Estatística

Conceitos

Estatística Descritiva: apresenta métricas sobre os dados.
Estatística Inferencial: realiza previsões a partir dos dados existentes.

C/qual conjunto será trabalhado?	
Conj. Completo	Conj. Incompleto
População	Amostra
Censo	Amostragem
Parâmetro	Estimativa

Tipos de variáveis			
Quantitativas		Qualitativa	
Discretas	Contínuas	Nominais	Ordinais
valores inteiros	valores inteiros e fracionários	s/ordem entre valores	c/ordem entre valores
Qtd. objetos	Altura, peso.	Gênero, cor, país	Grau de escolaridade

Métricas		
Elemento	População	Amostra
	Parâmetro populacional θ	Estimadores do par.populacional $\hat{\theta}$
Média	μ	$\bar{x}$
Variância	σ^2	s^2
Desvio Padrão	σ	s
Tamanho	N	n
Proporção	P	$\hat{p}$

Propriedades dos estimadores:
- não viesado: na média, acerta o parâmetro populacional
- eficiente: não viesado e o mais preciso possível (menor variância possível)
- consistente: à medida que a amostra cresce, será convergido para valor do parâmetro
- Máxima verossimilhança: estimador mantém a mesma distribuição de probabilidade da população

Estatística Descritiva

1. Formas de apresentação dos dados

☐ Dados brutos:listagem de todos os dados

{0,5, 10, 15, 15, 15, 20, 20, 30}

☐ Dados ponderados:tabela de frequência sem intervalo

* n de classes: amplitude/n
* n de intervalos: $\sqrt{n}$

Valor Observado (X_i)	Frequência Absoluta (f_i)	Frequência Relativa (fr_i)	Frequência Acumulada (F_i)	Frequência Acumulada Relativa (Fr_i)
0 ⊢ 10	2	2/9 ≅ 22%	2	2/9 ≅ 22%
10 ⊢ 20	4	4/9 ≅ 44%	6	6/9 ≅ 67%
20 ⊢⊣ 30	3	3/9 ≅ 33%	9	9/9 = 100%
Soma (Σ_i)	**9**	**9/9 = 100%**	-	-

☐ Dados agrupados: tabela de frequência com intervalo

Valor Observado (X_i)	Frequência Absoluta (f_i)	Frequência Relativa (fr_i)	Frequência Acumulada (F_i)	Frequência Acumulada Relativa (Fr_i)
0 ⊢ 10	2	2/9 ≅ 22%	2	2/9 ≅ 22%
10 ⊢ 20	4	4/9 ≅ 44%	6	6/9 ≅ 67%
20 ⊢⊣ 30	3	3/9 ≅ 33%	9	9/9 = 100%
Soma (Σ_i)	**9**	**9/9 = 100%**	-	-

Como calcular o número de classes?	
Op. 1: regra de Sturges	Op. 2: r. quadrada
$nc = 1 + 3,3 Log\, n$	$nc = \sqrt{n}$

Como calcular a amplitude do intervalo de classe(h)?

$$h = \frac{(Xmáx - Xmin)}{nc}$$

Formas de apresentação dos dados	
Uma variável	Duas ou mais variáveis
- Gráfico de frequência - Gráfico de barras - Histograma - Diagrama de pontos - Polígono de frequência - Curva de frequência - Diagrama de ramos e folhas	- Tabelas - Gráfico de colunas - Gráfico de barras - Gráfico de setores (pizza) - Gráfico de dispersão - Gráfico de linhas - Diagrama de ramos e folhas

Estatística Descritiva

*Quais dados devem ser ordenados? mediana, separatrizes

2. Medidas descritivas

A. Medidas de Dispersão:
A.1 Absoluta: amplitude total, amplitude interquartílica, desvio quartil, desvio médio , desvio padrão.
A.2 Relativa: coeficiente de variação, coeficiente de variação quartil.

B. Medidas de Posição:
B.1 Separatrizes: mediana(dados ordenados), decis, quartis, percentis.
B.2 Tendência central: média, mediana,, moda

C. Medidas de Forma:
C.1 Assimetria
C.2 Curtose

Transformação uniforme de dados:

Medidas de posição: acompanham a transformação do conjunto de dados (+,-,*,/)
Média, Mediana, Moda, Separatrizes (quartis e decis)
Medidas de dispersão: + e - não afetam
- variância e desvio padrão: * e / sofre modificações (pendente)
Coeficiente de variação: sofre alteração + ou -, * e / não afetam
coeficiente de variação

A. Medidas de dispersão

A.1 Absolutas: amplitude total, amplitude interquartílica, desvio quartil, desvio padrão, desvio médio.

Coeficiente de variação interquartílica	*Amplitude/ Intervalo interquartil (Aq):* *Amplitude semi-quartílica (As):*	*Desvio Quartil (Dq)*
$\frac{Q3-Q1}{Q3+Q1} = \frac{Aq}{Q3+Q1}$	$A_q = Q_1 - Q_3$ A_s = mediana	$D_q = (Q_3 - Q_1)/2$
Amplitude total (At):	*Desvio:*	Desvio médio
$A_t = X_{máx} - X_{min}$	Desvio = $x_i - \mu$	$D_m = \sum \lvert x_i - \mu \rvert / n$
Variância:	Desvio padrão: amostra	Desvio padrão população
$\sigma^2 = \sum (x_i - \bar{x})/n$ ou $s^2 = \sum (x_i - \bar{x})/n - 1$	$D_{p.} = \sqrt[2]{\sum (x_i - \bar{x})/n - 1}$	$D_{p.} = \sqrt[2]{\sum (x_i - \bar{x})/n}$

Pq o desvio médio deve ser apresentado em módulo?

xI	*Xi - μ*	*Desvio*	*Desvio*
2	2 - 6,4 = -4,4	-4,4	Soma -6,2
5	5 - 6,4 = -1,4	-11,4	
6	6-6,4 = -0,4	-0,4	
9	9-6,4 = 2,6	2,6	Soma +6,2
10	10-6,4 = 3,6	3,6	
Soma $(Xi - \mu)$	0	0	

Ao somarmos os valores de desvio o resultado será 0. Por conta disso, calcularemos o valor do desvio com **módulo** para dar continuidade ao cálculo.

xI	*Desvio*	$\lvert Xi - \mu \rvert$
2	-4,4	4,4
5	-11,4	1,4
6	-0,4	0,4
9	2,6	2,6
10	3,6	13,6
Somatório	0	12,4

D_m=12,4 / 5 = 2,48

B.2 Relativas: coeficiente de variação, coeficiente de variação quartil.

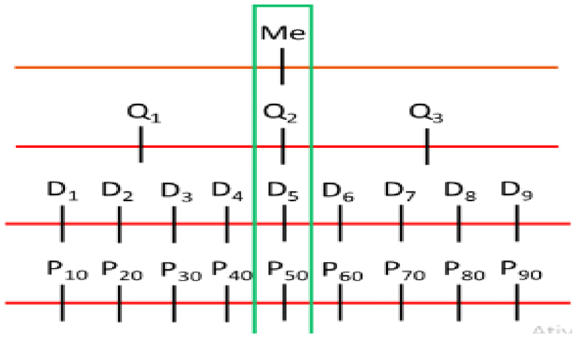
Coeficiente de variação: analisa a dispersão mas não depende de

unidade de medida
$Cv = \sigma/\mu$

Coeficiente de variação interquartil (pendente)

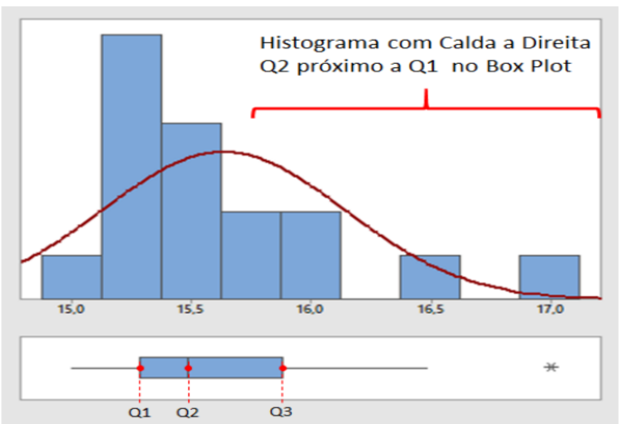
B. Medidas de posição
B.1 Separatrizes

> Quartis (Q)	> Decis (D)	> Percentis (P)
Q1: 0.25(n+1_ = (n+1)/4 Q2: 0.50(n+1_ = (n+1)/2 Q3:0.75(n+1_ = 3(n+1)/4	D1: 0.10 (n+1) D2: 0.20 (n+2) D3: 0.30 (n+3)	P5: 0.05 (n+1) P2: 0.20 (n+1) P50: 0.50(n+1)



Comparativo entre as separatrizes
Me = Q2 = D5 = P50

Figura 2. Relação de Histograma com Box Plot



a. Média

Considere x = {1,2,3}

Aritmética

$\overline{x}$ = 1+3+9 / 3 =13/3 = 4,33

Geométrica:

$\overline{G} = \sqrt[3]{1.3.\ 9} = \sqrt[2]{27} = 3$

Harmônica:

$\overline{H} = \frac{3}{1/1 + 1/3 + 1/9} = \frac{3}{13/9} = \frac{3.9}{13} = \frac{27}{13} = 2.08$

Importante $\overline{X} \geq \overline{G} \geq \overline{H}$

Média aritmética para dados brutos

Cálculo padrão

Média aritmética para dados ponderados

Cálculo	Nota	Peso
$\overline{X}_{p.} = \frac{7.3+6.3+8.2+9.1+7.1}{3+3+2+1+1}$ $\overline{X}_{p.}$ = 71/10=7,1	7	3
	6	3
	8	2
	9	1
	7	1

Média aritmética para dados agrupados

Passo 1: calcular o ponto médio para cada classe	Passo 2: repetir cálculo aplicado para dados ponderados
$Pm_1 = \frac{10+0}{2} = 5$ $Pm_2 = \frac{20+10}{2} = 15$ $Pm_3 = \frac{30+20}{2} = 25$	$\overline{X} = \frac{5.2 + 15.4 + 25.3}{9}$ $\overline{X}$ = 16,11 kg/semana

V observado	Freq Absoluta	Freq Relativa
0-10	2	2/9
10-20	4	4/9
20-30	3	3/9
Somatório	9	9/9

b. Mediana: valor central

Dados devem estar ordenados

Mediana para dados brutos

X = {15, 20, 10, 30, 20, 15, 0, 5, 15} n = 9

X = {0, 5, 10, 15, 15, 15, 20, 20, 30} (Rol crescente)

Me = 15

Mediana para dados ponderados

Identificar a classe central através da fórmula n/2.. O valor da moda é o valor observado para a classe.

V. Observado	F. Acumulada	F. Relativa
0	1	11%
5	2	22%
10	3	33%
15 (Mediana)	6	67%
20	8	89%
30	9	100%

Medida para dados agrupados

Passo 1: Identificar a classe central através da fórmula n/2.. O valor da moda é o valor observado para a classe.

V. Observado	F. Acumulada	F. Relativa
0 -10	2	22%
10 - 20 (Classe modal)	6	67%
20 - 30	9	100%

Passo 2: calcular o valor da interpolação linear para calcular a mediana.

$$\frac{20-10}{6-2} = \frac{Me-10}{4,5-2}$$

$$\frac{10}{4} = \frac{Me-10}{2,5}$$

$$2,5 = \frac{Me-10}{2,5}$$

$$2,5 \,.\, 2,5 = Me - 10$$

$$6,25 + 10 = Me$$

$$Me = 16,25$$

c. Moda: valor que mais se repete

Valor que mais se repete.

>Unimodal:

x={2,3,**4,4,4**,5,8}

>Bimodal:

x={2,3,**4,4,4**,6,**7,7,7**}

>Amodal:

x={2,4,7,8,9,10,15} ( )

Moda para dados brutos

Passo 1: observar o valor que mais se repete no conjunto de dados

X = {0,5,10,15,15,15,20,20,30}

Mo = 15 kg/ semana

Moda para dados ponderados

Observar o valor que mais se repete no conjunto de dados através da frequência absoluta (fi). A moda é o valor observado relativo à frequência

Moda para dados agrupados

Passo 1: observar o valor que mais se repete no conjunto de dados através da frequência absoluta (fi).

Passo 2: calcula-se o valor pontual da moda através. Há quatro métodos possíveis:

$$Mo = \frac{20+30}{2} = 25\ kg/semana$$

Moda de Czuber

$$Mo = Li + h\frac{fmodas - fant}{2fModas - (fAnt + fPost)}$$

Moda de Pearson

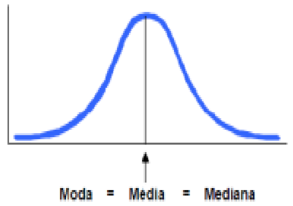
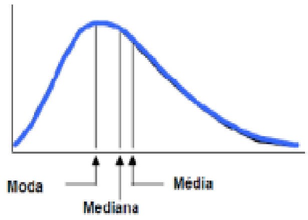
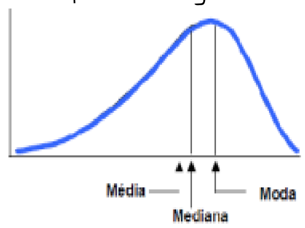
$$Mo = 3Me - 2\overline{x}$$

Moda de King

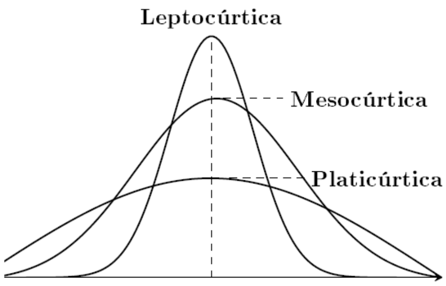
$$Mo = Li + h\frac{fpost}{(fant + fpost)}$$

C. Medidas de forma

C.1 Assimetria (As)

Distribuição simétrica	Distribuição Assimétrica a/ direita (positiva)	Distribuição Assimétrica a/ esquerda (negativa)	Assimetria
			$\frac{Q1 + Q3 - 2Me}{Q3 - Q1}$ As > 0 (ass. positiva) As = 0 (simétrica) As < 0 (ass. negativa)
Q2 -Q1 = Q3 - Q2	Q2 - Q1 < Q3 - Q2	Q3 - Q2> Q2 - Q1	

C.2 Curtose

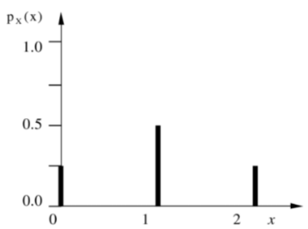
Mesocúrtica	Cálculo de Curtose (C)
	C = Q3 - Q1 / 2.(P90-P10) ou C = Q3 - Q1 / 2(D9 - D1)

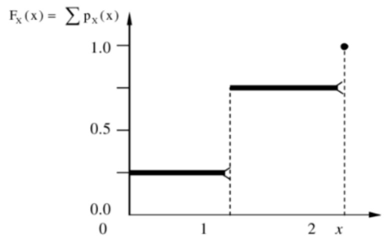
Distribuição de probabilidade discreta (adicionargeométrica)

Distribuição	Caso geral	Bernoulli	Binomial (*)	Poisson	Hipergeométrica (*)
Como identificar	Situações que podem ser quantificadas de forma discreta, isto é, valores inteiros.	1 tentativa 2 resultados (0 / 1)	"n" Bernoulli -independentes	"n" Bernoulli -independentes -intervalo contínuo (tempo/espaço)	"n" Bernoulli -dependentes
Valor Esperado (média)	$\Xi(x) = \sum_{i=1}^{n} x_i . P(x_i)$	$\Xi(x)$ = p	$\Xi(x)$ = n.p	$\Xi(x) = \lambda$	$\Xi(x)$ = n.p $\Xi(x)$ = sp
Variância	$\Xi(x) = \Xi(x_i - E(x)^2) .P(x_i)$ $\Xi(x) = \Xi(x^2) - [E(x)]^2$	$\Xi(x)$ = p.q	$\Xi(x)$ = n.p.q	$Var(x) = \Xi(x) = \lambda$	$Var(x)$ = np.q$(\frac{N-n}{N-1})$ + fator correção
Desvio padrão	$\sqrt[2]{variância}$	$\sqrt[2]{variância}$	$\sqrt[2]{variância}$	$\sqrt[2]{variância}$	$\sqrt[2]{variância}$
Coef. de Variação	$\frac{Desvio\ padrão}{\Xi(x)}$. 100	$\frac{Desvio\ padrão}{\Xi(x)}$. 100	$\frac{Desvio\ padrão}{\Xi(x)}$. 100	$\frac{Desvio\ padrão}{\Xi(x)}$. 100	$\frac{Desvio\ padrão}{\Xi(x)}$. 100
Função de Probabilidade	$\sum_{i=1}^{n} = P(x = x_i)$	$p^s * q^{1-s}$	$C_{n,s} . p^s . (q)^{n-k}$	$e^{-\lambda} . \lambda^k$ -------------- k! e = 2,72 λ = frequência média fornecida k = número de ocorrências em um intervalo variável (valor buscado)	$\frac{C^r_x . C^{N-r}_{n-x}}{C^N_n}$

Para todos os casos

Função Massa de probabilidade	Função Distribuição Acumulada de Probabilidade

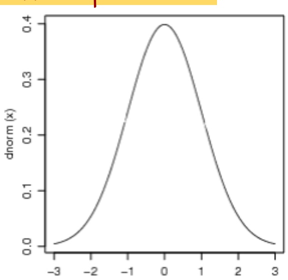
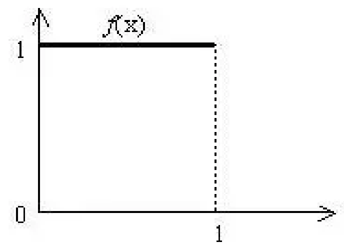
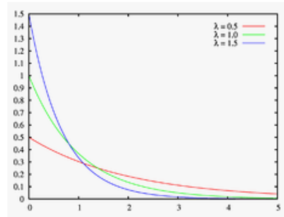
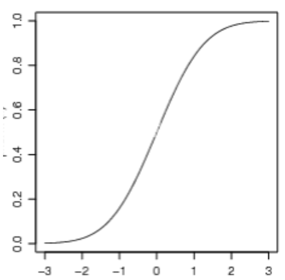
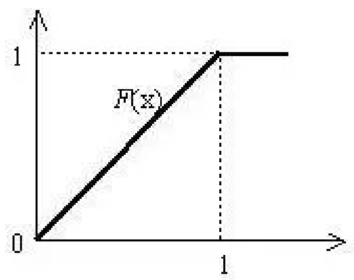
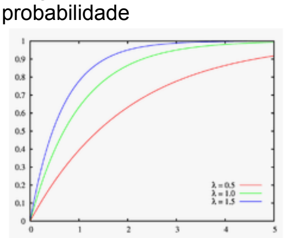




Distribuição de probabilidade contínua

Distribuição	Caso geral	Uniforme	Normal	Exponencial
Como identificar	Situações que podem ser quantificadas de forma contínua, isto é, valores decimais.	Probabilidade igualmente distribuída no intervalo		- Similar: a Poison, diferen porque queremos saber o 'tempo/ unidade contínua' em vez de qtd. de ocorrências - Eventos independentes - número de ocorrências por intervalo de tempo deve ser cte λ *tempo médio entre ocorrências*
Valor Esperado (média)		$\frac{Xmax - Xmin}{2}$	$\Xi(x) = \mu$	$E(x) = Dp(x) = 1/\lambda$ - E(x) = d.p(x)
Variância		$\frac{(Xmax - Xmin)^2}{12}$	$\Xi(x) = \sigma^2$	$E(x) = 1/\lambda^2$
Desvio padrão	$\sqrt[2]{variância}$	$\sqrt[2]{variância}$	$\sqrt[2]{variância}$	$\sqrt[2]{variância} = 1/\lambda$
COeficiente de variação	$\frac{Desvio\ padrão}{\Xi(x)} . 100$	$\frac{Desvio\ padrão}{\Xi(x)} . 100$	$\frac{Desvio\ padrão}{\Xi(x)} . 100$	$\frac{Desvio\ padrão}{\Xi(x)} . 100 = \frac{1/\lambda}{1/\lambda} = 1$
Função de Probabilidade (densidade)	$f(x) = P(x)/ x_{max} - x_{min}$ Probabilidade no ponto = 0 Caso haja necessidade de calcular a probabilidade de intervalo utiliza-se integral e derivada. Normalmente, utilizamos valores tabelados.	$\frac{1}{Xmax = Xmin}$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$f(x) = \lambda e^{\lambda x}$ Função acumulada de probabilidade: $f(x) = 1 - e^{-\lambda x}$

PARA DISTRIBUIÇÃO UNIFORME E NORMAL

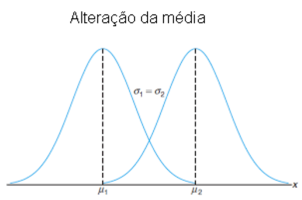
	UNIFORME	EXPONENCIAL
Função Densidade de Probabilidade **Atenção: em variáveis contínuas não há 'Função Massa de probabilidade** 	Função Densidade de Probabilidade 	Função Densidade de Probabilidade 
Função de distribuição acumulada de probabilidade 	Função densidade acumulada de probabilidade 	Função densidade acumulada de probabilidade 

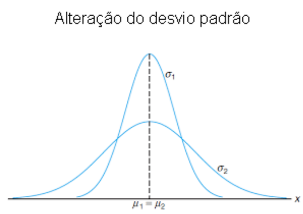
Atenção: além das distribuições apresentadas acima, há outras distribuições como T de Student , Qui Quadrado, entre outros.

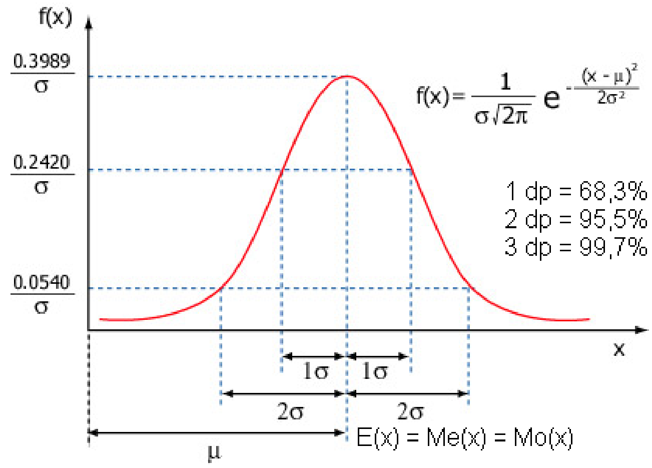
Sobre a distribuição normal

Além de ser utilizada para explicar a probabilidade de variáveis contínuas, também explica intervalos de segurança, teste de hipótese e tamanho amostral.

- Forma de sino
- Simétrica ($\overline{X}$ = Me = Mo)
- P(x<μ) = P(x>μ) = 50%
- É sempre unimodal
- Mesocúrtica
- Quartis equidistantes (Q2-Q1 = Q3-Q2)
- É especificada pela média e desvio padrão (raiz quadrada da variância)







Como aplicar a transformação normal nos dados:

1 - Transformar os valores para distribuição de probabilidade normal padrão (Z)

Z: indica o número de desvios padrões a partir da média

$Z = \frac{X-\mu}{\sigma}$	Quando aplicamos o teste Z $\mu = 0 \; e \; \sigma = 1$

2 - Identificar a área de Z sobre a distribuição normal.

3 - Encontrar valor tabelado que representa a probabilidade da área determinada..

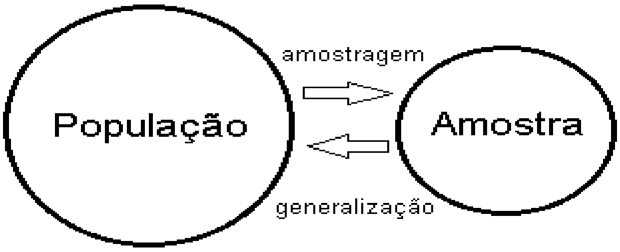
O valor final de Z indica a quantidade de desvios padrão em relação à média.

Testes para verificar a normalidade dos dados:

Numéricos:
-Shapiro-Wilk (limite de 5.000 amostras)
-Kolmogorov-Smirnov (Teste de lilliefors - independe do tamanho de amostras)
-Anderson-Darling
(14:31 - https://www.udemy.com/course/estatistica-para-analise-de-dados-com-python/learn/lecture/25824976?start=90#questions)

Gráficos:
-Histograma
-QQplot

Estatística Inferencial



proporção amostral = probabilidade amostral

Técnicas de amostragem

- Não probabilística (não aleatória)
- Probabilística (aleatória)

a. Simples

-Todas as unidades amostrais devem ter a mesma probabilidade de serem sorteadas

-Seleção das amostras deve permitir ser realizada com ou sem reposição

b. Estratificada

-proporcional

-não proporcional

* Apenas deve-se usar o processo de amostragem estratificada quando houver diferença significativa entre as médias dos estratos, caso contrário, utiliza-se amostragem simples. Nesse caso a amostragem estratificada aumentará a precisão em relação a amostragem simples.

c. Conglomerados: selecionar 2 de 10 turmas de uma escola

d. Sistemática: seleciona elementos com base em frequência pré-definida

$$Fração\ amostral\ =\ \frac{tamanho\ da\ amostra}{tamanho\ da\ população}$$

Estatística inferencial

OBS: Necessário saber qual o tipo de distribuição que os dados seguem

a - Estudar distribuição amostral

b - Calcular estimadores

c - Calcular Intervalo de confiança

d - Calcular tamanho da amostra

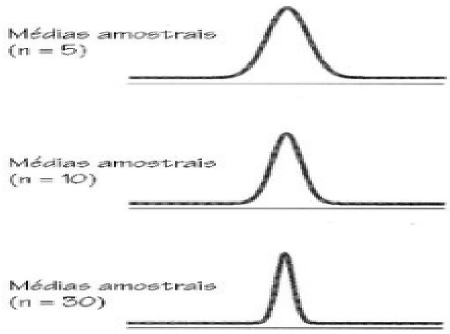
e - Calcular erro amostral

f - Teste de hipótese

Importante: como trabalhamos com amostras, os valores de $\bar{x}$, $\bar{s}^2$, $\hat{p}$, $\bar{s}$ são variáveis aleatórias

Importante: as variáveis aleatórias são quantitativas, não aceitam dados descritivos.

Teorema do Limite central
A amostra de dados tende a distribuição normal.m mesmo que a população apresente outra distribuição



∴ Não é necessário conhecer a distribuição da população para realizar inferências sobre a amostra

	Parâmetro população		Estimador amostra
1. Média	méd. população	=	méd. amostra
2. Variância da média (rever e comparar com variância da população)	var. população	≠	var. amostra
Variância da média das amostras	$\sigma^2 = \sum (x_i - \bar{x})/n$ (var. população)	≠	var. amostra = σ^2/n (variância da média da amostra)
	$\sigma^2 = \sum (x_i - \bar{x})/n$	≠	$s_{\bar{x}}^2 = \sigma^2/n$ (variância da média da amostra)
3. Desvio padrão da média	$\sigma = \sqrt[2]{\sum (x_i - \bar{x})/n}$	≠	$s = \sigma/\sqrt{n}$ (desvio padrão da média das amostras) "erro média amostral"
4. Probabilidade e proporção "estimativa pontual"	p = interesse/ total p = probabilidade	=	p = interesse/ total p = proporção
4.1 Esperança	esperança população	=	esperança amostra
Proporção			
4.2 Variância	var = (p.q)	≠	var = (p.q)/n "variância da proporção amostral"
4.3 Desvio padrão	$dp = \sqrt{p.q}$	≠	$dp = \sqrt{\frac{p.q}{n}} = \frac{\sqrt{p.q}}{\sqrt{n}}$ "erro padrão da proporção amostral" "erro da "estimativa de probabilidade!"
Para o valor máximo da proporção amostral adota-se p = 50%			

TEOREMA DO LIMITE CENTRAL X LEI DOS GRANDES NÚMEROS

Teorema do limite central: quanto maior a amostra, mais os dados se

aproximam da distribuição normal
>> Converge em distribuição

Lei dos grandes números: quanto maior a amostra, maior a probabilidade de se acertar o parâmetro populacional
>> Forte: converge quase certamente para a média
>> Fraca: converge em probabilidade para a média
∴ A lei dos grandes números não tem relação com a distribuição normal

Converter média amostral para forma normalizada	
Dist. Normal	**Dist. Tde Student**
$Z = (X - \mu_{\bar{x}}) / \sigma_{\bar{x}}$ $Z = (X - \mu_{\bar{x}}) / \sigma/\sqrt{n}$	$T = (X - \mu_{\bar{x}}) / s_{\bar{x}}$ $T = (X - \mu_{\bar{x}}) / s/\sqrt{n}$ obs: s = n - 1
Quando usar a normal - Desvio padrão populacional é conhecido - desvio padrão populacional é desconhecido, porém amostra>30	Quando usar t de Student: - desvio padrão desconhecido - amostras de n<30) ! Difere da normal porque possui distribuição de probabilidade para cada grau de liberdade (GL)
Quando usar T de Student? - dp populacional desconhecido - amostra de tamanho pequeno IMPORTANTE: tanto a distribuição normal como a distribuição t de Student tem média = 0. Na dist. normal o desvio padrão é igual a 1.	

Intervalo de confiança

P/ a média amostral (z | t)		P/ a proporção amostral (z)	
$\overline{X} \pm Z\frac{\sigma}{\sqrt{n}}$		$\hat{p} \pm Z\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}$	
P/média amostral	Intervalo de confiança p/ média	Erro de estimativa (margem de erro) p/média (*)	Erro padrão p/ média (*)
- P/ qlqr tamanho de amostra - D.p. pop. conhecido	$\overline{X} \pm Z\frac{\sigma}{\sqrt{n}}$	$Z\frac{\sigma}{\sqrt{n}}$	$\frac{\sigma}{\sqrt{n}}$
- amostra >=30 - D.p. pop. desc.	$\overline{X} \pm Z\frac{s}{\sqrt{n}}$	$Z\frac{s}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$
- amostra <30 - D.p. pop. desc.	$\overline{X} \pm t\frac{s}{\sqrt{n}}$	$t\frac{s}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$
P/ proporção amostral (b)	Intervalo de confiança p/ média	Erro de estimativa (margem de erro) p/média (*)	Erro padrão (*)
	$\hat{p} \pm Z\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}$	$Z\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}$	$\sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}$

Tamanho da amostra	
$E = Z\frac{\sigma}{\sqrt{n}}$ $\sqrt{n}.E = Z\sigma$ $\sqrt{n}.E = Z\sigma$ $\sqrt{n}.E = \frac{Z.\sigma}{E}$ $n = (Z\frac{\sigma}{E})^2$	$n = Z^2\frac{.p(1-p)}{e^2}$

Teste de Hipótese <u>para uma amostra</u> (utiliza valor Z ou T)

- As hipóteses são sempre calculadas para parâmetros populacionais

-Não precisa ser realizado com toda a população

-Similar a intervalo de confiança, no entanto, retorna sim ou não

-Coleta amostra e verifica se há compatibilidade com o valor informado

$H_0 = \textit{Hipótese nula} = \textit{deve conter sinal de igualdade obrigatoriamente}$

$H_0 = \mu = Mo$

$H_1 = \textit{Hipótese alternativa} = \textit{desigualdade}$

$\mu \neq Mo$ - bilateral

$\mu > Mo$ - unilateral à direita

$\mu < Mo$ - unilateral à esquerda

- Nível de significância: probabilidade do erro. Definido pelo investigador, assim sendo, trata-se de parâmetro subjetivo.. Área determinada pelo Z calculado. (chamado nível crítico). "Probabilidade máxima permitida para cometer erro do tipo I"
- Nível de confiança: complemento do nível de significância (lembrando que o total é 1)

Teste de hipótese		
Para a média (permite Z ou T)		Para a proporção (permite apenas Z)
$Z_{cal} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ Z_{cal} = Estatística do teste	$t_{cal} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$	$Z_{cal} \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
Etapas para a aplicação: 1. Calcula Z tabelado: define a área de rejeição/aceitação (costuma-se usar o valor de 5%, Z= 1,96) 2. Calcula Z calculado para a média : verifica posicionamento do valor dentro da distribuição normal		

Tipos de erro no teste de hipótese:

Erro do tipo I : rejeita H_0 quando deveria aceitar

$[\alpha = 1 - \textit{nível de confiança}]$

Erro do tipo 2 : aceita H_0 quando deveria rejeitar

$[\beta = 1 - \textit{potência do teste}]$

Valor - P (Φ)

Importante: o valor o também chamado de: probabilidade de significância do teste ou nível descritivo do teste (Varia de 0 a 1)
O nível de significância é o α (determinado pelo pesquisador)

Nível crítico: apresenta nível de aceitação/ rejeição.

Valor-p: é a área projetada pelo Z calculado. Representa a probabilidade de a amostra pertencer a distribuição do valor populacional testado. Pode ser utilizado como uma forma alternativa para que não se calcule a estatística de teste.

Quanto menor o valor de P, maior a probabilidade de rejeitar H_0

área valor p > área nível de significância: aceita H_0.

área valor p ≤ área nível de significância: não aceita H_0.

Para encontrar o valor P busca-se na tabela t/z

Teste de hipótese Qui - Quadrado

Variáveis qualitativas ou discretas

-Observações independentes entre si

-Observação contabilizada por contagem, frequência e proporções

-Cada observação pertence a apenas uma classe

-Não pode ser aplicada a amostra inferior a 5 observações por classe

Hipóteses do teste Qui - Quadrado

H0: Fobs = Fesp

H1: Fobs = Fesp

χ *tabelado* = encontrado na tabel

χ *calculado* = calculado pela fórmula

Teste qui- quadrado: não utiliza Z e nem t. Utiliza Qui-Quadrado

$\chi^2 = \sum Z_i^2$

-Qui- quadrado tabelado

-Qui- quadrado calculado

Fo: frequência observada

Fe: frequência esperada

H0: Fo - Fe

H1: Fo ≠Fe

Tipos de teste Qui-quadrado Aplicado apenas para variável categórica ou discreta, não pode ser aplicado em amostras muito pequenas. N<5	
Teste de adequação de ajustamento: 1 variável qualitativa	Teste de independência: 2 ou + variáveis qualitativas
$\chi^2 tabelado$ =GL GL = n-1	$\chi^2 tabelado$ =GL GL = (col-1).(lin-1)
$\chi^2 calculado = \sum \frac{(Fo-Fe)^2}{Fe}$	
Desvio = Fo - Fe (no entanto, o resultado dessa equação seria zero. Então, elevamos ao quadrado. Desvio quadrado = $(Fo - Fe)^2$ Proporção: $(Fo - Fe)^2$/Fe	1 - seleciona a variável com menor número de classes, calcula o % e desenvolve a tabela de valor esperado.

Variáveis bidimensionais

Covariância: + infinito a - infinito Correlação: +1 a -1 (indica grau de similaridade entre variáveis diferentes) Variância : apenas assume valores positivos Coeficiente de determinação R^2: apresenta apenas variação de 0 a 1

Distribuição de probabilidade conjunta e marginal

x	0	1	Total
1	⅓.1/4	⅔.1/4	1/4
2	⅓.1/4	⅔.1/4	1/4
3	...	...	1/4
4	...	...	1/4
Total	1/3	2/3	

	Conjunta
	Marginal

Probabilidade de variáveis discretas independentes
Como saber se as variáveis são independentes?
A probabilidade conjunta deve ser o resultado da multiplicação das respectivas marginais.
Nosse caso P(x) = P(x). p(y)

Probabilidade de variáveis discretas	
dependentes	independentes
$P_{(x=3\|\, y=1)} = \frac{P(x=3\ e\ y=1)}{P(y=1)}$	$P_{(x,y)} = p(x).\ p(y)$ p/ todos os pares x e y. Em variáveis independentes a covariância é igual a zero.
Como saber se as variáveis são dependentes? A probabilidade conjunta **não** é o resultado da multiplicação das respectivas marginais.	

Covariância [valor positivo ou negativo]
-: direção contrária
0: variáveis independentes OU desvios se anulam
+: mesma direção
* Não possui valor mínimo nem máximo

Medida descritiva que indica a relação de dependência entre duas variáveis. Pode ser aplicada tanto em estatística descritiva quanto inferencial.

Covariância como variável descritiva: a covariância é similar a variância, a diferença é que aqui analisamos 2 ou mais variáveis, enquanto na variância analisamos apenas uma.

x	y	(x - $\bar{x}$)	(y - $\bar{y}$)	(x - $\bar{x}$). (y - $\bar{y}$)
2	5	-3	-4	12
3	7	-2	-2	4
6	8	1	-1	-1
9	16	4	7	28
				43

$Cov\,(x, y) \;=\; \frac{43}{4}$ unidade: dominador/ denominado

$$Cov\,(x, y) \;=\; \frac{\sum(x-\bar{x}).\,(y-\bar{y})}{n}$$

Lembre-se de verificar se é n ou se é n-1

Fórmula alternativa da covariância Cov(X,Y) = Média Produto XY - Produto da média X e Y		
x	y	Produto X e Y
2	5	10
3	7	21
6	8	48
9	16	144

OBS: se fosse amostra seria n-1
Para corrigir a unidade de medida usamos a correlação.

Lembre-se: se as variáveis são independentes se a covariância é 0!

Covariância =0 não indica necessariamente que as variáveis são independentes!

Qual o problema da covariância? variáveis estão em unidades diferentes. Para solucionar o problema calculamos o coeficiente de correlação.

Covariância como variável aleatória

Correlação (r) [valor entre -1 e +1]: indica a "força" que mantém unidas duas variáveis.

$$r \;=\; \frac{Cov\,(X,Y)}{S_x \cdot S_y}$$

$$r = \frac{\sum(x_i-\bar{x}).(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 . \sum(y_i-\bar{y})^2}}$$

Coeficiente de correlação (r)	Correlação Positiva	Coeficiente de correlação (r)	Correlação Negativa
r = 1	Perfeita	r = - 1	Perfeita
0,95 ≤ r < 1	Muito forte	- 0,95 ≤ r < -1	Muito forte
0,8 ≤ r < 0,95	Forte	-0,8 ≤ r < -0,95	Forte
0,5 ≤ r < 0,8	Moderada	-0,5 ≤ r < -0,8	Moderada
0 ≤ r < 0,5	Fraca	0 ≤ r < -0,5	Fraca

1: variação da variável A é explicada por B
0: não há variação entre as variáveis
-1: correlação negativa entre as variáveis
"Força que mantém unidos dois conjuntos de valores, por ser adimensional podemos comparar os valores (o que não ocorre na covariância)"

0,0019 a 0,19 - muito fraca
0,20 a 0,39 - fraca
0,40 a 0,69 - moderada
0,70 a 0,89 - forte
0,90 a 1 - muito forte

IMPORTANTE: correlação não é causa !

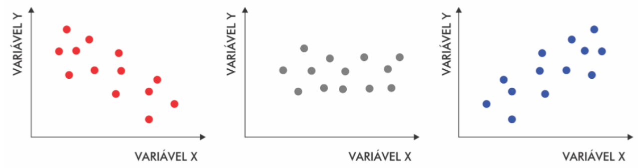
$r(x, y) = \frac{Cov(x,y)}{s_x s_y}$ [remove unidades de medidas]

0: não possuem dependência ou a correlação não é linear
s: desvio padrão de x e desvio padrão de y

A covariância não permite realizar comparações por conta da unidade de medida, nesse caso, utilizamos a correlação para poder comparar as variáveis. Dessa forma, como a correlação é adimensional, podemos realizar comparações.

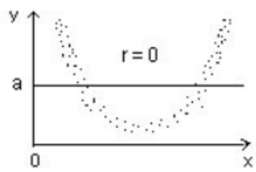
-Costuma usar gráfico de dispersão/ correlação

Correlação negativa, neutra, positiva



Correlação inexistente

f) Correlação inexistente



correlação positiva

Pesquisar sobre roat mean square error (RMSE)
- Interpretação facilitada

Pendente:
coeficiente de determinação: $(coeficiente\ de\ correlação)^2$

Correlações paramétricas:
Pearson
Correlação não paramétrica
Sperman: permite uso para dados lineares e não lineares
Kendall: utilizado para amostras de até 30 elementos ou para população com grande quantidade de empates

Teste de hipótese para correlação linear

Teste t

$$t = \frac{r}{\sigma_r} = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}}$$

Graus de liberdade

$$gl = n - 2$$

Coeficiente de Determinação:
Porcentagem da variação de y que pode ser explicada pela relação entre x e y

$$r^2 = \frac{Variação\ encontrada}{Variação\ total}$$

$$r^2 = \rho^2$$

$$y = m.x + b$$

Coeficientes

$$m = \frac{\Sigma(x_i - \bar{x}).(y_i - \bar{y})}{\Sigma(y_i - \bar{y})^2}$$

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

Cálculo do Coeficiente de Spearman

$$r_R = 1 - \frac{6\Sigma_i d_i^2}{n(n^2 - 1)}$$

n = número amostras.
di = diferença de alcance de cada elemento.

Coeficiente de correlação (r_R)	Correlação Positiva	Coeficiente de correlação (r_R)	Correlação Negativa
$r_R = 1$	Perfeita	$r_R = -1$	Perfeita
$0,95 \le r_R < 1$	Muito forte	$-0,95 \le r_R < -1$	Muito forte
$0,8 \le r_R < 0,95$	Forte	$-0,8 \le r_R < -0,95$	Forte
$0,5 \le r_R < 0,8$	Moderada	$-0,5 \le r_R < -0,8$	Moderada
$0 \le r_R < 0,5$	Fraca	$0 \le r_R < -0,5$	Fraca

Cálculo do Coeficiente de Kendall

$x_i > x_j$ e $y_i > y_j$ ou se $x_i < x_j$ e $y_i < y_j$.

$$\tau = \frac{(\text{quantidade de pares concordantes}) - (\text{quantidade de pares discordantes})}{n(n-1)/2}$$

Coeficiente de correlação (τ)	Correlação Positiva	Coeficiente de correlação (τ)	Correlação Negativa
$\tau = 1$	Perfeita	$\tau = -1$	Perfeita
$0,95 \le \tau < 1$	Muito forte	$-0,95 \le \tau < -1$	Muito forte
$0,8 \le \tau < 0,95$	Forte	$-0,8 \le \tau < -0,95$	Forte
$0,5 \le \tau < 0,8$	Moderada	$-0,5 \le \tau < -0,8$	Moderada
$0 \le \tau < 0,5$	Fraca	$0 \le \tau < -0,5$	Fraca

Análise de Regressão Linear

(explica o formato da regressão)

REGRESSÃO LINEAR SIMPLES

Variável X: 1 independente
Variável Y: 1 dependente
y = a + bx + e

REGRESSÃO LINEAR MÚLTIPLA

Variável X: n variáveis x
Variável y: 1 variável y
y = a + b1 + b2 +...+ bn + e

OBS: ao comparar mais de 3 variáveis, não é possível representar graficamente

a: constante da regressão, coeficiente linear, intercepto
valor de x que intercepta y

b: coeficiente de regressão, coeficiente angular, inclinação da reta.. Indica a inclinação da reta representa o crescimento de y a cada uma unidade de x. Indica se a reta será /, \ ou - .

Como calcular o valor de A e B?

Estimadores dos mínimos quadrados

b= $\frac{Cov\,(x,y)}{s_x^2}$ unidade x/ unidade y

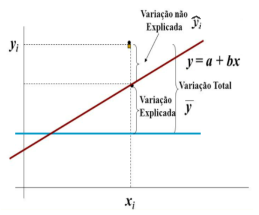
$a = \overline{Y} - b\overline{X}$

ERROS DA REGRESSÃO

ERROS DA REGRESSÃO

e = $y_i - \widehat{y}_i$

VARIAÇÃO TOTAL
Explicada: $\widehat{y}$ = a + bx
Não explicada: e = y = $\widehat{y}_i$



Erros da regressão:

$(y - \overline{y})$ = $(\widehat{y}$ - $\overline{y}$) + $(y$ - $\widehat{y})$

$\Xi(y - \overline{y})^2$ = $\Xi(\widehat{y} - \overline{y})^2$ + $\Xi(y$ - $\widehat{y})$

SQT = SQE + SQR

Var. total = Var. explicada + Var. erro

Variação = soma dos quadrados

Variância dos erros da regressão

$se^2 = \frac{\Sigma\,(yi - \widehat{yi})^2}{n-2}$

Desvio padrão dos erros da regressão
("Erro padrão da estimativa da regressão)

$se = \sqrt{se^2}$ = $\frac{\sqrt{\Sigma\,(yi - \widehat{yi})^2}}{\sqrt{n-2}}$

Teste de hipótese da regressão (usa apenas teste t):
Ho: não existe regressão linear
H1: existe regressão linear
tcal = $\frac{\overline{X} - \mu}{s_x}$ = $\frac{b - 0}{s_b}$ = $\frac{b}{s_b}$ $= \frac{se^2}{\sqrt{(x - \overline{x})^2}}$

Coeficiente de Determinação (r):
$r^2 = \frac{SQE}{SQT} = \frac{explicada}{total}$

Análise de variância: (usa f de Snedecor):
$f\ de\ Snedecor = \frac{QME\ (explicada)}{QMR\ (residual)}$
Indica qtas vezes a explicada é maior q a residual.

Teste de hipótese dos erros da regressão:
Ho: valor explicado = valor residual ∴ não há linearidade entre x e y
H1: variância explicada > variância residual

Covariância x Correlação

Covariância: unidade medida em min/ kg ∴ não podemos comparar. Para comparar utilizamos a correlação. ®.
$\frac{\Xi(x - \overline{x}).(y - \overline{y})}{(n)\ ou\ (n-1)}$ = $\Xi\,xy - ((\Xi x\,.\,\Xi y)/n)$
Probabilidade = $\Xi\,\epsilon(xy) - \Xi\epsilon(x)\,.\,\Xi\epsilon(y)$

Correlação: (r): + 1 a - 1

$r = \frac{Cov\,(x,y)}{s_x z_y}$ =

Propriedades da covariância:
(+ e -): não altera
(* e /): * ou /

Variáveis dependentes:
Var (x± y) $= Var\,(x) \pm Var\,(y) \pm 2\,Cov\,(x,y)$

Variáveis independentes:
Var (x± y) = Var(x) + Var (y)

Propriedades da Correlação (s):
só pode ser alterada pela * ou / de valores negativos.
∴ altera apenas o sinal.

Teste paramétrico: **analisam variáveis dependentes com distribuição conhecida. Costuma-se dizer que os testes paramétricos são utilizados para distribuição normal por conta do Teorema do Limite Central.**

Pré-requisitos:

a. Dados em intervalo (variável dependente precisa ser numérica, discreta ou contínua)

Dados estão em intervalo			Dados não estão em intervalo	
Cat 1	7,2		Cat 1	Sete
Cat 2	6		Cat 2	Seis
Cat 3	5,4		Cat 3	Cinco

b. Independência dos dados amostrais

c. Normalidade

Testes para verificar normalidade:

>> Kolmovrov-Smirnov (amostra>50)

>> Shapiro-Wilk (amostra<50)

>> Anderson Darling

>> entre outros

Ho: distribuição é igual à normal ($p>0.05$)

H1: distribuição é diferente da normal ($p<0.05$)

(em caso de H1 usa-se testes não paramétricos ou realiza-se a transformação dos dados)

d.Homogeneidade da variância

>>Para verificar homogeneidade aplica-se teste de Levene

Ho: variância é homogênea ($p>0.05$)

H1: variância não é homogênea ($p<0.05$)

(em caso de H1 pode-se aplicar a correção de Weich)

Teste paramétrico::**analisam variáveis dependentes com distribuição NÃO conhecida**

Grupo Pareado: teste do mesmo grupo em intervalo de tempo ou submetido a diferentes testes. *Nesse caso, as amostras devem ser de tamanho igual.*

Grupo Independente: grupos iniciais são diferentes, ex: feminino e masculino. *Nesse caso as amostras podem ser de tamanhos diferentes.*

Testes paramétricos: análise descritiva realizada com média e desvio padrão

Testes não paramétricos: análise descritiva realizada com média (2 quartil), e variabilidade (1 e 3 quartil)

!para calcular o intervalo de confiança é necessário saber a distribuição dos dados.

Variável dependente: resposta

Variável independente: variável de grupamento

! Dados qualitativos não podem ser descritos com média e desvio padrão ∴ só pode ser descrita por mediana, primeiro e terceiro quartil.

! Na análise pareada o grupo deve ser do mesmo tamanho

! Em grupos independentes os grupos não precisam ser do mesmo tamanho

comumente utilizado quando uma suposição do teste paramétrico foi violada. Possuem menor probabilidade para detectar um efeito e não possuem intervalo de confiança.

Ideal para dados distorcidos, ou que não estão em distribuição normal ou que apresentam ou

Qual o melhor teste de comparação?	**Para uma amostra**		
	Testes paramétricos		Testes não paramétricos
	Teste Z Desvio padrão populacional conhecido ou n>=30	Teste t Desvio padrão populacional desconhecido e n<30	Teste de sinais

Qual o melhor teste de comparação?			**VARIÁVEL INDEPENDENTE**			
			Qualitativa nominal			
VARIAVEL	QUALIT	grupo	2 grupos		3 grupos ou mais	
		Distribuição	Independentes e não pareados	dependentes /pareados	independente e não pareado	dependentes/ pareados
		Teste	Qui-quadrado/ Fisher	MC Nemar	Qui-quadrado/ Fisher	Q de Cochram
			Qualitativa Ordinal			
DEPENDENTE	QUALIT	grupo	2 grupos		3 grupos ou mais	
		Distribuição	Independentes e não pareados	dependentes /pareados	Independentes e não pareados	dependentes /pareados
		Teste	Mann - Whitney (versão não paramétrica do T de Student)	Wilcoxon	Kruskal - walls +comparação múltipla	Friedman +comparação múltipla
			Discreta/Continua			
	QUANTIT	grupo	2 grupos		3 grupos ou mais	
		Distribuição	Independentes e não pareados	dependentes /pareados	independente e não pareado	dependentes/ pareados
		Teste	- Paramétrico T de Student	-Paramétrico T de Student Pareado	- Paramétrico: ANOVA	- Paramétrico ANOVA de medidas repetidas

A T I V A	- Não paramétrico Teste de Mann-Whitney ("T de Student não paramétrico")	-Não paramétrico Teste de Wilcoxon ("Teste T pareado não paramétrico")	+comp. múltipla (Tukey) - Não paramétrico Kruska Walls + comp. múltipla (Dunn's)	+comp. múltipla (Tukey) - Não paramétrico Friedman + comp. múltipla (*Rever porque foi usado Tukey em vez de Dunn's - https://www.youtube.com/watch?v=piC-hsYaz8k)
	OBS: teste de Mann- Whitney e podem ser aplicados tanto a variáveis quantitativas não paramétricas quanto a variáveis qualitativas ordinais		OBS: teste de Kruska Walls e Friedman podem ser aplicados tanto a variáveis quantitativas não paramétricas quanto a variáveis qualitativas ordinais	
Estudo pendente: Análise de Variância (ANOVA) a dois critérios (SigmaPlot 12.0) +Tukey [Revisar] fator 1: 2, 3 ou + grupos; fator 2: 2, 3 ou + grupos Alguns testes de comparação múltipla: Tukey, Bonferroni, Dunn, Dunnet, Fisher LSD, Sidak, Student-Newman-Keuls, Duncan, Scheffé				

Qual o melhor teste de correlação a ser utilizado?		
Teste de correlação de Pearson	não há variáveis independentes	2 variáveis dependentes quantitativas (apenas 2) *Cada unidade amostral deve ser analisada pelas duas variáveis dependentes
Teste de correlação de Spearman		2 variáveis dependentes, sendo: - quantitativa + qualitativa ordinal ou - qualitativa ordinal + qualitativa ordinal pelo menos uma variável avaliada não tem distribuição normal (ex: qualitativa)

Comparação, correlação ou regressão?		Variável Dependente		
		Qualitativa nominal	Quantitativa nominal	Qualitativa
Variável Independente	Qualitativa nominal	Associação	Comparação não paramétrica	Comparação paramétrica ou Regressão Logistica
	Quantitativa ordinal	Regressão logistica	Correlação não paramétrica ou Regressão logistica ordinal	Comparação não paramétrica ou Regressão simples
	Quantitativa	Regressão logistica	Regressão logistica ordinal	Correlação paramétrica ou Regressão simples