FIAP

Statistics Fundamentals – Turma 1DTS
Trabalho 4

Habilidades desenvolvidas: Teste de hipóteses.Teste Qui-Quadrado. Correlação de Pearson.

Parte 1

1) Utilize as tabelas, os gráficos e a saída do RStudio para responder as questões:

a) Qual a diferença entre as Tabelas 1 e 2?

A diferença está no percentual linha (Tabela 1) e percentual coluna (Tabela 2). Tanto é que o título dos gráficos também são diferentes.

b) Existe associação entre Produto e Região?
Hipótese do analista:

Tabela 1:
Produto A o percentual da média é diferente do percentual da média total na região X;

Percentual da média do produto A na região X = 15%
Percentual da média total = 11%

Produto B e C o percentual da média é diferente do percentual da média total na região Y.

Percentual da média do produto B na região Y = 42%
Percentual da média total = 41%

Percentual da média do produto C na região Y = 50%
Percentual da média total = 48%

Tabela 2:

Produto A o percentual da média é diferente do percentual da média total na região X;

Percentual da média do produto A na região X = 60%
Percentual da média total = 45%

Produto B e C o percentual da média é diferente do percentual da média total na região Y.

Percentual da média do produto B na região Y = 56%
Percentual da média total = 55%

Percentual da média do produto C na região Y = 57%
Percentual da média total = 55%

c) Apresente as hipóteses H0, H1 e o nível de significância do teste (erro).

H0: independentes
H1: dependentes
Erro de decisão: 5%

d) Use o Teste Qui-quadrado para verificar a associação. Apresente a Tabela do valor esperado e calcule o ^{2}.

Tabela do valor observado

Produto		Região			Total	
	X	%	Y	%		%
A	300	60	200	40	500	100
B	800	44	1000	56	1800	100
C	900	43	1200	57	2100	100
Total	2000	45	2400	55	4400	100
		0.454545		0.545455		

Tabela do valor esperado

Produto	Região		Total
	X	Y	
A	**227**	**273**	500
B	**818**	**982**	1800
C	**955**	**1145**	2100
Total	2000	2400	4400

Cálculo da estatística ^{2}:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{k}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

a= 0.454545*500 = **227,2727273**
b= 0.454545*1800 = **818,1818182**
c= 0.454545*2100 = **954,5454545**
d= 0.545455*500 = **272,7272727**
e= 0.545455*1800 = **981,8181818**
f= 0.545455*2100 = **1145,454545**

Produto	X	Y	Total
A	**227,2727273**	**272,7272727**	500
B	**818,1818182**	**981,8181818**	1800
C	**954,5454545**	**1145,454545**	2100
Total	2000	2400	4400

e) Qual a conclusão?

Tabela 1 – Distribuição de vendas segundo produto e região. 2018

Produto	Região				Total	
	X		Y			
	N	%	N	%	N	%
A	300	15	200	8	500	11
B	800	40	1000	42	1800	41
C	900	45	1200	50	2100	48
Total	2000	100	2400	100	4400	100

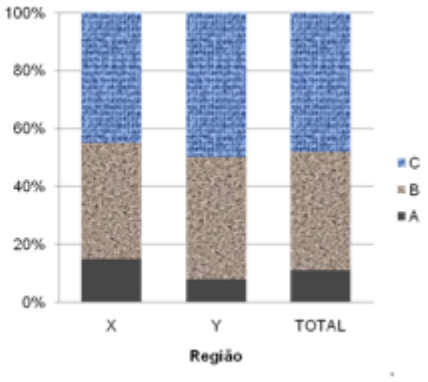
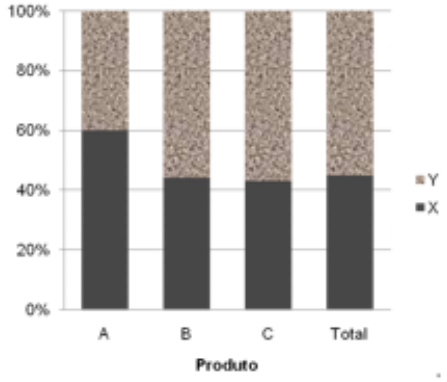


Tabela 2 – Distribuição de vendas segundo região e produto. 2018

Produto	Região				Total	
	X		Y			
	N	%	N	%	N	%
A	300	60	200	40	500	100
B	800	44	1000	56	1800	100
C	900	43	1200	57	2100	100
Total	2000	45	2400	55	4400	100

Fonte:zzzz



O produto A está associado à região X e os produtos B e C estão associados à região Y, pois o percentual da média do produto A na região X é diferente do que o percentual da média total e o percentual da média dos produtos B e C na região Y são diferentes do percentual da média total.

Saída do Teste Qui-quadrado do RStudio:

```
> x<-matrix(c(300,800,900,200,1000,1200),nc=2)
> x
     [,1] [,2]
[1,]  300  200
[2,]  800 1000
[3,]  900 1200
> chisq.test(x)

        Pearson's Chi-squared test

data:  x
X-squared = 49.122, df = 2, p-value = 2.155e-11
```

p-value = 0.00000000002155

FIAP

Células	O	E	O-E	O-E^2	O-E^2/E
a	300	227,2727273	72,72727273	5289,256198	23,27272727
b	800	818,1818182	-18,18181818	330,5785124	0,404040404
c	900	954,5454545	-54,54545455	2975,206612	3,116883117
d	200	272,7272727	-72,72727273	5289,256198	19,39393939
e	1000	981,8181818	18,18181818	330,5785124	0,336700337
f	1200	1145,454545	54,54545455	2975,206612	2,597402597
Somatória					**49,12**

2) Utilize a tabela, o gráfico e a saída do RStudio para responder as questões:

a) Existe associação entre Resposta e Carta?

Hipótese do analista:

O percentual de média de resposta "Sim" na carta "Não" é diferente do percentual de média total.

Percentual da média da resposta "Sim" na cartão "Não" = 52.6%
Percentual da média total = 51.9%

O percentual de média de resposta "Não" na carta "Sim" é diferente do percentual de média total.

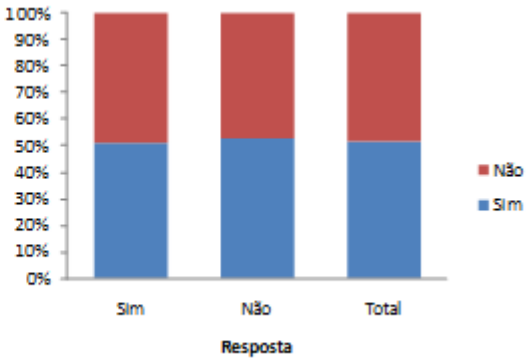
Percentual da média da resposta "Não" na cartão "Sim" = 48.8%
Percentual da média total = 48.1%

b) Apresente as hipóteses H0, H1 e o nível de significância do teste (erro).

H0: independentes
H1: dependentes
Erro de decisão: 5%

c) Use o Teste Qui-quadrado para verificar a associação.
d) Qual a conclusão?
.

Resposta	Carta				Total	
	Sim		Não			
	N	%	N	%	N	%
Sim	2570	51.2	2645	52.6	5215	51.9
Não	2448	48.8	2384	47.4	4832	48.1
Total	5018	100.0	5029	100.0	10047	100.0



Saída do RStudio:

```
> dados1 <-matrix(c(2570,2448,2645,2384),nc=2)
> dados1
     [,1] [,2]
[1,] 2570 2645
[2,] 2448 2384
```

```
> # Calcule o teste qui-quadrado
> chisq.test(dados1)

        Pearson's Chi-squared test with Yates' continuity correction

data:  dados1
X-squared = 1.8594, df = 1, p-value = 0.1727
```

Cálculo da estatística ^{2} com correção de continuidade:

$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Onde:

O_i = uma frequência observada

E_i = uma frequência esperada (teórica), afirmada pela hipótese nula

N = número de eventos distintos

Essa tabela é uma matriz 2,2, sendo assim o grau de liberdade igual a 1, usando a tabela do chi quadrado o valor de teste é 3.841, no resultado do chi quadrado na saída R é 1.8594. Dessa forma, podemos dizer que aceitamos H0, logo as variáveis (Resposta e Carta) não tem relação.

Parte 2)

1) Use os dados *Bike_Sharing.csv* para construir as análises descritivas, correlação, associação e modelo preditivo para previsão do número de bikes alugadas por mês.

Descrição:

FIAP

Os sistemas de compartilhamento de bicicletas são uma nova geração de aluguel de bicicletas tradicionais, onde todo o processo de associação, locação e devolução tornou-se automático. Através destes sistemas, o usuário pode facilmente alugar uma bicicleta a partir de uma determinada posição e retornar em outra posição. Atualmente, existem cerca de 500 programas de compartilhamento de bicicletas em todo o mundo, compostos por mais de 500 mil bicicletas. Hoje, existe um grande interesse nestes sistemas devido ao seu importante papel no trânsito, questões ambientais e de saúde.

Fonte de dados: https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset

Atividades:

a) Classifique o tipo de variável

Variável	Descrição	Tipo de Variável
instant	record index	ID
dteday	date	ID
season	season (~~1:springer, 2:summer, 3:fall, 4:winter~~) 1:winter, 2:spring, 3:summer, 4:fall	Qualitativa Ordinal
yr	year (0: 2011, 1:2012)	Qualitativa Ordinal
mnth	month (1 to 12)	Qualitativa Ordinal
hr	hour (0 to 23)	Qualitativa Ordinal
holiday	weather day is holiday or not	Qualitativa Nominal
weekday	day of the week	Qualitativa Ordinal
workingday	if day is neither weekend nor holiday is 1, otherwise is 0.	Qualitativa Nominal
weathersit	1: Clear, Few clouds, Partly cloudy, Partly cloudy; 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist;3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds;4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog	Qualitativa Nominal
temp	Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)	Quantitativa Contínua
atemp	Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)	Quantitativa Contínua

hum	Normalized humidity. The values are divided to 100 (max)	Quantitativa Contínua
windspeed	Normalized wind speed. The values are divided to 67 (max)	Quantitativa Contínua
casual	count of casual users	Quantitativa Discreta
registered	count of registered users	Quantitativa Discreta
cnt	count of total rental bikes including both casual and registered	Quantitativa Discreta

b) Quais variáveis foram normalizadas? Por quê? Apresente a fórmula utilizada.

As variáveis são temp, atemp, hum e windspeed, o motivo para padronização é aproximação das escalas.

$$X_p = \frac{X - X_{mínimo}}{X_{máximo} - X_{mínimo}}$$

c) Apresenta a análise de associação e correlação de Pearson. Quais variáveis têm correlação com a variável resposta?

a. TEMP

i.

```
> cor.test(arquivo$cnt,arquivo$temp  , method=c("pearson"))

        Pearson's product-moment correlation

data:  arquivo$cnt and arquivo$temp
t = 21.759, df = 729, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5814369 0.6695422
sample estimates:
      cor
0.627494
```

b. ATEMP

i.

```
> cor.test(arquivo$cnt,arquivo$atemp, method=c("pearson"))

        Pearson's product-moment correlation

data:  arquivo$cnt and arquivo$atemp
t = 21.965, df = 729, p-value < 0.00000000000000022
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5853376 0.6727918
sample estimates:
      cor
0.6310657
```

1) Existe associação entre a demanda de bike e estação do ano? Apresente a tabela do valor observado, a tabela do valor esperado e o valor da estatística o ^{2}. Qual a conclusão do teste?

R: Sim, existe associação.

a. Tabela do valor observado:

Faixa aluguel de bikes	Estações do ano								
	1	2	3	4	1%	2%	3%	4%	Total
22-3152	122	28	4	28	0,67033	0,153846	0,021978	0,153846	182
3152-4548	43	43	41	56	0,234973	0,234973	0,224044	0,306011	183
4548-5956	11	51	64	56	0,06044	0,28022	0,351648	0,307692	182
5956-8714	5	62	79	37	0,027322	0,338798	0,431694	0,202186	183
Total	**181**	**184**	**188**	**177**	0,247945	0,252055	0,257534	0,242466	**730**

b. Tabela do valor esperado:

Faixa aluguel de bikes	Estações do ano				
	1	2	3	4	Total
22-3152	45,12603	45,87397	46,87123	44,12877	182
3152-4548	45,37397	46,12603	47,12877	44,37123	183
4548-5956	45,12603	45,87397	46,87123	44,12877	182
5956-8714	45,37397	46,12603	47,12877	44,37123	183

c. `X-squared = 287.21, df = 9, p-value < 0.00000000000000022`

d. Rejeita hipótese H0 e aceita H1.

2) Existe associação entre CNT e dia da semana? Apresente a tabela do valor observado, a tabela do valor esperado e o valor da estatística o ^{2}. Qual a conclusão do teste?

a. Não existe associação

b. Tabela de valores observados:

Faixa aluguel de bikes	Dias da semana							
	0	1	2	3	4	5	6	Total
22-3152	29	24	24	29	25	20	31	182
3152-4548	27	34	27	21	23	29	22	183
4548-5956	27	21	30	27	25	29	23	182
5956-8714	22	25	23	27	31	26	29	183
Total	**105**	**104**	**104**	**104**	**104**	**104**	**105**	**730**

0%	1%	2%	3%	4%	5%	6%
0,159340659	0,131868132	0,131868132	0,159340659	0,137362637	0,10989011	0,17032967
0,147540984	0,18579235	0,147540984	0,114754098	0,12568306	0,158469945	0,120218579
0,148351648	0,115384615	0,164835165	0,148351648	0,137362637	0,159340659	0,126373626
0,120218579	0,136612022	0,12568306	0,147540984	0,169398907	0,142076503	0,158469945
0,143835616	**0,142465753**	**0,142465753**	**0,142465753**	**0,142465753**	**0,142465753**	**0,143835616**

c. Tabela dos valores esperados:

Faixa aluguel de bikes	Dias da semana						
	0	1	2	3	4	5	6
22-3152	26,17808	25,92876712	25,92876712	25,92876712	25,92877	25,92877	26,17808
3152-4548	26,32192	26,07123288	26,07123288	26,07123288	26,07123	26,07123	26,32192
4548-5956	26,17808	25,92876712	25,92876712	25,92876712	25,92877	25,92877	26,17808
5956-8714	26,32192	26,07123288	26,07123288	26,07123288	26,07123	26,07123	26,32192

d. `X-squared = 12.863, df = 18, p-value = 0.7997`

e. Aceita H0 e rejeita H1.

3) Apresente a matriz de correlação das variáveis quantitativas. Interprete os resultados.

```
                cnt       temp      atemp        hum  windspeed
cnt        1.0000000  0.6274940  0.6310657 -0.1006586 -0.2345450
temp       0.6274940  1.0000000  0.9917016  0.1269629 -0.1579441
atemp      0.6310657  0.9917016  1.0000000  0.1399881 -0.1836430
hum       -0.1006586  0.1269629  0.1399881  1.0000000 -0.2484891
windspeed -0.2345450 -0.1579441 -0.1836430 -0.2484891  1.0000000
```

R: Observando a tabela, podemos interpretar que a variável "**cnt**" tem forte correlação com as variáveis "**temp**" e "**atemp"**. Podemos concluir que a quantidade de aluguel de bicicletas aumentam ou diminuem com base na condição do tempo e na sua sensação térmica.

Entrega do exercício no formato word.
Data de entrega: 03/06/2022
Regina Bernal
25/05/2022