DEPARTAMENTO DE MATEMÁTICA DA UNIVERSIDADE DE COIMBRA

Estatística Computacional

(Licenciatura em Matemática)

Duração: 2h Frequência 24-05-2011

NOME: _____

Observação: A resolução completa das perguntas inclui a justificação do raciocínio utilizado.

I

Uma equipa médica desenvolveu um estudo envolvendo 271 doentes com perturbações gastrointestinais, de ambos os sexos e de várias faixas etárias. Para cada doente registou-se o sexo, a idade (em anos), o facto de ser ou não diabético e o grau de gravidade de alguns sintomas (Ausente, Ligeiro, Moderado, Grave, Muito grave, Excessivo).

1. Os valores observados das idades conduziram aos seguintes resultados:

```
Idade Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     5,00 Extremes    (=<27)
     1,00        3 .  3
     7,00        3 .  6777899
    16,00        4 .  0111122333444444
    24,00        4 .  55555666777788889999999
    42,00        5 .  000000000111122222222223333333344444444444
    53,00        5 .  55555555555666666677777777777777888888888899999999999
    40,00        6 .  0000000011122222222222333333333444444444
    53,00        6 .  55555555556666666666666677777777778888888888999999999
    24,00        7 .  000000000011111112222334
     4,00        7 .  5579
     2,00        8 .  11

 Stem width:        10
 Each leaf:       1 case(s)
```

Statistics

Idade

N	Valid	271
	Missing	0
Mean		57,75
Median		58,00
Mode		66
Std. Deviation		10,387
Skewness		-,668
Std. Error of Skewness		,148

Percentiles

		Percentiles						
		5	10	25	50	75	90	95
Weighted Average (Definition 1)	Idade	39,60	44,00	52,00	58,00	66,00	70,00	71,40

a) Que idade tinha o doente mais velho?
81 anos

b) Qual a percentagem de doentes com menos de 40 anos?

((5+1+7)/271)*100% = 4.8% (igualdade aproximada)

c) Interprete o valor 44 que figura no quadro *Percentiles*.

Trata-se do quantil de ordem 0.1 (percentil 10) da amostra: aproximadamente 10% dos doentes tem idade inferior ou igual a 44 anos.

d) Descreva a amostra no que diz respeito à localização central e à dispersão.

Localização central: a idade média dos doentes da amostra é 57.75 anos, a idade mediana é 58 anos e a moda é 66 anos.

Dispersão: O desvio padrão corrigido é igual a 10.387 anos

(Pode ainda referir-se a amplitude inter-quartis como medida de dispersão)

e) Que pode afirmar sobre a assimetria da amostra?

Assimetria negativa porque "média da amostra < mediana<moda". Esta afirmação é confirmada pelo sinal negativo do coeficiente de assimetria (-0.668).

Mais, como |Skewness / Std. Error of Skewness| = |-0.668/0.148| = 4.5 (aprox.), que é superior a 2, conclui-se que a assimetria é acentuada.

2. No quadro seguinte figuram os resultados do teste de Kolmogorov-Smirnov, com a correcção de Lilliefors, efectuado com base na amostra observada das idades.

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Idade	,072	271	,002	,970	271	,000

a. Lilliefors Significance Correction

a) Quais são as hipóteses em teste?

H0: X tem distribuição normal (N(57.75,10.387)) vs H1: X não tem distribuição normal,

onde X é a v.a.r. que representa a idade das pessoas que sofrem das perturbações referidas.

b) Qual a decisão a tomar relativamente às hipóteses referidas na alínea anterior?

Como a dimensão da amostra é muito elevada (271>30), a decisão é baseada no p-valor do teste de KS com a correcção de Lilliefors, que é igual a 0.002. Este valor é inferior aos níveis de significância usuais, pelo que rejeitamos H0, ou seja, *X* não segue uma lei normal.

c) Considera legítima a aplicação do teste de Student para verificar se a média das idades é inferior a 59 anos?

Sim, porque apesar de *X* não ter distribuição normal, a dimensão da amostra é muito elevada (271>30).

d) Apresenta-se a seguir o *output* do teste de Student.

One-Sample Test

	Test Value = 59					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
Idade	-1,977	270	,049	-1,247	-2,49	,00

(i) A que parâmetro corresponde o intervalo de confiança apresentado no quadro acima?

Trata-se de um i.c. para m-59, com $m=E(X)$.

(ii) Qual é o p-valor do teste referido na alínea c)? A que conclusão conduz este valor?

As hipóteses em teste são H0: $m = 59$ vs H1: $m < 59$. Como a média da amostra é inferior a 59, a amostra aponta no sentido de H1. Então o p-valor pedido é 0.049/2 = 0.0245. Considerando os níveis de significância (n.s.) mais usuais (0.01 e 0.05) tem-se que

- se n.s. = 0.05, então rejeita-se H0 porque p-valor < n.s. (neste caso, aceitamos que a idade média dos doentes é inferior a 59 anos);

- se n.s. = 0.01, então aceita-se H0 porque p-valor > n.s. (neste caso, não podemos afirmar que a idade média dos doentes é inferior a 59 anos).

3. Neste estudo investigaram-se possíveis relações entre a presença dos referidos sintomas e a presença da diabetes. Em particular, no que diz respeito a um determinado sintoma, designado *sintoma A*, obteve-se a seguinte tabela de contingência e o correspondente teste do qui-quadrado.

Diabetes * Sintoma A Crosstabulation

Count

		Sintoma A		
		Ausente	Presente	Total
Diabetes	Não diabético	94	37	131
	Diabético	89	51	140
Total		183	88	271

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
Pearson Chi-Square	2,067[a]	1	,150	,156
Continuity Correction[b]	1,711	1	,191	
N of Valid Cases	271			

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 42,54.
b. Computed only for a 2x2 table

a) Indique

(i) a percentagem de doentes que apresentam o sintoma A;

(88/271)*100% = 32.47% (igualdade aproximada)

(ii) a percentagem de doentes diabéticos que apresentam o sintoma A.

(51/140)*100% = 36.43% (igualdade aproximada)

b) O teste efectuado permite concluir que há associação entre a presença do sintoma A e a presença da diabetes?

As hipóteses em teste são H0: A presença do sintoma A e a presença da diabetes são independentes vs H1: A presença do sintoma A e a presença da diabetes não são independentes.

p-valor = 0.156 é superior aos n.s. usuais, logo aceita-se H0, isto é, o teste não permite concluir que há associação entre a presença do sintoma A e a presença da diabetes.

4. A análise da relação entre o grau de gravidade do sintoma A e o o grau de gravidade de outro sintoma, designado *sintoma B*, nos doentes diabéticos, conduziu ao seguinte *output*:

Sintoma A * Sintoma B Crosstabulation[a]

			Sintoma B						Total
			Ausente	Ligeiro	Moderado	Grave	Muito grave	Excessivo	
Sintoma A	Ausente	Count	62	9	7	6	2	3	89
		Expected Count	55,9	8,3	6,4	7,0	5,1	6,4	89,0
	Ligeiro	Count	7	1	2	1	3	0	14
		Expected Count	8,8	1,3	1,0	1,1	,8	1,0	14,0
	Moderado	Count	12	1	0	2	0	2	17
		Expected Count	10,7	1,6	1,2	1,3	1,0	1,2	17,0
	Grave	Count	4	1	0	1	3	2	11
		Expected Count	6,9	1,0	,8	,9	,6	,8	11,0
	Muito grave	Count	3	1	1	1	0	3	9
		Expected Count	5,7	,8	,6	,7	,5	,6	9,0
Total		Count	88	13	10	11	8	10	140
		Expected Count	88,0	13,0	10,0	11,0	8,0	10,0	140,0

a. Diabetes = Diabético

Chi-Square Tests[b]

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	40,083[a]	20	,005
Likelihood Ratio	34,179	20	,025
Linear-by-Linear Association	13,956	1	,000
N of Valid Cases	140		

a. 20 cells (66,7%) have expected count less than 5. The minimum expected count is ,51.
b. Diabetes = Diabético

a) Qual é o significado dos valores *Count* e *Expected Count* que figuram no primeiro quadro?

Sejam A_i: modalidades do atributo "grau de gravidade do sintoma A", i=1,…,5, e B_j: modalidades do atributo "grau de gravidade do sintoma B", i=1,…,6.

Count: frequências absolutas observadas para cada par (A_i, B_j), i=1,…,5; j=1,…,6.

Expected Count: frequências absolutas esperadas para cada par (A_i, B_j), i=1,…,5; j=1,…,6, sob a hipótese da independência dos atributos.

b) Será recomendável tomar uma decisão com base neste *output*?

Não, porque há frequências observadas inferiores a 5.

5. Pretendeu-se também tirar conclusões sobre a percentagem de diabéticos que não apresentam o sintoma B. Para tal, foi efectuado um teste binomial, cujos resultados figuram a seguir.

Binomial Test[a]

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
Sintoma B	Group 1	Presente	52	,37	,35	,326
	Group 2	Ausente	88	,63		
	Total		140	1,00		

a. Diabetes = Diabético

a) Quais são as hipóteses em teste subjacentes a este output?

H0: $p = 0.35$ vs H1: $p>0.35$, onde p designa a proporção de doentes diabéticos (na população de doentes com as perturbações em causa) que apresentam o sintoma B.

b) A percentagem de diabéticos que não apresentam o sintoma B pode ser considerada superior a 65%?

As hipóteses referidas na alínea anterior são equivalentes a H0: 1-p = 0.65 vs H1: 1-p < 0.65. Note-se que 1-p representa a a proporção de doentes diabéticos (na população de doentes com as perturbações em causa) que não apresentam o sintoma B. Este teste (unilateral) tem p-valor=0.326. Agora pretendemos testar H0: 1-p = 0.65 contra H1': 1-p > 0.65. A hipótese H1' aponta no sentido contrário ao da amostra observada, pelo que o p-valor deste teste é igual a 1-0.326 = 0.674, que é superior aos níveis de significância usuais. Portanto, não podemos afirmar que a proporção de diabéticos que não apresentam o sintoma B é superior a 0.65.

II

O gestor de uma agência de viagens fez vários estudos para melhor compreender as opções dos seus clientes relativamente a viagens turísticas.

1. No que diz respeito a uma amostra de tempos de duração das viagens ao estrangeiro (em dias), obteve o seguinte *output*:

Runs Test

	Duração
Test Value[a]	10
Cases < Test Value	23
Cases >= Test Value	27
Total Cases	50
Number of Runs	24
Z	-,529
Asymp. Sig. (2-tailed)	,597

a. Median

a) Como são definidos os *Runs*?

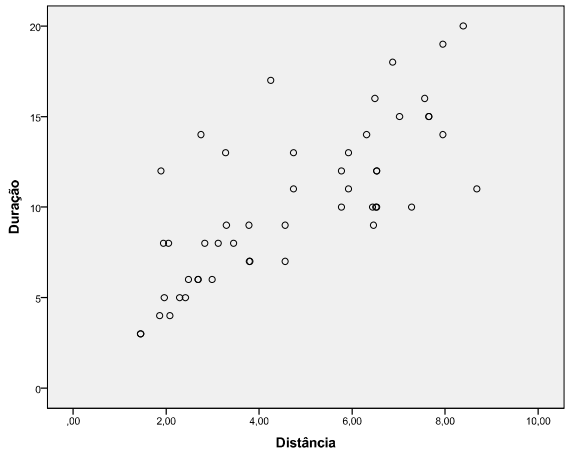
Os *Runs* são sequências de símbolos, por exemplo A e B, sendo que A corresponde (neste caso) a um valor inferior à mediana da amostra e B corresponde a um valor superior ou igual à mediana da amostra na amostra inicial (não ordenada).

b) Quais as hipóteses em confronto? Que pode concluir aos níveis de significância usuais?

H0: a amostra dos tempos de duração das viagens é aleatória vs H1 : a amostra dos tempos de duração das viagens não é aleatória.

O p-valor tem o valor aproximado de 0.597 (superior aos níveis de significância usuais), logo aceita-se H0, isto é, a amostra é aleatória.

2. Para cada viagem, foi também registada a região de destino. Para relacionar a duração da viagem com a distância ao seu destino (em milhares de quilómetros), foram construídos o diagrama de dispersão e o quadro que se seguem.



Dos quadros seguintes, contendo valores dos coeficientes de correlação de Pearson e de Spearman, qual poderá corresponder a este diagrama de dispersão?

Symmetric Measures

		Value
Interval by Interval	Pearson's R	,022
Ordinal by Ordinal	Spearman Correlation	,036
N of Valid Cases		50

Quadro 1

Symmetric Measures

		Value
Interval by Interval	Pearson's R	,751
Ordinal by Ordinal	Spearman Correlation	,766
N of Valid Cases		50

Quadro 2

Symmetric Measures

		Value
Interval by Interval	Pearson's R	-,679
Ordinal by Ordinal	Spearman Correlation	-,634
N of Valid Cases		50

Quadro 3

Symmetric Measures

		Value
Interval by Interval	Pearson's R	,679
Ordinal by Ordinal	Spearman Correlation	,898
N of Valid Cases		50

Quadro 4

A nuvem de pontos apresenta alguma dispersão, mas nota-se uma tendência crescente passível de ser aproximada por uma relação linear. Assim, não pode ser o quadro 1 porque os dois coeficientes apresentam valores próximos de 0. Também não pode ser o quadro 3, porque os valores negativos dos coeficientes são indicadores de uma relação funcional decrescente entre as variáveis. Por outro lado, quando uma relação linear se afigura adequada, os dois coeficientes têm valores próximos. Assim, o quadro 4 também não corresponde à nuvem de pontos apresentada. Solução: quadro 2.

3. Com o objectivo de verificar se as regiões de destino das viagens eram igualmente preferidas pelos clientes da agência, foi efectuado um teste de ajustamento do qui-quadrado, que forneceu os seguintes resultados:

Destino

	Observed N	Expected N	Residual
África	14	12,5	1,5
Brasil	11	12,5	-1,5
Caraíbas	9	12,5	-3,5
Europa Central	16	12,5	3,5
Total	50		

Test Statistics

	Destino
Chi-Square	2,320
df	3
Asymp. Sig.	,509

Relativamente ao objectivo referido, que pode concluir?

As hipóteses em teste são H0: as regiões de destino das viagens são igualmente preferidas pelos clientes da agência *vs* H1: as regiões de destino das viagens não são igualmente preferidas pelos clientes da agência.

p-valor aproximado = 0.509 é superior aos n.s. usuais, logo aceita-se H0, isto é, podemos concluir que as regiões de destino das viagens são igualmente preferidas pelos clientes da agência.

III

Sejam X e Y duas variáveis aleatórias absolutamente contínuas e independentes. Encontram-se abaixo alguns resultados da análise descritiva de duas amostras, uma da variável X e outra da variável Y, bem como o resultado do teste de Mann-Whitney aplicado a estas amostras.

Descriptives

Variável		Statistic	Std. Error
X	Mean	,9081	,05565
	Median	**,6630**	
	Std. Deviation	,87997	
	Minimum	,00	
	Maximum	4,77	
	Range	4,76	
	Interquartile Range	1,03	
	Skewness	1,569	,154
	Kurtosis	2,574	,307
Y	Mean	,4618	,05008
	Median	**,6630**	
	Std. Deviation	,77581	
	Minimum	-2,14	
	Maximum	2,00	
	Range	4,14	
	Interquartile Range	,98	
	Skewness	-1,200	,157
	Kurtosis	1,056	,313

Test Statistics[a]

Mann-Whitney U	24974,500
Z	-3,207
Asymp. Sig. (2-tailed)	,001

a. Grouping Variable: Variável

1. Como explica o resultado do teste de Mann-Whitney, face à igualdade das medianas das amostras observadas?

O teste de Mann-Whitney para igualdade de medianas pressupõe, em particular que as duas distribuições em causa tenham a mesma forma. Tal pressuposto não se verifica aqui, pois a distribuição de *X* é fortemente assimétrica positiva (|1.569/0.154| é muito superior a 2) enquanto a distribuição de *Y* é fortemente assimétrica negativa (|-1.2/0.157| é muito superior a 2). Portanto o teste não tem condições de aplicabilidade para testar a igualdade das medianas.

2. Pretende-se agora testar a hipótese H_0: $\mu_Y = 0.75$, onde μ_Y representa a mediana de *Y*.

a) Qual é a hipótese alternativa que está de acordo com a tendência da amostra?

H1: $\mu_Y < 0.75$, porque a mediana observada de *Y* é 0.663 < 0.75.

b) Que teste usaria neste caso?

Usaria o teste dos sinais, uma vez que o teste de Wilcoxon exige a simetria da distribuição, o que não se verifica neste caso, como foi referido na alínea a).

c) O SPSS fornece o valor 0.107 para o p-valor assintótico do teste bilateral adequado para esta situação. Para que níveis de significância aceitaria a hipótese alternativa a que se refere a alínea a)?

O p-valor do teste é 0.107/2 = 0.0535. Aceitar H1 é equivalente a rejeitar H0, o que acontece quando o p-valor é inferior ou igual ao nível de significância. Assim, aceitaria a hipótese em causa para níveis de significância superiores ou iguais a 0.0535.

3. Seja *Z* uma outra variável aleatória real absolutamente contínua da qual se dispõe uma amostra tal que o coeficiente de correlação amostral entre *X* e *Z* é igual a 0.00004. Este valor permite concluir que *X* e *Z* não estão relacionadas?

Apenas permite concluir que não estão relacionadas linearmente, não excluindo a possibilidade de existência de outro tipo de relação entre elas. Portanto, este valor não permite concluir que *X* e *Z* não estão relacionadas.