Keck_STAT630_HW5

Brandon Keck

2024-10-23

1)

```
# install.packages("MASS")
library(MASS)
library(tidyverse)
library(ggplot2)
data("survey")
```

(a)

```
survey <- survey %>%
  mutate(no_smoke = ifelse(Smoke == "Never", "No", "Yes"))
  # Create a new variable called 'no_smoke'

table(survey$no_smoke) # Check to see if we get yes no
```

```
##
##  No Yes
## 189  47
```

```
table(survey$Sex, survey$no_smoke)
```

```
##
##          No Yes
##   Female 99  19
##   Male   89  28
```

(b)

Hypotheses Test

$$H_0: The\ proportion\ of\ non-smokers\ is\ the\ same\ for\ females\ and\ males$$

$$p_f = p_m$$

$$H_A: The\ proportion\ of\ non-smokers\ is\ different\ between\ females\ and\ males$$

$$p_f \neq p_m$$

Checking the conditions:

1. Independent observations and independent groups:
2.

```
n1 <- 118 # female sample size
n2 <- 118 # male sample size

pf <- 99/118 # proportion of female non-smokers
pm <- 89/118 # proportion of male non-smokers


n1*pf
```

```
## [1] 99
```

```
n2*(1-pf)
```

```
## [1] 19
```

```
n2*pm
```

```
## [1] 89
```

```
n2*(1-pm)
```

```
## [1] 29
```

```
se <- sqrt((pf*(1-pf)/n1) + (pm*(1-pm)/n2))

z <- (pf - pm - 0) / se

2 *(pnorm(-abs(z)))
```

```
## [1] 0.1039053
```

Decision: Fail to reject because our p-value is 0.1039053 which is higher than our threshold of 0.05.

We don't have enough evidence to conclude that the proportion of female smokers is different than the true proportion of male non-smokers.

2

```
# install.packages('openintro')
library(openintro)
data("mariokart")
```

Hypotheses Test:

$$H_0: \mu_{new} = \mu_{used}$$

The mean price of new games is the same as the mean price of used games

$$H_A: \mu_{new} \neq \mu_{used}$$

The mean price of new games is different from the mean price of used games

(b)

Check the conditions: 1. Independence: The prices of new and used games do not affect one anther.

check conditions

```
# Pull data for new games
price_new <- mariokart %>%
  filter(cond == "new", !is.na(total_pr)) %>%
  dplyr::select(total_pr) %>%
  pull()

# Pull data for used games
price_used <- mariokart %>%
  filter(cond == "used", !is.na(total_pr)) %>%
  dplyr::select(total_pr) %>%
  pull()

# Sample sizes
n1 <- length(price_new)   # Sample size for new games
n2 <- length(price_used)  # Sample size for used games

# Display the sample sizes
n1
```

```
## [1] 59
```

```
n2
```

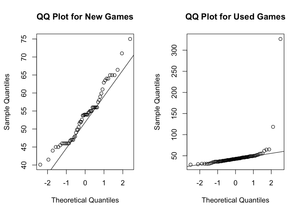
```
## [1] 84
```

```
mariokart %>%
  filter(!is.na(total_pr)) %>%
  ggplot(aes(x = total_pr, fill = cond)) +
  geom_histogram(binwidth = 10, col = "white", alpha = 0.7, position = "identity") +
  labs(x = "Total Price", y = "Frequency", title = "Distribution of Prices by Condition") +
  theme_light() +
  facet_wrap(~cond) +
  xlim(0, 150)  # Set the limit for the x-axis to focus on the relevant range
```


```
par(mfrow = c(1,2))

qqnorm(price_new, main = "QQ Plot for New Games")
qqline(price_new)


qqnorm(price_used, main = "QQ Plot for Used Games")
qqline(price_used)
```



(c)

I believe that we should not remove these outliers because they represent legitimate market transactions for different types of products such as the Wii bundle Guitar Hero and the 10 Wii games.

(d)

Remove the outliers

```
mariokart_full <- mariokart
# remove the outliers
# Remove outliers with total_pr greater than 100
mariokart <- mariokart_full %>%
  filter(total_pr <= 100)
```

Calculate a confidence interval

```
t.test(price_new, price_used, var.equal = FALSE, conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  price_new and price_used
## t = 1.7893, df = 94.902, p-value = 0.07676
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.725468 13.970633
## sample estimates:
## mean of x mean of y
##  53.77068  47.14810
```

Conclusion: We do not have enough evidence that the true differences of the mean new price compared to used price is different from 0.

3.

```
n1 <- 1200 # sample size n = 1200
sample_a <- 600 # sample size of group a
sample_b <- 600 # sample size of group b

prop_a <- 200/600  # proportion infected after vaccine a
prop_b <- 150/600 # proportion infected after vaccine b
```

$$H_0: p_a = p_b$$

$$H_A: p_a \neq p_b$$

Conditions: Independence: Groups A and B are independent of each other and each participant is independent.

```
sample_a * prop_a; sample_a*(1-prop_a)
```

```
## [1] 200
```

```
## [1] 400
```

```
sample_b * prop_b; sample_b*(1-prop_b)
```

```
## [1] 150
```

```
## [1] 450
```

```
se <- sqrt((prop_a*(1-prop_a)/sample_a) + (prop_b*(1-prop_b)/sample_b))

z <- (prop_a - prop_b - 0)/se

2*pnorm(-abs(z))
```

```
## [1] 0.001427836
```

Decision:

(b)

(c)

```
sample_a <- 48
sample_b <- 48

se <- sqrt((prop_a*(1-prop_a)/sample_a) + (prop_b*(1-prop_b)/sample_b))
z <- (prop_a - prop_b - 0)/se

2*pnorm(-abs(z))
```

```
## [1] 0.36707
```

Decision:

(d)

4.

5.

E-