UNIVERSIDADE ESTADUAL DE CAMPINAS
CIÊNCIAS DA NUTRIÇÃO ESPORTE E METABOLISMO
FACULDADE DE CIÊNCIAS APLICADAS

APOSTILA

EN004 – BIOESTATISTICA

Prof. Diogo Thimoteo da Cunha
Profa. Ligiana Pires Corona

Elaboração e supervisão

Diogo Thimoteo da Cunha

Ligiana Pires Corona

Apoio na elaboração

Camila de Mello Marsola

Luis D'Avoglio Zanetta

Giovana Santarosa Cassiano

Mariana Bessi Pereira

Mariana Piton Hakim

Limeira

Ano 2020

Atualizado 2022

Cronograma 2022

[redacted]

[redacted]

29/11	Encerramento da disciplina

Avaliação

As avaliações terão caráter formativo, contínuo e processual e estarão pautadas na exigência de 75% (setenta e cinco) de presença e média final igual ou superior a 5,0 (cinco). O aluno com nota igual ou superior a 5,0 está dispensado da realização do Exame. O aluno com nota entre 2,5 e 4,9 poderá fazer o Exame Final na data prevista no calendário acadêmico.

Os conceitos serão atribuídos conforme nota final sendo eles:

[redacted]

A	8,00 a 10,00

A nota final será composta por: Trabalho de análise de dados (50%) + Avaliação escrita (50%)

A disciplina não prevê prova substitutiva ao aluno que não tenha realizado alguma prova por motivo de falta bem como entrega posterior de exercícios.

Objetivos da disciplina

- Compreender a importância da bioestatística na formação e prática do

pesquisador - Compreender conceitos básicos de interpretação de dados

estatísticos - Conhecer as diferentes técnicas e testes estatísticos para aplicação

em estudos da saúde

Bibliografia básica

MANN, Prem S. Introdução à estatística. 8. ed. Rio de Janeiro, RJ: Livros Técnicos e Científicos, 2015. 765 p.

Sugestões gerais

1) A disciplina de bioestatística é densa e com diversos conceitos novos. Portanto sugerimos que os alunos dediquem pelo menos uma hora de estudos semanalmente, não deixando a matéria se acumular.
2) Utilize o gabarito apenas após ter feito todo exercício.
3) Formem grupos de estudos com os demais alunos. Grupos de três a cinco alunos são uma ótima forma de estudar. Assim, um pode tirar a dúvida do outro.
4) Em caso de dúvidas, procure os professores nos horários determinados. 5) Leia criticamente a seção “análise de dados” dos artigos da sua área. Verifique a descrição e testes mais utilizados.
6) Don’t panic!

Introdução a estatística

A palavra estatística vem do latim status e significa estado. Inicialmente, era utilizada para compilar dados que descrevem características de países (Estados), fortalecendo-se como ciência entre os séculos XVI e XVIII, pois com a emergência do estado moderno, juntamente com a implantação do modo capitalista de produção, havia a necessidade de contar o povo e o exército. Em 1662, John Graunt publicou estatísticas de nascimentos e mortes. A partir de então, o estudo dos eventos vitais e da ocorrência de doenças e óbitos impulsionou o desenvolvimento da Estatística nos campos teórico e aplicado (Triola, 1999).

Atualmente, índices e indicadores estatísticos fazem parte do dia a dia, tais como taxa de inflação, índice de desemprego, taxa de natalidade, taxa de crescimento populacional, índice de poluição atmosférica, índice de massa corporal, entre outros.

Estatística: é uma coleção de métodos para planejar experimentos, obter e organizar dados, resumi-los, analisá-los, interpretá-los e deles extrair conclusões (Triola, 1999).

Bioestatística = Estatística aplicada às ciências da vida.

1 Conceitos básicos de estatística

Neste capítulo, são apresentados alguns conceitos básicos que serão retomados em todas as aulas ao longo do curso, portanto, utilize-o sempre para consultas e revisões.

1.1. Por quê conhecer a bioestatística?

a. Para avaliar a literatura: não é possível ler, interpretar e criticar adequadamente os artigos científicos que usamos sem compreender os métodos estatísticos utilizados;

b. Para interpretar estatísticas vitais (como taxas de mortalidade, natalidade, etc);

c. Para aplicar os resultados dos estudos no cuidado com pacientes: os artigos científicos publicados com estatística adequada nos permitem chegar à conclusões que serão aplicadas diretamente na prática profissional, desde que corretamente

interpretadas;

d. Para entender problemas de saúde da população e utilizar a informação para planejamento de serviços e tratamento;

e. Para desenvolver novas técnicas em saúde (tecnologias, equipamentos, métodos diagnóstico, medicamentos) a partir da adequada interpretação de dados de pesquisa;

f. Para planejar e desenvolver pesquisas na área de biológicas

1.2 Tipos de análise estatística

- ANÁLISE DESCRITIVA: Análise exploratória
  - Tabelas e gráficos
  - Medidas numéricas de resumo de dados: média e desvio padrão, mediana, quartis, percentis, moda, extremos, frequências
- INFERÊNCIA ESTATÍSTICA: Fazer afirmações sobre características de uma população, baseando-se em resultados de uma amostra.
  - Estimação por ponto
  - Intervalo de confiança
  - Testes de hipóteses

2.3 Dado ou informação?

Dados são frutos de observações às quais se atribuem significados. São resultantes de enumerações (contagens), na tentativa de aproximação ou de descrição de eventos ou fenômenos diversos. Portanto, o conceito de dado é um valor quantitativo não trabalhado,

isto é, sem ter sido submetido a algum tipo de tratamento matemático ou estatístico que irá agregar valor ao seu significado.

O tratamento matemático ou estatístico transforma dado em informação, com o objetivo de complementar a interpretação e permitir comparações.

A coleta de dados é a observação e registro da categoria ou medida de variáveis relacionadas ao objeto de estudo que ocorrem em elementos (indivíduos) de uma amostra ou população.

- Elementos: são unidades de análise; podem ser pessoas, domicílios, escolas, creches, células, dentes, ou qualquer outra unidade

● Variáveis: São as características que queremos estudar na nossa amostra / população Em estatística, é importante diferenciar os conceitos de parâmetro e estimativa:

● Parâmetros: são os valores populacionais (desconhecidos): As características ou atributos são observados em cada elemento da população e agregados por meio de medidas estatísticas

● Estimativas: são os valores “aproximados” pela amostra. São as medidas estatísticas que agregam os valores com base na informação disponível, ou seja, da amostra

1.3. População e Amostra

● População alvo: aquela que detém as características clínicas e demográficas que queremos estudar

● População acessível: deve conter características geográficas e temporais específicas que determinarão uma parte da população alvo

● População em estudo: representativa da população acessível e fácil de trabalhar

● População: é o conjunto de elementos que apresentam em comum determinadas características definidas para o estudo. Pode ser finita e pequena (ex: população de pacientes internados no hospital Santa Casa), mas na maioria das vezes, apesar de finita, ela é incontável, ou até infinita (ex: População do município; população de pacientes portadores de diabetes no Brasil).

● Amostra: é um subconjunto de elementos de uma população, selecionado segundo algum critério. Pode ser de vários tipos, representativas ou não da população.

- IMPORTANTE: ela pode determinar o erro presente na estimativa e também a confiabilidade de um estudo.

A amostra adequada precisa representar todos os segmentos da população total a ser estudada. Para tanto, é necessário discutir os conceitos de critérios de inclusão e exclusão, validade, e erro amostral.

Os critérios de inclusão são aqueles que visam a caracterização da população alvo, e a definição (no tempo e no espaço) da população acessível. Os critérios de exclusão são aqueles que determinam a eliminação de indivíduos cuja inclusão diminuiria a qualidade dos dados e/ou da interpretação dos resultados.

Em relação à validade, uma boa amostra precisa garantir validade interna e externa, o que vai determinar o quanto os resultados da população de estudo podem ser inferidos para outras populações. Validade externa é obtida quando a população acessível é representativa da população alvo, em relação ao fenômeno de interesse. Validade interna é obtida quando a população de estudo é representativa da população acessível.

A aplicação de técnicas adequadas de amostragem pode garantir a validade externa e interna buscando, a partir de estimativas obtidas numa amostra, EXTRAPOLAR seus resultados para a população, ou seja, fazer uma INFERÊNCIA.

1.3.1 Cálculo da amostra e Erro amostral

O erro amostral é a diferença entre o valor que a estatística pode acusar e o verdadeiro valor do parâmetro que se deseja estimar, ou seja, é o erro causado por observar uma amostra em vez da população inteira. Como a amostra não inclui

todos os membros da população, é esperado que os parâmetros sejam diferentes das estimativas.

Ex: medir a altura de mil indivíduos de um país com um milhão de habitantes, a altura média dos mil indivíduos é tipicamente diferente da altura média de todos os habitantes no país, pois não se pode esperar que duas amostras, independentemente retiradas da mesma população, forneçam resultados iguais valores reais da população são desconhecidos.

Para determinação do tamanho da amostra, o pesquisador precisa especificar seu erro Amostral tolerável, ou seja, o quanto ele admite errar na avaliação dos parâmetros de interesse (Ex: pesquisas eleitorais em geral apresentam: esta pesquisa tolera erro de 2%).

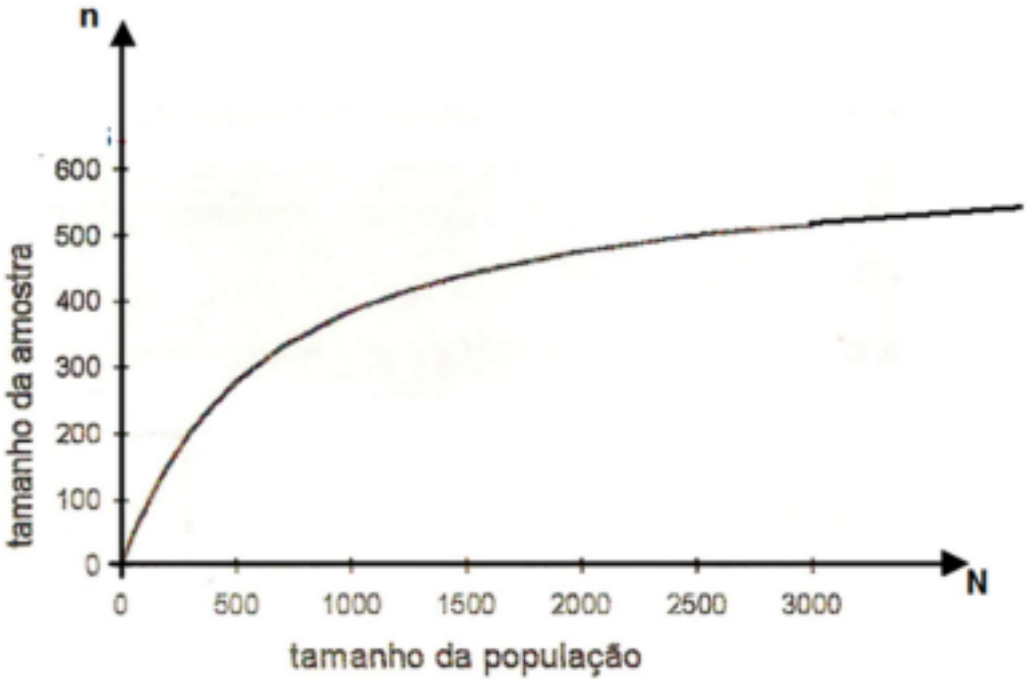
O erro amostral pode ser estimado por modelagem probabilística da amostra, e é controlável por ações como:

- Técnicas de amostragem: optando por aquela que, no caso concreto, se revelar mais eficiente (exemplo: amostragem aleatória e aumento do tamanho da amostra), para assegurar a representatividade e associar os resultados com grau de confiança elevado;
- Estimadores: optando por aquele que for mais eficiente, isto é, com menor variabilidade.

O erro aleatório é praticamente inevitável em estudos, pois as estimativas se comportam aleatoriamente em torno do verdadeiro valor do parâmetro, i.e., se concentram em torno de um valor central que coincide com o verdadeiro valor do parâmetro. Se o erro não é aleatório, temos então um viés de seleção - AMOSTRA ENVIESADA (Ex: utilizar amostra do time de basquete para estimar a média de estatura da população geral).

É importante notar que em geral é necessário saber o número total da população alvo para cálculo da amostra (Ex: número total de idosos do município de Limeira para calcular uma amostra de idosos para um determinado estudo). Quando não se conhece o total da população, pode ser necessário o uso de estudo piloto para conhecer os parâmetros necessários para o cálculo (média, desvio padrão, proporção, etc).

Em relação ao tamanho da amostra, é errônea a ideia de que para uma amostra ser REPRESENTATIVA ela deva abranger uma percentagem fixa da população, pois o n não é uma Função Linear de N, mas sim a seguinte função:



Alguns cuidados importantes para a determinação do tamanho da amostra (Miot, 2011):

- Leia artigos da mesma área! Eles em geral descrever como foi feita a amostragem;
- No caso de grandes bancos de dados (estudos populacionais): o aumento da amostra pode reduzir os intervalos de confiança das estimativas e permite a detecção de diferenças entre subgrupos que, apesar de estatisticamente significantes, não possuem relevância clínica. O olhar crítico do pesquisador sobre o resultado estatístico ainda é a melhor estratégia;
- Em estudos clínicos (intervenções cirúrgicas, medicamentos, etc), por motivos éticos, utiliza-se o menor n possível. Estudos com testes de equivalência, de não-inferioridade e de concordância utilizam dimensionamentos amostrais próprios, distintos dos testes de diferenças de médias e de proporções comumente usados.

TIPOS DE ERROS

O cálculo amostral para comparação de subgrupos (testes de hipóteses) dentro de uma amostra depende:

- do teste estatístico escolhido;
- das diferenças entre os grupos;
- da tolerância do pesquisador à detecção de diferenças quando elas não existem (erro tipo I - erro alfa) ou da falha na detecção de diferenças entre os subgrupos

quando elas realmente existem (erro tipo II - erro beta)

1.3.2 Tipo de amostra

Amostragem não probabilística: é uma amostra de conveniência, que não garante a mesma probabilidade de todos os indivíduos da população de pertencerem à amostra.

A. Amostragem consecutiva (cumulativa): São selecionados os indivíduos que preenchem o(s) critério(s) de inclusão, em um específico intervalo de tempo ou tamanho de amostra. A limitação deste tipo de amostra é que, em caso de tempo curto de coleta de dados, pode não incluir todos os indivíduos que seriam representativos da população.
B. Amostragem por conveniência: São selecionados os indivíduos que estão facilmente disponíveis. A maior limitação deste tipo de amostra é que os voluntários poderão não ser representativos da população acessível (p. ex.: só são coletados dados de indivíduos mais sadios ou dos mais doentes, ou dos mais expostos, etc).
C. Amostragem intencional: Da população acessível são escolhidos aqueles indivíduos que se julga serem mais apropriados para o estudo. A limitação é que na maioria das vezes não é possível inferir as conclusões para outras populações, já que a amostra foi determinada pelo pesquisador.

Amostragem probabilística: Cada elemento na população tem probabilidade conhecida e diferente de zero de pertencer à amostra. Ela pode ser realizada de várias maneiras:

A. Casual simples ou equiprobabilística: Cada elemento da população tem a mesma chance de entrar na amostra. Ex: Sorteio simples. Pode ser utilizada tabela de números aleatórios, programas computacionais, aplicativos, etc.
B. Sistemática: Seleciona-se qualquer unidade amostral e, a partir dela, escolhem-se as seguintes de acordo com o intervalo determinado, utilizando a ordenação natural dos elementos da população (prontuários, residência, ordem de nascimento, etc).

Exemplo 1: para uma escola com salas de 30 alunos, sorteou-se os números 14, 21 e 27. Em todas as salas, esses serão os alunos parte da amostra.

Exemplo 2: Intervalo de amostragem: $k = N / n$, onde:

N= tamanho da população;

n = tamanho da amostra;

Início casual i, sorteado entre 1 e k, inclusive.

Amostra sorteada é composta pelos elementos: i, i+k, i+2k,..., i+(n-1)k

OBS: É necessário ter cuidado com a periodicidade dos dados, p.ex. se for feito sorteio de dia no mês, pode cair sempre em um domingo onde o padrão do evento pode ser diferente.

Suponha: N=80; n=10;

k = N / n = 80/10

k = 8; início casual: $1 \leq i \leq 8$

Começo casual sorteado: i=4

Amostra composta dos elementos:

i	4
i+k	12
i+2k	20
i+3k	28
i+4k	36
i+5k	44
i+6k	52
i+7k	60
i+8k	68
i+(n-1)k	76

C. Estratificada: De acordo com algum fator como sexo, faixa etária, nível social. Os sorteios ou seleção são realizados dentro destes grupos separadamente.

TABELA 1 – Tamanho das amostras segundo sexo e grupo etário.

GRUPO(h)	SEXO	POP	f1	n1	n2	f2
60-64	MASC	119.066	0.0018	213,5	214	0,0018
65-69	MASC	95.938	0.0018	172,0	172	0,0018
70-74	MASC	64.834	0.0018	116,2	116	0,0018
75-79	MASC	36.112	0.0018	64,77	258*	0,0071
80 ou+	MASC	30.271	0.0018	54,29	273*	0,0090
60-64	FEM	150.884	0.0018	270,6	271	0,0018
65-69	FEM	27.926	0.0018	229,4	229	0,0018
70-74	FEM	92.614	0.0018	166,1	166	0,0018
75-79	FEM	57.641	0.0018	103,3	258*	0,0045
80 ou+	FEM	60.937	0.0018	109,3	273*	0,0045
TOTAL		836.223		1.500	2.230	

* n2 = n1x2,5 n2 = grupos etários femininos

Fonte: SILVA, 2003.

D. Por conglomerado: O conglomerado é um conjunto de elementos formando uma unidade amostral. Todos os indivíduos sorteados devem fazer parte do conglomerado.

Ex: Estudo Saúde, Bem-Estar e Envelhecimento (Silva, 2003). Critério de partilha proporcional ao tamanho (PPT), utilizando 72 Setores censitários sorteados entre os 263 setores do município de São Paulo. Número mínimo de 90 domicílios sorteados em cada setor, calculado pela média (5882/72). Dividiu-se o total de endereços de cada setor em segmentos de 10 domicílios e em cada setor sortearam-se 9 segmentos.

Fonte: SILVA, N. Aspectos metodológicos - Processo de amostragem. In: LEBRÃO, M. e DUARTE, Y. (Ed.). O projeto SABE no município de São Paulo: uma abordagem inicial. Brasilia: Organização Pan-Americana da Saúde, 2003. p.47-58.

Atividade 1

Os dados a seguir são de peso (kg) de 80 mulheres identificadas pela variável id (identificação). Utilize estes dados para realizar os exercícios 1 e 2.

Id	Peso	Id	peso	Id	Peso	Id	Peso	Id	Peso	Id	Peso
1	65	16	71	31	70	46	75	61	68	76	75
2	65	17	84	32	72	47	79	62	69	77	79
3	58	18	63	33	75	48	79	63	76	78	73
4	59	19	64	34	76	49	82	64	77	79	82
5	67	20	65	35	77	50	83	65	80	80	76
6	68	21	74	36	78	51	65	66	81		
7	74	22	81	37	80	52	68	67	59		
8	81	23	66	38	82	53	75	68	64		
9	66	24	69	39	63	54	76	69	70		
10	61	25	71	40	66	55	78	70	80		
11	64	26	71	41	72	56	78	71	85		
12	65	27	72	42	72	57	81	72	70		
13	67	28	73	43	72	58	85	73	71		
14	68	29	75	44	73	59	66	74	72		
15	70	30	77	45	73	60	68	75	72		

Fonte: Osborn JF. *Statistical Exercises in Medical Research.* John Wiley & Sons Inc., 1979. (adaptado).

EXERCÍCIO 1.a: AMOSTRA ALEATÓRIA

a) Sorteie uma amostra aleatória de tamanho 20 utilizando a tabela dos números equiprováveis.

b) Apresente os valores do peso dos indivíduos sorteados.

c) Some os valores e divida pelo tamanho da amostra (número de valores).

d) Este valor é o parâmetro ou a estimativa do peso médio?

EXERCÍCIO 1.b: AMOSTRA SISTEMÁTICA

a) Sorteie uma amostra sistemática de tamanho 20. Indique o intervalo de amostragem e o começo casual sorteado. Indique o número de identificação de cada elemento da amostra.

b) Some os valores e divida pelo tamanho da amostra (número de valores).

c) Compare com o peso médio obtido no exemplo 2. Você esperaria o mesmo resultado? Justifique.

d) Qual dos dois valores você diria que representa melhor o conjunto de dados? Justifique.

2 Variáveis

2.1 O que é variável?

Variáveis são valores que assumem determinadas características dentro de uma pesquisa.

Ou seja, tudo aquilo que varia dentro do seu universo de pesquisa pode se tornar uma variável. Cabe a você pesquisador decidir se aquela variação é relevante para o que você estuda ou não.

Exemplos de variáveis:

Estudos com humanos	Gênero (masculino ou feminino)	Raça
		Idade e faixa etária
		Nível de escolaridade
		Características da alimentação
		Características de atividade física
Estudos com alimentos	Quantificação de nutrientes	Quantificação de compostos
		Atividade antioxidante

Estudos com animais	Quantificação de proteínas Quantificação de citocinas Ingestão calórica/ hídrica Taxa de migração celular

Do ponto de vista estatístico, as variáveis se dividem em dois grandes grupos: variáveis quantitativas e variáveis qualitativas. Essas classificações são importantes pois determinados testes estatísticos são adequados para determinados tipos de variáveis.

Variáveis quantitativas: Expressam valores numéricos. Exemplos: peso (em kg), glicemia (em mg/dL), idade, número de filhos, índice de massa corporal (em kg/m²), quantidade de vitamina A (em RE) etc.

Variáveis qualitativas: Expressam classificações por tipos ou atributos. Exemplos: gênero, profissão, classificação do índice de massa corporal (de desnutrido a obesidade), faixa etária etc.

Além disso, cada tipo de variável se divide em duas categorias cada. As variáveis quantitativas podem ser: contínuas ou discretas. As variáveis qualitativas podem ser: nominais ou ordinais.

Variável quantitativas contínuas: Podem assumir qualquer valor numérico em um determinado intervalo. Geralmente são oriundas de instrumentos de medidas.

Escreva exemplos:

_____
_____
_____
_____
_____

Variável quantitativas discretas: São medidas de contagem. São valores pertencentes ao conjunto enumerável e são números inteiros. Valores discretos com decimais podem ser difíceis de serem interpretados

Escreva exemplos:

_____
_____
_____
_____

_____

Variável qualitativa nominal: São grupos, categorias ou classificações sem ordem definida ou graus de superioridade/ inferioridade

Escreva exemplos:

_____
_____
_____
_____
_____
_____

Variável qualitativa ordinal: São classificações que envolvem ordens de grandeza, sequências ou graus.

Escreva exemplos:

_____
_____
_____
_____
_____
_____

Veja esse artigo do JAMA

Clinical Review & Education

JAMA | Special Communication

Life Expectancy and Mortality Rates in the United States, 1959-2017

Steven H. Woolf, MD, MPH; Heidi Schoomaker, MAEd

IMPORTANCE US life expectancy has not kept pace with that of other wealthy countries and is now decreasing.

OBJECTIVE To examine vital statistics and review the history of changes in US life expectancy and increasing mortality rates; and to identify potential contributing factors, drawing insights from current literature and an analysis of state-level trends.

EVIDENCE Life expectancy data for 1959-2016 and cause-specific mortality rates for 1999-2017 were obtained from the US Mortality Database and CDC WONDER, respectively. The analysis focused on midlife deaths (ages 25-64 years), stratified by sex, race/ethnicity, socioeconomic status, and geography (including the 50 states). Published research from January 1990 through August 2019 that examined relevant mortality trends and potential contributory factors was examined.

FINDINGS Between 1959 and 2016, US life expectancy increased from 69.9 years to 78.9 years but declined for 3 consecutive years after 2014. The recent decrease in US life expectancy culminated a period of increasing cause-specific mortality among adults aged 25 to 64 years that began in the 1990s, ultimately producing an increase in all-cause mortality that began in 2010. During 2010-2017, midlife all-cause mortality rates increased from 328.5 deaths/100 000 to 348.2 deaths/100 000. By 2014, midlife mortality was increasing across all racial groups, caused by drug overdoses, alcohol abuse, suicides, and a diverse list of organ system diseases. The largest relative increases in midlife mortality rates occurred in New England (New Hampshire, 23.3%; Maine, 20.7%; Vermont, 19.9%) and the Ohio Valley (West Virginia, 23.0%; Ohio, 21.6%; Indiana, 14.8%; Kentucky, 14.7%). The increase in midlife mortality during 2010-2017 was associated with an estimated 33 307 excess US deaths, 32.8% of which occurred in 4 Ohio Valley states.

CONCLUSIONS AND RELEVANCE US life expectancy increased for most of the past 60 years, but the rate of increase slowed over time and life expectancy decreased after 2014. A major contributor has been an increase in mortality from specific causes (eg, drug overdoses, suicides, organ system diseases) among young and middle-aged adults of all racial groups, with an onset as early as the 1990s and with the largest relative increases occurring in the Ohio Valley and New England. The implications for public health and the economy are substantial, making it vital to understand the underlying causes.

JAMA. 2019;322(20):1996-2016. doi:10.1001/jama.2019.16932

Editorial page 1963

Supplemental content

CME Quiz at jamanetwork.com/learning

Author Affiliations: Center on Society and Health, Department of Family Medicine and Population Health, Virginia Commonwealth University School of Medicine, Richmond (Woolf); Center on Society and Health, Virginia Commonwealth University School of Medicine, Richmond (Schoomaker); Now with Eastern Virginia Medical School, Norfolk (Schoomaker).

Corresponding Author: Steven H. Woolf, MD, MPH, Center on Society and Health, Department of Family Medicine and Population Health, Virginia Commonwealth University School of Medicine, 830 E Main St, Ste 5035, Richmond, VA 23298-0212 (steven.woolf@vcuhealth.org).

Com base no resumo apresentado, classifique as variáveis em um dos quatro grupos de classificação. As variáveis estão grifadas em amarelo.

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

_____

2.2 Categorização de variáveis

Uma variável única pode assumir diversas classificações, depende da forma como foi coletada. Por exemplo a variável idade:

-Quantitativa contínua: 15,5 anos; 10 anos; 60 anos.

-Qualitativa ordinal: Criança (0 a 10 anos); pré-adolescente (10 a 13); adolescente (14 a 18); adulto jovem(19 a 30); adulto (30 a 59); idoso (60 ou mais).

-Qualitativa nominal: Menor de idade (criança; pré-adolescente; adolescente – até 18 incompleto) e Maior de idade (adulto e idoso – 18 ou mais).

Dica rápida
Sempre que possível colete a variável na sua forma quantitativa (discreta ou contínua). Se você precisar separar em qualitativas posteriormente, é possível. Entretanto, se você coletar de forma qualitativa, não conseguirá retornar a forma quantitativa, perdendo a informação

Categorizar uma variável é criar uma nova variável no seu banco de dados transformando uma variável em uma outra (geralmente qualitativa). Para categorizar uma variável é necessário utilizar pressupostos, quando eles existirem, por exemplo:

-Idade: até 18 anos (menor de idade); acima de 18 anos (maior de idade) -Peso ao nascer: até 2500g (baixo peso); acima de 2500g (peso adequado) -Glicemia: 90 a 110mg/dL (adequada); acima de 110mg/dL (elevada)

Importante verificar na literatura científica se existe alguma classificação, ponto de corte ou valor limite para sua variável de interesse.

Quando não temos nenhuma classificação na literatura podemos:

-Utilizar a mediana (valor central) - Dividindo em dois grupos.

-Dividir em grupos proporcionais como: 3 grupos (tercis); 4 grupos (quartis); 5 grupos (quintis) etc.

Geralmente fazemos a categorização de variáveis quando:

1) Nosso "n" amostral é pequeno para análises com variáveis quantitativas
2) Quanto a classificação é mais interessante do que o valor absoluto 3) Quanto queremos realizar análises estatísticas específicas para variáveis qualitativas (ex: qui-quadrado, regressão logística, regressão ordinal etc.)

Categorize a variável IMC conforme os pressupostos da Organização Mundial da Saúde.

ID IMC (kg/m²) Classificação do IMC Sedentários (1= sim; 0= não)

#1 27,1 1 #2 22,1 0 #3 23,6 0 #4 19,9 0 #5 24,1 1 #6 28,1 1 #7 22,9 0 #8

27,5 0

#9 30,7 1

#10	31,8		1

Quantos % das pessoas sedentárias são eutróficas, estão com sobrepeso e com obesidade?

Classificação do IMC (em kg/m²)

Desnutrição <18,5

Eutrofia: 18,5 a 24,9

Sobrepeso: 25 a 29,9

Obesidade: >30

2.3 Dependência entre variáveis

Variáveis independentes são aquelas que geralmente são manipuladas em um experimento. São aquelas que a gente acredita que podem ter efeitos em uma ou mais variáveis. Geralmente são representadas pela letra x.

Por exemplo: Se um rato ganha uma recompensa (1 biscoito) cada vez que executa uma tarefa (apertar um botão), executar a tarefa é a variável independente.

As variáveis dependentes são aquelas que dependem das variáveis independentes. No exemplo acima seria a recompensa. Geralmente são representadas pela letra y.

Veja o resumo do estudo: Monteiro et al. Causas do declínio da desnutrição infantil no Brasil, 1996-2007. Rev Saude Publica 2009;43(1):35-43.

RESUMO

OBJETIVO: Estabelecer a evolução da prevalência de desnutrição na população brasileira de crianças menores de cinco anos de idade entre 1996 e 2007 e identificar os principais fatores responsáveis por essa evolução.

MÉTODOS: Os dados analisados procedem de inquéritos *"Demographic Health Surveys"* realizados no Brasil em 1996 e 2006/7 em amostras probabilísticas de cerca de 4 mil crianças menores de cinco anos. A identificação dos fatores responsáveis pela variação temporal da prevalência da desnutrição (altura-para-idade inferior a -2 escores z; padrão OMS 2006) considerou mudanças na distribuição de quatro determinantes potenciais do estado nutricional. Modelagem estatística da associação independente entre determinante e risco de desnutrição em cada inquérito e cálculo de "frações atribuíveis parciais" foram utilizados para avaliar a importância relativa de cada fator na evolução da desnutrição infantil.

RESULTADOS: A prevalência da desnutrição foi reduzida em cerca de 50%: de 13,5% (IC 95%: 12,1%;14,8%) em 1996 para 6,8% (5,4%;8,3%) em 2006/7. Dois terços dessa redução poderiam ser atribuídos à evolução favorável dos quatro fatores estudados: 25,7% ao aumento da escolaridade materna; 21,7% ao crescimento do poder aquisitivo das famílias; 11,6% à expansão da assistência à saúde e 4,3% à melhoria nas condições de saneamento.

CONCLUSÕES: A taxa anual de declínio de 6,3% na proporção de crianças com déficits de altura-para-idade indica que em cerca de mais dez anos a desnutrição infantil poderia deixar de ser um problema de saúde pública no Brasil. A conquista desse resultado dependerá da manutenção das políticas econômicas e sociais que têm favorecido o aumento do poder aquisitivo dos mais pobres e de investimentos públicos que permitam completar a universalização do acesso da população brasileira aos serviços essenciais de educação, saúde e saneamento.

Com base nos dados apresentados no resumo, qual é a variável dependente e quais são as independentes?

_____

Dica rápida
A dependência de variáveis deve sempre ser baseadas em pressupostos biológicos! Ou seja, antes de elaborar modelos e testes estatísticos faça buscas na literatura científica para verificar se seus pressupostos fazem sentido.

Atividade 2

Exercício 2.a - Classifique quanto a natureza (qualitativa nominal ou ordinal; quantitativa contínua ou discreta) as seguintes variáveis:

Condição de saúde (doente, não doente):_____

Tipo de parto (normal, cesáreo) :_____

Glicose Sanguínea (mg/dL) :_____

Tempo de um procedimento cirúrgico (minutos)

:_____ Número de praias consideradas

poluídas:_____

Custo do procedimento (reais) :_____

Peso (g) :_____

Estado nutricional (desnutrição, eutrofia, sobrepeso, obesidade)

:_____ Consumo de energia (Kcal) :_____

Realização da refeição café da manhã (sim/não)

:_____ Número de escolares por

série:_____

Classificação de atividade física (sedentário, ativo, muito

ativo):_____ Realização de atividade física diária (sim/não)

:_____ Tempo assistindo TV/dia (< 2h, 2 a 4h, >4h)

:_____ Porções consumidas por grupo de

alimentos:_____ Percentual de gordura corporal

(%):_____

Exercício 2.b - Considerando o banco de dados abaixo complete as lacunas:

	Peso (g)	mais alto)	Glicose_classific
Animal Treinament o (sim/ não)	Peso_duas_categor ias (mais baixo/	Glicose_sanguín ea (mg/dL)	acao
1 13 100			
2 15 99			
3 18 98			
4 19 115			
5 20 117			
6 19 111			
7 21 120			

8 13 99

9 16 98

10 15 90

11 14 97

12		22		100	

Os animais 2,7,8,9,10,11 fizeram treinamento na esteira diariamente por

20min. Ponto de corte da glicose sanguínea: 110 mg/dL.

a) responda: a média de peso do grupo treinamento (sim) é maior do que o grupo treinamento (não)

b) Responda: Qual o n (%) de cada uma das categorias após cruzamentos das variáveis Treinamento e Peso_duas_categorias

c) Responda: Qual o n (%) de cada uma das categorias após cruzamentos das variáveis Treinamento e Glicose_classificacao

d) Defina possíveis variáveis dependentes e independentes do estudo

3 Medidas de tendência central e dispersão

Medidas de tendência central são medidas de um valor que possa melhor representar a tendência de um conjunto de números ou uma variável.

Mais usadas = Média, Mediana e Moda

3.1 Média aritmética

Se dá pela soma de todos os valores de X divididos pelo n (tamanho da amostra).

= média de X de uma amostra (usamos μ para representar a média de uma população)

$\sum$= Somatória

x_i = observações da variável X aleatória

n = tamanho da amostra

$$\bar{x} = \frac{\sum x_i}{n}$$

Observações importantes:

• só existe para variáveis quantitativas e seu valor é único;

• é da mesma natureza da variável considerada; e

• sofre influência dos valores aberrantes.

3.2 Mediana

É o valor que ocupa a posição central da distribuição.

Primeiramente se ordena de forma crescente todos os valores. Se o conjunto tem número ímpar, é o valor que fica exatamente no meio. Se o conjunto tem número par, é a média dos dois valores centrais.

Exemplo para número de n par(n= 10):

Pesos (kg): 10,5 / 12,4 / 9,8 / 11,6 / 8,9 / 13,9 / 10,2 / 9,7 / 14,1 /

10,8 • Ordenados: 8,9 / 9,7 / 9,8 / 10,2 / 10,5 / 10,8 / 11,6 / 12,4

/13,9 / 14,1 Como são 10 números, os valores centrais são o 5º e

o 6º número.

• Mediana: (10,5 + 10,8) / 2 = 10,65 kg

Exemplo para número de n ímpar (n= 9)

Pesos (kg): 10,5 / 12,4 / 9,8 / 11,6 / 8,9 / 13,9 / 10,2 / 9,7 / 10,8 /

• Ordenados: 8,9 / 9,7 / 9,8 / 10,2 / 10,5 / 10,8 / 11,6 / 12,4 /13,9

Mediana = 10,5 (5º valor)

Descreva aqui quando uma média não é adequada:

Descreva aqui quando uma mediana não é adequada:

3.3 Moda

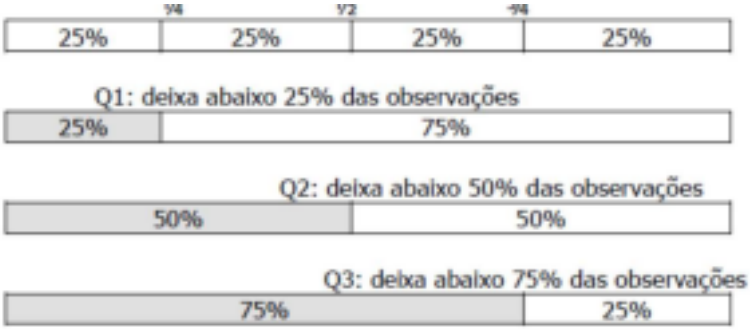
É o valor mais frequente na sua variável X.

Exemplo para número de filhos: 1 / 1 / 2 / 1 / 2 / 3 / 4 / 1 / 1 / 0 / 0 /

0 / 0 A moda para essa variável é 1 filho.

3.4 Quartis

Valores da variável que dividem a distribuição em quatro partes iguais.

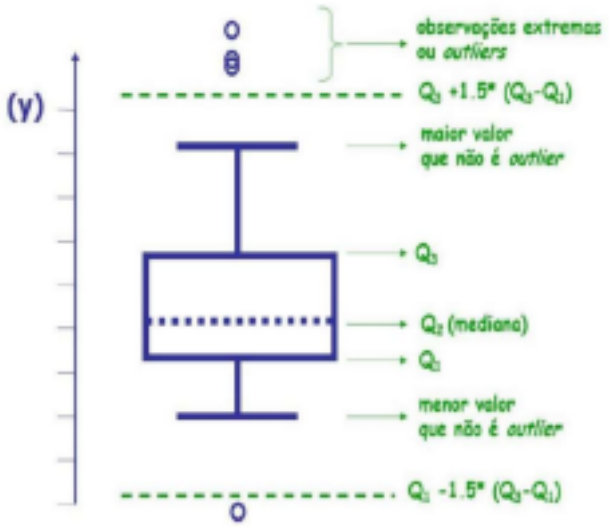


$$Q_1 = x_{(\frac{1}{4}(n+1))} \quad \text{e} \quad Q_3 = x_{(\frac{3}{4}(n+1))}$$

onde x é o valor da variável e $(\frac{1}{4}(n+1))$ e $(\frac{3}{4}(n+1))$ são índices que representam as posições ocupadas por x.

Boxplot

Representa graficamente dados de forma resumida em um retângulo onde as linhas da base e do topo são o Q1 e Q3, respectivamente. A linha entre estas é a mediana.



Valor adjacente superior é o maior valor das observações que é menor ou igual a Q3+1,5(Q3-Q1).

valor adjacente inferior é definido como o menor valor que é maior ou igual a Q1-1,5(Q3-Q1).

A diferença Q3-Q1 denominada intervalo inter-quartil (IIQ).

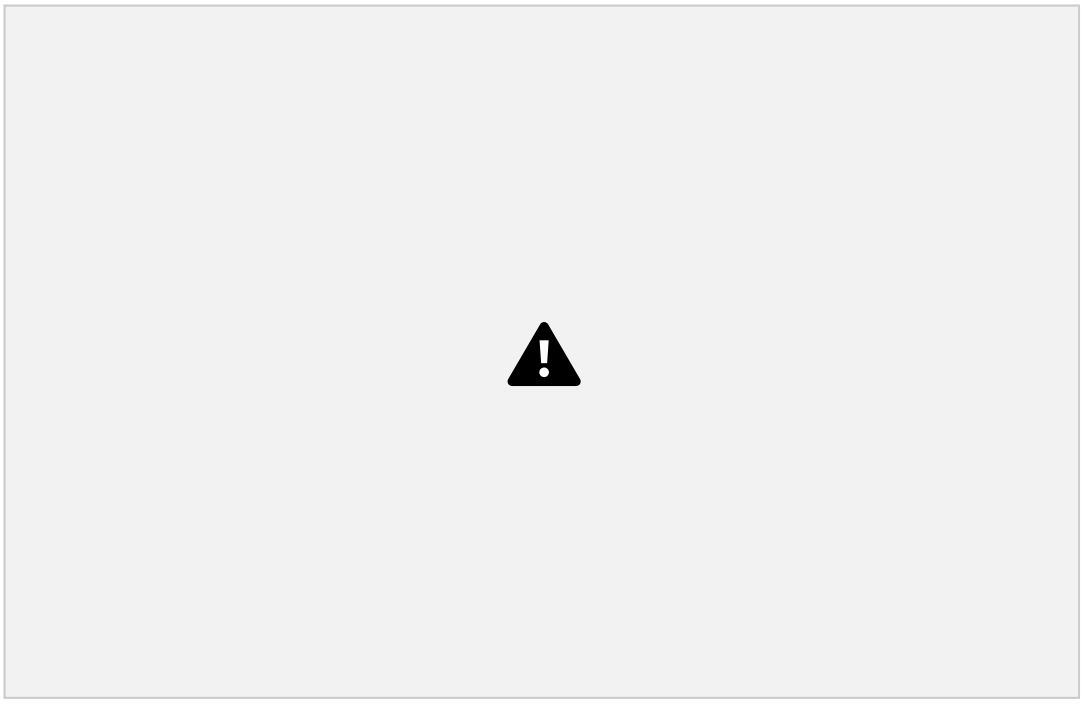
3.5 Distribuição em Centis

Quintil -> Valores da variável que dividem a distribuição em cinco partes iguais

Decil -> Valores da variável que dividem a distribuição em dez partes iguais

Percentil -> Valores da variável que dividem a distribuição em cem partes iguais

Exemplo de Percentil:



3.4 Medidas de dispersão

3.4.1 Amplitude

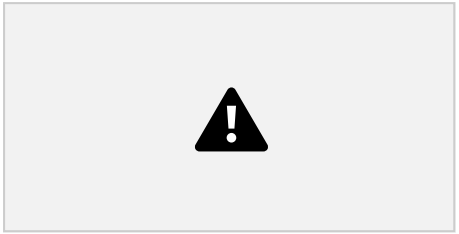
É a diferença entre os valores mínimo e máximo da distribuição

Ex: Peso: 14,1 -8,9 = 5,2

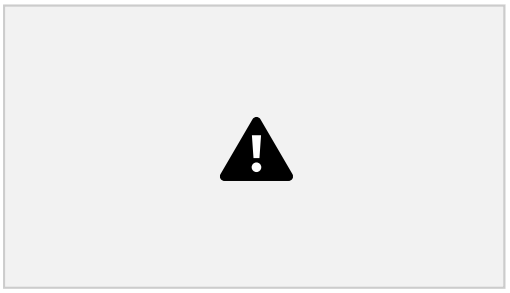
Boa medida para conjuntos pequenos.

3.4.2 Variância e desvio padrão

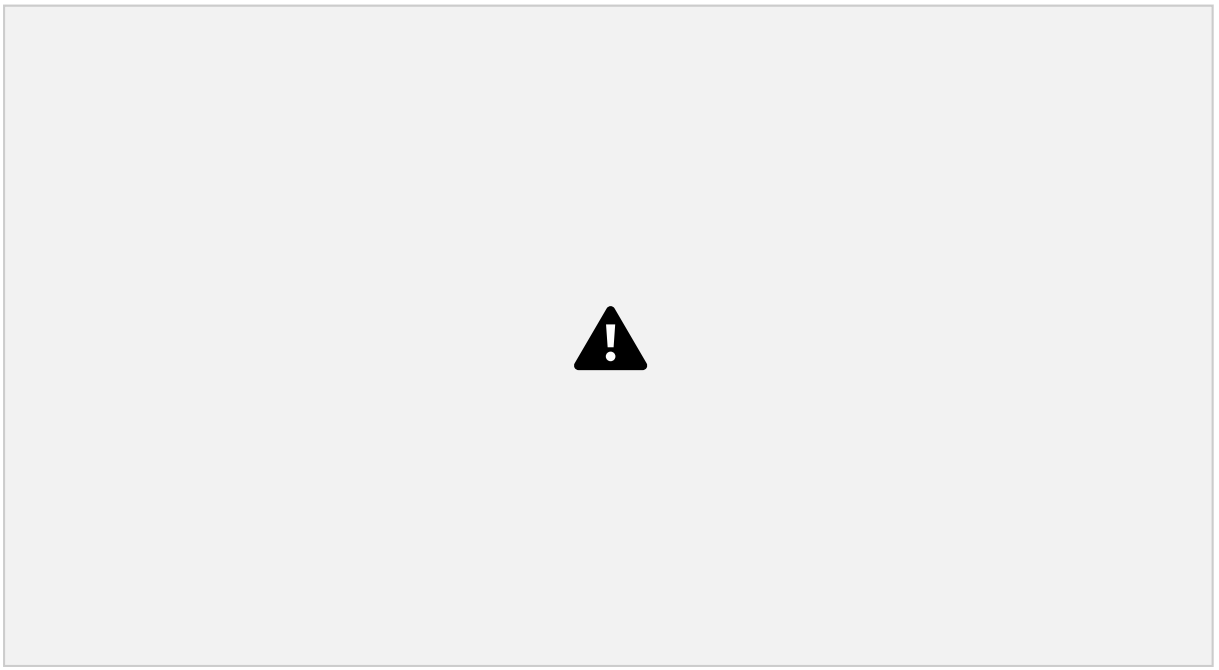
É a medida dos quadrados dos desvios em relação a média. Distância de cada valor da distribuição da média. Geralmente representada pela letra S^2 ou σ (sigma).



Já o desvio padrão é raiz quadrada da variância



Exemplo de cálculo do desvio padrão



Atividade 3

Exercício 3.a: Os dados a seguir são provenientes do grupo Western Collaborative Group Study, sobre níveis de colesterol de acordo com o comportamento de cada grupo. O Grupo tipo A era formado por pessoas caracterizadas pela urgência, agressividade e ambição. Os participantes de tipo B são mais relaxados, não competitivos e menos preocupados.

a) Calcule a média, mediana, desvio padrão e variância geral do estudo e para cada grupo. Comente as diferenças.

Exercício 3.b: Os dados a seguir são provenientes de um estudo que avalia o crescimento de crianças de 7 a 10 anos de uma escola pública do município de São Paulo no ano de 2008. Os dados apresentados são de 16 meninos e 16 meninas para os quais foram aferidos a circunferência do braço (CB) (cm):

a) Calcule a média, mediana, desvio padrão e variância geral do estudo e para cada grupo. Comente as diferenças.

b) Desenhe o box plot da circunferência braquial (cm) segundo sexo e comente o gráfico quanto a dispersão dos dados, existência de valores aberrantes e igualdade de medianas.

4 Apresentação de dados

A forma de se apresentar os dados deve ser pensada estrategicamente. Caso a apresentação de bons dados seja feita de maneira inadequada, a compreensão e confiança no resultado fica seriamente comprometida.

Após a apuração, há necessidade de os dados e os resultados serem dispostos de forma ordenada e resumida. Tal procedimento ajuda o pesquisador e os leitores a compreender os resultados. Geralmente os resultados numéricos são apresentados por meio de tabelas ou gráficos.

4.1 Apresentação tabular

Qualquer tabela deve sempre ser auto-suficiente, ou seja, deve ter significado próprio sem ser necessário consultas ao texto para ser compreendida. Nesse sentido

sugere se que uma tabela tenha

Elementos essenciais

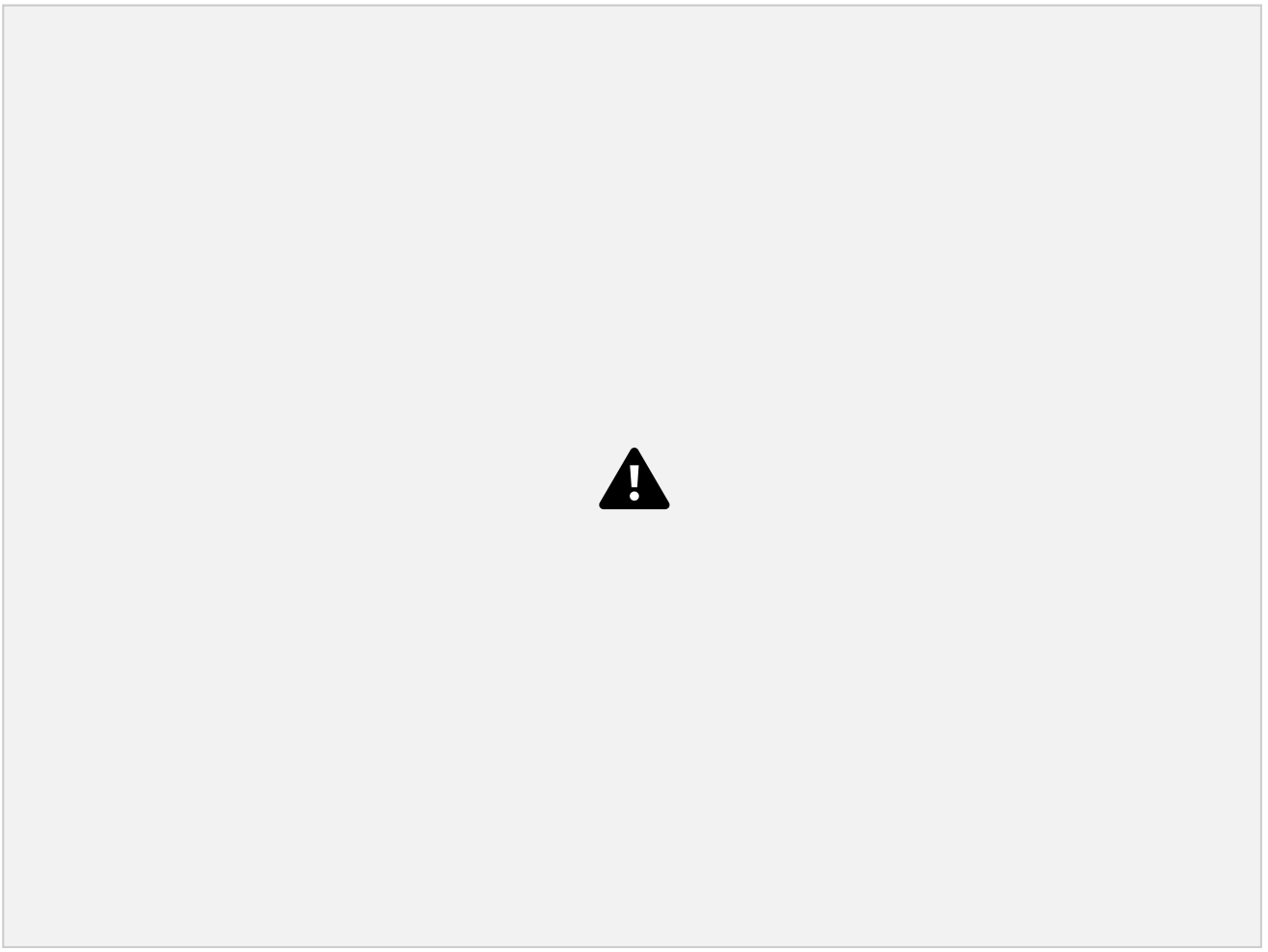
- Título - É a indicação que é colocada na parte superior da tabela. Deve ser preciso, claro e conciso, indicando o fato estudado (O quê?), as variáveis escolhidas na análise do fato (como?), o local (onde?), e a época (quando?).
- Corpo da tabela - Conjunto de linhas e colunas que contêm as informações. Casa, cela ou célula é o cruzamento de uma linha com uma coluna onde se tem o valor aferido.
- Cabeçalho - É uma parte da tabela que é designada a natureza do conteúdo de cada coluna
- Coluna indicadora - É a uma parte da tabela que é designada a natureza do conteúdo de cada linha.
- Linhas superior e inferior (fechadas no alto e embaixo). Não devem ser fechadas à direita e à esquerda por linhas verticais. É facultativo o emprego de traços verticais para separação de colunas no corpo da tabela, bem como de linhas.

Elementos complementares

- Fonte: É o indicativo, no rodapé da tabela, da entidade responsável pela organização ou fornecedora dos dados primários. Geralmente utilizada quando os dados do estudo são oriundo de base de dados e não dados observados
- Notas: Notas colocadas no rodapé da tabela para esclarecimentos de ordem geral. São numeradas, podendo-se também usar símbolos gráficos.
- Chamadas: Também colocadas no rodapé, servem para esclarecer minúcias em relação às casas, colunas ou linhas.

Nenhuma casa da tabela deve ficar em branco, apresentando sempre um número ou sinal.

Exemplo:



Geralmente se utilizam símbolos para identificar as notas. O mais comum é utilizar a seguinte ordem de símbolos: *, †, ‡, §, ¶, #. Entretanto, sempre verifique as normas da revista em que irá submeter o artigo.

Outras considerações:

- Mantenha sempre a uniformidade quanto ao número de casas decimais;
- Toda tabela deve ser apresentada no texto. Ou seja, inclua uma frase dizendo "A tabela 1 apresenta os dados X, Y, Z.";

4.2 Apresentação gráfica

Os dados também podem ser apresentados de forma de figuras, geralmente gráficos ou diagramas.

Assim como as tabelas o gráfico deve ser auto-explicativo e de fácil compreensão, de preferência sem comentários inseridos. Devem ser utilizados para atrair a atenção do leitor e destacar valores.

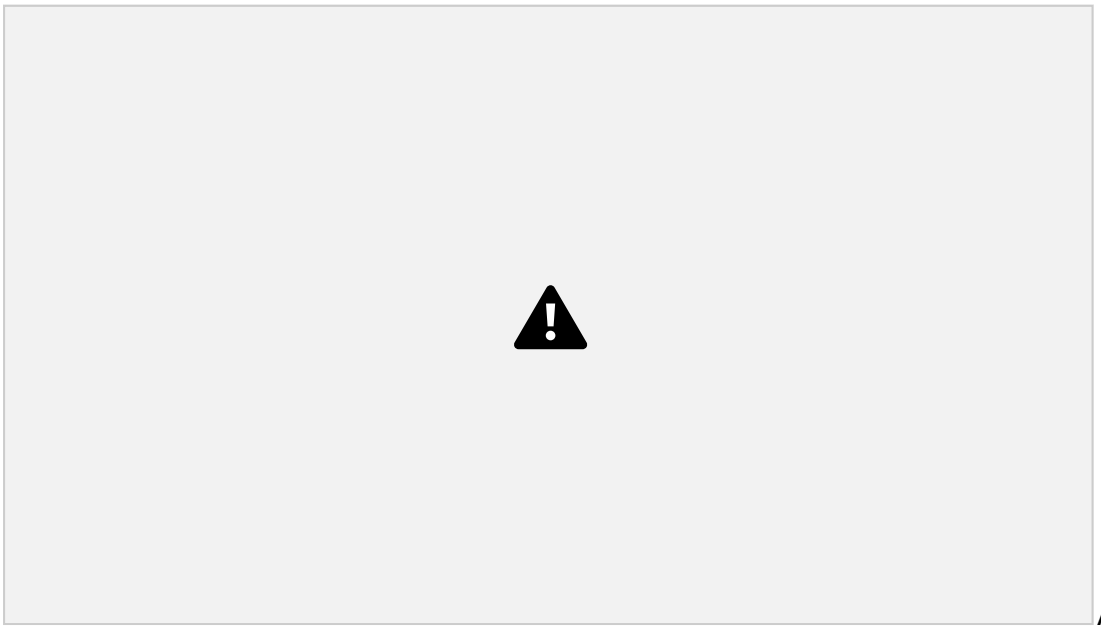
Elementos essenciais

Título - É a indicação que é colocada junto a figura (acima ou abaixo dependendo da formatação). Deve ser preciso, claro e conciso, indicando o fato estudado (O quê?), as variáveis escolhidas na análise do fato (como?), o local (onde?), e a época (quando?).

Escala - Deve sempre indicar de alguma forma a escala que foi utilizada para construção da figura de forma a não desfigurar o os fatos observados.

Exemplos de gráficos

Cartograma



Apostila de Bioestatística – Prof. Diogo Cunha e Profa. Ligiana Corona Página 34

Histograma

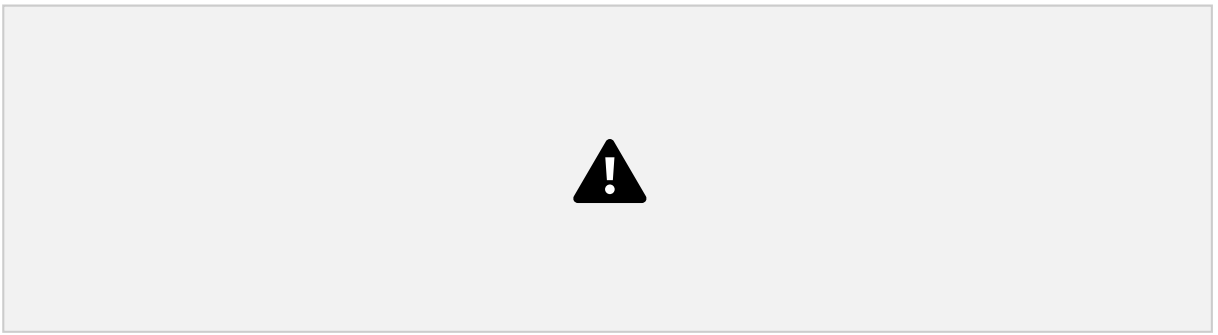


Gráfico de barras (variável quantitativa x ordinal)

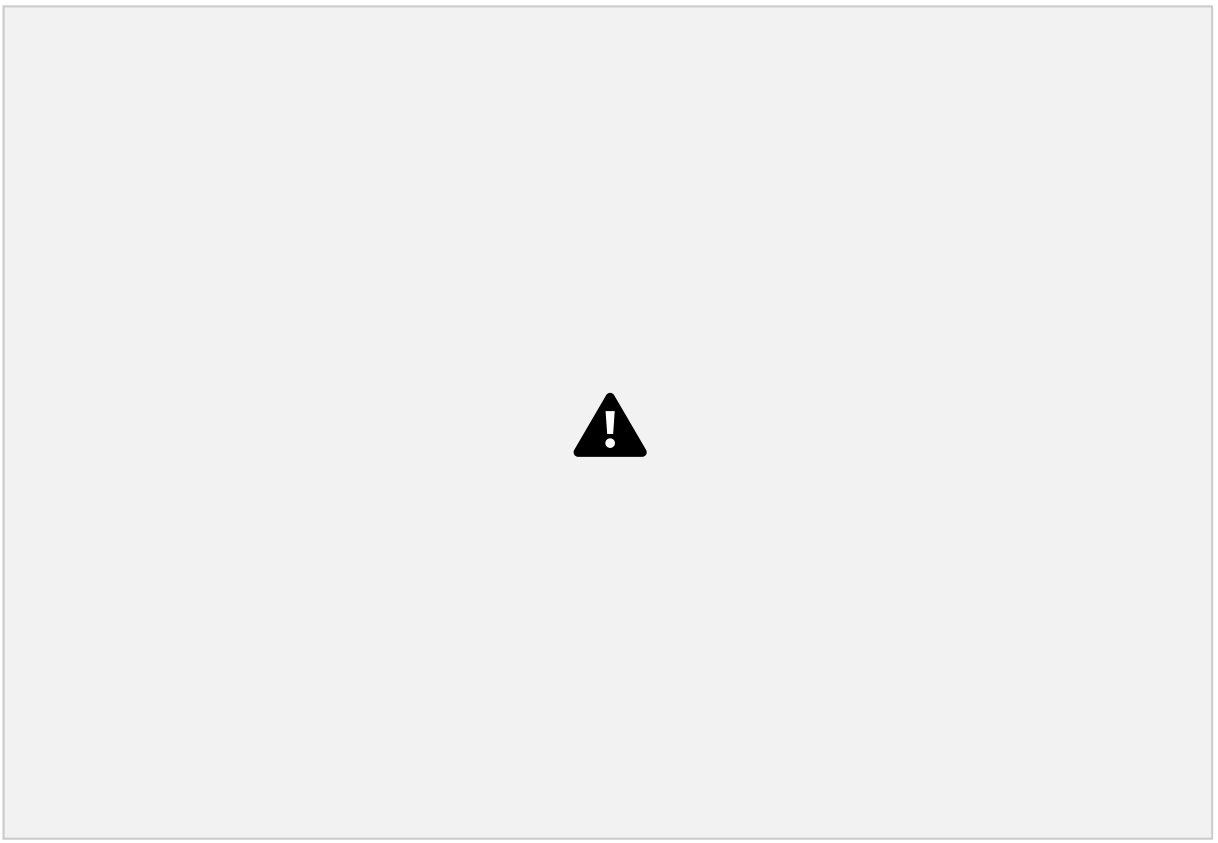


Gráfico de barras (variável quantitativa x nominal)

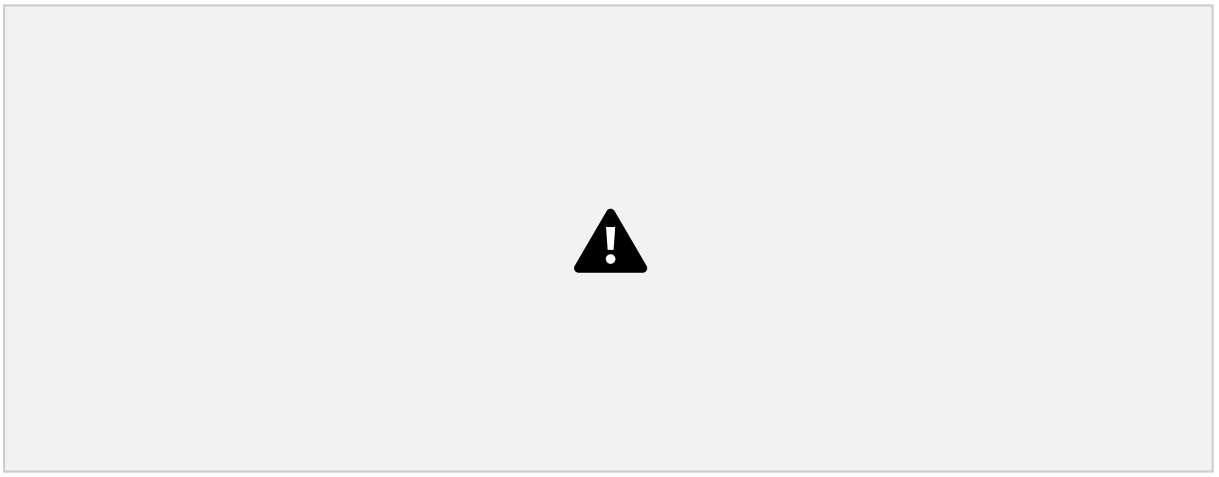
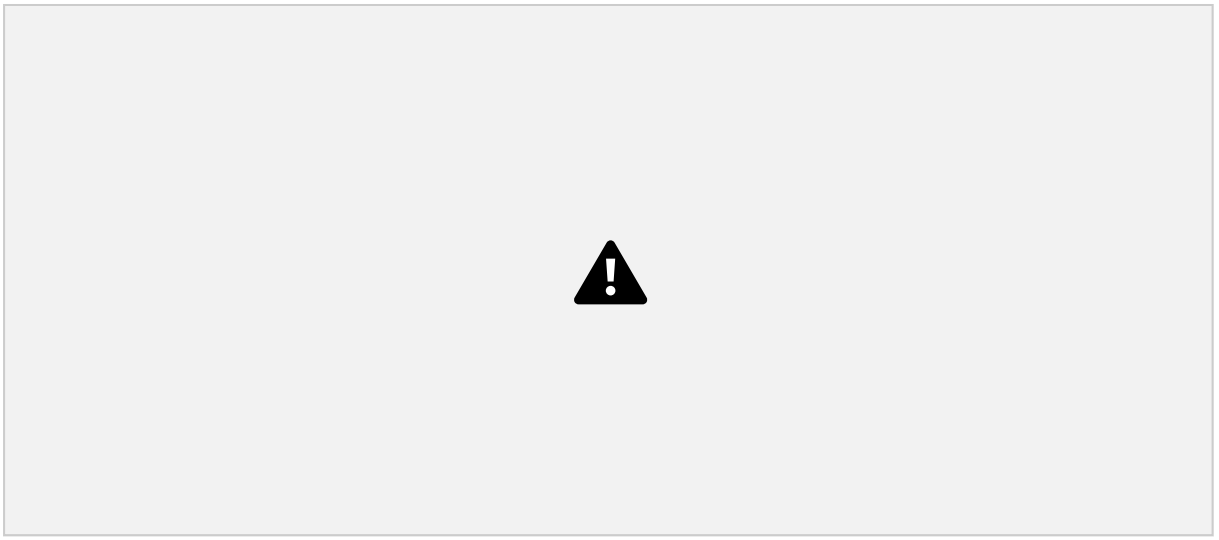


Gráfico de dispersão (variável quantitativa x quantitativa)



Boxplot

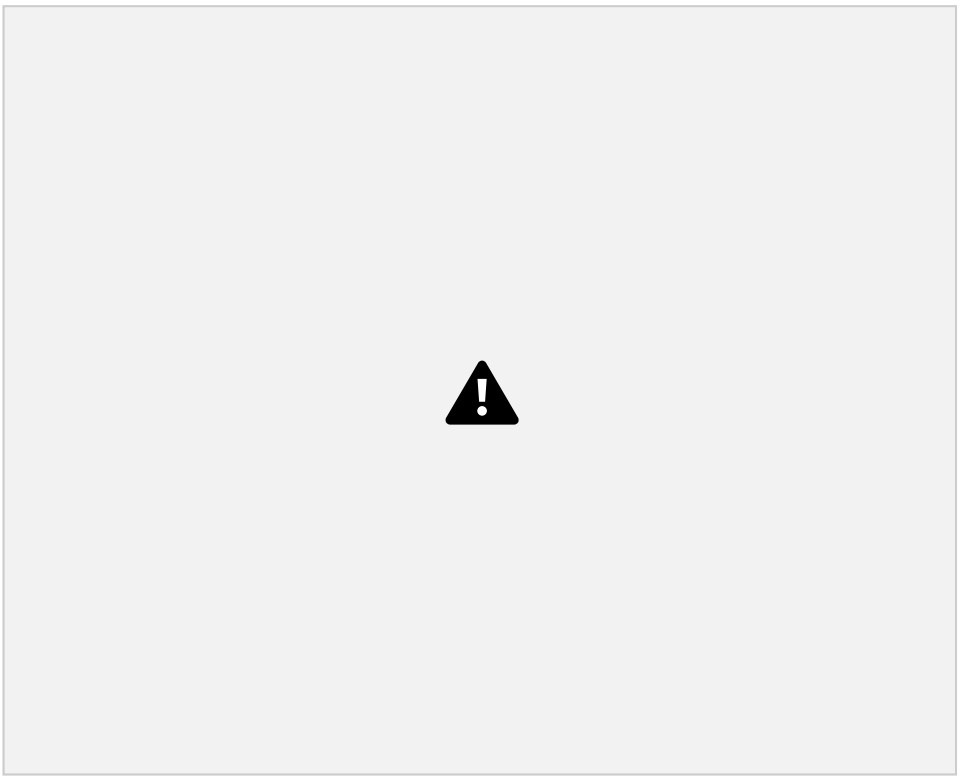
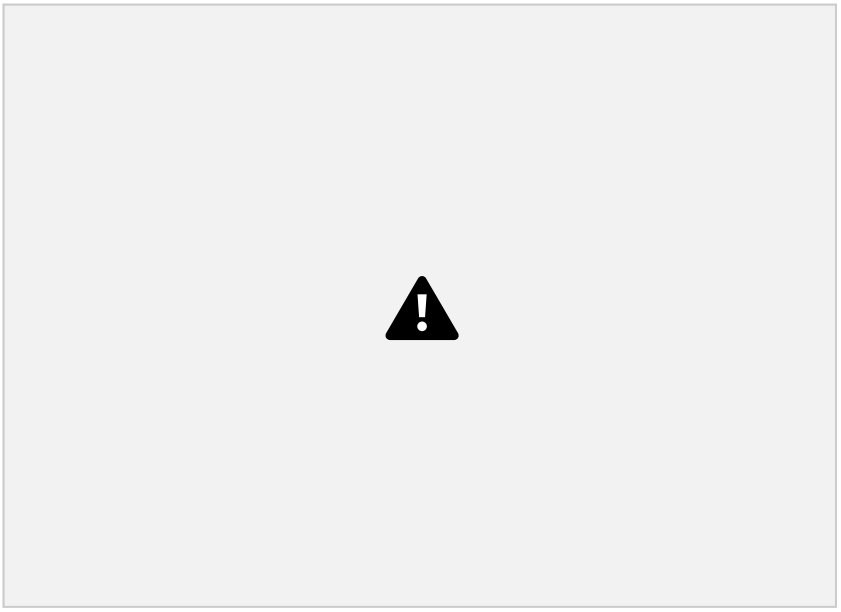


Diagrama circular



Atividade 4

Exercício 4.a - Construa um histograma de distribuição considerando a idade de alunos ingressantes no curso de música da Unicamp em 2017.

17	19	18	19	19	19	20

4.b Esse histograma tem tendência a distribuição normal? Discuta

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

4.c Considerando os dados abaixo, monte dois gráficos de dispersão (consumo de sertralina x BECK) e (consumo de sertalina x Kcal). Discuta a relação entre as variáveis.

Consumo de antidepressivo (Sertralina	mg/dia)	Escore de depressão BECK	Consumo de Kcal/ dia
50	51	1800	
250	15	1150	
120	27	1750	
150	39	1850	
350	7	930	
190	35	1750	
400	0	910	
100	47	1880	
300	8	980	
200	22	1450	
390	0	900	
180	38	1780	
220	21	1200	

5 Análise inferencial de dados

Retomando:

INFERÊNCIA: Generalizar afirmações sobre determinada população, baseadas em dados retirados de uma amostra.

HIPÓTESE CONCEITUAL: É uma forma de especulação relativa a um fenômeno estudado (qualquer que seja), escrita em relação ao objeto do estudo.

HIPÓTESE ESTATÍSTICA: É uma especulação feita em relação a uma proposição, porém relativa à uma população definida, escrita em relação ao teste utilizado.

5.1 Teste de hipóteses, tipos de erros e valor p

Neyman e Pearson propuseram uma abordagem para a tomada de decisão que envolve a fixação, ANTES da realização do experimento, das hipóteses nula e alternativa, e fixação de valores de probabilidade de ocorrência de erros de decisão, ou seja, o quanto eu aceito errar na estimativa de efeito do meu experimento.

Exemplos de hipóteses:

- Dados contínuos: O peso médio dos alunos desta sala é representa (é igual) ao peso médio da população de estudantes de São Paulo?
- Dados categóricos: Ex: A proporção de obesos desta sala representa (é igual) a proporção de obesos na população de estudantes de São Paulo?

As hipóteses estatísticas sempre partem da hipótese nula, mas a hipótese alternativa é onde se observa a diferença. É exatamente o que é feito em processos criminais, onde um acusado é inocente até que se prove o contrário (a pressuposição de inocência é uma hipótese nula).

● Hipótese nula (H0): hipótese tida como verdadeira. Um parâmetro da população é igual a um valor hipotético (não há diferença estatística entre os grupos testados) ● Hipótese alternativa (H1): hipótese que se apresenta como verdadeira, refutando a hipótese nula. Um parâmetro da população é diferente de um valor hipotético (há diferença estatística entre os grupos)

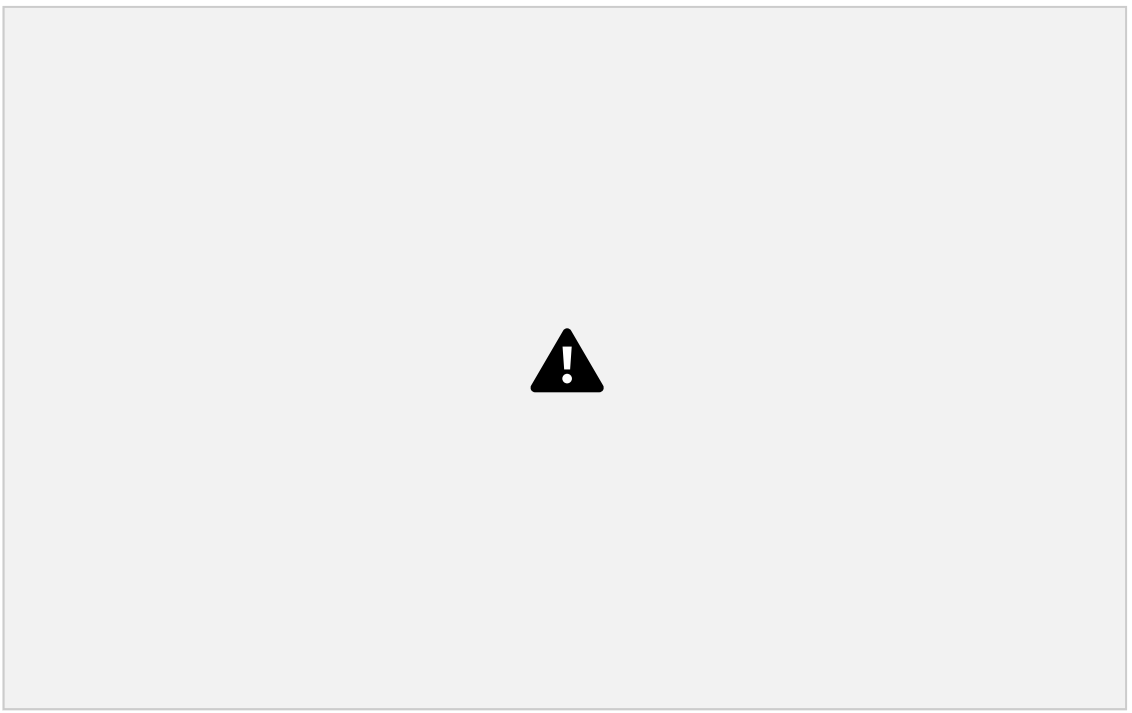
Por exemplo, seja a seguinte pergunta de pesquisa: Qual o efeito da cafeína sobre parâmetros inflamatórios de animais treinados e não treinados?

● Hipótese conceitual: Há diferença nos parâmetros inflamatórios entre os animais quanto a treinamento e consumo de cafeína

● Hipótese estatística nula (H0): Não há diferença entre as médias dos parâmetros inflamatórios quanto a treinamento e consumo de cafeína (ou somente $\mu_1 = \mu_2$)

● Hipótese estatística alternativa (H1): Há diferença entre as médias dos parâmetros inflamatórios quanto a treinamento e consumo de cafeína (ou somente $\mu_1 \neq \mu_2$)

O teste de hipóteses pode ser realizado de maneira unicaudal ou bicaudal, ou seja, o pesquisador pode investigar se há qualquer diferença entre os grupos, seja para valores maiores ou menores (bicaudal), ou determinar previamente uma direção específica da diferença ele deseja testar. Veja os exemplos abaixo:



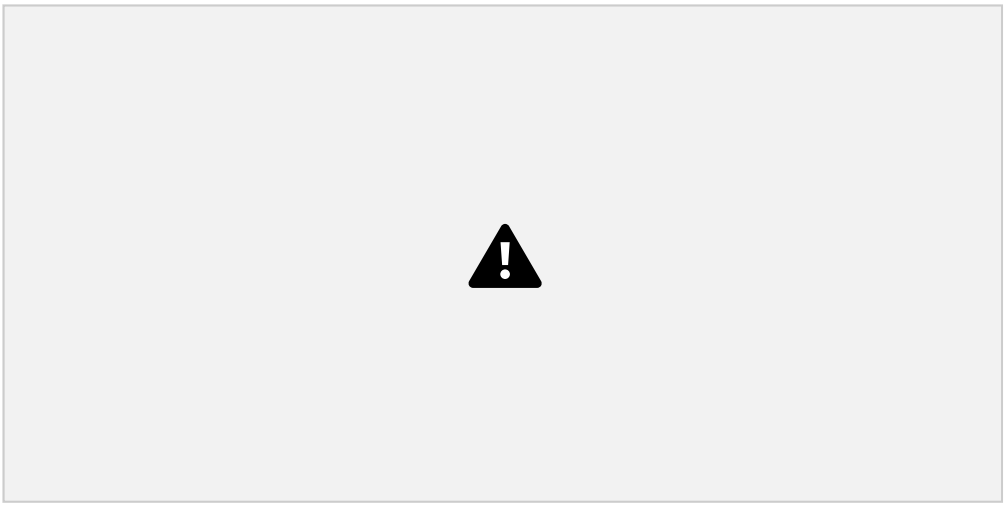
Para a realização do teste de hipótese, é necessário retomar o conceito de erros tipo I e tipo II, já mencionados na parte 2 desse material (conceitos básicos). Esses erros refletem a tolerância do pesquisador à detecção de diferenças entre os subgrupos quando elas não existem (erro tipo I ou erro α) ou da falha na detecção de diferenças quando elas realmente existem (erro tipo II ou erro β).

Assim, o erro α é a probabilidade de rejeitar H0 e H0 é verdade. A convenção em estatística clássica é dizer que ela deve ser de 0,05 (5%). Por exemplo: imagine um teste estatístico comparando os resultados de duas drogas que sejam semelhantes. Esse teste tem uma chance de 5% de dizer que uma é melhor do que a outra apesar de as duas serem semelhantes, ou seja, aceita-se 5% de chance que está errado ao rejeitar a H0.

O erro β então é a probabilidade de aceitar H0 se H0 é falsa. A convenção é que essa probabilidade seja estabelecida em 20% (0,20). Nesse caso, poder estatístico é, simplesmente, 1 menos β, ou seja: 1 - 0,2 = 0,8 = 80%.

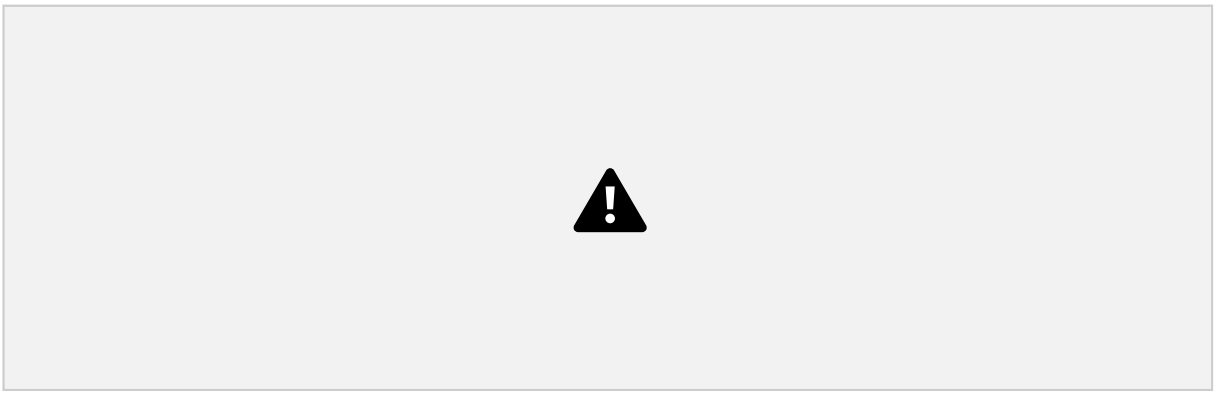
Esses valores não são rígidos, eles devem ser estabelecidos a priori pelo pesquisador em níveis diferentes dos comumente utilizados, mas valores muito discrepantes podem dificultar a publicação dos resultados pela dificuldade em

comparação com a literatura.



O poder do teste é uma outra maneira de revelar a falsidade de H0 quando a verdade é H1. Em geral pelo menos 80%. Mas atenção, lembre-se que o tamanho de amostra depende de α e β. Quanto menor o valor de α e β (e portanto maior o poder estatístico), maior o tamanho de amostra necessária para demonstrar uma associação estatisticamente significativa (ou rejeitar a hipótese nula). Amostras pequenas costumam gerar aumento no erro tipo II.

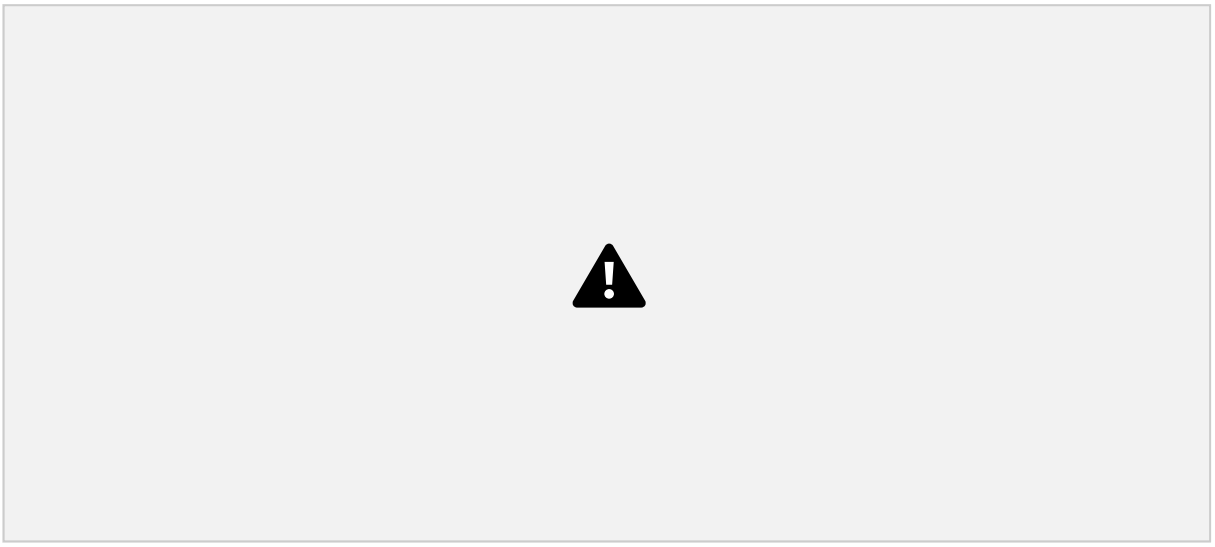
A situação ideal é aquela em que ambas as probabilidades, α e β, são próximas de zero. No entanto, podemos ver na figura abaixo que a medida que diminuímos α, β aumenta:



Fonte: Portal Action

O valor de p, também denominado nível descritivo do teste, é a probabilidade de que a estatística do teste tenha valor extremo (maior ou menor) em relação ao valor observado (estatística) quando a hipótese H0 é verdadeira. Tradicionalmente, o valor de corte para rejeitar a hipótese nula é de 0,05, o que significa que, quando não há nenhuma diferença, um valor tão extremo para a estatística de teste é esperado em menos de 5% das vezes (Ferreira e Patino, 2015). Ou seja, é a probabilidade do resultado observado ter ocorrido “ao acaso” considerando que a

hipótese nula é verdadeira.



Fonte: Portal Action

Ex: Medida de Glicemia de diabéticos X não diabéticos (considerando

p=0,001) H0 : Glicemia diabéticos = Glicemia não diabéticos

Se eu repetir mil vezes o teste em um deles indicará que a glicemia é igual

O uso do valor p exige alguns cuidados, conforme citado por Ferreira e Patino (2015): considerar o valor p mais importante que a relevância clínica da diferença; acreditar que, se o valor-p está próximo de 5%, há uma tendência de haver uma diferença entre os grupos; o valor p não é necessariamente o valor mais importante a ser relatado - outras medidas (como o IC95%) podem mostrar mais a confiabilidade da estimativa que o valor do p isolado.

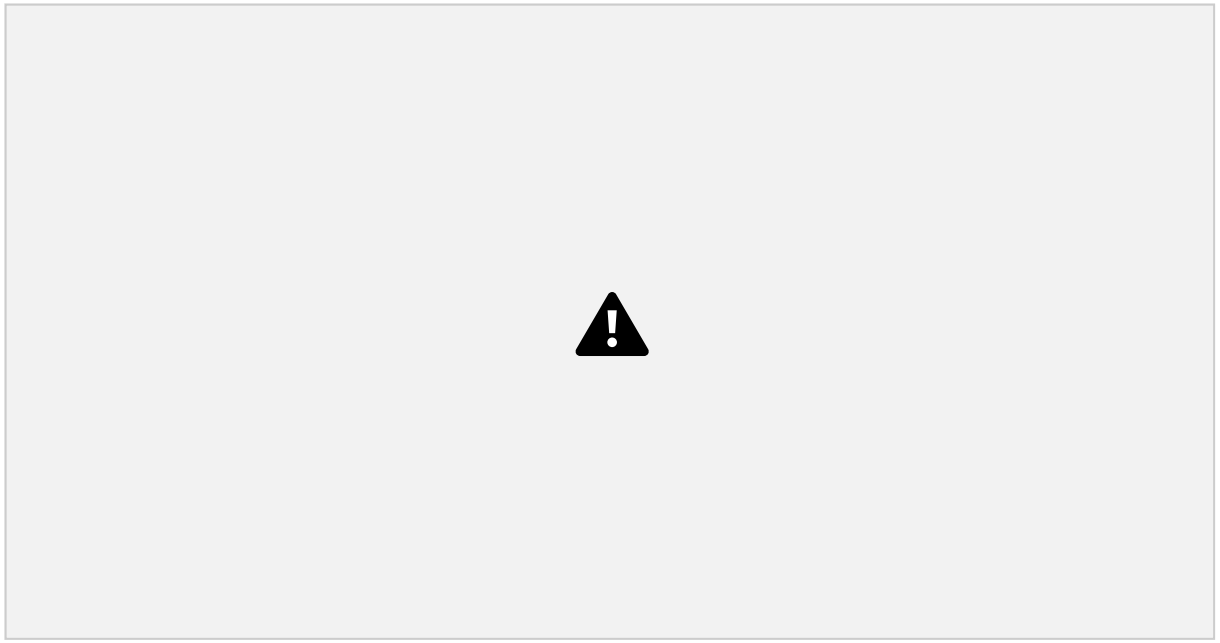
Para a tomada de decisão em estatística e para realizar a inferência corretamente são necessários vários passos:

• Formular as hipóteses conceitual e estatística;
• Identificar a distribuição de probabilidade da estatística do teste;
• Fixar o nível de significância do teste (a) e calcular o tamanho da amostra; • Determinar a região de rejeição/aceitação de H0;
• Realizar o estudo, observar os resultados, calcular a estatística do teste; • Tomar a decisão e apresentar a conclusão.

5.2 Intervalo de Confiança

O intervalo de confiança(IC) é um intervalo estimado de um parâmetro de interesse da população. Em vez de estimar o parâmetro por um único valor, é dado um intervalo de estimativas prováveis. É usado para indicar a confiabilidade de uma estimativa (Sendo todas as estimativas iguais, uma pesquisa que resulte num IC

pequeno é mais confiável do que uma que resulte num IC maior). O valor comumente adotado para um IC é de 95%, ou seja, podemos esperar que, se repetirmos o experimento inúmeras vezes, em 95% delas o verdadeiro valor se encontrará dentro intervalo estimado.



Assim, podemos esperar que aproximadamente 95% destes intervalos devem conter o verdadeiro valor da média populacional (Fonte: Portal Action).

Exemplo (Fonte: Portal Action): O projetista de uma indústria tomou uma amostra de 36 funcionários para verificar o tempo médio gasto para montar um determinado brinquedo. Foram encontrados os dados: Média: 19,9 e DP: 5,73 minutos.

Para um α = 0,05, portanto, (α/2)=0,05/2=0,025 ⇒ Na tabela da distribuição normal P = 0,50 – 0,025 = 0,475 (��α/2=1,96)

Ou seja: IC(95%)= 18,02; 21,77 ⇒ De 100 vezes que fizermos a média dos funcionários desta empresa, em 95 delas a média estará entre 18,02 e 21,77 minutos.

Exercicio 5

5.a) Em um julgamento jurídico o júri tem que decidir sobre a culpa ou inocência de um réu. Considere dois fatos: 1) o sistema jurídico admite que toda pessoa é inocente até que se prove o contrário. 2) só vai a julgamento pessoas sobre as quais existe dúvida de sua inocência. Fazendo uma analogia com

teste de hipóteses, responda:

a) Apresente as hipóteses nula e alternativa sobre a culpa ou inocência do réu.

b) O júri pode errar se decidir que o réu é culpado quando na verdade ele é inocente. Qual é o outro erro de decisão que o júri pode cometer?

c) Qual dos dois erros é o mais sério?

d) Na terminologia de teste de hipótese, qual tipo de erro (I ou II) pode-se vincular a cada uma das decisões do item b?

2) Calcule o intervalo de confiança (95%) considerando os dados abaixo:

5.b) Construa um intervalo de 95% de confiança para estimar a pressão diastólica média populacional (μ), sabendo que em uma amostra de 36 adultos a pressão média amostral (x) foi igual a85mmHg e o desvio padrão populacional (σ) foi 9 mm de Hg. Interprete o significado desse intervalo.

5.c) Uma amostra de 25 adolescentes meninos apresenta peso médio de 56 kg e desvio padrão 8 kg. Encontre o intervalo de confiança de 95% para o peso médio da população da qual esta amostra foi sorteada e interprete o intervalo de confiança encontrado.

5.d) Os dados abaixo mostram o tempo médio (em minutos) que uma amostra de corredores levou para correr 5 kilometros. Com os dados indique o intervalo de confiança e se a distribuição tem tendência a normalidade analisando média, mediana, desvio padrão e histograma (construir com escala de 1 minuto).

n minutos n minutos 1 14 13 17,6 2 14,2 14 17,8 3 15,4 15 17,9 4 15,4 16 17,9 5 15,5 17 18,2 6 16 18 18,3 7 16,5 19 18,3 8 16,7 20 18,7 9 16,8 21 19,1 10 16,9 22 19,5 11 17,5 23 19,6 Já ordenamos pra vocês! =)

12	17,6

24	20,1

6 Distribuições de probabilidade

A probabilidade é o ato de atribuirmos pesos aos eventos. Entretanto, para que cada um não defina probabilidade de sua forma, esta função deve algumas propriedades. Quando lançamos uma moeda não hesitamos em associar probabilidade 1/2 para o evento "cara" e também 1/2 para o evento "coroa". Da mesma forma, quando lançamos uma moeda n vezes todos os possíveis resultados deste experimento tem a mesma probabilidade (Portal Action).

A distribuição de probabilidade descreve o comportamento aleatório de um fenômeno dependente do acaso. Note que flutuações e variabilidade estão presentes em quase todo valor que pode ser medido durante a observação de um fenômeno, ou seja, pode modelar incertezas e descrever fenômenos físicos, biológicos, econômicos, etc. Desta forma, a distribuição é um modelo matemático que relaciona um certo valor da variável em estudo com sua probabilidade de ocorrência. Há várias distribuições de probabilidade no estudo da estatística:

Distribuições Discretas: quando a variável que está sendo medida pode assumir valores inteiros, de contagem. Os principais modelos são: uniforme discreto, Bernoulli, binomial, geométrico, Poisson e o Hipergeométrico. Nesta disciplina abordaremos a distribuição Bernoulli e Binomial, que medem eventos baseado em sucesso X falha.

Distribuições Contínuas: quando a variável que está sendo medida é contínua. Há muitos exemplos destas distribuições, como descrito na página específica deste

tópico no Portal Action: https://www.portalaction.com.br/probabilidades/modelos-probabilisticos continuos. Nesta disciplina, abordaremos as mais conhecidas e utilizadas: distribuição normal, Distribuição t de Student, e distribuição qui-quadrado (atenção à esta última, que apesar de tratar dados categorizados, é uma distribuição contínua).

6.1 Distribuição Discreta - Bernoulli e Binomial

A estrutura básica da distribuição Bernoulli é baseada em duas possibilidades de resultado: sucesso e fracasso. Por exemplo, joga-se uma moeda uma vez (A moeda é equilibrada = os lados possuem peso igual), e define-se como sucesso sair cara. Define-se então que a variável aleatória X que assume valor 1 se ocorrer sucesso e 0 se ocorrer fracasso.

Parâmetro: probabilidade da variável assumir valor 1 ⇒ X: 0,1

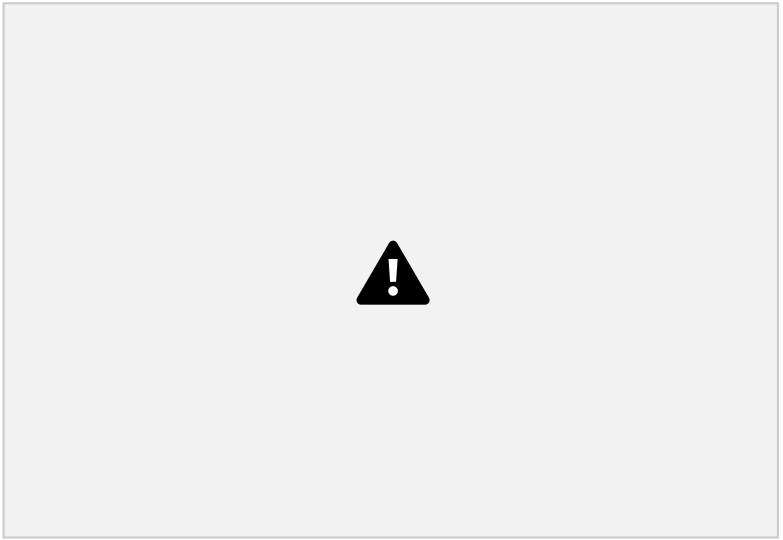
Notação: ﷺ ou p.

Se probabilidade de sucesso = p, a probabilidade de fracasso é q=(1-p),

porque p+q=1. Então:

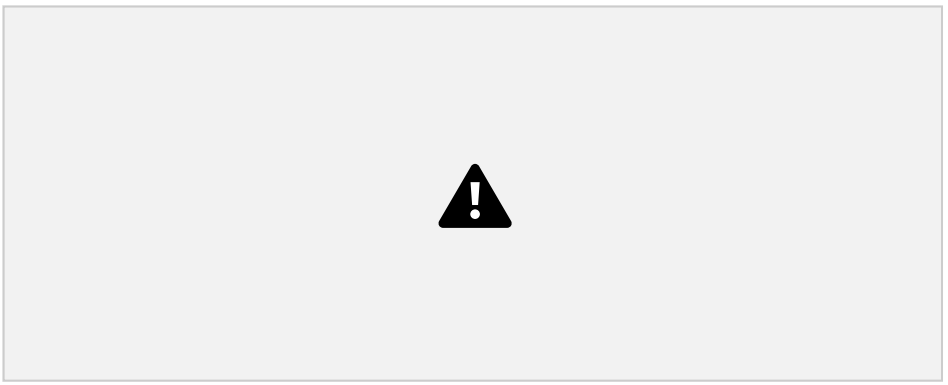
Probabilidade de sair cara = P(X=1) = p(1) = p = 0,5

Probabilidade de sair coroa = P(X=0) = p(0) = q = 1-p = 0,5



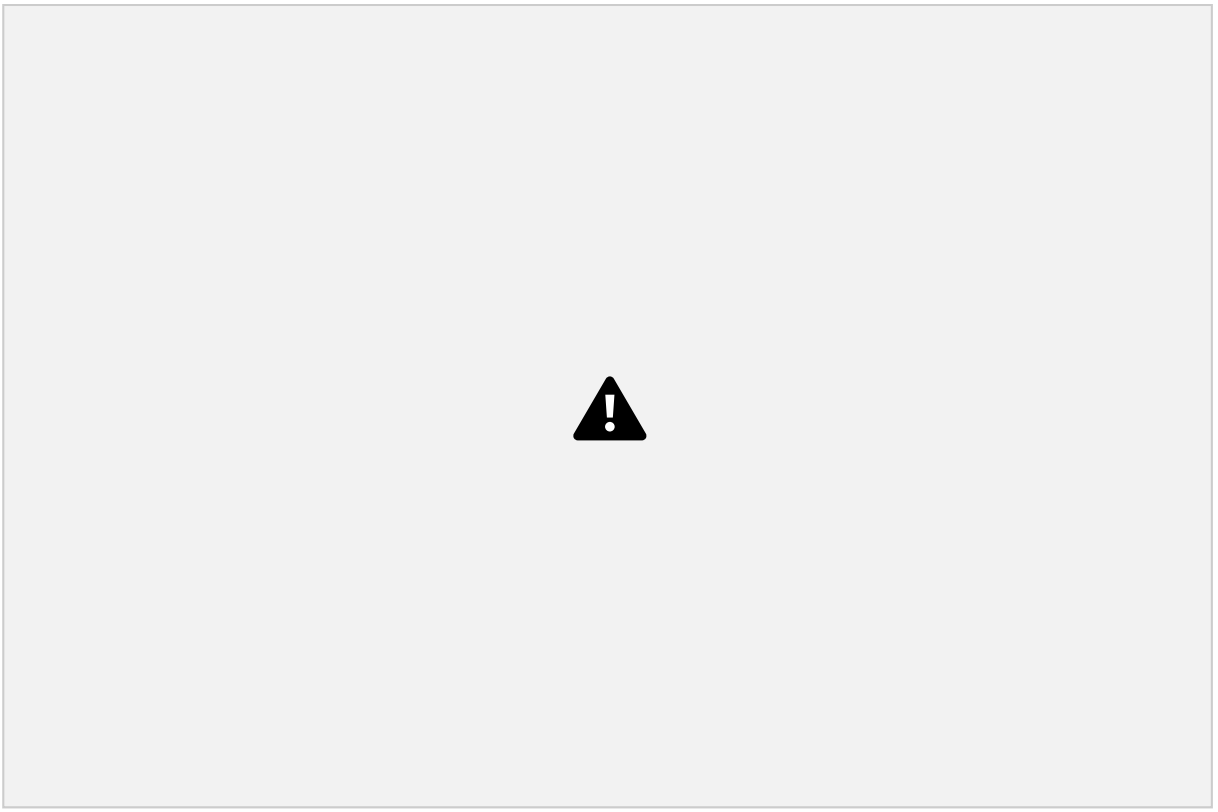
Exemplo:

Uma droga cura 15% dos pacientes. Administra-se a droga a um paciente. Qual a probabilidade do paciente ficar curado (sucesso)? Qual a probabilidade de fracasso (paciente não ficar curado)?



A distribuição binomial é a soma de n distribuições Bernoulli, sempre tratando de população / variáveis com 2 categorias. Exemplos: sexo (masculino, feminino); faces de uma moeda (cara, coroa); desfecho de um tratamento (cura, não cura).

Realiza-se o experimento n vezes, onde cada ensaio é independente do outro e os resultados são mutuamente exclusivos.



Então, X é uma variável aleatória (v.a.) Binomial, com parâmetros n e p e k. A probabilidade de ocorrerem k sucessos após n repetições é definida pela fórmula

abaixo (com média e variância nas fórmulas apresentadas):

Exemplo: Supondo que 20% de certa população tem sangue tipo B. Para uma amostra de tamanho 18, retirada desta população, calcule a probabilidade de que sejam encontradas:

a) 3 pessoas com sangue tipo B

N = 18 e P (sucesso) = 0,20 ⇒ P (3) = 0,230 ou 23,0%

b) 3 ou mais pessoas com sangue tipo B

N = 18 e P (sucesso) = 0,20 ⇒ P(≥3) = P(3) + P(4) + P(5)... + P(18) ⇒

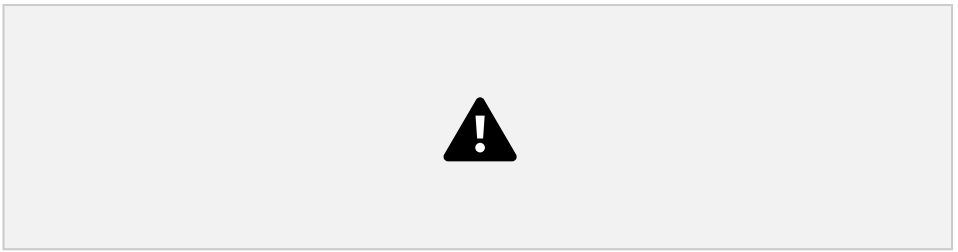
P(≥3) = 72,9% Ou ⇒ P(≥3) = 1 – P(0) – P(1) – P(2)

c) no máximo 3 pessoas com sangue tipo B

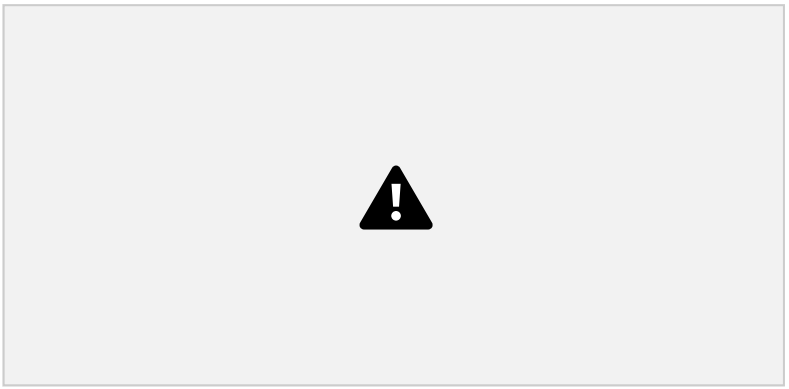
N = 18 e P (sucesso) = 0,20 ⇒ P(≤3) = P(0) + P(1) + P(2) + P(3) ⇒ P(≤3) = 50,1%

6.2 Distribuição contínua -Distribuição Normal

Também chamada gaussiana, de Gauss ou Laplace–Gauss, é sem dúvida a mais importante distribuição contínua, principalmente pelo fato de que muitos dos fenômenos estudados, características físicas, químicas, processos industriais, etc, seguem esta distribuição. Além disso, podemos citar o teorema central do limite, que garante que mesmo que os dados não sejam distribuídos segundo uma normal, a média dos dados converge para uma distribuição normal conforme o número de dados aumenta. Ela representa a distribuição de todos os valores da população, onde o ápice é indicado pela média populacional.

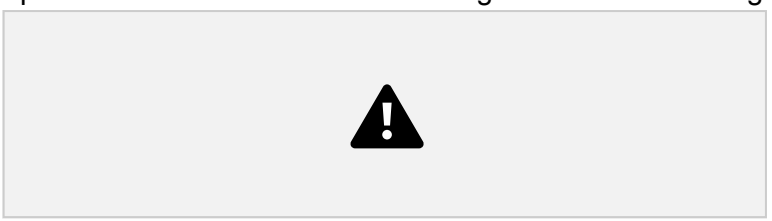


Se a variável aleatória X é normalmente distribuída com média μ e desvio padrão σ (variância σ2), a função densidade de probabilidade de X é dada por:

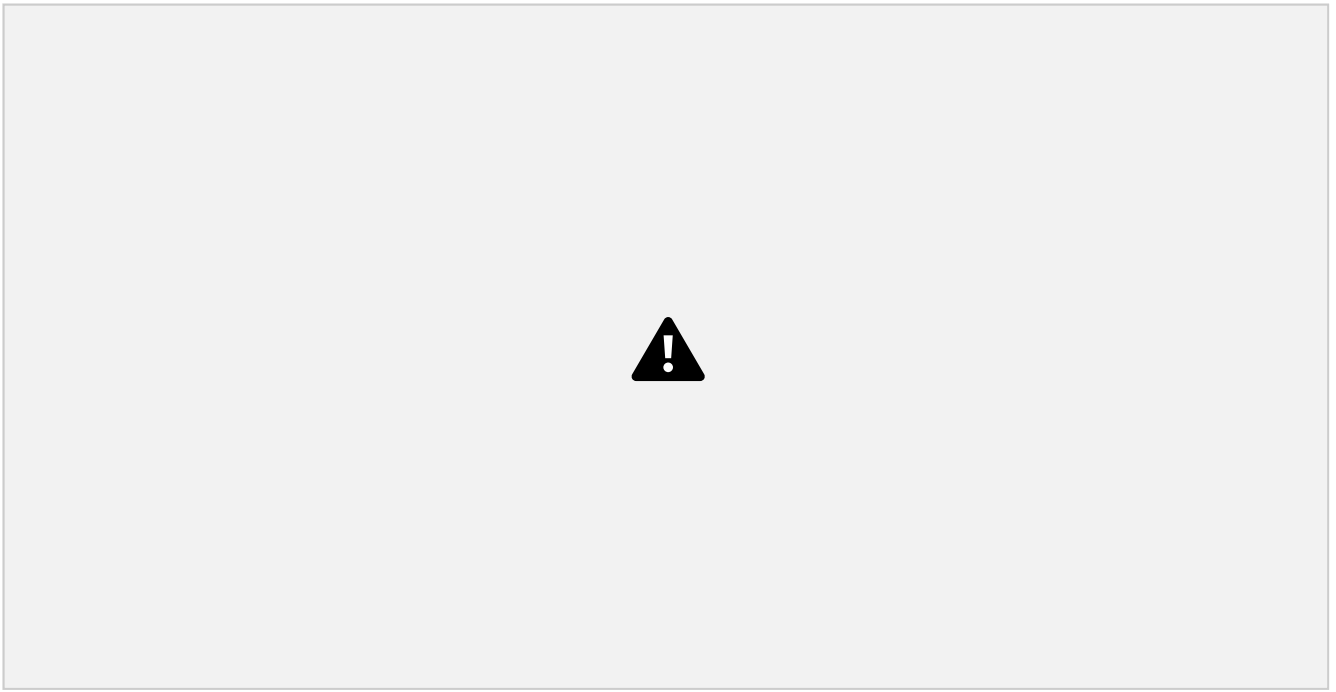


Principais propriedades:

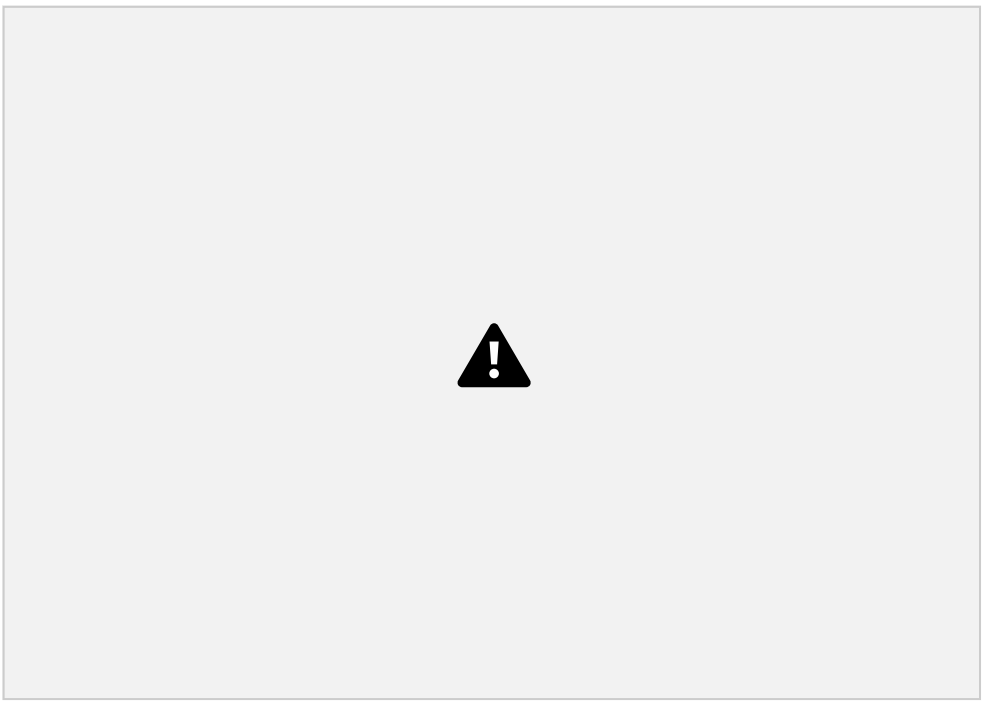
- Campo de variação : - ∞ < X < +∞
- A curva é simétrica em torno da média e tem forma de “sino”
- A média e mediana iguais
- A área entre a curva e o eixo horizontal (eixo x) é 1 ou 100% → pode ser entendida como probabilidade e é possível trabalhar com percentis
- A área entre a média e um ponto qualquer pode ser medida em de desvio padrão



Note que não existe somente uma curva normal. Em uma população, pode ser possível observar diferentes curvas para diferentes grupos da população (por exemplo, diferentes sexos, faixas etárias, estratos sociais, etc).

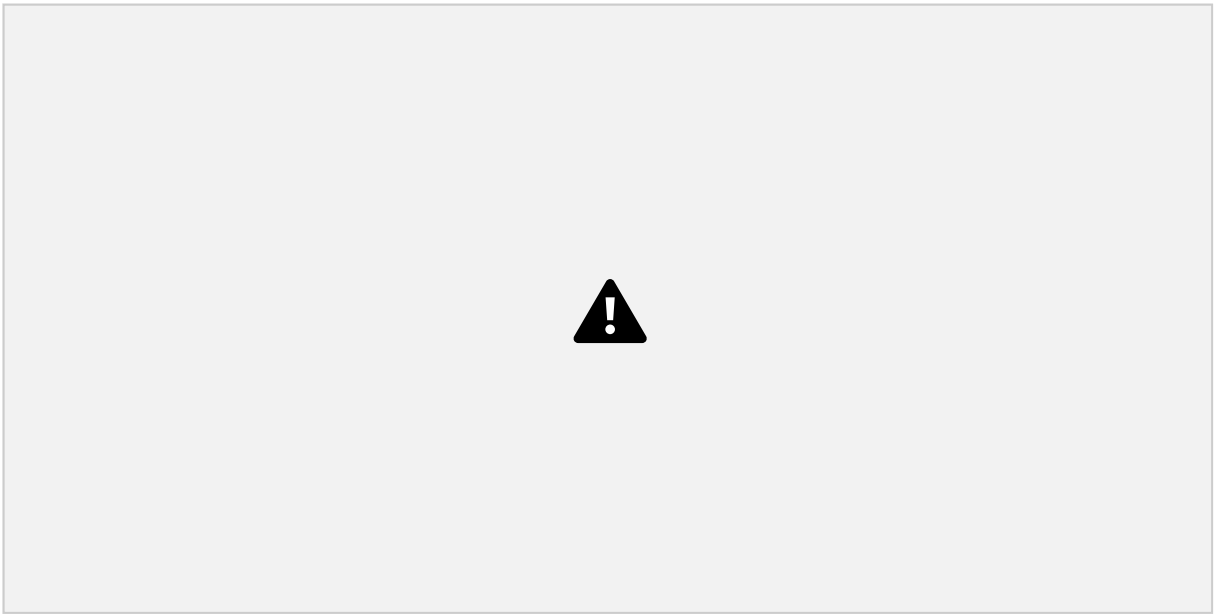


IMPORTANTE: A distribuição só será normal se seguir os pressupostos mencionados acima. Curvas sem o formato de “sino”, não simétricas em torno da média, possivelmente não apresentam distribuição normal.



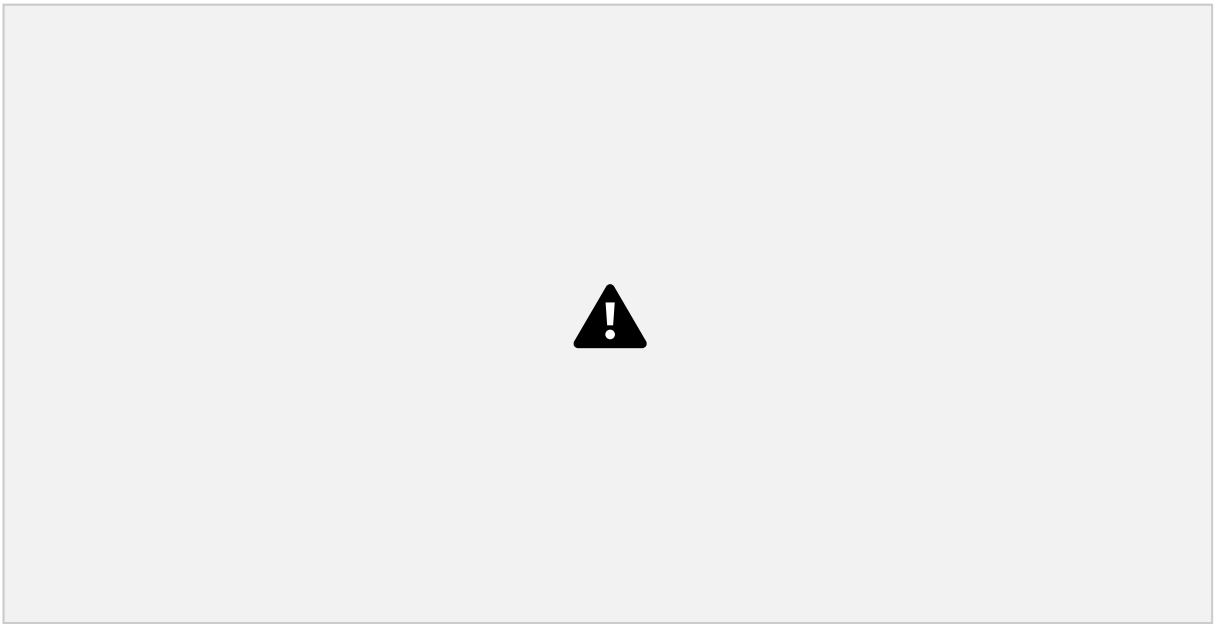
Assim, podemos verificar se nossa distribuição segue uma distribuição teórica normal através de alguns passos:

1. Verifique a distância entre a média, mediana, moda e desvio padrão 2. Verifique a histograma de distribuição

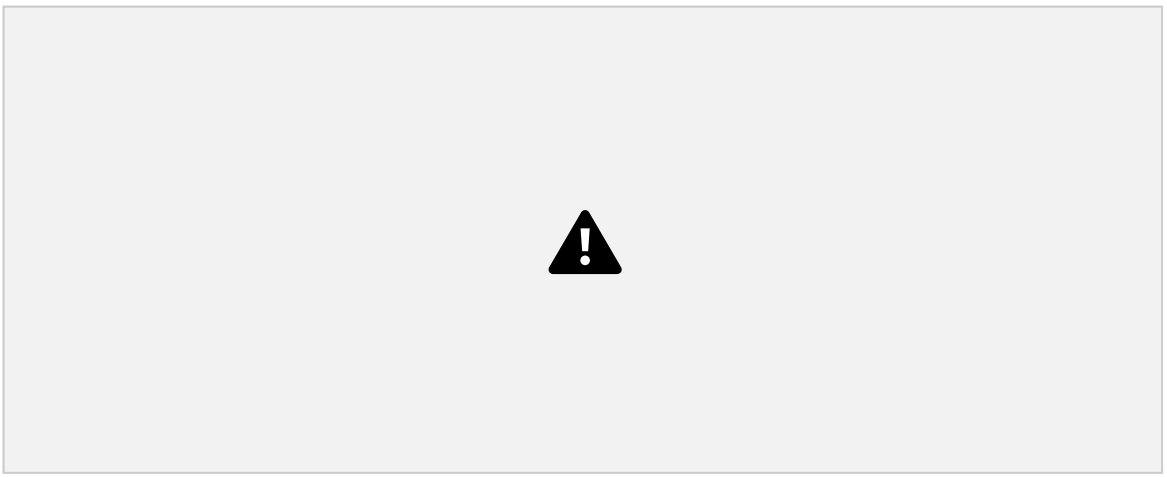


3. Aplique um teste para verificar a normalidade (será abordado no capítulo 8)

Exemplo: Depois de tomarmos várias amostras, decidiu-se adotar um modelo para as medidas de perímetro do tórax de uma população de homens adultos com os parâmetros: média (μ) = 40 polegadas e desvio padrão (s) = 2 polegadas. Quantos desvio padrão 43 está em torno da média? Qual a probabilidade de um indivíduo, sorteado desta população, ter um perímetro de tórax entre 40 e 43 polegadas?



Com base na mesma distribuição, qual a probabilidade de um indivíduo, sorteado desta população, ter um perímetro de tórax maior ou igual a 43 polegadas.



Exemplo 2 (Fonte: Portal Action):

Suponha que o peso médio de 800 porcos de uma certa fazenda é de 64 kg, e o desvio padrão é de 15 kg. Supondo que este peso seja distribuído de forma normal, quantos porcos pesarão entre 42 kg e 73 kg?

Então o valor padronizado de kg é de e de kg é de

. Assim a probabilidade é de

Portanto, o

número aproximado que se espera de porcos entre kg e kg é

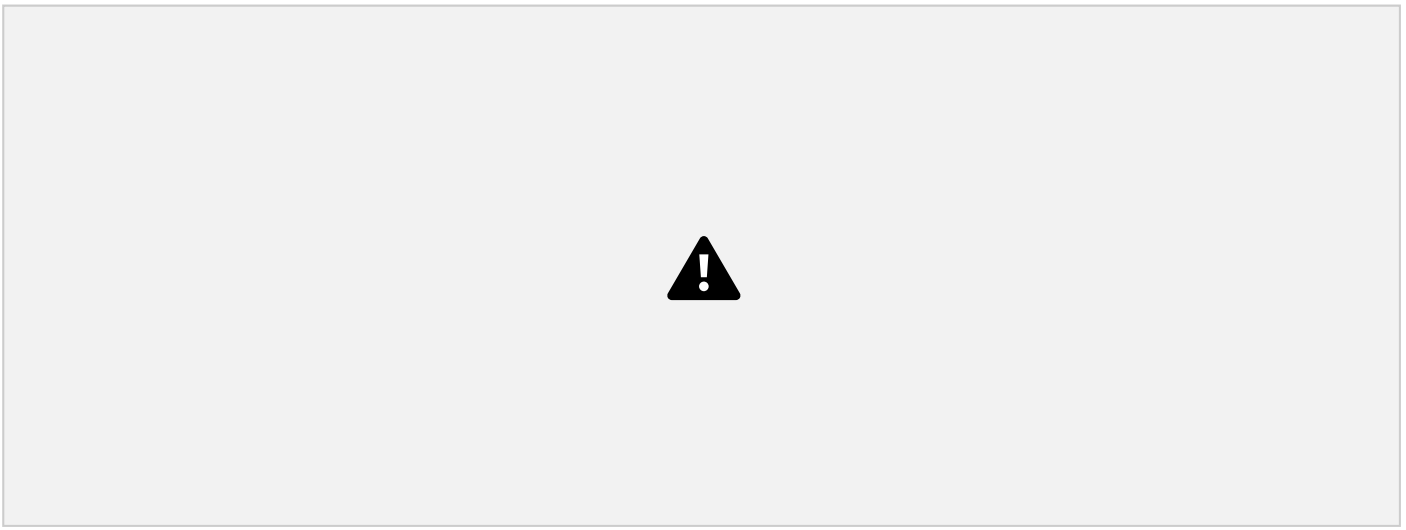
6.3 Distribuição contínua - Distribuição t Student

Esta é uma distribuição também muito utilizada em estatística e em teste de hipótese. Student é o pseudônimo de W. S. Gosset que, em 1908, propôs a distribuição t, que é muito parecida com a distribuição normal, também é centrada no zero e possui formato em sino. No entanto, a curva não é tão alta quanto a curva da distribuição normal e as caudas da distribuição t são mais altas que as da distribuição normal, pois reflete a maior variabilidade (com curvas mais alargadas) que é de se esperar em amostras pequenas.

Esta família t não descreve o que acontece na natureza mas sim o que aconteceria se selecionarmos milhares de amostras aleatórias de uma população normal com

média μ e fosse calculado:

O parâmetro que determina a altura e largura da distribuição t depende do tamanho da amostra (n) e é denominado graus de liberdade (gl), denotado pela letra grega (v) (lê-se ni). A notação da distribuição t é nt .Quando o número de graus de liberdade da distribuição t aumenta, a distribuição se aproxima de uma distribuição normal.



Nesta disciplina, utilizaremos amplamente as distribuições normal e t para testes de hipóteses e tomar decisões estatísticas, portanto seus conceitos devem estar bastante claros.

EXERCÍCIO 6

6.a) A probabilidade que uma pessoa que sofre de enxaqueca obter alívio utilizando certo medicamento é de 0,9. São selecionados 5 pacientes que sofrem de enxaqueca e recebem o medicamento. Quanto ao número de pessoas que vai ter

alívio, encontre a probabilidade de:

a) nenhuma pessoa ter alívio.

b) mais do que uma pessoa tenha alívio.

c) três ou mais pessoas tenham alívio.

d) no máximo duas pessoas tenham alívio.

6.b) Os valores de ácido úrico em homens adultos sadios seguem distribuição Normal com média 5,7mg% e desvio padrão 1mg%. Encontre a probabilidade de que uma amostra aleatória de tamanho 9, sorteada desta população, tenha média

a) maior do que 6 mg%.

b) menor do que 5,2 mg%.

6.c) Suponha que o peso em gramas do conteúdo de pacotes de salgadinho siga uma distribuição normal com média 500g e desvio padrão 85g. Sorteia-se uma amostra de 50 pacotes. Calcule:

a) a probabilidade de obter peso médio entre 500 e 530 gramas.

b) a probabilidade de obter peso médio entre 450 e 500 gramas.

7 Testes de aderência a normalidade

Interesse: Verificar se a distribuição das variáveis segue a distribuição normal

Existem disponíveis alguns testes para verificar a suposição de normalidade dos dados, Anderson-Darling, Cramer-von Mises, Kolmogorov-Smirnov e Shapiro-Wilk, bem como recursos gráficos, como histograma e normal plot. Esses testes também são chamados de Testes Goodness of fit. Geralmente verificam a discrepância dos valores observados com os valor esperados para aquela distribuição ou modelo.

Os testes de Anderson-Darling, Cramer-von Mises e Kolmogorov-Smirnov são baseados na função de distribuição empírica. Já o teste de Shapiro-Wilk baseia-se nos valores amostrais ordenados elevados ao quadrado e tem sido o teste de normalidade preferido por mostrar ser mais poderoso que diversos testes alternativos.

7.1 Quais testes de normalidade o software que eu uso tem?

Kolmogorov-Smirnov (com correção de Lilliefors)

- Software: SPSS, XLSTAT, BioEstat, SAS, GraphPad

D'Agostino- Pearson

- Software: GraphPad

Anderson-Darling

- Software: SAS

Shapiro-Wilk

- Software: SAS, Stata, SPSS, GraphPad

Cramér-von Mises

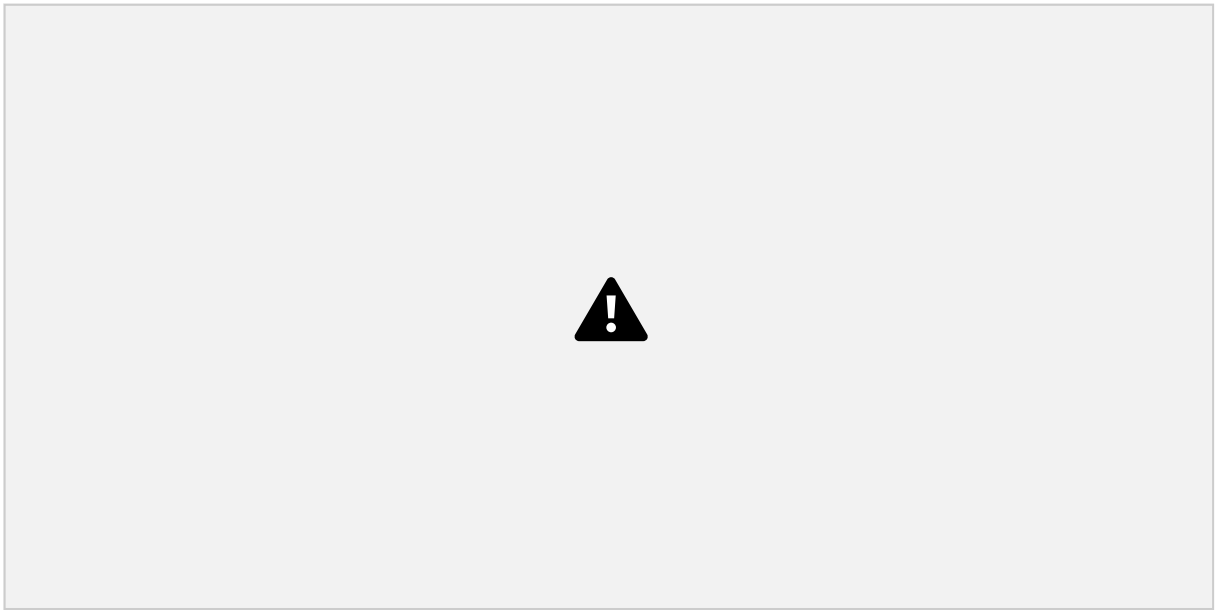
- Software: SAS

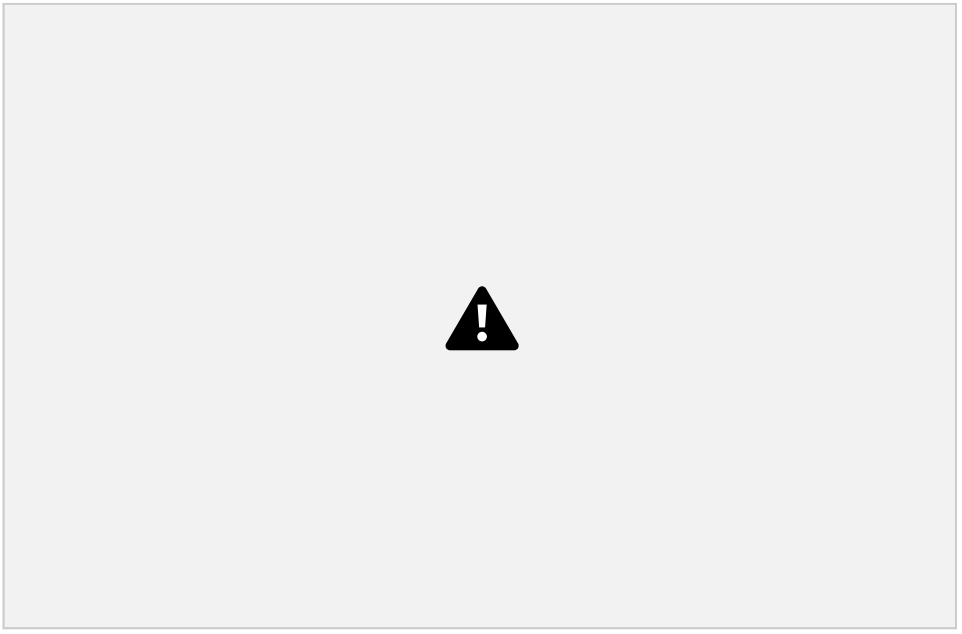
Teste de hipóteses dos testes de aderência à normalidade

- H0 : A amostra provém de uma população Normal
- H1 : A amostra não provém de uma população Normal

Ou seja caso o valor de "p" do teste seja <0,05 significa que a distribuição da amostra não é normal.

É sempre interessante detalhar no método estatístico do artigo, dissertação ou tese como foi verificada a normalidade, como nos exemplos abaixo:





Vamos apresentar com mais detalhe dois testes de aderência a normalidade: Shapiro-Wilk e Kolmogorov Smirnov.

7.1 teste de Shapiro-wilk

Shapiro-Wilk é o teste de normalidade que apresenta melhor poder estatístico seguido do Anderson-Darling (Razali and Wah, 2011).

O teste de Shapiro-Wilk gera uma estatística W, com base na fórmula abaixo:

Caso o valor de W seja maior que o valor crítico de W (chama do Wα) eu aceito H0. Caso o valor de W seja menor que Wα eu rejeito H0 e aceito H1.

H0 : W > Wα

H1 : W < Wα

Sugere-se que os pesquisadores não façam o teste Shapiro-wilk “na mão”. É um teste com muitas etapas, com alta chance de erro.

7.2 Teste de Kolmogorov Smirnov

Também conhecido como teste K-S

A estatística de Kolmogorov–Smirnov quantifica a distância entre a função distribuição empírica da amostra e a função distribuição acumulada da distribuição de referência

H0 : A amostra provém de uma população Normal

H1 : A amostra não provém de uma população Normal

Aceito H0 Rejeito H0

Valor crítico

Passo a passo

- Calcule as frequências acumuladas
- Compare com a frequência de referencia da normalidade
- Verifique o valor mais elevado de D
- Compare o valor D na tabela de valores críticos de D na prova de kolmogorov Smirnov

Exemplo

Teste de avaliação do Stress de alunos de pós do CNEM que cursam a disciplina de estatistica dos profs. Ligiana e Diogo. Como você classifica seu grau de stress durante as aulas de estatística, sendo: 5 Muito alto; 4 alto; 3 regular; 2 Baixo; 1 Muito baixo. Verifique se o grau de stress da turma segue a distribuição normal pelo teste de Kolmogorv Smirnov

Nível de Stress

1 2 3 4 5

No de pessoas

Freq. esperada acumulada 20 40 60 80 100 Fo (X) 0,20 0,40 0,60 0,80 1,00 Freq.

Observada acumulada

Sn(X)

|Fo(x)-Sn(X)|

Veja os resultados abaixo

Variável Média DP Mediana p de Shapiro

wilk

A 12,54 1,51 13 0,06

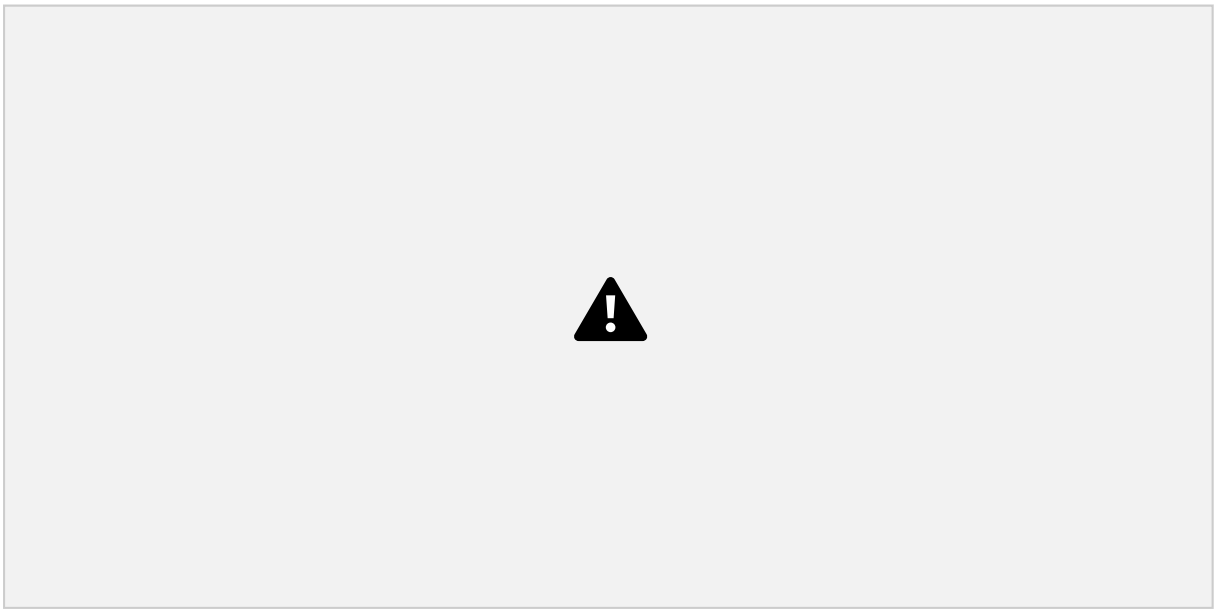
B 1,98 1,99 2,3 <0,001

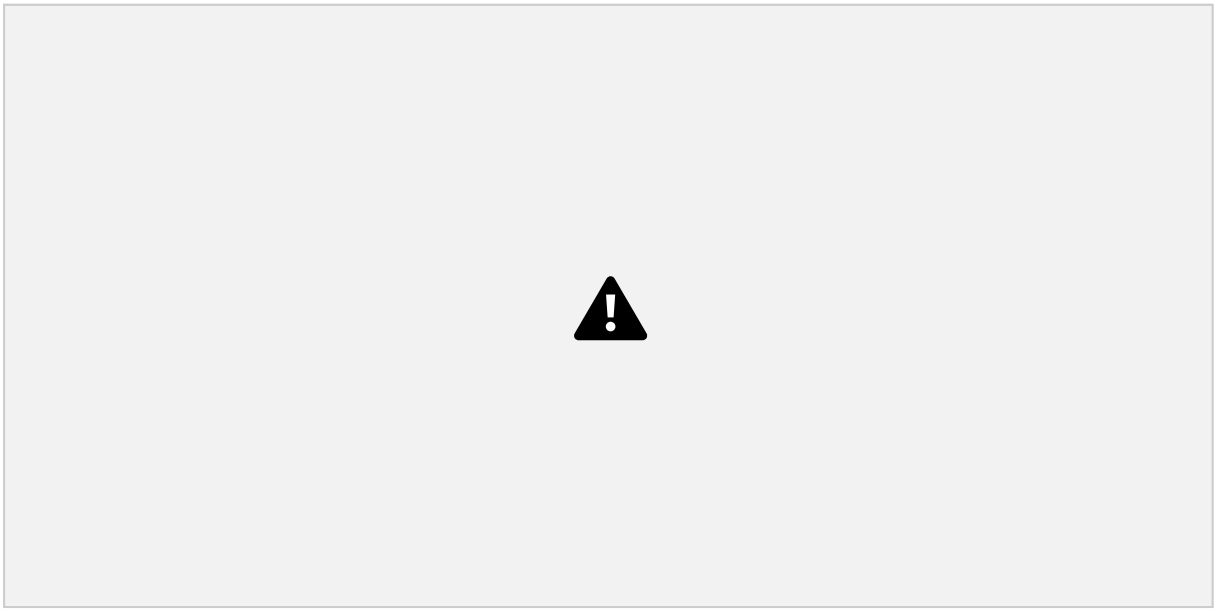
C 47,3 12,14 50,1 0,04

D 189,9 2,1 189,9 0,99

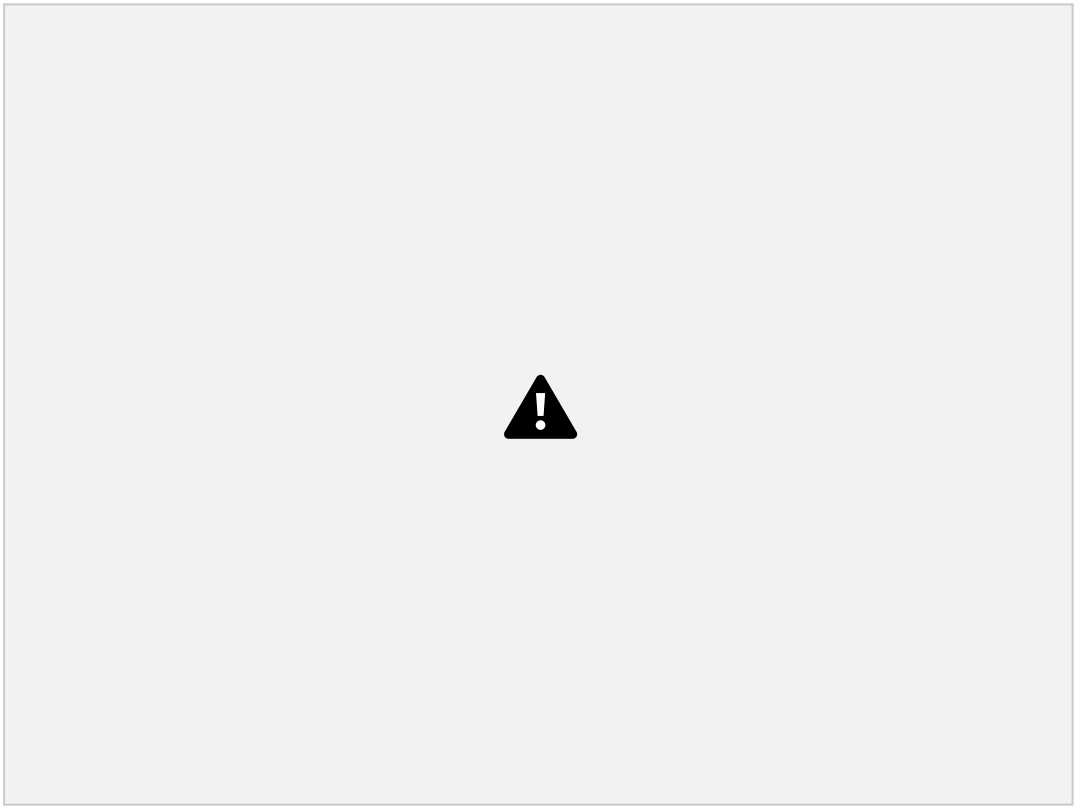
E 98,57 59,87 110,1 0,03

Que tipo de teste (paramétrico ou não-paramétrico) vocês usariam em cada variável?

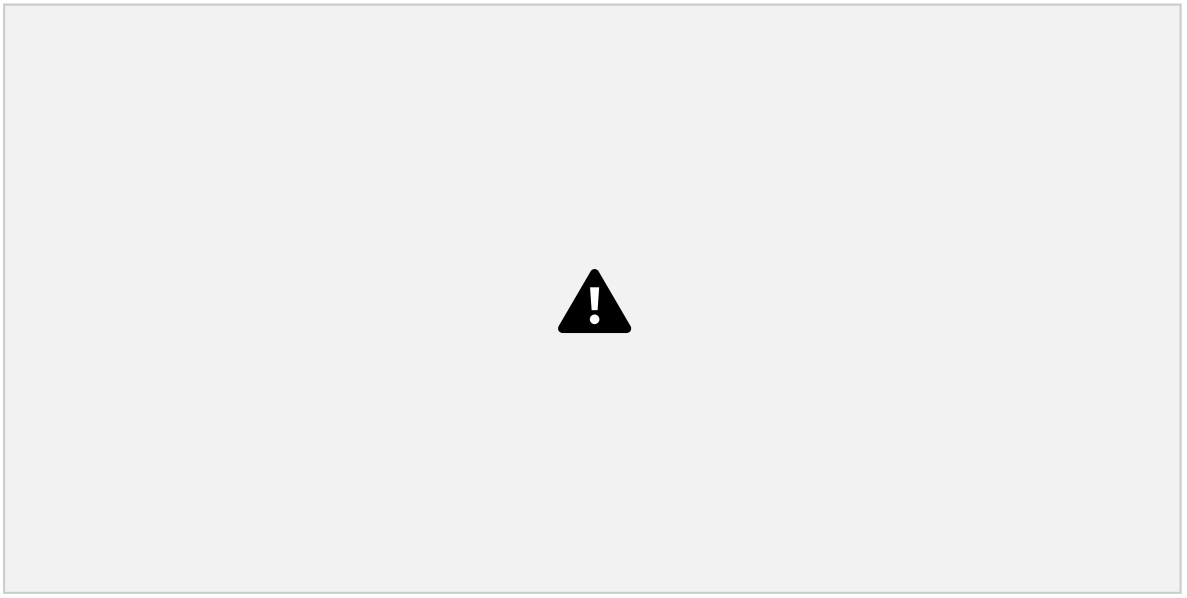




Saída no SAS



Saída no SPSS



Exercício7

7.a) Verifique se a variável idade, na tabela abaixo, segue a distribuição normal segundo o teste de Kolmogorov Smirnov.

N=308

Idade	Freq. %	Obser. %	acumulada Sn(X) %	esperada %	acumulada esperada	Fo(X) Fo*X) – Sn(X)

<50 6 1,9848 5,635

50-55 64 20,779 11,221

56-60 76 24,675 20,080

61-65 53 17,207 24,579

66-70 51 16,558 20,581

71-75 37 12,013 11,789

76-80 18 5,844 4,618

81-85 2 0,649 1,236

86 +	1	0,3247			0,257			

8 Associação de variáveis qualitativas

Interesse: Comparar variáveis dicotômicas.

8.1 Qui-Quadrado

O teste Qui-Quadrado é a melhor medida para se avaliar as diferenças entre uma distribuição de frequência teórica (E = esperada) e a obtida através de uma amostra (O = observada). É um teste não paramétrico, sendo possível aplicar em qualquer distribuição.

Pressupostos mínimos:

- As variáveis devem ser dicotômicas;
- Os grupos devem ser independentes;
- Não pode haver casela = 0;
- Não pode haver caselas<5 em amostras pequenas (n<20).

Teste de hipótese:

H0: O = E

H1: O ≠ E

Passo a passo do teste:

Verificar se as variáveis são qualitativas e independentes;

Identificar a variável de exposição e o desfecho e montar a tabela de distribuição, com a exposição na linha e o desfecho na coluna;

Desfecho 1 Desfecho 2 TOTAL

Exposição 1 A_O B_O T_{E1}

Exposição 2 C_O D_O T_{E2}

TOTAL	T_{D1}	T_{D2}	T_T

Calcular os valores esperados para ambas as variáveis;

Valores Observados Valores Esperados

A_O $A_E = T_{E1} * T_{D1} / T_T$

B_O $B_E = T_{E1} * T_{D2} / T_T$

C_O $C_E = T_{E2} * T_{D1} / T_T$

D_O	$D_E = T_{E2} * T_{D2} / T_T$

Calcular a diferença entre Valores Observados e Esperados (O - E); Elevar a diferença entre os Valores Observados e Esperados ao quadrado $(O - E)^2$; Dividir os valores da diferença entre Valores Observados e Esperados ao quadrado pelo respectivo Valor Esperado $(O-E)^2/E$

O E (O - E) $(O - E)^2$ $(O - E)^2 / E$

A_O A_E $A_O - A_E$ $(A_O - A_E)^2$ $(A_O - A_E)^2 / A_E$

B_O B_E $B_O - B_E$ $(B_O - B_E)^2$ $(B_O - B_E)^2 / B_E$

C_O C_E $C_O - C_E$ $(C_O - C_E)^2$ $(C_O - C_E)^2 / C_E$

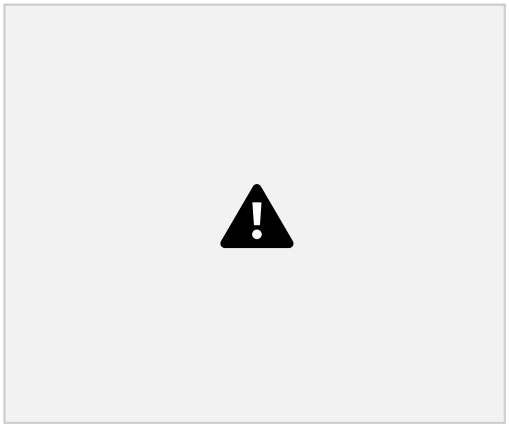
D_O	D_E	$D_O - D_E$	$(D_O - D_E)^2$	$(D_O - D_E)^2 / D_E$

Calcular os graus de liberdade (GL): ~$\chi^2_{(rows-1)(columns-1)}$. Para tabelas 2x2 o GL será sempre 1. Calcular a estatística do teste Qui-Quadrado $\chi^2 = \sum(O - E)^2 / E$

Verificar se o valor de Qui-Quadrado obtido é maior que o valor Crítico e o n<40 ou o valor de Qui-Quadrado obtido é maior que o valor Crítico e há pelo menos uma classe de frequência Esperada<5. Caso algum dos parâmetros anteriores seja verdadeiro, aplicar Qui Quadrado com Correção de Yates (somente para tabelas 2x2). $\chi^2 = \sum(|O - E| - 0{,}5)^2 / E$

Fixar a probabilidade de erro α (tipo I): Nível de significância (a)=0,05 (5%)

Limitar a área de rejeição do teste.



Valor crítico para 1 GL e 5% de significância = 3,841

Valor crítico para 1 GL e 0,5% de significância = 7,879

Decidir se aceita ou rejeita H0.

χ^2 > que o valor crítico ➨ Rejeita-se H0 (Há associação entre as variáveis)

χ^2 < que o valor crítico ➨ Aceita-se H0 (Não há associação entre as variáveis)

Exemplo:

Com objetivo de investigar a associação entre história de bronquite na infância e presença de tosse diurna ou noturna em idades mais velhas, foram estudados 1319 adolescentes com 14 anos. Fonte: Holland, WW et al., Long-term consequences of respiratory disease in infancy. Journal of Epidemiology and Community Health 1978; 32: 256-9

Tosse

TOTAL

Bronquite Sim Não

Sim 26 247 273

Não 44 1002 1046

TOTAL	70	1249	1319

Valores Observados | Valores Esperados (O - E) | (O - E)² | (|O - E|-0,5)² / E (O - E)² / E

26 14,488 11,512 132,526 9,147 8,370 247 258,512 -11,512 132,526

0,513 0,469

44 55,512 11,512 132,526 2,387 2,184

1002	990,448	-11,512	132,526	0,134	0,122

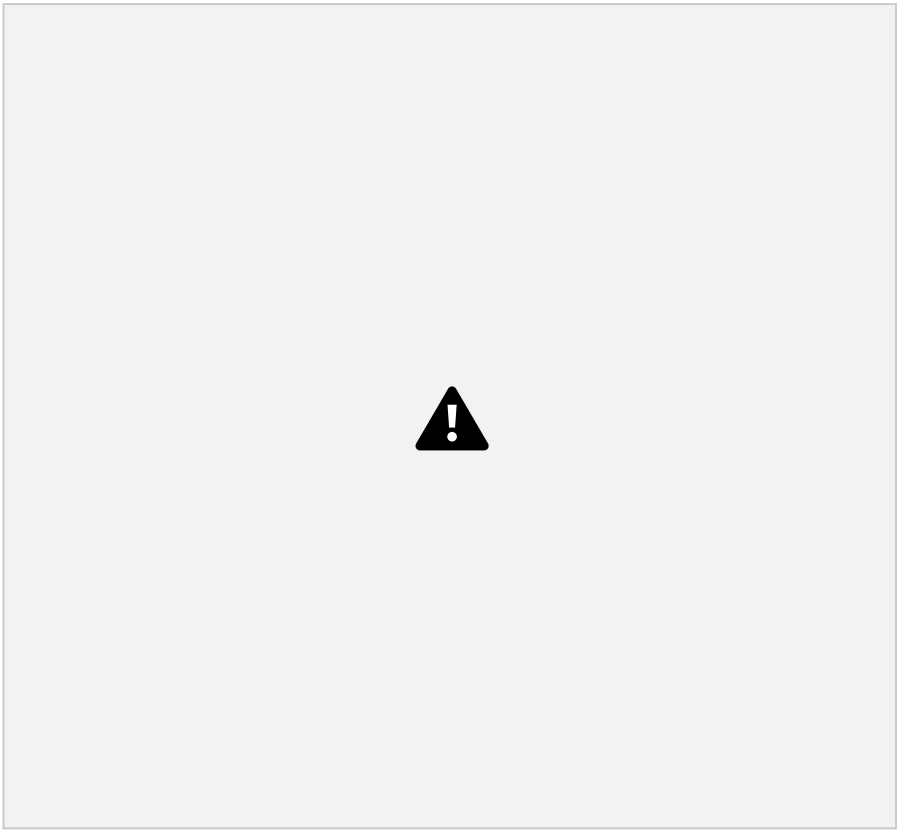
$\chi^2 = \frac{(O - E)^2}{E}$ = 9,147 + 0,513 + 2,387 + 0,134 = 12,181

$\chi^2_{\text{correção de Yates}} = \frac{(O - E - 0,5)^2}{E}$ = 8,370 + 0,469 + 2,184 + 0,122 = 11,145 χ^2 > Valor crítico ➨ Rejeita-se H0

12,181 e 11,145 são maiores que 7,789 (valor crítico para 1 grau de liberdade e p<0,001) Ou seja, existe associação (p<0,001)

Exato de Fisher

O Exato de Fisher é utilizado para se avaliar a significância entre associações de variáveis categóricas. É aplicado em amostras pequenas, n<20 ou que contenha caselas <5.



Teste McNemar

O Teste de McNemar é utilizado para se avaliar a significância entre associações de variáveis categóricas, porém para amostras pareadas, ou seja, dependentes. Ex.: Estudos caso controle ou estudos clínicos de intervenção pareados.

H0 = não existem diferenças na presença de A entre amostras;

H1 = existe diferença na presença de A entre as amostras.

Amostra I

Total

I Presente I
Ausente
c+d

Amostra II
II Presente a b a+b II Ausente c d

Total	a+c	b+d	n=a+b+c+d

O teste também é utilizado em medidas repetidas (desenvolveu ou não desenvolveu determinado atributo).

H0 = indivíduos “não

mudaram”; H1 = indivíduos

“mudaram”.

Antes

Depois

\+ -

\+ a b

-	c	d

Mudança: a e d; Não Mudança: b e c.

A fórmula do Teste de McNemar se origina da fórmula do Qui-Quadrado:

$$\chi^2 = \sum (O - E)^2 / E$$

As células de interesse nesse Teste são somente a A e a D. Desta forma A é o número de casos observados na célula A e D é o número de casos observados na célula D e (A + D)/2 é o número esperado de casos em cada uma das células, temos:

Simplificando:

$$\chi^2 = (O - E)^2 / (O + E)$$

Com correção de Yates:

$$\chi^2 = (|O - E| - E)^2 / (O + E)$$

Exemplo:

Para verificar a eficácia de um tratamento contra o ebola, uma amostra de 25 pacientes infectados foi analisada antes e depois do tratamento.

Antes

Depois

BEM MAL

MAL A B

BEM	C	D

H0 = não existe diferença antes e Depois

BEM MAL

após o tratamento H1: existe

diferença antes e após o

tratamento

TOTAL

Antes

MAL 14 4 18

BEM 3 4 7

	17	8	25

A = 14; D=4

$$\chi^2 = (|A - E| - 0,5)^2 / (A + E)$$

$$\chi^2 = (|14 - 10| - 0,5)^2 / (14 + 10)$$

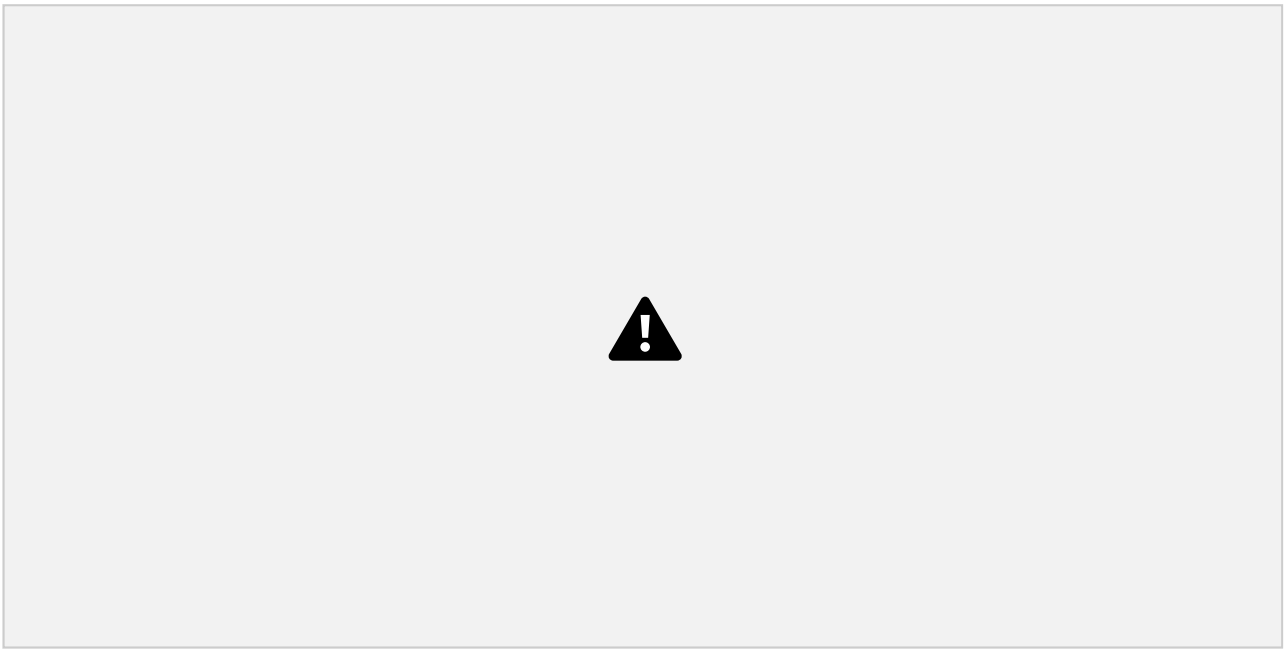
$$\chi^2 = 4,5$$

χ^2 > que o valor crítico ➨ Rejeita-se H0

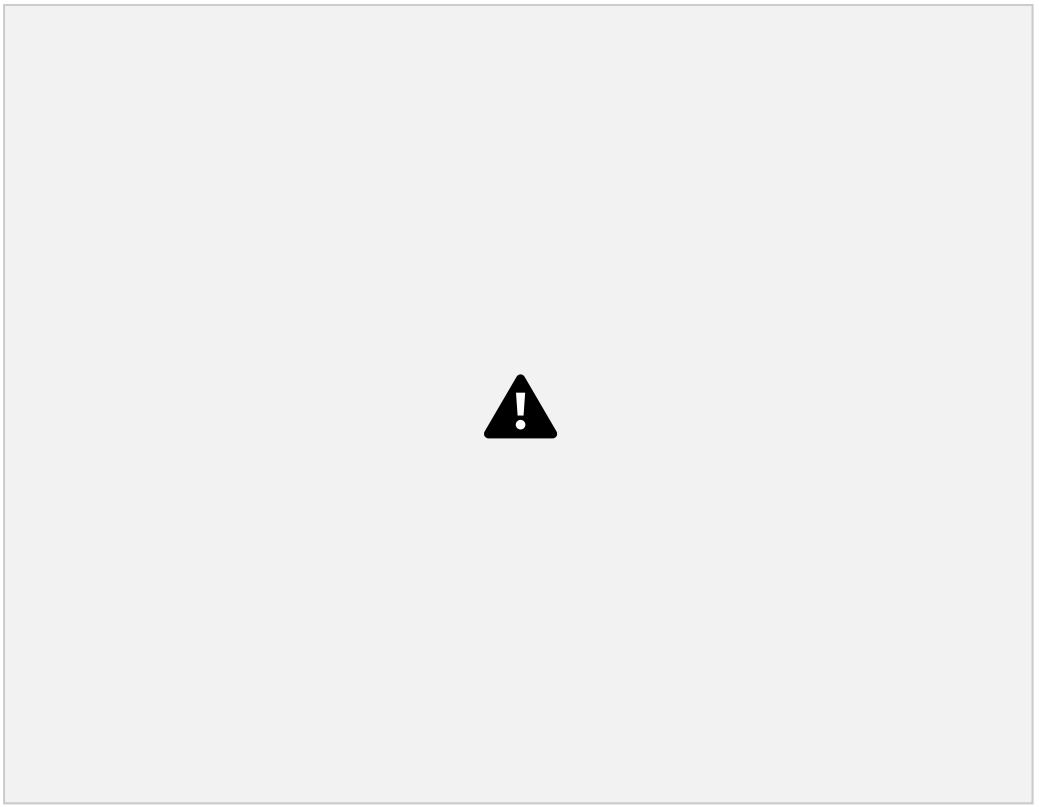
χ^2 < que o valor crítico ➨ Aceita-se H0

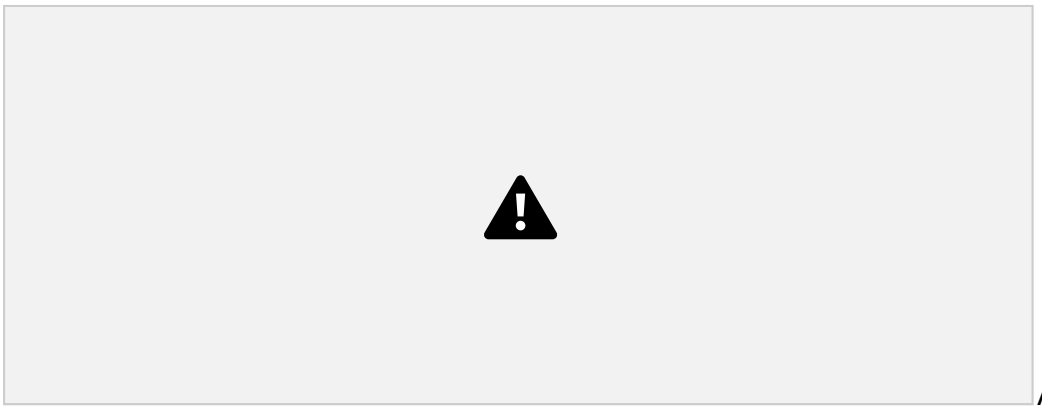
4,5 > 3,841 (valor crítico para nível de significância de 5% ou 0,05)

Rejeita-se H0, ou seja, o tratamento faz diferente. Neste caso a diferença indica piora dos pacientes.



Exemplo: artigo utilizando Teste Qui-Quadrado





9 Comparação de médias de dois grupos independentes

Interesse: Comparar uma variável quantitativa com uma variável qualitativa, envolvendo 2 grupos independentes.

Exemplos de grupos independentes: homens e mulheres; tratados com a droga X e grupo controle; estudantes da rede pública e particular.

9.1 Teste t-Student

Pressupostos mínimos:

- Os dois grupos possuem distribuição normal;
- Há igualdade de variância (homocedasticidade);
- Os dois grupos comparados são independentes;

● As variáveis devem ser quantitativas contínuas ou discretas.

Passo a passo do teste:

- Verificar a aderência à normalidade:
    - Observar a distância entre a média e mediana (costumam ser próximas em distribuição normal);
    - Observar o desvio padrão;
    - Observar o histograma de distribuição;
    - Aplicar o teste de distribuição (ex.: Shapiro Wilk ou KS).
- Verificar homocedasticidade:
    - Teste de Levene, Bartlett ou Cochran.
- Se apresentar aderência à normalidade e homocedasticidade, então seguir os passos abaixo:
    - Ordenar os grupos;
    - Calcular a média dos dois grupos;
    - Calcular a variância dos dois grupos;
    - Aplicar a fórmula;
    - Verificar o valor t na tabela de distribuição com crítico de 5% (com GL = (n1+n2) - 2); - Tomar a decisão do teste: aceitar ou rejeitar H0.

Fórmulas:

Hipóteses do teste

Bicaudal:
H0 = μ1 = μ2
H1 = μ1 ≠ μ2

Unicaudal:
H0= μ1 = μ2
H1 = μ1 < μ2
H1 = μ1 > μ2

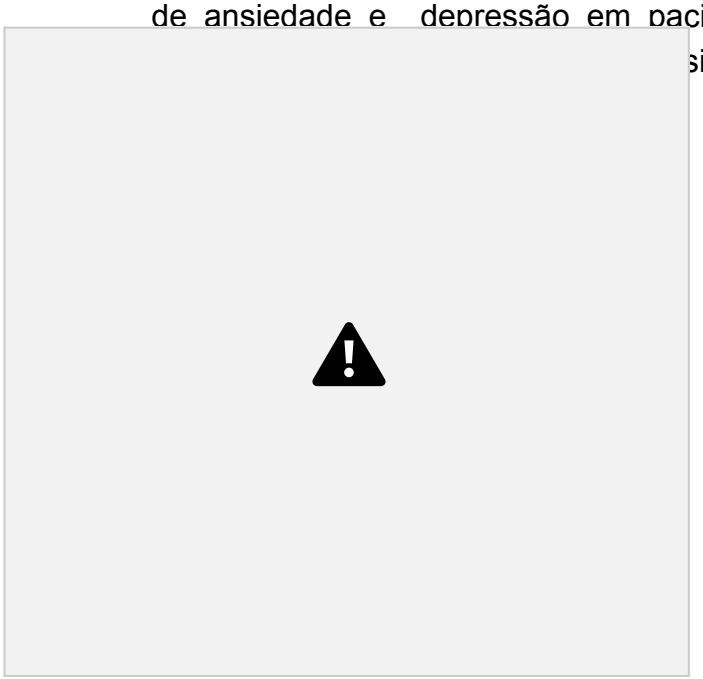
*Rejeita H0 quando o valor encontrado está na área hachurada do gráfico.

Dica rápida
Quando realizar o teste desenhe o gráfico para visualizar o valor t calculado e o t

crítico, isso ajuda a tomar a decisão!

Exemplo:

Um estudo transversal realizado no Brasil com 84 indivíduos verificou, entre outros aspectos, a diferença entre homens e mulheres na apresentação de sintomas de ansiedade e depressão em pacientes em pré-operatório de cirurgia cardíaca siedade e Depressão. Para esta comparação foi



"A Tabela 2 apresenta as médias e desvios-padrão dos sintomas de ansiedade e depressão, segundo o sexo e a idade categorizada. As mulheres apresentaram maiores médias que os homens tanto para os sintomas de ansiedade como de depressão, e essas diferenças foram estatisticamente significantes."

Fonte: RODRIGUES, H. F. et al. Ansiedade e depressão em cirurgia cardíaca: diferenças entre sexo e faixa etária. Escola Anna Nery, v. 20, n. 3, 2016.

E se a distribuição não for normal ou não houver homocedasticidade?

9.2 Teste U de Mann Whitney

Versão não paramétrica do teste t-Student para comparar médias de dois grupos independentes.

Passo a passo do teste:

- Ordenar os grupos (não separar os grupos para ordenar);
- Testes não paramétricos se baseiam em Ranks e não em médias! Então: atribuir os Ranks aos valores ordenados;
- Calcular R1 e R2 (soma dos Ranks do grupo 1 e 2);
- Estimar o valor de U;
- O menor valor entre U1 e U2 será o Ucrit;
- Olhar na tabela específica (o maior "n" dos grupos e o menor U observado e encontra o valor de p ou olha n1 e n2 e encontra o Ucrit no tabela);

- Tomar a decisão do teste.

U → número de vezes que o grupo 1 (menor dos grupos) antecede o grupo 2

Fórmulas:

Hipóteses do teste:

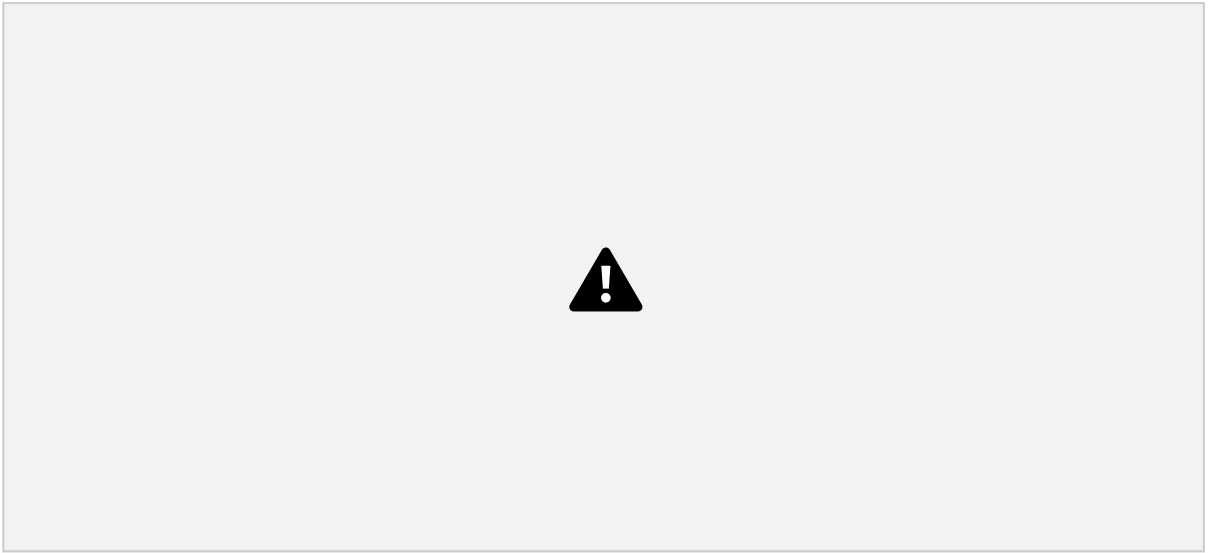
H0 = Md1 = Md2

H1 = Md1 ≠ Md2

Rejeita H0 quando Ucrit > Ucalculado ou p<0,05.

Exemplo:

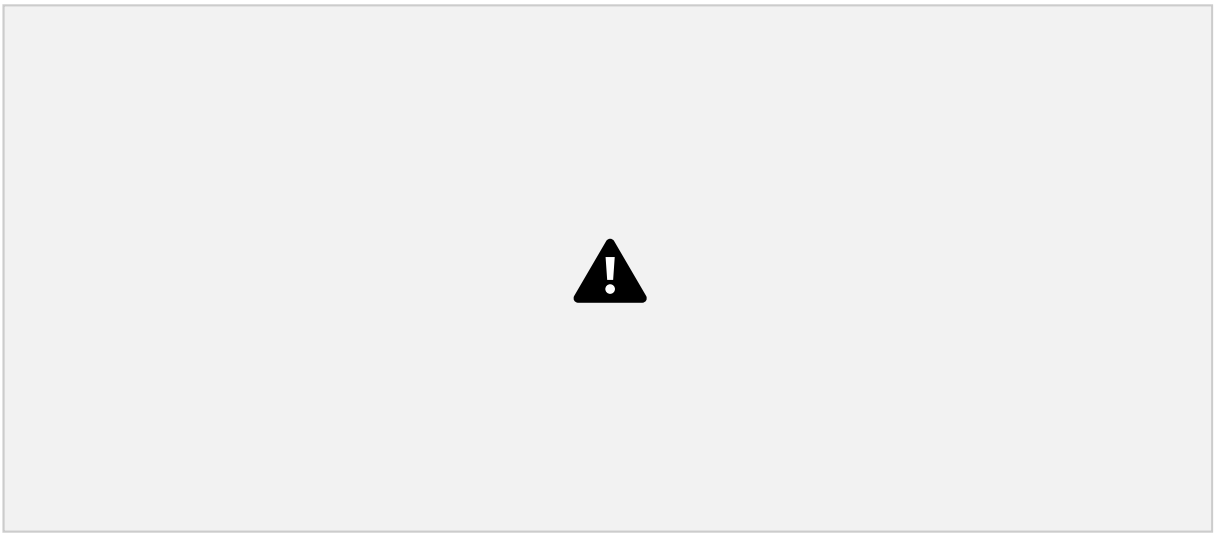
Um estudo de intervenção realizado com escolares matriculados em duas escolas públicas no Brasil realizou medidas antes e depois da intervenção, realizando comparações entre e intragrupos. Para comparar os grupos intervenção (GI) e controle (GC) nos dois momentos do estudo foi utilizado o teste U de Mann Whitney:



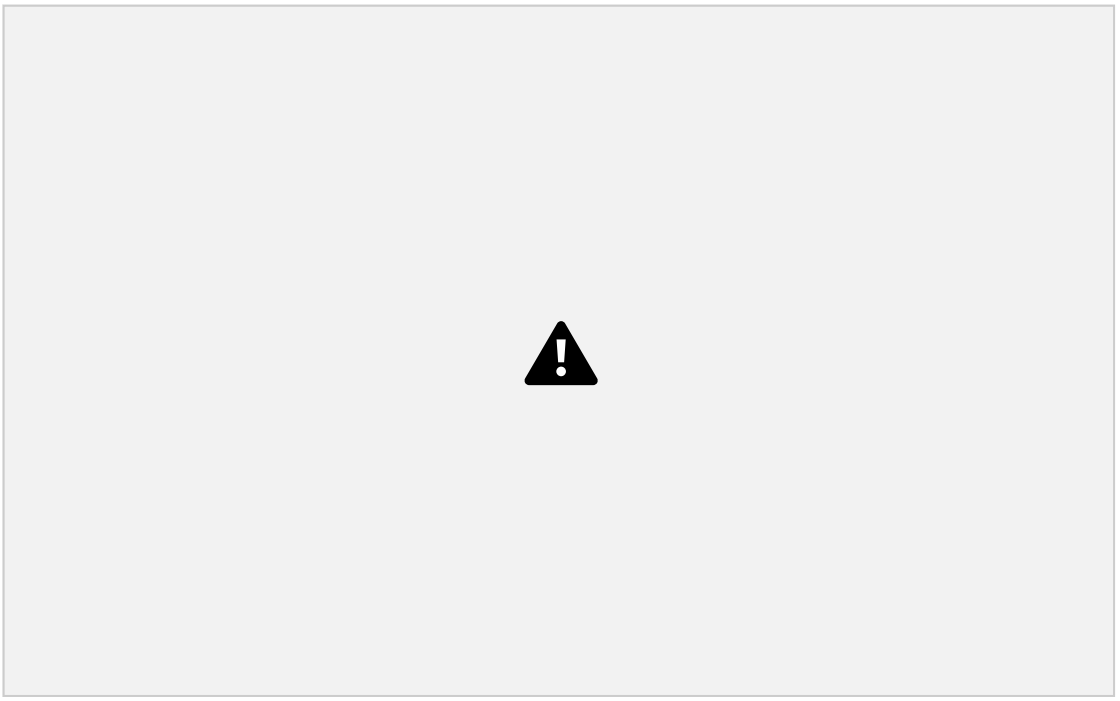
"A Tabela 4 apresenta as médias das variáveis antropométricas e de hábitos

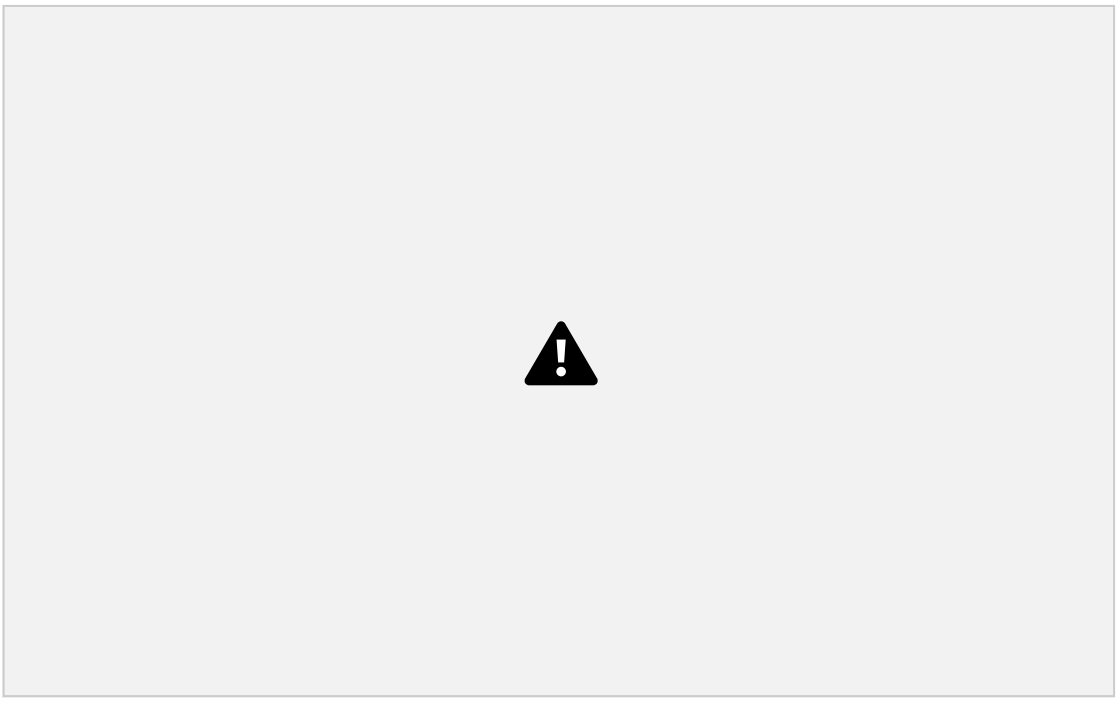
de vida nos dois momentos de avaliação por grupo de alocação. Houve diferença entre os grupos tanto na avaliação inicial e final em relação ao peso (p=0,033 e p=0,030, respectivamente), PC (p=0,007 e p=0,028, respectivamente) e ao IMC na avaliação final (p=0,024). Com relação ao tempo de tela, a média de ambos os grupos foi cerca de duas vezes maior que a recomendação nos dois momentos. A média da prática de atividade física também foi superior à recomendação, cerca de três vezes."

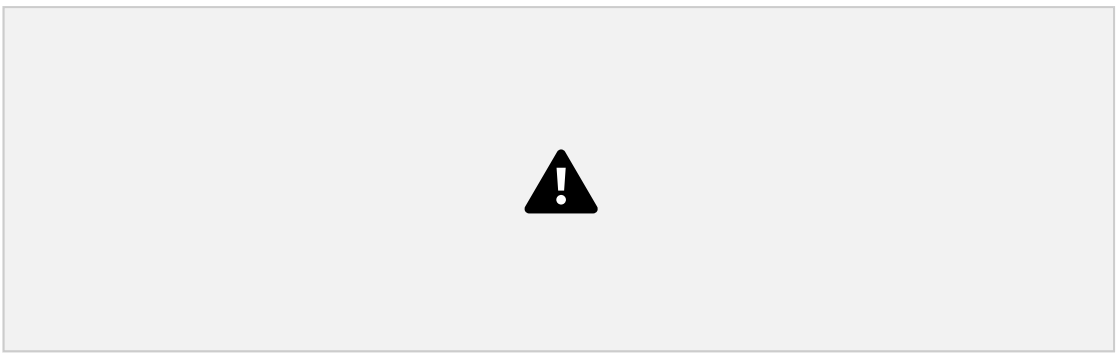
Fonte: COELHO, L. F.; SIQUEIRA, J. H.; MOLINA, M. del C. B. Estado Nutricional, Atividade Física E Tempo De Tela Em Escolares De 7-10 Anos: Um Estudo De Intervenção Em Vitória-Es. DEMETRA: Alimentação, Nutrição & Saúde, v. 11, n. 4, p. 1067–1084, 2016.



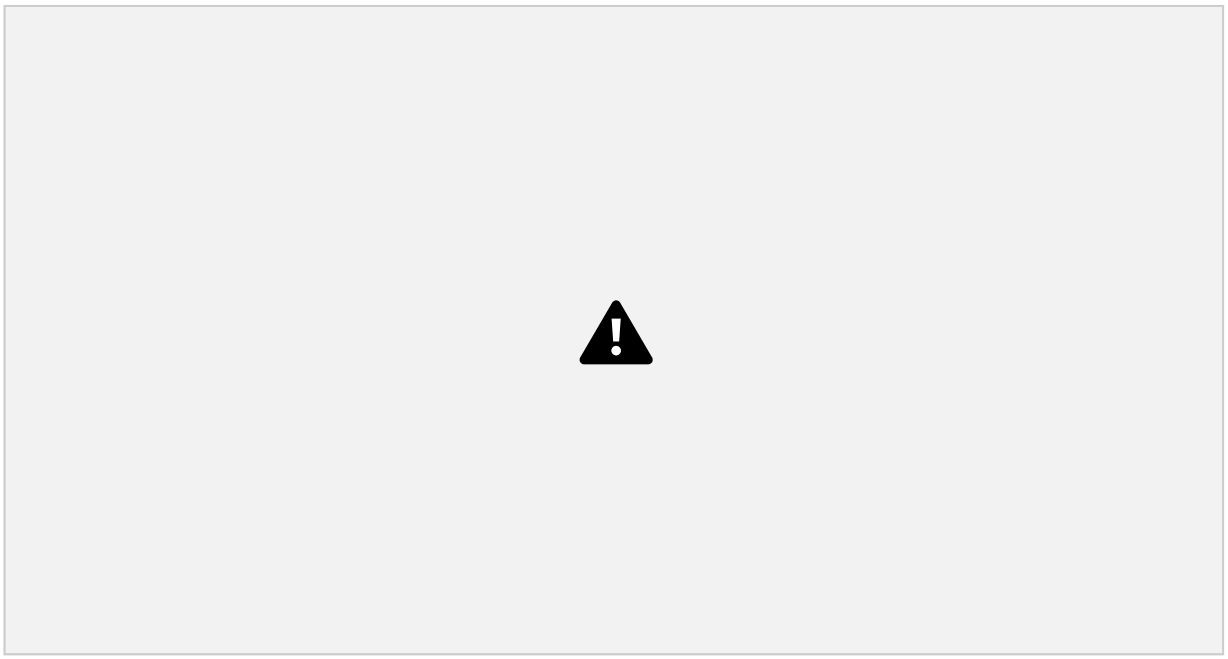
Valores críticos de Distribuição T

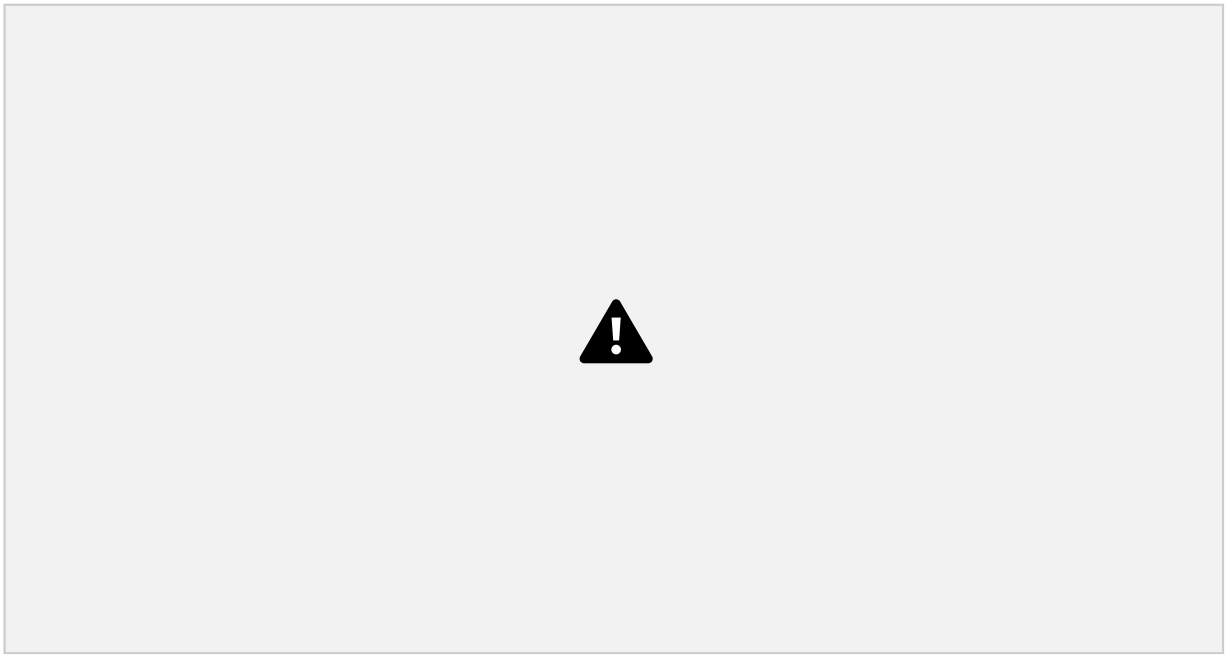


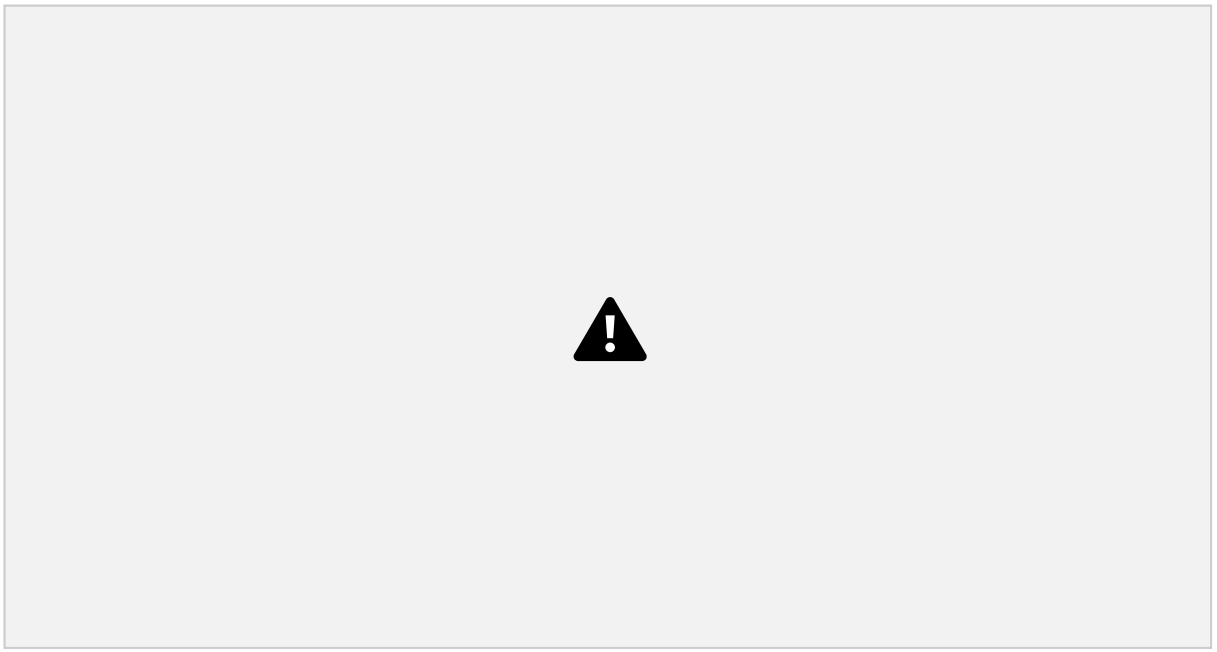


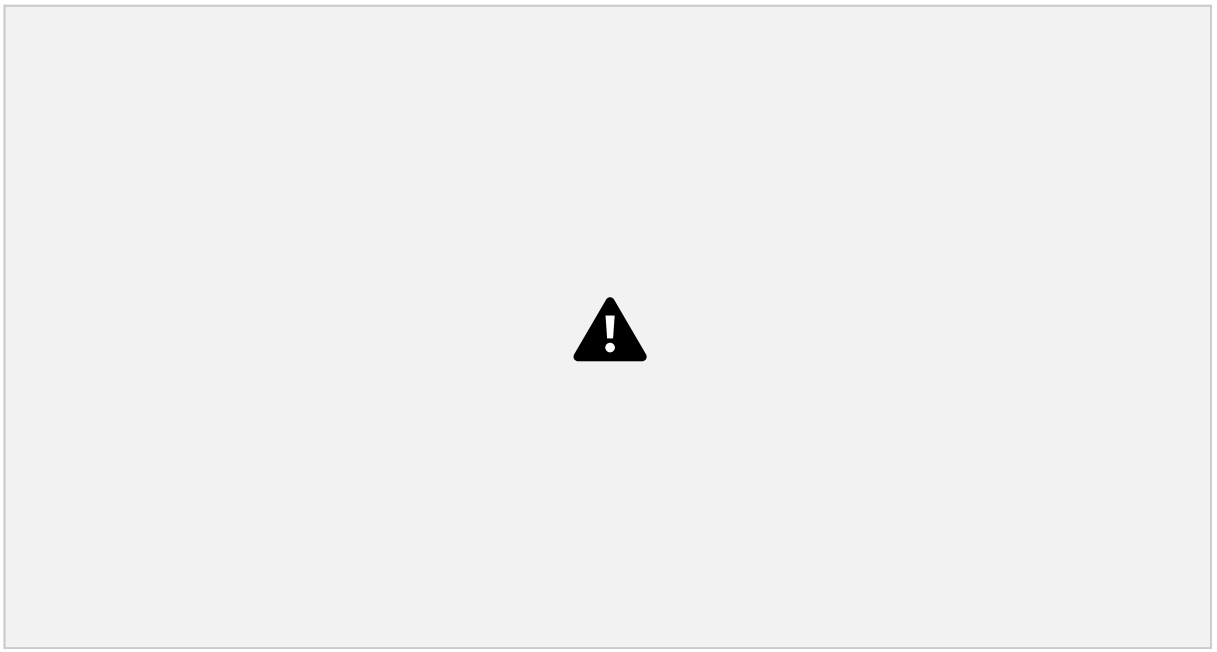


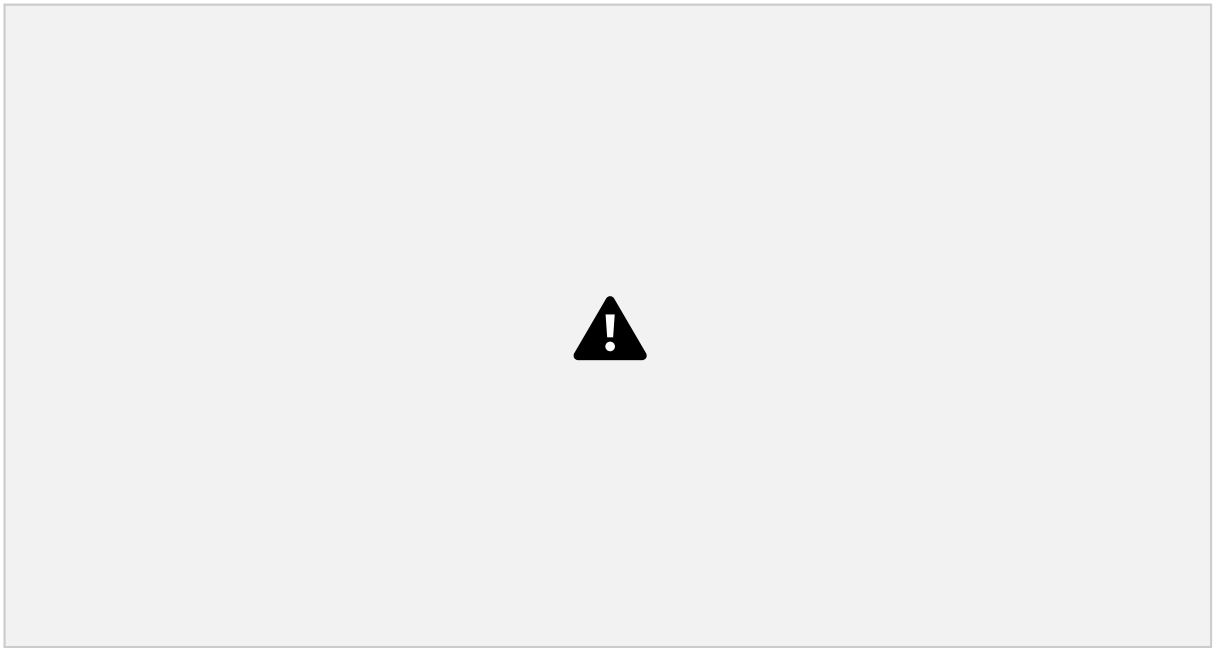
Valores críticos U de Mann Whitney:

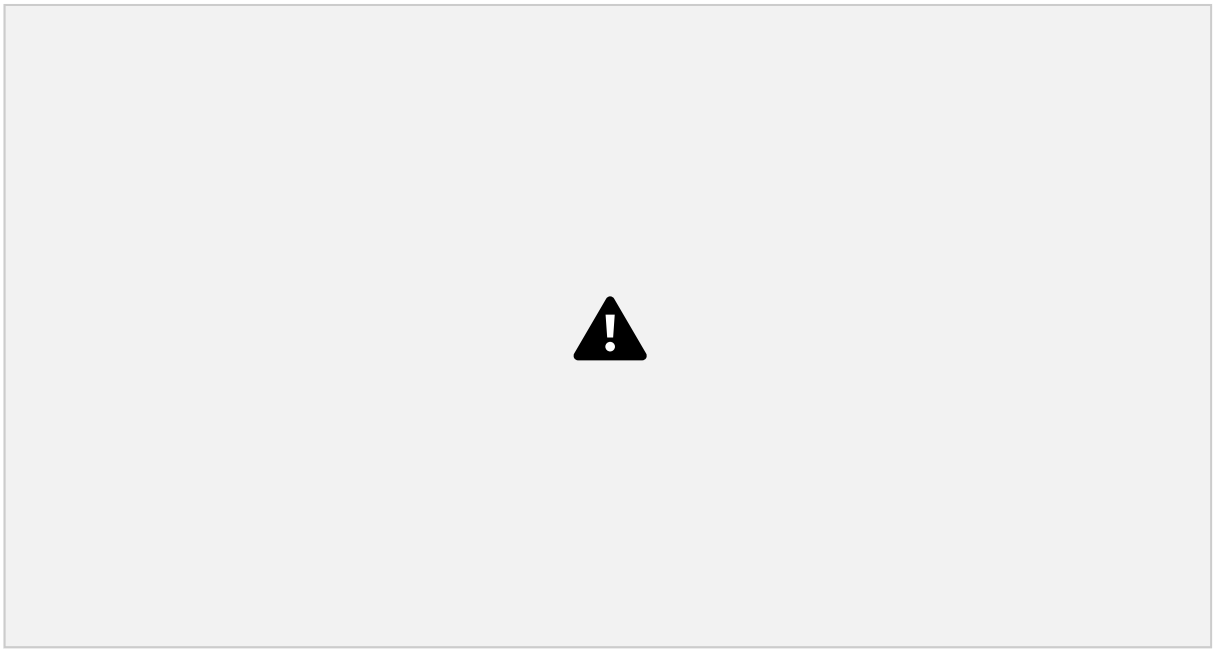


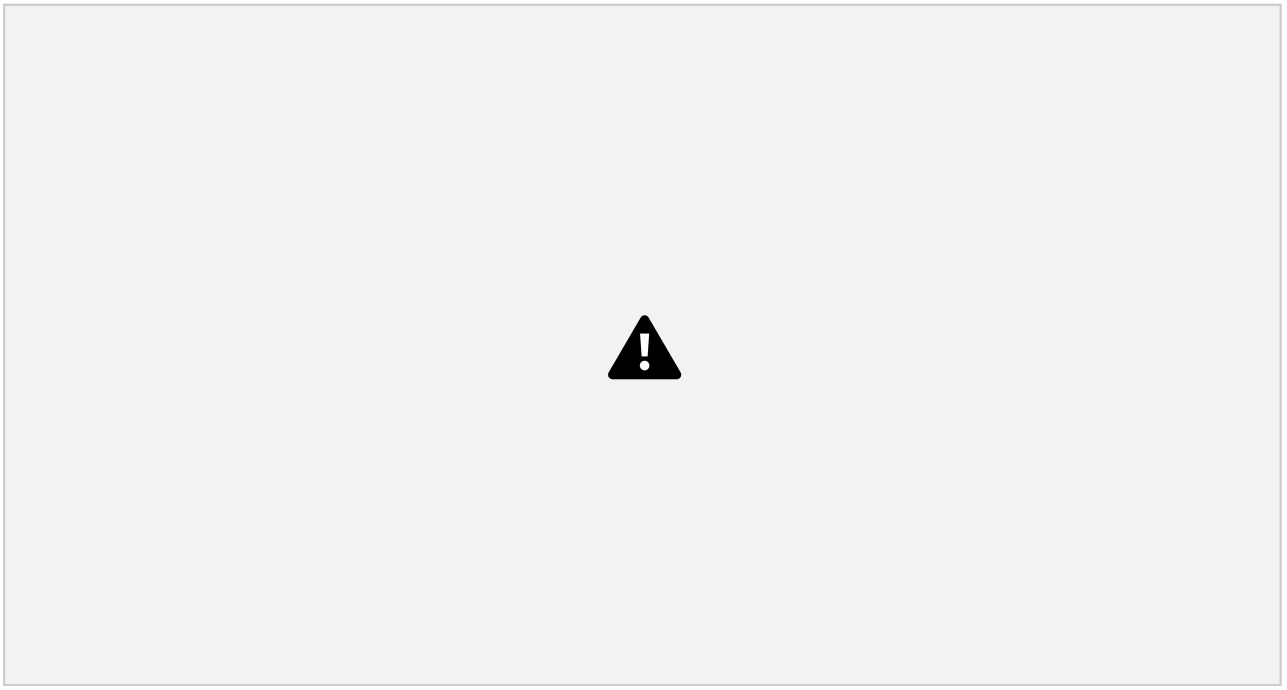


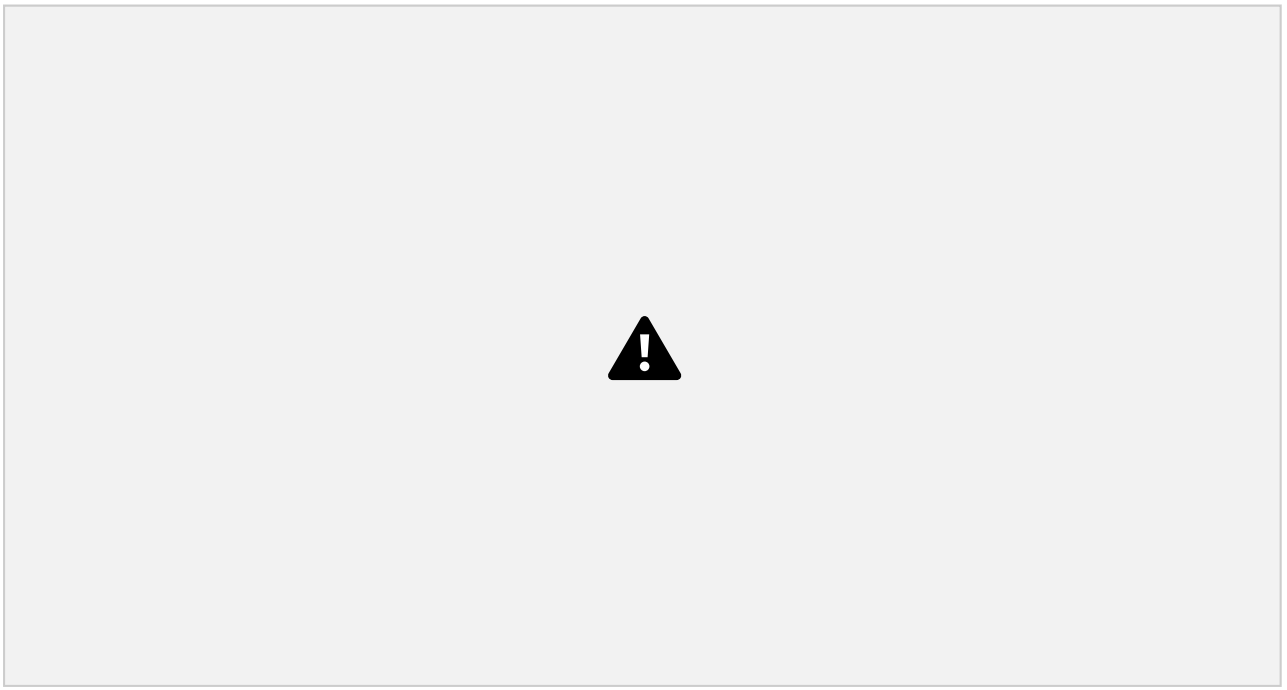


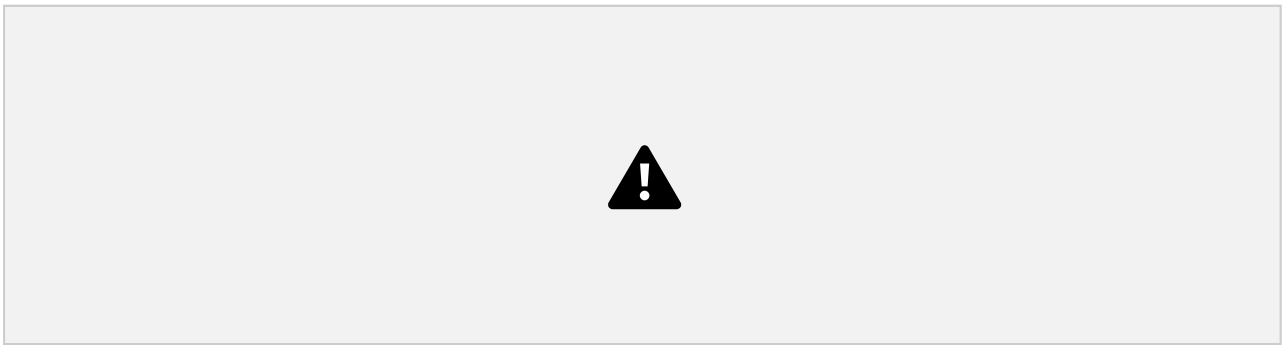










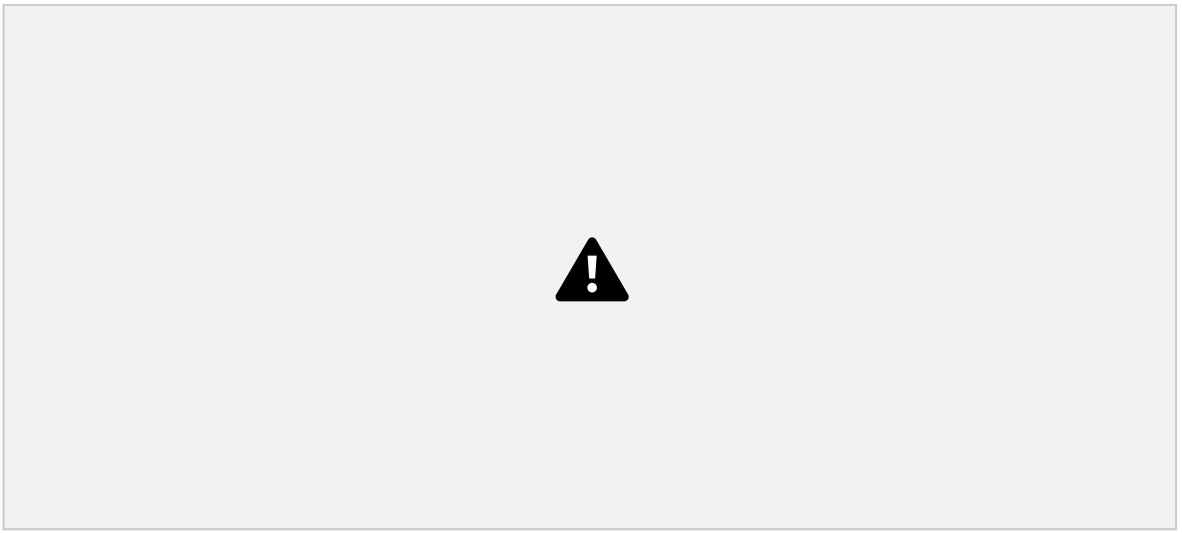


Atividades:

Considere os dados dos exercícios 1 e 2. Faça a comparação dos dois grupos escolhendo o teste mais adequado.

1) Variável quantitativa = Frequência cardíaca

Variável qualitativa = Fumo 1=sim; 0 =não



2) Variável quantitativa = escore de satisfação corporal (variação -1 a 6)