QSCI 381 - AUT 22
Homework #5 - Vinsensius

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

Question 1

In this question we will be using the faithful dataset

data("faithful")
head(faithful)
?faithful

The data represents information on the waiting time (waiting: minutes) and eruption duration (eruptions: minutes) of eruptions from the Old Faithful Geyser in Yellowstone National Park. There are 272 observations, and we will be using this dataset to illustrate sampling distributions and the Central Limit Theorem.

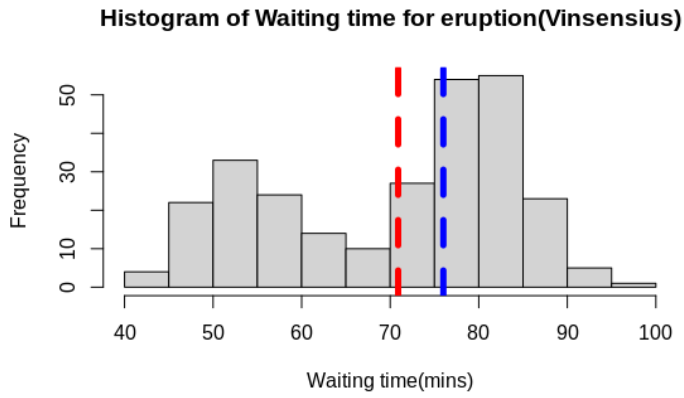
1a

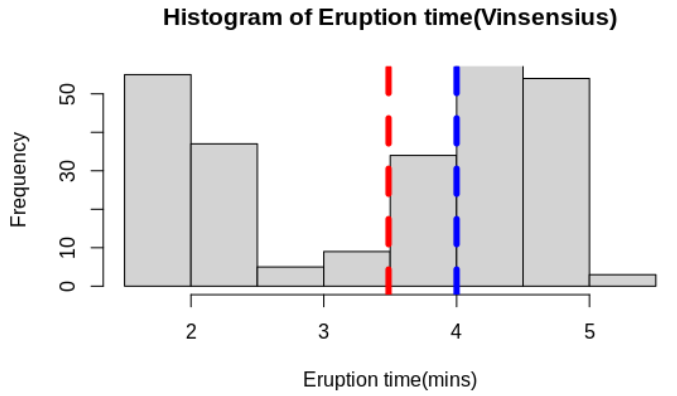
Calculate the mean and standard deviation of each of the two variables (waiting , eruptions), treating these data as the whole population. Copy and paste your code for these calculations below, as well as the numeric answers, rounded to 2 decimal places (6 pts)

	Eruption Time (mins)	Waiting time (mins)
Mean	3.49	70.90
Standard Dev	1.14	13.57

1b

Create histograms of each of the two variables (waiting , eruptions), remembering to label the axes appropriately, titling each histogram after the measurement shown. On each histogram add the mean (red) and median (blue) as vertical lines using abline (8 pts)





1c

For each variable comment on the distribution shape and characteristics, and from that, under what conditions would you be able to apply the Central Limit Theorem to samples drawn from these data. (4 pts)

Both histogram plots have two peaks on them, so the distribution is mostly be bi-modal. Both histogram have median larger than mean.

The condition to apply the central limit theorem on the data if $n \geq 30$.

1d

Use the sample function to create a sample from the waiting column with a sample size, n = 45. From this sample, calculate the sample mean and standard deviation, including your code, and the statistic values calculated below. (4 pts)

Sample Mean	74.76 mins
Sample Standard Dev	10.52 mins

1e

Given the sample size in (1d) and the population parameters you calculated in (1a), what does the Central Limit Theorem tell us about the sampling distribution of the sample mean statistic of waiting time (waiting)? In your answer provide a written description, as well as numeric values of the sampling distribution parameters, rounded to 3 decimal places (6 pts)

Our sample size The sampling distribution is roughly normal based on central limit theorem since $n > 30$. The sampling distribution of mean is the population mean, which is 70.90 mins. The standard deviation of the sampling distribution would be

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = 2.02 \text{ mins}$$

1f

Given the sampling distribution properties reported above, comment on how unusual your sample mean statistic was, using relevant calculations to support your answer. Include your code. (3 pts)

$$z_{\bar{x}} = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{74.76 - 70.90}{2.02} = 1.95$$

Based on the z-score of the sample mean, the sample mean statistic is not that unusual because it is still within the 95% confidence range, i.e $\pm$ 2 from the mean of sampling distribution.

1g

Given the sampling distribution properties reported in (1e) calculate:

a the probability of obtaining a sample mean within ± 0.5 of the sample mean you obtained in (1d). Include your code. (3 pts)

b the range of sample means (with the mid-range set to the sample mean you obtained in (1d)) you would expect to obtain 60% of the time with a sample size n = 45. Include your code.(4 pts)

1g(a)

$$P(\bar{x} - 0.5 < x < \bar{x} + 0.5) = P(x = \bar{x} + 0.5) - P(x = \bar{x} - 0.5) = 0.0328$$

1g(b)

Based on the sampling distribution 1(e), the 60% range is between $69.19 < p < 72.6$. We will shift the range such that the mid range is the sample mean from 1(d), $73.05 < \hat{p} < 76.46$.

Q2

In this question we will be using the Hearing dataset from canvas. Go to canvas and download this file.

```
hearing <- read.csv(file="Hearing_data.csv")
head(hearing)
```

Audiologists use standard lists of words to test hearing; the words are calibrated to make all words on the list equally hard to hear. In this study 24 subjects were asked to listen to 200 such words when played at low volume with a noisy background, and to identify what word was played. The Correct column is the number of correct responses across the 200 words played to them.

2a

Calculate the mean probability (i.e. across individuals) of correctly identifying the word that was played. Round your answer to 3 decimal places (2 pts)

The mean of the data is 56.6 words. Therefore the mean probability is $\frac{56.625}{200} = 0.283$.

2b

> Evaluate the conditions for applying the Normal approximation to the Binomial, and comment on each criteria and whether they are met (3 pts)

The conditions for applying normal approximation are when $np \geq 10$ and $n(1-p) \geq 10$. The value for $np = 56.6 > 10$ and $n(1-p) = 143.4 > 10$. Thus, the conditions are met and the binomial distribution can be approximated as normal.

2c

> Given your answers to (2a) and (2b), calculate the expected mean and standard deviation of the Normal approximation for this data. Round your answers to 2 decimal places. (3 pts)

Expected mean ($n\hat{p}$) is 56.63 and expected standard deviation $\left(\sqrt{n\hat{p}(1-\hat{p})}\right)$ is 6.37.

2d

> Given the values you calculated in (2c), apply the normal approximation to calculate the expected proportion of individuals that would get a score between 40 and 60 (inclusive), remembering to apply continuity corrections. Round your answer to 3 decimal places (4 pts)

Applying the continuity correction, we will be looking at the probability between 39.5 and 60.5.

$$\text{P}(39.5<\text{x}<60.5) = \text{P}(\text{x}=60.5) - \text{P}(\text{x}=39.5) = 0.725$$

Q3

In this question we will be using the National Parks Images dataset available on canvas

nat.parks <- read.csv(file="nationalparks_images.csv")

head(nat.parks)

A researcher wanted to find out if people are more likely to see and photograph wildlife (bears, marmots, elk, etc) inside vs outside National Parks in Washington State. She collected images from Instagram taken throughout the Cascade mountain range in Washington, and identified whether images were captured inside a National Park or in National Forest areas, which have different legal protections (e.g. dogs and timber harvest are not permitted in National Parks), as well as whether images contained wildlife.

3a

For the National Park and non-National Park areas, calculate and report the

- Sample size
- Number of positive detections of wildlife
- Probability wildlife has shown in an image

(6 pts)

	National Park	Non National Park
Sample Size	266	248
Number wildlife detected	29	21
$\hat{p}$=P(wildlife shown in image)	0.109	0.0847

3b

Given the answers in (3a) evaluate whether these data meet the criteria to apply the central limit theorem to the sample proportion statistics (5 pts)

The criteria is the sample size ≥ 30, $np \geq 10$ and $n(1-p) \geq 10$.

Park type / Criteria	National Park	Non National Park
Sample Size ≥ 30	satisfied (n= 266)	satisfied (n=248)
$np \geq 10$	satisfied ($np = 29$)	satisfied (np=21)
$n(1-p) \geq 10$	satisfied ($n(1-p) = 237$)	satisfied ($n(1-p) = 227$)

Therefore, these data meet the criteria to apply central limit theorem to the sample proportion statistics.

3c

Using your answer in (3a), calculate a 90% confidence interval for the proportion of times images contained wildlife for each of the National Parks, and National Forest areas (non-National Park areas). In doing so, report the

- Margin of error
- Confidence interval bounds

Show your working and round your answers to 3 decimal places (12 pts)

	National Park	Non National Park
E	0.031	0.029
Confidence interval bounds	$0.078 < p < 0.140$	$0.056 < p < 0.114$

3d

Compare the estimates and confidence intervals you obtained for the proportion of images containing wildlife in National Parks compared to National Forests, and describe any differences and whether you think those differences are indicative of a genuine difference in Wildlife photography between the two locations (3 pts)

The confidence interval of National forest is wider than National park because it has smaller sample size. The confidence interval of both data overlap each other with overlap p , $0.078 < p < 0.114$. This means that there is certain p in the data set such that a wildlife photograph can not be distinguish, whether it is taken in National Park or National Forest.

```
sd_population <- function(v) {
  sum_difference = var(v)
    #sum((v-mean(v))^2)
  N = length(v)
  sig = sqrt(sum_difference*(N-1)/N)
  return(sig)
}
### 1

geyser_dat = datasets::faithful

mean_wait = mean(geyser_dat$waiting)
mean_erup = mean(geyser_dat$eruptions)

wait_dat = geyser_dat$waiting
sd_wait = sd_population( geyser_dat$waiting)
sd_erup = sd_population(geyser_dat$eruptions)

hist(geyser_dat$eruptions,main = "Histogram of Eruption time(Vinsensius)"
     ,ylim = c(0,55),xlab = "Eruption time(mins)")
abline(v = mean(geyser_dat$eruptions),                # Add vertical line
       col = "red",            # Modify color
       lty = "dashed",         # Modify line type
       lwd = 5)                # Modify thickness
abline(v = median(geyser_dat$eruptions),                # Add vertical line
       col = "blue",            # Modify color
       lty = "dashed",         # Modify line type
       lwd = 5)                # Modify thickness
hist(geyser_dat$waiting,main = "Histogram of Waiting time for eruption(Vinsensius)"
     ,ylim = c(0,55),xlab = "Waiting time(mins)")
abline(v = mean(geyser_dat$waiting),                # Add vertical line
       col = "red",            # Modify color
       lty = "dashed",         # Modify line type
       lwd = 5)                # Modify thickness
abline(v = median(geyser_dat$waiting),                # Add vertical line
       col = "blue",            # Modify color
       lty = "dashed",         # Modify line type
       lwd = 5)                # Modify thickness
set.seed(321)
sample_wait = sample(geyser_dat$waiting,45)
mean_sample_wait = mean(sample_wait)
sd_sample_wait = sd(sample_wait)
z_score_sample_mean = (mean_sample_wait-mean_wait)/sd_sample_wait
SE_error = sd_wait/sqrt(45)
ans_1f = (mean_sample_wait - mean_wait)/SE_error
#1g
ans_1g_1 = pnorm(mean_sample_wait+0.5,
                 mean = mean_wait,sd= sd_wait/sqrt(45)) -
  pnorm(mean_sample_wait-0.5,
        mean = mean_wait,sd= sd_wait/sqrt(45))

##
# for 1g part 2,use upper limit qnorm(0.8,sample distribution statistics)
# lower limit qnorm(0.2,sample distribution stats)
upper_limit_1g_2=qnorm(0.8,mean_wait,SE_error)
lower_limit_1g_2=qnorm(0.2,mean_wait,SE_error)
#rm(list = ls(all.names = TRUE)) # remove all variables
### 2
#rm(list = ls(all.names = TRUE))
hearing <- read.csv(file="Hearing_data.csv")
mean_prob_correct = mean(hearing$Correct)/200

sd_expected = sqrt(mean_prob_correct*200*(1-mean_prob_correct))
mean_expected = mean_prob_correct*200
P_40_60 = pnorm(60.5,mean_expected,sd_expected) - pnorm(39.5,mean_expected,sd_expected)

### 3
nat_parks <- read.csv(file="nationalparks_images.csv")
nat_parks_yes = nat_parks[nat_parks$NatPark == 'Y',]
```

```
nat_parks_no = nat_parks[nat_parks$NatPark != 'Y',]
nat_parks_yes_wildlife = sum(nat_parks_yes$Wildlife > 0 )
nat_parks_no_wildlife = sum(nat_parks_no$Wildlife > 0 )

p_hat_nat_park_wildlife = nat_parks_yes_wildlife/length(nat_parks_yes$Wildlife)
p_hat_non_park_wildlife = nat_parks_no_wildlife/length(nat_parks_no$Wildlife)

expected_mean_park = p_hat_nat_park_wildlife*NROW(nat_parks_yes)
expected_mean_non_park = p_hat_non_park_wildlife*NROW(nat_parks_no)
z_star = abs(qnorm(0.05))
SE_park = sqrt(p_hat_nat_park_wildlife*(1- p_hat_nat_park_wildlife)/NROW(nat_parks_yes))
E_park = SE_park*z_star

SE_non_park = sqrt(p_hat_non_park_wildlife*(1- p_hat_non_park_wildlife)/NROW(nat_parks_no))
E_non_park = SE_non_park*z_star
###
```