STA305/1004 - Homework #1 Solutions

Nathan Taback

January 25, 2017

Question 1

(a)

The following set of equations models the weighing scheme.

$$
\begin{aligned}
y_1 &= \beta_1 x_{11} + \beta_2 x_{12} + \epsilon_1 \\
y_2 &= \beta_1 x_{21} + \beta_2 x_{22} + \epsilon_2 \\
y_3 &= \beta_1 x_{31} + \beta_2 x_{32} + \epsilon_3
\end{aligned}
$$

where,

$$
x_{ij} = \begin{cases} 1 & \text{if the } i^{th} \text{ measurement of the } j^{th} \text{ object is in the left pan} \\ -1 & \text{if the } i^{th} \text{ measurement of the } j^{th} \text{ object is in right pan,} \\ 0 & \text{if the } i^{th} \text{ measurement of the } j^{th} \text{ object is in neither pan.} \end{cases}
$$

$Var(\epsilon_i) = \sigma^2, i = 1, 2, 3, j = 1, 2.$

NB: The pan that's coded as 1 or -1 is arbitrary.

In this case the design indicates ceratin values for x_{ij}. So that the observed measurements y_i are related to the unknown weights β_i via the four equations:

$$
\begin{aligned}
y_1 &= \beta_1 + \beta_2 + \epsilon_1 \\
y_2 &= \beta_1 - \beta_2 + \epsilon_2 \\
y_3 &= \beta_1 + \epsilon_3
\end{aligned}
$$

In matrix form this could be written as $\mathbf{y} = X\beta + \epsilon$, where

$$
\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, X = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 0 \end{pmatrix} \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}.
$$

(b) The least squares estimates can be found using $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$. Using R we find $\left(X^T X\right)^{-1} X^T$ then multiply it by y.

```r
X <-rbind(c(1,1), # weigh two objects in one pan
          c(1,-1), # weigh one object in one pan and one object on another pan
          c(1,0)) # pick object 1 and weigh it (or pick object 2)
solve( t(X) %*% X ) %*% t(X)
```

```
          [,1]       [,2]      [,3]
[1,] 0.3333333  0.3333333 0.3333333
[2,] 0.5000000 -0.5000000 0.0000000
```

So

$$\begin{aligned}\hat{\beta}_1 &= (1/3)(y_1 + y_2 + y_3)\\ \hat{\beta}_2 &= (1/2)(y_1 - y_2 + 0y_3).\end{aligned}$$

(c)

The standard error of $\hat{\beta}$ is the square-root of the diagnal entries of the covariance matrix of $\hat{\beta}$, namely, $\left(X^T X\right)^{-1} \sigma^2$. The covariance matrix can be found using R.

```
X <-rbind(c(1,1),
          c(1,-1),
          c(1,0))
solve( t(X) %*% X )
```

```
          [,1] [,2]
[1,] 0.3333333  0.0
[2,] 0.0000000  0.5
```

The $se\left(\hat{\beta}_1\right) = \sigma/\sqrt{3}, se\left(\hat{\beta}_2\right) = \sigma/\sqrt{2}$.

(d) If she measured each object three times then a sensible way to combine the three measurements is to take their average $\bar{Y}$. The precision is the standard deviation of the average of $\bar{Y}$. Since each measurement has precision σ the standard deviation of the average is $\sqrt{Var\left(\bar{Y}\right)} = \sigma/\sqrt{3}$. So, the precision for the object that has weight β_1 is the same compared to the proposed design using a pan balance, but the precision for the object that has weight β_2 is greater than the design using the pan balance.

(e) In this case we have

$y = X\beta + \epsilon$, where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}, X = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \end{pmatrix} \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}.$$

So the least squares estimates are:

```
X <-rbind(c(1,1), #both objects in same pan measurement 1
          c(1,1), #both objects in same pan measurement 2
          c(1,-1)) #objects in opposite pans
solve( t(X) %*% X ) %*% t(X)
```

```
##      [,1] [,2] [,3]
## [1,] 0.25 0.25  0.5
## [2,] 0.25 0.25 -0.5
```

$$\begin{aligned}\hat{\beta}_1 &= (1/4)(y_1 + y_2 + 2y_3)\\ \hat{\beta}_2 &= (1/4)(y_1 + y_2 - 2y_3).\end{aligned}$$

NB: If you used the matrix

```
X <-rbind(c(1,-1), #both objects in same pan measurement 1
          c(1,1), #both objects in same pan measurement 2
          c(1,1)) #objects in opposite pans
solve( t(X) %*% X ) %*% t(X)
```

```
##      [,1] [,2] [,3]
## [1,]  0.5 0.25 0.25
## [2,] -0.5 0.25 0.25
```

the least-squares estimates would be:

$$\hat{\beta}_1 = (1/4)(2y_1 + y_2 + y_3)$$
$$\hat{\beta}_2 = (1/4)(-2y_1 + y_2 + y_3).$$

The standard errors can be found by calculating $\left(X^T X\right)^{-1} \sigma^2$.

```
X <-rbind(c(1,1),
          c(1,1),
          c(1,-1))
solve( t(X) %*% X )
```

```
       [,1]   [,2]
[1,]  0.375 -0.125
[2,] -0.125  0.375
```

The standard errors are: $se\left(\hat{\beta}_1\right) = se\left(\hat{\beta}_2\right) = \sqrt{\frac{3}{8}}\sigma$.

(f) In **DESIGN I** $se\left(\hat{\beta}_1\right) = 0.577\sigma, se\left(\hat{\beta}_2\right) = 0.707\sigma$. In **DESIGN II** $se\left(\hat{\beta}_1\right) = se\left(\hat{\beta}_2\right) = 0.612\sigma$. Therefore, **DESIGN I** is more precise for object 1 and less precise for object 2 compared to **DESIGN II**.

Question 2

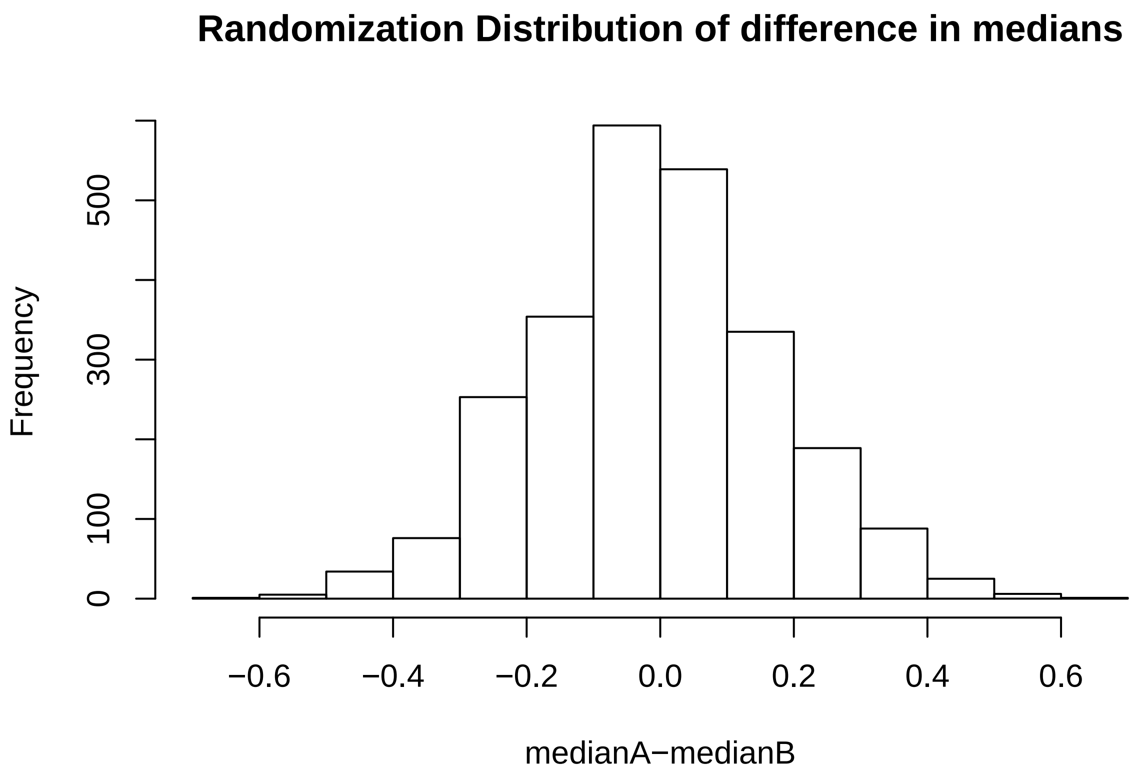
(a) The randomization distribution has $\binom{10000}{5000} > 10^{308}$ values. Many computers available today have a floating point range of $\pm 10^{308.25}$ and values greater than $\approx 10^{308}$ will be denoted by Infinity.

(b) Calculate a one-sided P-value to test if B has higher sales compared to A. Use Monte-Carlo simulation since the randomization distribution is too large to evaluate every possible difference.

```
attach(ABtest)
# I used 2500 to decrease the run time,
# but ideally you would use 250000 so SE
# of estimated p-value is small
N <- 2500
res <- numeric(N) # store the results

for (i in 1:N)
{
  index <- sample(length(sales),size=length(sales[page=="A"]),replace=F)
  res[i] <- median(sales[index])-median(sales[-index]) }

observed <- median(sales[page=="B"])-median(sales[page=="A"])

(sum(res >= observed)+1)/(N+1) # Randomization P-value
```

```
## [1] 0.0003998401
```

```
hist(res,xlab="medianA-medianB", main="Randomization Distribution of difference in medians")
```



(c) Yes there is significant statistical evidence that B leads to higher sales compared to A. The difference in median sales is $0.72 over a one-day period. Given the large number of visitors to the website this small gain may be practically significant. For example, if 10,000 visitors spent $0.72 more per day then over the period of one month the company would increase sales by $216,000. But, we would need more information to evaluate the practical significance of this result.

Question 3

(a) This is a randomized paired design. Treatments are randomly assigned to the two subplots within a plot.

(b) The probability is 1/2 since the randomization was done using a fair coin.

(c) The number of treatment allocations is 2^6. Therefore, each treatment alllocation has probability $\frac{1}{2^6}$. The observed treatment allocation has the same probability of occuring as the allocation where the first subplots all received fertilizer F. Under the null hypothesis that the yield for fertilzer F is the same as Fertilizer G each treatment allocation the randomization distribution consists of the $2^6 = 64$ arrangements of signs in:

$$\frac{\pm 6 \pm 12 \pm 27 \pm 20 \pm 8 \pm 5}{6}$$

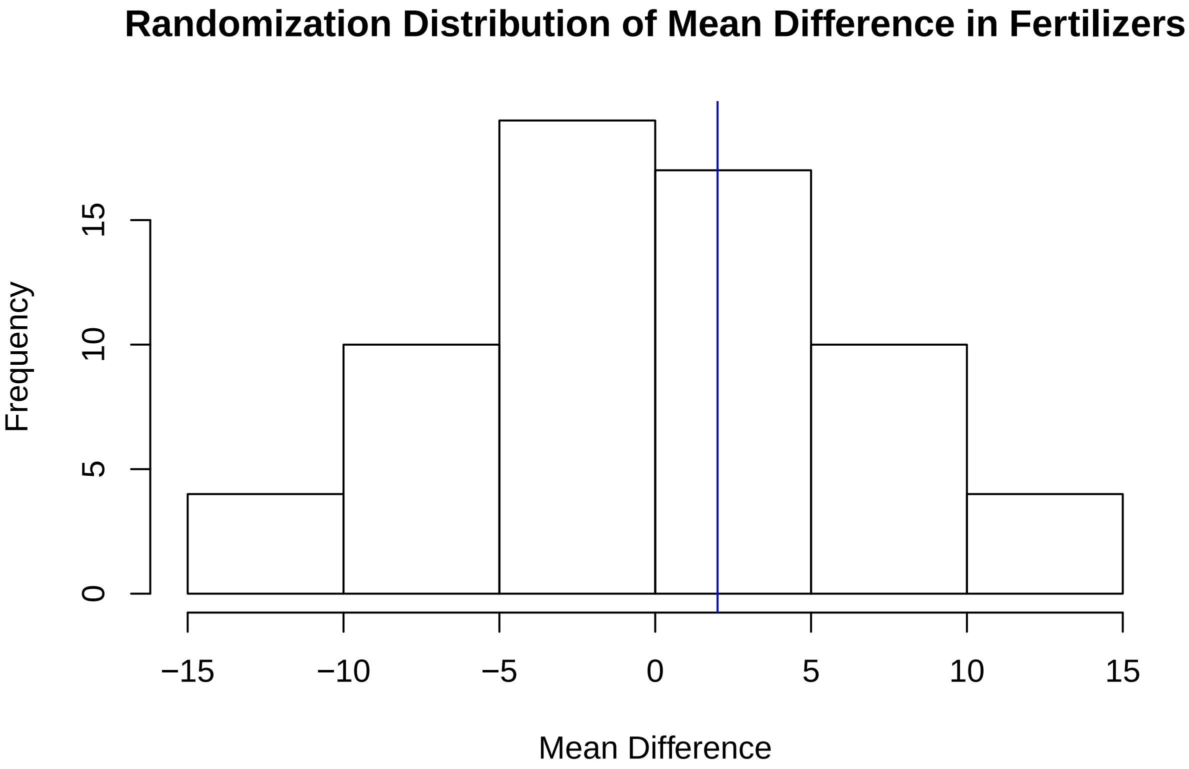
(d)

```
FertF <- c(78, 82, 82, 65, 51, 75)
FertG <- c(72, 70, 55, 85, 59, 80)
diff <- FertF-FertG
meandiff <- mean(diff) # the observed mean difference

N <- 2^(6) # number of treatment assignments
```

```r
res <- numeric(N) #vector to store results
LR <- list(c(-1,1)) # difference is multiplied by -1 or 1
trtassign <- expand.grid(rep(LR, 6)) # generate all possible treatment assign

for(i in 1:N){
  res[i] <- mean(as.numeric(trtassign[i,])*diff)
}

tbar <- mean(res) # mean of the randomization distribution
pval3d <- sum(abs(res-tbar)>=abs(meandiff-tbar))/N # the p-value
pval3d
```

```
## [1] 0.8125
```

```r
hist(res, xlab="Mean Difference",main="Randomization Distribution of Mean Difference in Fertilizers")
abline(v = meandiff,col="blue")
```

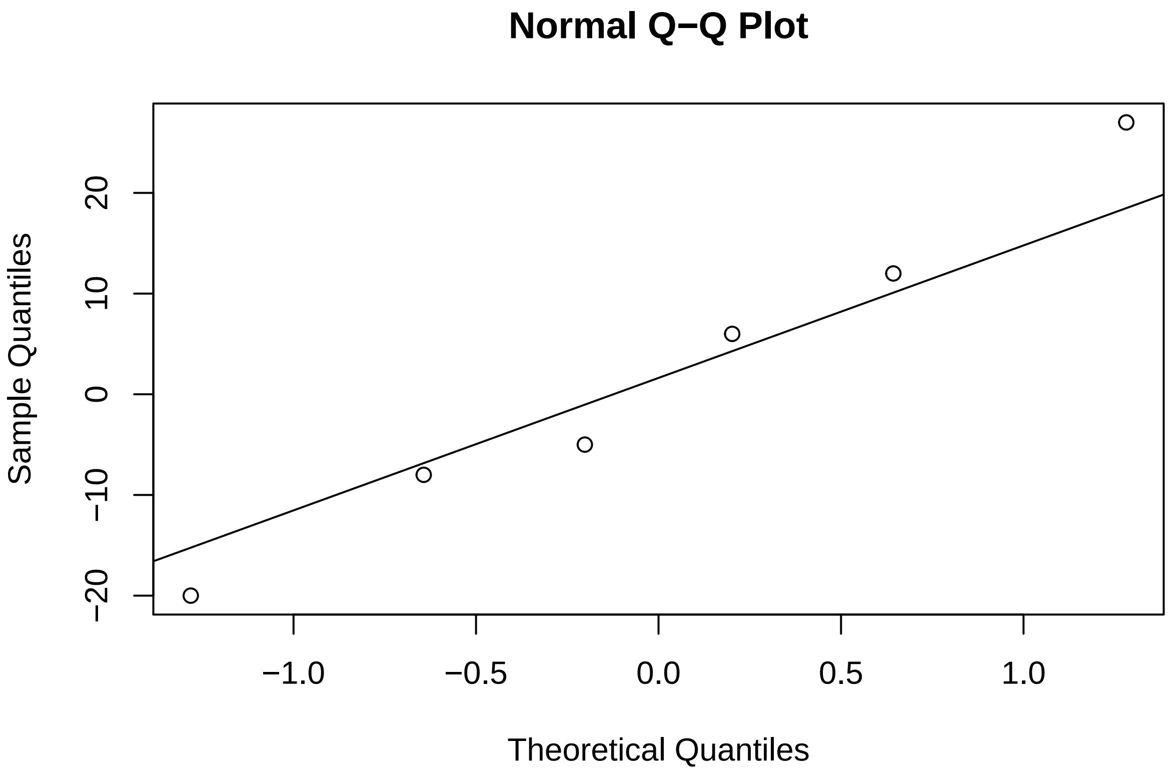


The large P-value of 0.8125 indicates that there is no evidence of a difference in mean yield between the two fertilizers.

Question 4

(a) We must assume that the differences in yield between the subplots are independent and that these differences are normally distributed. The latter can be checked using a normal quantile plot.

```r
FertF <- c(78, 82, 82, 65, 51, 75)
FertG <- c(72, 70, 55, 85, 59, 80)
diff <- FertF-FertG
qqnorm(diff);qqline(diff)
```



The plot indicates that the assumptions of normality for the differences is not violated.

(b) A paired t-test is conducted below.

```
t.test(FertF,FertG,paired = T)
```

```
##
##  Paired t-test
##
## data:  FertF and FertG
## t = 0.29553, df = 5, p-value = 0.7795
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -15.39659  19.39659
## sample estimates:
## mean of the differences
##                       2
```

The P-value for the paired t-test is close to the P-value for the randomziation test. Both tests indicate that there is no evidence of a mean difference between the fertilizers.