

Modelagem Preditiva

Filipe J. Zabala

Escola Politécnica
PUCRS
filipezabala.com

2024-08-09

Sumário

- 1 Minibio
- 2 Para começar
- 3 Sobre modelagem preditiva
- 4 Inferência bayesiana
- 5 Exemplos
- 6 Publicidade de dados PÚBLICOS
- 7 Privacidade de dados PESSOAIS
- 8 Para saber mais

Filipe J. Zabala · filipe.zabala@pucrs.br

- 2000-2004 Bacharel em Estatística [IME-UFRGS](#)
- 2006-2009 Mestre em Estatística [IME-USP](#)
- 2007-2009 Analista do [Banco Itaú S.A.](#)
- 2009- Sócio da [ZN Consultoria Estatística](#)
- 2010- Professor da [Escola Politécnica da PUCRS](#)
- 2019- Doutorando no [PPG Psiquiatria e C.C. UFRGS](#)

Para começar

- *Every once in a while there is house cleaning in mathematics. Some old names are discarded, some dusted off and refurbished; new theories, new additions to the household are assigned a place and name. Kasner and Newman (1940,3)*

Para começar

- *Every once in a while there is house cleaning in mathematics. Some old names are discarded, some dusted off and refurbished; new theories, new additions to the household are assigned a place and name. Kasner and Newman (1940,3)*
- Estatística vs Ciência de Dados vs Analytics vs IA vs ...

Para começar

- *Every once in a while there is house cleaning in mathematics. Some old names are discarded, some dusted off and refurbished; new theories, new additions to the household are assigned a place and name.* Kasner and Newman (1940,3)
- Estatística vs Ciência de Dados vs Analytics vs IA vs ...
- Teoria da Decisão vs Aprendizado por Reforço vs Aprendizado de Máquina

Para começar

- *Every once in a while there is house cleaning in mathematics. Some old names are discarded, some dusted off and refurbished; new theories, new additions to the household are assigned a place and name.* Kasner and Newman (1940,3)
- Estatística vs Ciência de Dados vs Analytics vs IA vs ...
- Teoria da Decisão vs Aprendizado por Reforço vs Aprendizado de Máquina
- Maximizar a utilidade esperada vs Maximizar a recompensa vs Minimizar o erro

Sobre modelagem preditiva

- Do Latim *praedicere*, anunciar antecipadamente

Sobre modelagem preditiva

- Do Latim *praedicere*, anunciar antecipadamente
- Métodos para prever novos valores de X
 - X : variável de interesse
 - θ : parâmetro associado a X

Sobre modelagem preditiva

- Do Latim *praedicere*, anunciar antecipadamente
- Métodos para prever novos valores de X
 - X : variável de interesse
 - θ : parâmetro associado a X
- *As duas culturas* de Leo Breiman (2001):
 - interpretar θ vs prever X

Sobre modelagem preditiva

- Do Latim *praedicere*, anunciar antecipadamente
- Métodos para prever novos valores de X
 - X : variável de interesse
 - θ : parâmetro associado a X
- *As duas culturas* de Leo Breiman (2001):
 - interpretar θ vs prever X
- Debabrata Basu (1988): *Information is what information does. It changes opinion (about θ).*

Sobre modelagem preditiva

- Do Latim *praedicere*, anunciar antecipadamente
- Métodos para prever novos valores de X
 - X : variável de interesse
 - θ : parâmetro associado a X
- *As duas culturas* de Leo Breiman (2001):
 - interpretar θ vs prever X
- Debabrata Basu (1988): *Information is what information does. It changes opinion (about θ).*
- George Box (1979): *All models are wrong but some are useful.*

Inferência bayesiana

- Priori: opinião (sobre θ) em forma de probabilidade antes de observar os dados

$$\pi(\theta)$$

Inferência bayesiana

- Priori: opinião (sobre θ) em forma de probabilidade antes de observar os dados

$$\pi(\theta)$$

- Verossimilhança: função (de θ) com informação dos dados

$$L(\theta|x)$$

Inferência bayesiana

- Priori: opinião (sobre θ) em forma de probabilidade antes de observar os dados

$$\pi(\theta)$$

- Verossimilhança: função (de θ) com informação dos dados

$$L(\theta|x)$$

- Posteriori: opinião (sobre θ) em forma de probabilidade depois de observar os dados

$$\pi(\theta|x)$$

Inferência bayesiana

- Operação bayesiana: calibrar a opinião à luz dos dados

$$\pi(\theta|x) = \frac{\pi(\theta)L(\theta|x)}{P(X=x)}$$

Inferência bayesiana

- Operação bayesiana: calibrar a opinião à luz dos dados

$$\pi(\theta|x) = \frac{\pi(\theta)L(\theta|x)}{P(X = x)}$$

- ‘A posteriori de hoje é a priori de amanhã’ (Máxima bayesiana)

Inferência bayesiana

- Operação bayesiana: calibrar a opinião à luz dos dados

$$\pi(\theta|x) = \frac{\pi(\theta)L(\theta|x)}{P(X = x)}$$

- 'A posteriori de hoje é a priori de amanhã' (Máxima bayesiana)
- Preditiva: distribuição de X

$$P(X = x) = \int_{\theta} \pi(\theta)L(\theta|x)d\theta$$

Inferência bayesiana

- Operação bayesiana: calibrar a opinião à luz dos dados

$$\pi(\theta|x) = \frac{\pi(\theta)L(\theta|x)}{P(X=x)}$$

- ‘A posteriori de hoje é a priori de amanhã’ (Máxima bayesiana)
- Preditiva: distribuição de X

$$P(X=x) = \int_{\theta} \pi(\theta)L(\theta|x)d\theta$$

- A probabilidade de o próximo resultado da moeda ser ‘cara’

$$Pr(X_{n+1} = \text{cara}) = \frac{r+1}{n+2}$$

Inferência bayesiana

- Variáveis permutáveis: a ordem das observações é indiferente

$$Pr(X_1 = x_1, \dots, X_N = x_N) = Pr(X_{\pi(1)} = x_1, \dots, X_{\pi(N)} = x_N)$$

Inferência bayesiana

- Variáveis permutáveis: a ordem das observações é indiferente

$$Pr(X_1 = x_1, \dots, X_N = x_N) = Pr(X_{\pi(1)} = x_1, \dots, X_{\pi(N)} = x_N)$$

- Teorema da representação de de Finetti (1930)

$$Pr(X_1 = x_1, \dots, X_N = x_N) = \int_{\theta} \theta^a (1 - \theta)^b \mu(d\theta)$$

Inferência bayesiana

- Variáveis permutáveis: a ordem das observações é indiferente

$$Pr(X_1 = x_1, \dots, X_N = x_N) = Pr(X_{\pi(1)} = x_1, \dots, X_{\pi(N)} = x_N)$$

- Teorema da representação de de Finetti (1930)

$$Pr(X_1 = x_1, \dots, X_N = x_N) = \int_{\theta} \theta^a (1 - \theta)^b \mu(d\theta)$$

- Flexibiliza a suposição de independência

Inferência bayesiana

- Variáveis permutáveis: a ordem das observações é indiferente

$$Pr(X_1 = x_1, \dots, X_N = x_N) = Pr(X_{\pi(1)} = x_1, \dots, X_{\pi(N)} = x_N)$$

- Teorema da representação de de Finetti (1930)

$$Pr(X_1 = x_1, \dots, X_N = x_N) = \int_{\theta} \theta^a (1 - \theta)^b \mu(d\theta)$$

- Flexibiliza a suposição de independência
- Trata θ apenas como uma variável de integração

Ex. 1: Previsão de demanda

- Objetivo: prever a demanda de bebida em função da temperatura máxima do dia
- Seção 7.2 de [Zabala \(2024\)](#) - Estatística Básica
- Y : número de garrafas de bebida consumidas
- X : temperatura máxima do dia em $^{\circ}\text{C}$
- Modelo: $y = \hat{\beta}_0 + \hat{\beta}_1 x$

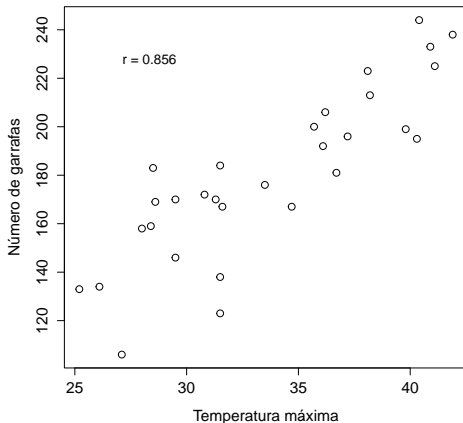
Ex. 1: Previsão de demanda

Obtendo dados e estatísticas descritivas.

```
dr <- read.table('https://filipezabala.com/data/drinks.txt',  
                 header = TRUE)  
str(dr) # estrutura dos dados  
  
## 'data.frame': 30 obs. of 2 variables:  
## $ temp: num 29.5 31.3 34.7 40.4 28.4 40.3 41.1 36.2 35.7 26.1 ...  
## $ gar : int 146 170 167 244 159 195 225 206 200 134 ...  
  
summary(dr)  
  
##           temp           gar  
## Min.      :25.20   Min.      :106.0  
## 1st Qu.:29.50     1st Qu.:161.0  
## Median :32.55     Median :178.5  
## Mean    :33.66     Mean    :180.0  
## 3rd Qu.:37.88     3rd Qu.:199.8  
## Max.    :41.90     Max.    :244.0
```

Ex. 1: Previsão de demanda

```
plot(dr, xlab = 'Temperatura máxima', ylab = 'Número de garrafas',  
      cex = 1.3, cex.axis = 1.3, cex.lab = 1.3)  
legend(26, 235, legend=paste0('r = ', round(cor(dr$temp, dr$gar), 3)),  
       cex = 1.2, box.lty = 0)
```



Ex. 1: Previsão de demanda

```
fit <- lm(gar ~ temp, data = dr) # modelo clássico
summary(fit)

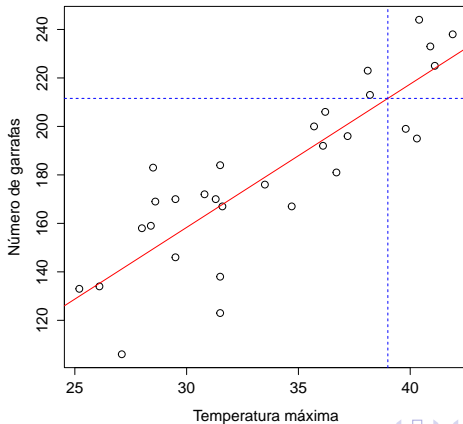
##
## Call:
## lm(formula = gar ~ temp, data = dr)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -44.204  -8.261   3.518  10.796  33.540
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -19.1096    22.9195  -0.834    0.411
## temp         5.9147     0.6736   8.780 1.57e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.24 on 28 degrees of freedom
## Multiple R-squared:  0.7336, Adjusted R-squared:  0.7241
## F-statistic: 77.1 on 1 and 28 DF,  p-value: 1.565e-09

(pr <- predict(fit, newdata = data.frame(temp = 39)))

##      1
## 211.5649
```

Ex. 1: Previsão de demanda

```
plot(dr, xlab = 'Temperatura máxima', ylab = 'Número de garrafas',  
      cex = 1.3, cex.axis = 1.3, cex.lab = 1.3)  
abline(fit$coefficients[1], fit$coefficients[2], col = 'red')  
abline(v = 39, col = 'blue', lty = 2)  
abline(h = pr, col = 'blue', lty = 2)
```



Ex. 1: Previsão de demanda

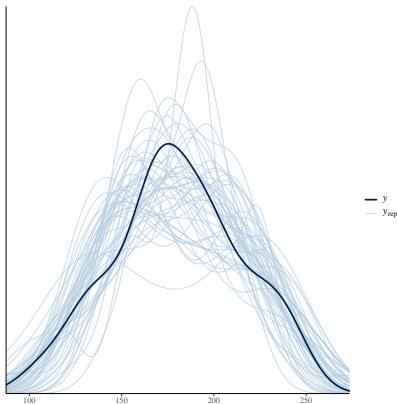
```
# modelo bayesiano
options(mc.cores = parallel::detectCores())
fit_stan <- rstanarm::stan_glm(gar ~ temp, data = dr, family = gaussian)
fit_stan

## stan_glm
## family:          gaussian [identity]
## formula:         gar ~ temp
## observations:    30
## predictors:      2
## -----
##               Median MAD_SD
## (Intercept) -18.9    23.1
## temp         5.9     0.7
##
## Auxiliary parameter(s):
##               Median MAD_SD
## sigma 18.6     2.4
##
## -----
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

Ex. 1: Previsão de demanda

If the model fits, then replicated data generated under the model should look similar to observed data. Gelman et al (2013,143)

```
rstanarm::pp_check(fit_stan) # posterior predictive checking
```



Ex. 2: Previsão de volume processual

- `jurimetrics`: ferramentas para Jurimetria

Ex. 2: Previsão de volume processual

- `jurimetrics`: ferramentas para Jurimetria
- `tjrs_2000_2017`: série mensal de jan/2000 a dez/2017 da cota inferior de volume processual no segundo grau do TJ-RS

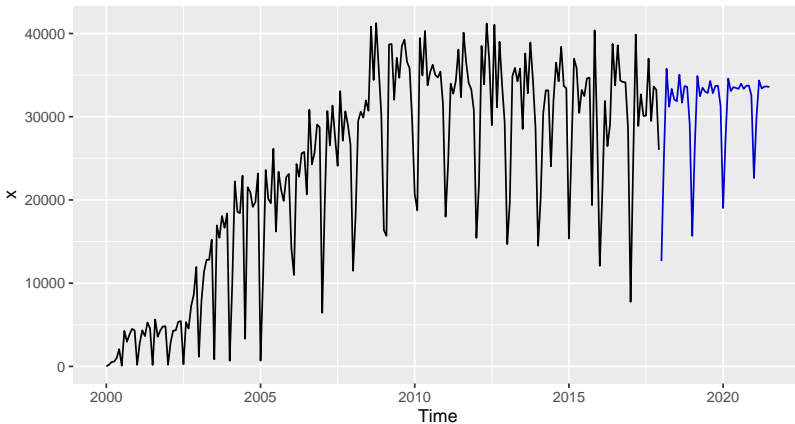
```
jurimetrics::tjrs_00_17$count_adjusted
```

##	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
## 2000	10	199	517	581	987	2074	96	4264	2967	3824	4510	4332
## 2001	197	2832	4334	3644	5275	4571	178	5638	3558	4330	4804	4824
## 2002	196	2863	4309	4326	5359	5447	266	5317	4546	7253	8607	11955
## 2003	1162	7845	11361	12819	12804	15229	856	16947	15460	18073	16656	18377
## 2004	677	10013	22227	18562	18420	22918	3314	21522	20948	19164	19747	23187
## 2005	683	10787	23602	20146	19612	26153	16213	23389	21239	19886	22740	23127
## 2006	14199	10986	24322	22791	25609	25769	20681	30839	24275	25657	29061	28746
## 2007	6447	19618	30692	26569	31335	27542	24096	33065	27126	30646	29016	26661
## 2008	11490	18158	29504	30598	29892	31958	30728	40828	34408	41238	35749	30162
## 2009	16367	15698	38680	38735	32054	37082	34679	38500	39260	36593	35815	29585
## 2010	20663	18758	39455	34965	40298	33780	35349	36246	34999	34714	35417	31460
## 2011	17990	24421	33980	32770	34319	38069	32344	40096	36581	34035	33303	30798
## 2012	15440	21913	38503	33899	41206	36746	28989	41048	31124	39001	33790	29317
## 2013	14693	19745	34876	35865	34252	35780	28517	37606	32825	38911	34410	28663
## 2014	14497	20139	30409	33173	33156	24049	31945	36489	34256	38395	33705	33406
## 2015	15362	26139	36961	35838	30456	33209	32480	34602	34688	19377	40370	27943
## 2016	12086	20802	31884	26466	29055	38739	33769	38571	34335	34177	34116	28583
## 2017	7754	24462	39869	28881	32684	30093	30144	36987	29486	33621	33241	26020

Ex. 2: Previsão de volume processual

```
library(jurimetrics)
y <- ts(tjrs_00_17$count_adjusted, start = c(2000,1), frequency = 12)
fits(y, show.sec.graph = FALSE, show.value = FALSE)
```

Forecasts from NNAR(3,1,2)[12]



Ex. 3: Lei de Newcomb-Benford

- *That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones.* Simon Newcomb (1881,39)

Ex. 3: Lei de Newcomb-Benford

- *That the ten digits do not occur with equal frequency must be evident to any one making much use of logarithmic tables, and noticing how much faster the first pages wear out than the last ones.* **Simon Newcomb** (1881,39)

Dig.				First Digit.	Second Digit.
0	0.1197
1	.	.	.	0.3010	0.1139
2	.	.	.	0.1761	0.1088
3	.	.	.	0.1249	0.1043
4	.	.	.	0.0969	0.1003
5	.	.	.	0.0792	0.0967
6	.	.	.	0.0669	0.0934
7	.	.	.	0.0580	0.0904
8	.	.	.	0.0512	0.0876
9	.	.	.	0.0458	0.0850

1

¹ Newcomb (1881,40) - Note on the Frequency of Use of the Different Digits in Natural Numbers

Ex. 3: Lei de Newcomb-Benford

TABLE I
PERCENTAGE OF TIMES THE NATURAL NUMBERS 1 TO 9 ARE USED AS FIRST
DIGITS IN NUMBERS, AS DETERMINED BY 20,229 OBSERVATIONS

Group	Title	First Digit									Count
		1	2	3	4	5	6	7	8	9	
A	Rivers, Area	31.0	16.4	10.7	11.3	7.2	8.6	5.5	4.2	5.1	335
B	Population	33.9	20.4	14.2	8.1	7.2	6.2	4.1	3.7	2.2	3259
C	Constants	41.3	14.4	4.8	8.6	10.6	5.8	1.0	2.9	10.6	104
D	Newspapers	30.0	18.0	12.0	10.0	8.0	6.0	6.0	5.0	5.0	100
E	Spec. Heat	24.0	18.4	16.2	14.6	10.6	4.1	3.2	4.8	4.1	1389
F	Pressure	29.6	18.3	12.8	9.8	8.3	6.4	5.7	4.4	4.7	703
G	H.P. Lost	30.0	18.4	11.9	10.8	8.1	7.0	5.1	5.1	3.6	690
H	Mol. Wgt.	26.7	25.2	15.4	10.8	6.7	5.1	4.1	2.8	3.2	1800
I	Drainage	27.1	23.9	13.8	12.6	8.2	5.0	5.0	2.5	1.9	159
J	Atomic Wgt.	47.2	18.7	5.5	4.4	6.6	4.4	3.3	4.4	5.5	91
K	n^{-1} , \sqrt{n} , ...	25.7	20.3	9.7	6.8	6.6	6.8	7.2	8.0	8.9	5000
L	Design	26.8	14.8	14.3	7.5	8.3	8.4	7.0	7.3	5.6	560
M	<i>Digest</i>	33.4	18.5	12.4	7.5	7.1	6.5	5.5	4.9	4.2	308
N	Cost Data	32.4	18.8	10.1	10.1	9.8	5.5	4.7	5.5	3.1	741
O	X-Ray Volts	27.9	17.5	14.4	9.0	8.1	7.4	5.1	5.8	4.8	707
P	Am. League	32.7	17.6	12.6	9.8	7.4	6.4	4.9	5.6	3.0	1458
Q	Black Body	31.0	17.3	14.1	8.7	6.6	7.0	5.2	4.7	5.4	1165
R	Addresses	28.9	19.2	12.6	8.8	8.5	6.4	5.6	5.0	5.0	342
S	$n^1, n^2 \dots n!$	25.3	16.0	12.0	10.0	8.5	8.8	6.8	7.1	5.5	900
T	Death Rate	27.0	18.6	15.7	9.4	6.7	6.5	7.2	4.8	4.1	418
Average		30.6	18.5	12.4	9.4	8.0	6.4	5.1	4.9	4.7	1011
Probable Error		± 0.8	± 0.4	± 0.4	± 0.3	± 0.2	± 0.2	± 0.2	± 0.2	± 0.3	—

2

Ex. 3: Lei de Newcomb-Benford

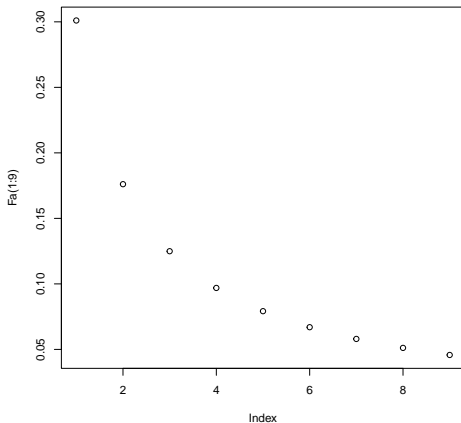
- Frequência do dígito a na 1^a posição

$$F_a = \log_{10} \left(\frac{a+1}{a} \right)$$

$$a = 1, 2, \dots, 9$$

Ex. 3: Lei de Newcomb-Benford

```
Fa <- function(a){log((a+1)/a, base = 10)}  
plot(Fa(1:9))
```



Ex. 3: Lei de Newcomb-Benford

- Frequência do dígito b na 2^a posição seguindo um dígito a na 1^a posição

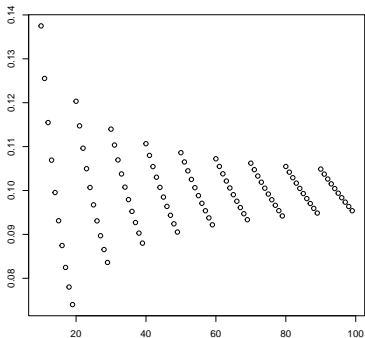
$$F_{ab} = \frac{\log_{10} \left(\frac{ab+1}{ab} \right)}{\log_{10} \left(\frac{a+1}{a} \right)}$$

$$ab = \{1, 2, \dots, 9\} \times \{0, 1, \dots, 9\}$$

```
Fab <- function(a,b){  
  ab <- as.numeric(paste0(a,b))  
  fab <- log((ab+1)/ab, base = 10)/Fa(a)  
  return(list(ab=ab, Fab=fab))  
}  
Fab(5,0)  
  
## $ab  
## [1] 50  
##  
## $Fab  
## [1] 0.1086137
```

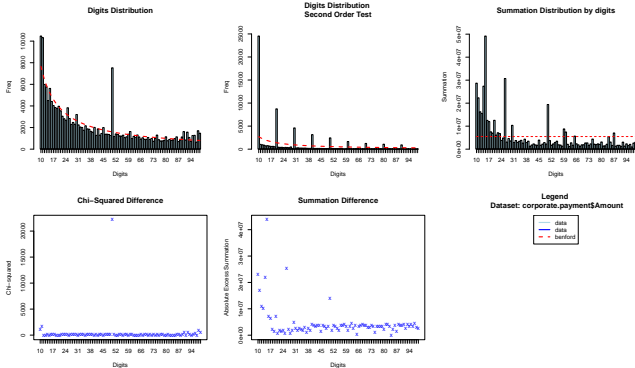
Ex. 3: Lei de Newcomb-Benford

```
grade <- expand.grid(1:9,0:9)
grade <- sort(as.numeric(paste0(grade[,1],grade[,2])))
prob <- data.frame(green=grade, Fab=NA)
k <- 0
for(i in 1:9){
  for(j in 0:9){
    k <- k+1
    prob[k,2] <- Fab(i,j)$Fab
  }
}
plot(prob[,1], prob[,2], xlab = '', ylab='')
```



Ex. 3: Lei de Newcomb-Benford

```
library(benford.analysis)
data(corporate.payment)
bfd <- benford(corporate.payment$Amount)
plot(bfd)
```



Ex. 4: voice

- **voice**: ferramentas para análise de voz, reconhecimento de falantes e inferência de humor

Ex. 4: voice

- **voice**: ferramentas para análise de voz, reconhecimento de falantes e inferência de humor

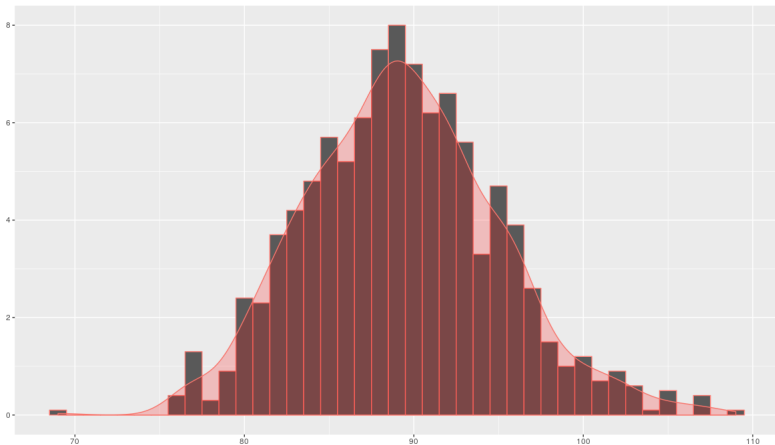
```
library(voice)
path2wav <- list.files(system.file('extdata', package = 'wrassp'),
                       pattern = glob2rx('*.wav'), full.names = TRUE)
E <- dplyr::tibble(subject_id = c(1,1,1,2,2,2,3,3,3), wav_path = path2wav)

# resume o áudio por sujeito
voice::tag(E, groupBy = 'subject_id')
```



```
## # A tibble: 3 x 7
##   subject_id f0_tag_mean f0_tag_sd f0_tag_vc f0_tag_median f0_tag_iqr f0_tag_mad
##   <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
## 1         1         85.1         15.3         0.180         78.3         26.8         11.9
## 2         2         84.6         14.9         0.176         76.4         28.3          7.97
## 3         3         81.0         14.6         0.180         75.6         21.6          8.68
```

Ex. 5: Tempo de operação de ATMs



Publicidade de dados PÚBLICOS

- Brasil (2011) Lei 12.527 de 18/11/2011
- Brasil(2012) Brasil. Decreto 7.724 de 16/05/2012



Presidência da República
Casa Civil
Subchefia para Assuntos Jurídicos

DECRETO Nº 7.724, DE 16 DE MAIO DE 2012

Vigência

Regulamenta a Lei nº 12.527, de 18 de novembro de 2011, que dispõe sobre o acesso a informações previsto no inciso XXXIII do **caput** do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição.

A PRESIDENTA DA REPÚBLICA, no uso das atribuições que lhe confere o art. 84, **caput**, incisos IV e VI, alínea "a", da Constituição, e tendo em vista o disposto na Lei nº 12.527, de 18 de novembro de 2011,

DECRETA:

CAPÍTULO I

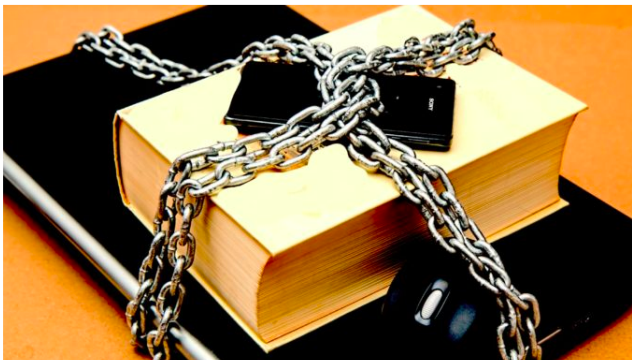
DISPOSIÇÕES GERAIS

Art. 1º Este Decreto regulamenta, no âmbito do Poder Executivo federal, os procedimentos para a garantia do acesso à informação e para a classificação de informações sob restrição de acesso, observados grau e prazo de sigilo, conforme o disposto na Lei nº 12.527, de 18 de novembro de 2011, que dispõe sobre o acesso a informações previsto no inciso XXXIII do caput do art. 5º, no inciso II do § 3º do art. 37 e no § 2º do art. 216 da Constituição.

Art. 2º Os órgãos e as entidades do Poder Executivo federal assegurarão, às pessoas naturais e jurídicas, o direito de acesso à informação, que será proporcionado mediante procedimentos objetivos e ágeis, de forma transparente, clara e em linguagem de fácil compreensão, observados os princípios da administração pública e as diretrizes previstas na Lei nº 12.527, de 2011.

France Bans Judge Analytics, 5 Years In Prison For Rule Breakers

© 4th June 2019  artificiallawyer  Litigation Prediction  17

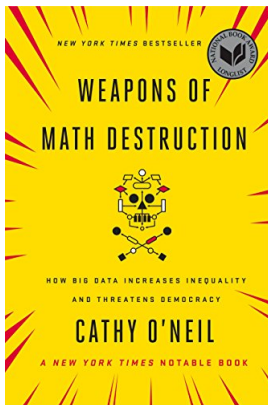


Privacidade de dados PESSOAIS

- Dwork (2006) - Differential Privacy

Privacidade de dados PESSOAIS

- Dwork (2006) - Differential Privacy
- O'Neil (2016) - Weapons of Math Destruction



Privacidade de dados PESSOAIS

- Lei 13.709, de 14 de agosto de 2018



Presidência da República
Secretaria-Geral
Subchefia para Assuntos Jurídicos

LEI Nº 13.709, DE 14 DE AGOSTO DE 2018

[Texto compilado](#)

[Mensagem de veto](#)

[Vigência](#)

LGPDP

~~Dispõe sobre a proteção de dados pessoais e altera a Lei nº 12.965, de 23 de abril de 2014 (Marco Civil da Internet):~~

Lei Geral de Proteção de Dados Pessoais (LGPD).
~~(Redação dada pela Lei nº 13.853, de 2019)~~ [Vigência](#)

O PRESIDENTE DA REPÚBLICA Faço saber que o Congresso Nacional decreta e eu sanciono a seguinte Lei:

CAPÍTULO I
DISPOSIÇÕES PRELIMINARES

Art. 1º Esta Lei dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural.

Parágrafo único. As normas gerais contidas nesta Lei são de interesse nacional e devem ser observadas pela União, Estados, Distrito Federal e Municípios. [\(Incluído pela Lei nº 13.853, de 2019\)](#) [Vigência](#)

Para saber mais

- ① Newcomb (1881) - Note on the Frequency of Use of the Different Digits in Natural Numbers
- ② De Finetti (1930) - Funzione Caratteristica di un Fenomeno Aleatorio
- ③ Benford (1938) - The Law of Anomalous Numbers
- ④ Kasner and Newman(1940) - Mathematics and the Imagination
- ⑤ Aitchison & Dunsmore (1975) - Statistical Prediction Analysis
- ⑥ Box (1979) - Robustness in the Strategy of Scientific Model Building
- ⑦ Ghosh (1988) - Statistical Information and Likelihood - A collection of critical essays by Dr. D. Basu
- ⑧ Seymour Geisser (1993) - Predictive Inference - An Introduction
- ⑨ Breiman (2001) - Statistical Modeling: The Two Cultures
- ⑩ Zabala (2009) - Desempate Técnico

Para saber mais

- 11 Fewster (2009) - A Simple Explanation of Benford's Law
- 12 Zabala & Silveira (2014) - Jurimetria: Estatística Aplicada ao Direito
- 13 Clarke & Clarke (2018) - Predictive Statistics - Analysis and Inference Beyond Models
- 14 Hyndman & Athanasopoulos (2018) - Forecasting: Principles and Practice
- 15 Zabala & Silveira (2019) - Decades of Jurimetrics
- 16 Izbicki & Santos (2020) - Aprendizado de Máquina: uma abordagem estatística
- 17 Zabala (2020) - Vídeos: Ciência de dados em software livre
- 18 Zabala (2020) - Código: Ciência de dados em software livre
- 19 Azevedo et al (2021) - A Benford's Law based methodology for fraud detection in social welfare programs - Bolsa Familia analysis
- 20 Zabala (2023). voice: Tools for Voice Analysis, Speaker Recognition and Mood Inference