

# Laboratório de Fonologia



## Estatística para Linguística

**Prof. Dr. Adelino Pinheiro Silva**

# Sumário

Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências

Sumário

**Introdução**

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências

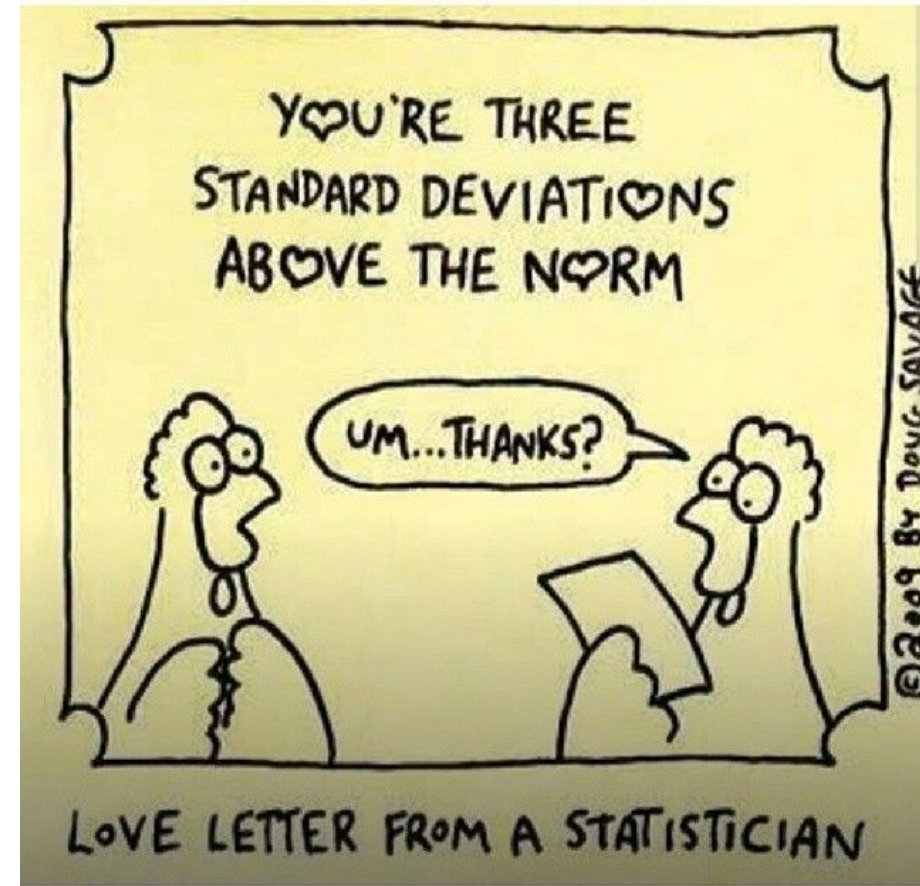
# In a hole in the ground there lived a...

Por que estudar estatística?

- Compreender **fatores** que afetam um resultado.
- Julgar de forma crítica as informações recebidas.
- Argumentar estatisticamente.

O que é estatística (Agresti, 2018)?

- Conjunto de métodos para se **obter** e **analisar** dados.
- Metodologia baseada na **ocorrência** para realizar **previsão**.



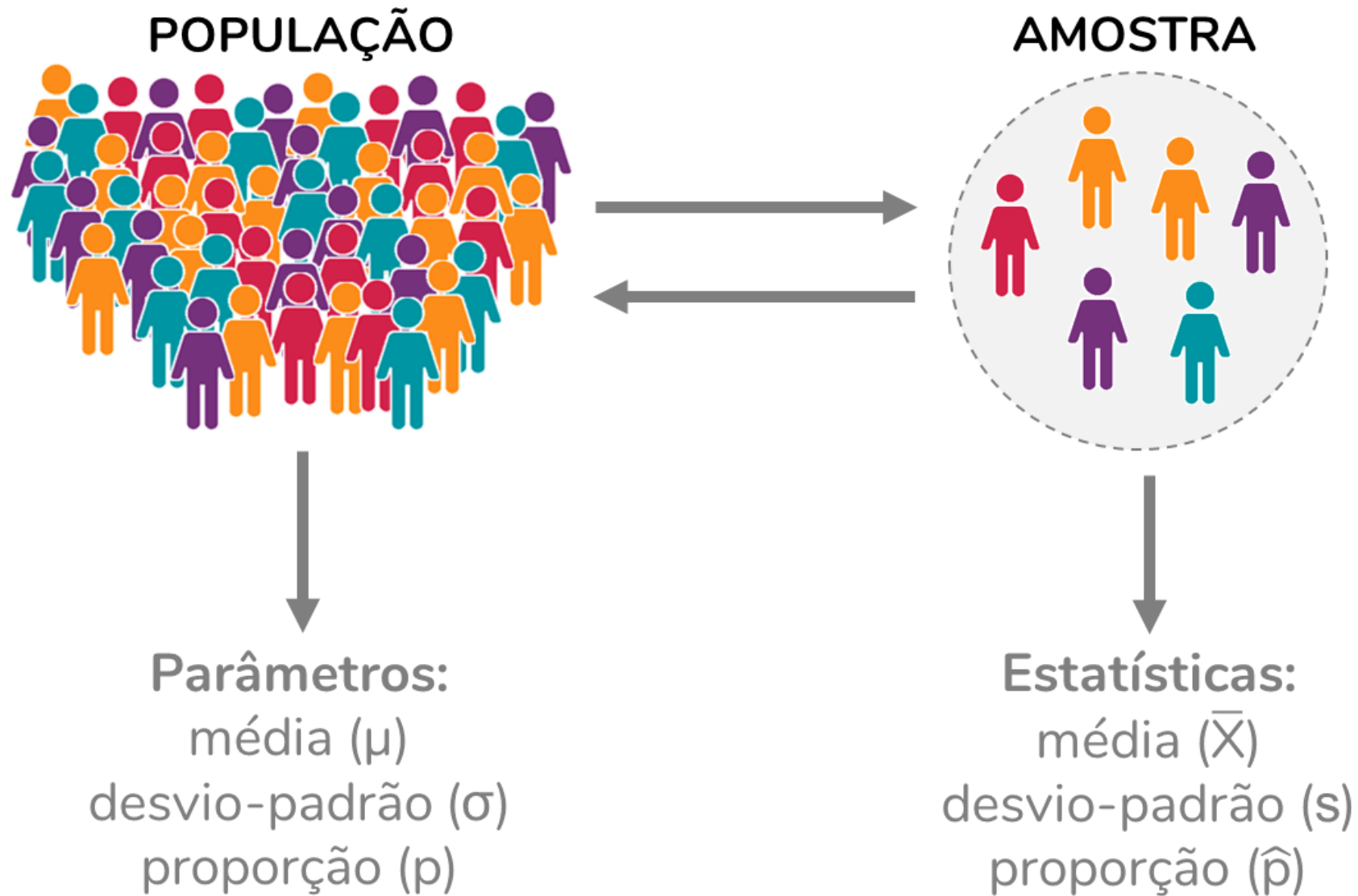
# I Have the High Ground

Alguns termos para começar

- **Dado:** Observação obtida sobre o objeto de interesse.
- **Observação:** Medida, ou informação coletada (sujeita a ruído e erros).
- **Base de dados:** Conjunto de dados, e.g., *general social survey*.
- **População:** Conjunto total dos elementos (desconhecido, inacessível).
- **Amostra:** subconjunto da população, dados (medidas) coletados.
- **Parâmetro:** Fator (resumo) numérico da população (dica: letras gregas).
- **Estatística:** Valor obtido da amostra !!!!!
- **Ferramental:** R-studio



# I Have the High Ground



# Medida e amostra

Maneiras de extrair informações de interesse.

- **Variável aleatória:** Característica que pode variar com os elementos da população ou amostra.
- **Escala de medição:** Extensão onde a variável aleatória pode ser medida. Exemplos:
  - Categóricas: (cara, coroa), (derrota, empate, vitória); ou
  - Quantitativas:  $\{x \in \mathbb{R} | 0 \leq x \leq 1\}$ ,  $[0, 1]$

Se caracteriza a variável aleatória como um resultado de uma experiência aleatória, que pode ser classificada como:

- **Categóricas:** valores aceitos dentro de um limite de categorias (qualitativos?).
- **Quantitativas:** valores numéricos de qualquer conjunto, e.g.,  $\mathbb{N}$ ,  $\mathbb{R}$ ,  $\mathbb{C}$

Escalas:

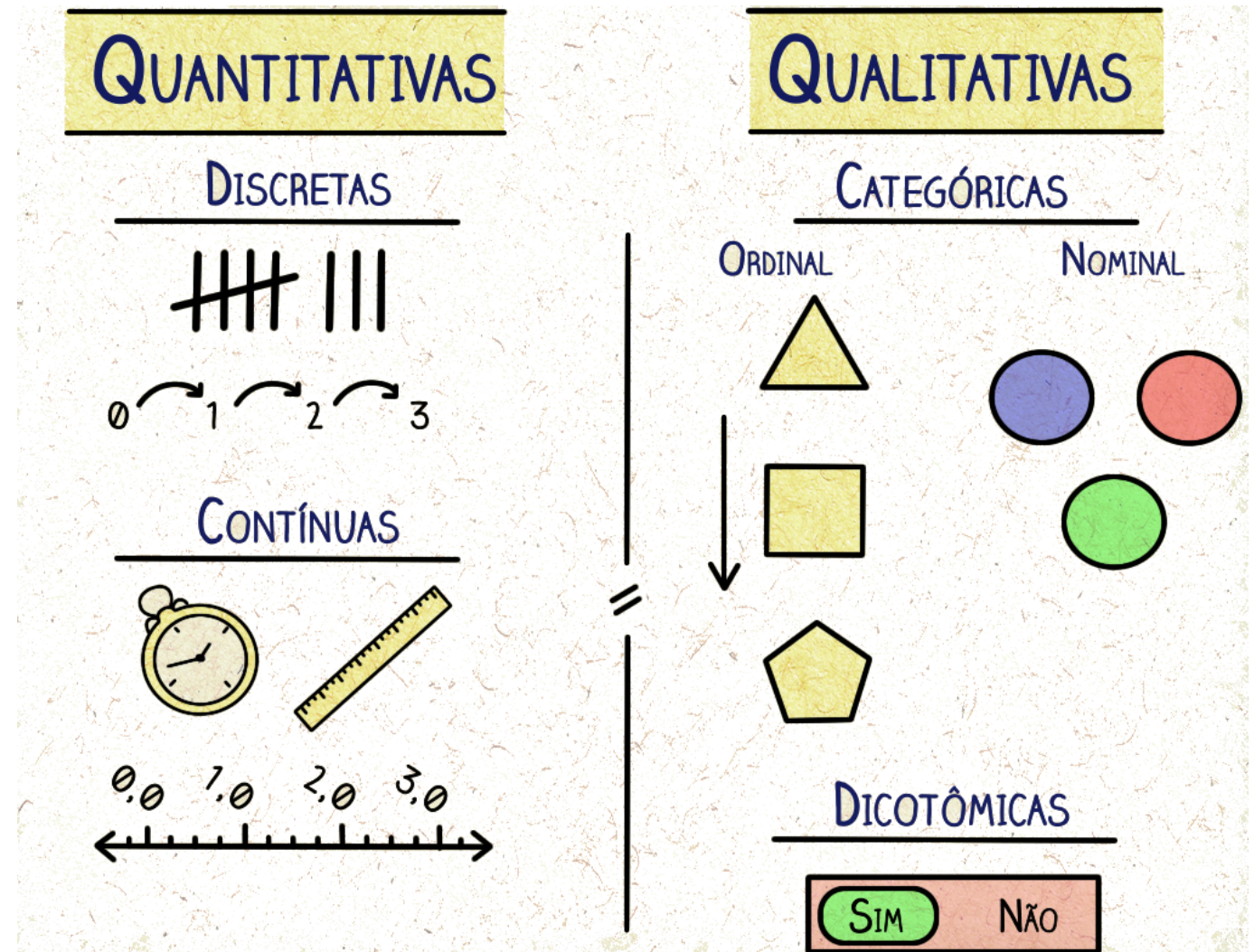
- **Intervalar:** delimitação numérica.
- **Nominal:** Nomes/categorias “não ordenáveis”, e.g., preferência de cores;
- **Ordenáveis:** Nomes/categorias que podem ser ordenadas em níveis, e.g., expectativa do curso (baixa, sem expectativa, alta).

Detalhe: Em escalas categóricas é muito difícil garantir uma homogeneidade dos intervalos, i.e., se os intervalos das categorias possuem escalas de mesmo tamanho.



# Variáveis estatísticas

- **Amostra aleatória simples:** todas amostras de mesmo tamanho possuem a mesma “chance”. Seria um retrato da população(?).
- **Métodos de amostragem, *sample survey*:** Sistemática, estratificada, grupo (*cluster*), multiestágios.
- **Amostra enviesada:** alunos de uma sala de aula (?).



# Estudo experimental

**Experimento:** Controlar variáveis independentes e observar a variação de variáveis dependentes para dar suporte ou refutar uma hipótese.

- Compara “tratamentos”.
- Unidades de testes.
- Grupos, pelo menos, “controle” e “tratamento”.
- Variáveis estranhas (predatórias).

## Problemas experimentais

- Variação do instrumento (ou pessoa que conduz parte dele).
- Regressão analítica.
- Viés de seleção.
- Perda de unidade



# Estudo experimental

Efeitos do teste: principal e interativo

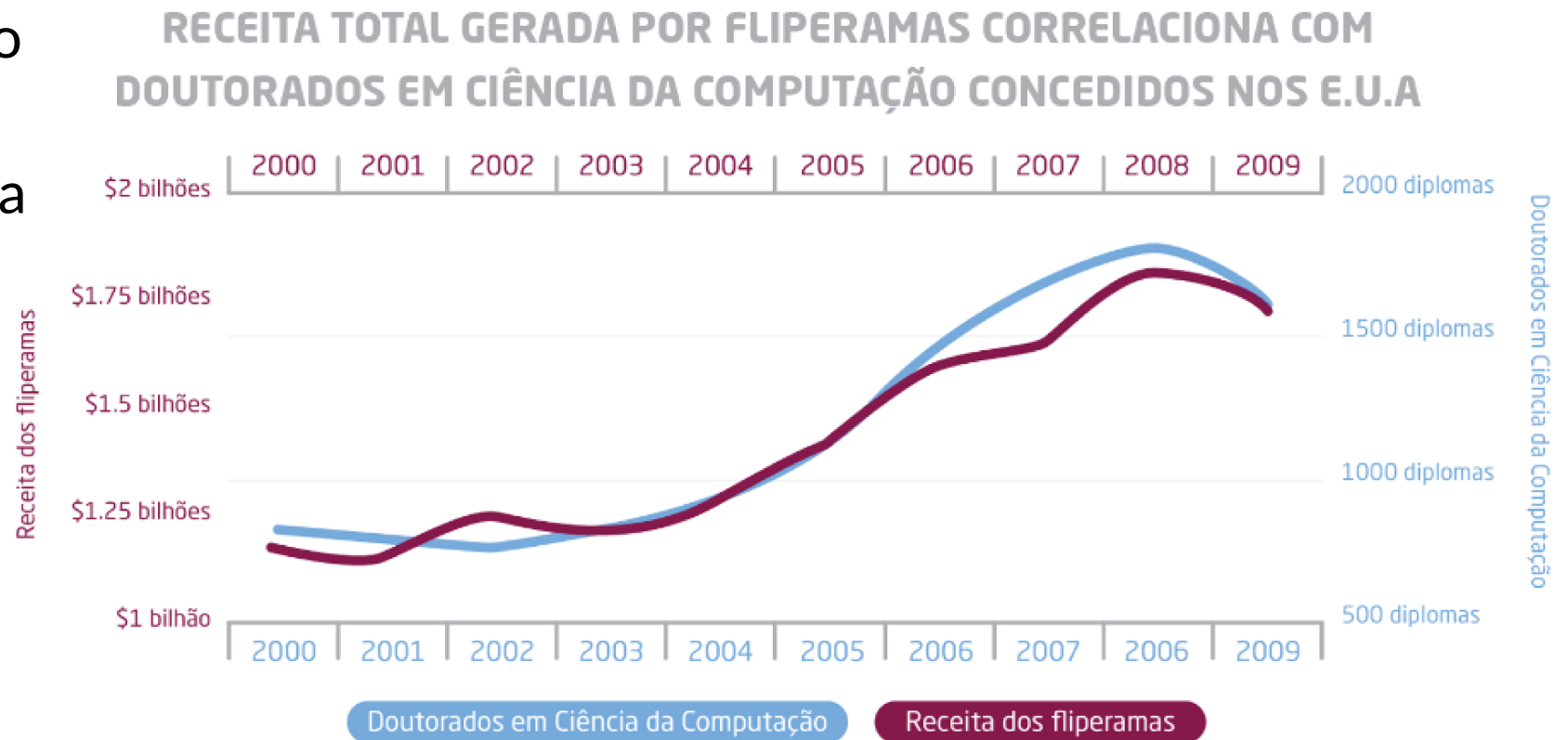
## Soluções para experimentos:

- Aleatorização.
- Emparelhamento.
- controle estatístico.
- Planejamento.
- Medições *a posteriori*.



# Estudo de Observação

- Sem manipulação do objeto de estudo.
- Grupos desbalanceados, difícil de realizar uma comparação adequada.
- **Não permite estabelecer causa e efeito.**
- Pode indicar uma relação entre variáveis.
- Uma variável não medida pode ser responsável pelo padrão observado.



# Variabilidade amostral e viés

**Erro de amostragem:** erro ocorrido ao utilizar uma estatística da amostra para prever um parâmetro da população. Exemplo: Erro da pesquisa eleitoral com  $n = 100$  de  $\pm 3\%$ .

**Viés:** erro quando a amostra é enviesada, e.g., voluntários ou respostas de carta.

- **Viés de resposta** ocorre quando a pergunta é confusa, e.g., referendo do desarmamento;
- **viés de falha de dados** apenas uma fatia da amostra responde.

# Fim da introdução - Dever de casa

---

## **Exercícios do livro Agresti (2018):**

- Capítulo 1: 1.1, 1.3, 1.5-1.8, 1.14, 1.16;
- Capítulo 2: 2.2-2.10, 2.27, 2.35-2.37, 2.39

## **Preparação do terreno**

- Instalar o R-studio.



Sumário

Introdução

**Estatística Descritiva**

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências



# Estatística descritiva

Primeiro passo para entender os dados coletados

Facilitar a assimilação de informação

Medidas de:

- tendência central (média),
- variabilidade
- associação

Análise e regressão: predizer uma variável a partir de outras.

Um pouco de código R para tratar com dados

```
data_lemas <- read.table("../Dados/lexporbr_alfa_lemas_txt.txt", header = TRUE,  
                           sep = "\t", dec = ",", quote = "\"")  
  
head(data_lemas)  
dim(data_lemas)  
summary(data_lemas)
```

Dados de Corpus Léxico do português

FALE





# Tabelas e gráficos

## Extraindo o cabeçalho dos dados

```
> head(data_lemas)
```

## Gera a saída:

	id	ortografia	cat_gram	inf_gram	freq_orto	freq_orto.M	log10_freq_orto	zipf_escala	nb_letras
1	1	o	gram	det	4364416	139093.06	6.6399	8.1433	1
2	2	de	gram	prp	2553292	81372.90	6.4071	7.9105	2
3	3	,	gram	pu	2133025	67979.08	6.3290	7.8324	1
4	4	.	gram	pu	1603184	51093.15	6.2050	7.7084	1
5	5	em	gram	prp	1044260	33280.36	6.0188	7.5222	2
6	6	e	gram	kc	667736	21280.61	5.8246	7.3280	1

A dimensionalidade dos dados, onde cada linha indica uma medição com as colunas indicando as informações

```
> dim(data_lemas)
```

Que é um total de 169.606 linhas com 9 colunas

```
[1] 169606      9
```

# Dever de casa

## Construindo uma tabela

```
> tab <- table(data_lemas$cat_gram, data_lemas$nb_letras)
```

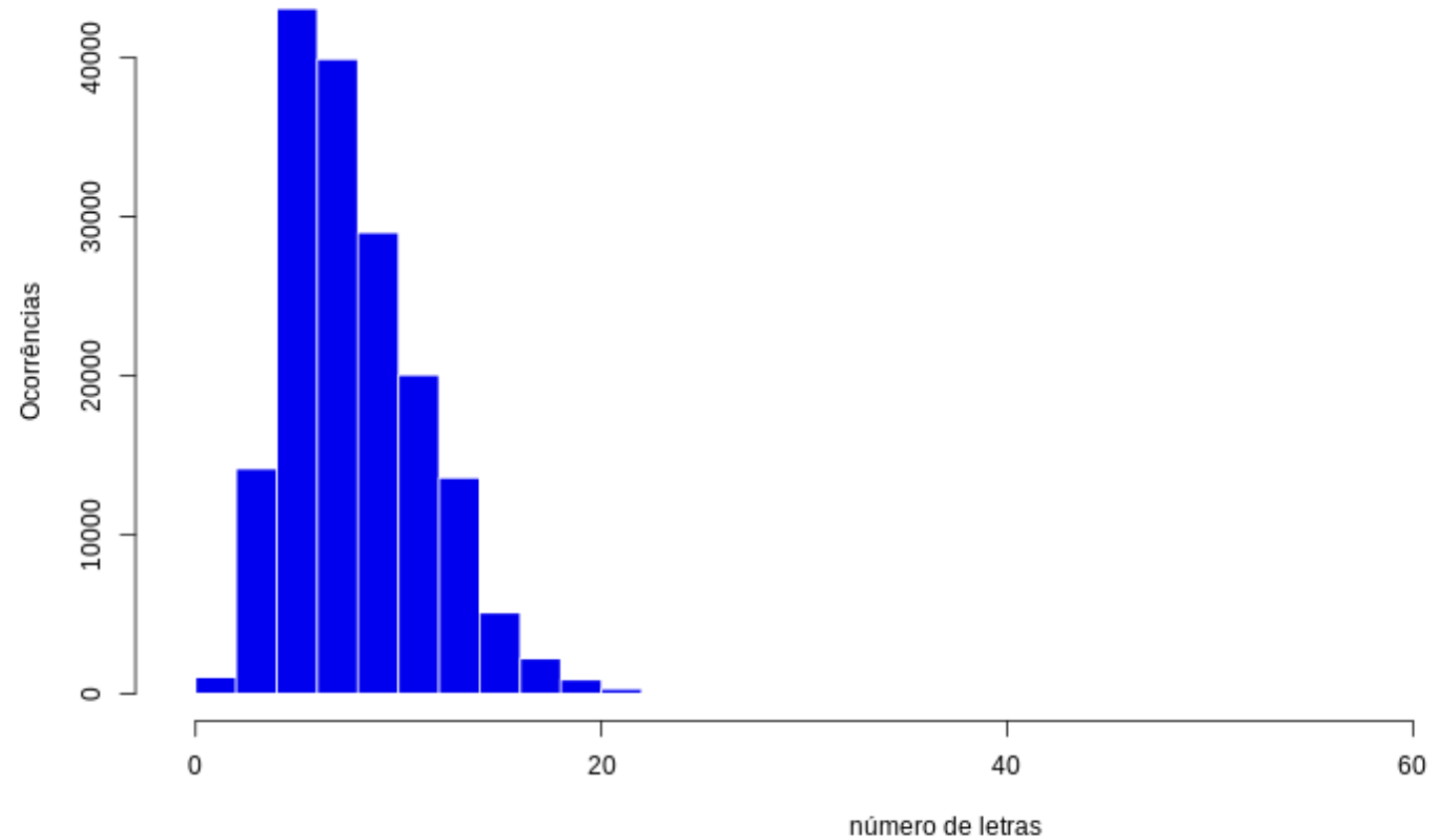
	1	2	3	4	5	6	7	8	9	10	11	12	13
adj	3	41	355	699	1061	1490	2320	2817	3117	3106	2672	2099	1575
adv	2	15	31	57	80	96	109	115	148	235	255	313	330
gram	56	60	96	75	75	47	22	12	14	8	1	2	0
nom	57	500	1517	3375	4992	5924	7396	7504	7397	6719	5912	4278	3151
num	11	275	2045	5678	14704	11792	8534	6535	3553	913	941	1630	657
ver	1	11	51	175	1037	1796	2099	2442	2115	1678	1123	815	475

# Histograma

## Histograma em uma figura PNG...

```
png(file = "../Imagens/histograma.png",width = 864, height = 486, units = "px")
hist(data_lemas$nb_letras,main="Histograma do numero de letras em cada ocorrencia",
breaks=40,xlab = "numero de letras", ylab="Ocorrencias",col = "blue2",
border="white")
dev.off()
```

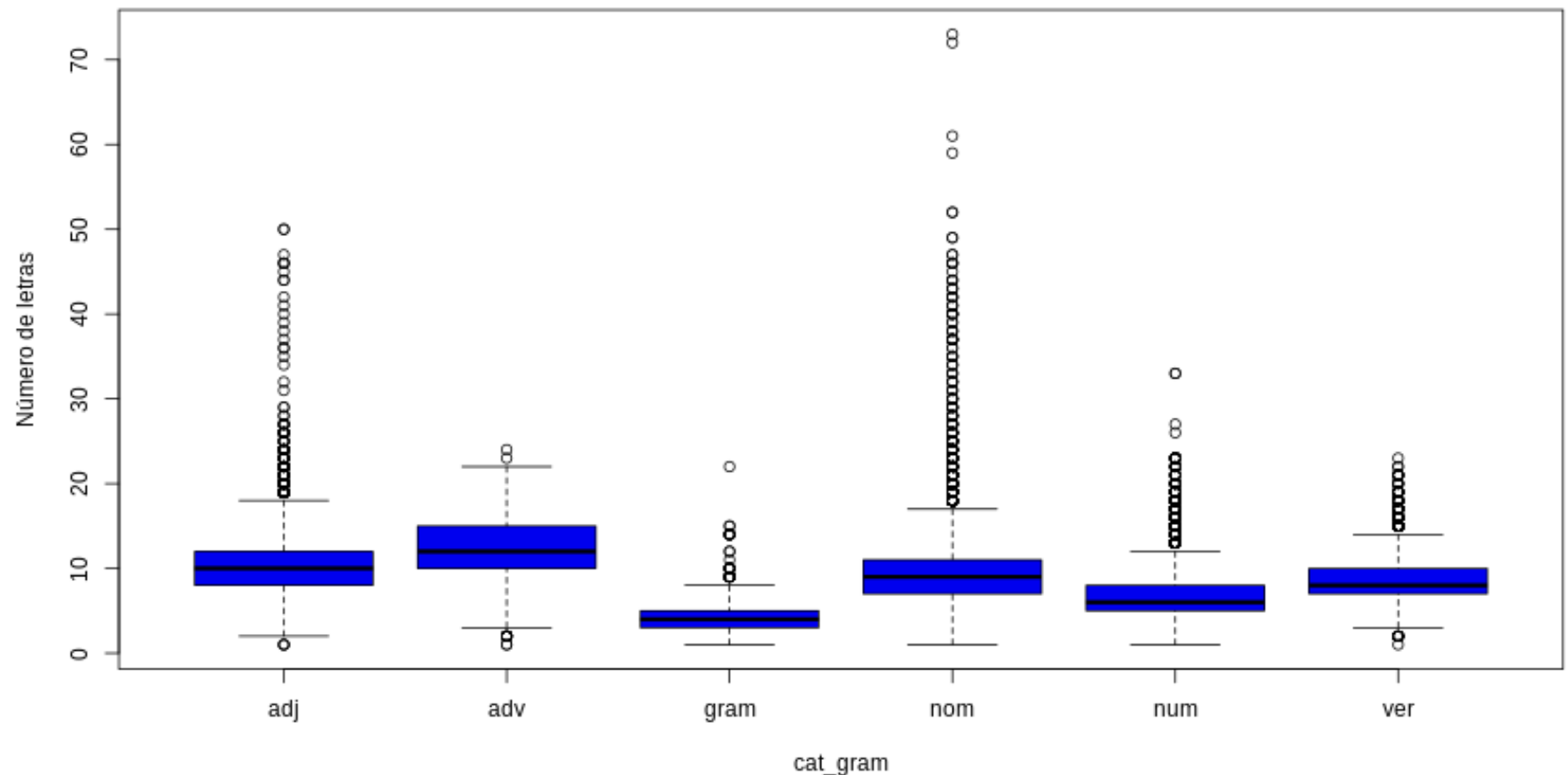
Histograma do numero de letras em cada ocorrência



# Diagrama de caixa

Diagrama de caixa (*boxplot*) em uma figura PNG...

```
png(file = "../Imagens/Box_plot.png", width = 864, height = 486, units = "px")
boxplot(data_lemas$nb_letras ~ cat_gram, data = data_lemas, ylab = "Numero de letras",
col = "blue2", border="black")
dev.off()
```

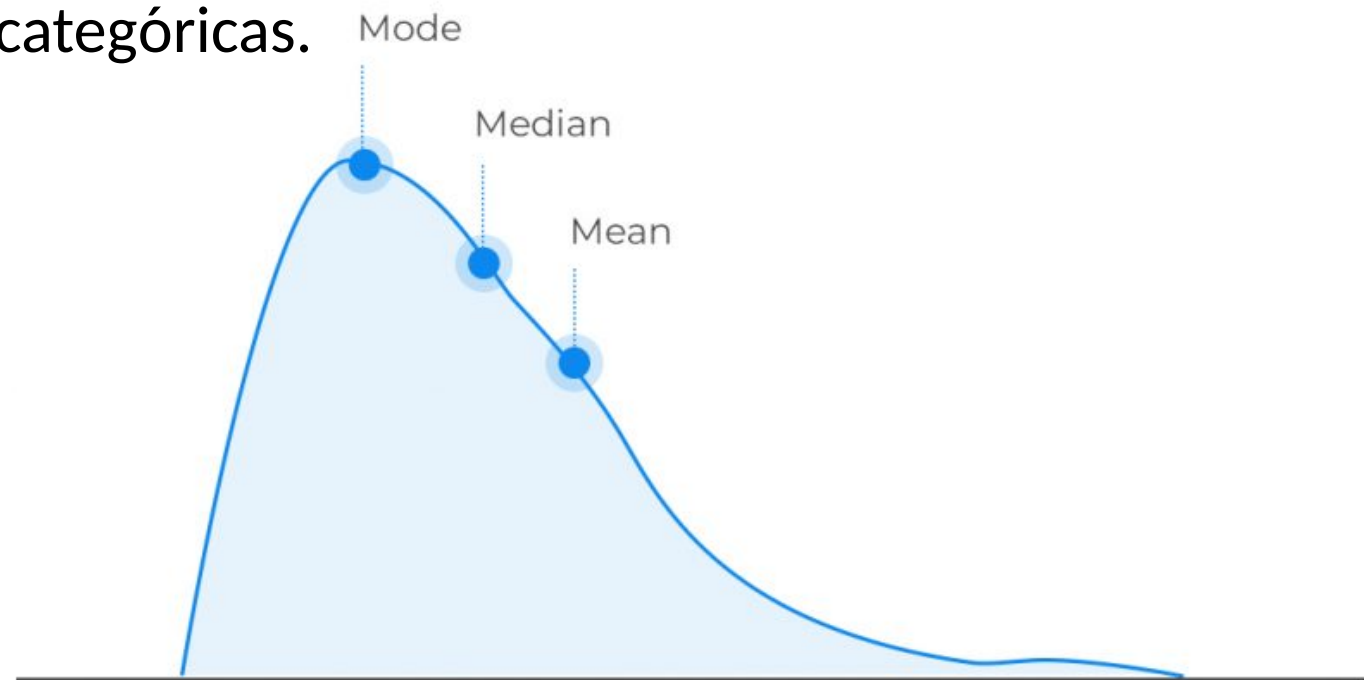


```
> tab <- stem(data_lemas$nb_letras[1:300])
```

[illegible]

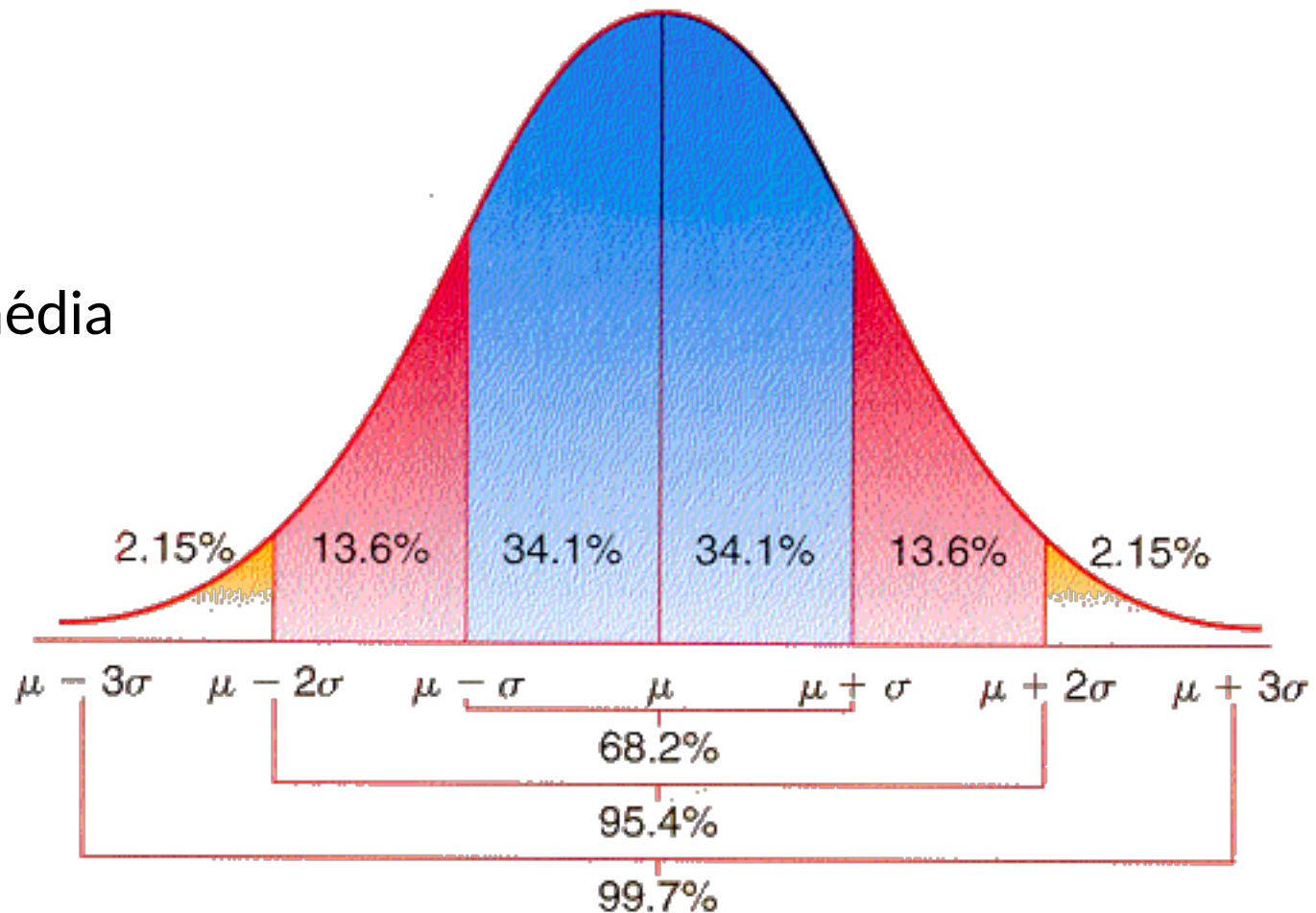
# Medidas de tendência central

- Média
  - Aritmética: problema que *outliers* podem alavancar.
  - Truncada (*winsorized*)
  - Ponderada
- Mediana: menos problemas com *outliers*.
- Moda: bem indicada para variáveis categóricas.
- Tri-média: utiliza quartis.



# Medidas de dispersão

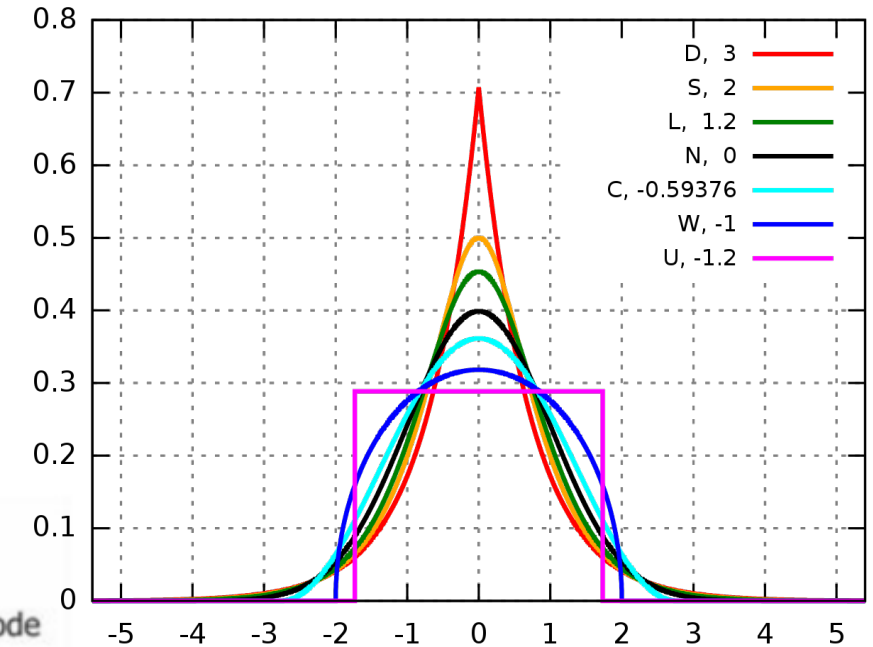
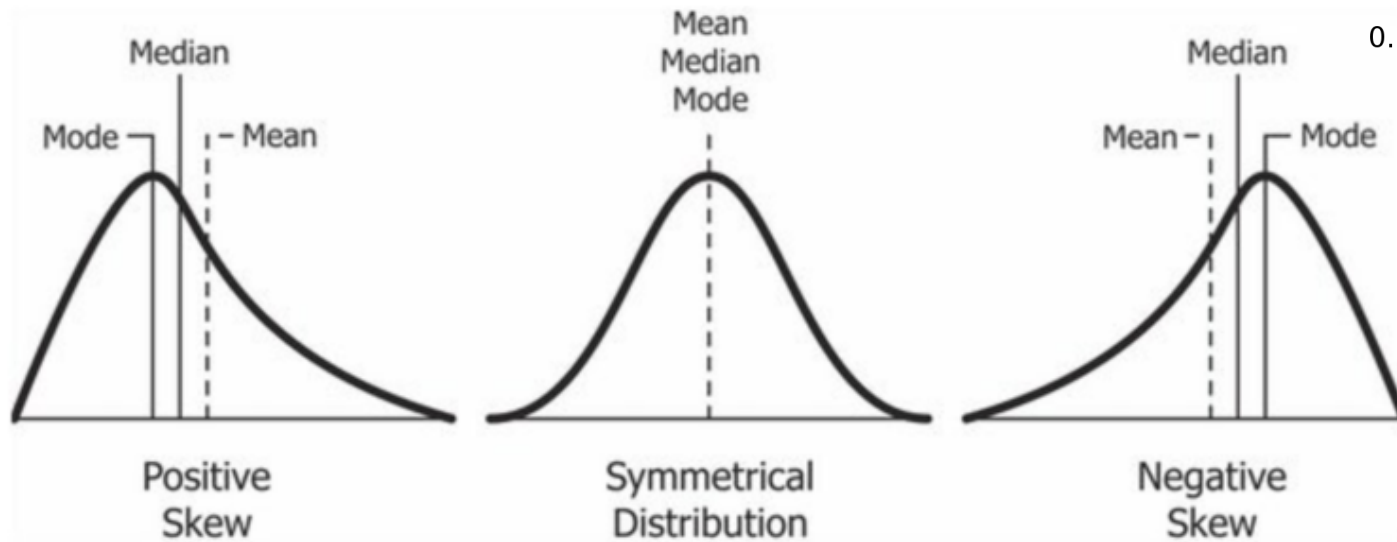
- *Range*, alcance, diferença entre mínimo e máximo.
- desvios
  - variância e desvio padrão.
  - Soma dos desvios quadrados.
- Distância inter-percentis.
- **Erro padrão**: desvio padrão da média
- Regra do  $\sigma$ :
  - $0,67\sigma \rightarrow 50\%$
  - $1\sigma \rightarrow 68,3\%$
  - $1,96\sigma \rightarrow 95\%$
  - $2\sigma \rightarrow 95,4\%$
  - $3\sigma \rightarrow 99,7\%$



# Achatamento(curtose) e (As)simetria

## Mais medidas de caracterização dos dados

```
library(moments)
curt_data <- kurtosis(vec_n_letras_sel)
assi_data <- skewness(vec_n_letras_sel)
```



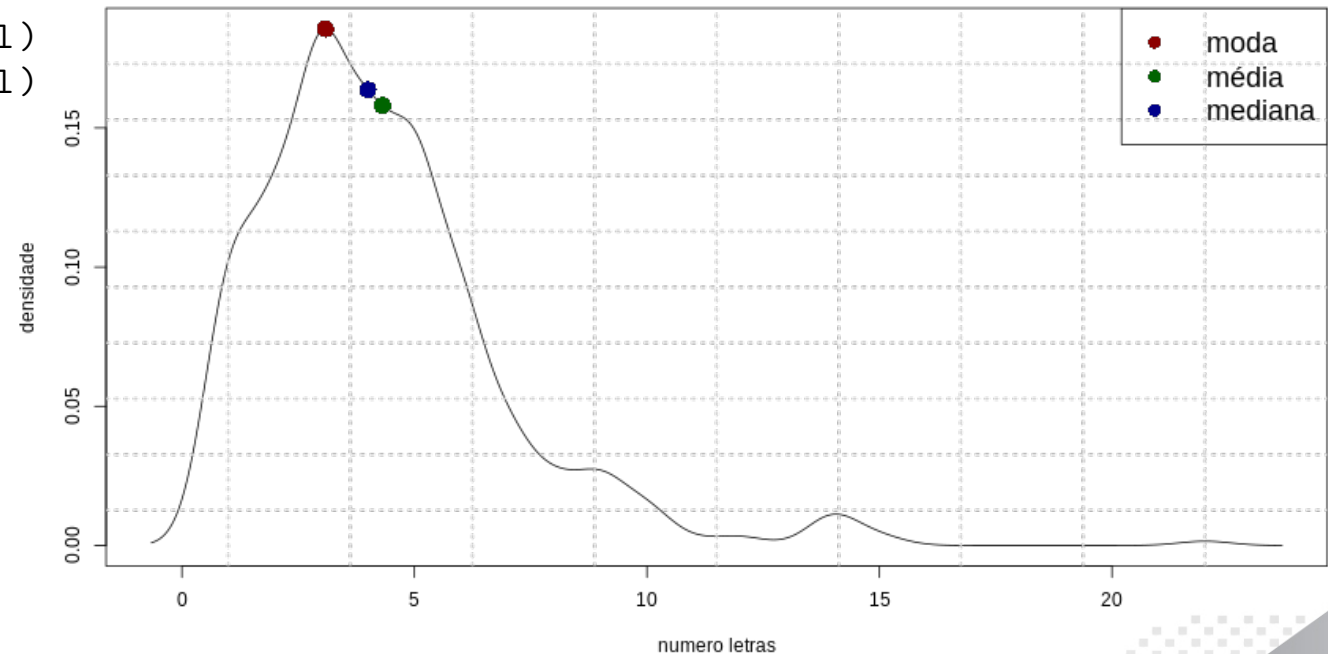


# De uma seleção de dados

## Extraindo algumas estatísticas dos dados:

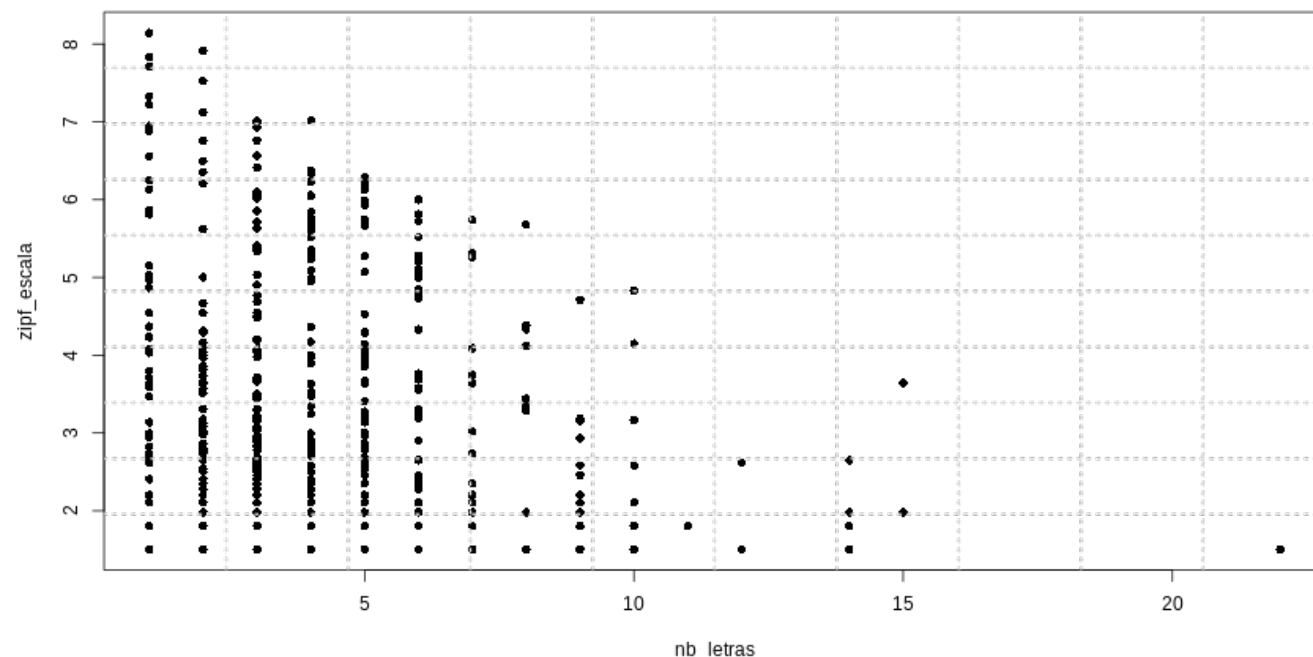
```
library(moments)
vec_n_letras_sel <- data_lemas[data_lemas$cat_gram
                             %in% 'gram',]$nb_letras
data_density <- density(vec_n_letras_sel,n=4096,
                        bw=1.2*bw.nrd(vec_n_letras_sel))
idx_max <- which.max(data_density$y)
moda_data <- data_density$y[idx_max]
mean_data <- mean(vec_n_letras_sel)
medi_data <- median(vec_n_letras_sel)
stdv_data <- sd(vec_n_letras_sel)
curt_data <- kurtosis(vec_n_letras_sel)
assi_data <- skewness(vec_n_letras_sel)
```

0.1855834  
4.313808  
4  
2.719188  
8.536706  
1.746169



# Representações bivariadas

- Tabelas de contingência
- Gráficos de dispersão
- Correlação
  - Pearson
  - Kendall
  - Spearman
- Informação mútua



Extraindo dados de duas variáveis:

```
vec_X_sel <- data_lemas[data_lemas$cat_gram %in% 'gram',]$nb_letras
vec_Y_sel <- data_lemas[data_lemas$cat_gram %in% 'gram',]$zipf_escala
png(file = "../Imagens/scatter_plot_02.png", width = 864, height = 486, bg = "transparent")
plot(x=vec_X_sel, y=vec_Y_sel, type='p', pch=16, xlab="log10_freq_orto", ylab="zipf_escala")
grid(10, lwd = 2)
dev.off()
cor(vec_X_sel, vec_Y_sel)
```

-0.2453342

# Fim da Estatística Descritiva - Dever de casa

## Exercícios do livro Agresti (2018):

- Capítulo 3: 1.1, 1.3, 1.5-1.8, 1.14, 1.16;

## Preparação do terreno

- Reproduzir os exemplos no R-Studio.

Lembrete:

**Parâmetros** de populações geralmente são representados por letras gregas, e.g.,  $\mu$  (média),  $\sigma^2$  (variância),  $\pi$  (proporção), etc...

**Estatísticas** são extraídas das amostras e representadas por letras latinas, com ou sem complemento, e.g.,  $m$ ,  $s^2$ ,  $p$ .

Sumário

Introdução

Estatística Descritiva

**Probabilidades**

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

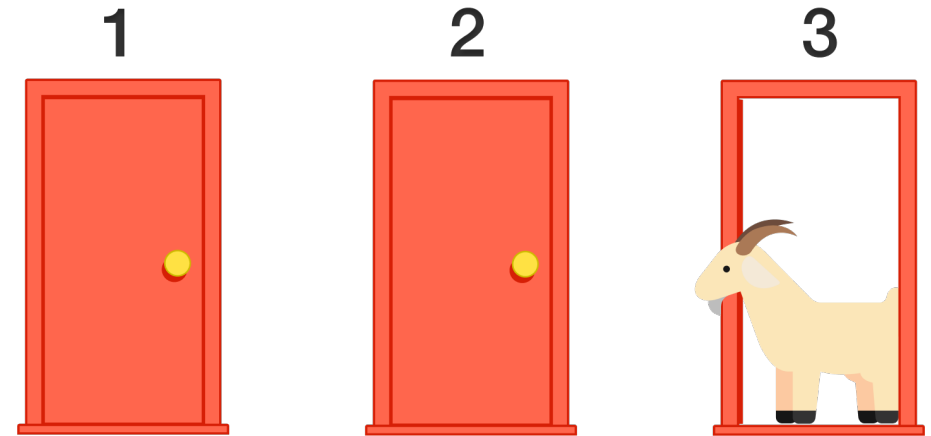
Introdução a métodos aprofundados

Encerramento

Referências

# Algunas definiciones

- Tabelas de contingência
- Gráficos de dispersão
- Correlação
  - Pearson
  - Kendall
  - Spearman
- Informação mútua





Sumário

Introdução

Estatística Descritiva

Probabilidades

**Estimação de Parâmetros**

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências







Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

**Teste de Significância**

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências





Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

**Comparação de dois grupos**

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências





# Assunto

Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

**Associação de Variáveis Categóricas**

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências









Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

**Regressão Linear e Correlação**

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências





Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

**Relação Multivariável**

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências







Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

**Regressão Múltipla e Correlação**

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências





Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

**Análise de Variância - ANOVA**

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências





Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

**Preditores Quantitativos e Categóricos**

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências







Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

**Modelos com Regressão Múltipla**

Regressão Logística

Introdução a métodos aprofundados

Encerramento

Referências





Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

**Regressão Logística**

Introdução a métodos aprofundados

Encerramento

Referências

# ▶ Dever de casa

---





Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

**Introdução a métodos aprofundados**

Encerramento

Referências





# ▶ Dever de casa

---



Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

**Encerramento**

Referências

# Sobre este material

Esta obra está licenciada sob a licença *Creative Commons* CC BY-NC-SA 4.0 (mais detalhes neste *link*)

Favor fazer referência a este trabalho como:

Silva, A. P. (2022), *Notas de Aulas de Estatística para Linguística*. Online:  
<https://github.com/adelinocpp/estatistica-para-linguistica>

```
@Misc{Silva2022,  
  title={Notas de Aulas de Notas de Aulas de Estatística para Linguística},  
  author={Adelino Pinheiro Silva},  
  howPublished={\url{https://github.com/adelinocpp/estatistica-para-linguistica}},  
  year={2022},  
  note={Version 1.0; Creative Commons BY-NC-SA 4.0.},  
}
```

Sumário

Introdução

Estatística Descritiva

Probabilidades

Estimação de Parâmetros

Teste de Significância

Comparação de dois grupos

Associação de Variáveis Categóricas

Regressão Linear e Correlação

Relação Multivariável

Regressão Múltipla e Correlação

Análise de Variância - ANOVA

Preditores Quantitativos e Categóricos

Modelos com Regressão Múltipla

Regressão Logística

Introdução a métodos aprofundados

Encerramento

**Referências**

# Referências I

---

Agresti, A. (2018). *Statistical methods for the social sciences*. Number 300.72 A3. Pearson.

