Estatística para Ciência de Dados
Resolução do trabalho 06

Eduardo Façanha Dutra
Giovanni Brígido

Conteúdo

1 Enunciado

Você é um analista de dados no setor de compras da Casa Pio, uma grande rede de lojas de calçados.

Ela é uma empresa com sede em Fortaleza estabelecido há 70 anos. Atualmente, também opera em várias cidades do interior do Ceará, A empresa vende sapatos de preços médios, variando de 120 a 200 reais. Embora os sapatos sejam de alta qualidade, existem lotes e lotes no estoque que nunca são vendidos. Em outras palavras, os sapatos estragando nas prateleiras das lojas.

Por isto, o gerenciamento de estoque é um problema muito comum e importante.

Muitas, senão a maioria das lojas, não conseguem determinar número correto de itens que precisam manter em estoque. O problema oposto também surge: lojas sem estoque necessário de mercadorias não irão atender à demanda de mercado. Por exemplo, você certamente já entrou em uma loja de sapatos, mas não conseguiu comprar porque eles não tinham o número em estoque.

Neste estudo, vamos examinar o problema oposto - ter muito estoque.

Este é um problema mais significativo para a empresa, pois indica que investiu na produção ou compra de um produto, mas não foi capaz de vendê-lo.

Temos um banco de dados com as informações de vendas das lojas durante os anos 2014 a 2016, com várias informações uteis.

Nosso problema é estimar o volume de compra de sapatos para não ficar com estoque muito alto, nem perder muitas vendas.

Tarefa 1 Responda as seguintes perguntas:

```
1- Os dados fornecidos são amostrais ou populacionais?
2- Você separaria os dados em quantos grupos?
3- Esses grupos são dependentes ou independentes?
4- Como você agrupará os dados para ter uma melhor visão do problema a ser resolvido?
5- Os dados fornecidos representam uma distribuição Normal?
6- Qual o intervalo de confiança que será usado?
7- Qual a estatística será usada? Z ou t?
8- Com base na sua resposta a pergunta 7, qual a sua justificativa?
9- Quantos pares de cada sapato devemos ter em estoque?
```

Tarefa 2

Digamos que queremos usar um intervalo de confiança para ver se duas lojas estão vendendo o mesmo número de sapatos. Além disso, queremos saber com 90% confiança quanto uma loja supera a outra em termos de vendas. Você tem duas tabelas representando as vendas de calçados femininos em duas lojas. Seus códigos são ARA 1 e ARA 2. Temos dados para 2016.

Justifique suas escolhas e aplique uma estatística para determinar qual das lojas vende mais produtos, e quais.

Verifique com 90% de confiança, se ambas lojas podem ter um estoque único.

2 Resolução

2.1 Tarefa 1

1- Os dados fornecidos são amostrais ou populacionais?

Os dados fornecidos são amostrais, uma vez que se referem a apenas 3 anos, de 2014 a 2016 e Casa Pio já está estabelecida em Fortaleza há 70 anos.

2- Você separaria os dados em quantos grupos?

Separaria os dados por gênero, por tamanho de sapato, por mês e ano. Nesse caso específico foram utilizados os dados de todo Ceará.

3- Esses grupos são dependentes ou independentes?

Os grupos são independentes.

4- Como você agrupará os dados para ter uma melhor visão do problema a ser resolvido?

Os dados serão agrupados por meses, por ano, e por tamanho de sapato, em ordem crescente, sendo criadas duas tabelas, uma para cada gênero. Dessa maneira, obtém-se uma visão melhor do problema a ser resolvido.

Tabela 1: Vendas de sapatos masculinos de 2014 a 2016 em todo o Ceará, por tamanho.

	6	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12	13	14	15	Total
jan14	1	0	2	2	4	6	13	27	15	18	11	7	4	0	2	0	112
fev14	3	0	2	4	0	7	16	26	17	31	6	8	0	0	0	2	122
mar14	0	0	1	3	13	10	21	31	16	20	11	8	4	0	6	5	149
abr14	0	5	2	0	4	9	27	33	17	23	10	7	4	1	1	4	147
mai14	0	0	2	5	4	5	24	26	17	20	5	3	5	2	4	0	122
jun14	1	0	0	0	5	13	18	28	25	26	10	3	3	1	6	4	143
jul14	0	0	1	1	13	4	17	33	21	14	10	9	2	3	5	1	134
ago14	3	3	3	1	6	7	20	26	27	23	11	5	2	0	1	2	140
set14	0	3	1	1	8	11	22	27	22	31	8	2	3	5	2	4	150
out14	3	0	3	4	4	16	19	35	29	23	7	10	7	2	2	1	165
nov14	1	0	1	5	5	5	11	15	20	18	4	5	6	3	0	1	100
dez14	0	2	1	2	3	9	27	34	24	22	9	4	0	0	0	1	138
jan15	1	6	0	2	9	14	26	53	36	19	9	8	6	1	1	1	192
fev15	1	4	0	2	13	11	26	41	30	25	13	6	7	0	2	1	182
mar15	0	2	2	2	8	12	24	42	28	31	15	11	3	2	4	0	186
abr15	0	3	1	1	6	15	35	43	25	34	18	7	6	1	3	2	200
mai15	5	0	1	1	9	14	35	58	23	38	17	12	5	3	2	3	226
jun15	3	1	5	6	8	26	51	61	37	34	12	19	9	6	0	1	279
jul15	1	5	2	5	9	17	52	44	40	54	14	10	5	2	3	3	266
ago15	4	5	1	2	3	32	44	66	38	34	14	11	8	2	7	6	277
set15	7	1	5	9	10	20	40	57	39	43	19	8	12	0	1	4	275
out15	9	5	5	7	11	22	52	52	36	32	17	14	7	3	4	1	277
nov15	2	4	4	6	9	16	26	67	42	36	11	13	8	4	4	7	259
dez15	3	3	6	5	9	14	42	76	43	37	14	15	6	3	4	4	284
jan16	6	4	2	9	17	30	58	53	45	49	12	15	5	3	5	2	315
fev16	2	3	1	5	17	31	29	60	54	38	30	12	7	2	11	7	309
mar16	3	3	7	3	14	32	43	61	52	49	29	11	2	4	4	0	317
abr16	2	1	1	2	15	32	49	76	62	60	21	6	6	2	4	4	343
mai16	6	1	7	17	17	30	52	63	49	59	29	12	14	4	8	5	373
jun16	4	2	11	5	11	24	79	93	72	53	24	29	17	6	4	6	440

	6	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12	13	14	15	Total
jul16	3	3	9	14	23	41	72	77	69	62	32	17	10	2	7	10	451
ago16	7	11	4	6	10	43	57	91	70	57	19	18	25	2	4	3	427
set16	3	3	10	13	17	26	89	70	56	60	29	16	15	1	1	3	412
out16	14	4	4	12	16	34	60	96	63	60	19	30	10	5	6	2	435
nov16	6	2	4	3	8	22	32	73	50	45	18	17	4	7	5	6	302
dez16	1	1	3	6	16	18	39	76	38	30	15	11	3	5	3	5	270

Tabela 2: Vendas de sapatos femininos de 2014 a 2016 em todo o Ceará, por tamanho.

	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10	10.5	11	11.5	12	Total
jan14	0	0	0	3	1	4	22	32	9	15	19	2	1	0	0	0	108
fev14	2	0	1	0	0	7	3	23	10	16	12	4	4	0	0	0	82
mar14	0	1	0	1	0	11	4	20	5	17	9	6	1	0	0	0	75
abr14	0	0	2	3	2	13	3	22	23	12	13	9	4	0	0	0	106
mai14	0	2	0	2	2	13	4	16	39	25	6	7	0	0	0	0	116
jun14	0	1	0	1	3	9	6	22	12	17	0	11	0	0	0	0	82
jul14	0	0	1	0	7	8	5	16	18	13	5	10	0	1	1	0	85
ago14	0	0	0	0	4	12	2	30	17	12	4	13	5	2	0	0	101
set14	0	0	2	0	3	12	8	16	14	21	2	13	2	0	6	1	100
out14	3	0	0	0	0	17	10	26	16	15	8	14	0	0	0	0	109
nov14	0	0	0	0	2	14	3	14	15	11	11	9	2	0	0	0	81
dez14	2	1	1	0	2	14	3	20	10	19	4	10	0	0	0	0	86
jan15	0	1	0	1	7	8	21	25	14	20	5	8	3	0	0	0	113
fev15	5	0	2	6	7	4	23	33	14	13	9	5	2	0	3	0	126
mar15	1	0	0	0	5	6	30	29	20	19	15	8	9	0	0	0	142
abr15	0	0	2	7	9	7	10	23	15	22	17	7	8	0	0	0	127
mai15	1	0	0	3	16	9	21	33	29	19	13	2	9	2	1	6	164
jun15	6	0	0	1	11	13	30	36	39	30	13	0	8	0	0	0	187
jul15	1	2	0	0	9	12	27	42	29	28	19	6	5	1	4	2	187
ago15	0	3	3	0	6	8	28	28	30	28	13	2	15	2	3	0	169
set15	0	4	2	2	9	13	20	35	23	35	13	2	6	0	0	1	165
out15	1	0	2	2	9	12	24	46	31	25	17	10	11	0	1	2	193
nov15	0	0	4	1	6	11	27	40	41	31	18	4	5	0	1	6	195
dez15	2	0	7	2	9	15	22	40	35	22	10	5	2	6	0	0	177
jan16	1	4	5	0	15	10	31	48	34	38	14	5	8	8	2	1	224
fev16	1	0	0	12	10	12	35	57	31	31	10	11	14	3	9	4	240
mar16	2	0	0	6	15	9	25	43	45	31	18	5	4	2	9	0	214
abr16	0	0	2	12	10	21	47	58	28	52	20	6	7	5	6	0	274
mai16	1	1	1	7	5	12	28	57	52	48	18	4	4	2	6	2	248
jun16	11	1	3	1	16	19	44	36	42	41	9	7	9	2	4	5	250
jul16	4	4	5	0	18	8	31	59	69	49	16	6	11	1	0	9	290
ago16	3	4	1	1	6	14	29	71	49	51	22	8	7	1	2	10	279
set16	2	6	1	3	8	22	56	52	51	42	19	4	4	0	1	4	275
out16	1	1	3	2	4	18	50	66	45	30	17	2	12	2	0	15	268
nov16	1	0	4	0	9	12	27	52	43	28	18	2	9	2	1	9	217
dez16	0	0	9	2	7	14	21	36	41	28	20	4	7	3	0	1	193

5- Os dados fornecidos representam uma distribuição Normal?

De acordo com o Teorema do Limite Central, as médias de amostras grandes e aleatórias são aproximadamente normais. Logo, pode-se aplicar técnicas usadas em dados que seguem uma distribuição normal.

6- Qual o intervalo de confiança que será usado?

O intervalo de confiança será de 95%

7- Qual a estatística será usada? Z ou t?

A estatística que será usada será a t.

8- Com base na sua resposta a pergunta 7, qual a sua justificativa?

Como são três anos de dados e os dados estão divididos em meses, totalizando 36 meses de dados, optou-se pela distribuição t, uma vez que a amostra tem tamanho próximo de 30. Ademais, como a variância populacional é desconhecida, sugere-se o uso da estatística t.

9- Quantos pares de cada sapato devemos ter em estoque?

Aplicando uma estratégia agressiva de mercado, adotou-se o arredondamento para cima do limite superior do intervalo de confiança para definir quantos pares de sapato deve-se ter em estoque, para cada mês, em média.

Tabela 3: Intervalo de confiança para as vendas de 2014 a 2016 de calçados para homens.

	Média	ErroPadrão	MargemErro	-95% CI	+95% CI	NumEscolhido
6	2.92	0.51	1.04	1.88	3.96	4
6.5	2.50	0.39	0.79	1.71	3.29	4
7	3.17	0.47	0.95	2.22	4.12	5
7.5	4.75	0.68	1.38	3.37	6.13	7
8	9.83	0.87	1.77	8.06	11.60	12
8.5	18.83	1.78	3.61	15.22	22.44	23
9	37.42	3.18	6.46	30.96	43.88	44
9.5	52.50	3.61	7.33	45.17	59.83	60
10	37.42	2.79	5.66	31.76	43.08	44
10.5	36.33	2.43	4.93	31.40	41.26	42
11	15.33	1.24	2.52	12.81	17.85	18
11.5	11.08	1.06	2.15	8.93	13.23	14
12	6.67	0.83	1.68	4.99	8.35	9
13	2.42	0.32	0.65	1.77	3.07	4
14	3.50	0.42	0.85	2.65	4.35	5
15	3.08	0.40	0.81	2.27	3.89	4

Tabela 4: Intervalo de confiança para as vendas de 2014 a 2016 de calçados para mulheres.

	Média	ErroPadrão	MargemErro	-95% CI	+95% CI	NumEscolhido
4.5	1.42	0.37	0.75	0.67	2.17	3
5	1.00	0.26	0.53	0.47	1.53	2
5.5	1.75	0.36	0.73	1.02	2.48	3
6	2.25	0.52	1.06	1.19	3.31	4
6.5	7.00	0.80	1.62	5.38	8.62	9
7	11.75	0.71	1.44	10.31	13.19	14
7.5	21.67	2.40	4.87	16.80	26.54	27
8	36.17	2.55	5.18	30.99	41.35	42

	Média	ErroPadrão	MargemErro	-95% CI	+95% CI	NumEscolhido
8.5	28.83	2.53	5.14	23.69	33.97	34
9	26.50	1.96	3.98	22.52	30.48	31
9.5	12.67	0.96	1.95	10.72	14.62	15
10	6.42	0.59	1.20	5.22	7.62	8
10.5	5.50	0.69	1.40	4.10	6.90	7
11	1.25	0.31	0.63	0.62	1.88	2
11.5	1.67	0.43	0.87	0.80	2.54	3
12	2.17	0.61	1.24	0.93	3.41	4

2.2 Tarefa 2

2.2.1 Desenvolvimento

Para a solução dessa tarefa serão filtrados os dados de vendas de sapatos femininos nas lojas ARA1 e ARA2 (Aracati 1 e 2), e separados em tabelas distintas, uma para cada loja referente às vendas em 2016.

Tabela 5: Vendas de sapatos femininos em 2016 da loja 1 de Aracati, por tamanho.

	jan16	fev16	mar16	abr16	mai16	jun16	jul16	ago16	set16	out16	nov16	dez16
4.5	0	0	0	0	1	3	0	0	0	0	1	0
5	0	0	0	0	0	0	2	0	0	0	0	0
5.5	0	0	0	0	0	0	0	0	0	0	1	0
6	0	2	0	0	0	0	0	0	0	0	0	0
6.5	3	3	1	2	1	0	2	0	2	1	3	4
7	0	3	3	4	1	0	1	0	2	0	0	1
7.5	1	2	4	1	2	6	4	3	5	8	2	1
8	6	10	3	9	1	3	6	8	3	12	3	9
8.5	10	10	10	7	14	4	7	7	4	8	7	9
9	1	3	8	6	3	1	4	4	0	2	4	2
9.5	4	1	2	1	2	2	2	4	5	2	3	2
10	0	1	1	1	1	1	3	1	0	0	0	1
10.5	1	0	0	0	2	2	4	1	0	3	1	1
11	1	0	0	1	0	0	0	0	0	0	0	0
11.5	0	0	0	2	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	1	0	1	1	0

Tabela 6: Vendas de sapatos femininos em 2016 da loja 2 de Aracati, por tamanho.

	jan16	fev16	mar16	abr16	mai16	jun16	jul16	ago16	set16	out16	nov16	dez16
4.5	0	0	0	0	0	0	0	0	0	1	0	0
5	0	0	0	0	0	0	0	2	0	0	0	0
5.5	0	0	0	0	0	0	0	1	0	2	0	1
6	0	1	3	1	2	0	0	0	0	0	0	0
6.5	2	0	2	1	1	2	0	1	2	1	3	0
7	0	0	0	4	1	3	1	1	1	3	1	4
7.5	2	1	1	3	2	7	9	8	14	8	6	3
8	13	6	5	13	5	3	11	6	6	9	8	3
8.5	8	5	10	4	5	5	9	7	3	7	9	8
9	5	2	2	9	3	1	1	7	2	1	4	2
9.5	0	1	1	0	1	2	2	1	7	2	4	2
10	0	1	1	0	0	0	2	3	0	2	0	0
10.5	0	2	0	0	0	1	0	0	0	0	2	1
11	1	0	0	1	1	0	0	0	0	0	0	1
11.5	0	0	5	0	0	0	0	1	0	0	0	0
12	0	0	0	0	1	0	2	1	0	0	1	0

Após a seleção e organização dos dados são aplicados testes estatísticos para cada tamanho de sapato vendido, com o objetivo de saber se há diferença no perfil de venda da loja em cada tipo de produto. O teste

selecionado é o t de student pelos seguintes motivos:

- Temos variância populacional desconhecida;
- Temos uma amostra pequena.
- Por se tratar de uma amostra grande podemos utilizar testes paramétricos.

O teste t será utilizado para saber se há diferença significativa entre as vendas em 2016 de cada tamanho de sapato. Serão utilizados os seguintes parâmetros para a realização dos testes:

- Amostra 1: vendas em 2016 dos sapatos de um determinado número da loja ARA1;
- Amostra 2: vendas em 2016 dos sapatos, do mesmo número da amostra 1, da loja ARA2;
- Variância populacional será considerada igual, embora desconhecida;
- Intervalo de confiança de 90%;
- Graus de liberdade: 22, pois são testadas duas amostras de 12 meses cada com variâncias populacionais consideradas iguais.

A hipótese nula do teste t aplicado é de que não há diferença significativa entre as médias das vendas mês a mês dos mesmos tipos produtos de cada loja.

A variância conjunta das amostras é calculada a partir da fórmula abaixo:

$$s_p^2 = \frac{(n_x - 1)\, s_x^2 + (n_y - 1)\, s_y^2}{n_x + n_y - 2}$$

A variância conjunta é utilizado para calcular o intervalo de confiança, para o nível especificado, da diferença entre as médias, de acordo com a expressão seguinte:

$$(\bar{x} - \bar{y}) \pm t_{n_x+n_x-2,\alpha/2} \sqrt{\frac{s_p^2}{n_x} + \frac{s_p^2}{n_y}}$$

A chamada da função do teste t em R:

```
t.test(ara1[,"num"],ara2[,"num"], var.equal = TRUE,conf.level = 0.90)
```

O resultado dos teste é mostrado na tabela 7.

Tabela 7: Comparação das vendas das lojas ARA1 e ARA2 por tamanho no ano de 2016

	estatística t	valor p	gl	mediaARA1	mediaARA2	-90% CI	+90% CI
4.5	1.221	0.2349	22	0.42	0.08	-0.14	0.80
5	0.000	1.0000	22	0.17	0.17	-0.40	0.40
5.5	-1.216	0.2370	22	0.08	0.33	-0.60	0.10
6	-1.254	0.2232	22	0.17	0.58	-0.99	0.15
6.5	1.268	0.2179	22	1.83	1.25	-0.21	1.37
7	-0.558	0.5827	22	1.25	1.58	-1.36	0.69
7.5	-1.575	0.1295	22	3.25	5.33	-4.35	0.19
8	-0.875	0.3912	22	6.08	7.33	-3.70	1.20
8.5	1.378	0.1822	22	8.08	6.67	-0.35	3.18
9	-0.085	0.9333	22	3.17	3.25	-1.77	1.61
9.5	0.881	0.3880	22	2.50	1.92	-0.55	1.72
10	0.215	0.8321	22	0.83	0.75	-0.58	0.75
10.5	1.715	0.1004	22	1.25	0.50	0.00	1.50
11	-0.920	0.3676	22	0.17	0.33	-0.48	0.14
11.5	-0.742	0.4662	22	0.17	0.50	-1.11	0.44
12	-0.715	0.4820	22	0.25	0.42	-0.57	0.23

2.2.2 Análise dos resultados

Considerando um valor p de referência de `0.05`, pode-se observar na tabela `7` que nenhum valor p dos testes realizados indica que há diferenças estatisticamente relevantes entre as vendas dos sapatos de cada tamanho, portanto a hipótese nula de que não há diferença entre as médias das vendas não pode ser rejeitada.

Pode-se tirar a mesma conclusão ao analisar os intervalos de confiança gerados pelos testes pois uma vez que, para todos os tamanhos de sapatos, o valor `0.0` está contido entre o limite inferior e o limite inferior do intervalo de confiança de 90%.

Em posse dessas conclusões, já que não foram encontradas diferenças significativas entre o perfil de vendas das lojas 1 e 2 de Aracati, é possível sugerir que as lojas possuam um estoque único.