Formulário e Definições de Estatística

Daniel Vartanian

August 22, 2022

Contents

Part I

Estatística Descritiva

Part II

Probabilidades

Part III

Inferência Estatística

10 Introdução à Inferência Estatística

11 Estimação

12 Intervalos de Confiança

13 Testes de Hipótese

14 Inferência para Duas Populações

15 Análise de Aderência e Associação

16 Inferência para Várias Populações

17 Regressão Linear

Procedimentos Univariados

Teste K-S e Shapiro-Wilk

Análise de Outlier - Boxplot

Verificar os slides "Introdução ao SPSS"

17.1 Método da Linha Reta

"É o tipo mais simples de ajustamento de curvas, cuja equação é

$$Y = a + bX \ ,$$

onde X e Y são variáveis e a e b são as constantes.

Assim, dados dois pontos quaisquer $(X1, Y1)$ e $(X2, Y2)$ dessa reta, as constantes a e b podem ser determinadas" [7] .

17.1.1 Equação da Reta

"Podemos também considerar a forma mais comum de representar a função de 1° grau (função linear), ou seja,

$$y = mx + n \ ;$$

onde $b = m$ é coeficiente angular e $a = n$ é o coeficiente linear" . [6]

$$m = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

17.2 Método dos Mínimos Quadrados

"A reta dos mínimos quadrados que se ajusta ao conjunto de pontos, tem a equação:

$$Y = a + bX + e \ ,$$

sendo $a + bX$ a equação da reta, e e o termo erro. Este último termo tem de ser incluído porque o valor de Y não será dado exatamente pelo ponto da reta a ser encontrada" [7].

"Podemos dizer então, que o erro dá conta de todos os eventos que são difíceis de medir, mas que são (supostamente) aleatórios. Mais do que isso, se o modelo (no nosso caso uma reta) estiver corretamente especificado, podemos supor que o erro, em média, será zero. Isto é, a probabilidade do erro ser x unidades acima da reta é a mesma de ser x unidades abaixo. Essa é a primeira hipótese: $E(e_i) = 0$ " [7].

"O próximo passo é estimar a reta de regressão:

$$a = \frac{\Sigma y - b \Sigma x}{n}$$

$$b = \frac{n \Sigma xy - \Sigma x \Sigma y}{n \Sigma x^2 - (\Sigma x)^2}$$

$a =$ o valor de y_i , quando o $x_i = 0$ ou o intercepto da reta no eixo y .

$b =$ o valor do coeficiente angular, que indica a inclinação da reta" [7].

17.3 Coeficiente de Correlação Amostral de Pearson (r)

"O coeficiente de correlação amostral de Pearson é indicado por r e calculado através da fórmula:

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[n(\Sigma X^2) - (\Sigma X)^2][n(\Sigma Y^2) - (\Sigma Y)^2]}}$$

onde,

$r = 1 \Rightarrow$ Correlação perfeita positiva;

$r = 0 \Rightarrow$ Correlação nula;

$r = -1 \Rightarrow$ Correlação perfeita negativa". [7]

O coeficiente de correlação **populacional** de Pearson é indicado por ρ.

Premissas [7]

(a) as duas variáveis envolvidas são **aleatórias e contínuas**;

(b) as duas variáveis **apresentam uma distribuição normal**.

17.4 Coeficiente de Determinação (r^2 ou R^2)

"O coeficiente de determinação r^2 (amostral) ou R^2 (populacional) mede o grau de ajustamento da reta de regressão aos dados observados.

O coeficiente de determinação representa a relação entre a variação explicada pelo modelo e a variação total, ou em outras palavras, indica a proporção da variação total da variável dependente y que é explicada pela variação da variável independente x ". [7]

17.5 Erro Padrão da Estimativa Amostral (s_e)

"O erro padrão da estimativa [amostral] calcula a dispersão dos resíduos (diferença entre valores reais e preditos) dos valores amostrados ao redor da reta de regressão. Quanto maior a dispersão, menor a precisão das estimativas". [7]

"Pode ser calculado pela fórmula que segue:

$$s_e = \sqrt{\frac{\Sigma y^2 - a\Sigma y - b\Sigma xy}{n-2}}$$

onde,

s_e = o erro padrão associado a y ;

n = número de observações". [7]

17.6 Erro Padrão do Coeficiente Angular Amostral (s_b)

"O cálculo do erro padrão do coeficiente angular amostral b é importante para poder construir o intervalo de confiança e efetuar os testes de hipóteses apropriados para o coeficiente angular β ". [7]

"Algebricamente, o erro padrão de b (coeficiente angular) pode ser apresentado por meio da seguinte equação:

$$s_b = \frac{s_e}{\sqrt{(n-1)\cdot S_x^2}}$$

onde,

s_e = o erro padrão associado a y ;

n = número de observações;

S_x^2 = variância de x (variável independente)". [7]

Vale relembrar a forma de calcular S_x^2 e $\bar{X}$.

$$S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$$

$$\bar{X}_x = \frac{\Sigma X_i}{n}$$

17.7 Erro Padrão do Coeficiente Linear Amostral (s_a)

"Algebricamente, o erro padrão [amostral] de a (coeficiente linear) pode ser apresentado por meio da seguinte equação:

$$s_a = s_e \cdot \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{(n-1) \cdot S_x^2}}$$

onde,

s_e = o erro padrão associado a y ;

n = número de observações;

$\bar{X}$ = média de x (variável independente);

S_x^2 = variância de x (variável independente)". [7]

17.8 Erro Padrão do Coeficiente de Correlação Populacional (s_ρ)

"O erro padrão do coeficiente de correlação populacional, geralmente expresso pela letra rô (ρ), pode ser calculado pela seguinte expressão:

$$s_\rho = \sqrt{\frac{1 - r^2}{n - 2}}$$

onde,

r^2 = o erro padrão associado a y ;

n = número de **pares analisados**". [7]

17.9 Utilizando a Distribuição t (*Student*) para Testar a Nulidade dos Estimadores [7]

17.9.1 Premissas

(a) as duas variáveis envolvidas são **aleatórias e contínuas**;

(b) as duas variáveis **apresentam uma distribuição normal**.

17.9.2 Coeficiente Angular (β)

$$t = \frac{b - \beta_0}{s_b}$$

Intervalo de Confiança $b \pm t \cdot s_b$

Teste de Hipóteses $\begin{cases} H_0: & \beta = 0 \\ H_a: & \beta \neq 0 \end{cases}$

17.9.3 Coeficiente Linear (α)

$$t = \frac{b - \alpha_0}{s_a}$$

Intervalo de Confiança $a \pm t \cdot s_a$

Teste de Hipóteses $\begin{cases} H_0: & \alpha = 0 \\ H_a: & \alpha \neq 0 \end{cases}$

17.9.4 Coeficiente de Correlação (ρ)

$$t = \frac{b - \rho_0}{s_\rho}$$

Intervalo de Confiança $r \pm t \cdot s_\rho$

Teste de Hipóteses $\begin{cases} H_0: & \rho = 0 \\ H_a: & \rho \neq 0 \end{cases}$

17.9.5 Projeção ($\hat{Y}_i$)

Intervalo de Confiança $\hat{y}_i \pm t \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{(x_i - \bar{X})^2}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}}$

17.10 Hipóteses do teste K-S e Shapiro-Wilk [7]

Ambos são testes de aderência à distribuição normal. Para os testes de Kolmogorov-Smirnov (**K-S**) e **Shapiro-Wilks** as hipóteses normalmente utilizadas são as abaixo. [7]

$$\begin{cases} H_0: & \text{os dados apresentam distribuição normal} \\ H_a: & \text{os dados não apresentam distribuição normal} \end{cases}$$

Logo, para esses testes, quando o sigma estiver acima de 5% (caso seja esse o nível de significância utilizado) deve se considerar que os dados apresentação distribuição normal.

O teste de Kolmogorov-Smirnov **K-S** é utilizado para amostras grandes, quando $n > 50$. Para amostras pequenas utilize o teste de **Shapiro-Wilks**. [7]

17.11 Avaliação do Modelo Utilizando o SPSS

Pressupostos da Regressão

(a) Qualidade do Ajustamento;

(b) Teste t;

(c) Teste F;

(d) Normalidade dos Resíduos;

(e) Homocedasticidade dos Resíduos (Variância Constante);

(f) Linearidade;

(g) Ausência de Autocorrelação dos Resíduos;

(h) Ausência de Multicolinearidade entre as Variáveis Independentes. [7]

17.11.1 Qualidade do Ajustamento

Verifique o valor do R^2 **ajustado** ou *Adjusted* R^2.
Quanto maior o ajustamento mais o modelo explica o que está se estudando. [7]

Sumarização do modelo[b]

Modelo	R	R quadrado	R quadrado ajustado	Erro padrão da estimativa	Mudança de R quadrado	Mu
1	,361[a]	,131	,120	1,0897	,131	

a. Preditores: (Constante), v25, v22

b. Variável Dependente: v84

17.11.2 Teste t

Verifique a nulidade das variáveis.

Caso exista, uma opção é retirar as variáveis nulas e rodar o modelo de novo. Outra opção é rodar uma regressão fatorial.

É possível manter a variável mesmo nula caso ela seja muito importante para o modelo. [7]

Coeficientes[a]

Modelo		Coeficientes não padronizados		Coeficientes padronizados	t	Sig.	95,0% Intervalo de Confiança para B	
		B	Erro Padrão	Beta			Limite inferior	Limite superior
1	(Constante)	5,273	,632		8,343	,000	4,025	6,521
	v22	,173	,071	,202	2,450	,015	,034	,312
	v25	,200	,076	,217	2,629	,009	,050	,350

a. Variável Dependente: v84

17.11.3 Teste F

Ao verificar os resultados da ANOVA (*Analysis of Variance*), veja se o sigma de F está abaixo de 5% (caso seja esse o nível de significância desejado).

ANOVA[a]

Modelo		Soma dos Quadrados	gl	Quadrado Médio	F	Sig.
1	Regressão	30,156	2	15,078	12,698	,000[b]
	Resíduo	200,677	169	1,187		
	Total	230,833	171			

a. Variável Dependente: v84

b. Preditores: (Constante), v25, v22

As premissas dos testes de hipóteses de F são as mesma do teste t. Seguem elas abaixo.

$$\begin{cases} \mathrm{H}_0: \quad R^2 = 0 \\ \mathrm{H}_a: \quad R^2 \neq 0 \end{cases}$$

17.11.4 Normalidade dos Resíduos

É importante olhar a normalidade dos resíduos para verificar se eles estão ajustados ao modelo (neste caso linear) e também para verificar a presença/ausência de outliers.

No SPSS isso se faz explorando a explorando os **resíduos standardizados Z** ou *Standardized Residual Z - ZRE*. [7]

Resumo de processamento de casos

	Casos					
	Válido		Omisso		Total	
	N	Porcentagem	N	Porcentagem	N	Porcentagem
Standardized Residual	172	100,0%	0	0,0%	172	100,0%

Testes de Normalidade

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	gl	Sig.	Estatística	gl	Sig.
Standardized Residual	,095	172	,001	,961	172	,000

a. Correlação de Significância de Lilliefors

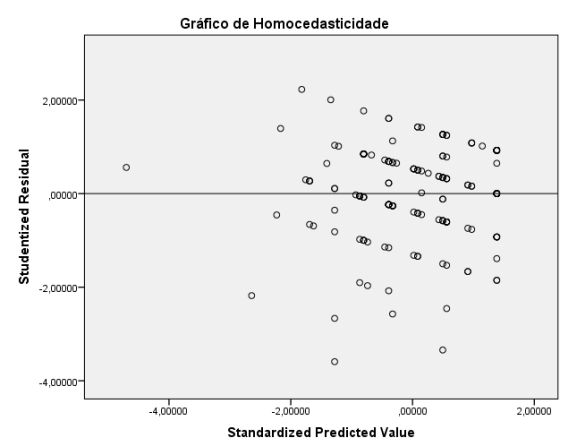
17.11.5 Homocedasticidade dos Resíduos (Variância Constante)

"A presença de variâncias não homogêneas é uma violação de um dos pressupostos da regressão, conhecida como heterocedasticidade.

Possíveis causas: outliers; erro de especificação das variáveis; **erro na função matemática** (Ela pode ser não-linear. No marketing as disciplinas não passam por regressão não-linear, logo, nos exercícios a opção nunca será essa. O correto é testar outros modelos antes de retirar os outliers.), entre outros.

Soluções possíveis: transformar variáveis ou estimação da regressão via mínimos quadrados ponderados; retirada de outliers". [7]

Para analisar a homocedasticidade no SPSS, **gere um gráfico de dispersão simples** com a variável *Studentized Residual - SRE* no eixo *x* e a variável *Standardized Predicted Value - ZPR* no eixo *y* .



17.11.6 Linearidade

O diagnóstico de linearidade pode ser feito pelo diagrama de dispersão [o mesmo realizado para a avaliação de homocedasticidade, que dá uma boa ideia sobre sua linearidade em torno das observações das variáveis dependentes e independentes.

Suas possíveis causas e soluções são as mesmas apresentadas para Variância Constante (pressuposto anterior).

17.11.7 Ausência de Autocorrelação dos Resíduos

"A análise da autocorrelação dos resíduos pode ser feita através do Teste de Durbin-Watson, cujas hipóteses são:

$$\begin{cases} H_0: & \text{não existe autocorrelação dos resíduos} \\ H_a: & \text{existe autocorrelação dos resíduos} \end{cases}$$

Quero que não exista autocorrelação, pois a violação leva a erro na estimação dos parâmetros.

A idéia da chamada autocorrelação serial é que os resíduos contém mais informação sobre a variável dependente do que aquilo que foi "filtrado" pelas variáveis explicativas. Em termos técnicos, o resíduo ainda pode ser sistematizado.

Exemplos de autocorrelação são normalmente encontrados em trabalhos que utilizam séries de tempo como dados de análise.

A autocorrelação dos resíduos depende do valor do teste de Durbin-Watson, cuja interpretação é:

Valores próximos de 2, não existe autocorrelação dos resíduos;

Valores próximos de Zero, significa autocorrelação positiva;
Valores próximos de 4, significa autocorrelação negativa". [7]

Sumarização do modelo[b]

Erro padrão da estimativa	Estatísticas de mudança					Durbin-Watson
	Mudança de R quadrado	Mudança F	gl1	gl2	Sig. Mudança F	
1,0897	,131	12,698	2	169	,000	2,062

17.11.8 Ausência de Multicolinearidade entre as Variáveis Independentes

"A multicolinearidade ocorre quando duas ou mais variáveis independentes do modelo apresentam **correlação alta (superiores em termos absolutos a 0,9)**, pois significa que contêm informações similares.

As consequências são: erros-padrão maiores, menor eficiências dos estimadores, estimativas imprecisas, entre outras". [7]

Correlações

		v84	v22	v25
Correlação de Pearson	v84	1,000	,308	,316
	v22	,308	1,000	,492
	v25	,316	,492	1,000
Sig. (unilateral)	v84	.	,000	,000
	v22	,000	.	,000
	v25	,000	,000	.
N	v84	172	172	172
	v22	172	172	172
	v25	172	172	172

17.12 Análise de Outlier

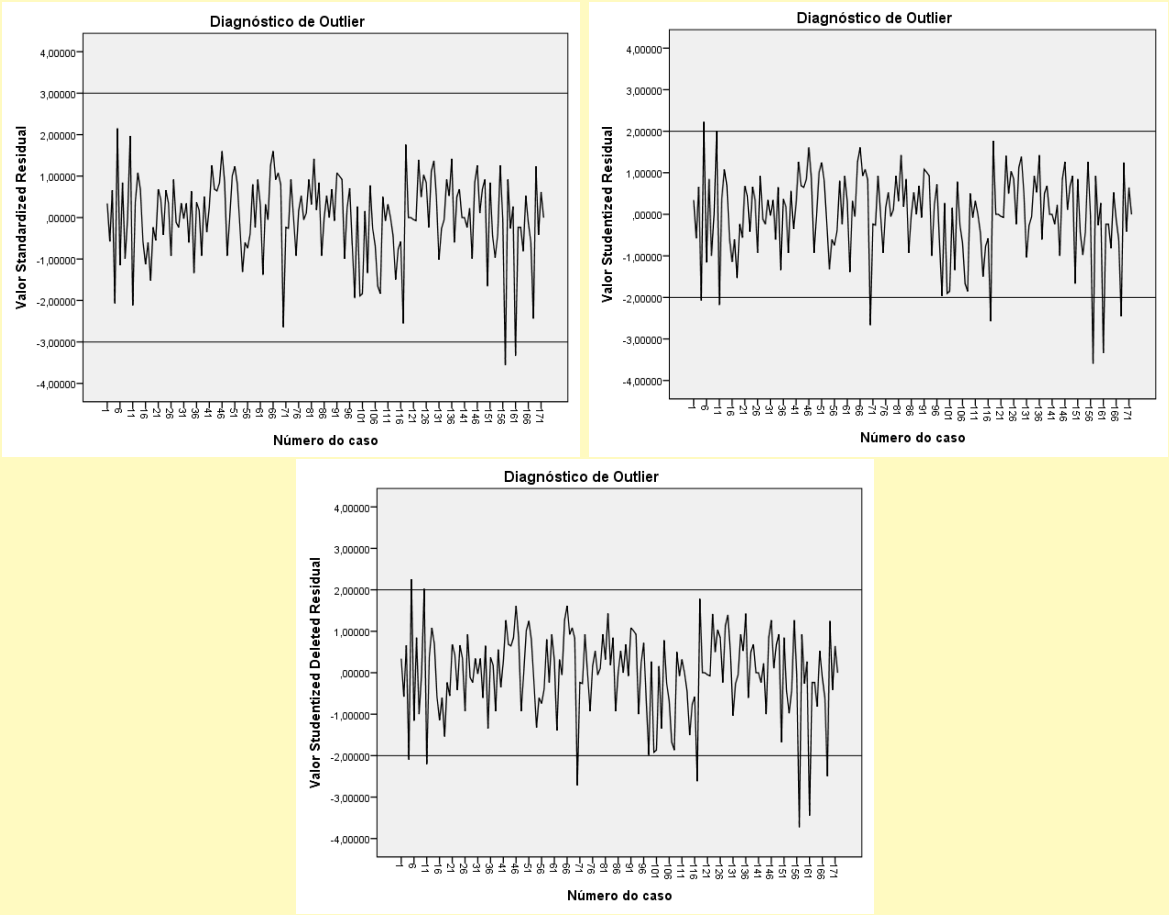
"A identificação de outliers no modelo de regressão linear é feita essencialmente através dos **resíduos standardizados** (*Standardized Residual - ZRE*), **studantizados** (*Studentized Residual - SRE*) e **studantizados deletedos** (*Studentized Deleted Residual - SDR*), pela verificação de pelo menos uma das condições:

(a) Resíduos standardizados terem valores absolutos superiores a 3;

(b) Resíduos studantizados terem valores absolutos superiores a 2;

(c) resíduos studantizados deleted terem valores absolutos superiores a 2". [7]

Para isso gere **gráficos de linha simples** com as variáveis acima mencionadas com as seguintes configurações:
Escala do gráfico: Mínimo: -4; Máximo: 4; Incremento: 1;
Linhas de referências no eixo Y: ZRE: 3 e -3; SRE e SDR: 2 e -2.



Como remover os Outliers

Comece tirando os outllies que aparecerem mais de um vez OU aqueles que a distância da linha de referência for maior - retire do *n* de número maior para o menor. Isso é importante pois o SPSS altera o número das linhas conforme você deleta o outlier.
Rode o modelo de novo e veja se melhorou. Se não, retire mais outliers e repita o processo.
Se você tirar muitos outliers você pode acabar com seu banco de dados. Caso isso aconteça um outra solução deve ser aplicada. Pode ser que colocando mais variáveis esse outliers melhorem (PODE SER).

18 Análise de Variância (ANOVA)

"A ANOVA é empregada para verificar se há diferença sistemática entre as médias de resultados **normalmente distribuídos** de experimentos randômicos.
Trata-se de um método estatístico, que por meio de **teste de igualdade de médias**, verifica se fatores (variáveis independentes) produzem mudanças sistemáticas em alguma variável de

interesse (variável dependente).

Os fatores propostos podem ser variáveis quantitativas ou qualitativas, enquanto a variável dependente deve ser quantitativa.

Todos os sujeitos (participantes ou unidades experimentais) de determinado grupo recebem o mesmo tratamento, assegurando que as diferenças sistemáticas entre médias de grupos possam ser atribuídas aos efeitos dos diferentes tratamentos" [7] .

18.1 Pressupostos da ANOVA [7]

18.1.1 As observações dentro de cada grupo têm distribuição normal

"A normalidade não é restritiva ao uso da ANOVA quando o número de elementos em cada grupo é relativamente elevado ($n \geq 30$) ;

A não normalidade tem consequências mínimas na interpretação dos resultados, a não ser que a distribuição seja muito viesada".

Outras Observações

A normalidade é função da combinação de 2 medidas: assimetria e curtose. Se ela é normal ela é simétrica e meso-cúrtica;

Caso ela não dê normal, uma possibildiade é calcular assimetria e a curtose. Então se divide as duas pelo desvio padrão;

Se for normal, a divisão da assimetria/desvio padrão e curtose/desvio padrão deve ficar entre $(-1,96; 1,96)$.

18.1.2 As observações são independentes entre si

Se as observações foram coletadas de maneira independente, logo os tratamentos serão independentes entre si (e.g. grupo experimental x grupo de controle).

18.1.3 As variâncias de cada grupo são iguais entre si, ou seja, há homocedasticidade

"O teste F é robusto a violações de homocedasticidade quando o número de observações em cada grupo é igual ou aproximadamente igual (considera-se grupos de dimensões semelhantes quando o quociente entre a maior dimensão e a menor for inferior a 1,5 - **se a razão entre a**

amostra de tamanho maior dividido pela amostra de tamanho menor der até 1,5 = OK);

Quando os n não são iguais ou semelhantes e há grande afastamento tanto da normalidade como da homocedasticidade, põe-se em risco as conclusões tidas na análise de variância. Nesta situação recomenda-se utilizar testes alternativos não paramétricos de Kruskal-Wallis".

18.2 Cálculo da ANOVA

18.2.1 Quadro da ANOVA

Fonte de Variação	Soma dos Quadrados	Grau de Liberdade	Quadrados Médios	Teste F
Entre Tratamentos	Q_e	$k-1$	$S_e^2 = \frac{Q_e}{k-1}$	
Dentro das Amostras (Residual)	$Q_r = Q_t - Q_e$	$n-k$	$S_r^2 = \frac{Q_t - Q_e}{n-k}$	$F_{\text{cal}} = \frac{S_e^2}{S_r^2}$
Total	Q_t	$n-1$		

18.2.2 Fórmulas

Soma dos Quadrados Totais $Q_t = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}^2 - C$

Soma dos Quadrados Entre Tratamentos $Q_e = \sum_{i} \left[\frac{\left(\sum_{j} x_{ij} \right)^2}{n_i} \right] - C$

Constante $C = \dfrac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} \right)^2}{n}$

18.2.3 Tabela de Contingência [2]

Indivíduo	Variável					
	X_1	X_2	...	X_j	...	X_p
1	X_{11}	X_{12}	...	X_{1j}	...	X_{1p}
2	X_{21}	X_{22}	...	X_{2j}	...	X_{2p}
.	.	.		.		.
.	.	.		.		.
.	.	.		.		.
i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ip}
.	.	.		.		.
.	.	.		.		.
.	.	.		.		.
n	X_{n1}	X_{n2}	...	X_{nj}	...	X_{np}

18.2.4 Utilizando a Distribuição F (Fisher–Snedecor) para Testar a Nulidade dos Estimadores

"A distribuição "F" é apropriada para a **razão das variâncias de duas amostras**, tomadas independentemente da mesma população normalmente distribuída.

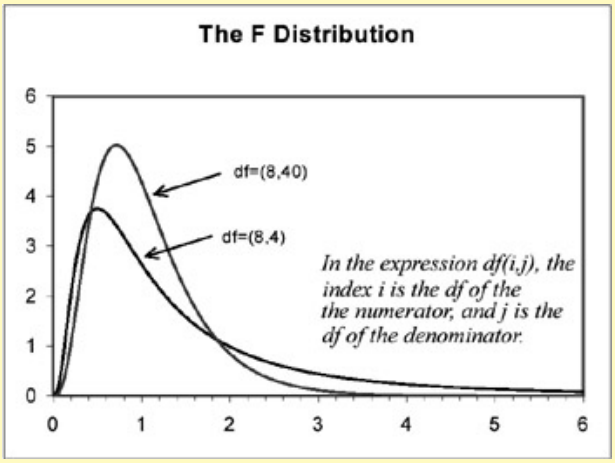
A estatística usada para testar a hipótese nula de que não existe diferença entre as variâncias é " [7]:

$$F_{v_1,v2} = \frac{S_1^2}{S_2^2}\ ,\ \text{onde}\ F_{v_1,v2,\text{inferior}} = \frac{1}{F_{v_2,v1,\text{superior}}}$$

Teste de Hipóteses $\begin{cases} H_0: & \mu_A = \mu_B = \mu_C \\ H_a: & \text{As médias de pelo menos dois grupos são diferentes} \end{cases}$

Ponto crítico $F_c\ \alpha - 1\ ;\ k - 1\ ;\ n - k = ?$

Distribuições F [4]



18.3 ANOVA 1 Fator no SPSS [7]

Para executar a ANOVA no SPSS vá em "Analisar > Comparar Médias > Análise de Variância Unidirecional...", porém não esqueça de verificar a normalidade dentro de cada grupo.
Para verificar a normalidade utilizando o SPSS, faça os testes de Kolmogorov-Smirnov (**K-S**) e **Shapiro-Wilks** (verifique as hipóteses desses testes na seção de Regressão Linear).

18.3.1 Teste de Levene

O teste de Levene serve para verificar a homocedasticidade dentre os grupos, ou seja, se as variâncias de cada grupo são iguais entre si. "Ele é tão sensível à violações de normalidade como o Teste de Bartlett. Porém, quando os "n" são iguais em cada grupo, a ANOVA é robusta às violações da homocedasticidade e o Teste de Levene se torna pouco útil" [7].
Caso não haja homocedasticidade, olhe o tamanho das amostras. Se elas não forem de tamanhos iguais ou semelhantes, não há teste paramétrico alternativo para utilizar.

Teste de Hipóteses

$$\begin{cases} H_0: & \text{As variâncias dentro dos grupos são homogêneas} \\ H_a: & \text{As variâncias dentro dos grupos não são homogêneas} \end{cases}$$

Tabela

Teste de Homogeneidade de Variâncias

Vendas

Estatística de Levene	gl1	gl2	Sig.
2,126	2	15	,154

18.3.2 Teste de Tukey

O teste de Tukey é um dos testes de comparação de médias mais utilizados por ser bastante rigoroso. Ele não permite comparar grupos entre si e é utilizado para testar toda e qualquer diferença entre duas médias de tratamento. É aplicado quando o teste F para tratamentos (1, 2, 3, ...) da ANOVA é significativo.

Tabela

Vendas

Tukey HSD[a]

Embalagem	N	Subconjunto para alfa = 0.05	
		1	2
Embalagem B	6	11,967	
Embalagem A	6	12,600	12,600
Embalagem C	6		12,950
Sig.		,056	,365

São exibidas as médias para os grupos em subconjuntos homogêneos.

a. Usa o Tamanho da Amostra de Média Harmônica = 6,000.

O teste separa as médias dos grupos em subconjuntos. Quando dois ou mais grupos se encontram no mesmo subconjunto isso mostra que suas médias são iguais - a partir do nível de significância (α) aplicado.

Neste caso, as embalagens A e B são iguais, assim como A e C, mas as embalagens B e C são diferentes, pois elas não aparecem juntas em nenhum subconjunto.
A depender do que está sendo analisado, escolhe-se o tratamento (k) que se destaca mais (não é igual aos outros) e que melhor se aplica à investigação (neste caso a embalagem C, pois sua média de vendas é maior do que A).

18.3.3 Tabela de Comparações Múltiplas

A tabela de comparações múltiplas compara as médias de diferentes tratamentos. Os testes seguem as seguintes hipóteses:

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_a: & \mu_1 \neq \mu_2 \end{cases} \quad \begin{cases} H_0: & \mu_1 = \mu_3 \\ H_a: & \mu_1 \neq \mu_3 \end{cases} \quad \begin{cases} H_0: & \mu_2 = \mu_1 \\ H_a: & \mu_2 \neq \mu_1 \end{cases} \quad \begin{cases} H_0: & \mu_2 = \mu_3 \\ H_a: & \mu_2 \neq \mu_3 \end{cases}$$

E por ae vai, comparando todos os tratamentos entre si.

Tabela

Comparações múltiplas

Variável dependente: Vendas

Tukey HSD

(I) Embalagem	(J) Embalagem	Diferença média (I-J)	Erro Padrão	Sig.	Intervalo de Confiança 95%	
					Limite inferior	Limite superior
Embalagem A	Embalagem B	,6333	,2498	,056	-,016	1,282
	Embalagem C	-,3500	,2498	,365	-,999	,299
Embalagem B	Embalagem A	-,6333	,2498	,056	-1,282	,016
	Embalagem C	-,9833*	,2498	,004	-1,632	-,334
Embalagem C	Embalagem A	,3500	,2498	,365	-,299	,999
	Embalagem B	,9833*	,2498	,004	,334	1,632

*. A diferença média é significativa no nível 0.05.

Neste caso podemos verificar que as médias de B e C são diferentes entre si, pois a significância é menor que 5% ($\alpha < 0.05$).

18.3.4 Tabela ANOVA

A tabela ANOVA comparada as médias dos grupos estudados, conforme já descrito mais acima. As hipóteses utilizadas são:

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \mu_3 \\ H_a: & \text{as médias de pelo menos dois grupos são diferentes} \end{cases}$$

Ela pode analisar dois ou mais tratamentos. A lógica é sempre a mesma.

Tabela

ANOVA

Vendas

	Soma dos Quadrados	gl	Quadrado Médio	F	Sig.
Entre Grupos	2,981	2	1,491	7,961	,004
Nos grupos	2,808	15	,187		
Total	5,789	17			

Neste caso rejeitamos a hipótese nula, pois a significância é menor que 5% ($\alpha < 0.05$).

18.4 ANOVA 2 Fatores [7]

"Sendo uma extensão da ANOVA, a ANOVA dois fatores permite analisar modelos de efeitos fixos ou mistos, analisar as tendências dos dados, proceder comparações múltiplas, analisar o efeitos das variáveis e controlar variáveis externas (fatores de segmentação da amostra)". É necessário primeiro fazer a ANOVA 1 fator para cada fator e seus tratamentos, antes de fazer uma ANOVA de 2 (ou mais) fatores, para verificar se os pressupostos da ANOVA estão satisfeitos.

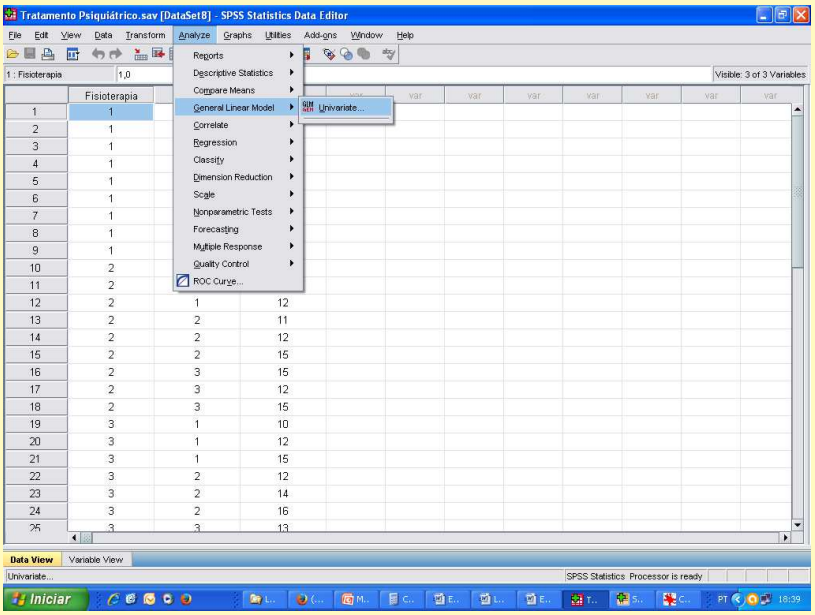
Teste de Hipóteses

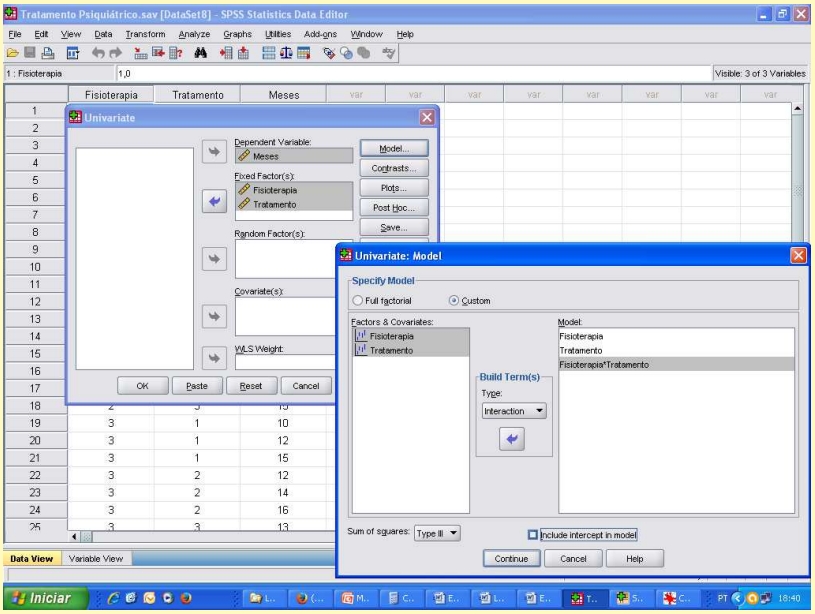
$$\begin{cases} H_0: & \text{não existe interação entre os fatores} \\ H_a: & \text{existe interação entre os fatores} \end{cases}$$

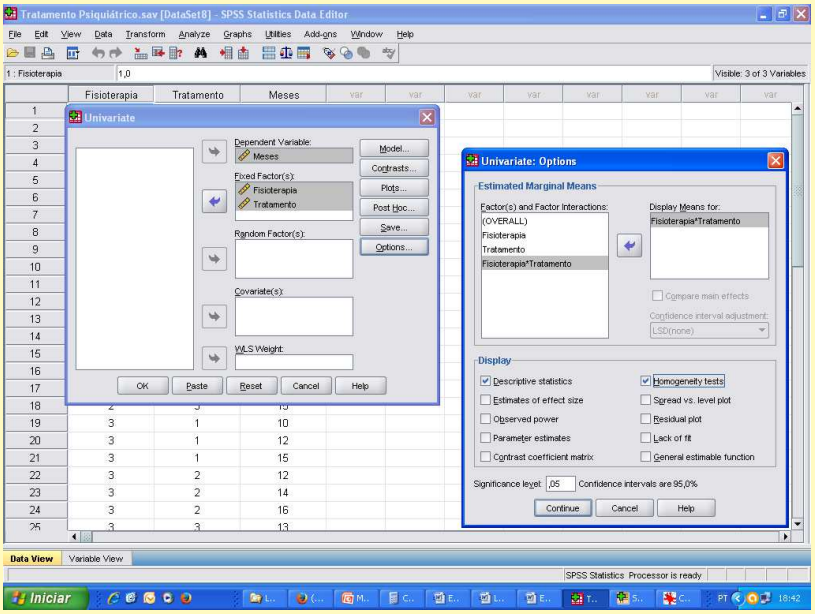
18.5 ANOVA 2 Fatores no SPSS [7]

Para realizar a ANOVA 2 fatores no SPSS vá em Analisar > Modelo linear geral > Univariado.

18.5.1 Passo a Passo







18.5.2 Análise

Primeiro verifique a homocedasticidade dente os fatores pelo teste de Levene.

O teste de hipótese da ANOVA 2 fatores se encontra na tabela de "Testes de efeitos entre sujeitos". O que nos importa aqui é a interação entre os fatores ("Fator 1*Fator 2"), logo, verifique a significância da interação pelas hipóteses da ANOVA 2 fatores.

19 Análise Discriminante

"A análise discriminante é uma **técnica multivariada** que **cria funções discriminantes**, provenientes de combinações lineares das variáveis iniciais, que maximizam as diferenças entre as médias dos grupos e minimizam a probabilidade de classificações incorretas dos casos nos grupos.

É aplicada quando a **variável dependente é qualitativa** (grupos) e as **variáveis independentes são quantitativas**. As variáveis dicotômicas, como sexo, podem também ser incluídas nas variáveis explicativas.

A análise discriminante **tem por objetivo escolher as variáveis que distinguem os grupos**, de modo que, conhecendo-se as características de um novo caso, se possa prever a que grupo pertence.

A análise discriminante pode ser usada também para validar a análise de cluster e confirmar os resultados da análise fatorial". [7]

19.1 Hipóteses

$$\begin{cases} H_0: & \text{Não há diferença. As variáveis não discriminam os objetos investigados} \\ H_a: & \text{Há diferença. As variáveis discriminam os objetos investigados} \end{cases}$$

Observações

Se as variáveis têm uma diferença de grandeza entre elas, será preciso então padronizá-las para que a análise discriminante não as discrimine por essa razão.

19.2 Pressupostos

19.2.1 Cada grupo é uma amostra aleatória de uma população normal multivariada

"A sua violação pode levar a decisões incorretas, principalmente quando as amostras são pequenas. **Quando a violação da normalidade se deve apenas à não simetria da distribuição, a potência do teste não é afetada**, contrariamente ao que acontece se a **distribuição não for mesocúrtica** e, de forma mais acentuada, se for platicúrtica, caso em que devemos optar pela regressão logística". [7]

Para medir a curtose, divida a "estatística" pelo "erro padrão" (SPSS). Para verificar se ela é mesocúrtica, veja se o valor fica entre $-1,96 < x > +1,96$. (Padronização Z: $1,96 = 0,475 \rightarrow 95\%$ no total (bicaudal)).

Verifique os outliers. A remoção de outliers pode aproximar as variáveis para a normal. Outliers moderados não costumam criar grande impacto nas análises.

19.2.2 Dentro dos grupos a variabilidade é idêntica, isto é, as matrizes de variância e covariância são iguais para todos os grupos

Teste homocedasticidade, equivalente multivariado ao teste de Levene.

"A verificação deste pressuposto é feita na própria análise discriminante, através do **teste Box's M**. Caso seja violado, aumenta a probabilidade dos casos serem classificados no grupo com maior dispersão.

A violação deste pressuposto afeta a análise principalmente quando os grupos não têm igual dimensão, mesmo que as diferenças sejam moderadas" [7].

Dicas e Recomendações [7]

- O número mínimo de observações por variável independente: 5 ;
- Número recomendado de observações por variável: 20 ;
- Número de observações por grupo: o menor grupo deve ter um tamanho que exceda o número de variáveis independentes;
- É recomendável que o número mínimo de casos em cada grupo seja 20, e que os grupos tenham dimensões semelhantes.

19.3 Análise Discriminante no SPSS [7]

19.3.1 Teste M de Box(*Box's M*)

"Este teste verifica um dos pressupostos na análise discriminante (**pressuposto da homocedasticidade**). Ele testa se as diferentes dispersões são ou não estatisticamente significativas. É muito sensível a afastamentos da normalidade".

O teste Box's M é muito "sensível à desvios de normalidade e à dimensão das amostras (amostras grandes conduzem geralmente à rejeição de Ho)". Para solucionar uma possível violação do pressuposto da homocedasticidade é importante ter uma amostra grande com o mesmo número de observações em cada grupo. "Segundo alguns autores, a análise discriminante é uma técnica bastante robusta à violação dos pressupostos, desde que a **dimensão do menor grupo seja superior ao número de variáveis independentes em estudo**".

Hipóteses

$$\begin{cases} H_0: & \text{As matrizes observadas de variância-covariância são iguais entre grupos} \\ H_a: & \text{As matrizes observadas de variância-covariância não são iguais entre grupos} \end{cases}$$

Tabela

Resultados do teste

M de Box		42,092
F	Aprox.	13,877
	gl1	3
	gl2	7056720,000
	Sig.	,000

Testa hipótese nula de matrizes de covariâncias de população igual.

19.3.2 Teste Lambda de Wilks (*Wilk's Lambda*)

"O Wilk's Lambda dá informação sobre as diferenças entre os grupos, para cada variável individualmente. Obtém-se pela razão entre a variação dentro dos grupos e a variação total. Este teste é robusto a violação do Teste Box's M quando os grupos têm dimensões semelhantes. A não rejeição da hipótese de igualdade da média de uma variável nos grupos ($\alpha > 0,05$), aumenta a probabilidade de ser classificada incorretamente em outro grupo".

Tabela de Estatísticas de Grupo

Será que essas médias são significativamente diferentes (pela variabilidade das mesmas) ou elas podem ser consideradas iguais?

Estatísticas de grupo

X25 -- Concorrente		Média	Desvio Padrão	N válido (listwise)	
				Não ponderado	Ponderado
Samouel's	X6 -- Funcionários simpáticos	2,89	1,091	100	100,000
	X11 -- Funcionários cordiais	1,96	,871	100	100,000
	X12 -- Funcionários competentes	1,62	,663	100	100,000
Gino's	X6 -- Funcionários simpáticos	4,42	,878	100	100,000
	X11 -- Funcionários cordiais	2,84	,813	100	100,000
	X12 -- Funcionários competentes	2,75	,880	100	100,000
Total	X6 -- Funcionários simpáticos	3,66	1,251	200	200,000
	X11 -- Funcionários cordiais	2,40	,949	200	200,000
	X12 -- Funcionários competentes	2,19	,962	200	200,000

Hipóteses

$$\begin{cases} H_0: & \text{Não existe diferença entre as médias dos grupos} \\ H_a: & \text{Existe diferença entre as médias dos grupos} \end{cases}$$

Tabela Lambda de Wilks

Testes de igualdade de médias de grupo

	Lambda de Wilks	F	gl1	gl2	Sig.
X6 -- Funcionários simpáticos	,624	119,366	1	198	,000
X11 -- Funcionários cordiais	,783	54,821	1	198	,000
X12 -- Funcionários competentes	,653	105,073	1	198	,000

"Neste caso, como temos 3 variáveis, devemos comparar a significância com $\alpha/3 = 0,05/3 = 0,017$ e não com $\alpha = 0,05$. Como todas as significâncias são $< 0,017$, devemos rejeitar H_0. A tabela mostra que existe diferenças significativas nas médias de cada variável nos dois grupos (significâncias $= 0,000$), não informando, entretanto, sua importância para discriminar grupos".

19.3.3 Contribuição das Variáveis

Matrizes intragupos em pool *Pooled within-groups matrices*

"Na sua interpretação tem de se levar em consideração a correlação entre as variáveis explicativas, pois se duas variáveis tiverem correlação 1, incluir ambas não fornece mais informação do que incluir uma só. Por isso o uso da opção *Stepwise*".

Matrizes intragrupos em pool

		X6 -- Funcionários simpáticos	X11 -- Funcionários cordiais	X12 -- Funcionários competentes
Correlação	X6 -- Funcionários simpáticos	1,000	,527	,454
	X11 -- Funcionários cordiais	,527	1,000	,340
	X12 -- Funcionários competentes	,454	,340	1,000

Nessa tabela, "como podemos observar, não existe problema de **multicolinearidade**, uma vez que nenhum coeficiente de correlação entre variáveis independentes é superior, em termos absolutos, a $0,9$.

Quando há multicolinearidade, não se deve analisar a importância de cada variável para a análise, visto que sua elevada correlação com outras a torna redundante. Nesta situação apenas se usa o procedimento *Stepwise*".

Estatísticas *Stepwise*

Variáveis Inseridas/Removidas[a,b,c,d]

Passo	Inseridas	Lambda de Wilks							
		Estatística	gl1	gl2	gl3	F exato			
						Estatística	gl1	gl2	Sig.
1	X6 -- Funcionários simpáticos	,624	1	1	198,000	119,366	1	198,000	,000
2	X12 -- Funcionários competentes	,561	2	1	198,000	76,942	2	197,000	,000

Em cada passo, a variável que minimiza o Lambda de Wilks geral é inserida.

a. O número máximo de passos é 6.

b. O F parcial mínimo a ser inserido é 3.84.

c. O F parcial máximo a ser removido é 2.71.

d. Nível f, tolerância ou VIN insuficiente para cálculos adicionais.

$$\begin{cases} H_0: & \text{A separação dos grupos não foi bem sucedida (em função das médias)} \\ H_a: & \text{A separação dos grupos foi bem sucedida (em função das médias)} \end{cases}$$

"A tabela acima resume o procedimento Stepwise, indicando para cada passo que variáveis foram adicionadas/removidas, o valor de Wilk's Lambda e a significância. Note que a cada passo, a variável "escolhida" é aquela que minimize o valor de Wilk's Lambda, isto é, aquela para a qual ocorrem maiores diferenças entre as médias dos grupos, até que não ocorram variação significativas de Lambda.

O Teste Wilk's Lambda avalia se o modelo consegue separar e classificar bem os grupos.

Neste caso, com uma significância menor que $0,05$, podemos dizer que a separação dos grupos foi bem sucedida".

Variáveis na análise

Passo		Tolerância	F a ser removido	Lambda de Wilks
1	X6 -- Funcionários simpáticos	1,000	119,366	
2	X6 -- Funcionários simpáticos	,794	32,236	,653
	X12 -- Funcionários competentes	,794	21,912	,624

Variáveis não presentes na análise

Passo		Tolerância	Mín. Tolerância	F a ser inserido	Lambda de Wilks
0	X6 -- Funcionários simpáticos	1,000	1,000	119,366	,624
	X11 -- Funcionários cordiais	1,000	1,000	54,821	,783
	X12 -- Funcionários competentes	1,000	1,000	105,073	,653
1	X11 -- Funcionários cordiais	,723	,723	2,342	,617
	X12 -- Funcionários competentes	,794	,794	21,912	,561
2	X11 -- Funcionários cordiais	,710	,638	,747	,559

"O quadro acima apresenta, para cada passo, as variáveis que foram consideradas como discriminantes na análise. Essas variáveis foram escolhidas como as "melhores", com base na matriz de correlação, seu poder de discriminação entre grupos e cálculo da tolerância (avaliação da multicolinearidade. Valores muito baixos demonstrariam multicolinearidade).

Sumarização de funções discriminantes canônicas

Autovalores

Função	Autovalor	% de variância	% cumulativa	Correlação canônica
1	,781[a]	100,0	100,0	,662

a. As primeiras 1 funções discriminantes canônicas foram usadas na análise.

"A tabela acima indica que a variância (em termos da diferença entre os grupos) é explicada em 100% pela função discriminante".

Lambda de Wilks

Teste de funções	Lambda de Wilks	Qui-quadrado	gl	Sig.
1	,561	113,719	2	,000

$$\begin{cases} H_0: & \text{A função discriminante não é significativa} \\ H_a: & \text{A função discriminante é significativa} \end{cases}$$

Neste caso "a função discriminante é significativa para separar os grupos".

Coeficientes de função discriminante canônica

	Função
	1
X6 -- Funcionários simpáticos	,642
X12 -- Funcionários competentes	,688
(Constante)	-3,848

Coeficientes não padronizados

A tabela acima é "usada para escrever a(s) função(ões) discriminante(s)".

$$\text{Função} = -3,848 + (0,642 * \text{X6}) + (0,688 * \text{X12})$$

Considere X6 = 2 e X12 = 1 → Função = −1, 8768 → Pertence ao Samouel's (Olhe a tabela do centróide mais abaixo)

Considere X6 = 4 e X12 = 3 → Função = 0, 7818 → Pertence ao Gino's (Olhe a tabela do centróide mais abaixo)

Estatísticas de classificação

Resultados da classificação[a,c]

			Associação ao grupo predita		
		X25 -- Concorrente	Samouel's	Gino's	Total
Original	Contagem	Samouel's	76	24	100
		Gino's	8	92	100
	%	Samouel's	76,0	24,0	100,0
		Gino's	8,0	92,0	100,0
Com validação cruzada[b]	Contagem	Samouel's	76	24	100
		Gino's	8	92	100
	%	Samouel's	76,0	24,0	100,0
		Gino's	8,0	92,0	100,0

a. 84,0% de casos agrupados originais classificados corretamente.

b. A validação cruzada é feita apenas para os casos da análise. Na validação cruzada, cada caso é classificado pelas funções derivadas de todos os casos diferentes desse caso.

c. 84,0% de casos agrupados com validação cruzada classificados corretamente.

"Finalmente, a tabela acima apresenta os resultados da classificação. Repare que 76% dos clientes do Samouel's foram classificados corretamente e que 24% dos clientes foram classificados como clientes do Ginos's. Além disso, 92% dos clientes do Gino's foram classificados corretamente e 8% não".

Coeficientes de função de classificação

	X25 -- Concorrente	
	Samouel's	Gino's
X6 -- Funcionários simpáticos	2,512	3,641
X12 -- Funcionários competentes	1,219	2,428
(Constante)	-5,310	-12,078

Funções discriminantes lineares de Fisher

A tabela acima é utilizada para escrever a(s) função(ões) para alocação dos dados.

$$\text{Função Samouel's} = -5,310 + (2,512 * \text{X6}) + (1,219 * \text{X12})$$

$$\text{Função Gino's} = -12,078 + (3,641 * \text{X6}) + (2,428 * \text{X12})$$

Condidere X6 = 2 e X12 = 1

Função Samouel's $= -5,310 + (2,512 * 2) + (1,219 * 1) = 0,933$

Função Gino's $= -12,078 + (3,641 * 2) + (2,428 * 1) = -2,368$

Condidere X6 = 4 e X12 = 3

Função Samouel's $= -5,310 + (2,512 * 4) + (1,219 * 3) = 8,359$

Função Gino's $= -12,078 + (3,641 * 4) + (2,428 * 3) = 9,77$

Funções em centroides de grupo

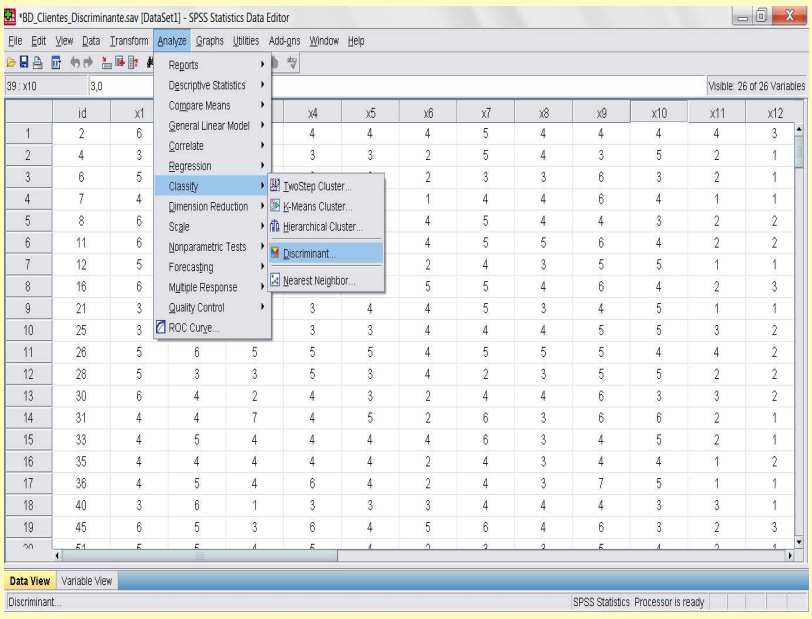
	Função
X25 -- Concorrente	1
Samouel's	-,879
Gino's	,879

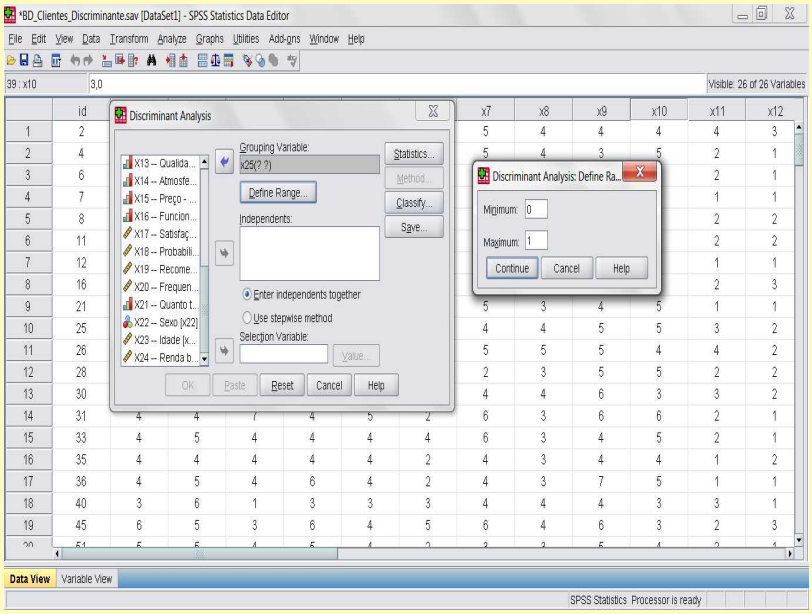
Funções discriminantes canônicas não padronizadas avaliadas em médias de grupo

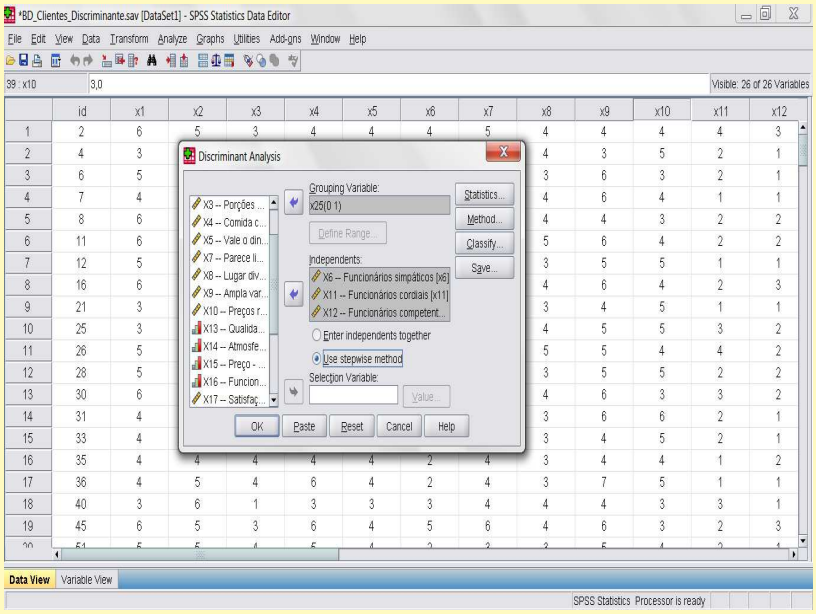
"Uma abordagem final para exame de diferenças entre grupos é o centróide. Em uma análise discriminante com dois grupos temos dois centróides, com três grupos três centróides, etc. Os centróides para nosso exemplo mostra que o do Samouel's é $-0,879$ e do Gino's $0,879$. Esta é uma medida de síntese global que indica que o Gino's é percebido de maneira muito mais positiva do que o Samuel's".

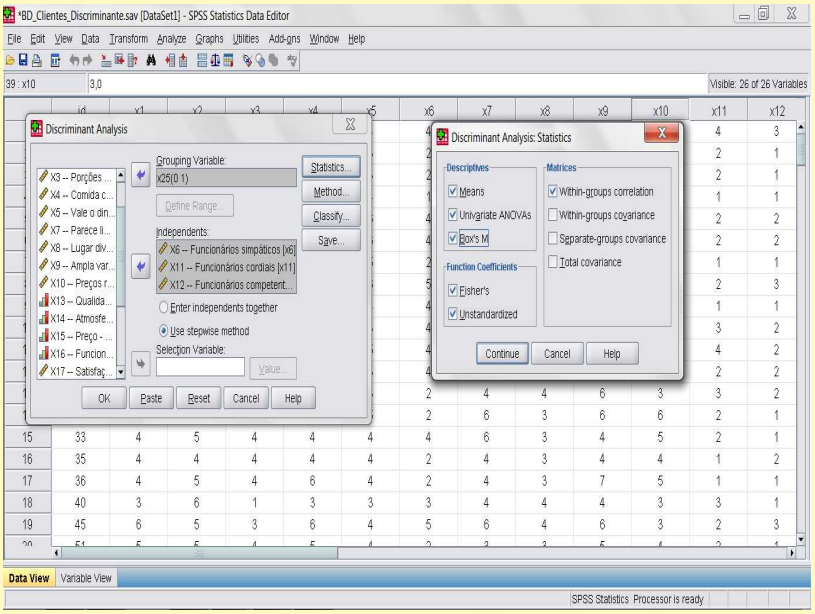
19.3.4 Passo a Passo

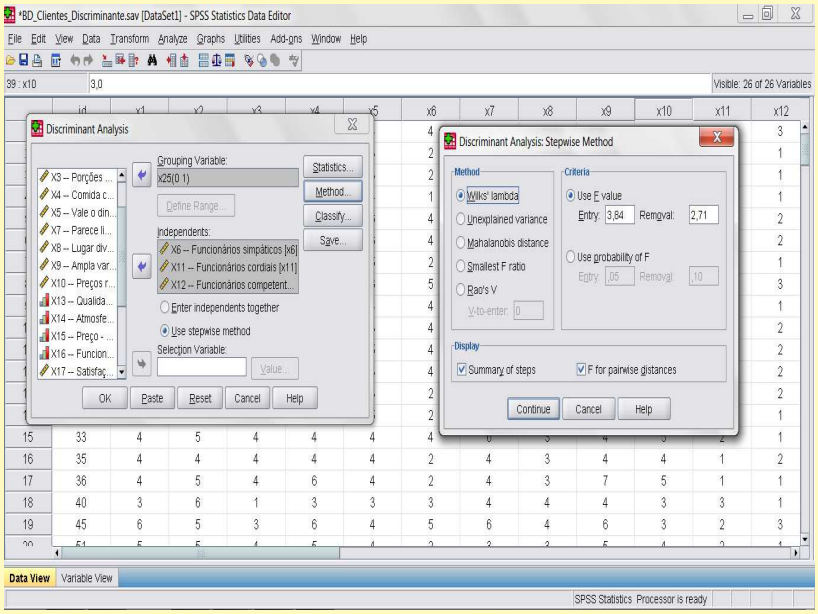
Não esqueça de verificar os pressupostos da análise antes de fazê-la.

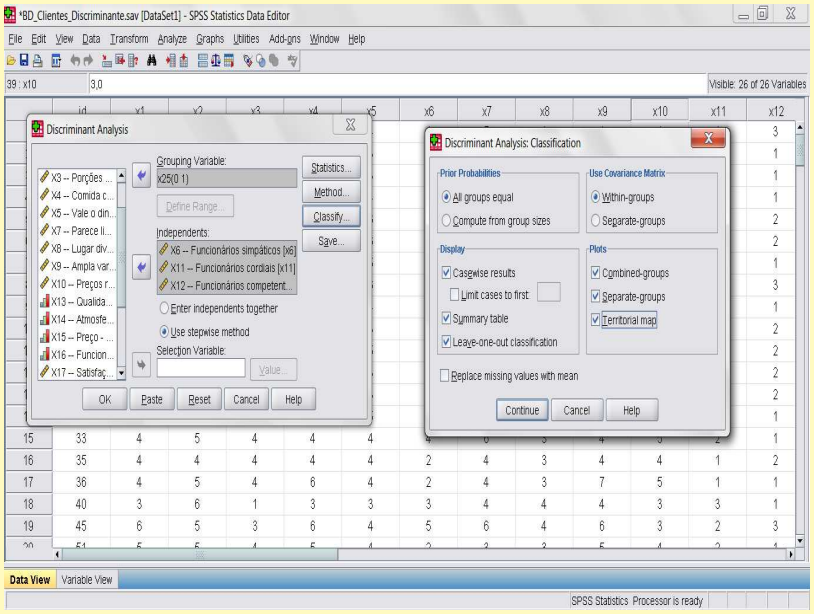


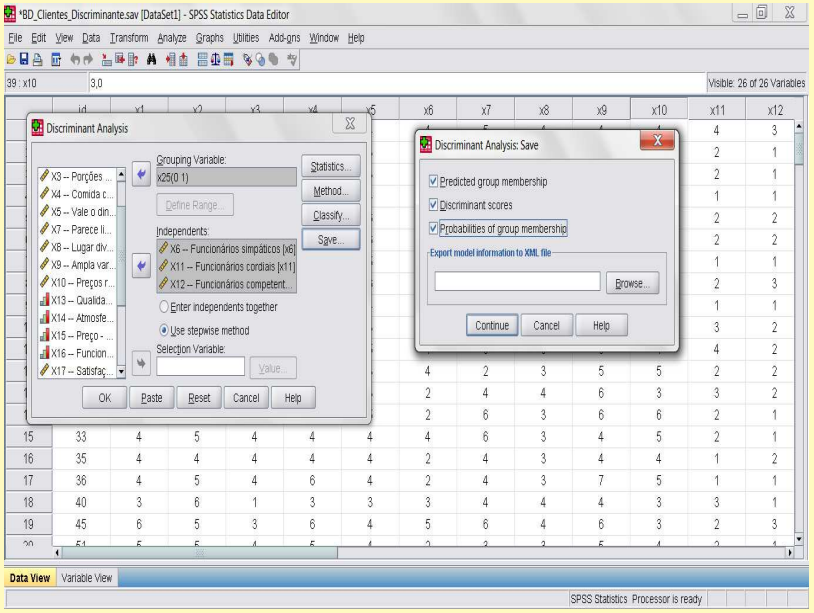

*BD_Clientes_Discriminante.sav [DataSet1] - SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Discriminant Analysis
Grouping Variable:
x25(? ?)
Define Range...
Independents:
Enter independents together
Use stepwise method
Selection Variable:
Statistics...
Method...
Classify...
Save...
OK Paste Reset Cancel Help
Discriminant Analysis: Define Ra...
Minimum: 0
Maximum: 1
Continue Cancel Help
Data View Variable View
SPSS Statistics Processor is ready


*BD_Clientes_Discriminante.sav [DataSet1] - SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Discriminant Analysis
Grouping Variable:
x25(0 1)
Define Range...
Independents:
X6 -- Funcionários simpáticos [x6]
X11 -- Funcionários cordiais [x11]
X12 -- Funcionários competent...
Enter independents together
Use stepwise method
Selection Variable:
Statistics...
Method...
Classify...
Save...
OK Paste Reset Cancel Help
Data View Variable View
SPSS Statistics Processor is ready


*BD_Clientes_Discriminante.sav [DataSet1] - SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Discriminant Analysis
Grouping Variable:
x25(0 1)
Define Range...
Independents:
X6 -- Funcionários simpáticos [x6]
X11 -- Funcionários cordiais [x11]
X12 -- Funcionários competent...
Enter independents together
Use stepwise method
Selection Variable:
Value...
Statistics...
Method...
Classify...
Save...
OK
Paste
Reset
Cancel
Help
Discriminant Analysis: Statistics
Descriptives
Means
Univariate ANOVAs
Box's M
Matrices
Within-groups correlation
Within-groups covariance
Separate-groups covariance
Total covariance
Function Coefficients
Fisher's
Unstandardized
Continue
Cancel
Help
Data View
Variable View
SPSS Statistics Processor is ready


*BD_Clientes_Discriminante.sav [DataSet1] - SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Discriminant Analysis
Grouping Variable:
x25(0 1)
Define Range...
Independents:
X6 -- Funcionários simpáticos [x6]
X11 -- Funcionários cordiais [x11]
X12 -- Funcionários competent...
Enter independents together
Use stepwise method
Selection Variable:
Value...
Statistics...
Method...
Classify...
Save...
OK
Paste
Reset
Cancel
Help
Discriminant Analysis: Stepwise Method
Method
Wilks' lambda
Unexplained variance
Mahalanobis distance
Smallest F ratio
Rao's V
V-to-enter: 0
Criteria
Use F value
Entry: 3,84
Removal: 2,71
Use probability of F
Entry: .05
Removal: .10
Display
Summary of steps
F for pairwise distances
Continue
Cancel
Help
Data View
Variable View
SPSS Statistics Processor is ready





20 Teoria da Amostragem [7]

$\overline{X} \pm Z * \sigma_{\overline{X}}$

$\overline{X} \pm Z * cfraco\sigma\sqrt{n}$

20.1 Tamanho da Amostra

20.1.1 Amostra Aleatória Simples

Para estimar a média: com reposição

Sabemos que Z (Erro amostral) = Erro absoluto, isto é:

$Z_x \sigma_{\overline{X}} = e$

onde

...

21 Análise Fatorial [7]

(...)

21.1 O que é?

É o conjunto de técnicas estatísticas, que procura explicar a correlação entre as variáveis observáveis, simplificando os dados através da redução do número...

21.2 Objetivos

O objetivo principal da análise fatorial é a redução da massa de dados altamente relacionados. Entretanto, a redução não implica em admitir perda do comportamento das variáveis originais. Nesse sentido quer-se reduzir a massa de dados, mas ao mesmo tempo, preservar ao máximo as características comportamentais do conjunto de variáveis.

A análise fatorial também é utilizada para gerar fatores compostos pelas variáveis originais, que podem ser utilizados em outras técnicas de análise multivariada.

Como os fatores gerados em uma análise fatorial podem não apresentar correlação entre si, então podem ser utilizados para resolver o problema de variáveis multi-correlacionadas na regressão.

21.3 Dimensão da Amostra

O mínimo de respostas válidas (N) por variáveis (K) é:

$N = 50 \text{se} K <= 5 N = 10 * K se 5 < K <= 15 N = 5 * K se K > 15$

(Gráfico)

21.4 Análise Fatorial Exploratória e de Confirmação

21.5 Exemplo:

Phil gostaria de saber se poderia simplificar seu entendimento das percepções dos dois restaurantes, reduzindo o número de variáveis para menos de 12 (tratar como amostra única). Isto é, se ele pode representar as 12 variáveis originais de percepção (X_1 até X_{12}) com um número menor de fatores significativos. A que resultados chegou o consultor?

21.5.1 Variáveis do banco de dados

X1 - Comida de excelente qualidade

X2 - Um interior atraente

X3 - Porções generosas

X4 - Comida de gosto excelente

X5 - Bom valor para o dinheiro

X6 - Funcionários simpáticos

X7 - Aparência limpa e organizada

X8 - Um ambiente divertido

X9 - Grande variedade de pratos

X10 - Preços razoáveis

X11 - Funcionários gentis

X12 - Funcionários competentes

21.6 Premissas:

A análise fatorial possui algumas premissas:

1a **As variáveis devem ser medidas em escala intervalar ou razão**;

2a As variáveis envolvidas na análise devem ser normalmente distribuídas;

3a Se as variáveis envolvidas na análise estiverem medidas em escalas diferentes, as mesmas devem ser padronizadas antes de proceder a análise.

21.6.1 Quando os Dados não são Normais

Neste caso, devemos proceder da seguinte maneira: (1) usar uma transformação matemática que normalize a variável e usar a nova variável no lugar da original não normal; ou (2) alguns

autores dizem que uma variável mesmo que não seja normalmente distribuída, pode ser incluída na análise fatorial, se apresentar simetria.
Para verificar se as variáveis, mesmo não sendo normais, tem distribuição simétrica: calcular a assimetria e dividir (Stat/Std. Error). Se o valor estiver contido no intervalo [-1.96; 1,96] podemos supor que exista uma distribuição simétrica.

21.6.2 Próximo Passo

21.7 Rotação de Fatores

A matriz de fatores contém os coeficientes utilizados para expressar as variáveis em termos dos fatores. Esses coeficientes, chamados de cargas fatoriais, representam as correlações entres os fatores e as variáveis. Um coeficiente com valor absoluto grande, indica que o fator e a avariável estão estritamente relacionados.
Ao fazer a rotação dos fatores, seria interessante que cada variável tivesse coeficiente diferente de zero ou significativos com poucos fatores, sendo o ideal com apenas um, pois se vários fatores têm altas cargas com a mesma variável, tona difícil interpretá-los.

21.7.1 Tipo de Rotação de Fatores:

Ortogonais: rotação de fatores em que os eixos são mantidos em ângulos retos. A rotação ortogonal tem como resultado fatores não-correlacionados (varimax, quartimax e equamax). O mais utilizado é o varimax, que minimiza o número de variáveis com altas cargas sobre um fator, facilitando assim, a interpretação dos fatores.
Oblíquas: rotação de fatores em que os eixos não são mantidos em ângulos retos. Deve-se utilizar rotação oblíqua quando os fatores na população tendem a ser fortemente correlacionados (direct oblimin e promax)

21.8 Tipos de Extração de Fatores:

Componentes principais: Geralmente usamos o padrão: **Análise de Componentes Principais**. No SPSS, bem como em outros pacotes de software de estatística, o PCA é o método padrão para a extração de análise fatorial. Ele não é especificamente um método de análise fatorial , mas é amplamente usado como um método de extração.

É um procedimento estatístico multivariado, que permite transformar um conjunto de variáveis **quantitativas iniciais**, correlacionadas entre si, em outro conjunto com um menor número de variáveis não correlacionadas reduzindo a complexidade dos dados. **Não pressupõe a normalidade**, mas sua ausência ou presença de *outliers* **pode** provocar distorções.
As componentes principais são calculadas por ordem decrescente de importância, isto é, a primeira explica a máxima variâncias dos dados, segunda a máxima variância ainda não explicada pela primeira, e assim sucessivamente. A última componente será a que menos explica a variância total dos dados.
Máxima Verossimilhança: a **normalidade é um pressuposto exigido** por esse método, isto é, ele assume que os dados provém de uma distribuição normal **multivariada**.
Não é simplesmente fazer o teste de Komolgrov, é preciso calcular o teste "PK de Mardia" - que junto as variáveis em uma só e testa sua normalidade;
Eixo Principal de Fatoração ou Análise Fatorial Comum: Opte pela Análise Fatorial Comum **se os seus dados são significativamente anormais**. Esse método de extração ...

21.9 KMO

Para se poder aplicar o modelo fatorial, deve haver correlação entre as variáveis. Se essas correlações forem pequenas é pouco provável que partilhem fatores comuns.
O KMO é um procedimentto estatístico que permite aferir a qualidade das correlações entre as variáveis de forma a prosseguir com a análise fatorial
...

21.10 Teste de Esfericidade de Bartlett

O teste de Esfericidade de Bartlett testa a hipótese de que há correlação entre algumas variáveis.
Este teste requer que os dados provenham de uma população normal multivariada.
No entanto, este teste é muito influenciado pelo tamanhos da amostra, e eleva a rejeitar a hipótese nula em grandes amostras. Neste caso, é preferível o uso do KMO.

21.11 Extração de Fatores

O número de fatores necessários para descrever os dados, pode ser obtido através de um dos seguintes procedimentos:

*K = número de variáveis.

1. **para** $K <= 30$, usar o critério de Kaiser, pelo qual se escolhem os fatores cuja variância explicada é superior a 1 (eigenvalues > 1);

2.**para** $K > 30$m usar o Scree Plot, isto é, o gráfico da variância pelo número de componente, onde os pontos no maior declive são indicativos do número apropriado de componentes a reter.

Quando o número de casos é superior a 250 e o valor médio das comunalidades é grande ($>= 0,6$), ambos os critérios fornecem o mesmo resultado.

Quando as comunalidades são pelo menos $0,6$ e o número de variáveis é inferior a 30 ou o número de observações é superior a 250, tanto o critério de Kaiser, como o Scree Plot, geram soluções confiáveis quanto ao número de fatores a reter. Essa credibilidade é aumentada quando o quociente entre o número de fatores retidos e o número de variáveis iniciais é inferior a $0,3$.

21.12 Matriz de Anti-Imagem

A matriz de anti-imagem é uma medida de adequação amostral de cada variável (MAS) para uso da análise fatorial onde pequenos valores (menores que $0,5$) na diagonal principal nos levam a considerar a eliminação da variável.

21.13 Comunalidade

Pelo quadro das comunalidades podemos observar que todas as variáveis possuem uma forte relação cpom os fatores encontrados, exceto as variáveis X7 e X10 que estão bem próximas do limite $0,6$.

21.14 Matriz de Componentes

A matriz dos componentes mostra as coeficientes ou pesos que correlacionam as variáveis aos fatores antes da rotação.

Espera-se que não hajam pesos elevados, em mais de um fator, para uma mesma variável, pois isto dificultaria a interpretação. Muitas vezes a extração inicial ou anterior à rotação não

fornece fatores interpretáveis.

Na análise fatorial, quando há variáveis com baixos pesos, não se controla a sua influência eliminando-as e usando apenas as variáveis com pesos elevados. Cabe ao pesquisador excluí-las ou não da análise, de acordo com o fundamento teórico subjacente.

21.15 Matriz de Componentes após Rotação

A matriz dos componentes após a rotação ortogonal é útil para designar o significado dos fatores, essencialmente quando as variáveis têm pesos elevados ...

21.16 Nomeando os Fatores

...

22 Alpha de Cronbach [7]

22.1 Consistência Interna

22.1.1 O Que é?

Consistência interna de um teste (ou questionário) é uma medida que vida detectar se os itens que o compões mede o mesmo conceito (ou constructo). Por exemplo, se dez questões foram projetos para medir o mesmo conceito, reo respondentes ...

Para medir a consistência interna de um teste ou uma escala Lee J; Cronbach desenvolveu em 1951 o coeficiente alpha que hoje é a estatística ....

...

As opões de resposta para cada item podem ser dicotômicas como "Sim" e "Não" ou escalonadas ...

Seja x_{ij} o i-ésimo...

...

22.2

23 Análise de Cluster [7]

23.1 O que é?

A análise de grupos ou de clusers, é uma técnica exploratória de análise multivariada que permite agrupar sujeitos ...

A análise de cluster é um bom procedimento para exploração dos dados, quando eiste a suspeita de uqe a amostra não é homegênea.

23.2 Semelhanças

A análise de **cluster de variáveis** assemelha-se à análise fatorial porque ambos os procedimentos identificam grupos de variáveis relacionadas entre si. No entanto, nesta situação, torna-se preferível utilizat a análise faotrial, pois é uma modelo térico enquanto análise de cluster...

A análise de cluster de cados é semelhante à análise discriminante pois procura classificar um conjunto de dados iniciais...

23.3 Vale Lembrar

...

23.4 Exemplo

...

23.5 Tipos de Procedimento

Agrupamento Hierárquico

24 Séries Temporais [7]

24.1 O que é?

Uma série temporal consiste em um conjunto de observações de variáveis quantitativas coletadas ao longo do tempo. (...)

24.2 Componentes de Séries Temporais

- Tendência - Crescente ou decrescente
- Variações cíclicas - Período longo
- Variações sazonais - Período curto (1 ano)
- Variações irregulares ou aleatórias

$ST = T + \times VS + \times VC + \times VA$

24.2.1 Tendência (T):

24.2.2 Variações Cíclicas (C):

24.2.3 Variações Sazonais (S):

24.2.4 Variações Irregulares ou Aleatórias (I):

24.3 Processo de Análises

(...)

24.3.1 Forma aditiva:

Considera que a série temporal é uma soma dos quatro componentes:

$Y = T + C + S + I$

24.3.2 Forma multiplicativa:

Considera que a série temporal é um produto dos quatro componentes:

$Y = T \times C \times S \times I$

24.4 Forma Aditiva:

(...)

24.5 Formas Multiplicativa:

(...)

24.6 Técnicas de Suavização

24.6.1 Média Móvel Simples (MMS):

Uma média móvel (MM) erá o efeito de "alisar" os dados, produzindo um movimento com menos picos e vales. (...)
(...)

24.6.2 Análise da Qualidade da Previsão

(...)
Desvio Médio Absoluto:
Erro Quadrático Médio (EQM)

24.6.3 Média Móvel Ponderada (MMP)

(...)

References

[1] BERTOLO, L. A. **Probabilidades**: teorema da probabilidade total e teorema de Bayes. 2012 Disponível em: <http://www.bertolo.pro.br/AdminFin/AnalInvest/Aula040912Revisao.pdf>. Acesso em: 14 mai. 2013.

[2] BUSSAB, Wilton de O; MORETTIN, Pedro A. **Estatística básica**. 6 ed. São Paulo: Saraiva, 2010.

[3] CRESPO, Antônio A. **Estatística fácil**. 10 ed. São Paulo: Saraiva, 1993.

[4] GOOGLE IMAGENS. Disponível em: <https://images.google.com.br>.

[5] MAGALHÃES, Marcos N.; LIMA, Antonio C. P. **Noções de probabilidade e estatística**. 6 ed. São Paulo: Edusp, 2008.

[6] MORETTIN, Pedro A; HAZZAN, Samuel; BUSSAB, Wilton de O. **Cálculo** - funções de uma e várias variáveis. 1 ed. São Paulo: Saraiva, 2003.

[7] TORRES, Rosane Rivera. Pode conter adaptações.

[8] VENTURA, Marcelo Freire. Pode conter adaptações.