INTRODUÇÃO À INFERÊNCIA ESTATÍSTICA

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

Tipos de Pesquisa

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

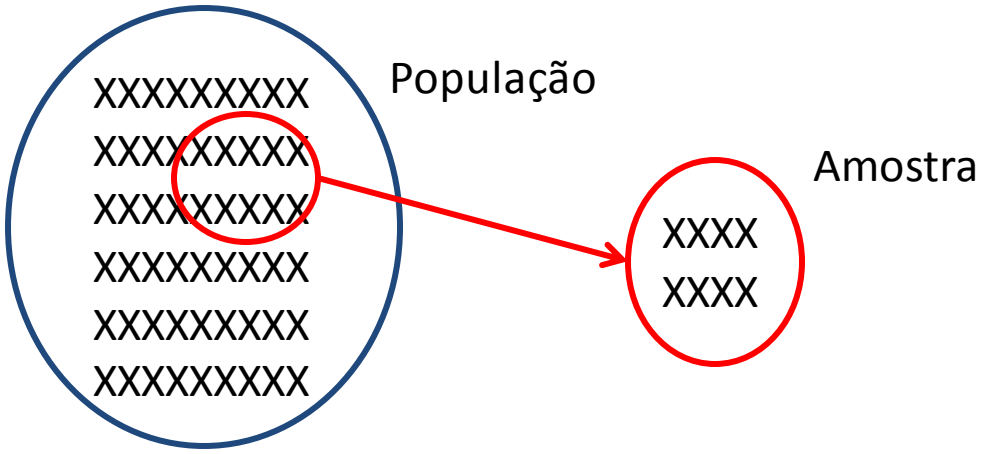
- **Amostragem:** coleta de observações sobre um grupo de indivíduos de uma população.

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

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

1. Amostragem

Importância da amostragem

- Torna possível fazer afirmações sobre características da população, com base nos resultados da amostra.

- Para inferências confiáveis, a amostra precisa ser representativa da população da qual foi retirada.

- Em relação ao censo, o processo de amostragem representa duas vantagens: redução de tempo e custos.

Tipos de Amostragem

- **Probabilísticas:** a seleção é aleatória de tal forma que cada elemento da população tem uma probabilidade conhecida. Ex. N é o tamanho da população e 1/N é a probabilidade de cada elemento participar da amostra.

- **Não probabilísticas ou intencionais:** há escolha deliberada dos elementos da amostra. Geralmente, amostras intencionais são usadas em alguns tipos de pesquisa de mercado.

OBS.: Para se fazer *inferências estatísticas*, há necessidade de que o processo seja *probabilístico*! Isso para que se possa avaliar a probabilidade de erro.

Amostragem probabilística

Alguns tipos de amostragem probabilística mais conhecidos são:

- **Simples ao acaso**
- **Sistemática**
- **Estratificada**

Amostragem simples ao acaso

- Processo bastante fácil e muito usado

- Todos os elementos da população tem igual probabilidade de serem escolhidos

- Procedimento:
  - Numerar todos os elementos da população (criar um índice). Ex.: se a população tem N=1000 elementos, atribuímos um índice variando de 0 a 999 ou 1 a 1000 aos seus elementos.
  - Efetuar sucessivos sorteios com reposição até completar o tamanho da amostra (n).

OBS.: Para realizar o sorteio dos elementos, podemos utilizar as *tabelas de números aleatórios* ou um programa computacional, tal como o *Excel* com a função =ALEATÓRIOENTRE().

Amostragem simples ao acaso

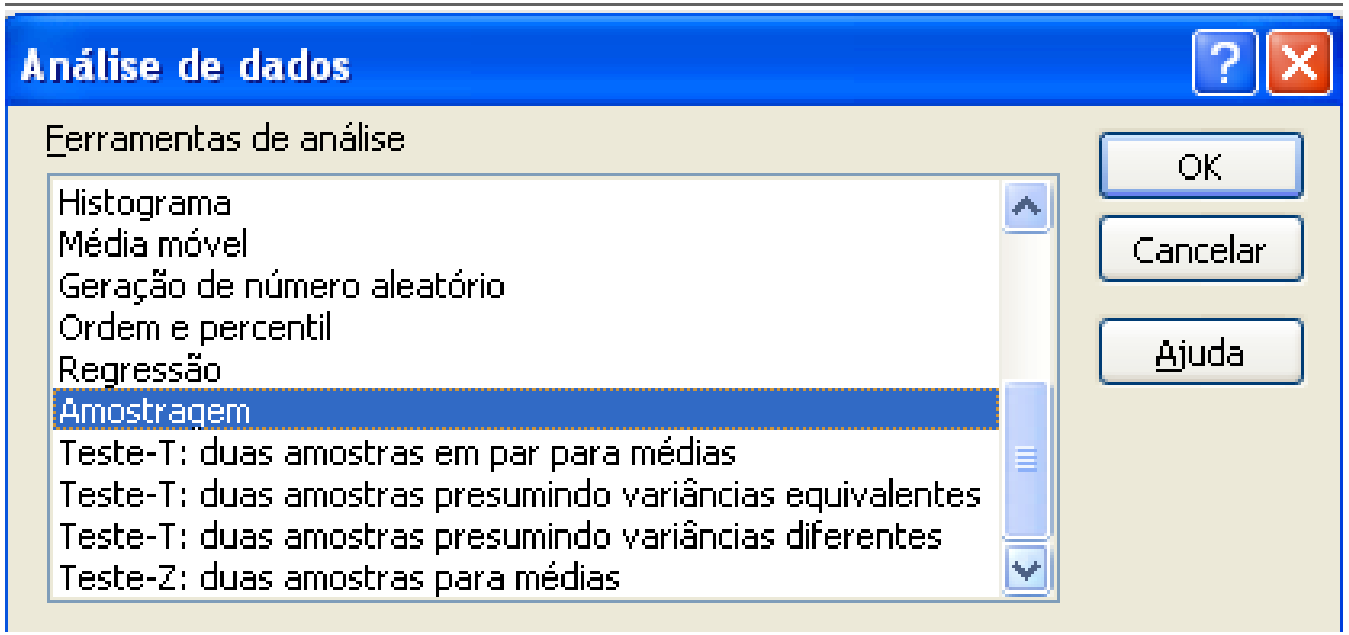
O7

	A	B	C	D	E
1	índice pop.	pop.	selecionado	amostra	
2	1	4.70	6	6.73	
3	2	3.72	14	4.02	
4	3	5.24	6	6.73	
5	4	6.28	9	6.10	
6	5	6.20	10	3.91	
7	6	6.73			
8	7	2.82			
9	8	4.77			
10	9	6.10			
11	10	3.91			
12	11	4.31			
13	12	3.31			
14	13	3.15			
15	14	4.02			
16	15	4.23			
17	16	2.88			
18	17	4.43			
19	18	4.60			
20	19	5.13			
21	20	4.63			
22					

Tamanho da população: N = 20
Tamanho da amostra: n = 5
p = 1/20 = 0,05

=ALEATÓRIOENTRE(1;20)

Amostragem simples ao acaso usando o *Excel*



Amostragem sistemática

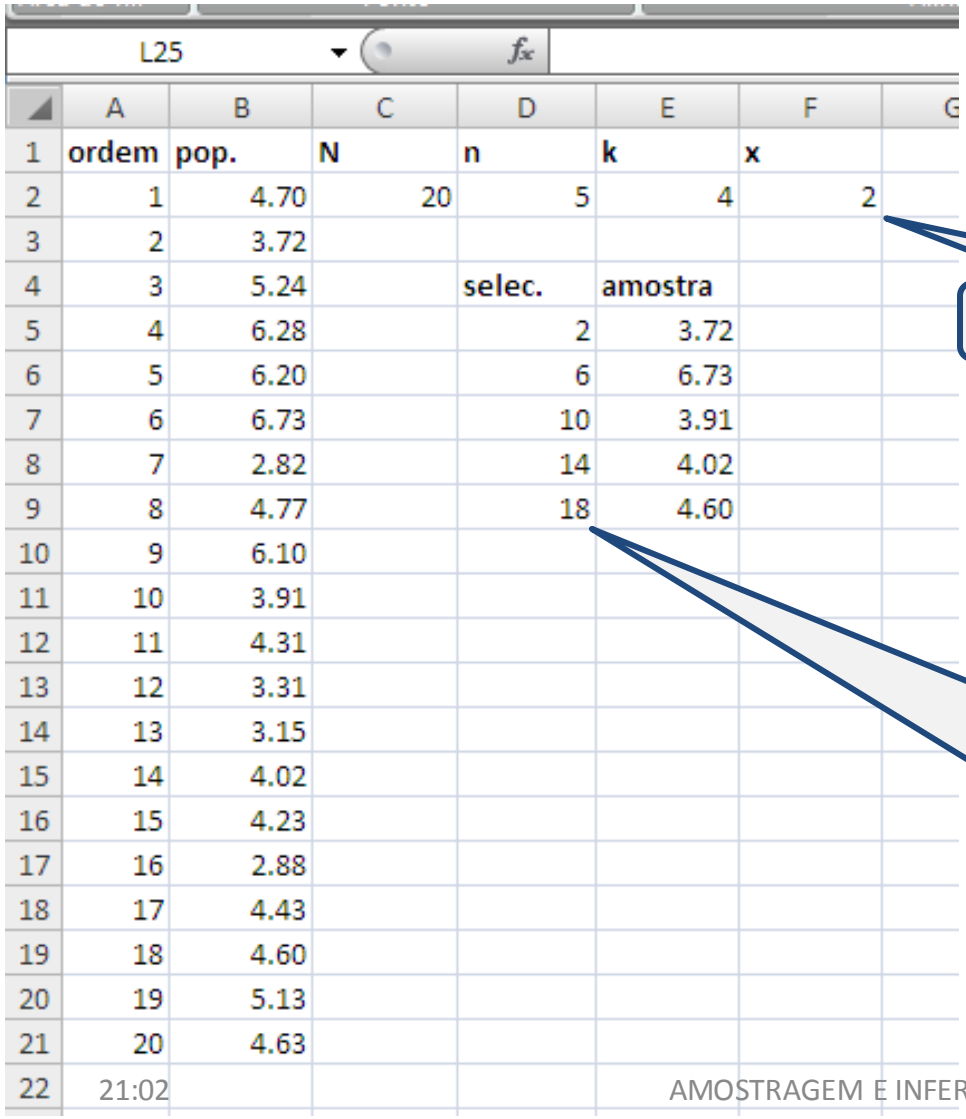
- É uma variação da amostragem simples ao acaso

- Conveniente quando a população está naturalmente ordenada, como fichas em um fichário, listas telefônicas etc.

- Procedimento:
  - Determina-se: N o tamanho da população e; n o tamanho da amostra.
  - Calcula-se o intervalo (ou período) da amostragem: **k = N/n** ou o inteiro mais próximo.
  - Sorteia-se um inteiro x entre 1 e k.
  - Forma-se a amostra dos elementos correspondentes aos números:

    x, x + k, x + 2k, ..., x + (n - 1)k.

Amostragem sistemática



L25

	A	B	C	D	E	F	G
1	ordem	pop.	N	n	k	x	
2	1	4.70	20	5	4	2	
3	2	3.72					
4	3	5.24		selec.	amostra		
5	4	6.28		2	3.72		
6	5	6.20		6	6.73		
7	6	6.73		10	3.91		
8	7	2.82		14	4.02		
9	8	4.77		18	4.60		
10	9	6.10					
11	10	3.91					
12	11	4.31					
13	12	3.31					
14	13	3.15					
15	14	4.02					
16	15	4.23					
17	16	2.88					
18	17	4.43					
19	18	4.60					
20	19	5.13					
21	20	4.63					
22	21:02						

Tamanho da população: N = 20
Tamanho da amostra: n = 5
k = 20/5 = 4

=ALEATÓRIOENTRE(1;k)

x = 2
x + k = 6
x + 2k = 10
x + 3k = 14
x + 4k = 18

Amostragem estratificada

- Indicado quando temos uma população heterogênea, na qual podemos distinguir subpopulações homogêneas (estratos).

- Estratificar a população significa dividi-la em S estratos mutuamente exclusivos, tais que: $n_1 + n_2 + ... + n_S = n$ (tamanho da amostra).

- Procedimento:
  - Determina-se os S estratos
  - Seleciona-se uma amostra aleatória (simples ao acaso) de cada estrato.
  - O tamanho das amostras de cada estrato pode ou não ser proporcional ao tamanho do estrato. Em caso positivo, temos a estratificação proporcional.

Estratificação proporcional

B25

	A	B	C	D	E
1	índice pop.	pop.	estrato	amostra	
2	1	9.46	1	8.33	
3	2	10.42	1	10.42	
4	3	9.83	1		
5	4	8.33	1		
6	5	9.20	1		
7	6	11.26	1		
8	7	9.38	1		
9	8	9.99	1		
10	9	6.78	2	8.99	
11	10	8.06	2	8.22	
12	11	8.61	2	8.22	
13	12	8.99	2		
14	13	6.30	2		
15	14	8.22	2		
16	15	8.56	2		
17	16	7.67	2		
18	17	6.23	2		
19	18	8.00	2		
20	19	8.64	2		
21	20	8.46	2		
22					

Tamanho da população: N = 20
Tamanho da amostra: n = 5

$N_1 = 8$
$p_1 = 8/20 = 0,40$
$n_1 = 5 \times 0,40 = 2$

$N_2 = 12$
$p_2 = 12/20 = 0,60$
$n_2 = 5 \times 0,60 = 3$

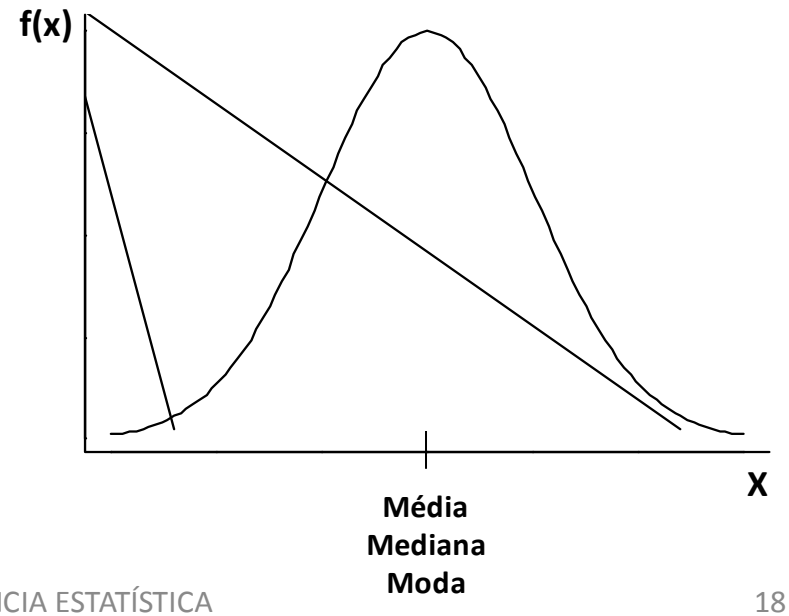
2. Inferência estatística

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% e o desvio padrão é 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:

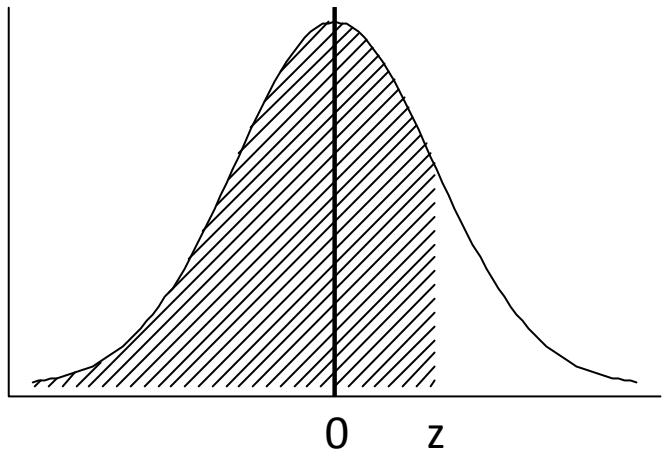
$$X \sim \text{Normal}(\mu, \sigma^2) \quad \Rightarrow \quad Z = \frac{X - \mu}{\sigma} \sim \text{Normal}(0, 1)$$

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.

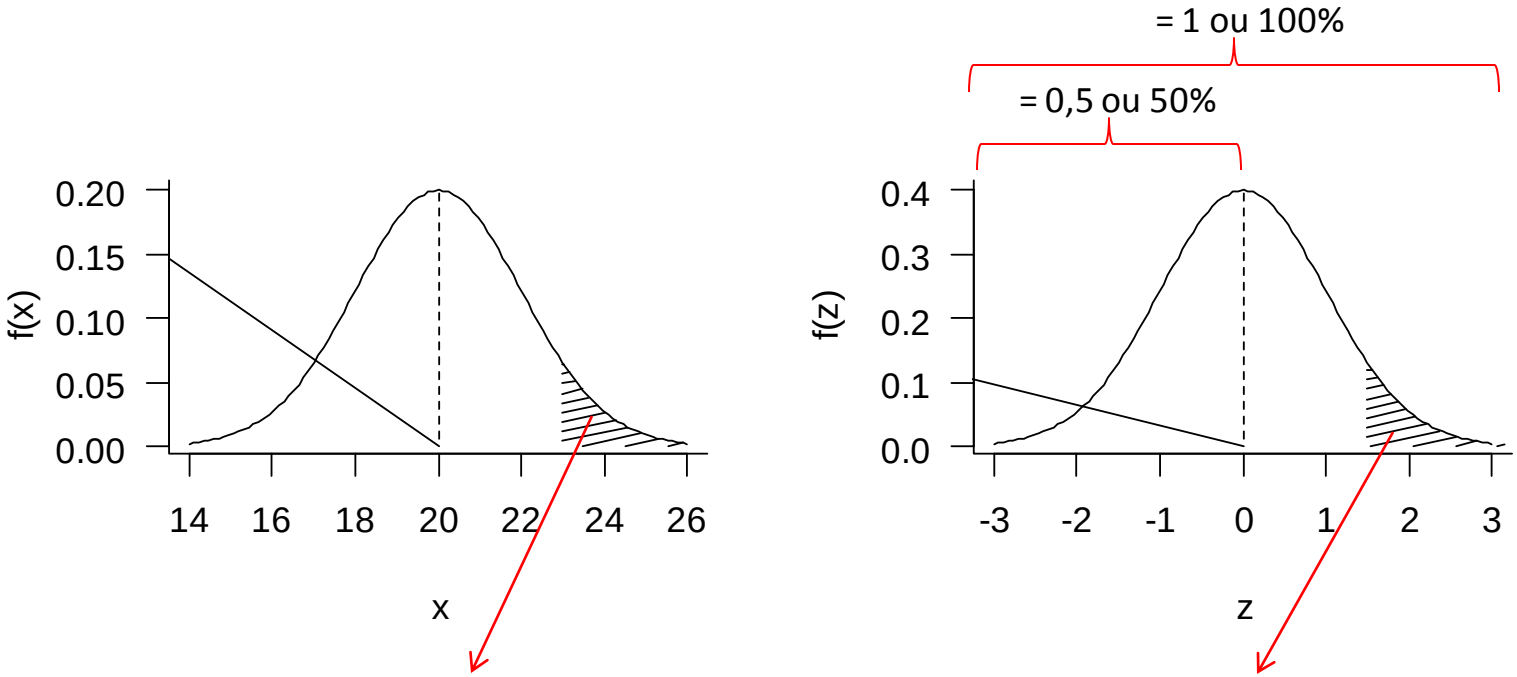


Valores obtidos com a função **=DIST.NORMP()** do Excel.

Cálculo de probabilidades a partir da distribuição Normal: exemplo 1

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

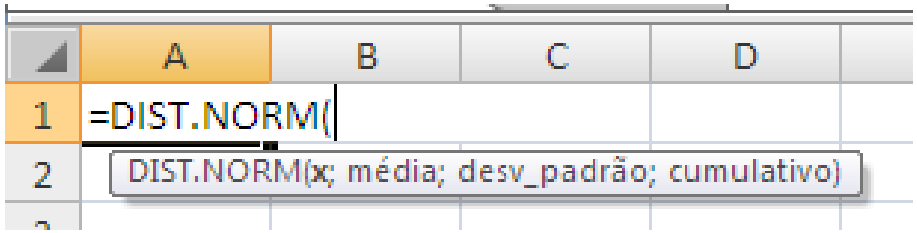
$$X = 23 \rightarrow Z = \frac{23 - 20}{2} = 1{,}5$$

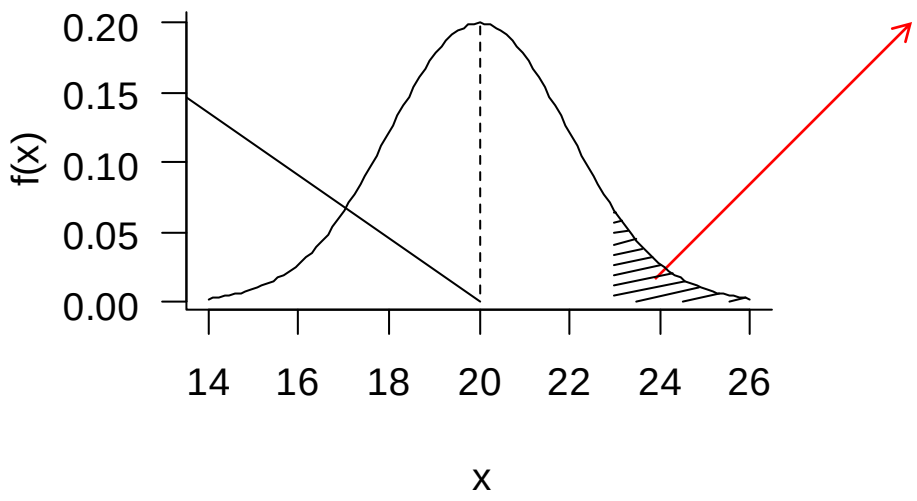


$$P(X > 23) = P(Z > 1{,}5) = 1 - 0{,}9332 = 0{,}0668 \quad \text{ou } 6{,}68\%$$

Exemplo 1

Usando o Excel ...





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

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

Distribuição da média amostral

Se X é uma variável com **distribuição normal** de média μ e variância σ^2, então dada uma amostra de tamanho n desta variável, temos que:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \text{ é um estimador de } \mu$$

$$X \sim \text{Normal}(\mu, \sigma^2) \quad \Rightarrow \quad \bar{x} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

Conhecendo a distribuição da média amostral da variável, podemos calcular probabilidades exatas e fazer inferências sobre o parâmetro μ.

Testes de Hipóteses

Suponha que o nível crítico (dano econômico) de infestação por um inseto-praga agrícola é de 10% das plantas infestadas. Você decide fazer um levantamento em nove lotes, selecionados aleatoriamente, de uma área de produção e calcula o percentual de plantas infestadas em cada lote, obtendo o seguintes valores:

5.1	6	8.8	11.5	4.4	8.4	8	7.5	9.5

Como estabelecer um critério para saber se a área de produção está ou não abaixo do nível crítico?

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 5%.

- 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).

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

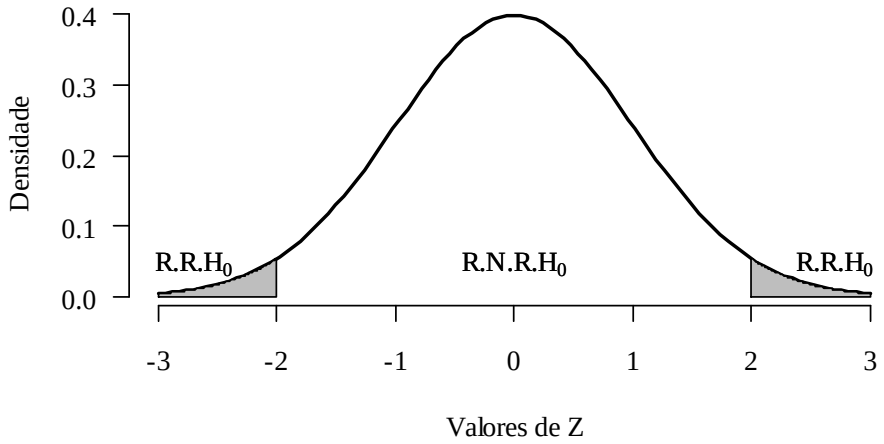
2) Caso em que não se conhece a variância (σ^2).

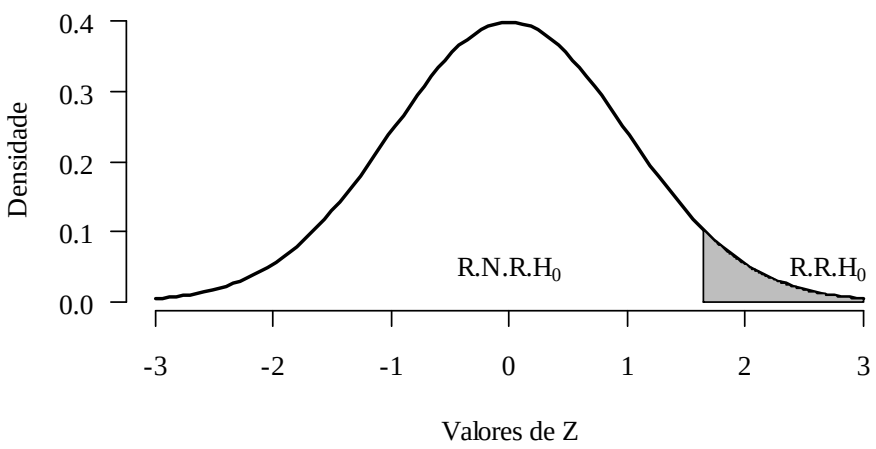
$$t = \frac{\overline{X} - \mu}{s / \sqrt{n}} \sim t(n - 1 \text{ g.l.})$$

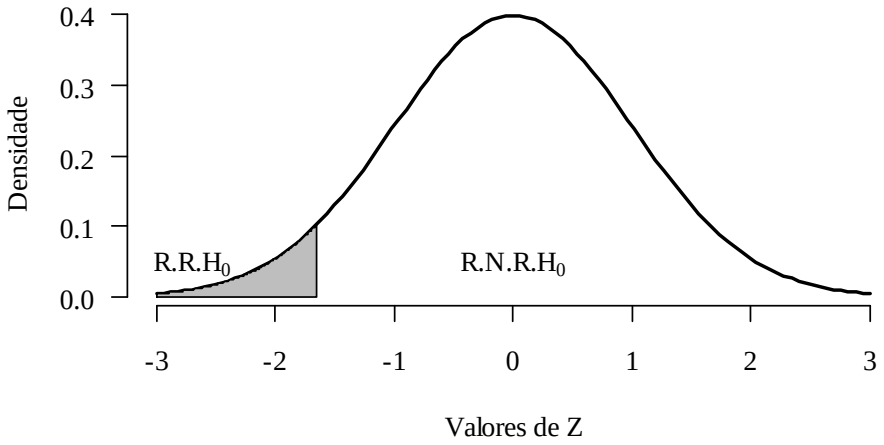
Erros associados aos testes de hipóteses

- **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







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...

Para o exemplo anterior, suponha que não conhecemos σ.
É possível dizer que a média da amostra a seguir é estatisticamente inferior a 1 litro?

$$Hipóteses: \begin{cases} H_0 : \mu = 10 \\ H_1 : \mu < 10 \end{cases}$$

Podemos usar uma estimativa de σ, obtida pelo estimador s:

$$s = 2{,}2374$$

A estatística de teste nesse caso é a t-Student com n-1 graus de liberdade:

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Conceito de valor-p

- O **valor-p** quantifica o quão estranho é o resultado da amostra, supondo H_0 verdadeira. Para o exemplo, estamos supondo que a infestação na área total é 10%.

- 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.

- Valor-p é a probabilidade de H_0 ser verdadeira, com base na amostra.

Intervalo de confiança para a média

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

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

IC para o exemplo

Relembrando: $\begin{cases} \bar{x} = 7{,}69\% \\ s = 2{,}2374\% \\ n = 9 \end{cases}$

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.

$$t_{\frac{\alpha}{2},(n-1)} = t_{\frac{0{,}05}{2},(9-1)} = 2{,}30$$

IC para o exemplo

$$IC(\mu)_{1-\alpha} = \bar{x} \pm t_{\frac{\alpha}{2}}(n-1)\frac{s}{\sqrt{n}} = 7{,}69 \pm 2{,}30\frac{2{,}2374}{\sqrt{9}}$$
$$= 7{,}69 \pm 1{,}71$$

$$5{,}98 < \mu < 9{,}40$$

Dado o nível de 95% de confiança, é possível afirmar que μ difere de 10?

Dimensionamento do tamanho amostral

Suponha que Δ = 1% tolerância máxima aceitável quando do cálculo da média do nível de infestação. Para detectar tal diferença com probabilidade de erro de, no máximo, 5%, qual o tamanho (n) necessário a ser considerado para as amostras?

Suponha σ = 0,05 (conhecido!).

$$|t_{\frac{\alpha}{2}}| \leq \frac{\left|\overline{X} - \mu\right| = \Delta}{s/\sqrt{n}} \quad \Rightarrow \quad n \geq \frac{|t_{0,025}|^2 \, s^2}{\Delta^2}$$

$$n \geq \frac{|2{,}30|^2 \, (2{,}2374)^2}{1^2} =$$

$$n \geq 26$$

Estatísticas de teste para 2 médias

O teste de hipótese do tipo H_0: $\mu_1 = \mu_2$ pode ser feito por meio das seguinte estatística de teste:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_c^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2 \;\; g.l.)$$

Estatísticas de teste para 2 médias

Considere comparar dados de comprimento de sépala (cm) das duas espécies:

I. setosa	I. versicolor
5.1	7.0
4.9	6.4
4.7	6.9
4.6	5.5
5.0	6.5
5.4	5.7
4.6	--