5

Teste Inferencial para Análise Fatorial

Ao interpretar fatores, é preciso tomar a decisão sobre quais cargas fatoriais valem à pena considerar. A discussão a seguir detalha formalmente questões relativas à significância estatística e prática, bem como ao número de variáveis que afetam a interpretação de cargas fatoriais.

5.1
Formalização do Teste Inferencial

Seja abaixo uma matriz usual de fatores rotada pelo método varimax, de uma análise fatorial por componentes principais:

	F_1	F_2	F_3	F_4	F_5	....	F_n
X_1	c_{11}	c_{12}	c_{13}	c_{14}	c_{15}	....	c_{1n}
X_2	c_{21}	c_{21}	c_{23}	c_{24}	c_{25}	....	c_{2n}
X_3	c_{31}	c_{32}	c_{33}	c_{34}	c_{35}	....	c_{3n}
X_4	c_{41}	c_{42}	c_{43}	c_{44}	c_{45}	....	c_{4n}
X_5	c_{51}	c_{52}	c_{53}	c_{54}	c_{55}	....	c_{5n}
.....							
X_n	c_{n1}	c_{n2}	c_{n3}	c_{n4}	c_{n5}	....	c_{nn}

X_i é o atributo ou a variável considerada no modelo, F_j é a componente principal ou fator da rodada da análise fatorial e C_{ij} é a carga fatorial rotada para X_i e F_j.

Considere os fatores 1 e 2, as duas componentes principais que reúnem a maior porcentagem de variância explicada do problema (a inferência pode envolver mais de 2 fatores, conforme interesse do analista).

Seja C_{i1} a carga fatorial da componente 1(F_1), então C_{i1} é a correlação entre F_1 e X_i. Seja C_{i2}, a carga fatorial da componente 2(F_2), então C_{i2} é a correlação entre F_2 e X_i.

As cargas de F_1 e as cargas de F_2 definem a natureza dos respectivos fatores, os nomeiam, e caracterizam a análise fatorial encontrada e a matriz principal de fatores formada por correlações constitui o modelo usual de análise fatorial.

A inferência estatística proposta nesta tese é testar a significância das cargas fatoriais da matriz principal da análise fatorial, utilizando-se dos intervalos de confiança *bootstrap* e *jackknife*, percentílico principalmente e do ***Valor-p*** *bootstrap* e *jackknife*. Por simplificação, serão testadas somente as significâncias das cargas dos fatores 1 e 2.

A inferência estatística sugerida, então, é testar a significância de todas as cargas fatoriais de um determinado fator de interesse, os fatores 1 e 2, utilizando-se dos intervalos de confiança *bootstrap* e *jackknife*, percentílico principalmente, e dos **v*alores-p*** *bootstrap* e *jackknife*.

Os intervalos de confiança *e* os ***valores-p***, portanto, são construídos através da distribuição por amostragem avaliada, empírica, das variáveis aleatórias cargas fatoriais de uma componente principal de interesse. A distribuição por amostragem assim obtida tem por objetivo estabelecer um modelo adequado à interpretação do comportamento regular da estatística investigada e de seus parâmetros característicos. A experiência nesta investigação é a base para se montar o modelo, ou para ajustá-lo ao modelo ideal (teórico). A distribuição por amostragem e os seus respectivos intervalos de confiança e valor-p são obtidos a posteriori, com base na experiência, e não da maneira clássica, a priori, através de informações teóricas existentes sobre a distribuição por amostragem e seu respectivo erro-padrão.

Para realizar o teste de significância supra mencionado, se utilizará a idéia de que o intervalo de confiança pode ser usado imediatamente, sem qualquer outro cálculo para testar qualquer hipótese: o intervalo de confiança pode ser considerado como um conjunto de hipóteses aceitáveis.

Qualquer hipótese nula que esteja fora do intervalo de confiança deve ser rejeitada. Por outro lado, qualquer hipótese que esteja dentro do intervalo de confiança deve ser aceita.

Nesta tese, a hipótese nula é de que não existe correlação entre a variável X_i e o fator F_j, contra a hipótese alternativa de que existe tal correlação:

$$H_0: C_{ij} = 0,00$$
$$H_1: C_{ij} \neq 0,00$$

Na verdade, a hipótese alternativa que está sendo testada é:

- $C_{ij} > 0,00$, se o sentido da correlação do fator for direto com a variável testada ou.
- $C_{ij} < 0,00$, se o sentido da correlação do fator for inverso com a variável testada.

Matematicamente, pode-se exprimir por $C_{ij} = 0,00$ a hipótese de X_i e F_j serem não-correlacionados. Para testar esta hipótese, basta ver se o valor 0 está contido no intervalo de confiança.

Portanto, se o zero estiver contido no intervalo de confiança, aceita-se a hipótese nula de que não existe correlação, resultado não significante para a carga fatorial C_{ij}. Por outro lado, se o zero estiver fora do intervalo de confiança considera-se a hipótese alternativa de que existe correlação, resultado significante

para C_{ij}. Este teste passa a se chamar "***Teste Inferencial para Análise Fatorial pelo Intervalo de Confiança -TIAFIC".***

Em lugar de aceitar ou rejeitar simplesmente , uma forma de teste mais adequada é o cálculo do valor-p. Os passos para se calcular o valor-p *bootstrap* e *jackknife* são:

1º) Localizar as distribuições amostrais das cargas fatoriais por variável:

2º) Calcular a **distribuição G da variável 1(G_1)** da seguinte forma: diminuindo cada valor constante da coluna V_1 de sua média e dividindo de seu desvio-padrão;

- Realizar a mesma operação com as outras variáveis da base de dados;

- Este passo gerará os arquivos das distribuições amostrais com as variáveis padronizadas.

3º) Calcular a estatística **g** do teste da variável 1 da seguinte forma:

$$g_1 = \frac{C_1}{S_1}$$

Onde:

C_1 = a carga fatorial da variável 1(V_1), com o fator 1, obtida da rotação da **amostra original**;

S_1 = o desvio-padrão da variável 1(V_1) com o fator 1.

- Realizar a mesma operação com as outras variáveis da base de dados.

4º) Calcular o ***VALOR-P TIAF*** da seguinte forma:

Se $g_1 \geq 0$:

VALOR-P = P ($G \geq g_1$) = **(nº de valores da coluna $G_1 \geq g_1$) \ nº de reamostragens**

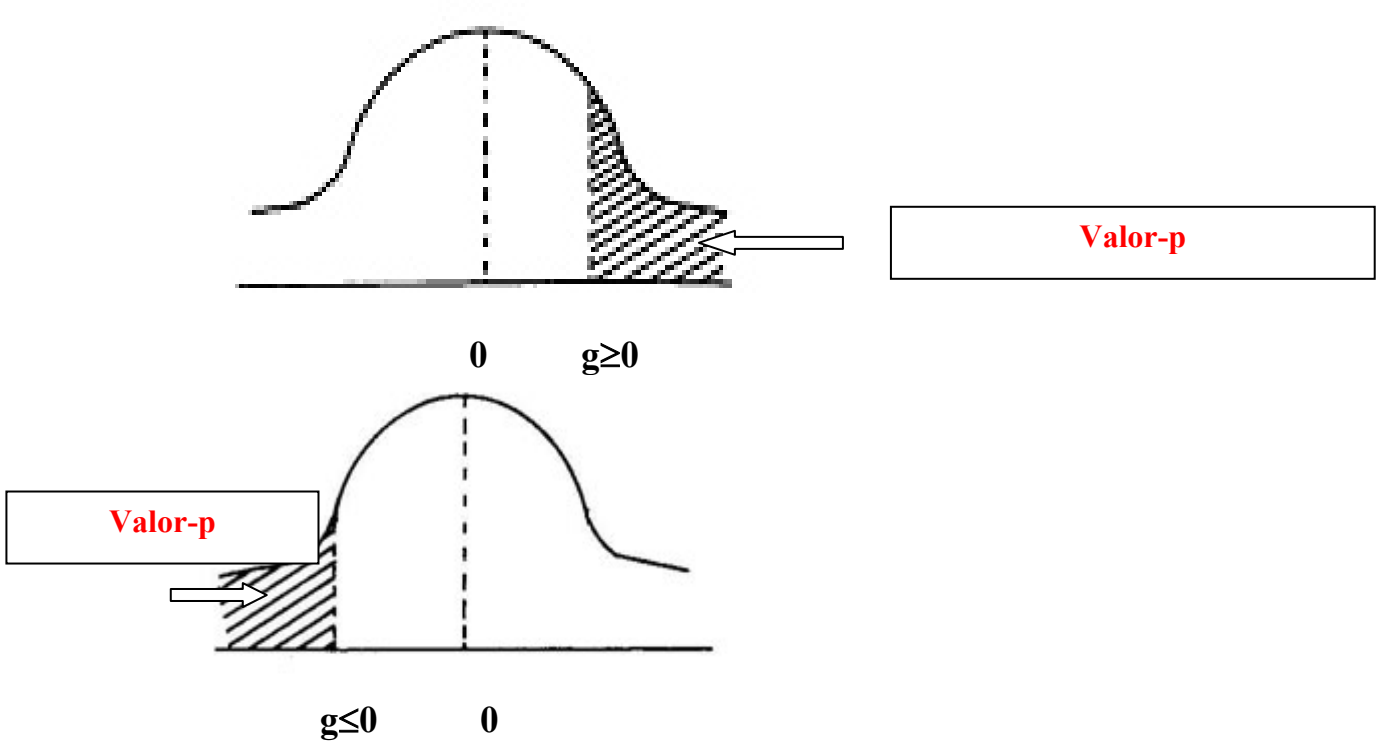
Se $g_1 \leq 0$:

VALOR-P = P ($G \leq g_1$) = **(nº de valores da coluna $G_1 \leq g_1$) \ nº de reamostragens**

Procura-se, então, a probabilidade na cauda, além deste valor observado de g: este é o ***valor-p***.

- Realizar a mesma operação com as outras variáveis da base de dados.

Visualização das áreas correspondentes aos *VALORES-P* calculados:



Este teste passa a se chamar "***Teste Inferencial para Análise Fatorial pelo Valor-p -TIAFVP".***

Cargas não significantes indicam que suas respectivas variáveis não participam ou não devem participar estatisticamente da nomeação/interpretação do fator em foco.Em contrapartida, cargas significantes pressupõem que a variável correspondente contribui estatisticamente para a formação do fator e esta contribuição é proporcional à sua significância ordinária ou magnitude.

Hair e Anderson (2005) sugerem como uma das etapas para a realização de uma boa análise fatorial, além da significância estatística, a verificação da significância prática. Vejam no trecho abaixo o que os referidos autores relatam sobre a questão:

> *"A força da análise multivarida é sua forma aparentemente mágica de ordenar um grande número de possíveis alternativas e encontrar as que têm significância estatística. Muitos pesquisadores ficam míopes ao se concentrar somente na significância alcançada dos resultados sem compreender suas interpretações, sejam boas ou ruins. Ao invés disso, o pesquisador deve olhar não apenas a significância estatística dos resultados, mas também sua significância prática. A significância prática faz a pergunta : 'E daí?' para qualquer aplicação administrativa, os resultados devem ter um efeito demonstrável que justifique a ação. Em termos acadêmicos, a pesquisa está se concentrando não apenas em resultados estatisticamente significantes, mas também em suas implicações substantivas e teóricas, as quais são muitas vezes extraídas de sua significância prática".*

Uma regra ou sugestão utilizada Hair e Anderson (2005) para se decidir se uma determinada variável pode ser considerada na nomeação de um fator sob investigação (significância prática) seria decidir pela sua significância ordinária toda vez que a variável de carga significante estatisticamente estiver entre ***/ 0,3 a 1,0 /***, pelo motivo indicado no Quadro3:

Quadro 3: Verificação da Significância Prática

Valores C_{ij} de um Determinado Fator (F_i)	Correlação ou Relação
| 0,00 |— 0,30 |	***Ausência de correlação ou grau de correlação não tolerável***
| 0,30 |— 1,00 |	***Existência de correlação ou grau de correlação tolerável***

Matematicamente, a hipótese desta tese consiste que se uma determinada carga de uma variável, $C_{ij,}$ não for significativa estatisticamente, então a correlação da variável X_i com o fator F_j, $C_{ij,}$ é nula, e da natureza ou nomeação de F_j não deve participar a variável X_i.Conseqüentemente, a análise fatorial será interpretada de acordo com um processo inferencial estatístico.

Denomina-se de vetor θ_1 o parâmetro populacional definido como as cargas fatoriais da primeira componente principal da matriz varimax(F_1), obtido da amostra original e denomina-se de vetor θ_2 o parâmetro populacional definido como as cargas fatoriais da segunda componente principal da matriz varimax(F_2), obtido da amostra original.

Adotando-se como estimador de θ_1, a matiz aleatória formadas pelas estimativas das cargas fatoriais de F_1, cujas colunas são as variáveis do problema e as linhas, as **B** ou as **n** reamostragens obtidas através das técnicas *bootstrap* e *jackknife* e como de θ_2 , a matiz aleatória formadas pelas estimativas das cargas fatoriais de F_2 cujas colunas são as variáveis do problema e as linhas, as **B** ou as **n** reamostragens obtidas através das técnicas bootstrap e *jackknife*; o problema é encontrar a distribuição por amostragem dos estimadores dos vetores θ_1 e θ_2, isto é, das matrizes aleatórias de F_1 e F_2, ou seja, a sua variância, viés e EMQ. Com essas estimativas, é possível construir intervalos de confiança para as estimativas dos vetores θ_1 e θ_2 , calcular valores-p e realizar o ***TIAF.***

A relevância deste procedimento sugerido é que em análise fatorial não é tão fácil encontrar pela teoria tradicional as exatas distribuições de amostragem

dos estimadores dos vetores θ_1 e θ_2. As técnicas ***CIS*** surgem originalmente como uma alternativa para este tipo de situação, mas se tornaram mais populares e viáveis devido em grande parte ao avanço computacional das últimas décadas.

Para estimar a distribuição por amostragem dos estimadores dos vetores θ_1 e θ_2 para qualquer pesquisa em que se utilize de uma análise fatorial por componente principal com rodada varimax serão oferecidas nesta tese as opções *bootstrap* e *jackknife*.

A teoria assintótica dos estimadores dos vetores θ_1 e θ_2 deve ser estabelecida,isto é, deve verificar se a aplicação das aproximações *bootstrap* e *jackknife* são válidas para estas estatísticas e se são consistentes. Esta fase será desenvolvida empiricamente com base em estudos de casos no capítulo 8: Estudo de Casos: Aplicações do TIAF.

Deverão ser desenvolvidos para cada base de pesquisa dois algoritmos, um *bootstrap* e outro *jackknife* para o cálculo dos respectivos intervalos de confiança, valores-p e resultados TIAF'S e que foram programados através dos softwares ***R 2.1.1*** e ***SAS Versão 8*** para obtenção computacional das estimativas.

Compara-se pelo EMQ, por exemplo, o melhor procedimento e assim pode-se utilizar o intervalo de confiança e o valor-p da melhor técnica para testar a significância de estimativas geradas, dos estimadores dos vetores θ_1 e θ_2, utilizando-se o intervalo de confiança e o valor-p.

5.2
Algoritmo *Bootstrap* para Realização do TIAF

Os algoritmos para realização computacional dos procedimentos inferênciais utilizando-se das técnicas *bootstrap* seguem abaixo.

Uma amostra existente de tamanho ***n*** é tomada como uma amostra original:

1º) Da amostra original, seleciona-se a primeira amostra *bootstrap*;

2º) Com base na amostra selecionada, gera-se a matriz principal de fatores, com base na matriz de correlações de tamanho igual a n;

3º) Da matriz principal de fatores, obtém-se a matriz rodada de fatores ortogonal (rotação *varimax*);

4º) Com base na matriz rodada de fatores, selecionam-se dos fatores 1 e 2 as cargas fatoriais, isto é, as estimativas dos vetores θ_1 e θ_2;

5º) Repetem-se os passos de 1º a 4º B vezes, por exemplo, 1000 vezes;

6º) Ordenam-se os valores obtidos para as cargas fatoriais, do menor ao maior. Determinam-se limites de confiança para uma especificada probabilidade α, igual ao nível de significância, de acordo com as expressões abaixo:

$\hat{\theta}(q_1)$= *limite inferior, onde* $q_1 = B\ \alpha\backslash 2$

$\hat{\theta}(q_2)$= *limite superior , onde ,* $q_2 = B - q_1 + 1$

Se forem 1000 iterações(B) e $\alpha = 0.05$, o limite inferior será o valor na 25ª posição (q_1=B.α/2)e o limite superior o na 976ª posição (q_2=B-q_1+1).

Pode-se então afirmar, com uma probabilidade α de se estar errado, que o intervalo de confiança construído tem alta probabilidade de conter o verdadeiro valor da carga fatorial sobre o qual a estimação foi baseada.

7º) Calculam-se os **valores-p bootstrap** para cada variável da pesquisa baseada na cauda direita ou esquerda da distribuição por amostragem empírica e real da variável em foco, isto é, a probabilidade na cauda, além deste valor observado da estimativa: este é o ***valor-p bootstrap***. Este valor corresponde à credibilidade da hipótese nula.

8º) Calculam-se: o valor esperado, a variância, o viés e o EMQ das distribuições amostrais obtidas, como indicado abaixo:

$$Viés = [\, E(\hat{\theta}) - \theta \,]$$

$$EMQ = S^{2}_{\hat{boot}} + Viés^{2}$$

As estatísticas erro-padrão, viés e EMQ são indicadores da eficácia e da qualidade da estimação efetuada.

A distribuição por amostragem de $\hat{\theta}$ corresponde ao histograma dos B valores determinados para as estimativas de $\boldsymbol{\theta_{(1)}, \theta_{(2)}, \theta_{(3)}, \ldots, \theta_{(B)}}$.

θ é o vetor de cargas fatoriais dos fatores F_1 e F_2 da matriz varimax da amostra original.

O algoritmo Jackknife segue procedimentos análogos.

5.3
Computação dos Algoritmos *Bootstrap* e *Jackknife*

Os programas propostos para os métodos *bootstrap* e *jackknife* com base nos algoritmos elaborados acima para reamostragem das B estatísticas no método *bootstrap* e nas *n*, no esquema *jackknife*, com uma confiança de 95% e com base nos softwares *R 2.1.1* e *SAS Versão 8.*

Os programas computacionais elaborados para realizar os procedimentos inferênciais para a análise fatorial utilizando as técnicas *bootstrap* e *jackknife* encontram-se nos Anexos 1 e 2 (versão R e versão SAS) e constituem a materialização, o produto concreto de todo o estudo realizado neste trabalho. Nos estudos de caso, utilizou-se dos resultados submetidos ao programa R 2.1.1, que é um programa *free,* mais disponível e de mais fácil execução.

Os programas em R 2.1.1 também estarão em uma biblioteca desenvolvida para o R, disponibilizando-a na rede. A vantagem é que a biblioteca pode ser usada por diferentes pessoas, que irão eventualmente reportar críticas ao procedimento, que pode então ser atualizado e aperfeiçoado posteriormente.

A escolha entre *bootstrap* e *jackknife* não é imediata, depende do usuário. Neste contexto, aponta-se que aquele que fornecer menor estimativa de precisão e viés, isto é, que minimize o erro médio quadrático (EMQ).

5.4
O Método do TIAF

Esta sessão objetiva estabelecer o método científico para realizar o TIAF.

O método científico é a ferramenta colocada à disposição do cientista que, através da pesquisa, pretende penetrar no segredo de seu objeto de estudo.

O método, em sentido amplo, é a ordem que se deve impor aos diversos processos necessários para atingir um fim dado ou um resultado desejado.

Método Científico é um instrumento de que se serve a inteligência para descobrir relações, verdades e leis referentes aos diversos objetos de investigação.

O método científico é um dispositivo ordenado, um conjunto de procedimentos sistemáticos que o pesquisador emprega para obter o conhecimento adequado do problema que se propõe resolver.

O método é constituído de um conjunto de processos ou técnicas que formam os passos do caminho a percorrer na busca da verdade.

Toda investigação nasce da observação cuidadosa de fatos que necessitam de uma maior explicação. Esta é imaginada através da hipótese. Em seguida, procura-se verificar a veracidade da solução sugerida. Nas ciências experimentais, isto é feito através de ensaios e experiências; nas ciências humanas, é feito através de demonstrações racionais e lógicas por meio da argumentação. Descoberta a explicação do fato, achada a relação de causalidade entre os fenômenos ou sua coexistência ou ainda sua finalidade, formula-se a lei. É a tarefa da indução: aplicar a relação necessária descoberta a casos não observados da mesma espécie.

Esta explicação parcial e fracionada de uma realidade não satisfaz a curiosidade científica. Por isso, o cientista reúne as tentativas de explicação, os princípios e leis particulares numa visão unificadora, mais ampla e globalizada, através da teoria ou do sistema.

Em resumo, o desenvolvimento do método científico se faz pelos processos ou técnicas da observação, hipótese, demonstração (experimental ou racional), indução da lei e teoria. Além disso, e simultaneamente com os processos referidos, o pesquisador sempre estará usando as técnicas da análise, da síntese e da indução.

O método científico é, pois, um meio imprescindível com a qual o espírito científico do pesquisador, com ordem e rigor, procura penetrar no sentido dos fatos e fenômenos que pretende conhecer.

Fundamentado em tudo o que foi dito em parágrafos acima e para solucionar o problema proposto nesta tese é necessário estabelecer o método científico para testar a significância de cargas fatoriais de componentes principais. O método científico da investigação deste estudo chama-se **"Método TIAF"** .

O "Método ***TIAF***" consiste nas seguintes etapas que devem ser seguidas nesta ordem para garantir o máximo de eficácia na obtenção do objetivo proposto:

1º) *Selecione uma boa amostra da população, de preferência probabilística, de tamanho proporcional ao número de variáveis e que reúna um conjunto o mais completo possível de características sobre o problema proposto (amostra original ou base de dados);*

2º) *Efetue os testes de validação da análise fatorial;*

3º) *Rode a análise fatorial e obtenha a matriz de fatores rodada varimax;*

4º) *Submeta os programas* bootstrap *e jackknife;*

5º) *Verifique comparativamente qual o melhor método para seus dados (o bootstrap ou o jackknife);*

6º) *Identifique para quais variáveis as cargas dos fatores de interesse são significantes, isto é, onde existe correlação significativa com o fator;*

7º) *Identifique qual das variáveis significante estatisticamente podem participar da nomeação e/ou interpretação dos fatores (significância prática);*

8º) *De acordo com um critério do analista, selecione as variáveis que nomearão os fatores em foco;*

9º) *Nomeie os fatores em estudo;*

10º) *Tomada de decisão administrativa, tendo uma confiança alta de que os resultados amostrais levados em consideração na análise de dados não são acidentais, frutos do acaso.*

A superioridade da abordagem inferencial para as cargas fatoriais empregadas neste trabalho, em relação ao apresentado pelo Método Hair e Anderson (2005), é que o número de variáveis contempladas e o fator específico em exame são considerados no processo de significância das cargas fatoriais.As reamostragens são realizadas levando em consideração todas as variáveis do problema (quanto mais completa for a lista de variáveis melhor o teste de significância da análise fatorial) e para um fator específico F_j gerado em cada amostra *bootstrap* ou *jackknife* conforme seu poder de explicação dentro do modelo fatorial.

No capítulo 9, se discutirá com mais detalhes aspectos comparativos e de desempenho entre o "Método Tradicional de Hair e Anderson" e o "Método TIAF".