



Modelos de Regressão e Previsão (ACA228)

Revisão de MAD211

Prof. Carlos Trucíos

 ctruciosm.github.io

 carlos.trucios@facc.ufrj.br

Faculdade de Administração e Ciências Contábeis,
Universidade Federal do Rio de Janeiro

Revisão

Antes de estudarmos com detalhe métodos mais sofisticados, precisamos entender/revisar alguns conceitos básicos que nos acompanharão ao longo do curso:

Estatística descritiva

Distribuições de probabilidade

Testes de Hipóteses



ctruciosm.github.io — Carlos Trucíos (FACC/UFRJ)

Estatística descritiva



Estatística descritiva

Dataset

Importando dados no R

promoted	sales	customer_rate	performance
0	594	3.94	2
0	446	4.06	3
1	674	3.83	4
0	525	3.62	2
1	657	4.40	3
1	918	4.54	2
0	318	3.09	3
0	364	4.89	1
0	342	3.74	3
0	387	3.00	3
0	527	2.43	3
1	716	3.16	3
0	557	3.51	2
0	450	3.21	3
0	344	3.02	2
0	372	3.87	3
0	258	2.49	1

Estatística descritiva

Sejam x_1, \dots, x_n, n observações de uma caracteristica (numérica) de interesse. A **média** é o valor promedio das observações e é definido como:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}.$$

```
mean(salespeople$sales, na.rm = TRUE)
```

```
## [1] 527.0057
```

| A média é intuitiva e fácil de entender/explicar, mas é afetada por observações extremas.

Estatística descritiva

Sejam $x_{(1)}, x_{(2)}, \dots, x_{(n)}$ os valores ordenados (de menor a maior) de x_1, \dots, x_n . A **médiana** é o valor *do meio* dos dados ordenados.

$$\text{Mediana}(x) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{se } n \text{ for ímpar} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{se } n \text{ for par.} \end{cases}$$

```
median(salespeople$sales, na.rm = TRUE)
```

```
## [1] 475
```

| A mediana é robusta a observações atípicas

Estatística descritiva

A **variância** é uma medida de variabilidade (em torno da média) dos dados. Quanto maior a variância, maior é a variabilidade.

$$\hat{\sigma}_x^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}$$

```
var(salespeople$sales, na.rm = TRUE)
```

```
## [1] 34308.11
```

A variância, por estar em **unidades ao quadrado** é difícil de interpretar e na prática é preferido o desvio padrão (raiz quadrada da variância).

Estatística descriptiva

O **desvio padrão** é outra medida de variabilidade. Quanto maior o valor, maior é a variabilidade. É definido como a raiz quadrada da variância.

$$\hat{\sigma}_x = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

```
sd(salespeople$sales, na.rm = TRUE)
```

```
## [1] 185.2245
```

| A vantagem do desvio padrão sobre a variância é que esta medida de variabilidade está na escala original dos dados.



Estatística descriptiva

A **covariância** mede o grau de associação entre 2 variáveis,

$$\text{cov}(x, y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

```
cov(salespeople, use = "complete.obs")
```

```
##          promoted      sales customer_rate performance
## promoted   0.21924683  73.81763  0.07561293  0.11891117
## sales     73.81763406 34308.11458  55.81769120 49.40687679
## customer_rate  0.07561293   55.81769  0.79581959  0.05008596
## performance  0.11891117   49.40688  0.05008596  0.90974212
```

Covariância igual a zero indica que as variáveis não tem nenhum grau de associação. A desvantagem da covariância é que não tem valor mínimo nem máximo.



Estatística descriptiva

O **coeficiente de correlação de Pearson** ($\rho_{x,y}$) também mede o grau de associação entre 2 variáveis (numéricas), mas $-1 \leq \rho_{x,y} \leq 1$

$$\rho(x, y) = \frac{cov(x, y)}{\hat{\sigma}_x \hat{\sigma}_y}$$

```
cor(salespeople, use = "complete.obs")  
  
##           promoted      sales customer_rate performance  
## promoted    1.0000000  0.8511283   0.18101815  0.26625444  
## sales       0.8511283  1.0000000   0.33780504  0.27965966  
## customer_rate 0.1810182  0.3378050   1.00000000  0.05886397  
## performance  0.2662544  0.2796597   0.05886397  1.00000000
```

| Correlação zero indica ausencia de associação entre as variáveis, valores próximos a 1(-1) indicam associação direta (inversa) entre as variáveis.



Estatística descriptiva

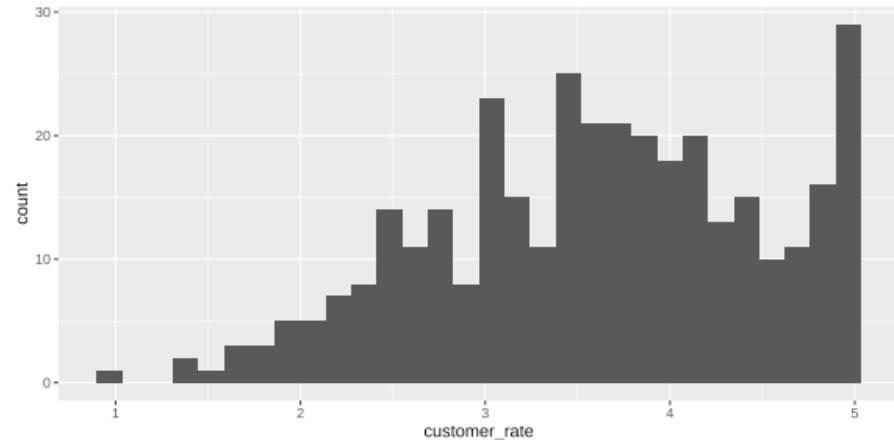
- Os quantis também podem ser utilizados
- Quando trabalhamos com variáveis qualitativas, calcular tabelas de frequência pode ser bastante útil.
- Histogramas, gráficos de barras, boxplots e gráficos de dispersão são bastante úteis no EDA.

Histograma

Boxplot

Gráfico de Barras

Gráfico de dispersão



ctruciosm.github.io — Carlos Trucíos (FACC/UFRJ)

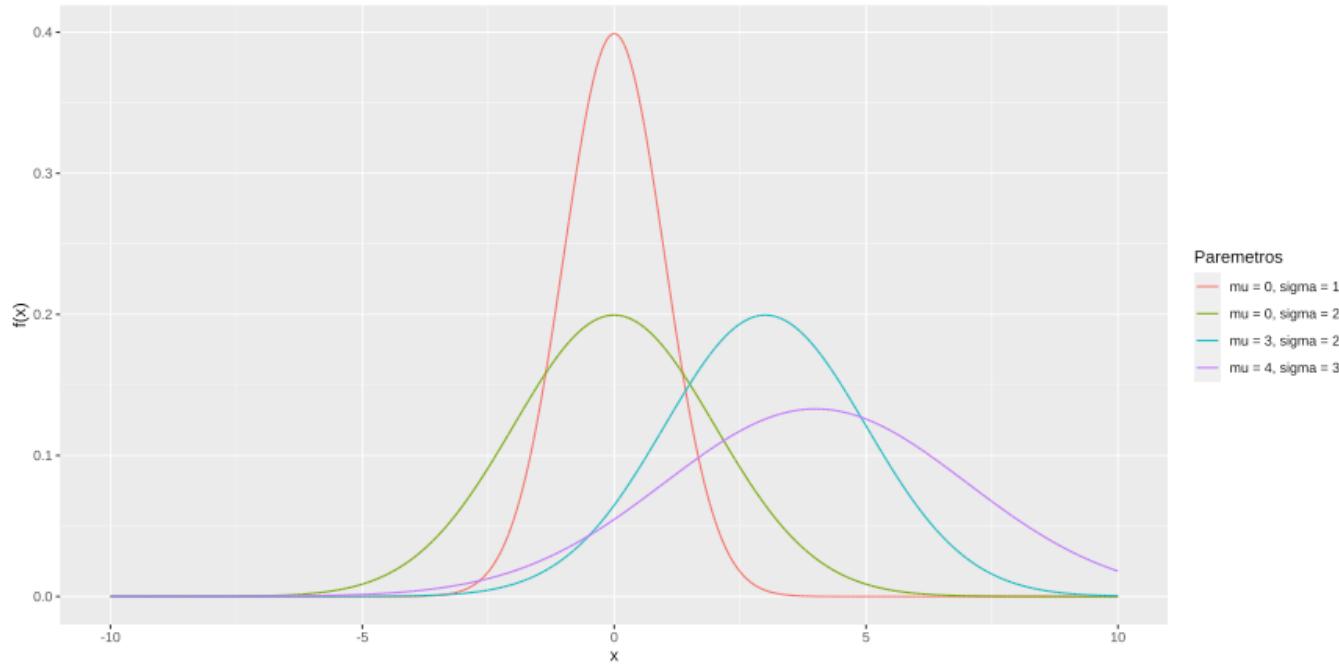
Distribuições de probabilidade.

Distribuições de probabilidade

Dist. Normal

Dist. T

Dist. F

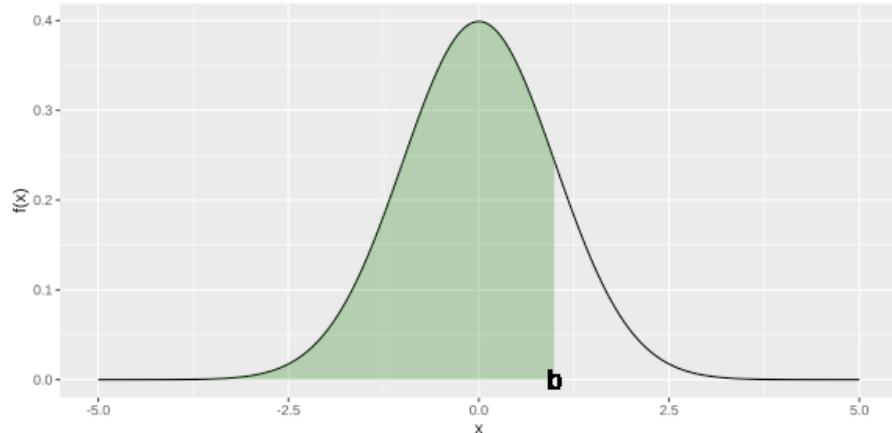


Distribuições de probabilidade

Dist. Normal

Dist. T com 7 g.l

Dist. F com 7 e 13 g.l



```
c(qnorm(0.05), qnorm(0.5), qnorm(0.975)) # R
```

```
## [1] -1.644854  0.000000  1.959964
```

Distribuições de probabilidade

- Sejam X_1, X_2, \dots, X_n v.as $\sim N(\mu, \sigma)$, então

$$\frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} \sim N(0, 1)$$

- Quando não conhecemos σ e substituimos este valor por $\hat{\sigma}$, temos que

$$\frac{(\bar{X}_n - \mu)}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

- Sejam $X_1, \dots, X_{n_x} \sim N(\mu_x, \sigma_x)$ e sejam $Y_1, \dots, Y_{n_y} \sim N(\mu_y, \sigma_y)$. Então

$$F = \frac{\hat{\sigma}_x^2/\sigma_x^2}{\hat{\sigma}_y^2/\sigma_y^2} \sim F_{n_x-1, n_y-1},$$

em que $\hat{\sigma}_x^2$ e $\hat{\sigma}_y^2$ são a variâncias amostral de X_1, \dots, X_{n_x} e Y_1, \dots, Y_{n_y} , respectivamente.

Distribuições de probabilidade

Teorema Central do Limite (TCL)

Sejam X_1, X_2, \dots, X_n (para n grande) v.as **independentes** e **identicamente distribuidas** com $E(X_1) = \mu$ e $V(X_1) = \sigma^2 < \infty$. Então,

$$\frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} \sim_{aprox} N(0, 1)$$

ctruciosm.github.io — Carlos Trucíos (FACC/UFRJ)

Testes de hipóteses:

Testes de hipóteses

Estamos interessados em verificar se, com base nos dados da nossa amostra, podemos **rejeitar ou não rejeitar** um determinada afirmação (**hipótese**) sobre um estado da natureza (parâmetro de interesse).

Para fazer um teste de hipótese precisamos:

1. Definir um nível de significância α
2. Construir a estatística de teste
3. Comparar o valor da estatística de teste o quantil teórico da distribuição (sob H_0)
4. Tomar uma decisão em função da comparação feita anteriormente

Teste para a média populacional.

Sejam as hipóteses

$$H_0 : \mu = \mu_0 \quad vs. \quad H_1 : \mu \neq \mu_0$$

Nossa estatística de teste é da forma

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

Assim, para um nível de significância α , rejeitamos H_0 se

$$\left| t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} \right| > t_{1-\alpha/2, n-1}$$

| Podemos também fazer testes unilaterais

Teste para a média populacional.

Ejemplo: Sejam as hipóteses

$$H_0 : \mu = 4 \quad vs. \quad H_1 : \mu \neq 4, \quad \text{em que } \mu \text{ é a média do customer_rate}$$

```
alpha = 0.05
t.test(salespeople$customer_rate, mu = 4, alternative = "two.sided", conf.level = 1-alpha)
```

```
##
##      One Sample t-test
##
## data: salespeople$customer_rate
## t = -8.2232, df = 349, p-value = 3.955e-15
## alternative hypothesis: true mean is not equal to 4
## 95 percent confidence interval:
##  3.514101 3.701670
## sample estimates:
## mean of x
## 3.607886
```

```
qt(1-alpha/2, 349)
```

```
## [1] 1.966785
```

Teste para a média populacional.

Sejam as hipóteses

$$H_0 : \mu \leq \mu_0 \quad vs. \quad H_1 : \mu > \mu_0$$

$$H_0 : \mu \geq \mu_0 \quad vs. \quad H_1 : \mu < \mu_0$$

Nossa estatística de teste é da forma

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} \sim t_{n-1}$$

Assim, para um nível de significância α , rejeitamos H_0 se

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} > t_{1-\alpha, n-1}$$

$$t = \frac{\bar{x} - \mu_0}{\hat{\sigma}/\sqrt{n}} < t_{\alpha, n-1}$$

Teste para a média populacional.

Ejemplo: Sejam as hipóteses

$$H_0 : \mu \leq 3 \quad vs. \quad H_1 : \mu > 3, \quad \text{em que } \mu \text{ é a média do customer_rate}$$

```
alpha = 0.05
t.test(salespeople$customer_rate, mu = 3, alternative = "greater", conf.level = 0.95)
```

```
##
##      One Sample t-test
##
## data: salespeople$customer_rate
## t = 12.748, df = 349, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 3
## 95 percent confidence interval:
##  3.529244     Inf
## sample estimates:
## mean of x
## 3.607886
```

```
qt(1-alpha, 349)
```

```
## [1] 1.649231
```

ctruciosm.github.io — Carlos Trucíos (FACC/UFRJ)

Esperança Condicional:

Esperança Condisional

Seja X e Y duas variáveis aleatórias com função de densidade conjunta

$$f_{X,Y}(x, y),$$

e marginais $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$ e $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$.

A função de densidade condicional de Y dado $X = x$ é definida como

$$f_{Y|X=x}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)}$$

A esperança condicional de Y dado X , denotada como $E(Y|X)$, é uma função de X , cujo valor quando $X = x$ é dado por

$$E(Y|x)) \int_{-\infty}^{\infty} y f_{Y|X=x}(y|x) dy$$

ou seja $E(Y|x)$ é média da distribuição condicional de Y dado $X = x$

Esperança Condisional

Como $E(Y|X)$ depende de X e X é uma variável aleatória, $E(Y|X)$ também é uma variável aleatória e podemos então calcular seu valor esperado.

Teorema

Para quaisquer variáveis aleatórias X e Y , temos que

$$E(E(Y|X)) = E(Y) \quad \text{ou} \quad E(E(X|Y)) = E(X)$$

Propriedades

- $E(Y + Z|X) = E(Y|X) + E(Z|X)$
- $E(cY|X) = cE(Y|X)$
- $E(XY|X) = XE(Y|X)$
- $E(g(X)Y|X) = g(X)E(Y|X)$

Esperança Condicional

Para ilustrar melhor o conceito de **esperança condicional**, veremos um exemplo simples no caso discreto.

Ejemplo: Sejam X e Y duas variáveis aleatórias discretas com função de probabilidade conjunta dada da seguinte forma

	$X = 0$	$X = 1$	
$Y = 0$	2/20	6/20	
$Y = 1$	6/20	9/20	
<hr/>			
1			

Calcule $E(Y)$,

$$E(Y) = \sum_y y p(y) = 0 \times 2/5 + 1 \times 3/5 = 3/5$$

Calcule $E(Y|X = 0)$,

Esperança Condisional

1. Precisamos calcular a função de probabilidade de $Y|X = 0$
2. Calcular o valor esperado $E(Y|X = 0)$

Calculando a função de probabilidade condicional

$$p_{X,Y}(Y = y|X = 0) = \frac{p_{X,Y}(y, X = 0)}{p_X(X = 0)}$$

A função de probabilidade de $Y|X = 0$ é dada por

Y = 0 / X = 0	(2/20)/(8/20) = 1/4
Y = 1 / X = 0	(6/29)/(8/20) = 3/4

$$E(Y|X = 0) = \sum_y y p_{Y|X=0}(y|X = 0) = 0 \times 1/4 + 1 \times 3/4 = 3/4$$

$$E(Y|X = 0) = 3/4 \neq 3/5 = E(Y)$$

ctruciosm.github.io — Carlos Trucíos (FACC/UFRJ)

Os conceitos vistos aqui são a base para ACA228

Caso precisar, pode acessar à playlist de MAD211/2020.2 [Aqui](#)