ME414 - Estatística para Experimentalistas

Parte 22

Análise de Variância (ANOVA)

Introdução

Já vimos anteriormente como testar se existe diferença entre duas médias μ_1 e μ_2 de duas populações independentes. Ou seja:

$$H_0 : \mu_1 = \mu_2 \quad \text{vs} \quad H_1 : \mu_1 \neq \mu_2$$

Considerando o caso das variâncias iguais e desconhecidas, usamos S_p^2 como estimador da variância σ^2 e temos a estatística do teste:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} \overset{H_0}{\sim} t_{n+m-2}.$$

Mas e se quiséssemos comparar as médias de 3 ou mais populações (grupos)?

Análise de Variância

Exemplo: O Departamento de Estatística oferece o curso ME414 todo semestre para várias turmas. Suponha que queremos saber se existe diferença significativa no desempenho na P1 entre as turmas A, B, C e I.

Poderíamos comparar as médias duas a duas, certo?

No entanto, isso não é muito viável quando temos muitos grupos.

A técnica estatística adequada para esse tipo de problema, com a qual pode-se comparar se as médias de várias populações (grupos) são todas iguais com um único teste é chamada de **Análise de Variância (ANOVA)**.

Análise de Variância - ANOVA

Objetivo: Comparar se as médias de 3 ou mais populações (grupos) são iguais.

Hipóteses:

$$
\begin{aligned}
H_0 &: \text{as médias são as mesmas para todos os grupos} \\
H_1 &: \text{pelo menos uma média é diferente das demais}
\end{aligned}
$$

Em termos estatísticos:

$$
\begin{aligned}
H_0 &: \mu_1 = \mu_2 = \ldots = \mu_k \\
H_1 &: \text{pelo menos uma média é diferente das demais}
\end{aligned}
$$

A **estatística do teste**, chamada de F, é conceitualmente o seguinte:

$$
F = \frac{\text{Variação entre as médias amostrais dos grupos}}{\text{Variação média dentro dos grupos}}
$$

ANOVA - Condições

Devemos checar três condições nos dados onde iremos realizar a ANOVA:

- as observações são independentes dentro dos grupos e entre os grupos;
- os dados dentro de cada grupo são aproximadamente normais; e
- a variância é aproximadamente constante entre os grupos.

Detalhes da ANOVA

Um conceito fundamental em Análise de Variância é que a variação total dos dados, considerando todas as amostras como vindas de uma única população, pode ser separadas em duas partes:

- variação devido às diferenças entre as médias dos grupos
- variação das observações dentro de cada grupo

Ou seja, escrevendo como uma equação:

$$\text{Variação Total} = \text{Variação Entre Grupos} + \text{Variação Dentro dos Grupos}$$

Iremos ver agora como medir cada uma dessas variações.

Estrutura dos Dados

Grupos	Observações	Média
Grupo 1	$X_{11}, X_{12}, X_{13}, \ldots, X_{1n}$	$\bar{X}_1$
Grupo 2	$X_{21}, X_{22}, X_{23}, \ldots, X_{2n}$	$\bar{X}_2$
$\vdots$	$\vdots$	$\vdots$
Grupo k	$X_{k1}, X_{k2}, X_{k3}, \ldots, X_{kn}$	$\bar{X}_k$

Veja que a média e variância amostral para cada grupo são calculadas como:

$$\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij} \quad \text{e} \quad s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2$$

Notação

Considere a seguinte notação:

k: número de populações ou grupos

n: tamanho de cada grupo

X_{ij}: a j-ésima observação dentro do i-ésimo grupo, $i = 1, \ldots, k$ e $j = 1, \ldots, n$

$\bar{X}_i$: média amostral do i-ésimo grupo

$\bar{X}$: média amostral considerando todas as observações como parte de um único grupo/população.

s_i: desvio padrão amostral do i-ésimo grupo

Variação Total

A variação total das observações é chamada de **Soma de Quadrados Total** ou SQ_T e é calculada como o numerador da variância amostral se todas as observações fossem combinadas em um único grupo. Ou seja,

$$SQ_T = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X})^2$$

Analiticamente pode-se mostrar que:

$$\begin{aligned} SQ_T = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X})^2 &= n\sum_{i=1}^{k}(\bar{X}_i - \bar{X})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2 \\ &= SQ_G + SQ_E \end{aligned}$$

Veremos agora o que são SQ_G e SQ_E.

Variação Entre Grupos

A variação entre as médias dos grupos é chamada de **Soma de Quadrados Entre Grupos** ou SQ_G e é calculada da seguinte forma:

$$SQ_G = n\sum_{i=1}^{k}(\bar{X}_i - \bar{X})^2 = n(\bar{X}_1 - \bar{X})^2 + \ldots + n(\bar{X}_k - \bar{X})^2$$

Veja que é a soma ponderada das diferenças entre as médias dos grupos $\bar{X}_i$ e a média geral $\bar{X}$ ao quadrado.

O numerador da estatística F é chamado de **Quadrado Médio Entre Grupos** ou QM_G e pode ser visto como sendo a variância amostral das médias dos grupos:

$$QM_G = \frac{SQ_G}{k-1}$$

Variação Dentro dos Grupos

A variação das observações dentro dos grupos é chamada de **Soma de Quadrados do Erro** ou SQ_E e é calculada da seguinte forma:

$$SQ_E = \sum_{i=1}^{k}\sum_{j=1}^{n}(X_{ij} - \bar{X}_i)^2 = \sum_{i=1}^{k}(n-1)s_i^2$$

Ou seja, é a soma ponderada das variâncias amostrais para o i-ésimo grupo.

O denominador da estatística F é chamado de **Quadrado Médio do Erro** ou QM_E e é a estimativa da variância populacional para k grupos:

$$QM_E = \frac{SQ_E}{k(n-1)} = \frac{(n-1)s_1^2 + \ldots + (n-1)s_k^2}{kn - k}$$

Teste de Igualdade das Médias para k Grupos

Resumindo, estamos interessados em testar as hipóteses:

$$
\begin{aligned}
H_0 &: \mu_1 = \mu_2 = \ldots = \mu_k \\
H_1 &: \text{pelo menos uma média é diferente das demais}
\end{aligned}
$$

A **estatística do teste** é dada por:

$$F = \frac{QM_G}{QM_E} = \frac{\frac{SQ_G}{k-1}}{\frac{SQ_E}{k(n-1)}}$$

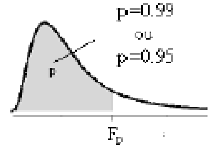
Sob a hipótese H_0 de igualdade das médias, a estatística do teste segue uma distribuição F com $k-1$ graus de liberdade no númerador e $k(n-1)$ graus de liberdade no denominador. Ou seja,

$$F \overset{H_0}{\sim} F_{k-1,k(n-1)}$$

Tabela F

Os valores críticos da distribuição F para $\alpha = 0.05$ ou $\alpha = 0.01$ estão na tabela abaixo. As linhas e colunas representam os graus de liberdade do denominador (ν_2) e numerador (ν_1), respectivamente.

Tabela III: Distribuição F de Fischer-Snedecor



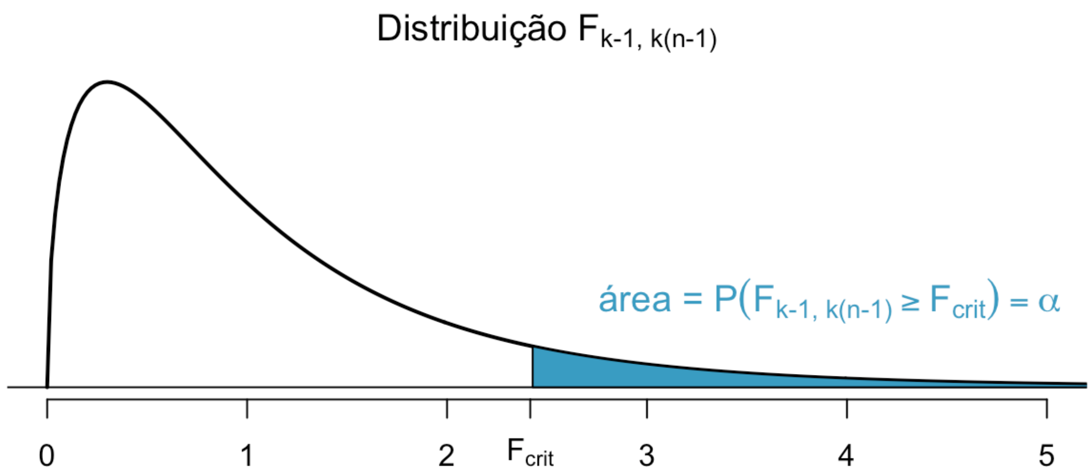
Fornece os quantis $F_{0,95}$ (em cima) e $F_{0,99}$ (em baixo) em função do nº de g.l. numerador ν_1 (coluna) e do nº de g.l. denominador ν_2 (linha)

F tem distribuição F com ν_1 g.l. no numerador e ν_2 g.l. no denominador $P(F \le F_{0,95}) = 0,95$ e $P(F \le F_{0,99}) = 0,99$

$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	7	8	9	10	20	40	60	120	∞
1	161,45	199,50	215,71	224,58	230,16	233,99	236,77	238,88	240,54	241,88	248,01	251,14	252,20	253,25	254,31
	4052,18	4999,50	5403,35	5624,58	5763,65	5858,99	5928,36	5981,07	6022,47	6055,85	6208,73	6286,78	6313,03	6339,39	6365,76
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,45	19,47	19,48	19,49	19,50
	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,45	99,47	99,48	99,49	99,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,66	8,59	8,57	8,55	8,53
	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	26,69	26,41	26,32	26,22	26,13
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,80	5,72	5,69	5,66	5,63
	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,02	13,75	13,65	13,56	13,46
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,56	4,46	4,43	4,40	4,37
	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,55	9,29	9,20	9,11	9,02
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	3,87	3,77	3,74	3,70	3,67
	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,40	7,14	7,06	6,97	6,88
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,44	3,34	3,30	3,27	3,23
	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,16	5,91	5,82	5,74	5,65
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,15	3,04	3,01	2,97	2,93
	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,36	5,12	5,03	4,95	4,86
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	2,94	2,83	2,79	2,75	2,71
	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	4,81	4,57	4,48	4,40	4,31
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,77	2,66	2,62	2,58	2,54
	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,41	4,17	4,08	4,00	3,91
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,12	1,99	1,95	1,90	1,84
	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	2,94	2,69	2,61	2,52	2,42
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	1,84	1,69	1,64	1,58	1,51
	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,37	2,11	2,02	1,92	1,81
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,75	1,59	1,53	1,47	1,39
	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,20	1,94	1,84	1,73	1,60
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,66	1,50	1,43	1,35	1,25
	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,03	1,76	1,66	1,53	1,38
∞	3,84	3,00	2,61	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,57	1,39	1,32	1,22	1,02
	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	1,88	1,59	1,47	1,33	1,03

Teste de Igualdade das Médias para k Grupos

Valor Crítico: Para um nível de significância α, encontrar o valor crítico F_{crit} na tabela F com $k-1$ graus de liberdade no numerador e $k(n-1)$ graus de liberdade no denominador tal que $P(F_{k-1,k(n-1)} \geq F_{crit}) = \alpha$.



Conclusão: Rejeitamos H_0 se $F_{obs} \geq F_{crit} = F_{k-1,k(n-1),\alpha}$

Tabela ANOVA

Tudo o que discutimos até agora pode ser resumido na tabela abaixo. Essa tabela é chamada de **Tabela ANOVA**

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	Estatística F
Grupos (Entre)	SQ_G	$k - 1$	QM_G	$F = \frac{QM_G}{QM_E}$
Erro (Dentro)	SQ_E	$k(n - 1)$	QM_E	
Total	SQ_T	$kn - 1$		

Na prática, basta calcular SQ_T e SQ_G e obter a SQ_E por subtração:

$$SQ_T = SQ_G + SQ_E \qquad \Longrightarrow \qquad SQ_E = SQ_T - SQ_G$$

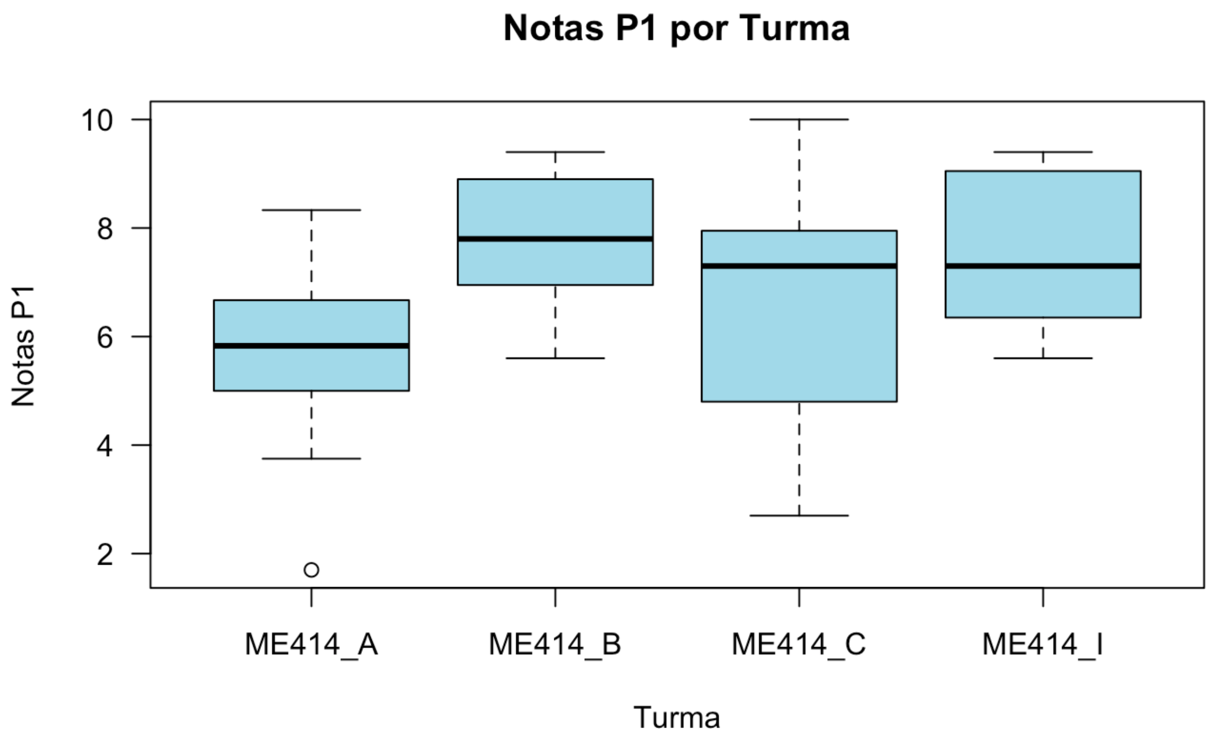
Turmas de ME414 - Notas P1

Voltando no exemplo das notas da P1 para as turmas A, B, C e I. Selecionamos ao acaso 15 alunos de cada turma e anotamos sua respectiva nota na P1.

A tabela abaixo mostra as notas dos primeiros 5 alunos.

Aluno	ME414_A	ME414_B	ME414_C	ME414_I
1	5.00	7.8	9.6	9.4
2	8.33	5.6	7.3	8.5
3	5.00	6.7	2.7	5.6
4	6.67	9.4	10.0	6.0
5	6.67	9.4	5.5	6.7

ME414 - Notas P1



Existe diferença do desempenho na P1 entre as turmas?

Estatísticas Descritivas

Resumo das Notas P1 por Turma

	n	Média	Variância	Desvio Padrão
ME414_A	15	5.71	3.02	1.74
ME414_B	15	7.71	1.75	1.32
ME414_C	15	6.45	6.10	2.47
ME414_I	15	7.66	2.18	1.48

A média geral, considerando todas as notas como sendo de uma única turma é $\bar{X} = 6.88$.

Cálculo das Somas de Quadrados

$$SQ_T = \sum_{i=1}^{4} \sum_{j=1}^{15} (X_{ij} - \bar{X})^2 = 225.31$$

$$\begin{aligned} SQ_G &= n \sum_{i=1}^{4} (\bar{X_i} - \bar{X})^2 \\ &= 15 \left[(5.71 - 6.88)^2 + (7.71 - 6.88)^2 + (6.45 - 6.88)^2 + (7.66 - 6.88)^2\right] \\ &= 42.59 \end{aligned}$$

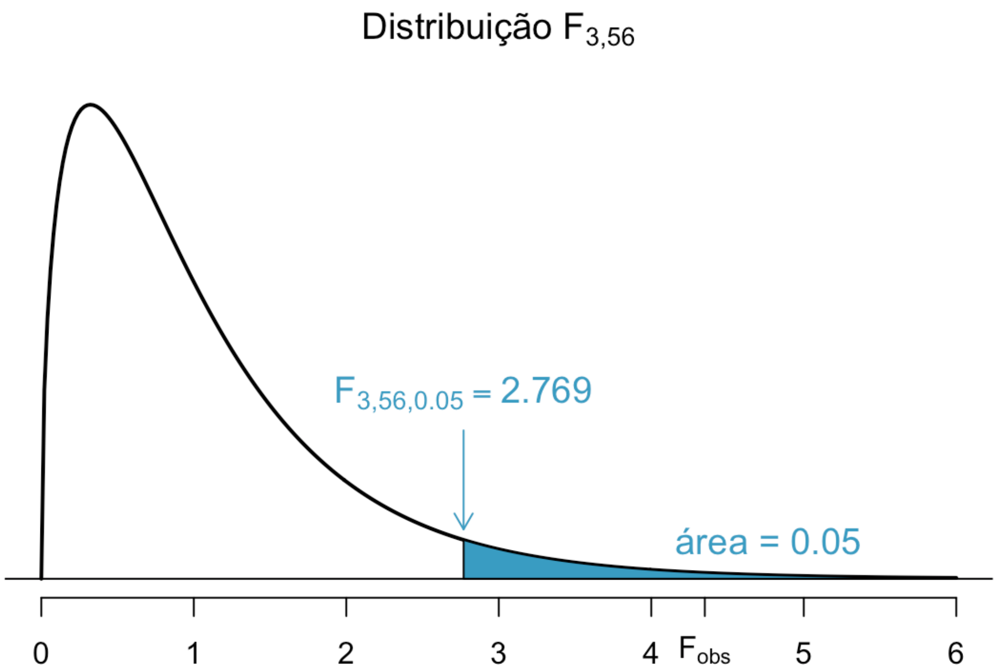
$$\begin{aligned} SQ_E &= SQ_T - SQ_G \\ &= 225.31 - 42.59 = 182.72 \end{aligned}$$

ANOVA - Notas P1 por Turma

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	Estatística F
Grupos (Turma)	42.59	3	14.2	$F = \frac{14.2}{3.26} = 4.351$
Erro	182.72	56	3.26	
Total	225.31	59		

Para $\alpha = 0.05$, olhando na tabela F com 3 e 56 graus de liberdadeo, o valor crítico é $F_{crit} = F_{3,56,0.05} = 2.769$.

ANOVA - Distribuição F



Conclusão: Para $\alpha = 0.05$, como $F_{obs} = 4.351 > 2.769 = F_{crit}$, rejeitamos a hipótese de que as médias da P1 para todas as turmas são iguais.

Exemplo: Qual dieta você faria?

Uma nutricionista quer comparar a perda de peso para três tipos diferentes de dieta. Ela selecionou 12 de seus pacientes e escolheu 4 ao acaso para fazer cada uma das dietas. Depois de um período de três meses os pacientes foram pesados e a perda de peso (em Kg) foi a seguinte:

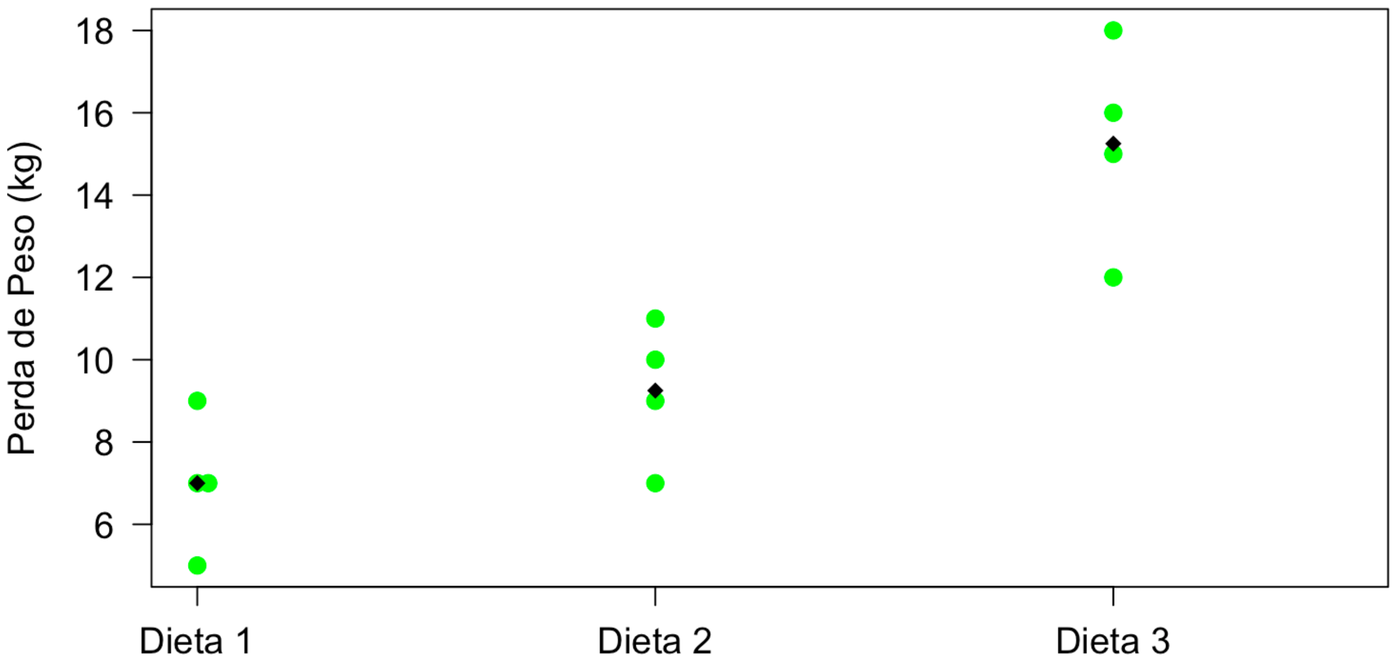
Paciente	Dieta 1	Dieta 2	Dieta 3
1	7	9	15
2	9	11	12
3	5	7	18
4	7	10	16

Exemplo: Dieta

Resumo das Perdas de Peso por Dieta

	n	Média	Variância	Desvio Padrão
Dieta 1	4	7.00	2.67	1.63
Dieta 2	4	9.25	2.92	1.71
Dieta 3	4	15.25	6.25	2.50

Exemplo: Dieta



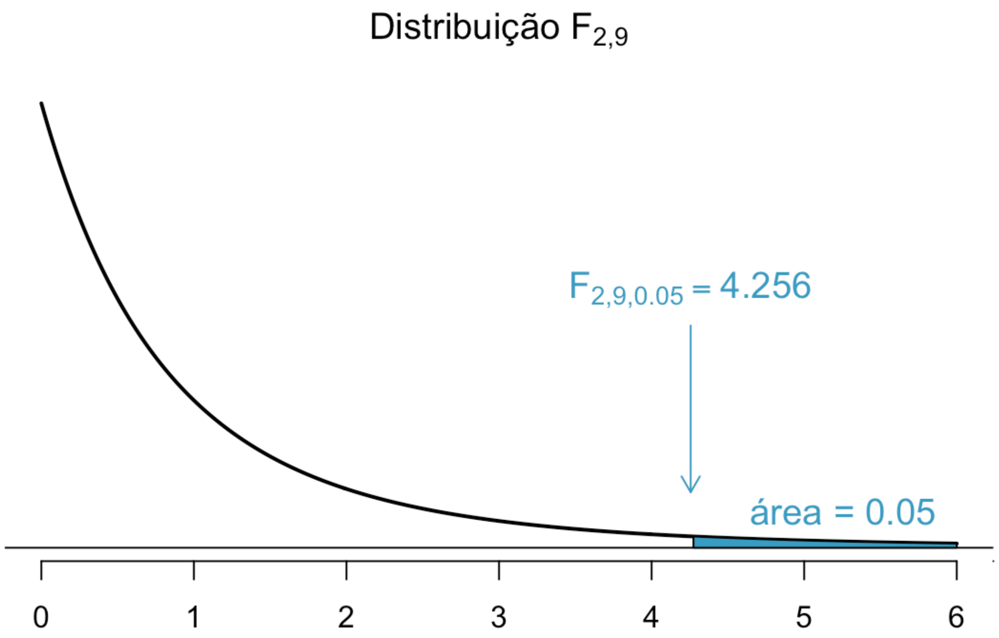
Exemplo: Dieta

Fonte de Variação	Soma de Quadrados	Graus de Liberdade	Quadrado Médio	Estatística F
Dieta	145.5	2	72.75	$F = 18.444$
Erro	35.5	9	3.94	
Total	181	11		

Para $\alpha = 0.05$, olhando na tabela F com 2 e 9 graus de liberdadeo, o valor crítico é $F_{crit} = F_{2,9,0.05} = 4.256$.

Conclusão: Para $\alpha = 0.05$, como $F_{obs} = 18.444 > 4.256 = F_{crit}$, rejeitamos a hipótese de que as perdas de peso médias para todas as dietas são iguais.

Exemplo: Dieta



Conclusão: Para $\alpha = 0.05$, como $F_{obs} = 18.444 > 4.256 = F_{crit}$, rejeitamos a hipótese de que as perdas de peso médias para todas as dietas são iguais.

Leituras

- OpenIntro: seção 5.5
- Magalhães: seção 9.4

Slides produzidos pelos professores:

- Samara Kiihl
- Tatiana Benaglia
- Benilton Carvalho