

# Integrando o com o L<sup>A</sup>T<sub>E</sub>X

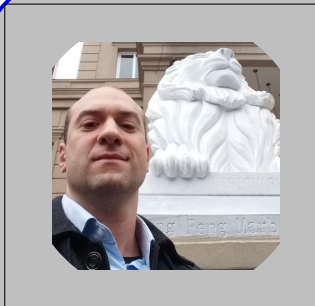
## Livro de Exercícios

Rodrigo Hermont Ozon\*

Maio, 2020

---

\*Economista e Mestre em Desenvolvimento Econômico pela UFPR.




## Sobre o Autor:

*Rodrigo Hermont Ozon, economista e apaixonado por econometria, pelas aplicações de modelos econômicos a problemas reais e cotidianos vivenciados na sociedade e na realidade das empresas.*

Seus contatos podem ser acessados em:

-  [Github](#)
-  [Linkedin](#)

## Resumo

Trata-se de um livro de exercícios rodado no overleaf para os chunks do  no L<sup>A</sup>T<sub>E</sub>X



À minha amada esposa, Idiane *"Porque sou eu que conheço os planos que tenho para vocês", diz o Senhor, "planos de fazê-los prosperar e não de lhes causar dano, planos de dar-lhes esperança e um futuro."*

[Jeremias 29:11](#)

















## Sumário

<b>1</b>	<b>Introdução</b>	<b>7</b>
<b>2</b>	<b>Funções Matemáticas</b>	<b>7</b>
2.1	Terceiro vídeo da primeira semana: . . . . .	7
<b>3</b>	<b>Variáveis escalares</b>	<b>8</b>
3.1	Quarto vídeo da primeira semana . . . . .	8
<b>4</b>	<b>Vetores coluna</b>	<b>10</b>
4.1	Quinto vídeo da primeira semana . . . . .	10
<b>5</b>	<b>Dataframes (bases de dados)</b>	<b>12</b>
5.1	Sexto vídeo da primeira semana . . . . .	12
<b>6</b>	<b>Importando um dataframe</b>	<b>16</b>
6.1	Sétimo vídeo da primeira semana . . . . .	16
<b>7</b>	<b>Acessando o help no  Studio</b>	<b>16</b>
7.1	Oitavo vídeo da primeira semana . . . . .	16
<b>8</b>	<b>Entrevista com Monica Penagos – Intro a satisfação do cliente</b>	<b>17</b>
8.1	Nono vídeo da primeira semana . . . . .	17
<b>9</b>	<b>Leituras do Módulo 1</b>	<b>18</b>
<b>10</b>	<b>Introdução ao Marketing Analytics Comum</b>	<b>20</b>
10.1	Décimo vídeo da primeira semana . . . . .	20
<b>11</b>	<b>Marketing Analytics - Um briefing da história</b>	<b>20</b>
11.1	Décimo primeiro vídeo da primeira semana . . . . .	20
<b>12</b>	<b>Satisfação dos clientes</b>	<b>22</b>
12.1	Décimo segundo vídeo da primeira semana . . . . .	22
<b>13</b>	<b>Métricas e técnicas de escalonamento</b>	<b>24</b>
13.1	Décimo terceiro vídeo da primeira semana . . . . .	24
<b>14</b>	<b>Técnicas de medição e dimensionamento - Escalas primárias de medição</b>	<b>24</b>
14.1	Décimo quarto vídeo da primeira semana . . . . .	24
<b>15</b>	<b>Técnicas de medição e dimensionamento - dimensionamento não comparativo</b>	<b>26</b>
15.1	Décimo quinto vídeo da primeira semana . . . . .	26
<b>16</b>	<b>Design de experimentos: conceitos-chave</b>	<b>27</b>
16.1	Décimo sétimo vídeo da primeira semana . . . . .	27
<b>17</b>	<b>Design de experimentos: controle de erros experimentais</b>	<b>29</b>
17.1	Décimo oitavo vídeo da primeira semana . . . . .	29
<b>18</b>	<b>Entrevista com Monica Penagos - Teste A / B e ANOVA na Prática</b>	<b>30</b>
18.1	Primeiro vídeo da segunda semana . . . . .	30
<b>19</b>	<b>Exercício da semana</b>	<b>31</b>
19.1	Leitura do material didático . . . . .	31
<b>20</b>	<b>Teste A/B: Introdução</b>	<b>38</b>
20.1	Quarto vídeo da segunda semana . . . . .	38
<b>21</b>	<b>Teste A/B: Tipos de teste</b>	<b>39</b>
21.1	Sexto vídeo da segunda semana . . . . .	39

<b>22</b>	<b>Teste A/B: Exemplo no </b>	<b>41</b>
22.1	Sétimo vídeo da segunda semana . . . . .	41
<b>23</b>	<b>Introdução a ANOVA</b>	<b>45</b>
23.1	Oitavo vídeo da segunda semana . . . . .	45
<b>24</b>	<b>One-Way ANOVA – Insect Spray Example</b>	<b>47</b>
24.1	Nono vídeo da segunda semana . . . . .	47
<b>25</b>	<b>One-Way ANOVA – Insect Spray Example in </b>	<b>48</b>
25.1	Décimo vídeo da segunda semana . . . . .	48
<b>26</b>	<b>Two-Way ANOVA – Tooth Growth Example</b>	<b>50</b>
26.1	Décimo primeiro vídeo da segunda semana . . . . .	50
<b>27</b>	<b>Entrevista com Monica Penagos – Modelos de escolha na prática</b>	<b>53</b>
27.1	Primeiro vídeo da terceira semana . . . . .	53
<b>28</b>	<b>Leitura da semana</b>	<b>53</b>
28.1	K means clustering . . . . .	53
<b>29</b>	<b>Modelo de Escolha Binária: Modelo Logit</b>	<b>53</b>
29.1	Segundo vídeo da terceira semana . . . . .	53
<b>30</b>	<b>Modelo de Escolha binária: Exemplo do modelo Logit</b>	<b>60</b>
30.1	Terceiro vídeo da terceira semana . . . . .	60
<b>31</b>	<b>Modelo de Escolha Binária – Modelo Logit: Exemplo 2</b>	<b>64</b>
31.1	Quarto vídeo da terceira semana . . . . .	64
<b>32</b>	<b>Escalonamento Multidimensional: Introdução</b>	<b>66</b>
32.1	Quinto vídeo da terceira semana . . . . .	66
<b>33</b>	<b>Procedendo uma análise de escalonamento multidimensional</b>	<b>68</b>
33.1	Sexto vídeo da terceira semana . . . . .	68
<b>34</b>	<b>Escalonamento multidimensional: Exemplo no </b>	<b>69</b>
34.1	Sétimo vídeo da terceira semana . . . . .	69
<b>35</b>	<b>Entrevista com Monica Penagos: Análise conjunta na prática</b>	<b>74</b>
35.1	Primeiro vídeo da quarta semana . . . . .	74
<b>36</b>	<b>Leituras da semana</b>	<b>75</b>
36.1	Leitura semanal (última semana) . . . . .	75
<b>37</b>	<b>Análise Conjunta: Introdução</b>	<b>75</b>
37.1	Segundo vídeo da quarta semana . . . . .	75
<b>38</b>	<b>Análise Conjunta: Métodos de coleta de dados</b>	<b>76</b>
38.1	Terceiro vídeo da quarta semana . . . . .	76
<b>39</b>	<b>Tipos de Análise Conjunta</b>	<b>78</b>
39.1	Quarto vídeo da quarta semana . . . . .	78
<b>40</b>	<b>Análise Conjunta: Utilidade com valor parcial</b>	<b>78</b>
40.1	Quinto vídeo da quarta semana . . . . .	78
<b>41</b>	<b>Conjoint Analysis Example – Ferry Fares</b>	<b>79</b>
41.1	Quinto vídeo da quarta semana . . . . .	79
<b>42</b>	<b>Análise Conjunta no </b>	<b>80</b>
42.1	Sexto vídeo da quarta semana . . . . .	80

<b>43 Análise Conjunta no  Exemplo parte 1</b>	<b>81</b>
43.1 Sétimo vídeo da quarta semana . . . . .	81
<b>44 Conjoint Analysis  Example – Part 2</b>	<b>88</b>
44.1 Último vídeo da última semana . . . . .	88
<b>45 Resumo do curso: Aplicando Data Analytics no Marketing</b>	<b>90</b>
45.1 Entrevista com Monica Penagos . . . . .	90

## Listings

1	Algumas funções matemáticas do 	7
2	Variáveis escalares no 	9
3	Vetores coluna no 	10
4	Dataframes no 	13
5	Imputando Dataframes no 	16
6	Acessando o help no  Studio	17
7	Teste A/B no 	31
8	ANOVA original no 	34
9	Teste A/B no 	42
10	One-Way ANOVA no 	49
11	Two-Way ANOVA no 	52
12	Regressão Logística no 	55
13	Exemplo de aplicação do modelo Logit no 	62
14	Rodando o modelo de regressão logística no 	64
15	Script de análise conjunta no 	83
16	Análise Conjunta no 	89


# 1 Introdução

Esse livro de exercícios é oriundo do curso *Applying Data Analytics in Marketing*.

Para que esses scripts funcionem redondinho recomendo que você faça a integração do seu  Studio com o Overleaf observando esse tutorial aqui.

## 2 Funções Matemáticas

### 2.1 Terceiro vídeo da primeira semana:

OK. Portanto, nesta lição, começaremos a examinar alguns comandos R elementares. Sei que alguns de vocês que participam desta aula podem não ter nenhuma experiência em programação e alguns podem ter uma vasta experiência em programação. Mas para aqueles que não têm experiência em programação em nenhuma linguagem, não se preocupe. Nesta classe, o nível de dificuldade será o mesmo que escrever uma função no Microsoft Excel. Daremos pequenos passos ao longo do caminho e ficaremos à vontade para pausar o vídeo a qualquer momento, caso você queira examinar suas anotações. Então vamos começar. A primeira coisa que eu gostaria de falar é usar  com funções matemáticas, quase como uma calculadora. Portanto, nesta lição, vou me concentrar principalmente na área de console. Então, eu vou maximizar essa tela. Aqui vamos nós.

Acho que mostrei a você, no vídeo anterior, alguns comandos básicos, como 2 mais 4. Então é isso, e estamos rolando. Subtração é com o hífen. Então, podemos usar 3 menos 5, e isso é menos 2. Obviamente, podemos fazer a multiplicação, 3 vezes 3 é 9. Até agora tudo bem. A divisão é com uma barra, então 5 acima de 2 é 2,5. Se você deseja fazer um expoente, por exemplo, duas a terceira potência, duas vezes duas vezes duas que seria 8. Outra maneira de escrever a mesma coisa é usar duas e, em vez de um asterisco, use asterisco duplo, dois para o terceiro poder, e esse também é oito. Portanto, essas são suas funções aritméticas básicas, mas você pode ir além disso e usar funções mais complicadas que aprendeu no passado. Portanto, podemos pegar o valor absoluto de um número que essencialmente retira o sinal. Portanto, o comando lá é abs menos 7, digamos, digitei 5, pressione Enter e o valor absoluto de menos 5 é 5. Então, uma coisa a ser observada sobre as funções, e isso vale para todo R, é que existe uma função name, nesse caso, é abs, existe um parêntese aberto e, em seguida, você coloca seus argumentos, as variáveis de interesse e os parênteses próximos. Essa é a forma geral de uma função. Outras funções podem ter outros argumentos que você pode colocar lá, mas esses geralmente são um número de cada vez. OK. Portanto, temos o valor absoluto de cinco. Podemos pegar o log de 5. Podemos pegar o log de 5, a base é igual a 3, se você quiser fazer dessa maneira. Expoente, exp. Raiz quadrada. A raiz quadrada de 4 é 2. O fatorial de 9 é 362.000 mais. Então, suas funções trigonométricas tradicionais, seno de zero, seno de pi. Aí está. Então, essas são suas funções básicas. Convido você a olhar para a folha de dicas R e praticar algumas dessas funções por conta própria. Se você não sabe o que essas funções significam, porque nunca foi exposto a algo como, por exemplo, fatorial, não se preocupe. Seu conhecimento direcionará o que você precisa saber em termos de função. Então, talvez um dia você encontre a necessidade de usar a função fatorial. Talvez você esteja fazendo algumas combinações e, quando estiver estudando isso, isso se tornará útil. Isso envolve funções aritméticas em R.

```

1 > 2+4
2 > 3-5
3 > 3*3
4 > 5/2
5 > 2^3
6 > 2**3
7 > abs(-5)
8 > log(5)
9 > log(5, base=3_)
10 > exp(3)
11 > sqrt(4)
12 > factorial(9)
13 > sin(0)
14 > sin(pi)

```

Listing 1: Algumas funções matemáticas do 

```

2+4

## [1] 6

```

```
3-5
## [1] -2

3*3
## [1] 9

5/2
## [1] 2.5

2^3
## [1] 8

2**3
## [1] 8

abs(-5)
## [1] 5

log(5)
## [1] 1.609438

log(5, base=3)
## [1] 1.464974

exp(3)
## [1] 20.08554

sqrt(4)
## [1] 2


factorial(9)
## [1] 362880

sin(0)
## [1] 0

sin(pi)
## [1] 1.224647e-16
```

## 3 Variáveis escalares

### 3.1 Quarto vídeo da primeira semana

No vídeo anterior, descrevemos como usar funções matemáticas em . Mas observe, neste exemplo aqui, eu tenho 2 mais 2. Se eu executar esse código, recebo a resposta 4, mas a resposta não é armazenada em nenhum lugar. Então, nesta palestra, vamos falar sobre variáveis em R, basicamente, como armazenar informações. A maneira de fazer isso é realmente simples.

Primeiro, você cria um nome de variável, vou chamá-lo de x. Não é um nome muito criativo, mas serve. Então você usa essa coisa que se parece com uma flecha. Portanto, é um sinal menor que e um hífen. Então



eu vou colocar um valor lá quatro.

Ainda não foi executado, então deixe-me executar esse código abaixo. Como mencionei em um vídeo anterior, quando clico no botão Executar, está cortando e colando esse comando no console e pressionando Enter. Portanto, no console, você vê que `x` recebe um valor de quatro. O que isso significa? Criar uma variável em R é criar, é como criar espaço no computador.

Imagine que você tenha uma Shoebox, e esse é o seu espaço no computador, e depois rotule essa Shoebox com as coisas que você colocará nele. Nesse caso, eu o chamei de `x`. Pode ser coisas como altura. Portanto, se você tem uma lista de equipes, esta é a lista de todas as alturas de todos os jogadores, podem ser os pesos, podem ser as datas de nascimento, pode ser o que você quiser. Então, você deseja rotular esta Shoebox e colocar dados leves dentro dessa Shoebox.

Aqui, eu apenas coloquei o valor de quatro. Observe também que, no canto superior direito do RStudio, você pode ver na guia ambiente global; portanto, clique nele se não estiver ativo, há sua variável `xe` aqui está o conteúdo dos quatro Shoebox.

Agora, vamos criar uma segunda variável `y` com valor três. Lá, agora você pode ver que eu executei esse código aqui, e você pode vê-lo no ambiente global aqui em cima, criei um pouco de espaço para ele e `y` é igual a um valor de três. Outra maneira de ver o valor do que está dentro de uma variável. Estou no console aqui agora e posso apenas digitar o nome da variável `x`, e ele responderá com quatro nesse caso ou `y`, e responderá com três. Para que eu possa vê-lo dentro da caixa, dessa forma, posso vê-lo no ambiente. Agora, eu posso executar alguns comandos como observamos anteriormente. Então, vamos fazer `x` mais `y`. Lá vamos nós, e executamos isso, e `x` mais `y` nesse caso são sete, o que faz sentido porque `x` contém um valor de quatro e `Y` contém o valor de três. Mas observe que a soma de `xey` não é armazenada em nenhum lugar. Para que eu possa criar uma terceira variável. Vamos chamá-lo de `z`. Use o operador de atribuição, que é o menos que o hífen do sinal, e `x` mais `y`. Vamos rodar essa linha, e você pode ver agora no canto superior direito, na seção Ambiente do RStudio, criamos a variável `z` e, ao mesmo tempo, colocamos o valor de `x` mais `y`. Então é assim que você cria uma variável escalar.

Uma coisa a notar é que as variáveis podem ser de qualquer tipo. Então, se você deseja inserir texto, deve colocá-lo entre aspas. Então, vamos colocar, chamarei essa variável `t1` para o exemplo de texto um, e posso escrever "oi mãe" e aí está. Você pode ver no ambiente que há "oi mãe". Se eu descer aqui e digitar `t1`, ele mostrará "oi mãe". Mas essa é uma sequência de caracteres. Então, o que você esperaria se eu fizesse algo como `x` mais `t1`? Tentamos isso, não tenha medo de pressionar "Enter", mas você pode ver que é um erro porque está tentando adicionar uma string de texto.

String é um monte de caracteres com um numérico. Uma nota final sobre a variável Existem várias maneiras de escrever nomes de variáveis. Quase não há limitações sobre o que você pode chamar de variável. Mas, em termos de estilo e legibilidade, você deseja escolher um estilo consistente em todos os seus programas. gostaria de usar é chamado Lower camel case. Então, o que isso significa? Digamos que você tenha uma coluna de dados ou variável com o tamanho de um sapato. A caixa Camel se pareceria com isso. `SHOESIZE`. Observe, eu uso `S` aqui. você pode encadear uma frase e colocar em maiúscula cada palavra. Algumas pessoas também gostam de colocar em maiúscula a primeira letra. e `S` maiúsculo para sapato.

Tamanho do sapato, algo assim. Não importa, basta escolher um estilo e ficar com ele para facilitar a leitura. Observe também que, se você entrar no ambiente corporativo, eles podem ter seus próprios padrões de nomeação de variáveis de dados. Portanto, é algo em que você deve pensar e tentar se manter consistente com o ambiente corporativo. Isso envolve variáveis escalares.

```
1 > 2+2
2 > x <- 4
3 > y <- 3
4 > z <- x+y
5 > t1 <- "Hi mom"
6 > x+t1
```

Listing 2: Variáveis escalares no 

```
2+2

## [1] 4

x <- 4
y <- 3
z <- x+y
z

## [1] 7
```

```
t1 <- "Hi mom"
x+t1

## Error in x + t1: non-numeric argument to binary operator
```

## 4 Vetores coluna

### 4.1 Quinto vídeo da primeira semana

Nesta palestra, vou falar sobre vetores de coluna em `R`. Os vetores de coluna são outro tipo de dados básico em `R` que nos ajudará a armazenar dados e poder manipular dados. O conceito de um vetor de coluna não é muito difícil, tenho certeza que você já viu isso antes.

Aqui no Microsoft Excel, tenho uma coluna de pesos. Chamo a coluna `wt`s de pesos e tenho alguns pesos de objetos que variam de 75 na parte inferior a um máximo de 280 libras ou quilogramas ou o que quer que seja. Então, aqui estão alguns pesos, e podemos querer criar algo semelhante em `R`.

Então, é nesta Linha 2 aqui. O nome do vetor da coluna é `wt`s. Você ainda tem esse hífen de sinal menor que o operador de atribuição `<-`, para criar um vetor de coluna, usa `C`, abre parênteses e há todo o conjunto de números, dados 214, 226 etc., até 75, feche pai. É assim que você cria um vetor de coluna. Então vamos fazer isso. Eu executo o código. Isso aparece no meu ambiente global. Aí está. Eu também posso digitar `wt`s, e aí está. Está listado, 214, 226, 280 etc. Uma coisa a observar, finalmente posso falar sobre esse item de colchete aqui. Se eu criei apenas uma variável escalar, `x` é igual a 4 e mostro `x`, tenho esse colchete um. De fato, uma variável escalar é um vetor de coluna com apenas um elemento, e isso indica qual elemento está no vetor de coluna. Então, aqui estou eu no Excel, e 214 é o primeiro elemento no vetor da coluna.

226 é o segundo elemento no vetor da coluna e assim por diante. Setenta e cinco é o último elemento ou o oitavo elemento no vetor de coluna. Se formos para o RStudio, podemos ver 214, 226 etc. Se eu quiser apenas obter os 226 libras, `wt`s colchete dois, posso acessar isso. Então aí está. Eu posso até colocá-lo em outra variável `y`. `y` é `wt`s colchete 2 mais 5, e você pode ver que agora é um valor de 231. Se eu quisesse acessar mais do que apenas a segunda célula, se eu desejar talvez da segunda à quarta célula, seria `wt`s square colchete dois dois pontos quatro, e lá estão eles. Este colchete na resposta dirá que este é o primeiro elemento. Se atropelar, você verá o próximo número. Então, aqui, eu acessei as variáveis de segunda a quarta das ponderações e recebo 226, 280, 185, que podemos verificar aqui de duas a quatro, 226, 280 e 185, para que funcione.

Este colchete é o endereço do seu vetor de coluna. Então, aqui está o primeiro elemento, este é o seu segundo elemento, este é o seu terceiro elemento, etc. Se ele se envolver, você verá o endereço do vetor maior. Então, vamos criar um vetor grande, e eu vou chamá-lo de vetor grande, e vou usar um gerador de números aleatórios. Vamos criar 100 números aleatórios. Se eu quiser ver um vetor grande agora, você pode ver que aqui o colchete, esse é o primeiro elemento, 1 menos 1,67 etc., esse é o primeiro elemento. Então temos aqui um colchete 10, esse item aqui é o décimo elemento e assim por diante até obtermos o centésimo elemento.

Então é isso que esse endereçamento aqui é. A próxima coisa que quero falar são alguns comandos básicos que podemos usar em um vetor. Eu os tenho aqui nas linhas 6 a 10. A primeira é que podemos obter um resumo dos pesos. Aqui estão suas estatísticas descritivas, o valor mínimo, o valor máximo, a média, a mediana e os intervalos do quartil. Se eu quisesse apenas a média, eu poderia fazer isso com o comando da média. Aí está, e a variação e o desvio padrão também.

```
1 > wt <- c(214, 226, 280, 185, 130, 146, 165, 750)
2 > bigvector <- rnorm(100)
3 > bigvector
4 > x2 <- c(1,2,3,4,5,6,7,8)
5 > summary(wt)
6 > mean(wt)
7 > var(wt)
8 > sd(wt)
9 > wt[2]+5
10 > wt[2] <- wt[2]+5
11 > wt
12 > wt*3
13 > wtstimes3 <- wt*3
14 > wtstimes3
```

Listing 3: Vetores coluna no 

```

wts <- c(214, 226, 280, 185, 130, 146, 165, 750)
bigvector <- rnorm(100)
bigvector

##      [1] -1.40988943 -0.05161684  1.10053367  0.94710049 -0.09241373 -0.12446197
##      [7] -0.91298779  1.06259478 -0.78132547 -0.09445940 -0.28390975  1.25780260
##     [13]  0.25159855  0.76872735  0.46710579  1.14802579 -0.57814666 -0.88179739
##     [19] -0.21864046 -1.08256458 -0.44852525 -1.63851211 -0.41178982  0.08033096
##     [25] -0.19932419 -0.27100666 -0.81904809 -0.75186704 -0.83654656  1.35577258
##     [31] -2.27524441 -0.52600025  1.04097381 -1.80907508  0.87888081 -0.52429226
##     [37]  1.05002341  0.60707779 -0.24318349 -0.66164939 -0.45452204  1.50197306
##     [43]  1.95168357  0.47428626 -1.17610159 -0.32755461 -0.36056346 -0.06970120
##     [49]  0.42165063  1.29190287  1.34220934 -0.10130120  1.59828321 -0.57558484
##     [55] -0.67107142  0.52228045 -0.38915103  0.39850276  0.41310161  1.85754487
##     [61] -1.03096879  0.75996995  0.86518429 -1.17607992  0.27714983 -0.50548589
##     [67]  0.11331399 -1.37384649  3.13079966  0.36341759  0.20151762  0.20772604
##     [73] -0.63240062 -0.02916549 -1.26319628 -0.72506144  1.30239874 -0.56453164
##     [79]  1.70879379  0.34089589 -0.50953588 -0.09608883 -0.29161406  1.00173897
##     [85]  0.71179821  0.98371677  3.31083342  1.05678358  0.04375154  0.13597274
##     [91]  0.58274449 -0.51448439  0.25739834 -0.43363149 -1.08923612  1.22886186
##     [97] -0.04456067  0.76999725 -0.77670103 -0.59328684

x2 <- c(1,2,3,4,5,6,7,8)
summary(wts)

##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     130.0   160.2   199.5   262.0   239.5   750.0

mean(wts)

## [1] 262

var(wts)

## [1] 41183.71

sd(wts)

## [1] 202.9377

wts[2]+5

## [1] 231

wts[2] <- wts[2]+5
wts

## [1] 214 231 280 185 130 146 165 750

wts*3

## [1]  642  693  840  555  390  438  495 2250


wtstimes3 <- wts*3
wtstimes3

## [1]  642  693  840  555  390  438  495 2250

```

## 5 Dataframes (bases de dados)

### 5.1 Sexto vídeo da primeira semana

Neste vídeo, gostaria de falar sobre DataFrames, outro tipo de estrutura de dados em . Não é nada além de uma tabela que você pode ver em uma planilha do Excel ou algo assim. Mas vamos construir um. A primeira linha aqui, neste código, existe vários nomes de família. É um vetor de coluna e os nomes são criativamente pai, mãe, irmão, irmã e cachorro. Então, vamos criar esse vetor. Se dermos uma olhada, podemos ver que há o vetor da coluna. Da mesma forma, temos o vetor da coluna de crédito de suas idades. Podemos dar uma olhada. Lá estão eles, seus gêneros e seus pesos familiares. Então agora, eu tenho essas quatro colunas. Um é o nome, a idade, a coluna, o sexo e o peso. Portanto, esses são alguns dados que você normalmente pode encontrar. Vou criar um DataFrame. Para criar um DataFrame, você precisa desta impressão aberta do Dataframe e, em seguida, dos nomes de cada coluna que você colocará lá. Agora, vou nomear o DataFrame, a família. Então, vamos executar essa linha de código. Aí está. Agora, podemos realmente dar uma olhada e você pode ver que parece quase uma planilha do Excel. Os nomes das colunas estão no topo. Lá estão eles. Aqui estão os nomes, as idades correspondentes, os gêneros e os pesos.

Se você se deparar com o código de outra pessoa ou algo com o qual alguma estrutura de dados não esteja familiarizada, você pode usar este comando `str`, que significa estrutura. Eu posso até colocar isso nos comentários aqui, estrutura.

Isso informará que tipo de dados você possui. Então, aqui fizemos a estrutura da família, e você pode ver que existem cinco observações de quatro variáveis e as listamos.

Deixe-me ir aqui embaixo, a estrutura de talvez os pesos, apenas os pesos de vetor de colunas.

Você pode ver aqui que é um vetor de coluna numérica que varia de um a cinco elementos e existem os valores reais. Portanto, o comando `structure` informa um pouco sobre os dados com os quais você está trabalhando. Se você quiser acessar apenas uma das colunas em um DataFrame, poderá usar esta notação de nome de cifra. Portanto, no comando de estrutura, você vê o nome é a família e, em seguida, os nomes de família em dólares, a idade da família etc. Lá estão eles e há o vetor da coluna. Agora, podemos fazer coisas como calcular a média. Eu já tenho o comando aqui em cima. Qual é a média ou qual é a idade média na família? Aí está. Uma coisa que você precisa fazer quando tem uma tabela é poder acessar uma linha, coluna ou célula individual. Existem diferentes maneiras de abordar cada um desses componentes. Aqui está um conjunto de dados pré-carregado em R. É chamado `mtcars`. `Mt` significa motor trends Magazine, `mtcars`. Aqui está um DataFrame. Tem o modelo do carro, milhas por galão, cilindrada, etc. Podemos querer olhar para a estrutura, isso é um quadro de dados? Aí está. É um DataFrame, possui 32 observações de 11 variáveis neste caso. Se quiséssemos obter a coluna de milhas por galão, como mostrei no exemplo de família, isso seria `$ mpg`. Lá estão eles.


Se eu quisesse obter apenas a primeira coluna, usaria esse colchete de notação, algo como vírgula. Então, vamos olhar para a linha 10. Esta é a linha 1, coluna 2. Portanto, se eu executar essa linha de código, obtenho o valor de seis, que é a linha 1, coluna 2, é esse valor lá. Se eu quiser apenas a primeira coluna, não coloco nada nesse primeiro elemento antes da vírgula e apenas digo a primeira coluna. Aí está. Acontece que é o mesmo que a coluna por milhas por galão. Essa é a primeira coluna. Se eu quisesse obter os pesos das colunas, aí estão eles, `mtcars $ weight`. Lá estão eles. Se eu quisesse a sexta coluna, posso fazer isso. Se eu quisesse a primeira linha. Então agora, estou trocando colchete, linha um, me dê todas as colunas na primeira linha. Lá estão eles. Nesse caso, esse DataFrame possui alguns rótulos, para que possamos fazê-lo pelo rótulo ou pela linha. Aí está. Isso basicamente o ajudará a contornar os conjuntos de dados. De um modo geral para esta aula, apenas me dê a coluna.

Portanto, esse primeiro conjunto de comandos é o que você realmente precisa saber, seja no nome da coluna. Então, nome do DataFrame e nome da coluna. Se você está acostumado ao Excel, esse seria o nome do arquivo do Excel e, em seguida, o nome da sua coluna ou, em seguida, basta escolher as colunas individuais. Isso envolve os DataFrames.

```

1 #Dataframes
2 > famNames <- c("dad","mom","bro","sis","dog")
3 > famNames
4 > famAges <- c(42,41,12,8,5)
5 > famAges
6 > famGender <- c("M","F","M","F","F")
7 > famGender
8 > famWeight <- c(188,135,83,61,44)
9 > famWeight
10 #Create dataframe
11 > TheFamily <- data.frame(famNames, famAges, famGender, famWeight)
12 > TheFamily
13 #Structure
14 > str(TheFamily)
15 > str(famGender)
16 > summary(TheFamily)
17 > mean(TheFamily$famAges)
18 > View(TheFamily)
19 > mtcars
20 > str(mtcars)
21 > mtcars$mpg
22 > mtcars[1,2]
23 > mtcars[,1]
24 > mtcars$wt

```

Listing 4: Dataframes no 

```

#Dataframes
famNames <- c("dad","mom","bro","sis","dog")
famNames

## [1] "dad" "mom" "bro" "sis" "dog"

famAges <- c(42,41,12,8,5)
famAges

## [1] 42 41 12 8 5

famGender <- c("M","F","M","F","F")
famGender

## [1] "M" "F" "M" "F" "F"

famWeight <- c(188,135,83,61,44)
famWeight

## [1] 188 135 83 61 44

#Create dataframe
TheFamily <- data.frame(famNames, famAges, famGender, famWeight)
TheFamily

##   famNames famAges famGender famWeight
## 1     dad     42         M        188
## 2     mom     41         F        135
## 3     bro     12         M         83
## 4     sis      8         F         61
## 5     dog      5         F         44

#Structure
str(TheFamily)

## 'data.frame': 5 obs. of 4 variables:
## $ famNames : Factor w/ 5 levels "bro","dad","dog",...: 2 4 1 5 3
## $ famAges : num 42 41 12 8 5
## $ famGender: Factor w/ 2 levels "F","M": 2 1 2 1 1
## $ famWeight: num 188 135 83 61 44

```

```

str(famGender)

## chr [1:5] "M" "F" "M" "F" "F"

summary(TheFamily)

## famNames    famAges    famGender    famWeight
## bro:1      Min.      : 5.0    F:3          Min.      : 44.0
## dad:1      1st Qu.: 8.0    M:2          1st Qu.: 61.0
## dog:1      Median :12.0          Median : 83.0
## mom:1      Mean      :21.6          Mean    :102.2
## sis:1      3rd Qu.:41.0          3rd Qu.:135.0
##           Max.      :42.0          Max.    :188.0

mean(TheFamily$famAges)

## [1] 21.6

View(TheFamily)

## Warning in View(TheFamily): unable to open display
## Error in .External2(C_dataviewer, x, title): unable to start data viewer

mtcars

##           mpg cyl  disp  hp drat   wt  qsec vs am gear carb
## Mazda RX4      21.0   6 160.0 110 3.90 2.620 16.46  0  1   4    4
## Mazda RX4 Wag  21.0   6 160.0 110 3.90 2.875 17.02  0  1   4    4
## Datsun 710     22.8   4 108.0  93 3.85 2.320 18.61  1  1   4    1
## Hornet 4 Drive  21.4   6 258.0 110 3.08 3.215 19.44  1  0   3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0   3    2
## Valiant        18.1   6 225.0 105 2.76 3.460 20.22  1  0   3    1
## Duster 360     14.3   8 360.0 245 3.21 3.570 15.84  0  0   3    4
## Merc 240D      24.4   4 146.7  62 3.69 3.190 20.00  1  0   4    2
## Merc 230       22.8   4 140.8  95 3.92 3.150 22.90  1  0   4    2
## Merc 280       19.2   6 167.6 123 3.92 3.440 18.30  1  0   4    4
## Merc 280C      17.8   6 167.6 123 3.92 3.440 18.90  1  0   4    4
## Merc 450SE     16.4   8 275.8 180 3.07 4.070 17.40  0  0   3    3
## Merc 450SL     17.3   8 275.8 180 3.07 3.730 17.60  0  0   3    3
## Merc 450SLC    15.2   8 275.8 180 3.07 3.780 18.00  0  0   3    3
## Cadillac Fleetwood 10.4   8 472.0 205 2.93 5.250 17.98  0  0   3    4
## Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0   3    4
## Chrysler Imperial 14.7   8 440.0 230 3.23 5.345 17.42  0  0   3    4
## Fiat 128       32.4   4  78.7  66 4.08 2.200 19.47  1  1   4    1
## Honda Civic    30.4   4  75.7  52 4.93 1.615 18.52  1  1   4    2
## Toyota Corolla 33.9   4  71.1  65 4.22 1.835 19.90  1  1   4    1
## Toyota Corona  21.5   4 120.1  97 3.70 2.465 20.01  1  0   3    1
## Dodge Challenger 15.5   8 318.0 150 2.76 3.520 16.87  0  0   3    2
## AMC Javelin    15.2   8 304.0 150 3.15 3.435 17.30  0  0   3    2
## Camaro Z28     13.3   8 350.0 245 3.73 3.840 15.41  0  0   3    4
## Pontiac Firebird 19.2   8 400.0 175 3.08 3.845 17.05  0  0   3    2
## Fiat X1-9      27.3   4  79.0  66 4.08 1.935 18.90  1  1   4    1
## Porsche 914-2  26.0   4 120.3  91 4.43 2.140 16.70  0  1   5    2
## Lotus Europa   30.4   4  95.1 113 3.77 1.513 16.90  1  1   5    2
## Ford Pantera L  15.8   8 351.0 264 4.22 3.170 14.50  0  1   5    4
## Ferrari Dino    19.7   6 145.0 175 3.62 2.770 15.50  0  1   5    6
## Maserati Bora   15.0   8 301.0 335 3.54 3.570 14.60  0  1   5    8
## Volvo 142E     21.4   4 121.0 109 4.11 2.780 18.60  1  1   4    2

str(mtcars)

```

```
## 'data.frame': 32 obs. of 11 variables:
## $ mpg : num 21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
## $ cyl : num 6 6 4 6 8 6 8 4 4 6 ...
## $ disp: num 160 160 108 258 360 ...
## $ hp : num 110 110 93 110 175 105 245 62 95 123 ...
## $ drat: num 3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
## $ wt : num 2.62 2.88 2.32 3.21 3.44 ...
## $ qsec: num 16.5 17 18.6 19.4 17 ...
## $ vs : num 0 0 1 1 0 1 0 1 1 1 ...
## $ am : num 1 1 1 0 0 0 0 0 0 0 ...
## $ gear: num 4 4 4 3 3 3 3 4 4 4 ...
## $ carb: num 4 4 1 1 2 1 4 2 2 4 ...

mtcars$mpg

## [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4
## [16] 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7
## [31] 15.0 21.4

mtcars[1,2]

## [1] 6

mtcars[,1]





## [1] 21.0 21.0 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4
## [16] 10.4 14.7 32.4 30.4 33.9 21.5 15.5 15.2 13.3 19.2 27.3 26.0 30.4 15.8 19.7
## [31] 15.0 21.4

mtcars$wt

## [1] 2.620 2.875 2.320 3.215 3.440 3.460 3.570 3.190 3.150 3.440 3.440 4.070
## [13] 3.730 3.780 5.250 5.424 5.345 2.200 1.615 1.835 2.465 3.520 3.435 3.840
## [25] 3.845 1.935 2.140 1.513 3.170 2.770 3.570 2.780
```

## 6 Importando um dataframe

### 6.1 Sétimo vídeo da primeira semana

Falamos sobre DataFrames como um dos cavalos de trabalho, um dos principais tipos de dados que usaremos no , e é essencialmente uma tabela de dados. No vídeo sobre quadros de dados, mostrei como criar um do zero no ambiente . Mas, na realidade, no ambiente corporativo, geralmente você recebe um conjunto de dados em algum tipo de tabela, geralmente uma tabela do Excel ou uma tabela de valores separados por vírgula e, em seguida, deseja importá-los para o , o que significa que deseja trazer esses dados no ambiente . É assim que se faz. Uma é usar este comando de importação aqui em cima. Como você pode ver, existem algumas opções.

Você pode importar de arquivos de texto, de arquivos do Excel e também de estruturas de dados provenientes de outros pacotes estatísticos, como SPSS, SAS ou Stata. O segundo método é ir para essa guia no canto inferior direito em arquivos, clicar nas reticências para navegar até o diretório em que você deseja estar. Então, aqui estão alguns diretórios, já naveguei para o diretório, e há um arquivo do Excel que mostrarei como exemplo. Posso clicar nele, importar o conjunto de dados e agora posso ver uma prévia do conjunto de dados e posso importá-lo. Uma das opções é a primeira linha como nomes. Agora, neste conjunto de dados, não coloquei nomes na primeira coluna, por isso vou desmarcar isso.

```
1 > library(readxl)
2 > ex1 <- read_excel("C:/Users/Rodri.../ex1.xlsx", +colnames = FALSE
3 > View(ex1)
4 > str(ex1)
5 > ex1[,2]
```


Listing 5: Importando Dataframes no 

```
library(readxl)
ex1 <- read_excel("C:/Users/Rodri.../ex1.xlsx", +colnames = FALSE
View(ex1)
str(ex1)
ex1[,2]

## Error: <text>:2:60: unexpected '='
## 1: library(readxl)
## 2: ex1 <- read_excel("C:/Users/Rodri.../ex1.xlsx", +colnames =
##
```

## 7 Acessando o help no Studio


### 7.1 Oitavo vídeo da primeira semana

OK. Portanto, você pode ter problemas com um comando  e precisa de ajuda. Existem várias maneiras de fazer isso. Por exemplo, você pode estar tendo problemas com uma função. Vamos usar a função soma que adiciona vários números, mas você realmente não sabe como usá-la. Então, vamos tentar obter ajuda. Uma é usar um ponto de interrogação se você souber o nome exato da função que está usando, e observe aqui no canto inferior direito, há alguma documentação sobre como usar essa função. Se você usar ponto de interrogação, ponto de interrogação SUM, ele fará uma pesquisa desses termos, S-U-M, e fará uma pesquisa mais ampla.

O outro lugar que você pode procurar, aqui é o [r-project.org/help.html](http://r-project.org/help.html), e isso fornece uma lista de comandos que você pode usar para procurar ajuda na pesquisa na documentação do R ou pedir ajuda. Aqui está um link para o estouro da pilha, outro ótimo recurso. O Google é outro ótimo recurso. Além disso, no ambiente do RStudio, se você clicar em 'Cheatsheets' da ajuda dela, há várias dicas diferentes criadas pelo pessoal do RStudio. Deixe-me clicar neste primeiro, 'RStudio IDE Cheat Sheets' e este PDF,



```
1 > sum()
2 > ?sum()
3 > ??summary.aov
```

Listing 6: Acessando o help no  Studio

```
sum()

## [1] 0

?sum()
??summary.aov
```

## 8 Entrevista com Monica Penagos – Intro a satisfação do cliente

### 8.1 Nono vídeo da primeira semana

Temos Monica conosco novamente hoje. Nas palestras desta semana, falaremos sobre satisfação do cliente, medição da satisfação do cliente e talvez um pouco sobre regressão linear.

Monica, você tem alguma experiência no mundo real usando essas técnicas na Procter and Gamble?

Definitivamente. Portanto, a regressão linear é muito usada na P&G como parte do modelo de mix de marketing. Mais do que a satisfação do cliente, geralmente em preços e promoções para determinar isso. Por isso, lembro-me muito cedo de minha carreira, quando eu estava na Columbia, fui o representante do negócio de tecidos e cuidados domésticos. Queríamos que ele ganhasse liderança de categoria. Naquela época, nossa concorrência realizava inúmeras promoções no mercado e sabíamos que tínhamos um produto melhor, melhor remoção de manchas e melhor criatividade. Mas, como parte do mix de marketing, as promoções e os preços foram mais limpos. Por isso, passamos dois anos com meu colega de marketing usando regressão para entender o que nossos preços precisavam ser e nossa promoção. Acabamos eliminando as promoções, favorecendo um preço mais baixo, e conquistamos a liderança da categoria no final desses dois anos. Então isso é realmente ótimo.

Você usa regressão linear para ajudar a identificar se o preço ou a promoção foi mais eficaz para obter a liderança da categoria. Vamos abordar esses tópicos no conjunto de aulas desta semana.

## 9 Leituras do Módulo 1

- [Introduction to marketing analytics and customer satisfaction](#)
- [Baixar os arquivos de dados de exemplo](#)

```
newdata = read.csv(file="videogamesales.txt") #lendo os dados

which(colnames(newdata)%in%c("Global_Sales","Critic_Score")) #escolha dos dados

## [1] 10 11

summary(newdata[,c(10,11)]) #descriptive statistics

##   Global_Sales      Critic_Score
##   Min.      : 0.0000   Min.      : 0.00
##   1st Qu.: 0.0900   1st Qu.:60.00
##   Median : 0.2400   Median :71.00
##   Mean    : 0.6883   Mean    :68.75
##   3rd Qu.: 0.6500   3rd Qu.:79.00
##   Max.    :82.5300   Max.    :98.00
##   NA's    :27       NA's    :27

model = lm(Global_Sales ~ Critic_Score, data = newdata)

summary(model)

##
## Call:
## lm(formula = Global_Sales ~ Critic_Score, data = newdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.415 -0.615 -0.292  0.136 81.621
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.400484   0.094901  -14.76  <2e-16 ***
## Critic_Score  0.030381   0.001351   22.49  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.763 on 8135 degrees of freedom
## (27 observations deleted due to missingness)
## Multiple R-squared:  0.05855, Adjusted R-squared:  0.05843
## F-statistic: 505.9 on 1 and 8135 DF, p-value: < 2.2e-16
```

Ver no slide [página 47](#)

### Exercício:

Using the same dataset, determine how critic score affect the NA\_Sales

```
newdata = read.csv(file="videogamesales.txt") #lendo os dados

which(colnames(newdata)%in%c("NA_Sales","Critic_Score")) #escolha dos dados

## [1] 6 11

summary(newdata[,c(6,11)]) #descriptive statistics

##      NA_Sales      Critic_Score
## 0      : 629   Min.      : 0.00
## 0.03    : 310   1st Qu.:60.00
## 0.05    : 309   Median :71.00
## 0.04    : 307   Mean    :68.75
## 0.07    : 303   3rd Qu.:79.00
## 0.02    : 295   Max.     :98.00
## (Other):6011   NA's     :27

model = lm(NA_Sales ~ Critic_Score, data = newdata)

## Warning in model.response(mf, "numeric"): using type = "numeric" with a factor response
## will be ignored
## Warning in Ops.factor(y, z$residuals): '-' not meaningful for factors

summary(model)

## Warning in Ops.factor(r, 2): '^' not meaningful for factors

##
## Call:
## lm(formula = NA_Sales ~ Critic_Score, data = newdata)
##
## Residuals:

## Error in quantile.default(resid): factors are not allowed
```

## 10 Introdução ao Marketing Analytics Comum

### 10.1 Décimo vídeo da primeira semana

Olá. Neste vídeo, vou começar a introduzir alguns conceitos, conceitos elementares na análise de marketing, e que ajudarão a impulsionar o restante deste módulo. Espero que você goste da palestra de hoje. A primeira coisa sobre a qual quero falar um pouco é o que é a satisfação do cliente e o que é a análise da satisfação do cliente.

Esse é o processo para medir e identificar seus clientes e verificar se eles estão felizes ou não, satisfeitos ou insatisfeitos com algum tipo de produto ou serviço que você está oferecendo.

É obviamente um grande motivador dos interesses de uma empresa, pois queremos saber quem são nossos clientes e garantir que eles gostem do nosso produto ou serviço e voltem a buscá-lo novamente. Portanto, se seus clientes ficarem satisfeitos com o seu produto, é mais provável que comprem de você novamente do que do seu concorrente e, da mesma forma, se seu serviço for satisfatório ou acima de satisfatório, eles provavelmente voltarão. Portanto, essa análise nos fornece as ferramentas para acessar os sentimentos dos clientes sobre nossos produtos e serviços.

Podemos ajustar nossos produtos e serviços seguindo as respostas que recebemos dos clientes. Se eles não gostarem de um produto ou de uma nova versão de um produto, podemos revertê-lo ou ajustar o produto e nosso serviço apenas para manter nossos clientes satisfeitos. É por isso que é realmente importante fazer. Então, como fazemos isso? Como conduzimos a análise de satisfação do cliente? A maneira mais comum de acessar a satisfação do cliente é uma combinação de pesquisas quantitativas e qualitativas. Então pedimos a eles.

## 11 Marketing Analytics - Um briefing da história

### 11.1 Décimo primeiro vídeo da primeira semana

Este gráfico mostra uma breve história dos dados de marketing e análise de dados de marketing. Há duas faixas aqui. Se primeiro focarmos na parte dos dados, há uma linha do tempo. No início da virada da pesquisa do século passado dados eram bastante populares.

Então começamos a procurar no rastreamento ocular, vimos dados no nível da transação. Com o advento de sistemas de ponto de vendas, começamos a olhar dados do ponto de venda. Agora com a internet, nós temos dados de fluxo de cliques, o que significa como as pessoas são clicando em uma página da web.

Pesquisar dados, o que as palavras-chave estão sendo usadas nos mecanismos de pesquisa, vídeos, mídias sociais. Agora, especialmente com nossos smartphones móveis, também podemos rastrear localizações das pessoas. Junto com esse desenvolvimento de diferentes tipos de dados, também desenvolvemos novas técnicas analíticas ANOVA, regressão.

Estes são os seus modelos lineares, e eles eram muito populares e ainda são muito populares hoje. Eles são a força de trabalho que acredito, prevendo e análise. Nos anos cinquenta, os bayesianos as estatísticas realmente se tornaram importantes e também em torno disso pesquisa de operações temporais, escala multi-dimensional, modelos de participação de mercado para se tornar importante. Nos anos setenta, começamos a analisar a análise conjunta. Nós meio que tivemos filosofias sobre classes latentes, classes não observáveis e modelos estruturais.

Vamos nos aprofundar um pouco mais em algumas dessas linhas do tempo principais. Eu acho importante para entender a história de análise e dados. Apenas um oferece uma abordagem para o seu estudos de análise. Se você estiver olhando, deseja aprender análises de um contexto histórico, por exemplo, aprender ANOVA, regressão primeiro e então seguindo em frente, acho que é uma ótima caminho a percorrer, porque especialmente desde essas técnicas se tornam cada vez mais sofisticadas à medida que avançamos no tempo.

Também acho que nos dá como um sentido relativo de onde o mercado está indo em termos de análise. Em 1910, Charles Parlin começou a coletar dados e mercados para fins publicitários. Isso encorajou grandes empresas para estabelecer comercial departamentos de pesquisa ou marketing. Obviamente, na época, se você queria saber sobre seus clientes, uma ótima maneira de fazer é simplesmente conduzir um questionário pesquisa, saia e pergunte a eles. Isso ainda é um método popular hoje. Eu sei que temos variantes de coisas como um grupo de foco, mas estamos essencialmente pedindo diretamente ao consumidor suas opiniões.

Em 23, A. C. Nielson fundou uma empresa de pesquisa baseada em produtos em uma loja. Em 24, eles começaram a procurar nos dados de rastreamento ocular, onde os olhos das pessoas nós estamos olhando. Nos anos 30 e 50, Nielson começou a procurar rádio e televisão públicos como um meio de dados.

Na década de 1940, as empresas começaram a usar os dados do painel para registrar suas compras ao consumidor. Os dados do painel são longitudinais dados e dados em vários níveis. Essas técnicas realmente

começaram a se tornar mais sofisticadas. Um grande marco foi o advento dos computadores em pesquisa de marketing. Agora que tínhamos computadores, poderíamos fazer mais sofisticação tipos de análise. Então, em 72, outro o principal marco foi que o produto universal código ou código de barras, foi inventado pela IBM. Foi realmente adotado em lojas de varejo e seus sistemas de ponto de venda.

Isso é significativo porque, inicialmente, eu acho a ideia era realmente sobre controle de estoque. O que está sendo comprado na loja e o que é sendo vendido na loja. Mas o mercado começou a perceber que isso era um ótimo.

Fonte de dados. Poderíamos descobrir não apenas o que foi vendido e a que horas, mas o que foi vendido em conjunto com esse produto. Esse foi um grande marco real. Mais tarde, eles foram capazes para alavancar cartões de fidelidade.

Para que eles pudessem adicionar o dimensão adicional de quem estava comprando essas coisas e a que horas? Esse foi um grande marco. Em 79, a publicidade realmente começou a ser medida e se tornou mais importante. Em 1981, o computador pessoal realmente começou a decolar, e realmente permitiu mercados para armazenar dados em nível local e pesquisa facilitada pesquisa através de pessoal, através de indivíduos e entrevistas por telefone. Eu acho que a razão pela qual isso é importante é porque, antigamente, quando alguém dizia computador ou o que era falando sobre onde esses computadores gigantes de mainframe que foram desenvolvidos em 66, eles foram máquinas multimilionárias, você tinha tempo para compartilhar. Então você tinha que entrar neste computador gigante para poder acessar qualquer desses algoritmos. Mas com um PC, muito parecido com o laptop, você pode estar usando para assistir a esses vídeos, podemos armazenar localmente alguns dados e executar análises técnicas. Em 95, o mundo da Internet nasceu. Eu poderia apenas ressaltar que era realmente Aberto ao público. Agora temos muitos dados.

Em 98, o Google realmente começou a fazer isso informações na web, por meio de pesquisas de palavras-chave e as próprias pesquisas de palavras-chave se tornaram uma fonte de dados. Então, essas são as grandes mudanças, no final do século passado.

Em 2004, o Facebook foi lançado. Agora sabemos mais sobre nossos consumidores, não apenas suas preferências, mas suas preferências particulares, seus hobbies, coisas eles gostam de fazer, no que estão clicando, nos dá indicações de suas preferências. Também sabemos quem seus amigos são? Como são os amigos deles? Agora temos o YouTube, que nos permite entender que tipo de vídeo que as pessoas estão assistindo? Isso também se tornou uma fonte de dados.

Então, em 2007, com o iPhone da Apple, agora com smartphones em geral, temos sistemas de GPS em nossos telefones e permite não apenas capturar o que as pessoas são fazendo, mas onde eles estão e quais palavras-chave está procurando? Então, se você está no meio da Times Square e está procurando um restaurante, esse é outro ponto de dados que os profissionais de marketing podem usar. Em termos de análise, acadêmicos, bem como profissionais realmente começaram a investigar dados em técnicas diferentes. Modelos de difusão foram desenvolvidos e estavam enraizados nas estatísticas e envolveu alguns premissas distributivas dos indivíduos. Quais eram as preferências deles? Começamos a ver coisas como análise conjunta. A análise conjunta é realmente algo que veremos nesses vídeos. Isso foi do trabalho de luce em psicométrica.

Tínhamos dados em painel e agora usamos modelos de logit multinomiais para olhar para este tipo de dados. O modelo de logit foi também uma ferramenta popular em análise de marketing e foi uma ótima maneira de escolha de entendimento que os consumidores possam fazer. Dados de séries temporais. Vimos séries temporais dados um pouco nos módulos financeiros, mas também é algo frequentemente usado em análise de marketing. Algo para estar ciente, mas não vamos realmente ter tempo de olhar agora são modelos bayesianos.

Isso é totalmente novo classe de estatística, toda uma nova filosofia de estatística baseada no teorema de Bayes. Então, para alguns de vocês que podem ter alguma teoria das probabilidades, provavelmente estão cientes do teorema de Bayes. Esse tipo de abordagem ajuda incorporar a aprendizagem, esta missão de aprendizagem.

É tudo o que vou dizer sobre Estatísticas Bayes e Bayesianas. Também da economia, decisões de lucro também foram aplicadas a pesquisa de marketing. Então, coisas como decisões sobre publicidade ou decisões sobre o tamanho da sua força de vendas ou em quais mercados sua alvo e coisas assim. Agora, esses são alguns dos tipos reais de análise sendo usado hoje. Temos também estruturas modelos que entraram em jogo. O que isso significa é, se há uma mudança na política ou uma mudança repentina no atitudes do consumidor, que tipo de diferenças acontecem e como entender essas diferenças?

Vimos uma espécie de cronograma das diferentes fontes de dados que os profissionais de marketing eram capazes de usar, além de, uma linha do tempo dos diferentes técnicas analíticas que têm avenidas ao longo do tempo. Agora estamos realmente começando a ver algumas novas técnicas sendo usadas. Por exemplo, agora estamos olhando coisas como, nos dias de [inaudível], nós olhamos para a massa personalização, o que significa que fizemos uma oferta ao público de massa para atrair clientes. Mas agora, com a análise,

somos capazes de segmentar o mercado para que possamos direcionar nossa publicidade para baixo para um pequeno grupo de pessoas. Agora no online especialmente o ambiente, somos capazes de atingir indivíduos e seus preferências de nível individual. Tenho certeza que você provavelmente viu isso, talvez você esteja olhando sua página de mídia social e você é direcionado a isso e você pensa consigo mesmo, como eles sabiam que eu queria isso?

Então agora eles estão mirando você no nível individual e tudo isso é baseado no Analytics. No próximo conjunto de vídeos, apresentarei algumas das técnicas básicas de análise de marketing. Espero que você tenha uma compreensão do marketing de hoje ambiente de análise.

## 12 Satisfação dos clientes

### 12.1 Décimo segundo vídeo da primeira semana

Neste vídeo, eu gostaria para falar um pouco sobre como entendemos satisfação do cliente e o que é cliente satisfação. Então vamos começar. E daí é a satisfação do cliente? Ouvimos esse termo muito em marketing, mas o que isso realmente significa? Satisfação do cliente é uma medida, é algo que medimos dos produtos e serviços fornecido por uma empresa.

Portanto, seja qual for a oferta da empresa, a fim de gerar receita, seus produtos, serviços e é uma medida de se uma empresa é ou não capaz de atender ou não expectativa de um cliente. Então tem um cliente expectativa, e nós encontramos isso expectativa ou não? É importante porque satisfação do cliente é um indicador de se ou não consumidor; um, pretende comprar alguma coisa. Dois, eles vão voltar e comprar novamente. É algo que queremos entender como profissionais de marketing. Existe um modelo importante usado para nos ajudar a entender e desenvolver nossas equações e é chamado de modelo de desconfirmação, e é baseado em a comparação da expectativa do cliente e sua taxa de desempenho percebida.

Nós encontramos os seus expectativa ou não. Antes do cliente compra algo ou utiliza um serviço da empresa, eles têm algum tipo de expectativa prioritária ou expectativa que eles tinham antes. Isso pode vir de recentes publicidade de uma empresa, a comunicação da empresa recente. Poderia ser apenas o reputação da marca.

Algumas pessoas comprem produtos apenas com base no marca de uma empresa. Poderia ser boca a boca apoio de seus amigos. Poderia ser da mídia comentários ou críticas online. Vemos isso muito agora, especialmente em coisas como Amazon.com, onde estamos Ao comprar um produto, lemos os comentários. Também poderia basear-se em sua experiência passada com essa empresa ou esse produto. Talvez você tenha algo, um par de sapatos, que você realmente gostou e eles se desgastaram, então você acabou de comprar os mesmos novamente. Então é assim que o cliente expectativas são definidas. Mas depois que eles comprem, agora estão avaliando ou refletindo sobre sua compra. Eles pensaram em eles mesmos, "Sim. Isso foi um bom negócio. Essa foi uma ótima compra."

As coisas que podem impacto que são a qualidade geral de o produto ou serviço, talvez a interação deles com a empresa, suporte ao cliente, o pessoal da venda, etc. Foi o processo de compra o bem ou serviço fácil? Isso foi difícil? Valeu a pena? Qual foi o preço e valeu a pena? Talvez um concorrente tenha um produto ou serviço semelhante oferecido a um preço mais baixo. Existem intangíveis como status social ou imagem, coisas assim. Antes de eu aludir às expectativas sendo cumpridos ou não, mas há um positivo e um aspecto negativo para isso. Se a empresa exceder minha expectativa em termos de seu produto ou serviço, então eu estou realmente feliz e estou muito satisfeito.

No outro lado da moeda, se eles não atenderem ao expectativas e há uma percebida falta de qualidade no produto ou serviço, ou talvez o cliente serviço não foi tão bom, então você está insatisfeito. Portanto, a expectativa não é atendida. Então, se você estiver satisfeito, isso significa que suas expectativas foram atendidas e você está muito feliz com a transação.

Então, as expectativas são as antecipação dos clientes, o que eles esperam divulgar de um produto ou serviço. O desempenho é o percepção dos clientes reais desse produto e desconfirmação é a diferença. Como mencionei antes, pode ser positivo ou pode ser negativo. Geralmente, isso é medido em algum tipo de escala. Tenho certeza que você já viu isso questionários antes. Você esta muito feliz com o produto? Você é apenas feliz com o produto? Você é [inaudível] ok com o produto? Um pouco insatisfeito e extremamente insatisfeito. Então, isso chega a positivo e desconfirmação negativa. Então, como avaliamos as expectativas de um cliente e avaliar satisfação do cliente?

Existem vários métodos. Estes são apenas alguns. Nós poderíamos construir um sistema de reclamações ou um sistema de sugestões para coletar o feedback do cliente. Amazon obviamente tem um ótimo mecanismo de feedback, onde ele pode descrever e classificar o produto que você compra. A Starbucks tem MyStarbucks.com, que também coleta os comentários de seus clientes. Ir às compras é outra método potencial. É aí que ele chega compradores secretos para entrar em um ambiente de varejo e fingir ser um

comprador e fazer perguntas e pergunte muitas direções.

Basicamente, isso é para entender como limpar a loja é, ou talvez obter uma compreensão do nível de serviço pela equipe de vendas e coisas assim. Um incidente crítico método pode estar olhando crítica incidências ou falhas, talvez retornos de um produto e com que frequência eles estão sendo devolvidos e por que estão sendo devolvidos. Talvez possa ser um defeito do produto ou pode ser que os clientes estejam começando a usar isso de uma maneira que não era intencional para esse produto. Então, essas são as coisas você vai querer olhar.

Satisfação do cliente pesquisa em escala é simplesmente pedir ao cliente algo como "Em uma escala de um a sete, você está satisfeito com o produto?" Então, esses são apenas alguns dos métodos básicos, e há uma infinidade de métodos lá fora.

Então, vamos dar uma olhada no processo de pesquisa de clientes. Então esses são os quatro etapas básicas. Mas cada passo é bonito crítico se você deseja obter uma completa informações precisas e lembre-se também de que você está tempo do cliente. Então, você quer ser eficiente sobre isso e certifique-se de cobrir o máximo de terreno possível com o menor tempo possível questionário possível.

Então, uma coisa a considerar é qual é o propósito da pesquisa? Você está tentando controlar um novo produto e entender quais são as alterações e se isso é ou não algo que é desejado, ou você está olhando para algo eles compraram no passado e provavelmente repetir a compra? Coisas assim. Então isso é o objetivo de uma pesquisa. Próxima coisa que você quer pensar é: quem você deve perguntar? Então, quem chega ao seu público-alvo é para esta pesquisa.

Você quer entender talvez talvez jovens adolescentes e suas opiniões sobre um produto ou está procurando em novas mães, ou você está olhando pais adolescentes de adolescentes? Coisas assim. Então quem é o seu público-alvo e como você vai provar esse público?

Então você quer realmente comece a pensar nas questões individuais. O que deve ser perguntado? Como eu disse antes, você quer tê-lo curto e doce. Você não quer queimar o cliente porque seu tempo é valioso também. Então você quer ser respeitoso disso, mas ao mesmo tempo, tentando obter o máximo de informação possível. Então, finalmente, uma vez que você coletou os dados, você realizará sua análise dos dados e comece a interpretar os dados e tirar conclusões. Observe que esse processo é não necessariamente linear como essas quatro setas parece implicar. Você pode analisar os dados, tire algumas conclusões.

Pode abrir novas perguntas e você pode querer ir de volta ao começo e pense em uma nova pesquisa para refinar o seu pensamento. Então aqui está um pouco exemplo simplificado de um questionário. Isso tem cinco níveis de satisfação, de extremamente insatisfeito muito satisfeito. Esta é uma escala de 5 pontos. Você pode querer pensar sobre uma escala de 7 pontos.

Algumas escalas terão um número par de categorias para que haja nenhuma categoria intermediária, onde o respondente pode classificar de apenas escolher diferentes, eles têm que escolher de um jeito ou de outro. Então você tem que pensar sobre o tipo de escala que você deseja. Queremos medir satisfação do cliente. Então nós queremos entender sua medida ou suas expectativas antes de comprarem o produto ou serviço, e depois medir o desempenho e, em seguida, medir o diferença entre os dois.

Isso realmente nos ajuda a entender se as expectativas são ou não foram excedidos, atendidos ou não atendidos. Então, perguntas em os questionários nos ajudarão a avaliar isso. Então, a primeira coisa que queremos medir é as expectativas. O que o cliente pensou antes de comprar o produto ou serviço? Às vezes, eles podem ser convidados para relembrar a situação pouco antes de comprá-la e perguntar o que eles pensaram. Eles acham que a bebida seria doce ou não doce? Eles achavam que o carro teria uma ótima milhagem ou não?

Coisas assim. Em seguida, você desejará medir o desempenho. Os clientes são solicitados a avaliar o produto e serviço e estamos tentando entender como eles sentiu sobre essa decisão. Por exemplo, a bebida é doce, eles estavam satisfeitos. Eles queriam doce e eles ficaram doces. Então esse é um exemplo de como medir o desempenho do seu produto ou serviço. Então você quer meça a diferença.

Quão longe você está? Durante o processo, o cliente pode atenda às suas expectativas, supere suas expectativas ou não fique satisfeito. Então isso é algo que você quer descobrir e até que ponto a diferença existe. Você realmente excedeu a expectativas ou foi apenas, é melhor do que eles pensavam? Um pouco melhor do que eles pensavam. Então isso é algo você quer medir.

Então o resto é bastante simples. Você apenas implementará um modelo de regressão linear e lembre-se de que este é o modelo para regressão linear e que  $y$  é o seu variável dependente ou sua variável de escolha ex são as diferentes fatores que entram nele.

## 13 Métricas e técnicas de escalonamento

### 13.1 Décimo terceiro vídeo da primeira semana

Neste vídeo e no próximo conjunto de vídeos, gostaria de falar um pouco sobre técnicas de medição e dimensionamento. E realmente é sobre entender o que as pessoas estão pensando, quais são suas atitudes?

E no marketing, quais são as atitudes em relação a um produto ou serviço específico. Se você fez algumas aulas de psicologia, pode ter visto essas técnicas para entender atitudes e emoções. E no marketing, é praticamente a mesma idéia, exceto que estamos tentando entender como eles pensam sobre um produto ou serviço. Então vamos começar.

Então, a primeira coisa que quero falar um pouco é: o que é medição? E a medição é um processo para atribuir números ou outro tipo de símbolo às características de um objeto, de acordo com algum conjunto de regras pré-especificado.

E a segunda qualidade de uma medição é que ela deve ser isomórfica. Isso significa que deve haver algum tipo de correspondência individual entre as atribuições de números ou as atribuições de símbolos à característica real que está sendo medida. E a seguir, as regras para atribuir os números devem ser padronizadas e aplicadas de maneira uniforme, para que você trate cada medição da mesma forma. E isso é importante principalmente porque queremos poder usar técnicas estatísticas ou técnicas analíticas para avaliar esses números. E a maioria das técnicas usava números como dados.

Embora, eu diria que sempre é esse o caso, mas até recentemente, fizemos grandes avanços no aprendizado de máquina que podem receber dados de texto ou análise de texto. Mas esse é o tipo de abordagem tradicional e mais tradicional. E então, um conceito relacionado de medir algo é escala, e a escala de algumas maneiras pode ser considerada apenas uma extensão da medição. Portanto, é também o processo de atribuir um número a vários graus de opinião. Quão forte você se sente sobre algo? Quanto você gosta de algo, de suas atitudes e de outros conceitos semelhantes? E então essa é basicamente a idéia, de dimensionar. Então, se você quiser dar um exemplo do mundo real, diga para seus amigos, o que você achou do filme? Eles podem dizer sim, é ótimo, eu gostei. Eu não gostei. Mas o que você está procurando é algum tipo de resposta padronizada entre todas as pessoas na platéia, para que você possa realizar algumas análises e tirar algumas conclusões sobre isso. A escala envolve um continuum no qual as medidas estão localizadas. Portanto, há uma escala de frio a quente, de um a dez. Há algum tipo de continuum lá.

Consiste em um ponto alto e um ponto baixo, bem como em qualquer ponto intermediário, de frio a quente. Gostei um pouco, gostei muito, coisas assim. Então agora que temos uma sensação de escala, [COUGH] vamos abordar esses tópicos com mais detalhes no próximo conjunto de vídeos.

## 14 Técnicas de medição e dimensionamento - Escalas primárias de medição

### 14.1 Décimo quarto vídeo da primeira semana

Então, nós queremos coletar algumas medições.

Primeiro, precisamos entender algumas escalas primárias de medição. Isso é útil quando você está começando a pensar em projetar uma pesquisa ou um questionário, tentando coletar dados sobre as atitudes de um determinado produto ou serviço. Então vamos mergulhar as escalas primárias. Essencialmente, existem quatro escalas primárias que todos vocês devem conhecer. Há nominal, dados ordinais, de intervalo e de escala de proporção. Eles são todos um pouco diferentes. Os dois primeiros são considerados variáveis categóricas e, em seguida, as duas últimas são escalas numéricas.

Eu não deveria ter dito variáveis, mas são escalas categóricas e escalas numéricas. Então, vamos olhar para eles, por sua vez.

Primeiro, o nominal escala é provavelmente uma das mais grosseiras e a mais simples das escalas de medição. É uma rotulação figurativa esquema no qual os números servirão apenas como rótulo ou etiqueta para classificar um objeto. Então, um exemplo comum seria atribuir um número ao sexo de uma pessoa. Portanto, você pode ter zero no sexo masculino, um no feminino ou vice-versa. Realmente não importa. Mas é apenas algo para distinguir entre os dois grupos.

Às vezes, uma escala nominal é considerada um identificador único para uma pessoa ou um objeto. Nesse caso, há uma rigorosa correspondência um-para-um entre o número e esse objeto, então este é o objeto um, este é o objeto dois, o objeto três. Alguns exemplos comuns de um identificador exclusivo podem ser seus alunos número de identificação.

Nos Estados Unidos, seria o seu número de segurança social, algo parecido. Em seguida, temos o que é chamado de escala ordinal. Então uma escala ordinal é um mecanismo de classificação de variáveis



categóricas que é usado para atribuir classificação aos objetos com base em alguma característica. Permite determinar se um objeto tem mais ou menos de alguma característica do que outro, mas não diz você por quanto. Um exemplo de uma escala ordinal seria algo como classificação de classe. Então você é um calouro na faculdade, um estudante do segundo ano, um júnior ou um sênior.

Para podermos codificar isso com números; 0, 1, 2, 3, 4, mas isso realmente não diz você a que distância eles estão. Não temos concreto compreensão das regras subjacentes que determine seu status de classe. Assim é a quantidade de distância entre um calouro e no segundo ano o mesmo que entre um junior e senior?

Nós realmente não temos um bom senso disso. Mas a ordem importa. Então sabemos que os alunos do segundo ano vêm atrás de calouros, os juniores vêm depois do segundo ano e os seniores vêm depois dos juniores, e é por isso que a ordem importa. Você pode querer pausar o vídeo por um segundo e ver se você pode venha com outros exemplos. Vou fazer uma pausa por um segundo e deixe você pensar sobre isso. Então, outro exemplo pode ser posto militar, privado, cabo, sargento. Outro exemplo pode ser sua classificação em uma organização, um funcionário novato, supervisor, gerente, gerente sênior, aquele tipo de coisa. Então isso encerra as variáveis categóricas.

Existe o nominal e o ordinal. Realmente, a única diferença é que dados de escala nominal, ordem não importa; masculino, feminino, para zero e um, é um exemplo comum, cor dos olhos, vermelho ou azul, vermelho, quem tem olhos vermelhos? Preto, marrom ou azul. Então, essas são três cores nominais dos olhos que você pode ver. Realmente não importa se o azul vem antes do preto ou marrom vem depois do azul, etc. Tudo o que sei também são variáveis categóricas em que a ordem importa; classe, militar classificação, coisas assim. Agora temos dados de escala de intervalo.

É aqui que os números se tornam importantes. Os dados de uma escala de intervalo são classificados os atributos de modo que as distâncias sejam igualmente distanciados na balança, e é assim que a característica é medida. Uma coisa a notar é que todas as informações que está contido em uma escala ordinal também é contido em uma escala de intervalo, mas a principal diferença é que com uma escala de intervalo, você pode comparar as diferenças entre dois pontos, entre objetos.

No entanto, a localização do ponto zero ou do ponto de referência zero não é importante, não é fixo. Então, o que é um exemplo de algo que é considerado dados de escala de intervalo? A temperatura é uma dados de escala de intervalo. Então eu vou falar Fahrenheit por um minuto. Se algo é 32 graus Fahrenheit, e algumas horas depois, digamos que é a temperatura, é 34 graus Fahrenheit, podemos dizer que havia uma diferença de dois graus Fahrenheit.

Isso é dois graus Fahrenheit é a mesma diferença entre 60 e 62 graus Fahrenheit. Então podemos pegar as diferenças e comparar essas diferenças, e comparar essas diferenças teria significado. Mas se você pensar sobre o ponto zero, é arbitrário. O que é zero graus Fahrenheit? É um arbitrário ponto na escala.

Realmente não tem significado. Enquanto na escala Celsius, o ponto zero, está atrelado a o ponto de congelamento da água, mas também é um ponto arbitrário. A razão pela qual é chamado dados de escala de intervalo é que você só pode observe os intervalos, a diferença entre dois pontos de dados. O que você não pode fazer é usar proporções.

Por exemplo ontem estava 10 graus lá fora, hoje está 20 graus lá fora, e você não faria, normalmente conversa ou sempre, digamos que hoje é duas vezes mais frio como estava ontem. Você pode dizer que foi 10 graus mais quente, mas você não diria que é duas vezes mais quente ou duas vezes mais frio. Então você não diria esses tipos de coisas. O ponto zero é arbitrário e não realmente tem significado. Se o ponto zero tem significado, você tem algo que é chamados dados de escala de proporção. Portanto, os dados da escala de proporção possuem todas as características de dados nominais, dados ordinais e dados de escala de intervalo. Então tem ordem. Você pode comparar o diferença entre dois pontos em um intervalo escala e escala de proporção, mas zero tem significado. Existe um elemento chave para a escala zero.

Assim, com a escala de proporção, podemos classificar os objetos, podemos observar as diferenças. Alguns exemplos comuns inclua algo como peso. Realmente não importa qual unidade de medida você usa. Então, se algo é um quilograma e algo mais é dois quilogramas, você diria o segundo objeto pesando dois quilos pesa o dobro do peso como o primeiro objeto.

Então você pode fazer isso tipos de comparações. Zero tem significado. Nesse sentido, ele faz tem um ponto de referência. A idade é outro ponto de referência. Então, se alguém tem cinco anos e alguém tem 10 anos, pode-se dizer que o garoto de 10 anos é duas vezes mais velha como a criança de cinco anos. Então, quando você pensa em fazer essas proporções de mais de b, então você sabe que tem algo que está em formato de escala de proporção.

Então, por que isto é importante? Falamos sobre dados nominais, dados ordinais, escala de proporção dados e dados de intervalo. Esses são os quatro tipos de dados. Os dois primeiros; nominal e ordinal são categóricos e os dois seguintes; intervalo e proporção escala são numéricos. A forma ideal de dados que você quer, porque você tem mais coisas que pode fazer com esses dados dados de escala de proporção, é o melhor

e aquele que tem a menor quantidade de granularidade são dados nominais, que têm apenas categorias que nem têm ordem. Uma coisa a notar é que você sempre pode ir de dados de escala de proporção e trabalhar cada vez menos e menos refinamento. Significado, você pode ir de dados de escala de proporção para dados de escala de intervalo para dados ordinais e nominais dados, se assim o desejar. Portanto, pode haver, por exemplo, você pode ter um escala de proporção contínua de informações e, em seguida, você pode definir arbitrariamente pontos de interrupção.

Digamos, se estiver abaixo de 100, isso é considerado bom. Então, se for mais de 100, isso é considerado ótimo. Então agora você tem esses dois ordinais categorias, ok e ótimo. Obviamente, você pode quebrar mais um pouco. Então, você pode até fazer tudo bem e ótimo e não tem mais detalhe em torno disso. A chave é que se você estiver dado um dado em forma ordinal; bom mau. Vamos fazer em ordem; justo, bom, ótimo. Então, essas são três categorias que você pode querer avaliar atitudes sobre um produto ou serviço. Você realmente não pode ficar mais granular compreensão desses dados e você não pode ir desse ponto de dados para dados de escala de proporção, mas você pode ir de dados de escala de proporção até ordinais ou mesmo classes nominais. Então isso termina as escalas primárias.

## 15 Técnicas de medição e dimensionamento - dimensionamento não comparativo

### 15.1 Décimo quinto vídeo da primeira semana

Neste vídeo, eu gostaria de falar sobre alguns não comparativos técnicas de dimensionamento. Tão não comparativo técnicas de dimensionamento, algumas vezes referidas como escalas métricas, cada objeto ou serviço é escalado independentemente. Você não está comparando com outra coisa. É por isso que eles são chamados de não competitivos técnicas de dimensionamento. De um modo geral, os dados podem ser contínuos, por isso é intervalo ou dados de escala de proporção.

Eles geralmente podem ser contínuos, mas também podem ser escalas de classificação discriminadas. Vamos examinar algumas das técnicas mais usadas na pesquisa de marketing. A primeira escala que eu gostaria de falar é conhecido como a classificação contínua escala e às vezes é referido como uma escala de classificação gráfica. Veremos isso em um minuto.

Basicamente, pede-se aos entrevistados que joguem algum tick marca ou alguma marca em uma linha contínua e que onde eles colocam essa marca reflete suas atitudes em relação a algum produto ou serviço. Observe que eles não estão realmente restrito a selecionar marcas anteriormente definido pelos pesquisadores, eles são livres para colocar uma marca em qualquer lugar nessa linha e a forma do

A escala contínua pode variar consideravelmente, então há muitas variações sobre isso. Mas, basicamente, parece algo assim. Existe uma linha horizontal. Você pode até colocar de 0 a 100 para dar a eles pontos de referência, menos preferidos mais preferido.

A pergunta seria algo sobre uma bebida gaseificada, a efervescência, por exemplo, o sabor, o teor de açúcar, a cor da lata, etc. Em seguida, quero falar sobre uma escala de classificação detalhada. É aqui que os entrevistados recebem uma escala com números ou uma breve descrição associada a eles. As categorias são ordinal em que eles são ordenados em termos de sua posição de escala. O entrevistado é perguntado para selecionar uma categoria que eles sintam mais refletida suas atitudes. Portanto, existem alguns comumente utilizou escalas de classificação discriminadas. Nós falaremos sobre eles.

Há a escala Likert, há a semântica escala diferente, e há algo conhecida como escala Stapel. Então esses são os três escalas mais usadas. Eu imagino que você vi essas escalas em suas próprias vidas ou pode até ter tomado essas tipos de pesquisas. Então a primeira escala que eu gostaria de falar é conhecido como a escala Likert. Solicita ao entrevistado que indique em que grau concorda ou discorda de uma declaração sobre o produto ou serviço. Geralmente, tem cinco respostas que variam de discordo totalmente concordo plenamente. Às vezes, é conhecida como escala somada porque você pode somar a pontuação para obter um total pontuação para o respondente. Então, vamos olhar um exemplo. Então, eu tenho certeza que você já viu algo assim antes.

Aqui estão as cinco categorias, discordo totalmente de concordar, e há uma coluna do meio que é a nossa categoria neutra. Então, aqui estão algumas declarações que você pode pedir a um respondente. A empresa faz ótimos produtos; concorda, discorda, em algum lugar no meio, tem ótimos relacionamentos com seus fornecedores, ótimas relações com seus clientes, o produto é realmente durável? O preço é razoável? Então, esses são os tipos de perguntas que você pode fazer. Em seguida, temos uma semântica escala diferencial.

É semelhante a uma escala Likert. Geralmente, é um ponto de sete classificação com os terminais que representam extremos ou rótulos bipolares que têm significado semântico. Portanto, os entrevistados classificam os objetos em uma escala de sete pontos e os adjetivos pode ser alto-baixo, quente-frio, algo

dos extremos. Aqui estão alguns exemplos de uma escala diferencial semântica. Então, avalie o monitor de frequência cardíaca. Essa é a nossa teoria exemplo de produto que vamos fazer perguntas sobre. Aqui estão alguns atributos.

O monitor de frequência cardíaca é preciso ou não. Então existem os extremos polares, e então o respondedor será solicitado a marcar algo entre isso; perto de preciso, ou não preciso, ou em algum lugar no meio. Confortável de usar, não é confortável, é barato ou caro, é confiável, não é confiável. Esses são alguns exemplos de escala diferencial semântica. Finalmente, temos a escala Stapel. Esta é uma escala de classificação que geralmente tem 10 categorias. Varia de menos cinco a mais cinco. Geralmente, tem um ponto neutro zero, para que eles tenham que decidir um de um jeito ou de outro. A escala é geralmente apresentado verticalmente. Solicita-se aos entrevistados que indiquem selecionando categoria apropriada, como eles se sentem produto ou serviço. Uma das vantagens é que não exigir um pré-teste dos adjetivos ou frases para garantir que eles sejam verdadeiro por polaridade. Você pode apenas usar o escala imediatamente. Aqui está um exemplo disso. Equipe amigável, excelente comida, ótimo ambiente. Essas são coisas que você pode pensar são as qualidades de um restaurante. Então, a escala aqui é menos cinco a mais cinco. Então, o que você acha do precisão da declaração? Com uma noção pré-teste, de volta à semântica escala diferencial e monitor de frequência cardíaca, falamos sobre preciso, não preciso, então você pode ter que pré-testar essas palavras para portanto, é necessário pré-testar essas palavras para garantir que você esteja realmente medindo o que pensa estar medindo. Confortável, não confortável, o que significa estar confortável? Isso significa que é leve e eu realmente não percebo que quando eu coloco isso, ou isso significa que é macio e mole? Então, o que é confortável realmente significa? Esses são os tipos de coisas você pode querer fazer um pré-teste.

Enquanto na escala Stapel, você pode colocar os termos e então eles são solicitados a avaliar esse item. Então isso envolve tudo escalas não competitivas. Analisamos a escala Likert. Vimos a semântica escala e escala Stapel. Eu mostrei alguns casos básicos. Mas quando você está realmente Ao criar uma pesquisa, há outras coisas que você pode querer considerar. Por exemplo, o número de escalar categorias para usar. Sugerir 10 aqui, mas pode haver casos em que você pode querer mais ou menos granularidade. Na escala Likert, existem apenas cinco categorias. Algumas pessoas argumentam que sete é melhor semelhante ao a escala semântica. Você geralmente quer fazer certifique-se de que suas escalas sejam equilibradas que, por exemplo, na escala Stapel você tem cinco; um, dois, três, quatro, cinco no positivo direção, um, dois, três, quatro, cinco em a direção negativa e você quer fazer verifique se há uma quantidade igual nos dois lados. Você não quer ter cinco na boa direção e apenas três na direção negativa, porque isso enviesar seus resultados.

Algo para pensar também é se você tem ou não um número ímpar de categorias ou um número par de categorias. Além disso, se deve ou não você quer forçá-los a fazer uma escolha ou não faça uma escolha. Você os força a fazer uma escolha, não dando-lhes um neutro categoria ou resposta. Então, também o físico forma da escala. Então, se isso é uma caneta e escala de papel, você pode usar essa linha onde eles marcam ou é um tipo verbal de pesquisa. Então, essas são algumas das coisas que você convém pensar enquanto cria suas pesquisas.

## 16 Design de experimentos: conceitos-chave

### 16.1 Décimo sétimo vídeo da primeira semana

Neste vídeo eu gostaria de conversar um pouco pouco sobre o design de experimentos, então vamos começar. E antes de começar, eu quero mencione que o que estou prestes a discutir é uma espécie de ouro padrão de desenho experimental. Mas isso nos ajuda crie uma linha de base da aparência da experiência ideal. E você quer ao projetar um experimento que você deseja estabelecer uma estrutura para comparar tratamentos ou grupos em termos de alguns resposta mensurável. E você pode criar experimentos em maneiras diferentes baseadas em condições diferentes, baseadas em interesses diferentes, tudo depende dos seus objetivos.

Mas realmente o que é chave e essencial que você desenvolva um plano sistemático para avaliar antes de executar o experimento. Então, aqui estão os contornos básicos de um de criar um plano. Você deseja determinar primeiro seu objetivo de pesquisa. Qual é o objetivo do estudo? O que você realmente está tentando descobrir? Determine os tratamentos e selecione quais fatores serão variados. Então, se isso veio de um ambiente médico você pode variar o tipo de medicamento e, em seguida, variar o tipo e o nível de dosagens.

Em um estudo de marketing, você pode testar se uma página da web é ou não mais eficaz em atrair clientes para clicar no botão comprar, se estiver vermelho, ou azul, ou uma fonte grande ou uma fonte pequena, esses são os tipos de coisas que você pode olhar. E você também quando está meio que no estágio

inicial de planejamento, você deve identificar quaisquer fatores estranhos que possam estar em seu ambiente que possa afetar o resultado de seu experimento.

algo que você vai querer considerar. Em seguida, você deseja determinar o características de sua variável de resposta a serem medidas nas unidades experimentais.

Em seguida, você desejará determinar seu procedimento de amostragem. E o que quero dizer com procedimentos de amostragem, como você coletará os dados e quem serão seus súditos? Em seguida, você vai adquirir uma gravação das respostas, basicamente execute sua experiência, e colete os resultados. E então determine o número de unidades experimentais para cada tratamento para manter o poder do teste e confiabilidade dos intervalos de confiança. Ok, então vamos falar sobre estes em termos de primeiro o vocabulário. Mas antes de falar sobre a terminologia é bom ter um exemplo em mente que ajude você tem algo para agarrar. Então, digamos que você é um mercado pesquisador e você está projetando um experimento para estudar a taxa de cliques de uma página da web do produto e, nesse caso, variaremos o preço e vamos variar o tamanho da fonte. Portanto, existem dois níveis de preços: US\$ 25 e \$ 35 e também podemos variar o tamanho da fonte em 12 pontos fonte, fonte de 20 pontos e fonte de 30 pontos. E nós vamos gravar automaticamente a taxa de cliques, digamos, por um mês.

Então, os primeiros fatores, essas são as variáveis controladas selecionadas pelo pesquisador para comparação e ajudam a formar a comparação grupos definidos pela hipótese. No nosso exemplo aqui, o que você acha que são os fatores? Bem, eles são o preço e o tamanho da fonte, então esses são seus fatores, estamos interessados em preço e tamanho da fonte. Certamente, em uma página da web, tenho certeza você pode pensar em outros fatores, mas eles serão constantes, nós não vamos mudar isso. Mas você pode querer fazer um estudo de acompanhamento e observe não apenas o tamanho da fonte, mas tipo de fonte, por exemplo, cor, etc. Medição e observações, essas são as variáveis de resposta que são registradas, mas não controlado pelos cientistas.

Então, o que vamos medir? O que vamos observar? A taxa de cliques da página da web, então essa é a nossa resposta. Tratamentos são as condições construído pelos fatores. No exemplo, a combinação de diferentes níveis de preços e tamanho da fonte no título são os tratamentos. Então você poderia ter para exemplo, fonte de 12 pontos \$ 25, fonte de 12 pontos \$ 35, 20, Estou ficando confuso. Fonte de 20 pontos \$ 25, 20 pense em fonte de US \$ 35, etc, etc, então todas as diferentes combinações para seus tratamentos.

[COUGH] Alguma terminologia, unidirecional classificação envolve um único fator, daí o tratamento e os níveis dos fatores seriam os mesmos, então é uma classificação unidirecional. E no próximo, agora não sei se é o próximo vídeo ou o seguinte, mas falaremos sobre ANOVA unidirecional que analisa esse tipo de projeto experimental. O próximo vocabulário terminológico word é design de tratamento fatorial. E o tratamento projetou um experimento [COUGH] que envolve vários fatores e os tratamentos formados pela combinação de todos os diferentes níveis de seus fatores. Então, voltando ao exemplo, há 12 fonte de ponto, fonte de 20 e 30 pontos e 25 e US\$ 30 e assim você misturaria toda a combinação. Um experimento fatorial fracionário é aquele em que apenas um número fracionário dos tratamentos possíveis são realmente usados no experimento porque o número de fatores é muito grande. No nosso exemplo simples, somos apenas olhando para dois fatores preço e tamanho da fonte, e o preço tem apenas dois níveis e frente tem três níveis. Mas se você tivesse vários fatores que você está olhando para o tamanho da fonte, preço, localização da compra botão na parte superior da página, no final da página, no meio da página. Se você estiver olhando para talvez a cor ou o plano de fundo, você poderia pensar muitas combinações diferentes. E o número de combinações se tornar muito, muito grande, então você vai querer projetar o experimento de forma a chamar apenas os principais tratamentos que você está interessado. E um tratamento de controle [COUGH] é um tipo especial de tratamento e essa é sua referência para comparar a eficácia do seu tratamento. Então você pode dizer na web exemplo de página fonte de 12 pontos \$ 25, que será sua linha de base e todo o resto será comparado com esse benchmark. Em seguida é a unidade experimental e tenho uma definição longa aqui, mas não deixe que isso te confunda. É a entidade física à qual o tratamento é designado aleatoriamente ou o sujeito selecionado aleatoriamente de uma das populações de tratamento. Por exemplo, neste estudo da página da web que Eu tenho usado para orientar os exemplos, se um pesquisador decidir exibir a página da web em 24 cidades, com cada cidade sendo um tratamento diferente. Por exemplo, a cidade A teria os 12 fonte de ponto com o preço de US\$ 20 e a cidade B teria a fonte de 12 pontos com o preço de US\$ 35, mediríamos a taxa de cliques para cada cidade e essa seria a unidade experimental. Às vezes você ouvirá o termo, qual é a sua unidade de análise? Qual é a sua unidade de análise? Qual é a sua unidade experimental? Então, qual é a unidade que você está medindo? É uma pessoa? Nesse caso, é uma cidade. É um código postal? Ai está. Ok, a seguir, replicação assim que tivermos. Depois que o tratamento for atribuído a uma unidade experimental, um único Ocorreu replicação do tratamento. Então, mas podemos querer repita isso várias vezes, por isso é chamado de replicação. E uma unidade de medida, unidade de medida é distinto da unidade experimental. A unidade de medida é a física entidade na qual é feita uma medição. Portanto, para o exemplo da taxa de cliques de nas páginas individuais, a unidade experimental

será a cidade e a medição unidade será a página individual. Então a unidade experimental é a cidade, a unidade de medida que estamos vai medir por página.

## 17 Design de experimentos: controle de erros experimentais

### 17.1 Décimo oitavo vídeo da primeira semana

Erro experimental é a variação nas respostas entre unidades experimentais, sob o mesmo tratamento e nas mesmas condições, então esse é o seu erro experimental. Então, o que é experimental erro causado por? Bem, diferenças naturais nas unidades experimentais antes de receber sua alocação ou o tratamento.

Eles também podem ser variações nos dispositivos que registram as medições. No caminho do clique, temos uma medição muito limpa, sabemos que o número de cliques. Mas se você está medindo, diga o rendimento de uma colheita em peso, as escalas podem ser diferente e então você pode ter leve diferenças de medida. Além disso, você terá variações nas condições de tratamento, essa é outra causa de erro experimental. Todos esses fatores estranhos além dos fatores de tratamento tudo isso aleatoriedade no mundo. Poderia ser algo como simples como alguém batendo na mesa e eles clicou no botão por acidente ou eles clique no botão não intencionalmente, mas por outros motivos. Então isso pode confundir seus resultados. OK. Então, como os cientistas controlam erro experimental? Queremos minimizar o quantidade de erros para que possamos fazer isso através do uso de procedimentos experimentais estritos e seguindo os procedimentos exatamente da mesma maneira, sempre, para minimizarmos os riscos experimentais erro nessa dimensão.

Nossa escolha de unidades experimentais e unidades de medida também faz a diferença. Se estamos tentando talvez medir o quão longe um objeto é se medirmos em quilômetros e arredondarmos para apenas números inteiros. Nós vamos ter menos resultados precisos do que se medíssemos em milímetros e podemos descer para o nível em milímetros.

Como gravamos os dados? Que tipo de dispositivos usamos? Eles são dispositivos precisos? Também o tipo de projeto experimental, agora estou falando de uma forma muito rigorosa e controlada desenho experimental que pode ser usado por cientistas em geral. Mas existem outras projetos experimentais dos quais você deve estar ciente, por exemplo desenho quase experimental. São coisas usadas nas ciências sociais quando não podemos realmente fazer experimentos no verdadeiro sentido da palavra. Finalmente, variáveis de controle, quais são algumas das coisas que podem afetar nossa variável de resultado, mas realmente não são interesse para o nosso estudo. OK.

Então vamos conversar um pouco sobre procedimentos experimentais e são as condições sob as quais um experimento é executado. Eles devem tentar ser o mais constante possível durante o experimento. Então a temperatura pode ser algo que você deseja pensar sobre. No exemplo da página da web, você pode querer o grau de controle possível para saber se algumas pessoas gostariam de usar a Internet quando o tempo não está tão bom lá fora. Então, se uma cidade tem muita chuva, como o noroeste, eles podem ser endossados com mais frequência vendo essas páginas da web. Se o experimental procedimentos não são seguidos rigorosamente a variação da resposta pode ser inflada e a precisão de nossas inferências ou intervalos de confiança podem ser comprometidos corretamente. Então, teremos grandes variações de variação, então estamos não será tão confiante em nossas estimativas. Por fim, o experimental procedimento deve ser conduzido ou executado da mesma maneira pela duração do experimento na medida do possível. Caso contrário, você pode obter uma variação inflada ou algum viés nos seus resultados. Um viés é consistente superestimar ou subestimar a estatística que você está olhando. Na maioria dos casos, você está olhando para algo como a média, qual é a resposta média e você pode ultrapassar essa média ou ultrapassar essa média de forma consistente e isso é o que é conhecido como viés. Variação experimental de erro Variação de erro pode aumentar se as unidades no experimento não são semelhantes em relação a essas características.

Então, se você não está escolhendo cidades do mesmo tamanho ou talvez escolher uma cidade grande e alguma cidade rural que pode fazer a diferença. Então você quer ter certeza que as unidades experimentais estão selecionadas para corresponder. Observe que se o teste experimental Como as unidades são excessivamente uniformes, as generalizações para a população podem ser restritas. Então, se você é apenas escolhendo cidades grandes, seus resultados podem se aplicar apenas para grandes cidades e eles podem não se aplicar às áreas rurais cidades por exemplo. Então, no caso de o exemplo de página da web, se estamos tentando ver como nossa página chama a atenção de uma população estudantil.

Se escolhermos alunos da mesma série ou o mesmo sistema escolar das unidades experimentais que selecionamos pode alcançar um conjunto mais homogêneo de unidades de medida. Mas as inferências de como podemos generalizar isso para a população em geral pode ser afetado pela escolha de escola ou nível escolar. Portanto, isso se aplica a um aluno da primeira série que não sei se a primeira série será clicando em uma página da web, mas o que se aplica a um estudante do ensino médio pode não necessariamente se

aplica ao comportamento de um graduado aluno por exemplo. Randomização de tratamentos, portanto, alguns desses procedimentos estatísticos baseiam-se na condição de que os dados foram extraídos de uma população que foram coletados dados de uma população que distribuído normal. Uma distribuição normal é a curva em forma de sino.

Quando você estuda um novo procedimento estatístico, há sempre uma lista de premissas,  $X$  é extraído de uma distribuição normal. Há um linear relação entre  $X$  e  $Y$  no caso de regressão, etc. Há sempre essas suposições e quaisquer que sejam essas suposições são sua amostra real dos dados pode não seguir o real Distribuição populacional. Então isso também é algo ter em mente. Então, como nós randomizamos os tratamentos? Há um pequeno processo aqui. Suponha que temos  $N$  unidades experimentais. Portanto, esse é o número total de unidades experimentais que estamos usando. No caso do exemplo de página da web, é o número de cidades que foram decididos.  $T$  é o número de tratamentos lembre-se que é uma mistura do fatores e seus níveis. Queríamos executar aleatoriamente atribuir o tratamento a alguns unidade experimental.

Observe que  $r$  é o número de replicação de replicações por unidade experimental. Então, quando somamos todas as horas, quantas vezes repita o estudo deve ser igual a  $N$  e isso é delineamento inteiramente casualizado. Você faz isso imaginando o número de unidades experimentais de um a  $N$  gera uma lista de números aleatórios que é uma permutação do números de um a  $N$ . Em seguida, atribua-lhes o número um para experimentar um para a primeira replicação, atribua o tratamento duas das unidades experimentais a a segunda replicação, etc. Depois, continue repetindo esse processo até que todas as tratamentos são atribuídos. Então isso te dá um pouco uma visão geral sobre a forma mais pura de projeto experimental em variáveis de termos de controle, suas variáveis de resultado e coisas que podem afetar sua variação e seu viés.

Então no próximo alguns vídeos, falaremos sobre o testes estatísticos envolvidos e, esperançosamente, vendo isso e, em seguida, vendo os testes estatísticos reais tudo virá juntos no final.

## 18 Entrevista com Monica Penagos - Teste A / B e ANOVA na Prática

### 18.1 Primeiro vídeo da segunda semana

Esta é a semana 2 de Análise de Marketing e nós temos conosco Monica novamente. Monica, esta semana vou falar sobre testes A / B e ANOVA. Talvez você possa dar alguns exemplos sobre como você pode usar isso em o ambiente corporativo.

Sim. Então o teste A / B é o pão com manteiga hoje no que faço. Então, como você sabe, eu sou inovação em dinheiro líder e parte do meu papel é ajudar nossas equipes para lançar suas startups. Então todas as equipes que estão trabalhando no lançamento de startups usam Teste A / B o tempo todo. Deixe-me apenas dar vocês dois exemplos. Não sei se você já ouviu falar dessas marcas, mas são marcas públicas agora e espero que todos comprem. Então o primeiro é o Zevo, é natural repelente de inseticidas. O outro que eu quero para abordar é chamado Gemz é um shampoo sem água.

Então, usamos testes A / B no Zevo, para entender a publicidade interna se erros reais ou ilustração de erros fossem mais eficazes para os consumidores. Eles vão clicar mais? Eles vão converter mais e comprar o produto? Adivinha o que eu encontrei, real erros são muito melhores.

Em nossa publicidade, fizemos Testes A / B para promover e entender como a demonstração de o produto funciona. Para Gemz, no site, essa é uma marca de beleza. Então é um tratamento capilar que vem com água que você pode escolher e escolher o coleção que você deseja. Então hoje você pode tomar banho com um perfume, amanhã com outro. Quando estávamos lançando o produto, queremos entender se avaliações e críticas impactariam a taxa de conversão que nós tínhamos no site.

Então tivemos mais de 140 classificações com 4,8 estrelas. Deve ser incluído isso em a página de destino que temos ou não deveria ser incluído nele? Então fizemos testes A / B e nossas taxas de conversão aumentaram depois que incluímos as classificações e revisar com os consumidores.

Pode parecer que sim, você deveria ter feito isso de qualquer maneira porque tinha ótimas classificações e resenhas, mas não sabíamos se valia a pena fazer essa alteração e quanto isso afetará a taxa de conversão. Direita. Então esses são alguns ótimos exemplos. Eu realmente gosto do exemplo sobre o Zevo? Zevo. Zevo, o inseticida. Eu posso ver a hipótese oposta em que erros reais podem ser tão assustadores que as pessoas simplesmente clique fora do site.

Então essa é uma visão fascinante. Esse é o tipo de coisas que você poderia fazer com testes A / B e ANOVA, e estes são os tópicos nós estaremos cobrindo esta semana.

## 19 Exercício da semana

### 19.1 Leitura do material didático

Disponível em [aqui](#) \*ver página 33.

```

1 # install.packages("mosaicData")
2 library(mosaicData)
3 data(SaratogaHouses)
4
5 SaratogaHouses
6 #####
7 ## One-Sample t-test Example ##
8 #####
9
10 mean(SaratogaHouses$price)
11
12 t.test(SaratogaHouses$price)
13
14 # Do a t-test to test whether the true mean housing price in Saratoga is $200000.
15 t.test(SaratogaHouses$price, mu = 200000)
16
17 # The p-value is 4.804e-07, which is smaller than 0.05.
18 # So, we can reject the null hypothesis under the significance level of 0.05
19 # and conclude that the true mean housing price in Saratoga is not $200000.
20
21
22 #####
23 ## Two-Sample t-test Example ##
24 #####
25
26 # Split the price data into two groups.
27 # Houses in one group have central air and
28 # houses in the other group do not have central air.
29 x <- SaratogaHouses$price[SaratogaHouses$centralAir == "Yes"]
30 y <- SaratogaHouses$price[SaratogaHouses$centralAir == "No"]
31
32 # The two groups have unequal sample sizes.
33 length(x)
34 length(y)
35
36 # Generate a boxplot of the housing prices of the two groups.
37 boxplot(x, y)
38
39 # Performed an unpaired two-sample t-test to test
40 # whether there is a difference in housing prices between the two groups.
41 # Note that the t.test(x, y) assumes unequal variances by default.
42 t.test(x, y, paired = F)
43
44 # The p-value is 2.2e-16, which is smaller than 0.05.
45 # So, we can reject the null hypothesis under the significance level of 0.05
46 # and conclude that there is a difference in housing prices between the two groups

```

Listing 7: Teste A/B no 

```

install.packages("mosaicData")

## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)
## Warning in install.packages("mosaicData"): 'lib = "/usr/local/lib/R/site-library"' is
not writable
## Error in install.packages("mosaicData"): unable to install packages

library(mosaicData)

## Error in library(mosaicData): there is no package called 'mosaicData'

data(SaratogaHouses)

## Warning in data(SaratogaHouses): data set 'SaratogaHouses' not found

SaratogaHouses

## Error in eval(expr, envir, enclos): object 'SaratogaHouses' not found

#####
## One-Sample t-test Example ##
#####

mean(SaratogaHouses$price)

## Error in mean(SaratogaHouses$price): object 'SaratogaHouses' not found

t.test(SaratogaHouses$price)

## Error in t.test(SaratogaHouses$price): object 'SaratogaHouses' not found

# Do a t-test to test whether the true mean housing price in Saratoga is £200000.
t.test(SaratogaHouses$price, mu = 200000)

## Error in t.test(SaratogaHouses$price, mu = 2e+05): object 'SaratogaHouses' not found

# The p-value is 4.804e-07, which is smaller than 0.05.
# So, we can reject the null hypotheis under the significance level of 0.05
# and conclude that the true mean housing price in Saratoga is not £200000.

#####
## Two-Sample t-test Example ##
#####

# Split the price data into two groups.
# Houses in one group have central air and
# houses in the other group do not have central air.
x <- SaratogaHouses$price[SaratogaHouses$centralAir == "Yes"]

## Error in eval(expr, envir, enclos): object 'SaratogaHouses' not found

y <- SaratogaHouses$price[SaratogaHouses$centralAir == "No"]

## Error in eval(expr, envir, enclos): object 'SaratogaHouses' not found

# The two groups have unequal sample sizes.
length(x)

## [1] 1

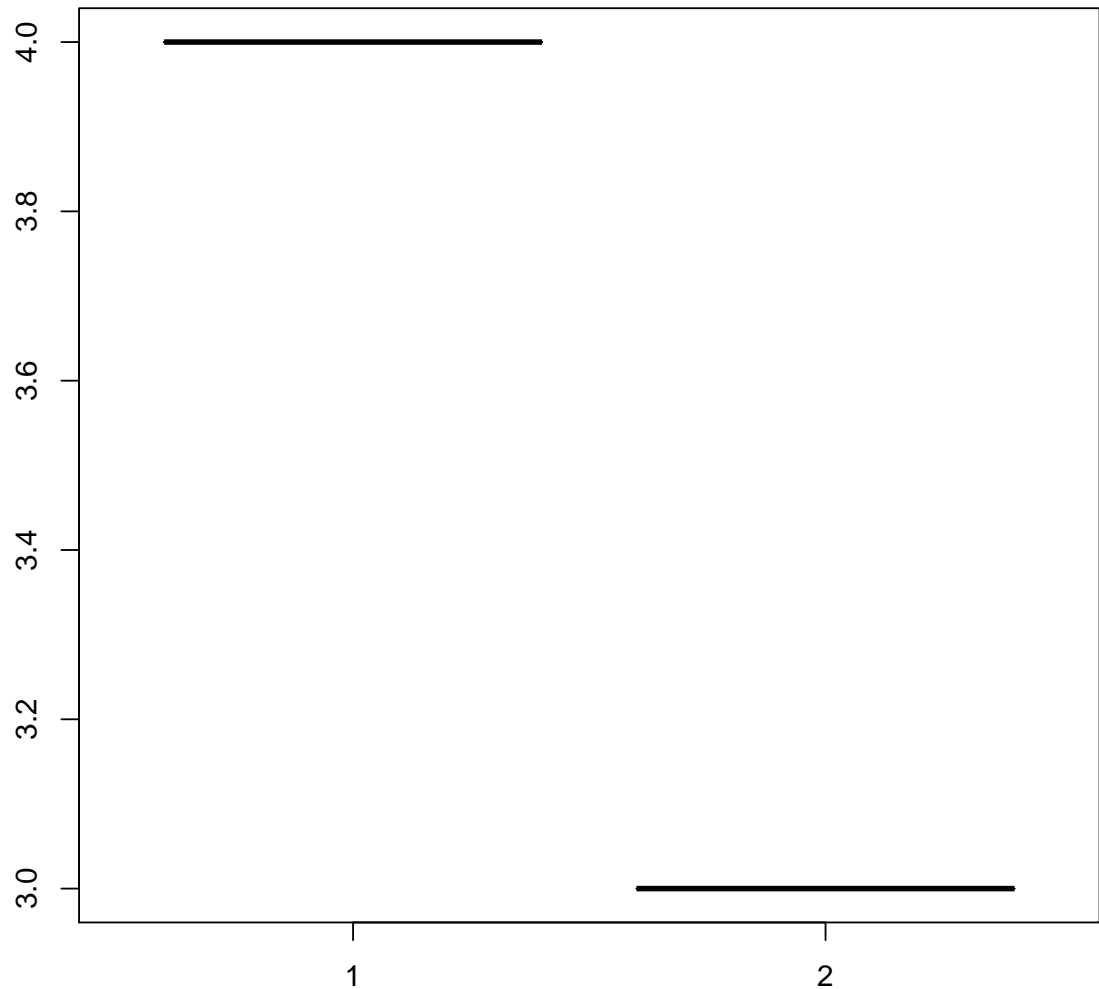
length(y)

## [1] 1

# Generate a boxplot of the housing prices of the two groups.
boxplot(x, y)

```





```
# Performed an unpaired two-sample t-test to test
# whether there is a difference in housing prices between the two groups.
# Note that the t.test(x, y) assumes unequal variances by default.
t.test(x, y, paired = F)

## Error in t.test.default(x, y, paired = F): not enough 'x' observations

# The p-value is 2.2e-16, which is smaller than 0.05.
# So, we can we can reject the null hypothesis under the significance level of 0.05
# and conclude that there is a difference in housing prices between the two groups
```

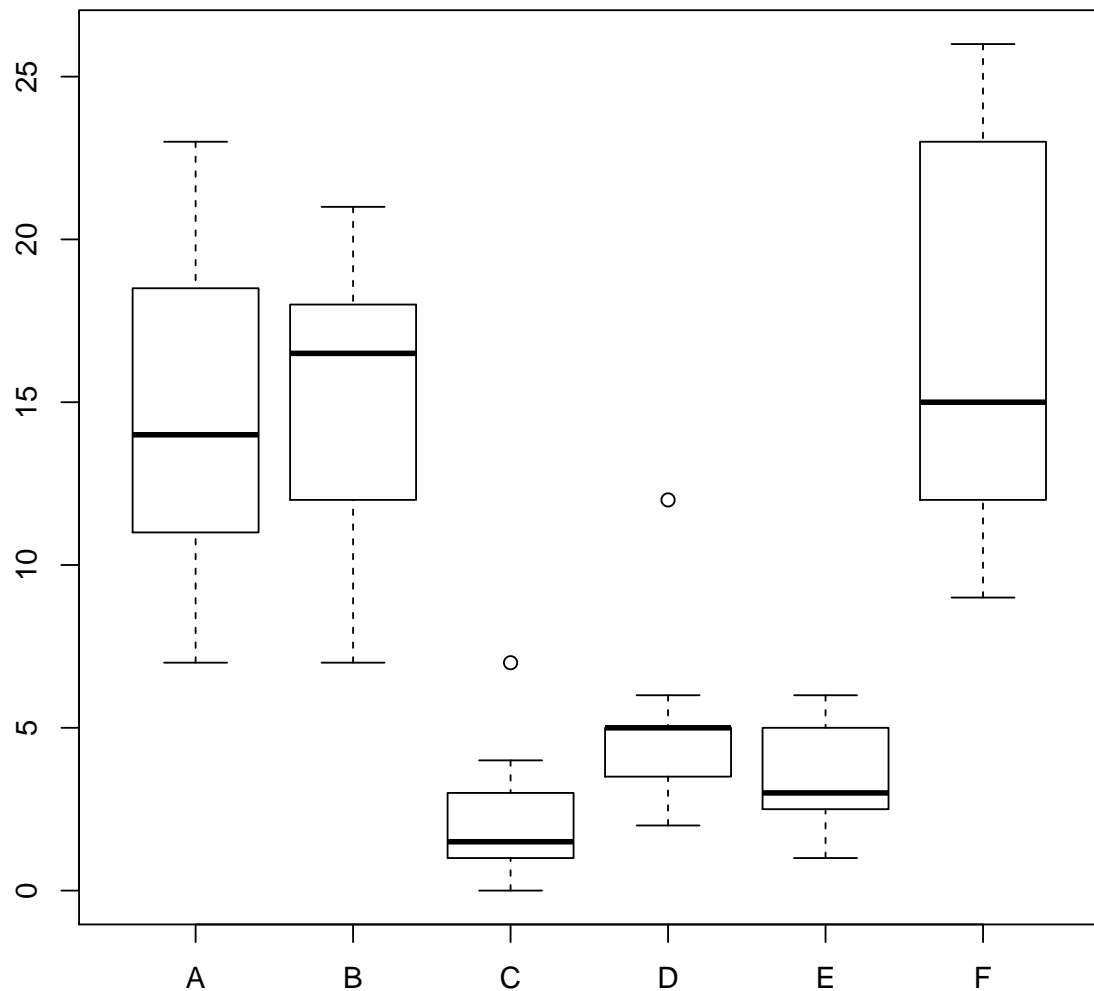
```

1 #####
2 ## One-Way ANOVA Example ##
3 #####
4
5
6 attach(InsectSprays)
7
8 # We can get a summary of this dataset
9 # The first column is the counts of insects in agricultural experimental units
10 # The second column contains 6 types of sprays
11 dim(InsectSprays)
12 str(InsectSprays)
13
14 # Each type of sprays has the same sample size of 12
15 tapply(count, spray, length)
16
17 # We can generate a boxplot of each type of the spray
18 boxplot(count ~ spray)
19
20 # Run One-Way Anova using aov()
21 anova <- aov(count ~ spray, data = InsectSprays)
22 summary(anova)
23
24 # Manually Calculate the F statistic
25 N <- nrow(InsectSprays) # number of Observations
26 n <- 12 # number of observations per group
27 k <- length(unique(InsectSprays$spray)) # Number of groups
28 grand_mean <- mean(InsectSprays$count) # Grand Mean
29 group_mean <- tapply(count, spray, mean) # Mean of each group
30
31 SST <- sum(n*(group_mean - grand_mean)^2)
32 MST <- SST/(k-1)
33
34 SSE <- sum((InsectSprays$count - rep(group_mean, each = n))^2)
35 MSE <- SSE/(N-k)
36
37 F_statistic <- MST/MSE
38 F_statistic
39
40 p_value <- pf(F_statistic, df1=(k-1), df2=(N-k), lower.tail = FALSE)
41 p_value
42
43
44 #####
45 ## Two-Way ANOVA Example ##
46 #####
47
48 # install.packages("mosaicData")
49 library(mosaicData)
50 data(SaratogaHouses)
51
52 # The response variable in this example is the price.
53 # We want to test whether heating type and sewer type
54 # will affect the housing price in Saratoga.
55 head(SaratogaHouses)
56
57 # We can check the sample sizes in each combination of the two variables
58 table(SaratogaHouses$heating, SaratogaHouses$sewer)
59
60 # If we assume the two variables are independent, we can use the additive model.
61 anova2 <- aov(price ~ heating + sewer, data = SaratogaHouses)
62 summary(anova2)
63
64 # If we also want to test whether the two variables interact to affect the housing price,
65 # we can use an interaction term.
66 anova3 <- aov(price ~ heating * sewer, data = SaratogaHouses)
67 summary(anova3)
68
69 # The command below is equivalent to anova3
70 anova4 <- aov(price ~ heating + sewer + heating:sewer, data = SaratogaHouses)
71 summary(anova4)

```

Listing 8: ANOVA original no 

```
#####  
## One-Way ANOVA Example ##  
#####  
  
attach(InsectSprays)  
  
# We can get a summary of this dataset  
# The first column is the counts of insects in agricultural experimental units  
# The second column contains 6 types of sprays  
dim(InsectSprays)  
  
## [1] 72 2  
  
str(InsectSprays)  
  
## 'data.frame': 72 obs. of 2 variables:  
## $ count: num 10 7 20 14 14 12 10 23 17 20 ...  
## $ spray: Factor w/ 6 levels "A","B","C","D",...: 1 1 1 1 1 1 1 1 1 1 ...  
  
# Each type of sprays has the same sample size of 12  
tapply(count, spray, length)  
  
## A B C D E F  
## 12 12 12 12 12 12  
  
# We can generate a boxplot of each type of the spray  
boxplot(count ~ spray)
```



```
# Run One-Way Anova using aov()
anova <- aov(count ~ spray, data = InsectSprays)
summary(anova)

##           Df Sum Sq Mean Sq F value Pr(>F)
## spray      5   2669   533.8    34.7 <2e-16 ***
## Residuals  66   1015    15.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Manually Calculate the F statistic
N <- nrow(InsectSprays) # number of Observations
n <- 12                  # number of observations per group
k <- length(unique(InsectSprays$spray)) # Number of groups
grand_mean <- mean(InsectSprays$count) # Grand Mean
group_mean <- tapply(count, spray, mean) # Mean of each group

SST<- sum(n*(group_mean - grand_mean)^2)
MST <- SST/(k-1)
```

```

SSE <- sum((InsectSprays$count- rep(group_mean, each = n))^2)
MSE <- SSE/(N-k)

F_statistic <- MST/MSE
F_statistic

## [1] 34.70228

p_value <- pf(F_statistic, df1=(k-1), df2=(N-k), lower.tail = FALSE)
p_value

## [1] 3.182584e-17

#####
## Two-Way ANOVA Example ##
#####

install.packages("mosaicData")

## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)
## Warning in install.packages("mosaicData"): 'lib = "/usr/local/lib/R/site-library"' is
not writable
## Error in install.packages("mosaicData"): unable to install packages

library(mosaicData)

## Error in library(mosaicData): there is no package called 'mosaicData'

data(SaratogaHouses)

## Warning in data(SaratogaHouses): data set 'SaratogaHouses' not found

# The response variable in this example is the price.
# We want to test whether heating type and sewer type
# will affect the housing price in Saratoga.
head(SaratogaHouses)

## Error in head(SaratogaHouses): object 'SaratogaHouses' not found

# We can check the sample sizes in each combination of the two variables
table(SaratogaHouses$heating, SaratogaHouses$sewer)

## Error in table(SaratogaHouses$heating, SaratogaHouses$sewer): object 'SaratogaHouses'
not found

# If we assume the two variables are independent, we can use the additive model.
anova2 <- aov(price ~ heating + sewer, data = SaratogaHouses)

## Error in terms.formula(formula, "Error", data = data): object 'SaratogaHouses' not found
summary(anova2)

## Error in summary(anova2): object 'anova2' not found

# If we also want to test whether the two variables interact to affect the housing price,
# we can use an interaction term.
anova3 <- aov(price ~ heating * sewer, data = SaratogaHouses)

## Error in terms.formula(formula, "Error", data = data): object 'SaratogaHouses' not found
summary(anova3)

## Error in summary(anova3): object 'anova3' not found

# The command below is equivalent to anova3
anova4 <- aov(price ~ heating + sewer + heating:sewer, data = SaratogaHouses)

## Error in terms.formula(formula, "Error", data = data): object 'SaratogaHouses' not found
summary(anova4)

## Error in summary(anova4): object 'anova4' not found

```

## 20 Teste A/B: Introdução

### 20.1 Quarto vídeo da segunda semana

Olá, vamos ver, neste conjunto de vídeos, vamos conversar um pouco pouco sobre o teste A / B. É comumente usado técnica de marketing, e tornou-se ainda mais com a linha recursos de marketing. Então, por que não apenas mergulhe direto e comece. Então primeiro, deixe-me fale um pouco sobre o que é o teste A / B. Esse é realmente apenas o processo de comparar dois grupos. Você pode ter ouvido um teste grupo e um grupo controle. Isso é frequentemente usado na ciência ou na literatura científica. Um grupo de controle ser seleção aleatória, amostra aleatória onde você não faz nada e depois em outro metade do grupo, o grupo de teste, você faria realizar algum tratamento.

Então, um grupo não receberá a droga e um grupo vai pegar a droga, e esse será o Grupo de tratamento. Frequentemente no marketing mundo da análise, isso é referido como teste A / B, mais do que testar grupo de controle. É realmente uma questão de vocabulário. Dois grupos geralmente significa duas versões de algum ativo do mercado. Então, o que eu quero dizer com isso? Um ativo de mercado é algo você deseja pesquisar como, se for uma página da web, um título, um layout de um design, talvez você esteja testando para obter um novo marketing slogan ou frase, um recurso como a cor de um botão de site, uma fonte, todos esses as coisas são comparadas.

Então você pode executar a página da web, sua página original que você está usando há algum tempo, e então você pode ajustar para testar alguma coisa. A única diferença entre os dois grupos é que o grupo de teste tem o tratamento e o grupo controle não. Isso geralmente é referido para como teste A / B. Então, aqui está um exemplo que eu aludi anteriormente marketing digital. Este é um exemplo de um experimento de página da web. Então diga que você tem nessas duas páginas da web e você está tentando decida se deve ou não usar o botão verde que você pode ver abaixo aqui, lá vamos nós, um botão verde ou um botão azul, e você está tentando para entender qual deles atrai mais cliques. Qual deles realmente recebe o usuário envolvido.

Em algum nível, isso pode parecer como uma mudança superficial, mas estes são os tipos de coisas que você irá querer considerar se você está trabalhando em um campanha de marketing digital. Então, o procedimento geral para testes A / B, você tem algum design de produto. Então, se você está projetando algum widget, você pode adicionar um recurso ou não adicionar um recurso. Realizando teste A / B, faça a análise e veja se o que você esperava que acontecesse, ou não acontece. Então, se der certo, você pode iniciar o produto. Nos testes A / B, você projete seu experimento, você conduz sua análise de dados, desenvolve suas hipóteses e se suas alterações suportam sua hipótese, você pode entrar no produto revisão e lançamento do produto.

Este slide basicamente diz o que eu acabei de dizer. Você quer selecionar o tratamento de teste. Nos testes A / B, você está testando apenas um recurso por vez. Você quer ter certeza de que tenha um objetivo claro em mente. Qual é o seu objetivo? O que você está tentando fazer? Então associado a isso, qual é a sua métrica? Como você vai medir a mudança? Você quer que a gente crie um teste grupo e um grupo de controle? De um modo geral, você deseja dividir seus grupos de maneira uniforme e aleatória e deseja determinar o tamanho da amostra para que você tem dados suficientes e depois coleta o dados e você realiza sua análise, incluindo testando hipóteses. Então, primeiro passo, vamos perfurar abaixo em detalhes. Selecione uma variável independente que você deseja testar. No exemplo verde azul, a cor do botão "Saiba mais" é a única variável.

Você pode testar várias variáveis, mas deseja fazer este de cada vez. Há muitas empresas que apenas exploram isso especialmente em o erro de marketing digital. Eles testarão um recurso para um conjunto de clientes, outro recurso para outro um conjunto de clientes compara, e eles iteram através esse processo continuamente. O segundo passo é identificar a meta e a métrica. Parece fácil, mas, na verdade, é provavelmente um dos aspectos mais desafiadores de um bom programa de análise e você deseja identificar a meta, além de como está indo para medir esse teste. As vezes no azul exemplo de botão verde, é apenas o número de cliques. Portanto, a métrica é muito óbvia, mas às vezes não é tão óbvio. Se você alterar sua métrica, poderá configurar um procedimento de teste diferente. Então, por si só não é muito difícil de entender, mas você quer manter isso em mente e ter bons registros quando você compara através dos testes. Então você tem testes A / B mais antigos que você está olhando.

Você está olhando para os dados deles e, de alguma forma, você mude de objetivo ou talvez sua métrica e você não poderá comparar os testes com a mesma facilidade. Então o próximo passo no processo é criar um grupo de teste e um grupo de controle.

Idealmente, em um verdadeiro experimento, queremos fazer isso usando tarefa aleatória. Estamos avaliando o proposição se x então y. Então, no verde azul exemplo de botão, se o botão estiver verde, os cliques aumentarão. Se não x, então não y. Se o botão não estiver verde, os cliques não vai aumentar. Isso prova a causa da eficácia. De um modo geral, você queremos dois grupos equivalentes e atribuímos aleatoriamente

sujeitos ao controle grupo ou grupo de teste. De um modo geral, por convenção, o grupo de controle é A e o grupo de tratamento é B.

A ideia fundamental aqui está a randomização, que aleatoriamente aloca os assuntos nos dois grupos. Este é o mais ideal forma de um experimento, e é chamado de verdadeiro experimento. Principalmente porque controla muitos outros fatores estranhos. Outra coisa que você está vai querer considerar é o tamanho da amostra. Quantos dados você tem? vai colecionar? O teste A / B usa essa noção de significância estatística para determinar se um tratamento funciona ou não ou não funciona. Portanto, o tamanho da amostra é uma variável-chave. Basicamente, quanto maior o tamanho da amostra, maior sua confiança O nível está em seu experimento. Então, antes do teste, você deve pensar em o tamanho ideal da amostra, bem como a estatística nível de significância. De um modo geral, nós use um nível de significância de cinco por cento e uma confiança intervalo de 95 por cento.

As coisas a considerar aqui é que, com o tamanho da amostra, lembre-se de que você está em um ambiente corporativo e a coleta de dados não é necessariamente gratuita. Então, se você está fazendo em um experimento de campo, você está testando mercados em um ambiente de varejo. Se você deseja muitos dados, isso pode custar muito de dinheiro para coletar.

Então, tentando encontrar o máximo quantidade de poder estatístico e usando o mínimo tamanho da amostra para obter esse poder é algo que você vai quer considerar. Então, um último ponto sobre significância estatística nível de cinco por cento, são convenções científicas.

Mas se você vir algo talvez 5,01 por cento, ou mesmo seis por cento, vale a pena olhar para isso um pouco mais para ver se isso é algo que é vale a pena investigar. Então, como você consegue o tamanho da amostra? Está bem aqui. As fórmulas, vamos falar através desta fórmula bem rápido.

Então aqui temos Alpha. Vamos começar por aí, e é isso seu nível de significância. Seu 0,05, por exemplo, e esta é uma distribuição normal e vai ser um teste bicaudal, então é isso por que eu divido em dois. Então, se você tem uma curva normal aqui, normal normal. Z é distribuído, 0,1 normal normal. Qual é esse valor aqui? Então esse é z, e então você soma esse termo como uma estimativa da soma de sua variação. Se você pode realmente calcular a variação que é ideal.

Às vezes as pessoas usam isso estimativa de regra de ouro, que é o intervalo. O valor máximo menos o valor mínimo, dividido por quatro. Se você não pode melhorar nada, é esse o termo.

Então você divide isso pelo tamanho do seu efeito, e o tamanho do seu efeito é como distantes seus testes. Então, se você pensa que é vai fazer o teste é mais 100 cliques com o botão verde, o tamanho do seu efeito será seja esse valor lá. Então se você apenas conecte a fórmula, você obterá o valor n aqui ou o número do seu tamanho da amostra. Coletar e analisar dados é o quarto passo na sua análise. Verifique se você tem tempo suficiente para obter um tamanho de amostra sólido. Evite conflitos entre seus teste e teste, projeto ou empresa de outra equipe planos de negócios.

Então pode haver outros esforços de análise de dados em andamento na sua empresa e você não deseja colidir com seus dados, se você estiver fazendo experimentos de campo ou ensaios, você deseja para garantir que você esteja livre de influências externas no seu experimento.

Você também pode ter um abrangente plano de coleta de dados, um plano de backup para garantir reprodutibilidade. Reprodutibilidade é algo isso é realmente importante. Você quer poder para garantir que, se alguém voltar para você em um mês e diz: "Você pode executar novamente essa análise?"

Você pode fazer isso e deve tem os mesmos resultados. Eles podem solicitar que você estenda sua análise estatística usando diferentes testes e técnicas diferentes. Então é isso que o seu forma ideal é. Então agora que eu falei sobre o processo geral geral de executar um teste A / B, no próximo vídeo eu sou vai falar sobre teste estatístico específico sob diferentes cenários.

## 21 Teste A/B: Tipos de teste

### 21.1 Sexto vídeo da segunda semana

Neste vídeo, eu vou para falar sobre o teste A/B especificamente em termos dos diferentes tipos de testes envolvidos.

Isso é realmente uma função do tipo de dados você coletou. Então, vamos nos aprofundar. Em todos esses testes, há uma hipótese nula e uma hipótese alternativa. A hipótese nula afirma que não há diferença nos resultados.

Basicamente, em outras palavras, o grupo A e o grupo B não tem diferença. Embora a hipótese alternativa forneça evidências de que há uma diferença nos resultados. Então, vamos olhar para o caso de uma amostra. Então, isso não é realmente assistência de teste A/B. Você tem um grupo do conjunto de dados e você deseja saber se a média desse grupo é diferente de zero.

Essa é geralmente a coisa que você está testando. O comando em R é `t.tests`.

Veremos um pouco mais exemplos no próximo vídeo. Mas a hipótese nula afirma que não há diferença de zero, e a hipótese alternativa diz que a média do grupo é diferente de zero.

Neste slide, eu tenho a fórmula para o teste t de uma amostra. Existe o teste t. Você leva sua amostra média. Então você tem seu dados, você calcula a média da amostra, que é  $\bar{x}$ , e subtrai a média da população que você está testando contra.

Nesse caso, já que é testes de uma amostra, geralmente testamos para ver se a média é diferente de zero. Então você pode simplesmente colocar zero aqui para  $\mu$ , e então você vai dividir por  $s$  sobre  $n$ .  $S$  é a estimativa do seu desvio padrão da sua população, e  $n$  representa seu tamanho da amostra.

As premissas do teste t são  $\bar{x}$  e seguem uma distribuição normal com uma média de  $\mu$ . Portanto, a barra  $x$  é igual a  $\mu$ . No entanto, o padrão desvio como este Sigma sobre o raiz quadrada de  $n$ . Então a variância da amostra segue uma distribuição de qui-quadrado com graus de liberdade  $n-1$ .

Eu quero discutir, é se você tiver variações iguais e tamanhos de amostra iguais dos seus dois grupos. Portanto, as condições, tamanhos de amostra dos dois grupos são iguais e denotamos que matematicamente por  $n_A$  é igual a  $n_B$ . Então o número de elementos em cada grupo é o mesmo. As duas populações tem a mesma variação. Então, se eles são atraídos por na mesma população, você poderá fazer essa suposição.

Então, aqui estão as fórmulas para tamanhos iguais de amostra, variâncias iguais e aqui está um teste contra  $\mu$ . Você está testando  $\bar{x}_A$ , uma média da amostra do grupo A e menos a amostra média do grupo B, e vamos apenas olhar para o denominador por um segundo. Se são iguais, é vai ser igual a zero. Se eles são iguais a zero, isso significa que as duas amostras meios são os mesmos.

A hipótese alternativa diz que existem diferentes. Então, se  $\bar{x}_B$  é maior ou menor que A, então eles são considerados ser diferente. É escalado por este padrão valor do desvio aqui em baixo. No teste de uma cauda você teve  $s$  sobre a raiz quadrada de  $n$ . Mas aqui você tem isso desvio padrão combinado, porque a variação é igual. Então, o jeito que funciona é pegar  $s_A$  quadrado mais  $s_B$  ao quadrado, então a variação do grupo A, a variação do grupo B, você as adiciona e pega as média e divide por 2, e então você faz o raiz quadrada disso. É assim que você obtém o desvio padrão combinado. Então, você toma a diferença entre as duas médias e você escala por esse fator do desvio padrão combinado. Os graus de liberdade são agora  $2n$  ou o número total de elementos em ambos ou assuntos em ambos os grupos, menos 2. Com o qual você costumava calcular as médias da amostra.

Então aqui está o código R que eu mostrarei a você em um vídeo posterior. Aqui está outro tipo de teste A/B, um teste t chamado teste t de Welch.

Aqui, relaxamos a suposição de que a amostra tamanho precisa ser o mesmo. Então os tamanhos das amostras dos dois grupos podem ser iguais ou desiguais, e as duas populações têm distribuições normais, mas podem não ter a mesma variação. Então você pode estar se retirando duas populações diferentes. Novamente, você está testando isso, o numerador é a diferença entre as duas médias do grupo. Então, se eles são zero, são iguais, se não são zero, são vai ser diferente. Mas então você está escalando por esse valor agregado, que é a variação do grupo A, dividido por  $n$ , o número de elementos em A, mais a variação do grupo B, dividido pelo número de elementos no Grupo B e você os adiciona e você pegou a raiz quadrada.

Aqui vamos nós. Direita há. É assim que você faz o teste t de Welch. Os graus de liberdade são como fórmula complicada. Eu não vou entrar com muitos detalhes, em parte porque R calcule isso para você.

Mas se você estiver interessado, é assim que você calcularia isso. Neste último tipo de teste t que eu quero discutir, é chamado de pareado teste t da amostra.

É aí que o amostras são dependentes. Uma amostra foi testada duas ou duas amostras foram emparelhadas. Então a diferença entre os pares deve ser calculada. Então, o que eu quero dizer por uma amostra emparelhada? A maneira mais fácil de pensar sobre isso, diga que você tem um conjunto de assuntos e avalia eles em alguma métrica A maneira mais fácil de pensar sobre isso, diga que você tem um conjunto de assuntos e avalia eles em alguma métrica. Você os testa. Dizer um monte de estudantes, você faz um pré-teste. Então você realiza seu experimento. Então eles recebem tratamento e você faz um pós-teste. Então agora você está testando os mesmos alunos que foram testados antes e depois de algum tratamento, e você quer saber se o tratamento funcionou.

Então é aí que o emparelhamento entra. São os alunos A antes alunos A depois, esses são pareados em dados de amostra, o aluno B está emparelhado com aluno B antes e depois etc.

Então aqui você pode ver a média em que  $\bar{x}_D$  é a amostra emparelhada das diferenças, e aqui está o desvio padrão, graus de liberdade é  $n-1$ . Aqui, novamente, há outro exemplo para um indivíduo, um pré-teste e um pós-teste de uma campanha publicitária. Eu acho que isso termina as principais formas de estatísticas de teste AB que estão por aí, e agora quero mostrar como implementar esses testes em R.



## 22 Teste A/B: Exemplo no

### 22.1 Sétimo vídeo da segunda semana

Neste vídeo, vou demonstrar como executar o AB Testing no Ambiente RStudio. Então, por que não mudo para lá, e aqui está minha pequena programa que vai demonstrar algumas das principais elementos do teste A/B.

Nós vamos instalar isso mosaico de biblioteca que possui alguns conjuntos de dados e o Dados Saratoga Houses. Então está carregado. Vamos dar uma olhada, e esse comando e você pode veja aqui no canto superior direito, existem 1.728 observações de 16 variáveis.

Podemos olhar para a linha superior aqui, aqui está o preço, o tamanho do lote, idade. Claramente, isso é real tipo de informação imobiliária. Número de quartos, lareira [inaudível] à beira-mar e et cetera, et cetera.

Então, vamos focar nessa primeira coluna, preço. Podemos dar uma média, 211.966,7. Então essa é a média preço das casas em, acredito que este é Saratoga County, em Nova York. Agora podemos executar um teste t. Agora, se eu não colocar quaisquer valores aqui, para ver se a média da nossa amostra é ou não diferente de zero. Nós apenas olhamos a média, 211.000.


Para que possamos executar esse teste t e por que não olhamos por um segundo. Há uma amostra teste t e os resultados. Isso nos fornece os dados que usado para calcular o teste t. Então você vai quer olhar para isso. Aqui, esta é a chave número que você deseja observar, os valores-p e ele tem menos do que o valor crítico valor de 0,05, neste caso, é muito, muito pequeno.

Portanto, a alternativa da hipótese, a verdadeira média é não é igual a zero. Lembre-se de que a hipótese nula é que é igual a zero e aqui está a nossa confiança intervalo da média. Então aqui em cima, nós calcule a média. É 211.966. De acordo com isso intervalo de confiança, deve cair entre 207322 e 216611, e esse é o nosso 95 por cento intervalo de confiança. Em seguida, é meio óbvio, você não precisaria de um teste estatístico para perguntar se os preços da habitação são ou não diferente de zero, especialmente considerando os dados. Mas é diferente de algum valor de seu interesse, digamos 200.000. Se você estiver testando habitação preços, por exemplo, novamente zero, você não esperaria que o média para ser zero. Mas você pode querer pensar, é a média de 200.000.

Talvez eu tenha ficado estranho amostra dos meus dados. Então você pode executar isso teste, lá vai você, e você tem um valor de P de 4,8 vezes 10 para menos sete. Então isso é definitivamente inferior a 0,05 e a alternativa hipótese é a verdadeira média, não é igual a e isso é notação científica, mas isso é 200.000. Então, novamente, há o intervalo de confiança e também lhe dá a média da sua amostra. Então, para finalizar esse teste, temos um valor P que é menor que 0,05.

Podemos rejeitar o nulo hipótese de que os valores, desculpe, a amostra média é igual a 200.000. Se você quiser brincar com isso, você poderia. Apenas por diversão, vamos colocar 211.000. Aqui temos um valor de P de 0,68 e não podemos rejeitar a hipótese nula. A verdadeira média é a hipótese alternativa é que a média da amostra seja diferente de 211.000. Mas neste caso, não podemos rejeitar o nulo. Ok, deixe-me mudar isso de volta para que eu não atrapalhe esse arquivo. Aqui está uma amostra de duas exemplo de teste t.

Lembre-se do Saratoga dados de habitação, há o centro ar condicionado, variável sim ou não. Lá está na última coluna.

Então, algumas casas têm ar condicionado central, algumas casas não e nós apenas vamos quebrar em dois grupos. Essa notação aqui, o colchete, isso obtém os índices onde o ar central é igual a sim e este fornece os índices em que o ar central é igual a não, então o divide em dois grupos. Esta é uma convenção , nossa técnica de programação, que eu encorajo você a estudar e olhar por um tempo. Então, basicamente, você tem uma coluna de preços e outra coluna, ar central, Sim, Não, e nós vamos dividir em dois grupos.

Outra maneira de fazer isso é realmente criar duas colunas vetores e uma tabela. Uma coluna Sim, uma coluna Não e então poderíamos apenas compare esses dois grupos. Então, vamos criar estes duas variáveis xey representando os dois grupos. A próxima coisa que você é vai querer saber é, eles têm o mesmo tamanho de amostra?

Para que possamos dar uma olhada no comprimento de x é 635 e o comprimento de y é 1093. Então, claramente, eles não têm o mesmo tamanho de amostra. Podemos querer olhe para um gráfico de caixa. Há um gráfico de caixa lá e lembre-se de um gráfico de caixa, que a linha superior e inferior ou o intervalos interquartis e a linha no meio é a mediana e mostra como os dados são distribuídos.

Então o último ponto Eu devo mencionar sobre o box-plot é, você acha que essas médias são iguais ou não são iguais e é realmente difícil de dizer. Mas podemos executar o teste t. Nesse caso, é não é um teste t emparelhado. São apenas duas amostras assumindo variação desigual. Vamos fazer o teste. Aqui temos um P valor de 2,2 vezes 10 ao menos 16 que é bem menor que 0,05. Para que possamos rejeitar a hipótese

nula de que os meios são os mesmos, e aqui está a intervalo de confiança, 57881,43 a 77883. Aqui estão as duas amostras significa 254.000 e 187.000. E daí

```

1 install.packages("mosaicData")
2 library(mosaicData)
3 data(SaratogaHouses)
4 SaratogaHouses
5
6 #Exemplo de teste t para uma amostra
7
8 mean(SaratogaHouses$price)
9 t.test(SaratogaHouses$price)
10
11 #Rode o teste t para testar se a media de precos das casas em Saratoga eh de $200000
12
13 t.test(SaratogaHouses$price, mu=200000)
14
15 #O valor-p eh de 4.8 o que eh menor que 0.05 o que nos leva a concluir que a media
    verdadeira de precos das casas em Saratoga nao eh $200000.
16
17 #Mudamos o valor de mu para 211000
18
19 t.test(SaratogaHouses$price, mu=211000)
20
21 #Veja os resultados acima e entenda que nao podemos rejeitar a hipotese nula como no caso
    de mu = 200000
22
23 x <- SaratogaHouses$price[SaratogaHouses$centralAir=="Yes"]
24 y <- SaratogaHouses$price[SaratogaHouses$centralAir=="No"]
25
26 #Os dois grupos nao possuem o mesmo tamanho de amostra
27
28 lenght(x)
29 lenght(y)
30
31 #Gera o boxplot dos pre os das casas dos dois grupos
32
33 boxplot(x,y)
34
35 #Dado um teste t de duas amostras nao pareadas
36 #se tem uma diferenca entre os precos das casas dos dois grupos
37 #Perceba que o test.t(x,y) assume vari ncias desiguais por default
38
39 t.test(x,y, paired=F)
40
41 #O valor-p eh 2.2e que eh menor do que 0.05
42 #Entao, podemos rejeitar a hipotese nula abaixo do nivel de significancia de 5% e
    concluimos que ha diferen as nos precos das casas desses dois grupos.

```

Listing 9: Teste A/B no 

```

install.packages("mosaicData")

## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)
## Warning in install.packages("mosaicData"): 'lib = "/usr/local/lib/R/site-library"' is
not writable
## Error in install.packages("mosaicData"): unable to install packages

library(mosaicData)

## Error in library(mosaicData): there is no package called 'mosaicData'

data(SaratogaHouses)

## Warning in data(SaratogaHouses): data set 'SaratogaHouses' not found

SaratogaHouses

## Error in eval(expr, envir, enclos): object 'SaratogaHouses' not found

```

```

#Exemplo de teste t para uma amostra

mean(SaratogaHouses$price)

## Error in mean(SaratogaHouses$price): object 'SaratogaHouses' not found

t.test(SaratogaHouses$price)

## Error in t.test(SaratogaHouses$price): object 'SaratogaHouses' not found

#Rode o teste t para testar se a media de precos das casas em Saratoga eh £200000

t.test(SaratogaHouses$price, mu=200000)

## Error in t.test(SaratogaHouses$price, mu = 2e+05): object 'SaratogaHouses' not found

#O valor-p eh de 4.8 o que eh menor que 0.05 o que nos leva a
#concluir que a media verdadeira de precos das casas
#em Saratoga nao eh £200000.

#Mudamos o valor de mu para 211000

t.test(SaratogaHouses$price, mu=211000)

## Error in t.test(SaratogaHouses$price, mu = 211000): object 'SaratogaHouses' not found

#Veja os resultados acima e entenda que
#nao podemos rejeitar a hipótese nula
#como no caso de mu = 200000

x <- SaratogaHouses$price[SaratogaHouses$centralAir=="Yes"]

## Error in eval(expr, envir, enclos): object 'SaratogaHouses' not found

y <- SaratogaHouses$price[SaratogaHouses$centralAir=="No"]

## Error in eval(expr, envir, enclos): object 'SaratogaHouses' not found

#Os dois grupos nao possuem o mesmo tamanho de amostra

length(x)

## Error in length(x): could not find function "length"

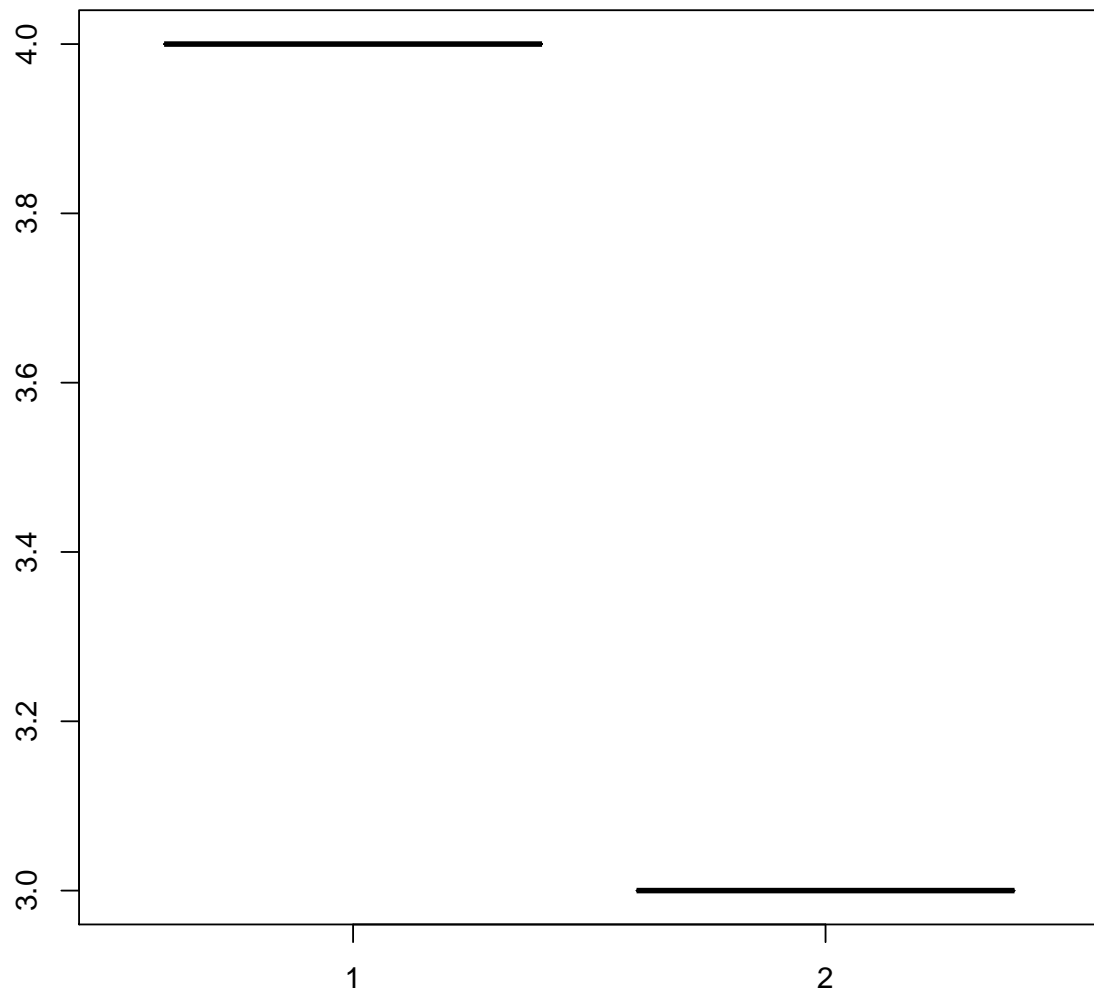
length(y)

## Error in length(y): could not find function "length"

#Gera o boxplot dos precos das casas dos dois grupos

boxplot(x,y)

```



```
#Dado um teste t de duas amostras não pareadas  
#se tem uma diferenca entre os preços das casas dos dois grupos  
#Perceba que o test.t(x,y) assume variancias desiguais por default
```

```
t.test(x,y, paired=F)
```

```
## Error in t.test.default(x, y, paired = F): not enough 'x' observations
```

```
#O valor-p eh 2.2e que eh menor do que 0.05  
#Entao, podemos rejeitar a hipótese nula  
#abaixo do nivel de significancia de 5%  
#e concluimos que ha diferencas nos preços  
#das casas desses dois grupos.
```

## 23 Introdução a ANOVA

### 23.1 Oitavo vídeo da segunda semana

No próximo conjunto de vídeos, vou falar sobre pouco sobre ANOVA, ou análise de variância. Então, vamos nos aprofundar. Então, o que é ANOVA ou análise de variação? É uma suíte ou um métodos estatísticos para determinar se existem diferenças nas médias entre duas ou mais categorias grupos ou dois tratamentos.

No conjunto anterior de vídeos com teste AB, analisamos apenas um grupo ou talvez dois grupos e foi isso. Aqui, podemos olhar em vários grupos. Algumas aplicações típicas eu listei é, um sorvete tem vendas diferentes nas quatro estações do ano.

Então, os grupos de estações, então eu tenho sorvete vendas na primavera, o sorvete é vendido na verão, inverno e outono. Você está vendendo a mesma quantia em cada uma dessas quatro estações ou variando de montantes de diferença? É baseado em uma comparação das variações e não em comparação dos meios, então foi assim que você conseguiu nome Análise de variância.

Então, primeiro, vamos olhar a estrutura dos dados e isso nos ajuda a entender o que está acontecendo.

Então aqui está uma completamente design aleatório, cada  $y$  é o seu ponto de dados então poderia ser vendas. Você tem tratamentos diferentes, no exemplo de sorvete são as quatro estações do ano.

Então, se você observar, quais são as vendas médias para tratamento um ou estação um, estação 2, 3, 4? Aqui, o nulo hipótese é que, os meios para todas as grupos são iguais. No exemplo do sorvete, a maneira de pensar é: vendemos a mesma quantidade de sorvete em cada das quatro estações do ano? A alternativa hipótese é que, pelo menos dois dos grupos são diferentes.

Isso não significa que tudo os grupos são diferentes, significa apenas pelo menos dois dos grupos são diferentes. Então você pode ter três grupos, e A e B são iguais, e B e C são diferentes, então lá está. A outra coisa a considere é a notação. Então, vamos nos concentrar por um momento. Aqui está  $\mu_i$ , e eu sou os indicadores do grupo, então o grupo 1, 2, 3, seja o que for, e  $t$  é o número total de grupos, então  $i$  varia de um a  $t$ , e  $n_i$  é o número de assuntos dentro de cada grupo. Então você tem grupos, quantos grupos você tem e quantos elementos estão em cada grupo? Portanto, esses são os elementos-chave que você deve ter em mente. Então, quais são as suposições do modelo ANOVA? Primeiro é a normalidade, que as amostras de cada um dos grupos categóricos são coletadas de uma população que é distribuído normal. No normal, chamamos isso a curva em forma de sino. Independência. que significa que cada amostra é coletada independentemente das outras amostras, portanto a existência de um elemento não afeta a existência de outro, eles são apenas independentes de cada um.

Também assumimos que as variações são as mesmo entre os grupos. Portanto, esse é um componente-chave e o dependente variável é contínua. Portanto, a variável alvo no exemplo de sorvete de que falei é a vendas do creme, e isso é dividido por estações. Então aqui está a matemática representação do modelo ANOVA.

$Y_{ij}$  é o cada um dos assuntos. Então, o assunto reside em algum grupo  $i$ , e aqui está o  $j^o$  observação no grupo  $i$ . Então, pessoa um no grupo um, pessoa dois no grupo um, pessoa três no grupo um, etc., e então você vai para o grupo dois e você tem uma pessoa um no grupo dois, pessoa dois no grupo dois, etc.  $\mu$  é o que se sabe como a grande média. Então, se você pegar todos os seus pontos de dados e basta calcular a média, esse é o seu grande quer dizer aqui, e essa é a sua média de todos os pontos de dados.  $\alpha_i$  indica o efeito do grupo  $i$ .

Portanto, cada um desses grupos tem um tratamento diferente e  $\alpha_i$  é o efeito de esse tratamento e, em seguida,  $e_{ij}$  é o aleatório erro que ocorre para a  $j$ -ésima observação no grupo  $i$ .

Também assumimos que  $e_{ij}$  tem uma distribuição normal de uma média de zero, e algumas comuns e variação constante padrão ou desvio padrão, e os termos de erro são independentes um do outro.  $N_i$  indica o número total de observações no grupo  $i$ . Então este é o geral forma do modelo. Portanto, o modelo ANOVA implica que, o  $j$ -ésimo sujeito ou a  $j$ -ésima resposta no grupo  $i$  é normalmente distribuído com uma média do grande significa  $\mu$  mais  $\alpha_i$ , que é o impacto desse tratamento nesse grupo, mais alguns variação constante para os termos do erro.

Então, a maneira de pensar sobre isso, você tem a grande média que é um número e então você está muito longe da média geral, de alguma forma, de algum grupo de fato eu e é assim que você chega à média do grupo. Então, para chegar à média do grupo, primeiro você encontra o média geral e depois mais ou menos algo para chegar à média do grupo. Então, vamos decompor isso em termos da soma dos quadrados, e espero que isso comece a esclarecer as coisas. O principal a lembrar é que  $j$  é o assunto e eu sou o grupo. Então deixe-me escrever isso aqui em baixo, então temos isso à mão enquanto avançamos através dessas fórmulas. Eu sou grupo Lá vamos nós, o grupo  $iej$  é a resposta.

$I$  é um grupo, G-R-O-U-P,  $ej$  é igual ao número do sujeito ou a resposta. Portanto, soma total de quadrados, lembre-se de que estes são apenas distâncias. Portanto, soma total de quadrados, lembre-se de que

estes são apenas distâncias agora. Você está tomando a soma de a distância de  $y_{ij}$ , então esse é um ponto de dados, para a grande média. Então, onde quer que você estão no seu gráfico, você encontrará o grande média e essa é a distância que você está interessado. Essa distância pode ser quebrada da seguinte maneira. Vamos olhar para essa pessoa aqui,  $y_{ij}$ , é a mesma coisa,  $y_{ij}$ . Então esse é um ponto de dados e você vê a notação  $y_{ij}$ .

Isso significa que a média do grupo  $i$ . assim é isso que é isso. Então, primeiro, este é a distância entre o ponto de dados e o grupo significa e então aqui, você tem o grande estimativa média aqui,  $y_{..}$  e então aqui está o grupo significa. Então, para ir de  $y_{ij}$  para  $y_{..}$  você faz a diferença entre a média geral e a média do grupo e a média do grupo e a ponto de dados individual. Então os dados apontam para agrupar significa para a grande média. Este  $n_i$  representa o fato de que em um ano, você só tem média de grupo menos a grande média, e você tem que fazer isso  $n$  vezes para cada um dos pontos de dados apenas para equilibrar as equações. Quando  $n_i$  é o mesmo para em todos os grupos de tratamento, você pode fatorar o  $n$ , o que é feito aqui, e você deixado com esta equação. Então, para ir de  $y_{ij}$ , o ponto de dados individual, a média geral, você vá de longe. Essa distância é igual à distância entre a média geral e a média do grupo e o grupo significa para o ponto de dados, que é esse pedaço aqui.

Então, em termos de soma dos quadrados, sua soma total de quadrados, essa é a sua distância do seu ponto de dados para sua média principal é composto de duas partes. A primeira parte é a "entre a soma dos quadrados do tratamento ou SST", soma tratamento de quadrados. Que mede a variabilidade devido às diferenças nos tratamentos, e esse é o seu grupo, você pode ver que é o distância entre a média do seu grupo ou o seu tratamento média e sua grande média. Então cada grupo tem seu próprio grupo significa, e a que distância está da média geral e, em seguida, você multiplica por  $n$ , porque não há  $n$  assuntos que você precisa explicar. Então aqui o erro da soma dos quadrados mede a variabilidade que não é explicada pela diferença de o tratamento significa. Portanto, o tratamento explica a diferença do grande significado para o tratamento, e essas são as diferenças no nível individual. Então aqui está o indivíduo assunto,  $y_{ij}$  e a diferença ou a distância do grupo significa, e essa é a sua soma erros de quadrados. Então, é a soma de quadrados significativos? Então a estatística do teste aqui está o quadrado médio do tratamentos entre grupos.

Então é isso, você pega sua soma de quadrados tratamento e você divide por basicamente o número de grupos menos um, porque você calculou o grupo significa, então esses são seus graus de liberdade lá, e então seu quadrado médio erro dentro do grupo, essa é a média da soma dos erros ao quadrado divididos pelo número de elementos dentro de cada grupo menos  $k$  pelos graus de liberdade.

Sob o nulo hipótese para o teste, você está assumindo que tudo o que significa na população são os mesmos. Portanto, o tratamento para A, B, C, D é o mesmo e, se dois deles são diferentes, você sabe que algo aconteceu com um dos tratamentos. Os graus de liberdade é uma distribuição F, aí está. O numerador tem um graus de liberdade  $k$  menos 1 e o denominador tem graus de liberdade  $N$  menos  $k$ ) Você pode procurar, eu apenas fornecerei você as estatísticas. A principal coisa a observar é que não é uma distribuição simétrica como a distribuição normal. É uma distribuição que reside no final positivo de a escala de linha, digamos, varia de zero a infinito positivo, e não é completamente simétrico, tem algum peso de um lado. Mas como um normal interpretação da distribuição, você deseja encontrar o áreas de cinco por cento. Então o valor sombreado lá é um valor  $p$  de 0,05.

Então, se você encontrar uma estatística  $F$ , qual é a razão dos quadrados médios para os tratamentos e os quadrado médio para os erros, então você encontra suas estatísticas lá fora, isso significa que mais da variação total é explicado pelo tratamento, tratamento quadrado médio, do que pelo erro quadrático médio. Então é isso que faz esse número é grande. Se você quer pensar de outra maneira, e veja essa razão, se o erro médio quadrático for realmente grande e o quadrado médio o tratamento é pequeno, então essa proporção total se torna pequena e vai abaixo da linha numérica e você estará longe dessa área de valor- $p$  ou da região de rejeição de a hipótese nula.

Então nossa estatística é a média quadrado para o tratamento, a média ao quadrado para os erros, seus graus de liberdade para sua estatística  $F$ . Se o seu valor  $P$  for para o direito da sua estatística, está lá fora à direita aqui, então você pode rejeitar, e a rejeição é que todos os meios de os grupos são iguais e  $\alpha$  é o seu nível de significância e geralmente usamos um número como 0,05.

## 24 One-Way ANOVA – Insect Spray Example

### 24.1 Nono vídeo da segunda semana

Neste exemplo, quero ver um exemplo simples de ANOVA. É uma ANOVA de mão única e o conjunto de dados precisa ser pulverizadores de insetos.

Portanto, esse conjunto de dados, Insect Sprays, é incorporado ao R e existem algumas colunas neste conjunto de dados que nós vamos olhar.

Eu vou apresentar os slides primeiro, mostram como calcular a ANOVA unidirecional manualmente e, em seguida, veremos no exemplo R, e você verá que é muito fácil de implementar em R. Portanto, os conjuntos de dados estão em R.

Esse conjunto de dados possui duas colunas, uma é a contagem de variáveis do número de insetos que eles contaram e há seis tipos de pesticidas A, B, C, D, E e F. Em seguida, dentro de cada tipo de spray inseticida, existem 12 números. Então você pode imaginar um gigante plotagem que eles dividiram em seis subparcelas e depois em cada subtrama eles mediram ou pulverizaram cada um os diferentes produtos químicos e, quando fizeram medições, fizeram 12 medições de A, 12 medições de B, 12 medições de C etc.

Queremos testar se houve ou não alguma diferença nesses lugares. Então, abaixo estão alguns dos exemplos de seus dados. Então, se olharmos para o spray A tem uma contagem de 10, 7 e 20 e depois há 12 deles.

Spray B; 11, 17, 21 etc. Lembre-se de que o teste ANOVA que estamos testando para verificar se o grupo significa que são os mesmos. Então essa é a nossa hipótese nula. Portanto, a média do grupo de A é igual à A média do grupo de B é igual à média do grupo de. A média do grupo de A é a o mesmo que B é o mesmo que C, D, E e F. Então a hipótese alternativa é que pelo menos dois dos os meios de grupo são diferentes.

Ao menos dois. Então em menos dois são diferentes,  $y_{.i}$  indica a média de inseticida  $i$ , lembre-se que eu varia de A, B, C, D, E, F. Vamos usar um nível de significância de 0,05. Isso é algo você sempre desejará se comunicar quando apresentando um relatório. Nesse caso, a amostra tamanhos para cada grupo é de 12 e que é denotado  $n$  sub A é o mesmo que  $n$  sub B é o mesmo que yada yada e sadaf.

Então todo o grupo tamanhos são iguais, mas meu número total de pontos de dados coletados é 72  $ek$  é o número de grupos, e isso é seis. Portanto, verifique se você tem essa corda na sua cabeça. Eu tenho seis grupos, 12 itens dentro de cada grupo e existem 72 itens completamente. Aqui está a soma de cálculos de quadrados.

Em primeiro lugar, queremos calcular a média geral. Aí está. É isso que você leva para todos os  $i$ 's  $ej$  de cada elemento, você os adiciona e então você divide por  $n$  ou o número total de observações que você tem. Nesse caso, passa a ser 75. Então, dentro de cada grupo, você vai querer calcule a média do grupo. Então observe que o diferença entre esta fórmula aqui e uma dessas fórmulas de grupo significa que estamos segurando o grupo corrigido denotando A e depois passando todos os elementos do grupo A, é isso que notação está me dizendo aqui e então você divide por  $n$ . Nós sabemos disso caso  $n$  é constante.

Se não houvesse constante, você teria que conta para isso. Portanto, temos  $y_{.A}$  ou o A média do grupo A é 14,5, B é 15,33 e você deseja fazer isso para cada um dos grupos. Sua soma total quadrada, lembre-se, é a distância entre a média do grupo  $y_{.i}$  e a média geral denotado ponto  $y$  ponto.  $y_{.i}$  significa para grupo eu e todos os  $j$ 's. É isso que o pequeno ponto significa. Então nós temos 12 em cada amostra. Então, há a fórmula. Então, sim, este é o grupo A, este é o grupo B, este aqui é 2,08 menos 9,5, é C e você faz o resto deles. Então, espero que você possa ver isso aqui. Deixe-me apontar isso com uma caneta. Aqui está 14,5? Veio deste número aqui. Então esse é o significado do grupo menos sua grande média.

Aqui está o próximo grupo significa menos sua média e assim por diante. Você considera a diferença entre a média do grupo e a grande média, você equivale a isso, você adiciona todos juntos, e então você multiplica por  $n$ . Aqui está a soma do erro quadrático e isso vai do ponto de dados real. Então  $y_{.ij}$  é o grupo  $i$ , o  $j$ 'th elemento. Então aqui está a média do grupo. Então aqui o grupo significa que calculei para A, B e C, 14,5, 15,33, e você pode vê-los aqui, 14,5, 14,5 etc.

O primeiro elemento no grupo A é 10, o segundo elemento é 7, e você só precisa calcule esta longa equação. Então queremos agora calcular os quadrados médios para o tratamento e a média quadrados para os termos do erro. Então soma dos quadrados total menos  $k$  menos um, esses são o número de grupos menos um, e então você obtém esse valor. Você está calculando a média da soma dos erros ao quadrado sobre o número total de elementos menos o número de grupos lá e obtém isso 15,38. Então essa é a sua média erro ao quadrado. Então sua estatística F é a relação desses dois valores. Você recebe 34,7 e lembra se o número é à direita, você tem um valor-p menor e precisará conhecer seus graus de liberdade para procurar.

A distribuição F muda de forma com base nesses dois parâmetros DF1 e DF2. Mas neste caso, nós obtive um valor-p significativamente menor que 0,05. Para que possamos rejeitar o hipótese nula, que dois dos seis meios grupo significa que são os mesmos.



Então é isso que tabela pareceria em uma tabela ANOVA padrão. Entre grupo, esse é o seu significa, seu tratamento, sua soma de quadrados, aqui está sua média tratamento quadrado, e quando você vê uma mesa, pode observar isso a soma dos quadrados menos seus graus de liberdade deve ser igual ao quadrado médio, e aqui está sua estatística F, e aqui está sua dentro dos valores do grupo.

Ai está. Então em Nesse caso em particular, obtemos a seguinte tabela. Então deixe-me mostrar como você corre isso em .

## 25 One-Way ANOVA – Insect Spray Example in

### 25.1 Décimo vídeo da segunda semana

OK, então estamos aqui no ambiente R e o InsectSprays é um conjunto de dados pré-carregado em R, para que possamos anexar isso. E isso traz o conjunto de dados das bibliotecas R e coloca-o em nosso ambiente. Mas podemos querer obter um identificador nesse conjunto de dados ou um entendimento desse conjunto de dados.

Dois comandos comuns que você pode usar, um é dimensões. Então tem 72 linhas e 2 colunas, dimensões 72, 2. E lembre-se, sempre são linhas e colunas, linhas e colunas. E então este comando de estrutura, que nos fala um pouco sobre a estrutura do conjunto de dados.

A primeira coluna é denominada contagem, a segunda coluna é denominada spray e esses são os valores reais. Se nós, InsectSprays, podemos dar uma olhada nesse conjunto de dados. E você pode ver, aqui está a linha 1, é do inseticida A e conta 10 insetos nesta amostra, etc, etc, e sabemos que existem 12 para cada uma dessas amostras.

Ok, então vamos voltar ao código, então esta próxima linha de código, aplique, esse é um dos pacotes de aplicar funções em R. Eu sei que não discuti isso nos vídeos tutoriais, mas deixe-me explicar você o que está fazendo. Lembre-se de que os nomes das colunas InsectSprays são contagem e pulverização, e o comando length obtém o comprimento.

Então, nós temos dimensões, mostramos que anteriormente, InsectSprays, Abaixo e você pode ver que é 72 por 2. Eu mostrei o comando de estrutura, Eu deveria apenas executá-lo aqui. Há o comando de estrutura, e tem contagem. E então podemos usar esse comando de comprimento, comprimento (InsectSprays \$ count), e é 72, então esse é o comprimento dessa coluna. Também posso fazer o comprimento do spray, S-P-R-A-Y, que também deve ser 72.

O que este comando tApply está fazendo é contar, qual é o tamanho da contagem se eu separar por cada um dos valores de spray? Então eu vou contar quantos Como existem, quantos tipos de spray B, spray tipo C, etc etc.

Então, quando executo esse comando tApply, é x, ye a função, Eu recebo 12 em cada categoria. Então agora eu sei que lá são 12 elementos em cada grupo. Em seguida, vamos fazer uma pequena visualização, aqui está o nosso gráfico de caixa.

A linha inferior é o mínimo, o máximo, a linha escura no meio é a mediana. E a caixa representa o quartil varia de 25 a 75% e deixe-me expandir isso. Só de olhar para isso gráfico à direita, você pode ver que alguns dos pesticidas teve um efeito melhor do que outros.

Claramente, C, D e E têm contagens mais baixas, o que significa que o inseticida era forte, bom, funcionou. Considerando que estes estão aqui em cima, então eles não foram tão eficazes. Então esse é o gráfico da caixa, e vamos executar uma ANOVA unidirecional. Aqui está o comando, aov, e estamos interessados em contar como nossa variável de resposta, e o tipo de spray é o nosso tratamento, de modo que colocamos isso como uma variável independente, data = InsectSprays.

E aqui está sua técnica R padrão de executar sua função e colocar os resultados na função em algum nome de variável. Aqui, eu chamei de anova, eu poderia chamei de qualquer coisa que eu quisesse. E então eu uso o comando de resumo para realmente ver os resultados. Então eu corro isso, e então agora quero ver os resultados e uso esse comando de resumo. E aí está, deixe-me esclarecer essa coisa de baixo, claro que, para ficar mais fácil de ver execute o resumo novamente. Então você pode ver os graus de liberdade para o spray é cinco, certo, então esse é o número de grupos menos 1.

A soma dos quadrados para o tratamento é 2669, se você dividir 2669 por 5, você deve obter 533. E na parte inferior de a razão da estatística F, é esse número, 1015 menos 66, 66 é 72 menos 6, o número de grupos e, se você dividir 1015 por 66, você deve obter 15.4.

Este é o seu valor real de a estatística T, mas realmente, este é o número chave aqui, que é menor que 0,05. Então sabemos que pelo menos dois do grupo significa, podemos rejeitar a hipótese nula que todos os meios são iguais e pelo menos dois são diferentes. Eu queria mostrar um código sobre como para calcular esses números manualmente.



Certo, na prática, você apenas faz isso e deve saber ler uma tabela ANOVA unidirecional. Mas eu mostrei algumas fórmulas no slide do PowerPoint, pensei em fazer isso manualmente.  $n$  é o número de observações e deve ser 72, então  $n$  é 72. Número de observações por grupo, vimos isso antes, vou codificá-lo como um 12.

$k$  é o tamanho de sprays exclusivos, certo, então eu sei que vai A, B, C, D, E e F. Mas aqui eu recebo os únicos, e então conto o número de únicos. Então, se você apenas executar isso parte do comando, você obtém um vetor parecido com este. E então qual é o tamanho desse vetor, deve ser 6, e você pode ver isso aqui em cima,  $k$  é igual a 6. Aqui está a média geral, e essa é apenas a média de todos os dados nesse conjunto de dados. Então aqui está `InsectSprays $ count`, tomamos a média, e essa é agora a grande média. E agora eu quero entender a média cada grupo, então eu uso esse comando `tApply`.

Aqui está a função `significa`, eu estou usando isso função, e eu vou obter a média da contagem, e eu vou quebrar pelo tipo de spray. Então isso me dá seis grupos me Então isso me dá seis meios de grupo, boom, e lá estão eles. Podemos dar uma olhada neles, `group_mean`, para que o grupo A seja 14,5, 15,33, 2,08, etc. Tudo bem, e a soma dos quadrados para o tratamento é  $n$  vezes o grupo `significa` menos a grande média ao quadrado, lá estão eles. E então, para obter a média ao quadrado do tratamento é apenas a soma do tratamento quadrado menos  $k$  menos 1, Aí está.

E então a soma do quadrado erros é semelhante a isso. A única coisa estranha aqui neste implementação aqui é esta parte, onde repito o grupo `significa` 12 vezes, Para cada um dos elementos. Tenho certeza que há um mais elegante maneira de fazer isso, mas funcionou. E então eu recebo o erro quadrático médio, que é a soma do erro quadrático menos  $n$  menos  $k$ , aí está. A estatística  $F$  é apenas a proporção dos dois, e então eu não vou passar esse comando com muitos detalhes. Mas isso basicamente parece o valor da estatística  $F$ , procura o valor  $p$  com base nessa estatística  $F$ . Para aqueles de vocês que podem ter visto algo assim antes em uma aula de estatística mais tradicional, é aqui que você calcularia a estatística  $F$  e depois vá para o final do livro e procure o valor- $p$  em o índice nessas tabelas.

Mas R pode fazer isso por você. Na verdade, não é tão ruim, é só pf, e então essa é sua estatística  $F$  real que você está procurando e seus dois graus de liberdade. E nós podemos fazer isso, Opa, Eu não corri esse código. E então temos um valor  $p$  de 3,18 vezes 10 ao menos 17. Portanto, é muito, muito, muito, muito pequeno e deve estar alinhado com o resumo Comando ANOVA que eu mostrei anteriormente. E isso envolve tudo uma ANOVA unidirecional e como ler uma tabela. Na prática, é essa linha aqui, até a linha 22, que você precisa ser capaz de fazer. E a seguir, falaremos sobre ANOVAs bidirecionais.

```

1 attach(InsectSprays)
2 dim(InsectSprays)
3 str(InsectSprays)
4
5 tapply(count, spray, length)
6
7 boxplot(count~spray)
8
9 anova <- aov(count~spray, data=InsectSprays)
10 summary(anova)
11
12 #Calculando o teste F manualmente
13 N <- nrow(InsectSprays)
14 n <- 12 #nro de obs por grupo
15 k <- length(unique(InsectSprays$spray)) #Nro de grupos
16 grand_mean <- mean(InsectSprays$count)
17 group_mean <- tapply(counts=spray, mean) #Media de cada grupo
18
19 SST <- sum(n*(group_mean-grand_mean)^2)
20 MSE <- SSE/(N-k)
21
22 F_statistic <- MST/MSE
23 F_statistic
24
25 p_valor <- pf(F_statistic, df1=(N-k), lower.tail=FALSE)
26 p_valor

```

Listing 10: One-Way ANOVA no 

```

attach(InsectSprays)
dim(InsectSprays)
str(InsectSprays)

tapply(count, spray, length)

```

```

boxplot(count ~ spray)

anova <- aov(count ~ spray, data=InsectSprays)
summary(anova)

#Calculando o teste F manualmente
N <- nrow(InsectSprays)
n <- 12 #nro de obs por grupo
k <- length(unique(InsectSprays$spray)) #Nro de grupos
grand_mean <- mean(InsectSprays$count)
group_mean <- tapply(countspray,mean) #Media de cada grupo

SST <- sum(n*(group_mean-grand_mean)^2)
MST <- SST/(k-1)

SSE <- sum(InsectSprays$count - rep(group_mean, each=n))^2)
MSE <- SST/(N-k)

F_statistic <- MST/MSE
F_statistic

p_valor <- pf(F_statistic,df1=(N-k), lower.tail=FALSE)
p_valor

## Error: <text>:22:59: unexpected ')
## 21:
## 22: SSE <- sum(InsectSprays$count - rep(group_mean, each=n))^2)
##

```

## 26 Two-Way ANOVA – Tooth Growth Example

### 26.1 Décimo primeiro vídeo da segunda semana

Neste último vídeo da ANOVA, eu vou falar sobre uma ANOVA de duas vias. Nós vamos falar sobre isso conjunto de dados de crescimento dos dentes. Então, qual é a ANOVA bidirecional? Recordando o caminho único ANOVA, comparamos a variância de grupo significa considerar apenas uma variável independente. Então, no exemplo do spray de insetos, falamos apenas sobre spray como a variável independente e não foram seis tratamentos de A a F para cada um dos diferentes tipos de spray.

Na ANOVA bidirecional, podemos dividir em mais do que apenas uma variável independente. Nós podemos ter dois variáveis independentes. É um estudo se existe é uma interação entre 200 variáveis independentes na variável de resposta. A variável de resposta é nossa variável de interesse. Você pode pensar nisso como a variável dependente. As variáveis independentes são dois tratamentos. Por exemplo, podemos estudar se um tipo de spray e dosagem de spray afetará o tipo de insetos através uma ANOVA de duas vias.

Então, antes de assumirmos que havia um quantidade igual de spray, de A a F em cada dos grupos, mas quando agora podemos também variar a dose. Isso pode ser um exemplo de uma ANOVA bidirecional. Então, quais são as suposições de uma ANOVA bidirecional. Primeiro, todas as observações são independentes uma da outra, portanto a existência de uma observação não afeta a existência de outra observação.

O dependente variável é algum tipo de variável contínua. A variável dependente também é distribuído normal para cada combinação de grupos das duas variáveis independentes. Eles também têm cada uma das duas variáveis independentes composta por duas ou mais grupos categóricos. Portanto, dentro de uma variável independente, digamos, repelente de insetos, há pelo menos dois ou mais grupos categóricos.

Portanto, nessa ANOVA unidirecional, temos repelente de insetos e o grupos eram A, B, C a F. A variação para cada combinação de grupos são homogêneos, o que significa que são os mesmos. Não há significativo outliers que você poderia fazer olhar através inspeção visual. Eu acho que ajuda olhar em uma tabela de dados. Então, vamos olhar para o conjunto de dados SaratogaHouses que analisamos antes no teste AB.

Aqui nós vamos veja o tipo de aquecimento. Então esse é um dos nossos variáveis independentes e tipo de esgoto. Nossa variável de resultado ou variável dependente é o preço da habitação. Então, vamos olhar no tipo de aquecimento e tipo de esgoto e veja se essas duas categorias de variáveis têm efeito nos preços da habitação.

Então aqui está uma mesa. O tipo de aquecimento pode levar três valores possíveis. Aqui estão eles. Você pode ter ar quente como um sistema de aquecimento. Você pode ter água quente ou sistema de vapor ou sistema elétrico. Então esses são três tipos de sistemas de aquecimento. Então, lá em cima, você tem o tipo de esgoto naquele pedaço de propriedade se você teve uma fossa séptica, se está usando um público ou algum tipo de sistema de esgoto comercial ou se não houver esgoto sistema em tudo.

Cada célula tem um preço. Então, se assumirmos os dois variáveis são independentes, podemos usar este comando aqui. É o mesmo comando AOV. Agora, esta é uma ANOVA de mão dupla. Então, preço no aquecimento e esgoto. Podemos ajustar esse modelo aditivo, que é o resultado são mostrados aqui. Nós vamos ver isso em R em apenas um momento. Mas a principal coisa notar aqui é a soma dos quadrados para cada um dos tratamentos. Aqui está o quadrado médio.

Lembre-se de cada categoria de aquecimento e esgoto, existem três tipos. Portanto, os graus de liberdade para cada grupo menos um são dois. Se você pegar esta soma de quadrados dividida pelo graus de liberdade, você obtém a média ao quadrado. Aqui estão os valores F. Aqui os valores de P e estes são os números-chave que você deseja examinar aqui. Ambos são inferiores a 0,05, que é o nosso valor crítico. Então sabemos que ambos o tipo de aquecimento e o tipo de esgoto acabarão afetar o preço da habitação. Se assumirmos ou modelarmos um efeito de interação entre os dados de aquecimento e esgoto, usamos essa estrela deste aditivo plus.

O que isso faz é que parece no aquecimento como um item, o tipo de esgoto como um item e depois o efeito de interação. A maneira como eu tive que pensar sobre os efeitos da interação quando estava estudando ANOVA pela primeira vez. Gosto da ideia ou metáfora de um experimento médico em que você tinha dois medicamentos, A e B. Você pode tomar o medicamento A o efeito da droga A? Você pode tomar a droga B, o que é o efeito da droga B? Mas se você os pegasse ambos ao mesmo tempo, alguns de seus benefícios pode ser da droga A, alguns dos benefícios pode ser da droga B.

Então há alguma interação efeito A e B juntos. Ambos estão trabalhando ao mesmo tempo. Qual é o efeito da interação? Um efeito de interação pode ser executado nos dois sentidos, positivo e negativo. Então, olhando para isso efeito de interação, observando a combinação, podemos ver quais são os valores de P. Mais uma vez, eu olho para isso última coluna e posso ver que a interação termo é significativo, é menor que 0,05. Então o efeito de interação também afeta o preço da habitação. Então, vamos olhar para um exemplo em nosso estúdio. Então, aqui estamos no RStudio ambiente e eu vou carregar o Conjunto de dados SaratogaHouses.

Esse é o mesmo conjunto de dados que vimos antes. Podemos olhar para os primeiros elementos, aí estão eles. Clico duas vezes no valor aqui em cima, no canto superior direito, é o mesmo conjunto de dados. Para esta análise nós vamos olhar para o tipo de esgoto aqui e o tipo de aquecimento aqui, essas duas colunas.

Lembre-se de que o comando head nos fornece o primeiro número de elementos. Parece seis elementos. Vejamos os tamanhos das amostras. Então aqui está uma tabela do aquecimento. Você pode vê-lo no canto inferior direito da minha tela. Então aqui está o aquecimento digite e aqui está o tipo de esgoto e aqui são as variáveis de contagem. Portanto, existem 319 casas quentes com sistema de aquecimento de ar quente e uma fossa séptica contra 791 casas com água quente aquecimento de ar e usa esgoto público e 11 casas com aquecimento de ar quente e nenhum sistema de esgoto. Podemos olhar para cada desses elementos.

Portanto, este comando da tabela é uma tabela útil. Então aqui estão seus dois variáveis de interesse. Agora, se assumirmos que os dois variáveis são independentes, ou seja, esgoto e aquecimento, eles não estão relacionados. Nós podemos usar este aditivo modelo. Aqui está. Então, eu vou chamar isso anova2 e execute isso. Observe aqui que é aquecimento e esgoto. Então, para obter a saída, vou executar o comando de resumo. Aqui estão os valores. Podemos ver que o aquecimento é um componente importante de determinar o preço de uma casa ao longo com o tipo de esgoto. Se quiséssemos ter efeito de interação, usamos esse comando A de B, mas criamos o modelo dessa maneira. A diferença aqui é que eu uso o asterisco para indicar que quero inclua o termo de interação. Eu vou fazer isso, coloque os resultados em anova3, resumo de anova3. Você pode ver que todos os três agora são significativos. Aquecimento, o tipo de o aquecimento é significativo, o tipo de esgoto usado é significativo, bem como o efeito de interação. Existem significativas no nível 0,05.

Finalmente anova4. Este é essencialmente o mesmo que esta notação aqui. Mas tem aquecimento, esgoto e então eu coloquei explicitamente no efeito de interação usando a coluna de essas duas variáveis. Essa notação é útil usar se você tiver talvez mais de dois variáveis independentes. Então você tem três variáveis independentes, você terá mais de esse termo de interação. Você vai ter termo de interação com, digamos que você tenha três variáveis A/B e C, você terá o interação da interação A e B em A e C e a

interação de B e C. Se você usar este modelo multiplicativo, você terá todos os termos de interação.

Aqui, você pode excluir explicitamente os termos de interação, se tiver alguma teoria razoável para fazê-lo. Então é por isso que notação é útil.

```

1 install.packages("mosaicData")
2 library(mosaicData)
3 data(SaratogaHouses)
4
5 head(SaratogaHouses)
6
7 table(SaratogaHouses$heating, SaratogaHouses$sewer)
8 anova2 <- aov(price ~ heating, sewer, data = SaratogaHouses)
9 summary(anova2)
10
11 anova3 <- aov(price ~ heating*sewer, data = SaratogaHouses)
12 summary(anova3)
13
14 anova4 <- aov(price ~ heating+sewer:sewer, data = SaratogaHouses)
15 summary(anova4)

```

Listing 11: Two-Way ANOVA no 

```

install.packages("mosaicData")

## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)
## Warning in install.packages("mosaicData"): 'lib = "/usr/local/lib/R/site-library"' is
not writable
## Error in install.packages("mosaicData"): unable to install packages

library(mosaicData)

## Error in library(mosaicData): there is no package called 'mosaicData'

data(SaratogaHouses)

## Warning in data(SaratogaHouses): data set 'SaratogaHouses' not found

head(SaratogaHouses)

## Error in head(SaratogaHouses): object 'SaratogaHouses' not found

table(SaratogaHouses$heating, SaratogaHouses$sewer)

## Error in table(SaratogaHouses$heating, SaratogaHouses$sewer): object 'SaratogaHouses'
not found

anova2 <- aov(price ~ heating, sewer, data = SaratogaHouses)

## Error in terms.formula(formula, "Error", data = data): object 'SaratogaHouses' not found

summary(anova2)

## Error in summary(anova2): object 'anova2' not found

anova3 <- aov(price ~ heating*sewer, data = SaratogaHouses)

## Error in terms.formula(formula, "Error", data = data): object 'SaratogaHouses' not found

summary(anova3)

## Error in summary(anova3): object 'anova3' not found

anova4 <- aov(price ~ heating+sewer:sewer, data = SaratogaHouses)

## Error in terms.formula(formula, "Error", data = data): object 'SaratogaHouses' not found

summary(anova4)

## Error in summary(anova4): object 'anova4' not found

```

## 27 Entrevista com Monica Penagos – Modelos de escolha na prática

### 27.1 Primeiro vídeo da terceira semana

Oi bem vindo. Temos Monica novamente, e nesta semana vamos falar sobre o modelo de escolha binária, modelos específicos de logit e nós também vamos cobrir um pouco sobre escala multidimensional.

Então Monica, você tem algum experiência com essas ferramentas? Definitivamente. Então obrigado por me receber de novo.

Nós usamos um binômio escolha realmente prever respostas dos consumidores e prever uma resposta do consumidor. Vamos dizer para entender o talento de compra ou a pontuação líquida do promotor. Então, deixe-me dar um exemplo de como eu usei no passado.

Digamos assumir que vamos lançar uma nova formação extensão de crosta, na verdade eu fiz isso com coisas que estão no mercado hoje.

Temos uma crosta que é chamado de limpeza profunda. Essa é a pasta de dente? Pasta de dentes. Sim, um crosta chamada Deep Clean e outra crosta isso é chamado Outlast. Então, fazemos testes de conceito para entender o que são as alegações de que os consumidores gostariam de ver e eles vão aumentar sua compra talento dessas escolhas.

Portanto, não apenas as reivindicações tem que ser verdade, mas temos muitas opções de quais reivindicações colocamos em um pacote. Mas em um pacote há único espaço para uma reivindicação.

Então, qual reivindicação você escolhe? Usamos um binômio e, mais cedo, no processo usamos multinomial também quando temos muitos opções para tentar diminuir e obter para uma escolha binomial. Bem, então você usa binomial modelos de escolha em testes de conceito e que ajuda a identificar?

Reivindicações que irão no mercado. Reivindicações que irão no mercado?

Sim. Isso é ótimo. Então em aula desta semana, vamos falar sobre essas técnicas. Espero que você goste do próximo conjunto de vídeos.

## 28 Leitura da semana

### 28.1 K means clustering

## 29 Modelo de Escolha Binária: Modelo Logit

### 29.1 Segundo vídeo da terceira semana

Nesta palestra, vamos conversar um pouco sobre regressão logística. Estes são modelos de resultados binários. Então vamos começar. Então, primeiro de tudo, temos que discutir o que é um modelo de logit.

E um modelo de logit é um modelo binário, binário que significa dois, usado em marketing, e usa uma função logística para modelar uma variável dependente binária. Então, o que é uma variável dependente binária? A variável dependente  $y$ , essa é a nossa variável dependente, é uma escolha binária. Ou será 0 ou 1. Portanto, você pode pensar em alguns exemplos do que é uma variável 0 ou 1.

Ou se um cliente não compra um widget, é possível codificá-lo como 1 se ele comprar, um 0 se eles não comprarem. Gênero geralmente é descrito como uma variável binária, homem-mulher, 0, 1. Partido político republicano e democrata afiliação, algo assim.

Então aqui está a fórmula para a função logit. E de várias maneiras, estudando regressão logística, trata-se de entender essa função de logit. O lado direito da equação é essencialmente o mesmo que regressão linear. Então, é sobre ser capaz de converter esses  $Y$  iguais a 1,  $Y$  iguais a 0 e suas probabilidades nesta função de logit. E então, também poder voltar para fins de interpretação. Então, vamos começar aqui. Aqui está a função logit. A função logit é a probabilidade de que  $Y$  seja igual a 1. E assumimos a probabilidade de que  $Y$  é igual a 1 no numerador aqui. E então, tomamos a probabilidade de  $Y$  não é igual a 1 no denominador. Então, vamos estudar isso primeiro parte um pouco. Aqui vamos nós. Então, se tivermos a probabilidade de  $Y$  igual a 1, igual a 1 é dizer 80%, ou 0,8. Então a probabilidade que  $Y$  não é igual a 1, probabilidade de que  $Y$  não seja igual a 1, é igual a 1 menos 0,8, certo? E é isso que está acontecendo aqui, menos um a probabilidade de que  $Y$  é igual a 1. Lembre-se dos axiomas de Kolmogorov que o a soma das probabilidades tem que ser igual a 1. Em outras palavras, algo acontece no seu espaço de probabilidade. Ok, então e então, você pega o log desse número. Então, há muitas coisas acontecendo.

Então, vamos primeiro olhar a probabilidade de probabilidades. E o que isso significa? Então as probabilidades, como eu mencionei, é a probabilidade de um evento ocorrer mais de 1 menos a probabilidade de o evento o evento que está ocorrendo. E se você quiser ir no outro direção, aqui está a fórmula. Se você deseja calcular a probabilidade, você assume as probabilidades acima de 1 mais as probabilidades. Então,

vamos dar um exemplo de a probabilidade de algum evento acontecer, digamos que é 0,8. Isso significa que a probabilidade de o evento que não está acontecendo é 0,2.

E então, as chances são iguais a 4, ok? E então, se você quiser passar de 4 de volta à probabilidade, são 4 sobre 5. E que, se você fizer as contas, vai sair para 0,8, certo? Então, eu tenho aqui neste slide, nesta primeira coluna, só tenho probabilidades variando de 0 a 1. E acabei de aumentar em 10%. Então, 0% de probabilidade, 10% de probabilidade, 20, 30, 40, até 100% de probabilidade. E então, calculei as chances de cada um desses valores. Então 0 sobre 1 menos 0 é 0. Este é o que eu acabei de fazer no slide anterior.

A probabilidade de 0,8 sobre 0,2 é uma probabilidade de 4. E então, este é apenas todos os cálculos. E podemos ver o relacionamento entre a probabilidade e as probabilidades. Portanto, as probabilidades devem variar entre 0 e 1 inclusive. E podemos ver o relacionamento essas probabilidades têm contra a probabilidade. Neste slide, eu tenho chances aqui. E na outra coluna, Eu calculei o log das probabilidades. E esse é apenas o logaritmo natural.

Observe que isso é indefinido aqui, assim como aqui. Mas você pode ver como probabilidades ir nessa direção, as chances de log, é uma curva côncava. A seguir, chegamos à função logística, que é a probabilidade de log aqui, e aqui estão as probabilidades. Então, no eixo y, tenho probabilidades. E observe, as probabilidades de log aqui podem ir do infinito no lado direito e menos infinito. Portanto, pode variar a linha numérica completa. Mas no eixo y, você verá as probabilidades variam de 0 a 1. E é basicamente isso que estamos tentando descobrir com variáveis binárias.

Eles vão comprar, não por? E qual é a probabilidade que eles comprem? E a probabilidade que eles comprem é em algum lugar entre 0 e 1, ok? Então agora que limitamos o alcance as probabilidades de nossa variável de resultado, e podemos usar essas probabilidades de log que vai de 0 a menos infinito a mais infinito, agora podemos executar algum tipo de regressão. Antes de mostrar como executar uma regressão logística, eu queria mostrar como Eu criei esses gráficos.

E espero que isso ajude você a uma sensação melhor do que está acontecendo. Então probabilidades. Eu apenas peguei uma sequência de números de 0 a 1 e incrementado em 0,1. Então isso é 0, 0,1, 0,2, 0,3. Então é isso que esse comando faz aqui. E se eu olhar para isso, você pode ver os valores reais abaixo. 0,0, 0,1, 0,2, etc., etc. E então, eu tenho probabilidades. E é exatamente como eu descrevi, está tomando a probabilidade mais de 1 menos a probabilidade. E se olharmos para essa variável, probabilidades, aí estão elas.

E se olharmos para essa variável, probabilidades, aí estão elas. E então, tomar o log das probabilidades é apenas essencialmente tomando exatamente isso. Para calcular o natural logaritmo de um número em R, você usa esta função de log, log. Aqui vamos nós. E então, esses próximos três três comandos, Estou apenas criando mesinhas que você viu no lado esquerdo desses slides do PowerPoint.

Então, C significa Columbine. Toma a coluna das probabilidades e a coluna das probabilidades e reuni-las um pequeno quadro de dados. Então vamos fazer isso. E eu farei todos eles para que você possa ver. E eu vou te mostrar uma, e lá estão eles. Eles são essencialmente como eu consegui o pequeno tabelas nos slides do PowerPoint. Mas agora, eu quero te mostrar como criar esses gráficos. Então, aqui, eu fiz a mesma sequência para probabilidades de 0 a 1. Dessa vez, em vez de aumentar por 0,1, estou usando 100. Isso significa dar-me 100 números entre 0 e 1. E a razão pela qual faço isso é apenas para as funções gráficas. Isso me dá um pouco mais de granularidade. Mas você pode ver lá, vai de 0 a 1, e há 100 números desses. Novamente, as probabilidades são calculadas da mesma maneira.

Agora, eu vou traçar probabilidade contra probabilidades. E este é o gráfico que mostrei você nos slides do PowerPoint. Vou criar as probabilidades de log número novamente e traçar isso. Aqui vamos nós. Aqui vamos nós. E depois, aqui estão as probabilidades de log contra a probabilidade. E observe que isso é o gráfico que eu mostrei antes.

E essa é outra maneira de criar isso função logística com essa curva S, e isso parece um pouco melhor. E, novamente, os principais pontos a serem mantidos a mente é que x pode variar de menos infinito a infinito positivo, enquanto o eixo y varia de 0 a 1. Agora, eu vou falar sobre a regressão logística. Aqui está a nossa função de logit. E no lado direito da o seu seu modelo de regressão, você tem os mesmos betas e x é o que você vê. Então aqui está beta 0 mais beta 1x, estritamente falando, 1.

E como o Y é agora uma função logit, a interpretação dos betas torna-se um pouco mais complicado. Então beta 0 é a probabilidade de log do evento quando x é 0. Portanto, seja qual for o seu x, se x for 0, você ficará com essa quantidade beta 0 lá. E então, se seu x1 é binário, beta 0 mais beta 1 são as probabilidades de log de o evento em que x é igual a 1. Se x é algum tipo de variável contínua, então isso realmente se resume a beta 0 mais beta 1 vezes x. Então beta 1 é a diferença de as probabilidades de log do evento quando x é 1 em comparação com x é 0 se x é uma variável binária.

As probabilidades, novamente, são a razão da probabilidade de um evento ocorrer contra a probabilidade de que isso não aconteça. Nós conversamos sobre isso antes. E este é o exemplo Eu te dei mais cedo. É exatamente a mesma coisa. O odds ratio é o odds ratio de um evento que ocorre em um grupo em comparação

com as chances dele ocorrendo em outro grupo.

Então  $P_1$  acima de 1 menos  $P_1$ . Então essas são as chances de acontecer no grupo 1 versus as chances de isso acontecer no grupo 2. E isso é indicado no denominador. O que você deve ter em mente é que, se o odds ratio for igual a 1, isso significa que o evento é igualmente provável acontecer no grupo 1 ou no grupo 2. É igualmente provável. Se o odds ratio for maior que 1, isso significa que há um maior grau de probabilidade de que ocorrerá no primeiro grupo. E se for menor que 1, será menos provável de ocorrer no primeiro grupo e mais provável de ocorrer no segundo grupo.

```

1 probs <- seq(0.1, by = 0.1)
2 probs
3 odds <- probs/(1-probs)
4 odds
5 logodds <- log(odds)
6 logodds
7 tab1 <- cbind(probs, odds)
8 tab1
9 tab2 <- cbind(odds, logodds)
10 tab2
11 tab3 <- cbind(logodds, probs)
12 tab3
13
14 #Repete para os plots
15 probs <- seq(0.1, length=100)
16 probs
17 odds <- probs/(1-probs)
18 odds
19 plot(probs, odds, type="l")
20 logodds <- log(odds)
21 lododds
22 plot(odds, logodds, type="l")
23 plot(logodds, probs, type="l")
24
25 #Gráfico Logit
26 x <- seq(-6,6, length=100)
27 x
28 logistifn <- exp(x)/(1+exp(x))
29 logistifn
30 plot(x, logistifn, type="l")
31
32 #Exemplo 1
33 discount <- c(0,0,0,0,0,0,0,50,50,50,50,50,100,100,100,100,150,150,150,200,200)
34 buy <- c(0,0,0,1,0,0,0,1,0,0,1,1,1,0,1,1,0,1,1,1)

```

Listing 12: Regressão Logística no 

```

probs <- seq(0.1, by = 0.1)
probs

## [1] 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

odds <- probs/(1-probs)
odds

## [1] 0.1111111 0.2500000 0.4285714 0.6666667 1.0000000 1.5000000 2.3333333
## [8] 4.0000000 9.0000000      Inf

logodds <- log(odds)
logodds

## [1] -2.1972246 -1.3862944 -0.8472979 -0.4054651 0.0000000 0.4054651
## [7] 0.8472979 1.3862944 2.1972246      Inf

tab1 <- cbind(probs, odds)
tab1

##      probs      odds
## [1,] 0.1 0.1111111
## [2,] 0.2 0.2500000

```

```
## [3,] 0.3 0.4285714
## [4,] 0.4 0.6666667
## [5,] 0.5 1.0000000
## [6,] 0.6 1.5000000
## [7,] 0.7 2.3333333
## [8,] 0.8 4.0000000
## [9,] 0.9 9.0000000
## [10,] 1.0 Inf

tab2 <- cbind(odds, logodds)
tab2

##          odds    logodds
## [1,] 0.1111111 -2.1972246
## [2,] 0.2500000 -1.3862944
## [3,] 0.4285714 -0.8472979
## [4,] 0.6666667 -0.4054651
## [5,] 1.0000000 0.0000000
## [6,] 1.5000000 0.4054651
## [7,] 2.3333333 0.8472979
## [8,] 4.0000000 1.3862944
## [9,] 9.0000000 2.1972246
## [10,] Inf Inf

tab3 <- cbind(logodds, probs)
tab3

##          logodds probs
## [1,] -2.1972246 0.1
## [2,] -1.3862944 0.2
## [3,] -0.8472979 0.3
## [4,] -0.4054651 0.4
## [5,] 0.0000000 0.5
## [6,] 0.4054651 0.6
## [7,] 0.8472979 0.7
## [8,] 1.3862944 0.8
## [9,] 2.1972246 0.9
## [10,] Inf 1.0

#Repete para os plots
probs <- seq(0.1, lenght=100)

## Warning: In seq.default(0.1, lenght = 100) :
## extra argument 'lenght' will be disregarded

probs

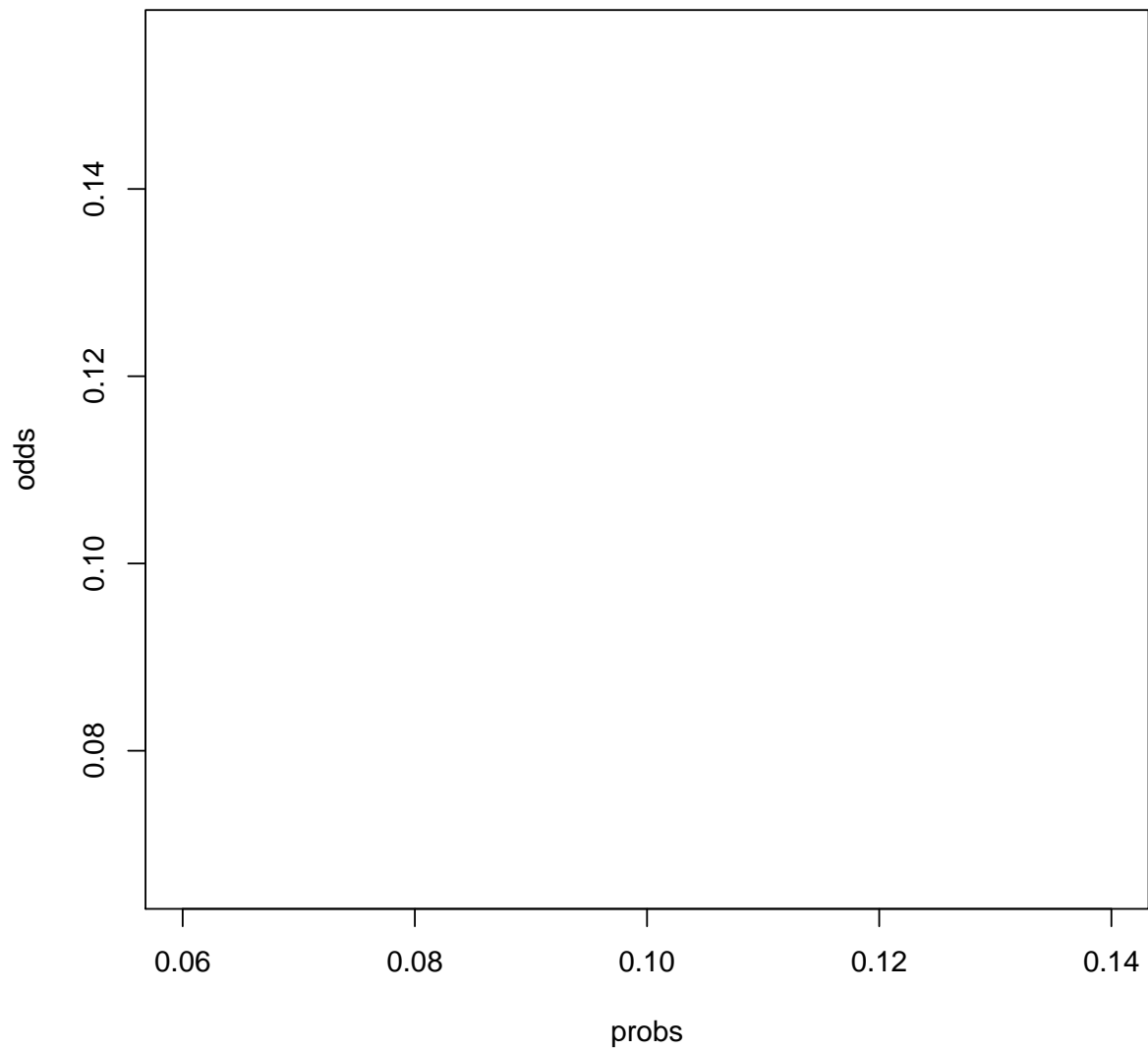
## [1] 0.1

odds <- probs/(1-probs)
odds

## [1] 0.1111111

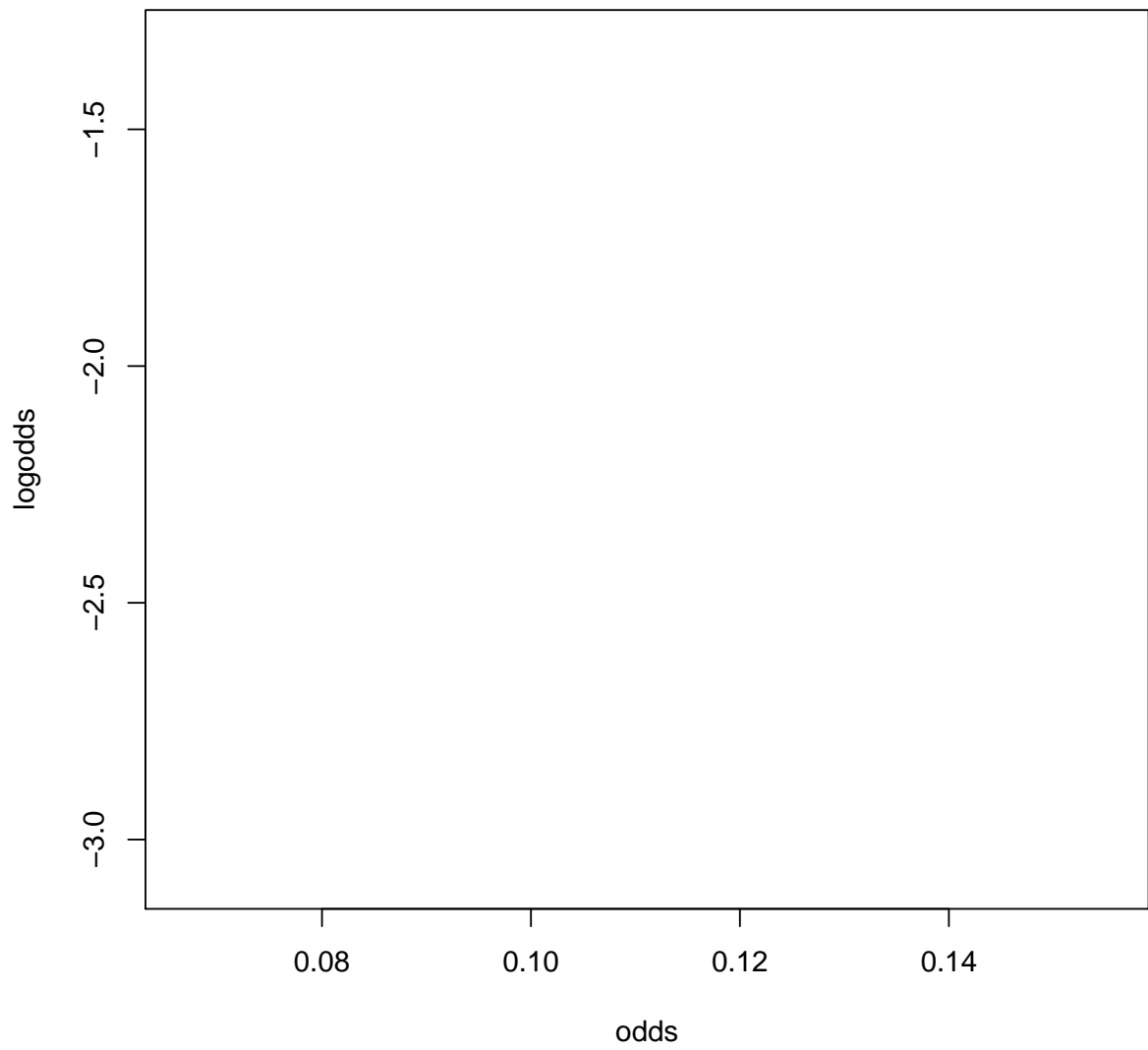
plot(probs,odds,type="l")
```



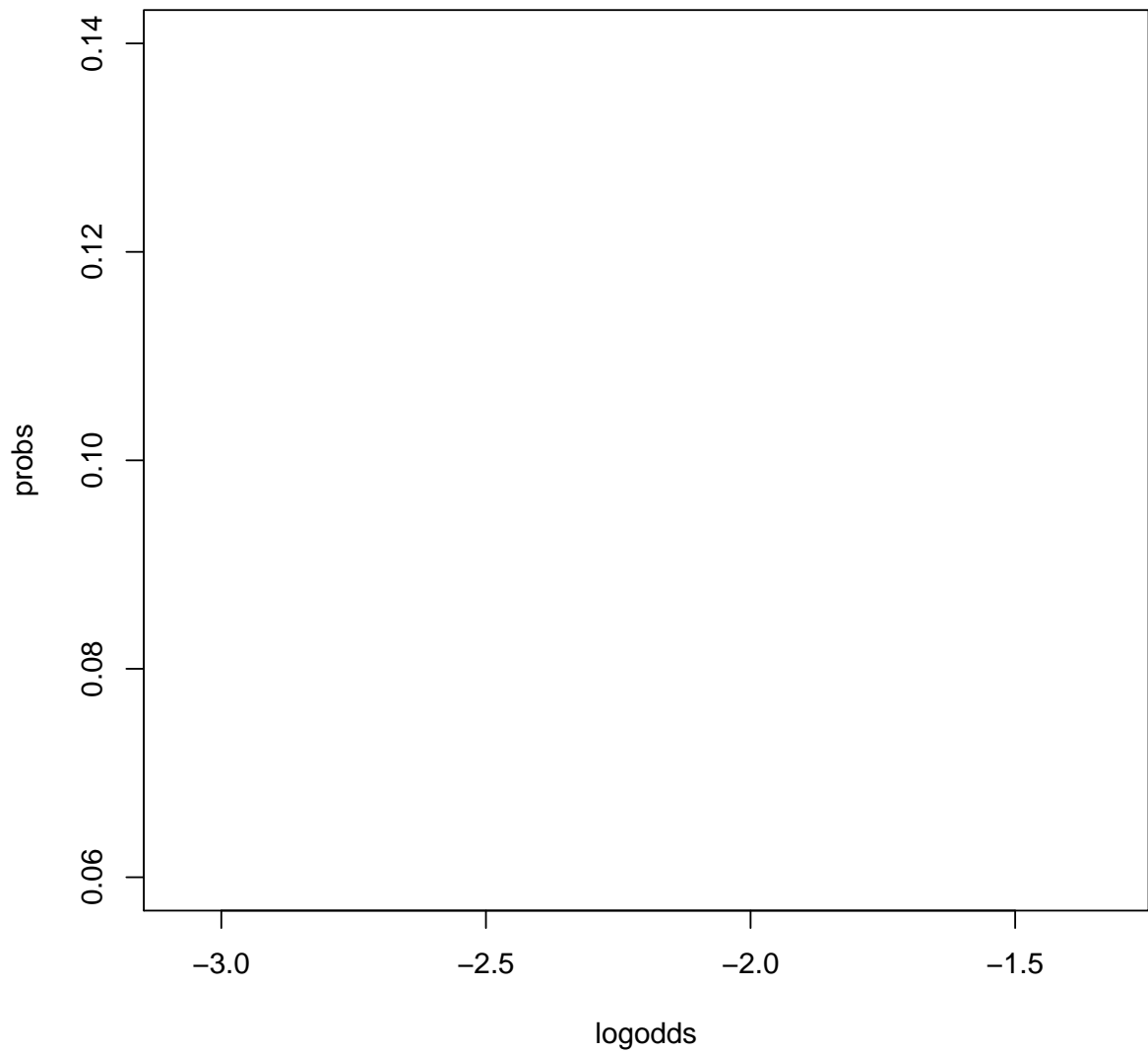


```
logodds <- log(odds)
lododds

## Error in eval(expr, envir, enclos): object 'lododds' not found
plot(odds,logodds,type="l")
```



```
plot(logodds,probs,type="l")
```



```
#Gráfico Logit
x <- seq(-6,6,lenght=100)

## Warning: In seq.default(-6, 6, lenght = 100) :
## extra argument 'lenght' will be disregarded

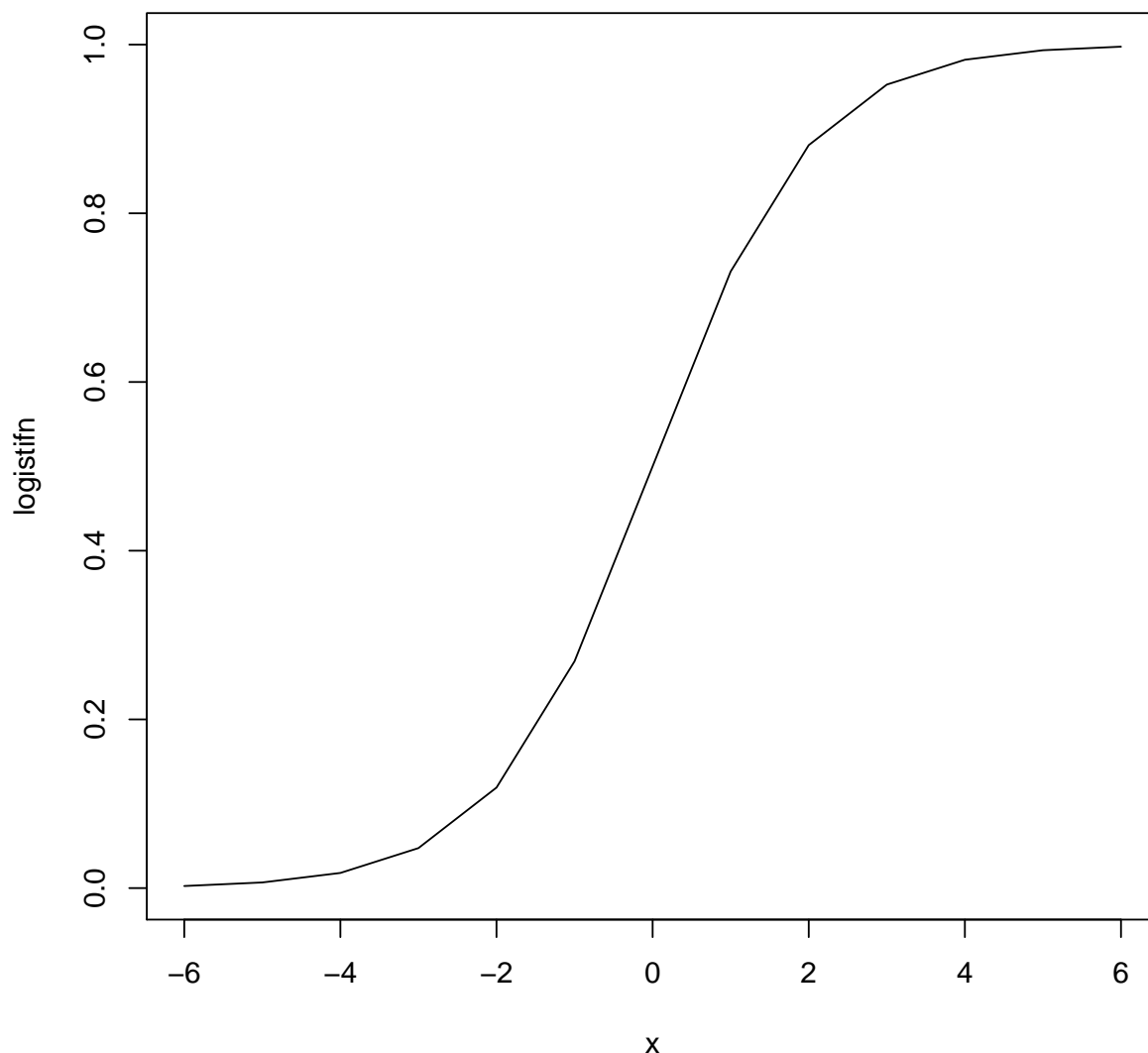
x

## [1] -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6

logistifn <- exp(x)/(1+exp(x))
logistifn

## [1] 0.002472623 0.006692851 0.017986210 0.047425873 0.119202922 0.268941421
## [7] 0.500000000 0.731058579 0.880797078 0.952574127 0.982013790 0.993307149
## [13] 0.997527377

plot(x,logistifn,type="l")
```



## 30 Modelo de Escolha binária: Exemplo do modelo Logit

### 30.1 [Terceiro vídeo da terceira semana](#)

Neste ponto, eu acho é melhor começar a olhar para um exemplo real. Espero que solidifique e esclareça sua compreensão de regressão logística. Então, aqui estão alguns dados inventados da empresa Acme Widget. Eles fizeram isso por pesquisa de escolha do cliente, de comprar algum widget em 20 ocasiões, obtivemos 20 pontos de dados. Decidimos usar um número de descontos. Sem desconto aqui, sem desconto, um desconto de US\$ 50, Desconto de US\$ 100, etc. Então, incrementos de 50 de 0-200 e aqui, você pode ver o caminho esses dados estão configurados. O primeiro grupo de itens não tem desconto. Nesse caso, 1, 2, 3, 4, 5, 6, um em cada seis indivíduos decidiu fazer a compra sem nenhum desconto.

Agora, sua intuição deve dizer que, quanto maior o desconto ou maior, maior a probabilidade de compra, mantendo outras coisas constantes. Portanto, com o desconto de US\$ 200 aqui, você pode ver que havia duas opções para comprar com desconto de US\$ 150, dois terços das pessoas selecionado para comprar, e aqui está o desconto de US\$ 100 e depois diminui. Então, nossa amostra de dados, temos 20 ocasiões ou n é igual a 20, nossa variável independente é o desconto que varia de 0-200 em incrementos de 50.

Não precisa ser em incrementos de 50. Nossa variável de escolha binária é aquela quando o cliente compra e é zero quando eles não compram. Então, um significa comprar e zero significa não comprar. Então, vamos tentar descobrir quais são as probabilidades usando esse modelo aqui. Então aqui está a nossa Função Logit,

Beta 0, Beta 1 vezes o desconto. Então, antes de irmos a esses resultados, vamos olhar para o nosso nota para calcular. Tudo certo. Então aqui estou no ambiente R.

Deixe-me limpar alguns dessas coisas de antes. Limpe o meu também, limpar o meio ambiente. Eu tenho duas variáveis; o desconto e a variável buy, e posso cindir desconto, buy. Você pode ver aqui que eu essencialmente codifiquei o que estava acontecendo o slide do PowerPoint. Então uma pessoa comprou com um desconto de zero, e então você vê para baixo aqui com um desconto de US\$ 200, todo mundo comprou, etc., então está. Aqui está a linha para realmente execute a regressão. Em alguns níveis, em alguns aspectos, o código real para executá-lo em R não é complicado.

O truque é tudo a interpretação. Então aqui, olhe para este lado. GLM é um modelo linear geral e seu dependente A variável é buy e, em seguida, você coloca seu til e, em seguida, suas variáveis independentes à direita do til. Esse é o rabisco linha, é o til. Então família é igual a binômio. É assim que sabe que é um modelo de escolha binária. Eu vou correr isso e eu vou colocá-lo na variável ex1. Esta é a técnica R padrão de executar uma regressão de qualquer tipo.

Você coloca os resultados em algum nome de variável e então você me diz o resumo de esse nome de variável. Então vamos fazer isso. Aqui, eu nomeei a variável ex1 para o exemplo 1, e aí está. Podemos olhar, há a fórmula. Aqui estão as estimativas do Beta, então o Beta 1 é estimado seja menos 1,45 e 0,02 para o número Beta 1. Então, vamos voltar ao Slides do PowerPoint e eu vou falar a interpretação. Então aqui, eu basicamente apenas cortei e cole os resultados de R aqui. Existem os estimativas aqui.

Eles são significativos no nível 0,05, e o termo de interceptação é significativo no nível 0,1. Bem, vamos lá por enquanto. Aqui está o nosso modelo. Portanto, a função logit, o log das probabilidades é igual a menos 1,46, vem daqui, mais duas vezes o desconto, então copiei na parte superior. Portanto, Beta 0 é 1,46 são as chances de log de comprar um widget quando houver não tem desconto. Beta 1 é igual a 0,02 vezes o valor do desconto, lembre-se de que foi 50, 100, 150, 200. É a diferença no as probabilidades de log do widget quando o desconto é comparado a quando não há desconto. Você pode pensar nisso dessa maneira. Então, vamos dar um exemplo concreto. Então eu acho que isso faz muito mais sentido. Se o desconto for de US\$ 100, a versão Beta 0 mais a versão 1 é igual a menos 1,46, esse é o termo de interceptação, mais 0,02 vezes 100 que passa a ser igual a 0,54.

Essa é a probabilidade de log de comprar um widget quando houver um desconto de US\$ 100. Então, teremos que ter em mente essas duas fórmulas, se o log de a for igual a b, se você exponenciar ambos os lados, e para o logaritmo de a é igual ao e ao b. Probabilidades de log do Beta 0, as probabilidades de log que temos determinado a partir de nossa regressão, se houver sem desconto é Beta 0, então nossas chances são de e para o Beta 0. Então, vamos calcular as probabilidades. Quando não há desconto, x é 0, as probabilidades de log são Beta 0 mais Beta 1 vezes 0 ou Beta 0. Então apenas exponenciamos, então e para menos 1,46 é 0,23.

Então essas são as chances. Eu acho que a maioria das pessoas é mais confortável em pensar nessas coisas em termos de probabilidade do que de probabilidades. Portanto, a probabilidade de comprar um widget sem desconto, basta conectá-lo ao essa fórmula de P, Probabilidade é igual à odds acima de 1 mais odds. Então é isso que é, as probabilidades 0,2322 sobre um mais as probabilidades, e isso é igual a 18,8 ou cerca de 19%. Então há 19% chance de alguém comprar este widget sem desconto. Sua intuição deve dizer você, à medida que os descontos aumentam, esse número também deve subir. Então, vamos dar um exemplo onde x é igual a 100. Então, aqui vamos nós, o as probabilidades de log são menos 1,46 mais 0,02 vezes 100 e obtemos esse valor de 0,54. Isso é igual a 0,54 e então e para o 0,54 é 1,716. Lembre-se de que essas são as probabilidades, então queremos converter em probabilidade e tem esse valor aqui. Portanto, há 63% de chance de uma pessoa comprar o widget com desconto de US\$ 100. Vamos recapitular bem rápido. Até agora, devemos entender o que o modelo de logit é, a função de logit é, e essa curva S acentuada e por que escolhemos uma função sem curva S. Devemos saber como interpretar os coeficientes e a função logit, especialmente através de exemplo específico. Em certo sentido, é não é tão fácil quanto interpretar regressão linear.

A regressão linear é apenas uma mudança de uma unidade em x. É igual a uma mudança Beta em Y. Você tem que passar por essas pequenas questões matemáticas transformações para obter a probabilidade. Com o tempo, você deve ser capaz pensar em termos de chances, mas vamos começar com os fundamentos. Eu acho que as pessoas entendem melhor a probabilidade, então vamos começar por aí. Você deve ser capaz de calcular as probabilidades, a probabilidade, e o odds ratio. Então eu quero olhar em outro exemplo. Vou usar esse conjunto de dados do CAFE para ilustrar a função logit em R.

Portanto, este conjunto de dados contém informações sobre o voto de 50 democratas e 49 republicanos, e isso é para o 11ª alteração na média corporativa padrão de economia de combustível para carros e caminhões. Portanto, neste conjunto de dados, há um binário variável de resposta, uma se eles votaram pela emenda, zero se eles votaram contra a emenda. Temos um binário variável independente, sua afiliação política, um se o senador for republicano e zero se Eles são democratas.

Então nós temos isso valor variável, e essa é a contribuição total vitalícia feita a partir de a indústria automobilística. Vamos transformar isso tomando o log natural por causa de alguns outliers. Note que eu tirei isso de um artigo, estou replicando o estudo no trabalho de precedência, as citações no Slide do PowerPoint, e eu só quero que você saiba que eu só vou pegar um log natural para a transformação, mas no artigo, ele usa essa transformação de tomar a base de log 10 em vez do log natural, e ele o multiplica por 10. Pega a quantidade, multiplica por 10 e, em seguida, isso dá a você uma compreensão de quantos números são nos montantes em dólares. Então, se for \$ 1.000, ele tem quatro algarismos, o número de dígitos. Ele argumenta em seu trabalho que é uma maneira de conseguir um lidar com a interpretação, mas acho que a matemática é um pouco complicada, então só vou ao registro natural por enquanto. Os resultados são virtualmente idêntico como veremos.

```

1
2 #Exemplo 1
3 discount <- c(0,0,0,0,0,0,50,50,50,50,50,100,100,100,100,150,150,150,200,200)
4 buy <- c(0,0,0,1,0,0,0,1,0,0,1,1,1,0,1,1,0,1,1,1)
5
6 cbind(discount,buy)
7
8 ex1 <- glm(buy~discount, family = binomial)
9 summary(ex1)
10
11 #Busque os dados
12 current_working_dir <- dirname(rstudioapi::getActiveDocumentContext()$path)
13 setwd(current_working_dir)
14
15 data <- read.csv("cafe.txt", header=TRUE, sep = ",")
16 head(data, n=5)

```

Listing 13: Exemplo de aplicação do modelo Logit no 

```

#Exemplo 1
discount <- c(0,0,0,0,0,0,50,50,50,50,50,100,100,100,100,150,150,150,200,200)
buy <- c(0,0,0,1,0,0,0,1,0,0,1,1,1,0,1,1,0,1,1,1)

cbind(discount,buy)

##          discount buy
## [1,]           0   0
## [2,]           0   0
## [3,]           0   0
## [4,]           0   1
## [5,]           0   0
## [6,]           0   0
## [7,]          50   0
## [8,]          50   1
## [9,]          50   0
## [10,]         50   0
## [11,]         50   1
## [12,]        100   1
## [13,]        100   1
## [14,]        100   0
## [15,]        100   1
## [16,]        150   1
## [17,]        150   0
## [18,]        150   1
## [19,]        200   1
## [20,]        200   1

ex1 <- glm(buy~discount, family = binomial)
summary(ex1)

##
## Call:
## glm(formula = buy ~ discount, family = binomial)

```

```
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -1.8848  -0.7338  -0.1346   0.9453   1.8268
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.459653    0.838949  -1.740   0.0819 .
## discount     0.020336    0.009696   2.097   0.0360 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 21.459  on 18  degrees of freedom
## AIC: 25.459
##
## Number of Fisher Scoring iterations: 4

#Busque os dados
current_working_dir <- dirname(rstudioapi::getActiveDocumentContext())$path

## Error: RStudio not running

setwd(current_working_dir)

## Error in setwd(current_working_dir):  object 'current_working_dir' not found

# The data is taken from
# DePaolo, C.A., & Robinson, D. F. (2011). "Cafe data". Journal of Statistics Education, 19(1).
# https://www.tandfonline.com/doi/pdf/10.1080/10691898.2006.11910586

# I have modified their model by only taking the natural log of the dollar amount.
# I add one to the amount as log(0) undefined.

data <- read.csv("cafe.txt", header=TRUE, sep = ",")

## Error in read.table(file = file, header = header, sep = sep, quote = quote, : duplicate
'row.names' are not allowed

head(data, n=5)

##
## 1 function (... , list = character(), package = NULL, lib.loc = NULL,
## 2     verbose = getOption("verbose"), envir = .GlobalEnv)
## 3 {
## 4     fileExt <- function(x) {
## 5         db <- grepl("\\\\.([.]+\\\\.\\.(gz|bz2|xz))$", x)
```

## 31 Modelo de Escolha Binária – Modelo Logit: Exemplo 2

### 31.1 Quarto vídeo da terceira semana

Então, vamos olhar para o nosso exemplo. Então aqui estou eu no R Ambiente de estúdio. Essas uma, duas, três, primeiras três linhas de código, que apenas nos dizem qual é o diretório ativo está no seu computador, defina esse diretório como seu diretório de trabalho e, em seguida, procura esse arquivo, `cafe.csv`, e importa-o nesses dados variáveis.

Então, se olharmos para o dados, aí está. A coisa que você tem que lembrar é garantir que seu programa R e o arquivo Excel, `.csv` estejam na mesma pasta, e deve funcionar bem. Então aqui estão os dados e aqui são os nomes dos senadores, seu estado, não vamos estar usando essa variável.

REP significa republicano. Um é um sim. Zero é um não. Como eles votaram nisso alteração específica e o valor em dólar que eles recebem da indústria automobilística. Então eu vou dar uma transformação da quantidade de dados. Nós vamos pegar o log e eu vou colocá-lo nesta variável, `Amt1`. Note que eu adicionei um para este número. Esta é uma técnica comprometida. A ideia aqui é que alguns dos valores no conjunto de dados são zero e o log de zero é indefinido. Então eu adiciono um apenas para contornar isso.

Vamos executar esse código bem rápido. Eu posso olhar para o primeiros cinco valores dessa quantidade e lá estão eles. Então esse é o log de os montantes mais um. Aqui está a regressão. Note, `vote` é nosso binário variável dependente, `Tilda`. A variável independente está do lado direito. Este exemplo é legal porque Eu tenho uma variável binária.

Portanto, sejam eles republicanos ou democratas. Então, aqui estão as quantidades neste vetor de coluna do doações que receberam, contribuições políticas que receberam da indústria automobilística. Mais uma vez, vamos pegue, execute essa regressão, coloque-a em algum nome de variável e, em seguida, execute o resumo do nome da variável, a técnica R padrão. Então, eu executei a regressão e você pode ver aqui a tabela dos coeficientes, todos são significativos, pelo menos até o nível 0,05 e ainda mais então no nível 0,01.

O termo de interceptação é menos 4,47. O Beta-1 é estimado em 1,8479 e o coeficiente de o valor é 0,49. Então, vamos ver como interpretar isso. Essa é a equação. Aqui está o termo de interceptação. Aqui está o Beta-1 para Republicano ou não republicano, e há 0,4915 para a quantidade. Portanto, a partir da função logit, devemos poder calcular o logit, as probabilidades, a probabilidade de um voto no SIM e também calcular e interpretar a razão de chances. Então vamos fazer isso. Eu reescrevi a equação aqui em cima na linha superior. Vamos dar uma primeira exemplo de um democrata.

Então isso significaria REP é igual a zero, que não recebeu nenhum valor do fabricante de automóveis. Então `Amount1` é igual a zero e, na verdade, `Valor1` era, na verdade, o log de um nesse caso. Então aqui temos. Então, realmente, é o logit de  $y$  é igual a um é igual a menos 4.4. As probabilidades de log de um voto no SIM de um democrata é menos 4.4. Exponenciamos  $E$  para o menos 4,4 é 0,01. Então, podemos calcular a probabilidade aqui.

Então há um por cento chance de votar SIM neste projeto, uma vez que são democratas e eles não receberam nada de a indústria automobilística. Então, vamos olhar mais dois casos no esquerda e direita. Um é o caso onde o senador é republicano e o outro caso é onde o senador é democrata. Nesse caso, eles ambos receberam US\$ 25.000 do fabricantes de automóveis. Então aqui está o modelo. Ambos têm o termo de interceptação. Ambos têm 0,4915 vezes o valor de US\$ 25.000. Na verdade, é o log mais 1, a quantidade mais 1 e pegue o log.

Então, observe o democrata aqui tem um zero para a afiliação do partido enquanto aqui é 1,8479 vezes 1 para o Filiação do Partido Republicano. Então, se você calcular o número deste lado, você recebe 235. Aqui, é quase 0,5. Exponenciados das probabilidades de uma votação republicana são 10,48 enquanto as chances de um voto democrata é 1,65. Então, passamos das probabilidades para probabilidade usando a fórmula, as probabilidades acima de 1 mais as probabilidades. Portanto, neste caso, se você é republicano e você recebeu US\$ 25.000 da fabricantes de automóveis, você obterá o probabilidade de você votar sim como 91%, enquanto, aqui, se você é democrata, a probabilidade de você votar é 0,59 ou 59 por cento. Então isso termina regressão logística. Encorajo-vos a passar pelo código. No código R, eu tenho realmente mostrou a metodologia usado para transformar a variável de quantidade usada no artigo de precedência e também incluí um link para o artigo de precedência dentro do código R. Espero que isso dê você uma compreensão da regressão logística.

```

1
2 # Get data
3 current_working_dir <- dirname(rstudioapi::getActiveDocumentContext()$path)
4 setwd(current_working_dir)
5 data <- read.csv(file="cafe.txt",header = TRUE, sep = ",")
6 head(data, n=5)
7
8 # The data is taken from

```



```

9 # DePaolo, C.A., & Robinson, D. F. (2011). "Cafe data". Journal of Statistics Education,
   # 19(1).
10 # https://www.tandfonline.com/doi/pdf/10.1080/10691898.2006.11910586
11
12 # I have modified their model by only taking the natural log of the dollar amount.
13 # I add one to the amount as log(0) undefined.
14
15 Amt1 = log(data$Amount+1)
16 head(Amt1,n=5)
17
18 # fit logistic regression model
19 lr_model1 <- glm(Vote ~ REP + Amt1, family = binomial, data)
20 summary(lr_model1)
21
22
23 ### Below is the code to run the model run in Preston's paper.
24 ### See https://www.tandfonline.com/doi/pdf/10.1080/10691898.2006.11910586
25 # Transform Amount to log-transformation. This transformation "express"
26 # how many digits are in the dollar amount.
27
28 Amt2 = log10(10*data$Amount+1)
29 head(Amt2,n=5)
30
31 # fit logistic regression model
32 lr_model2 <- glm(Vote ~ REP+AMT2, family = binomial, data)
33 summary(lr_model2)

```

Listing 14: Rodando o modelo de regressão logística no 

```

# Get data
current_working_dir <- dirname(rstudioapi::getActiveDocumentContext())$path

## Error: RStudio not running

setwd(current_working_dir)

## Error in setwd(current_working_dir): object 'current_working_dir' not found

data <- read.csv(file="cafe.txt",header = TRUE, sep = ",")

## Error in read.table(file = file, header = header, sep = sep, quote = quote, : duplicate
'row.names' are not allowed

head(data, n=5)

##
## 1 function (... , list = character(), package = NULL, lib.loc = NULL,
## 2     verbose = getOption("verbose"), envir = .GlobalEnv)
## 3 {
## 4     fileExt <- function(x) {
## 5         db <- grepl("\\\\.([^.]+\\\\\\\\.(gz|bz2|xz)$", x)

# The data is taken from
# DePaolo, C.A., & Robinson, D. F. (2011). "Cafe data". Journal of Statistics Education, 19(1).
# https://www.tandfonline.com/doi/pdf/10.1080/10691898.2006.11910586

# I have modified their model by only taking the natural log of the dollar amount.
# I add one to the amount as log(0) undefined.

Amt1 = log(data$Amount+1)

## Error in data$Amount: object of type 'closure' is not subsettable

head(Amt1,n=5)

## Error in head(Amt1, n = 5): object 'Amt1' not found

# fit logistic regression model
lr_model1 <- glm(Vote ~ REP + Amt1, family = binomial, data)

```

```
## Error in as.data.frame.default(data, optional = TRUE): cannot coerce class "'function'"
to a data.frame

summary(lr_model1)

## Error in summary(lr_model1): object 'lr_model1' not found

### Below is the code to run the model run in Preston's paper.
### See https://www.tandfonline.com/doi/pdf/10.1080/10691898.2006.11910586
# Transform Amount to log-transformation. This transformation "express"
# how many digits are in the dollar amount.

Amt2 = log10(10*data$Amount+1)

## Error in data$Amount: object of type 'closure' is not subsettable

head(Amt2,n=5)

## Error in head(Amt2, n = 5): object 'Amt2' not found

# fit logistic regression model
lr_model2 <- glm(Vote ~ REP+AMT2, family = binomial, data)

## Error in as.data.frame.default(data, optional = TRUE): cannot coerce class "'function'"
to a data.frame

summary(lr_model2)

## Error in summary(lr_model2): object 'lr_model2' not found
```

## 32 Escalonamento Multidimensional: Introdução

### 32.1 Quinto vídeo da terceira semana

Neste vídeo, estamos vai introduzir os conceitos e as idéias por trás escala multidimensional. Então, primeiro de tudo, o que é escala multidimensional? É um multivariado técnica estatística que leva um conjunto de dissimilaridades ou medidas ou conjunto de pontos e olha para a distância entre esses pontos. O que estamos tentando fazer é obter uma imagem ou representação geométrica desses pontos, que é chamado de mapa espacial. Muitas vezes, também estamos reduzindo o número de dimensões de e a dimensão é normalmente representada repentinamente em uma página. Então, como é multidimensional dimensionamento usado em marketing?

É frequentemente usado em da seguinte maneira. Leva o número na natureza dessas dimensões de um produto ou serviço específico e, em seguida, tenta colocar consumidores nesse espaço. Por exemplo, se você pensar em algumas das características da cerveja, o sabor pode ir do claro ao escuro. Então, algo como um logotipo, até um olheiro muito forte ou algo assim, e então, em outra dimensão, você pode ter o teor alcoólico de um baixo teor alcoólico a um alto teor alcoólico. A partir disso, você pode coloque, obviamente, diferentes marcas de cerveja.

Então, algumas cervejas são muito fortes em sabor e alto teor de álcool ou algumas cervejas podem ser leves e baixas teor alcoólico e você pode descobrir o resto. Há algum continuum lá em uma grade x-y. Então o que você está tentando fazer é descobrir onde as pessoas com base no preferências estão nessa grade. A partir daí, você pode ver quais marcas posicionam de acordo com o gosto do consumidor.

Então, o que me leva a os segundos dois pontos; o posicionamento de as marcas ao longo dessas duas dimensões dizem sabor e teor alcoólico, e também onde é a marca ideal baseada em preferências do consumidor? Esses estudos de atitudes são freqüentemente usado em psicologia, sociologia e pesquisa de mercado, e também agora mais recentemente, eu já vi isso em comportamento economia também. Então, em marketing, foi usado para várias aplicações, incluindo medição de imagem. O que isso significa, compara os clientes e os não clientes percepção da firma, e da empresa percepção consigo mesma.

Portanto, temos uma visão interna da empresa e um visão externa da empresa. Às vezes, estamos interessados em comparar as duas diferenças para ver onde elas podem estar e revelar uma lacuna estratégica que deve ser endereçado.

Outra área em que é usado é avaliar publicidade eficácia. Então os mapas espaciais também pode determinar se uma campanha publicitária foi bem-sucedida ou não. Através da campanha, se muda ou não as preferências do consumidor em relação à sua marca ou, no caso negativo, muda as preferências do consumidor longe da sua marca. Escala multidimensional cria algo chamado mapa espacial. O que está fazendo é tomando todas as dimensões e reduzindo-o a algo que você pode colocar em uma página, e isso ajuda você a entender bem onde as preferências das pessoas mentem.

Então, neste exemplo, um exemplo simples, este é o resultado de uma análise dos padrões de votação. Nesse gráfico, eles são rotulados um pouco genericamente como Coordenada 1 e Coordenada 2, mas você pode pensar em estas são duas questões importantes para o eleitorado. Então, fazendo isso, você pode ver que existem dois grupos que estão acontecendo, vermelho e azul para Republicanos e democratas.

Se você quer um pouco De uma maneira mais concreta de pensar sobre isso, você também pode imaginar que, em vez de padrões de votação, temos preferências de cerveja. Do jeito que eu gosto pense nisso é ao longo do eixo 0 lá. Essa grade é um pouco um pouco fora do centro, mas tudo bem. Vamos veja outra cor. Então, em vez da Coordenada 1, você pode provar, da luz, L-I-G-H-T para escuro, D-A-R-K. Então, cerveja light a cerveja escura e, em vez da coordenada 2, você pode ter algo como porcentagem de álcool.

Então, talvez um baixo percentual de álcool e uma alta porcentagem de cerveja forte, e então você veja onde as pessoas mentem. Então no fundo coordenada esquerda, as pessoas podem preferir cerveja light com baixo teor alcoólico, acima, algumas pessoas podem preferir cerveja light com alto teor de álcool teor de álcool, etc. Então você posicionaria suas marcas particulares. Vou marcá-los como X's aqui. É aqui que suas marcas estão. É assim que você pode identificar seu segmento de mercado e onde as pessoas gostam de cerveja.

Portanto, está usando o marketing para comparar não apenas reduza as preferências do consumidor, mas como seus clientes ver a empresa ou marca e como as pessoas dentro da empresa vê sua marca ou produto? Isso ajuda a identificar talvez diferenças estratégicas. Por exemplo, uma empresa online. Eles podem estar enfrentando alguma pressão de seus clientes sobre dados segurança e privacidade de dados, onde as pessoas dentro da empresa podem sentir que a privacidade dos dados não é tão importante quanto aspectos da empresa, e isso pode mostrar uma desconexão entre esses dois grupos.

Outra área onde O MDS é realmente útil, está tentando avaliar a eficácia de um campanha publicitária. Outra área onde O MDS é realmente útil, está tentando avaliar a eficácia de um campanha publicitária. Então você pode ter conduzido algumas ações multidimensionais processo de dimensionamento para identificar uma lacuna na o espaço do mercado.

Você pode então executar campanha publicitária e, em seguida, faça uma análise posterior para ver se houve alguma mudança na preferências do consumidor. É também, como acabei de mencionar, ideal para identificar lacunas no mapa espacial, e isso pode ser áreas de oportunidade, bem como a compreensão como segmentar o mercado. Então deixe-me voltar a isso deslize com o mapa espacial.

Então aqui você pode ver isso pode haver uma oportunidade. Deixe-me ver. Então diga, por exemplo, apague este, você tem um marca neste espaço aqui, uma marca neste espaço aqui. Você pode ver que esta marca aqui, não há muitas pessoas que querem uma cerveja escura com um alto teor alcoólico, de modo que pode ser algo que um marketing possa quer ficar longe. Você pode ver que existe um grupo aqui que você pode segmentar oferecendo uma cerveja escura com um teor médio de álcool. Então essa é uma oportunidade para entender as lacunas no mercado.

Em termos de campanhas publicitárias, por exemplo, você está fazendo uma publicidade para essa cerveja aqui, isso não é tão popular e você está tentando para que mais pessoas gostem de uma cerveja mais escura com com alto teor alcoólico, você realiza alguma publicidade campanha e espero que você veja mais pontos consumidores, suas preferências começam a mover nessa direção ou agrupar-se essa cerveja e isso representa uma mudança de preferência. Então, antes de mergulharmos escala multidimensional, eu gostaria de falar sobre alguns termos importantes ou os dados usados com frequência. Um é um julgamento de similaridade e essa é a classificação de todos os pares possíveis de uma marca ou algumas marcas de estímulo, serviço, produto, em termos de similaridade com alguma escala Likert.

Então nós conversamos sobre Likert escala antes, e podemos usá-lo em escala multidimensional. Classificação de preferência. Estes são rank encomenda de marcas. Podemos descobrir em um negócios como Pepsi sobre Coca-Cola, ou Coca-Cola sobre Pepsi, ou alguma outra marca de refrigerante. Então também podemos olhar para o estresse, que é um medida de falta de ajuste, valores mais altos de estresse indicam ajuste inadequado. Então, queremos ter certeza de que nosso modelo se encaixa em nosso dados razoavelmente bem, e então essa noção de desdobramento.

Essa é uma representação da marca e o entrevistado, como pontos no mesmo espaço. Então você sobrepõe os dois. R-quadrado, você já viu antes. É o mesmo quadrado R você viu na regressão linear e outras áreas, e é uma medida de qualidade de ajuste que vamos usar. O mapa espacial, Eu lhe mostrei que dois bytes nessa grade no eixo x e y, e se você desenhar pontos médios, obterá uma grade de dois por dois, e é isso que

se sabe como um mapa espacial. Isso ajuda você a entender a percepção relações entre as marcas e outros estímulos, as coordenadas são na verdade as coordenadas ao longo do eixo x, y.

## 33 Procedendo uma análise de escalonamento multidimensional

### 33.1 Sexto vídeo da terceira semana

Então, quando estamos conduzindo uma multidimensional análise de escala, temos que definir o problema, obtenha os dados e selecione nosso procedimento MDS, existem algumas variações disso. Decida o número de dimensões que você vai usar. Normalmente, as pessoas tendem a usar duas dimensões para que cabe em uma página.

Às vezes eles podem usar três dimensões, que também é facilmente compreensível para a maioria das pessoas e você pode visualizá-lo. Quando você chegar a dimensões mais altas, torna-se um problema de interpretabilidade. As pessoas têm muito o que processar, não são realmente capazes de entender dimensões superiores. Então, depois de rotular, talvez uma escala multidimensional de dois por dois ou três por três, você os rotula de alguma forma e, obviamente, avalia a validade do seu modelo em termos de confiabilidade e adequação. Então, vamos falar sobre as etapas individualmente e em um um pouco mais de detalhes. O primeiro passo é formular o problema. Pesquisadores e profissionais de marketing devem realmente entender o objetivo desse esforço, esse processo multidimensional esforço de escala.

Como eles vão usar as informações? Como eles vão selecionar as marcas que eles vão colocar em seu esforço de escala? Quais outros fatores ou estímulos devem ser incluído na análise? Esse é provavelmente um dos aspectos mais desafiadores de uma escala multidimensional e, nesse caso, qualquer esforço de análise, e isso está estruturando o problema e o que você inclui no problema e o que você exclui no problema. Se você optar por ter algum ponto de dados ou variável em sua análise, é o ponto de dados correto? Se você optar por excluir dados da sua análise, como isso limita a interpretabilidade dos seus resultados? Então isso é algo isso requer um entendimento profundo da sua área de assunto, nesse caso, marketing e alguma experiência.

Então, o próximo número de marcas ou estímulos selecionados em sua análise. Isso também afetará seus resultados em interpretabilidade dos resultados. Mas geralmente como regra geral, você gostaria de ter pelo menos oito marcas em sua análise e que ajudarão você a criar um mapa espacial bem definido. Uma marca não é um número estrito, é uma regra geral de dados de entrada. Portanto, os dados de entrada podem ser obtido com os entrevistados em termos preferências e pode ser coletado tanto diretamente ou indiretamente. Assim, por observação ou por pesquisa. Dados de percepção ou abordagem direta, pede-se aos entrevistados que julguem quão similar ou produtos diferentes são.

Você pode usar coisas como uma escala Likert ou algum tipo de classificação. Outra maneira de chegar a dados de percepção é através de uma abordagem indireta onde os entrevistados são solicitados a classificar as marcas com base em atributos identificáveis usando algum tipo de escala diferencial semântica ou novamente uma escala Likert. Os dados de preferência são onde os entrevistados é solicitado que classifique a marca ou o produto do que eles gostam para o que eles não gostam. O próximo passo é selecione um procedimento MDS.

Isso realmente depende da tipo de dados que você está usando, seja ele escalado ou não, seja ordenado ou não, e realmente a natureza dos dados de entrada é o fator decisivo. Uma maneira de pensar sobre isto é, se os dados são dados de escala de intervalo ou proporção ou se é apenas classificar dados ordenados. Nós conversamos sobre isso em um vídeo anterior. Mas uma coisa em que pensar é que, com pedidos dados ou dados ordinais, você pode saber que um cliente prefere a marca A sobre a marca B. Você pode saber que eles preferem a marca B ao invés da marca C.

Mas você realmente não sabe o que isso distancia entre essas preferências, quão forte um cliente prefere a marca A sobre B e B sobre C? Mas com dados de escala de intervalo ou proporção, você pode essa diferença numérica. Agora você tem alguns tipo de medidas. Portanto, se eles classificam a marca A a 10 e marca B e oito, isso é uma diferença de dois. Se eles classificarem a marca B um oito e marca C um, isso é uma diferença de sete. Então você sabe que eles realmente odiavam a marca C e B preferido sobre C, mas eles preferem ligeiramente marca A sobre marca B.

Então é isso que eu quero dizer sobre medidas de distância ou um dados de escala de intervalo. Então existem diferentes Procedimentos MDS para isso. Neste exemplo, estamos só vamos olhar para a métrica clássica versão que usa intervalo ou escala de proporção, mas os conceitos se aplicam dados da escala ordinal como apenas outra técnica de comando. Então, como eu mencionei antes, o objetivo de escala multidimensional é obter um mapa espacial, algo como eu mostrei a você no primeiro conjunto de slides. É melhor ter entrada em um pequeno número de dimensões, dois, talvez três. Mapas espaciais são calculado

para melhorar o ajuste conforme o número de dimensões aumenta. Então, usamos estresse ou falta de ajuste como uma medida do mapa. Quanto maior o estresse ou menor a qualidade do ajuste, o pior do mapa.

Então isso é algo ter em mente. Algumas diretrizes sobre como decidir o número de dimensões. Isso é algo que você decide com antecedência quando executa o algoritmo. Então você pode ter algum experiência prévia com o conhecimento prévio de uma número, dois ou três. A segunda consideração que você quer é interpretabilidade. Se você ultrapassar três, fica realmente difícil entender ou entender sua compreensão os mapas espaciais. Então você também pode usar um Critério de cotovelo ou um gráfico do estresse versus dimensionalidade.

Então você teria junto uma dimensão de um gráfico, você tem a bondade do inverso do medida da qualidade do ajuste, e então você tem a número de dimensões. Você verá algum tipo de quebrar lá. Então esse é outro critério. Mas acho que um é mais um critério técnico do que uma base de conhecimento ou critérios interpretáveis. Então, tenha isso em mente. Depois de desenvolver o mapa, você precisará rotule as dimensões. Nesse primeiro slide, tivemos apenas dimensão 1, dimensão 2, e depois tentei dar um exemplo concreto com alguns características da cerveja, mas é um julgamento ligue especialmente quando você tiver pouco mais de duas características.

Você pode ter uma multidão de características, e então você meio que tem agrupá-los. Isso também requer alguma reflexão e conhecimento em sua área. Isso ocorre porque as dimensões representam mais de um atributo. Então, quando eu dei o exemplo de cerveja, cada dimensão tinha um atributo, mas você pode ter uma coleção de atributos que você está tentando montar, é interpretado com base na posição relativa em suas coordenadas. Para avaliar a validade e confiabilidade de seu esforço de dimensionamento multidimensional, os dados de entrada estão sujeitos a algum ruído aleatório.

Então isso é algo que você precisa ter em mente. Então você quer olhar medidas de qualidade. R-quadrado, por exemplo, é um ótimo para começar. Também obtém especificadores que são indicativos da qualidade do seu escala multidimensional. Como mencionei antes, especialmente quando você está observando os efeitos de um campanha publicitária, você poderia fazer esse esforço diferentes pontos no tempo. Portanto, antes e depois, um teste reteste o esforço que ajudará você a avaliar sua confiabilidade se você não a colocar em um campanha publicitária ou se você estiver olhando para tentar entender os esforços de uma campanha publicitária que você poderia fazê-lo antes e depois. Então esses são alguns coisas para pensar.

## 34 Escalonamento multidimensional: Exemplo no

### 34.1 Sétimo vídeo da terceira semana

Então agora que eu descrevi escala multidimensional, quero mostrar a você o processo e são. Nós vamos olhar para o tradicional escala multidimensional, reduzindo a número de dimensões até um número predefinido. Para usar este comando, o clássico multidimensional comando de dimensionamento, que é denotado `cmdscale()` os parâmetros necessários são `K` ou o número de dimensões e, em seguida, o valor `eig` que mostra os valores Eig.

Se você optar por vê-los. O conjunto de dados que vamos usar está contido em `R`. É chamado `eurodist` e possui uma matriz das distâncias entre cidades europeias. Eu meio que quero conversar sobre isso por um segundo. Então aqui você tem um Atenas a Atenas e, obviamente, a distância entre essas duas cidades são zero. Mas de Atenas a Barcelona são 3.313 quilômetros etc. Então, isso tem todas as distâncias. Se você receber um tabela como esta, você pode imaginar que, se tivesse um lápis, uma régua e uma bússola, poderia de alguma forma, recrie este mapa. Então você pode arbitrariamente escolha um ponto, digamos que é Atenas e dizer que há Atenas, e então Barcelona é 3.313 quilômetros de distância. Então você pode levar uma bússola e faça um círculo em torno disso.

Vamos desenhar um círculo, e isso deve estar no meio. Mas em algum lugar Barcelona está nesse círculo. Então você pode ver Bruxelas é 2.900 quilômetros longe de Atenas, mas também é se você olhe para Barcelona e Bruxelas são 1.300 quilômetros longe de Barcelona. Então, novamente, triangulando você pode descobrir as possibilidades e então com o tempo, preencha o mapa. OK. Observe que isso não necessariamente orienta seu mapa na direção certa, mas você terá pelo menos a relação espacial entre as cidades e um boa ideia disso.

Então você terá alguns tipo de saída, aqui está uma tabela de, vamos chamá-los de longitude e coordenadas de latitude. Então latitude-longitude coordenadas. Eles não são reais longitude e latitude. Eles são apenas números em um escala e onde colocar as coisas. Então é aí que estamos vai inventar. Então vamos fazer isso. Então, para este exemplo, você precisará para se certificar de que você instalou e carregou a biblioteca de estatísticas. Então deixe-me executar esse código. Aqui vamos nós.

Então, vamos obter os dados que você está dist. Nós vamos fazer claro que é uma matriz. Então aí está e você pode ver aqui tem 21 linhas e 21 colunas. Deixe-me olhar para isso dados bem rápido. Então aqui está a tabela Eurodist, e aqui estão as distâncias, Atenas em si é obviamente zero quilômetros. Considerando que Barcelona, Atenas é 3.300 quilômetros mais etc. etc. distâncias entre duas cidades. O comando é realmente simples. Mais uma vez, vamos use esta técnica de colocar os resultados de um função em uma variável. Nesse caso, o nome da variável é MDS. Muitas vezes, se for regressão, você verá algo como o ajuste.

Aqui está o comando de função CAN classic escala multidimensional. Então é isso que `cmds` significa. O conjunto de dados é a distância em euros. `K` é igual a dois, e eu desliguei os valores eig. Estes são os parâmetros necessários e vamos tem duas dimensões. Então vamos correr esses comandos `Data`. Execute a escala multidimensional. Então agora nós executamos e nós pode olhar para esta variável, e você pode ver aqui nós temos essas duas colunas de dados. Então Atenas tem esses coordenadas `x` e `y`.

Pense nisso dessa maneira 2.200 e 1.798 etc. Para colocá-los, vou colocá-los em essas duas variáveis `x` e `y`, `x` `y`. Então a primeira coluna entrará no `X`. A segunda coluna vai entrar no porquê. Ignore este comentário por enquanto, eu vou lhe mostrar o que significa em um segundo. Então, basicamente, o que esses dois linhas 14 e 15 fazem é pegar a coluna um e colocar isso na coluna `x` dois e coloque isso em `y`. `X` e `y`. Agora eu posso traçá-los. Então, deixe-me traçá-los. Lá estão eles. Então, essas são baseadas nesse conjunto de dados, nessa tabela de distâncias. É isso que a saída de a escala multidimensional.

Saiu com esses pontos de dados. Então esses são os parentes distâncias conforme indicado. Posso adicionar a cidade títulos aqui. Deixe-me expandir um pouco pouco, então é mais fácil ver. Aí está. Aqui está a distância entre Lisboa e Madrid, Barcelona e Madrid. Aqui está Marselha, Estocolmo aqui está no fundo.

Então, uma coisa que você deve ter notado se estiver familiarizado com a geografia européia é que Estocolmo é realmente em direção ao norte, e Atenas aqui em cima, que está no topo do mapa, é realmente no sul. Então, para corrigir esse pequeno problema, vou inverter o eixo `y` subtraindo-o de zero e agora, se eu traçar isso, coloque os rótulos. Você pode ver que Atenas fica na parte inferior, Estocolmo está no norte. Eu acho que isso tem uma bonita boa representação das principais cidades da Europa, e acho que se você olhasse um mapa nacional e este mapa, você verá que os pontos alinhar com bastante precisão.

Então, para finalizar, quero observar alguns dos as limitações do exemplo simples que eu te mostrei. Mas isso não quer dizer que escala multidimensional é restrita a esses exemplos bidimensionais. Mas isso não quer dizer que escala multidimensional é restrita a esses exemplos bidimensionais. A primeira coisa que eu quero mencionar é que há esse conjunto de dados aqui que era o mapa das distâncias entre duas cidades. Aqui é realmente ponto a ponto. Você pode imaginar de ter cidades, você poderia ter marcas de refrigerantes ou marcas de batatas fritas ou biscoitos ou o que você quiser. Então pergunte você prefere marca A sobre a marca B.

Marca B sobre a marca C. Então isso seria uma classificação preferência de pedido. Você também pode perguntar para colocá-lo em algum tipo de escala numérica. Se você está tentando obter mais compreensão granular de suas preferências e você não precisa ter apenas essas ordens de classificação. Você pode ter vários dimensões em termos de. Se houvesse algum tipo de biscoito, você pode perguntar sobre doçura, textura, cor, sabores etc. Para que você possa número de dimensões reduzidas para duas dimensões.

Então é aí que o vem a interpretação. Portanto, se ele descobriu que cor, marrom, doçura, doçura alta, uma espécie de biscoito de cor marrom com cobertura são preferidos. Você pode realmente olhar para eles. Marcas de cookies e diga que estes são os biscoitos de chocolate com algum tipo de gelo. Considerando que há outra grupo de cookies que são meio que iluminados sabor, batata frita delicada. Eles podem ser outra classe de cookies que você reconhece e, em seguida, observando preferências do consumidor, você pode identificar pessoas em esses vários grupos. Portanto, esse é um resumo muito breve da escala multidimensional. Mas depois que você entender nesses conceitos básicos, as outras formas de dimensionamento multidimensional são igualmente fáceis de interpretar e executar. Espero que tenham gostado disso.

```

1 library(stats)
2
3 #Obtem o dado e converte pro formato matricial
4 data(eurodist)
5 eurodist <- as.matrix(eurodist)
6 eurodist
7
8 #Usa a funcao cmdscale para converter a matriz na
9 #cl ssica multidimensional escala
10
11 mds <- cmdscale(eurodist, k=2, eig = FALSE)
12 mds
13
14 #muda o eixo y para representar a posicao
15 #relativa dos pontos
16 x <- mds[,1]
17 x
18 y <- mds[,2]
19 y
20
21 #plota as coordenadas reduzidas para obter o mapa espacial
22 plot(x,y,pch=19,xlim=range(x)+c(0,600))
23 text(x,y,pos=4, labels=colnames(eurodist))

```

```

library(stats)

#Obtem o dado e converte pro formato matricial
data(eurodist)
eurodist <- as.matrix(eurodist)
eurodist

##           Athens Barcelona Brussels Calais Cherbourg Cologne Copenhagen
## Athens           0       3313      2963   3175      3339      2762      3276
## Barcelona       3313          0      1318   1326      1294      1498      2218
## Brussels        2963      1318          0     204       583       206       966
## Calais           3175      1326      204        0       460       409      1136
## Cherbourg       3339      1294      583      460        0       785      1545
## Cologne         2762      1498      206      409       785        0       760
## Copenhagen      3276      2218      966     1136      1545       760        0
## Geneva          2610       803      677      747       853      1662      1418
## Gibraltar       4485      1172     2256     2224      2047      2436      3196
## Hamburg         2977      2018      597      714      1115       460       460
## Hook of Holland  3030      1490      172      330       731       269       269
## Lisbon          4532     1305     2084     2052      1827      2290      2971
## Lyons           2753       645      690      739       789       714      1458
## Madrid          3949       636     1558     1550      1347      1764      2498
## Marseilles      2865       521     1011     1059      1101      1035      1778
## Milan           2282     1014      925     1077      1209       911      1537
## Munich          2179     1365      747      977      1160       583      1104
## Paris           3000     1033      285      280       340       465      1176
## Rome            817      1460     1511     1662      1794      1497      2050
## Stockholm       3927     2868     1616     1786      2196      1403       650
## Vienna          1991     1802     1175     1381      1588       937      1455
##
##           Geneva Gibraltar Hamburg Hook of Holland Lisbon Lyons Madrid
## Athens          2610       4485      2977           3030     4532     2753     3949
## Barcelona       803       1172      2018           1490     1305      645      636
## Brussels        677       2256      597           172     2084      690     1558
## Calais           747       2224      714           330     2052      739     1550
## Cherbourg       853       2047     1115           731     1827      789     1347
## Cologne         1662       2436      460           269     2290      714     1764
## Copenhagen      1418       3196      460           269     2971     1458     2498
## Geneva           0         1975     1118           895     1936      158     1439
## Gibraltar       1975          0     2897           2428     676     1817      698
## Hamburg         1118       2897          0           550     2671     1159     2198

```



```
## Hook of Holland      895      2428      550              0  2280  863  1730
## Lisbon               1936      676      2671             2280    0  1178   668
## Lyons                158      1817      1159             863  1178    0  1281
## Madrid              1439      698      2198             1730  668  1281    0
## Marseilles          425      1693      1479             1183  1762  320  1157
## Milan               328      2185      1238             1098  2250  328  1724
## Munich              591      2565      805              851  2507  724  2010
## Paris               513      1971      877              457  1799  471  1273
## Rome                995      2631      1751             1683  2700  1048  2097
## Stockholm           2068      3886      949              1500  3231  2108  3188
## Vienna              1019      2974      1155             1205  2937  1157  2409
##
##      Marseilles Milan Munich Paris Rome Stockholm Vienna
## Athens                2865  2282   2179  3000  817    3927   1991
## Barcelona             521  1014   1365  1033 1460    2868   1802
## Brussels             1011   925    747   285 1511    1616   1175
## Calais               1059  1077    977   280 1662    1786   1381
## Cherbourg           1101  1209   1160   340 1794    2196   1588
## Cologne             1035   911    583   465 1497    1403    937
## Copenhagen          1778  1537   1104  1176 2050     650   1455
## Geneva               425   328    591   513  995    2068   1019
## Gibraltar           1693  2185   2565  1971 2631    3886   2974
## Hamburg             1479  1238    805   877 1751     949   1155
## Hook of Holland      1183  1098    851   457 1683    1500   1205
## Lisbon               1762  2250   2507  1799 2700    3231   2937
## Lyons                320   328    724   471 1048    2108   1157
## Madrid              1157  1724   2010  1273 2097    3188   2409
## Marseilles           0    618   1109   792 1011    2428   1363
## Milan                618    0    331   856  586    2187    898
## Munich              1109   331    0    821  946    1754    428
## Paris                792   856   821    0 1476    1827   1249
## Rome                1011   586   946  1476    0    2707   1209
## Stockholm           2428  2187   1754  1827 2707     0    2105
## Vienna              1363   898   428  1249 1209    2105    0
```

*#Usa a função cmdscale para converter a matriz na  
#clássica multidimensional escala*

```
mds <- cmdscale(eurodist, k=2, eig = FALSE)
mds
```

```
##      [,1]      [,2]
## Athens  2290.274680  1798.80293
## Barcelona -825.382790   546.81148
## Brussels   59.183341 -367.08135
## Calais    -82.845973 -429.91466
## Cherbourg -352.499435 -290.90843
## Cologne   293.689633 -405.31194
## Copenhagen  681.931545 -1108.64478
## Geneva    -9.423364   240.40600
## Gibraltar -2048.449113   642.45854
## Hamburg    561.108970 -773.36929
## Hook of Holland  164.921799 -549.36704
## Lisbon   -1935.040811    49.12514
## Lyons    -226.423236   187.08779
## Madrid   -1423.353697   305.87513
## Marseilles -299.498710   388.80726
## Milan     260.878046   416.67381
## Munich    587.675679    81.18224
## Paris    -156.836257  -211.13911
## Rome      709.413282  1109.36665
```



```
## Stockholm      839.445911 -1836.79055
## Vienna         911.230500  205.93020

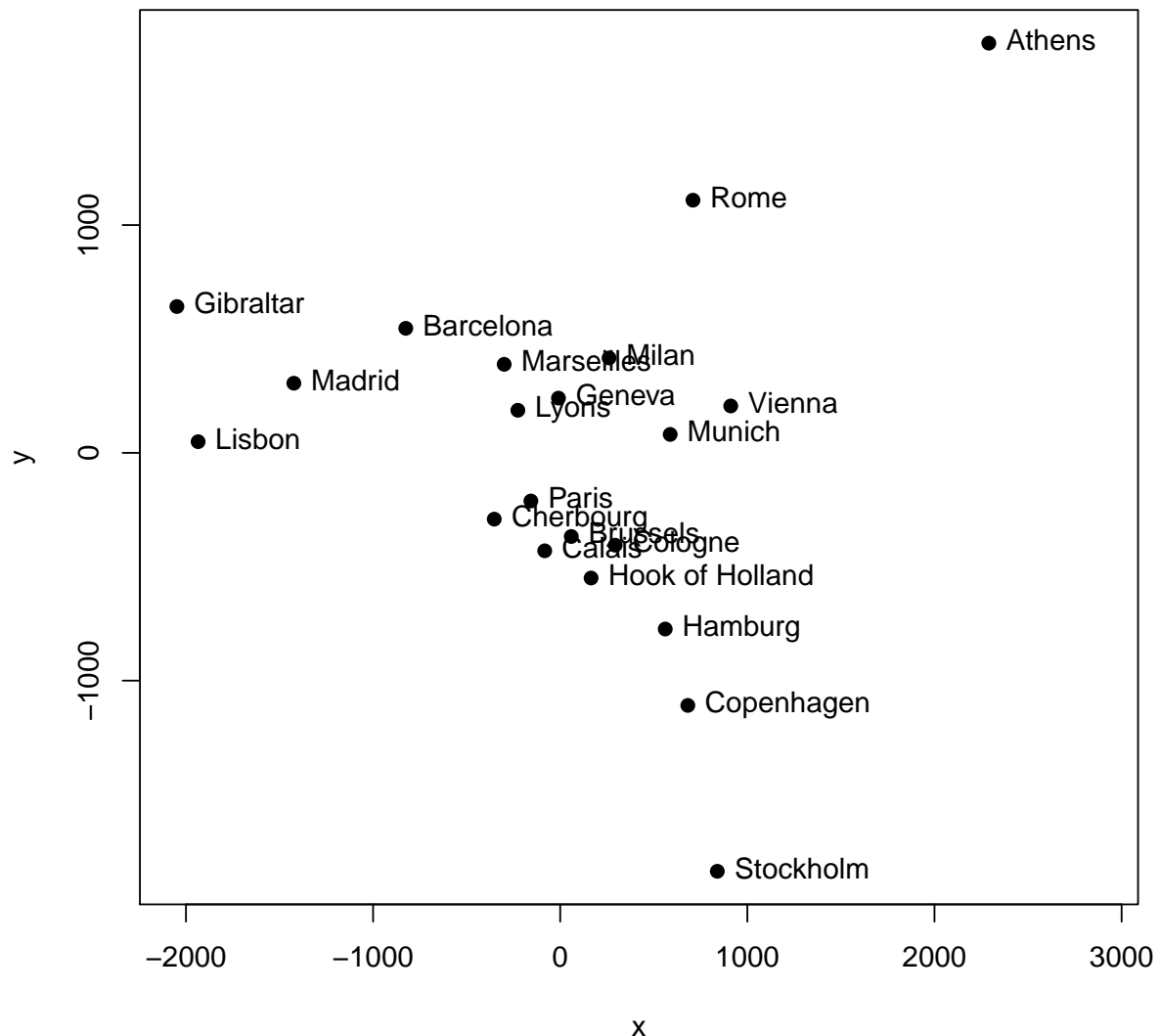
#muda o eixo y para representar a posição
#relativa dos pontos
x <- mds[,1]
x

##           Athens      Barcelona      Brussels      Calais      Cherbourg
##      2290.274680    -825.382790     59.183341    -82.845973    -352.499435
##           Cologne      Copenhagen      Geneva      Gibraltar      Hamburg
##      293.689633     681.931545    -9.423364    -2048.449113     561.108970
## Hook of Holland      Lisbon      Lyons      Madrid      Marseilles
##      164.921799    -1935.040811    -226.423236    -1423.353697    -299.498710
##           Milan      Munich      Paris      Rome      Stockholm
##      260.878046     587.675679    -156.836257     709.413282     839.445911
##           Vienna
##      911.230500

y <- mds[,2]
y

##           Athens      Barcelona      Brussels      Calais      Cherbourg
##      1798.80293     546.81148    -367.08135    -429.91466    -290.90843
##           Cologne      Copenhagen      Geneva      Gibraltar      Hamburg
##     -405.31194    -1108.64478     240.40600     642.45854    -773.36929
## Hook of Holland      Lisbon      Lyons      Madrid      Marseilles
##     -549.36704      49.12514     187.08779     305.87513     388.80726
##           Milan      Munich      Paris      Rome      Stockholm
##      416.67381      81.18224    -211.13911    1109.36665    -1836.79055
##           Vienna
##      205.93020

#plota as coordenadas reduzidas para obter o mapa espacial
plot(x,y,pch=19,xlim=range(x)+c(0,600))
text(x,y,pos=4, labels=colnames(eurodist))
```



## 35 Entrevista com Monica Penagos: Análise conjunta na prática

### 35.1 Primeiro vídeo da quarta semana

Neste último conjunto de vídeos, vamos estar cobrindo uma técnica bem grande.

Nós vamos cobrir os elementos de análise conjunta. Eu tenho comigo de novo Monica que trabalha na Procter & Gamble. Monica, você tem alguma experiência com análise conjunta?

Obrigado por me receber. E em termos de análise conjunta, deixe-me dar um exemplo de como usamos isso na seleção de produtos. E quantos produtos nós quer lançar em um mercado.

Então, em um produto, podemos oferecer uma variedade de aromas. Por exemplo, vamos usar o Febreze. E Febreze podemos decidir quantas aromas com os quais queremos lançar no mercado. Bem, usamos em conjunto para entender quais são os aromas que nossos consumidores provavelmente comprarão? Quais são os aromas que os consumidores menos provável vai gostar de comprar?

E então usamos um conjunto para entender a otimização entre o número de SKUs e o número de perfumes que abordará o maior público para isso. Uau, então você Número de produtos ou Use o SK e tente também descobrir qual sensor é popular ou não popular descobrir a demanda e otimizar essas duas variáveis usando análise conjunta. E isso está no espaço do produto.

Você tem outros exemplos definitivamente eles anunciando espaço na publicidade na mídia, por exemplo, no Facebook quando você estiver criando um anúncio no Facebook e indo para o Facebook gerente de

negócios, você tem uma combinação de ofertas de vídeo ou imagens. Você pode colocar suas reivindicações, você pode colocar seu preço, você pode colocar um botão de ação e analisamos quando criamos anúncios diferentes usamos análise conjunta para entender.

Qual é o ideal dessas combinações em uma publicidade, porque isso otimizará nossa publicidade, nossas taxas de cliques e diminuímos nosso custo por aquisição. Ok, isso é muito o que acontece que no espaço de marketing digital que você olha oferece nomes de nomes, você olha para o tipo de vídeo ou se tem imagens e depois usa análise conjunta para descobrir qual é a melhor história em quadros conjuntos de combinações. Isso é ótimo. É um grande tópico sobre análise e espero que você goste do próximo conjunto de vídeos.

## 36 Leituras da semana

### 36.1 Leitura semanal (última semana)

- [R for marketing students](#)
- [R pubs análise conjunta](#)
- [Pacote conjoint](#)
- [Apresentação do material](#)

## 37 Análise Conjunta: Introdução

### 37.1 Segundo vídeo da quarta semana

Oi. No próximo conjunto de vídeos, discutiremos uma abordagem comumente usada e relativamente técnica sofisticada usada frequentemente em marketing chamado análise conjunta.

Então, neste vídeo, falarei sobre o que é análise conjunta e, posteriormente, vídeos, falarei sobre métodos de coleta de dados, utilitários que valem a pena, selecionando atributos, como realmente executar a análise em R e como interpretar os resultados. Embora esse termo possa parecer um pouco complicado a princípio, o que isso significa? Na verdade, é muito simples técnica e utiliza algumas ferramentas básicas que você já tem em seu kit de ferramentas, especificamente, regressão linear.

Então vamos começar. Então, primeiro, eu quero descrever o que é análise conjunta.

Em primeiro lugar, é um estudo baseado em pesquisas técnicas estatísticas e é frequentemente usado em marketing, pesquisa ou análise. Então a fonte primária de dados são pesquisas. Você vai sair para o campo e pergunte às pessoas, você e sua amostra, um monte de perguntas e, a partir disso, você tentará identificar suas preferências. Essa técnica ajuda a encontrar a combinação ideal de recursos em um produto ou serviço.

Então, por exemplo, então você tem dois tênis, um é bem alto preço e alta qualidade, um é baixo preço e talvez de qualidade inferior e, em seguida, um sapato tenha preço médio, mas é realmente bom para correr, por exemplo. Você pegaria essas combinações de sapatos e tentaria descobrir qual atributo é o mais importante. É a qualidade do sapato, é o preço do sapato ou é para isso que o sapato é projetado para correr ou alguma outra atividade, e você tenta quebrar os pesos que as pessoas têm em suas cabeças. Qual é o valor relativo de preço versus qualidade, preço versus o tipo de uso do sapato, a qualidade versus o tipo de uso do sapato? Tente quebrar todos aqueles pequenos componentes para identificar qual é o sapato mais popular.

Então, algumas coisas importantes a saber é que a análise conjunta é feita de fatores e níveis, e os fatores são os variáveis que afetam a compra e os níveis são os valores atribuídos para cada fator. Então, no exemplo do tênis que acabei de discutir, pode haver um fator, o preço seria um fator, por exemplo, e quais são os vários níveis? Você pode ter um alto preço, preço médio, preço baixo, e você pode até ter um pouco mais de dados em nível granular.

Então, no conjunto de análise, tenta medir os efeitos dos níveis desses fatores, misturando-os. Na análise conjunta, você descreve os recursos que você está interessado e é significativo aos entrevistados e, em seguida, peça que eles classifiquem a importância relativa dessas combinações.

Portanto, é uma técnica baseada em pesquisa. Você vai subir com seus recursos, faça uma pesquisa e depois faça-lhes as perguntas. Pode ser considerado, especialmente na forma que estamos fazendo, uma extensão de múltipla análise de regressão e em nossa análise conjunta estaremos usando vários tipos de regressão como o mecanismo. Mas observe que a análise conjunta não está realmente ligada a um método estatístico específico.

É um problema geral conjunto de tentar obter uma compreensão de como as pessoas tomam decisões com base em diferentes atributos de um produto.

Nesta aula, discutiremos a análise conjunta usando regressão múltipla, mas observe que existem outras técnicas por aí que são um pouco mais sofisticado.

Saiba também que isso é uma classe de linha de base, então quando você entender esses conceitos, passando para o avançado conceitos mais tarde em sua carreira deve seja uma transição suave.

Então, usando o múltiplo análise de regressão, tentamos encontrar as melhores pesos ou combinações das variáveis a produzir o resultado desejado, que estamos procurando. Subjacente a tudo isso é uma premissa básica de que os compradores veem os produtos como compostos de vários fatores e eles estão pesando os fatores e os níveis em fazer a sua decisão de compra. Tenho certeza que se você refletir sobre sua própria decisão de compra, especialmente se você pensa em uma grande compra em que realmente pensar sobre isso, como comprar um carro ou uma bicicleta nova ou algo assim, você pesará toda a recursos e tente descobrir qual combinação é a que você deseja no preço certo.

Então deixe-me antes de mergulhar falar sobre o vocabulário, as definições que nós use na análise conjunta.

Primeiro é atributo e fator. Estes são os subjetivos avaliações de um produto. Eles podem ser coisas como preço, peso, cor, qualidade, uso, coisas assim. Então o nível está dentro o fator e é assim que você mede cada uma delas atributos ou fatores. Alto ou baixo, é o preço alto ou baixo? O peso é pesado ou leve? Cor, é vermelho ou azul?

Uso, é para correr ou caminhadas, coisas assim? O próximo termo que eu gostaria de descrever é conhecido como utilitários com valor parcial. Essa é a extensão em que um fator contribui para a toda a utilidade do produto.

Então, quando compro, por exemplo, um carro novo, valorizo o conforto, valorizo o manuseio, os recursos de segurança e o preço, entre outros diferentes tipos de atributos. Então, o que a análise conjunta faz é tentar descobrir quanto pesa o preço na minha tomada de decisão processo ou quanto pesa a segurança no meu processo de tomada de decisão?

Existe uma noção conhecida como pares proibidos e esses são dois fatores que devem nunca aparecem juntos. Por exemplo, uma polegada de seis polegadas A Apple telefona por US\$ 100. Portanto, estes estão "fora do domínio normal possibilidades do dia-a-dia". Isso não é algo que você seria considerar realisticamente. Comprando uma Apple de seis polegadas telefone por US\$ 100. O custo supera em muito o preço, de modo que realmente nunca acontecer na vida real.

Então isso termina a introdução à análise conjunta e vamos mergulhar na próxima conjunto de vídeos.

## 38 Análise Conjunta: Métodos de coleta de dados

### 38.1 Terceiro vídeo da quarta semana

Neste vídeo, eu vou para falar um pouco sobre várias coletas de dados metodologias.

Observe que, embora existam diferentes metodologias, na prática, você pode usar vários métodos ao realizar uma análise conjunta ou você pode usar um abordagem híbrida que você acha que pode ser um valor à sua empresa ou firma.

Mas vamos falar sobre o métodos diferentes clássicos. Os métodos que eu vou estar falando é a abordagem de perfil completo, e é a que eu vou usar nos meus exemplos nesta aula. Então, aqui estão alguns outros métodos. Há uma troca método matricial, método de comparação pareada, método de autoexplicação, um método adaptativo e um método híbrido, que leva parte do características de cada um deles.

Então vamos passar eles em detalhes. O primeiro método que eu gostaria de falar é o método de perfil completo. Um perfil é uma mistura dos atributos em diferentes níveis. Por exemplo, alta preço, alta qualidade, alto preço, média qualidade, alto preço, baixa qualidade, baixo preço, alta qualidade, preço baixo de qualidade média.

Então você faz todas as combinações. Então você tem algum produto completo. Então você pergunta aos entrevistados para avaliar esses produtos. A avaliação dessas respostas produzirá muitos dados. Então você terá muitas informações sobre cada um dos produtos. Uma vantagem, é fácil para o entrevistado visualizar o especialmente se você tiver protótipos que eles podem olhar ou fotos.

Então é fácil para eles avaliar o produto e para avaliação porque todos os atributos estão incluídos nesse protótipo, na imagem, no sabor teste, o que você tem. Outro método é o que é conhecida como matriz de trade-off.

Os entrevistados são solicitados a avaliar os conceitos de produtos em combinações de dois atributos por vez. Então, um problema com a abordagem de perfil completo é que você poderia ter um gazilhão combinações e se tornaria muito cansativo para um entrevistado, você gosta do carro vermelho com tração nas quatro rodas, você gosta do carro azul com tração nas quatro rodas com freios anti-trava?

Então, existem todas essas combinações e é quase impossível perguntar eles para avaliar tudo isso. Então, esses são os métodos de tentar descobrir maneiras de fazer isso em uma maneira mais parcimoniosa.

Então poderíamos pedir combinações envolvendo dois atributos de cada vez e, em seguida, todo o possível par de atributos no estudo são eventualmente avaliados pelo respondente. Uma dificuldade é que é difícil agregar os resultados de diferentes entrevistados como as suposições feitas por cada entrevistado podem ser diferentes sobre os atributos que não são especificados, pois você está apenas procurando às duas de cada vez.

Portanto, existem alguns problemas com o método da matriz de trade-off. Outro método é conhecido como o método de comparação emparelhado. Nesse caso, o Solicita-se aos entrevistados que forneçam uma preferência entre dois perfis. Significado, duas versões do seu produto. Versões completas completas. A vantagem de esse método é que o entrevistado é apenas concentrando-se em duas coisas ao mesmo tempo e, portanto, as avaliações podem ser mais significativas simplesmente porque o entrevistado pode olhar para os dois produtos lado a lado e tomar uma decisão.

Prefiro esta versão do produto a esta versão do produto. Uma desvantagem é que o número de pares pode ser muito grande. Portanto, isso é semelhante ao problema que tivemos antes. Você gosta de produto A ou produto B? Você gosta de produto A ou produto C? Você gosta de produto B ou produto C? Então, todas essas combinações se tornam muitas trabalhar para administrar.

Lembre o os entrevistados podem se cansar depois de um tempo ou eles podem não querer responder a uma longa sessão de perguntas como essa. Então você quer ser consciente disso. Para quebrá-lo, outra abordagem sugerida é conhecida como o método de auto-explicação. Aqui, oferece uma simples método que não requer o desenvolvimento de conceitos completos do produto ou perfis completos.

Especificamente, pedimos sobre a preferência para cada nível de recurso em vez da preferência por um pacote de recursos. Então, em vez de perguntar, você quer o carro vermelho, a um preço alto, com freios anti-trava? Você pergunta, você gosta de vermelho ou azul? Você gosta de um alto preço ou preço baixo? Você gosta de segurança recursos ou não? Então, separando o perguntas por recursos, é que você combina isso. Fornece aproximadamente abordagem comparável com uma abordagem de perfil completo. É apenas o, eu não quer dizer padrão ouro, mas é o ouro padrão, mas não é realista perguntar que muitas perguntas.

O próximo método é o análise conjunta adaptativa. É aí que variamos os conjuntos de opções com base em suas preferências. Portanto, torna um pouco mais eficiente. Então, como eles estão indo para baixo e eles fazem algumas escolhas, então pedimos a eles Perguntas relacionadas. Então não perguntamos perguntas e níveis com pouco ou nenhum apelo. Então, se você sabe que eles como produtos de qualidade, e a qualidade é importante para essa pessoa, continuamos nesse caminho de pedir variações de qualidade com outros atributos.

Nós não nos incomodamos fazendo perguntas de baixa qualidade porque eles apenas não estão interessados. Este método obviamente reduz a duração da pesquisa, o tempo que leva para administrar a pesquisa. Cada pacote é apresentado para avaliação. Então a conta da pesquisa para escolha e, em seguida, faz a próxima pergunta mais eficiente. Por isso, nos adaptamos às respostas do entrevistado para agilizar o tempo da pesquisa mais eficiente.

Finalmente, aqui está um método híbrido. Foi realmente desenvolvido como você pode ver com o problema de lidar com um grande número de atributos em um estudo conjunto. Que você poderia ter, nos exemplos que eu vou mostrar na aula, teremos apenas dois ou três atributos que são ótimos exemplos para mexer com sua cabeça.

Mas, na prática, um produto terá muitos, muitos, muitos atributos que você pode querer olhar. Então, se você pensa em um produto alimentar, eles olham para a nutrição conteúdo, o teor de açúcar, o número de calorias, o sabor, a cor do pacote, eles são embalados individualmente? Eles estão embrulhados em pacotes? Nós podemos fazer uma lista e continuar e continuar e continuar. Então isso se torna bastante problema complexo para lidar.

Então esse método, o método híbrido, ajuda a resolver esse problema. É dividido em fases. Então, na primeira fase, o entrevistado é solicitado a fornecer dados sobre o que eles acham importante e a importância relativa desses atributos. Então, se eu estou pensando em voltar para o exemplo de salgadinhos, eu quero alta fibra, Quero saudável, quero pouco açúcar, esse pode ser o meu conjunto de preferências. Ou outro conjunto pode ser, eu quero algo é divertido de comer, mais fácil de transportar, tem gosto bom, algo assim.

Então, dados aqueles conjunto de atributos, na fase 2 pedimos a eles uma número limitado de perfis para avaliar em vez de administrando todos os perfis. Então, como eu disse a você que quero talvez alimentos com pouco açúcar, saudáveis e fáceis de transportar, vou fazer perguntas sobre produtos que se encaixam nessas categorias.

Em seguida, os perfis apresentados são desenhados usando um desenho ortogonal. Falaremos sobre isso em vídeos posteriores. A pesquisa não passa de uma escala de classificação, algo assim.

Então, aqui estão as propostas sobre os diferentes produtos você pode oferecer. Aqui, este é um exemplo

de balsa que eu vou usar para esse conjunto de vídeos. Então aqui você pode ver o a tarifa da balsa é a preços diferentes, US\$ 3,10, US\$ 3,70 e US\$ 4,30. Então, também o duração do ferry. Então é uma balsa de 10 minutos? Uma balsa de 20 minutos? Ou eu acho que só tenho esses dois, um ferry de 20 minutos, Ferry de 10 minutos.

Então, essas são as coisas que você pode querer considerar. Por exemplo, se a duração do ferry é rápido, sua empresa pode precisa atualizar seus barcos de balsa com motores maiores, talvez seja necessário aumento do uso e combustível. Portanto, há um custo associado à qualidade desse passeio de balsa ou à duração de o passeio de balsa. Então a tarifa também se torna importante porque o cliente pode ser sensível ao preço. Eles estão dispostos a negociar 10 minutos em sua viagem de balsa por um pouco de poupança, por exemplo? Se assim for, por quanto? Esses são os tipos de perguntas que estamos tentando responder usando a análise conjunta.

## 39 Tipos de Análise Conjunta

### 39.1 Quarto vídeo da quarta semana

Olá, neste vídeo, gostaria brevemente de fale sobre algumas classes de análise conjunta e alguns dos métodos que você pode estar usando para executar esta análise.

O primeiro é tradicional análise conjunta. Esse será o tipo que eu serei focando nesses tipos de vídeos, mas achei que você deveria esteja ciente dos outros. Há análise conjunta tradicional, usuário indicar as preferências por pesquisas. Você pode pensar nisso como um problema de regressão múltipla e você verá isso mais tarde quando eu mostro o exemplo. E as classificações do produto são observações e essas são da variável dependente, e isso é aquele em que vamos nos concentrar na aula.

Há uma escolha baseada em que os entrevistados escolherão seu perfil mais preferido, como combinação de atributos que eles mais gostam. E a vantagem deste é que pensa-se simular um ambiente de compras real. E a última aqui que eu meio que quero mencionar é o adaptável. Você pergunta primeiro quais são os seus preferências no conjunto de opções e, dadas as preferências, você mostra perfis que correspondem a essa preferência.

Uma coisa a notar é que aqui nesta classe, usaremos a versão baseada em regressão. Mas à medida que você continua em sua carreira, outras versões, outras técnicas estatísticas subjacentes a esses objetivos são usados. Então existem coisas como logit multinomial, pode haver algumas técnicas bayesianas usado para capturar processos de aprendizagem.

Então, o que eu quero que você ande afastado é que não é uma técnica estatística, é uma análise conjunta objetiva e existem várias maneiras de chegar lá.

Portanto, avaliando o tipo de análise e garantir que você tenha a técnica estatística correta é algo que só virá com você através da experiência e também através aprendendo essas novas técnicas. Mas primeiro, vamos ao básico conceitos para baixo para você começar.

## 40 Análise Conjunta: Utilidade com valor parcial

### 40.1 Quinto vídeo da quarta semana

Neste vídeo, eu gostaria de descrever uma componente da análise conjunta, e esse é o utilitário de valor parcial.

Então, o que estamos tentando fazer na análise conjunta? Quais são os recursos de análise conjunta? É uma técnica de medição para quantificar o preferência do comprador, é uma técnica para prever o que um novo comprador fará diante com um novo produto, é também uma segmentação técnica para identificar grupos de compradores que compartilham preferências semelhantes, é uma técnica de simulação; portanto, quando criamos o modelo podemos avaliar novas variações de produtos em um ambiente competitivo, também é uma otimização técnica para procurar perfis ou produtos isso pode maximizar o retorno.

Então, sabemos o custo subjacente desses vários produtos, para que possamos descobrir qual obterá mais lucro e talvez não seja aquele que é o preferido pelo cliente. Então, o que é um utilitário que vale a pena? Isso é realmente o cerne da análise conjunta está tentando fazer.

Valor da peça significa utilidade de nível para todos os diferentes atributos. Então você pode ter um preço alto, baixo preço e alta qualidade e baixa qualidade, e aqueles que são diferentes atribui preço e qualidade, mas você pode querer para descobrir qual é o peso relativo de cada um. Em termos de minha utilidade total, quanto valor recebo do produto? Quão importante é apenas o preço ou quão importante é essa qualidade? Se tivermos outros atributos, qual a importância da cor? Quão importante é o ajuste?

Quando você combina tudo esses fatores juntos, o vocabulário usado é um perfil de produto e o valor do utilitário que separa os atributos do perfil é o valor da peça.

O valor da peça em nível permite mergulhar mais fundo entender quais são os recursos específicos que orientam a escolha do cliente. Você pode achar que, talvez no desenvolvimento de produtos há algo que você pensou que seria seja uma ótima idéia, são novos sinos e assobios, e quando você realmente pergunte ao cliente, eles podem não estar interessados nele de qualquer maneira.

Então, esses são os tipos de coisas que você está tentando entender. Então, vejamos uma equação e, novamente, vamos assumir um modelo linear para o utilitário de valor parcial para um dado fator. Então, vamos supor que nós tem três fatores;  $x$ ,  $y$  e  $z$  com três níveis. Então aqui está o nível 1, aqui está o nível 2. Falaremos sobre isso mais tarde, mas você não precisa especificar o terceiro nível, e aqui está o valor da parte utilitário para os  $y$ s e aqui o valor da parte utilitários para os  $z$ s.

Então você tem os  $x$ s, os  $y$ s e os  $z$ s, e é isso que isso está tentando fazer. Então, quando você corre sua regressão você coloca todos eles em uma equação, e então você pode descobrir cada um dos diferentes fatores deste modelo linear. Então, para o fator  $x$ , é esta parte do modelo total, para o fator  $y$  é esta parte e para o fator  $z$  é esta parte aqui. Portanto, isso corresponde a isso, pois a parte  $y$  corresponde com isso, e a porção  $z$  combina com isso.

## 41 Conjoint Analysis Example – Ferry Fares

### 41.1 Quinto vídeo da quarta semana

Até agora, eu introduzi muitos conceitos para você; atributos, perfis, utilitários com valor parcial etc., e tenho certeza que está começando a se transformar em um grande nuvem em sua cabeça.

Então, vamos tentar limpar o nevoeiro usando um exemplo concreto. Vou discutir esse caso inventado de tarifas de balsa. Então, vamos considerar uma tarifa de ferry perto de uma cidade metropolitana.

Por exemplo, Hong Kong tem muitos ferries que vão e voltam entre as diferentes ilhas. Nós vamos usar dois fatores para decidir tarifas de ferry; o tempo de viagem entre dois pontos, A e B, e o valor real custo da tarifa do bilhete. Nós vamos perguntar usuários em potencial para classificar em uma escala de 0 a 100, sendo 100 o melhor e zero sendo o pior, como eles se sentem combinações diferentes.

Então aqui está a tarifa; US\$ 3,10, US\$ 3,70, US\$ 4,30 e duração da viagem; 10 minutos e 20 minutos. Agora, é fácil imaginar podemos fazer com que esse gráfico se expanda para a direita, tendo diferentes níveis de tempo e diferentes níveis de preços, mas vamos ficar com esses três básicos por enquanto.

Portanto, temos pela tarifa, um preço baixo, um preço médio preço e preço alto, e durante a viagem, 10 minutos e 20 minutos ou rápido e lento. A tarifa ajuda você a entender as preferências do cliente e as compensações que eles estão dispostos a fazer em termos de duração de tempo que leva.

Então, se estudarmos esta tabela com um pouco mais de cuidado, podemos fazer algumas observações com base na pesquisa. Esse entrevistado em particular, obviamente, gosta a tarifa mais barata. Esta, a tarifa mais barata no menor tempo, eles gostaram que a melhor e eles avaliaram que 100. O pior é a tarifa mais cara e o tempo mais longo. Então, tudo em entre nós alguma mistura. Um interessante observação, se você olhar, são esses dois números aqui, 92 e 94, eles são quase idêntico.

Então esse particular o entrevistado está disposto a doar \$ 0,60 por um 10 minutos extras de tempo. Portanto, a diferença de preço e a diferença de tarifa é de US \$ 0,60 entre esses dois números e a diferença de o tempo é de 10 minutos. Então, se eles estão montando tarifa barata por 20 minutos, eles alcançam sua escala pessoal, classificam-na como 94, mas estão quase dispostos a pagar US \$ 3,70 por um período mais curto, e estes são quase equivalentes.

Então é aí que esses trade-offs começam a ocorrer. Quanto mais dados você coletar, mais você terá uma compreensão do mercado. Então, como codificamos esta tabela para um modelo de regressão? Temos que configurar variáveis fictícias. Então, as variáveis fictícias aqui são como segue; temos  $F_1$ , e se a tarifa for \$ 3,10 e zero caso contrário. Então, o que isso significa é  $F_1$  será igual a um ou zero e depende da tarifa. Se a tarifa for de US \$ 3,10, então é um sim. Caso contrário, é um não.

Portanto, a tarifa um é de US \$ 3,10, a tarifa dois é, da mesma forma, de US \$ 3,70. Então, se a tarifa é US \$ 3,70, sim ou não. OK. Então,  $D_1$  é o tempo de viagem, se for 10 minutos e zero caso contrário.

A codificação fictícia apenas define o utilitários com valor parcial dentro dos atributos a serem estimados e observe que não temos configurar especificamente uma variável fictícia para a última categoria desde que está implícito. Então, no lado esquerdo desta tabela, podemos ver o original dados e a classificação.

Portanto, essas são apenas as tarifas listadas e as durações em cada um dos as combinações. Então ainda há seis caixas; um dois três quatro cinco, seis, para as classificações. A tarifa mais barata e a balsa mais rápida a viagem ainda é 100, e a tarifa mais cara e a viagem mais lenta de balsa são 60, e todas são preenchidas.



Para codificar para a regressão, aqui estão os três variáveis; F\_1, F\_2, F\_3. Lembre-se de F\_1 é se for uma feira de US \$ 3,10. Portanto, observe que estes são US \$ 3,10 e estes dizem que sim. F\_2 é para a tarifa média de US \$ 3,70 e diz sim. Ai está. Então, não precisamos para codificar os US\$ 4,30, uma vez que isso está implícito. Então, finalmente, para o D\_1, a duração, vamos definir o código D\_1 para um sim ou um se a duração for 10. Então 10, sim, é esse.

Dez, sim, é esse. Dez, sim, e é esse. Então, acho que é isso. Sim, é tudo. Então agora, temos todas as combinações em nossos dados codificadas como variáveis fictícias.

Eu gostaria algum tempo, talvez pausando o vídeo e encarando isso para garantir que você entendesse o que está acontecendo. Então aqui está a equação de regressão. Aqui ainda podemos ver nossas variáveis fictícias. Tem aqui; F\_1, F\_2, D\_1, e vamos estimar o modelo de preferências com base em uma interceptação termo, Y interceptar termo.

Então, aqui estão as variáveis fictícias e estamos estimando este X\_1, X\_2 e X\_3. Essas são as variáveis estamos estimando. Y é a combinação total da classificação do indivíduo. X\_1, X\_2, X\_3 são os coeficientes que nos estamos procurando por. Portanto, considerando os dados, aqui está o nosso resultado de regressão.

Podemos ver a interceptação O termo é 61 aqui. Todos os valores-P parecem bons. Eles são todos inferiores a 0,05. O coeficiente para X\_1 é 31,5, para X\_2 é 21,5 e para X\_3 é nove. Então, aqui estão as variáveis fictícias e estamos estimando este X\_1, X\_2 e X\_3. Essas são as variáveis estamos estimando. Y é a combinação total da classificação do indivíduo. X\_1, X\_2, X\_3 são os coeficientes que nos estamos procurando por.

Portanto, considerando os dados, aqui está o nosso resultado de regressão. Podemos ver a interceptação O termo é 61 aqui. Todos os valores-P parecem bons. Eles são todos inferiores a 0,05. O coeficiente para X\_1 é 31,5, para X\_2 é 21,5 e para X\_3 é nove. Então, aqui estão os erros padrão. Podemos ver que a Praça R parece muito boa.

Na realidade, parece muito bom. Eu acho que isso é apenas devido ao número de observações que temos. Aqui está o modelo. Então Y, é a utilidade total, é igual a 61, isso é nosso termo de interceptação, mais 31,5 para F\_1, e F\_1 é se é uma tarifa barata ou não. Então, se é barato, sim, aos 31 anos. Se é mais caro tarifa, talvez uma tarifa média, então você adiciona 21,5 ao 61 e depois a duração de a viagem lhe dará, se for de curta duração, outros nove pontos.

Então vamos dar uma olhada através de um exemplo. Lembre-se de que, na tabela, a viagem mais lenta de 20 minutos e uma tarifa de US\$ 4,30 foi classificado como 60. Certo? Então, US\$ 4,30 e lento foram de 60. Nós os codificamos como zeros.

Portanto, se você olhar o modelo aqui, F\_1 é zero, F\_2 é zero e D\_1 é zero. Direita? Porque isso é a tarifa mais alta, não F\_3, e é lenta. Então, isso foi codificado como D\_1 é zero. Então, se você olhar isso, estes três termos aqui são zero e nós tem modelo interno, nossa estimativa é Y tem um valor de 61, que é bem próximo de o valor real de 60. Se olharmos, vamos desacelerar novamente e faremos uma tarifa, que é a tarifa barata.


Temos 61 mais 31,5 e isso seria 92,5. O valor real para barato tarifa e devagar foi de 94. Então isso é bem próximo também. Encorajo-vos a passar por este modelo aqui, esta equação Y, e calcular o utilitários diferentes e compará-los aos dados reais que tivemos em nossa pesquisa.

O próximo passo é entender quais são os utilitários com valor parcial. Nós vamos decompô-lo e podemos dividir em dois funções para a tarifa. O valor da parte do utilitário para a tarifa, U de F, é 31,5 para a tarifa barata mais 21,5 para a tarifa cara. Você pode ver isso por duração depende se é uma curta duração ou de longa duração e você obtém esse coeficiente de nove com base nisso.

Portanto, essas são as funções de utilidade parcial ou parcial. É assim que você pode separar as preferências do cliente com base no custo da viagem versus a duração do passeio.

## 42 Análise Conjunta no

### 42.1 Sexto vídeo da quarta semana

Neste vídeo, eu gostaria de percorrer um exemplo de como realmente executar o análise conjunta em R. Então, vamos olhar para algum código R, a fim de realizar uma análise conjunta. É um trava-língua. Ok, vamos começar. O pacote que vamos o uso é chamado de conjunto. Então você vai precisa instalar esse pacote no seu Ambiente  Studio.

É realmente simples pacote que é usado para tradicional análise conjunta. Então, dentro do pacote, aqui estão algumas dicas úteis funções e as funções principais que estaremos usando.


Você pode ver que começa com muita CA, e há o modelo, os utilitários, o utilitários parciais. As funções de utilitários parciais é o que descrevi como parte do valor nos slides. Também é conhecido às vezes como



utilidades parciais, o planejamento fatorial, e o design codificado. Então, por que não basta pular direto. Uma das primeiras coisas que tem que fazer é descobrir que tipo de design nós vamos usar.

Como eu mencionei no vídeo sobre atributos, um fatorial completo design é aquele em que todas as respostas e todos os diferentes níveis são contabilizados e temos todo o combinações disponíveis, e é isso que é conhecido como um planejamento fatorial completo. Na prática, isso é realmente não é realista porque você pode ter centenas, senão milhares de combinações diferentes.

Então você pode querer para fazer algum tipo de design fracionário ou um desenho ortogonal. Esse processo seleciona os atributos limitando a quantidade informações sobre perdas, minimizando a dependências entre os fatores e vimos que com o desenho ortogonal.

Como você verá,  executará esse processo para você. Portanto, neste exemplo, é um exemplo hipotético. Vamos criar uma pesquisa para apreciar a probabilidade de lendo um romance, certo? Qual é a sua propensão para escolher ou comprar um romance em particular?


Então, nós vamos ter alguns fatores envolvidos. Nós vamos escolher três fatores, páginas, o gênero e o autor. Para o número de páginas, teremos apenas três níveis amplos ou três categorias amplas; menos de 500 páginas, entre 500 e 1000 páginas ou mais de 1000 páginas.

Então, outra maneira de pense que este é o potencial comprador deste livro, eles gostam de livros curtos? Eles gostam de comprimento médio livros ou livros longos? Agora, pode depender eles gostam de histórias completas da saga ou podem ser apenas uma questão de peso, o peso de carregar o livro por aí? Eles gostam de ficção ou não ficção? Sim, você pode ter mais categorias, mas, para o nosso exemplo, vamos continuar com esses dois.


Mas se você quiser, pode ter mistério, culinária, hobbies, etc. Então, para o autor, acabamos de nomear autores e autores anônimos. Nós vamos usar isso projeto fatorial que nos ajuda a olhar para diferentes maneiras de nivelar os atributos.

Nesse caso, estamos vai criar um planejamento fatorial completo e será uma matriz 12 por 3 com todos as combinações possíveis. Então, o que o CA codificou função de design faz em R é que temos no lado esquerdo da tabela das categorias, o número de páginas; curta, média, longa, ficção ou não ficção tipo de autor, conhecido, desconhecido. Então ele apenas codifica tem números 1,2,3,1,3 para as diferentes categorias então você está olhando.

Isso permite que o computador para lidar com as coisas com mais facilidade para análise. Uma coisa a ter em mente é que quando você está olhando para o resultado ou algo assim, você só vai ver esses atributos codificados. Então, você pode se lembrar ou lembrar o que o tradução é. OK.


Então, respostas. Nós conversamos sobre isso antes de um pouco, mas vou revisar novamente. As respostas geralmente são respostas a uma pesquisa questionário, que pode ser obtido em muitos formatos. Podemos fazer rankings ou classificações dos termos. Podemos olhar em termos de qual é a probabilidade de você está comprando este produto? Você gosta disso produto ou você está inclinado a recomendar este produto a um amigo? Então, tudo isso chega a essa noção de você gosta deste produto? Se sim, quanto você gosta? Para o nosso estudo, criaremos nossas próprias respostas aleatoriamente usando uma função de amostra em .

As respostas serão em termos de classificações distintas de todas as 12 combinações e devido à randomização, você deve observar que os resultados podem não refletir os resultados reais da vida real. Então aqui, eu mostrei a você o código em um minuto, mas os perfis são as diferentes combinações do produto. Então o perfil um seria que primeira linha que o desenho fatorial que foi menos de 500 páginas, ficção e autor desconhecido. Então o perfil dois seria a segunda combinação de atributos.

Perfil três é o próximo combinação de atributos. Então pedimos 10 pessoas aqui o que eles pensavam em termos daqueles diferentes tipos de livros. Então, vamos olhar o código .

## 43 Análise Conjunta no Exemplo parte 1

### 43.1 Sétimo vídeo da quarta semana

Então aqui estou eu no Ambiente  Studio, e vamos usar esta biblioteca conjunta, que mencionei anteriormente. Então, vamos carregar isso executando essa biblioteca de linhas.

Se não responder corretamente, você precisará pacotes aqui e instale e, em seguida, verifique se você baixa e instala esse pacote em seu ambiente. Isso exigirá isso biblioteca de clustering, fpc.

Portanto, verifique se você tem instalado isso também. As três primeiras linhas de código são os atributos, páginas e então criamos três vetores de coluna: menos de 500 páginas, 500-1000 páginas e mais de 1000 páginas. Criamos gênero: ficção, não-ficção e autor tipo: conhecido e desconhecido.

Então, basicamente criamos essas três colunas vetores 1, 2, 3. Lá estão eles. Se você quiser inspecioná-los, vamos inspecionar páginas. Você pode ver as páginas e ver que é apenas tem esses três elementos, e se

you quiser acessar apenas um deles, colchete, digamos o segundo, aí está. Então esse é o segundo.

Nós vamos criar estes níveis de fator como DataFrames, e é nossa coluna vetor. Então vamos fazer isso. Aqui vamos nós. Está sete observações de uma variável. isto se parece com isso. Se digitarmos níveis de fator de sublinhado, aí estão eles.

Então, essas são todas as diferentes níveis que temos, os três comprimentos do livro, os dois tipos de gênero e os dois tipos de autor. Isso apenas nomeia o níveis de colunas. Mas agora vamos ver nossos dados. Dados é o nome de a variável que estamos criando ou os dados estrutura que estamos criando e vamos usar Neste comando, expanda.grid.

Isso criará nossa grade de dados. Então, vamos olhar para isso, e você pode ver a estrutura e isso fará mais sentido. Então variável 1. Este é basicamente o mesa que tínhamos antes.

Curto médio ou longo livros, ficção, livros de não ficção, autor conhecido ou desconhecido. Então esses são os três categorias, e então isso coloca a coluna nomes nesses dados.

Então, em vez de ver as variáveis 1, 2 e 3, veremos páginas, gênero e autores e isso dará mais significado ao humano leitor. Então vamos fazer isso. Eu posso olhar para isso dados novamente, D-A-T-A, e você pode ver em vez de acima, apenas tinha variáveis 1, variável 2, variável 3 e agora lista a tabela com o títulos de coluna corretos.

Nós vamos criar um planejamento fatorial, dados é igual a dados. Então isso significa que vamos use essa grade que acabamos de criar e esses serão os dados usados neste desenho fatorial. Que tipo de fatorial design que estamos criando? Estamos criando um completo desenho fatorial.

Outra opção seria ortogonal. Então vamos fazer isso. Você pode olhar para isso, se quiser, e então isso é basicamente a mesa novamente. Vamos codificar o design que cria a tabela em números 1, 2 e 3, codificam o design e se olharmos para isso, podemos ver a tabela 1, 2 etc. etc. uma forma numérica mais fácil para o nosso pacote de software a ser usado. Podemos olhar para correlações. Você pode ver que, neste caso, as páginas são obviamente correlacionado consigo mesmo, mas não está correlacionado com qualquer outra coisa, e da mesma forma para o outras duas variáveis.

Então vamos na verdade crie algumas respostas. Então, no mundo real, você sairia e faça uma pesquisa com várias pessoas, mas vamos ver algumas respostas. Linha 20, set.seed, que define a semente para o acaso gerador de números.

Portanto, neste próximo comando, há um comando de amostra, que descreverei em um minuto, mas que basicamente cria selecionar aleatoriamente as coisas. Mas o software realmente não selecionar aleatoriamente as coisas. Ele usa algum tipo de semente para iniciar o gerador de números aleatórios. Então, o que set.seed faz é você pode especificar um número aqui.

Realmente não Não importa qual seja esse número, mas contanto que você e outra pessoa estejam usando a mesma semente, você obterá exatamente os mesmos resultados. Você terá o mesmo "Números aleatórios. Então, neste caso, eu apenas usei um. Poderia ser 10. Poderia ser 100.

Vamos criar o comando de amostra. Então, vamos apenas executar essa parte primeiro e veja o que faz. Então amostra significa o primeiro argumento aqui é uma amostra de 1 a 12. Portanto, escolha um número de 1 a 12.

Quantas vezes? 12 vezes aqui e substituir é igual a FALSO. Então isso significa, imagine você tem 12 números em uma sacola, 12 bolas de pingue-pongue, todas numeradas de 1 a 12, estão em uma sacola e você pega um número e depois pega outro número e então você pega outro número até você tirar todos os 12 números.

Isso é o que isso representa aqui. Se você diz que substituir é igual a VERDADEIRO, isso significa que o número da sacola, eu olho, gravo e coloco a bola de pingue-pongue de volta para a bolsa.

Portanto, se substituir for igual a VERDADEIRO, você poderá ter o mesmo número que ocorre duas vezes. Então, se você tivesse uma amostra até 12 substituem igual TRUE, que pode emular rolando dois dados. Então você poderia rolar você recebe sete. Você poderia rolar novamente, você recebe sete.

Nesse caso, é mais como depois de retirá-lo, você não pode usá-lo novamente. Então isso pode ser como distribuir cartas ou algo assim. Então essa é outra maneira para pensar sobre isso. Então, nós apenas estamos indo para pegue 12 bolas de pingue-pongue, puxe-as em alguma ordem, colete-as e essas serão nossas respostas. Isso apenas leva isso vetor de coluna de amostra e o coloca como um DataFrame. Então, vamos executar isso. Aí está.

Então, esse loop, o que isso faz é que vamos fazer isso mais nove vezes para eu indo de 2-10. Nós apenas estamos indo para faça a mesma coisa e vincule-o ao DataFrame. Então, o que está acontecendo aqui? Bem, deixe-me mostrar-lhe.

Deixe-me executar isso. Bem, deixe eu te mostro a resposta primeiro. Aqui está a resposta, R-E-S-P-O-N-S-E. Você pode ver aqui, é apenas um DataFrame. Aí está. Existem os números no lado direito. O lado esquerdo é o ordem neste DataFrame. Quando eu corro o loop, eu vou fazer um monte mais vezes, e depois ligar apenas anúncios nessa coluna. Eu crio uma coluna e adicione-o a essa tabela. Crie uma coluna

e vinculá-lo a essa tabela.

Então é isso que está acontecendo aqui. Então agora podemos olhar para isso resposta, R-E-S-P-O-N, novamente, e aqui você pode veja muito mais exemplos. Deve ter dez no total. Um dois três quatro cinco seis sete, oito, nove, 10. Então, nós temos 10 respostas e eles estão todos nesta coluna. Isso é ótimo. Então agora temos fatores. Onde estou? Respostas, lá estão eles. Eu cliquei duas vezes aqui à direita e agora nós pode ver esta tabela e aqui estão as valores diferentes para 10. Você lê para cada pesquisa, para cada pessoa.

Então agora que eu tenho minhas respostas, o que esse comando `t` faz é, aqui estão as respostas. Eu só vou levar a transposição disso, e você pode vê-lo aqui, para que as colunas se tornem linhas e o linhas se tornam colunas. Colunas se tornam linhas, linhas se tornam, essa é a transposição aqui. Então, os nomes das linhas, tenho de um a 10. Esses são cada um os entrevistados e os nomes das colunas chamaremos o perfil 1, perfil 2, perfil 3, etc.

Então, vamos executar esses pequenos comandos aqui e então nós pode olhar para isso agora. Vejamos a resposta aqui. Agora, esta é a tabela que Eu mostrei a você nos slides, perfil um através do perfil 12 e os 10 respondentes. Então lembre-se disso cada linha é um respondente, uma pessoa que responde a uma pesquisa e esses são seus resultados da pesquisa e o perfil é o tipo de livro.

Então o perfil um era um livro curto, ficção com um autor conhecido, etc. Então, essas são todas as combinações diferentes.

```


1 library(conjoint)
2 require(fpc)
3
4 pages <- c("less than 500", "500 to 1000", "more than 1000")
5 genre <- c("fiction", "non-fiction")
6 author <- c("known", "unknown")
7 factor_levels <- as.data.frame(c(pages, genre, author))
8 colnames(factor_levels) <- "levels"
9 data <- expand.grid(pages, genre, author)
10 colnames(data) <- c("Pages", "Genre", "Author")
11
12
13 #Creating a set of profiles for the factors
14 facdesign <- caFactorialDesign(data = data, type = "full")
15 encdesign <- caEncodedDesign(facdesign)
16
17 cor(encdesign)
18
19 #Fill in random responses for the 10 respondents for all the 12 profiles created
20 set.seed(1)
21 response <- as.data.frame(sample(1:12, 12, rep= FALSE))
22 for (i in 2 : 10) {
23   temp <- as.data.frame(sample(1:12, 12, rep = FALSE))
24   response <- cbind(response, temp)
25 }
26
27 response <- t(response)
28 row.names(response) <- c(1:10)
29 colnames(response) <- c(paste("Profile", c(1:12)))
30 response <- as.data.frame(response)
31
32
33 #Partial utility
34 partutil <- caPartUtilities(y = response[1:5,], x = encdesign, z = factor_levels)
35
36 util <- caUtilities(y = response, x = encdesign, z = factor_levels)
37
38 #Conjoint Analysis
39 analysis <- Conjoint(response, encdesign, factor_levels)
40
41
42 #Obtaining the importance of each of the factors
43 Importance <- caImportance(response, encdesign)
44 Factor <- c("Pages", "Genre", "Author")
45
46 FactorImp <- as.data.frame(cbind(Factor, Importance))
47
48
49 #Segmentation and Cluster Plot
50 pref <- as.vector(t(response))
51 seg <- caSegmentation(pref, encdesign, c = 2)
52

```

```

53 cluster <- as.data.frame(seg$sclu)
54 colnames(cluster) <- "Cluster"
55
56 plotcluster(seg$util, seg$sclu, pch = 20, xlab = " ", ylab = " ", main = "K-Means Clustering
    Result")

```

Listing 15: Script de análise conjunta no 

```

install.packages("cojoint")

## Installing package into '/usr/local/lib/R/site-library'
## (as 'lib' is unspecified)
## Warning in install.packages("cojoint"): 'lib = "/usr/local/lib/R/site-library"' is not
writable
## Error in install.packages("cojoint"): unable to install packages

library(conjoint)

## Error in library(conjoint): there is no package called 'conjoint'

require(fpc)

## Loading required package: fpc
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE, logical.return =
TRUE, : there is no package called 'fpc'

pages <- c("less than 500", "500 to 1000", "more than 1000")
pages

## [1] "less than 500" "500 to 1000" "more than 1000"

genre <- c("fiction", "non-fiction")
genre

## [1] "fiction" "non-fiction"

author <- c("known", "unknown")
author

## [1] "known" "unknown"

factor_levels <- as.data.frame(c(pages, genre, author))
factor_levels

## c(pages, genre, author)
## 1 less than 500
## 2 500 to 1000
## 3 more than 1000
## 4 fiction
## 5 non-fiction
## 6 known
## 7 unknown

colnames(factor_levels) <- "levels"
data <- expand.grid(pages, genre, author)
data

## Var1 Var2 Var3
## 1 less than 500 fiction known
## 2 500 to 1000 fiction known
## 3 more than 1000 fiction known
## 4 less than 500 non-fiction known
## 5 500 to 1000 non-fiction known
## 6 more than 1000 non-fiction known

```

```
## 7   less than 500      fiction unknown
## 8     500 to 1000      fiction unknown
## 9   more than 1000      fiction unknown
## 10  less than 500 non-fiction unknown
## 11     500 to 1000 non-fiction unknown
## 12 more than 1000 non-fiction unknown

colnames(data) <- c("Pages", "Genre", "Author")
colnames(data)

## [1] "Pages" "Genre" "Author"

#Creating a set of profiles for the factors
facdesign <- caFactorialDesign(data = data, type = "full")

## Error in caFactorialDesign(data = data, type = "full"): could not find function "caFactorialDesign"
facdesign

## Error in eval(expr, envir, enclos): object 'facdesign' not found
encdesign <- caEncodedDesign(facdesign)

## Error in caEncodedDesign(facdesign): could not find function "caEncodedDesign"
encdesign

## Error in eval(expr, envir, enclos): object 'encdesign' not found
cor(encdesign)

## Error in is.data.frame(x): object 'encdesign' not found

#Fill in random responses for the 10 respondents for all the 12 profiles created
set.seed(1)
response <- as.data.frame(sample(1:12, 12, rep= FALSE))
response

##      sample(1:12, 12, rep = FALSE)
## 1                                4
## 2                                5
## 3                                6
## 4                                9
## 5                                2
## 6                                7
## 7                               10
## 8                               12
## 9                                3
## 10                               1
## 11                               11
## 12                               8

for (i in 2 : 10) {
  temp <- as.data.frame(sample(1:12, 12, rep = FALSE))
  response <- cbind(response, temp)
}

response <- t(response)
response

##               [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
## sample(1:12, 12, rep = FALSE)  4  5  6  9  2  7  10  12  3
## sample(1:12, 12, rep = FALSE)  9  5  8  11  6  7  3  4  12
## sample(1:12, 12, rep = FALSE)  4  5  1  12  7  3  8  6  2
```

```

## sample(1:12, 12, rep = FALSE) 10 2 8 4 7 5 12 3 6
## sample(1:12, 12, rep = FALSE) 9 8 5 11 4 2 1 6 7
## sample(1:12, 12, rep = FALSE) 11 4 5 3 6 2 9 12 1
## sample(1:12, 12, rep = FALSE) 5 4 12 9 7 3 10 6 2
## sample(1:12, 12, rep = FALSE) 10 3 8 2 9 12 7 1 11
## sample(1:12, 12, rep = FALSE) 6 5 9 12 10 3 2 11 7
## sample(1:12, 12, rep = FALSE) 12 7 10 11 3 4 1 6 8
##
##           [,10] [,11] [,12]
## sample(1:12, 12, rep = FALSE) 1 11 8
## sample(1:12, 12, rep = FALSE) 1 2 10
## sample(1:12, 12, rep = FALSE) 10 9 11
## sample(1:12, 12, rep = FALSE) 9 1 11
## sample(1:12, 12, rep = FALSE) 12 3 10
## sample(1:12, 12, rep = FALSE) 8 10 7
## sample(1:12, 12, rep = FALSE) 8 1 11
## sample(1:12, 12, rep = FALSE) 4 6 5
## sample(1:12, 12, rep = FALSE) 1 4 8
## sample(1:12, 12, rep = FALSE) 5 9 2

row.names(response) <- c(1:10)
row.names

## function (x)
## UseMethod("row.names")
## <bytecode: 0x561c07e45d18>
## <environment: namespace:base>

colnames(response) <- c(paste("Profile", c(1:12)))
colnames

## function (x, do.NULL = TRUE, prefix = "col")
## {
##   if (is.data.frame(x) && do.NULL)
##     return(names(x))
##   dn <- dimnames(x)
##   if (!is.null(dn[[2L]]))
##     dn[[2L]]
##   else {
##     nc <- NCOL(x)
##     if (do.NULL)
##       NULL
##     else if (nc > 0L)
##       paste0(prefix, seq_len(nc))
##     else character()
##   }
## }
## <bytecode: 0x561c080fdcf8>
## <environment: namespace:base>

response <- as.data.frame(response)
response

##   Profile 1 Profile 2 Profile 3 Profile 4 Profile 5 Profile 6 Profile 7
## 1         4         5         6         9         2         7         10
## 2         9         5         8        11         6         7         3
## 3         4         5         1        12         7         3         8
## 4        10         2         8         4         7         5        12
## 5         9         8         5        11         4         2         1
## 6        11         4         5         3         6         2         9
## 7         5         4        12         9         7         3        10

```

```
## 8      10      3      8      2      9      12      7
## 9      6      5      9      12     10      3      2
## 10     12      7     10     11      3      4      1
##      Profile 8 Profile 9 Profile 10 Profile 11 Profile 12
## 1      12      3      1      11      8
## 2      4      12      1      2      10
## 3      6      2      10      9      11
## 4      3      6      9      1      11
## 5      6      7      12      3      10
## 6      12      1      8      10      7
## 7      6      2      8      1      11
## 8      1      11      4      6      5
## 9      11      7      1      4      8
## 10     6      8      5      9      2

#Partial utility
partutil <- caPartUtilities(y = response[1:5,], x = encdesign, z = factor_levels)

## Error in caPartUtilities(y = response[1:5, ], x = encdesign, z = factor_levels): could
not find function "caPartUtilities"

partutil

## Error in eval(expr, envir, enclos): object 'partutil' not found

util <- caUtilities(y = response, x = encdesign, z = factor_levels)

## Error in caUtilities(y = response, x = encdesign, z = factor_levels): could not find function
"caUtilities"

util

## Error in eval(expr, envir, enclos): object 'util' not found

#Conjoint Analysis
analysis <- Conjoint(response, encdesign, factor_levels)

## Error in Conjoint(response, encdesign, factor_levels): could not find function "Conjoint"

analysys

## Error in eval(expr, envir, enclos): object 'analysys' not found

#Obtaining the importance of each of the factors
Importance <- caImportance(response, encdesign)

## Error in caImportance(response, encdesign): could not find function "caImportance"

Importance

## Error in eval(expr, envir, enclos): object 'Importance' not found

Factor <- c("Pages", "Genre", "Author")
Factor

## [1] "Pages" "Genre" "Author"

FactorImp <- as.data.frame(cbind(Factor, Importance))

## Error in cbind(Factor, Importance): object 'Importance' not found

FactorImp

## Error in eval(expr, envir, enclos): object 'FactorImp' not found

#Segmentation and Cluster Plot
pref <- as.vector(t(response))
pref
```

```
##      [1]  4  5  6  9  2  7 10 12  3  1 11  8  9  5  8 11  6  7  3  4 12  1  2 10  4
##     [26]  5  1 12  7  3  8  6  2 10  9 11 10  2  8  4  7  5 12  3  6  9  1 11  9  8
##     [51]  5 11  4  2  1  6  7 12  3 10 11  4  5  3  6  2  9 12  1  8 10  7  5  4 12
##     [76]  9  7  3 10  6  2  8  1 11 10  3  8  2  9 12  7  1 11  4  6  5  6  5  9 12
##    [101] 10  3  2 11  7  1  4  8 12  7 10 11  3  4  1  6  8  5  9  2

seg <- caSegmentation(pref, encdesign, c = 2)

## Error in caSegmentation(pref, encdesign, c = 2): could not find function "caSegmentation"

seg

## Error in eval(expr, envir, enclos): object 'seg' not found

cluster <- as.data.frame(seg$sclu)

## Error in as.data.frame(seg$sclu): object 'seg' not found

cluster

## Error in eval(expr, envir, enclos): object 'cluster' not found

colnames(cluster) <- "Cluster"

## Error in colnames(cluster) <- "Cluster": object 'cluster' not found

plotcluster(seg$util, seg$sclu, pch = 20, xlab = " ", ylab = " ", main = "K-Means Clustering Result")

## Error in plotcluster(seg$util, seg$sclu, pch = 20, xlab = " ", ylab = " ", : could not
find function "plotcluster"
```

## 44 Conjoint Analysis Example – Part 2

### 44.1 Último vídeo da última semana

Até agora, estamos apenas configurando os dados e configurando os resultados da pesquisa. Se você estivesse fazendo isso na vida real, você os colocaria em uma tabela que parece algo assim. Em uma planilha do Excel, importe a planilha do Excel e você poderá acessar a próxima função aqui que está recebendo o utilidades parciais.

Vamos olhar para isso. Aqui estão eles. Então aqui estão as utilidades parciais. Autores conhecidos é mais preferível do que autores desconhecidos. Ficção não-ficção. Parece que a não-ficção é mais popular.

Gênero, temos ficção não-ficção. Nós criamos autores, onde está o outro gráfico? Oh, está bem aqui no canto inferior direito da tela.

Então, para cada um dos atributos de amor, temos um gráfico como este. Este tem três níveis. Podemos ver que parece livros curtos são livros mais populares e de tamanho médio não são tão populares, e livros longos são populares aqui. Você pode ver isso utilidades não são.

Isso é quase zero aqui. Podemos olhar para a análise. Estas são as médias importância, esses gráficos. Deixe-me executar estes próximos códigos e eu vou te mostrar o que eles descrevem em um pouco mais detalhes nos slides. Existe a importância. Esses são meus fatores e aqueles são minhas parcelas preferenciais.

Permite obter a importância de cada um desses fatores, esses atributos. Então é isso que o a importância faz. Deixe-me executar o todo linha. Aqui vamos nós. Aqui estão meus fatores e essa é a importância do fator que será algo como isto. Vejamos os slides que os descrevi em um um pouco mais de detalhes. Então eu coloquei os resultados em esses slides do PowerPoint. É um pouco mais fácil de descrever.

As utilidades de peças aqui estão eles. Esses valores representam o valor da parte do comprimento do livro, gênero, autor desconhecido e conhecido, e existe esse interceptar termo aqui. Este é o que aqueles gráficos que estamos mostrando esses gráficos de peças. Aqui temos o real resultados da análise conjunta. Observe que é apenas um modelo linear. Aqui abaixo, você pode ver a importância de cada um desses fatores. Então lá estão eles, eles estão listados numericamente.

Depois à direita aqui, listei todos os utilidades com valor parcial. A saída mostra o impacto estimado na utilidade por cada um desses níveis de fator. Então, se você olhar fator X páginas um, se for um livro



pequeno, tem um negativo. Livros curtos não tão bons, livros de tamanho médio não tão bons. Gênero 1, eu acho isso foi ficção.

Portanto, neste conjunto de dados de ficção foi um resultado positivo. Aqui estão os gráficos, e é aqui que você realmente quer concentre sua atenção.

Páginas, gênero, autor, eles têm peso igual neste exemplo, para que os consumidores olhem principalmente páginas, mas não muito. Lembre-se, este é um conjunto de dados fabricados para que alguns desses resultados são um pouco diretamente.

Você poderia imaginar que uma dessas categorias seria de mais importante para as pessoas. Aqui estão as utilidades de peça para o nível de páginas como os resultados anotados.

Livros médios curtos são não são preferidos, livros longos são preferidos. Então podemos olhar para um agrupamento, realmente o que isso faz é tentar agrupar segmentos que gostam tipos semelhantes de livros. Aqui você pode ver com dois grupos, podemos ver que existem pontos pretos e laranja e aqueles são os dois grupos.

Portanto, há algum agrupamento. Veremos isso em um minuto ou no próximo vídeo. Então, como interpretamos os resultados? Podemos ver que o O número de páginas nos livros é um pouco mais importante que o outros dois fatores, quando olhamos para o classificando os atributos, mas o atributo mais preferido para procurar é o número de páginas.

Em conclusão, os resultados indicam que os mais livro popular seriam aqueles que são escritos por ficção, autores conhecidos e mais de 1.000 páginas.

```

1
2 #Partial utility
3 partutil <- caPartUtilities(y = response[1:5,], x = encdesign, z = factor_levels)
4
5 util <- caUtilities(y = response, x = encdesign, z = factor_levels)
6
7 #Conjoint Analysis
8 analysis <- Conjoint(response, encdesign, factor_levels)
9
10
11 #Obtaining the importance of each of the factors
12 Importance <- caImportance(response, encdesign)
13 Factor <- c("Pages", "Genre", "Author")
14
15 FactorImp <- as.data.frame(cbind(Factor, Importance))
16
17
18 #Segmentation and Cluster Plot
19 pref <- as.vector(t(response))
20 seg <- caSegmentation(pref, encdesign, c = 2)
21
22 cluster <- as.data.frame(seg$sclu)
23 colnames(cluster) <- "Cluster"
24
25 plotcluster(seg$util, seg$sclu, pch = 20, xlab = " ", ylab = " ", main = "K-Means Clustering
    Result")

```

Listing 16: Análise Conjunta no 

```

#Partial utility
library(conjoint)

## Error in library(conjoint):  there is no package called 'conjoint'

require(fpc)

## Loading required package: fpc
## Warning in library(package, lib.loc = lib.loc, character.only = TRUE, logical.return =
TRUE, :  there is no package called 'fpc'

partutil <- caPartUtilities(y = response[1:5,], x = encdesign, z = factor_levels)

## Error in caPartUtilities(y = response[1:5, ], x = encdesign, z = factor_levels):  could
not find function "caPartUtilities"

util <- caUtilities(y = response, x = encdesign, z = factor_levels)

```

```
## Error in caUtilities(y = response, x = encdesign, z = factor_levels): could not find function
"caUtilities"

#Conjoint Analysis
analysis <- Conjoint(response, encdesign, factor_levels)

## Error in Conjoint(response, encdesign, factor_levels): could not find function "Conjoint"

#Obtaining the importance of each of the factors
Importance <- caImportance(response, encdesign)

## Error in caImportance(response, encdesign): could not find function "caImportance"

Factor <- c("Pages", "Genre", "Author")

FactorImp <- as.data.frame(cbind(Factor, Importance))

## Error in cbind(Factor, Importance): object 'Importance' not found

#Segmentation and Cluster Plot
pref <- as.vector(t(response))
seg <- caSegmentation(pref, encdesign, c = 2)

## Error in caSegmentation(pref, encdesign, c = 2): could not find function "caSegmentation"

cluster <- as.data.frame(seg$sclu)

## Error in as.data.frame(seg$sclu): object 'seg' not found

colnames(cluster) <- "Cluster"

## Error in colnames(cluster) <- "Cluster": object 'cluster' not found

plotcluster(seg$util, seg$sclu, pch = 20, xlab = " ", ylab = " ", main = "K-Means Clustering Result")

## Error in plotcluster(seg$util, seg$sclu, pch = 20, xlab = " ", ylab = " ", : could not
find function "plotcluster"
```

## 45 Resumo do curso: Aplicando Data Analytics no Marketing

### 45.1 Entrevista com Monica Penagos

Então, eu quero agradecer a Monica por passar um tempo conosco e compartilhar sua experiência e sua sabedoria sobre o uso de análises no mundo real.

E também quero agradecer por hospedando-nos em seu site e em suas salas de conferência aqui em a sede da Procter Gamble.

Monica, você tem alguma pérola de sabedoria que você gostaria de compartilhar com a nossa turma antes de encerrarmos? *Li* Obrigado por me receber aqui. Foi um prazer falar para você sobre análises. Não tenho pérolas de sabedoria, mas nos perguntam o tempo todo sobre a P&G e como a PG vê as análises.

E o que eu diria é quando você estiver pensando em análise e tecnologia, muitas pessoas pensam em empresas de tecnologia. Então eles pensam no Facebook, Google, Amazon. E o que eu encorajaria você e seus alunos são, pense no consumidor empresas de produtos embalados.

Nenhum dos produtos que temos em hoje essa sala seria possível sem ela. Nenhum deles. Todos eles passaram por todo o metodologias que estávamos conversando hoje para possibilitar e toda a publicidade.

Nosso futuro existe em dados e análises para tomar as decisões de torná-lo mais personalizado para nós? Eu sou meu próprio consumidor e quero ser segmentado de maneira personalizada com produtos relevantes que realmente falam comigo, e isso só será possível com análises, dados e tecnologia em que estamos vivendo hoje. Esse é um ponto excelente. A análise de dados não é apenas para empresas de alta tecnologia, mas para empresas de produtos como como Procter & Gamble.

Eu também vi isso em análises esportivas, e eu também vi isso na área da saúde. Portanto, a análise está ao nosso redor, como você disse. Muito obrigado pelo seu tempo. Obrigado.

## Referências

- [1] Gujarati, D.,N. **Basic Econometrics**, fourth edition, McGraw-Hill/Irwin, 2003.
- [2] Disponível em Overleaf.com
- [3] Xie, Y. **Dynamic Documents with R and knitr** 2nd edition, 2015.
- [4] **Reproducible Research using RMarkdown and Overleaf** Disponível em [Reproducible Research using RMarkdown and Overleaf](#)