Estatística aplicada

Tópicos especiais em Estatística Aplicada

Prof. Celso J. Munaro (cjmunaro@gmail.com)

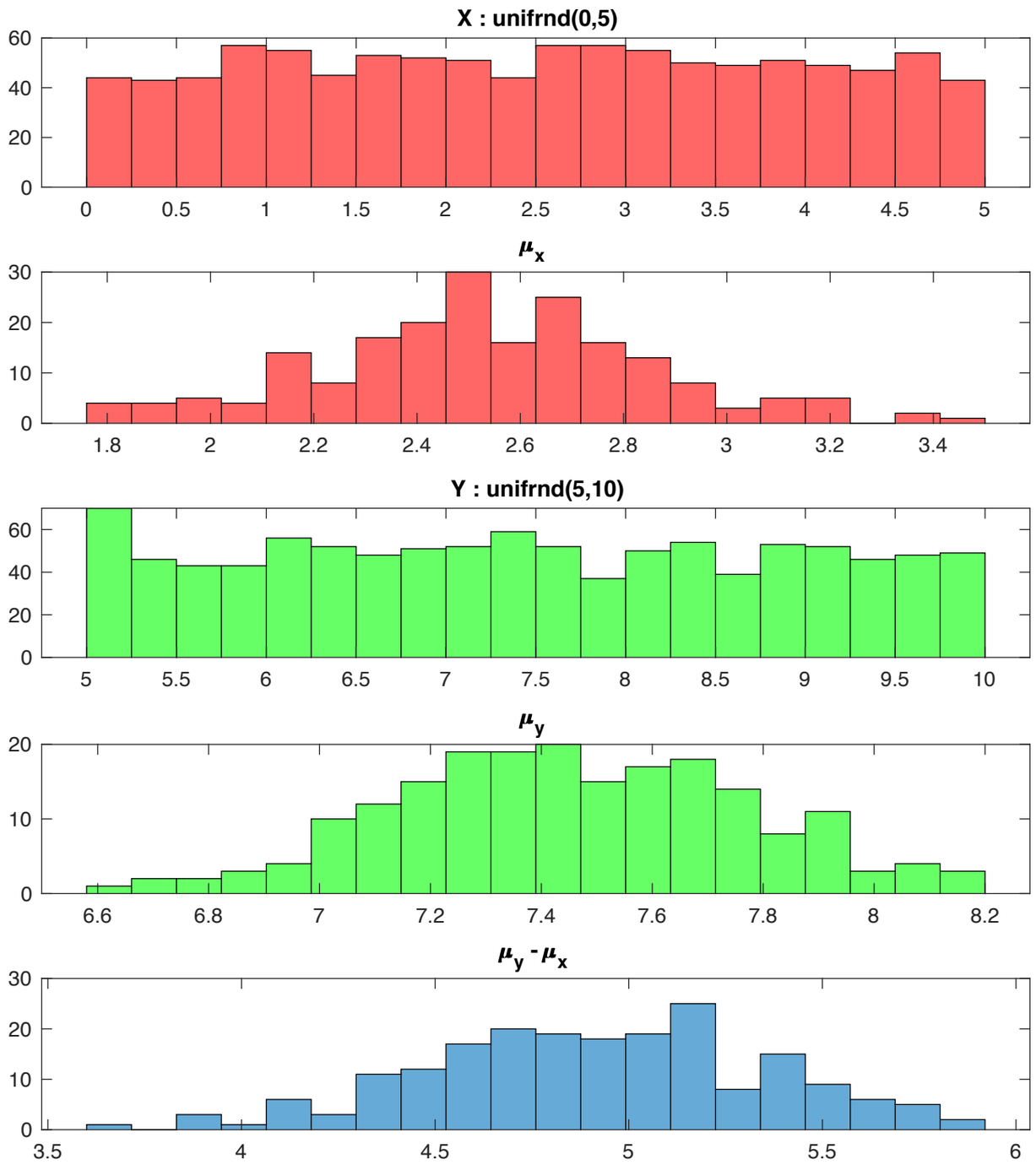
VIII - Inferência estatística para duas amostras

Cap 10 de [1]

Introdução

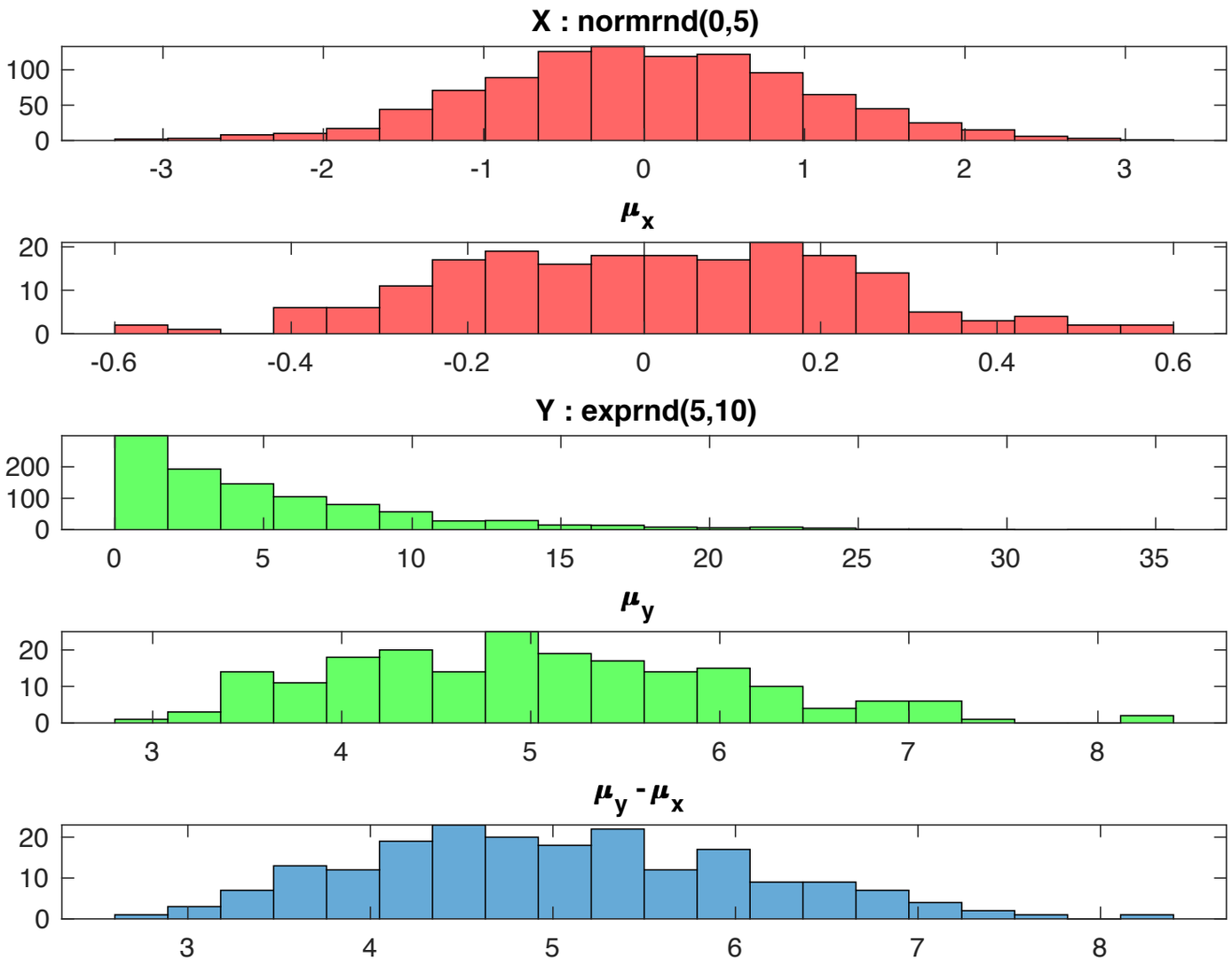
O objetivo agora é comparar os parâmetros das duas populações

Introdução



Observe os histogramas

Introdução

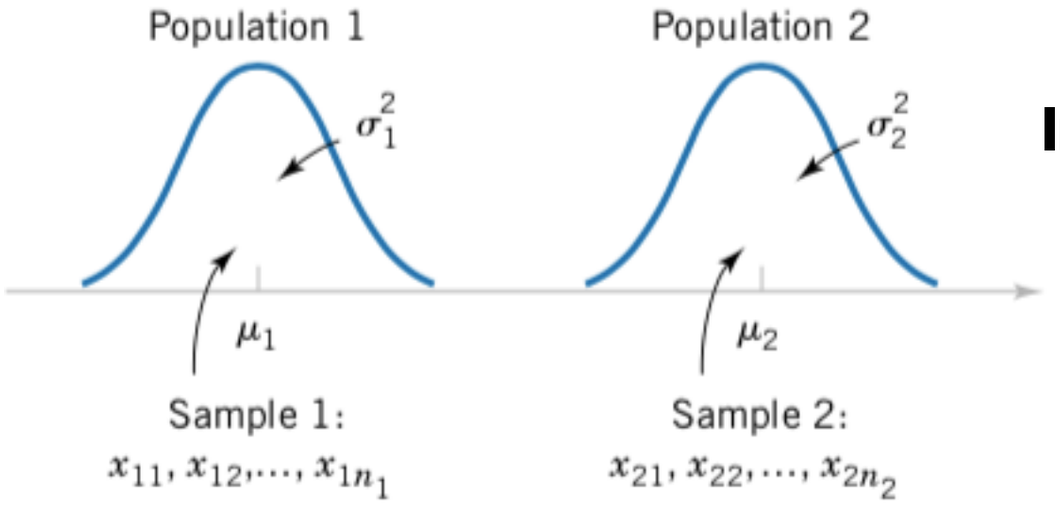


Observe os histogramas

Introdução

Suposição

1. $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1.
2. $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample from population 2.
3. The two populations represented by X_1 and X_2 are independent.
4. Both populations are normal.



Inferências sobre as médias, variâncias conhecidas

Introdução

Como

$$E(\bar{X}_1 - \bar{X}_2) = E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Introdução

Portanto,

The quantity

$$Z = \frac{X_1 - X_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

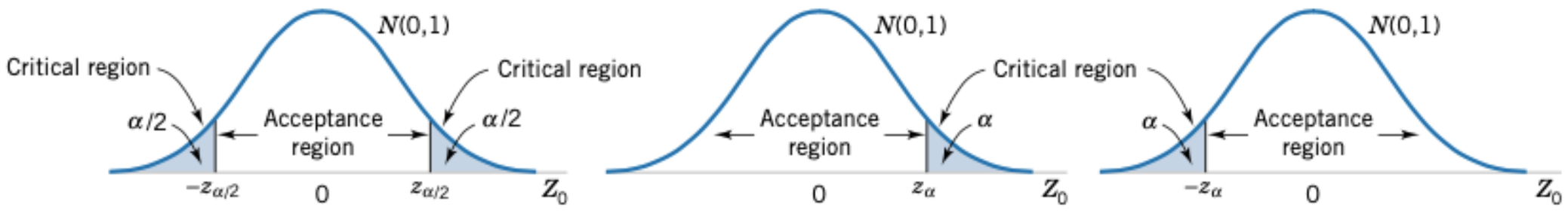
has a $N(0, 1)$ distribution.

Teste de hipóteses: diferença de médias, variâncias conhecidas

Null hypothesis: $H_0: \mu_1 - \mu_2 = \Delta_0$

Test statistic:
$$Z_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

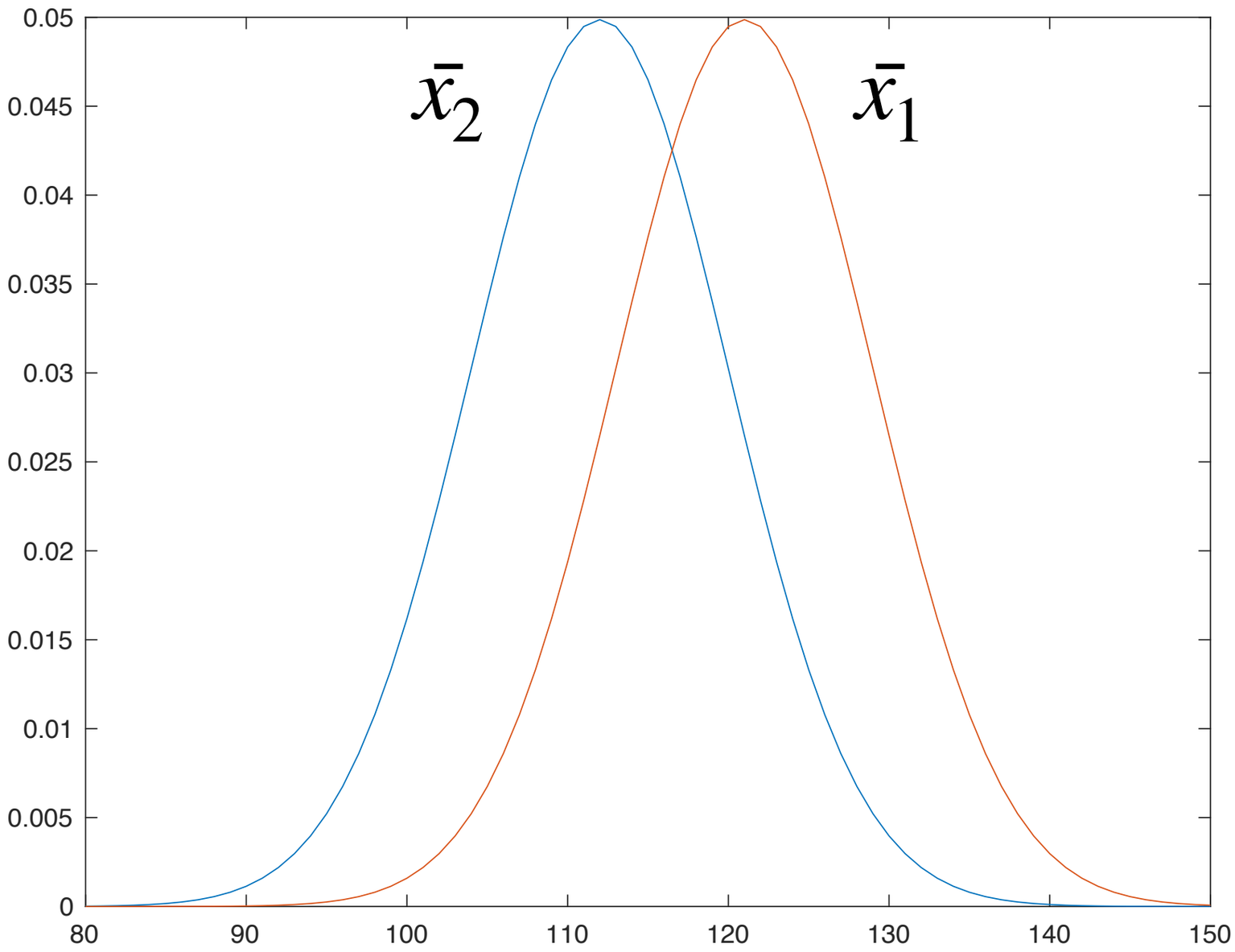
Alternative Hypotheses	Rejection Criterion
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	$z_0 > z_\alpha$
$H_1: \mu_1 - \mu_2 < \Delta_0$	$z_0 < -z_\alpha$



Exemplo 10-1

Duas tintas são feitas com diferentes formulações: deseja-se avaliar se o tempo para secagem é diferente. São feitas 10 pinturas com cada tinta e medidos os tempos de secagem. Sabe-se por experiência que o desvio padrão é 8min. As médias de tempo de secagem foram $\bar{x}_1 = 121$ e $\bar{x}_2 = 112$ minutos. Considerando $\alpha = 0.05$, pode-se afirmar que há diferença entre o tempo de secagem das tintas?

Exemplo 10-1


$\bar{x}_2$
$\bar{x}_1$
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
80
90
100
110
120
130
140
150

Exemplo 10-1

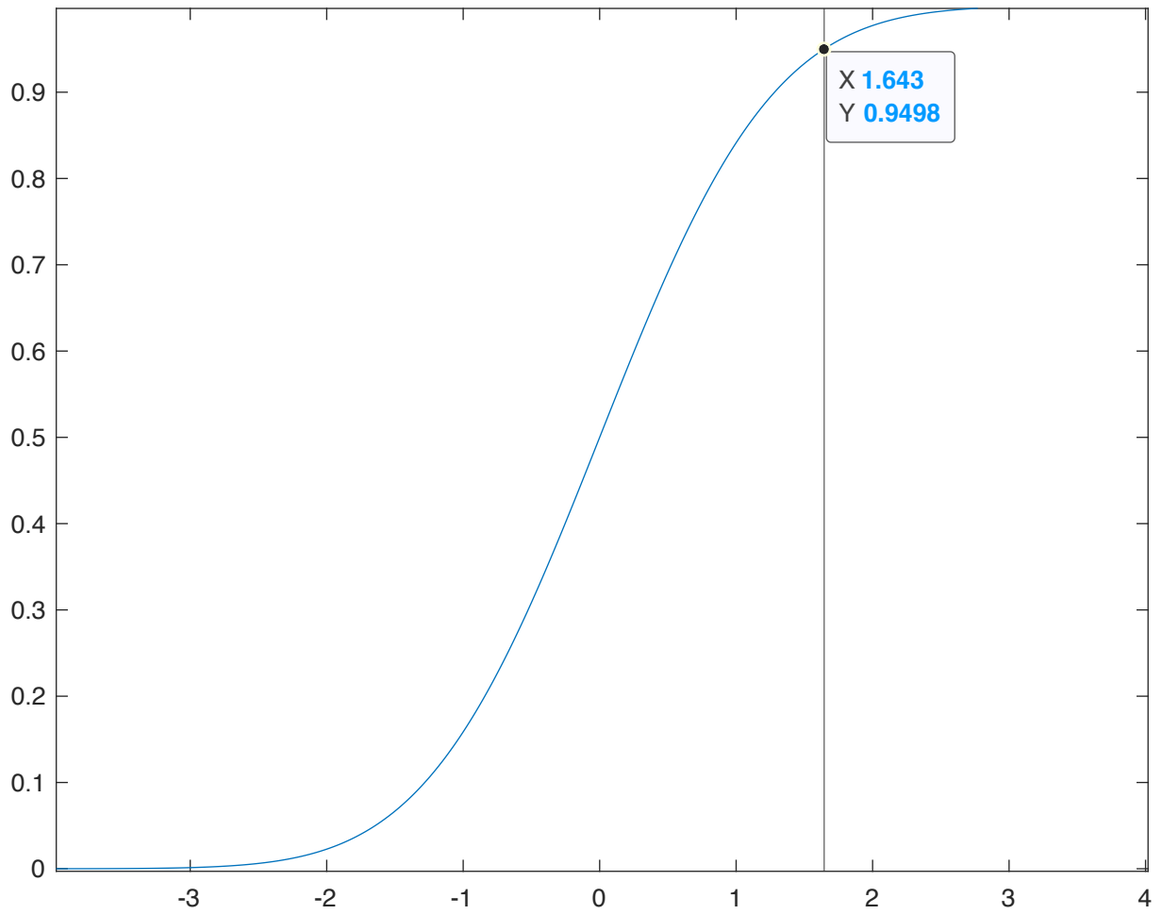
1. The quantity of interest is the difference in mean drying times, $\mu_1 - \mu_2$, and $\Delta_0 = 0$.
2. H_0: $\mu_1 - \mu_2 = 0$, or H_0: $\mu_1 = \mu_2$.
3. H_1: $\mu_1 > \mu_2$. We want to reject H_0 if the new ingredient reduces mean drying time.
4. $\alpha = 0.05$
5. The test statistic is

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

where $\sigma_1^2 = \sigma_2^2 = (8)^2 = 64$ and $n_1 = n_2 = 10$.

Exemplo 10-1

6. Reject H_0: $\mu_1 = \mu_2$ if $z_0 > 1.645 = z_{0.05}$.



Exemplo 10-1

6. Reject H_0: $\mu_1 = \mu_2$ if $z_0 > 1.645 = z_{0.05}$.

7. Computations: Since $\bar{x}_1 = 121$ minutes and $\bar{x}_2 = 112$ minutes, the test statistic is

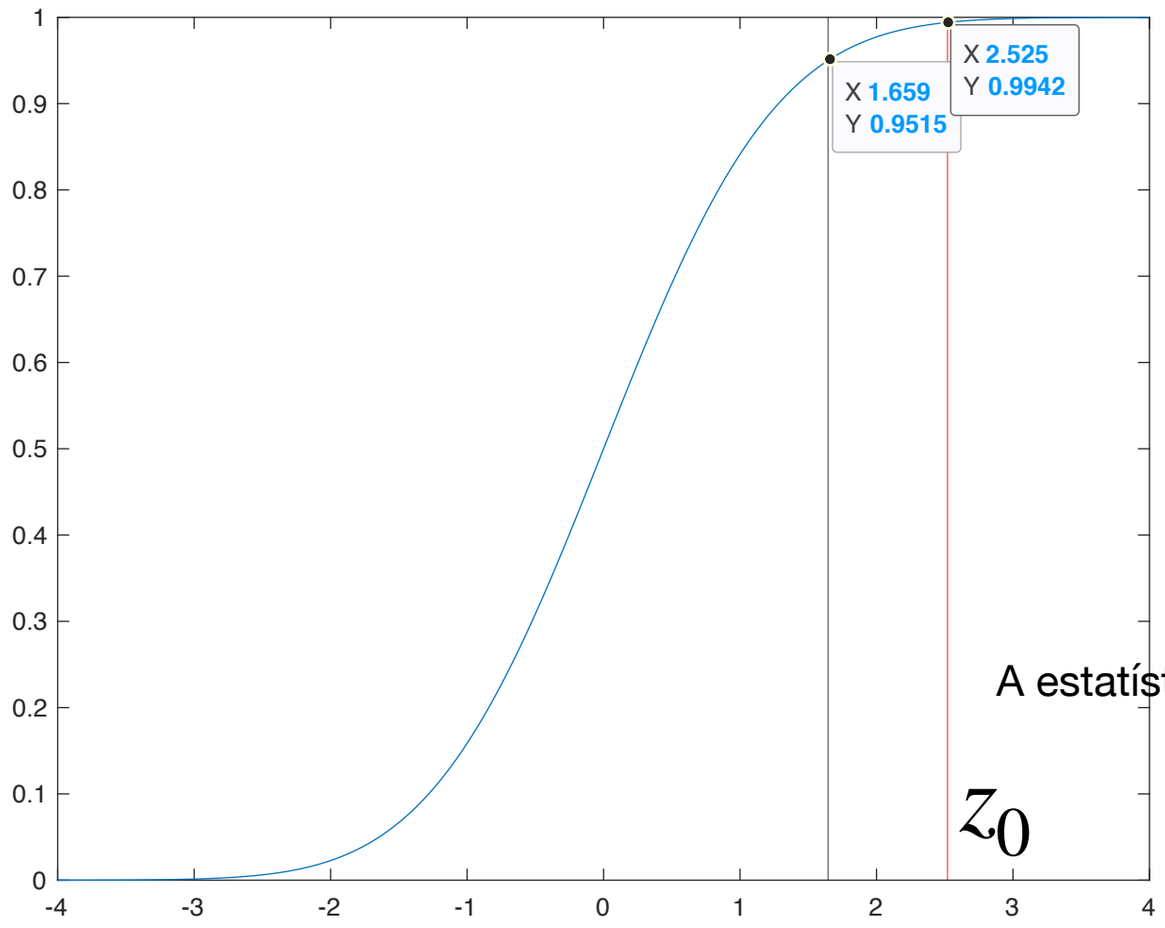
$$z_0 = \frac{121 - 112}{\sqrt{\frac{(8)^2}{10} + \frac{(8)^2}{10}}} = 2.52$$

8. Conclusion: Since $z_0 = 2.52 > 1.645$, we reject H_0: $\mu_1 = \mu_2$ at the $\alpha = 0.05$ level and conclude that adding the new ingredient to the paint significantly reduces the drying time. Alternatively, we can find the P-value for this test as

$$P\text{-value} = 1 - \Phi(2.52) = 0.0059$$

Therefore, H_0: $\mu_1 = \mu_2$ would be rejected at any significance level $\alpha \geq 0.0059$.

Exemplo 10-1



A estatística z_0 está dentro da região crítica e longe dos limites

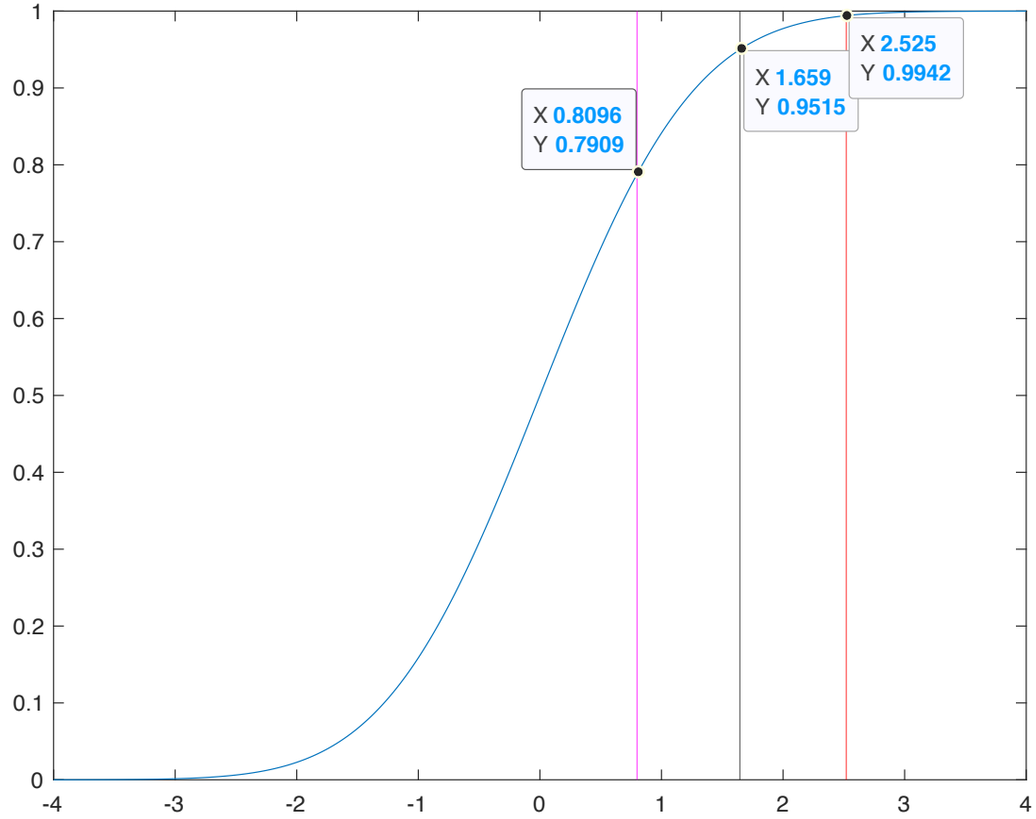
Exemplo 10-1

Se a média de $\bar{x}_2$ fosse 115 ao invés de 112

$$z_0 = \frac{121 - 115}{3.57} = 1.67$$

$$\phi(1.67) = normcdf(1.67,0,1) = 0.9525$$

valor-p = 1-0.9525 = 0.0475



Se a média de $\bar{x}_2$ fosse 118

valor-p = 1-0.7991 = 0.2

Intervalo de confiança para diferença de médias

If $\bar{x}_1$ and $\bar{x}_2$ are the means of independent random samples of sizes n_1 and n_2 from two independent normal populations with known variances σ_1^2 and σ_2^2, respectively, a **100(1 − α)% confidence interval for $\mu_1 - \mu_2$** is

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (10\text{-}7)$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

Exemplo 10-4

Testes de resistência à tração de longarinas de alumínio para aviões

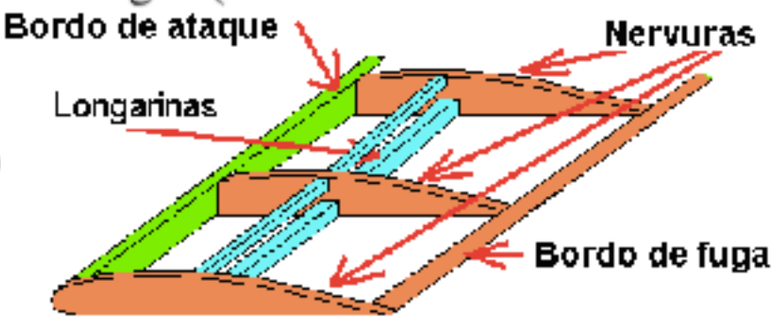
Tensile strength tests were performed on two different grades of aluminum spars used in manufacturing the wing of a commercial transport aircraft. From past experience with the spar manufacturing process and the testing procedure, the standard deviations of tensile strengths are assumed to be known. The data obtained are as follows: $n_1 = 10$, $\bar{x}_1 = 87.6$, $\sigma_1 = 1$, $n_2 = 12$, $\bar{x}_2 = 74.5$, and $\sigma_2 = 1.5$. If μ_1 and μ_2 denote the true mean tensile strengths for the two grades of spars, we may find a 90% confidence interval on the difference in mean strength $\mu_1 - \mu_2$ as follows:

$$\bar{x}_1 - \bar{x}_2 - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \bar{x}_1 - \bar{x}_2 + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$87.6 - 74.5 - 1.645\sqrt{\frac{(1)^2}{10} + \frac{(1.5)^3}{12}} \leq \mu_1 - \mu_2 \leq 87.6 - 74.5 + 1.645\sqrt{\frac{(1^2)}{10} + \frac{(1.5)^2}{12}}$$

Therefore, the 90% confidence interval on the difference in mean tensile strength (in kilograms per square millimeter) is

$$12.22 \leq \mu_1 - \mu_2 \leq 13.98 \text{ (in kilograms per square millim}$$



O que vêm agora?

Variância desconhecida

Comparar variância de duas populações

Teste de hipóteses: diferença de médias, variâncias desconhecidas

Caso 1: $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Neste caso, deve-se ter n_1 e n_2 maiores que 40.

Duas populações:

$X_{11}, X_{12}, \ldots, X_{1n_1}$ e $X_{21}, X_{22}, \ldots, X_{2n_1}$

$$E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2$$

$$V(\bar{X}_1 - \bar{X}_2) = V(\bar{X}_1) + V(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \sigma(\frac{1}{n_1} + \frac{1}{n_2})$$

The **pooled estimator** of σ^2, denoted by S_p^2, is defined by

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Ponderacão de S_1^2 eS_2^2 usando n_1 e n_2 para estimar σ

Teste de hipóteses: diferença de médias, variâncias desconhecidas

Substituindo S_p em

$$Z = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

A variável

$$T = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Tem distribuição t com $n_1 + n_2 - 2$ graus de liberdade.

Teste de hipóteses: diferença de médias, variâncias desconhecidas

Null hypothesis: $H_0: \mu_1 - \mu_2 = \Delta_0$

Test statistic:
$$T_0 = \frac{\bar{X}_1 - \bar{X}_2 - \Delta_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Alternative Hypothesis	Rejection Criterion
$H_1: \mu_1 - \mu_2 \neq \Delta_0$	$t_0 > t_{\alpha/2,n_1+n_2-2}$ or $t_0 < -t_{\alpha/2,n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 > \Delta_0$	$t_0 > t_{\alpha,n_1+n_2-2}$
$H_1: \mu_1 - \mu_2 < \Delta_0$	$t_0 < -t_{\alpha,n_1+n_2-2}$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Teste de hipóteses: diferença de médias, variâncias desconhecidas

Caso 2: $\sigma_1^2 \neq \sigma_2^2$

Usar estatística

$$T_0^* = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Com distribuição t e número de graus de liberdade dado por

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}}$$

Teste t pareado

Empregado quando as observações de duas populações de interesse são coletadas aos pares.

O que se quer é controlar fontes de variação que poderiam influenciar os resultados da comparação.

Exemplos de amostras pareadas:

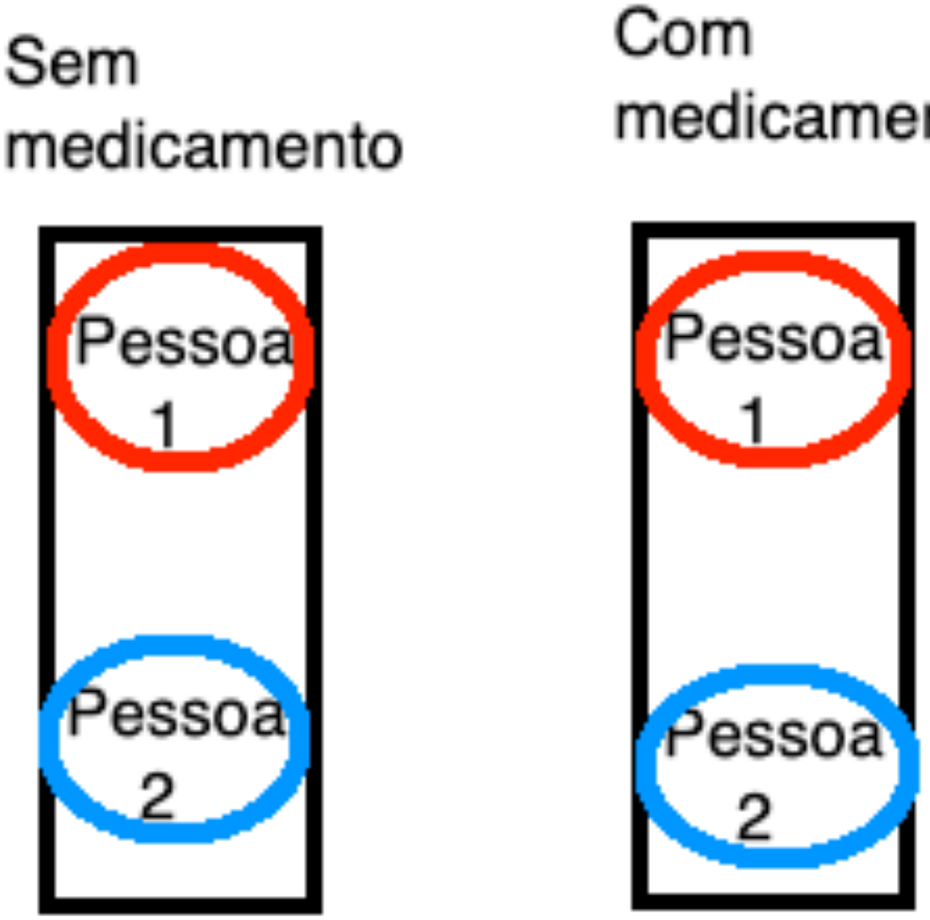
Estudo feito do antes e depois no mesmo indivíduo

Estudo feito com gêmeos

Não são consideradas estatisticamente independentes, pois as duas observações são mais prováveis de serem similares

Teste t pareado

Seja um teste com medicamento aplicado em diferentes pessoas. Se testar o medicamento comparando um grupo com e outro sem, a diferença entre as pessoas pode afetar o resultado



Teste t pareado

Exemplo: teste de um medicamento em 5 indivíduos, para verificar se uma dada droga baixa a taxa de colesterol (TC);

Individuo	1	2	3	4
TC antes	217	252	229	200
TC depois	209	241	230	208

Para cada indivíduo,

$$D_i = TC_{antes} - TC_{depois}, i = 1,2,3,4$$

A fonte de variação seria a diferença entre indivíduos

$$D_i = N(\mu_d, \sigma_d^2)$$

Teste t pareado

$X_1 : \mu_1, \sigma_1^2$

$X_2 : \mu_2, \sigma_2^2$

$D_j = X_{1j} - X_{2j}, j = 1,...,n$ normalmente distribuídos com média

$\mu_d = E(X_1 - X_2) = E(X_1) - E(X_2) = \mu_1 - \mu_2$

Teste t pareado

$\bar{D}$ é a média das diferenças

$D_1, D_2, \ldots, D_n$ e S_D é a variância dessas diferenças

$$H_0: \mu_D = \Delta_0$$
$$H_1: \mu_D \neq \Delta_0$$

Teste de hipótese:

Null hypothesis: $H_0: \mu_D = \Delta_0$

Test statistic: $$T_0 = \frac{\overline{D} - \Delta_0}{S_D/\sqrt{n}} \qquad (10\text{-}22)$$

Alternative Hypothesis	Rejection Region
$H_1: \mu_D \neq \Delta_0$	$t_0 > t_{\alpha/2,n-1}$ or $t_0 < -t_{\alpha/2,n-1}$
$H_1: \mu_D > \Delta_0$	$t_0 > t_{\alpha,n-1}$
$H_1: \mu_D < \Delta_0$	$t_0 < -t_{\alpha,n-1}$

Teste t pareado

Exemplo 10-9: predição da resistência ao cisalhamento de vigas (girder) de aço usando dois métodos

Table 10-2 Strength Predictions for Nine Steel Plate Girders (Predicted Load/Observed Load)

Girder	Karlsruhe Method	Lehigh Method	Difference d_j
S1/1	1.186	1.061	0.119
S2/1	1.151	0.992	0.159
S3/1	1.322	1.063	0.259
S4/1	1.339	1.062	0.277
S5/1	1.200	1.065	0.138
S2/1	1.402	1.178	0.224
S2/2	1.365	1.037	0.328
S2/3	1.537	1.086	0.451
S2/4	1.559	1.052	0.507

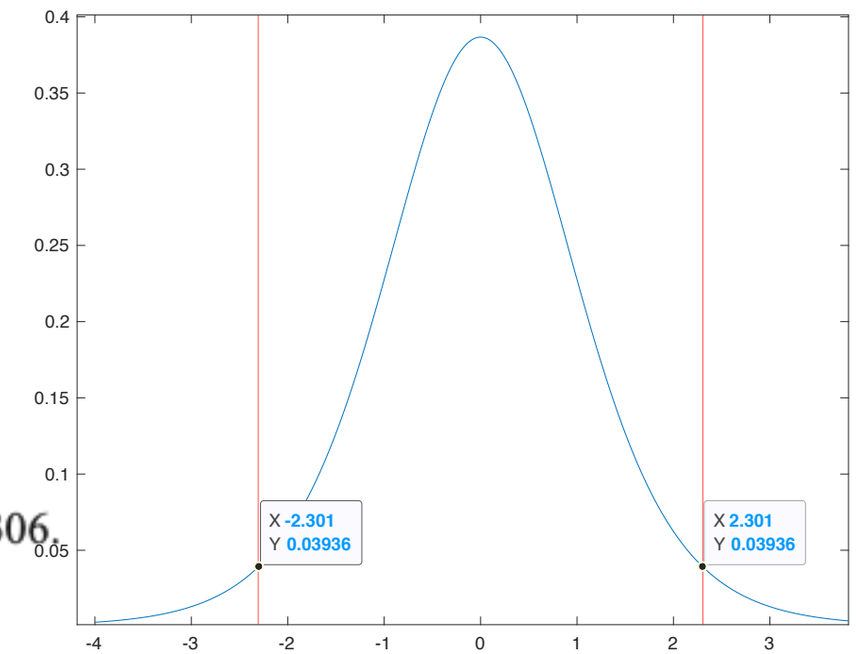
1. The parameter of interest is the difference in mean shear strength between the two methods, say, $\mu_D = \mu_1 - \mu_2 = 0$.
2. H_0: $\mu_D = 0$

Teste t pareado

Exemplo 10-9: predição da resistência ao cisalhamento de vigas (girder) de aço usando dois métodos

1. The parameter of interest is the difference in mean shear strength between the two methods, say, $\mu_D = \mu_1 - \mu_2 = 0$.
2. H_0: $\mu_D = 0$
3. H_1: $\mu_D \neq 0$
4. $\alpha = 0.05$
5. The test statistic is

$$t_0 = \frac{\bar{d}}{s_D/\sqrt{n}}$$



6. Reject H_0 if $t_0 > t_{0.025,8} = 2.306$ or if $t_0 < -t_{0.025,8} = -2.306$.

7. Computations: The sample average and standard deviation of the differences d_j are $\bar{d} = 0.2736$ and $s_D = 0.1356$, so the test statistic is

$$t_0 = \frac{\bar{d}}{s_D/\sqrt{n}} = \frac{0.2736}{0.1356/\sqrt{9}} = 6.05$$

$$\phi(6.05) = icdf(6.05,8) = 0.9998$$

valor-p=1.6e-04

Inferência sobre variância de duas populações normais

População 1: n_1 amostras, μ_1, σ_1^2 desconhecidos

População 2: n_2 amostras, μ_2, σ_2^2 desconhecidos

Variâncias amostrais: S_1^2, S_2^2

Desejamos testar a hipótese:

$$H_0: \sigma_1^2 = \sigma_2^2$$
$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Inferência sobre variância de duas populações normais

Distribuição F:

Define-se uma variável F com a razão de duas variáveis aleatórias com distribuição chi-quadrado divididas por seus graus de liberdade,

$$F = \frac{W/u}{Y/\nu}$$

Inferência sobre variância de duas populações normais

Distribuição F:

Let W and Y be independent chi-square random variables with u and v degrees of freedom, respectively. Then the ratio

$$F = \frac{W/u}{Y/v} \tag{10-26}$$

has the probability density function

$$f(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{2}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x+1\right]^{(u+v)/2}}, \qquad 0 < x < \infty \tag{10-27}$$

and is said to follow the F distribution with u degrees of freedom in the numerator and v degrees of freedom in the denominator. It is usually abbreviated as $F_{u,v}$.

Inferência sobre variância de duas populações normais

A média e variância da distribuição F são:

$$\mu = \frac{\nu}{\nu - 2}, \nu > 2$$

$$\sigma^2 = \frac{2\nu^2(u + \nu - 2)}{u(\nu - 2)^2(\nu - 4)}$$

Inferência sobre variância de duas populações normais

Distribuição F

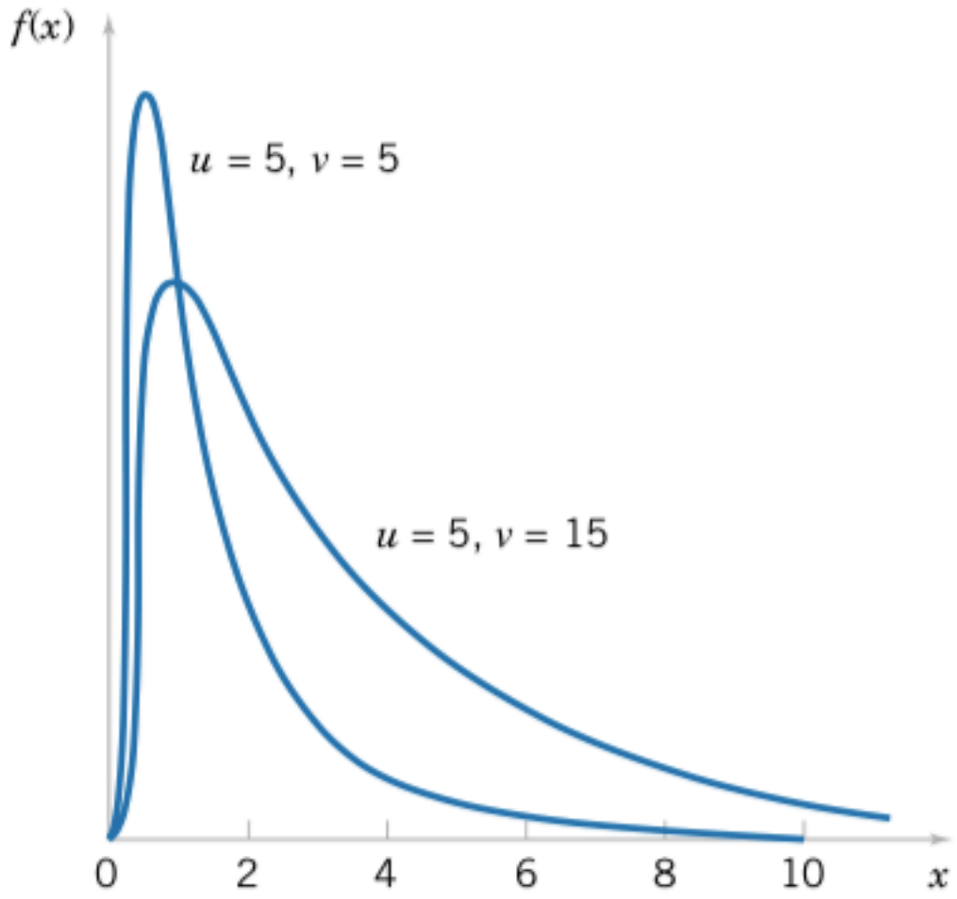


Figure 10-4 Probability density functions of two F distributions.

$$P(F > f_{\alpha,u,v}) = \int_{f_{\alpha,u,v}}^{\infty} f(x)\,dx = \alpha$$

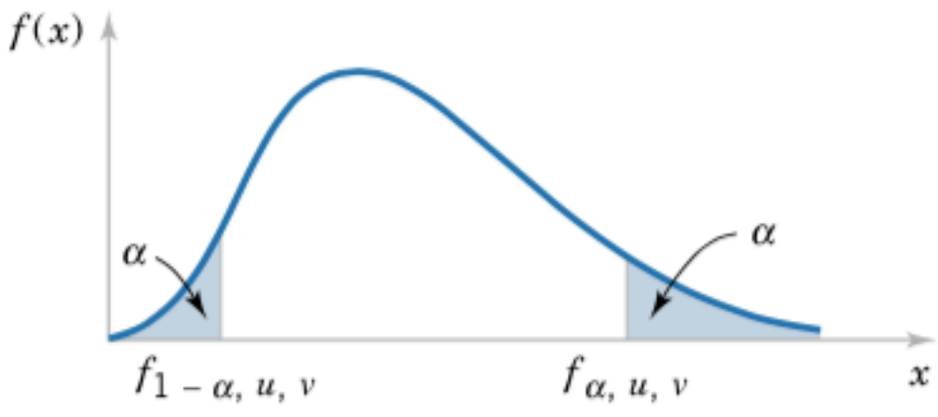


Figure 10-5 Upper and lower percentage points of the F distribution.

Teste de hipóteses da razão de duas variâncias

Let $X_{11}, X_{12}, \ldots, X_{1n_1}$ be a random sample from a normal population with mean μ_1 and variance σ_1^2, and let $X_{21}, X_{22}, \ldots, X_{2n_2}$ be a random sample from a second normal population with mean μ_2 and variance σ_2^2. Assume that both normal populations are independent. Let S_1^2 and S_2^2 be the sample variances. Then the ratio

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

has an F distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom.

Teste de hipóteses da razão de duas variâncias

Null hypothesis: $H_0: \sigma_1^2 = \sigma_2^2$

Test statistic: $F_0 = \dfrac{S_1^2}{S_2^2}$ (10-29)

Alternative Hypotheses	Rejection Criterion
$H_1: \sigma_1^2 \neq \sigma_2^2$	$f_0 > f_{\alpha/2,n_1-1,n_2-1}$ or $f_0 < f_{1-\alpha/2,n_1-1,n_2-1}$
$H_1: \sigma_1^2 > \sigma_2^2$	$f_0 > f_{\alpha,n_1-1,n_2-1}$
$H_1: \sigma_1^2 < \sigma_2^2$	$f_0 < f_{1-\alpha,\, n_1-1,n_2-1}$

Exemplo 10-11

Camadas de óxido são gravadas em semicondutores com uma mistura de gases para obter menor espessura da camada. Dois diferentes gases estão sendo testados e 20 camadas são gravadas em cada gás.

Os desvios padrão da espessura do óxido são: $s_1 = 1.96$ angstroms e $s_2 = 2.13$angstroms. Há evidências sobre qual gás é preferível? Considere $\alpha = 0.05$

Exemplo 10-11

1. The parameters of interest are the variances of oxide thickness σ_1^2 and σ_2^2. We will assume that oxide thickness is a normal random variable for both gas mixtures.
2. H_0: $\sigma_1^2 = \sigma_2^2$
3. H_1: $\sigma_1^2 \neq \sigma_2^2$
4. $\alpha = 0.05$
5. The test statistic is given by Equation 10-29:

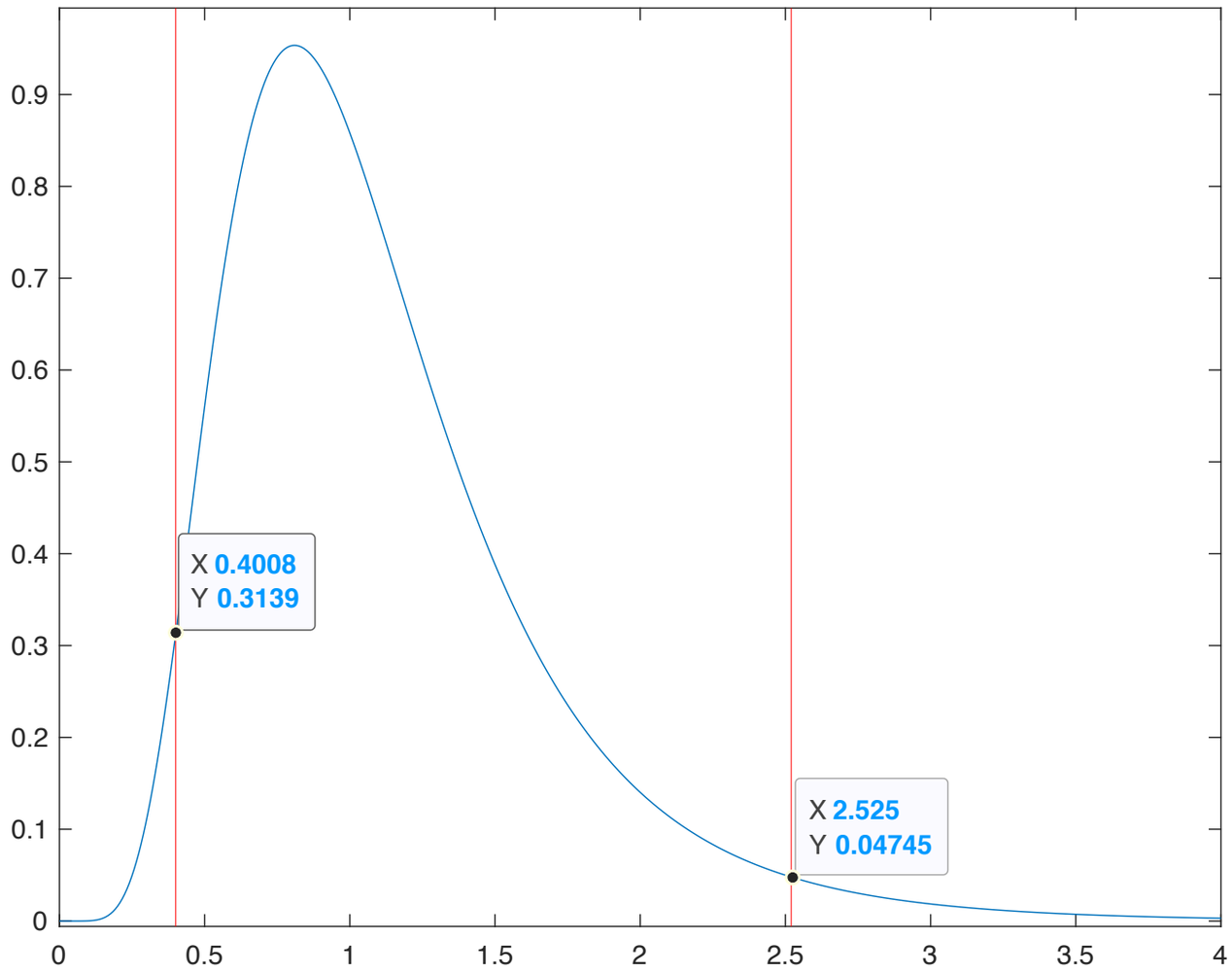
$$f_0 = \frac{s_1^2}{s_2^2}$$

finv(0.975,19,19)=2.5265

finv(0.025,19,19)=0.3958

6. Since $n_1 = n_2 = 20$, we will reject H_0: $\sigma_1^2 = \sigma_2^2$ if $f_0 > f_{0.025,19,19} = 2.53$ or if $f_0 < f_{0.975,19,19} = 1/f_{0.025,19,19} = 1/2.53 = 0.40$.

Exemplo 10-11



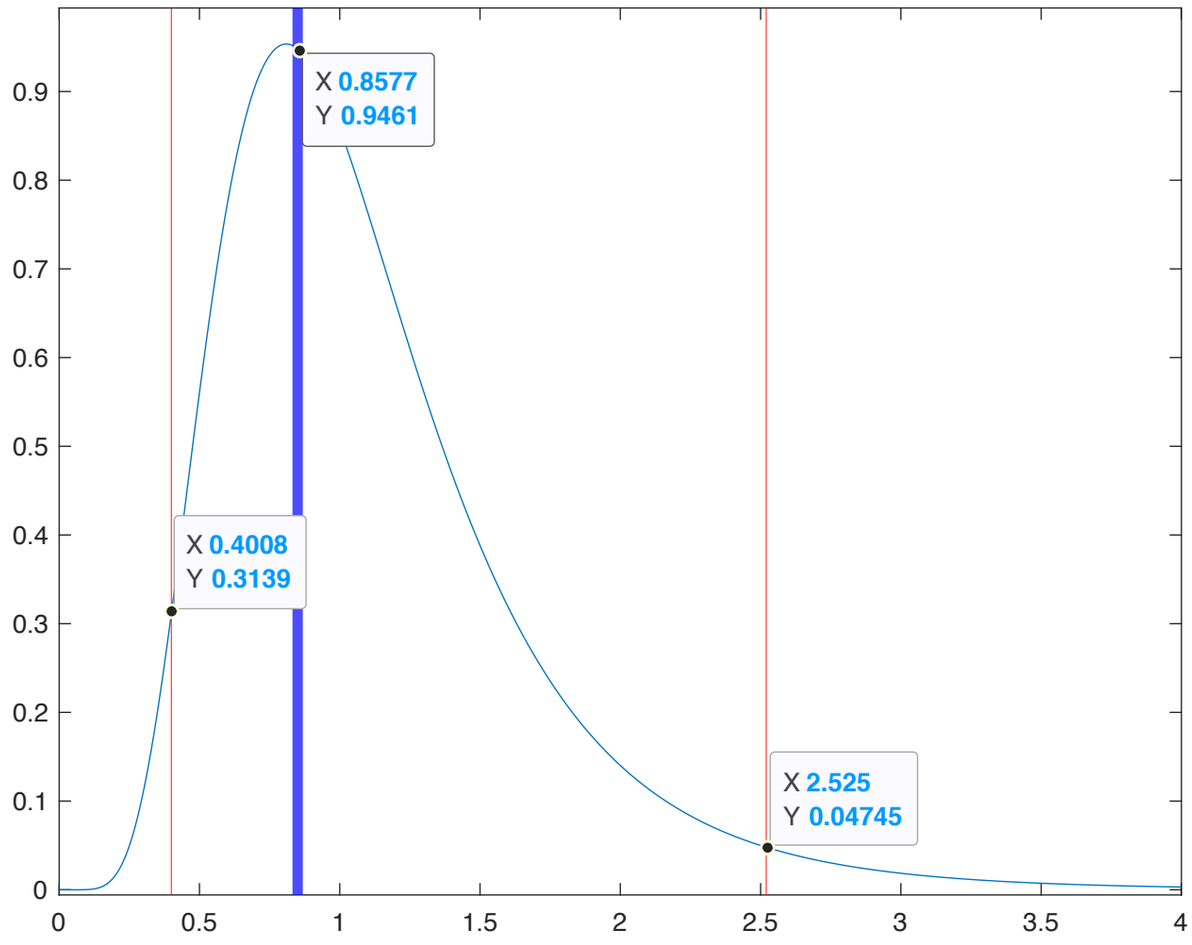
finv(0.975,19,19)=2.5265

finv(0.025,19,19)=0.3958

7. Computations: Since $s_1^2 = (1.96)^2 = 3.84$ and $s_2^2 = (2.13)^2 = 4.54$, the test statistic is

$$f_0 = \frac{s_1^2}{s_2^2} = \frac{3.84}{4.54} = 0.85$$

Exemplo 10-11



Não podemos rejeitar a hipótese

$$H_0 : \sigma_1^2 \neq \sigma_2^2 para \alpha = 0.05$$

7. Computations: Since $s_1^2 = (1.96)^2 = 3.84$ and $s_2^2 = (2.13)^2 = 4.54$, the test statistic is

$$f_0 = \frac{s_1^2}{s_2^2} = \frac{3.84}{4.54} = 0.85$$

Intervalo de confiança da razão de duas variâncias

If s_1^2 and s_2^2 are the sample variances of random samples of sizes n_1 and n_2, respectively, from two independent normal populations with unknown variances σ_1^2 and σ_2^2, then a **100(1 − α)% confidence interval on the ratio σ_1^2/σ_2^2** is

$$\frac{s_1^2}{s_2^2} f_{1-\alpha/2,n_2-1,n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2}{s_2^2} f_{\alpha/2,n_2-1,n_1-1} \tag{10-31}$$

where $f_{\alpha/2,n_2-1,n_1-1}$ and $f_{1-\alpha/2,n_2-1,n_1-1}$ are the upper and lower $\alpha/2$ percentage points of the F distribution with $n_2 - 1$ numerator and $n_1 - 1$ denominator degrees of freedom, respectively. A confidence interval on the ratio of the standard deviations can be obtained by taking square roots in Equation 10-31.