STAT630 Homework 5

David Teng

by Thursday, October 24th at 9:00pm

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

1. (a)

```
#  Create a binary factor variable 'no_smoke' (Regular vs. Non-regular)
survey2 <- survey %>%
  mutate(no_smoke = as.integer(Smoke == "Regul"))
```

(b)

1.Write the hypotheses.

H_0 : There is no significant difference in the proportion of non-smokers between males and females. H_A : There is a significant difference in the proportion of non-smokers between males and females.

2. Check conditions.

(1) Check the independence condition: Assuming the survey dataset consists of randomly selected observational units

2. Large Counts

```
#  Create a binary factor variable 'non_smoker' (Never vs. Non-never)
survey3 <- survey %>%
  mutate(`non_smoker` = as.integer(Smoke == "Never"))
```

```
survey3 %>%
  filter(Sex=="Male",non_smoker=="1")
survey3 %>%
  filter(Sex=="Female",non_smoker=="1")
```

```
# Not Pooled
n1 = 118 # sample size of male students
x1 = 89 # male students who are non-smokers
phat1 <- x1 / n1

n2 = 118 # sample size of female students
x2 = 99 # female students who are non-smokers
phat2 <- x2 / n2

cat("# Not Pooled:\n")
```

```
## # Not Pooled:
```

```r
cat("n1 * phat1:", n1 * phat1, "\n")
```

```
## n1 * phat1: 89
```

```r
cat("n1 * (1 - phat1):", n1 * (1 - phat1), "\n")
```

```
## n1 * (1 - phat1): 29
```

```r
cat("n2 * phat2:", n2 * phat2, "\n")
```

```
## n2 * phat2: 99
```

```r
cat("n2 * (1 - phat2):", n2 * (1 - phat2), "\n\n")
```

```
## n2 * (1 - phat2): 19
```

```r
# Pooled
phat_pooled <- (phat1*n1 + phat2*n2)/ (n1 + n2)

cat("#  Pooled:\n")
```

```
## #  Pooled:
```

```r
cat("n1*phat_pooled:", n1*phat_pooled, "\n")
```

```
## n1*phat_pooled: 94
```

```r
cat("n1*(1-phat_pooled):", n1*(1-phat_pooled), "\n")
```

```
## n1*(1-phat_pooled): 24
```

```r
cat("n2*phat_pooled:", n2*phat_pooled, "\n")
```

```
## n2*phat_pooled: 94
```

```r
cat("n2*(1-phat_pooled):", n2*(1-phat_pooled), "\n\n")
```

```
## n2*(1-phat_pooled): 24
```

```r
# Check conditions
  condition1 <- n1 * phat_pooled >= 5
  condition2 <- n1 * (1 - phat_pooled) >= 5
  condition3 <- n2 * phat_pooled >= 5
  condition4 <- n2 * (1 - phat_pooled) >= 5
  condition1
```

```
## [1] TRUE
```

```r
  condition2
```

```
## [1] TRUE
```

```r
  condition3
```

```
## [1] TRUE
```

```r
  condition4
```

```
## [1] TRUE
```

3. Calculate test statistic.

```
est <- phat1 - phat2

# Not pooled
se_phats <- sqrt((phat1*(1-phat1))/n1 + (phat2*(1-phat2))/n2)
se_phats
```

```
## [1] 0.05211248
```

```
# Pooled
se_pooled <- sqrt( (phat_pooled * (1 - phat_pooled)) * (1/n1 + 1/n2) )
se_pooled
```

```
## [1] 0.05240365
```

```
# Test statistic not pooled
z_stat <- (est - 0) / se_phats
z_stat
```

```
## [1] -1.626208
```

```
# Test statistic pooled
z_pooled <- (est - 0) / se_pooled
z_pooled
```

```
## [1] -1.617173
```

4. Calculate p-value.

```
prop_test <- prop.test(c(89, 99), c(118, 118), alternative = "two.sided")
prop_test
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(89, 99) out of c(118, 118)
## X-squared = 2.1184, df = 1, p-value = 0.1455
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1953589  0.0258674
## sample estimates:
##    prop 1    prop 2
## 0.7542373 0.8389831
```

```
p_value <- prop_test$p.value
p_value
```

```
## [1] 0.1455432
```

5. Make a decision and conclude in the context of the problem.

```
alpha <- 0.05
decision <- if (p_value < alpha) {
  "Reject the null hypothesis."
} else {
  "Fail to reject the null hypothesis."
}
decision
```

```
## [1] "Fail to reject the null hypothesis."
```

Decision: Fail to reject H_0 Conclusion: We do not have enough evidence that there is a significant difference in the proportion of non-smokers between males and females.

2. (a)

```
library(openintro)
data("mariokart")

mariokart_new <- mariokart%>%
filter(cond=="new")
mariokart_used <- mariokart%>%
filter(cond=="used")
```

1.Write the hypotheses.

$H_0 : \mu_{\text{new}} - \mu_{\text{used}} = 0$ There is no significant difference in the price of new versus used Mario Kart games for Nintendo Wii's

$H_A : \mu_{\text{new}} - \mu_{\text{used}} \neq 0$ There is a significant difference in the price of new versus used Mario Kart games for Nintendo Wii's

(b)

2. Check conditions.

(1) Check the independence condition: Assuming the survey dataset consists of randomly selected observational units
(2) Normality

```
n1 <- nrow (mariokart_new)
n2 <- nrow (mariokart_used)
cat( "check if n1 > 30?", n1>30, "\n")
```

```
## check if n1 > 30? TRUE
```

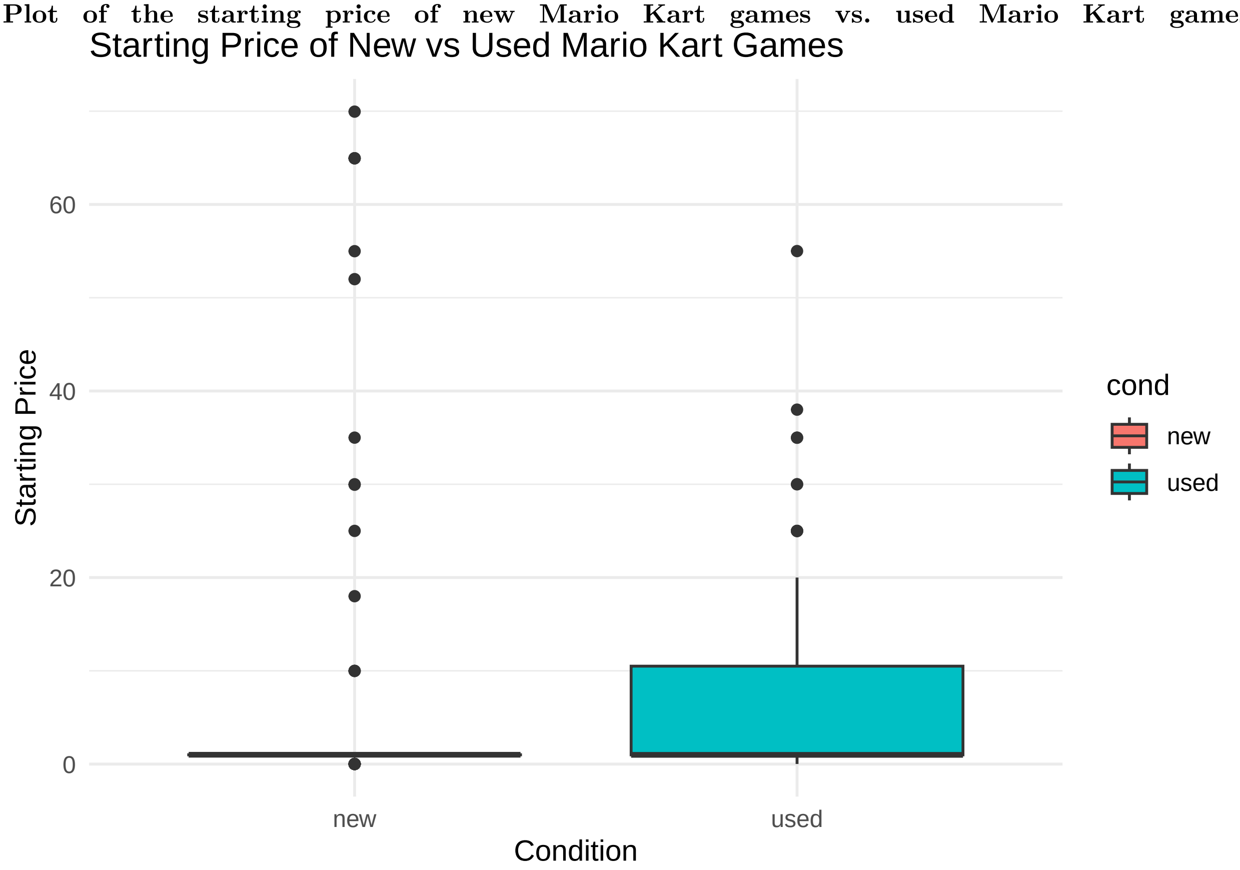
```
cat( "check if n2 > 30?", n2>30, "\n")
```

```
## check if n2 > 30? TRUE
```

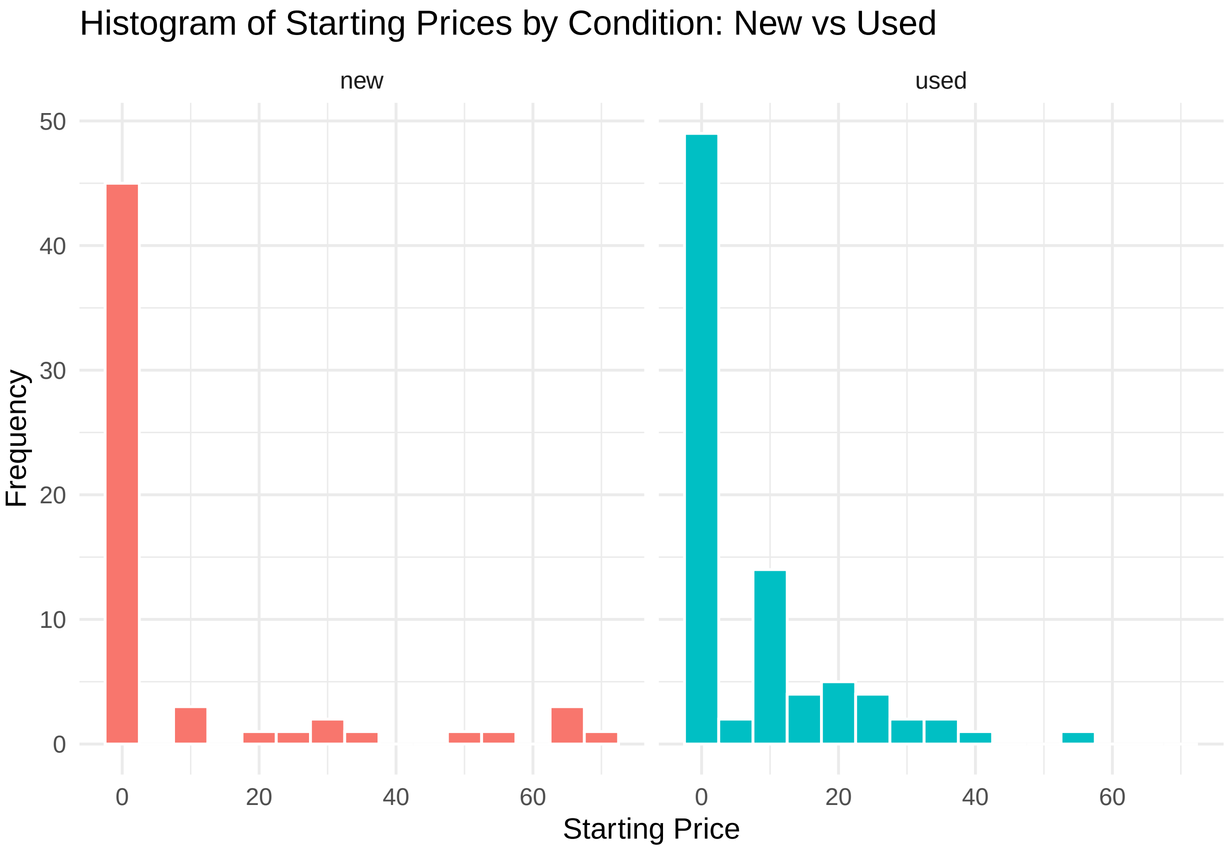
```

#Check the Normality via boxplots
mariokart %>%
  ggplot(aes(x = cond, y = start_pr, fill = cond)) +
  geom_boxplot() +
  labs(title = "Starting Price of New vs Used Mario Kart Games",
       x = "Condition",
       y = "Starting Price") +
  theme_minimal()
```

Plot of the starting price of new Mario Kart games vs. used Mario Kart games



```r
#Check the Normality via histograms
mariokart %>%
  ggplot(aes(start_pr, fill = cond)) +
  geom_histogram(binwidth = 5, col = "white", show.legend = FALSE) +
  facet_wrap(~ cond) +
  labs(title = "Histogram of Starting Prices by Condition: New vs Used",
       x = "Starting Price",
       y = "Frequency") +
  theme_minimal()
```



### The normality condition of Starting Price of New vs Used Mario Kart Games is not satisfied because of outliers.

(c)

```
# Function to investigate outliers
find_outlier <- function(x) {
  Q1 <- quantile(x, 0.25)
  Q3 <- quantile(x, 0.75)
  IQR <- Q3 - Q1
  x < (Q1 - 1.5 * IQR) | x > (Q3 + 1.5 * IQR)
}
# Identify outliers in start_pr grouped by condition
outliers <- mariokart %>%
  group_by(cond) %>%
  filter(find_outlier(start_pr))
outliers
```

```
## # A tibble: 35 x 12
## # Groups:   cond [2]
##            id duration n_bids cond  start_pr ship_pr total_pr ship_sp seller_rate
##         <dbl>    <int>  <int> <fct>    <dbl>   <dbl>    <dbl> <fct>         <int>
##  1   1.70e11        1     20 new       0.01     0        71   media           820
##  2   1.80e11        1     15 new       0.01     0        56.0 media           820
##  3   1.80e11        1     16 new       0.01     0        56   media           820
##  4   2.51e11        7      6 used     25.0      4        43.3 standa~         154
##  5   3.50e11        1      1 new      65.0      0        65.0 standa~      118345
##  6   1.80e11        1     19 new       0.01     0        55   media           820
##  7   3.20e11        7     15 new       9.99     4        47   parcel           62
## 8   2.60e11        7      5 new      30       0        46.0 firstC~          555
## 9   3.30e11        3      2 new      52.0     0        52.0 other         223861
## 10  2.90e11        5      7 used     25       2.99     48   priori~          239
## # i 25 more rows
## # i 3 more variables: stock_photo <fct>, wheels <int>, title <fct>
```

Comment on outliers and implications of removing them:

Removing outliers helps achieve normality, which is essential for many hypothesis tests that assume data follows a normal distribution. Outliers can skew the results, making the test statistics unreliable and leading to incorrect conclusions. Ensuring normality improves the accuracy and validity of the hypothesis test!

(d)

```
# Remove outliers above and create a cleaned dataset
cleaned_mariokart <- mariokart %>%
  group_by(cond) %>%
  filter(!find_outlier(start_pr))
```

```
t_test <- t.test(start_pr ~ cond, cleaned_mariokart)
t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  start_pr by cond
## t = -5.4208, df = 73, p-value = 7.314e-07
## alternative hypothesis: true difference in means between group new and group used is not equal to 0
## 95 percent confidence interval:
##  -5.261961 -2.432888
## sample estimates:
##  mean in group new mean in group used
##          0.9967647          4.8441892
```

```
ci <- t_test$conf.int
cat("the confidence interval after removing the outliers is",ci)
```

```
## the confidence interval after removing the outliers is -5.261961 -2.432888
```

(e)

```
H_0 <- 0
# Conclusion based on confidence interval
decision <- if (ci[1] > H_0 | ci[2] < H_0) {
  "We reject the null hypothesis."
} else {
  "We fail to reject the null hypothesis."
}
decision
```

```
## [1] "We reject the null hypothesis."
```

Decision: Reject H_0 Conclusion: We have enough evidence that there is a significant difference in the price of new versus used Mario Kart games for Nintendo Wii's.

3. (a)

1.Write the hypotheses.

$H_0 : p_A - p_B = 0$ (There is no significant difference in infection rates between Vaccine A and B) $H_A : p_A - p_B \neq 0$ (There is a significant difference in infection rates between Vaccine A and B)

2. Check conditions.

1. Check the independence condition: Assuming the data is randomly sampled and each observation is independent
2. Check the large counts condition pooled option:

```
# Not Pooled
n1 = 600 # sample size for participants who received Vaccine A
x1 = 200 # participants who received Vaccine A were infected
phat1 <- x1 / n1

n2 = 600 # sample size for participants who received Vaccine B
x2 = 150 # participants who received Vaccine B were infected
phat2 <- x2 / n2

cat("# Not Pooled:\n")
```

```
## # Not Pooled:
```

```
cat("n1 * phat1:", n1 * phat1, "\n")
```

```
## n1 * phat1: 200
```

```
cat("n1 * (1 - phat1):", n1 * (1 - phat1), "\n")
```

```
## n1 * (1 - phat1): 400
```

```
cat("n2 * phat2:", n2 * phat2, "\n")
```

```
## n2 * phat2: 150
```

```
cat("n2 * (1 - phat2):", n2 * (1 - phat2), "\n\n")
```

```
## n2 * (1 - phat2): 450
```

```
# Pooled
phat_pooled <- (phat1*n1 + phat2*n2)/ (n1 + n2)

cat("#  Pooled:\n")
```

```
## #  Pooled:
```

```
cat("n1*phat_pooled:", n1*phat_pooled, "\n")
```

```
## n1*phat_pooled: 175
```

```
cat("n1*(1-phat_pooled):", n1*(1-phat_pooled), "\n")
```

```
## n1*(1-phat_pooled): 425
```

```
cat("n2*phat_pooled:", n2*phat_pooled, "\n")
```

```
## n2*phat_pooled: 175
```

```
cat("n2*(1-phat_pooled):", n2*(1-phat_pooled), "\n\n")
```

```
## n2*(1-phat_pooled): 425
```

```r
# Check conditions
  condition1 <- n1 * phat_pooled >= 5
  condition2 <- n1 * (1 - phat_pooled) >= 5
  condition3 <- n2 * phat_pooled >= 5
  condition4 <- n2 * (1 - phat_pooled) >= 5
  condition1
```

```
## [1] TRUE
```

```r
  condition2
```

```
## [1] TRUE
```

```r
  condition3
```

```
## [1] TRUE
```

```r
  condition4
```

```
## [1] TRUE
```

3. Calculate test statistic.

```r
est <- phat1 - phat2

# Not pooled
se_phats <- sqrt((phat1*(1-phat1))/n1 + (phat2*(1-phat2))/n2)
se_phats
```

```
## [1] 0.02613179
```

```r
# Pooled
se_pooled <- sqrt( (phat_pooled * (1 - phat_pooled)) * (1/n1 + 1/n2) )
se_pooled
```

```
## [1] 0.02624228
```

```r
# Test statistic not pooled
z_stat <- (est - 0) / se_phats
z_stat
```

```
## [1] 3.188964
```

```r
# Test statistic pooled
z_pooled <- (est - 0) / se_pooled
z_pooled
```

```
## [1] 3.175537
```

4. Calculate p-value.

```r
# Calculate the p-value for the two-tailed test
p_value <- 2 * pnorm(abs(z_pooled), lower.tail = FALSE)
p_value
```

```
## [1] 0.001495596
```

5. Make a decision and conclude in the context of the problem.

```
# Decision based on p-value and significance level
alpha <- 0.01
decision <- if (p_value < alpha) {
  "Reject the null hypothesis."
} else {
  "Fail to reject the null hypothesis."
}
decision
```

```
## [1] "Reject the null hypothesis."
```

Decision: Reject H_0 Conclusion: We have enough evidence that there is a significant difference in infection rates between Vaccine A and Vaccine B at the 1% significance level.

(b)

```
infection_rate_difference <- abs(phat1 - phat2)
cat("The difference in infection rates between the two vaccines is", infection_rate_difference, "\n")
```

```
## The difference in infection rates between the two vaccines is 0.08333333
```

Discussion on Statistical Significance vs Practical Importance

Statistical significance tells us if the difference in infection rates is likely due to chance. A p-value < 0.01 shows strong evidence that the two vaccines differ. Since the difference in infection rates between the two vaccines is 0.083 which is > 0.05, this difference might be of practical importance, indicating that one vaccine might be more effective in real-world applications and could influence clinical decisions.

(c)

```
# Original proportions
phat1 <- 200 / 600
phat2 <- 150 / 600

# New sample size
n_A_new <- 48 # new sample size for participants who received Vaccine A
n_B_new <- 48 # new sample size for participants who received Vaccine B

# Same proportions but with smaller sample sizes
infected_A_new <- round(n_A_new * phat1)
infected_B_new <- round(n_B_new * phat2)

prop_test_new <- prop.test(c(infected_A_new, infected_B_new), c(n_A_new, n_B_new), alternative = "two.s
prop_test_new
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  c(infected_A_new, infected_B_new) out of c(n_A_new, n_B_new)
## X-squared = 0.45378, df = 1, p-value = 0.5005
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1185807  0.2852474
## sample estimates:
##    prop 1    prop 2
## 0.3333333 0.2500000
```

```r
p_new <-prop_test_new$p.value
cat("New P-value:",p_new)
```

```
## New P-value: 0.5005446
```

```r
if (p_new < alpha) {
  cat("With smaller sample size, we reject the null hypothesis: There is a significant difference.\n")
} else {
  cat("With smaller sample size, we fail to reject the null hypothesis: There is no significant differe
}
```

```
## With smaller sample size, we fail to reject the null hypothesis: There is no significant difference.
```

Decision: Fail to reject H_0 Conclusion: We do not have enough evidence that there is a significant difference in infection rates between Vaccine A and Vaccine B at the 1% significance level.

(d)

Comparison of Results:

Original Sample (1,200 participants): Z Statistic: 3.1755 P-value: 0.0015 (significant at the 1% level)

New Sample (48 participants): Z Statistic: 0.8982 P-value: 0.3691 (not significant at the 1% level)

Conclusion: In the smaller sample size scenario (48 participants), the test couldn't reject the null hypothesis, even with the same infection proportions, due to higher variability and reduced power. This highlights the importance of sample size in interpreting results, as it significantly affects statistical significance and practical implications.

4. I found checking the normality condition difficult but it's very important.

5. E - Excellent