Introdução a Métodos Estatísticos para a Bioinformática

Profa. Júlia Maria Pavan Soler

pavan@ime.usp.br

IBI 5086 – Bioinformática - IME/USP

2º Sem/2023

Programa

- Revisão

- Estrutura de Dados e Estatísticas Descritivas
- Álgebra linear básica: cálculo matricial, determinantes, sistemas lineares, produto interno, norma, ortogonalidade, autovalores e autovetores

1. **Comparação de Grupos** (2 ou mais): Testes Clássicos (teste t, Wilcoxon, modelosANOVA e Regressão), Testes de Aleatorização, Comparações Múltiplas, Modelos de Efeitos Fixos e Aleatórios (Modelos Mistos), Análise de Tabelas de Contingência (Testes Qui-Quadrado), Regressão Logística.
2. **Análise Multivariada de Dados**: Componentes Principais, Análise Discriminante e Classificação, Agrupamento, Correlação Canônica, modelos MANOVA
3. **Métodos Computacionalmente Intensivos**: Simulação de Monte Carlo, Intervalos de Confiança Bootstrap, Testes de permutação e Aleatorização

Atividade

Pontue o seu conhecimento (domínio) sobre os seguintes conteúdos relacionados à Estatística:

Mentimeter

Acesse www.menti.com
Use o código: ??????

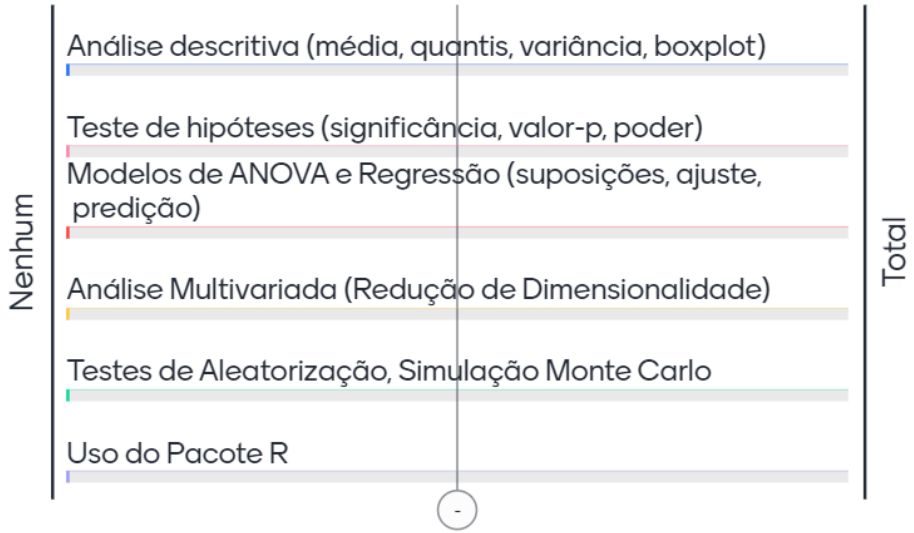
Atividade

Mentimeter

Acesse menti.com e use o código 3892 2599

Mentimeter

Pontue o seu conhecimento (domínio) sobre os seguintes conteúdos relacionados à Estatística:

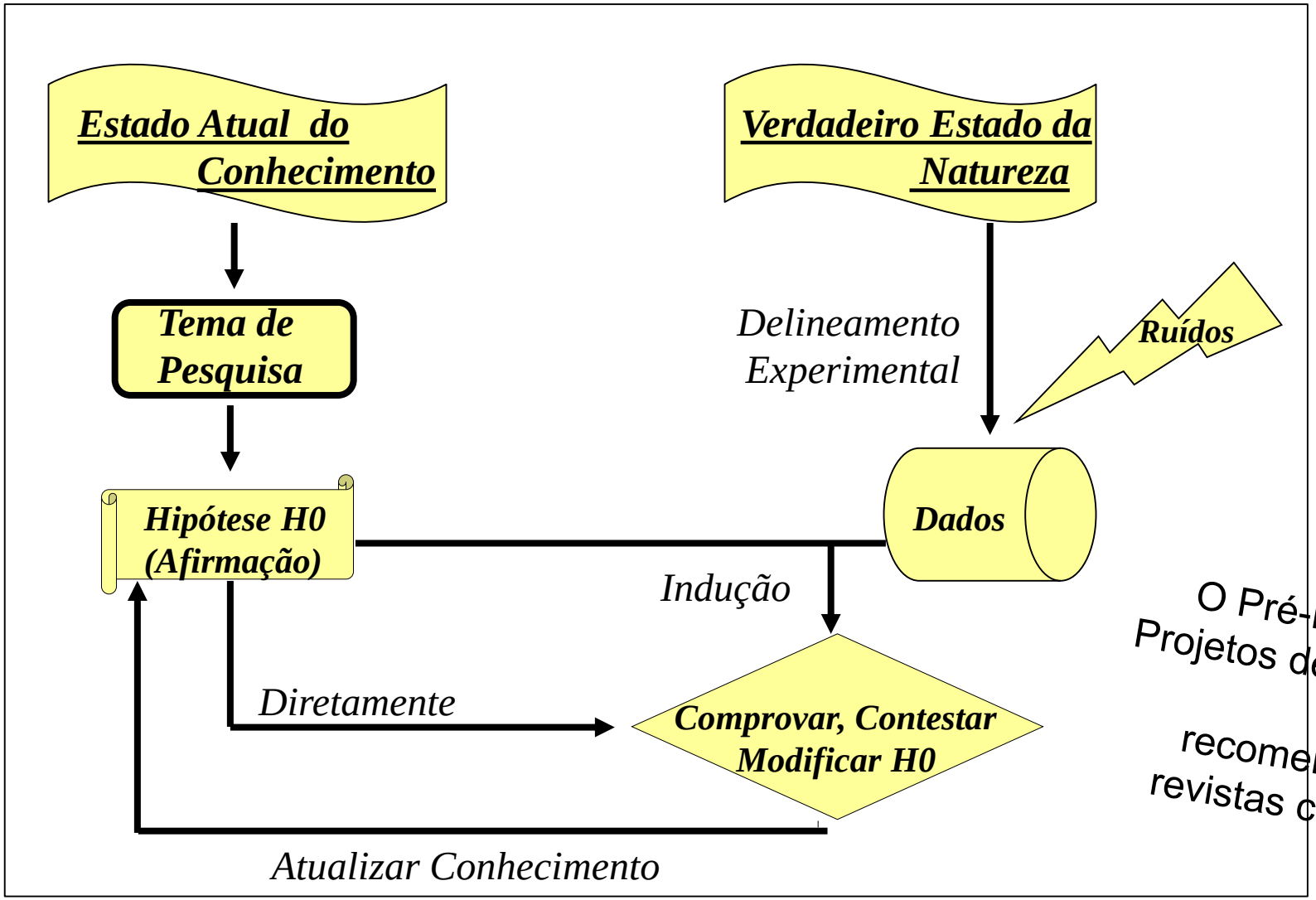


https://www.menti.com/al58sqnuscdg

Motivação

- Por que coletamos dados?
- Como coletar os dados?
- É preciso caracterizar a estrutura dos dados que estamos coletando, gerando ou analisando
- O que estamos fazendo com os dados que coletamos?
- Os resultados obtidos são robustos e reprodutíveis?
- Qual é o papel da Estatística na Pesquisa Científica?

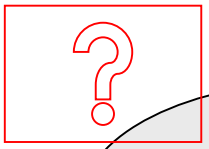
Estatística e Conhecimento Científico



O Pré-Registro de Projetos de Pesquisa tem sido recomendado por revistas científicas!

Box, Hunter and Hunter, 1978.

Estatística e Conhecimento Científico



Como fazer Ciência?

Observar fatos ⟶ **Criar Teorias**

Investigação Científica

- **Não existe um procedimento único**
- **Não é uniforme mas é convergente**

Assegurar Convergência

- **Hipóteses bem elaboradas**
- **Plano amostral e experimental eficientes**
- **Análise de dados sensível e legítima**

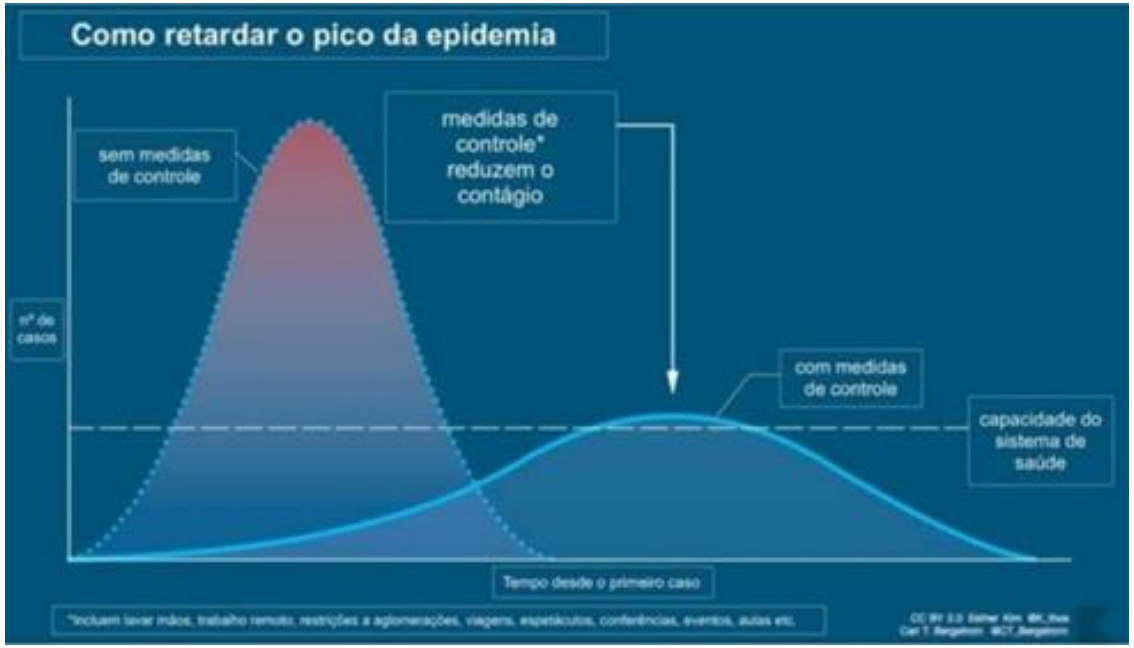
A Estatística no Método Científico e na nossa vida cotidiana!

2020: Epidemia causada pelo Coronavírus
Mudanças inestimáveis, em nível mundial, em diferentes setores da sociedade (Saúde, Educação, Família, Comunicação, Economia, etc.)

Grande valorização e popularização do Método Científico, com reconhecimento do papel fundamental da Estatística (na tomada de decisões).

O uso da Estatística buscando minimizar as consequências da Pandemia ao Sistema de Saúde (Março, 2020)

Monitoramento da Curva de Contágio do Coronavirus



BOAS PRÁTICAS

Para prevenir novas crises

Documento propõe mudanças em práticas de pesquisa a fim de reduzir a publicação de artigos cujos resultados ninguém consegue repetir

A Real Academia de Artes e Ciências da Holanda lançou um documento propondo mudanças em práticas de pesquisa a fim de enfrentar o que se convencionou chamar de "crise da reprodutibilidade" – uma sucessão de casos de artigos científicos que caíram em descrédito por seus resultados não terem sido confirmados em experimentos subsequentes. As recomendações do relatório divulgado em janeiro, intitulado *Estudos de replicação – Melhorando a reprodutibilidade nas ciências empíricas*, buscam aumentar o rigor com que o trabalho científico é realizado e apoiar pesquisadores interessados em verificar resultados obtidos por colegas. Uma das propostas apresentadas consiste em estimular o financiamento de estudos voltados para ratificar outros estudos, seguindo o exemplo da Organização Holandesa para Pesquisa Científica (NWO), que destinou no ano passado € 3 milhões a um programa-piloto para projetos dessa natureza. As sugestões também incluem reforçar o treinamento de cientistas e estudantes em tópicos como desenho de experimentos e análise estatística, e incentivar periódicos científicos a publicar pesquisas que não confirmaram as hipóteses testadas ou então chegaram a resultados nulos.

"O conhecimento só avança se os pesquisadores puderem confiar nos resultados de estudos feitos anteriormente", escreveu, no prefácio do relatório, a presidente da Real Academia, José van Dijck, pesquisadora de mídia e cultura da Universidade de Utrecht. Na avaliação da entidade, produzir dados fidedignos é essencial para evitar o desperdício de recursos na pesquisa e garantir a confiança do público na ciência. "O relatório conclui que estudos para reproduzir outros estudos devem ser conduzidos de forma mais frequente e sistemática, o que exige um esforço conjunto de agências de fomento, pesquisadores, periódicos e instituições científicas", afirmou Van Dijck.

O panorama da "crise da reprodutibilidade" apresentado no relatório mostra a relevância do problema. À procura de novos medicamentos contra o câncer, a empresa farmacêutica Amgen tentou confirmar os achados de 53 estudos pré-clínicos publicados que pareciam ter grande potencial. Apenas 11% dos resultados foram corroborados. A Bayer fez um esforço semelhante para tentar validar dados sobre alvos potenciais de novos remédios, obtidos por 67 projetos de pesquisa, e só teve sucesso em 25% dos casos. Uma colaboração internacional para investigar estudos de

ILUSTRAÇÃO ROBILLO

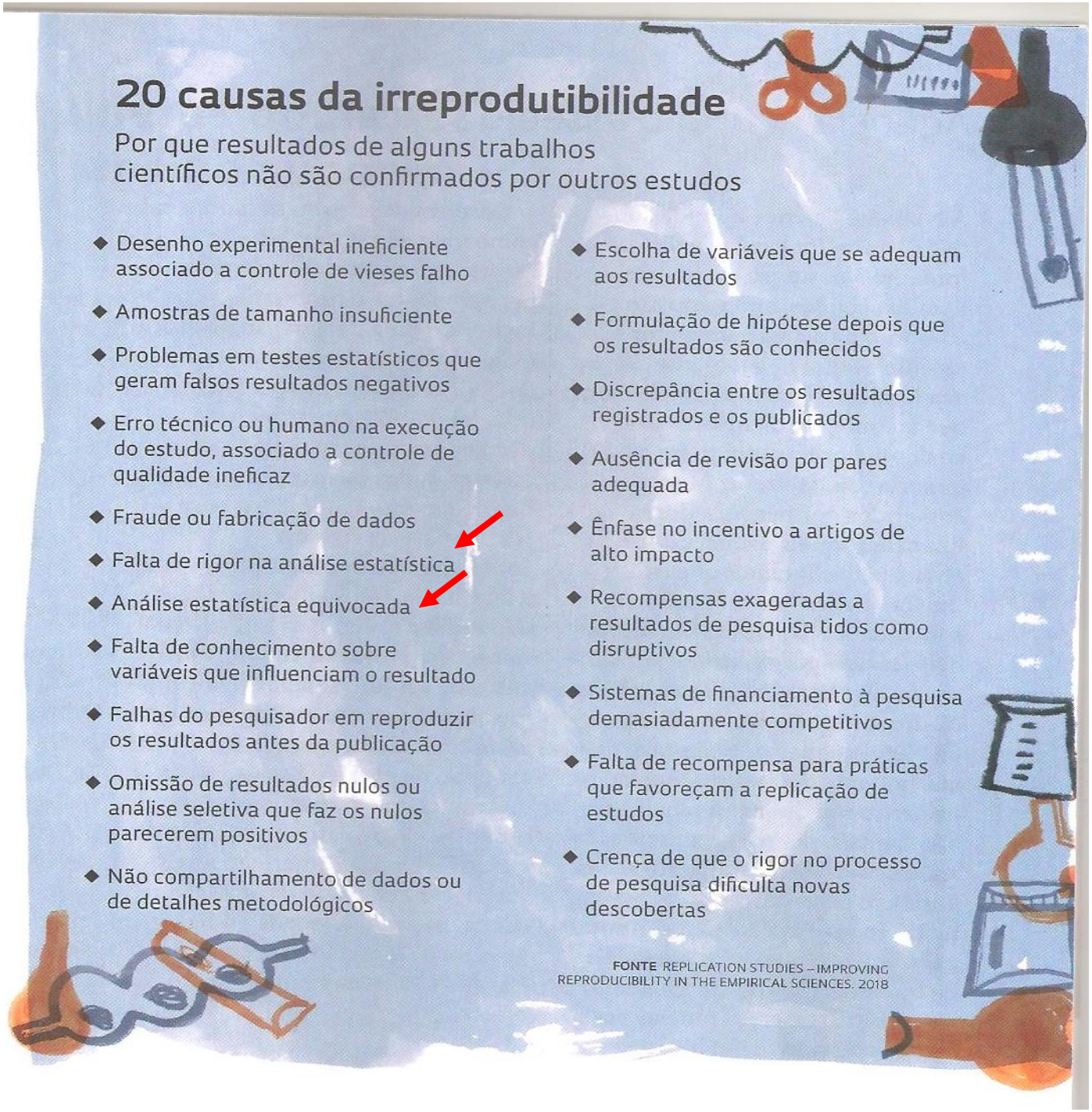
Mas a Aplicação da Estatística deve ser feita de forma apropriada, usando

BOAS PRÁTICAS,

para garantir

REPRODUTIBILIDADE NAS PESQUISAS.

Revista FAPESP
Março/2018



Leia também:

Artigo da BBC-News (Feb/2019): ‘Machine learning causing science crisis’

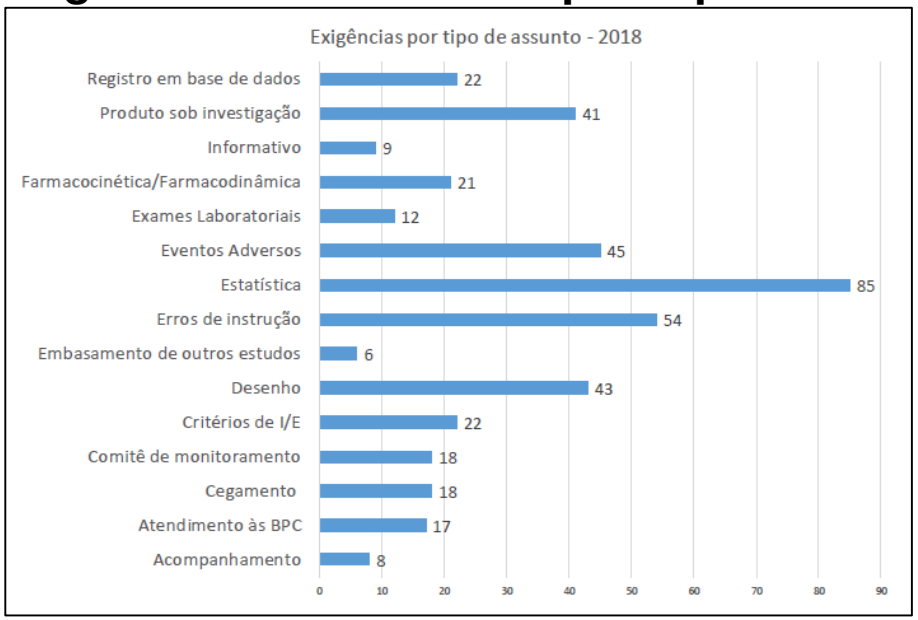
Artigo da Nature: ‘Six factors affecting reproducibility in life science research and how to handle them’

→ Desenho amostral e experimental pobre
→ Inabilidade de processor, analisar e interpretar dados complexos

A Estatística no Método Científico
-Estudos Clínicos-

ANVISA

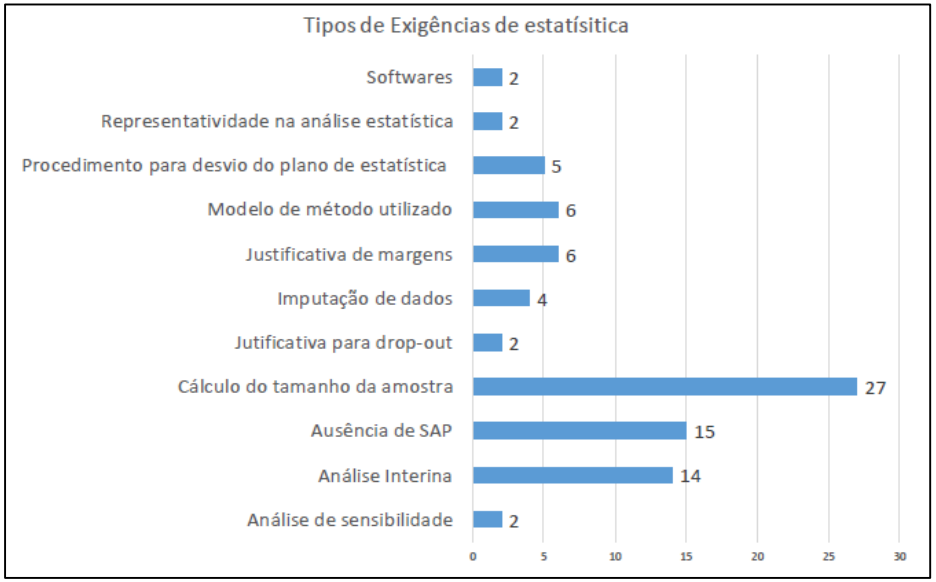
Exigências de Protocolo por Tipo - 2018



426 EXIGÊNCIAS

20% - Estatística
12,7% - Erros de Instrução
10,6% – Eventos Adversos
10,1% - Desenho
9,6% - Produto sob Investigação

Protocolo – Exigências de Estatística



85 EXIGÊNCIAS

31,8% - Cálculo do tamanho da amostra
17,6% - Ausência de SAP
16,5% – Análise Interina
7,1% - Modelo de Método Utilizado
7,1% - Justificativas de margens

A Estatística no Método Científico
-Atualização do Conhecimento-

ASA AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics

The ASA's Statement on p-Values: Context, Process, and Purpose

Ronald L. Wasserstein American Statistical Association

The American Statistician (2015), 70(2):129-133
DOI:10.1080/00031305.2016.1154108#tabModule

Of course, it was not simply a matter of responding to some articles in print. The statistical community has been deeply concerned about issues of *reproducibility* and *replicability* of scientific conclusions. Without getting into definitions and distinctions of these terms, we observe that much confusion and even doubt about the validity of science is arising. Such doubt can lead to radical choices, such as the one taken by the editors of *Basic and Applied Social Psychology*, who decided to ban *p*-values (null hypothesis significance testing) (Trafimow and Marks 2015). Misunderstanding or misuse of statistical inference is only one cause of the "reproducibility crisis" (Peng 2015), but to our community, it is an important one.

Recomendações de Leituras

Benjamin, DJ; Berger, JO. (2019) Three Recommendations for Improving the Use of *p*-Values, The American Statistician, 73:sup1, 186--91, DOI: 10.1080/00031305.2018.1543135

✓ Box, GEP; Hunter, WG; Hunter, JS. (1978) Statistics for Experimenters. John Wiley & Sons (ver 2th ed., 2006). "Cap. 3" (2006)

Cooper, RA. (2019). Making Decisions with Data: Understanding Hypothesis Testing & Statistical Significance. The American Biology Teacher 81(8): pp. 535–542.

Cowles, M; Davis, C. (1982). On the Origins of the .05 Level of Statistical Significance. *American Psychologist* 37(**5**): pp. 553-558.

✓ Oehlert, GW. (2010) A first course in Design and Analysis of Experiments. Univ. of Minnesota, Licensed by Creative Commons. "Caps. 1 e 2"

Wasserstein, RL; Lazar, NA. (2016) The ASA's statement on p-values: context, process, and purpose. The American Statistician, DOI: 10.1080/00031305.2016.1154108

Wasserstein, RL; Lazar, NA. (2019) Moving to a World Beyond "$p < 0.05$", The American Statistician, 73 (1): 1-19 DOI: 10.1080/00031305.2019.1583913

Tipos de Estudos

Estudos Observacionais

A população sob estudo é amostrada sem que haja intervenção, nenhum fator está sendo controlado → Estudos epidemiológicos: Prospectivo, Retrospectivo e Transversal

- Casos de Covid19 no município
- Distribuição de um CNV em pacientes

Categoria do CNV	Frequência
-2	10
-1	50
0	340
1	120
2	30

Estudos Experimentais

Estudos de intervenção: há fatores sendo controlados (causas, var. independentes) para se obter os efeitos (var. dependentes) → **Ensaios Clínicos Controlados e Aleatorizados** (padrão ouro em estudos de causalidade)

- Efeito do tratamento e dose no tempo de cura (resposta y)

Trat	T1		T2		T3	
Dose	Baixa	Alta	Baixa	Alta	Baixa	Alta
	_	_		_	_	_
	_	_	y	_	_	_
	_	_		_	_	_

Tipos de Estudos

Estudos Observacionais	Estudos Experimentais
▪ Estudos Prospectivos ▪ Estudos Retrospectivos (Caso-Controle) ▪ Estudos Transversais	▪ Ensaios Clínicos Controlados e Aleatorizados ▪ Ensaios de Campo ▪ Ensaios de Bancada

Em geral, os dados desses estudos podem ser dispostos em ***Tabelas de Contingência*** para análises de associação (testes Qui-Quadrado, modelos de regressão logística)

Ex. GWAS (Genome Wide Association Studies)

Em geral, esses dados são analisados por meio de modelos de ***Análise de Variância (ANOVA) ou Regressão***.

Ex. Experimentos de Microarrays

Estudos Observacionais

Estudo Transversal (**N fixado**)

	Fator de Risco		
Doença	**-**	**+**	**Total**
D	37	80	117
ND	45	38	83
Total	82	118	**200**

$\chi^2 = 10.25$, **p-value = 0.0014**

Estudo Prospectivo (**n- e n+ fixados**)

	Fator de Risco		
Doença	**-**	**+**	**Total**
D	45	70	115
ND	55	30	85
Total	**100**	**100**	200

$\chi^2 = 12.79$, **p-value = 0.0003**

Estudo Retrospectivo

(Caso-Controle) (**n_D e n_{ND} fixados**)

	Fator de Risco		
Doença	**-**	**+**	**Total**
D	38	62	**100**
ND	67	33	**100**
Total	105	95	200

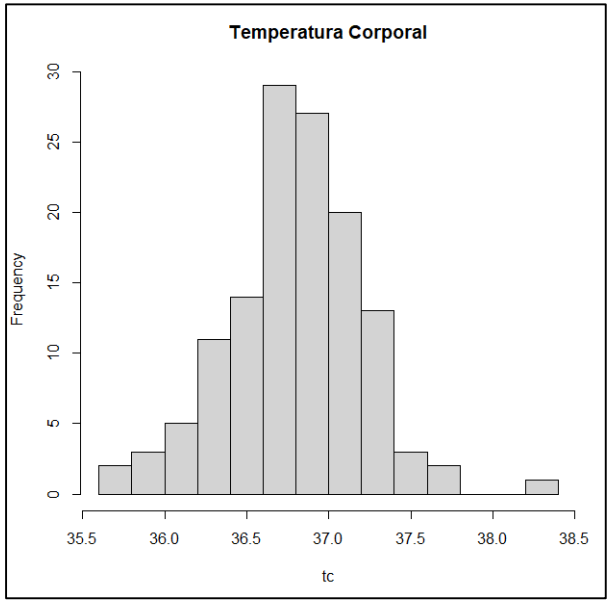
$\chi^2 = 16.86$, **p-value = 0.00004**

Estudo Observacional

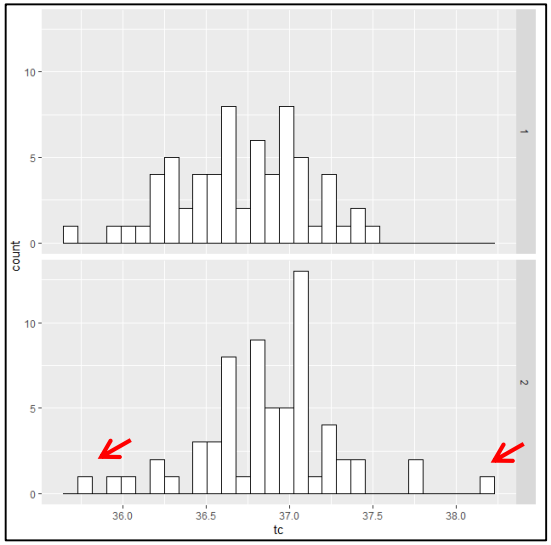
Dados de Temperatura Corporal (pacote R)

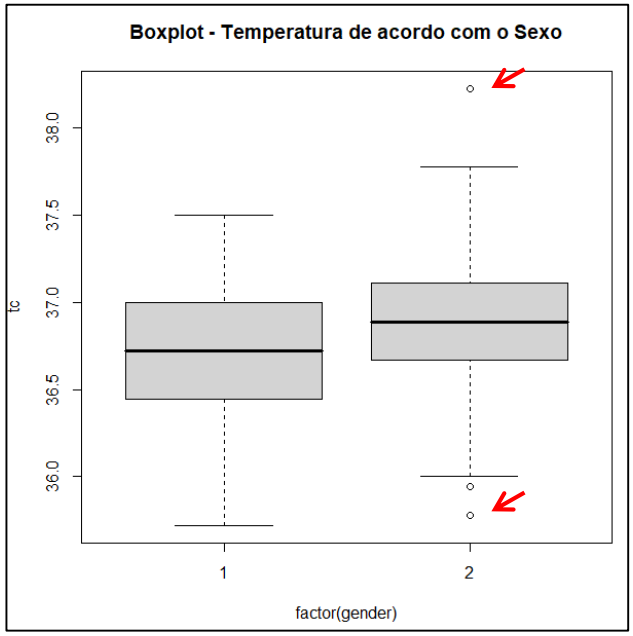
Dados simulados gerados a partir dos dados reais em Mackowiak, P. A., Wasserman, S. S., and Levine, M. M. (1992), "A Critical Appraisal of 98.6 Degrees F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich," *Journal of the American Medical Association*, 268, 1578-1580.

	temperature	gender	hr	tc
1	96.3	1	70	35.72222
2	96.7	1	71	35.94444
3	96.9	1	74	36.05556
...				
64	99.4	1	70	37.44444
65	99.5	1	75	37.50000
66	96.4	2	69	35.77778
67	96.7	2	62	35.94444
...				
129	100.0	2	78	37.77778
130	100.8	2	77	38.22222



Histograma da Temperatura de acordo com sexo





Var	n	média	dp
tc	130	36.81	0.407

Alguns resultados descritivos. Comente!

Estudo Observacional

Dados de Temperatura Corporal:

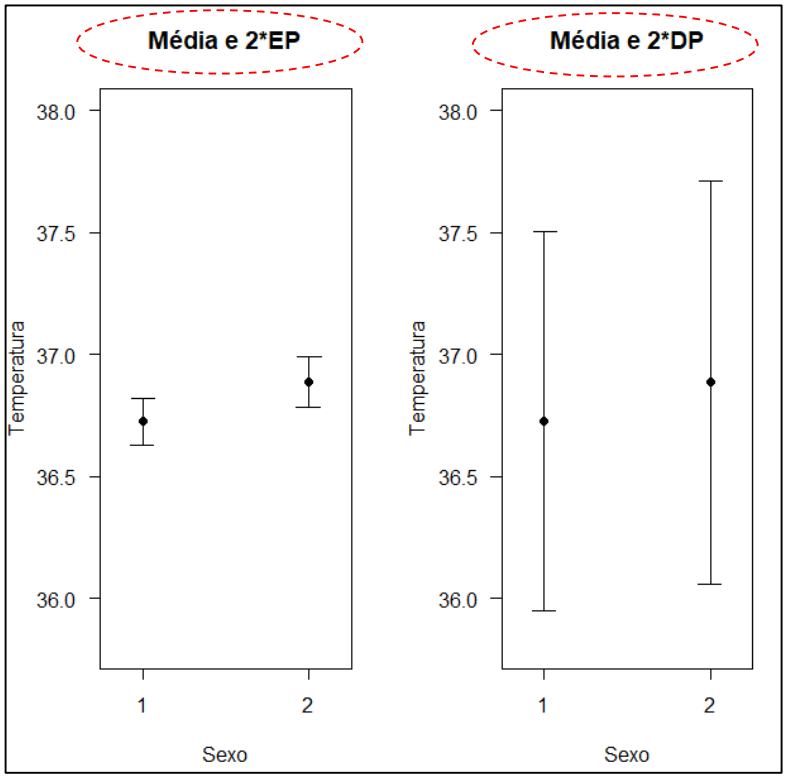
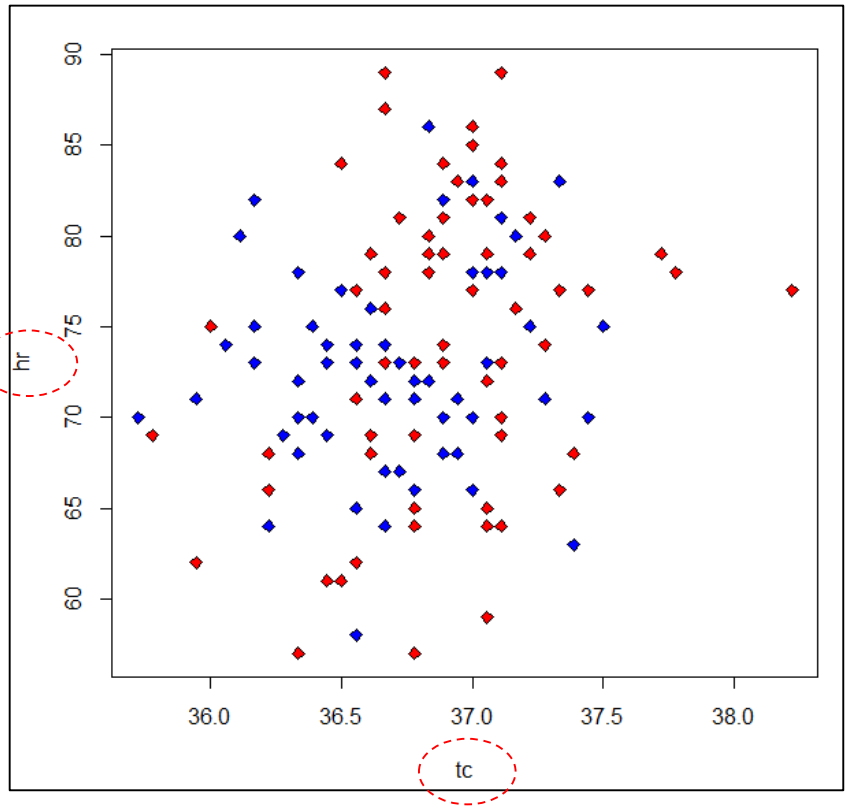


Gráfico de dispersão



Var	Sex	n	média	dp	mediana	ep
tc	1	65	36.72	0.39	36.72	0.05
tc	2	65	36.89	0.41	36.89	0.05

Interprete estes resultados descritivos!
Interprete os intervalos: Média±2DP
Média±2EP

Estudo Observacional - Dados Simulados

Dados de câncer de ovário foram simulados a partir do projeto TCGA

https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

As unidades amostrais correspondem a pacientes com tempo de sobrevida inferior a 3 anos (casos – outcome = 1) ou superior a 3 anos (controle – outcome = 0). Os dados foram gerados de acordo com o modelo causal na Figura ao lado.

Foram simulados dados para 1000 pacientes, sendo 500 casos e 500 controles. Estão disponíveis 50 réplicas desse cenário de simulação.

Para as mesmas unidades amostrais, estão disponíveis informações de 4 bancos de dados: CNV Metilação, Expressão Gênica e Proteína.

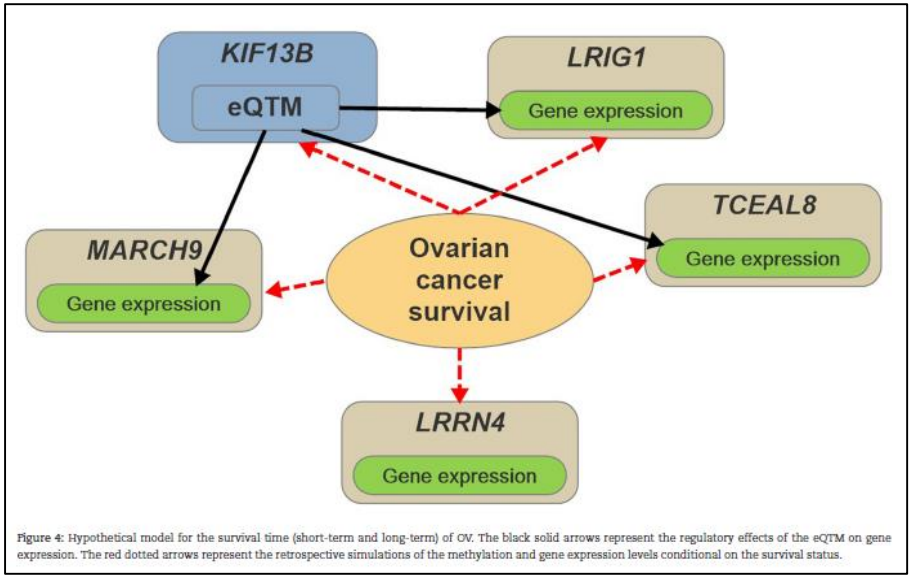


Figure 4: Hypothetical model for the survival time (short-term and long-term) of OV. The black solid arrows represent the regulatory effects of the eQTM on gene expression. The red dotted arrows represent the retrospective simulations of the methylation and gene expression levels conditional on the survival status.

Referência : Ren Hua Chung and Chen Yu Kang. A multi-omics data simulator for complex disease studies and its application to evaluate multi-omics data analysis methods for disease classification. *Giga Science* 8(5): 1–12, 2019. ISSN 2047217X. doi: 10.1093/gigascience/giz045.

No link (https://figshare.com/articles/Ovarian_cancer_profile_for_OmicsSIMLA/7763627) é posível baixar os dados.

No link (https://omicssimla.sourceforge.io/simuomicsTCGA.html) há uma explicação do que significa cada coluna de cada banco de dados.

Estudo Observacional – Dados Simulados

Dados de câncer de ovário foram simulados a partir do projeto TCGA
https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

- **CNV**: variação no número de cópias para 2884 regiões. As variáveis estão codificadas como -2, -1, 0, 1 e 2. Os valores negativos indicam a perda de duas ou uma cópia da região cromossômica, os valores positivos indicam o ganho de duas ou uma cópia e o valor nulo indica que a região cromossômica é normal.

- **Exp**: intensidade de expressão gênica dos genes LRIG1, TCEAL8, MARCH9, LRRN4 e 2000 outros.

- **NorExp**: intensidade de expressão gênica normalizados, em que foi eliminado ruídos aleatórios.

- **Methy**: metilação de 2752 locais cromossômicos, além do eQTM. Os dados indicam o percentual de metilação em cada local.

- **Protein**: valores da expressão proteica normalizada para os mesmos genes em que foi avaliada a expressão.

Discuta:

1. **Estrutura desses dados:** tipo de estudo, tamanho amostral, natureza das unidades amostrais, variáveis envolvidas
2. **Objetivos da pesquisa**
3. **Possíveis análises estatísticas** (descritiva, inferencial e preditiva)

Estudo Experimental

Exemplo: Avaliar a eficácia de 3 Medicamentos (A, B e C) para cefaleia.

Variável resposta: tempo (min) até o alívio da dor

Amostra: 15 voluntários, que atenderam aos critérios de inclusão do estudo, foram randomizados para um de 3 tratamentos (A, B e C).

Dados:

	Trat A	Trat B	Trat C
	24,5	28,4	26,1
	23,5	34,2	28,3
	26,4	29,5	24,3
	27,1	32,2	26,2
	29,9	30,1	27,8
Média	**26,28**	**30,88**	**26,54**
DP	**2,48**	**2,31**	**1,58**

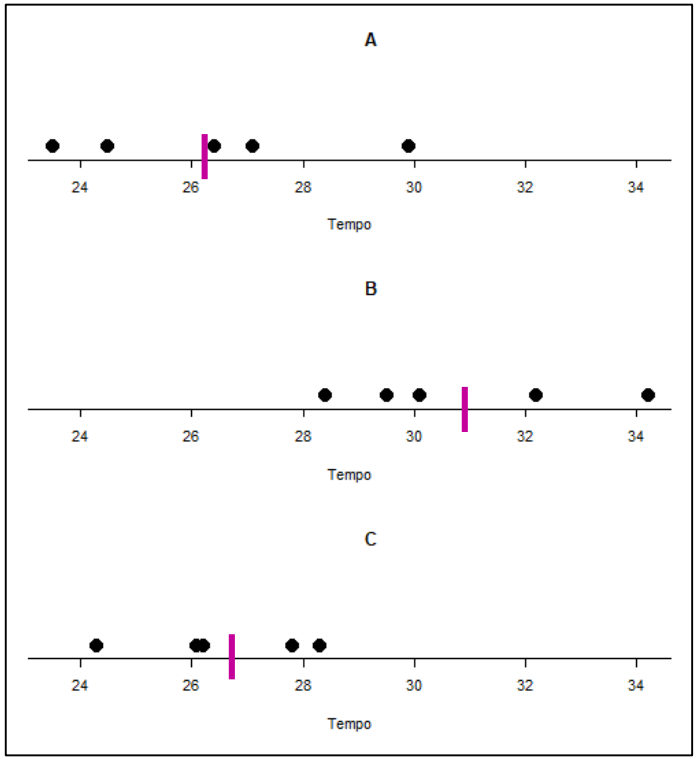


Gráfico de pontos

Hipóteses estatísticas para a Média da Resposta:

H_0: $\mu_A = \mu_B = \mu_C$

H_1: $\mu_i \neq \mu_j$ para i,j=A, B ou C

Estudo Experimental

Dados Pulse: Avaliação da pulsação de 92 estudantes na condição inicial em repouso → P1

Os estudantes foram aleatoriamente distribuídos (jogaram uma moeda: cara=submetidos a corrida (Grupo Ran1); coroa=permaneceram em repouso (Grupo Ran2). A pulsação dos estudantes foi novamente obtida → P2

	P1	P2	Ran	Fu	Sex	Altura	Peso	Ativ
1	64	88	1	2	1	66.00	140	2
2	58	70	1	2	1	72.00	145	2
…								
34	62	98	1	1	2	62P15	112	2
35	80	128	1	2	2	68.00	125	2
36	62	62	2	2	1	74.00	190	1
37	60	62	2	2	1	71.00	155	2
…								
91	86	84	2	2	2	67.00	150	3
92	76	76	2	2	2	61.75	108	2

	Sex		
Ran	1	2	Total
1	24	11	35
2	33	24	57

Var	Ran	n	mean	sd	median	se
P1	1	35	73.6	11.44	70	1.93
	2	57	72.42	10.82	72	1.43
P2	1	35	92.51	18.94	88	3.20
	2	57	72.32	9.95	70	1.32

Discuta:

1. **Estrutura desses dados:** tipo de estudo, tamanho amostral, variáveis sob estudo, unidades amostrais são dependentes ou independentes?
2. **Possíveis objetivos da pesquisa**
3. **Possíveis análises estatísticas** (descritiva, inferencial e preditiva)

Estudo Experimental – Dados Pulse

Há indicação de efeito da corrida na pulsação?

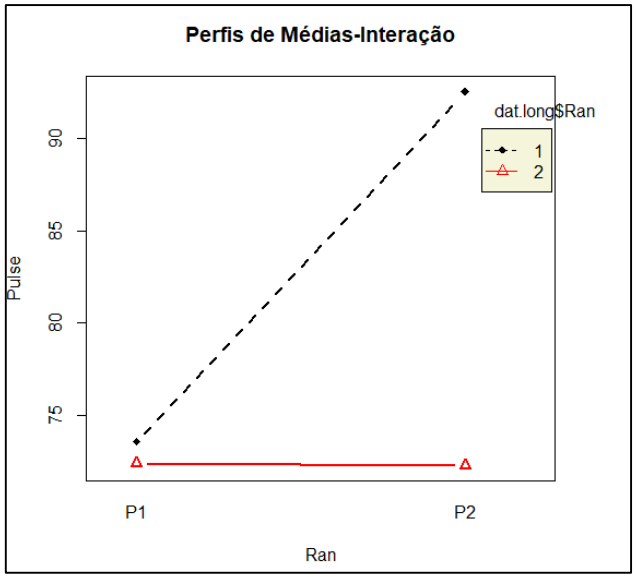
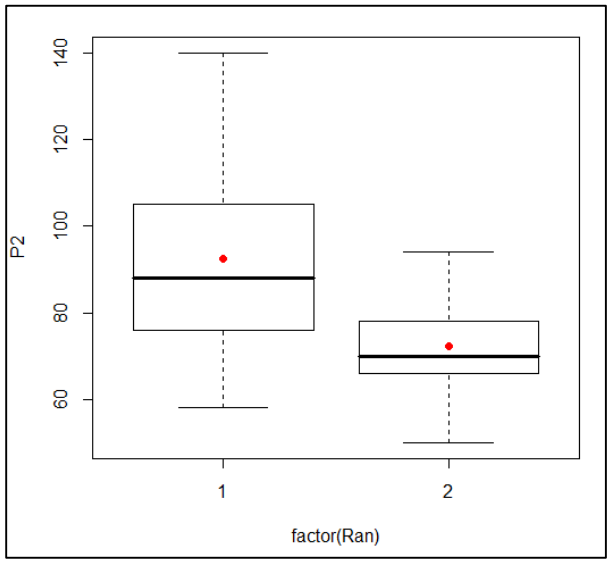


Gráfico com os perfis médios de Pulse de acordo com a condição de avaliação e RAN (amostras dependentes)



Boxplot de P2 (amostras independentes)

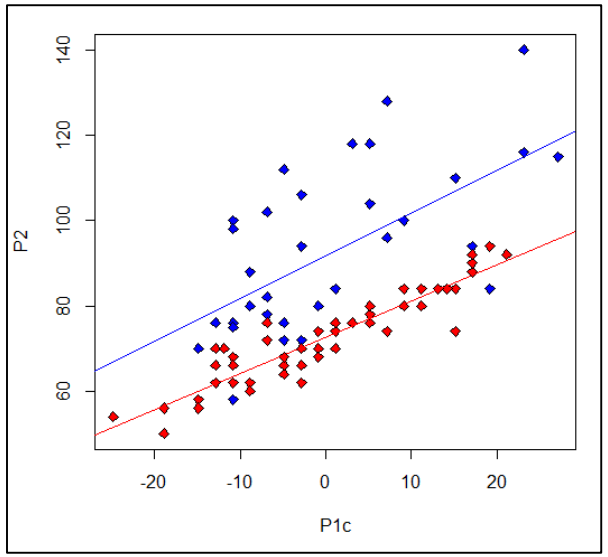


Gráfico de dispersão de P1 e P2 de acordo com RAN (amostras dependents)

Estudo Experimental

Dados Pulse - Convertendo o format dos dados

Formato "wide"

	P1	P2	Ran	Fu	Sex	Altura	Peso	Ativ
1	64	88	1	2	1	66.00	140	2
2	58	70	1	2	1	72.00	145	2
…								
34	62	98	1	1	2	62.75	112	2
35	80	128	1	2	2	68.00	125	2
36	62	62	2	2	1	74.00	190	1
37	60	62	2	2	1	71.00	155	2
…								
91	86	84	2	2	2	67.00	150	3
92	76	76	2	2	2	61.75	108	2

Formato "long"

	Ran	Fu	Sex	Altura	Peso	Ativ	variable	value
1	1	2	1	66.00	140	2	P1	64
2	1	2	1	72.00	145	2	P1	58
3	1	1	1	73.50	160	3	P1	62
4	1	1	1	73.00	190	1	P1	66
…								
180	2	1	2	64.00	125	1	P2	92
181	2	2	2	68.00	133	1	P2	80
182	2	2	2	62.00	110	2	P2	68
183	2	2	2	67.00	150	3	P2	84
184	2	2	2	61.75	108	2	P2	76

Estudos Clínicos para Aprovação de Medicamentos (Vacinas)

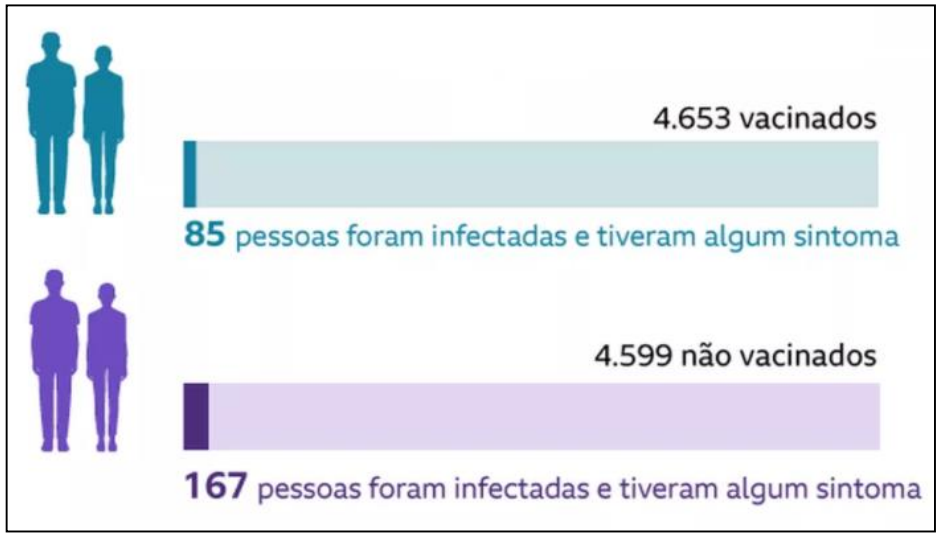
Estudo COV-02-IB - Vacina Coronavac: Submetido à ANVISA para Aprovação Emergencial

Ensaio Clínico fase III duplo-cego, randomizado, controlado com Placebo para Avaliação de Eficácia e Segurança em Profissionais da Saúde da Vacina COVID-19 (Inativada) produzida por Sinovac – Instituto Butantan

⇒ Envolveu 16 centros de pesquisa brasileiros

⇒ Início: 21/07/2020 Previsão de término: Outubro/2021

⇒ Data da primeira coorte: 16/12/2020.



Com base nestes resultados há evidência da Eficácia da Vacina?

Estudos Clínicos para Aprovação de Medicamentos (Vacinas)

Eficácia da Vacina e Razão de Risco (RR) – Covid-19

- Razão de Risco(RR) mede a incidência de casos de Covid-19 no braço da vacina em relação ao placebo.

$$RR = \frac{85/4653}{167/4599} = 50,31\%$$

?

- Eficácia da Vacina (VE) = 1 - RR

$$1 - RR = 49,69\%$$

tvBrasil

CORONAVÍRUS • COVID-19 • VACINA

Resultados publicados do Estudo:

Resultado Primário de Eficácia

Eficácia total (conforme desfecho primário): **50,39% (IC 95 : 35,26 – 61,98)***

*p=0,0049

Hazard Ratio (HR)

Inclui o Tempo e Covariáveis (Centro)

- Modelo de Regressão de Cox

$$\frac{h(t)}{h_0(t)} = \exp\left(b_1X_1 + b_2X_2 + \ldots b_pX_p\right)$$

- Eficácia da Vacina = 1 - HR

Atividade

Para discussão do Estudo Clínico da Coronavac:

1. **Qual população foi amostrada?**
2. É um estudo observacional ou experimental? Justifique.
3. Quem é a unidade amostral?
4. Quem é a unidade experimental (aquela que recebeu o tratamento)?
5. Quem é a unidade de mensuração da resposta de interesse?
6. Qual é a estrutura dos tratamentos?
7. Houve aleatorização? Como deve ter sido realizada?
8. Qual é o desfecho primário do estudo (resposta de interesse)?
9. Como deve ter sido definido o tamanho amostral dos grupos?
10. Como avaliar o efeito do tratamento (eficácia da vacina)? Há variáveis de controle (reduzir confundimentos, vícios)?
11. **As inferências são válidas para qual população?**

Planejamento de Experimentos

- Experimentos na Agricultura - Fisher (1920s): Princípios da experimentação (aleatorização, replicação, blocagem)
- Experimentos Industriais - Box e Wilson (1950s), Box (1999): Modelos de superfície de respostas
- Experimentos para o Controle da Qualidade industrial – Taguchi (1980s): processos robustos, resistentes a variações transmitidas de componentes, Modelos fatoriais fracionais, Gráficos de Controle
- Ensaios Clínicos Controlados e Aleatorizados (90's) - ICH-Guidelines on General Considerations dor Clinical Trials: agências reguladoras da Indústria Farmacêutica (FDA-USA, EMA-Europa, ANVISA-Brasil)
- ...
- Planejamento de Experimentos na era Omics: Experimentos em Microarrays (Bioconductor-R); Experimentos em Proteômica (Espectrometria de Massas, MSstats-R, MixOmics), etc.

Planejamento de Experimentos

Por que PLANEJAR ?

Controlar e Evidenciar o efeito de Fontes de Variação (FV) conhecidas

Reduzir o efeito de FV Desconhecidas

modelo estatístico

$$Y = FV\ conhecidas + e$$

Variável resposta de interesse

Fontes de Variação conhecidas (variáveis preditoras X)

Erro (fontes de variação desconhecidas)

Atenção à estrutura dos dados:

- Tipo da variável resposta Y
- Tipo da variáveis X (fatores de interesse, covariáveis)
- Erro: é a fonte ALEATÓRIA em Y, traz informação sobre a falta de ajuste
- Y, X e e são avaliadas em unidades amostrais, em geral supostas independentes.

Modelo Estatístico

$$Y = FV\ conhecidas + e$$

Princípio do Planejamento: Planejar para ser capaz de mensurar a "importância" (significância estatística) de possíveis fontes de variação conhecidas (de significância factual) e reduzir a influência de fontes de variação desconhecidas (desvio ou erro aleatório) que atuam sobre a resposta de interesse (Y).

A Estatística é uma lição de humildade, pois temos que compreender o significado do termo "erro"!

Planejamento de Experimentos

- População Alvo do estudo e Plano amostral
  - População de Interesse (N)
  - Amostra (n): unidades amostrais independentes ou correlacionadas
- Estruturas do Experimento
  - Tratamentos: Fatores de interesse
  - Unidades Experimentais: Aleatorização
  - Blocos e Covariáveis: variáveis de controle
- Variável resposta de interesse
  - Tipo de variável (unidade de medida, escala)
  - Réplicas (genuínas), replicatas
  - Tamanho amostral efetivo

Atenção: Matriz de Dados nxp
"n" (tamanho amostral) e "p" (variáveis avaliadas)
Clássico: n>p (?) Big-data: n<<"p"

Classificação de Variáveis
Modelo Estatístico

$$Y = f(X) + e$$

FV Conhecida
Componente Fixo

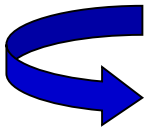
FV desconhecida
Componente Aleatório

Y: variável resposta (dependente, aleatória)

X: variáveis preditoras ou fatores de interesse (independente, em geral fixa)

f : função associando X a Y, em geral, linear e assumida conhecida, mas pode ser não linear ou desconhecida (*spline*)

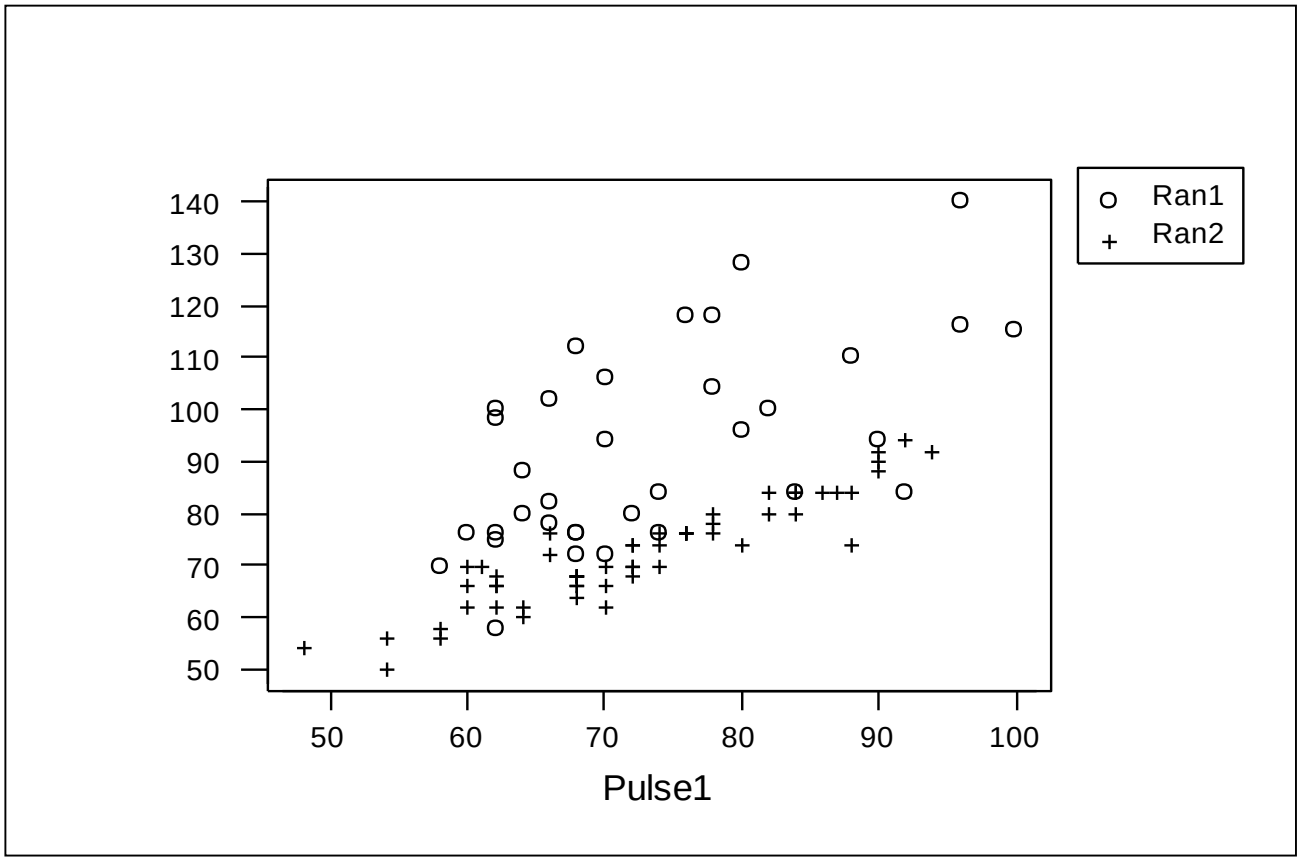
e : variável aleatória (em geral, envolve somente o termo de erro, mas pode ser decomposta em efeitos aleatórios de interesse no estudo + erro)



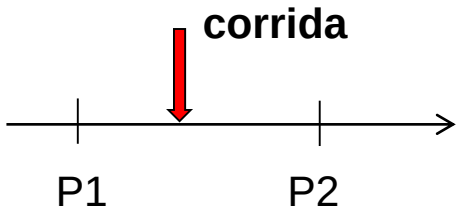
Discutiremos diferentes situações ao longo do curso!

Modelo Estatístico

$$Y = f(X) + e$$



Dados Pulse: a pulsação de 92 estudantes foi avaliada Antes de Depois de uma intervenção (corrida)
Ran1: corrida
Ran2: repouso



Pense em possíveis modelos estatísticos para avaliar o efeito da corrida na Pulsação!

Modelos Estatísticos - Efeitos Fixos

- **M1:** $Y = \beta_0 + \beta_1 X_1 + e$ Y = P2; X1=(P1-Média de P1)
- **M2:** $Y = \beta_0 + \beta_2 X_2 + e$ X2 = 0 se em repouso; = 1 se correu
- **M3:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$
- **M4:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \left(X_1 * X_2 \right) + e$

Em cada modelo, supondo que o erro aleatório tem média 0, qual é o valor esperado de Y (P2) para estudantes em repouso e para aqueles que correram?

Modelo Estatístico

$$Y = f(X) + e$$

Modelo de regressão com variável preditora *dummy*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$$

Y: P2 X_1 = (P1 – 73): P1corrigida pela média (=73 batimentos/minuto)

X_2: Ran (0 = em repouso, 1 = correu)

Betas: coeficientes de regressão

β_0: valor esperado de P2 para indivíduos em repouso (Ran=0) e com P1=73

β_1: desvio esperado em P2 para variações de 1 batimento em P1, considerando estudantes do mesmo grupo (em repouso ou que correram)

β_2: desvio esperado em P2 para indivíduos que correram, comparados àqueles em repouso, considerando que tenham a mesma P1.

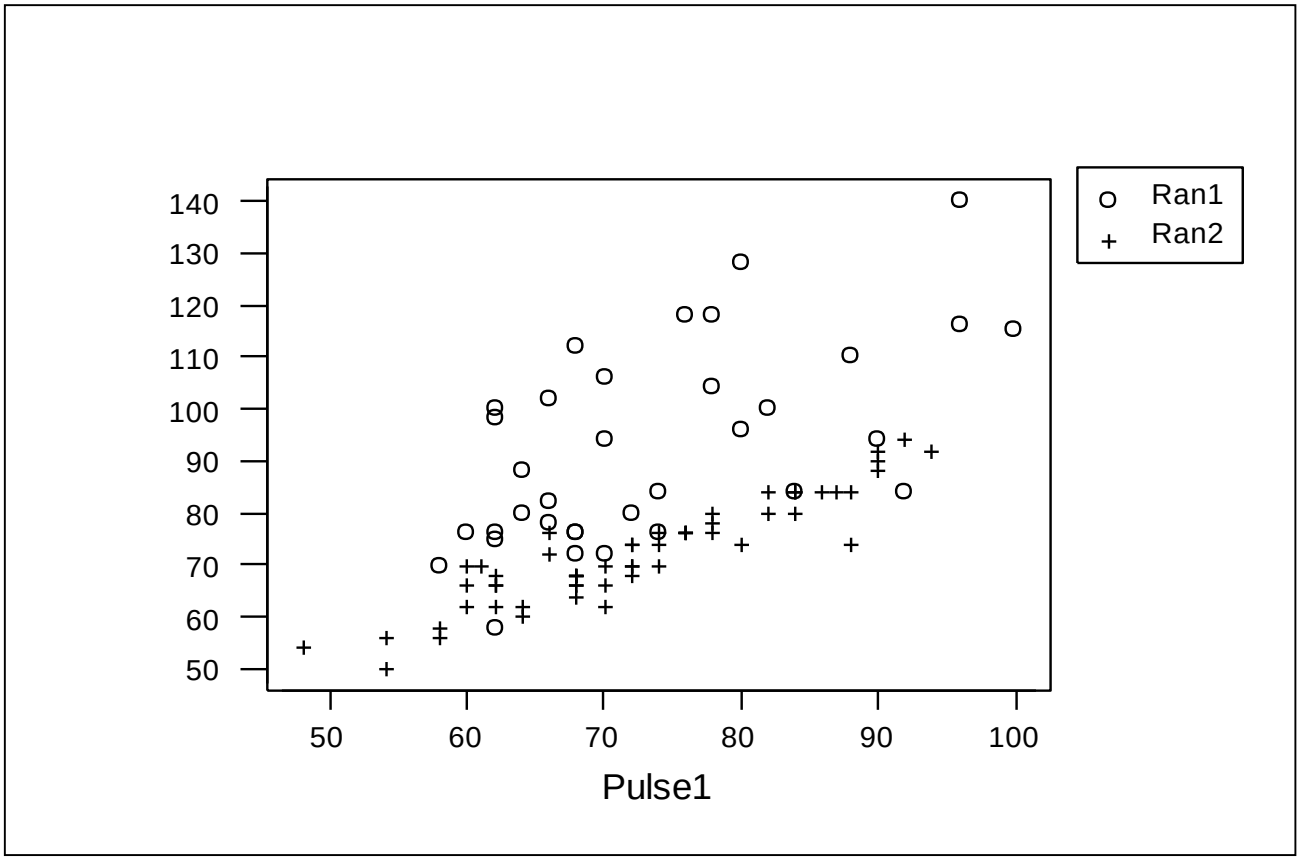
e: erro aleatório com media 0 e variância constante.

Ajustado para n=92 →

Modelo Ajustado

P2 = 72,7 + 0,912 (P1-73) + 19,1 Ran

Codificação: 0 1



Em repouso: E(P2|P1,Ran) = 72,7 + 0,912 (P1-73)

Corrida: E(P2|P1,Ran) = (72,7+19.1) + 0,912 (P1-73)

Classificação de Variáveis no Modelo Estatístico

$$Y = f(X) + e$$

Qualitativas	Nominais Ordinais
Quantitativas	Discretas Contínuas

Definir modelos para os dados de Pulsação.

Proponha formulações de modelos para os dados de câncer de ovário (TCGA)!

Modelos ANOVA: Y é variável resposta quantitativa univariada e X variável independente qualitativa (define o efeito de tratamentos, pode incluir covariáveis).
Modelos de Regressão Linear: Y é variável resposta quantitativa univariada e X variável independente quantitativa. MAS, nos modelos gerais, as variáveis preditoras X podem ser tanto quantitativas como categóricas.
Modelos de Regressão Linear Generalizado: no caso de regressão logística, Y é variável resposta binária. MAS, no caso geral, Y pode ser definida para outras escalas (multinomial, Poisson, etc.). As variáveis preditoras poder ser definidas como quantitativas, bem como categóricas.

Princípios do Planejamento de Experimentos e da Análise de Dados

O que queremos "EXTRAIR" dos dados?

→ Inferências válidas (sobre os efeitos de X em Y)

Como planejar um experimento para a coleta de dados?

→ Evitando erros sistemáticos (vícios de seleção, desgaste de aparelho, diferentes fornecedores, etc.) e controlando fontes de variação indesejáveis e conhecidas.

Contudo, como é possível controlar flutuações (aleatórias) nos dados que são devidas a fontes de variação desconhecidas?

→ **A priori (antes da coleta de dados) usar Aleatorização e Cegamento:** técnicas que evitam erros sistemáticos e servem para "balancear" o efeito do erro entre os grupos de tratamentos (gerar grupos comparáveis, com características semelhantes)

→ **A posteriori (depois da coleta de dados)*:*** *é possível* eliminar o efeito de fontes desconhecidas, por exemplo, ajustando modelos específicos e usando os resíduos como as respostas filtradas, corrigidas (normalizadas).

Princípios do Planejamento de Experimentos e da Análise de Dados

Como extrair INFORMAÇÃO dos dados?

O Planejamento do Experimento pode garantir qualidade dos dados, controlar fontes indesejáveis (de erro) e evidenciar o efeito do fator de interesse.

$$Y = f(X) + e$$

Como realizar Análise dos dados "Válidas"?

Entender a ESTRUTURA dos dados e o OBJETIVO do estudo. Então adotar Modelos estatísticos apropriados, que satisfazem a essa estrutura e atendam aos objetivos. Os modelos devem respeitar a escala da variável resposta (Y) bem como características do planejamento (aleatorização, variáveis preditoras fixas ou aleatórias, variáveis de controle, independência ou não entre as observações, etc.). Os pressupostos do modelo adotado devem ser verificados por meio de **análises de diagnóstico**.

Princípios do Planejamento de Experimentos e da Análise de Dados

Como garantir decisões válidas sobre hipóteses científicas?

- Pré-Registro da Pesquisa (antes da coleta dos dados), com definição clara e rigorosa do Plano Amostral, Experimental, Armazenamento dos dados, Processamento, Análise Estatística e Análises de Sensibilidade (como serão tratadas observações faltantes?).

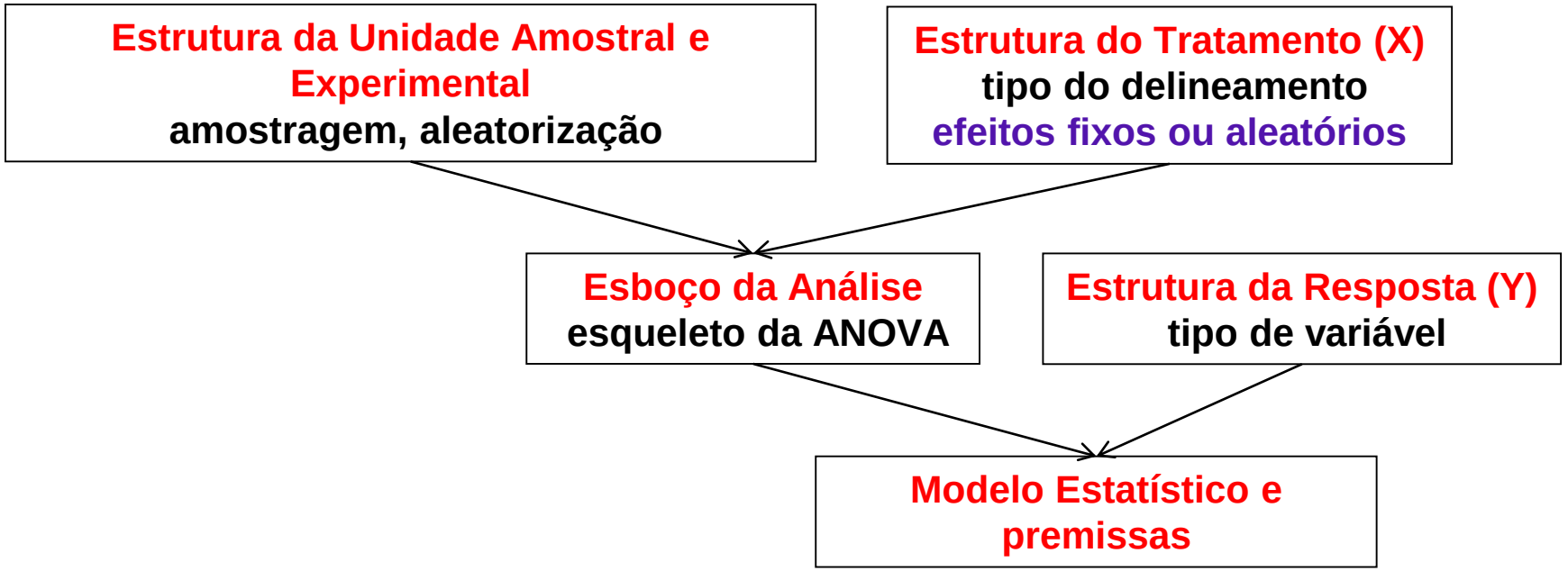
- Realizar análises estatísticas dos dados bem fundamentadas (Testes de hipóteses seguidos de Intervalos de Confiança) e asseguradas por análises de diagnóstico.

- Interpretação: Realizar a análise factual profunda e detalhada dos resultados, em coerência com o Plano de Pesquisa Inicial.

Acompanhe discussões
sobre decisões baseadas
em significância estatística
e significância biológica!

Estratégia Geral de Análise de Dados

(Goos and Gilmour, 2012; Technometrics)



Pesquise a Pesquisa

Desconfie e seja crítico com:

Garantir reprodutibilidade na pesquisa!

- Resultados que não podem ser explicados por fatos
- Resultados que destoam de resultados de outras pesquisas
- Variações muito grandes entre períodos amostrais curtos
- Procure interpretar (factualmente) um efeito significante de tratamento. Procure entender por que um efeito de tratamento não foi significante (variabilidade residual muito alta, tamanho amostral insuficiente?)
- Discuta significância estatística versus significância biológica!
- ...

Pesquise a Pesquisa

Questione, entenda, desconfie e seja crítico com:

- Objetivo do estudo
- Plano amostral usado na coleta dos dados. Qual é a população alvo?
- Como foi feita a aleatorização das unidades experimentais aos tratamentos?
- Como estão definidas as estruturas do planejamento?
- Pressupostos do modelo usado para ajustar os dados são válidos?

As observações são independentes entre e dentro de grupos?

O efeito do tratamento pode ser definido em termos de diferenças entre Médias?

A Variância da resposta é constante entre os tratamentos?

Para modelar, entenda primeiro a distribuição dos dados!

…

Atividade

1. Simulação de dados

$\Rightarrow$ Simule Dados sob a estrutura do Estudo PULSE (para facilitar, considere apenas as variáveis P1, P2 e RAN)

Pesquise os seguintes comandos do R: `rnorm` e `mvrnorm`

2. Recomende cenários de modelagem para os dados de câncer de ovário (TCGA)