STA305/1004 - Class 4

January 18, 2017

Today's Class

- Hypothesis testing via randomization
- Two-sample t-test
- Paired t-test

Example: Wheat Yield

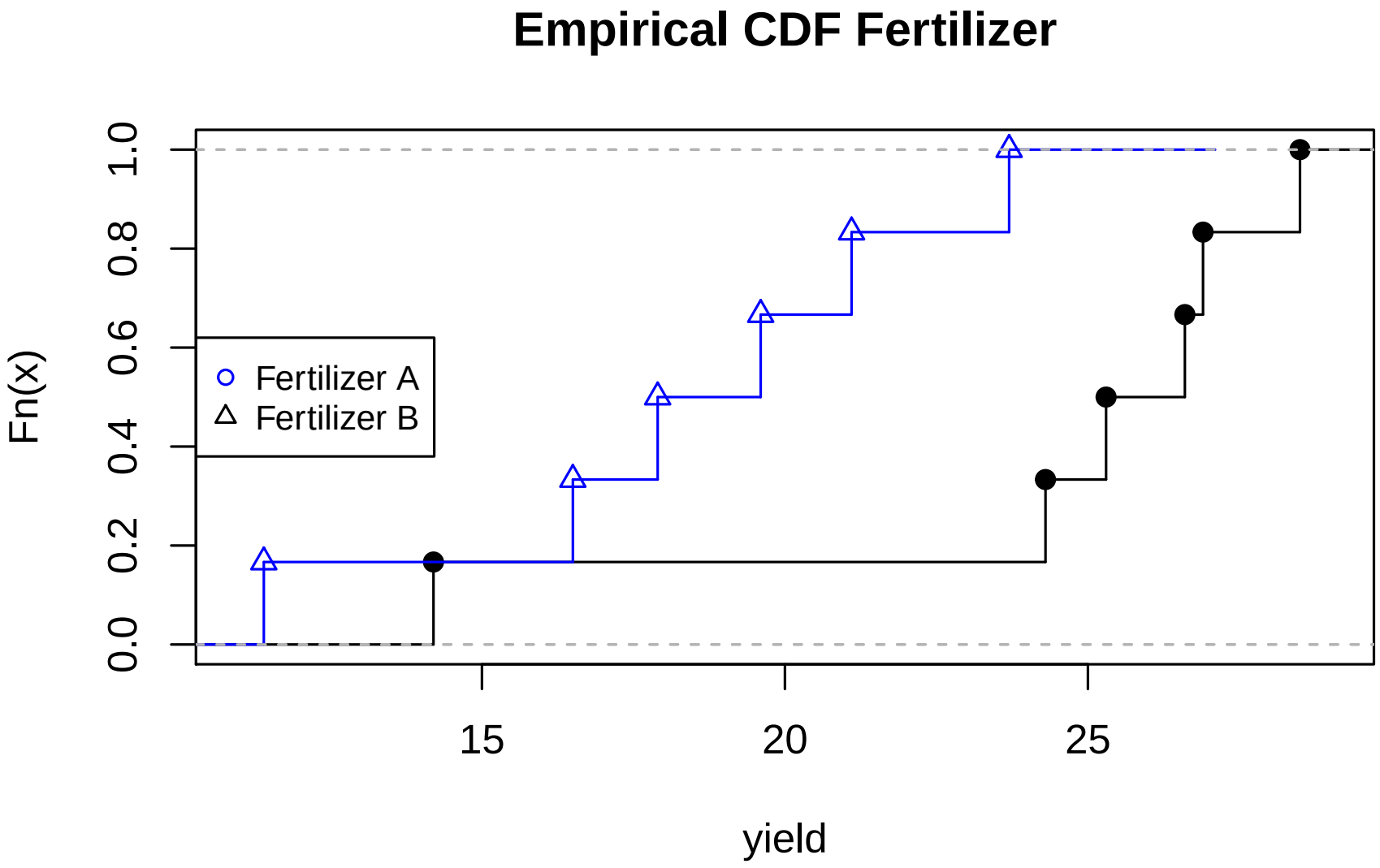
- Assigning treatments randomly avoids any pre-experimental bias.
- 12 playing cards, 6 red, 6 black were shuffled (7 times??) and dealt
- 1st card black $\rightarrow 1^{st}$ plot gets B
- 2nd card red $\rightarrow 2^{nd}$ plot gets A
- 3rd card black $\rightarrow 3^{rd}$ plot gets B
- Completely randomized design

Wheat Yield Example

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

- Evidence that fertilizer type is a source of yield variation?
- Evidence about differences between two populations is generally measured by comparing summary statistics across two sample populations.
- A statistic is any computable function of the observed data.

Wheat Yield Study



Wheat Yield Study

```
summary(yA); sd(yA);quantile(yA,prob=c(0.25,0.75))
```

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

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

```
##    25%    75%
## 16.850 20.725
```

```
summary(yB); sd(yB); quantile(yB,prob=c(0.25,0.75))
```

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

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

```
##    25%    75%
## 24.550 26.825
```

Results

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

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

- So there is a moderate/large difference in mean yield for these fertilizers.
- Would you recommend B over A for future plantings?
- Do you think these results generalize to a larger population?
- Could the result be due to chance?

Hypothesis Testing Via Randomization

- Are the observed differences in yield due to fertilizer type?
- Are the observed differences in yield due to plot-to-plot variation?

Hypothesis Testing Via Randomization

Hypothesis tests:

- H_0 (null hypothesis): Fertilizer type does not affect yield.
- H_1 (alternative hypothesis): Fertilizer type does affect yield.
- A statistical hypothesis evaluates the compatibility of H_0 with the data

Test Statistics and Null Distributions

We can evaluate H_0 by answering:

- Is a mean difference of -5.93 plausible/probable if H0 true?
- Is a mean difference of -5.93 large compared to experimental noise?

Test Statistics and Null Distributions

- Compare $\bar{y}_a - \bar{y}_b$=-5.93 (observed difference in the experiment) to values of $\bar{y}_a - \bar{y}_b$ that could have been observed if H_0 were true.
- Hypothetical values of $\bar{y}_a - \bar{y}_b$ that could have been observed under H_0 are referred to as samples from the null distribution.

Test Statistics and Null Distributions

- $\bar{y}_a - \bar{y}_b$ is a function of the outcome of the experiment.
- If a different experiment were performed then we would obtain a diffrent value of $\bar{y}_a - \bar{y}_b$.

Test Statistics and Null Distributions

- In this experiment we observed $\bar{y}_a - \bar{y}_b$=-5.93.
- If there was no difference between fertilizers then what other possible values of $\bar{y}_a - \bar{y}_b$ could have been observed?

Experimental Procedure and Potential Outcomes

The cards were shuffled and we were dealt B, R, B, R, ...

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

Under this treatment assignment we oberved the yields:

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

Experimental Procedure and Potential Outcomes

Another potential treatment assignment under H_0 is:

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

The yields obtained under this assignment are:

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

This data could occur if the experiment were run again.

Experimental Procedure and Potential Outcomes

- Under this hypothetical assignment the mean difference is:

```
yA <- c(11.4,25.3,28.5,14.2,21.1,24.3)
yB <- c(26.9,26.6,23.7,17.9,16.5,19.6)
mean(yA-yB)
```

```
## [1] -1.066667
```

This represents an outcome of the experiment in a universe where:

1. The treatment assignment is B, A, B, B, A, A, A, B, B, A, A, B
2. H_0 is true (i.e., $\mu_A = \mu_B$, where μ_A, μ_B are the mean yields of fertilizers A and B).

The Null distribution

- What potential outcomes **could** we see if H_0 is true?
- Compute $\bar{y}_a - \bar{y}_b$ for each possible treatment assignment.

The Null Distribution

- For each treatment assignment compute

$$\delta_i = \bar{y}_a - \bar{y}_b, i = 1, 2, \ldots, 924.$$

- $\{\delta_1, \delta_2, \ldots, \delta_{924}\}$ enumerates all pre-randomisation outcomes assuming no treatment effect.
- Since each treatment assignment is equally likely under the null distribution, a probability distribution of experimental results if H_0 is true can be described as

$$\begin{aligned}\hat{F}(y) &= \frac{\#(\delta_i \leq y)}{924} \\ &= \frac{\sum_{k=1}^{\binom{12}{6}} I(\delta_k \leq y)}{\binom{12}{6}}\end{aligned}$$

This is called the randomisation distribution.

Randomization Distribution

- The yield is not random since the plots were not chosen randomly.
- Their assignment to treatments is random.
- The basis for building a probability distribution for $\bar{y}_a - \bar{y}_b$ comes from the randomization of fertilizers to plots.

Randomization Distribution

- This randomization results in 6 plots getting fertilizer A and the remaining 6 plots receiving fertilizer B.
- This is one of $\binom{12}{6} = 924$ equally likely randomizations that could have occured.

Experimental Procedure and Potential Outcomes

This represents an outcome of the experiment in a universe where:

1. H_0 is true.
2. The yield will be the same regardless of which fertilizer a plot received.

For example a plot that had a yield of 26.9 given fertilizer B would have the same yield if the plot received fertilizer A if H_0 is true.

R Code for Randomization Distribution

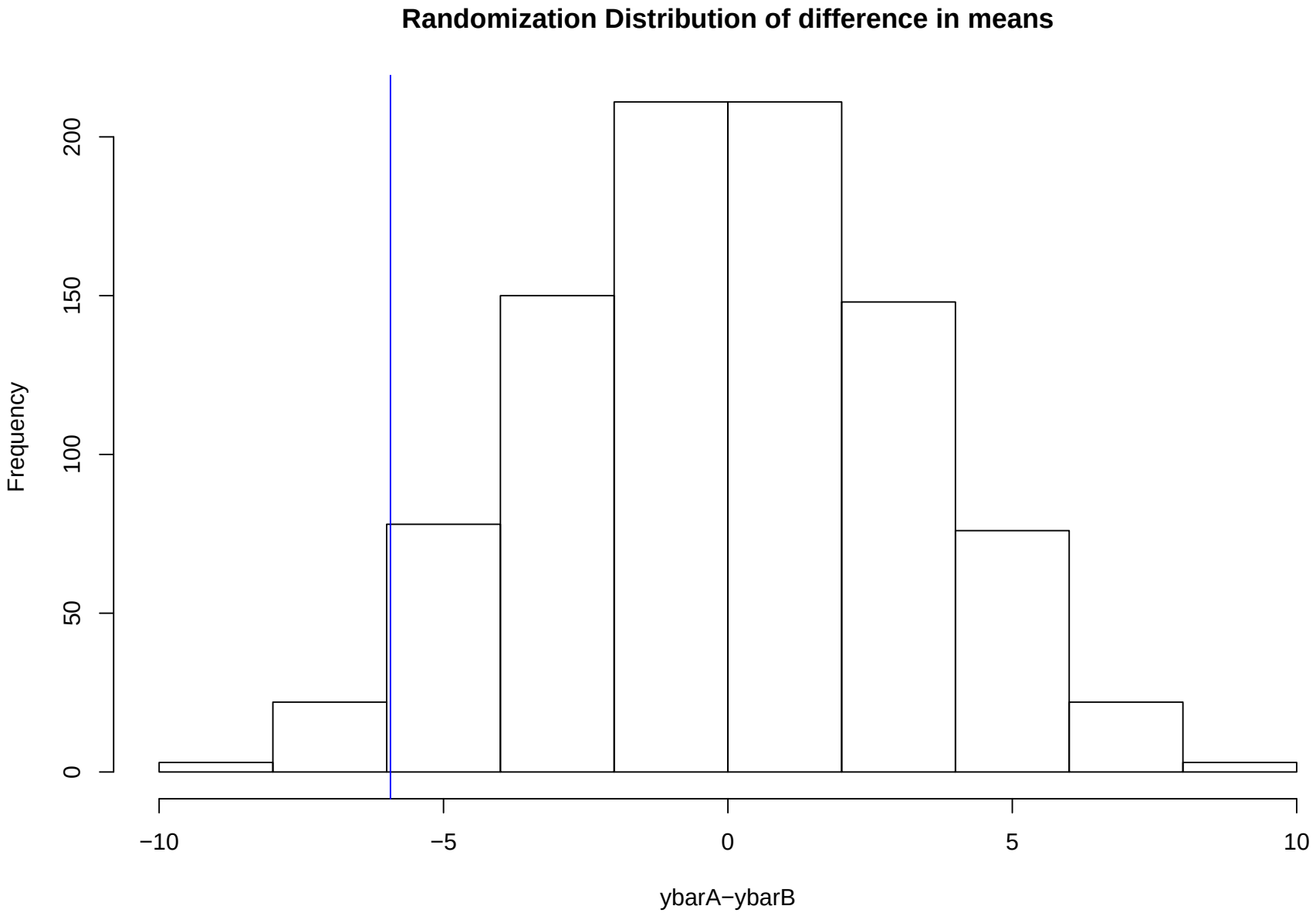
```
yA <- c(11.4,23.7,17.9,16.5,21.1,19.6);yB <- c(26.9,26.6,25.3,28.5,14.2
fert <- c(yA,yB); N <- choose(12,6)
res <- numeric(N) # store the results
index <-combn(1:12,6) #Generate N treatment assignments
for (i in 1:N)
{res[i] <- mean(fert[index[,i]])-mean(fert[-index[,i]])}
index[,1:2] #output first two randomizations
```

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

```
res[1:2] #output first two mean diffs
```

```
## [1] -5.933333 -3.500000
```

Randomization Distribution



Hypothesis Testing

- Is there any contradiction between H_0 and the observed data?
- A **P-value** is the probability, under the null hypothesis of obtaining a more extreme than the observed result.

$$\text{P-value} = P(\delta \leq -5.93) = \hat{F}(-5.93)$$

- A small P-value implies evidence **against** null hypothesis.
- If the P-value is large does this imply that the null is true?

Randomization Test

- Assume H_0 is true.
- Calculate the difference in means for every possible way to split the data into two samples of size 6.
- This would result in $\binom{12}{6} = 924$ differences.
- Calculate the probability of observing a value as extreme of more extreme than the observed value of the test statistic (*P-value*).
- If the P-value is small then there are two possible explanations:

1. An unlikely value of the statistic has occurred, or
2. The assumption that H_0 is true is incorrect.

- If the P-value is large then the hypothesis test is inconclusive.

Computing the P-value

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 observed 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
```

Interpretation of P-value

- 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.
- Therefore it's unlikely (if we consider 3% unlikely) that an observed mean difference as extreme or more extreme than -5.93 would be observed if $\mu_A = \mu_B$.

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

$$\bar{t} = \left(1/\binom{N}{N_A}\right) \sum_{i=1}^{\binom{N}{N_A}} t_i$$

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

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

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

Two-Sided Randomization P value

```
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
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
```

Randomization Test

- We could calculate the difference in means for every possible way to split the data into two samples of size 6.
- This would result in $\binom{12}{6} = 924$ differences.
- If there were 30 observations split evenly into two groups then there are $\binom{30}{15} = 155,117,520$ differences.
- So unless the sample sizes are small these exhaustive calculations are not practical.

Randomization Test

Instead we can create a permutation resample (Monte Carlo Sampling).

1. Draw 6 observations from the pooled data without replacement. (fert A)
2. The remaining 6 observations will be the second sample (fert B)
3. Calculate the difference in means of the two samples
4. Repeat 1-3 at least 250000 times.
5. P-value is the fraction of times the random statistics exceeds the original statistic.

Estimate P-value via Monte Carlo Sampling

If M test statistics, t_i, $i = 1, ..., M$ 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.

Estimate P-value via Monte Carlo Sampling

```
N <- 250000 # number of times to repeat this process
result <- numeric(N) # space to save random diffs.
for (i in 1:N)
{ #sample of size 6, from 1 to 12, without replacement
  index <- sample(12,size=6,replace=F)
  result[i] <- mean(fert[index])-mean(fert[-index])
}

#store observed mean difference
observed <- mean(yA)-mean(yB)

#P-value - mean - results will vary
pval <- (sum(result <= observed)+1)/(N+1)
round(pval,4)
```

```
## [1] 0.0279
```

Basic Decision Theory

	H_0 True	H_0 False
Accept H_0	correct	type II error
Reject H_0	type I error	correct

$$\text{P-value} = P(\text{test statistic} \geq \text{observed value of test statistic})$$

$$\begin{aligned} \alpha &= P(\text{type I error}) \\ \beta &= P(\text{type II error}) \\ 1 - \beta &= \text{power} \end{aligned}$$

The Randomization P-value

- An achievable P-value of the randomization test must be a multiple of $\frac{k}{\binom{12}{6}} = \frac{k}{924}$, where $k = 1, 2, \ldots, 924$.
- If we choose a significance level of $\alpha = \frac{k}{924}$ that is one of the achievable P-values then $P(\text{type I error}) = \alpha$.
- The randomization test is an exact test.
- If α is not chosen to be one of the achievable P-values but $\alpha = \frac{k}{924}$ is the largest acheivable P-value less than α then $P(\text{type I error}) < \alpha$.

Choosing a Test Statistic

A test statistic should be able to differentiate between H_0 and H_a in ways that are scientifically relevant.

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

$$median(Y_A) - median(Y_B)$$

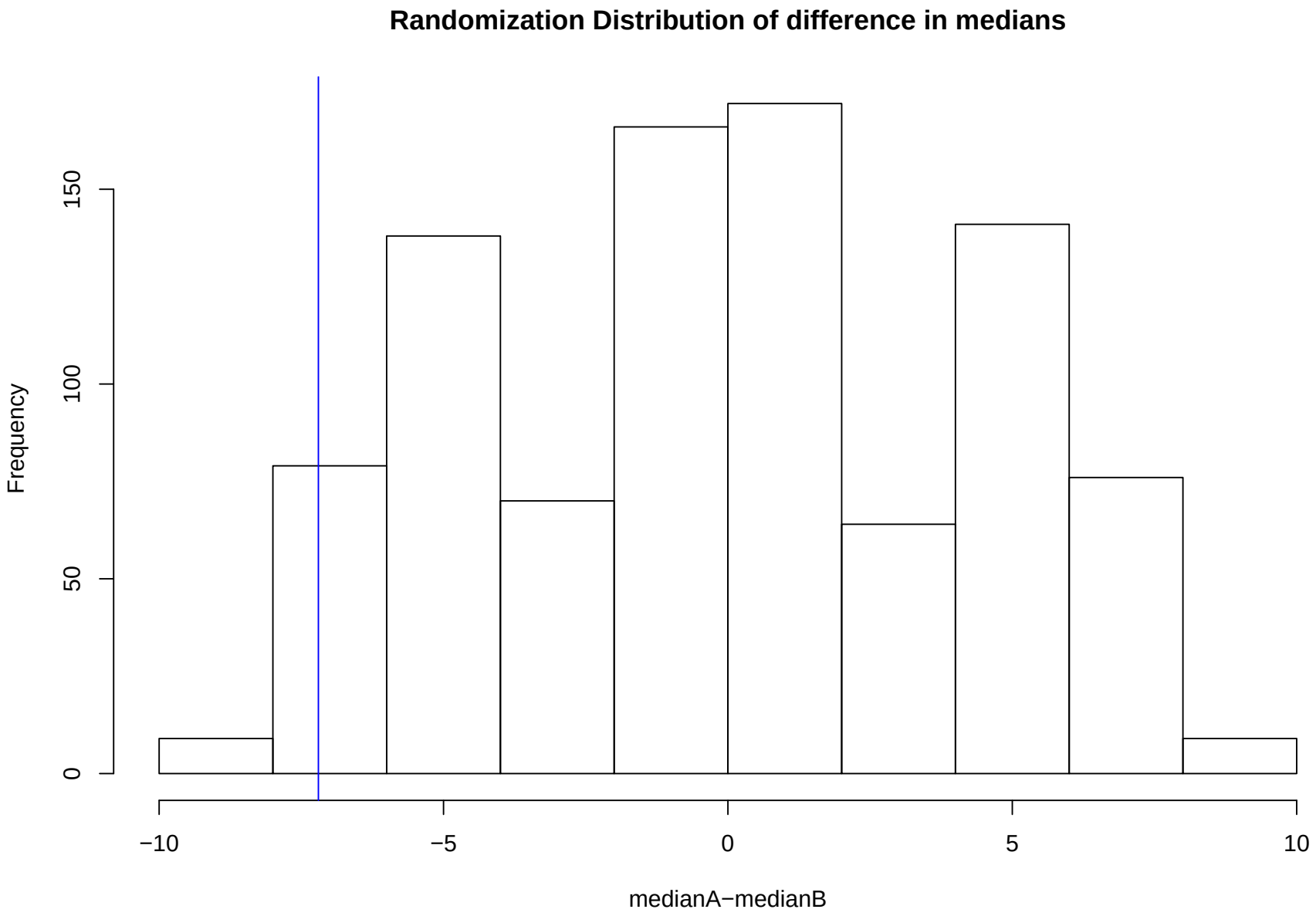
or trimmed means are examples of other test statistics.

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
index <-combn(1:12,6) # Generate N treatment assignments
for (i in 1:N)
{
  res[i] <- median(fert[index[,i]])-median(fert[-index[,i]])
}
```

Other Test Statistics



Other Test Statistics

The p-value of the randomization test can be calculated

```
# of times values from the median randomization
# distribution less than observed 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
```

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.

The two-sample t-test

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,5)
```

```
## [1] 0.02715
```

The two-sample t-test

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

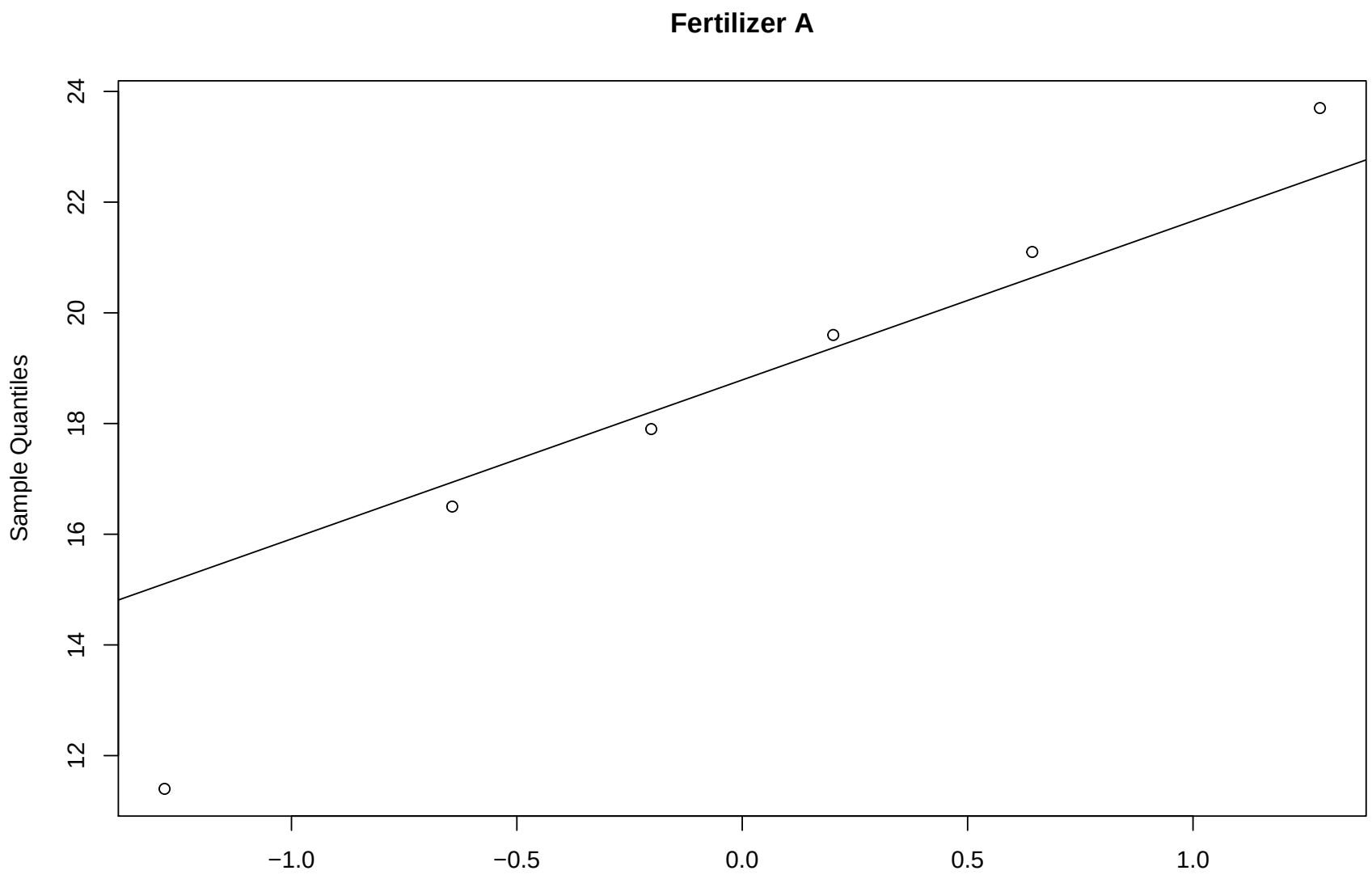
```
t.test(yA,yB,var.equal = TRUE,alternative = "less")
```

```
##
##  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 two-sample t-test

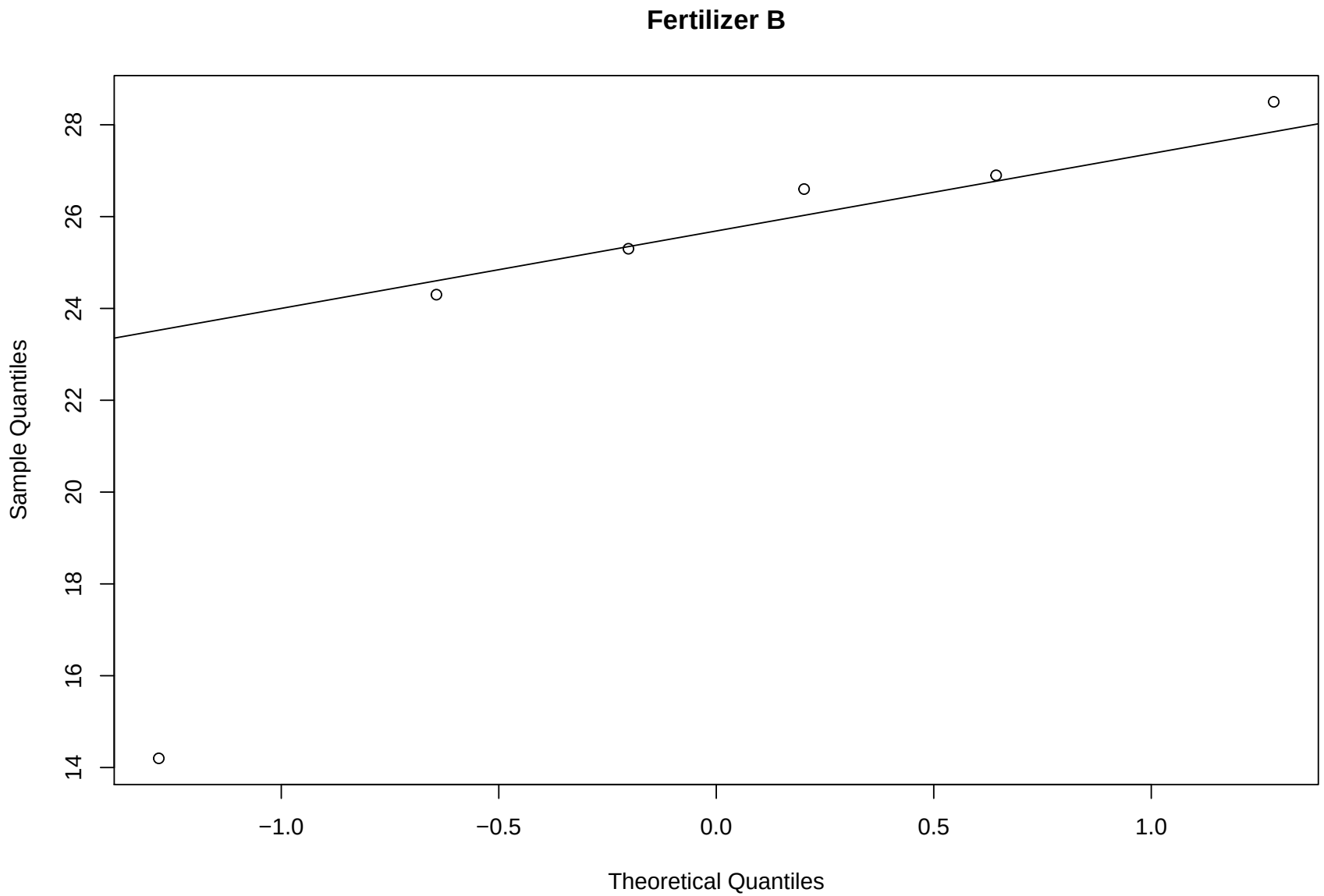
The assumption of normality can be checked using normal quantile plots, although the t-test is robust against non-normality.

```
qqnorm(yA,main = "Fertilizer A");qqline(yA)
```



The two-sample t-test

```
qqnorm(yB,main = "Fertilizer B");qqline(yB)
```



Two-Sample t-test versus Randomization Test

- The p-value from the randomization test and the p-value from two-sample t-test are almost identical.
- The randomization test does not depend on normality or independence.

Two-Sample t-test versus Randomization Test

- 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.

Paired Comparisons

- Increase precision by making comparisons within matched pairs of experimental material.
- Randomize within a pair.

Boy's Shoe Experiment

- Two materials to make boy's shoes, A and B, are tested to evaluate if B is more sturdy compared to A.
- During the experimental test some boys scuffed their shoes more than others.
- Each boy's two shoes were subjected to the same treatment by having each boy wear both materials.
- Working with 10 differences B-A most of the boy-to-boy variation could be eliminated.
- Called a randomized paired comparison design.

Boy's Shoe Experiment

- Toss a coin to randomize material to L/R foot of a boy.
- Head: Material A used on right foot.
- Null hypothesis: amount of wear associated with material A and B are the same.
- So labelling given to a pair of results only affects the sign of the difference.

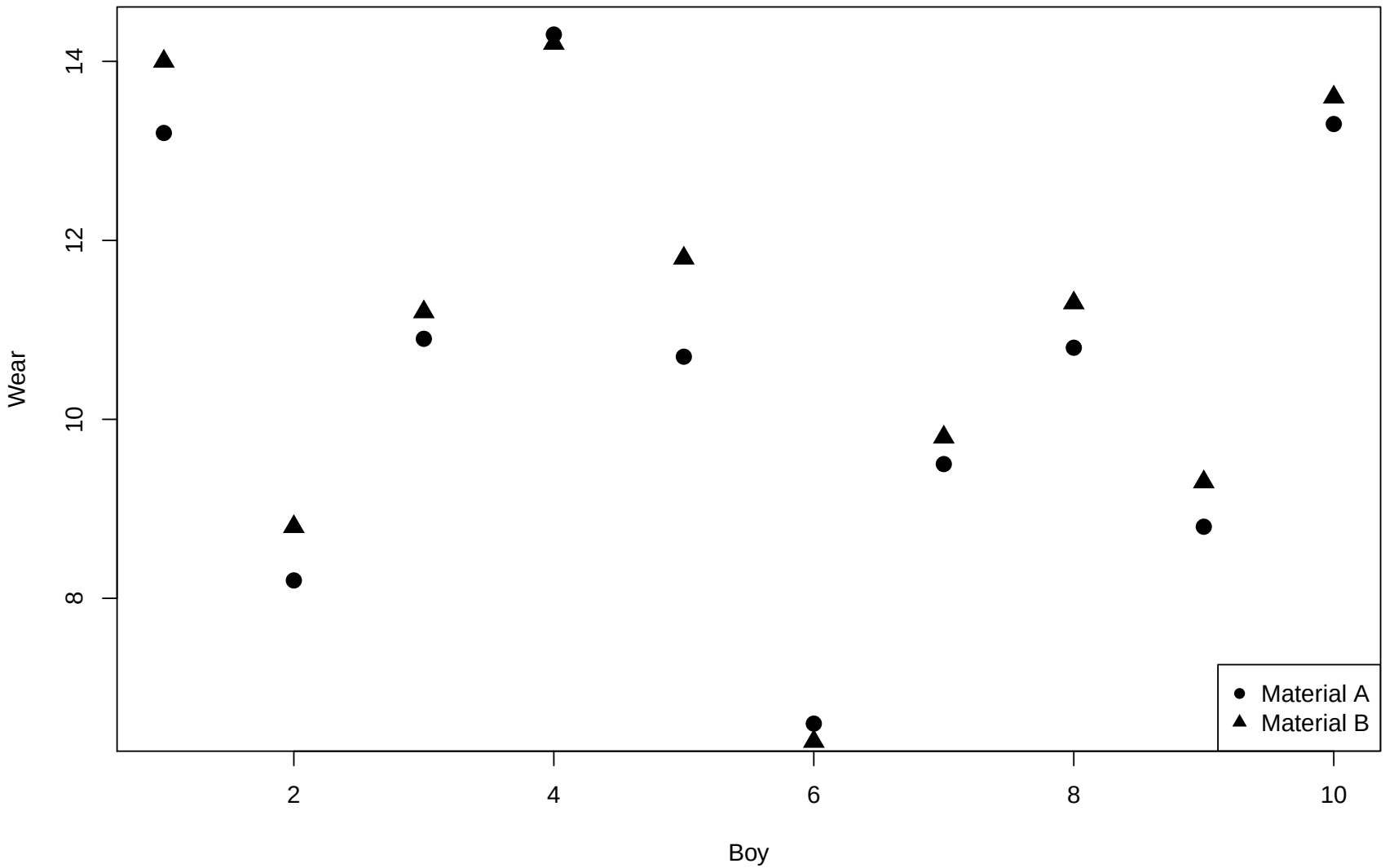
Randomized paired comparison

```
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
```

Randomized paired comparison

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



Randomized paired comparison

```
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
```

Boy's Shoe Experiment

- The sequence of coin tosses is one of $2^{10} = 1024$ equiprobable outcomes.
- To test H_0 the average difference of -0.41 observed observed can be compared with the other 1023 averages by calculating the average difference for each of 1024 arrangements of signs in:

$$\bar{d} = \frac{\pm 0.8 \pm 0.6 \cdots \pm 0.3}{10}$$

Randomized paired comparison

```
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
# generate all possible treatment assign
trtassign <- expand.grid(rep(LR, 10))

for(i in 1:N){
res[i] <- mean(as.numeric(trtassign[i,])*diff)
}
trtassign[1:2,]
```

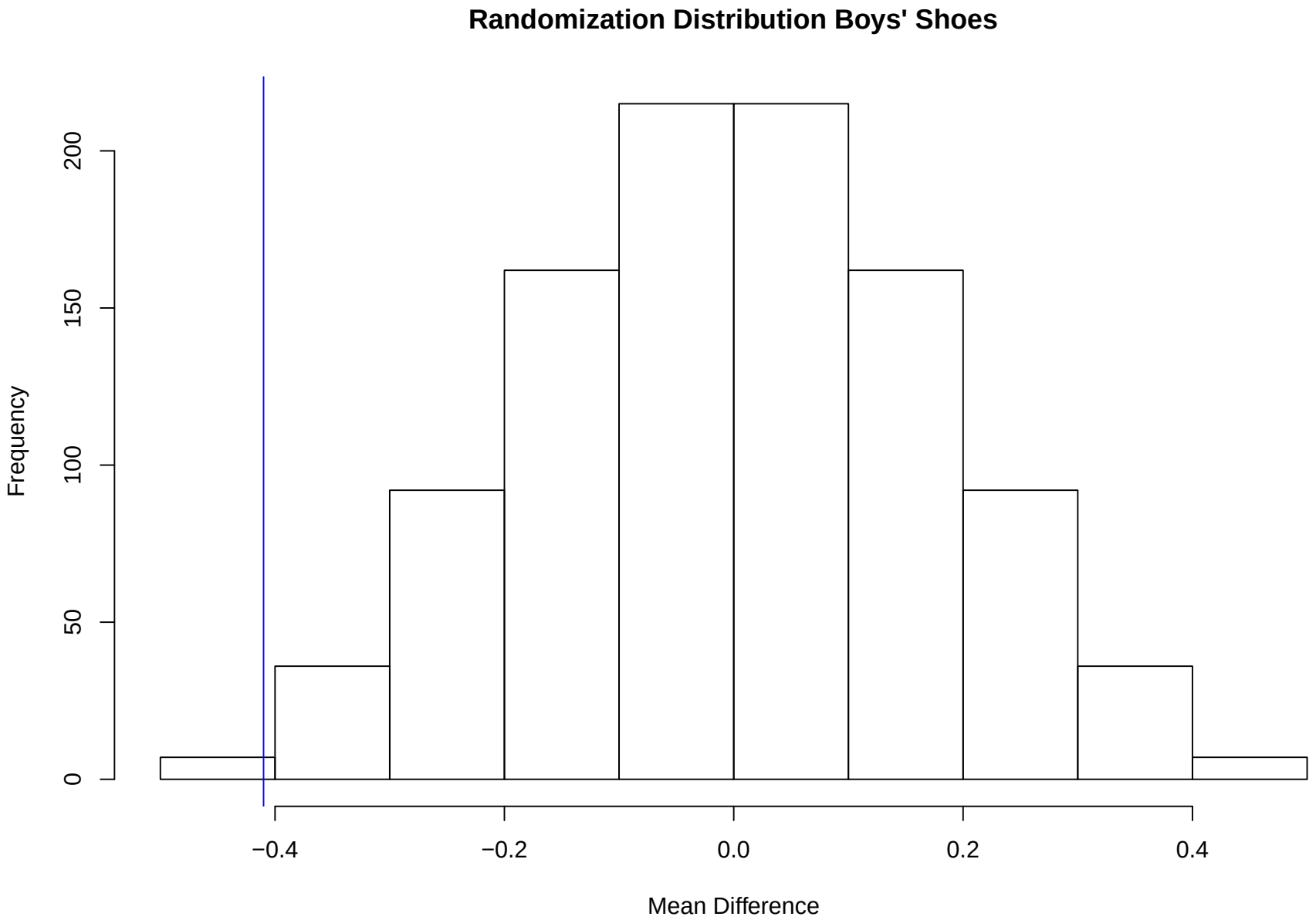
```
##   Var1 Var2 Var3 Var4 Var5 Var6 Var7 Var8 Var9 Var10
## 1   -1   -1   -1   -1   -1   -1   -1   -1   -1    -1
## 2    1   -1   -1   -1   -1   -1   -1   -1   -1    -1
```

```
res[1:2]
```

```
## [1] 0.41 0.25
```

Randomized paired comparison

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



Randomized paired comparison

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

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

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

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

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

$$P(t_9 < t).$$

Paired t-test

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.

Paired t-test

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

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

```
##
##  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
```

Paired t-test

This is the same as a one-sample t-test on the difference.

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

```
##
##  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
```

Paired t-test

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