Stat 502 HW3

Dongyang Wang

10/25/2021

1

(a)

```
#rm(list = ls())
sleep <- sleep
sleep1 <- sleep[sleep$group == 1,]
sleep2 <- sleep[sleep$group == 2,]
t.test(sleep1$extra, sleep2$extra, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  sleep1$extra and sleep2$extra
## t = -1.8608, df = 18, p-value = 0.07919
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.363874  0.203874
## sample estimates:
## mean of x mean of y
##      0.75      2.33
```

The confidence interval is [-3.363874, 0.203874].

(b)

Since Y_A, Y_B are normally distributed, $Y_A - Y_B \sim N(\delta, \frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B})$. To form a standard normal, we have $\frac{Y_A - Y_B - \delta}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}}}$. Since we also know $\frac{n_A + n_B - 2}{\sigma^2} s_P^2$ follows chi-squared distribution with $n_A + n_B - 2$ as the dof, and

X and Z are independent. So, by definition of t-statistic, we have $t(Y_A, Y_B) = \frac{\frac{Y_A - Y_B - \delta}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}}}}{\sqrt{\frac{n_A + n_B - 2}{\sigma^2} s_P^2 / (n_A + n_B - 2)}} = \frac{Y_A - Y_B}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}} - \frac{\delta}{s_p \sqrt{\frac{1}{n_A} + \frac{1}{n_B}}}$. As we can observe, this is a non-central t-distribution.

The components are (i) Z $= \frac{Y_A - Y_B - \delta}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}}}$. (ii) X $= \frac{n_A + n_B - 2}{\sigma^2} s_P^2$. (iii) The non-centrality parameter is $\gamma = \frac{-\delta}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}}}$.

(c) Since power is an increasing function of absolute value of the non centrality parameter $\gamma = |\frac{\delta}{\sqrt{\frac{\sigma^2}{n_A} + \frac{\sigma^2}{n_B}}}|$, we want to maximize γ. Rewriting with $n_B = N - n_A$, we have $\gamma = |\frac{\delta}{\sigma \sqrt{\frac{1}{n_A} + \frac{1}{N - n_A}}}| = |\frac{\delta}{\sigma \sqrt{\frac{N}{n_A (N - n_A)}}}| = |\frac{\delta \sqrt{n_A (N - n_A)}}{\sigma \sqrt{N}}|$. To maximize this, we let $N - n_A = n_A$, that is $n_A = \frac{1}{2} N$. Hence, we maximized power.

Dongyang Wang
Stat 502 Q2

2. $E(MST) = E\left[\frac{\sum_{i=1}^{m} n(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{m-1} \mid \mu\right]$

$\Leftarrow \begin{cases} \bar{Y}_{i\cdot} = \mu_i + \bar{\varepsilon}_i \\ \bar{Y}_{\cdot\cdot} = \bar{\mu} + \bar{\varepsilon} \end{cases}$

$= \frac{n}{m-1} E\left(\frac{\sum_{i=1}^{m}(\mu_i + \bar{\varepsilon}_i - \bar{\mu} - \bar{\varepsilon})^2}{1} \mid \mu\right]$

$= \frac{n}{m-1} E\left[\sum_{i=1}^{m}\left((\mu_i - \bar{\mu}) + (\bar{\varepsilon}_i - \bar{\varepsilon})\right)^2 \mid \mu\right]$

Since $E(\bar{\varepsilon}_i) = E(\bar{\varepsilon})$ the term $2(\mu_i - \bar{\mu})\cdot(\bar{\varepsilon}_i - \bar{\varepsilon})$ is canceled out with a value of 0.

$= \frac{n}{m-1} E\left[\sum_{i=1}^{m}(\mu_i - \bar{\mu})^2 + (\bar{\varepsilon}_i - \bar{\varepsilon})^2 \mid \mu\right]$

$\sigma_{\varepsilon_i}^2 = \frac{\sum_{i=1}^{m}(\varepsilon_i - \bar{\varepsilon})^2}{m-1} = \sigma_{\varepsilon_i}^2$

because of that the sample variance (w/ denominator $m-1$) is an unbiased estimator of true variance.

$= \frac{n}{m-1} E\left((m-1)\sigma_{\varepsilon_i}^2\right) + \frac{nE\left[\sum_{i=1}^{m}(\mu_i - \bar{\mu})^2\right]}{m-1}$

$E(\sigma_{\varepsilon_i}^2) = \frac{\sigma^2}{n}$

$= n\frac{\sigma^2}{n} + \frac{n\sum_{i=1}^{m}(\mu_i - \bar{\mu})^2}{m-1}$

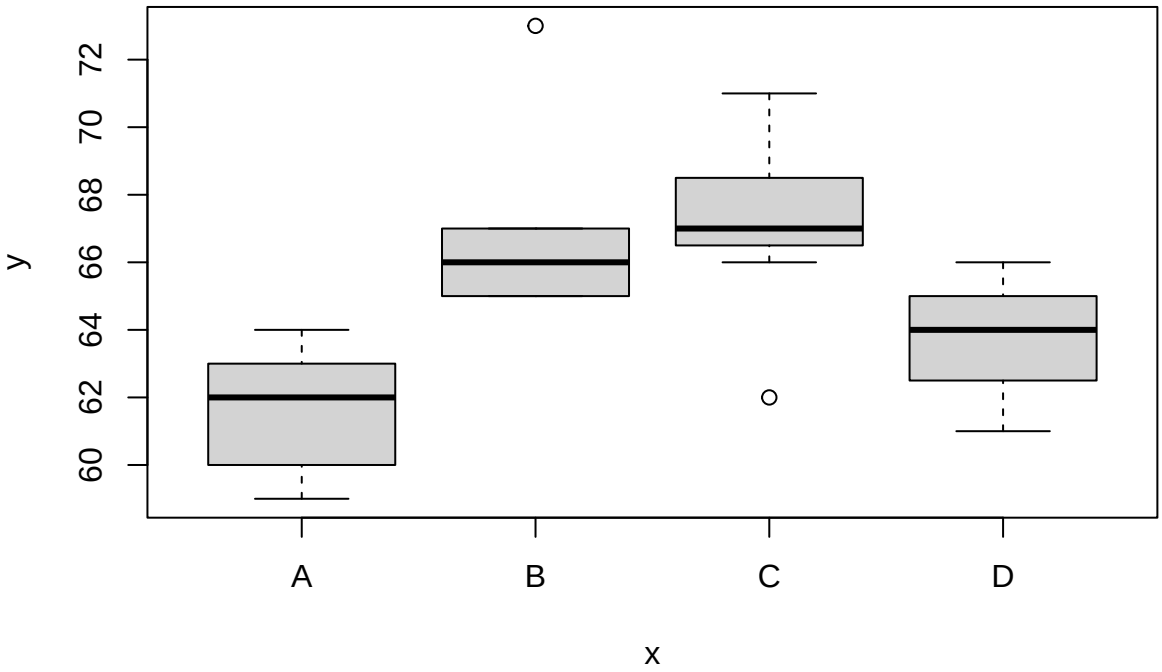
$= \sigma^2 + \frac{n\sum_{i=1}^{m}(\mu_i - \bar{\mu})^2}{m-1}$

3

(a) The model is $y_{ij} = \mu + \tau_i + \epsilon_{ij}$. The meaning of μ means the mean when no treatment is in place. τ_i represent the within treatment variation for the four treatments. Key assumptions include $E(\epsilon_{ij}) = 0$, $Var(\epsilon_{ij}) = \sigma^2$. We also want the sum of τ_i is 0.

(b)

```
A <- c(62,60,63,59,64)
B <- c(65,67,73,65,66)
C <- c(69,66,71,67,67,68,62)
D <- c(66,62,65,61,64,65,63)
effect <- c(A,B,C,D)
diet <- c(rep('A', 5), rep('B', 5),rep('C', 7), rep('D', 7))
data <- data.frame(effect, diet)
plot(as.factor(data$diet), data$effect)
```




```
mean(A)
```

```
## [1] 61.6
```

```
mean(B)
```

```
## [1] 67.2
```

```
mean(C)
```

```
## [1] 67.14286
```

```
mean(D)
```

```
## [1] 63.71429
```

```
mean(data)
```

```
## Warning in mean.default(data): argument is not numeric or logical: returning NA
```

```
## [1] NA
```

There is no big difference between C and B, but among others theire is a difference.

(c)

```
#Group sample variances
sA <- var(A)
sB <- var(B)
sC <- var(C)
sD <- var(D)
sA;sB;sC;sD
```

```
## [1] 4.3
```

```
## [1] 11.2
```

```
## [1] 7.809524
```

```
## [1] 3.238095
```

```
# MSE
mse <- (4*sA+4*sB+6*sC+6*sD)/(24 - 4)
mse
```

```
## [1] 6.414286
```

Group sample variances are 4.3, 11.2, 7.809524, 3.238095. The MSE is 6.414286.

(d)

```
mst <- 5/3*(mean(A) - mean(data$effect))^2 + 5/3*(mean(B) - mean(data$effect))^2
+7/3*(mean(C) - mean(data$effect))^2 + 7/3*(mean(D) - mean(data$effect))^2
```

```
## [1] 14.57143
```

```
mst
```

```
## [1] 27.33333
```

MST is 41.90476. It is way larger than MSE, so we can expect to reject the null that the diets are no different.

(e)

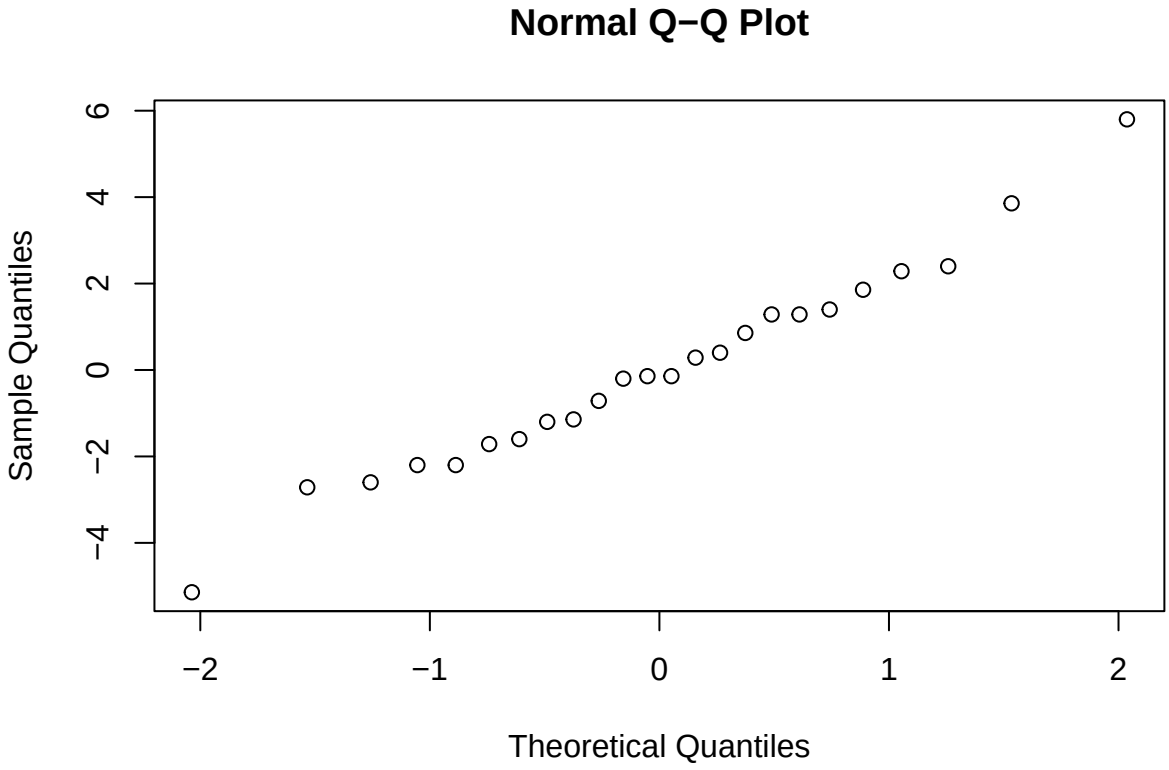
```
anova(lm(data$effect ~ data$diet))
```

```
## Analysis of Variance Table
##
## Response: data$effect
##           Df Sum Sq Mean Sq F value   Pr(>F)
## data$diet  3 125.71  41.905   6.533 0.002937 **
## Residuals 20 128.29   6.414
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, since the p-value is 0.002937, we can safely reject the null and say that there is a difference for these four diets.

(f) Since the fitted values are the mean for each diet, we have

```
data$miu.hat <- c(rep(mean(A), 5), rep(mean(B), 5),rep(mean(C), 7), rep(mean(D), 7))
data$residual <- data$effect - data$miu.hat
qqnorm(data$residual)
```



Yes, the residuals appear to follow a normal distribution, since the points on the plot are pretty much along the diagonal.

4

(a)

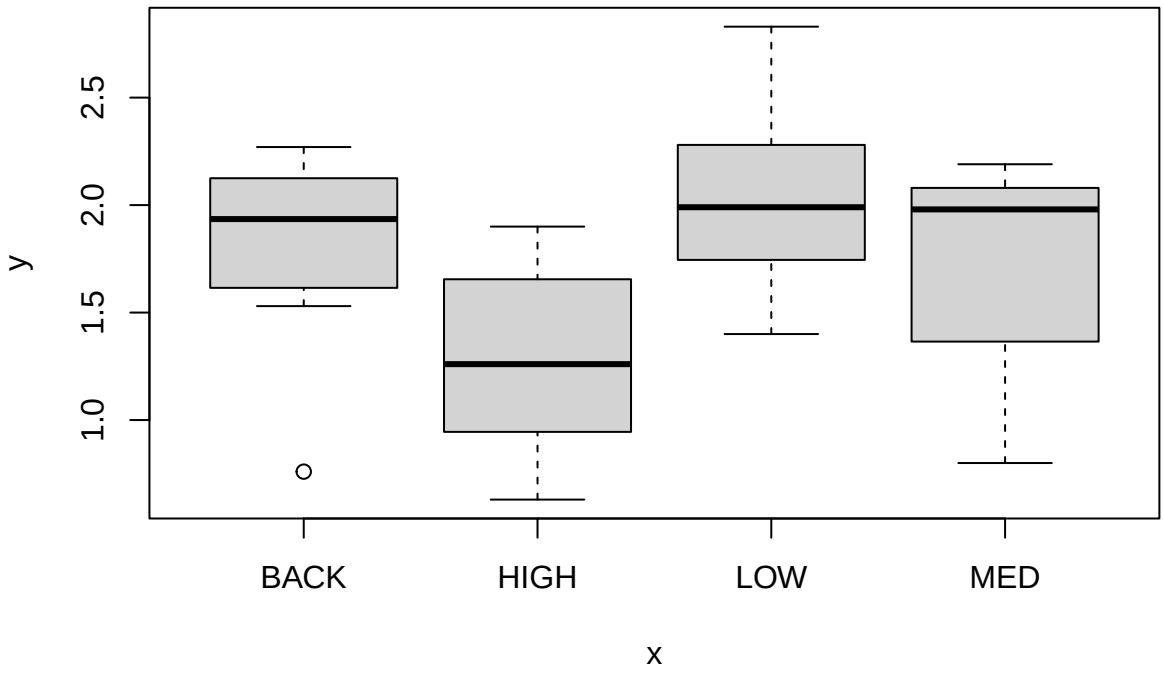
```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
zinc <- readRDS("zinc.RDS")
plot(as.factor(zinc$ZINC), zinc$DIVERSITY)
```




```
mean(zinc$DIVERSITY)
```

```
## [1] 1.710312
```

```
# Method 1
meanZ <- zinc %>%
  group_by(ZINC) %>%
  summarize(mean(DIVERSITY))
# 2
aggregate(x= zinc$DIVERSITY,
          # Specify group indicator
          by = list(zinc$ZINC),
          # Specify function (i.e. mean)
          FUN = mean)
```

```
##   Group.1       x
## 1    BACK 1.79750
## 2    HIGH 1.28125
## 3     LOW 2.03250
## 4     MED 1.73000
```

Yes, there appears to be a difference in biodiversity of rivers with different Zinc levels.

(b)

```
#Group sample variances
varZ <- zinc %>%
  group_by(ZINC) %>%
  summarize(var(DIVERSITY))
varZ
```

```
## # A tibble: 4 x 2
##   ZINC  `var(DIVERSITY)`
##   <fct>            <dbl>
## 1 BACK             0.235
## 2 HIGH             0.208
```

```
## 3 LOW              0.198
## 4 MED              0.288
```

```
#count
numZ <- zinc %>%
  count(ZINC)
#MSE
mse1 <- (0.2354786+0.2081268+0.1980214  +0.2876286
)/4
mse1
```

```
## [1] 0.2323139
```

Group variances are BACK 0.2354786, HIGH 0.2081268, LOW 0.1980214, MED 0.2876286. MSE is 0.2323139.

(c)

```
meanZ
```

```
## # A tibble: 4 x 2
##   ZINC  `mean(DIVERSITY)`
##   <fct>             <dbl>
## 1 BACK               1.80
## 2 HIGH               1.28
## 3 LOW                2.03
## 4 MED                1.73
```

```
mst1 <- 8/3*((1.79750 - mean(zinc$DIVERSITY))^2 + (1.28125 - mean(zinc$DIVERSITY))^2 +
               (2.03250 - mean(zinc$DIVERSITY))^2 + (1.73000 - mean(zinc$DIVERSITY))^2 )
mst1
```

```
## [1] 0.7890365
```

```
fratio <- mst1 / mse1
fratio
```

```
## [1] 3.396425
```

```
#verify
anova(lm(zinc$DIVERSITY ~ zinc$ZINC))
```

```
## Analysis of Variance Table
##
## Response: zinc$DIVERSITY
##           Df Sum Sq Mean Sq F value  Pr(>F)
## zinc$ZINC  3 2.3671 0.78904  3.3964 0.03151 *
## Residuals 28 6.5048 0.23231
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Yes, I would have expected this F ratio – it is large such that we can reject the null that there is no difference in biodiversity in zones with different Zinc levels. This is consistent with my observation in the plots and early data exploration of the mean.

(d)

```
zinc.level <- zinc$ZINC
fr <- c()
for(i in 1:1000){
  zinc.sim <- sample(zinc.level)
  fr <- c(fr, anova(lm(zinc$DIVERSITY ~ zinc.sim))$F[1] )
```

```
}
mean(fr >= 3.3964)
```

```
## [1] 0.027
```

With a p-value of 0.027, we can reject the null and say that Zinc levels do affect biodiversity.