

1

Introdução

Técnicas estatísticas e matemáticas são ferramentas úteis para o desenvolvimento, melhoria e otimização de processos produtivos. Essas técnicas são igualmente importantes para o projeto e desenvolvimento de novos produtos, bem como para a melhoria de produtos existentes.

Nesse contexto, é particularmente importante aplicar essas técnicas na identificação de fatores que influenciam alguma **característica de qualidade** de processos ou produtos. É comum representar uma característica de qualidade por uma variável, denominada variável **resposta**. Igualmente, é comum representar um fator por uma variável, denominada variável **independente**.

A variável resposta pode ser contínua ou discreta. As variáveis contínuas são utilizadas quando a característica de qualidade pode ser representada por uma grandeza definida num intervalo dos números reais, como é o caso das medidas de comprimento, superfície, volume, resistência, velocidade, voltagem, pureza, etc.

Muitas das características de qualidade não podem ser representadas por medidas numéricas de grandezas físicas ou químicas. Em tais casos, usualmente classificam-se os produtos inspecionados como **conforme** ou **não conforme** com as especificações. A terminologia “**não defeituoso**” e “**defeituoso**” é freqüentemente empregada para estas duas classificações do produto. Um exemplo típico é o de um dispositivo que funciona ou não. Características de qualidade desse tipo são denominadas de **atributos**. A variável resposta é o **número de unidades não conformes** por amostra inspecionada.

Outro tipo característica de qualidade discreta é o **número de defeitos por unidade de inspeção**. A contagem de defeitos é feita em um intervalo contínuo (uni, bi, ou tridimensional). Um exemplo típico é o de um produto onde se observa o número de defeitos por unidade de comprimento, como é o caso de cabos, tecidos, etc.

Quanto às variáveis independentes, ou fatores que influenciam as características de qualidade, podemos classificá-los como **quantitativos** ou **qualitativos**. Um fator quantitativo é aquele cujos níveis (valores) podem ser associados com pontos de uma escala numérica, como temperatura, pressão, porcentagem, etc. Fatores qualitativos são aqueles em que não é possível esta associação. Máquinas, operadores, tipos de materiais, são fatores qualitativos porque não correspondem a valores numéricos, ainda que se possa representá-los por números (Operador 1, Operador 2), por exemplo. Esses números não representam quantidades, sendo apenas códigos identificadores.

Quando, para uma mesma combinação de valores das variáveis independentes, não se pode prever com certeza qual será o resultado da variável resposta, esta é uma variável aleatória.

Podem-se usar modelos matemáticos para medir a influência das variáveis independentes sobre a variável resposta, quando então se estabelece uma relação entre elas. Podem-se, também, aplicar testes estatísticos para avaliar quais são as variáveis independentes cuja influência sobre a variável resposta é significativa.

Com os modelos podemos estabelecer qual é a combinação de níveis das variáveis independentes que resulta em uma resposta desejada, que pode corresponder, quase sempre, às vizinhanças de um valor máximo, de um valor mínimo ou de um “valor-alvo”.

Para construir esses modelos e aplicar os testes, são realizados experimentos. Nos experimentos se estabelecem os níveis das variáveis independentes e se observa o valor da variável resposta. Com esses dados pode-se construir o modelo e aplicar os testes de significância dos parâmetros do modelo.

Uma estratégia de experimentação é estabelecer os níveis dos fatores de um modo arbitrário, com base em palpites; experimentar e medir a resposta. Esta prática é comum entre engenheiros. Muitas vezes se obtêm bons resultados devido aos conhecimentos teóricos e tecnológicos, bem como experiência prévia sobre o processo ou produto em questão. Entretanto, se um palpite não der um resultado satisfatório, experimenta-se outro palpite, e assim por diante. Isso pode continuar por muito tempo sem que haja garantia de sucesso. Ademais, pode-se chegar a um

resultado satisfatório, interromper os experimentos, sem que possa saber se não haveria uma solução melhor ainda.

Outra estratégia de experimentação é variar um fator de cada vez. Nessa estratégia se estabelece um valor inicial para o nível de cada fator. Então, escolhe-se um fator, passa-se a variar o nível desse fator, enquanto o nível dos outros fatores é mantido constante. Para cada nível variado observa-se a resposta. Isso é feito sucessivamente com todos os fatores. Após a realização dos testes, tem-se uma medida do efeito da variação do nível de cada fator sobre a resposta. O principal problema com esta estratégia é que ela não detecta interação entre fatores. Uma interação ocorre entre dois fatores quando um fator não produz um mesmo efeito na resposta quando o outro fator está em diferentes níveis. Ademais, esta estratégia é menos eficiente (i. e. demanda maior número de observações) do que outras baseadas em experimentos fatoriais.

A estratégia recomendável é a baseada em experimentos fatoriais. Nesse tipo de experimentos os fatores variam *juntos* de um teste para outro. Eles têm propriedades que permitem, entre outros objetivos desejáveis, a construção de modelos que se ajustam bem aos dados e que reduzem ao mínimo as probabilidades de erros dos testes de significância dos parâmetros do modelo. Livros excelentes sobre o planejamento e análise de experimentos fatoriais são fornecidos como referência no final deste capítulo.

Na otimização de processos produtivos e de projetos de produtos, os métodos inicialmente utilizados tinham como foco principal a identificação de fatores que afetam a **média** da variável resposta, produzindo nestes efeitos denominados efeitos na posição (*location effects*). Nesses métodos iniciais, a variância da resposta era considerada constante, ou seja, não era afetada pelos fatores nem variava quando a média mudava de valor. Mais recentemente passou-se a utilizar métodos para identificar fatores que afetam a **variância** da resposta, produzindo nestes efeitos denominados efeitos na dispersão (*dispersion effects*).

Dois tipos de fatores podem ser considerados. **Fatores de controle** são aqueles que se pode manter em um dado nível durante o processo de produção. **Fatores de ruído** são aqueles que oscilam em torno de um dado nível durante o processo, embora possam ser controlados durante a fase experimental do projeto do processo. Os dois fatores podem produzir efeitos na média e na dispersão.

Quando se faz o ajuste no modelo não é necessário distinguir os dois tipos de fatores. Apenas na análise dos resultados será levada em conta a distinção.

Na década de 80 o consultor Genichi Taguchi caracterizou os fatores de ruído como importantes para a análise de efeitos na dispersão. Em sua conceituação é necessário identificar os fatores de controle que afetam a dispersão da resposta e, então, o nível do fator deve ser ajustado de modo que resulte em produtos e processos robustos às variações dos fatores de ruído. Os conceitos por ele estabelecidos são considerados de grande importância, embora as técnicas estatísticas de análise por ele propostas tenham sido criticadas por vários autores e desde então tenham surgido técnicas mais eficientes. Foi cunhado o termo **Projeto Robusto de Parâmetros** (*Robust Parameter Design*) para caracterizar a busca das condições ótimas no desenvolvimento de produtos e processos produtivos, i. e., menos sensíveis aos fatores de ruído.

Na identificação dos fatores importantes é desejável utilizar experimentos replicados porque, com replicações, pode-se construir um modelo para a dispersão, através da variância das observações replicadas, independentemente do modelo utilizado para a média. Não havendo replicações, o modelo para a dispersão é obtido através dos resíduos provenientes do modelo da média, o que exige uma identificação correta dos fatores que afetam a média, sob pena de identificação de efeitos espúrios na dispersão. Entretanto, experimentos replicados podem ter alto custo ou longo tempo de experimentação. Daí o interesse em experimentos sem replicações.

A investigação dos fatores que afetam um processo e a identificação das condições ótimas de operação do processo geralmente atravessa vários estágios. No início, freqüentemente o número de fatores é elevado e as condições de operação geralmente estão longe das condições ótimas. Nestas circunstâncias, são indicados os modelos de primeira ordem (polinômios do primeiro grau) para a resposta e, conseqüentemente, experimentos fatoriais em dois níveis, completos (2^k) ou fracionados (2^{k-p}). Neste estágio seriam descartados os fatores que não afetam a média nem a dispersão e identificado qual seria o próximo estágio da investigação, deslocando a região de experimentação em direção às vizinhanças do ótimo. No estágio final da investigação geralmente torna-se necessário utilizar modelos de segunda ordem (polinômios do segundo grau) e experimentos fatoriais

em três níveis; completos (3^k) ou os denominados Experimentos Compostos Centrados (*Central Composite Design*).

Cabe salientar que, nesse tipo de investigação é muito comum se deparar com escolhas em lugar de respostas. Em muitos estágios, quer seja pela estrutura de “confundimento” associada ao projeto do experimento, quer seja pela própria incerteza inerente a um modelo estatístico, não fica claro quais os fatores que devem ser descartados. Em muitos casos a escolha só é possível com conhecimentos teóricos e tecnológicos, bem como experiência prévia sobre o processo ou produto em questão. Caso contrário, necessita-se de dados adicionais.

O modelo linear clássico, com o método dos mínimos quadrados, tem sido a técnica estatística mais difundida para estabelecer a relação entre as variáveis de um experimento. A sua simplicidade, seu apelo intuitivo e as propriedades do método dos mínimos quadrados explicam a difusão de seu uso. Três considerações são feitas para a aplicação dessa técnica: i) a resposta tem distribuição normal, ii) a variância da resposta é constante e iii) os efeitos dos fatores sobre a variável resposta se combinam aditivamente. Essas condições são encontradas em muitas aplicações industriais. Porém, nem sempre isso ocorre.

Como foi visto, em certos processos a variável resposta é discreta: por exemplo o número de produtos defeituosos em uma amostra aleatória segue a distribuição binomial; e o número de defeitos por unidade de inspeção segue a distribuição de Poisson. Em outros processos a variável resposta segue a distribuição gama, como é o caso da resistividade de semicondutores ou o tempo até a falha de certos dispositivos, ambas variáveis essencialmente positivas, com distribuição assimétrica e longas caudas à direita. Nesses casos (binomial, Poisson e gama), a variância da resposta não é constante e sim função da média. Ademais, nem sempre os efeitos se combinam de uma forma aditiva. Nos exemplos acima citados com distribuições Poisson e gama, geralmente, os efeitos se combinam multiplicativamente.

Caso um ou mais dos três pressupostos — normalidade, aditividade e variância constante — não possam ser considerados, uma alternativa é alterar a escala da variável resposta, transformando-a e usando o modelo linear e o método dos mínimos quadrados com os dados transformados.

A transformação da resposta muitas vezes pode ser satisfatória. Em muitos casos os três pressupostos são alcançados com uma única transformação. Entretanto, nem sempre se consegue isso. Nestes casos devemos usar os modelos lineares generalizados (MLG).

Com os MLG podemos resolver os três problemas independentemente: podemos considerar outras distribuições que não a normal, não é necessário considerar a variância constante (ela pode ser função da média) e conseguimos linearidade através de uma função que faz a ligação entre a média da variável resposta e o polinômio linear das variáveis independentes.

Nos modelos lineares generalizados usam-se as distribuições da família exponencial e a estimativa dos coeficientes do modelo é obtida com a maximização da função de log-verossimilhança. (No caso da distribuição normal, efeitos aditivos e variância constante, a maximização da função de log-verossimilhança leva aos estimadores de mínimos quadrados, de modo que os modelos lineares ordinários podem ser considerados um caso particular dos MLG).

Muitas vezes os membros da família exponencial não são distribuições adequadas para representar a variável resposta. Para esses casos há um método que, para estimar os coeficientes, maximiza a função de quase-verossimilhança (QV). No método da QV não é necessário definir a distribuição de probabilidade; apenas se define uma função entre a média e a variância da variável resposta.

Quando modelamos a média com o método da QV ou dos MLG, a variância é modelada como uma função da média, multiplicada ainda por um *parâmetro de dispersão* constante. Quando os fatores afetam a dispersão, essa relação entre a variância e a média não é suficiente para “explicar” a variância da resposta. Nesses casos — como em certos modelos com as distribuições binomial e Poisson —, existe uma “sobre-dispersão”, que é explicada pelo parâmetro de dispersão, que não é constante e sim função dos fatores. Portanto, os métodos da QV e dos MLG não são indicados para modelar a média nos casos em que há efeitos na dispersão. O método indicado é o da maximização da função de quase-verossimilhança estendida (QVE).

Nesse método o parâmetro de dispersão não é considerado constante e, portanto, é necessário para estimar os coeficientes do modelo da média. As estimativas do parâmetro de dispersão para cada resposta são obtidas através do modelo para a dispersão.

O objetivo desta tese é realizar uma síntese, ou consolidação, das técnicas estatísticas indicadas para a modelagem da média e da variância das características de qualidade de processos e produtos, em experimentos fatoriais não replicados, em um roteiro coerente, detalhado e exaustivo de análise.

A motivação para isso é que, apesar de haver várias publicações sobre regressão linear clássica, modelos lineares generalizados, transformação da resposta e planejamento de experimentos, não existe um texto que reúna e descreva em detalhe todos os aspectos da modelagem da média e da variância em experimentos fatoriais. Os poucos textos sobre esse assunto não descrevem vários aspectos importantes em estudos dessa natureza. Particularmente, não deixam claro como são aplicados os testes de significância dos coeficientes dos MLG, nem quais são as estatísticas e os gráficos indicados para verificar a adequação do modelo. Ademais, nada existe sobre a identificação de modelos para a média em experimentos fatoriais.

Serão descritos os modelos de regressão indicados para a média e a variância das características de qualidade de processos e produtos, em experimentos fatoriais não replicados, os métodos de estimativa dos coeficientes dos modelos, os testes de significância para seleção dos fatores importantes, e quais as estatísticas e os gráficos indicados para verificar a adequação do modelo.

Uma vez construído o modelo, será mostrado como obter as condições ótimas de operação dos processos e produtos.

Contribuição da Pesquisa

1. Reunir em um só texto as técnicas de regressão e de avaliação de modelos, indicadas para a modelagem conjunta da média e da variância das características de qualidade de processos e produtos.

2. Mostrar, através de exemplos, como testar a significância dos coeficientes dos modelos lineares generalizados, aplicando a análise de *deviance*, que está sugerida, mas não detalhada nos textos disponíveis.
3. Mostrar claramente quais as estatísticas e gráficos indicados para verificar a adequação do modelo.
4. Apontar limitações em todos quatro métodos que se propõem a escolher a transformação mais adequada para a resposta. Esses métodos não produziram resultados satisfatórios quando houve interações significativas entre os fatores.
5. Propor o uso dos métodos de transformação da resposta como fonte de indicação da função de ligação a ser usada nos modelos lineares generalizados.
6. Propor o uso da função de log-verossimilhança para uma escolha da distribuição de probabilidade e da função de ligação, nos modelos lineares generalizados.

O restante desta tese é organizada da seguinte forma: os capítulos 2 e 3 fornecem a fundamentação conceitual para a compreensão dos capítulos 4 e 5, que são o cerne da tese. Especificamente: no Capítulo 2 apresentamos o método de regressão linear clássica. Seção 2.1: o método dos mínimos quadrados. Seção 2.2: os testes de significância dos coeficientes. Seção 2.3: as técnicas de verificação da adequação do modelo. Seção 2.4: o método dos mínimos quadrados ponderados. Seção 2.5: a *Forward Search* (FS), que é um procedimento gráfico que monitora os resíduos e outras estatísticas importantes.

No Capítulo 3 apresentamos os modelos lineares generalizados. Na Seção 3.1 apresentamos as distribuições de probabilidade usadas nos MLG. Na Seção 3.2 mostramos a estrutura formal dos MLG. Na Seção 3.3 mostramos como é feita a estimativa dos coeficientes e apresentamos o algoritmo iterativo usado para calcular essa estimativa. Na Seção 3.4 detalhamos como aplicar o teste de significância dos coeficientes. Na Seção 3.5 apresentamos as estatísticas e os gráficos indicados para verificar a adequação do modelo. Na Seção 3.6 descrevemos o método de quase-verossimilhança. Na Seção 3.7 descrevemos o

método de quase-verossimilhança estendida. Na Seção 3.8 descrevemos o método de quase-verossimilhança restrita.

No Capítulo 4 definimos um roteiro de análise que permita a escolha do modelo mais adequado para representar o comportamento da média de uma característica de qualidade de um processo ou produto. Na Seção 4.1 descreveremos tipos de variável resposta que caracterizam a qualidade em produtos e processos produtivos. Na Seção 4.2 consideraremos, para modelagem destas variáveis, a metodologia usada apresentada no Capítulo 2, ou seja: usaremos os modelos lineares gaussianos juntamente com o método dos mínimos quadrados (MQ). Na Seção 4.3 usaremos a alternativa de transformar a resposta para outra escala, usando com os dados transformados ainda os mesmos procedimentos da Seção 4.2. Na Seção 4.4 será descrita a utilização dos modelos lineares generalizados (MLG) apresentados no Capítulo 3, incluindo aí a quase-verossimilhança (QV).

No Capítulo 5 apresentamos métodos para a modelagem conjunta da média e da dispersão. Na Seção 5.1 mostramos a importância da elaboração do modelo para média. Na Seção 5.2 apresentamos um método gráfico muito simples e útil para identificar efeitos na dispersão. Na Seção 5.3 apresentamos o método de Box e Meyer (1986), usado para identificar fatores que afetam a dispersão. Na Seção 5.4, apresentamos o método iterativo de Lee e Nelder (1998). Esses autores aplicam modelos lineares generalizados para modelar tanto a média quanto a dispersão explicitamente. Na Seção 5.5, veremos como analisar experimentos onde são encontrados fatores de ruído. Na seção 5.6 apresentamos um roteiro que sintetiza a metodologia geral para a análise da média e da dispersão apresentada ao longo dos capítulos 4 e 5.

No Capítulo 6 apresentamos as conclusões da tese.

Quanto a revisão bibliográfica (trabalhos precedentes), pela própria natureza particular dessa tese, que pretende ser uma síntese que consolide os métodos existentes em um roteiro coerente e detalhado de análise, ela permeia todo o texto, distribuindo-se nos diversos capítulos conforme o assunto, aspecto, método ou parte da análise.

Serão, de qualquer forma, relacionadas a seguir as principais referências bibliográficas sobre os diversos tópicos, métodos e modelos aqui abordados, à guisa de recomendação para o leitor interessado em um aprofundamento dos mesmos.

Serão também relacionados os softwares utilizados nas análises de casos presentes nesta tese.

Bibliografia Indicada

1. Planejamento e Análise de Experimentos:

Box *et al.* (1978), Box e Draper (1987), Daniel (1976), Montgomery (2001), Myers e Montgomery (2002)

2. Regressão Linear Clássica

Atkinson (1985), Atkinson e Riani (2000), Cook e Weisberg (1999), Draper e Smith (1998), Montgomery (2001, Cap 10), Montgomery *et al.* (2001), Myers e Montgomery (2002, Cap 2), Myers *et al.* (2002, Cap 2).

3. Transformação da Resposta

Atkinson (1985), Atkinson e Riani (2000), Box *et al.*(1978), Cook e Weisberg (1999), Cook e Olive (2001).

4. Modelos Lineares Generalizados

Atkinson e Riani (2000), Dobson (1990), Lindsey (1999), McCullagh e Nelder (1989), Myers *et al.* (2002).

Software Utilizado

1. ARC

Software gratuito que acompanha o livro de Cook e Weisberg (1999). Pode ser obtido no *site* do Departamento de Estatística da Universidade de Minnesota, EUA. Excelente para construir os gráficos para diagnóstico dos modelos. Na parte de modelos lineares generalizados as funções de ligações oferecidas são apenas: canônica, inverso, identidade e logarítmica.

2. Design-Expert

Software comercial desenvolvido e distribuído pela empresa Stat-Ease, especificamente para Planejamento de Experimentos conjuntamente com regressão linear. Possui um módulo de otimização.

3. S-Plus

Software comercial de uso geral, distribuído pela empresa Insightful. Dentre os softwares utilizados por nós é o único que tem a opção de Quase-Verossimilhança. Para os modelos lineares generalizados, em algumas distribuições de probabilidade, oferece poucas opções de função de ligação.

4. Genstat

Software comercial de uso geral, distribuído pela empresa VSN International. Oferece todas as opções de função de ligação para os modelos lineares generalizados.