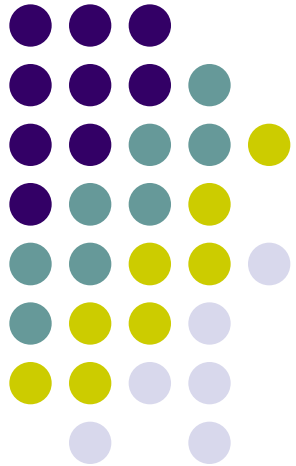
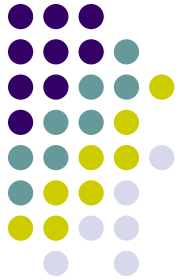
Análise de Variância e outras análises

2018

Airlane P. Alencar



Introdução: Análise de Variância

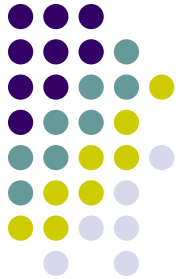


- Caso particular do modelo de regressão:
  - As variáveis explicativas são em geral, de natureza qualitativa (chamadas de fatores);
  - Comparação entre duas populações → Teste t
- Comparação entre mais de duas populações: Análise de variância → ANOVA (Analysis of Variance)

Conceitos

- Terminologia
  - Variável resposta: variável de interesse no estudo
  - Fator: variável explicativa (em geral de natureza qualitativa – variável categorizada)
  - Nível do fator: caracterização do fator, ou seja, as categorias que formam o fator
  - Tratamento:
    - Apenas um fator: tratamento = níveis do fator
    - Dois ou mais fatores: tratamentos = combinação dos níveis dos fatores

Exemplo – Hot Dog

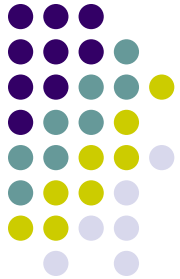


- Verificar se o tipo de salsicha (carne bovina, frango ou mista) influi no número de calorias do hot dog.
  - Variável resposta: número de calorias do hot dog;
  - Fator: tipo de salsicha;
  - Níveis do fator: carne bovina, frango e mista.
- Veja que, como há apenas um fator, os tratamentos serão os próprios níveis do fator: carne bovina, frango ou mista.

Exemplo – Teste de paladar

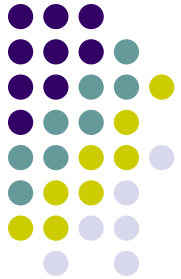
- Um empresa de alimentos quer verificar se o tipo de biscoito (simples ou recheado) e o sabor (morango ou chocolate) tem efeito na preferência de seus consumidores. A empresa faz uma pesquisa em que os consumidores experimentam os biscoitos e dão uma nota de 1 a 10 para o biscoito experimentado.
  - Variável resposta: nota dada pelo consumidor (para mensurar sua preferência);
  - Fatores:
    - Tipo do biscoito – dois níveis (simples ou recheado)
    - Sabor do biscoito (chocolate ou morango)
  - Tratamentos: quatro possíveis tratamentos – biscoito recheado de chocolate, biscoito simples de chocolate, biscoito recheado de morango e biscoito simples de morango.

Objetivos da ANOVA



- Avaliar o efeito dos fatores sobre a **média** da variável resposta;
- Comparar os efeitos dos diferentes tratamentos sobre a **média** da variável resposta

Formulação do modelo de ANOVA (com um fator)



- Vamos pensar em um caso que desejamos verificar o efeito de um fator com *k* níveis.
- Modelo

$$y_{ij} = \mu_i + \varepsilon_{ij},\ i = 1,\ldots,k \text{ e } j = 1,\ldots,n$$

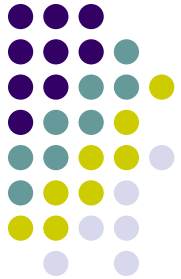
em que

μ_i é a média populacional da variável resposta para o i-ésimo nível do fator estudado;

erros aleatórios independentes ε_{ij} ~Normal(0, σ^2).

- **Objetivo**: comparar o efeito dos níveis do fator em estudo = testar a igualdade das médias μ_i's

Suposições do modelo de ANOVA

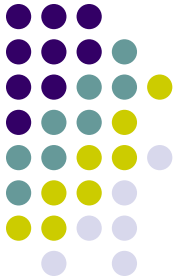


- As suposições do modelo de ANOVA são semelhantes às do modelo de regressão:

1. ε_{ij} com média zero
2. ε_{ij} são independentes entre si → Suposição de independência
3. $Var(\varepsilon_{ij}) = \sigma^2$ → Suposição de homocedasticidade
4. $\varepsilon_{ij} \sim Normal(0, \sigma^2)$ → Suposição de normalidade

- A validade das suposições é checada em análise de resíduos.

Hipótese testada pela ANOVA



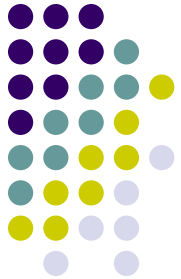
- Nosso objetivo é testar as hipóteses:

H_0: $\mu_1 = \mu_2 = ... = \mu_k = \mu$

H_a: os μ_i's não são todos iguais

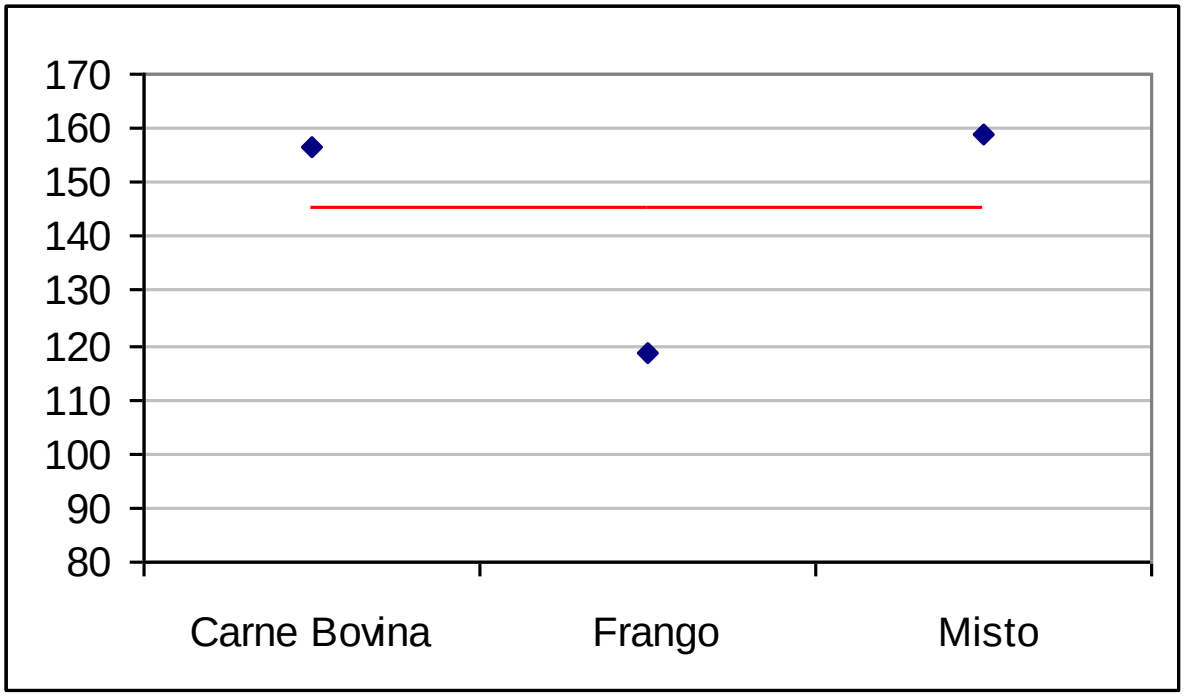
  - Queremos verificar se a média da variável resposta é igual para todos os níveis do fator estudado.

Análise dos efeitos dos tratamentos



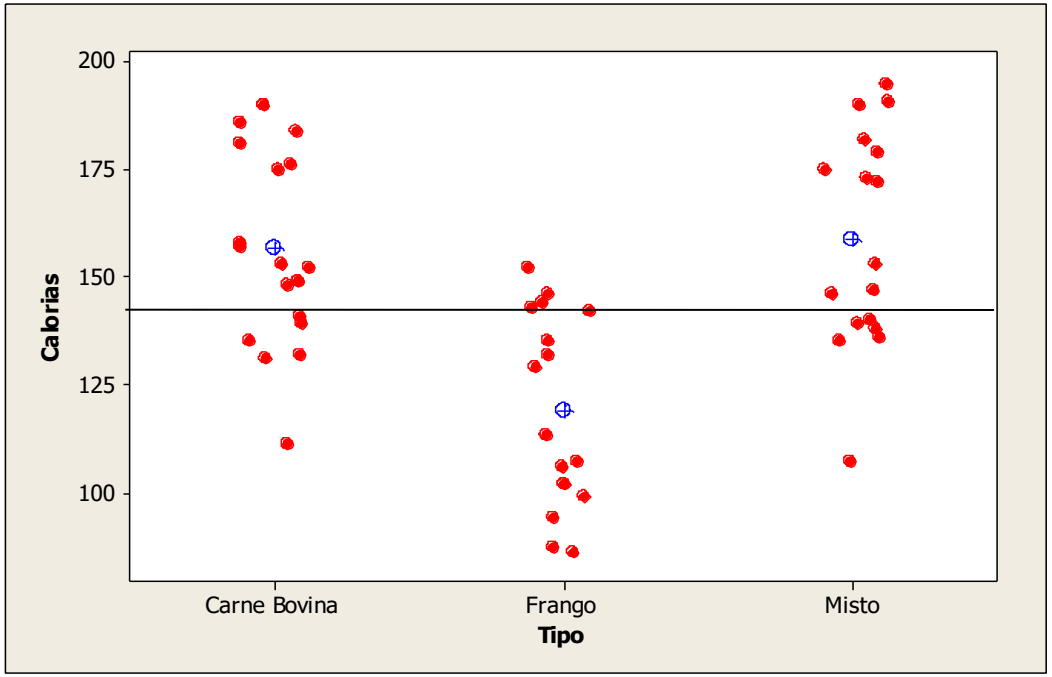
- Se H_0 não é rejeitada: não há evidências de relação entre a variável resposta e o fator;
- Se H_0 é rejeitada: devemos localizar as diferenças entre as médias da variável resposta sob os diferentes níveis do fator – Qual ou quais médias são diferentes?
  - Comparações múltiplas (Tukey, Scheffé, Bonferroni, etc.)

Exemplo: Hot Dog - Número médio de calorias por tipo de salsicha



- Médias diferentes?
- Média geral em vermelho

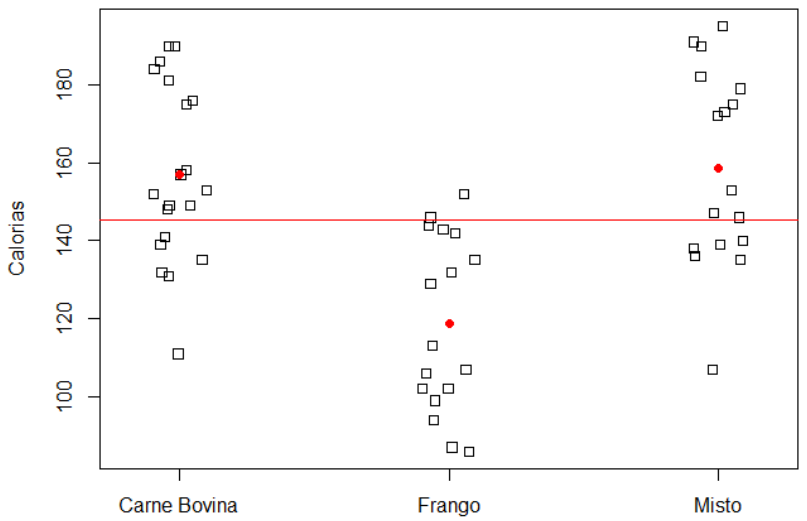
Calorias



- Levando-se em conta a variabilidade do número de calorias dos vários tipos de salsicha, mas médias não parecem tão distantes...

Dotplot

```
(Table1 <- aggregate(Calorias ~
Tipo, data=d, FUN="mean"))

stripchart(Calorias ~ Tipo,
data=d, vertical=TRUE,
method="jitter")

points(c(1,2,3),Table1[,2],
col=2, pch=16)

abline(h=mean(d$Calorias),
col=2)
```



```
f<-lm(Calorias ~ Tipo, data=d)
summary(f)
anova(f)
Estimate Std. Error t value Pr(>|t|)
(Intercept)  156.850      5.246  29.901  < 2e-16 ***
TipoFrango   -38.085      7.739  -4.921 9.39e-06 ***
TipoMisto      1.856      7.739   0.240    0.811
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 23.46 on 51 degrees of freedom
Multiple R-squared:  0.3866,      Adjusted R-squared:  0.3626
F-statistic: 16.07 on 2 and 51 DF,  p-value: 3.862e-06

Analysis of Variance Table

Response: Calorias
          Df Sum Sq Mean Sq F value    Pr(>F)
Tipo       2  17692  8846.1  16.074 3.862e-06 ***
Residuals 51  28067   550.3
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Resultados

	DF	SS	MS	F	P
Tipo	2	17692	8846	16.07	0.000
Error	51	28067	550		
Total	53	45759			

- A estatística F é a razão entre as medidas de variabilidades entre os grupos (explicada) e intra-grupo (Resíduo ou erro). Para **g** grupos:

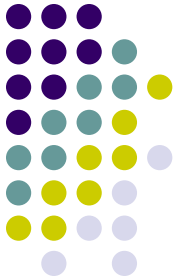
$$F = \frac{SQExp/g}{SQRes/(n-g)} = \frac{17692/2}{28067/51} = \frac{8846}{550} = 16{,}07$$

- A estatística sob H_0: $\mu_1 = \mu_2 = \mu_3$ tem dist $F_{g,n-g}$.

- No caso acima, $F_{3,51}$, p=P($F_{3,51}$>16,07)<0,001, logo, com os níveis de significância usuais (5%), rejeitamos H_0, então as médias não são todas iguais.

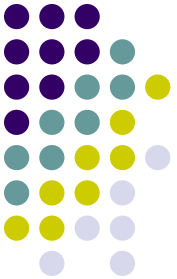
- Mas onde estão as diferenças? https://rcompanion.org/rcompanion/d_05.html

Comparações 2 a 2 - Tukey



- $\bar{y}_i - \bar{y}_j \pm \frac{q_{\propto,k,N-k}}{\sqrt{2}} \hat{\sigma} \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$,
- comparando as médias dos grupos i e j
- usando a distr q proposta por Tukey por exemplo em https://www2.stat.duke.edu/courses/Spring98/sta110c/qtable.html
- Para tamanhos de amostras diferentes: Tukey-Kramer

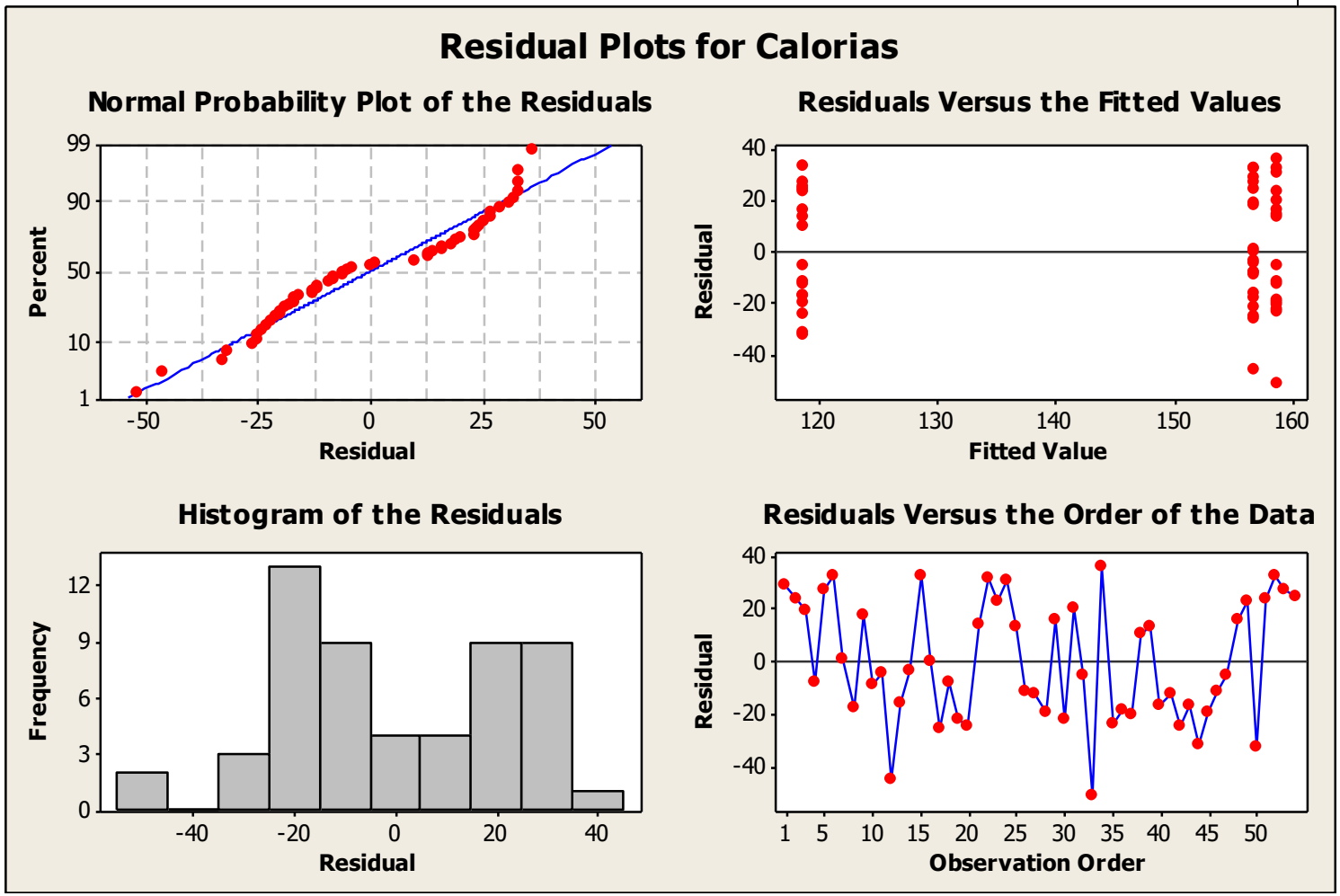
Comparações Múltiplas: ICs simultâneos de 95% de Tukey



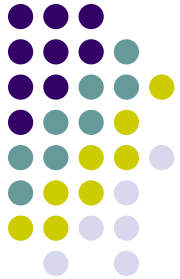
```
> comp <- emmeans(f, ~ Tipo, data=d)
> comp
 Tipo            emmean       SE df lower.CL upper.CL
 Carne Bovina 156.8500 5.245646 51 146.3189 167.3811
 Frango       118.7647 5.689702 51 107.3422 130.1873
 Misto        158.7059 5.689702 51 147.2833 170.1284

Confidence level used: 0.95
> pairs(comp)
 contrast                estimate       SE df t.ratio p.value
 Carne Bovina - Frango  38.085294 7.738831 51   4.921  <.0001
 Carne Bovina - Misto   -1.855882 7.738831 51  -0.240  0.9688
 Frango - Misto        -39.941176 8.046454 51  -4.964  <.0001

P value adjustment: tukey method for comparing a family of 3
   estimates
```

- Só o intervalo para a diferença Misto-Carne é que contém o zero.

Análise de Resíduos

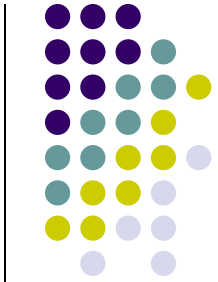


Alternativas para quando as suposições não são válidas



- Se as suposições do modelo não são válidas, podemos corrigir a heterocedasticidade ou utilizar testes não-paramétricos.
- Os testes não paramétricos não se baseiam nas médias, sendo que essas podem ser muito influenciadas por pouco valores discrepantes.
- Por exemplo, o teste Kruskal-Wallis pode ser utilizado para testar se as distribuições da variável respostas nos 3 grupos são semelhantes. O teste utiliza os postos referentes a cada observação e calcula as médias dos postos em cada grupo.

Kruskal-Wallis

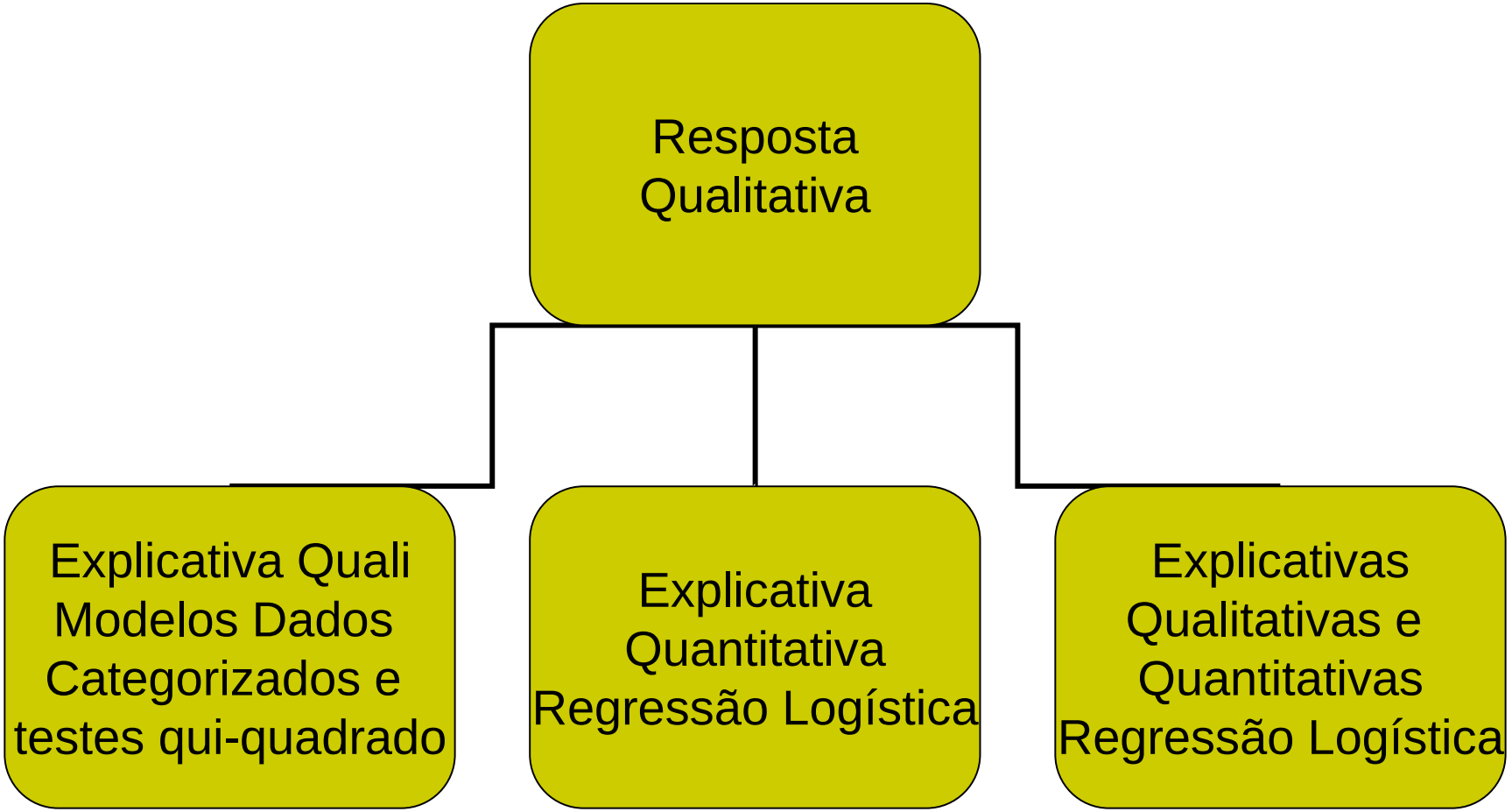


Tipo	N	Median	Ave Rank	Z
Carne Bovina	20	152,5	33,8	2,25
Frango	17	113,0	13,6	-4,39
Misto	17	153,0	34,0	2,05
Overall	54		27,5	

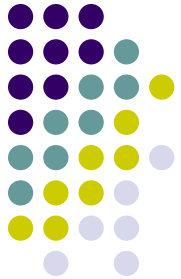
H = 19,24 DF = 2 P = 0,000
H = 19,25 DF = 2 P = 0,000 (adjusted for ties)

- A média dos postos é menor para a salsicha de frango.
- Rejeita-se a igualdade da distribuição das calorias nos 3 grupos (p<0,001).
- Também são propostas comparações múltiplas utilizando postos.

Variável Resposta Qualitativa



Regressão Logística



- Em um estudo para investigar a incidência de dengue numa determinada cidade da costa mexicana, um total de 196 indivíduos (Paula, 2004 e Neter et al. ,1996), escolhidos aleatoriamente em dois setores da cidade, respondeu às seguintes perguntas:
- Idade
- Nível sócio-econômico: 1 = Baixo, 2= Médio, 3= Alto
- Setor da cidade onde mora o entrevistado: 1 ou 2
- Dengue: 1 se contraiu a doença recentemente e 0 caso, contrário.

- Dos 196 entrevistados, 57 (29%) tiveram dengue.
- Será que a probabilidade de contrair dengue depende da idade, nível sócio-econômico ou setor da cidade?

Dengue e Setor

	Setor		
	1	2	Total
Dengue	22	35	57
Sem Dengue	95	44	139
Total	117	79	196

	Setor		
	1	2	Total
Dengue	19%	44%	29%
Sem Dengue	81%	56%	71%
Total	100%	100%	100%

- 44% dos entrevistados do Setor 2 tiveram dengue, enquanto essa proporção é de 19% no Setor 1.
- A chance de ter dengue com relação a não ter para quem é do setor 2 é 3,43 vezes a chance de quem mora no Setor 1. IC=[*1,81; 6,53*], p<001.

Dengue e Nível Sócio-Econômico

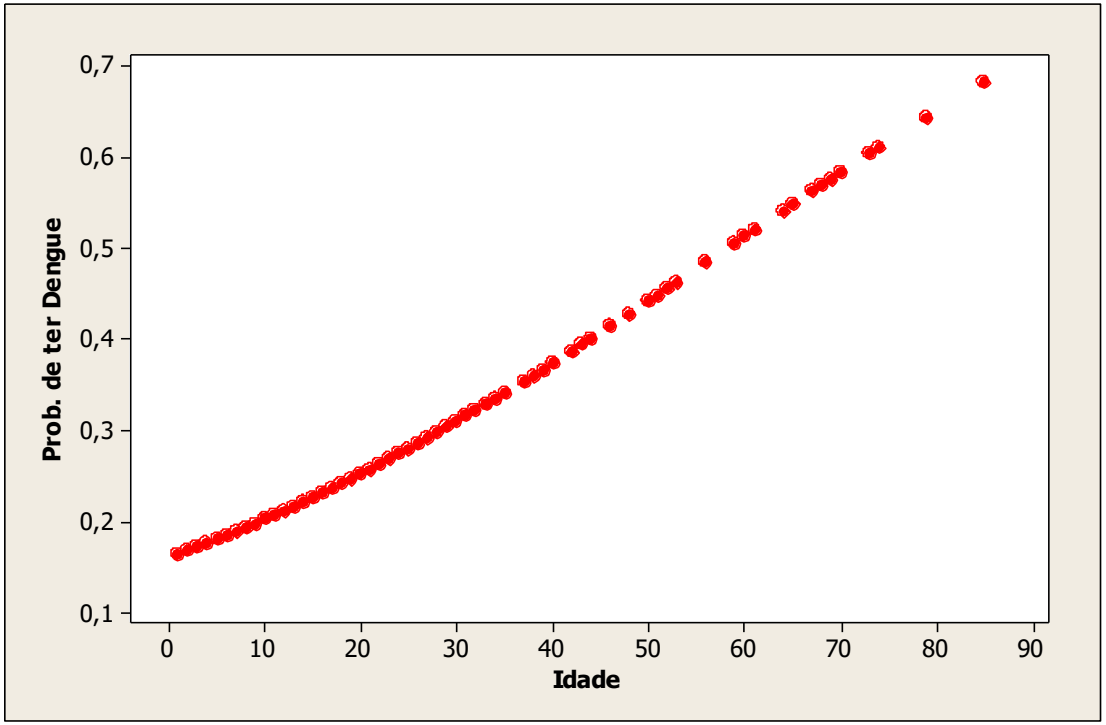
	Nível Sócio-econômico			
	Baixo	Médio	Alto	Total
Dengue	19	14	24	139
Sem Dengue	51	35	53	57
Total	70	49	77	196

	Nível Sócio-econômico			
	Baixo	Médio	Alto	Total
Dengue	27%	29%	31%	71%
Sem Dengue	73%	71%	69%	29%
Total	100%	100%	100%	100%

- Quanto maior o nível sócio-econômico, maior a proporção de pessoas com dengue.
- OR (Médio/Baixo)=1,07, p= 0,07 => IC=[***0,48 2,42***] => NS a 5%
- OR (Alto/Baixo)=1,22, p= 0,59 => IC=[***0,60 2,48***] => NS a 5%

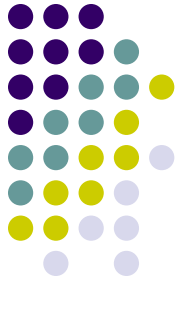
Dengue e Idade

- Probabilidade de ter dengue depende da idade?
- Considerando somente idade, temos:



- OR= 1,03 => IC=[***1,01 1,05***], p=0,001.

Modelo Logístico Múltiplo



Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-2,04050	0,389478	-5,24	0,000			
Setor							
2	1,24363	0,352291	3,53	0,000	3,47	1,74	6,92
Nível sócio-ec.							
2	-0,208825	0,454527	-0,46	0,646	0,81	0,33	1,98
3	-0,253433	0,405552	-0,62	0,532	0,78	0,35	1,72
Idade	0,026991	0,008675	3,11	0,002	1,03	1,01	1,04

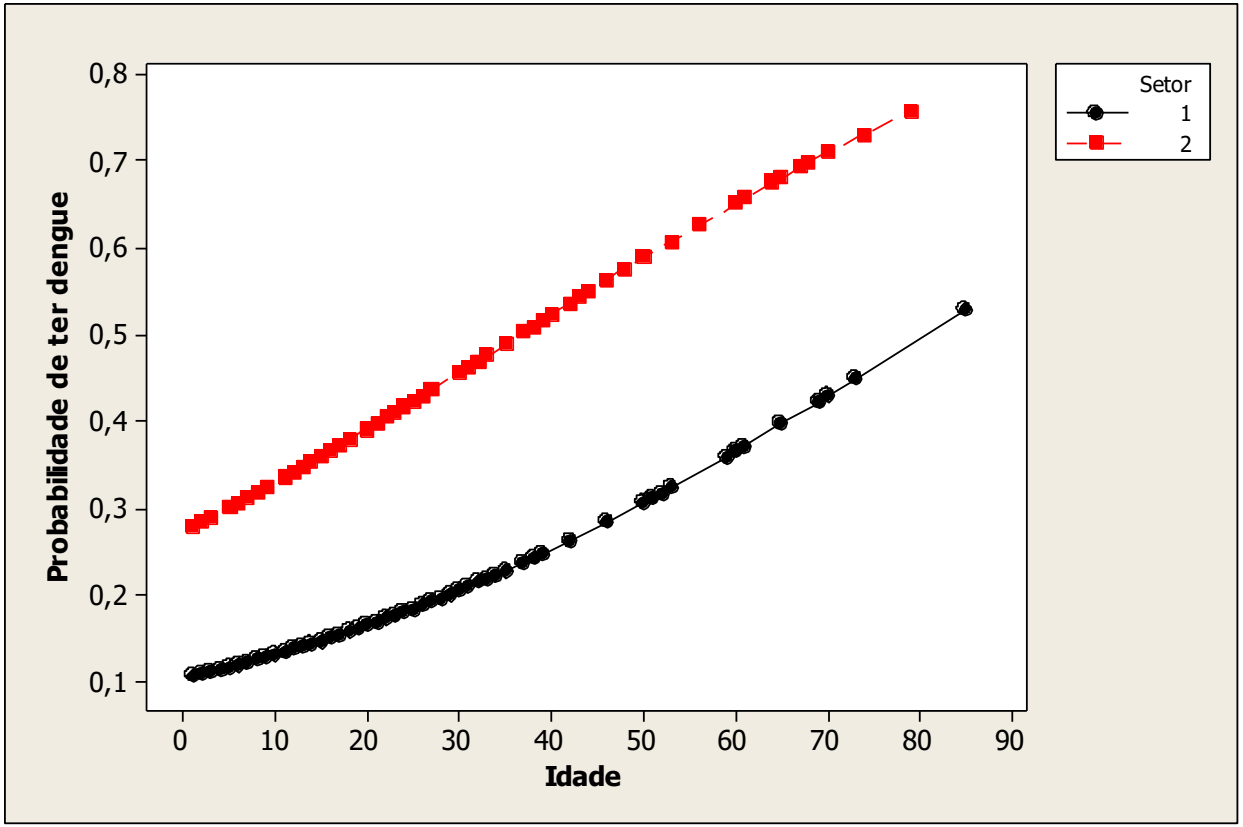
Nível sócio-econômico não apresenta efeito significativo.

Modelo com Idade e Setor

Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	-2,15966	0,343904	-6,28	0,000			
Idade	0,0268129	0,0086501	3,10	0,002	1,03	1,01	1,04
Setor							
2	1,18169	0,336963	3,51	0,000	3,26	1,68	6,31

- A cada ano a mais na idade, a chance de ter dengue com relação a não ter sobe 3% em média, mantendo setor constante.
- A chance de ter dengue (com relação a não ter) para quem é do setor 2 é 3,26 vezes a chance de quem é do setor 1, mantendo idade constante.

Probabilidade de ter Dengue em função da Idade e Setor



OR x RR

$$RR = \frac{OR}{(1 - P_0) + (P_0 OR)}$$

- Zhang e Yu. (1998). What´s the relative risk? JAMA, 18, 1690-1.

Referências

- Conover, W.J. (1980). **Practical Nonparametric Statistics.** Second Edition, New York: John Wiley & Sons, Inc.
- Kutner, Michael H; Nachtshein, Chistopher J; Neter, John; Li, William (2005). **Applied Linear Statistical Models**, Fifth Edition. Boston : McGraw-Hill Irwin.
- Soares, J. F. e Siqueira, A. L. (2002). **Introdução à estatística médica**. 2ª edição. Belo Horizonte: COOPMED.