Projeto IMDB - Enunciado 2*

Natanael Magalhães Cardoso†

31/05/2021

1 Intervalo de confiança da média para variância populacional conhecida

O parâmetro μ será estimado pelo estimador $\bar{x}$, que é definido pela Equação (1).

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

Para um valor de n suficientemente grande, a distribuição de $\bar{x}$ é normal. E, o intervalo que tem a probalidade $1 - \alpha$ de conter o verdadeiro valor do parâmetro μ da população é dado pela Equação (2).

$$P(\bar{x} - e_0 \leq \mu \leq \bar{x} + e_0) = 1 - \alpha \tag{2}$$

Sendo assim, o intervalo de confiança da média considerando variância populacional σ conhecida é expresso pela Equação (3).

$$IC = \bar{x} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}} \tag{3}$$

A Listagem a seguir mostra a implementação da função `IC_media_Z`, que calcula o intervalo de confiança da média para variância populacional conhecida. Aplicando a função nos valores da coluna *Lucro*, para um nível de confiança de 95% e considerando a média populacional $\mu = 55$M, temos que $IC = [14.853.397; 20.182.038]$.

```
IC_media_Z <- function(dados, confianca, dp) {
  xbarra = mean(dados, na.rm=TRUE)
  z = qnorm((1 + confianca) / 2)
  n = length(dados)
  e = z * (dp / sqrt(n))
  LS = xbarra + e
  LI = xbarra - e
  return(c(LI, LS))
}
IC_media_Z(imdb2$Lucro, 0.95, 55e6)
```

```
## [1] 14853397 20182038
```

*os números das seções são equivalentes aos números das perguntas.

†nUSP: 8914122. Usando dados para o segundo grupo (datas entre 2000-2010).

2 Intervalo de confiança da média para variância populacional desconhecida

É possível estimar o parâmetro μ de uma distribuição com variância populacional desconhecida a partir da estimação da variância pelo estimador variância amostral, s, que é definido na Equação (4), para $n \geq 30$.

$$s = \sqrt{\frac{\sum_i^n (x_i - \bar{x})^2}{n-1}} \tag{4}$$

Esta estimativa segue uma distribuição t-Student e o intervalo de confiança para média com variância populacional desconhecida é mostrado na Equação (5).

$$IC = \bar{x} \pm t_{n-1;\alpha/2} \times \frac{s}{\sqrt{n}} \tag{5}$$

A Listagem a seguir mostra a implementação da função `IC_media_t`, que calcula o intervalo de confiança da média para variância populacional desconhecida. Aplicando a função nos valores da coluna *Lucro*, para um nível de confiança de 95%, temos que o intervalo de confiança calculado é $IC = [14.467.782; 20.567.653]$.

```
IC_media_t <- function(dados, confianca) {
  xbarra = mean(dados, na.rm=TRUE)
  n = length(dados)
  t = qt((1 + confianca) / 2, (n - 1))
  e = t * (sd(dados, na.rm=TRUE) / sqrt(n))
  LS = xbarra + e
  LI = xbarra - e
  return(c(LI, LS))
}
IC_media_t(imdb2$Lucro, 0.95)
```

```
## [1] 14467782 20567653
```

3 Criação da variável `Aprovado`

A Listagem a seguir mostra a inclusão de uma nova coluna com valores booleanos no conjunto de dados principal e no subconjunto com filmes lançados entre 2000 e 2010. O campo recebe valor `verdadeiro` se a nota for maior que 6,5 e `falso` caso contrário.

```
imdb2 <- imdb2 %>% mutate(
  Aprovado=ifelse(nota_imdb >= 6.5, TRUE, FALSE)
)
df_g2 <- imdb2 %>% mutate(
  Aprovado=ifelse(nota_imdb >= 6.5, TRUE, FALSE)
)
```

4 Intervalo de confiança da proporção

O estimador p', definido na Equação (6), será usado para calcular a proporção.

$$p' = \frac{f}{n} \tag{6}$$

A distribuição de p' é Binomial, mas, se $np \geq 5$ e $n(1-p) \geq 5$, ela pode ser aproximanda por uma Normal. Desta forma, o intervalo de confiança pode ser descrito pela Equação (7).

$$IC = p' \pm z_{\alpha/2}\sqrt{\frac{p'(1-p')}{n}} \tag{7}$$

‘

A Listagem a seguir mostra a implementação da função `IC_prop`, que calcula o intervalo de confiança da proporção amostral. Para `Aprovado = TRUE`, os intervalos de confiança para proporção são $IC_{80\%} = [0,499; 0,531]$, $IC_{95\%} = [0,491; 0,539]$ e $IC_{99\%} = [0,483; 0,547]$ para os níveis de confiança de 80, 95 e 95%, respectivamente. Foi notado que o intervalo **diminuiu** com o aumento do nível de confiança. Consequentemente, para `Aprovado = FALSE`, o intervalo **aumentou** com o aumento do nível de confiança.

```
IC_prop <- function(dados, confianca) {
  n = length(dados)
  z = qnorm((1 + confianca) / 2)
  p_linha = table(dados) / n
  e = z * sqrt((p_linha * (1 - p_linha)) / n)
  LS = p_linha + e
  LI = p_linha - e
  return(c(LI, LS))
}
IC_prop(imdb2$Aprovado, 0.80)
```

```
##     FALSE      TRUE     FALSE      TRUE
## 0.4692034 0.4991362 0.5008638 0.5307966
```

```
IC_prop(imdb2$Aprovado, 0.95)
```

```
##     FALSE      TRUE     FALSE      TRUE
## 0.4608234 0.4907562 0.5092438 0.5391766
```

```
IC_prop(imdb2$Aprovado, 0.99)
```

```
##     FALSE      TRUE     FALSE      TRUE
## 0.4532159 0.4831488 0.5168512 0.5467841
```

5 Intervalo de confiança da variância

O estimador s, definido na Equação (8), será usado para estimar o parâmetro populacional μ. Da Equação (4), temos

$$s^2 = \frac{\sum_i^n (x_i - \bar{x})^2}{n-1} \tag{8}$$

Mas, se $X \sim \mathcal{N}(\mu; \sigma) \quad \Rightarrow \quad s^2 \sim \chi^2_{n-1}$. Então

$$s^2 = \frac{\sigma^2}{n-1}\chi^2_{n-1} \tag{9}$$

Assim, o intervalo de confiança IC da variância é dado por

$$IC = \left[\sqrt{\frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}}; \sqrt{\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}}\right] \tag{10}$$

A Listagem a seguir mostra a implementação da função `IC_var`, que calcula o intervalo de confiança da variância. O intervalo de confiança da variância do lucro, considerando um nível de confiança de 95%, é $IC = [60.830.134; 65.146.228]$.

```
IC_var <- function(dados, confianca) {
  gl = length(dados) - 1
  S = var(dados, na.rm=TRUE)
  chi_lower = qchisq((1 + confianca) / 2, gl)
  chi_upper = qchisq((1 - confianca) / 2, gl)
  LI = sqrt(gl * S / chi_lower)
  LS = sqrt(gl * S / chi_upper)
  return(c(LI, LS))
}
IC_var(imdb2$Lucro, 0.95)
```

```
## [1] 60830134 65146228
```

6 Análise visual da distribuição Aprovado

```
p1 <- df_g2 %>% ggplot(aes(x=Aprovado, y=Lucro)) +
  geom_boxplot(aes(fill=Aprovado)) +
  theme(legend.position="none") +
  scale_y_continuous(labels=unit_format(unit="M", scale=1e-6)) +
  labs(
    x="Aprovado",
    y="Lucro",
    title="Gráfico de Boxplot",
    subtitle="Lucro em função de aprovação"
  )

p2 <- df_g2 %>% ggplot(aes(x=Aprovado, y=Lucro)) +
  theme(legend.position="none") +
  geom_jitter(aes(colour=Aprovado)) +
  scale_y_continuous(labels=unit_format(unit="M", scale=1e-6)) +
  labs(
    x="Aprovado",
    y="Lucro",
    title="Stripchart",
    subtitle="Lucro em função de aprovação"
  )

ggarrange(p1, p2, ncol=2)
```

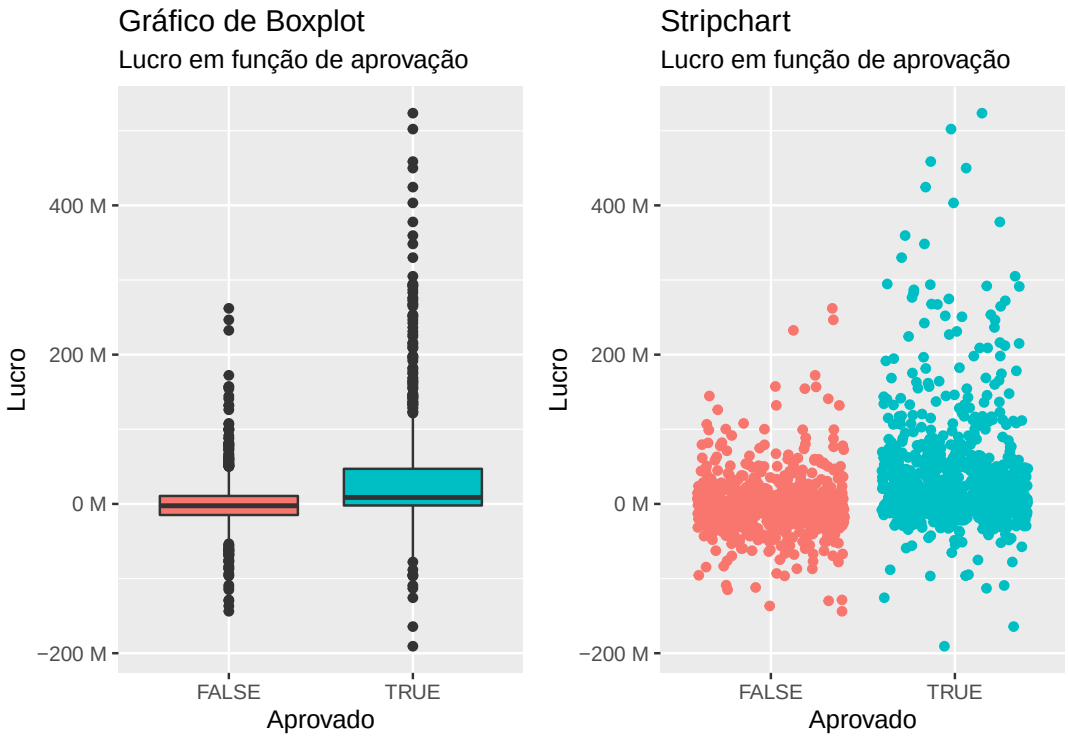

Figure 1: Boxplot e Stripchart para o lucro dos filmes em função da aprovação.

Pelos gráficos da Figura 1, é possível observar diferenças de lucro em função da aprovação de um filme, visto que a dispersão vertical dos dados em função da aprovação é diferente, mais visível no *Stripchart*. Nota-se, a

partir do *Boxplot*, que a tendência central (mediana) de cada conjunto é parecida e fica em torno de 0, com uma pequena vantagem para os filmes com aprovação. Além disso, é possível notar que a distribuição de filmes reprovados está mais concentrada em torno da mediana do que a distribuição de filmes aprovados. Isso é notado pelo tamanho da caixa, que é o IIQ (intervalo interquartil) entre o primeiro e o terceiro quartil, representando 50% de uma distribuição normal, e pelo tamanho dos *whiskers*, que são os limites superiores ($Q3 + 1.5 \times IIQ$) e inferiores ($Q1 - 1.5 \times IIQ$) e representam $\approx 99,3\%$ de uma distribuição normal. Estes valores de porcentagem não são exatos para este conjunto, mas a aproximação pelos valores da distribuição normal é rasoável dada a análise da Seção 9. Os espaços entre as diferentes partes da caixa indicam o grau de dispersão, a obliquidade nos dados. Observando isso e os *outliers*, é possível notar visualmente que as distribuições possuem variâncias diferentes.

7 Teste de igualdade de variâncias: Teste-F

```
var.test(
  x=df_g2[df_g2$Aprovado==TRUE,]$Lucro,
  y=df_g2[df_g2$Aprovado==FALSE,]$Lucro
)
```

```
##
##  F test to compare two variances
##
## data:  df_g2[df_g2$Aprovado == TRUE, ]$Lucro and df_g2[df_g2$Aprovado == FALSE, ]$Lucro
## F = 3.9333, num df = 842, denom df = 793, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  3.428333 4.511308
## sample estimates:
## ratio of variances
##            3.93328
```

Descrição dos campos retornados pela função:

data são os dados amostrais utilizados para o teste de variância, o primeiro valor apresentado representa o valor de x e o segundo o valor de y.

F é a o resultado principal do teste-F e representa a razão entre as variâncias, usada para distingir se as variâncias são iguais ou não. Representa $F_{CALC} = s_x^2/s_y^2 \approx 3.93$.

num df são os graus de liberdade do numerador (dado pelo conjunto x). Onde $df = n_1 - 1 = 842$.

denom df são os graus de liberdade do denominador (dado pelo conjunto y). Onde $df = n_2 - 1 = 793$.

p-valor é a probabilidade de entrar os resultados observados quando a hipótese nula (H_0) é verdadeira. Enquanto que o p-valor é uma probabilidade calcualda, α é uma probabilidade pré-determinada. Comparando o p-valor com α, que é o erro de tipo I (ou seja, a probabilidade de rejeitar H_0 dado H_0 verdadeira), é possível aceitar ou rejeitar a hipótese nula.

alternative hypothesis é a hipótese alternativa (H_1). Neste caso, H_0 é a hipótese das variâncias serem iguais (razão verdadeira $= 1$) e H_1 é a hipótese das variâncias serem diferentes (razão verdadeira $\neq 1$).

95 percent confidence interval é o intervalo de confidência a um nível de confiança de 95%. Mostrando que o valor de F_{CALC}, para um nível de confiança de 95%, está no intervalo IC = [3.428333, 4.511308].

sample estimates são as estimativas para a amostra. Neste teste, a única estimativa retornada é a **F**.

A partir do resultado deste teste, é possível conclcuir que a variância conjunto dos filmes aprovados é de 3.43 a 4.51 vezes maior que a variância do conjunto formado pelos filmes não aprovados, a um nível de

significância de 95%. Como p-valor $< \alpha$ ($\alpha = 0.05$), a hipótese nula, H_0, deve ser rejeitada. Alternativamente, é possível rejeitar H_0 comparando F_{CALC} do teste-F contra F_{CRIT} da distribuição F-Snedecor, visto que $F_{CALC} > F_{CRIT}$. F_{CRIT} é calculado pela Listagem a seguir.

```
F_crit <- qf(
  p=0.05,
  df1=length(df_g2[df_g2$Aprovado==TRUE,]$Lucro) - 1,
  df2=length(df_g2[df_g2$Aprovado==FALSE,]$Lucro) - 1,
  lower.tail=FALSE
)
F_crit
```

```
## [1] 1.1222
```

8 Teste de igualdade de médias: Teste-t

```
t.test(
  x=df_g2[df_g2$Aprovado==TRUE,]$Lucro,
  y=df_g2[df_g2$Aprovado==FALSE,]$Lucro,
  var.equal=FALSE
)
```

```
##
##  Welch Two Sample t-test
##
## data:  df_g2[df_g2$Aprovado == TRUE, ]$Lucro and df_g2[df_g2$Aprovado == FALSE, ]$Lucro
## t = 11.622, df = 1260.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  28423190 39967902
## sample estimates:
##   mean of x   mean of y
## 34103706.08   -91839.63
```

```
mean(df_g2[df_g2$Aprovado==TRUE,]$Lucro) - mean(df_g2[df_g2$Aprovado==FALSE,]$Lucro)
```

```
## [1] 34195546
```

Descrição dos campos retornados pela função:

data são os dados amostrais utilizados para o teste-t.

t é o resultado principal do teste-t e representa o valor de t_{CALC}.

df é o grau de liberdade da estatística para variâncias diferentes.

p-value é descrito na Seção 7.

alternative hypothesis é a hipótese alternativa (H_1). Neste caso, H_0 é a hipótese do verdadeiro valor das médias serem iguais (diferença = 0) e H_1 é a hipótese do valor verdadeiro das médias serem diferentes.

95 percent confidence interval é o intervalo de confidência da estatística a um nível de confiança de 95%. Mostrando que o valor de t_{CALC}, para um nível de confiança de 95%, está no intervalo IC = [28423190, 39967902].

sample estimates são as estimativas para a amostra.

9 Inspeção visual da normalidade: Papel de probabilidade normal

```
fit <- fitdistr(df_g2$nota_imdb, "normal")
MI <- fit$estimate[1]
SIGMA <- fit$estimate[2]

plot3 <- df_g2 %>% ggplot(aes(x=nota_imdb)) +
  geom_histogram(
    aes(y=..density..),
    bins=10,
    colour="black",
    fill="tan1"
  ) +
  labs(
    x="Nota imdb",
    y="Densidade de frequência",
    title="Histograma de densidade\nde frequência",
    subtitle="Nota IMDB"
  ) +
  stat_function(
    fun=dnorm,
    args=list(mean=MI, sd=SIGMA),
    size=1,
    col="black",
    linetype=1
  )

plot4 <- df_g2 %>% ggplot(aes(sample=nota_imdb)) +
  geom_qq_line(colour="darkorange4") +
  geom_qq(shape=21, colour="tan4", fill="tan1", size=3) +
  labs(
    x="Quantis teóricos",
    y="Quantis amostrais",
    title="Papel de probabilidade normal",
    subtitle="Nota IMDB"
  )

ggarrange(plot3, plot4, ncol=2)
```

Os gráficos da Figura 2 ajudam na análise visual sobre a aderência do conjunto à distribuição normal. No primeiro painel, é mostrado o histograma de densidade de frequência juntamente com a sobreposição da PDF $\varphi \sim \mathcal{N}(\mu, \sigma^2)$, onde μ e σ são, respectivamente, a média e o desvio padrão do conjunto formado pelas notas do IMDB. No segundo painel, é mostrado o papel de probalidade normal, onde os dados do conjunto são representados pelos círculos e a distribuição normal pela reta. No eixo horizontal, é mostrado os quantis normais teóricos, sendo que cada quantil é equivalente à um desvio padrão da média.

Como pode ser observado no papel de distribuição normal, os dados são bem próximos da distribuição normal para quantis normais entre -2 e 2. E, considerando que, em uma distribuição normal, aproximadamente 95% das ocorrências estão entre -2σ a 2σ, é possivel identificar que, na região central, a distribuição do conjunto tende à normal.

Em relação aos extremos, para os quantis normais maiores que 2 (ou seja, dois desvios padrão acima da média), as notas no IMDB estão abaixo da linha, o que indica que a cauda direita possui uma frequência muito baixa quando comparada com a distribuição normal. Isso pode ser identificado na última classe do

histograma. Para quantis normais que estão abaixo de -2 (ou seja, 2 desvios padrão abaixo da média), as notas no IMDB também estão abaixo da linha, o que indica que a cauda esquerda possui maior frequência frequência quando comparada com a distribuição normal. Isso poder ser identificado nas primeiras classes do histograma.

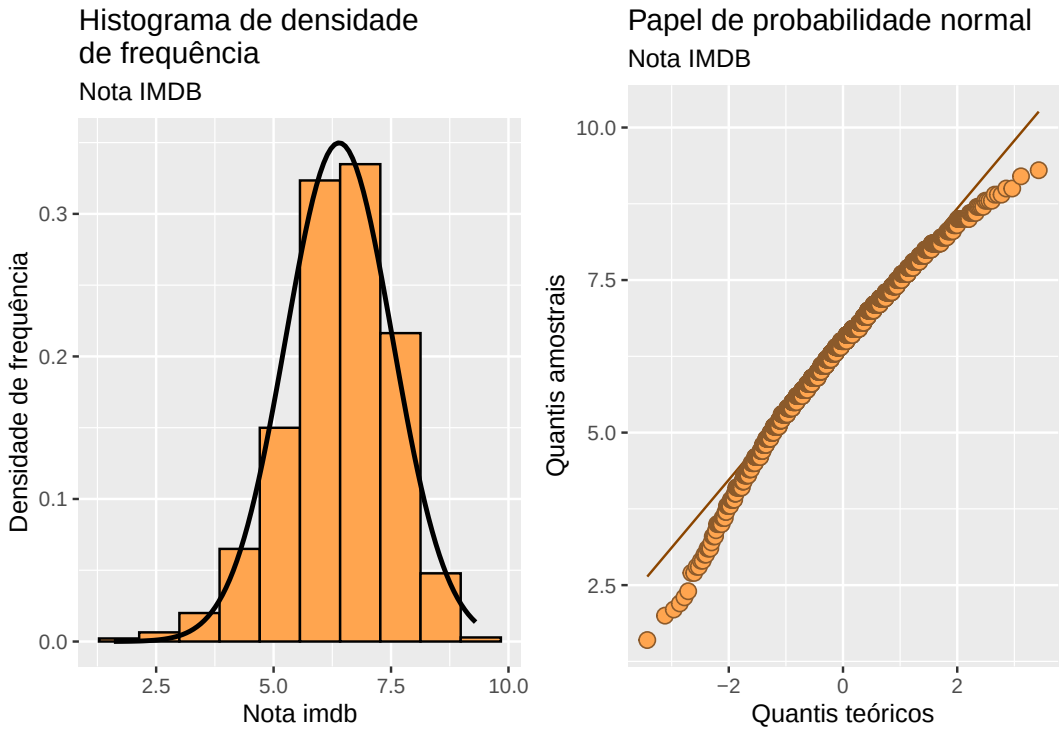

Figure 2: Da esquerda para direita, o primeiro painel mostra o histograma de densidade de frequência e o segundo painel mostra o papel de probabilidade normal.

10 Teste de normalidade: Teste Kolmogorov-Smirnov

Kolmogorov-Smirnov é um teste de aderência de distribuições de probabilidade contínuas e unidimensionais que pode ser usado para comparar uma amostra com uma distribuição de probabilidade de referência (teste K–S uniamostral) ou duas amostras uma com a outra (teste K–S biamostral). Neste caso, será feito um teste uniamostral entre a amostra de notas do IMDB e uma distribuição normal $\Phi \sim \mathcal{N}(\mu, \sigma^2)$, onde μ e σ são, respectivamente, a média e o desvio padrão da amostra de notas do IMDB.

```
x = df_g2$nota_imdb
KS_teste <- ks.test(x, "pnorm", mean=mean(x), sd=sd(x))
KS_teste
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  x
## D = 0.052888, p-value = 0.0002108
## alternative hypothesis: two-sided
```

Descrição dos campos retornados pela função

data é o conjunto de dados de entrada

D é a variável de teste e representa $D_{CALC} := \max|\Phi_A(x) - \Phi_B(x)|$, onde Φ_A e Φ_B são duas CDF's.

p-value já foi definido na Seção 7.

H_0 é a hipótese dos dados aderirem ao modelo e H_1 é a hipótese de não aderirem ao modelo. Para um nível de confiança de 95%, o valor de α vale 0.025. Sendo assim, evidências amostrais para rejeitar H_0, visto que p-valor $< \alpha$. Isto é, o teste K-S nesta amostra indica que os dados aderem à distribuição normal, a um nível de significância de $\approx$ 99% (p-valor). Alternativamente, é possível obter o mesmo resultado, à um nível de significância de 95%, por exemplo, comparando o valor de D_{CALC} e D_{CRIT}, como mostra a Listagem a seguir. Já que $D_{CRIT} > D_{CALC}$

```
D_crit <- 1.63 / sqrt(length(x))
D_crit
```

```
## [1] 0.04028685
```