3
Incerteza Técnica, Valor da Informação e Processo de Revelação

3.1.
Incerteza Técnica e Valor da Informação

3.1.1.
Introdução sobre Incerteza Técnica e a Teoria de Finanças

A incerteza técnica é aquela relacionada às características específicas de um projeto. Uma classe relacionada de incerteza é aquela descrita por Zeira (1987) como "incerteza estrutural", na qual a firma não conhece totalmente a sua própria função lucro e pode descobri-la através de investimentos. Dessa forma, "*investimento sob incerteza estrutural cria uma interação entre a acumulação de capital e a acumulação de informação*" (Zeira, 1987, p.204-205). Embora Zeira estivesse mais preocupado em efeitos macroeconômicos de acumulação de capital e em aplicações sobre incentivos para fazer testes de mercado para conhecer a sua função demanda, essa incerteza estrutural – tal como a incerteza técnica aqui focada, incentiva o investimento em processos de aprendizagem da função lucro.

Incerteza técnica tem um papel muito importante na valoração de muitos ativos especialmente em E&P em petróleo. Aqui, será focada a incerteza técnica sobre a existência, o volume e a qualidade de um campo de petróleo. Mas a metodologia pode ser estendida para incertezas técnicas de outras indústrias[104], por ex., em mineração, em P&D de inovação tecnológica (especialmente os realizados em fases, como na indústria farmacêutica), etc.

Em E&P de petróleo, na primeira fase exploratória (rever Figura 1), onde se tem a opção de perfurar o poço pioneiro, estão presentes os três tipos de incerteza técnica da reserva (existência, volume e qualidade), com destaque ao mais primário de todos, que é a existência de um campo de petróleo. Essa incerteza é

[104] Devem ser também observadas as características específicas que demandam adaptações.

modelada pelo fator de chance exploratório (distribuição de Bernoulli), que será particularmente discutida nesse capítulo (item 3.4).

Já nas fases de delimitação e desenvolvimento, as incertezas remanescentes são o volume e a qualidade. Em geral, a seqüência de investimento em exploração e explotação da reserva, e o histórico de produção, vai reduzindo essa incerteza.

Na fase de produção, em geral a incerteza técnica do reservatório ainda existe, mas é muito menor que antes do início do investimento e da produção. Mas mesmo assim essa incerteza técnica pode ser ainda suficientemente importante para justificar investimentos em informação adicional nessa fase, tais como a sísmica 4D[105] e conversão de poços convencionais em poços inteligentes.

Um dos principais objetivos dessa tese (senão o principal) é prover os modelos de OR de uma abordagem mais rigorosa – mas ao mesmo tempo prática, para tratar a incerteza técnica. Em contraste com a incerteza de mercado, a modelagem de incerteza técnica tem sido pobremente analisada na literatura de OR, com raras exceções. Assim, a necessidade motivou o autor para desenvolver novos caminhos para modelar a incerteza técnica visando aplicações de OR.

Dada a importância do tema, a literatura mais recente de OR em petróleo começa a abordar com freqüência os temas de incerteza técnica e processos de aprendizagem. Ver por ex., Chorn & Carr (1997); Chorn & Croft (2000); Dias (2001b); Gallant, Kieffel & Chatwin (1999); Galli, Armstrong & Jehl (1999); Galli, Armstrong & Dias (2004); Sharma et al. (2002); Whiteside & Drown & Levy (2001), entre outros. Como foi visto no cap. 2, incerteza técnica é o foco de várias aplicações de OR em petróleo tais como a avaliação econômica de poços inteligentes e de opção de expansão.

Berk & Green & Naik (2004) analisam a questão do prêmio de risco em projetos com incerteza técnica (P&D), mas considerando: (a) que o valor do projeto (ou da firma de P&D) tem componentes de risco não-sistemáticos (incerteza técnica), mas também components de risco sistemático (fluxos de caixa após a completação do projeto de P&D); (b) o efeito seqüencial das OR altera o prêmio de risco do projeto. Mas eles chegam a conclusões (p.3) já esperadas pela própria literatura de opções financeiras: o risco da opção de compra é maior do

[105] Análises de VOI na Petrobras em 2003 mais que justificaram os investimentos em sísmica 4D para vários campos marítimos.

que o risco do ativo básico para o caso de opção "out-of-money" e igual ao do ativo básico para opções "deep-in-the-money".

Na literatura teórica de finanças, um artigo famoso que já foi usado em aplicações de OR relacionadas à incerteza técnica é o de Merton (1987) sobre *mercado incompleto*. Mais recentemente Merton (1998) voltou a abordar o problema de mercado incompleto com um modelo de "tracking error" que segue um movimento Browniano. No entanto, essas abordagens parecem mais úteis do ponto de vista da decisão de um acionista na avaliação de uma ação com razoável risco específico (ex.: obter mais informações sobre uma firma pouco conhecida ou avaliar discrepâncias econométricas de retorno para obter algum retorno anormal) do que do ponto de vista de uma firma, que deve decidir se investe ou não em informação num projeto com incertezas técnicas bem específicas.

Num mercado incompleto existem ativos (ações de firmas) em que a informação disponível no mercado a respeito do seu retorno é incompleta, ou distribuída de forma assimétrica entre os investidores, ou flui numa velocidade menor que a requerida para um mercado eficiente e completo. Isso pode fazer com que os retornos esperados de certas firmas dependam não só do risco de mercado como também da variância total, fenômeno que é chamado por Merton (1987) de discrepância α_k da ação k (que é zero se o mercado é completo). Mas essa discrepância é um fenômeno relevante apenas para certos tipos de firma. Por ex.: "*firmas menores tendem a ter maior variância total e menor correlação de seus retornos com o mercado geral*", "*firmas menores tendem a ter muito menos acionistas que firmas grandes*", "*firmas mais conhecidas e com maior base de acionistas terão menores* α_k" (Merton, 1987, p.497).

Nessa tese, em que o foco está em aplicações em petróleo do ponto de vista de corporações com grande base de acionistas, e que em geral são firmas antigas e conhecidas no mercado, essa discrepância é muito pequena (ou muito menor que para firmas da "nova economia", por ex.) e assim será considerado a premissa razoável de que o mercado é suficientemente completo para se usar a teoria clássica de finanças. A visão de que riscos técnicos (ou únicos ou diversificáveis) não demandam prêmio de risco é assumida rotineiramente em livros texto de finanças corporativas tais como, por ex., Brealey & Myers (1999, p.167-169).

Proposição 4: A incerteza técnica não demanda prêmio de risco por parte de corporações com acionistas diversificados.

Prova: A teoria neoclássica de finanças, em especial a teoria de portfólio e o CAPM ("capital asset pricing model") mostra a distinção entre *risco diversificável* e *risco não-diversificável*. Pelo CAPM, a taxa ajustada ao risco $\mu = r + \pi$, onde π é o prêmio de risco demandado por investidores diversificados, dado pelo produto do fator β pelo spread de retorno do mercado ($r_M - r$), onde r_M é o retorno do portfólio de mercado e β a covariância do retorno do ativo com o retorno do mercado, normalizada (dividida) pela variância de mercado. Mas a incerteza técnica é *independente* da evolução do mercado (flutuações do mercado não afetam a probabilidade de ocorrência de petróleo, o volume, etc.) e, logo, tem *correlação zero* com o retorno do portfólio de mercado (é um *risco diversificável*). Assim, pela definição, tem-se $\beta = 0$ e, portanto $\pi = 0$. Logo, a incerteza técnica não demanda prêmio de risco por parte de investidores diversificados de corporações – como os acionistas de companhias de petróleo. □

A Proposição 4 tem importantes implicações na valoração de OR, especialmente nos métodos que usam a valoração neutra ao risco (Bingham & Kiesel, 1998), tipicamente aplicando a teoria dos martingales (Musiela & Rutkowski, 1997). Exemplos de métodos neutros ao risco são os métodos *binomial* (Cox, Ross & Rubinstein, 1979) e o da *simulação de Monte Carlo de processos neutros ao risco*[106]. Enquanto que os processos estocásticos são ajustados ao risco através da subtração de um prêmio de risco (ver a eq. 8), as distribuições de probabilidade da incerteza técnica não necessitam de nenhum ajustamento ao risco, pois o prêmio de risco requerido por investidores diversificados é zero.

Assim, a combinação de incertezas técnicas com incerteza de mercado em modelos de OR pode se dar de uma forma mais simples se for usado um método neutro ao risco. Especificamente, pode-se usar a taxa livre de risco r para descontar valores futuros, caso se trabalhe com processos estocásticos neutros ao risco (i. é, penalizados por prêmios de risco) para as variáveis de mercado, combinados com as distribuições de probabilidade que representam as incertezas técnicas, as quais já são *naturalmente* neutras ao risco.

[106] Já o método da EDP (ver Proposição 3) parte de processos reais e não neutros ao risco. Lá a neutralização ao risco é feita depois, através da construção de um portfólio livre de risco.

O fato da incerteza técnica não demandar prêmio de risco é reconhecido não só na teoria como na prática das empresas de petróleo. Por ex., num mesmo país, as companhias de petróleo usam a mesma taxa de desconto ajustada ao risco para projetos de exploração (que têm elevada incerteza técnica) e para projetos de desenvolvimento da produção (que têm incerteza técnica muito menor). No entanto, muitos gerentes e praticantes de análise econômica ainda confundem esse ponto, achando que essa incerteza deveria ser penalizada na taxa de desconto e que não penalizá-la seria ignorar esse risco. Essa seção introdutória tem a intenção de esclarecer esse ponto.

Existe também o problema conhecido na literatura como *agente x principal* ou problema de *agência*, indesejável, mas comum em corporações. O problema é que o gerente (agente) nem sempre está alinhado com os acionistas (principal). Em geral o gerente não é diversificado como os acionistas e assim não representa de forma perfeita os seus interesses. A *assimetria de informações* (agente é mais bem informado), dificulta o principal a monitorar o agente. No caso da incerteza técnica, o gerente tem a tendência de aversão maior por não ser diversificado e isso pode levá-lo a considerar suas preferências ao risco técnico nas suas decisões de investimento em detrimento das preferências agregadas dos acionistas (que em geral podem ser medidas no mercado). Isso é um desvio que é enfrentado através do desenho de incentivos adequados, ver por ex. Mas-Colell, Whinston & Green (1995). Mas a discussão de prêmio de risco não é tudo no caso da incerteza técnica e os gerentes têm um papel fundamental para maximizar o valor do acionista com o gerenciamento da incerteza técnica.

Não demandar prêmio de risco é um importante e incontestável resultado da teoria neoclássica de finanças corporativas, mas é apenas *um dos aspectos* do papel da incerteza técnica na valoração de ativos. Não demandar prêmio de risco não quer dizer que a incerteza técnica não seja importante ou que seja menos importante que a incerteza de mercado. Os acionistas esperam que os gerentes otimizem o valor da firma inclusive através do gerenciamento ótimo da incerteza técnica. Ao contrário dos acionistas, os gerentes podem fazer muito melhor do que apenas *diversificar*, eles podem *alavancar* o valor da firma através do gerenciamento ótimo da incerteza técnica. Em outras palavras, uma teoria sobre o prêmio de risco, como o CAPM, é válida e necessária, mas *não é suficiente* para maximizar o valor da firma sob incertezas.

A incerteza técnica em projetos de investimento tem dois lados. Um é o lado da *ameaça*, pois a incerteza técnica leva quase certamente a decisões sub ótimas de investimento. O outro lado é o da *oportunidade*, pois cria oportunidades de investimento em informação, em que o exercício de *opções de aprendizagem* ("learning options") pode ser muito valioso.

O lado negativo da incerteza técnica é que, apenas por sorte, o projeto otimizado *ex ante* se mostrará também otimizado na realidade *ex post*. Para ver isso, suponha que existe uma distribuição de probabilidades representando a incerteza do volume de reservas B. Suponha que o investimento ótimo (quantidade de poços, capacidade de processo, diâmetro de dutos, etc.) para desenvolver o campo foi calculado usando E[B] (o valor esperado minimiza o erro). Suponha que o investimento ótimo é uma função monotônica crescente de B, por ex. linear como na eq. (28). Assim, se ex post o cenário verdadeiro revelado de B for maior que E[B], então o investimento será *insuficiente* (sub-ótimo), enquanto que se o verdadeiro cenário revelado for B menor que E[B], então o investimento será *excessivo* (sub-ótimo também) para o tamanho da reserva. Assim, se a distribuição de probabilidades de B é contínua, então se pode dizer que a incerteza técnica leva *quase certamente*[107] a investimentos sub-ótimos.

Além disso, a incerteza técnica pode levar tanto ao exercício prematuro da opção (quando, para o *verdadeiro* cenário técnico, o ótimo seria esperar) como ao não exercício da opção quando o melhor seria exercê-la de imediato (para o *verdadeiro* cenário técnico de volume e qualidade da reserva). Assim, a incerteza técnica diminui o valor do projeto por levar a decisões sub ótimas de investimento e não devido a uma aversão ao risco técnico ou devido à "função utilidade do gerente", como alegado por alguns pesquisadores da escola tradicional de *análise de decisão*, que ignoram conceitos básicos da teoria de finanças corporativas, em especial a distinção de riscos diversificável e não-diversificável.

O investimento em informação pode alavancar o valor do projeto, pois pode levar a melhorar bastante a rentabilidade do projeto em caso de informações favoráveis e evitar fazer um mau projeto (ou de investir em excesso) em caso de informações desfavoráveis. Ou seja, em ambos os casos a informação é valiosa.

---

[107] *Quase certamente* significa que só não é válida em um conjunto de *medida* (de probabilidade) *igual a zero*. No caso, só é ótimo em *um* cenário num universo de *infinitos* cenários.

Assim a incerteza técnica provoca dois efeitos de sentidos opostos no valor de projetos. Em especial, o lado do benefício pode ser mais bem avaliado com métodos de OR, que valoriza a flexibilidade de resposta com a chegada de novas informações. Nesse contexto, surge a chamada análise de valor da informação, a ser vista a seguir, inicialmente desde um ponto de vista tradicional, e depois visto de uma nova forma para se adaptar ao contexto mais dinâmico de OR.

3.1.2. Introdução à Análise de Valor da Informação

A análise de valor da informação (VOI ou "value of information") é bem antiga (ver item 3.1.3 para revisão da literatura) e uma das principais aplicações da escola de *análise de decisão* (Raiffa, 1968). Um problema de decisão é definido pelo espaço de ações disponíveis e pela função que relaciona o resultado ("payoff") a essas ações e aos estados da natureza (Arrow, 1992, p.169).

Problemas de valor da informação são de grande importância não só em aplicações econômicas, mas também em várias outras áreas tão distintas como medicina, P&D (pesquisa e desenvolvimento) em geral, e até no dia a dia das decisões individuais. Por isso a literatura de aplicações de VOI ocorre em diversas disciplinas. A seguinte definição do objetivo da análise de valor da informação é tirada da literatura de *saúde ambiental*, Yokota & Thompson (2004), exceto o grifo, mas é suficientemente geral para ser válida também em outros contextos tais como o econômico: "*A análise de valor da informação avalia o benefício de coletar informação adicional para reduzir ou eliminar incerteza num contexto de tomada de decisão específico*".

O grifo na definição em "reduzir ou eliminar incerteza" destaca o fato que essa definição sugere que uma análise de VOI demanda medidas de redução de incerteza (medidas de aprendizagem) e também a modelagem de como um processo seqüencial de aquisição de informação pode resultar na eliminação total, ou parcial, da incerteza. Essas demandas serão especialmente atendidas nesse capítulo com uma metodologia inovadora sobre *medidas de aprendizagem* e *processos de revelação*. Esses conceitos serão diretamente relacionados com o conceito de *expectativa condicional* e será, portanto, de grande aplicabilidade em finanças e em outras disciplinas, ao simplificar de forma adequada a solução de problemas de VOI.

Embora o desenvolvimento teórico de análise de VOI tenha se iniciado nas décadas de 50 e 60, sua aceitação e/ou aplicação ainda é bastante limitada especialmente devido à complexidade tanto na sua modelagem como na solução de problemas de VOI, (Yokota & Thompson, 2004, abstract). Esses autores ainda ressaltam a grande complexidade prática de resolver problemas de VOI trabalhando com distribuições contínuas (como inputs) "*apesar da simulação permitir o analista a resolver problemas mais complexos e realistas*". Essas mesmas preocupações são compartilhadas nessa tese[108] e é a motivação de alguns conceitos e proposições. Por exemplo, as proposições sobre *distribuições de revelações* irão facilitar tanto o uso de simulação de Monte Carlo em problemas de VOI com qualquer tipo de distribuição (exige-se apenas média e variância finitas) bem como a integração do VOI em modelos clássicos de opções reais. Isso também será objeto desse cap. 3.

O VOI é sempre avaliado ex-ante porque ele é calculado antes de decidir se investe ou não em informação. Ou seja, é necessário avaliar o VOI antes, a fim de decidir se o benefício da informação supera (o suficiente) o custo de adquiri-la.

Informação é em geral valiosa na presença de incerteza. Conforme Arrow (1973, p.138), "*quando existe incerteza, existe usualmente a possibilidade de reduzi-la através da aquisição de informação. Logo, informação é a medida negativa de incerteza*" (grifos dessa tese). Aqui será particularmente explorado esse tema ligando o valor da informação com a redução esperada da incerteza (ver item 3.3 sobre medidas de aprendizagem).

A seguinte definição é adequada no contexto de valor econômico da informação (Lawrence, 1999, p.2):

Definição. **Informação**: é qualquer estímulo que muda o conhecimento do receptor da informação, i. é, que mude a distribuição de probabilidade do receptor a respeito de um bem-descrito conjunto de estados. Por razões de conveniência para o comportamento matemático, será incluído nessa definição o caso trivial do *limite* de pouca relevância da informação, limite esse chamado *não-informativo*.

Naturalmente, a definição acima engloba o caso de *limite* oposto, de muita relevância, em que a distribuição de probabilidade do receptor muda radicalmente,

[108] Apesar dos grandes avanços nos últimos 3 anos na introdução de metodologias de valor da informação na Petrobras (particularmente no E&P), a complexidade de análises mais realistas é o grande desafio que essa tese procura lidar.

colapsando em um ponto de probabilidade 1. Esse caso será chamado de *informação perfeita*, que permite a *revelação total* do verdadeiro estado da natureza. Esse limite será discutido em detalhes, devido à sua importância.

A distinção entre informação e *conhecimento* pode ser mais bem entendida no contexto de uma analogia com a *teoria econômica do capital* feita por Boulding (1966) e discutida em Lawrence (1999, p.3). Capital físico é *estoque* (ativo) e investimento é *fluxo* que, quando ocorre, muda o estoque de capital. Da mesma forma, conhecimento é estoque e informação é o fluxo que altera o conhecimento. Esse conhecimento é descrito por uma distribuição de probabilidades e a (nova) informação altera essa distribuição e o conhecimento. Também de forma análoga à *depreciação* de capital, está o *esquecimento* do conhecimento. Em ambos os casos, pode haver um custo para repor o capital ou para manter o conhecimento.

Seguindo Lawrence (1999, p.5), num problema de decisão sob incerteza, a distribuição a priori engloba todo o conhecimento inicial do decisor em relação à realização de cada estado. Uma *mensagem* ou *sinal* é a forma final de saída de uma *fonte de dados*. Por ex., o médico é a fonte de dados e a mensagem/sinal é o seu diagnóstico. Como em Arrow (1992), aqui se usará preferencialmente o termo "sinal". Dados são símbolos, imagens, sons e idéias que podem ser codificados, estocados e transmitidos. Um sinal pode levar ao decisor alterar a distribuição a priori através da eliminação de alguns estados, redução da probabilidade de outros estados e aumento da probabilidade dos restantes. O uso do sinal para atualizar o conhecimento gera a distribuição posterior.

Esse autor faz também a importante distinção entre *informação estatística* e *informação pragmática*. A primeira é relativa somente às propriedades das distribuições de probabilidade envolvidas e a segunda é a aplicação da primeira, ou seja, o impacto potencial da informação estatística em um problema de decisão. Os atributos da primeira são, por ex., coerência, formato e acurácia, enquanto que na segunda os atributos são, por ex., relevância, completitude e tempo de ocorrência. Uma mensagem pode trazer informação estatística, mas não trazer informação pragmática, ou seja, não ter relevância para alterar as decisões.

Um *modelo probabilístico de aprendizagem* trabalha com o relacionamento de variáveis aleatórias de interesse (vetor de estados **X**) com variáveis aleatórias que provém informação ou *sinais* (vetor **S**) para se conhecer melhor X. Assim, o

modelo probabilístico de aprendizagem trabalha com *distribuições a priori* de variáveis aleatórias G(**X**), *distribuições conjuntas* de variáveis aleatórias J(**X**, **S**) e *distribuições condicionais* de variáveis aleatórias H(**X** | **S**). Por isso, é necessário discutir aspectos teóricos desses relacionamentos de variáveis aleatórias para poder elaborar uma metodologia consistente de OR envolvendo incerteza técnica e investimento em informação.

Dado o conhecimento inicial representado pela densidade de probabilidades p(x), o conjunto de decisões ou ações $\mathcal{A}$ e a função valor ("payoff") que aqui é dada pela opção real F(**X**, t, a), onde a $\in \mathcal{A}$, o VDP, valor da decisão a priori (sem a informação) é aquele obtido com a ação que maximiza o valor esperado da OR:

$$\mathbf{VDP = Max_a\{E[F(X, t, a)]\} = Max_a \{\int_X F(x, t, a)\, p(x)\, dx\}} \qquad \mathbf{(52)}$$

Ou seja, é necessário testar todas as possíveis decisões e escolher aquela que maximiza o valor esperado da OR, o qual é função do vetor de variáveis aleatórias **X** (incertezas técnica e/ou de mercado) e do tempo t,

Já o valor esperado da decisão informada (VDI) considera que, para cada possível sinal s $\in$ S, o decisor irá tomar a ação ótima (como antes) e assim o VDI considera o valor esperado (em relação a S) das ações que maximizam a OR dada a informação revelada. Ou seja:

$$\mathbf{VDI = E[Max_a\{E[F(X, t, a) \mid S]\}] = \int_S Max_a\{\int_X F(x, t, a)\, p(x \mid s)\, dx\}\, p(s)\, ds} \quad \mathbf{(53)}$$

A eq.(52) e a eq.(53) consideram que as distribuições p(x), p(s) e p(x | s) são contínuas, mas o conceito vale também para distribuições discretas (usa-se somatório em vez da integral) e para casos mistos que combinem distribuições discretas e contínuas (usa-se a integral de Lebesgue-Stieltjes[109], em vez da integral de Riemann). Nas aplicações, como é usual na literatura de OR e também para simplificar a notação, em geral será subentendido que o valor da OR já considera a ação ótima sob incerteza (em muitos casos a decisão binária de exercer ou não

[109] Uma maneira intuitiva de ver a diferença entre essas integrais é que as somas de Riemann são feitas com retângulos *verticais* e as somas de Lebesgue são feitas com retângulos *horizontais* (Shorack, 2000, Figure 1.1, p.2). A integral de Lebesgue garante a convergência geral de integrais de funções quando $f_n \rightarrow f$, enquanto na integral de Riemann isso nem sempre ocorre.

uma opção) e assim será suprimido da notação o operador de maximização e também será usado simplesmente F(X, t) em vez de F(X, t, a).

Em geral o VOI é a diferença entre o valor da decisão informada e o valor da decisão sem essa informação (ou a priori), i. é[110]:

$$\mathbf{VOI = VDI - VDP} \qquad \mathbf{(54)}$$

Essa diferença é sempre positiva ou zero para problemas sem interação estratégica, i. é, para um único decisor maximizador de riquezas. É fácil ver isso, pois com mais informação o decisor pode escolher tanto as ações que ele faria sem a informação como ações mais adequadas à nova informação, sendo que essas últimas só seriam escolhidas se forem melhor que as outras ações sob o ponto de vista de maximização de riquezas.

Já para múltiplos decisores, como na teoria dos jogos, existem casos em que mais informação *diminui* o valor desses jogadores. Um exemplo clássico é o mercado de seguros de automóveis (e de seguros em geral). Se tanto o segurador como os segurados tivessem, antes de assinar o contrato, acesso à informação *perfeita* ("bola de cristal") sobre o que ocorrerá com o bem segurado no período de contrato (sinistro ou não-sinistro), então um dos dois lados não iria querer assinar o contrato de seguro. Isso faria desaparecer o mercado de seguros, o que seria pior tanto para a seguradora (deixaria até de existir) como para os consumidores que demandam seguro (que não poderiam fazê-los). Assim, essa informação perfeita seria pior para ambas as partes.

Segundo Hilton (1981), os quatro determinantes do valor da informação são:

1. A estrutura do conjunto de ações, i. é, a *flexibilidade* de ação;
2. A estrutura da função valor resultante ("payoff"). Tradicionalmente, é a função valor monetário e sua relação com a *função utilidade*;
3. O grau de incerteza inicial, dado pela *distribuição a priori*; e
4. A percepção do decisor para o *mapeamento dos sinais* S para os estados da natureza da variável de interesse X. Tradicionalmente, é a função verossimilhança, i. é, a distribuição $g(s|x)$ como função de x.

[110] Assim como em geral se faz na literatura tradicional de VOI, inicialmente não está sendo considerado o tempo para adquirir a informação. Mas no modelo dinâmico será considerado o *tempo de aprendizagem*, para poder comparar as alternativas de investimento em informação, uma vez que em finanças o valor do dinheiro no tempo é muito importante.

A metodologia dessa tese concorda com a importância da lista acima, mas com algumas importantes alterações. Os itens 1 e 3 terão um papel mais importante do que tem normalmente na literatura tradicional. A questão do valor da flexibilidade não tem sido explorada pela literatura de análise de decisão como tem sido pela literatura de OR, especialmente em contextos dinâmicos. Na análise de decisão tradicional, raramente as aplicações envolvem a variável tempo. Processos estocásticos e controle ótimo estocástico também em geral não fazem parte da literatura de análise de decisão[111]. O papel da distribuição a priori será maior do que na literatura tradicional tanto por causa da aplicação em petróleo – onde o conhecimento a priori é muito importante e tem de ser bem modelado, como por causa da metodologia aqui usada, onde a distribuição a priori é usada como limite de um processo de revelação.

Já nos itens 2 e 4 da lista de Hilton (1981) acima, a palavra "tradicionalmente" foi usada para indicar que nessa tese, em contraste, será seguido um caminho diferente. No caso do item 2, em vez da função utilidade de valores monetários, será usado o valor da opção real para calcular os "payoffs" do valor da informação[112]. Essa combinação de teorias (OR + VOI) está dentro do espírito *híbrido* dessa tese, mas não seria a primeira vez que se faria.

Para o item 4 essa tese apresenta uma *novidade metodológica*, em que será feita uma análise não usual em problemas de VOI para a relação entre as variáveis aleatórias de interesse (X) e os sinais (S). Aqui não será preciso usar a função verossimilhança. Em vez disso, jogarão um papel maior as distribuições a priori e a de expectativas condicionais (ou de *revelações*). Será mostrado que a verossimilhança não é uma medida de aprendizagem adequada, sendo melhor para isso uma medida de *redução esperada de variância* que, além de melhores propriedades para esse tipo de aplicação, é mais prática, intuitiva e de fácil obtenção com os mais populares métodos estatísticos, desde as simples regressões, até métodos um pouco mais sofisticados como os de análise de

[111] Uma notável exceção é o relativamente recente livro de Bather (2000).

[112] Não que as funções utilidade e opções reais sejam excludentes em geral. No cap. 6 será visto aplicações em que se usa a função utilidade para obter o valor da OR, por ex., no desenho de *OR para o consumidor*, onde é necessário avaliar as utilidades do tipo específico de consumidor que se busca atingir. Mas na maioria das aplicações, OR usa valores de mercado refletindo preferências agregadas em relação ao risco e retorno dos investidores.

variância (ANOVA). Será mostrado que essa medida de aprendizagem está diretamente ligada à distribuição de revelações.

A literatura distingue dois tipos de VOI, o *valor esperado da informação perfeita* (VEIP) e o *valor esperado da informação imperfeita* (VEII).

Dados os valores das opções reais $E[F(\mathbf{X}, t) \mid S_{perf}]$ e $F(\mathbf{X}, t)$, para os casos de valor esperado de OR com a informação perfeita e OR sem a informação (ou incondicional, ou antes da informação), então o VEIP é obtido por:

$$\mathbf{VEIP = E[F(X, t) \mid S_{perf}] - F(X, t)} \quad \mathbf{(55)}$$

O caso de informação imperfeita é mais complicado, mas bem mais relevante na prática por ser mais realista. Naturalmente o VEII é menor ou igual que o VEIP. De forma análoga ao caso do VEIP, o VEII é dado por:

$$\mathbf{VEII = E[F(X, t) \mid S_{imperf}] - F(X, t)} \quad \mathbf{(56)}$$

Nos capítulos 3 e 5 será detalhado o cálculo dos termos das equações (55) e (56). O exemplo a seguir será útil inicialmente para introduzir uma série de conceitos importantes de uma maneira simples e intuitiva, para poderem depois ser usado em casos bem mais complexos. Depois ele será usado para exemplificar o cálculo do VEIP e as suas dificuldades e características.

Seja um caso simples de um campo delimitado, mas não desenvolvido, com incerteza residual sobre o volume de reservas B. Suponha que existam duas alternativas, a primeira é desenvolver o campo imediatamente e a segunda é investir previamente em informação. Com a primeira alternativa, para minimizar o erro, o investimento (capacidade de processo, etc.) é dimensionado para o valor esperado do volume E[B], pois a teoria elementar de probabilidade ensina que dessa forma se minimiza o erro de dimensionamento. Na segunda alternativa, suponha que o investimento em informação *revela toda a verdade* a respeito do volume da reserva B, i. é, a nova informação é perfeita. Até quanto se pagaria por essa informação perfeita, i. é, qual é o VEIP?

Para simplificar esse exemplo, suponha que a incerteza em B é representada por apenas três cenários equiprováveis B^+, $B^=$ e B^-, ou seja, se tem uma *distribuição a priori* de B discreta e com três cenários. Se a informação é perfeita então o investimento em informação provoca uma *revelação total* ("full revelation") de B. Nesse caso então, um desses três cenários será revelado como o verdadeiro valor de B, i. é, B^+, $B^=$ e B^- são, respectivamente, (o verdadeiro valor

de) B dado boas notícias, B dado notícias neutras e B dado más notícias. A Figura 27 ilustra esse exemplo de informação perfeita que provoca a revelação total de B.

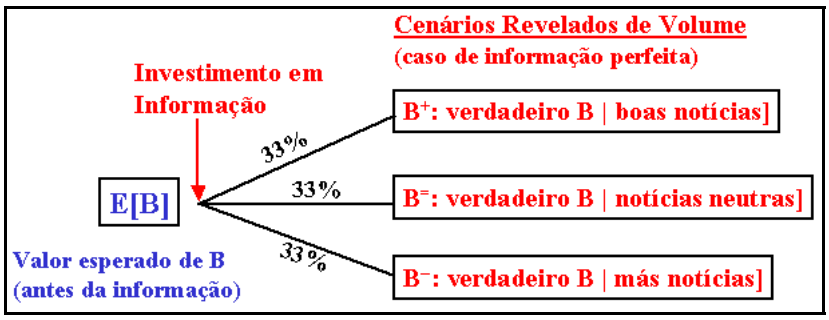

Figura 27 – Exemplo de Informação Perfeita ou Revelação Total

Dessa forma, fica claro que, por consistência, a distribuição de cenários revelados com a informação tem de ser igual à distribuição a priori, *se* a informação é perfeita (revelação total). Se não fosse assim, por ex., se o analista achar que o investimento em informação pode revelar um cenário que não está na distribuição à priori, então ele deveria mudar a distribuição a priori para incluir esse cenário. A definição seguinte formaliza esse ponto.

Definição. **Distribuição a priori**: representa, em forma de uma distribuição de probabilidades, todo o conhecimento inicial (a priori) que o decisor tem de uma variável aleatória (baseado em Lawrence, 1999, p.5)[113]. Ou seja, o suporte[114] da distribuição à priori inclui todos os possíveis valores (cenários) que pode assumir essa variável, enquanto que a densidade de probabilidade representa a melhor estimativa possível da densidade de probabilidade de ocorrência desses cenários, usando o conhecimento corrente (a informação a priori).

Definição. **Distribuição posterior** (ou distribuição a posteriori): é a distribuição a priori atualizada com o sinal ou (nova) informação. Expressa o que é conhecido da variável *depois* de observar o dado (O'Hagan, 1994, p.10).

Assim, se a informação é perfeita, então a distribuição posterior é um cenário ou número ("singleton") que ocorre com probabilidade igual a 1 (Lawrence, 1999, p.69). Nesse caso, a distribuição posterior é dita degenerada. Ou seja, a informação perfeita faz a variância da distribuição posterior colapsar para zero. Nesse exemplo, existem três distribuições posteriores possíveis (de variância

[113] Equivale à definição de O'Hagan (1994, p.10): A distribuição a priori de X expressa o que é conhecido sobre X, antes de observar o dado.

[114] Suporte de uma distribuição p(x) é o conjunto de valores onde p(x) > 0.

zero). Para distribuições a priori contínuas, a quantidade de distribuições posteriores é infinita.

Já a distribuição de cenários revelados com a informação, mostrada na Figura 27, é única e será aqui chamada de distribuição de revelações (a ser definida num contexto mais geral). Note que, se a distribuição a priori fosse contínua, a distribuição de revelações continuaria sendo única, embora com infinitos cenários (distribuição contínua) em vez de três. Esses pontos serão retomados num contexto mais geral.

Com a discussão acima fica claro que, em caso de informação perfeita (revelação total), a distribuição de revelações é igual à distribuição a priori. Isso decorre diretamente da definição de distribuição a priori. No exemplo da Figura 27, p(B) é uma distribuição a priori discreta com três cenários equiprováveis que nesse caso de informação perfeita é igual à distribuição de revelações.

Formalmente, seja um espaço de estados $\mathbf{X} = \{x\}$ de interesse (que pode ser enumerável ou não-enumerável), descrito por uma distribuição a priori de probabilidades p(x), e um espaço de mensagens (ou de sinais) $\mathbf{S} = \{s\}$.

Se a informação é perfeita então o espaço de mensagens é igual ao espaço de estados, i. é, $\mathbf{X} = \mathbf{S}$ (Lawrence, 1999, p. 69). Para ver isso, considere a Figura 27. Nesse exemplo $\mathbf{X} = \{ B^{+}, B^{=}, B^{-}\}$, isto é, os volumes de reservas possíveis de ocorrer. Se a informação é perfeita, $\mathbf{S} = \{ B^{+}, B^{=}, B^{-}\}$ também pois a informação revela o verdadeiro estado da natureza de X (ou seja, uma distribuição posterior degenerada de B). Se a informação fosse imperfeita, então a mensagem revelaria uma distribuição não degenerada, i. é, uma distribuição posterior p(x | s) com variância estritamente positiva.

Defini-se informatividade como a qualidade de um sinal em ser informativo. Seja θ uma medida genérica de grau de informatividade de uma estrutura de informação. No caso de aplicações estatísticas como a de *planejamento de experimentos* ("experimental design"), θ é o tamanho da amostra. Nos casos clássicos de VOI, se usa muito a função verossimilhança. Nessa tese será usada uma medida de aprendizagem relacionada com a redução esperada de incerteza. A informatividade estatística varia desde o caso $\theta = 0$ (nenhuma informação) até o caso de informação perfeita (máxima informatividade). No interior desse intervalo se tem a chamada informação imperfeita.

Definição. **Estrutura de informação**: compreende o espaço de mensagens (sinais) mais a medida conjunta de estados e mensagens (Lawrence, 1999, p.16)[115]. Assim, uma estrutura de informação $\mathcal{I}$ é definida por:

$$\mathcal{I} = \{ S, p(x, s)\} \quad (57)$$

Essa medida conjunta é a *distribuição conjunta de probabilidades* de duas variáveis aleatórias p(x, s). Isso sugere que é necessário estudar distribuições bivariadas e multivariadas de probabilidades para a análise de VOI. A definição acima sugere também uma *comparação de estruturas de informação*, para determinar em que medida existe uma estrutura mais informativa que a outra. Esse foi o tema dos dois artigos clássicos de Blackwell (1951) e (1953), ver também Blackwell & Girshick (1954). Esse tema será comentado no item 3.1.4.2.

Uma estrutura de informação é completamente não-informativa se a distribuição de probabilidades condicional do sinal g(s | x) é a mesma para todos os estados da natureza (Radner & Stiglitz, 1984, p. 34). Ou seja, com uma estrutura não-informativa, o decisor toma a mesma decisão independentemente do sinal recebido, já que qualquer que seja o verdadeiro valor de X, o sinal é o mesmo. Uma outra maneira de definir estrutura não-informativa é através da distribuição posterior, i. é, p(x | s) = p(x) se a estrutura é não informativa.

Para calcular o VEIP do caso simples mostrado na Figura 27, considere a equação do VPL de desenvolvimento, eq. (1), o modelo de negócios, eq. (23), e a equação do investimento (ótimo) em função do volume da reserva, eq. (28). Combinando essas três equações, se obtém a seguinte equação do VPL:

$$\mathbf{VPL = V - I_D = (q\ P\ B) - (c_f + c_v\ B)} \quad (58)$$

Se a informação é perfeita, então ela permite conhecer o verdadeiro valor de B nesse exemplo e assim dimensionar otimamente o investimento, conforme a eq. (28). Em geral, para cada B se faria um projeto que, além de dimensionar o investimento ótimo (eq. 28), permitiria calcular o novo valor de q através do fluxo de caixa, conforme descrito no cap. 2 (ver comentário após a Figura 13). Assim, em geral q é uma função de B. Tentar descobrir essa função é factível, mas muito trabalhosa, pois teria de fazer vários projetos (para vários B), estimar os fluxos de

[115] Radner & Stiglitz (1984, p.34) definem estrutura de informação como a matriz de Markov de funções (distribuições) verossimilhança g(s | x). No contexto da tese, onde a função verossimilhança será desnecessária, ficará claro que a definição de Lawrence é mais adequada.

caixa *de cada um deles*, para depois calcular os vários q e assim obter a função q(B), dado que o investimento ótimo é obtido com eq. (28).

Para manter o exemplo o mais simples possível, suponha que o valor de q é independente de B *se* o dimensionamento da capacidade for feito adequadamente, i. é, o valor de q é constante para todos os cenários de B se o investimento for dimensionado otimamente usando a eq. (28). Esse ponto será detalhado depois, mas a idéia é simples. Como q dá a velocidade com que a reserva B é produzida, então se a planta de processamento estiver dimensionada adequadamente, ela não será restrição e os poços poderão produzir o seu potencial. Entretanto, se a planta estiver sub-dimensionada, essa velocidade de produção da reserva B será menor devido à restrição de capacidade que impedirá que os poços produzam o seu potencial máximo. Essa redução de velocidade da produção de B irá reduzir o valor presente das receitas e por conseqüência o valor de q.

Dimensionar otimamente só é plenamente possível se a informação sobre B for perfeita. Ou seja, para se usar a eq. (28) é necessário saber o verdadeiro valor de B. A falta dessa informação (i. é, se B é incerto), leva a usar E[B] no lugar de B como dado de entrada na eq. (28), já que E[B] minimiza o erro técnico de dimensionamento[116]. Dessa forma, sem informação perfeita, o investimento será dado por:

$$\mathbf{I_D = c_f + (c_v\ E[B])} \qquad \mathbf{(59)}$$

Mas nesse caso (sem informação perfeita sobre B), nos cenários ex-post em que B é maior que E[B], a capacidade de processo será uma restrição e assim q será menor. De forma geral, se B é incerto então $E[q\ B] \leq E[q]\ E[B]$, e logo:

$$\mathbf{E[V] = E[q\ P\ B] \leq E[q]\ E[B]\ E[P]} \qquad \mathbf{(60)}$$

Dessa maneira, sem a informação, nos cenários de boas notícias (B^+), o investimento é o mesmo (dimensionado para E[B]), mas o valor de V é penalizado devido à restrição de capacidade que reduz o valor presente das receitas e, portanto, reduz o valor de q. Já no cenário de más notícias, V em geral não é penalizado, mas perde-se VPL devido ao investimento I_D maior que o necessário.

Pode-se também criticar a eq. (23) no contexto de incertezas argumentando que B é função de P, ou seja, se P for maior então a data esperada de abandono é

[116] Eventualmente pode ser ótimo *economicamente* dimensionar a planta de processo para um valor um pouco diferente de E[B], a depender de um estudo detalhado. Na falta desse estudo econômico, recomenda-se usar E[B] para minimizar o erro *técnico* de dimensionamento.

postergada, aumentando B. Em alguns casos isso não ocorre porque existe uma data legal de final da concessão da fase de produção. Mas o mais importante é que essa variação, na imensa maioria dos casos, só ocorre em fluxos de caixa (produção) muito distantes da data em que os fluxos de caixa são avaliados (geralmente entre 25 e 30 anos), e assim o impacto em V (que está em valor presente na data de início de investimento) é muito pequeno e não seria bem capturado pela eq. (23) a menos que se reduzisse o valor de q, fazendo q(B(P)). Assim, para usar a eq. (23) no contexto de incertezas, comete-se um erro menor se for considerado que B não é função do preço P. Já a restrição de dimensionamento tem impacto logo no início de produção (o pico de produção de um projeto de desenvolvimento em geral ocorre em um ou dois anos) e assim é muito importante considerar esse efeito para problemas de decisão de investimentos em desenvolvimento da produção, ao contrário do efeito de P sobre B.

Para levar em consideração o problema de dimensionamento, i. é, q(B), será introduzido um fator multiplicativo $\boldsymbol{\gamma_+(B) < 1}$ para corrigir o valor de q e logo o valor de V em cenários que a capacidade limitada se torna uma restrição técnica que reduz V, ou seja:

$$\mathbf{V = \gamma_+ q P B} \quad , \qquad \text{se } B > E[B] \tag{61}$$

No exemplo simples em que apenas um cenário é maior que E[B], será usado apenas um valor para $\gamma_+ = 0{,}8$. No caso mais geral esse fator depende da diferença entre B e E[B], i. é, quanto maior for essa diferença, mais severo (menor) deve ser esse fator, pois mais sub-dimensionada estará a capacidade de processo e o investimento. O caso mais geral e como calcular com simulações de FCD será visto depois. Para o caso do cenário revelado de B ser exatamente E[B], esse fator claramente é igual a 1. Já no caso do cenário revelado de B ser *menor* que E[B], não haveria restrição de capacidade, ao contrário. O excesso de capacidade, no entanto, não aumentaria o valor presente da produção e, sendo assim, se assume que o um fator de correção γ_- seria igual a 1[117].

Com isso, já é possível calcular o VEIP do exemplo simples da Figura 27. Suponha nesse exemplo simples que a decisão é "agora ou nunca" t = T, ou seja,

[117] Está se assumindo que a planta de processo é dimensionada para atender ao *pico* de produção. Se o ótimo econômico mostrar que o melhor é dimensionar a planta com um pouco de restrição (para reduzir um pouco os investimentos), o fator γ_- pode ser um pouco *maior* que 1, pois aliviaria a restrição de capacidade nos cenários em que B se revelar *menor* que E[B].

F(B, T) = Max(E[VPL], 0). Assuma também os valores numéricos P = 25 US$/bbl, q = 20%, c_1 = 180 MM US$, c_2 = 2 US$/bbl, { B^+, $B^=$, B^-} = {150, 100, 50} MM bbl; e γ_+ = 0,8. Os valores do VPL nos diversos cenários revelados de B são calculados com a eq. (58) para o caso com informação perfeita. Já os VPLs dos cenários de B para o caso sem informação, são calculados com a eq. (1) acoplada com a eq. (59) para o investimento, com a eq. (28) para os casos de cenários de B ≤ E[B] e com a eq. (61) para o caso de cenário de B > E[B]. A Tabela 8 a seguir sumariza os cálculos do VEIP, mostrando os valores de V, I_D e VPL tanto para o caso condicional à informação perfeita como para o caso incondicional (ou sem a informação).

Tabela 8 – Cálculo dos VPLs Com e Sem a Informação Perfeita (em MM US$)

Cenários de B		Com Informação Perfeita			Sem Informação		
Prob.	MM bbl	V | B	I_D | B	Max[VPL | B, 0]	V(E[B])	I_D(E[B])	VPL ex-post
33,3%	B^+ = 150	750	480	270	600	380	220
33,3%	$B^=$ = 100	500	380	120	500	380	120
33,3%	B^- = 50	250	280	0	250	380	– 130
E[F(B, T) | B] =				**130**	**F(B, T) =**		**Max[70, 0]**

Aplicando a eq. (55) nesse exemplo simples, o valor esperado da informação perfeita VEIP é simplesmente:

VEIP = E[F(B, t) | B] – F(B, t) = 130 – 70 **= US$ 60 MM**.

Assim, nesse exemplo se pagaria até US$ 60 MM para obter a informação perfeita. Esse exemplo mostrou que existem sutilezas para calcular o VOI mesmo nos casos mais simples, com poucos cenários, informação perfeita e sem ainda considerar interações com incertezas de mercado, valor da espera da opção, etc. Em geral, otimizar sob certeza é muito mais fácil do que otimizar sob incerteza.

Considere agora o caso mais realista de aquisição de informação imperfeita. Seja o exemplo anterior baseado na Figura 27, mas agora com a informação imperfeita ou revelação parcial do verdadeiro valor de B. A Figura 28 representa esse caso, mostrando que a informação revela cenários ainda incertos que podem ser representados por um conjunto de distribuições posteriores ou de forma mais simples pelos seus valores esperados, através de uma distribuição de expectativas condicionais, que aqui será chamado de *distribuição de revelações*.

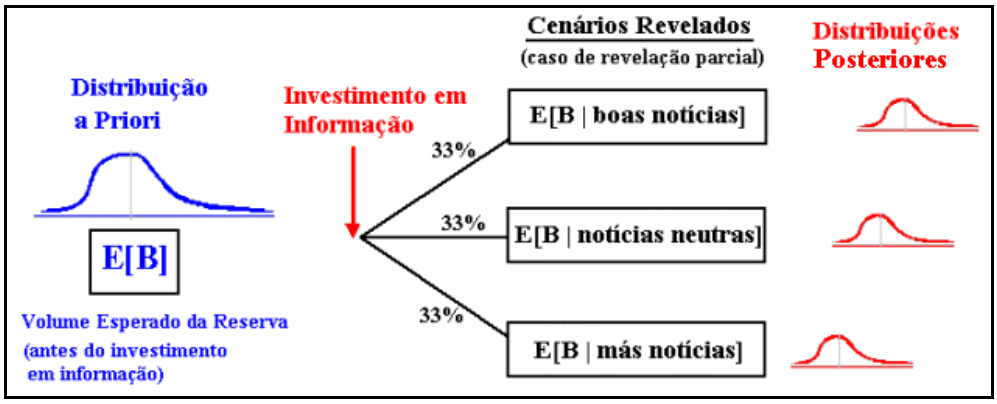

Figura 28 – Exemplo de Informação Imperfeita ou Revelação Parcial

Na Figura 28, para simplicidade de exposição, o investimento em informação revela apenas três cenários. Em geral, no entanto, existe uma distribuição contínua de sinais S, originando infinitos cenários ou infinitas distribuições posteriores de densidades $p(B \mid S = s_i)$, $i = 1, 2, \ldots \infty$. Como a informação imperfeita não revela o verdadeiro cenário da distribuição posterior – apenas a distribuição, após essa informação o decisor terá de tomar a decisão ótima (investimento ótimo) dada uma distribuição posterior obtida para cada sinal $S = s_i$. Essa decisão pode ser baseada em momentos probabilísticos dessas (infinitas) distribuições posteriores, especialmente as medidas baseadas no primeiro momento (valor esperado condicional ou expectativa condicional) e no segundo momento (variância condicional). Em geral note que, enquanto ex-ante existem infinitas distribuições posteriores, a distribuição de expectativas condicionais – ou distribuição de revelações, é única.

Nessa tese se trabalhará especialmente com a distribuição de revelações (expectativas condicionais) que tem propriedades matemáticas convenientes, intuitivas e simples de trabalhar. Também será usada em muitos casos a variância *esperada* das distribuições posteriores, que dá a incerteza técnica residual esperada após o investimento em informação.

Uma propriedade conhecida (que depois será mostrada) é que, se a informação for relevante em termos estatísticos, então a variância *esperada* das distribuições posteriores é menor que a variância da distribuição a priori. Ou seja, a informação ou sinal gera uma *redução esperada na variância*, reduzindo a incerteza técnica residual esperada. No entanto, pode ocorrer o caso de um ou mais sinais específicos gerarem algumas distribuições posteriores com variância *maior* que a variância da distribuição a priori. Mas, em média, a variância

posterior nunca é maior que a variância da distribuição a priori. Assim como ocorre com a distribuição de expectativas condicionais, a distribuição de variâncias condicionais é única. No entanto, ao que consta, ela não tem propriedades matemáticas intuitivas, convenientes ou fáceis de trabalhar. Por isso, essa tese irá trabalhar apenas com a distribuição de expectativas condicionais e com a variância residual esperada (média das variâncias posteriores), que são fáceis de obter e convenientes para trabalhar, como será visto.

O método tradicional de VOI da escola de análise de decisão usa geralmente a representação diagramática denominada *árvore de decisão*. Nela existem os nós de decisão (representados por pequenos quadrados) e nós de chance (pequenas circunferências). O exemplo a seguir ilustrará os problemas de trabalhar com árvores de decisão. Baseado em Raiffa (1968, p.241), a Figura 29 mostra o clássico problema de perfuração do poço pioneiro sem e com teste sísmico prévio.

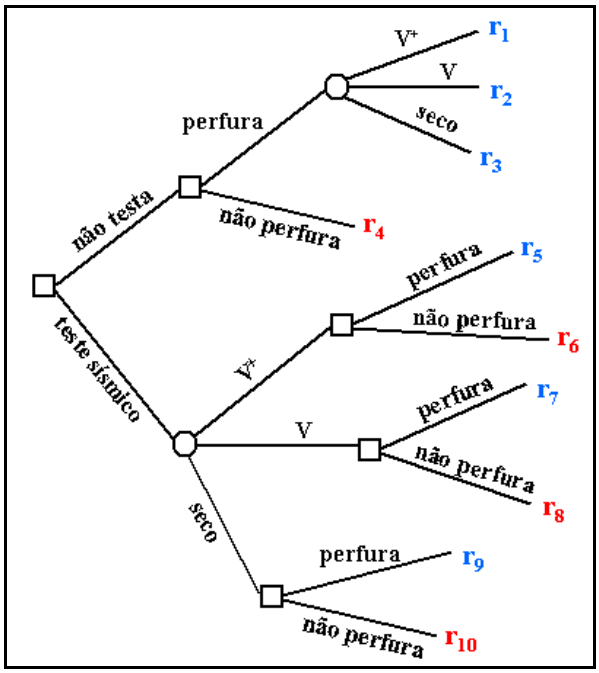

Figura 29 – Árvore de Decisão com Teste Sísmico Perfeito

Os cenários V^+ e V representam descobertas de reservas respectivamente de grande e médio tamanho[118], enquanto "seco" significa inexistência de petróleo. Nesse caso estilizado, a sísmica indica o verdadeiro cenário (informação é perfeita), e logo o número de cenários apontados pela sísmica tem de ser igual a

[118] A terminologia é um pouco diferente de Raiffa, que usou "wet" para V, "soaking" para V^+, "no structure" para a indicação "seco" pela sísmica, "open structure" para indicação V pela sísmica e "closed structure" para indicação V^+ pela sísmica.

três, pois, como foi visto, S = X para informação perfeita. Além disso, é intuitivo que a distribuição a priori de X (probabilidade dos ramos r_1, r_2, r_3) tem de ser igual à distribuição de sinais S (probabilidades do teste indicar V^+, V, seco). Veremos que a *intercambiabilidade* das distribuições de X e S é uma condição necessária para a informação ser perfeita (haver revelação total). A Figura 30 apresenta o mesmo caso, mas com o teste sísmico revelando informação imperfeita.

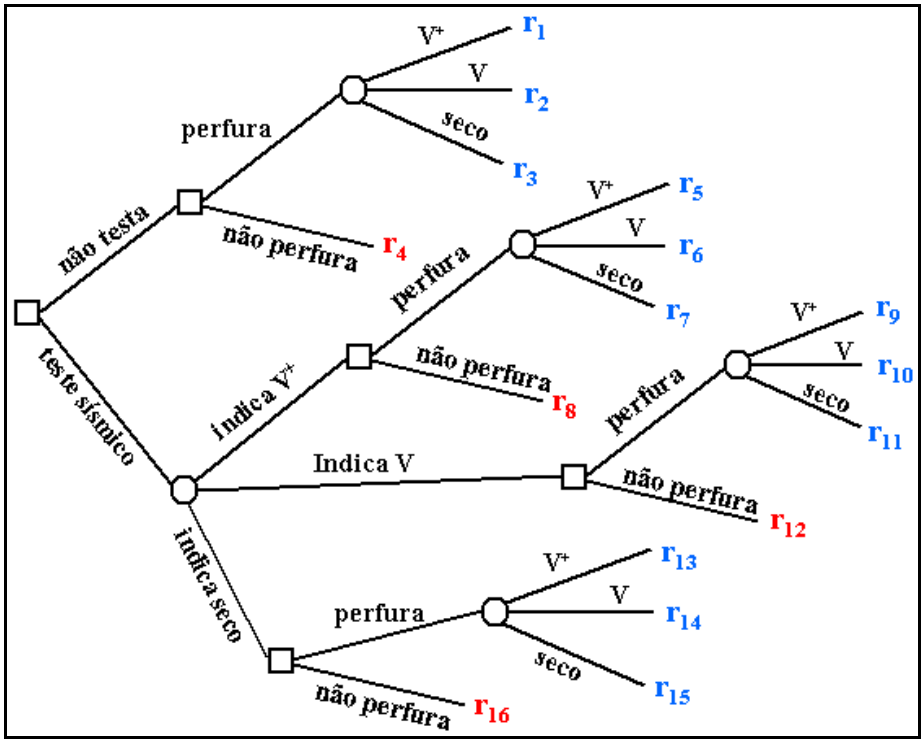

Figura 30 - Árvore de Decisão com Teste Sísmico Imperfeito

Aqui está se supondo que, embora imperfeito (com ruído), a sísmica dá indicações para todos os cenários da distribuição a priori X = {seco, V, V^+}. Comparando as figuras 29 e 30, note que as partes de cima ("não testa") das árvores são iguais. A diferença entre as árvores está na parte de baixo da árvore ("teste sísmico"), onde o caso de informação imperfeita tem muito mais ramos.

Os ramos terminais das árvores em vermelho são aqueles de cálculo trivial de seus valores (valor igual a zero na parte de cima e de menos o custo da sísmica na parte de baixo). Os que dão mais trabalho para calcular são os ramos terminais azuis especialmente os da parte de baixo da Figura 30, principalmente pela necessidade de computar probabilidades condicionais inversas de forma consistente. Comparando essas duas árvores, a de informação perfeita tem três ramos azuis na parte de baixo, enquanto que a de informação imperfeita tem 3^2 = 9 ramos azuis. É fácil verificar que se o número de cenários fosse 4, esses ramos

azuis passariam de 4 para $4^2 = 16$, etc. Ou seja, apesar da vantagem intuitiva da árvore de decisão, com alguns poucos cenários a mais o número de ramos da árvore "explodiria", tornando o método de difícil manuseio prático.

Os ramos terminais azuis da parte de baixo da árvore da Figura 30 usam *probabilidades inversas* p(s | x) que são medidas de *confiabilidade* da informação. Por ex., a probabilidade de ocorrer o ramo r_9 é a probabilidade da sísmica indicar V dado que o verdadeiro cenário é V^+. A função que descreve essa confiabilidade da informação é a chamada *função verossimilhança*. Essas probabilidades inversas podem ser difíceis de estimar na prática por profissionais da indústria.

Além disso, a função verossimilhança requer regras de consistência probabilística em geral mais complicadas do que as que serão propostas nessa tese. Por ex., na Figura 30, os três primeiros ramos têm as probabilidades iguais às da distribuição a priori. Já as probabilidades (inversas) dos ramos r_5, r_6, r_7, r_9, r_{10}, r_{11}, r_{13}, r_{14} e r_{15}, têm de ser consistentes tanto com as probabilidades a priori (ou dos ramos r_1, r_2 e r_3) como com as probabilidades dos sinais. Ex.: $p(r_5 + r_9 + r_{13}) = p(r_1)$; $p(r_9 + r_{10} + r_{11}) = p(s = V)$, etc. Ou seja, é necessário montar uma *matriz de probabilidades inversas* p(s | x) consistentes, que nesse exemplo seria 3 x 3.

Problemas de compatibilidade de probabilidades ou distribuições inversas são freqüentes na literatura tradicional Bayesiana. Ver, por ex., Arnold & Castillo & Sarabia (2001) para discussão e para um detalhado exemplo discreto.

As aplicações profissionais de valor da informação usando a literatura clássica de análise de decisão, em geral apresentam importantes limitações:

- Muitas vezes usam árvores de decisão e assim precisam limitar o número de cenários (ex.: três) revelados pela informação;
- Muitas vezes assumem que toda a incerteza será resolvida com o investimento em informação (revelação total, informação perfeita);
- Não quantificam e nem comparam alternativas de diferentes custos e diferentes potenciais de revelar informações relevantes;
- Não consideram as interações das incertezas técnicas e de mercado, apesar de ambas afetarem o valor econômico da reserva; e
- Ignoram o tempo legal de expiração dos direitos de investir e o tempo de aprendizagem, em geral diferente para cada alternativa.

Essa última limitação é particularmente relevante em problemas de OR, ou seja, é necessário obter o valor *dinâmico* da informação, onde o adjetivo "dinâmico" está relacionado ao tempo como variável de estado.

3.1.3. Revisão da Literatura de VOI e Opções Reais Relacionadas

3.1.3.1. Revisão da Literatura de Valor da Informação

Na literatura tradicional de VOI, o único livro texto avançado visando cursos de mestrado ou doutorado é o de Lawrence (1999), o qual tem sido aqui citado algumas vezes. Mas mesmo assim ele não é todo satisfatório, por ex., ele desconhece a literatura de opções reais, ele cita a medida de dependência chamada cópula (que será definida) mas não cita a razão de correlação η^2 que será aqui usada, aplicações dinâmicas são raras, etc. Já o livro de Bather (2000), embora também para pós-graduação e sobre teoria da decisão, é mais um livro de programação dinâmica com aplicações em decisões seqüenciais onde inexistem aplicações diretamente relacionadas a VOI (embora o método de cálculo possa contribuir nessa análise). Livros clássicos de análise de decisão como o de Raiffa (1968) e o de Lindley (1985, primeira edição de 1971), abordam bastante o tema de VOI, mas a nível mais elementar (graduação) e não consideram problemas de VOI num contexto mais dinâmico. Em todos esses livros, tanto a literatura neoclássica de finanças como a teoria de OR, são ignoradas.

Apesar da intuição de que a informação poderia ser vista como uma commodity, com custo de suprimento e preço de mercado baseado na demanda e oferta de informação, as tentativas de assim caracterizá-la e desenvolver uma teoria geral da economia da informação, não tem sido bem sucedidas. Por ex., Moscarini & Smith (2002) consideram que o problema de como valorar/apreçar a informação é ainda "*um problema não-resolvido na teoria Bayesiana de decisão*".

Conforme assinala Arrow (1984, Preface), apesar de não ser difícil perceber que num mundo de incertezas a informação é valiosa no sentido econômico, "*tem sido provado difícil conceber uma teoria geral de informação como uma commodity econômica, porque diferentes tipos de informação não têm uma unidade comum que tenha ainda sido identificada*".

Tentativas, no entanto, foram feitas. Em particular, atraiu muita atenção a revolucionária *Teoria da Informação* oriunda da área de engenharia de comunicação, desenvolvida por Shannon nos anos 40 e sintetizada no livro de Shannon & Weaver (1949). Essa abordagem matemática baseada no conceito de *entropia*, levou à criação de um novo ramo na estatística e levou eminentes economistas como Arrow e Marschak a investigarem a aplicabilidade dessa teoria em problemas de VOI, especialmente na década de 50. A teoria da informação (incluindo o conceito de entropia) será discutida no item 3.1.4.3.

Conforme reportam diversos autores, tais como Lawrence (1999, p.x), Lindstädt (2001, p.352) e Arrow (1978), Jacob Marschak foi o grande contribuidor das bases teóricas e da formalidade matemática/probabilística para a análise econômica da informação. Seus artigos nessa área, Marschak (1954), Marschak (1959) e Marschak & Miyasawa (1968), e seus livros Marschak (1974, coletânea) e Marschak & Radner (1972), são os mais conhecidos. Arrow (1978, p.xiii) destaca que a maior novidade de Marschak foi "*a síntese da teoria estatística da decisão, com a teoria econômica da decisão sob incerteza e a teoria da organização, a cuja combinação ele deu o nome de teoria das equipes*". Assim, a maior contribuição de Marschak segundo o Prêmio Nobel K. Arrow, foi obtida através de uma *hibridização* de teorias econômica, estatística e organizacional! O livro Marschak & Radner (1972) sobre teoria econômica de equipes teve grande impacto na literatura da economia da informação, em especial a primeira parte do livro (decisão individual), pela clareza e rigor da exposição.

Para Lawrence (1999, p.x), Marschak foi o maior contribuidor dessa área e ele está para a economia de informação como Alfred Marshall está para a microeconomia clássica. Laffont (1986, p.68) afirma que o primeiro tratamento importante sobre o papel da informação na teoria econômica foi Hayek (1945), mas ele baseou o seu capítulo sobre estrutura de informação na "*abordagem fundamental*" do cap. 2 de Marschak & Radner (1972).

Nessa tese será discutido de forma crítica um exemplo numérico de Marschak (1959), o qual será usado para defender uma medida de redução esperada de incerteza como superior à função verossimilhança usada por ele.

Em paralelo com a teoria econômica da informação, surgia a *teoria estatística da decisão*, sendo particularmente pioneiro e fundamental para a teoria do VOI os artigos clássicos de Blackwell (1951 e 1953) que fundou a teoria da

comparação de experimentos ou *comparação de estruturas de informação*. Hoje esse é um ramo da teoria estatística, sendo que o tratado de Torgersen (1991) é dedicado totalmente a ele e com grande rigor matemático (com o uso da teoria da medida). Marschak & Miyasawa (1968) usaram as idéias de Blackwell para comparar sistemas econômicos de informação. Esse tópico será ainda discutido no item 3.1.4.2.

Na teoria estatística ajudaram também trabalhos sobre *decisões seqüenciais* de Wald (1947) e os *axiomas Bayesianos* introduzidos por Savage (1954). Na década de 50 também houve grandes avanços na área de *métodos de otimização dinâmica*, com a teoria da *política s-S* para regra de decisão ótima (ver Arrow & Harris & Marschak, 1951)[119] e métodos de retro-indução (indução "backwards") e de parada ótima em Arrow & Blackwell & Girshick (1949)[120]. Ambos foram precursores do método mais geral chamado de *programação dinâmica* por Bellman na década de 50, sintetizado no seu livro texto (Bellman, 1957).

A popularização dos métodos de VOI para estudantes de pós-graduação e para o público profissional se deu apenas na década de 60, com a emergente (na época) teoria de *análise de decisão*, especialmente com a publicação do livro texto de Raiffa (1968), difundindo métodos mais intuitivos baseados em árvores de decisão, aproveitando o grande impacto que teve o artigo de Hertz (1964). Um artigo muito citado e representativo dessa literatura é o de Howard (1966).

Historicamente, pode-se dizer que o livro do Raiffa (1968) está para a *análise de decisão* assim como o livro de Dixit & Pindyck (1994) está para *opções reais*. Assim, a teoria das opções reais é cerca de uma geração mais nova que a teoria de análise de decisão. OR cresceu num ambiente em que a teoria moderna de finanças corporativas estava mais bem estabelecida, o que não ocorreu com a teoria de análise de decisão, especialmente nas primeiras décadas.

Da literatura econômica clássica, tem sido muito citado em artigos acadêmicos de OR (e também em Dixit & Pindyck, 1994, p.352) o conhecido artigo de Roberts & Weitzman (1981) sobre VOI. Eles consideram investimentos seqüenciais onde o custo é incerto e sua variância vai sendo reduzida com o

[119] A idéia é escolher dois níveis, inferior (s) e superior (S) para uma política ótima sob incerteza. No caso de estoques, o nível s é o *gatilho* que detona a reposição de estoques através de uma compra que eleva o estoque de s para S. A escolha ótima de s-S minimiza uma *função perda*.

[120] Conforme assinala Bather (2000, p.166), eles provaram um resultado fundamental da análise seqüencial, a otimalidade do "sequential probability ratio test", usando retro-indução.

investimento acumulado[121]. Mas para simplificar o modelo, os estágios exploratórios são modelados como um *processo contínuo*. Já em exploração de petróleo, por ex., o foco da valoração é em *eventos pontuais*: podem-se passar anos sem nenhum evento relevante e depois se perfurar um poço revelando a existência de petróleo. Em Roberts & Weitzman (1981) a incerteza técnica não interage com a de mercado, um efeito importante que é um dos focos do modelo dessa tese.

Conforme mencionado antes, são poucos os artigos de OR relacionados a VOI que tem real interesse e que não cometem erros conceituais relevantes. Um deles é o de Martzoukos & Trigeorgis (2001). Eles modelam um processo de aprendizagem custoso (investimento em informação) como um processo endógeno de "saltos" de tamanho aleatório. Os saltos são ativados pelo gerente e assim é um processo indexado por *eventos*, como será feito nessa tese, e não pela simples passagem do tempo (erro mais comum da literatura de OR + VOI). Na metodologia que eles usaram, a aprendizagem é relacionada somente com o valor do ativo básico (nessa tese é incluído também o fundamental efeito no preço de exercício ótimo da opção, por ex., na eq. 28, I_D é função de B) e o foco deles foram o momento ótimo de aprendizagem ("timing of learning") e a aprendizagem multi-estágio. O mérito do artigo é que eles tiveram a intuição correta de que a revelação de informação técnica gera saltos no valor do ativo básico, assim como na indexação do aprendizado a eventos opcionais (e não ao tempo). Mas ao contrário do modelo apresentado nessa tese, eles não teceram considerações sobre como obter a distribuição do tamanho dos saltos (nessa tese será visto que os tamanhos dos saltos devem ser amostrados da *distribuição de revelações*), nem consideram o problema de seleção da melhor alternativa de investimento em informação e nem outros detalhes como a inclusão do tempo de aprendizado.

Outro artigo importante de OR aplicado a VOI é o recente de Murto (2004), que usa justamente o exemplo de incerteza técnica na decisão de desenvolvimento de um campo de petróleo. No caso ele usa a incerteza técnica no volume de reservas, além da incerteza de mercado modelada com um MGB.

Murto (2004) critica dois artigos de OR que consideram incerteza técnica no parâmetro de tendência ("drift") do processo estocástico: "*a aprendizagem é*

[121] Pindyck (1993) faz a mesma coisa mas num modelo mais completo que será aqui visto.

passiva, pois as firmas atualizam as suas crenças continuamente enquanto esperam, meu foco é em aquisição ativa de informação discreta". Essa tese concorda plenamente com isso e tem o mesmo foco.

As divergências dessa tese com o modelo de Murto (2004)[122] começam com os detalhes da modelagem, que reflete em parte o pouco conhecimento dele sobre a indústria de E&P de petróleo. Primeiro, ele assume oportunidade perpétua de desenvolvimento para eliminar o tempo como variável de estado. Com exceção de países com monopólios ou campos terrestres em terrenos particulares em alguns países, existe um tempo legal finito para essa opção. A simplificação adicional de que o aprendizado é instantâneo (na tese será considerado o caso mais realista de tempo de aprendizagem, em geral igual ou maior que um mês), prejudica um pouco o objetivo dele de tirar conclusões gerais sobre o momento ótimo de aprender. Outra simplificação feita por Murto foi de que a atividade de aprendizagem "*revela o verdadeiro valor do parâmetro*" a um certo custo. Ou seja, foi considerado o caso pouco realista de revelação total (informação perfeita). No modelo dessa tese a informação é em geral imperfeita e a revelação de informação é parcial. Murto evita definir a forma funcional da função VPL de exercício da opção de desenvolvimento, o que só faz na sua seção 5 quando faz um exemplo ilustrativo, onde assume uma distribuição a priori *uniforme*[123] do volume de reservas e que o custo de desenvolvimento é constante e independente da variável com incerteza técnica. Essa última premissa é muito inadequada, pois o investimento ótimo claramente depende de B (na seção 3.1.3.2 será visto que essa simplificação pode levar a erros no VOI muito grandes, maior até que 50%).
Mas Murto (2004) tem também interessantes contribuições, pois modela de forma endógena a questão do momento ótimo de investir em informação vis a vis com o momento ótimo de desenvolver o campo[124]. Com isso, seu modelo faz o balanço entre postergar o custo de aprender versus o possível benefício de que a revelação antecipada da informação pode ser mais benéfica ao projeto.

---

[122] Após elogiá-lo pelas importantes contribuições, o autor dessa tese expôs todas as críticas aqui listadas no debate público que se seguiu após a apresentação desse artigo em Montreal.

[123] A distribuição uniforme significa quase total ignorância do volume B. Isso não é realista e não é usado nem em prospectos exploratórios pela indústria. Distribuições lognormais ou triangulares, por ex., seriam bem mais adequados ao conhecimento prévio dessa fase de E&P.

[124] Na tese essa questão será analisada no cap. 5 como uma *sensibilidade* à data de investimento em informação, resolvendo o modelo para diferentes datas de investimento em informação. A menor importância da tese a essa questão é de ordem prática: o montante do investimento em informação é apenas cerca de 1 a 5% do investimento em desenvolvimento.

Da literatura de OR + VOI deve-se mencionar também artigos como o de Mayor et al. (1999) e principalmente o de Childs & Ott & Riddiough (2001) que usam a *teoria de filtração ótima* ("optimal filtering theory"), ver, por ex., os livros texto de Jazwinski (1970) ou de Bensoussan (1992). A aplicação dessa teoria assim como esses artigos serão comentados abaixo, no subitem 3.1.3.2.

Antes de discutir a teoria da filtração, é necessário comentar o artigo clássico de Pindyck (1993) que combina de forma elegante a incerteza técnica com a de mercado. Esse modelo está sumarizado em Dixit & Pindyck (1994, p. 345-353) e tem algumas conexões com o modelo defendido nessa tese.

No modelo de Pindyck (1993), o custo total do projeto K é incerto e a taxa ótima de investimento é zero ou k. Quando ocorre o investimento, existe uma revelação de informação de forma que a variância da incerteza técnica é reduzida. Assim, em consonância com essa tese e em oposição ao modelo de Cortazar & Schwartz & Casassus (2001), a variância da incerteza técnica só é reduzida na ocorrência de um evento como o investimento. Ou seja, tanto aqui como em Pindyck (1993), a variância da incerteza técnica não muda pela simples passagem de tempo como em Cortazar & Schwartz & Casassus (2001). Da mesma forma, em caso de apenas incerteza técnica, os valores dos parâmetros técnicos mudam somente se a firma está investindo (também em oposição ao modelo de Cortazar et al., 2001, que muda pela simples passagem do tempo).

No modelo dessa tese será visto que a variância da distribuição de revelações aumenta linearmente com a redução esperada de variância. De forma similar, no modelo de Pindyck a variância instantânea da taxa de variação da variável técnica aumenta linearmente com a razão I/K. Lá existe um máximo quando essa razão é igual a 1, o que também é similar ao caso de revelação total (informação perfeita) usada na tese, porque será visto que o máximo da variância da distribuição de revelações ocorre no caso de revelação total quando a medida de aprendizagem $\eta^2 = 1$. No caso de Pindyck (1993), o próprio investimento seqüencial de desenvolvimento do projeto gera informações (*modelo de tempo de construção* aplicado a usinas nucleares). Já nessa tese, o investimento em informação não reduz necessariamente o custo de desenvolvimento do projeto (mas será visto no cap.5 casos em que reduz).

Pindyck (1993) é um modelo rigoroso e que captura os fatos estilizados da incerteza técnica, mas ele é focado em um tipo de aplicação onde o aprendizado

ocorre durante a construção do projeto. Aqui será visto casos em que se pode investir antes em informação para depois decidir se desenvolve ou não o projeto e, caso desenvolva, o investimento deve se ajustar a essa realidade técnica revelada.

3.1.3.2. Teoria da Filtração Ótima Aplicada a Opções Reais e VOI

Antes de comentar os artigos de Mayor et al. (1999) e de Childs & Ott & Riddiough (2001), é necessário fazer a seguinte caracterização para problemas de estimação em geral em que o processo de incerteza é indexado pelo tempo (contínuo ou discreto, abaixo será usado o discreto por simplicidade). Essa definição é baseada em Jazwinski (1970, p.143-145) e em Kellerhals (2004, p.24).

Seja o problema de estimar um parâmetro, fator ou variável ξ_t, usando a informação disponível até o tempo discreto s, representada pelo *conjunto de informação* $\mathcal{F}_s = \{y_s, \ldots, y_2, y_1\}$, onde y_τ é uma observação mensurável na data τ.

Distinguem-se então três casos, dependendo do conjunto de informação usado:

1. Problema de Filtração: para $t = s$,
2. Problema de Suavização ("smoothing"): para $t < s$, e
3. Problema de Predição: para $t > s$.

Como destaca Jazwinski (1970, p.143), filtração e predição são usualmente associadas com operações em tempo real, ou seja, as observações são disponíveis imediatamente. Isso já dá uma idéia de aplicações típicas dessa abordagem. A mais óbvia é a estimativa de parâmetros de processos estocásticos que representam a evolução de preços ou outras variáveis da economia que podem ser medidas em tempo real. No caso de incerteza técnica em OR em petróleo, raramente existe a chegada contínua ou mesmo periódica de informações em tempo real para se fazer uma atualização contínua de uma variável com incerteza técnica usando essa informação em tempo real.

Uma exceção, no entanto, pode ser o caso de poços inteligentes (ver último tópico do cap. 2), onde existe uma incerteza técnica residual no reservatório (e logo na produção dos poços), sendo que esses poços são equipados com sensores que enviam informações em tempo real a uma plataforma (ou a um sistema inteligente) que usa essas informações para atualizar o modelo e eventualmente tomar ações corretivas (abertura ou fechamento de zonas do poço) à luz dessas novas informações. Nesse caso poderia se pensar em modelos tais como os

desenvolvidos por Childs & Ott & Riddiough (2001). No entanto, na maioria dos problemas, a nova informação técnica só ocorre se for feito um projeto de investimento em informação que dará uma resposta pontual. Tipicamente, a perfuração de um poço exploratório ou de delimitação, só gera informação quando o poço atinge a zona de interesse (no final da perfuração), e qualquer decisão só é tomada após a análise do relatório do poço, que dará informações sobre a existência, quantidade e qualidade do petróleo. Ou seja, na maior parte dos problemas de relevantes incertezas técnicas em E&P de petróleo, a revelação de informação técnica só ocorre pontualmente, após semanas ou meses após o início de algum investimento em informação.

É demasiadamente grosseira a alegação que, numa bacia exploratória com muitas firmas atuando, a freqüência de perfurações exploratórias poderia tornar razoável um modelo com revelação contínua de informação exploratória. Parece bem mais realista nesse caso, modelar essa chegada de informação como um processo de Poisson, por ex., com $\lambda = 2$ meses para muitas firmas explorando a bacia. Além disso, do ponto de vista de uma firma que tem um prospecto nessa bacia e quer usar essa informação, muitas dessas perfurações exploratórias não têm correlação (ou são muito fracas) com o prospecto de interesse, seja por pertencerem a outro "play" geológico, seja por serem muito distantes.

Mas deve se reconhecer, que fora da área de petróleo, pode haver aplicações em que a informação técnica pode chegar em tempo real ou com grande freqüência e assim aplicar os modelos que serão comentados nessa seção. Por ex., um teste de mercado de um novo produto, onde a incerteza técnica é a função demanda e/ou a reação de consumidores ao produto, etc., pode ser observado diariamente e assim justificar o uso desses modelos para esses tipos de aplicação. Outra aplicação *adequada* desse modelo é no mercado imobiliário ("real estate"), que possivelmente estava na mente dos autores quando escreveram o primeiro artigo, Childs & Ott & Riddiough (2001), a julgar pelos dois artigos relacionados subseqüentes, Childs & Ott & Riddiough (2002a e 2002b)[125]. Num mercado imobiliário amplo, pode-se pensar que existem transações freqüentes que provêem informações sobre o verdadeiro valor de um ativo específico que o modelo está

[125] Grenadier (1999) também modela a revelação de informação no mercado imobiliário, mas num contexto estratégico em que existe assimetria de informação.

analisando. Mas mesmo assim, apenas as transações de "ativos comparáveis" são relevantes para reduzir o nível de ruído do verdadeiro valor do ativo, que ao mesmo tempo oscila devido à incerteza de mercado (oferta x demanda). Assim, modelos em que informações sobre incerteza técnica (ruído) chegam continuamente ao longo do tempo, assumem uma premissa que nem sempre é válida para apreçar um imóvel específico. Além disso, é bem complicado separar o que é uma oscilação de preço devido ao mercado daquela devido a uma revisão do ruído técnico (eles procuram separar esses efeitos através de medidas de autocorrelação da série, usando técnicas de cointegração).

Mas infelizmente, esses modelos de OR com filtração não têm a generalidade que os autores desses modelos em geral supõe. Por ex., em P&D de um novo remédio, existem estágios bem definidos[126], sendo que a informação relevante (ou processada) só se dá ao fim de cada estágio, com um relatório técnico analisando e tirando conclusões que são usadas para exercer ou não a opção de realizar o próximo estágio. Assim, também em projetos de P&D de novos remédios, as metodologias que prevêem chegadas freqüentes de informação técnica relevante em tempo real, não são adequadas para a decisão ótima de exercícios de OR. A exceção é a análise *agregada* de investimentos em P&D.

Uma importante conexão da teoria de filtração e de predição com a abordagem de expectativas condicionais que será usada nessa tese, é que para *todos* os problemas de filtração e predição, e independentemente das propriedades da função densidade de probabilidade condicional, a expectativa condicional é a melhor estimativa desses problemas (no sentido de menor erro quadrático médio, que é o critério usualmente aplicado), ver Jazwinski (1970, p.149-150). Por isso, esses problemas trabalham basicamente com equações de evolução da densidade condicional e de evolução da expectativa (ou média) condicional.

O artigo de OR de Mayor et al. (1999) talvez tenha sido o primeiro a usar a teoria de filtração ótima em problemas de OR. O foco de aplicação é justamente a aprendizagem através de pesquisas de mercado (que é um nicho adequado de aplicação, ver discussão acima). Num único modelo, o artigo combina os aspectos da aprendizagem do verdadeiro estado da economia e o clássico problema do

[126] Tipicamente (Rogers, 2002, p.25-26): 1) descoberta; 2) testes pré-clínicos; 3) testes clínicos, fase I; 4) idem, fase II; 5) idem, fase III; 6) aprovação da agência de saúde; 7) testes pós-aprovação (para desenvolver extensões do produto, dosagens para crianças, etc.).

momento ótimo de investimento. Eles estendem a abordagem de Roberts & Weitzman (1981) não apenas com ingredientes de OR como assumindo que o processo de aprendizagem e a sua intensidade são endógenos ao modelo (i. é, opções da firma). Assim, a firma pode "*controlar a quantidade de informação que ela compra ao longo do tempo*". A firma tem três opções: (a) desenvolver o projeto; (b) continuar observando sinais com ruídos a um certo custo (compra de informação); (c) abandonar o projeto.

Mayor et al. (1999) argumentam que a opção de aprendizagem é valiosa "*porque permite à firma reduzir o risco de desenvolver um projeto com valor ('payoff') negativo*". Como em Martzoukos & Trigeorgis (2001), a opção de aprendizagem tem impacto apenas no valor do ativo básico V, não no preço de exercício. O investimento I para desenvolver o projeto é fixo e conhecido. Essa é uma simplificação matemática muito conveniente para resolver o modelo. No entanto, é uma simplificação inaceitável em muitas aplicações, pois as reais características técnicas de um projeto (ex.: o real tamanho do mercado para um produto) é um dado de entrada fundamental para dimensionar o investimento adequado para desenvolver V. Logo, em geral, o preço de exercício I não pode independer das verdadeiras características técnicas de V que forem reveladas.

Para se ter uma idéia do erro em não reconhecer esse fato elementar do dimensionamento de um projeto, considere novamente o exemplo dado no item 3.1.2, especialmente a Tabela 8. Se a informação perfeita fosse útil apenas para decidir se a opção deve ou não ser exercida, não influenciando o investimento I, então a primeira linha de dados da Tabela 8, cenário B^+ seria igual para o caso com ou sem informação, i. é, VPL(B^+) = 220 (em vez de 270 para informação perfeita). O único benefício da informação seria com o cenário B^- onde a informação perfeita evitaria um exercício equivocado da opção de desenvolver o projeto. Nesse caso o VEIP cairia de US$ 60 MM para US$ 43,3 MM, um erro de 28% nesse caso simples. Se aumentar o número de cenários e/ou se o projeto for um pouco melhor, o erro tende a ser ainda maior e não é difícil mostrar exemplos onde esse erro superaria o patamar de 50%. Se o projeto fosse melhor, por ex., se

a qualidade aumentasse de q = 0,2 para q = 0,25, então o erro de ignorar o dimensionamento ótimo pularia de 28% para 64%[127]!

O artigo de Childs & Ott & Riddiough (2001) também assume investimento fixo independente da incerteza técnica e assim também deixa de considerar uma das principais fontes de valor da aprendizagem. Eles fazem outras simplificações que limitam a sua aplicabilidade, mas que permitem ter soluções analíticas, como o caso de opção perpétua de desenvolvimento. Mas é um artigo que trás inúmeras contribuições que podem ser úteis, mas num intervalo mais restrito de aplicações onde a incerteza técnica possa ser indexada pelo tempo (o que em geral não ocorre com aplicações em E&P de petróleo e P&D). Essa incerteza ou ruído tanto pode diminuir como aumentar pela simples passagem do tempo. Além disso, a informação pode ser comprada em pequenas quantidades e o dono da opção pode escolher o número de vezes que a informação é adquirida e o nível de redução de ruído (leia-se redução esperada de variância). O artigo usa uma terminologia que guarda similaridade com o usado em Dias (2002) e nessa tese. Exemplos, eles usam o termo "revealed variance", enquanto aqui se usa "variância da distribuição de revelações"; eles usam o termo "full information" e aqui é usado "revelação total". Childs & Ott & Riddiough (2002b, eq.8a) usa uma equação em que "informação total = variância revelada + variância residual"[128], que é similar a uma equação que será usada nessa tese no item 3.2. Como nessa tese, a variância revelada tem um papel fundamental para "*a determinação do valor e política de exercício das opções*" (Childs & Ott & Riddiough, 2002b, p.440).

O artigo de Childs & Ott & Riddiough (2001) é bastante rigoroso na aplicação da teoria de filtração ótima, mas ao mesmo tempo é didático na explicação dos termos. Eles obtêm equações para a expectativa condicional do valor do projeto (e também para a variância residual) que, conforme Jazwinski (1970), é típico em aplicações da teoria de filtração. Esse é um ponto de contato interessante da teoria de filtração com a metodologia da tese, que também trabalha basicamente com distribuições de expectativas condicionais. Eles obtêm uma série de resultados analíticos que permitem tirar algumas conclusões interessantes. Por

[127] Com q = 0,25, o VEIP seria de US$ 62,5 MM considerando dimensionamento ótimo como benefício da informação e de apenas US$ 22,5 MM se considerar apenas o benefício de exercer ou não a opção. Ver no CD-Rom as planilhas ex_inf_perf.xls e ex_inf_perf2.xls.

[128] Nesses artigos, a *variância residual* mede a precisão da estimativa do valor verdadeiro e a *variância revelada* mede a quantidade de informação que chega num certo intervalo de tempo.

ex., o valor da opção é menor na presença de ruído (incerteza técnica) do que sem ruído. Na tese isso ocorre porque o ruído diminui o VPL de exercício, pois com ruído quase certamente o investimento será sub-ótimo. Na "Figure I" do artigo, eles mostram um exemplo em que o gatilho para o desenvolvimento ótimo diminui com o aumento da variância da incerteza técnica. Interpretar esse resultado é difícil, pois no modelo deles o ruído pode crescer com a simples passagem de tempo e assim o exercício antecipado evitaria exercer uma opção com mais ruído. Essa conclusão do artigo é até um pouco perigosa, pois se com muito ruído o gatilho é baixo, pode-se querer desenvolver logo um projeto em que o melhor seja investir em informação. Mas no artigo, o aumento do ruído também aumenta o valor de adquirir informação, amenizando o problema.

Como nessa tese, Childs & Ott & Riddiough (2001, p.59) usam uma medida de aprendizagem (que eles chamam de fator β) em que o valor 1 corresponde ao caso de "fully revealing signal" e o valor zero ao caso completamente não informativo. No entanto, ao contrário dessa tese, eles não desenvolvem uma teoria de medida de aprendizagem, não estabelecem propriedades para esse fator e nem percebem a ligação direta dessa medida com o conceito de expectativas condicionais, como aqui será mostrado. Além disso, no artigo subseqüente Childs & Ott & Riddiough (2002a, eq.9) eles usam o clássico *coeficiente de correlação de Pearson*, que só é válido em casos específicos (como distribuição Normal, ver item 3.3) para estabelecer uma medida que eles chamam de "instantaneous information ratio", enquanto que nessa tese a medida será bem mais geral em termos de distribuições que podem ser usadas (só exige média e variância finitas).

3.1.4.
Temas Clássicos de VOI e Conexões com Outras Teorias

3.1.4.1.
Concavidade e Não-Concavidade no Valor da Informação

Na literatura econômica, especialmente a *teoria da demanda de informação*, o artigo de Radner & Stiglitz (1984) sobre a não-concavidade no VOI, causou grande impacto na comunidade acadêmica e foi objeto de muito debate (exs.: Lindstädt, 2001; Kihlstrom, 1984b; Chade & Schlee, 2002; Lawrence, 1999).

Esse tema tem um interesse adicional nessa tese, devido à metodologia que emprega uma medida de aprendizagem para estimar o VOI. Dessa forma é

importante saber como a função VOI se comporta (toda côncava, parcialmente côncava, linear, etc.) entre os limites de nenhuma informação até o de informação perfeita. Conhecendo as características dessa função, é possível fazer melhores interpolações entre esses limites extremos de informatividade, simplificando dramaticamente o problema de VOI especialmente em contextos dinâmicos.

O teorema de Radner & Stiglitz (1984) diz que, sob certas condições razoáveis e ao menos que a informação seja sempre inútil, *o valor marginal de uma pequena quantidade de informação é zero e a informação deve exibir retornos marginais crescentes sobre algum intervalo*, pelo menos num intervalo vizinho ao caso não informativo. Mas o próprio artigo de Radner & Stiglitz (1984) mostra exemplos[129] no qual a informação exibe retornos marginais decrescentes, ou seja, "*claramente o VOI não é sempre não-côncavo*" (Chade & Schlee, 2002). Isso tem implicações na teoria econômica do tipo: "a demanda por informação é uma função descontínua de seu preço", "agentes nunca comprarão pequenas quantidades de informação"[130] e "agentes tenderão se especializar na produção de informação" (para produzir uma quantidade suficientemente grande), etc.

Além disso, a não-concavidade complica qualquer análise de aquisição de informação pois a condição de primeira ordem não é suficiente para maximizar o valor de um agente demandando informação. Isso pode, por ex., impedir a existência de equilíbrio competitivo, ou o equilíbrio em expectativas racionais lineares, assim como prejudica toda a literatura de aprendizagem ativa, inclusive porque pode impedir a prova da existência de equilíbrio perfeito em estratégias puras em alguns modelos (ver Chade & Schlee, 2002, p.422-423).

À primeira vista, as condições de validade do teorema de Radner & Stiglitz são apenas condições que freqüentemente ocorrem em modelos aplicados, i. é, condições de *suavidade* e de *continuidade* da estrutura de informação com um certo índice de informatividade θ. Em particular eles assumem que existe a derivada da estrutura de informação para $\theta = 0$. Radner & Stiglitz também assumem um número finito tanto de cenários da variável de interesse (ou de estado) X, como para cenários do sinal S. Mas eles trabalharam com a função verossimilhança como estrutura de informação, mais especificamente uma matriz

[129] Um dos exemplos de Radner & Stiglitz (1984) é justamente em exploração de petróleo.
[130] Isso prejudica bastante os modelos econômicos que usam análise *diferencial* em geral.

de Markov de distribuições condicionais inversas, i. é, uma matriz p(s | x). Na discussão de um exemplo de Marschak no item 3.4, será visto que com verossimilhança ou confiabilidade da informação p(s | x), o VOI nem sempre é definido para todo o intervalo de p(s | x), i. é, em todo o intervalo [0, 1], contrastando com a medida alternativa η^2 a ser proposta. Assim, uma sugestão dessa tese é repensar o teorema de Radner & Stiglitz usando outra abordagem em termos de estrutura de informação.

Como aponta Chade & Schlee (2002, p.423), existe uma rica classe de problemas para o qual o VOI é côncavo. Por ex., o modelo dinâmico de aquisição de informação em tempo contínuo de Moscarini & Smith (2002) e modelos de outros artigos, sugerem que o VOI pode ser em geral uma função côncava do *tamanho da amostra*. Esses modelos em geral assumem ou um número infinito de cenários para a variável de estado X ou um número infinito de cenários para o sinal S. Por outro lado, também existem casos de não-concavidade que não atende as premissas de Radner & Stiglitz. Chade & Schlee (2002, p.424) mostram ainda que, num modelo de muita generalidade, é difícil descartar a não-concavidade.

Nessa tese o índice de informatividade θ é uma medida de aprendizagem chamada redução (percentual) esperada de variância η^2. A pergunta natural é: essa medida também exibe não concavidade para regiões próximas de $\theta = 0$ como em Radner & Stiglitz (1984)? Nas aplicações dessa tese, a exigência de realismo faz a solução ser numérica e assim dificulta tirar conclusões gerais que seriam mais fáceis de serem obtidas em modelos simplificados teóricos com solução analítica. Mas experimentos foram feitos com os modelos usados nessa tese, tendo sido encontrado não concavidade para regiões próximas de $\theta = \eta^2 = 0$. Uma dessas aplicações que será mostrada no cap. 5 exibe esse efeito, mas de forma não severa, e uma linearidade com o valor de η^2. Esse caso é mostrado na Figura 31 a seguir.

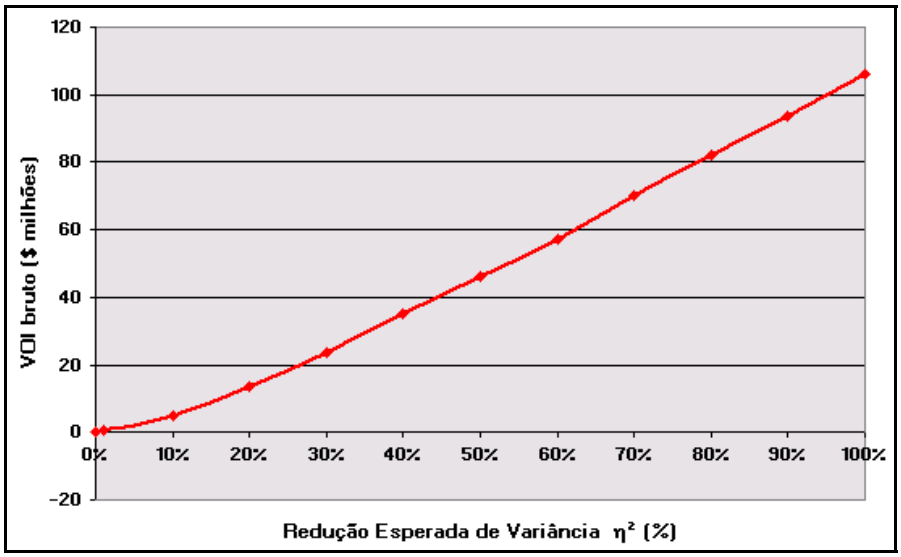

Figura 31 – Não-Concavidade do Valor da Informação

Note na Figura 31 que a não concavidade ocorre apenas na região próxima do caso não informativo. Em muitas aplicações práticas (mas não todas), essa região é de menor interesse prático, pois as reduções de variância são maiores que 30%[131]. Na Figura 31, o VOI é bruto, i. é, não considera o custo da informação e nem o efeito do tempo de aprendizagem. Além disso, o gráfico acima considera apenas a incerteza técnica (incerteza de mercado tem volatilidade igual a zero). Interessante é que continua existindo essa linearidade[132] na *região de maior interesse prático*, para nessa mesma aplicação mesmo quando se considera incerteza de mercado, como será visto no cap. 5. A qualidade dessa aproximação linear dependerá do tipo de aplicação, mas isso facilita algumas aplicações rápidas (gerenciais) em que se resolve apenas o caso de informação perfeita $\theta = \eta^2 = 1$, e com o outro ponto na *origem* se descreve uma reta que dá uma aproximação do VOI em relação a uma medida de aprendizagem. Mas o modelo dessa tese não dependerá dessa linearidade. Mas não deixa de ser um bom indício prático a ocorrência de linearidade do VOI com a medida de aprendizagem proposta.

Lawrence (1999, p.226) apresenta a seguinte conclusão prática geral do teorema de Radner & Stiglitz: "*nós podemos simplesmente concluir que os efeitos econômicos benéficos da informação começam lentamente*".

Um caso que não é mencionado nos artigos citados é a análise do efeito de escala no VOI, por ex., o efeito da aquisição de informação com o mesmo θ para

[131] Mas serão vistas também aplicações para o fator de chance exploratório com $\eta^2 \leq 10\%$.
[132] Com excelente ajuste linear, $R^2 > 0{,}998$.

projetos de tamanho diferentes (jazida pequena x jazida grande, etc.) pode fazer diferença na concavidade da função VOI x θ.

Isso foi analisado por Wilson (1975), que apresenta um modelo de economia de escala advinda da aquisição de informação para dimensionamento ótimo, i. é, uma informação positiva justifica um aumento na escala de operações, etc. Formalmente, ele considera o conjunto de possíveis estados da natureza, sendo a informação descrita como uma partição desse conjunto. Assim, a aquisição de informação é uma partição de um intervalo unitário em subconjuntos. Por exemplo, considere n *sucessivas biseções*, onde cada subintervalo tem igual comprimento $\Delta = 2^{-n}$ e assim adquirir informação significa a *revelação*[133] de algum inteiro $k \in [1, 2, \ldots 2^n]$ de forma a reduzir a incerteza de uma variável X com incerteza tecnológica. Por exemplo, se $X \sim U[0, 1]$, então dado k pode-se saber que X cairá num subintervalo (a, b] onde $a = (k - 1)\Delta$ e $b = k\Delta$ (Wilson, 1975, p.186). No caso de nenhuma informação então n = 0, Δ = 1, k = 1 (pois k vale no máximo 2^{-n}) e assim a = 0 e b = 1, ou seja, a distribuição *a posteriori* de X é igual a sua distribuição *a priori*.

Mas o mais interessante no trabalho de Wilson (1975) é que ele mostrou vários exemplos em que o valor da informação *por unidade de escala* é estritamente côncavo, enquanto que o valor da informação *líquido do seu custo*[134] de aquisição é estritamente convexo. Isso é devido ao fato que, enquanto o custo *unitário* da informação diminui com o aumento de escala, o valor unitário da informação não diminui. Assim, Wilson conclui que a demanda por informação pode ser ilimitada, embora na prática a firma não possa aumentar a escala indefinidamente e nem possa refinar sua informação perfeitamente nesse contexto.

Desde um ponto de vista mais prático, quando se considera informação custosa, pode haver aquisições *inúteis* de informação – informação insuficiente para mudar a ação ótima que se tomaria sem essa informação; aquisições *antieconômicas* de informação – informação é útil para mudar a ação ótima, mas não o suficiente comparado a seu custo; aquisições *econômicas* de informação – quando o valor supera o custo; e *desperdício* ou aquisição excessiva de

[133] Nota histórica: o termo "*revelation*" foi usado pelo próprio Wilson (1975, p.186).

[134] Assume que o custo é uma função convexa da informação, por ex., o custo de n biseções é proporcional a n. Usando o conceito de *entropia* (ver item 3.1.4.3), o custo de descobrir um subintervalo de comprimento Δ é proporcional a – log(Δ), Wilson (1975, p.189).

informação – quando é adquirida mais informação do que a necessária (também é antieconômico). Seja um índice de *grau de informação* ("*informativeness*") θ, que pode ser a redução percentual esperada de variância (ou *poder de revelação*), o tamanho da amostra, um índice baseado na função verossimilhança, ou outra qualquer. Seja $\theta \in [0, 1]$. A Figura 32 a seguir, baseada em Lawrence (1999, p.274), ilustra essas diferentes situações de economicidade do valor líquido da informação.

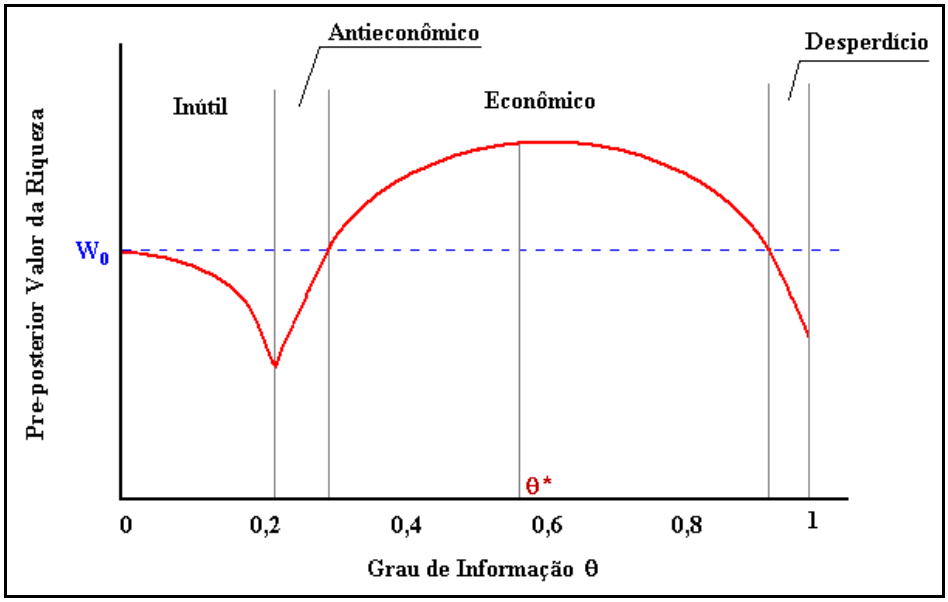

Figura 32 – Valor Líquido da Informação para Vários Graus de Informação

Na Figura 32 se vê que só a partir de $\theta > 0{,}2$ é que a informação pode ser útil e só a partir de $\theta > 0{,}3$ é que a aquisição de informação custosa (custo proporcional a θ) se torna econômica. Sendo essa função côncava na região em que $\theta > 0{,}2$, então existe uma aquisição ótima de informação em θ*. Note que a partir de $\theta > 0{,}9$ ocorre um desperdício de informação, pois devido ao seu custo a revelação total ($\theta = 1$) é antieconômica nesse exemplo. No ponto ótimo de aquisição da informação em $\theta^* \cong 0{,}6$, a riqueza é maximizada com a aquisição de informação. Mas na prática, nem sempre as alternativas de aquisição (ou investimento) em informação são disponíveis para qualquer valor θ. De qualquer maneira, o gráfico acima sugere o objetivo da primeira aplicação do cap. 5 que é selecionar a alternativa ótima de investimento em informação, de um conjunto discreto de alternativas. Em E&P de petróleo, esse conjunto é constituído de alternativas de projetos de investimento em informação, com diferentes custos, tempos de aprendizagem e poderes de revelação (diferentes η^2).

3.1.4.2.
Visão Estatística do VOI e Comparação de Experimentos

A teoria estatística em geral e o ramo chamado *teoria estatística da decisão*, em particular, sempre se preocuparam com questão da informação, por ex., o tamanho ótimo de uma amostra para fazer uma previsão dentro de um certo intervalo de confiança. A teoria estatística da decisão, especialmente a visão Bayesiana, teve grande influência na escola de *análise de decisão*. Por ex., Raiffa é autor não só do clássico de análise de decisão como co-autor de um clássico de teoria estatística da decisão, em Raiffa & Schlaifer (1961) e numa versão mais moderna com a colaboração de Pratt em Pratt & Raiffa & Schlaifer (1995).

Um dos mais influentes trabalhos sobre VOI desde um ponto de vista estatístico, mas com relevância em aplicações econômicas, foram os artigos de Blackwell (1951 e 1953), ver também o livro de Blackwell & Girshick (1954, cap.12). Blackwell foi um dos pioneiros[135] da teoria da *comparação de experimentos*, que compara estruturas de informação que possam ser preferidas a outras estruturas, independentemente da distribuição a priori assumida.

Como nota Arrow (1992, p.169), "*seria útil ter pelo menos um ordenamento parcial dos sinais que seja independente do problema de decisão*". Esse foi o ambicioso objetivo teórico de Blackwell. Ele obteve, no entanto, condições restritas em que se pode realmente fazer esse ordenamento genérico. Antes de examinar os resultados de Blackwell é oportuno fazer algumas definições estatísticas que serão úteis nesse e/ou em outros itens dessa tese.

A visão estatística de valor da informação é originalmente derivada do conceito de estatística suficiente ("sufficient statistics") T(S) onde "*o valor da informação a qual está contida na estatística T(S) é a mesma que o valor da informação da amostra original S*" (Jammernegg, 1988). Dessa forma, pode-se usar a estatística T(S) em vez de toda a amostra S, já que T(S) sintetiza o essencial da amostra S nesses problemas. Formalmente, uma estatística T(S) é suficiente para X se a distribuição de probabilidade condicional dos dados S dada a estatística T(S), não depender de X. O conceito de estatística suficiente tem um papel importante na teoria original de Blackwell.

[135] Segundo o próprio Blackwell (1951, summary) e Torgersen (1991, p.222), essa literatura começou em 1949 com um artigo não publicado de Bohnenblust & Shapley & Sherman.

O conceito de estatística suficiente tem a seguinte limitação apontada por DeGroot & Schervish (2002, p.387): "*a existência e forma de uma estatística suficiente em um problema particular depende criticamente da forma da função assumida para a distribuição de probabilidade*". Ou seja, uma estatística suficiente para uma distribuição $f(x \mid \theta)$ pode não ser estatística suficiente para uma outra distribuição $g(x \mid \theta)$.

Informação de Fisher $I(\xi)$ é um indicador estatístico clássico da *quantidade de informação* que uma amostra de dados contém sobre um parâmetro ξ de valor incerto. Menor incerteza significa dados mais informativos e logo um valor $I(\xi)$ maior. Uma propriedade importante de $I(\xi)$ é que numa amostra aleatória de n observações iid (independentes e identicamente distribuídos), a informação de Fisher $I_n(\xi)$ é simplesmente n vezes a informação de Fisher de uma única observação $I(\xi)$, isto é, $I_n(\xi) = n\, I(\xi)$. Prova: DeGroot & Schervish (2002, p.438). No item 3.4.1 sobre distribuição de Bernoulli, se mostrará que *não* é conveniente trabalhar com $I(\xi)$ como medida de aprendizagem em problemas de VOI.

A idéia da teoria de comparação de experimentos é responder a pergunta: dados dois experimentos ou duas estruturas de informação S e S', quando se pode dizer que uma a estrutura S é sempre melhor que a estrutura S'? Uma outra maneira de dizer a mesma coisa é com o conceito de estatística suficiente: quando que a estrutura de informação S é estatisticamente suficiente para S'? Torgersen (1991) usa também a frase: quando que a estrutura S é mais informativa que S'? Em Blackwell (1953), um experimento S é suficiente (ou mais informativo) que o experimento S' se existe uma transformação estocástica de S para uma v.a. Z(S) tal que, para toda a distribuição a priori p(x), as v.a. Z(S) e S' tem distribuições idênticas. Nem sempre é fácil conseguir um caso real que pode ser assim provado.

Conforme aponta Arrow (1992), pode-se ver claramente que um sinal é pelo menos tão bom quanto outro, quando o outro sinal é apenas um sinal misturado ("garbling") do primeiro. Ou seja, é apenas um *sinal do* primeiro *sinal*, como se pegasse o primeiro sinal e se adicionasse apenas ruído. Arrow argumenta ainda que em vez do conceito de mistura/alteração de sinal, pode ser mais conveniente trabalhar com os conceitos de suficiência e quase-suficiência estatística, esse último definido por ele usando a função verossimilhança.

Essas comparações de experimentos são feitas com base apenas em mérito estatístico, ignorando o custo do experimento (ou custo da aquisição de informação), ou seja, avalia-se o mérito de uma estrutura de informação antes de considerações sob seu custo. Se as estruturas tiverem o mesmo custo, a estrutura mais informativa será preferível. Mas a teoria de comparação de experimentos sozinha não ajuda muito na valoração do "trade-off" em caso de uma estrutura mais informativa, porém mais cara que outra. Também não considera que uma estrutura mais informativa poderia ser um projeto de investimento em informação não só mais caro como com maior tempo de aprendizagem. Essas importantes questões estarão contempladas no modelo proposto nessa tese.

Arrow (1992) mostra a interpretação econômica dessa teoria apontando que um sinal é pelo menos tão bom quanto outro no sentido da informação, se o resultado ("payoff") esperado baseado no primeiro sinal for pelo menos tão grande quanto o resultado esperado baseado no outro sinal, para todo problema de decisão. A seguinte definição (adaptada de Lawrence, 1999, p.198) é útil no contexto de OR: dada a opção real F(X, t), detida por um único decisor, a estrutura $\mathcal{I}(\theta)$ é estatisticamente mais informativa que $\mathcal{I}(\theta')$ se e somente se:

$$\mathbf{E[F(X, t) \mid \mathcal{I}(\theta)] \geq E[F(X, t) \mid \mathcal{I}(\theta')]} \qquad \mathbf{(62)}$$

Para todos os problemas de decisão relacionados, i. é, independente da distribuição a priori p(x). Em palavras, o valor esperado da opção real dada uma estrutura de informação mais informativa é igual ou maior que o valor esperado dessa opção real para uma estrutura menos informativa. Como antes, X é um vetor de variáveis aleatórias de interesse e t é o tempo.

Lawrence (1999, p.197-204) mostra que existem pelo menos 4 condições equivalentes para caracterizar que uma estrutura é mais informativa que a outra, valendo para distribuições a priori arbitrárias. Primeiro, a comparação via verossimilhança, que foi a abordagem original de Blackwell. A segunda condição equivalente é a comparação via probabilidades posteriores. A terceira é a comparação via possíveis utilidades. A quarta é a comparação via funções incerteza. Conforme aponta Kihlstrom (1984a), embora equivalentes, nem todas

são igualmente convenientes em cada aplicação. Por ex., a literatura de *planejamento seqüencial de experimentos* usa o critério da função incerteza[136].

A quarta condição, que usa uma função incerteza côncava (ex.: entropia, ver item 3.1.4.3), foi mostrada por Blackwell e por DeGroot (1962) e pode ser enunciada da seguinte maneira: *uma estrutura de informação é mais informativa que outra se e somente se reduz mais a incerteza esperada do que a outra*. Ver também a demonstração em Kihlstrom (1984a, definition 4 e theorems 4, 5 e 6).

O detalhamento desse tema foge do escopo dessa tese, mas a última condição sugere justamente fazer o que se propõe nessa tese e considerar uma estrutura de informação relacionada à redução esperada de incerteza. Na tese será considerada a seguinte *flexível* estrutura de informação alternativa para o caso de uma variável de interesse X e um sinal específico S.

$$\mathcal{I} = \{\eta^2(\mathbf{X} \mid \mathbf{S}), \bullet\} \qquad \textbf{(63)}$$

Onde $\eta^2(X \mid S)$ é a redução esperada de variância em X devido à revelação de informação S (a ser discutida no item 3.3) e o símbolo $\bullet$ indica um critério adicional para definir totalmente a estrutura da informação, sendo que o critério mais conveniente depende da aplicação. Em alguns casos (aplicação do cap. 5) esse critério pode ser simplesmente assumir que o *tipo* (se normal, triangular, etc.) de distribuição de revelações é o mesmo do caso *limite* de informação perfeita; em outras aplicações (caso exploratório, item 3.4) o critério pode ser assumir que as variáveis aleatórias X e S são *intercambiáveis*. Isso será discutido ainda nesse capítulo. Mas $\bullet$ pode ser também o critério mais convencional, i. é, a distribuição conjunta p(x, s). Em todos esses casos, o critério $\bullet$ junto com η^2, definem totalmente o impacto de S sobre X em qualquer problema de VOI. Mas outros critérios podem ser mais fáceis ou mais práticos do que a distribuição conjunta p(x, s). O importante é que a estrutura seja consistente, i. é, a combinação de η^2 com $\bullet$ deve ser feita dentro de certos limites, pois a distribuição conjunta p(x, s) tem de obedecer aos chamados limites de Hoeffding-Fréchet (ver item 3.3.1).

A estrutura de informação da eq. 63 é mais flexível pois dá liberdade para trabalhar com critérios mais convenientes para a aplicação, mas sempre

[136] O problema de planejamento de experimentos é seqüencial quando a escolha de um experimento em qualquer estágio é feita com o conhecimento dos resultados dos experimentos anteriores. O critério de escolha é a minimização da incerteza esperada medida por essa função.

acompanhada de uma unidade de medida de aprendizagem η^2. Essa estrutura flexível será usada em diversas aplicações.

Para fechar esse item é oportuno estabelecer o famoso teorema de Bayes e os fatos elementares sobre distribuições condicionais. Sejam os eventos $X_1, X_2, \ldots X_k$, partições do espaço X tal que $Pr(X_j) > 0$ para j = 1, 2, … k e seja S um evento tal que $Pr(S) > 0$. Então, para todo i = 1, 2, … k, o teorema de Bayes estabelece:

$$\mathbf{Pr(X_i \mid S) = \frac{Pr(X_i)\, Pr(S \mid X_i)}{\sum_{j=1}^{k} Pr(X_j)\, Pr(S \mid X_j)}} \qquad \mathbf{(64)}$$

Como se pode notar, o teorema de Bayes trabalha com probabilidades inversas $Pr(S \mid X_i)$ para determinar a probabilidade condicional $Pr(X_i \mid S)$.

Seja a distribuição de probabilidades conjunta da variável de interesse X e do sinal S, dada pela distribuição conjunta p(x, s). Seja p(s) > 0 a distribuição (marginal) de probabilidades do sinal S. A distribuição posterior é dada por:

$$\mathbf{p(x \mid s) = \frac{p(x, s)}{p(s)}} \qquad \mathbf{(65)}$$

Seja a distribuição a priori p(x) > 0. A *probabilidade inversa* ou *confiabilidade* de S ou ainda verossimilhança de S, é dada analogamente por:

$$\mathbf{p(s \mid x) = \frac{p(x, s)}{p(x)}} \qquad \mathbf{(66)}$$

Fixando s e vendo p(s | x) como uma função de x, se define a chamada função verossimilhança. Combinando as eqs. 65 e 66, chega-se à seguinte equação que também é chamada de teorema de Bayes (Lawrence, 1999, p.59):

$$\mathbf{p(x \mid s) = \frac{p(s \mid x)\, p(x)}{p(s)}} \qquad \mathbf{(67)}$$

A combinação da distribuição a priori p(x) com a verossimilhança p(s | x) é uma abordagem comum e popular entre os estatísticos para obter a distribuição conjunta p(x, s), mas certamente não é a única (Lawrence, 1999, p.59). Apesar do merecido sucesso da função verossimilhança na teoria e prática estatística (ex., a estimação por *máxima verossimilhança* é uma das mais importantes em estatística) e seu amplo uso em testes de hipóteses, não se pode pensar que isso automaticamente a credencia como o melhor elemento para estruturas de informação em problemas de valor econômico da informação. Apesar do seu amplo uso na literatura acadêmica *econômica* de VOI por Arrow, Marschak,

Radner & Stiglitz, etc., essa tese irá propor trabalhar com outra estrutura de informação que usa a redução esperada de incerteza em vez da função verossimilhança. As várias vantagens de se fazer isso serão discutidas nos itens 3.2 (expectativas condicionais) e no item 3.3 (medidas de aprendizagem).

Uma maneira de facilitar a matemática envolvida com essas distribuições em aplicações da teoria estatística da decisão é o uso de *distribuições conjugadas*. Distribuições conjugadas pertencem a uma família de distribuições com a seguinte propriedade: se a distribuição a priori escolhida é um membro dessa família conjugada, então a distribuição posterior também pertence a essa família (ver DeGroot & Schervish, 2002, seção 6.3).

3.1.4.3. O Conceito de Entropia e a Teoria da Informação

A teoria da informação (também chamada de *teoria estatística da comunicação*) foi desenvolvida por Shannon nos anos 40 tendo em mente aplicações em engenharia de comunicação. Mas o caráter inovador dessa teoria aliado a sua elegância matemática, a fez ter grande impacto não só na engenharia como em áreas diversas tais como estatística teórica e economia da informação. Essa teoria, que usa o conceito de entropia como medida de incerteza, foi sintetizada no livro de Shannon & Weaver (1949). Uma discussão crítica das bases matemáticas dessa teoria é feita por Khinchin (1953).

Na área de matemática, o livro texto de estatística teórica de Kullback (1959) teve um grande impacto e iniciou um novo ramo na estatística, que continua com grande prestígio nos dias de hoje. Kullback usou esses conceitos para estabelecer uma famosa medida de divergência entre distribuições que dá a distância entre duas distribuições. Outra importante contribuição é o livro texto de Jaynes (2003), um inovador e crítico tratado sobre os fundamentos da teoria da probabilidade, que discute também o famoso *princípio da máxima entropia* para a distribuição a priori (será vista aqui). Poderiam ser mencionadas muitas outras importantes contribuições do conceito de entropia e da teoria da informação[137].

[137] Por ex., Kolmogorov na década de 50 usou fatos matemáticos sobre entropia para mostrar que nem todos os processos iid são mutuamente isomórficos. Isso iniciou um grande desenvolvimento na *teoria ergódica* em que a entropia tem um papel chave. A *teoria da complexidade* de Kolmogorov, também usa a teoria da informação como base matemática.

A seguir serão apresentadas as definições de entropia e informação mútua para o caso discreto. Considere uma distribuição a priori discreta univariada p(x). No sentido de Shannon, entropia (H) é a medida de incerteza definida por[138]:

$$\mathbf{H(X) = -\sum_i p(x_i)\log[p(x_i)]} \quad \mathbf{(68)}$$

A base do logaritmo na equação da entropia é arbitrária. Conforme mostra Shannon (p.32-33), uma escolha conveniente poderia ser a base 2, de forma que a entropia é convenientemente medida em "bits"[139]. Já em trabalhos analíticos onde se trabalha com operações de diferenciação e integração, a base neperiana "*e*" é mais útil. Na definição de entropia se poderia multiplicar o somatório por uma constante positiva para mudança de escala, que foi a equação original de Shannon.

Note na eq. 68 que para o caso de revelação total (ou informação perfeita), a entropia colapsa para zero pois log(1) = 0. Mais precisamente, $H(X) = 0 \Leftrightarrow$ um dos números $p(x_i)$ é igual a um e todos os outros $p(x_j)$, $j \neq i$, são iguais a zero (Shannon & Weaver, 1949, p.51). Para todos os outros casos a entropia é estritamente positiva.

Uma aplicação de interesse da tese é o estabelecimento da distribuição a priori usando o chamado *princípio da máxima entropia*. Esse princípio é muito usado em exploração de petróleo, especialmente na área de geofísica (interpretação de registros sísmicos). Esse princípio diz que as probabilidades a priori devem ser atribuídas de forma a ser a de maior entropia dentre aquelas consistentes com o conhecimento a priori. A intuição é que dessa forma se obtém uma distribuição que assume o mínimo da variável incerta (distribuição *honesta*).

Jaynes (1968) recomenda que a determinação da distribuição a priori seja feita com o princípio da máxima entropia conjugada com a chamada *teoria de grupos*[140], muito usada em física teórica. A forma da distribuição a priori é unicamente determinada se o domínio fundamental do grupo se reduz a um ponto. A aplicação de métodos analíticos da física em finanças e em economia tem crescido muito, geralmente com enfoque muito mais teórico que prático e muita

[138] Assuma que 0 . log(0) = 0. Entropia também pode ser interpretada como o valor esperado de log[1/p(x)].

[139] Bits = "binary digits". Uma chave liga-desliga armazena um bit de *informação*, ver Shannon & Weaver, 1949, p.32.

[140] Grupo é um par (A, •) onde A é um conjunto não vazio e • é uma operação binária em A que obedece três propriedades (associativa, identidade e inversa), ver Kholodnyi (1998, p.13-14).

vezes competindo com métodos tradicionais mais simples. Mas essa hibridização sempre abre novas possibilidades. A teoria de grupos, por ex., foi usada em Kholodnyi (1998) para formalizar matematicamente o conceito de *simetria*[141] na valoração de ativos contingentes, especialmente em mercados cambiais.

Conforme aponta Arrow (1972, p.134), Shannon vê H(X) também como *quantidade de informação*. Isso porque uma proposição de Shannon diz que um *canal de comunicação* com capacidade H não será uma restrição, i. é, tem capacidade de enviar uma mensagem sobre o real estado da natureza de X com um erro arbitrariamente pequeno. No entanto, é muito mais útil e adequada a interpretação da entropia como medida de incerteza.

Na definição de entropia (eq. 68), note que não interessa os valores dos cenários, apenas as probabilidades dos diferentes cenários são consideradas. Assim, a entropia é uma medida adimensional, o que trás vantagens matemáticas importantes, mas também desvantagens. A entropia tem a vantagem da simplicidade para representar a incerteza. Mas, em problemas de *valoração econômica* onde a *magnitude* de perdas e ganhos tem importância[142], essa característica pode ser freqüentemente uma desvantagem ou será preciso uma variável adicional para conjugar probabilidades e valores dos cenários.

Para o caso de distribuição de probabilidades p(x) contínua, a entropia é de forma análoga definida por Shannon & Weaver (1949, p.87) como:

$$\mathbf{H(X) = -\int_X p(x)\ log[p(x)]\ dx} \qquad \mathbf{(69)}$$

No entanto, usando o limite do caso discreto, Jaynes (1968, eq.40) argumenta que é preferível usar log[p(x)/m(x)] no lugar de log[p(x)] na eq.69, onde m(x) é uma *função de medida invariante*. O livro texto de Cover & Thomas (1991, cap.9) usa a eq.69, mas chamando-a de *entropia diferencial* e alerta de que são necessários certos cuidados no uso do conceito de entropia para distribuições contínuas. Por ex., no caso de uma distribuição uniforme contínua no intervalo [0, a], a qual tem p(x) = 1/a, aplicando a eq.69 obtém-se h(X) = log(a). Mas se a < 1 $\Rightarrow$ h(X) < 0. Assim, ao contrário da entropia discreta, a entropia diferencial pode

[141] A teoria de grupos e conceitos como "gauge symmetry" deram a formalidade matemática que permitiu uma revolução na física no século XX (Kholodnyi, 1998).

[142] No projeto de um *canal de comunicação* isso não é problema pois os cenários são mensagens medidas em bits e não quantidades numéricas.

ser negativa. Isso complica o uso desse conceito em aplicações com distribuições contínuas.

A *entropia condicional* de X dado S, também conhecida como *entropia esperada* ou *equivocação* ("equivocation", ver McEliece, 2002, p.20), é a entropia média das distribuições posteriores p(x | s) para todos os possíveis valores do sinal S, i. é, para o caso discreto o conceito de entropia condicional é definida como:

$$\mathbf{H(X \mid S) = -\sum_{x}\sum_{s} p(x, s) \log[p(x \mid s)]} \quad \mathbf{(70)}$$

Em palavras, a entropia condicional mede a incerteza esperada remanescente sobre X depois de S ter sido observado. Também pode ser vista como o valor esperado de log[1/p(x | s)]. Note que H(X | S) ≤ H(X), com igualdade ocorrendo se e somente se X e S forem independentes (prova: Cover & Thomas, 1991, p.27-28). Em geral essa medida é *assimétrica*, e por isso seria uma candidata a medida de aprendizagem (ver item 3.3.2).

A diferença entre a entropia (incondicional) H(X) e a entropia condicional H(X | S) é uma medida de redução esperada de incerteza devido à informação revelada por S, onde a medida de incerteza é a entropia. Essa quantidade é chamada de *informação mútua*, ou *informação transmitida* ou *incerteza removida* (Lawrence, 1999, p.62) ou ainda *taxa de transmissão* (Shannon), é definida por:

$$\mathbf{I(X; S) = H(X) - H(X \mid S)} \quad \mathbf{(71)}$$

A informação mútua pode ser vista como uma medida de quantidade de informação que uma variável aleatória (v.a.) tem da outra. Esse é o conceito mais importante da teoria da informação no contexto dessa tese por ser uma medida de redução de incerteza. Além disso, num contexto teórico de utilidade logarítmica, I(X; S) é o próprio VOI (Arrow, 1972, p.136-137). Será visto que a medida de aprendizagem proposta na tese está intuitivamente associada à idéia da eq. 71, mas usando variância no lugar da entropia para medir incerteza. No caso de variáveis aleatórias discretas, a informação mútua pode ser escrita como:

$$\mathbf{I(X;S) = \sum_{x}\sum_{s} p(x, s) \log\left[\frac{p(x, s)}{p(x)\, p(s)}\right]} \quad \mathbf{(72)}$$

Para o caso de distribuições contínuas, a informação mútua é escrita como:

$$\mathbf{I(X;S) = \iint_{X\,S} p(x, s) \log\left[\frac{p(x, s)}{p(x)\, p(s)}\right] dx\, ds} \quad \mathbf{(73)}$$

Uma propriedade importante da informação mútua é a simetria, que vale tanto no caso discreto como contínuo, i. é:

$$\mathbf{I(X; S) = I(S; X)} \tag{74}$$

Além disso, I(X; S) ≥ 0, com a igualdade ocorrendo se e somente se X e S forem independentes (prova: Cover & Thomas, 1991, p.27). Outra propriedade é que se T(S) for estatística suficiente, então I(X; T(S)) = I(X; S), ou seja, a estatística suficiente preserva a informação mútua e vice-versa.

Um conceito muito usado em probabilidade, inclusive no estudo de medidas de dependência de variáveis aleatórias, e que é relacionado a I(X; S) é o de *entropia relativa*, mais conhecida como distância de Kullback-Leibler entre duas distribuições de probabilidade p(x) e q(x). Essa "distância" no caso discreto é[143]:

$$\mathbf{D(p \,||\, q) = \sum_x p(x) \log\left[\frac{p(x)}{q(x)}\right]} \tag{75}$$

Entretanto, D(p || q) *não* é uma verdadeira distância pois ela não é uma medida simétrica, ao contrário da mútua informação. Tem-se que D(p || q) ≥ 0, com a igualdade ocorrendo se e somente se p(x) = q(x) para todo x. Para o caso de distribuições contínuas, a medida de Kullback-Leibler é dada por:

$$\mathbf{D(p \,||\, q) = \int_x p(x) \log\left[\frac{p(x)}{q(x)}\right] dx} \tag{76}$$

Tem-se que D(p || q) é finito apenas se o suporte de p(x) está contido no suporte de q(x). Ela pode ser definida para um vetor X de n v.a. (na eq. 76, substituir q(x) por um produtório de n distribuições $q(x_i)$), e por isso alguns autores dizem que D(p || q) é uma generalização da informação mútua para o caso multivariado. A medida chamada *divergência* (Kullback, 1959, p.6) é definida pela soma das medidas de Kullback-Leibler:

$$\mathbf{J(p; q) = D(p \,||\, q) + D(q \,||\, p)} \tag{77}$$

Arrow (1972) faz uma análise crítica do uso da teoria da informação para a análise de VOI no contexto da teoria econômica. Conforme também apontado por Marschak (1959, p.81), se for assumido que o custo de um canal de comunicação é proporcional à sua capacidade, então a entropia é uma medida de preço de *suprimento* da informação e não o valor da *demanda* por informação. Mas Arrow (1972, p.134) aponta que foram feitos esforços para tentar interpretar a entropia

[143] Com as convenções: 0 log(0/q) = 0 e p log(p/0) = ∞.

como valor (no sentido de demanda) da informação, inclusive por Marschak (1959, p.92-95). A limitação da entropia nesse caso vem da sua própria definição que só considera as *probabilidades* dos cenários e não o *valor* dos cenários. Por isso que a entropia só pode ser usada diretamente em problemas de VOI quando a função utilidade é logarítmica, pois *se o VOI é independente das remunerações, então a função utilidade tem de ser logarítmica* (prova: Arrow, 1972, p.134-135).

No caso de utilidade logarítmica, Arrow (1972, p.136-137) que o VOI é precisamente a informação mútua (ou incerteza removida) definida anteriormente. Isso é mais um indício da ligação direta e estreita do VOI com o conceito de redução esperada de incerteza. Mas Arrow (1972, p.138-139) aponta que a simplicidade da solução do caso logarítmico não se generaliza e assim o VOI não irá depender *apenas* da informação mútua/remoção de incerteza.

Lawrence (1999, p.6) faz uma crítica mais aguda à aplicação da teoria da informação de Shannon no contexto econômico dizendo que ela é mais apropriada para o estudo de armazenagem e transmissão de *dados* e não de *informação*. Muitos matemáticos discordarão dele. Mas apesar das inegáveis limitações da teoria da informação no contexto econômico, a sua simplicidade permite fazer úteis e interessantes analogias como a que será feita comparando o conceito de informação mútua com a variância da distribuição de revelações (item 3.2).

3.1.4.4. Análise de Sensibilidade Global

Uma literatura que pode ser relacionada com medidas de aprendizagem (de interesse em aplicações de VOI) recentemente descoberta pelo autor é com área de modelagem científica chamada de *análise de sensibilidade* (Saltelli et al., 2004). Essa literatura é proveniente da área de *física computacional* (exs.: Hofer, 1999; e Jansen, 1999) e de *engenharia de confiabilidade* (ex.: Saltelli, 2002).

Quando se pensa em análise de sensibilidade se imagina a *derivada parcial* da variável de interesse (X) em relação a uma variável de entrada (S_j), o que teria pouco ou nada a ver com incerteza e com os problemas dessa tese. Mas a derivada parcial é uma medida de sensibilidade *local*, enquanto que Saltelli et al. (2004)

está preocupado com medidas e técnicas *globais* de sensibilidade e envolvendo incerteza nos dados de entrada e no valor de saída[144].

De acordo com Saltelli et al. (2004, preface), a literatura de análise de sensibilidade global (ASG) em modelagem científica está preocupada com problemas tais como: escolha de um modelo versus outro; seleção de tamanho de malha; seleção de diferentes conceituações sobre um sistema, etc. Técnicas de ASG são usadas para acessar a importância relativa dos fatores de entrada do modelo. Por isso esses autores chamam as medidas de ASG de *medidas de importância*. A literatura sobre ASG procura responder perguntas tais como: "*qual dos fatores de entrada (inputs) é o mais importante na incerteza da variável de interesse (output)?*"; e "*se pudermos eliminar a incerteza em um dos fatores de entrada, qual fator deveremos escolher para reduzir ao máximo a variância do resultado?*". Essa última pergunta tem uma clara conexão com a questão de medida de aprendizagem que é um dos focos dessa tese.

3.2. Expectativas Condicionais e Distribuição de Revelações

3.2.1. Expectativas Condicionais

Expectativa condicional E[X | S] é uma variável aleatória que assume o valor E[X | S = s] com probabilidade Pr(S = s)[145]. Ela pode ser vista também como uma função, i. é, E[X | S] é uma função de S[146]. Será mostrada a definição rigorosa de E[X | S] para o caso geral de variáveis aleatórias arbitrárias usando a integral de Lebesgue-Stieltjes e alguns outros conceitos da teoria da medida. Mas antes são mostradas abaixo as equações de E[X | S = s], respectivamente para os casos de variáveis aleatórias ambas discretas e ambas contínuas.

$$\mathbf{E[X \mid S = s\,]} = \sum_{x} \mathbf{x\ p(x \mid s)} \qquad \mathbf{(78)}$$

[144] Infelizmente a literatura de análise de sensibilidade global (ASG), em especial o primeiro livro texto de Saltelli et al. (2004), só chegou ao conhecimento do autor em novembro de 2004, com a tese praticamente pronta. Assim, a sugestão é que no futuro seja explorado melhor a conexão e aplicabilidade de métodos de ASG em problemas de valor da informação.

[145] Ver o livro de Sheldon Ross (1998) para uma boa introdução ao conceito de expectativa condicional sem usar teoria da medida e o livro de Williams (1991) para o caso mais geral com a teoria da medida.

[146] O item 3.2 é parcialmente baseado em Dias (2002).

$$\mathbf{E[X \mid S = s\,]} = \int_{-\infty}^{+\infty} \mathbf{x\ p(x \mid s)\ ds} \qquad \mathbf{(79)}$$

Conforme foi visto antes (rever a discussão sobre a Figura 28), uma revelação parcial de informação através de um sinal S trazendo informação imperfeita sobre o verdadeiro estado da natureza da variável de interesse X, gera um conjunto (infinito no caso de sinais com distribuição contínua) de distribuições posteriores p(x | s). A média de cada distribuição posterior (revelada no cenário S = s) é uma expectativa condicional E[X | S = s], que ex-ante é uma variável aleatória (v.a.) que depende do sinal revelado s. Foi visto que é muito mais simples trabalhar com a (única) distribuição de expectativas condicionais do que com as (infinitas) distribuições posteriores. Além disso, após receber a informação S, o decisor não sabe qual cenário da distribuição posterior é o verdadeiro, conhece apenas a distribuição posterior. Assim, tudo que ele pode fazer é usar alguns momentos dessa distribuição para tomar decisões econômicas (ainda sob incerteza) adaptadas a essa informação. O principal momento de uma distribuição posterior é exatamente a expectativa condicional E[X | S]. Nas aplicações de VOI, além de E[X | S], pode-se usar a distribuição de Var[E[X | S]]. Mas como é um efeito de segunda ordem na otimização, será *muito* mais simples usar a variância *esperada* das distribuições posteriores E[Var[E[X | S]]] em conjunção com a distribuição de E[X | S].

A alternativa com métodos Bayesianos tradicionais para modelar a incerteza técnica em problemas de VOI usando a *função verossimilhança* para determinar as possíveis distribuições posteriores, é bem mais complexa para ser inserida em modelos dinâmicos de opções reais e/ou apresenta computação muito mais intensiva, além de outras desvantagens que serão mostradas. Mas mesmo nessa literatura tradicional, é comum trabalhar com a expectativa condicional dentro da equação de maximização de valor, ver, por ex., McCardle (1985, eqs.3 e 4)[147].

Em economia é muito usada a expectativa condicional em várias aplicações, especialmente em econometria. Goldberger (1991, preface) adota a função expectativa condicional "*as the key feature of a multivariate population for*

[147] Entretanto McCardle não desenvolveu ou usou as propriedades da distribuição de expectativas condicionais, como é feito aqui.

economists who are interested in relations among economic variables… – a very simple concept"[148].

Em finanças esse conceito é fundamental, de forma que livros texto de cálculo estocástico aplicado a finanças dedicam espaço significativo para discutir o conceito de expectativa condicional (exs.: Mikosch, 1998; Shreve, 2004). Conforme Tavella (2002, p. vii, viii, 4), esse conceito tem um lugar natural em engenharia financeira computacional onde joga um papel chave. Isso porque o preço de um derivativo é simplesmente um valor presente esperado de um valor futuro condicional a um conjunto de informação.

O uso de expectativa condicional como base para decisões tem também uma forte base teórica. Imagine uma variável X com incerteza técnica e seja a nova informação ou sinal S uma v.a. definida no mesmo espaço de probabilidades (Ω, Σ, P). O objetivo é obter a melhor estimativa de X observando S, através do uso de uma função g(S). A medida mais usada para a qualidade de um previsor g(S) é o seu *erro quadrático médio* ("mean square error") definido por $MSE(g) = E[X - g(S)]^2$. A escolha da função g* que minimiza MSE(g) é exatamente a função expectativa condicional $E[X \mid S]$. Essa é uma propriedade muito conhecida em econometria (ver, por ex., Gallant, 1997, p.64-65). Em adição, o erro $\varepsilon = X - E[X \mid S]$ tem covariância zero com toda *função* de S (ver Goldberger, 1991, p. 53).

A expectativa condicional $E[Y \mid X = x]$ é também o (melhor) valor da *regressão* de Y versus X para X = x. A melhor regressão pode ser *linear* mas em geral é *não-linear* (ver Whittle, 2000, p.89)[149]. Ex.: uma regressão *quadrática* Y versus X minimizará o $MSE(a, b, c) = E[Y - a - bX - cX^2]^2$ escolhendo os parâmetros a*, b*, c*. Se essa função quadrática for a melhor função no sentido de minimizar o MSE, então para X = x tem-se $g(x; a^*, b^*, c^*) = E[Y \mid X = x]$. Gallant (1997) provou que o melhor *polinômio* previsor (*regressão polinomial*) de ordem N é a expectativa condicional quando N tende ao infinito.

Em adição, conforme apontam DeGroot & Schervish (2002, p.348), a expectativa condicional é o (melhor) estimador de Bayes quando é usado o critério do erro quadrático como função perda ("loss function") na avaliação dos

[148] Além disso, os estatísticos sabem que em certos problemas estatísticos uma regressão linear é equivalente, mas bem mais simples que os métodos Bayesianos tradicionais e os que usam a máxima verossimilhança. Será vista a ligação direta entre regressões e expectativas condicionais.

[149] Ver também Gallant (1997, p.109-112) para o caso não-linear.

estimadores. Esses autores lembram que essa função perda é de longe a mais usada em problemas de estimação. Além disso, se S é estatística suficiente para X, então o estimador E[X | S] domina *qualquer* outro estimador de X (DeGroot & Schervish, 2002, p.385) e, assim, qualquer outro estimador é chamado de *inadmissível* (DeGroot & Schervish, 2002, p.386).

Já que a informação S = s *revela* E[X | S = s] e dado o grande uso da palavra "revelation" e "information revelation" na literatura de economia de informação (exs.: Wilson, 1975, p.186; Drazen & Sakellaris, 1995; Dutta & Morris, 1997; Chamley & Gale, 1994) e na literatura de opções reais (exs.: Grenadier, 1999; Childs & Ott & Riddiough, 2001, 2002a, 2002b), se adotará a notação "revelação" $R_X(S)$ para denotar a função expectativa condicional E[X | S], de forma que a distribuição de revelações será usada para caracterizar a distribuição da v.a. R_X. Assim o termo "revelação" enfatiza a mudança de expectativas com o novo cenário revelado pela informação, i. é, ressalta o *processo de aprendizagem* ou *processo de descoberta* em direção ao verdadeiro valor da variável[150]. Assim, será usada a seguinte equação para definir a função $R_X(S)$:

$$\mathbf{R_X(S) = E[\, X \mid S]} \qquad \mathbf{(80)}$$

O conceito de "revelação" aqui usado tem similaridades e diferenças com o famoso *princípio da revelação* ("revelation principle") da literatura de *assimetria de informação*, mais especificamente da *teoria de desenho de mecanismos*[151] (ex., Salanié, 1994) e de *jogos Bayesianos* (ex.: a vasta literatura de *jogos de sinalização*). Nessa tese, a revelação é do verdadeiro valor de um parâmetro técnico X (verdadeiro estado da natureza de X), enquanto que o conceito usado em desenho de mecanismos é relacionado ao verdadeiro tipo de um agente.

É oportuno formalizar matematicamente o conceito de expectativa condicional de uma variável X dado o sinal S. Para efeito de generalidade, essa formalização usará um pouco de *teoria da medida* e da visão axiomática da probabilidade de Kolmogorov, mas será dada também a interpretação intuitiva e informal. Kolmogorov (1933) *primeiro* definiu expectativa condicional e *depois*

[150] O autor usa esse termo desde Dias (1997) e foi incentivado por uma conversa com o Prof. Dixit em Princeton (em março de 1997) sobre "revelation". O autor tem usado esse termo na Petrobras desde 1998 com o sentido que apareceu em Dias (2001b). Entretanto, somente em Dias (2002) é que foi apresentado esse conceito de forma mais formal, com proposições sobre as propriedades da distribuição de revelações no contexto de OR e VOI.

[151] Usa-se o princípio de revelação para desenhar um mecanismo direto (exs.: contratos, regras de um leilão) de forma a ser ótimo para um jogador revelar toda a verdade sobre o seu tipo.

(usando essa definição) definiu probabilidade condicional como caso particular. A abordagem axiomática permite trabalhar com eventos condicionantes com probabilidade muito pequena ou zero (ver Rao, 1993, p.25-26), o que não é possível com a definição tradicional de probabilidade condicional (teria divisão por zero). O livro clássico de Kolmogorov (1933), trazendo os axiomas fundamentais da teoria da probabilidade, causou uma revolução nessa teoria e é hoje a abordagem padrão de textos avançados de probabilidade[152].

Sejam as variáveis aleatórias X e S definidas no mesmo *espaço de probabilidades* $(\Omega, \Sigma, \mathbb{P})$, onde Ω é o *espaço amostral* (conjunto de todas as possíveis realizações ω), Σ é uma *sigma-álgebra*[153] e $\mathbb{P}$ é uma *medida de probabilidade* definida[154] no intervalo [0, 1]. Na teoria da medida, uma v.a. é vista como uma função de um espaço amostral $\Omega \rightarrow \mathbb{R}$ (conjunto dos números reais). Condicionando uma probabilidade a um conjunto A, significa uma *contração* do espaço amostral. Em vez de tomar o valor esperado sobre todo o espaço Ω, o valor esperado é tomado apenas sobre um subconjunto $A \subseteq \Omega$. Algumas definições são possíveis dependendo se o condicionante é um evento (conjunto) sobre uma v.a. discreta, ou uma v.a. contínua, ou mesmo uma v.a. arbitrária (discreta + contínua, ex.: distribuição do valor da opção mostrada na Figura 4). Será apresentado abaixo o caso mais geral. Seja X uma v.a. *integrável*[155] mapeando o espaço de probabilidade $(\Omega, \Sigma, \mathbb{P})$ em um espaço mensurável. Seja Ψ uma *sub-sigma-álgebra* de Σ (i. é, $\Psi \subseteq \Sigma$)[156]. A expectativa condicional de X dado Ψ, $\mathbf{E[X \mid \Psi]}$ é uma função Ψ-mensurável[157] que satisfaz a equação abaixo (chamada de "partial averaging" por autores como Shreve, 2004, p.68) para todo conjunto $Y \in \Psi$:

---

[152] Confome Spanos (1999, p.412), a abordagem de Kolmogorov se tornou um sucesso instantâneo por clarear toda a "bagunça" criada com o estudo do movimento Browniano.

[153] Sigma-álgebra Σ em Ω é uma família de eventos E (sub-conjuntos de Ω que são eventos), compreendendo o conjunto vazio $\varnothing$, complementos de conjuntos que pertencem a Σ e uniões contáveis de seqüências de conjuntos $E_n \in \Sigma$.

[154] Mais precisamente, P é uma função mapeando Σ em [0, 1].

[155] Se X é integrável, então X tem valor esperado finito, ou X é integrável $\Leftrightarrow E[|X|] < \infty$ (James, 1996, p.110). Uma função f que é μ-integrável no sentido de Lebesgue, é muitas vezes escrita como $f \in \boldsymbol{L}^1(\Omega, \Sigma, \mu)$ e assim $X \in \boldsymbol{L}^1$, onde $\boldsymbol{L}^1$ é simplesmente um vetor sobre $\mathbb{R}$ (reais).

[156] Ver Shyraiev (1996, p.212) ou Williams (1991, p.84). Uma sub-sigma-álgebra representa simplesmente "*menos (ou igual) informação*" ou um sub-conjunto de informação.

[157] Se X é Ψ-mensurável, então a informação em Ψ determina totalmente o valor de X. (Shreve, 2004, p.66). Ver, Williams (1991, p.29-30) para a definição mais formal de funções Ψ-mensuráveis em um espaço mensurável (Ω, Ψ).

$$\int_Y \mathbf{E[X \mid \Psi](\omega)}\, d\mathbb{P}(\omega) = \int_Y \mathbf{X(\omega)}\, d\mathbb{P}(\omega) \qquad \textbf{(81)}$$

Onde as integrais acima são as de Lebesgue-Stieltjes. Em finanças, primeiro se estabelece o espaço amostral Ω, que é o conjunto de possíveis cenários futuros, sobre os quais se impõe uma medida ("real") de probabilidade $\mathbb{P}$. No entanto, no cálculo de opções se usa muito uma medida de probabilidade equivalente chamada de medida *neutra ao risco* $\tilde{\mathbb{P}}$. Duas medidas de probabilidade são ditas equivalentes no espaço mensurável (Ω, Σ) se elas concordam em quais conjuntos em Σ que têm probabilidade zero (Shreve, 2004, p.34-35). Nos conjuntos de probabilidade positiva, essas medidas irão discordar apenas nos valores de quão prováveis são cada cenário. No caso de incerteza técnica, as probabilidades já são naturalmente neutras ao risco (ver item 3.1.1) e assim essas duas medidas se confundem. Shreve (2004, p.35) não gosta das denominações "mundo real" e "mundo neutro ao risco", muito mencionadas na literatura de finanças, já que o "mundo" é o mesmo e representado por Ω.

Denota-se por σ(S) a menor sigma-álgebra gerada pela v.a. S com respeito ao qual S é mensurável (Karlin & Taylor, 1975, p.300). Nesse caso tem-se E[X | S] = E[X | σ(S)] (Mikosch, 1998, p.69). A sigma-álgebra σ(S) gerada por S contém toda a informação essencial sobre a estrutura da variável (ou vetor) aleatória S como função de $\omega \in \Omega$. Nas aplicações se trabalhará apenas com valores pertencentes ao conjunto dos números reais $\mathbb{R}$, i. é, com a sigma-álgebra Ψ com todos os conjuntos $\{S \in \mathcal{B}\}$, onde $\mathcal{B}$ é a sigma-álgebra de Borel[158].

A expectativa condicional R_X pode ser vista de forma *alternativa* como a projeção de X no *espaço L^2 de Hilbert*[159] sobre o subespaço linear fechado $\boldsymbol{L}^2(\Omega, \Psi, \mathbb{P})$ do espaço $\boldsymbol{L}^2(\Omega, \Sigma, \mathbb{P})$ e assim a expectativa condicional existe (ver

[158] A sigma-álgebra de Borel contém sub-conjuntos gerais da reta $\mathbb{R}$, i. é, se $\Omega = \mathbb{R}$, então $\mathcal{B}$ é gerada pelos sub-conjuntos $C^1 = \{(a, b] : -\infty < a < b < \infty\}$. Já se **S** for um vetor e $\Omega = \mathbb{R}^{\mathbf{n}}$, usa-se os sub-conjuntos chamados de retângulos (**a**, **b**], onde **a** e **b** são vetores. Qualquer sub-conjunto *razoável* de $\mathbb{R}^{\mathbf{n}}$ (bolas, esferas, curvas suaves, superfícies, conjuntos abertos, fechados, etc.) é um *conjunto de Borel* (Mikosch, 1998, p.64-65) e pertence a $\mathcal{B}$.

[159] Espaço de Hilbert $\boldsymbol{L}^2$ é o espaço de variáveis aleatórias com finitos segundos momentos, o qual é gerado por combinações *lineares* de funções integráveis ao quadrado de $\Omega \rightarrow \mathbb{R}$. Uma variável aleatória X é dita pertencer a $\boldsymbol{L}^2$ se $E[X^2] < \infty$ (ver, por ex., Williams, 2001, p.65).

Jacod & Protter, 2000, p.196). A Figura 33 ilustra essa visão de E[X | Ψ], como a projeção ortogonal de X no espaço $\boldsymbol{L}^2(\Omega, \Psi, \mathbb{P})$ de Hilbert[160].

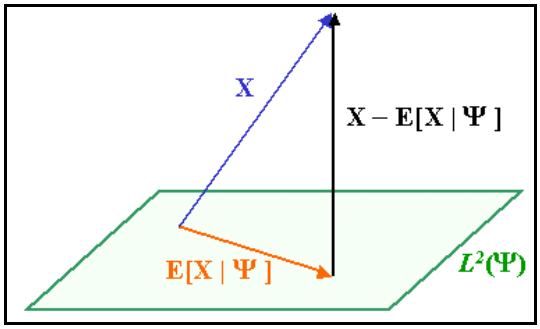

Figura 33 - E[X | Ψ] como Projeção no Espaço L^2 de Hilbert

Existem outras maneiras de definir a expectativa condicional, por ex., através da *função indicador*[161] $\mathbb{I}_S$ (ver Mikosch, 1998, p.58; ou Shiryaev, 1996, p.79). O lema a seguir (baseado em Mikosch, 1998, p.194) seguirá a ótica de Kolmogorov (1933, p.53) no sentido que a existência de E[X] é também suficiente para a existência de R_X.

Lema 1 (existência da expectativa condicional R_X): Seja o espaço de probabilidade $(\Omega, \Sigma, \mathbb{P})$ onde X e S são definidos e Ψ uma sub-sigma-álgebra de Σ. Se $E[|X|] < \infty$, então existe uma v.a. R_X definida pela eq. 80 tal que a sigma-álgebra gerada por R_X está contida em Ψ e, para todo $S \in \Psi$, essa variável atende a definição de expectativa condicional dada pela eq. 81. Além disso, essa expectativa condicional é *quase certamente*[162] finita e é única a menos de *versões* de R_X que diferem de R_X apenas em conjuntos de medida igual a zero.

Prova: Conforme Glivenko (ver Kolmogorov, 1933, p.40), $E[\,|X|\,] < \infty$ é uma condição necessária e suficiente para a existência de E[X]. A prova da existência da expectativa condicional, i. é, $E[\,|X|\,] < \infty \Rightarrow E[\,|R_X|\,] < \infty$, é baseada no famoso teorema de Radon-Nikodým[163], e é mostrada em Kolmogorov (1933, p. 53). Se a eq. 81 vale tanto para R_X como para R'_X e ambas geram sigma-álgebras

[160] Ver figura similar em Mikosch (1998, p.75) e mais detalhada em Saporta (1990, p.78).

[161] Função indicador $\mathbb{I}_S(\omega)$ é igual a 1 se $\omega \in S$ e igual a zero se $\omega \notin S$.

[162] Uma assertiva é válida quase certamente (q.c.) se ela é sempre válida exceto num conjunto de medida igual a zero. Ex.: é *possível* sortear o número exato 1 do conjunto de números reais, mas como existem infinitos números reais, Pr(X = 1 exato) = 0.

[163] Esse teorema (na época recente, de 1930) permite a *diferenciação de funções-conjuntos* ("set functions"). Esse teorema não será demonstrado já que mesmo em livros avançados de finanças, como em Shreve (2004, p.39), a prova é considerada "além do escopo desse texto".

contidas em Ψ então, pelo teorema de Radon-Nikodým, R_X é única no sentido que a probabilidade de $R_X \neq R'_X$ é zero (ver também Shreve, 2004, p.69). □

Serão listadas sem provas[164] no Lema 2, as 5 propriedades fundamentais das expectativas condicionais (Shreve, 2004), todas válidas quase certamente.

Lema 2 (propriedades fundamentais das expectativas condicionais): Seja o espaço de probabilidade $(\Omega, \Sigma, \mathbb{P})$ e seja Ψ uma sub-sigma-álgebra de Σ.

(a) Se X e Y são variáveis aleatórias integráveis e c_1 e c_2 são constantes, então é válida a propriedade de linearidade das expectativas condicionais:

$$\mathbf{E[c_1 X + c_2 Y \mid \Psi] = c_1 E[X \mid \Psi] + c_2 E[Y \mid \Psi] = c_1 R_X + c_2 R_Y} \quad \mathbf{(82)}$$

(b) Se X e Y são variáveis aleatórias integráveis, e Y é Ψ-mensurável, então é válida a propriedade "tirando para fora o que é conhecido":

$$\mathbf{E[Y X \mid \Psi] = Y\ E[X \mid \Psi]} \quad \mathbf{(83)}$$

(c) Se X é uma variável aleatória integrável e $\mathcal{H}$ é uma sub-sigma-álgebra da sub-sigma-álgebra Ψ, então é válida a propriedade de condicionamento iterado:

$$\mathbf{E[E[X \mid \Psi] \mid \mathcal{H}] = E[X \mid \mathcal{H}]} \quad \mathbf{(84)}$$

Essa propriedade será demonstrada no item 3.2.2 por ser fundamental para a teoria dessa tese. Ela pode ser escrita com a seguinte notação mais compacta:

$$\mathbf{E[R_X(\Psi) \mid \mathcal{H}] = R_X(\mathcal{H})} \quad \mathbf{(85)}$$

(d) Se X é uma variável aleatória integrável e independente de Ψ, então é válida a propriedade de independência:

$$\mathbf{E[X \mid \Psi] = E[X]} \quad \mathbf{(86)}$$

(e) Se X é uma variável aleatória integrável e φ(x) é uma função *convexa* $\mathbb{R} \rightarrow \mathbb{R}$ tal que $E[|\varphi(x)|] < \infty$ e $x \in X$ (x é uma variável "dummy"), então vale a propriedade da desigualdade condicional de Jensen:

$$\mathbf{E[\varphi(X) \mid \Psi] \geq \varphi(E[X \mid \Psi])} \quad \mathbf{(87)}$$

Em adição a essas propriedades fundamentais, outras serão vistas no item 3.2.2. Mas é útil mencionar antes mais duas propriedades e uma definição. A primeira propriedade (ver Lawrence, 1999, p.63) diz que a ordem de integração não importa quando se toma o valor esperado de uma função real φ(x, s):

$$\mathbf{E_{x,s}[\varphi(x,s)] = E_x[E_{s|x}[\varphi(x,s)]] = E_s[E_{x|s}[\varphi(x,s)]]} \quad \mathbf{(88)}$$

[164] Para as provas, ver Shreve, (2004, pp.69-72) e/ou Williams (1991, pp.88-90).

Onde os subscritos de E denotam as densidades de probabilidade de onde os valores esperados são tomados (distribuições conjuntas ou distribuições condicionais ou distribuições marginais, dependendo dos subscritos).

A segunda propriedade diz como a densidade marginal de S, p(s), e a densidade condicional de X dado S, p(x | s), determinam a densidade conjunta de X e S, p(x, s), conforme James (1996, p.185):

$$\mathbf{p(x, s) = p(s)\ p(x \mid s)} \qquad \mathbf{(89)}$$

A qual é bem similar à eq. 65, que dizia como calcular a distribuição condicional dada as distribuições conjunta e marginal. Aqui se diz que, se p(s) é zero, também será zero a distribuição conjunta p(x, s).

Definição. **Variância condicional a uma sub-sigma-álgebra Ψ de Σ**: Seja a v.a. X, definida no espaço (Ω, Σ, P) com expectativa condicional E[X | Ψ] em relação à sub-sigma-álgebra Ψ dos subconjuntos de Ω. A variância condicional de X é uma v.a. definida por[165]:

$$\mathbf{Var(X \mid \Psi) = E[(\ X - E[X \mid \Psi]\ )^2 \mid \Psi\ ]} \qquad \mathbf{(90)}$$

3.2.2. Processo de Revelação e Distribuição de Revelações

Após a introdução e fórmulas sobre expectativas condicionais do item 3.2.1 (que em parte serão aqui usadas), nesse item será estabelecido um teorema fundamental para o tratamento da incerteza técnica em problemas dinâmicos de opções reais, de forma que esse item será de grande importância prática, embora ainda com algumas formalizações e teoria. No item 3.2.3 será visto um exemplo simples para ilustrar o teorema, além de figuras para reforçar a intuição. Esse teorema será usado nas aplicações do cap. 5, assim como no estabelecimento tanto da medida de aprendizagem proposta (detalhada no item 3.3) como nos processos de revelação de Bernoulli que são aplicados em exemplos exploratórios (item 3.4).

De forma consistente com a estrutura de informação flexível definida no item 3.1.4.2 (ver eq. 63), no Teorema 1 abaixo a distribuição de revelações será definida exceto por um grau de liberdade (o que dá flexibilidade ao usuário do teorema). O Teorema 1 ajudará a usar a estrutura de informação flexível (eq. 63),

[165] Ver, por ex., Shyraiev (1996, p.214).

orientando na escolha do critério aberto •. Além disso, η^2, a medida de aprendizagem proposta, será diretamente relacionada com esse teorema.

Definição. **Processo de revelação**: é uma *seqüência* de variáveis aleatórias $\{R_{X,1}, R_{X,2}, R_{X,3}, \ldots\}$[166] geradas por uma seqüência de informações ou sinais S_1, S_2, S_3, … sobre uma variável de interesse X, que tem como principal característica a redução esperada de incerteza advinda da chegada de nova informação. Processo de revelação é um *processo de aprendizagem* probabilístico. Na literatura matemática, é às vezes chamado de "*acumulando dados sobre uma variável aleatória*" (Williams, 1991, p.96).

Processos de revelação podem ser vistos como processos estocásticos, mas geralmente os processos de revelação devem ser indexados por eventos e não pelo *tempo* como ocorre com a maioria dos processos estocásticos estudados. Essa tese está interessada especialmente nos processos com eventos sendo exercícios de opções de aprendizagem, a fim de modelar a evolução (redução esperada) da incerteza técnica com o processo de investimento em informação. Um exemplo de processo de revelação indexado por eventos é a perfuração seqüencial de poços de delimitação em uma reserva de petróleo com volume B incerto. Nesse caso, cada poço corresponde a um evento no processo de revelação de B e entre esses eventos podem decorrer meses ou até ano(s). No intervalo entre esses eventos não há porque alterar as expectativas sobre B, i. é, não há porque rever R_B.

Esses eventos geralmente são *endógenos* no modelo, são ativados pelo decisor (ou decisores em caso de interação estratégica), i. é, eventos são exercícios de opções. Na literatura tradicional de VOI, não foi dada atenção ou não foi considerada importante a distinção entre os processos indexados por eventos e indexados pelo tempo. Por ex., Lawrence (1999, p.156) argumenta que uma seqüência de eventos pode ser vista como um "*time-driven process with different chronological length*" porque os eventos são sucessivos ao longo do tempo. Mas no contexto de opções reais isso não funciona porque, nos modelos mais importantes e realistas, os eventos são opcionais, ativados pela firma e em paralelo com outro processo (incerteza de mercado representada pelos preços do

[166] Poderia se definir esse processo como uma seqüência de momentos das distribuições posteriores, com a distribuição de expectativas condicionais sendo um deles. Ou até pensar numa seqüência de distribuições de distribuições posteriores. Mas dificilmente essa teoria teria a utilidade prática e simplicidade que será obtida trabalhando focado em expectativas condicionais.

petróleo) no qual a informação chega continuamente ao longo do tempo e que não é opcional. A complexidade desses dois processos superpostos (processo de revelação e processo de mercado) requer uma metodologia específica. Uma questão de importância prática na superposição desses processos é o tempo de aprendizagem. É uma necessidade prática que o modelo permita comparar alternativas de investimento em informação com diferentes tempos de aprendizagem. O modelo proposto fará isso de uma forma simples (ver cap.5).

Em alguns outros casos os eventos podem ser modelados como processos exógenos, por ex., um processo de Poisson com freqüência λ. Numa bacia petrolífera com muitas companhias atuando, pode ser mais simples modelar a revelação de informação exploratória (perfuração de poços pioneiros) com um processo exógeno de Poisson do que um modelo endógeno de equilíbrio entre as firmas petrolíferas. O primeiro problema é que essa informação é mais relevante e *geral*[167] justamente no caso de bacias pouco exploradas, onde geralmente o número de firmas é pequeno. O segundo problema é que em bacias mais conhecidas (ex.: Golfo do México), a revelação de informação tem mais relevância *localmente* e assim novamente os decisores relevantes são duas ou poucas firmas. Com poucas firmas, é melhor usar o modelo endógeno para considerar a interação estratégica entre as firmas, ou seja, deve-se usar a *teoria dos jogos de opções* (que será vista no cap.4) para analisar o exercício estratégico de opções que gera um processo de revelação de informações.

Um exemplo de processo de revelação indexado pelo tempo é o que ocorre geralmente com o retorno de ações de *novas* empresas que são lançadas no mercado (IPO, *Initial Public Offering*). A volatilidade dessas ações é geralmente alta (causando muitas falências prematuras, vide o estouro da "bolha das ações da internet" nos anos 2000/2001). Com o tempo as empresas que sobreviverem se tornam mais conhecidas e assim sua volatilidade tende a cair até se estabilizar dinamicamente, indicando um certo *equilíbrio* numa situação próxima de um mercado com informação completa sobre o retorno dessa ação. Nesse caso se tem um *processo de revelação apenas num transiente temporal* em que houve uma

[167] A incerteza sobre a existência de *rochas geradoras* de petróleo numa bacia tem um caráter *geral* no sentido de ser relevante para blocos distantes de vários quilômetros em uma bacia. Numa bacia mais conhecida essa incerteza pode inexistir e assim a informação revelada por um poço pioneiro tem mais relevância *local*, i. é, revela mais sobre a *migração do petróleo* gerado e assim é importante para prospectos localizados em blocos vizinhos.

difusão de informação sobre a capacidade de essa ação gerar retorno e sobre o seu valor justo (ou de equilíbrio dinâmico no mercado). Após esse transiente, o processo de revelação deixa de existir e o mercado tem informação completa.

Exemplos de processo que não são de revelação são o de preços de ações seguindo um MGB (novas informações não reduzem a incerteza esperada) e o modelo de "ruído acumulativo" descrito em Childs & Ott & Riddiough (2001). Não há redução esperada de incerteza (volatilidade) nesses casos.

Antes de definir processo de revelação *convergente*, é oportuno definir o caso limite mais importante que é o de revelação total ou *aprendizagem total* sobre a variável com incerteza técnica X. Esse é o máximo benefício que se pode obter com um investimento em informação. Um exemplo de revelação total com um *único* investimento em informação é quando a variável de interesse X é a existência ou não de petróleo (v.a. de Bernoulli, dada pelo fator de chance) e o investimento em informação é a perfuração do poço pioneiro, que tirará toda a dúvida a respeito da existência de petróleo (mas existirá ainda incerteza sobre o volume B da reserva e sua qualidade). Mas na prática, em geral, se precisa de uma seqüência (muito grande) de investimento em informação para se atingir valores próximos desse limite de revelação total da variável com incerteza técnica X.

Definição. **Revelação total da variável X**: significa a revelação de um cenário c tal que $Pr(X = c) = 1$, onde c é uma constante pertencente ao suporte de p(x). Em termos gerais, se a informação disponível é dada pela sub-sigma-álgebra Ψ, revelação total de X significa que X é Ψ–mensurável e, logo, é válido escrever $E[X \mid \Psi] = X$ quase certamente[168]. Intuitivamente, significa que existe informação perfeita sobre o verdadeiro estado da natureza da variável X. Em termos matemáticos, "*uma variável aleatória é Ψ–mensurável se e somente se ela assume valores constantes sobre os átomos de Ψ*" (Shiryaev, 1996, p.80).

Será visto que todo processo de revelação converge[169] para uma variável aleatória integrável denotada por X_∞ quando $n \to \infty$. Mas nem sempre converge para uma revelação total de X, i. é, nem sempre $X_\infty = X$. Intuitivamente (serão vistos exemplos) pode ocorrer que a seqüência de informações $S_1, S_2, \ldots, S_n$

[168] Além disso, operações algébricas ordinárias tais como soma, multiplicação e divisão, não destroem a mensurabilidade (Gallant, 1997, p.47).

[169] Converge quase certamente (com probabilidade 1), o que implica que converge em probabilidade, o que também implica convergência em distribuição (Karlin & Taylor, 1974, p.18).

reduza a incerteza esperada cada vez mais com o progresso de n, mas essa redução pode ser cada vez menor de forma que a redução esperada de variância no limite seja, por ex., de 50%, i. é, a cada informação S_n, a redução esperada de variância se aproxima cada vez mais de uma redução de 50%, mas não de 100%. Serão vistos exemplos práticos em exploração de petróleo para ambos os casos (revelação parcial no limite e revelação total no limite $n \to \infty$) de forma mais detalhada no item 3.4. Matematicamente um processo de revelação é igual ao chamado *processo de Doob* (Karlin & Taylor, 1975, p.246 e 295).

Nem toda seqüência de v.a. converge no limite para uma variável integrável X_∞. Todo processo de revelação converge porque o processo $\{R_{X,1}, R_{X,2}, \ldots R_{X,n}\}$ é *uniformemente integrável* (definido a seguir). Será visto na Proposição 5 abaixo que, como o processo de revelação é uniformemente integrável, então ele converge para X_∞ (que sob certas condições pode ser X) quando $n \to \infty$. Como esse processo sempre converge para algum X_∞, seria redundante o qualificativo "convergente" para esse processo, de forma que se usará o adjetivo "convergente" apenas no caso limite de revelação total, i. é, quando $X_\infty = X$.

A explanação da definição abaixo e da Proposição 5 seguirá Brzezniak & Zastawniak (1999, cap. 4) e Karlin & Taylor (1975, p.295-297 e 309-312), mas de forma menos detalhada pode ser vista em Williams (1991, p.96 ["accumulating data about a random variable"] e o caso particular das p.166-167 ["noisy observation of a single random variable"]), ou em Ross (1996, p.297 e 318-319).

Definição. **Seqüência uniformemente integrável**: Uma seqüência de variáveis aleatórias $R_{X,1}, R_{X,2}, \ldots R_{X,n}$ é chamada de uniformemente integrável se para todo $\varepsilon > 0$ existe um $M > 0$ tal que para todo $n = 1, 2, \ldots$:

$$\int_{\{|R_{X,n}| > M\}} |\, R_{X,n} \,|\; dP < \varepsilon \qquad (91)$$

Integração uniforme é uma condição necessária para uma seqüência de variáveis aleatórias $\{R_{X,n}\}$ integráveis convergir em $\boldsymbol{L}^1$, i. é, convergir em média para uma v.a. integrável, que aqui seria $\lim_{n \to \infty} E[|R_{X,n} - X_\infty|\,] = 0$. Na eq. (91) a integral é a de Lebesgue-Stieltjes, ou seja, **P** é a medida de probabilidade ou mais simplesmente a função distribuição acumulada $\mathbf{P}(R_{X,n})$.

Proposição 5: Seja $\{R_{X,1}, R_{X,2}, \ldots R_{X,n}\}$ um processo de revelação, i. é, $R_{X,k} = E[X \mid \Im_k]$ são definidos no mesmo espaço de probabilidade (Ω, Σ, P),

sendo X integrável e onde $\Im_k$ é uma *filtração*[170] $\{\Im_k: k \geq 0\}$, com $\Im_k$ sendo gerada pela seqüência de sinais $\{S_k\}$. Então o processo de revelação é uniformemente integrável e assim quando $n \to \infty$, existe q.c. um limite de $R_{X,n}$ em $\boldsymbol{L}^1$ que é uma v.a. integrável, denotada por X_∞, que é também uma expectativa condicional, i. é:

$$\mathbf{lim_{n \to \infty} R_{X,n} = X_\infty = E[X \mid S_1, S_2, \ldots] = E[X \mid \Im_\infty]} \qquad \mathbf{(92)}$$

Prova: Primeiro tem de ser provado que *qualquer* processo de revelação é um martingale. Isso será mostrado no Teorema 1(d) abaixo. A prova de que integrabilidade uniforme é *suficiente* para a convergência de um martingale em $\boldsymbol{L}^1$ é dada pelo famoso *Teorema da Convergência de Martingale de Doob*[171] (ver, por ex., Brzezniak & Zastawniak, 1999, theorem 4.2, p.71-73). A prova de que o processo de revelação (um "martingale de Doob") é uniformemente integrável é dada por ex. em Ross (1996, p.319) ou Karlin & Taylor (1975, p.295-296) e então existe um limite dado por uma v.a. integrável X_∞. A prova que esse limite é uma expectativa condicional $E[X \mid \Im_\infty]$ é dada por Karlin & Taylor (1975, p.310)[172]. □

Na verdade essa proposição poderia ser ainda mais forte. Pode ser provado que *todo* martingale uniformemente integrável pode ser escrito como um processo de revelação (i. é, um processo de Doob), ver Karlin & Taylor (1975, p.311-312, em especial o Lemma 7.3) ou Brzezniak & Zastawniak (1999, theorem 4.4, p.77) ou até mesmo em Doob (1953, cap.VII). Conforme Karlin & Taylor (1975, p.247-248), os dois principais resultados da teoria dos martingales são o teorema da convergência de martingales e o *teorema da amostragem opcional* (ou *parada ótima*)[173]. Mas enquanto esse último encontra freqüentes aplicações em problemas de decisão opcional (inclusive problemas de OR), o teorema de convergência é usado mais de forma teórica[174] para determinar a distribuição assintótica de

---

[170] Pode-se interpretar a filtração $\Im_n$ gerada pela informação seqüencial $\{S_1, S_2, \ldots S_n\}$ como um conjunto contendo toda a informação disponível no estágio n. Em termos técnicos, é uma crescente família de sub-sigma-álgebras gerada pelas informações. Ex.: $R_{X,2} = E[X \mid S_1, S_2]$.

[171] Esse teorema tem sido mais usado para provar teoremas tais como a lei 0-1 de Komolgorov e do filtro de Kalman. A aplicação no contexto da tese não tem sido explorada.

[172] Uma prova simples é ver que $R_{X,n}$ é uma função de $S_1, S_2, \ldots S_n$, de forma que o limite X_∞ é uma função mensurável dessa seqüência de sinais e portanto mensurável com respeito a $\Im_\infty$.

[173] Esse teorema diz que, sob condições bem gerais, se $\{R_{X,n}\}$ é um martingale, então também será uma seqüência indexada por T_n em vez de n, onde $\{T_n\}$ é uma seqüência de *tempos de Markov* ou tempos de parada. Tempo de Markov tem a propriedade que o evento T = n é determinado apenas pela história $S_1, S_2, \ldots, S_n$ até o estágio n (Karlin & Taylor, 1975, p.247).

[174] Whittle (2000, p.299) afirma que o teorema da convergência de martingales não trás resultados muito interessantes em processos que são *absorvidos* (ex., exercício de uma opção) em algum estado e por isso o teorema de parada ótima (ou opcional) é mais significativo.

funções de processos estocásticos gerais. Aplicações do teorema de convergência em OR como a dessa tese (por ex., no item 3.4), não têm sido vistas na literatura.

Definição. **Processo de revelação convergente**: é um *processo de aprendizagem* que no limite converge ao verdadeiro valor da variável (ou parâmetro). Formalmente, se $n \to \infty$ então $Var[X \mid S_1, S_2, \ldots, S_n] \to 0$. Em termos da Proposição 5, é quando $X_\infty = X$. A Proposição 5 diz que isso é possível, mas não diz sob que condições irá ocorrer $X_\infty = X$ (será visto depois).

Um exemplo de processo de revelação convergente é o da perfuração de poços de delimitação para reduzir a incerteza do volume da reserva B. É convergente, pois se perfurar um número n muito grande (infinito) de poços, se conhecerá totalmente o volume de reservas B, i. é, se $n \to \infty$ então $Var[B_n] \to 0$. Já o mencionado caso do IPO, o processo de revelação é não convergente pois mesmo que se passe um período muito grande de tempo, a volatilidade do retorno não converge a zero. Outro exemplo de processo de revelação não convergente é a venda de um novo produto no mercado, em que existe incerteza sobre a aceitação do produto, i. é, sobre a função demanda. Com testes de mercado e ao longo da produção e venda do produto, a incerteza diminui mas nunca vai a zero pois existe a incerteza sobre o nível de crescimento da economia de um país, região, etc., que influi na demanda de produtos em geral da economia.

No Teorema 1 será mostrada a propriedade fundamental de martingale dos processos de revelação relacionada à distribuição de revelações. Além disso, será rediscutida a questão de processos de revelação no item 3.4, especificamente para o caso de fator de chance exploratório modelado com o processo de revelação de Bernoulli. O fator de chance exploratório pode ser modelado tanto com processos de revelação (totalmente) convergentes como não (totalmente) convergentes.

O Teorema 1 a seguir descreve as 4 principais propriedades da distribuição de revelações (distribuição da variável R_X). As 4 propriedades são: a média de R_X, a variância de R_X, R_X no caso limite de revelação total e a propriedade de martingale dos processos de revelação (seqüências de R_X). Ele será formulado, provado e depois será discutida a intuição por trás e suas conseqüências.

Teorema 1 (Distribuição de Revelações): Sejam as variáveis aleatórias X e S com média e variância finitas[175] definidas no espaço de probabilidade $(\Omega, \Sigma, \mathbb{P})$. X é a variável de interesse e tem probabilidade a priori p(x). S é chamado de sinal ou nova informação e gera a sigma-álgebra Ψ, onde Ψ é uma sub-sigma-álgebra de Σ, i.é, $\Psi \subseteq \Sigma$. Seja $p(R_X)$ a distribuição[176] de probabilidades da variável $R_X = E[X \mid S]$, denominada *distribuição de revelações* de X. Então, a distribuição de revelações tem as seguintes propriedades que *quase*[177] a definem:

(a) No caso limite de revelação total da variável X, a variância de qualquer distribuição posterior é zero e a distribuição de revelações $p(R_X)$ é igual à distribuição a priori p(x).

(b) A média da distribuição de revelações é igual à média de X, i. é:

$$\mathbf{E[R_X] = E[X]} \qquad \mathbf{(93)}$$

(c) A variância da distribuição de revelações é dada simplesmente pela redução esperada da variância de X induzida pelo sinal S, i. é:

$$\mathbf{Var[R_X] = Var[X] - E[\,Var[X \mid S\,]\,]} \qquad \mathbf{(94)}$$

(d) Seja um processo seqüencial de informação $S_1, S_2, S_3, \ldots$ e as variáveis aleatórias $R_{X,n} = E[X \mid S_1, S_2, \ldots S_n]$, n = 1, 2, … Então, o *processo de revelação* $\{R_{X,1}, R_{X,2}, R_{X,3}, \ldots\}$ é um martingale.

Prova:

(a) A Proposição 5 acima garante que *um* limite *sempre* existe com probabilidade 1. Além disso, sabe-se que **Prob(X = c) = 1 ⇔ Var[X] = 0** (prova: DeGroot & Schervish, 2002, theorem 4.3.1, p.198). Então, para cada possível cenário c do suporte da distribuição a priori que possa ser revelado, estará associada uma distribuição posterior com variância zero. O restante da prova decorre diretamente da definição de distribuição a priori. Para ver isso, como a distribuição de revelações é a distribuição das médias das distribuições posteriores

[175] Um matemático poderia preferir dizer apenas que X e S são duplamente integráveis, i. é, $E[\,|X^2|\,] < \infty$ e $E[\,|S^2|\,] < \infty$, ou de forma equivalente que X e S $\in \boldsymbol{L}^2$, pois isso automaticamente implica que X e S $\in \boldsymbol{L}^1$ (Williams, 2001, p.65). Mas aqui isso não é restritivo pois as aplicações práticas do teorema são focadas apenas em problemas em que as médias e variâncias são finitas.

[176] Quando se menciona apenas "distribuição" e/ou se usa letra minúscula, significa a função densidade de probabilidades (caso contínuo), etc. Para a função distribuição acumulada, se usará letra maiúscula e em geral o texto será mais explícito.

[177] Definição: distribuição quase-definida é aquela em que se conhece pelo menos a *média*, a *variância* e que pertence a um processo seqüencial de distribuições com *origem* definida e *convergente* para uma distribuição totalmente definida. Então o Teorema 1 quase-define $R_X(S)$. Uma distribuição *totalmente* definida é simulável e tem função distribuição acumulada conhecida.

e essas médias colapsam para os cenários c (pois E[c | S] = c, se c é constante) do suporte da distribuição a priori, então o suporte da distribuição de revelações é simplesmente o mesmo suporte da distribuição a priori. Que as probabilidades de ocorrência (massa ou densidade) desses cenários são iguais às descritas pela distribuição a priori decorre diretamente da definição de distribuição a priori (ver item 3.1.2). Assim, em caso de revelação total, a distribuição de revelações é igual à distribuição a priori. □

(b) Essa propriedade é conhecida na literatura por *lei das expectativas iteradas*. Esse item pode ser formulado de forma mais geral, com a sub-sigma-álgebra Ψ (em vez da v.a. S) e da seguinte forma: se R_X é qualquer *versão*[178] de E[X | Ψ] então $E[R_X] = E[X]$. Para efeito intuitivo, serão mostradas as provas sem conceitos de teoria da medida para os casos particulares de v.a. discretas e contínuas. A prova mais geral é dada, por ex., em Williams (1991, p.89). Primeiro o caso em que ambas, X e S, são variáveis aleatórias discretas (a prova segue Ross, 1998, p.338). Quer-se provar que:

$$\sum_s R_X(s)\ \Pr(S = s) = E[X]$$

Pois o lado esquerdo é $E[R_X]$ por definição. Pela definição de R_X o lado esquerdo pode ser escrito como:

$$E[R_X] = \sum_s R_X(s)\ \Pr(S = s) = \sum_s \sum_x x\ \Pr(X = x \mid S = s)\ \Pr(S = s) =$$

$$E[R_X] = \sum_s \sum_x x\ \frac{\Pr(X = x,\ S = s)}{\Pr(S = s)}\ \Pr(S = s) = \sum_s \sum_x x\ \Pr(X = x,\ S = s)$$

$$\Rightarrow E[R_X] = \sum_x x \sum_s \Pr(X = x, S = s) = \sum_x x\ \Pr(X = x) = E[X] \qquad \square$$

Agora a prova para o caso de X e S como variáveis aleatórias contínuas (segue James, 1996, p.176), com densidade de probabilidade conjunta p(x, s), densidade condicional p(x | s) = p(x, s)/f(s), e densidade de S igual a p(s) > 0.

$$R_X(s) = \int x\, dF_X(x \mid S = s) = \int_{-\infty}^{+\infty} x\ p(x \mid s)\, dx = \int_{-\infty}^{+\infty} x\ \frac{p(x, s)}{p(s)}\, dx$$, como X é integrável:

[178] Se R^*_X é uma versão de R_X, então $R^*_X = R_X$ quase certamente (Williams, 1991, p.84).

$$E[R_X] = \int R_X(s)\, dF_S(s) = \int_{-\infty}^{+\infty} \left(\int_{-\infty}^{+\infty} x \, \frac{p(x, s)}{p(s)} \, dx \right) f(s)\, ds = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x\, p(x, s)\, dx\, ds$$

$$E[R_X] = \int_{-\infty}^{+\infty} x \left(\int_{-\infty}^{+\infty} p(x, s)\, ds \right) dx = \int_{-\infty}^{+\infty} x\, p(x)\, dx = E[X] \qquad \square$$

(c) Considere a definição de variância condicional dada pela eq. (90) mas para o caso mais intuitivo de S no lugar da sub-sigma-álgebra Ψ como condicionante[179]. Uma propriedade bastante conhecida da variância é:

$Var[Y] = E[Y^2] - (E[Y])^2$. Logo, como X é duplamente integrável, tem-se:

$Var[R_X] = Var[E[X \mid S]] = E[(E[X \mid S])^2] - (E[E[X \mid S]])^2 \Rightarrow$

$$\mathbf{Var[R_X] = E[(E[X \mid S])^2] - (E[X])^2} \qquad \mathbf{(95)}$$

Usando a mesma equação bastante conhecida para $Var[X \mid S]$, se obtém:

$Var[X \mid S] = E[X^2 \mid S] - (E[X \mid S])^2$. Tomando o valor esperado, tem-se:

$E[Var[X \mid S]] = E[E[X^2 \mid S]] - E[(E[X \mid S])^2] \Rightarrow$

$$\mathbf{E[Var[X \mid S]] = E[X^2] - E[(E[X \mid S])^2]} \qquad \mathbf{(96)}$$

Somando as eqs. (95) e (96) e rearranjando, se completa a prova:

$Var[E[X \mid S]] = Var[R_X] = Var[X] - E[Var[X \mid S]]$ $\square$

(d) De forma simples, martingale[180] significa preservar a mesma média e assim se pode mostrar o item (d) simplesmente aplicando de forma reiterada o resultado do item (b) para a seqüência $\{R_{X,1}, R_{X,2}, R_{X,3}, \ldots\}$. Essa seqüência é conhecida como sendo um processo ou martingale de Doob. A aplicação reiterada da eq. (93) é conhecida por *propriedade de torre* que em termos simples pode ser vista como:

$$\mathbf{E[X \mid informação] = E[\, E[X \mid \textit{mais} \text{ informação}] \mid informação]} \qquad \mathbf{(97)}$$

Mas uma prova mais rigorosa com a teoria da medida é útil no item (d). Conforme Williams (1991, p.94) um martingale tem de preencher três condições. A primeira condição, $R_X \in \boldsymbol{L}^1$, é uma conseqüência da premissa que o parâmetro com incerteza técnica X é integrável e assim, como foi visto, pelo teorema de

[179] A idéia de usar essa propriedade chave ocorreu ao resolver o problema 2 de Shyraiev (1996, p.83). A prova é similar à apresentada em Ross (1998, p.348). Para uma prova com condicionante sendo uma sub-sigma-álgebra, ver, por ex., Fristedt & Gray (1997, p.454).

[180] Pode-se interpretar martingale como um jogo "justo". Seja K_n o capital dum jogador após a aposta n e se todas as apostas são justas no sentido que o ganho esperado é zero, então para $n \geq 0$, K_n são martingales. Já aqui, K_n seria uma expectativa condicional de um parâmetro técnico X depois de n investimentos seqüenciais em informação.

Radon-Nikodým isso implica que $E[\,|R_X|\,] < \infty$ quase certamente (existência do valor esperado da distribuição de revelações mencionada antes)[181].

A segunda condição é que o processo de revelação seja *adaptado* à filtração $\{\Im_n: n \geq 0\}$, i. é, $R_{X,n}$ deve ser $\Im_n$–mensurável a uma crescente família de sub-sigma-álgebras de Ψ: $(\Im_0 \subseteq \Im_1 \subseteq \Im_2 \subseteq \ldots \subseteq \Im_n = \Psi)$. A expressão[182] $\Im_{n-1} \subseteq \Im_n$ significa um aumento de informações com o progresso de n (ex.: uma seqüência de n investimentos em informação, a perfuração seqüencial de n poços de delimitação, etc.). Isso decorre da definição de expectativa condicional, pois conforme a eq. (80), $R_{X,n} = E[X \mid \Im_n]$ que é $\Im_n$-mensurável por definição (ou seja, é conhecido ao observador no estágio n, Williams, 2001, p.406), para todo n.

A terceira condição (a mais importante e específica), diz que para $\{R_{X,n}\}$ ser martingale deve-se ter $E[R_{X,n} \mid \Im_{n-1}] = R_{X,n-1}$ quase certamente (q.c). Ou seja, as médias dessas seqüências $R_{X,n}$ devem ser iguais. Para provar isso, seja a versão mais geral da eq. 91 que foi apresentada antes com as eqs. 84 e 85, chamada (ver Williams, 1991, p.88; ou Williams, 2001, p.405) *propriedade geral de torre* ("general tower property"). Se Υ é uma sub-sigma-álgebra de Ψ, então q.c. se tem:

$$E[R_X \mid \Upsilon\,] \,(= E[E[X \mid \Psi] \mid \Upsilon\,]) = E[X \mid \Upsilon\,]$$

Essa propriedade é imediata da definição de expectativa condicional (Williams, 1991, p.90)[183] ou uma aplicação do item (b) desse teorema para o caso dos condicionantes serem sub-sigma-álgebras. Para ver isso de forma mais clara, seja o exemplo de martingale chamado "acumulando dados sobre uma variável aleatória" (Williams, 1991, p.96), seja a variável $\xi \in L^1(\Omega, \Psi, P)$ e defina $R_n = E[\xi \mid \Im_n]$. Então, pela propriedade de torre, se tem, q.c.:

$$E[R_n \mid \Im_{n-1}] = E[E[\xi \mid \Im_n] \mid \Im_{n-1}] = E[\xi \mid \Im_{n-1}] = R_{n-1}$$

E assim $R_{X,n}$ são martingales. □

[181] Uma prova simples: $R_{X,n} = E[X \mid S_0, S_1, \ldots S_n] \Rightarrow E[\,|R_{X,n}|\,] = E[\,|E[X \mid S_0, S_1, \ldots S_n]|\,] \leq E[\,E[\,|X|\,] \mid S_0, S_1, \ldots S_n]\,] = E[\,|X|\,] < \infty$. A primeira desigualdade é conseqüência de $|E[X]| \leq E[|X|]$ (ver Williams, 2001, p.61 para a prova), i. é, o módulo de uma integral ≤ integral do módulo.

[182] Em muitos textos se usa $\subset$ em vez de $\subseteq$. Aqui segue a notação $\subseteq$ de Shiryaev (1996). A sigma-álgebra trivial $\Im_0 = \{\varnothing, \Omega\}$ é o caso limite de não carregar nenhuma informação.

[183] Brzezniak & Zastawniak (1999, p.30) apresenta esse resultado "imediato" em 4 linhas.

3.2.3.
Discussão do Teorema e Exemplo Ilustrativo em Petróleo

Agora é conveniente uma discussão dos 4 itens do Teorema 1 sob um ponto de vista mais prático. Em muitos problemas de OR, a importância da distribuição de revelações é que através de uma simulação de Monte Carlo ela pode ser facilmente combinada com distribuições neutras ao risco de variáveis com incerteza de mercado, geradas por *processos estocásticos neutros ao risco* dessas variáveis ou em qualquer metodologia de finanças quantitativas que trabalhe com martingales. Os valores resultantes da combinação desses cenários podem ser descontados pela taxa livre de risco e assim pode ser calculado o valor da OR, etc. Por isso é da maior relevância prática saber as propriedades básicas da distribuição de revelações.

O item (a) do Teorema 1 é relacionado a um caso limite fundamental. O conceito de *revelação total* ("full revelation") tem um papel chave tanto na teoria – como limite de um processo de aprendizagem, como na prática – é muito mais simples otimizar quando não existe incerteza. Nesse caso limite, a distribuição de revelações está totalmente definida – é igual à distribuição a priori.

É oportuno fazer uma conexão entre o conceito de *revelação total* e (não) *arbitragem*: em ambos os casos há uma *eliminação de incerteza* quando uma v.a. é relacionada com outra v.a. (no caso, X e S). Mas no caso de arbitragem se monta um portfólio livre de risco $\mathbf{\Pi = a\,X + b\,S}$ (a e b são constantes reais)[184] – ou seja uma *relação aritmética* de variáveis aleatórias, enquanto que no conceito de revelação total, a eliminação de incerteza é feita através de uma *relação de condicionamento* de variáveis aleatórias, na qual se obtém o valor verdadeiro (determinístico) de X. O conceito de revelação total no entanto é *estável* no tempo enquanto que a arbitragem é *instável*, pois as constantes a e/ou b da equação $\mathbf{\Pi = a\,X + b\,S}$ variam com o tempo para o portfólio permanecer livre de risco (o "hedge" é dinâmico).

O item (b) do Teorema 1 pode ser interpretado intuitivamente como: a média ponderada do valor esperado condicional de X dado $S = s_i$ sendo cada termo $R_X(i) = E[X \mid S = s_i]$ (calculado pelas eqs. 78 e 79, para os casos de v.a.

[184] O portfólio livre de risco deve ter retorno igual à taxa livre de risco, caso contrário dá oportunidades de ganhos por arbitragem. Ver, por ex., Dixit & Pindyck (1994).

discretas e contínuas, respectivamente) ponderado pela probabilidade de ocorrência de cada evento s_i sobre o qual X é condicionado, é simplesmente igual ao valor esperado original (incondicional) de X, i. é, o valor esperado da distribuição a priori. Ou seja, a média das médias é igual à média original. Essa lei das expectativas iteradas foi usada num exemplo do cap. 1.

O item (c) do Teorema 1, que dá a variância da distribuição de revelações, é um resultado que não é óbvio, mas é uma propriedade extraordinária que faz a distribuição de revelações muito útil para propósitos práticos, como será visto a seguir. Note que o lado direito da eq. (94) é simplesmente a redução esperada de variância devido ao investimento em informação, i. é, é a diferença entre a variância a priori (incondicional ou antes da informação) e a variância residual *esperada* depois da informação. Em outras palavras, é a variância da distribuição a priori menos a variância média do conjunto de possíveis distribuições posteriores[185].

Note na eq. (94) que, se $Var[R_X]$ for normalizada (dividida) por Var[X], então o lado direito da eq. (94) é uma redução esperada percentual de variância, que é a medida de aprendizagem η^2 mencionada antes, que pode ser chamada também de poder de revelação de uma (alternativa de investimento em) informação. Isso significa que, conhecendo apenas a variância a priori (incerteza original, Var[X]) e essa medida de aprendizagem, se obtém a variância da distribuição de revelações simplesmente multiplicando Var[X] por η^2. Como a média da distribuição de revelações é obtida ainda mais facilmente (pelo item (b) do teorema diz que é igual a E[X]), então para obter a média e a variância da distribuição de revelações só precisa conhecer a (média e variância da) distribuição a priori e a medida de aprendizagem η^2. Como o item (a) dá o tipo de distribuição (igual à distribuição a priori) no limite de revelação total (informação perfeita), então o problema de VOI é totalmente definido para o caso de informação perfeita e *quase* definido para o caso de informação imperfeita.

No caso de informação imperfeita será necessário mais um dado de entrada (ex.: tipo de distribuição), representado por • na estrutura de informação flexível, eq. (63). Isso será discutido ainda nesse item.

---

[185] Como discutido antes, em geral "*existe uma infinidade de tais distribuições condicionais*" Goldberger (1991, p.40), uma distribuição posterior para cada possível resultado do investimento em informação (ou seja, para cada cenário $S = s_i$).

O Teorema 1 permite uma maneira prática de perguntar ao especialista técnico as informações necessárias para modelar a incerteza técnica com o modelo proposto. Bastam dois dados de entrada a serem perguntados (além de assumir algum critério • no caso geral de informação imperfeita):

- *Incerteza inicial* (distribuição a priori): Qual a incerteza total de um parâmetro particular de interesse (ex., o volume da reserva B)? O especialista precisa especificar essa distribuição a priori, i. é, a média, variância e a classe (ou tipo) de distribuição (Triangular, LogNormal, Beta, etc.).
- *Poder de revelação* (η^2): Qual a percentagem esperada de redução de incerteza técnica (leia-se *redução de variância*) com uma *alternativa específica* de investimento em informação adicional?

Com essas duas respostas dos especialistas, se especificam a média e a variância da distribuição de revelações, sendo que a variância será diferente para cada alternativa de investimento em informação. Quanto maior o poder de revelação de uma alternativa, maior a variância da distribuição de revelações. Essas distribuições serão usadas no modelo integrado de VOI dinâmico a ser mostrado com mais detalhe no cap. 5.

Note também a consistência dos itens (a) e (c) do Teorema 1. Na eq. (94), em caso de revelação total, a variância de todas as distribuições posteriores vai a zero e logo E[Var[X | S]] = 0. Assim, Var[R_X] = Var[X] em caso de revelação total, o que é consistente com o item (a).

É oportuno notar a semelhança conceitual da variância da distribuição de revelações, eq. (94), com o conceito de *informação mútua* ou *incerteza removida*, eq. (71), da teoria da informação de Shannon. Ambas equações expressam uma redução esperada de incerteza, sendo a incerteza medida pela entropia (eq. 68 ou 69) no caso da eq. (71) e medida pela variância no caso da eq. (94). Comparando o lado direito de cada uma dessas equações, a variância da distribuição a priori Var[X] é análoga à entropia incondicional de X, i. é, H(X). Já a variância esperada das distribuições posteriores (ou *condicionais*) E[Var[X | S]], é análoga ao conceito de *entropia condicional* H(X | S), já que essa também é uma entropia posterior (à informação S) *média* em relação aos possíveis sinais S. Para ver isso, rever a eq. 70 (entropia condicional), onde aparece um somatório a mais (comparada com a eq. 68) exatamente para a variável S. Assim, existe uma forte

analogia entre as eqs. (71) e (94), i. é, entre os conceitos de variância da distribuição de revelações e a informação mútua (ou incerteza removida) da teoria da informação.

O curioso é que essa analogia nunca foi notada antes (ao melhor do conhecimento do autor da tese). A explicação é que a eq. (94) é usada mas num contexto e num *formato* diferente (exs.: Williams, 2001, p.392; Ross, 1996, p.51), e é chamada de "fórmula da variância condicional" ou "lei da variância total" (Bertsekas & Tsitsiklis, 2002, p.229), mostrada a seguir:

$$\mathbf{Var[X] = Var[E[X \mid S]] + E[\, Var[X \mid S\,]\,]} \qquad \mathbf{(98)}$$

Em livros-texto, quando essa fórmula é mencionada, ela é usada para calcular a variância de X de forma mais fácil que outros métodos, em alguns tipos de problemas, por ex., quando se quer computar a variância de X sem conhecer a distribuição de X (ver, por ex., Wackerly & Mendenhall III & Scheaffer, 2002, p.273; e Gut, 1995, p.39). Não é esse o caso dessa tese. Com o formato da eq. (98), ela também pode ser interpretada num contexto de *análise de variância*, como caracterizado por Goldberger (1991, p.48). A eq. (98) nesse caso pode ser vista como uma *decomposição da variância* de X em duas partes, uma variância do (melhor) estimador dado S, i. é, E[X | S] e uma variância residual de X após observar S. Mesmo assim é difícil haver menção dessa equação na grande maioria dos livros de estatística. Um motivo é que a ANOVA parte de um modelo paramétrico, mais especificamente linear, e depois se analisa a variância através de várias "somas de quadrados". Note que nessa tese não está sendo imposto modelos paramétricos para a incerteza técnica. O Teorema 1 é válido para praticamente todas as distribuições[186] e é válido para qualquer relação, não-linear ou linear, entre as variáveis aleatórias de interesse X e a informação S.

O formato da eq. (94) é muito similar ao de uma equação de DeGroot (1970, p.432) para o que ele chamou de "*quantidade de informação que pode ser obtido de um experimento*", usando o conceito de *função incerteza* (rever o item 3.1.4.2 sobre *comparação de experimentos*), denotada por U(.). Essa quantidade de informação $\mathfrak{I}(S, U, p(x))$ de DeGroot é dada por (na notação da tese):

$$\mathbf{\mathfrak{I}(S, U, p(x)) = U(p(x)) - E[U(p(x \mid s))]} \qquad \mathbf{(99)}$$

[186] Existem distribuições com média infinita, onde não seria válido o Teorema 1. Mas não se usa esse tipo de distribuição em qualquer das possíveis aplicações imaginadas em petróleo.

A idéia é escolher um experimento que maximiza essa quantidade de informação. Isso é obtido escolhendo um sinal S que tenha o *menor* E[U(p(x | s))] e dessa forma, se S é estatística suficiente para S', então (DeGroot, 1970, p.436):

$$\mathbf{E[U(p(x \mid s))] \leq E[U(p(x \mid s'))]} \qquad \mathbf{(100)}$$

Note que se a função incerteza U(.) for a variância, a eq. (99) é exatamente igual à eq. (94) e no formato usado nessa tese. Nesse caso, $\Im$ (S, U, p(x)) seria exatamente a variância da distribuição de revelações. Infelizmente DeGroot menciona apenas os casos de U(.) ser a função entropia (eqs. 68 e 69) e a informação de Fisher (ver item 3.1.4.2), não mencionando o caso da variância e nem fazendo a conexão que aqui é feita com a eq. 94. Em DeGroot (1970, p.431 e 122-123) a função incerteza é um *risco de Bayes* que está ligada à *função perda*, a qual é vista como o *negativo* da função *utilidade*. Como foi visto antes (item 3.2.1), a função perda mais usada é a *perda de erro quadrático*, e nesse caso a função perda está associada à variância (idéia é minimizar a variância posterior), conforme assinala Trottini (2001, ex. 1, p.11; 2003, ex.6, p.144). Nesse contexto, pode-se inferir que a idéia seria escolher a informação S que maximiza a variância da distribuição de revelações ou, de forma equivalente, a de maior medida η^2.

A propriedade de martingale do item (d) do Teorema 1 é útil para avaliar planos alternativos de investimento seqüencial em informação, inclusive se beneficiando da vasta literatura existente sobre martingales. Por ex., na análise de quando parar otimamente uma seqüência de investimentos em informação, existe uma teoria bem desenvolvida sobre o (mencionado antes) teorema de parada ótima ou teorema de amostragem opcional de martingales.

Com o Teorema 1 pode-se reforçar as críticas, mas também reconhecer uma concordância com o artigo de Cortazar & Schwartz & Casassus (2001) sobre a modelagem da incerteza técnica em OR. Deve-se reconhecer que o artigo acerta ao modelar o processo de incerteza técnica como um martingale, consistente com o item (d) do Teorema 1 dessa tese. Mas eles justificam isso apenas de forma intuitiva ("(end)notes 2") e sem o rigor matemático dessa tese. Uma outra concordância, mas relacionada à Proposição 4, é que a incerteza técnica não demanda prêmio de risco e se pode integrá-la num processo neutro ao risco de mercado. Mas as críticas são agora reforçadas, já que a variância só muda com a

chegada de uma nova informação dada por S (item c) e não pela simples passagem de tempo como nesse artigo. Além disso, os itens (a) e (c)[187] mostram que a variância do processo é limitada (pela variância da distribuição a priori) e não ilimitada como no modelo que eles propõe (eles usam um MGB com $\alpha = 0$).

O teorema da distribuição de revelações deixa de propósito um grau de liberdade para a determinação da seqüência de distribuições de forma a dar *flexibilidade ao modelador*. Exceto para os casos extremos (inicial e revelação total), tem-se a média e a variância da distribuição de revelações, mas não o *tipo* de distribuição (se normal, triangular, etc.). Qual a classe (ou tipo) da distribuição de revelações para o caso de revelação parcial (informação imperfeita)? A resposta geral é que ela depende da *função densidade de probabilidade conjunta* de X e S, i. é, de p(x, s). Ver, por ex., Goldberger, (1991, p.49). Ou seja, no caso geral o item • na estrutura de informação flexível, eq. (63), será p(x, s). Para obter essa distribuição, o método geral usa a simulação de Monte Carlo.

Embora os textos com capítulos sobre expectativa condicional sejam muito comuns, o estudo da *distribuição* de expectativas condicionais (distribuição de revelações), mesmo quando o condicionante é v.a. discreta, é muito difícil de achar (Steckley & Henderson, 2003, p.383). Algumas exceções são: Lee & Glynn (1999), que estimam a distribuição *acumulada*; e Steckley & Henderson (2003), que estimam a função *densidade* dessa distribuição.

Claro que é possível usar uma abordagem mais precisa para determinar a classe dessa distribuição, mas existem os custos adicionais de complexidade e tempo computacional. A abordagem usada nas aplicações dessa tese é mais simplificada, mas é bem mais prática e flexível, como será visto.

A abordagem mais prática usa uma premissa adicional razoável para •, convenientemente escolhida pelo modelador, que definirá totalmente a distribuição de revelações e solucionará o problema. É conveniente, pois essa premissa adicional irá depender do problema. Por ex., no caso exploratório (item 3.4), será assumido que as variáveis de Bernoulli são *intercambiáveis*. Em muitos outros casos pode-se, numa aproximação prática, considerar que *o tipo da distribuição de revelação é do mesmo tipo da distribuição a priori*. Isso será

[187] Como o valor esperado de uma variável que assume apenas valores não-negativos (caso da variância) é sempre não negativo, pelo item (c) a máxima variância de R_X é Var[X].

assumido num exemplo de investimento em informação (Cap.5), mas a rigor só é verdade no limite de revelação total, em que as duas distribuições não apenas são do mesmo tipo como são iguais. Note que distribuição de revelações começa com um ponto[188] (a média da distribuição a priori) e depois evolui com distribuições intermediárias até virar distribuição à priori. A Figura 34 a seguir ilustra isso e também resolve o aparente paradoxo apontado em Martzoukos & Trigeorgis (2001): Se o valor de uma OR aumenta com a volatilidade e se investir em informação reduz a incerteza, então por que aprender? Com o conceito de distribuição de revelações essa questão é facilmente respondida.

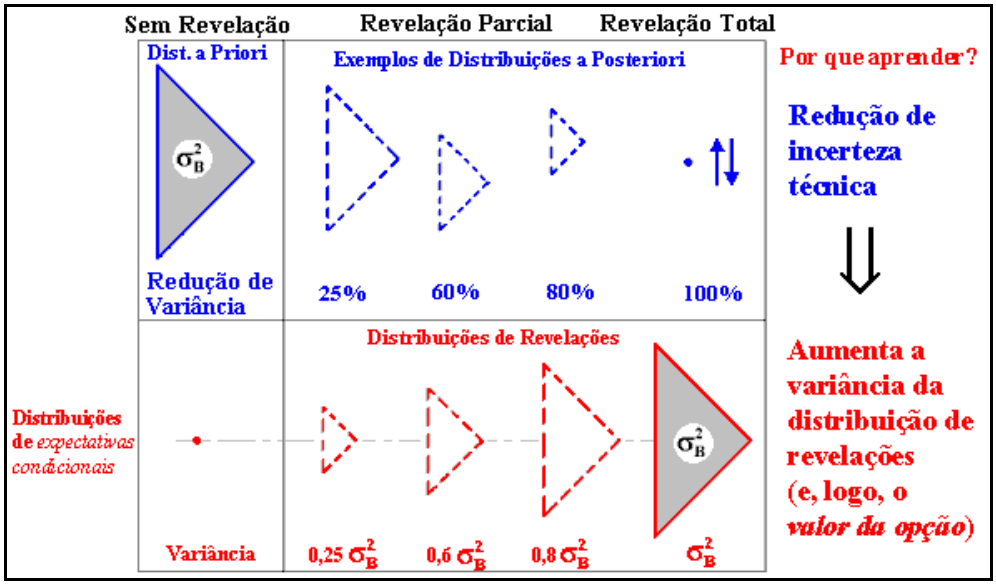

Figura 34 - Exemplo Ilustrando o Teorema 1 sobre Distribuição de Revelações

Na Figura 34, suponha que existe incerteza técnica sobre o volume de reserva B, sendo que os especialistas indicaram uma distribuição a priori triangular para B, com variância σ_B^2. Análises de geologia e reservatórios indicam que as perfurações seqüenciais de três poços de delimitação podem reduzir essa variância inicial em 25%, 60% e 80%. Essas percentagens são as medidas de aprendizagem, i. é, $\eta^2(B \mid S_1)$, $\eta^2(B \mid S_1, S_2)$ e $\eta^2(B \mid S_1, S_2, S_3)$, respectivamente. Essas perfurações reduzem a variância esperada das distribuições posteriores (a parte de cima da Figura 34 mostra exemplos delas, já que existe uma infinidade de distribuições posteriores a cada S_n), mas aumenta a variância da distribuição dos momentos das distribuições posteriores, particularmente da distribuição de

[188] Na eq.(94), em caso de não haver redução de variância, E[Var[X | S]] = Var[X] e assim Var[R_X] = 0. Logo, a distribuição de revelações é um ponto nesse caso trivial sem informação.

expectativas condicionais (ou de revelações), mostrada na parte de baixo da Figura 34. Note que, para cada nova informação ou sinal (S_n), a distribuição de revelações é única (ao contrário das distribuições posteriores). Os 4 itens do Teorema 1 estão ilustrados na Figura 34: se fosse perfurado um número infinito de poços (revelação total), o item (a) do teorema indica que as (inúmeras) distribuições posteriores colapsam para pontos (variância zero) enquanto que a (única) distribuição de revelações se transforma na distribuição a priori (consistência ex-ante). O item (b) do teorema indica que (cada) distribuição de revelações terá valor esperado igual ao valor esperado original de X (da distribuição a priori), ver parte de baixo da Figura 34. O item (c) do teorema diz que as variâncias das distribuições de revelações são dadas pelas reduções esperadas de variâncias que, conforme discutido acima, é obtido simplesmente pela multiplicação da medida de aprendizagem pela variância a priori σ_B^2. Isso está mostrado na parte de baixo da Figura 34. Finalmente, o item (d) do teorema diz que a seqüência das distribuições de revelações é um processo martingale, o que também pode ser visto na parte de baixo da Figura 34: as distribuições de revelações (ex-ante sempre) têm a mesma média e assim visualmente estão centralizadas num mesmo "eixo" mostrada na Figura 34.

O paradoxo de Martzoukos & Trigeorgis (2001) é facilmente respondido com a Figura 34, pois nos modelos de OR se irá trabalhar com a distribuição de revelações e não com uma particular distribuição posterior. Dessa forma, quanto maior a redução esperada de incerteza, maior a variância da distribuição de revelações e maior o valor da opção. Dessa forma, a variância (ou a sua raiz quadrada, o desvio-padrão) da distribuição de revelações joga um papel parecido com a volatilidade nos modelos de OR tradicionais. Ou seja, a volatilidade está para a incerteza de mercado como a (raiz) da variância da distribuição de revelações está para a incerteza técnica no modelo proposto nessa tese. De forma consistente com a teoria tradicional de VOI, o maior valor de informação (antes de computar os custos de adquiri-la) é obtido para o caso de informação perfeita (revelação total), que é justamente o caso em que a distribuição de revelações tem máxima variância.

Na Figura 34 é mostrado que, nos casos de revelação parcial, o tipo de distribuição de revelação é igual ao da distribuição a priori (todas são triangular), mas isso só é absoluta verdade no caso de revelação total. Por isso foram usadas

linhas tracejadas para caracterizar essas distribuições de revelação intermediárias. Essa aproximação é tão melhor quanto mais próximo o problema estiver do caso *limite* de revelação total. Além disso, no caso de S ser v.a. contínua, essa aproximação é melhor do que se S for discreta. Um exemplo numérico tornará isso mais claro.

O exemplo numérico estilizado a seguir[189] sobre a delimitação de um campo de petróleo permitirá "enxergar" melhor os 4 itens do Teorema 1 e dará também uma intuição maior da medida de aprendizagem proposta. Considere um campo de petróleo da Figura 35 a seguir, em que um poço descobriu e provou a existência e um volume de petróleo de 100 milhões de barris (bbl) na área a, mas existe incerteza sobre esse volume B nas demais áreas desse campo (áreas b, c, d).

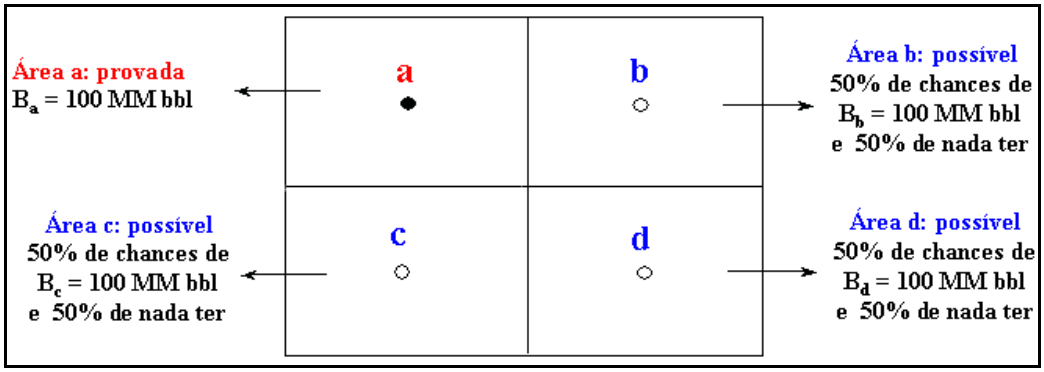

Figura 35 – Exemplo Estilizado de Delimitação de um Campo (Teorema 1)

Cada um dos três poços de delimitação (áreas b, c, d) revelam toda a verdade sobre esse volume na sua área específica. Por ex., na área b existem 50% de chances de ter 100 MM bbl e 50% de nada ter, e essa incerteza (na área b) será totalmente resolvida perfurando o poço no centro dessa área. Por simplicidade, todas as outras áreas com incerteza têm as mesmas características numéricas. Além disso, S_1, S_2, S_3 são independentes, ou seja, só revelam informação relevante nas suas áreas específicas e nada sobre as outras áreas.

Nesse exemplo se pode tanto imaginar um processo seqüencial de investimento em informação (perfuração seqüencial dos três poços – a ordem é irrelevante devido à simetria) ou três planos mutuamente exclusivos de delimitação do campo: alternativa 1 consiste em perfurar um poço; alternativa 2 consiste em perfurar dois poços; e a alternativa 3 consiste em perfurar três poços.

[189] Na verdade foi analisando esse exemplo conceitual e algumas "coincidências" numéricas obtidas, que fez o autor investigar a generalidade das suas conclusões, resultando nos 4 itens do Teorema 1.

Sempre existe também a alternativa trivial (ou base) que é a alternativa zero de não investir em informação, que sempre deve ser considerada já que a informação em geral não é grátis. Note que, por construção do exemplo, a alternativa 3 é a de revelação total, pois perfurando os três poços se revela toda a verdade sobre B.

Note que a distribuição a priori (ou incondicional ou incerteza inicial) é representada pelos seguintes cenários discretos:

- 100 MM bbl com 12,5 % chances;
- 200 MM bbl com 37,5 % chances;
- 300 MM bbl com 37,5 % chances; e
- 400 MM bbl com 12,5 % chances.

Logo, o valor esperado da distribuição a priori é E[B] = 250 milhões bbl e a variância é Var[B] = 7500 $(MM\ bbl)^2$.

Agora será visto o que ocorre com as distribuições posteriores e com as distribuições de revelações em decorrência das diferentes alternativas de investimento em informação. A distribuição de revelações gerada por uma alternativa tem como cenários as médias das distribuições posteriores. Essas são distribuições condicionais à informação S_n gerada pela alternativa A_n.

No caso da alternativa A_1, a perfuração de um poço (ex.: área b) gera que distribuição de revelações? A alternativa 1 revela um de dois cenários, cada um com 50% de chances: o poço b pode resultar em sucesso (S_1 = boas notícias), provando mais 100 MM bbl ou num poço seco (S_1 = más notícias), provando zero barril (inexistência de petróleo nessa área). Esses dois cenários geram os seguintes cenários de médias da incerteza remanescente em B, i. é, cenários da distribuição de revelações (aqui uma distribuição discreta, de dois cenários) dados abaixo:

- E(B | S_1 = boas notícias) = 100 + 100 + (0,5 x 100) + (0,5 x 100) = 300 MM bbl com 50% chances;
- E(B | S_1 = más notícias) = 100 + 0 + (0,5 x 100) + (0,5 x 100) = 200 MM bbl com 50% chances.

Note que, com a alternativa 1 (apenas um poço) é impossível alcançar cenários mais extremos tais como o de 100 milhões de bbl ou 400 milhões de bbl. Isso ocorre porque o *poder de revelação* da alternativa 1 não é suficiente para

mudar de forma tão radical a expectativa do volume de toda a reserva. Os limites também podem ser caracterizados de maneira mais formal[190].

A alternativa 1 alcança apenas uma revelação parcial sobre a v.a. B e assim existe incerteza residual dado pelas variâncias estritamente positivas das distribuições posteriores p(x | s). Qual é a variância *esperada* das distribuições posteriores no caso da alternativa 1? No caso de revelação positiva, a distribuição posterior é {200 MM bbl com 25 % chances; 300 MM bbl com 50 % chances; e 400 MM bbl com 25 % chances}. No caso de revelação negativa, a outra distribuição posterior é {100 MM bbl com 25% chances; 200 MM bbl com 50% chances; e 300 MM bbl com 25% chances}. É fácil calcular e concluir que as variâncias das distribuições posteriores ambos os cenários são 5000 $(\text{MM bbl})^2$, e assim a variância *esperada* das distribuições posteriores também é 5000 $(\text{MM bbl})^2$. Logo, em média a alternativa 1 reduz a incerteza (variância) em 33% (de 7500 para 5000), ou seja, $\eta^2(B \mid S_1) = 33\%$. Isso dá uma intuição para η^2 nesse exemplo simples (simétrico), já que a relação entre o volume revelado e o volume total também é igual a 1/3.

Agora, com a distribuição de revelações da alternativa 1 e as suas correspondentes distribuições posteriores, se poderá checar os itens (b) e (c) do Teorema 1. O valor esperado da distribuição de revelações com a alternativa 1 é:

$\mathbf{E[R_B(S_1)]}$ = 50% x E(B | S_1 = boas notícias) + 50% x E(B | S_1 = más notícias) **= 250 MM bbl**

Igual à média a priori de B, como esperado pelo item (b) do Teorema 1!

A variância da distribuição de revelações com a alternativa 1 é:

$\mathbf{Var[R_B(S_1)]} = 50\% \times (300 - 250)^2 + 50\% \times (200 - 250)^2 =$ **2500 (MM bbl)^2**

Como esperado pelo item (c) do Teorema 1! A variância da distribuição de revelações é igual à redução esperada de variância (= 7500 – 5000) causada pelo investimento em informação na alternativa 1.

De forma similar pode-se checar os itens (b) e (c) do Teorema 1 para os casos das alternativas 2 e 3, assim como checar o item (a) do Teorema 1 para o caso da alternativa 3 (revelação total). Pode-se verificar também que a redução

[190] Por ex., pela *desigualdade de Markov* (ver, por ex., James, 1996, p.125) para o caso de uma v.a. qualquer ou pela versão análoga para o caso de um *processo martingale*, a *desigualdade de Komolgorov-Doob* (ver, por ex., Motwani & Raghavan, 1995, p.92).

esperada de variância da alternativa 2 é 66% (de 7500 para 2500), enquanto que a redução esperada de variância da alternativa 3 é 100% (de 7500 para zero).

A Figura 36 mostra as distribuições de revelações para as três alternativas (ou se preferir, para a seqüência de três investimento em informação).

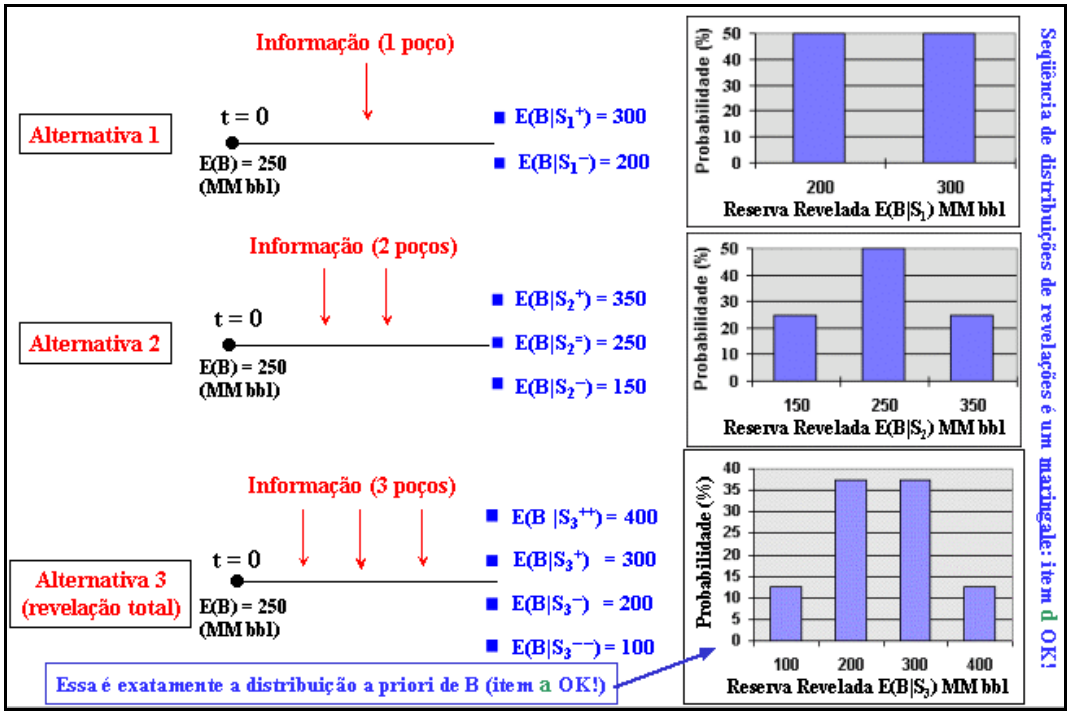

Figura 36 – Distribuições de Revelações para o Exemplo do Teorema 1

Note na Figura 36 que quanto maior o poder de revelação (aqui número de poços), maior a variância da distribuição de revelações. Note também que a distribuição de revelações para a alternativa 3 (revelação total) é exatamente a distribuição a priori, como esperado pelo item (a) do Teorema 1. Em adição, a seqüência vertical de distribuições de revelações tem a mesma média e assim são martingales, como esperado pelo item (d) do Teorema 1. Essas médias são todas iguais à média original da distribuição a priori (250 MM bbl), como esperado pelo item (a) do Teorema 1. Todas as distribuições de revelações são do tipo discreto, embora o número de cenários seja diferente: o intervalo do suporte da distribuição de revelações vai se alargando quanto mais se aproxima do caso de revelação total. Compare também a evolução das distribuições de revelações da Figura 36 com a da Figura 35.

A seqüência de investimentos em informações, com suas distribuições posteriores e de revelações, também podem ser vistas em *diagramas de árvore*. Para efeito de visualização, serão mostrados na árvore apenas dois investimentos

seqüenciais em informação (dois poços do exemplo estilizado anterior). Para ficar mais compacto, se suprimirá os nós de decisão que têm numa árvore de decisão tradicional, assumindo que serão feitos os dois investimentos em informações. A Figura 37 ilustra esse diagrama de árvore com os cenários das distribuições posteriores e de revelação, após a perfuração de dois poços.

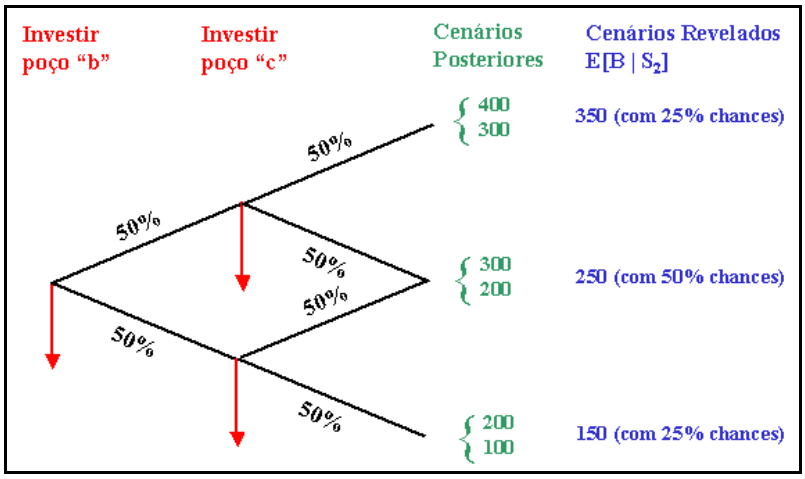

Figura 37 – Diagrama de Árvore para o Exemplo do Teorema 1

Essas distribuições de revelações irão ser aplicadas em problemas mais complexos, mas sua utilidade em OR pode ser vista de forma intuitiva, através da *equação visual de opções reais* (rever a Figura 4), pois ao ser combinado com distribuições neutras ao risco (provenientes de processos estocásticos neutros ao risco), ele aumenta a variância dos possíveis valores do ativo básico (projeto) no futuro e assim aumenta o valor da OR. Como foi discutido antes, quanto maior a variância da distribuição de revelações, maior deve ser o VOI (antes de considerar o custo da informação). Apesar das excelentes propriedades teóricas, o modelo de distribuição de revelações será ainda mais importante na prática de OR.

O mérito do Teorema 1 não é relacionado às demonstrações, já que individualmente cada um dos 4 itens desse teorema já existia na literatura de teoria da probabilidade. O mérito (ou inovação) principal foi selecioná-los e reuni-los para dar suporte ao modelo proposto, assim como o tipo de aplicação (VOI e OR) em que os mesmos foram usados. Ao melhor do conhecimento do autor da tese, isso nunca foi feito antes.

A abordagem proposta poderia ser chamada de *opções reais bayesianas*, já que a distribuição a priori joga um papel fundamental no modelo (daí se tira a média das distribuições de revelações, o limite, etc.). No entanto, não é uma

combinação automática, já que algumas ferramentas são diferentes: na tese não se usará a função verossimilhança, muito usada na abordagem bayesiana tradicional. O item 3.3 a seguir reforçará esses argumentos com a defesa da medida de aprendizagem proposta nessa tese.

3.3. Medidas de Dependência e Medidas de Aprendizagem

3.3.1. Medidas de Dependência de Variáveis Aleatórias

3.3.1.1. Introdução e Limites de Hoeffding-Fréchet para Dependência

Como foi visto no item 3.1, no caso geral é necessário conhecer a distribuição conjunta para definir a *estrutura de informação* do problema de decisão. Além disso, o estudo desse item 3.3 irá reforçar bastante a importância da medida de aprendizagem η^2 proposta, já comentada anteriormente e diretamente ligada ao item (c) do Teorema 1. As excelentes propriedades dessa medida para problemas de investimento em informação serão bastante analisadas aqui.

Sejam duas variáveis aleatórias X e S. A estrutura de dependência entre essas variáveis é definida pela distribuição conjunta bivariada J(X, S). Esse conceito pode ser estendido para o caso de múltiplas variáveis aleatórias, com a estrutura de dependência sendo definida pela distribuição conjunta multivariada.

Na prática, geralmente são dadas as distribuições marginais univariadas e se deseja um parâmetro (ou medida) para estabelecer o grau de dependência entre essas variáveis. No entanto, como será visto, a medida de dependência "ideal" irá depender do tipo de aplicação. Por exemplo, em algumas aplicações medidas simétricas são mais úteis, enquanto que em outras aplicações (como em aplicações de valor da informação) serão mais úteis as medidas assimétricas.

Estudos de dependência usando o popular *coeficiente de correlação* ρ datam de 1885, com F. Galton, medida essa que dominou a estatística praticamente sozinha durante os primeiros 70 anos do século XX (Mari & Kotz, 2001, preface). O primeiro livro texto sobre conceitos de dependência só foi aparecer na segunda metade dos anos 90 (Joe, em 1997). Devido às contribuições de Pearson, no início do século XX, ρ é conhecido por coeficiente de correlação de Pearson. Mas será visto que ρ só tem significado como medida de dependência em *relações lineares*

e/ou *distribuições normais*[191]. Por isso, um erro comum em análise de risco e outras aplicações é o uso desse coeficiente em situações em que ele não se aplica.

A teoria de medidas de dependência é estudada na teoria da probabilidade principalmente no ramo chamado de *distribuições conjuntas dada as marginais* ("distributions with given marginals"), ramo também conhecido como *classes de Fréchet* (Kotz, 1991). Tem raízes nos trabalhos do matemático alemão Hoeffding (1940, 1941) e do matemático francês Fréchet no início dos anos 50. Na década de 50 ocorreram as principais inovações dessa teoria, com o estabelecimento dos *limites de Fréchet-Hoeffding* para distribuições bivariadas dada as marginais (que será visto a seguir) por Fréchet em 1950/51 e Hoeffding (1940, 1941, que só se tornou conhecido nos anos 50)[192]; a teoria dos *espaços métricos probabilísticos* por Schweizer & Sklar em 1957/58, e a *teoria das cópulas* em 1959 por Sklar (respondendo uma questão colocada por Fréchet, ver Schweizer, 1991, ou Nelsen, 1999, p.2). Esse desenvolvimento despertou grande interesse na comunidade acadêmica, já que combinava desenvolvimentos da matemática pura (espaços métricos) com a teoria aplicada de probabilidade.

A função cópula será vista no próximo subitem como uma medida de dependência, sendo que seu sucesso é em parte devido à propriedade de ser uma função de variáveis aleatórias *invariante sob transformações monotônicas*. Por isso Nelsen (1999, p.2) considera os artigos de Hoeffding (1940 e 1941) como precursores da teoria de cópulas de Sklar.

Espaço métrico consiste de um conjunto C e de uma métrica d que mede distâncias entre pontos do conjunto C (ex.: entre a e b). No caso do espaço métrico probabilístico, a distância d(a, b) é substituída por uma função distribuição G_{ab} cujo valor $G_{ab}(x)$ para qualquer número real x é a probabilidade que a distância entre a e b seja menor que x (Nelsen, 1999, p.3). Note que quando se usa o termo "distância" se tem uma métrica necessariamente simétrica, i.é, d(a, b) = d(b, a). Será visto que a métrica copula é simétrica, o que é uma vantagem em algumas

[191] De forma mais geral para *distribuições elípticas* (a distribuição normal multivariada é um caso particular). Outras distribuições elípticas conhecidas são a de Student t, a logística e a de Laplace. Um fato pouco conhecido é que o CAPM e a análise de média-variância em geral são válidos não só para retornos normais como para outros retornos elípticos (Ingersoll, 1987, p.104).

[192] Muitos livros, no entanto, usam apenas o qualificativo "Fréchet". A tese seguirá Nelsen (1999, p.3), que prefere os termos "limites de Fréchet-Hoeffding" e "classes de Fréchet-Hoeffding", já que ambos desenvolveram esses conceitos de forma independente do outro. Hoeffding (1940, 1941) publicou artigos durante a guerra em revistas alemãs pouco conhecidas.

aplicações – por ex. para *distâncias entre distribuições* [193] de probabilidade para variáveis dependentes, mas é uma desvantagem em aplicações dessa tese em que as variáveis aleatórias têm uma relação de condicionamento.

A fórmula que estabelece os chamados limites de Fréchet-Hoeffding para o caso bivariado é dada abaixo. Seja a distribuição (acumulada) conjunta bivariada G com duas distribuições marginais (acumuladas) G_1 e G_2. Fréchet e Hoeffding provaram que:

$$\mathbf{L(x, y) \leq G(x, y) \leq U(x, y)} \qquad \mathbf{(101)}$$

Sendo L(x, y) denominado limite inferior de Fréchet-Hoeffding e U(x, y) chamado de limite superior de Fréchet-Hoeffding, os quais (a prova é simples, ver por ex., Mari & Kotz, 2001, p.68-69) são dados por:

$$\mathbf{L(x, y) = Max\{0, G_1(x) + G_2(y) - 1\}} \qquad \mathbf{(102)}$$

$$\mathbf{U(x, y) = Min\{ G_1(x), G_2(y) \}} \qquad \mathbf{(103)}$$

Uma utilidade desses limites é que se pode construir uma família de distribuições bivariadas dada as distribuições marginais (ou, de forma equivalente, dado os limites L e U) através da combinação convexa ($w_1, w_2 \geq 0$; $w_1 + w_2 = 1$):

$$G(x, y) = w_1 L(x, y) + w_2 U(x, y)$$

Para o caso geral de distribuição multivariada com n variáveis aleatórias de distribuições marginais $G_1(x_1)$, $G_2(x_2)$, … $G_n(x_n)$, os limites de Fréchet-Hoeffding inferior e superior são, respectivamente (Müller & Stoyan, 2002, p.86):

$$\mathbf{L(x_1, x_2, ... , x_n) = Max\{0, G_1(x_1) + G_2(x_2) + ... + G_n(x_n) - (n - 1)\}} \qquad \mathbf{(104)}$$

$$\mathbf{U(x_1, x_2, ...x_n) = Min\{ G_1(x_1), G_2(x_2), \dots , G_n(x_n) \}} \qquad \mathbf{(105)}$$

Essas equações são bem gerais e impõe limites na estrutura de dependência, *qualquer* que seja a maneira que a mesma seja medida. Com essas equações, ou seja, dados os limites de Fréchet-Hoeffding, e com as distribuições marginais, podem-se calcular os limites para cada medida de dependência nesse contexto. Uma aplicação desses limites será vista na discussão de processos de Bernoulli.

[193] A idéia é que a medida de dependência meça a distância em relação à independência, ou seja, dada uma distribuição conjunta com dependência entre as marginais, qual é a distância em relação à outra distribuição conjunta com as mesmas marginais mas expressando independência?

3.3.1.2. Principais Medidas de Dependência

Existem dezenas (talvez centenas) de medidas de dependência entre v.a. analisadas na literatura. Aqui serão mostradas apenas algumas mais importantes. É de se notar que os livros textos de dependência probabilística mal mencionam (se tanto) a medida defendida nessa tese (η^2)[194], mas dá grande espaço para funções cópulas e algumas medidas de correlação de ordem ("rank correlation"), tais como Spearman-ρ e Kendall-τ. No entanto, grandes matemáticos do passado deram atenção a η^2, como Pearson, Kolmogorov, Fréchet e Rényi. Essa medida é útil em muitas aplicações, especialmente as dessa tese. Algumas medidas de dependência são úteis em outras aplicações específicas, por ex., medidas de *dependência de caudas* ("tail-dependence") de distribuições, são úteis em aplicações da área de seguros (onde se está interessado em eventos extremos), mas não no contexto dessa tese.

A seguir serão definidas algumas medidas de dependência, iniciando pela medida defendida nessa tese, a redução (percentual) esperada de variância η^2 em uma variável de interesse X causada pelo conhecimento do sinal (informação) S.

Definição: Sejam duas variáveis aleatórias X e S com médias e variâncias finitas, definidas no espaço de probabilidade $(\Omega, \Sigma, \mathbb{P})$. Define-se a *redução percentual esperada de variância de X dado S* como:

$$\eta^2(X \mid S) = \frac{Var[X] - E[\,Var[\,X \mid S\,]\,]}{Var[X]} \qquad (106)$$

A notação η^2 é adotada por duas razões: (a) para facilitar a conexão com sua interpretação estatística, a *razão de correlação* ("correlation ratio"), também chamada de razão de *correlações* e também conhecida por "*eta-squared*" em alguns livros de estatística; e (b) em algumas situações (ex.: processos de Bernoulli) é mais intuitivo (para efeito da interpretação) usar a raiz positiva de η^2, ou seja, simplesmente η. Isso será discutido em detalhes ainda no tópico 3.3. Muitos autores denotam η "correlation ratio", em vez de η^2 adotado nessa tese. Aqui se segue a nomenclatura adotada por Kolmogorov (1933, p.60) e Stuart &

[194] Joe (1997) e Nelsen, (1999) nem mencionam; Mari & Kotz (2001) dedicam menos de meia página. O principal motivo parece ser a indesejável (para esses autores) falta de simetria de η^2. Mas aqui assimetria será uma vantagem e uma qualidade duma boa medida de aprendizagem.

Ord & Arnold (1999, p.501). A razão de correlação η^2 foi introduzida por Pearson em 1903, conforme reporta Sampson (1984).

Aplicando o item (c) do Teorema 1 sobre a variância da distribuição de revelações na equação que define $\eta^2(X \mid S)$, se obtém:

$$\eta^2(X \mid S) = \frac{Var[\, E[\, X \mid S]\,]}{Var[X]} = \frac{Var[R_X]}{Var[X]} \quad (107)$$

Ou seja, a medida de aprendizagem proposta é a variância normalizada da distribuição de revelações, sendo normalizada pela variância inicial, i. é, pela variância da distribuição a priori.

De maneira análoga, pode-se definir a *redução percentual esperada de variância de S dado X* como:

$$\eta^2(S \mid X) = \frac{Var[S] - E[\, Var[\, S \mid X\,]\,]}{Var[S]} \quad (108)$$

No caso geral, $\eta^2(X \mid S) \neq \eta^2(S \mid X)$ Logo essa é uma *medida assimétrica de dependência* entre variáveis aleatórias. Assim, em geral a medida η^2 não exibe a propriedade de *mutualidade de informação* que ocorre com a métrica da entropia. Isso é uma desvantagem? Não! No contexto da tese de aplicações de valor da informação, aprendizagem, etc., isso é uma vantagem! Isso será mostrado no próximo subitem sobre as propriedades desejadas para medidas de aprendizagem. Em outros contextos, no entanto, a assimetria de uma medida é considerada uma desvantagem: para alguns autores (ex.: Rényi, 1959), a propriedade de simetria é desejável para medidas de dependência. Essa tese irá discordar da generalização desses autores, dizendo que isso depende da classe de aplicação.

Uma medida relacionada que é sempre simétrica é a média aritmética entre $\eta^2(X \mid S)$ e $\eta^2(S \mid X)$. Essa medida simétrica será aqui chamada de *redução média percentual esperada mútua de variância* das variáveis X e S, definida como:

$$\bar{\eta}^2(X, S) = \frac{\eta^2(X \mid S) + \eta^2(S \mid X)}{2} \quad (109)$$

Já a média *geométrica* não é tão interessante pois se apenas um dos dois valores de $\eta^2(. \mid .)$ for zero, a média geométrica também será igual a zero (i. é, poderia ser igual a zero em casos de clara relação de dependência). Lembrar que a média geométrica é sempre menor ou igual que a média aritmética.

No próximo sub-item serão vistas várias propriedades para η^2. Mas é oportuno mencionar que as propriedades de $\eta^2(X|S)$ para o caso que S é um vetor de variáveis aleatórias (em vez de uma variável) são basicamente as mesmas do caso mais simples (ver Hall, 1970, ou Sampson, 1984). No entanto, quando *ambos* X e S são *vetores* de v.a. é necessário trabalhar com uma extensão do conceito de η^2 chamada de *razão de correlação multivariada* (um indicador agregado do impacto da informação num conjunto de variáveis de interesse). Essas extensões são discutidas em Sampson (1984), em Kabe & Gupta (1990) e em Shaffer & Gillo (1974). Mas aqui se trabalhará com um η^2 para *cada* variável de interesse (com o condicionante S podendo ou não ser um vetor). Dessa forma, se terá simulações e distribuições de revelações separadas para cada variável de interesse, não necessitando trabalhar com a versão multivariada de η^2.

Agora será definido o clássico coeficiente de correlação de Pearson. Para isso, é necessário primeiro escrever a equação da covariância entre X e S:

$$\mathbf{Cov[X, S] = E[(X - E[X]) (S - E[S])]} \qquad \mathbf{(110)}$$

Note que Var[X] é simplesmente Cov[X, X]. Desenvolvendo o produto interno da eq. (110), chega-se facilmente à conhecida equação para a covariância:

$$\mathbf{Cov[X, S] = E[X S] - E[X] E[S]} \qquad \mathbf{(111)}$$

O popular coeficiente de correlação de Pearson é a *covariância normalizada* pela média geométrica das variâncias de X e S, i. é, o coeficiente de correlação é uma covariância *adimensional* dada por:

$$\mathbf{\rho(X, S) = \frac{Cov[X, S]}{\sqrt{Var[X]\ Var[S]}}} \qquad \mathbf{(112)}$$

O coeficiente de correlação não depende das origens e das unidades de medida (Feller, 1968, p.236), i. é, é invariante em relação a transformações lineares monotônicas. Ou seja, dadas as constantes a, b, c, d, com a > 0 e c > 0:

$$\mathbf{\rho(aX + b, cS + d) = \rho(X, S)} \qquad \mathbf{(113)}$$

Será visto que o coeficiente de correlação não é uma boa medida geral de dependência, mas é uma boa medida de dependência linear entre X e S.

Lema 3: Sejam X e S v.a. definidas no mesmo espaço de probabilidades. O coeficiente de correlação exibe quase certamente as propriedades:

(a) $\mathbf{-1 \le \rho(X, S) \le +1}$ **(114)**

(b) $\mathbf{\rho(X, S) = +1 \Leftrightarrow X = a S + b\ ,\ a > 0}$ **(115)**

(c) $\rho(X, S) = -1 \Leftrightarrow X = a S + b , a < 0$ **(116)**

Prova: Feller (1968, p.236-237)

Para *relações lineares* entre as variáveis aleatórias ou se as mesmas tiverem *distribuição normal* (ou de forma mais geral, distribuição elíptica), o popular coeficiente de correlação de Pearson ρ pode ser aplicado/interpretado como uma medida de dependência de variáveis aleatórias. A métrica η tem a vantagem prática de ser interpretada como sendo igual ao módulo de ρ *apenas* quando ρ tem realmente significado enquanto medida de dependência e pode ser corretamente aplicado/interpretado. Esse conhecido resultado é resumido no seguinte lema que reforça a métrica defendida nessa tese.

Lema 4: Sejam duas v.a. X e S definidas no espaço de probabilidade $(\Omega, \Sigma, \mathbb{P})$, sendo que X tem média e variância finitas. Então,

(a) $\eta^2(X|S)$ pode ser visto como o supremo do coeficiente de correlação ao quadrado, sendo o supremo tomado em relação a todas as possíveis funções reais g(S), i. é:

$$\eta^2(X \mid S) = \sup_g \rho^2(g(S), X) \quad \textbf{(117)}$$

(b) Em particular, se a função entre X e S for linear, X = a + b E[X | S], com b ≠ 0, *e/ou* se X e S tem distribuição conjunta p(x, s) bivariada normal (logo as marginais têm distribuições normais)[195], então η(X | S) é igual ao módulo do coeficiente de correlação:

$$\eta(X|S) = |\rho(X, S)| \quad \textbf{(118)}$$

(c) $\eta^2(X | S)$ é igual ao quadrado do coeficiente de correlação entre X e o valor esperado condicional $E[X | S] = R_X(S)$:

$$\eta^2(X|S) = \rho^2(X, R_X(S)) \quad \textbf{(119)}$$

Prova: (a) A prova da eq.(117) pode ser vista em Rényi (1970, p.278-279), que usa a desigualdade de Cauchy-Schwarz, $(E[X S])^2 \leq E[X^2] E[S^2]$.

(b) Hall (1970, p.364) demonstra a eq. (118).

(c) Pode ser visto como um caso particular da eq. (117). Segundo Kruskal (1958, p.817), foi provado por Fréchet na década de 30. A prova não aparece nos

[195] Mas o inverso não é verdade, por isso se especifica p(x, s). Kowalski (1973) mostra que é possível ter as distribuições marginais X e S normais, mas a bivariada não-normal. Por ex., a soma de duas *bivariadas* normais escritas nas mesmas variáveis normais X e S, mas cada bivariada com *diferentes* coeficientes de correlação, é uma nova distribuição bivariada não-normal de X e S.

textos talvez por ser muito simples. Como será usada numa proposição, ela será aqui demonstrada. Por definição de variância e usando o Teorema 1 (b), tem-se:

$$\mathrm{Var}[R_X(S)] = \mathrm{Var}[E[X \mid S]] = E[(E[X \mid S] - E[X])^2]$$

Elevando ao quadrado e dividindo por $\mathrm{Var}[R_X(S)]\ .\mathrm{Var}[X]$, vem:

$$\frac{\mathrm{Var}[R_X(S)]}{\mathrm{Var}[X]} = \frac{(\mathrm{Cov}[R_X(S), X])^2}{\mathrm{Var}[R_X(S)]\ \mathrm{Var}[X]}$$

Mas o lado esquerdo da equação acima é $\eta^2(X \mid S)$, enquanto que o lado direito é $\rho^2(X, R_X(S))$, provando a eq. (119). □

A eq. (118) permite uma ligação prática importante entre a medida proposta η^2 e o popular coeficiente de correlação ρ. Isso é uma vantagem prática, pois existem diversos modelos lineares que são populares em diversas aplicações.

Mas o mundo é muitas vezes não-linear. Um mundo não-linear significa que uma variação *infinitesimal* num dado de entrada S ("*input*") pode produzir um efeito *macroscópico* na variável de saída X ("*output*"). Ou vice-versa. A métrica defendida nessa tese tem as vantagens de não ser igual a ρ no mundo não-linear, quando se sabe que ρ não se aplica como medida de dependência, e de ser igual a essa métrica popular, quando ela se aplica como medida de dependência.

Segundo Kruskal (1958, p.817), também foi provado por Fréchet na década de 30 que:

$$\rho^2(X, S) = \eta^2(X|S)\ \rho^2(S, E[X \mid S]) = \eta^2(S|X)\ \rho^2(X, E[S \mid X]) \qquad (120)$$

Uma generalização teórica importante da razão de correlação η^2 é apresentada por Hall (1970) se as variáveis X e S forem complexas (reais como caso particular). Várias propriedades de η^2 para variáveis reais valem para variáveis complexas, em particular a média e a variância de $R_X(S)$ se X e S são v.a. complexas são as dadas nos itens (b) e (c) do Teorema 1. Hall (1970) define o índice de dependência característica $\eta^2(t)$ como a razão de correlação de uma *função complexa* $f(X, t) = \exp(i\ t\ X)$ dado S, onde i é o número imaginário (raiz de – 1) e t é um número real positivo. Essa função complexa tende para X (e logo $\eta^2(t)$ tende para o η^2 tradicional) quando t tende a zero, de forma que $\eta^2(0+)$ é uma generalização de η^2 que sempre existe (mesmo com variâncias *infinitas*) e tem outras propriedades favoráveis.

Existe uma família de medidas de dependência advindas da *teoria da informação* e do conceito de *entropia* (ver item 3.1.4.3). Aqui não serão repetidas

as definições/equações, mas serão destacadas as duas principais medidas de dependência derivadas da entropia: a informação mútua (ou *incerteza removida*) dada pela eq. (71) (ver também as eqs. 72, 73 e 74), que como foi visto acima é muito similar à medida proposta η^2, e a distância de Kullback-Leibler entre duas distribuições (eqs. 75 e 76) e sua variante chamada divergência (eq. 77).

Duas medidas de dependência bivariadas e não-paramétricas muito conhecidas são o Kendall-τ e o Spearman-ρ (conhecida como "rank correlation" de Spearman). Elas são classificadas como medidas de *correlação de ordem* (Kendall, 1962) e medem uma forma de dependência chamada de *concordância* (Nelsen, 1999, p.125). Intuitivamente, um par de v.a. (X, S) é concordante se os valores elevados de uma v.a. tende a serem associados com os valores elevados da outra v.a. (e vice-versa, baixos com baixos). No caso oposto (elevados com baixos), o par é denominado *discordante*. Enquanto o Kendall-τ usa apenas a ordem relativa, o Spearman-ρ usa a diferença numérica entre as ordens ("ranks").

A medida Kendall-τ ("Kendall's tau") das v.a. X e S, denotada por $\tau_{X,S}$, é definida como a probabilidade de concordância menos a probabilidade de discordância dessas duas v.a. Sejam (X_1, S_1) e (X_2, S_2) dois vetores aleatórios independentes e identicamente distribuídos, então a medida Kendall-τ é dada por:

$$\tau_{X,S} = \Pr[(X_1 - X_2)(S_1 - S_2) > 0] - \Pr[(X_1 - X_2)(S_1 - S_2) < 0]$$

Na prática da estimativa estatística, em vez da definição acima (versão de "população"), usa-se a definição chamada de versão de amostra do Kendall-τ:

$$\hat{\tau}_{X,S} = \frac{\textbf{número de pares concordantes} - \textbf{número de pares discordantes}}{\textbf{número total de pares}} \quad \textbf{(121)}$$

Note na eq. (121) que a medida $\tau_{X,S}$ é não-paramétrica (independe da distribuição), simétrica e o intervalo é $[-1, +1]$[196].

A medida Spearman-ρ ("rank correlation"), denotada por $\rho_S(X, S)$, proposta pelo psicólogo inglês Spearman em 1904 (Kruskal, 1958, p.854), tem características similares ao $\tau_{X,S}$, i. é, é não-paramétrica, simétrica, o intervalo é $[-1, +1]$ e é uma medida de concordância. O procedimento de cálculo é dado a seguir. Sejam duas amostras de dados X e S, cada uma com N dados. Em cada

[196] No caso de distribuições conjuntas *descontínuas*, a presença de *empates* ("ties") no ordenamento demanda uma adaptação nas equações de τ. Kendall criou o "tau-b" em 1945 para esses casos. O leitor interessado nesses detalhes pode consultar, por ex., Liebetrau (1983, p.68-72).

vetor, atribua o número 1 ao menor valor e continue atribuindo o número de ordem para cada valor até atribuir N ao maior valor. Para cada par de dados (x_i, s_i) desses vetores, calcule as diferenças quadradas $d_i^2 = (x_i - s_i)^2$. Calcule a soma dessas diferenças e use a fórmula abaixo (caso sem "empates", ver Press et al, 2002, p.645-646 para todos os casos):

$$\rho_S(X,S) = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2 - 1)} \qquad (122)$$

Pela eq. (122), dá para ver que ρ_S só é igual a 1 se todos os pares tiverem o mesmo ordenamento (todos os d_i iguais a zero).

Muitas medidas acima *não* valem para as v.a. categóricas[197], i. é, variáveis *ordinais* (ex.: v.a. pode assumir os valores excelente, bom, regular, ruim, péssimo) e variáveis *nominais* (exs.: sexo: masculino ou feminino; doença X: tem ou não tem, etc.). Essa última é a mais restrita em termos de medidas de dependência já que, além de não ser numérica, não pode ser ordenada. No entanto, se verá que a medida de dependência proposta nessa tese *pode e é* usada para v.a. nominais.

No caso de variáveis ordinais se pode comparar se um valor é maior ou menor que outro, mas não tem muito significado se falar em distâncias entre v.a. Algumas medidas de concordância podem ser usadas para v.a. ordinais, já que o conceito de monotonicidade faz sentido (medir se uma variável tende a aumentar quando uma outra variável aumenta, etc.). Várias medidas de dependência entre v.a. ordinais, inclusive o clássico (data de 1900) coeficiente Q de Yule, são mostradas em Spanos (1999, p.284-286).

No caso de variáveis nominais o conceito de monotonicidade não faz sentido. Mas se quer uma medida de dependência para responder perguntas do tipo: "como o conhecimento da classificação da v.a. S pode ajudar na conjectura da classificação da v.a. X?". A distribuição de probabilidades de v.a. nominais é feita normalmente através de tabelas de contingência, que contém as probabilidades de ocorrência conjunta de v.a. nominais. No caso bivariado (X, S), a tabela de contingência é uma matriz m x n contendo as probabilidades conjuntas π_{ij} dos pares (x_i, s_j) para i = 1, 2, … m e j = 1, 2, … n. Uma conhecida medida de

[197] As v.a. discretas podem ser divididas em três sub-classes (escalar, ordinal e nominal, ver Liebetrau, 1983, p.7) ou em duas sub-classes, escalar e categórica (ver Spanos, 1999, p.25).

dependência para v.a. nominais é devido a Theil que em 1950 usou o conceito de entropia para calcular o coeficiente de incerteza U (ver Spanos, 1999, p.287-288).

Uma classe importante de medidas de dependência para v.a. nominais foi desenvolvida em uma série de artigos por Goodman & Kruskal, especialmente o primeiro artigo (Goodman & Kruskal, 1954). Um dos índices que eles propuseram foi o chamado coeficiente de concentração ou "tau de Goodman & Kruskal", denotado por τ_{GK}. Margolin & Light (1974, p.757) aponta que essa medida tem boas propriedades, em particular ela é invariante a permutações de linhas ou colunas na tabela de contingência, e assim τ_{GK} "*tem sido amplamente adotado na literatura de ciências sociais*". Spanos (1999, p.287) mostra que se pode calcular as variâncias (marginais, condicionais, etc.) numa tabela de contingência através de somatórios de π $(1 - \pi)$. Curiosamente, levou cerca de 20 anos[198] para se reconhecer que o coeficiente de concentração τ_{GK} de Goodman & Kruskal em termos computacionais é exatamente a redução percentual esperada de variância[199] η^2! Isso é mais um indicador da força e da abrangência da medida de dependência η^2 que, infelizmente, não tem sido notado nos livros textos de medidas de dependência dos últimos 10 anos (Joe, Nelsen, Mari & Kotz). Apontando essas conexões (além de ótimas propriedades e outras vantagens), essa tese procura corrigir essa injustificável e grave omissão da literatura recente de medidas de dependência probabilística.

Cópula é uma função que descreve a dependência entre variáveis aleatórias e que vem se tornando muito popular na literatura de finanças. Ver, por ex., o recente livro-texto de Cherubini & Luciano & Vecchiato (2004), dedicado a aplicações de cópulas em finanças. Uma função cópula C é uma função distribuição de probabilidade conjunta de N variáveis aleatórias (multivariada), definida no hipercubo unitário $[0, 1]^N$, isto é, C: $[0, 1]^N \rightarrow [0, 1]$. Ela é obtida da transformação das distribuições marginais através de suas funções distribuições. Isso significa que *todas* as distribuições marginais unidimensionais de C são distribuições uniforme [0, 1]. No caso bivariado, a distribuição conjunta J das v.a.

[198] Margolin & Light (1974) mostra que τ_{GK} é computacionalmente equivalente ao R^2, coeficiente de ajuste duma regressão, que será visto é também equivalente a η^2.

[199] Spanos (1999, p.287) mostra que τ_{GK} é a redução esperada percentual de variância, mas não faz nenhuma conexão com espectativa condicional ou que τ_{GK} é a razão de correlação η^2.

com distribuições (acumuladas) marginais G e H, pode ser representada por uma função distribuição chamada cópula C (marginais uniformes), através da equação:

$$\mathbf{J = C(G, H)} \qquad \mathbf{(123)}$$

O teorema principal para aplicabilidade de cópulas, chamado teorema de Sklar, estabelece que *qualquer* função distribuição conjunta que tem distribuições marginais (G_1, G_2, … G_N) de qualquer tipo, pode ser obtida por uma escolha adequada da função cópula C. Isso significa que, dadas as distribuições marginais, pode-se obter qualquer estrutura de dependência através das cópulas. No entanto, o mesmo teorema estabelece que a função cópula é única apenas se a distribuição conjunta for contínua. Isso porque a função copula trabalha com a função inversa G^{-1} da função distribuição G. Para v.a. discretas, a função G teria degraus e não poderia ser invertida. Como se sabe, para v.a. contínuas essa função inversa G^{-1} é uma distribuição uniforme sobre o intervalo [0, 1], mas o mesmo não ocorre para v.a. discretas. Isso limita a aplicabilidade desse conceito em casos importantes onde se precisa trabalhar com distribuições discretas. Assim, para o caso de distribuições discretas, o conceito de cópula é *insuficiente* para definir dependência entre variáveis aleatórias. O uso de cópulas para distribuições discretas necessita de um critério complementar (tal como a *interpolação bilinear*, ver Nelsen, 1999, p.16-18) e por isso é usado quase exclusivamente em aplicações com distribuições contínuas.

Mais precisamente, se ao menos *uma* distribuição marginal for descontínua, então a cópula não é única. Conforme mostra o artigo crítico de Marshall (1996), isso significa que a *mesma* medida de cópula pode indicar casos opostos de dependência, i. é, indicar o caso extremo de limite inferior de Fréchet-Hoeffding (eq.102) para um par de marginais G_1, H_1 e indicar o limite superior de Fréchet-Hoeffding (eq.103) para outro par de marginais G_2, H_2. Assim, é muito difícil se interpretar essa medida. Marshall (1996) comenta ainda que, embora Nelsen tenha conseguido expressar as medidas de Spearman-ρ e Kendall-τ em termos só de cópulas, em caso de descontinuidade de um dos marginais, não se pode obter essas medidas através da cópula. Em geral, não existe medida de dependência de v.a. que dependa só da cópula. Isso inclui o popular coeficiente de correlação, que também não pode ser obtido através da cópula. Marshall (1996, p.217-218) mostra que dado uma cópula C, o coeficiente de correlação em geral não é único, é todo

um intervalo: se $\rho > 0$, então a cópula determina o intervalo $(0, \rho]$; e se $\rho < 0$, então a cópula determina o intervalo $[\rho, 0)$. Ou seja, a cópula não diz muita coisa sobre o coeficiente de correlação, o que é uma limitação prática importante. Marshall (1996) ainda mostra outros problemas, inclusive problemas de não-convergência de seqüências de cópulas de seqüências de marginais, mesmo quando as distribuições conjuntas convergem e todas as suas marginais são contínuas. Ou seja, apesar da popularidade, as cópulas têm muitos problemas.

O fato do conceito de cópula ser insuficiente no caso de v.a. discretas (e, é óbvio, se aplica menos ainda a v.a. ordinais ou nominais) é uma limitação prática *muito* importante para aplicações de VOI e assim, apesar de toda a popularidade que vem adquirindo nos últimos 10 anos, o conceito de cópula não é adequado para as aplicações dessa tese. Por ex., nas avaliações de projetos de exploração de petróleo, a variável mais básica é uma v.a. discreta de Bernoulli: o fator de chance FC, que fornece a probabilidade de existência de petróleo.

Uma tentativa de usar o conceito de cópula para variáveis de Bernoulli associado a um critério adicional de definição (já que a cópula não é única para v.a. discretas), foi feita por Tajar & Denuit & Lambert (2001). Como não dá para trabalhar com a distribuição uniforme contínua no intervalo [0, 1], foi usada uma distribuição uniforme *discreta* com dois cenários (zero e um), cada cenário com probabilidade ½ (uniforme pois todos os cenários têm mesma probabilidade), i. é, uma distribuição de Bernoulli de parâmetro $p = ½$, que também pode ser vista como distribuição uniforme discreta, servindo como referência da cópula. No entanto, isso gerou uma medida de dependência bem questionável, por ex., a medida não é zero para o caso de independência e sim igual a ¼ o que faz a medida não ser intuitiva e nem prática.

Medidas baseadas em conceitos de dependência positiva, tais como PQD ("positive quadrant dependence"), são muito analisadas em livros texto (ex.: Mari & Kotz, 2001, p.33-36), mas em geral não tem interesse para essa tese pois não tem significado em termos de medida de aprendizagem. Mari & Kotz (2001, p.171) mostram um exemplo em que $\rho = + 0{,}53$, mas que não é PQD. O mesmo exemplo pode ser usado para mostrar que se aprende muito conhecendo uma das variáveis para prever a outra, *apesar* de não ser PQD.

Medidas como a informação de Fisher (ver item 3.1.4.2), apesar da grande popularidade em estatística, não tem sido considerada como medida de dependência. Ela é paramétrica (depende da distribuição), não é normalizada, etc. Mas no limite ela pode ser vista como a medida de Kullback-Leibler, item 3.1.4.3.

3.3.2. Propriedades Desejadas para Medidas de Aprendizagem

Para motivar e ilustrar o estudo das propriedades desejadas (axiomas) para medidas de aprendizagem, considere o seguinte exemplo numérico simples. Esse exemplo foi usado num livro-texto de *teoria da informação* (McEliece, 2002, p.23-24 e 45) para mostrar a superioridade da medida de *informação mútua* (ver eqs. 71 a 74) sobre o popular coeficiente de correlação ρ (eq. 112). Aqui nessa tese o mesmo exemplo é usado mas para mostrar a superioridade da medida aqui defendida (η^2) sobre a medida de informação mútua! Em geral, esse exemplo ilustrará a superioridade de medidas assimétricas sobre medidas simétricas (informação mútua é simétrica), quando o foco de aplicação é valor da informação ou processos de aprendizagem.

Sejam duas variáveis aleatórias A e B, sendo que A pode assumir os valores – 1, +1, – 2, + 2, cada cenário com probabilidade de ¼, enquanto que $B = A^2$. A Figura 38 ilustra esse exemplo[200].

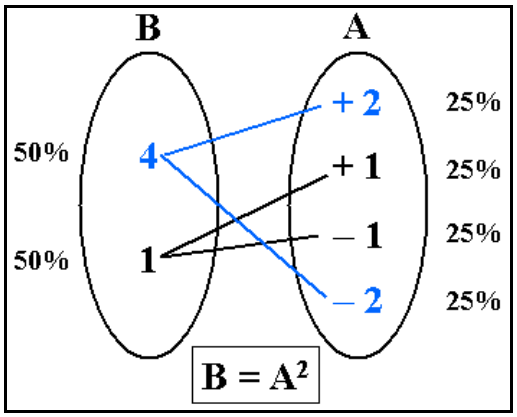

Figura 38 – Exemplo Sobre Medidas de Dependência

Embora claramente as v.a. A e B não sejam independentes – ao contrário, são bastante dependentes pois existe uma dependência funcional – pode-se verificar que A e B tem correlação ρ = zero. McEliece (2002) aponta então a

[200] Ver a planilha dependência_VoI.xls com esse exemplo no CD-Rom.

superioridade da métrica baseada em entropia, já que a informação mútua I(A, B) é diferente de zero, no caso I(A, B) = 1 bit.

Pode-se ver também que embora $\eta^2(A|B) = 0$, tem-se que $\eta^2(B|A) = 100$ %, ou seja, revelação total de B dado A: se for informado o valor de A (por exemplo, A = – 2) a incerteza (variância) em B colapsa para zero (B = 4, com certeza). Esse exemplo mostra a importância da medida de aprendizagem poder ser assimétrica: claramente o conhecimento de A para prever B é muito maior (pois é o limite superior do conhecimento, já que revela toda a verdade sobre A) do que o conhecimento de B para prever A (cuja redução esperada de variância é zero, apesar de reduzir o número de cenários). Assim, métricas simétricas não servem como medidas de aprendizagem para casos gerais, embora muito simples, como no exemplo acima.

Assim, essa tese descarta como medida de aprendizagem todas as medidas simétricas tais como a informação mútua, o coeficiente de correlação, medidas baseadas em cópulas, correlações de ordem (Spearman-ρ e Kendall-τ), etc. No caso de medidas baseadas em entropia, a medida assimétrica chamada *entropia condicional* (ver eq. 70), poderia ser usada como medida de aprendizagem. No exemplo acima, pode-se verificar que as entropias condicionais são H(B | A) = 0 e H(A | B) = 1. Ou seja, distingue a assimetria da relação de A e B, mas de forma oposta à medida η^2, (quando um é zero o outro é 1 e vice-versa). Poderia se pensar então em usar 1 – H(X | S) como medida de aprendizagem, mas enquanto a medida η^2 é sempre normalizada entre 0 e 1, a entropia condicional em geral não é (embora dependa da base do logaritmo). O livro de Cover & Thomas (1991, p.17) mostra um exemplo que tanto H(X|Y) como H(Y|X) são maiores que 1 (logaritmo na base 2). Assim, fica difícil também trabalhar com a entropia condicional como medida de aprendizagem.

E a função verossimilhança[201], que tanto sucesso faz na teoria estatística, tanto clássica como Bayesiana, seria uma boa medida de aprendizagem? Ela é assimétrica, já que é derivado de probabilidades condicionais inversas. Como foi visto, a verossimilhança dá a confiabilidade da informação S para prever X, enquanto a função verossimilhança mapeia essa confiabilidade, i. é, fixando $S = s_i$ e vendo p(s | x) como uma função de X. Para ver que a confiabilidade não é uma

[201] Sobre a função verossimilhança, rever itens 3.1.2 e 3.1.4 e a discussão da eq. (66).

boa base para medidas de aprendizagem, considere dois "expert infalíveis" que podem ser consultados para saber se a ação da companhia X subirá ou não no pregão do dia seguinte da Bolsa de Valores de São Paulo. Suponha que um dos "expert infalíveis" sempre diz a verdade e o outro sempre diz uma mentira, i. é, as confiabilidades são $p_1(s \mid x) = 100\%$ e $p_2(s \mid x) = 0\%$. Do ponto de vista do aprendizado de X, os conselhos são equivalentes pois se aprenderá tudo sobre X (se o "expert infalível 2" dizer que a ação não vai subir, então se saberá que a ação vai subir com probabilidade 1) em ambos os casos. Assim essa medida atribuiria dois valores diferentes para uma mesma aprendizagem (no caso, a aprendizagem máxima), e o que é pior, usando até o valor zero para o caso de aprendizagem máxima. Ou seja, a confiabilidade da informação (verossimilhança) igual a zero pode significar revelação total, que é a máxima aprendizagem possível! Assim, medidas baseadas na confiabilidade da informação (verossimilhança) $p(s \mid x)$ não poderão atender aos axiomas mais elementares sobre medidas de aprendizagem, que serão vistas ainda nesse sub-item.

Voltando ao exemplo da Figura 38, no caso da informação ser A e se querer prever B, a confiabilidade de A é 50% ou zero a depender do valor de B, por ex., se A = 2, então $p(A = 2 \mid B = 4) = 50\%$, $p(2 \mid 1) = 0$. De forma similar, se obteria os valores de 0 e 50% para A = 1, – 1 e – 2. Um valor médio dessa função verossimilhança para A como sinal seria 25%, contrastando com $\eta^2(B \mid A)$ que é de 100%. Já no caso do sinal ser B e a variável de interesse ser A, as probabilidades inversas são 100% ou zero, por ex., se B = 1, então $p(B = 1 \mid A = 2) = 0$, $p(1 \mid 1) = 100\%$, $p(1 \mid -1) = 100\%$ e $p(1 \mid -2) = 0$. O caso de B = 4 seria similar (100% ou 0) e o valor médio dessa função seria de 50%, contrastando com a medida $\eta^2(A \mid B)$ que é de 0%. Assim a verossimilhança não serve como escala de aprendizagem, já que esse exemplo mostrou que ela atribuiu um valor menor para um sinal claramente mais informativo que um outro menos informativo.

Uma crítica baseada na teoria da informação para a medida η^2 seria que ela foi igual a zero no exemplo anterior, $\eta^2(A \mid B) = 0$, apesar do conhecimento de B, tanto no caso de B = 1 como B = 4, reduzir o número de cenários de 4 para 2. Ou seja, para $\eta^2(A \mid B)$, não é relevante a informação dada pelo conhecimento do conjunto B. A resposta a essa crítica é que esse conhecimento é irrelevante no sentido de que não reduz a incerteza (medida pela variância) esperada. Na teoria

da informação a incerteza é medida pela entropia, que se reduz com a redução do número de cenários: H(A) = 2 e reduz para H(A | B) = 1 com o conhecimento de B, enquanto que H(B) = 1 e reduz para H(B | A) = 0 com o conhecimento de A. Em ambos os casos essa redução (que é igual à informação mútua) é igual a 1 (o que não é surpresa, já que a informação mútua é simétrica). Um exemplo de VOI irá indicar que a crítica da teoria da informação não é pertinente.

Seja um problema simples de VOI, em que o valor de um prospecto é baseado no valor esperado do benefício $E[V_k]$ e no investimento I_k. O valor da informação a priori (VDP, ver eq. 52) é dado por:

$$\mathbf{VDP = Max[0, E[V_k] - I_k]} \quad \mathbf{(124)}$$

O benefício é incerto mas o investimento é conhecido. Se k = A, então os possíveis benefícios são dados pelos cenários do conjunto A da Figura 38, e assim $E[V_A] = 0$. Se k = B, então os possíveis benefícios são dados pelos cenários do conjunto B da Figura 38, e assim $E[V_B] = 2,5$. Para efeito de clareza, considere os problemas simétricos em que, sem a informação, o valor da informação a priori é zero em ambos os casos. Para tal, considere $I_A = 0$ e $I_B = 2,5$. Também considere que em ambos os casos o custo de aquisição da informação é zero (já que o interesse é comparar a informatividade dos sinais). Qual seria o VOI nos dois casos (conhecendo A para k = B e conhecendo B para k = A)? Como o VDP nos dois casos é zero, pela eq. (54) o VOI é o próprio VDI (valor da decisão informada, eq. 53). Esses são dados pelas equações:

$$\mathbf{VDI(B \mid A) = E[Max[0, E[V_B] - I_B] \mid A]} \quad \mathbf{(125)}$$

$$\mathbf{VDI(A \mid B) = E[Max[0, E[V_A] - I_A] \mid B]} \quad \mathbf{(126)}$$

Calculando as expectativas condicionais, vem:

$VDI(B \mid A) = 0,25\ (4 - 2,5) + 0,25\ (0) + 0,25\ (0) + 0,25\ (4 - 2,5) = 0,75$

$VDI(A \mid B) = 0,5\ [Max[0, 0 - 0]] + 0,5\ [Max[0, 0 - 0]] = 0$

Assim, de forma consistente com a medida η^2, mas não com medidas simétricas tais como as derivadas da entropia e outras, a informação proporcionada por A para aprender sobre B é relevante, VOI > 0, enquanto que a informação proporcionada por B para aprender sobre A é irrelevante, VOI = 0. Medidas simétricas não notariam a diferença entre esses dois problemas. Note que nesse exemplo simples não foi nem necessário penalizar a decisão baseada em B,

pelo fato de estar calculando um VOI ainda com incerteza residual[202] (enquanto que a decisão baseada na informação A não tem incerteza residual, é informação perfeita). Assim, esse exemplo mostrou que a simetria é indesejável como propriedade de medidas de informatividade em problemas de VOI.

Embora vários autores de livros texto (Joe, Nelsen, e de forma mais comedida em Mari & Kotz) coloquem a simetria como uma propriedade desejável de uma medida de dependência, além dessa tese, existem alguns autores que discordam. Como ressalta Liebetrau (1983, p.18-19), "*se o objetivo é usar uma variável para aumentar a previsibilidade de outra, então medidas assimétricas ... são melhores*". Ele ainda reforça essa conclusão na p. 87. Duas das medidas propostas por Goodman & Kruskal (1954) são assimétricas, propriedade desejável para eles (p.736). Uma delas, o coeficiente de concentração, foi comentada no item 3.3.1.2. Somers (1962) propôs uma medida assimétrica para associação de v.a. ordinais, destacando a "*virtude da assimetria*" (p.803), que distingue a variável "independente" da "dependente" (linguagem usada em regressão).

Lancaster (1963) também defende medidas assimétricas de dependência, mas vai mais longe. Ele destaca que o conceito de *dependência* é assimétrico por natureza (semântica)[203]. Isso significa que não só a medida de aprendizagem, mas também uma medida de *dependência*, deveria ser assimétrica como pré-requisito, contrariando várias listas de "propriedades desejadas para medidas de *dependência*". Ele mostra, dentre outros, o seguinte exemplo em que uma variável Y é *completamente* dependente duma variável X, mas não vice-versa: seja $Y = -1$ quando $X \leq ½$ e $Y = +1$ quando $X > ½$. Assim, conhecer X determina completamente Y, mas o oposto não vale. O conceito *completamente dependente* de Lancaster é similar ao conceito de *revelação total* obtido quando $\eta^2 = 1$.

Outra questão é saber se é desejável para medidas de aprendizagem o sinal (ou a direção) da dependência entre as variáveis aleatórias A e B, se positiva ou negativa (se *concordante* ou *discordante*). Por exemplo, para a teoria de portfólio, introduzir um ativo que tem correlação negativa com o resto da carteira é preferível do que um ativo de correlação positiva (com o mesmo retorno) para efeito de *diversificação*. Assim, para aplicações de portfólio é necessário

[202] Lembrar do fator gama da eq. (61), usado na Tabela 8, que penalisa o VPL e/ou o VDI, em caso de se tomar decisões baseados no valor esperado quando ainda existe incerteza residual.

[203] Exemplo: uma filha pode ser dependente financeiramente do pai, mas não o contrário.

distinguir o sinal da correlação para a correta tomada de decisão de gestão da carteira, pois o efeito da diversificação é bem diferente. Quando essa propriedade é fundamental, se busca um tipo especial de medida de dependência chamada de *medida de concordância* entre variáveis aleatórias[204]. Conforme Kruskal (1958, p.818), enquanto ρ é uma medida de concordância, η é uma *medida de conexão*.

Seria a distinção do sinal da dependência importante também para medidas de aprendizagem? Não! Para ver isso considere um exemplo estilizado simples do mercado financeiro. Sejam dois consultores infalíveis, que sempre acertam o percentual de variação da bolsa de valores para o dia seguinte (a informação que revela toda a verdade). Um investidor compra o conselho do consultor A e um outro investidor compra o conselho do consultor B. O primeiro diz que a bolsa irá subir 2 %, pois a taxa de juros caiu 0,25% e as demais notícias se anulam. O segundo consultor também diz que a bolsa irá subir 2 %, mas porque o índice de satisfação do consumidor subiu 3% e as demais notícias se anulam. Do ponto de vista do investidor, os dois conselhos são equivalentes (eles aprendem que a bolsa vai subir 2%) embora os sinais das correlações da performance da bolsa com indicadores econômicos usados pelos consultores para inferir isso sejam opostos.

Ou seja, para efeito de aprendizagem não importa o sinal de dependência e sim o quanto se pode aprender a respeito da variável de interesse X observando o sinal S. Ou seja, o que interessa é a redução de incerteza (ou quão próximo se está da verdade) e não o sinal da dependência. O caso de portfólio, onde se detém os ativos de risco A e B (*diversificação*), é uma situação bastante diferente do caso de valor da informação, onde se usa uma v.a. B para aprender e tomar decisões ótimas a respeito de outra variável A (*inferência* e *alavancagem*). No caso portfólio tem-se uma *relação aritmética* entre variáveis aleatórias, por ex., $\mathbf{\Pi = w_A \,.\, A + w_B \,.\, B}$, enquanto que em aprendizagem tem-se uma *relação de condicionamento* entre variáveis aleatórias, por ex., se quer conhecer sobre A, dado o conhecimento de B, obtendo-se o valor esperado E[A | B].

Apesar do estudo das medidas de relacionamento entre variáveis aleatórias expressa em distribuições conjuntas se chamar estudo de medidas de dependência, essa tese irá preferir introduzir o nome mais geral de *medida de relacionamento*

[204] Ver item anterior na discussão de correlação de ordem. Para uma definição mais rigorosa de medidas de concordância, ver Scarsini (1984) ou Cherubini et al (2004, p. 95-96).

entre variáveis aleatórias para caracterizar essa classe mais geral. Isso porque, apesar dos argumentos de Lancaster (1963), o nome medida de dependência tem sido usado para um tipo particular de relacionamento entre essas variáveis, relativo a uma referência de v.a. independentes. Ou seja, se deseja medir distância da independência ou distância entre distribuições, no sentido da distância entre uma distribuição conjunta qualquer e a distribuição conjunta que denota *independência* entre as mesmas variáveis. Nesse caso, o valor zero é atribuído *apenas* ao caso extremo de independência e o valor 1 é atribuído apenas ao caso extremo de dependência total, i.é, relação funcional. Dentro dessa visão, muitas vezes essas medidas são usadas para testes estatísticos de independência e assim se deseja que a hipótese nula (medida de dependência igual a zero) use uma medida de dependência que seja zero se e somente se as v.a. forem independentes. Ver, por ex., Tjøsthein (1996). No entanto, como ressalta Goodman & Kruskal (1954, p.740), o fato de uma medida permitir fazer um excelente teste estatístico de independência, não quer dizer que seja uma medida adequada de grau de associação entre v.a.

Para medir aprendizagem, em vez da usar como referência a independência, é em geral preferível que o valor zero seja atribuído ao caso de *nenhuma aprendizagem*, o que pode ocorrer mesmo se as variáveis não sejam independentes. O exemplo anterior (Figura 38, eqs. 125 e 126) mostrou isso claramente: o VOI pode ser zero mesmo com as v.a. não sendo independentes. Goodman & Kruskal (1954, p.742) também ressalta esse ponto, discordando que a medida de associação seja zero apenas no caso de independência. Kruskal (1958) analisou as principais medidas de associação e todas elas podem ser zero mesmo se as v.a. não forem independentes (p.856).

Revendo a literatura das áreas de finanças, economia, e de diversos ramos das teorias de probabilidade e estatística, do ponto de vista de classe de aplicação podem ser identificados pelo menos os seguintes tipos de medidas de relacionamento (ou medidas de dependência) entre variáveis aleatórias:

- Medidas de distâncias entre distribuições;
- Medidas de concordância; e
- Medidas de aprendizagem.

Por isso é muito difícil se fazer uma lista suficientemente geral de propriedades desejadas de medidas de relacionamento entre variáveis aleatórias se essas variáveis se relacionam de forma tão distintas (ex.: aritmético versus condicionamento), ou em aplicações tão distintas como testes de independência (onde a medida tem de ser simétrica, i. é, uma medida de distância entre distribuições) e problemas de predição ou de VOI (medida tem de ser em geral assimétrica, i. é, medida de condicionamento entre distribuições).

Olhando de forma *crítica* para as listas de propriedades desejadas de medidas de dependência, pode-se ter uma idéia do que seriam as propriedades desejadas para medidas de aprendizagem entre variáveis aleatórias. As listas conhecidas são fortemente influenciadas por certos tipos de aplicação. Os exemplos simples acima serão usados para mostrar que as listas encontradas na literatura sobre as propriedades desejáveis de medidas de dependência entre variáveis aleatórias (ex., simetria), são aplicáveis mais para o caso de distâncias entre distribuições e aplicações relacionadas, não para os casos de aprendizagem probabilística. Para que a lista seja geral (válida para vários tipos de aplicações) o número de propriedades desejadas listadas teria de ser o menor possível. Hoeffding (1942) propôs uma lista bem curta para que uma quantidade α sirva como medida do grau da relação entre duas variáveis aleatórias X e Y. Claramente tendo em mente as aplicações de distância entre distribuições, ele considerou apenas as três propriedades listadas a seguir:

1. α deve estar entre dois limites finitos e fixos (ex.: 0 e 1);
2. α deve ser igual ao limite inferior se e somente se X e Y são independentes.
3. α deve ser igual ao limite superior se e somente se X e Y são funcionalmente dependentes.

A primeira propriedade tem grande importância prática especialmente com $\alpha \in [0, 1]$, quando essa medida pode ser interpretada em termos percentuais. As duas outras propriedades parecem também muito atrativas, mas do jeito que estão formuladas implica que o autor imaginava uma medida simétrica para α. Independência é um conceito simétrico (X é independente de Y $\Leftrightarrow$ Y é independente de X). Já a terceira propriedade, para ser válida no contexto de medida simétrica, terá de existir tanto a função Y = f(X), quanto a função inversa

$f^{-1}(X) = X = g(Y)$, ou seja, tem de ser uma função que admita inversa (função 1-1), tal como uma função estritamente crescente (ou uma estritamente decrescente). Isso é mais uma restrição prática gerada pelo "desejo de simetria".

Hoeffding (1942) ainda fez a exigência adicional que as medidas de dependência fossem suavemente em direção dos dois limites, por ex., se o valor de α for só um pouco diferente do limite inferior isso deve implicar que a distribuição de (X, Y) deve diferir só ligeiramente do caso de independência.

Hoeffding (1942) foi modesto ao não colocar na sua lista a principal propriedade da medida de dependência que ele propôs em 1940: ser invariante em relação à mudança de escala. Em geral a invariância é uma propriedade desejável mesmo se o efeito econômico de escala for importante, desde que se use de forma consistente um fator de escala separadamente da medida invariante de dependência em problemas de valor da informação.

Agora será apresentada a lista clássica de sete axiomas de Rényi (1959) para medidas de dependência de v.a., que é similar a outras listas encontradas na literatura, como Nelsen (1999, p.170), Schweizer & Wolff (1981) e Bell (1962). Bell (1962) analisou a informação mútua (eqs. 71-74) à luz dos axiomas de Rényi. Em seguida serão apresentadas algumas alternativas desses axiomas que foram propostas por Hall (1970) e pelo interessante, embora irregular, livro-texto de Mari & Kotz (2001). Todas essas referências mencionam a lista clássica de Rényi (1959). Rényi mostra que apenas o chamado *máximo coeficiente de correlação* $\overline{\rho}$ atende os sete axiomas da sua lista. Hall (1970, p.340-341, 360, 363-364) critica a medida de dependência $\overline{\rho}$ indicando que ela só é aplicável sob certas condições de regularidade e mostra exemplos em que ela não pode ser calculada, além de outros exemplos em que ela é igual a 1 com muita facilidade, inclusive quando X e S são *condicionalmente* independentes[205].

Lista de axiomas de Rényi (1959) e alternativas para uma medida de dependência entre as v.a. X e S, denotada por m(X, S):

A) m(X, S) é definida para qualquer par de v.a. não triviais (i. é, com variâncias estritamente positivas);

B) m(X, S) = m(S, X);

[205] X e S são *condicionalmente* independentes dado Z, se Pr(X | S, Z) = Pr(X | Z). Ou seja, caso se conheça Z, então S não provê informação adicional para X (pois S se torna irrelevante).

C) $m(X, S) \in [0, 1]$;

D) $m(X, S) = 0 \Leftrightarrow$ X e S são independentes;

E) $m(X, S) = 1 \Rightarrow$ existe uma dependência estrita entre X e S, i. é, ou X = f(S) ou S = f(X), onde f(.) e g(.) são funções reais (Borel) mensuráveis;

F) Se as funções do axioma anterior são 1-1 (admitem inversas) $\Rightarrow$ a medida é invariante, i. é, $m(X, S) = m(f(S), g(X))$; e

G) Se a distribuição conjunta de X e S for bivariada normal $\Rightarrow$ $m(X, S) = |\rho(X, S)|$.

Conforme apontado por Rényi (1959, p.444) é bem conhecido o fato de que η^2 atende o axioma G. Mas ele aponta também que η^2 não atende os axiomas A (por causa dos casos de variância infinita); B (η^2 é assimétrica); D (só vale a volta para η^2, i. é, X e S são independentes $\Rightarrow \eta^2(X \mid S) = 0$ e $\eta^2(S \mid X) = 0$; e F (mas será visto que η^2 é invariante num sentido mais interessante, inclusive no caso mais geral de função que não é 1-1). A medida η^2 atende aos axiomas C, E e G.

As alterações dessa lista em Nelsen (1999, p.170) e em Schweizer & Wolff (1981), por defenderem medidas baseadas em cópulas, restringem para v.a. contínuas. No contexto dessa tese, essa é uma restrição inaceitável.

Já o livro de Mari & Kotz (2001, p.150-151) apresenta sua lista de axiomas para medidas de dependência, que parece bem mais razoável que a de Rényi. Eles não mencionam o axioma A; no axioma B eles indicam a como desejável a propriedade de simetria, mas apenas para variáveis intercambiáveis (ver item 3.4 para a definição e discussão); concordam com a normalização do axioma C; relaxam o axioma D ao dizer que, em caso de independência $\Rightarrow$ medida deve ser zero (e não $\Leftrightarrow$, como em Rényi); concordam com o axioma E; relaxam o axioma F ao desejar invariância apenas em relação a transformações lineares (e não para funções 1-1 em geral); e não consideram o axioma G. Ou seja, as modificações de Mari & Kotz melhoram a classificação de η^2 como medida de dependência.

Além disso, Mari & Kotz acrescentam os axiomas: H) medida tem de ter a propriedade de aumentar com o aumento da dependência; I) relação com v.a. ordinais (se a medida é definida tanto para v.a. contínuas como ordinais, as medidas devem ter uma forte relação entre si); e J) as medidas devem ser interpretáveis, i. é, o valor numérico da medida deve ter um significado qualitativo. Claramente medidas baseadas em cópulas não possuem as

propriedades dos axiomas "extras" I e J de Mari & Kotz, enquanto que a medida η^2 atende aos mesmos, assim como ao axioma H.

Hall (1970) propôs algumas modificações nos axiomas de Rényi, a maioria mais favorável para η^2. No axioma A, no entanto, ele é mais rigoroso que Rényi propondo que seja válido também para X e S determinísticos (a tese não concorda, pois não teria interesse prático uma medida de dependência probabilística num caso determinístico). No axioma E, Hall propõe uma versão mais forte, a medida é 1 se X = f(S), enquanto que em Rényi pode ser também S = g(X). Será visto que η^2 atende a uma condição mais forte que em Hall e Rényi (vale o "se e somente se"). No caso do axioma F, Hall propõe uma condição em parte mais forte do que a de Rényi. Esse axioma alternativo é atendido por η^2 e corresponde aos itens (f) e (g) da Proposição 6 abaixo. Hall também substitui o axioma G de Rényi por um outro axioma que também faz (mas de forma bem indireta) uma relação com o coeficiente de correlação. Assim como o axioma original, esse axioma alternativo também é atendido por η^2. Ele corresponde ao item (h) da Proposição 6 abaixo.

A Proposição 6 a seguir terá os seus itens seguindo a seqüência (e as letras, mas minúsculas) dos axiomas de Rényi, para facilitar a comparação, exceto o axioma F, aqui dividido nos itens (f) e (g); e o axioma G, que corresponde ao (h).

Proposição 6: Sejam duas variáveis aleatórias não triviais[206] X e S com médias e variâncias finitas, definidas no espaço de probabilidade $(\Omega, \Sigma, \mathbb{P})$. Seja a medida de aprendizagem $\eta^2(X \mid S)$ definida pela equação (106). Então, essa medida de aprendizagem tem seguintes propriedades:

(a) A medida $\eta^2(X \mid S)$ existe sempre que a variância de X for estritamente positiva (i. é, o problema for não-trivial) e finita;

(b) A medida η^2 é em geral assimétrica, i. é, $\eta^2(X \mid S) \neq \eta^2(S \mid X)$

(c) A medida η^2 é normalizada no intervalo unitário, ou seja[207],

$$\mathbf{0 \leq \eta^2 \leq 1} \qquad \mathbf{(127)}$$

(d) Se X e S são v.a. independentes, então a medida η^2 é zero:

$$\mathbf{X \text{ e } S \text{ independentes } \Rightarrow \eta^2(X \mid S) = \eta^2(S \mid X) = 0} \qquad \mathbf{(128)}$$

[206] Por "v.a. não-triviais", se assume que as variâncias são estritamente positivas. O caso trivial só interessa na análise de limites de processos. Assim, quando se refere a v.a. em geral elas são não-triviais. Como sempre, a proposição é válida com probabilidade 1 (i. é, quase certamente).

[207] Poderia se acrescentar que η^2 é *realmente* uma medida, pois $\eta^2 \geq 0$.

Além disso, η^2 é zero se e somente se a variância da distribuição de revelações for zero também:

$$\eta^2(X \mid S) = 0 \Leftrightarrow Var[R_X(S)] = 0 \quad (129)$$

(e) Se $\eta^2(X \mid S) = 1 \Leftrightarrow$ existe uma função real, a v.a. g(S), em que X = g(S);

(f) A medida $\eta^2(X \mid S)$ é invariante sob transformações lineares de X, i. é, para quaisquer números reais a e b, com $a \neq 0$, tem-se:

$$\eta^2(a\,X + b \mid S) = \eta^2(X \mid S) \quad (130)$$

(g) A medida $\eta^2(X \mid S)$ é invariante sob transformações lineares e não-lineares de S se a transformação g(S) for uma função 1-1 (injetiva ou com função inversa). Em geral, se g(S) é uma função mensurável pela sigma-álgebra gerada por S, então vale a desigualdade:

$$\eta^2(X \mid g(S)) \leq \eta^2(X \mid S)\text{, com igualdade se g(s) for uma função 1-1} \quad (131)$$

(h) Se as v.a. $Z_1, Z_2, \ldots$ são independentes e identicamente distribuídas (iid) e se $S = Z_1 + Z_2 + \ldots + Z_j$ e $X = Z_1 + Z_2 + \ldots + Z_{j+k}$ para quaisquer inteiros não-negativos j e k, com $j + k > 0$, a medida $\eta^2(X \mid S)$ proposta é dada diretamente por:

$$\eta^2(X \mid S) = \frac{j}{j+k} \quad (132)$$

Prova:

(a) Essa propriedade é trivial, considerando a premissa de Var[X] > 0 mas finita e a premissa de Var[S] finita. □

(b) A medida η^2 é *em geral* assimétrica se pelo menos um exemplo mostrar assimetria. Isso foi mostrado em um exemplo nesse sub-item (Figura 38). □

(c) As eqs. (106) e (107) mostram isso: pela eq. (107) η^2 é uma razão de variâncias e assim não pode ser menor que zero devido a definição de variância. A eq. (106) mostra que o valor máximo é igual a 1, pois η^2 é maximizado minimizando E[Var[X | S]], cujo mínimo ocorre para Var[X | S] = 0 para todo $s \in S$, i. é, para $E[Var[X \mid S]] = 0 \Rightarrow \eta^2(X \mid S) = 1$. □

(d) Se X e S são v.a. independentes, então E[X | S] = E[X] e E[S | X] = E[S] (Lema 2(d), eq. (86), propriedade elementar da independência, ver, por ex.,

Williams, 1991, p.88)[208]. Além disso, a variância duma variável Y qualquer é *definida* por $Var[Y] = E[(Y - E[Y])^2]$. Logo, as variâncias das distribuições de revelações $R_X(S)$ e $R_S(X)$ são ambas iguais a zero, pois:

$$Var[R_X(S)] = E[(E[X \mid S] - E[E[X \mid S]])^2] = E[(E[X \mid S] - E[X])^2] = 0$$

$$Var[R_S(X)] = E[(E[S \mid X] - E[E[S \mid X]])^2] = E[(E[S \mid X] - E[S])^2] = 0$$

Onde foram usados o item (b) do Teorema 1 e a propriedade elementar da independência mencionada acima. Se as variâncias das distribuições de revelações são iguais a zero, então também o serão as medidas $\eta^2(X \mid S)$ e $\eta^2(S \mid X)$, pois essas medidas de aprendizagem são variâncias normalizadas das distribuições de revelações (ver eq. 107). Assim fica provado a eq. (128) e a volta ($\Leftarrow$) da eq. (129). Para provar a ida ($\Rightarrow$) da eq. (129), note na eq. (107) que $\eta^2(X \mid S) = 0 \Rightarrow Var[R_X(S)] = \eta^2(X \mid S)\, Var[X] = 0$ se $Var[X] > 0$. □

(e) Se $\eta^2(X|S) = 1$, a eq.(106) $\Rightarrow E[Var[X|S]] = 0 \Rightarrow E[(X - E[X|S])^2] = 0$ e assim com probabilidade 1 $\Rightarrow X = E[X \mid S] \Rightarrow$ X é mensurável pela sigma-álgebra de S e portanto pode ser escrito X = g(S).

Em Hall (1970, p.342), usa-se o axioma que se uma medida de dependência é igual a 1 $\Rightarrow$ X = g(S), mas sem a volta ($\Leftarrow$). Em Hall, a volta X = g(S) $\Rightarrow$ medida de dependência igual a 1, não é considerada necessária. A explicação é que nesses estudos mais teóricos se quer incluir o caso de variância de X igual a infinito (o que não é o caso dessa tese). Aqui, assumindo que Var[X] é finito, então vale a volta para a medida η^2. A prova da volta ($\Leftarrow$) é ainda mais simples: se S é revelado então, pela definição de função, X = f(S) é *unicamente* determinado. Assim $E[Var[X \mid S]] = 0$ e logo $\eta^2(X \mid S) = 1$ pela eq. (106). Conforme Hall (1970, p.342), no caso de variância infinita de X, quando o conhecimento de S *reduz a variância de infinito para um valor finito*, isso poderia ser considerado um *completo* estado de dependência de maneira análoga ao caso de variância de X finita, em que S reduz a variância de X de um valor finito para um valor igual a zero. Nas aplicações da tese em que Var[X] é sempre finito, escrever a propriedade mais forte ($\Leftrightarrow$) é muito mais conveniente. □

(f) A prova é simples. Aplicando a eq. (106) para a variável Y = aX + b:

[208] Mas se $E[X \mid S] = E[X]$ e $E[S \mid X] = E[S]$ não implica que X e S são independentes (embora sejam na grande maioria dos casos). Por ex., se X e Y tem distribuição uniforme sobre um círculo centrado na origem, claramente são dependentes, mas $E[X \mid Y] = E[X]$ e $E[Y \mid X] = E[Y]$.

$$\eta^2(a\,X + b \mid S) = \frac{Var[a\,X + b] \;-\; E[Var[a\,X + b \mid S]}{Var[a\,X + b]} \Rightarrow$$

$$\eta^2(a\,X + b \mid S) = \frac{a^2\;Var[X] \;-\; a^2\;E[Var[X \mid S]]}{a^2\;Var[X]}$$

Como $a \neq 0$, pode-se simplificar (cortar a^2) e obter $\eta^2(X \mid S)$. □

(g) O caso de igualdade se a função $Y = g(S)$ for 1-1 é porque a inversa $S = g^{-1}(S)$ existe, ou seja, g^{-1} é uma *função* que por definição determina S de forma única se for conhecido $Y = g(S)$. Assim, conhecer g(S) equivale a conhecer S.

Se g(S) não fosse 1-1, o conhecimento de g(S) seria menor do que o conhecimento de S, por ex., se $Y = S^2$ então o valor $Y = 1$ poderia ser tanto devido a $S = 1$ como a $S = -1$. Por isso, a intuição diz que $\eta^2(X \mid g(S))$ deve ser menor que $\eta^2(X \mid S)$. Para provar formalmente essa desigualdade, considere a eq. (117) do Lema 4, onde $\eta^2(X \mid S)$ é o supremo de $\rho^2(X, f(S))$ para *qualquer* função real f(S). Isso implica que $\eta^2(X \mid S) \geq \rho^2(X, f(S))$ para *qualquer* função real f(S). Se essa função é qualquer, então isso inclui a função $f(S) = E[X \mid g(S)]$. Logo:

$$\eta^2(X \mid S) \geq \rho^2(X, E[X \mid g(S)])$$

Mas o Lema 4(c), eq. (119), diz que $\eta^2(X \mid Y) = \rho^2(X, E[X \mid Y])$. Fazendo $Y = g(S)$ obtém-se $\eta^2(X \mid g(S)) = \rho^2(X, E[X \mid g(S)])$. Substituindo na desigualdade anterior, vem:

$$\eta^2(X \mid S) \geq \eta^2(X, g(S))$$ □

Essa propriedade vale também para o caso de X e S serem v.a. complexas, conforme mostrou Hall (1970, p.353-354 e 349), numa demonstração similar[209].

(h) Essa propriedade foi sugerida por Hall (1970) como um dos axiomas que uma medida de dependência deveria ter para se relacionar com o coeficiente de correlação, indicando que a razão de correlação η^2 atende esse axioma[210], mas sem apresentar a prova. Assim, é conveniente apresentar uma demonstração, a qual é bem simples se usar a eq. (107), i. é, η^2 como a variância normalizada da distribuição de revelações. Como Z_1, Z_2, … tem as mesmas distribuições, suas médias e variâncias são iguais e conhecidas com a informação corrente, ou seja:

[209] Existe uma propriedade para v.a. complexas que é parecida com a mostrada no Lema 4.

[210] Na verdade tem uma pequena imprecisão em Hall (1970) que menciona que η (e não η^2) atende a esse axioma. Como é mostrado aqui, η^2 é que atende ao seu axioma. O curioso é que Kruskal (1958, p.818) comete a mesma imprecisão para ρ (o correto é que a eq. 132 vale para ρ^2).

$E[Z_1] = E[Z_2] = \ldots = E[Z]$

$Var[Z_1] = Var[Z_2] = \ldots = Var[Z]$

Como as v.a. Z_i são independentes, a variância da soma é igual à soma das variâncias. Assim, a variância de X (soma de j + k v.a. Z_i) é:

$$\mathbf{Var[X]} = Var[Z_1] + Var[Z_2] + \ldots + Var[Z_{j+k}] = \mathbf{(j + k)\ Var[Z]} \qquad \mathbf{(133)}$$

Além disso, o Lema 2(a) (eq. 82) permite decompor E[X | S] como:

$$E[X \mid S] = E[Z_1 + Z_2 + \ldots + Z_j \mid S] + E[Z_{j+1} + \ldots + Z_{j+k} \mid S]$$

Como as v.a. $Z_{j+1}, \ldots Z_{j+k}$ são independentes das v.a. $Z_{i<j+1}$, então pelo Lema 2(d) (eq. 86), pode-se escrever:

$$E[X \mid S] = E[Z_1 + Z_2 + \ldots + Z_j \mid S] + E[Z_{j+1} + \ldots + Z_{j+k}]$$

Além disso, $Z_1 + Z_2 + \ldots + Z_j$ é S-mensurável (pois $S = Z_1 + Z_2 + \ldots + Z_j$) e assim pelo Lema 2(b) (eq. 83), pode-se tirar para fora do operador E[. | S], i. é:

$$\mathbf{E[X \mid S] = Z_1 + Z_2 + \ldots + Z_j + k\ E[Z]} \qquad \mathbf{(134)}$$

Logo, a variância da distribuição de revelações é dada por:

$$Var[E[X \mid S]] = Var[Z_1 + Z_2 + \ldots + Z_j] + k^2\ Var[E[Z]]$$

Mas como as v.a. Z_i são independentes e E[Z] é uma constante (variância igual a zero), pode-se escrever a variância da distribuição de revelações como:

$$\mathbf{Var[E[X \mid S]] = j\ Var[Z]} \qquad \mathbf{(135)}$$

Dividindo a eq. (135) pela eq.(133) se obtém a medida de aprendizagem η^2:

$$\eta^2(X \mid S) = \frac{j\ Var[Z]}{(j+k)\ Var[Z]} = \frac{j}{j+k} \qquad \square$$

Agora serão apresentados os axiomas ou propriedades desejáveis em uma medida de aprendizagem probabilística. Elas são adequadas em especial para problemas de análise econômica, mas possivelmente não apenas (medicina?).

Axiomas para Medidas de Aprendizagem Probabilística: As seguintes propriedades são desejáveis para uma medida de aprendizagem M(X | S):

A) M(X | S) existe pelo menos para todas as v.a. X e S com incerteza não-trivial (estritamente positiva) e todas as v.a. com incerteza finita;

B) M(X | S) deve em geral ser assimétrica para poder medir os casos em que se aprende mais sobre X conhecendo S, do que se aprende sobre S conhecendo X, e vice versa;

C) M(X | S) deve ser normalizada no intervalo unitário para facilitar a interpretação, i. é,

$$\mathbf{0 \leq M(X \mid S) \leq 1} \qquad \mathbf{(136)}$$

D) Se X e S forem <u>independentes</u> $\Rightarrow$ M(X | S) = M(S | X) = 0, pois não há aprendizagem probabilística. Além disso,

$$\mathbf{M(X \mid S) = 0 \Rightarrow \underline{\text{não haver aprendizagem}}} \qquad \mathbf{(137)}$$

Onde não aprender pode ocorrer não apenas por casos de independência. O conceito *aprender* é definido na medida, mas o seu sentido tem de ser <u>invariável</u> (ex.: medida de incerteza igual a zero para todas as aplicações);

E) Em caso de <u>dependência funcional</u> a medida é máxima, i. é, para toda função real f(.):

$$\mathbf{X = f(S) \Rightarrow M(X \mid S) = 1} \qquad \mathbf{(138)}$$

Além disso, se a medida é máxima, M(X | S) = 1, então:

$$\mathbf{M(X \mid S) = 1 \Rightarrow \underline{\text{aprendizagem é máxima}}} \qquad \mathbf{(139)}$$

Onde aprendizagem máxima significa não ser possível aprender mais sobre X e *aprender* é definido na medida, desde que o seu sentido seja <u>invariável</u>;

F) M(X | S) deve ser invariante a transformações lineares (mudança de escala) da v.a. X ou da v.a. S, i. é, para a e b constantes reais, $a \neq 0$:

$$\mathbf{M(a\,X + b \mid S) = M(X \mid S)} \qquad \mathbf{(140)}$$

$$\mathbf{M(X \mid S) = M(X \mid a\,S + b)} \qquad \mathbf{(141)}$$

G) M(X | S) deve ser <u>prática</u> no sentido de ser de *fácil interpretação* (intuitivo) e *fácil de ser quantificada e estimada*.

H) M(X | S) deve ser <u>aditiva</u> no sentido de, caso a informação S possa ser decomposta numa soma de n fatores <u>independentes</u> $S_1 + S_2 + \ldots + S_n$, de forma que o conhecimento de todos esses fatores proporcione uma <u>aprendizagem máxima</u> (aprender definida na medida, mas invariável), então a soma das medidas de aprendizagem deve ser unitária, i. é:

$$\mathbf{M(X \mid S_1) + M(X \mid S_2) + \ldots + M(X \mid S_n) = 1} \qquad \mathbf{(142)}$$

<u>Teorema 2</u> (<u>Medida de Aprendizagem η^2</u>): A medida η^2 atende a todos os axiomas de medidas de aprendizagem acima.

<u>Prova</u>: Note que η^2 tem a variância como medida de incerteza. Assim, *aprender* significa que é esperada uma redução de incerteza medida pela variância. *Aprendizagem máxima* significa reduzir para zero a variância posterior (revelação total) e *não aprender* é não se esperar uma redução de variância. A Proposição 6 prova todos os itens do Teorema 2, sendo que em alguns casos a

medida η^2 atende condições até mais fortes do que os axiomas acima. Os axiomas A, B e C são provados pela Proposição 6 (a), (b) e (c). O Axioma D é provado pela Proposição 6 (d) e pelo fato da eq. (129) implicar em E[Var[X | S]] = Var[X], logo provando a eq. (137) pela definição de "não aprender" dessa medida. O Axioma E é provado pela Proposição 6 (e) e pela definição de "aprendizagem máxima", que implica em E[Var[X | S]] = 0, que implica em $\eta^2(X \mid S) = 1$. O Axioma F é provado pela Proposição 6 (f) e (g), sendo que η^2 atende condições mais restritas: g(S) pode ser uma função 1-1 qualquer (não apenas linear). O Axioma G é o mais subjetivo, mas a medida η^2 atende amplamente no sentido que tem uma interpretação intuitiva de redução de incerteza, e será visto no item 3.3.4 que ele é facilmente estimado com métodos estatísticos populares tais como a regressão e ANOVA. O Axioma H será provado numa versão mais forte (mais geral) no Teorema 3 (ver item 3.3.3 a seguir), em que não apenas a soma de sinais independentes, mas a soma de *qualquer função* (não necessariamente 1-1) real de variáveis independentes, tem a propriedade de aditividade igual a 1 em caso de aprendizagem máxima. □

3.3.3. Outras Aplicações de η^2 e Decomposição do Aprendizado

A medida de aprendizagem η^2 será usada como elemento da *estrutura de informação* nas aplicações de opções reais híbridas ainda nesse capítulo e também no capítulo 5. Nessas aplicações sempre haverá a presença de $\eta^2(X \mid S)$ e da distribuição a priori da variável de interesse X, já que a variância da distribuição de revelações é o produto de η^2 pela variância a priori. A presença constante e particularmente importante da *distribuição a priori*, sugere caracterizar o método como sendo de *opções reais Bayesianas*. Mas os elementos adicionais da modelagem da incerteza técnica dependerão do problema. Em alguns casos, não será necessário conhecer a distribuição ou os momentos do sinal S, mas assumir que as distribuições são do mesmo tipo da distribuição limite (ver parte de baixo da Figura 34 e discussão). Já nos problemas com v.a. discretas como o fator de chance (Bernoulli), isso não é assumido, mas se conhece a distribuição do sinal S (outra v.a. de Bernoulli) e mais uma premissa, por ex., que as v.a. X e S são intercambiáveis, de forma a definir totalmente o problema.

Mas nesse item serão citadas algumas aplicações fora do contexto de opções reais para a medida de aprendizagem η^2, assim como será apresentado um teorema de decomposição do aprendizagem que usa a medida proposta na tese. Isso servirá para reforçar o caráter prático e intuitivo da medida.

A Tabela 9 mostra os principais nomes da medida de aprendizagem η^2 encontrados em diversas literaturas, assim como o seu uso.

Tabela 9 – Nomes e Usos da Medida de Aprendizagem Proposta

Nome de η^2	Principal Uso	Principal Literatura
Redução esperada percentual de variância	Decisão econômica: VOI, OR	Econômica (essa tese)
Razão de correlação	Predição, teste de hipótese	Estatística
Eta-squared	**Associação, poder estatístico de explicação de fator ou modelo**	Ciências sociais e comportamentais

O que Tabela 9 resume é que em alguns contextos η^2 é usada para predição (regressão) ou teste de hipótese de linearidade (ANOVA). Na literatura de ciências comportamentais (principalmente psicologia) e sociais, ela é denominada "eta-squared" (o nome "correlation ratio" não é usado) e serve como indicador de medida de associação para mostrar os fatores mais importantes para explicar um comportamento ou um indicador social. É usado também para verificar o poder explicativo de um modelo (exs.: uma regressão simples ou múltipla). No caso dessa tese, ela é usada como indicador de poder de revelação de um sinal em problemas de decisão econômica de VOI. Mas as aplicações de η^2 são mais amplas do que a Tabela 9 sugere, conforme será visto a seguir.

É interessante observar que essa métrica (η^2) é aplicada de forma intuitiva na literatura econômica sem mencionar nenhum dos nomes conhecidos, o que mostra o seu apelo prático e intuitivo. Por ex., no contexto econômico da teoria dos jogos, Medin & Rodriguez & Rodriguez (2003, p.195-196) analisam um duopólio de Cournot com informação incompleta a respeito da demanda, em que se estuda a atratividade dos duopolistas trocarem informações privadas sobre a demanda. Eles usam um "índice G" para medir a fração da variância do erro de previsão da demanda que pode ser eliminada através da informação que o seu competidor pode fornecer. Esse índice G nada mais é do que o η^2, mas os autores não mencionaram nenhum dos nomes mais conhecidos desse índice, que foi visto apenas como uma medida *natural* para "*mostrar quando as firmas poderiam ter*

incentivos para compartilhar suas informações". Esse tipo de aplicação, modelagem do valor da informação para decisões estratégicas no contexto da teoria dos jogos, será desenvolvida no cap. 5 dessa tese, usando as equações proposições e teoremas desse cap. 3, proporcionando uma abordagem mais detalhada sobre o VOI do que geralmente se vê na literatura de jogos.

A idéia intuitiva de redução de variância encontra aplicações em ciências do gerenciamento, especialmente na área de gerenciamento da qualidade de produtos, serviços e processos. Em 1951, Deming provocou uma revolução nessa área ao introduzir métodos para melhorar a qualidade do produto através justamente da redução da variância do processo. Os métodos de Deming têm sido usados amplamente pelas indústrias e assim parece interessante usar a medida η^2 para planejamento ou para analisar investimentos na melhoria da qualidade.

A medida η^2 tem sido usada nas mais diferentes áreas de conhecimento. O mais antigo uso é na estatística (onde nasceu), por ex., para testar a linearidade dos dados num modelo de regressão, pois se a diferença entre η^2 e ρ^2 for significativa em termos estatísticos, se rejeitaria a hipótese nula da linearidade (ver Lema 4 (a) e (b), para entender o motivo). A discussão e as referências dessa visão estatística serão analisadas no próximo sub-item (3.3.4).

Na área biomédica, tem sido recentemente usada na análise de imagens 3D de ultra-som, que são ferramentas de cirurgias assistidas por computador, ver Pennec et al (2005). Nessa aplicação, se busca uma transformação espacial T, sendo dada a imagem J, chamada de imagem gabarito ("template image", que corresponde ao sinal S), através da imagem transformada J o T que deve ser a mais similar possível da imagem "verdadeira" I. A medida η^2 serve para quantificar a similaridade de imagens, explorando suas características intuitivas e também pelo fato de ser válido no mundo não-linear e ser não-paramétrica. Aplicações biomédicas similares usando η^2 pode ser vista em Roche et al (2001).

Além disso, já foram mencionadas (e serão dadas mais referências dessas literaturas no item 3.3.4) que η^2 tem sido muito usado como medida de associação em psicologia e sociologia, assim como tem sido usado recentemente em física computacional no ramo denominado análise de sensibilidade global (item 3.1.4.4).

Existe potencial para outras aplicações, que aparentemente ainda não foi usada. Por ex., η^2 poderia ser usada como indicador de *causação*, se valendo da

assimetria para ver qual variável é mais provável causar a outra. Mas devem-se ter alguns cuidados. Embora seja clara a ligação entre correlação e causação, a primeira não implica na segunda, embora possa aumentar as chances de haver alguma causação. Por ex., lama e chuva tem uma forte correlação, mas lama não causa chuva e sim a chuva é que causa a lama. No entanto, a freqüência em que ocorrem juntas (lama e chuva) sugere que existe uma boa chance de haver alguma causação entre essas variáveis. Mas a inferência em estudos de causação deve ter alguns cuidados. Como aponta Pearl (2000, p. 342), "*correlações podem ser estimadas diretamente em um estudo simples sem controle, enquanto que conclusões sobre causação requerem experimentos controlados*". Embora η^2 não possa provar uma direção causal, ela pode medir o nível de causação *dada uma premissa* do analista de existir uma certa direção causal.

A seguir será apresentada uma propriedade surpreendente e prática quando a variável de interesse pode ser escrita como uma soma de funções reais de n variáveis aleatórias independentes. Se essas n variáveis explicam totalmente a variável de interesse X, então a aprendizagem pode ser decomposta como a soma das medidas $\eta^2(X \mid S_i)$, i = 1, 2, … n, soma essa que será sempre igual a 1. Essa propriedade será válida para a medida η^2, mas não para medidas concorrentes tais como medidas de cópulas, correlação, informação mútua, etc. A surpresa é que, após uma pesquisa exaustiva na literatura de probabilidade e estatística, essa propriedade não foi reportada anteriormente[211].

O conceito chave a ser usado aqui é a *independência*. Seguindo Breiman (1969, cap. 4), independência de variáveis aleatórias é uma condição forte e tem a propriedade de *família* ou propriedade *hereditária*. Isso é ilustrado com o seguinte resultado conhecido (ver, por ex., Breiman, 1969, p.91), o qual será usado para demonstrar o teorema da decomposição do aprendizado para sinais independentes.

Lema 5 (<u>independência é uma família</u>): Sejam os sinais $S_1, S_2, \ldots, S_n$, variáveis aleatórias *independentes*. Então:

(a) Qualquer grupo menor dessas variáveis também é independente;

[211] Apesar do título, o livro de Vind (2002) não trata desse tipo de assunto e sim de uma visão teórica da incerteza usando topologia – espaços de funções com relações de pré-ordem, onde as medidas de probabilidades são substituídas por um conceito de incerteza na forma de conjuntos convexos.

(b) Para *quaisquer funções* reais f, g, … , h, as variáveis $f(S_1)$, $g(S_2)$, . . . , $h(S_n)$, são independentes; e

(c) Funções de grupos *disjuntos* dessas variáveis são independentes (ex., $f(S_1)$ e $g(S_2, S_3)$ são independentes).

O teorema da decomposição do aprendizado para sinais independentes é apresentado a seguir. Ele mostra diretamente a participação de cada variável no processo de revelação total da variável de interesse X, desde que seja usada como medida de aprendizagem a redução esperada de variância η^2. Será visto com um exemplo simples que outras métricas não exibem essa propriedade.

Teorema 3 (Decomposição do Aprendizado): Sejam os sinais S_1, S_2, … , S_n, variáveis aleatórias *independentes*. Seja X a v.a. de interesse para a aprendizagem, com Var[X] > 0 (não-trivial) e finita[212], assim como todas as variâncias dos sinais. Seja X igual a uma soma de funções desses sinais, isto é:

$$\mathbf{X = f(S_1) + g(S_2) + \ldots + h(S_n)} \qquad \mathbf{(143)}$$

Onde f, g, . . . , h, são *quaisquer* funções reais de v.a., constituindo novas v.a. com variâncias finitas. Então:

$$\mathbf{\eta^2(X \mid S_1, \ldots, S_n) = \eta^2(X \mid S_1) + \eta^2(X \mid S_2) + \ldots + \eta^2(X \mid S_n) = 1} \qquad \mathbf{(144)}$$

Prova: Por definição, $\eta^2(X \mid S_i) = Var[E(X \mid S_i)\,] / Var[X]$, i = 1, 2, … n. Então, a soma dos $\eta^2(X \mid S_i)$ da eq. (144) pode ser escrita como:

$$\mathbf{\eta^2(X \mid S_1) + \eta^2(X \mid S_2) + \ldots + \eta^2(X \mid S_n) =}$$

$$\frac{Var[\,E(X \mid S_1)]}{Var[X]} + \frac{Var[\,E(X \mid S_2)]}{Var[X]} + \ldots + \frac{Var[\,E(X \mid S_n)]}{Var[X]} =$$

$$\frac{Var[\,E[f(S_1) + g(S_2) + \ldots + h(S_n) \mid S_1]] + \ldots + Var[\,E[f(S_1) + g(S_2) + \ldots + h(S_n) \mid S_n]]}{Var[f(S_1) + g(S_2) + \ldots + h(S_n)]}$$

Mas $E[f(S_i) \mid S_i] = f(S_i)$. Além disso, devido à independência entre S_i e S_j, tem-se $E[f(S_i) \mid S_j] = E[f(S_i)]$ e $Var[f(S_i) + g(S_j)] = Var[f(S_i)] + Var[g(S_j)]$, pois funções de v.a. independentes também são independentes (Lema 5). Logo, a expressão anterior pode ser escrita como:

[212] Com a premissa razoável de que a variância da distribuição a priori Var[X] é finita, a variância da distribuição de revelação é também finita e limitada por Var[X], mesmo se $n \to \infty$. Logo, não é necessário trabalhar com variância infinita no limite $n \to \infty$. Isso contrasta com o movimento geométrico Browniano, cuja variância é ilimitada: ela tende a infinito quando $t \to \infty$.

$$\frac{\text{Var}[\, f(S_1) + E[g(S_2)] + ... + E[h(S_n)]] \; + ... + \; \text{Var}[\, h(S_n) + \; E[f(S_1)] + \; E[g(S_2)] + ... + \; E[k(S_{n-1})]]}{\text{Var}[f(S_1)] \; + \; \text{Var}[g(S_2)] \; + \; ... \; + \; \text{Var}[h(S_n)]}$$

Entretanto, embora a função $f(S_i)$ seja uma variável aleatória, o seu valor esperado é conhecido (e finito) no momento inicial. Logo o valor esperado *não-condicional* $E[f(S_i)]$ não é uma variável aleatória, é um número conhecido (i. é, ele é S_i-mensurável). Logo, $\text{Var}[E[f(S_i)]] = 0$. Com isso, vários termos da equação anterior desaparecem:

$$\eta^2(X \mid S_1) + \eta^2(X \mid S_2) + \ldots + \eta^2(X \mid S_n) \; =$$

$$\frac{\text{Var}[\, f(S_1)] + \text{Var}[\, g(S_2)] + ... \; + \text{Var}[\, h(S_n)]}{\text{Var}[f(S_1)] + \text{Var}[g(S_2)] + ... + \text{Var}[h(S_n)]}$$

$$\Rightarrow \; \eta^2(X \mid S_1) + \eta^2(X \mid S_2) + \ldots + \eta^2(X \mid S_n) \quad = \mathbf{1} \qquad \square$$

Essa propriedade dá uma interpretação intuitiva em termos de decomposição do aprendizado, mostrando a importância relativa de cada variável no processo de aprendizagem. Uma aplicação prática imediata é a indicação de vantagens de algumas alternativas de investimento em informação sobre outras, por reduzirem a incerteza em variáveis mais importantes na decomposição do aprendizado.

A demonstração acima serve também para complementar a prova do Teorema 2 em relação ao último axioma para medidas de aprendizagem (ver eq. 142). Note que a eq. (144) é bem mais forte que o exigido pelo axioma. No axioma H (aditividade da medida de aprendizagem), se requer a aditividade igual a 1 apenas se X for a soma de sinais independentes. Isso é um caso particular do Teorema 3, já que para η^2 vale para a soma de funções de sinais e não apenas quando $f(S_i) = S_i$. Essas funções não precisam nem mesmo ser 1-1 (ter inversas). O Teorema 3 pode ter outras conseqüências particulares em outros contextos, como a decomposição da variância em modelos teóricos "exatos" (sem erro residual) de ANOVA para v.a. independentes. No entanto, a literatura de ANOVA não especifica nada similar (talvez por ter predominantemente a visão "amostra" e não a de "população", ver item 3.3.4).

Comentário: Métodos que usam probabilidades inversas $p(s \mid x)$, como os de verossimilhança, não podem ter propriedades similares às da eq. (144), já que nem todas as funções admitem a inversa (a menos que se restrinja a funções 1-1).

Corolário do Teorema 3: O Teorema 3 vale também para o caso de X ser a soma de funções de grupos disjuntos dessas variáveis independentes. Um exemplo seria: $X = f(S_1, S_2) + g(S_3) + h(S_4, S_5, S_6)$.

Prova: similar ao do Teorema 3, mas usando também o item (c) do Lema 5.

A seguir, é mostrado um exemplo simples que ilustra o Teorema 3 e mostra que outras medidas/métricas de dependência não exibem essa propriedade. Considere um campo de petróleo já descoberto, mas com incerteza no verdadeiro volume de reserva, conforme a esquemática Figura 39 abaixo.

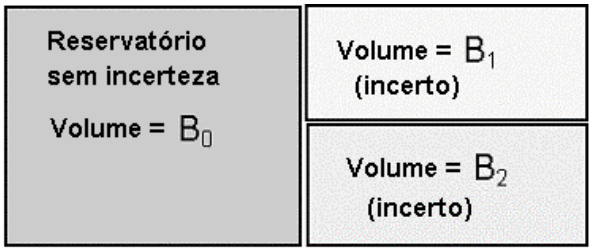

Figura 39 - Exemplo do Teorema da Decomposição do Aprendizado

Denote B o número de barris (ou volume) dessa reserva, onde B é uma v.a.. Nesse campo existe um reservatório com B_0 milhões de barris de reservas provadas, sem incerteza, e dois reservatórios independentes (diferentes idades geológicas) com incertezas, tendo reservas B_1 e B_2 se esses reservatórios estiverem preenchidos com petróleo e zero caso contrário. Embora não existam dúvidas do tamanho dos reservatórios, existem dúvidas se eles estão preenchidos com água ou com óleo. O primeiro reservatório tem probabilidades q de ter petróleo e (1 – q) de ter água, enquanto que o segundo reservatório tem probabilidades p e (1 – p) para petróleo e água, respectivamente. Perfurando um poço em cada área com incerteza se pode revelar toda a verdade a respeito da variável de interesse B. Ou seja, o volume total da reserva B é uma função de apenas duas variáveis aleatórias independentes S_1 e S_2:

$$\mathbf{B(S_1, S_2) = B_0 + (B_1 \times S_1) + (B_2 \times S_2)} \qquad \mathbf{(145)}$$

Onde os sinais S_i são v.a. independentes[213] de Bernoulli de parâmetros q e p, i.é, $S_1 \sim Be(q)$ e $S_2 \sim Be(p)$. Aqui se está interessado em $\eta^2(B \mid S_1)$ e $\eta^2(B \mid S_2)$, ou seja, na relevância dos sinais S_1 e S_2 para prever a variável de interesse B.

[213] Independência significa (para variáveis de Bernoulli vale até "se e somente se") que ambas $\eta^2(S_2 \mid S_1)$ e $\eta^2(S_1 \mid S_2)$ são iguais a zero.

Para facilitar a intuição, considere os valores numéricos: $B_0 = 100$; $B_1 = 50$; $B_2 = 50$; $q = p = 50\%$. O valor esperado de B antes de qualquer nova informação é: $E[B] = B_0 + (B_1 \times q) + (B_2 \times p) = 150$. A variância incondicional é $Var[B] = 1250$. Também é fácil ver/calcular que a revelação do sinal S_1 (perfurando o primeiro poço) reduz Var[B] pela metade, i.e., $\eta^2(B \mid S_1) = 50\%$. Analogamente, $\eta^2(B \mid S_2) = 50\%$. Pode-se verificar que o processo obedece ao Teorema 3, i. é, $\eta^2(X \mid S_1) + \eta^2(X \mid S_2) = \eta^2(f(S_1, S_2) \mid S_1) + \eta^2(f(S_1, S_2) \mid S_2) = 0{,}5 + 0{,}5 = 1$.

Pode-se com esse exemplo verificar que essa propriedade da medida η^2 não se verifica para métricas "competidoras" tais como o coeficiente de correlação $(0{,}71 + 0{,}71 > 1)$[214] e a mútua informação $(1 + 1 > 1)$[215]. Adicionalmente, devido à insensibilidade do valor do cenário, mútua informação não muda se o volume B_1 for o dobro de B_2, quando claramente o sinal S_1 tornar-se-ia mais relevante que o sinal S_2 em termos de redução de incerteza do volume B para aplicações de valor da informação. Ou seja, se $B_1 = 100$ (permanecendo $B_2 = 50$, etc.), a propriedade de η^2 (teorema acima) mostra que o sinal S_1 tem maior peso, i.e., $\eta^2(X \mid S_1) + \eta^2(X \mid S_2) = 0{,}8 + 0{,}2 = 1$.

Esse exemplo simples pode ser estendido para aplicações mais complexas devido à flexibilidade da variável de Bernoulli gerar outras distribuições (Binomial, Normal, etc.) e/ou descrever a incerteza em situações complexas.

Esse tipo de propriedade pode ser útil também quando a variável de interesse é uma função de um *produto* ou de um *quociente* de (funções de) variáveis aleatórias independentes, pois através de uma transformação logarítmica é possível usar o teorema anterior. Por exemplo, nas companhias de petróleo e na literatura profissional o volume da reserva B é estimado considerando B como uma função de várias variáveis independentes. A função mais usada para estimar o volume de reservas é a função $B = f(FR, GV, NTG, \phi, Sw, Bo)$ dada por:

$$\mathbf{B = FR \times [GV \times NTG \times \phi \times (1 - Sw)] / Bo} \qquad \mathbf{(146)}$$

Onde FR = fator de recuperação (% do volume total que é recuperável economicamente); GV = volume de rocha reservatório ("gross volume"); NTG = % da espessura do reservatório com óleo ("net to gross thickness"); ϕ =

[214] Nesse exemplo simples, devido a função $B(S_1, S_2)$ ser *linear*, o quadrado da correlação (*coeficiente de determinação*) é igual ao η^2 (Lema 4). Isso é uma vantagem prática de η^2.

[215] Considerando a base 2 nos logaritmos da equação entropia. Outras bases não terão melhor performance no caso geral.

porosidade da rocha reservatório; Sw = saturação de água; e Bo = fator volume da formação para o óleo[216]. Uma outra função usada na literatura para volume de reservas é B(FR, A, h, ϕ, Sw, Bo) dada por:

$$\mathbf{B = FR \times (A \times h \times \phi \times (1 - Sw)] / Bo} \qquad \mathbf{(147)}$$

Onde A = área; h = espessura com óleo ("net pay") e as outras variáveis são como antes. Com a transformação logarítmica X = ln(B), pode-se escrever X como a soma de funções de variáveis independentes e logo irá obedecer o teorema da decomposição do aprendizado se for usado η^2 como medida de aprendizagem.

3.3.4. Métodos Estatísticos Populares: Outra Vantagem de η^2

Uma das várias vantagens da medida de aprendizagem proposta é que ela pode ser estimada com os métodos estatísticos mais populares. Em particular, η^2 pode ser obtida através dos métodos paramétricos de estimação mais populares: regressões (lineares ou não-lineares) – já que ela pode ser interpretada como o fator de ajuste R^2 da regressão, e a análise de variância (ANOVA), pois é diretamente relacionada com duas *somas de quadrados* usadas nesse método. Esse item irá discutir sucintamente essas estimativas de η^2.

Nenhuma outra medida de dependência é tão fácil de ser obtida através de métodos estatísticos populares. Além disso, como o conceito de η^2 é *não-paramétrico*, pode-se estimar η^2 sem assumir uma distribuição para X ou S, e também sem assumir uma função específica (ou modelo) para X = f(S). Com isso, pode-se estimar η^2 também com métodos não-paramétricos.

Os estatísticos costumam separar os conceitos de *população* e *amostra* na definição das estimativas de parâmetros. Uma população pode ser vista como uma amostra de tamanho infinito. Como em Kruskal (1958), a ênfase aqui é visão de população (visão mais probabilística, que permite interpretação), mas será mostrada sucintamente a visão de amostra (visão mais estatística), sendo dadas referências para trabalhos empíricos (que não é o foco da tese).

A discussão de estimadores (visão amostra) para η^2 é bem antiga, por ex., Pearson (1911) aponta que a estimativa comum de η^2 é tendenciosa para cima[217]

[216] Devido à menor pressão na superfície, o volume de óleo do reservatório se expande.

(maior que o verdadeiro η^2), em geral quanto menor for o tamanho da amostra. Assim, Pearson introduz um fator de correção que assintoticamente tende para zero tanto quando o tamanho da amostra tende a infinito, como quando η^2 tende para 1. Outra discussão estatística clássica, ainda mais detalhada, é de Wishart (1932). Também Kelley (1935) propôs uma fórmula para estimar η^2 usando uma amostra, a fim de corrigir o seu erro sistemático.

A eq. (107) de η^2 pode ser interpretada no contexto estatístico de uma regressão (linear ou não-linear) como o coeficiente R^2, que dá o ajuste do modelo em relação aos dados. O modelo de regressão é uma função X = f(S), onde f(S) é chamado de preditor de X baseado em S. Assim, a eq. (107) é interpretada como a razão entre a variação explicada pelo modelo e a variação total de X. Na versão "população", isso significa que (ver, por ex., Pedhazur, 1997, p.355):

$$\eta^2(X \mid S) = R^2(X = f(S)) = \frac{\text{variância explicada}}{\text{variância total}} \qquad (148)$$

No caso da regressão linear ser o modelo que dá a melhor previsão de X baseada em S, como foi visto, η^2 é igual ao quadrado do coeficiente de correlação ρ. Dada a regressão linear que *explica* X, i. é, X = a S + b, dada a regressão total que inclui o erro ε, i. é, X = a S + b + ε, e se realmente o modelo linear é o melhor modelo, então pode-se escrever a eq. (148) como (ex.: Kruskal, 1958, p.817):

$$\eta^2(X|S) = \rho^2(X, S) = \frac{\text{Var}[a\,S + b]}{\text{Var}[a\,S + b + \varepsilon]} = \frac{a^2\ \text{Var}[S]}{a^2\ \text{Var}[S] + \text{Var}[\varepsilon]} \qquad (149)$$

Caso a amostra mostre uma divergência significativa entre η^2 e ρ^2, então é uma indicação que a regressão linear não é um bom modelo para explicar X. Esse teste de linearidade usando η^2 e ρ^2 é discutido em Stuart & Ord & Arnold (1999, p.501-502). Serão mostradas as equações para estimar η^2 a partir de amostras de dados S. O valor de amostra de R^2 duma regressão é mostrada rotineiramente em software populares (ex.: Excel), dada a regressão que, se for linear, então $\rho^2 = R^2$.

No caso da versão amostra, será usado um *acento circunflexo* em η^2 para caracterizar que é uma medida baseada numa amostra. Nessa versão, devem ser

---

[217] Uma maneira intuitiva de ver isso é que η^2 é uma razão de variâncias, sendo que a variância maior fica no denominador. Como se sabe, na estimativa de variância de uma amostra de tamanho N, se divide a soma $(x_i - E[x])^2$ por N – 1 (em vez de N) para corrigir a tendenciosidade (ao custo de erro esperado maior). Se dividisse por N, o estimador seria tendencioso *para baixo*.

analisados alguns detalhes típicos da literatura estatística tais como a questão dos *graus de liberdade* a serem usados para estimar as variâncias da eq. (148).

Uma regressão também pode ser interpretada num contexto mais geral de análise de variância (ANOVA), desenvolvida pelo estatístico inglês R.A. Fisher na década de 20 do século XX. No entanto, Fisher atacou o uso estatístico de η^2 e assim teve divergências famosas com Pearson. Conforme reporta Levine & Hullet (2002, p.620 e n.5), a implicância de Fisher era devido à estimativa de η^2 ser tendenciosa para cima especialmente para pequenas amostras e devido a ele não conhecer a distribuição de η^2 (Fisher só sabia que tinha a desvantagem de não ser uma distribuição normal) e assim não poder calcular o *intervalo de confiança* da estimativa. Mas existem métodos para corrigir a tendenciosidade (como visto, o próprio Pearson apresentou um)[218] e hoje em dia se sabe que a distribuição empírica de η^2 é uma distribuição beta (isso foi mostrado por Hotelling em 1931, ver Keeping, 1962, p.345-346), tornando datadas as críticas de Fisher.

Maxwell & Camp & Arvey (1981, p. 526) aponta que a ANOVA pode ser conceituada como um caso especial de análise de regressão. Miles & Shevlin (2001, p.33) ressalta que, embora a ANOVA seja mais usada para examinar as diferenças entre *médias de grupos* (grupos de fatores expressando diferentes modelos para explicar a variável X), a ANOVA examina variabilidade e assim pode ser aplicada para ver quanto da variância total de X é explicada pelos fatores S_i (por cada um ou por grupos de fatores). Em geral, a resposta (dos dados) de qualquer modelo pode ser escrita como a predição do modelo mais um erro, i. é:

$$\textbf{Resposta} = \textbf{Modelo} + \textbf{Erro} \qquad \textbf{(150)}$$

Dessa forma, a variação total da resposta dos dados pode ser decomposta em termos de variação devido ao modelo mais a variação não explicada pelo modelo. Nessas análises estatísticas, a variação é dada pela *soma de quadrados* ("sum of squares", SS). Essas somas de quadrados estão diretamente relacionadas às respectivas variâncias (já que a variância é uma média de soma de quadrados, por definição). No contexto de regressão, a variação da equação anterior é dada por:

$$\mathbf{SS_{total}} = \mathbf{SS_{regressão}} + \mathbf{SS_{erro}} \qquad \textbf{(151)}$$

[218] Além disso, como reporta Levine & Hullet (2002, p.620), as pesquisas geralmente usam um tamanho suficientemente grande de amostra, de forma que o problema de tendenciosidade "é de pequena importância na prática".

Onde SS_{total} é a soma dos quadrados dos desvios de cada resposta observada (cada dado) x_i em relação à (grande) média das respostas $\bar{x}$. De forma similar, $SS_{regressão}$ é a soma dos quadrados dos desvios de cada resposta prevista (pelo modelo de regressão) $\hat{x}_i$ em relação à média das respostas $\bar{x}$. Finalmente, SS_{erro}, também chamada de soma dos quadrados dos resíduos[219], é a soma dos quadrados dos desvios de cada resposta observada (cada dado) x_i em relação à sua respectiva resposta prevista (ou explicada) pelo modelo de regressão $\hat{x}_i$.

No contexto mais comum em que a ANOVA é usada para comparar médias entre grupos, a soma total dos quadrados é decomposta na soma dos quadrados *entre os grupos* (modelo) mais a soma dos quadrados *dentro dos grupos* (erro):

$$\mathbf{SS_{total} = SS_{entre} + SS_{dentro}} \qquad \mathbf{(152)}$$

As três últimas equações são equações estatísticas análogas à *lei da variância total* (ou fórmula da variância condicional), eq. (98), onde a variância a priori é a média do SS_{total}, a variância de E[X | S] pode ser vista como uma média do $SS_{regressão}$ (se o modelo é uma regressão) e E[Var[X | S]] pode ser vista como uma média do SS_{erro}. No caso de grupos de dados, do ponto de vista de população, a medida η^2 pode ser vista como:

$$\mathbf{\eta^2(X \mid S) = \frac{Var[total] - Var[dentro]}{Var[total]} = \frac{Var[entre]}{Var[total]}} \qquad \mathbf{(153)}$$

No caso do modelo ser uma regressão, na eq. (153) a razão seria a variância da regressão em relação à variância total (regressão + erro), comparar a eq. (153) com a eq. (148). No caso de uma regressão, com uma amostra de N dados i.i.d., as somas dos quadrados são dadas por:

$$\mathbf{SS_{total} = \sum_{i=1}^{N}(x_i - \bar{x})^2} \qquad \mathbf{(154)}$$

$$\mathbf{SS_{regressão} = \sum_{i=1}^{N}(\hat{x}_i - \bar{x})^2} \qquad \mathbf{(155)}$$

$$\mathbf{SS_{erro} = \sum_{i=1}^{N}(x_i - \hat{x}_i)^2} \qquad \mathbf{(156)}$$

[219] Em ANOVA, esse erro quadrático é também chamado de variação "dentro dos grupos".

Uma tabela de (um-fator) ANOVA tipicamente mostra[220] essas somas de quadrados e o teste estatístico baseado na distribuição-F, dado pela estatística f_{stat}, definido por uma razão de *médias quadradas* (que são variâncias). No caso de uma regressão:

$$\mathbf{f_{stat} = \frac{SS_{regressão} / df_{regressão}}{SS_{erro} / df_{erro}}} \quad \textbf{(157)}$$

Onde df ("degree of freedom") significa os *graus de liberdade*, os quais são funções do tamanho da amostra e do número de fatores[221]. Ex., numa regressão linear, os fatores ou parâmetros a estimar são dois, o coeficiente angular e o coeficiente de interseção. Nesse caso, $df_{regressão} = 2 - 1 = 1$ e $df_{erro} = N - 2$. No caso mais comum de comparação de médias entre grupos, a razão f_{stat} é similar, bastando trocar os subscritos "regressão" por "entre" (os grupos) e "erro" por "dentro" (dos grupos). Note que variância entre as médias dos grupos é uma média dos quadrados $(\hat{x}_k - \overline{x})^2$, análoga à variância das médias condicionais (i. é, variância da distribuição de revelações).

A estatística f_{stat} é diretamente relacionada com a razão de correlação η^2. Essa termos de *população*, a relação entre η^2 e a razão f é dada por (ver, por ex., Cohen, 1988, p.281):

$$\mathbf{\eta^2 = \frac{f^2}{1 + f^2}} \quad \textbf{(158)}$$

O valor de amostra de $\hat{\eta}^2$ pode ser estimado de uma tabela típica de ANOVA usando diretamente a razão das somas quadradas entre os grupos (ou da regressão) e a total, ou seja, o estimador é dado por (Pedhazur, 1997, p.355):

$$\mathbf{\hat{\eta}^2 = \frac{SS_{entre}}{SS_{total}}} \quad \text{ou} \quad \mathbf{\hat{\eta}^2 = \frac{SS_{regressão}}{SS_{total}}} \quad \textbf{(159)}$$

O valor R^2 (eq. 148, geralmente reportado na tabela de ANOVA) é outro estimador de η^2 equivalente ao da eq. (159), ver, por ex., Maxwell & Camp & Arvey (1981, p. 529). Nas análises tradicionais de ANOVA, apesar da facilidade

[220] Mostra outros indicadores também, por ex., o *p-valor*: se o p-valor for menor que um certo nível de significância, então a hipótese nula (ex.: de não ter relação linear) pode ser rejeitada.

[221] O número de graus de liberdade total é N – 1, o do modelo (ou "entre grupos") é k – 1 e o do erro (ou "dentro do grupo") é N – k, de forma que se tem N – 1 = (k – 1) + (N – k). Ver qualquer texto popular de estatística, por ex., Levine & Berenson & Stephan (1999, p.604-610).

de calcular $\hat{\eta}^2$, esse valor geralmente não é reportado[222], uma vez que o interesse é fazer um teste de hipótese para rejeitar ou não um determinado modelo. Nesse contexto, basta calcular f_{stat}, com isso determinar o p-valor[223], e em seguida comparar o p-valor com o nível de significância desejado no teste de hipótese. Em alguns textos, em vez do p-valor se compara a estatística obtida f_{stat} com um $f_{crítico}$ (também tirado de tabela). Nesses testes de hipótese, o cálculo de η^2 não é necessário.

No entanto, não apenas as aplicações dessa tese podem se beneficiar da estimativa $\hat{\eta}^2$, como também existem aplicações tradicionais em estatística que se beneficiam desse indicador. É o caso da literatura de "poder estatístico" que usa η^2 como um *índice de efeito de tamanho* ("effect size index"), ver, por ex., Cohen (1988) ou Pedhazur (1997). Essas análises são usadas especialmente em ciências humanas comportamentais (como psicologia) e ciências sociais. A idéia é determinar se um fator (ou grupo de fatores) é importante para prever a realidade. Exemplos: (a) pode-se querer saber a importância dos fatores "renda" e "idade" na demanda de um produto a fim de desenhar uma campanha de marketing; (b) pode-se querer saber a relevância dos fatores "classe social" e "educação" nos índices de criminalidade, para desenhar políticas sociais.

Aqui é oportuno apontar as diferenças de estimativa de η^2 quando existe apenas um sinal S (um fator A na linguagem de ANOVA) e quando existe um vetor de sinais $S = (S_1, S_2, \ldots S_n)^T$, por ex., os fatores A, B, C influenciando a variável dependente X, na linguagem de ANOVA. No caso de vetor, existem duas estimativas de η^2 na literatura, o η^2 como aqui apresentado e o parcial-η^2 ("partial eta squared"). É importante fazer essa distinção já que existem software que fazem confusão e usam o parcial-η^2 como se fosse o η^2, por ex., o pacote estatístico SPSS, conforme reportado por Levine & Hullet (2002, abstract). Dentro do contexto de ANOVA para três fatores A, B, C, a estimativa de η^2 e parcial-η^2 para um dos fatores (por ex., A) para uma amostra grande, são respectivamente dados em Cohen (1973, eq.1 = eq.6 e eq. 3) por:

[222] Ex., na ferramenta de análise de dados do Excel, a tabela ANOVA mostra também o valor de R^2 e R^2-ajustado (ambos são estimadores de η^2), mas não menciona diretamente o η^2.

[223] O p-valor é obtido em tabelas de ANOVA que usa a distribuição F (F de Fisher).

$$\hat{\eta}^2(X \mid A) = \frac{SS_A}{SS_{total}} = \frac{SS_A}{SS_A + SS_B + SS_C + SS_{erro}} \quad (160)$$

$$\text{parcial-}\hat{\eta}^2(X \mid A) = \frac{SS_A}{SS_A + SS_{erro}} \quad (161)$$

Como ressalta Cohen (1973), o parcial-η^2 é maior que o η^2 (a não ser para o caso de apenas um fator, em que as duas medidas são iguais), o que pode levar a uma posição muito otimista do efeito do fator A para predizer X. Tanto Kennedy (1970) quanto Levine & Hullet (2002, p.620) criticam[224] o uso do parcial-η^2 por não ter a interpretação intuitiva de ser a percentagem da soma total dos quadrados, não ser aditivo e não ser equivalente aos conceitos familiares de R^2 (regressão geral, ver a seguir) e ρ^2 (regressão linear), como ocorre com o (não-parcial) η^2. Por isso, essa tese recomenda trabalhar com η^2 e ter cuidado com software que pode estar reportando parcial-η^2 em vez de η^2.

A ligação de $\hat{\eta}^2$ com o coeficiente R^2 de uma regressão é reportada em vários textos e artigos (por ex., em Levine & Hullet, 2002, p.619; Pedhazur, 1997, etc.). O coeficiente $R^2(X = f(S))$ no caso da *regressão linear* é o quadrado do coeficiente de correlação de Pearson, ou seja, a versão amostra da eq. (149) é:

$$\hat{\eta}^2(X \mid S) = R^2(X = f(S) \text{ linear}) = \rho^2(X, S) = \frac{SS_{regressão}}{SS_{total}} = 1 - \frac{SS_{erro}}{SS_{total}} \quad (162)$$

No caso de regressão não-linear, $\hat{\eta}^2$ também pode ser visto como o coeficiente R^2 da amostra, que no caso não-linear é diferente de ρ^2. A versão amostra da eq. (148) é dada por:

$$\hat{\eta}^2(X \mid S) = R^2(X = f(S) \text{ não-linear}) = \frac{SS_{regressão}}{SS_{total}} \quad (163)$$

No caso de regressões múltiplas (vários S_k), a interação entre os fatores faz com que o valor de R^2 seja tendencioso para cima (como esperado, já que η^2 é tendencioso para cima, como foi visto). Essa tendenciosidade é maior, quanto menor for o tamanho da amostra N e maior for o número k de fatores interagindo. Nesse caso, se usa o R^2-ajustado, denotado por R^2-adj, que é dado pela equação (exs.: Miles & Shevlin, 2001, p.33; ou Maxwell & Camp & Arvey, 1981, p. 528):

[224] No entanto, como mostra Cohen (1973) e citado por Levine & Hullet (2002, p.620-621) como "notável exceção", o parcial-η^2 pode ser mais interessante que o η^2 sob certas condições.

$$\mathbf{R^2\text{-}adj = 1 - (1 - R^2)\ \frac{N - 1}{N - k - 1}} \qquad \mathbf{(164)}$$

É fácil ver que (dado k finito) quando $N \to \infty \Rightarrow R^2\text{-adj} \to R^2$. Quando η^2 é estimado dessa forma, essa estimativa é chamada muitas vezes de "epsilon-squared", $\hat{\varepsilon}^2$. Por ex., Maxwell & Camp & Arvey (1981, p. 527-528) e outros na literatura de psicologia, associam "epsilon-squared" ao R^2-adj (eq. 164) e "eta-squared" ao R^2 da eq. (163). Outros, como o artigo clássico de Kelley (1935), associam o uso da eq. (164) combinado com a eq. (163) à estimativa "não tendenciosa da razão de correlação". Maxwell & Camp & Arvey (1981, p. 532) concluem que "eta-squared" (como medido pela eq. 163) é um bom índice descritivo de grau de associação entre a v.a. "dependente" (X) e a v.a. "independente" (S) em uma amostra, enquanto que "epsilon-squared" (e uma outra chamada "omega-squared") são melhores em termos de inferência desse grau de associação. Mas eles mesmos indicam (p.526) que não existe consenso entre os autores sobre o melhor estimador para medida de associação, por ex., Friedman recomenda o uso do "eta-squared", Cohen recomenda ambos "eta-squared" e "epsilon-squared", dependendo da situação, etc[225].

Esses estudos de ANOVA ou de múltipla regressão consideram que o vetor de sinais são variáveis estatisticamente independentes[226]. E para o caso do vetor de sinais com variáveis dependentes entre si? McKay & Morrison & Upton (1999) analisam esse caso no contexto de *análise de sensibilidade global* (item 3.1.4.4), ou seja, o vetor de sinais são as variáveis de "input" e a variável de interesse é a variável de "output". Eles usam uma fórmula de variância total para variáveis dependentes conhecida como fórmula de Panjer (ver Panjer, 1973) para justificar o método de estimativa de η^2. Em seguida eles usam métodos computacionais iterativos para estimar o valor de η^2 da amostra com sinais dependentes, incluindo a conhecida amostragem "Latin hypercubic", muito usada em simulação de Monte Carlo (co-desenvolvida pelo próprio McKay no final da década de 70). O livro de

[225] Até na simples estimativa de uma variância não há consenso. Embora a maioria divida a soma de quadrados por N – 1, para o estimador ser *não tendencioso*, há quem prefira dividir pelo tamanho da amostra N, pois dessa forma o estimador terá *menor variância* (menor erro). São dois critérios razoáveis que levam a diferentes estimadores. Se N for grande, a discussão é irrelevante.

[226] Em estimativa Bayesiana, se assume que as v.a. são *condicionalmente* independentes ou *intercambiáveis*, relaxando um pouco a hipótese de independência.

Saltelli et al (2004), também discute métodos para estimar η^2 com sinais ("inputs") dependentes.

O objetivo desse item foi principalmente mostrar a facilidade de obter estimativas de η^2 com métodos populares como regressão e ANOVA. Mas foram apresentadas também algumas equações de estimativa assim como os conceitos (populacionais) por trás dessas equações. Diversas referências foram mostradas para o leitor interessado em trabalho empírico nessa área. Agora será apresentado um tema mais aplicado, de fundamental importância em exploração de petróleo.

3.4. Processo de Revelação de Bernoulli

3.4.1. Distribuição de Bernoulli e Fator de Chance Exploratório

A incerteza técnica mais elementar e básica em exploração de petróleo é a incerteza na existência de petróleo num prospecto exploratório. Essa incerteza é expressa pela variável denominada fator de chance (FC), que dá a probabilidade de existir petróleo em um prospecto[227]. O fator de chance exploratório é uma v.a. discreta definida por uma distribuição de Bernoulli com parâmetro p (onde p é chamado de probabilidade de sucesso), ou seja, em termos de notação:

$$\mathbf{FC \sim Be(p)} \qquad \mathbf{(165)}$$

Pela sua importância prática no contexto de exploração de petróleo, inclusive na aplicação do cap. 5 de jogo de opções reais, a distribuição de Bernoulli será estudada em detalhes, assim como a distribuição bivariada de Bernoulli, em que a v.a. fator de chance FC interage com a v.a. sinal S, que também é uma distribuição de Bernoulli. Ainda no item 3.4 serão estudados os processos de revelação de Bernoulli, que pertencem a uma categoria específica de processos dependentes de Bernoulli que ainda não foi estudada na literatura[228].

A distribuição de Bernoulli é a distribuição mais simples que existe: é uma distribuição discreta com apenas um parâmetro e apenas dois cenários. Um dos

[227] Essa parte é baseada parcioalmente em notas de aulas de cursos ministrados pelo autor na Petrobras e em Dias (2003).

[228] Ao melhor do conhecimento do autor da tese. O processo dependente de Bernoulli dado pela distribuição hipergeométrica (sorteios sem reposição) é totalmente diferente do tipo de dependência que será aqui analisado. Para outros detalhes, ver item 3.4.3.

cenários tem valor[229] 1 (com probabilidade p), chamado "sucesso", e o outro tem valor zero (com probabilidade 1 – p), chamado "falha" (ou insucesso). Nas aplicações essas duas possíveis realizações de FC são *fatores multiplicativos* do VPL na equação do valor monetário esperado (VME), eq. (3). A Figura 40 mostra a distribuição de Bernoulli, sendo que o lado esquerdo mostra a função probabilidade (massa) com parâmetro p = 30% e o lado direito mostra a função distribuição acumulada G(x) com parâmetro genérico p.

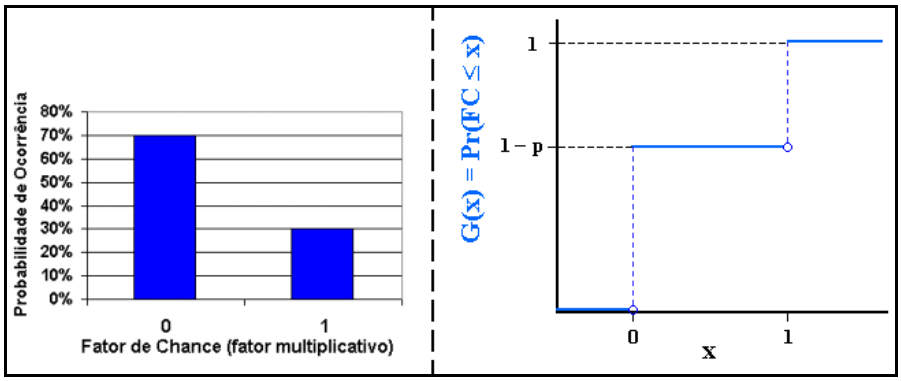

Figura 40 – Distribuição de Bernoulli: Massa (esquerda) e Acumulada (direita)

Formalmente, a função probabilidade (massa) de Bernoulli pode ser escrita como uma das duas equações abaixo:

$$\mathbf{Be(x)} = \begin{cases} \mathbf{1 - p} & \mathbf{para \quad x = 0} \\ \mathbf{p} & \mathbf{para \quad x = 1} \end{cases} \tag{166}$$

$$\mathbf{Be(x)} = \mathbf{p^x (1-p)^{1-x}} \quad \text{, sendo x = 0 ou 1} \tag{167}$$

A distribuição acumulada G(x) da Figura 40 é formalmente escrita como:

$$\mathbf{G(x)} = \begin{cases} \mathbf{0} & \mathbf{para \quad x < 0} \\ \mathbf{1 - p} & \mathbf{para \quad 0 \le x < 1} \\ \mathbf{1} & \mathbf{para \quad x \ge 1} \end{cases} \tag{168}$$

A distribuição de Bernoulli é também de grande importância teórica, tanto por razões históricas[230] como por poder gerar outras distribuições conhecidas tais como as distribuições binomial, hipergeométrica, Pascal, geométrica, e no limite até distribuições contínuas como a normal (através da binomial). Discussões sobre a distribuição de Bernoulli, assim como as suas propriedades podem ser

[229] Existem distribuições de Bernoulli com pequenas variações: (a) cenários nominais, ex: sucesso é x = sim e falha é x = não; (b) cenários numéricos mas não 0-1, ex: – 1 e + 1.

[230] Foi uma das primeiras distribuições estudadas. Além disso, a primeira versão da *lei dos grandes números*, o clássico teorema Bayesiano sobre seqüências de v.a. intercambiáveis de De Finetti, dentre outros, usaram a distribuição de Bernoulli, tomando vantagem de sua simplicidade.

encontradas especialmente em Balakrishnan & Nevzorov (2003, cap. 4) e alguma discussão também em Evans & Hastings & Peacock (2000, cap.4). Aqui serão mostradas algumas das propriedades da distribuição de Bernoulli. As propriedades mais importantes são o valor esperado e variância de FC ~ Be(p):

$$\mathbf{E[FC] = p} \tag{169}$$

$$\mathbf{Var[FC] = p\,(1 - p)} \tag{170}$$

A Tabela 10 resume as principais propriedades da distribuição de Bernoulli com parâmetro p.

Tabela 10 – Propriedades da Distribuição de Bernoulli (Univariada)

Propriedade/ Função	**Equação**
Média	$\mathbf{E[Be(p)] = p}$
Variância	$\mathrm{Var[Be(p)]} = p\,(1-p)$
Assimetria	$\gamma_3[\mathrm{Be(p)}] = \dfrac{1-2p}{\sqrt{p\,(1-p)}}$
Curtose	$\gamma_4[\mathrm{Be(p)}] = \dfrac{1}{p\,(1-p)} - 3$
Função Característica	$\phi(t) = 1 + p\,(e^{i\,t} - 1)$
Função Geradora de Momentos	$M(t) = (1-p) + p\,e^{t}$
Estimador de p (n sucessos na amostra N)	$\hat{p} = \dfrac{n}{N}$

Quando a medida de incerteza é a entropia H (eq. 68), então a entropia da distribuição de Bernoulli é dada por:

$$\mathbf{H[Be(p)] = -p\log(p) - (1-p)\log(1-p)} \tag{171}$$

Note que tanto a variância (eq. 170) como a entropia (eq. 171) são máximas quando p = ½ e mínimas (iguais a zero) para os casos de p = 0 e p = 1. Ou seja, em certos aspectos a variância e a entropia têm comportamentos similares no caso da distribuição de Bernoulli, o que pode levar a soluções similares.

Assim como ocorre com a distribuição Normal, a informação de Fisher I(p) para a distribuição de Bernoulli é simplesmente o inverso da variância (prova: ver DeGroot & Schervish, 2002, p.437). Ou seja, a informação de Fisher é:

$$\mathbf{I(p) = 1 / [p\,(1-p)]} \tag{172}$$

Assim a informação de Fisher é mínima para p = 50% (mas não é zero, é inclusive maior que 1) e vai a infinito nos casos de revelação total, i. é, $I(p) = +\infty$ para os casos de p = 0% e p = 100%. Assim, claramente a informação de Fisher não serve como medida de aprendizagem, pois não atende ao axiomas mais elementares das medidas de aprendizagem, ver item 3.3.2.

Para outras propriedades, ver Balakrishnan & Nevzorov (2003, cap. 4). Por ex., todos os momentos não centrais são iguais a p, i. é, $E[X^n] = p$, n = 1, 2, … A variância é igual a $E[X^1] - (E[X^2])^2$, etc.

No caso de p = 0 ou de p = 1, a distribuição de Bernoulli é dita *degenerada*. Nesse caso se tem a *revelação total* do verdadeiro cenário da distribuição de Bernoulli, pois se p = 0 então a variável de Bernoulli FC = 0 com probabilidade 1 e se p = 1 então FC = 1 com probabilidade 1. Como esse também é o caso limite de uma distribuição de revelações de uma v.a. de Bernoulli, ver Teorema 1(a), será a seguir discutido esse caso no contexto de exploração de petróleo.

O que significa *revelação total* para o fator de chance exploratório? Existe algum investimento em informação que revele toda a verdade sobre uma locação específica? Sim, a própria perfuração do poço pioneiro! Essa perfuração revelaria ou o cenário $FC^+ = 1$ ou o cenário $FC^- = 0$. Por consistência, ex-ante esses cenários revelados teriam probabilidade de ocorrência iguais a p e (1 – p), respectivamente. Ou seja, a distribuição de revelações no caso de revelação total é exatamente igual à distribuição a priori de FC. Isso é exatamente o Teorema 1 (a), que diz que a distribuição de revelações é igual a distribuição a priori, no caso limite de revelação total!

Outra maneira de ver é através de um sinal S que proporcione uma *aprendizagem máxima*. O caso limite de revelação total através de um sinal poderia ser idealizado através de um "expert infalível" (ou "profeta") que saberia *toda a verdade* sobre uma bacia. Isso significaria que toda vez que alguém apresentasse um prospecto de um bloco exploratório, esse "expert infalível" diria que ou a locação tem FC = 100% ou a locação tem FC = 0%. Se um geólogo estima que um certo prospecto tem uma probabilidade de sucesso p então, por consistência, ele espera que se o "expert infalível" for convocado a se pronunciar em relação a esse prospecto, ex-ante a probabilidade de ele(a) revelar que FC = 1 é igual a p e a probabilidade de ele(a) revelar que FC = 0 é (1 – p). Ou seja, a distribuição de revelações é exatamente a distribuição a priori, consistente com o

Teorema 1 (a). Se o geólogo achar que o "expert infalível" dirá FC = 1 com uma probabilidade diferente (ex.: q), então ele deve revisar sua distribuição a priori de FC e usar q como probabilidade de sucesso.

Note que a má notícia é também muito valiosa para uma companhia de petróleo, pois evita um gasto inútil em um poço "seco" (petróleo inexistente) com probabilidade 1, i. é, evita o exercício sub-ótimo da opção de perfurar um poço pioneiro naquela locação. Lembrar do exemplo simples do cap. 2 (Figura 3).

Bacias muito conhecidas (ex.: águas rasas do Golfo do México) tem locações ou com FC muito alto (> 40%) ou muito baixo. Embora a própria perfuração do poço pioneiro seja a única maneira conhecida de revelar toda a verdade sobre a existência de petróleo em um prospecto (pois infelizmente não existe o "expert infalível"), isso não seria vantagem já que se deseja obter uma informação mais barata sobre a chance de sucesso, antes de gastar algumas dezenas de milhões de dólares na perfuração do poço pioneiro. Pesquisa sísmica e os resultados das perfurações de poços vizinhos (inclusive de outras firmas – agindo como "*free rider*") proporcionam *revelação parcial* sobre FC que permite revisar as expectativas em relação ao FC de um prospecto específico. Isso será estudado no item 3.4.2 sobre distribuição bivariada de Bernoulli, pois é necessário descrever a distribuição conjunta do FC do prospecto de interesse com o sinal S, geralmente também uma v.a. de Bernoulli.

É oportuno estabelecer a seguinte definição sobre "play geológico", muito útil tanto para desenho da seqüência ótima de perfuração de prospectos, como para análise de interação estratégica entre companhias de petróleo atuando numa mesma bacia. Essa definição é baseada em Rose (2001, p. 3 e 57) e é similar, mas mais detalhada do que a usada em McCray (1975, p.222).

Definição. **Play geológico**: é uma família de campos, descobertas, prospectos e caminhos/regiões ("leads") que tem similaridade geológica, i. é, compartilham a mesma origem geológica. A similaridade se dá principalmente em termos de tipo de rocha reservatório, geometria da trapa (para *aprisionar* o petróleo no reservatório) e rocha geradora de petróleo.

A análise de play geológico (chamada de "play analysis") utiliza estudos de *geologia regional* (que estuda uma determinada área de uma bacia), de *geoquímica* (que dá indicações principalmente sobre migração de petróleo) e *geofísica* (especialmente a sísmica, que dá indicações sobre a existência e o

tamanho de rochas e falhas geológicas). Prospectos no mesmo play geológico são correlacionados, i. é, por compartilharem as mesmas características geológicas a informação resultante da perfuração de um prospecto afeta (para cima ou para baixo) os fatores de chance de todos os prospectos situados no mesmo play geológico. Por isso esse conceito é fundamental para a análise da interação estratégica entre companhias de petróleo em aplicações de jogos de opções reais.

Para aprofundar a questão de correlação entre prospectos é necessário detalhar o significado do fator de chance de existência de petróleo, decompondo-o em outros fatores. Desde o ponto de vista da exploração de petróleo[231], o fator de chance de existência de uma reserva de petróleo em um prospecto pode ser visto como uma função de 6 fatores de chance (que individualmente também são v.a. de Bernoulli). O FC total é o produto dos seis fatores abaixo, já que esses fatores geralmente são considerados independentes ou condicionalmente independentes:

1) Probabilidade de existência da *rocha geradora* ("source rock") de petróleo;
2) Probabilidade de existência de *migração*, i. é, uma falha geológica ou outro caminho ligando a rocha geradora com a rocha reservatório;
3) Probabilidade de existência de *rocha reservatório*;
4) Probabilidade de existência da *trapa geométrica*[232] ("closure chance") que dá a geometria da relação entre rocha reservatório e rocha selante;
5) Probabilidade de *retenção*, i. é, uma *rocha selante* cercando o reservatório e *preservação* do petróleo retido ("containment chance"); e
6) Probabilidade de existência do *sincronismo geológico* ("timing"), *condicional* à ocorrência dos fatores anteriores. Ou seja, coincidência temporal da seqüência geológica: geração, migração, enchimento do reservatório, aprisionamento e preservação do petróleo.

Se qualquer desses fatores não ocorrer, não haverá reserva de petróleo. Ou seja, a existência de petróleo é uma grande coincidência. A lista acima pode variar um pouco, a depender do autor. Uma das referências mais citadas por geólogos-economista é Rose (2001, p.34-36), que recomenda os 5 primeiros fatores de

[231] Agradeço ao Consultor Sênior da Petrobras, Geofísico Paulo Johann, pelas discussões sobre a decomposição do fator de chance e sobre a comparação entre as revelações de informações obtidas com um sinal sísmico e com o resultado da perfuração de um poço exploratório vizinho.

[232] Existem diversos tipos de trapas geológicas: estruturais, estratigráfica, diagenéticas, hidrodinâmicas e em centro de bacias. Ver Rose (2001, p.35).

chance da lista acima[233]. Em Rose (2001) a probabilidade de sincronismo é considerada dentro do fator de chance de migração principalmente. Mas aqui será interessante destacar o fator de sincronismo geológico, pois pode haver um caminho claro de migração (indicado pela sísmica, ex.: técnica de "ray tracing"), mas não ter havido sincronismo para alguns objetivos geológicos numa bacia e ter havido sincronismo em outras (o sexto fator pode facilitar a análise desses casos, para separar o que a sísmica não vê). Além disso, esse fator é importante para análise estratégica de prospectos vizinhos correlacionados (ver discussão abaixo).

Aliás, o próprio Rose (2001, p.80-82) destaca que na análise de play geológico é conveniente distinguir entre os fatores de chance que são compartilhados ("shared") por todos os prospectos do mesmo play, dos fatores de chance específicos do prospecto (fatores de chance locais ou independentes). Os fatores de chance compartilhados aqui serão referidos como fatores de chance globais ou regionais. Rose mostra como exemplos de fatores "shared", os fatores relativos à rocha geradora, migração em direção à área do play geológico e o sincronismo (!), fator esse que antes ele não tinha destacado da migração. Isso reforça a decisão dessa tese de adicionar o fator sincronismo à lista de Rose.

Agora serão discutidos os cinco primeiros fatores da lista acima. Conforme destaca Rose (2001, p.35), a probabilidade na geração de petróleo é muito alta em bacias produtivas (bacias relativamente conhecidas), mas pode ser baixa em bacias de fronteira (bacias pouco exploradas). Assim, esse é um fator global, pois afeta praticamente todos os prospectos de uma bacia. Em bacias de fronteira, mesmo um sucesso num poço distante afeta bastante o prospecto, pois confirma a existência de rochas geradoras na bacia. Mas se tiver muito distante dessas rochas geradoras, o petróleo pode não ter chegado até a área do prospecto.

O segundo fator, a migração, é um fator regional, pois pode afetar diversos prospectos, mas menos globalmente que o caso da rocha geradora. Por ex., pode ter havido migração em direção a plays geológicos da área norte de uma bacia, mas não na área sul da mesma. No caso de uma bacia bem explorada, com poços relativamente pertos (vizinhos) confirmando a existência de petróleo, esse índice pode ser bastante alto. Caso contrário, ele pode ser muito baixo. Esse fator é de grande relevância em prospectos vizinhos e assim pode levar a interação

[233] Mas, conforme reporta o Geólogo Luciano Costa, a Petrobras usa 6 fatores desde 1984.

estratégica de firmas diferente que possuem prospectos correlacionados (no mesmo play geológico) em áreas (blocos) vizinhos. Ou seja, a revelação de informação do resultado do poço vizinho afeta bastante esse fator de migração.

Os outros fatores 3, 4 e 5 são mais locais, mas um sucesso num poço vizinho correlacionado também pode melhorar um pouco a chance de ocorrência desses fatores. Ou seja, tem alguma dependência para prospectos do mesmo play, mas menos do que os fatores anteriores. Para esses fatores, os registros de sísmica dão uma boa indicação de sua ocorrência. Os registros sísmicos (especialmente a sísmica 3D) dão boas indicações sobre a existência de rochas e aspectos estruturais em geral, mas virtualmente nenhuma (ou muito fraca) indicação sobre os fluidos na rocha reservatório (especialmente se óleo ou água) ou sincronismo.

Em geral, a sísmica pode dar razoáveis ou boas indicações para os 5 primeiros fatores listados acima, mas não dá quase nenhuma indicação sobre o sexto fator (sincronismo). Já o resultado da perfuração de poços vizinhos correlacionados revela principalmente (e fortemente) o sexto fator. Visto dessa forma, o resultado de poços adjacentes complementa bem a sísmica. Entretanto, mesmo com ambas as fontes de informação alguma incerteza permanece no prospecto de interesse (apenas a sua própria perfuração revelaria toda a verdade sobre esse prospecto). A análise matemática de sinais imperfeitos (revelação parcial) será desenvolvida no item 3.4.2.

Embora tenham casos em que se reporta alguma dependência entre esses fatores (por ex., entre reservatório e geometria de trapa e entre rocha selante e geometria de trapa), a grande maioria dos analistas concorda em assumir a premissa da independência entre esses 5 fatores num mesmo prospecto (Rose, 2001, p.41). Mas pode ser mais preciso considerar como Delfiner (2000, p.2), que alguns fatores são *condicionalmente* independentes e aí sim eles podem ser multiplicados. Ele dá como exemplo que só há migração se houver geração e assim o fator migração deve ser estimado como Pr(migração | existe a rocha geradora), para poder multiplicar pelos outros fatores. Delfiner (2000) considera migração e sincronismo no mesmo fator, como Rose. Mas a dependência mais forte é com o sincronismo: só tem sentido falar em sincronismo de eventos caso esses eventos tenham ocorrido (sincronizados ou não). Por isso esse fator foi listado como condicional à ocorrência das rochas, geometrias, etc. A diferença em relação a Delfiner é que em alguns casos a sísmica pode indicar um *caminho/falha*

da possível migração (com o fator refletindo a chance dessa existência), mesmo que não exista a rocha geradora.

Mas em muitos casos, considerar os fatores condicionais pode ser mais interessante quando se decompõe os mesmos em *outros* fatores, onde alguns sub-fatores são comuns e outros são independentes (Delfiner, 2000). Dessa forma, quais os fatores destacar como principais e quais os sub-fatores que compõem cada fator principal, *depende da aplicação*. Aqui (análise econômica) vale a pena destacar o sincronismo, já que o poder de revelação da perfuração de poços correlacionados é bem maior que a da sísmica 3D para esse fator específico.

Assim, em geral se calcula o fator de chance total FC simplesmente multiplicando esses fatores de chance específicos. Aqui está sendo assumido que existe apenas um objetivo num prospecto. Em caso de múltiplos objetivos (múltiplas zonas de interesse sendo atravessadas pelo mesmo poço pioneiro), então deve ser considerada a dependência no sentido que a revelação de informação do resultado da perfuração no objetivo mais raso irá mudar a probabilidade de sucesso nos objetivos mais profundos do poço. Isso *pode* sugerir a interrupção da perfuração (exercício da opção de abandono), caso os objetivos mais rasos revelem notícias negativas.

Para determinar o fator de chance FC, cada um dos seis fatores deve ser determinado por estudos de especialistas[234]. Rose (2001, p.34-36) recomenda decompor cada um desses fatores principais em sub-fatores de chance, i. é, elementos mais básicos dentro de cada tópico, aprofundando a análise de cada item com estudos qualitativos e/ou quantitativos. No item 3.4.4 será mostrada a metodologia de *inferência Bayesiana* (assim como as suas limitações) para estimar a probabilidade de sucesso de uma distribuição de Bernoulli que, em tese, poderia ser aplicada para sub-fatores de um mesmo play geológico desde que existam dados. Pode-se usar a seguinte escala subjetiva de probabilidade, seja para fatores principais, seja para sub-fatores, que associa frases a cada uma das diversas faixas de valores de probabilidade de sucesso. A idéia é os especialistas se guiarem por essa escala para atribuir probabilidades para cada fator que irá

[234] Em alguns casos são usadas estimativas baseadas na taxa de sucesso média de uma bacia ou área. Mas elas são indicadores pobres do fator de chance específico de um prospecto. Assim, na maioria dos casos, é preferível fazer uma análise específica dos vários fatores (Rose, 2001, p.46).

compor o fator de chance total FC. Essa escala subjetiva de probabilidade é mostrada na Figura 41 (baseada em Rose, 2001, p.37).

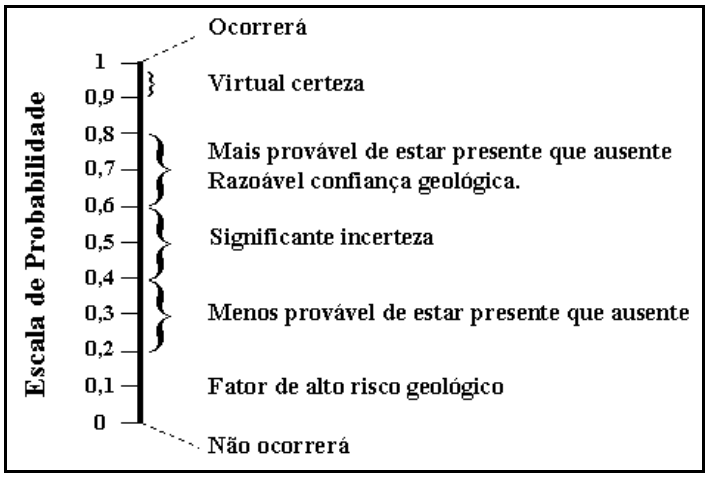

Figura 41 – Escala de Probabilidade Subjetiva para Fator de Chance

Na estimativa dessas probabilidades, o caso mais complicado (ou mais tendencioso) é a análise de prospectos de alto risco, i. é, quando FC ≤ 10%. Conforme Rose (2001, p.47), um resultado "perturbador" tem sido reportado por grandes companhias tais como a Shell, Amoco e Mobil nos anos 80, onde os prospectos com FC ≤ 10% (ex-ante) resultaram em sucessos ex-post em menos de 1% dos casos. Essa tendenciosidade (para cima) de avaliação de prospectos de alto risco é explicada de várias maneiras por Rose, duas delas são: (a) gerentes têm a tendência de querer "desesperadamente" achar grandes campos que esses prospectos podem conter[235], confundindo a necessidade de *adicionar valor* com a necessidade (?) de *perfurar poços*; e (b) em geral os geólogos (e as pessoas) são melhores em julgar a diferença entre FC de 50% e FC de 25% do que entre FCs de 5% e 10%. Essa última razão é consistente com a chamada *teoria dos prospectos* (ver item 6.3.4), que explica porque as pessoas investem em jogos de loterias que têm baixíssimas chances de sucesso e VME claramente negativo desde um ponto de vista "racional" (i. é, VME < 0 para os *verdadeiros* FCs desses jogos, que podem ser calculados com precisão no caso de jogos lotéricos).

Porém, parte dos prospectos de alto risco *pode* ter valores de opções reais elevados devido ao poder de revelação maior de um prospecto que tem como objetivo um novo play geológico e/ou usa um modelo geológico totalmente diferente do que vem sendo usado. A revelação de um novo play geológico

[235] Se o FC é baixo, o VME só será maior que zero se o volume B esperado for grande.

produtivo numa bacia tem efeito global ou regional, pois afeta positivamente inúmeros prospectos em diferentes graus. Assim, uma ou mais companhias de petróleo com prospectos naquela região tem o valor do seu portfólio alavancado. Um exemplo simples disso é apresentado em Smith & Thompson (2004, p.2-3 e 6n), de um portfólio com três prospectos dependentes (eles usaram 8 probabilidades condicionais, além das 3 probabilidades marginais)[236] e mesmos prêmios. O de menor chance de sucesso foi escolhido ser perfurado primeiro por ter maior grau de dependência com os outros (maior poder de revelação).

Prospectos de alto risco podem gerar um valor *agregado* de OR maior, mas a *distribuição* de valor dessa externalidade positiva (em caso de sucesso) pode ser desigual. Nesse caso, pode ser necessária a *análise estratégica* dos "jogadores" com maior benefício de uma revelação positiva num certo play ou bacia, a fim de estimar que firma deve perfurar primeiro os prospectos de alto risco.

E o comportamento dos FCs de uma bacia ao longo do tempo? Rose (2001, p.46) mostra um gráfico típico de uma área dos EUA (norte de Michigan) onde o FC realizado (ex-post) tem um grande crescimento durante um período de três anos e depois uma clara tendência de declínio (entremeados com breves períodos de crescimento). Jones & Smith (1983, Fig. 4) mostram um gráfico similar em outra área dos EUA (sul de Louisiana), onde o FC ex-post cresce durante três anos atingindo um valor de 80% de sucesso, seguindo-se um longo período de declínio. O período de grande crescimento em geral está associado à revelação de um novo play geológico produtivo numa bacia, fazendo revisar para cima não só os FCs *previstos* (ex-ante), como sendo confirmado *ex-post* com um aumento da taxa de descoberta nesse play. Esse é o grande foco de análises como as que serão efetuadas nessa tese, devido ao potencial de alavancagem com opções reais exploratórias. A fase de declínio ocorre depois que as estruturas maiores e mais evidentes nesse play já foram encontradas e assim os prospectos remanescentes têm tanto ex-ante como ex-post FCs menores, pois são menores e/ou mais difíceis de serem encontrados (Rose, 2001, p.47). Isso é razoavelmente consistente com o método chamado de *modelagem de processo de descoberta* (ver o item 3.4.3 e White, 1992, p.88, para referências e discussão), que assume que a existência de pequenos campos é mais freqüente do que a de grandes campos, sendo que estes

[236] Embora a distribuição Bernoulli trivariada só precise de 8 parâmetros para ser definida.

são geralmente descobertos antes do que os pequenos campos. Esse método também assume que as descobertas posteriores (menores) tendem a preencher os vazios da distribuição de volumes descobertos de uma distribuição de probabilidades conhecida (geralmente lognormal).

Um aspecto prático sobre o FC. Geralmente se considera como descoberta (FC = 1 ex-post) apenas se o volume da acumulação (ou da reserva B) for maior que um certo valor mínimo. Isso pode ser conveniente para se trabalhar com distribuições conhecidas (ex.: triangular) para o volume de reservas B, pois se incluir volumes muito pequenos a distribuição poderia ser bi-modal (Delfiner, 2000, Fig.7), dificultando a análise. O importante é a consistência entre o FC e a distribuição a ser usada para o volume. Mas colocar um volume mínimo *muito alto* (apenas os que são hoje econômicos) é inconveniente, pois esquece a natureza dinâmica dos preços do petróleo e da evolução da tecnologia (que pode viabilizar pequenas descobertas inicialmente taxadas de inviáveis). A indústria de petróleo está cheia de exemplos de campos considerados inviáveis que se tornaram atrativos[237]. Por isso é preferível considerar no FC apenas o sucesso geológico (de um volume não desprezível, por ex., a partir de 5 milhões de barris) e não o sucesso comercial (que poderia demandar volumes mínimos de algumas *centenas* de milhões de barris, no caso de campos marítimos).

Mas quem usa estatísticas de sucessos para estimar o FC de uma bacia ou para outros estudos, deve considerar que geralmente é reportado o sucesso comercial (na época) e não o sucesso geológico[238]. Assim, a taxa de sucesso exploratório *comercial*, ao contrário da *geológica*, deve ser correlacionada com o preço do petróleo. Isso é exatamente o que demonstra o conhecido artigo de Forbes & Zampelli (2000), que usa testes econométricos (regressão logística) para mostrar a correlação entre os FC reportados e o preço do petróleo, com dados de

[237] Um exemplo é o campo de Mars em águas profundas do Golfo do México, reportado em Chen & Conover & Kensinger (2001). O bloco foi ganho em leilão pela Shell em 1985, mas com a queda do preço do óleo e a falta de tecnologia a baixo custo só foi perfurado e descoberto em janeiro de 1989. Mas foi considerado inviável economicamente devido ao alto custo. Com o passar do tempo voltou a ser viável e os investimentos iniciaram no segundo semestre de 1993. Outro exemplo ocorreu no campo de Jubarte na Bacia do Espírito Santo, que no ano 2000 era considerado inviável economicamente. Investimentos em VOI e a melhora no preço do petróleo o tornaram viável e estimulou a exploração naquela área. Hoje, as reservas descobertas nesse play geológico (conjunto de campos chamado de "parque das baleias") superam os 2 billhões de barris.

[238] Para o American Petroleum Institute, um poço pioneiro é classificado como sucesso se houver a completação de um poço para produção nesse campo (Forbes & Zampelli, 2000, p.115).

1978 a 1995. Eles mostram que a tecnologia só teve impacto (positivo) nos FCs reportados no período após 1986 (particularmente nos anos 90 com o uso de sísmica 3D, como será visto), sendo que antes o aumento no sucesso reportado foi devido principalmente ao aumento do preço do petróleo[239].

3.4.2. Distribuição Bivariada de Bernoulli: Fator de Chance e Sinal

Esse tópico irá analisar matematicamente o modelo probabilístico para um importante caso prático: como um sinal S altera o fator de chance FC de um prospecto de interesse? O sinal S aqui é especialmente a informação gerada pela perfuração de um poço pioneiro vizinho ou de um registro sísmico. O exemplo discutido no cap. 2 (ver Fig. 3) dá uma idéia da utilidade prática de se aprofundar nessa análise. Lá foram assumidos valores revisados (com o sinal) do fator de chance de um prospecto, sem discuti-los. Isso será feito nesse item.

A abordagem aqui será baseada no uso da medida de aprendizagem η^2 (ou a sua raiz positiva η), que tornará esse cálculo mais simples e intuitivo do que o método tradicional Bayesiano, que usa probabilidades inversas (verossimilhança). A discussão dessas probabilidades inversas na árvore de decisão da Fig. 30 deu uma idéia do método Bayesiano e mostrou que é muito trabalhoso calcular e verificar a consistência dessas probabilidades inversas, além de não ser muito intuitivo. Além disso, a estimativa de probabilidades condicionais não permite comparar diferentes sinais em termos de aprendizagem ou *poder de revelação*. Isso é importante até para a gerência da exploração ter uma métrica que possa comparar o efeito de uma sísmica 3D sobre diferentes prospectos de uma mesma área. Intuitivamente, o poder de revelação da sísmica 3D sobre esses prospectos deve ser mais ou menos o mesmo, mesmo que os FC sejam diferentes. Ou seja, a informatividade deve independer da distribuição a priori da v.a. de interesse, ver discussão sobre a comparação de experimentos (Blackwell). Uma medida de aprendizagem como η^2 resolve o problema de comparação, já as diferentes probabilidades inversas são difíceis de comparar. As métricas de verossimilhança não resolvem, pois foi visto que uma confiabilidade igual a zero, $p(s \mid x) = 0$, pode significar um aprendizado máximo igual ao obtido com $p(s \mid x) = 100\%$.

[239] Com o colapso ("jump-down") do preço do petróleo em 1986, o percentual de campos com menos de 2 milhões de barris caiu de 30% para 11% (Forbes & Zampelli, 2000, p.112 n2).

Há muito tempo que a literatura de exploração de petróleo reconhece o fato probabilístico que a revelação de um sinal positivo (ex.: descoberta feita em um bloco vizinho) aumenta a chance de sucesso de encontrar petróleo em prospectos correlacionados e vice-versa (sinal negativo diminui o FC). Isso motivou estudos de comportamento estratégico de firmas numa bacia (exs.: Hendricks & Kovenock, 1989; Hendricks & Porter, 1996), problema a ser analisado no cap.5.

A análise da relação entre as v.a. de Bernoulli FC e S é importante não só em aplicações de jogos de opções reais, como também no caso *sem* interação estratégica da análise de portfólio de exploração de uma companhia de petróleo, a fim de *priorizar* e determinar a *seqüência ótima* de investimentos em informação, i. é, em sísmica 3D e perfuração de poços pioneiros. Dentro da visão da tese, deve-se fazer uma análise integrada por grupos de prospectos correlacionados.

Para revisar o valor do fator de chance, em vez de usar diretamente uma medida de aprendizagem como nessa tese, geralmente são usados os métodos tradicionais Bayesianos (nem mesmo é usado o popular coeficiente de correlação ρ) que em geral demandam uma posterior verificação de *consistência* e/ou cálculo preliminar de probabilidades inversas, além de ter de estimar valores para algumas probabilidades condicionais. Isso é ilustrado no livro texto de Lerche & MacKay (1999, p.287-291) através de um exemplo para calcular o FC após um ou mais sinais, que é bem mais trabalhoso do que o método aqui proposto. Aqui só será necessário estimar a medida de aprendizagem e as probabilidades a priori do FC e do sinal (que ainda poderá ser simplificado se assumir v.a. intercambiáveis). Além disso, se for fácil estimar apenas *uma* dessas probabilidades condicionais, por ex., o valor de $FC^+ = Pr(FC = 1 \mid Sinal = 1)$, então poderá ser diretamente obtida (uma fórmula simples) a medida de aprendizagem η^2. Isso tudo será visto nesse tópico.

O artigo de Wang et al (2000) usa métodos de consistência, mais simples do que o Bayesiano tradicional, para estimar os FC revisados com a revelação do sinal. Além disso, eles também se preocuparam em introduzir uma medida de dependência para estabelecer um processo de revelação (não foi usado esse nome) para um fator de chance. Por isso, será dada uma atenção especial a esse artigo.

Wang et al (2000) estabelecem o valor do fator de chance condicional a um sinal positivo, FC^+, e o valor de FC condicional a um sinal negativo, FC^-, e depois verificam se esses valores são consistentes com o valor inicial. Nesse teste de

consistência, é usado o que eles chamam de "conservação de risco" (nome pouco adequado), que nada mais é do que a *lei das expectativas iteradas* aqui mostrada no Teorema 1 (b) (ou eq. 93). Eles não mencionam o nome mais conhecido, mas a aplicam corretamente. Wang et al (2000, p.3) também assumem outra premissa, embora de forma implícita, i. é, sem mencioná-la: que o fator de chance e o sinal têm a mesma probabilidade de sucesso, i. é, são v.a. intercambiáveis. Isso é um indicador que a premissa de v.a. intercambiáveis é considerada intuitiva pela indústria. Aqui será estudado tanto esse caso (intercambiável) como casos em que isso não é assumido. Enquanto a lei das expectativas iteradas é *necessária* para a consistência probabilística, a premissa de v.a. intercambiáveis é razoável mas *não* é necessária, a não ser em casos extremos que serão vistos[240]. No exemplo de Wang et al (2000, p.3) um o fator de chance inicial (incondicional) $FC_0 = 60\%$ e um sinal pode fazer ele subir para $FC^+ = 76\%$ ou $FC^- = 36\%$. Será visto que isso implica que o sinal também é uma v.a. Be(60%). Se o sinal fosse Be(50%), então se $FC^+ = 76\%$ seria consistente ter $FC^- = 44\%$ (pois 0,6 = 0,5 x 0,76 + 0,5 x 0,44).

Mas Wang et al (2000, p.4) sentem a necessidade de estabelecer alguma medida de dependência para determinar ou o cenário FC^+, ou o cenário FC^-. No caso eles escolhem esse último e criam uma medida de dependência que varia de 0 (independência) a 5 (total dependência) para calcular FC^-. Essa medida entra numa equação simples para o caso de n sinais:

$$\mathbf{(FC^-)^n = FC_0\,(1 - dep/5)^n} \qquad \mathbf{(173)}$$

Ou seja, em caso de n sinais negativos, o fator de chance inicial decai do fator $(1 - dep/5)^n$. Ele não se altera em caso de independência e vai a zero se um sinal negativo for totalmente dependente em relação ao FC. Essa é uma tentativa interessante que vai ser analisado no item 3.4.3, onde será visto que essa equação induz um processo de revelação convergente, recombinante, além de intercambiável, mas com algum problema de consistência do modelo de sinais.

Outra tentativa de modelar a revelação de informação por meio de sinal correlacionado com o FC foi Delfiner (2000). Ele modela a dependência através de uma decomposição de cada fator que compõe o FC total. Essa decomposição é tal que os sub-fatores ou são independentes ou são comuns (totalmente dependentes) e dessa forma se estabelece a estrutura de dependência. A crítica é

[240] Um sinal S só consegue reduzir 100% da variância do FC se for intercambiável com FC.

que nem sempre se pode decompor em fatores extremos como ele faz. Muitas vezes o fator é correlacionado, mas não totalmente dependente. Outra crítica ao artigo de Delfiner foi feita por Smith & Thompson (2004, p. 6), pois ele ao invés de enxergar a oportunidade (valor da flexibilidade de alterar as decisões subseqüentes), conclui apenas que a "*dependência aumenta o risco exploratório*".

Para avaliar o efeito de um *sinal* binário[241] S correlacionado (informação relevante) no fator de chance exploratório FC, é necessário estudar a relação de dependência entre duas distribuições de Bernoulli: o sinal binário S e o FC do prospecto de interesse. A maneira geral de estudar essa dependência é através da especificação da *distribuição de probabilidade conjunta* entre sinal e fator de chance. No caso, a distribuição bivariada de Bernoulli, será estudada a seguir. Essa seção será importante para entender a construção de processos de revelação de Bernoulli, uma classe *especial* de experimentos *dependentes* de Bernoulli.

Uma distribuição multivariada de Bernoulli com k distribuições marginais de Bernoulli, tem $2^k - 1$ parâmetros (Marshall & Olkin, 1985). No caso de interesse da tese, k = 2, e assim a distribuição bivariada de Bernoulli é totalmente definida com três parâmetros, que serão discutidos a seguir. Distribuições trivariadas de Bernoulli ou com k > 2 em geral, não são do interesse da tese, pois a evolução do fator de chance (processo de revelação de Bernoulli) se dará com seqüências de distribuições bivariadas de Bernoulli.

Os três parâmetros usados para definir a distribuição bivariada de Bernoulli são os dois parâmetros das distribuições marginais ou incondicionais (probabilidades de sucesso p e q) e um terceiro parâmetro, que estabelecerá a correlação entre as distribuições marginais. Esse último pode ser, por ex., a probabilidade de sucesso conjunto $p_{11} = Pr(X = 1 \textbf{ e } S = 1)$. Posteriormente, p_{11} será substituído por η. A Tabela 11 apresenta a distribuição bivariada de Bernoulli, assim como as suas distribuições univariadas marginais.

[241] Exs.: poço vizinho pode revelar sucesso (encontrou petróleo) ou insucesso; sísmica 3D pode indicar sucesso (indicação de presença de rocha selante) ou insucesso.

Tabela 11 – Distribuição Bivariada de Bernoulli e Distribuições Marginais

		Sinal S (ex.: sísmica)		Distribuição Marginal de X (FC)
		S = 1	S = 0	
Variável X (ex.: fator de chance)	X = 1	p_{11}	p_{10}	p
	X = 0	p_{01}	p_{00}	1 – p
Distribuição Marginal de S		q	1 – q	

Conforme Spanos (1999, p.288), a densidade (massa) de probabilidade bivariada de Bernoulli pode ser escrita analiticamente como:

$$\mathbf{p(x\,,\,s) \;=\; (p_{00})^{(1-s)\,(1-x)}\,(p_{01})^{(1-s)\;x}\,(p_{10})^{s\,(1-x)}\,(p_{11})^{s\;x}}\;,\; x \text{ e } s = 0,\,1 \qquad \mathbf{(174)}$$

Por conveniência de notação, especialmente no item sobre processos de revelação de Bernoulli, em vez de trabalhar com p será adotada a notação FC_0 para a probabilidade de sucesso do FC inicial (antes da informação), i. é:

$$\mathbf{FC_0 = p} \qquad \mathbf{(175)}$$

A distribuição de revelações (distribuição de expectativas condicionais) nesse caso tem dois cenários, por conveniência chamados de FC^+ e FC^-:

$$\mathbf{FC^+ = \Pr[FC = 1 \mid S = 1] \;=\; E[FC \mid S = 1]} \qquad \mathbf{(176)}$$

$$\mathbf{FC^- = \Pr[FC = 1 \mid S = 0] \;=\; E[FC \mid S = 0]} \qquad \mathbf{(177)}$$

Assim, FC_0 *evolui* para FC^+ ou FC^-, a depender do sinal S. Esses cenários da distribuição de revelações têm probabilidades de ocorrência de q e (1 – q), respectivamente. A Figura 42 mostra a distribuição de revelações para um sinal, sendo à esquerda como um processo de evolução dependente de um sinal e à direita a função probabilidade (massa) da distribuição de revelações de Bernoulli.

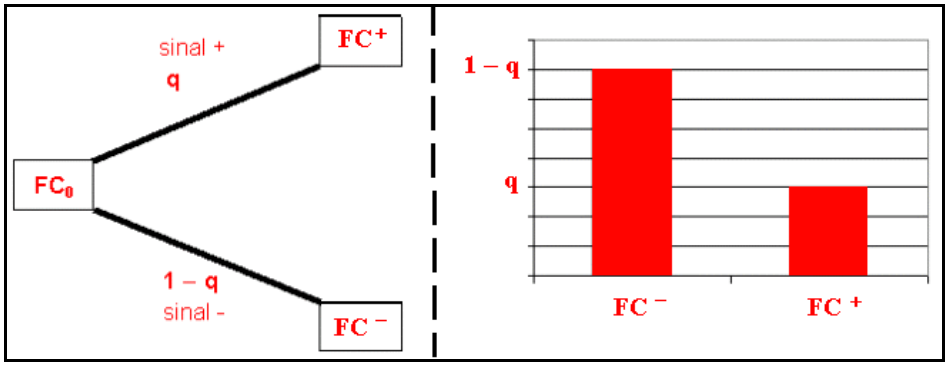

Figura 42 – Distribuição de Revelações de Bernoulli com Um Sinal

A variância da distribuição de revelações $Var[R_{FC}(S)]$ é, por definição de variância, igual a:

$$\mathbf{Var[R_{FC}(S)] = q\,(FC^{+} - FC_0)^2 + (1 - q)\,(FC^{-} - FC_0)^2} \quad \mathbf{(178)}$$

A variância da distribuição de revelações também pode ser escrita em função da medida de aprendizagem η^2 usando a eq. (107), lembrando que variância da distribuição a priori é $Var[FC] = FC_0\,(1 - FC_0)$. Logo,

$$\mathbf{Var[R_{FC}(S)] = \eta^2(FC \mid S)\;\; FC_0\,(1 - FC_0)} \quad \mathbf{(179)}$$

A teoria elementar de probabilidade sobre distribuição marginais e conjuntas, permite escrever as seguintes equações para as probabilidades de sucesso marginais (ver Tabela 11), as quais serão usadas em demonstrações:

$$\mathbf{FC_0 = p = p_{11} + p_{10}} \quad \mathbf{(180)}$$

$$\mathbf{q = p_{11} + p_{01}} \quad \mathbf{(181)}$$

Para a distribuição conjunta p(x, s) ser uma distribuição de probabilidade, os seus quatro cenários tem de ter probabilidades somando 1, ou seja:

$$\mathbf{p_{11} + p_{10} + p_{01} + p_{00} = 1} \quad \mathbf{(182)}$$

Os valores dos cenários da distribuição de revelações (eqs. 176 e 177) podem ser escritos em termos das variáveis mais básicas FC_0, q e p_{11}. Para isso, basta usar a definição de probabilidade condicional, $P(A \mid B) = P(A \cap B) / P(B)$:

$$\mathbf{FC^{+} = \frac{p_{11}}{q}} \quad \mathbf{(183)}$$

$$\mathbf{FC^{-} = \frac{p_{10}}{1 - q} = \frac{FC_0 - p_{11}}{1 - q}} \quad \mathbf{(184)}$$

Onde foi usada a eq. (180) na eq. (184). Combinando as eqs. (180) e (181) na eq. (182), pode-se tirar o valor da probabilidade p_{00} em termos das variáveis mais básicas FC_0, q e p_{11}, o que será útil em seguida:

$$\mathbf{p_{00} = 1 + p_{11} - FC_0 - q} \quad \mathbf{(185)}$$

Agora serão apresentadas as equações de confiabilidade (verossimilhança) da informação S, chamadas de L(S = 1) e L(S = 0). Como antes, serão usadas ambas, a definição de probabilidade condicional e a Tabela 11:

$$\mathbf{L(S = 1) = Pr(S = 1 \mid FC = 1) = \frac{p_{11}}{FC_0}} \quad \mathbf{(186)}$$

$$\mathbf{L(S = 0) = Pr(S = 0 \mid FC = 0) = \frac{p_{00}}{1 - FC_0} = \frac{1 + p_{11} - FC_0 - q}{1 - FC_0}} \quad \mathbf{(187)}$$

Onde foi usada a eq. (185) na eq. (187), a fim de expressá-la em termos das variáveis mais básicas FC_0, q e p_{11}.

Agora será estabelecida a condição de independência entre as v.a. de Bernoulli FC e S. Da literatura elementar de probabilidade (ex. DeGroot & Schervish, 2002, p.56), sabe-se que FC e S são independentes se e somente se ocorrer P(A | B) = P(A) e P(B | A) = P(B). Como P(A | B) = P(A ∩ B) / P(B), é fácil provar (ou ver Kocherlakota & Kocherlakota, 1992, p.57) que FC e S são v.a. independentes de Bernoulli se e somente se a probabilidade conjunta de sucesso Pr(FC∩S) = p_{11} for igual ao produto das probabilidades marginais de sucesso, i.é:

$$\textbf{FC e S independentes} \Leftrightarrow p_{11} = FC_0\, q \qquad (188)$$

A covariância entre FC e S (ver eq. 111) pode ser facilmente deduzida (ou ver Kocherlakota & Kocherlakota, 1992, p.57) para essas v.a. de Bernoulli como:

$$\textbf{Cov(FC, S)} = p_{11} - FC_0\, q \qquad (189)$$

Como Cov(FC, S) = 0 se e somente se $p_{11} = FC_0\, q$, então no caso de distribuição bivariada de Bernoulli, covariância igual a zero implica em independência, o que não ocorre em casos mais gerais de outras distribuições.

O coeficiente de correlação ρ(FC, S), que é a covariância normalizada, para o caso de v.a. de Bernoulli, é dada por (usar eq. 112 ou ver Kocherlakota & Kocherlakota, 1992, p.57):

$$\rho(FC, S) = \frac{p_{11} - FC_0\, q}{\sqrt{FC_0\,(1 - FC_0)\, q\,(1 - q)}} \qquad (190)$$

De forma análoga à covariância, para a distribuição bivariada de Bernoulli, ao contrário da regra geral, v.a. não-correlacionadas implicam em v.a. independentes (prova: eqs. 188 e 190):

$$\rho(FC, S) = 0 \Leftrightarrow \textbf{FC e S independentes} \qquad (191)$$

Assim, dependência implica em |ρ(X, S)| > 0 e os termos "correlação" e "dependência" se equivalem no caso de distribuição bivariada de Bernoulli. Por isso é correto dizer "prospectos correlacionados" para expressar dependência.

Lema 6 (sinal da correlação na distribuição bivariada de Bernoulli): Sejam as v.a. não-triviais FC ~ Be(FC_0) e S ~ Be(q). Então os casos não-correlacionados, com correlação positiva e com correlação negativa são, respectivamente:

$$\rho(FC, S) = 0 \Leftrightarrow p_{11} = FC_0\, q \qquad (192)$$

$$\rho(FC, S) > 0 \Leftrightarrow p_{11} > FC_0\, q \qquad (193)$$

$$\rho(FC, S) < 0 \Leftrightarrow p_{11} < FC_0\, q \qquad (194)$$

Prova: Direto da eq. (190).

Em adição, a menos que seja especificado em contrário, o interesse das aplicações dessa tese é para o caso de correlação positiva entre FC e S, de forma que FC^+ não será menor que FC_0 e FC^- não será maior que FC_0. Ou seja:

$$\mathbf{FC^+ \geq FC_0 \quad e \quad FC^- \leq FC_0} \tag{195}$$

O caso da correlação negativa é um problema análogo ao caso aqui analisado, com a notação trocada, já que FC^+ passaria a ser menor do que FC_0, etc. As raízes da medida de aprendizagem η^2 são flexíveis para indicar correlação positiva (+ η) ou negativa (– η). Mas em geral o que importa é o poder de revelação do sinal para revisar os valores de FC, e não o sentido (positivo ou negativo) da correlação, já que ambos os sentidos geram ganhos de aprendizagem.

Os limites de Fréchet-Hoeffding (ver item 3.3.1.1 e eqs. 101 a 103) para a probabilidade de duplo sucesso p_{11} e para o coeficiente de correlação são dadas pelas equações abaixo (prova: Joe, 1997, p.210):

$$\mathbf{Max\{0, FC_0 + q - 1\} \leq p_{11} \leq Min\{ FC_0, q\}} \tag{196}$$

$$\mathbf{Max}\left\{-\sqrt{\frac{FC_0\ q}{(1-FC_0)\ (1-q)}}\ ,\ -\sqrt{\frac{(1-FC_0)\ (1-q)}{FC_0\ q}}\right\} \leq \rho \leq \sqrt{\frac{Min\{FC_0\ ,\ q\}\ (1-Max\{FC_0\ ,\ q\})}{Max\{FC_0\ ,\ q\}\ (1-Min\{FC_0\ ,\ q\})}} \tag{197}$$

A prova da eq. (196) é simples tanto pelas eqs. (101) a (103) como observando a Tabela 11 e notando que as probabilidades tem de ser não-negativas (ex.: $p_{00} \geq 0$ na eq. 185). Note que seria válida a eq. (196) também para p_{00} no lugar de p_{11}. Já a prova da eq. (197) precisa de colocar p_{11} em função de ρ e substituir na eq. (196), além de álgebra cuidadosa.

Note na eq. (197) que é sempre possível o caso de $\rho = 0$, que aqui significa independência, pois ρ fica entre um número positivo e outro negativo. Agora será apresentado o Teorema 4 para a medida de aprendizagem η^2.

Teorema 4 (medida de aprendizagem η^2 para distribuição bivariada de Bernoulli): Sejam as distribuições marginais FC ~ Be(FC_0) e S ~ Be(q) não-triviais de uma distribuição bivariada de Bernoulli e seja a medida de aprendizagem η^2 definida pela eq. (106) ou pela eq.(107), então:

(a) As probabilidades de sucesso reveladas por S, i. é, FC^+ e FC^-, em caso de correlação *não*-negativa, são:

$$\mathbf{FC^+ = FC_0 + \sqrt{\frac{1-q}{q}}\ \sqrt{FC_0\ (1-FC_0)}\ \sqrt{\eta^2(FC\mid S)}} \qquad \mathbf{(198)}$$

$$\mathbf{FC^- = FC_0 - \sqrt{\frac{q}{1-q}}\ \sqrt{FC_0\ (1-FC_0)}\ \sqrt{\eta^2(FC\mid S)}} \qquad \mathbf{(199)}$$

Em caso de correlação não-positiva, vale as eqs. (198) e (199), mas com o sinal após FC_0 invertido.

(b) A medida de aprendizagem η^2 é igual ao quadrado do coeficiente de correlação ρ:

$$\mathbf{\eta^2(FC\mid S) = \rho^2(FC, S) = \frac{(p_{11} - FC_0\ q)^2}{FC_0\ (1-FC_0)\ q\ (1-q)}} \qquad \mathbf{(200)}$$

(c) A medida de aprendizagem η^2 é simétrica:

$$\mathbf{X \text{ e } S \sim \text{Bernoulli} \Rightarrow \eta^2(FC\mid S) = \eta^2(S\mid FC)} \qquad \mathbf{(201)}$$

(d) A medida η^2 é igual a zero se e somente se FC e S são independentes:

$$\mathbf{\eta^2(FC\mid S) = 0 \Leftrightarrow FC \text{ e } S \text{ independentes}} \qquad \mathbf{(202)}$$

(e) Os limites de Fréchet-Hoeffding, expressos em termos de η^2, para existência da distribuição bivariada de Bernoulli, sendo possíveis aprendizados tanto com correlação positiva como com negativa, são:

$$\mathbf{0 \le \eta^2 \le Max\left\{ Max\left\{ \frac{FC_0\ q}{(1-FC_0)\ (1-q)}\ ,\ \frac{(1-FC_0)\ (1-q)}{FC_0\ q} \right\},\ \frac{Min\{FC_0\ , q\}\ (1-Max\{FC_0\ , q\})}{Max\{FC_0\ , q\}\ (1-Min\{FC_0\ , q\})} \right\}} \qquad \mathbf{(203)}$$

Prova: (a) Pelo Teorema 1 (b), a média da distribuição de revelações é igual à média da distribuição a priori, que no caso é FC_0. Logo:

$FC_0 = q\ FC^+ + (1-q)\ FC^- \Rightarrow$

$$\mathbf{FC^- = \frac{FC_0 - q\ FC^+}{1-q}} \qquad \mathbf{(204)}$$

Substituindo na eq. (178) que dá a variância da distribuição de revelações:

$$\text{Var}[R_{FC}(S)] = q\ (FC^+ - FC_0)^2 + (1-q)\left(\frac{FC_0 - q\ FC^+}{1-q} - FC_0\right)^2$$

Após alguma álgebra, a equação acima se reduz para:

$$\mathbf{Var[R_{FC}(S)]} = \frac{\mathbf{q}}{\mathbf{1-q}} \mathbf{(FC^{+} - FC_0)^2} \quad \mathbf{(205)}$$

Mas a variância da distribuição de revelações, baseada no Teorema 1 (c), também pode ser escrita como na eq. (179). Combinando as eqs. (179) e (205):

$$\eta^2(FC \mid S) \;\; FC_0\,(1 - FC_0) \;\; = \frac{q}{1-q}\,(FC^{+} - FC_0)^2$$

Tirando o valor de FC^{+}, existem duas possibilidades (a depender do sinal ±):

$$\mathbf{FC^{+} = FC_0 \pm \sqrt{\frac{1-q}{q}}\,\sqrt{FC_0\,(1-FC_0)}\,\sqrt{\eta^2(FC \mid S)}} \quad \mathbf{(206)}$$

Como o segundo termo da eq. (206) é sempre positivo ou zero, a condição de correlação positiva (ou melhor, não-negativa) é dada pela eq. (195) e assim o sinal da eq. (206) é positivo em caso de correlação positiva, e é negativo em caso de correlação negativa. A prova da eq. (199) é análoga a essa prova. □

(b) Substituindo a eq. (206) na eq. (183) e elevando ao quadrado para explicitar η^2 (note que tanto faz usar o sinal positivo ou negativo da eq. 206), se obtém a eq. (200). □

(c) Pode-se provar através dos mesmos passos usados para obter a eq. (200) mas partindo de $\eta^2(S \mid FC)$ em vez de $\eta^2(FC \mid S)$. Mas bem mais simples é notar que se permutar as v.a. q e FC_0 na eq. (200), a eq.(200) permanece exatamente a mesma. Outra maneira é, já que foi provado que para v.a. de Bernoulli η^2 é igual a ρ^2, e esse último é simétrico, então também será simétrico η^2. □

(d) Basta usar a eq.(200) na eq. (191). Ou seja, o que vale para ρ^2, vale para η^2 no caso de distribuição bivariada de Bernoulli. □

(e) Primeiro note que o limite inferior de η^2 foi dado pela Proposição 6 (c), eq. (127). Ele é sempre viável pois no comentário à eq. (127) foi visto que o caso de independência é sempre viável (é sempre possível construir a distribuição bivariada de Bernoulli). Assim, apenas o limite superior é que *pode* não ser igual ao seu máximo, i. é, 1. Para deduzi-lo, um caminho é tirar a probabilidade p_{11} em função de η^2, usando a eq. (206) e substituindo na eq. (183):

$$\mathbf{p_{11} = FC_0\,q \pm \sqrt{FC_0\,(1-FC_0)}\,\sqrt{q\,(1-q)}\,\sqrt{\eta^2(FC \mid S)}} \quad \mathbf{(207)}$$

A seguir, substituir a eq. (207) na eq. (196) e após uma álgebra cuidadosa se obtém o lado direito da eq. (203). Note que a eq. (207) tem de ser substituída para os dois casos de sinais. Como são possíveis aprendizados tanto com correlação positiva como com negativa, vale o máximo dos dois casos. Se fosse permitido aprendizado apenas com correlação positiva, na eq. (203), em vez do máximo externo, teria de verificar qual das suas duas parcelas tem correlação positiva, que passaria a ser o limite *restrito* de Fréchet-Hoeffding para η^2.

Outra maneira é usar diretamente a eq. (197), notando que foi provado pela eq. (200) que η^2 é o quadrado de ρ para v.a. de Bernoulli[242]. □

A eq. (207) serve também para substituir a variável básica p_{11} por η^2, que tem um significado mais intuitivo de redução de incerteza, além das outras vantagens já discutidas. O limite superior de Fréchet-Hoeffding para η^2 da eq. (203) muitas vezes é igual a 1, o seu limite natural. Isso será visto em breve.

As eqs. (198) e (199) são não-lineares em η^2, mas <u>são lineares em η</u>, a raiz positiva. As eqs. (198) e (199) também podem ser vistas do ponto de vista da variância da distribuição de revelações, eq. (179), ou seja:

$$\mathbf{FC^{+} = FC_0 + \sqrt{\frac{1-q}{q}}\ \sqrt{Var[R_{FC}(S)]}} \qquad \mathbf{(208)}$$

$$\mathbf{FC^{-} = FC_0 - \sqrt{\frac{q}{1-q}}\ \sqrt{Var[R_{FC}(S)]}} \qquad \mathbf{(209)}$$

As eqs. (198) e (199) serão usadas em aplicações, mas com uma premissa a ser vista, elas ficarão ainda mais simples. Isso é discutido a seguir.

Um caso particular importante é quando as v.a. de Bernoulli FC e S são intercambiáveis. Nesse tipo de problema (fator de chance exploratório), a intercambialidade é considerada intuitiva, como por exemplo foi assumido de forma implícita em Wang et al (2000), comentado no item 3.4.1.

<u>Definição</u>. **Variáveis aleatórias intercambiáveis**: Duas ou mais v.a. são ditas intercambiáveis ("*exchangeable*" ou "*interchangeable*" em inglês) se sua distribuição conjunta é a mesma não importa a ordem em que as variáveis são observadas (Ross, 1998, p.288). Ou seja, as v.a. discretas $X_1, X_2, \ldots, X_n$ são

[242] A planilha joint-dist_Bernoulli.xls do CD-Rom permite verificar experimentalmente a eq. (203), i. é, os limites de Fréchet-Hoeffding, e também as condições de correlações positivas.

intercambiáveis se para toda permutação $i_1, \ldots, i_n$ dos inteiros 1, …, n, a igualdade abaixo permanece válida:

$$\mathbf{Pr\{Xi_1 = x_1, Xi_2 = x_2, \ldots, Xi_n = x_n\} = Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\}} \quad \mathbf{(210)}$$

Variáveis intercambiáveis, também chamadas de v.a. *simetricamente dependentes* (Rotar, 1997, p.231), ocorrem naturalmente em muitas situações práticas de interesse e por isso tem sido focado na literatura de distribuição multivariada e principalmente na estatística Bayesiana. Alguns exemplos: Fréchet (1943, p.107-121), O'Hagan (1994, p. 112-118, 156 e 290), Joe (1997, p. 211, 216-217, 222, 227-229), Jaynes (2003, p. 62, 620), Rotar (1997, p.231-232). O conhecido *modelo de urna de Pólya* é outro exemplo de modelo com variáveis intercambiáveis (Ross, 1998, p.289-290). O sorteio sem reposição de bolas de uma urna finita faz os sorteios seqüenciais não serem independentes, mas os sorteios são condicionalmente independentes, i. é, são intercambiáveis (nesse caso sem reposição, o número de sucessos é dado pela distribuição hipergeométrica).

O conceito de *intercambialidade* é atribuído a De Finetti, com um famoso teorema publicado em 1937 justamente para seqüências de Bernoulli. Na visão Bayesiana, esse é um dos mais importantes conceitos para modelos probabilísticos pois "*evita conceitos fortes de independência ... justifica a adoção de modelos paramétricos com base em condições fracas de simetria*" (French & Insua, 2000, p.70). Além disso, "*...provê uma ligação entre as idéias frequentistas e Bayesiana ...sugere que os limites de freqüências são probabilidades condicionais, dada a informação ainda não disponível*" (French & Insua, 2000, p.71).

No caso da distribuição bivariada de Bernoulli, as variáveis "sinal" e "fator de chance" são intercambiáveis se e somente se as suas probabilidades de sucesso incondicionais (marginais) são iguais. Isso é expresso na proposição abaixo.

Proposição 7: Sejam FC e S v.a. não triviais de Bernoulli intercambiáveis e a distribuição bivariada de Bernoulli definida por suas probabilidades de sucesso FC_0 e q e pela medida de aprendizagem η^2. Então:

(a) As distribuições marginais são iguais. Também, se as distribuições marginais são iguais, então as v.a. são intercambiáveis, i. é:

$$\mathbf{FC \text{ e } S \text{ intercambiáveis} \Leftrightarrow FC_0 = q} \quad \mathbf{(211)}$$

(b) É possível a revelação total, $\eta^2 = 1$, para qualquer valor não trivial de FC_0 e q. Ou seja, os limites de Fréchet-Hoeffding não restringem o uso de η^2:

$$\textbf{Limites de Fréchet-Hoeffding:} \quad 0 \leq \eta^2 \leq 1 \tag{212}$$

(c) As probabilidades de sucesso reveladas por S, i. é, FC^+ e FC^-, em caso de correlação *não*-negativa, são:

$$FC^+ = FC_0 + (1 - FC_0)\,\eta \tag{213}$$

$$FC^- = FC_0 - FC_0\,\eta \tag{214}$$

Em caso de correlação não-positiva, vale as eqs. (213) e (214), mas com o sinal após FC_0 invertido.

Prova: (a) Se S e FC são variáveis intercambiáveis então $p_{0,1} = p_{1,0}$ por definição. Logo, é possível igualar as eqs. (180) e (181) e obter o sentido $\Rightarrow$ da eq. (211). A volta é similar, se $FC_0 = q$ então pode-se igualar as eqs. (180) e (181) e concluir que $p_{0,1} = p_{1,0}$. Nesse caso, as v.a. são intercambiáveis por definição. □

(b) Como o limite superior de η^2 é 1, dado pela Proposição 6 (c), eq. (127), se uma das parcelas do lado direito da desigualdade da eq. (203) for igual a 1 ele será também o máximo. Como $FC_0 = q$, basta substituir isso na segunda parcela do máximo externo (parcela da linha de baixo) da eq. (207), que se obtém o valor de 1. O limite inferior já tinha sido provado no Teorema 4. □

(c) Basta substituir $FC_0 = q$ nas eqs. (198) e (199), que se obtém as eqs. (213) e (214). □

Note nas eqs. (213) e (214) que se está usando a raiz positiva da medida η^2. Assim se obtém equações lineares bem simples em η. Essas duas equações serão as mais usadas em aplicações. Além disso, note que a diferença de probabilidades de sucessos reveladas FC^+ e FC^- é exatamente η:

$$FC^+ - FC^- = \eta \tag{215}$$

Assim, no caso de v.a. de Bernoulli, a raiz positiva η parece ser mais intuitiva que η^2. Isso não foi verdade por exemplo quando se analisou exemplo do item 3.2.3 (Figuras 35 e 36), onde η^2 teve mais apelo intuitivo (cada um dos três poços com incerteza reduziam 1/3 da incerteza). Já aqui, a raiz positiva η é mais intuitiva pois dá o "spread" percentual entre os cenários FC^+ e FC^-. Na literatura

de psicologia também existem aplicações em que se reporta que o "eta" é mais intuitivo ou melhor, e outras em que o "eta squared" é melhor.

Para ilustrar as eqs. (213) e (214), seja o caso de um sinal de um poço vizinho intercambiável com o fator de chance do prospecto de interesse. Seja a probabilidade de sucesso (para ambas as v.a.) $FC_0 = 30\%$. Então, a depender do poder de revelação do sinal do poço vizinho, podem ser reveladas as seguintes probabilidades de sucesso FC^+ e FC^-, mostradas na Figura 43.

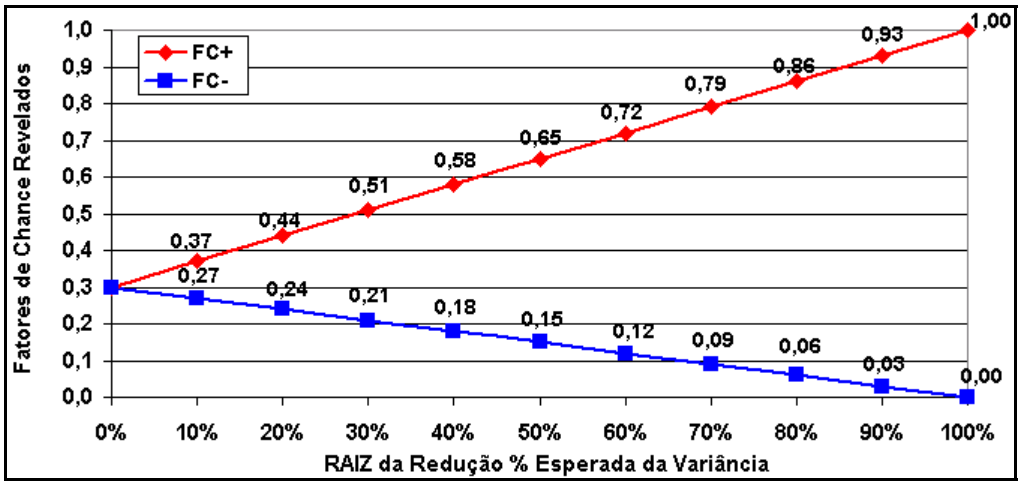

Figura 43 - Fatores de Chance Revelados x Raiz da Redução Esperada da Variância

Note que qualquer valor de η é possível, conforme a Proposição 7 (b). Em caso de revelação total, então se saberá com certeza se existe ou não petróleo na jazida, i. é, se o FC revelado é igual a 1 ou a zero. Note também a equação do "spread" (eq. 215) na diferença entre as linhas da Figura 43. Esse gráfico é representativo da simplicidade do método e seu apelo intuitivo.

Lema 7 (condição necessária para haver revelação total): Sejam FC e S v.a. não triviais de Bernoulli e seja a distribuição bivariada de Bernoulli definida por suas probabilidades de sucesso FC_0 e q e pela medida de aprendizagem η^2. A condição necessária para haver revelação total de uma variável de interesse FC é que o sinal S seja intercambiável com FC, i. é:

$$\eta^2(FC \mid S) = 1 \;\Rightarrow\; \textbf{FC e S v.a. intercambiáveis} \qquad \textbf{(216)}$$

Prova: Se $\eta^2(FC \mid S) = 1$ então as variâncias posteriores são iguais a zero, i. é, $FC^+ = 1$ e $FC^- = 0$ se a correlação é positiva ou $FC^+ = 0$ e $FC^- = 1$ se a correlação é negativa. Suponha o primeiro caso. Substituindo $FC^+ = 1$, $FC^- = 0$ e $\eta^2 = 1$ nas eqs. (198) e (199) e depois fazendo a diferença entre essas duas equações, i. é, eq. (198) – eq. (199), se obtém que FC0 = q. A suposição de

correlação negativa leva à mesma conclusão com os mesmos passos. Logo, FC e S são v.a. intercambiáveis. □

Assim, apenas um sinal intercambiável é forte o suficiente para poder ser *candidato* a um sinal que revele toda a verdade sobre FC. Ou seja, se ele for intercambiável é viável que $\eta^2(FC \mid S) = 1$, caso contrário é inviável. Isso pode ser visto como uma conseqüência dos limites de Fréchet-Hoeffding.

Um exemplo clássico de VOI em que se pode aplicar a teoria acima é o de Marschak (1959, p. 90). Marschak usa o critério de confiabilidade (verossimilhança) para determinar as probabilidades de sucesso após o sinal. Esse sinal é o parecer de um consultor que tem uma confiabilidade igual a p^c válida para os dois cenários revelados FC^+ e FC^-. Por ex., se a confiabilidade do consultor é de 70 % e ele diz que vai haver sucesso, então o comprador do conselho revisa a sua probabilidade de sucesso para $FC^+ = 70\%$, e se ele diz que vai haver insucesso, então o comprador do conselho revisa a sua probabilidade de sucesso no cenário revelado (más notícias) para $FC^- = 30\%$ (pois existe 70% de chances do consultor estar certo e a chance de fracasso ser $70\% = 1 - FC^-$). Sem o conselho do consultor, ele usa a sua probabilidade de sucesso a priori igual a FC_0. O comprador decide se vai investir usando uma probabilidade de sucesso e quer saber se o valor do conselho desse consultor, i. é, o VOI.

Nesse esquema de Marschak que usa a confiabilidade para estipular as probabilidades reveladas de sucesso, para existir a distribuição conjunta FC e S de Bernoulli, ou seja, ser consistente em termos de probabilidade (atender aos limites de Fréchet-Hoeffding), é necessário ajustar a probabilidade de revelação dos sinais q (boas notícias do consultor) e 1 – q (más notícias). Apesar do atrativo da simplicidade, pode-se mostrar facilmente que o intervalo válido para p^c, i. é, que permite uma probabilidade válida $q \in [0, 1]$ é muito mais restrito do que o caso em que se usa uma medida de aprendizagem (η^2 ou η) para estabelecer essas probabilidades de sucesso reveladas. Isso porque a *lei das expectativas iteradas* (ou o Teorema 1 b) diz que para haver consistência a equação abaixo deve ser obedecida (o lado esquerdo usa o esquema de confiabilidade de Marschak):

$$\mathbf{FC_0 = q\,FC^+ + (1-q)\,FC^- = q\,p^c + (1-q)(1-p^c)} \qquad \mathbf{(217)}$$

Tirando o valor de q, vem:

$$q = \frac{FC_0 + p^c - 1}{2\,p^c - 1} \qquad (218)$$

Por ex., se $FC_0 = 40\%$ e $p^c = 70\%$, então a eq. (218) indica que q tem de ser igual a 25% para existir consistência. Mas se $p^c = 60\%$, então q = 0, indicando que com certeza ele dará uma má notícia nesse modelo. Já se $p^c = 55\%$, então q dá um valor negativo (– 50%), i. é, uma inconsistência. Ou seja, é muito mais fácil trabalhar com uma medida de aprendizado η^2 do que com a confiabilidade. Para ilustrar, em cima desse exemplo de Marschak (1959, p.90), foi *adaptado* um problema de VOI para fator de chance exploratório, usando o VME (eq. 3), considerando o VOI como a diferença do VME com a informação e sem a informação[243]. A Figura 44 mostra o VOI para o esquema de confiabilidade de Marschak, considerando apenas consultores não mentirosos (i. é, apenas confiabilidades iguais ou superiores a 50%), para diversos probabilidades de sucesso a priori FC_0. Como antes, esse VOI é bruto (sem o custo do conselho).

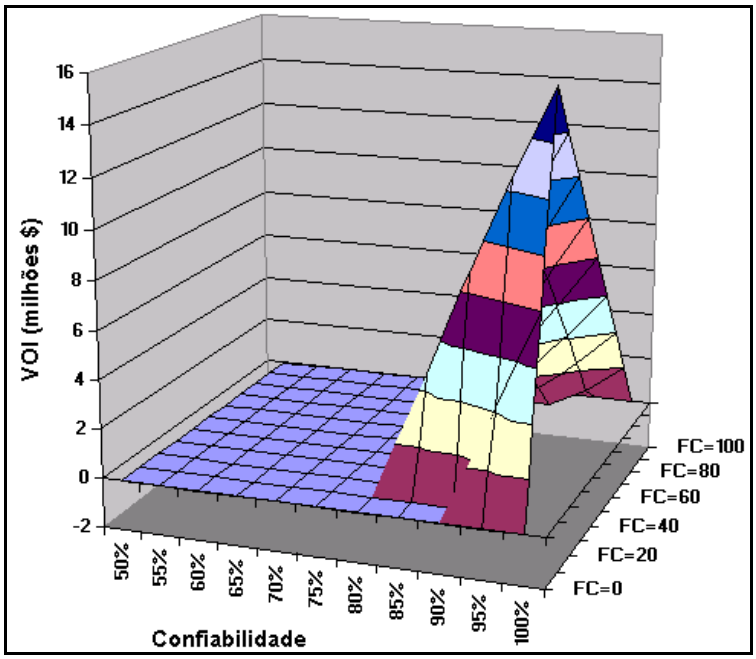

Figura 44 – VOI versus Confiabilidade para Diversos FC_0

Para o mesmo problema, a Figura 45 mostra o VOI usando a medida de aprendizagem η^2 em vez da confiabilidade, para diversos.

[243] Ver planilhas do CD-Rom Marschak_inf_Bernoulli.xls e Marschak_inf_Bernoulli-controle_eta.xls, com os detalhes numéricos e demais fórmulas.

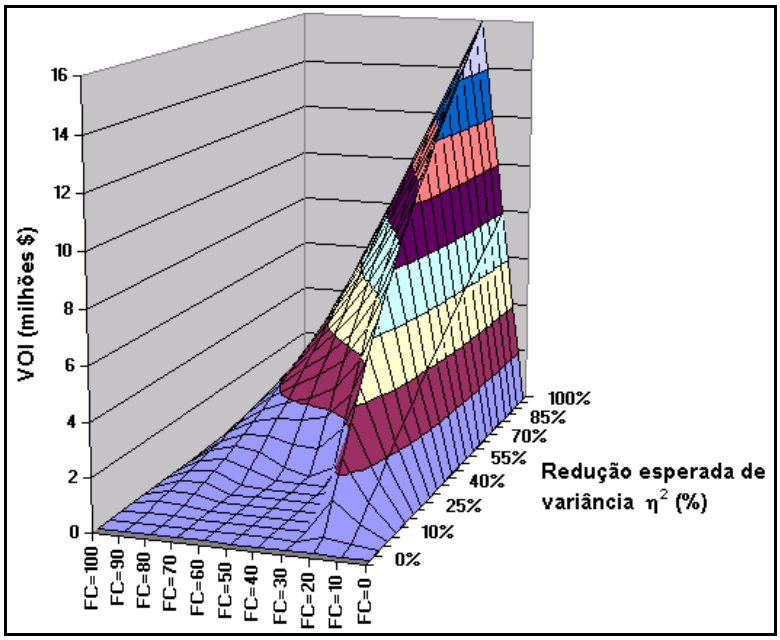

Figura 45 - VOI versus Medida η^2 para Diversos FC_0

Note que o intervalo de aplicabilidade da medida η^2 é bem maior do que a confiabilidade. Mesmo assim, se a aprendizagem for muito pequena, o VOI pode ser zero, por ex., um prospecto com VME ex-ante negativo, pode continuar a ser negativo após a informação e assim o prospecto continuaria a não ser perfurado em nenhum dos dois possíveis cenários revelados. Uma figura parecida é apresentada em Lawrence (1999, p.218) para um certo índice de informatividade θ. Em Lawrence (1999, ver p.209) esse índice pode ser o quadrado do coeficiente de correlação, para o caso de distribuição bivariada normal (devido a ρ^2 ser suficiente no sentido de Blackwell no caso de distribuição normal). Nesse caso específico, a abordagem recomendada pela tese coincide com a de Lawrence[244].

3.4.3. Alguns Processos de Revelação de Bernoulli

3.4.3.1. Processos de Descoberta e Processos de Revelação de Bernoulli

Em termos de análise econômica de projetos de valor da informação, e em particular de projetos de exploração de petróleo, pode-se caracterizar pelo menos dois tipos de processos interligados:

[244] Pena que Lawrence não tenha desenvolvido mais esse ponto, entrando nas complicações de matrizes de verossimilhança para os outros casos. Lawrence também não discutiu a questão de consistência da distribuição conjunta do ponto de vista dos limites de Fréchet-Hoeffding.

1) *processo de descoberta*, que consiste numa seqüência de exercícios de opções de aprendizagem até ocorrer uma descoberta; e
2) *processo de revelação* que modela o efeito probabilístico na variável de interesse de um processo de descoberta (exercícios de opções).

A seguinte definição *detalhada* de processo de descoberta é aplicada ao caso de exploração de petróleo e é baseada no artigo clássico dos professores do MIT e Harvard, Kaufman & Balcer & Kruyt (1975). Aqui a definição é adaptada para o contexto de opções reais e de valor dinâmico da informação na exploração.

Definição. **Processo de Descoberta Exploratória**: é uma seqüência de exercícios de opções reais de aprendizagem com diferentes custos e tempos de aprendizagem e diferentes poderes de revelação que culminam na descoberta de depósitos de petróleo. Essa seqüência de atividades pode ser: reconhecimento de superfície (afloramento geológico, etc.), pesquisas magnéticas, graviométricas, sísmicas e perfuração de um ou mais poços pioneiros.

Para operacionalizar o seu modelo, Kaufman & Balcer & Kruyt (1975) dividem o processo de descoberta em quatro módulos ou principais componentes: (a) submodelo de *volumes* descobertos por ordem de descoberta; (b) submodelo de *sucessos e falhas* (realizações da v.a. fator de chance); (c) submodelo *econômico* de um simples prospecto; e (d) submodelo de *mercado*. Assim, o modelo desses autores resulta em volumes de reservas em função de preços e características técnicas das reservas.

Nessa tese o processo de descoberta será modelado nas aplicações (ver cap.5) considerando o exercício ótimo de opções de aprendizagem, cujas conseqüências são modeladas com processos de revelação. Além disso, no processo de descoberta será considerada a natureza dinâmica dos preços do petróleo e outros fatores, inclusive a interação estratégica entre firmas de petróleo.

Como o conceito de processo de revelação foi estabelecido no item 3.2.2, o objetivo principal desse tópico é mais específico. O objetivo é descrever um processo de revelação em termos do impacto de uma seqüência de sinais em um fator de chance específico de um prospecto. Essa análise pode ser estendida para se estimar o número de descobertas esperadas num intervalo de tempo ou para um dado esforço (investimento) exploratório acumulado.

Nesse aspecto, McCray (1975, p.223-224) aponta quatro métodos para estimar a quantidade de descobertas: (1) *volume de sedimentos*, calcula-se esse

volume em uma bacia ou região e multiplica-se por um fator de chance *unitário* (por m^3 de sedimentos); (2) *binomial* (FC Constant), onde se assume um fator de chance médio e usa-se a distribuição binomial, i. é, considera n experimentos *independentes* de Bernoulli com a *mesma* probabilidade de sucesso; (3) *quantidade fixa de campos* numa região, com o número de descobertas sendo modelada com uma distribuição hipergeométrica (experimentos de Bernoulli *sem* reposição), i. é, assumindo uma forma *fraca de dependência*; e (4) *predições de tendências* ("trend predictions"), através da identificação de fatores os quais as descobertas estão correlacionadas, de forma a gerar tendências baseadas na extrapolação do passado de outras bacias para a bacia em estudo[245].

O método sugerido nessa tese apresenta uma alternativa aos quatro métodos descritos por McCray (1975, p.223-224), embora possa ser integrado com o quarto método. O método proposto foi testado em um projeto PUC-Petrobras[246], usando a simulação de Monte Carlo e a teoria de processos de revelação para modelar e avaliar o exercício ótimo de opções reais numa bacia ao longo de um período de 20 anos. Em alguns caminhos simulados ocorrem boas notícias (descobertas) que produzem revelações positivas de informações, elevando os fatores de chance de outros prospectos de um bloco ou região que estejam no mesmo play geológico. Nos casos de insucesso, os fatores de chance de prospectos do mesmo play geológico são reduzidos, o que por sua vez reduz a probabilidade de novos exercícios ótimos de opções nesse play. Ou seja, é assumida uma dependência bem mais forte e realista do que a mencionada por McCray com a distribuição hipergeométrica.

O processo de revelação é em geral dependente do caminho ("path-dependent") e por isso é geralmente necessário usar a simulação de Monte Carlo. Isso significa que, em geral, um processo de revelação não é Markoviano. Por isso, em geral não podem ser usados métodos tais como as *cadeias de Markov*. Em casos específicos, nem sempre convenientes, um processo de revelação de um fator de chance pode ser Markoviano.

[245] Ex.: assume que o número de campos descobertos é proporcional ao número de poços perfurados vezes a área alvo (McCray, 1975, p.227). Calibra-se o modelo com dados passados.

[246] Do Pravap-14, usado para avaliar uma região petrolífera marítima na África num horizonte de 20 a 30 anos de exercícios de opções reais de aprendizagem (sísmica e perfuração principalmente) e de desenvolvimento. Os detalhes são omitidos por razões de confidencialidade.

Embora o uso de simulação seja mais trabalhoso do que os métodos analíticos que usam fórmulas das distribuições binomiais e hipergeométricos, essa tese considera que as simplificações que permitem usar essas distribuições são em geral inaceitáveis por não ter a menor representatividade de fatos estilizados de exploração de petróleo[247] e com a teoria probabilística de fatores de chance correlacionados desenvolvidas no item 3.4.2. No entanto, pode-se pensar em combinar o método desenvolvido nessa tese com o método de *predições de tendências*, usado por McCray (1975, p.224-239), a fim de calibrar melhor as condições de contorno dos processos de descoberta e conseqüentemente os processos de revelação, com correlações baseadas em dados de outras bacias. Isso é deixado como sugestão de pesquisa futura.

Uma outra abordagem de interesse, mas para o processo de revelação, foi mencionado no item 3.4.2 e agora será analisado. O artigo de Wang et al (2000) analisou o processo de dependência entre um fator de chance e sinal e apresentou um processo de revelação do fator de chance com dois sinais seqüenciais. A eq. (173) de Wang et al sugeriu que, em caso de n sinais negativos, o fator de chance inicial decai do fator $(1 - dep/5)^n$. No exemplo, foi usado dep = 2. É curioso notar que, no caso particular de v.a. intercambiáveis, a raiz positiva η da medida de dependência proposta na tese é exatamente igual ao dep/5 de Wang et al, basta comparar as eqs. (173) e (214). No caso, para dep = 2 tem-se $\eta = 40\%$.

Os valores do processo de revelação construído por Wang et al (2000, Fig.4) são mostrados na Figura 46, incluindo algumas informações adicionais tais como as probabilidades assumidas de forma implícita para os sinais e as probabilidades de ocorrência dos cenários finais da distribuição de revelações.

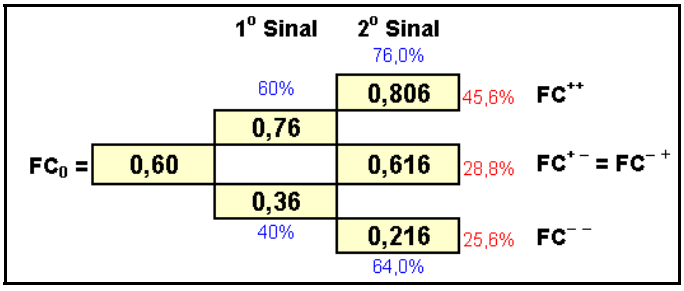

Figura 46 – Processo de Revelação em Wang et al

[247] Lembrar a discussão no item 3.4.1 sobre a evolução do fator de chance ao longo do tempo, comentando os gráficos típicos de Rose (2001, p.46) e de Jones & Smith (1983, Fig. 4), com períodos cerca de três anos de grande aumento do FC e depois um longo período de declínio.

Dentro das células da Figura 46 estão os fatores de chance inicial (FC_0 = 0,6) e os FC revelados. As probabilidades azuis mostram as probabilidades *condicionais* de ocorrência dos cenários (probabilidades do sinal ser positivo ou negativo) assumindo implicitamente v.a. intercambiáveis. As probabilidades terminais (vermelhas) são as probabilidades de ocorrência dos cenários, após dois sinais (essas têm de somar 1, pois *não* são condicionais). Note na Figura 46 que após dois sinais, a distribuição de revelações tem três cenários discretos, cada cenário pode ocorrer segundo as probabilidades terminais mencionadas.

Aproveitando o exemplo, são introduzidas as notações para os FC revelados após dois sinais: FC^{++}, FC^{+-} (que é igual ao FC^{-+} no caso recombinante) e FC^{--}. Pode-se verificar que o processo de revelação sugerido pela eq. (173) tende a ser convergente para o caso de revelação total, pois em caso de dependência (dep > 0) então se $n \to \infty \Rightarrow (FC^{-})^n \to 0$. Já o cálculo dos outros cenários $(FC^{+})^n$ e os FC's para as combinações de sinais positivos e negativos é feita com a equação de consistência mencionada no item 3.4.2 (que é nada mais é do que a lei das expectativas iteradas, também usada na tese).

No exemplo de Wang et al (Fig. 4 deles), além da intercambialidade, pode ser visto que eles assumiram uma outra premissa, mas novamente sem explicitá-la: o processo de revelação é recombinante, i. é, se houver um sinal positivo e outro negativo, o novo FC é igual ao caso do primeiro sinal ser negativo e o segundo sinal positivo. Uma vantagem prática do processo recombinante é que reduz de forma significativa o número de cenários da distribuição de revelações: após k sinais o número de cenários no caso da árvore recombinate é de k + 1, enquanto que no caso da árvore não-recombinate esse número de cenários é de 2^k, e é fácil ver que $2^k > k + 1$ se $k > 1$. Embora tenha sua lógica (e aqui serão estudados os processos recombinantes), em geral *não* é necessário ser recombinante para ser consistente e pode ser não realista. Isso porque os sinais intercambiáveis terão diferentes probabilidades em cada cenário. Nesse exemplo, caso tenha ocorrido $FC^{+} = 76\%$ com o primeiro sinal, o segundo sinal terá 76% de chances de ser uma nova boa notícia (devido a intercambialidade), enquanto que caso tenha ocorrido $FC^{-} = 36\%$, a chance do segundo sinal ser uma nova má notícia é de 64%.

Em Wang et al (2000), para que ao *mesmo tempo* possam haver intercambialidade entre FC e sinal S, recombinação de cenários e convergência do processo de revelação para a revelação total (indicada pela eq. 173), eles implicitamente assumiram que o impacto em termos de aprendizagem (redução esperada de variância) do segundo sinal é variável com FC. No exemplo deles, esse impacto é $\eta^2 = 3{,}6\%$ (ou $\eta = 18{,}9\%$) se for sobre o cenário $FC^{+} = 76\%$ e de $\eta^2 = 16\%$ (ou $\eta = 40\%$) se for sobre o cenário $FC^{-} = 36\%$. Ou seja, o *mesmo sinal* S_2 provocaria um impacto (redução de variância) no *mesmo prospecto*, após ter sido revelado *o mesmo* sinal S_1, de forma muito diferente (mais de 4 vezes) a depender da correção feita com o primeiro sinal (a depender do caminho). Isso não parece ser muito consistente, apesar de todos os atrativos da conjugação de premissas simplificadoras (intercambialidade, recombinação e convergência).

O relaxamento de uma das premissas implícitas (ex.: recombinação), poderia ter levado a um resultado mais consistente em Wang et al (2000). Um processo de revelação não-recombinante intercambiável é mostrado na Figura 47 a seguir, em que foram usadas as premissas de Wang et al (2000) para o cenário FC^{--}, mas considerando que o poder de revelação do segundo sinal é o mesmo, quer o primeiro sinal tenha sido positivo ou negativo.

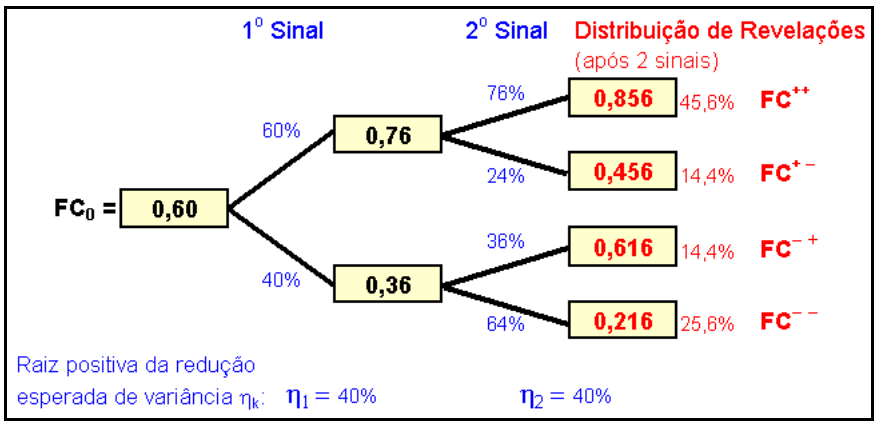

Figura 47 – Processo de Revelação Não-Recombinante

Algumas questões importantes podem ser discutidas observando a Figura 47. Primeiro, seguindo a premissa de Wang et al para determinar FC^{--}, o sinal S_1 tem o mesmo poder de revelação do sinal S_2 ($\eta_1 = \eta_2 = 40\%$), o que nem sempre é realista (em muitos casos pode ocorrer uma redução do poder de revelação ou mesmo um aumento). Segundo, o FC em caso do *caminho* sinal positivo + sinal negativo (FC = 0,456) é menor que em caso do *caminho* sinal negativo + sinal positivo (FC = 0,616), i. é, não-recombina. Essas duas questões estão ligadas. Para

recombinar é necessário que $\eta_2 < \eta_1$ em pelo menos um cenário após o 1º sinal. Na maioria dos casos existe correlação (positiva) entre os sinais S_1 e S_2, e assim, *parte da informação* que S_2 iria revelar para o FC do prospecto já foi revelada por S_1. É mais fácil ver isso com um caso limite: se $\eta^2(S_2 \mid S_1) = 100\%$, então a informação proveniente de S_1 tornaria S_2 determinístico e dessa forma o segundo sinal não traria nenhuma informação adicional para o FC do prospecto de interesse após a revelação de S_1. Mas se o sinal S_2 fosse revelado primeiro ele seria relevante para FC e S_1 é que passaria a ser irrelevante. Esse raciocínio pode sugerir uma seqüência η_k com valores decrescentes. No entanto, a medida η_2 é sobre um FC com uma variância esperada menor devido à ação do 1º sinal e assim pode ocorrer que η_2 seja até maior que η_1. Por ex., se a correlação entre S_1 e S_2 não for perfeita e o conhecimento de ambos S_1 e S_2 prover a *revelação total* sobre FC, então pode ocorrer, por ex., $\eta^2(FC \mid S_1) = 66{,}6\%$ e $\eta^2(FC \mid S_1 \mid S_2) = 100\%$, apesar de, ex-ante, S_1 poder ser mais forte do que S_2[248], ver exemplo da Figura 35. Assim, a especificação mais adequada do poder de revelação de sucessivos sinais depende do tipo de problema e várias alternativas são possíveis.

Será estudado em especial o caso de sinais com η_k decrescentes, i. é, uma queda no poder de revelação dos sucessivos sinais, pois apenas nesse caso pode haver uma seqüência infinita de sinais que não converge para o limite de revelação total (a ser visto). Além disso, no caso de se querer a recombinação na árvore de cenários, é necessário especificar η_k decrescentes. O valor dessa queda em η_k depende do tipo de sinal. Por ex., foi comentado que o sinal de uma sísmica é complementar em muitos aspectos ao sinal da perfuração de um poço vizinho[249]. Dessa forma, se a sísmica for o primeiro sinal, o poder de revelação de um poço vizinho não deve diminuir muito comparado ao caso do poço vizinho ser o primeiro sinal. Já no caso de dois poços vizinhos (no mesmo play geológico do prospecto), esses sinais são em certo grau sinais substitutos, ou seja, a revelação do primeiro sinal diminui a relevância *absoluta* do segundo sinal.

Para ilustrar esse ponto, no caso do exemplo modificado (não-recombinante) de Wang et al, suponha que $\eta_2 = 20\% < \eta_1 = 40\%$. O processo de revelação da

[248] $\eta^2(FC \mid S_1) > \eta^2(FC \mid S_2)$, mas, uma vez conhecido S_1, o sinal S_2 passa a ter $\eta^2 = 100\%$.

[249] Ver item 3.4.1: a sísmica revela muito da estrutura, geometria, rochas, mas quase nada de sincronismo, enquanto que um poço vizinho revela o sincronismo e pouco da geometria local.

Figura 47, mas com esse valor menor para a segunda revelação é mostrada na Figura 48 a seguir.

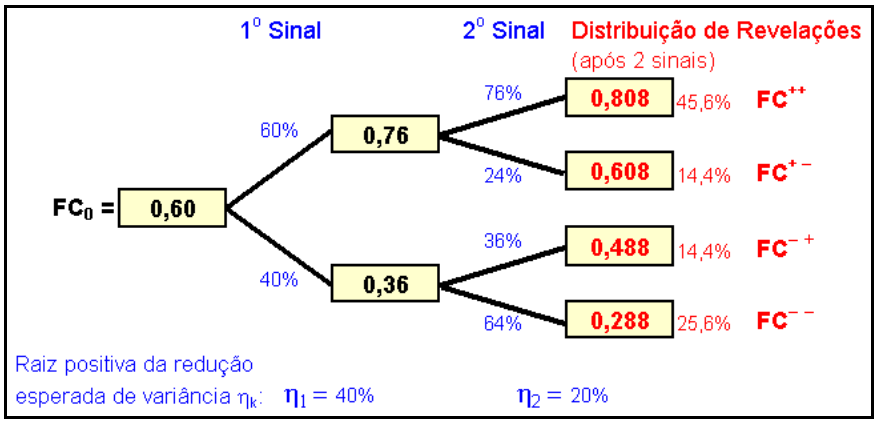

Figura 48 – Processo de Revelação Decrescente Não-Recombinante

Note que, comparado com a Figura 47, a Figura 48 mostra uma inversão: o FC em caso do caminho sinal positivo + sinal negativo (FC = 0,608) passa a ser *maior* que em caso do caminho sinal negativo + sinal positivo (FC = 0,488). Isso reflete o fato do segundo sinal ter sido *bem* menos informativo que o primeiro. Esse raciocínio e o exemplo também sugerem que é possível um processo intercambiável ser também recombinante, sem ter de variar o poder de revelação com o caminho (como implicitamente fizeram Wang et al)[250]: basta variar η_k (reduzir) ao longo da seqüência, de forma adequada, assumindo a lógica que esses sinais compartilham uma parte da informação relevante para o prospecto de interesse. Essa forma mais adequada/consistente de recombinação será analisada no sub-item 3.4.3.2, onde será visto que no limite implicará em uma convergência parcial (que não será a revelação total). A convergência apenas parcial pode ser não ser inconveniente na prática: mesmo em bacias muito exploradas, não se sabe toda a verdade sobre um novo prospecto, mesmo num play geológico conhecido.

Um outro processo de revelação de fator de chance foi estudado por Jaynes (2003, p. 75-82), mas num contexto de crítica à teoria da amostragem, em especial ao esquema de urnas (binomial, hipergeométrico) para estimativa de um parâmetro. A pergunta dele é: o que ocorre se, ao retirar e repor uma bola vermelha numa urna, a sua probabilidade para o próximo sorteio aumentar de um valor $\varepsilon > 0$? Ele assume também que, caso no primeiro sorteio saia uma bola

[250] Note que Wang et al assumiram que S_2 permanece com o mesmo poder de revelação, apesar de já ter sido revelado S_1, desde que a primeira revelação fosse negativa. Mas o poder de revelação cairia muito se a primeira revelação fosse positiva. Parece mais lógico que haja uma queda do poder de revelação dos sinais só devido aos mesmos serem correlacionados entre si.

branca, a probabilidade de tirar uma bola vermelha no próximo sorteio diminui de um valor $\delta > 0$. Ou seja, existe um processo de revelação do fator de chance de sair bola vermelha em que os sinais são sorteios sucessivos. Assim, o processo de revelação de Jaynes (2003) pode ser visto com as equações:

$$\mathbf{FC^{+} - FC_0 = \varepsilon > 0} \tag{219}$$

$$\mathbf{FC_0 - FC^{-} = \delta > 0} \tag{220}$$

Com as eqs. (183) e (184) é fácil mostrar que esse esquema é viável se:

$$\varepsilon = \frac{p_{11} - q\, FC_0}{q} \tag{221}$$

$$\delta = \frac{p_{11} - q\, FC_0}{1 - q} \tag{222}$$

Daí pode-se concluir que $\varepsilon = \delta$ somente se q = 50% (Jaynes, p.76, chega a essa conclusão de uma forma totalmente diferente). Mas isso eliminaria a possibilidade geral de intercambialidade. Com as eqs. (221) e (222) pode-se obter a seguinte equação (Jaynes, p.81, usando outra abordagem):

$$q\,\varepsilon = (1 - q)\,\delta \tag{223}$$

Jaynes (2003, p.81), comentando a eq. (223), reconhece que "*a distribuição ainda não é intercambiável*", mas aponta outras vantagens. A idéia dele foi fazer uma aproximação Markoviana para estimar um parâmetro. Embora o processo de Jaynes seja uma alternativa, ela não parece ser a melhor ao não permitir, no caso geral, v.a. intercambiáveis (o que implica na impossibilidade de revelação total ou informação perfeita, conforme visto no item 3.4.2) e também por não trabalhar com uma medida de aprendizagem que tivesse uma interpretação mais intuitiva e permitisse comparar sinais de diferentes problemas relacionados, mas com diferentes distribuições a priori (i. é, diferentes probabilidades de sucesso FC_0).

Um caso similar à medida de dependência de Jaynes (2003), particularmente a eq. (219), foi feita por Smith (2004), justamente para o caso de exploração de petróleo. Ele, no entanto, usa uma versão percentual da eq. (219) ao dividir a diferença por FC_0. Dessa forma, após o k-ésimo poço perfurado seco, é revelado o fator FC^{-}_{k+1}, a medida de dependência (na notação da tese) é:

$$d_k = \frac{FC^{-}_{k} - FC^{-}_{k+1}}{FC^{-}_{k}} \tag{224}$$

Esse modelo tem, em geral, os mesmos problemas do modelo de Jaynes. Smith (2004) analisa um caso que ele considera realista de aumento da dependência com o número de fracassos. Conforme visto antes, isso pode não ser o mais realista, dada a correlação entre sinais e dada a premissa razoável de que a probabilidade de receber sinais positivos (q_k) deve variar com o caminho (ex.: ser intercambiável)[251]. Smith também analisa os casos limites de independência e dependência total (que equivale ao conceito de revelação total) e o impacto no valor da opção "de perfurar de novo", que é o foco do artigo dele. Assim como Wang et al (2000), Smith focaliza a seqüência de falhas (uma aplicação é quando parar). Embora seja uma aplicação relevante, parece mais interessante analisar também (e principalmente) os caminhos de revelação positiva que, mesmo com menor probabilidade, levam a uma seqüência de elevadas chances de sucesso (efeito *cascata* da dependência), onde se pode alavancar *muito* o valor de OR de uma firma se ela deter prospectos opcionais e puder se aproveitar dos sinais. Esse foi, por ex., o foco do projeto PUC-Petrobras mencionado antes. Seqüências de sucesso da história da exploração (tem vários exemplos na Bacia do Espírito Santo, Bacia de Santos e Bacia de Campos) é que permitem alavancagens de valor das OR de companhias de petróleo que tem atuação importante no E&P.

Wang et al (2000) foram quem chegaram mais perto da formulação adequada de um processo de revelação de um fator de chance exploratório, usando quase que apenas a intuição e a lógica. Para melhorar esse resultado é necessário detalhar a discussão da relação probabilística entre FC e sinais, como foi feita aqui no item 3.4.2. De qualquer modo, Wang et al, Smith e Jaynes são exceções na literatura, ao abordar o problema de *processos* dependentes de Bernoulli sem recorrer a esquemas conhecidos do tipo hipergeométrico (sorteio sem reposição), que daria apenas uma dependência muito fraca e não representativa. Como esse é um tema praticamente inexplorado, a tese ficará bem longe de esgotar esse assunto, mas tem a intenção de dar algumas contribuições relevantes.

A discussão de propriedades tais como intercambialidade, recombinação e convergência em processos de revelação de Bernoulli, será feita no próximo item

[251] Smith (2004) assume uma premissa de uma probabilidade condicional inversa constante. Isso leva o processo a não ser intercambiável o que *pode* levar a problemas de consistência (limite de Fréchet-Hoeffding) a depender da medida d_k usada. Smith focou num *ramo* da árvore (sucessivas falhas). Aqui se olha a toda a árvore, o que dá uma visão mais global de consistência.

(3.4.3.2), onde serão feitas mais comparações sobre as alternativas de processos de revelação para fatores de chance.

3.4.3.2. Processos Convergentes, Recombinantes e Outros

O objetivo da análise de processos de revelação é estudar seqüências de distribuições de revelações. Cada sinal gera uma distribuição de revelações. As distribuições de revelações de Bernoulli permitem resolver problemas de opções reais de uma maneira direta. O valor ex-ante do prospecto de interesse, onde se espera a revelação (positiva ou negativa) de dois sinais, como na Figura 48, deve considerar o problema de exercício ótimo da opção de perfurar o prospecto, em *cada um dos 4 cenários revelados* (4 fatores de chance diferentes) da distribuição de revelações. O fator de chance de cada cenário é usado como parâmetro de equações de valor do prospecto (ativo básico) tais como a eq. (3). O valor da OR para cada cenário, deve ser ponderado pelas probabilidades dos cenários da distribuição de revelações (percentuais em vermelho na Figura 48)

Dessa forma, a seqüência de distribuições de revelações ajuda a avaliar ex-ante problemas de sinais seqüenciais, i. é, decidir por um plano de avaliação exploratória. Exs.: investir em sísmica e depois perfurar um poço pioneiro correlacionado a um outro; ou, após a sísmica, calcular o valor da alternativa de esperar por um sinal (comportamento estratégico) em um jogo de opções reais exploratórias; ou a análise do valor de um bloco exploratório com vários prospectos correlacionados; ou o valor de entrar numa nova bacia de alto risco, mas que se espera haver perfurações que podem alavancar portfólio de prospectos na bacia, etc. Assim, processos de revelação têm grande importância prática.

No item 3.4.2 foi analisado com detalhes (Proposição 7) o caso intuitivo de intercambialidade entre FC e sinal S, que permite uma grande simplicidade de tratamento. Normalmente essa premissa será adotada por essas razões, além do que permite um tratamento mais uniforme e consistente para todo um processo de revelação. A idéia intuitiva é que se houver uma revelação positiva numa área, aumenta não só a chance de sucesso FC^+ do prospecto de interesse, como também aumenta a chance de um novo sinal positivo. Esse novo sinal pode ser a perfuração de um novo poço na área ou até a reinterpretação da sísmica e de modelos geológicos, incluindo mapas. De forma análoga, em caso de fracasso

(sinais negativos), além de reduzir a probabilidade de exercícios de novas opções exploratórias, quando isso voltar a ocorrer é de se esperar que o *novo sinal* tenha menor probabilidade de ser positivo do que tinha antes da má notícia.

A intercambialidade é a maneira natural de capturar esse processo de mudança de probabilidades de sinal positivo $q_k(c)$ no k-ésimo sinal e no caminho c. Deve-se lembrar que, em geral, o processo de revelação de Bernoulli é dependente do caminho (não-Markoviano). Além disso, a intercambialidade é uma condição necessária para se poder garantir a convergência de um processo de revelação de Bernoulli em direção do caso limite de revelação total.

No tópico anterior foram comentadas algumas outras características de processos de revelação, em especial ao se referir ao artigo de Wang et al (2000). Duas dessas características são *recombinação* de cenários e *convergência* do processo de revelação para a revelação total num critério de limite após um número muito grande de sinais.

A recombinação de cenários permite que o processo de revelação seja Markoviano, i. é, independente do caminho. Nesse caso, o valor do fator de chance revelado $FC_k(p, n)$, após k sinais, sendo p sinais positivos e n sinais negativos (logo $k = p + n$) é determinado apenas por p e n (ou p e k ou n e k), dada a condição inicial (FC_0) e dada a *estrutura de informação* do processo de revelação. A Figura 46 mostrou um exemplo do caso mais simples (dois sinais) de um processo de revelação recombinante que, em adição, era intercambiável.

Como foi comentado antes, nem sempre é adequado ou consistente usar a recombinação, pois o impacto do mesmo sinal no mesmo prospecto (poder de revelação do sinal sobre um prospecto) pode variar muito para poder forçar a recombinação. A comparação dos exemplos das Figuras 47 e 48, mostra que é possível manter uma estrutura de informação com certas características consistentes (mesmo poder de revelação do k-ésimo sinal, naqueles exemplos seria um valor de η entre 20 e 40%) mas isso irá impedir o processo de convergir, apesar da intercambialidade. A intercambialidade é condição necessária (Lema 7), mas *não suficiente* para a revelação total ser alcançada. A vantagem da recombinação em qualquer caso é a simplicidade de independer do caminho. Nesse aspecto, ele lembra o modelo binomial de Cox & Ross & Rubinstein (1979), mas suas probabilidades são diferentes (no binomial das opções financeiras, o valor *descontado* dos cenários é que se comporta como martingale).

A estrutura de informação (ver eq. 57 e a versão mais flexível, eq. 63) de um processo de revelação de Bernoulli é dada pela medida de aprendizagem (η^2 ou η) e pelas probabilidades marginais $FC_k(c)$ e $p_k(c)$, no k-ésimo sinal, caminho c. Como foi visto, com apenas três parâmetros é possível definir a distribuição bivariada, desde que ela seja viável. E essa viabilidade é dada pelo critério adicional dos limites de Fréchet-Hoeffding. Em suma, a estrutura de informação é definida por 4 elementos, as probabilidades $FC_k(c)$ e $p_k(c)$, η, e o limite superior de Fréchet-Hoeffding (já que o inferior é sempre zero).

A estrutura de informação intercambiável proposta para várias aplicações, é tremendamente simplificada, pois em vez de 4 elementos, são necessários apenas dois elementos: a probabilidade $FC_k(c)$ e a medida de aprendizagem η. Isso porque sendo v.a. intercambiáveis, pela Proposição 7 (a) e (b), $FC_k(c) = p_k(c)$ e o limite de Fréchet-Hoeffding é irrelevante pois não restringe a medida de aprendizagem η, que pode ser qualquer valor em que ela é válida (intervalo unitário). Assim, assumindo apenas dois elementos (FC_0 e η), se pode construir todo um processo de revelação, obtendo todos os valores de $FC_k(c)$ que, para vários casos tais como η constante (independente de k), será mostrado que converge para o limite teórico de revelação total.

Em relação à convergência de um processo de revelação de Bernoulli, deve ser lembrado que revelação total significa que o prospecto de interesse terá o seu FC convergindo para zero ou para um. Wang et al (2000) apresentaram um processo que converge pois tem a condição necessária de intercambialidade, e porque o cenário extremo analisado (seqüência de insucessos) converge para zero. Isso deve induzir cenários de revelação positiva a convergir para 1. No entanto, será apresentado um critério mais claro de convergência para revelação total que não olha uma probabilidade específica (k fracassos ou k sucessos) e sim na variância esperada das distribuições posteriores, que deve convergir para zero.

A Figura 46 mostrou que após 2 sinais, a distribuição de revelações tinha 3 cenários. No caso recombinante, é fácil ver após k sinais a distribuição de revelações terá k + 1 cenários. Isso parece ser inconsistente com o Teorema 1 (a) que diz que no limite a distribuição de revelações é igual à distribuição a priori (que tem só dois cenários, já que é uma distribuição de Bernoulli), se o processo for convergente. Mas essa contradição é só aparente. O que ocorre na verdade

(simulações e a discussão abaixo) é se k for um número relativamente grande, dos k + 1 cenários, se terá um grupo de cenários com valores muito perto de FC = 0 (massa de probabilidade se concentrando em cenários próximos de zero) e outro grupo de cenários com valor muito perto de FC = 1. Os cenários intermediários (ex., FC próximo de 0,5) terão cada vez menor probabilidade, até que, no limite, se terá apenas dois cenários FC = 1 e FC = 0, com a massa de probabilidade concentrada nesses dois cenários. Essas massas de probabilidade serão exatamente iguais a FC_0 e $(1 - FC_0)$, respectivamente, como prevê o Teorema 1 (a)!

Seja um processo de revelação com valor inicial FC_0 não trivial. Seja o poder de revelação (qualidade) do sinal k, $\eta^2(FC \mid S_k)$, denotado por simplicidade por η_k^2, que depende da quantidade e qualidade dos sinais anteriores, mas que independe do caminho (se foram reveladas boas ou más notícias), de forma a ser o mesmo para todos os possíveis FC revelados após o sinal k – 1 (i. é, o mesmo para todos os cenários da distribuição de revelações após o sinal k – 1). A seguinte condição geral garante a convergência de um processo de revelação de Bernoulli:

$$\lim_{n \to \infty} \prod_{k=1}^{n} (1 - \eta_k^2) = 0 \qquad (225)$$

É fácil ver porque o processo converge, caso a eq. (225) seja satisfeita: se o valor inicial do FC é $FC_0 \in (0, 1)$, então a sua variância a priori é não nula e igual a $FC_0 (1 - FC_0)$. A cada revelação do sinal S_k, a variância esperada do FC é reduzida em η_k^2 %, ou seja, a variância esperada decai por um fator multiplicativo $(1 - \eta_k^2) < 1$. Logo, a variância média do FC posterior a n sinais, cada um com poder de revelação η_k^2, dado que a variância inicial (a priori) é $FC_0 (1 - FC_0)$, é:

$$E[Var[FC \mid S_1, S_2, ..., S_n]] = FC_0 (1 - FC_0) \prod_{k=1}^{n} (1 - \eta_k^2) \qquad (226)$$

Assim, após n sinais, a variância inicial de FC foi reduzida pelo produtório de $(1 - \eta_k^2)$. Se esse produtório tende a zero, então o seu produto pela variância inicial (finita) também irá a zero. Se a variância esperada no limite tende a zero, então se tem a revelação total por definição.

Note que é muito *fácil* obter o limite de revelação total. Em particular, se os valores η_k^2 forem iguais (sinais do mesmo tipo ou com o mesmo poder de revelação), então é trivial ver que $(1 - \eta^2)^n$ tende a zero se n tende a infinito, pois

como $(1 - \eta^2)$ é um número positivo menor que 1, o produtório converge. Por ex., se $\eta^2 = 10\%$, após 50 sinais a variância esperada (posterior) é apenas 0,5 % do valor inicial. Com 100 sinais, a variância média das distribuições posteriores é de apenas 0,00265 % do valor inicial (ou praticamente zero).

A eq. (225) também se verifica para o caso de sinais com poderes de revelação crescentes $\eta^2_{k+1} > \eta^2_k$. De forma geral, se obtém convergência para qualquer seqüência infinita de η^2_k não decrescentes. E se a seqüência η^2_k for decrescente? Nesse caso o processo *pode* se estabilizar (ter um limite) num patamar de variância média posterior maior ou bem maior que zero, não convergindo para a revelação total. Na eq. (226), se para k grande o fator $(1 - \eta^2_k)$ convergir para 1, o produtório não converge para zero e o processo não converge para a revelação total.

Para analisar a convergência do produtório usado nas eqs. (225) e (226), pode ser conveniente lembrar que o logaritmo do produto é igual à soma dos logaritmos, reduzindo o problema para a análise da soma infinita de uma série:

$$\prod_{k=1}^{\infty}(1-\eta_k^2) \;=\; \exp\left(\sum_{k=1}^{\infty}\ln(1-\eta_k^2)\right) \qquad \textbf{(227)}$$

O produtório acima converge para zero se o somatório dentro do exponencial divergir para $-\infty$, lembrando que $\ln(1 - \eta^2_k)$ é um número negativo. Assim, para haver revelação total (produtório convergir para zero), deve-se ter:

$$\lim_{n\to\infty}\sum_{k=1}^{n} -\ln(1-\eta_k^2) \;=\; +\infty \qquad \textbf{(228)}$$

Se o limite dessa soma for *menor que infinito* (i. é, a eq. 228 convergir), então não haverá convergência do processo de revelação para a revelação total. Mas o processo de revelação irá se estabilizar convergindo para uma variância esperada *intermediária* entre o valor da distribuição a priori FC_0 $(1 - FC_0)$ e o valor zero (revelação total). Note que <u>não</u> existe a possibilidade do processo de revelação <u>oscilar</u> em termos de variância esperada (i. é, a média das distribuições posteriores após infinitos sinais), pois a cada sinal a variância posterior *esperada* nunca aumenta (conseqüência do Teorema 1 c). Ou seja, com infinitos sinais, um processo de revelação sempre converge ou para a revelação total (variância

posterior esperada igual a zero) ou para a revelação parcial (variância posterior esperada converge para um valor entre a variância da distribuição a priori e zero).

Principalmente para mostrar que um processo de revelação de Bernoulli *não* converge para a revelação total, é útil a seguinte desigualdade (Lima, 1996, p.84):

$$\eta_k^2 \;<\; -\ln(1-\eta_k^2) \;<\; \frac{\eta_k^2}{1-\eta_k^2} \qquad \textbf{(229)}$$

Se um somatório infinito das frações do lado direito da eq. (229) convergir para um número menor que infinito, então o somatório infinito de $-\ln(1-\eta_k^2)$, um número menor (eq. 229), também irá convergir para um número menor que infinito. Assim, como condição suficiente, o critério para um processo de revelação de Bernoulli *não* convergir para a revelação total (e sim para parcial) é:

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{\eta_k^2}{1-\eta_k^2} \;<\; \infty \qquad \textbf{(230)}$$

A eq. (230) pode ser usada para provar a não convergência total usando critérios de convergência de somas infinitas tais como o *teste da razão* ("ratio test") ou o *teste da integral*, ver, por ex., Sydsaeter & Strøm & Berck (2000, p.43). Para ilustrar o teste da razão para convergência, seja o seguinte esquema simples de decaimento constante (= $\gamma < 1$) dos poderes de revelação sucessivos:

$$\eta_k^2 \;=\; \gamma^{k-1}\;\eta_1^2\;,\;\; \gamma \in (0,\,1) \qquad \textbf{(231)}$$

Ou seja, η_2^2 é menor que η_1^2 por um fator γ, etc. Para provar que o processo de revelação da eq. (231) não converge, basta provar que a eq. (230) converge. Para isso, o teste da razão diz que uma soma infinita de $\{a_n\}$ converge se:

$$\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| \;<\; 1 \;\Rightarrow\; \sum_{n=1}^{\infty} a_n \text{ converge} \qquad \textbf{(232)}$$

Substituindo os termos η_{k+1}^2 e η_k^2 dados pela eq. (231) na eq. (232), simplificando e chamando a razão obtida de R, vem:

$$R \;=\; \gamma\;\frac{1-\gamma^k\,\eta_1^2}{1-\gamma^{k+1}\,\eta_1^2} \qquad \textbf{(233)}$$

Na eq. (233) pode-se observar que a fração é sempre menor que 1 e, no limite para $k \to \infty$, a fração tende a 1. Como o valor de γ (constante que multiplica a fração da eq. 233) é menor que 1, então o limite de R quando $k \to \infty$ é menor que 1 (igual a γ) e, portanto, o limite da eq. (230) converge se usar o esquema de

decaimento de poderes de revelação da eq. (231). Ou seja, o esquema da eq. (231) converge apenas para revelação parcial, não para a revelação total. Muitos outros esquemas de decaimento podem ser sugeridos e analisados dentro dessa metodologia de análise de convergência. Isso é deixado para futuros trabalhos.

Agora serão apresentadas as equações de um processo recombinante. Para analisar a recombinação dos cenários da distribuição de revelações após um sinal S_k, o qual tem um poder de revelação η_k^2, só é necessário analisar o conjunto relevante de cenários mostrado na Figura 49 abaixo.

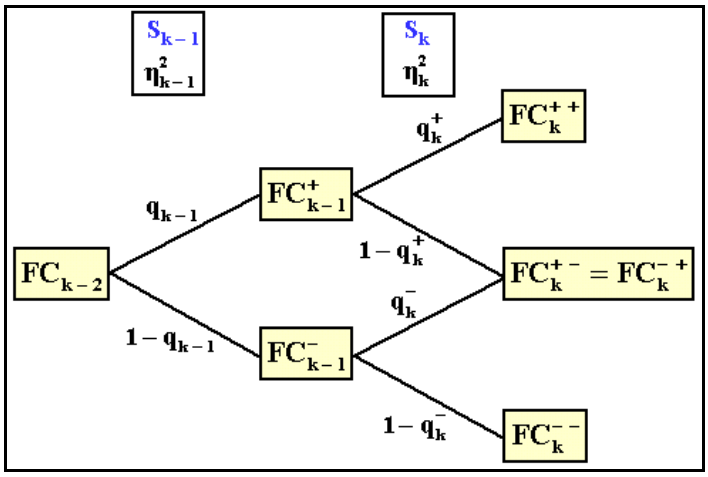

Figura 49 – Processo de Revelação Recombinante para um Sinal S_k

A Figura 49 mostra um cenário da distribuição de revelações após k – 2 sinais, dois cenários da distribuição de revelações após k – 1 sinais e três cenários da distribuição de revelações após o sinal k. A notação é similar às figuras anteriores. Para permitir a recombinação, i. é, que $FC^{+-} = FC^{-+}$, primeiro deve-se escrever as equações para esses cenários no caso geral, baseada nas eqs. (199) e (198) do Teorema 4 (a), que podem ser usadas recursivamente:

$$FC_k^{+-} = FC_{k-1}^{+} - \sqrt{\frac{q_k^+}{1-q_k^+}}\ \sqrt{FC_{k-1}^{+}\,(1-FC_{k-1}^{+})}\ \sqrt{\eta_k^2} \qquad (234)$$

$$FC_k^{-+} = FC_{k-1}^{-} + \sqrt{\frac{1-q_k^-}{q_k^-}}\ \sqrt{FC_{k-1}^{-}\,(1-FC_{k-1}^{-})}\ \sqrt{\eta_k^2} \qquad (235)$$

Essas equações são gerais, valem para processos recombinantes ou não, intercambiáveis ou não. Se forem recombinantes, então as eqs. (234) e (235) terão de ser iguais. Como discutido quando se comparou a Figura 48 com a Figura 47, se for assumida a intercambialidade entre FC e S, para recombinar os cenários terá de ser ajustado o poder de revelação do sinal S_k para um valor η_k^2 menor que o

poder de revelação do sinal anterior η^2_{k-1}. A alternativa seria relaxar a premissa de intercambialidade e escolher as probabilidades do sinal k ser positivo, q^+_k e q^-_k combinado ou não com uma escolha do poder de revelação η^2_k (que nesse caso pode ser igual ao anterior). Ou seja, existem infinitas combinações η^2_k, q^+_k e q^-_k que fariam as eqs. (231) e (232) serem iguais e o processo recombinar. A mais razoável, por tudo que foi discutido antes[252], parece ser a premissa de intercambialidade que dessa forma reduziria as equações anteriores para equações similares as eqs. (214) e (213) da Proposição 7, que nesse caso ficariam:

$$\mathbf{FC^{+\,-}_k = \ FC^+_{k-1} \ - \ FC^+_{k-1} \ \eta_k} \tag{236}$$

$$\mathbf{FC^{-\,+}_k = \ FC^-_{k-1} \ + \ (1 - FC^-_{k-1}) \ \eta_k} \tag{237}$$

Para recombinar basta igualar as eqs. (236) e (237) e lembrar que, pela eq. (215), para processos intercambiáveis vale a equação simples:

$$\mathbf{FC^+_{k-1} - FC^-_{k-1} \ = \ \eta_{k-1}} \tag{238}$$

Logo, igualando as eqs. (236) e (237) e substituindo a eq. (238), obtém-se a relação entre poderes de revelação ou condição de recombinação intercambiável:

$$\mathbf{\eta_k \ = \ \frac{\eta_{k-1}}{1 + \eta_{k-1}}} \tag{239}$$

Que é uma equação bastante simples. Para o caso não trivial, $\eta_{k-1} > 0$, pode-se ver facilmente que $\eta_k < \eta_{k-1}$. Além disso, pode-se tirar o valor de η_k em função da medida de aprendizagem inicial η_1. Para isso, basta substituir repetidamente a eq. (239), que se obtém a equação (para k = 1, 2, …):

$$\mathbf{\eta_k \ = \ \frac{\eta_1}{1 + (k-1)\,\eta_1}} \tag{240}$$

No exemplo baseado em Wang et al (2000), que foi discutido comparando a Figura 48 com a Figura 47, pode agora ser concluído usando a eq. (239) ou a eq. (240) para mostrar que a medida de aprendizagem que faz o processo recombinar é $\eta_2 = 28{,}6\%$. Assim, o processo recombinante intercambiável *modificado* (mais consistente) de Wang et at (2000) é apresentado na Figura 50.

[252] Pode ser interessante estudar processos que violem essa condição, mas de forma controlada (fugindo pouco da intercambialidade) e que tenham alguma vantagem. Nesse caso, sugere-se ter uma boa teoria ou importantes eventuais vantagens que justifiquem o abandono da simplicidade da intercambialidade e suas grandes vantagens matemáticas e intuitivas.

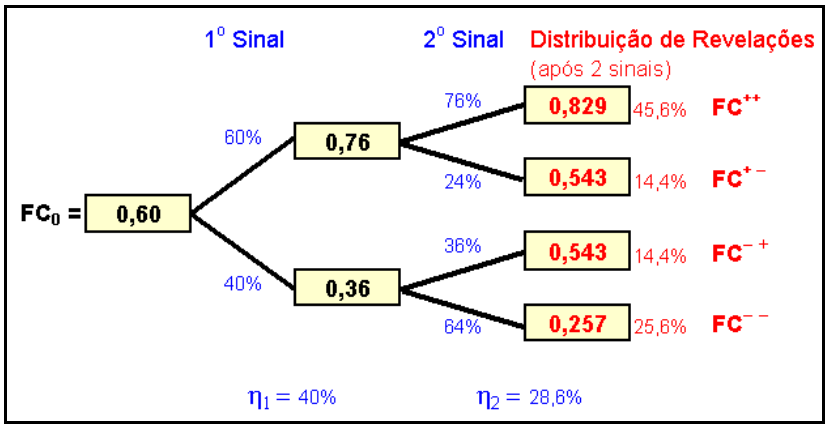

Figura 50 – Processo de Revelação Decrescente Recombinante

Note que as probabilidades dos cenários revelados e dos sinais são as mesmas nas três figuras (47, 48 e 50), apenas os valores dos cenários revelados (após o segundo sinal) é que são diferentes, a depender do poder de revelação do segundo sinal.

Com o esquema da eq. (240), os fatores de redução da variância média posterior ao sinal k (fatores do produtório da eq. 226), $(1 - \eta_k^2)$, rapidamente tendem a 1, pois esse fator é:

$$1 - \eta_k^2 = 1 - \frac{\eta_1^2}{[1 + (k-1)\,\eta_1]^2} \qquad (241)$$

Assim, o denominador da fração da eq. (241) rapidamente (ao quadrado) vai a infinito quando k vai a infinito e, portanto, a fração vai a zero e o fator vai a 1. Assim, a variância média posterior não converge a zero. Para ilustrar, a Figura 51 mostra o percentual acumulado de redução da variância média posterior (i. é, o produtório da eq. 225) após k sinais que decaem conforme a eq. (240).

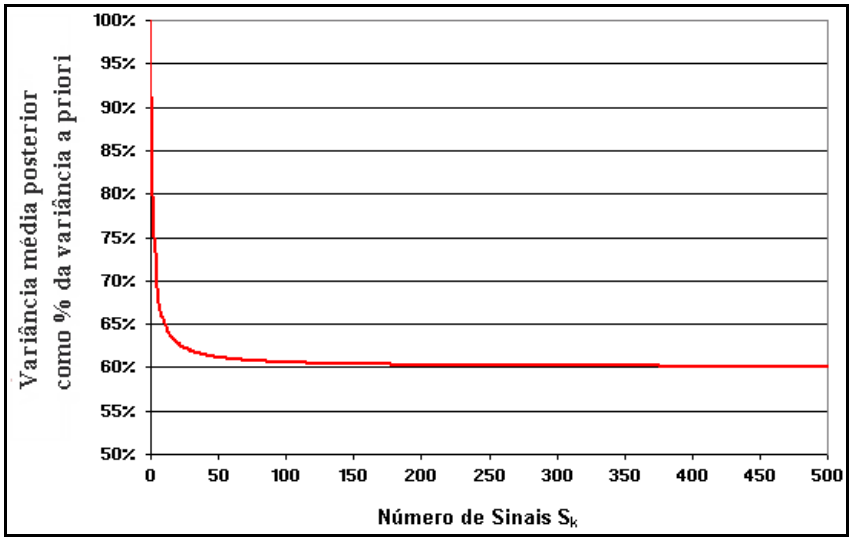

Figura 51 – Fator Redutor da Variância Esperada no Processo Recombinante

Na Figura 51 foi usada $\eta_1 = 40\%$, como em Wang et al (2000). Assim, existe uma grande redução de variância no início, mas depois de um número razoavelmente grande de sinais (~50), a redução adicional de incerteza esperada é muito pequena e converge de forma assintótica para um valor de cerca de 60% da variância inicial (variância da distribuição a priori), i. é, uma redução de 40%. Em outro ângulo, a variância da distribuição de revelações converge para 40% da variância da distribuição a priori, i. é, 40% de $FC_0\ (1 - FC_0)$.

Não foi coincidência que, para $\eta_1 = 40\%$, o limite do produtório da eq. (226) foi $1 - \eta_1 = 60\%$. Uma álgebra tediosa mostra que esse limite é simplesmente:

$$\lim_{n \to \infty} \prod_{k=1}^{n} \left(1 - \frac{\eta_1^2}{[1 + (k-1)\,\eta_1]^2} \right) = 1 - \eta_1 \qquad (242)$$

A Figura 52 ilustra o caso recombinante e intercambiável para 5 sinais seqüenciais, mostrando as 5 distribuições de revelações, tanto os valores dos cenários (FC revelados, dentro das células amarelas) como as probabilidades de ocorrência desses cenários (em azul, debaixo de cada cenário), que somam 1.

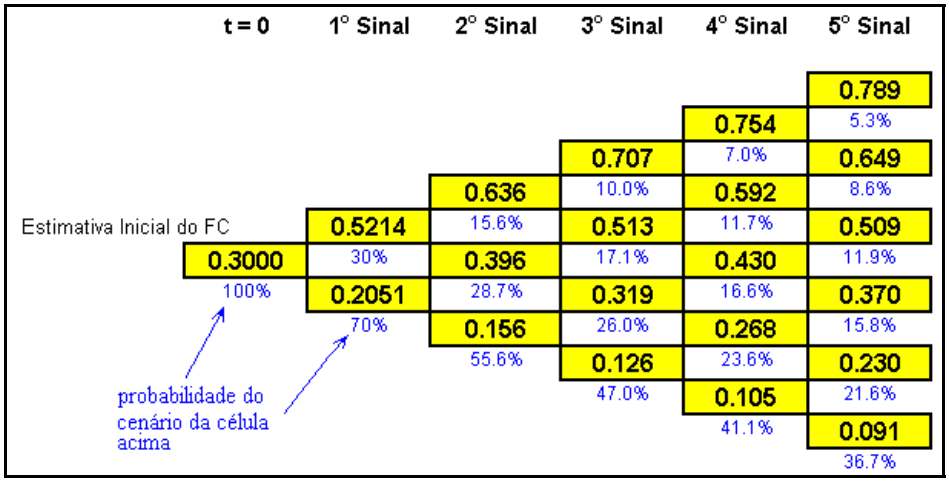

Figura 52 – Processo de Revelação de Bernoulli Recombinante com 5 Sinais

Nesse processo, a medida de aprendizagem inicial foi relativamente pequena, $\eta_1^2 = 10\%$. Um valor inicial maior para a medida de aprendizagem pode fazer o processo ficar próximo da revelação total e dessa forma pode-se observar o que ocorre com a distribuição de revelações perto desse limite. A Figura 53 ilustra isso mostrando o histograma da distribuição de revelações, após 10 sinais (logo, a distribuição de revelações tem 11 cenários), mas com um valor maior para a medida de aprendizagem inicial, no caso com $\eta_1^2 = 80\%$, mas mantendo o mesmo fator de chance inicial ($FC_0 = 30\%$).

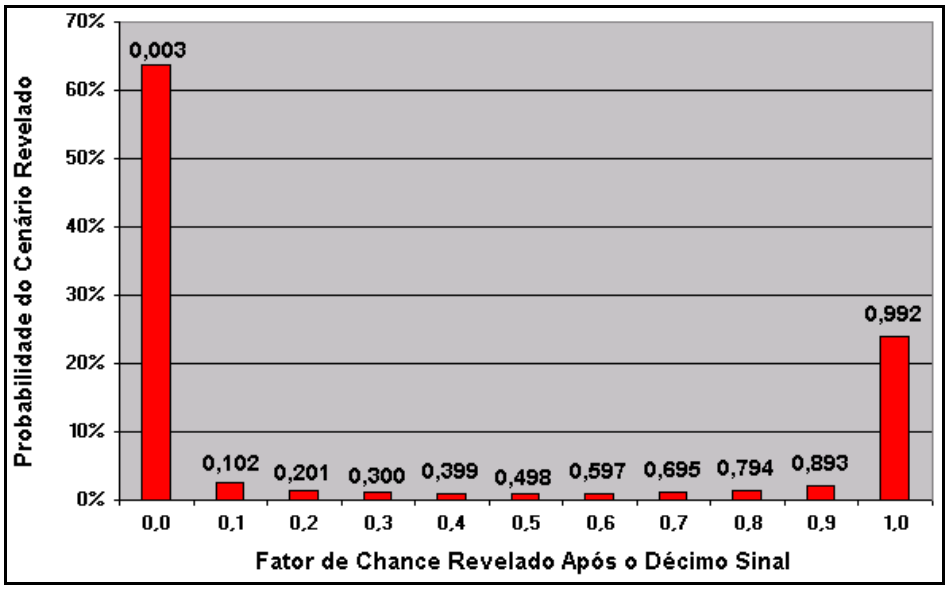

Figura 53 – Distribuição de Revelações de Bernoulli Após 10 Sinais

A Figura 53 é instrutiva no sentido de ilustrar o Teorema 1 (a), i. é, como uma distribuição de revelações converge no limite para a distribuição a priori, no caso para uma distribuição de Bernoulli (dois cenários apenas). É possível intuir isso, apesar do processo acima só ficar perto, não convergindo totalmente, e apesar de usar apenas 10 sinais. Para ver isso, note que os cenários de FC intermediários têm baixas probabilidades de ocorrência (no limite de revelação total teriam probabilidades iguais a zero), enquanto que os cenários extremos de revelação total (perto de FC = 0 e perto de FC = 1) tem elevadas probabilidades. Além disso, note que as *probabilidades* de ocorrência desses cenários extremos se aproximam, respectivamente, de $(1 - FC_0) = 70\%$ e de $FC_0 = 30\%$. Ou seja, apesar da distribuição de revelações ter k + 1 cenários, no limite haverão k – 1 cenários com probabilidade zero e apenas dois cenários (FC = 0 e FC = 1) com probabilidades positivas, que são exatamente os cenários da distribuição a priori, uma Bernoulli com parâmetro $FC_0 = 30\%$. É o que prevê o Teorema 1 (a)[253].

É oportuno apresentar as equações recursivas para obter as probabilidades dos cenários da distribuição de revelações (mostradas em azul na Figura 52). O caso de processo *não*-recombinante é muito simples, pois basta multiplicar as probabilidades dos caminhos. Por ex., na Figura 47, a probabilidade do cenário $FC^{+-} = 14{,}4\%$ (em vermelho) é obtida pela multiplicação das probabilidades dos

[253] Esses exemplos e um código VBA para gerar cenários e probabilidades da distribuição de revelações recombinante, podem ser simulados com a planilha proc_Bernoulli_recombina.xls do CD-Rom (número de sinais possíveis é limitado apenas pelo número de colunas do Excel).

caminhos (= probabilidade dos sinais) que chegam até esse cenário, i. é, as probabilidades (em azul) 60% e 24%.

O caso recombinante é que é um pouco mais complicado, pois na maioria dos cenários, existem vários caminhos que levam a um cenário específico (efeito da recombinação). Mas, tendo apenas as probabilidades dos dois cenários *imediatamente anteriores* que levam ao cenário de interesse, pode ser calculada a probabilidade desse cenário da forma simples a seguir. Seja "i" o indexador que dá a posição vertical da célula de cima para baixo, por ex., na Figura 50 o cenário FC^{++} tem i = 1, $FC^{+-} = FC^{-+}$ tem i = 2 e FC^{--} tem i = 3. Como antes, a posição horizontal é dada simplesmente pelo indexador "k" (cenários após o k-ésimo sinal). Com a ajuda da Figura 49, é fácil concluir que as probabilidades dos cenários da distribuição de revelações $p_{i,k}$ são obtidas recursivamente com a seguinte equação, sendo i – 1 o cenário mais acima, i o cenário mais abaixo e quando não existir a probabilidade é zero:

$$\mathbf{p_{i,k} = p_{i-1,k-1}\,(1 - FC_{i-1,k-1}) + p_{i,k-1}\,FC_{i,k-1}} \qquad \mathbf{(243)}$$

Um ponto importante em um processo de revelação de Bernoulli, pelo menos em aplicações de exploração de petróleo, é que a escala de indexação por eventos (sinais) é *muito* diferente da escala de tempo. Observando um processo de *difusão de eventos* como o da Figura 52, em caso de ocorrência de um resultado positivo proveniente de um sinal qualquer, aumenta aumenta a chance de haver *logo* um novo sinal, um efeito cascata. Isso porque a revisão proporcionada no FC de prospectos de uma bacia, aumenta a chance de exercício da opção de perfurar prospectos exploratórios na área, que serão novos sinais para os prospectos restantes correlacionados na bacia. No entanto, em caso de fracasso, além dessa aceleração de investimentos exploratórios não ocorrer, ainda reduz as chances de novos exercícios de opções naquela área e/ou bacia. Assim, o tempo corre mais rápido na parte de cima do processo de difusão da Figura 52 do que na parte de baixo. Isso é mais um motivo para classificar de grosseiros os modelos de incerteza técnica que substituem o índice de eventos por índice de tempo.

Processos de revelação de Bernoulli é um tema fascinante do ponto de vista teórico e de grande importância prática. Aqui foi apresentado apenas um pouco mais do que o necessário para as aplicações que serão vistas no cap. 5. Esse é um campo aberto para futuras pesquisas, que aqui foram apenas iniciadas.

3.4.4.
Inferência para Distribuição de Bernoulli e Suas Limitações

No item 3.4.1 foi recomendado que o fator de chance exploratório fosse estimado decompondo o mesmo em novos fatores, também v.a. de Bernoulli, e depois eventualmente em novos sub-fatores. Dessa forma, os fatores de chance podem ser ligados mais facilmente a hipóteses geológicas que contribuem para estimar a chance da descoberta. Dentro dessa abordagem, são usados programas estatísticos para estimar esses fatores usando abordagens populares tais como regressões, etc. Isso é feito rotineiramente por companhias de petróleo. Foi visto que os mesmos métodos (e dados) podem ser usados para estimar estatisticamente a medida η^2 proposta.

O propósito desse item é analisar métodos mais diretos de estimar o fator de chance exploratório FC e o poder de revelação η^2 de um sinal como sísmica 3D ou poço vizinho. No caso do FC, uma alternativa é observar dados binários, i. é, amostras da distribuição empírica do parâmetro θ, o estimador da probabilidade de sucesso p da distribuição teórica de Bernoulli Be(p). Isso é feito de uma forma *crítica* a essas abordagens. Além disso, será dado um exemplo de estimativa de η^2 com dados *reais*[254].

Nesse item será mostrado o método quantitativo de inferência Bayesiana para estimar a probabilidade de sucesso de um fator de chance (distribuição de Bernoulli), assim como será discutida a limitação dos métodos tradicionais de inferência estatística. O caso real que será analisado é uma estimativa de poder de revelação de um sinal de sísmica 3D, i. é, uma estimativa de η^2, a medida de aprendizagem proposta. Para tal, serão usados dados ex-ante e ex-post de fatores de chance de prospectos *reais*, estimados antes e depois de uma sísmica 3D.

Na inferência estatística Bayesiana é muito usado o conceito de *distribuições conjugadas* (mencionada no item 3.1.4.2), o qual simplifica bastante o trabalho de estimativa da distribuição a posteriori. Quando as distribuições a priori e a posteriori pertencem à mesma *família de distribuições*, essas distribuições são chamadas de conjugadas. No caso da distribuição de Bernoulli de parâmetro p, se a estimativa a priori desse parâmetro é θ_0 e a estimativa a

[254] Esse item é empírico e pode ser pulado se o interesse for apenas na metodologia.

posteriori é θ_1, i. é, o novo fator de chance (posterior) será também uma distribuição de Bernoulli mas com parâmetro $p = \theta_1$. No entanto, sendo o foco da inferência a estimativa do valor desse parâmetro p, considerado fixo mas incerto (ex., a probabilidade p de dar "cara" numa moeda *não-confiável*), normalmente se usa o conceito de distribuições conjugadas para relacionar a distribuição a priori desse parâmetro p com a distribuição posterior desse parâmetro após o aprendizado com a incorporação de novos dados (*atualização Bayesiana*).

No caso da distribuição de Bernoulli de parâmetro p, é usada a distribuição conjugada Beta para modelar a incerteza sobre o verdadeiro valor de p. A distribuição Beta (B) tem dois parâmetros $B(\gamma, \eta)$ e é muito flexível. Dada uma distribuição Beta que reflita a incerteza a priori do verdadeiro parâmetro p de uma distribuição de Bernoulli, após um novo dado, a distribuição posterior da incerteza desse parâmetro é também uma distribuição Beta com novos parâmetros γ_1 e η_1, isto é, $B(\gamma_1, \eta_1)$ que tem menor variância (refletindo o aprendizado sobre p com o novo dado). As distribuições de Bernoulli e Beta são ditas conjugadas pois a distribuição Beta – que tem intervalo [0, 1] e assim é apropriada para representar a incerteza do parâmetro p duma distribuição de Bernoulli, permite uma atualização Bayesiana que gera uma distribuição posterior que também é uma distribuição Beta, mas com novos parâmetros (prova: DeGroot & Schervish, 2002, p.336, teorema 6.3.1).

Note que, nesse contexto de estimação, existem duas distribuições a priori e duas posteriores: a distribuição a priori da incerteza técnica – no caso o fator de chance que é uma distribuição de Bernoulli, e a distribuição a priori que descreve a incerteza do *parâmetro* da distribuição de Bernoulli – no caso uma distribuição Beta. As distribuições posteriores dessas duas distribuições são respectivamente uma outra distribuição de Bernoulli e uma outra distribuição Beta. A Tabela 12 a seguir (adaptada de Jammernegg, 1988, p.22) resume as distribuições a priori e posteriores da incerteza técnica e da incerteza do parâmetro p, sendo que a atualização Bayesiana é feita com a informação de uma amostra com n observações, das quais y são sucesso (observações de y iguais a 1) e em conseqüência (n – y) observações são falhas (observações de y iguais a zero).

Tabela 12 – Distribuições a Priori e Posterior Conjugadas (Bernoulli e Beta)

Distribuição a Priori da Incerteza Técnica (FC)	Bernoulli: Be(p) Média = p Variância = p (1 – p)
Distribuição a Priori do Parâmetro p	Beta: B(γ, η) Média = γ / (γ + η) Variância = $\gamma\ \eta$ / [(γ + η)^2 (γ + η +1)]
Distribuição Posterior do Parâmetro p	Beta: B(γ + y, η + n – y) Média = (γ + y) /(γ + η+ n) Variância = (γ + y) . (η + n – y) / [(γ + η+ n)^2 . (γ + η + n + 1)]
Distribuição Posterior da Incerteza Técnica (FC)	Bernoulli: Be($E[p]_{post}$) = Be(E[B(γ + y, η + n – y)]) Média = (γ + y) / (γ + η+ n) Variância = (γ + y) . (η + n – y) / (γ + η+ n)^2

Na Tabela 12, pode-se mostrar que a variância da distribuição Beta posterior é sempre igual ou menor que a da distribuição Beta a priori. Isso nem sempre irá ocorrer com a variância da distribuição de Bernoulli posterior se a nova estimativa de p for um valor mais próximo de 0,5 do que o valor de p da distribuição de Bernoulli a priori[255].

No problema de inferência Bayesiana muitas vezes o objetivo é obter uma estimativa do parâmetro p tal que a variância da distribuição posterior seja menor que um certo valor. Para isso, vai se obtendo dados (observações) até que a essa variância seja menor ou igual a certo valor. Por ex., no caso do parâmetro da distribuição de Bernoulli, DeGroot & Schervish (2002, p. 337) assinalam que para reduzir a variância da distribuição Beta posterior do parâmetro p para 0,01 ou menos, não é necessário selecionar mais que n = 22 itens de amostra.

No entanto, pode-se estar interessado na melhor estimativa (um valor único) sem especificar toda a distribuição que representa a incerteza desse parâmetro. Ou seja, para o nível de informação atual, qual o melhor estimador de Bayes? O estimador Bayesiano ótimo é escolhido de forma a minimizar o erro entre a estimativa e o valor verdadeiro desse parâmetro. Para isso terá de ser usado o conceito de *função perda* (mencionada nos itens 3.2.1 e 3.2.3) que mede esse erro

[255] No entanto, foi visto que no *processo de revelação* sobre o fator de chance é resgatada a propriedade de redução de variância para a variância *média* das distribuições posteriores de Bernoulli, refletindo a *aprendizagem em direção à verdade*.

de estimativa. Conforme foi visto, a função perda mais usada é a função perda de erro quadrático[256]. Essa função é definida por $L(\theta, a) = (\theta - a)^2$, onde θ é o *verdadeiro* valor do parâmetro e "a" é o *estimador* desse parâmetro.

Sendo o objetivo *minimizar o valor esperado* dessa função perda, conforme mencionado antes, prova-se (DeGroot & Schervish, 2002, p.224) que o valor de a que minimiza E[L(.)] é exatamente a *média da distribuição posterior*, i. é, cenários da distribuição de *expectativas condicionais* $E[\theta \mid S]$. Já no processo de revelação, como foi visto, para avaliar ex-ante um investimento em informação se trabalha também com uma distribuição de expectativas condicionais, mas a de revelações E[FC | S] e não a distribuição do parâmetro p da distribuição original[257]. É importante lembrar que as decisões de investimento exploratório (cálculo do VME) usam a distribuição de Bernoulli e não a Beta.

O exemplo a seguir irá ilustrar a inferência Bayesiana do parâmetro p da distribuição de Bernoulli usando distribuições conjugadas Beta para estimar esse parâmetro (baseado em DeGroot & Schervish, 2002, pp.336-337). Suponha que não haja qualquer informação a priori sobre o valor desse parâmetro p de uma distribuição de Bernoulli sobre um problema qualquer (exs.: proporção de peças defeituosas num lote, sub-fator de um FC dum prospecto exploratório, etc.). Assim, como distribuição a priori desse parâmetro foi adotada uma distribuição não-informativa, a distribuição uniforme U(0, 1), que é também uma distribuição Beta com parâmetros $\gamma = 1$ e $\eta = 1$, isto é, B(1,1), o que mostra a versatilidade da distribuição Beta. Suponha que se obteve uma amostra com n observações (n peças inspecionadas, n poços perfurados, etc.), das quais se contabilizou y "sucessos" (e logo n − y falhas). Qual é o estimador de Bayes atualizado do parâmetro p da distribuição de Bernoulli usando essas n observações? Utilizando a Tabela 12 é fácil ver que a estimativa de p dado essas n observações é a média da distribuição Beta posterior, i. é, $E[p]_{post} = (1 + y) / (2 + n)$.

[256] A alternativa de função perda igual ao *erro absoluto* (em vez de erro quadrático) levaria a usar um estimador Bayesiano igual a *mediana* da distribuição posterior. Além de ser bem menos usada em estimação, a mediana não é linear como o operador valor esperado E[.] e é inútil no contexto de *finanças* que trabalha com valores esperados de fluxos de caixa.

[257] Por ex., distribuição das médias de distribuições de Bernoulli posteriores e não da distribuição Beta do parâmetro p. No entanto, como as médias das possíveis distribuições posteriores Beta são usadas como estimadores para as médias das possíveis distribuições de Bernoulli, então as distribuições das médias das duas distribuições posteriores são iguais (já a variância média das distribuições a posteriori de Bernoulli e Beta são em geral diferentes).

Qual seria a estimativa estatística *clássica* (não-Bayesiana) desse exemplo? Na escola *frequentista* clássica não se usa a distribuição a priori para o parâmetro, apenas as observações da amostra. O método clássico mais usado é o *estimador de máxima verossimilhança*, isto é, o estimador que maximiza a função verossimilhança. Para a distribuição de Bernoulli o estimador de máxima verossimilhança é simplesmente a média da amostra (prova: ver DeGroot & Schervish, 2002, p.357). No caso do exemplo do parágrafo anterior, essa estimativa seria simplesmente y / n. Note que, quando o tamanho da amostra n cresce, o estimador Bayesiano tende ao estimador de máxima verossimilhança, o que é um resultado típico e geral.

Apesar da larga aplicabilidade desses métodos em diferentes problemas práticos, existem importantes aplicações em que os mesmos deixam a desejar. Em geral se assume que na amostra as observações são *independentes e identicamente distribuídas* (iid). Assim, se existe dependência entre as observações ou se as distribuições de Bernoulli da amostra tem diferentes parâmetros p, a amostra teria de ser reduzida para uma quantidade de observações realmente iid, sendo que nessa redução de amostra pode-se estar jogando fora informação relevante[258]. Outro problema é como estimar um parâmetro (ex.: fator de chance exploratório) através de uma observação *indireta*, por ex., uma sísmica 3D na área do prospecto de interesse em vez do resultado de um poço exploratório vizinho. Não se pode simplesmente somar a observação da sísmica como sendo mais uma observação de uma amostra que inclui resultados de poços exploratórios e aplicar as metodologias de inferências acima descritas. Para esses problemas será necessário considerar os diferentes poderes de revelação de cada informação (ou observação) adicional. Esse *poder de revelação* é definido como a capacidade de reduzir a variância esperada das possíveis distribuições posteriores, i. é, η^2.

[258] Ex: quer se avaliar o FC de um prospecto de idade do Eoceno numa bacia em que 25 poços foram perfurados com 5 sucessos (20% de sucesso). Se desses 5 sucessos, 2 ocorreram em prospectos do Eoceno e se 6 prospectos foram perfurados com objetivo no Eoceno, então o FC do prospecto aumenta. Se em adição um dos sucessos se deu numa área próxima (ex., a 5 km) do prospecto, e foi o único poço do Eoceno perfurado num raio de 5 km, então o FC do prospecto aumenta ainda mais. Usar apenas essa última informação poderia ser uma saída, mas se estaria jogando fora a informação de outros poços mais distantes do Eoceno que são relevantes, embora menos do que o do poço mais próximo. O *poder de revelação* de cada observação não é bem capturado nem com a metodologia clássica de inferência e nem com a metodologia tradicional Bayesiana. Com a medida η^2, pode-se modular as diferenças de relevância dos vários sinais.

O exemplo real a seguir, da *literatura profissional* de petróleo, permite fazer uma estimativa da medida de aprendizagem η^2 no caso da sísmica 3D. Esse exemplo também mostrará a insuficiência dos métodos estatísticos de inferência para avaliar o impacto econômico duma nova informação no valor de um projeto com incerteza técnica.

Aylor Jr. (1999) reporta um progresso substancial, em termos de sucesso exploratório, da firma Amoco a partir de 1994 com o uso da então nova tecnologia[259] de sísmica 3D. Ele relaciona o benefício da sísmica 3D diretamente à capacidade de separar bons prospectos de maus prospectos, reportando o efeito da sísmica 3D na probabilidade de sucesso de existência de rocha selante (ou selo, ver item 3.4.1) em 8 prospectos num bloco do Mar do Norte:

"*Uma ilustração muito boa da habilidade da pesquisa 3D em separar bons prospectos exploratórios de maus prospectos ... um bloco exploratório no Mar do Norte mostra 8 prospectos com probabilidade de falha de 'selo' geralmente entre 20% e 50% antes de adquirir dados 3D. Depois do 3D, dois dos prospectos tem sido convincentemente confirmados com 10% de probabilidade de falha (P(f)), e seis tem sido condenados com P(f) de 80-90%. ... Os prospectos condenados têm economizado caros poços secos. Os (prospectos) confirmados podem ser perfurados com baixo risco de falha.*"

Aylor Jr. reporta a probabilidade p de sucesso na existência de rocha selante avaliada por uma equipe de técnicos da exploração da Amoco antes e depois da sísmica 3D. Esses dados reais são mostrados na Tabela 13.

Tabela 13 - Probabilidades de Sucesso dos Prospectos Terem Selo

Prospecto	1	2	3	4	5	6	7	8
Probabilidade de Sucesso de Selo Antes da Sísmica 3D	**70 %**	50%	50%	50%	80%	90%	50%	75%
Probabilidade de Sucesso de Selo Depois da Sísmica 3D	10%	10%	10%	20%	90%	10%	20%	90%

Note que o impacto da informação proveniente da sísmica 3D não foi ter uma melhor avaliação do parâmetro p da distribuição de Bernoulli, ou seja, não

[259] A sísmica tradicional de reflexão 2D é usada há várias décadas, mas a sísmica 3D só começou a ser usada nos anos 90.

foi reduzir a variância da distribuição Beta a priori da incerteza desse parâmetro. O impacto foi "separar o joio do trigo", isto é, fazer com que cada prospecto ficasse ou com alta chance de sucesso (p mais próximo de 100%, o que ocorreu com dois prospectos, veja tabela) ou com alta chance de falha (p mais próximo de 0%). Em outras palavras, o processo de aprendizagem caminhou no sentido da revelação total (embora tenha ficado longe disso, pois precisaria que parte dos prospectos ficasse com FC = 0 e a outra parte com FC = 1). Ou seja, o parâmetro p da distribuição de Bernoulli não é fixo e sim evolui (aumenta ou diminui) com o aprendizado trazido com a nova informação. Assim o estimador Bayesiano tradicional com o uso da distribuição Beta para estimar o parâmetro p da distribuição de Bernoulli, etc., é inútil nesse caso[260].

Ainda é possível usar a Lei de Bayes para estimar a revisão de probabilidades de sucesso dada as informações da sísmica, como fazem alguns autores. No entanto, deve-se notar um fato importante com os dados da Tabela 13 que é o *processo de redução da incerteza* (variância) com a informação proveniente da sísmica 3D. Aplicando a fórmula da variância da distribuição de Bernoulli, pode-se verificar que a variância *média* das distribuições de Bernoulli a priori (antes da sísmica 3D) era de 0,20594 $\%^2$, enquanto que a variância média das distribuições de Bernoulli posteriores (depois da sísmica 3D) caiu para 0,10750 $\%^2$, uma redução de quase 50% na variância a priori[261]. Ou seja, a estimativa da medida de aprendizagem $\eta^2 = 47{,}8\%$ ou $\eta = 69{,}1\%$. Note que esse valor não é para o fator de chance total e sim para *um dos fatores* que compõe esse FC. O impacto no FC total em termos de η^2 deve ser menor que 47,8%, a menos que os outros fatores sejam iguais a 1 (haja certeza nos outros fatores).

O exemplo real mostrou que a sísmica 3D proporciona uma revelação parcial sobre o FC nos prospectos, que está associada a um processo de redução de incerteza em direção à verdade sobre cada prospecto (p = 100% ou p = 0%). O valor obtido para a da medida de aprendizagem η^2 deve ser olhado com cuidado já que foi apenas um caso em que dados reais foram reportados (o que é raro).

[260] No caso do problema de exploração de petróleo o parâmetro de Bernoulli muda com a informação. É um caso diferente do problema mais simples de estimar a percentagem de produtos defeituosos num lote através de amostras aleatórias desse lote. Se a retirada de amostras for com reposição (ou se o lote for muito grande) a probabilidade de Bernoulli não muda.

[261] Lembrar que nem sempre a variância é reduzida com a informação, mas é um resultado típico, pois em *média* a variância se reduz (no mínimo nunca aumenta) com a nova informação.

Outras estimativas podem e devem ser feitas por companhias de petróleo baseadas nos seus próprios bancos de dados, não só para o fator de existência da rocha selante, como para os outros fatores discutidos no item 3.4.1.