M358K - Project 3

Arun Krishnaraj - ak37738, Keven Li - kl32584, Lizbeth Rayas - ltr369,

11-22-2020

Problem 3.1.

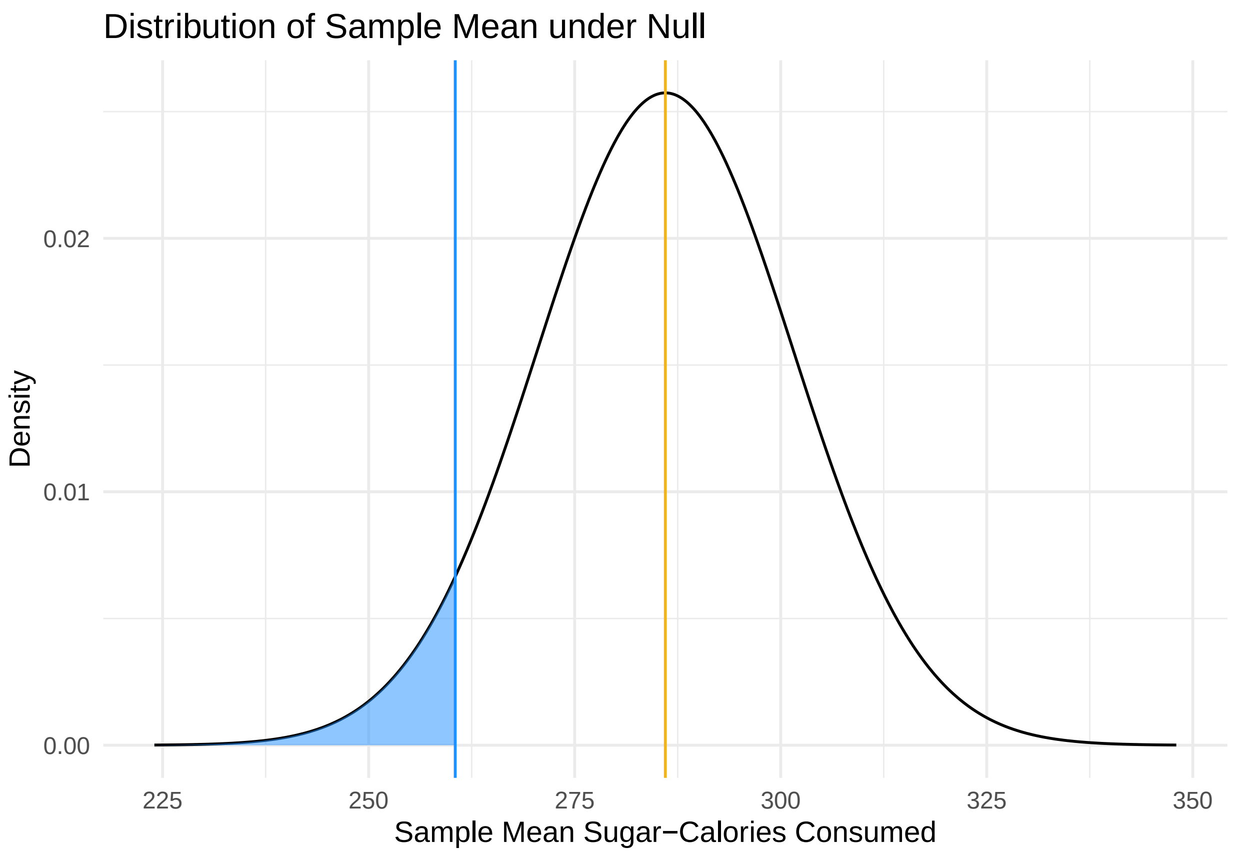
Part i.

We are interested in exploring the mean consumption of sugar-sweetened beverages at our university. We have formulated the null and alternative hypotheses for the population average caloric consumption from sugary-beverages as

$$H_o : \mu = 286 \ vs. \ H_a : \mu < 286$$

We are interested in the rejection region under this null hypothesis for a sample of 100 students; we assume in this case that the population standard deviation is 155 calories. We can first visualize the null sampling distribution and the left-sided rejection region at the 0.05 significance level.

```r
null.mean = 286; samp.size = 100; pop.sd = 155
norm.sd = pop.sd/(samp.size^.5)
x <- seq(null.mean-4*norm.sd, null.mean+4*norm.sd, length = 1000)
z.star <- qnorm(0.05)
hx <- dnorm(x, mean = null.mean, sd = norm.sd)
left.alt.cut = null.mean + z.star*norm.sd

dmerge <- data.frame(x,hx)

library(ggplot2)
## Warning: package ’ggplot2’ was built under R version 3.6.2
ggplot(data = dmerge, aes(x = x, y = hx)) + geom_line() + theme_minimal() +
  xlab("Sample Mean Sugar-Calories Consumed") + ylab("Density") +
  ggtitle("Distribution of Sample Mean under Null") +
  geom_vline(xintercept = null.mean, col = "goldenrod2") +
  geom_vline(xintercept = left.alt.cut, col = "dodgerblue") +
  geom_ribbon(data = dmerge[dmerge$x < left.alt.cut,] ,
              aes(ymin = 0, ymax = hx), fill = "dodgerblue", alpha = 0.5)
```

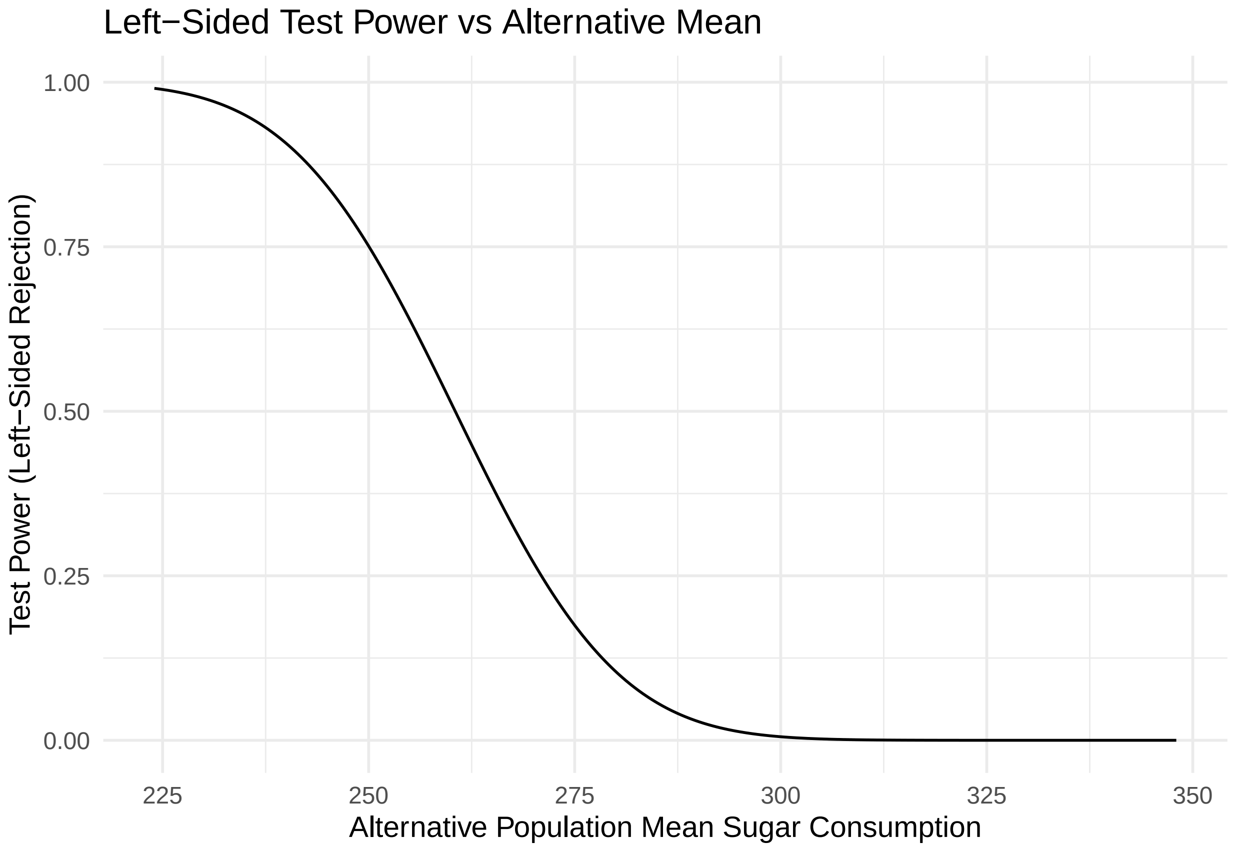


From this figure, we can see that the population mean under the null hypothesis is 286 calories, shown in yellow. The upper-bound of the rejection rejection occurs at $286 - 1.645 \times \frac{155}{\sqrt{100}} = 260.5048$ at the 5% significance level; the corresponding rejection region for a significance level of 0.05 is $(-\infty, 260.5048]$, shown in blue. If the null hypothesis were correct, then there would be a 5% probability of obtaining a sample result as or more negative as the blue cutoff line.

Part ii.

Now that we know the rejection region, we can calculate the power of the test for any alternative population mean; the power of the test is the probability of obtaining a sample result in the rejection region under various specific alternative population means.

```r
test_pow <- function(alternative){
  pnorm(left.alt.cut, mean = alternative, sd = norm.sd)
}

hx.pow <- test_pow(x)

dmerge.pow <- data.frame(x,hx.pow)
ggplot(data = dmerge.pow, aes(x=x,y=hx.pow)) + geom_line() + theme_minimal() +
  xlab("Alternative Population Mean Sugar Consumption") +
  ylab("Test Power (Left-Sided Rejection)") +
  ggtitle("Left-Sided Test Power vs Alternative Mean")
```



The power of the test decreases as alternative value of the population mean increases; the power of the test experiences a large initial drop, followed by a gradual decrease in slope as alternative population mean increases past 260 calories.

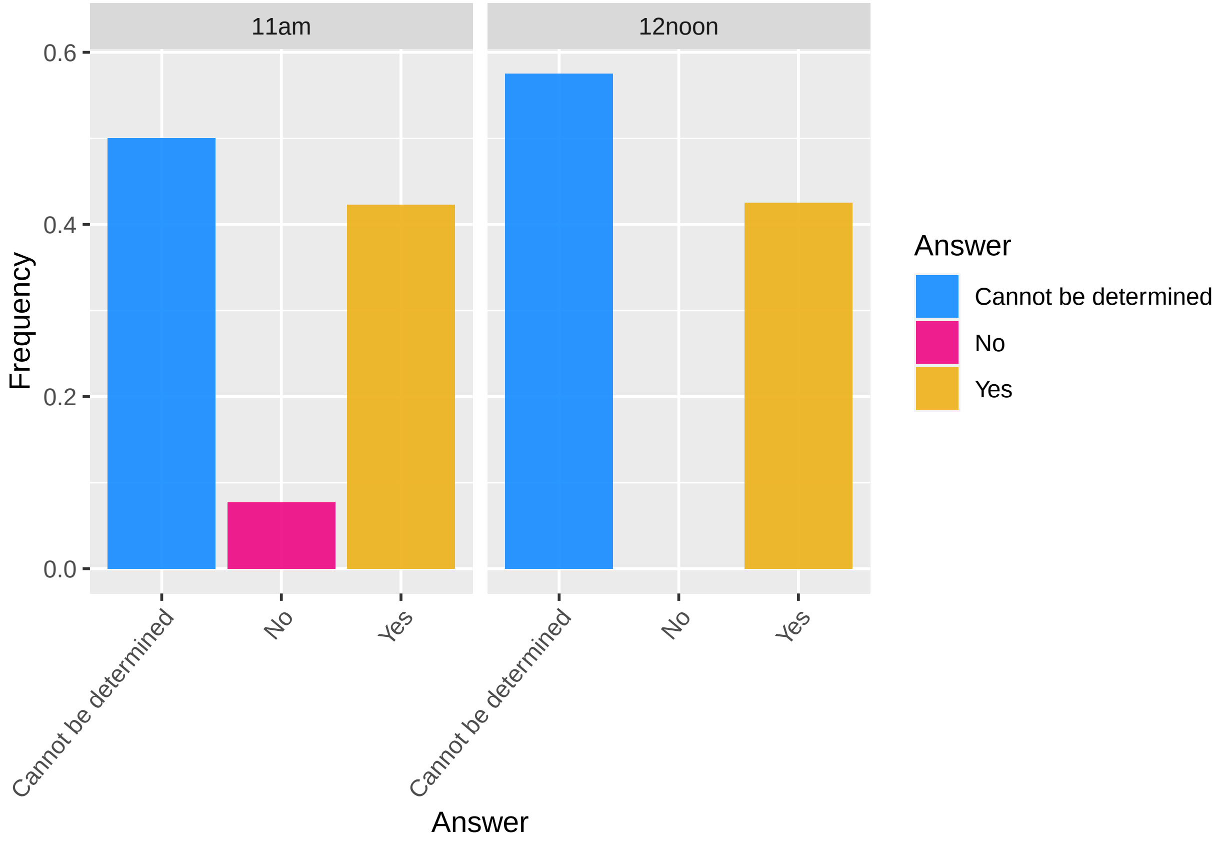
Problem 3.2.

Part i.

We are interested in comparing the results of the logic survey between classes to determine if any difference in proportion correct exists.

```
library(tidyverse)
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
## v purrr   0.3.4
## Warning: package 'tibble' was built under R version 3.6.2
## Warning: package 'tidyr' was built under R version 3.6.2
## Warning: package 'readr' was built under R version 3.6.2
## Warning: package 'purrr' was built under R version 3.6.2
## Warning: package 'dplyr' was built under R version 3.6.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
marriage <- read.csv("logic.csv", stringsAsFactors = F)

marriage <- marriage[,2:3]

colnames(marriage) <- c("Answer", "Section")

marriage %>% group_by(Section, Answer) %>% summarise(n = n()) %>%
  mutate(freq = n/sum(n)) %>% ggplot(aes(x = Answer, y = freq)) +
  geom_bar(aes(fill = Answer), alpha = 0.95, stat = "identity") +
  facet_wrap(~Section) + scale_fill_manual(values =
                                        c("dodgerblue", "deeppink2", "goldenrod2")) +
  theme(axis.text.x = element_text(angle = 50, hjust = 1)) + ylab("Frequency")
## `summarise()` regrouping output by 'Section' (override with `.groups` argument)
```



After cleaning up the logic survey data to include the information needed for hypothesis testing, we can visually compare the proportion of correct responses between classes; it appears that nearly identical proportions selected the correct answer 'Yes' between classes, though the breakdown of wrong answers varies.

```
marriage$Answer <-ifelse(marriage$Answer=="Yes", TRUE, FALSE)
prop.11 <-mean(marriage[marriage$Section=="11am",]$Answer)
prop.12 <-mean(marriage[marriage$Section=="12noon",]$Answer)
num.11 <-nrow(marriage[marriage$Section=="11am",])
num.12 <-nrow(marriage[marriage$Section=="12noon",])

prop.overall <- mean(marriage$Answer)
```

```
(prop.12-prop.11)/(sqrt(prop.overall*(1-prop.overall)*((1/num.12)+(1/num.11)))) -> marriage.z
marriage.z
## [1] 0.01850342

pnorm(marriage.z, lower.tail = FALSE) + pnorm(-marriage.z, lower.tail = TRUE)
## [1] 0.9852372
```

We can compare the proportion of correct responses between the classes. We formulate the following hypotheses for the test,

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

After calculating the pooled sample proportion and using it to calculate the z-statistic for the observed difference in sample proportions, we find that the difference is not significant; with a z-score of 0.0185, we obtain a p-value of 0.985 for the bidirectional alternative hypothesis. There is not sufficient evidence to reject the null hypothesis that the proportion of right answers between classes is equal.

Part ii.

We can now compare the overall proportion of correct answers to the previously claimed proportion of 20%. We formulate our hypotheses as

$$H_o : p = 0.20 \ vs. \ H_a : p > 0.20$$

```
null.prop = 0.2
samp.size = nrow(marriage)

(prop.overall - null.prop)/(sqrt((null.prop * (1-null.prop))/(samp.size))) -> logic.z
logic.z
## [1] 5.369246

pnorm(logic.z, lower.tail = F)
## [1] 3.953322e-08
```

We can calculate the z-statistic by using the null proportion and sample variance; this test results in a z-score of 5.369, which corresponds to a right-tailed alternative p-value of 1.08×10^{-8}. We can thus reject the null hypothesis that the population proportion for applied statistics students is equal to 20%, since there is strong evidence to suggest that the population proportion is larger than 20%.

Problem 3.3.

Part i.

We are interested in whether or not pizza and icecream preferences are independent among survey respondants; we first import and clean the data to make later visualizations easier.

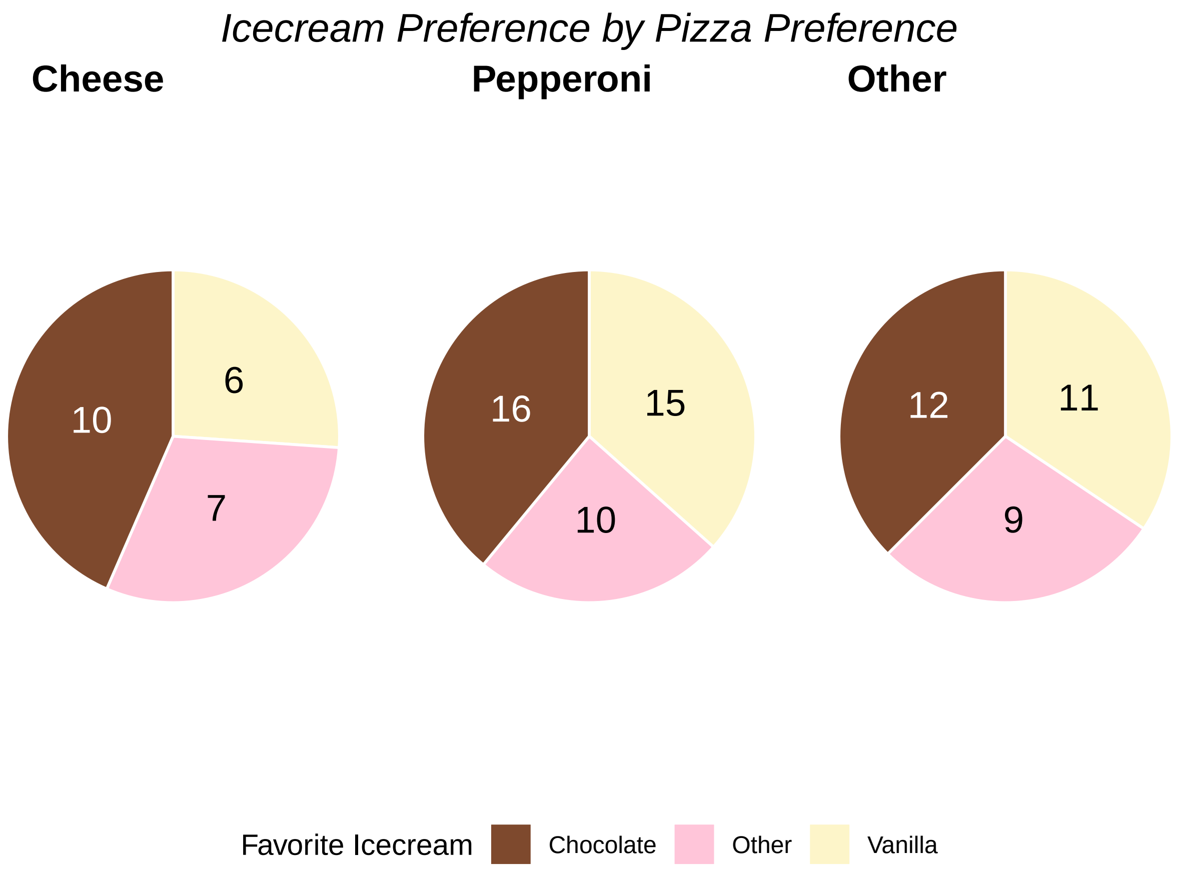
```
pizza.ice <- read.csv("pizza.csv", stringsAsFactors = F)

pizza.ice <- pizza.ice[,2:3] #Removing time-stamp column
colnames(pizza.ice) <- c("Fav.Icecream", "Fav.Pizza") #Shortening column names
pizza.ice$Fav.Icecream <- gsub(" .*$", "", pizza.ice$Fav.Icecream) #Simplifying Icecream Responses
pizza.ice$Fav.Pizza <- gsub(" .*$","", pizza.ice$Fav.Pizza) #Simplifying Pizza Responses

pizza.ice %>% table() -> tab.pizza
tab.pizza %>% addmargins() %>% knitr::kable()
```

	Cheese	Other	Pepperoni	Sum
Chocolate	10	12	16	38
Other	7	9	10	26
Vanilla	6	11	15	32
Sum	23	32	41	96

Part ii.

```
library(tidyverse)
library(ggpubr)
## Warning: package ’ggpubr’ was built under R version 3.6.2
data.frame(tab.pizza) -> pizza.frame

#Separating Responses by Pizza Preference
pizza.frame <- pizza.frame %>% arrange(desc(Fav.Icecream))
pizza.frame.a <- pizza.frame[pizza.frame$Fav.Pizza == "Cheese",]
pizza.frame.b <- pizza.frame[pizza.frame$Fav.Pizza == "Pepperoni",]
pizza.frame.c <- pizza.frame[pizza.frame$Fav.Pizza == "Other",]

#Calculating pie-chart label positions
pizza.frame.a %>% mutate(lab.ypos = cumsum(Freq) - 0.5*Freq) -> pizza.frame.a
pizza.frame.b %>% mutate(lab.ypos = cumsum(Freq) - 0.5*Freq) -> pizza.frame.b
pizza.frame.c %>% mutate(lab.ypos = cumsum(Freq) - 0.5*Freq) -> pizza.frame.c

#Creating individual icecream pizza charts
a <- pizza.frame.a %>% ggplot(aes(x ="", y= Freq, fill = Fav.Icecream)) +
  geom_bar(stat = "identity", width =1, color = "white") +
  coord_polar("y", start = 0) +scale_fill_manual(values =
                                                   c("#7e492d","#ffc5d9", "#fdf5c9" )) +
  theme_void() + geom_text(aes(y = lab.ypos, label = Freq), color =
                             c("black", "black", "white"), size = 5) +
  labs(fill = "Favorite Icecream")

b <- pizza.frame.b %>% ggplot(aes(x ="", y= Freq, fill = Fav.Icecream)) +
  geom_bar(stat = "identity", width =1, color = "white") +
  coord_polar("y", start = 0) +scale_fill_manual(values =
                                                   c("#7e492d","#ffc5d9", "#fdf5c9" )) +
  theme_void() + geom_text(aes(y = lab.ypos, label = Freq), color =
                             c("black", "black", "white"), size = 5) +
  labs(fill = "Favorite Icecream")

c <- pizza.frame.c %>% ggplot(aes(x ="", y= Freq, fill = Fav.Icecream)) +
  geom_bar(stat = "identity", width =1, color = "white") +
  coord_polar("y", start = 0)+scale_fill_manual(values =
                                                  c("#7e492d","#ffc5d9", "#fdf5c9" )) +
  theme_void() + geom_text(aes(y = lab.ypos, label = Freq), color =
                             c("black", "black", "white"), size = 5) +
  labs(fill = "Favorite Icecream")

#Merging and Arranging individual charts
ggarrange(a,b,c, labels = c("Cheese", "Pepperoni", "Other"), ncol= 3, common.legend = T,
          legend = "bottom") -> icecream_pizzas
```

```r
annotate_figure(icecream_pizzas,
                top = text_grob("Icecream Preference by Pizza Preference",
                                face = "italic", size = 15)) -> icecream_pizzas
icecream_pizzas
```



While icecream preferences seem to be somewhat equivalent across pizza preferences, we can notice that total counts for some pizza preferences are higher than others; this indicates that the food preferences may be independent, which we can test by the χ^2 test.

Part iii.

We are interested in testing for association between icecream and pizza preferences; we formulate the hypotheses as

H_o : Icecream and Pizza preferences are independent $vs.$ H_a : Icecream and Pizza preferences are dependent

```r
chisq.test(tab.pizza)
##
##  Pearson's Chi-squared test
##
## data:  tab.pizza
## X-squared = 0.84729, df = 4, p-value = 0.932
```

The χ^2 test for association returns a statistic of 0.847 on the pizza-icecream preference data; this corresponds to a p-value of 0.932 for the χ^2 distribution with 4 degrees of freedom, indicating that we cannot reject the null hypothesis that pizza and icecream preferences are independent.