Estatística Básica

Lista 2 GABARITO - Medidas Resumo

Luan Fiorentin

2019-03-05

1. Um exame de vestibular para uma faculdade tem 80 questões, sendo 40 de português e 40 de matemática. Para os 20 melhores classificados, apresentamos o número de acertos em cada disciplina, com os valores já ordenados.
   - Português: (8; 10; 11; 12; 12; 14; 17; 20; 20; 23; 23; 24; 26; 26; 30; 31; 32; 34; 35; 35)
   - Matemática: (13; 20; 20; 20; 21; 21; 23; 23; 25; 25; 26; 27; 28; 28; 28; 29; 30; 31; 31; 32)

   (a) Calcule as medidas de centro (média, mediana e moda) para cada grupo.
   (b) Calcule as medidas de variabilidade (variância, desvio-padrão, e coeficiente de variação) para cada grupo.
   (c) Calcule os quartis para cada grupo.
   (d) Construa um gráfico de caixa (box plot) para cada grupo (em um mesmo gráfico para comparação).
   (e) Com todos os resultados obtidos, descreva comparativamente os dois grupos em termos de medidas de tendência central, variabilidade, amplitude e distribuição (simetria) dos dados.
   (f) Você acha que os aprovados são melhores em português ou matemática?

```r
# (a)
# Média
mean(da1$Por)
```

```
## [1] 22.15
```

```r
mean(da1$Mat)
```

```
## [1] 25.05
```

```r
# Mediana
median(da1$Por)
```

```
## [1] 23
```

```r
median(da1$Mat)
```

```
## [1] 25.5
```

```r
# Moda
# Para variável Português há 5 valores modais: 12; 20; 23; 26; 35.
# Para variável Matemática há 2 valores modais: 20; 28.

# (b)
# Variância
var(da1$Por)
```

```
## [1] 80.13421
```

```r
var(da1$Mat)
```

```
## [1] 23.83947
```

```r
# Desvio Padrão
sd(da1$Por)
```

```
## [1] 8.951771
```

```r
sd(da1$Mat)
```

```
## [1] 4.882568
```

```r
# Coeficiente de Variação
coevar(da1$Por)
```

```
## [1] 40.41432
```

```r
coevar(da1$Mat)
```

```
## [1] 19.49129
```

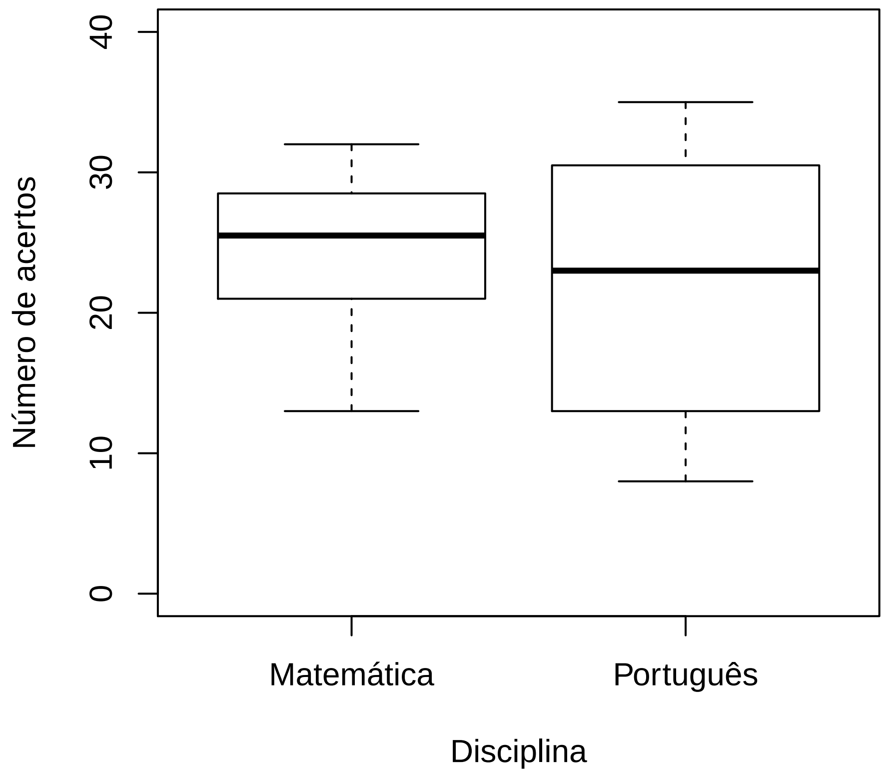
```r
# (c) Quantis
quantile(da1$Por)
```

```
##    0%   25%   50%   75%  100%
##  8.00 13.50 23.00 30.25 35.00
```

```r
quantile(da1$Mat)
```

```
##    0%   25%   50%   75%  100%
## 13.00 21.00 25.50 28.25 32.00
```

```r
# (d)
boxplot(dab$val ~ dab$var,
        ylim = c(0,40),
        xlab = "Disciplina",
        ylab = "Número de acertos")
```




```
# (e)
# Em média, o número de acertos em
# matemática (25,05) foi maior do que o número de
# acertos em português (22,15). A diferença entre
# os valores médios e a mediana mostra que existe uma leve assimetria
# negativa (ou à esquerda) para os dois casos (média < ,mediana),
# embora esta diferença seja mais evidente nas notas de
# português. A amplitude do número de acertos em português foi de
# 35-8=27, maior do que a amplitude obervada
# para o número de acertos em matemática, que foi de
# 32-13=19. A variabilidade dos acertos em torno
# da média também foi maior para as notas de português, com variância
# de 80,134 e desvio-padrão de 8,951. Já
# para a matemátiva, a variabilidade em torno da média foi menor, com
# 23,839 e desvio-padrão 4,883. Resumindo
# estas informações sobre variabilidade, nota-se que o coeficiente de
# variação para português foi de 40,4%, enquanto para a
# matemática foi menor, com aproximadamente 19,5%. Através do resumo
# dos cinco números e do gráfico de caixa, percebe-se que 50% dos
# acertos foram entre 13 e 30,5 em português (diferença entre Q1 e
# Q3), e entre 21 e 28,5 em matemática, mostrando novamente a menor
# variabilidade observada para a matemática.
```

2. Considere a amostras de alturas de uma espécie de árvore (metros), coletadas em quatro áreas diferentes.
   - Área A: (9,2; 10,8; 10,6; 11,1; 12,1; 9,6; 11,2; 8,4; 12,9; 12,1; 14,4; 11,1; 11,1; 9,7; 8,4; 12,3; 10,7; 12,9; 9,1; 12,8).
   - Área B: (12,5; 18,5; 21,3; 14,3; 18,5; 19,0; 10,8; 23,1; 17,4; 10,7; 14,3; 16,3; 18,0; 7,1; 12,8; 14,7; 11,3; 8,2; 13,8).
   - Área C: (21,3; 28,7; 15,8; 24,0; 13,7; 18,1; 12,6; 14,6; 6,1; 19,8; 22,3; 15,7; 16,3; 18,2; 15,7; 6,6; 9,3; 1,3; 19,0).
   - Área D: (13,7; 8,6; 14,9; 10,2; 14,0; 10,5; 15,0; 5,2; 10,0; 11,7; 18,7; 9,3; 7,9; 6,5; 11,5; 12,0; 8,3; 8,3; 9,8; 4,7).

   (a) Calcule as medidas de amplitude, média, mediana, variância, desvio padrão e coeficiente de variação para as quatro áreas.

   (b) Descreva comparativamente as quatro áreas quanto à altura das árvores, utilizando as estatísticas que você calculou.

```r
# (a) Estatísticas calculadas
round(apply(da2, 2, estat_descritiva), 3)
```

```
##                AreaA  AreaB  AreaC  AreaD
## Mínimo         8.400  7.100  1.300  4.700
## Média         11.025 14.665 15.805 10.540
## Mediana       11.100 14.300 16.050 10.100
## Máximo        14.400 23.100 28.700 18.700
## Amplitude      6.000 16.000 27.400 14.000
## Variância      2.658 18.451 41.709 12.318
## Desvio Padrão  1.630  4.295  6.458  3.510
## CV            14.787 29.290 40.862 33.299
```

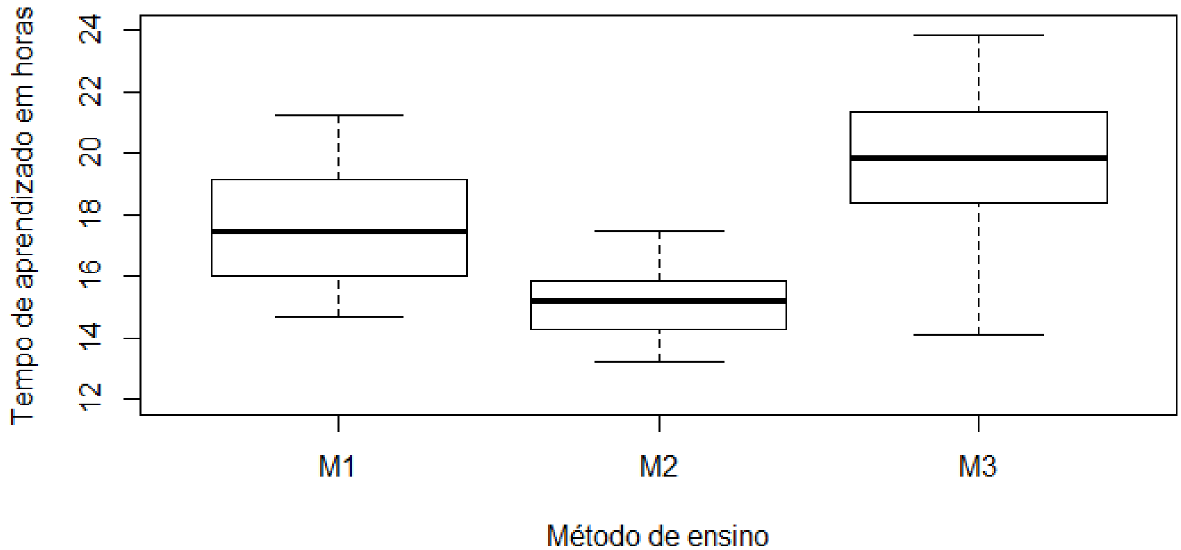
```r
# (b)
# Em média, a área C possui as árvores
# mais altas (18,8), enquanto a área D possui as
# árvores mais baixas (10,5). Em todas as áreas, o valor
# da mediana está muito próximo do valor da média, o que
# indica que a distribuição das alturas em todas as áreas é
# aproximadamente simétrica. A maior amplitude de variação de alturas
# foi observada na área C, que também apresentou a maior variabilidade
# das observações em relação à média, como pode ser observado pelos
# valores da variância (43,943), e do desvio-padrão (6,629).
# A área com árvores de alturas mais homogêneas foi a A, pois a
# amplitude foi de 6 m, e a varibilidade das alturas em torno da média
# foi a menor quando comparada com as demais áreas (A = 2,66 e
# 1,63). Estas diferenças de variabilidade podem ser observadas
# através do coeficiente de variação, que foi de 40,9% para a área C, e
# de 14,8% para a área A. A área D apresentou um CV de 33,3%, enquanto
# o CV da área B foi de 29,3%.
```

3. Deseja-se comparar três métodos de ensino no aprendizado de estatística. Cada método foi aplicado a 30 alunos e os resultados estão apresentados nos boxplot abaixo.

   (a) Encontre os valores (aproximados) para a mediana, os quartis, máximo e mínimo.

   (b) Discuta a variabilidade do tempo de aprendizado de cada método.

   (c) Se você é otimista, qual método escolheria?




```r
# (a)
# Observação: São valores verdadeiros.
round(apply(daresul, 2, quantile), 3)
```

```
##          M1     M2     M3
## 0%   14.675 13.219 14.114
## 25%  16.087 14.302 18.507
## 50%  17.499 15.227 19.872
## 75%  19.136 15.841 21.219
## 100% 21.225 17.473 23.869
```

```r
#(b)
# A amplitude do tempo de aprendizado para
# os métodos M1, M2 e M3 são: M1 = 6,550, M2 = 4,254, e M3 = 9,755.
# Com isso, percebe-se que:
# o método M1 apresenta um tempo aprendizado (aproximado) entre 14 e 21
# horas, o método M2 entre 13 e 17 horas, e o método M3 entre 14 e 23
# horas. A maior variação de tempo de recuperação foi então do método M3
# (maior amplitude), enquanto o método M2 apresentou a menor
# variação. Nota-se, em termos de mediana, que o método M2 apresentou o
# menor tempo de aprendizado (Me = 15,227), e com uma pequena
# variabilidade. Além disso, percebe-se através do gráfico boxplot, que
# os métodos M1 e M3 possuem uma distribuição simétrica, como pode
# ser observado pela proximidade da mediana com o primeiro (Q1) e terceiro quartil
# (Q3). Percebe-se também que a distribuição do tempo de aprendizado do método M2
# é levemente assimétrica à direita, pois a mediana está deslocada para cima
# dentro da caixa. O tempo de de aprendizado mediano para o método M2 foi
# de 15,227. Para esse método, 50\% das pessoas tiveram um tempo de
# recuperação entre 14,302 (Q1) e 15,841 (Q3) horas.
```

4. A distribuição das estaturas, em centímetros, de alunos de um curso colegial está representada na tabela

de frequência abaixo. Calcule a média, a variância, e o desvio-padrão das estaturas.

Classes	Frequência
[135-145)	15
[145-155)	150
[155-165)	250
[165-175)	70
[175-185)	10
[185-195]	5

```
# Tabela auxiliar
data.frame(xi, fi, xifi, xi2, xi2fi)
```

```
##    xi  fi  xifi   xi2   xi2fi
## 1 140  15  2100 19600  294000
## 2 150 150 22500 22500 3375000
## 3 160 250 40000 25600 6400000
## 4 170  70 11900 28900 2023000
## 5 180  10  1800 32400  324000
## 6 190   5   950 36100  180500
```

```
# Média
sum(xifi / sum(fi))
```

```
## [1] 158.5
```

```
# Variância
(1/(sum(fi)-1)) * (sum(xi2fi) - sum(xifi)^2/sum(fi))
```

```
## [1] 70.89178
```

```
# Desvio Padrão
sqrt((1/(sum(fi)-1)) * (sum(xi2fi) - sum(xifi)^2/sum(fi)))
```

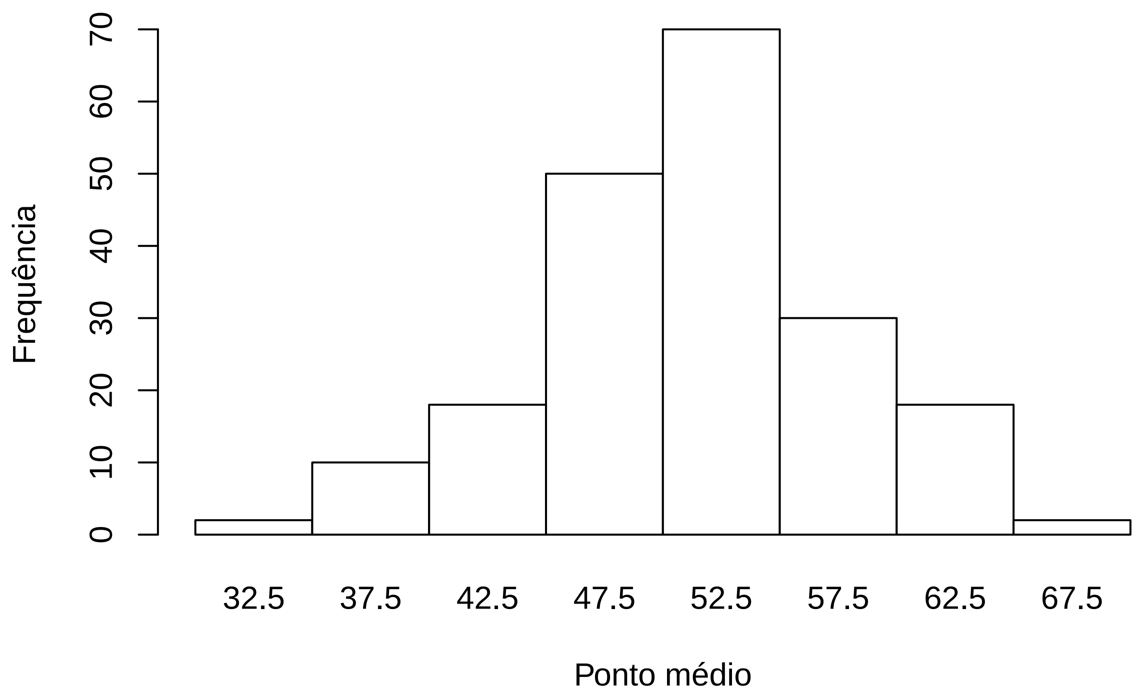
```
## [1] 8.419726
```

5. Os dados abaixo representam as vendas semanais, em classes de salários mínimos, de vendedores de gêneros alimentícios:
   (a) Faça o histograma das observações.
   (b) Calcule a média da amostra ($\bar{x}$).
   (c) Calcule o desvio-padrão da amostra.
   (d) Calcule o primeiro quartil (Q_1), mediana (Q_2) e o terceiro quartil (Q_3).

Vendas	Vendedores
[30-35)	2
[35-40)	10
[40-45)	18
[45-50)	50
[50-55)	70
[55-60)	30
[60-65)	18
[65-70)	2

```
# (a)
barplot(fi,
        space = 0,
```

```r
        names.arg = xi,
        xlab = "Ponto médio",
        ylab = "Frequência",
        col = "white")
```




```r
# Tabela auxiliar
data.frame(xi, fi, xifi, xi2, xi2fi)
```

```
##     xi fi xifi     xi2    xi2fi
## 1 32.5  2   65 1056.25   2112.5
## 2 37.5 10  375 1406.25  14062.5
## 3 42.5 18  765 1806.25  32512.5
## 4 47.5 50 2375 2256.25 112812.5
## 5 52.5 70 3675 2756.25 192937.5
## 6 57.5 30 1725 3306.25  99187.5
## 7 62.5 18 1125 3906.25  70312.5
## 8 67.5  2  135 4556.25   9112.5
```

```r
# (b)
# Média
sum(xifi / sum(fi))
```

```
## [1] 51.2
```

```r
# (c)
# Variância
(1/(sum(fi)-1)) * (sum(xi2fi) - sum(xifi)^2/sum(fi))
```

```
## [1] 44.03015
```

```r
# Desvio Padrão
sqrt((1/(sum(fi)-1)) * (sum(xi2fi) - sum(xifi)^2/sum(fi)))
```

```
## [1] 6.635522
```

```r
# (d)
cumsum(prop.table(fi))
```

```
## [1] 0.01 0.06 0.15 0.40 0.75 0.90 0.99 1.00
```

```r
# Considerando que
# (55-50)/0,35 = (Q2-50)/0,10
# Então a mediana é
(Q2 <- ((55-50)/0.35) * 0.10 + 50)
```

```
## [1] 51.42857
```

```r
(Q1 <- ((50-45)/0.25) * 0.10 + 45)
```

```
## [1] 47
```

```r
(Q3 <- ((55-50)/0.35) * 0.35 + 50)
```

```
## [1] 55
```

6. O departamento pessoal de uma certa firma fez um levantamento dos salários dos 120 funcionários do setor administrativo, obtendo os resultados (em salários mínimos) da tabela abaixo.
   (a) Calcule a média, a mediana, a variância e o desvio-padrão
   (b) Se for concedido um aumento de 100% para todos os 120 funcionários, haverá alteração na média? E na variância? Justifique sua resposta.
   (c) Se for concedido um abono de dois salários mínimos para todos os 120 funcionários, haverá alteração na média? E na mediana? E na variância? Justifique sua resposta.

Salários	Freq. relativa
[0-2)	0,25
[2-4)	0,40
[4-6)	0,20
[6-10]	0,15

```r
# (a)
# Tabela auxiliar
data.frame(xi, fi, xifi, xi2, xi2fi)
```

```
##   xi fi xifi xi2 xi2fi
## 1  1 30   30   1    30
## 2  3 48  144   9   432
## 3  5 24  120  25   600
## 4  8 18  144  64  1152
```

```r
# Média
sum(xifi / sum(fi))
```

```
## [1] 3.65
```

```r
# Variância
(1/(120)) * (sum(xi2fi) - (sum(xifi)^2)/120)
```

```
## [1] 5.1275
```

```r
# Desvio Padrão
sqrt((1/(120)) * (sum(xi2fi) - (sum(xifi)^2)/120))
```

```
## [1] 2.264398
```

```r
# Mediana
(Q2 <- ((4-2)/0.40) * 0.25 + 2)
```

```
## [1] 3.25
```

```r
# (b)
# Haverá alteração na média, pois é multiplicada por 2
sum(xifi / sum(fi)) * 2
```

```
## [1] 7.3
```

```r
# Haverá alteração na variância, pois é multiplicada por 4
(1/(120)) * (sum(xi2fi) - (sum(xifi)^2)/120) * 4
```

```
## [1] 20.51
```

```r
#(c)
# Haverá alteração na média, pois é somada em 2 unidades
sum(xifi / sum(fi)) + 2
```

```
## [1] 5.65
```

```r
# Haverá alteração na mediana, pois é somada em 2 unidades
(Q2 <- (((4-2)/0.40) * 0.25 + 2) + 2)
```

```
## [1] 5.25
```

```r
# Não haverá mudança na variância, pois apenas acrescentou-se um
# um valor constante
(1/(120)) * (sum(xi2fi) - (sum(xifi)^2)/120)
```

```
## [1] 5.1275
```

7. O resultado de uma prova de estatística aplicada à 25 alunos foi o seguinte:

$$(9, 9, 8, 8, 9, 10, 8, 8, 9, 8, 10, 7, 7, 9, 9, 7, 8, 9, 4, 7, 7, 8, 10, 9, 9)$$

Como os alunos possuiam diferentes níveis educacionais, decidiu-se calcular o desempenho relativo de cada candidato, para facilitar a interpretação dos resultados. Essa medida de desempenho relativo será obtida da seguinte forma: 1. Calcula-se a média ($\bar{x}$) e o desvio-padrão (s) da amostra. 2. A nota x_i de cada aluno será padronizada da seguinte forma:

$$z_i = \frac{x_i - \bar{x}}{s}.$$

Assim, criou-se uma nova variável Z que corresponde ao conjunto de notas padronizadas. Com isso:

(a) Calcule as notas padronizadas de todos os funcionários.
(b) Com os resultados obtidos acima, calcule a média ($\bar{z}$) e o desvio-padrão (s_z) de Z.
(c) Se alguma das notas padronizadas estiver acima de $2s_z$ ou abaixo de $-2s_z$, esse aluno deve ser considerado "atípico". Existe algum nessa situação?
(d) Interprete o significado de Z.

```r
# (a)
# Média
mean(da7)
```

```
## [1] 8.24
```

```r
# Variância
sd(da7)
```

```
## [1] 1.3
```

```r
# Padronizando
(xi_pad <- (da7 - mean(da7)) / sd(da7))
```

```
##  [1]  0.5846154  0.5846154 -0.1846154 -0.1846154  0.5846154  1.3538462
##  [7] -0.1846154 -0.1846154  0.5846154 -0.1846154  1.3538462 -0.9538462
## [13] -0.9538462  0.5846154  0.5846154 -0.9538462 -0.1846154  0.5846154
## [19] -3.2615385 -0.9538462 -0.9538462 -0.1846154  1.3538462  0.5846154
## [25]  0.5846154
```

```r
# (b)
# Média
mean(xi_pad)
```

```
## [1] -1.587923e-16
```

```r
# Desvio Padrão
sd(xi_pad)
```

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

```r
# (c)
# Sim, o valor atípico é -3,33,
# e corresponde a nota 4.

# (d)
# A variável Z é uma variável padronizada, que mede o número de desvios padrões
# que cada observação se afasta de média.
```