CENTRO UNIVERSITÁRIO

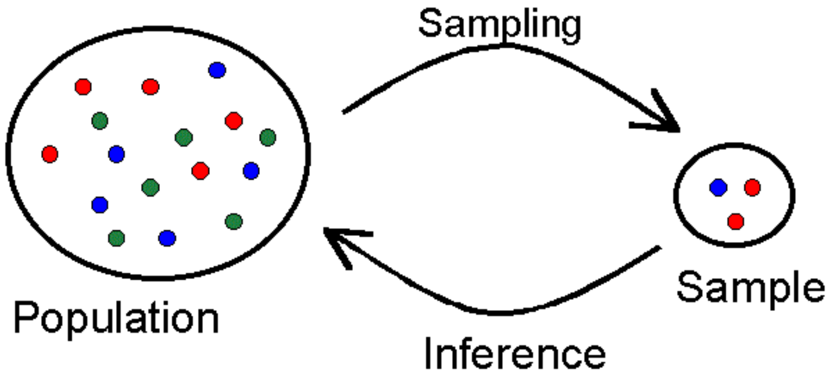
Análise de Dados Aplicada à Computação

DISTRIBUIÇÃO DE DADOS E AMOSTRAS

Prof. M.Sc. Howard Roatti

Distribuição de Dados e Amostras

- O vasto conjunto de dados disponível reforça a necessidade da amostragem para um trabalho eficiente e minimização de viés
- A amostragem é utilizada para testes diversos como: precificação e web treatments
- Geralmente a Ciência de Dados se preocupa com os procedimentos de amostragem e nos dados que possui, independente de ser uma população ou uma amostra



Amostragem Aleatória e Viés de Amostra

- **Amostra:** um subgrupo de uma população
- **População:** grupo de dados maior
- **N(n):** o tamanho da população e da amostra
- **Amostragem Aleatória:** seleção de elementos aleatórios para formação de uma amostra
- **Amostragem Estratificada:** divide a população em partes, então realiza uma amostragem aleatória em cada parte
- **Amostra Aleatória Simples:** amostra aleatória sem estratificar
- **Viés de Amostragem:** uma amostra que não representa a população

Amostragem Aleatória

- Trata de um processo onde cada membro da população possui as mesmas chances de ser selecionado
- O resultado é a amostra aleatória simples
- É possível realizar amostragem com reposição, onde as observações selecionadas são devolvidas à população
- E sem reposição, onde uma vez selecionada, não será possível selecionar novamente

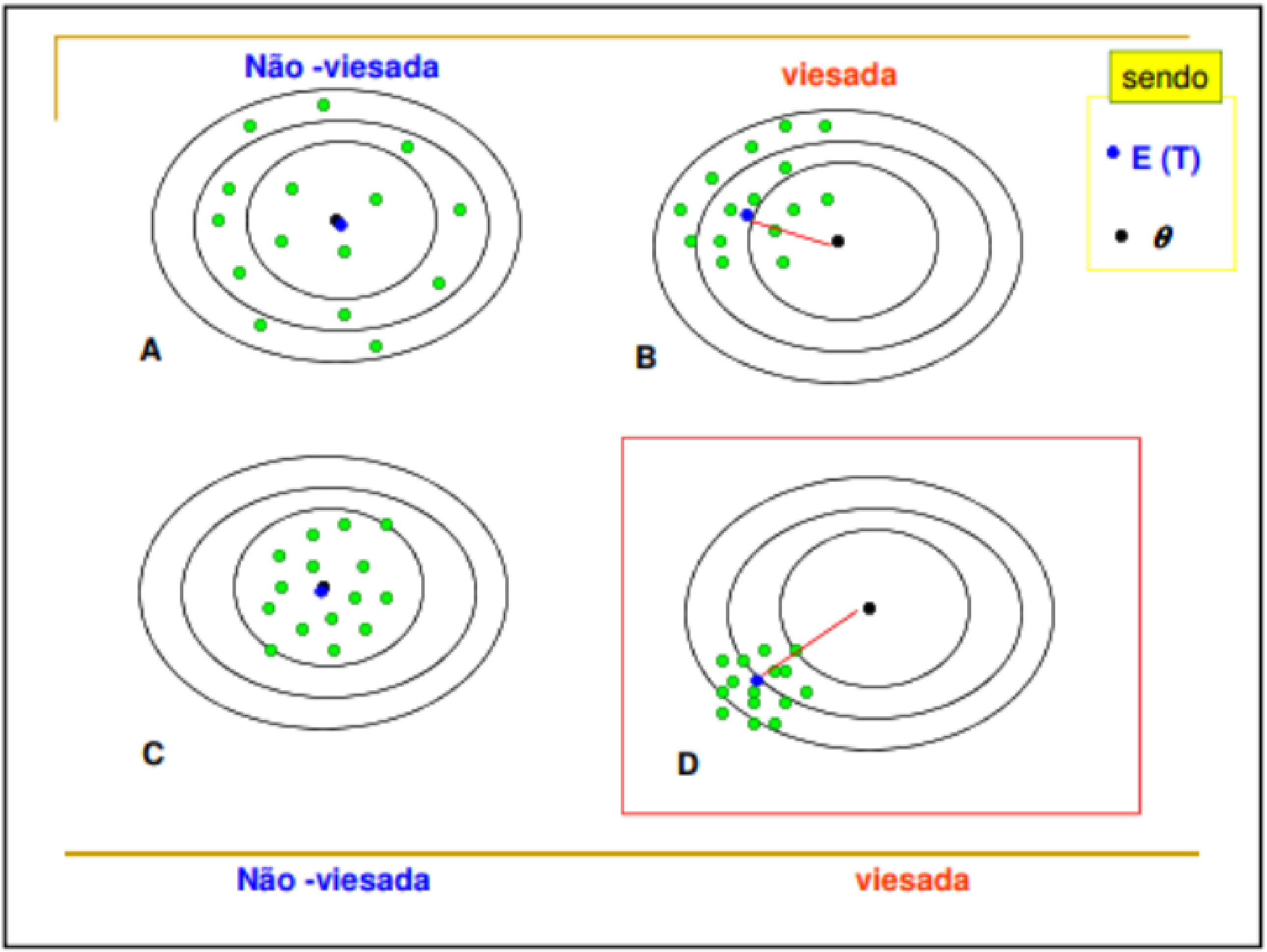
Amostras de Autosseleção

- Avaliações em mídias sociais como Google Maps estão propensas a serem viciadas
- Pessoas que avaliam não são selecionados de forma aleatória, tomam a iniciativa
- **Motivação**: alguma experiência negativa ou ter relacionamento com o estabelecimento
- Apesar de não ser confiável como indicador do estabelecimento, pode ser eficaz para comparar estabelecimentos similares, já que o vício de autosseleção pode acontecer nos dois

Viés Estatístico

- Viés estatístico trata de **erros de medição** ou **amostragem produzidas sistematicamente**
- O viés pode acontecer de diversas formas e pode ou não ser observado, é necessário atenção para identificação em uma análise
- O viés ocorre quando há uma má interpretação ou a falta de alguma variável importante no conjunto de dados

Viés Estatístico



Viés de um estimador: No caso A, um estimador não-viesado e com baixa precisão. No caso B, um estimador viesado e com baixa precisão. No caso C, um estimador não-viesado e com alta precisão. No caso D, um exemplo de um estimador viesado e com alta precisão.

Amostragem Aleatória

- Mesmo na era do Big Data, a seleção aleatória é importante
- O viés ocorre quando as observações não representam a população
- Amostras aleatórias podem reduzir o viés, ou seja, a qualidade dos dados é mais importante que a quantidade
- Para uma melhor seleção de amostras aleatórias que represente a população, a estratificação ajuda na seleção aleatória de qualidade

Viés de Seleção

- “Se você não sabe o que está procurando, procure bastante e vai encontrar.”, Yogi Berra
- O viés de seleção ocorre quando na escolha dos dados é seletiva levando a uma conclusão enganosa, mesmo que seja uma escolha inconsciente
- Se uma hipótese é especificada, conduzida a um experimento e testada, pode ser concluída com alta confiança
- Isso não ocorre com frequência, as pessoas olham os dados e tentam encontrar padrões que muitas vezes é fruto de insistência
- “Se você torturar os dados o bastante, cedo ou tarde eles vão confessar.”

Distribuição Amostragem

- **Estatística de Amostra:** métrica calculada para uma amostra
- **Teorema do Limite Central:** Conforme uma amostra cresce, a tendência da frequência amostra é ter uma forma normal
- **Erro Padrão:** A variabilidade de uma métrica em várias amostras

Distribuição Amostragem

- As amostras são extraídas com o objetivo de mediar ou modelar algo estatístico ou para aprendizagem de máquina, por exemplo
- Por se basear em uma amostra, existe a chance de conter erro. E outra amostra pode obter um resultado diferente
- Distribuições de estatísticas como a média tendem a ser mais regulares e no formato de sino do que o dado em si

Distribuição Amostragem

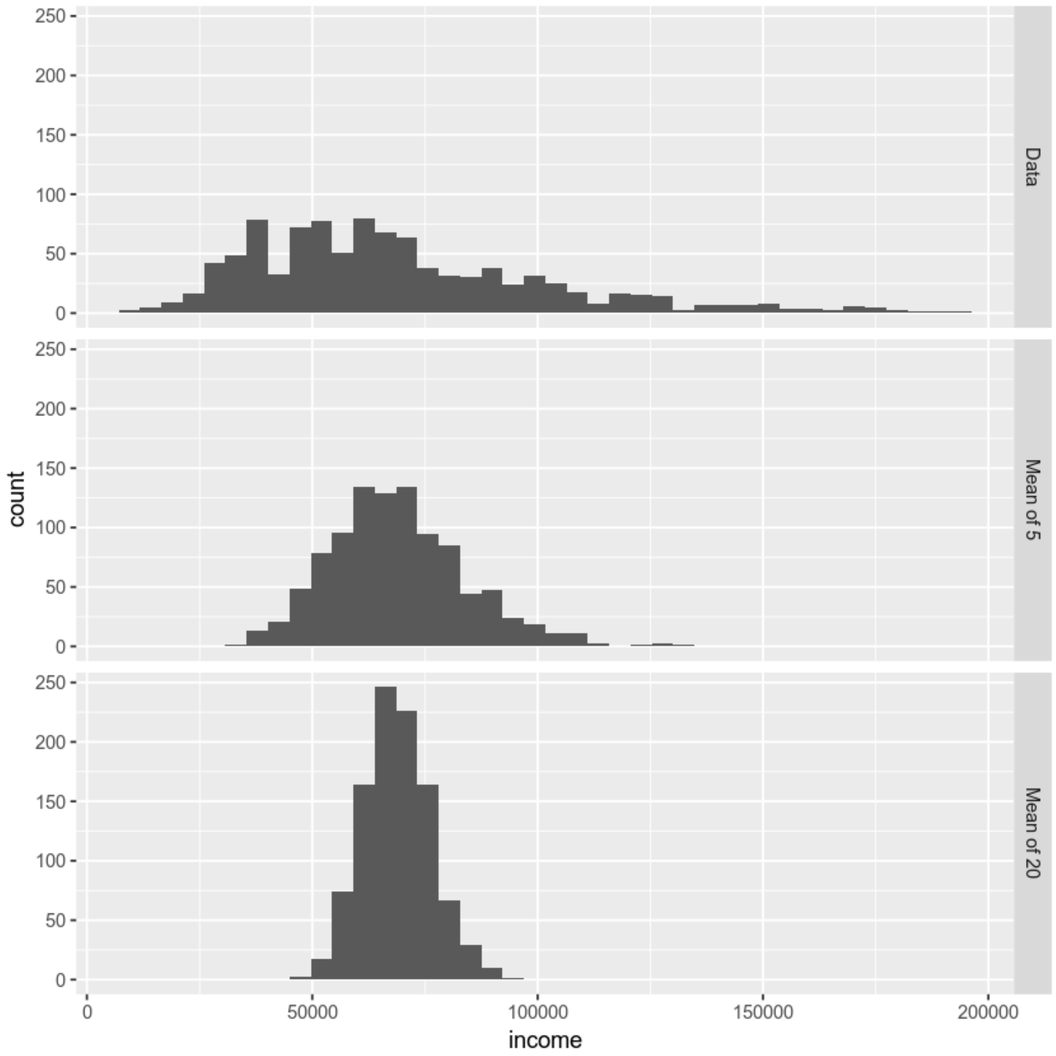
```
library(ggplot2)
loans_income <- read.csv(file.path('.\\',
'DataSets','loans_income.csv')
                         )[,1]
# take a simple random sample
samp_data <- data.frame(income=sample(loans_income, 1000),
                        type='data_dist')
# take a sample of means of 5 values
samp_mean_05 <- data.frame(
                income = tapply(sample(loans_income, 1000*5),
                rep(1:1000, rep(5, 1000)), FUN=mean),
                type = 'mean_of_5')
# take a sample of means of 20 values
samp_mean_20 <- data.frame(
                income = tapply(sample(loans_income, 1000*20),
                rep(1:1000, rep(20, 1000)), FUN=mean),
                type = 'mean_of_20')
```

Distribuição Amostragem

```
# bind the data.frames and convert type to a factor
income <- rbind(samp_data, samp_mean_05, samp_mean_20)
income$type <- factor(income$type,
                      levels=c('data_dist', 'mean_of_5', 'mean_of_20'),
                       labels=c('Data', 'Mean of 5', 'Mean of 20'))

ggplot(income, aes(x=income)) +
geom_histogram(bins=40) +
facet_grid(type ~ .)
```

Distribuição Amostragem



Teorema de Limite Central

- As médias tiradas de muitas amostras lembram a **curva normal**, mesmo que a população não seja normalmente distribuída
- Permite aproximação de fórmulas como **distribuição t**
- Está por traz de **testes de hipótese** e **intervalos de confiança**

Erro-Padrão

- Informa a precisão da média de qualquer amostra em comparação com a média real
- Se o erro-padrão aumenta, significa que a amostra extraída está distante de uma representação fiel da população
- Ela resume a variabilidade na distribuição amostral
- Pode ser estimada baseando-se no desvio padrão e no tamanho da amostra

$$\text{Standard Error} = SE = \frac{s}{\sqrt{(n)}}$$

Distribuição Amostragem

- A frequência de distribuição de uma métrica nos mostra como ela se comporta de amostra para amostra
- A distribuição amostral pode ser calculada pelo bootstrap ou por formulas que dependem do teorema de limite central
- Uma métrica para variabilidade entre amostras é o erro padrão

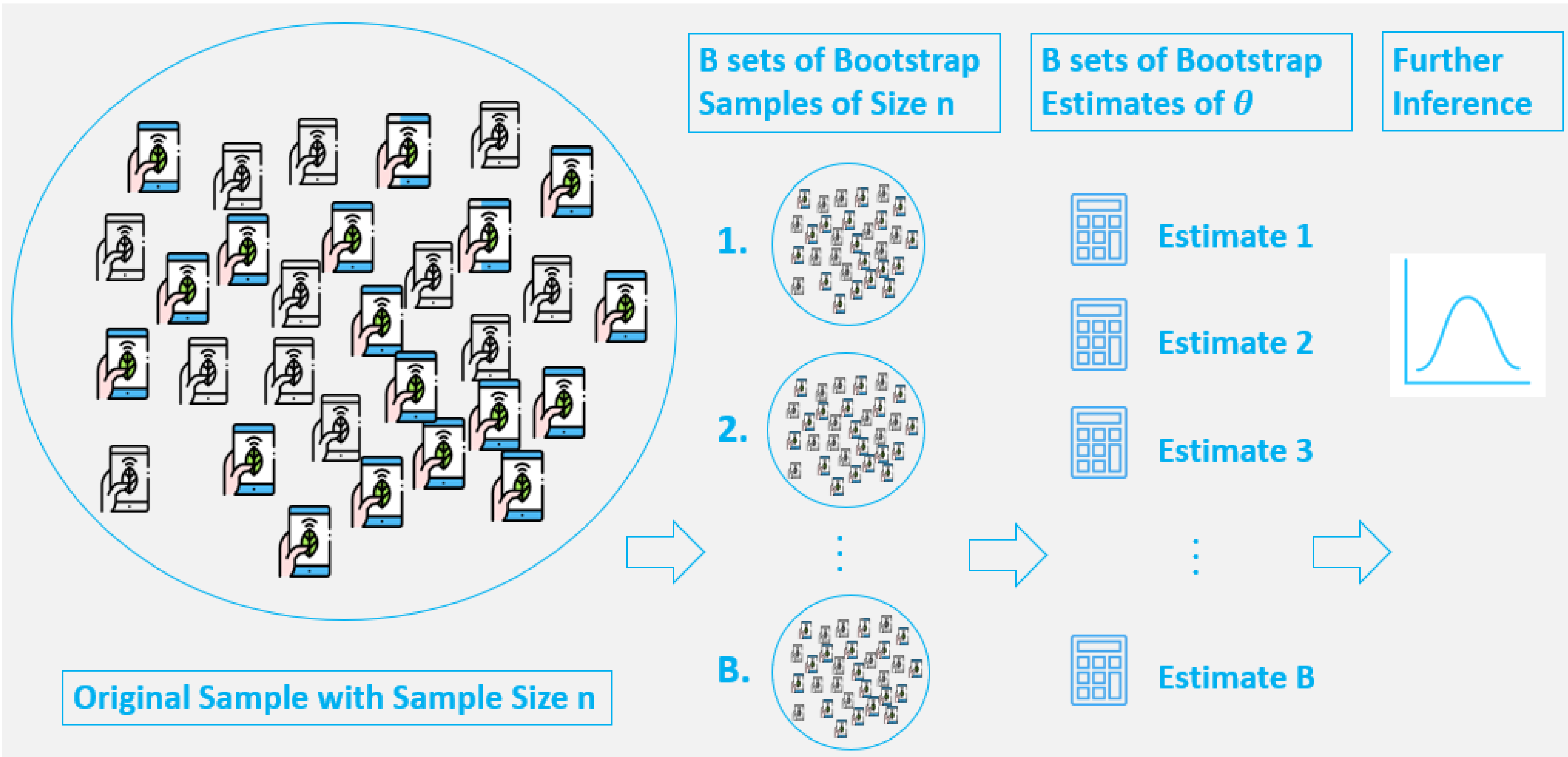
Bootstrap

- Uma maneira fácil e eficaz de estimar estatísticas de distribuição amostral é a extração das amostras, com reposição, então recalcular a estatística ou modelo para cada nova amostra
- Não é necessário que os dados ou estatísticas sigam a distribuição normal
- **Amostra Bootstrap:** uma amostra extraída com reposição dos dados observados
- **Reamostragem:** é o processo de repetidas extrações amostrais, podendo incluir bootstrap ou permutação

Bootstrap

- Conceitualmente imagine que o bootstrap replica a amostra original milhares ou milhões de vezes para que tenha uma população hipotética
- Na prática não precisamos replicar as amostras, basta retirar com reposição
- Assim, a probabilidade de um elemento ser retirado permanece sempre igual

Bootstrap


B sets of Bootstrap Samples of Size n
B sets of Bootstrap Estimates of θ
Further Inference
1.
2.
B.
Estimate 1
Estimate 2
Estimate 3
Estimate B
Original Sample with Sample Size n

Bootstrap

```
library(boot)
stat_fun <- function(x, idx) median(x[idx])
boot_obj <- boot(loans_income, R=1000, statistic=stat_fun)

boot_obj
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = loans_income, statistic = stat_fun, R = 1000)


Bootstrap Statistics :
    original  bias    std. error
t1*    62000 -74.017    219.9868
```

Estimativa Original: $62.000 | Viés: -81.90 | Erro-Padrão: $229.86

Bootstrap

- É possível usar para dados multivariados, onde as linhas são amostradas como unidades
- Então um modelo pode ser aplicado para estimar a estabilidade ou aumentar a capacidade preditiva

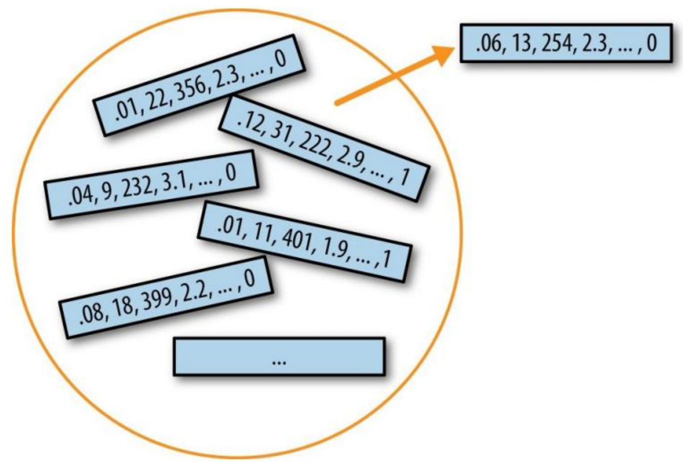


Figure 2-8. Multivariate bootstrap sampling

- As árvores de decisão pode ser aplicadas nas amostras, então calcula-se a média de suas previsões trazendo resultados melhores que uma única árvore

Esse processo é chamado de *bagging*, abreviação de *bootstrap aggregating*

Bootstrap

- O bootstrap é uma poderosa ferramenta para verificar a variabilidade de uma estatística
- Permite estimar estatísticas que não possuem aproximação matemática
- Quando aplicado a modelos preditivos, é mais eficiente que um modelo único

Intervalos de Confiança

- Tabelas de frequência, histogramas, boxplots e erros-padrão são formas de entender o potencial erro de uma estimativa de amostra.
- Intervalos de confiança são outro.
- **Nível de Confiança:** A porcentagem de confiança de uma estatística de interesse
- **Extremidades de Intervalo:** o topo e a base do intervalo de confiança

Intervalos de Confiança

- Os intervalos sempre vem com a probabilidade de cobertura
- Um intervalo de 90% indica que uma pesquisa é 90% confiável e que haverá um risco baixo de coincidir com o resultado obtido pela população
- Em se tratando do Bootstrap, significa que um intevalor de x% em torno de uma estimativa amostral, deveria na média conter estimativas amostrais de x% do tempo

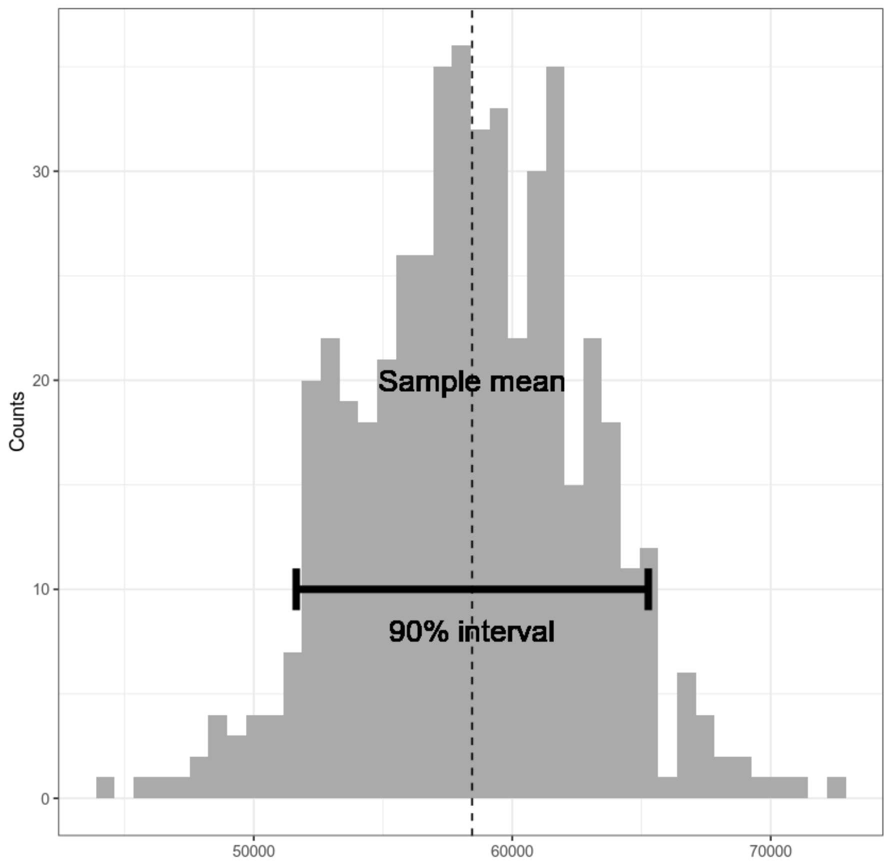
Intervalos de Confiança

```
set.seed(5)
set.seed(7)
sample20 <- sample(loans_income, 20)
sampleMean <- mean(sample20)
stat_fun <- function(x, idx) mean(x[idx])
boot_obj <- boot(sample20, R=500, statistic=stat_fun)
boot_ci <- boot.ci(boot_obj, conf=0.9, type='basic')
X <- data.frame(mean=boot_obj$t)
ci90 <- boot_ci$basic[4:5]
ci <- data.frame(ci=ci90, y=c(9, 11))
ci
ggplot(X, aes(x=mean)) +
    geom_histogram(bins=40, fill='#AAAAAA') +
    geom_vline(xintercept=sampleMean, linetype=2) +
    geom_path(aes(x=ci, y=10), data=ci, size=2) +
    geom_path(aes(x=ci90[1], y=y), data=ci, size=2) +
    geom_path(aes(x=ci90[2], y=y), data=ci, size=2) +
    geom_text(aes(x=sampleMean, y=20, label='Sample mean'), size=6) +
    geom_text(aes(x=sampleMean, y=8, label='90% interval'), size=6) +
    theme_bw() +
    labs(x='', y='Counts')
```

Intervalos de Confiança

ci	y
51643.09	9
65262.95	11
Média	57573.0

O Bootstrap é uma ferramenta geral que pode ser usada para gerar intervalos de confiança para a maioria das estatísticas ou parâmetros de modelo.

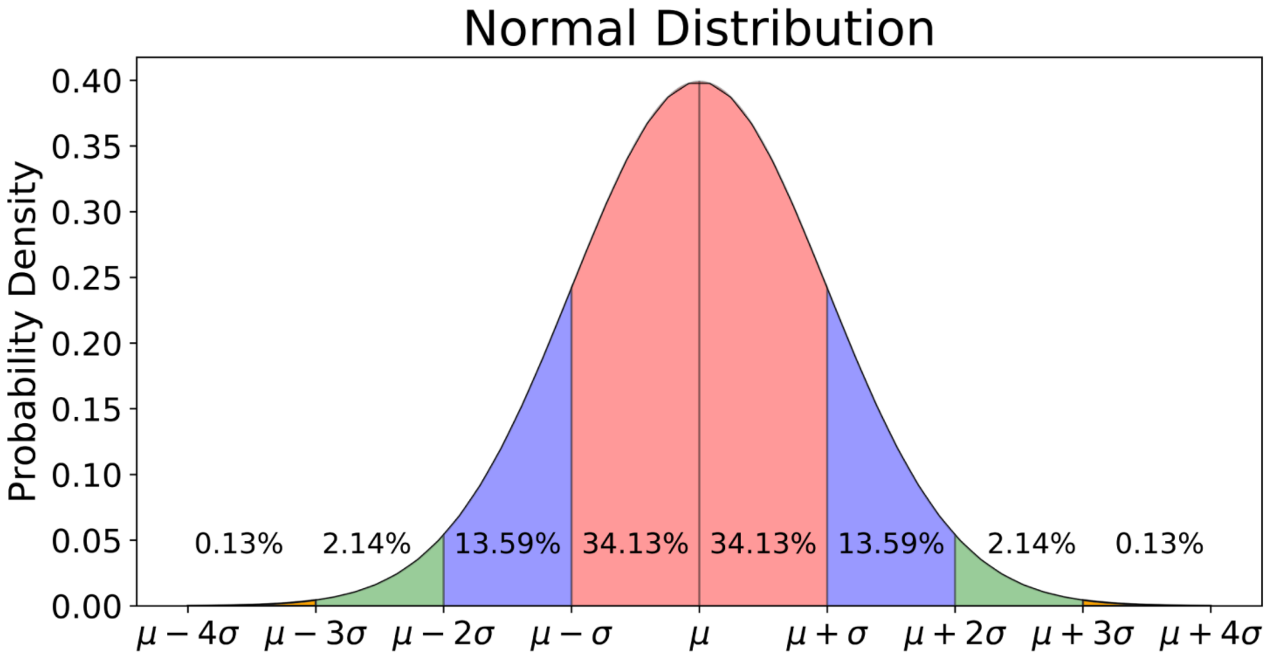


Intervalos de Confiança

- O que devemos nos perguntar é:
- Qual a probabilidade de um valor real estar dentro de um certo intervalo?
- A porcentagem associada ao intervalo é chamado de **nível de confiança**
- Quanto maior o nível, maior o intervalo
- Para um cientista de dados, é uma ferramenta para ter noção da variabilidade da amostra e verificar a necessidade de uma amostra maior

Distribuição Normal

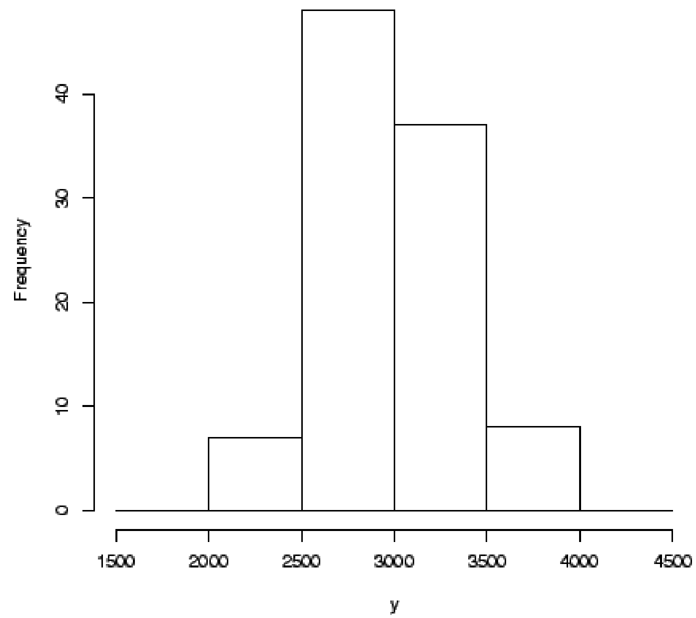
- É uma ferramenta de aproximações que tem como objetivo auxiliar no cálculo de probabilidade de um evento ocorrer



Distribuição Normal

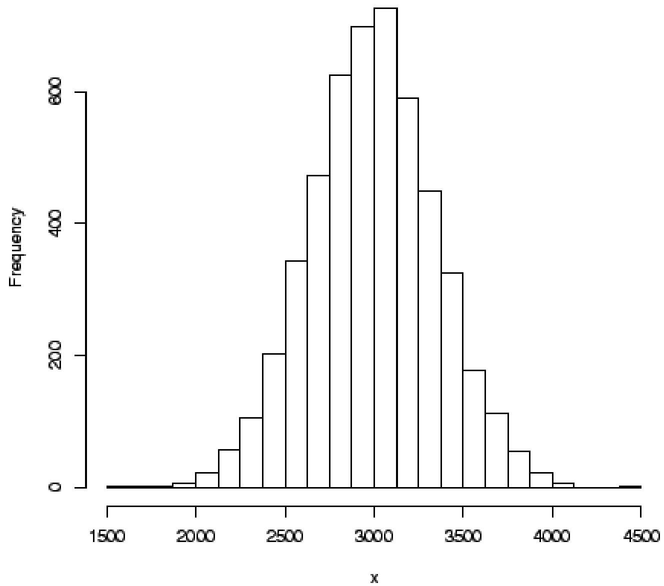
- É comum pensar que a maior parte dos dados segue a distribuição normal, geralmente não seguem
- Termos Importantes:
  - **Erro**: diferença entre um ponto de dado e um valor previsto
  - **Padronizar**: calculo para padronização do dado aproximando da curva normal, subtrai a média e divide pelo desvio padrão
  - **Escore-Z**: resultado obtido pela padronização
  - **Normal Padrão**: distribuição com média = 0 e desvio-padrão = 1
  - **Gráfico QQ**: permite visualizar o quão próximo uma distribuição amostral está de uma distribuição normal

Distribuição Normal



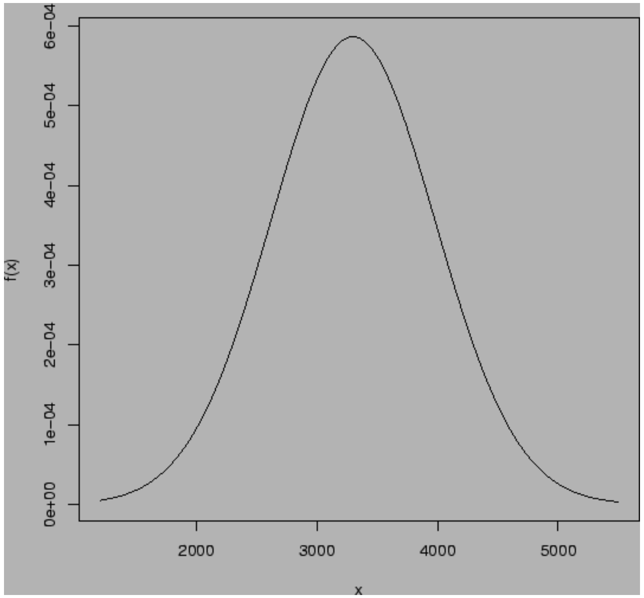
-
- **Exemplo:** O peso de recém-nascidos é uma variável aleatória contínua.
- Histograma de frequências relativas a 100 pesos de recém-nascidos com intervalo de 500 gramas

Distribuição Normal



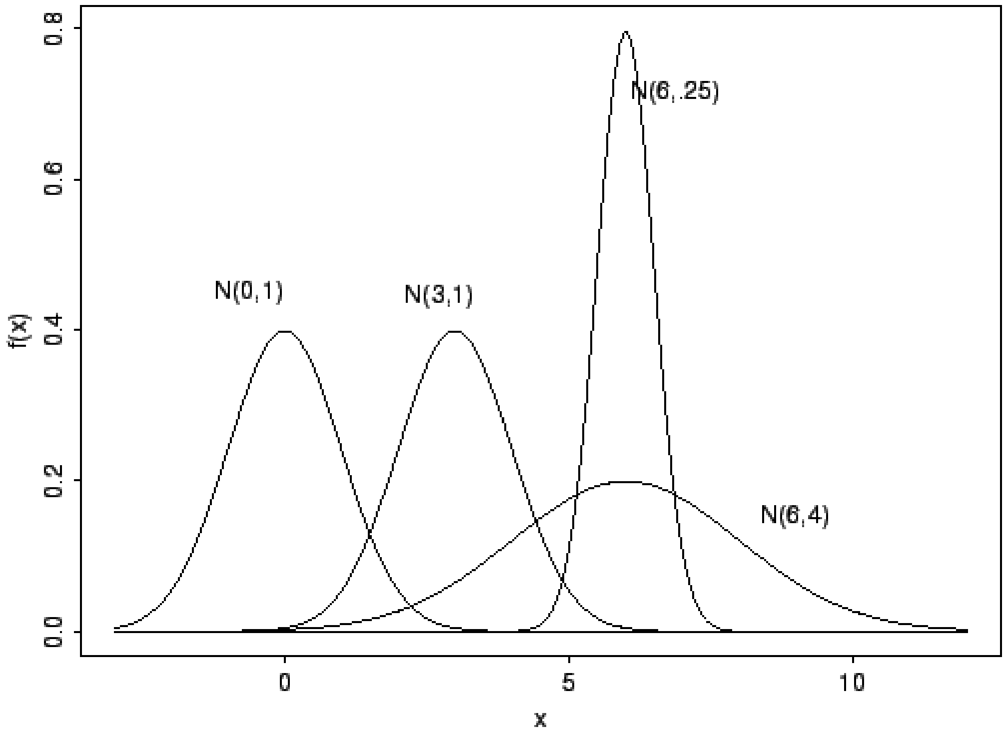
- **Exemplo:** O peso de recém-nascidos é uma variável aleatória contínua.
- Histograma de frequências relativas a 5000 pesos de recém-nascidos com intervalo de 125 gramas

Distribuição Normal



- **Exemplo:** O peso de recém-nascidos é uma variável aleatória contínua.
- Função de densidade de probabilidade para a variável aleatória contínua X=peso do recém-nascido (g)

Distribuição Normal



Para a distribuição Normal, a proporção de valores caindo dentro de um, dois, ou três desvios padrão da média são:

Range	Proportion
$\mu \pm 1\sigma$	68.3%
$\mu \pm 2\sigma$	95.5%
$\mu \pm 3\sigma$	99.7%

Distribuição Normal

- Isso posto podemos estimar algumas probabilidades. Supondo que a **média é 2800g** e o **desvio padrão é 500g**
- Qual a probabilidade de um recém nascido pesar entre 2300g e 3300g?
- É o mesmo que um desvio padrão da média =>
  - **P(2300 <= X <= 3300) = 0,683**
- Qual a probabilidade de um recém nascido pesar entre 1800g e 3800g?
- É o mesmo que dois desvios padrão da média =>
  - **P(1800 <= X <= 3800) = 0,955**
- **O peso de aproximadamente 95% dos recém-nascidos está entre 1800g e 3800g.**

Distribuição Normal

- O R inclui funcionalidade para operações com distribuições de probabilidades. Para cada distribuição há 4 operações básicas indicadas pelas letras:
  - d calcula a densidade de probabilidade $f(x)$ no ponto
  - p calcula a função de probabilidade acumulada $F(x)$ no ponto
  - q calcula o quantil correspondente a uma dada probabilidade
  - r retira uma amostra aleatória da distribuição
- Para usar os funções deve-se combinar uma das letras acima com uma abreviatura do nome da distribuição. Por exemplo, para calcular probabilidades usamos: **pnorm()** para normal, **pexp()** para exponencial, **pbinom()** para binomial, **ppois()** para Poisson e assim por diante.

Distribuição Normal

- Por default as funções para distribuição normal assumem a distribuição normal padrão ***N*(*μ* = 0, *σ*² = 1)**.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

- Supondo um média = 100 e desvio-padrão = 10, vamos calcular:
- *P*[*X* < 95]
  - `pnorm(95, 100, 10)`
- *P*[90 < *X* < 110]
  - `pnorm(110, 100, 10) - pnorm(90, 100, 10)`
- *P*[*X* > 95]
  - `1 - pnorm(95, 100, 10)`
- Em dos nosso exemplos, P[2300 <= X <= 3300] com média = 2800 e desvio-padrão = 500
  - `pnorm(3300, 2800, 500) - pnorm(2300, 2800, 500)`

Distribuição Normal

- A distribuição normal foi essencial para o desenvolvimento de estatísticas, permitindo aproximação matemática
- Apesar de dados não serem normalmente distribuídos, os erros, médias e totais em amostras grandes são

Distribuição Binomial

- Resultados sim ou não estão no centro da análise: comprar/não comprar, clicar/não clicar, sobreviver/morrer
- Primordial para entender é ter a ideia de um conjunto de ensaios, cada um com dois resultados possíveis com possibilidades definidas
- **Ensaio** evento com resultado discreto
- **Sucesso** o resultado de interesse. Cara em uma moeda ou 1.
- **Binomial** dois resultados possíveis
- **Ensaio binomial** Ensaio com dois resultados

Distribuição Binomial

- Exemplificando:
  - Jogar uma moeda 10 vezes é um experimento com 10 ensaios, cada um podendo ter 2 resultados
  - Não é necessário que a probabilidade seja 50/50
  - Qualquer probabilidade cuja soma dê 1 é suficiente
  - Normalmente o resultado 1 é sucesso, e também costuma ser o evento mais raro
  - O uso do termo sucesso não quer dizer que o resultado é algo desejável ou benéfico

Distribuição Binomial

- A distribuição binomial é a distribuição de frequências do número de sucessos
- A distribuição binomial responde perguntas como:
  - Se a probabilidade de converter um clique em venda é de 0.02, qual a probabilidade de não observar nenhuma venda em 200 clicks?
  - `dbinom(x=0,size=200,p=0.02)`

Distribuição Binomial

- Muitas vezes estamos interessados em *x* ou menos sucessos, nesse caso
- `pbinom(10,size=200,p=0.02)`
- Ou seja, 99% de chance de 10 vendas ou menos
- A média da distribuição binomial é *n* x *p*. Sendo *n* o número de ensaios e *p* a probabilidade de sucesso
- A variância é de *n* x *p*(1-*p*)
- A distribuição binomial é quase indistinguível da distribuição normal, muitos procedimentos estatísticos usam a normal.

Outras Distribuições

- **t-Student**: é uma distribuição de probabilidade teórica. É semelhante à curva normal padrão com caudas mais largas
- **Poisson**: é uma distribuição de probabilidade de variável aleatória discreta que expressa a probabilidade de uma série de eventos ocorrer num certo período de tempo se estes eventos ocorrem independentemente de quando ocorreu o último evento.
- **Exponencial**: A distribuição exponencial é um tipo de distribuição contínua de probabilidade, representada por um parâmetro λ (lambda).
- **Weibull**: versão generalizada da exponencial, onde o taxa de eventos pode mudar com o tempo

Resumo

- Seleção aleatória de amostras pode reduzir o viés e produzir dados com melhor qualidade
- Conhecer os métodos de amostragem e as distribuições permite estimar erros
- Bootstrap é um método atrativo para determinar erros nas amostras

Referências

- Bruce, P.; Bruce, A.; **Estatística Prática para Cientista de Dados**: 50 Conceitos Essenciais; Rio de Janeiro; Alta Books; 2019.
- Morettin, P. A.; Bussab, W. O.; **Estatística Básica**. 8 ed. São Paulo: Saraiva, 2013.
- http://leg.ufpr.br/~shimakur/CE701/node36.html Acessado em: 16/03/2021
- http://www.leg.ufpr.br/~fernandomayer/aulas/ce083-2015-02/ce083_aula7_2015-02.html Acessado em: 16/06/2021

Análise de Dados Aplicada à Computação

PROF. M.SC HOWARD ROATTI