Mohammad Javad Ranjbar

810101173

Homework 2

Statistical Inference, Fall 2022

Question #1:

$$A = (1,3), (2,2), (3,1) \Rightarrow P(A) = \frac{3}{36}$$

$$B = \{(1,5), (2,4), (3,3), (4,2), (5,1)\} \Rightarrow P(B) = \frac{5}{36}$$

$$C = \{(1,2), (2,2), (3,2), (4,2), (5,2), (6,2), (2,1), (2,3), (2,4), (2,5), (2,6)\} \Rightarrow P(C) = \frac{11}{36}$$

$$P(A, C) = (\text{at least one of the dice shows a } 2 \cap \text{ sum of two dice equals } 4) = \frac{1}{36}$$

$$P(A|C) = \frac{P(A, C)}{P(C)} = \frac{P(A, C)}{P(C)} = \frac{\frac{1}{36}}{\frac{11}{36}} = \frac{1}{11}$$

If A and C are independent the following equation should be correct: $P(A, C) = P(A) * P(C)$ however based on the below equation we can determine they are not independent.

$$\frac{3}{36} * \frac{11}{36} \neq \frac{1}{36} \Rightarrow P(A, C) \neq P(A) * P(C)$$

$$P(B, C) = (\text{sum of two dice equals } 6 \cap \text{ sum of two dice equals } 4) = \frac{2}{36}$$

$$P(B|C) = \frac{P(B, C)}{P(C)} = \frac{P(B, C)}{P(C)} = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11}$$

If A and B are independent the following equation should be correct. $P(A, B) = P(A) * P(B)$. however, based on the below equation we can determine they are not independent.

$$\frac{3}{36} * \frac{5}{36} \neq \frac{2}{36} \Rightarrow P(A, B) \neq P(A) * P(B)$$

Question #2:

$$P(F) = \frac{55}{100}, \quad P(CS) = \frac{8}{100}, P(F, CS) = \frac{3}{100}$$

a) The student is female given that the student is majoring in computer science.

$$P(F|CS) = \frac{P(F, CS)}{P(CS)} = \frac{\frac{3}{100}}{\frac{8}{100}} = \frac{3}{8}$$

b) The student is majoring in computer science given that the student is female.

$$P(CS|F) = \frac{P(F, CS)}{P(F)} = \frac{\frac{3}{100}}{\frac{55}{100}} = \frac{3}{55}$$

Question #3:

a) The variable is binomial. People either approve or disapprove of President George W. Bush and their choice is independent of other people choices. Therefore, there are only two choices and X is a binomial variable.

b) $P(X) = \binom{n}{x} (\frac{92}{100})^n (\frac{8}{100})^{n-x}$

$P(X \leq 358) = 1 - P(X \geq 358) = 0.044$

```
> pbinom(358,400,0.92)
[1] 0.04410268
```

c) For binomial distribution, the expected number is equal to $np = 400 * \frac{92}{100} = 368$ and the standard deviation is equal to $\sqrt{npq} = \sqrt{400 * \frac{92}{100} * \frac{8}{100}} = 5.42$.

Proof:

$$E(x) = \sum x_i P(x_i) = \sum_{x=0}^{n} x * \binom{n}{x} (p)^x (1-p)^{n-x} = \sum_{x=0}^{n} x * \frac{n!}{(n-x)!\, x!} (p)^x (1-p)^{n-x}$$

$$= \sum_{x=1}^{n} \frac{n!}{(n-x)!\,(x-1)!} (p)^x (1-p)^{n-x}$$

$$= np \sum_{x=1}^{n} \frac{(n-1)!}{(n-x)!\,(x-1)!} (p)^{x-1} (1-p)^{n-x}$$

$$\xrightarrow{\sum_{x=1}^{n} \frac{(n-1)!}{(n-x)!(x-1)!} (p)^{x-1}(1-p)^{n-x} = (p+q)^{n-1}} E(x) = np \sum_{x=1}^{n} (p+q)^{n-1} = np$$

$$\sigma^2 = E(x^2) - E(x)^2 \xrightarrow{E(x^2)=E(x(x-1)+x)} \sigma^2 = E(x(x-1)) + E(x) - (np)^2$$

$$= \sum_{x=0}^{n} x * (x-1) * \frac{n!}{(n-x)!\,x!}(p)^x(1-p)^{n-x} + np - (np)^2$$

$$= \sum_{x=2}^{n} x * (x-1) * \frac{n!}{(n-x)!\,x!}(p)^x(1-p)^{n-x} + np - (np)^2$$

$$= n * (n-1)p^2 \sum_{x=2}^{n} \frac{(n-2)!}{(n-x)!\,(x-2)!}(p)^{x-2}(1-p)^{n-x} + np - (np)^2$$

$$\xrightarrow{\sum_{x=2}^{n} \frac{(n-2)!}{(n-x)!(x-2)!}(p)^{x-2}(1-p)^{n-x}=(p+q)^{n-2}} \sigma^2 = n * (n-1)p^2 + np - (np)^2$$

$$= np(p-1) = npq \rightarrow \sigma = \sqrt{npq}$$

d) Because $np = 358 > 10 \Rightarrow$ we can use the normal distribution.

$$P(X \leq 358) \approx P(Z \leq \frac{358 - np}{\sqrt{npq}}) = P(Z \leq \frac{358 - 368}{5.42}) = P(Z \leq 1.8) = 0.035$$

```
> pnorm(1.8, mean=0, sd=1, lower.tail=FALSE)
[1] 0.03593032
```

Yes, the approximation is very close to the exact value. Therefore, it is satisfactory.

Question #4:

$$P(W|T1) = \frac{3}{10}, P(W|T2) = \frac{4}{10}, P(W|T3) = \frac{5}{10}$$

$$P(T1) = \frac{1}{2}, P(T2) = \frac{1}{4}, P(T3) = \frac{1}{4}$$

$$P(W) = P(W,T1) + P(W,T2) + P(W,T3) = \frac{3}{10} * \frac{1}{2} + \frac{4}{10} * \frac{1}{4} + \frac{5}{10} * \frac{1}{4} = \frac{3}{8}$$

The probability of winning against a random player is equal to $\frac{3}{8}$.

Question #5:

a) In reality, words in a sentence are dependent. For instance, some words only come when they are adjacent to some other special words. However, for the sake of simplicity, they can be considered as Poisson random variables. Because our data are counts of events and All events are independent. The average rate of occurrence does not change during the period of interest. If the sops' lengths are equal they both will have the same parameters.

b) We want to calculate the $P(X_i > 0, Y_i = 0)$ and Because they are independent, we have $P(X_i > 0, Y_i = 0) = P(X_i > 0) * P(Y_i = 0)$. Therefore:

$$(1 - P(y > 0)) * P(y = 0) = (1 - \frac{\lambda^0 e^{-\lambda}}{0!}) \frac{\lambda^0 e^{-\lambda}}{0!} = e^{-\lambda} - e^{-2\lambda}$$

Question #6:

a) $P(X = k) = (1-p)^{k-1} * p \overset{k=1}{\Longrightarrow} P(X = 1) = p$
b) Because the incident is independent of the past we have $P(X = k_1, X = k_2) = P(X = k_1) * P(X = k_2)$
$P(X = 5, X = 8) = P(X = 5) * P(X = 8) = p * p = p^2$

Question #7:

The question is describing a geometric random variable. Therefore, the mean and standard deviation of geometric distribution are $\mu = \frac{1}{p} \sigma = \sqrt{\frac{1-p}{p^2}}$

Proof:

$$E(x) = \sum x_i P(x_i) = \sum_{k=0}^{n} k * (1-p)^{k-1} * p = p \sum_{k=0}^{n} k * (1-p)^{k-1}$$
$$= p(0 + 2 * (1-p) + 3 * (1-p)^2 + \cdots)$$
$$= p * \frac{-d \sum_{k=1}^{n} (1-p)^k}{dp} \overset{\sum_{k=0}^{n} (1-p)^k = (1 + (1-p) + (1-p)^2 + \cdots) = \frac{1}{1-(1-p)}}{\Longrightarrow} E(x) = p * \frac{1}{p^2}$$
$$= \frac{1}{p}$$

$$\sigma^2 = E(x^2) - E(x)^2 \overset{E(x^2) = E(x(x-1)+x)}{\Longrightarrow} \sigma^2 = E(x(x-1)) + E(x) - \frac{1}{p^2}$$
$$= \sum_{k=0}^{n} k * (k-1) * (1-p)^{k-1} * p + \frac{1}{p} - \frac{1}{p^2}$$
$$= p(1-p) \sum_{k=0}^{n} k * (k-1) * (1-p)^{k-2} + \frac{1}{p} - \frac{1}{p^2}$$
$$= p * (1-p) * \frac{-d \frac{-d \sum_{k=0}^{n} (1-p)^k}{dp}}{dp} + \frac{1}{p} - \frac{1}{p^2} = p * (1-p) * \frac{2}{p^3} + \frac{1}{p} - \frac{1}{p^2}$$
$$= \frac{2(1-p)}{p^2} - \frac{1}{p^2} + \frac{1}{p} = \frac{1-p}{p^2} \Rightarrow \sigma = \sqrt{\frac{1-p}{p^2}}$$

Question #8:

In this question, it has been mentioned that variance is equal to 7 (minutes). However, for the original question, the standard deviation is equal to 7. We solve this question based on both of these assumptions:

a) for $\sigma = \sqrt{7}$:

95 percent of the time Negar can be sure with the confidence of 0.95 that the latest time that she arrives. Based on qnorm(0.95, 0, 1)=1.644854 is 44.35187 minutes

$$t < 40 + 1.64 * \sigma \Rightarrow t < 44.35187$$

or we can just use: $\text{qnorm}(0.95, 40, \text{sqrt}(7)) = 44.35187$

Now in order to have 95 percent confidence she can get there at 1 pm she has to leave at $13 - 44.35187 = 12{:}15$.

a) for $\sigma = 7$:

$$t < 40 + 1.64 * \sigma < 51.51398$$

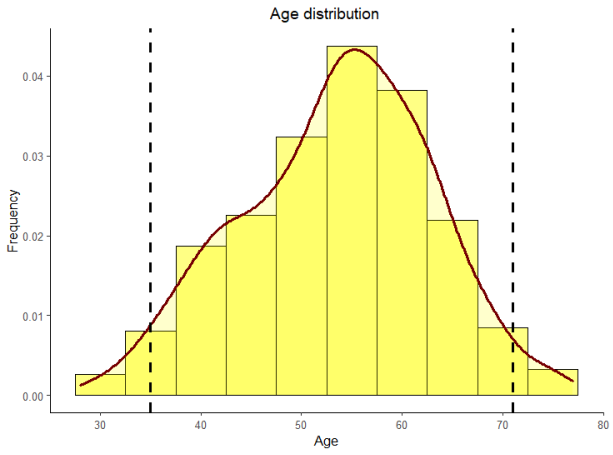
or we can just use: $\text{qnorm}(0.95, 40, 7) = 51.51398$

Now in order to have 95 percent confidence she can get there at 1 pm she has to leave at $13 - 51.51398 = 12{:}08$.

Question #9:

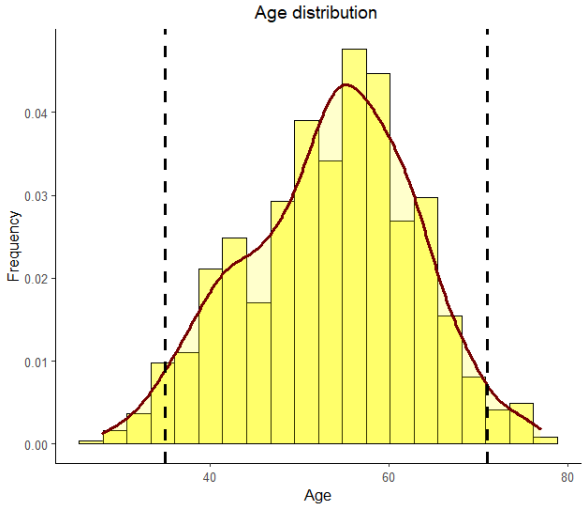
a) I choose 5 for the bin size. Because for this dataset that we are analyzing age effects on Heart conditions the 5-year bin size makes sense.

```
ggplot(df, aes(x=age)) + geom_histogram(aes(y=..density..),binwidth=5, color="black", fill="#ffff84") +
  geom_density(aes(y=stat(density)),alpha = .2, fill= "yellow", color = '#760002',size=1)+
  geom_vline(xintercept= quantile(df$age,c(0.975,0.025)),size=1,linetype = "dashed")+ labs(title ="Age distribution ")+
  xlab("Age") + ylab("Frequency")+theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 0.5),
  panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



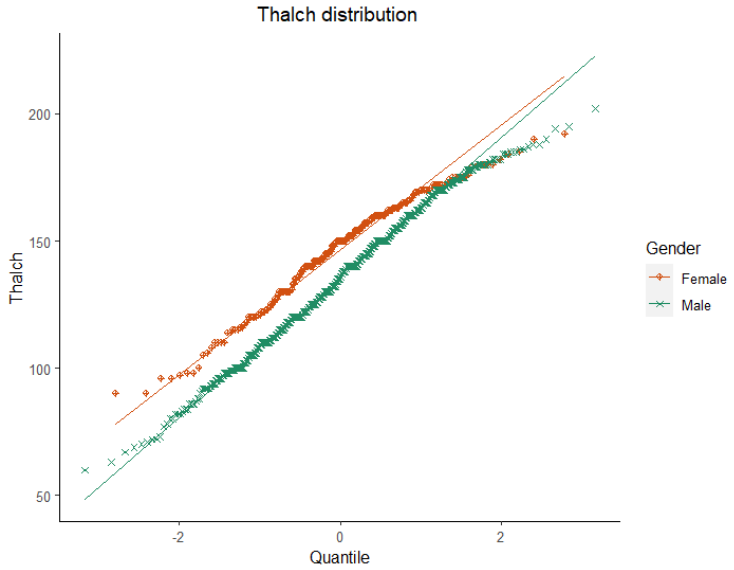
Also, we can use rules such as Freedman–Diaconis rule and calculate $\text{binwidth} = \frac{2*\text{IQR}}{n^3}$

And with this bandwidth the histogram would be as followed:



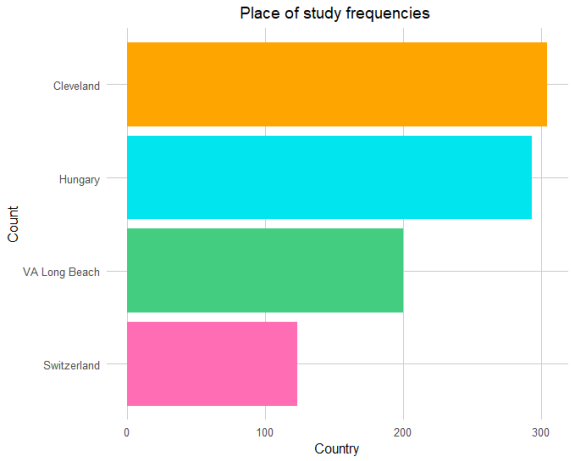
b) The points are close to straight lines, and because, there is a one-to-one relationship between the data and the theoretical quantiles, then the data follow a nearly normal distribution. The Q-Q plot is short tails. Because points follow an S-shaped curve.

```
qplot(sample = thalch,data=df, colour =sex,shape=sex)+
  stat_qq_line()+
  scale_shape_manual(values=c(10,4))+ labs(title ="Thalch distribution ",colour='Gender',shape='Gender')+
  xlab("Quantile") + ylab("Thalch")+  scale_color_manual(values=c("#d25010", "#208d64"))+
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 0.5),
        panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank(), axis.line = element_line(colour = "black"))
```



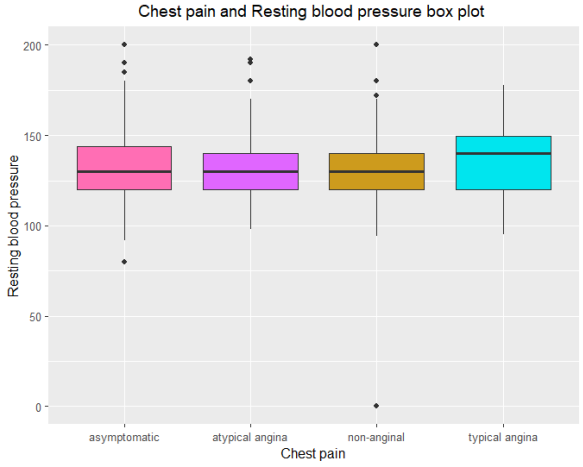
c)

```
ggplot(df, aes(x=reorder(origin,origin, function(x)+length(x)),fill=origin)) +
  geom_bar()+coord_flip()+ labs(title ="Place of study frequencies ")+
  xlab("Count") + ylab("Country")+scale_fill_manual(values = c("#ffa500","#00e5ee","#ff6eb4","#43cd80"), guide = "none") +
  theme(panel.grid.major = element_line(colour = "grey80"), panel.grid.minor = element_blank(),
        panel.background = element_rect(fill = NA),axis.ticks.x=element_blank(),axis.ticks.y=element_blank(),
        plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 0.5))
```



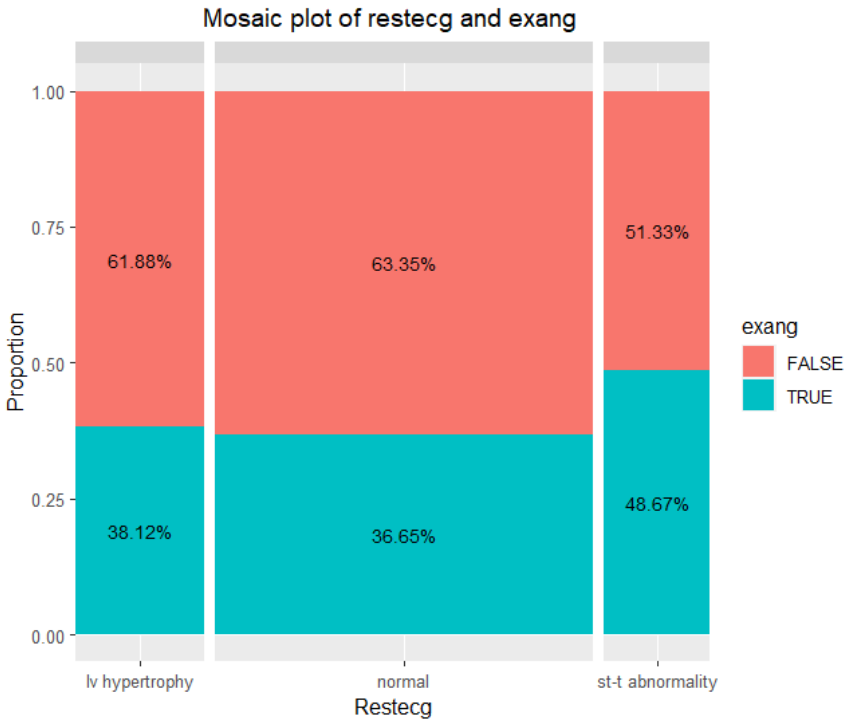
d)

```
ggplot(df, aes(x=trestbps, y=cp,fill=cp)) +
  geom_boxplot()+ coord_flip()+ labs(title ="Chest pain and Resting blood pressure box plot ")+
  xlab("Resting blood pressure") + ylab("Chest pain")+scale_fill_manual(values = c("#ff6eb4","#e066ff","#cd9b1d","#00e5ee")
                                                                        , guide = "none")+
  theme(plot.title = element_text(hjust = 0.5),plot.subtitle = element_text(hjust = 0.5))
```



e)

```
df2 <- with(df, df[!(restecg == "" | is.na(restecg)), ])
df2 <- with(df2, df2[!(exang == "" | is.na(exang)), ])
v <- aggregate(df2$exang~df2$restecg + exang, data = df2, FUN = length)
colnames(v) <- c('restecg','exang','counts')
v<-transform(v, rel1 = round(ave(counts, restecg, FUN = prop.table),digit=4))
v<-transform(v, grpsize = aggregate(v$counts, by=list(restecg =v$restecg ), FUN=sum))
colnames(v) <- c('restecg','exang','counts','rel1','test','grpSize')
names=c('lv hypertrophy'="",'normal'="",'st-t abnormality'="")
graphics.off()
ggplot(v, aes(x=restecg,y=rel1,fill=exang,width = grpSize)) +
  geom_bar(stat='identity') +
  scale_x_discrete(expand = c(0, 0)) +
  facet_grid(~restecg, scales = "free", space = "free",labeller = as_labeller(names))+
  geom_text(aes(label = paste(round(100*rel1, 2), "%", sep="")),size=3.34,position = position_stack(vjust = 0.5)) +
  labs(title ="Mosaic plot of restecg and exang ")+
  xlab("Restecg") + ylab("Proportion")+
  theme(plot.title = element_text(hjust = 0.5), plot.subtitle = element_text(hjust = 0.5),
        strip.text = element_text(size = 1))
```


Mosaic plot of restecg and exang
1.00
0.75
0.50
0.25
0.00
Proportion
61.88%
38.12%
63.35%
36.65%
51.33%
48.67%
lv hypertrophy
normal
st-t abnormality
Restecg
exang
FALSE
TRUE