QSCI 381 Homework #4 - Vinsensius

The PDF file of the code is attached at the end of the HW

Question 1

In this short question we will work a little bit more with using the binomial distribution using an example covered in lectures. Globally, 8% of people have blue eyes, and we are going to evaluate a range of scenarios related to this.

1a

> The US women's soccer squad has 23 players, 7 of which have blue eyes. Using the binomial distribution (*dbinom*), calculate the probability of observing exactly this many blue-eyed individuals out of 23 if the probability was 8% (2 pts)

Using *dbinom* function, we get 0.00135 as probability of getting 7 people having blue eyes.

1b

> If the US women's soccer squad was representative of the global population (i.e. 8% chance of having blue eyes), what would be the most likely number of players with blue eyes that you would expect to observe (4 pts) [Hint: try using *dbinom* for a range of different scenarios informed by the expected value]

By varying the values for n for *dbinom*, we got 1 person is the most likely observe in team since it has the highest probability.

1c

> Using the cumulative binomial distribution function (*qbinom*), calculate the upper limit of the range of players with blue eyes (still out of 23) that you would expect to observe with 0.95 probability if the true proportion was 8% (3 pts)

The number that we would expect is 4.

1d

Using your answers (1a-1c) comment on how anomalous the observed count of 7 blue-eyed individuals out of 23 is, and possible reasons for this discrepancy (2 pts)

From part (c), we know that getting 4 people of 23 having blue eye is 0.95. From part (b), we expect 1 people of 23 people having blue eyes if the soccer team is the population. So, having 7 people in the soccer team is anomalous since the probability is very small based on (a). The reason for the discrepancy is because that the sample size is not representative of the global population.

Question 2

In this example we will be using the ShipAccidents dataset available from canvas:

ship $<-$ read.csv("shipaccidents.csv")

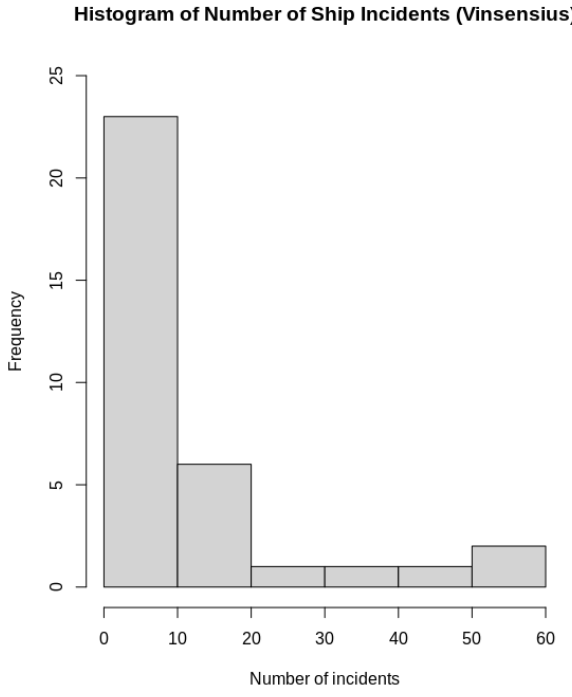
head(ship)

This data describes the number of accidents recorded for a range of ship types, along with their service period. We will mostly consider the following variables:

- incidents: the total number of accidents recorded for each ship type
- service: the collective number of months that ships of that type were active for

2a

Plot a histogram of the number of incidents recorded in this dataset. Label the x and y-axes accordingly and define the plot title to be your name. Paste your plot below, and provide a written description of the main features you would interpret from this data (5 pts)



The main features that we would like to see in the histogram is how data is being distributed. The histogram is right-skewed because there are high values on the left side of the graph while there is tail extending to the right.

2b

Using the number of incidents and the time that the ship type was active for, calculate the rate (number per month) of incident occurrence per ship type (i.e. create a new column called rate), and from this, report the mean and range of incident rates across the different ship types (4 pts)

Mean	0.00309
Range	0.0160

2c

Using the mean incident rate from (2b), if an individual ship was in service for 50 years (600 months), what would be the expected value for the number of incidents (2 pts)

Assuming the data has poisson distribution, expected value is 600*rate = 1.85

2d

Using the rate of incident occurrence for a ship of type E3, what would be the probability of observing

(a) 1 incident in the first ten years (120 months) of service (3 pts)

(b) At least 2 incidents in the first ten years (120 months) of service (4 pts)

(a) We will assume poisson distribution. The rates of incident will be the parameter λ. To calculate the probability of the first ten years, we will need to multiply the rate by 120 to change the unit from per month to per 10 years. Then we will use *dpois* function to calculate the probability with parameter n=1 and the new rate. Hence, the P(1 incident in the first 10 years) = 0.281

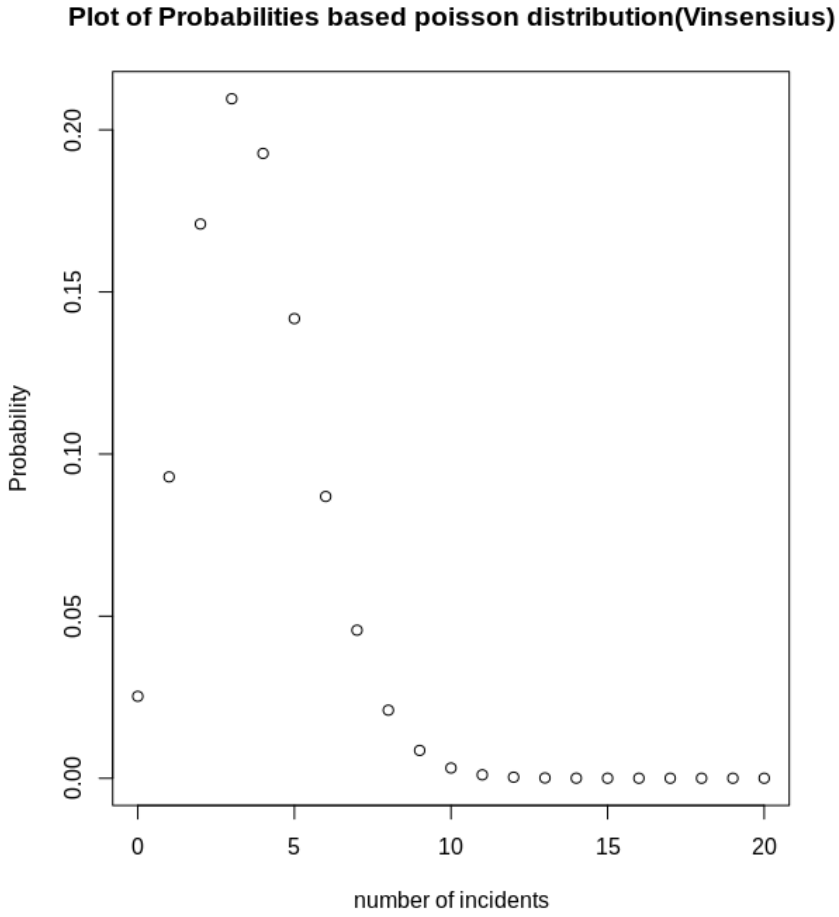
(b) We will use the assumption in previous part. To find P(x $\geq$ 2),

$$P(x \geq 2) = 1 - P(x < 2) = 0.573$$

2e

Assuming that incident counts are distributed according to a poisson distribution, calculate the most likely (i.e. count with highest probability) incident count for a vessel of type C5 that had been in service for 40 years (480 months), reporting the count AND it's probability of occurrence (4 pts) [Hint: use *dpois* for a range of counts]

We can see the result from the plot



So, the count = 3 and the P(x=3) = 0.210

Question 3

In this question we will be using the AudioVisual dataset available from canvas.

Subjects in a reaction time study were asked to press a button as fast as possible after being exposed to either an auditory stimulus (a burst of white noise) or a visual stimulus (a circle flashing on a computer screen). Average reaction times (ms) were recorded for between 10 and 20 trials for each type of stimulus for each subject.

We will be using this dataset to look at reaction time differences between auditory or visual stimuli

3a

Calculate the mean and standard deviation of response times for auditory and visual stimuli. Comment on which stimulus results in the fastest response times, and which results in the most consistent response times. Round your answers to 1 decimal place (6 pts)

	Auditory	Visual
Mean	215.9	288.8
Standard Dev.	86.9	60.4

Visual data is more consistent due to lower standard deviation. Audio data has the fastest response.

3b

In athletics, if an Athlete responds in less than 100 ms to the starting gun it is deemed a false start and they are disqualified. Using your answer in (2a) and assuming the data follow a normal distribution, what proportion of responses would be deemed a false start. You can assume that they respond to an auditory stimulus. Round your answer to 3 decimal places. (3 pts)

$P(x<100) = 0.0912$

3c

At the 2004 Olympics, the fastest reaction time during the womens 100 m hurdles final was 112 ms. Assuming reaction times to auditory stimuli are normally distributed, what proportion of people respond quicker than 112 ms, but would not be excluded due to false starts (< 100 ms). Round your answer to 3 decimal places. (4 pts)

P(100 < x < 112) = P(x=112) - P(x=100) = 0.024

3d

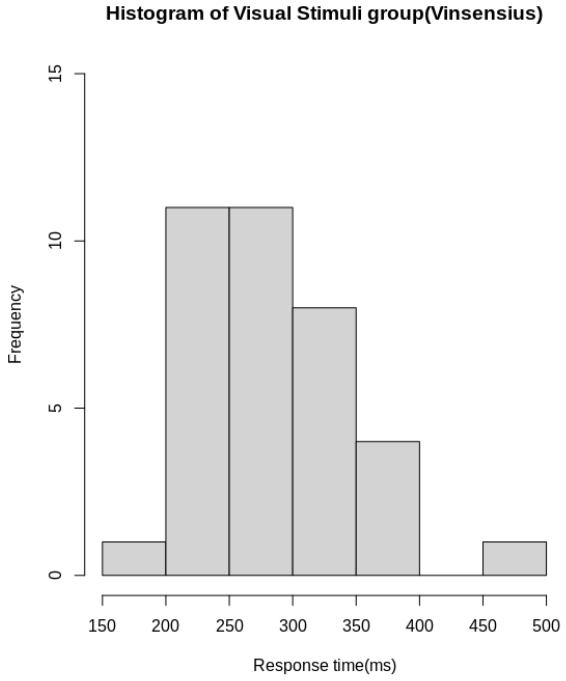
Assuming that response times are normally distributed, for each stimulus type identify the expected fastest 10% and slowest 10% (i.e. 90th percentile) of response times using the means and standard deviations reported in (3a) (6 pts)

	Auditory	visual
10% percentile response time(ms)	104	211
90% percentile response time(ms)	327	366

3e

Compare the fastest 10% of reaction times given a visual stimuli calculated in 3d (i.e. assuming a perfect normal distribution) to the corresponding empirical (i.e. observed) value in the dataset, and comment on any difference that you observe (3 pts)

The fastest 10% for visual stimuli is 235 ms. The difference is because the data for visual stimuli is not exactly normal distribution so there will be a difference. We can see the histogram of the visual stimuli data below.



We can see that based on the histogram that the data is somewhat close to normal if the response time between 450-500 is removed.

```
## Number 1

ans_1a = dbinom(7,23,0.08)
x = c(0:23)
prob_1b = dbinom(x,23,0.08)
ans_1b = x[match(max(prob_1b),prob_1b)]
ans_1c = qbinom(0.95,23,0.08)

## Number 2

ship_dat <- read.csv("shipaccidents.csv")

hist(ship_dat$incidents, main = "Histogram of Number of Ship Incidents (Vinsensius)"
     ,ylim = c(0,25),xlab = "Number of incidents")

ship_dat$rate = ship_dat$incidents/ship_dat$service

summary(ship_dat$rate)

ans_2c = mean(ship_dat$rate)*600

rate_e3 = ship_dat$rate[ship_dat$type == 'E3']

ans_2d_1 = dpois(1,rate_e3*120)

ans_2d_2 = 1-ppois(1,rate_e3*120)

rate_c5 = ship_dat$rate[ship_dat$type == 'C5']
prob_c5 = dpois(0:20,rate_c5*480)
plot(0:20,prob_c5,main="Plot of Probabilities based poisson distribution(Vinsensius)",
     ylab="Probability",xlab="number of incidents of E5")
ans_2e  = c(3,max(prob_c5))

## Number 3

av_dat <- read.csv("AudioVisual.csv")
av_dat_audio = av_dat[av_dat$Stimulus == 'auditory',]
av_dat_visual = av_dat[av_dat$Stimulus == 'visual',]
summary(av_dat_audio$ResponseTime)
summary(av_dat_visual$ResponseTime)
mean_audio = mean(av_dat_audio$ResponseTime)
mean_viz = mean(av_dat_visual$ResponseTime)
sd_audio = sd(av_dat_audio$ResponseTime)
sd_visual = sd(av_dat_visual$ResponseTime)
ans_3b = pnorm(100,mean= mean(av_dat_audio$ResponseTime), sd_audio)
z_score_100 = (100- mean(av_dat_audio$ResponseTime))/sd_audio
z_score_112 = (112- mean(av_dat_audio$ResponseTime))/sd_audio
ans_3c = pnorm(z_score_112) - pnorm(z_score_100)

ans_3d_1 = qnorm(0.1,mean_audio,sd_audio)
ans_3d_11 = qnorm(0.9,mean_audio,sd_audio)
ans_3d_2 = qnorm(0.1,mean_viz,sd_visual)
ans_3d_21 = qnorm(0.9,mean_viz,sd_visual)

decile1 <- quantile(av_dat_visual$ResponseTime, probs = seq(.1, .9, by = .1))
hist(av_dat_visual$ResponseTime, main = "Histogram of Visual Stimuli group(Vinsensius)"
     ,ylim = c(0,15),xlab = "Response time(ms)")
```