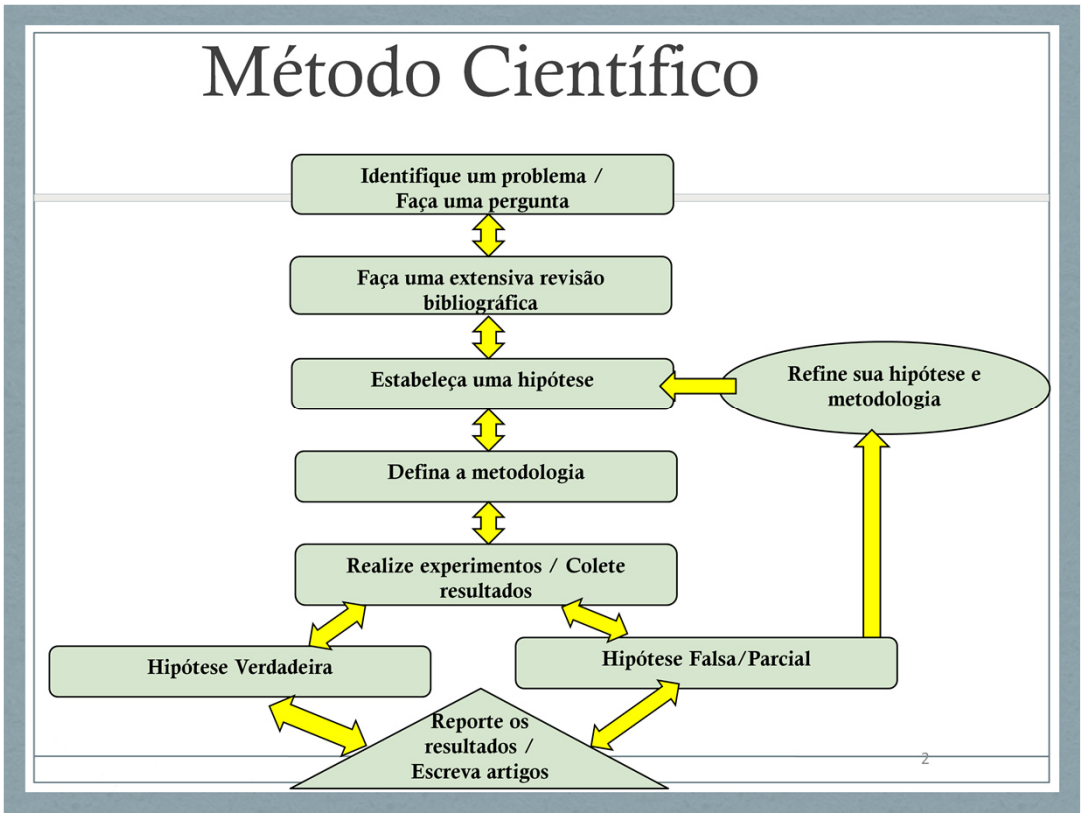
Análise Estatística

Profa. M. Cristina
cristina@icmc.usp.br



Estudos experimentais como forma de validar uma contribuição de pesquisa.

Há outras maneiras de validar...

Em visualização (e talvez também em outras disciplinas), a validação está, muitas vezes, associada com a avaliação (*evaluation*) de uma nova técnica, ou de um novo sistema.

Um estudo experimental pode ser um componente de um processo de avaliação, muitas vezes não é o único.

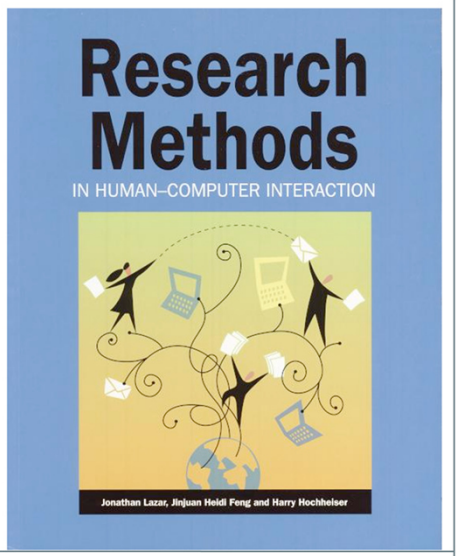
Em visualização, se a contribuição é uma técnica, ou um algoritmo, em geral a validação se dá por meio da extração de métricas (p.ex., de qualidade visual), ou de estudos experimentais controlados comparando a nova com as existentes. Esse tipo de pesquisa é direcionado à técnica...

Se a contribuição é um sistema, proposto para resolver um problema, a pesquisa é 'orientada ao problema'. Nesse caso, em geral um estudo experimental controlado não é forma mais adequada de fazer validação, ou não é suficiente – porque você só consegue 'validar' aspectos muito específicos.
Aí é preciso recorrer a técnicas de avaliação de sistemas/interfaces, muito estudadas em HCI.

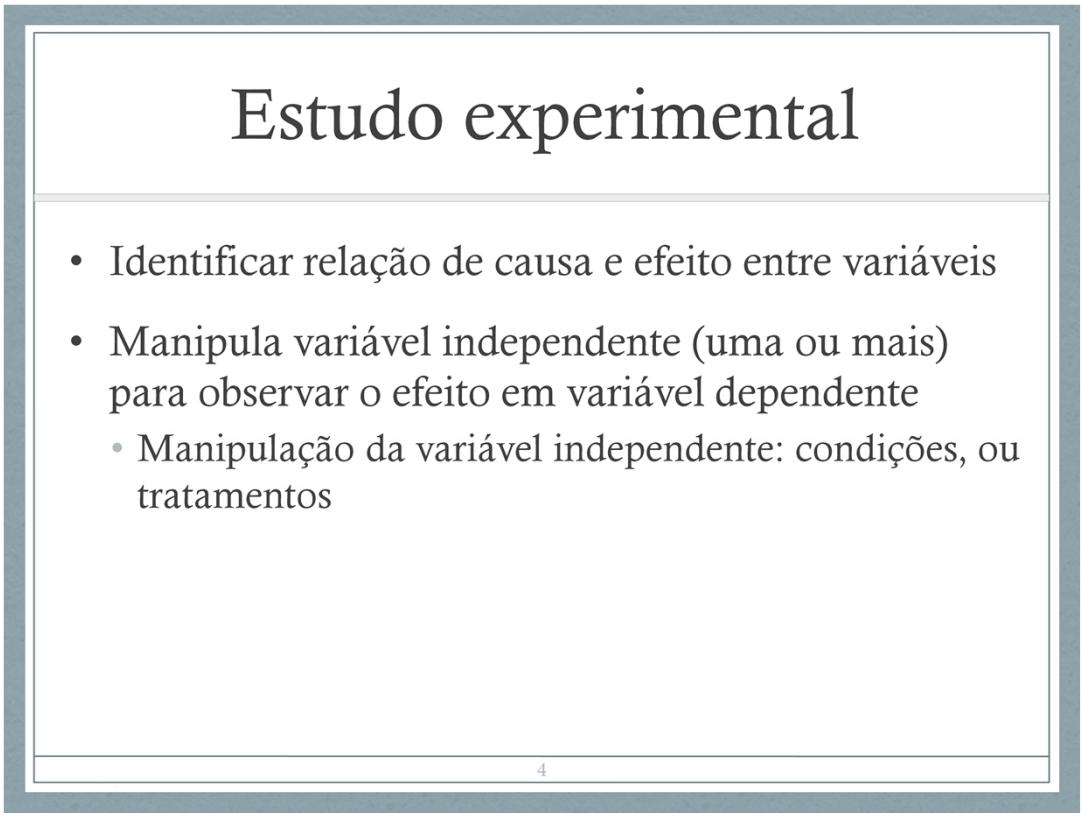
Leitura recomendada 1

- Capítulo 4:
  Statistical Analysis

  (Jonathan Lazar, Jinjuan Heidi Feng, & Harry Hochheiser - Research Methods in Human-Computer Interaction, Wiley, 2010. ISBN 0-470-72337-8, 978-0-470-72337-1)



3



Ex. 1: Suponha que você quer comparar a efetividade de dois motores de busca para tarefas de recuperação de informação, ou que você desenvolveu um novo motor de busca e quer comparar com os que já existem;

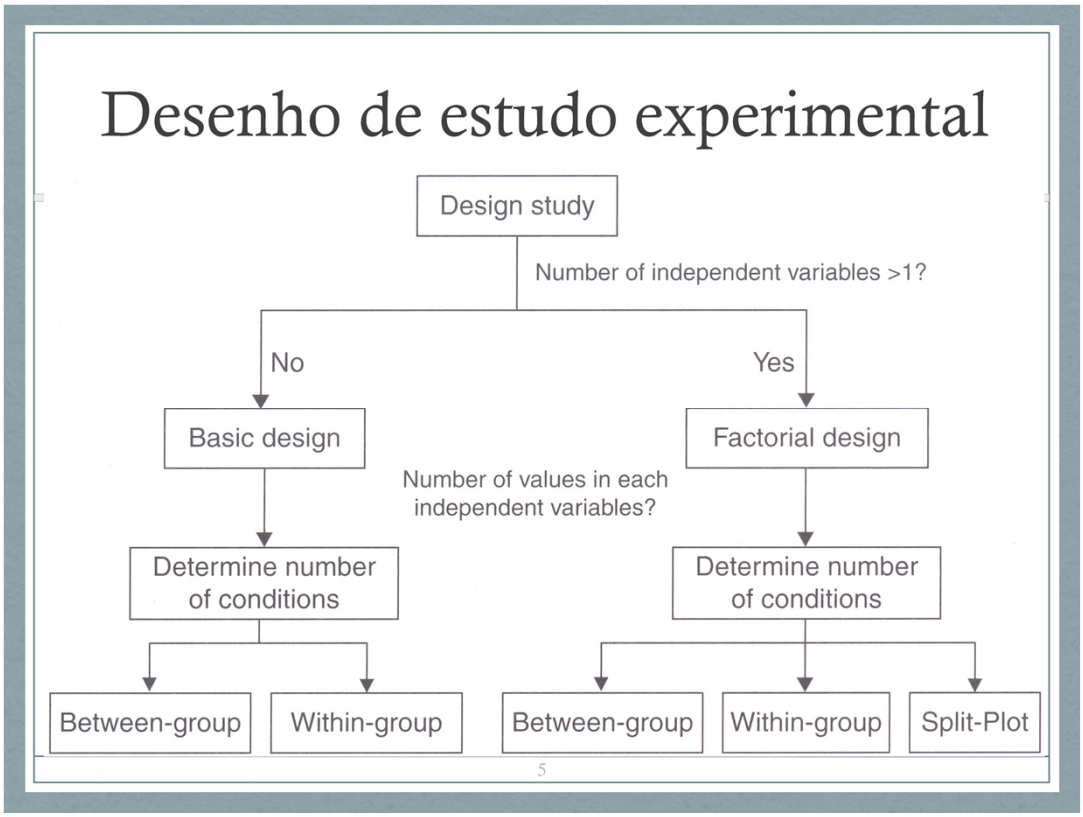
v.i.: a escolha do motor de busca

v.d.: alguma medida de efetividade em tarefas de r.i.

As condições, ou tratamentos, são os 2 motores de busca.

Ex. 2 Por exemplo, quero analisar se uma nova

técnica de visualização é mais efetiva para executar uma certa tarefa do que uma técnica já existente.
Suponha que vamos considerar que o tempo de execução da tarefa é a medida de efetividade.
v.i. a técnica escolhida. São duas condições, ou tratamentos, a serem comparados, a técnica A e a técnica B.
v.d.: a efetividade (medida pelo tempo de execução).



Desenho básico: uma única v.i. (como nos exemplos anteriores)

Desenho fatorial: múltiplas v.i.

Ex. Desenho básico:

Efeito do tipo de teclado no tempo de digitação de textos
QWERTY, DVORAK e Alfanumérico (v.i., 3 condições)
Tempo de digitação (v.d.)

Ex. Desenho fatorial

Efeito do tipo de teclado (QWERTY, DVORAK e Alfanumérico) e do tipo de tarefa executada (p.ex., tarefas de composição e de transcrição de texto)

no tempo de digitação

v.i. tipo de teclado: 3 valores
v.i. tipo de tarefa: 2 valores
Numero de condições = 3 x 2 = 6

Roteiro

- Tratamento e análise inicial dos dados: estatística descritiva
- Testes de significância estatística: qual usar e como interpretar
  - Independent-samples t test
  - Paired samples t test
  - One-way analysis of variance (ANOVA)
  - Factorial ANOVA
  - Repeated measures ANOVA
  - Correlation
  - Regression
  - Chi-square test

Análise exploratória/descritiva

- Importante verificar (olhar!) os dados antes de fazer qualquer análise estatística!
- Estatísticas descritivas, gráficos típicos…
- Verificar se os pressupostos dos testes estatísticos são satisfeitos
  - *Garbage in, garbage out…*

Tratamento inicial dos dados

- Limpeza: inspeção para identificar erros óbvios
  - entrada de dados, e.g., idade = 223
  - Data logging, e.g., *logged start time* > *logged end time*
- Codificação: mapear em valores numéricos adequados para tratamento pelos métodos estatísticos
  - p. ex. codificar dados categóricos
- Organização: adequar à estrutura ou formato exigido por métodos ou sistemas específicos

8

Várias questões aqui:

Dados errôneos (limpeza): como tratar?

Dados faltantes (*missing data*): como tratar?

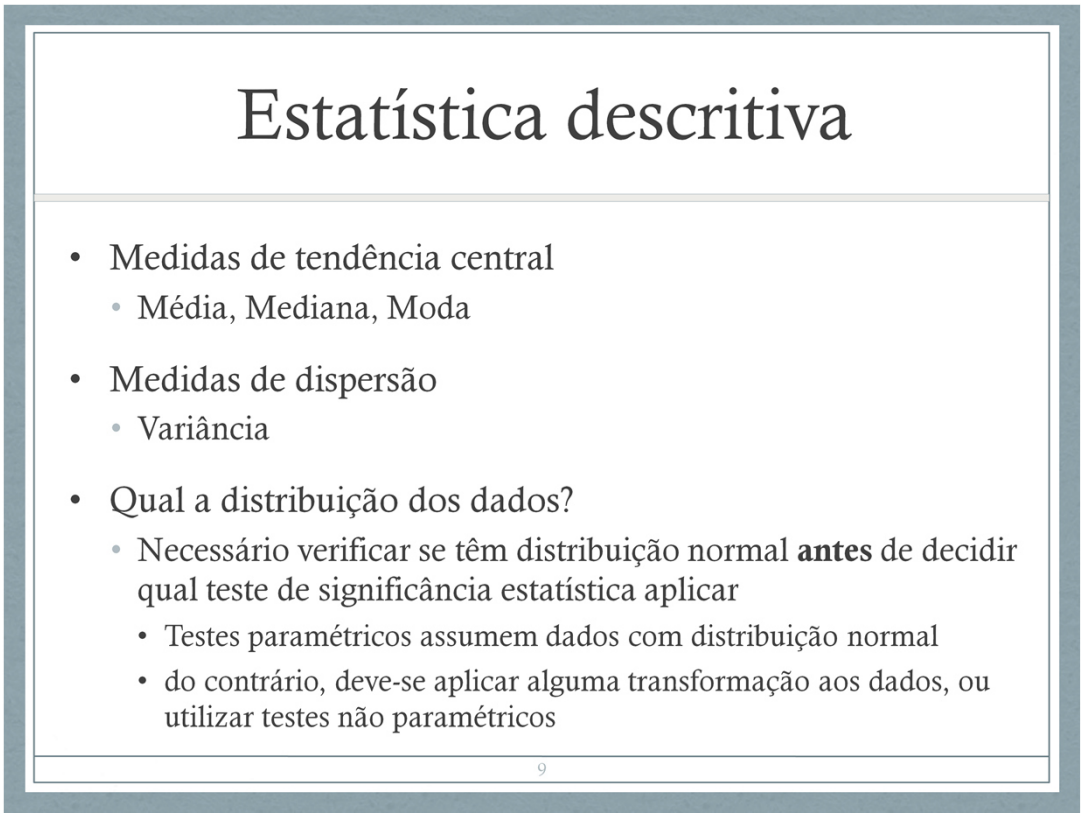
Valores extremos (*outliers*): como tratar?

Qualquer decisão demanda conhecer se o problema existe, no seu caso. Ou seja, tem que 'olhar' os dados, ou ao menos uma amostra deles.

Normalização dos dados: diversas técnicas de mineração de dados requerem algum tipo de normalização dos dados, várias estratégias de normalização... como escolher?

Mais comuns: min-max e z-score...

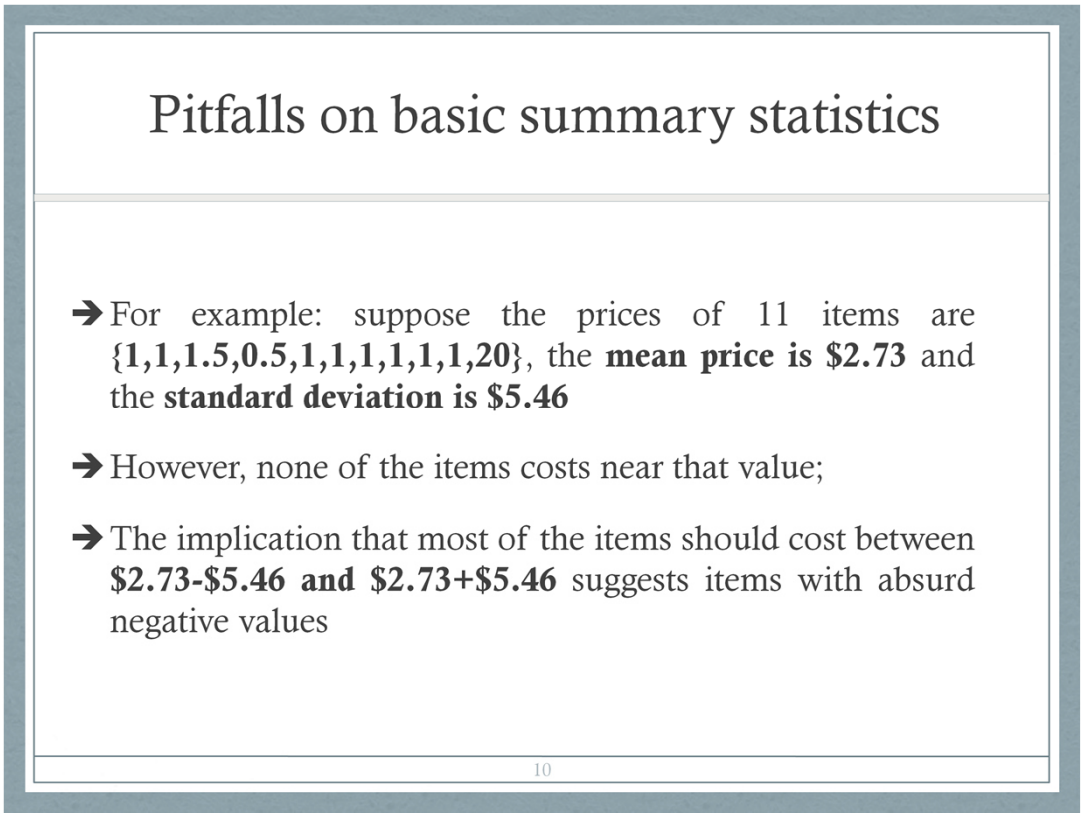
https://medium.com/@urvashilluniya/why-data-normalization-is-necessary-for-machine-learning-models-681b65a05029



medidas da tendência central – em que valor está concentrado o ´grosso´ dos dados: média, mediana, moda

Mediana: valor que divide o conjunto de dados (ordenado) em dois subconjuntos de igual tamanho

Moda: valor que ocorre com maior frequência.

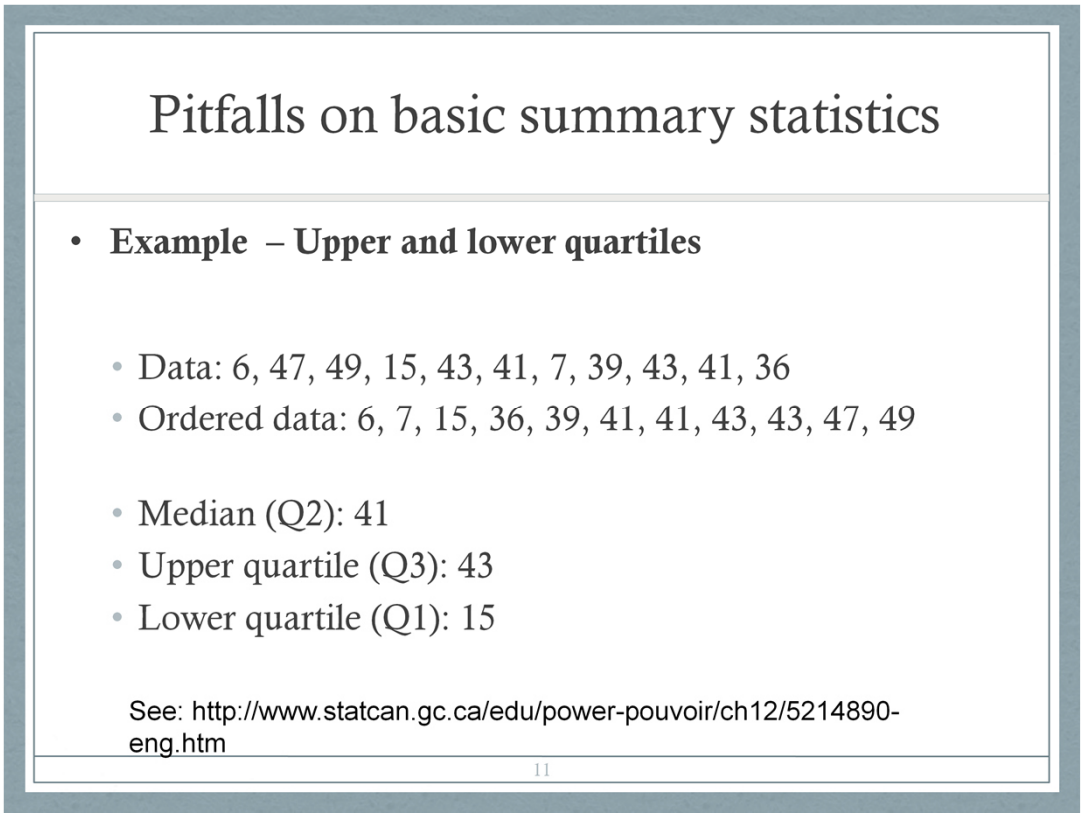


A plot of a normal distribution (or bell curve). Each band has a width of 1 standard deviation.

Cumulative probability of a normal distribution with expected value (mean) 0 and standard deviation 1.

Função distribuição de probabilidade acumulada: A função de nome "F" é igual à probabilidade de que a variável aleatória *X* assuma um valor inferior ou igual a determinado *x* i.e., F(x) = P(X <=x)

Note que, via de regra, para cada *x*, a função F assumirá um valor diferente.



A plot of a normal distribution (or bell curve). Each band has a width of 1 standard deviation.

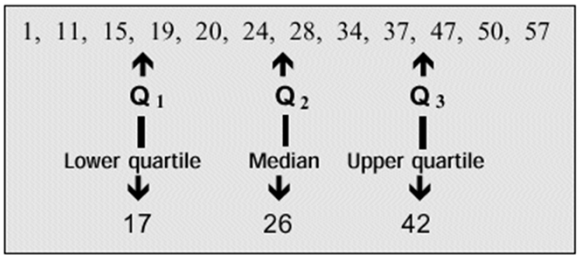
Cumulative probability of a normal distribution with expected value (mean) 0 and standard deviation 1.

Função distribuição de probabilidade acumulada: A função de nome "F" é igual à probabilidade de que a variável aleatória *X* assuma um valor inferior ou igual a determinado *x* i.e3., F(x) = P(X <=x)

Note que, via de regra, para cada *x*, a função F assumirá um valor diferente.

Pitfalls on basic summary statistics

- **Example 2 – Upper and lower quartiles**



See: http://www.statcan.gc.ca/edu/power-pouvoir/ch12/5214890-eng.htm

12

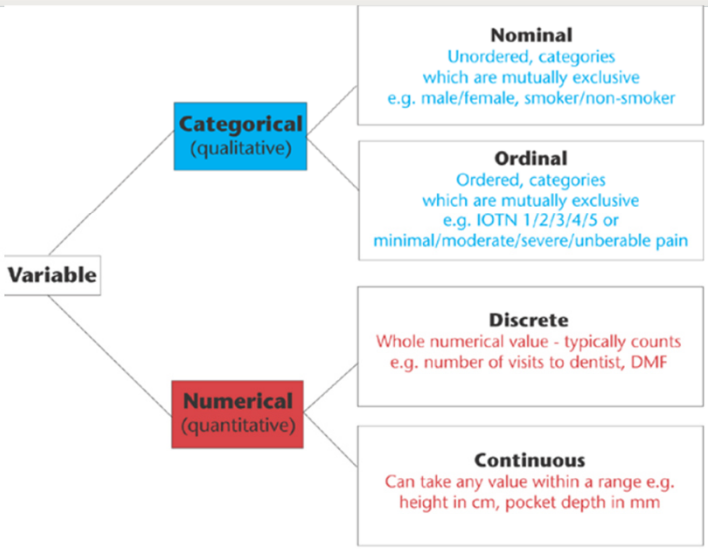
A plot of a normal distribution (or bell curve). Each band has a width of 1 standard deviation.

Cumulative probability of a normal distribution with expected value (mean) 0 and standard deviation 1.

Função distribuição de probabilidade acumulada: A função de nome "F" é igual à probabilidade de que a variável aleatória *X* assuma um valor inferior ou igual a determinado *x* i.e3., F(x) = P(X <=x)

Note que, via de regra, para cada *x*, a função F assumirá um valor diferente.

Parenthesis: variable types



https://www.graphpad.com/support/faqid/1089/

This type of classification can be important to know in order to choose the correct type of statistical analysis.

For example, the choice between regression (quantitative X) and ANOVA (qualitative X) is based on knowing this type of classification for the X variable(s) in your analysis.

Quantitative variables can be further classified into Discrete and Continuous. Discrete variables can take on either a finite number of values, or an infinite, but countable number of values. The number of patients that have a reduced tumor size in response to a treatment is an example of a discrete random variable that can take on a finite number of values. The number of car accidents at an intersection is an example of a discrete random variable that can take on a countable infinite number of values (there is no fixed upper limit to the count).

Continuous variables can take on infinitely many values, such as blood pressure or body temperature. Even though the actual measurements might be rounded to the nearest whole number, in theory, there is some exact body temperature going out many decimal places That is what makes variables such as blood pressure and body temperature continuous.

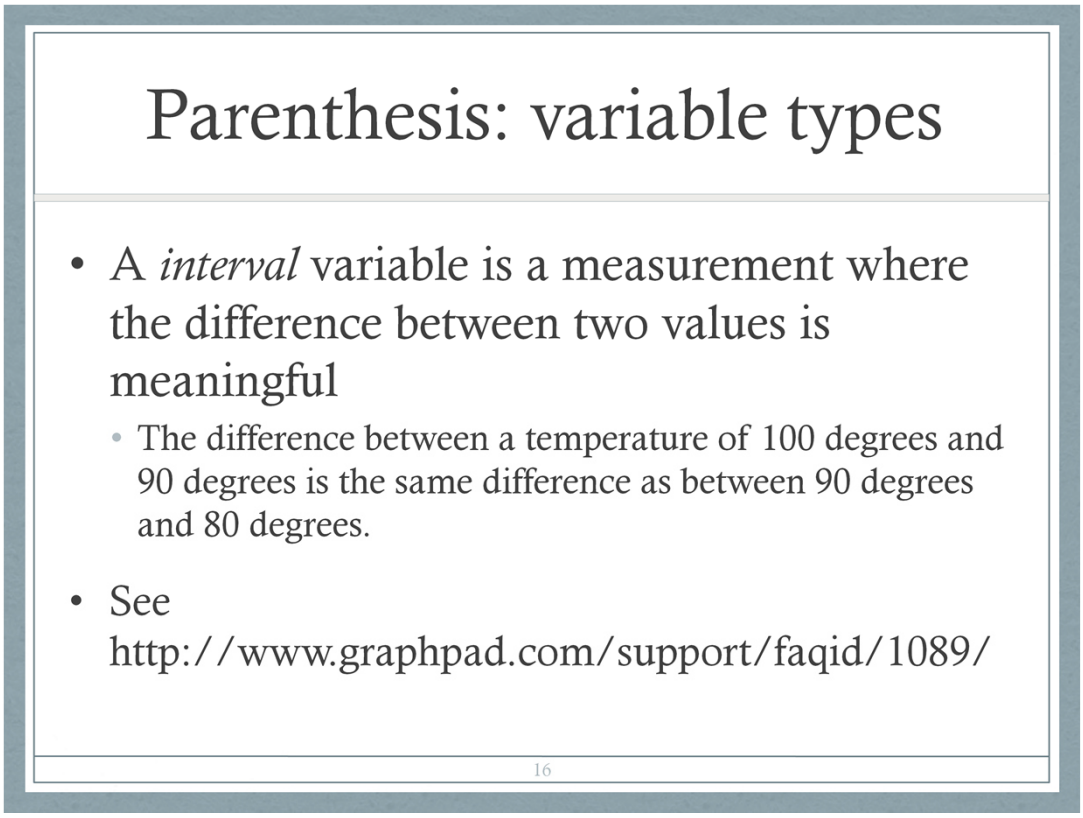
It is important to know whether you have a discrete or continuous variable when selecting a distribution to model your data. The Binomial and Poisson distributions are popular choices for discrete data while the Gaussian and Lognormal are popular choices for continuous data.

Parenthesis: variable types

- A *categorical* variable, also called a *nominal* variable, is for mutual exclusive, but not ordered, categories.
  - e.g. your study might compare five different genotypes. You can code the five genotypes with numbers if you want, but the order is arbitrary
  - any calculations (for example, computing an average) would be meaningless.
- See http://www.graphpad.com/support/faqid/1089/

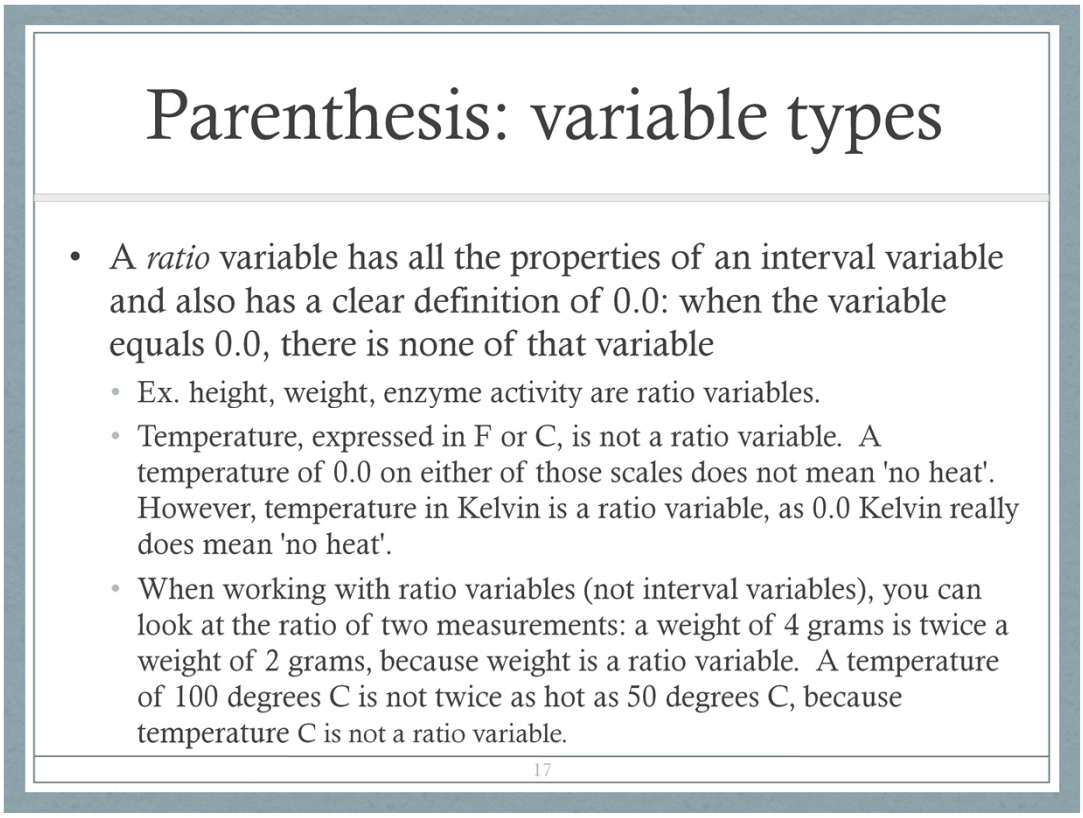
Parenthesis: variable types

- A *ordinal* variable is one where the order matters but not the difference between values
  - e.g., you might ask patients to express the amount of pain they are feeling on a scale of 1 to 10. A score of 7 means more pain that a score of 5, and that is more than a score of 3. But the difference between the 7 and the 5 may not be the same as that between 5 and 3. The values simply express an order.
  - Another example would be movie ratings, from * to *****.
- See http://www.graphpad.com/support/faqid/1089/



Note that the categories are not as clear cut as they sound. What kind of variable is color? In a psychological study of perception, different colors would be regarded as nominal. In a physics study, color is quantified by wavelength, so color would be considered a ratio variable. What about counts? If your dependent variable is the number of cells in a certain volume, what kind of variable is that. It has all the properties of a ratio variable, except it must be an integer. Is that a ratio variable or not? These questions just point out that the classification scheme appears to be more comprehensive than it is. **Read more**

about these problems.



Note that the categories are not as clear cut as they sound. What kind of variable is color? In a psychological study of perception, different colors would be regarded as nominal. In a physics study, color is quantified by wavelength, so color would be considered a ratio variable. What about counts? If your dependent variable is the number of cells in a certain volume, what kind of variable is that. It has all the properties of a ratio variable, except it must be an integer. Is that a ratio variable or not? These questions just point out that the classification scheme appears to be more comprehensive than it is. **Read more**

about these problems.

Parenthesis: variable types

OK to compute...	**Nominal**	**Ordinal**	**Interval**	**Ratio**
frequency distribution	Yes	Yes	Yes	Yes
median and percentiles	No	Yes	Yes	Yes
add or subtract	No	No	Yes	Yes
mean, standard deviation, standard error of the mean	No	No	Yes	Yes
ratio, or coefficient of variation	No	No	No	Yes

Representações gráficas

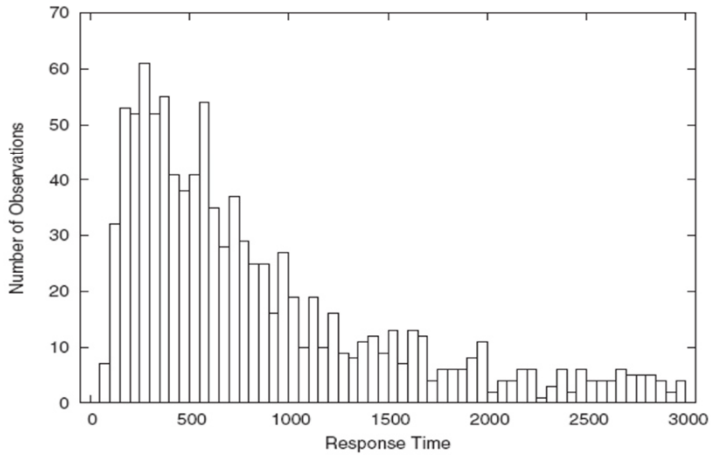
- Para informar sobre distribuição dos dados:
  - Histogramas
  - *Box-plot charts*

Histogramas

➔ Suppose a dataset with 1,000 data points, each one corresponding to the time (ms) a web server takes to answer to a given service request:

- 452.42
- 318.58
- 144.82
- 129.13
- 1216.45
- 991.56
- 1476.69
- 662.73
- 1302.85
- 1278.55
- 627.65
- 1030.78
- 215.23
- 44.50
- ...

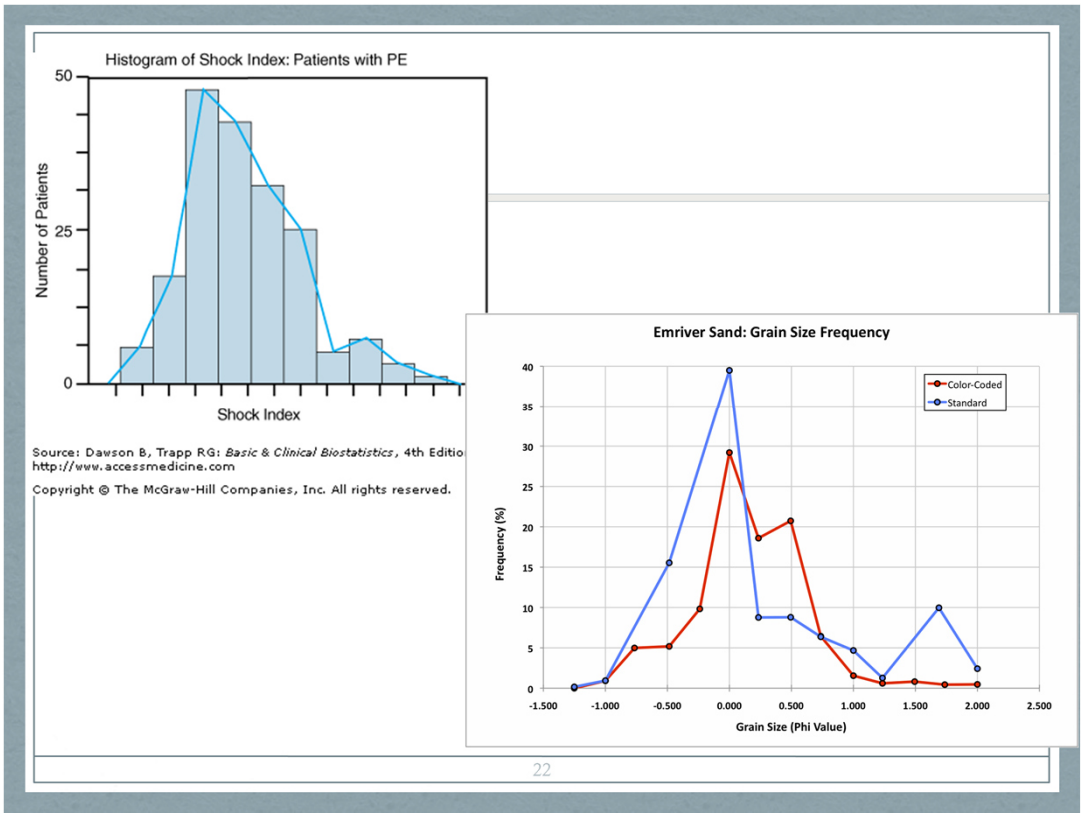
Histogramas



21

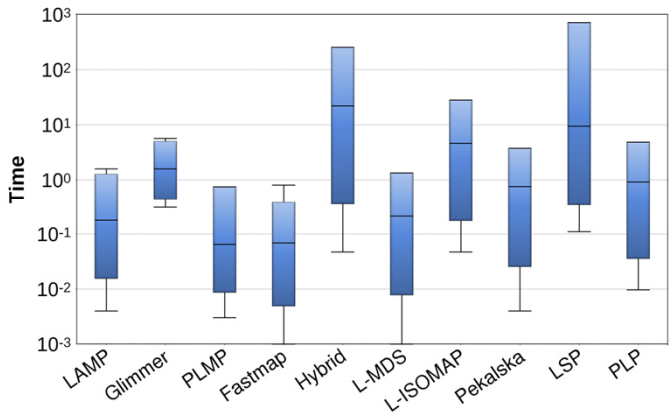
conjunto de dados com 1.000 observações, que foram agregadas em 60 intervalos (´bins´) de 50ms cada.

O que eu posso afirmar sobre o funcionamento do servidor web?



https://datavizcatalogue.com/methods/histogram.html

Box-Plots (Box-and-Whisker Plots)



Fonte: Comparação de técnicas de projeção multidimensional para identificação de grupos e busca por similaridade em dados multidimensionais. Tese de doutorado, P. Jóia Filho, 2015.

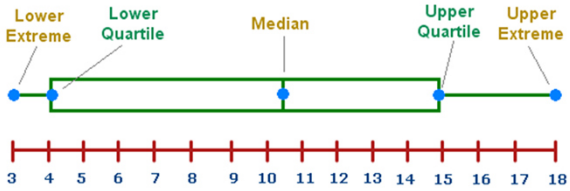
23

A Figura 4.4(b) mostra os tempos computacionais das técnicas comparadas. Note que
a LAMP é bastante competitiva, equiparada a métodos do estado da arte como a PLP.
De fato, a LAMP só tem desempenho inferior a PLMP e FASTMAP, técnicas conhecidas
por seu baixo custo computacional.

https://datavizcatalogue.com/methods/box_plot.html

Box plots (Box-and-Whisker Plots)

- A box plot can adequately express median and IQR
- It consists of
  - A marker or symbol for the median – the location of the distribution
  - A box spanning the inter-quartile range – the width of the distribution
  - A set of whiskers extending from the center to the upper and lower extremes – the tails of the distribution



Source figure: http://www.mathcaptain.com/statistics/box-and-whisker-plot.html

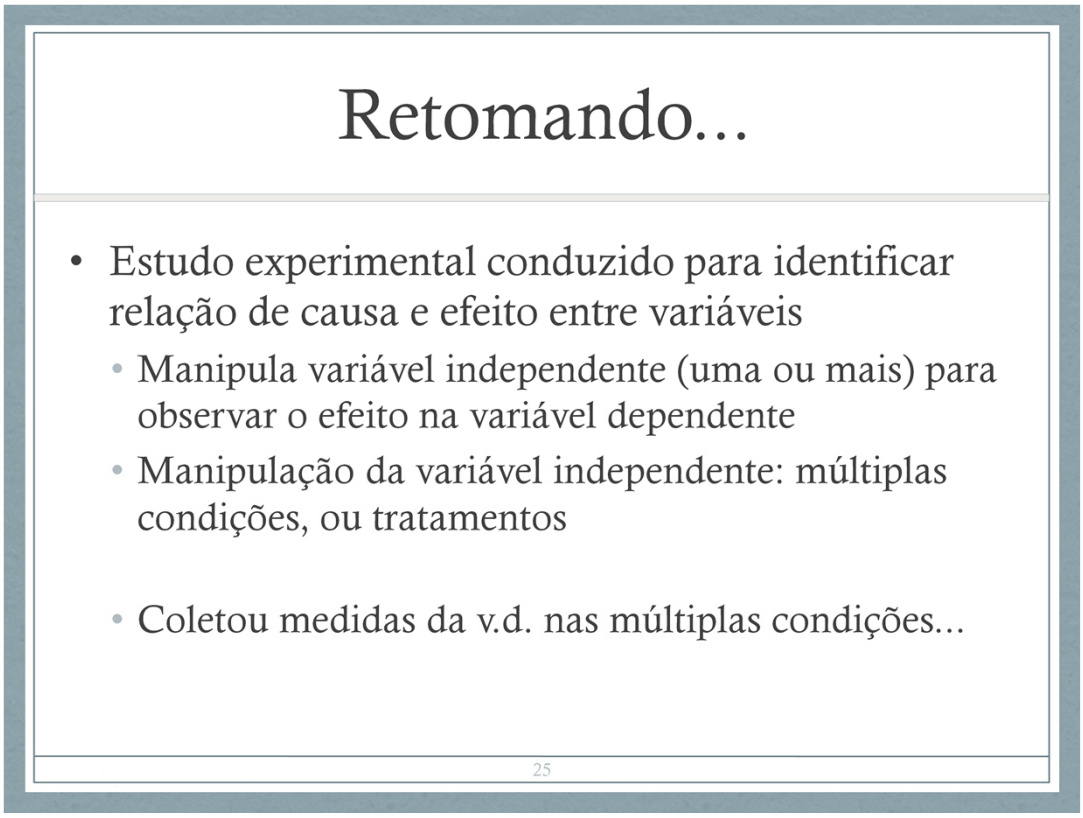
24

IQR: amplitude interquartil: 50% do conjunto: representado no ´comrimento´ da caixa

Lower adjacent value: valores < (1,5 x IQR) + Q1

Upper adjacent value: valores > (1,5 x IQR) + Q3

Valores < (>) que o Lower (Upper) adjacent values são considerados ´outliers´ (valores espúrios) (muito distantes dos valores que ocorrem com
maior frequencia (50% da frequencia…)
de valores)



Ex. 1: Suponha que você quer comparar a efetividade de dois motores de busca para tarefas de recuperação de informação, ou que você desenvolveu um novo motor de busca e quer comparar com os que já existem;

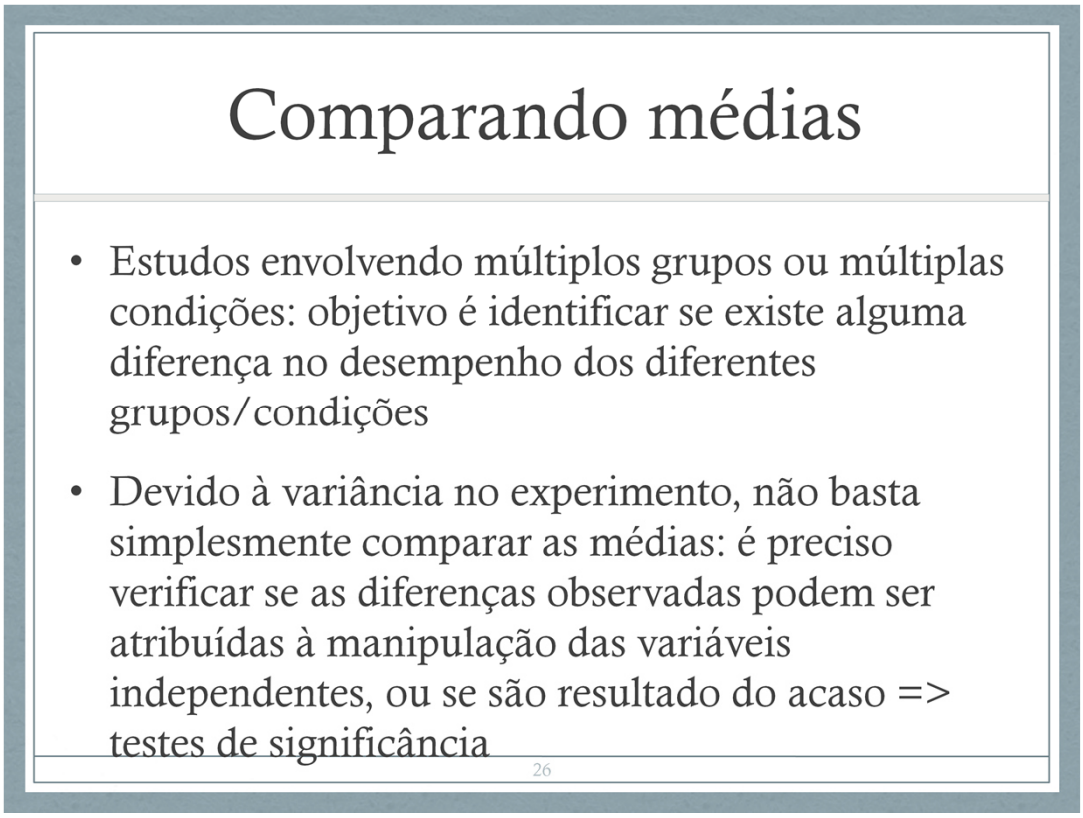
v.i.: a escolha do motor de busca

v.d.: alguma medida de efetividade em tarefas de r.i.

As condições, ou tratamentos, são os 2 motores de busca.

Ex. 2 Por exemplo, quero analisar se uma nova

técnica de visualização é mais efetiva para executar uma certa tarefa do que uma técnica já existente.
Suponha que vamos considerar que o tempo de execução da tarefa é a medida de efetividade.
v.i. a técnica escolhida. São duas condições, ou tratamentos, a serem comparados, a técnica A e a técnica B.
v.d.: a efetividade (medida pelo tempo de execução).



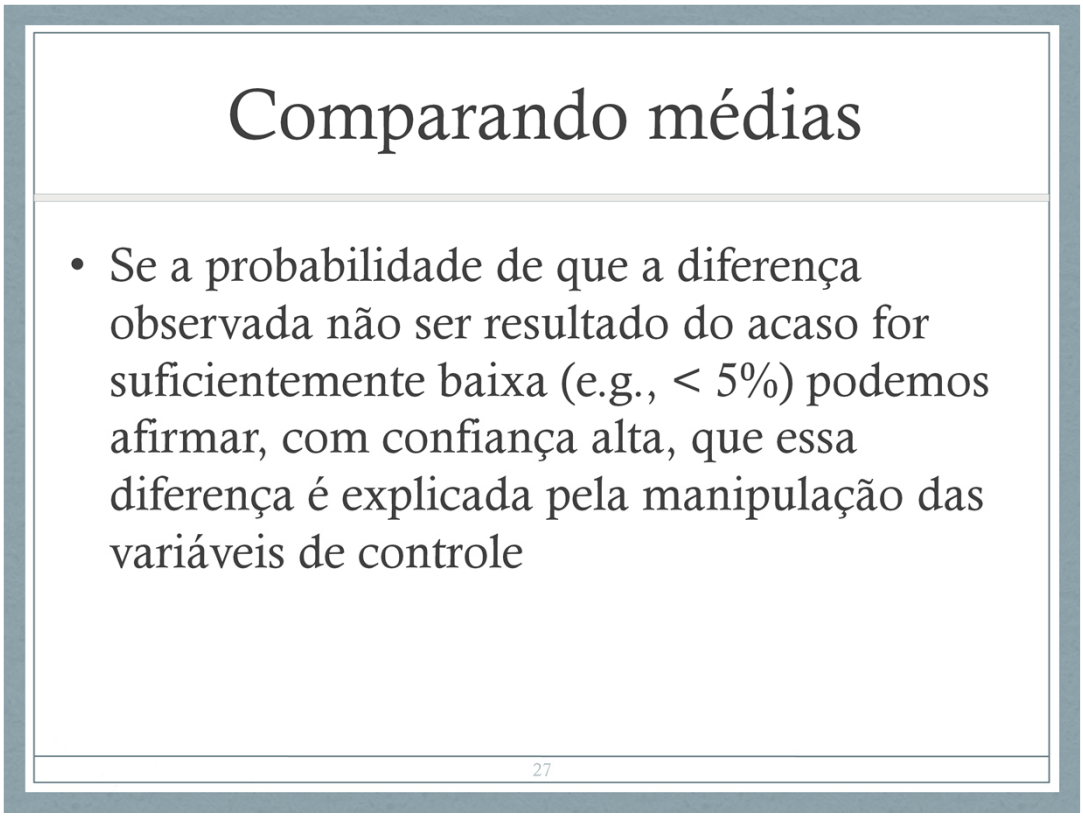
Suponha que você está avaliando a efetividade de dois motores de busca para tarefas de recuperação de informação;

Se você decidiu adotar um ´between-group design´: recrutou dois grupos de participantes, cada um dos grupos vai usar um dos motores de busca para completar uma sequencia de tarefas de busca.

Se você optou por um ´within-group´design´, você recrutou um único grupo de participantes, e todos vão executar uma sequência de tarefas de busca usando ambos os motores.

Em qualquer dos casos, você quer comparar alguma medida de desempenho dos dois grupos, ou das duas condições, para verificar se existe uma diferença que pode ser considerada estatisticamente significativa.

Muitos estudos envolvem três ou mais condições sendo comparadas.



Suponha que você está avaliando a efetividade de dois motores de busca para tarefas de recuperação de informação;

Se você decidiu adotar um ´between-group design´: recrutou dois grupos de participantes, cada um dos grupos vai usar um dos motores de busca para completar uma sequencia de tarefas de busca.

Se você optou por um ´within-group´design´, você recrutou um único grupo de participantes, e todos vão executar uma sequência de tarefas de busca usando ambos os motores.

Em qualquer dos casos, você quer comparar alguma medida de desempenho dos dois grupos, ou das duas condições, para verificar se existe uma diferença que pode ser considerada estatisticamente significativa.

Muitos estudos envolvem três ou mais condições sendo comparadas.

Testes de significância

Experiment design	Independent variables (IV)	Conditions for each IV	Types of test
Between-group	1	2	Independent-samples *t* test
	1	3 or more	One-way ANOVA
	2 or more	2 or more	Factorial ANOVA
Within-group	1	2	Paired-samples *t* test
	1	3 or more	Repeated measures ANOVA
	2 or more	2 or more	Repeated measures ANOVA
Between- and within-group	2 or more	2 or more	Split-plot ANOVA

Table 4.3 Commonly used significance tests for comparing means and their application context.

28

Existem diversos testes de significância para comparar as médias de múltiplos grupos, sendo que dois muito comuns são o t-test e a análise de variância (ANOVA).

(esses testes analisam se a variância observada nos dados pode ser atribuída ao acaso, ou se pode ser atribuída à manipulação das v.i.s (e com que confiança).

P. ex., que confiança eu posso ter que a variância observada nos tempos de digitação de textos se deve à escolha do teclado, ou que o sucesso na execução das tarefas de r.i. se deve à escolha do motor do busca.

O t-test é uma análise de variância simplificada aplicável quando são só 2 condições a serem comparados. No exemplo dos motores de busca, você usaria o independent samples t-test se tivesse adotado o between-group design, ou o paired-samples t test se tivesse optado pelo within-group design.

Se você precisa comparar mais de 2 grupos ou condições, é preciso usar um teste ANOVA.

Comparando médias

- Dado o valor retornado pelo teste, busca na tabela o valor indicado para o nível de confiança (*alpha*) e '*degree-of-freedom*' (grau de liberdade) correspondente: esse é o valor mínimo esperado para considerar a diferença observada significativa
  - se o valor retornado pelo teste é maior do que o indicado, existe diferença significativa (hipótese nula é rejeitada),
  - se o valor retornado é menor, conclui-se que a diferença não é significativa (hipótese nula é aceita)

29

http://conteudo.icmc.usp.br/pessoas/francisco/SME0123/listas/Tabela_Dist_t.pdf

Tabela informa o valor mínimo esperado pelo teste para que a diferença observada seja considerada significativa.

Lembrando

- O *p-value* (*alfa*) é a probabilidade do resultado observado ocorrer, partindo da premissa que H_0 seja verdadeira. **Quanto menor o p-value, maior é a evidência contra H_0.**
  - $p < 0.05$ = este resultado não ocorreria com frequência maior do que 5% (1 vez em 20 amostras), **caso a hipótese nula seja verdadeira**
  - $p < 0.01$ = este resultado não ocorreria com frequência maior do que 1% (1 vez em 100 amostras), **caso a hipótese nula seja verdadeira**
- Se o teste resulta em um valor de significância inferior ao de referência, é possível rejeitar a hipótese nula.

30

P-value, ou alpha: probabilidade dos dados observados ocorrerem, dado que H_0 é verdadeira. Por isso, é possível rejeitar a hipótese nula se o teste de significância estatística for inferior ao p-value.

Não é a probabilidade de incorrer no erro Tipo I (falso positivo), i.e., rejeitar H_0,

Em outras palavras, é a probabilidade de eu observar essa diferença entre os grupos (condições) nos meus dados, se a hipótese nula for verdadeira.

Vídeo https://www.youtube.com/watch?v=-MKT3yLDkqk sobre p-values

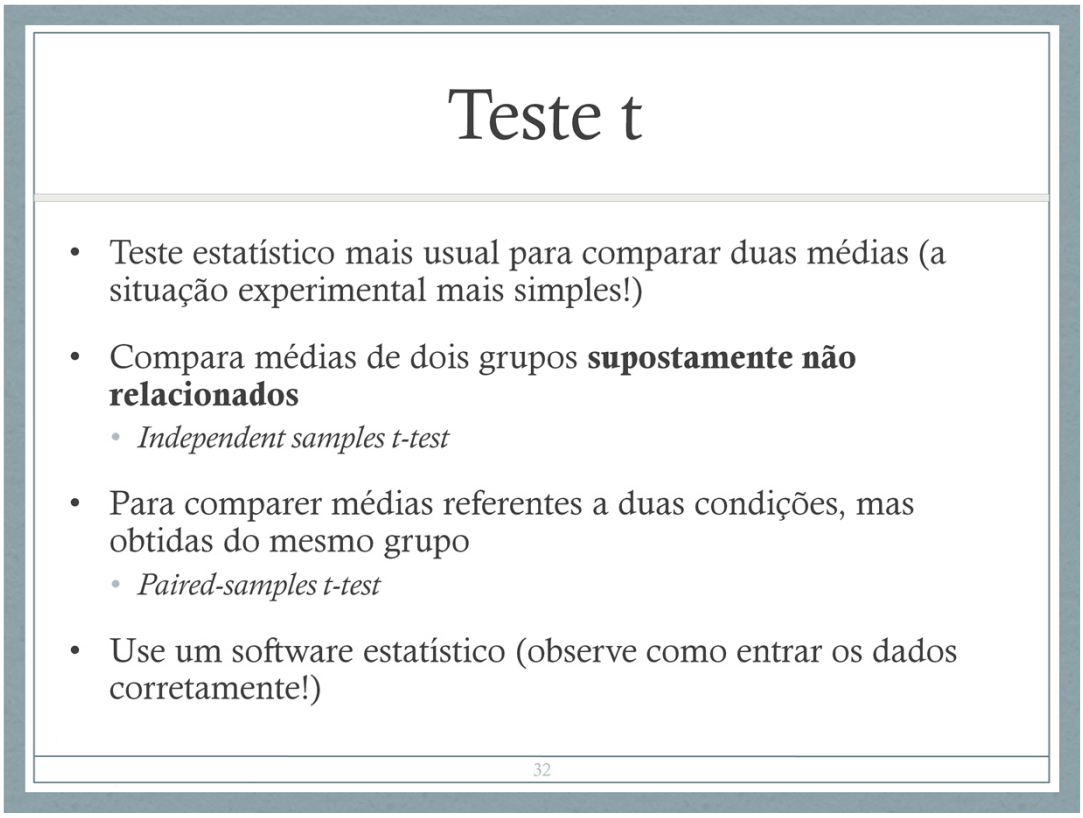
Testes de significância

Experiment design	Independent variables (IV)	Conditions for each IV	Types of test
Between-group	1	2	Independent-samples *t* test
	1	3 or more	One-way ANOVA
	2 or more	2 or more	Factorial ANOVA
Within-group	1	2	Paired-samples *t* test
	1	3 or more	Repeated measures ANOVA
	2 or more	2 or more	Repeated measures ANOVA
Between- and within-group	2 or more	2 or more	Split-plot ANOVA

Table 4.3 Commonly used significance tests for comparing means and their application context.

31

Tabela 4.3



O valor t é uma razão: a diferença entre as médias (i.e., o efeito experimental observado) dividida por uma estimativa do erro padrão da diferença entre as médias das duas amostras

Erro padrão: é uma medida de quão representativa é uma amostra, em relação à toda a população.

https://blog.minitab.com/blog/adventures-in-statistics-2/understanding-t-tests-1-sample-2-sample-and-paired-t-tests

Comparando duas médias

- independent-samples t-test: between-group design

Group	Participants	Task completion time	Coding
No prediction	Participant 1	245	0
No prediction	Participant 2	236	0
No prediction	Participant 3	321	0
No prediction	Participant 4	212	0
No prediction	Participant 5	267	0
No prediction	Participant 6	334	0
No prediction	Participant 7	287	0
No prediction	Participant 8	259	0
With prediction	Participant 1	246	1
With prediction	Participant 2	213	1
With prediction	Participant 3	265	1
With prediction	Participant 4	189	1
With prediction	Participant 5	201	1
With prediction	Participant 6	197	1
With prediction	Participant 7	289	1
With prediction	Participant 8	224	1

Table 4.4 Sample data for independent-samples *t* test.

33

p. ex. H_0: Não há diferença no tempo nos tempos para completar uma tarefa entre indivíduos que usam um software com o recurso de ´word-prediction´ e indivíduos que usam um software sem esse recurso.

Suponha que você executou o estudo experimental, com um *design between group:* 2 tratamentos, 2 grupos, cada grupo foi submetido a um tratamento.

Esse seria o formato de entrada de dados em um software estatístico como o SPSS. A coluna *Coding* indica o grupo do participante.

Entradas as colunas 3 e 4, o SPSS retorna um valor t: quanto maior esse valor, maior a probabilidade da hipótese nula ser falsa (e de você estar acertando ao rejeitá-la). Precisa verificar na tabela da t-distribution se esse valor é suficientemente alto para permitir rejeitar a hipótese nula com um determinado nível de confiança (em HCI, é comum usar 95%, ou $p <= 0.05$)

Nessa configuração, o SPSS retornaria (entre outras informações) o valor $t = 2.169$, que é maior do que o valor t para o *degree-of-freedom* específico ($df = 15$, porque temos 16 amostras) para um intervalo de confiança de 95%.

http://conteudo.icmc.usp.br/pessoas/francisco/SME0123/listas/Tabela_Dist_t.pdf

Em termos estatísticos, o resultado seria reportado assim:

An independent samples t-test suggests that there is significant difference in the task completion time between the group who used the standard word-processing software and the group who used the prediction software

($t(15) = 2{,}169$, $p < 0{,}05$)

"degrees of freedom" means that number of observations which are independent from parameter. For example, there are 25 people and 25 chairs in a room. People will sit these chairs. 24 people have chance to choose their chairs. But 25th person has not chance, he can sit only one chair. So in this case degree of freedom is 24.
To this end,
-if you use chi-square to test homogeneity or independence in contingency tables, df calculates with (row-1)*(column-1)
-if you use chi-square to test suitability of data to a specific probability distribution (goodness of fit), df calculates with (number of class -number of parameter which are estimated-1)
-if you use chi-square to test difference between expected and observed data, df calculates with (number of category-1)

When we test hypothesis with t-tests, we use some formulas to decide df. Formulas change to situations. For example: Are samples dependent/independent? Do we know variances of populations? etc… Also when we test equality of variances, we use F test and df change to variances of sample.
Firstly you should know why you are using these tests and what do you know about your data and then you can decide df.

Comparando duas médias

- paired-samples t test: within-group design

Participants	No prediction	With prediction
Participant 1	245	246
Participant 2	236	213
Participant 3	321	265
Participant 4	212	189
Participant 5	267	201
Participant 6	334	197
Participant 7	287	289
Participant 8	259	224

Table 4.5 Sample data for paired-samples *t* test.

34

p. ex. H_0: Não há diferença no tempo nos tempos para completar uma tarefa entre indivíduos que usam um software com o recurso de ´word-prediction´ e indivíduos que usam um software sem esse recurso.

Suponha que você executou um estudo experimental, com um design within group (mesmo grupo foi submetido aos 2 tratamentos).

Esse seria o formato de entrada de dados no SPSS se o design for within-group

Teste t

- Interpretação do teste t
  - o teste retorna um valor-*t* (*t-value*), **que indica a probabilidade da hipótese nula ser falsa**, i.e., quanto maior o *t-value*, maior a probabilidade da diferença observada entre as médias ser significativa (real), i.e., não ser fruto do acaso.
- Pressuposto do teste t
  - amostras com distribuição normal
  - homogeneidade da variância: variâncias das amostras nas duas condições experimentais são similares

35

quanto maior o valor t retornado, maior a probabilidade da hipótese nula ser falsa (e de você estar acertando ao rejeitá-la).

Testes de significância

Experiment design	Independent variables (IV)	Conditions for each IV	Types of test
Between-group	1	2	Independent-samples *t* test
	1	3 or more	One-way ANOVA
	2 or more	2 or more	Factorial ANOVA
Within-group	1	2	Paired-samples *t* test
	1	3 or more	Repeated measures ANOVA
	2 or more	2 or more	Repeated measures ANOVA
Between- and within-group	2 or more	2 or more	Split-plot ANOVA

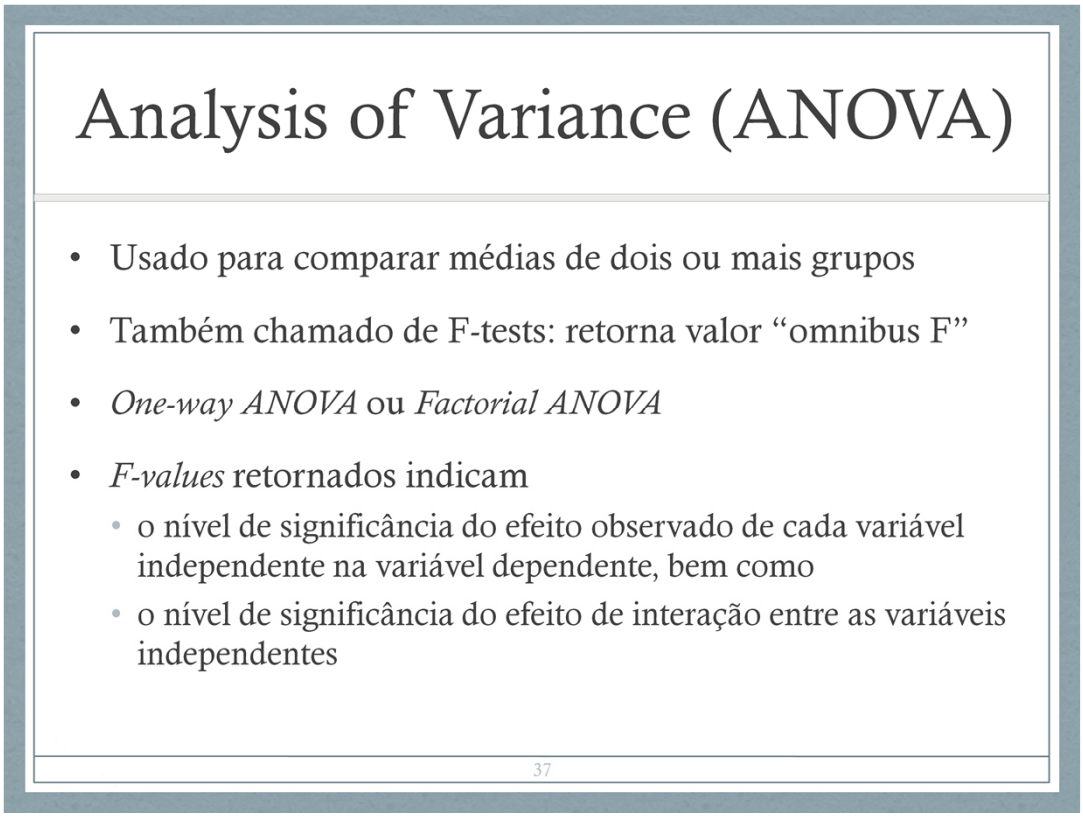
Table 4.3 Commonly used significance tests for comparing means and their application context.

36

Existem diversos testes de significância para comparar as médias de múltiplos grupos, sendo que dois muito comuns são o t-test e a análise de variância (ANOVA).

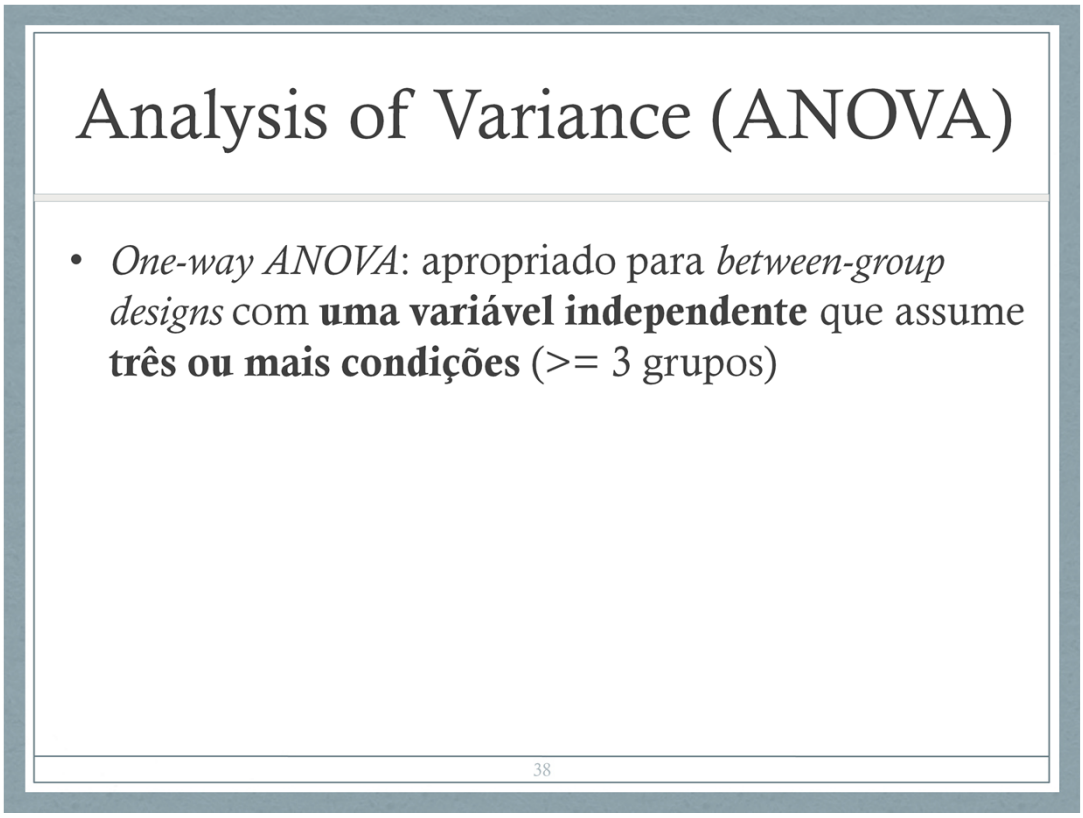
O t-test é um análise de variância simplificada aplicável quando são só 2 grupos a serem comparados. No exemplo dos motores de busca, você usaria o independent samples t-test se tivesse adotado o between-group design, ou o paired-samples t test se tivesse optado pelo within-group design.

Se você precisa comparar mais de duas condições, é preciso usar um teste ANOVA.



Retorna num valor F – em homenagem a Sir Ronald Fisher

Sabemos que a variância é uma medida de dispersão, i.e., de quão próximos ou distantes os valores observados estão, em relação à média.



v. Excelente explicação e exemplo em: https://blog.minitab.com/blog/adventures-in-statistics-2/understanding-analysis-of-variance-anova-and-the-f-test

Valor F é uma razão entre duas variâncias que pode ser usada, entre outras coisas, para testar a equivalência (igualdade) entre as médias obtidas de diferentes grupos.

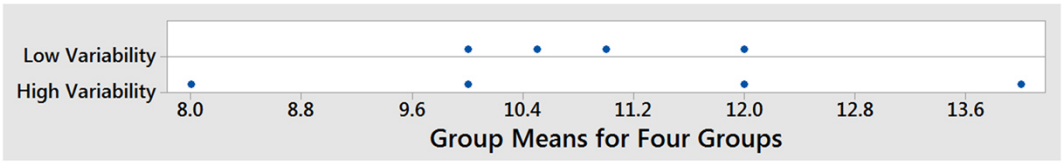
variação entre as médias amostrais
variation between sample means
F = ---------- = ----------

variação interna às amostras
variation within the samples

Ex. tenho 40 observações referentes à força (*strength*) de 4 amostras de um material plástico (10 observações por amostra). Quero determinar se os 4 grupos têm valores médios de força distintos. Suponha que as médias das observações em cada um dos 4 grupos são:

11.203, 8.938, 10.683, 8.838

F-value: numerador



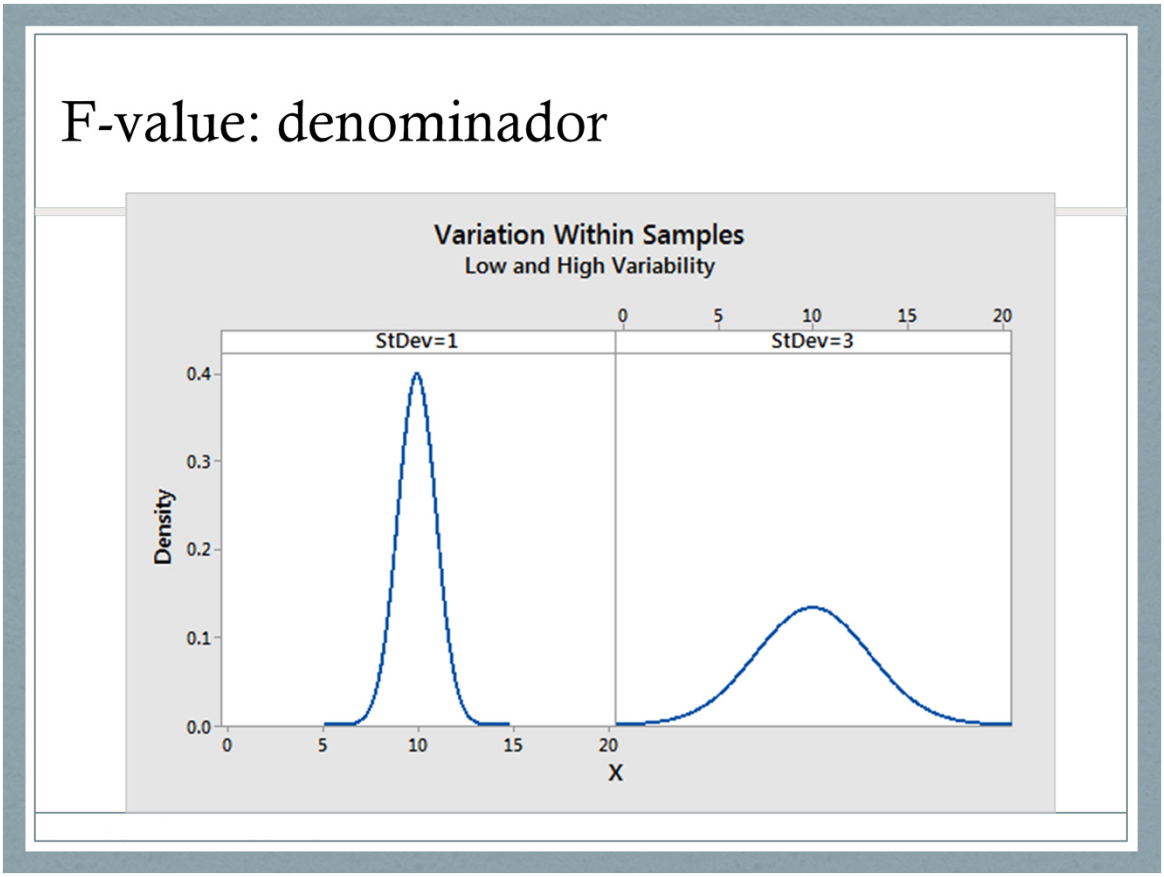
variabilidade entre as amostras variabilidade devida ao Fator Adj MS Factor = 14.540

F = ---------- = ---------- = ---------- =

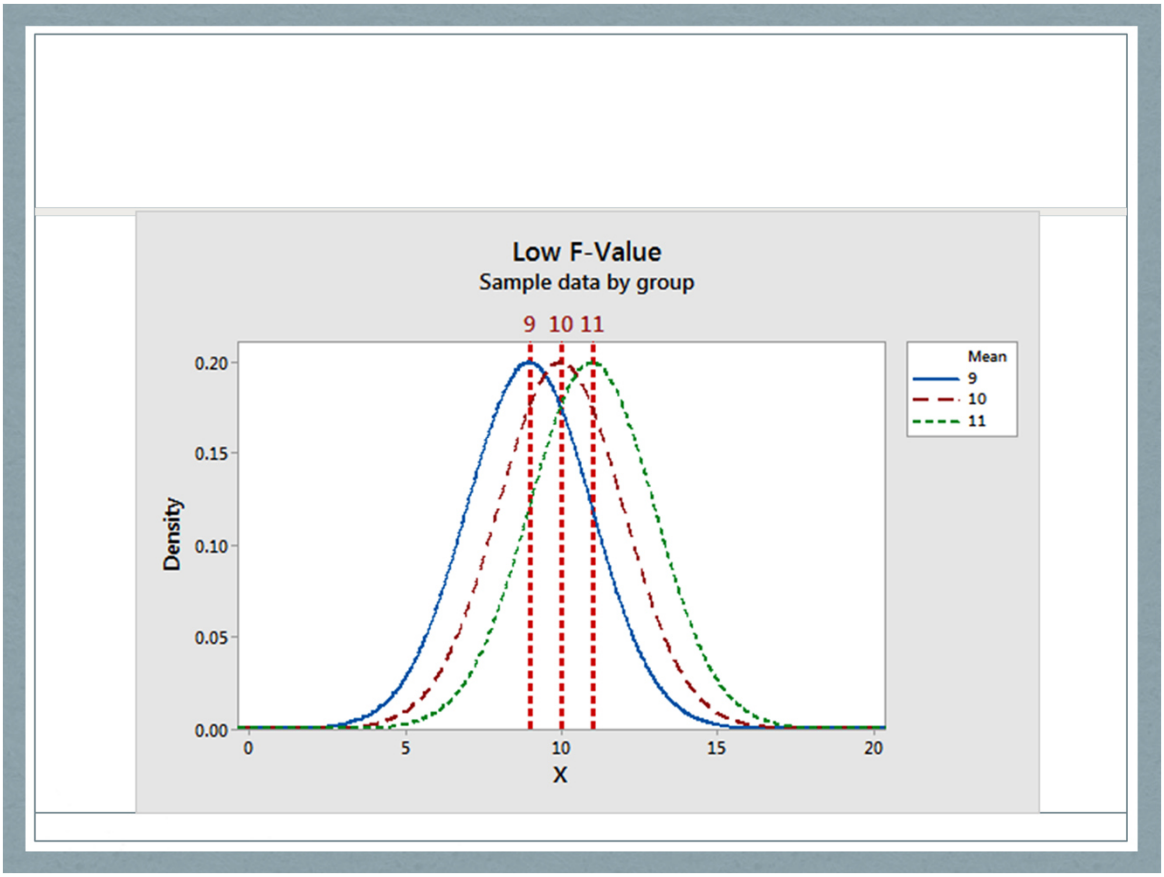
variabilidade dentro da amostra variabilidade ´natural´ Adj Ms Error = 4.402

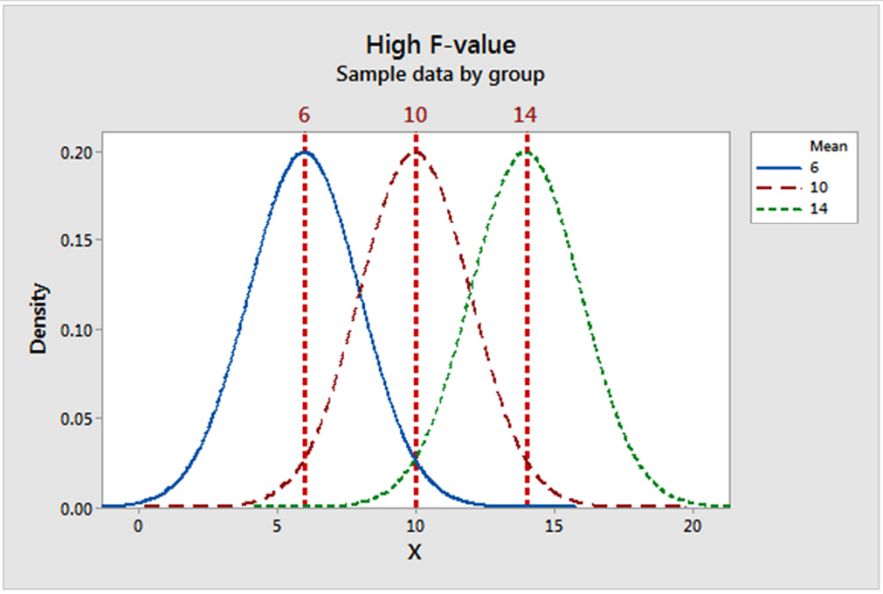
Quanto maior a variabilidade natural dentro de cada amostra, maior vai ser o denominador...

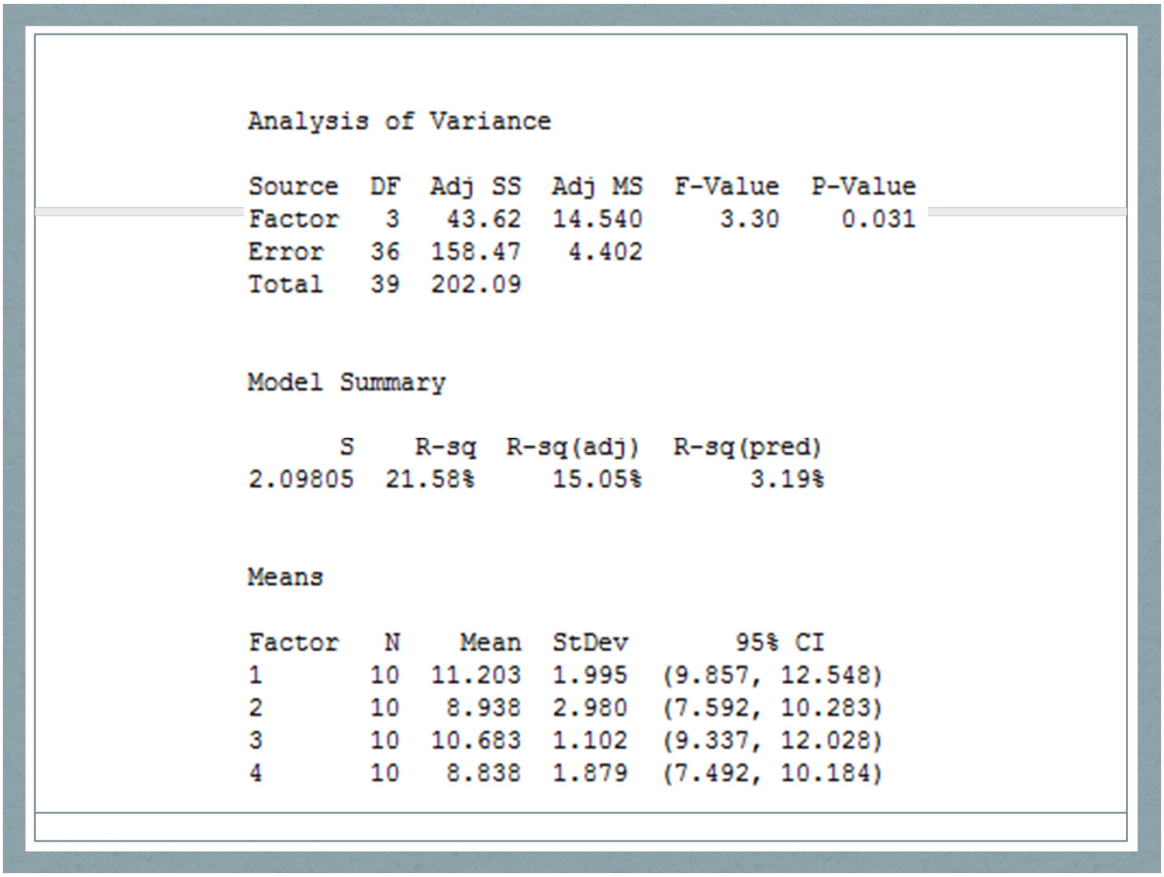
F-value: denominador



A variabilidade dentro da amostra é equivalente a um ruído que vai dificultar a percepção da variabilidade


Low F-Value
Sample data by group
9 10 11
Density
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
X
Mean
9
10
11


High F-value
Sample data by group
6
10
14
Mean
6
10
14
Density
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
X



Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Factor	3	43.62	14.540	3.30	0.031
Error	36	158.47	4.402		
Total	39	202.09			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
2.09805	21.58%	15.05%	3.19%

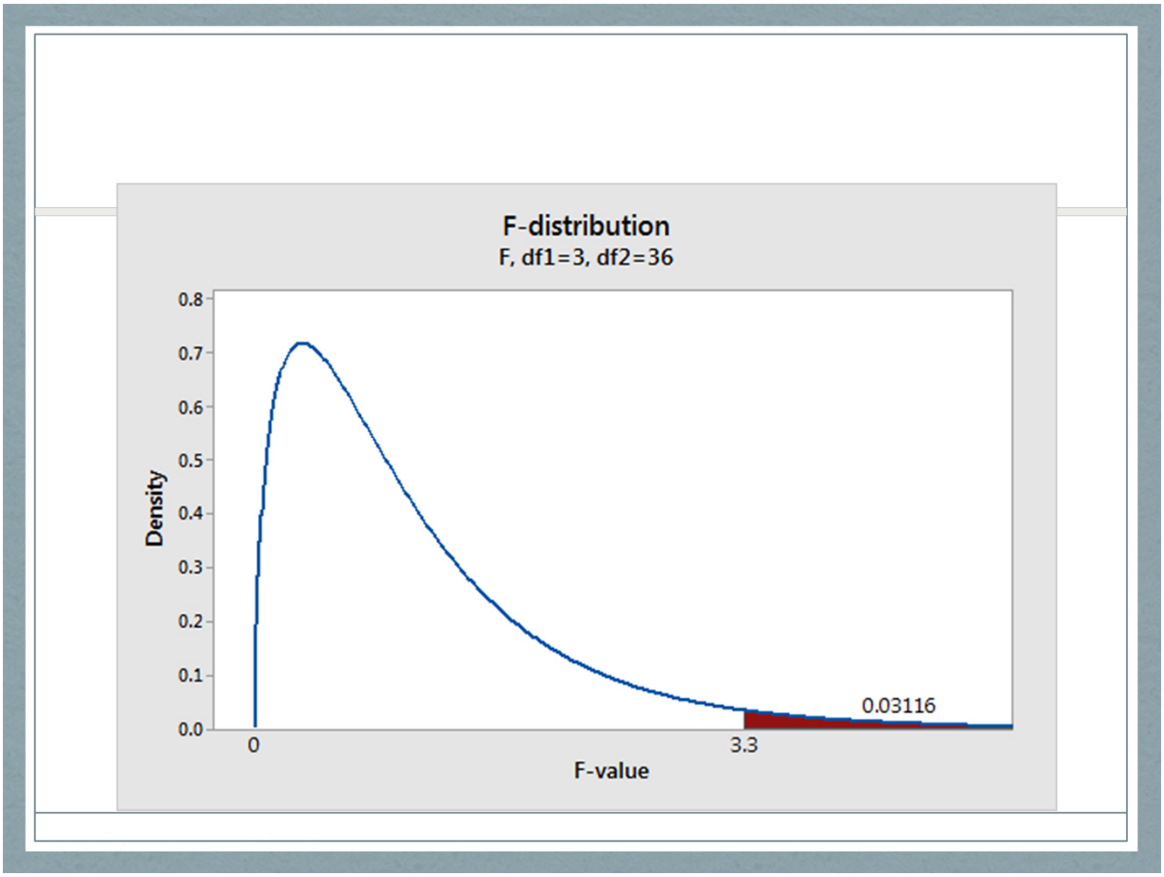
Means

Factor	N	Mean	StDev	95% CI
1	10	11.203	1.995	(9.857, 12.548)
2	10	8.938	2.980	(7.592, 10.283)
3	10	10.683	1.102	(9.337, 12.028)
4	10	8.838	1.879	(7.492, 10.184)

O que significa um valor F = 3.30? As médias são iguais ou diferentes (do ponto de vista estatístico)?

Para a one-way ANOVA (1 variável independente, 4 “condições”) a razão entre a between-group variability e a within-group variability segue uma distribuição F quando a hipótese nula ´verdadeira.

Pa



Esse gráfico mostra a distribuição dos F-values que iriamos obter se repetirmos nosso estudo muitas vezes (fazendo outras amostragens dos 4 grupos), caso a hipótese nula seja verdadeira.
A área marcada representa a probabilidade de observar um valor F igual ou maior do que o obtido nesse estudo. A probabilidade de observar um valor F igual ou maior do que este é de 3,1% quando a hipótese nula é verdadeira: essa probabilidade é suficientemente pequena para permitir que a hipótese nula (de que as médias são iguais) seja rejeitada) com nível de significância (p-value) de 0.05 (95% de confiança de estar tomando a decisão correta)

Significância = 0.03116 < 0.05 => posso rejeitar a hipótese nula

Group	Participants	Task completion time	Coding
Standard	Participant 1	245	0
Standard	Participant 2	236	0
Standard	Participant 3	321	0
Standard	Participant 4	212	0
Standard	Participant 5	267	0
Standard	Participant 6	334	0
Standard	Participant 7	287	0
Standard	Participant 8	259	0
Prediction	Participant 1	246	1
Prediction	Participant 2	213	1
Prediction	Participant 3	265	1
Prediction	Participant 4	189	1
Prediction	Participant 5	201	1
Prediction	Participant 6	197	1
Prediction	Participant 7	289	1
Prediction	Participant 8	224	1
Speech-based dictation	Participant 1	178	2
Speech-based dictation	Participant 2	289	2
Speech-based dictation	Participant 3	222	2
Speech-based dictation	Participant 4	189	2
Speech-based dictation	Participant 5	245	2
Speech-based dictation	Participant 6	311	2
Speech-based dictation	Participant 7	267	2
Speech-based dictation	Participant 8	197	2

Table 4.6 Sample data for one-way ANOVA test.

Desenho experimental:

Comparar se existe diferença nos tempos de execução de tarefas de entrada de texto, dependendo do processador de texto utilizado:
Convencional (standard), com recurso de previsão de palavras (auto-complete) e entrada por voz.

3 grupos, cada um submetido a uma das 3 condições.

Organização dos dados para rodar um teste

one-way ANOVA no SPSS: parecida com a do t-test.

Em geral, o código 0 é associado ao grupo de controle. (o SPSS requer entrada das colunas 3 e 4 apenas)

One-way ANOVA

- Resultado do teste one-way ANOVA para os dados da Tabela 4.6

Source	Sum of squares	df	Mean square	*F*	Significance
Between-group	7842.250	2	3921.125	2.174	0.139
Within-group	37880.375	21	1803.827		

Table 4.7 Result of the one-way ANOVA test.

46

Assume que os grupos testados são independentes, i.e., o desenho do experimento é *between-group* – o F-teste considera variação entre-grupos e intra-grupos para decidir se o efeito observado é significativo. Nesse caso, o valor F retornado (F = 2,174) é menor que o valor no intervalo de confiança de 95%, sugerindo que a diferença observada entre os grupos não é significativa.

Esses resultados poderiam ser reportados da seguinte forma:

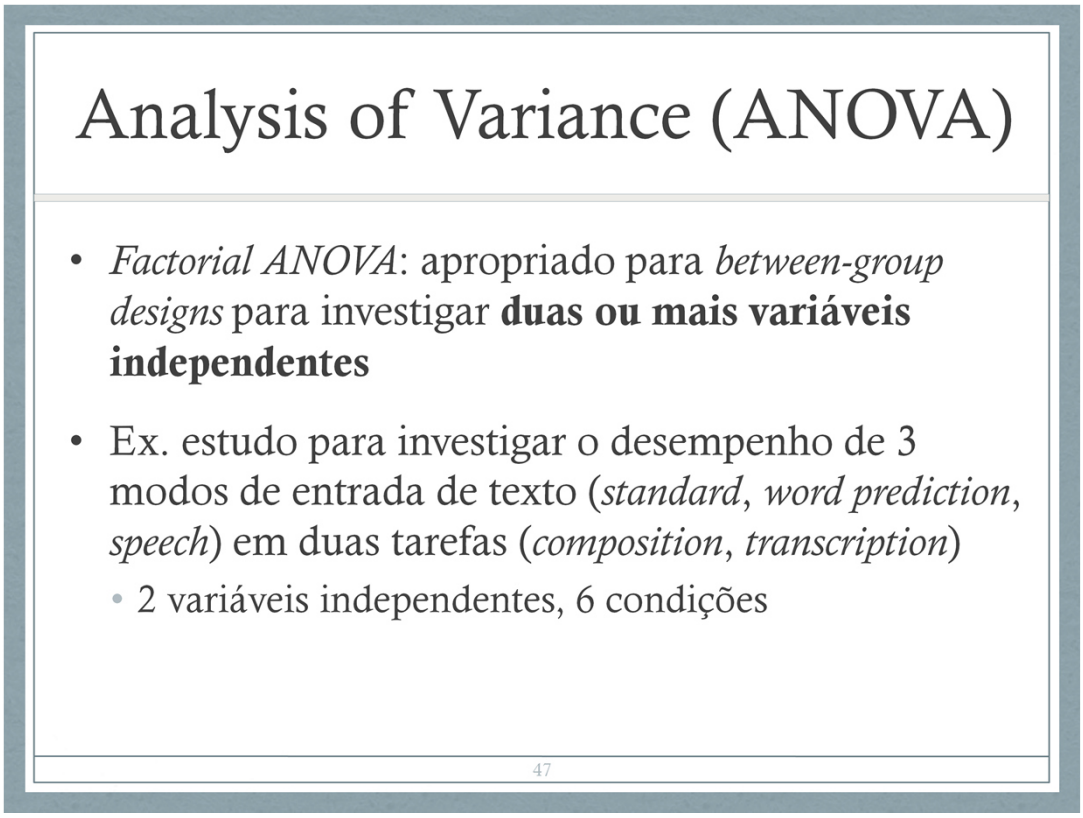
A one-way ANOVA test using task completion time as the dependent variable and group as the independent variable suggests that there is no significant diference among the three conditions (F(2, 21) = 2.174, n. s.)

A significância p = 0,139 > 0,05 => a hipótese nula **não pode ser rejeitada**.

https://blog.minitab.com/blog/adventures-in-statistics-2/understanding-analysis-of-variance-anova-and-the-f-test

F-distributions and Hypothesis Testing

https://www.youtube.com/watch?v=FPqeVhtOXEo **How to read F Distribution Table used in Analysis of Variance (ANOVA)**



Vamos considerar novamente o estudo sobre diferentes formas de entrada de texto. Você também pode estar interessado em considerar como a tarefa a ser executada afeta o desempenho.

Analysis of Variance (ANOVA)

- *between-group design*: 6 grupos de participantes

	Standard	Prediction	Speech
Transcription	Group 1	Group 2	Group 3
Composition	Group 4	Group 5	Group 6

Table 4.8 A between-group factorial design with two independent variables.

48

Task type	Entry method	Participant number	Task time	Task type coding	Entry method coding
Transcription	Standard	Participant 1	245	0	0
Transcription	Standard	Participant 2	236	0	0
Transcription	Standard	Participant 3	321	0	0
...	...	...	...	...	...
Transcription	Prediction	Participant 9	246	0	1
Transcription	Prediction	Participant 10	213	0	1
Transcription	Prediction	Participant 11	265	0	1
...	...	...	...	...	...
Transcription	Speech-based dictation	Participant 17	178	0	2
Transcription	Speech-based dictation	Participant 18	289	0	2
Transcription	Speech-based dictation	Participant 19	222	0	2
...	...	...	...	...	...
Composition	Standard	Participant 25	256	1	0
Composition	Standard	Participant 26	269	1	0
Composition	Standard	Participant 27	333	1	0
...	...	...	...	...	...
Composition	Prediction	Participant 33	265	1	1
Composition	Prediction	Participant 34	232	1	1
Composition	Prediction	Participant 35	254	1	1
...	...	...	...	...	...
Composition	Speech-based dictation	Participant 41	189	1	2
Composition	Speech-based dictation	Participant 42	321	1	2
Composition	Speech-based dictation	Participant 43	202	1	2
...	...	...	...	...	...

Table 4.9 Sample data for the factorial ANOVA test.

Analysis of Variance (ANOVA)

- Resultado do teste *factorial* ANOVA para os dados da Tabela 4.9

Source	Sum of square	Df	Mean square	*F*	Significance
Task type	2745.188	1	2745.188	1.410	0.242
Entry method	17564.625	2	8782.313	4.512	0.017
Task*entry	114.875	2	57.437	0.030	0.971
Error	81751.625	42	1946.467		

Table 4.10 Result of the factorial ANOVA test.

50

Na tabela, as linhas 1 e 2 apresentam as informações para as duas v.i., respectivamente. A linha 3 mostra os resultados para a interação entre as duas variáveis.

O resultado sugere que a diferença de desempenho (tempo) entre os participantes executando as duas tarefas distintas não é significativa: (F(1,42) = 1.41, n.s., p = 0.242 > 0.05 não permite rejeitar a hipótese nula)

E sugere a diferença de desempenho (tempo)

entre os participantes utilizando os 3 modos de entrada distintos é significativa ($F(2,42) = 4.512$, $p = 0.017 < 0.05$) permite rejeitar a hipótese nula no intervalo de confiança de 95%

Testes de significância

Experiment design	Independent variables (IV)	Conditions for each IV	Types of test
Between-group	1	2	Independent-samples *t* test
	1	3 or more	One-way ANOVA
	2 or more	2 or more	Factorial ANOVA
Within-group	1	2	Paired-samples *t* test
	1	3 or more	Repeated measures ANOVA
	2 or more	2 or more	Repeated measures ANOVA
Between- and within-group	2 or more	2 or more	Split-plot ANOVA

Table 4.3 Commonly used significance tests for comparing means and their application context.

51

Existem diversos testes de significância para comparar as médias de múltiplos grupos, sendo que dois muito comuns são o t-test e a análise de variância (ANOVA).

O t-test é um análise de variância simplificada aplicável quando são só 2 grupos a serem comparados. No exemplo dos motores de busca, você usaria o independent samples t-test se tivesse adotado o between-group design, ou o paired-samples t test se tivesse optado pelo within-group design.

Se você precisa comparar mais de duas condições, é preciso usar um teste ANOVA.

Analysis of Variance (ANOVA)

- *Repeated measures ANOVA*: apropriado para *within-group designs* envolvendo **uma ou mais variáveis independentes**, 2 ou mais tratamentos
  - *One-way repeated measures*: uma variável independente
  - *Multiple-level repeated measures*: duas ou mais variáveis

- Ex. estudo para investigar o desempenho de 3 modos de entrada de texto (*standard*, *word prediction*, *speech*)
- 1 variável independente, 3 condições

Analysis of Variance (ANOVA)

- *Within-group design*: 1 grupo, cada participante executa 3 tarefas, uma em cada condição

	Standard	Prediction	Speech
Participant 1	245	246	178
Participant 2	236	213	289
Participant 3	321	265	222
Participant 4	212	189	189
Participant 5	267	201	245
Participant 6	334	197	311
Participant 7	287	289	267
Participant 8	259	224	197

Table 4.11 Sample data for one-way repeated measures ANOVA.

Analysis of Variance (ANOVA)

- Resultado do teste *one-way repeated measures* ANOVA para os dados da Tabela 4.11

Source	Sum of square	Df	Mean square	*F*	Significance
Entry method	7842.25	2	3921.125	2.925	0.087
Error	18767.083	14	1340.506		

Table 4.12 Result of the one way repeated measures ANOVA test.

54

Valor F retornado F(2,14) = 2.925, p = 0.087 > 0.05: não permite rejeitar a hipótese nula no intervalo de 95% de confiança, sugerindo que a diferença observada entre os 3 modos de entrada de texto não é significativa.

Veja que, se houvesse diferença significativa, o teste não nos informa muito sobre como são essas diferenças... Há diferenças entre os 3 grupos? Entre dois deles? Quais?

A conclusão seria que existe um efeito, mas não sabemos aonde ele está. Para saber, seria necessário fazer outros testes (chamados post hoc tests), que vão comparar

cada condição experimental com todas as demais. Ex. Bonferroni correction.

Analysis of Variance (ANOVA)

- *Within-group design*: se o estudo investiga 2 ou mais variáveis independentes, é utilizado o *repeated measures* ANOVA

	Standard	Prediction	Speech
Transcription	Group 1	Group 1	Group 1
Composition	Group 1	Group 1	Group 1

Table 4.13 Experiment design of a two-way, repeated measures ANOVA.

55

p.ex., se você está interessado no impacto dos 2 fatores: método de entrada do texto (standard, prediction software, speech) e tipo de tarefa realizada (composition, transcription): são 6 condições distintas: teste é *two-way repeated measures* ANOVA.

Analysis of Variance (ANOVA)

	Transcription			Composition		
	Standard	Prediction	Speech	Standard	Prediction	Speech
Participant 1	245	246	178	256	265	189
Participant 2	236	213	289	269	232	321
Participant 3	321	265	222	333	254	202
Participant 4	212	189	189	246	199	198
Participant 5	267	201	245	259	194	278
Participant 6	334	197	311	357	221	341
Participant 7	287	289	267	301	302	279
Participant 8	259	224	197	278	243	229

Table 4.14 Sample data for two-way, repeated measures ANOVA test.

56

Entrada de dados no SPSS: os valores de cada participante são fornecidos na mesma linha da tabela.

Analysis of Variance (ANOVA)

Source	Sum of square	df	Mean square	*F*	Significance
Task type	2745.187	1	2745.187	14.217	0.007
Error (task type)	1351.646	7	193.092		
Entry method	17564.625	2	8782.313	2.923	0.087
Error (entry method)	42067.708	14	3004.836		
Task type * entry method	114.875	2	57.438	0.759	0.486
Error (task type * entry method)	1058.792	14	75.628		

Table 4.15 Result of the two-way, repeated measures ANOVA test.

57

O que se observa nos resultados: o tipo de tarefa tem um impacto significativo no tempo gasto para executar a tarefa (F(1,7) = 14,217, p = 0.007 < 0,01, permite rejeitar a hipótese nula, com confiança de 99%)

Não há diferença significativa entre os 3 modos de entrada de texto (F(2,14) = 2,923, n.s.). p = 0.087 > 0,01 não permite rejeitar a hipótese nula com confiança de 99%)

O efeito da interação entre as 2 variáveis independentes também não é significativo (F(2,14) = 0.759, n.s.) p = 0.486 > 0,01 não

permite rejeitar a hipótese nula com confiança de 99%)

ANOVA com “split-plot” design

- Quando o mesmo experimento analisa um fator com *between-group design* e outro fator com *within-group design*
- Como no caso anterior, F-values são computados para cada fator, e para a interação entre os dois fatores
- Ex. Estudo anterior com 2 grupos
  - Grupo 1 executa tarefas de transcrição com os 3 métodos
  - Grupo 2 executa tarefas de composição com os 3 métodos

58

No estudo anterior

ANOVA com “split-plot” design

- Estudo: 2 tarefas x 3 modos de entrada de texto
  - Tarefa: *between group*
  - Modo de entrada de texto: *within group*

	Keyboard	Prediction	Speech
Transcription	Group 1	Group 1	Group 1
Composition	Group 2	Group 2	Group 2

Table 4.16 Split-plot experiment design.

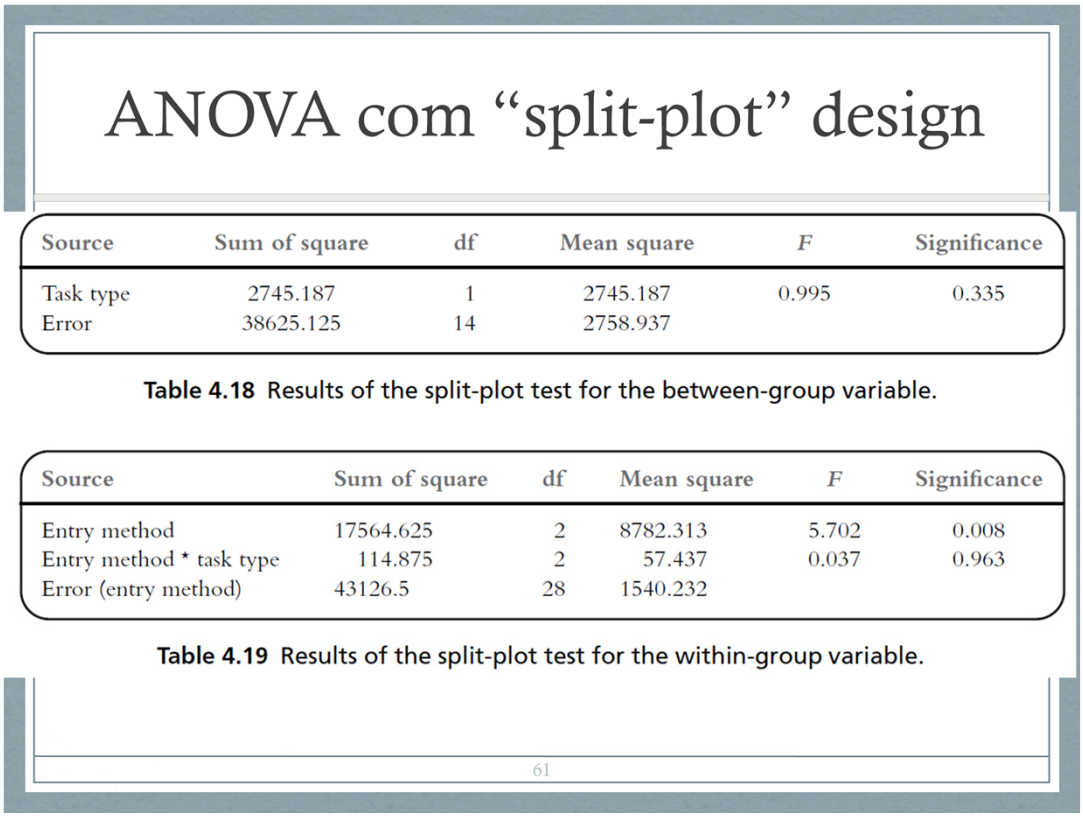
59

Vantagens desse design: reduz o tempo de execução das tarefas, maior controle sobre o efeito de aprendizado

ANOVA com "split-plot" design

Task type	Participant number	Task type coding	Standard	Prediction	Speech
Transcription	Participant 1	0	245	246	178
Transcription	Participant 2	0	236	213	289
Transcription	Participant 3	0	321	265	222
Transcription	Participant 4	0	212	189	189
Transcription	Participant 5	0	267	201	245
Transcription	Participant 6	0	334	197	311
Transcription	Participant 7	0	287	289	267
Transcription	Participant 8	0	259	224	197
Composition	Participant 9	1	256	265	189
Composition	Participant 10	1	269	232	321
Composition	Participant 11	1	333	254	202
Composition	Participant 12	1	246	199	198
Composition	Participant 13	1	259	194	278
Composition	Participant 14	1	357	221	341
Composition	Participant 15	1	301	302	279
Composition	Participant 16	1	278	243	229

Table 4.17 Sample data for the split-plot ANOVA test.



Resultados no SPSS:

Tabela 4.18, resultados para o fator **Tipo de tarefa**. Conclusão: não existe diferença significativa entre os participantes que completam uma ou outra tarefa (F(1,14) = 0,995, n.s.) p = 0.335 > 0,01 não permite rejeitar a hipótese nula)

Tabela 4.19, resultados para o fator **Modo de entrada de texto**. Conclusão: existe uma diferença significativa nos resultados obtidos com os 3 modos distintos de entrada de texto (F(2,28) = 5.702, p = 0.008 < 0, 01 permite

rejeitar a hipótese nula)

O efeito da interação entre os 2 fatores não é significativo: F(2,28) = 0,037, n.s.) p = 0.963 > 0,01

Pressupostos de *t-tests* e *F-tests*

- Os erros associados aos dados coletados/medidos são independentes entre si
- Os erros são identicamente distribuídos
  - homogeneidade da variância: ao comparar as médias de múltiplos grupos, os resultados do *t-test* ou *F-test* serão mais precisos se as variâncias nas populações forem simlares
  - Como saber? Levene´s test para verificar se as variâncias de dois grupos são equivalentes
- Os erros satisfazem uma distribuição normal

62

Populações normalmente distribuidas•
Populações tem mesma variância (ou mesmo desvio padrão).•
Amostras são aleatórias e mutuamente independentes.•
As diferentes amostras são obtidas de populações classificadas em apenas uma categoria.

Testes de significância

- Ex.

"*On average, participants performed significantly better* (F(1, 25)=20,83, p < 0.01)... *in condition A than in condition B...*"

"*A t test showed that there was a significant difference in the number of lines of text entered (t(11) = 6.28, p < 0.001) with more entered in the tactile condition. ...* "

Correlação

- Medidas de correlação entre fatores
  - Dois fatores (variáveis) são correlacionados se existe uma relação entre eles
  - Pode ser verificado, p.ex., pelo teste de Correlação de Pearson (coeficiente *r*, varia no intervalo [-1,+1])
  - *r square* (r^2) representa a proporção da variância compartilhada pelas duas variáveis: quanto da variância da variável Y pode ser explicada pela variável X

64

Muitos estudos têm por objetivo identificar se 2 ou mais fatores estão relacionados...

Correlação

- Correlação não implica em causalidade

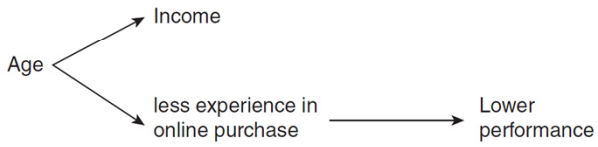


Figure 4.2 Relationship between correlated variables and an intervening variable.

65

A correlação observada entre 2 fatores não implica necessariamente em relação de causa->efeito

Exemplo: experimento para estudar como usuários interagem com um website de e-comércio. Observou-se que participantes com maior renda gastam mais tempo buscando um item especifico, e erram mais durante a navegação.
Pode-se afirmar que ganhar mais é a causa?

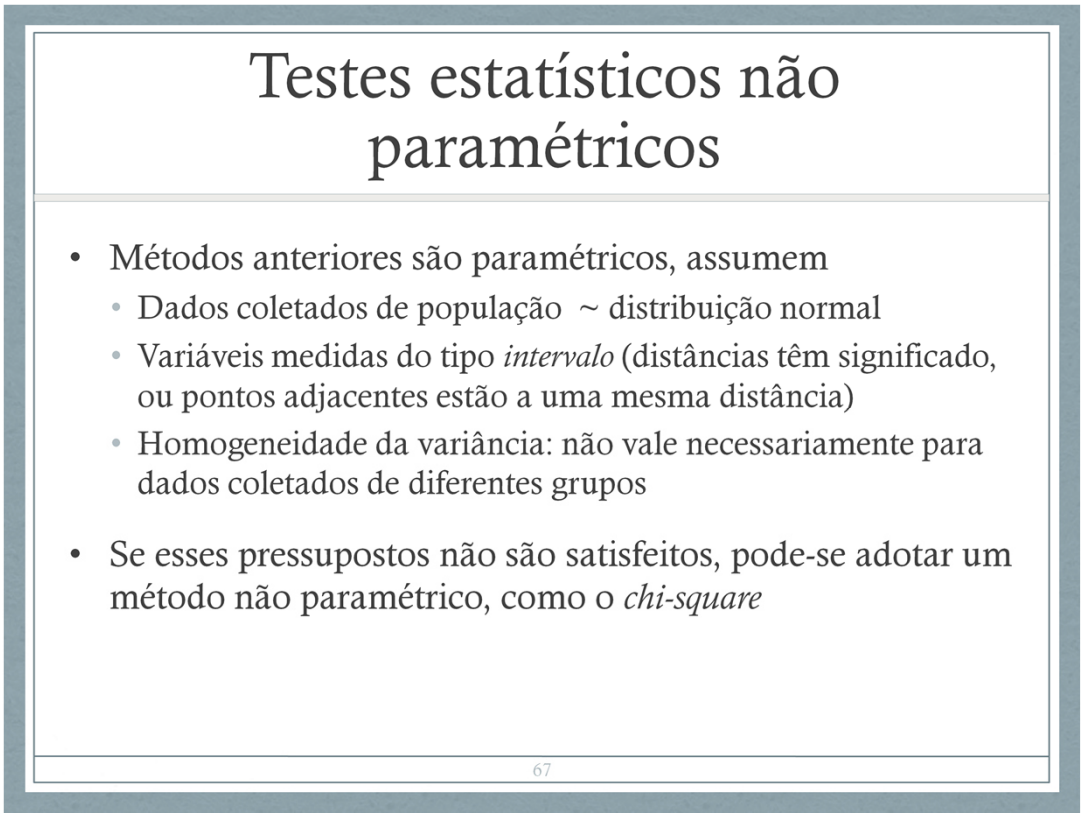
Obviamente não! O fato é que pessoas com maior renda tendem a ser mais velhas, e muitas

delas usam menos o computador para acessar sites do que os mais jovens... No caso, a idade é uma variável oculta que possivelmente está influenciando esse comportamento!

Intervening variable – confouding variable

Regressão

- Análise de regressão: objetivo é criar um modelo que explique a relação de uma variável dependente com uma ou mais variáveis independentes
  - **Regressão simultânea**: análise da variável dependente em relação a grupo de variáveis independentes. Determina a porcentagem da variância da variável dependente que pode ser explicada pelo grupo de variáveis independents
  - **Regressão hierárquica**: análise da variável dependente em relação a cada variável independente



Métodos não paramétricos fazem menos suposições a respeito do comportamento dos dados do que os testes paramétricos

Chi-square test

- Aplicável quando os dados descrevem contagens (tabelas de contingência) (categóricos)
- O teste retorna um *chi-square value* e um *p-value* que permitem determinar se as diferenças entre os grupos são significativas

Tabela de contingência 2 x 2

		Preferred device	
		Mouse	Touch Screen
Age	< 65	14	6
	>=65	4	16

$X^2(1) = 10.1$, $p < 0.005$: probabilidade dos valores observados serem resultado do acaso é inferior a 0.005

69

Tabela indica que mais participantes mais jovens preferem o mouse, enquanto os mais velhos preferem tela sensível ao toque. Para verificar se esses resultados não são apenas resultado do acaso assumindo que a hipótese nula é verdadeira (i.e., H_0 = não há relação entre a idade e a preferência por um dispositivo), você pode rodar um teste Chi-square. Os resultados SPSS para esses dados:

$X^2(1) = 10.1$, $p < 0.005$ sugerindo que a probabilidade de que os valores observados sejam resultado do acaso são inferiores a

0.005. Usando o intervalo de confiança de 95% você pode rejeitar a hipótese nula e concluir que existe uma relação entre a idade e a preferência por um dispositivo.

O grau de liberdade do Chi-square é dado por (n-1) x (m – 1), n = número de linhas, m = número de colunas na tabela

Chi-square test

- Suposições sobre os dados
  - valores inseridos na tabela de contingência são independentes entre si, i.e.,
    - cada participante contribui com uma única entrada na tabela
  - a amostra não pode ser muito pequena
    - (> 20?)

Outros testes não paramétricos

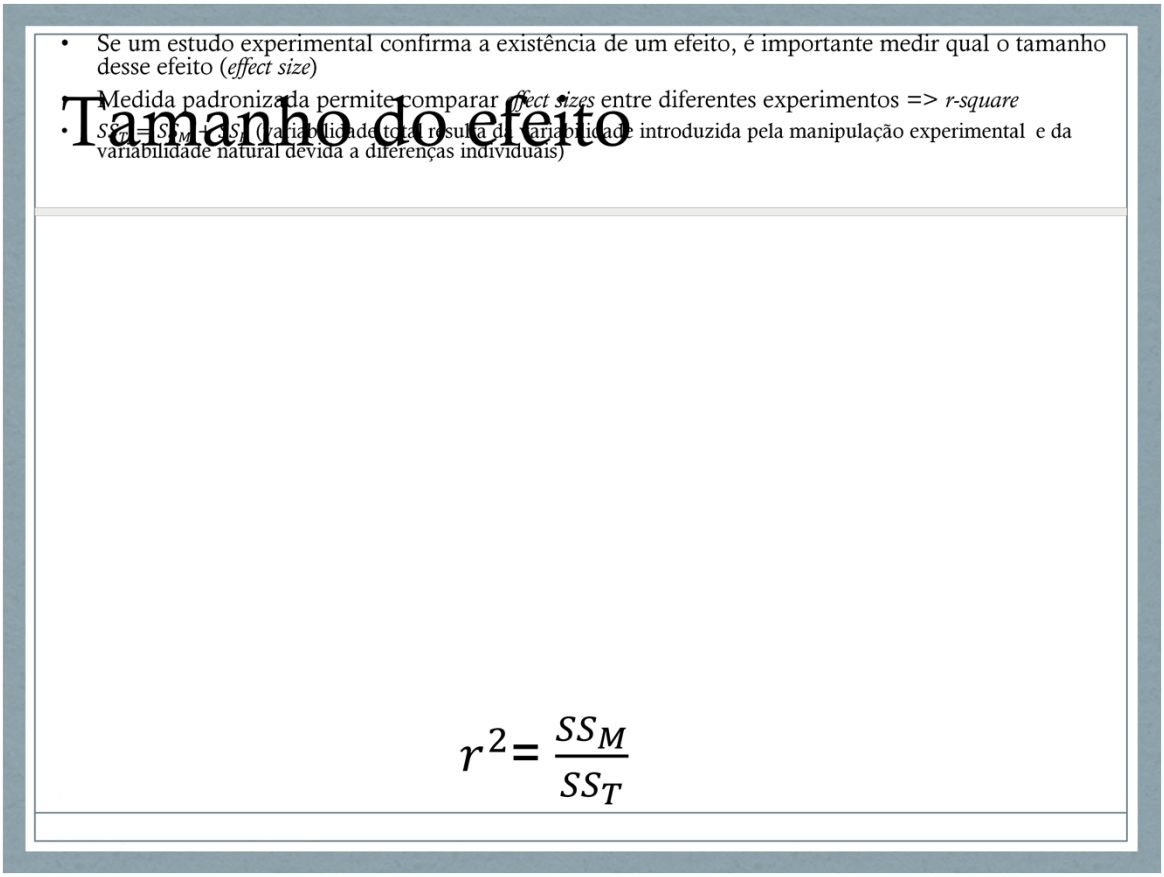
- *Within-group design*, porém pressupostos para o *paired samples t-test* não são satisfeitos => *Wilcoxon signed ranks test*
- Comparando três ou mais grupos, porém pressupostos para *one-way* ANOVA não satisfeitos
  - dados independentes entre si => *Kruskal-Wallis one-way analysis of variance by ranks*
  - dados dependentes entre si => *Friedman's two-way analysis of variance test*


http://marcomello.casadosmorcegos.org
Qual teste estatístico devo usar?
por Marco Mello (adaptado de Jutta Schmid)
Qual é a distribuição dos meus dados?
Há uma relação entre as minhas variáveis?
variável independente qualitativa ou quantitativa
variável independente nominal/ordinal, variável dependente quantitativa³
todas as variáveis quantitativas³
categorias independentes
categorias ligadas
sem causalidade
com causalidade
distrib. normal
distrib. outra
distrib. normal
distrib. outra
distrib. normal
distrib. outra
distrib. normal
distrib. outra
1 variável independente
Shapiro-Wilk¹
Kolmogorov-Smirnov¹
1. Costumam ser usados para testar a normalidade dos dados.
1 variável independente
2 categorias
teste t
Mann-Whitney
teste t pareado
Wilcoxon
1 variável independente
>2 categorias
ANOVA unifatorial
Kruskal-Wallis
ANOVA de medidas repetidas
Friedman
1 variável independente
correlação de Pearson
correlação de Spearman
regressão linear/não-linear simples
regressão não-paramétrica
>1 variável independente
qui-quadrado
teste G
teste exato de Fisher
teste binomial
>1 variável independente
ANOVA multifatorial
GLM ou GLzM²
2. Pode-se usar esses modelos quando há mistura de variáveis independentes quantitativas e qualitativas.
>1 variável independente
correlação parcial/ múltipla
correlação de Kendall
regressão múltipla/ stepwise
análise de caminhos
3. Se a a variável independente for quantitativa, mas a variável dependente for nominal e binária (e.g., sim ou não), você pode usar uma regressão logística.

implica necessariamente que o efeito observado seja relevante → o efeito pode ser muito pequeno

O que se pode ou não concluir de um teste estatístico

- Se os dados permitem rejeitar a hipótese nula, não implica que ela seja necessariamente verdadeira.



Suponha que fizemos um experimento para determinar qual de dois filmes de terror é mais assustador, o filme A e o filme B. Vamos descobrir medindo o batimento cardíaco das pessoas.

Claro que as medições vão variar: diferentes pessoas têm diferentes taxas de batimento cardíaco, e além disso elas assistiram filmes diferentes, e achamos que isso vai influenciar o seu batimento
(o filme mais assustador vai gerar taxas maiores...) .

Se quisermos descobrir o quanto os batimentos variaram entre as pessoas, podemos calcular (i) a média dos batimentos, e (ii) a soma dos erros quadrados entre as medidas e a média

SST = Total Sum of Squares: uma medida ´crua´ de quanto as medidas variaram nessa amostra.

Existem basicamente duas explicações para a variabilidade observada nas medidas dos batimentos: a manipulação experimental (no caso, assistir o filme A ou o filme B), e a variabilidade natural entre os batimentos das pessoas.

A parcela de variabilidade natural pode ser medida olhando a soma dos erros quadrados entre a medida de cada pessoa e a média do grupo no qual ela está (medida dentro do grupo)

SSR = Sum of Residual Squares

A parcela de variabilidade devida à manipulação experimental (pode ser medida)

SSM = Sum of Model Squares

SST = SSM + SSR

O efeito é grande se uma parcela grande da variabilidade observada é devida à manipulação experimental

report experiments. Sage Publications,
2010.

Bibliografia adicional