Project #5

Ben Cartwright, Heewon Huh, Sally Kyong, Harvey Lee

2021-11-14

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*

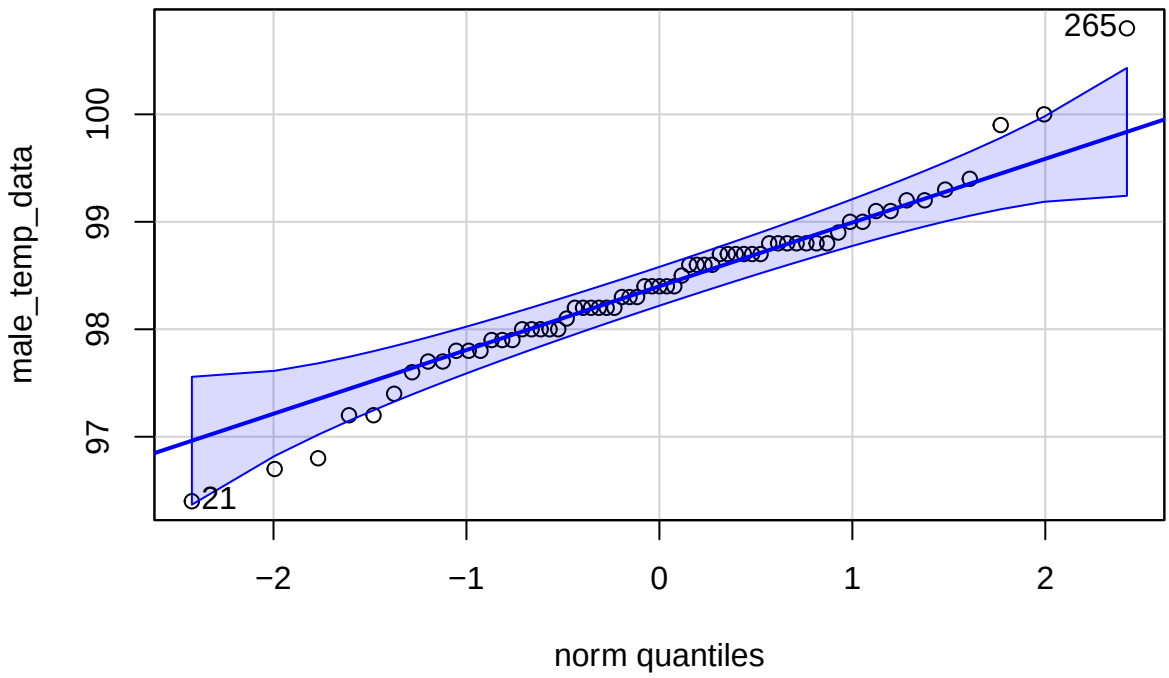
```
data<- read.csv("temperatures-heart.csv")
```

(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.

Solution:

```
# Male Temperature Data Normal?

gendered_temp_data = split(data$temp, data$gender)
male_temp_data = unlist(gendered_temp_data[2])
install.packages("car", , repos = "http://cran.us.r-project.org")
## Installing package into 'C:/Users/heewo/Documents/R/win-library/4.1'
## (as 'lib' is unspecified)
##
##   There is a binary version available but the source version is later:
##     binary source needs_compilation
## car 3.0-11 3.0-12             FALSE
## installing the source package 'car'
library(car)
## Loading required package: carData
qqPlot(male_temp_data)
```




```
## 265  21
##  65   1

# qqPlot function shows that the most points fall approximately along the straight line,
# so we can assume normality.
```

(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.

Solution:

```
# 80% confidence interval

# Start with finding a  critical value Z*.
z.star0.8<- qnorm(0.8+(1-0.8)/2)
print(z.star0.8)
## [1] 1.281552
#Calculate the margin of error using the Z* and the information given in the problem.
me<- (z.star0.8*0.7)/sqrt(length(male_temp_data))
#Calculate the upper and lower bounds.
upperbound<- mean(male_temp_data) + me
lowerbound<- mean(male_temp_data) - me
print(upperbound)
## [1] 98.50512
print(lowerbound)
## [1] 98.28258
```

The 80% confidence interval is (98.28258, 98.50512).

(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.

```
# Hypothesis Test

# Our null hypothesis is that the population mean is 98.6
# Our alternative hypothesis is that the population mean does not equal 98.6

n = length(male_temp_data)
mean = mean(male_temp_data)
z = (mean - 98.6)/(0.7/sqrt(n))
p_value = pnorm(z) + (1 - pnorm(-z))
print(p_value)
## [1] 0.01757849

# With a p-value of less than 0.05, the data exhibits significant evidence against the null
# hypothesis. This evidence is strong enough to reject the null hypothesis in favor of
# the alternative.
```

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**:

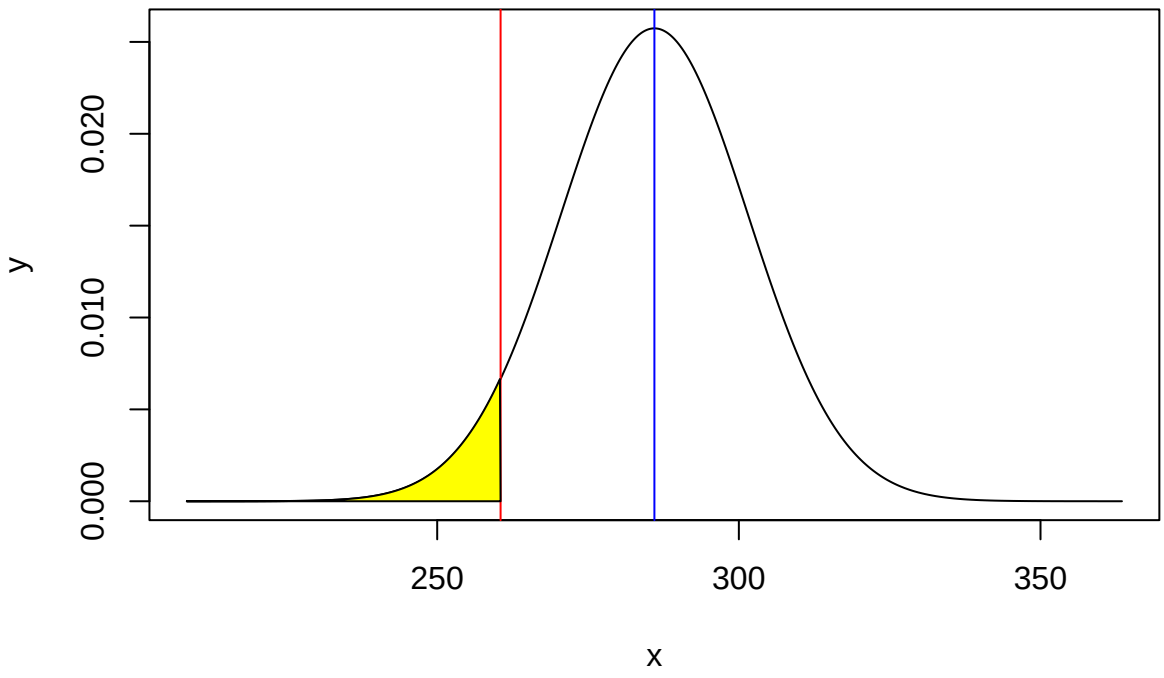
- (3 points) Draw the density of the **sampling distribution** of the sample mean under the null hypothesis.
- (2 points) Draw the vertical line indicating the **mean** of the population distribution under the null hypothesis (preferably in a different color).
- (2 points) Draw the vertical line indicating the **upper bound** of the rejection region for a significance level of 0.05 (preferably in a different color).
- (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.

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

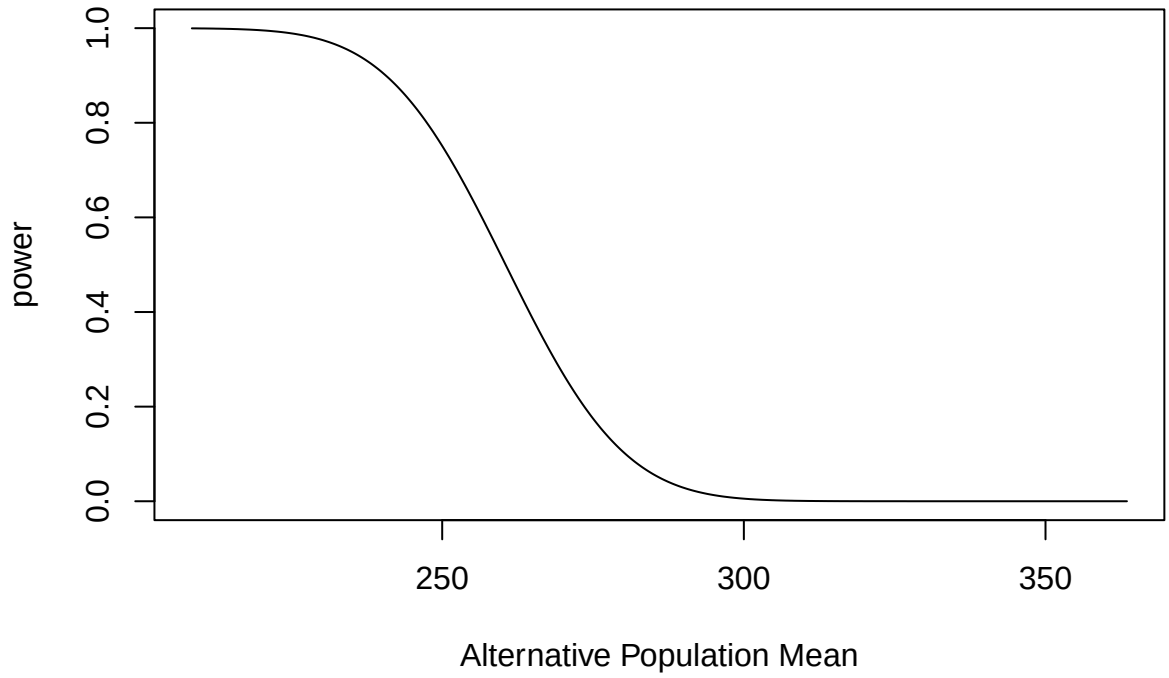
Solution:

```
# i
x = seq(286 - 5 * 15.5, 286 + 5 * 15.5, 0.1)
y = dnorm(x, mean = 286, sd = 15.5)
plot(x, y, type = "l")
abline(v = 286, col = "blue")
upperbound = qnorm(0.05, 286, 15.5)
abline(v = upperbound, col = "red")
polygon(c(min(x), x[x <= upperbound]), c(y[x <= upperbound], y[x == min(x)]),
        col = "yellow")
```




```
# ii
power.function = function(alt.pop.mean){
  z.alpha = qnorm(0.05)
  return(pnorm((286-alt.pop.mean)*(sqrt(100)/155)+z.alpha))
}
power.function
## function(alt.pop.mean){
##   z.alpha = qnorm(0.05)
##   return(pnorm((286-alt.pop.mean)*(sqrt(100)/155)+z.alpha))
## }

x = seq(286-5*15.5, 286+5*15.5, length = 470)
y = rep(0, 470)
for(i in 1:470){
  y[i] = power.function(x[i])
}
plot(x, y, xlab = "Alternative Population Mean", ylab = "power", type = "l")
```


power
0.0
0.2
0.4
0.6
0.8
1.0
250
300
350
Alternative Population 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.
- (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!*

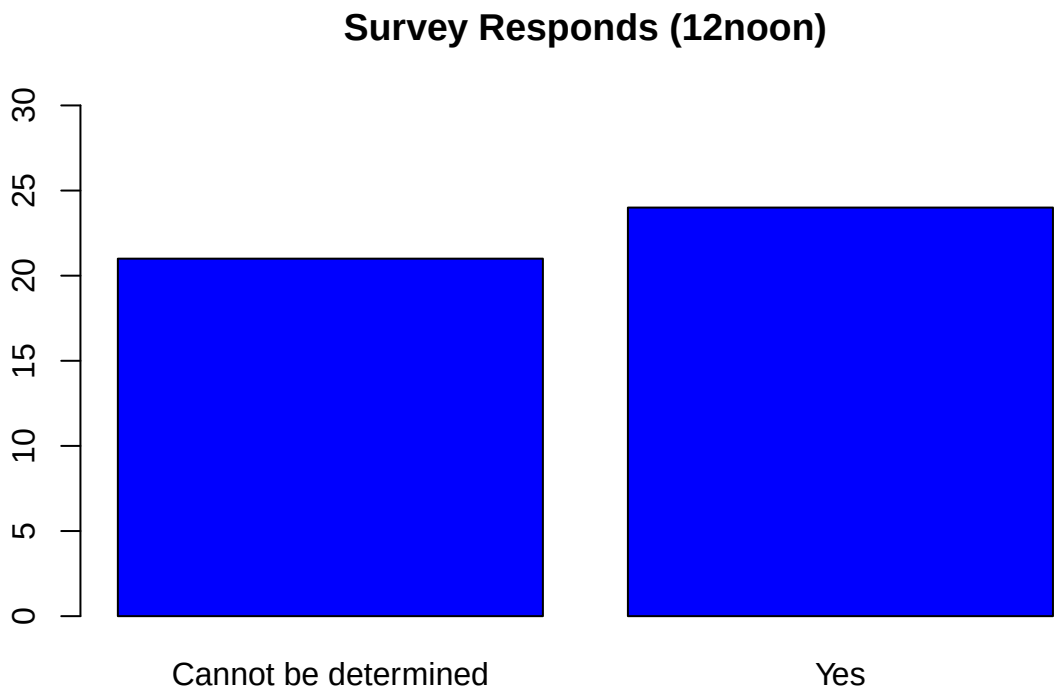
(ii) (4 points) The reported proportion of 20% correct answers in the general population was what prompted the survey talked about in the above videos. One would assume that the students taking applied statistics are more capable of logical thought than the general population. So, for p denoting the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

Solution:

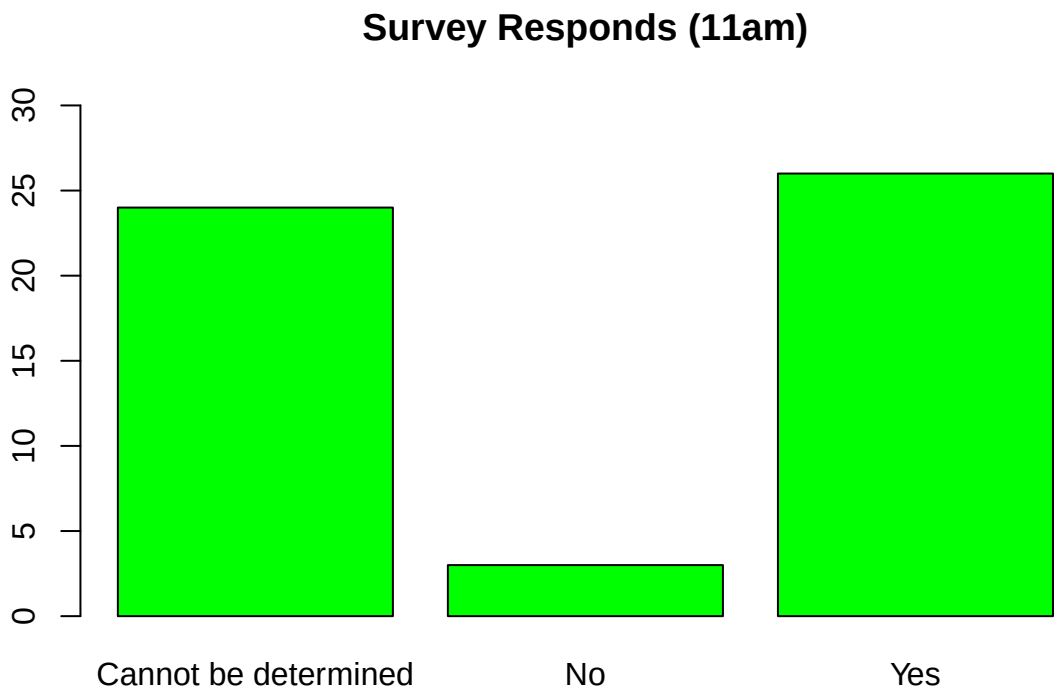
```
# i
# Clean up and visualize data
poll<- read.csv("married-folks-anonymized-1.csv")
noon<- poll[which(poll$My.section.in.M358K.is.=="12noon"),]
noon.table<- table(noon$What.s.the.answer.to.the.above.logic.puzzle.)
eleven<- poll[which(poll$My.section.in.M358K.is.=="11am"),]
eleven.table<- table(eleven$What.s.the.answer.to.the.above.logic.puzzle.)
columns<- c("Response", "Count")
barplot(noon.table, main="Survey Responds (12noon)", col="blue", ylim = c(0,30))
```




```
knitr::kable(noon.table, col.names = columns)
```

Response	Count
Cannot be determined	21
Yes	24

```
barplot(eleven.table, main="Survey Responds (11am)", col="green", ylim = c(0,30))
```




```
knitr::kable(eleven.table, col.names = columns)
```

Response	Count
Cannot be determined	24
No	3
Yes	26

```
# Test hypotheses
# Find p, the proportion of students who answered correct, for each section
p.am<-eleven.table[3]/(eleven.table[1]+eleven.table[2]+eleven.table[3])
n.am<- eleven.table[1]+eleven.table[2]+eleven.table[3]
p.noon<-noon.table[2]/(noon.table[1]+noon.table[2])
n.noon<- noon.table[1]+noon.table[2]
print(p.am)
##      Yes
## 0.490566
print(p.noon)
##       Yes
## 0.5333333

# Find p.hat and standard deviation
p.hat<- (n.am*p.am+n.noon*p.noon)/(n.am+n.noon)
sd<- sqrt(p.hat*(1-p.hat)*((1/n.am)+(1/n.noon)))
```

```r
# Find Z statistics
z<- (p.am-p.noon)/sd
print(z)
##        Yes
## -0.4220495

# Calculate p-value
p.value<- pnorm(-abs(z))+(1-pnorm(abs(z)))
print(p.value)
##       Yes
## 0.6729889

# Since the p-value is large, it indicates weak evidence against the null hypothesis.

# ii
# Find p.hat
poll.table<- table(poll$What.s.the.answer.to.the.above.logic.puzzle.)
p.hat<-poll.table[3]/(poll.table[1]+poll.table[2]+poll.table[3])
n<- (poll.table[1]+poll.table[2]+poll.table[3])
print(p.hat)
##       Yes
## 0.5102041

# Find Z statistics
z<- (p.hat-0.2)/sqrt((0.2*0.8)/n)
print(z)
##      Yes
## 7.677159

# Find the p-value
p.value<- 1-pnorm(z)
print(p.value)
##          Yes
## 8.104628e-15

# Since p-value is small, it indicates strong evidence against the null hypothesis.
```

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.

(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.

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

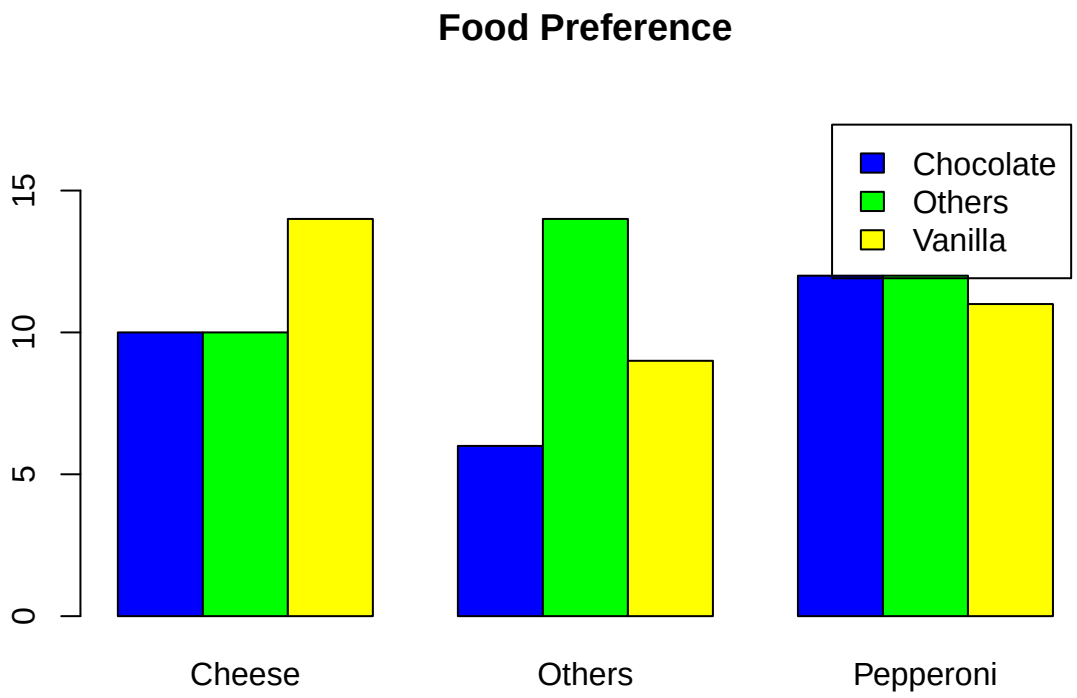
i

Solution:

```
# Clean up and visualize data
food_pref = read.csv("food-prefs-anonymized-1.csv")
t = table(food_pref$What.is.your.favorite.ice.cream.flavor.,
          food_pref$What.is.your.favorite.kind.of.pizza.)
colnames(t)<- c("Cheese", "Others", "Pepperoni")
rownames(t)<- c("Chocolate", "Others", "Vanilla")
knitr::kable(t)
```

	Cheese	Others	Pepperoni
Chocolate	10	6	12
Others	10	14	12
Vanilla	14	9	11

```
# Create the graph
barplot(t, beside = TRUE, main = "Food Preference", legend = TRUE, ylim = c(0,18),
        col = c("blue", "green", "yellow"))
```



ii

Solution:

Set up the χ^2-test:

$H_\circ$: Pizza and ice cream preferences are independent. Pizza preferences do not vary by ice cream preferences.

H_a : Pizza and ice cream preferences are dependent. Pizza preferences vary by ice cream preferences.

```
# Find degree of freedom
df<- (3-1)*(3-1)

# Find total count food preferences
columns<- c("Type", "Count")
pizza<-table(food_pref$What.is.your.favorite.kind.of.pizza.)
knitr::kable(pizza, col.names = columns)
```

Type	Count
Cheese (includes any number of cheeses, Margherita, and the like).	34
Other (includes any other type of pizza, and the unlikely "I do not like pizza" response).	29
Pepperoni (includes salami, artisanal, and the like).	35

```r
ice.cream<-table(food_pref$What.is.your.favorite.ice.cream.flavor.)
knitr::kable(ice.cream, col.names = columns)
```

Type	Count
Chocolate (includes Belgian, dark, white, and the like).	28
Other (includes any other type of ice cream, and the unlikely "I do not like ice cream" response).	36
Vanilla (includes Mexican, French, vanilla bean, and the like).	34

```r
total<- 98

# Find the expected counts of each preference combination
e.chocolate.cheese<- (28*34)/total
e.chocolate.others<- (28*29)/total
e.chocolate.pepperoni<- (28*35)/total
e.others.cheese<- (36*34)/total
e.others.others<- (36*29)/total
e.others.pepperoni<- (36*35)/total
e.vanilla.cheese<- (34*34)/total
e.vanilla.others<- (34*29)/total
e.vanilla.pepperoni<- (34*35)/total

# Calculate the test statistic
a<- (10-e.chocolate.cheese)^2/e.chocolate.cheese
b<- (6-e.chocolate.others)^2/e.chocolate.others
c<- (12-e.chocolate.pepperoni)^2/e.chocolate.pepperoni
d<- (10-e.others.cheese)^2/e.others.cheese
e<- (14-e.others.others)^2/e.others.others
f<- (12-e.others.pepperoni)^2/e.others.pepperoni
g<- (14-e.vanilla.cheese)^2/e.vanilla.cheese
h<- (9-e.vanilla.others)^2/e.vanilla.others
i<- (11-e.vanilla.pepperoni)^2/e.vanilla.pepperoni
chi.squre<- a+b+c+d+e+f+g+h+i
print(chi.squre)
## [1] 3.275281

# Find the p-value
1-pchisq(chi.squre, df= df)
## [1] 0.5128582
```

Conclusion: Since p-value is large, there is a weak evidence against the null hypothesis, so we fail to reject the null hypothesis. Pizza and ice cream preferences are independent. Pizza preferences do not vary by ice cream preferences.