Problem Set 2

Applications of probability theory

Michael Ghattas

Notes

Other students who I worked with on this assignment (if any): None.

```
library(knitr)
knitr::opts_chunk$set(echo = TRUE)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
library(HistData)
# install.packages('Lock5Data')
library(Lock5Data)
library(foreign)
```

```
## Warning: package 'foreign' was built under R version 4.1.2
```

Introduction

These questions were rendered in R markdown through RStudio (https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf, http://rmarkdown.rstudio.com).

Please generate your solutions in R markdown and upload a knitted pdf document to Gradescope. Please put your name in the "author" section in the header.

The questions in this problem set use material from the slides on discrete and continuous probability spaces and the `Rmds` `Discrete_Probability_Distributions_2_3_3.Rmd` and `02_continuous_probability_distributions_in_r.`

Load Data

```
data('PolioTrials')
dat<-PolioTrials
dat
```

```
##          Experiment                 Group Population Paralytic NonParalytic
## 1 RandomizedControl            Vaccinated     200745        33           24
## 2 RandomizedControl               Placebo     201229       115           27
```

```
## 3 RandomizedControl          NotInoculated     338778       121           36
## 4 RandomizedControl IncompleteVaccinations       8484         1            1
## 5   ObservedControl             Vaccinated     221998        38           18
## 6   ObservedControl               Controls     725173       330           61
## 7   ObservedControl    Grade2NotInoculated     123605        43           11
## 8   ObservedControl IncompleteVaccinations       9904         4            0
##   FalseReports
## 1           25
## 2           20
## 3           25
## 4            0
## 5           20
## 6           48
## 7           12
## 8            0
```

Question 1

Please carry out the analysis below and answer the questions that follow. For this assignment, please do all calculations in R and show the code and the results in the knit document.

Context

Question 2 on problem set 1 addresses the question of whether the `NotInoculated` and `Placebo` groups in the `Randomized Control` experiment had statistically significantly different rates of paralytic polio.

Recall that the `NotInoculated` and `Placebo` groups differ in that the children in the `Placebo` group had been enrolled in the vaccine trial while the parents of the children in the `NotInoculated` group did not enroll their children.

The approach, using the `rbinom` function, implemented the idea that populations in the `NotInoculated` and `Placebo` groups in the `RandomizedControl` experiment were the same in regards to paralytic polio cases by using the `rbinom` function to assign paralytic polio cases in the combined `NotInoculated` and `Placebo` groups of the `RandomizedControl` experiment to the `Placebo` group with probability equal to the ratio of the size of the `Placebo` group to the size of pooled `Placebo` group and `NotInoculated` group.

Note that the function `rbinom(x,size,prob)` simulates drawing x random samples from `Binom(size,prob)`.

The computations for that analysis are reproduced here:

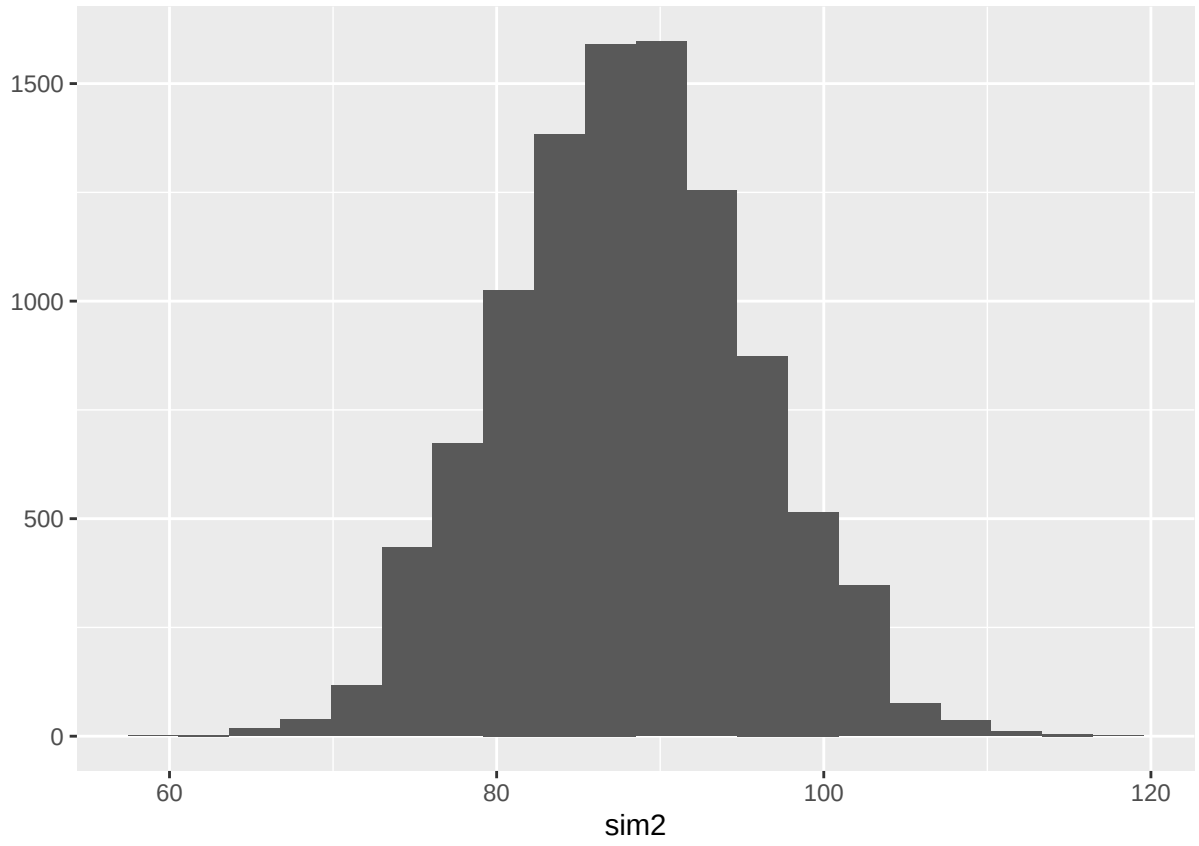
```
n<-10000 # number of simulations

# Calculate the number of paralytic polio cases in the pooled "Placebo" and "NotInoculated" group.
ct<-sum(dat$Paralytic[2:3])

# Calculate the proportion "prop" of the the pooled "Placebo" and "NotInoculated" group that are in the
prop<-dat$Population[2]/sum(dat$Population[2:3])

# Generate 10,000 counts of paralytic polio cases in the "Placebo" group under the model that each para
set.seed(45678765)
sim2<-rbinom(n,ct,prop)
qplot(sim2,bins=20)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```




```
# proportion of the simulated counts of paralytic polio in the "Placebo" that are less than or equal to
mean(sim2<=dat$Paralytic[2])
```

```
## [1] 0.9997
```

```
# proportion of the simulated counts of paralytic polio in the "Placebo" that are greater than or equal
mean(sim2>=dat$Paralytic[2])
```

```
## [1] 4e-04
```

Question 1.1

Using the same null model described above, please calculate the probability that the count of paralytic polio cases in the `Placebo` group under the null model is less than or equal to `dat$Paralytic[2]` directly rather than by simulating it. Recall that the function `pbinom(x,size,prob)` returns the probability of the event that the number of successes is in the set $\{0, 1, ...x\}$.

Your answer here:

```r
## Please be sure that your computed probability shows in your knitted solutions
# Number of paralytic polio cases in Placebo group
observed_placebo <- dat$Paralytic[2]

# Total number of paralytic polio cases in pooled "Placebo" and "NotInoculated" group
ct <- sum(dat$Paralytic[2:3])

# Proportion of the pooled group that are in the "Placebo" group
prop <- (dat$Population[2] / sum(dat$Population[2:3]))

# Probability calculation using pbinom
p_value <- pbinom(observed_placebo, size = ct, prob = prop)
p_value
```

```
## [1] 0.9998718
```

Question 1.2

Using the same null model described above, please calculate the probability that the count of paralytic polio cases in the `Placebo` group under the null model is greater than or equal to `dat$Paralytic[2]` directly rather than by simulating it. Hint: Denote the value in part 1 by p. This answer is not $1 - p$. The value $1 - p$ is the probability of the event that count of paralytic polio cases in the `Placebo` group under the null model is strictly greater than `dat$Paralytic[2]`.

Your answer here:

```r
## Please be sure that your computed probability shows in your knitted solutions
# Probability calculation for less than or equal using pbinom
p_value_less_equal <- (1 - pbinom(observed_placebo - 1, size = ct, prob = prop))
p_value_less_equal
```

```
## [1] 0.0002119741
```

Question 1.3

Is the value computed in part 2 strong evidence against the null model?

Your answer here:

(Yes, the value computed in part 2 provides strong evidence against the null model. The p-value is less than 0.05, indicating that the observed number of paralytic polio cases in the Placebo group is statistically significant and unlikely to have occurred by chance. This suggests a significant difference in paralytic polio rates between the Placebo and `NotInoculated` groups.)

Question 2

Context

This question concerns the uniform distribution on $[0, 1]$, the continuous probability space $(\mathcal{S}, \mathcal{M}, \mathcal{P})$ with $\mathcal{S} = [0, 1]$ and $\mathcal{P}$ defined by $\mathcal{P}(A) = \int_{A \cap [0,1]} 1 dx$ for measurable sets A as described in the week 2 slides. This distribution will be important in hypothesis testing.

Are the events $A = \{s \in S | 0 \leq s \leq \frac{1}{2}\}$ and $B = \{s \in S | \frac{1}{4} \leq s \leq \frac{3}{4}\}$ independent? To answer this, please address the following questions:

Question 2.1

What is $\mathcal{P}(A)$?

Your answer:

$$\mathcal{P}(A) = \frac{1}{2}$$

Question 2.2

What is $\mathcal{P}(B)$?

Your answer:

$$\mathcal{P}(B) = \frac{1}{2}$$

Question 2.3

What is $\mathcal{P}(A \cap B)$?

Your answer:

$$\mathcal{P}(A \cap B) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

Question 2.4

Are the events A and B independent? Please answer yes or no and explain your response using the calculations above.

Your answer: Yes, the events A and B are independent because:

$$\mathcal{P}(A \cap B) = \frac{1}{4} = \mathcal{P}(A) \cdot \mathcal{P}(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

So, $\mathcal{P}(A \cap B) = \mathcal{P}(A) \cdot \mathcal{P}(B)$.

Question 3

Data

This data set contains data on body temperatures, sex, and pulse rates for a sample of 50 healthy adults. We will focus on body temperatures as this can for important indicate of health and sickness.

```
attach(BodyTemp50)
head(BodyTemp50)
```

```
##   BodyTemp Pulse Sex
## 1     97.6    69   0
## 2     99.4    77   1
## 3     99.0    75   0
```

```
## 4     98.8    84   1
## 5     98.0    71   0
## 6     98.9    76   1
```

```
head(BodyTemp)
```

```
## [1] 97.6 99.4 99.0 98.8 98.0 98.9
```

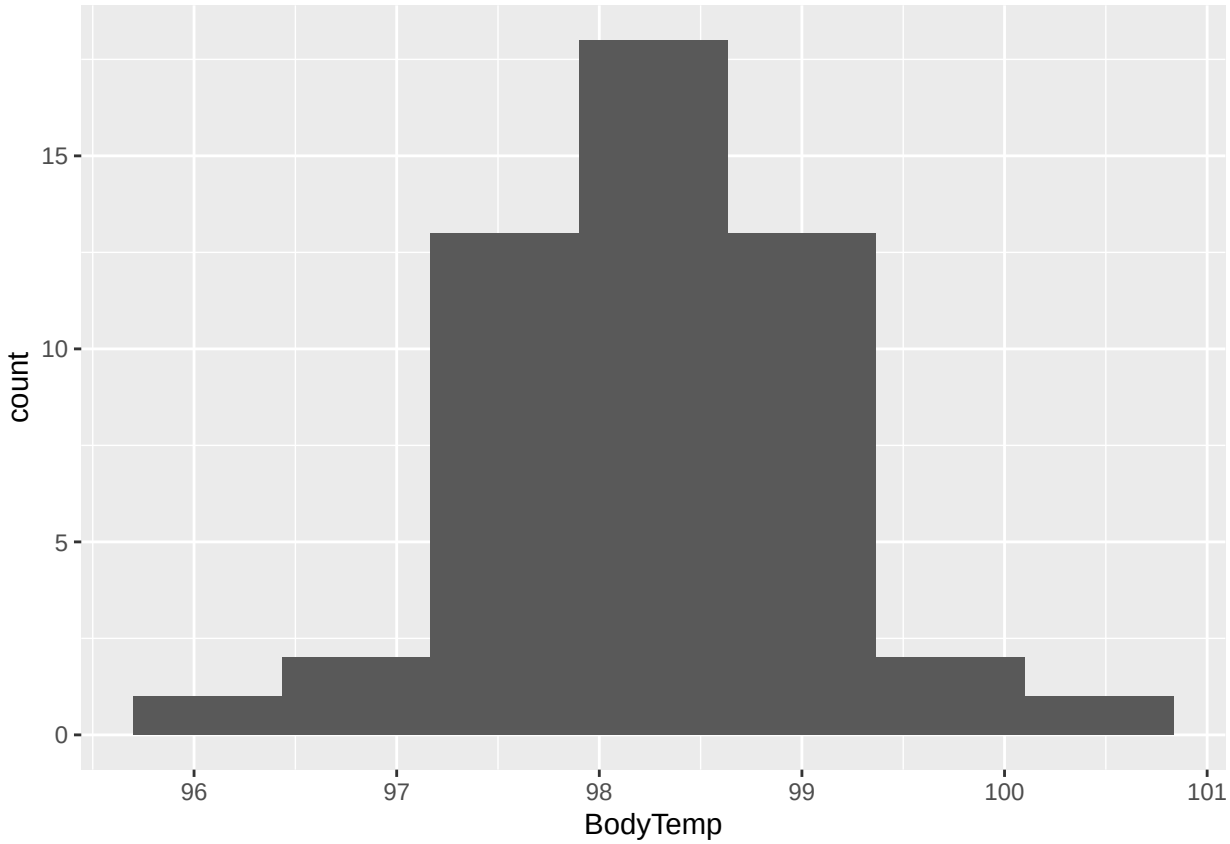
Context

This question concerns (the normal) distributions, the continuous probability space $(\mathcal{S}, \mathcal{M}, \mathcal{P})$ with $\mathcal{S} = (-\infty, \infty)$ and $\mathcal{P}$ defined by $\mathcal{P}(A) = \int_A \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx$ for measurable sets A as described in the week 2 material. This distribution will be essential in future methods. We will use R to assess if data could be considered normally distributed, calculating probability with (R and) the normal distribution, and reasoning with data that is normally distributed.

Question 3.1

Use ggplot to make a histogram of body temperature, and use an appropriate number of bins for the histogram. Based on the histogram do you believe the data to be normally distributed?

```
ggplot(data = BodyTemp50 , aes(x = BodyTemp )) +  geom_histogram(bins = 7)
```

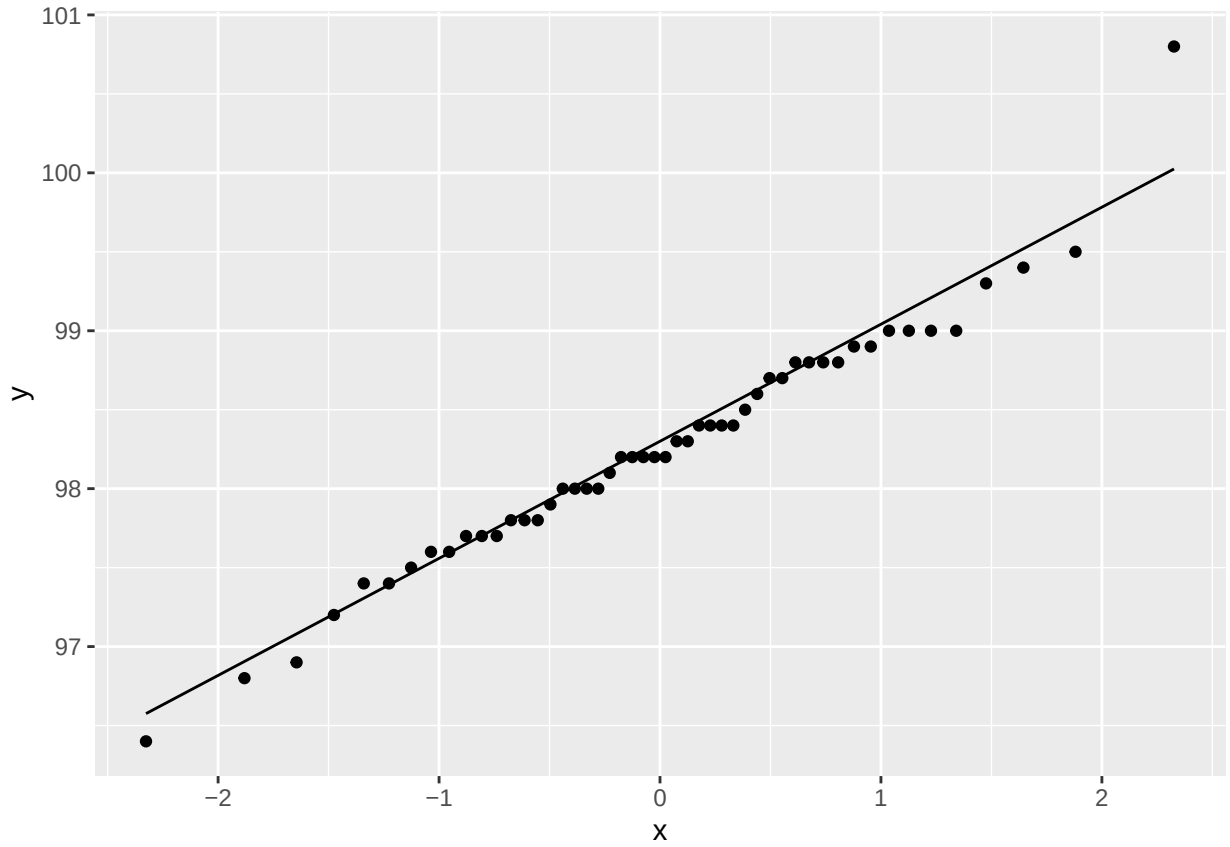


Your answer: Based on the histogram, the data appears to be approximately normally distributed, as it shows a bell-shaped curve centered around the mean body temperature.

Question 3.2

In the plot below we have made a qqplot of body temperature. In a qqplot the data is considered normally distributed if it lies approximately on the qqline. This is a common type of visual assessment to assess normality. It is important to keep in mind that due to randomness of the data the points do not lie perfectly on a line, but the extent to which we accept these deviations from the qqline is subjective. State if you think the data set could be normally distributed according to the qqplot.

```
ggplot(BodyTemp50, aes(sample = BodyTemp)) +
  stat_qq() +
  stat_qq_line()
```



Your answer: According to the QQ plot, the data points mostly lie along the QQ line, suggesting that the body temperatures could be normally distributed.

Question 3.3

A common numerical assessment of normality is called the Shapiro test. In the Shapiro test the null hypothesis is that the data is normally distributed, and the alternative is that the data is not normally distributed. State if the data for body temperatures has a normal distribution based on the p-value.

```
shapiro.test(BodyTemp)
```

```
##
```

```
##  Shapiro-Wilk normality test
##
## data:  BodyTemp
## W = 0.97322, p-value = 0.3115
```

Your answer: Since the p-value (0.3115) is greater than 0.05, we fail to reject the null hypothesis. Therefore, we conclude that the body temperatures are normally distributed.

Question 4

Context

We will use the same data set as problem 3. We will calculate probabilities using the normal distribution, and use the parameters $\mu = 98.26$, and $\sigma = .765$ to model the human body temperature.

Question 4.1

In the code below we calculate the probability an individual has a body temperature between 98.1 and 98.6 degrees units using (1) the normal distribution, and (2) using the data. Are the two methods in relative agreement?

```
mu = mean(BodyTemp)
s = sd(BodyTemp)
#using the data
mean( BodyTemp < 98.6 & BodyTemp >  98.1  )
```

```
## [1] 0.24
```

```
#using the normal distribution
pnorm(98.6, mean = mu, sd= s) - pnorm(98.1, mean = mu, sd= s)
```

```
## [1] 0.2543727
```

Your answer: The two methods are in relative agreement, as the probabilities calculated using the data (0.24) and the normal distribution (0.2543727) are very close. This indicates that the normal distribution is a good model for the body temperature data.

Events & additional probability

Consider the events A the human body temperature is less than 98.6 degrees, and B the human body temperature is more than 98.1 degrees. Answer these questions

Question 4.2

Please give a numerical approximation to $\mathcal{P}(A)$ using the normal distribution.

Your answer:

```r
mu <- 98.26
sigma <- 0.765
prob_A <- pnorm(98.6, mean = mu, sd = sigma)
prob_A
```

```
## [1] 0.6716394
```

Question 4.3

Please give a numerical approximation to $\mathcal{P}(B)$ using the normal distribution?

Your answer:

```r
mu <- 98.26
sigma <- 0.765
prob_B <- 1 - pnorm(98.1, mean = mu, sd = sigma)
prob_B
```

```
## [1] 0.5828346
```

Question 4.4

Please give a numerical approximation to $\mathcal{P}(A \cap B)$ using the normal distribution.

Your answer:

```r
prob_A_and_B <- pnorm(98.6, mean = mu, sd = sigma) - pnorm(98.1, mean = mu, sd = sigma)
prob_A_and_B
```

```
## [1] 0.2544739
```

Question 4.5

Are the events A and B independent? Please answer yes or no and explain your response using the calculations above.

Your answer: No, the events A and B are not independent. - $\mathcal{P}(A) = \text{pnorm}(98.6, \text{mean} = 98.26, \text{sd} = 0.765) = 0.6976$ - $\mathcal{P}(B) = 1 - \text{pnorm}(98.1, \text{mean} = 98.26, \text{sd} = 0.765) = 0.5636$ - $\mathcal{P}(A \cap B) = \text{pnorm}(98.6, \text{mean} = 98.26, \text{sd} = 0.765) - \text{pnorm}(98.1, \text{mean} = 98.26, \text{sd} = 0.765) = 0.2540$ For independence, we need:

$$\mathcal{P}(A) \cdot \mathcal{P}(B) = 0.6976 \times 0.5636 = 0.3931$$

Since:

$$\mathcal{P}(A \cap B) = 0.2540 \neq 0.3931$$

The actual value of $\mathcal{P}(A \cap B)$ is not equal to the product of $\mathcal{P}(A)$ and $\mathcal{P}(B)$. Therefore, the events A and B are not independent.

Question 4.6

Use the 'qnorm' function to determine what body temperature would be unusually large. Quantify what you mean by unusually large.

Your answer: An unusually large body temperature can be defined as one that is in the top 5% of the distribution. This can be found using the 95th percentile.

```
#?qnorm
mu <- 98.26
sigma <- 0.765

# Calculate the 95th percentile
unusually_large_temp <- qnorm(0.95, mean = mu, sd = sigma)
unusually_large_temp
```

```
## [1] 99.51831
```

A body temperature above 99.52 degrees would be considered unusually large. This is because it is in the top 5% of the body temperature distribution, indicating it is significantly higher than the average.