1. [10 marks] The following data set $y = (y_1, y_2, \ldots, y_{19})$ consists of the heights in centimeters of nineteen STAT 231 female students in the winter term 2016:

156	158	158	158	158	159	159	160	160	161
161	161	162	162	164	164	165	166	167	

$$\sum_{i=1}^{19} y_i = 3059 \qquad \sum_{i=1}^{19} y_i^2 = 492667$$

(a) Circle the letter corresponding to your choice.

(i) Which of the following commands in R computes the sample mean for these data?

A: average(y)
[✓ **B**]: mean(y)
C: mn(y)
D: sum$(y)\,/n$
E: None of the Above

(ii) Which of the following commands in R generates 23 observations from a $G\,(160, 4)$ distribution?

A: rgauss$(23, 160, 4)$
B: rnorm$(160, 4, 23)$
C: rnorm$(23, 160, 16)$
[✓ **D**]: rnorm$(23, 160, 4)$
E: None of the Above

(b) **Write your numerical answer only in the space provided. Carry as many decimal places as possible while making your calculations. Your final answers should be given to 3 decimal places unless the answer is exact to less than 3 decimal places. Use the definition given in the Course Notes for calculating quantiles.**

(i) Determine the following:

-0.5 marks for answers not rounded to 3 decimal places and for incorrect answers due to rounding errors to a maximum of 2 marks

sample mode = 158

IQR (interquartile range) = 6 $\qquad IQR = y_{(15)} - y_{(5)} = 164 - 158 = 6$

$\hat{F}\,(159)$ = 0.368 $\qquad \hat{F}\,(159) = \dfrac{\text{no. observations } \le 159}{19} = \dfrac{7}{19} = 0.36842105$

(ii) The person measuring the students was using a measuring tape for which the first 5 centimeters were missing. To correct for this error, 5 centimeters was added to every height. Find the following for the corrected data set (that is, the data set with 5 added to every observation):

sample mean = 166 $\qquad \dfrac{3059}{19} + 5 = 161 + 5 = 166$

sample standard deviation = 3.055 $\qquad \left\{\dfrac{1}{18}\left[492667 - \dfrac{1}{19}(3059)^2\right]\right\}^{1/2} = 3.0550505$

(iii) The **original** heights were converted from centimeters to inches by dividing by 2.54. For the data set recorded in inches find:

sample mean = 63.386 $\qquad 161/2.54 = 63.385827$

sample variance = 1.447 $\qquad \left(\dfrac{1}{2.54}\right)^2 \dfrac{1}{18}\left[492667 - \dfrac{1}{19}(3059)\right] = 1.44666956$

(iv) It was discovered that the smallest observation 156 was recorded incorrectly and that the number should have been 157. Find the following for the corrected data set:

sample variance = 8.830 $\qquad \dfrac{1}{18}\left[492667 + (157)^2 - (156)^2 - \dfrac{1}{19}(3059 + 157 - 156)^2\right] = 8.8304094$

2. [10 marks] Fill in the blanks below. Use one of the following **at most once**:

prediction, estimation, hypothesis testing, deductive, inductive, statistical inference, descriptive statistics, applied statistics, sample proportion, sample mean, sample median, sample variance, sample standard deviation, interquartile range (IQR), relative risk, sample correlation, Poisson, Binomial, Exponential, Gaussian, θ, $\theta/\sqrt{n}$, θ^2/n, θ/n.

(a) Statistical inference is a form of inductive reasoning.

(b) Numerical summaries such as the sample median and the IQR are examples of

descriptive statistics.

(c) When we use the data obtained in a study of a population or process to draw general conclusions about the population or process we call this

statistical inference.

(d) A researcher wishes to design a study to decide whether or not a new migraine medication reduces the severity of migraine pain.

This is an example of a hypothesis testing problem.

(e) A researcher in the Ontario Ministry of Transportation has data on the number of deaths per year between 2000 and 2015 on Highway 401 between two intersections. Based on these data the researcher would like to say something about the number of deaths on the same stretch of highway during the year 2017.

This is an example of a prediction problem.

(f) Suppose the data set $y_1, y_2, \ldots, y_n$ is the realization of n independent copies of a random variable $Y \backsim G(\mu, \sigma)$.

The value of σ may be estimated using the sample standard deviation.

(g) Suppose $Y_1, Y_2, \ldots, Y_n$ are independent random variables all with the same standard deviation $\sqrt{\theta}$.

The variance of $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is equal to θ/n.

(h) The measure of location which is robust against outliers (extreme observations) is the

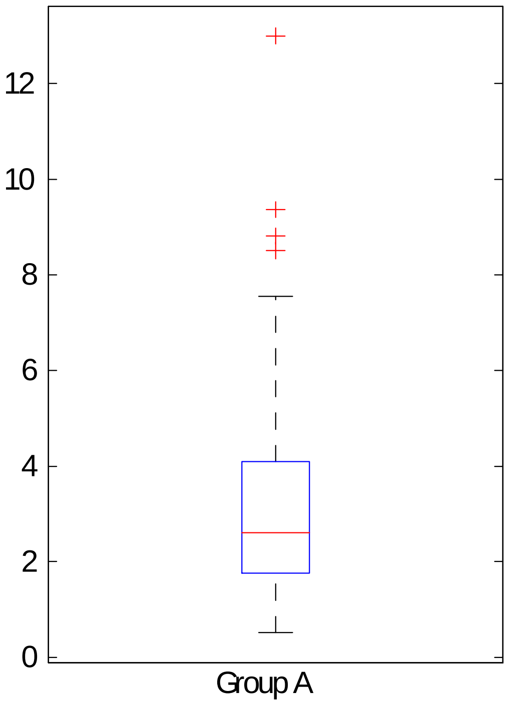
sample median.

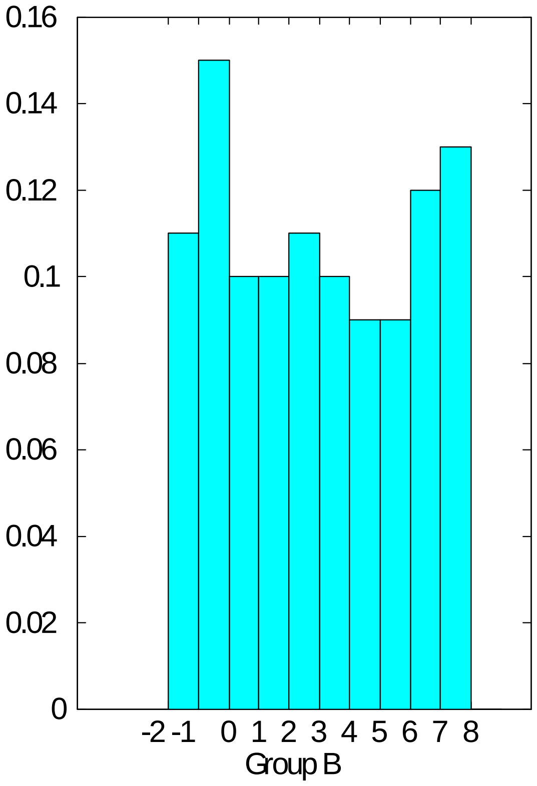
(i) For continuous bivariate data the sample correlation is a measure of the linear relationship between two variates.

(j) The Traffic and Road Department of the City of Waterloo has collected data on the number of bicycle accidents on city roads in Waterloo per month for each of the last 36 months.

A Poisson random variable could be used to model these data.

3. [5] Two different data sets, Group A and Group B, consist of 100 observations each. Below is a boxplot for the data in the Group A data set and a relative frequency histogram for the data in the Group B data set.





Circle the Roman numeral corresponding to the correct answer:

(a) The value of the sample skewness for Group A is:

$\boxed{(i)}$ greater than 0 (ii) approximately equal to 0 (iii) less than 0

Reason: Boxplot indicates a long right tail.

(b) The value of the sample mean and the sample median are approximately equal for the data in

(i) Group A $\boxed{(ii)}$ Group B

Reason: The relative frequency histogram is reasonably symmetric for Group B.

(c) The sample kurtosis for Group B is:

(i) greater than 3 (ii) approximately equal to 3 $\boxed{(iii)}$ less than 3

Reason: The relative frequency histogram is very uniform and has fewer observations in the tails than expected with Gaussian data.

(d) The lower or first quartile for Group B would be in the interval:

(i) $[-2, -1]$ $\boxed{(ii)}$ $[-1, 0]$ (iii) $[0, 1]$

Reason: The height of the first rectangle is $0.11 < 0.25$ while the sum of the heights of the first 2 rectangles $= 0.11 + 0.15 = 0.26 > 0.25$ so the first quartile must be in the interval $[-1, 0]$.

(e) The IQR for Group A is approximately equal to:

$\boxed{(i)}$ 2.3 (ii) 7.0 (iii) 12.5

Reason: Top of box is at approximately 4 while the bottom of the box is at approximately 1.7 so the difference is approximately $4 - 1.7 = 2.3$.

4. [5] Researchers at the Get Well Hospital were interested in studying the relationship between cholesterol levels and smoking in adult women. A sample of n women was selected at random from a list of all female nurses employed at the hospital. Cholesterol levels for each nurse were determined using a blood sample and each nurse was asked how many cigarettes she smoked per day.

Fill in the blanks below.

(a) In this study the units are (adult) women.

(b) The response variate is cholesterol levels

and the explanatory variate is number of cigarettes smoker per day.

Note: A variate is a characteristic measured for each unit in the study.

(c) The type of the explanatory variate is discrete.

(d) What type of study is this? observational (sample survey)

Note: This is an observational study since information was collected without any attempt to change the units in any way.

(e) One attribute of interest for this study would be

mean cholesterol level of adult women

or mean number number of cigarettes smoked per day by adult women.

Note: An attribute is a function of a variate which is defined for all units in the population or process of interest.