

STA 305/1004

Jan. 20, 2016

# TODAY'S CLASS

TA office hours in STAT Aid  
Centre 12-2 on Jan. 26

- Hypothesis testing via randomization

Hw #1

Q.

① Brand A to have 8 judges  
and Brand B to have 7 judges.

For question #2 in Hw#1

We want \*

2. (c) What is the question?

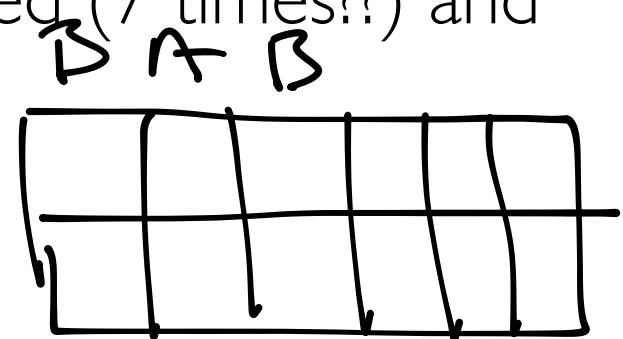
Is A better than B or

B better than A or  
A equal to B?

# EXPERIMENTAL DESIGN

- Assigning treatments randomly avoids any pre-experimental bias.
- 12 playing cards, 6 red, 6 black were shuffled (7 times??) and dealt

1st card black → 1st plot gets B



2nd card red → 2nd plot gets A

3rd card black → 3rd plot gets B

Completely randomized design

12 plots  
two trts.  
A / B.

# WHEAT YIELD DATA

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.

# EXPERIMENTAL PROCEDURE AND POTENTIAL OUTCOMES

Shuffled cards were dealt B, R, B, R, ..., fertilizers assigned to subplots

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

(3)  
= 924

Crops were grown and yields obtained:

exchangeability

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

- Imagine re-doing the experiment if  $H_0$  is true (no treatment effect)

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

- Crops are grown and wheat yields obtained:

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

$B$  under  
 $H_0$

# THE NULL DISTRIBUTION

- For each one of these compute  $\delta = \bar{y}_a - \bar{y}_b$
- treatment assignment possibilities*

- $\{\delta_1, \delta_2, \dots, \delta_{924}\}$  enumerates all potential pre-randomisation outcomes assuming no treatment effect.

*assume  $H_0$  is true*

$$H_0: \mu_A = \mu_B.$$

# THE NULL DISTRIBUTION

The null distribution, a probability distribution of experimental results if  $H_0$  were true can be described as

$$\begin{aligned}\hat{F}(x | H_0) &= P(|\bar{y}_b - \bar{y}_a| < x | H_0) \\ &= \#\{\delta_k \leq x\} / 924 \\ &= \frac{\sum_{i=1}^{\binom{12}{6}} I(\delta_k \leq x)}{\binom{12}{6}}\end{aligned}$$

$$I(\delta_n \leq x) = \begin{cases} 1 & \delta_n \leq x \\ 0 & \delta_n > x \end{cases}$$

Called the randomization 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 plots to the fertilizers.

# 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 occurred.
- A probability distribution for  $\bar{y}_a - \bar{y}_b$  can be calculated for each of the possible randomizations.

# 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.

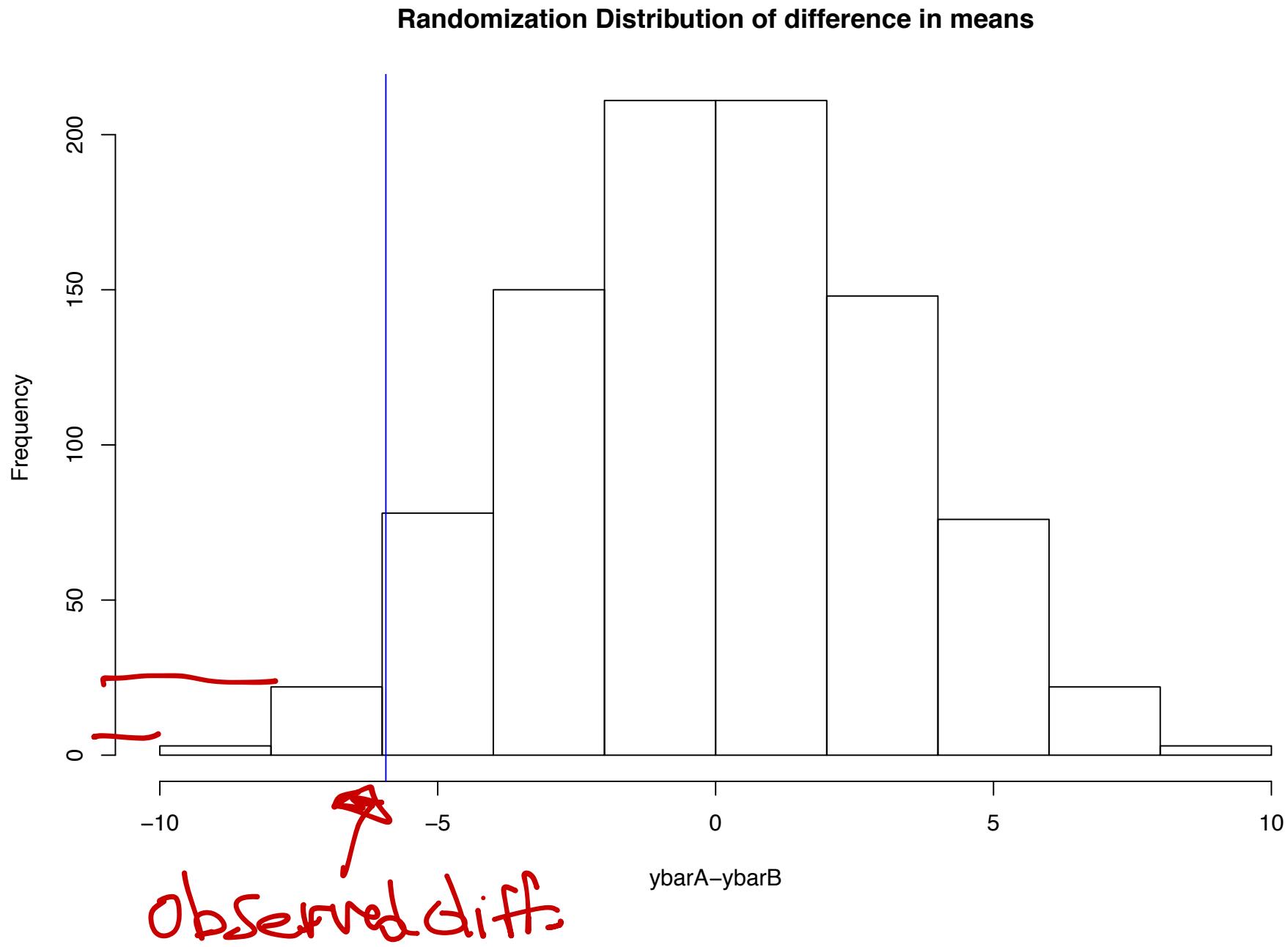
For example a plot that had a yield of 26.9 given fertilizer B would have the same yield if the plot received fertilizer A if  $H_0$  is true.

# R Code for Randomization Distribution

```
yA <- c(11.4,23.7,17.9,16.5,21.1,19.6)           ← Yield under A
yB <- c(26.9,26.6,25.3,28.5,14.2,24.3)          ← Yield under B
fert <- c(yA,yB) #pool data
N <- choose(12,6)                                     ← A → size of rand. clust.
res <- numeric(N) # store the results
#Generate N treatment assignments
library(combinat); index <- combn(1:12,6)           ← all possible
for (i in 1:N) {                                     ← B → treat assign.
  res[i] <- mean(fert[index[,i]])-mean(fert[-index[,i]]) }   ↑
hist(res,xlab="ybarA-ybarB",
     main="Randomization Distribution of difference in means")
observed <- mean(yA)-mean(yB) #store observed mean difference
abline(v=observed,col="blue") #add line at observed mean diff
```

B A B D B B      A A A A A AB  
Index[, 1]      - index[, 1]

# Randomization Distribution



# HYPOTHESIS TESTING

Is there any contradiction between  $H_0$  and the observed data?

$$\text{Calculate } 1 - \hat{F}(5.93 | H_0) = P(\bar{y}_a - \bar{y}_b < -5.93 | H_0)$$

*observed value  
of the diff.*

# HYPOTHESIS TESTING

- A p-value is the probability, under the null hypothesis of obtaining a more extreme than the observed result.

$$P\text{-}value = P(\bar{y}_a - \bar{y}_b < -5.93 | H_0)$$

- Small P-value implies evidence **against** null hypothesis.
- Large P-value implies no evidence **against** null hypothesis.
- If the P-value is large does this imply that the null is true?

## 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

26/924,

```
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  $\mu_A = \mu_B$ .

$$\alpha = P(\text{falsely reject } H_0)$$

depends  
in context.

$$\alpha = .05 \text{ or } \frac{1}{n}$$

## 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 justification for simply doubling the probability in one tail.

Let

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

$$I(x \geq t) = \begin{cases} 1 & x \geq t \\ 0 & x < t \end{cases}$$

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

$\sigma \cdot \sigma_3 \times 2 -$

# 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

250,000 = M e.g.).

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

Divisible by  $M+1$   
∴ include  
original  
sample -

## 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] 0.0283

approx. the same as  
the exact P-value.

Given a true p-value of  $p^*$   
and  $M$  draws from the  
randomization blistrn.

the large sample Standard error

$$\text{LS} \sqrt{\frac{p^*(1-p^*)}{M}}$$

- Has max value at  $P^* = 1/2$
- Then  $1/\sqrt{M}$  use  $M =$ 
  - If want  $SE = 0.001$   $250,000$

# HYPOTHESIS TESTING

- Is a small p-value evidence in favour of  $H_a$ ?
- Is a large p-value evidence in favour of  $H_0$ ?
- What does the p-value say about the probability that the null hypothesis is true? Try using Bayes' rule to figure this out.

$P(H_0) = 0.03$  (could use to  
calc.  $P(H_0 \text{ blates})$ .)

# BASIC DECISION THEORY

	H <sub>0</sub> True	H <sub>0</sub> False
Stat. Test	accept H <sub>0</sub>	correct
	reject H <sub>0</sub>	type I error
		type II error

p-value =  $P(\text{test stat} \geq \text{observed test stat} | H_0 \text{ true})$

$\alpha = P(\text{type I error})$  and  $\beta = P(\text{type II error})$

$1 - \beta = \text{power of test} = \text{prob. test rejects } H_0 \text{ correctly}$

# THE RANDOMIZATION P-VALUE

- An achievable P-value of the randomization test must be a multiple of  $\binom{k}{\frac{12}{6}} = \frac{k}{924}$ , where  $k = 1, 2, 3, \dots, 924$ .
- 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  $\alpha$  is not chosen as one of the achievable P-values but  $\frac{k}{924}$  is the largest achievable P-value less than  $\alpha$  then  $P(\text{Type I error}) < \alpha$ . *Conservative*.

$$\text{P-value} = \frac{\sum_{i=1}^{924} I(t_i \leq t^*)}{924} = p$$

$$\alpha = \frac{\kappa}{924}$$

$$P(\text{Type I}) = P(\text{P-value} \leq \alpha \mid H_0)$$

Shows that the rand-test (exact.)  $\equiv$   $= P\left(\frac{\sum_{i=1}^{924} I(t_i \leq t^*)}{924} \leq \frac{\kappa}{924}\right) = P\left(\sum_{i=1}^{924} I(t_i \leq t^*) \leq \kappa\right) = \kappa / 924 = \alpha.$

# 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 or trimmed means are examples of other test statistics.

## Other Test Statistics

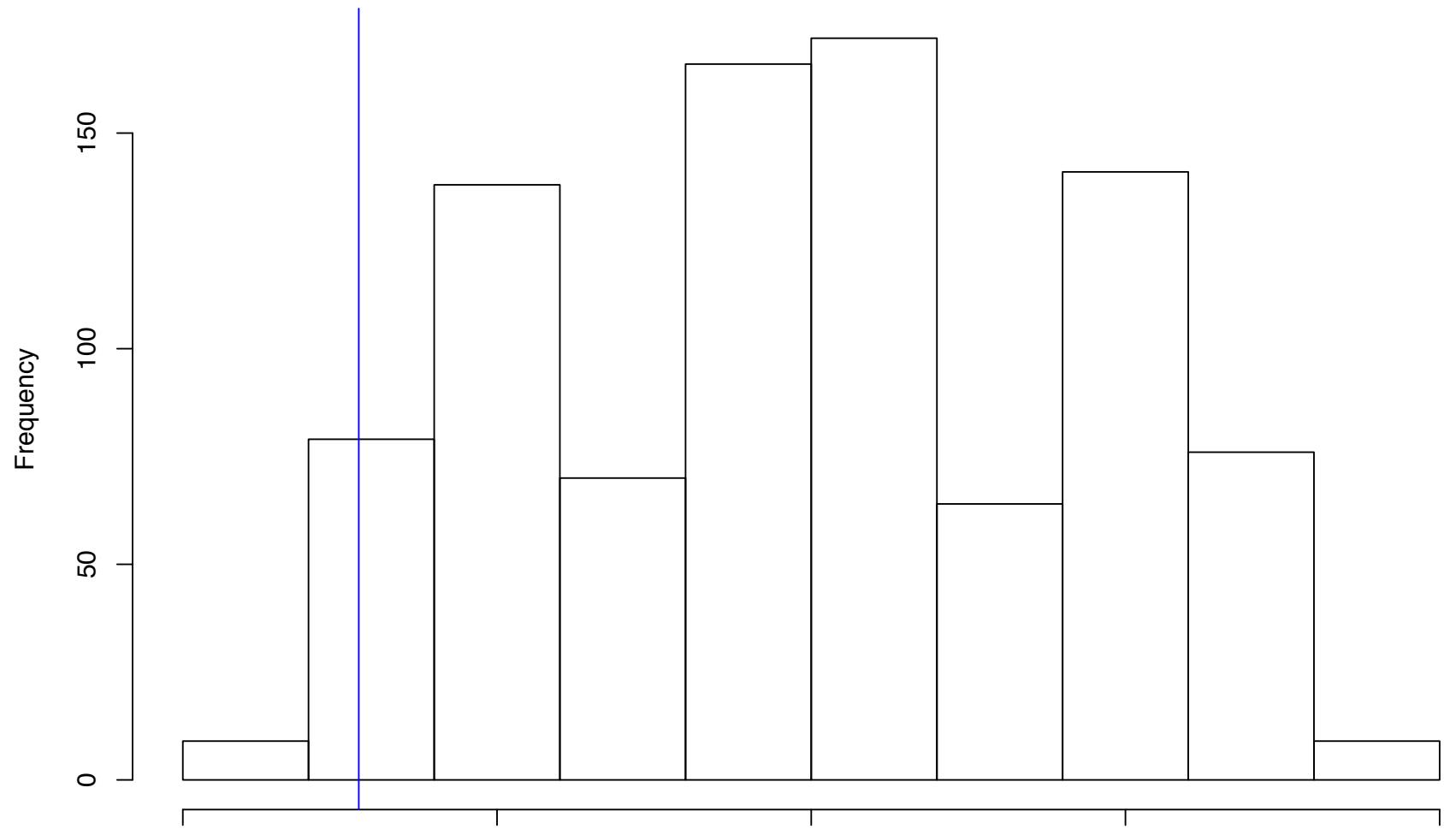
The randomization distribution of the difference in group medians can be obtained by modifying the R code used for the difference in group means.

```
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]])
}
```

Swap mean  
for median .

## Other Test Statistics

Randomization Distribution of difference in medians



Obs. median diff

## The two-sample t-test

If the two wheat yield samples are independent random samples from a normal distribution with means  $\mu_A$  and  $\mu_B$  but the same variance then the statistic

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

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  $\sigma^2$  then

$n_A = \text{Sample Size A}$

$$n_B = \text{Sample Size for } B: \frac{\bar{y}_A - \bar{y}_B - \delta}{S \sqrt{(1/n_A + 1/n_B)}} \sim t_{n_A + n_B - 2},$$

is called the two sample t-statistic.

$$\begin{aligned} Y_{A_1} &\sim \dots & Y_{A_6} \\ &\sim N(\mu_A, \sigma) \\ Y_{B_1} &\sim \dots & Y_{B_6} \\ &\sim N(\mu_B, \sigma). \end{aligned}$$

Sample  
SD  
Assuming  
 $\mu_A = \mu_B$ .

## 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  $H_0$ .

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

$$\sigma = \sigma$$

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
```

cdf of  $t_{10}$   $\frac{6+6-2}{11}$

$n_A + n_B - 2$

## 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")
```

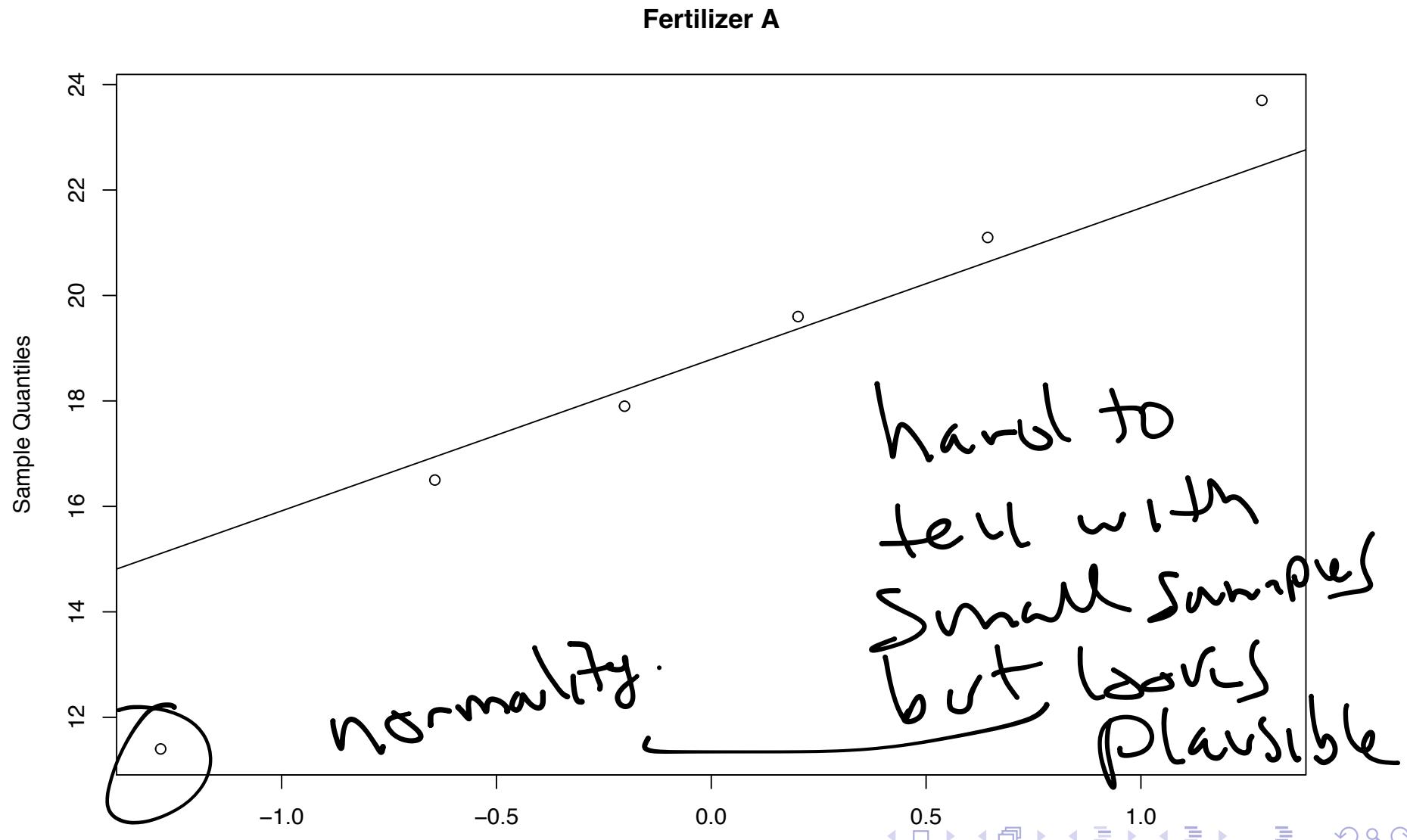
```
##  
## Two Sample t-test  
##  
## data: yA and yB  
## t = -2.1793, df = 10 p-value = 0.02715  
## alternative hypothesis: true difference in means is less than 0  
## 95 percent confidence interval:  
##       -Inf -0.9987621  
## sample estimates:  
## mean of x mean of y  
## 18.36667 24.30000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

Very Similar to  
randomization  
test

## The two-sample t-test

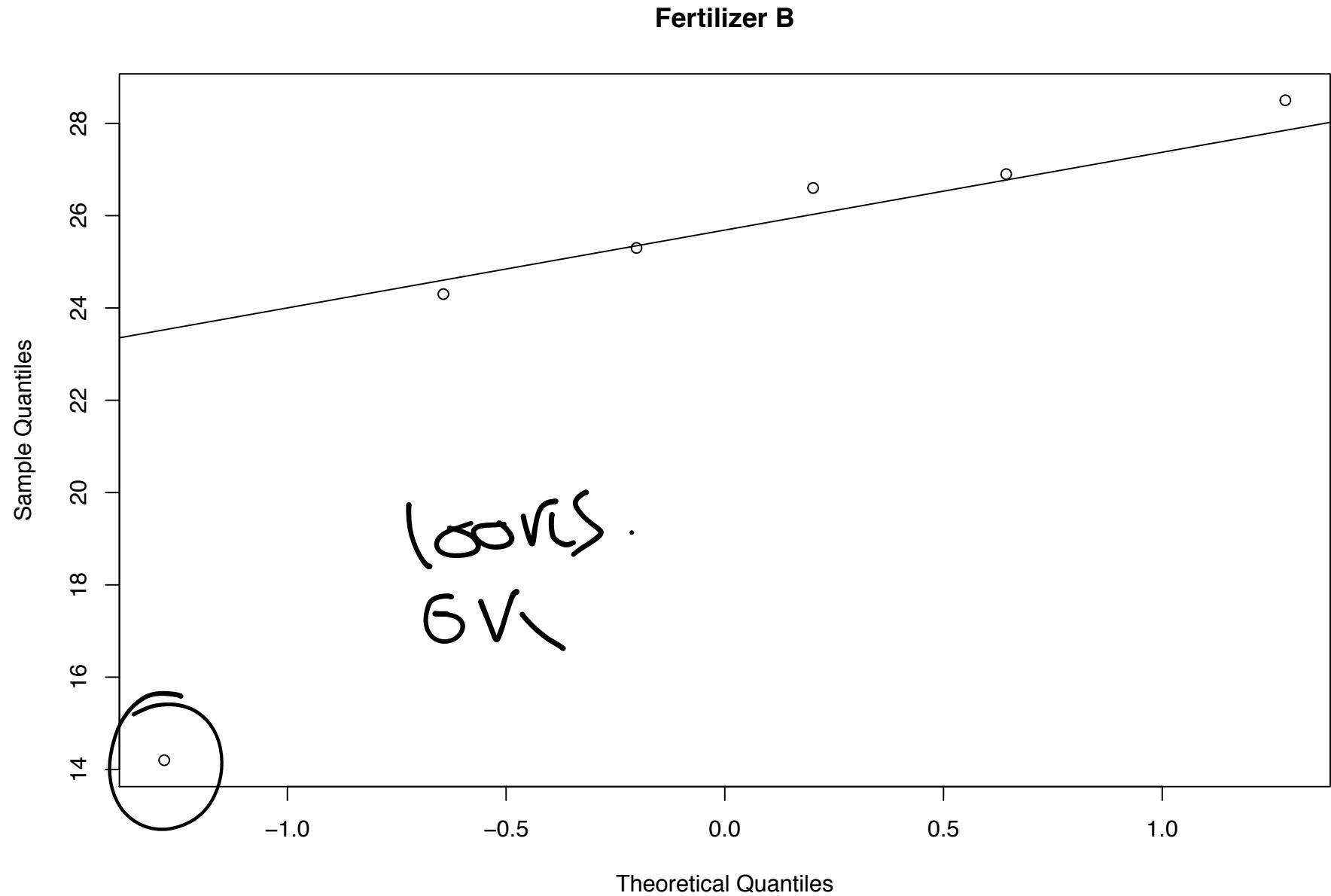
The assumption of normality can be checked using normal quantile plots, although the t-test is robust against non-normality.

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



## The two-sample t-test

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



# TWO-SAMPLE T-TEST VS. RANDOMIZATION TEST

The p-value from the randomization test and the p-value from two-sample t-test are almost identical.

The randomization test neither depends on normality nor independence.

# TWO-SAMPLE T-TEST VS. 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

Objective: Is material B better than material A?

- During test some boys scuffed their shoes more than others.
- each boy's two shoes were subjected to the same treatment.
- 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

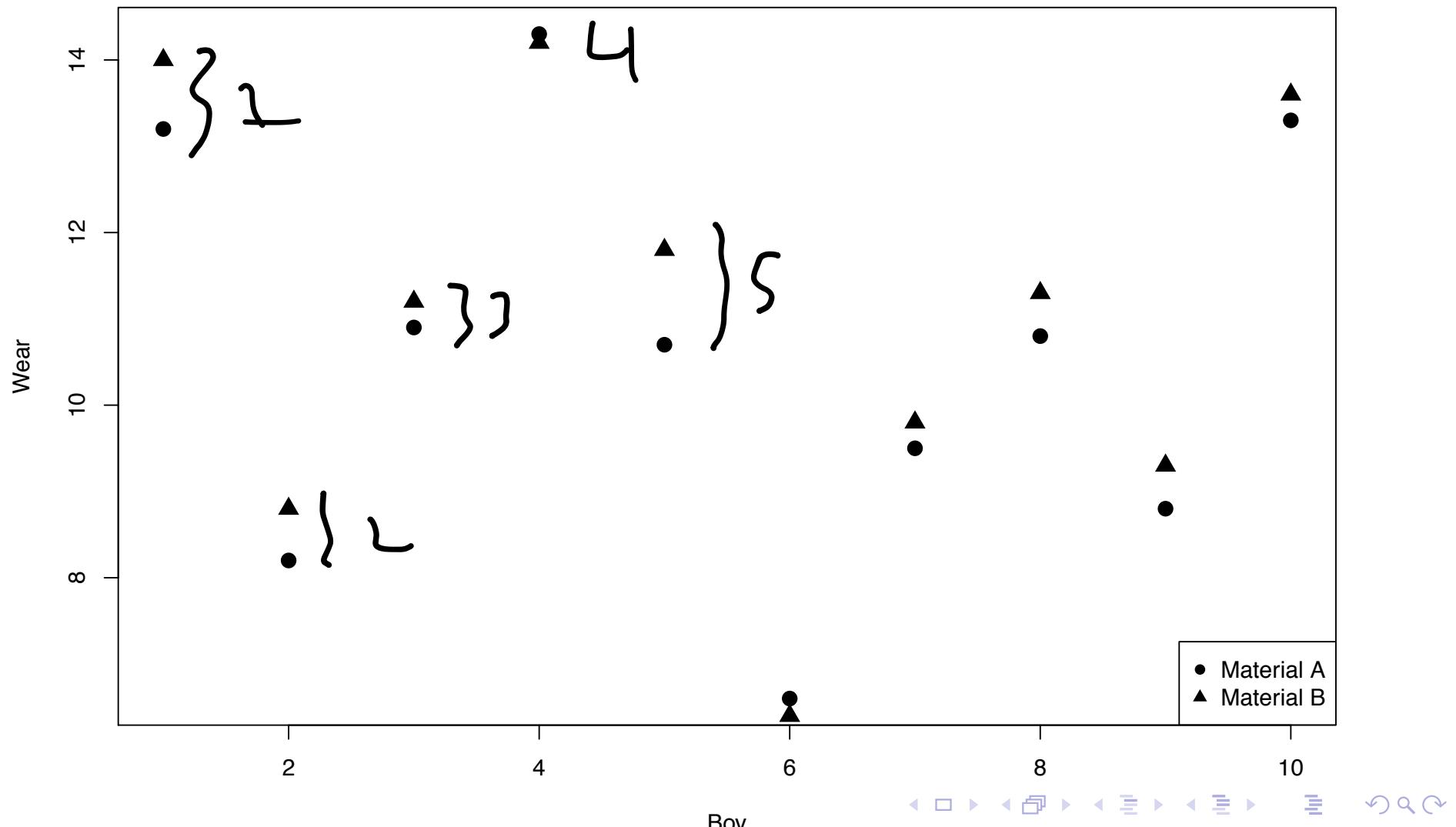
```
library(BHH2)
data(shoes.data)
shoes.data
```

##	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

$\text{matA} \rightarrow L$   
 $\text{MATB} \rightarrow R$   
 $\text{MATA} \rightarrow R$   
 $\text{MATB} \rightarrow L$

## 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

$H \rightarrow R$   
 $T \rightarrow L$

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

## [1] -0.41

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

$A - B$   $\neq$

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

mean of diff  
 $-0.8 + -0.6$   
 $+ -0.3$   

---

 $= -0.41$

Experimental units Left / Right

Show.

- how to assign treatments to expt. units?
- flip a coin.  
 $\begin{cases} H \rightarrow R \\ T \rightarrow L \end{cases}$

# BOY'S SHOE EXPERIMENT



$$2 \times 2 \times 2 \cdot \cdot \cdot \times 2 = 2^{10} = 1024$$

arrang(m)nt  
of +/-

In the  
diff.

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

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

## 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))
```

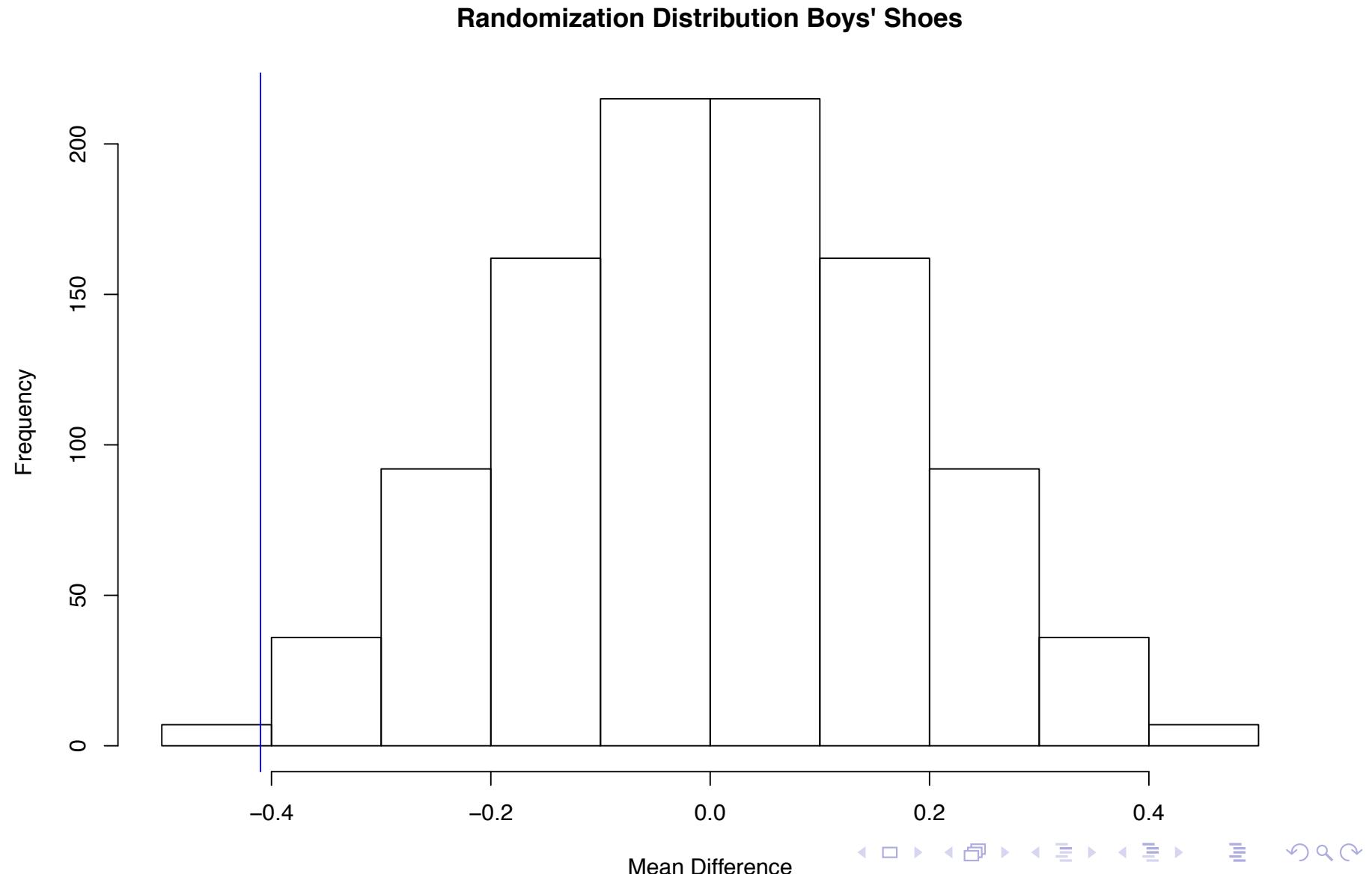
```
for(i in 1:N){
  res[i] <- mean(as.numeric(trtassign[i,])*diff)
}
```

(0.24  
values)  
diff.  
(in rand. diffn..)

↑  
Multiplies by +/− 1.  
 $\pm 0.8 \pm 0.6 \pm \dots \pm 0.3$   
1024

## Randomized paired comparison

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



## Randomized paired comparison

-0.411

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

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

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

1024 .  
= 7 / 1024 .

## 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.

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

```
##  
## One Sample t-test → data = diff  
##  
## 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)
```

