INTRODUÇÃO À INFERÊNCIA ESTATÍSTICA

Prof. Anderson Rodrigo da Silva
anderson.silva@ifgoiano.edu.br

Exemplo 1: cigarrinha das raízes

Nº ninfas/m de *M. frimbriolata* em 10 pontos amostrais de talhão com de cana-de-açúcar:

X = { 2 1 4 1 3 1 3 1 5 2}

- *Nível de controle = 3 ninfas/m*
- *Com base na amostra, realizar controle?*

Exemplo 2: produtividade de cana

Dados de produtividade (t/ha) de duas variedades de cana-de-açúcar (A e B), em seis talhões:

A = { 78 80 77 81 90 150 }
B = { 61 65 66 64 63 88 }

- *É possível diferenciar A de B utilizando média e desvio padrão?*
- *Seria razoável afirmar que elas diferem estatisticamente?*
- *Qual teste deveria ser aplicado?*

Exemplo 3: nematoides em soja

Dados de densidade populacional de *P. brachyurus* (indivíduos/100 cm^3 solo) antes e após cultivo de soja.

Pop. incial	Solo + Raiz (Soja R6)
66	284
128	914
54	925
171	887
357	742
986	771
100	1043

- *Houve aumento significativo (alfa = 5%)?*
- *Qual teste deveria ser aplicado?*
- *Há mais de uma forma apropriada de se testar o aumento populacional?*

Tipos de Pesquisa

- **Censo**: é o levantamento de toda população. Aqui não se faz inferência e sim uma descrição dos resultados.

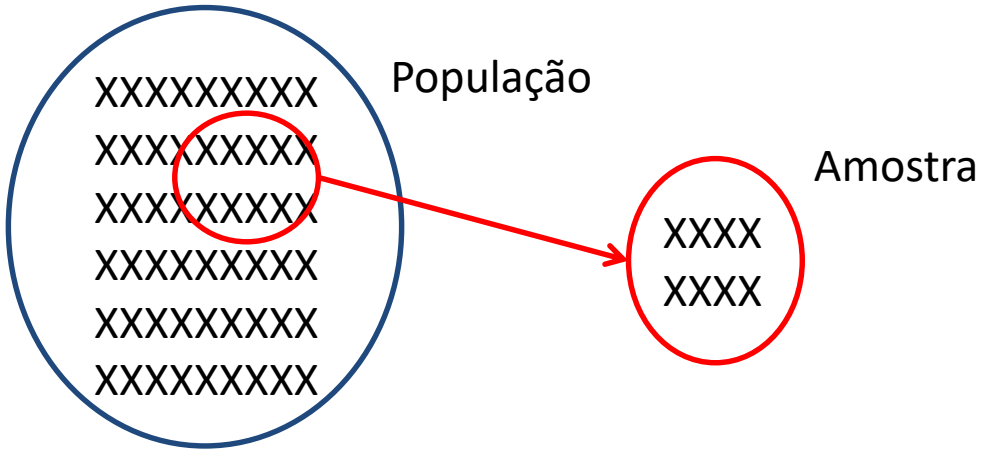
- **Amostragem:** coleta de observações sobre um grupo de indivíduos de uma população.

OBS.: **Inferência estatística** é o ato de inferir sobre o comportamento de uma população a partir do conhecimento da amostra por meio de um conjunto de métodos.

Conceitos importantes: População e Amostra

- **População ou universo estatístico**: é o conjunto de todos as possíveis unidades observacionais de uma variável.
  - Ex: Pesquisa sobre a composição do leite de vacas de uma fazenda. População: todas as vacas leiteiras da fazenda.

- **Amostra:** é uma parte ou subconjunto da população. Obs.: Em geral toma-se a amostra para estudar (inferir sobre) a população.



Conceitos importantes: Parâmetro e Estimador

- **Parâmetro**: é uma quantidade desconhecida (geralmente) que caracteriza a população, tal como a média ou a variância populacional.

- **Estimador**: é uma regra ou método de estimar um parâmetro. Geralmente uma fórmula. Um valor particular assumido pelo estimador em uma dada amostra é uma *estimativa*.

Parâmetros

$$\mu_X$$
$$\sigma_X^2$$
$$\sigma_X$$

Estimadores

$$\bar{x} = \sum x / n$$
$$s^2 = \sum (x - \bar{x})^2 / (n - 1)$$
$$s = \sqrt{s^2}$$

Importância

- Permite fazer mais que uma simples descrição da amostra.

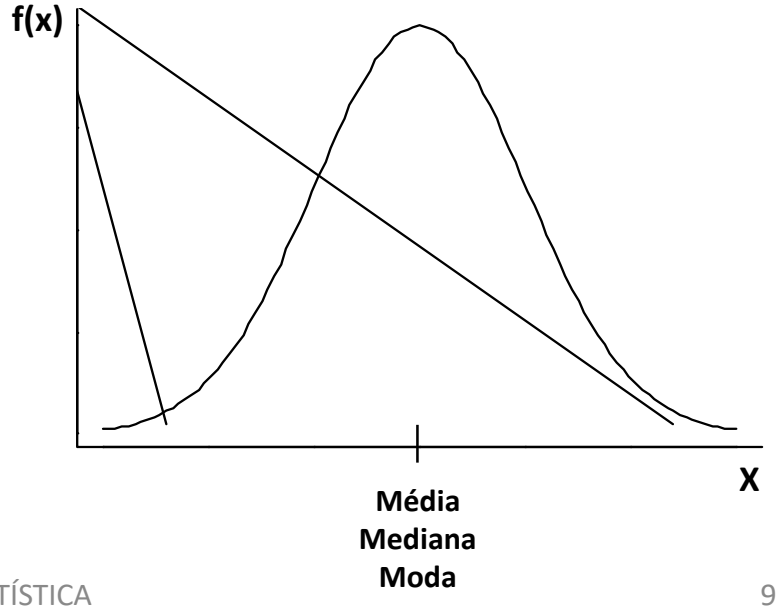
- Permite obter probabilidades exatas da ocorrência de certos valores devido ao acaso.

- **Exemplo:** se a média amostral do grau de toxicidade em gramíneas por certo herbicida é de 20%, com desvio padrão de 2%, qual a probabilidade de encontrar uma parcela experimental com escore de toxicidade acima de 23%?

- Admitindo que esta variável tenha distribuição normal, podemos calcular essa probabilidade por meio do modelo de distribuição normal.

Distribuição Normal

- Grande importância em inferência estatística
- A distribuição de probabilidades de uma variável aleatória normal tem a forma de sino, sendo simétrica em torno da média.
- Uma variável contínua X normalmente distribuída, é completamente caracterizada pela sua média (μ) e pela sua variância (σ^2).
- O domínio da distribuição é: $(-\infty, \infty)$.



Distribuição Normal

Um resultado importante:

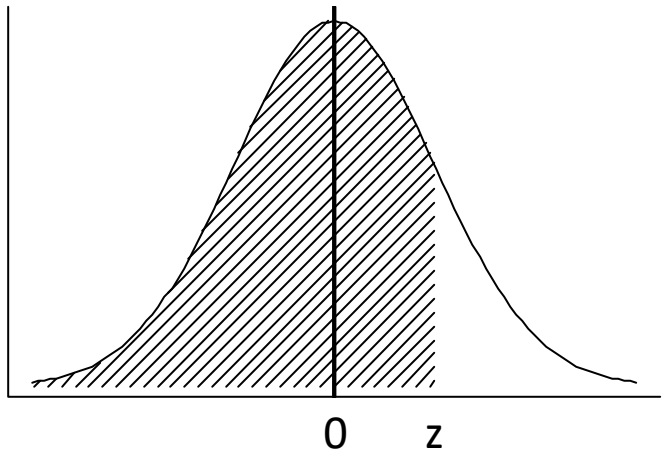
$$X \sim \text{Normal}(\mu, \sigma^2) \quad \Rightarrow \quad Z = \frac{X - \mu}{\sigma} \sim \text{Normal}(0, 1)$$

Que nos permite usar valores tabelados da **distribuição normal padrão (Z)** para calcular probabilidades associadas a valores de X.

Tabela da distribuição Normal Padrão

Z	0.00	0.01	0.02	0.03	...
0.0	0.5000	0.5040	0.5080	0.5120	
0.1	0.5398	0.5438	0.5478	0.5517	
0.2	0.5793	0.5832	0.5871	0.5910	
0.3	0.6179	0.6217	0.6255	0.6293	
0.4	0.6554	0.6591	0.6628	0.6664	
0.5	0.6915	0.6950	0.6985	0.7019	
0.6	0.7257	0.7291	0.7324	0.7357	
0.7	0.7580	0.7611	0.7642	0.7673	
0.8	0.7881	0.7910	0.7939	0.7967	
0.9	0.8159	0.8186	0.8212	0.8238	
1.0	0.8413	0.8438	0.8461	0.8485	
1.1	0.8643	0.8665	0.8686	0.8708	
1.2	0.8849	0.8869	0.8888	0.8907	
1.3	0.9032	0.9049	0.9066	0.9082	
1.4	0.9192	0.9207	0.9222	0.9236	
1.5	0.9332	0.9345	0.9357	0.9370	
1.6	0.9452	0.9463	0.9474	0.9484	
1.7	0.9554	0.9564	0.9573	0.9582	
...					

Cada célula na tabela dá a proporção acumulada sob a curva até um valor z.

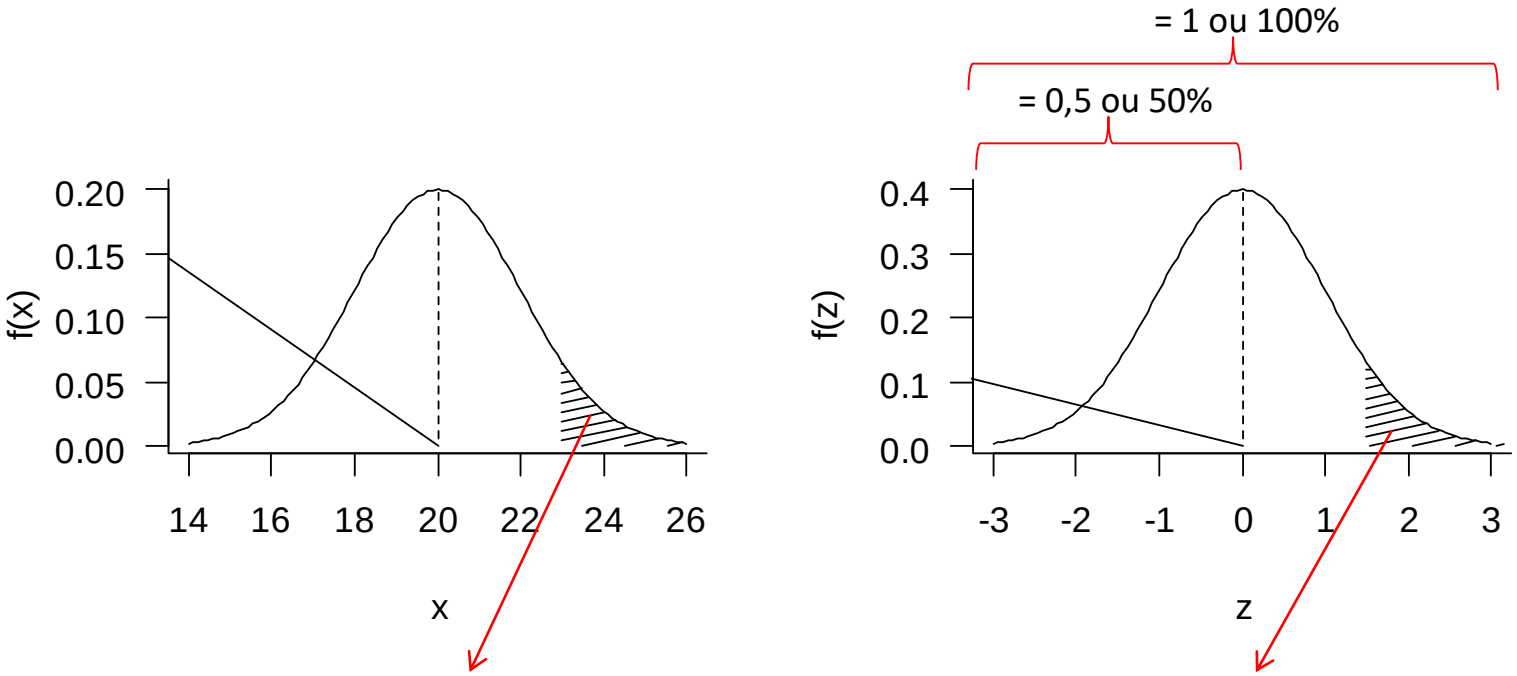


Valores obtidos com a função **=DIST.NORMP()** do Excel.

Cálculo de probabilidades a partir da distribuição Normal: exemplo

Se X ~ Normal(μ = 20, σ = 2), qual a probabilidade de obter um valor superior a 23?

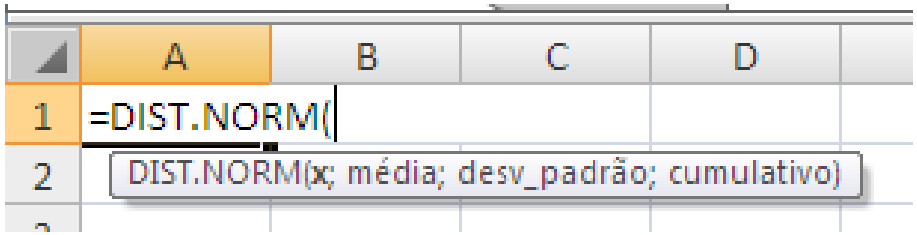
$$X = 23 \rightarrow Z = \frac{23-20}{2} = 1{,}5$$

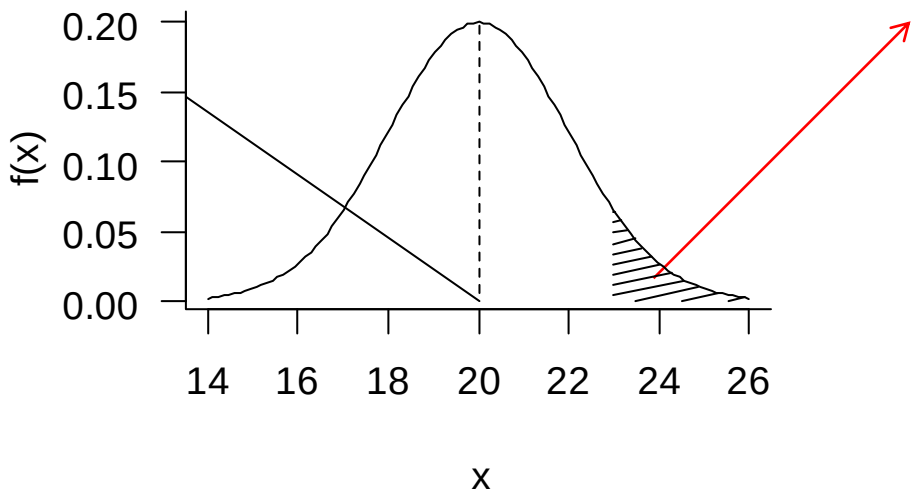


$$P(X > 23) = P(Z > 1{,}5) = 1 - 0{,}9332 = 0{,}0668 \quad \text{ou } 6{,}68\%$$

Exemplo

Usando o Excel ...





=DIST.NORM(23; 20; 2; FALSO)
0,0648
ou 6,48%

Agora calcule: $P(17 < X < 22)$

Testes de Hipóteses

Suponha que você é o agrônomo encarregado pelo manejo de pragas de uma usina de cana-de-açúcar. Admita que o nível de controle da broca-da-cana (Diatrea sacchralis) é de 3% de infestação (lagartas recém-eclodidas). Para tal, um monitoramento deve ser feito, amostrando 13 pontos de raio 2 m num talhão e computando o percentual de internódios atacados.

Como estabelecer um critério para saber se é preciso realizar algum tipo de controle, isto é, se em média o nível de infestação é igual ou superior a 3%?

Testes de Hipóteses

- As principais áreas da inferência estatística são: estimação de parâmetros, cálculos de probabilidade e testes de significância ou *testes de hipóteses*.

- Hipótese é uma afirmação sobre a população. Uma suposição quanto a um parâmetro desta ou quanto a forma da população. Exemplos:
  - A média populacional da produtividade de alho é 10 t ha^{-1}
  - A proporção de plantas de cana-de-açúcar infestadas com a broca gigante numa usina é 0,05.

- O objetivo de um teste de hipótese é construir uma regra que permita validar ou rejeitar uma hipótese através dos resultados da amostra.

Testes de Hipóteses

Todo teste de hipótese é baseado em duas hipóteses:

- Hipótese de nulidade ou afirmativa (H_0). Exemplos:
  - H_0: $\mu = 10$ t ha^{-1}
  - H_0: $p = 0,05$

- Hipótese alternativa (H_1). Exemplos:
  - H_1: $\mu \neq 10$ t ha^{-1} (bilateral) ou,
  - H_1: $\mu > 10$ t ha^{-1} (unilateral à direita) ou,
  - H_1: $\mu < 10$ t ha^{-1} (unilateral à esquerda)

  - H_1: $p \neq 0,05$
  - H_1: $p > 0,05$
  - H_1: $p < 0,05$

Estatísticas de teste para 1 média

O teste de hipótese do tipo H_0: $\mu = \mu_0$ é feito por meio de uma das seguintes estatísticas de teste:

1) Caso em que se conhece a variância (σ^2).

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

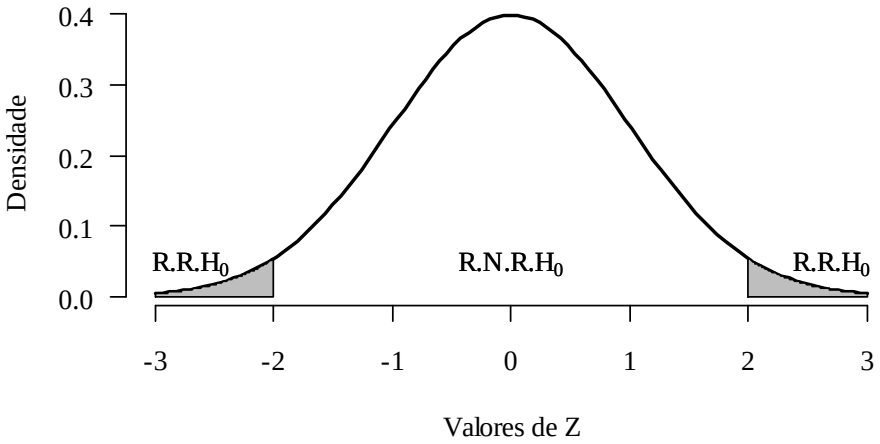
2) Caso em que não se conhece a variância (σ^2).

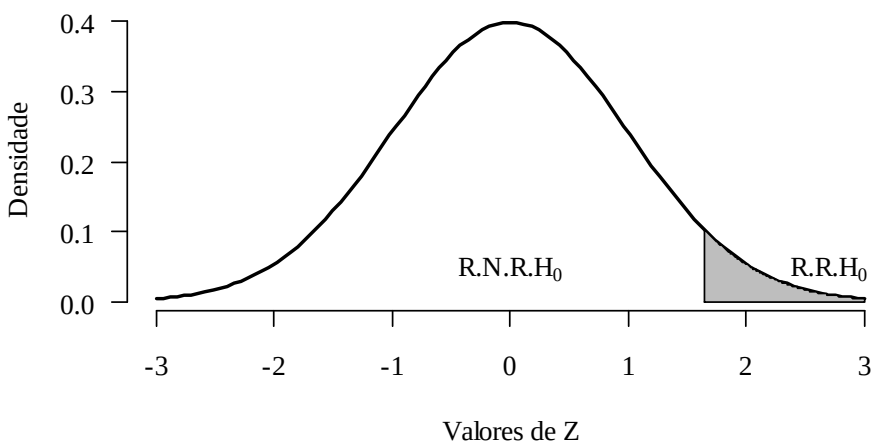
$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}} \sim t(n - 1 \text{ g.l.})$$

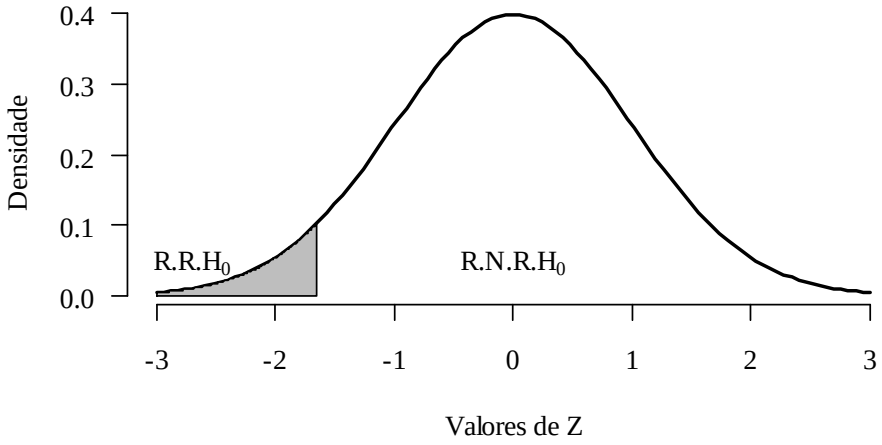
Erros associados aos testes de hipóteses

- **Erro tipo I (alfa):** é caracterizado pelo fato de rejeitar H_0 quando esta é verdadeira.
- **Erro tipo II (beta):** erro tipo II é caracterizado pelo fato de aceitar H_0 quando esta é falsa.
- Alfa e beta são inversamente relacionados e não é possível fazer o controle de ambos ao mesmo tempo. Prioriza-se um deles, modificando o erro tipo I.
- Em geral, os valores adotados para alfa são: 0,01, 0,05 ou 0,10.
- A quantidade ***1 – alfa*** é conhecida como *nível de confiança* do teste.

Região crítica ou de rejeição de H_0







Os 5 passos para executar um teste de hipóteses

1. Enunciar as hipóteses H_0 e H_1
2. Identificar a estatística de teste
3. Fixar o limite de erro alfa e a região crítica do teste
4. Com os dados amostrais, calcular a estatística de teste
5. Concluir pela aceitação ou rejeição de H_0 pela comparação do valor obtido no passo (4) com a RC do passo (3)

Exemplo

i	X
1	2.61
2	2.76
3	3.06
4	2.91
5	3.06
6	2.94
7	2.97
8	2.73
9	3.24
10	3.09
11	2.73
12	3.18
13	2.94
média	***2.94***
desvio	***0.189***

Broca-da-cana: Admita que a variável (X) nível de infestação tem distribuição normal. Avalie as hipótestes:

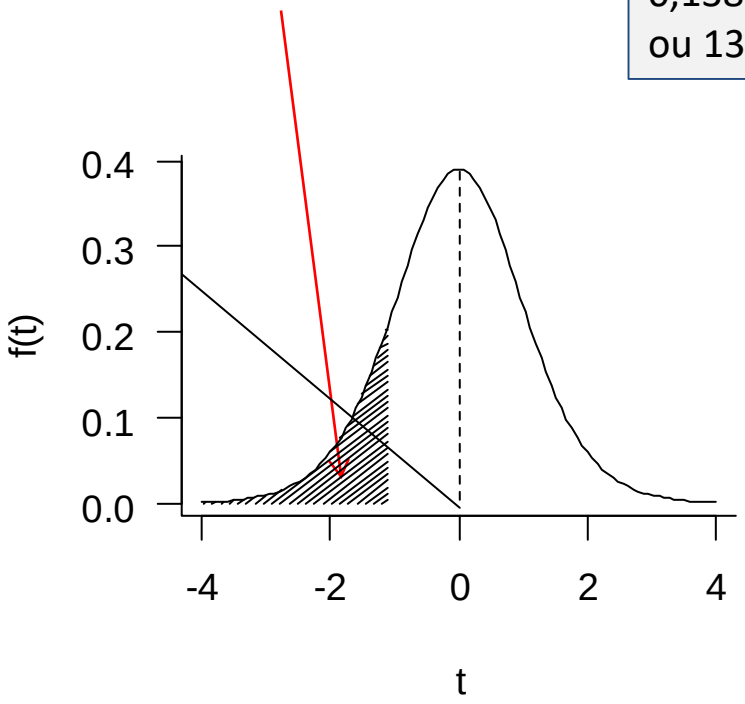
$$Hipóteses: \begin{cases} H_0 : \mu = 3 \\ H_1 : \mu < 3 \end{cases}$$

Exemplo

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{2{,}94 - 3}{0{,}189/\sqrt{13}} = -1{,}14 \quad \sim \quad t(13-1)$$

$$P(X < 2{,}94) = P(t < -1{,}14) = 0{,}138$$

=DISTT(1.14, 12, 1)
0,138
ou 13,8%



Valor-p !

Conceito de valor-p

- O **valor-p** quantifica o quão estranho é o resultado da amostra, supondo H_0 verdadeira. Para o exemplo 2, estamos supondo que o nível de infestação é de 3%.

- Dado o resultado da amostra e a distribuição sob H_0, calcula-se o valor-p computando a probabilidade de ocorrer um resultado tão ou mais extremo do que aquele que efetivamente ocorreu.

Intervalo de confiança para a média

Considerando ainda que $\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$

Para um dado coeficiente de confiança (1 - α), existem quantidades q1 e q2 tais que

$$P\left(\overline{X} - \frac{q_2 \sigma}{\sqrt{n}} < \mu < \overline{X} - \frac{q_1 \sigma}{\sqrt{n}} \right) = 1 - \alpha$$

Em que: $\begin{cases} q_1 = -t_{\frac{\alpha}{2}}(n-1 \text{ g.l.}) \\ q_2 = t_{\frac{\alpha}{2}}(n-1 \text{ g.l.}) \end{cases}$

Intervalo de confiança para a média

Assim,

$$IC(\mu)_{1-\alpha} = \overline{X} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$$

Como veremos no exemplo a seguir, um IC tem uma relação direta com um teste de hipóteses.

IC para o exemplo

Relembrando:

$$\begin{cases} \bar{x} = 2{,}94\ \% \\ s = 0{,}189\ \% \\ n = 13 \end{cases}$$

Construindo um IC com 95% de confiança para a verdadeira média (μ):

$$100(1-\alpha)\% = 95 \Rightarrow \alpha = 0{,}05$$

Como será utilizado o desvio padrão amostral, devemos utilizar os quantis da distribuição t-Student para construir o IC.

$$t_{\frac{\alpha}{2},(n-1)} = t_{\frac{0{,}05}{2},\,(13-1)} = 2{,}18$$

No Excel...
=INVT(0.05; 12)
2.18

IC para o exemplo

$$IC(\mu)_{1-\alpha} = \bar{x} \pm t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}} = 2{,}94 \pm 2{,}18\frac{0{,}189}{\sqrt{13}}$$
$$= 2{,}94 \pm 0{,}114$$

$$3{,}054 < \mu < 2{,}826$$

Dado o nível de 95% de confiança, é possível afirmar que μ difere de 3?

Interpretação:
Aproximadamente 95% das médias amostrais devem estar entre 3,054 e 2,826%.

Dimensionamento do tamanho amostral

Suponha que o seu grau de tolerância para com a média do nível de infestação seja Δ = 0,1%, para mais ou para menos. Para detectar tal diferença com probabilidade de erro de, no máximo, 5%, quantos pontos deve o agrônomo amostrar?

$$n \geq \frac{1{,}96^2 \times s^2}{\Delta^2}$$