Unicamp
[Estatística]
[Samara F. Kiihl]

Estatística

para experimentalistas

Erik Yuji Goto

Campinas
2021

Sumário

1 Estatística Descritiva

São métodos para resumir e descrever os dados.

É o primeiro passo antes de qualquer análise estatística. O resumo dos dados pode ser feito por:

- Métricas quantitativas: média, mediana, desvio padrão, proporções;
- Ferramentas visuais: gráfico.

Estrutura básica de dados - Variável:

- Qualitativa
  - Nominal: Não existe ordenação. Ex: sexo, estado civil, profissão;
  - Ordinal: Existe uma certa ordem. Ex: escolaridade, estágio da doença, classe social.
- Quantitativa
  - Discreta: Os valores possíveis formam um conjunto finito ou enumerável. Ex: nº filhos, nº ovos de páscoa;
  - Contínua: Os valores possíveis estão dentro de um intervalo dos números reais. Ex: peso, altura, salário.

1.1 Medidas Resumo(estatísticas sumárias)

O **objetivo** é resumir os dados, através de valores que representem o conjunto de dados em relação à alguma característica.

1.1.1 Média Aritmética

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

1.1.2 Mediana

Valor que separa os dados em dois grupos de tamanhos iguais, ou seja, 50% das observações em cada, de acordo com **seus valores ordenados**.

Para n ímpar e para n par respectivamente:

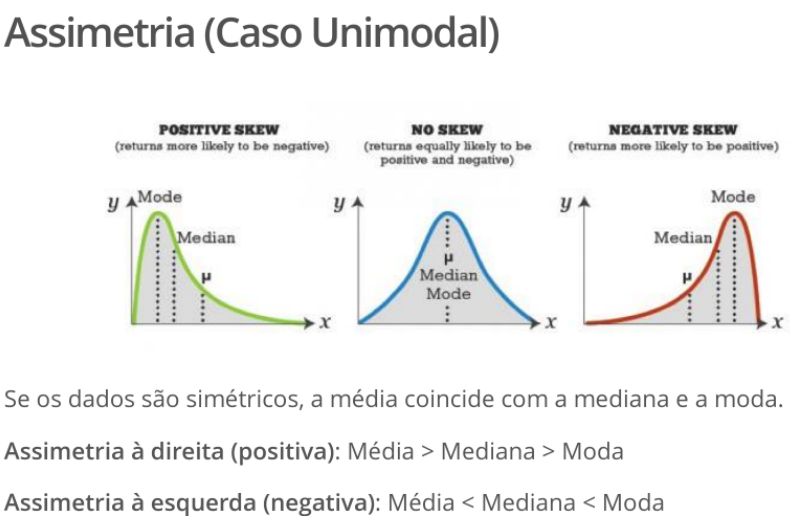
$$Q_2 = x_{(\frac{n+1}{2})}$$

$$Q_2 = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}$$

1.1.3 Moda

É o valor com maior número de ocorrência nos dados.

A moda não precisa ser única



1.1.4 Desvio

O desvio de uma observação x_i da média $\overline{x}$ é a diferença entre a observação e a média dos dados. A média dos desvios ao quadrado é denominada **variância**:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Desvio padrão é a raiz quadrada da variância:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

1.1.5 Coeficiente de Variação

É a razão do desvio padrão **s** pela média $\overline{x}$

$$c_v = \frac{s}{\overline{x}}$$

1.2 Quartis

Consiste em dividir os dados em 4 partes iguais. Para se obter os quartis:

1. Ordene os dados em ordem crescente;
2. Encontre a mediana Q_2;
3. Considere o subconjunto de dados abaixo da mediana. Q_1 é a mediana deste subconjunto de dados;
4. Considere o subconjunto de dados acima da mediana. Q_3 é a mediana deste subconjunto.

Para uma relação simétrica(ou quase), temos as seguintes relações:

$$Q_2 - x_{(1)} \approx x_n - Q_2$$
$$Q_2 - Q_1 \approx Q_3 - Q_2$$
$$Q_1 - x_1 \approx x_n - Q_3$$

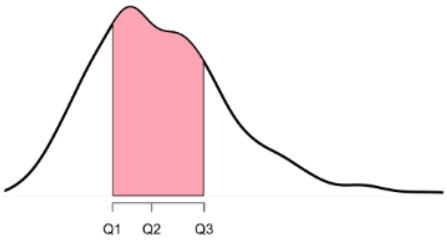
x_1 *é o minimo*
x_n *é o máximo*

Distância entre as medianas e Q_1, Q_3 menores que as distâncias entre os extremos $Q_1 e Q_3$

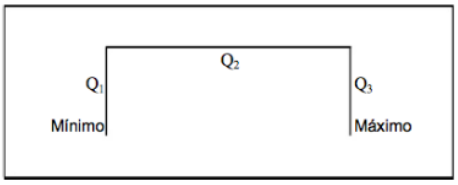
1.2.1 Intervalo Interquartílico

A vantagem do uso de quartis sobre o desvio padrão é que os quartis são mais resistentes a dados extremos, são mais robustos. O intervalo quartílico representa 50% dos dados localizados na parte central da distribuição.

$$IQ = Q_3 - Q_1$$



1.2.2 Esquema dos 5 números e Boxplot



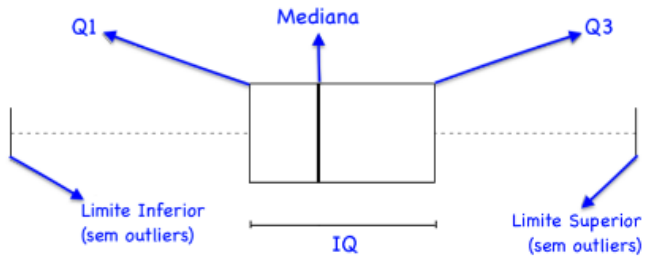
Notação:
$x_{(1)}$: mínimo
$x_{(k)}$: k-ésima observação depois de ordenar os dados
$x_{(n)}$: máximo

Dados discrepantes(outliers): Como regra geral, dizemos que uma obsevação é uma potencial outlier se está:

Abaixo de $Q_1 - 1,5.IQ$
Acima de $Q_3 + 1,5.IQ$

O **boxplot** é a representação gráfica do esquema dos 5 números.

Esse gráfico permite resumir visualmente importante características dos dados (posição, dispersão, assimetria) e identificar a presença de *outliers*.



Para descobrir os limites superior e inferior usamos a fórmula para encontrar os *outliers*.

1.3 Análise Descritiva Bivariada

1.3.1 Coeficiente de Correlação

É uma medida resumo que representa associação linear entre duas variáveis quantitativas.

Dado n pares de observações $(x_1, y_1), (x_2, y_2)...(x_n, y_n)$

$$Corr(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\overline{x}}{s_x}\right)\left(\frac{y_i-\overline{y}}{s_y}\right)$$
$$-1 \leq Corr(X,Y) \leq 1$$

Onde $s_x e s_y$ são os desvios padrões de X e Y.

Corr(X,Y) próximo de 1: X e Y estão positivamente associados e o tipo de associação entre as variáveis é linear.

Corr(X,Y) próximo de 0: X e Y não estão correlacionados.

Atenção: Correlação não implica causa

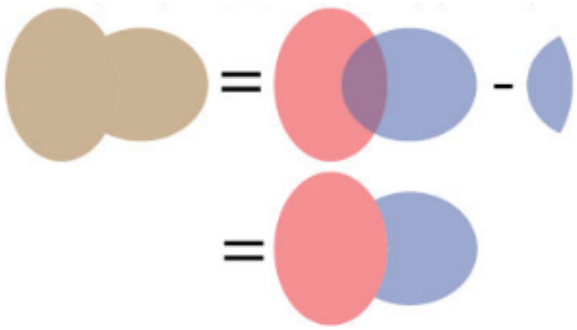
2 Probabilidade

Espaço amostral: Todos os resultados possíveis do experimento, denotado por $\Omega = \{w_1, w_2, ...\}$.

Evento: Dizemos que o evento A ocorreu sempre que o resultado observado pertencer ao subconjunto de elementos do evento A.

2.0.1 Probabilidade de União de Eventos

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



2.0.2 Evento Complementar

Sejam A e B subconjuntos de Ω. Eles são complementares se $A \cap B = \phi$ e $A \cup B = \Omega$. Como $P(A) + P(B) = 1$, então $P(B) = 1 - P(A)$.

2.1 Regras de Contagem

2.1.1 Regra da adição

Suponha que temos dois procedimentos possíveis para executarmos uma tarefa, ou seja, basta executarmos um dos procedimentos para que a tarefa tenha sido executada.

O total de maneiras para executarmos a tarefa é $n_1 + n_2$.

2.2 Regra da multiplicação

Suponha que para realizarmos uma tarefa temos que executar dois procedimentos, denotados $p_1 e p_2$.

O total de maneiras para executarmos a tarefa é dado por $p_1 * p_2$.

2.3 Permutação

Suponha que temos n caixas e queremos dispor os n objetos nessas caixas.

Aplicando a regra da multiplicação, temos que o número de maneiras de permutar n é: $n!$

2.4 Arranjo

O número de maneiras de arranjar n elementos em r caixas é:

$$A(n, r) = \frac{n!}{(n-r)!}$$

2.5 Combinação

A combinação é semelhante ao arranjo, com exceção de que a ordem não importa:

$$C(n, r) = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

2.6 Probabilidade Condicional

Encontrar a probabilidade de um evento quando você tem alguma outra informação sobre o evento.

A probabilidade condicional de **B** dado **A** é expressa por $P(B \mid A)$.

Supondo que o resultado do experimento esteja contido no evento A.

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}$$

2.7 Independência

Quando $P(B \mid A) = P(B)$, informação sobre A *não altera* a probabilidade do evento B, dizemos que A e B são **independentes**.

$$P(A \cap B) = P(A)P(B)$$

2.8 Teorema das Probabilidades Totais

A probabilidade do evento A é uma média ponderada de $P(A \mid B)$ e $P(A \mid B^c)$. O peso de cada probabilidade condicional é a probabilidade do evento que está sendo levado em conta ao calcular a probabilidade condicional de A.

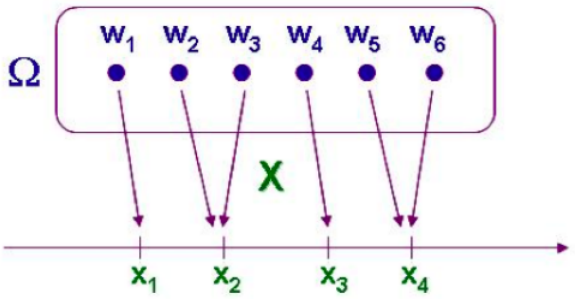
$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$$

2.9 Teorema de Bayes

$$P(B_i \mid A) = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^{n} P(A|B_i)P(B_i)}$$

3 Variável Aleatória

Uma função X que associa a cada elemento do espaço amostral um valor num conjunto enumerável de pontos da reta é denominada *variável aleatória discreta*.



3.1 Distribuição de probabilidade

Associa uma probabilidade $P(X = x)$ para cada valor possível, x, da variável aleatória X.

- Seja X uma v.a. discreta com n valores possíveis denotados por $x_1, x_2, \ldots, x_n$.
- $P(X = x_i)$ denota a probabilidade de que a v.a. X assuma o valor x_i.
- O conjunto de todas essas probabilidades (para cada x_i) representa a **distribuição de probabilidade** de X.

X	x_1	x_2	x_3	...
$P(X = x)$	$P(X = x_1)$	$P(X = x_2)$	$P(X = x_3)$	...

- Como X só pode assumir valores entre $x_1, x_2, \ldots, x_n$, temos que:

$$\sum_{i=1}^{n} P(X = x_i) = 1$$

3.2 Função de Distribuição Acumulada

É definida por

$$F(x) = P(X \leq x), x \in \Re$$

Exemplo:

Doses (X)	1	2	3	4	5
$P(X = x)$	0.245	0.288	0.256	0.145	0.066

Note que a f.d.a. de X = número de doses é definida para qualquer valor real, logo:

$$F(x) = \begin{cases} 0 & x < 1 \\ 0.245 & 1 \leq x < 2 \\ 0.533 & 2 \leq x < 3 \\ 0.789 & 3 \leq x < 4 \\ 0.934 & 4 \leq x < 5 \\ 1 & x \geq 5 \end{cases}$$

3.3 Esperança

A esperança(ou valor esperado) da variável X é dada por:

$$E(X) = \mu = \sum_{i=1}^{n} x_i P(X = x_i)$$

Não podemos interpretar E(X) como o valor que esperamos que X irá assumir, mas sim como uma **média** dos valores observados de X ao longo de muitas repetições do experimento aleatório.

Propriedades da Esperança

1. Para qualquer v.a. X e constantes a e b:
   $E(aX + b) = aE(X) + b$
2. $E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$

3. Esperança de uma funçao de uma variável aleatória - útil na hora de calcular a variância:
   $E[g(X)] = \sum_i g(x_i)p(x_i)$

3.4 Variância

Queremos uma medida para quantificar quão distantes os valores da v.a. X estão da sua esperança. A variância é calculada por:

$$Var(X) = E(x^2) - [E(X)]^2$$

Propriedades da Variância

1. Para qualquer v.a. X e ctes a e b:
   $Var(aX + b) = a^2 Var(X)$

2. Se $x_1, x_2, ...x_n$ são variáveis aleatórias independentes:
   $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$

3.5 Mediana e Moda

Sabemos que a média é encontrada pela esperança.
A *mediana(Md)* é o valor que satisfaz:

$$P(X \geq Md) \geq \frac{1}{2} \text{ e } P(X \leq Md) \geq \frac{1}{2}$$

A *moda(Mo)* é o valor da variável X com maior probabilidade de ocorrência.

$$P(X = M_o) = max\{p_1, p_2, ...\}$$

3.6 Principais Modelos Discretos

3.6.1 Uniforme Discreta

Ocorre se cada valor possível tem a mesma probabilidade de ocorrer. Exemplo: Lançamento de um dado.

$$P(X = x_i) = \frac{1}{k}$$

Notação: $X \sim$ Uniforme Discreta$\{1, 2, ...\}$

3.6.2 Bernoulli

Quando cada observação de um experimento aleatório é **binária**, tem apenas dois resultados possíveis:

- Sucesso;
- Fracasso.

Seja X assumindo apenas valores 0 e 1, onde X= 1 corresponde a sucesso e seja p a probabilidade de sucesso

$$P(X = x) = p \text{ , se x} = 1$$
$$P(X = x) = 1 - p \text{ , se x} = 0$$

Ou de forma equivalente:

$$P(X = x) = p^x(1-p)^{1-x} \text{ , para x} = 0{,}1$$

Notação: $X \sim b(p)$

Além disso, calculamos a **esperança** e a **variância** como sendo

$$E(x) = p$$
$$Var(X) = p(1-p)$$

3.6.3 Modelo Binomial

Três condições para ser usado:

1. **n** ensaios de Bernoulli;
2. Os ensaios são independentes;
3. A probabilidade de sucesso **p** é a mesma em todo o ensaio.

Notação: $X \sim Bin(n, p)$

A probabilidade de se observar x é dada pela expressão geral:

$$p(X = x) = \binom{n}{x} p^x(1-p)^{n-x}, \text{ x} = 0,\ 1,\ \ldots\ \text{n}$$

A **esperança** e **variância** de uma v.a. Binomial são dados por:

$$E(X) = np$$
$$Var(X) = np(1-p)$$

3.6.4 Distribuição Geométrica

Seja X a v.a. que representa o número de ensaios de Bernoulli até a ocorrência do primeiro sucesso. Então dizemos que X segue uma distribuição **Geométrica** com parâmetro p.

Notação: $X \sim G(p)$

A probabilidade de se observar x é:

$$P(X = x) = (1-p)^{x-1}p \text{ , x} = 1,\ 2,\ \ldots$$

A esperança e variância são

$$E(X) = \frac{1}{p}$$
$$Var(X) = \frac{1-p}{p^2}$$

A função de distribuição acumulada de uma v.a. G(p) é dada por:

$$F(x) = P(X \leq x) = 1 - (1-p)^x$$

Propriedade de perda de Memória: O fato de já termos observado m fracassos sucessivos não muda a probabilidade do número de ensaios até o primeiro sucesso ocorrer.

3.6.5 Hipergeométrica

- População dividida em duas características;
- Extração casuais **sem reposição**.

O que queremos resolver:

- N objetos;
- r têm a característica A;
- N-r têm a característica B;
- Um grupo de n elementos é escolhido ao acaso, dentre os N possíveis, sem reposição.

Objetivo é calcular a probabilidade de que este grupo de n elementos contenha x elementos com característica A.

Notação: $X \sim Hip(N, n, r)$
N = tamanho da população
n = tamanho da amostra
r = número de indivíduos com a característica(em relação à população total)

A probabilidade de se observar x é dada por:

$$P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}} \text{ , } 0 \leq x \leq min\{r, n\}$$

Esperança e Variância

$$E(X) = \frac{nr}{N}$$
$$Var(X) = \frac{nr}{N}(1 - \frac{r}{N})\frac{(N-n)}{(N-1)}$$

3.6.6 Distribuição de Poisson

Muitas vezes, em problemas em que seria natural usar a distribuição binomial, temos **n** muito grande e p muito pequeno. Nestes casos podemos usar o teorema de Poisson:

$$P(X = x) \approx \frac{e^{-np}(np)^x}{x!} \text{ , x} = 0, 1, 2...$$

Considera-se o critério $np \leq 7$ para usar essa aproximação.

Outra regra prática: $n \geq 20$ e $p \leq 0,05$

Uma variável aleatória x tem distribuição de Poisson com parâmetro $\lambda > 0$, se sua função de probabilidade é dada por:

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ , x} = 0, 1, 2, ...$$

λ é chamado de taxa de ocorrência

Notação: $X \sim P(\lambda)$

$$E(X) = Var(X) = \lambda$$

4 Variáveis Aleatórias Contínuas

Variável Aleatória com valores possíveis em um intervalo no conjunto de números reais.

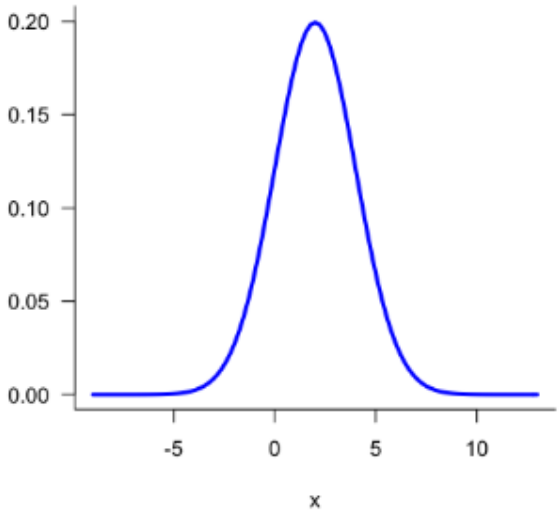


Figura 1: $\int_{-\infty}^{\infty} f(x)dx = 1$

A probabilidade de que uma v.a. x contínua pertença a um intervalo da reta (a, b], $a < b$ é:

$$P(a < x < b) = \int_a^b f(x)dx$$

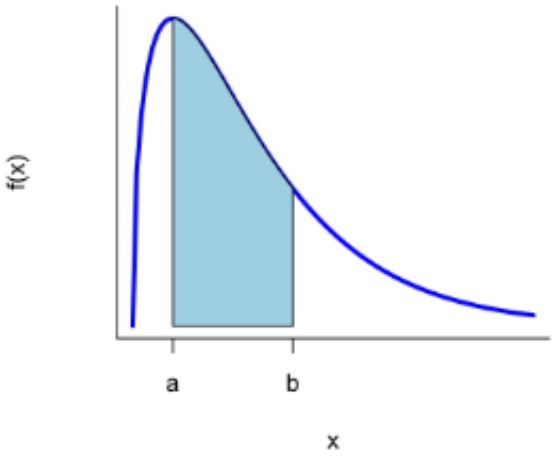


Figura 2: Intervalo a-b

4.1 Função de Distribuição Acumulada

$$F_x(x) = P(X \leq x) = \int_{-\infty}^{x} f_x(u)du$$

Propriedade: Toda v.a X contínua tem probabilidade pontual nula:

$$P(X = x) = 0$$

4.2 Esperança

$$E(X) = \int_{-\infty}^{\infty} x f_x(x) dx$$

Propriedade: Seja X uma v.a. contínua com densidade f_x o *k-ésimo momento* de X é dado por:

$$E(X^k) = \int_{-\infty}^{\infty} x^k f_x(x) dx$$

4.3 Variância

$$Var(x) = E(X^2) - [E(X)]^2$$
$$\sigma = \sqrt{Var(X)}$$

4.4 Distribuição Uniforme

A v.a X tem distribuição uniforme no intervalo [a, b], $a < b$ se:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b, \\ 0 & \text{caso contrário.} \end{cases}$$

Figura 3: Distribuição Uniforme

Notação: $X \sim U[a, b]$ ou $X \sim U(a, b)$

4.4.1 Distribuição Acumulada

$$F_X(x) = \begin{cases} 0 & x < a, \\ \int_a^x \frac{1}{b-a} dt = \frac{x-a}{b-a} & a \leq x \leq b, \\ 1 & x > b. \end{cases}$$

Figura 4: Distribuição Acumulada

4.4.2 Esperança e Variância

$$E(X) = \frac{(a+b)}{2}$$
$$Var(X) = \frac{(b-a)^2}{12}$$

4.5 Distribuição Exponencial

Uma v.a X possui distribuição exponencial com parâmetro $\lambda(\lambda > 0)$

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0 & \text{caso contrário.} \end{cases}$$

Figura 5: Distribuição Exponencial

Notação: $X \sim Exp(\lambda)$

4.5.1 Distribuição Acumulada

$$F_X(x) = \begin{cases} \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x} & x \geq 0, \\ 0 & \text{caso contrário.} \end{cases}$$

Figura 6: Distribuição Acumulada

4.5.2 Esperança e Variância

$$E(X) = \frac{1}{\lambda}$$
$$Var(X) = \frac{1}{\lambda^2}$$

4.6 Distribuição Normal - Gaussiana

Uma v.a X possui distribuição normal com parâmetros μ e σ^2 se:

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp[-\frac{(x-\mu)^2}{2\sigma^2}], \; -\infty < x < \infty$$

Notação: $X \sim N(\mu, \sigma^2)$

4.6.1 Esperança e Variância

$$E(X) = \mu$$
$$Var(X) = \sigma^2$$

4.6.2 Distribuição Normal Padrão

Se $X \sim N(\mu, \sigma^2)$

$$Z = \frac{x-\mu}{\sigma} \sim N(0, 1)$$

4.6.3 Simetria da Distribuição Normal

A distribuição normal é simétrica, portanto:

$$P(Z < -z) = P(Z > z)$$

Além disso, temos a seguinte propriedade para a função de distribuição acumulada:

$$\phi(x) = 1 - \phi(-x)$$

A probabilidade de um intervalo é dada por:

$$P(a < Z < b) = \phi(b) - \phi(a)$$

Onde $\phi(x)$ é a função de densidade acumulada e é encontrada a partir da tabela normal.

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
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
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998

Figura 7: Tabela para encontrar phi

4.7 Aproximação Normal para uma Binomial

Seja $X \sim Bin(n, p)$. Se **n** é suficientemente grande, a distribuição de X pode ser aproximada pela distribuição normal, isto é:

$$X \sim N(np, np(1-p))$$

Em geral, para que a aproximação para a normal seja utilizada:

$$np \geq 10$$
$$n(1-p) \geq 10$$

5 Distribuição Amostral

Para uma amostra de tamanho **n** a partir de uma população:

- Com média μ e variância σ^2:
  $\bar{X}$: $E(\bar{X}) = \mu$ e $Var(\bar{X}) = \frac{\sigma^2}{n}$
  Erro Padrão: $EP(\bar{X}) = \sqrt{Var(\bar{X})} = \frac{\sigma}{\sqrt{n}}$
- Com proporção populacional p
  $\hat{p}$: $E(\hat{p}) = p$ e $Var(\hat{p}) = \frac{p(1-p)}{n}$
  Erro Padrão: $EP(\hat{p}) = \sqrt{Var(\hat{p})} = \sqrt{\frac{p(1-p)}{n}}$

5.1 Teorema do Limite Central

Para uma amostra aleatória $X_1, ..., X_n$ coletada de uma população com média μ e variância σ.

A distribuição amostral de $\bar{X}$ aproxima-se de uma *Distribuição Normal* de média μ e variância $\frac{\sigma^2}{2}$, quando **n** for suficientemente grande:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

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

Obs: o resultado vale para $\hat{p}$, com $\mu = p$ e $\sigma^2 = p(1-p)$.

Figura 8: Teorema do Limite Central

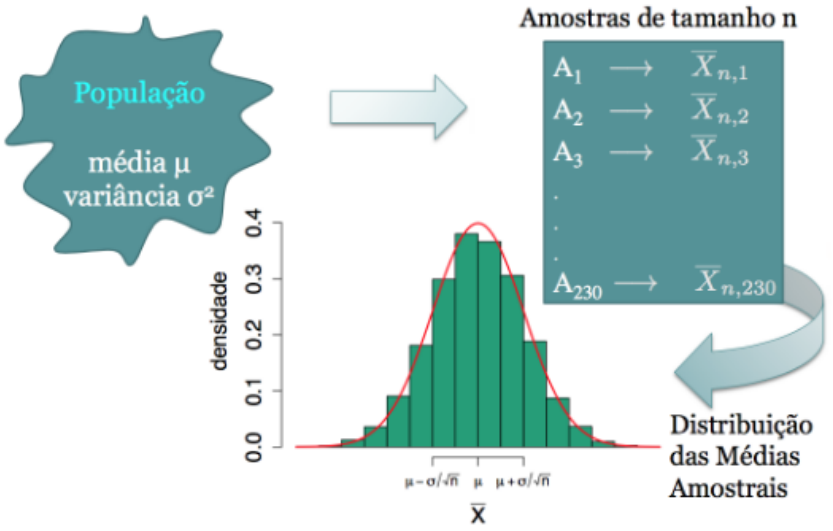


Figura 9: Teorema do Limite Central

6 Intervalo de Confiança

6.1 Intervalo de Confiança para p

Os intervalos de $100(1-\alpha)\%$ de confiança para p podem ser de duas formas:

1. Método Conservador

$$IC_1(p, 1-\alpha) = [\hat{p} - z_{\alpha/2}\sqrt{\frac{1}{4n}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{1}{4n}}]$$

2. Usando $\hat{p}$ para estimar o erro padrão

$$IC_2(p, 1-\alpha) = [\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$$

Veja que nos dois casos, os IC's são da forma $\hat{p}\pm$ margem de erro.

6.1.1 Como encontrar $z_{\alpha/2}$?

$$1-\alpha = P(|Z| \leq z_{\alpha/2}) = P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 2\phi(z_{\alpha/2}) - 1$$

Portanto,

$$\phi(z_{\alpha/2}) = 1 - \frac{\alpha}{2}$$

Procure na tabela o valor de z tal que a probabilidade acumulada até o valor de z seja $1-\alpha/2$.

6.1.2 Interpretação do Intervalo de Confiança

Se várias amostras aleatórias forem retiradas da população e calcularmos um IC de 95% para cada amostra, cerca de 95% desses intervalos irão conter a verdadeira proporção na população, p.

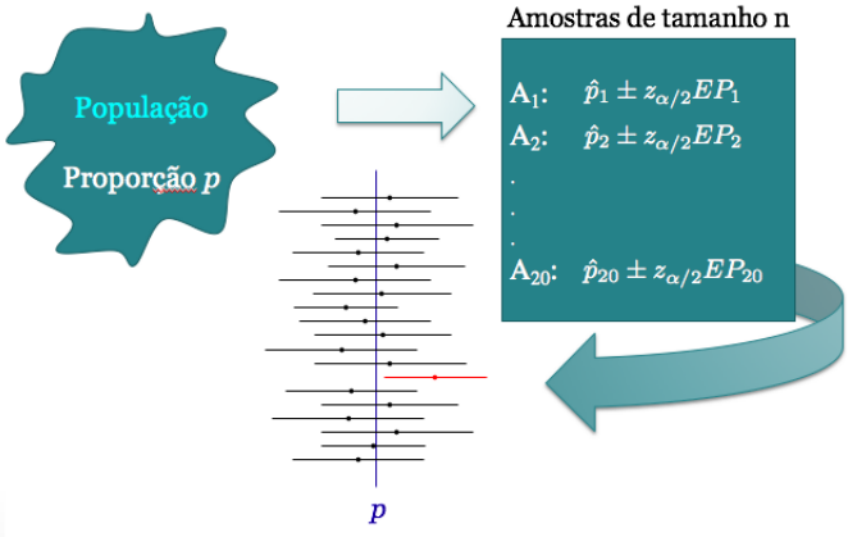


Figura 10: Interpretação IC

6.1.3 Tamanho da Amostra para Estimar p

Em geral, para uma margem de erro **m**

$$n = (\frac{z_{\alpha/2}}{2m})^2$$

6.2 Intervalo de Confiança para a média populacional μ

Seja $X_1, ..., X_n$ uma a.a de uma população com média μ e variância σ^2 conhecida. Então:

$$Z = \frac{\bar{X}_n - \mu}{\sigma\sqrt{n}} \sim N(0,1)$$
$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

Um intervalo de $100(1-\alpha)\%$ de confiança para mu é dado por:

$$IC(\mu, 1-\alpha) = [\bar{x} - z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{\alpha/2}.\frac{\sigma}{\sqrt{n}}]$$

6.2.1 Interpretação do Intervalo de Confiança para μ

Imagine que seja possível coletar uma amostra de tamanho n da população várias vezes. Para cada vez, você calcula $\bar{a}$ e constrói um IC de 95% para μ.Imagine também que você conhece μ e conte quantos dos intervalos contêm μ. A proporção de intervalos que contêm μ será próxima a 0.95.

6.2.2 Tamanho da Amostra

Em geral, para uma margem de erro m e confiança $100(1-\alpha)\%$

$$n = (\frac{z_{\alpha/2}}{m})^2\sigma^2$$

6.2.3 Interpretação do Intervalo de Confiança para μ: σ desconhecido

Neste caso, usaremos a variância amostral(s^2) como uma estimativa de σ^2:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Como consequência, não temos mais distribuição Normal, mas sim a distribuição t-student com n-1 graus de liberdade:

$$T = \frac{\bar{X}_n - \mu}{\sqrt{s^2/n}} \sim t_n - 1$$
$$P(-t_{n-1,\alpha/2} < T < t_{n-1,\alpha/2}) = 1 - \alpha$$

Um intervalo de $100(1-\alpha)\%$ de confiança para μ é dado por:

$$IC(\mu, 1-\alpha) = [\bar{x} - t_{n-1,\alpha/2}\cdot\frac{s}{\sqrt{n}}; \bar{x} + t_{n-1,\alpha/2}\cdot\frac{s}{\sqrt{n}}]$$

Os valores da distribuição t-student também encontram-se tabelados.

7 Teste de Hipóteses

1. Suposições
   O teste é válido sob algumas suposições. A mais importante assume que os dados foram coletados aleatoriamente;

2. Hipóteses

   - **Hipótese Nula(H_0)** afirma que o parâmetro populacional assume um dado valor;
   - Hipótese **Alternativa(H_A)** afirma que o parâmetro populacional assume outros valore, diferente do valor descrito na H_0.

   Em testes de hipóteses, mantêm-se a favor de H_0 a menos que os dados tragam grande evidência contra.

3. Estatística do Teste
   Descreve quão longe do parâmetro populacional usado na H_0 a estimativa está;

4. Valor-de-p
   Para interpretar uma estatística do teste, vamos usar uma probabilidade para resumir a evidência contra H_0, chamada valor-de-p;

5. Conclusão
   Geralmente, fixamos o **nível de significância** do teste(α), e usamos a seguinte regra - é comum usarmos $\alpha = 0{,}05$.

   ✓ valor-de-p $< \alpha$: rejeitamos H_0
   ✓ valor-de-p $> \alpha$: não rejeitamos H_0

7.0.1 Valor-de-p

$H_A : p \neq p_0$ (bilateral): valor-de-p=$P(|Z| \geq |z_{obs}|)$

$H_A : p < p_0$ (unilateral à esquerda): valor-de-p=$P(Z \leq z_{obs})$

$H_A : p > p_0$ (unilateral à direita): valor-de-p=$P(Z \geq z_{obs})$

Figura 11: Valor de p

7.0.2 Estatística do Teste

Baseada na distribuição amostral de $\hat{p}$.

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}, \sim N(0,1)$$

7.1 Teste de Hipóteses para a média σ conhecido

Usamos o mesmo passo a passo que foi feito para as proporções:

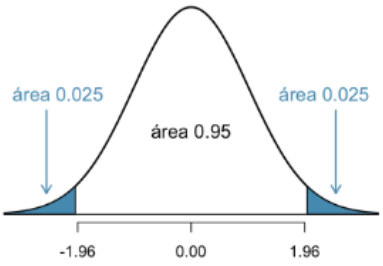
- Suposições;
- Hipóteses;
- Estatística do teste:
  $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
- Valor-de-p

7.1.1 Região Crítica

Conjunto de valores da estatística do teste para as quais a hipótese nula é rejeitada.

Quando o teste for bilateral: $H_0 : \mu = 500$ vs $H_a : \mu \neq 500$

A região critíca, para $\alpha = 0.05$, é a área em azul na figura abaixo:



Decisão: Rejeitamos H_0 se $z_{obs} < -1.96$ ou $z_{obs} > 1.96$.
No nosso exemplo, $z_{obs} = -7.5$. Portanto, rejeitamos H_0.

Figura 12: Região Crítica

7.2 Teste de Hipóteses para média(σ desconhecido)

Quando σ é desconhecido e a amostra é pequena ($n < 30$) devemos utilizar a distribuição t.

$$t = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

Valor-de-p: $P(\mid t_{n-1} \mid \geq \mid t_{obs} \mid) = 2p(t_{n-1} \geq t_{obs})$

8 Inferência para duas populações

8.1 Intervalo de Confiança para Duas Médias

Caso 1: Variâncias diferentes e conhecidas

$$\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m})$$

Normalizando,

$$Z = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} \sim N(0,1)$$

Intervalo de Confiança

$$IC(\mu_1 - \mu_2, 1-\alpha) = (\bar{X} - \bar{Y}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$$

Caso 2: Variâncias iguais e conhecidas
É um caso particular em que as variâncias são conhecidas e idênticas, isto é, $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Caso 3: Variâncias iguais e desconhecidas
Usamos uma estimativa de σ^2 e a distribuição normal é substituída pela *distribuição t.*

$$Sp^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$$

$$T = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{Sp^2(\frac{1}{n}+\frac{1}{m})}} \sim t_{n+m-2}$$

$$IC(\mu_1 - \mu_2, 1-\alpha) = (\bar{X} - \bar{Y}) \pm t_{n+m-2} \sqrt{Sp^2(\frac{1}{n} + \frac{1}{m})}$$

8.2 Intervalo de Confiança para Duas Proporções

Queremos encontrar um IC de confiança para a diferença entre as proporções p_1 e p_2 , ou seja, um IC para $p_1 - p_2$.

Normalizando,

$$Z = \frac{(\hat{p_1}-\hat{p_2})-(p_1-p_2)}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1}+\frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$$

Intervalo de Confiança é dado por:

$$IC(p_1 - p_2, 1-\alpha) = (\hat{p_1} - \hat{p_2}) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

8.3 Teste de Hipóteses para duas médias

Caso 1: Variâncias diferentes e conhecias
Temos interesse em testar as hipóteses

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$
$$\text{VS}$$
$$H_A : \mu_1 - \mu_2 \neq \Delta_0$$
$$H_A : \mu_1 - \mu_2 < \Delta_0$$
$$H_A : \mu_1 - \mu_2 > \Delta_0$$

Estatística do Teste

$$z_{obs} = \frac{(\bar{X}-\bar{Y})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}}} \sim N(0,1)$$

Depois de calcular a Estatística do Teste, o próximo passo é calcular o *valor-de-p* e comparar com o nível de significância.

Caso 3: Variâncias iguais e desconhecidas

$$Sp^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$$

Prosseguir usando **t de student**

8.4 Teste de Hipóteses para Duas Proporções

Considere $X_1, ..., X_{n1}$ e $Y_1, ..., Y_{n1}$ duas amostras independentes de ensaios de Bernoulli tal que $X \sim b(p_1)$ e $Y \sim b(p_2)$, com probabilidade p_1 e p_2 de apresentarem uma certa característica.

$$H_0 : p_1 - p_2 = 0$$
$$\text{VS}$$
$$H_A : p_1 - p_2 \neq 0$$
$$H_A : p_1 - p_2 < 0$$
$$H_A : p_1 - p_2 > 0$$

Para calcular a estatística do teste precisamos de:

$$\hat{p} = \frac{n_1\hat{p}_1+n_2\hat{p}_2}{n_1+n_2}$$

Estatística do Teste

$$z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}} \sim N(0,1)$$