Introdução a Métodos Estatísticos para a Bioinformática

Profa. Júlia Maria Pavan Soler

pavan@ime.usp.br

IBI 5086 – Bioinformática - IME/USP

2º Sem/2023

Programa de IBI5086

- Álgebra linear básica: cálculo matricial, determinantes, sistemas lineares, produto interno, norma, ortogonalidade, autovalores e autovetores
- Estrutura de Dados: variáveis (resposta, explicativas, covariáveis, tipo da var), unidades amostrais e experimentais, aleatorização, dependência das observações

1.1. Comparação de Grupos (2 ou mais): Testes Clássicos (teste t, Wilcoxon, modelos ANOVA) e Testes de Aleatorização, Comparações Múltiplas, Simulação de dados

1.2. Análise de Tabelas de Contingência: Testes Qui-Quadrado, Regressão Logística.

2. Análise Multivariada de Dados: Componentes Principais, Análise Discriminante e Classificação, Análise de Agrupamento, Correlação Canônica, modelos MANOVA

3. Simulação de Monte Carlo, Intervalos de Confiança Bootstrap

Consulta - Mentimeter

Arquivo Pulse

	P1	P2	Ran	Fu	Sex	Altura	Peso	Ativ
1	64	88	1	2	1	66.00	140	2
2	58	70	1	2	1	72.00	145	2
…								
34	62	98	1	1	2	62.75	112	2
35	80	128	1	2	2	68.00	125	2
36	62	62	2	2	1	74.00	190	1
37	60	62	2	2	1	71.00	155	2
…								
91	86	84	2	2	2	67.00	150	3
92	76	76	2	2	2	61.75	108	2

Estrutura dos dados!

$\Rightarrow$ Simulação de Dados

Como avaliar se existe efeito da corrida na pulsação de estudantes?

$\Rightarrow$ Análise descritiva e inferencial (comparação de dois grupos)

Estrutura de Dados

DADOS PULSE

População sob estudo: Estudantes (de Estatística Brasileiros)

- ✓ **Amostra**: 92 estudantes (unidades amostrais, experimentais e de mensuração)
- ✓ **Resposta de interesse**: pulsação (batimentos/minuto) dos estudantes (P1, P2)
- ✓ **Fatores sob estudo**: Corrida (avaliar seu efeito na pulsação dos estudantes)
  1 Fator em dois níveis (Correr: Ran=1, Não Correr: Ran=2)
- ✓ **Variáveis de controle (covar.)**: tabagismo, sexo, altura, peso, atividade física

⇒Delineamento Completamente Aleatorizado: atribuição aleatória dos "tratamentos" aos 92 estudantes (n=92: **n1=35** submetidos à corrida, **n2=57** repouso/controle)

	P1	P2	Ran	Fu	Sex	Altura	Peso	Ativ
1	64	88	1	2	1	66.00	140	2
2	58	70	1	2	1	72.00	145	2
…								
34	62	98	1	1	2	62.75	112	2
35	80	128	1	2	2	68.00	125	2
36	62	62	2	2	1	74.00	190	1
37	60	62	2	2	1	71.00	155	2
…								
91	86	84	2	2	2	67.00	150	3
92	76	76	2	2	2	61.75	108	2

Tipos de Variáveis:

Pulsação: quantitativa discreta

Ran, Fu, Sex: qualitativa nominal (categórica)

Altura: quantitativa contínua

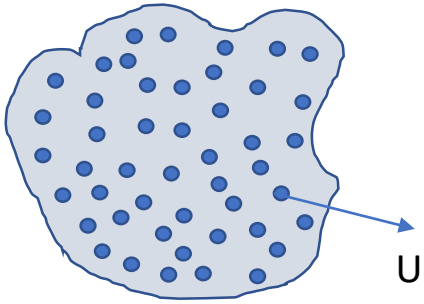
Peso: contínua (anotada como discreta)

Ativ: qualitativa ordinal

Discutir: Por que aleatorizar?
Dependência nas respostas

Racional da Análise Estatística

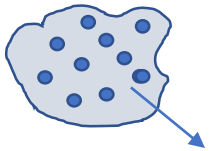
POPULAÇÃO Alvo do Estudo



Unidades a serem amostradas

Técnicas de Amostragem

AMOSTRA



Unidades já amostradas
Unidades experimentais
Unidades de mensuração

VARIÁVEL ALEATÓRIA (resposta, desfecho de interesse)

PARÂMETROS (valores fixos DESCONHECIDOS)

Dados: VALORES DA VARIÁVEL

ESTATÍSTICAS (valores calculados na amostra)

Contextualizando

Número de brasileiros com diabetes aumentou 31% nos últimos dois anos

No Dia Mundial do Diabetes, uma entidade divulga levantamento inédito sobre a doença feito em 138 países. E os dados do Brasil são especialmente negativos

Por **Maria Teresa Santos** 14 nov 2019, 11h59

9ª ed. Atlas de Diabetes, 2019

"A epidemia de diabetes diz respeito a todos", clamam cientistas em estudo

Relatório publicado no "The Lancet" por 44 especialistas traça estratégias para combater a doença, ressaltando que sua prevenção é de responsabilidade coletiva e evita milhões de mortes por ano

Brasil: terceira maior população de crianças e adolescentes no mundo (95.800 jovens ≤ 20 anos) com diabetes tipo 1.

Diabetes Overview

American Diabetes Association.

Learn the Genetics of Diabetes

Estima-se que 463 milhões de pessoas tenham diabetes no mundo.

POPULAÇÃO

Jovens (≤20 anos) brasileiros com diabetes tipo I (N=95.800?)

Variáveis de interesse: nível de HbA1c, Presença ou não de genes específicos

Parâmetros: (DESCONHECIDOS)

- Nível **Médio** de HbA1c (μ)
- **Prevalência** (π) de genes HLA-DR3, DR4, DR7 e DR9

AMOSTRA

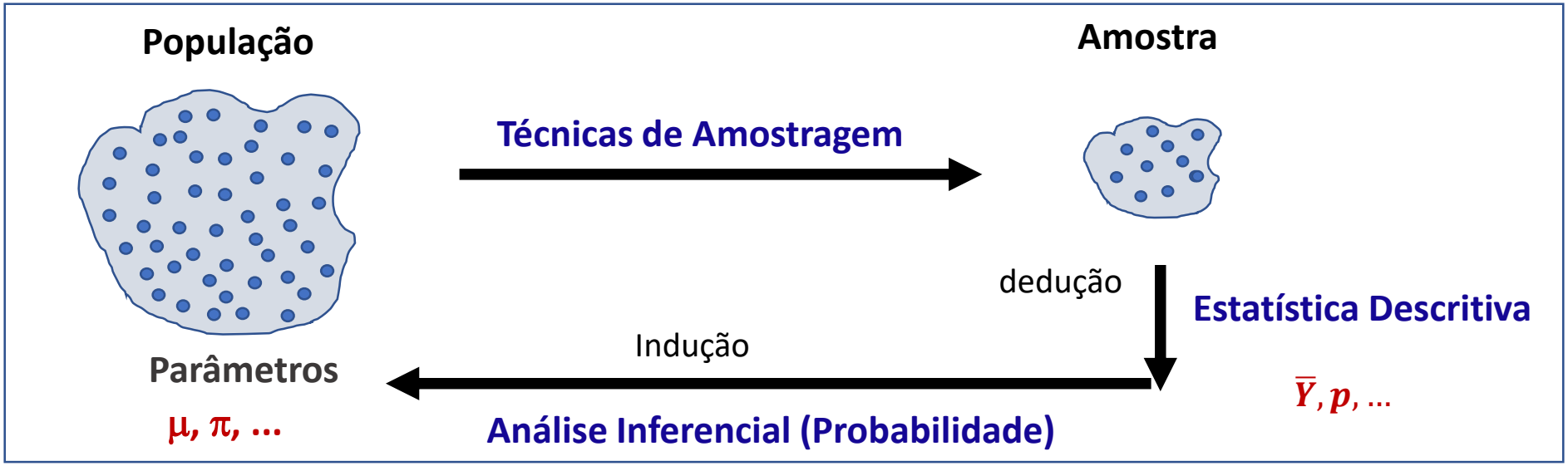
Estudo brasileiro multicêntrico com n=800 jovens (≤ 20 anos) com diabetes tipo I

Variáveis avaliadas: nível de HbA1c, presença de genes (entre outras)

Estatísticas:

- **Média de HbA1c na amostra** $(\bar{Y})$
- **Proporção** $(\hat{p})$ **de jovens brasileiros na amostra** que carregam os genes sob estudo

Grandes Áreas da Estatística



	Estística descritiva	Inferência
Variável Quantitativa Y: Média (μ)	Média amostral: $\bar{Y}$	H_0: $\mu=\mu_{Ref}$ x H_1: $\mu\neq\mu_{Ref}$ IC95%(μ) = $\bar{Y} \pm 1{,}96se$
Variavel Binária Y: Prevalência (π)	Proporção amostral: p	H_0: $\pi=0$ x H_1: $\pi>0$ IC95%(π) = $p \pm 1{,}96se$

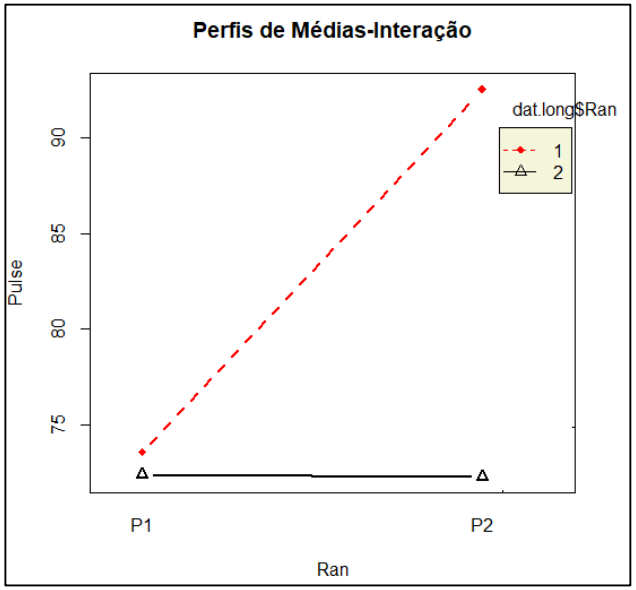
O tipo da VARIÁVEL sob estudo é decisivo na análise de dados!

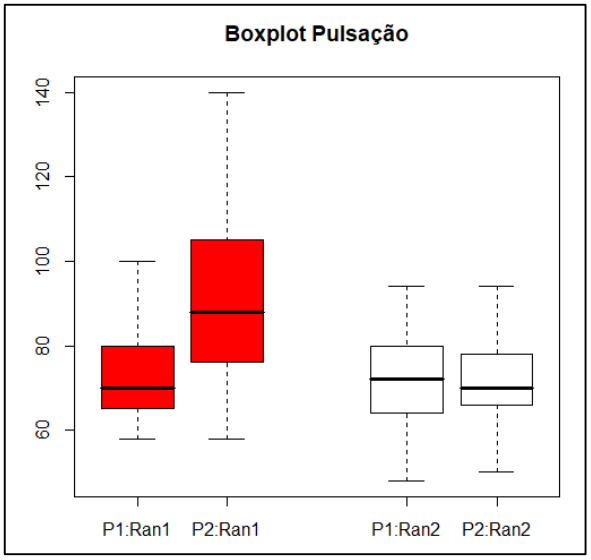
DADOS PULSE

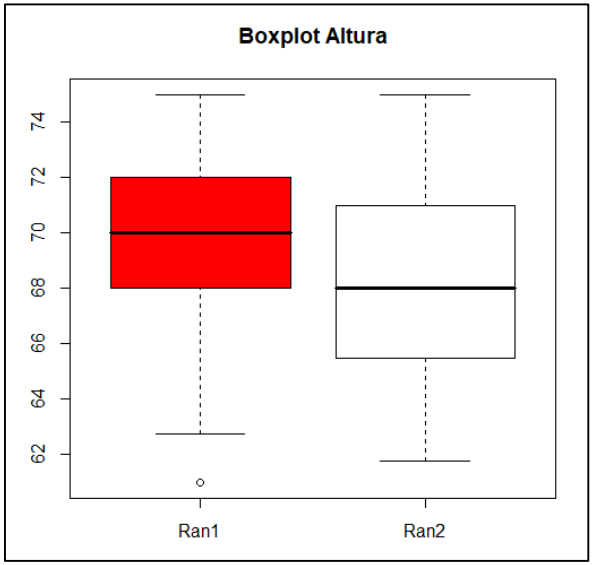
Var	Ran	n	mean	sd	med	min	max	range	se	cv
P1	1	35	73.60	11.44	70	58	100	42	1.93	0.16
	2	57	72.42	10.82	72	48	94	46	1.43	0.15
P2	1	35	92.51	18.94	88	58	140	82	3.20	0.21
	2	57	72.32	9.95	70	50	94	44	1.32	0.14

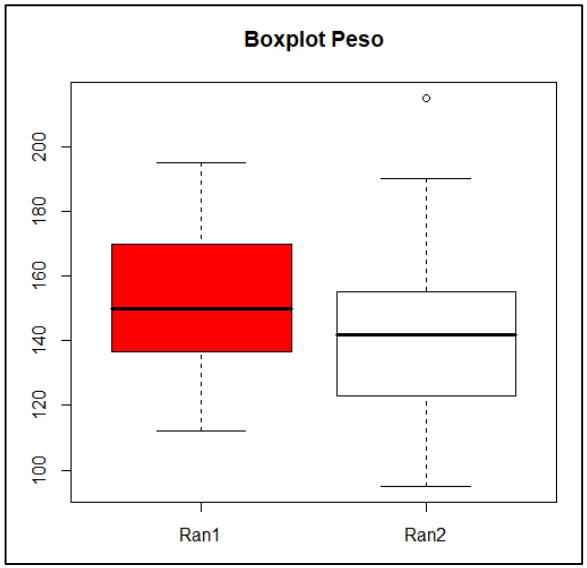
	Fu		Sex	
Ran	1	2	1	2
1	12(34%)	23(66%)	24(69%)	11(31%)
2	16(28%)	41(72%)	33(58%)	24(42%)

	Ativ			
Ran	1	2	3	4
1	0(0%)	3(9%)	25(71%)	7(20%)
2	1(2%)	6(10%)	36(63%)	14(25%)









Comparações de 2 Populações

DADOS PULSE

	P1	P2	Ran
1	64	88	1
2	58	70	1
…			
34	62	98	1
35	80	128	1
36	62	62	2
37	60	62	2
…			
91	86	84	2
92	76	76	2

Objetivo do Estudo: Há efeito da Corrida na pulsação dos estudantes?

Como os 2 grupos (Ran=1 e Ran=2) podem ser comparados?

Comparação de Médias

- **Dados Pareados: Ran=1 (n=35)**

Comparar as médias de pulsação Antes (P1) e Depois (P2) da corrida
(usar Ran=2 como Controle negativo)

- **Dados Independentes: Coluna P2 (n=92=35+57)**

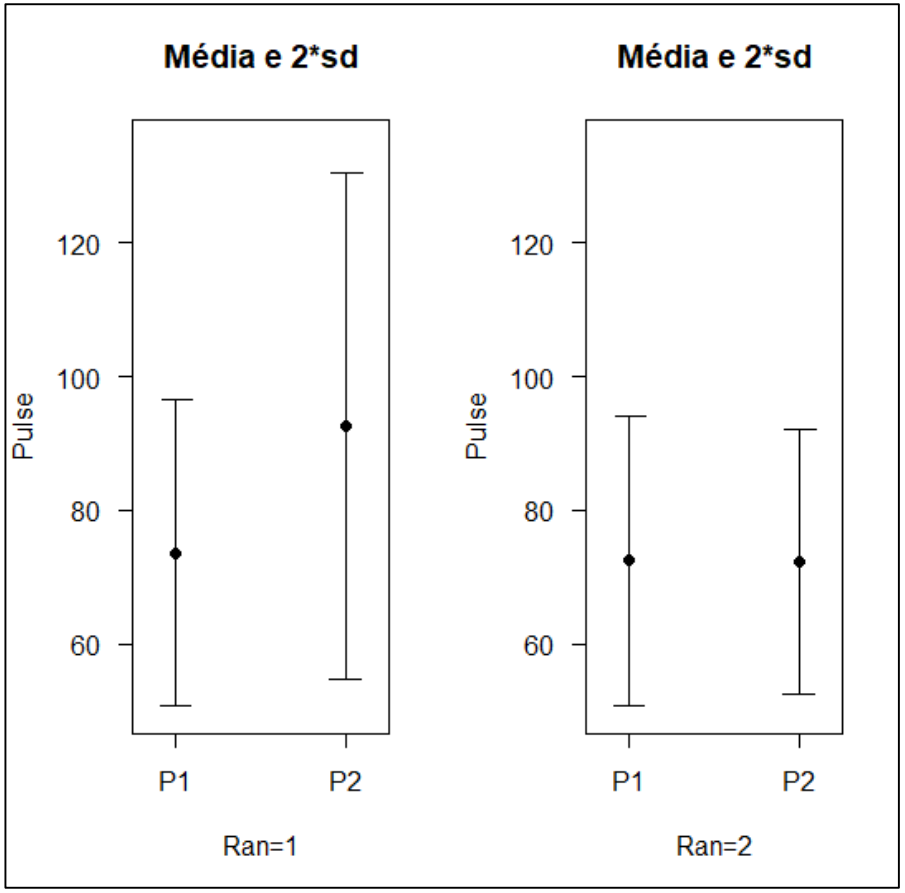
Comparar as médias de pulsação Antes (P1) e Depois (P2) da corrida
(usar P1 como controle negative)

Alternativa: Comparação de Médias (em P2) ajustada por P1

Análise Descritiva e Inferencial

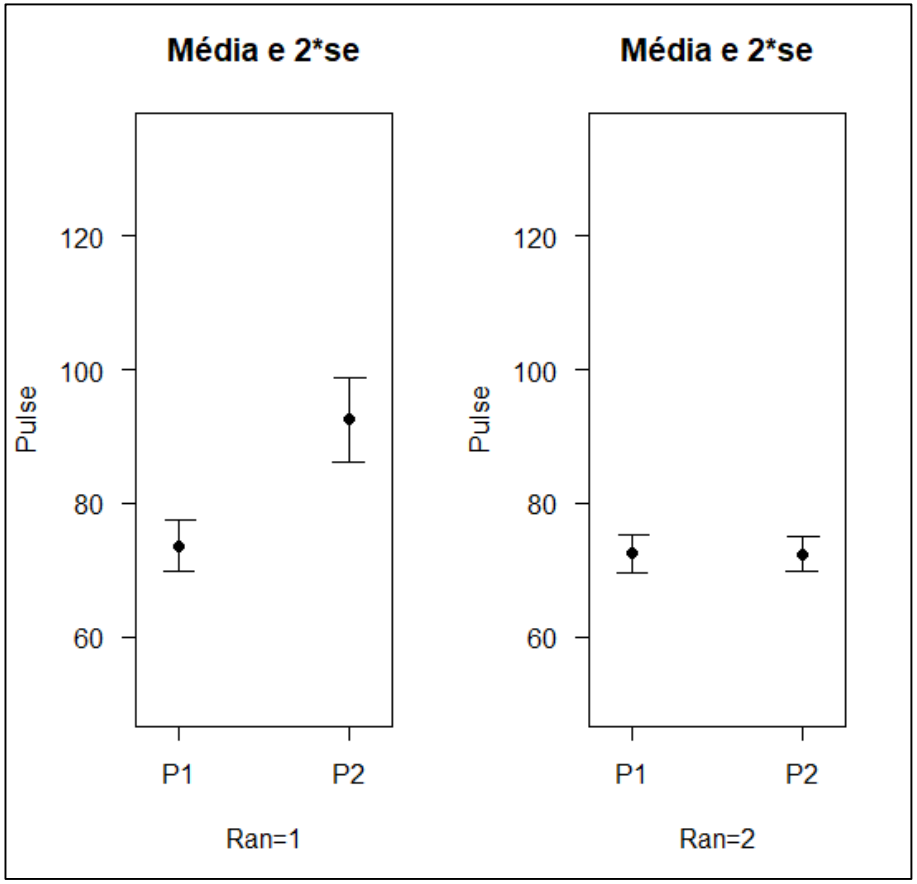
Entenda estes resultados!

Intervalo de Concentração dos Dados (Y)



$$P\left(\bar{Y} - 2s \leq Y \leq \bar{Y} + 2s\right) = 95\%$$

Intervalo de Confiança para a verdadeira Pulsação Média (μ)



$$IC95\%(\mu) = \left(\bar{Y} - 2\frac{s}{\sqrt{n}}; \bar{Y} + 2\frac{s}{\sqrt{n}}\right)$$

Variáveis Aleatórias

- Na pesquisa científica, os Estudos que realizamos são **Experimentos Aleatórios**
- No Experimento Aleatório, as variáveis (desfechos de interesse) a serem avaliadas nos indivíduos do estudo são **Variáveis Aleatórias**

Y é variável aleatória: Pulsação P1, P2, Ran, Sexo, Tabagismo, Altura, Peso, Nível de Atividade Física são exemplos de Variáveis Aleatórias

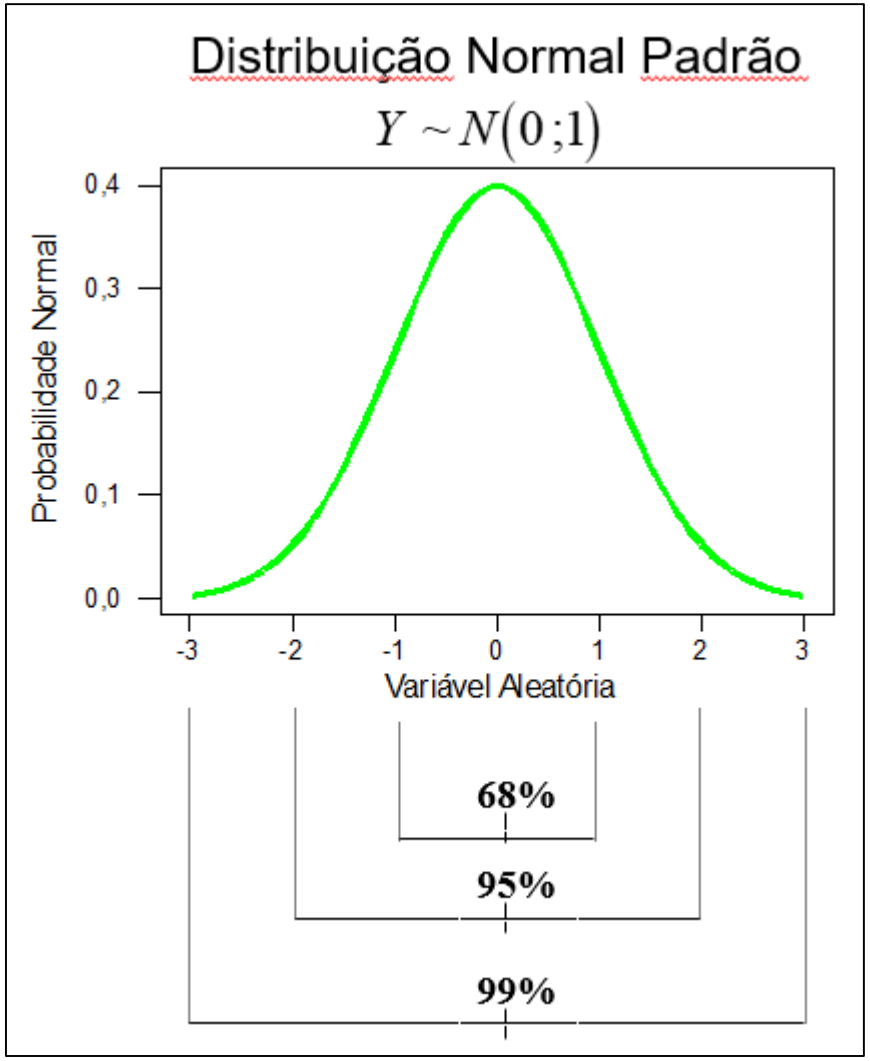
A **Amostra** efetivamente observada na realização do Estudo é um conjunto de valores da variável aleatória para os indivíduos amostrados da População sob investigação

Indivíduo 1	**Indivíduo 2**	**…**	**Indivíduo n**
$Y=y_1$	$Y=y_2$		$Y=y_n$

Amostra: $\{y_1, y_2, \ldots, y_n\}$

Dados Pulse:
P1= {64, 58, …, 76}
P2= {88, 70, …, 76}

Intervalos de Concentração dos Dados



Sob: $Y \sim N\left(\mu;\sigma^2\right)$

$$P\left(\mu-\sigma \le Y \le \mu+\sigma\right)=0{,}68$$

$$P\left(\mu-2\sigma \le Y \le \mu+2\sigma\right)=0{,}95$$

$$P\left(\mu-3\sigma \le Y \le \mu+3\sigma\right)=0{,}997$$

...

$$P\left(\mu-6\sigma \le Y \le \mu+6\sigma\right)\cong 1$$

Na prática:

Considere uma Amostra de n observações (aproximadamente Normal) de Y. É esperado que o interval

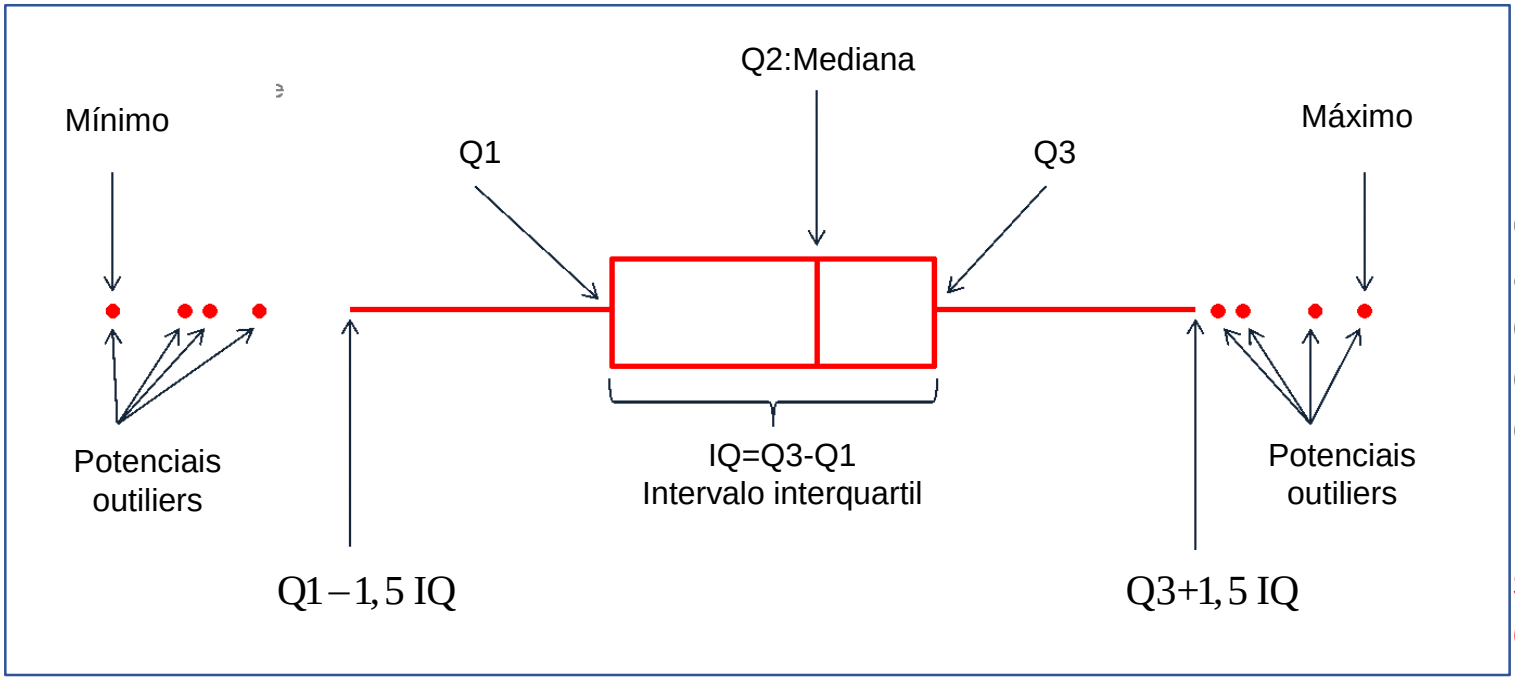
$$\left(\bar{Y}-2s\ ;\bar{Y}+2s\right)$$

concentre 95% dos valores da variável aleatória Y

Intervalos de Concentração dos Dados

Box-Plot: Sumariza a distribuição de uma variável contínua por meio de 5 valores

Valor mínimo	Q1: primeiro quartil 25º percentil	Q2: segundo quartil 50º percentil	Q3: terceiro quartil 75º percentil	Valor máximo



Representação útil para diagnóstico de observações atípicas (potenciais outliers) que podem ocorrer em um conjunto de dados.

Potenciais outliers são pontos que ocorrem abaixo de Li ou acima de Ls:
Li=Q1-1,5(Q3-Q1)
Ls=Q3+1,5(Q3-Q1)

IQ = Q3 – Q1: contém 50% das observações que ocupam o centro da distribuição

$$P(Y \leq Q1 - 1,5IQ) = 0,0034883 = P(Y \geq Q3 + 1,5IQ)$$

$$P(Q1 - 1,5IQ \leq Y \leq Q3 + 1,5IQ) = 0.9930234$$

Teorema Limite Central

Y: Variável Aleatória

Momentos Finitos: $E(Y) = \mu$ $V(Y) = \sigma^2$

$\{y_1, y_2, ..., y_n\}$ é Amostra Aleatória $\Rightarrow$ $\bar{Y}$ a média da amostra

Distribuição Amostral da Média

$$\bar{Y} \overset{n \to \infty}{\sim} N(\mu ; \sigma^2/n)$$

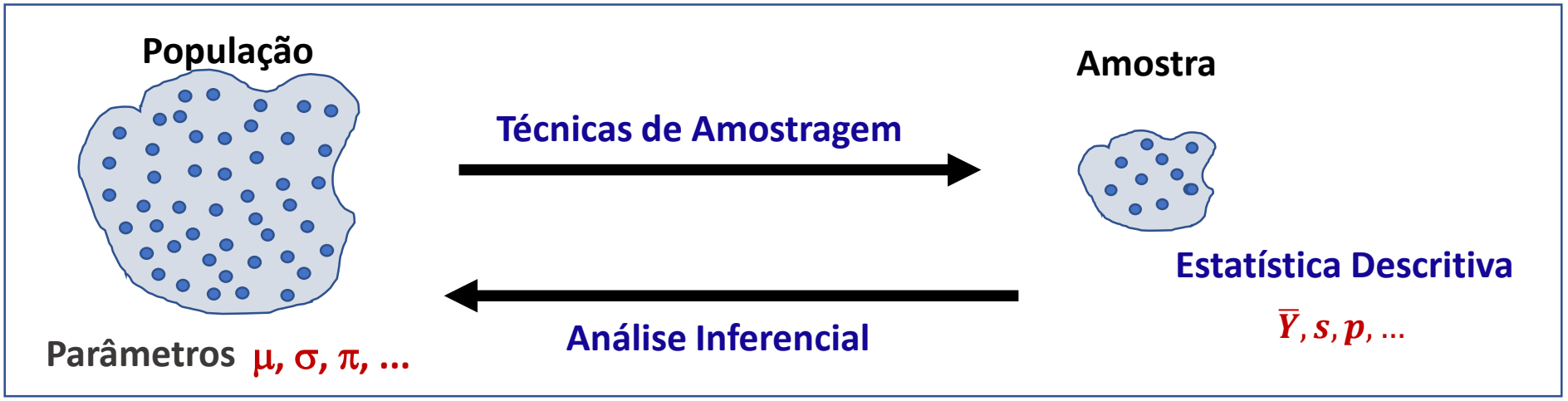
Normal

$E(\bar{Y}) = \mu$ $V(\bar{Y}) = \sigma^2/n$

A média amostral é Estimador não viciado de μ

Ganho em precisão

Distribuição Amostral de Estimadores



Para inferir resultados ***da*** Amostra (efetivamente observada) ***para*** a População é preciso considerar a **variação do Estimador de Amostra para Amostra**:

"Distribuição Amostral do Estimador do Parâmetro de interesse"

Amostra 1	Amostra 2	...	Amostra 100 ...
$\bar{Y}_1$	$\bar{Y}_2$		$\bar{Y}_{100}$
s_1	s_2		s_{100}
p_1	p_2		p_{100}

Suponha que o Experimento é realizado muitas vezes (10, 100 ou mais vezes). Em cada realização do Experimento uma amostra de tamanho n é obtida. O ideal é quea variação nas estimativas (ou erro amostral) seja pequena!

Distribuição Amostral da Média

$$\bar{Y} \overset{n\to\infty}{\sim} N(\mu ; \sigma^2/n)$$

Alguns resultados

Estatística z

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0 ; 1)$$

Estatística t

$$t = \frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{(n-1)}$$

Intervalos de Confiança para a verdadeira média μ

$$\left(\bar{Y} - z_{\alpha/2} * \sigma/\sqrt{n} ; \bar{Y} + z_{\alpha/2} * \sigma/\sqrt{n} \right)$$

$$\left(\bar{Y} - t_{(n-1),\alpha/2} * s/\sqrt{n} ; \bar{Y} + t_{(n-1),\alpha/2} * s/\sqrt{n} \right)$$

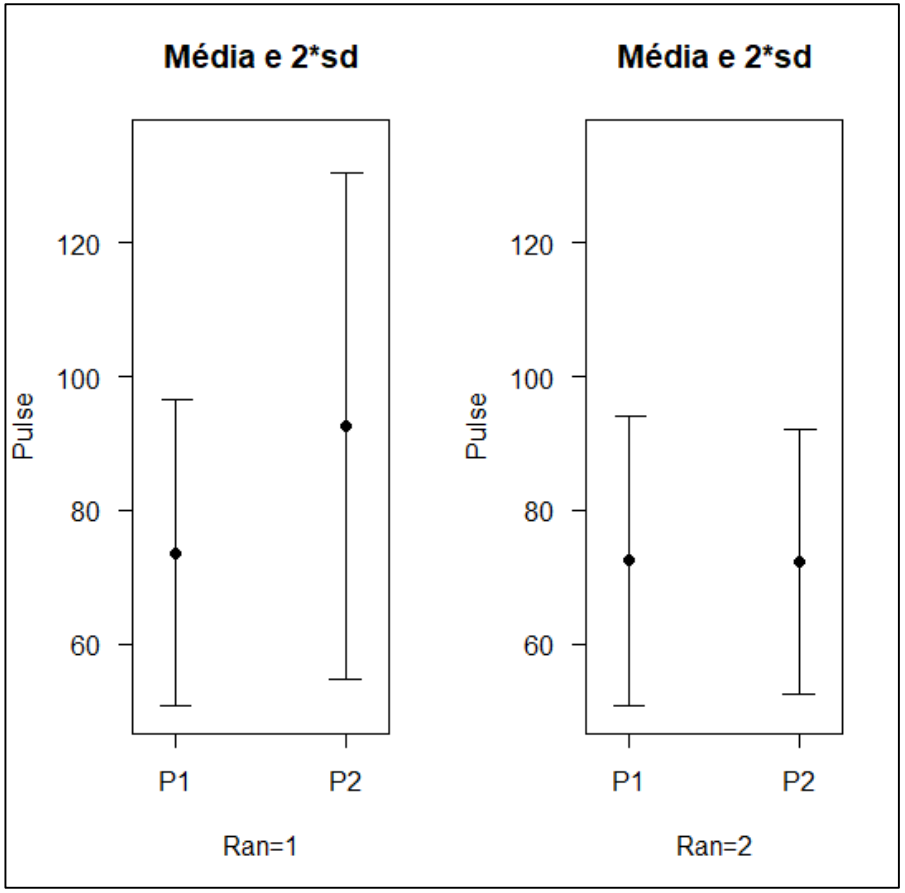
Para n "grande" são equivalentes

$$IC95\%(\mu) \overset{n\to\infty}{=} \left(\bar{Y} - 2\frac{s}{\sqrt{n}} ; \bar{Y} + 2\frac{s}{\sqrt{n}} \right)$$

Análise Descritiva e Inferencial

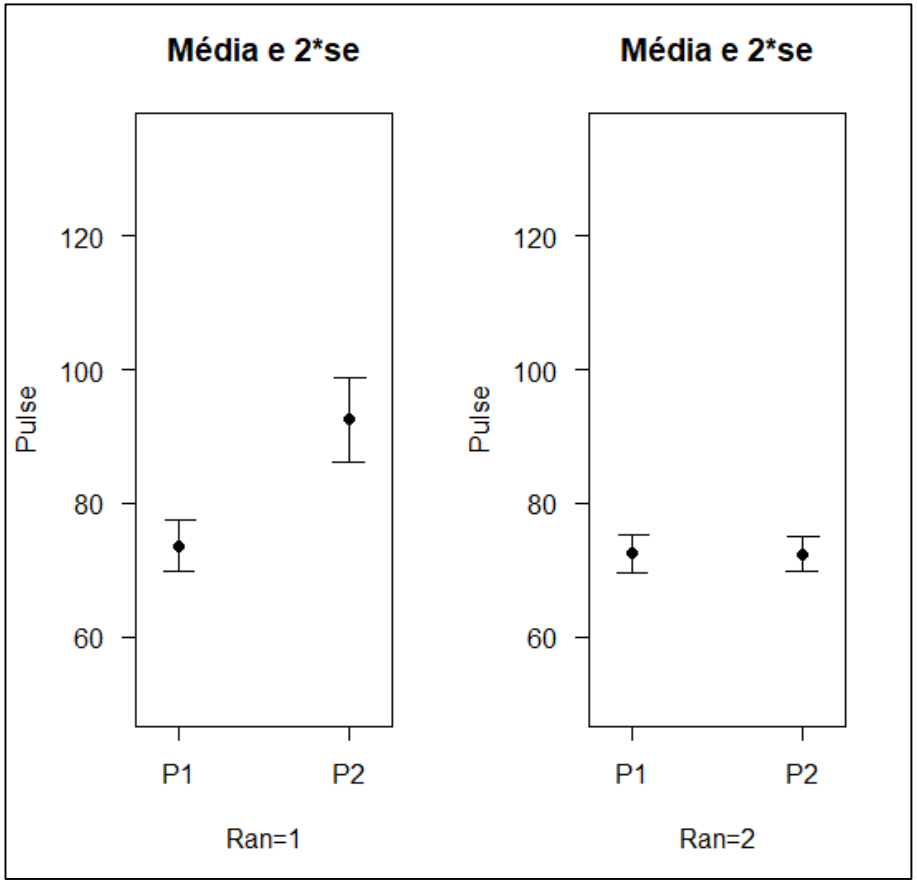
Entenda estes resultados!

Intervalo de Concentração dos Dados (Y)



$$P\left(\bar{Y} - 2s \leq Y \leq \bar{Y} + 2s\right) = 95\%$$

Intervalo de Confiança para a verdadeira Pulsação Média (μ)



$$IC95\%(\mu) = \left(\bar{Y} - 2\frac{s}{\sqrt{n}}; \bar{Y} + 2\frac{s}{\sqrt{n}}\right)$$

$$(P1_i; P2_i) \rightarrow d_i = P2_i - P1_i$$

	Ran	P1	P2	d
[1,]	1	64	88	24
[2,]	1	58	70	12
[3,]	1	62	76	14
[4,]	1	66	78	12
[5,]	1	64	80	16
...				
[29,]	1	100	115	15
[30,]	1	68	112	44
[31,]	1	96	116	20
[32,]	1	78	118	40
[33,]	1	88	110	22
[34,]	1	62	98	36
[35,]	1	80	128	48
[36,]	2	62	62	0
[37,]	2	60	62	2
[38,]	2	72	74	2
[39,]	2	62	66	4
[40,]	2	76	76	0
...				
[86,]	2	76	76	0
[87,]	2	87	84	-3
[88,]	2	90	92	2
[89,]	2	78	80	2
[90,]	2	68	68	0
[91,]	2	86	84	-2
[92,]	2	76	76	0

Análise Descritiva e Inferencial

- **Há efeito da corrida na Pulsação?**

Vamos considerar os dados P1 e P2 para Ran=1.
Usar Ran=2 como Controle negativo.

$$d_i = P2_i - P1_i$$

Dados "pareados" permitem o cálculo de *diferenças*

$$\bar{d} = \bar{Y}_{P1} - \bar{Y}_{P2}; \quad s_d; \quad s_{\bar{d}} = s_d / \sqrt{n}$$

$$\left(\bar{d} \mp 2s_d\right)$$

Intervalo de Concentração dos Dados de Diferença (P2 – P1): avaliar o que é esperado do comportamento da variável

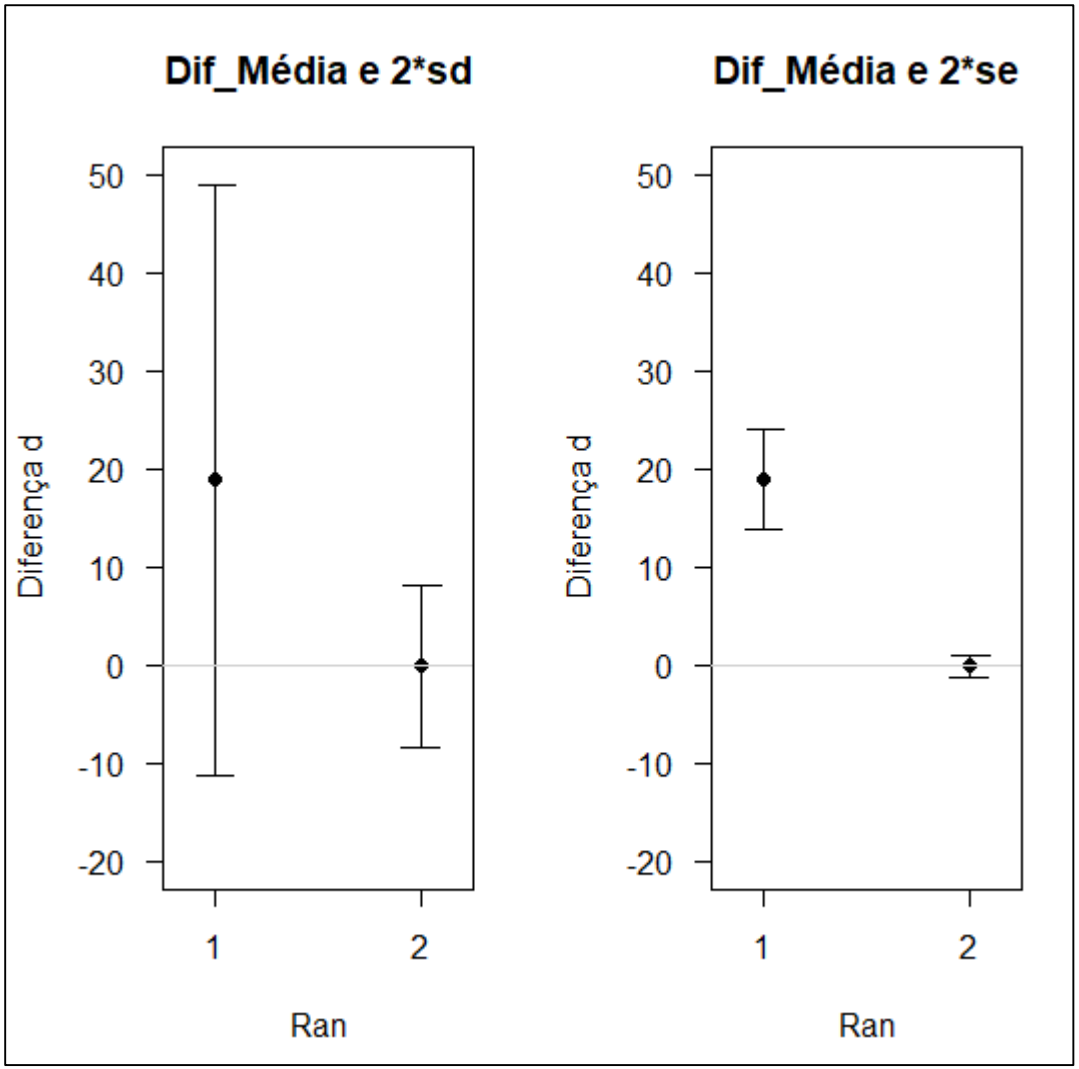
$$\left(\bar{d} \mp 2s_{\bar{d}}\right)$$

Intervalo de 95% de Confiança para a Verdadeira Diferença entre as Médias de P1 e P2

Análise Descritiva e Inferencial

Intervalo de Concentração dos Dados de Diferença (P2 – P1)

Intervalo de Confiança para a Diferença entre as Médias



Para o grupo que não correu (Ran2) o **IC INCLUI o "0"**, indicando que **não há evidência de diferença** entre as médias de P1 e P2 na população de estudantes (em repouso).

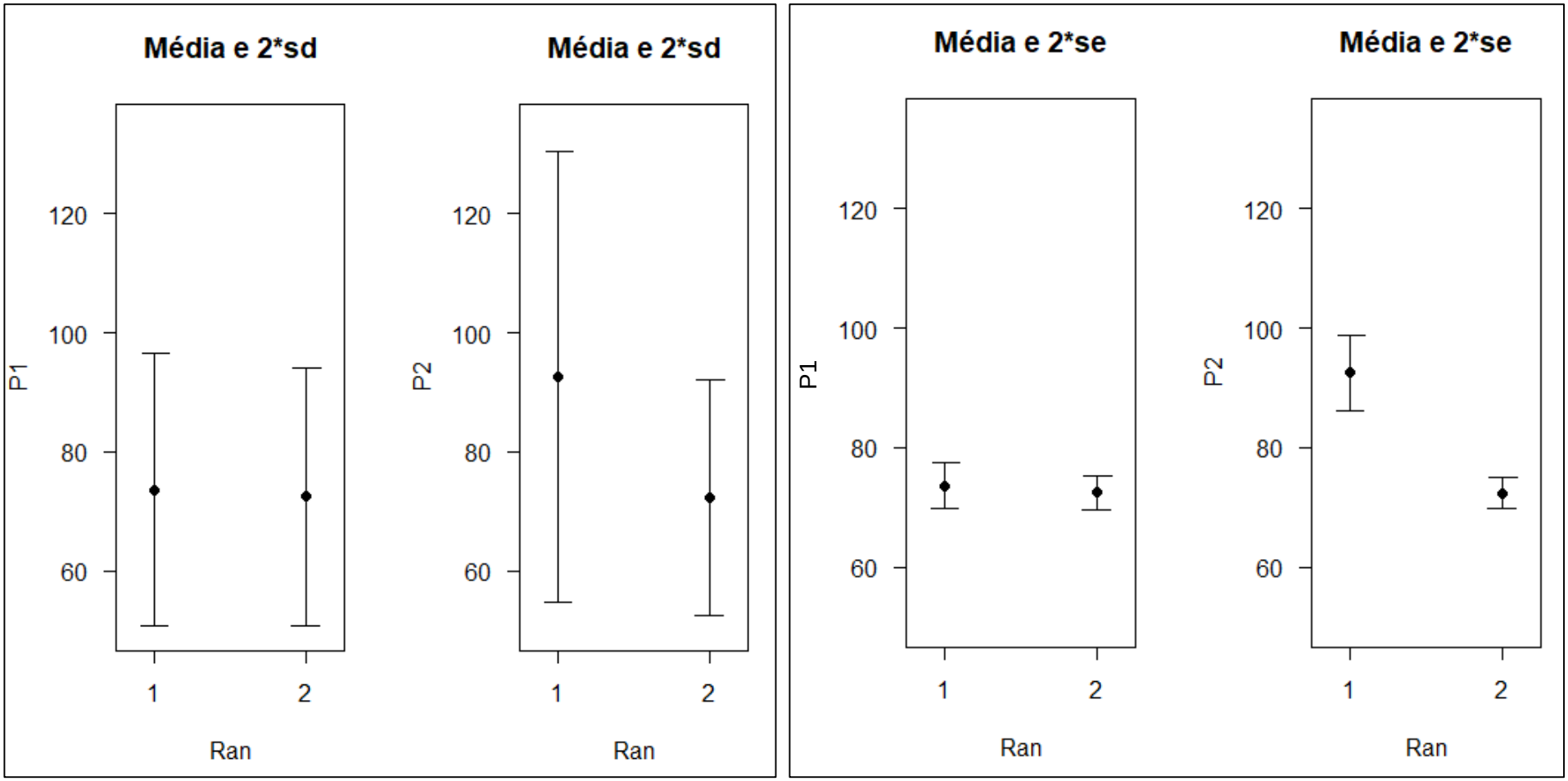
Para o grupo que correu (Ran1) o **IC NÃO INCLUI** o "0", indicando que **há evidência de diferença significante** entre as médias de P1 e P2 na população de estudantes, devido ao efeito da corrida.

Análise Descritiva e Inferencial

Entenda a diferença entre Experimentos com amostras pareadas e independentes

Comparação entre Populações (Ran1 e Ran2)

Adotar Amostras Independentes



Inferência sobre Diferenças entre Médias de Duas Populações

Intervalos de Confiança e Testes "t"

Observações Pareadas: $(Y_{1i}\,;Y_{2i}) \qquad i=1,2,...,n$

$$\Rightarrow d_i = Y_{1i} - Y_{2i} \quad \sim \quad N\left(\mu_d = \mu_1 - \mu_2; \sigma_d^2\right), \quad i=1,2,...,n$$

$$\Rightarrow IC(1-\alpha)100\% \text{ para } \mu_d = \left[\bar{d} - e; \bar{d} + e\right]; \quad e = t_{(n-1)}(1-\alpha/2)\frac{s_d}{\sqrt{n}}$$

$$\Rightarrow \begin{cases} H_0 : \mu_d = 0 \\ H_1 : \mu_d \neq 0 \end{cases} \quad \Rightarrow \quad t = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}; \quad \Rightarrow \quad t^2 = \frac{\bar{d}^2}{s_d^2/n} \sim F_{1,n-1}$$

Observações Independentes: $Y_{11}, Y_{12}, ..., Y_{1n_1} \sim N_1\left(\mu_1; {\sigma_1}^2\right); \quad Y_{21}, Y_{22}, ..., Y_{2n_2} \sim N_1\left(\mu_2; {\sigma_2}^2\right)$

$$\Rightarrow \bar{d} = \bar{Y}_1 - \bar{Y}_2 \quad \sim \quad N\left(\mu_d = \mu_1 - \mu_2; \sigma_d^2 = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right); \quad \sigma_1^2 = \sigma_2^2 = \sigma^2$$

$$\Rightarrow IC(1-\alpha)100\% \text{ para } \mu_d = \left[\bar{d} - e; \bar{d} + e\right]; \quad e = t_{(n_1+n_1-2)}(1-\alpha/2)s_c\left(\frac{1}{n_1} + \frac{1}{n_2}\right); \quad s_c^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

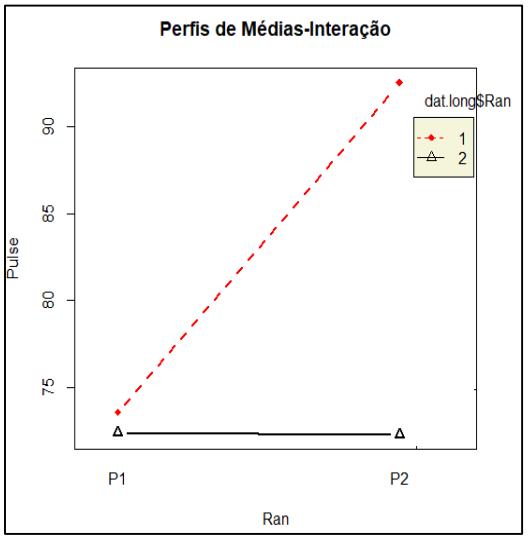
$$\Rightarrow \begin{cases} H_0 : \mu_D = 0 \\ H_1 : \mu_D \neq 0 \end{cases} ; \sigma_1^2 = \sigma_2^2 \quad \Rightarrow t = \frac{\bar{Y}_1 - \bar{Y}}{s_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}; \quad t^2 = \frac{\bar{Y}_1 - \bar{Y}_2}{s_c^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \sim F_{1;(n_1+n_2-2)}$$

Inferência sobre Diferenças entre Médias de Duas Populações

Dados Pulse - **Observações Pareadas**: Comparar P1 e P2 (Antes e Depois da Corrida), para cada Grupo Ran1 (n=35 observações pareadas e Ran2 (n=57 observações pareadas)

```
Teste t Bicaudal - Ran == 1
t = 7.4353, df = 34, p-value = 1.265e-08
IC95%: 13.74454 24.08403
Média da diferença: 18.91429


Teste t Bicaudal - Ran == 2
t = -0.19101, df = 56, p-value = 0.8492
IC95%: -1.2092048  0.9986785
Média da diferença: -0.1052632
```



Conclusão. Comparando as Médias de pulsação antes e depois da corrida (P1 e P2): Ran1: há diferença significante a 5% ($\mu_{P1} < \mu_{P2}$), isto é, há efeito da corrida (valor-p<0.05 $\Rightarrow$ Rejeitar H0; o IC95% não inclui o valor 0)

Ran2: não há evidência para diferença significante (valor-p>0.05 $\Rightarrow$ Não rejeitar H0; o IC95% inclui o 0)

Inferência sobre Diferenças entre Médias de Duas Populações

Dados Pulse - **Observações Independentes**: Comparar as médias de **P2** entre os grupos Ran=1 e Ran=2 (35 observações independentes das demais 57 observações). Repita para P1

Análise de P2

Teste de Bartlett para H0: homogeneidade das variâncias dos grupos

Bartlett's K-squared = 18.01, df = 1, **p-value = 2.198e-05**

Conclusão: Rejeitar H0. Grupos com variâncias heterogêneas

Teste t (Welch) Bicaudal, Amostras independentes, Variâncias Heterogêneas

t = 5.8335, df = 45.695, **p-value = 5.251e-07**

IC95% para a diferença entre Médias: 13.22755 27.16944

Médias dos Grupos: Ran1=92.51429 Ran2=72.31579

Variâncias: $s_1^2 = 18.94^2$ $s_2^2 = 9.95^2$

Análise de P1

Teste de Bartlett para H0: homogeneidade das variâncias dos grupos

Bartlett's K-squared = 0.13067, df = 1, **p-value = 0.7177**

Conclusão: Não Rejeitar H0. Grupos com variâncias homoggêneas

Teste t Bicaudal, Amostras independentes, Variâncias Homogêneas

t = 0.49663, df = 90, **p-value = 0.6207**

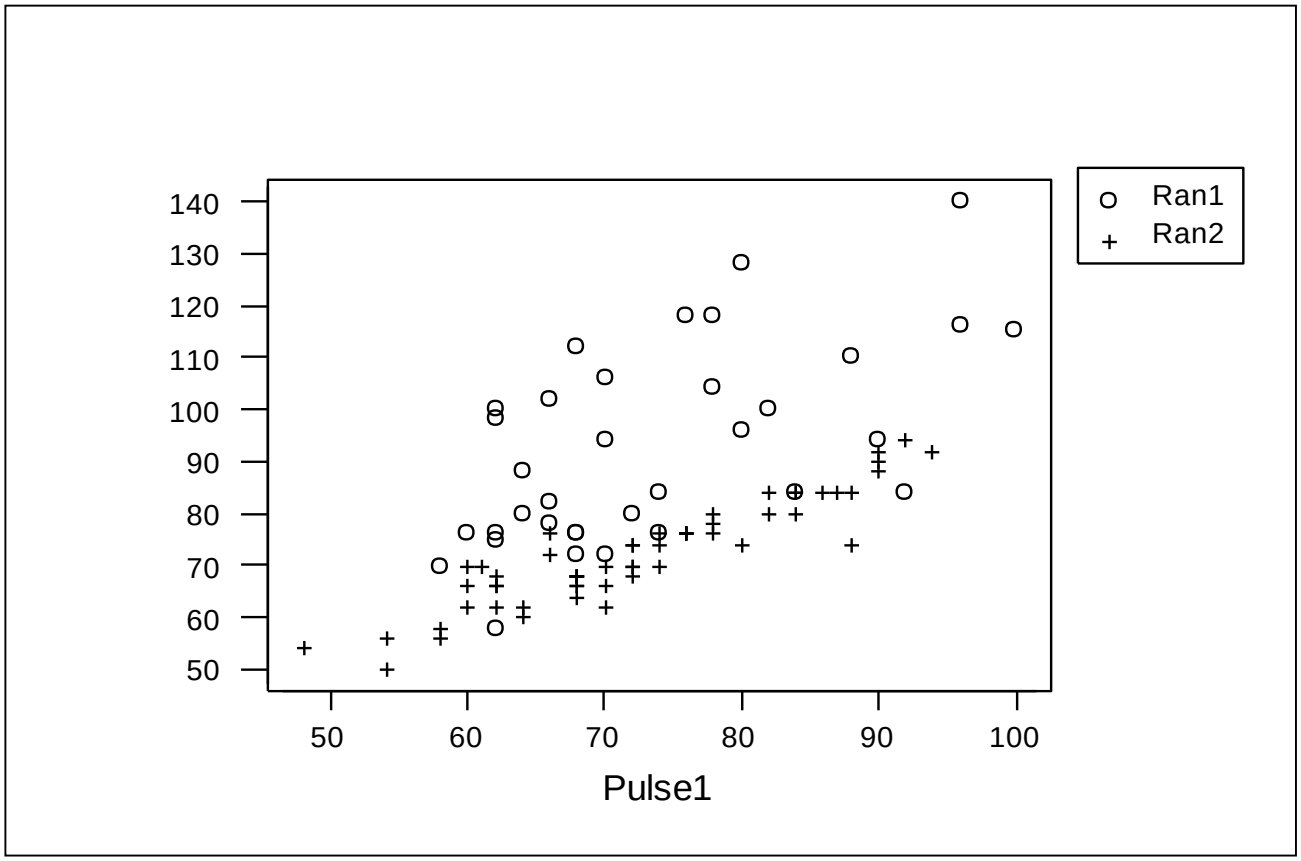
IC95% para a diferença entre Médias: -3.537235 5.895130

Médias dos Grupos: Ran1= 73.60000 Ran2= 72.42105

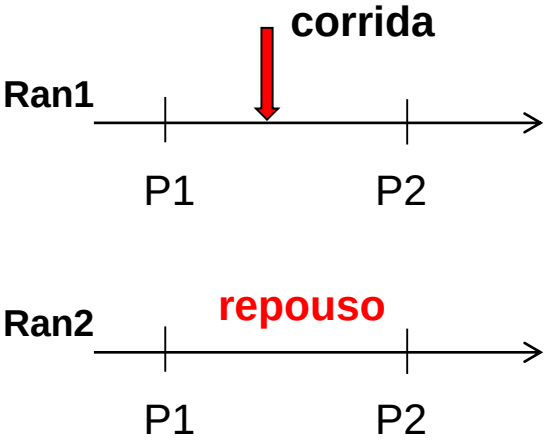
Variâncias: $s_1^2 = 11.44^2$ $s_2^2 = 10.82^2$ $s_c = 11.05456$

Modelos mais Gerais

$$Y = f(X) + e$$



Dados Pulse: a pulsação de 92 estudantes foi avaliada Antes de Depois de uma intervenção (corrida).
Ran1: corrida
Ran2: repouso



Pense em possíveis modelos para avaliar o efeito da corrida na Pulsação!

Modelos Estatísticos

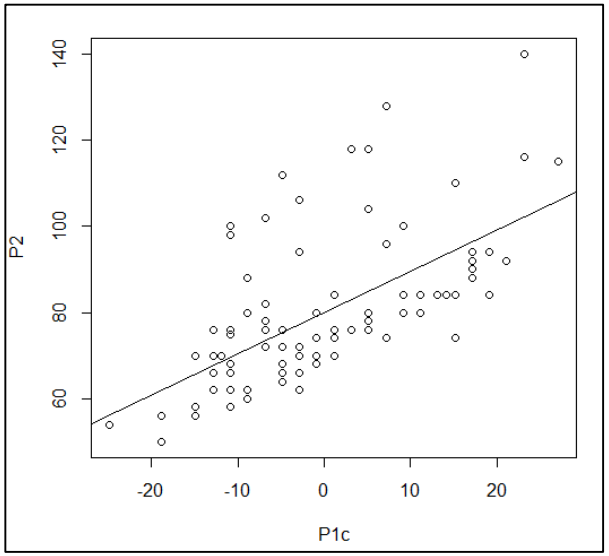
- **M1:** $Y = \beta_0 + \beta_1 X_1 + e$ Y = P2; X1=(P1-Média de P1)
- **M2:** $Y = \beta_0 + \beta_2 X_2 + e$ X2 = 0 se em repouso (Ran=2); = 1 se correu (Ran=1)
- **M3:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$
- **M4:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 \left(X_1 * X_2 \right) + e$

Em cada modelo, qual é o valor esperado de Y (P2) para estudantes em repouso e que correram?

Modelos mais Gerais

- **M1:** $Y = \beta_0 + \beta_1 X_1 + e$

Y = P2
X1=(P1-Média de P1)



Suposição: $e_i \overset{iid}{\sim} N\left(0;\sigma^2\right)$

$$\Rightarrow Y_i \overset{iid}{\sim} N\left(\beta_0 + \beta_1 X_{1i};\sigma^2\right)$$

Resultados do Ajuste: Analysis of Variance Table

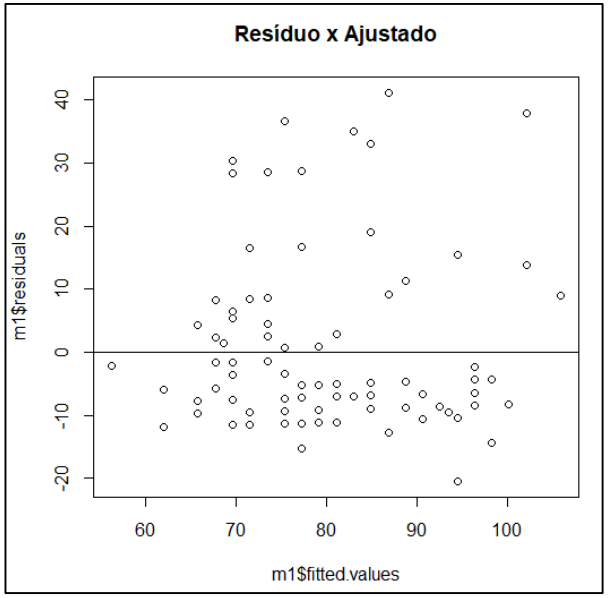
```
Response: P2
           Df Sum Sq Mean Sq F value    Pr(>F)
P1c         1  10096 10096.1   55.09 6.218e-11 ***
Residuals  90  16494   183.3
```

Coefficients:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.0000     1.4114  56.682  < 2e-16 ***
P1c           0.9568     0.1289   7.422 6.22e-11 ***
```

```
Residual standard error: 13.54 on 90 degrees of freedom
Multiple R-squared:  0.3797, Adjusted R-squared:  0.3728
F-statistic: 55.09 on 1 and 90 DF,  p-value: 6.218e-11
```

Conclusão:



Modelos mais Gerais

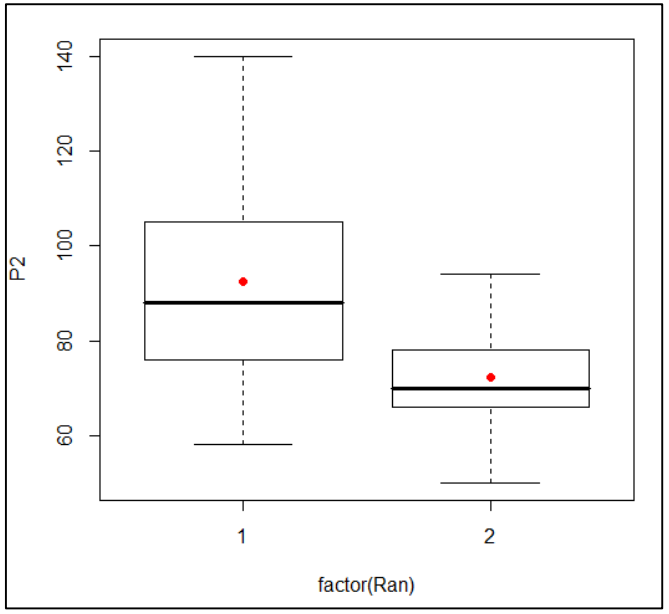
Equivalente ao teste t (amostras independentes, homocedasticidade

- **M2:** $Y = \beta_0 + \beta_2 X_2 + e$

X2=0: em repouso X2=1correu

Suposição: $e_i \overset{iid}{\sim} N\left(0;\sigma^2\right)$

$$\Rightarrow Y_i \overset{iid}{\sim} N\left(\beta_0 + \beta_2 X_{2i};\sigma^2\right)$$



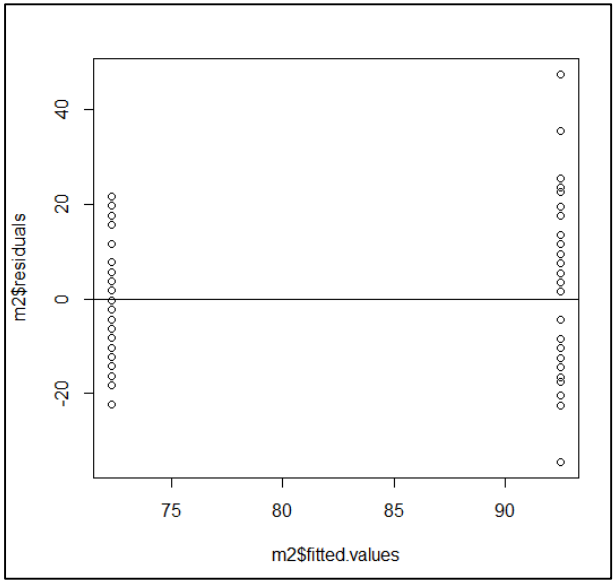
Resultados do Ajuste - Analysis of Variance Table

```
          Df  Sum Sq Mean Sq F value    Pr(>F)
Ran        1  8846.9  8846.9  44.875 1.768e-09 ***
Residuals 90 17743.1   197.1
```

Coefficients:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  112.713      5.098  22.109  < 2e-16 ***
Ran          -20.198      3.015  -6.699 1.77e-09 ***

Residual standard error: 14.04 on 90 degrees of freedom
Multiple R-squared: 0.3327,  Adjusted R-squared:  0.3253
F-statistic: 44.88 on 1 and 90 DF,  p-value: 1.768e-09
```



Conclusão:

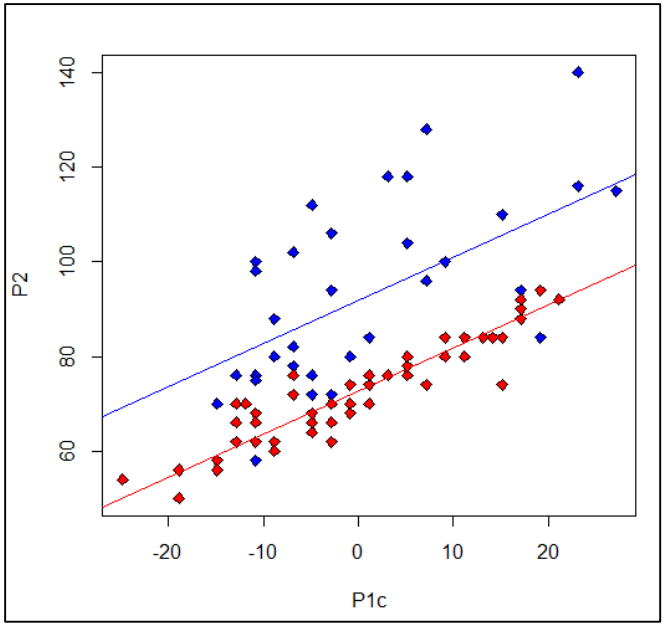
Modelos mais Gerais

Equivalente à comparação de médias de P2 (teste t) ajustada por P1

- **M3:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e$

Suposição: $e_i \overset{iid}{\sim} N\left(0; \sigma^2\right)$

$$Y_i \overset{iid}{\sim} N\left(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}; \sigma^2\right)$$



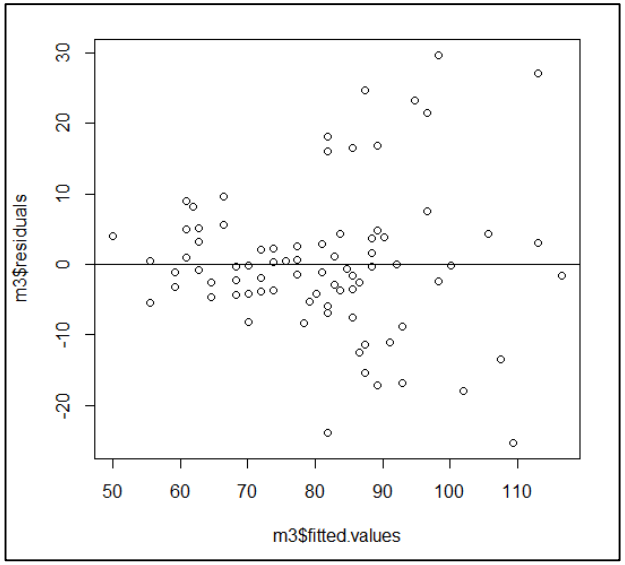
Resultados do ajuste - Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
P1c	1	10096.1	10096.1	104.656	< 2.2e-16	***
Ran	1	7908.0	7908.0	81.974	2.905e-14	***
Residuals	89	8585.8	96.5			

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	72.72504	1.30162	55.873	< 2e-16	***
P1c	0.91247	0.09366	9.743	1.09e-15	***
Ran	19.12274	2.11209	9.054	2.90e-14	***

Residual standard error: 9.822 on 89 degrees of freedom
Multiple R-squared: 0.6771, Adjusted R-squared: 0.6698
F-statistic: 93.31 on 2 and 89 DF, p-value: < 2.2e-16

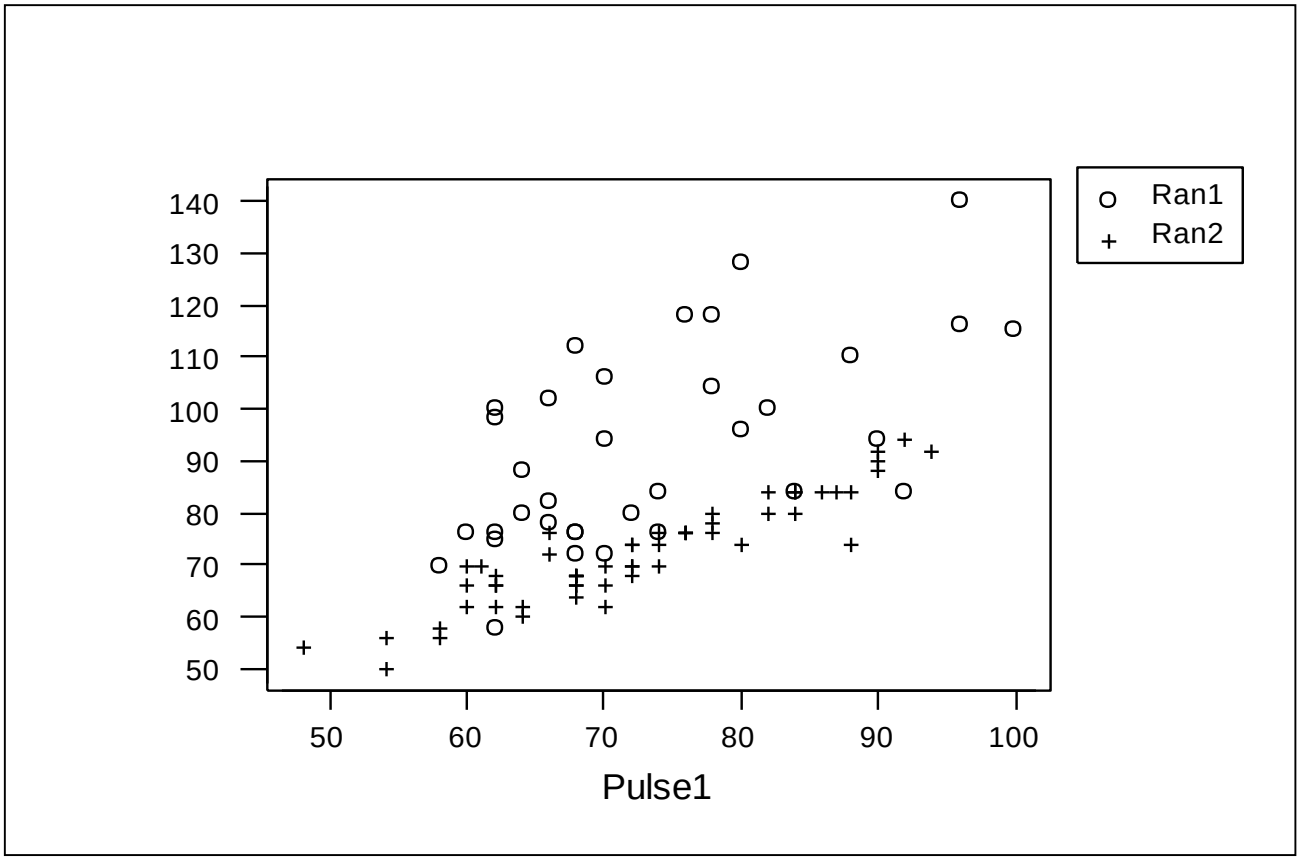


Conclusão:

Modelo Ajustado

P2 = 72,7 + 0,912 (P1-73) + 19,1 Ran

Codificação: 0 1



Em repouso: E(P2|P1,Ran) = 72,7 + 0,912 (P1-73)

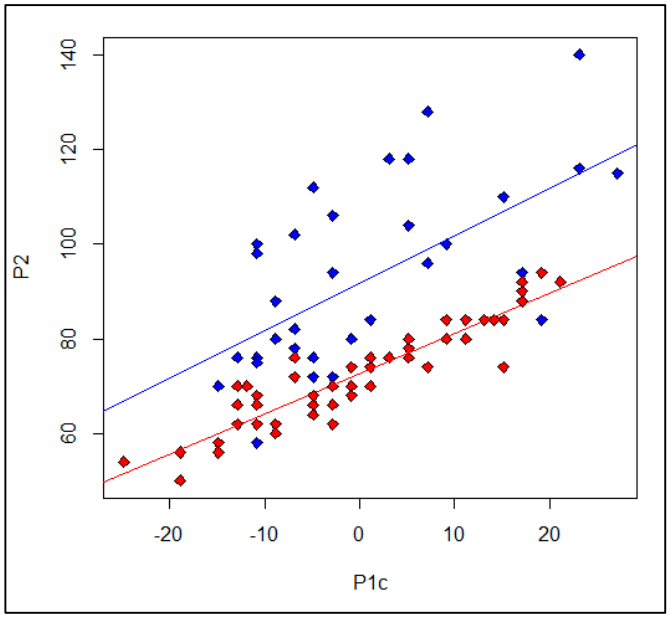
Corrida: E(P2|P1,Ran) = (72,7+19.1) + 0,912 (P1-73)

Modelos mais Gerais

- **M4:** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 * X_2) + e$

Suposição: $e_i \overset{iid}{\sim} N(0; \sigma^2) \Rightarrow$

$$Y_i \overset{iid}{\sim} N(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} * X_{2i}); \sigma^2)$$



Resultados do ajuste - Analysis of Variance Table

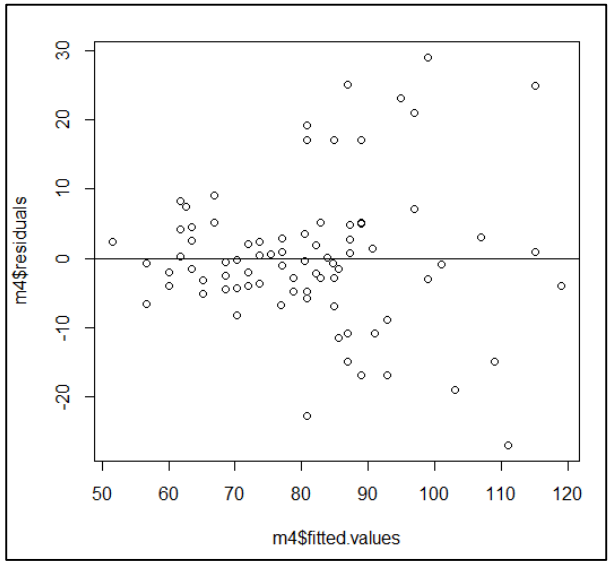
	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
P1c	1	10096.1	10096.1	104.2731	< 2.2e-16	***
Ran	1	7908.0	7908.0	81.6739	3.438e-14	***
P1cRan	1	65.3	65.3	0.6747	0.4136	
Residuals	88	8520.5	96.8			

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	72.6966	1.3045	55.729	< 2e-16	***
P1c	0.8490	0.1216	6.984	5.22e-10	***
Ran	19.0829	2.1165	9.016	3.80e-14	***
P1cRan	0.1570	0.1912	0.821	0.414	

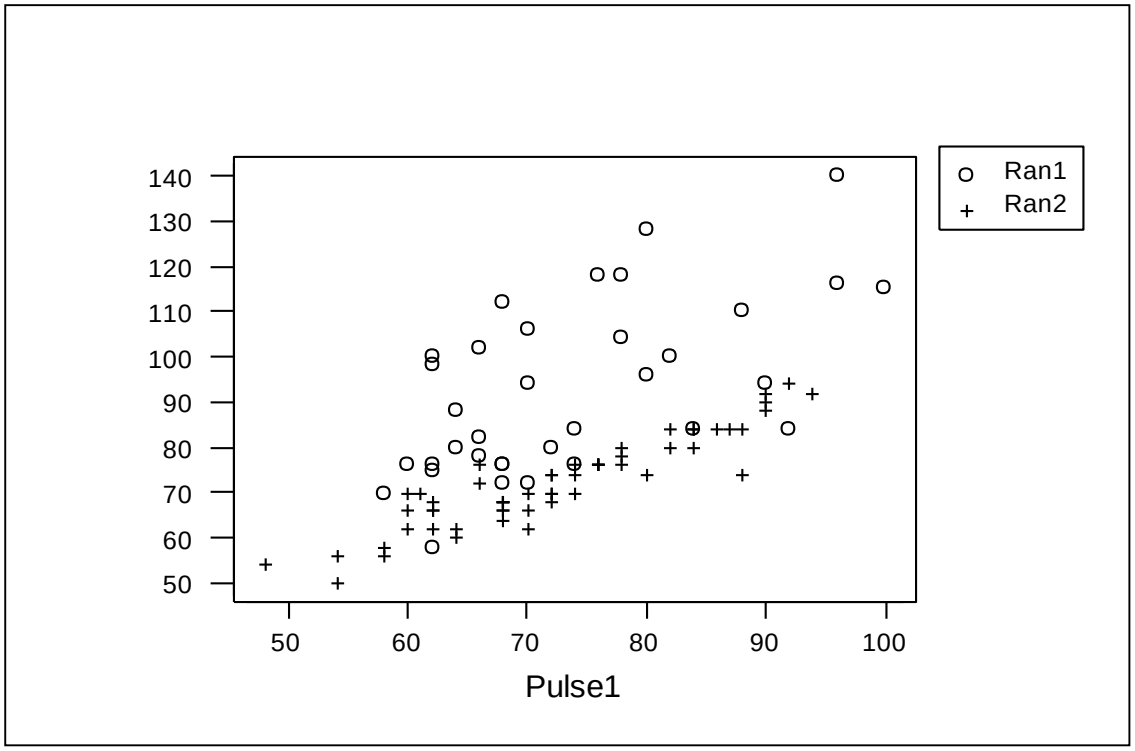
Residual standard error: 9.84 on 88 degrees of freedom
Multiple R-squared: 0.6796, Adjusted R-squared: 0.6686
F-statistic: 62.21 on 3 and 88 DF, p-value: < 2.2e-16

Conclusão:



Modelo Ajustado

P2 = 72,7+0,849(P1-73)+19,08 Ran + 0,157 P1c*Ran)



Em repouso:E(P2|P1,Ran,P1*Ran)= 72,7 + 0,849(P1-73)

Corrida:

E(P2|P1,Ran,P1*Ran)=(72,7+19.08) + (0,849+0,157)(P1-73)

Modelos Estatísticos

- **Intervalos de Confiança**
- **Testes de Hipóteses**

Caso de Duas Populações

- **Entender a estrutura dos dados**
- **Adotar um modelo estrutural e distribucional cujas suposições sejam válidas aos dados**
- **Realizar análises de diagnóstico**
- **Interpretar o ajuste**

Independência, normalidade, homocedasticidade

Próxima Aula

- **Testes de Aleatorização**

Take-away

Pontue o seu entendimento sobre os seguintes conteúdos considerados na Aula de IBI5086-170823:

- Estrutura de Dados
- Comparação de 2 Populações
- Análises descritivas e inferenciais
- Intervalos para os Dados e para a Média populacional
- Testes t (amostra pareada e independente)
- Modelos estatísticos, suposições, diagnóstico

Intervalos de Confiança

EXERCÍCIO. Suponha que:

- ✓ 100 experimentos independentes sejam realizados
- ✓ Que o verdadeiro valor da Média μ da resposta de interesse (Ex.: pulsação de estudantes em repouso) seja conhecida (igual a 72 batimentos/min)

$\Rightarrow$ Usando o aplicativo R, explore os recursos da `library` *`RcmdrPlugin.TeachingDemos`* e encontre intervalos de 95% de confiança para μ em amostras de tamanho 92. Adote um valor para σ.

$\Rightarrow$ Simule 100 destes experimentos

$\Rightarrow$ Em sua simulação, em quantos dos 100 experimentos o intervalo obtido incluiu o verdadeiro valor de μ?

$\Rightarrow$ Em quantos experimentos é esperado que o intervalo obtido inclua o verdadeiro valor de μ?