Estatística Inferencial

26-27 Outubro 2021

ICS-ULisboa

Alda Botelho Azevedo | Marcelo Camerlo

alda.azevedo@ics.ulisboa.pt | marcelo.camerlo@ics.ulisboa.pt

Programa

Programa

1. Fundamentos da Estatística Inferencial:
    1.1. Distribuições de variáveis aleatórias;
    1.2. Teorema do Limite Central.

2. Testes paramétricos:
    2.1. Teste t para uma média;
    2.2. Teste t para duas médias, amostras independentes;
    2.3. Comparação de várias médias (*One-way* ANOVA).

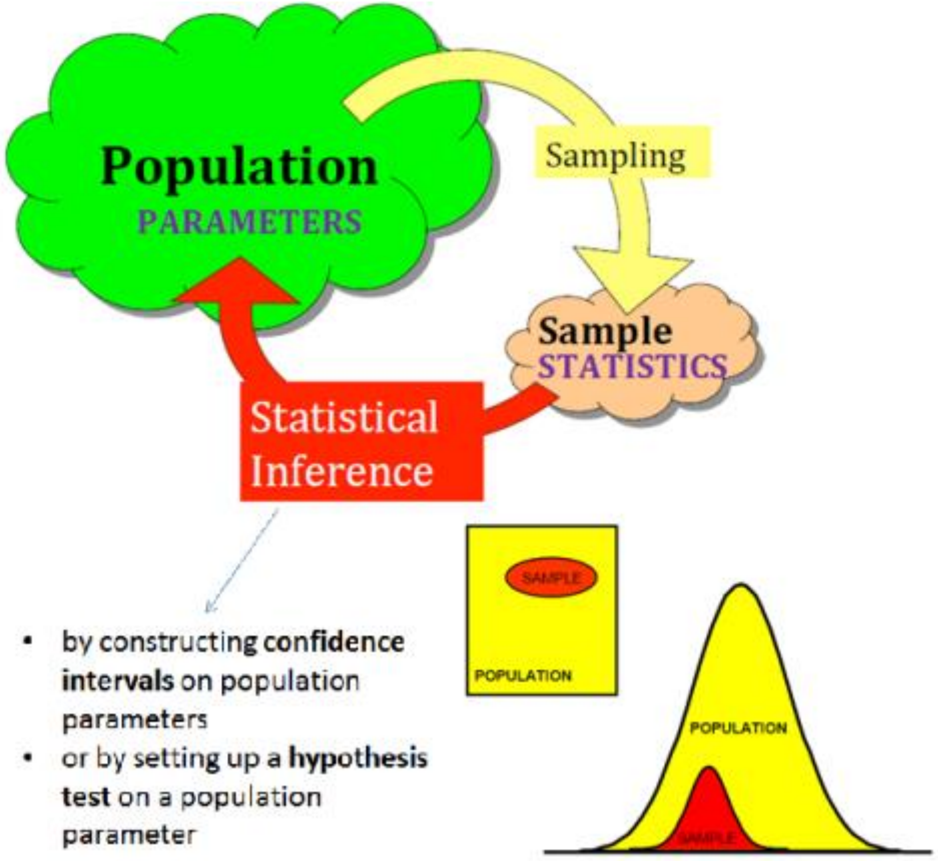
3. Testes não paramétricos:
    3.1. Testes de Normalidade (Kolmogorov-Smirnov e Shapiro-Wilk);
    3.2. Teste de independência do Qui-quadrado.
    3.3. Comparação de várias médias (Mann-Whitney e Kruskal-Wallis)

Bibliografia principal: Diez, Bar rand Çetinkaya-Rundel (2012). *OpenIntro Statistics*. Second Edition.

Disponível em: https://www.openintro.org/book/os/.

Fundamentos da Estatística Inferencial

Porquê Estatística Inferencial?



Estimação: pontual ou por intervalos

Ensaios de hipóteses: paramétricos ou não paramétricos

Parâmetros: características da população
Média (μ)
Variância (σ^2)
Desvio-padrão (σ)
Proporção (P)

Estatísticas: características da amostra
Média amostral ($\overline{X}$)
Variância amostral (s^2)
Desvio-padrão amostral (s)
Proporção (f ou p)

Fonte: https://mahritaharahap.wordpress.com/teaching-areas/inferential-statistics/

Objetivos da Estatística Inferencial

- Avaliar um parâmetro
  - Em que medida a média de uma amostra é representativa da média da população?
- Avaliar uma relação
  - Em que medida A e B estão relacionados? E em que medida essa relação é real ou simplesmente fruto do acaso?
- Predizer um valor
  - Existindo uma relação entre A e B, conhecendo o resultado de um sujeito em A é possível prever B, com uma margem de erro.
- Determinar se uma diferença observada entre duas amostras é devida a uma causa sistemática.

Fonte: Adaptado a partir de D'Hainaut, 1997.

População e amostra na Estatística Inferencial

Amostra aleatória ou probabilística

Todos os elementos da população uma probabilidade conhecida e diferente de zero de ser selecionados para a amostra.

Amostra não aleatória ou não probabilística

São escolhidos alguns elementos da população. Não é conhecida a probabilidade dos elementos serem selecionados para a amostra.

	Amostra aleatória (ou probabilística)	**Amostra não aleatória (ou não probabilística)**
Base de sondagem da população	Sim	Não
Representatividade da população	Sim	Não é possível determinar
Inferência de parâmetros populacionais, com uma margem de erro definida	Sim	Não

Distribuições de variáveis aleatórias

Distribuição | Conceito

- A distribuição é a expressão para uma variável aleatória específica.
- Tem por base uma função matemática parametrizada (processo moroso e complexo) que pode ser utilizada para calcular a probabilidade de qualquer observação individual a partir do espaço da amostra.
- As distribuições são frequentemente descritas em termos das suas funções de densidade.
  - A **função densidade de probabilidade** calcula a probabilidade de observar um determinado valor.
  - A **função densidade cumulativa** calcula a probabilidade de uma observação ter um valor igual ou inferior a um determinado valor.
- Uma vez conhecida uma função de distribuição, esta pode ser utilizada como um exemplo-tipo para descrever e calcular quantidades relacionadas, tais como a probabilidade de observações, e traçar a relação entre as observações no domínio.

Distribuição Nomal

Normal (ou Gaussian)

- É uma das distribuições mais utilizadas e por isso ocupa um lugar de destaque na estatística inferencial.

- Aplicável a variáveis aleatórias contínuas.

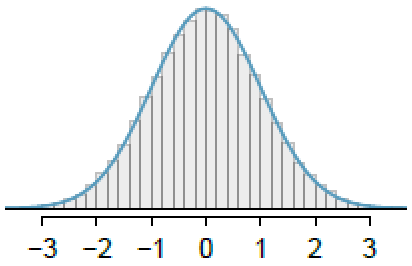
Exemplos

- Altura.

- Pressão arterial.

- Número de entregas de uma plataforma de distribuição por dia.

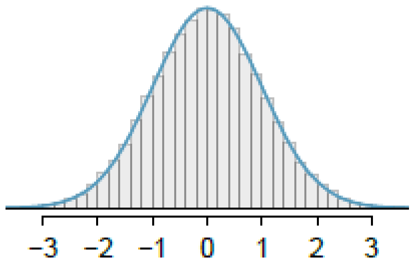
- …



Fonte: OpenIntro.

Distribuição normal (ou Gaussian)

- Simétrica, unimodal e em forma de sino.
- Denota-se N(μ, σ) a curva normal com média μ (miú) e desvio-padrão σ (sigma).
- As observações concentram-se em torno da média e a distância da média a qualquer ponto da distribuição é constante e mede-se em unidades desvio-padrão.
- A área sob a curva totaliza 1.
- Numa distribuição normal padrão a média é igual a 0 e o desvio-padrão é igual a 1.



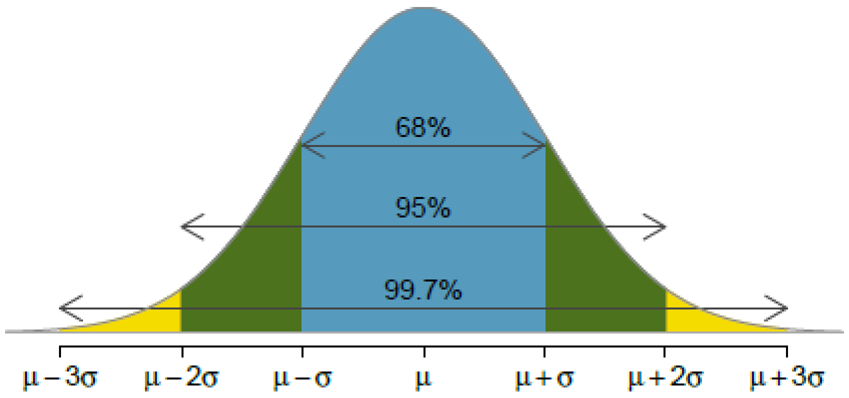
Fonte: OpenIntro.

Distribuição normal (ou Gaussian)

Regra 68 – 95 – 99.7

- Na distribuição normal, as observações concentram-se em torno da média e a distância da média a qualquer ponto da distribuição é constante e mede-se em unidades desvio-padrão.

Probabilidade de cair dentro de 1, 2 e 3 desvios-padrão da média numa distribuição normal



Fonte: OpenIntro.

Distribuição t-Student

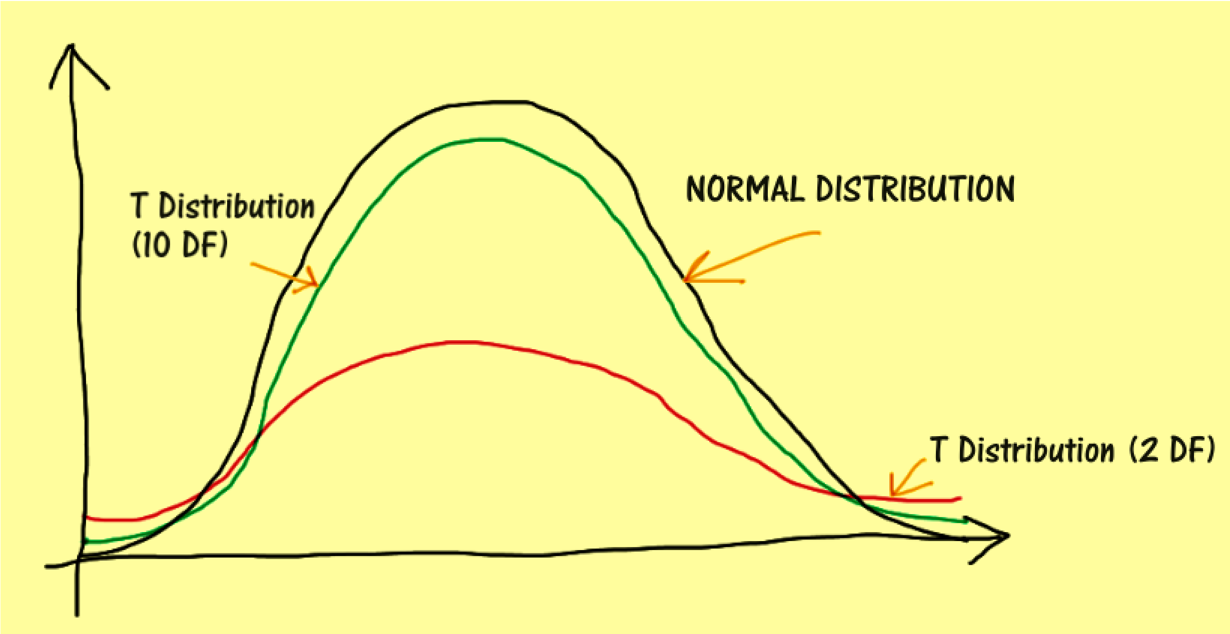
- Permite fazer inferências sobre a média de uma população com base uma amostra aleatória.
- A distribuição pode ser descrita com um único parâmetro, o número de graus de liberdade (v) que é um número positivo.
- O número de graus de liberdade descreve o número de elementos de informação utilizados para descrever uma quantidade populacional.
  - Por exemplo, a média tem *n* graus de liberdade, pois todas as *n* observações da amostra são utilizadas para calcular a estimativa da média da população.
  - Uma quantidade estatística que faz uso de outra quantidade estatística no seu cálculo deve subtrair 1 aos graus de liberdade.
- As observações na distribuição t-Student são calculadas a partir de observações numa distribuição normal, a fim de descrever o intervalo para as médias das populações na distribuição normal.
- É simétrica, semelhante à distribuição normal.

Distribuição do Qui-quadrado

- Permite fazer inferências sobre a **variância** de uma população a partir de uma amostra e é denotada por $\chi 2$.
- Tal como a distribuição t de Student, a distribuição qui-quadrado é também utilizada em métodos sobre dados retirados de uma distribuição normal para quantificar a incerteza. Por exemplo, é utilizada nos testes de independência do qui-quadrado.
- A distribuição do qui-quadrado tem um parâmetro, os graus de liberdade, denotados por k.
- Uma observação numa distribuição qui-quadrado é calculada como a soma das observações ao quadrado de k retiradas de uma distribuição normal.

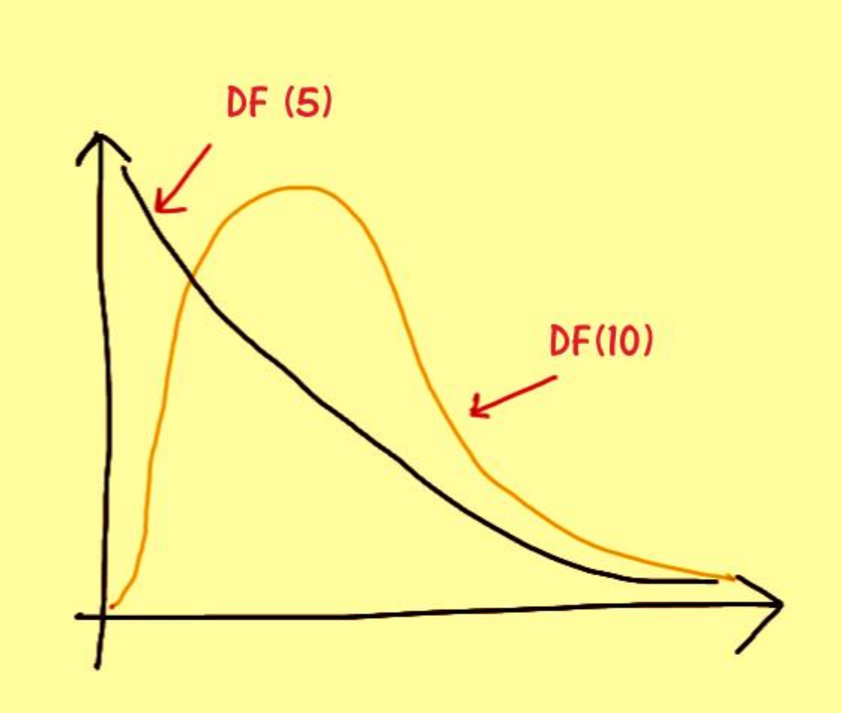
Formas das principais distribuições

Distribuição Normal e distribuição t-Student



As caudas da distribuição de T estreitam-se quando os graus de liberdade aumentam e a curva da distribuição começa a assemelhar-se à distribuição normal.

Distribuição Normal e distribuição do Qui-quadrado



À medida que os graus de liberdade aumentam, a distribuição começa a assemelhar-se à distribuição normal, mas sempre com valores positivos.

Fonte: https://towardsdatascience.com/understanding-probability-and-statistics-chi-squared-student-t-and-f-distributions-e46b4f802707

Teorema do Limite Central

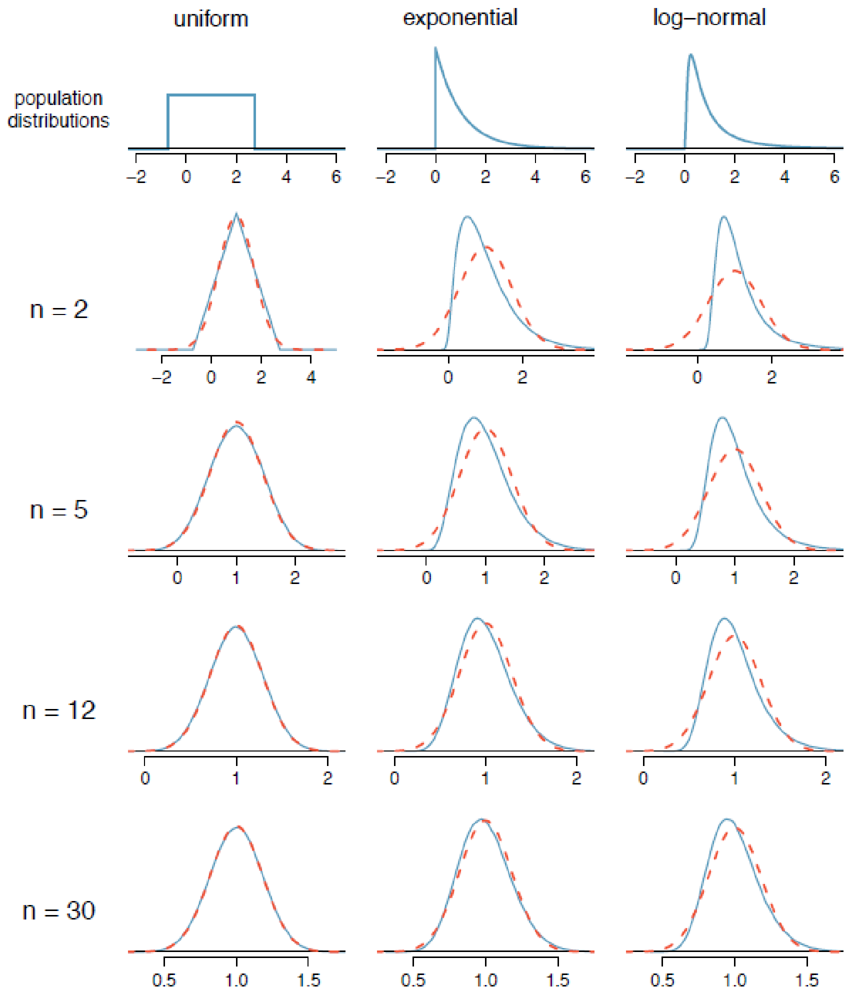
Teorema do Limite Central (TLC)

A distribuição da média de um grande número de variáveis aleatórias independentes e identicamente distribuídas será aproximadamente normal, independentemente da distribuição subjacente.

Quanto maior a dimensão das amostras, mais as médias amostrais se aproximam da distribuição normal e mais se concentram sobre a média na população.

As médias das amostras são uma variável que tem distribuição aproximadamente normal.

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{s\,\cancel{\sigma}}{\sqrt{n}}\right)$$



Fonte: OpenIntro.

Teorema do Limite Central (TLC)

Condições para que se aplique o TLC

Independência

As amostras devem ser independentes. Isso é difícil de verificar, mas é mais provável em amostras aleatórias, sem substituição e se n <10% da população.

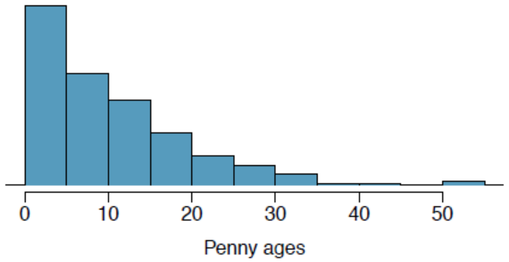
Normalidade

Quando uma amostra é pequena, as observações da amostra devem vir de uma população normalmente distribuída. Podemos relaxar essa condição cada vez mais para amostras cada vez maiores.

Fonte: OpenIntro.

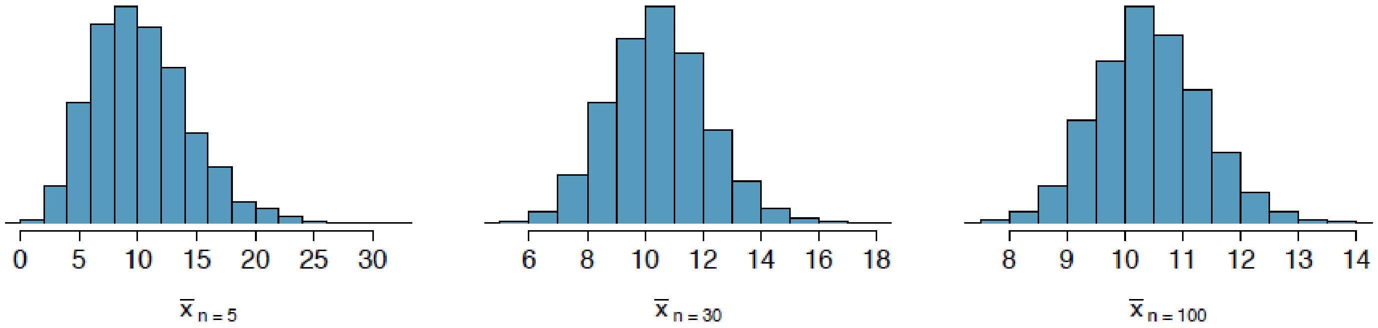
Exercício 1

O histograma que se segue mostra a distribuição das idades das moedas de centavos num banco. Descreva-a.



Os histogramas que se seguem referem-se a distribuições de amostras aleatórias de 5, 30 e 100 centavos.

Descreva as formas dessas distribuições e comente se elas se aproximam de que esperaria com base no Teorema do Limite Central.



Fonte: OpenIntro.

Exercício 2

Um estudo recolheu dados da frequência cardíaca, fenómeno que segue aproximadamente uma distribuição normal. São conhecidas três estatísticas:

- média=110
- valor mínimo=65
- valor máximo=155

Das seguintes opções, qual é o desvio-padrão mais provável da distribuição?

a) $\sigma = 5$

b) $\sigma = 15$

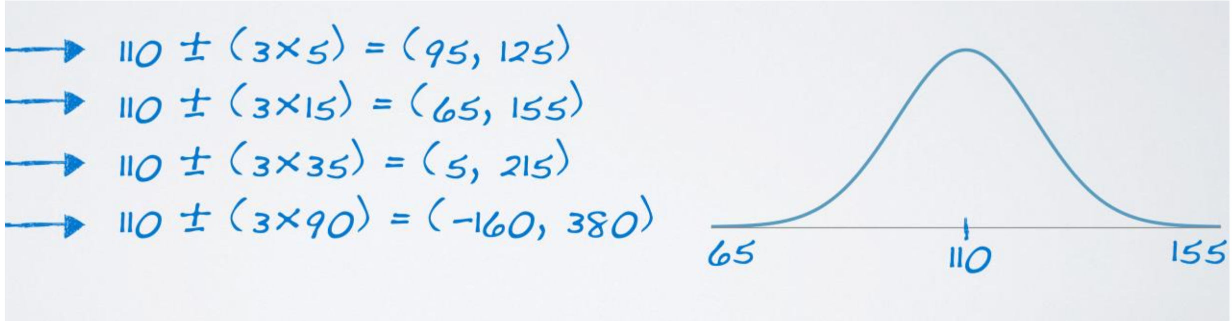
c) $\sigma = 35$

d) $\sigma = 90$

Desenhe a distribuição da frequência cardíaca.

Fonte: OpenIntro.

Resolução exercício 2



b) σ = 15

110 ± (3*15) = 45
110+45 = 155
110-45=65

Exercício 3

O responsável pela admissão dos alunos numa faculdade quer determinar qual de dois candidatos obteve o melhor resultado no seus testes. No entanto, os candidatos realizaram exames com classificações em escalas distintas. Pam, que obteve 1800 no seu SAT ou Jim, que pontou um 24 em seu ACT?

Sobre os resultados de cada um dos exames sabe-se que:

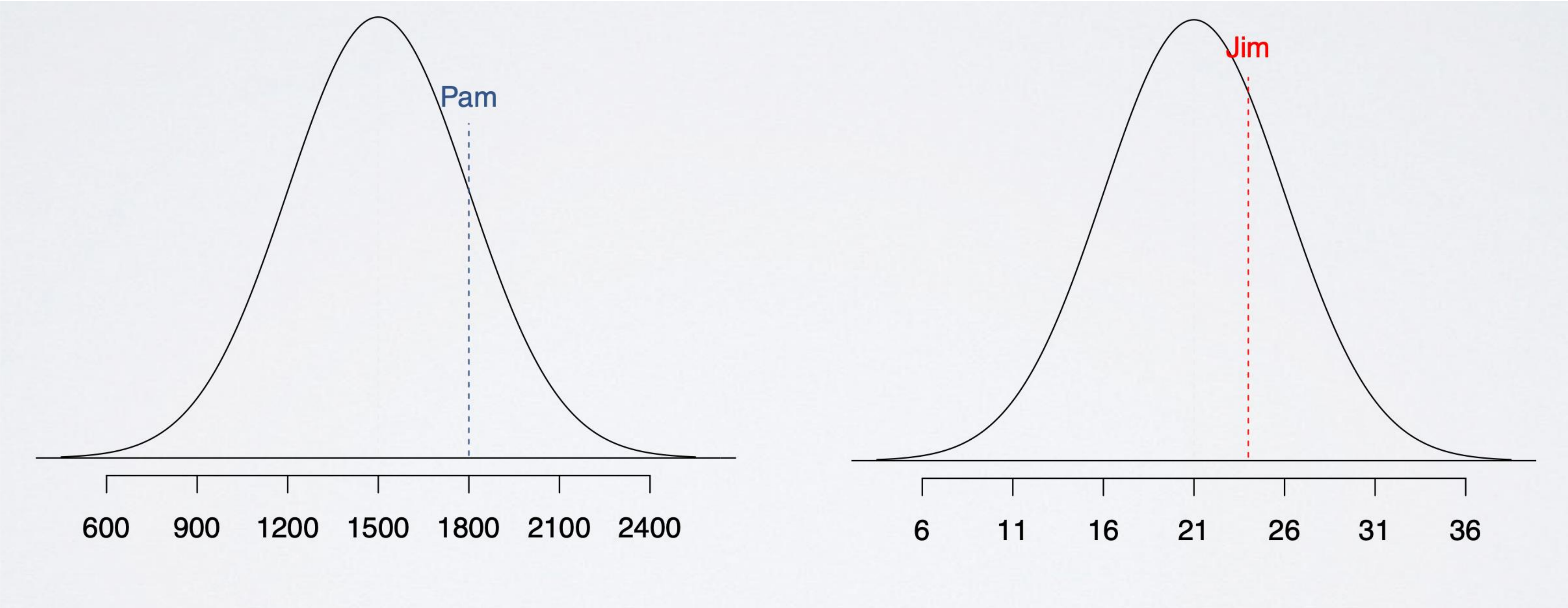
SAT ~ N(μ = 1500, σ = 300)

ACT ~ N(μ =21, σ = 5)

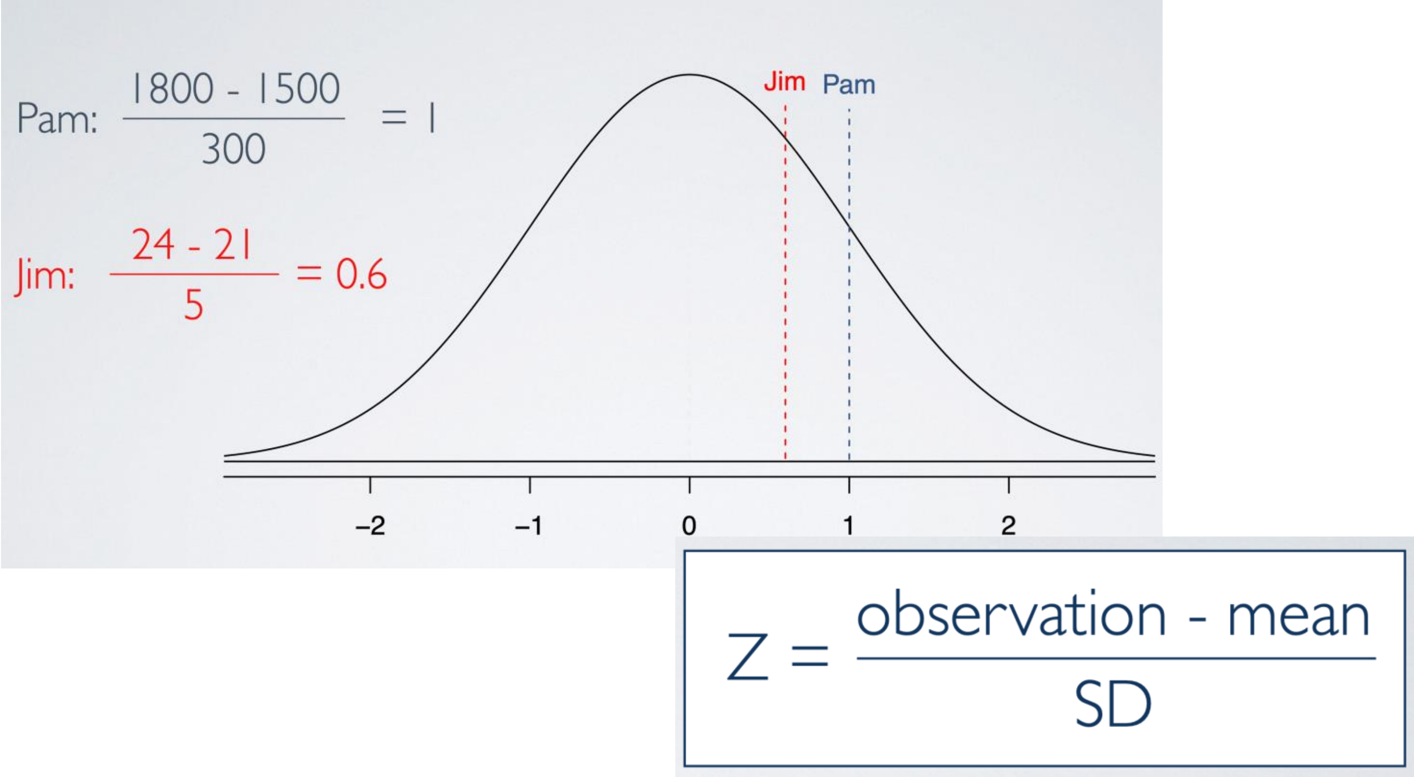
Desenhe as distribuições.

Fonte: OpenIntro.

Resolução exercício 3



Quem terá tido o melhor desempenho?
Distribuição normal e padronização de variáveis



Fonte: OpenIntro.

Distribuição normal e padronização de variáveis (Z-score)

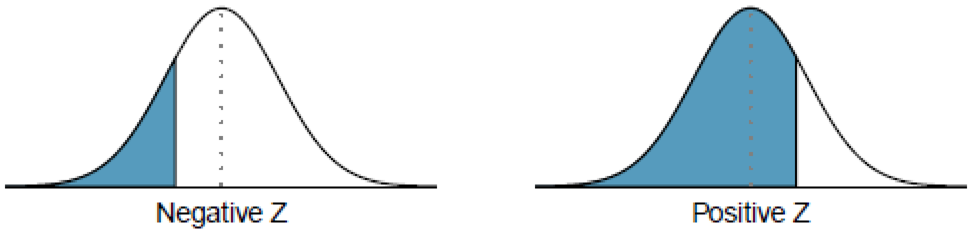
- O Z-score de uma observação é definido como o número de desvios-padrão que cai acima ou abaixo da média.
  - Se a observação se distribui 1 DP acima da média, então Z=1
  - Observações pouco frequentes: |Z| > 2
- Método mais frequentemente utilizado para observações quase normais, mas que pode ser usado com qualquer distribuição.

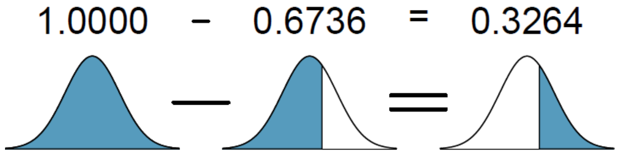
$$\boxed{Z = \frac{\text{observation - mean}}{\text{SD}}}$$

Fonte: OpenIntro.

Distribuição normal e padronização de variáveis (Z-score)

- Quando a distribuição é normal, o Z-score pode ser usado para encontrar áreas de distribuições.
- Essas áreas sob a curva correspondem aos percentis e aumentam da esquerda para a direita até somar 1.
- Um percentil indica o valor abaixo do qual se situa uma determinada percentagem de observações.





Fonte: OpenIntro.

Tabela da distribuição normal padrão

Que percentil corresponde a uma observação = 0? E igual a 1? E a 2?

Z	Second decimal place of Z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	*0.8413*	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

Fonte: OpenIntro.

Distribuição normal e padronização de variáveis (Z-score)

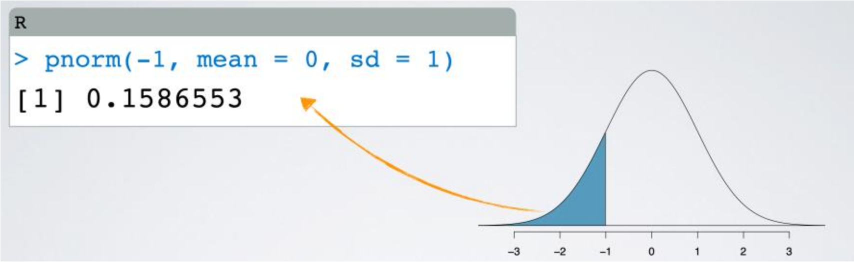
Usar R

?pnorm

“**dnorm** gives the density, **pnorm** gives the distribution function, **qnorm** gives the quantile function, and **rnorm** generates random deviates.”

```
> pnorm (0, 0,1)
[1] 0.5
> pnorm (1, 0,1)
[1] 0.8413447
> pnorm (2, 0,1)
[1] 0.9772499
> pnorm (-1, 0,1)
[1] 0.1586553
```

Um z-score é distribuído normalmente, o que significa que tem μ = 0 e σ = 1 , que, como vimos, são os parâmetros padrão para pnorm() . O z-score é apenas a transformação de qualquer valor normalmente distribuído para um valor padrão normalmente distribuído.



Fonte: OpenIntro.

Exercício 4

- Desenhe a percentagem da distribuição normal padrão ($\mu = 0$, $\sigma = 1$) que se encontra em cada uma das regiões:

a) $Z < -1.35$

b) $Z > 1.48$

c) $-0.4 < Z < 1.5$

d) $|Z| > 2$

Fonte: OpenIntro.

Resolução Exercício 4

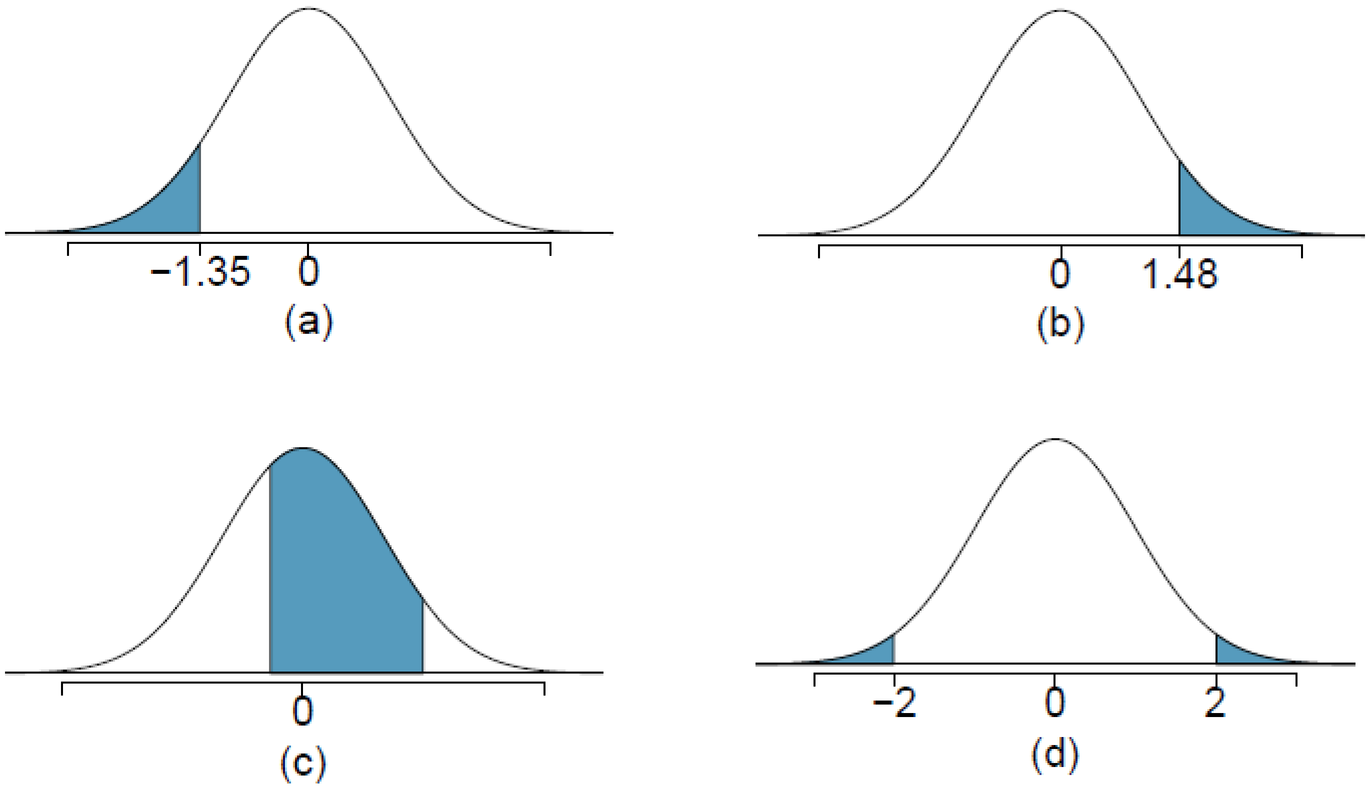
- Desenhe a percentagem da distribuição normal padrão ($\mu = 0$, $\sigma = 1$) que se encontra em cada uma das regiões:

a) $Z < -1.35$

b) $Z > 1.48$

c) $-0.4 < Z < 1.5$

d) $|Z| > 2$



Fonte: OpenIntro.

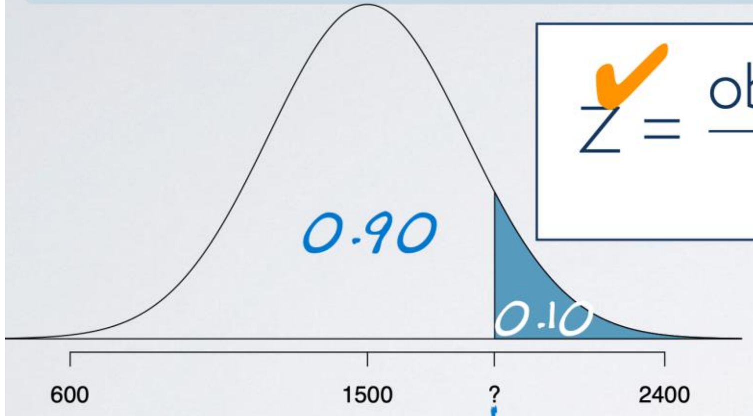
Exercício 5

Uma amiga diz-lhe ter obtido um resultado no exame SAT que se encontra entre os melhores 10%. As pontuações do SAT são distribuídas normalmente com média = 1500 e DP = 300. Qual é a avaliação mínima que a sua amiga pode ter obtido?

Fonte: OpenIntro.

Resolução Exercício 5

A friend of yours tells you that she scored in the top 10% on the SAT. What is the lowest possible score she could have gotten?



$$Z = \frac{\text{observation - mean}}{\text{SD}}$$

$$Z = 1.28 = \frac{X - 1500}{300}$$

$$X = (1.28 \times 300) + 1500 = 1884$$

Z	Second decimal place of Z 0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
			0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
			0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
			0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
			0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
			0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
			0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319

R

```
> qnorm(0.90, 1500, 300)
[1] 1884.465
```

Exercício 6

Um estudo reporta duas distribuições para pontuações num exame misto de literatura e matemática.

N (μ = 462, σ = 119) para a parte de literatura;

e

N (μ = 584, σ = 151) para a parte de matemática.

Com o recurso ao R, determine:

a) A pontuação de um aluno que ficou classificado no percentil 80 na parte de matemática.

b) A pontuação de um aluno que classificou abaixo de 70% dos alunos na parte de literatura.

Fonte: OpenIntro.

Resolução Exercício 6

a) A pontuação de um aluno que ficou classificado no percentil 80 na parte de matemática.

qnorm(.8,584,151) # = 711.0848

b) A pontuação de um aluno que classificou abaixo de 70% dos alunos na parte de literatura.

qnorm(.3, 462,119) # = 399.5963

Exercício 7

A temperatura diária média em Junho em Los Angeles é de 77° F com um desvio-padrão de 5° F. Suponha que as temperaturas em Junho seguem aproximadamente uma distribuição normal.

a) Qual é a probabilidade de observar uma temperatura de 83° F ou mais em Los Angeles durante um dia escolhido aleatoriamente em junho?

b) Quão frios são os 10% menos quentes dos dias do mês de junho em Los Angeles?

Fonte: OpenIntro.

Resolução Exercício 7

A temperatura diária média em junho em Los Angeles é de 77° F com um desvio-padrão de 5° F. Suponha que as temperaturas em junho seguem aproximadamente uma distribuição normal.

a) Qual é a probabilidade de observar uma temperatura de 83° F ou mais em Los Angeles durante um dia escolhido aleatoriamente em junho?

(83-77)/2 #= 1.2

pnorm(1.2) #=0.8849303

Z = 1.2, P(Z > 1.2) = 0.1151

a) Quão frios são os 10% menos quentes dos dias do mês de junho em Los Angeles?

qnorm (.1) # = -1.281552

qnorm (.1,77,5) # = 70.59224

Z = -1.28 -> 70.6º F ou menos

Fonte: OpenIntro.

Exercício 8

- Suponha que os pesos da bagagem despachada pelos passageiros das companhias aéreas seguem uma distribuição quase normal com média de 45 libras e desvio-padrão de 3,2 libras. A maioria das companhias aéreas cobra uma taxa pela bagagem que pesa mais de 50 libras. Determine o percentual de passageiros das companhias aéreas com essa taxa.

Fonte: OpenIntro.

Resolução Exercício 8

- Suponha que os pesos da bagagem despachada pelos passageiros das companhias aéreas seguem uma distribuição quase normal com média de 45 libras e desvio-padrão de 3,2 libras. A maioria das companhias aéreas cobra uma taxa pela bagagem que pesa mais de 50 libras. Determine o percentual de passageiros das companhias aéreas com essa taxa.

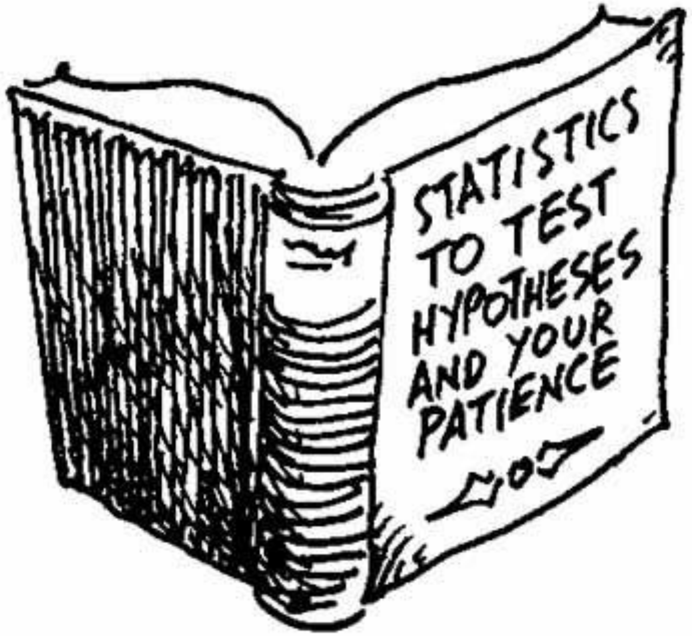
(50-45)/3.2 #=1.56

pnorm(1.56) #=0.9406201

Z = 1.56, P(Z > 1.56) = 0.0594, i.e., 6%

Fonte: OpenIntro.

Testes de hipóteses



Fonte: https://www.pinterest.ph/Jhieeseng/statistics-quotes/.

Testes de hipóteses

Os testes de hipóteses permitem determinar se existem evidências estatísticas suficientes a favor de uma certa hipótese sobre:

- Uma média;
- Uma comparação de médias;
- Uma comparação de distribuições;
- Uma comparação de variâncias;
- Uma comparação do valor global de dois ou vários grupos, não sendo possível comparar médias.

Fonte: Adaptado a partir de D'Hainaut, 1997.

Testes de hipóteses

Exemplos

- O número médio de idas ao cinema por ano na população portuguesa é igual a cinco?
- O número médio de idas ao cinema por ano é igual para os residentes em Faro e em Évora?
- Na população de atletas federados, o tempo médio de realização de uma prova de triatlo é igual antes e depois dos atletas terem feito exercícios de aquecimento?
- Na população de chefes de cozinha, a pontuação média em concursos nacionais é igual pontuação média em concursos internacionais?
- Não existe diferença entre o desempenho dos profissionais de saúde formados em diferentes universidades?

Hipóteses estatísticas

Hipótese nula (H_0): afirmação relativa a uma ou mais populações sobre a forma ou os parâmetros da distribuição dessas populações. Afirmação a ser testada, o *status quo.*

Hipótese alternativa (Ha): afirmação alternativa.

- Não rejeitar a H_0é admitir que as amostras são extraídas de populações cujas características haviam sido especificadas na hipótese.
- Rejeitar a H_0 é admitir que as amostras são extraídas de populações diferentes das que haviam sido especificadas na H_0, pelo menos no que respeita à característica comparada.

➢ A H_0 contém sempre a igualdade.

➢ Realizamos um ensaio de hipóteses sob o pressuposto de que a H_0 é verdadeira. Se os resultados do teste sugerem que os dados não fornecem evidências convincentes para a H_a, mantemos a hipótese nula. Se o fizerem, rejeitamos a H_0 em favor da H_a.

➢ As hipóteses referem-se sempre a uma ou mais populações, nunca às amostras.

Fonte: Adaptado a partir de D'Hainaut, 1997.

Erros de decisão

Nunca podemos estar inteiramente certos quanto à decisão a tomar relativamente à hipótese nula.

No sistema judicial, as pessoas inocentes às vezes são condenadas erroneamente e os culpados às vezes são libertados. Da mesma forma, também podemos tomar uma decisão errada nos testes de hipóteses estatísticas.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

H_0: O réu é inocente
H_a: O réu é culpado
Declarar inocente o réu quando ele é realmente culpado - **Erro tipo II**
Declarar o réu culpado quando ele é realmente inocente - **Erro tipo I**

"better that ten guilty persons escape than that one innocent suffer"
William Blackstone

Fonte: OpenIntro.

Níveis de significância

- A escolha de um nível de significância para um teste é importante em muitos contextos. O nível tradicional é **α = 0.05**.

- Pode adotar-se um nível de significância maior ou menor consoante as consequências das conclusões alcançadas no teste:

  - **α = 0.01**. Nesse cenário, queremos ser muito cautelosos ao rejeitar a H_0, portanto, exigimos evidências muito fortes a favor da H_a.

  - **α = 0.10**. Aqui queremos ser cautelosos ao não rejeitar H_0 quando a H_a for realmente verdadeira.

Fonte: Adaptado a partir de D'Hainaut, 1997.

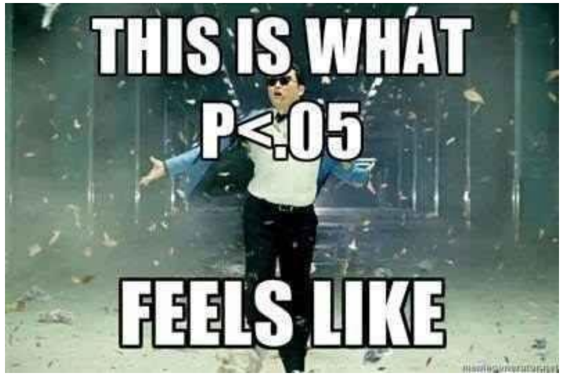
P-value

O *p-value* é a probabilidade de observar dados pelo menos tão favoráveis à hipótese alternativa no nosso conjunto de dados, se a hipótese nula for verdadeira.

Quando *p-value* é menor ou igual que o nível de significância (α), rejeitamos H_0 e concluímos que os dados fornecem fortes evidências apoiando a hipótese alternativa.

Quando o *p-value* for maior que o nível de significância (α), não rejeitamos H_0 e concluímos que não temos evidências suficientes para rejeitar a hipótese nula.

Fonte: OpenIntro.



Fonte: https://www.pinterest.ph/Jhieeseng/statistics-quotes/.

Testes unilaterais e bilaterais

Testes unilaterais

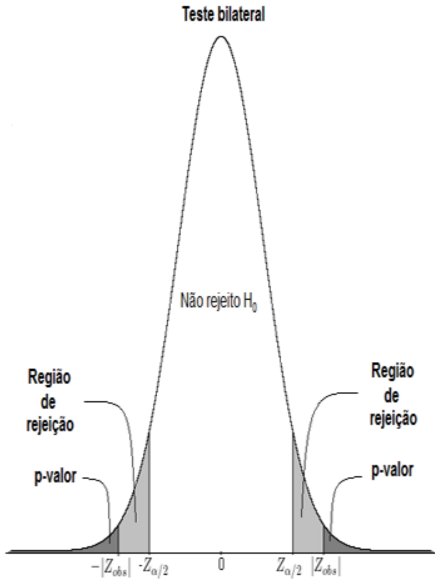
$H_0 : \mu \leq 1$ e $H_a: \mu > 1$ (teste unilateral direito)

$H_0 : \mu \geq 1$ e $H_a: \mu < 1$ (teste unilateral esquerdo)

Testes bilaterais

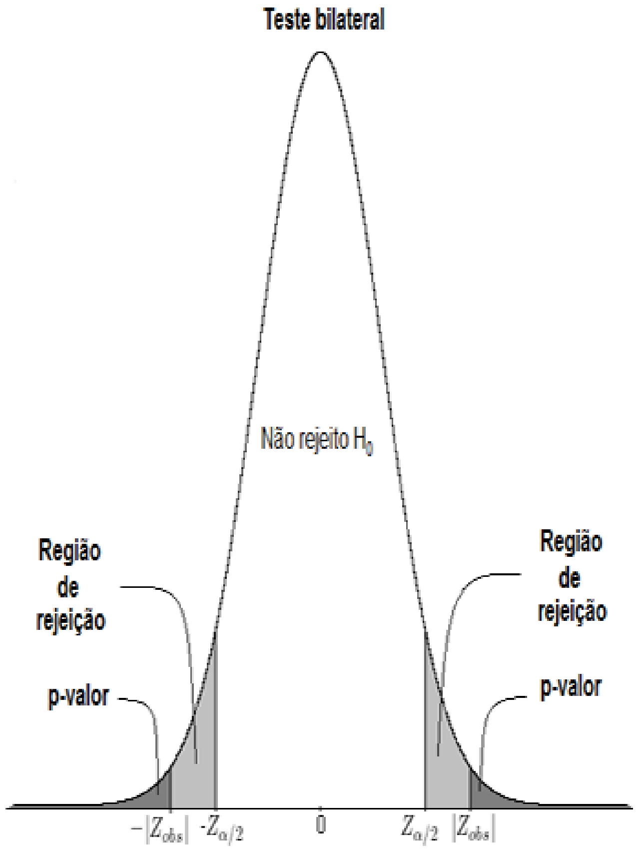
$H_0: \mu = 1$

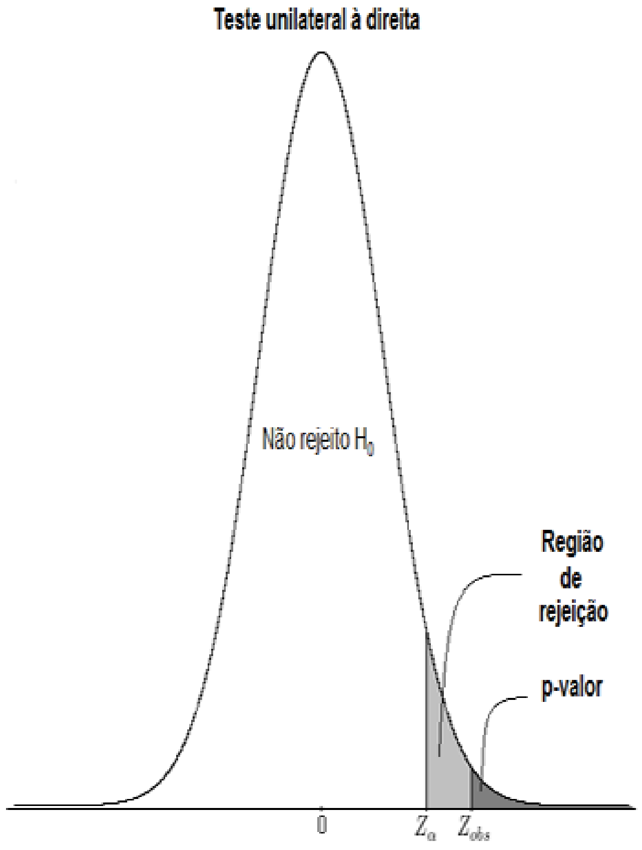
$H_a: \mu \neq 1$



Fonte: Adaptado a partir de D'Hainaut, 1997.

Testes bilaterais e unilaterais





Teste bilateral H_0: $\mu = 1$ e H_a: $\mu \neq 1$

Teste unilateral direito $H_0 : \mu \leq 1$ e H_a: $\mu > 1$

Teste unilateral esquerdo $H_0 : \mu \geq 1$ e H_a: $\mu < 1$

Fonte; http://www.portalaction.com.br/inferencia/512-calculo-e-interpretacao-do-p-valor

Testes de hipóteses

Etapas de um ensaio de hipóteses

1. Pergunta de investigação

1.1 Variáveis, tipo e relação

1.2. Teste estatístico adequado

2. Especificação formal do teste

2.1. Hipóteses

2.2. Significado dos parâmetros

3. Análise exploratória das variáveis

3.1. Condições de aplicabilidade

3.2. Análise exploratória das variáveis e sua relação

4. Aplicação do teste

5. Conclusões contextualizadas

Fonte: OpenIntro.

Teste t para uma média

Perguntou-se a uma amostra aleatória de 206 alunos da Universidade de Duke a quantas universidades eles se candidataram. Em média, os respondentes candidataram-se a 9,7 universidades, com um desvio-padrão de 7. O website da College Board diz que os conselheiros recomendam que os alunos se inscrevam em aproximadamente em 8 universidades.

1. Pergunta de investigação

Será que o número médio de universidades às quais a população de alunos da Duke se candidatam é igual a oito?

1.1 Variáveis envolvidas, tipo e relação

Variável dependente: número de universidades a que os alunos inquiridos se candidataram.

1.2. Teste estatístico pertinente

Teste t para uma média

Fonte: Adaptado a partir de OpenIntro.

Teste t para uma média: exemplo

2. Especificação formal do teste

2.1. Hipóteses

H_0: O número médio de candidaturas a universidades dos alunos da Duke é igual a oito. [H_0: $\mu = 8$]

H_a: O número médio de candidaturas a universidades dos alunos da Duke é diferente de oito. [H_0: $\mu \neq 8$]

2.2. Significado dos parâmetros

μ = número médio de universidades às quais a população de alunos da Duke se candidatam.

Fonte: OpenIntro.

Teste t para uma média: exemplo

3. Análise exploratória

3.1. Condições de aplicabilidade

✓ A amostra é uma amostra aleatória.

✓ As observações são independentes.

✓ A variável dependente é quantitativa.

✓ A variável dependente tem uma distribuição aproximadamente normal.

Fonte: OpenIntro.

Teste t para uma média: exemplo

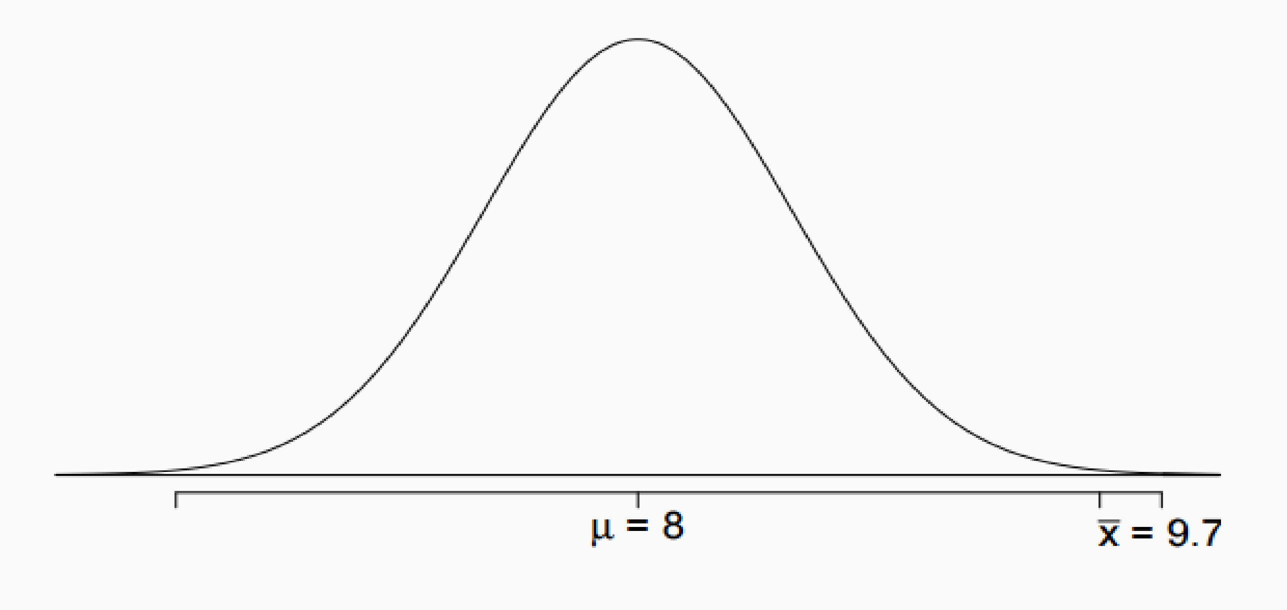
4. Aplicação do teste estatístico

$$Z = \frac{\text{observation - mean}}{\text{SD}}$$

$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

$P(\bar{x} > 9.7 \mid \mu = 8) = P(Z > 3.4) = 0.0003$



Em R: função t.test
Ver: ?t.test

Fonte: OpenIntro.

Teste t para uma média: exemplo

5. Conclusões contextualizadas

Com o objetivo de verificar se a **média real** do número de universidades a que se candidataram os alunos da Duke difere do número recomendado (8), foi recolhida uma **amostra aleatória de 206 estudantes** e realizado um **teste t para uma amostra**. Nesta amostra, o número médio de universidades a que se candidataram os alunos é de 9,7, com um desvio-padrão de 7. Os dados fornecem evidências suficientes de que, em média**, a população de alunos** da Duke se candidataram a mais de oito universidades (T(206)=3.4, p= 0.0003).

Fonte: OpenIntro.