Aula 01 - Descritiva I

População $\xrightarrow{\text{Técnicas de Amostragem}}$ Amostra

Amostra $\xrightarrow{\text{Análise Descritiva}}$ Informações contida nos dados

Informações contida nos dados $\xrightarrow{\text{Inferência Estatística}}$ Conclusões sobre as características da população

Conclusões sobre as características da população $\dashrightarrow$ População

Estatística Descritiva: Etapa inicial da análise utilizada para descrever, organizar e resumir os dados observados.

Probabilidade: A teoria das probabilidades auxilia na modelagem de fenômenos aleatórios, ou seja, aqueles em que está presente a incerteza.
É uma ferramenta fundamental para a inferência estatística.

Inf. Estatística: Cjto de técnicas que permite, a partir de dados amostrais, tirar conclusões sobre a população de interesse, controlando erros.

Estatística Descritiva

- Resumo dos Dados

Classificação das Variáveis
↓
qualquer caract. associada à uma população

Qualitativa $\begin{cases} \text{Nominal} & \text{sexo, cor dos olhos} \\ \text{Ordinal} & \text{classe social, grau de instrução} \end{cases}$

Quantitativa $\begin{cases} \text{Contínua} & \text{peso, altura, salário, idade} \\ \text{Discreta} & \text{nº de filhos, nº de carros} \end{cases}$

Para Variáveis ~~Qualitativas~~ Quantitativas

Medidas de Posição

Mínimo (min): a menor observação

Máximo (max): a maior observação

Moda (mo): é o valor (ou atributo) que ocorre em maior frequência

Média ($\bar{x}$):

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

Mediana: A mediana é o valor da variável que ocupa a posição central de um conjunto de n dados ordenados.

∴ Posição da ordem da mediana: $\frac{n+1}{2}$

Percentil: O percentil de ordem p×100 (0<p<1), em um conjunto de dados de tamanho n, é o valor da variável que ocupa a posição p×(n+1) do conjunto de dados ordenados.

Medidas de Dispersão.

Finalidade: Encontrar um valor que resume a variabilidade de um cjto de dados

- Amplitude: máx − min
- Intervalo-Interquartil. Q3 - Q1
- Variância: $\sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1) = S^2$ ou $S^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$
- Desvio padrão: $S = \sqrt{\text{Variância}}$
- Coeficiente de variação: $CV = \frac{S}{\bar{x}} \times 100\%$
  - é uma medida de disp. relativa
  - elimina o efeito da magnitude dos dados
  - exprime a variabilidade em relação à média

Aula 1B - Descritiva II

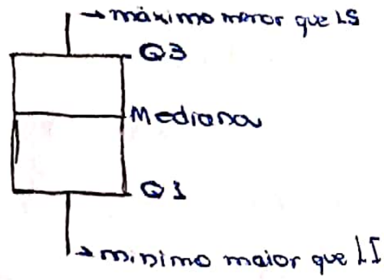
uma lista dos val. individuais ou dos intervalos de valores que a variável pode assumir com as respectivas frequências de ocorrência,

- Resumo de dados (Tabela de distribuição de frequências)

  Gráficos para variáveis quantitativas (Strip Chart ou Dotplot, Boxplot, Histograma)

  Boxplot: representa os dados através de um retângulo construído com os quartis e fornece várias informações, incluindo a existência de valores extremos.

$$LS = Q3 + 1{,}5\,(Q3 - Q1)$$



$$LI = Q1 - 1{,}5\,(Q3 - Q1)$$

x → discrepante

Histograma: Agrupar os dados em intervalos de classes

Bases iguais: construir um retângulo para cada classe, com base igual ao tamanho da classe e altura proporcional à frequência da classe.

Bases diferentes: construir um retângulo para cada classe, com base igual ao tamanho da classe e área do retângulo igual a freq. relativa da classe (fr). A altura será dada por $h = fr / base$ (dens. da freq.)

- Variáveis Qualitativas (Os dados também podem ser resumidos construindo-se uma tabela de dist. de frequências, que quantifica a freq. das distintas categorias)

  Gráficos (de setores, de barras)

Exercício. Produtor de leite → qtde de litros que cada uma de suas 20 vacas ao longo da semana

valores observados

22, 25, 27, 28, 29, 32, 32, 36, 37, 37, 37, 44, 45, 48, 52, 53, 55, 57, 60, 65

Medianas, Q_1, Q_3

a) Lembremos que a mediana é o valor que ocupa a posição central dos dados. Ademais, o percentil de ordem $p \times 100\%$ $(0 < p < 1)$ é o valor que ocupa a $p \times (n+1)$ posição no rol. Daí

Posições: Mediana ⇒ posição $= 0{,}5 \times (n+1) = 0{,}5 \times 21 = 10{,}5$

Q_1 ⇒ posição $= 0{,}25 \times (n+1) = 0{,}25 \times 21 = 5{,}25$

Q_3 ⇒ posição $= 0{,}75 \times (n+1) = 0{,}75 \times 21 = 15{,}75$

Valores: mediana ⇒ $Md = (37+37)/2 = 37$

pr. quartil ⇒ $Q_1 = (29+32)/2 = 30{,}5$

ter. " ⇒ $Q_3 = (52+53)/2 = 52{,}5$

b) Intervalo Interquartil e pontos atípicos

Basta lembrar que

$$IQ = Q_3 - Q_1 = 52{,}5 - 30{,}5 = 22,$$

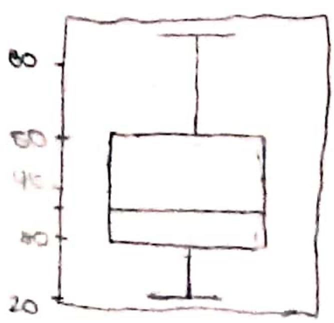
e que os limites inf. e superior são

$$LI = Q_1 - 1{,}5 \times IQ = 30{,}5 - 33 = -2{,}5$$

$$LS = Q_3 + 1{,}5 \times IQ = 52{,}5 + 33 = 85{,}5$$

Logo, não há pontos atípicos (basta ver os dados ordenados)

c)



Não há ptos atípicos.

∃ assimetria

... a variabilidade em relação à média

Aula 02 - Descritiva III

Associação entre Variáveis

- Associação entre variáveis qualitativas

Apenas mostrou a construção das tabelas de Contingência

Verificar se há associação através da

- porcentagem segundo as colunas, ou
- porcentagem " as linhas

Ex. Classe:

	Carrinho	Escapular	Lateral	Total
Feminino	53 (41,4%)	59 (46,1%)	16 (12,5%)	128 (100%)
Masculino	27 (27,6%)	64 (65,3%)	7 (7,1%)	98 (100%)
Total	80 (35,4%)	123 (54,4%)	23 (10,2%)	226 (100%)

Parece ∃ relação entre sexo e tipo de mochila

Maioria Meninos → escapular

Maioria Meninos → escap/carrinho

- Associação entre variáveis ~~qualitativas~~ quantitativas
- Correlação: quantifica a força da relação
- Regressão: explicita a forma da relação
- Diagrama de dispersão 

- Coeficiente de correlação linear. Uma medida que avalia o quanto a "nuvem de pontos" no diagrama de dispersão aproxima-se de uma reta.

O Coeficiente de correlação linear de Pearson é calculado por

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, S_x S_y}$$

, S_x e S_y os desvios padrão de X e Y, respectivamente.

Fórmula alternativa: $r = \dfrac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{(n-1)\, S_x S_y}$

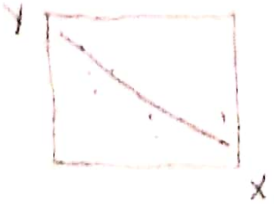
Prop: $-1 \le r \le 1$

$r = 1 \rightarrow$ correlação linear positiva e perfeita

$r = -1 \rightarrow$ " " negativa " "

$r = 0 \rightarrow$ inexistência de correlação linear

Análise de Regressão



Explicar a forma da relação por meio de uma função matemática:

$$Y = a + bX$$

Reta ajustada:

$$\hat{Y} = \underbrace{a}_{\text{intercepto}} + \overbrace{b}^{\text{inclinação ou coef. angular}}X$$

Interpretação de b: Para cada aumento de uma unidade em X, tem-se um aumento, em média, de b unidades em Y.

Mínimos quadrados:

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)\, s_x^2}$$

$$a = \bar{y} - b\bar{x}$$

Exercício: Qtde de força necessária (Y) ← comprimento da mola alongada (X)

a) Interpretar relacionamento entre as variáveis.

b) Interpretar coef. correl. 0,93

c) Interpretar coef. angular 23,8

d) Força nec. para alongar 4,5 cm (estimativa, segundo o modelo)

Aula 08 - Noções de Probabilidade

- Espaço Amostral: cjto de todos os res. possíveis

- Eventos (subconjuntos do espaço amostral). Operações com eventos (união e intersecção). Eventos disjuntos e complementares.

$A \cup B$: ocorrência de pelo menos um dos eventos, A ou B ; $A \cap B = \emptyset \Rightarrow$ A e B disj.
$A \cap B$: ocor ... simultânea $A \cap B = \emptyset$ e $A \cup B = \Omega \Rightarrow$ B e A são complementos.

- Probabilidade: medida de incerteza associada aos resultados de experimento aleatório (Freq. relativas, Suposições teóricas, Experiência de um especialista)

No caso discreto, experimento aleatório tem seu modelo prob. especificado qdo estabelecemos $\Omega = \{\omega_1, \omega_2, \ldots\}$ e $P(\omega)$ t.q. $0 \leq P(\omega_i) \leq 1$ e $\sum_{i=1}^{\infty} P(\omega_i) = 1$ e $P(A) = \sum_{\omega_j \in A} P(\omega_j)$.

- Regra da adição de probabilidades, Probabilidade Condicional e Independência

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

Prob. Condicional

- Dados dois eventos A e B, a prob. condicional de A dado que ocorreu B é denotada por $P(A/B)$ e def. por

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

Da def. acima, segue a regra do produto de probabilidades:

$$P(A \cap B) = P(B) \times P(A/B) \text{ e}$$
$$P(A \cap B) = P(A) \times P(B/A)$$

Independência

Dois eventos A e B são independentes se a informação da ocorrência (ou não) de B não altera a prob. de ocorrência de A, isto é

$$P(A/B) = P(A), \quad P(B) > 0$$

Equivalentemente: $P(A \cap B) = P(A) \cdot P(B)$

Exercício:

3 máquinas: A, B e C

- A faz o dobro de peças que B
- B faz e C produzem o mesmo nº de peças

1% das por A e B defeituosas
3% das por C defeituosas

a) Do depósito, uma peça ret. ao acaso. Qual a prob. de ser defeituosa?

Considere os seguintes eventos

D: a peça é def.

A: veio da máq. A

B:

C:

$P(A) = 50\%$

$P(B) = 25\% = P(C)$

$P(D/A) = P(D/B) = 1\%$ e $P(D/C) = 3\%$

teste do produto

$P(D) = P(D \cap A) + P(D \cap B) + P(D \cap C) =$

$P(A)P(D/A) + P(B)P(D/B) + \dots = 0{,}015 = 1{,}5\%$

b) prob. vir de A dado que é def.? $P(A/D) = \frac{P(D \cap A)}{P(D)} = \frac{P(A)P(D/A)}{P(D)} = \frac{0{,}005}{0{,}015} = \frac{1}{3}$

Um terço das def. vem de A.

Aula 04.

Variável aleatória: uma função X que associa a cada $\omega \in \Omega$ um valor $x \in \mathbb{R}$ (Ex. nº de vezes que ocorre um nº par em 2 lanç. de um dado honesto),

Variáveis aleatórias discretas (cjto de valores possíveis é finito ou infinito enumerável)

" " **contínuas** (cjto de valores possíveis não enumerável)

Função de probabilidade de uma variável aleatória discreta: atribui a cada valor x_i da v.a. X sua prob. de ocorrência e pode ser representada pela tabela

x	x_1	...	x_n	
$P(X=x)$	$P(X=x_1)$		$P(X=x_n)$	

- $0 \leq P(X=x_i) \leq 1$
- $\sum_{i=1}^{n} P(X=x_i) = 1$

- **Valor Esperado:** Dada a v.a. X, assumindo os valores $x_1, \ldots, x_n$, chamamos de valor médio, ou valor esperado, ou esperança matemática da distribuição de X o valor

$$E(X) = x_1 P(X=x_1) + \ldots + x_n P(X=x_n) = \sum_{i=1}^{n} x_i \times P(X=x_i)$$

Notação: $\mu = E(X)$

Variância: É o valor esperado da v.a. $(X - E(X))^2$, i.e.

$$Var(X) = \sum_{i=1}^{n} [x_i - E(X)]^2 \, P(X=x_i)$$

Not: $\sigma^2 = Var(X)$

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

Modelos Probabilísticos Discretos: Bernoulli, Binomial

$X \sim B(p)$ $\quad P(X=k) = p^k (1-p)^{1-k}, \; k=0,1$

$Y \sim Bin(n,p)$ $\quad P(Y=k) = \binom{n}{k} p^k (1-p)^{n-k}, \; k = 0, \ldots, n$

Exercício

Vacina contra o HPV para prev. inf. pelo vírus do papiloma humano
30% dos casos do câncer de colo do útero não são prev. pela essa vacina.

a) 15 pessoas selecionadas ao acaso que entraram em contato com o vírus no máximo 2

$X \sim b(15; 0,3)$ $P(X \leq 2)$

b) entre 2 e 5 (inclusive os extremos)

$P(2 \leq X \leq 5) =$

c) 100 indiv.

menos de 16

$X \sim b(100; 0,3)$ $P(X \leq 15) = P\left(\frac{X - 30}{\sqrt{21}} < \frac{15 - 30}{\sqrt{21}}\right) = P(Z < -3,275) = 0,0005$

Aula 05 - Distribuição Normal

- Des. por Gauss. Importante em aplicações no cálculo de prob.

- Uma v.a. contínua X é caracterizada por sua função densidade de prob. $f(x)$ que tem as seg. propriedades

  i. A área sob a curva da densidade é 1

  ii. $P(a \leq X \leq b)$ = área sob a curva da dens. $f(x)$ e acima do eixo x, entre os pontos a e b;

  iii. $f(x) \geq 0$, para todo x

  iv. $P(X = x_0) = 0$, para x_0 fixo

- $X \sim N(\mu, \sigma)$ se $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$, $-\infty < x < \infty$

- $\mu = E(X)$ e $\sigma^2 = Var(X)$. $\mu - \sigma$ e $\mu + \sigma$ são pontos de inflexão da curva. $x = \mu$ pto de máximo; $f(x) \to 0$ qdo $x \to \pm\infty$; simétrica em torno de μ.
- Normal reduzida ou padrão. $Z \sim N(0,1)$. $X = \mu + \sigma Z$. Tab. p/ Z;

- Aproximação da binomial pela normal. Quando n cresce, a dist. da primeira se aproxima da seg.

$$X \sim b(n,p), \; Y \sim N(np, np(1-p))$$

$P(a \leq X \leq b) \approx P(a < Y < b)$; $P(X \geq a) \approx P(Y \geq a)$; $P(X \leq b) \approx P(Y \leq b)$

- Aprox. boa qdo $np(1-p) \geq 3$. Validade: TLC.

Exercício: "Produção de energia por usinas numa certa região $\sim N(500; 100^2)$.

a) $P(X > 600)$; b. $P(X < 700)$; c. t tal que $P(X < t) = 0{,}25$

Aula 07. Estimação da Média Populacional.

•

Inferência: os val. dos parâmetros são desc. Queremos estimá-los!

Parâmetro: qtde desconhecida de uma caract. da população sobre a qual temos interesse.

1. X: var. de interesse
2. Observ. n elementos ao acaso de forma independente
3. Para cada elemento, obs. o valor de X
4. Temos uma a.a. de tamanho n $X_1, \ldots, X_n$
5. Construo um estimador, que terá uma distribuição amostral
6. A partir dele, estimativas pontuais e intervalares

Estimador: fção dos elementos da amostra, construído com a finalidade de representar, ou estimar, um parâmetro da característica de interesse X na população $f(X_1 \ldots X_n)$. São fções de v.a. e, ∴, tbm são v.a. Sua dist. é denominada Distribuição Amostral do Estimador.

Objetivo da Aula: Estimar a média μ de uma v.a. X, que representa uma característica de interesse de uma população, a partir de uma amostra de valores de X.

1. • **Estimação Pontual:** para μ, baseado numa amostra aleatória de tamanho n

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \sum_{i=1}^{n} X_i / n$$

2. **Estimativa por intervalo ou intervalo de confiança**

$$[\bar{X} - \varepsilon, \bar{X} + \varepsilon], \quad \varepsilon = \text{erro amostral (margem de erro)}$$

obtenho o IC por meio da dist. amostral do estimador pontual.

- $X \sim N(\mu, \sigma^2) \to \bar{X} \sim N(\mu, \sigma^2/n)$
  - σ conhecido, $Z = \bar{X} - \mu / \sqrt{\sigma^2/n} \sim N(0, 1)$

$\gamma = P(\varepsilon) =$ nível de confiança: prob. da média amostral $\bar{X}$ estar a uma distância de, no máximo, ε da média populacional μ (desconhecida), ou seja

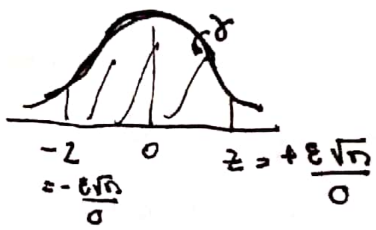
$$\gamma = P(\varepsilon) = P(|\bar{X} - \mu| \leq \varepsilon)$$: coeficiente de confiança do intervalo.

$\varepsilon = z\,\sigma/\sqrt{n}$

$$I.C(\mu;\gamma): \left[\bar{X} - z\frac{\sigma}{\sqrt{n}}\ ;\ \bar{X} + z\frac{\sigma}{\sqrt{n}}\right]$$, σ conhecido *

Dimensionamento da Amostra

$$\varepsilon = z\frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{z}{\varepsilon}\right)^2 \sigma^2$$



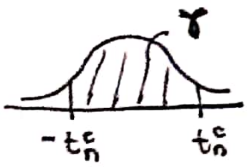
• Na prática, não conhecemos σ^2.

Resultado.

$X \sim N(\mu, \sigma^2)$, μ, σ^2 desc.

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}\ ,\quad S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

$$I.C.(\mu;\gamma): \left[\bar{X} - t^c_{n-1}\frac{S}{\sqrt{n}}\ ;\ \bar{X} + t^c_{n-1}\frac{S}{\sqrt{n}}\right]$$, σ desconhecido



Aula 7. Estimação da Média Populacional

1. Análise Descritiva
"Os dados efetivamente obs. parecem indicar que..."

2. Teoria de Probabilidades
"Se a dist. dos dados seguir uma certa lei, é esperado que..."

3. Inferência
"Ajuste de um modelo abstrato aos dados de uma amostra"
"Estimação e Teste de Hipóteses"

Probabilidade: parâmetros conhecidos

Inferência: os valores desses parâmetros são desconhecidos. Queremos estimá-los!

Parâmetro: qtde desconhecida de uma característica da população e sobre a qual temos interesse. Ex: μ: média da caract. da população

p: proporção de "indivíduos" em uma população com determinada característica

Procedimento

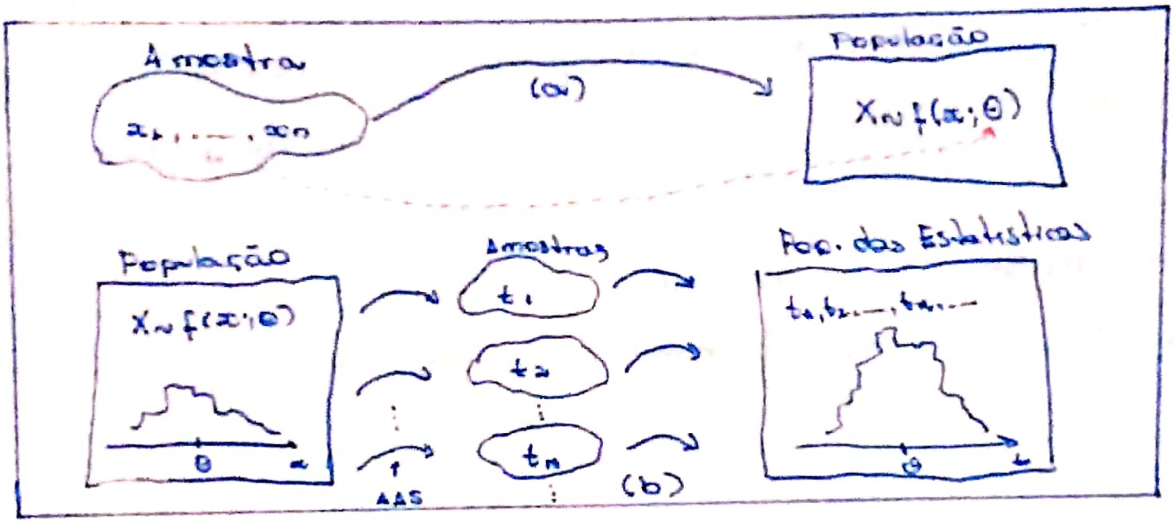
- X: variável de interesse (Renda)
- Vamos observar n elementos extraídos ao acaso da população, de forma independente
- Para cada elemento selecionado, observamos o valor da variável X de interesse.
- Temos amostra aleatória de tamanho n de X: $X_1, \ldots, X_n$

Estimador: função dos elementos da amostra, construída com a finalidade de representar, ou estimar, um parâmetro da característica de interesse X na população. $(\bar{X}, \hat{p}) = f(X_1, \ldots, X_n)$

Estimativa: valor numérico assumido pelo estimador, para a amostra selecionada $(\bar{x})$

Os estimadores $\bar{X}$ e $\hat{p}$ (prop amostral) são intuitivos e tem boas prop.

Estimadores são funções de v.a e, portanto, tbm são variáveis aleatórias. Conseq., tem uma distribuição de prob., denominada distribuição amostral do estimador.



a) Esquema de inf. sobre θ

b) Distribuição amostral da estatística T.

Objetivo da Aula: Estimar $\mu = \mathbb{E}X$,
1. Estimação Pontual
2. " Intervalar

que representa uma carac. de interesse de uma pop., a partir de uma amostra de valores de X.

1. Um estimador pontual para μ, baseado numa a.a de tamanho n, é dado pela média amostral

$$\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \sum_{i=1}^{n} X_i / n$$

$\bar{x}$ é a estimativa pontual para μ.

2. Estimativa por intervalo ou intervalo de confiança

Ideia: construir intervalos de confiança que incorporem à estimativa pontual informações a respeito de sua variabilidade (erro amostral).

Aula 07. Folha 2

Intervalos de confiança são obtidos por meio da distribuição amostral do estimador pontual.

Um estimador intervalar ou intervalo de conf. p/ μ tem a forma

$$[\bar{X} - \varepsilon, \bar{X} + \varepsilon]$$

sendo ε o erro amostral (margem de erro) calculado a partir da distribuição de probabilidade de $\bar{X}$.

. Distribuição de $\bar{X}$

1. Para qqe X com $EX = \mu$ e var $X = \sigma^2$, temos

$$E\bar{X} = \mu \quad e \quad Var\,\bar{X} = \frac{\sigma^2}{n}$$, n o tamanho da amostra

O desvio padrão $\sqrt{\sigma^2/n} = \sigma/\sqrt{n}$ é denominado <u>erro padrão</u> da média amostral.

2. Se $X \sim N(\mu, \sigma^2)$, então $\bar{X} \sim N(\mu, \sigma^2/n)$.

. Se σ é conhecido

$$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

Seja $P(\varepsilon) = \gamma$ a probabilidade da média amostral $\bar{X}$ estar a uma distância de, no máximo, ε, da média pop μ (desconhecida), ou seja

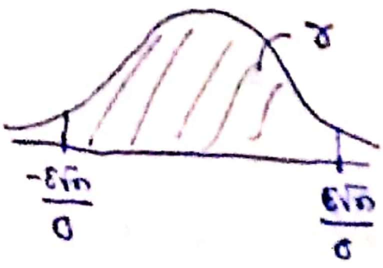
$$\gamma = P(\varepsilon) = P(|\bar{X} - \mu| \leq \varepsilon)$$: coeficiente de confiança do intervalo

Desse modo, temos

$$P(\varepsilon) = P(|\bar{X}-\mu| \leq \varepsilon) = P(\mu - \varepsilon \leq \bar{X} \leq \mu + \varepsilon) =$$

$$= P\left(\frac{-\varepsilon}{\sigma/\sqrt{n}} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq \frac{\varepsilon}{\sigma/\sqrt{n}}\right)$$

$$= P\left(\frac{-\varepsilon}{\sigma/\sqrt{n}} \leq Z \leq \frac{\varepsilon}{\sigma/\sqrt{n}}\right), \quad Z \sim N(0,1)$$



Erro na Estimativa Intervalar

$z = \frac{\varepsilon\sqrt{n}}{\sigma}$, segue que $\varepsilon = z\frac{\sigma}{\sqrt{n}}$

$$I.C.(\mu;\gamma) \quad \left[\bar{X} - z\frac{\sigma}{\sqrt{n}} \; ; \; \bar{X} + z\frac{\sigma}{\sqrt{n}}\right], \quad \sigma \text{ conhecido}$$

Dimensionamento da Amostra

$$\varepsilon = z\frac{\sigma}{\sqrt{n}} \Rightarrow n = \frac{z^2\sigma^2}{\varepsilon^2} = \left(\frac{z}{\varepsilon}\right)^2 \sigma^2$$

Conhecendo-se o desvio padrão σ de X, com erro da estimativa fixado ε e o coef. de conf. do int. γ, sendo z t.q. $\gamma = P(-z \leq Z \leq z)$, $Z \sim N(0,1)$

Na prática, em geral, não conhecemos σ^2.

Resultado 3.

X na pop. $\sim N(\mu, \sigma^2)$, μ e σ^2 desconhecidos

$$T = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1} \quad , \quad S^2 = \frac{\sum^n (X_i - \bar{X})^2}{n-1}$$

t_{n-1} : t-Student com n-1 graus de liberdade.

Tal como para o caso em que σ^2 é conhecido, prova-se facilmente que o intervalo de conf. com coef. de conf. γ

$$I.C(\mu; \gamma) : \left[\bar{X} - t^c_{n-1} \frac{S}{\sqrt{n}} \; ; \; \bar{X} + t^c_{n-1} \frac{S}{\sqrt{n}} \right]$$

t^c_{n-1} o pto crítico da dist. t-student com n-1 g.l. t.q

$P(-t^c_{n-1} \leq t_{n-1} \leq t^c_{n-1}) = \gamma$ e S o desv. padrão

Aula 08 - T. Central do Limite e estimação da proporção populacional p.

4. X v.a. qqer com média μ e var. σ^2.
Para amostras $X_1, \ldots, X_n$ retiradas ao acaso e com repos. de X, Para n grande

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right) \quad : \text{n grande, aproximadamente}$$

$$\left[\bar{X} - z\frac{\sigma}{\sqrt{n}} ; \bar{X} + z\frac{\sigma}{\sqrt{n}}\right], \sigma \text{ conhecido, n grande}$$

$$\left[\bar{X} - z\frac{S}{\sqrt{n}} ; \bar{X} + z\frac{S}{\sqrt{n}}\right], \sigma \text{ desconhecido, n grande}$$

ESTIMAÇÃO DA PROPORÇÃO

Objetivo: Estimar uma proporção p (desconhecida) de elementos em uma população, apresentando certa característica de interesse, a partir da informação fornecida por uma amostra.

1. Obs. n elementos, extraídos ao acaso da pop., de forma independente.
2. Para cada selec. da população, verif. a presença ("sucesso") ou não ("fracasso") da característica de interesse

Aqui, a.a tamanho n de X, $X \sim$ Bernoulli : $X_1, \ldots, X_n$, onde X_i vale "1" se ocorre sucesso, ou "0", se ocorre fracasso pro i-ésimo

- Estimador Pontual para p, tbm denominado proporção amostral,

$$\hat{p} = \frac{X_1 + \ldots + X_n}{n}$$

· Intervalo de Confiança para p

n grande, TCL.

$$[\hat{p} - \varepsilon ; \hat{p} + \varepsilon] \quad , \quad \varepsilon = z\sqrt{\frac{p(1-p)}{n}} \quad , z \text{ t.q}$$

$P(-z < Z < z) = \gamma$.

Na prática

$$\left[\hat{p} - z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} ; \hat{p} + z\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

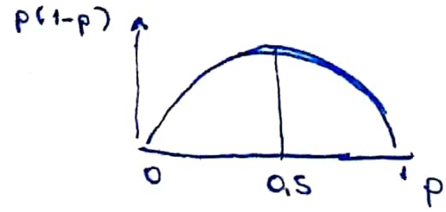
· Dimensionamento da Amostra

$$\varepsilon = z\sqrt{\frac{p(1-p)}{n}}$$

. Dados ε (margem de erro) e γ,

$$n = \left(\frac{z}{\varepsilon}\right)^2 p(1-p)$$

Se $p(1-p)$ é desconhecido → pego o máximo → $0{,}5 \times 0{,}5 = 0{,}25$



Aula 09 - Teste de Hipótese p/ proporção populacional

Métodos Estatísticos

- Descritivos
- Inferência Est.
  - Estimação
  - Teste de Hipóteses

Estimação ⇒ Objetivo: estimar o valor desconhecido.

Entretanto, se o obj for saber se a estimativa pontual obs. na amostra dá ou não suporte a uma conjectura sobre o valor do parâmetro, trata-se Testar Hipóteses. → Conjectura sobre um parâmetro populacional

Resumo.

0. Def. o parâmetro p de interesse no problema
1. Estabelecer as hipóteses estatísticas

$H_0: p = p_0$ contra umas das alternativas

$H_1: p \neq p_0$ (bilateral) $\quad H_1: p > p_0$ (unilateral) $\quad H_1: p < p_0$ (unilateral)

2. Escolher um nível de significância α.
3. Determinar a região crítica da forma

$$\{\hat{p} \leq a_1, \hat{p} \geq a_2\}, \ \{\hat{p} \geq a\}, \ \{\hat{p} \leq a\}$$

respect. às hipóteses alternativas

4. Sel. uma amostra casual simples e determinar $\hat{p}$

5. Decidir, usando a evidência $\hat{p}$, ao nível de significância α, e concluir

$\hat{p} \in RC \Rightarrow$ rejeitamos H_0

$\hat{p} \notin RC \Rightarrow$ não rejeitamos H_0

Aula 09. Teste de Hipótese p/ proporção populacional

M. Estatísticos
- Descritiva
- Inf.
  - Estimação
  - Teste de Hipótese

Estimação: Objetivo: Estimar o valor desconhecido de um parâmetro.

Entretanto, se o Objetivo for saber se a estimativa pontual dá ou não suporte a uma conjectura, trata-se de Testar Hipóteses.

Hipótese: Conjectura sobre um parâmetro populacional.

No caso de teste de Hip. sobre p:

Hipótese Nula: afirmação sobre p, ligada a um valor de referência, ou a uma especificação padrão ou histórica.

Hipótese Alternativa: afirmação sobre p que suspeitamos que seja verdadeira.

Testar uma hip. estatística é estabelecer uma regra que nos permita, com base na informação de uma amostra, decidir pela rejeição ou não de Ho.

Regra de decisão (teste)

$$\hat{p} \in RG \Rightarrow \text{rejeitamos } H_0$$
$$\hat{p} \notin RG \Rightarrow \text{não rejeitamos } H_0$$

Erros:

Tipo I: Rejeitar Ho quando Ho é verdadeira

Tipo II: Não rejeitar Ho quando Ho é falsa.

Prob. dos Erros

$P(\text{Erro I}) = P(\text{rej. } H_0 / H_0 \text{ é verd.}) = \alpha$: nível de sig. do teste

$P(\text{Erro II}) = P(\text{não rej. } H_0 / H_0 \text{ é falsa}) = \beta$: $1-\beta$ poder do teste

Resumo

0. Def. o parâmetro de interesse p no prob.

1. Establecer as hipóteses estatísticas

$H_0: p = p_0$

$H_1: p \neq p_0$ $H_1: p > p_0$ $H_1: p < p_0$
bilateral uni uni

2. Escolher um nível de significância α.

3. Determinar a Região crítica da forma

$\{\hat{p} \leq a_1, \hat{p} \geq a_2\}$, $\{\hat{p} \geq a\}$, $\{\hat{p} \leq a\}$

4. Selecionar uma amostra casual simples e determinar $\hat{p}$.

5. Decidir, usando a ev. $\hat{p}$, ao nível de sig. α, e concluir

- $\hat{p} \in RC \Rightarrow$ rej. H_0
- $\hat{p} \notin RC \Rightarrow$ não rej. H_0

Aula 10. Nível descritivo e Teste de Hipóteses para a média populacional μ

Nível Descritivo P. (ou valor P)

No caso da prop.

$$P = P(\hat{p} \leq p_{obs} \mid p = p_0) \text{ se } H_1: p < p_0$$

Essa prob. P mede a força de evidência contida nos dados, contra a hip. nula Ho.

Quão pequeno?

P pequeno ⇒ rejeitamos Ho → $P \leq \alpha$

P não pequeno ⇒ não rej. Ho → $P > \alpha$

dizemos que a amostra forneceu ev. suf. p/ rej. Ho.

Res. (via nível descritivo para p)

0. Def. o parâmetro p a ser testado
1. Est. as hipóteses

   $H_0: p = p_0$

   $H_1: p \neq p_0$, $H_1: p > p_0$, $H_1: p < p_0$

2. Escolher nível de sig. α
3. Selec. a.a. casual simples e obter $\hat{p}$.
4. Det. o nív. descritivo ou valor P

   $H_1: p > p_0$ $\quad P = P(\hat{p} \geq \hat{p}_{obs} \mid p = p_0)$

   $\quad\; p < p_0$ $\quad P = P(\hat{p} \leq \hat{p}_{obs} \mid p = p_0)$

   $\quad\; p \neq p_0$ $\quad P = 2 \times P(\hat{p} \leq \hat{p}_{obs} \mid p = p_0)$ se $\hat{p}_{obs} < p_0$

   $\qquad\qquad\qquad 2 \times P(\hat{p} \geq \hat{p}_{obs} \mid p = p_0)$ se $\hat{p}_{obs} > p_0$

5. Decidir com base em P com nível de sig. α e concluir

   $P \leq \alpha \Rightarrow$ rej. Ho

   $P > \alpha \Rightarrow$ não rej. Ho

Teste Hipóteses para a média populacional

1. Formular as hip. nula H_0 e a alternativa H_1
   - H. Nula: afirmação ou congectura sobre μ contra qual estamos estaremos buscando evidência nos dados
   - " Alternativa: " " " " " que suspeitamos (ou esperamos) ser verdadeira.

2. Esc. a Estatística de Teste a ser utilizada $\Rightarrow \bar{X}$

3. Fixar ní. sig. α do teste

4. Coletar os dados e calcular as medidas necessárias: $\bar{x}_{obs}$ e, se necessário, s.

5. Determinar RC ou P nível descritivo P

6. Tomar a dec. e concluir

   $P \leq \alpha$: reconhecemos na amostra evidência suf. p/ rej. H_0

Resumo

0. Descrever o parâmetro de int. μ
1. Est. as hipóteses

   $H_0: \mu = \mu_0$

   $H_1: \mu \neq \mu_0$; $H_1\ \mu > \mu_0$; $H_1: \mu < \mu_0$
2. Est. de Teste: $\bar{X}$
3. Esc. nív. sig. α
4. Sel. amostra cas. e computar $\bar{x}_{obs}$ e σ / s
5. Determinar

   $H_1: \mu > \mu_0 \Rightarrow P = P(\bar{X} > \bar{x}_{obs} \mid \mu = \mu_0)$

   $H_1: \mu < \mu_0 \Rightarrow P = P(X < \bar{x}_{obs} \mid \mu = \mu_0)$

   $H_1: \mu \neq \mu_0$ $\quad 2 \times P(\bar{X} > \bar{x}_{obs} \mid \mu = \mu_0)$ se $\bar{x}_{obs} > \mu_0$

   $2 \times P(\bar{X} < \bar{x}_{obs} \mid \mu = \mu_0)$ $\bar{x}_{obs} < \mu_0$
6. Decidir. $P \leq \alpha \Rightarrow$ rej. H_0

Lembrar

- σ conhecido, X normal ou n grande

  $$Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$
- σ desconh.

  $$T = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \sim t_{n-1},\ X \sim N(\mu, \sigma)$$

  $$T = \frac{\bar{X} - \mu}{\sqrt{s^2/n}} \sim t_{n-1},\ n \text{ grande}$$

Aula 11. Teste Qui-Quadrado. Aderência e Independência

1. Teste de Aderência: testar a adequabilidade de um modelo probabilístico a um cjto de dados observados.

Metodologia

Categorias	Freq. Obs	Fre. Esp
1	O_1	E_1
2	O_2	E_2
3	O_3	E_3
⋮	⋮	⋮
k	O_k	E_k
Total	n	n

, O_i: total de ind. obs. na categoria i.

p_i: prob. associada à cat. i.

Obs. Teste aderência:

$H_0: p_1 = p_{01}, \ldots, p_k = p_{0k}$

H_1: existe pelo menos uma diferença

com p_{0i} a prob. especificada para a cateria i, $i=1 \ldots k$ fixada por meio do modelo prob. de interesse.

E_i: # indivíduos esperados na categoria i, quando H_0 é verdadeira

$$E_i = n\, p_{0i}, \quad i = 1, \ldots, k$$

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \rightarrow \text{Est. do teste de Aderência} \sim \chi^2_q$$

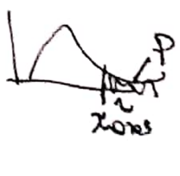
sob H_0

$q = k-1$.

→ Se H_0 é vdd, $\chi^2 \sim \chi^2_{k-1}$ aproximadamente.

$\boxed{\text{Importante}}$ Válido para n grande e para $E_i \geq 5$, $i=1,\ldots,k$

Regra decisão: $P = P(\chi^2_q \geq \chi^2_{obs})$



$P \leq \alpha \Rightarrow$ rejeitamos H_0

2. Teste de Independência

Obj: Verificar se existe Ind. entre duas variáveis med. nas mesmas unidades experimentais

Metodologia

$A \setminus B$	B_1	B_2	...	B_s	Total
A_1	O_{11}	O_{12}		O_{1s}	$O_{1.}$
A_2	O_{21}	O_{22}		O_{2s}	$O_{2.}$
$\vdots$					$\vdots$
A_r	O_{r1}	O_{r2}		O_{rs}	$O_{r.}$
Total	$O_{.1}$	$O_{.2}$		$O_{.s}$	n

H_0: A e B são variáveis indep.

H_1: As variáveis A e B não são indep.

$$E_{ij} = \frac{O_{i.} \times O_{.j}}{n}$$

$$\chi^2 = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$: dist. entr. val. obs e os val. esp. sob a sup. de ind.

Supondo H_0 verd.

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})}{E_{ij}} \sim \chi^2_q \,, \quad q = (r-1)(s-1)$$

Regra Decisão

$p = P(\chi^2 > \chi^2_{obs})$

χ^2_{obs}

$p \leq \alpha$ rej. a hip. H_0 de independência