HW7 - Statistical Inference

Madhu Peduri

July 4, 2021

0.0.1 1.Analyzing Data with a Categorical Outcome.

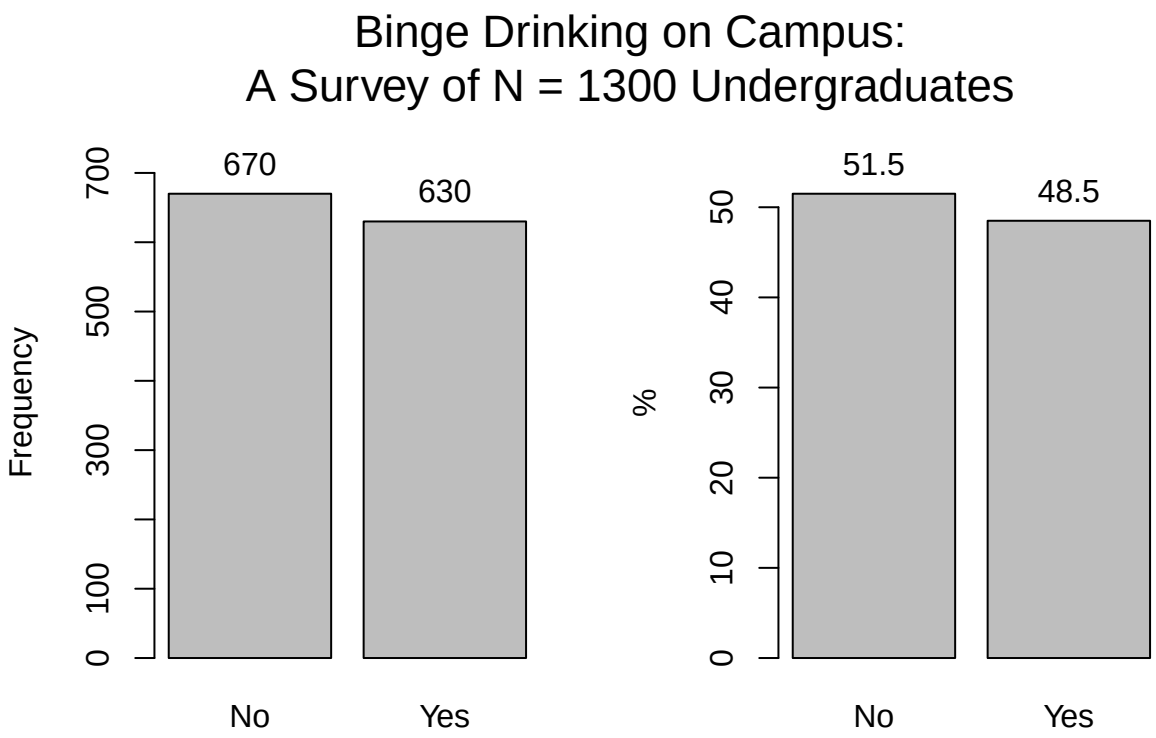
```
cat <- c(rep("Yes", 630), rep("No", 670))
vec <- factor(cat)
binge <- data.frame(vec)
names(binge) <- c("BingeDrink")
head(binge)
```

0.0.1.1 1.1 (a) Enter the data into R in an expanded form

```
##   BingeDrink
## 1        Yes
## 2        Yes
## 3        Yes
## 4        Yes
## 5        Yes
## 6        Yes
```

```
par(mfrow = c(1, 2))
t1 <- tab1(binge, main = "")
t2 <- tab1(binge, bar.values = "percent", main = "")
mtext("Binge Drinking on Campus:", side = 3, line = 2.4, at = -0.8,
    cex = 1.4)
mtext("A Survey of N = 1300 Undergraduates", line = 1, side = 3,
    at = -0.8, cex = 1.4)
```

0.0.1.2 1.1 (b) Frequency charts




```
prop.test(x = length(binge$BingeDrink[binge$BingeDrink == "Yes"]),
    n = length(binge$BingeDrink), p = 0.5)
```

0.0.1.3 1.2 (a) Prop & Binomial test

```
##
##  1-sample proportions test with continuity correction
##
## data:  length(binge$BingeDrink[binge$BingeDrink == "Yes"]) out of length(binge$BingeDrink), null prob
## X-squared = 1, df = 1, p-value = 0.3
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.457 0.512
## sample estimates:
##     p
## 0.485
```

```
binom.test(x = length(binge$BingeDrink[binge$BingeDrink == "Yes"]),
    n = length(binge$BingeDrink), p = 0.5)
```

```
##
##  Exact binomial test
##
## data:  length(binge$BingeDrink[binge$BingeDrink == "Yes"]) and length(binge$BingeDrink)
## number of successes = 630, number of trials = 1300, p-value = 0.3
```

```
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.457 0.512
## sample estimates:
## probability of success
##                  0.485
```

0.0.1.4 Observations:

- A binomial test compares a sample proportion to a hypothesized proportion. The test has the following null and alternative hypothesis.
- We have 630 number of successes in an experiment with total number of 1300 trials with a probability of 1/2 for a success for a given trial.
- We obtained a p-value = 0.3 (> 0.05) in our binomial test, which suggests that we can accept the null hypothesis and conclude that we got 'Yes' in our experiment.

```
# When probability of happeining Yes and No are equal to
# 1/2
prob <- c(length(binge$BingeDrink[binge$BingeDrink == "Yes"]),
    length(binge$BingeDrink[binge$BingeDrink == "No"]))
chisq.test(prob, p = c(1/2, 1/2))
```

0.0.1.5 1.2 (b)(i) Chi-square goodness

```
##
##  Chi-squared test for given probabilities
##
## data:  prob
## X-squared = 1, df = 1, p-value = 0.3
```

```
chisq.test(prob, p = c(1/3, 2/3))
```

```
##
##  Chi-squared test for given probabilities
##
## data:  prob
## X-squared = 134, df = 1, p-value <2e-16
```

0.0.1.6 Observations:

- The chi-square goodness of fit test is used to compare the observed distribution to an expected distribution, in a situation where we have two or more categories in a discrete data. In other words, it compares multiple observed proportions to expected probabilities.
- For the hypothesis $H_0 : p_1 = p_2$, with same probability for both categorical values, we have p-value = 0.3 (> 0.05). This says that, observed proportions are not different from the expected proportions.
- For the hypothesis $H_1 : p_1 \neq p_2$, with different probability for both categorical values, we have p-value very less. This says that, observed proportions are different from the expected proportions.

```
pi = 0.247
chisq.test(prob, p = c(pi, 1 - pi))
```

0.0.1.7 1.2 (b)(ii)

```
##
##  Chi-squared test for given probabilities
```

```
##
## data:  prob
## X-squared = 395, df = 1, p-value <2e-16
```

0.0.1.8 Observations:

- If we use the proportion = 0.247 by NSDUH as expected hypothesis, we are getting a less p-value. We can say this, our actual proportion do not comply with expected hypothesis.
- We can see 0.247*1300 = 321.1 (< 630) which is less than the actual number of successes.

```
prop.test(x = 630, n = 1300, p = 0.247, alternative = "g")
```

0.0.1.9 1.2 (c)

```
##
##  1-sample proportions test with continuity correction
##
## data:  630 out of 1300, null probability 0.247
## X-squared = 393, df = 1, p-value <2e-16
## alternative hypothesis: true p is greater than 0.247
## 95 percent confidence interval:
##  0.461 1.000
## sample estimates:
##     p
## 0.485
```

```
binom.test(x = 630, n = 1300, p = 0.247, alternative = "g")
```

```
##
##  Exact binomial test
##
## data:  630 and 1300
## number of successes = 630, number of trials = 1300, p-value <2e-16
## alternative hypothesis: true probability of success is greater than 0.247
## 95 percent confidence interval:
##  0.461 1.000
## sample estimates:
## probability of success
##                  0.485
```

0.0.1.10 Observations:

- We can see in both the cases we have p-value is less.
- If we have proportion = 0.5 (>0.247), we have p-value that is sufficient which can be equivalent to hypothesis.

```
sales <- read.csv("sales.csv")
tabs <- xtabs(~Region + Sport, data = sales)
tabs
```

0.0.1.11 1.3(a) Read Sales Data

```
##       Sport
## Region  A  S  W
##      A  9  6 22
```

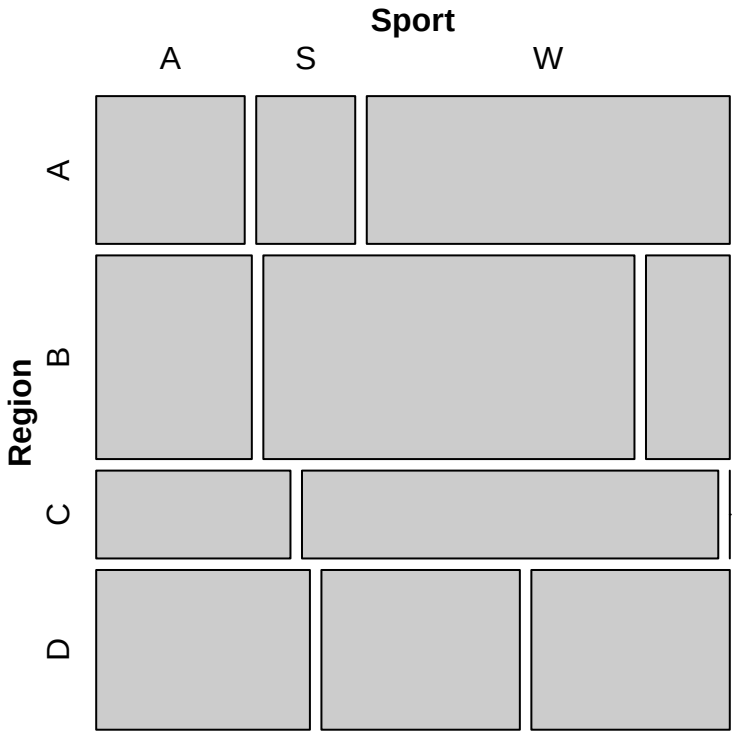
```
##       B 13 31  7
##       C  7 15  0
##       D 14 13 13
```

```
ptab <- prop.table(tabs)
ptab
```

0.0.1.12 1.3(b) Mosaic Plot

```
##       Sport
## Region      A      S      W
##      A 0.0600 0.0400 0.1467
##      B 0.0867 0.2067 0.0467
##      C 0.0467 0.1000 0.0000
##      D 0.0933 0.0867 0.0867
```

```
mosaic(ptab, zero_size = 0)
```



0.0.1.13 Observations:

- We can observe that Mosaic plot shows the equivalent densities.
- We can see that, Region B has highes sales for the Summer sport gear.

```
chisq.test(tabs)
```

0.0.1.14 1.3(c) Chi-square test of independence

```
##
##  Pearson's Chi-squared test
##
## data:  tabs
## X-squared = 38, df = 6, p-value = 9e-07
```

0.0.1.15 Observations:

- We can see very less p-value which shows the less dependency between Sales of sports gear and geographic region.

0.0.2 2. Analysis of Continuous Outcome Data

```
car <- c(19, 26, 24, 21, 24, 23, 26, 24, 23, 20, 21, 24, 18,
    21, 20, 23, 24, 26, 25, 19, 24, 23, 27, 24, 26, 25, 20, 21,
    19, 23)
ttest <- t.test(car, mu = 25)
sprintf("p-value of t-test: %.10f", ttest$p.value)
```

0.0.2.1 2.1(a) One-Sample Tests

```
## [1] "p-value of t-test: 0.0000343022"
```

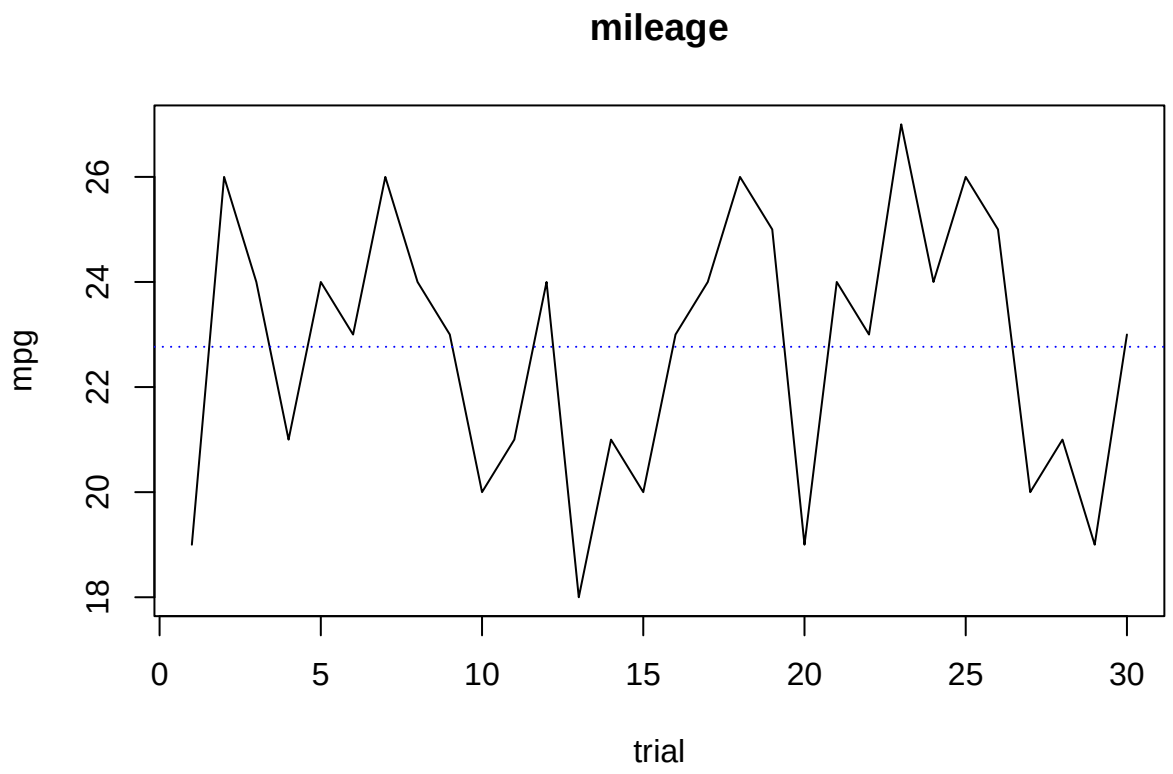
```
sprintf("Mean of the sample: %.2f", mean(car))
```

```
## [1] "Mean of the sample: 22.77"
```

```
sprintf("Standard deviation of the sample: %.2f", sd(car))
```

```
## [1] "Standard deviation of the sample: 2.50"
```

```
plot(car, type = "l", xlab = "trial", ylab = "mpg", main = "mileage")
par(new = TRUE)
abline(h = mean(car), col = "blue", lty = 3)
```

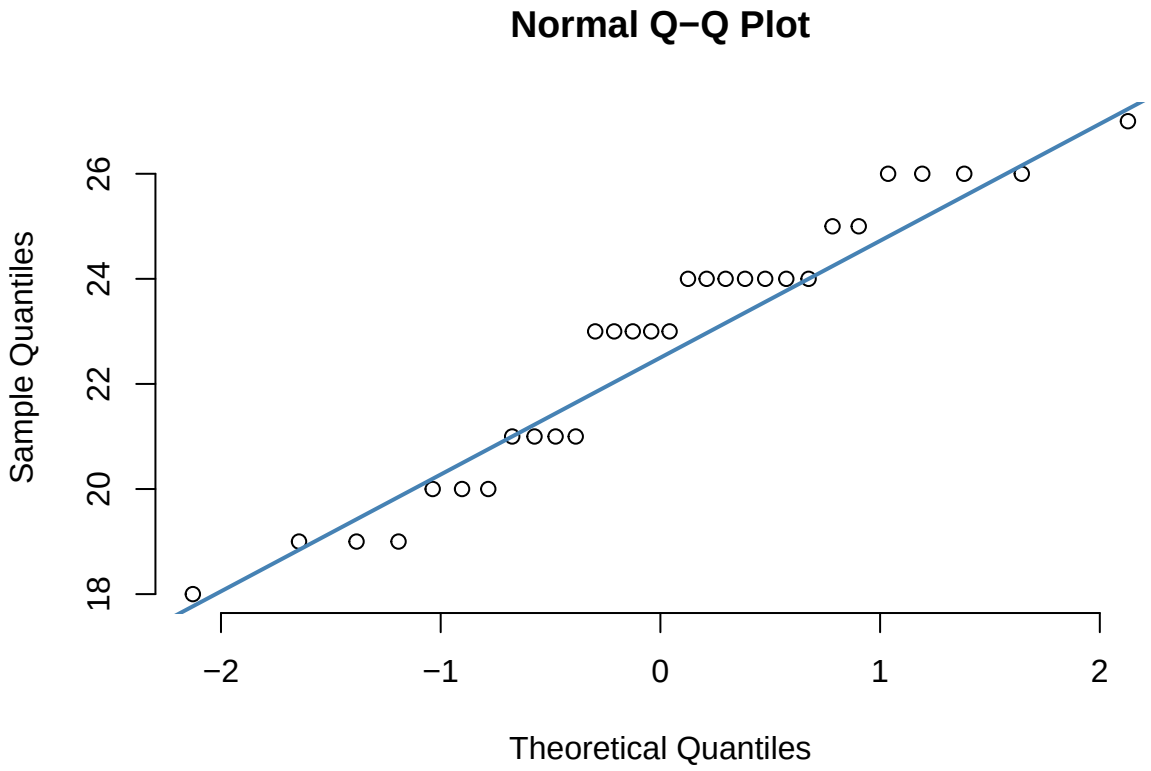


0.0.2.2 Observations:

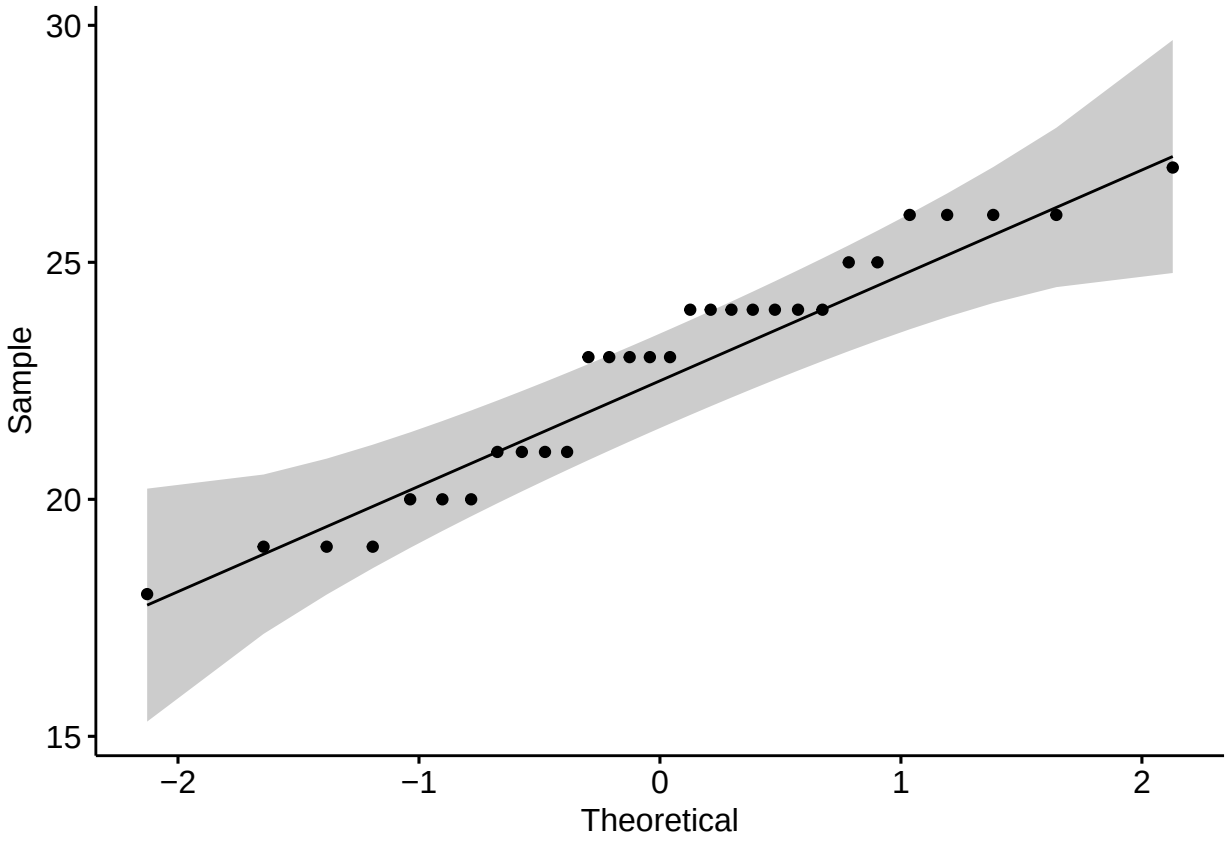
- We can see p-value of our t-test is less (inferior) to the significant value 0.05. This suggest that null hypothesis is incorrect and the hypothesis $H_1 : \mu \neq 25$ is correct.

```
qqnorm(car, pch = 1, frame = FALSE)
qqline(car, col = "steelblue", lwd = 2)
```

0.0.2.3 2.1(b) Shaprio test




```
ggqqplot(car)
```




```
shapirotest <- shapiro.test(car)
sprintf("P-value of Shapiro normality test:%.4f", shapirotest$p.value)
```

```
## [1] "P-value of Shapiro normality test:0.0928"
```

0.0.2.4 Observations:

- From qqplots, we can see the sample quantiles follow a straight line and fall with in the range.
- The p-value of shapiro.test is 0.09 which is greater than significant value 0.05 suggests that our data follow normal distribution.

```
ttest <- t.test(car, alternative = "less", mu = 25)
ttest
```

0.0.2.5 2.1(c) t - test

```
##
##  One Sample t-test
##
## data:  car
## t = -5, df = 29, p-value = 2e-05
## alternative hypothesis: true mean is less than 25
## 95 percent confidence interval:
##  -Inf 23.5
## sample estimates:
## mean of x
##      22.8
```

```
sprintf("P-value of t-test test:%.5f", ttest$p.value)
```

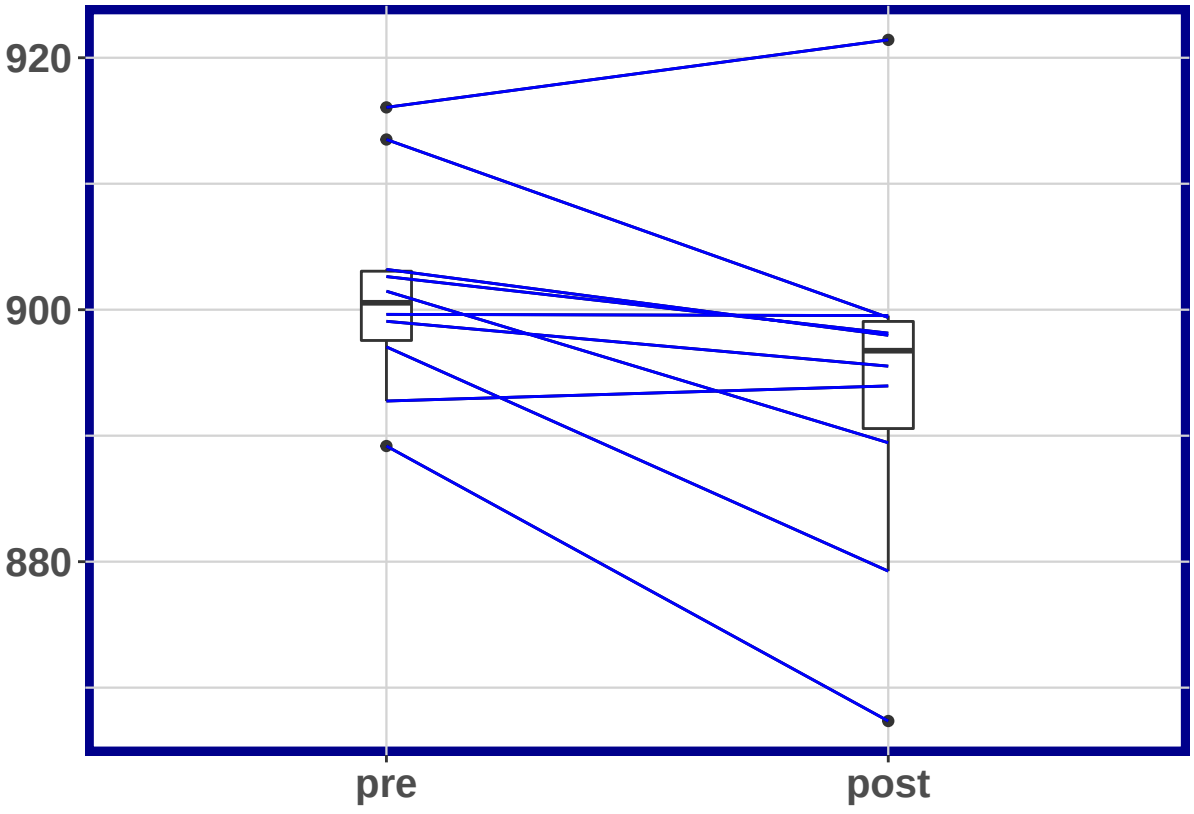
```
## [1] "P-value of t-test test:0.00002"
```

0.0.2.6 Observations:

- There is insufficient evidence to conclude that the average fuel economy of 2018 Sedans is less than the 25 mpg reported by Company A (p = 0.00002, 95%CI : -Inf 23.5).

0.0.3 2.2(a) Dependent Samples Tests

```
id <- c(1:10)
pre <- c(899.63, 913.51, 897.05, 889.18, 903.2, 916.06, 899.08,
    892.75, 901.47, 902.63)
post <- c(899.53, 899.38, 879.25, 867.35, 897.97, 921.42, 895.52,
    893.95, 889.44, 898.14)
datap <- data.frame(cbind(id, pre, post))
paired.plotProfiles(datap, "pre", "post") + geom_line(color = "blue") +
    theme(panel.background = element_rect(fill = "white", colour = "dark blue",
        size = 3), axis.text.x = element_text(size = 15, face = "bold"),
        axis.text.y = element_text(size = 15, face = "bold"),
        panel.grid.major = element_line(colour = "light gray",
            size = 0.4), panel.grid.minor = element_line(colour = "light gray",
            size = 0.4))
```



0.0.4 2.2(b) Dependent Samples Tests

```r
sprintf("Pre-swim time - Mean:%.2f and Standard deviation:%.2f",
    mean(datap$pre), sd(datap$pre))
```

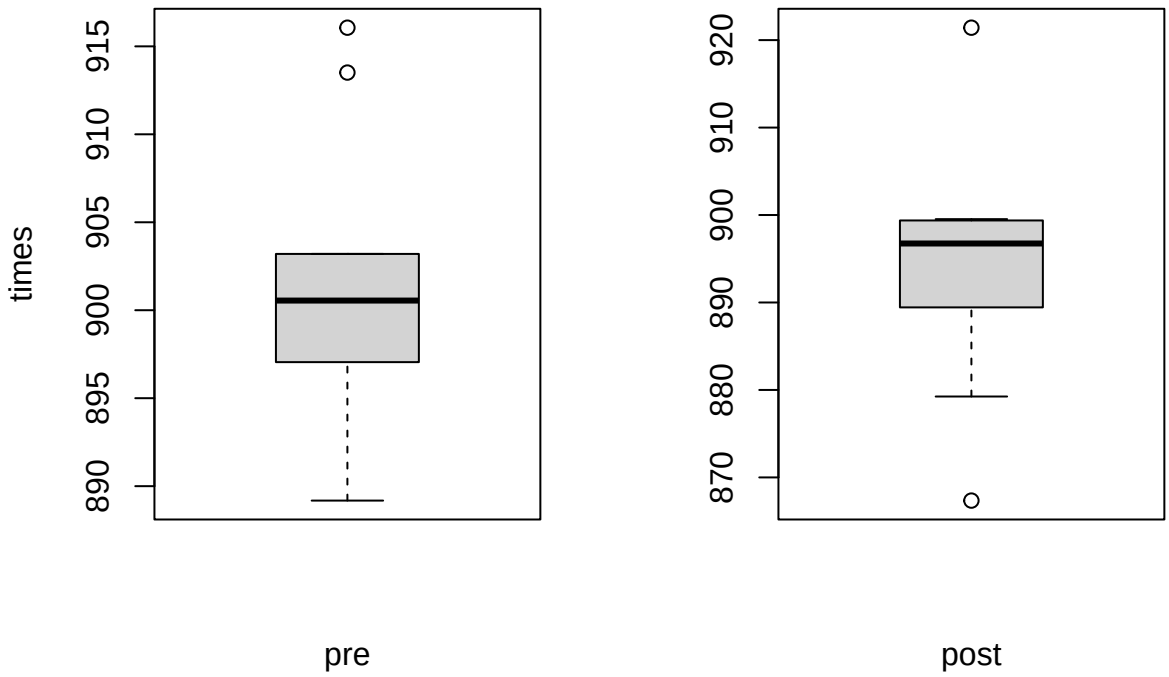
0.0.4.1 (i) Box plots

```
## [1] "Pre-swim time - Mean:901.46 and Standard deviation:8.29"
```

```r
sprintf("Post-swim time - Mean:%.2f and Standard deviation:%.2f",
    mean(datap$post), sd(datap$post))
```

```
## [1] "Post-swim time - Mean:894.20 and Standard deviation:14.12"
```

```r
par(mfrow = c(1, 2))
boxplot(x = datap$pre, xlab = "pre", ylab = "times")
boxplot(x = datap$post, xlab = "post")
```

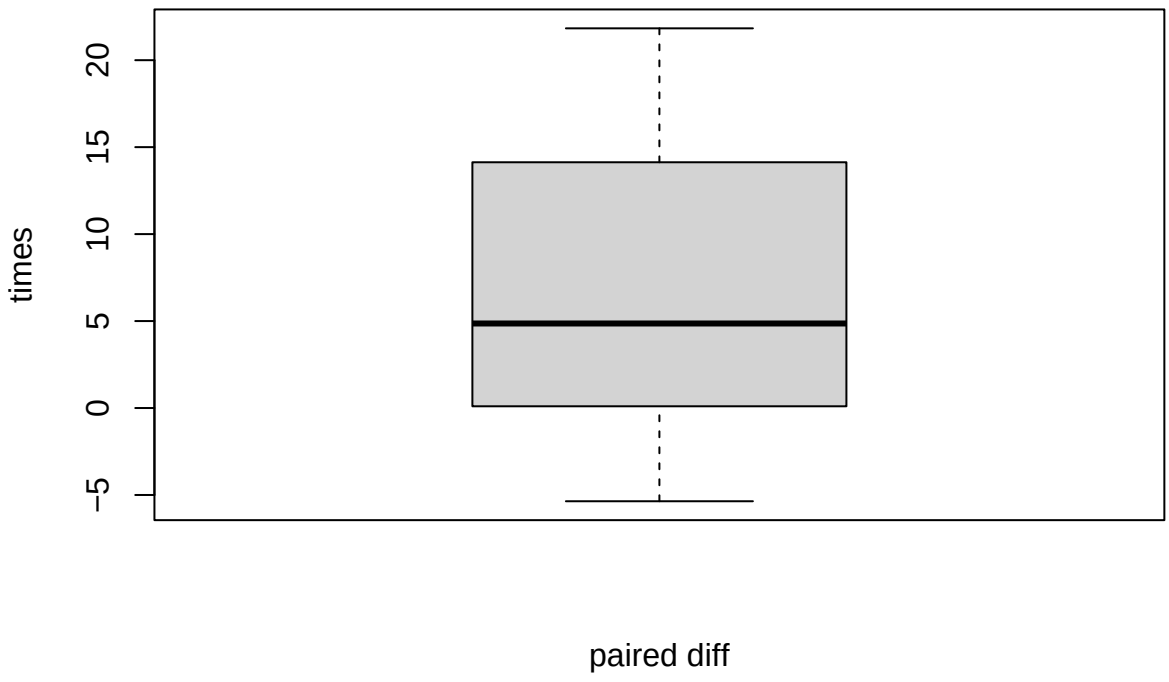



```r
paired_dif <- datap$pre - datap$post
sprintf("Paired differences - Mean:%.2f and Standard deviation:%.2f",
    mean(paired_dif), sd(paired_dif))
```

0.0.4.2 (ii) Box plots

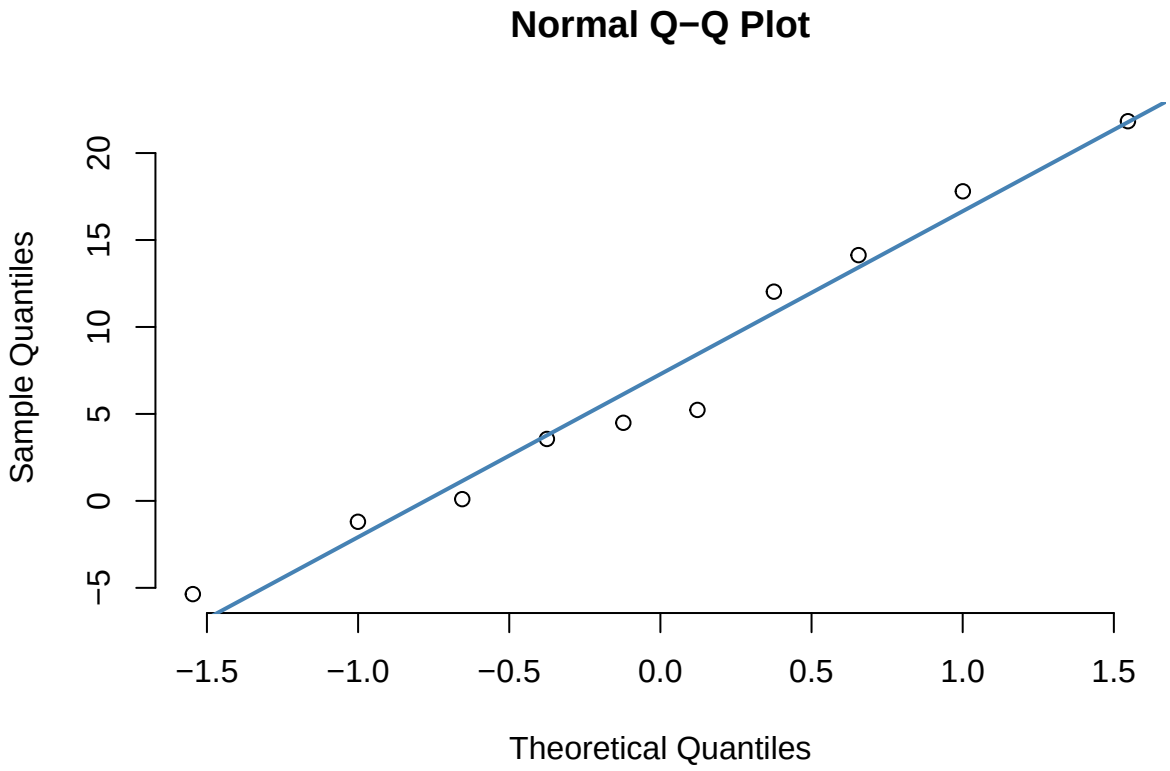
```
## [1] "Paired differences - Mean:7.26 and Standard deviation:8.82"
```

```r
boxplot(x = paired_dif, xlab = "paired diff", ylab = "times")
```

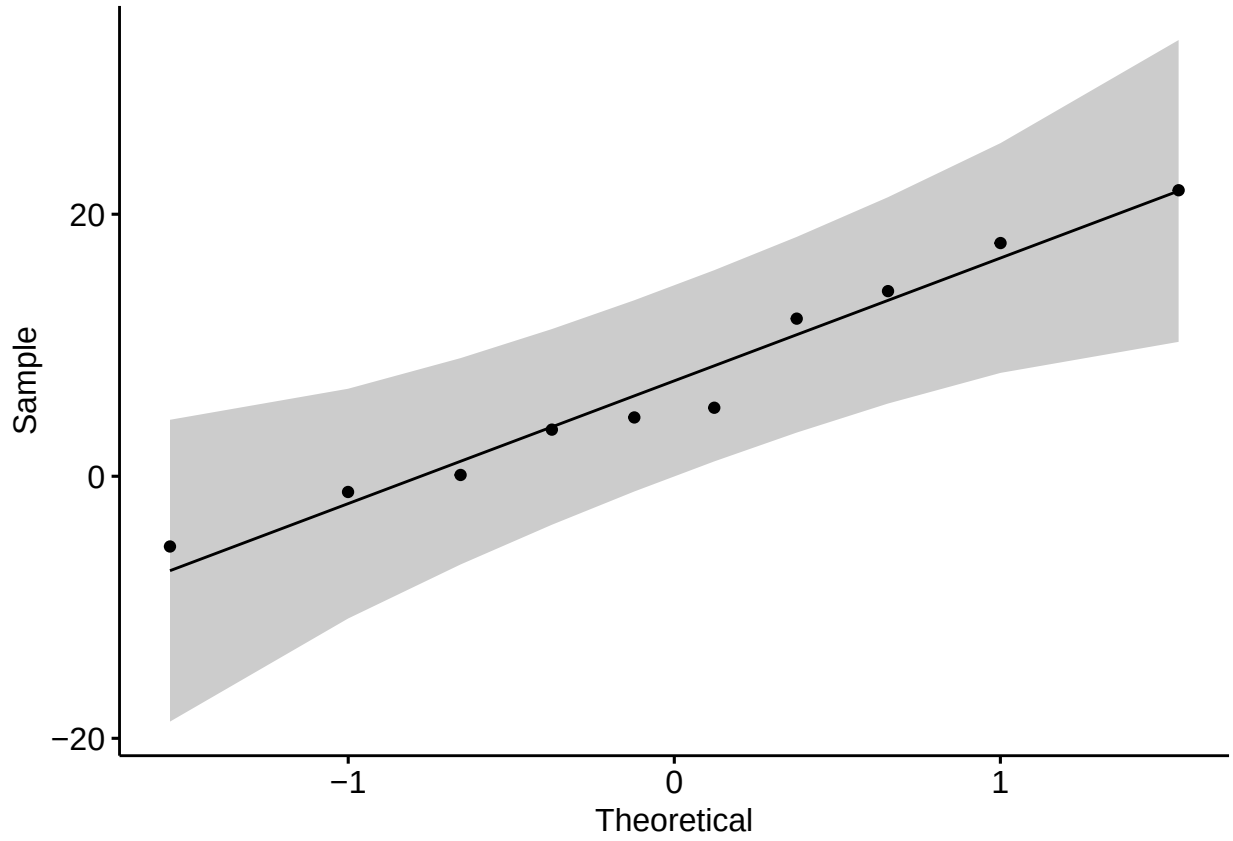



```
qqnorm(paired_dif, pch = 1, frame = FALSE)
qqline(paired_dif, col = "steelblue", lwd = 2)
```

0.0.4.3 (iii) Normality test




```
ggqqplot(paired_dif)
```




```
shapirotest <- shapiro.test(paired_dif)
sprintf("P-value of Shapiro normality test:%.4f", shapirotest$p.value)
```

```
## [1] "P-value of Shapiro normality test:0.7748"
```

0.0.4.4 Observation

- From the plots and shapiro test's p-value = 0.78 indicates the nature of normality in paired differences

```
datap$diff <- pre - post
t.test(pre, post, paired = TRUE, alternative = "greater")
```

0.0.4.5 (iii) t - test

```
##
##  Paired t-test
##
## data:  pre and post
## t = 3, df = 9, p-value = 0.01
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.15  Inf
## sample estimates:
## mean of the differences
##                    7.26
```

```r
t.test(datap$diff, alternative = "greater", mu = 0)
```

```
##
##  One Sample t-test
##
## data:  datap$diff
## t = 3, df = 9, p-value = 0.01
## alternative hypothesis: true mean is greater than 0
## 95 percent confidence interval:
##  2.15  Inf
## sample estimates:
## mean of x
##      7.26
```

0.0.4.6 Observation

- There is insufficient evidence to conclude that the training program is effective at reducing swim times for Men's 1500 Freestyle (p = 0.01). The program, on average, decreased swim time by 7.26 seconds (95% CI on difference: pre - post > 0: 2.15 Inf)."

0.0.5 2.3 Wilcoxon Signed-Rank Test

```r
wilcox.test(datap$diff, alternative = "greater", mu = 0)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  datap$diff
## V = 47, p-value = 0.02
## alternative hypothesis: true location is greater than 0
```

0.0.5.1 Observation

- For wilcox test, we have p-value=0.02 which is similar to t-test but less than the significant value 0.05. This suggest that, our initial hypothesis of improve in the swim times is incorrect.

0.1 Document Information.

All of the statistical analyses in this document will be performed using R version 4.1.0 (2021-05-18). R packages used will be maintained using the packrat dependency management system.

```r
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
```

```
##
## attached base packages:
## [1] grid      stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] PairedData_1.1.1   mvtnorm_1.1-2      gld_2.6.2          ggpubr_0.4.0
##  [5] car_3.0-11         carData_3.0-4      mnormt_2.0.2       vcd_1.4-8
##  [9] epiDisplay_3.5.0.1 nnet_7.3-16        foreign_0.8-81     Hmisc_4.5-0
## [13] Formula_1.2-4      survival_3.2-11    lattice_0.20-44    MASS_7.3-54
## [17] ggplot2_3.3.5      rmarkdown_2.8      knitr_1.33
##
## loaded via a namespace (and not attached):
##  [1] tidyr_1.1.3          splines_4.1.0        tmvnsim_1.0-2
##  [4] highr_0.9            lmom_2.8             latticeExtra_0.6-29
##  [7] cellranger_1.1.0     yaml_2.2.1           pillar_1.6.1
## [10] backports_1.2.1      glue_1.4.2           digest_0.6.27
## [13] RColorBrewer_1.1-2   ggsignif_0.6.2       checkmate_2.0.0
## [16] colorspace_2.0-1     htmltools_0.5.1.1    Matrix_1.3-3
## [19] pkgconfig_2.0.3      broom_0.7.8          haven_2.4.1
## [22] purrr_0.3.4          scales_1.1.1         jpeg_0.1-8.1
## [25] openxlsx_4.2.4       rio_0.5.27           proxy_0.4-26
## [28] htmlTable_2.2.1      tibble_3.1.2         farver_2.1.0
## [31] generics_0.1.0       ellipsis_0.3.2       withr_2.4.2
## [34] magrittr_2.0.1       crayon_1.4.1         readxl_1.3.1
## [37] evaluate_0.14        fansi_0.5.0          class_7.3-19
## [40] rstatix_0.7.0        forcats_0.5.1        tools_4.1.0
## [43] data.table_1.14.0    hms_1.1.0            formatR_1.11
## [46] lifecycle_1.0.0      stringr_1.4.0        munsell_0.5.0
## [49] cluster_2.1.2        zip_2.2.0            e1071_1.7-7
## [52] compiler_4.1.0       rlang_0.4.11         rstudioapi_0.13
## [55] htmlwidgets_1.5.3    labeling_0.4.2       base64enc_0.1-3
## [58] gtable_0.3.0         abind_1.4-5          curl_4.3.1
## [61] R6_2.5.0             gridExtra_2.3        zoo_1.8-9
## [64] dplyr_1.0.7          utf8_1.2.1           stringi_1.6.1
## [67] Rcpp_1.0.6           vctrs_0.3.8          rpart_4.1-15
## [70] png_0.1-7            tidyselect_1.1.1     xfun_0.23
## [73] lmtest_0.9-38
```