UNIVERSIDADE DA REGIÃO DE JOINVILLE - UNIVILLE

Bacharelado em Engenharia de Software (BES)

Estatística para Computação

Professora Priscila Ferraz Franczak

Engenheira Ambiental - UNIVILLE

Mestre em Ciência e Engenharia de Materiais - UDESC

Doutora em Ciência e Engenharia de Materiais - UDESC

priscila.franczak@gmail.com

Plano de Aula

Análise de variância (parte 1)

1. Delineamento inteiramente causalizado
2. Delineamento em blocos causalizados
3. Delineamento em quadrado latino
4. Experimentos fatoriais
5. Exercícios

- Os testes de hipóteses estudados até agora limitaram-se à comparação de duas médias ou duas proporções.

- Contudo, há situações onde se deseja comparar várias médias, cada uma oriunda de um grupo diferente.

A análise de variância é um método estatístico, que, por meio de teste de igualdade de médias, verifica se fatores (variáveis independentes) produzem mudanças sistemáticas em alguma variável de interesse (variável dependente).

Os fatores propostos podem ser variáveis quantitativas ou qualitativas, enquanto a variável dependente deve ser quantitativa e observada dentro das classes dos fatores – os tratamentos.

1. Delineamento inteiramente causalizado (DIC)

- Trata de experimentos em que os dados não são pré-separados ou classificados em categorias mais conhecidas como blocos.

- A ANOVA, associada a esse tipo de experimento, é muitas vezes chamada *One Way ANOVA* (classificação única ou experimento com um fator).

- Aplicado a projetos experimentais completamente aleatórios, em que amostras independentes são retiradas de k populações normais ($k > 2$) com médias μ_1, μ_2, μ_3... $\mu_{k,}$ respectivamente, e variância σ^2.

- As populações são supostas com variâncias iguais.

- As amostras podem ser de tamanhos diferentes, sendo o número total de observações da experiência igual a $n = n_1 + n_2 + ... + n_k$

- As populações são denominadas tratamentos – categorias ou níveis do fator.

- Por meio de um teste estatístico, procuramos verificar se determinado fator é possível causa dos efeitos observados em certa variável de estudo.

- A hipótese nula do teste é de que as médias dos k tratamentos são iguais, isto é:

$$H_o : \mu_1 = \mu_2 = ... = \mu_k$$

- A hipótese alternativa H_1 é a de que pelo menos duas médias sejam diferentes.

- Caso o teste estatístico indique a rejeição de H_o, pode-se concluir, com risco α, que o fator considerado tem influência sobre a variável de estudo.

Quadro de análise de variância

Fonte de variação	**Soma dos quadrados**	**Graus de liberdade**	**Quadrados médios**	**Teste F**
Entre tratamentos	Q_e	$k - 1$	$S_e^2 = \dfrac{Q_e}{k-1}$	$F_{cal} = \dfrac{S_e^2}{S_r^2}$
Dentro das amostras (residual)	$Q_r = Q_t - Q_e$	$n - k$	$S_r^2 = \dfrac{Q_t - Q_e}{n-k}$	
Total	Q_t	$n - 1$		

Q_e = variação entre os tratamentos
Q_r = variação dentro dos tratamentos (residual)
Q_t = variação total
k = quantidade de tratamentos

n = tamanho da amostra
S_e^2 = variância devido aos tratamentos
S_r^2= variância devido aos erros

- Para testar H_o contra H_1, comparamos o valor F_{cal} com o valor F tabelado com (k - 1) g.l. no numerador e (n - k) no denominador, fixando certo nível de significância.

- $F_{cal} \leq F_{tab}$: não se pode rejeitar H_o, concluindo, com risco α, que o fator considerado não causa efeito sobre a variável de estudo.

- $F_{cal} > F_{tab}$: rejeita-se H_o, concluindo, com risco α, pela diferença das médias, e consequente influência sobre a variável analisada.

Exemplo: O resultado das vendas efetuadas por três vendedores de uma indústria durante certo tempo é dado na tabela abaixo. Deseja-se saber, ao nível de 5% de significância, se há diferença de eficiência entre os vendedores.

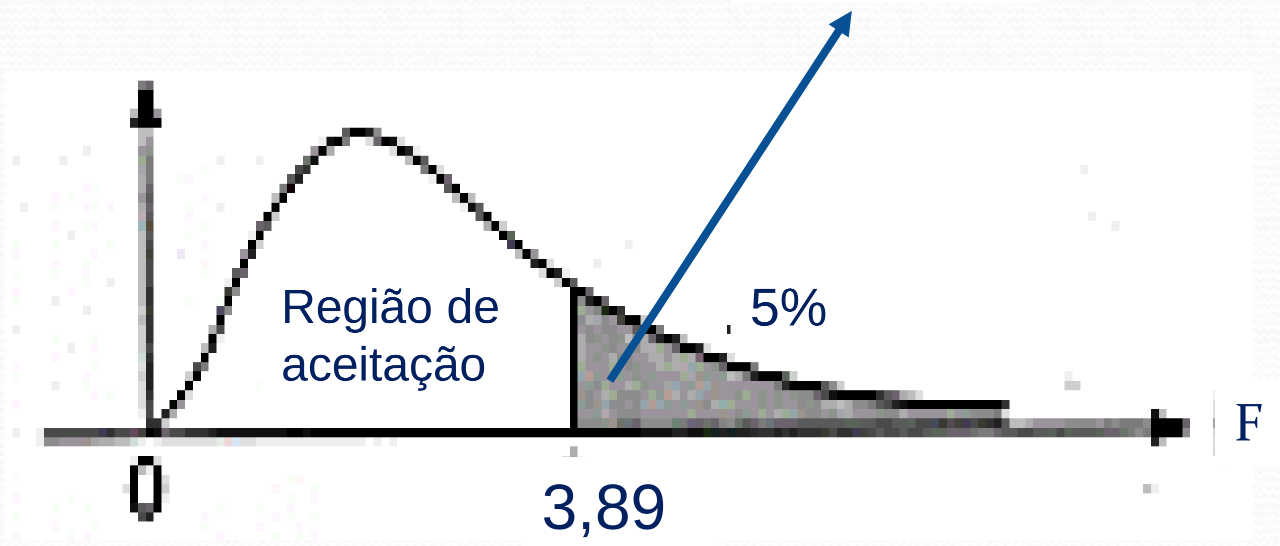
Vendedores		
A	**B**	**C**
29	27	30
27	27	30
31	30	31
29	28	27
32		29
30		

ANOVA em Excel, BioStat entre outros

Analysis of Variance (One-Way)						
Summary						
Groups	Sample size	Sum	Mean	Variance		
A	6	178,00000	29,6666667	3,0666667		
B	4	112,00000	28,00000	2,00000		
C	5	147,00000	29,400000	2,300000		
ANOVA						
Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	7,200000	2	3,600000	1,4148472	,2807335	3,8852938
Within Groups	30,5333333	12	2,5444444			
Total	37,7333333	14				

Anova: fator único						
RESUMO						
Grupo	Contagem	Soma	Média	Variância		
A	6	178	29,66667	3,066667		
B	4	112	28	2		
C	5	147	29,4	2,3		
ANOVA						
Fonte da variação	SQ	gl	MQ	F	valor-P	F crítico
Entre grupos	7,2	2	3,6	1,414847	0,280734	3,885294
Dentro dos grupos	30,53333	12	2,544444			
Total	37,73333	14				

F tabelado (2; 12) = 3,89



Compara-se F calculado com F tabelado, obtendo-se a conclusão:

F calculado = 1,41

F tabelado = 3,89

Como Fcal < F tab, não se rejeita a hipótese nula, concluindo, com nível de 5% que não há diferença na eficiência dos vendedores.

- A entrada dos dados no R pode ocorrer da seguinte maneira:

```
vendedores<-c(29,27,30,
              27,27,30,
              31,30,31,
              29,28,27,
              32,NA,29,
              30,NA,NA) # entrando com os dados
```

- Agora, criando os nomes dos tratamentos na ordem correspondente, tem-se:

```
> vended<-factor(rep(paste("vend",1:3,sep = ""),6))
> vended
 [1] vend1 vend2 vend3 vend1 vend2 vend3 vend1 vend2 vend3 vend1
[11] vend2 vend3 vend1 vend2 vend3 vend1 vend2 vend3
Levels: vend1 vend2 vend3
```

- Em todos os tipos de análise de variância, para todas as variáveis qualitativas, devem ser criados fatores e não vetores, ou seja, o objeto que contém os nomes (ou números) dos tratamentos, dos blocos etc. devem ser fatores e não vetores.

- Para criar fatores ou para a conversão de um vetor em um fator podemos usar as funções factor( ) ou as.factor( )

- Fazendo a análise de variância:

```
> resultado<-aov(vendedores~vended)
> resultado
Call:
   aov(formula = vendedores ~ vended)

Terms:
                   vended Residuals
Sum of Squares   7.20000  30.53333
Deg. of Freedom        2        12

Residual standard error: 1.595131
Estimated effects may be unbalanced
3 observations deleted due to missingness
```

- Note que o resultado é bem diferente do quadro da ANOVA. Para exibir o quadro da ANOVA, faça:

```
> anova(resultado)
Analysis of Variance Table

Response: vendedores
          Df Sum Sq Mean Sq F value Pr(>F)
vended     2  7.200  3.6000  1.4148 0.2807
Residuals 12 30.533  2.5444
```

A ANOVA pode ser interpretada da seguinte maneira: como o *p-value* (0,2807) foi maior que 5%, então não existe diferença significativa entre as médias de vendas feitas pelos vendedores.

ANOVA em Excel, BioStat entre outros

Analysis of Variance (One-Way)						
Summary						
Groups	Sample size	Sum	Mean	Variance		
A	6	178,00000	29,6666667	3,0666667		
B	4	112,00000	28,00000	2,00000		
C	5	147,00000	29,400000	2,300000		
ANOVA						
Source of Variation	SS	df	MS	F	p-value	F crit
Between Groups	7,200000	2	3,600000	1,4148472	,2807335	3,8852938
Within Groups	30,5333333	12	2,5444444			
Total	37,7333333	14				

2. Delineamento em blocos causalizados

- Este delineamento é bastante utilizado quando há heterogeneidade nas condições experimentais.

- Nesse caso, divide-se o material experimental, ou amostras, em blocos homogêneos, de forma a contemplar as diferenças entre os grupos.

- O que importa é a homogeneidade dentro de cada grupo e não entre os grupos.

- A ANOVA associada a este modelo de experimento é também conhecida como *Two Way ANOVA* (classificação dupla ou experimento com dois fatores).

Exemplo: Em uma experiência agrícola, foram usados cinco diferentes fertilizantes em duas variedades de trigo. A produção está indicada a seguir, em sacos. Verificar ao nível de 5% se:

a) Há diferença na produção devido ao fertilizante.
b) Há diferença na safra devido à variedade do trigo.

	Variedade de trigo	
Fertilizantes	**1**	**2**
A	54	57
B	38	42
C	46	45
D	50	53
E	44	50

* Os blocos são os fertilizantes e os tratamentos são as variedades de trigo. Criando o vetor de dados, blocos e tratamentos, temos:

```
dad<-c(54,57,
       38,42,
       46,45,
       50,53,
       44,50)
bloc<-gl(5,2,label=c(paste("fertilizante",LETTERS[1:5])))
trat<-rep(paste("var. trigo",1:2),5)
```

- Agora vamos criar um data.frame contendo todos os dados:

```
> tabela<-data.frame(blocos=bloc,
+                    tratam=factor(trat),
+                    dados=dad)
> tabela
          blocos        tratam dados
1  fertilizante A var. trigo 1    54
2  fertilizante A var. trigo 2    57
3  fertilizante B var. trigo 1    38
4  fertilizante B var. trigo 2    42
5  fertilizante C var. trigo 1    46
6  fertilizante C var. trigo 2    45
7  fertilizante D var. trigo 1    50
8  fertilizante D var. trigo 2    53
9  fertilizante E var. trigo 1    44
10 fertilizante E var. trigo 2    50
```

- Com o objeto contendo os dados devidamente criado, podemos proceder à ANOVA.

- Como visto para o caso do DIC, o comando que gera a análise de variância é o aov( ) , e o que exibe o quadro da ANOVA é o anova( ).

- Podemos proceder da seguinte forma:

```
>
> result<-aov(dados~tratam+blocos,
+             tabela)
> result
Call:
   aov(formula = dados ~ tratam + blocos, data = tabela)

Terms:
                tratam blocos Residuals
Sum of Squares    22.5  279.4      13.0
Deg. of Freedom      1      4         4

Residual standard error: 1.802776
Estimated effects may be unbalanced
>
> anova(result)
Analysis of Variance Table

Response: dados
          Df Sum Sq Mean Sq F value   Pr(>F)
tratam     1   22.5   22.50  6.9231 0.058115 .
blocos     4  279.4   69.85 21.4923 0.005754 **
Residuals  4   13.0    3.25
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

3. Delineamento em quadrado latino

- É como se fizéssemos um DBC com dois grupos de blocos, um horizontal (linhas) e outro vertical (colunas).

- É um modelo muito utilizado quando desejamos controlar duas fontes de variação sistemáticas conhecidas.

- Exemplo: a tabela abaixo resume os dados de teste de desgaste para 4 marcas diferentes de pneus, testadas em 4 carros diferentes, considerando a posição dos pneus:

Desgaste / Posição		Marcas A	B	C	D
Carros	I	3(17)	2(14)	1(12)	4(13)
	II	4(14)	3(14)	2(12)	1(11)
	III	1(13)	4(13)	3(11)	2(10)
	IV	2(13)	1(8)	4(9)	3(9)

```r
d<-c(17,14,12,13,
     14,14,12,11,
     13,13,11,10,
     13,8,9,9) #dados

li<-gl(4,4,label=c(paste("carro",(1:4))))
co<-factor(rep(paste("marca de pneu",LETTERS[1:4]),4))
tr<-factor(rep(paste("posição",c("3","2","1","4",
                                 "4","3","2","1",
                                 "1","4","3","2",
                                 "2","1","4","3"))))
tab<-data.frame(
  coluna=co,
  linha=li,
  tratamento=tr,
  dados=d
)
tab
```

```
> tab
              coluna   linha tratamento dados
1  marca de pneu A carro 1  posição 3    17
2  marca de pneu B carro 1  posição 2    14
3  marca de pneu C carro 1  posição 1    12
4  marca de pneu D carro 1  posição 4    13
5  marca de pneu A carro 2  posição 4    14
6  marca de pneu B carro 2  posição 3    14
7  marca de pneu C carro 2  posição 2    12
8  marca de pneu D carro 2  posição 1    11
9  marca de pneu A carro 3  posição 1    13
10 marca de pneu B carro 3  posição 4    13
11 marca de pneu C carro 3  posição 3    11
12 marca de pneu D carro 3  posição 2    10
13 marca de pneu A carro 4  posição 2    13
14 marca de pneu B carro 4  posição 1     8
15 marca de pneu C carro 4  posição 4     9
16 marca de pneu D carro 4  posição 3     9
```

```
> anova(aov(dados~coluna+linha+tratamento,
+           tab))
Analysis of Variance Table

Response: dados
           Df Sum Sq Mean Sq F value   Pr(>F)
coluna      3 30.688 10.2292 12.5897 0.005337 **
linha       3 38.688 12.8958 15.8718 0.002934 **
tratamento  3  6.688  2.2292  2.7436 0.135342
Residuals   6  4.875  0.8125
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ho rejeitada para carros e marcas. Posição não afeta desgaste.

4. Experimentos fatoriais

- São aqueles em que dois ou mais fatores são estudados simultaneamente. Cada um deles pode possuir dois ou mais níveis.

- A vantagem desse tipo de experimento é que além de termos o controle dos fatores individualmente, consideramos a interação entre eles.

- Em outras palavras, podemos estudar se os fatores atuam de forma independente ou se existem interações entre eles.

- Os experimentos fatoriais podem ser conduzidos segundo o DIC, DBC ou outros modelos

4.1 Experimentos com dois fatores segundo o DIC

- **Exemplo:** Um engenheiro agrimensor resolve estudar os efeitos da ***distância* e do *ângulo*** *de visada* ao alvo nos desvios radiais (valores absolutos em milímetros) encontrados nas observações (em relação a média).

- Então ele decide fazer um experimento fatorial segundo um DIC com duas repetições, conforme representado a seguir:

	Âng 1		Âng 2		Âng 3		Âng 4	
Dist 1	0,7	0,5	1,0	1,3	1,0	0,9	0,9	0,9
Dist 2	1,5	1,6	2,0	1,2	1,2	1,3	1,6	1,2
Dist 3	0,8	1,2	1,9	0,6	1,6	1,1	1,3	1,0

O engenheiro deseja saber se os fatores ***distância* e *ângulo de visada*** atuam independentemente e se suas influências são significativas ou não nos desvios radiais, a 5% de significância.

A ANOVA pode ser assim montada:

```
des.ra<-c(0.7,0.5,1.0,1.3,1.0,0.9,0.9,0.9,
          1.5,1.6,2.0,1.2,1.2,1.3,1.6,1.2,
          0.8,1.2,1.9,0.6,1.6,1.1,1.3,1.0) #observações
d<-gl(3,8,label=c(paste("Dist",1:3))) #distâncias
a<-rep(gl(4,2,label=c(paste("Ang",1:4))),3)#ângulos
dados<-data.frame(dist=d,ang=a,des.ra) #data.frame
```

```
> dados
      dist   ang des.ra
1  Dist 1 Ang 1    0.7
2  Dist 1 Ang 1    0.5
3  Dist 1 Ang 2    1.0
4  Dist 1 Ang 2    1.3
5  Dist 1 Ang 3    1.0
6  Dist 1 Ang 3    0.9
7  Dist 1 Ang 4    0.9
8  Dist 1 Ang 4    0.9
9  Dist 2 Ang 1    1.5
10 Dist 2 Ang 1    1.6
11 Dist 2 Ang 2    2.0
12 Dist 2 Ang 2    1.2
13 Dist 2 Ang 3    1.2
14 Dist 2 Ang 3    1.3
15 Dist 2 Ang 4    1.6
16 Dist 2 Ang 4    1.2
17 Dist 3 Ang 1    0.8
18 Dist 3 Ang 1    1.2
19 Dist 3 Ang 2    1.9
20 Dist 3 Ang 2    0.6
21 Dist 3 Ang 3    1.6
22 Dist 3 Ang 3    1.1
23 Dist 3 Ang 4    1.3
24 Dist 3 Ang 4    1.0
```

```
> anova(aov(des.ra~dist+ang+dist:ang,dados)) #ANOVA
Analysis of Variance Table

Response: des.ra
          Df  Sum Sq Mean Sq F value  Pr(>F)
dist       2 1.21083 0.60542  4.6127 0.03266 *
ang        3 0.24792 0.08264  0.6296 0.60974
dist:ang   6 0.34583 0.05764  0.4392 0.83909
Residuals 12 1.57500 0.13125
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

O quadro da ANOVA mostra que *distância* e *ângulo* atuam independentemente, uma vez que a interação representada por **dist:ang** não foi significativa *(p-value de 0,8309*).

- Podemos verificar também que houve diferença significativa apenas quanto ao fator *distância (p-value* de 0,03266), a 5% de significância.

- O restante da variação encontrada nos valores do erro deu-se ao acaso.

4.2 Fatorial usando o DBC

- Um engenheiro agrimensor quer testar diferentes *modelos de alvo,* a fim de avaliar se o desenho afeta a precisão, e também avaliar o *ângulo de visada ao alvo.*

- Conhecendo-se previamente a heterogeneidade nas *distâncias,* ele as separa em blocos.

- Foram selecionados três diferentes *modelos de alvo* (A, B e C) e coletados os valores para o desvio radial de cada observação (valores, absolutos em milímetros), conforme a seguir representados:

Dist 1	**Alvo A**	**Alvo B**	**Alvo C**
Ang 1	0,2	0,7	0,8
Ang 2	0,4	0,8	0,9
Ang 3	0,5	1,2	1,2

- Foram selecionados três diferentes *modelos de alvo* (A, B e C) e coletados os valores para o desvio radial de cada observação (valores, absolutos em milímetros), conforme a seguir representados:

Dist 2	Alvo A	Alvo B	Alvo C
Ang 1	0,6	0,8	1,1
Ang 2	0,9	1,3	1,5
Ang 3	1,2	1,4	1,8

- Foram selecionados três diferentes *modelos de alvo* (A, B e C) e coletados os valores para o desvio radial de cada observação (valores, absolutos em milímetros), conforme a seguir representados:

Dist 3	**Alvo A**	**Alvo B**	**Alvo C**
Ang 1	0,7	1,1	1,5
Ang 2	0,9	1,5	1,7
Ang 3	1,2	1,7	1,5

- O engenheiro deseja saber ainda se os três diferentes *modelos de alvo* influenciam significativamente o desvio radial, ou seja, se existem alvos melhores que outros, a 5% de significância.

```r
des.ra<-c(0.2,0.7,0.8,
          0.4,0.8,0.9,
          0.5,1.2,1.2,

          0.6,0.8,1.1,
          0.9,1.3,1.5,
          1.2,1.4,1.8,

          0.7,1.1,1.5,
          0.9,1.5,1.7,
          1.2,1.7,1.5)

d<-gl(3,9,label=c(paste("DIST",1:3))) #distâncias
al<-rep(paste("Alvo",LETTERS[1:3]),9) #alvos
an<-rep(gl(3,3,label=c(paste("Ang",1:3))),3) #ângulos

dados1<-data.frame(  #montando o data.frame
  bloco=d,           #blocos=distâncias
  alvo=al,           #modelo do alvo=fator 1
  ang=an,            #ângulo de visada=fator 2
  des.ra)            #observações

dados1
```

```
> anova(aov(des.ra~bloco+alvo+ang+alvo:ang,dados1))  #ANOVA
Analysis of Variance Table

Response: des.ra
          Df  Sum Sq Mean Sq F value
bloco      2 1.58000 0.79000 42.1333
alvo       2 1.72667 0.86333 46.0444
ang        2 0.98667 0.49333 26.3111
alvo:ang   4 0.03333 0.00833  0.4444
Residuals 16 0.30000 0.01875
             Pr(>F)
bloco     4.204e-07 ***
alvo      2.305e-07 ***
ang       8.735e-06 ***
alvo:ang     0.7748
Residuals
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05
  '.' 0.1 ' ' 1
```

- Percebemos, de acordo com a tabela da ANOVA, que o *modelo do alvo* e o *ângulo de visada* atuam independentemente, uma vez que a interação entre eles foi "não significativa" *com valor p* ~ 0,77).

- Percebemos também que o fator *ângulo de visada ao alvo* é significativo.

- Porém, o mais importante é que, de acordo com a ANOVA, pode-se afirmar que existe diferença na eficiência dos diferentes *modelos de alvo,* no que diz respeito ao desvio radial, que era o principal objetivo do engenheiro agrimensor.

5. Exercícios