Estatística para ciência de dados

Code

Aluno: Eduardo Façanha Dutra

Matrícula: 2016473

Resolução do trabalho 02: Ex-Teste-usabilidade-Gera-Graficos - PP

1ª Questão: Considere que testadores analisaram na etapa "atende aos requisitos?" se um participante teria a idade requerida (entre 18 e 24 anos). Explique o que significaria, em um gráfico XY, se ter no eixo Y idade e no eixo X o perfil da amostra (número de indivíduos selecionados)?

Resposta:

Um gráfico XY da [idade requerida] necessitaria que no eixo x seja representada a [idade] do perfil da amostra, já que a variável [idade requerida] é dependente da variável [idade].

2ª Questão: Considere que todos os indivíduos selecionados irão realizar as mesmas atividades usando cada um dos sistemas. Pede-se:

```
1- Justifique a necessidade de "sortear atividade" para o exemplo em estudo.

2- Dê exemplos de variáveis que podem ser riscos para os resultados senão houvesse o sorteamento
das atividades.

3- Considere que, é necessário que os sistemas sejam analisados com relação à 8 funcionalidades.
Tecnicamente, isto significaria que cada indivíduo deveria fazer 8 atividades usando sistema. Ma
s, mesmo sorteando a ordem das atividades, os indivíduos se cansariam em ter que realizar todas
e o experimento poderia também trazer viés.  Como eles resolveriam este problema? Quais as impli
cações da(s) decisão(ões)?</b>
```

Resposta:

```
1- A distribuição aleatória de atividades por voluntário evita, ou mitiga, o efeito de viés nos
efeitos em observação.

2- Atividades dependentes: Alguma atividade pode afetar de alguma forma a execução da próxima at
ividade;
  Nível de dificuldade: Atividades mais complexas que são deixadas para o final podem ser execut
adas com menor rigor, devido ao cansaço ou pressa do voluntário;
3- Diminuição da complexidade das atividades e compensar a falta de profundidade das atividades
com o aumento no número de amostras, oferecer recompensas de alguma natureza aos voluntários, tr
abalhs em grupo.
```

3ª. Questão. Sobre a etapa "receber treinamento", considere que ela se refere à forma como um experimento é esperado ser conduzido pelos investigadores. Por exemplo, todos os participantes devem receber o mesmo treinamento sobre cada sistema. O treinamento consistiria no testador mostrar a utilidade e estrutura do sistema, antes do usuário usar o sistema. Depois o testador diria que passaria X atividades para o usuário realizar usando o sistema. Para o exemplo em estudo, dê exemplos de variáveis que podem ser riscos para os resultados se os indivíduos não receberem este treinamento.

Resposta:

O avaliador não terá a certeza de que os dados coletados serão confiáveis para se extrair alguma conclusão, o que comprometerá toda a pesquisa.

4ª. Questão. Para cada um dos 4 problemas dados a seguir pede-se:

a. **defina a hipótese para um modelo de pesquisa da preferência de usuários por um sistema interativo;**
b. **coloque a amostra coletada para o problema no excel;**
c. **Faça a análise dos dados, usando dois gráficos: boxplot (gráfico de caixa), que é um gráfico utilizado para avaliar a distribuição empírica dos dados) e histograma, que é um gráfico de frequência, que tem como objetivo ilustrar como uma determinada amostra ou população de dados está distribuída)**

1- Problema: analisar as preferências de usuários entre dois sistemas: prefsAB.cvs

Desenvolvimento:

a)Hipóteses:

“Há diferença na preferência entre os sistemas A e B?”

b)Leitura dos dados:

Hide

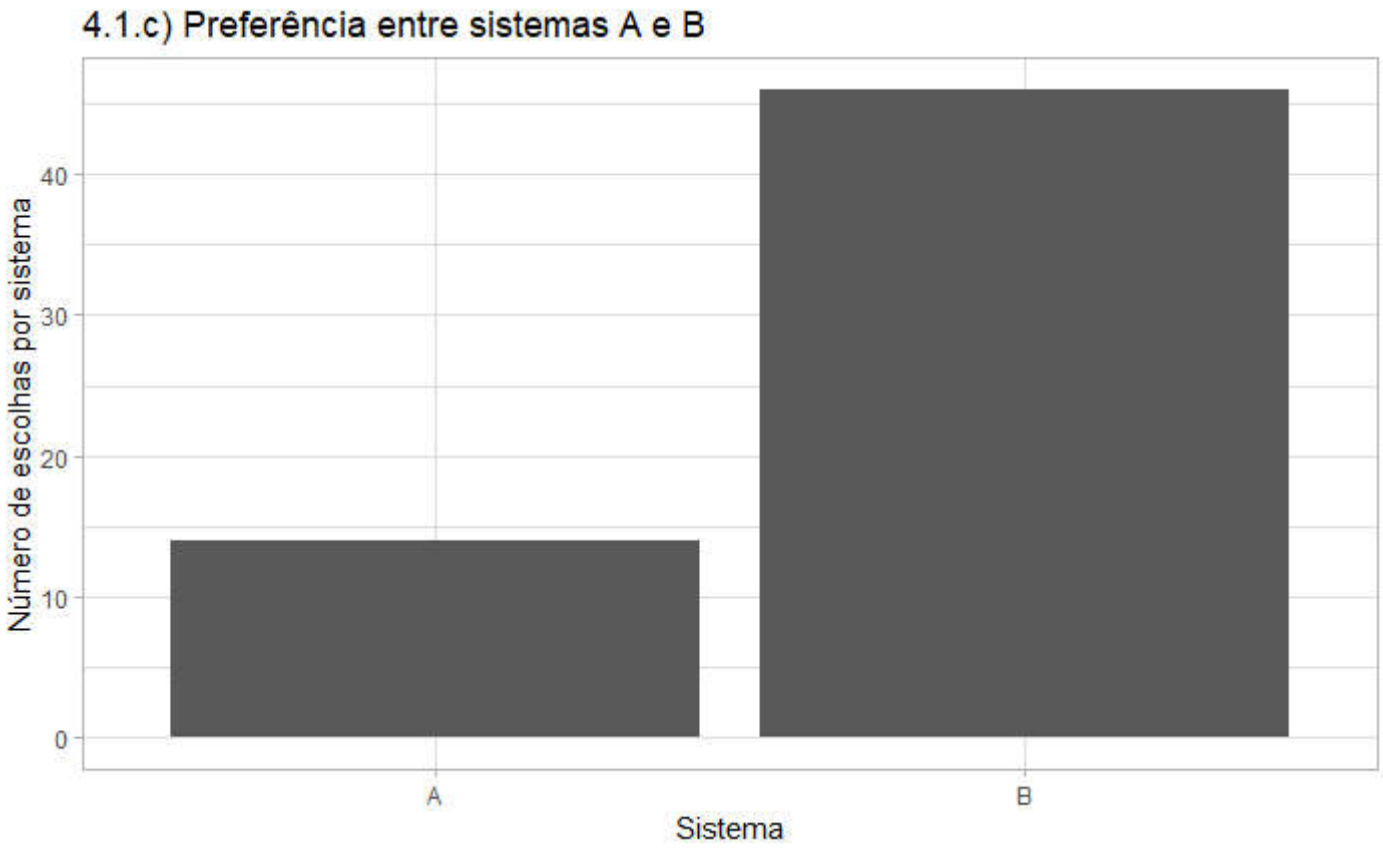
```
library(readr)
#install.packages("tidyverse") #<--- caso tidyverse não esteja instalado
library(tidyverse)
#A leitura dos dados é realizado excluindo-se a primeira coluna, pois se trata de um sequêncial
 sem significado para o problema em questão. Além disso é realizado o "factor" dos dados para qu
e seja possível realizar a contagem dos valores
prefsAB <- as_tibble(read_csv("Dados/prefsAB.csv",col_types = cols(Pref = col_factor(levels = c(
"A","B")), Subject = col_skip())))

summary(prefsAB)
```

```
 Pref
 A:14
 B:46
```

c)Geração de gráficos

Hide

```
library(ggplot2)
grafico4.1.c <- ggplot(prefsAB, aes(x=Pref))+  geom_bar() + theme_light()+ labs(x= "Sistema", y=
"Número de escolhas por sistema", title="4.1.c) Preferência entre sistemas A e B")


grafico4.1.c
```



2- Problema: analisar as preferências de usuários entre três sistemas: prefsABC.cvs

Desenvolvimento:

a)Hipóteses:

“Há diferença na preferência entre os sistemas A, B e C?”

b)Leitura dos dados:

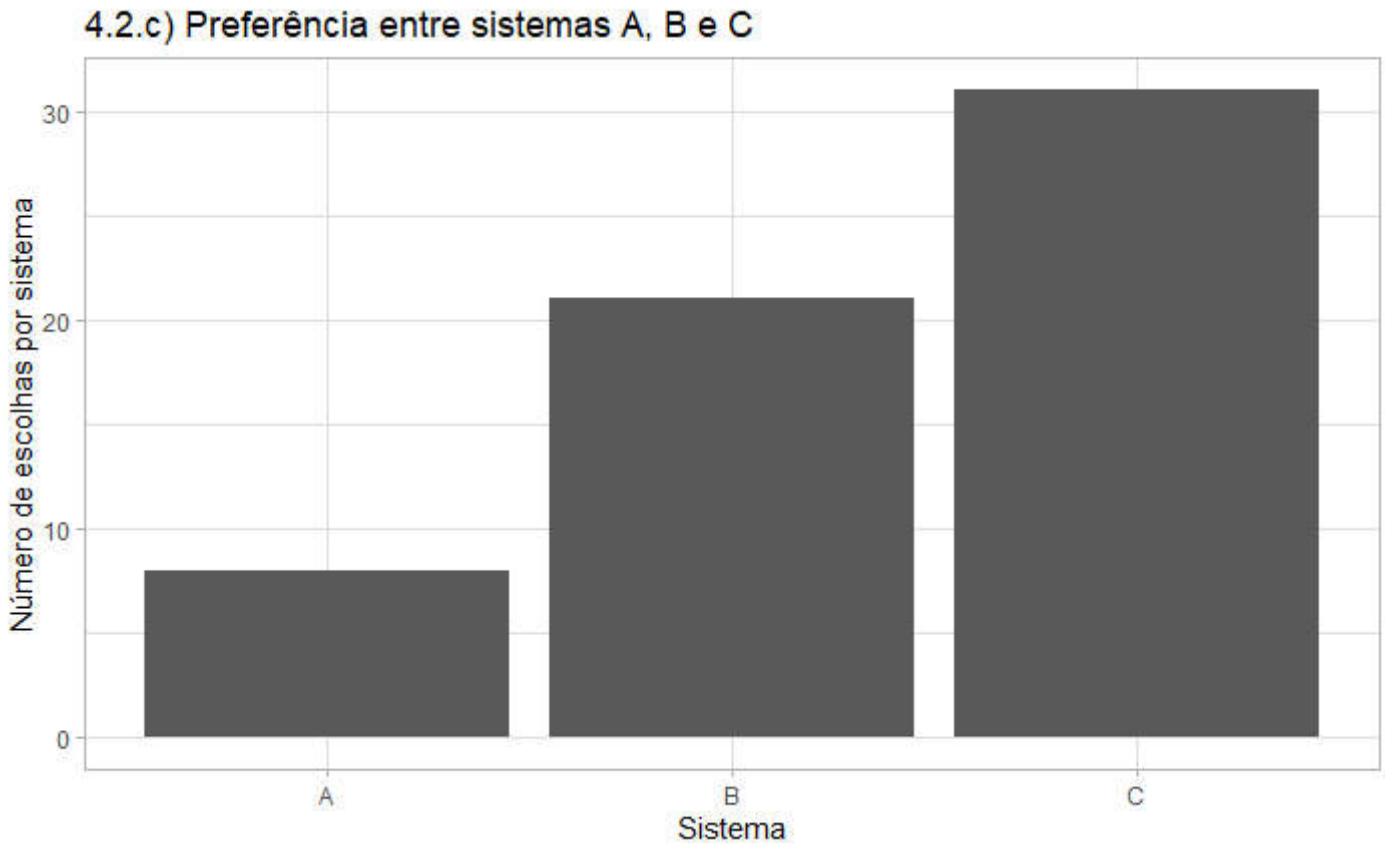
Hide

```
#A leitura dos dados é realizado excluindo-se a primeira coluna, pois se trata de um sequêncial
 sem significado para o problema em questão. Além disso é realizado o "factor" dos dados para qu
e seja possível realizar a contagem dos valores
prefsABC <- as.tibble(read_csv("Dados/prefsABC.csv",col_types = cols(Pref = col_factor(levels =
 c("A","B","C")), Subject = col_skip())))
```

c)Geração de gráficos

Hide

```
grafico4.2.c <- ggplot(prefsABC, aes(x=Pref))+  geom_bar() + theme_light()+ labs(x= "Sistema", y
="Número de escolhas por sistema", title="4.2.c) Preferência entre sistemas A, B e C")


grafico4.2.c
```



3- Problema: analisar as preferências de usuários, por gênero, entre dois sistemas Amostra é prefsABsex.cvs

Desenvolvimento:

a)Hipóteses:

“Há diferença na preferência do sistemas A por gênero?”

“Há diferença na preferência do sistemas B por gênero?”

b)Leitura dos dados:

Hide

```
#A leitura dos dados é realizado excluindo-se a primeira coluna, pois se trata de um sequêncial
 sem significado para o problema em questão. Além disso é realizado o "factor" dos dados para qu
e seja possível realizar a contagem dos valores
prefsABsex <- as.tibble(read_csv("Dados/prefsABsex.csv",
    col_types = cols(Pref = col_factor(levels = c("A",
        "B")),
        Sex = col_factor(levels = c("M", "F")),
        Subject = col_skip())))

#organização dos dados: agrupamento por preferência de sistema
tabelaPorPref.ABSex <- prefsABsex%>%
  group_by(Pref,Sex)%>%
  summarize(count = n())%>%
  spread(Sex, count)


print(tabelaPorPref.ABSex)
```

Pref <fctr>	**M** <int>	**F** <int>
A	12	2
B	17	29

2 rows

Hide

```
#organização dos dados: agrupamento por gênero do participante
tabelaPorSex.ABSex <- prefsABsex%>%
  group_by(Sex, Pref)%>%
  summarize(count = n())%>%
  spread(Pref, count)

print(tabelaPorSex.ABSex)
```

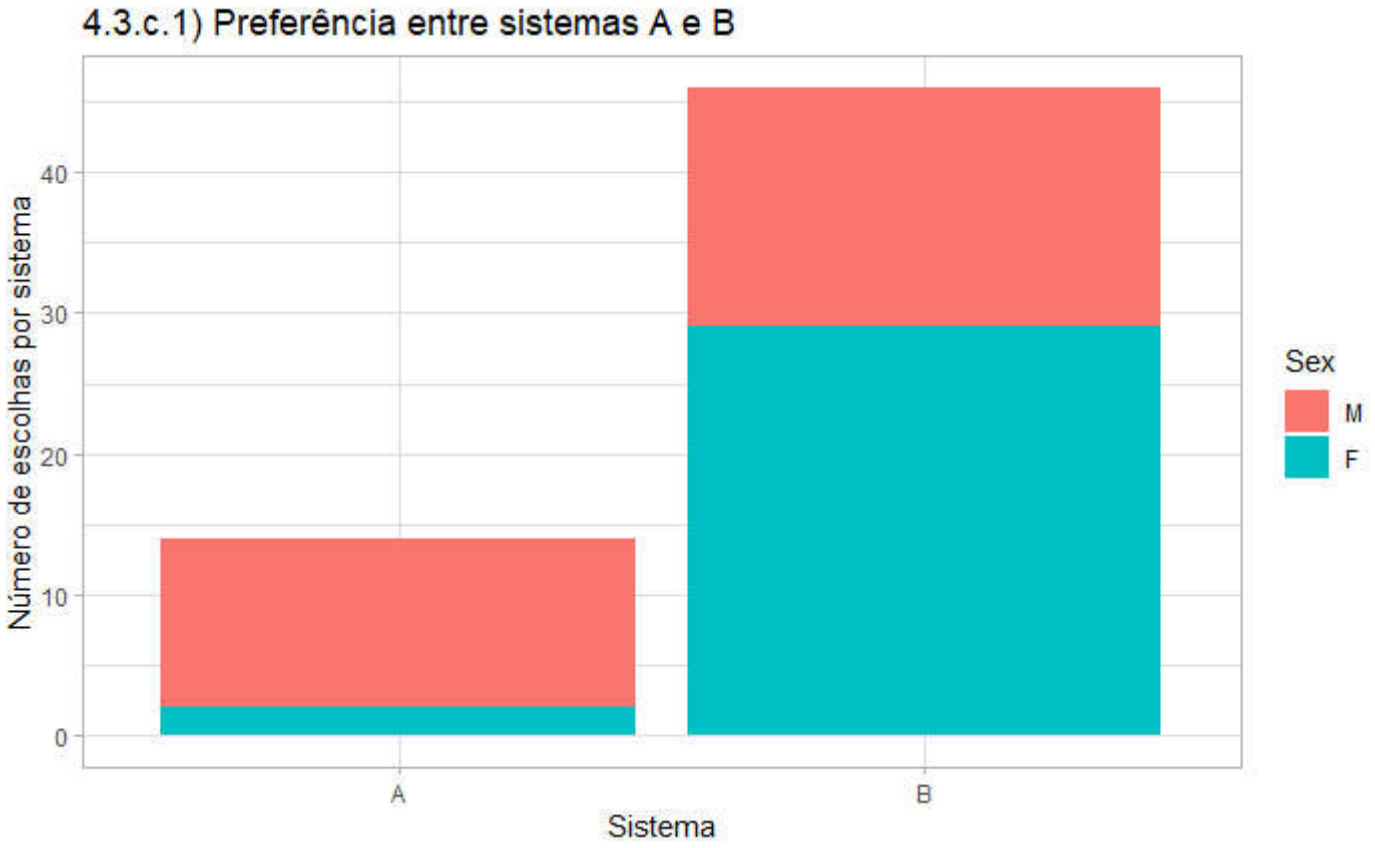
Sex <fctr>	**A** <int>	**B** <int>
M	12	17
F	2	29

2 rows

c)Geração de gráficos

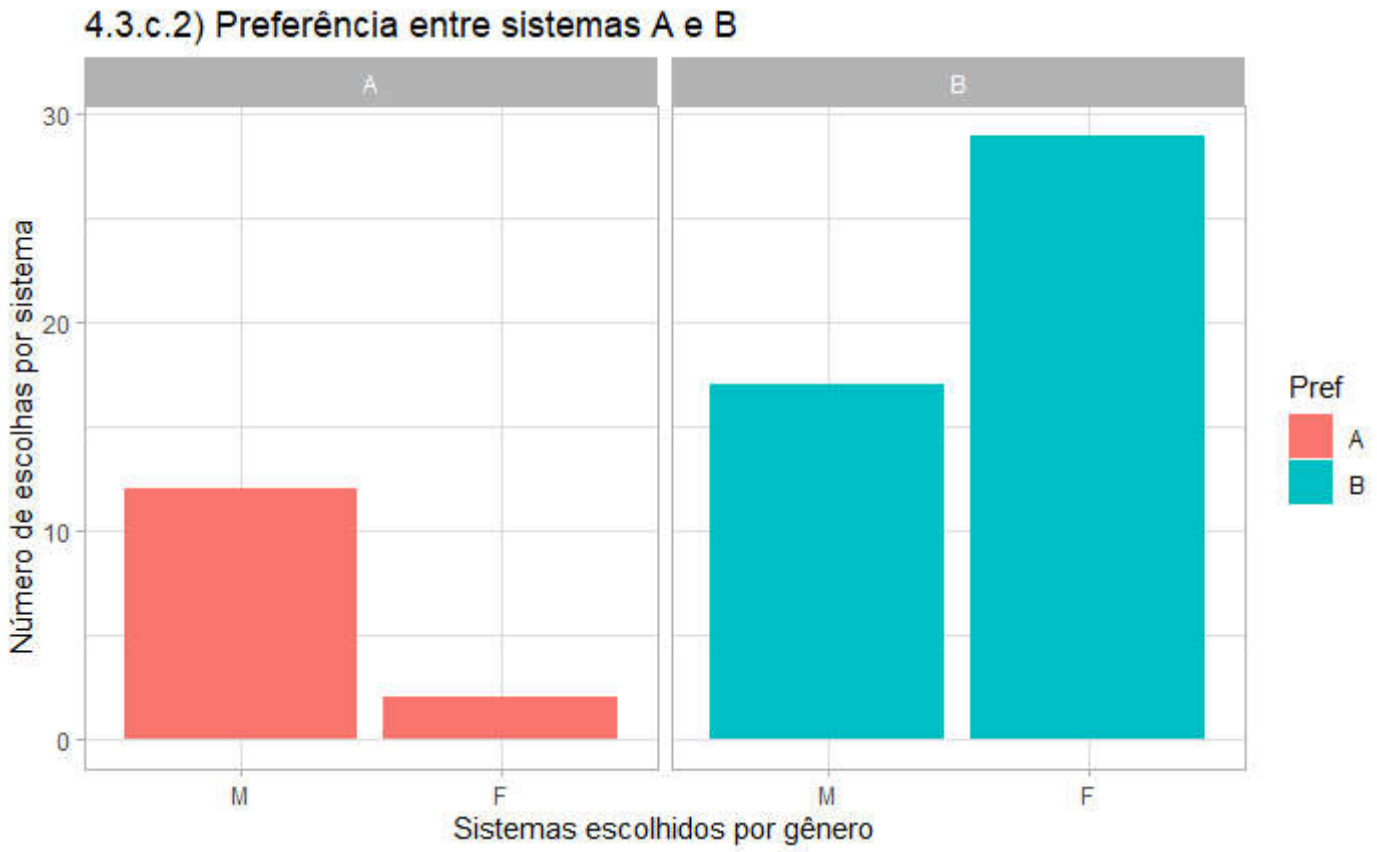
Hide

```
grafico4.3.c.1 <- ggplot(prefsABsex, aes(x=Pref, fill = Sex))+  geom_bar() + theme_light()+ labs
(x= "Sistema", y="Número de escolhas por sistema", title="4.3.c.1) Preferência entre sistemas A
 e B")


grafico4.3.c.2 <- ggplot(prefsABsex, aes(x=Sex, fill = Pref))+  geom_bar() + theme_light()+ labs
(x= "Sistemas escolhidos por gênero", y="Número de escolhas por sistema", title="4.3.c.2) Prefer
ência entre sistemas A e B")+facet_wrap(~Pref)

grafico4.3.c.1
```



Hide

```
grafico4.3.c.2
```



4- Problema: analisar as preferências de usuários, por gênero, entre dois sistemas Amostra é prefsABCsex.cvs

Desenvolvimento:

a)Hipóteses:

“Há diferença na preferência do sistemas A por gênero?”

“Há diferença na preferência do sistemas B por gênero?”

“Há diferença na preferência do sistemas C por gênero?”

b)Leitura dos dados:

Hide

```
#A leitura dos dados é realizado excluindo-se a primeira coluna, pois se trata de um sequêncial
 sem significado para o problema em questão. Além disso é realizado o "factor" dos dados para qu
e seja possível realizar a contagem dos valores
prefsABCsex <- as.tibble(read_csv("Dados/prefsABCsex.csv",
    col_types = cols(Pref = col_factor(levels = c("A",
        "B", "C")), Sex = col_factor(levels = c("F",
        "M")), Subject = col_skip())))

#organização dos dados: agrupamento por preferência de sistema

tabelaPorPref.ABCSex <- prefsABCsex%>%
  group_by(Pref,Sex)%>%
  summarize(count = n())%>%
  spread(Sex, count)

print(tabelaPorPref.ABCSex)
```

Pref <fctr>	**F** <int>	**M** <int>
A	3	5
B	15	6
C	11	20

3 rows

Hide

```
#organização dos dados: agrupamento por gênero do participante
tabelaPorSex.ABCSex <- prefsABCsex%>%
  group_by(Sex, Pref)%>%
  summarize(count = n())%>%
  spread(Pref, count)

print(tabelaPorSex.ABCSex)
```

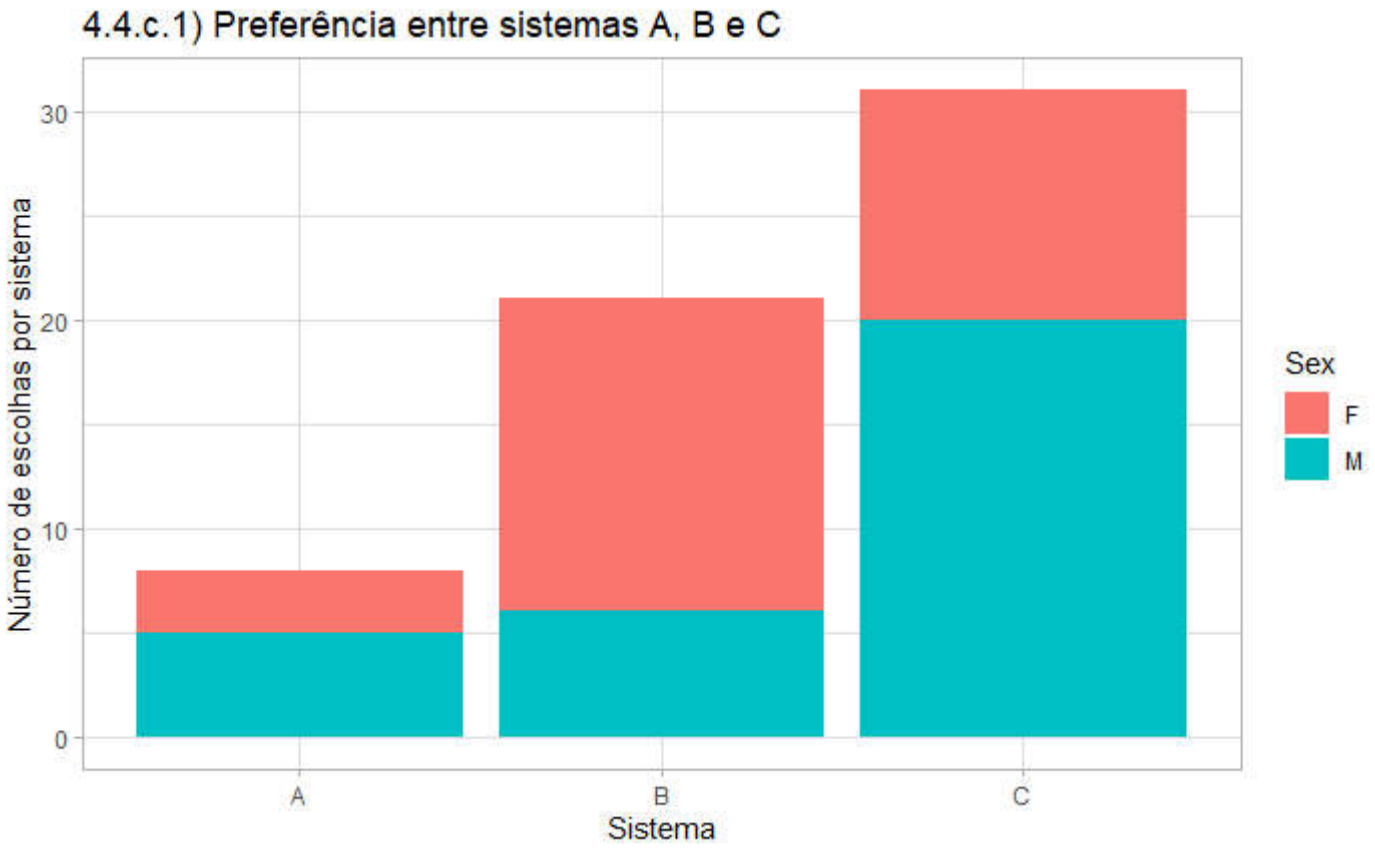
Sex <fctr>	**A** <int>	**B** <int>	**C** <int>
F	3	15	11
M	5	6	20

2 rows

c)Geração de gráficos

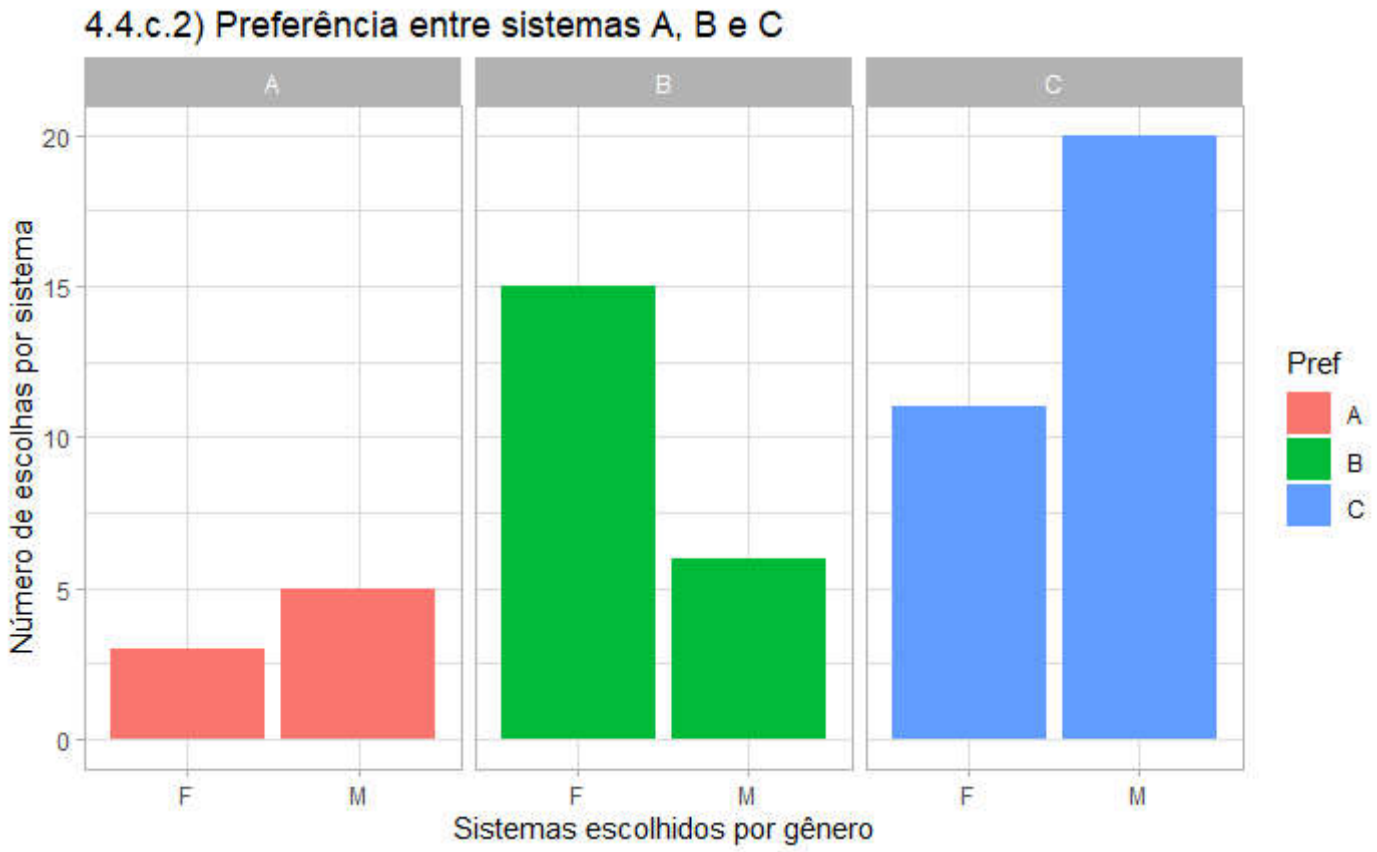
Hide

```
grafico4.4.c.1 <- ggplot(prefsABCsex, aes(x=Pref, fill = Sex))+  geom_bar() + theme_light()+
  labs(x= "Sistema", y="Número de escolhas por sistema", title="4.4.c.1) Preferência entre siste
mas A, B e C")


grafico4.4.c.2 <- ggplot(prefsABCsex, aes(x=Sex, fill = Pref))+  geom_bar() + theme_light()+
  labs(x= "Sistemas escolhidos por gênero", y="Número de escolhas por sistema", title="4.4.c.2)
 Preferência entre sistemas A, B e C")+facet_wrap(~Pref)
grafico4.4.c.1
```



Hide

```
grafico4.4.c.2
```



5ª. Questão. Considere o arquivo, tempoporsite.cvs, em que 600 participantes, usaram um dos dois sites A ou B, e o tempo de uso de cada um deles foi computado. Este arquivo descreve um teste A / B do site em que o tempo no site dos participantes foi medido em duas variações do mesmo site. Metade foram expostos ao site A e metade a variação de A, que é o site B. Nenhum deles tinha usado o site antes. A questão de investigação foi a seguinte: Qual site causa no usuário o interesse em ficar o maior tempo possível.

Pede-se:

a. **Analise os conceitos apresentados em aula (desvio padrão, mediana). Quantos sujeitos estavam neste teste A / B do site?**

Quantos sujeitos foram expostos a cada variação do site A e B?

Qual foi o tempo médio para o site "B"?

Qual foi o desvio padrão de tempo para o site "A"?

b. **Gere um ou mais gráficos mencionados na aula e ilustre seu gráfico. Analise a utilidade dos gráficos que você gerou.**

Faça os histogramas para cada um dos sites

Faça um boxplot comparando os sites.

Desenvolvimento:

Hide

```
#leitura dos dados
tempoporsite <- read_csv("Dados/tempoporsite.csv",
    col_types = cols(Site = col_factor(levels = c("A",
        "B")), Subject = col_skip(), Time = col_integer()))
resumo = summary(tempoporsite)
#filtro por observações de cada site
obsSiteA <- filter(tempoporsite, Site == "A")
obsSiteB <- filter(tempoporsite, Site == "B")

#ciração de uma tabela resumo
tabelaResumo <- tibble(Site = c("A", "B"),
                       qtdSujeitos = c(as.integer(count(obsSiteA)),as.integer(count(obsSiteB))),
                       mediaTempo = c(mean(obsSiteB$Time),mean(obsSiteB$Time)),
                       desvPadTempo = c(sd(obsSiteB$Time),sd(obsSiteB$Time)))
#qtdSujeitos <- tabelaResumo%>%select(qtdSujeitos)%>%colSums(na.rm = TRUE, dims = 1L)
print("a)")
```

```
[1] "a)"
```

Hide

```
#Cálculo da quantidade de sujeitos
qtdSujeitos <- tabelaResumo%>%select(qtdSujeitos)%>%colSums(na.rm = TRUE, dims = 1L)
sprintf("Quantidade de sujeitos no teste: %d", qtdSujeitos)
```

```
[1] "Quantidade de sujeitos no teste: 600"
```

Hide

```
#Cálculo da quantidade de sujeitos que utilizaram o Site A
qtdSujeitosA <- tabelaResumo%>%filter(Site== "A")%>%select(qtdSujeitos)%>%colSums(na.rm = TRUE,
 dims = 1L)
sprintf("Quantidade de sujeitos no teste expostos ao site A: %d", qtdSujeitosA)
```

```
[1] "Quantidade de sujeitos no teste expostos ao site A: 300"
```

Hide

```
#Cálculo da quantidade de sujeitos que utilizaram o Site B
qtdSujeitosB <- tabelaResumo%>%filter(Site== "B")%>%select(qtdSujeitos)%>%colSums(na.rm = TRUE,
 dims = 1L)
sprintf("Quantidade de sujeitos no teste expostos ao site B: %d", qtdSujeitosB)
```

```
[1] "Quantidade de sujeitos no teste expostos ao site B: 300"
```

Hide

```
#Cálculo da média de tempo gasto no Site B
mediaSiteB <- as.double(tabelaResumo%>%filter(Site == "B")%>%select(mediaTempo))
sprintf("Média de tempo gasto no site B: %.2f", mediaSiteB)
```

```
[1] "Média de tempo gasto no site B: 347.56"
```

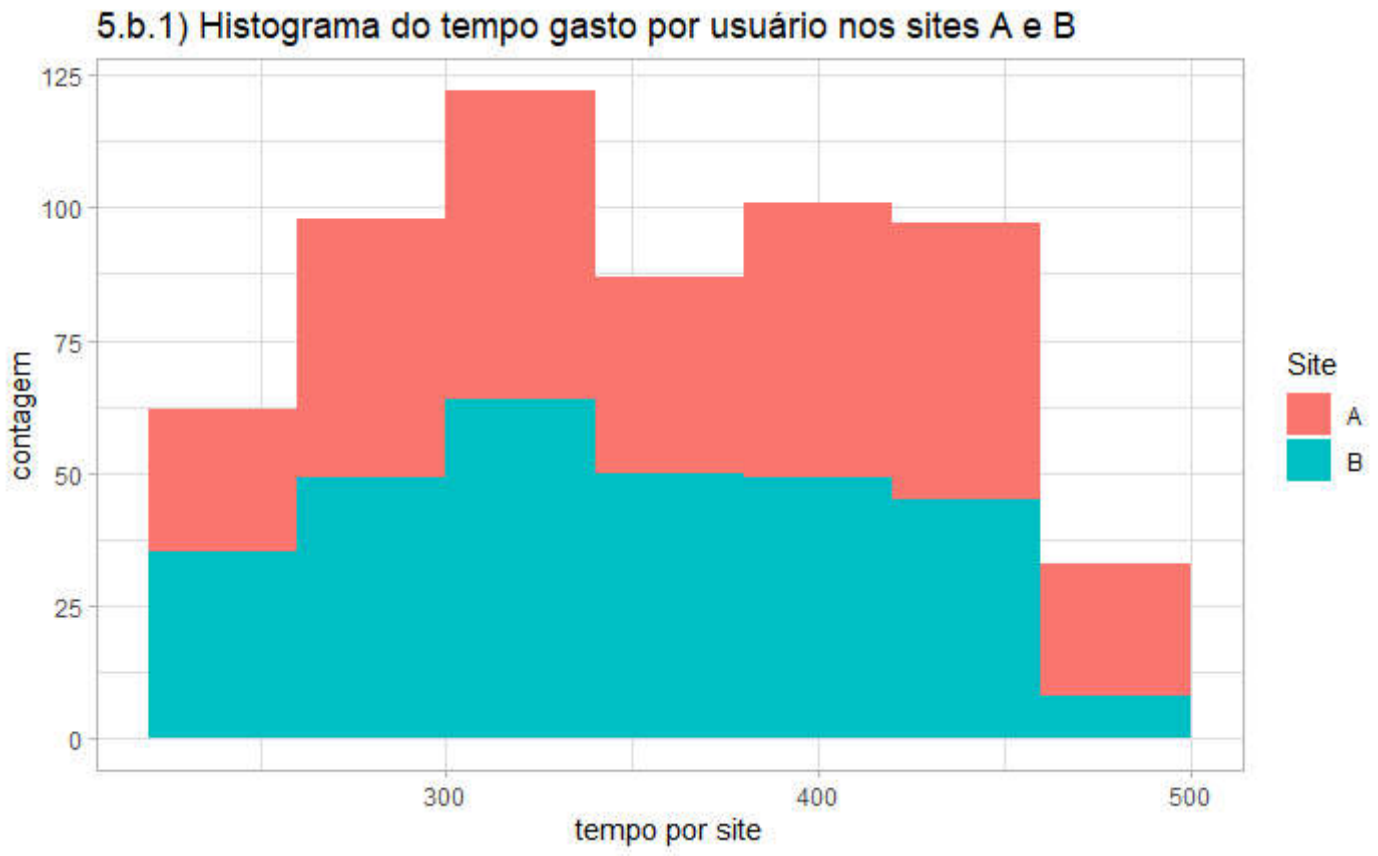
Hide

```
#Cálculo do desvio padrão do tempo gasto no Site A
desvPadSiteA <- as.double(tabelaResumo%>%filter(Site == "A")%>%select(desvPadTempo))
sprintf("Desvio padrão do tempo gasto no site A: %.3f", desvPadSiteA)
```

```
[1] "Desvio padrão do tempo gasto no site A: 65.878"
```

Hide

```
#geração de gráficos


grafico5.b.1 <- ggplot(tempoporsite, aes(Time, fill=Site))+  geom_histogram(binwidth=40) + theme
_light()+
  labs(x= "tempo por site", y = 'contagem',title="5.b.1) Histograma do tempo gasto por usuário n
os sites A e B")

grafico5.b.1
```



Hide

```
grafico5.b.2 <- ggplot(tempoporsite, aes(Site,Time, fill=Site))+  geom_boxplot() + theme_light()
+
  labs(x= "Site", y="Tempo por site", title="5.b.2) Tempo gasto por usuário nos sites A e B")+co
ord_flip()

grafico5.b.2
```