1. [10 marks] The following data set $y = (y_1, y_2, \ldots, y_{23})$ consists of the rent per month in dollars paid by twenty-three STAT 231 students in the winter term 2016:

375	396	429	432	513	522	528	544	561	564	589	672
699	717	735	762	764	846	926	935	977	1038	1178	

$$\sum_{i=1}^{23} y_i = 15702 \qquad \sum_{i=1}^{23} y_i^2 = 11784130$$

(a) Circle the letter corresponding to your choice.

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

[✓ **A**]: sqrt(var(y))

B: stdev(y)

C: var(y)

D: fivenum(y)

E: None of the Above

(ii) Which of the following commands in R generates 19 observations from an Exponential(4) distribution?

A: rexp(19, 4)

B: rpoisson(19, 4)

[✓ **C**]: rexp(19, 1/4)

D: rexponential(19, 1/4)

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 mean = 682.696 $\qquad \dfrac{15702}{23} = 682.695652$

IQR (interquartile range) = 324 $\qquad IQR = y_{(18)} - y_{(6)} = 846 - 522 = 324$

$\hat{F}(530)$ = 0.304 $\qquad \hat{F}(530) = \dfrac{\text{no. observations } \leq 530}{23} = \dfrac{7}{23} = 0.3043478$

(ii) Suppose 50 dollars was added to every rent. Find the following for the transformed data set (that is, the data set with 50 added to every observation):

sample mean = 732.696 $\qquad \dfrac{15702}{23} + 5 = 732.695652$

sample variance = 48383.767 $\qquad \dfrac{1}{22}\left[11784130 - \dfrac{1}{23}(15702)^2\right] = 48383.7668$

(iii) The **original** prices were converted from Canadian currency (CAD) to Chinese currency (CNY), by multiplying by 5.05. For the data set recorded in CNY find:

sample median = 3393.6 $\qquad (5.05)\, y_{(12)} = (5.05)\, 672 = 3393.6$

sample standard deviation = 1110.814 $\qquad (5.05)\left\{\dfrac{1}{22}\left[11784130 - \dfrac{1}{23}(15702)^2\right]\right\}^{1/2} = 1110.81367$

(iv) A 24th observations was added to the **original** data set. If the new observation was 530 find the following for the data set of 24 observations:

sample median = 630.5 $\qquad \dfrac{1}{2}(589 + 672) = 630.5$

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) 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 .

(b) Graphical summaries such as frequency histograms and boxplots are examples of

descriptive statistics .

(c) A proof of the Central Limit Theorem is a form of deductive reasoning.

(d) A researcher wishes to design a study to decide whether or not taking vitamin D reduces the risk of contracting influenza in children between the ages of 5 and 15.

This is an example of a hypothesis testing problem.

(e) A market research company conducts a poll to determine what proportion of Canadians feel safe living in Canada.

This is an example of a estimation problem.

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

The value of θ may be estimated using the sample mean .

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

The standard deviation of $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is equal to $\theta/\sqrt{n}$.

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

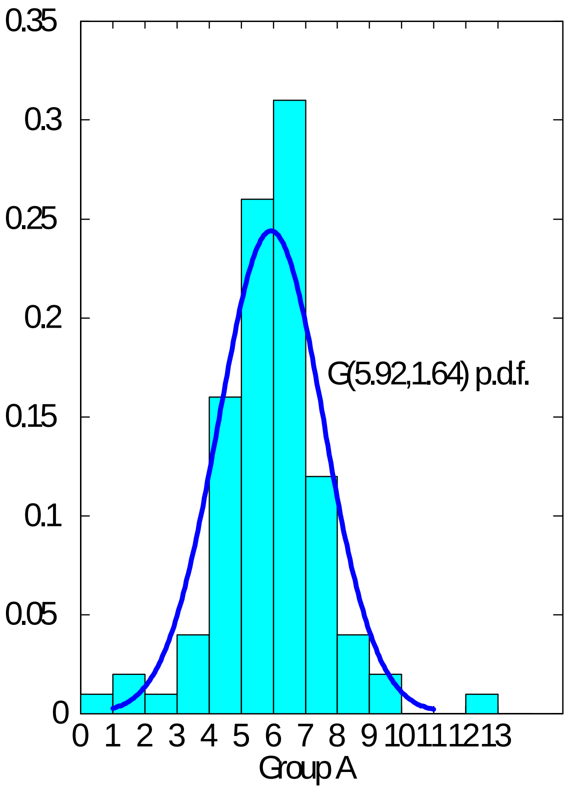
interquartile range (IQR) .

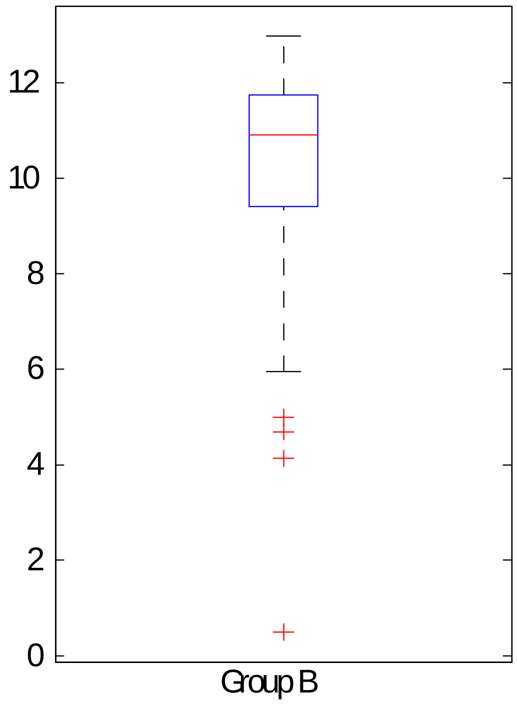
(i) For categorical bivariate data the relative risk is a measure of the relationship between two variates.

(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 relative frequency histogram for the data in the Group A data set and a boxplot for the data in the Group B data set.





Circle the Roman numeral correponding to the correct answer:

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

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

Reason: Boxplot indicates a long left tail.

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

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

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

(c) The sample kurtosis for Group A is:

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

Reason: The relative frequency histogram is more peaked in the center and has more observations in the tails than expected with Gaussian data.

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

(i) $[4, 5]$ $\boxed{(ii)}$ $[5, 6]$ (iii) $[6, 7]$

Reason: The sum of the heights of the first 5 rectangles $= 0.01 + 0.02 + 0.01 + 0.04 + 0.16 = 0.24 < 0.25$ while the sum of the heights of the first 6 rectangles $= 0.01 + 0.02 + 0.01 + 0.04 + 0.16 + 0.26 = 0.50 > 0.25$ so the first quartile must be in the interval $[5, 6]$.

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

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

Reason: Top of box is at approximately 11.8 while the bottom of the box is at approximately 9.5 so the difference is approximately $11.8 - 9.5 = 2.3$.

4. [5] In a recent British study on the the effectiveness of treatments for prostate cancer 1643 men were randomly assigned to one of three treatment groups: prostate surgery, radiation therapy, and regular testing to see if the cancer had spread. After ten years the number of deaths due to prostate cancer were observed in each of the treatment groups.

Fill in the blanks below.

(*a*) In this study the units are men.

(*b*) The response variate is whether or not the man died of prostate cancer (after 10 years).

and the explanatory variate is treatment group

(prostate surgery, radiation therapy, or regular testing to see if the cancer had spread)

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

(*c*) The type of the explanatory variate is categorical.

(*d*) What type of study is this? experimental

Note: This is an experimental study since the researchers controlled which treatment group (the explanatory variate) each man was assigned to.

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

the proportion of men in the prostate surgery treatment group
who had died of prostate cancer.

also: the proportion of men in the radiation therapy group who had died of prostate cancer

or: the proportion of men in the regular testing group who had died of prostate cancer

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