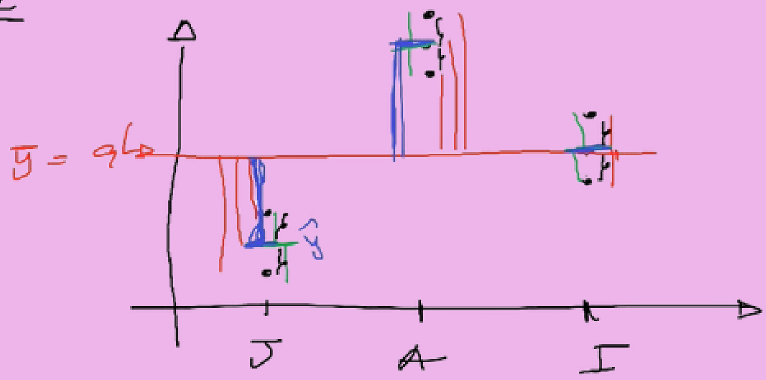
Resumo - TE - Aula 7

Na nossa sétima aula, fizemos os esclarecimentos das dúvidas e uma revisão no início da aula!

Revisando a história da patricinha!

gasto	perfil idade
4	Jovem
5	Jov
6	J
13	A
12	A
14	A
8	I
9	I
10	I



$$SQT = \Sigma(y-\bar{y})^2$$

$$SQE = \Sigma(y-\hat{y})^2$$

$$SQM = \Sigma(\hat{y}-\bar{y})^2$$

$$SQT = SQM + SQE$$

$$R^2 = \frac{SQM}{SQT}$$

Acompanhe o racional da aula! Provamos matematicamente que SQT = SQM + SQE

A	B	C	D	E	F	G	H	I	J
y	ybarra	erro para yba	SQT		ychapeu	erro para ychapeu			SQM
gasto	gastomedio	(y-ybarra)	(y-ybarra)^2	perfil_idade	chute sofisticado	(y - ychapeu)	(y - ychapeu)^2	(ychapeu - ybarra)	(ychapeu - ybarra)^2
4	9	-5	25	jovem	5	-1	1	-4	16
5	9	-4	16	jovem	5	0	0	-4	16
6	9	-3	9	jovem	5	1	1	-4	16
13	9	4	16	adulto	13	0	0	4	16
12	9	3	9	adulto	13	-1	1	4	16
14	9	5	25	adulto	13	1	1	4	16
8	9	-1	1	idoso	9	-1	1	0	0
9	9	0	0	idoso	9	0	0	0	0
10	9	1	1	idoso	9	1	1	0	0
		SQT	102			SQE	6	SQM	96
		variancia	20.4						

www.asn.rocks

Vamos conhecer algumas métricas de erro?

Mean Square Error

$$MSE = \frac{1}{n} \Sigma \left(y - \widehat{y} \right)^2$$

Root Mean Square Error

$$RMSE = \sqrt{\frac{1}{n} \Sigma \left(y - \widehat{y} \right)^2}$$

ychapeu chute sofisticado	erro para ychapeu (y - ychapeu)	(y - ychapeu)^2
5	-1	[illegible]
5	0	0
5	1	1
13	0	0
13	-1	1
13	1	1
9	-1	1
9	0	0
9	1	1
	SQE	6
	Mean Square Error	0.666666667
	Root MSE	0.816496581

Mean Absolute Error

$$MAE = \frac{1}{n} \sum_{i=1}^{n} \left| y_i - \widehat{y}_i \right|$$

SQM (ychapeu - ybarra)^2	modulo(y - ychapeu)
16	1
16	0
16	1
16	0
16	1
16	1
0	1
0	0
0	1
96	
Mean Absolut Error	0.666666667

Discutimos sobre a utilização de cada uma delas e a existência de outras métricas igualmente válidas, como o MAPE!

data science analytics
transformando negócios

ANOVA 1 fator

É um teste para a comparação de k médias, quando temos somente um fator (uma variável x categórica) para entendermos um y numérico.
Lembrando que esse é um teste paramétrico que exige homogeneidade de variâncias e normalidade das amostras!!

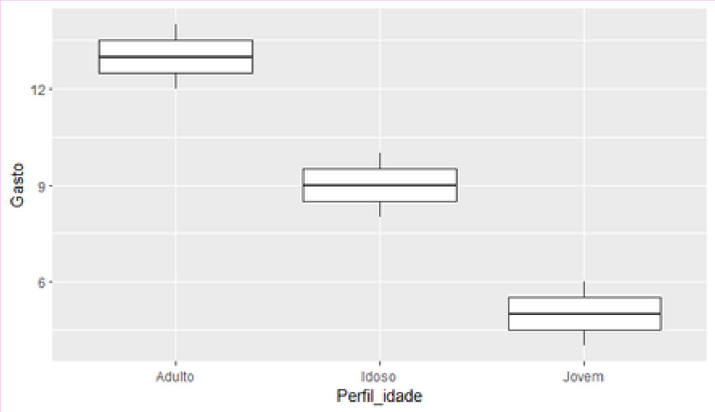
$$\begin{cases} H_0: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_1: pelo\ menos\ um\ é\ \neq \end{cases}$$

Faremos um exemplo em R com a base dados ANOVA_1fator:

```
ANOVA_1fator <- read_excel("dados/ANOVA_1fator.xlsx")
```

Plotando o boxplot para cada perfil_idade:

```
ANOVA_1fator %>%
  ggplot()+
  geom_boxplot(aes(x= Perfil_idade, y=Gasto))
```



Conseguimos notar que as variação dos três perfis são as mesmas, o que muda são as médias. Vamos fazer a ANOVA para confirmar?

```
anova_1fator = aov(Gasto ~ Perfil_idade, data=ANOVA_1fator)
summary(anova_1fator)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)    
Perfil_idade  2     96      48      48 0.000204 ***
Residuals     6      6       1                     
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Observe que na saída do R obtemos as métricas vistas anteriormente, SQT, SQM e SQE, mas vamos analisar o p-valor? Rejeitamos H0 ou não? O que a decisão significa?

ANOVA 2 fatores ou fatorial

É um teste para a comparação de k médias, quando temos dois ou mais fatores (variáveis x's categóricas) para entendermos um y numérico.
Lembrando que esse é um teste paramétrico que exige homogeneidade de variâncias e normalidade das amostras!!

$$\begin{cases} H_0^A: \mu_1 = \mu_2 = \cdots = \mu_a \\ H_1^A: pelo\ menos\ um\ é\ \neq \end{cases}$$

$$\begin{cases} H_0^B: \mu_1 = \mu_2 = \cdots = \mu_b \\ H_1^B: pelo\ menos\ um\ é\ \neq \end{cases}$$

$$\begin{cases} H_0^A: \gamma_{ij} = 0\ (não\ há\ interação\ entre\ A\ e\ B) \\ H_1^A: \gamma_{ij} \neq 0 (há\ interação\ entre\ A\ e\ B) \end{cases}$$

Voltando ao nosso exemplo, agora teremos dois fatores, o perfil_idade e o gênero!
Podemos fazer a concatenação dessas duas variáveis para tentar entender se a interação entre elas ajuda na estimação do y. Assim, podemos fazer três testes:

$$\text{fator 1}\begin{cases} H_0: \mu_J = \mu_A = \mu_I \\ H_1: \text{pelo menos } 1 \neq \end{cases}$$

$$\text{fator 2}\begin{cases} H_0: \mu_F = \mu_M \\ H_1: \mu_F \neq \mu_M \end{cases}$$

$$\begin{matrix}\text{interação} \\ \text{entre} \\ \text{vars}\end{matrix}\begin{cases} H_0: \mu_{JF} = \mu_{JM} = \mu_{AM} = \ldots \\ H_1: \text{pelo menos } 1 \neq \end{cases}$$

Faremos um exemplo em R com a base dados ANOVA_2fatores:

```
df_anova_2fatores <- read_excel("dados/ANOVA_2fatores.xlsx",
                                col_types = c("numeric", "text", "text"))
```

```
anova_2fatores = aov(gastos ~ sexo*renda, data=df_anova_2fatores)
summary(anova_2fatores)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)    
sexo          1  205.4   205.4  15.572 0.000231 ***
renda         2 2426.4  1213.2  92.000  < 2e-16 ***
sexo:renda    2  108.3    54.2   4.107 0.021860 *  
Residuals    54  712.1    13.2                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Repare que nesse teste resultaram 3 p-valores diferentes, pois utilizamos a sintaxe para a interação ser incluída (x1*x2).
Mas e aí, como você interpretaria cada um dos p-valores?

Sem acompanhar o racional da aula, não gera sentimento!!
Assista aula!!

Lembre-se da tarefa de casa!!