Project #5

Antoinette Lynne, Harini Shanmugam, Aanya Shrivastava

21-11-15

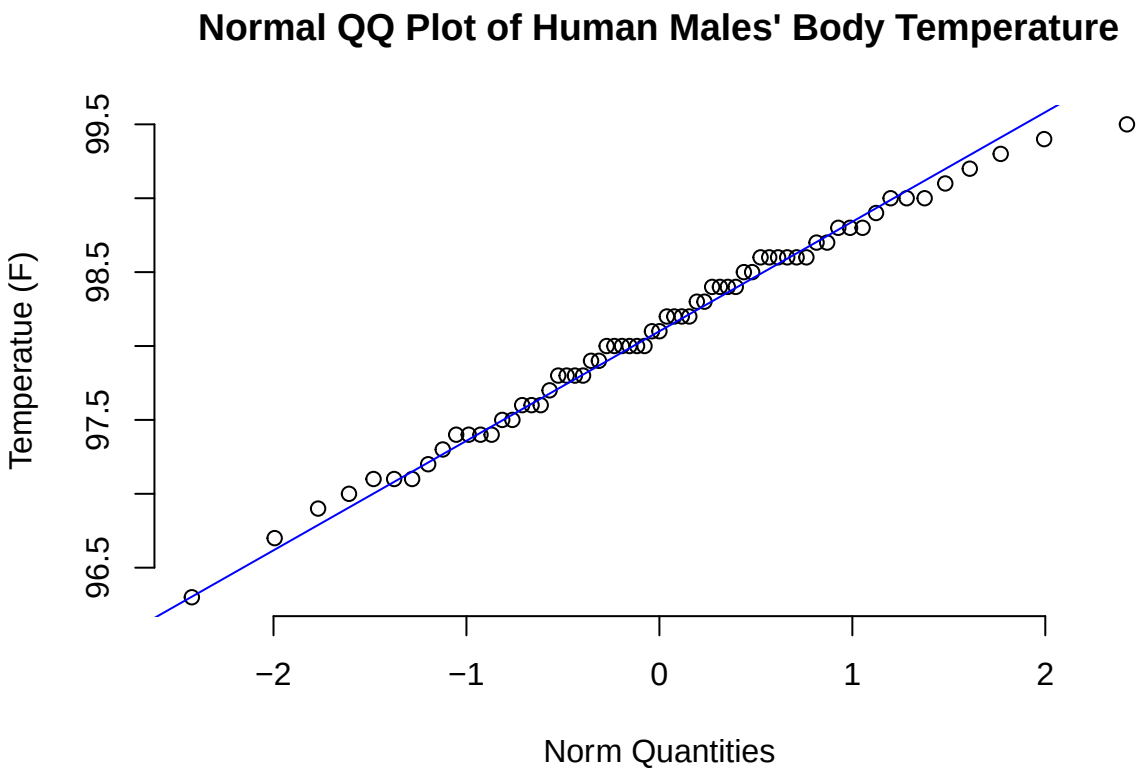
Problem #1 (10 points)

The "normal" temperature of the human body.

The data set provided on the course website was obtained from the following article: *L. Shoemaker Allen (1996) What's Normal? – Temperature, Gender, and Heart Rate, Journal of Statistics Education, 4:2, DOI: 10.1080/10691898.1996.11910512*

(2 points) In this data set, the second column encodes the gender. In this data set, all the cases were either male or female. The males were represented by "1" while the females were represented by "2". Create an appropriate plot in `R` which will help you determine whether the distribution of the male humans' temperatures can be modelled as normal.

```
tempdata = read.csv("temperatures-heart.csv")
maletemp = tempdata$temp[tempdata$gender == 1]
qqnorm(maletemp,
       main="Normal QQ Plot of Human Males' Body Temperature",
       xlab="Norm Quantities",
       ylab="Temperatue (F)",
       pch=1,
       frame=FALSE)
qqline(maletemp,
       col= "blue")
```



The data points are all very close to the line of best fit, which shows that the distribution of males' body temperature can be modeled as normal.

(5 points) From previous studies, you know that the population standard deviation of body temperatures of human males is 0.70. Create an 80%-confidence interval for the mean body temperature of human males.

```
mean_maletemp = mean(maletemp)
mean_maletemp
## [1] 98.10462
sd_maletemp = sd(maletemp)
sd_maletemp
## [1] 0.6987558
```

80% Confidence Interval: $= \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} = 98.10462 \pm 1.28 \frac{0.6987558}{\sqrt{65}} = (97.99368, 98.21555)$

We are 80% confident that the true proportion of the male "normal" body temperature is between 97.99368F and 98.21555F.

(3 points) The "traditionally accepted" normal temperature of a human body is 98.6F. Set the hypotheses for a test of whether that value is the true population mean. What is the $p-$value you obtain? Formulate a conclusion in accordance with the obtained $p-$value.
$H_0 : \mu = 98.6F$ *vs.* $H_a : \mu \neq 98.6F$
$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{98.10462-98.6}{\frac{0.6987558}{\sqrt{65}}} = -5.7157$
$P = \phi(-5.7157) = 5.4744 * 10^{-9}$

Since the P value is lower than 0.05, we reject the null hypothesis.

Problem #2 (20 points)

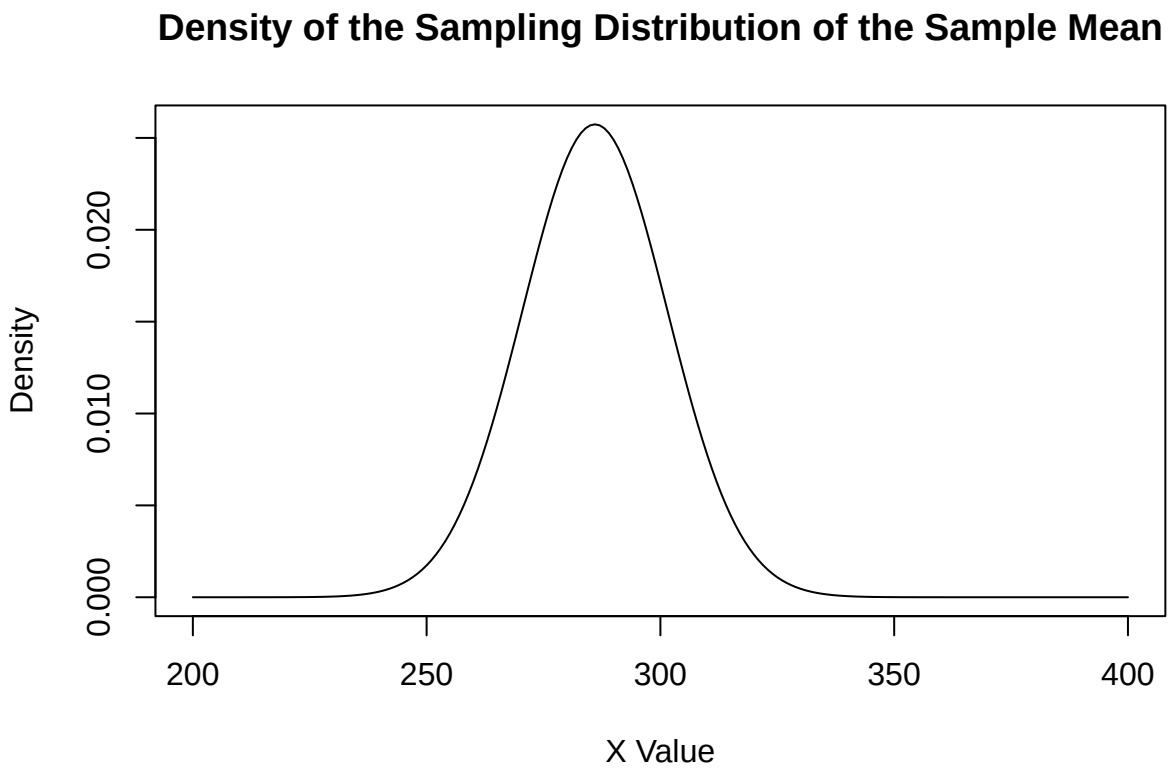
The operating characteristic curve: Power of a $z-$test.

Consider the following test of a hypothesis about the mean consumption of sugar-sweetened beverages at your university based on a sample of size of 100. In this case we'll assume that the population distribution is normal and that the population standard deviation is given at 155 calories. The hypotheses are

$$H_0 : \mu = 286 \quad vs. \quad H_a : \mu < 286$$

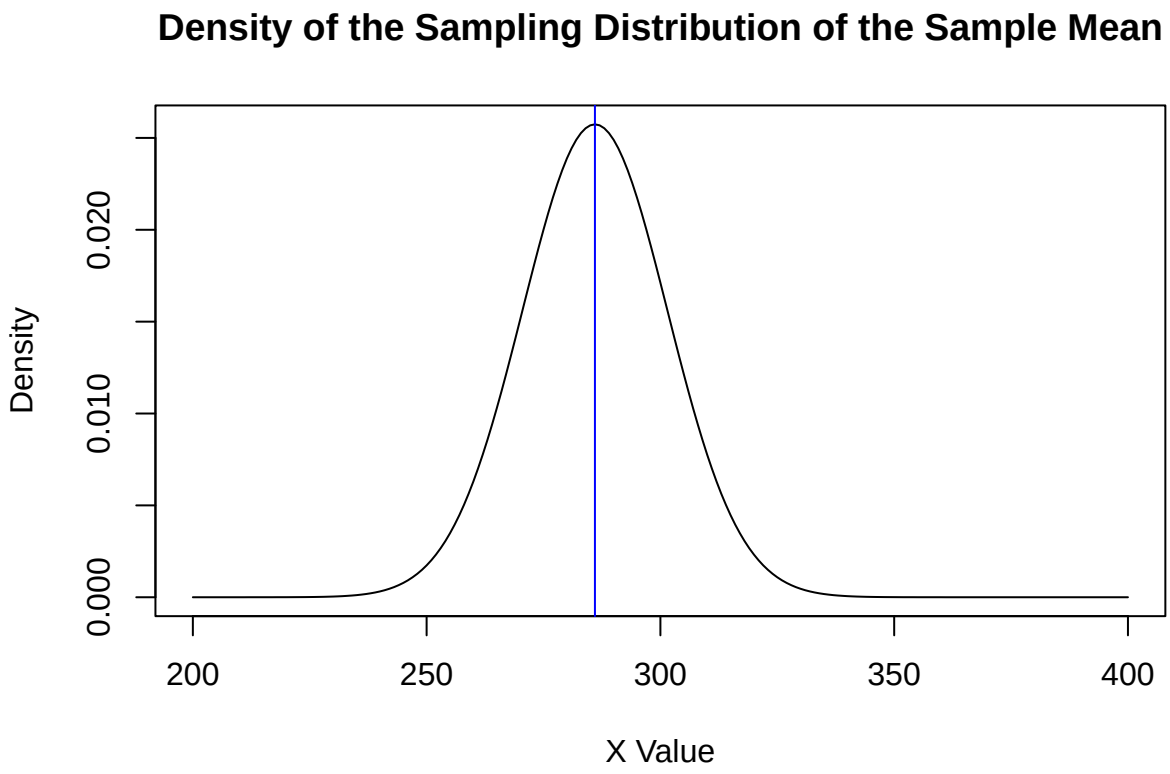
(i) (12 points) What is the *rejection region* under the above null hypothesis for a one-sided alternative with the significance level of 0.05? Complete the following steps in **R**:
(3 points) Draw the density of the **sampling distribution** of the sample mean under the null hypothesis.

```
mu=286
sd=155/sqrt(100)
z=qnorm(0.05)

bev = 200:400
dbev = dnorm(bev, mu, sd)
plot(bev, dbev,
     main="Density of the Sampling Distribution of the Sample Mean",
     xlab="X Value",
     ylab="Density",
     xlim=c(200,400),
     type="l",
     lwd=1)
```



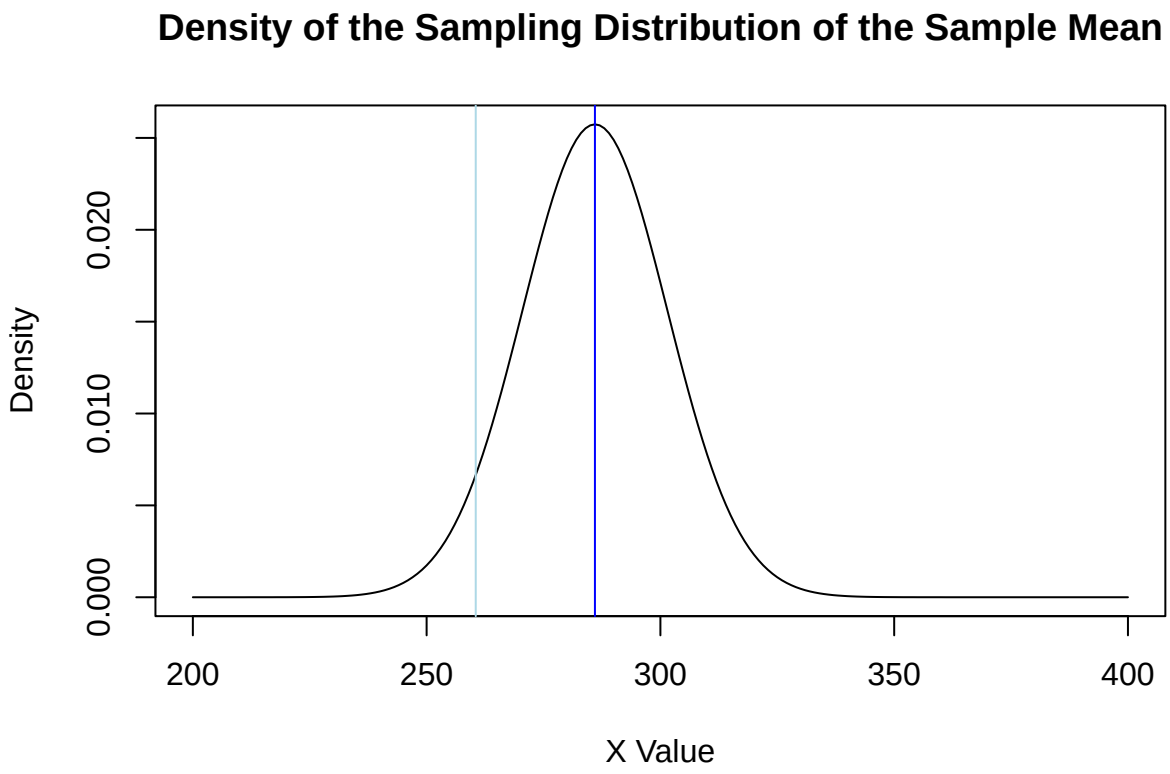
(2 points) Draw the vertical line indicating the **mean** of the population distribution under the null hypothesis (preferably in a different color).

```
plot(bev, dbev,
     main="Density of the Sampling Distribution of the Sample Mean",
     xlab="X Value",
     ylab="Density",
     xlim=c(200,400),
     type="l",
     lwd=1)

abline(v=286,
       col="blue")
```



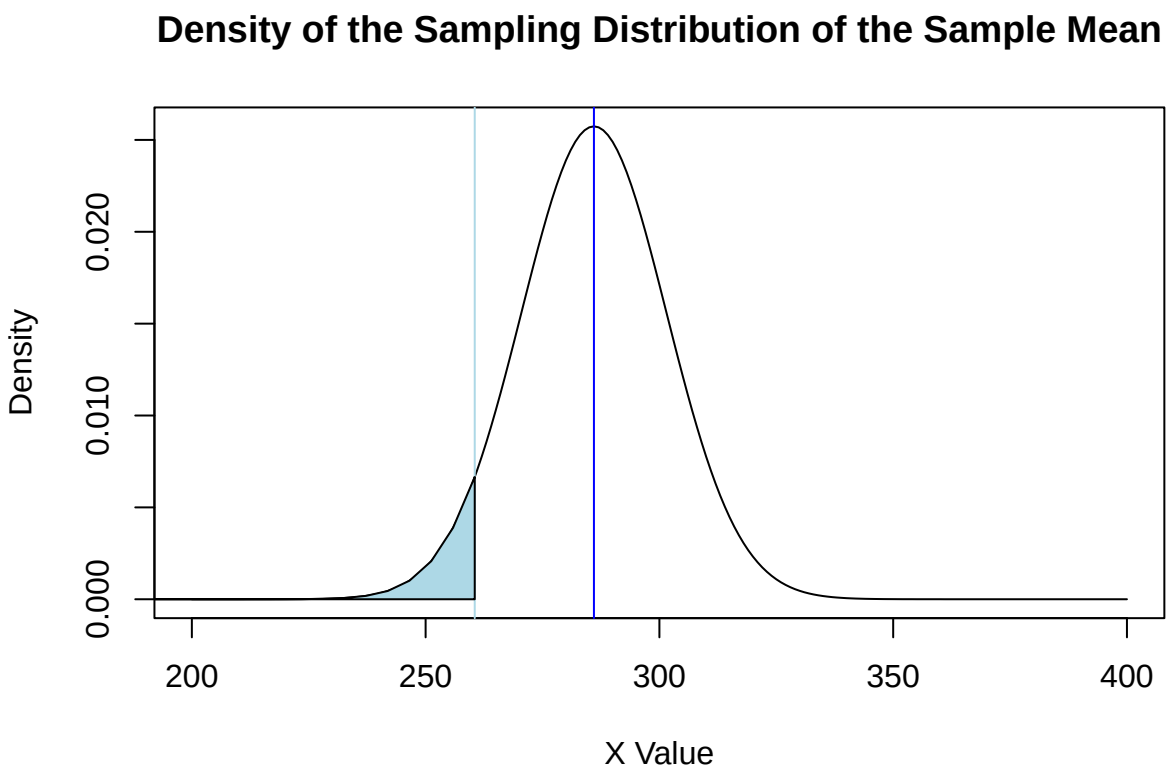
(2 points) Draw the vertical line indicating the **upper bound** of the rejection region for a significance level of 0.05 (preferably in a different color).

```
upperbound = mu+z*sd
upperbound
## [1] 260.5048
plot(bev, dbev,
     main="Density of the Sampling Distribution of the Sample Mean",
     xlab="X Value",
     ylab="Density",
     xlim=c(200,400),
     type="l",
     lwd=1)

abline(v=286,
       col="blue")
abline(v=upperbound,
       col="lightblue")
```



(5 points) Using the `polygon` command (not any of the packages you may have found on the internet), shade the region below the normal density function to the left of the upper bound you found in the previous task.

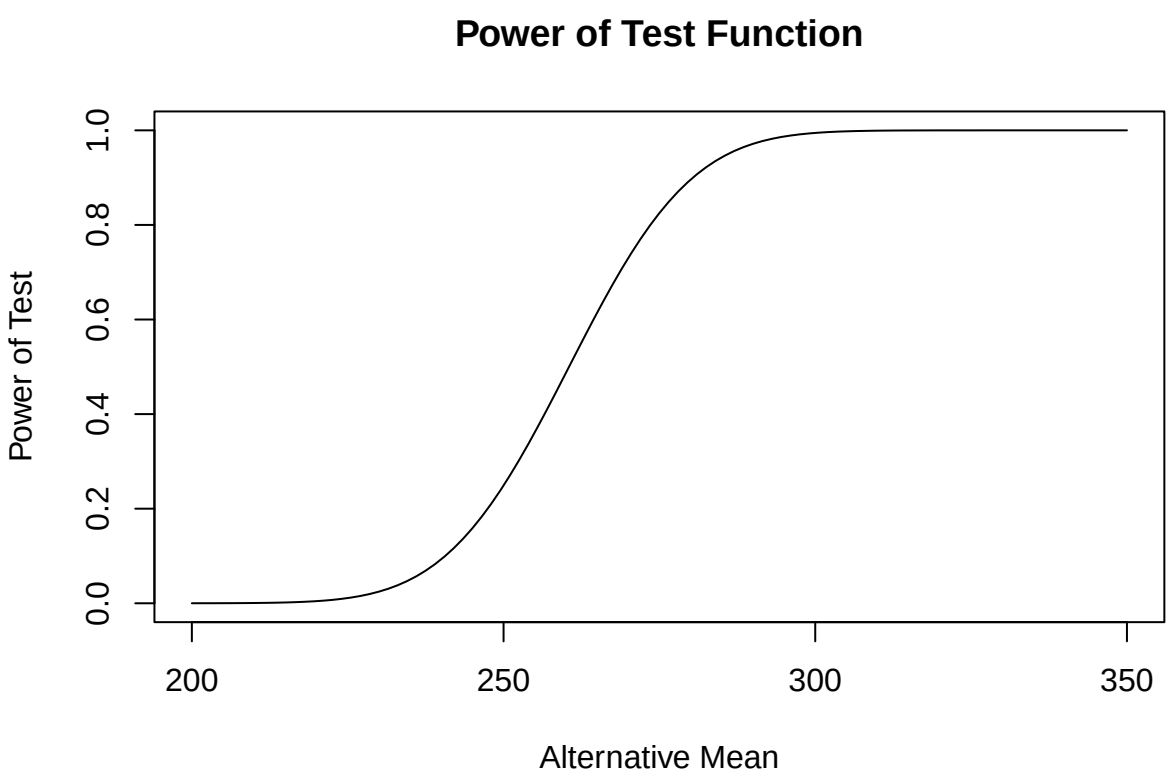
```
 plot(bev, dbev,
     main="Density of the Sampling Distribution of the Sample Mean",
     xlab="X Value",
     ylab="Density",
     xlim=c(200,400),
     type="l",
     lwd=1)

abline(v=286,
       col="blue")

abline(v=upperbound,
       col="lightblue")

x=seq(-200, upperbound, length=100)
y=dnorm(x, mean=286, sd=15.5)
polygon(c(-200,x, upperbound),
        c(0,y,0),
        col="lightblue")
```



(ii) (8 points) What is the correspondence between the alternative values of the population mean and the power of the above test?
(5 points) Define a **function** which will calculate (from first principles) the power of the above test as a function of the alternative population mean.
First principle function: $F_{\mu_a} = P_{\mu_a}[\text{Fail to reject } H_0] = \beta$
F_{μ_a} = 1 - $\phi(\frac{x_a - \mu_a}{\sigma/\sqrt{n}})$
F_{μ_a} = 1 - $\phi(\frac{260.5048 - \mu_a}{155/\sqrt{100}})$

(3 points) Draw the graph of the function you obtained in the previous task.

```
curve(1-pnorm(upperbound, x, sd),
      from=200,
      to=350,
      main="Power of Test Function",
      xlab="Alternative Mean",
      ylab="Power of Test")
```


Power of Test Function
Power of Test
0.0
0.2
0.4
0.6
0.8
1.0
200
250
300
350
Alternative Mean

Problem #3 (14 points)

Our logic survey.

After you have completed the surveys you received in an email, you can watch the following videos for fun. They are not necessary for the remainder of the problem, but they are entertaining and informative.

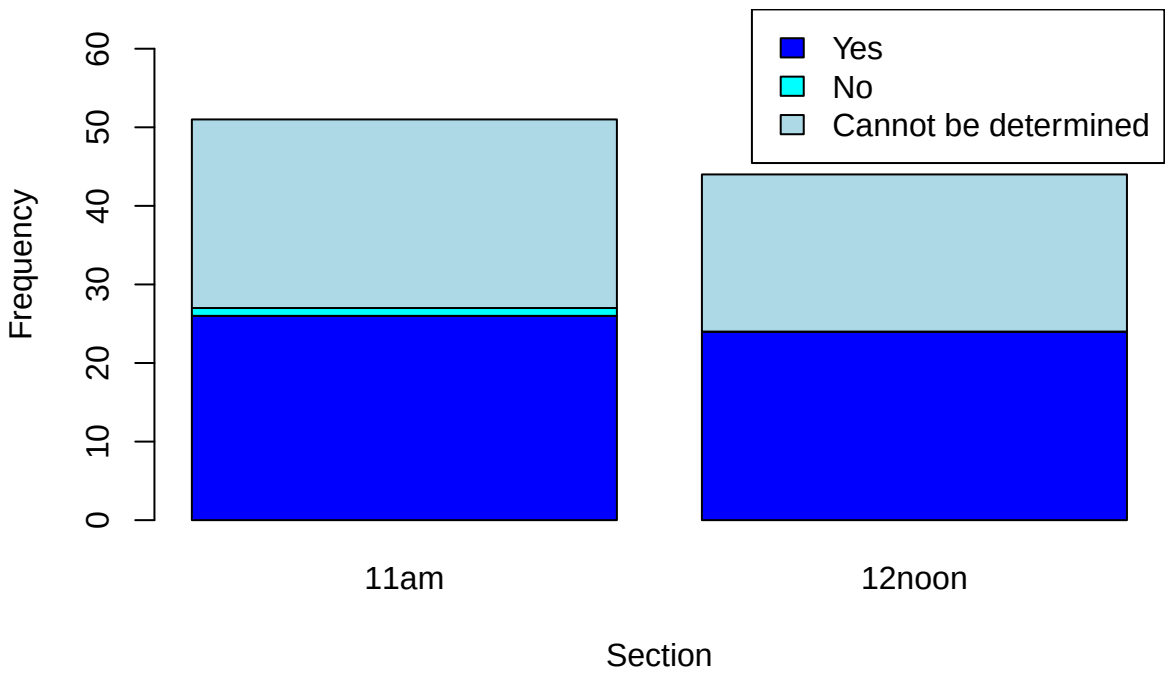
Video #1

Video #2

Video #3

(i) Let us first figure out if any of the two sections is doing demonstrably better. (4 points) "Clean up'' the data in the spreadsheet so that you have the information that you need for the test. Then, summarize the results of our survey visually.

```
puzzleresponse = matrix(c(26,1,24,24,0,20), ncol=2)
rownames(puzzleresponse) = c("Yes", "No", "Cannot be determined")
colnames(puzzleresponse) = c("11am", "12noon")
puzzleresponsetable = as.table(puzzleresponse)
puzzleresponse
##                      11am 12noon
## Yes                    26     24
## No                      1      0
## Cannot be determined   24     20
barplot(puzzleresponse,
        main="Responses to Logic Puzzle from Both Sections",
        ylab="Frequency",
        xlab="Section",
        ylim=c(0,65),
        col = c("blue", "cyan", "lightblue"))

legend("topright",
       legend = c("Yes", "No", "Cannot be determined"),
       fill = c("blue", "cyan", "lightblue"))
```



(6 points) Use **R** to test

$$H_0 : p_{am} = p_{noon} \quad vs. \quad H_a : p_{am} \neq p_{noon}$$

and report the $p-$value. Note: Do not use the built-in prop.test command here!

```
p_am = 26/53
p_noon = 24/45

p_tot = (26+24)/ (53+45)

z = (p_am-p_noon)/ (sqrt(p_tot*(1-p_tot)*((1/53)+(1/45))))

p_value = 2*pnorm(z)
p_value
## [1] 0.6729889
```

(ii) (4 points) The reported proportion of 20% correct answers in the general population was what prompted the survey talked about in the above videos. One would assume that the students taking applied statistics are more capable of logical thought than the general population. So, for p denoting the population proportion for applied statistics students, let us test the following hypothesis:

$$H_0 : p = 0.20 \quad vs. \quad H_a : p > 0.20$$

and report the $p-$value.

```r
correct = (26+24)/98
z = (correct-.2)/ (sqrt((.2*.8)/(98)))
p_value = 1-pnorm(z,0,1)
p_value
## [1] 8.104628e-15
```

Problem #4 (16 points)

Pizza & ice cream.

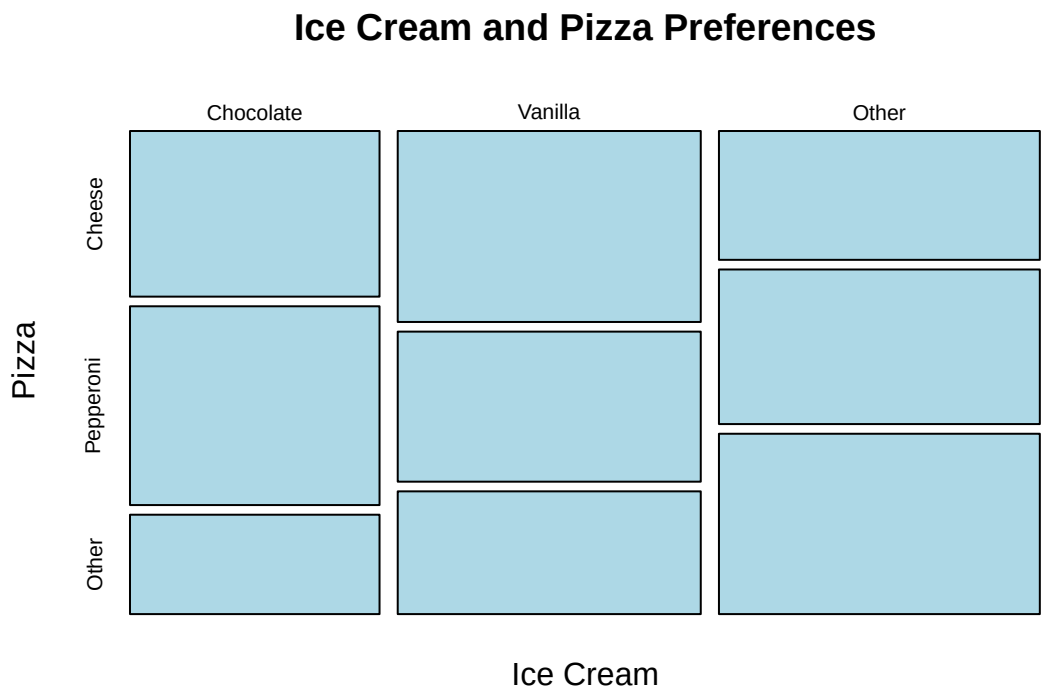
Recently you were sent a link to a survey regarding pizza and ice-cream preferences. The results, as Google reports them are in the spreadsheet you received as an attachment. To see if there is any evidence of association between pizza and ice-cream preferences, please do the following:

(i) (4 points) "Clean up'' the spreadsheet so that you have more manageable entries in the cells. You can do this in R or using some Excel-like software. Then, create and display a **two-way table** summarizing the results of our survey.

```
foodpref = matrix(c(10,14,10,12,11,12,6,9,14), ncol = 3)
rownames(foodpref) = c('Chocolate','Vanilla','Other')
colnames(foodpref) = c('Cheese', 'Pepperoni','Other')
foodpreftable = as.table(foodpref)
foodpreftable
##           Cheese Pepperoni Other
## Chocolate     10        12     6
## Vanilla       14        11     9
## Other         10        12    14
```

(ii) (8 points) Graph the data from the two-way table you obtained above. Creative data presentation will earn bonus points. Do not be afraid to download additional R libraries.

```
rownames(foodpref) = c('Chocolate','Vanilla','Other')
colnames(foodpref) = c('Cheese', 'Pepperoni','Other')

mosaicplot(foodpreftable,
           main="Ice Cream and Pizza Preferences",
           xlab="Ice Cream",
           ylab="Pizza",
           col="lightblue")
```



(iii) (4 points) Perform that χ^2-test to see if there is an association between your subjects' preferences.

```
chisq.test(foodpreftable)
##
##  Pearson's Chi-squared test
##
## data:  foodpreftable
## X-squared = 3.2753, df = 4, p-value = 0.5129
```

Since the X-squared statistic is relatively low, there is not much variation between the two categories (pizza and ice cream). Since the P-value is high, we fail to reject the null hypothesis (that there is no correlation between the choices of pizza and ice cream) and so the choices between the two are independent.