Design of Scientific Studies - Notes on Week 2

Nathan Taback

1 Comparing Two Treatments

- Is fertilizer A better than fertilzer B for growing wheat?
- Is drug A better than drug B for treating breast cancer?
- Is webpage A better than webpage B for selling a certain product?

These are all examples of comparing two *treatments*. In experimental design *treatments* are different procedures applied to *experimental units* - the things to which we apply *treatments*.

In the first example the treatments are two fertilizers and the experimental units might be plots of land. In the second example the treatments are two different drugs to treat breast cancer and the experimental units are breast cancer patients. In the third example the treatments are two webpage designs and the experimental units are potential customers (example from Google experiments (https://support.google.com/analytics/answer/1745147?hl=en)).

2 Randomizing Two Treatments to Experimental Units

In randomized experiments (pg. 20, Imbens and Rubin, 2015):

> "… the assignment mechanism is under the control of the experimenter, and the probability of any assignment of treatments across the units in the experiment is entirely knowable before the experiment begins."

Suppose, for example, that we have two breast cancer patients and we want to randomly assign these two patients to two treatments (A and B). Then how many ways can this be done?

1. patient 1 receives A and patient 2 receives A
2. patient 1 receives A and patient 2 receives B
3. patient 1 receives B and patient 2 receives A
4. patient 1 receives B and patient 2 receives B

There are 4 possible treatment assignments. The probability of a treatment assignment is 1/4, although the probability that an individual patient receives treatment A (or B) is 1/2. In general, if there are N experimental units then there are 2^N possible treatment assignments (provided there are two treatments).

A treatment assignment vector records the treatmemnt that each experimental unit is assigned to receive. If $N = 2$ then the possible treatment assignment vectors are:

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where 1= treatment A, and 0=treatment B. The first treatment assignment vector

$$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

means that the first experimental unit receives treatment A, and the second treatment B. The third treatment assignment vector

$$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

means that the first and second experimentals units both receive treatment A.

It wouldn't be a very imformative expriment if both patients received A or both received B. Therefore, it makes sense to rule out this scenario. If we rule out this scenario then we want to assign treatments to patients such that one patient receives A and the other receives B. So, the possibile treatment assignments are:

1. patient 1 receives A and patient 2 receives B or (in vector notation) $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$.
2. patient 1 receives B and patient 2 receives A or (in vector notation) $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

There are are two possible treatment assignments. The probability of a treatment assignment is 1/2, and the probability that an individual patient receives treatment A (or B) is still 1/2.

Notice that the probability of a treatment assignment is different from the probability that of an individual experimental unit receiving a treatment.

3 Completely Randomized Experiment

A completely randomized experiment has the number of units assigned to treatment A, N_A fixed in advance so that the number of units assigned to treatment B $N_B = N - N_A$ are also fixed in advance . In such a design, N_A units are randomly selected, from a population of N units, to receive the treatment A, with the remaining N_B assigned to the other treatment. In this case, each unit has probability N_A/N of being assigned to treatment A. There are $\binom{N}{N_A}$ distinct values of the treatment assignment vector with N_A units out of N assigned to treatment A. Therefore, the probability of any particular treatment assignment is:

$$\frac{1}{\binom{N}{N_A}}.$$

Is fertilizer A better than fertilzer B for growing wheat? It is decided to take one large plot of land and divide it into twelve smaller plots of land then treat some plots with fertilizer A or B. How should we assign fertilizers (*treatments*) to plots of land?

plot 1	plot 2	plot 3	plot 4	plot 5	plot 6
plot 7	plot 8	plot 9	plot 10	plot 11	plot 12

Some of the plots get more sunlight and not all the plots have the exact same soil composition which may affect wheat yield. In other words, the plots are not identical. Nevertheless, we want to make sure that we can identify the treatment effect even though the plots are not identical. Statisticians sometimes state this as being able to identify the treatment effect (*viz.* difference between fertilizers) in the presence of other sources of variation (*viz.* differences between plots).

Ideally we would assign fertilizer A to six plots and fertilizer B to six plots. How can this be done so that each plot has an equal chance of being assigned fertilizer A or B? One way to assign the two fertilizers to the plots is to use six playing cards labelled A (for fertilizer A) cards and six playing cards labelled B (for fertilizer B), shuffle the cards then assign the first card to plot 1, the second card to plot 2, etc.

In R we can represent the Red and Black cards as:

```
cards <- c(rep("A",6),rep("B",6))
cards # print cards
```

```
##  [1] "A" "A" "A" "A" "A" "A" "B" "B" "B" "B" "B" "B"
```

Now, to “shuffle” the “cards” we can use `sample()`

```
shuffle <- sample(cards,12)
shuffle
```

```
##  [1] "B" "B" "B" "B" "A" "B" "A" "A" "A" "B" "A" "A"
```

The first plot will be assigned B, the second plot will be assigned B, etc.

Exercise:

1. How many ways are there to assign six plots to fertilizer A and six plots to fertilizer B? In other words how many different treatment assignments are possible?
2. What is the probability that an individual plot receives fertilizer A?
3. What is the probability of choosing the treatment assignment A, A, A, A, A, A, B, B, B, B, B, B?

Answers:

1. $\binom{12}{6} = 924$. That is, there are 924 unique subsets of 6 plots that can be chosen from 12 plots. In R we can choose six plots from 12 using the `sample()` command.

```
sample(1:12,6)
```

```
## [1] 10  4  2 11  1  6
```

2. 1/2.
3. $P(\text{treatment assignment}) = \frac{1}{\binom{12}{6}} = 0.001$.

4 The Randomization Distribution

Let's consider the fertilzer example from the previous section. The treatment assignment that the experimenter used was

```
shuffle
```

```
##  [1] "B" "B" "B" "B" "A" "B" "A" "A" "A" "B" "A" "A"
```

A	B	B	B	B	A
A	A	A	B	A	B

This is one of the $\binom{12}{6} = 924$ possible ways of allocating 6 A's and 6 B's to the 12 plots. The probability of choosing any of these allocations is $\frac{1}{\binom{12}{6}} = 0.001$.

The data from this experiment is:

A(11.4)	B(26.9)	B(26.6)	B(25.3)	B(28.5)	A(23.7)
A(17.9)	A(16.5)	A(21.1)	B(14.2)	A(19.6)	B(24.3)

This can be stored in R. A summary of the distributions of the two samples is given below.

```
#Fertilizer data
yA <- c(11.4,23.7,17.9,16.5,21.1,19.6)
summary(yA); sd(yA)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.40   16.85   18.75   18.37   20.72   23.70
```

```
## [1] 4.234934
```

```
yB <- c(26.9,26.6,25.3,28.5,14.2,24.3)
summary(yB); sd(yB)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.20   24.55   25.95   24.30   26.82   28.50
```

```
## [1] 5.151699
```

```
mean(yA)-mean(yB)
```

```
## [1] -5.933333
```

The distributions of the two samples can also be described by the empirical cumulative distribution function (CDF):

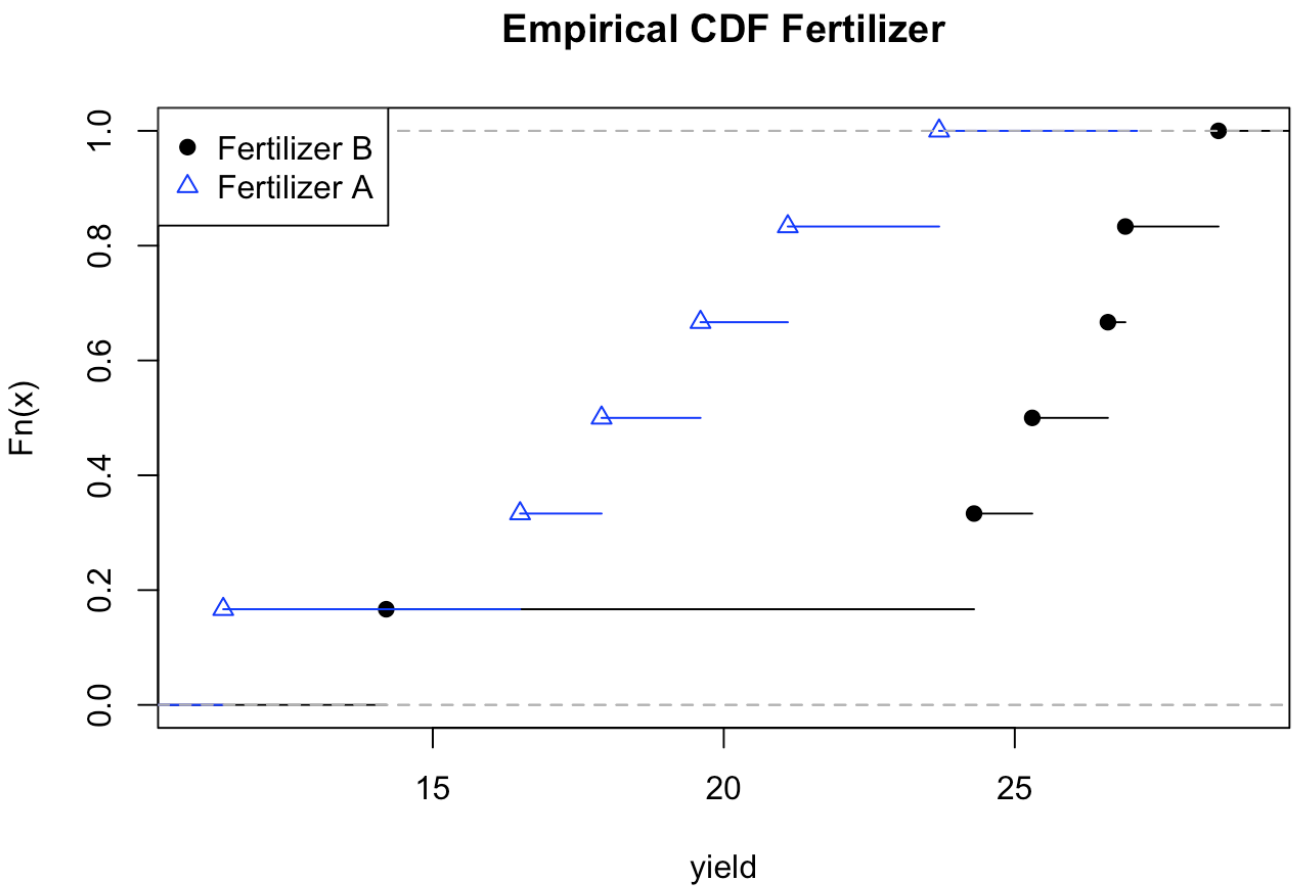
$$\hat{F}(y) = \frac{\sum_{i=1}^{n} I(y_i \leq y)}{n},$$

where n is the number of sample points and $I(\cdot)$ is the indicator function

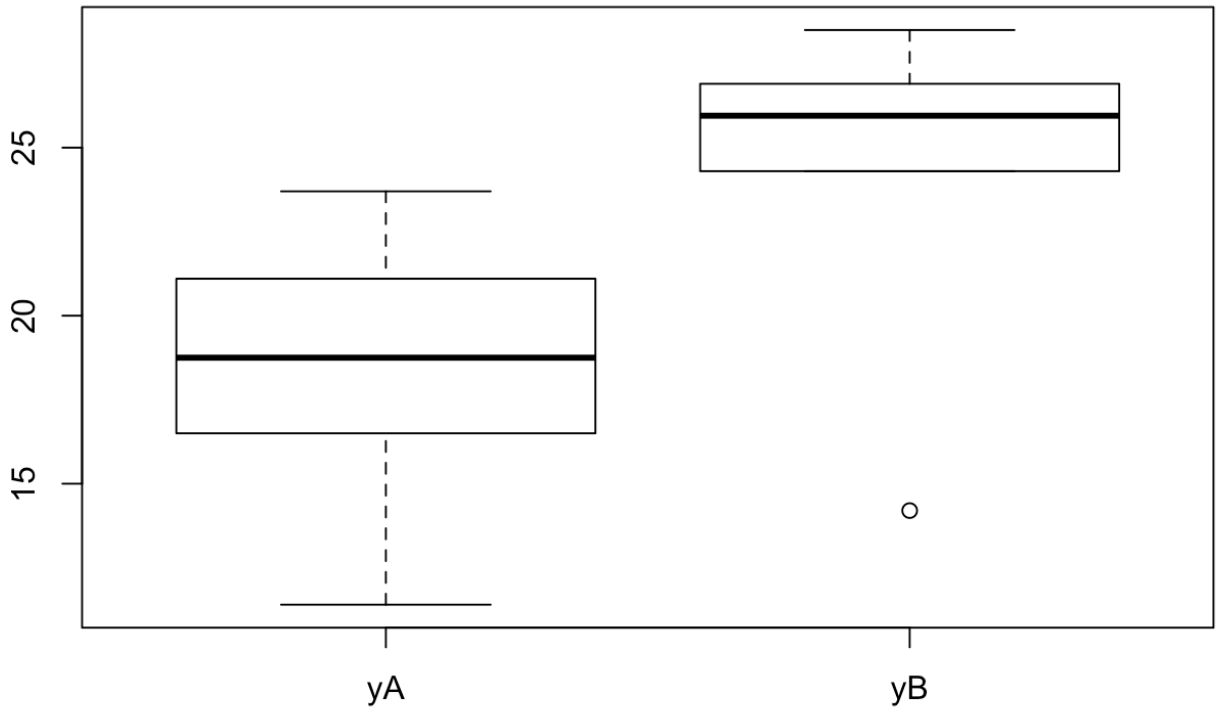
$$I(y_i \leq y) = \begin{cases} 1 & \text{if } y_i \leq y \\ 0 & \text{if } y_i > y \end{cases}$$

The R function `ecdf()` calculates the empirical CDF.

```
#plot empirical cdf for fertilizer B
plot.ecdf(yB,xlab="yield",xlim=c(11,29),main="Empirical CDF Fertilizer")
#add empirical cdf for fertilizer A to plot
plot.ecdf(yA,col="blue",pch=2,add=T)
# add legend
legend("topleft",legend=c("Fertilizer B","Fertilizer A"),col=c("black","b
lue"),pch=c(19,2))
```




```
wheatdat <- stack(data.frame(yA,yB)) #stack the columns
boxplot(values~ind,data=wheatdat) # create side-by-side boxplot
```



Is the difference in wheat yield due to the treatment or due to chance?

- Assume that there is no difference in the average yield between fertilzer A and fertilizer B.
- If there is no difference then the yield would be the same even if a different treatment allocation occured.
- Under this assumption of no difference between the treatments, if one of the other 924 treatment allocations occured such as B, B, B, B, A, B, A, A, A, B, A, A. Then the data from the experiment would have been:

B(11.4)	A(26.9)	B(26.6)	B(25.3)	B(28.5)	A(23.7)
A(17.9)	A(16.5)	A(21.1)	B(14.2)	A(19.6)	B(24.3)

```
#Fertilizer data
yA_1 <- c(26.9,23.7,17.9,16.5,21.1,19.6)
mean(yA_1); sd(yA_1)
```

```
## [1] 20.95
```

```
## [1] 3.844867
```

```
yB_1 <- c(11.4,26.6,25.3,28.5,14.2,24.3)
mean(yB_1); sd(yB_1)
```

```
## [1] 21.71667
```

```
## [1] 7.103638
```

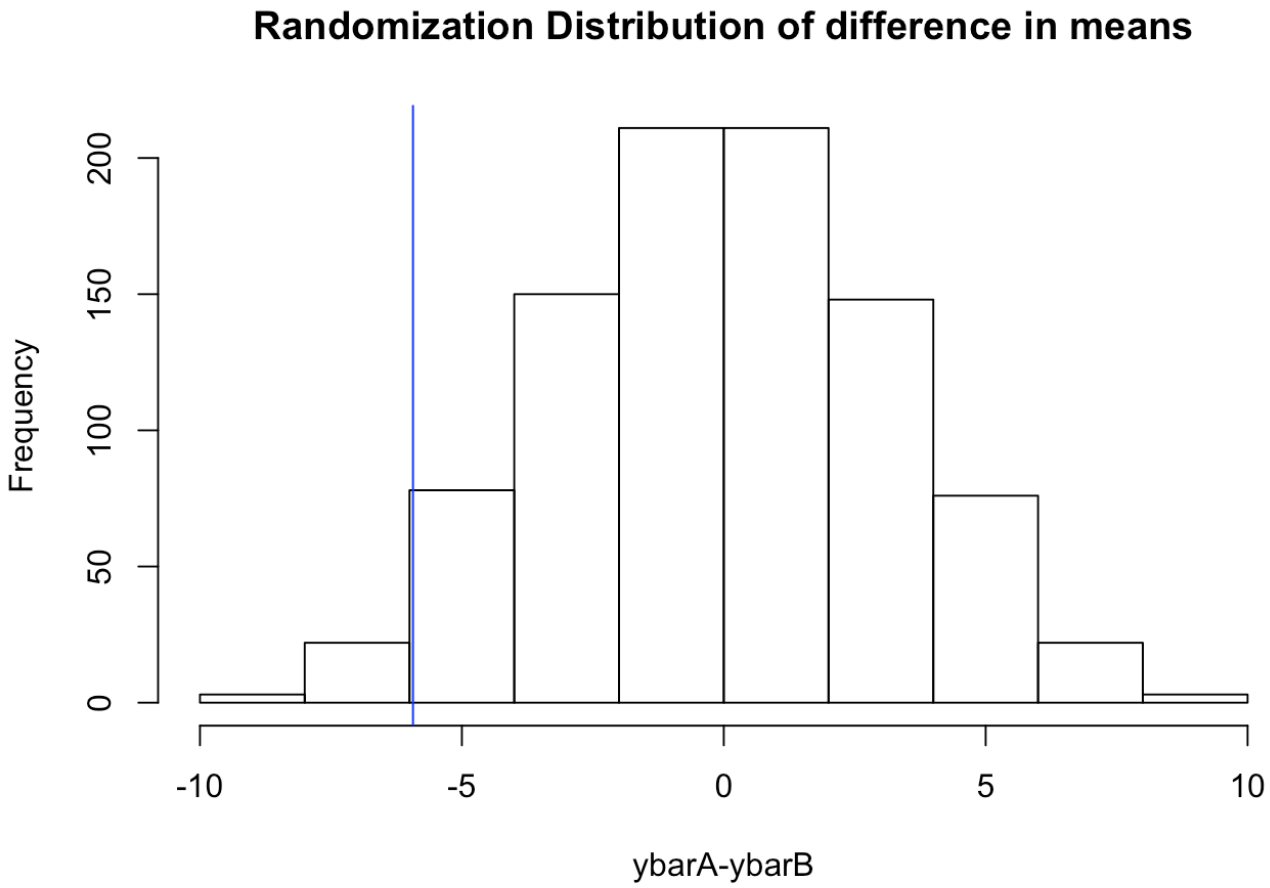
```
mean(yA_1)-mean(yB_1)
```

```
## [1] -0.7666667
```

Assume that there is no difference between the two treatments. The set of all possible differences that might have occured if a different treatment allocation was chosen is called the randomization distribution.

The randomization distribution can be obtained in R using the following code.

```
fert <- c(yA,yB) #pool data
N <- choose(12,6)
res <- numeric(N) # store the results
#install.packages("combinat") # if package not installed then remove comm
ent
library(combinat)
index <-combn(1:12,6) # Generate N treatment assignments
for (i in 1:N)
{
  res[i] <- mean(fert[index[,i]])-mean(fert[-index[,i]])
}
hist(res,xlab="ybarA-ybarB", main="Randomization Distribution of differen
ce in means")
observed <- mean(yA)-mean(yB) #store observed mean difference
abline(v=observed,col="blue") #add line at observed mean diff
```



The researchers are intrested in determining if fertilizer B produces a higher yield compared to fertilizer A.

The null and alternative hypotheses of interest are

$$H_0 : \text{There is no difference between treatments,}$$

$$H_1 : \text{Fertilizer B increases wheat yield.}$$

4.0.1 The Randomization p-value

The p value of the *randomization test* of H_0 can be calculated as the probability of getting a test statistic as extreme of more extreme than the observed value of the test statistic t^*. Since all of the $\binom{N}{N_A}$ randomizations are equally likely under H_0, the p value is

$$P(T \leq t^* | H_0) = \sum_{i=1}^{\binom{N}{N_A}} \frac{I(t_i \leq t^*)}{\binom{N}{N_A}},$$

where t_i is the value of the test statistic $T = \bar{Y}_A - \bar{Y}_B$ for the i^{th} randomization (Ernst, 2004).

The observed value of the test statistic is -5.93. So, the p-value is

```
# of times values from the mean randomization distribution less than obse
rved value
sum(res<=observed)
```

```
## [1] 26
```

```
N # Number of randomizations
```

```
## [1] 924
```

```
pval <- sum(res<=observed)/N # Randomization p value
round(pval,2)
```

```
## [1] 0.03
```

A p-value of 0.03 can be interpreted as: assume there is no difference in yield between fertilizers A and B then the proportion of randomizations that would produce an observed mean difference between A and B of at most -5.93 is 0.03. In other words, under the assumption that there is no difference between A and B only 3% of randomizations would produce an extreme or more extreme difference than the observed mean difference.

4.0.2 Two-Sided Randomization P value

If we are using a two-sided alternative then how do we calculate a p-value? The randomization distribution may not be symmetric so there is no justifcation for simply doubling the probability in one tail.

Let

$$\bar{t} = \frac{1}{\binom{N}{N_A}} \sum_{i=1}^{\binom{N}{N_A}} t_i$$

be the mean of the randomization distribution then we can define the two-sided p-value as

$$P(|T - \bar{t}| \geq |t^* - \bar{t}| \, | H_0) = \sum_{i=1}^{\binom{N}{N_A}} \frac{I(|t_i - \bar{t}| \geq |t^* - \bar{t}|)}{\binom{N}{N_A}},$$

this is the probability of obtaining an observed value of the test statistic as far, or farther, from the mean of the randomization distribution.

In R this can be calculated

```
yA <- c(11.4,23.7,17.9,16.5,21.1,19.6)
yB <- c(26.9,26.6,25.3,28.5,14.2,24.3)
fert <- c(yA,yB) #pool data
N <- choose(12,6)
res <- numeric(N) # store the results
#install.packages("combinat") # if package not installed then remove comm
ent
library(combinat)
```

```
##
## Attaching package: 'combinat'
```

```
## The following object is masked from 'package:utils':
##
##     combn
```

```
index <-combn(1:12,6)
for (i in 1:N)
{
  res[i] <- mean(fert[index[,i]])-mean(fert[-index[,i]])
}

tbar <- mean(res)
pval <- sum(abs(res-tbar)>=abs(observed-tbar))/N
round(pval,2)
```

```
## [1] 0.06
```

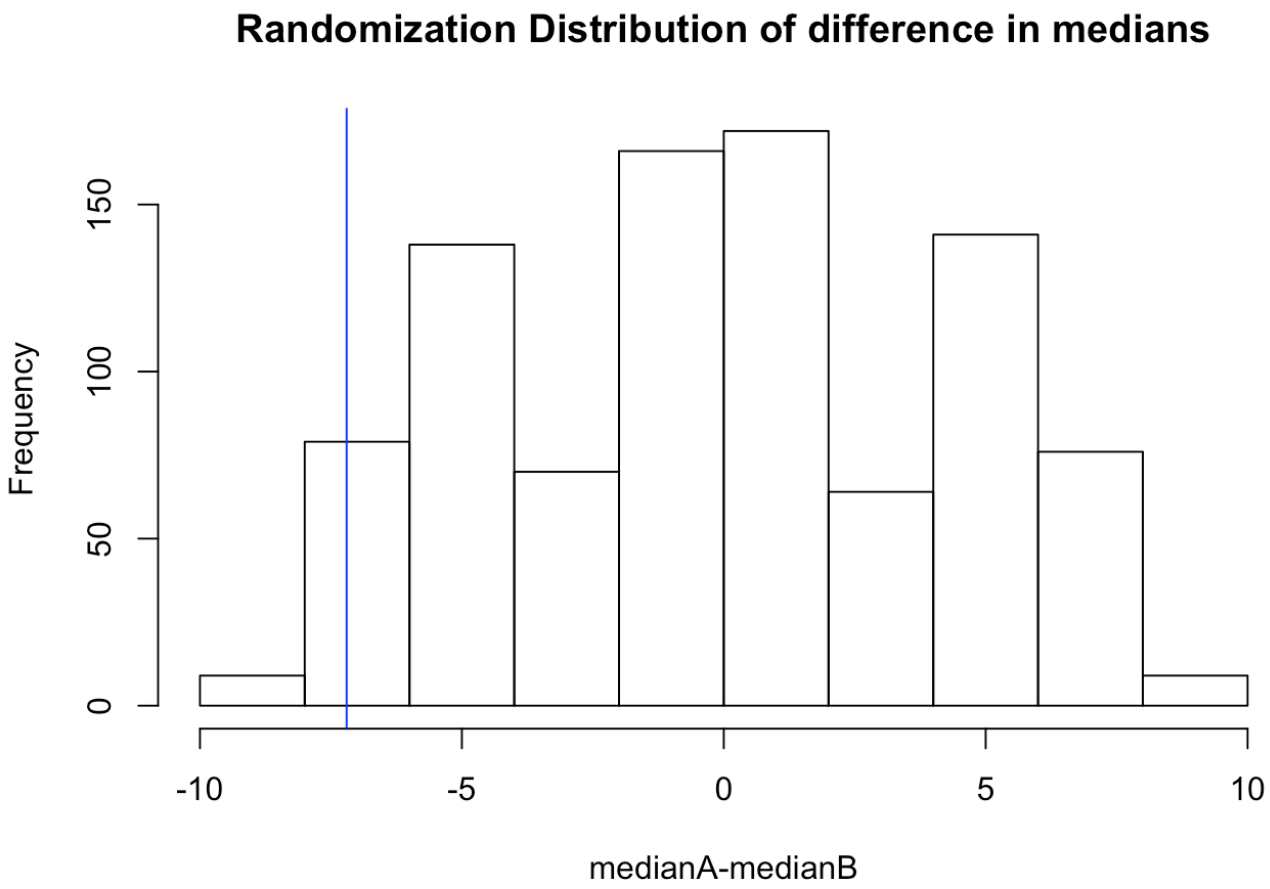
In this case, since the randomization distribution, is roughly symmetric the two-sided p-value is approximately half the one-sided p-value.

4.0.3 Other Test Statistics

Other test statistics could be used instead of $T = \bar{Y}_A - \bar{Y}_B$ to measure the effectiveness of fertilizer A. The difference in group medians or trimmed means are examples of other test statistics.

The randomiztion distribution of the difference in group medians can be obtained by modifying the R code used for the difference in group means.

```
fert <- c(yA,yB) #pool data
N <- choose(12,6)
res <- numeric(N) # store the results
#install.packages("combinat") # if package not installed then remove comm
ent
library(combinat)
index <-combn(1:12,6) # Generate N treatment assignments
for (i in 1:N)
{
  res[i] <- median(fert[index[,i]])-median(fert[-index[,i]])
}
hist(res,xlab="medianA-medianB", main="Randomization Distribution of diff
erence in medians")
observed <- median(yA)-median(yB) #store observed median difference
abline(v=observed,col="blue") #add line at observed median diff
```



The p-value of the randomization test can be calculated

```
# of times values from the median randomization distribution less than ob
served value
sum(res<=observed)
```

```
## [1] 36
```

```
N # Number of randomizations
```

```
## [1] 924
```

```
pval <- sum(res<=observed)/N # Randomization p value
round(pval,2)
```

```
## [1] 0.04
```

5 Monte Carlo Sampling

Computation of the randomization distributon involves calculating the difference in means for every possible way to split the data into two samples of size n_A each. If $N = 30$ and $N_A = 15$ this would result in $\binom{30}{15} = 155.12$ million differences. These types of calculations are not practical unless the sample size is small.

Instead we can resort to Monte Carlo sampling from the randomization distribution to estimate the exact p-value. The p-value is the proportion of test statistics as extreme or more extreme than the observed value.

The data set can be randomly divided into two groups and the test statistic calculated. Several thousand test statistics are usually sufficient to get an accurate estimate of the exact p-value and sampling can be done without replacement.

If M test statistics, $t_i, i = 1, \ldots, M$ are randomly sampled from the permutation distribution, a one-sided Monte Carlo p value for a test of $H_0 : \mu_T = 0$ versus $H_1 : \mu_T > 0$ is

$$\hat{p} = \frac{1 + \sum_{i=1}^{M} I(t_i \geq t^*)}{M + 1}.$$

Including the observed value t^* there are $M + 1$ test statistics.

A student conducted a study of hot chicken wings and beer consumption at a bar in Minneapolis (adapted from Chichara, Hesterberg, 2011). She decided to recruit 30 participants and randomly assign an equal number to two groups F and M. The M group was given a 2 minute description about the quality of the food served at the bar and a coupon for one free beer next time they visit the bar. The F group entered the bar as a regular customer. She asked patrons at the bar to record their consumption of hot wings and beer over the course of several hours. She wanted to know if the promotion had an impact on hot wings or beer consumption. The data are below:

```
##    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
## 8         8   24      F
## 11        9   24      F
## 12       11   24      F
## 14       12   30      F
## 15       12   30      F
## 16       13   24      F
## 17       13   36      F
## 20       14   30      F
## 21       14   36      F
## 7         7   24      M
## 9         8    0      M
## 10        8   12      M
## 13       11   24      M
## 18       13   30      M
## 19       13   30      M
## 22       14   48      M
## 23       16   36      M
## 24       16   36      M
## 25       17   36      M
## 26       17   42      M
## 27       18   30      M
## 28       18   30      M
## 29       21   36      M
## 30       21   42      M
```

Did the M group consume more chicken wings than the F group?

```
table(Beerwings$Gender) # the numbers of males and females
```

```
##
##  F  M
## 15 15
```

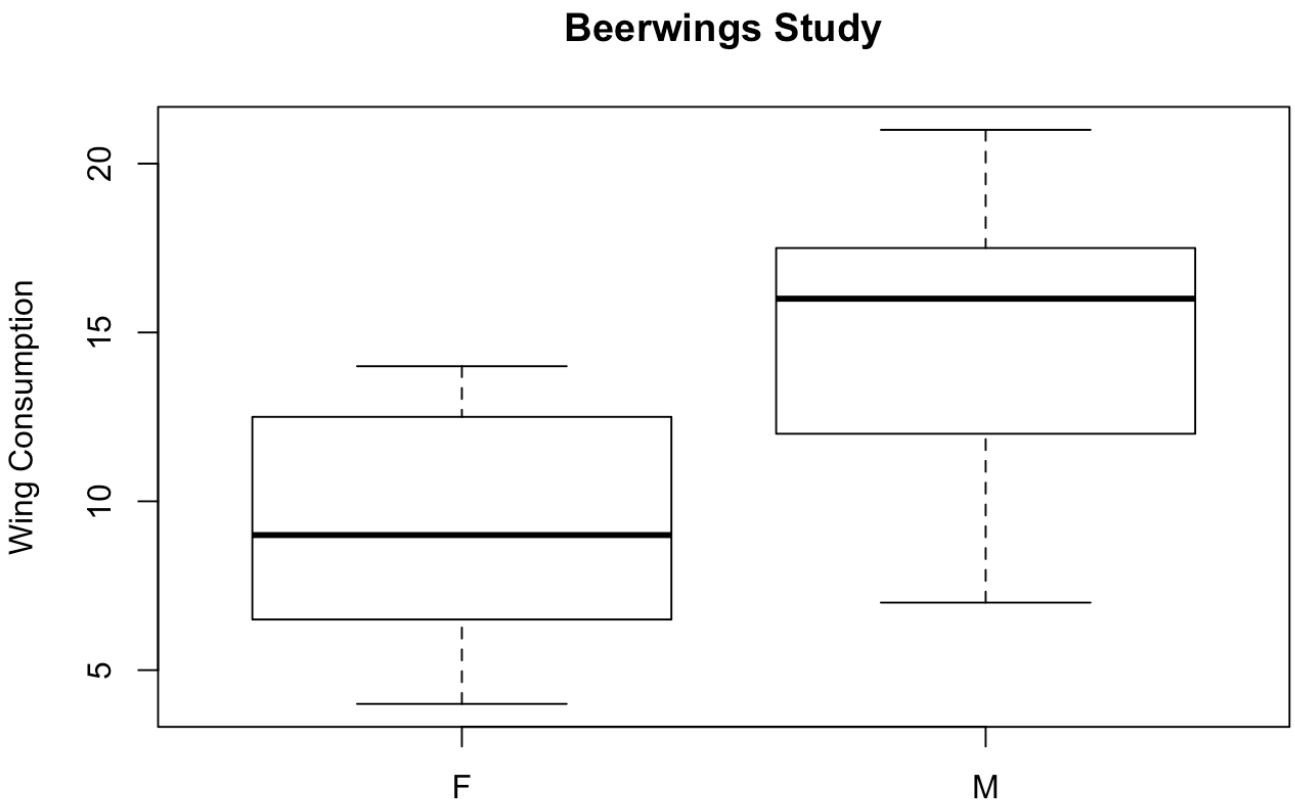
```
summary(Beerwings$Hotwings[Beerwings$Gender=="F"])  #Distribution for Fem
ales
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.000   6.500   9.000   9.333  12.500  14.000
```

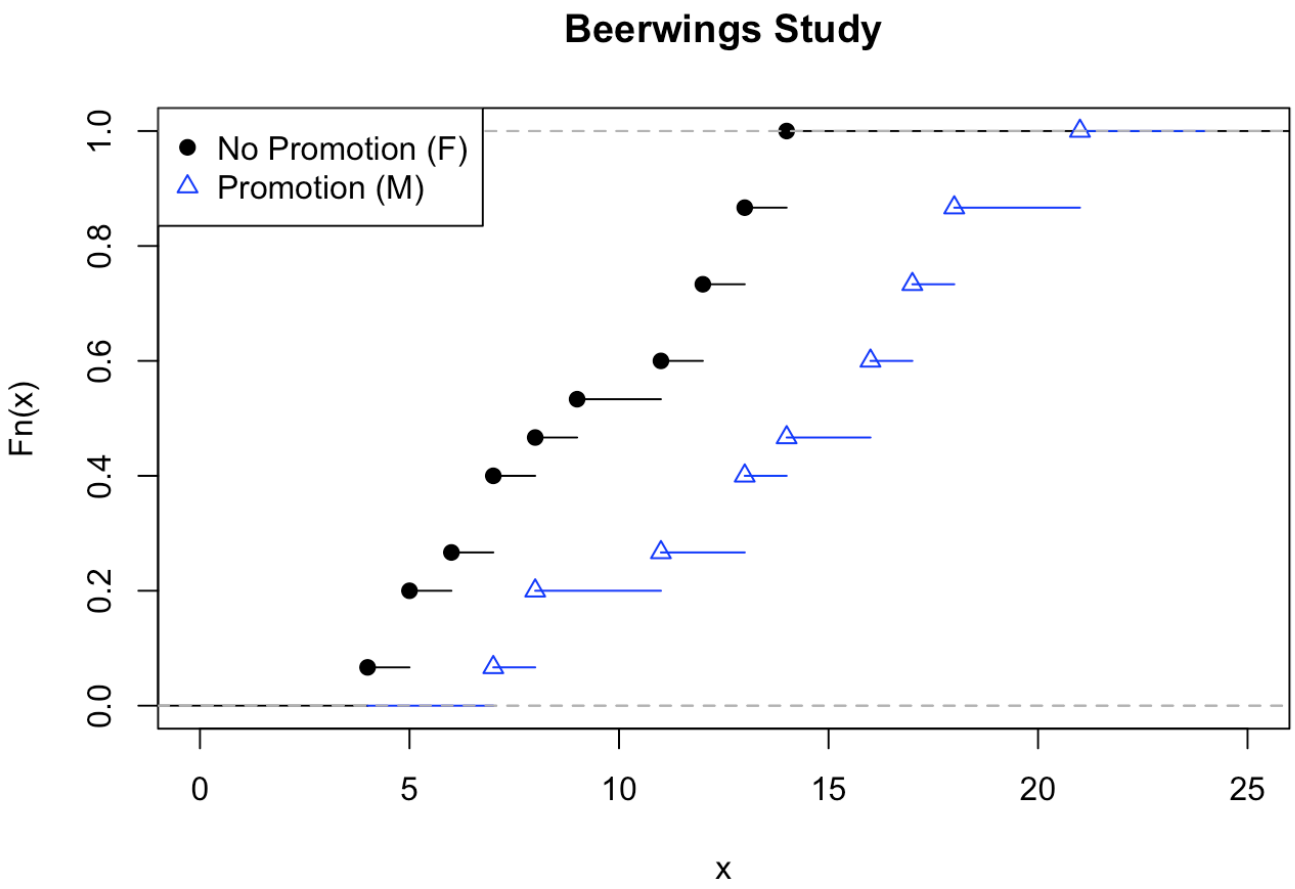
```
summary(Beerwings$Hotwings[Beerwings$Gender=="M"])  #Distribution for Males
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    7.00   12.00   16.00   14.53   17.50   21.00
```

```
boxplot(Hotwings~Gender,data=Beerwings,ylab="Wing Consumption",main="Beerwings Study")
```




```
plot.ecdf(Beerwings$Hotwings[Beerwings$Gender=="F"],xlim=c(0,25),main="Beerwings Study")
plot.ecdf(Beerwings$Hotwings[Beerwings$Gender=="M"],col="blue",pch=2, add=T)
legend("topleft",legend=c("No Promotion (F)","Promotion (M) "),col=c("black","blue"),pch=c(19,2))
```



The data indicate that the M group consumed an average of 5.2 more wings compared to the F group. Is the observed difference due to random chance? Let's conduct a randomization test to find out, but we will approximate the p-value using Monte Carlo simulation since there are 155.12 million possible differences.

```
mgrp <- Beerwings$Hotwings[Beerwings$Gender=='M']
fgrp <- Beerwings$Hotwings[Beerwings$Gender=='F']
wings <- c(mgrp,fgrp)
N <- 250000
result <- numeric(N)
set.seed(1701)
for (i in 1:N) {
index <- sample(length(wings),size=length(mgrp),replace=FALSE)
result[i] <- mean(wings[index]) - mean(wings[-index])
}
observed <- mean(mgrp) - mean(fgrp)

#P-value - results will vary depending on the sample chosen, but set.see
d() is fixed so
#same sample will be chosen unless this value changes.
phatval <- (sum(result >= observed)+1)/(N+1)
phatval
```

```
## [1] 0.001131995
```

The p-value 0.001132 is unusual under the null hypothesis. Thus, there is evidence that the promotion increased hot wings sales.

6 Basic Decision Theory in Hypothesis Testing

In hypothesis testing there are two types of errors that can be made. They are called type I and type II errors.

	H_0 true	H_1 true
Accept H_0	correct decision	type II error
Reject H_0	type I error	correct decision

The probabilities of type I and II errors are usually set in advance of running the experiment.

$$\alpha = P(\text{type I}),\ \beta = P(\text{type II}).$$

If the p-value $\leq \alpha$ then the test is statistically significant at level α. The power of the test is $1 - \beta$: the probability of rejecting H_0 when the alternative hypothesis H_1 is true.

7 Properties of the Randomization Test

The P-value of the one-sided randomization test must be a multiple of $\frac{1}{\binom{N}{N_A}}$. If a significance level of $\alpha = \frac{k}{\binom{N}{N_A}}$, where $k = 1, \ldots, \binom{N}{N_A}$ is chosen then

$$P(\text{type I}) = \alpha.$$

In other words the randomization test is an exact test.

If α is not chosen as a multiple of $\frac{1}{\binom{N}{N_A}}$, but $\frac{k}{\binom{N}{N_A}}$ is the largest p-value less than α, then $P(\text{type I}) = \frac{k}{\binom{N}{N_A}} < \alpha$ and the randomization test is conservative. Either way, the test is guarunteed to control the probability of a type I error under very minimal conditions: randomization of the experimental units to the treatments (Ernst, 2004).

8 The two-sample t-test

If the two wheat yield samples are independent random samples from a normal distribution with means μ_A and μ_B but the same variance then the statistic

$$\bar{y}_A - \bar{y}_b \sim N\left(\mu_A - \mu_B, \sigma^2(1/n_A + 1/n_B)\right).$$

So,

$$\frac{\bar{y}_A - \bar{y}_b - \delta}{\sigma\sqrt{(1/n_A + 1/n_B)}} \sim N(0,1),$$

where $\delta = \mu_A - \mu_B$.

If we substitute

$$S^2 = \frac{\sum_{i=1}^{n_A}(y_{iA} - \bar{y}_A) + \sum_{i=1}^{n_B}(y_{iB} - \bar{y}_B)}{n_A + n_B - 2}$$

for σ^2 then

$$\frac{\bar{y}_A - \bar{y}_b - \delta}{s\sqrt{(1/n_A + 1/n_B)}} \sim t_{n_A+n_B-2},$$

is called the two sample t-statistic.

In the wheat yield example $H_0 : \mu_A = \mu_B$ and suppose that $H_1 : \mu_A < \mu_B$. The p-value of the test is obtained by calculating the observed value of the two sample t-statistic under H_0.

$$t^* = \frac{\bar{y}_A - \bar{y}_b}{s\sqrt{(1/n_A + 1/n_B)}} = \frac{18.37 - 24.3}{4.72\sqrt{(1/6 + 1/6)}} = -2.18$$

The p-value is $P(t_{18} < -2.18) = 0.03$.

The calculation was done in R.

```
s <- sqrt((5*var(yA)+5*var(yB))/10)
tstar <- (mean(yA)-mean(yB))/(s*sqrt(1/6+1/6)); round(tstar,2)
```

```
## [1] -2.18
```

```
pval <- pt(tstar,10); round(pval,2)
```

```
## [1] 0.03
```

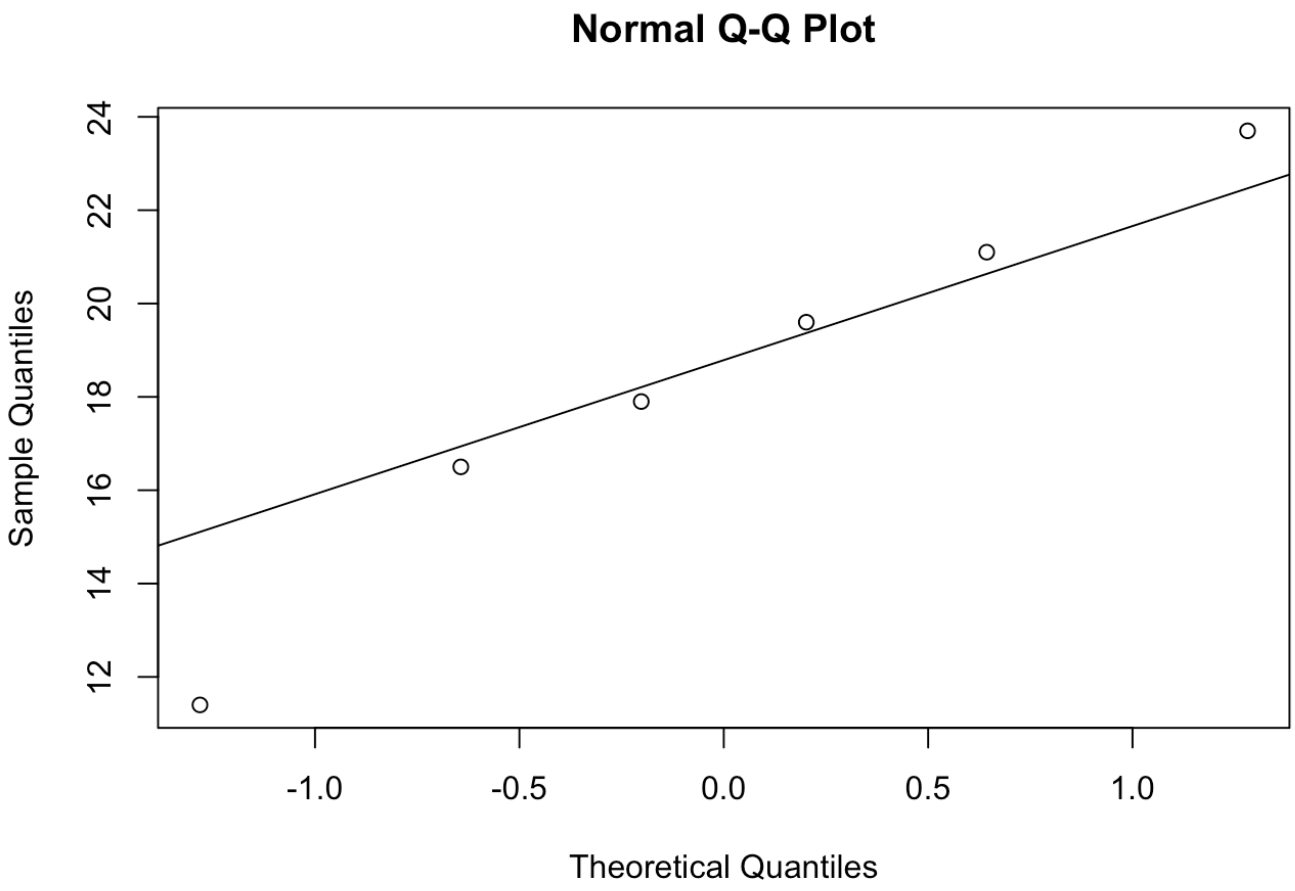
In R the command to run a two-sample t-test is `t.test()`.

```
t.test(yA,yB,var.equal = TRUE,alternative = "less")
```

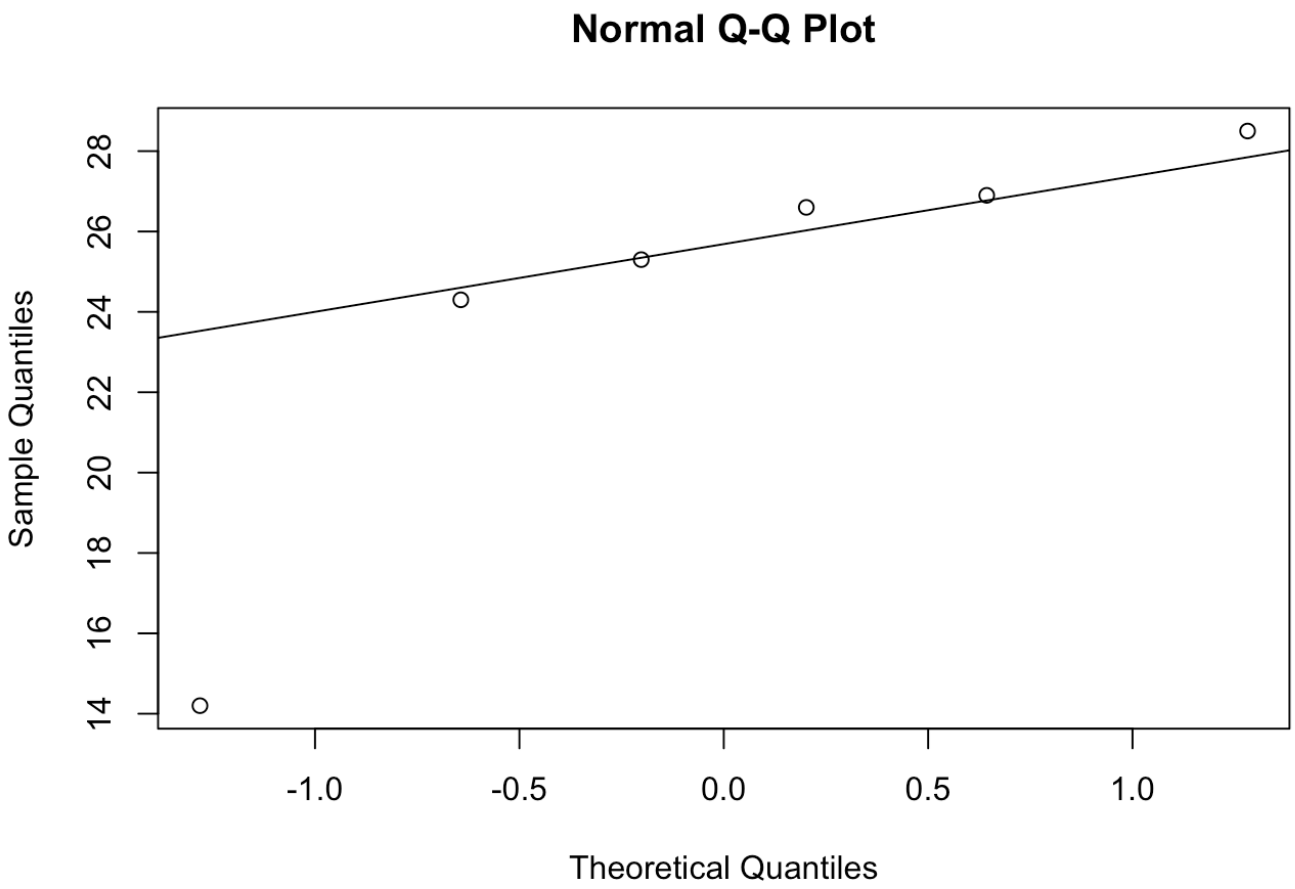
```
##
##  Two Sample t-test
##
## data:  yA and yB
## t = -2.1793, df = 10, p-value = 0.02715
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.9987621
## sample estimates:
## mean of x mean of y
##  18.36667  24.30000
```

The assumption of normality can be checked using normal quantile plots, although the t-test is robust against non-normality.

```
qqnorm(yA);qqline(yA)
```




```
qqnorm(yB);qqline(yB)
```



Both plots indicate that the normality assumption is satisfied.

Notice that the p-value from the randomization test and the p-value from two-sample t-test are almost identical. Although the randomization test neither depends on normality nor independence. The randomization test does depend on Fisher's concept that after randomization, if the null hypothesis is true, the two results obtained from each particular plot will be *exchangeable*. The randomization test tells you what you could say if exchangeability were true.

9 Randomized paired comparison

If a comparison is made within matched pairs of experimental units then randomization is straightforward to carry out within a matched pair.

This is illustrated with a study on the wear of boys' shoes (Box, Hunter, and Hunter, 2005).

Measurements on the amount of wear of the soles of shoes worn by 10 boys were obtained by the following design:

- Each boy wore a special pair of shoes with the soles made of two different synthetic materials, A (a standard material) and B (a cheaper material).
- The decision as to whether the left or right sole was made with A or B was determined by the flip of a fair coin.
- During the test some boys skuffed their shoes more than others, but each boys' shoes were subjected to the same amount of wear.

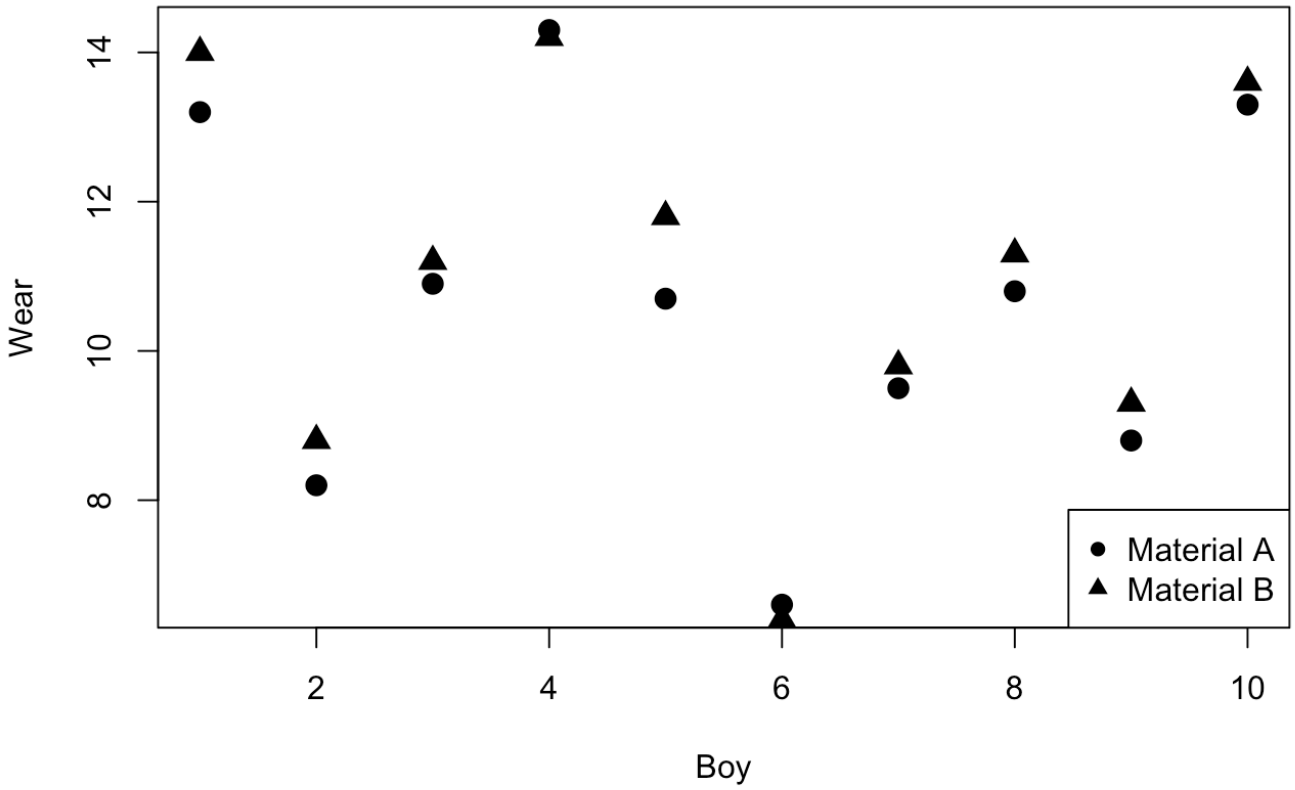
- Most of the boy-to-boy variation can be eliminated by working with the differences between A and B.

The data and a plot of the data are shown below .

```
library(BHH2)
data(shoes.data)
shoes.data
```

```
##    boy matA sideA matB sideB
## 1    1 13.2     L 14.0     R
## 2    2  8.2     L  8.8     R
## 3    3 10.9     R 11.2     L
## 4    4 14.3     L 14.2     R
## 5    5 10.7     R 11.8     L
## 6    6  6.6     L  6.4     R
## 7    7  9.5     L  9.8     R
## 8    8 10.8     L 11.3     R
## 9    9  8.8     R  9.3     L
## 10  10 13.3     L 13.6     R
```

```
plot(shoes.data$boy,shoes.data$matA,pch=16,cex=1.5,xlab="Boy",ylab="Wea
r")
points(shoes.data$boy,shoes.data$matB,pch=17,cex=1.5)
legend("bottomright",legend=c("Material A","Material B"),pch=c(16,17))
```



An experimental design of this kind is called a randomized paired comparison design. Later in the course we will see how this idea can be extended to compare more than two treatments using randomized block designs.

10 The Randomization Test for a Randomized Paired Design

The treatments were assigned to the boys left or right shoe by flipping a fair coin. If the coin toss was a tail then the left side received material A and the right side material B; if the coin toss was a head then the right side received material A and the left side received material B.

Exercise: Based on the boys shoe data above write down the treatment allocation for this experiment. Use “T” for tails and “H” for heads.

Answer: T T H T H T T T H T

The null hypothesis is that there is no difference in wear between A and B. This means that the treatment assignment (sequence of 10 coin tosess) is one of $2^{10} = 1024$ equiprobable treatment assignments.

In a paired design we can work with the difference between treatmnt for each experimental unit.

```
diff <- shoes.data$matA-shoes.data$matB
meandiff <- mean(diff); meandiff
```

```
## [1] -0.41
```

```
shoe.dat2 <- data.frame(shoes.data,diff)
shoe.dat2
```

```
##    boy matA sideA matB sideB diff
## 1    1 13.2     L 14.0     R -0.8
## 2    2  8.2     L  8.8     R -0.6
## 3    3 10.9     R 11.2     L -0.3
## 4    4 14.3     L 14.2     R  0.1
## 5    5 10.7     R 11.8     L -1.1
## 6    6  6.6     L  6.4     R  0.2
## 7    7  9.5     L  9.8     R -0.3
## 8    8 10.8     L 11.3     R -0.5
## 9    9  8.8     R  9.3     L -0.5
## 10  10 13.3     L 13.6     R -0.3
```

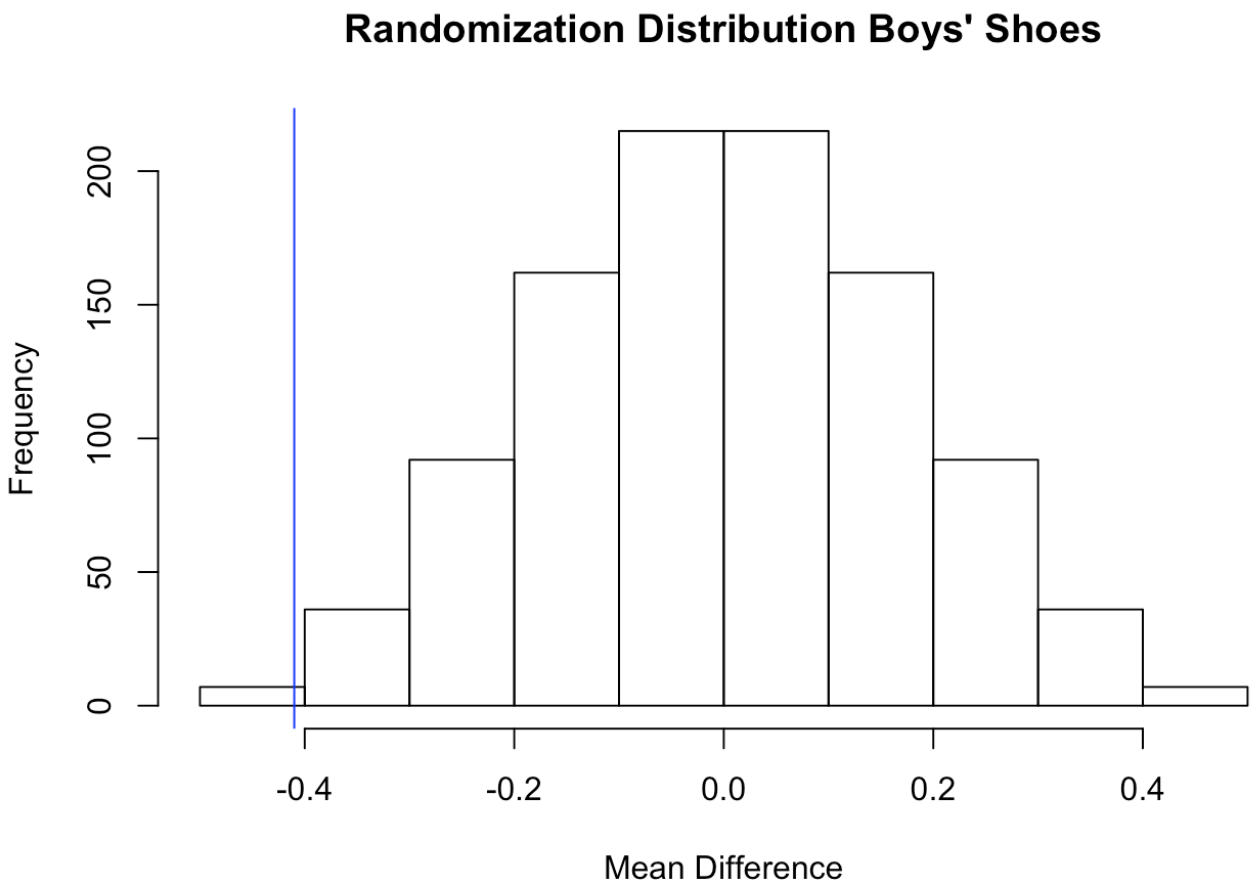
Under the null hypothesis the wear of boys left or right shoe is same regardless of what material he had on his sole. This means that if there was a different treatment assignment, say, H T H T H T T T H T then the difference for the first boy would have been +0.8 since he would have had his right side assigned to material A (14.0) and his left side assigned to material B (13.2).

The randomization distribution of the average difference is the distribution of the average differences, for all the different treatment assignments.

```
N <- 2^(10) # number of treatment assignments
res <- numeric(N) #vector to store results
LR <- list(c(-1,1)) # difference is multiplied by -1 or 1
trtassign <- expand.grid(rep(LR, 10)) # generate all possible treatment a
ssign

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

hist(res, xlab="Mean Difference",main="Randomization Distribution Boys' S
hoes")
abline(v = meandiff,col="blue")
```



The p-value for testing if B has more wear than A is:

$$P(D \leq d^* | H_0) = \sum_{i=1}^{2^{10}} \frac{I(d_i \leq d^*)}{2^{10}},$$

where $D = \bar{A} - \bar{B}$, and d^* is the observed mean difference.

This can be calculated in R

```
sum(res<=meandiff) # number of differences le observed diff
```

```
## [1] 7
```

```
sum(res<=meandiff)/N # p-value
```

```
## [1] 0.006835938
```

The value of d^* = -0.41 is unsusual under the null hypothesis since only 7 produced by the randomization distribution give d^* less than -0.41. Therefore, there is a statistically significant increase in the amount of wear with the cheaper material B.

11 Paired t-test

If we assume that the differences -0.8, -0.6, -0.3, 0.1, -1.1, 0.2, -0.3, -0.5, -0.5, -0.3 are a random sample from a normal distribution then the statistic

$$t = \frac{\bar{d}}{s_{\bar{d}}/\sqrt{10}} \sim t_{10-1},$$

where, $s_{\bar{d}}$ is the sample standard deviation of the paired differences. The p-value for testing if $\bar{D} < 0$ is

$$P(t_9 < t).$$

In general if there are n differences then

$$t = \frac{\bar{d}}{s_{\bar{d}}/\sqrt{n}} \sim t_{n-1},$$

where, $s_{\bar{d}}$ is the sample standard deviation of the paired differences. The p-value for testing if $\bar{D} < 0$ is

$$P(t_{n-1} < t).$$

NB: This is the same as a one-sample t-test of the differences.

In R a paired t-test can be obtained by using the command `t.test()`.

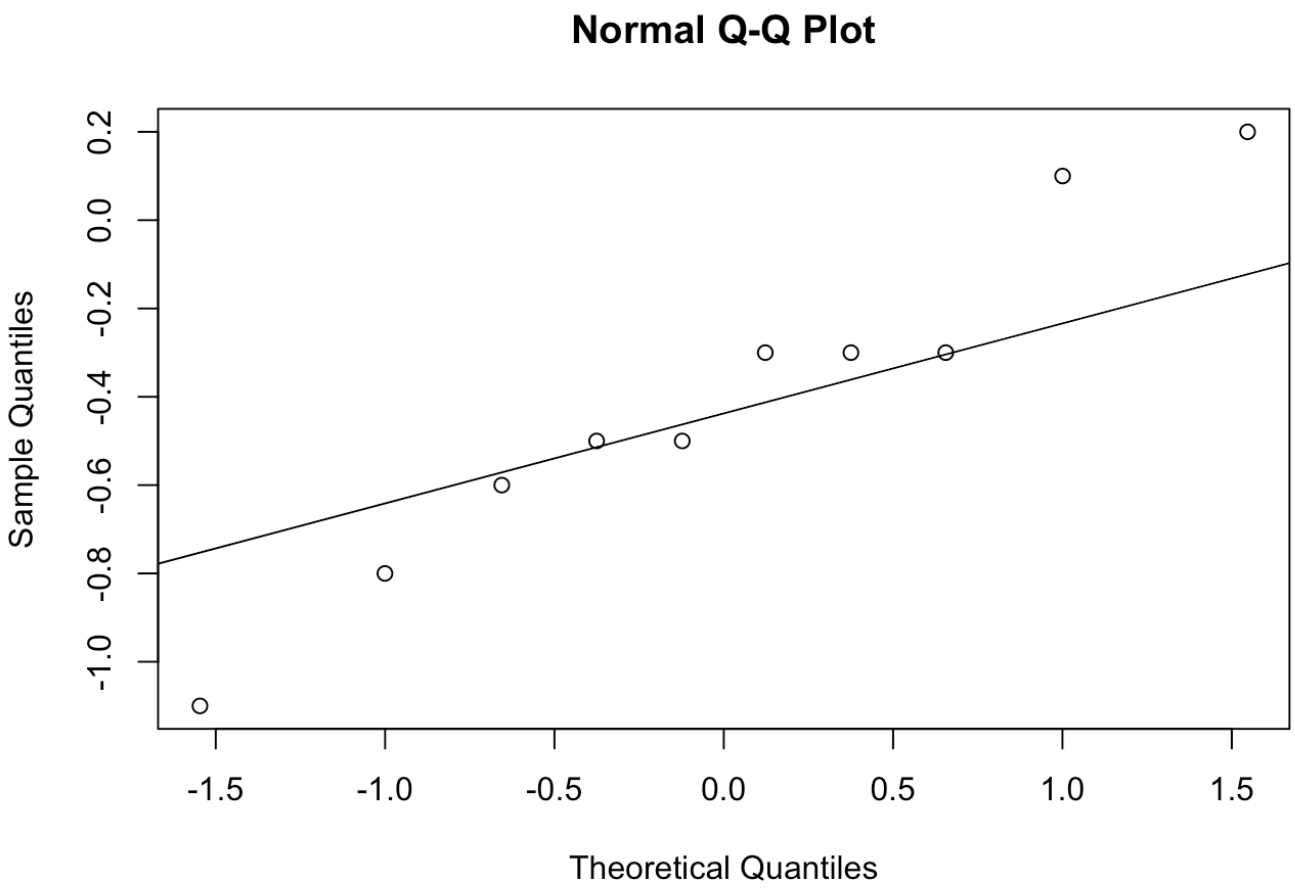
```
t.test(shoes.data$matA,shoes.data$matB,paired = TRUE,alternative = "les
s")
```

```
##
##  Paired t-test
##
## data:  shoes.data$matA and shoes.data$matB
## t = -3.3489, df = 9, p-value = 0.004269
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.1855736
## sample estimates:
## mean of the differences
##                   -0.41
```

```
t.test(diff,alternative = "less") # same as a one-sample t-test on the di
ff
```

```
##
##  One Sample t-test
##
## data:  diff
## t = -3.3489, df = 9, p-value = 0.004269
## alternative hypothesis: true mean is less than 0
## 95 percent confidence interval:
##        -Inf -0.1855736
## sample estimates:
## mean of x
##     -0.41
```

```
qqnorm(diff); qqline(diff)
```



Exercise: Calculate the test statistic and p-value of the paired t test using R

Answer:

```
tobs <- mean(diff)/(sd(diff)/sqrt(10)); tobs
```

```
## [1] -3.348877
```

```
pt(tobs,df = 9) # p-value using t-dist CDF
```

```
## [1] 0.00426939
```

12 Questions

1. Use the `Beerwings` study answer the following questions.

a. Explain the rationale behind using Monte Carlo simulation to calculate the p-value for the randomization test of mean hotwing consumption between the two groups.
b. Is this an experiment or observational study?
c. Use a two-sample t-test to test for a difference in hotwing consumption between men and women? Does your answer agree with the randomization test? State the assumptions that both tests rely upon.
d. Did the promotion increase beer sales? Conduct an appropriate statistical test. What do you conclude?
e. Suppose that the researcher recruited 40 subjects and decided to randomize 30 subjects to group M and 10 subjects to group F. How many possible treatment assignments are possible?

2. Suppose that two drugs A and B are to be tested on 12 subjects' eyes. The drugs will be randomly assigned to the left eye or right eye based on the flip of a fair coin. If the coin toss is heads then a subject will receive drug A in their right eye. The coin was flipped 12 times and the following sequence of heads and tails was obtained:

$$T \quad T \quad H \quad T \quad H \quad T \quad T \quad T \quad H \quad T \quad T \quad H$$

a. Create a table that shows how the treatments will be allocated to the 12 subjects' left and right eyes?
b. What is the probability of obtaining this treatment allocation?
c. What type of experimental design has been used to assign treatments to subjects? Explain.

13 Answers

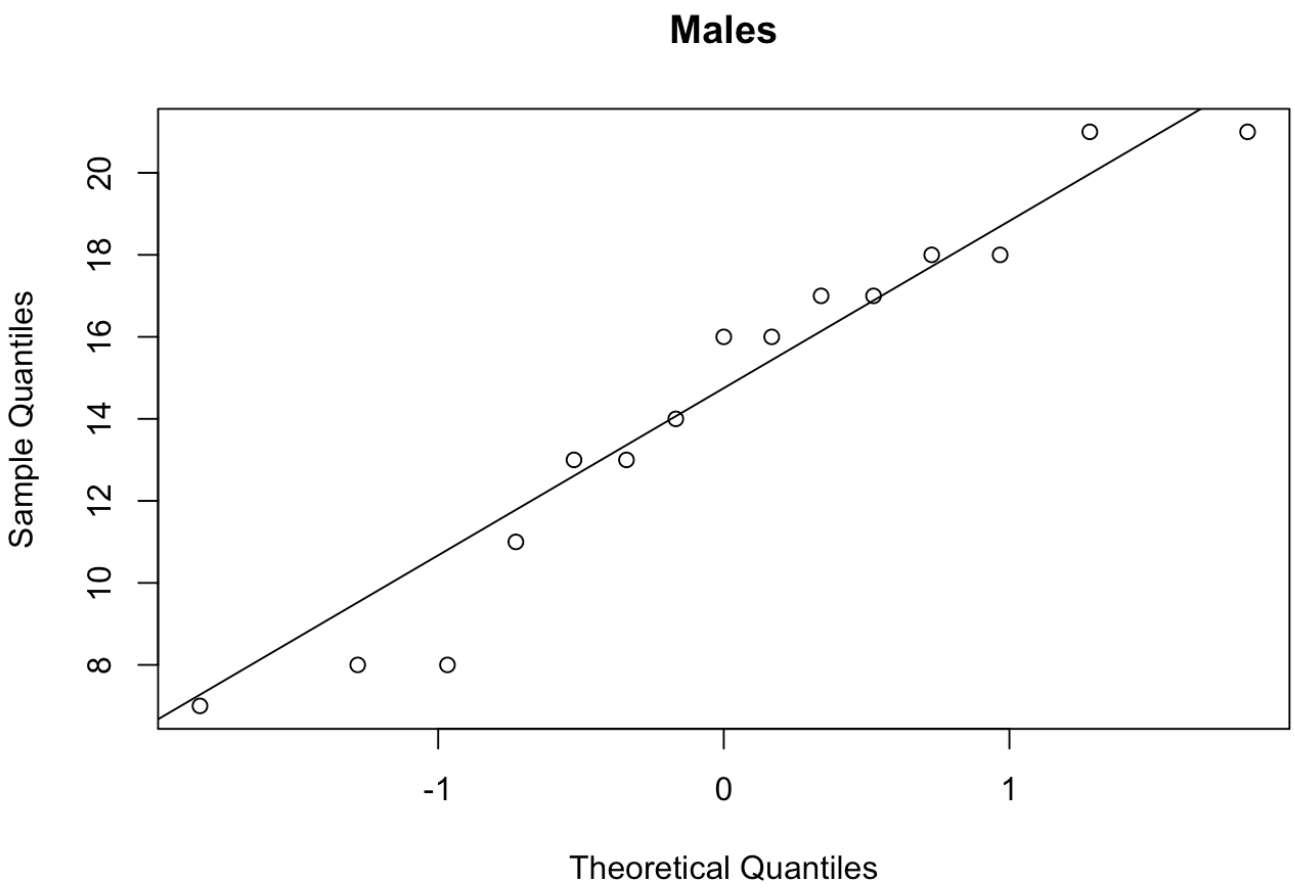
1.

a. The number of possible ways to split the data is $\binom{30}{15}$ which is very large.
b. Experiment. The assignment mechanism is controlled by the experimenter.
c.

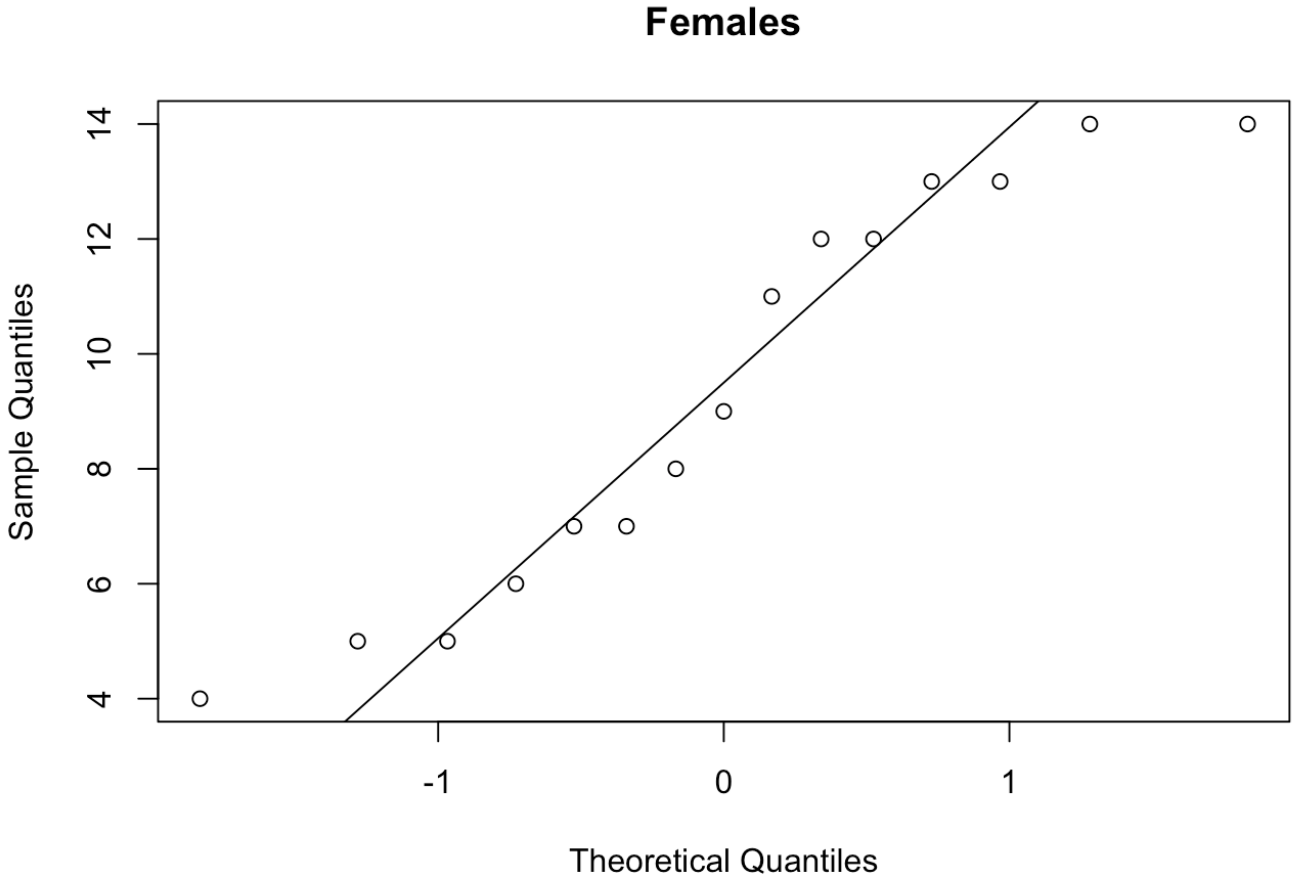
```
mgrp <- Beerwings$Hotwings[Beerwings$Gender=='M']
fgrp <- Beerwings$Hotwings[Beerwings$Gender=='F']
t.test(mgrp,fgrp,var.equal = T,alternative = "greater")
```

```
##
##  Two Sample t-test
##
## data:  mgrp and fgrp
## t = 3.5094, df = 28, p-value = 0.0007692
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  2.679365      Inf
## sample estimates:
## mean of x mean of y
## 14.533333  9.333333
```

```
qqnorm(mgrp,main="Males");qqline(mgrp)
```




```
qqnorm(fgrp,main="Females");qqline(fgrp)
```



The p-value is very close to the p-value from the randomizatiuon test. The t-test relies on each sample being normal with different means but the same SD, and the observations within each sample are independent. The qq plots indicate the normality assumption is plausible. The randomization test relies on randomization of treatments to experimental units, and exchangeability under the null hypothesis.

d. There is evidence that the promotion increased sales since the p-value is very small.

e. $\binom{40}{10} = \binom{40}{30} = 847660528.$

2. a.

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

b. $\frac{1}{2^{12}} = 0.0002441406.$

c. A randomized paired design. There are 12 subjects, but each subject receives both treatments in a paired fashion.