Disciplina: MAE0212: Introdução a Probabilidade e Estatística II
Professora: Silvia Ferrari
Nome: **N.USP**: **Nota**:

Use os espaços em branco, inclusive o verso das páginas, para responder às questões. Mostre os desenvolvimentos, não apenas as respostas. **ATENÇÃO: Identifique quais suposições são necessárias para a resolução de cada item**.

Provinha 4 - Gabarito

Para resolver as questões, você pode usar as informações abaixo. Se não encontrar exatamente o que precisa, pode dar o melhor resultado aproximado possível usando apenas o que está fornecido aqui. Lembre de justificar sua resposta!

1. Se $Z \sim N(0,1)$ então,

$\mathbb{P}(Z \leq 3.03) = 0.999$	$\mathbb{P}(Z \leq 1.96) = 0.975$	$\mathbb{P}(Z \leq 1.64) = 0.95$	$\mathbb{P}(Z \leq 2.60) = 0.995$
$\mathbb{P}(Z \leq 1.28) = 0.90$	$\mathbb{P}(Z \leq 2.33) = 0.99$	$\mathbb{P}(Z \leq 1.67) = 0.953$	$\mathbb{P}(Z \leq 1.47) = 0.93$

2. Se $T_\nu \sim t_{(\nu)}$, então

$\mathbb{P}(T_8 \leq 1.39) = 0.90$	$\mathbb{P}(T_8 \leq 1.86) = 0.95$	$\mathbb{P}(T_8 \leq 2.31) = 0.975$	$\mathbb{P}(T_8 \leq 2.89) = 0.99$	$\mathbb{P}(T_8 \leq 3.35) = 0.995$
$\mathbb{P}(T_9 \leq 1.38) = 0.90$	$\mathbb{P}(T_9 \leq 1.83) = 0.95$	$\mathbb{P}(T_9 \leq 2.26) = 0.975$	$\mathbb{P}(T_9 \leq 2.82) = 0.99$	$\mathbb{P}(T_9 \leq 3.25) = 0.995$
$\mathbb{P}(T_{10} \leq 1.37) = 0.90$	$\mathbb{P}(T_{10} \leq 1.81) = 0.95$	$\mathbb{P}(T_{10} \leq 2.23) = 0.975$	$\mathbb{P}(T_{10} \leq 2.76) = 0.99$	$\mathbb{P}(T_{10} \leq 3.17) = 0.995$
$\mathbb{P}(T_{11} \leq 1.36) = 0.90$	$\mathbb{P}(T_{11} \leq 1.79) = 0.95$	$\mathbb{P}(T_{11} \leq 2.20) = 0.975$	$\mathbb{P}(T_{11} \leq 2.71) = 0.99$	$\mathbb{P}(T_{11} \leq 3.10) = 0.995$
$\mathbb{P}(T_{35} \leq 1.31) = 0.90$	$\mathbb{P}(T_{35} \leq 1.69) = 0.95$	$\mathbb{P}(T_{35} \leq 2.03) = 0.975$	$\mathbb{P}(T_{35} \leq 2.44) = 0.99$	$\mathbb{P}(T_{35} \leq 2.72) = 0.995$
$\mathbb{P}(T_{38} \leq 1.30) = 0.90$	$\mathbb{P}(T_{38} \leq 1.68) = 0.95$	$\mathbb{P}(T_{38} \leq 2.02) = 0.975$	$\mathbb{P}(T_{38} \leq 2.43) = 0.99$	$\mathbb{P}(T_{38} \leq 2.71) = 0.995$
$\mathbb{P}(T_{39} \leq 1.30) = 0.90$	$\mathbb{P}(T_{39} \leq 1.68) = 0.95$	$\mathbb{P}(T_{39} \leq 2.02) = 0.975$	$\mathbb{P}(T_{39} \leq 2.42) = 0.99$	$\mathbb{P}(T_{39} \leq 2.70) = 0.995$
...				
$\mathbb{P}(T_{200} \leq 1.28) = 0.90$	$\mathbb{P}(T_{200} \leq 1.65) = 0.95$	$\mathbb{P}(T_{200} \leq 1.97) = 0.975$	$\mathbb{P}(T_{200} \leq 2.34) = 0.99$	$\mathbb{P}(T_{200} \leq 2.60) = 0.995$

3. Se $\chi^2_\nu \sim \chi^2_{(\nu)}$, então

$\mathbb{P}(\chi^2_1 \leq 2.70) = 0.90$	$\mathbb{P}(\chi^2_1 \leq 3.84) = 0.95$	$\mathbb{P}(\chi^2_1 \leq 5.02) = 0.975$	$\mathbb{P}(\chi^2_1 \leq 6.63) = 0.99$	$\mathbb{P}(\chi^2_1 \leq 7.88) = 0.995$
$\mathbb{P}(\chi^2_4 \leq 7.78) = 0.90$	$\mathbb{P}(\chi^2_4 \leq 9.49) = 0.95$	$\mathbb{P}(\chi^2_4 \leq 11.14) = 0.975$	$\mathbb{P}(\chi^2_4 \leq 13.28) = 0.99$	$\mathbb{P}(\chi^2_4 \leq 14.86) = 0.995$
$\mathbb{P}(\chi^2_9 \leq 14.68) = 0.90$	$\mathbb{P}(\chi^2_9 \leq 16.92) = 0.95$	$\mathbb{P}(\chi^2_9 \leq 19.02) = 0.975$	$\mathbb{P}(\chi^2_9 \leq 21.66) = 0.99$	$\mathbb{P}(\chi^2_9 \leq 23.59) = 0.995$
$\mathbb{P}(\chi^2_{10} \leq 15.99) = 0.90$	$\mathbb{P}(\chi^2_{10} \leq 18.31) = 0.95$	$\mathbb{P}(\chi^2_{10} \leq 20.48) = 0.975$	$\mathbb{P}(\chi^2_{10} \leq 23.21) = 0.99$	$\mathbb{P}(\chi^2_{10} \leq 25.19) = 0.995$
$\mathbb{P}(\chi^2_{19} \leq 27.20) = 0.90$	$\mathbb{P}(\chi^2_{19} \leq 30.14) = 0.95$	$\mathbb{P}(\chi^2_{19} \leq 32.85) = 0.975$	$\mathbb{P}(\chi^2_{19} \leq 36.19) = 0.99$	$\mathbb{P}(\chi^2_{19} \leq 38.58) = 0.995$
$\mathbb{P}(\chi^2_{20} \leq 28.41) = 0.90$	$\mathbb{P}(\chi^2_{20} \leq 31.41) = 0.95$	$\mathbb{P}(\chi^2_{20} \leq 34.17) = 0.975$	$\mathbb{P}(\chi^2_{20} \leq 37.57) = 0.99$	$\mathbb{P}(\chi^2_{20} \leq 40.00) = 0.995$
$\mathbb{P}(\chi^2_{35} \leq 46.06) = 0.90$	$\mathbb{P}(\chi^2_{35} \leq 49.80) = 0.95$	$\mathbb{P}(\chi^2_{35} \leq 53.20) = 0.975$	$\mathbb{P}(\chi^2_{35} \leq 57.34) = 0.99$	$\mathbb{P}(\chi^2_{35} \leq 60.27) = 0.995$
$\mathbb{P}(\chi^2_{36} \leq 47.21) = 0.90$	$\mathbb{P}(\chi^2_{36} \leq 51.00) = 0.95$	$\mathbb{P}(\chi^2_{36} \leq 54.44) = 0.975$	$\mathbb{P}(\chi^2_{36} \leq 58.62) = 0.99$	$\mathbb{P}(\chi^2_{36} \leq 61.58) = 0.995$

4. Se $W_{p,q} \sim F_{p,q}$, então

$\mathbb{P}(W_{9,9} \leq 2.44) = 0.90$	$\mathbb{P}(W_{9,9} \leq 3.18) = 0.95$	$\mathbb{P}(W_{9,9} \leq 4.02) = 0.975$	$\mathbb{P}(W_{9,9} \leq 5.35) = 0.99$	$\mathbb{P}(W_{9,9} \leq 6.54) = 0.995$
$\mathbb{P}(W_{10,15} \leq 2.06) = 0.90$	$\mathbb{P}(W_{10,15} \leq 2.54) = 0.95$	$\mathbb{P}(W_{10,15} \leq 3.06) = 0.975$	$\mathbb{P}(W_{10,15} \leq 3.80) = 0.99$	$\mathbb{P}(W_{10,15} \leq 4.42) = 0.995$
$\mathbb{P}(W_{15,10} \leq 2.24) = 0.90$	$\mathbb{P}(W_{15,10} \leq 2.86) = 0.95$	$\mathbb{P}(W_{15,10} \leq 3.52) = 0.975$	$\mathbb{P}(W_{15,10} \leq 4.56) = 0.99$	$\mathbb{P}(W_{15,10} \leq 5.47) = 0.995$
$\mathbb{P}(W_{19,19} \leq 1.82) = 0.90$	$\mathbb{P}(W_{19,19} \leq 2.17) = 0.95$	$\mathbb{P}(W_{19,19} \leq 2.53) = 0.975$	$\mathbb{P}(W_{19,19} \leq 3.03) = 0.99$	$\mathbb{P}(W_{19,19} \leq 3.43) = 0.995$
$\mathbb{P}(W_{20,20} \leq 1.79) = 0.90$	$\mathbb{P}(W_{20,20} \leq 2.12) = 0.95$	$\mathbb{P}(W_{20,20} \leq 2.46) = 0.975$	$\mathbb{P}(W_{20,20} \leq 2.94) = 0.99$	$\mathbb{P}(W_{20,20} \leq 3.32) = 0.995$

Figura 1: Algumas informações sobre probabilidades envolvendo as distribuições normal, t-Student, χ^2 e F.

1. Um estudo foi conduzido numa grande empresa para comparar dois métodos de treinamento de atendentes de *call center*. Quarenta funcionários recém contratados foram divididos aleatoriamente em dois grupos de 20 pessoas cada. Um grupo recebeu treinamento pelo método A e o outro, pelo método B. Após o término do treinamento, todos os funcionários passaram por um mesmo exame e suas notas (de 0 a 100) foram registradas. Algumas medidas-resumo são dadas na tabela a seguir.

	Média	Desvio-padrão
Grupo A	52.5	3.5
Grupo B	62.3	2.9

a) (1.0 ponto) Esses dados indicam evidência suficiente para afirmar que há diferença entre os métodos de treinamento quanto às suas variabilidades de notas? Considere α = 5%.

b) (1 ponto) Esses dados indicam evidência suficiente para afirmar que há diferença entre os métodos de treinamento quanto às suas notas médias? Considere α = 5%.

c) (0.5 ponto) Construa um intervalo de 95% de confiança para a diferença entre as notas médias sob os dois treinamentos.

Solução:

Considere as variáveis aleatórias X_A, que representa a nota de um funcionário submetido a treinamento sob o método A, e X_B, que representa a nota de um funcionário submetido a treinamento sob o método B.

Suposições:

- $X_A \sim N(\mu_A, \sigma_A^2)$ e $X_B \sim N(\mu_B, \sigma_B^2)$;
- $X_{A_1}, \ldots, X_{A_{20}}$ é uma amostra aleatória simples de X_A e $X_{B_1}, \ldots, X_{B_{20}}$ é uma amostra aleatória simples de X_B.
- As amostras são independentes.

Observe que estamos no contexto de populações normais com amostras independentes.

a) *Passo 1*: Note que o interesse é verificar se há diferença entre os métodos de treinamento quanto às suas **variabilidades de notas**. Dessa forma, estamos interessados nas seguintes hipóteses:

$$\mathcal{H}_0 : \sigma_A^2 = \sigma_B^2 \quad vs \quad \mathcal{H}_1 : \sigma_A^2 \neq \sigma_B^2.$$

Passo 2: Neste caso,

$$\frac{S_A^2/\sigma_A^2}{S_B^2/\sigma_B^2} \sim F_{(20-1,20-1)}.$$

Se $\sigma_A^2 = \sigma_B^2$, temos que $W = S_A^2/S_B^2 \sim F_{(19,19)}$.

Passo 3: Fixe $\alpha = 5\%$ e queremos encontrar uma região crítica da forma $RC = (0, f_1) \cup (f_2, \infty)$, em que f_1 e f_2 são tais que $\mathbb{P}(\text{erro I}) = 0.05$. Dessa forma, temos que

$$\alpha = \mathbb{P}(\text{rejeitar } \mathcal{H}_0 \mid \sigma_A^2 = \sigma_B^2) \Rightarrow \mathbb{P}(W < f_1 \text{ ou } W > f_2) = 0.05.$$

Os valores f_1 e f_2 são tais que $\mathbb{P}(W < f_1) = \alpha/2 = 0.025 = \mathbb{P}(W > f_2)$, em que $W \sim F_{(19,19)}$. Pelas informações da tabela da página 1, temos $\mathbb{P}(W > 2.53) = 0.025$. Temos também que $\mathbb{P}(1/W < 1/2.53) = 0.025$. Lembrando que, se $W \sim F(a, b)$, então, $1/W \sim F(b, a)$ e que, neste caso, $a = b = 19$, temos que W e $1/W$ têm a mesma distribuição. Logo, $\mathbb{P}(1/W < 1/2.53) = \mathbb{P}(W < 1/2.53) = \mathbb{P}(W < 0.40) = 0.025$.

Assim, a região crítica é

$$RC = (0, 0.40) \cup (2.53, \infty).$$

Passo 4: Considerando os dados da amostra, temos $w_0 = 3.5^2/2.9^2 = 12.25/8.41 = 1.46$.

Passo 5: Como o valor w_0 observado não pertence à região crítica, não temos evidências amostrais para rejeitar a hipótese nula, considerando $\alpha = 5\%$. Isto é, não temos evidências para dizer que há diferença entre os métodos de treinamento quanto às suas variabilidades de notas.

b) *Passo 1*: Note que o interese é verificar se há diferença entre os métodos de treinamento quanto às notas médias. Além disso, não rejeitamos a hipótese de igualdade das variâncias (ver item a).), assim, vamos considerar $\sigma_A^2 = \sigma_B^2$. Dessa forma, estamos interessados nas seguintes hipóteses:

$$\mathcal{H}_0 : \mu_A = \mu_B \quad vs \quad \mathcal{H}_1 : \mu_A \neq \mu_B.$$

Passo 2: Neste caso, a estatística de teste é

$$T = \frac{\overline{X}_A - \overline{X}_B}{S_p \sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}},$$

em que

$$S_p^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}.$$

E, sob $\mathcal{H}_0$, temos que $T \sim t_{(38)}$.

Passo 3: Fixe $\alpha = 5\%$ e queremos encontrar uma região crítica da forma $RC = (-\infty, t_1) \cup (t_2, \infty)$, em que t_1 e t_2 são tais que $\mathbb{P}(\text{erro I}) = 0.05$. Dessa forma, temos que

$$\alpha = \mathbb{P}(\text{rejeitar } \mathcal{H}_0 \mid \mu_A = \mu_B) \Rightarrow \mathbb{P}(T < t_1 \text{ ou } T > t_2) = 0.05.$$

em que $T \sim t_{(38)}$ e, pela simetria da distribuição t-Student, tomamos $t_1 = -t_2$ com t_2 tal que $\mathbb{P}(T < t_2) = 0.975 \Rightarrow t_2 = 2.02$. Assim, a região crítica é

$$RC = (-\infty, -2.02) \cup (2.02, \infty).$$

Passo 4: Considerando os dados da amostra, temos

$$t_{obs} = \frac{52.5 - 62.3}{3.21\sqrt{\dfrac{1}{20} + \dfrac{1}{20}}} = \frac{-9.8}{3.21/\sqrt{10}} = -9.65.$$

Passo 5: Como o valor t_{obs} pertence à região crítica, temos evidências amostrais para rejeitar a hipótese nula, considerando $\alpha = 5\%$. Isto é, temos evidências para dizer que há diferença entre os métodos de treinamento quanto às notas médias.

c) Seja $\Delta = \mu_A - \mu_B$, isto é, a diferença entre as notas médias sob os dois treinamentos. O intervalo de confiança para Δ com coeficiente de confiança γ é dado por

$$\text{IC}(\Delta;\gamma) = \left[(\overline{X}_A - \overline{X}_B) - t_\gamma S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}, (\overline{X}_A - \overline{X}_B) + t_\gamma S_p\sqrt{\frac{1}{n_A} + \frac{1}{n_B}}\right].$$

Assim, o intervalo de confiança observado para a diferença de médias, considerando $\gamma = 95\%$, é dado por

$$\text{IC}(\Delta; 95\%) = \left[-9.8 - 2.02\frac{3.21}{\sqrt{10}}, -9.8 + 2.02\frac{3.21}{\sqrt{10}}\right] = [-11.85, -7.75].$$

2. Um pesquisador conduziu um experimento com crianças gêmeas para avaliar o tempo de absorção de um medicamento. Sua conjectura é que o medicamento B, que tem um componente adicional relativamente ao medicamento A, tem um tempo médio de absorção menor. A tabela a seguir apresenta os tempos de absorção (em segundos) para para cada uma de 10 pares de gêmeos.

Par de gêmeos (i)	Medicamento A (x_i)	Medicamento B (y_i)
1	50	42
2	61	54
3	58	61
4	68	64
5	78	70
6	45	40
7	56	59
8	73	69
9	64	58
10	51	48
Total	604	565

a) (1,5 ponto) Teste a conjectura do pesquisador através de um teste de hipóteses adequado. Construa a região crítica e dê sua conclusão tomando o nível de significância de 5%.

b) (1 ponto) Qual é o valor-p do teste?

Se necessário, use $\displaystyle\sum_{i=1}^{10}(x_i - y_i)^2 = 297$.

Solução:

Considere a variável aleatória X, que representa o tempo de absorção (em segundos) do medicamento A e Y, que representa o tempo de absorção (em segundos) do medicamento B. Seja $(X_1, Y_1), \ldots, (X_{10}, Y_{10})$ uma amostra aleatória dos tempos de absorção dos medicamentos A e B.

Suposições:

- Não há dependência entre os pares de gêmeos com relação ao tempo de absorção dos medicamentos A e B;
- A variável aleatória $D = X - Y \sim N(\mu_D, \sigma_D^2)$.

Observe que temos duas medidas para um mesmo par de gêmeos $\Rightarrow$ amostras dependentes.

a) *Passo 1*: Note que o interese é verificar se o medicamento B tem um tempo médio de absorção menor, ou seja, se $\mu_D > 0$ (esta é a conjectura do pesquisador a ser testada; é o que vai aparecer na hipótese alternativa). Dessa forma, consideremos as seguintes hipóteses:

$$\mathcal{H}_0 : \mu_D \leq 0 \quad vs \quad \mathcal{H}_1 : \mu_D > 0.$$

Passo 2: Considerando uma amostra aleatória simples de 10 indivíduos, temos que

$$\frac{\overline{D} - \mu_D}{S_D/\sqrt{10}} \sim t_{(10-1)},$$

em que $\overline{D} = \sum_{i=1}^{n} D_i/n$, com $D_i = X_i - Y_i$. Sob $\mathcal{H}_0$, temos que

$$T = \frac{\overline{D}}{S_D/\sqrt{10}} \sim t_{(10-1)}.$$

Note que, agora, estamos trabalhando com a amostra $D_1, \ldots, D_{10}$.

Passo 3: Fixe $\alpha = 5\%$ e queremos encontrar uma região crítica da forma $RC = (t_c, \infty)$, em que t_c é tal que $\mathbb{P}(\text{erro I}) = 0.05$. Dessa forma, temos que

$$\alpha = \mathbb{P}(\text{rejeitar } \mathcal{H}_0 \mid \mu_D = 0) \Rightarrow \mathbb{P}(T > t_c) = 0.05 \Rightarrow t_c = 1.83,$$

em que $T \sim t_{(9)}$. Assim, a região crítica é

$$RC = (1.83, \infty).$$

Passo 4: O valor

$$t_{obs} = \frac{3.9}{4.01/\sqrt{10}} = 3.07.$$

Passo 5: Como o valor t_{obs} pertence à região crítica, temos evidências amostrais para rejeitar a hipótese nula, considerando $\alpha = 5\%$. Isto é, temos evidências amostrais em favor da conjectura do pesquisador.

c) Queremos obter o valor-p do teste. Note que, neste caso, o valor-p é definido como

$$\widehat{\alpha} = \mathbb{P}(T > t_{obs}) = \mathbb{P}(T > 3.07) < 0.01.$$

Pelas informações fornecidas na página 1, temos que $\mathbb{P}(T \leq 2.82) = 0.99$. Então $\mathbb{P}(T \leq 3.07) > 0.99 \Rightarrow \mathbb{P}(T > 3.07) < 0.01$. Observe que o valor-p é pequeno, ou seja, há forte evidência amostral para se concluir que o tempo de absorção do medicamento B é menor que o do medicamento A.

3. (2,5 pontos) Um aluno de Computação afirmou ter programado um algoritmo gerador de números aleatórios inteiros no intervalo 0-9. O professor resolveu testar o algoritmo gerando 1000 valores e dispondo na tabela abaixo.

	0	1	2	3	4	5	6	7	8	9
Frequência	89	103	70	100	96	112	109	86	109	126

O professor tem evidência suficiente para afirmar que o algoritmo programado pelo aluno é inadequado? Justifique sua resposta com base em um teste de hipóteses adequado usando um nível de significância de 5%.

Solução: Note que queremos testar se o algoritmo do aluno esta funcionando, e isso pode ser feito através de um teste de aderência. Note que, se o algoritmo funciona corretamente, a probabilidade de ocorrência do valor i deve ser $1/10$, para $i = 0, \ldots, 9$. Dessa forma, temos que

	0	1	2	3	4	5	6	7	8	9
Frequência observada	89	103	70	100	96	112	109	86	109	126
Frequência esperada	100	100	100	100	100	100	100	100	100	100

As hipóteses do teste de aderência são:

$$\mathcal{H}_0 : p_0 = p_1 = \cdots = p_9$$
$$\mathcal{H}_1 : \text{há pelo menos uma diferença nas igualdades em } \mathcal{H}_0,$$

em que p_i denota a probabilidade de um número gerado pelo algoritmo ser i.

A estatística do teste é

$$\chi^2 = \sum_{i=0}^{9} \frac{(O_i - E_i)^2}{E_i},$$

em que O_i é a contagem observada de valores iguais a i e E_i é a correspondente contagem esperada sob a hipótese nula. Como o número de observações é grande e as contagens esperadas não são pequenas, a estatística χ^2 tem distribuição χ^2 com $k-1$ graus de liberdade; k é o número de categorias, ou seja, 10. Valores grandes da estatística fornecem evidência contra $\mathcal{H}_0$. O valor da estatística na amostra observada é

$$\chi^2_{obs} = \frac{(89-100)^2}{100} + \cdots + \frac{(126-100)^2}{100} = 22.24$$

Pela tabela da página 1, temos que $\mathbb{P}(\chi^2_9 > x_c) = 0.05 \Rightarrow \mathbb{P}(\chi^2_9 \leq x_c) = 0.95 \Rightarrow x_c = 16.92$.

Como o valor observado da estatística, χ^2_{obs}, é maior do que 16.92, temos evidências amostrais, considerando $\alpha = 5\%$, para rejeitar a hipótese de que o algoritimo do aluno funciona corretamente.

Nota: o exercício pode também ser resolvido usando o p-valor, que é dado por p-valor $=$ $\mathbb{P}(\chi^2_9 \geq \chi^2_{obs})$.

Pela tabela da página 1, temos

$$\text{p-valor} = P(\chi^2_9 \geq 22.24) < P(\chi^2_9 \geq 21.66) = 0.01 < 0.05,$$

o que indica que $\mathcal{H}_0$ é rejeitada.

4. Pesquisadores de doenças parasitárias em animais de grande porte suspeitam que a incidência de certos parasitas esteja associada à raça do animal, pura ou não-pura. Num estudo para verificar essa suspeita, 950 animais selecionados aleatoriamente de uma grande fazenda foram classificados segundo sua raça e incidência de parasitose (berne). O resultado da pesquisa está na tabela a seguir (Sim representa "com incidência de parasitose" e Não representa "sem incidência de parasitose").

	Sim	Não	Total
Pura	200	600	800
Não-pura	47	103	150
Total	247	703	950

a) (1 ponto) Calcule as proporções segundo os totais das linhas. O que representam estas proporções?

b) (0,5 ponto) Sob independência entre raça e a incidência de parasitas, qual é o número esperado de animais de raça não-pura com incidência de parasitas? Quantos animais foram observados nesse caso?

c) (1 ponto) Formule as hipóteses adequadas para testar se existe dependência entre raça do animal e incidência de parasitas. Calcule o valor-p do teste e tome a decisão adequada para o nível de significância de 5% .

Solução: Inicialmente, note que o total de animais selecionados é fixado. Ou seja, estamos num contexto de teste de independência.

a) A tabela abaixo apresenta as proporções segundo os totais linhas.

	Sim	Não	Total
Pura	0.250	0.750	1
Não-pura	0.313	0.687	1
Total	0.260	0.740	1

Essas proporções representam as proporções **amostrais** de incidência e não incidência de parasitose por raça. A última linha contém as proporções amostrais de incidência e não incidência de parasitose, independentemente de raça.

b) Sob independência entre raça e a incidência de parasitas, o número esperado de animais da raça não-pura com incidência de parasitas é $150 * 247/950 = 39$, ou seja, 39 animais. Neste caso foram observados 47 animais.

c) Queremos testar as hipóteses

$$\mathcal{H}_0 : \text{Raça e incidência de parasitas são variáveis independentes}$$

versus

$$\mathcal{H}_1 : \text{Raça e incidência de parasitas são variáveis dependentes}$$

A tabela a seguir apresenta os valores esperados sob a hipótese $\mathcal{H}_0$.

	Sim	Não	Total
Pura	208	592	800
Não-pura	39	111	150
Total	247	703	950

A estatística do teste é

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}},$$

em que O_{ij} e E_{ij} são, respectivamente, as contagens observadas e esperadas (sob a hipótese nula) na i-ésima linha e j-ésima coluna. Como o número de observações é grande e as contagens esperadas não são pequenas, a estatística χ^2 tem distribuição χ^2 com $(2-1)*(2-1) = 1$ grau de liberdade. Valores grandes da estatística fornecem evidência contra $\mathcal{H}_0$. O valor da estatística na amostra observada é

$$\begin{aligned}\chi^2_{obs} &= \frac{(200-208)^2}{208}+\frac{(600-592)^2}{592}+\frac{(47-39)^2}{39}+\frac{(103-111)^2}{111}\\ &= \frac{64}{208}+\frac{64}{592}+\frac{64}{39}+\frac{64}{111} = 2.63.\end{aligned}$$

Assim, o p-valor do teste é dado por p-valor $= \mathbb{P}(\chi^2_1 > 2.63)$. Pelas informações da tabela da página 1 temos $\mathbb{P}(\chi^2_1 > 2.63) > \mathbb{P}(\chi^2_1 > 2.70) = 0.10$. Como o p-valor do teste é maior que $\alpha = 0.05$, não temos evidências amostrais (para $\alpha = 5\%$) para rejeitar a $\mathcal{H}_0$. Isto é, não temos evidências amostrais suficientes para dizer que há dependência entre raça e a incidência de parasitas.