Keck_STAT630_HW5

Brandon Keck

2024-10-25

1)

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

```
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
```

(a)

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

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

```
##
##           No Yes Sum
##   Female  99  19 118
##   Male    89  28 117
##   Sum    188  47 235
```

(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: The smoking status of the female groups has no effect on the smoking status of that of the male group.
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
```

Since all of these values are greater than 5, this satisfies both conditions for difference of proportions.

```
se <- sqrt((pf*(1-pf)/n1) + (pm*(1-pm)/n2)) # hand calculate standard error

z <- (pf - pm - 0) / se # hand calculate test statistic

2 *(pnorm(-abs(z))) # hand calculate p-value
```

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

Decision: Fail to reject because our p-value is 0.1039053 which is higher than our threshold of 0.05. Therefore, we are in favor of the H_0

Conclusion: 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")
```

Hyoptheses Test:

$$\textit{The mean price of new games is the same as the mean price of used games}$$

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

$$\textit{The mean price of new games is different from the mean price of used games}$$

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

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

```
# 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
```

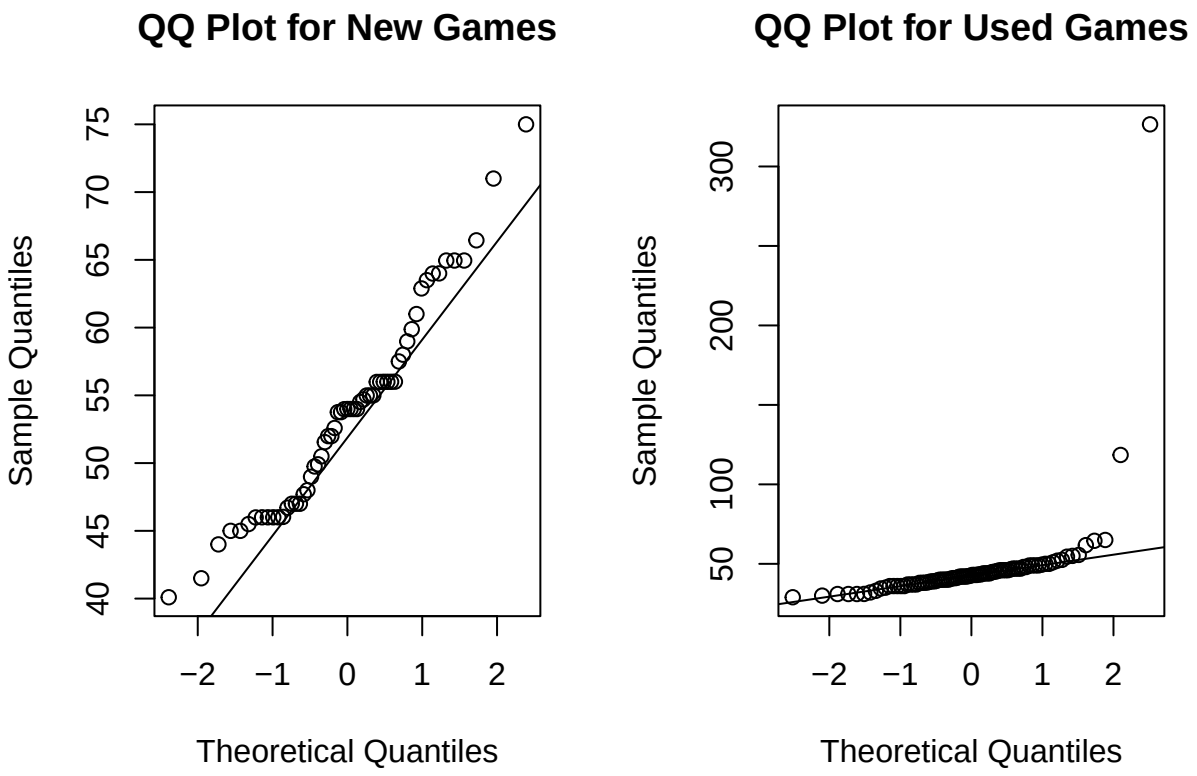
Since both of these values are greater than 30 this satisfies the conditions for a difference of means.

```r
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)
```


```r
par(mfrow = c(1,2)) # Two graphs on 1 grid

qqnorm(price_new, main = "QQ Plot for New Games") # Q-Q Plot for New Games
qqline(price_new) # Adding a line to Q-Q plot

qqnorm(price_used, main = "QQ Plot for Used Games") # Q-Q Plot for Used Games
qqline(price_used) # Adding a line to Q-Q plot
```



(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. These just are not single Wii games such as the rest of the dataset.

(d)

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

Remove the outliers Rechecking the conditions in order to see if we removed the outliers

```
# Pull data for new and used games from the filtered dataset
price_new <- mariokart_full %>%
  filter(cond == "new") %>%
  dplyr::select(total_pr) %>%
  pull()

price_used <- mariokart_full %>%
  filter(cond == "used") %>%
  dplyr::select(total_pr) %>%
  pull()
```

```r
# Sample sizes
n_new <- length(price_new)
n_used <- length(price_used)

n_new
```

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

```r
n_used
```

```
## [1] 82
```

We in fact removed the outliers. Before used games had a total of 84 products under totoal_pr, we now have a total of 82 products under total_pr which means we removed the two outliers.

```r
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 = 8.6406, df = 123.99, p-value = 2.349e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   8.402838 13.396322
## sample estimates:
## mean of x mean of y
##  53.77068  42.87110
```

Decision:Since our p-value is essentially 0 there is strong evidence against the null hypothesis indicating that we are in favor of the alternative hypothesis.

Conclusion: We have enough evidence to conclude that the true difference in mean prices between new and used Mario Kart games is significantly different from 0.

3.

```r
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
```

Hypthesis:

$$H_0 : p_a = p_b$$

$$H_A : p_a \neq p_b$$

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
```

Since these values are greater than 5 the conditions for a difference of proportions has been satisfied.

```
se <- sqrt((prop_a*(1-prop_a)/sample_a) + (prop_b*(1-prop_b)/sample_b))
# Hand calculate the standard error

z <- (prop_a - prop_b - 0)/se # hand calculate the test statistic

2*pnorm(-abs(z)) # Hand calculate p-value
```

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

Decision:Our p-value of 0.001427836 is less than the significance level of $\alpha = 0.1$, we reject the H_0 : in favor of H_A :

Conclusion: From our calculations we can conclude that there is significant evidence that the true proportion of people who get infected from vaccine A is different that the proportion of people who get infected from vaccine B.

(b) In part (a) we performed a hypothesis test and found a statistically significant result with a p-value of 0.001427836, which is lower than the significance level of 0.01. This indicates that there is strong evidence to reject the null hypothesis. However, statistical significance simply means that the observed difference is unlikely to have occured by chance, while practical importance refers to whether the difference is large enought to matter in real-world applications.

```
sample_a <- 48 # group a sample size
sample_b <- 48 # group b sample size

se <- sqrt((prop_a*(1-prop_a)/sample_a) + (prop_b*(1-prop_b)/sample_b)) # calculate standard error
z <- (prop_a - prop_b - 0)/se # calculate the test statistic

2*pnorm(-abs(z)) # hand calculate the p-value
```

(c)

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

Decision: Our p-value has now changed from 0.001427836 to 0.36707 which is greater than a significance level of $alpha = 0.1$. Therefore, we fail to reject $ H_0$

Conclusion: If the sample sizes of each group were to be the same as we had conducted in this experiment of (48), we would not have sufficient evidence to conclude that the true proportion of people who get infected from vaccine A is any different from the true proportion of people who get infected from vaccine B.

(d) The larger the sample sizes, the more likely we are to get statistically significant results than if we were to have a smaller sample size. We have essentially proven this with our comparison of having sample sizes from parts (a) and parts (c). By having a smaller sample size of 48 our p-value changed from being statistically significant to not being significant at all.

4.

I found removing the outliers to be a bit challenging for me. I was especially struggling with the part of rechecking the conditions to see whether or not the conditions for a difference of means was still met after we had removed the outliers. I'm still not 100% sure if I have done this correctly.

5.

M-Meets Expectation