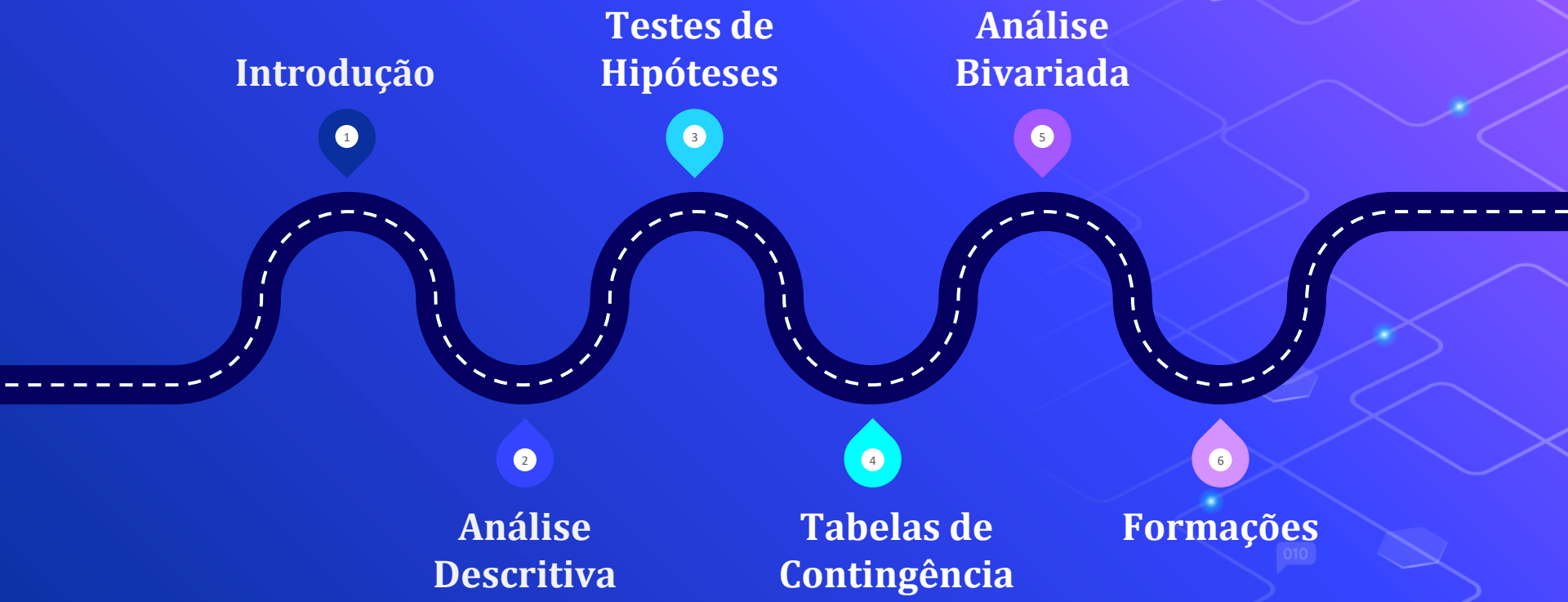
ENEMath 23

Análise de Dados em Saúde com o SPSS

Ricardo São João (PhD)
CEAUL – Centro de Estatística e Aplicações, Universidade de Lisboa
IPSantarém
ricardo.sjoao@esg.ipsantarem.pt

Roadmap IBM SPSS Statistics

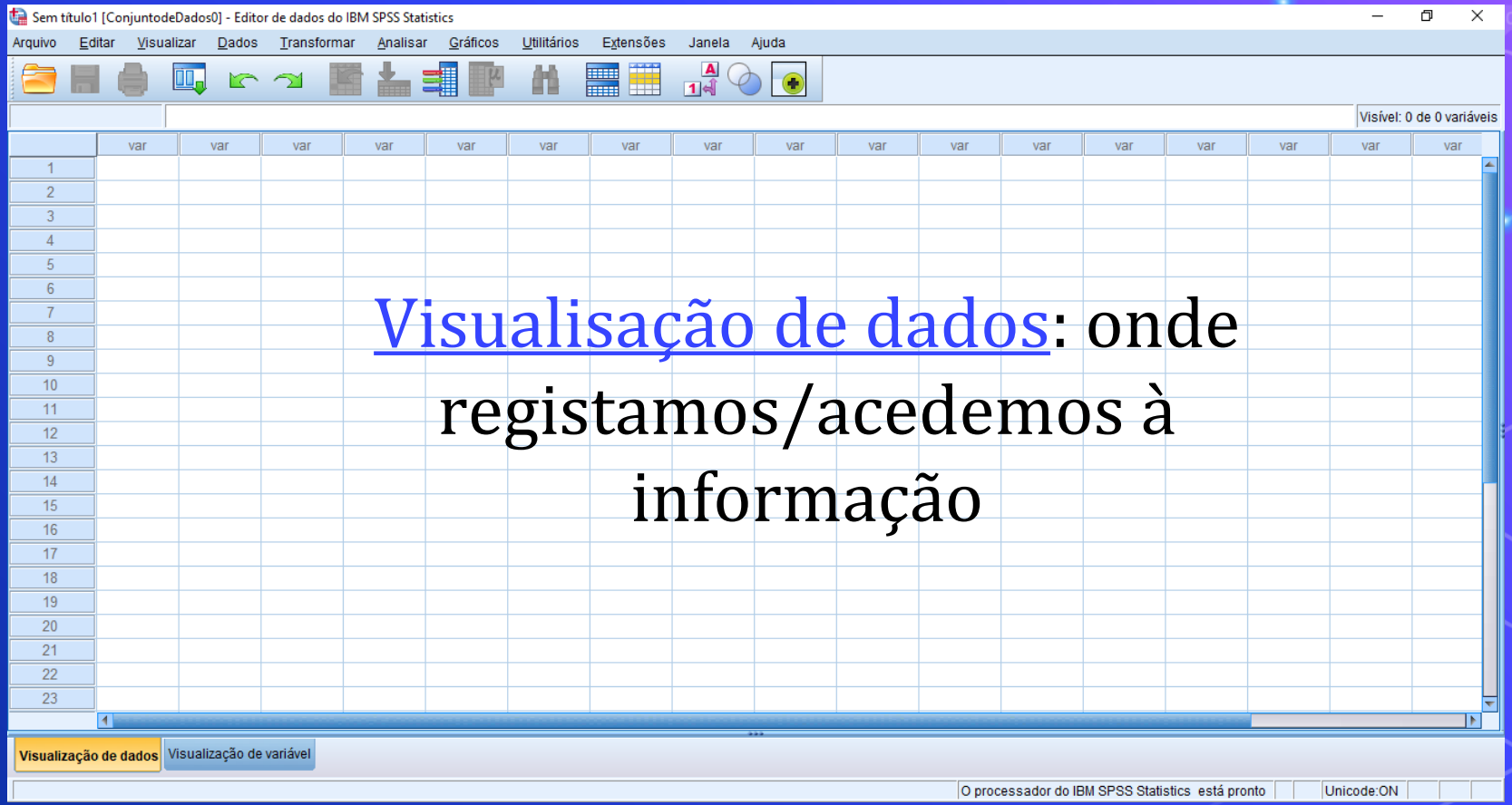


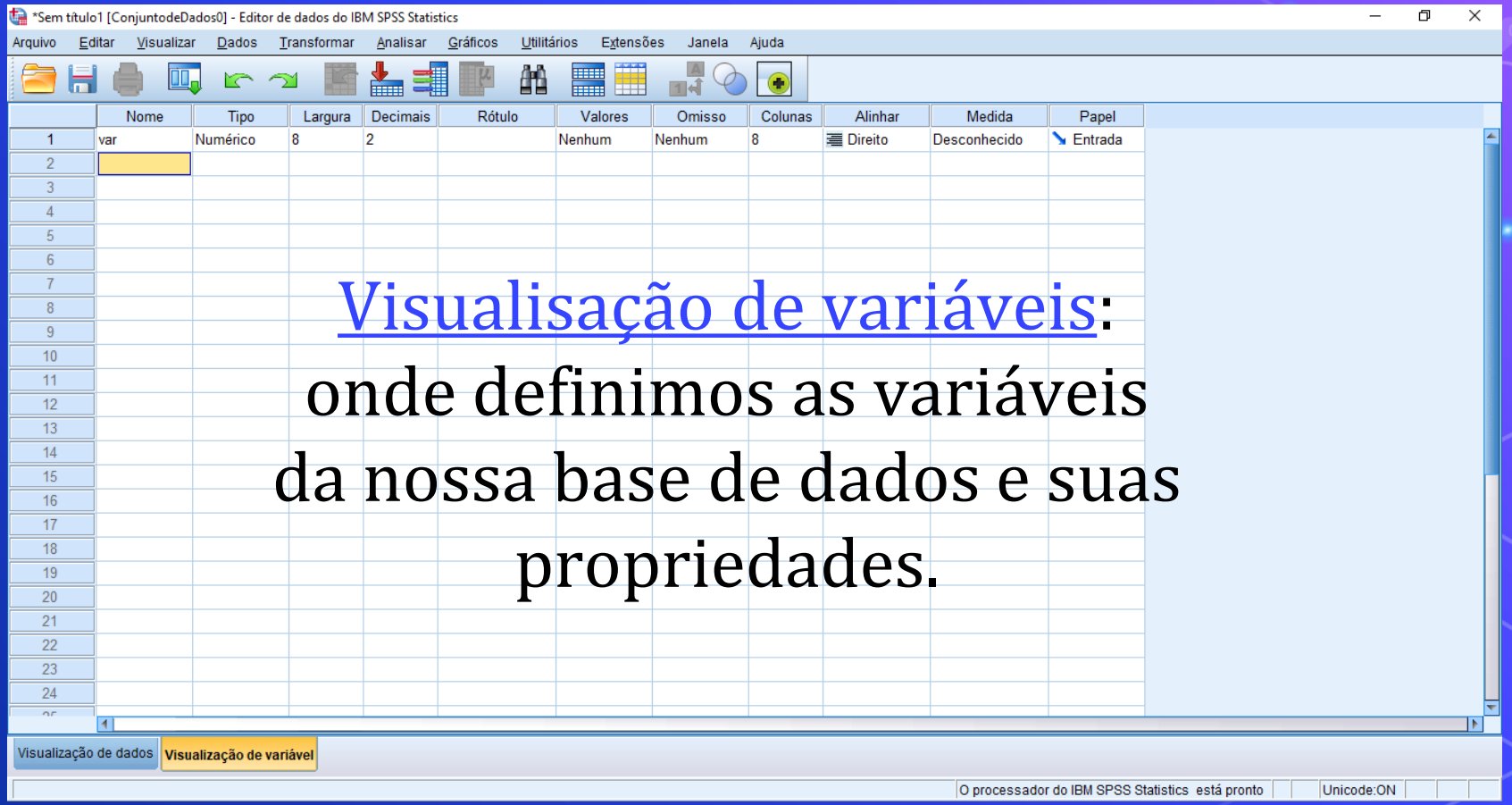
O ***SPSS*** é um software comercial lançado em 1968. Originalmente o seu nome era acrónimo de ***S****tatistical* ***P****ackage for the* ***S****ocial* ***S****ciences*. Posteriormente foi adquirido pela IBM passando a designar-se *IBM SPSS Statistics.*

Trata-se de um software "amigável" e "intuitivo" cujas diversas funcionalidades estão dispostas em menus.

Numa base de dados usualmente linhas representam registos (por exemplo indivíduos) e colunas variáveis (por exemplo características dos indivíduos).

Duas janelas:





Visualisação de variáveis:

onde definimos as variáveis da nossa base de dados e suas propriedades.

Análise Descritiva

Como seria introduzida a informação ?

- Num exame laboratorial, foi determinado o nível de glicose em 36 adultos em jejum aparentemente saudáveis. Os resultados obtidos, em mg/dL (miligramas por decilitro), foram registados assim como o género (F-feminino, M-masculino):

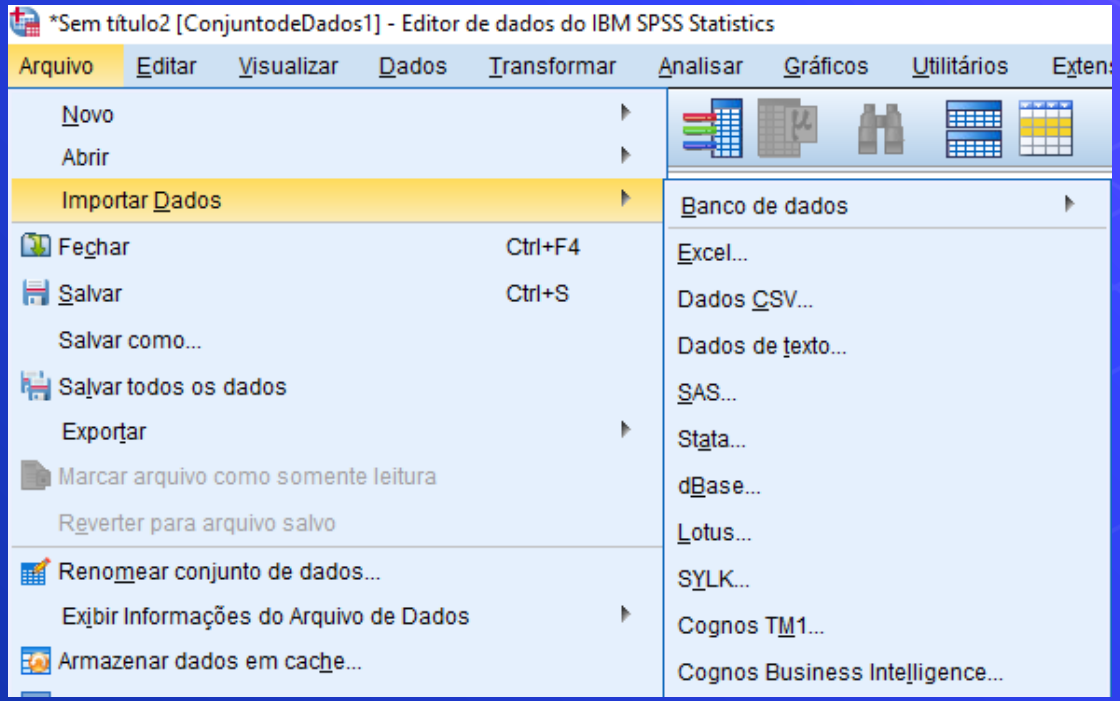
75-F	84-M	80-F	77-F	68-M	87-F	92-M	77-M	92-F	86-M	78-F	76-F
80-F	77-M	77-M	92-F	68-F	87-F	84-F	75-F	78-M	80-M	80-M	77-M
80-M	81-F	72-F	77-F	92-M	80-M	72-M	81-M	76-F	78-F	81-F	86-M

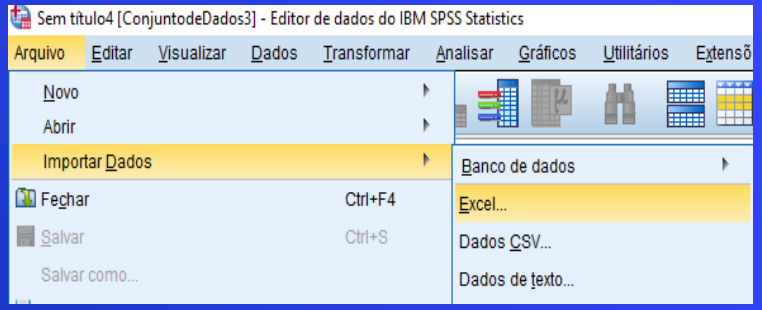
Atribuir uma codificação para as categorias da variável género. Por exemplo:
1 – Feminino;
0 - Masculino

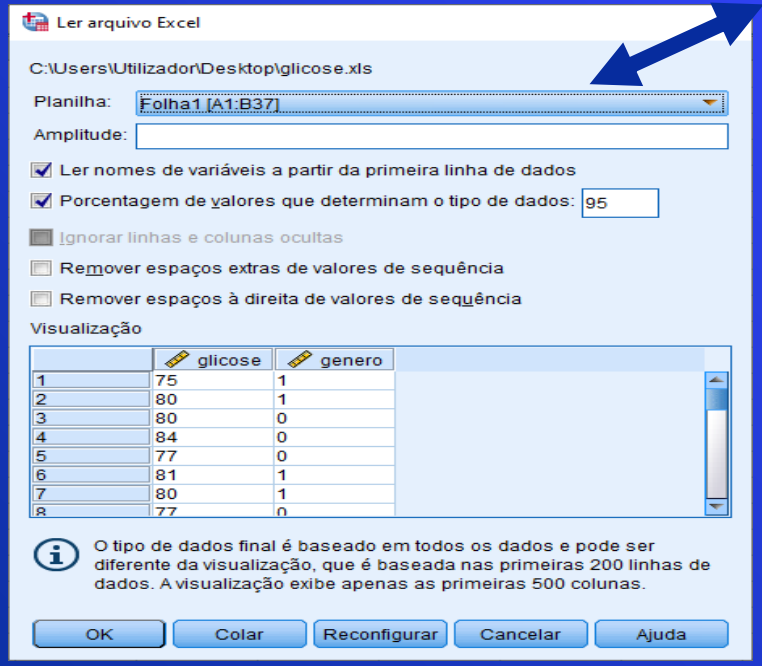
- Importe o ficheiro Excel **glicose.xlx** para o IBM SPSS.

E quando forem muitos dados disponíveis em outras plataformas ????

Menu Arquivo -> submenu Importar Dados

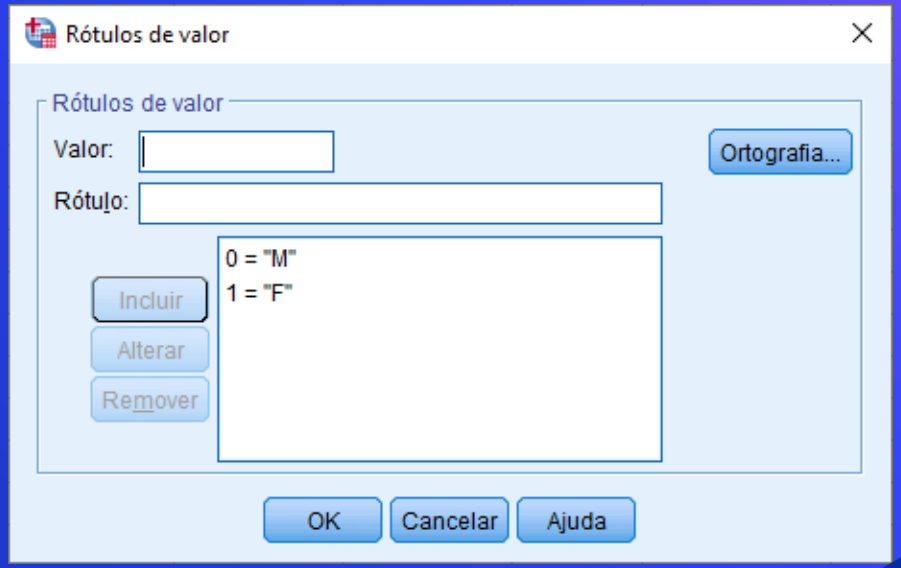


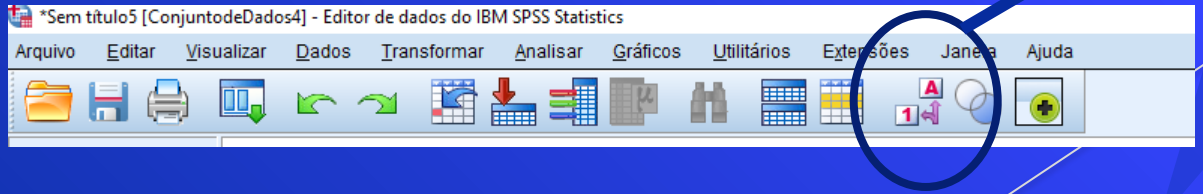




NOTA: Se o livro excel tiver mais de uma folha é necessário especificar qual

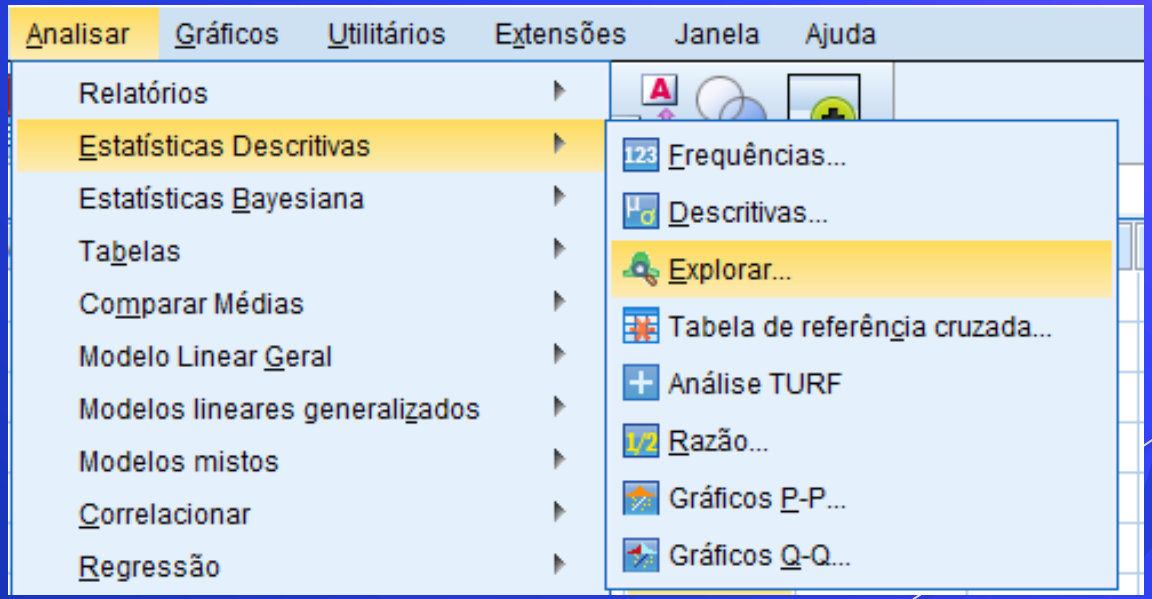
Codificação da variável gênero

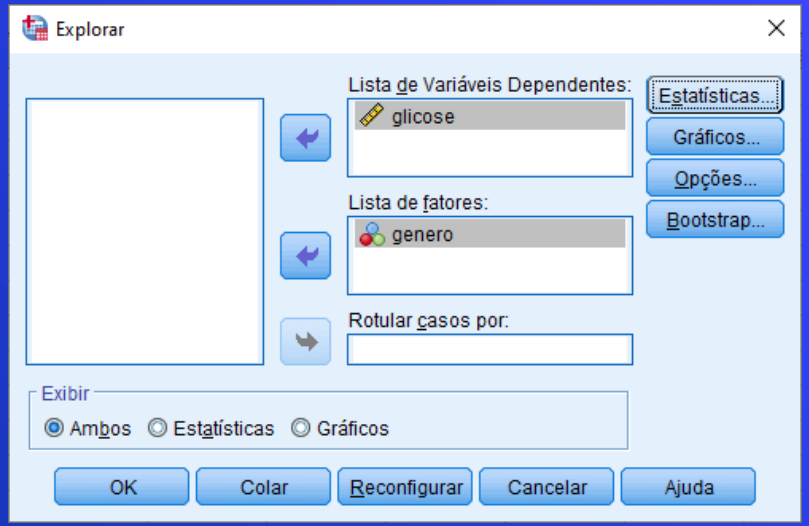


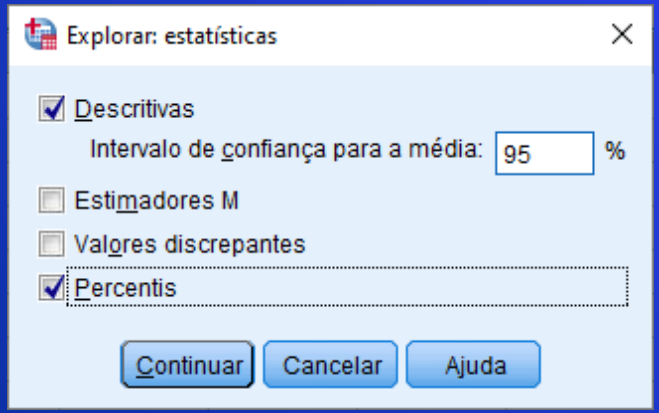


Mostrar os rótulos

Estratificação da análise estatística por género

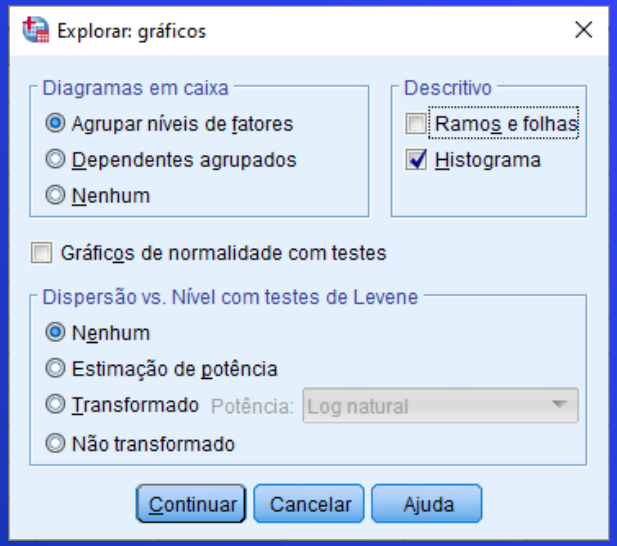


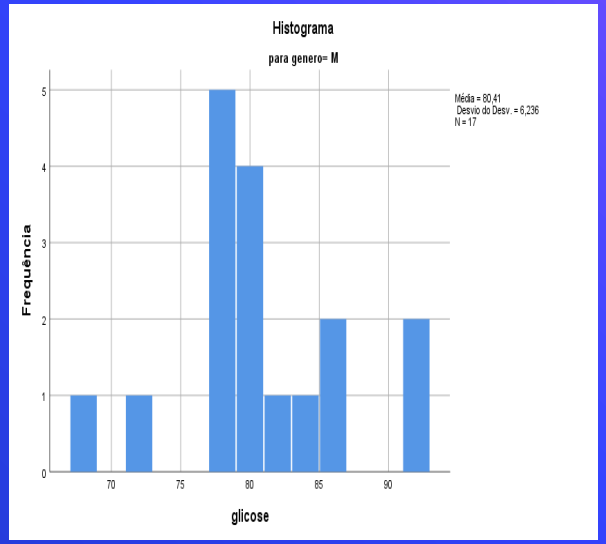


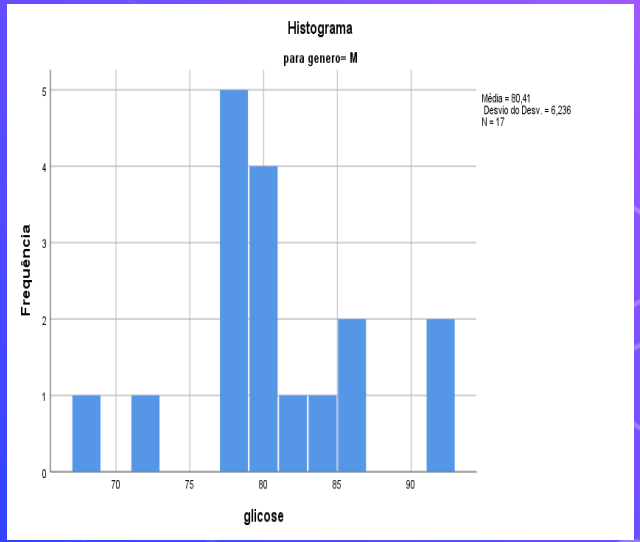


Descritivos

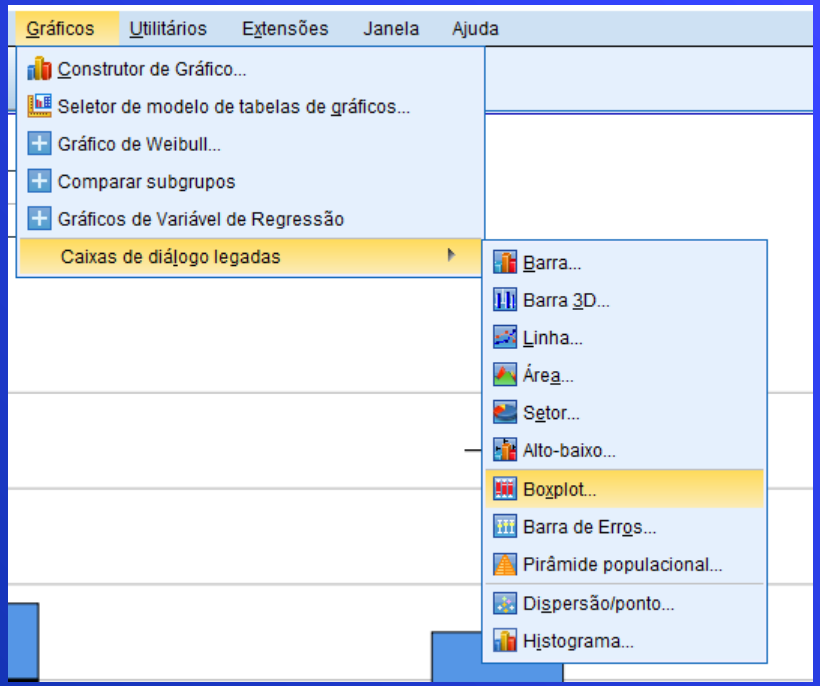
genero				Estatística	Desvio Padrão
glicose	M	Média		80,41	1,512
		95% de Intervalo de	Limite inferior	77,21	
			Limite superior	83,62	
		5% da média aparada		80,46	
		Mediana		80,00	
		Variância		38,882	
		Erro Padrão		6,236	
		Mínimo		68	
		Máximo		92	
		Amplitude		24	
		Amplitude interquartil		8	
		Assimetria		0,242	0,550
		Curtose		0,409	1,063
	F	Média		79,79	1,450
		95% de Intervalo de	Limite inferior	76,74	
			Limite superior	82,84	
		5% da média aparada		79,77	
		Mediana		78,00	
		Variância		39,953	
		Erro Padrão		6,321	
		Mínimo		68	
		Máximo		92	
		Amplitude		24	
		Amplitude interquartil		8	
		Assimetria		0,485	0,524
		Curtose		0,063	1,014

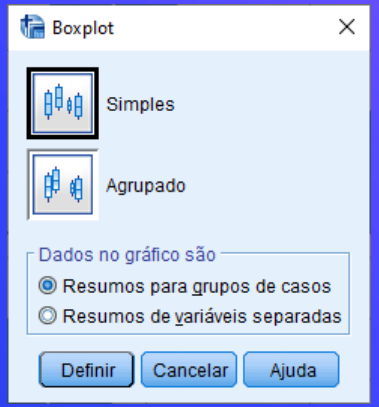


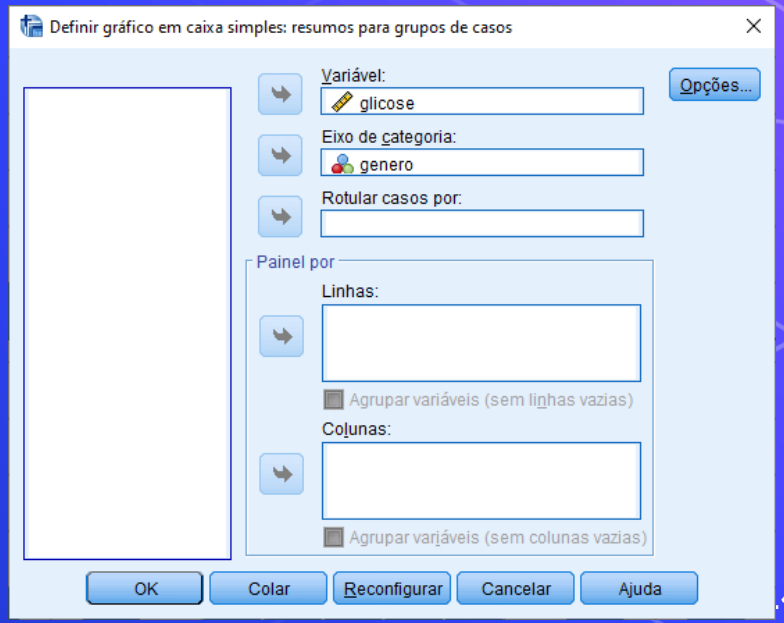




Vamos fazer um gráfico caixa de bigodes (boxplot) da distribuição da variável glicose estratificado pelo gênero.

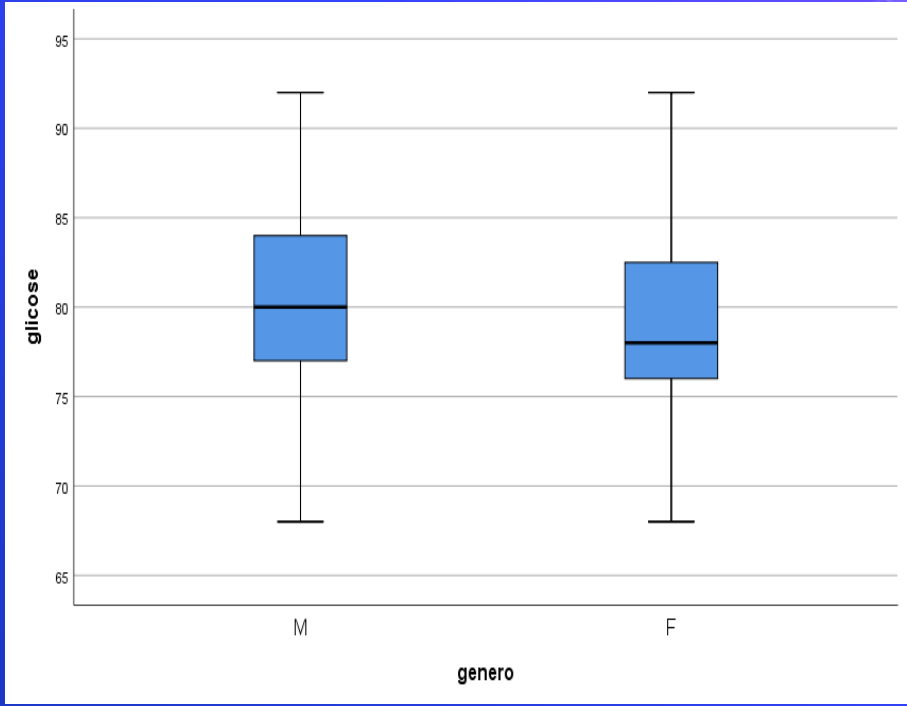

Gráficos
Utilitários
Extensões
Janela
Ajuda
Construtor de Gráfico...
Seletor de modelo de tabelas de gráficos...
Gráfico de Weibull...
Comparar subgrupos
Gráficos de Variável de Regressão
Caixas de diálogo legadas
Barra...
Barra 3D...
Linha...
Área...
Setor...
Alto-baixo...
Boxplot...
Barra de Erros...
Pirâmide populacional...
Dispersão/ponto...
Histograma...


Boxplot
Simples
Agrupado
Dados no gráfico são
Resumos para grupos de casos
Resumos de variáveis separadas
Definir
Cancelar
Ajuda


Definir gráfico em caixa simples: resumos para grupos de casos
Variável:
glicose
Opções...
Eixo de categoria:
genero
Rotular casos por:
Painel por
Linhas:
Agrupar variáveis (sem linhas vazias)
Colunas:
Agrupar variáveis (sem colunas vazias)
OK
Colar
Reconfigurar
Cancelar
Ajuda

Interpretação dos resultados

Dentro de cada uma das caixas espera-se encontrar 50% das observações



Testes de Hipóteses para uma população

por exemplo ...

Valor Médio-μ

$$\begin{cases} H_o\colon \mu = \mu_0 \text{ vs } H_1\colon \mu \neq \mu_o \\ H_o\colon \mu \geq \mu_0 \text{ vs } H_1\colon \mu < \mu_o \\ H_o\colon \mu \leq \mu_0 \text{ vs } H_1\colon \mu > \mu_o \end{cases}$$

Proporção-π

$$\begin{cases} H_o\colon \pi = \pi_0 \text{ vs } H_1\colon \pi \neq \pi_o \\ H_o\colon \pi \geq \pi_0 \text{ vs } H_1\colon \pi < \pi_o \\ H_o\colon \pi \leq \pi_0 \text{ vs } H_1\colon \pi > \pi_o \end{cases}$$

Teste ao Valor Médio-μ

EXEMPLO: Considere um estudo que procura estimar a frequência cardíaca em repouso (número de batimentos por minuto-bpm). Para tal, as frequências cardíacas de 20 indivíduos representativos desta população foram medidas, obtendo-se os valores presentes na tabela seguinte:

Frequências cardíacas de uma amostra de 20 indivíduos da população em estudo

67 67 68 68 68 69 69 69 69 69 70 70 70 70 71 71 72 72 73 74

Estudos anteriores estimaram que a frequência cardíaca média em repouso nos indivíduos desta população era igual a 73 bpm. Há evidência de que esta média (populacional) continue sendo a mesma, com base na presente amostra?

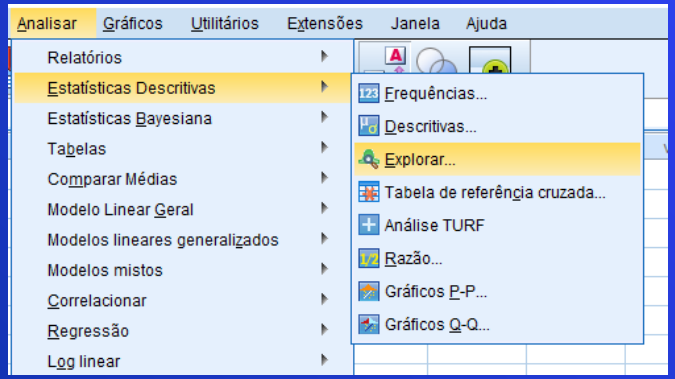
Nota: dados disponíveis no ficheiro **bpm.xlsx**

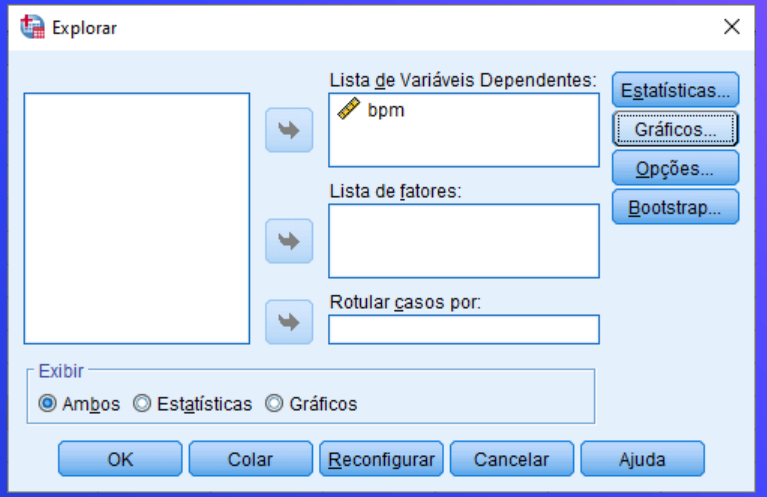
TESTE BILATERAL

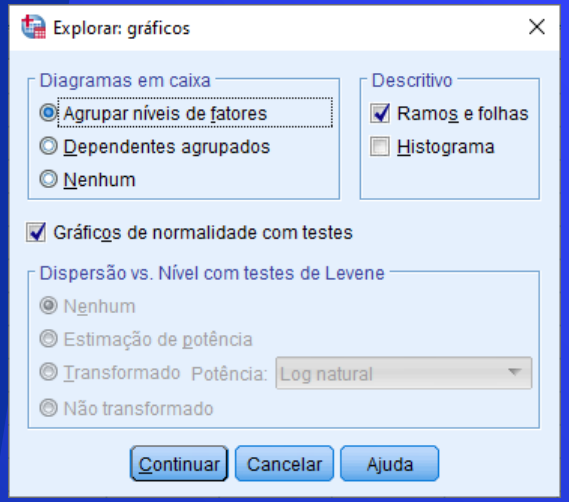
Objetivo: $\{H_o: \mu = 73$ vs $H_1: \mu \neq 73$

ETAPAS:

1º) Verificar a normalidade da distribuição (Se sim, Teste Paramérico).







Os testes de normalidade disponíveis no IBM SPSS são:

i) Teste de **Kolmogorov Smirnov** (dimensões amostrais maiores que 50 observações);
ii) Teste de **Shapiro Wilk** (dimensões amostrais menores que 50 observações);

As hipóteses subjacentes aos testes de normalidade são:

Ho: a distribuição do nº de bpm na população donde foi retirada a amostra é normalmente distribuída;
H1: a distribuição do nº de bpm na população donde foi retirada a amostra **não é** normalmente distribuída;

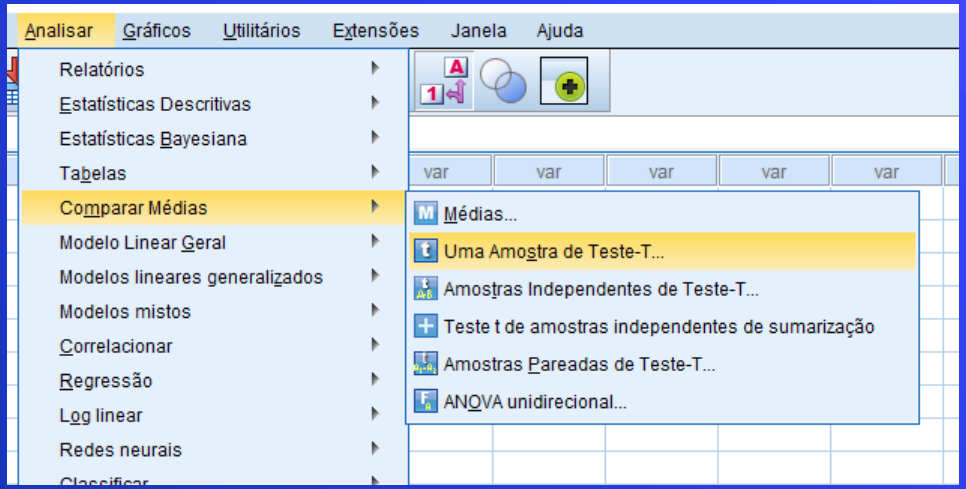
Com base no valor-p=0,346 não rejeitamos Ho, ou seja a distribuição dos bpm é normalmente distribuída

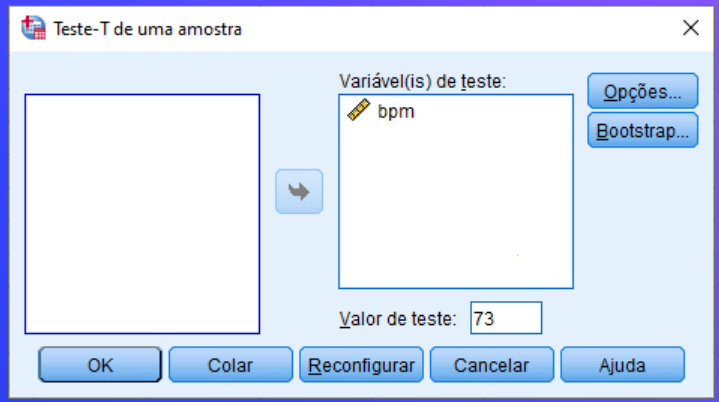
Testes de Normalidade

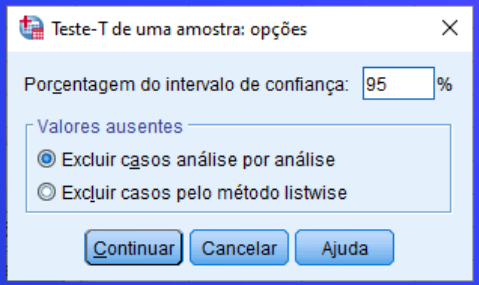
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	gl	Sig.	Estatística	gl	Sig.
bpm	0,162	20	0,175	0,949	20	0,346

a.

2º) Caso tenha sido verificada a normalidade da distribuição dos bpm, aplicar-se-á o teste t-student.







Teste de uma amostra

	Valor de Teste = 73					
	t	df	Sig. (2 extremidades)	Diferença média	95% Intervalo de Confiança da Diferença	
					Inferior	Superior
bpm	-7,499	19	0,000	-3,200	-4,09	-2,31

Como o valor-p≤α, rejeita-se Ho, ou seja o valor médio é estatisticamente diferente de 73 bpm.

Estatística Descritiva

	N	Mínimo	Máximo	Média	Erro Desvio
bpm	20	67	74	69,80	1,908
N válido (de lista)	20				

Poder-se-ia optar por um teste unilateral à esquerda.

Nesse caso, no IBM SPSS deverá ter-se em consideração o

$$\frac{valor-p}{2}$$

Teste à Proporção π

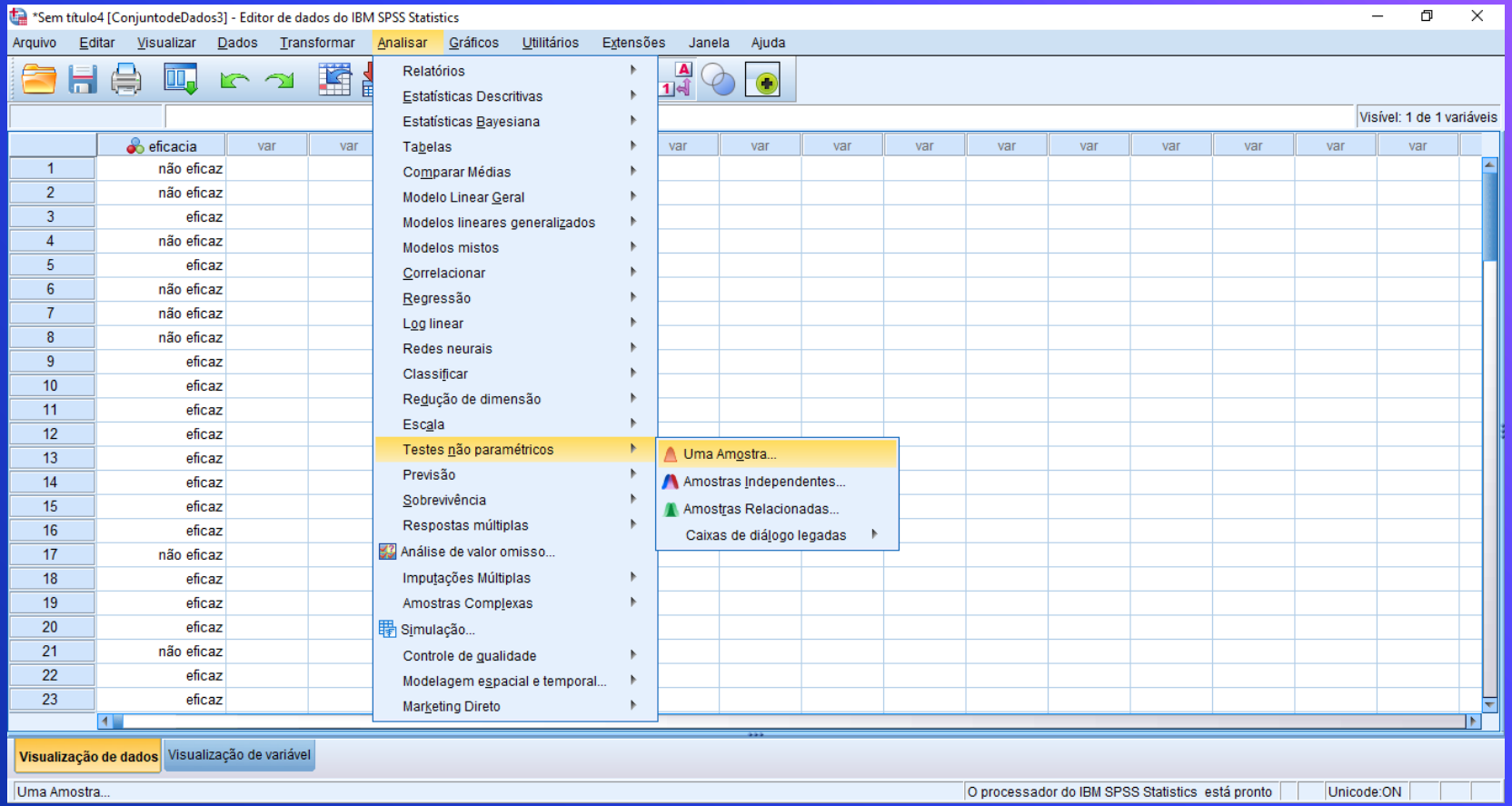
Sabe-se que a eficácia de uma vacina, após um ano da data de vacinação é de 40% (i.e., o efeito imunológico se prolonga por mais de um ano em 40% das pessoas vacinadas). Desenvolve-se uma nova vacina, mais cara, que foi administrada a cem voluntários. Os resultados experimentais apontaram para uma eficácia de 55%. Deseja-se saber se esta vacina é de fato melhor.

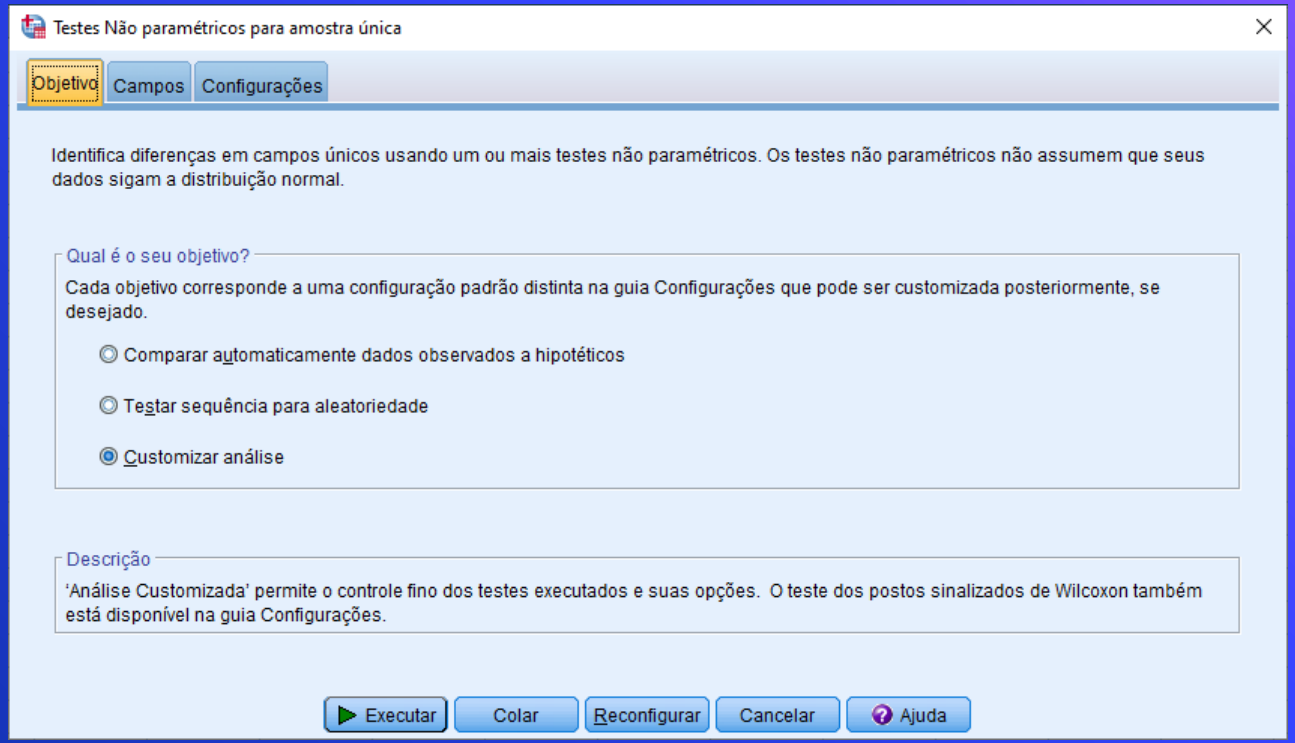
Nota: o ficheiro **vacina.xlsx** contém toda a informação amostral.

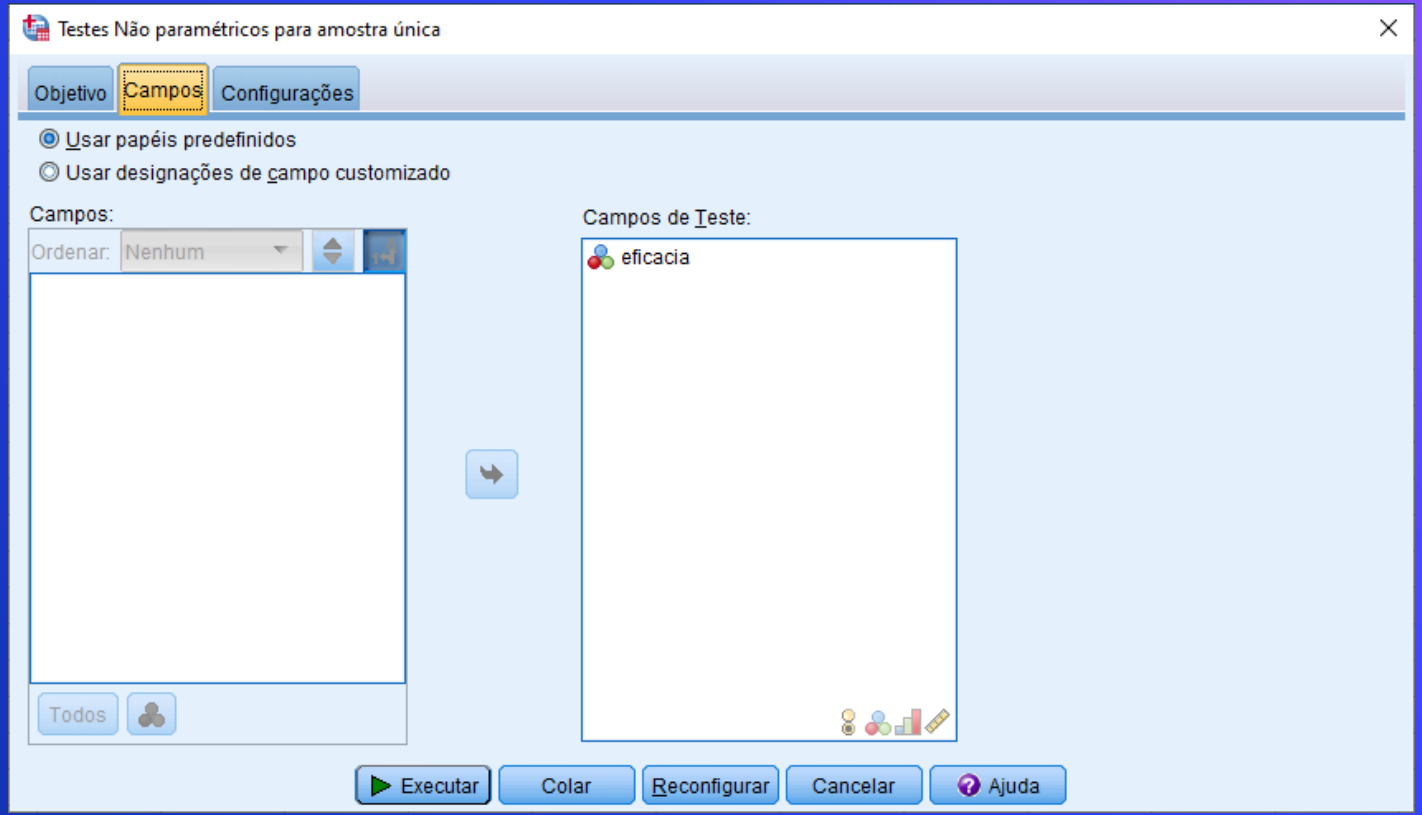
Codificação: 1-vacina eficaz; 0 – vacina não eficaz.

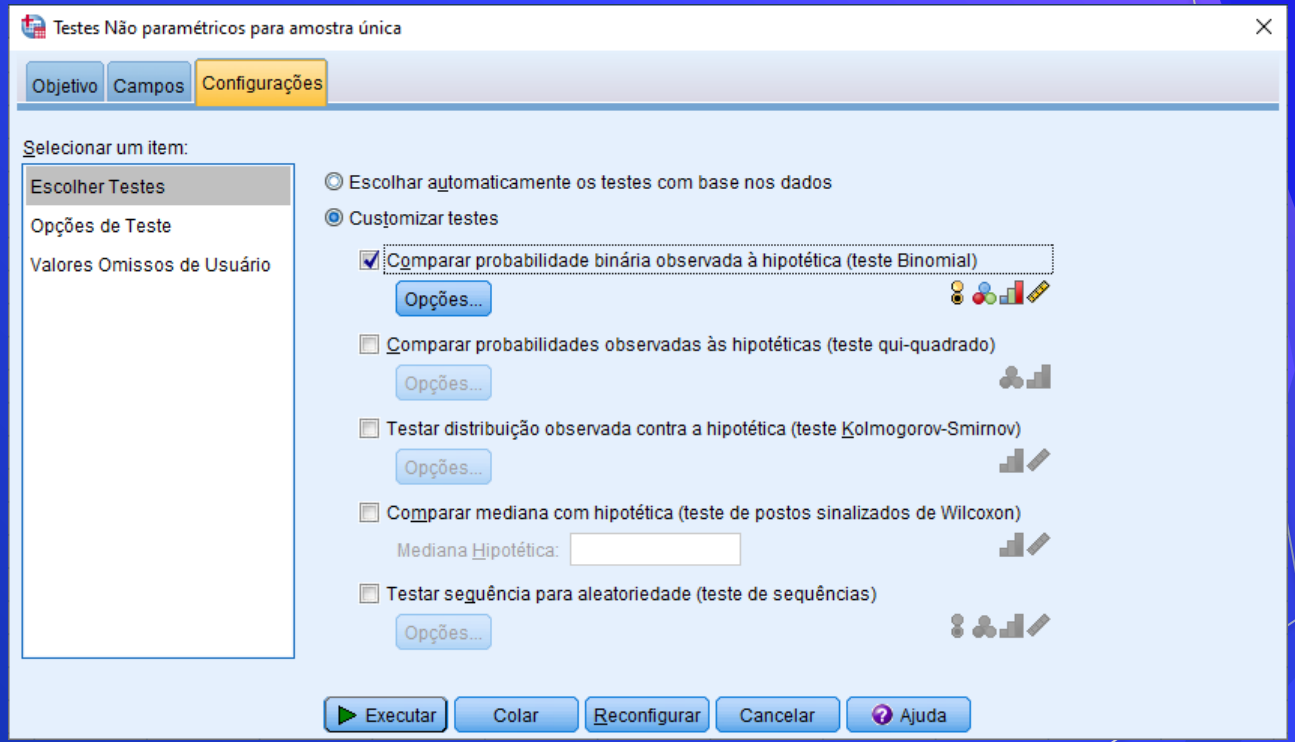
Formulação de Hipóteses

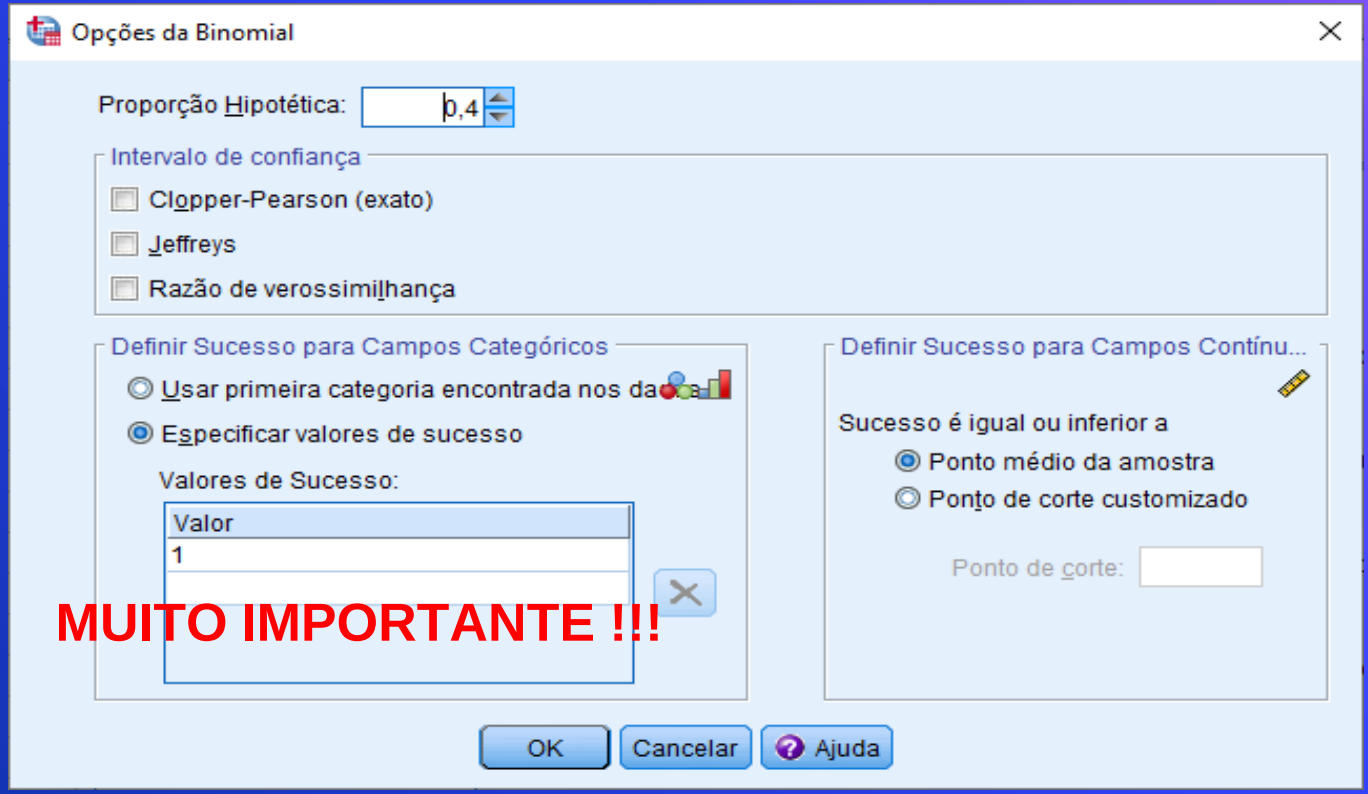
$$H_o: \pi \leq 0.4 \text{ vs } H_1: \pi > 0.4$$


*Sem título4 [ConjuntodeDados3] - Editor de dados do IBM SPSS Statistics
Arquivo Editar Visualizar Dados Transformar Analisar Gráficos Utilitários Extensões Janela Ajuda
Visível: 1 de 1 variáveis
eficacia
Relatórios
Estatísticas Descritivas
Estatísticas Bayesiana
Tabelas
Comparar Médias
Modelo Linear Geral
Modelos lineares generalizados
Modelos mistos
Correlacionar
Regressão
Log linear
Redes neurais
Classificar
Redução de dimensão
Escala
Testes não paramétricos
Previsão
Sobrevivência
Respostas múltiplas
Análise de valor omisso...
Imputações Múltiplas
Amostras Complexas
Simulação...
Controle de qualidade
Modelagem espacial e temporal...
Marketing Direto
Uma Amostra...
Amostras Independentes...
Amostras Relacionadas...
Caixas de diálogo legadas
Visualização de dados
Visualização de variável
Uma Amostra...
O processador do IBM SPSS Statistics está pronto
Unicode:ON


Testes Não paramétricos para amostra única
Objetivo
Campos
Configurações
Identifica diferenças em campos únicos usando um ou mais testes não paramétricos. Os testes não paramétricos não assumem que seus dados sigam a distribuição normal.
Qual é o seu objetivo?
Cada objetivo corresponde a uma configuração padrão distinta na guia Configurações que pode ser customizada posteriormente, se desejado.
Comparar automaticamente dados observados a hipotéticos
Testar sequência para aleatoriedade
Customizar análise
Descrição
'Análise Customizada' permite o controle fino dos testes executados e suas opções. O teste dos postos sinalizados de Wilcoxon também está disponível na guia Configurações.
Executar
Colar
Reconfigurar
Cancelar
Ajuda


Testes Não paramétricos para amostra única
Objetivo
Campos
Configurações
Usar papéis predefinidos
Usar designações de campo customizado
Campos:
Ordenar:
Nenhum
Todos
Campos de Teste:
eficacia
Executar
Colar
Reconfigurar
Cancelar
Ajuda


Testes Não paramétricos para amostra única
Objetivo
Campos
Configurações
Selecionar um item:
Escolher Testes
Opções de Teste
Valores Omissos de Usuário
Escolhar automaticamente os testes com base nos dados
Customizar testes
Comparar probabilidade binária observada à hipotética (teste Binomial)
Opções...
Comparar probabilidades observadas às hipotéticas (teste qui-quadrado)
Opções...
Testar distribuição observada contra a hipotética (teste Kolmogorov-Smirnov)
Opções...
Comparar mediana com hipotética (teste de postos sinalizados de Wilcoxon)
Mediana Hipotética:
Testar seguência para aleatoriedade (teste de sequências)
Opções...
Executar
Colar
Reconfigurar
Cancelar
Ajuda


Opções da Binomial
Proporção Hipotética:
0,4
Intervalo de confiança
Clopper-Pearson (exato)
Jeffreys
Razão de verossimilhança
Definir Sucesso para Campos Categóricos
Usar primeira categoria encontrada nos da
Especificar valores de sucesso
Valores de Sucesso:
Valor
1
MUITO IMPORTANTE !!!
Definir Sucesso para Campos Contínu...
Sucesso é igual ou inferior a
Ponto médio da amostra
Ponto de corte customizado
Ponto de corte:
OK
Cancelar
Ajuda

Uma Amostra de Resumo de Teste Binomial

N total	100
Estatística do teste	55,000
Erro padrão	4,899
Estatística de Teste Padronizado	2,960
Sinal assintótico (teste de um lado)	0,002

Como o valor-p≤α, rejeita-se Ho, ou seja a proporção da eficácia da nova vacina é superior a 0.4 (40%); logo é melhor.

Análise Bivariada

Como duas variáveis estão relacionadas ?

- O conhecimento/**identificação** de relações entre variáveis é uma mais valia na compreensão do fenómeno em estudo.

- **Mas o que isto significa?**

  É dito que existe uma **relação** entre duas variáveis quando a distribuição dos valores de uma variável está associada com a distribuição dos valores da outra variável.

Nem todas as variáveis são da mesma natureza !!!!

EXEMPLOS

- a mortalidade infantil é superior em países com baixo rendimento *per capita;*
- a intensão de voto está relacionada com a classe social do eleitor;
- a obesidade está relacionada com a ingestão em maior quantidade de elementos calóricos.

Exemplo: Estudo sobre o absentismo e satisfação em 30 empregados do setor da saúde.

Objetivo: identificar uma relação/associação entre variáveis.

empregado	satisfação	absentismo
1	1	0
2	1	1
3	0	1
4	1	1
5	0	1
...	...	...
30	1	0

Código: 1 - "Sim"; 0 - "Não".

Identificar relações por vezes não é fácil.

LIMITAÇÕES

- **↑ número de observações;**
- **≠ tipos de variáveis;**
- **Informação escassa.**

Sugestão:

utilização de **tabelas de contingência** !!!

- Uma tabela de contingência é um tipo de tabela em formato matricial que apresenta a distribuição de frequências das variáveis;

- Fornecem um retrato inicial entre as variáveis (em escala nominal ou ordinal) permitindo o auxílio na identificação de interações entre elas.

Associação NÃO PERFEITA

Satisfação *vs* Absentismo (Total coluna %)

Absentismo	Satisfação Sim	Não	Total
Sim	4 (29%)	11 (69%)	15
Não	10 (71%)	5 (31%)	15
Total	14	16	30

↑ *Satisfação* ↓ *Absentismo*
↓ *Satisfação* ↑ *Absentismo*

Existe uma associação mas NÃO É PERFEITA. [a]

[a] χ^2_p= 4.8214, df = 1, p-value =0.02811

Associação PERFEITA

Satisfação *vs* Absentismo (Total coluna %)

Absentismo	Satisfação Sim	Não	Total
Sim	0 (0%)	16 (100%)	16
Não	14 (100%)	0 (0%)	14
Total	14	16	30

Há uma PERFEITA associação. [a]

[a] χ^2_P= 30, df = 1, p-value =4.32e−08

AUSÊNCIA de relação

Satisfação *vs* Absentismo (Total coluna %)

Absentismo	Satisfação Sim	Não	Total
Sim	7 (50%)	8 (50%)	15
Não	7 (50%)	8 (50%)	15
Total	14	16	30

Não há associação entre as variáveis. [a]

[a] $\chi^2_p = 0$, df = 1, p-value =1

O que caracterizou esta tabela ?

Como identificar a existência/ausência de associações/relações em tabelas de contingência ???

Alguns testes: Teste Qui-Quadrado; Teste Exacto de Fisher e sua extensão; Teste de McNemar

Teste Qui-Quadrado (não paramétrico)

Objetivo

Avaliar a associação/relação entre 2 variáveis que se encontram em escala nominal e/ou ordinal dispostas numa tabela de contingência de qualquer dimensão (quadrada ou rectangular).

Condições de aplicabilidade (critérios de Cochran)

- 80% das células da tabela devem ter frequências esperadas > 5 e todas as células devem ter frequências esperadas > 1
- As unidades experimentais não são emparelhadas, ou seja são independentes.

Frequências esperadas → Frequências que esperaríamos ter, caso as variáveis não estivessem relacionadas (independentes), ou seja se Ho for verdadeira.

Caso as condições de aplicabilidade do teste qui-quadrado forem violadas poder-se-à agrupar classes e voltar a verificar os critérios de Cochran !

Hipóteses de investigação do teste qui-quadrado

Ho: Não há relação/associação entre a variável X e a variável Y

H1: Há relação/associação entre a variável X e a variável Y

Se a amostra for pequena:

- Junção de algumas das categorias de uma ou das duas variáveis
- Teste exacto de Fisher
- Correcção para a continuidade (correcção de Yates)

Consiste na redução da diferença entre as frequências observadas e as esperadas (à medida que a dimensão da amostra aumenta, a correção diminui).

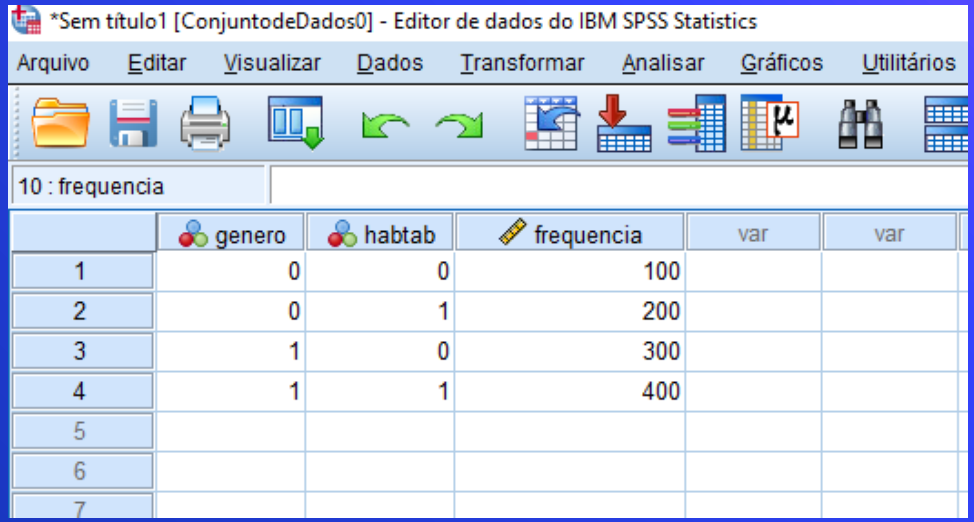
Exemplo:

Um investigador suspeita que existe uma relação/associação entre o género e os hábitos tabágicos. Com o intuito de verificar se tal se verifica, selecionou aleatoriamente 1000 pessoas da mesma faixa etária obtendo os seguintes resultados:

Sexo|Hab.Tabágicos	Fuma	Não fuma
Masculino	200	100
Feminino	400	300

Existirá alguma relação entre as variáveis em análise?
O que poderá concluir? Assuma α=5%.

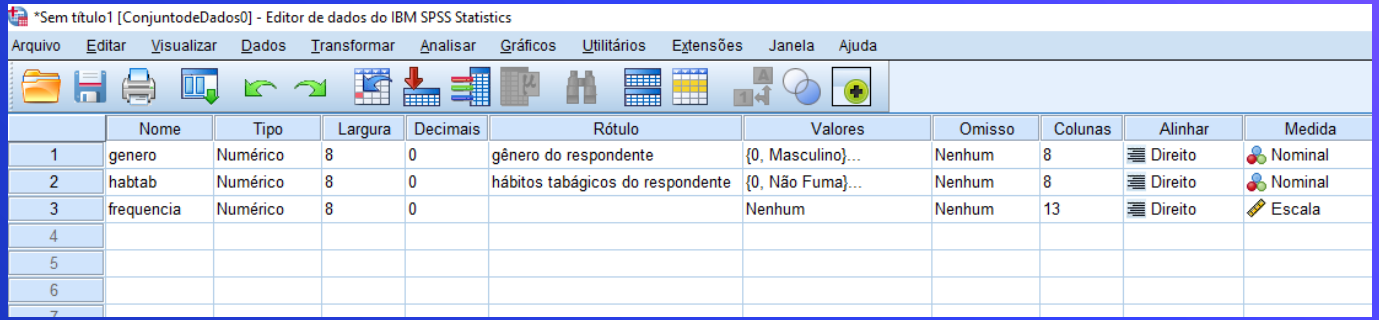
No IBM SPSS teremos de preparar a informação oriunda de uma tabela de frequências



	genero	habtab	frequencia	var	var
1	0	0	100		
2	0	1	200		
3	1	0	300		
4	1	1	400		
5					
6					
7					

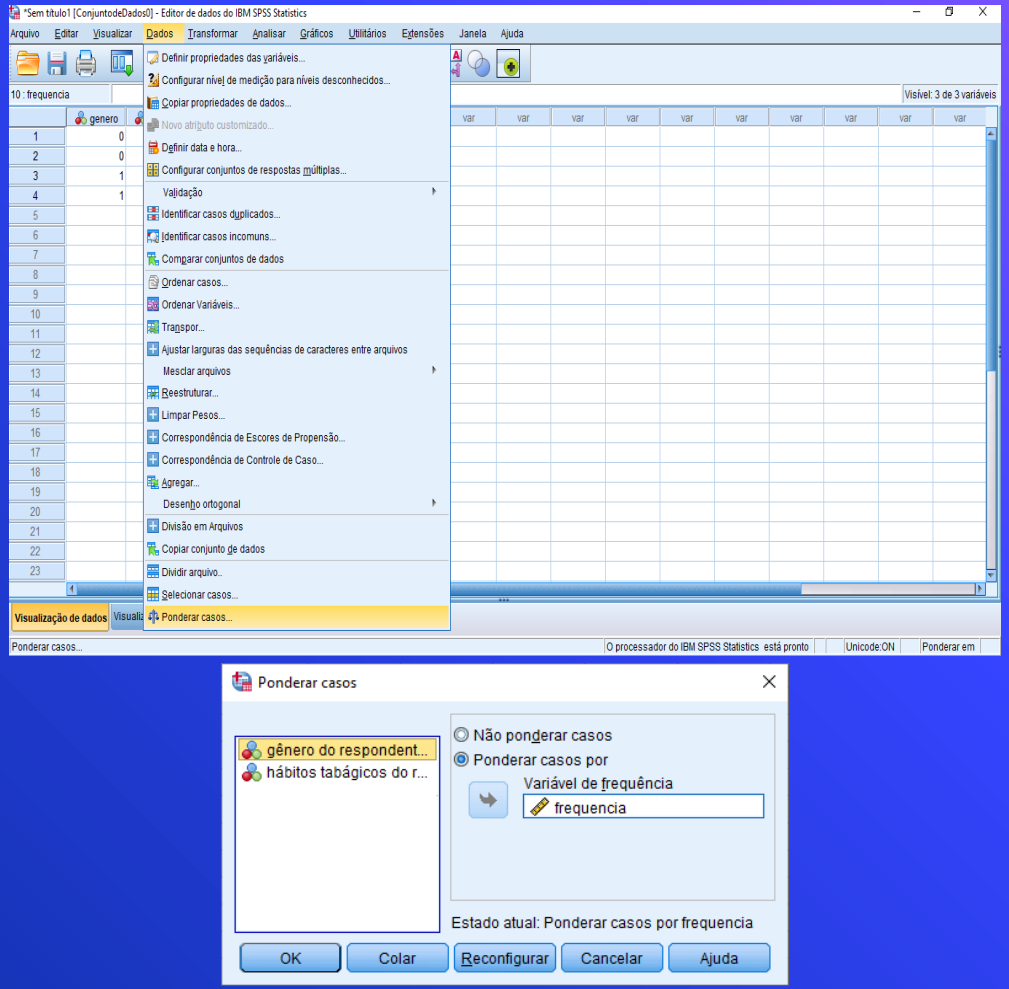
Codificação: Gênero 0-Masculino, 1-Feminino;
Hab. Tabágicos: 0-Não Fuma, 1-Fuma.

Definir os valores no IBM SPSS e especificar os rótulos



	Nome	Tipo	Largura	Decimais	Rótulo	Valores	Omisso	Colunas	Alinhar	Medida
1	genero	Numérico	8	0	gênero do respondente	{0, Masculino}...	Nenhum	8	Direito	Nominal
2	habtab	Numérico	8	0	hábitos tabágicos do respondente	{0, Não Fuma}...	Nenhum	8	Direito	Nominal
3	frequencia	Numérico	8	0		Nenhum	Nenhum	13	Direito	Escala
4										
5										
6										

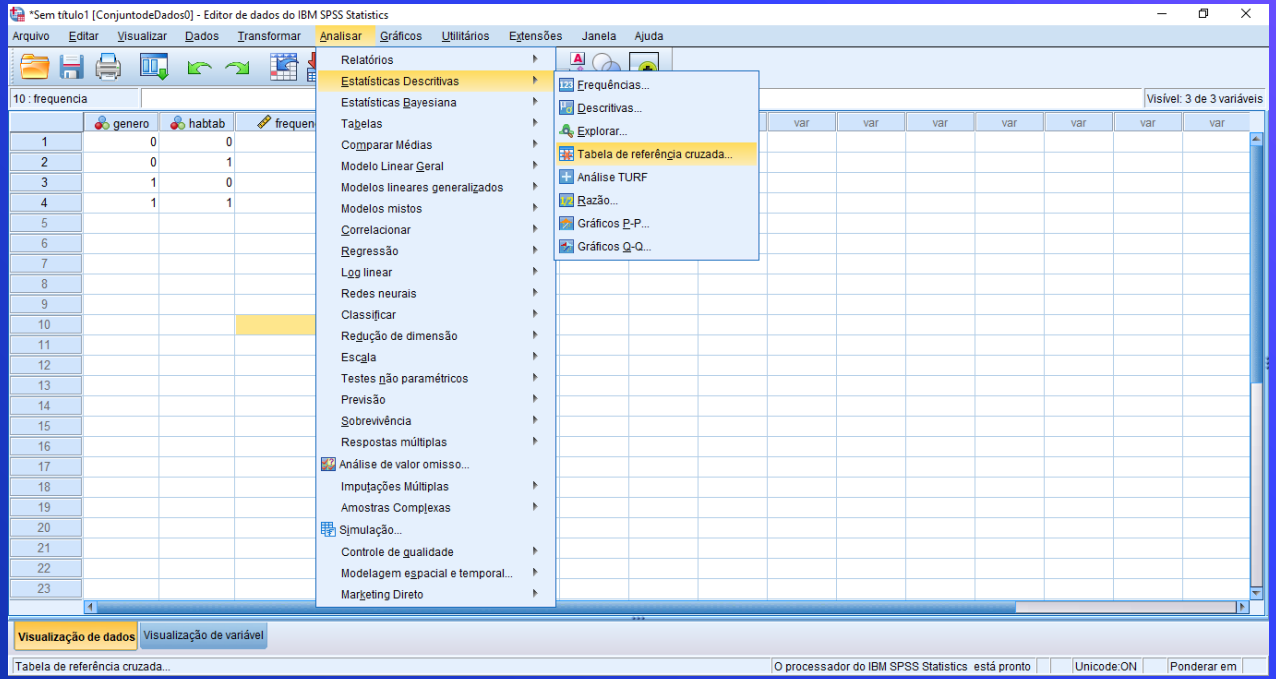
Será necessário ponderar os casos (informação) por frequência !!!

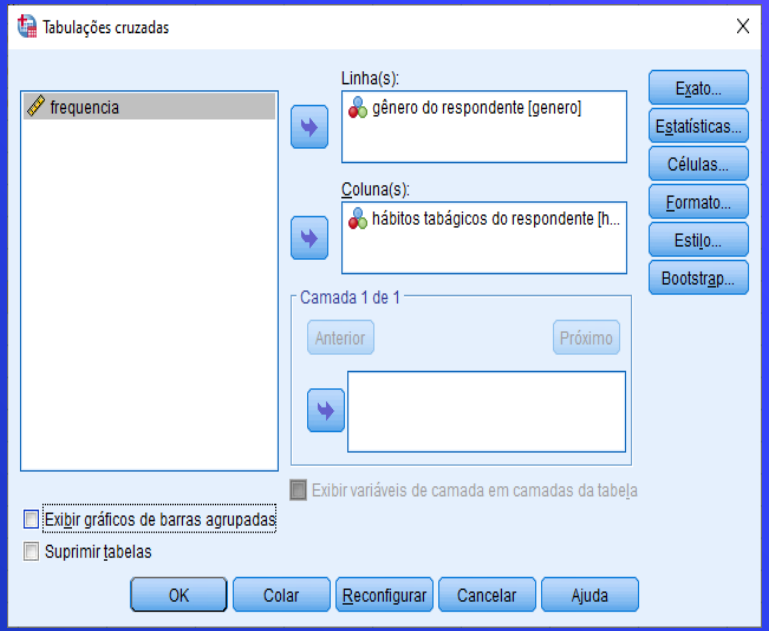

*Sem título1 [ConjuntodeDados0] - Editor de dados do IBM SPSS Statistics
Arquivo Editar Visualizar Dados Transformar Analisar Gráficos Utilitários Extensões Janela Ajuda
Definir propriedades das variáveis...
Configurar nível de medição para níveis desconhecidos...
Copiar propriedades de dados...
Novo atributo customizado...
Definir data e hora...
Configurar conjuntos de respostas múltiplas...
Validação
Identificar casos duplicados...
Identificar casos incomuns...
Comparar conjuntos de dados
Ordenar casos...
Ordenar Variáveis...
Transpor...
Ajustar larguras das sequências de caracteres entre arquivos
Mesclar arquivos
Reestruturar...
Limpar Pesos...
Correspondência de Escores de Propensão...
Correspondência de Controle de Caso...
Agregar...
Desenho ortogonal
Divisão em Arquivos
Copiar conjunto de dados
Dividir arquivo...
Selecionar casos...
Ponderar casos...
10 : frequencia
genero
Visível: 3 de 3 variáveis
Visualização de dados
Ponderar casos...
O processador do IBM SPSS Statistics está pronto
Unicode:ON
Ponderar em

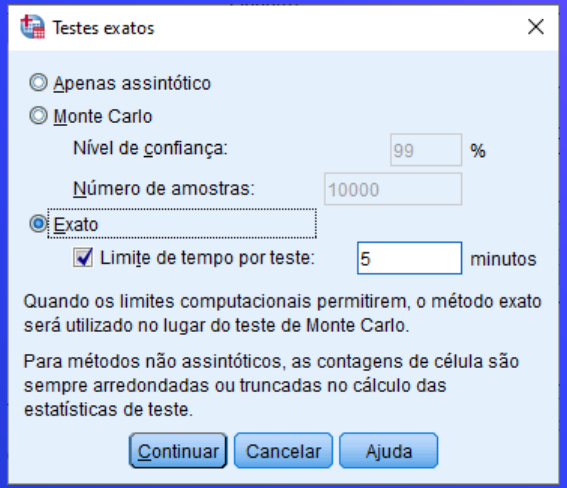
Ponderar casos
gênero do respondent...
hábitos tabágicos do r...
Não ponderar casos
Ponderar casos por
Variável de frequência
frequencia
Estado atual: Ponderar casos por frequencia
OK
Colar
Reconfigurar
Cancelar
Ajuda

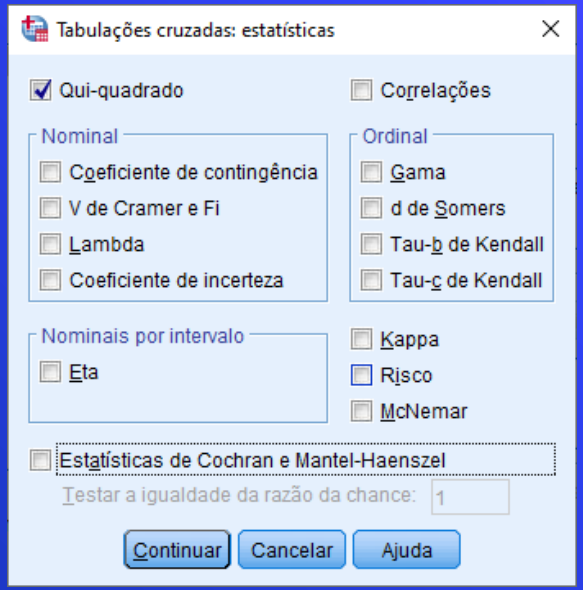
Definição de hipóteses

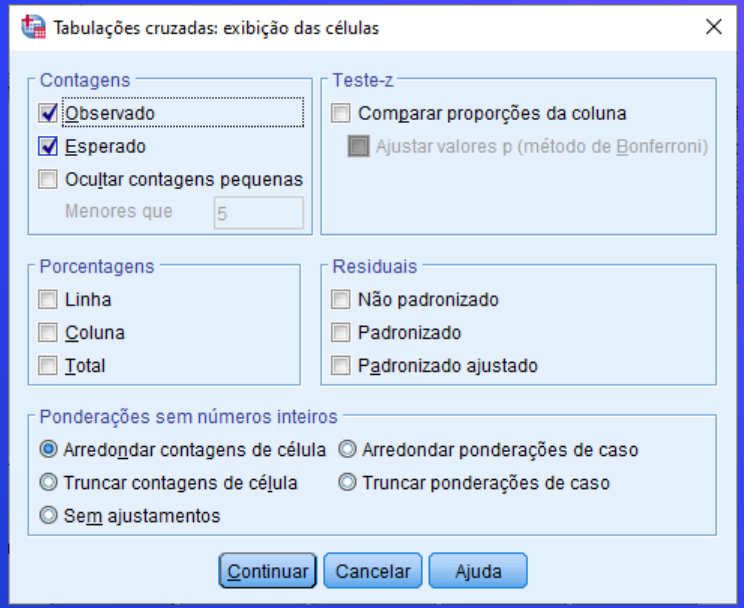
Ho: Não relação/associação entre o gênero e os háb. tabágicos;
H1: Há relação/associação entre o gênero e os hábitos tabágicos.




Tabulações cruzadas
frequencia
Linha(s):
gênero do respondente [genero]
Coluna(s):
hábitos tabágicos do respondente [h...
Camada 1 de 1
Anterior
Próximo
Exibir variáveis de camada em camadas da tabela
Exibir gráficos de barras agrupadas
Suprimir tabelas
Exato...
Estatísticas...
Células...
Formato...
Estilo...
Bootstrap...
OK
Colar
Reconfigurar
Cancelar
Ajuda


Testes exatos
Apenas assintótico
Monte Carlo
Nível de confiança:
99
%
Número de amostras:
10000
Exato
Limite de tempo por teste:
5
minutos
Quando os limites computacionais permitirem, o método exato será utilizado no lugar do teste de Monte Carlo.
Para métodos não assintóticos, as contagens de célula são sempre arredondadas ou truncadas no cálculo das estatísticas de teste.
Continuar
Cancelar
Ajuda


Tabulações cruzadas: estatísticas
Qui-quadrado
Correlações
Nominal
Coeficiente de contingência
V de Cramer e Fi
Lambda
Coeficiente de incerteza
Ordinal
Gama
d de Somers
Tau-b de Kendall
Tau-c de Kendall
Nominais por intervalo
Eta
Kappa
Risco
McNemar
Estatísticas de Cochran e Mantel-Haenszel
Testar a igualdade da razão da chance:
1
Continuar
Cancelar
Ajuda


Tabulações cruzadas: exibição das células
Contagens
Observado
Esperado
Ocultar contagens pequenas
Menores que
5
Teste-z
Comparar proporções da coluna
Ajustar valores p (método de Bonferroni)
Porcentagens
Linha
Coluna
Total
Residuais
Não padronizado
Padronizado
Padronizado ajustado
Ponderações sem números inteiros
Arredondar contagens de célula
Arredondar ponderações de caso
Truncar contagens de célula
Truncar ponderações de caso
Sem ajustamentos
Continuar
Cancelar
Ajuda

Verificação das Condições de aplicabilidade –critérios de Cochran.

Tabulação cruzada gênero do respondente * hábitos tabágicos do respondente

					Total
gênero do respondente					300
		Contagem Esperada	120,0	180,0	300,0
	Feminino	Contagem	300	400	700
		Contagem Esperada	280,0	420,0	700,0
Total		Contagem	400	600	1000
		Contagem Esperada	400,0	600,0	1000,0

(i) Todas as frequências esperadas são superiores a um;
(ii) 80% das frequências esperadas são superiores a 5.

Testes qui-quadrado

	Valor	gl	Significância Assintótica (Bilateral)	Sig exata (2 lados)	Sig exata (1 lado)	Probabilidade de ponto
Qui-quadrado de Pearson	7,937[a]	1	0,005	0,006	0,003	
Correção de continuidade[b]	7,545	1	0,006			
Razão de verossimilhança	8,043	1	0,005	0,005	0,003	
Teste Exato de Fisher				0,005	0,003	
Associação Linear por Linear	7,929[c]	1	0,005	0,006	0,003	0,001
N de Casos Válidos	1000					

a. 0 células (0,0%)

b. Computado apenas para

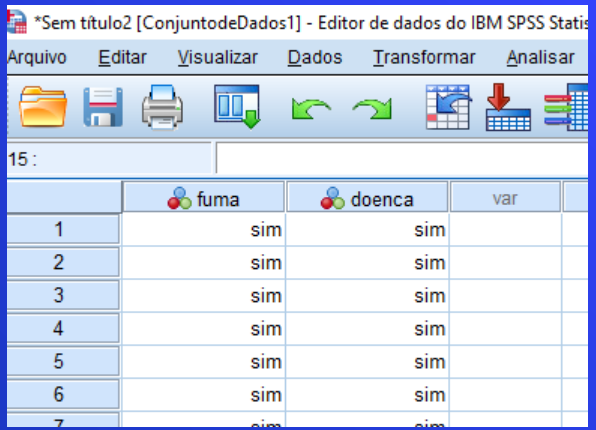
c. A estatística padronizada

Com base no valor-p, rejeita-se a Ho, ou seja: existe associação/relação estatisticamente significativa entre o gênero e os hábitos tabágicos **(neste caso forte evidência- valor p quase nulo).**

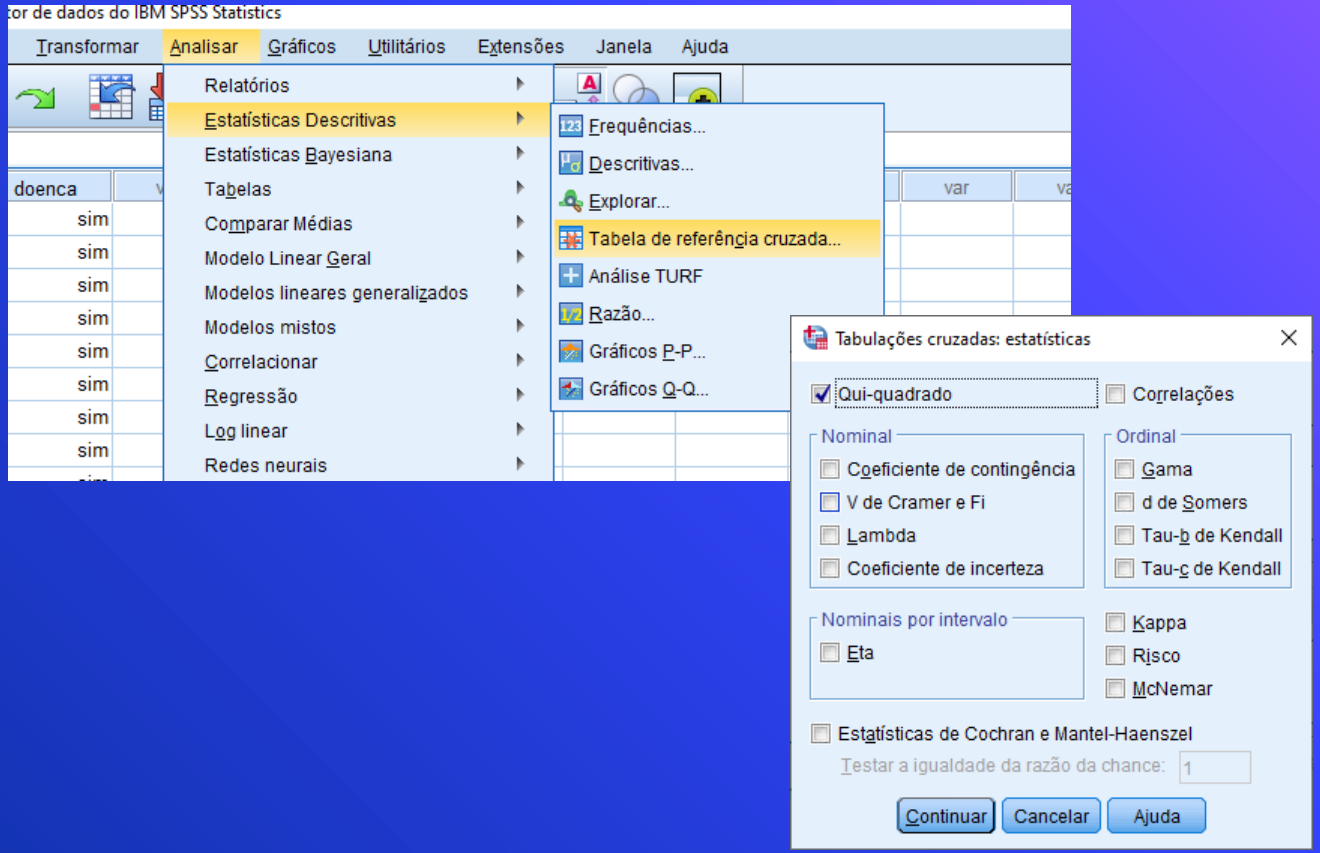
E se a informação não estiver na forma de uma tabela, como fazer?

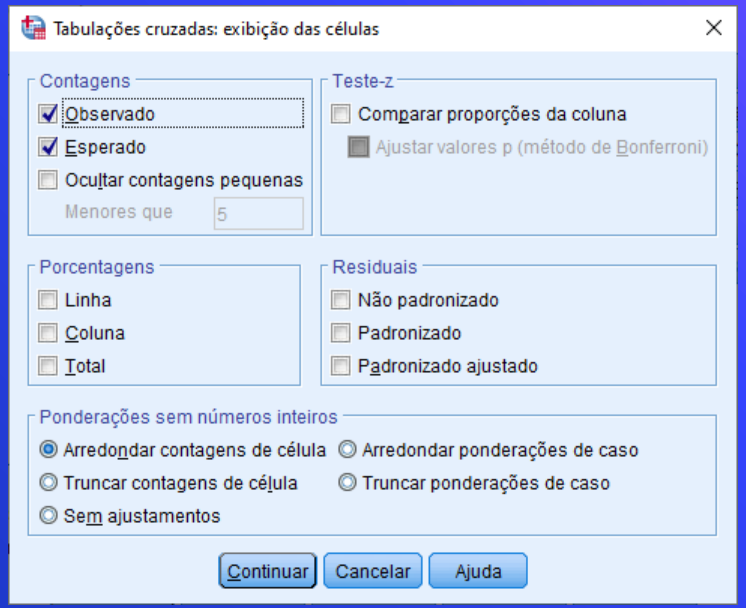
Exemplo: Foram questionados 85 indivíduos sobre hábitos tabágicos (1=fuma, 0=não fuma) e se sofrem ou não de uma determinada doença coronária (1=sim, 0=não). A informação encontra-se presente no ficheiro fumar_doencacoronaria.xls. Veja se existe alguma relação entre as variáveis em estudo. Assuma um α=5%.

Codificação: fuma: 0-não, 1-sim; doenca: 0-Não, 1-sim.



No IBM SPSS faríamos:




Tabulações cruzadas: exibição das células
Contagens
Observado
Esperado
Ocultar contagens pequenas
Menores que
5
Teste-z
Comparar proporções da coluna
Ajustar valores p (método de Bonferroni)
Porcentagens
Linha
Coluna
Total
Residuais
Não padronizado
Padronizado
Padronizado ajustado
Ponderações sem números inteiros
Arredondar contagens de célula
Arredondar ponderações de caso
Truncar contagens de célula
Truncar ponderações de caso
Sem ajustamentos
Continuar
Cancelar
Ajuda

Definição das hipóteses

$$\begin{cases} H_o: \text{Não há associação entre hábitos tabágicos e a existência de doença coronária} \\ \quad\quad (\text{independência entre as variáveis}) \\ \\ H_1: \text{Há associação entre hábitos tabágicos e a existência de doença coronária} \end{cases}$$

Verificação das Condições de aplicabilidade:critérios de Cochran.

Tabulação cruzada fuma * doenca

					Total
fuma					45
		Contagem Esperada	27,5	17,5	45,0
	sim	Contagem	15	25	40
		Contagem Esperada	24,5	15,5	40,0
Total		Contagem	52	33	85
		Contagem Esperada	52,0	33,0	85,0

OK!!!!

Testes qui-quadrado

	Valor	gl	Significância Assintótica (Bilateral)	Sig exata (2 lados)	Sig exata (1 lado)
Qui-quadrado de Pearson	17,833[a]	1	0,000		
Correção de continuidade[b]	16,000	1	0,000		
Razão de verossimilhança	18,506	1	0,000		
Teste Exato de Fisher				0,000	0,000
Associação Linear por Linear	17,623	1	0,000		
N de Casos Válidos	85				

a. 0 células (0,0%) esperavam uma contagem menor que 5. A contagem

b. Computado apenas para

Temos ainda coeficientes de associação (Cramer V; Phi, ...)

Com base no valor-p, rejeita-se a Ho, ou seja: existe associação/relação estatisticamente significativa entre os hábitos tabágicos e a existência de doença coronária (neste caso muito forte!)

Desafio

Exemplo: Um psicólogo tenciona verificar se existe alguma relação entre a dimensão da personalidade (extrovertido/introvertido) e o consumo de cerveja (regular/esporádico). Recrutou de forma aleatória 12 estudantes universitários que decidiram participar no estudo. Os resultados apresentam-se na tabela seguinte:

	Consumo Cerveja	
Personalidade	**Esporádico**	**Regular**
Introvertido	1	6
Extrovertido	4	1

Codificação:
Introvertido-0;Extrovertido-1
Esporádico-0;Regular-1

O que poderá concluir o psicólogo? Assuma α=10%.

Limitações das tabelas de contingência:

para dados qualitativos .....

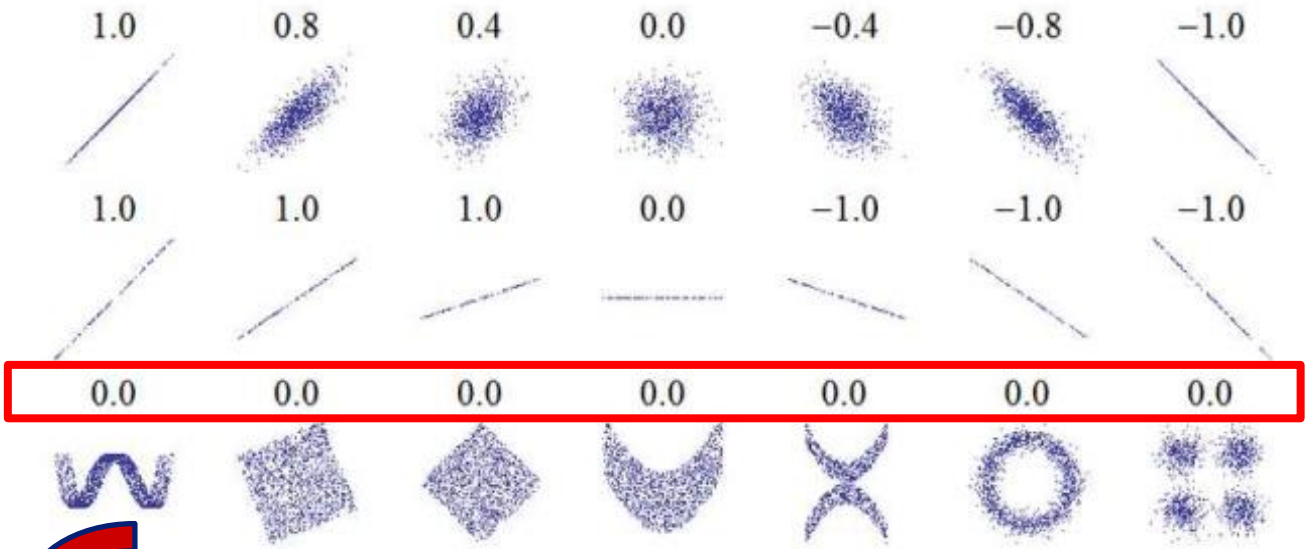
Surge então a CORRELAÇÃO para ultrapassar esta limitação!

1. A ideia de correlação é uma das mais importantes e básicas no estudo de relações bivariadas.
2. Ao contrário da abordagem anterior, a correlação indica a **"intensidade"** da relação das variáveis e a **direcção** da relação entre o par de variáveis.
3. Distinguem-se duas medidas de correlação: medidas de **correlação linear** utilizando **variáveis contínuas** e medidas de **correlação de postos** utilizando **variáveis ordinais**.
4. Apesar de partilharem algumas propriedades comuns, diferem em aspectos importantes.

Para avaliar a intensidade e a direcção da relação entre duas variáveis é determinado o coeficiente de correlação.

- Coeficiente de correlação de Pearson - r $\Rightarrow$ variáveis contínuas (O'Brien, 1979);
- Coeficiente de correlação de Spearman -ρ ou $r_s \Rightarrow$ variáveis ordinais. Nota: a par do coeficiente ρ também utiliza-se o coeficiente de correlação de Kendall -τ.

Dispersão *vs* ρ



As variáveis podem ter uma relação que não a linear !!!!

Figura retirada de https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Considere uma amostra aleatória de 16 estudantes do ensino secundário no qual foram registadas as alturas (em cm) e o peso (em kg), cujos dados estão presentes na tabela seguinte:

alturacm	pesokg
152,40	43,86
160,02	45,64
160,02	46,65
162,56	48,13
162,56	49,22
165,10	57,46
167,64	60,11
167,64	60,38
170,18	61,35
170,18	67,39
172,72	68,24
175,26	71,28
177,80	72,67
182,88	77,39
185,42	80,00
187,96	81,91

A informação encontra-se no ficheiro alturapeso.xls

Importe a informação para o IBM SPSS

Ir spersão

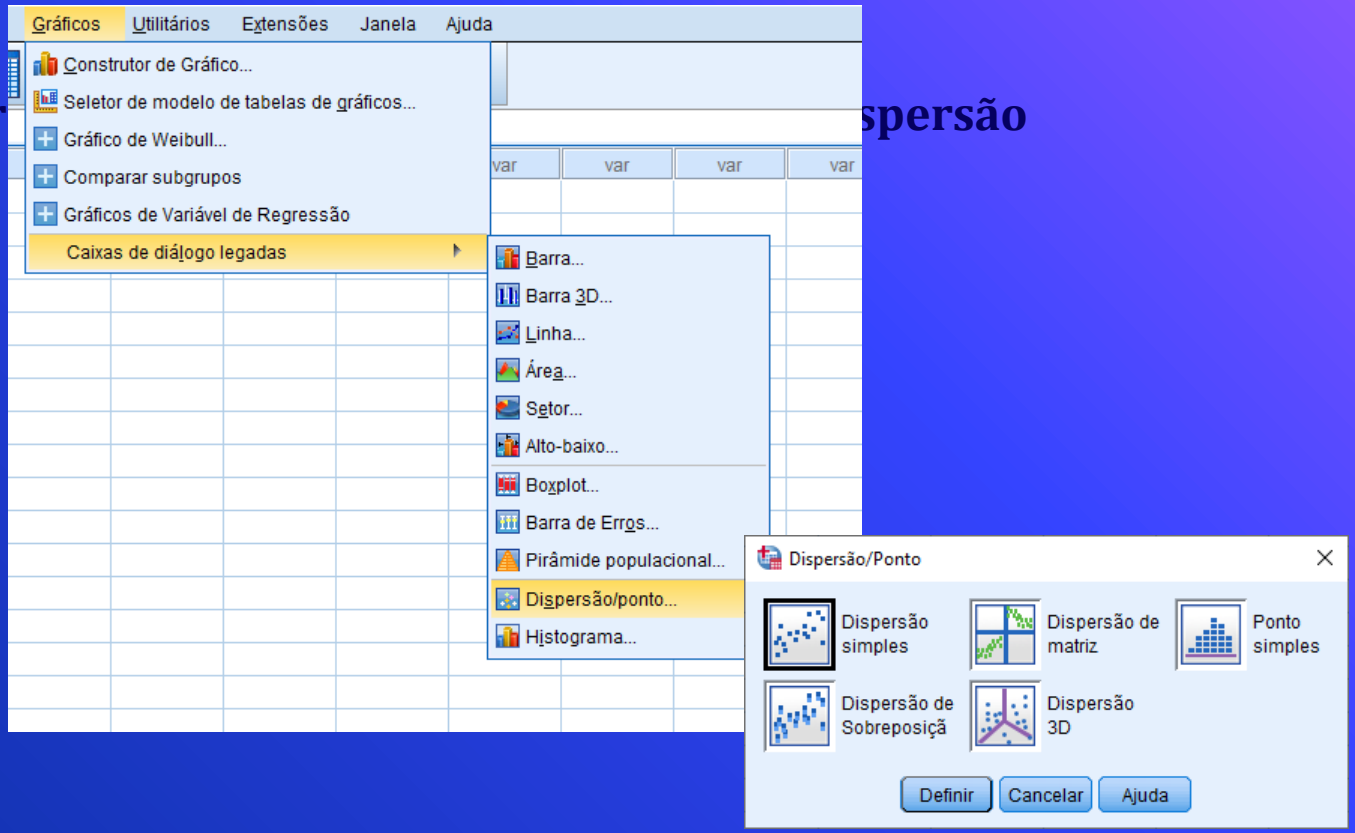

Gráficos
Utilitários
Extensões
Janela
Ajuda
Construtor de Gráfico...
Seletor de modelo de tabelas de gráficos...
Gráfico de Weibull...
Comparar subgrupos
Gráficos de Variável de Regressão
Caixas de diálogo legadas
Barra...
Barra 3D...
Linha...
Área...
Setor...
Alto-baixo...
Boxplot...
Barra de Erros...
Pirâmide populacional...
Dispersão/ponto...
Histograma...
var
var
var
var
Dispersão/Ponto
Dispersão simples
Dispersão de matriz
Ponto simples
Dispersão de Sobreposiçã
Dispersão 3D
Definir
Cancelar
Ajuda


Gráfico de dispersão simples
alturacm
pesokg
Eixo Y:
Eixo X:
Definir marcadores por:
Rotular casos por:
Painel por
Títulos...
Opções...

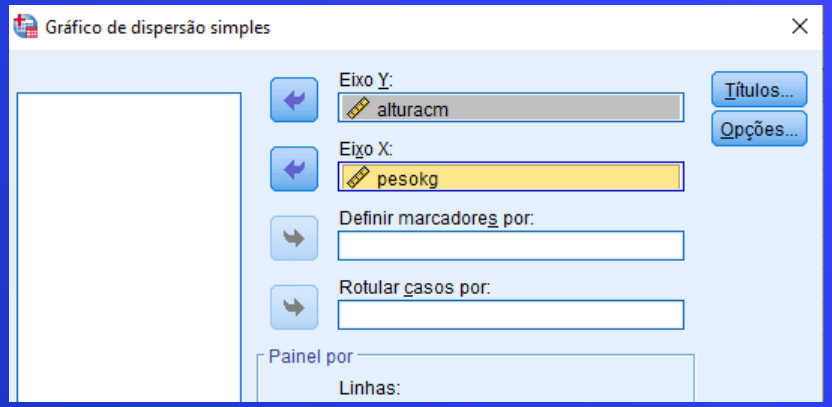
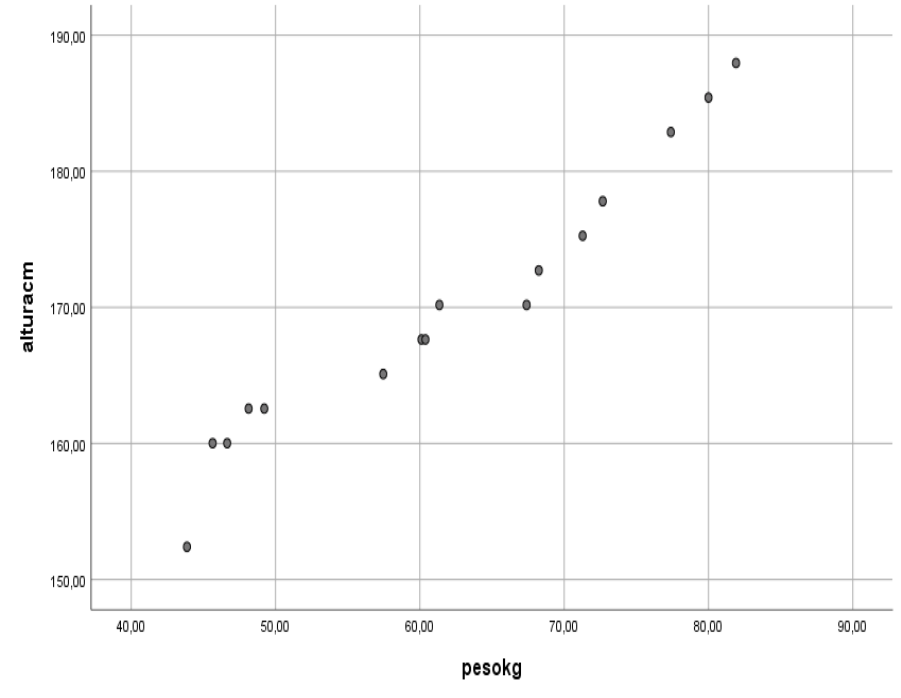


Gráfico de dispersão simples
Eixo Y:
alturacm
Eixo X:
pesokg
Definir marcadores por:
Rotular casos por:
Painel por
Linhas:
Títulos...
Opções...

O que o gráfico de dispersão

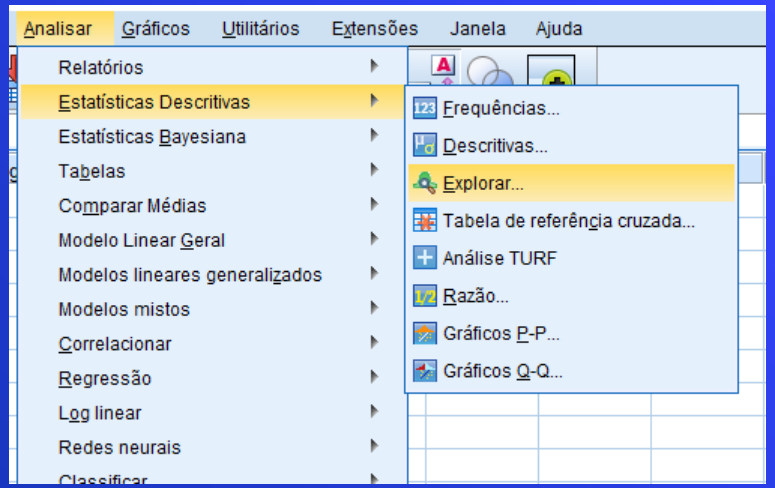


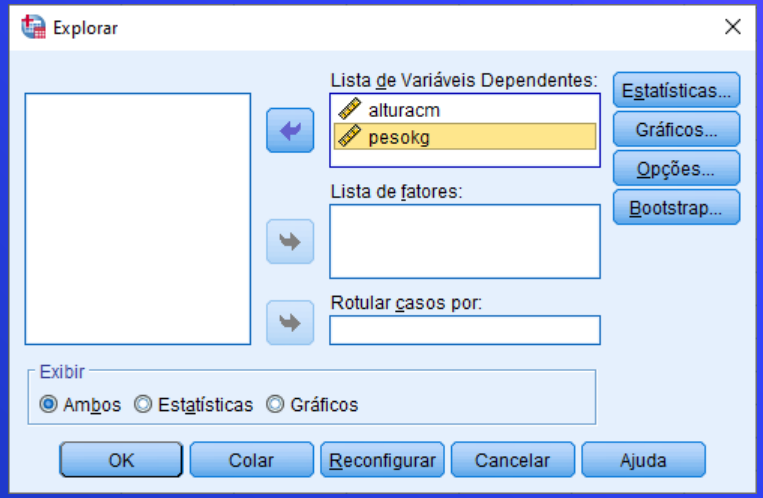
Existe algum indício de relação/correlação entre as variáveis ?
Em que sentido ?

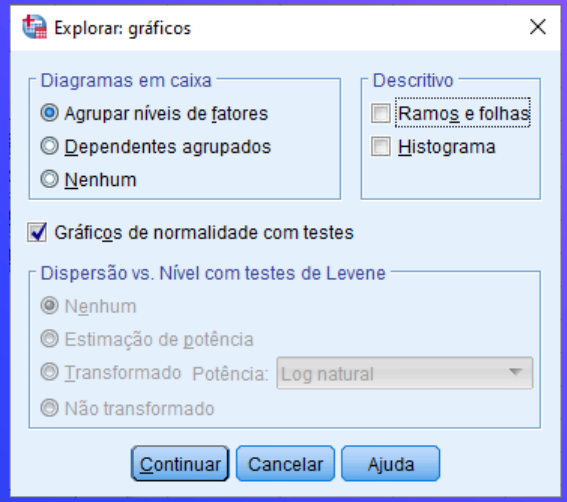
Testar a normalidade das distribuições do peso e da altura.

$\begin{cases} H_0: \text{A distribuição das alturas (em cm) da população de estudantes de onde foi retirada a amostra } \textbf{é} \text{ Normal;} \\ H_1: \text{A distribuição das alturas (em cm) da população de estudantes de onde foi retirada a amostra não } \textbf{é} \text{ Norm} \end{cases}$

$\begin{cases} H_0: \text{A distribuição do peso (em kg) da população de estudantes de onde foi retirada a amostra } \textbf{é} \text{ Normal;} \\ H_1: \text{A distribuição do peso (em kg) da população de estudantes de onde foi retirada a amostra não } \textbf{é} \text{ Normal;} \end{cases}$







Testes de Normalidade

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estatística	gl	Sig.	Estatística	gl	Sig.
alturacm	0,119	16	,200*	0,970	16	0,843
pesokg	0,154	16	,200*	0,935	16	0,295

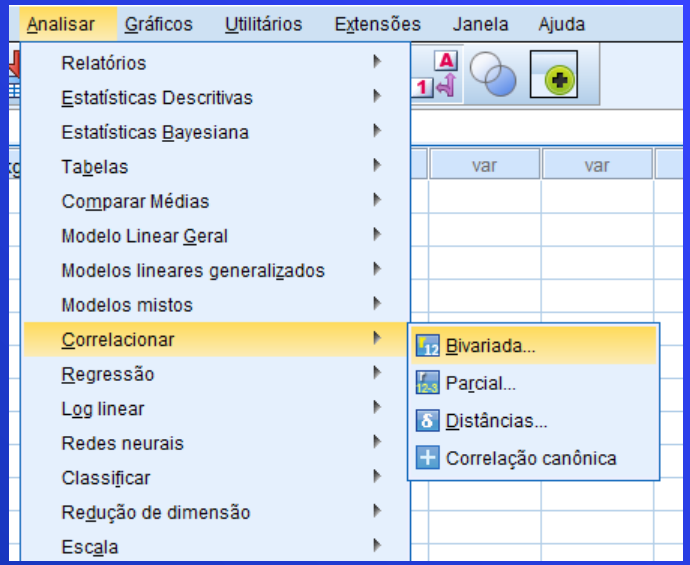
*. Este é

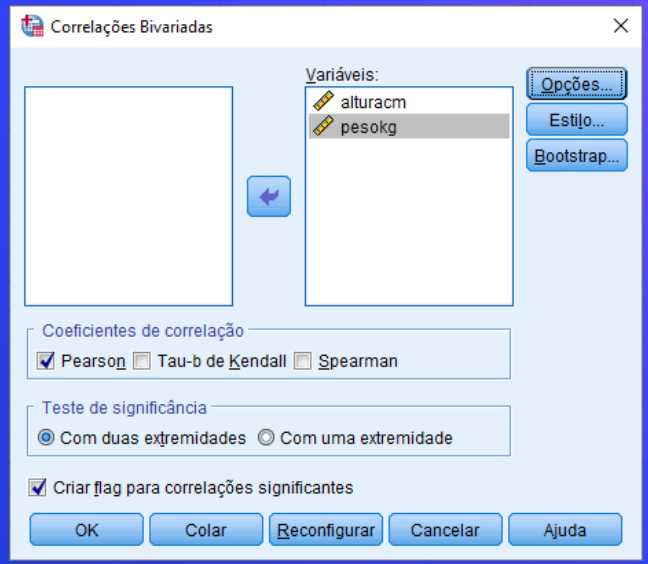
a.

Com base no valor-p associado ao teste de Shapiro Wilk, concluímos pela não rejeição da hipótese nula, ou seja pela normalidade de ambas as distribuições, para um α=5%.

Coeficiente de Correlação PARAMÉTRICO - PEARSON

$\begin{cases} H_0: \text{As variáveis peso (kg) e altura (cm) não estão relacionadas (são independentes)} \\ H_1: \text{As variáveis peso (kg) e altura (cm) estão relacionadas (não são independentes)} \end{cases}$ **OU** $\begin{cases} H_0: \rho = 0 \\ H_1: \rho \neq 0 \end{cases}$





$$\begin{cases} H_0: & \rho = 0 \\ H_1: & \rho \neq 0 \end{cases}$$

A matriz das correlações é simétrica (por exemplo como o mapa das distâncias entre capitais de distrito)

Correlações

		alturacm	pesokg
alturacm	Correlação de Pearson	1	,973**
	Sig. (2 extremidades)		0,000
	N	16	16
pesokg	Correlação de Pearson	,973**	1
	Sig. (2 extremidades)	0,000	
	N	16	16

**. A correlação

		alt
peso		
alt	1	0,973
peso	0,973	1

O valor do coeficiente de correlação amostral r =0,973 cujo valor é muito elevado segundo os citérios de Cohen & Holiday (1982). Com base no valor-p, rejeitamos Ho, ou seja as variáveis altura e peso estão fortemente relacionadas, para um α=5%.

Nota: caso uma das distribuições não fosse normalmente distribuída teríamos de utilizar o coeficiente não paramétrico de Spearman.

Timeline – Formação Certificada Gades Solutions

https://gades-solutions.com/lista-completa/

Estatística e Análise de Dados em Saúde com SPSS

16/05 a 07/06/2023 – Online / Duração: 16 horas / Formador: Ricardo São João / Em colaboração com CEAUL – Universidade de Lisboa

Modelos Multinível de Regressão para Dados em Painel com SPSS

Data a Anunciar – Online / Duração: 12 horas / Formador: Ricardo São João / Em colaboração com CEAUL – Universidade de Lisboa

JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV DEC

Estatística e Análise de Dados com R

20/06 a 06/07/2023 – Online / Duração: 18 horas / Formador: Ricardo São João / Em colaboração com CEAUL – Universidade de Lisboa

Obrigado pela Vossa PPA !!!

Presença, Participação e Atenção

Qualquer informação adicional pode ser enviada para o meu email
ricardo.sjoao@esg.ipsantarem.pt / rsj@net.sapo.pt