Teste Prático

André Filipe Gomes Silvestre

2022-11-06

GRUPO A (6 valores)

Devido às condições económicas de um determinado país, onde a subida de preços está a retirar poder de compra às famílias, uma determinada associação conseguiu angariar o apoio de três grandes grupos de distribuição (A, B e C) para fornecimento de frescos (a integrar cabazes básicos que serão entregues pela associação às famílias do programa de apoio).

Um dos voluntários está encarregue de monitorizar a qualidade dos pacotes de produtos frescos entregues, e classifica os pacotes de produtos em duas categorias: 1 - para entrega às famílias e 0 - para compostagem.

Assim, cada pacote pode ser classificado segundo duas vertentes: Origem (A, B ou C) e Estado (1 ou 0).

Considere então a experiência aleatória que consiste em observar um pacote de frescos e proceder a esta dupla classificação (origem, estado).

a)

Construa um dataframe que contenha todos os possíveis resultados da experiência aleatória descrita. (Mostre o df)

b)

Do histórico, sabe-se que as proporções de pacotes de frescos em condições de serem entregues às famílias variam segundo o distribuidor. Os registos do mês anterior apontam para **pe_a%**, **pe_b%** e **pe_c%** as percentagens de pacotes enviados para compostagem, com origem em, respetivamente, A, B ou C.

Sabe-se ainda que o grupo C é responsável por 20% do apoio à associação, sendo o restante apoio dado, em partes iguais, pelos grupos A e B.

Com base nesta informação, adicione ao dataframe que construiu na alínea anterior uma coluna, "prob", com a probabilidade de ocorrência de cada um dos resultados. (mostre o df)

c)

Obtenha uma simulação de **nreplica** observações da experiência aleatória descrita, que respeite as condições de ocorrência indicadas.

Obtenha a tabela de classificação cruzada (origem, estado) para essa simulação, quer com frequências absolutas, quer com frequências relativas.

```r
# 1.
nreplica <- 104532

# 2.
pe_a <-2
pe_b <-3
pe_c <-5

# 3.
npac <- 104532 %/% 2000

# 4.
nvis <- ceiling(1.5*104532)
```

GRUPO A. Resolução

a)

Cada pacote pode ser classificado segundo duas vertentes: Origem (A, B ou C) e Estado (1 ou 0).

Logo,

```r
origem <- c("A","B","C")
estado <- c("1","0")

# Resultados Possíveis da experiência aleatória descrita
esp_res <- expand.grid(origem,estado)
origem_estado<- data.frame(esp_res)
```

b)

- *Do histórico, sabe-se que as proporções de pacotes de frescos em condições de serem entregues às famílias variam segundo o distribuidor. Os registos do mês anterior apontam para* **pe_a%**, **pe_b%** e **pe_c%** *as percentagens de pacotes enviados para compostagem, com origem em, respetivamente, A, B ou C.*
- *Sabe-se ainda que o grupo C é responsável por 20% do apoio à associação, sendo o restante apoio dado, em partes iguais, pelos grupos A e B.*

Sabendo que o **Teorema da Probabilidade Total** é:

$$P\left[B\right] = \sum_{i=1}^{n} P\left[B|A_i\right] \times P\left[A_i\right] = \sum_{i=1}^{n} P\left[B \cap A_i\right]$$

e aplicando a **Fórmula de Bayes**,

$$P\left[A_j|B\right] = \frac{P\left[B|A_j\right] \times P\left[A_i\right]}{\sum_{i=1}^{n} P\left[B|A_i\right] \times P\left[A_i\right]} = \frac{P\left[B \cap A_j\right]}{P\left[B\right]}$$

Com base nesta informação, o dataframe com a coluna "prob", com a probabilidade de ocorrência de cada um dos resultados e dada por:

```r
p_A <- (1 - 0.2) /2
p_B <- (1 - 0.2) /2
p_C <- 0.2

p_A_sabendo_0 <- pe_a/100
p_B_sabendo_0 <- pe_b/100
p_C_sabendo_0 <- pe_c/100

origem_estado$prob <- c(p_A -(p_A*p_A_sabendo_0), p_B -(p_B*p_B_sabendo_0), p_C -(p_C*p_C_sabendo_0), p
origem_estado
```

```
##   Var1 Var2  prob
## 1    A    1 0.392
## 2    B    1 0.388
## 3    C    1 0.190
## 4    A    0 0.008
## 5    B    0 0.012
## 6    C    0 0.010
```

c)

Obtenha uma simulação de **nreplica** observações da experiência aleatória descrita, que respeite as condições de ocorrência indicadas.

Obtenha a tabela de classificação cruzada (origem, estado) para essa simulação, quer com frequências absolutas, quer com frequências relativas.

```r
# library(crosstable)
# library(dplyr)

# Simulação
# Origem <- sample(origem, nreplica, replace = TRUE, prob = c(p_A ,p_B ,p_C))
# Estado <- sample(estado, nreplica, replace = TRUE, prob = c(sum(origem_estado$prob[4:6]),
#                                                             sum(origem_estado$prob[1:3])))

# simul <- data.frame(Origem, Estado)

# crosstable(simul, Estado, by=Origem) %>%
#   as_flextable(keep_id=FALSE)

# ------------------------------- Correção TC ----------------------------------
linhas<-sample(1:nrow(origem_estado),size=nreplica,replace=TRUE,prob=origem_estado$prob)
simul<-origem_estado[linhas,]
tab1<-xtabs(~simul$Var1+simul$Var2)
knitr::kable(tab1, format="markdown", digits=3)

knitr::kable(round(prop.table(tab1)*100,1), format="markdown", digits=3)
```

GRUPO B (6 valores)

Uma determinada empresa é especialista na recolha de preços. Esta empresa recolhe preços e tem dois serviços principais: divulgação de preços para comparações online e fornecimento de pacotes de preços a especialistas que analisam a inflação.

Para a divulgação de preços para comparações online, a empresa usa um site.

Sabe-se que o valor que a empresa recebe, em euros, por cada visita ao site segue uma distribuição normal com valor médio 0.5 € e variância 0.2.

Já relativamente aos pacotes de preços fornecidos a especialistas, a empresa sabe que o valor que recebe por cada pacote fornecido, em euros, segue uma distribuição normal de valor médio 500 e variância 100.

Considere que em certo mês, m, a empresa tem **nvis** visitas e vende **npac** pacotes de preços.

a.

Calcule os parâmetros caracterizadores da variável Xm – valor recebido no mês m. Represente graficamente a função densidade de X_m.

b.

Calcule o quantil de probabilidade 0.9 dessa distribuição. Represente a área em causa no gráfico.

c.

Gere **nreplica** observações aleatórias da variável Xm e use essa simulação para obter um valor aproximado para o quantil referido na alínea anterior.

GRUPO B. Resolução

Seja

$X-$ divulgação de preços para comparações online

X tem valor médio 0.5 € e variância 0.2.

$$X \sim N(\mu = 0.5, \sigma = \sqrt{0.2})$$

$Y-$fornecimento de pacotes de preços a especialistas que analisam a inflação

Y tem valor médio 500 e variância 100

$$Y \sim N(\mu = 500, \sigma = \sqrt{100} = 10)$$

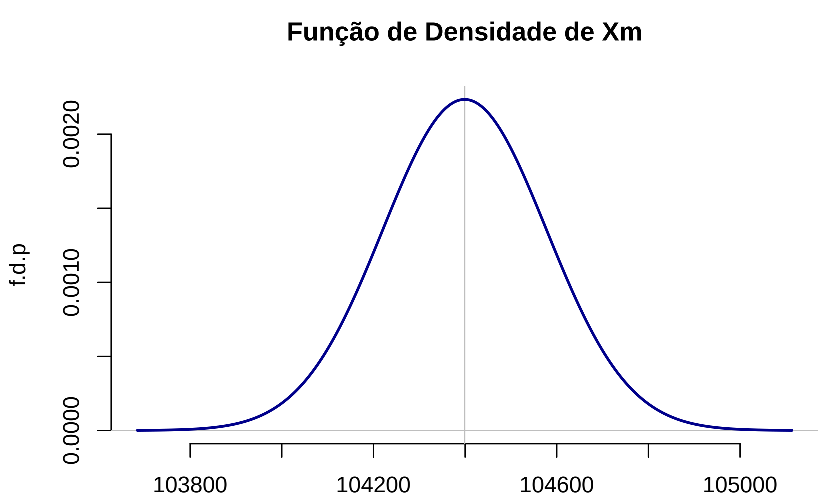
a)

Considere que em certo mês, m, a empresa tem *nvis* visitas e vende *npac* pacotes de preços.

Os parâmetros caracterizadores da variável $X_m \breve{}$ valor recebido no mês m. Represente graficamente a função densidade de X_m.

Pelo **TAN** conseguimos deduzir X_m

```r
miuXm <- nvis*0.5 + npac*500
sigmaXm <- sqrt(nvis*0.2 + npac*10)

eixo_x<-c(miuXm - 4* sigmaXm, miuXm + 4* sigmaXm)
eixo_y<-c(0,dnorm(miuXm,miuXm,sigmaXm))

# Preparar o Espaço
plot(1,
     xlim = eixo_x, ylim = eixo_y,
     type = "n",
     main = "Função de Densidade de Xm",
     ylab = "f.d.p", xlab = "",frame.plot=FALSE)

# Add x and y-axis lines
abline(h = 0 , col="grey")
abline(v = miuXm, col="grey")

# Desenhar a Função
curve(dnorm(x,miuXm,sigmaXm),
      from = eixo_x[1], to = eixo_x[2],
      n = 1000,
      col = "darkblue",
      lwd = 2,
      add=TRUE)
```



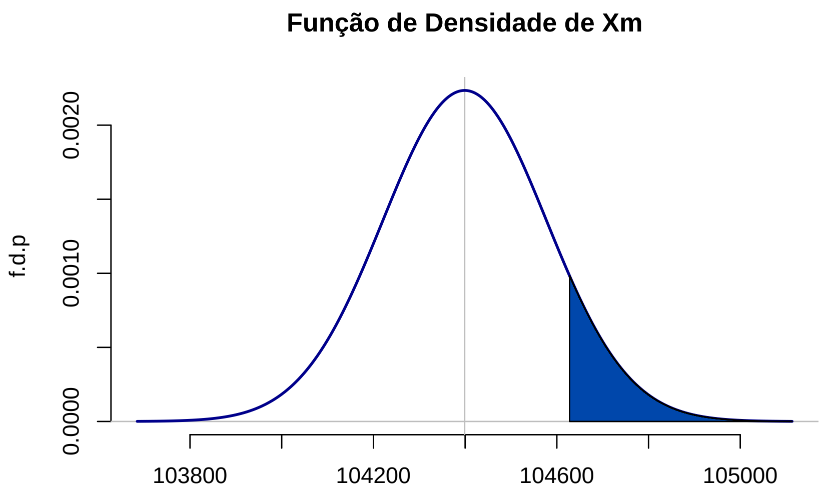
b.

Calcule o *quantil de probabilidade 0.9* dessa distribuição. Represente a área em causa no gráfico.

```
qnorm_0.9 <- qnorm(0.9,miuXm,sigmaXm)
qnorm_0.9
```

```
## [1] 104627.8
```

```
# Sombrear
miuXm <- nvis*0.5 + npac*500
sigmaXm <- sqrt(nvis*0.2 + npac*10)

eixo_x<-c(-4,4)* sigmaXm + miuXm
eixo_y<-c(0,dnorm(miuXm,miuXm,sigmaXm))

# Preparar o Espaço
plot(1,
     xlim = eixo_x, ylim = eixo_y,
     type = "n",
     main = "Função de Densidade de Xm",
     ylab = "f.d.p", xlab = "",frame.plot=FALSE)

# Add x and y-axis lines
abline(h = 0 , col="grey")
abline(v = miuXm, col="grey")

# Desenhar a Função
curve(dnorm(x,miuXm,sigmaXm),
      from = eixo_x[1], to = eixo_x[2],
      n = 1000,
      col = "darkblue",
      lwd = 2,
      add=TRUE)

# Sombrear a área
x1 <- seq(qnorm_0.9,eixo_x[2],0.01)
y1 <- dnorm(x1,miuXm,sigmaXm)

coord_x <- c(qnorm_0.9,x1,eixo_x[2])
coord_y <- c(0,y1,0)

polygon(coord_x,coord_y,col='#0047ab',border = NULL)
```



c.

Gere *nreplica* observações aleatórias da variável X_m e use essa simulação para obter um valor aproximado para o quantil referido na alínea anterior.

```
simul_1 <- rnorm(nreplica,miuXm,sigmaXm)

quantile(simul_1,0.9)
```

```
##      90%
## 104628.4
```

GRUPO C (5 valores)

Considere os dados em "Estudo_Oculos_Sol.rds", utilizados nos últimos TPC.

a.

Obtenha uma coluna adicional no dataframe que tenha o valor "Sup" caso o nível educacional seja "Tertiary" e "No Sup" caso contrário.

b.

Teste se os dois grupos acima definidos diferem quanto à importância concedida, em termos médios, à Qualidade dos óculos de sol.

c.

Teste se existe relacionamento entre ter ou não nível de educação superior (a variável que criou em a) e a possibilidade de vir a comprar óculos RB (will_buy_RB). Caso o relacionamento seja significativo, obtenha uma representação gráfica adequada.

GRUPO C. Resolução

Considerando a base de dados "Estudo_Oculos_Sol.rds", utilizados nos últimos TPC.

```r
# Leitura do ficheiro Estudo_Oculos_Sol.rds
bd_oculos_sol <-readRDS("Estudo_Oculos_Sol.rds")
```

a.

Obtenha uma coluna adicional no dataframe que tenha o valor "Sup" caso o nível educacional seja "Tertiary" e "No Sup" caso contrário.

```r
# a)
# library(dplyr)
#
# bd_oculos_sol <- bd_oculos_sol %>%
#   mutate(exercicio_a = ifelse(bd_oculos_sol$educ == "Tertiary", "Sup", "NoSup"))

# bd_oculos_sol

# ------------------------------- Correção TC ----------------------------------
bd_oculos_sol$exercicio_a<-ifelse(bd_oculos_sol$educ == "Tertiary","Sup","NoSup")
```

b. Teste de Hipóteses Paramêtrico

Teste se os dois grupos acima definidos diferem quanto à importância concedida, em termos médios, à Qualidade dos óculos de sol.

podemos definir as hipóteses como:

- $H_0 : \mu_{sup} \ = \mu_{no\ sup} \ \ \Leftrightarrow \ \ mu_{sup} - \mu_{no\ sup} = 0$ (hipótese nula)
- $H_1 : \mu_{sup} \neq \mu_{no\ sup} \ \ \Leftrightarrow \ \ mu_{sup} - \mu_{no\ sup} \neq 0$ (hipótese alternativa)

Pelo que, sendo $n_1 + n_2 = 640 > 30$ e não conhecendo σ^2, mas assumindo que são iguais, a VF a usar é

$$VF = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\sqrt{\frac{(n_1-1)s_1'^2 + (n_2-1)s_2'^2}{n_1+n_2-2}}} \sim t_{(n_1+n_2-2)}$$

```r
# b)
significancia = 0.05

# -------------------------- Através do t.test --------------------------
teste2 <- t.test(bd_oculos_sol$Quality ~ bd_oculos_sol$exercicio_a,
                 alternative = "two.sided",
                 mu = 0,
                 conf.level = 1- significancia,
                 var.equal = TRUE)
teste2
```

```
##
##  Two Sample t-test
##
## data:  bd_oculos_sol$Quality by bd_oculos_sol$exercicio_a
## t = -5.6429, df = 634, p-value = 2.522e-08
## alternative hypothesis: true difference in means between group NoSup and group Sup is not equal to 0
## 95 percent confidence interval:
##  -0.5765381 -0.2788620
## sample estimates:
## mean in group NoSup   mean in group Sup
##            7.607728            8.035428
```

c. Teste do Qui-Quadrado (χ^2)

Teste se existe relacionamento entre ter ou não nível de educação superior (a variável que criou em a) e a possibilidade de vir a comprar óculos RB (*will_buy_RB*). Caso o relacionamento seja significativo, obtenha uma representação gráfica adequada.

Hipóteses em teste + Estatística de teste

X - Educação (variável *educ*)

Y - Poder vir a comprar *SoleMio* (variável *Will_buy_RB*)

Hipóteses em teste

H_0 : O nível de educação é independente da possibilidade de vir a comprar óculos RB

H_1 : Existe relacionamento entre ambos

ou, teoricamente,

$H_0 : \forall(i,j) \in \{1:r\} \times \{1:c\} : p_{ij} = p_{i.} \times p_{.j}$

$H_1 : \exists(i,j) \in \{1:r\} \times \{1:c\} : p_{ij} \neq p_{i.} \times p_{.j}$

Teste Qui-Quadrado

Estatística de teste

$$ET = \sum_{i=1}^{3} \sum_{j=1}^{3} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} \dot{\sim} \chi^2_{(r-1)(c-1)}$$

Sendo ambas variáveis fatores, o estudo do relacionamento entre elas será feito através da análise do respetivo cruzamento (tabela de contingência), com a subsequente aplicação do **Teste Qui-quadrado de Pearson**.

```r
tab1<-table(bd_oculos_sol$educ, bd_oculos_sol$Will_buy_RB)  # crosstabs freq abs

teste<-chisq.test(tab1)
teste
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 12.493, df = 6, p-value = 0.05183
```

Como $p - value = 0.0518271 > \alpha$ de referência ($\alpha = 0.05$), então não se rejeita a H_0.

Logo, não existem divergências significativas entre as frequências observadas e as frequências esperadas (ou seja, o que esperaríamos observar numa situação de independência).

Correção TC

```
tab1<-table(bd_oculos_sol$educ, bd_oculos_sol$exercicio_a)  # crosstabs freq abs

teste_certo<-chisq.test(tab1)
teste_certo
```

```
##
##  Pearson's Chi-squared test
##
## data:  tab1
## X-squared = 636, df = 3, p-value < 2.2e-16
```

- Utilizando a variável que criou em a)

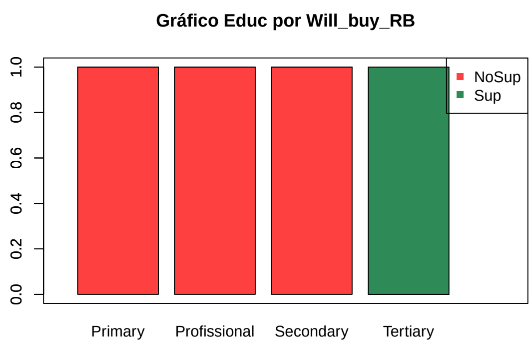
  Como $p - value = 1.580173 \times 10^{-137} < \alpha$ de referência ($\alpha = 0.05$), então rejeita-se a H_0.

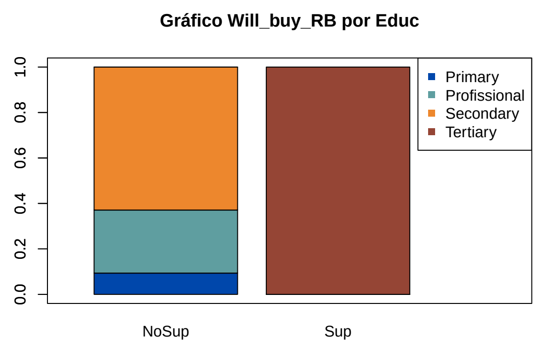
Logo, existem divergências significativas entre as frequências observadas e as frequências esperadas (ou seja, o que esperaríamos observar numa situação de independência).

Estas divergências podem ser observadas pelos seguintes gráficos

```
cores<-c('#ff4040','#2e8b57')

# preparar a área
plot(1,
     xlim = c(0,2.5), ylim = c(0,1),
     type = "n",                          # vazio
     main = "Gráfico Educ por Will_buy_RB",     # título
     ylab = "", xlab = "",                # sem nomes
     xaxt = "n")                          # sem marcas eixo x


barplot(tab1_byEduc_col,              # att dados org em colunas
        col =cores,                   # cores a usar
        width = 0.45,                  # largura das barras
        add=TRUE)                     # para dar espaço p legenda

legend("topright",
       legend = rownames(tab1_byEduc_col),
       pch = 15,
       col = cores)

# Representação gráfica - Will_buy_RB por Educ ------------------------------------
# preparar a área
tab1_by_Will_buy_RB <- prop.table(tab1,margin = 2)

cores<-c("#0047ab","#5f9ea0","#ed872d", "#954535")
plot(1,
     xlim = c(0,2.5), ylim = c(0,1),
     type = "n",                        # vazio
     main = "Gráfico Will_buy_RB por Educ",   # título
```

```
     ylab = "", xlab = "",                  # sem nomes
     xaxt = "n")                            # sem marcas eixo x


barplot(tab1_by_Will_buy_RB,                # att dados org em colunas
        col =cores,                         # cores a usar
        width = 0.8,                        # largura das barras
        # para dar espaço p legenda
        add=TRUE)

legend("topright",
       legend = rownames(tab1_by_Will_buy_RB),
       pch = 15,
       col = cores)
```





“‘