STA304 Assignment 1 Solutions

1. Recall that for the linear model $\boldsymbol{y} = X\beta + \epsilon$ the least squares estimate of β is $\hat{\beta} = (X'X)^{-1}X'\boldsymbol{y}$.

One mark removed for missing each X and $\hat{\beta}$. One mark removed for each incorrect variance, and one mark removed for a conclusion not based on relevant information.

a) The following R code generates X and $(X'X)^{-1}X'$:

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

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    1    1    1    1    1    1    1
## [2,]    1    1    1    0    0    0    0
## [3,]    1    0    0    1    1    0    0
## [4,]    1    0    0    0    0    1    1
## [5,]    0    1    0    1    0    1    0
## [6,]    0    1    0    0    1    0    1
## [7,]    0    0    1    1    0    0    1
## [8,]    0    0    1    0    1    1    0
```

```
##        [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]
## [1,] 0.0625  0.3125  0.3125  0.3125 -0.1875 -0.1875 -0.1875 -0.1875
## [2,] 0.0625  0.3125 -0.1875 -0.1875  0.3125  0.3125 -0.1875 -0.1875
## [3,] 0.0625  0.3125 -0.1875 -0.1875 -0.1875 -0.1875  0.3125  0.3125
## [4,] 0.0625 -0.1875  0.3125 -0.1875  0.3125 -0.1875  0.3125 -0.1875
## [5,] 0.0625 -0.1875  0.3125 -0.1875 -0.1875  0.3125 -0.1875  0.3125
## [6,] 0.0625 -0.1875 -0.1875  0.3125  0.3125 -0.1875 -0.1875  0.3125
## [7,] 0.0625 -0.1875 -0.1875  0.3125 -0.1875  0.3125  0.3125 -0.1875
```

The actual vector $\hat{\beta}$ is obtained by symbolic multiplication. For example, $\hat{\beta}_1 = (1/16)(y_1 + 5y_2 + 5y_3 + 5y_4 - 3y_5 - 3y_6 - 3y_7 - 3y_8)$.

b) The X matrix for Hotelling's procedure is Yates' X with the zeros replaced by -1:

```
Xh <- ifelse(X == 0,-1,1)
Xh; solve( t(Xh) %*% Xh ) %*% t(Xh)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7]
## [1,]    1    1    1    1    1    1    1
## [2,]    1    1    1   -1   -1   -1   -1
## [3,]    1   -1   -1    1    1   -1   -1
## [4,]    1   -1   -1   -1   -1    1    1
## [5,]   -1    1   -1    1   -1    1   -1
## [6,]   -1    1   -1   -1    1   -1    1
## [7,]   -1   -1    1    1   -1   -1    1
## [8,]   -1   -1    1   -1    1    1   -1
```

```
##       [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]
## [1,] 0.125  0.125  0.125  0.125 -0.125 -0.125 -0.125 -0.125
## [2,] 0.125  0.125 -0.125 -0.125  0.125  0.125 -0.125 -0.125
## [3,] 0.125  0.125 -0.125 -0.125 -0.125 -0.125  0.125  0.125
## [4,] 0.125 -0.125  0.125 -0.125  0.125 -0.125  0.125 -0.125
## [5,] 0.125 -0.125  0.125 -0.125 -0.125  0.125 -0.125  0.125
## [6,] 0.125 -0.125 -0.125  0.125  0.125 -0.125 -0.125  0.125
## [7,] 0.125 -0.125 -0.125  0.125 -0.125  0.125  0.125 -0.125
```

Note that $(X'X)^{-1} = 1/8\boldsymbol{I}$, so $(X'X)^{-1}X' = 1/8X'$. Again to compute $\hat{\beta}$, perform the relevant multiplication. For example, $\hat{\beta}_1 = (1/8)(y_1 + y_2 + y_3 + y_4 - y_5 - y_6 - y_7 - y_8)$.

c) Recall that $var(\hat{\beta}) = \sigma^2(X'X)^{-1}$:

```
#Yates:

solve( t(X) %*% X )
```

```
##         [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]
## [1,]  0.4375 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
## [2,] -0.0625  0.4375 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625
## [3,] -0.0625 -0.0625  0.4375 -0.0625 -0.0625 -0.0625 -0.0625
## [4,] -0.0625 -0.0625 -0.0625  0.4375 -0.0625 -0.0625 -0.0625
## [5,] -0.0625 -0.0625 -0.0625 -0.0625  0.4375 -0.0625 -0.0625
## [6,] -0.0625 -0.0625 -0.0625 -0.0625 -0.0625  0.4375 -0.0625
## [7,] -0.0625 -0.0625 -0.0625 -0.0625 -0.0625 -0.0625  0.4375
```

```
#Hotelling:

solve( t(Xh) %*% Xh )
```

```
##       [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
## [1,] 0.125 0.000 0.000 0.000 0.000 0.000 0.000
## [2,] 0.000 0.125 0.000 0.000 0.000 0.000 0.000
## [3,] 0.000 0.000 0.125 0.000 0.000 0.000 0.000
## [4,] 0.000 0.000 0.000 0.125 0.000 0.000 0.000
## [5,] 0.000 0.000 0.000 0.000 0.125 0.000 0.000
## [6,] 0.000 0.000 0.000 0.000 0.000 0.125 0.000
## [7,] 0.000 0.000 0.000 0.000 0.000 0.000 0.125
```

The variance of an estimated weight is the appropriate diagonal element of the above matrix multiplied by σ^2.

d) Pick the procedure with the lowest variance- Hotelling. Note that because Yates' estimates are negatively correlated, if you were interested in estimating some linear combination of the weights, you may actually acheive lower variance with his method.

2.

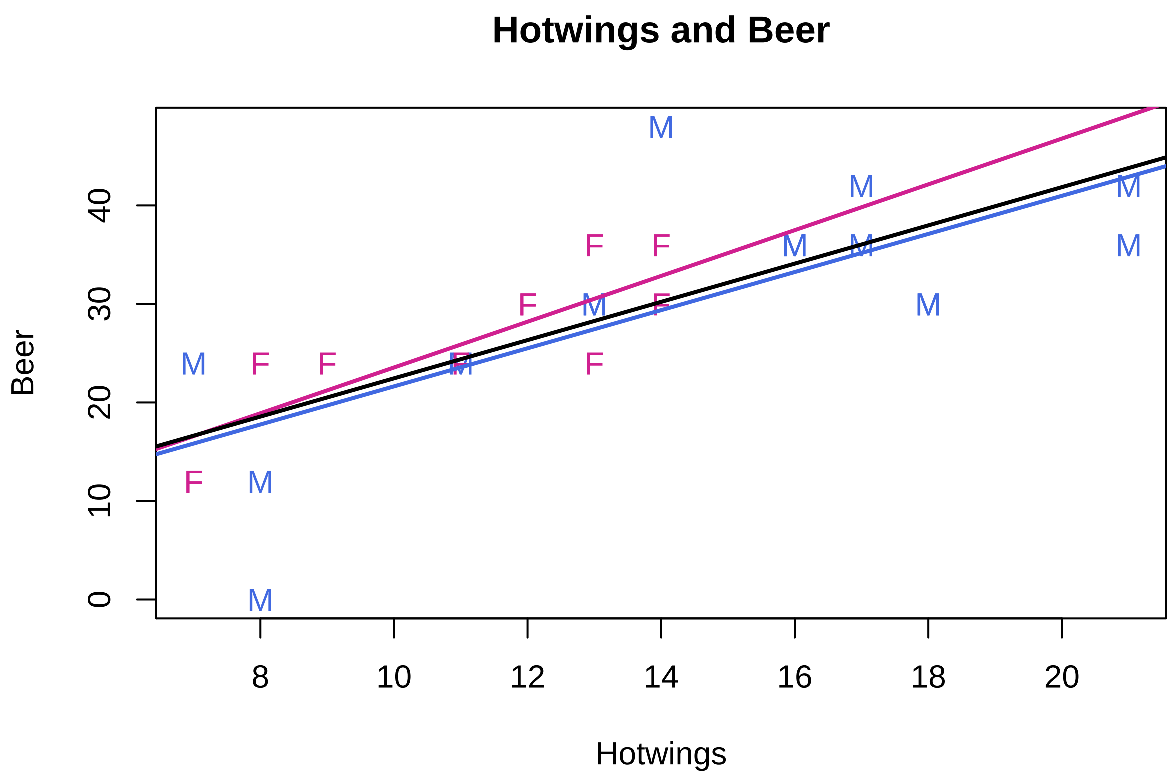
One mark for relevant conclusion, two marks for assumptions. One mark for correct number of possible assignments. Two marks for histogram, one for one-sided p-value. One mark for indicating both observational study and not appropriate to generalize to population. No mark given if the connection between these was not made.

a) Here is one possible analysis:

```
datta <- read.table(file.choose(),header=TRUE,sep=',')
datta <- datta[,-1]
head(datta)
```

```
##   Hotwings Beer Gender
## 1        4   24      F
## 2        5    0      F
## 3        5   12      F
## 4        6   12      F
## 5        7   12      F
## 6        7   12      F
```

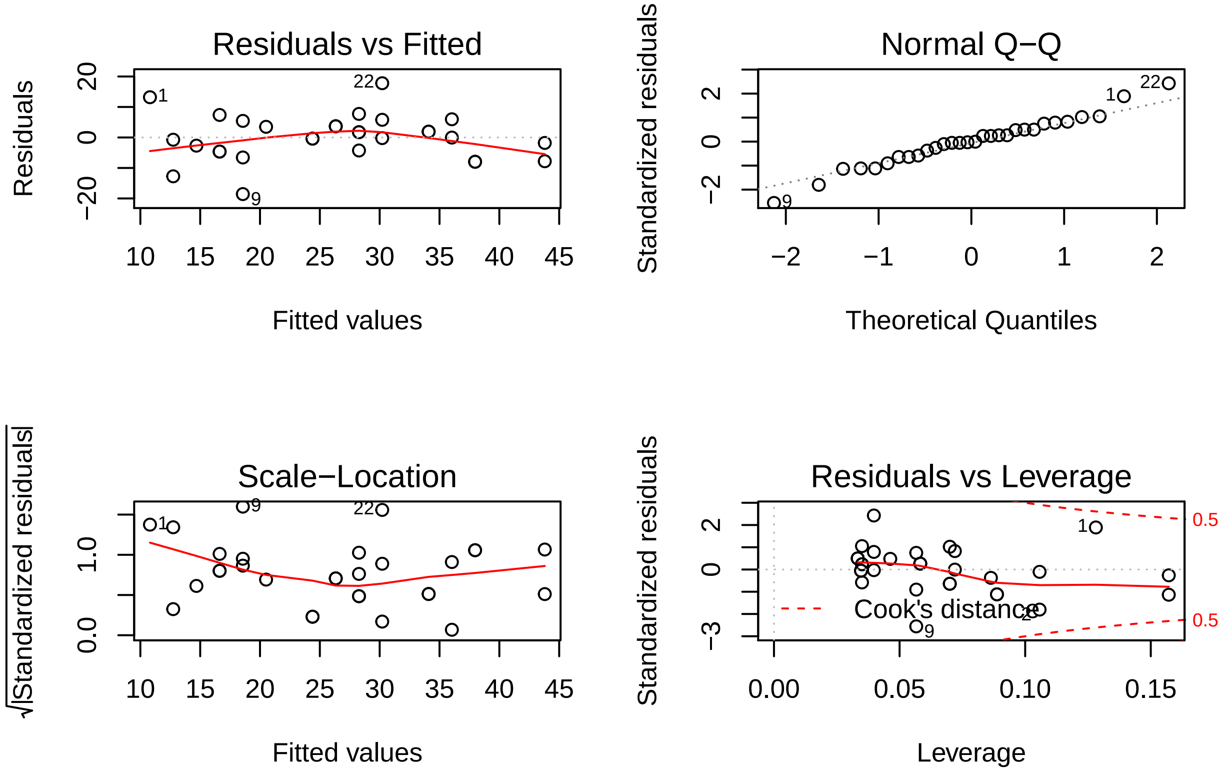
```
modM <- lm(Beer~Hotwings,data=datta[datta$Gender=='M',])
modF <- lm(Beer~Hotwings,data=datta[datta$Gender=='F',])
modT <- lm(Beer~Hotwings,data=datta)

with(datta,
    plot(Hotwings[Gender=='M'],Beer[Gender=='M'],pch='M',
        col='royalblue',xlab='Hotwings',ylab='Beer',main='Hotwings and Beer')
    )
with(datta,
    points(Hotwings[Gender=='F'],Beer[Gender=='F'],pch='F',col='violetred')
    )
abline(modM,col='royalblue',lwd=2)
abline(modF,col='violetred',lwd=2)
abline(modT,col='black',lwd=2)
```



```
summary(modT)
```

```
##
## Call:
## lm(formula = Beer ~ Hotwings, data = datta)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.566  -4.537  -0.122   3.671  17.789
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0404     3.7235   0.817    0.421
## Hotwings      1.9408     0.2903   6.686 2.95e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.479 on 28 degrees of freedom
## Multiple R-squared:  0.6148, Adjusted R-squared:  0.6011
## F-statistic:  44.7 on 1 and 28 DF,  p-value: 2.953e-07
```

```
par(mfrow=c(2,2))
plot(modT)
```



The assumptions of linearity and residuals being normally distributed with constant variance appear to hold. We may conclude that on average an increase of one hotwing corresponds to an increase of 1.94 units of beer, and this effect is not due to chance ($p \approx 0$).

b) i) There are $\binom{30}{15} = 155117520$ different assignments

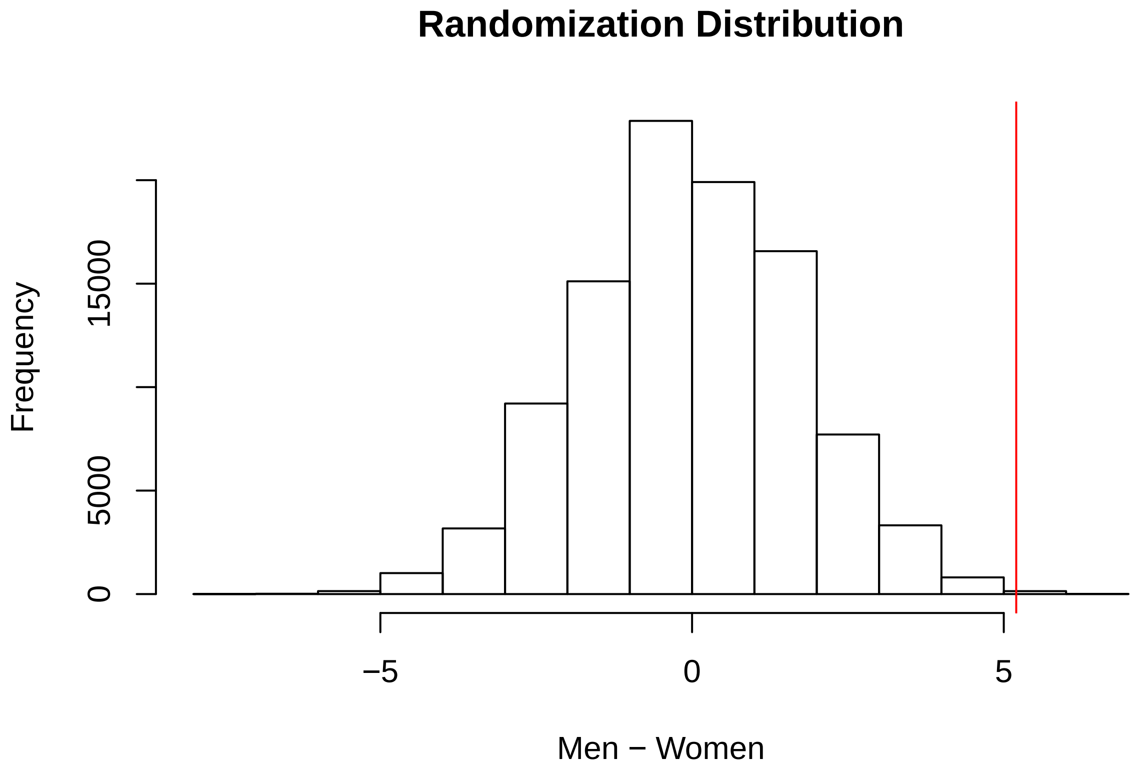
ii)

```
#Randomization test

men <- datta$Hotwings[datta$Gender=='M']
women <- datta$Hotwings[datta$Gender=='F']

wings <- c(men,women)

N <- 1e05-1
result <- numeric(N)
set.seed(72924)

for (i in 1:N)
{
    index <- sample(length(wings),size=length(men),replace=FALSE)
    result[i] <- mean(wings[index] - wings[-index])
}

observed <- mean(men - women)
```

```
#P-value - results will vary
( pval <- (sum(result >= observed)+1)/(N+1) )
```

```
## [1] 0.00112
```

```
hist(result,xlab='Men - Women',main='Randomization Distribution')
abline(v=observed,col='red')
```



iii) This is an observational study since it was not possible to randomly assign the treatment (gender) to the units (people). It is not appropriate to make a causal inference about the population.

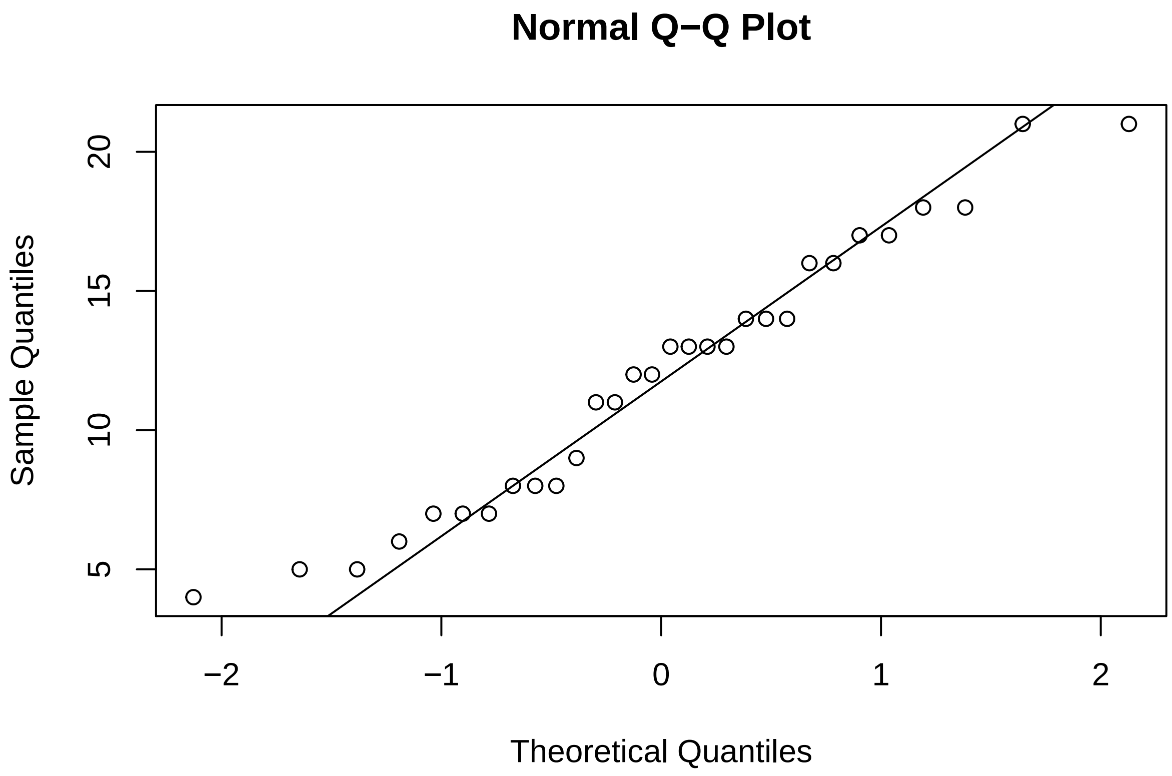
3.

Two marks for assumptions, one for test. One for correctly identifying whether the test agreed with the randomization test.

a) Assumptions:

- Independence: we don't know much about how she took these samples, have to hope they are roughly independent
- Normally Distributed:

```r
with(datta,qqnorm(Hotwings))
with(datta,qqline(Hotwings))
```



- Equal population variances: we don't need to assume this, but we should check:

```r
with(datta,sd(Hotwings[Gender=='M']))
```

```
## [1] 4.501851
```

```r
with(datta,sd(Hotwings[Gender=='F']))
```

```
## [1] 3.559026
```

The sample standard deviations are close enough (rule of thumb: ratio is less than 2) that we can assume the population variances are equal.

```r
with(datta,
    t.test(Hotwings[Gender=='M'],Hotwings[Gender=='F'],var.equal=TRUE)
)
```

```
##
##  Two Sample t-test
##
## data:  Hotwings[Gender == "M"] and Hotwings[Gender == "F"]
## t = 3.5094, df = 28, p-value = 0.001538
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.164792 8.235208
## sample estimates:
## mean of x mean of y
## 14.533333  9.333333
```

b) Yes, they do. There is strong evidence of a population difference in means, and both tests are able to detect it.

4.

One mark for probability. Three marks for correct table and design identification.

a)

Left	A	A	B	A	B	A	A	A	B	A	A	B
Right	B	B	A	B	A	B	B	B	A	B	B	A

b) There are 8 tails and 4 heads, in a specific order. The probability of this is $(0.5)^8(0.5)^4 = 0.000244$

c) This is a randomized paired design, with drug as treatment paired within subjects. That is, each individual subject receives both drugs.