Estatística e Probabilidade

Frases

"Uma probabilidade razoável é a única certeza"

Samuel Howe

"A experiência não permite nunca atingir a certeza absoluta. Não devemos procurar obter mais que uma probabilidade ."

Bertrand Russel

Roteiro

Variáveis Aleatórias

Variável Aleatória

- Variável que tem um valor numérico único para cada resultado de um experimento aleatório.
- Representada geralmente por uma letra maiúscula.

Variável Aleatória Discreta

- Ou admitem um número finito de valores ou tem uma quantidade enumerável de valores,
- ‘Enumerável’ refere-se ao fato que podem ser infinitos valores, mas que são resultados de um processo de contagem.

Função de Probabilidade

- Gráfico, tabela ou fórmula que dá a probabilidade de cada valor da variável aleatória.

Função de Probabilidade – Condições

- Para qualquer valor de x:

$$0 = P(x) = 1$$

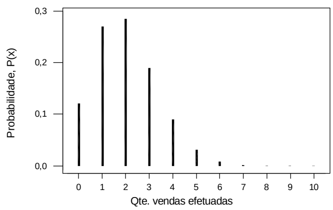
- A soma da probabilidade de todos os valores possíveis da variável é 1

$$?\ P(x) = 1$$

Exemplo – Distribuição Vendas Diárias

X	*P(X)*
0	O,122
1	0,270
2	0,285
3	0,190
4	0,090
5	0,032
6	0,009
7	0,002
8	0,000+
9	0,000+
10	0,000+

Exemplo – Gráfico da Distribuição de Probabilidades



Função de Probabilidade – Parâmetros

- Esperança (Média): $m = \sum_i x_i \, p(x_i)$
- Variância: $s^2 = \sum_i (x_i - m)^2 \, p(x_i)$
- Desvio padrão: $s = \sqrt{s^2}$

Variável Aleatória Contínua

- Toma um número infinito não-enumerável de valores.
- Esses valores podem ser associados a medições em uma escala contínua (sem lacunas ou interrupções).

Função de Densidade

- Gráfico de uma distribuição de probabilidade contínua
  - √ A área total sob a curva deve ser 1,
  - √ A curva não deve estar abaixo do eixo das abcissas (*x*)
- Há correspondência entre probabilidade e áreas da função de densidade (área total 1)

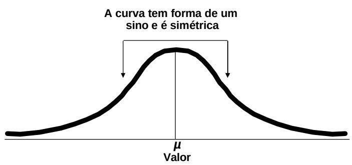
Função de Densidade – Parâmetros

- Esperança (Média): $m = \int_{R_X} x\, f(x)dx$
- Variância: $s^2 = \int_{R_X} (x - m)^2 f(x)dx$
- Desvio padrão: $s = \sqrt{s^2}$

Distribuição Normal

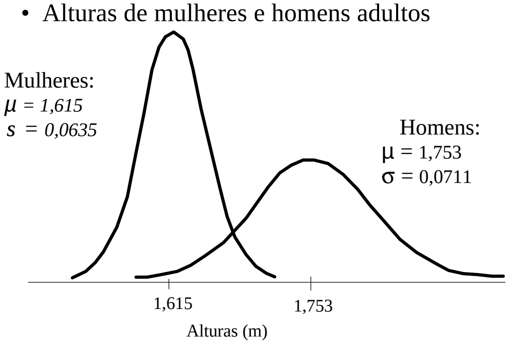
Distribuição Normal

- Variável aleatória contínua
- Função de densidade: $f(x) = \frac{1}{\sqrt{2p}\; s} \exp\left\{ -\frac{1}{2}\left(\frac{x-m}{s}\right)^2 \right\}$

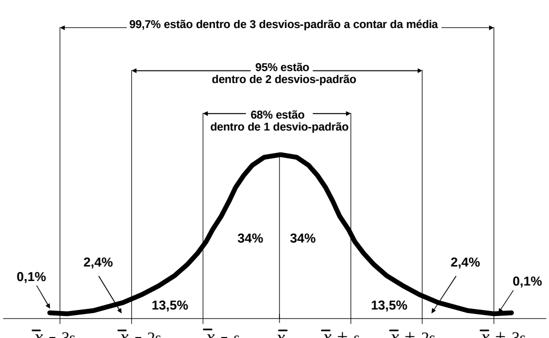


Exemplo

- Alturas de mulheres e homens adultos

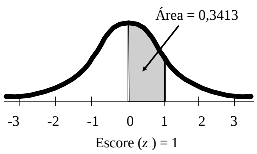


Regra Empírica



Distribuição Normal Padronizada

- Distribuição normal com:
  - √ Média: 0
  - √ Desvio padrão: 1



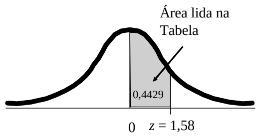


Tabela da Distribuição Normal Padrão

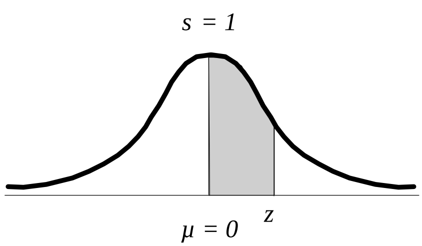
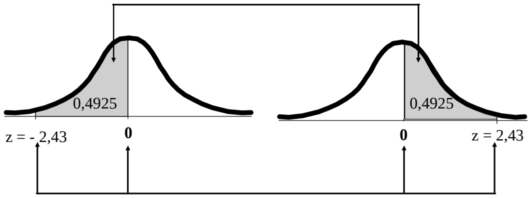


Tabela A-2 Distribuição Normal Padrão (*z*)

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

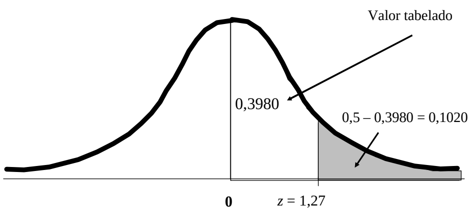
Simetria da Curva

Pela simetria, estas áreas são iguais,



Distâncias iguais a contar de 0

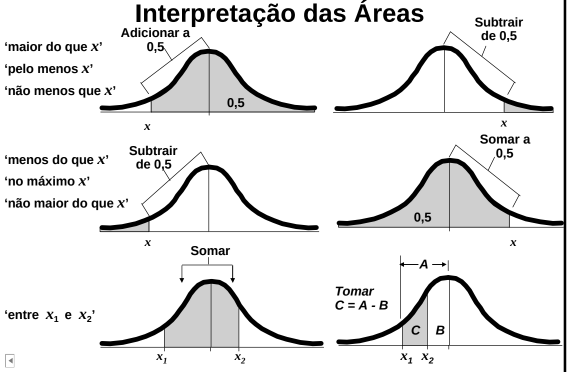
Área à Direita de *z*



Notação

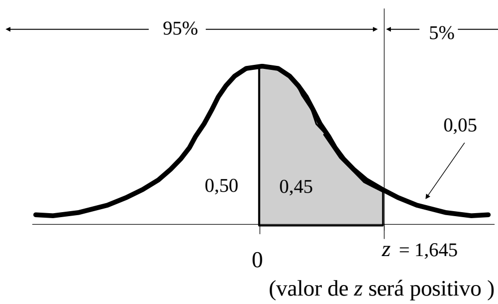
- P(a < Z < b): probabilidade de o valor de *z* estar entre *a* e *b*
- P(Z > a): probabilidade de o valor de *z* ser maior do que *a*
- P (Z < a): probabilidade de o valor de *z* ser menor do que *a*

Interpretação das Áreas



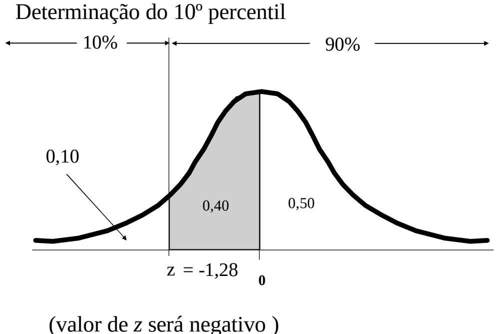
Determinação Quantil (1)

Determinação do 95º percentil



Determinação Quantil (2)

Determinação do 10º percentil



Outras Distribuições Normais

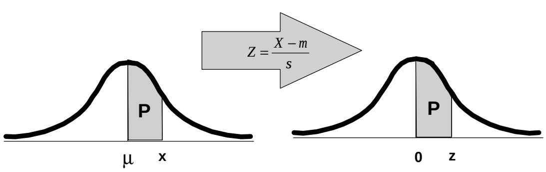
- Padroniza-se a variável, utilizando-se a tabela da normal padrão para os cálculos de interesse
- Para padronização:

$$Z = \frac{X - m}{s}$$

μ: média

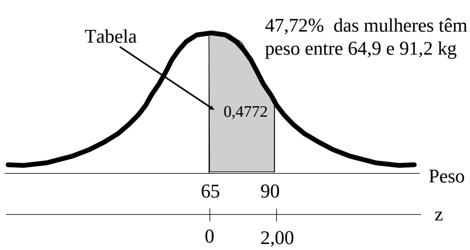
s : desvio padrão

Conversão para a Normal Padrão



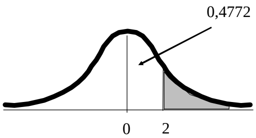
Exemplo (1)

- Peso população feminina
  - √ *μ = 65 kg* e *s = 12,5 kg*
  - √ Probabilidade de peso entre 65 e 90 kg



Exemplo (2)

- Duração de carga de bateria de celular:
  - √ $\mu = 20\ h$ e $s = 0,5\ h$
  - √ Probabilidade durar mais de 21 h



P (x > 2,00) = 0,0228

2,28% de baterias duram mais de 21 horas

Distribuição Amostral da Média

É a distribuição de probabilidade das médias amostrais, com todas as amostras de mesmo tamanho *n*, da mesma população

Teorema Central do Limite – Hipóteses

- A variável aleatória X tem uma distribuição qualquer (normal, ou não), com média μ e desvio padrão s;
- Amostras aleatórias de tamanho *n* extraídas dessa população.

Teorema Central do Limite

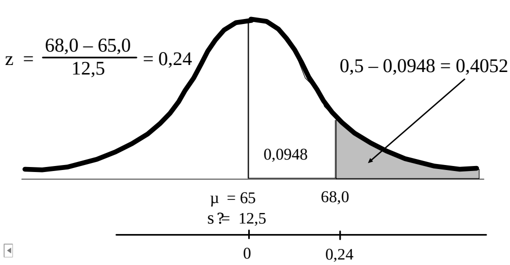
- Quando o tamanho da amostra aumenta, a distribuição das médias amostrais $\overline{X}$ tende para uma distribuição *normal*;
- A média das médias amostrais será a média populacional μ,
- O desvio padrão das médias amostrais é $s/\sqrt{n}$

Notação

- Média das médias amostrais: $\mu_{\overline{X}} = \mu$
- Desvio padrão das médias amostrais: $s_{\overline{X}} = \dfrac{s}{\sqrt{n}}$
  (*erro-padrão da média)*

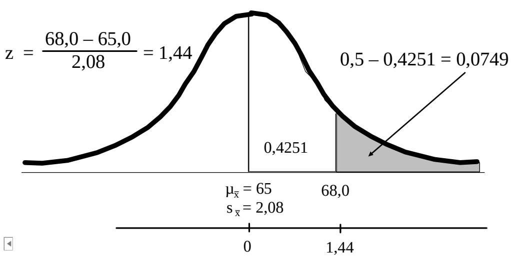
Exemplo

- Peso população feminina
  √ *μ = 65 kg* e *s = 12,5 kg*
- Probabilidade de peso maior que 68 kg



Exemplo

- Em amostras de tamanho *36*, probabilidade de peso médio maior que 68 kg



Comparação

- Probabilidade do peso de uma mulher ser maior que *68,0 kg*:

  0,4050

- Probabilidade da média de peso de 36 mulheres ser maior que *68 kg*:

  0,0749

- É mais fácil um elemento se desviar da média da população do que um <u>grupo de 36</u> elementos.

Inferência Estatística

Intervalo de Confiança

- Estimação intervalar do verdadeiro valor de parâmetro populacional

Menor # < parâmetro populacional < Maior #

Exemplo: Menor # < μ < Maior #

Grau de Confiança

- É a freqüência relativa de vezes que o processo de estimação gerará intervalo que contenha o parâmetro populacional.
- É denotado por *1 – a;*
- Geralmente:

Grau de confiança	a
90%	10%
95%	5%
99%	1%

$$98{,}08^{o} < \mu < 98{,}32^{o}$$

- Correto:

  Estamos 95% confiantes que o intervalo de 98,08 a 98,32 contenha o verdadeiro valor de μ.

  Se construíssemos intervalos de confiança a partir de muitas amostras de mesmo tamanho, 95% deles conteriam o parâmetro μ.

- Errado:

  É de 95% a probabilidade de o verdadeiro valor μ estar entre 98,08 e 98,32.

Exemplo – Intervalo de Confiança de 20 Amostras

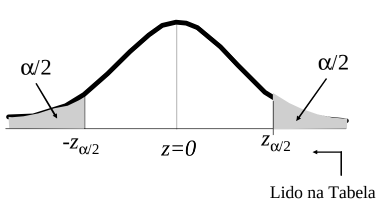
Valor Crítico

É o número na fronteira que separa os valores das estatísticas amostrais prováveis de ocorrer, dos valores que têm pouco chance de ocorrer.

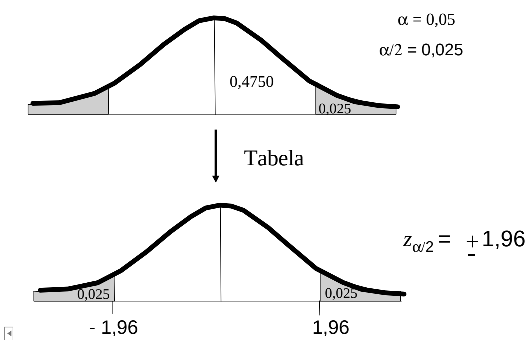
O Valor Crítico $Z_{\alpha/2}$

Considerados:

- Normalidade da estatística de teste
- Fronteira bilateral



$z_{a/2}$ para *95%* de Confiança

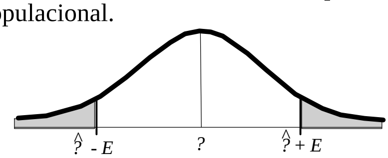


$z_{a/2}$ para vários Grau de Confiança

Grau de Confiança	a	Valor Crítico
90%	10%	1,645
95%	5%	1,960
99%	1%	2,575

Margem de Erro

- É a máxima diferença provável entre o estimador e o verdadeira valor do parâmetro populacional.

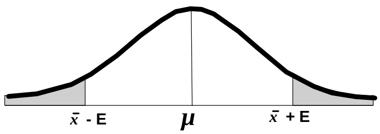


$$\underbrace{\hat{?} - E}_{\text{Limite inferior}} < ? < \underbrace{\hat{?} + E}_{\text{Limite superior}}$$

Margem de Erro da Média Amostral

Também chamada de <u>erro máximo da estimativa</u>

$$E = z_{a/2} \frac{s}{\sqrt{n}}$$



Intervalo de Confiança para a Média Populacional

$$\bar{x} - E < \mu < \bar{x} + E$$

$$\mu = \bar{x} \pm E$$

$$(\bar{x} + E, \bar{x} - E)$$

Exemplo

- Estimação da renda média do primeiro ano de trabalho de um bacharel.
  - √ Determinar tamanho amostral para um nível de 95% de confiança para que a média da amostra esteja a menos de $500 da média populacional?
  - √ Sabe-se por estudos prévios que s = $ 6.250

Exemplo (2)

$$n = \left(\frac{z_{a/2} s}{E}\right)^2$$

$$n = \left(\frac{1{,}96.6250}{500}\right)^2 = 600{,}25$$

Amostras de tamanho 601 oferecem 95% de confiança que sua média difiram em menos de $500 da verdadeira média populacional.

Distribuição *t* de Student

- Se a distribuição de uma população é essencialmente normal, então a distribuição de

$$t = \frac{\overline{X} - m}{\frac{s}{\sqrt{n}}}$$

  é uma Distribuição *t* de Student para todas amostras de tamanho *n*.
- Valores críticos denotados por $t_{a/2}$.

Graus de Liberdade

- Corresponde ao número de valores amostrais que podem variar após terem sido impostas certas restrições a todos os valores.
- No caso da *t-Student*:

$$gl = n - 1$$

Tabela Distribuição *t*

Graus de liberdade	.005 (unilateral) .01 (bilateral)	.01 (unilateral) .02 (bilateral)	.025 (unilateral) .05 (bilateral)	.05 (unilateral) .10 (bilateral)	.10 (unilateral) .20 (bilateral)	.25 (unilateral) .50 (bilateral)
1	63,657	31,821	12,706	6,314	3,078	1,000
2	9,925	6,965	4,303	2,920	1,886	,816
3	5,841	4,541	3,182	2,353	1,638	,765
4	4,604	3,747	2,776	2,132	1,533	,741
5	4,032	3,365	2,571	2,015	1,476	,727
6	3,707	3,143	2,447	1,943	1,440	,718
7	3,500	2,998	2,365	1,895	1,415	,711
8	3,355	2,896	2,306	1,860	1,397	,706
9	3,250	2,821	2,262	1,833	1,383	,703
10	3,169	2,764	2,228	1,812	1,372	,700
11	3,106	2,718	2,201	1,796	1,363	,697
12	3,054	2,681	2,179	1,782	1,356	,696
13	3,012	2,650	2,160	1,771	1,350	,694
14	2,977	2,625	2,145	1,761	1,345	,692
15	2,947	2,602	2,132	1,753	1,341	,691
16	2,921	2,584	2,120	1,746	1,337	,690
17	2,898	2,567	2,110	1,740	1,333	,689
18	2,878	2,552	2,101	1,734	1,330	,688
19	2,861	2,540	2,093	1,729	1,328	,688
20	2,845	2,528	2,086	1,725	1,325	,687
21	2,831	2,518	2,080	1,721	1,323	,686
22	2,819	2,508	2,074	1,717	1,321	,686
23	2,807	2,500	2,069	1,714	1,320	,685
24	2,797	2,492	2,064	1,711	1,318	,685
25	2,787	2,485	2,060	1,708	1,316	,684
26	2,779	2,479	2,056	1,706	1,315	,684
27	2,771	2,473	2,052	1,703	1,314	,684
28	2,763	2,467	2,048	1,701	1,313	,683
29	2,756	2,462	2,045	1,699	1,311	,683
Grande (z)	2,575	2,327	1,960	1,645	1,282	,675

Propriedades

- É diferente de acordo com o tamanho da amostra .
- Tem a mesma forma geral que a distribuição normal padrão (forma de sino, com média 0)

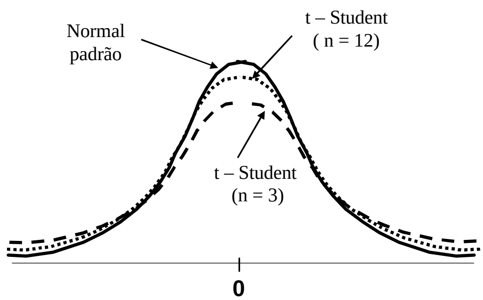
Propriedades (2)

- Reflete a maior variabilidade da média amostral, esperada em pequenas amostras (caudas mais pesadas).
- Seu desvio-padrão varia com o tamanho da amostra, sendo sempre superior ao da normal padrão (1).

Propriedades (3)

- A distribuição *t* de Student se aproxima da distribuição normal padrão à medida em que *n* cresce
- Em geral, para valores *n* > 30, as diferenças são tão pequenas que podemos utilizar os valores da normal padrão

Distribuição *t* – *Student* para *n = 3* e *n = 12*



Exemplo

- Objetivo: avaliação de tempo de treinamento foi selecionada
- Amostra: 15 empregados
- Resultados amostrais:
  - √ Média: 53,87 dias
  - √ Desvio-padrão: 6,82 dias
- Determinar intervalo com 95% de confiança para μ (tempo médio para treinamento de todos os empregados da empresa)

Exemplo (2)

$\bar{x} = 53{,}87$

$s = 6{,}82$

$t_{a/2} = 2{,}145$

$$E = \frac{t_{a/2}s}{\sqrt{n}} = \frac{2{,}145.6{,}82}{\sqrt{15}} = 3{,}78$$

$$\bar{x} - E < \mu < \bar{x} + E$$

$$53{,}87 - 3{,}78 < \mu < 53{,}87 + 3{,}78$$

$$50{,}09 < \mu < 57{,}65$$

Estamos 95% confiantes que este intervalo contenha a média de tempo de treinamento de todos os empregados.

Teste de Hipóteses

Hipótese

- Em Estatística, é uma alegação ou afirmação sobre uma propriedade de uma população

Regra do Evento Raro

- Analisar uma amostra para distinguir entre resultados que podem ocorrer facilmente daqueles que dificilmente ocorrem.
- Resultados altamente improváveis:

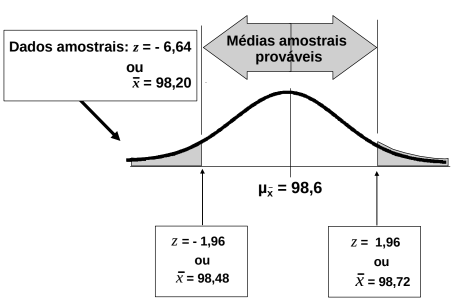
  Sua ocorrência pode ser explicada por ele ser efetivamente um evento raro, ou de que a suposição não está correta.

Exemplo – Médias Amostrais

- Objetivo: verificar se a amostra é originária de população com média: 98,6
- Média da amostra obtida: 98,2

  Suposição: a amostra é suficientemente grande para a aplicação do Teorema Central do Limite

Exemplo - Médias Amostrais



Teste de Hipóteses – Componentes

- Hipótese Nula;
- Hipótese Alternativa
- Estatística de Teste
- Região Crítica
- Nível de Significância

Hipótese Nula – H_0

- Afirmação sobre valor de parâmetro populacional
- Deve conter uma condição de igualdade

  =, ≥, ou ≤
- Testar diretamente a Hipótese Nula:

  Rejeitar H_0 ou não rejeitar H_0

Hipótese Alternativa – H_1

- Deve ser verdadeira se H_0 é falsa

  ?, <, >
- ‘oposto’ da Hipótese Nula.

Hipótese de Pesquisa

- Para montar um teste de hipótese para apoiar de pesquisa, ela deve ser formulada de maneira a ser a hipótese alternativa.

Estatística de Teste

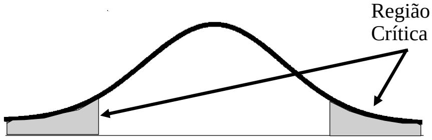
- Valor baseado nos dados amostrais que é usado para tomar uma decisão sobre a rejeição da hipótese nula.
- Exemplo:

  Para testar afirmações sobre médias populacionais, através de grandes amostras

$$z = \frac{\bar{x} - m_{\bar{x}}}{\frac{s}{\sqrt{n}}}$$

Região Crítica

- Conjunto de todos os valores da estatística de teste que levam à rejeição de H_0.



Região de Aceitação

- Valores da estatística de teste que não levam à rejeição de H_0
- É o conjunto complementar à região crítica.

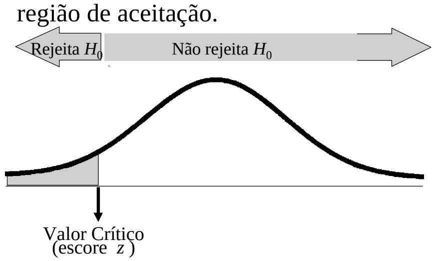
Nível de Significância

- É a probabilidade de rejeitar H_0 quando ela é verdadeira.
- É tipicamente pré-determinado, sendo comum as escolhas:

  0,05; 0,01 e 0,10
- Notação *a*

Valores Críticos

- Valor(es) que separa(m) a região crítica da região de aceitação.

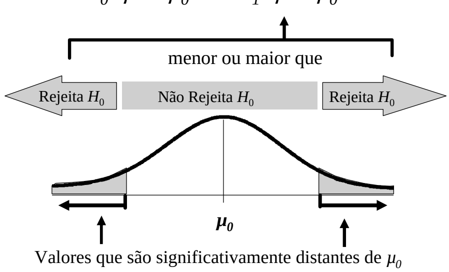


Tipos de Teste

- As caudas em uma distribuição são as regiões extremas delimitadas por valores críticos.
- Tipos de teste:
  - √ Bilateral
  - √ Unilateral esquerdo
  - √ Unilateral direito

Teste Bilateral

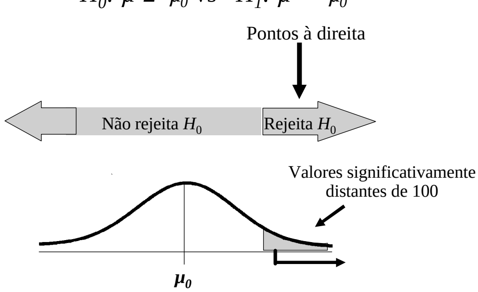
$$H_0: \mu = \mu_0 \ vs \ H_1: \mu \neq \mu_0$$



α é dividido igualmente entre as duas caudas da região crítica

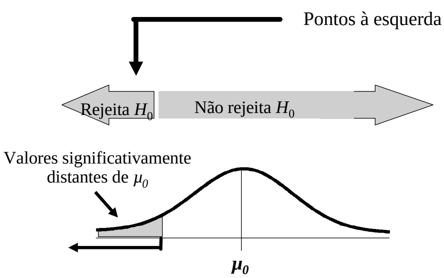
Teste Unilateral Direito

$$H_0: \mu \leq \mu_0 \text{ vs } H_1: \mu > \mu_0$$

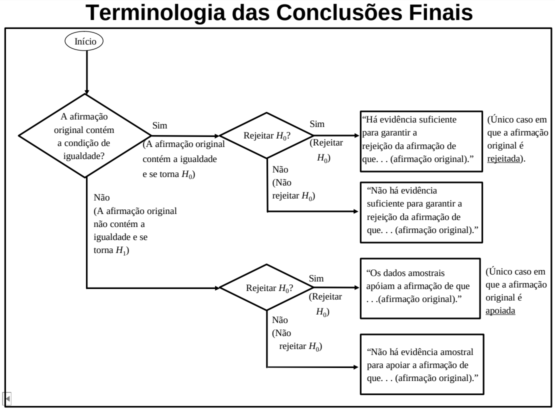


Teste Unilateral Esquerdo

H_0: $\mu = \mu_0$ vs H_1: $\mu < \mu_0$



Terminologia das Conclusões Finais



Erro Tipo I

- O erro de rejeitar a H_0 quando ela é verdadeira.
- O nível de significância a é a probabilidade de um erro tipo I.
- Exemplo:

  Rejeitar a afirmação de que a temperatura do corpo é 37ºC, quando ela é, de fato, 37ºC.

Erro Tipo II

- Erro de não rejeitar H_0 quando ela é falsa.
- A probabilidade do erro tipo II é *ß*.
- Exemplo:

  Não rejeitar a afirmação de que a temperatura do corpo é 37ºC, quando ela é, de fato, falsa (a média não é 37ºC).

Erros Tipo I e Tipo II

		Estado da Natureza	
		H_0 Verdadeira	H_0 Falsa
Decisão	Rejeição de H_0	Erro tipo I *(a)*	Acerto *(1 – ß)*
	Não rejeição de H_0	Acerto *(1 – a)*	Erro tipo II *(ß)*

Controle dos Erros Tipo I e Tipo II

- Para *a* fixo, aumentar o tamanho da amostra *n* reduz o valor de *ß*;
- Para *n* fixo, diminuir *a* leva a um aumento de *ß* (ou, vice-versa);
- Para redução de *a* e *ß* , deve-se aumentar *n*.

Poder do Teste

- É a probabilidade *(1 - ß)* de rejeitar uma H_0 falsa
- É calculada para:
  - √ um nível de significância *a*, e
  - √ um valor do parâmetro que seja alternativa para o valor de H_0.

Teste de Hipóteses – Método Clássico

- Utiliza uma <u>estatística de teste</u> (função dos dados da amostra), comparado-a com um <u>valor crítico</u>.
- Exemplo:
  Estatística de teste para Afirmações sobre μ (suposição de normalidade)

$$z = \frac{\bar{x} - m_{\bar{x}}}{\frac{s}{\sqrt{n}}}$$

Passos do Teste

- Identificar as hipóteses a ser testada;
- Colocá-las em forma simbólica;
- H_0 contém a condição de igualdade;
- Escolher o nível de significância *a*, baseando-se na gravidade de um erro tipo I (Em geral, 0,05 e 0,01):
  - √ . Tomar *a* pequeno se as conseqüências da rejeição de uma H_0 verdadeira são sérias.

Passos do Teste (2)

- Escolher a estatística relevante para este teste e determinar sua distribuição amostral.
- Determinar os valores críticos e a região crítica.
- Critério de Decisão:

  Rejeitar H_0 se a estatística de teste está na região crítica, caso contrário, não rejeitar H_0.
- Concluir sobre o texto no contexto do problema

Método do *Valor p*

- Similar ao método tradicional;
- Procedimento para tomada de decisão:

  Determinar a *probabilidade* (*Valor p*) de se obter um resultado mais favorável contra H_0,

  Se o *Valor p* é muito baixo rejeita-se H_0.

Valor p – Definição

- É a probabilidade de se obter um valor da estatística de teste *pelo menos tão extremo* como o observado, supondo-se que H_0 seja verdadeira.

Valor *P*	Interpretação
Valor p pequeno *(= 0,05)*	Resultados amostrais incomuns. Diferença significante de H_0.
Valor p grande *(> 0,05)*	Resultados amostrais não são incomuns. Não há diferença significante de H_0.

Teste de Hipóteses – Razão Subjacente

- Dada uma suposição observada, se a probabilidade de obtermos a amostra observada é pequena, então provavelmente a suposição não é correta.
- Ao se testar uma afirmação (H_0), supõe-se que ela contenha a igualdade. Essa suposição é comparada com os dados amostrais.

Teste de Hipóteses – Conclusões

- Se os resultados amostrais podem ocorrer facilmente quando a suposição (H_0) é verdadeira, atribue-se ao acaso a diferença pequena encontrada entre a suposição e os resultados amostrais.
- Se os resultados amostrais são improváveis de ocorrer (sob H_0), explica-se a diferença (relativamente grande) entre H_0 e os resultados amostrais supondo-se que ela não seja verdadeira.

Referências

Bibliografia Recomendada

- Freund, J. E. e Simon, G. A. (Artmed)
  Estatística aplicada: economia, administração e contabilidade
- Triola, M. F. (LTC)
  Introdução à estatística.
- Bussab, W. O. e Morettin, P. A. (Saraiva)
  Estatística básica