

MATH 203: PRINCIPLES OF STATISTICS 1

FALL 2024

ASSIGNMENT 2 – SOLUTIONS

1. Grades from a previous MATH 203 midterm exam are saved in the file

<https://www.math.mcgill.ca/dstephens/203/Data/Midterm203.txt>

The maximum grade available for this exam was 128 points.

(a) Compute the sample mean, $\bar{x}$, and the sample standard deviation, s , of the grades. Report your answers to two decimal places. 4 MARKS

(b) Based on the Empirical Rule, it is claimed that the interval

$$\bar{x} \pm 2s = (92.54, 146.16)$$

contains approximately 95% of the midterm grades.

(i) Identify one deficiency in this interval as representation of the spread of the sample data. 2 MARKS

(ii) Explain (in one sentence) why the interval may not contain approximately 95 % of the data. 2 MARKS

(c) Compute the sample z -score for a grade of 120 on the exam. 2 MARKS

Solutions: Using the R code

```
dtype<-'https://www.math.mcgill.ca/dstephens/203/Data/'
fname<-'Midterm203.txt'
MT<-scan(paste(dtype,fname,sep='/'))
mean.val<-mean(MT); sd.val<-sd(MT)
c(mean.val, sd.val)

## [1] 119.35071 13.40455

hist(MT,plot=F)$breaks; hist(MT,plot=F)$counts

## [1] 60 65 70 75 80 85 90 95 100 105 110 115 120 125 130
## [1] 2 2 5 4 1 5 12 18 16 15 5 46 105 186

zval<-(120-mean.val)/sd.val #Full precision
zrnd<-(120-round(mean.val,2))/round(sd.val,2) #With 2dp rounding
c(zval, zrnd)

## [1] 0.04843797 0.04850746
```

(a) $\bar{x} = 119.35, s = 13.40$. 4

(b) (i) This interval is deficient as the upper limit of the grades is 128, so having an upper limit of 146.16 is inappropriate. 2

(ii) The data are left-skewed, as evidenced by the histogram counts, and therefore the histogram is not symmetric. 2

(c) $z = (120 - 119.35)/13.40 = 0.05$. 2

2. A version of the Corporate sustainability data (see Assignment 1) are saved in the `Sustain.txt` file, and represent measured scores for the level of support for sustainability in the corporate sector.

`https://www.math.mcgill.ca/dstephens/203/Data/Sustain.txt`

- (a) Find the sample mean, $\bar{x}$, and sample standard deviation, s , of the sustainability scores. Report your answers to two decimal places. 4 MARKS
- (b) Construct an interval that contains approximately 95% of all the sustainability scores. Report the lower and upper ends of the interval to two decimal places. 4 MARKS
- (c) Compute the sample inter-quartile range, IQR, for the sustainability scores. 2 MARKS

Solutions: Using the R code

```
dname<-'https://www.math.mcgill.ca/dstephens/203/Data/'
fname<-'Sustain.txt'
Sus<-scan(paste(dname,fname,sep='/'))
mean.val<-mean(Sus); sd.val<-sd(Sus)
c(mean.val, sd.val)

## [1] 69.03427 26.43567

n<-length(Sus)
pL<-0.25
pU<-0.75
y<-sort(Sus) #Data sorted into ascending order
idL<-pL*n
idU<-pU*n
c(y[idL],y[idU]) #Lower and upper quartiles

## [1] 52 88

IQR<-y[idU]-y[idL] #Compute the IQR
quantile(Sus,probs=c(0.25,0.75)) #Using the quantile function

## 25% 75%
## 52 88
```

- (a) $\bar{x} = 69.03$, $s = 26.44$. 4

- (b) The interval is 4

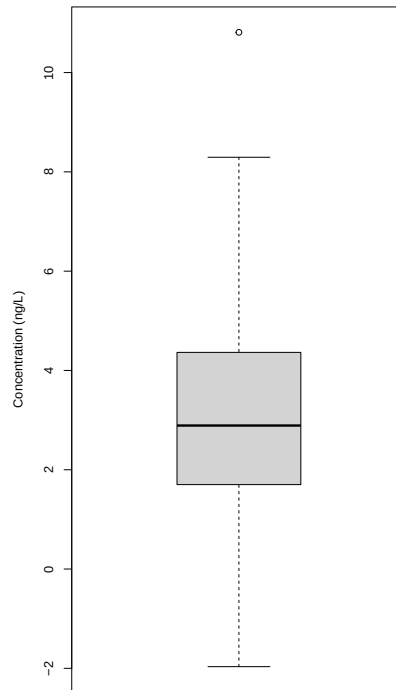
$$(\bar{x} - 2s, \bar{x} + 2s) = (69.03 - (2 \times 26.44), 69.03 + (2 \times 26.44)) = (16.16, 121.91)$$

- (c) The sample size is $n = 992$. Therefore, we need to identify the values in the sorted list that are 25% and 75% of the way through the list. 2

- Lower quartile q_L : $p = 0.25$, $n \times p = 248$, so $q_L = 52$.
- Upper quartile q_U : $p = 0.75$, $n \times p = 744$, so $q_U = 88$.
- $IQR = q_U - q_L = 36$.

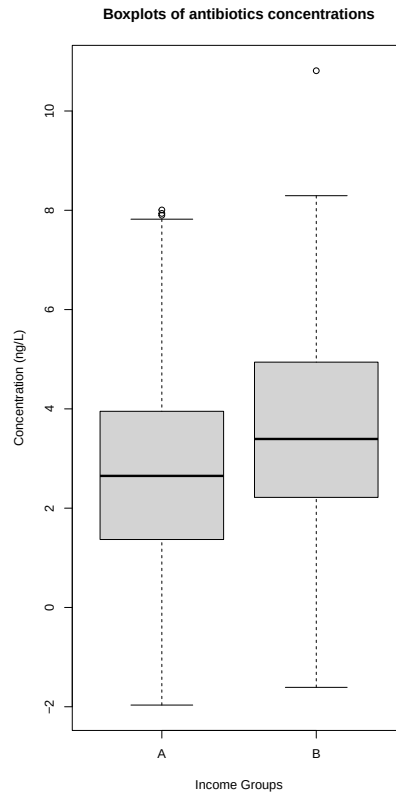
3. **Global occurrence of anti-infectives in contaminated surface waters: Impact of income inequality between countries** (Segura et al. *Environment International* Volume 80, July 2015, Pages 89-97).

(a) The following is a boxplot of concentrations of antibiotics from pooled data from different regions in the world (values are log-transformed).



- (i) Report (approximately) a measure of central tendency for the data; explain how you obtained this value. 2 MARKS
- (ii) Provide (approximately) a measure of spread or variability for the data; explain how you obtained this value. 2 MARKS
- (iii) What can you say about the 'shape' of the data? 2 MARKS
- (iv) What can you say about the presence of outliers? 2 MARKS
- (v) Is it possible to compute (approximately) how many of the log-transformed numerical values are negative? Justify your answer. 2 MARKS

- (b) The following are boxplots of concentrations of antibiotics from two different regions, A and B (values are log-transformed).



Compare in a few sentences these two distributions of log-transformed antibiotics concentrations, in terms of centre, spread or variability, shape, and the presence of outliers.

6 MARKS

Solutions:

- (a) (i) The median is approximately 2.8 to 2.9 (the middle line in the box). 2
- (ii) The IQR is approximately 2.6 (1.8 to 4.4) (the difference between upper and lower box limits). 2
- (iii) The distribution is quite symmetric; the median is central in the box, the whiskers are more or less the same length. 2
- (iv) There is a single outlier with value around 11. 2
- (v) This is not really possible, as the original data are not available, and we do not know the sample size (however, we can say that the proportion is below 25%, as the lower quartile is well above zero). 2
- (b) • Centre: Region B has a higher median than Region A;
- Spread: The spreads of the two regions are similar;
- Shape: Both distributions are reasonably symmetric;
- Outliers: Region A has a couple of outliers, Region B has one outlier. 6

4. A marketing research organization conducts a yearly survey collecting demographic information on respondents from each country that they survey.

The table below shows the cross-classification of a sample of $n = 7645$ people by country and maximum level of education they have received, in each of five countries.

	Post-graduate	College	High School	Primary
China	7	315	671	506
France	69	388	766	309
India	161	514	622	227
U.K.	58	207	1240	32
U.S.A.	84	486	896	87

A person is selected from this survey, with each of the surveyed individuals equally likely to be selected. Let events

- A_1, A_2, A_3, A_4, A_5 represent the events that the person originates from China, France, India, U.K. and U.S.A. respectively.
- B_1, B_2, B_3, B_4 represent the events that the person completed their education at the post-graduate, college, high school and primary education levels.

Express the following probabilities or conditional probabilities using this notation and the available data; that is, write

$$P(\cdot) = \text{answer} \quad \text{or} \quad P(\cdot|\cdot) = \text{answer}$$

where $\cdot$ is a specific event.

- (a) What is the probability that the selected person

- (i) is from France? 2 MARKS
- (ii) completed their education at college or after post-graduate study? 2 MARKS
- (iii) is from China or did post-graduate study? 2 MARKS
- (iv) is from the U.S.A and finished only primary school? 2 MARKS

You may leave your numerical answers as fractions.

- (b) Are the events

‘Person is from the U.K.’ and ‘Person’s maximum level of education is College’

independent? Justify your answer: use the above notation and give a numerical calculation.

2 MARKS

- (c) Given that the selected person is observed to be from India, compute the probability that they completed their education at high school. Use the above notation and give a numerical calculation. 2 MARKS

- (d) Given that the selected person is observed to have completed their education at the college level, compute the probability that they originated from the U.S.A. Use the above notation and give a numerical calculation. 2 MARKS

Solutions:

```
Market<-matrix(c(7,315,671,506,69,388,766,309,161,514,
                 622,227,58,207,1240,32,84,486,896,87),5,4,byrow=T)
rownames(Market)<-c('China','France','India','UK','USA')
colnames(Market)<-c('PG','C','HS','P')
Market;n<-sum(Market)
```

```
##          PG      C      HS      P
## China      7    315    671    506
## France    69    388    766    309
## India   161    514    622    227
## UK       58    207   1240     32
## USA      84    486    896     87
```

```
apply(Market,1,sum);
```

```
##  China France  India      UK      USA
##   1499   1532   1524   1537   1553
```

```
apply(Market,2,sum)
```

```
##      PG      C      HS      P
##   379  1910  4195  1161
```

(a) (i) is from France: $P(A_2) = 1532/7645 = 0.200$. 2

(ii) completed their education at college *or* after post-graduate study: 2

$$P(B_1 \cup B_2) = P(B_1) + P(B_2) = \frac{379 + 1910}{7645} = \frac{2289}{7645} = 0.299.$$

(iii) is from China *or* did post-graduate study: 2

$$P(A_1 \cup B_1) = P(A_1) + P(B_1) - P(A_1 \cap B_1) = \frac{1499 + 379 - 7}{7645} = \frac{1871}{7645} = 0.245.$$

(iv) is from the U.S.A *and* finished only primary school: 2

$$P(A_5 \cap B_4) = 87/7645 = 0.011.$$

(b) These events are not independent: we have 2

$$P(A_4) = 1537/7645 = 0.201 \quad P(B_2) = 1910/7645 = 0.250 \quad P(A_4 \cap B_2) = 207/7645 = 0.027$$

so evidently $P(A_4 \cap B_2) \neq P(A_4)P(B_2)$.

(c) This is 2

$$P(B_3|A_3) = \frac{P(A_3 \cap B_3)}{P(A_3)} = \frac{622/7645}{1524/7645} = \frac{622}{1524} = 0.408.$$

(d) This is 2

$$P(A_5|B_2) = \frac{P(A_5 \cap B_2)}{P(B_2)} = \frac{486/7645}{1910/7645} = \frac{486}{1910} = 0.254.$$