# Crash Course de amostragem

Alex Rodrigues do Nascimento
Igor Ferreira do Nascimento

LAMFO/UnB

26 de setembro de 2020

# Estrutura da Oficina

# Estrutura da Oficina

# *Etapas*

[Cochran, 2007]

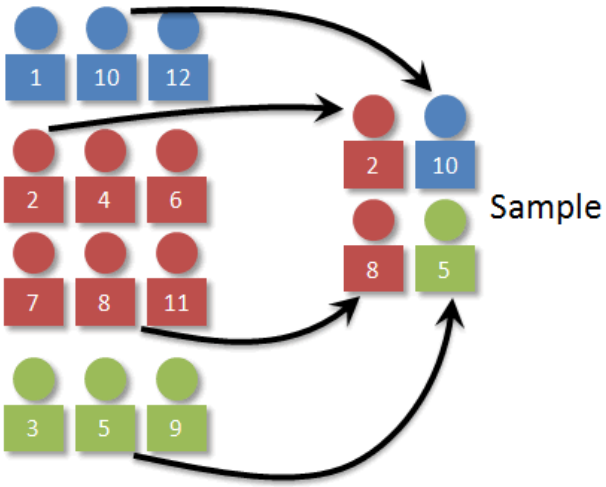
- Objectives of the Survey
- Population to be Sampled
- Data to be Collected
- Degree of Precision Desired
- Methods of Measurement (forma da coleta)
- The Frame (unidades)
- Selection of the Sample (plano amostral)
  https://faculty.elgin.edu/dkernler/statistics/ch01/1-4.html
- The Pretest
- Summary and Analysis of the Data

# *AAS*

[Cochran, 2007]

| | Population | Sample | standard errors |
|---|---|---|---|
| Total: | $Y = \sum^{N} y_i = y_1 + y_2 + \cdots + y_N$ | $\sum^{n} y_i = y_1 + y_2 + \cdots + y_n$ | $s_{\bar{y}} = \frac{s}{\sqrt{n}}\sqrt{1-f},$ |
| Mean: | $\bar{Y} = \frac{y_1 + y_2 + \cdots + y_N}{N} = \frac{\sum^{N} y_i}{N}$ | $\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum^{n} y_i}{n}$ | $s_{\hat{Y}} = \frac{Ns}{\sqrt{n}}\sqrt{1-f}$ |

# *Estratificada*




Fonte: https://faculty.elgin.edu/dkernler/statistics/ch01/1-4.html

# *Estratificada*

[Cochran, 2007]

$$N = N_1 + N_2 + \cdots + N_L.$$

$$\bar{y} = \frac{\sum_{h=1}^{L} n_h \bar{y}_h}{n}$$

stratification with *proportional* allocation of the $n_h$.

$$\frac{n_h}{n} = \frac{N_h}{N}$$

# *Sopa de letrinhas*

- Instituto Brasileiro de Geografia e Estatística - IBGE
- Sistema Integrado de Pesquisas Domiciliares por amostragem (SIPD)
- CNEFE, Cadastro Nacional de Endereços para Fins Estatísticos
- Pesquisa de Orçamentos Familiares (POF),
- Pesquisa de Orçamentos Familiares Simplificada (POFs)
- a Pesquisa Nacional de Saúde (PNS) e,
- PME (Pesquisa Mensal de Emprego)
- PNAD (Pesquisa Nacional por Amostra de Domicílios)
- Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua).

# *Amostra Mestra*

## Multistage Sampling

Often one technique isn't possible, so many professional polling agencies use a technique called **multistage sampling**. The strategy is relatively self-explanatory - two or more sampling techniques are used.

For example, consider the light-bulb example we looked at earlier with cluster sampling. Let's suppose that the bulb come off the assembly line in boxes that each contain 20 packages of four bulbs each. One strategy would be to do the sample in two stages:

**Stage 1:** A quality control engineer removes every 200th box coming off the line. (The plant produces 5,000 boxes daily. (This is *systematic* sampling.)

**Stage 2:** From each box, the engineer then samples three packages to inspect. (This is an example of **cluster** sampling.)

The US Census also uses multistage sampling. If you haven't already (you should have!), read Section 1.4 in your te for more details.

Fonte: https://faculty.elgin.edu/dkernler/statistics/ch01/1-4.html

# *Amostra Mestra*

- Estratos: Unidades da Federação (UFs)
  1. Capital;
  2. Demais municípios pertencentes à RM ou à RIDE;
  3. colar ou expansão metropolitana ou a outra RM;
  4. à RIDE com sede em outra UF e
  5. Demais municípios da UF.
- Unidades Primárias de Amostragem (UPAs): selecionadas com probabilidade proporcional ao tamanho, medido pelo número de domicílios particulares permanentes ocupados e vagos (DPPs).
- selecionado 14 de domicílios particulares permanentes ocupados dentro de cada UPA da amostra, por amostragem aleatória simples (CNEFE)

# *Estimador total - PNAD*

O estimador para o total é de2014amostra:

$$\hat{Y}_r^{tri} = \sum_g \sum_i \sum_j \sum_k w_{gij}^{**} \times y_{gijk} \tag{1}$$

$y_{gijk}$ é o valor da variável de interesse $y$ para a pessoa $k$ do domicílio $j$ da UPA $i$ do grupo de rotação $g$

Peso ajustado por calibração (totais populacionais):

$$w_{gij}^{**} = \frac{1}{m_g} \times \frac{N_g}{N_{gi}} \times \frac{N_g^*}{n_{gi}} \times \frac{n_{gi}^*}{n_{gi}^{**}} \times \frac{P_a^{tri}}{\hat{P}_a^{tri}} \tag{2}$$

$P_a^{tri}$ é a estimativa populacional produzida pela COPIS para o nível geográfico a para o dia 15 do mês do meio do trimestre $\hat{P}_a^{tri}$ é a estimativa

populacional obtida com os dados da pesquisa para o nível geográfico ano trimestre

# *Variância do estimador - PNAD*

Considerando a aproximação [de Freitas and de Abreu Antonaci, 2014]:

$$V(\hat{\theta}) = \sum_h \frac{m_h}{m_h - 1} \sum_i (\hat{Z}_{hi} - \bar{Z}_h)^2 \tag{3}$$

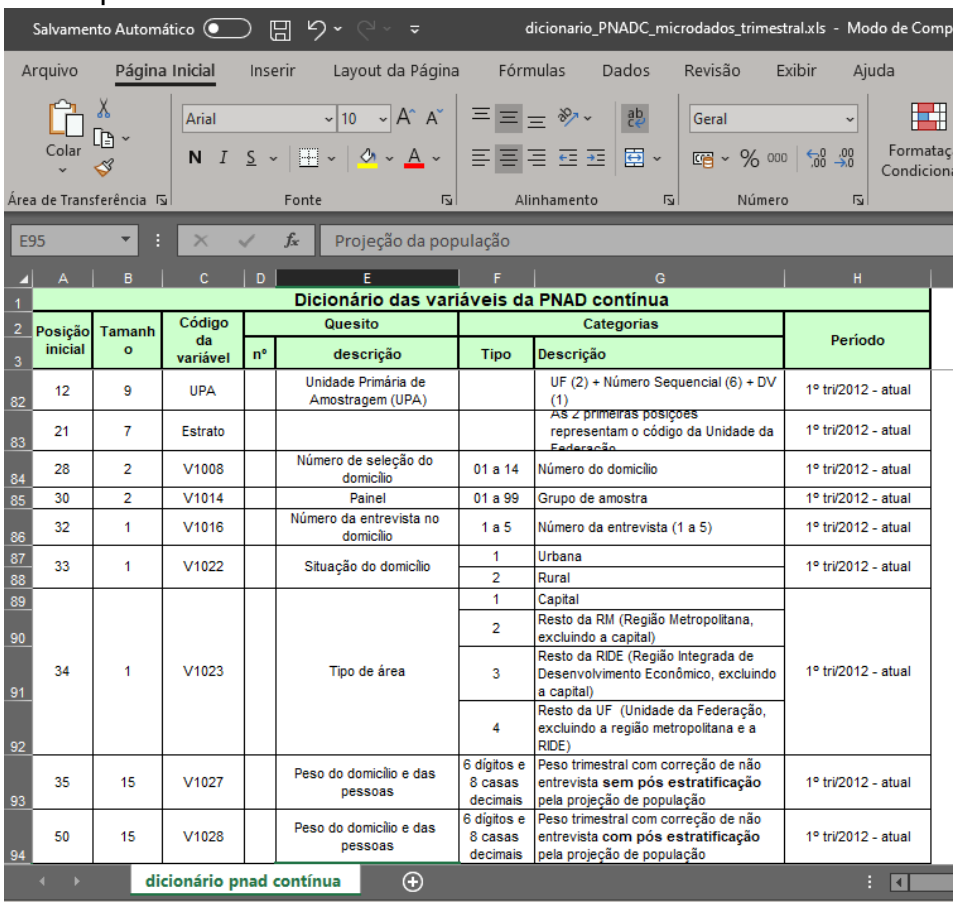
Onde

$$\hat{Z}_{hi} = \sum_j \sum_k w^{**}_{gij} \times z_{gijk} \tag{4}$$

$$\bar{Z}_h = \frac{1}{m_h} \sum_i \hat{Z}_{hi} \tag{5}$$

# *Aplicação*

RMD para PNAD contínua



Salvamento Automático — dicionario_PNADC_microdados_trimestral.xls - Modo de Comp

Arquivo | Página Inicial | Inserir | Layout da Página | Fórmulas | Dados | Revisão | Exibir | Ajuda

Área de Transferência | Fonte | Alinhamento | Número

E95 — Projeção da população

**Dicionário das variáveis da PNAD contínua**

| Posição inicial | Tamanho | Código da variável | Quesito nº | Quesito descrição | Categorias Tipo | Categorias Descrição | Período |
|---|---|---|---|---|---|---|---|
| 12 | 9 | UPA | | Unidade Primária de Amostragem (UPA) | | UF (2) + Número Sequencial (6) + DV (1) | 1º tri/2012 - atual |
| 21 | 7 | Estrato | | | | As 2 primeiras posições representam o código da Unidade da Federação | 1º tri/2012 - atual |
| 28 | 2 | V1008 | | Número de seleção do domicílio | 01 a 14 | Número do domicílio | 1º tri/2012 - atual |
| 30 | 2 | V1014 | | Painel | 01 a 99 | Grupo de amostra | 1º tri/2012 - atual |
| 32 | 1 | V1016 | | Número da entrevista no domicílio | 1 a 5 | Número da entrevista (1 a 5) | 1º tri/2012 - atual |
| 33 | 1 | V1022 | | Situação do domicílio | 1 | Urbana | 1º tri/2012 - atual |
| | | | | | 2 | Rural | |
| 34 | 1 | V1023 | | Tipo de área | 1 | Capital | 1º tri/2012 - atual |
| | | | | | 2 | Resto da RM (Região Metropolitana, excluindo a capital) | |
| | | | | | 3 | Resto da RIDE (Região Integrada de Desenvolvimento Econômico, excluindo a capital) | |
| | | | | | 4 | Resto da UF (Unidade da Federação, excluindo a região metropolitana e a RIDE) | |
| 35 | 15 | V1027 | | Peso do domicílio e das pessoas | 6 dígitos e 8 casas decimais | Peso trimestral com correção de não entrevista **sem pós estratificação** pela projeção de população | 1º tri/2012 - atual |
| 50 | 15 | V1028 | | Peso do domicílio e das pessoas | 6 dígitos e 8 casas decimais | Peso trimestral com correção de não entrevista **com pós estratificação** pela projeção de população | 1º tri/2012 - atual |

dicionário pnad contínua

# Estrutura da Oficina

# *Bootstrap*

- O *bootstrap* foi desenvolvido por Efron [Efron, 1979] e ajuda a aprender sobre as características da amostra pela obtenção de reamostragem (reamostragem da amostra original com reposição) e utiliza-se essa informação para inferir sobre à população [CASELLA and BERGER, 2011].
- Método utilizado, em geral, para estimar o erro padrão de estimadores. Suponha que tenhamos uma amostra $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ e uma estimativa $\hat{\theta}(x_1, x_2, \ldots, x_n) = \hat{\theta}$, ao selecionarmos $B$ reamostras (ou amostras *bootstrap*) podemos calcular

$$Var^*_B(\hat{\theta}) = \frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}^*_i - \overline{\hat{\theta}^*})^2.$$

# *Bootstrap*

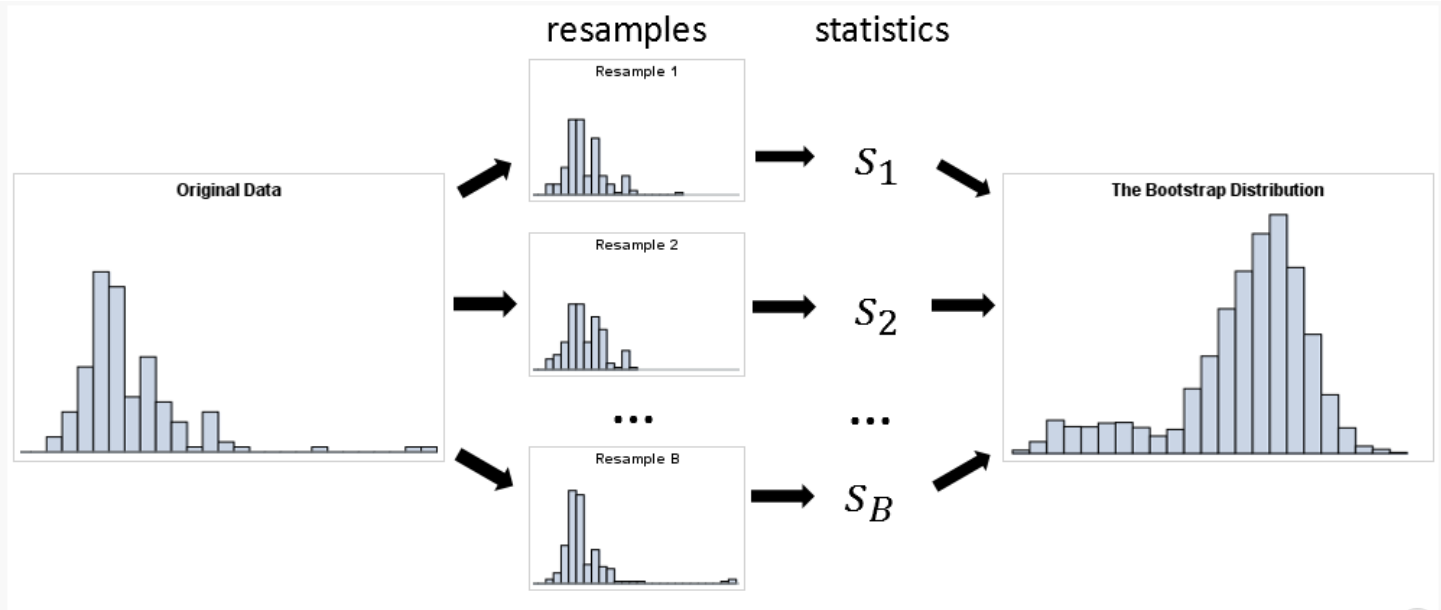



Figura: Análise Bootstrap: (1) Iniciar com amostra (2)Reamostrar da amostra original B vezes (Não-paramétrico) (3) Computar estatística para cada amostra (4)Utilizar distribuição *bootstrap* para fazer inferência.
Fonte: [Wicklin, 2018]

- Sua fácil aplicabilidade torna o método uma poderosa ferramenta, com possibilidade de ser utilizado em muitas técnicas de *machine learning* (incluindo algumas as quais uma medida de variabilidade é difícil de se obter).

# *Bootstrap*

Apesar do conceito ser intuitivo e de fácil aplicabilidade, a teoria por trás do *bootstrap* é bastante sofisticada, tendo como base as expansões de *Edgeworth*, expansões de funções de distribuição acumulada em torno de uma distribuição normal. A teoria recebe abordagem completa em [Hall, 2013].

Em muitos casos, *bootstrap* oferece um estimador razoável, que é consistente.

$$Var^*_B(\hat{\theta}) \xrightarrow{B\to\infty} Var^*(\hat{\theta})$$

$$Var^*(\hat{\theta}) \xrightarrow{n\to\infty} Var(\hat{\theta})$$

no qual,

$Var^*(\hat{\theta}) = \frac{1}{n^n-1}\sum_{i=1}^{n^n}(\hat{\theta}^*_i - \overline{\hat{\theta}^*})^2$.

# Qualidade do ajuste

O objetivo de muitas aplicações de *machine learning* (ML) é fazer previsões para observações não vistas anteriormente.

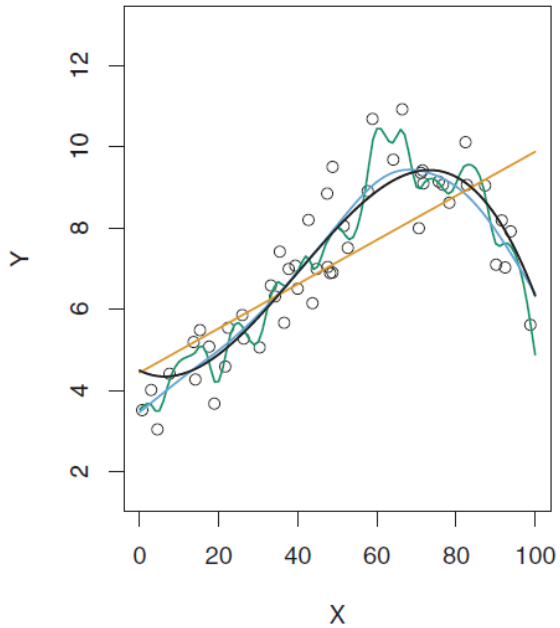



Figura: *Overfitting.*
Fonte: [James et al., 2013]

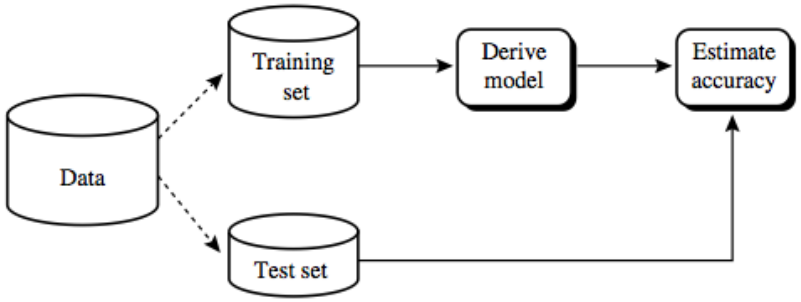



Figura: Amostra de treino *versus* Amostra de Teste.

# *Cross Validation*

A estimativa do erro de teste pode apresentar alta variabilidade porque é dependente da amostragem realizada no conjuntos de dados. As técnicas de *Cross Validation* refinam o conceito de amostra de teste e treino.

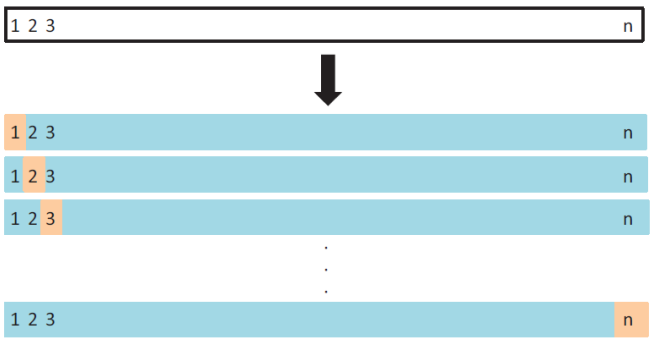



Figura: *Leave-one-out cross-validation.*
Fonte: [James et al., 2013]

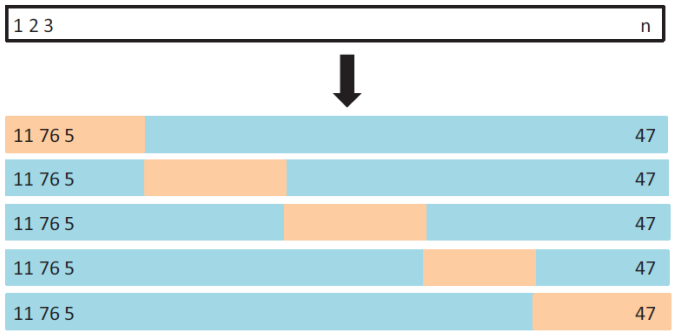



Figura: *K-fold cross-validation.*
Fonte: [James et al., 2013]

# Árvore de Decisão

Técnica de ML que consiste em dividir o espaço preditor em regiões ($X|X_j > s$ e $X|X_j < s$) que fornecem a maior redução de alguma medida de erro. O método tende a fornecer boa interpretabilidade mas baixo poder de previsão.

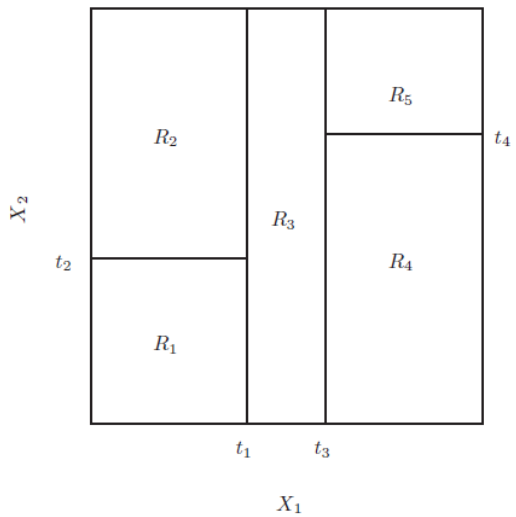



Figura: *Divisão de um espaço preditor bidimensional.*
Fonte: [James et al., 2013]

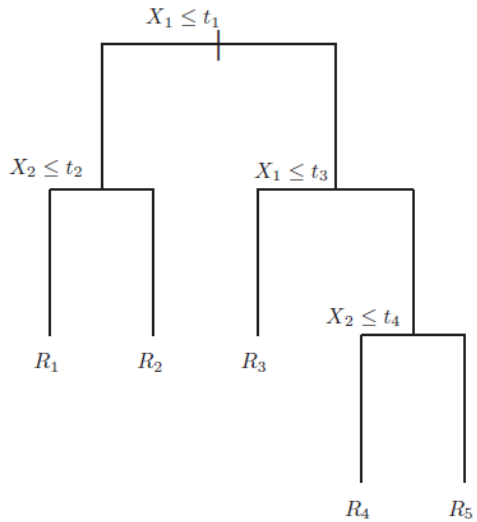



Figura: *Exemplo de árvore de decisão.*
Fonte: [James et al., 2013]

# *Bagging*

Procedimento de propósito geral para reduzir a variância de um método de ML. Utilizando conceitos de inferência estatística (dado $x_1, x_2, ..., x_n$ $iid$ então $\overline{x} \sim N(\mu, \frac{\sigma^2}{\mathbf{n}})$), aplica-se o método de ML em B amostras *Bootstrap* e utiliza-se a média dos resultados como previsão.

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}^b(x).$$

Aplicado em árvore de decisão fornece melhorias na precisão ao combinar centenas ou mesmo milhares de árvores em um único procedimento.

# Floresta Aleatória

O método de *Bagging* tende a utilizar as mesmas variáveis para construir as árvores de decisão em cada amostra *Bootstrap*, gerando árvores correlacionadas o que impacta a redução de variância. As florestas aleatórias (FA) superam esse problema, forçando cada divisão a considerar apenas um subconjunto dos preditores.

$$\hat{f}^*(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x).$$

Em geral utiliza-se um grupo de $m = \sqrt{p}$ variáveis para construir cada árvore de decisão.

# Floresta Aleatória em *Big Data*

- A redução de correlação entre as árvores de decisão é a motivação para construção de FA. Entretanto o desempenho do método é dependente da performance de cada árvore de decisão, ou seja, se muitas árvores apresentarem baixo poder de previsão, o método irá apresentar estimativas ruins (ou classificações ruins).
- No contexto de *Big Data*, uma grande proporção de preditores não são informativos para a análise e a utilização de amostragem aleatória simples tende a gerar árvores de decisão com baixo poder de previsão.
- [Ye et al., 2013] apresentam uma metodologia que utiliza amostragem **estratificada** para construção da floresta aleatória.

# Floresta Aleatória com Amostra Estratificada

- A ideia dos autores se baseia em dividir as $p$ variáveis em dois grupos: muito informativo e pouco informativo. Em seguida, selecionamos aleatoriamente variáveis de cada grupo, garantindo que temos características representativas de ambos. Esta abordagem garante que cada subespaço contenha informação suficiente para a finalidade.
- Considera-se uma função não negativa $\phi$ que capture o grau de explicação da variável $p_i$ com respeito a variável resposta $Y$. Normaliza-se $\phi_i$:

$$\theta_i = \frac{\phi_i}{\sum_{k=1}^{N} \phi_k}$$

- As variáveis preditoras são ordenadas segundo $\theta_i$ e determina-se um limiar $\alpha$ que irá dividir as variáveis em dois grupos. O algoritmo proposto constrói um modelo de floresta aleatório considerando os estratos para produzir cada árvore de decisão.

# Estrutura da Oficina

# Bibliography I


CASELLA, G. and BERGER, R. L. (2011).
Inferência estatística-tradução da 2ª edição norte-americana.
*Centage Learning*.
15

Cochran, W. G. (2007).
*Sampling techniques*.
John Wiley & Sons.
4, 5, 7

de Freitas, M. P. S. and de Abreu Antonaci, G. (2014).
Amostra mestra 2010 e amostra da pnad contínua.
12

# Bibliography II


Efron, B. (1979).
The 1977 rietz lecture.
*The annals of Statistics*, 7(1):1–26.
15

Hall, P. (2013).
*The bootstrap and Edgeworth expansion*.
Springer Science & Business Media.
17

# Bibliography III


James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013).
*An introduction to statistical learning*, volume 112.
Springer.
18, 19, 20


Wicklin, R. (2018).
*The essential guide to bootstrapping in SAS*.
16

# Bibliography IV


Ye, Y., Wu, Q., Huang, J. Z., Ng, M. K., and Li, X. (2013). Stratified sampling for feature subspace selection in random forests for high dimensional data. *Pattern Recognition*, 46(3):769–787. 23