www.estudar.com.vc

Resumo e Lista de Exercícios

Estatística

Fuja do Nabo P3 2018.2

As tabelas necessárias encontram-se em anexo ao final do documento

Resumo

1. Testes de Aderência e Associação

Em testes de hipóteses comuns, a distribuição da população é conhecida, e são testados parâmetros populacionais, com base na amostra.

Em testes de aderência (ou testes não paramétricos), a distribuição populacional é desconhecida, e o objetivo é identificar se a população segue uma determinada distribuição.

Neste caso, a hipótese nula diz que a população segue uma determinada distribuição P_0, enquanto a hipótese alternativa refuta essa afirmação:

$$\begin{cases} H_0: P = P_0 \\ H_1: P \neq P_0 \end{cases}$$

a. Inspeção Visual

Esse método consiste em comparar visualmente observações amostrais com os valores obtidos pelas equações e curvas da distribuição em questão.

O método é menos preciso, pois não fornece uma análise matemática que rejeite ou afirme as hipóteses.

b. Teste Qui-Quadrado (*Pearson*)

Neste teste, ocorre o cálculo da variável Qui-Quadrado, com a seguinte equação:

$$\chi_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n$$

Onde O_i é a frequência observada de uma determinada classe/categoria, n é o número total de observações, k é o número de classes e E_i é a frequência esperada da categoria i, calculada por:

$$E_i = np_i$$

Onde p_i é a probabilidade de obter determinado valor da classe i (segundo a distribuição testada).

É importante que os valores E_i sejam maiores ou iguais a **5**. As classes que não satisfazem essa condição devem se juntar às classes adjacentes.

A condição de rejeição de H_0 é a seguinte:

$$\chi_0^2 > \chi_{\alpha;k-1-m}^2$$

Onde χ_0^2 foi calculado e $\chi_{\alpha;k-1-m}^2$ está localizado na tabela Qui-Quadrado, lembrando que α é o nível de significância e m é o número de parâmetros da distribuição ajustada.

c. Teste de *Kolmogorov-Smirnov*

Este teste possui hipóteses diferentes dos anteriores. Dessa vez, o objetivo é checar se duas amostras diferentes, de populações X e Y diferentes, possuem a mesma distribuição.

As hipóteses são:

$$H_0: X \text{ e } Y \text{ têm a mesma distribuição}$$

$$H_1: X \text{ e } Y \text{ não têm a mesma distribuição}$$

Sendo $F(x)$ o valor observado e $G(x)$ o valor conhecido para um valor de x, assumindo uma distribuição específica, o seguinte cálculo é feito:

$$d = \max(|F(x_i) - G(x_i)|; |F(x_i) - G(x_{i+1})|)$$

Ou seja, d é o módulo da maior diferença entre valor observado atual e valor conhecido atual, ou entre valor observado atual e valor conhecido seguinte.

Uma tabela especial (com valores críticos do teste de *Kolmogorov-Smirnov*) indica o valor $d_{crítico}$, e H_0 é rejeitada se:

$$d_{calculado} > d_{crítico}$$

d. Papel de Probabilidade Normal

O papel de probabilidade Normal (PPN) é utilizado para verificar se uma determinada população possui distribuição aproximadamente Normal.

Para verificar isso, os seguintes passos devem ser seguidos:

I. Ordenamos os valores do menor para o maior, atribuindo um índice i, que começa em 1;

II. Aplicamos a seguinte fórmula para cada $i \leq n$, onde n é o número de elementos:

$$y = \frac{50(2i - 1)}{n}$$

III. Inserimos os valores y_i (dados em porcentagem) no eixo vertical do PPN (respeitando a escala), e os valores de x_i fornecidos no eixo horizontal (estabelecendo uma escala);

IV. Plotamos os pontos e criamos uma reta que minimize a distância entre reta e pontos.

Se a reta for suficientemente próxima aos pontos, então pode-se dizer que a distribuição é aproximadamente Normal.

A média μ pode ser encontrada. Basta traçar uma linha horizontal no símbolo, e encontrar o valor na reta horizontal correspondente à intersecção entre a reta criada e a linha traçada.

O desvio padrão σ também pode ser obtido; basta traçar uma linha horizontal no símbolo $+1\sigma$, encontrar o valor na reta horizontal no ponto de intersecção entre a reta criada e a linha traçada, e subtrair a média já obtida.

2. Análise de Variância (ANOVA)

A Análise de Variância, ANOVA, é uma ferramenta poderosa para comparar diferentes médias populacionais, a partir de médias amostrais. As hipóteses são:

$$H_0: \mu_1 = \mu_2 = \ ... \ = \mu_i = \mu$$

H_1: Existe pelo menos uma média diferente

Para realizar a análise de variância, algumas premissas devem ser adotadas sempre:

I. As populações são independentes e possuem distribuição Normal;

II. As médias populacionais são todas desconhecidas;

III. As populações são homocedásticas (ou seja, elas possuem a mesma variância).

A tabela ANOVA se estrutura em uma coluna de soma de quadrados, uma de graus de liberdade, uma de quadrados médios, uma de $F_{calculado}$ e outra de $F_{tabelado}$.

A comparação entre os valores calculado e tabelado de F de *Snedecor* permite realizar o teste de hipótese.

a. ANOVA Simples

O caso principal de ANOVA é quando há apenas um fator, mas diversas amostras desse mesmo fator.

Assim, sejam x_{ij} a observação j da amostra i, n_i o tamanho da amostra i, k o número de amostras (linhas) e n o número total de elementos.

Se H_1 for verdadeira, a variável resposta x_{ij} pode ser aproximada por um modelo linear:

$$x_{ij} = \mu + \alpha_i + e_{ij}$$

Onde μ é o efeito médio geral, α_i é o efeito comum a todas as amostras e e_{ij} um efeito não controlado (resíduo).

As hipóteses também podem ser escritas como:

$$H_0: \alpha_1 = \alpha_2 = \ \ldots \ = \alpha_i = \alpha$$

$$H_1: \alpha_i \neq 0 \ (i = 1, \ldots, a)$$

A equação fundamental usada nas análises de variâncias é a que envolve soma dos quadrados.

Verifica-se uma relação entre a soma dos quadrados totais (SQT), soma dos quadrados entre as amostras (SQE) e soma dos quadrados residuais (SQR):

$$SQT = SQE + SQR$$

Onde:

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij} - \bar{x}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i} x_{ij}^2 - \frac{\left(\sum_{i=1}^{k} x_i\right)^2}{n}$$

$$SQE = \sum_{i=1}^{k} n_i(\overline{x_i} - \bar{x})^2 = \left(\sum_{i=1}^{k}\frac{(x_i)^2}{n_i}\right) - \frac{\left(\sum_{i=1}^{k} x_i\right)^2}{n}$$

Os graus de liberdade estão na tabela abaixo. Os quadrados médios são calculados pela divisão entre soma de quadrado por grau de liberdade (apenas para os casos entre amostras e residual).

O $F_{calculado}$ é dado pela razão entre os quadrados médios, e o $F_{tabelado}$ é encontrado, na tabela de probabilidade α, na linha $k-1$ e na coluna $n-k$:

Fonte de variação	Soma de quadrados	Graus de liberdade	Quadrado médio	F	$F_{k-1;(\Sigma n)-k;\alpha}$
Entre amostras	$SQE = \left(\sum_{i=1}^{k}\frac{(x_i)^2}{n_i}\right) - \frac{\left(\sum_{i=1}^{k} x_i\right)^2}{N}$	$k-1$	$s_E^2 = \frac{SQE}{k-1}$	$F = \frac{s_E^2}{s_R^2}$	$F_{Critico}$
Dentro das amostras Residual	$SQR = SQT - SQE$	$\left(\sum_{i=1}^{k} n_i\right) - k$	$s_R^2 = \frac{SQR}{\left(\sum_{i=1}^{k} n_i\right) - k}$		
Total	$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(x_{ij}\right)^2 - \frac{\left(\sum_{i=1}^{k} x_i\right)^2}{N}$	$\left(\sum_{i=1}^{k} n_i\right) - 1$			

A hipótese H_0 é rejeitada se:

$$F_{calculado} > F_{tabelado}$$

Se rejeitamos H_0, então pode-se concluir que existe efeito do fator em questão nos resultados.

b. ANOVA de Dois Fatores Sem Repetição

Outro caso importante é quando há dois ou mais fatores e diversas amostras de cada um (sem repetições).

Sejam x_{ij} a observação da amostra na linha i e coluna j, k o número de linhas e n o número de colunas.

Se H_1 for verdadeira, a variável resposta x_{ij} pode ser aproximada por um modelo linear:

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ijk}$$

Onde β_i é o efeito comum a todas as observações no nível j.

Além das hipóteses comentadas no caso anterior, será analisado o efeito causado pelo novo fator. Portanto, são testadas também as seguintes hipóteses:

$$H_0: \beta_1 = \beta_2 = \ \ldots \ = \beta_i = \beta$$

$$H_1: \beta_i \neq 0 \ (i = 1, \ldots, b)$$

A equação da soma dos quadrados envolve, nesse caso, a soma dos quadrados entre linhas (SQL) e entre colunas (SQC), além da total e da residual:

$$SQT = SQC + SQL + SQR$$

Onde:

$$SQT = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(x_{ij} - \bar{x}\right)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}^2 - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2}{nk}$$

$$SQC = \left(\sum_{j=1}^{n}\frac{\left(x_j\right)^2}{k}\right) - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2}{nk}$$

$$SQL = \left(\sum_{i=1}^{k}\frac{\left(x_i\right)^2}{n}\right) - \frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}\right)^2}{nk}$$

Na tabela abaixo, estão os graus de liberdade, os quadrados médios e os valores de F:

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	$F_{Calculado}$	$F_{Crítico}$
Entre Colunas	$SQC = \left(\sum_{j=1}^{n} \frac{(x_j)^2}{k} \right) - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2}{nk}$	$n-1$	$S_C^2 = \frac{SQC}{n-1}$	$F_C = \frac{S_C^2}{S_R^2}$	$F_{n-1;(n-1)(k-1);\alpha}$
Entre Linhas	$SQL = \left(\sum_{i=1}^{k} \frac{(x_i)^2}{n} \right) - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2}{nk}$	$k-1$	$S_L^2 = \frac{SQL}{k-1}$	$F_L = \frac{S_L^2}{S_R^2}$	$F_{k-1;(n-1)(k-1);\alpha}$
Residual	$SQR = SQT - SQL - SQC$	$(n-1)(k-1)$	$s_R^2 = \frac{SQR}{(n-1)(k-1)}$		
Total	$SQT = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij})^2 - \frac{\left(\sum_{i=1}^{k} \sum_{j=1}^{n} x_{ij} \right)^2}{nk}$	$nk-1$			

A hipótese H_0 é rejeitada se:

$$F_{calculado} > F_{tabelado}$$

Se rejeitamos H_0, então pode-se concluir que existe efeito do fator em questão nos resultados. No caso, isso pode ocorrer na análise entre linhas ou entre colunas.

Se rejeitamos a hipótese relacionada com a análise entre linhas, então é preciso considerar o efeito do fator na coluna.

Se rejeitamos a hipótese relacionada com a análise entre colunas, então é preciso considerar o efeito do fator na linha.

c. ANOVA de Dois Fatores Com Repetição

Neste caso, há dois ou mais fatores em consideração, com diversas amostras realizadas para cada.

No entanto, existe repetição do experimento, de forma que existem r repetições para cada linha i e coluna j.

Por isso, o modelo linear é:

$$x_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Onde $(\alpha\beta)_i$ é o efeito da interação entre os fatores.

Além das hipóteses comentadas no caso anterior, deve ser considerada a possibilidade de efeito da interação. Portanto, são testadas também as seguintes hipóteses:

Para a tabela abaixo, r é o número de repetições. São calculados valores de F entre linhas, colunas e avaliando interação.

Observação: Na tabela abaixo, i e k são referentes às colunas e j e n se referem às linhas:

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	$F_{Calculado}$	$F_{Crítico}$
Entre Linhas	$SQL=\left(\sum_{j=1}^{n}\frac{(x_j)^2}{kr}\right)-\frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ijr}\right)^2}{nkr}$	$n-1$	$s_L^2=\frac{SQL}{n-1}$	$F_L=\frac{s_L^2}{s_R^2}$	$F_{n-1;nk(r-1);\alpha}$
Entre Colunas	$SQC=\left(\sum_{i=1}^{k}\frac{(x_i)^2}{nr}\right)-\frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ijr}\right)^2}{nkr}$	$k-1$	$s_C^2=\frac{SQC}{k-1}$	$F_C=\frac{s_C^2}{s_R^2}$	$F_{k-1;nk(r-1);\alpha}$
Interação	$SQI=SQTr-SQL-SQC$	$(n-1)(k-1)$	$s_I^2=\frac{SQI}{(n-1)(k-1)}$	$F_I=\frac{s_I^2}{s_R^2}$	$F_{(n-1)(k-1);nk(r-1);\alpha}$
Residual	$SQR=SQT-SQTr$	$nk(r-1)$	$s_R^2=\frac{SQR}{nk(r-1)}$		
Entre Tratamentos	$SQTr=\sum_{i=1}^{k}\sum_{j=1}^{n}\frac{(x_{ij})^2}{r}-\frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ijr}\right)^2}{nkr}$	$nk-1$	$s_{Tr}^2=\frac{SQTr}{nk-1}$		
Total	$SQT=\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ijr})^2-\frac{\left(\sum_{i=1}^{k}\sum_{j=1}^{n}x_{ijr}\right)^2}{nkr}$	$nkr-1$			

3. Correlação

A correlação estabelece uma relação entre duas populações – com base nas amostras obtidas delas.

Usamos o conceito de covariância na definição do coeficiente de correlação:

$$S_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})=\sum x_i y_i-\frac{\sum x_i \sum y_i}{n}$$

$$S_{xx}=\sum(x_i-\bar{x})^2=\sum x_i^2-\frac{(\sum x_i)^2}{n}$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n}$$

O coeficiente de correlação (ou coeficiente de *Pearson*) é calculado como:

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

Podemos aplicar um teste de hipóteses para esse coeficiente (buscando verificar sua validade a partir das amostras); as hipóteses são:

$$\begin{cases} H_0 : \rho = \rho_0 \\ H_1 : \rho \neq \rho_0 \end{cases}$$

Encontramos um $t_{calculado}$ usando a seguinte expressão:

$$t_{calc} = R \sqrt{\frac{n-2}{1-R^2}}$$

E comparamos seu valor com um $t_{tabelado}$ (pela tabela t de *Student*), tal que:

$$t_{tab} = t_{\nu;\alpha}$$

Onde:

$$\nu = n - 2$$

Se $t_{calc} > t_{tab}$, então rejeitamos H_0, ou seja, aceitamos H_1 (existe correlação).

4. Regressão

A regressão permite estabelecer uma relação entre diferentes variáveis. Com ela, é possível estimar resultados ainda não vistos ou analisados.

A reta de regressão é a reta formada pelo método de regressão linear. Ela tem o formato:

$$y = a + bx$$

Com parâmetros populacionais, o modelo para a relação linear é:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Assim, a é estimativa do coeficiente angular α, enquanto b é estimativa do coeficiente linear β.

Existe um valor ε_i que corresponde ao resíduo, ou seja, a diferença entre os valores amostrais e populacionais.

As relações descritas abaixo surgem do Método dos Mínimos Quadrados (MMQ). Este método é capaz de fornecer a reta que minimiza o resíduo.

O coeficiente angular é:

$$b = \frac{n\sum(x_i y_i) - (\sum x_i)(\sum y_i)}{n\sum x_i^2 \ - (\sum x_i)^2} = \frac{S_{xy}}{S_{xx}}$$

E o coeficiente linear (ou intercepto):

$$a = \frac{\sum y_i \ - b\sum x_i}{n} = \bar{y} - b\bar{x}$$

O coeficiente de determinação é uma medida do quanto da variação a regressão é capaz de explicar, e vale:

$$R^2 = \frac{\sum(f(x_i) - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

Onde $f(x_i)$ é o valor calculado pela reta de regressão:

$$f(x_i) = a + bx_i$$

O termo $\sum(f(x_i) - \bar{y})^2$ é chamado de variação explicada, enquanto $\sum(y_i - \bar{y})^2$ é a variação total.

Existe uma tabela ANOVA para regressão. As hipóteses estabelecidas são:

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

Desta forma, se o valor de $F_{calculado}$ for maior do que $F_{tabelado}$, a hipótese H_0 é rejeitada, então existe correlação entre a variável resposta e a variável explicativa.

Abaixo está a tabela completa da ANOVA para regressão:

Fonte de variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	$F_{Calculado}$
Regressão	1	$SQR = \sum (f(x) - \bar{y})^2$	$QMR = \frac{SQR}{1}$	$F_{Calc} = \frac{QMR}{QME}$
Erro	N-2	$SQE = \sum (y - f(x))^2$	$QME = \frac{SQE}{N-2}$	
Total	N-1	$SQT = \sum (y - \bar{y})^2$	$QMT = \frac{SQT}{N-1}$	

O seguinte valor é importante para os próximos cálculos:

$$S_R^2 = \frac{S_{yy} - bS_{xy}}{n-2}$$

O coeficiente angular teórico está no seguinte intervalo de confiança:

$$\beta = b \pm t_{n-2;\alpha/2} \frac{S_R}{\sqrt{S_{xx}}}$$

Por sua vez, o coeficiente linear teórico está no seguinte intervalo de confiança:

$$\alpha = a \pm t_{n-2;\alpha/2} \sqrt{\frac{S_R^2}{S_{xx}} \cdot \frac{\sum x_i^2}{n}}$$

Para realizar um teste de hipóteses (com a variável t de *Student*) para o β, o seguinte valor corresponde ao $t_{calculado}$ (ou observado):

$$t_{calc} = \frac{b - \beta_0}{\frac{S_R}{\sqrt{S_{xx}}}}$$

Para o α:

$$t_{calc} = \frac{a - \alpha_0}{\sqrt{\frac{S_R^2}{S_{xx}} \cdot \frac{\sum x_i^2}{n}}}$$

O valor esperado tem seu próprio intervalo de confiança, dado por:

$$\mu(y) = f(x_0) \pm t_{n-2;\alpha/2}\, S_R \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Por fim, a predição (ou seja, o valor de y previsto) tem o seguinte intervalo de confiança:

$$y = f(x_0) \pm t_{n-2;\alpha/2}\, S_R \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

5. Linearização

Algumas funções que relacionam as variáveis x e y não são lineares. No entanto, para criar uma reta de regressão, é possível obter um formato linear dessas funções.

a. Função Potência

Por exemplo, a função potência pode ser linearizada. Ela é dada por:

$$y = Cx^b$$

Aplicando as propriedades de logaritmo, chegamos em:

$$\log y = \log C + b \log x$$

Nesse caso, definindo $z = \log y$, $a = \log C$ e $w = \log x$, temos:

$$z = a + bw$$

b. Função Exponencial

Em funções exponenciais, como:

$$y = Ce^{bx}$$

Aplicando as propriedades de logaritmo natural, chegamos em:

$$\ln y = \ln C + bx$$

Nesse caso, definindo $z = \log y$ e $a = \log C$, temos:

$$z = a + bx$$

c. Função Hipérbole

Considere a função hipérbole, definida como:

$$y = a + \frac{b}{x}$$

Aplicando a transformação $z = \frac{1}{x}$, chegamos em:

$$y = a + bz$$

d. Função Denominadora
Considere a seguinte função:

$$y = \frac{1}{a + bx}$$

Aplicando a transformação $z = \frac{1}{y}$, chegamos em:

$$z = a + bx$$

6. Regressão Linear Múltipla

A regressão linear múltipla é semelhante à regressão simples, mas busca-se obter uma relação entre uma variável resposta (y) e várias variáveis controladas $(x_1, x_2, \dots, x_n)$.

O modelo teórico para essa regressão é:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ \dots + \beta_k x_k + \varepsilon$$

Usando o Método dos Mínimos Quadrados, a equação obtida a partir das amostras:

$$f(x) = \hat{y} = a + b_1x_1 + b_2x_2 + \ ... \ + b_kx_k$$

Se $x_1 = x$, $x_2 = x^2$, até $x_k = x^k$, obtemos uma regressão polinomial, com forma geral:

$$f(x) = \hat{y} = a + b_1x + b_2x^2 + \ ... \ + b_kx^k$$

O coeficiente linear é obtido pela equação:

$$a = \ \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ ... \ - b_k\bar{x}_k$$

E os coeficientes b_i são obtidos a partir do seguinte sistema linear:

$$\begin{cases} S_{1y} = b_1S_{11} + b_2S_{12} + \cdots + b_kS_{1k} \\ S_{2y} = b_1S_{21} + b_2S_{22} + \cdots + b_kS_{2k} \\ \qquad\qquad ... \\ S_{ky} = b_1S_{k1} + b_2S_{k2} + \cdots + b_kS_{kk} \end{cases}$$

Tal que:

$$S_{ii} = S_{x_ix_i} = \sum(x_i - \bar{x}_i)^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$S_{ij} = S_{x_ix_j} = \sum(x_i - \bar{x}_i)\left(x_j - \bar{x}_j\right) = \sum x_ix_j - \frac{(\sum x_i)\left(\sum x_j\right)}{n}$$

$$S_{iy} = S_{x_iy} = \sum(x_i - \bar{x}_i)(y - \bar{y}) = \sum x_iy - \frac{(\sum x_i)(\sum y)}{n}$$

A avaliação do modelo se faz com uma tabela ANOVA. Para verificar se a regressão é significativa, as hipóteses são:

$$\begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1: \beta_i \neq 0, 1 \leq i \leq k \end{cases}$$

A tabela ANOVA está abaixo; o valor de $F_{tabelado}$, neste caso, é:

$$F_{tabelado} = F_{k;n-k-1;\alpha}$$

Fonte de variação	Graus de liberdade	Somas dos quadrados	Quadrados médios	Estatística F
Regressão	k	$SQE = \Sigma b_i S_{iy}$	$s_E^2 = SQE/k$	$F = \frac{s_E^2}{s_R^2}$
Resíduo	$n-k-1$	$SQR = S_{yy} - \Sigma b_i S_{iy}$	$s_R^2 = SQR/(n-k-1)$	
Total	$n-1$	$SQT = S_{yy}$		

O coeficiente de determinação é dado por:

$$R^2 = \frac{SQE}{SQT} = \frac{\sum b_i S_{iy}}{S_{yy}}$$

No entanto, para melhor comparar o ajuste de dois modelos, é preferível a comparação entre os coeficientes de determinação ajustados, calculados por:

$$R^2_{ajustado} = \frac{\frac{SQT}{n-1} - \frac{SQR}{n-k-1}}{\frac{SQT}{n-1}} = 1 - \frac{n-1}{n-k-1}(1-R^2)$$

O modelo com maior $R^2_{ajustado}$ é o que possui melhor ajuste.

Por fim, suponha que um modelo possui $k-1$ variáveis independentes. Podemos analisar se a entrada de uma nova variável independente, x_k, produz melhoria significativa no ajuste.

Isso ocorre com o auxílio da Análise de Melhoria. É construída uma tabela ANOVA, onde:

H_0: Não ocorre melhoria significativa com a entrada de x_k

H_1: Ocorre melhoria significativa com a entrada de x_k

O teste é realizado com a seguinte tabela:

Fonte de variação	Graus de liberdade	Soma de quadrados	Quadrados médios	Estatística F
Devida à melhoria de ajuste	1	SQM	SQM	$F = \frac{SQM}{s_k^2}$
Residual para o modelo com k variáveis	$n-k-1$	$SQR(k)$	$s_k^2 = \frac{SQR(k)}{n-k-1}$	
Residual para o modelo com $(k-1)$ variáveis	$n-k$	$SQR(k-1)$		

Onde:

$$SQM = \sum(\bar{y}_{P_i} - \bar{y}_i)^2$$

$$SQR(k) = \sum\left(y_i - \bar{y}_{P\,i}\right)^2$$

$$SQR(k-1) = S_{yy} - b^2 S_{xx}$$

Além disso:

$$F_{tabelado} = F_{1;n-k-1;\alpha}$$

Lista de Exercícios

1. Teste Qui-Quadrado

Slides de Aula

O número de defeitos por unidade observado em uma amostra de 100 aparelhos de televisão produzidos em uma linha de montagem apresentou a distribuição de frequências como na tabela abaixo.

Verifique se o número de defeitos por unidade segue uma Distribuição de Poisson. Considere um nível de significância de 5%.

Nº de defeitos (X_i)	0	1	2	3	4	5	6	7	Total
Frequência Observada (O_i)	25	35	18	13	4	2	2	1	100

2. ANOVA com Um Fator

Slides de Aula

Considere os dados abaixo e verifique se há evidências de que as médias não sejam iguais.

A	B	C
12	3	4
14	4	5
17	5	3
15	4	
11		

3. ANOVA com Dois Fatores

P2 2015 Estatística Poli USP

Um experimento foi realizado visando avaliar o efeito de dois fatores na conversão (produtividade) de um processo químico: temperatura e pressão. Por falta de tempo, não foi possível fazer o experimento com repetição.

Um funcionário descuidado deixou cair café sobre a tabela, perdendo alguns de seus valores. Felizmente, alguns cálculos já haviam sido realizados. Os resultados obtidos foram:

	$1\ atm$		$2\ atm$		$3\ atm$		$Soma\ dos\ X_{jj}$ $valores\ da\ linha$	$Soma\ dos (X_{jj})^2$ $valores\ da\ linha$
$220ºC$	5,4	29				36	17,1	97,65
$250ºC$	3,2					38,44	13,8	68,04
$270ºC$	3,8	14,				34,81	13,9	66,89
$300ºC$	4,6	21,16	5,2	27,04	6,1	37,21	15,9	85,41
$Soma\ dos\ X_{jj}$ $valores\ da\ coluna$	17		19,5		24,2			
$Soma\ dos (X_{jj})^2$ $valores\ da\ coluna$		75		96,53		146,6		

a. Ao nível de significância de 5%, existem diferenças significativas entre os fatores Pressão e Temperatura?
b. Para responder à questão anterior, que hipóteses você teve que assumir?

4. Regressão

P2 2015 Estatística Poli USP

A primeira prova de *PRO3200* ocorreu no dia 19 de outubro de 2015, as 7:30 da manhã. A tabela com os acessos nos intervalos medidos está abaixo. "Acesso" aqui quer dizer um clique de algum aluno online em um dos recursos do AVA.

Período	Tempo faltante para a prova (Horas)	Número de Acessos
18/10 14:00 *a* 14:59	17	37
18/10 15:00 *a* 15:59	16	96
18/10 16:00 *a* 16:59	15	
18/10 17:00 *a* 17:59		
18/10 18:00 *a* 18:59		
18/10 19:00 *a* 19:59		
18/10 20:00 *a* 20:59		
18/10 21:00 *a* 21:59		
18/10 22:00 *a* 22:59		490
18/10 23:00 *a* 23:59	8	557
19/10 00:00 *a* 00:59	7	542
19/10 01:00 *a* 01:59	6	558
19/10 02:00 *a* 02:59	5	621
19/10 03:00 *a* 03:59	4	703

Novamente o funcionário descuidado deixou cair café sobre a tabela, perdendo alguns de seus valores, mas, felizmente, alguns cálculos já haviam sido realizados.

$$\sum x_i = 147 \qquad \sum y_i = 4921$$

$$\sum x_1^2 = 1771 \qquad \sum y_i^2 = 2360653$$

$$\bar{x} \cong 10{,}5 \qquad \bar{y} \cong 351{,}50$$

$$n = 14 \qquad \sum x_i \,.\, y_i = 39934$$

a. Determine a reta de regressão.
b. Ao nível de significância de 5%, é possível verificar correlação negativa entre proximidade de prova e número de acessos?
c. Determine o intervalo com 95% de confiança para os parâmetros Alfa e Beta da reta teórica.

5. Regressão

Elaboração Própria

Determine o intervalo de confiança, a partir da estimativa da reta da regressão $(\hat{y} = 0{,}174 + 0{,}217x)$, com 95% de confiança, para o valor previsto de y, quando $x = 8$, e para o valor médio de y, quando $x = 6$.

São dados: $n = 8, S_x = 42, \bar{x} = 4{,}5, S_R^2 = 0{,}0142$. Considere que $x = 8$ não pertence à amostra, mas pertence ao intervalo de variação estudado. Considere que $x = 6$ pertence à amostra.

6. Regressão Linear Múltipla

P2 2015 Estatística Poli USP

O consumo de energia por um ser humano para de manter vivo é denominado Taxa Metabólica Basal ou TBM. Reflete o consumo mínimo diário de *kcal* para o corpo de manter em repouso.

Um pesquisador estudou o TBM e quer correlacionar o TBM com a Massa Corporal Magra e a Massa de Gordura, medidas por bioimpedância. Observe os seguintes dados:

Y	1745,3	1684,3	2095,1	1645,2	1803,3	1774,5	2395,3	1985,4
MG	10	5	21	16	18	7,5	44	16
MCM	60	55	72	45	62	63	56	67

Y: TBM $kcal$, MG: Massa gorda kg, MCM: massa magra kg.

Avaliando a relação entre Y e MG encontrou-se a reta média $\tilde{Y} = 1573 + 18{,}5\ MG$ com a seguinte tabela ANOVA:

Fonte	Soma Quad.	Graus lib.	Quad. Médio	F calc.	F tab.
Regressão	353969	1	353969	22,00	5,99
Residual Reta	96552	6	16092		
Total	450521	7			

Para melhorar o modelo, resolveu incluir a variável MCMe encontrou o modelo (plano) para $\tilde{Y} = 853 + 11{,}9\ MCM + 18{,}7\ MG$ com a tabela ANOVA.

Fonte	Soma Quad.	Graus lib.	Quad. Médio	F calc.	F tab.
Regressão	421200	2			
Residual Reta					
Total	450521				

a. Complete a tabela de ANOVA preenchendo os valores faltantes nos espaços vazios.

b. Faça a análise de melhoria e verifique se a introdução da variável MCM no modelo traz melhoria significativa.

c. Qual o valor do coeficiente de determinação para o modelo de Regressão Linear Múltipla (plano)?

Papel de Probabilidade Normal

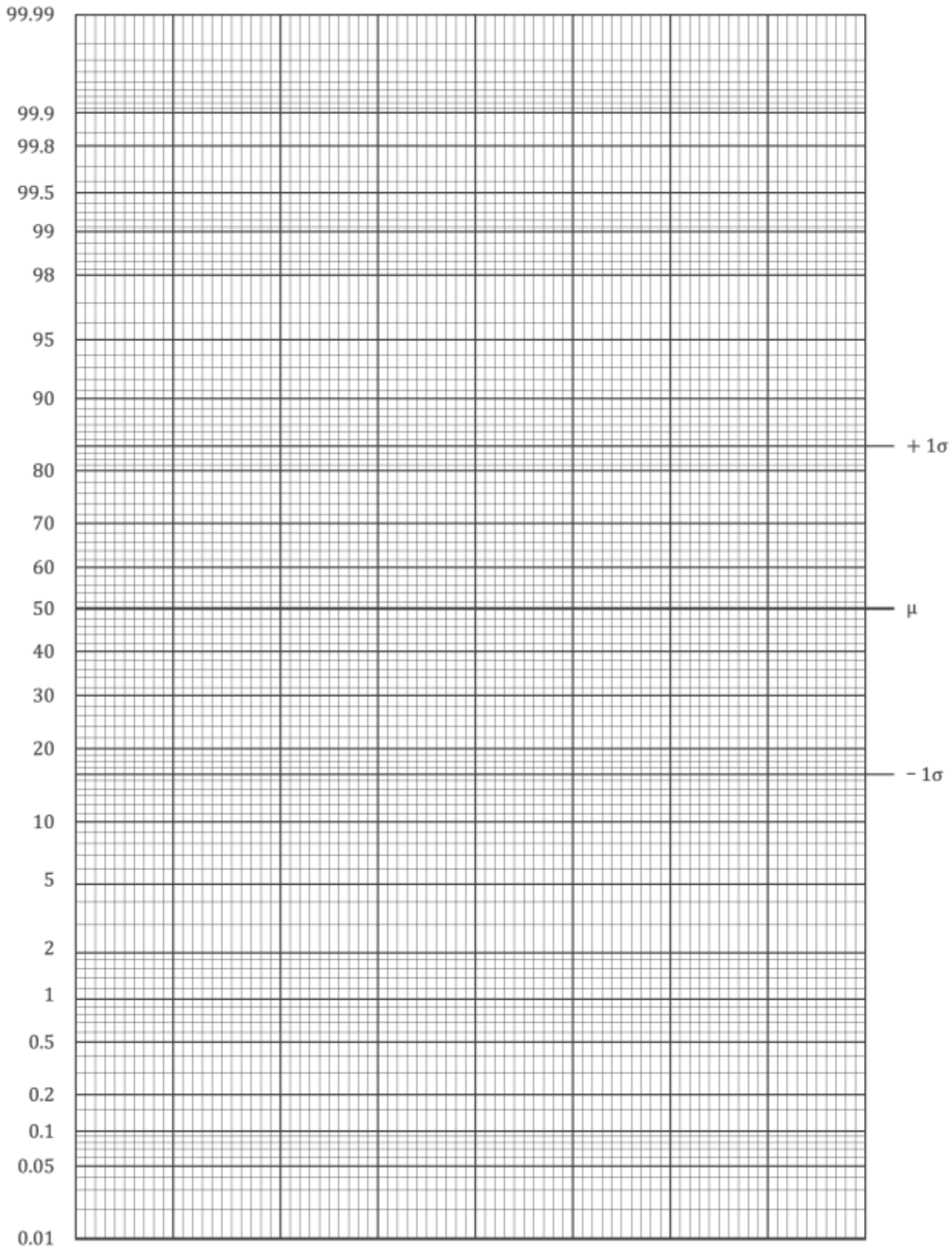
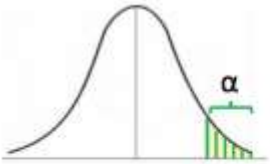
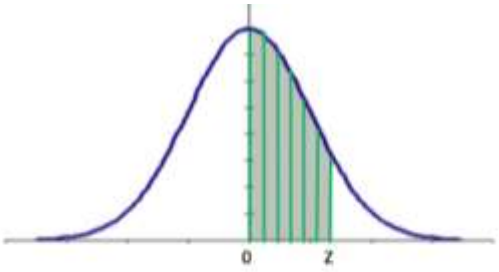


Tabela da Distribuição t-Student



	α										
Graus de Liberdade	**25%**	**12,5%**	**10,0%**	**5,0%**	**2,5%**	**1,25%**	**1,0%**	**0,5%**	**0,25%**	**0,1%**	**0,05%**
1	1,000	2,414	3,078	6,314	12,706	25,452	31,821	63,657	127,321	318,309	636,619
2	0,816	1,604	1,886	2,920	4,303	6,205	6,965	9,925	14,089	22,327	31,599
3	0,765	1,423	1,638	2,353	3,182	4,177	4,541	5,841	7,453	10,215	12,924
4	0,741	1,344	1,533	2,132	2,776	3,495	3,747	4,604	5,598	7,173	8,610
5	0,727	1,301	1,476	2,015	2,571	3,163	3,365	4,032	4,773	5,893	6,869
6	0,718	1,273	1,440	1,943	2,447	2,969	3,143	3,707	4,317	5,208	5,959
7	0,711	1,254	1,415	1,895	2,365	2,841	2,998	3,499	4,029	4,785	5,408
8	0,706	1,240	1,397	1,860	2,306	2,752	2,896	3,355	3,833	4,501	5,041
9	0,703	1,230	1,383	1,833	2,262	2,685	2,821	3,250	3,690	4,297	4,781
10	0,700	1,221	1,372	1,812	2,228	2,634	2,764	3,169	3,581	4,144	4,587
11	0,697	1,214	1,363	1,796	2,201	2,593	2,718	3,106	3,497	4,025	4,437
12	0,695	1,209	1,356	1,782	2,179	2,560	2,681	3,055	3,428	3,930	4,318
13	0,694	1,204	1,350	1,771	2,160	2,533	2,650	3,012	3,372	3,852	4,221
14	0,692	1,200	1,345	1,761	2,145	2,510	2,624	2,977	3,326	3,787	4,140
15	0,691	1,197	1,341	1,753	2,131	2,490	2,602	2,947	3,286	3,733	4,073
16	0,690	1,194	1,337	1,746	2,120	2,473	2,583	2,921	3,252	3,686	4,015
17	0,689	1,191	1,333	1,740	2,110	2,458	2,567	2,898	3,222	3,646	3,965
18	0,688	1,189	1,330	1,734	2,101	2,445	2,552	2,878	3,197	3,610	3,922
19	0,688	1,187	1,328	1,729	2,093	2,433	2,539	2,861	3,174	3,579	3,883
20	0,687	1,185	1,325	1,725	2,086	2,423	2,528	2,845	3,153	3,552	3,850
21	0,686	1,183	1,323	1,721	2,080	2,414	2,518	2,831	3,135	3,527	3,819
22	0,686	1,182	1,321	1,717	2,074	2,405	2,508	2,819	3,119	3,505	3,792
23	0,685	1,180	1,319	1,714	2,069	2,398	2,500	2,807	3,104	3,485	3,768
24	0,685	1,179	1,318	1,711	2,064	2,391	2,492	2,797	3,091	3,467	3,745
25	0,684	1,178	1,316	1,708	2,060	2,385	2,485	2,787	3,078	3,450	3,725
26	0,684	1,177	1,315	1,706	2,056	2,379	2,479	2,779	3,067	3,435	3,707
27	0,684	1,176	1,314	1,703	2,052	2,373	2,473	2,771	3,057	3,421	3,690
28	0,683	1,175	1,313	1,701	2,048	2,368	2,467	2,763	3,047	3,408	3,674
29	0,683	1,174	1,311	1,699	2,045	2,364	2,462	2,756	3,038	3,396	3,659
30	0,683	1,173	1,310	1,697	2,042	2,360	2,457	2,750	3,030	3,385	3,646
40	0,681	1,167	1,303	1,684	2,021	2,329	2,423	2,704	2,971	3,307	3,551
50	0,679	1,164	1,299	1,676	2,009	2,311	2,403	2,678	2,937	3,261	3,496
60	0,679	1,162	1,296	1,671	2,000	2,299	2,390	2,660	2,915	3,232	3,460
100	0,677	1,157	1,290	1,660	1,984	2,276	2,364	2,626	2,871	3,174	3,390
150	0,676	1,155	1,287	1,655	1,976	2,264	2,351	2,609	2,849	3,145	3,357
300	0,675	1,153	1,284	1,650	1,968	2,253	2,339	2,592	2,828	3,118	3,323
∞	0,67	1,15	1,28	1,64	1,96	2,24	2,33	2,58	2,81	3,09	3,29

Tabela da Distribuição Normal



z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,0000	0,0040	0,0080	0,0120	0,0160	0,0199	0,0239	0,0279	0,0319	0,0359
0,1	0,0398	0,0438	0,0478	0,0517	0,0557	0,0596	0,0636	0,0675	0,0714	0,0753
0,2	0,0793	0,0832	0,0871	0,0910	0,0948	0,0987	0,1026	0,1064	0,1103	0,1141
0,3	0,1179	0,1217	0,1255	0,1293	0,1331	0,1368	0,1406	0,1443	0,1480	0,1517
0,4	0,1554	0,1591	0,1628	0,1664	0,1700	0,1736	0,1772	0,1808	0,1844	0,1879
0,5	0,1915	0,1950	0,1985	0,2019	0,2054	0,2088	0,2123	0,2157	0,2190	0,2224
0,6	0,2257	0,2291	0,2324	0,2357	0,2389	0,2422	0,2454	0,2486	0,2517	0,2549
0,7	0,2580	0,2611	0,2642	0,2673	0,2704	0,2734	0,2764	0,2794	0,2823	0,2852
0,8	0,2881	0,2910	0,2939	0,2967	0,2995	0,3023	0,3051	0,3078	0,3106	0,3133
0,9	0,3159	0,3186	0,3212	0,3238	0,3264	0,3289	0,3315	0,3340	0,3365	0,3389
1,0	0,3413	0,3438	0,3461	0,3485	0,3508	0,3531	0,3554	0,3577	0,3599	0,3621
1,1	0,3643	0,3665	0,3686	0,3708	0,3729	0,3749	0,3770	0,3790	0,3810	0,3830
1,2	0,3849	0,3869	0,3888	0,3907	0,3925	0,3944	0,3962	0,3980	0,3997	0,4015
1,3	0,4032	0,4049	0,4066	0,4082	0,4099	0,4115	0,4131	0,4147	0,4162	0,4177
1,4	0,4192	0,4207	0,4222	0,4236	0,4251	0,4265	0,4279	0,4292	0,4306	0,4319
1,5	0,4332	0,4345	0,4357	0,4370	0,4382	0,4394	0,4406	0,4418	0,4429	0,4441
1,6	0,4452	0,4463	0,4474	0,4484	0,4495	0,4505	0,4515	0,4525	0,4535	0,4545
1,7	0,4554	0,4564	0,4573	0,4582	0,4591	0,4599	0,4608	0,4616	0,4625	0,4633
1,8	0,4641	0,4649	0,4656	0,4664	0,4671	0,4678	0,4686	0,4693	0,4699	0,4706
1,9	0,4713	0,4719	0,4726	0,4732	0,4738	0,4744	0,4750	0,4756	0,4761	0,4767
2,0	0,4772	0,4778	0,4783	0,4788	0,4793	0,4798	0,4803	0,4808	0,4812	0,4817
2,1	0,4821	0,4826	0,4830	0,4834	0,4838	0,4842	0,4846	0,4850	0,4854	0,4857
2,2	0,4861	0,4864	0,4868	0,4871	0,4875	0,4878	0,4881	0,4884	0,4887	0,4890
2,3	0,4893	0,4896	0,4898	0,4901	0,4904	0,4906	0,4909	0,4911	0,4913	0,4916
2,4	0,4918	0,4920	0,4922	0,4925	0,4927	0,4929	0,4931	0,4932	0,4934	0,4936
2,5	0,4938	0,4940	0,4941	0,4943	0,4945	0,4946	0,4948	0,4949	0,4951	0,4952
2,6	0,4953	0,4955	0,4956	0,4957	0,4959	0,4960	0,4961	0,4962	0,4963	0,4964
2,7	0,4965	0,4966	0,4967	0,4968	0,4969	0,4970	0,4971	0,4972	0,4973	0,4974
2,8	0,4974	0,4975	0,4976	0,4977	0,4977	0,4978	0,4979	0,4979	0,4980	0,4981
2,9	0,4981	0,4982	0,4982	0,4983	0,4984	0,4984	0,4985	0,4985	0,4986	0,4986
3,0	0,4987	0,4987	0,4987	0,4988	0,4988	0,4989	0,4989	0,4989	0,4990	0,4990

Tabela da Distribuição Qui-Quadrado

	α					
ν	$\mathbf{0,99}$	$\mathbf{0,95}$	$\mathbf{0,90}$	$\mathbf{0,10}$	$\mathbf{0,05}$	$\mathbf{0,01}$
1	0,0002	0,0039	0,0158	2,706	3,841	6,635
2	0,0201	0,103	0,211	4,605	5,991	9,210
3	0,115	0,352	0,584	6,251	7,815	11,345
4	0,297	0,711	1,064	7,779	9,488	13,277
5	0,554	1,145	1,610	9,236	11,070	15,086
6	0,872	1,635	2,204	10,645	12,592	16,812
7	1,239	2,167	2,833	12,017	14,067	18,475
8	1,646	2,733	3,490	13,362	15,507	20,090
9	2,088	3,325	4,168	14,684	16,919	21,666
10	2,558	3,940	4,865	15,987	18,307	23,209
11	3,053	4,575	5,578	17,275	19,675	24,725
12	3,571	5,226	6,304	18,549	21,026	26,217
13	4,107	5,892	7,042	19,812	22,362	27,688
14	4,660	6,571	7,790	21,064	23,685	29,141
15	5,229	7,261	8,547	22,307	24,996	30,578
16	5,812	7,962	9,312	23,542	26,296	32,000
17	6,408	8,672	10,085	24,769	27,587	33,409
18	7,015	9,390	10,865	24,769	27,587	33,409
19	7,633	10,117	11,651	27,204	30,144	36,191
20	8,260	10,851	12,443	28,412	31,410	37,566
21	8,897	11,591	13,240	29,615	32,671	38,932
22	9,542	12,338	14,041	30,813	33,924	40,289
23	10,196	13,091	14,848	32,007	35,172	41,638
24	10,856	13,848	15,659	33,196	36,415	42,980
25	11,524	14,611	16,473	34,382	37,652	44,314
26	12,198	15,379	17,292	35,563	38,885	45,642
27	12,879	16,151	18,114	36,741	40,113	46,963
28	13,565	16,928	18,939	37,916	41,337	48,278
29	14,256	17,708	19,768	39,087	42,557	49,588
30	14,953	18,493	20,599	40,256	43,773	50,892
40	22,164	26,509	29,051	51,085	55,758	63,691
50	29,707	34,764	37,689	63,167	67,505	76,154
60	37,485	43,188	46,459	74,397	79,082	88,379

Tabela F-Snedecor ($\alpha = 5\%$)

	ν_1									
ν_2	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32