

**Estatística para Economia e Gestão**  
**Licenciatura em Economia e Licenciatura em Gestão**

Nova School of Business and Economics

Exame Final de 2<sup>a</sup> Época

23 de junho de 2012

Duração: 2 horas e 15 minutos

**INSTRUÇÕES**

Material autorizado: Caneta e este enunciado.

Escreva o seu nome e número de aluno na primeira página deste enunciado.

Este enunciado deve permanecer sempre agrafado.

As respostas às questões devem ser escritas neste enunciado nos locais indicados.

Pode utilizar o verso de cada folha como rascunho.

Qualquer situação de plágio (como sejam a utilização de material não autorizado, comunicação com colegas, etc.) terá como consequência imediata a reprovação à disciplina neste semestre.

Não é permitido tirar dúvidas durante o exame.

Antes de iniciar o exame confirme que este enunciado tem 19 páginas numeradas de 1 a 19.

No final deste enunciado são incluídos um formulário e tabelas estatísticas que podem ser necessárias para responder a algumas das questões deste exame.

Deve permanecer sentado no seu lugar até ao final do exame.

A recolha final do enunciado será feita pelos vigilantes.

Nota: Os dados e factos mencionados neste exame são meramente exemplificativos e não correspondem necessariamente ao mundo real.

**Para cada questão dos Grupos I-III indique uma só resposta na tabela da página 9**

Cada resposta certa vale 0,9 valores.

Cada resposta errada vale 0,3 valores negativos (-0,3).

Cada resposta em branco ou mal assinalada vale 0 valores.

### Grupo I

Todas as casas da cidade de Enerpólis são abastecidas de gás natural e eletricidade. Para uma casa escolhida aleatoriamente desta cidade, o consumo de gás natural (em metros cúbicos) e o consumo de eletricidade (em milhares de kWh) durante um ano são variáveis aleatórias com distribuições normais, com médias iguais a 2 000 e 10 000, e desvio-padrões iguais a 1 000 e 4 000, respectivamente. A moeda da cidade de Enerpólis é o tostão. Para calcular as despesas em tostões a partir dos consumos, sabemos que o preço do gás natural é igual a 3 tostões por metro cúbico, enquanto que o preço da eletricidade é igual a 1 tostão por kWh. Responda às seguintes questões relativas a uma casa escolhida ao acaso nesta cidade.

1. Qual a probabilidade do consumo anual de eletricidade ser inferior a 5 mil kWh.

- a) 0,11
- b) 0,31
- c) 0,69
- d) 0,89

$$E \sim N(10\ 000, 4\ 000^2)$$

$$P(E < 5000) = P(Z < \frac{5000 - 10\ 000}{4\ 000}) =$$

$$= P(Z < -\frac{5}{4}) = P(Z < -1,25) = 1 - P(Z < 1,25) = 1 - 0,8944$$

2. Existe uma probabilidade de 5% do consumo anual em gás natural exceder um certo valor. Qual é esse valor?

- a) 3645 metros cúbicos
- b) 4020 metros cúbicos
- c) 9692 metros cúbicos
- d) 12845 metros cúbicos

$$G \sim N(2\ 000, 1\ 000^2)$$

$$P(G > k) = 5\% \Leftrightarrow P(Z > \frac{k - 2\ 000}{1\ 000}) = 5\%$$

$$\Leftrightarrow \frac{k - 2\ 000}{1\ 000} = 1,645 \Rightarrow k = 2\ 000 + 1,645 \cdot 1\ 000 = 3645$$

3. Qual o valor esperado da soma das despesas anuais de eletricidade e gás natural?

- a) 24 000 tostões
- b) 16 000 tostões
- c) 8 000 tostões
- d) 4 000 tostões

$$E(3G + 1 \cdot E) = 3 \cdot 2\ 000 + 1 \cdot 10\ 000$$

$$= 16\ 000$$

4. Se a correlação entre os consumos de eletricidade e de gás natural for igual a

-0,5, qual a variância da soma das despesas anuais em eletricidade e gás natural?

- a) 37 000 000 tostões<sup>2</sup>
- b) 25 000 000 tostões<sup>2</sup>
- c) 17 000 000 tostões<sup>2</sup>
- d) 13 000 000 tostões<sup>2</sup>

$$\begin{aligned} V(3G + 1 \cdot E) &= 3^2 V(G) + 1^2 V(E) + 2 \cdot 3 \cdot 1 \cdot \text{cov}(G, E) \\ &= 9 \cdot 1\ 000^2 + 4\ 000^2 + 6 \cdot (-0,5) \cdot 1\ 000 \cdot 4\ 000 \\ &= (9 + 16 - 12) \cdot 1\ 000^2 = 13\ 000\ 000 \end{aligned}$$

### Grupo II

A maior empresa de aluguer de automóveis vai ser inspecionada por uma agência de certificação de qualidade. Um dos critérios para se receber um certificado de qualidade está relacionado com as condições dos automóveis. A empresa de aluguer sabe que 20% da sua frota de vários milhares de automóveis tem defeitos que não são aceitáveis de acordo com os critérios da agência de certificação. No

entanto, a agência de certificação apenas irá inspecionar uma amostra aleatória de 100 automóveis. Se a proporção de automóveis com defeito nessa amostra exceder 22%, a agência de qualidade não atribuirá o certificado de qualidade à empresa

1. Qual o valor esperado da proporção de automóveis com defeito nessa amostra?

- a) 2%
- b) 20%  $E(\hat{P}) = P = 0,20$
- c) 22%
- d) 50%

2. Qual a probabilidade de se conseguir cumprir o critério dos 22% necessário para poder ser atribuído o certificado de qualidade?

- a) 98%  $P(\hat{P} < 0,22) = P(Z < \frac{0,22 - 0,20}{\sqrt{\frac{0,2 \times 0,8}{100}}})$
- b) 69%  $\rightarrow P(Z < \frac{0,02}{0,04}) = P(Z < 0,5) = 0,6915$
- c) 26%
- d) 11%

3. Como se alterava a resposta à alínea anterior se a amostra aleatória tivesse uma dimensão superior a 100?

- a) não se alterava.
- b) a probabilidade seria inferior à obtida na alínea anterior.
- c) a probabilidade seria superior à obtida na alínea anterior.
- d) a probabilidade seria diferente da obtida na alínea anterior, podendo vir inferior ou superior.

4. A agência de certificação analisou uma amostra aleatória de 100 automóveis tendo encontrado 30 com defeito. Sabendo apenas este resultado, o responsável

$H_0: p \leq 0,2$  pelo controle de qualidade na empresa de aluguer de automóveis quer testar se a percentagem de automóveis com defeito em toda a população de automóveis na empresa excede os 20%. Qual das seguintes afirmações está correta?

- a) A estatística de teste é igual a 0,5 pelo que, para um nível de significância de 5%, a evidência estatística não é suficiente para se concluir que a percentagem de automóveis com defeito em toda a população de automóveis na empresa excede os 20%.
- b) A estatística de teste é igual a -2,5 pelo que, para um nível de significância de 5%, a evidência estatística é suficiente para se concluir que a percentagem de automóveis com defeito em toda a população de automóveis na empresa excede os 20%.
- c) A estatística de teste é igual a 2,5 pelo que, para um nível de significância de 5%, a evidência estatística é suficiente para se concluir que a percentagem de automóveis com defeito em toda a população de automóveis na empresa excede os 20%.
- d) A estatística de teste é igual a 2,5 pelo que, para um nível de significância de 5%, a evidência estatística não é suficiente para se concluir que a percentagem de automóveis com defeito em toda a população de automóveis na empresa excede os 20%.

$$Z = \frac{(0,3 - 0,2)}{\sqrt{\frac{0,2 \times 0,8}{100}}} = \frac{0,10}{0,04} = 2,5 > Z_{5\%} = 1,645$$

### Grupo III

Será que os alunos em turmas mais pequenas têm uma melhor aprendizagem e conseguem atingir melhores resultados? Este é um problema importante que tem sido estudado em muitos países e com diversas abordagens. Neste exercício vamos utilizar uma base de dados de 220 distritos escolares da rede pública de ensino do estado do Massachusetts nos EUA. Para cada um dos distritos escolares estão disponíveis dados para as seguintes variáveis:

- **RESULT**: resultado (numa escala de 600 a 840 em que o pior resultado corresponde a 600 e o melhor desempenho corresponde a 840) das provas de aferição dos alunos do 4º ano do ensino básico ( dado pela média dos resultados de todos os alunos de 4º ano no distrito escolar);
- **RACIO**: rácio nº alunos / nº professores de 4º ano no distrito escolar;
- **REND\_ALTO**: = 1 se o rendimento per capita nesse distrito escolar é superior à mediana de todos os distritos do estado do Massachusetts; = 0 se não;
- **NECESS**: percentagem de alunos de 4º ano com necessidades educativas especiais no distrito escolar;
- **ALMOCO**: percentagem de alunos de 4º ano com direito a almoço grátis ou a preço reduzido no distrito escolar;
- **DESPESA**: despesas totais em equipamentos de educação por aluno no distrito escolar (em dólares);
- **N\_INGLES**: percentagem dos alunos de 4º ano no distrito escolar cuja língua mãe não é o inglês.

Com base nestes dados estimou-se a seguinte regressão linear em que a variável dependente é RESULT.

#### Regressão 1:

<i>Regression Statistics</i>	
Multiple R	0.551002
R Square	0.303604
Adjusted R Square	0.297185
Standard Error	12.68113
Observations	220

#### ANOVA

	<i>Df</i>	<i>SS</i>	<i>MS</i>	<i>Significance F</i>	
				<i>F</i>	<i>F</i>
Regression	2	15213.41	7606.705	47.30209	8.91E-18
Residual	217	34896.03	160.8112		
Total	219	50109.44			

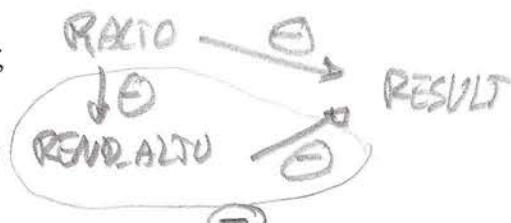
	<i>Coefficients</i>	<i>Standard</i>		
		<i>Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	720.9795	6.932372	104.0019	4.1E-187
RACIO	-1.07469	0.383764	-2.8004	0.005565
REND_ALTO	14.9747	1.743429	8.589225	1.72E-15

1. Tendo em conta apenas os resultados da Regressão 1, responda às seguintes questões.
  - 1.1. Os sinais das estimativas dos coeficientes estimados indicam que:
    - a) Os distritos em que os alunos de 4º ano têm turmas mais pequenas têm em média resultados piores, *ceteris paribus*.
    - b) Os distritos em que os alunos de 4º ano têm turmas mais pequenas têm em média resultados melhores, *ceteris paribus*.
    - c) Os distritos em que os alunos de 4º ano têm turmas mais pequenas têm em média resultados piores apenas se o rendimento per capita no distrito for mais elevado que a mediana do estado.
    - d) Os distritos em que os alunos de 4º ano têm turmas mais pequenas têm em média resultados melhores apenas se o rendimento per capita no distrito for mais elevado que a mediana do estado.
  - 1.2. Para um nível de significância de 5%, quais as variáveis explicativas que explicam os resultados das provas de aferição dos alunos de 4º ano?
    - a) Apenas RACIO.
    - b) Apenas REND\_ALTO.
    - c) RACIO e REND\_ALTO.
    - d) Nenhuma das variáveis explica os resultados das provas de aferição.
  - 1.3. O intervalo de confiança a 95% para o coeficiente de RACIO é dado por:
    - a) [-1,5 ; -0,7];
    - b) [-1,8 ; -0,3];
    - c) [-2,1 ; 0,0];
    - d) [-3,9 ; 1,7].

$-1,07 \pm 1,96 \times 0,38$   
 $\approx -1,07 \pm 0,76 = [-1,8 ; -0,3]$
  - 1.4. É plausível que haja um impacto negativo nos resultados dos testes de aferição dos alunos de 4º ano caso se aumente em uma unidade o rácio nº alunos / professor. No entanto seria interessante estudar a possibilidade de esse impacto ser menos pronunciado em turmas maiores. Qual das seguintes afirmações dá a melhor resposta para se poder concluir algo sobre essa possibilidade?
    - a) Para um nível de significância de 5% pode-se concluir que tal possibilidade é rejeitada pois o impacto estimado é fixo e significativamente negativo;
    - d) Os resultados apresentados para a Regressão 1 não confirmam essa possibilidade uma vez que o coeficiente da variável REND\_ALTO não é significativo a 5%;
    - c) Os resultados apresentados para a Regressão 1 nada permitem concluir sobre essa possibilidade. Para se poder concluir algo sobre essa afirmação poder-se-ia estimar um novo modelo em que se acrescentaria uma nova variável dada pelo produto das variáveis RACIO e REND\_ALTO;
    - d) Os resultados apresentados para a Regressão 1 nada permitem concluir sobre essa possibilidade. Para se poder concluir algo sobre essa afirmação poder-se-ia estimar um novo modelo em que se acrescentaria uma nova variável dada por RACIO ao quadrado;

1.5. Sabe-se que os distritos com rendimento per capita mais elevado têm em média menos alunos por turma. Caso seja estimada uma nova regressão em que a variável dependente é a mesma da Regressão 1 mas única variável explicativa é RACIO, o que se espera que aconteça ao coeficiente estimado dessa variável explicativa? Espera-se que o coeficiente estimado de RACIO na nova regressão seja:

- a) aproximadamente igual ao obtido na Regressão 1;
- b) menos negativo que o obtido na Regressão 1;
- c) ainda mais negativo que o obtido na Regressão 1;
- d) positivo.



1.6. Qual o resultado do teste da hipótese nula de que o coeficiente da variável REND\_ALTO não é superior a 5?

- $H_0: \beta \leq 5$
- $H_1: \beta > 5$
- a) Rejeita-se a hipótese nula para um nível de significância de 10%
  - b) Não se rejeita a hipótese nula para um nível de significância de 1%
  - c) Não se rejeita a hipótese nula para um nível de significância de 2,5%
  - d) Não se rejeita a hipótese nula para um nível de significância de 5%

$$t = (15 - 5) / 1,74 = 10 / 1.74 > z_{10\%} = 1,28$$

2. Por forma a controlar o efeito de outros fatores que possam influenciar os resultados nas provas de aferição foi estimada a seguinte regressão em que REND\_ALTO \*RACIO designa uma variável dada pelo produto das variáveis REND\_ALTO e RACIO.

## Regressão 2

Regression Statistics	
Multiple R	0.820563
R Square	0.673323
Adjusted R Square	0.662537
Standard Error	8.787211
Observations	220

## ANOVA

	df	SS	MS	Significance	
				F	F
Regression	7	33739.84	4819.977	62.42274	4.28E-48
Residual	212	16369.6	77.21508		
Total	219	50109.44			

	Coefficients	Standard				Upper 95%
		Error	t Stat	P-value	Lower 95%	
1 Intercept	726.837	9.028919	80.50099	6.7E-161	709.039	744.6349
20 RACIO	-0.80481	0.353857	-2.27439	0.023943	-1.50234	-0.10728
0 REND_ALTO	-26.2292	10.28178	-2.55103	0.011445	-46.4968	-5.96153
0 REND_ALTO *RACIO	1.665732	0.588666	2.82967	0.005107	0.505343	2.826121
0 NECESS	-0.4267	0.173909	-2.45357	0.014952	-0.76951	-0.08389
0 ALMOCO	-0.66492	0.065963	-10.0803	8.97E-20	-0.79495	-0.5349
1000 DESPESA	0.002504	0.000783	3.198934	0.001591	0.000961	0.004047
0 N_INGLES	-0.41432	0.295241	-1.40333	0.161982	-0.9963	0.167664

2.1. De acordo com os resultados desta regressão, qual a estimativa para o impacto no resultado das provas de aferição de haver mais um aluno por professor num distrito com rendimento per capita superior à mediana assumindo tudo o resto constante?

- a) decréscimo médio de 0,8
- b) aumento médio de 0,9
- c) decréscimo médio de 24,6
- d) aumento médio de 26,2

$$-0,8 + 1,66 = 0,86 \approx 0,9$$

2.2. Qual das seguintes afirmações está correta?

- a) Caso a Regressão 2 fosse re-estimada incluindo como variáveis explicativas adicionais o número de computadores por aluno e o salário médio dos professores, o novo  $R^2$  viria maior apenas se essas variáveis fossem significativas a 5%.
- b) Caso a Regressão 2 fosse re-estimada excluindo todas as variáveis não significativas a 1% o novo  $R^2$  viria maior.
- c) Caso a Regressão 2 fosse re-estimada incluindo como variáveis explicativas adicionais o número de computadores por aluno e o salário médio dos professores, o novo  $R^2$  nunca viria inferior.
- d) Caso a Regressão 2 fosse re-estimada excluindo todas as variáveis com estatísticas t inferiores a 2 em valor absoluto, o novo  $R^2$  viria superior.

2.3. Qual a melhor estimativa da média dos resultados de aferição dos alunos de 4º ano num distrito com rendimento per capita inferior à mediana em que: o número médio de alunos por professor seja de 20, não existam alunos com necessidades especiais nem com direito a almoço grátis ou com desconto, em que todos os alunos tenham como língua mãe o inglês, e em que as despesas totais em equipamentos de educação por aluno seja igual a 1000?

- a) 713
- b) 726
- c) 739
- d) 777

$$\begin{aligned} & 726,8 + (-0,8) \times 20 + 0,0025 \times 1000 \\ & = 726,8 - 16 + 2,5 \\ & = 713 \end{aligned}$$

3. (1 valor) Noutro estado dos EUA decidiu-se estudar até que ponto o rácio nº alunos/ nº professores tinha impacto na aprendizagem dos alunos através de outra abordagem. Em primeiro lugar, para medir a aprendizagem dos alunos utilizou-se um sistema de avaliação mais eficiente baseado em entrevistas pessoais com cada aluno. O resultado final da avaliação foi expresso numa escala de 0 a 100. Deste modo, no ano letivo de 2005-2006 foram escolhidas de forma aleatória 25 escolas do ensino básico e no final desse ano letivo foram avaliados todos os seus alunos de 4º ano. Nos anos letivos seguintes essas escolas reduziram em 10% a dimensão média das turmas e no final do ano letivo de 2009-2010, todos os alunos de 4º ano dessas escolas foram avaliados pelo mesmo sistema de avaliação. Para cada escola calculou-se quanto é que variou a avaliação média dos seus alunos do ano letivo 2005-2006 para o ano letivo 2009-2010. Finalmente, calculou-se para o conjunto

das 25 variações (das 25 escolas) a média e o desvio-padrão para se poder levar a cabo um teste estatístico. Será que esta abordagem permitiria concluir algo sobre se a redução do tamanho das turmas **causa** uma melhoria do desempenho dos alunos? Se a sua resposta for positiva, indique qual seria a forma estatística correta de testar essa causalidade. Se a sua resposta for negativa, explique porquê e proponha uma modificação à abordagem seguida que permitisse realizar no final desse estudo um teste de causalidade. Responda no espaço seguinte.

NÃO PERMITE TESTAR CAUSALIDADE PORQUE  
PARA ALÉM DA REDUÇÃO EM 10% DA  
DIMENSÃO MÉDIA DAS TURMAS, PODERÃO TER  
OCORRIDO OUTRAS ALTERAÇÕES COM  
IMPACTO NA AVALIAÇÃO MÉDIA DOS ALUNOS.

UMA SOLUÇÃO SERIA DEFINIR TAMBÉM UM  
GRUPO DE CONTROLE COMO UMA amostra  
ALEATÓRIA DE ESCOLAS, E GARANTIR QUE  
ESSAS ESCOLAS NÃO ALTERAVAM O  
TAMANHO DAS TURMAS.

O TESTE DE CAUSALIDADE PODERIA SER  
FEITO ATRAVÉS DE UM TESTE DE  
DIFERENÇA DE MÉDIAS DOS DOIS  
GRUPOS.

**Tabela de respostas às questões dos Grupos I a III****Assinale as suas respostas com um X**

<b>Questão</b>	<b>Resposta</b>			
	<b>a</b>	<b>b</b>	<b>C</b>	<b>d</b>
I.1				
I.2				
I.3				
I.4				
II.1				
II.2				
II.3				
II.4				
III.1.1				
III.1.2				
III.1.3				
III.1.4				
III.1.5				
III.1.6				
III.2.1				
III.2.2				
III.2.3				
III.3	<i>Responda no espaço delimitado no final da página anterior</i>			
IV	<i>Responda nos espaços delimitados nas páginas seguintes</i>			
V	<i>Responda nos espaços delimitados nas páginas seguintes</i>			
VI	<i>Responda nos espaços delimitados nas páginas seguintes</i>			

***Não esquecer de responder aos Grupos IV, V e VI nas páginas seguintes***

**Grupo IV**

Um grupo de professores de Marketing realizou uma experiência para analisar a memória das crianças quando expostas a anúncios de televisão com som ou sem som. A população alvo consiste em crianças com 6 anos. Foram constituídas de forma independentes duas amostras aleatórias dessa população. A cada criança na primeira amostra foi mostrado um anúncio de 1 minuto de uma marca de Cereais de Mel com imagem e som. Às crianças na segunda amostra foi mostrado o mesmo anúncio mas sem som. No final, cada criança foi questionada sobre 50 itens que apareceram no anúncio. O número de itens que cada criança se lembrou corretamente foi registado. Um resumo dos resultados é apresentado na tabela seguinte.

	Amostra	
	Com Som	Sem Som
Número de Crianças	30	40
Média	33	36
Variância	15	20

1. (1 valor) Será que a memória das crianças é diferente com som vs. sem som? Apresente uma resposta rigorosa e completa do ponto de vista estatístico, sem esquecer de mencionar as hipóteses a serem testadas.

$$H_0: \mu_{CS} = \mu_{SS}$$

$$H_1: \mu_{CS} \neq \mu_{SS}$$

$$Z = \frac{\bar{X}_{CS} - \bar{X}_{SS}}{\sqrt{\frac{s_{CS}^2}{n_{CS}} + \frac{s_{SS}^2}{n_{SS}}}}$$

$$\sim N(0,1)$$

$$\alpha = 5\%$$

SE:

- $H_0$  VERDADEIRA

- $n_{CS}, n_{SS}$  GRANDES

$$Z = \frac{33 - 36}{\sqrt{\frac{15}{30} + \frac{20}{40}}} = \frac{-3}{\sqrt{\frac{1}{2} + \frac{1}{2}}} = -3$$

$$-z_{2,5\%} = -1,96 \text{ E } z_{2,5\%} = 1,96$$

$$Z < -1,96 \Rightarrow \text{REJ. } H_0$$

CONCLUI-SE QUE O SOM TEM IMPACTO NA MEMÓRIA.

2. (0,7 valores) Num outro estudo realizado nos mesmos moldes do anterior mas com crianças de 10 anos, a única informação disponível acerca dos resultados é que o *p-value* foi de 0,004. Utilizando apenas essa informação é possível concluir alguma coisa sobre a memória das crianças de 10 anos? Se sim, o que é que se pode concluir e porquê? Se não, explique a razão e que informação está em falta.

*SIM, PARA UM NÍVEL DE SIGNIF. DE 5%  
TEMOS P-VALUE < 5%, PELD QUE  
SE REJEITA H<sub>0</sub> TAL COMO NO  
CASO DOS 6 ANOS. LOGO O SOM  
AFFECTA A MEMÓRIA. NO ENTANTO  
FICA SEM SE SABER QUAL O SINAL  
DA DIFERENÇA.*

#### Grupo V (1 valor)

Tendo em vista uma utilização mais racional dos transportes na área metropolitana de Autopolis, foi levada a cabo durante o ano de 2010 uma grande reforma no sistema de transportes que envolveu alterações dos preços e dos percursos dos transportes públicos, dos lugares de estacionamento automóvel dentro e fora da cidade, das vias prioritárias para autocarros, entre outras. Um dos objetivos era que no final do ano de 2010 o principal meio de transporte que cada pessoa utilizava para se deslocar para o seu emprego se distribuisse de acordo com as seguintes percentagens: 40% de automóvel, 20% de comboio, 20% de autocarro, 10% de bicicleta, e 10% a pé. Para verificar se essas percentagens foram alcançadas, recolheu-se uma amostra aleatória de 200 pessoas tendo sido obtidas as seguintes percentagens: 50% de automóvel, 10% de comboio, 30% de autocarro, 5% de bicicleta, e 5% a pé. Os resultados desta amostra põem em causa as percentagens definidas no referido objetivo? Utilize um nível de significância de 5% para realizar o teste estatístico.

	O	E	$(O-E)^2$	$(O-E)^2/E$
AUTOM	100	80	400	5
CDMB	20	40	400	10
AUTOC	60	40	400	10
BIC	10	20	100	5
PÉ	10	20	100	5
TOTAL	200	200		$\chi^2 = 35$

$$\Rightarrow \alpha = 5\%$$

$$\rightarrow \text{VALOR CRÍTICO A } 5\% = \chi^2_{(5-1) = (4)} = 9,49$$

$$\rightarrow \chi^2 = 35 > 9,49 \Rightarrow$$

$\Rightarrow$  REJEITAR  $H_0$  DE QUE

A DISTRIBUIÇÃO NA POPULAÇÃO

É IGUAL À ESPERADA DE

ACORDO COM O OBJETIVO

**Grupo VI (1 valor)**

O designers de uma agência de publicidade propuseram uma nova embalagem para um conhecido sumo de laranja (vamos designar a embalagem atual por A, e a embalagem nova por B). Para decidir qual das duas embalagens é preferida pelos consumidores vai ser realizada uma experiência em que cada consumidor vai dar notas de 0 a 100 às embalagens.

Pretende-se estimar qual seria a diferença média na população das notas dadas às duas embalagens.

Existem duas formas de implementar esta experiência. A primeira forma consiste em recolher de forma independente duas amostras aleatórias de  $n$  consumidores cada (no total serão necessários  $2 \times n$  consumidores) e pedir aos consumidores de uma das amostras que dê uma nota à embalagem A e aos consumidores da outra amostra que dê uma nota à embalagem B. O estimador natural neste caso é a diferença das duas médias amostrais.

Outra forma de implementar a experiência consiste em recolher uma amostra aleatória de  $n$  consumidores e pedir a cada um deles que dê uma nota à embalagem A e uma nota à embalagem B. O estimador natural neste caso é a média amostral das  $n$  diferenças de notas, de A vs. B, dadas por cada consumidor.

Pode-se assumir que para um mesmo consumidor, as notas dadas às duas embalagens (A vs. B) não são independentes entre si. Isto porque em geral quem dá notas mais altas tende a dar notas altas tanto a A como a B, e quem em geral dá notas mais baixas tende a dar notas baixas tanto a A como a B.

Qual dos dois estimadores descritos acima é preferível em termos de: i) valor esperado e ii) variância? Justifique apresentando os cálculos necessários.

$$X_{Ai} = \text{NOTA DADA A } \textcircled{A} \text{ PELO CONSUMIDOR } i$$

$$X_{Bi} = " " " " \textcircled{B} " "$$

$$\Rightarrow \text{COV}(X_{Ai}, X_{Bi}) > 0$$

$$\begin{aligned} \text{PRETENDE-SE ESTIMAR } & E(X_{Bi}) - E(X_{Ai}) \\ & \equiv M_B - M_A \end{aligned}$$

1<sup>A</sup> FORMAAMOSTRAS ALEATÓRIAS INDEPENDENTES

$$X_{A1}, \dots, X_{Am}$$

$$X_{B1}, \dots, X_{Bm}$$

ESTIMADOR:  $\bar{X}_B - \bar{X}_A$ 

$$\begin{aligned} \rightarrow E(\bar{X}_B - \bar{X}_A) &= E(X_{Bi}) - E(X_{Ai}) \\ &= \mu_B - \mu_A \end{aligned}$$

$$\begin{aligned} \rightarrow V(\bar{X}_B - \bar{X}_A) &= V(\bar{X}_B) + V(\bar{X}_A) + 2 \times 0 \\ &= \frac{V(X_{Bi})}{m} + \frac{V(X_{Ai})}{m} \end{aligned}$$

2<sup>A</sup> FORMA:

AMOSTRA ALEATÓRIA:

$$Y_1 = X_{B1} - X_{A1}, Y_2 = X_{B2} - X_{A1}, \dots, Y_m = X_{Bm} - X_{Am}$$

ESTIMADOR:  $\bar{Y}$ 

$$\rightarrow E(\bar{Y}) = E(Y_i) = E(X_{Bi} - X_{Ai}) = \mu_B - \mu_A$$

$$\begin{aligned} \rightarrow V(\bar{Y}) &= \frac{V(Y_i)}{m} = \frac{V(X_{Bi} - X_{Ai})}{m} \\ &= \frac{V(X_{Bi})}{m} + \frac{V(X_{Ai})}{m} - 2 \frac{\text{COV}(X_{Bi}, X_{Ai})}{m} \end{aligned}$$

$$< V(\bar{X}_B - \bar{X}_A)$$

(O)

i) AMBAS AS FORMAS DÃO ESTIMADORES CENTRADOS.

ii) 2<sup>A</sup> FORMA É PREFERÍVEL EM TERMOS DA VARIÂNCIA.

**SOME USEFUL FORMULAS****Difference Between Population Means or Proportions (Independent Samples)**

Parameter	Assumption	Confidence Interval Endpoints
$\mu_X - \mu_Y$	$N(\mu_X, \sigma_X^2)$	$\bar{x} - \bar{y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_X^2}{n_x} + \frac{\sigma_Y^2}{n_y}}$
	$N(\mu_Y, \sigma_Y^2)$	
	$\sigma_X^2, \sigma_Y^2$ known	
$\mu_X - \mu_Y$	$N(\mu_X, \sigma_X^2)$	$\bar{x} - \bar{y} \pm t_{n_x+n_y-2, \alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$
	$N(\mu_Y, \sigma_Y^2)$	
	$\sigma_X^2 = \sigma_Y^2$ unknown	$s_p^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x + n_y - 2}$
$\mu_X - \mu_Y$	$N(\mu_X, \sigma_X^2)$	$\bar{x} - \bar{y} \pm t_{v, \alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$
	$N(\mu_Y, \sigma_Y^2)$	
	$\sigma_X^2, \sigma_Y^2$ unknown	$v = \left[ \frac{s_x^2}{n_x} + \frac{s_y^2}{n_y} \right]^2 / \left[ \left( \frac{s_x^2}{n_x} \right)^2 / (n_x-1) + \left( \frac{s_y^2}{n_y} \right)^2 / (n_y-1) \right]$
$\mu_X - \mu_Y$	Large samples $\sigma_X^2, \sigma_Y^2$ unknown	$\bar{x} - \bar{y} \pm z_{\alpha/2} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$
$p_X - p_Y$	Large samples Bernoulli $p_X$ Bernoulli $p_Y$	$\hat{p}_x - \hat{p}_y \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_x(1-\hat{p}_x)}{n_x} + \frac{\hat{p}_y(1-\hat{p}_y)}{n_y}}$

Note: The first two intervals in the table are exact. The other three intervals are approximations.

**Multiple Linear Regression:**  $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$

Total S.Sq.=SST=  $\sum (y_i - \bar{y})^2$ , Regression S.Sq.=SSR=  $\sum (\hat{y}_i - \bar{y})^2$ , Error S.Sq.=SSE=  $\sum (y_i - \hat{y}_i)^2$

$$R^2 = \text{SSR/SST} \text{ and } \bar{R}^2 = 1 - \frac{\text{SSE}/(n-K-1)}{\text{SST}/(n-1)}$$

$$\text{Var}(\varepsilon_i) = \sigma^2 \text{ is estimated as } s_e^2 = \sum_{i=1}^n e_i^2 / (n-k-1)$$

$$\text{Confidence interval for } \beta_j: b_j \pm t_{n-k-1, \alpha/2} s_{b_j}$$

$$\text{Test for } H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0 \text{ is } F = \frac{\text{SSR}/k}{\text{SSE}/(n-k-1)} \sim F_{k, n-k-1} \text{ under } H_0$$

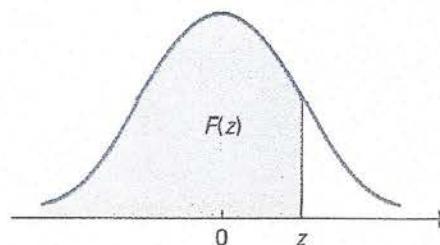
**Simple Linear Regression:**  $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$

$$b_0 = \bar{y} - b_1 \bar{x} \text{ and } b_1 = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) / \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{Var}(b_1) = \sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2 \text{ is estimated as } s_{b_1}^2 = s_e^2 / \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{Prediction interval for } y_{n+1}: b_0 + b_1 x_{n+1} \pm t_{n-2, \alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

**Chi-square Tests:**  $\chi^2 = \sum_{i=1}^k \frac{(Obs_i - Exp_i)^2}{Exp_i} \sim \chi^2_{(k-1)}$  and  $\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(Obs_{ij} - Exp_{ij})^2}{Exp_{ij}} \sim \chi^2_{((r-1)(c-1))}$

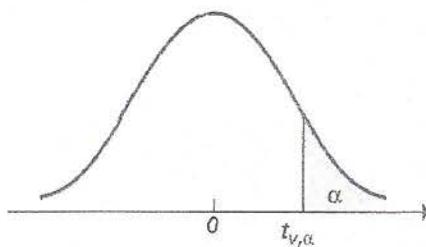
**Table 1** Cumulative Distribution Function of the Standard Normal Distribution

$z$	$F(z)$	$z$	$F(z)$	$z$	$F(z)$	$z$	$F(z)$	$z$	$F(z)$
.00	.5000								
.01	.5040	.31	.6217	.61	.7291	.91	.8186	1.21	.8869
.02	.5080	.32	.6255	.62	.7324	.92	.8212	1.22	.8888
.03	.5120	.33	.6293	.63	.7357	.93	.8238	1.23	.8907
.04	.5160	.34	.6331	.64	.7389	.94	.8264	1.24	.8925
.05	.5199	.35	.6368	.65	.7422	.95	.8289	1.25	.8944
.06	.5239	.36	.6406	.66	.7454	.96	.8315	1.26	.8962
.07	.5279	.37	.6443	.67	.7486	.97	.8340	1.27	.8980
.08	.5319	.38	.6480	.68	.7517	.98	.8365	1.28	.8997
.09	.5359	.39	.6517	.69	.7549	.99	.8389	1.29	.9015
.10	.5398	.40	.6554	.70	.7580	1.00	.8413	1.30	.9032
.11	.5438	.41	.6591	.71	.7611	1.01	.8438	1.31	.9049
.12	.5478	.42	.6628	.72	.7642	1.02	.8461	1.32	.9066
.13	.5517	.43	.6664	.73	.7673	1.03	.8485	1.33	.9082
.14	.5557	.44	.6700	.74	.7704	1.04	.8508	1.34	.9099
.15	.5596	.45	.6736	.75	.7734	1.05	.8531	1.35	.9115
.16	.5636	.46	.6772	.76	.7764	1.06	.8554	1.36	.9131
.17	.5675	.47	.6803	.77	.7794	1.07	.8577	1.37	.9147
.18	.5714	.48	.6844	.78	.7823	1.08	.8599	1.38	.9162
.19	.5753	.49	.6879	.79	.7852	1.09	.8621	1.39	.9177
.20	.5793	.50	.6915	.80	.7881	1.10	.8643	1.40	.9192
.21	.5832	.51	.6950	.81	.7910	1.11	.8665	1.41	.9207
.22	.5871	.52	.6985	.82	.7939	1.12	.8686	1.42	.9222
.23	.5910	.53	.7019	.83	.7967	1.13	.8708	1.43	.9236
.24	.5948	.54	.7054	.84	.7995	1.14	.8729	1.44	.9251
.25	.5987	.55	.7088	.85	.8023	1.15	.8749	1.45	.9265
.26	.6026	.56	.7123	.86	.8051	1.16	.8770	1.46	.9279
.27	.6064	.57	.7157	.87	.8078	1.17	.8790	1.47	.9292
.28	.6103	.58	.7190	.88	.8106	1.18	.8810	1.48	.9306
.29	.6141	.59	.7224	.89	.8133	1.19	.8830	1.49	.9319
.30	.6179	.60	.7257	.90	.8159	1.20	.8849	1.50	.9332
									.9641

**Table 1** Cumulative Distribution Function of the Standard Normal Distribution Continued

$Z$	$F(z)$										
1.81	.9649	2.21	.9864	2.61	.9955	3.01	.9987	3.41	.9997	3.81	.9999
1.82	.9656	2.22	.9868	2.62	.9956	3.02	.9987	3.42	.9997	3.82	.9999
1.83	.9664	2.23	.9871	2.63	.9957	3.03	.9988	3.43	.9997	3.83	.9999
1.84	.9671	2.24	.9875	2.64	.9959	3.04	.9988	3.44	.9997	3.84	.9999
1.85	.9678	2.25	.9878	2.65	.9960	3.05	.9989	3.45	.9997	3.85	.9999
1.86	.9686	2.26	.9881	2.66	.9961	3.06	.9989	3.46	.9997	3.86	.9999
1.87	.9693	2.27	.9884	2.67	.9962	3.07	.9989	3.47	.9997	3.87	.9999
1.88	.9699	2.28	.9887	2.68	.9963	3.08	.9990	3.48	.9997	3.88	.9999
1.89	.9706	2.29	.9890	2.69	.9964	3.09	.9990	3.49	.9998	3.89	1.0000
1.90	.9713	2.30	.9893	2.70	.9965	3.10	.9990	3.50	.9998	3.90	1.0000
1.91	.9719	2.31	.9896	2.71	.9966	3.11	.9991	3.51	.9998	3.91	1.0000
1.92	.9726	2.32	.9898	2.72	.9967	3.12	.9991	3.52	.9998	3.92	1.0000
1.93	.9732	2.33	.9901	2.73	.9968	3.13	.9991	3.53	.9998	3.93	1.0000
1.94	.9738	2.34	.9904	2.74	.9969	3.14	.9992	3.54	.9998	3.94	1.0000
1.95	.9744	2.35	.9906	2.75	.9970	3.15	.9992	3.55	.9998	3.95	1.0000
1.96	.9750	2.36	.9909	2.76	.9971	3.16	.9992	3.56	.9998	3.96	1.0000
1.97	.9756	2.37	.9911	2.77	.9972	3.17	.9992	3.57	.9998	3.97	1.0000
1.98	.9761	2.38	.9913	2.78	.9973	3.18	.9993	3.58	.9998	3.98	1.0000
1.99	.9767	2.39	.9916	2.79	.9974	3.19	.9993	3.59	.9998	3.99	1.0000
2.00	.9772	2.40	.9918	2.80	.9974	3.20	.9993	3.60	.9998		
2.01	.9778	2.41	.9920	2.81	.9975	3.21	.9993	3.61	.9998		
2.02	.9783	2.42	.9922	2.82	.9976	3.22	.9994	3.62	.9999		
2.03	.9788	2.43	.9925	2.83	.9977	3.23	.9994	3.63	.9999		
2.04	.9793	2.44	.9927	2.84	.9977	3.24	.9994	3.64	.9999		
2.05	.9798	2.45	.9929	2.85	.9978	3.25	.9994	3.65	.9999		
2.06	.9803	2.46	.9931	2.86	.9979	3.26	.9994	3.66	.9999		
2.07	.9808	2.47	.9932	2.87	.9979	3.27	.9995	3.67	.9999		
2.08	.9812	2.48	.9934	2.88	.9980	3.28	.9995	3.68	.9999		
2.09	.9817	2.49	.9936	2.89	.9981	3.29	.9995	3.69	.9999		
2.10	.9821	2.50	.9938	2.90	.9981	3.30	.9995	3.70	.9999		
2.11	.9826	2.51	.9940	2.91	.9982	3.31	.9995	3.71	.9999		
2.12	.9830	2.52	.9941	2.92	.9982	3.32	.9996	3.72	.9999		
2.13	.9834	2.53	.9943	2.93	.9983	3.33	.9996	3.73	.9999		
2.14	.9838	2.54	.9945	2.94	.9984	3.34	.9996	3.74	.9999		
2.15	.9842	2.55	.9946	2.95	.9984	3.35	.9996	3.75	.9999		
2.16	.9846	2.56	.9948	2.96	.9985	3.36	.9996	3.76	.9999		
2.17	.9850	2.57	.9949	2.97	.9985	3.37	.9996	3.77	.9999		
2.18	.9854	2.58	.9951	2.98	.9986	3.38	.9996	3.78	.9999		
2.19	.9857	2.59	.9952	2.99	.9986	3.39	.9997	3.79	.9999		
2.20	.9861	2.60	.9953	3.00	.9986	3.40	.9997	3.80	.9999		

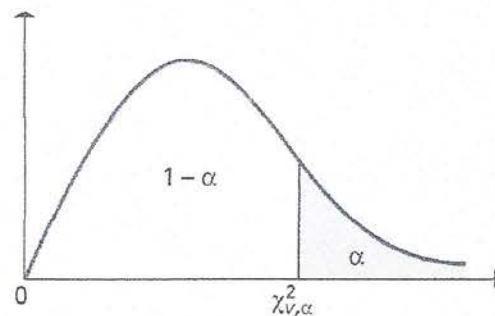
Reproduced with permission of the trustees of Biometrika, from *Biometrika Tables for Statisticians*, vol. 1 (1966).

**Table 8** Cutoff Points for the Student's  $t$  Distribution

For selected probabilities,  $\alpha$ , the table shows the values  $t_{v,\alpha}$  such that  $P(t_v > t_{v,\alpha}) = \alpha$ , where  $t_v$  is a Student's  $t$  random variable with  $v$  degrees of freedom. For example, the probability is .10 that a Student's  $t$  random variable with 10 degrees of freedom exceeds 1.372.

v	$\alpha$				
	0.100	0.050	0.025	0.010	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
$\infty$	1.282	1.645	1.960	2.326	2.576

Reproduced with permission of the trustees of Biometrika, from *Biometrika Tables for Statisticians*, vol. 1 (1966).

**Table 7** Cutoff Points of the Chi-Square Distribution Function

For selected probabilities  $\alpha$ , the table shows the values  $\chi^2_{v,\alpha}$  such that  $P(\chi^2_v > \chi^2_{v,\alpha}) = \alpha$ , where  $\chi^2_v$  is a chi-square random variable with  $v$  degrees of freedom. For example, the probability is .100 that a chi-square random variable with 10 degrees of freedom is greater than 15.99.

v	α									
	.995	.990	.975	.950	.900	.100	.050	.025	.010	.005
1	0.04393	0.03157	0.03982	0.02393	0.0158	2.71	3.84	5.02	6.63	7.88
2	0.0100	0.0201	0.0506	0.103	0.211	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.064	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.145	1.61	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.4	104.2
80	51.17	53.54	57.15	60.39	64.28	96.58	101.9	106.6	112.3	116.3
90	59.20	61.75	65.65	69.13	73.29	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	118.5	124.3	129.6	135.8	140.2

Reproduced with permission from C. M. Thompson, "Tables of percentage points of the chi-square distribution," *Biometrika* 32 (1941).