Homework Assignment 2

Jing Su, PhD

Oct 9, 2022

Problem set policies. *Please provide concise, clear answers for each question. Note that only writing the result of a calculation (e.g., "SD = 3.3") without explanation is not sufficient. For problems involving* R*, include the code in your solution, along with any plots.*

Please submit your problem set via Canvas as a PDF, along with the R Markdown source file.

We encourage you to discuss problems with other students (and, of course, with the course head and the TFs), but you must write your final answer in your own words. Solutions prepared "in committee" are not acceptable. If you do collaborate with classmates on a problem, please list your collaborators on your solution.

Unit 1

Problem 1. (*10 pionts*)

Since states with larger numbers of elderly residents would naturally have more nursing home residents, the number of nursing home residents in a state is often adjusted for the number of people 65 years or older (65+). That adjustment is usually given as the number of nursing home residents age 65+ per 1,000 members of the population age 65+. For example, a hypothetical state with 200 nursing home residents age 65+ and 50,000 people age 65+ would have the same adjusted number of residents as a state with 400 residents and a total age 65+ population of 100,000 – 4 residents per 1,000.

The data file `nursing.home.Rdata` contains this adjusted number of residents for each state in the United States. The state names are saved under the variable name `state` and the adjusted number of residents under the variable name `resident`.[1]

Hint: use the R functions `setwd('your work directory')` to set your work directory to where the data file locates. Then use `load('nursing.home.Rdata')` to load the data file. To find where your current work directory is, use `getwd()`.

a) How many variables are included in this data file? Please specify the data types of these variables. Hint: options for data types are: discrete numeric, continous numeric, ordinal categoric, and nominal categoric. (*2 pionts*)

b) Which row has the smallest number of nursing home residents per 1000 population 65 years of age and over? Which raw has the largest number? Hint: use the R functions `which.min()` and `which.max()` to find the index of the row. (*2 pionts*)

c) Which state has the smallest number of nursing home residents per 1000 population 65 years of age and over? Which state has the largest number? Hint: use the indexes found by the R functions `which.min()` and `which.max()` to index the state names. (*2 pionts*)

d) Construct a boxplot for the number of nursing home residents per 1,000 population. (*2 pionts*)

e) According to the boxplot, is the distribution of nursing home resident per 1000 population symmetric or skewed? Are there any states that could be considered outliers? (*2 pionts*)

f) Display the number of nursing home residents per 1000 population using a histogram. Explain your choice of bin numbers. (*2 pionts*)

Answwer:

```
#load the nursing.home dataset
setwd('/users/sreyatummala/downloads/')
load('nursing.home.Rdata')
getwd()
```

[1]The data originally appeared in Chapter 12 of *Case Studies in Biometry*, 1994, by Lange et al.

```
## [1] "/Users/sreyatummala/Downloads"
```

a) There are two variables namely state and resident in the data file nurseing.home.Rdata. State is an ordinal categorical variable and resident is discrete numerical variable.

b) Row 12 has the smallest number of nursing home residents and row 42 has the maximum.

```
#smallest number
which.min(nursing.home$resident)
```

```
## [1] 12
```

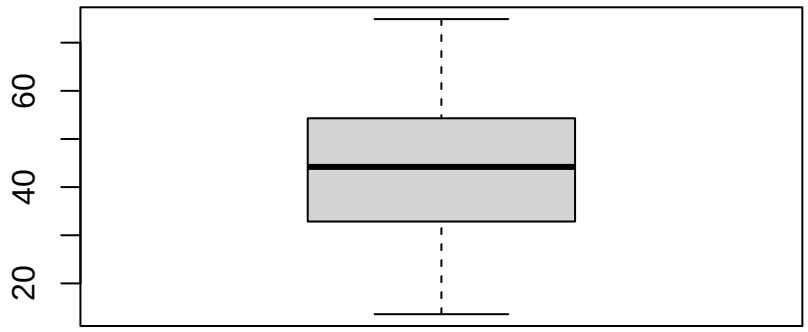
```
#maximum
which.max(nursing.home$resident)
```

```
## [1] 42
```

c) Hawaii- smallest South Dakota- largest

d) Code:

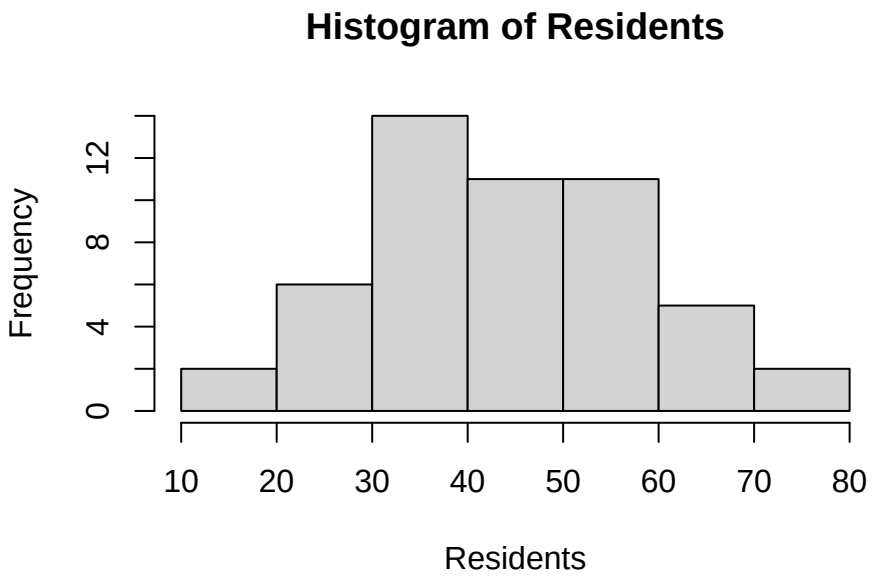
```
#construct a boxplot
boxplot(nursing.home$resident)
```



e)The boxplot represents the distribution to be symmetric as the median is present at the center. There aren't any states to be considered as outliers as no points are present above the whiskers.

f) I have taken 7 as the bin number as the functions whichmin() and whichmax() gave the least count and maximum count of residents.

```
#construct a histogram
Residents <- nursing.home$resident
hist(Residents,breaks = 7)
```



Problem 2. (*10pionts*)

The file `adolescent.fertility.Rdata` contains data on the number of children born to women aged 15-19 from 189 countries around the world for the years 1997, 2000, 2002, 2005, and 2006.[2] The data are defined using a scaling similar to that used in the nursing home data. The values for the annual adolescent fertility rates represent the number of live births among women aged 15-19 per 1,000 women members of the population of that age.

For the years 2000-2006, the adolescent fertility rate for Iraq is coded `NA`, or missing. When calculating a mean or standard deviation in R for a variable `x` which has missing data, add `na.rm=TRUE` to the argument to perform the calculations without the missing observations: `mean(x, na.rm=TRUE); sd(x, na.rm=TRUE)`.

a) Calculate the mean, standard deviation, and five-number summary for the distribution of adolescent fertility in 2006 (`fert_2006`). (*4 pionts*)

b) Note that the `summary()` command in R produces six numbers; specify which five belong in the five-number summary as defined in lecture. (*2 pionts*)

c) What is the 75^{th} percentile of the distribution? Write a sentence explaining the 75^{th} percentile in the context of this data. (*2 pionts*)

d) Use a single `boxplot` command to produce side-by-side boxplots of the fertility rates for each of the five years in the dataset. What pattern do you see? (*2 pionts*)

Answwer:

a) Mean = 53.58395 Code: mean(adolescent.fertility$fert_2$006, $na.rm = TRUE$)$StandardDeviation$ = 46.97848$Code$: $sd(adolescent.fertility$fert_2006, na.rm = TRUE) 5-number summary: Min = 1.453, 1st Qu.= 17.876, Median= 40.068, 3rd Qu.= 75.727, Max. = 223.834 Code: fivenum(adolescent.fertility$fert_2006)

```
#load adolescent.fertility dataset
setwd('/users/sreyatummala/downloads/')
load('adolescent.fertility.Rdata')
getwd()
```

```
## [1] "/Users/sreyatummala/Downloads"
```

```
#calculate the mean, SD, and 5-number summary
#adolescent.fertility$fert_2006
#mean
mean(adolescent.fertility$fert_2006, na.rm = TRUE)
```

```
## [1] 53.58395
```

[2] Data from the CIA World Factbook

```
#standard deviation
sd(adolescent.fertility$fert_2006, na.rm = TRUE)
```

```
## [1] 46.97848
```

```
#5-number summary
fivenum(adolescent.fertility$fert_2006, na.rm = TRUE)
```

```
## [1]   1.4534  17.8518  40.0682  76.0476 223.8336
```

b) R summary() produces, Min, 1st Qu., Median, Mean, 3rd Qu., Max, and any NA's if the data set contains. Min = 1.453, 1st Qu.= 17.876, Median= 40.068, Mean= 53.584, 3rd Qu.= 75.727, Max. = 223.834, NA's = 1

The fivenumber summary contains Min, 1st Qu., Median, 3rd Qu. And Max Min = 1.453, 1st Qu.= 17.876, Median= 40.068, 3rd Qu.= 75.727, Max. = 223.834

```
#summary
summary(adolescent.fertility$fert_2006)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.453  17.876  40.068  53.584  75.727 223.834       1
```

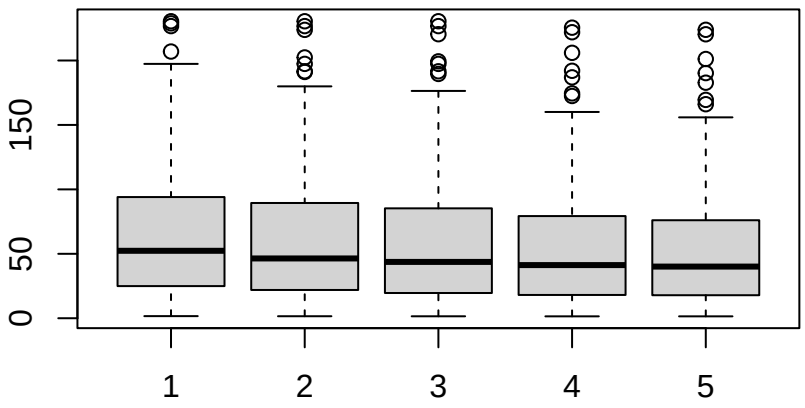
```
fivenum(adolescent.fertility$fert_2006)
```

```
## [1]   1.4534  17.8518  40.0682  76.0476 223.8336
```

c) The 3rd Quartile represents the 75th percentile of the distribution. The five number summary represents 3rd Qu. = 75.727 as the 75th percentile for the given data.

d) The representation of boxplots show that from 1997-2006. There isn't much difference among the adolescent fertility rates as the median is almost similar in all years. The adolescent fertility rates have reduced from 1998 to 2006 as the maximum and interquartile ranges show a decreased number, Overall, positive skewness is observed across all years. Also, the number of outliers have increased from 1997 to 2006.

```
#graphical summary
boxplot(adolescent.fertility$fert_1997,adolescent.fertility$fert_2000,adolescent.fertility$fert
```


0
50
150
1
2
3
4
5

Problem 3. (*8 pionts*)

Suppose that you are interested in determining whether a relationship exists between the fluoride content in a public water supply and the dental caries experience of children using this water. The file `water.Rdata` contains the data from a study examining 7,257 children in 21 cities from the Flanders region in Belgium.

The fluoride content of the public water supply in each city, measured in parts per million (ppm), is saved under the variable name `fluoride`; the number of dental caries per 100 children examined is saved under the name `caries`. The total dental caries number is obtained by summing the numbers of filled teeth, teeth with untreated dental caries, teeth requiring extraction, and missing teeth.[3]

a) How many variables are included in this data file? Please specify the data types of these variables. Hint: options for data types are: discrete numeric, continous numeric, ordinal categoric, and nominal categoric. (*2 pionts*)

b) Construct a two-way scatterplot for these data, with `fluoride` as the x-variable and `caries` as the y-variable. (*2 pionts*)

c) Calculate the correlation between `fluoride` and `caries`. (*2 pionts*)

d) Do `fluoride` and `caries` appear to be positively or negatively associated? Explain your answer. (*2 pionts*)

Answwer:

a) There are two variables fluoride and caries in the given dataset Water.Rdata Fluoride- continuous numerical variable Caries- discrete numerical variable

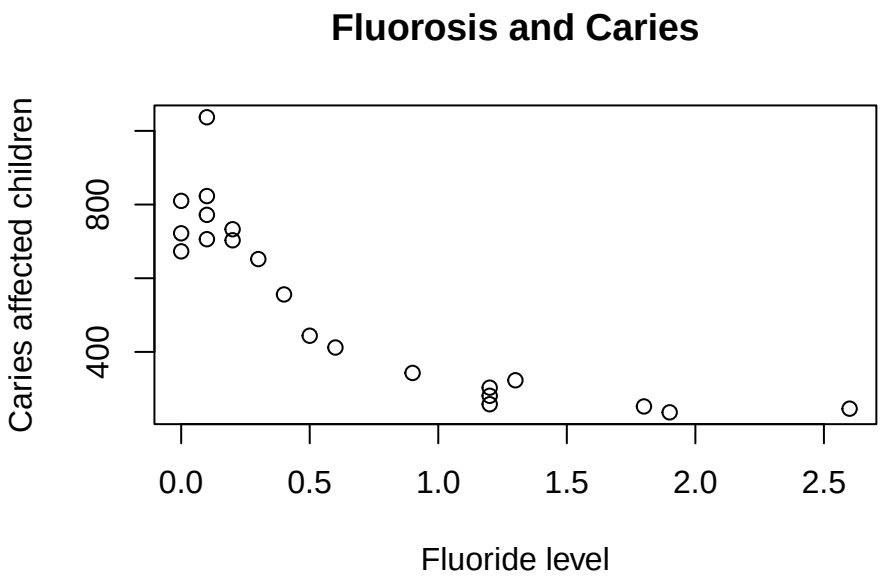
```
#load water dataset
setwd('/users/sreyatummala/downloads/')
load('water.Rdata')
getwd()
```

```
## [1] "/Users/sreyatummala/Downloads"
```

b)

```
#construct a two-way scatterplot
x <- water$fluoride
y <- water$caries
plot(x,y, main = "Fluorosis and Caries", xlab = "Fluoride level",ylab = "Caries affected childr
```

[3] These data appear in Table B21 in *Principles of Biostatistics*, 2nd ed. by Pagano and Gauvreau.



c) Code for Pearson's correlation: cor(x,y), value = -0.857029 Code for Spearman's correlation: cor(x,y, method = "spearman"), value= -0.914098

```
#calculate the correlation
cor(x,y)
```

```
## [1] -0.857029
```

```
cor(x,y, method = "spearman")
```

```
## [1] -0.914098
```

d) The given variables are negatively correlated to each other. This means that, increase in the amount of Fluoride content is inversely related to the total number of children affected with caries. It indicates that the variables are strongly negatively correlated to each other as the value is close to -1.

Problem 4. (*24 pionts*)

This problem features data from the *FAMuSS* (*Functional SNPs Associated with Muscle Size and Strength*) study discussed in lecture. The study examined the possible genetic determinants of skeletal muscle size and strength, before and after training.

This problem uses the following variables from the FAMuSS data:

- `ndrm.ch`: the percent change in strength in a participant's non-dominant arm, from before training and after.
- `drm.ch`: the percent change in strength in a particpant's dominant arm.
- `actn3.r577x`: the genotype at residue *r577x* within the *ACTN3* gene.
- `race`: race of the participant, with values stored as text strings.

The `famuss` dataset is in the `oibiostat` package.

a) How many variables are included in this data file? Please specify the data types of these variables. Hint: options for data types are: discrete numeric, continous numeric, ordinal categoric, and nominal categoric. (*6 pionts*)

b) Make a table of the genotypes for the SNP `actn3.r577x`. (*2 pionts*)

c) Construct a table of `actn3.r577x` by race, with the genotypes in the columns of the table and races in the rows. The command for creating a two-way table of categorical variables x and y is: `table(x, y)`. (*2 pionts*)

d) Provide numerical summaries to describe the `ndrm.ch` variable. Use both the mean and standard deviation and the five-number summary. (*4 pionts*)

e) Provide graphic summaries to describe the `ndrm.ch` variable. Use both boxplot and histogram. (*4 pionts*)

f) If you were to use numerical summaries to describe the `ndrm.ch` variable, would you prefer the mean and standard deviation or the five-number summary? Why? (*2 pionts*)

g) Produce a graphical summary that shows the association between `age` and genotype at the SNP `actn3.r577x`. Describe what you see. (*4 pionts*)

Answwer:

a) 9 variables are included in the data file Ndrm.ch – Continuous numerical data Drm.ch – Continuous numerical data Sex – Nominal categorical data Age – Discrete numerical data Race – Nominal categorical data Height – Continuous numerical Weight – Continuous numerical Actn3.r577x – Nominal categorical BMI – Continuous numerical

```r
#load the data
library(oibiostat)
data("famuss")
getwd()
```

```
## [1] "/Users/sreyatummala/Downloads"
```

b) CC CT TT 173 261 161

```r
#make table for actn3.r577x
table(famuss$actn3.r577x)
```

```
##
##  CC  CT  TT
## 173 261 161
```

c)

```r
#make table of actn3.r577x by race
a <- famuss$race
b <- famuss$actn3.r577x
table(a,b)
```

```
##              b
## a              CC  CT  TT
##   African Am  16   6   5
##   Asian       21  18  16
##   Caucasian  125 216 126
##   Hispanic     4  10   9
##   Other        7  11   5
```

d) Numerical summaries: Mean = 53.29109

Standard deviation = 33.13923

Fivenumber summary = 0.0 ,30.0 , 45.5, 66.7, 250.0

```r
#numeric summaries
mean(famuss$ndrm.ch)
```

```
## [1] 53.29109
```

```
sd(famuss$ndrm.ch)
```

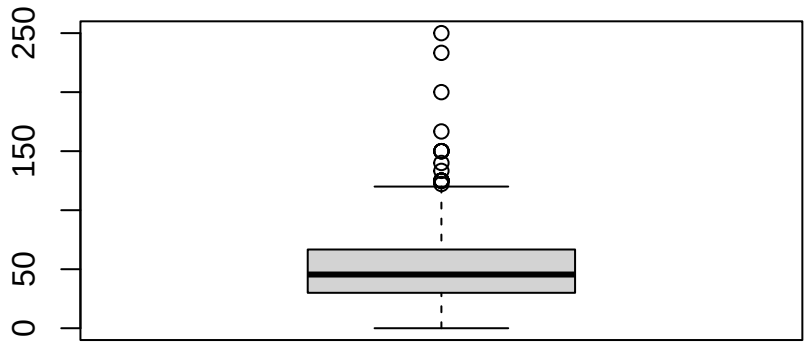
```
## [1] 33.13923
```

```
fivenum(famuss$ndrm.ch)
```

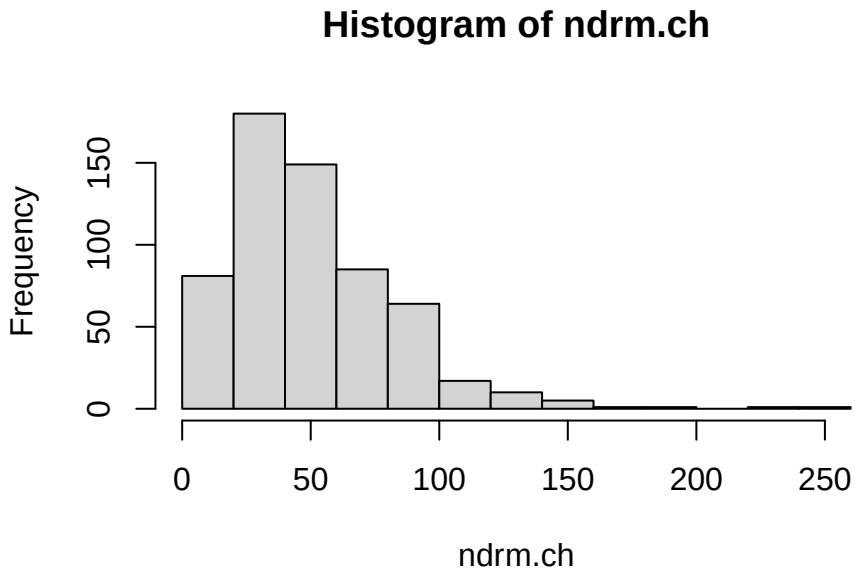
```
## [1]   0.0  30.0  45.5  66.7 250.0
```

e) Boxplot and Histogram

```
#graphic summaries
boxplot(famuss$ndrm.ch)
```



```
hist(famuss$ndrm.ch, main = "Histogram of ndrm.ch", xlab = "ndrm.ch")
```



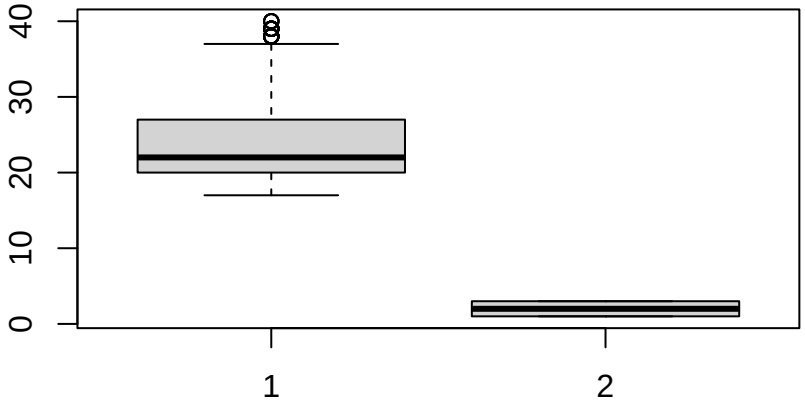
f) I would prefer to use fivenumber summary as there is detailed information on the quartiles (1st and 3rd). The minimum and maximum data points are also well known.

```
#hint: is the data skewed?
fivenum(famuss$ndrm.ch)
```

```
## [1]   0.0  30.0  45.5  66.7 250.0
```

g) It shows that the median of genotype of SNP at actn3.r577x is located outside the box of age. This shows that there is no relation or association between the two variables. Also, there are no minimum maximum values, whiskers and outliers for genotype of SNP at actn3.r577x. However, the boxplot of age shows positive skewness with potential outliers at the maximum value.

```
#graphical summary
boxplot(famuss$age,famuss$actn3.r577x)
```


0
10
20
30
40
1
2

Unit 2

Useful Formatting Notes.

In the following problems, you may need to show your work by including equations.

It is best to enclose any in-line equations, including math operators, within two $ symbols, e.g. $0.40+0.02=0.42$. The following operators may be useful: $\times$, $\cdot$, $\cap$, $\cup$, and $\neq$. To create a superscript, A^C. To create a subscript, A_X.

To typeset fractions, use the command $\frac{numerator}{denominator}$.

For your convenience, the syntax for generating the PPV equation and Bayes' Rule is given:

$$P(D|T^+) = \frac{P(D) \cdot P(T^+|D)}{P(T)} = \frac{P(T^+|D) \cdot P(D)}{[P(T^+|D) \cdot P(D)] + [P(T^+|D^C) \cdot P(D^C)]}$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

Problem 5. (*10 pionts*)

Suppose you are helping the police develop a new test for blood alcohol levels. The potential advantage of the test is that it can be used during a routine traffic stop, but neither you nor the police are sure how accurate the test is. By design, the test should be positive when blood alcohol level is above 0.05%.

The test has been used on a large number of subjects where a more expensive, less convenient test is known to give perfectly accurate results. The table below shows the joint distribution for the outcome of the new test and true blood alcohol status:

	Alcohol level > 0.05%	
Test Result	**Yes**	**No**
Positive	0.08	0.30
Negative	0.02	0.60

a) What is the probability that a randomly selected driver has both a positive test and a blood alcohol level higher than 0.05%? (*2 pionts*)

b) For a randomly selected driver, what is the probability that the test will be positive? (*2 pionts*)

c) Among drivers who have a positive test, what is the probability that a driver has a blood alcohol level higher than 0.05%? (*2 pionts*)

d) Suppose A is the event that a driver has a positive test result and B is the event that the driver has a blood alcohol level above 0.05%.

   i. Why might it be reasonable to expect that A and B are not independent? Explain your answer. (*2 pionts*)

   ii. Are A and B independent? Justify your answer. (*2 pionts*)

Answwer:

a) This statement is an example of conditional probability. Hence the expression is $P(A|B) = \frac{P(A \cap B)}{P(B)}$ $P(Positivetest|Alcohollevel > 0.05) = 0.08$ The probability that a randomly selected driver has both a positive test and a blood alcohol level higher than 0.05 is 0.08.

b) The blood alcohol level above 0.05 and below 0.05 won't happen at the same time. So, they are two disjoint or mutually exclusive events. Thus the expression for this scenarios is $P(A \cup B) = P(A) + P(B)$ $P(Postive) = P(0.08) + P(0.30) = 0.38$ For a randomly selected driver,the probability that the test will be positive is 0.38.

c)This concept in this question is an example of conditional probability. Hence denoted as P(A|B).

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(Alcohollevel > 0.05 \cap Positivetestresult)}{P(Positivetestresult)} = \frac{0.08}{0.38} = 0.2105$$

d) i. According to the given question, the test is designed in such a way that the test should be positive when the blood alcohol level is greater than 0.05%. But the data given suggests that the test results are negative for 0.02% of the people when the alcohol levels are still greater than 0.05%. This suggests that the positive test results and blood alcohol levels are not independent events.

ii. From the given question, P(A) is the probabilty of positive test result, i.e P(Positive test result) $P(A) = 0.38$ P(B) is the probability of alcohol greater than 0.05, i.e $P(Alcohollevel > 0.05)$ $P(B) = 0.08 + 0.02 = 0.10$ So the expression for two independent events is $P(A \cap B) = P(A) \cdot P(B)$ $P(A \cap B) = P(Positivetestresult \cap Alcohollevel \geq 0.05) = 0.08$ $P(A) \cdot P(B) = (0.38) \times (0.10) = 0.038$ $AsP(A \cap B) \neq P(A) \cdot P(B)$, A and B are not independent events Hence Positive test result and blood alcohol level greater than 0.05% are not independent events.

Problem 6. (*10 pionts*)

The strongest risk factor for breast cancer is age; as a woman gets older, her risk of developing breast cancer increases. The following table shows the average percentage of American women in each age group who develop breast cancer, according to statistics from the National Cancer Institute. For example, approximately 3.56% of women in their 60's get breast cancer.

Table 1: Prevalence of Breast Cancer by Age Group

Age Group	**Prevalence**
30 - 40	0.0044
40 - 50	0.0147
50 - 60	0.0238
60 - 70	0.0356
70 - 80	0.0382

A mammogram typically identifies a breast cancer about 85% of the time, and is correct 95% of the time when a woman does not have breast cancer.

a) If a woman in her 60's has a positive mammogram, what is the likelihood that she has breast cancer? Solve this problem algebraically. (*4 pionts*)

b) Using whatever methods you wish, calculate the PPV for each age group; show your work. Describe the trend in PPV values as prevalence changes and explain the reasoning behind the relationship between prevalence and PPV. (*4 pionts*)

c) Suppose that two new mammogram imaging technologies have been developed which can improve the PPV associated with mammograms; one improves sensitivity to 99% (but specificity remains at 95%), while the other improves specificity to 99% (while sensitivity remains at 85%). Which technology offers a higher increase in PPV? Explain your answer. (*2 pionts*)

Some notes on including answers from an image:

To include an image in your solutions, upload the image file to your pset_02 folder, then use the following syntax in the comment block (i.e. remove the). You can either take a picture of a diagram/table drawn on paper or use software such as MS PowerPoint to draw a diagram/table.

Some helpful syntax for including work done with the table method:

	Cancer	No cancer	Total
Positive	number	number	number
Negative	number	number	number
Total	number	number	number

Answwer:

a) Given, Prevalence = 0.0356 Sensitivity = 0.85 Specificity = 0.95 False Negative Rate = 1 - Sensitivity = 1 - 0.85 = 0.15 False Positive Rate = 1 - Specificity = 1 - 0.95 = 0.05 According to Baye's rule,

$$P(D|T^+) = \frac{P(D) \cdot P(T^+|D)}{P(T)} = \frac{P(T^+|D) \cdot P(D)}{[P(T^+|D) \cdot P(D)] + [P(T^+|D^C) \cdot P(D^C)]}$$

$$P(BreastCancer|Positivemammogram) = \frac{Sensitivity \times Prevalence}{[Sensitivity \times Prevalence] + [(1 - Specificity) \times (1 - Prevalence)]}$$

$$= \frac{0.85 \times 0.0356}{[0.85 \times 0.0356] + [(0.05) \times (1 - 0.0356)]} = \frac{0.03026}{0.03026 + [(0.05) \times (0.9644)]} = \frac{0.03026}{0.03026 + 0.04822}$$

$$= \frac{0.03026}{0.07848} = 0.3855 = 38.56$$

The likelihood(probability) for a women in her 60s with breast cancer and positive mammogram test is 38.56%.

b)PPV for 30-40 age group is 0.06988042 PPV for 40-50 age group is 0.2023154 PPV for 50-60 age group is 0.2930185 PPV for 60-70 age group is 0.3855759 PPV for 70-80 age group is 0.4030536 The PPV values increased with increase in prevalence.To explain the relation, let us consider the PPV formula taken in the question. 1-Prevalence value influences the PPV values. If prevalence is small and close to zero, numerator also would be small and the PPV values are reduced. Vice verasa happens when prevalence values are large or close to 1.

Table 2: Prevalence of Breast Cancer by Age Group

Age Group	**Prevalence**	**PPV**
30 - 40	0.0044	0.070
40 - 50	0.0147	0.202
50 - 60	0.0238	0.293
60 - 70	0.0356	0.386
70 - 80	0.0382	0.403

```
#calculations
# 30-40 age group
sensitivity <- 0.85
specificity <- 0.95
prevalence <- 0.0044
```

```r
numerator <- prevalence*sensitivity
denomenator <- prevalence*sensitivity + (1-specificity)*(1-prevalence)
ppv <- numerator/denomenator
print(ppv)
```

```
## [1] 0.06988042
```

```r
#40-50
prevalence <- 0.0147
numerator <- prevalence*sensitivity
denomenator <- prevalence*sensitivity + (1-specificity)*(1-prevalence)
ppv <- numerator/denomenator
print(ppv)
```

```
## [1] 0.2023154
```

```r
#50-60
prevalence <- 0.0238
numerator <- prevalence*sensitivity
denomenator <- prevalence*sensitivity + (1-specificity)*(1-prevalence)
ppv <- numerator/denomenator
print(ppv)
```

```
## [1] 0.2930185
```

```r
#60-70
prevalence <- 0.0356
numerator <- prevalence*sensitivity
denomenator <- prevalence*sensitivity + (1-specificity)*(1-prevalence)
ppv <- numerator/denomenator
print(ppv)
```

```
## [1] 0.3855759
```

```r
#70-80
prevalence <- 0.0382
numerator <- prevalence*sensitivity
denomenator <- prevalence*sensitivity + (1-specificity)*(1-prevalence)
ppv <- numerator/denomenator
print(ppv)
```

```
## [1] 0.4030536
```

c) Increase in specificity raises the PPV values. The prevalence of the given disease ranges from 1-4 percent. The reasons for the low PPV values are due to a high number of false positive cases. This value roots back to specificity. The number of true negative cases increases with a decrease in false positives when specificity increases. Sensitivity can increase the number of true positive cases and rise PPV inturn but this effect is very minimal. This can be demonstrated with the example below.

```
# prevalence for age group 50-60 is taken as example
#change in sensitivity
sensitivity <- 0.99
specificity <- 0.95
prevalence <- 0.0238
numerator <- prevalence*sensitivity
denomenator <- prevalence*sensitivity + (1-specificity)*(1-prevalence)
ppv <- numerator/denomenator
print(ppv)
```

```
## [1] 0.3255679
```

```
#change in specificity
sensitivity <- 0.85
specificity <- 0.99
prevalence <- 0.0238
numerator <- prevalence*sensitivity
denomenator <- prevalence*sensitivity + (1-specificity)*(1-prevalence)
ppv <- numerator/denomenator
print(ppv)
```

```
## [1] 0.6745132
```

Unit 3

Useful Formatting Notes.

It is best to enclose any in-line equations, including math operators, within two $ symbols, e.g. $0.40+0.02=0.42$. The following operators may be useful: $\times$, $\cdot$, $\cap$, $\cup$, $\neq$, $\geq$, and $\leq$. To create a superscript, A^C. To create a subscript, P_X. To use the square root symbol, $\sqrt{x}$.

To typeset fractions, use the command $\frac{numerator}{denominator}$.

For your convenience, the following syntax is given:

$$\text{Var}(X) = E(X-\mu)^2 = \sum_i^k P(X=x_i)(x_i-\mu)^2$$

$$P(X=k) = \frac{e^{-\lambda}\cdot\lambda^k}{k!}$$

Problem 7. (*8 pionts*)

Let X be a random variable with the following probability mass function:

$X=x$	0	1	2	3
$P(X=x)$	0.10	0.20	0.30	0.40

a) Find $P(X \geq 2)$. (*2 pionts*)

b) Find $P(X \geq 2 | X \geq 1)$. (*2 pionts*)

c) Find $E(X)$. (*2 pionts*)

d) Find Var(X). (*2 pionts*)

Answwer:

a) $P(X \geq 2) = P(X=2) + P(X=3) = 0.30 + 0.40 = 0.70$

b) $P(X \geq 2 | X \geq 1) = \frac{P(X \geq 2) \cap P(X \geq 1)}{P(X \geq 1)}$ From the law of conditional probability $P(X \geq 2 | X \geq 1) = \frac{P(X \geq 2)}{P(X \geq 1)} = \frac{P(X \geq 2)}{1 - P(X < 1)} = 0.70/1 - 0.10 = 0.70/0.90 = 0.7778$

c) The expected value is 2.

```r
#use r as a calculator
#defining values
values <- c(0,1,2,3)

#defining probability
probability <- c(0.10,0.20,0.30,0.40)

# calculating expected value
expval<-sum(values*probability)
print(expval)
```

```
## [1] 2
```

d) The variance of X is 1

```r
#use r as a calculator

var <- sum(values*values*probability)
variance <- var-(expval*expval)
print(variance)
```

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

Problem 8. (*8 pionts*)

According to data from the CDC, about 37.1% of adults (individuals 18 years of age or older) in the United States and 57.9% of children (individuals between 6 months and 17 years of age) in the United States received a flu vaccine during the 2017-2018 flu season.

a) Consider a random sample of 50 adults.

   i. Calculate the probability that exactly 20 adults received a flu vaccine. (*2 pionts*)
   ii. Calculate the probability that exactly 30 adults did not receive a flu vaccine. (*2 pionts*)

b) Consider a random sample of 20 children.

   i. What is the probability that at most 10 children received a flu vaccine? (*2 pionts*)
   ii. What is the probability that at least 11 children received a flu vaccine? (*2 pionts*)

Answwer:

a) Here, x denotes vector of events n denotes number of trails p denotes probability

   i. x = 20 n = 50 p = 0.371

Hence probability that exactly 20 adults received a flu vaccine is 0.1047823.

```
dbinom(20,50,0.371)
```

```
## [1] 0.1047823
```

   ii. x = 30 n = 50 p = 0.371

Hence, the probability that exactly 30 adults did not receive a flu vaccine is 0.0005339743.

```
dbinom(30,50,0.371)
```

```
## [1] 0.0005339743
```

b) Here, x denotes the vector of events n denotes number of trails p denotes probability

   i. x = 10 (at most 10) n = 20 p = 0.579 Thus the probability that at most 10 children received a flu vaccine is 0.309678.

```
pbinom(10,20,0.579)
```

```
## [1] 0.309678
```

ii. x = 10 (at least 11) n = 20 p = 0.579

Thus the the probability that at least 11 children received a flu vaccine is 0.690322.

```
pbinom(10,20,0.579,lower.tail = FALSE)
```

```
## [1] 0.690322
```

Problem 9. (*8 pionts*)

Consider a senior Statistics concentrator with a packed extracurricular schedule, taking five classes, and writing a thesis. Each time she takes an exam, she either scores very well (at least two standard deviations above the mean) or does not. Her performance on any given exam depends on whether she is operating on a reasonable amount of sleep the night before (more than 7 hours), relatively little sleep (between 4 - 7 hours, inclusive), or practically no sleep (less than 4 hours).

When she has had practically no sleep, she scores very well about 30% of the time. When she has had relatively little sleep, she scores very well 40% of the time. When she has had a reasonable amount of sleep, she scores very well 42% of the time. Over the course of a semester, she has a reasonable amount of sleep 50% of nights, and practically no sleep 30% of nights.

a) What is her overall probability of scoring very well on an exam? (*2 pionts*)

b) What is the probability she had practically no sleep the night before an exam where she scored very well? (*2 pionts*)

c) Suppose that one day she has three exams scheduled. What is the probability that she scores very well on exactly two of the exams, under the assumption that her performance on each exam is independent of her performance on another exam? (*2 pionts*)

d) What is the probability that she had practically no sleep the night prior to a day when she scored very well on exactly two out of three exams? (*2 pionts*)

Answwer:

a) Given: A represents the event of reasonable amount of sleep B represents the event of practically no sleep C represents the event of relatively little sleep D represents the the event of scoring very well in the exam All these are independent events.So, P(A) = 0.50 P(B) = 0.30 P(C) = 0.20

P(D|A) = 0.42 P(D|B) = 0.30 P(D|C) = 0.40 We multiply the probability and add them as the events are independent and disjoint. $P(D) = P(D|A)P(A) + P(D|B)P(B) + P(D|C)P(C)$ $= (0.42) \times (0.50) + (0.30) \times (0.30) + (0.40) \times (0.20) = 0.38$

Hence the overall probability of scoring well on an exam is 0.38 or 38%

b) $P(B|D) = \dfrac{P(B \cap D)}{P(D)} = \dfrac{P(D|A) \times P(B)}{P(D)} = \dfrac{(0.30) \times (0.30)}{0.38} = 0.2368$

Hence the probability she had practically no sleep the night before an exam where she scored very well is 0.2368.

c) The probability that she scores very well on exactly two of the exams is 0.268584.

```
dbinom(2,3,0.38)
```

```
## [1] 0.268584
```

d) The probability that she had practically no sleep the night prior to a day when she scored very well on exactly two out of three exams is 0.1283876.

```
dbinom(2,3,0.2368)
```

```
## [1] 0.1283876
```

Problem 10. (*2 pionts*)

This is a simple exercise in computing probabilities for a Poisson random variable. Suppose that X is a Poisson random variable with rate parameter $\lambda = 2$.

Use R to calculate each of the 3 probabilities in part a). (*2 pionts*)

Answwer:

```
# Pr(X = 2)
first <- dpois(2,lambda = 2)
print(first)
```

```
## [1] 0.2706706
```

```
# Pr(X \leq 2)
second <- ppois(2,lambda = 2)
print(second)
```

```
## [1] 0.6766764
```

```
# Pr(X \geq 3)
third <- ppois(2, lambda = 2, lower.tail = FALSE)
print(third)
```

```
## [1] 0.3233236
```

Problem 11. (*8 pionts*)

Osteosarcoma is a relatively rare type of bone cancer. It occurs most often in young adults, age 10 - 19; it is diagnosed in approximately 8 per 1,000,000 individuals per year in that age group. In New York City (including all five boroughs), the number of young adults in this age range is approximately 1,400,000.

a) What is the expected number of cases of osteosarcoma in NYC in a given year? (*2 pionts*)

b) What is the probability that 15 or more cases will be diagnosed in a given year? (*2 pionts*)

c) The largest concentration of young adults in NYC is in the borough of Brooklyn, where the population in that age range is approximately 450,000. What is the probability of 10 or more cases in Brooklyn in a given year? (*2 pionts*)

d) Suppose that over five years, there was one year in which 10 or more cases of osteosarcoma were observed in Brooklyn. Is the probability of this event equal to the probability calculated in part c)? Explain your answer. (*2 pionts*)

Answwer:

a) The expected number of cases of osteosarcoma in NYC in a given year is 12 (we rounded off 11.2 to 12).

```
#use r as a calculator
la = 8/1000000
print(la)
```

```
## [1] 8e-06
```

```
1400000*la
```

```
## [1] 11.2
```

b) The probability that 15 or more cases will be diagnosed in a given year is 0.2279755

```
1-ppois(14,lambda = 12 )
```

```
## [1] 0.2279755
```

c) The probability of 10 or more cases in Brooklyn in a given year is 0.008132243

```
nycla <- 450000*la
print(nycla)
```

```
## [1] 3.6
```

```r
1- ppois(9, lambda = 4)
```

```
## [1] 0.008132243
```

d) No, the probabilities are not equal.Over the years the cases might accumulate as events occur independently at the rate lambda.

```r
#5 years
#lambdafive <- 18
1-ppois(9,18)
```

```
## [1] 0.9846189
```

```r
#one year
1-ppois(9,4)
```

```
## [1] 0.008132243
```

Problem 12. (*10 pionts*)

Consider the standard normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$.

a) What is the probability that an outcome z is greater than 2.30? (*2 pionts*)

b) What is the probability that z is less than 1.45? (*2 pionts*)

c) What is the probability that z is between -1.60 and 3.10? (*2 pionts*)

d) What value of z cuts off the upper 15% of the distribution? (*2 pionts*)

e) What value of z marks off the lower 20% of the distribution? (*2 pionts*)

Answwer:

a) The probability that an outcome z is greater than 2.30 is 0.01072411.

```
pnorm(q=2.30, lower.tail = FALSE)
```

```
## [1] 0.01072411
```

b) The probability that z is less than 1.45 is 0.9264707.

```
pnorm(q=1.45)
```

```
## [1] 0.9264707
```

c) The probability that z is between -1.60 and 3.10 is 0.9442331.

```
pnorm(3.10)- pnorm(-1.60)
```

```
## [1] 0.9442331
```

d) Hence z = 1.036433 cuts off the upper 15% of the distribution

```
qnorm(p=0.15, lower.tail = FALSE)
```

```
## [1] 1.036433
```

e) Hence z = -0.8416212 marks off the lower 20% of the distribution

```r
qnorm(p=0.20)
```

```
## [1] -0.8416212
```

Problem 13. (*2 pionts*)

The World Health Organization defines osteoporosis in young adults as a measured bone mineral density 2.5 or more standard deviations below the mean for young adults. Assume that bone mineral density follows a normal distribution in young adults. What percentage of young adults suffer from osteoporosis according to this criterion?

Answwer:

```
pnorm(q=-2.5)
```

```
## [1] 0.006209665
```

Problem 14. (*6 pionts*)

(Based on Problem 1.136 in IPS, 6^{th} edition.) High blood cholesterol levels increase the risk of heart disease. Young women are generally less afflicted with high cholesterol than other groups. The cholesterol levels for women aged 20 to 34 years follow an approximately normal distribution with mean 185 milligrams per deciliter (mg/dl) and standard deviation 39 mg/dl.

a) Cholesterol levels above 240 mg/dl demand medical attention. What percent of young women have levels above 240 mg/dl? (*2 pionts*)

b) Levels above 200 mg/dl are considered borderline high. What percent of young women have blood cholesterol between 200 and 240 mg/dl? (*2 pionts*)

c) Among a random sample of 150 women in this age group, what is the probability that no more than 5 women have cholesterol levels that demand medical attention? (*2 pionts*)

Answwer:

a) The percent of young women have levels above 240 mg/d is 0.07923199

```
pnorm(240,185,39, lower.tail = FALSE)
```

```
## [1] 0.07923199
```

b) The percent of young women have blood cholesterol between 200 and 240 mg/dl is 0.2710292

```
pnorm(240,185,39) - pnorm(200,185,39)
```

```
## [1] 0.2710292
```

c) The probability that no more than 5 women have cholesterol levels that demand medical attention is 0.0182394.

```
# 0.07923199 is obtained from 14.a)
pbinom(5,150,0.07923199)
```

```
## [1] 0.0182394
```

Problem 15. (*6 pionts*)

Hemophilia is a sex-linked bleeding disorder that slows the blood clotting process. In severe cases of hemophilia, continued bleeding occurs after minor trauma or even in the absence of injury. Hemophilia affects 1 in 5,000 male births. In the United States, there are approximately 4,000,000 births per year. Assume that there are equal numbers of males and females born each year.

a) What is the probability that at most 390 newborns in a year are born with hemophilia? (*2 pionts*)

b) What is the probability that 425 or more newborns in a year are born with hemophilia? (*2 pionts*)

c) Consider a hypothetical country in which there are approximately 2 million births per year. If the incidence rate of hemophilia is equal to that in the US, as well as the sex ratio at birth, how many newborns are expected to have hemophilia over five years, and with what standard deviation? (*2 pionts*)

Answwer:

a) The probability that at most 390 newborns in a year are born with hemophilia is 0.3197029.

```
lam <- (1/5000)*(4000000/2)
print(lam)
```

```
## [1] 400
```

```
ppois(390,lambda = lam)
```

```
## [1] 0.3197029
```

b) The probability that 425 or more newborns in a year are born with hemophilia 0.1020334.

```
ppois(425, lambda = lam, lower.tail = FALSE)
```

```
## [1] 0.1020334
```

c) 1000 newborns are expected to have hemophilia and with a standard deviation of 31.62278.

```
#use r as a calculator
newlam <- (1/5000)*(2000000/2)
print(newlam)
```

```
## [1] 200
```

```r
fiveyears <- newlam*5
print(fiveyears)
```

```
## [1] 1000
```

```r
sqrt(fiveyears)
```

```
## [1] 31.62278
```