Regression and Clinical prediction models

Session 4
Introducing statistical modeling – Part 1 (Analysis of variance)

Pedro E A A do Brasil
pedro.brasil@ini.fiocruz.br
2019

Objetivos

- Introduzir a ideia de que eventos em saúde podem ser previstos por modelos estatísticos.
- Apresentar conceitos, passos e interpretações de uma análise de variância.

Teste de hipótese

- Revisando: Testes de hipóteses paramétricos
  - **Hipótese:** alegação ou afirmação sobre uma propriedade de uma população.
  - **Testes de hipóteses:** representam uma **regra de decisão** que permite rejeitar ou não uma hipótese questionada, sendo a decisão tomada em função de valores obtidos em uma amostra.

Teste de hipótese

- Os elementos básicos de um teste de hipóteses são:
  - Hipóteses (nula e alternativa)
  - Erros tipo I e II
  - Nível de significância (α)
  - Estatística de teste
  - Região crítica
  - p-valor
  - Regra de decisão

Teste de hipótese

- Resumo dos procedimentos
  1. Enunciar as hipóteses H_0 e H_1.
  2. Determinar um nível de significância (α) aceitável.
  3. Estabelecer a estatística de teste.
  4. Devemos de terminar a região crítica em função da variável tabelada.
  5. Calculamos o valor da estatística de teste obtido na amostra.
  6. Rejeitar ou não rejeitar a hipótese nula de acordo com a estimativa obtida no 4º item em comparação com a região crítica estabelecida no 3º item.

- Opcionalmente, podemos pular a etapa 4 e calcular o p-valor na etapa 5.

Teste de hipótese

- Exemplificando testes de hipóteses para a média
  - **Testes para uma amostra**
    - σ conhecido: estatística-teste Z
    - σ desconhecido: estatística teste t-student
  - **Testes para duas amostras (teste t)**
    - σ^2 desconhecida: estatística teste t-student
      - Dados pareados
      - Dados independentes, σ^2 supostas iguais
      - Dados independentes, σ^2 supostas diferentes
  - **Teste para três ou mais amostras**
    - Análise de variância (ANOVA)

Análise de variância (ANOVA)

- É um método para testar a igualdade de três ou mais médias populacionais.
- Exemplo:
  - A temperatura média é diferente segundo a faixa etária dos indivíduos (18 a 20 anos, entre 21 a 29 anos e 30 anos ou mais)?
  - ou
  - Esses três grupos etários têm a mesma temperatura corporal média?

Análise de variância (ANOVA)

- Hipóteses:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

$$H_1 : \text{pelo menos uma média difere}$$

- Fator (**preditor** ou determinante): é uma característica que permite distinguir diferentes populações umas das outras.

- Exemplo: Temperatura corporal (°C)
  - Trata-se de um único fator, pois as três populações se distinguem segundo uma única característica (grupo etário).

Análise de variância (ANOVA)

- Exemplo 1:
  - Temperatura corporal em (°F) classificados em três grupos etários diferentes.

18 a 20	21 a 29	30 ou mais
98,0	99,6	98,6
98,4	98,2	98,6
97,7	99,0	97,0
98,5	98,2	97,5
97,1	97,9	97,3
97,0	97,0	98,3
98,2	99,2	98,0
97,3	99,6	97,7
97,7	99,0	97,0
98,0		98,5
		98,6
$n_1 = 10$	$n_2 = 9$	$n_3 = 11$
$\bar{x}_1 = 97,79$	$\bar{x}_2 = 98,63$	$\bar{x}_3 = 97,92$
$s_1 = 0,53$	$s_2 = 0,87$	$s_3 = 0,65$

Análise de variância (ANOVA)

- Pressupostos da Análise de variância (ANOVA)
  - Aleatoriedade e independência
  - Normalidade – os dados das amostras seguem a distribuição normal
  - Homogeneidade da variância – A variância dos grupos são iguais (teste de Levene)
  - *O teste ANOVA não é muito afetado por desvios moderados dos pressupostos.*

- *Perguntas: O que fazer caso a normalidade seja rejeitada? E caso a homogeneidade seja rejeitada? E se ambos forem rejeitados?*

Análise de variância (ANOVA)

- *Perguntas:*
- *O que fazer caso a normalidade seja rejeitada?*
  - *Realizar um teste não-paramétrico*
- *E caso a homogeneidade seja rejeitada?*
  - *Realizar um teste de comparação múltipla apropriado a variâncias desiguais*
- *E se ambos forem rejeitados?*
  - *Transformação da variável por logaritmo, raiz quadrática, dentre outros.*

Análise de variância (ANOVA)

- Como montar a tabela ANOVA?
  - 1º Passo: devem ser calculadas as variâncias dentro dos grupos, entre os grupos e total
  - 2º Passo: calcular a estatística F
  - 3º Passo: Obter o p-valor

  - *Onde n é o número de observações totais e c é o número de grupos*

Análise de variância (ANOVA)

- Variação total ou soma dos quadrados (SQT)

$$SQT = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(X_{ij} - \overline{\overline{X}})^2$$

$$\overline{\overline{X}} \rightarrow Grande\ média \left(\frac{\sum_{j-1}^{c}\sum_{i-1}^{n_j} X_{ij}}{n} \right)$$

$X_{ij} \rightarrow$ i – ésimo valor no grupo j

$n_j \rightarrow$ número de valores no grupo j

$n \rightarrow$ número total de valores em todos os grupos combinados (ou seja, $n = n_1 + n_2 + ... + n_c$)

$c \rightarrow$ número de grupos

Análise de variância (ANOVA)

- Variação entre os grupos (SQE)

$$SQE = \sum_{j=1}^{c} n_j (\overline{X}_j - \overline{\overline{X}})^2$$

$$\overline{\overline{X}} \rightarrow Grande\ média \left(\frac{\sum_{j=1}^{c} \sum_{i=1}^{n_j} X_{ij}}{n} \right)$$

$\overline{X}_j \rightarrow$ média aritmética da amostra do grupo j

$n_j \rightarrow$ número de valores no grupo j

$c \rightarrow$ número de grupos

Análise de variância (ANOVA)

- Variação dentro dos grupos (SQD)

$$SQD = \sum_{j=1}^{c} n_j (X_{ij} - \overline{X}_j)^2$$

$X_{ij} \rightarrow$ i – ésimo valor no grupo j

$\overline{X}_j \rightarrow$ média aritmética da amostra do grupo j

Análise de variância (ANOVA)

- Media dos Quadrados (MQ)

$$MQE = \frac{SQE}{c-1}$$

$$MQD = \frac{SQD}{n-c}$$

$$MQT = \frac{SQT}{n-1}$$

Análise de variância (ANOVA)

- **Para testar as hipóteses:**
  - H_0 : $\mu_1 = \mu_2 = \mu_3 = \ \ldots = \mu_n$
  - H_1 : $pelo\ menos\ uma\ média\ difere$

  - A estatística F é uma razão entre duas variâncias

$$F = \frac{Variância\ entre\ as\ asmostras}{Variância\ dentro\ das\ amostras} = \frac{MQE}{MQD}$$

  - *F tabelado:* F com (c-1)(n-c) graus de liberdade e nível de significância (α)
  - Se $F_{calculado} > F_{tabelado}$ → Rejeita-se H_0

Análise de variância (ANOVA)

Fontes de variação	Soma dos quadrados (SQ)	Graus de liberdade (gl)	Quadrados médios (QM)	F	p-valor
Entre grupos	SQE	k-1	$QME = \frac{SQE}{k-1}$	$F = \frac{QME}{QMD}$	p-valor
Dentre grupos	SQD	n-k	$QMD = \frac{SQD}{n-k}$		
Total	SQT	n-1	$QMT = \frac{SQT}{n-1}$		

- **Notas:** SQT = SQE + SQD
- Se as populações têm médias iguais F → 1
- O F tabelado dependerá dos graus de liberdade (k-1)(n-k) e do nível de significância.
- O p-valor exato só pode ser obtido pelo software estatístico.

Análise de variância (ANOVA)

- Exemplo: Temperatura corporal (ºF) classificados em três grupos etários diferentes.

```
> by(banco$temperatura,banco$grupo,summary)
banco$grupo: 18 a 20 anos
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  97.00   97.40   97.85   97.79   98.15   98.50
------------------------------------------------------------------------
banco$grupo: 21 a 29 anos
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  97.00   98.20   99.00   98.63   99.20   99.60
------------------------------------------------------------------------
banco$grupo: 30 anos ou mais
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  97.00   97.40   98.00   97.92   98.55   98.60
```

Análise de variância (ANOVA)

- Exemplo: Temperatura corporal (ºF) classificados em três grupos etários diferentes.

```
> resultado <- aov(temperatura ~ grupo, data = banco)
> summary(resultado)
             Df Sum Sq Mean Sq F value Pr(>F)
grupo         2  3.882  1.9408   4.124 0.0274 *
Residuals    27 12.705  0.4706
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Rejeita-se a hipótese nula. Há evidências estatísticas de que as médias de temperatura corporal são diferentes entre as faixas etárias, considerando o nível de significância de 5%. Entretanto, não conseguimos identificar quais faixas etárias apresentam médias de temperaturas diferentes. Utilizamos, então um teste de comparação múltipla.

Análise de variância (ANOVA)

- Testes de comparação múltiplas:
  - Após a ANOVA mostrar significância devemos testar todos os pares de grupos para descobrir quais as médias que diferem.
  - Entretanto, devemos empregar vários testes t para duas amostras independentes.
  - O teste de comparação múltipla serve para corrigir (ponderar) o p-valor obtido no teste-t pelo número de comparações efetuadas
  - Caso contrário, podemos encontrar facilmente significância estatística numa relação inexistente.
  - Existem diversos testes para comparação múltiplas das médias:
  - No caso de variâncias iguais, os mais utilizados são o Tukeye Bonferroni
  - No caso de variâncias desiguais, podem ser utilizados testes alternativos (ex: Dunnett)
  - Não existem regras sobre quais os testes devem ser preferidos.

Análise de variância (ANOVA)

- Testes de comparação múltiplas:

```
> t.test(temperatura ~ grupo, data = banco[banco$grupo != "30 anos ou mais",])

            Welch Two Sample t-test

data:  temperatura by grupo
t = -2.5251, df = 12.898, p-value = 0.02548
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.5654389 -0.1212278
sample estimates:
mean in group 18 a 20 anos mean in group 21 a 29 anos
                  97.79000                   98.63333

> t.test(temperatura ~ grupo, data = banco[banco$grupo != "18 a 20 anos",])

            Welch Two Sample t-test

data:  temperatura by grupo
t = 2.0487, df = 14.503, p-value = 0.05904
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.03111679  1.46141982
sample estimates:
   mean in group 21 a 29 anos mean in group 30 anos ou mais
                     98.63333                      97.91818
```

Análise de variância (ANOVA)

- Testes de comparação múltiplas:

```
> t.test(temperatura ~ grupo, data = banco[banco$grupo != "21 a 29 anos",])

        Welch Two Sample t-test

data:  temperatura by grupo
t = -0.50038, df = 18.794, p-value = 0.6226
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6647471  0.4083835
sample estimates:
   mean in group 18 a 20 anos mean in group 30 anos ou mais
                     97.79000                      97.91818
```

Análise de variância (ANOVA)

- Testes de comparação múltiplas (Bonferroni):

```
> pairwise.t.test(temperatura, grupo, p.adj = "bonf", data = banco, pool.sd = T)

        Pairwise comparisons using t tests with pooled SD

data:  temperatura and grupo

                18 a 20 anos 21 a 29 anos
21 a 29 anos    0.038        -
30 anos ou mais 1.000        0.085

P value adjustment method: bonferroni
```

Análise de variância (ANOVA)

- Testes de comparação múltiplas (Bonferroni):

```
> tu <- TukeyHSD(resultado)
> tu
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = temperatura ~ grupo, data = banco)

$`grupo`
                                    diff        lwr        upr     p adj
21 a 29 anos-18 a 20 anos      0.8433333  0.06185506 1.62481161 0.0324820
30 anos ou mais-18 a 20 anos   0.1281818 -0.61496508 0.87132871 0.9044577
30 anos ou mais-21 a 29 anos -0.7151515 -1.47961855 0.04931552 0.0700558
```

- Então podemos dizer que há evidências estatísticas suficientes ao nível de 5% para afirmar que há diferença entre as médias de temperatura das faixas etárias de 18 a 20 anos e 21 a 29 anos.

fim

Session 4
Introducing statistical modeling – Part 1 (Analysis of variance)

Pedro E A A do Brasil
pedro.brasil@ini.fiocruz.br
2019