Trabalho\_02\_Amostragem

Carlos Alberto Alves de Meneses,20180003202

2023-03-12

PRÁTICA DA AMOSTRAGEM ALEATÓRIA SIMPLES, SISTEMÁTICA E ESTRATIFICADA EM BANCO DE DADOS.

**Introdução**

Uma amostra perfeita seria como Grandview: uma versão “reduzida” da população, refletindo todas as características de toda a população. Claro, nenhuma amostra perfeita pode existir para populações complicadas (mesmo que existisse, não saberíamos que era uma amostra perfeita sem medir toda a população). Mas uma boa amostra será representativa no sentido de que características de interesse na população podem ser estimadas a partir da amostra com um grau conhecido de precisão.

Os termos amostra aleatória simples, amostra estratificada, amostra por conglomerado e amostra sistemática são básicos para qualquer discussão sobre pesquisas por amostragem, então vamos análisar um banco de dados utilizando esses tipos de amostras probabilísticas.

**Objetivos**

* Extrair amostras aleatórias simplificada, sistemática e por estratos de tamanho n = 30, do banco de dados *TRAB02*
* Realizar a limpeza do banco excluindo as doenças que não serão utilizadas
* Analisar a variável *Tempoint* em relação a sua média, variância e tamanho populacional

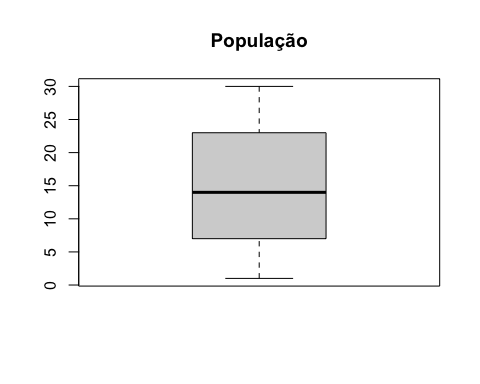
**Amostra aleatória simples**

Uma amostra aleatória simples (SRS) é a forma mais simples de amostra probabilística. Um SRS de tamanho n é considerado quando todo subconjunto possível de n unidades na população tem a mesma chance de ser a amostra. Ao pegar uma amostra aleatória, o investigador está, na verdade, misturando a população antes de pegar n unidades.

**1. Carregando a base de dados**

**Análise exploratória da população**

Como nesse estudo temos acesso a nossa população, então, podemos obter algumas estatísticas da nossa variável *Tempoint* dessa população.



##   
## One Sample t-test  
##   
## data: TRAB02$Tempoint  
## t = 23.401, df = 199, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 13.72222 16.24778  
## sample estimates:  
## mean of x   
## 14.985

## [1] 14.985

## [1] 82.01485

## [1] 9.056205

A análise nos mostra que a média da variável *Tempoint* na nossa população é =15, a variância é =82 e o desvio padrão é = 9,1.

Agora, iremos extrair amostras dessa população de tamanho n = 30 e realizar as análises para observar qual técnica de amostragem mais se aproxima dos reais valores das estatísticas obtidas da população.

Nota, esse tipo de experimento só é possível porque estamos tendo acesso aos dados de toda a população, porém, na vida real, na maioria das vezes isso não é possível.

**1.2 Extrair amostras aleatória simples do banco de dados de tamanho n =30**

Ao fixarmos uma semente garantimos obter os mesmos resultados em análises futuras.

## # A tibble: 6 × 31  
## ...1 Pacie…¹ Sexo Escol…² Unidade Idade Pesoi…³ Altura Peso30 Peso60 Tempo…⁴  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 68 68 2 2 1 37 64 1.79 57 47 9  
## 2 167 167 2 2 3 56 93 1.58 52 51 20  
## 3 129 129 1 1 2 26 55 1.59 49 62 8  
## 4 162 162 1 2 1 63 50 1.79 48 58 5  
## 5 43 43 2 1 2 33 80 1.89 55 57 14  
## 6 14 14 1 2 1 48 61 1.78 53 69 18  
## # … with 20 more variables: Dor1 <dbl>, Dor2 <dbl>, Dor3 <dbl>, Acred1 <dbl>,  
## # Acred2 <dbl>, Acred3 <dbl>, Pressao1 <dbl>, Pressao2 <dbl>, Pressao3 <dbl>,  
## # Satisfacao <dbl>, Hipertensao <dbl>, Diabetes <dbl>, Tabagismo <dbl>,  
## # Alcoolismo <dbl>, Obesidade <dbl>, Asma <dbl>, Alergias <dbl>,  
## # Cardiopatias <dbl>, Outras <dbl>, Nenhuma <dbl>, and abbreviated variable  
## # names ¹​Paciente, ²​Escolaridade, ³​Pesoinic, ⁴​Tempoint

## # A tibble: 6 × 31  
## ...1 Pacie…¹ Sexo Escol…² Unidade Idade Pesoi…³ Altura Peso30 Peso60 Tempo…⁴  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 7 7 1 4 2 53 67 1.58 76 50 10  
## 2 15 15 1 2 3 42 87 1.46 61 57 9  
## 3 16 16 2 4 1 45 92 1.72 61 63 19  
## 4 29 29 1 4 2 20 49 1.43 72 55 22  
## 5 48 48 1 1 3 35 78 1.81 79 61 28  
## 6 50 50 1 3 2 30 86 1.83 68 52 24  
## # … with 20 more variables: Dor1 <dbl>, Dor2 <dbl>, Dor3 <dbl>, Acred1 <dbl>,  
## # Acred2 <dbl>, Acred3 <dbl>, Pressao1 <dbl>, Pressao2 <dbl>, Pressao3 <dbl>,  
## # Satisfacao <dbl>, Hipertensao <dbl>, Diabetes <dbl>, Tabagismo <dbl>,  
## # Alcoolismo <dbl>, Obesidade <dbl>, Asma <dbl>, Alergias <dbl>,  
## # Cardiopatias <dbl>, Outras <dbl>, Nenhuma <dbl>, and abbreviated variable  
## # names ¹​Paciente, ²​Escolaridade, ³​Pesoinic, ⁴​Tempoint

**1.3 Manipulando dados com o pacote dplyr**

Usaremos o comando *select()* do pacote *dplyr* para extrai as colunas desejada,*Paciente, Sexo, Escolaridade, Unidade, Idade, Pesoinic, Altura, Peso30, Peso60, Tempoint, Dor1, Dor2, Dor3, Acred1, Acred2, Acred3, Pressao1, Pressao2, Pressao3, Satisfacao* e por útimo *Diabetes* da nossa tabela de dados *trab02* na forma de tabela.

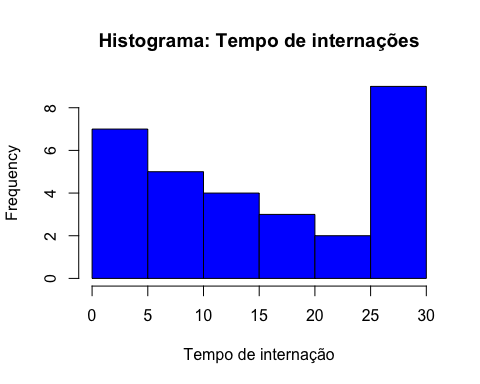
## # A tibble: 6 × 21  
## Pacie…¹ Sexo Escol…² Unidade Idade Pesoi…³ Altura Peso30 Peso60 Tempo…⁴ Dor1  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 7 1 4 2 53 67 1.58 76 50 10 4  
## 2 15 1 2 3 42 87 1.46 61 57 9 3  
## 3 16 2 4 1 45 92 1.72 61 63 19 5  
## 4 29 1 4 2 20 49 1.43 72 55 22 3  
## 5 48 1 1 3 35 78 1.81 79 61 28 2  
## 6 50 1 3 2 30 86 1.83 68 52 24 3  
## # … with 10 more variables: Dor2 <dbl>, Dor3 <dbl>, Acred1 <dbl>, Acred2 <dbl>,  
## # Acred3 <dbl>, Pressao1 <dbl>, Pressao2 <dbl>, Pressao3 <dbl>,  
## # Satisfacao <dbl>, Diabetes <dbl>, and abbreviated variable names ¹​Paciente,  
## # ²​Escolaridade, ³​Pesoinic, ⁴​Tempoint

**2.Amostragem Aleatória Simples**

Os dados de uma amostra aleatória simples (SRS) podem ser analisadsos usando funções R que são projetadas para dados que podem ser considerados independentes e distribuições de forma idêntica, e um SRS pode ser selecionado usando a função de amostra do R.

**2.1 Histograma**

Primeiro, desenhamos um histograma da variável *Tempoint*.



**2.2 Usando fórmulas em SDA**

Funções como t.test calcularão intervalos de confiança sem uma correção de população finita (fpc).

##   
## One Sample t-test  
##   
## data: dados1$Tempoint  
## t = 9.033, df = 29, p-value = 6.292e-10  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 12.40310 19.66356  
## sample estimates:  
## mean of x   
## 16.03333

Portanto, como p-valor = 0.00000000001081, nós rejeitamos a hipótese nula, logo,a média verdadeira não é igual a zero. Observamos um intervalode confiança de 95% = 14.40718 21.12615 e uma média estimada de 17,77.

Erros padrão e intervalos de confiança que incorporam o fpc podem ser calculados diretamente com as fórmulas.

* **Cálculo da média amostral**

Com relação à , a média amostral

*é um estimador não viesado da média populacional*.

## [1] 16.03333

com variância amostral dada por:

onde

*é um estimador não viesado da variância amostral*.

## [1] 13.7737

e com o desvio padrão dado por:

## [1] 3.711294

**Cálculo do intervalo de confiança**

Calculando o intervalo de confiança pela fórmula direta usando a distribuição t

## Lim Inf Lim Sup   
## 8.442885 23.623781

Para obter estimativas para o total da população, multiplicamos cada um de ybar, seybar e mean\_CI por N=200

## [1] 742.2588

## Lim Inf Lim Sup   
## 1688.577 4724.756

Calculando o coeficiente de variação da média.

## [1] 0.2314736

**3. Amostragem Aleatória Sistemática-AAS**

Às vezes, a amostragem sistemática é usada como um substituto para a amostragem aleatória simples, quando não existe uma lista da população ou quando a lista está em ordem aproximadamente aleatória. Para obter uma amostra sistemática, escolha um tamanho de amostra n. Se N/n for um número inteiro, seja k = N/ n; caso contrário, seja k o próximo inteiro após N/ n. Em seguida, encontre um inteiro aleatório R entre 1 e k, que determina a amostra como sendo as unidades numeradas R, R + k, R + 2k, etc.

Utilizando o mesmo banco de dados *TRAB02* agora para extrair uma amostra aleatória sistemática.

## # A tibble: 200 × 31  
## ...1 Paciente Sexo Escolaridade Unidade Idade Pesoinic Altura Peso30 Peso60  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 1 2 1 3 22 58 1.59 76 61  
## 2 2 2 1 1 3 52 82 1.59 64 56  
## 3 3 3 1 3 3 54 90 1.68 48 57  
## 4 4 4 2 1 2 57 67 1.6 48 54  
## 5 5 5 2 2 3 40 69 1.82 56 66  
## 6 6 6 2 2 2 55 78 1.59 60 51  
## 7 7 7 1 4 2 53 67 1.58 76 50  
## 8 8 8 1 4 1 38 89 1.6 46 62  
## 9 9 9 1 3 1 37 91 1.88 53 53  
## 10 10 10 1 2 3 36 71 1.58 67 55  
## # … with 190 more rows, and 21 more variables: Tempoint <dbl>, Dor1 <dbl>,  
## # Dor2 <dbl>, Dor3 <dbl>, Acred1 <dbl>, Acred2 <dbl>, Acred3 <dbl>,  
## # Pressao1 <dbl>, Pressao2 <dbl>, Pressao3 <dbl>, Satisfacao <dbl>,  
## # Hipertensao <dbl>, Diabetes <dbl>, Tabagismo <dbl>, Alcoolismo <dbl>,  
## # Obesidade <dbl>, Asma <dbl>, Alergias <dbl>, Cardiopatias <dbl>,  
## # Outras <dbl>, Nenhuma <dbl>

Vamos selecionar uma amostra de tamanho n = 30 utilizando o método da amostragem aleatória sistemática.

## # A tibble: 6 × 31  
## ...1 Pacie…¹ Sexo Escol…² Unidade Idade Pesoi…³ Altura Peso30 Peso60 Tempo…⁴  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 4 4 2 1 2 57 67 1.6 48 54 10  
## 2 11 11 2 3 3 34 50 1.46 79 68 29  
## 3 18 18 1 2 2 35 53 1.84 77 57 8  
## 4 25 25 1 3 3 64 69 1.76 54 47 10  
## 5 32 32 1 3 2 52 45 1.52 48 58 14  
## 6 39 39 1 3 1 59 94 1.83 79 46 24  
## # … with 20 more variables: Dor1 <dbl>, Dor2 <dbl>, Dor3 <dbl>, Acred1 <dbl>,  
## # Acred2 <dbl>, Acred3 <dbl>, Pressao1 <dbl>, Pressao2 <dbl>, Pressao3 <dbl>,  
## # Satisfacao <dbl>, Hipertensao <dbl>, Diabetes <dbl>, Tabagismo <dbl>,  
## # Alcoolismo <dbl>, Obesidade <dbl>, Asma <dbl>, Alergias <dbl>,  
## # Cardiopatias <dbl>, Outras <dbl>, Nenhuma <dbl>, and abbreviated variable  
## # names ¹​Paciente, ²​Escolaridade, ³​Pesoinic, ⁴​Tempoint

**3.1 Manipulando dados com o pacote dplyr**

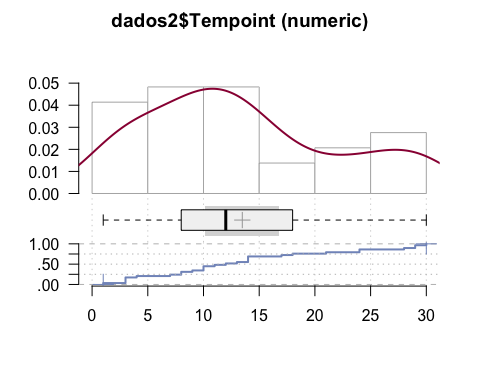
Usaremos o comando *select()* do pacote *dplyr* para extrai as colunas desejada para obter o nosso banco de dados com a amostra aleatória sistemática.

## # A tibble: 6 × 21  
## Pacie…¹ Sexo Escol…² Unidade Idade Pesoi…³ Altura Peso30 Peso60 Tempo…⁴ Dor1  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 4 2 1 2 57 67 1.6 48 54 10 2  
## 2 11 2 3 3 34 50 1.46 79 68 29 2  
## 3 18 1 2 2 35 53 1.84 77 57 8 4  
## 4 25 1 3 3 64 69 1.76 54 47 10 5  
## 5 32 1 3 2 52 45 1.52 48 58 14 2  
## 6 39 1 3 1 59 94 1.83 79 46 24 1  
## # … with 10 more variables: Dor2 <dbl>, Dor3 <dbl>, Acred1 <dbl>, Acred2 <dbl>,  
## # Acred3 <dbl>, Pressao1 <dbl>, Pressao2 <dbl>, Pressao3 <dbl>,  
## # Satisfacao <dbl>, Diabetes <dbl>, and abbreviated variable names ¹​Paciente,  
## # ²​Escolaridade, ³​Pesoinic, ⁴​Tempoint

**3.2 Estatística descritiva com o pacote desctools**

O código a seguir realiza a análise exploratória da variável *Tempoint*, tempo de internação, da nossa amostra aleatória sistemática.

## ------------------------------------------------------------------------------   
## dados2$Tempoint (numeric)  
##   
## length n NAs unique 0s mean meanCI'  
## 29 29 0 18 0 13.48 10.17  
## 100.0% 0.0% 0.0% 16.80  
##   
## .05 .10 .25 median .75 .90 .95  
## 3.00 3.00 8.00 12.00 18.00 28.20 29.00  
##   
## range sd vcoef mad IQR skew kurt  
## 29.00 8.72 0.65 7.41 10.00 0.50 -0.92  
##   
## lowest : 1.00, 3.00 (4), 4.00, 7.00, 8.00 (2)  
## highest: 21.00, 24.00 (2), 28.00, 29.00 (2), 30.00  
##   
## ' 95%-CI (classic)



Ao analisarmos a saída do comando, vemos que é apresentado o resumo estatístico da variável *Tempoit* , tempo de internação: temos o número de observações (29), quantas foram válidas (n = 29), quantas foram perdidas (= 0), quantas observações únicas (unique = 19 n), quantos zeros aparecem (=0), o valor da média (mean = 15,34), o intervalo de confiança da média (meanCI = 11,91 e 18,78) indicando que o tempo médio de internação é estimado entre 12 e 19 dias.

Os percentis são apresentados em (.05 = 2,00; .10 = 3,60; …; .95 = 28,60), são os valores dos percentis 5, 10, 25, 50 (mediana), 75, 90 e 95. As medidas de variabilidade são a amplitude (range = 28,00), o desvio-padrão (sd = 9,02), o coeficiente de variação (vcoef = 0,59), o desvio médio absoluto (mad = 10,38), a amplitude interquartil (IQR = 14,00), o coeficiente de assimetria (skew = -0,07) e o coeficiente de Curtose (kurt= -1,35).

Após as medidas descritivas, uma tabela de frequência se apresenta para as variáveis quantitativas.

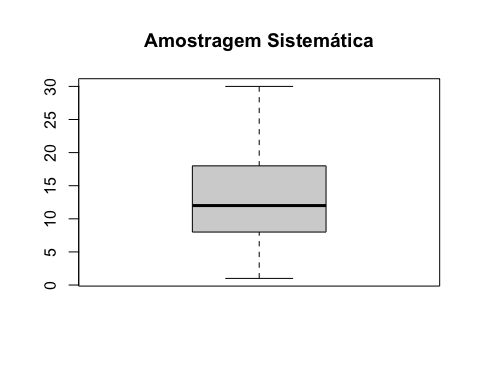
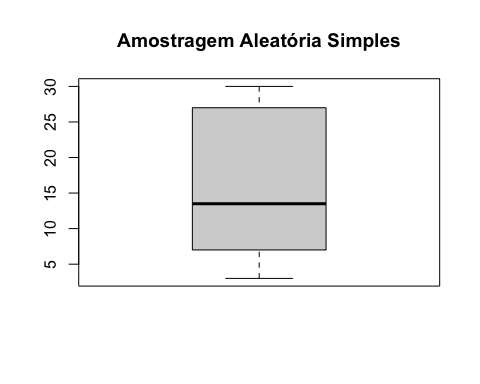
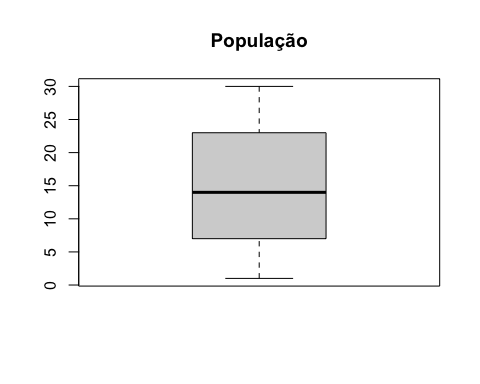
**3.3 Estimador HT para o total populacional para Tempoint**

## N dados2.Tempoint  
## Estimation 203 2737.00000  
## Standard Error 0 302.08145  
## CVE 0 11.03695  
## DEFF NaN 1.00000

**3.4 Análise Descritiva**

Análise do tempo de internação

## named list()



Através dos gráficos boxplots, observa-se que a tempo de internação varia conforme o plano amostral utilizado.

Funções como t.test calcularão intervalos de confiança sem uma correção de população finita (fpc).

##   
## One Sample t-test  
##   
## data: dados2$Tempoint  
## t = 8.3262, df = 28, p-value = 4.651e-09  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 10.16572 16.79980  
## sample estimates:  
## mean of x   
## 13.48276

Portanto, como p-valor = 0.0000000006431, nós rejeitamos a hipótese nula, logo,a média verdadeira não é igual a zero. Observamos um intervalode confiança de 95% = 11,91346 18,77620 e uma média estimada de 15,34483.

**3.5 Cálculos da média e variância amostral e populacional**

* **Cálculo da média amostral**

Com relação à , a média amostral

*é um estimador não viesado da média populacional*.

## [1] 13.48276

com variância amostral dada por:

onde

*é um estimador não viesado da variância amostral*.

## [1] 10.69971

e com o desvio padrão dado por:

## [1] 3.27104

**Cálculo do intervalo de confiança**

Calculando o intervalo de confiança pela fórmula direta usando a distribuição t

## Lim Inf Lim Sup   
## 6.79273 20.17279

Para obter estimativas para o total da população, multiplicamos cada um de ybar, seybar e mean\_CI por N=200

## [1] 654.2081

## Lim Inf Lim Sup   
## 1358.546 4034.557

Calculando o coeficiente de variação da média.

## [1] 0.2426091

**4. Amostragem estratificada**

Nesse tipo de amostragem, uma população heterogênea é estratificada ou dividida em subpopulações ou estratos homogênios,e, em cada estrato, uma amostra é retirada. Desta forma, definimos, inicialmente, o número de estratos e obtemos, assim, o tamanho de cada um deles; para cada estrato, espercificamos quantos elementos serão retirados da subpopulação, podendo ser uma alocação uniforme ou proporcional.

Utilizando o mesmo banco de dados *TRAB02* agora para extrair uma amostra estratificada.

**4.1 Limpeza do banco de dados**

Iremos realizar a limpeza do banco de dados retirando as variáveis que não iremos utilizar.

## # A tibble: 6 × 21  
## Pacie…¹ Sexo Escol…² Unidade Idade Pesoi…³ Altura Peso30 Peso60 Tempo…⁴ Dor1  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 2 1 3 22 58 1.59 76 61 1 2  
## 2 2 1 1 3 52 82 1.59 64 56 2 2  
## 3 3 1 3 3 54 90 1.68 48 57 4 3  
## 4 4 2 1 2 57 67 1.6 48 54 10 2  
## 5 5 2 2 3 40 69 1.82 56 66 27 4  
## 6 6 2 2 2 55 78 1.59 60 51 11 3  
## # … with 10 more variables: Dor2 <dbl>, Dor3 <dbl>, Acred1 <dbl>, Acred2 <dbl>,  
## # Acred3 <dbl>, Pressao1 <dbl>, Pressao2 <dbl>, Pressao3 <dbl>,  
## # Satisfacao <dbl>, Diabetes <dbl>, and abbreviated variable names ¹​Paciente,  
## # ²​Escolaridade, ³​Pesoinic, ⁴​Tempoint

## # A tibble: 6 × 22  
## Pacie…¹ Sexo Escol…² Unidade Idade Pesoi…³ Altura Peso30 Peso60 Tempo…⁴ Dor1  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 2 1 3 22 58 1.59 76 61 1 2  
## 2 2 1 1 3 52 82 1.59 64 56 2 2  
## 3 3 1 3 3 54 90 1.68 48 57 4 3  
## 4 4 2 1 2 57 67 1.6 48 54 10 2  
## 5 5 2 2 3 40 69 1.82 56 66 27 4  
## 6 6 2 2 2 55 78 1.59 60 51 11 3  
## # … with 11 more variables: Dor2 <dbl>, Dor3 <dbl>, Acred1 <dbl>, Acred2 <dbl>,  
## # Acred3 <dbl>, Pressao1 <dbl>, Pressao2 <dbl>, Pressao3 <dbl>,  
## # Satisfacao <dbl>, Diabetes <dbl>, strwt <dbl>, and abbreviated variable  
## # names ¹​Paciente, ²​Escolaridade, ³​Pesoinic, ⁴​Tempoint

Vamos selecionar uma amostra utilizando o método da amostragem estratificada.

**4.2 Método de Alocação**

## # A tibble: 6 × 22  
## Pacie…¹ Sexo Escol…² Unidade Idade Pesoi…³ Altura Peso30 Peso60 Tempo…⁴ Dor1  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 2 1 3 22 58 1.59 76 61 1 2  
## 2 2 1 1 3 52 82 1.59 64 56 2 2  
## 3 3 1 3 3 54 90 1.68 48 57 4 3  
## 4 4 2 1 2 57 67 1.6 48 54 10 2  
## 5 5 2 2 3 40 69 1.82 56 66 27 4  
## 6 6 2 2 2 55 78 1.59 60 51 11 3  
## # … with 11 more variables: Dor2 <dbl>, Dor3 <dbl>, Acred1 <dbl>, Acred2 <dbl>,  
## # Acred3 <dbl>, Pressao1 <dbl>, Pressao2 <dbl>, Pressao3 <dbl>,  
## # Satisfacao <dbl>, Diabetes <dbl>, strwt <dbl>, and abbreviated variable  
## # names ¹​Paciente, ²​Escolaridade, ³​Pesoinic, ⁴​Tempoint

## [1] "Paciente" "Sexo" "Escolaridade" "Unidade" "Idade"   
## [6] "Pesoinic" "Altura" "Peso30" "Peso60" "Tempoint"   
## [11] "Dor1" "Dor2" "Dor3" "Acred1" "Acred2"   
## [16] "Acred3" "Pressao1" "Pressao2" "Pressao3" "Satisfacao"   
## [21] "Diabetes" "strwt"

## # A tibble: 6 × 22  
## Pacie…¹ Sexo Escol…² Unidade Idade Pesoi…³ Altura Peso30 Peso60 Tempo…⁴ Dor1  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 2 1 3 22 58 1.59 76 61 1 2  
## 2 2 1 1 3 52 82 1.59 64 56 2 2  
## 3 3 1 3 3 54 90 1.68 48 57 4 3  
## 4 4 2 1 2 57 67 1.6 48 54 10 2  
## 5 5 2 2 3 40 69 1.82 56 66 27 4  
## 6 6 2 2 2 55 78 1.59 60 51 11 3  
## # … with 11 more variables: Dor2 <dbl>, Dor3 <dbl>, Acred1 <dbl>, Acred2 <dbl>,  
## # Acred3 <dbl>, Pressao1 <dbl>, Pressao2 <dbl>, Pressao3 <dbl>,  
## # Satisfacao <dbl>, Diabetes <dbl>, strwt <dbl>, and abbreviated variable  
## # names ¹​Paciente, ²​Escolaridade, ³​Pesoinic, ⁴​Tempoint

## [1] 200

## [1] 3 2 1

##   
## 1 2 3   
## 42 103 55

Usaremos as informações sobre as *unidades* para alocar uma amostra estratificada.

Lembrando que pelo questionário temos as unidades assim distribuídas:

1. Mangabeira 2. Bancários 3. Valentina

**4.3 Alocação proporcional**

Com alocação proporcional, os tamanhos de amostra de estrato são proporcionais aos tamanhos de estrato populacional. Uma alocação proporcional é fácil de calcular no R; simplesmente multiplique pelo tamanho de amostra desejado. Nesse trabalho por exemplo, a unidade 2-Bancários tem 103 unidades e a população tem 200 observações. Para uma amostra de tamnaho n = 30, a alocação proporcional selecionará 16 amostra da unidade Bancários, 06 da unidade de Mangabeira e 08 da unidade do Valentina.

## [1] 15.45

## [1] 6.3

## [1] 8.25

Os valores em *propalloc* são frações, então arredondamos para os inteiros mais próximos para obter o tamanho da amostra.

##   
## 1 2 3   
## 6.30 15.45 8.25

##   
## 1 2 3   
## 6 15 8

## [1] 30

## [1] 29

**4.4 Alocação de Neyman**

Para a alocação de Neyman, nós precisamos fornecer informações adicionais sobre as variações do estrato. Por exemplo, a variação na unidade dos Bancários é duas vezes a variação da unidade do Valentina. Utilizando o vetor *popsize* que foi calculado no código anterior, temos:

##   
## 1 2 3   
## 4.732612 18.175351 7.092037

##   
## 1 2 3   
## 5 18 7

## [1] 30

**4.5 Alocação ótima**

A alocação ótima pode ser feita de forma semelhante definindo custos ou custos relativos para amostragem em cada estrato.

##   
## 1 2 3   
## 6.30 15.45 8.25

##   
## 1 2 3   
## 6 15 8

## [1] 29

**4.6 Usando a função strata do pacote de amostragem**

Uma alternativa é usar estratos de função para selecionar uma amostra aleatória estratificada.

* Ordenando a população por estratos

Usaremos a função *strata* para selecionar as unidades para a amostra.

Vejamos o número de unidades selecionados dentro de cada estratos.

##   
## 1 2 3   
## 42 103 55

## Unidade ID\_unit Prob Stratum  
## 1 1 1 1 1  
## 2 1 2 1 1  
## 3 1 3 1 1  
## 4 1 4 1 1  
## 5 1 5 1 1  
## 6 1 6 1 1

Vamos extrair a cabeça da amostra (strsample2)

## Paciente Sexo Escolaridade Idade Pesoinic Altura Peso30 Peso60 Tempoint Dor1  
## 1 8 1 4 38 89 1.60 46 62 25 2  
## 2 9 1 3 37 91 1.88 53 53 13 3  
## 3 14 1 2 48 61 1.78 53 69 18 1  
## 4 16 2 4 45 92 1.72 61 63 19 5  
## 5 39 1 3 59 94 1.83 79 46 24 1  
## 6 44 2 3 34 60 1.84 48 58 23 2  
## Dor2 Dor3 Acred1 Acred2 Acred3 Pressao1 Pressao2 Pressao3 Satisfacao Diabetes  
## 1 1 4 0 1 1 16 15 12 1 1  
## 2 2 1 1 0 0 14 17 12 1 1  
## 3 4 1 0 0 0 16 13 16 2 0  
## 4 2 1 0 1 1 11 13 10 3 0  
## 5 2 3 0 0 1 18 13 14 2 0  
## 6 3 2 0 0 1 16 11 10 2 0  
## strwt Unidade ID\_unit Prob Stratum  
## 1 6.666667 1 1 1 1  
## 2 6.666667 1 2 1 1  
## 3 6.666667 1 3 1 1  
## 4 6.666667 1 4 1 1  
## 5 6.666667 1 5 1 1  
## 6 6.666667 1 6 1 1

O dataframe *index2* contém as variáveis do estrato, os identificadores das unidades selecionadas para compor a amostra e a probabilidade de inclusão de cada unidade na amostra. A função *getdata* então extrai as unidades amostradas dos dados populacionais.

A função *strata* fornece as probabilidades de inclusão para as unidades amostrais, mas não os pesos.

Podemos calcular os pesos de amostragem tomando o recíproco das probabilidades de inclusão. Ao calcular pesos para uma amostra aleatória estratificada, sempre verifique se os pesos somam os tamanhos da população do estrato. Se eles não somam os tamanhos populacionais do estrato, então, cometemos um erro em algum lugar nos cálculos de peso.

* Calculando os pesos amostras.

Verificamos primeiro se nenhuma probabilidade é zero.

## [1] 0

Agora, verificamos se os pesos amostrais somam os tamanhos populacionais para cada estrato.

## 1 2 3   
## 42 103 55

**4.7 Calculando Estatísticas de uma Amostra Aleatória Estratificada**

A função *svydesign* do pacote de pesquisa pode ser usada para inserir as informações da amostra aleatória estratificada, e as funções *svymean* e *svytotal* calcularão médias e totais estimados de uma amostra aleatória estratificada.

O conjunto de dados agstrat é uma amostra aleatória estratificada retirada dos dados populacionais agpop com alocação proporcional. Primeiro, vamos olhar para os dados.

## tibble [200 × 22] (S3: tbl\_df/tbl/data.frame)  
## $ Paciente : num [1:200] 1 2 3 4 5 6 7 8 9 10 ...  
## $ Sexo : num [1:200] 2 1 1 2 2 2 1 1 1 1 ...  
## $ Escolaridade: num [1:200] 1 1 3 1 2 2 4 4 3 2 ...  
## $ Unidade : num [1:200] 3 3 3 2 3 2 2 1 1 3 ...  
## $ Idade : num [1:200] 22 52 54 57 40 55 53 38 37 36 ...  
## $ Pesoinic : num [1:200] 58 82 90 67 69 78 67 89 91 71 ...  
## $ Altura : num [1:200] 1.59 1.59 1.68 1.6 1.82 1.59 1.58 1.6 1.88 1.58 ...  
## $ Peso30 : num [1:200] 76 64 48 48 56 60 76 46 53 67 ...  
## $ Peso60 : num [1:200] 61 56 57 54 66 51 50 62 53 55 ...  
## $ Tempoint : num [1:200] 1 2 4 10 27 11 10 25 13 17 ...  
## $ Dor1 : num [1:200] 2 2 3 2 4 3 4 2 3 2 ...  
## $ Dor2 : num [1:200] 1 3 2 2 1 3 5 1 2 1 ...  
## $ Dor3 : num [1:200] 2 4 4 3 4 3 2 4 1 4 ...  
## $ Acred1 : num [1:200] 0 1 0 1 0 1 0 0 1 0 ...  
## $ Acred2 : num [1:200] 1 1 0 0 0 1 1 1 0 1 ...  
## $ Acred3 : num [1:200] 0 1 0 1 0 1 0 1 0 0 ...  
## $ Pressao1 : num [1:200] 14 13 15 10 11 18 14 16 14 12 ...  
## $ Pressao2 : num [1:200] 12 18 17 16 14 12 15 15 17 17 ...  
## $ Pressao3 : num [1:200] 12 11 13 14 14 16 10 12 12 14 ...  
## $ Satisfacao : num [1:200] 3 5 4 3 4 2 3 1 1 3 ...  
## $ Diabetes : num [1:200] 0 1 1 1 0 1 1 1 1 0 ...  
## $ strwt : num [1:200] 6.67 6.67 6.67 6.67 6.67 ...

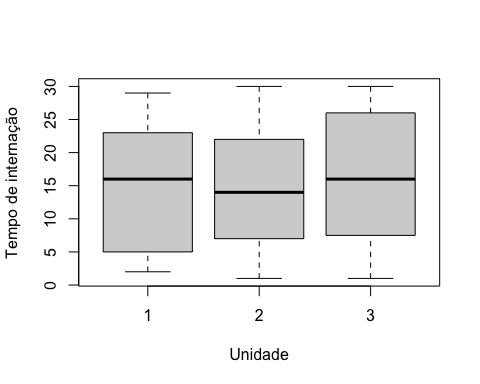
## [1] "Paciente" "Sexo" "Escolaridade" "Unidade" "Idade"   
## [6] "Pesoinic" "Altura" "Peso30" "Peso60" "Tempoint"   
## [11] "Dor1" "Dor2" "Dor3" "Acred1" "Acred2"   
## [16] "Acred3" "Pressao1" "Pressao2" "Pressao3" "Satisfacao"   
## [21] "Diabetes" "strwt"

## [1] 200

## [1] 3 2 1

##   
## 1 2 3   
## 42 103 55

O gráfico abaixo apresenta um boxplot para a variável *unidade*.



Podemos observar que a unidade 2 (Bancários) possui a maior mediana e a menor variabilidade, enquanto a unidade 1 (Mangabeira) e 3 (Valentina) possuem a mesma mediana, porém a unidade Valentina posui uma maior variabilidade.

**4.8 Calculando as estimativas da amostra aleatória estratificada**

Primeiro, configuramos as informações para o desing da pesquisa.

##   
## 42 55 103   
## 42 55 103

Informações de projeto de entrada para *dados4* dstr

## Stratified Independent Sampling design  
## svydesign(id = ~1, strata = ~Unidade, weigts = ~strwt, fpc = ~popsize,   
## data = dados4)

Todo o trabalho de especificação das informaçõs do projeto é feito na função *svydesing*; depois de definir o projeto lá, as funções *svymean* e *svytotal* são usados exatamente como fizemos para Amostra Aleatória Simples (SRS).

* Cáculo da média

## mean SE  
## Tempoint 14.985 0

Portanto, obtemos uma média de 15 dias de tempo de interanação.

* Cálculo do intervalo de confiança

## 2.5 % 97.5 %  
## Tempoint 14.985 14.985

* Cálculos do total, SE e CI

## total SE  
## Tempoint 2997 0

* Mostrando os graus de liberdade para o projeto

## [1] 197

* Calcula o intervalo de confiança usando os graus de liberdade

## 2.5 % 97.5 %  
## Tempoint 2997 2997

##   
## One Sample t-test  
##   
## data: dados4$Tempoint  
## t = 23.401, df = 199, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 13.72222 16.24778  
## sample estimates:  
## mean of x   
## 14.985

Se a amostra tiver poucas observações, precisamos especificar os graus de liberdade e usar a distribuição t para calcular os intervalos de confiança.

Portanto, pelo IC, a verdadeira média está localizada entre 14 e 16 dias de internação.

* Cálculo da média e da variância de estratos

A função *svyby* calculará estatísticas e seus erros padrão para subgrupos de dados. Aqui ele foi usado para calcular as médias e totais do estrato.

## Unidade Tempoint se  
## 1 1 14.45238 0  
## 2 2 14.55340 0  
## 3 3 16.20000 0

Portanto, temos as médias do tempo de internação de:

15 dias para a unidade de Mangabeira, 15 dias para a unidade dos Bancários e de 16 dias para a unidade de Valentina.

* Cálculo do total e SE

## Unidade Tempoint se  
## 1 1 607 0  
## 2 2 1499 0  
## 3 3 891 0

Portanto, obtemos os totais de tempo de internações para cada unidade.

* Usando a função tapply

Podemos calcular as estatísticas resumidas diretamente para cada estrato usando a *função tapply* e, em seguida, usar as fórmulas do SDA para calcular os erros padrão para cada média ou total estimado do estrato.

* Calculando a média dentro de cada Unidade

## 1 2 3   
## 14.45238 14.55340 16.20000

* Cálculo da variância dentro de cada Unidade

## 1 2 3   
## 79.96109 77.46526 93.12593

* Verificando os erros padrão pela fórmula direta

##   
## 1 2 3   
## 0 0 0

**4.9 Estimando proporções de uma Amostra Aleatória Estratificada**

Uma proporção é um caso especial de média de uma variável que assume os valores 1 e 0.

## [1] 30

* Criando a variavel lt30d

Execute novamente o svydesing porque o conjunto de dados agora tem uma nova variável

* Calcula proporção, SE e intervalo de confiança

## mean SE  
## lt30d 0.975 0

* Cálculo do intervalo de confiança

## 2.5 % 97.5 %  
## lt30d 0.975 0.975

* Cálculo total, SE e CI

## total SE  
## lt30d 195 0

**Conclusão**

Através das análises observadas, podemos concluir que a seleção da amostra através do método da amostra estratificada, conseguimos obter estatísticas que mais se aproximam dos reais parâmetros da nossa população.

**Referência**

LU, Yan; LOHR, Sharon L.. Using R Software with Sampling: design and analysis. 3. ed. New York: Crc Press, 2022. 205 p.