qbs120_ps8_gibran

Gibran Erlangga

11/7/2021

```
library(pwr)
library(tidyverse)
```

Question 1

Pancreatic Cancer Example: Consider two pancreatic cancer gene expression experiments (A and B). Both experiments measure transcript abundance for a set of human protein coding genes in bulk tissue samples taken from pancreatic cancer patients with either non-metastatic (M0) or metastatic disease (M1). The goal of these experiments is to identify a set of genes whose expression differs between M0 and M1 tumors and may therefore represent potential therapeutic targets and/or prognostic biomarkers.

- **Experiment A**: Measures expression of 1000 randomly selected protein coding genes. Assume that 10 of these genes are differentially expressed (DE) between M0 and M1 pancreatic tumors.

- **Experiment B**: Measures expression of 100 genes with prior evidence of DE between M0 and M1 solid human cancers. Assume that 20 of these genes are DE between the M0 and M1 pancreatic tumors.

For both experiments, assume that:
- A total of n bulk tissue samples from separate individuals are analyzed (n/2 from M0 tumors and n/2 from M1 tumors).
- The non-DE genes have expression values that follow a N(μM0,1) distribution in both M0 and M1 samples. (Note: a normal distribution is being assumed to simplify the computations in the questions below; transcript abundance computed using methods like RNA-seq is count data and is typically modeled by a Poisson or negative binomial dis- tribution.)
- The DE genes have expression values that follow a N(μM0,1) distribution among M0 samples and a N(μM1,1) distribution among M1 samples.
- Researchers are interested in analyzing the expression values for each gene to test the following null and alternative hypotheses:
– H0 : μM0 = μM1
– HA :μM0 = μM1

Questions:

(a) What is the minimum number (i.e., n) of tumor samples required to achieve a power of 0.8 for testing a single DE gene in Experiment A with a type I error rate of $\alpha = 0.05$ and the specific HA of μM0 = 0.5 and μM1 = 1.0? (OK to use pwr R package)

```
pwr.2p.test(h = 0.5, power = 0.8, sig.level = 0.05)
```

```
##
##      Difference of proportion power calculation for binomial distribution (arcsine transformation)
```

```
##
##                 h = 0.5
##                 n = 62.79088
##         sig.level = 0.05
##             power = 0.8
##       alternative = two.sided
##
## NOTE: same sample sizes
```

(b) Does the required n calculated for Experiment A in a) differ for Experiment B?

```
pwr.t.test(d=0.5, power = 0.8, sig.level = 0.05)
```

```
##
##      Two-sample t test power calculation
##
##                 n = 63.76561
##                 d = 0.5
##         sig.level = 0.05
##             power = 0.8
##       alternative = two.sided
##
## NOTE: n is number in *each* group
```

value for n is similar to (a).

(c) If only 50 samples are available, i.e., $n = 50$, what is the power to detect the effect size of $\mu M0 = 0.5$ and $\mu M1 = 1.0$ with $\alpha = 0.05$?

```
pwr.t.test(d = 0.5, n = 25, sig.level = 0.05)
```

```
##
##      Two-sample t test power calculation
##
##                 n = 25
##                 d = 0.5
##         sig.level = 0.05
##             power = 0.4101003
##       alternative = two.sided
##
## NOTE: n is number in *each* group
```

so, power is 0.41.

(d) Confirm the theoretical power computed in c) via simulation. Hint: remember that power is defined under HA.

```
n_sim <- 10000
power <- c()

for (i in 1:n_sim) {
```

```r
  mu_0 <- rnorm(25, 0.5, 1)
  mu_1 <- rnorm(25, 1, 1)
  power[i] <- t.test(mu_0, mu_1, alternative = "two.sided", var.equal = TRUE)$p.value
}

(length(which(power < 0.05))/n_sim)
```

```
## [1] 0.4064
```

close enough.

(e) How could a researcher increase the power for the analysis of a single DE gene?

- increase significance level
- increase effect size
- decrease standard deviation
- increase sample size

(f) Estimate the empirical power if the value of μM1 for each DE gene is modeled as a random draw from U(0,1), n = 50,α = 0.05 and μM0 = 0.1.

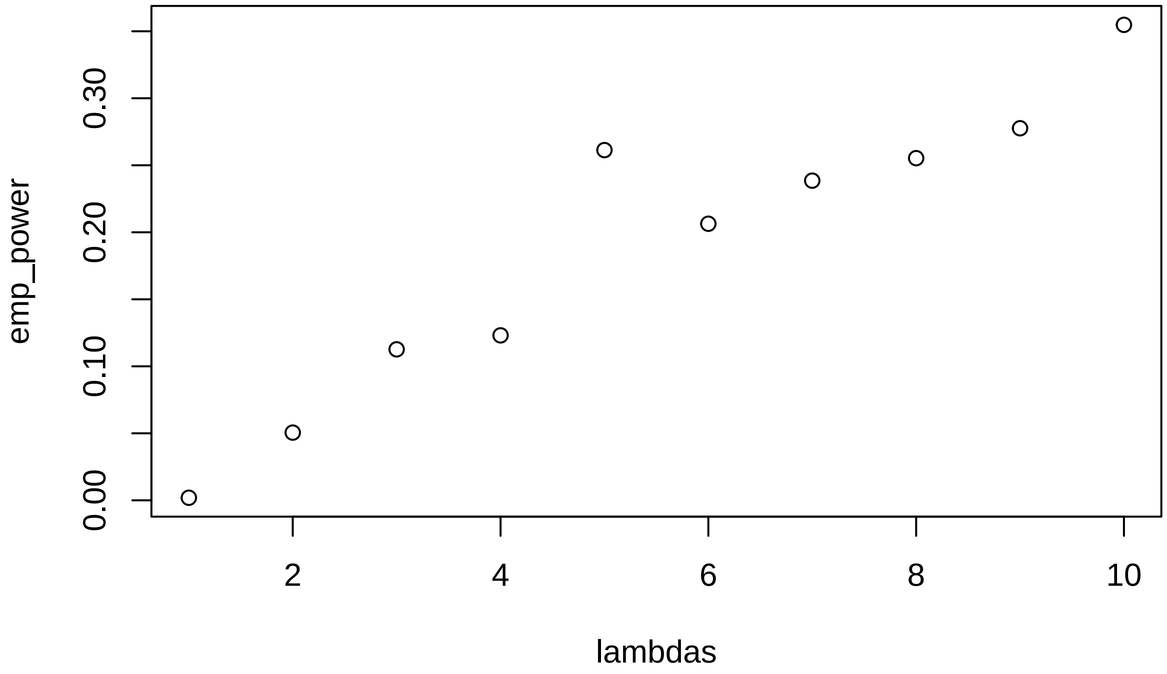
```r
n_sim <- 10000
emp_power <- c()

for (i in 1:n_sim) {
  mu_0 <- rnorm(25, runif(25, 0, 1), 1)
  mu_1 <- rnorm(25, 1, 1)
  emp_power[i] <- t.test(mu_0, mu_1, alternative = "two.sided", var.equal = TRUE)$p.value
}

(length(which(emp_power < 0.05))/n_sim)
```

```
## [1] 0.3923
```

Question 2

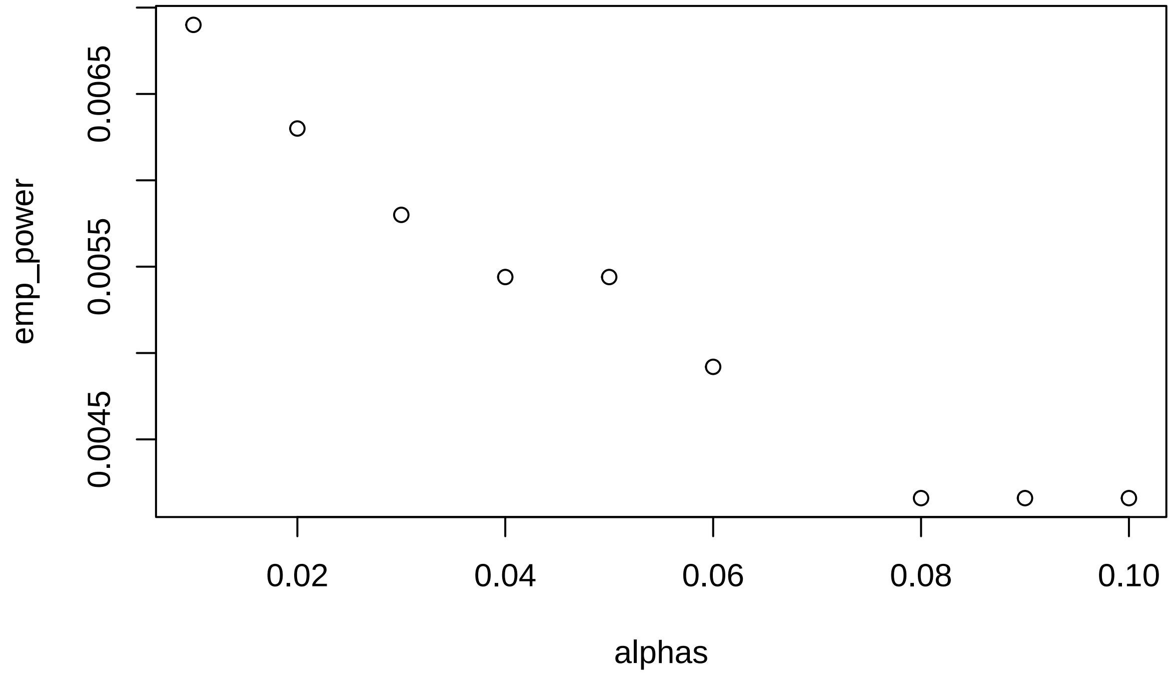
For the coefficient of skewness question in Problem Set 6, calculate the empirical power of your normality test for the following cases:
(a) Type I error rate of 0.05 and HA of 100 iid Poisson RVs (i.e., part d)) with values of 1 to 10 (incrementing by 1). Plot empirical power vs. λ. Do the results match your expectations? Explain.

```r
# add helper functions
biased.sd = function(x) {
  biased.var = mean((x-mean(x))^2)
  return (sqrt(biased.var))
}

coef.of.skewness = function(x) {
  b.1 = mean((x - mean(x))^3)/biased.sd(x)^3
  return (b.1)
}

simPVal = function(x, ranked.sim.values) {
  n = length(ranked.sim.values)
  smaller.vals =which(ranked.sim.values <= x)
  if (length(smaller.vals) == 0) {
    alpha.low = 0
    } else{
      alpha.low = length(smaller.vals)/n
    }
  larger.vals = which(ranked.sim.values >= x)
  if (length(larger.vals) == 0) {
    alpha.hi = 0
  } else {
    alpha.hi = length(larger.vals)/n
  }
  p.val = 2*min(alpha.low, alpha.hi)
  return(p.val)
}

# initiate data from PS6
sim.data = matrix(rnorm(1000*100), nrow=1000, ncol=100)
sim.b.1 = apply(sim.data, 1, coef.of.skewness)
ranked.sim.b.1 = sort(sim.b.1)

lambdas <- seq(from=1,to=10,by=1)
emp_power = c()

# iterate for all lambdas
for (i in lambdas) {
  test.data = matrix(rpois(10000, lambda=i), nrow=100, ncol=100)
  test.b.1 = apply(test.data, 1, coef.of.skewness)
  p.values = sapply(test.b.1, function(x) simPVal(x, ranked.sim.b.1))
  power <- p.values[(p.values > 0.05)]
  emp_power[i] <- sum(power)/length(p.values)
}

plot(lambdas, emp_power)
```



(b) Type I error rate values between 0.01 and 0.1 (incrementing by 0.01) and HA of 100 iid Poisson RVs with $\lambda = 1$. Plot empirical power vs. type I error rate. Do the results match your expectations? Explain.

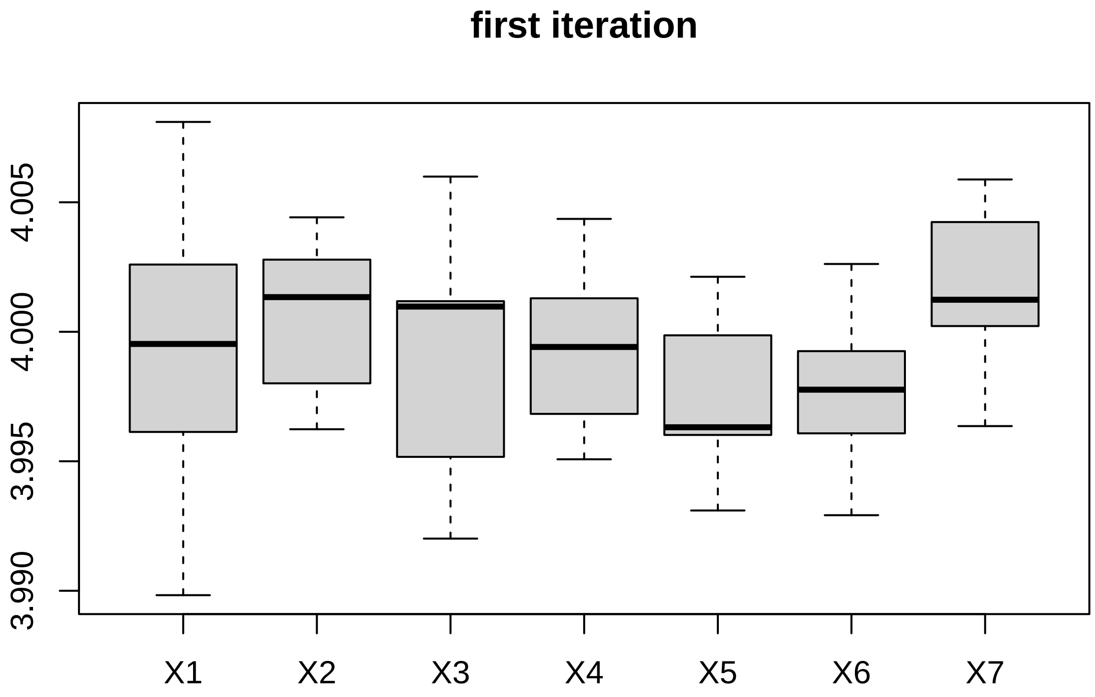
```
alphas <- seq(from=0.01,to=0.1,by=0.01)
emp_power = c()

test.data = matrix(rpois(10000, lambda=1), nrow=100, ncol=100)
test.b.1 = apply(test.data, 1, coef.of.skewness)
p.values = sapply(test.b.1, function(x) simPVal(x, ranked.sim.b.1))

# iterate for all alphas
for (i in alphas) {
  power <- p.values[(p.values > i)]
  emp_power[i*100] <- sum(power)/length(p.values)
}

plot(alphas, emp_power)
```



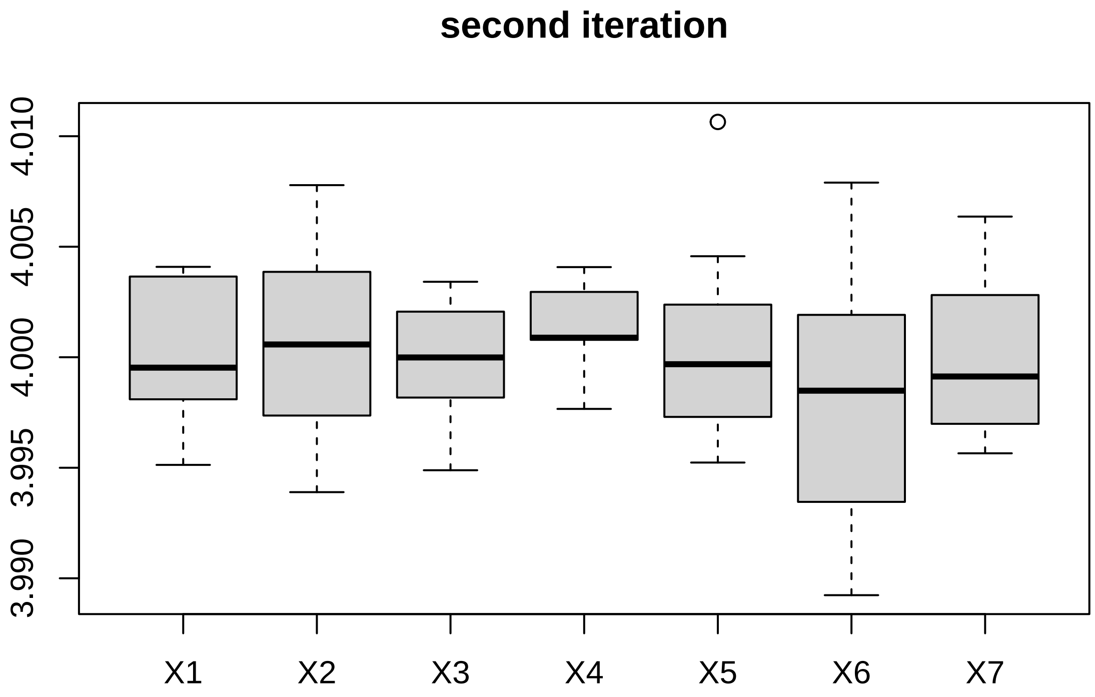
Question 3

(Based on Rice 12.1) Simulate observations like those of Figure 12.1 under the H0 of no treatment effects. That is, simulate seven batches of ten normally distributed random numbers with mean 4 and variance 0.0037. Make parallel boxplots of the seven batches like those of Figure 12.1. Do this twice. Your figures display the kind of variability that random fluctuations can cause; do you see any pairs of labs that appear different in either mean level or dispersion?

```
storage_1 <- data.frame(matrix(nrow=10, ncol=7))
storage_2 <- data.frame(matrix(nrow=10, ncol=7))
n_iter = 7

# get values for first iteration
for (i in 1:7) {
  storage_1[i] = rnorm(10, 4, 0.0037)
}

# get values for second iteration
for (i in 1:7) {
  storage_2[i] = rnorm(10, 4, 0.0037)
}

boxplot(storage_1, main='first iteration')
```



```
boxplot(storage_2, main='second iteration')
```



```
summary(storage_1)
```

```
##        X1              X2              X3              X4
##  Min.   :3.990   Min.   :3.996   Min.   :3.992   Min.   :3.995
##  1st Qu.:3.997   1st Qu.:3.999   1st Qu.:3.996   1st Qu.:3.997
##  Median :4.000   Median :4.001   Median :4.001   Median :3.999
##  Mean   :3.999   Mean   :4.001   Mean   :3.999   Mean   :3.999
##  3rd Qu.:4.002   3rd Qu.:4.003   3rd Qu.:4.001   3rd Qu.:4.001
##  Max.   :4.008   Max.   :4.004   Max.   :4.006   Max.   :4.004
##        X5              X6              X7
##  Min.   :3.993   Min.   :3.993   Min.   :3.996
##  1st Qu.:3.996   1st Qu.:3.996   1st Qu.:4.000
##  Median :3.996   Median :3.998   Median :4.001
##  Mean   :3.997   Mean   :3.998   Mean   :4.002
##  3rd Qu.:3.999   3rd Qu.:3.999   3rd Qu.:4.004
##  Max.   :4.002   Max.   :4.003   Max.   :4.006
```

```
summary(storage_2)
```

```
##        X1              X2              X3              X4
##  Min.   :3.995   Min.   :3.994   Min.   :3.995   Min.   :3.998
##  1st Qu.:3.998   1st Qu.:3.998   1st Qu.:3.998   1st Qu.:4.001
##  Median :4.000   Median :4.001   Median :4.000   Median :4.001
##  Mean   :4.000   Mean   :4.001   Mean   :4.000   Mean   :4.001
##  3rd Qu.:4.004   3rd Qu.:4.004   3rd Qu.:4.002   3rd Qu.:4.002  
##  Max.   :4.004   Max.   :4.008   Max.   :4.003   Max.   :4.004  
##        X5              X6              X7       
##  Min.   :3.995   Min.   :3.989   Min.   :3.996  
##  1st Qu.:3.998   1st Qu.:3.994   1st Qu.:3.997  
##  Median :4.000   Median :3.998   Median :3.999  
##  Mean   :4.001   Mean   :3.998   Mean   :4.000  
##  3rd Qu.:4.002   3rd Qu.:4.001   3rd Qu.:4.003  
##  Max.   :4.011   Max.   :4.008   Max.   :4.006  
```

Yes, there are differences in the lab pairs shown above between sample 1 and sample 2 due to the nature of randomization of the normality function. We can have same parameters for the random variable, but different results.

Question 4

(Based on Rice 12.3) For the one-way analysis of variance with I = 2 treatment groups, show that the F statistic is t^2, where t is the usual t statistic for a two-sample case.

Getting F-statistic with I=2:

$$\begin{aligned}
F &= \frac{\frac{SS_B}{I-1}}{\frac{SS_W}{I(J-1)}} \\
&= \frac{SS_B}{\frac{SS_W}{2(J-1)}} \\
&= 2(J-1) \cdot \frac{J\sum_{i=1}^{I}(\bar{Y}_i - \bar{Y})^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}(\bar{Y}_{i,j} - \bar{Y}_i)^2} \\
&= 2(J-1) \cdot \frac{J(\bar{Y}_1 - \bar{Y})^2 + J(\bar{Y}_2 - \bar{Y})^2}{\sum_{j=1}^{J}(\bar{Y}_{1,j} - \bar{Y}_1)^2 + \sum_{j=1}^{J}(\bar{Y}_{2,j} - \bar{Y}_2)^2}
\end{aligned}$$

with definition of $\bar{Y}$ being:

$$\begin{aligned}
\bar{Y} &= \frac{\sum_{j=1}^{J} Y_{1,j} + Y_{2,j}}{2J} \\
&= \frac{\bar{Y}_1 + \bar{Y}_2}{2}
\end{aligned}$$

and simplify the denominator of F-statistic as:

$$\sum_{j=1}^{J}(\bar{Y}_{1,j} - \bar{Y}_1)^2 + \sum_{j=1}^{J}(\bar{Y}_{2,j} - \bar{Y}_2)^2 = (J-1) \cdot Js^2_{\bar{Y}_1 - \bar{Y}_2}$$

Plugging both formulas back to F-statistic formula, we get:

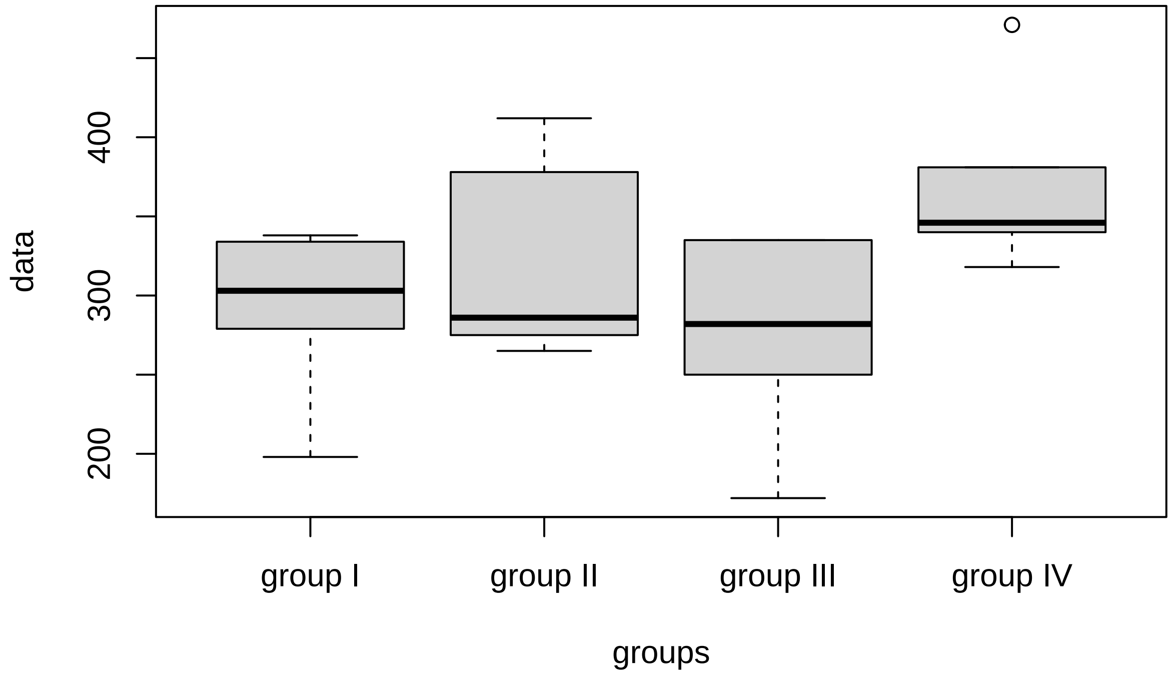
$$
\begin{aligned}
F &= 2(J-1)\cdot\frac{J(\bar{Y}_1-\frac{\bar{Y}_1+\bar{Y}_2}{2})^2+J(\bar{Y}_2-\frac{\bar{Y}_1+\bar{Y}_2}{2})^2}{(J-1)\cdot Js^2_{\bar{Y}_1-\bar{Y}_2}}\\
&= 2(J-1)\cdot\frac{J(\frac{\bar{Y}_1-\bar{Y}_2}{2})^2+J(\frac{\bar{Y}_1-\bar{Y}_2}{2})^2}{(J-1)\cdot Js^2_{\bar{Y}_1-\bar{Y}_2}}\\
&= 2(J-1)\cdot\frac{\frac{J}{2}\cdot(\bar{Y}_1-\bar{Y}_2)^2}{(J-1)\cdot Js^2_{\bar{Y}_1-\bar{Y}_2}}\\
F &= \frac{(\bar{Y}_1-\bar{Y}_2)^2}{s^2_{\bar{Y}_1-\bar{Y}_2}}\\
F &= t^2
\end{aligned}
$$

Question 5

(Based on Rice 12.21) During each of four experiments on the use of carbon tetrachloride as a worm killer, ten rats were infested with larvae. Eight days later, five rates were treated with carbon tetrachloride; the other five were kept as controls. After two more days, all the rats were killed and the numbers of worms were counted. The data.frame below contains the counts of worms for the four control groups:

Significant differences, although not expected, might be attributable to changes in experimental conditions. A finding of significant differences could result in more carefully controlled experimentation and thus greater precision in later work. Use both graphical techniques and the F test to test whether there are significant differences among the four groups. Use a nonparametric technique as well.

```r
worms = data.frame(group = as.factor(c(rep("Group I", 5), rep("Group II", 5),
                                      rep("Group III", 5), rep("Group IV", 5))),
                   count = c(279, 338,334,198,303,378,275,412,265,286,
                             172,335,335,282,250,381,346,340,471,318))

group_1 <- c(279, 338, 334, 198, 303)
group_2 <- c(378, 275, 412, 265, 286)
group_3 <- c(172, 335, 335, 282, 250)
group_4 <- c(381, 346, 340, 471, 318)

data <- c(group_1, group_2, group_3, group_4)
groups <- factor(rep(c("group I", "group II", "group III", "group IV"), each=5))

boxplot(data~groups)
```



F-test

```
summary(aov(count~group, worms))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        3  27234    9078   2.271  0.119
## Residuals   16  63954    3997
```

based on the result above, p-value = 0.119 > 0.05 so we cannot reject the null hypothesis.

kruskal-wallis test for non-parametric ones:

```
kruskal.test(worms$count, worms$group)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  worms$count and worms$group
## Kruskal-Wallis chi-squared = 6.2047, df = 3, p-value = 0.1021
```

based on the result above, p-value = 0.1021 > 0.05 so we cannot reject the null hypothesis.