Fatemeh Nadi
810101285
fatemehnadi@ut.ac.ir

Homework 2
Statistical Inference , Fall 1401

November 18, 2022

1

$P(A) = \frac{3}{36},\ P(B) = \frac{5}{36},\ P(C) = \frac{11}{36}$

$P(A \cap C) = \frac{1}{36},\ P(B \cap C) = \frac{2}{36}$

Events A and B are independent if:

$$P(A|B) = P(A) \tag{1}$$

- a
  $P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{\frac{1}{36}}{\frac{11}{36}} = \frac{1}{11},\quad P(A|C) \neq P(C) \xrightarrow{1} A, C$ are dependent.

- b
  $P(B|C) = \frac{P(B \cap C)}{P(C)} = \frac{\frac{2}{36}}{\frac{11}{36}} = \frac{2}{11},\quad P(B|C) \neq P(C) \xrightarrow{1} B, C$ are dependent.

2

$P(Women) = 55\%, P(Men) = 1 - P(Women) = 45\%$
$P(CS) = 8\%$
$P(Women \cap CS) = 3\%$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2}$$

- a
  $P(CS|Women) \stackrel{2}{=} \frac{P(Women \cap CS)}{P(Women)} = \frac{\frac{3}{100}}{\frac{55}{100}} = \frac{3}{55} \simeq 5.4\%$

- b
  $P(Women|CS) \stackrel{2}{=} \frac{P(CS \cap CS)}{P(Women)} = \frac{\frac{3}{100}}{\frac{8}{100}} = \frac{3}{8} \simeq 37.5\%$

3

- a
X is a discrete random variable and refers to the number of people who approve of George W. Bush's response to the World Trade Center terrorist attacks in September 2001.
The binomial distribution is used to describe the number of successes in a fixed number of trials, so in this case, people have two choices: approved (success) Bush's response after the incident or not(fail).
X is binomial.

- b
Suppose the probability of a single trial being a success is p. Then the probability of observing exactly k successes in n independent trials is given by:

$$P\left(X=k\right)=\binom{n}{k}p^{k}\left(1-p\right)^{n-k}$$

$n=400, k=358, p=0.92$

$P\left(X=358\right)=\binom{400}{358}0.92^{358}\left(0.08\right)^{42}=0.0136$

In fact $p\left(X\leq 358\right)$ must be calculated.
The binomial cumulative distribution function lets you obtain the probability of observing less than or equal to k successes in n trials, with the probability p of success on a single trial.
The binomial cumulative distribution function for a given value k and a given pair of parameters n and p is:

$$\begin{aligned} P\left(X\leq k\right) &= \sum_{i=0}^{k}\binom{n}{i}p^{i}\left(1-p\right)^{n-i} \\ &= \sum_{i=0}^{358}\binom{400}{i}0.92^{i}\left(0.08\right)^{400-i} \\ &= \boxed{0.044} \end{aligned}$$

Also can use this in R: $sum(dbinom(0:358, 400, 0.92))$

- c
The binomial distribution with probability of success p is nearly normal when the sample size n is sufficiently large that np and $n(1p)$ are both ***at least 10***.
The approximate normal distribution has parameters corresponding to the mean and standard deviation of the binomial distribution:

$$\mu=np, \quad \sigma=\sqrt{np(1-p))}$$

$\mu=400\times 0.92=\boxed{368}$
$\sigma=\sqrt{400\times 0.92\times 0.08}=\boxed{5.426}$

- d
We verify that both np and n(1 p) are at least 10 : $np=368$ and $n(1-p)=32$
we use the normal approximation in place of the binomial distribution using the mean and standard deviation from the binomial model: $N\left(\mu=365, \sigma=5.426\right)$

$$P\left(X\leq 358\right)=P\left(Z\leq\frac{358-368}{5.426}\right)=P\left(Z\leq -1.843\right)=\boxed{0.032}$$

Binomial: 0.044, Normal: 0.032

The normal approximation to the binomial distribution for intervals of values is usually improved if cutoff values are modified slightly. And in this case, because p is near one so width of bins in the binomial is vast, so this discrepancy causes the area between the binomial and normal distribution.

4

The total probability rule can be written in the following equation:

$$P(A) = \sum_{i=1}^{N} P(A \cap B_i) = \sum_{i=1}^{N} P(A|B_i)\, P(B_i)$$

We assume that:

$P(winning|type1) = 0.3,\ P(type1) = 0.5$
$P(winning|type2) = 0.4,\ P(type1) = 0.25$
$P(winning|type3) = 0.5,\ P(type3) = 0.25$

$$\begin{aligned} P(winning) &= P(winning|type1)\, P(type1) \ + \ P(winning|type2)\, P(type2) \ + \ P(winning|type3)\, P(type3) \\ &= 0.3 \times 0.5 + 0.4 \times 0.25 + 0.5 \times 0.25 \\ &= \boxed{0.375} \end{aligned}$$

5

- a
  A Poisson distribution models the number of events occurring in a fixed interval of time or space when the events are independent, and the average rate of the events is known.

  Here we have an SOP text which contains, on average, a specific and discrete number of words in a particular space, in this case, is a text.
  Furthermore, one word's existence does not affect another word's existence. Therefore, Poisson distribution is ideal to use in this case.

  A good candidate distribution would be the Poisson because the number of times that specific word appears is a discrete number that can only be 0 or take positive values.
  The assumption of constant probability and events' independence meets the Poisson distribution's characteristics. so X_i and Y_i are Poisson random variables.
  They have different parameters because the number of times a specific word appears in the first SOP differs from the second SOP.

- b
  The probability mass function for the Poisson distribution with rate parameter $\lambda > 0$ is:

  $$P(x|\lambda) = \frac{\lambda^x\ e^{-\lambda}}{x!}$$

  where x may take a value $0, 1, 2,$ and so on.

The number of "machine learning" occurring on the first SOP is a Poisson random variable with parameter λ, so the probability that "machine learning" didn't occur in the first SOP is:

$$P\left(Y = 0|\lambda\right) = e^{-\lambda}$$

The number of "machine learning" occurring on the second SOP is a Poisson random variable with parameter λ, so the probability that "machine learning" occurs in the first SOP is:

$$P\left(X \geq 1|\lambda\right) = 1 - P\left(X = 0|\lambda\right) = 1 - e^{-\lambda}$$

These two events are independent, so the total probability is:

$$\begin{aligned} P\left(\text{"}machinelearning\text{" } is\ used\ in\ the\ second\ SOP\ but\ not\ in\ the\ first\ SOP\ \right) &= P\left(Y = 0|\lambda\right) \cdot P\left(X \geq 1|\lambda\right) \\ &= \boxed{e^{-\lambda} \cdot \left(1 - e^{-\lambda}\right)} \end{aligned}$$

6

- a

  If the probability of a success in one trial is p and the probability of a failure is $1 - p$, then the probability of finding the first success in the n^{th} trial is given by:

  $$P\left(success\ in\ the\ n^{th}\ trial\right) = \left(1 - p\right)^{n-1} p$$

  but in this exercise X be an arrival process, treating rainy days as arrivals. so the iid inter-arrival times' $Geo\left(p\right)$ implies X is $Bernoulli\left(X\right)$.
  We know the probability that it rains on the 1th day of the month is independent of the past:

  $P\left(X_{1th} = 1\right) = \boxed{p}$

- b

  If A and B are independent, $P\left(A \cap B\right) = P\left(A\right) \cdot P\left(B\right)$
  Same as part a, we know the probability that it rains on the 5th and the 8th day of the month is independent.

  $P\left(X_{5th} = 1\ \cap X_{8th} = 1\right) = P\left(X_{5th} = 1\right) \cdot P\left(X_{8th} = 1\right) = \boxed{p^2}$

7

We recognize X as a geometric random variable, which X is the program works correctly without error, so we can calculate mean and variance with the geometric distribution:

$$E\left[X\right]=\boxed{\frac{1}{p}}$$

$$Var\left[X\right]=\boxed{\frac{1-p}{p^2}}$$

For proof of these formulas, we must calculate these equations, the mean and variance of X are given by:

$$E\left[X\right]=\sum_{k=1}^{\infty}k\left(1-p\right)^{k-1}p$$

$$Var\left[X\right]=\sum_{k=1}^{\infty}\left(k-E\left[X\right]\right)^2\left(1-p\right)^{k-1}p$$

8

I get X as the time Negar must take her cycle from home to school.

$$X\sim Normal\left(\mu=40,\sigma=7\right)$$

We have a random variable with a Normal Distribution. We should standardize it with Z score:

$$\begin{aligned}z&=\frac{x-\mu}{\sigma}\\&=\frac{x-40}{7}\end{aligned}$$

We want to know what time she should leave her house with 95% confidence interval,

$$P\left(X\leq x\right)=0.95$$

from normal table $P\left(Z\leq z=1.65\right)=0.95$

$$\frac{x-40}{7}=1.65\rightarrow x=51.55\simeq 52min$$

If Negar wants to have 95% confidence that she can attend her class at 1 p.m., the latest time she should leave her house is at least 12:08 p.m.

9 R

Please see this file: "Q9-R.Rmd" and "Q9-R.html"
Code and explanation are provided.

- a

```
heart <- data.frame(heart)

ggplot(heart, aes(x = age)) +
    geom_histogram(aes(y = ..density..), colour = "black", fill = "#F7D302",
                   bins = round((max(heart$age)-min(heart$age))/2)+1) +
    labs(
        title = "Histogram and Density of Age",
        caption = "Dashed lines show the 2.5% and the 97.5% percentiles on the diagram.",
        x = "Age",
        y = "Count"
    ) +
    geom_density(color = "#631919", size = 1) +
    geom_vline(aes(xintercept = quantile(age,.025)), linetype = "dashed", size = 1) +
    geom_vline(aes(xintercept = quantile(age,1-.025)), linetype = "dashed", size = 1) +
    theme_classic() +
    theme(
      plot.title = element_text(size = 12, face = "bold", hjust = 0.5),
      plot.caption = element_text(face = "italic")
    )
```

Figure 1: Q9-a code

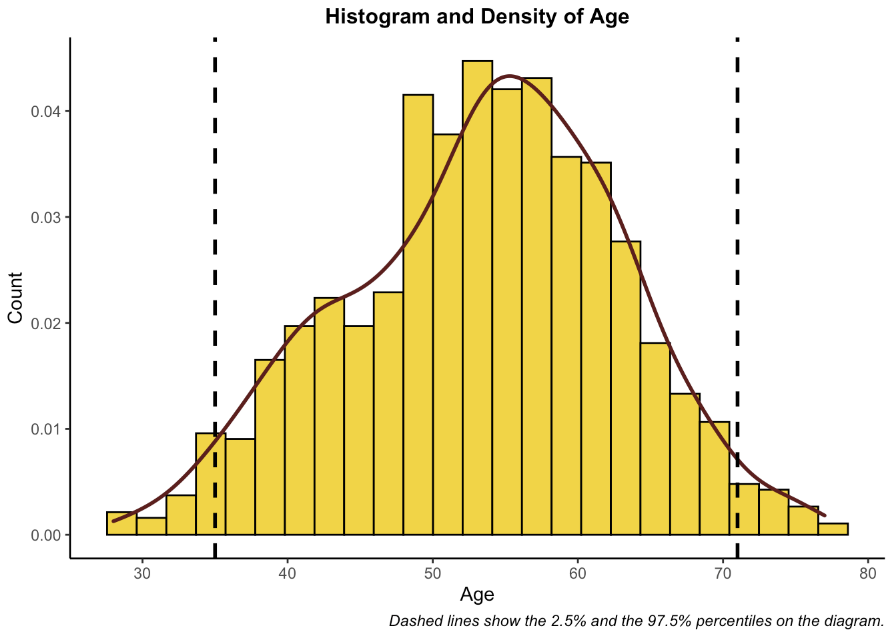

Figure 2: Q9-a age histogram and density line

- b

```
heart$sex <- as.factor(heart$sex) #set sex as categorical variable

qplot(sample = thalch, data = heart , color = sex ,shape = sex) +
  labs(title="QQ-plot of Maximum heart rate achieved for each gender") +
  theme(
    plot.title = element_text( size = 12, face = "bold", hjust = 0.5),  #change title font size, bold and center
    panel.grid.major = element_blank(),         #remove grid
    panel.grid.minor = element_blank(),         #remove grid
    axis.line = element_line(colour = "black"), #set border lines
    legend.title = element_text(size=12),       #change legend title font size
    legend.text = element_text(size=10)         #change legend text font size
    ) +
  guides(color = guide_legend(title = "gender")) +
  guides(shape = guide_legend(title = "gender")) +
  scale_shape_manual(values = c(20, 4)) +       #set shape of each groups
  stat_qq() +   #fit line
  stat_qq_line()
```

Figure 3: Q9-b code

Figure 4: Q9-b thalch qqplot

- c

```
# sorting origin
heart_sorted =  as.data.frame(sort(table(heart$origin)))

ggplot(heart_sorted) +
  geom_col(aes(Var1,Freq, fill = Var1)) +
  coord_flip() +
  xlab('Origin') +
  ylab('Count') +
  ggtitle("Horizontal Bar Plot of Origin") +
  theme_bw() +
  theme(legend.position = "none",
        plot.title = element_text( face = "bold", hjust = 0.5))
```

Figure 5: Q9-c code

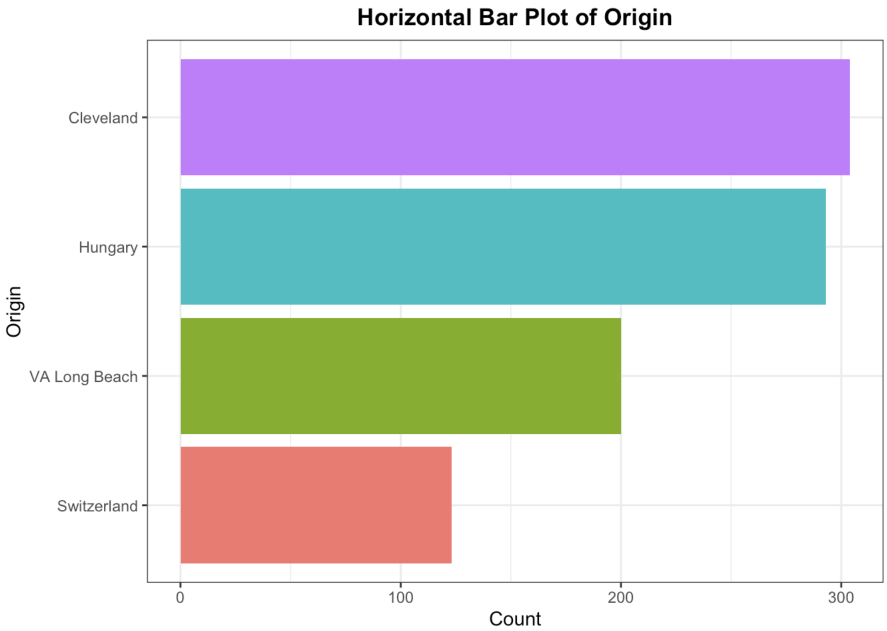

Figure 6: Q9-c Origin bar plot

- d

```
ggplot(heart, aes(x=cp, y=trestbps, fill=cp)) +
  geom_boxplot() +
  xlab('chest pain type') +
  ylab('resting blood pressure') +
  ggtitle("Boxplots of resting blood pressure") +
  theme_gray() +
  theme(legend.position = "none",
          plot.title = element_text( face = "bold", hjust = 0.5))
```

Figure 7: Q9-d code

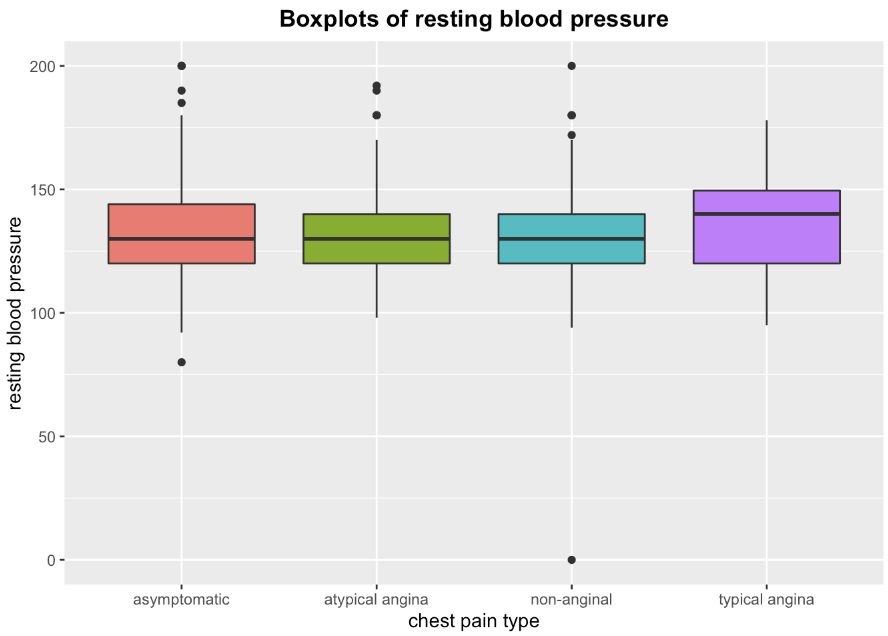

Figure 8: Q9-d rbp vs. cp boxplots

- e

```
#preprocess
heart <- heart[!(is.na(heart$exang) | heart$exang=="") ,]
heart <- heart[!(is.na(heart$restecg) | heart$restecg=="") ,]

#primary mosaic plot
p <- ggplot(data = heart) +
  geom_mosaic(aes(x = product(restecg), fill = exang))
```

Figure 9: Q9-e code

```
# heart_sumerice contains percentages of each restecg
heart_sumerice <- heart %>%
  count(exang,restecg) %>%
  group_by(restecg) %>%
  mutate(percentage = prop.table(n));

#
p_label <- ggplot_build(p)$data %>% as.data.frame() %>% filter(.wt > 0)
```

Figure 10: Q9-e code

```
p_label$percentage = heart_sumerice$percentage
```

```
p +
  geom_text(data = p_label,
            aes(x = (xmin + xmax)/2,
                y = (ymin + ymax)/2,
                label = scales::percent((percentage))),
            position = "identity",
            check_overlap = TRUE) +
  ggtitle("Mosaic plot of \nResting electrocardiographic results & Exercise-induced angina") +
  xlab("resting electrocardiographic results") +
  ylab("Proportion") +
  theme_grey() +
  theme(plot.title = element_text(face = "bold", hjust = 0.5))
```

Figure 11: Q9-e code

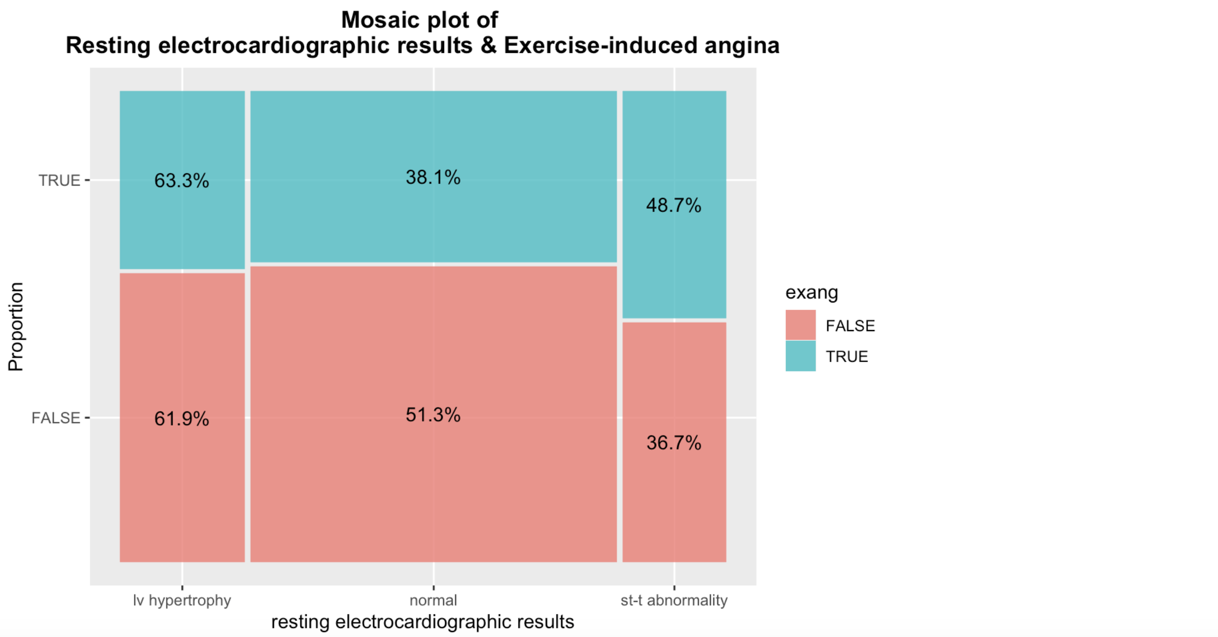

Figure 12: Q9-e mosaic plot of restecg and exang