STA 305/1004

Jan. 20, 2016

TODAY'S CLASS

TA office hour in STAT Aid center 12-2 on Jan 26.

- Hypothesis testing via randomization

2. HW #1

(*) Brand A to have 8 judges and Brand B to have 7 judges.

For q2 in HW#1 we want (*)

2. (c) What is the question? Is A better than B or B better than A or A differ from B?

EXPERIMENTAL DESIGN

- Assigning treatments randomly avoids any pre-experimental bias.

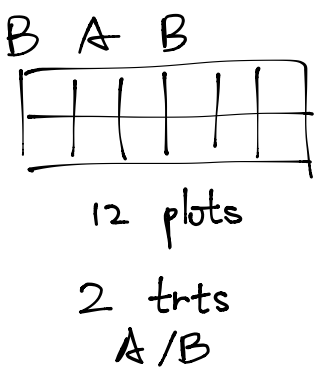
- 12 playing cards, 6 red, 6 black were shuffled (7 times??) and dealt

1st card black ⟶ 1st plot gets B

2nd card red ⟶ 2nd plot gets A

3rd card black ⟶ 3rd plot gets B

Completely randomized design



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

$\binom{12}{6} = 924$

Crops were grown and yields obtained:

exchangability

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 under H_0

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

THE NULL DISTRIBUTION

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

- $\{\delta_1, \delta_2, ..., \delta_{924}\}$ 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 be described as

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

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

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.

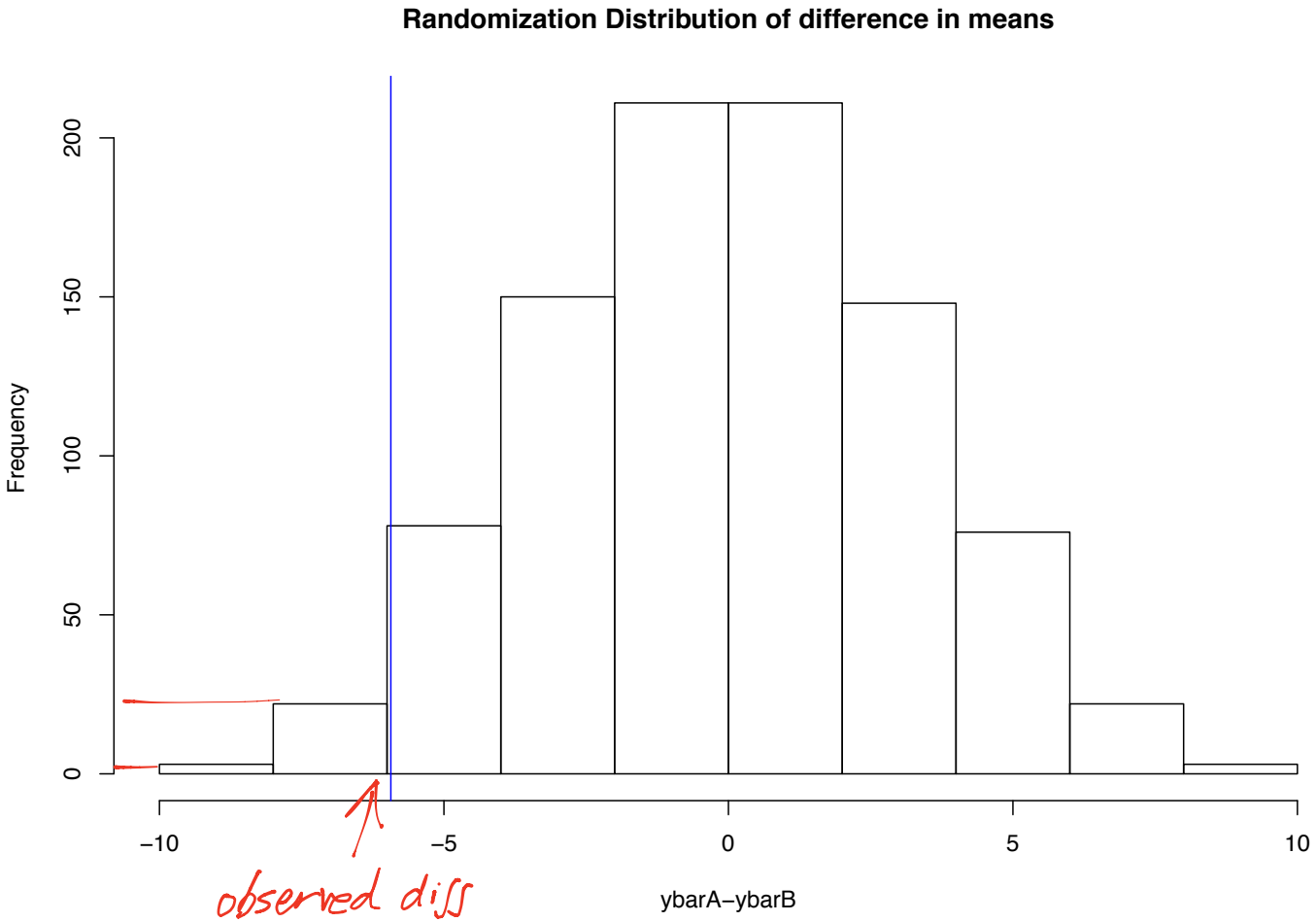
R Code for Randomization Distribution

```
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)  <- size of random data
res <- numeric(N) # store the results
#Generate N treatment assignments
library(combinat); index <-combn(1:12,6)  - all possible treatment assign
for (i in 1:N)                  A          B
{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
```

BABBBB AAAAAB

index[,1] -index[,1]

Randomization Distribution



HYPOTHESIS TESTING

Is there any contradiction between H_0 and the observed data?

Calculate $1-\hat{F}(5.931 \mid H_0) = P(\bar{y}_a - \bar{y}_b < -5.931 \mid H_0)$

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.

$$P-value = P\left(\bar{y}_a - \bar{y}_b < -5.93 \mid H_0\right)$$

- 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
```

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

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

$\frac{26}{924}$

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 $\mu_A = \mu_B$.

α = P(falsely reject H_0)

α = .05 or .01

.1 ??

depends on context

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

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

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

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
```

0.03 ×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, $i = 1, ..., M$ 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

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

obviously $M+1$ since include original sample

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] 0.0283
```

approx. the same as the exact p-value

Given a true p-value of p^* and M draws from the randomization distn. the large sample standard error is $\sqrt{\frac{p^*(1-p^*)}{M}}$

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

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.

$P(H_0 \mid \text{data})$

BASIC DECISION THEORY

Truth

	H_0 True	H_0 False
accept H_0	correct	type II error
reject H_0	type I error	correct

stat test

p-value=P(test stat>=observed test stat|H_0 true)

α=P(type I error) and β=P(type II error)

1-β=power of test – prob test rejects correctly

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, 3, …,924.
- If we choose a significance level of $\alpha = \frac{k}{924}$ that is one of the achievable P-values then P(Type I error) = α.
- The randomization test is an exact test.
- If α is not chosen as one of the achievable P-values but $\frac{k}{924}$ is the largest achievable P-value less than α then P(Type I error) < α.

Conservative

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

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

$$P(\text{Type I error}) = P(\text{p-value} \le \alpha \mid H_0) = P\left(\frac{\sum I(t_i \le t^*)}{924} \le \frac{K}{924}\right) = P(\Sigma I(t_i \le t^*) \le K) = \frac{K}{924} = \alpha$$

Basically shows that randomization test is exact.

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.

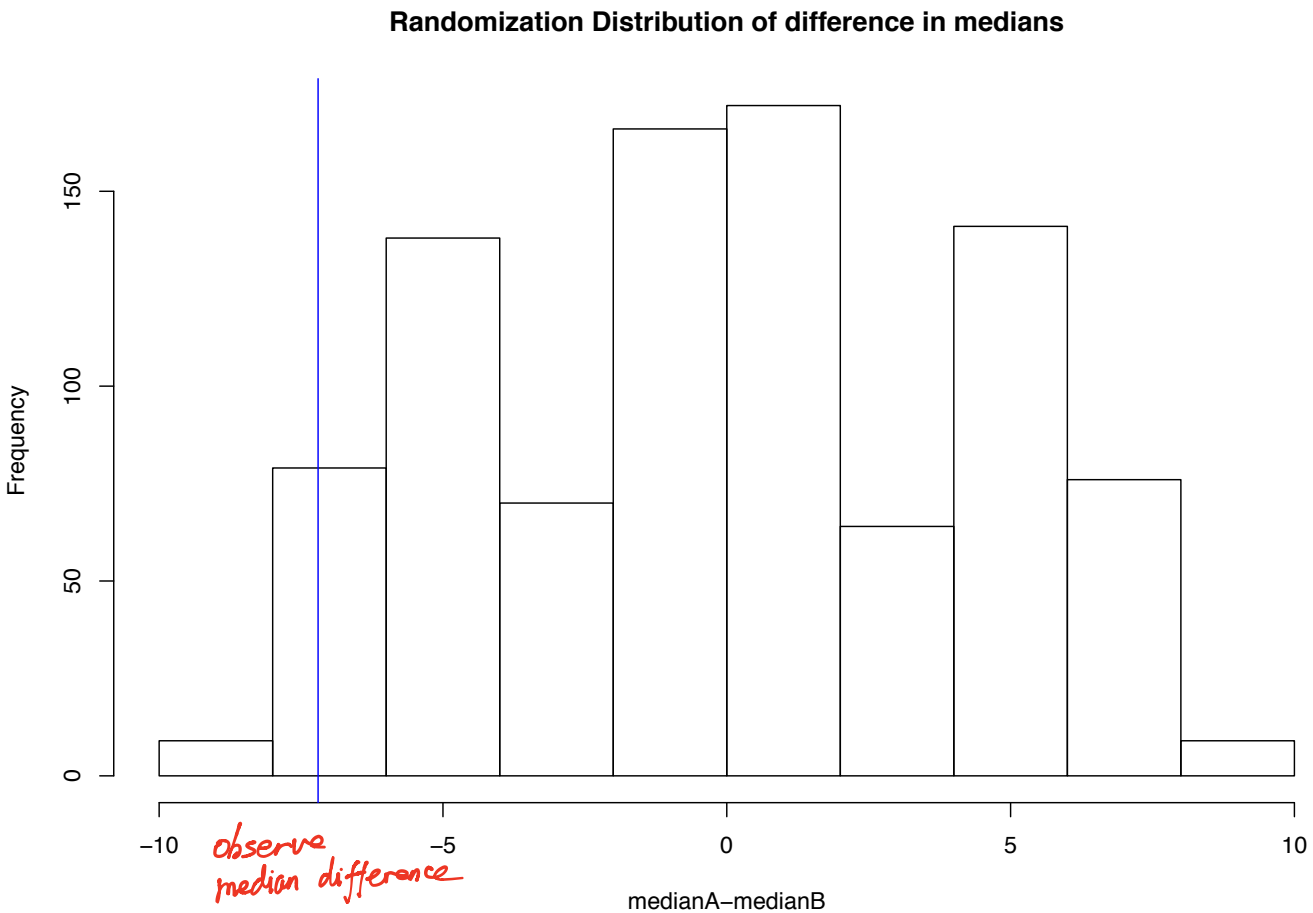
Other Test Statistics

The 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]])
}
```

swap mean for median

Other Test Statistics



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

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

$Y_{A1}, \cdots, Y_{A6} \sim N(\mu_A, \sigma)$

$Y_{B1}, \cdots, Y_{B6} \sim N(\mu_B, \sigma)$

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

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

sample SD assuming $\mu_A = \mu_B$.

for σ^2 then

n_A = sample size for A

n_B = " " " 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.

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.

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

$\delta = 0$

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
```

deg of freedom of t_{10}

$6+6-2=11$

$n_A + n_B - 2$

```
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")
```

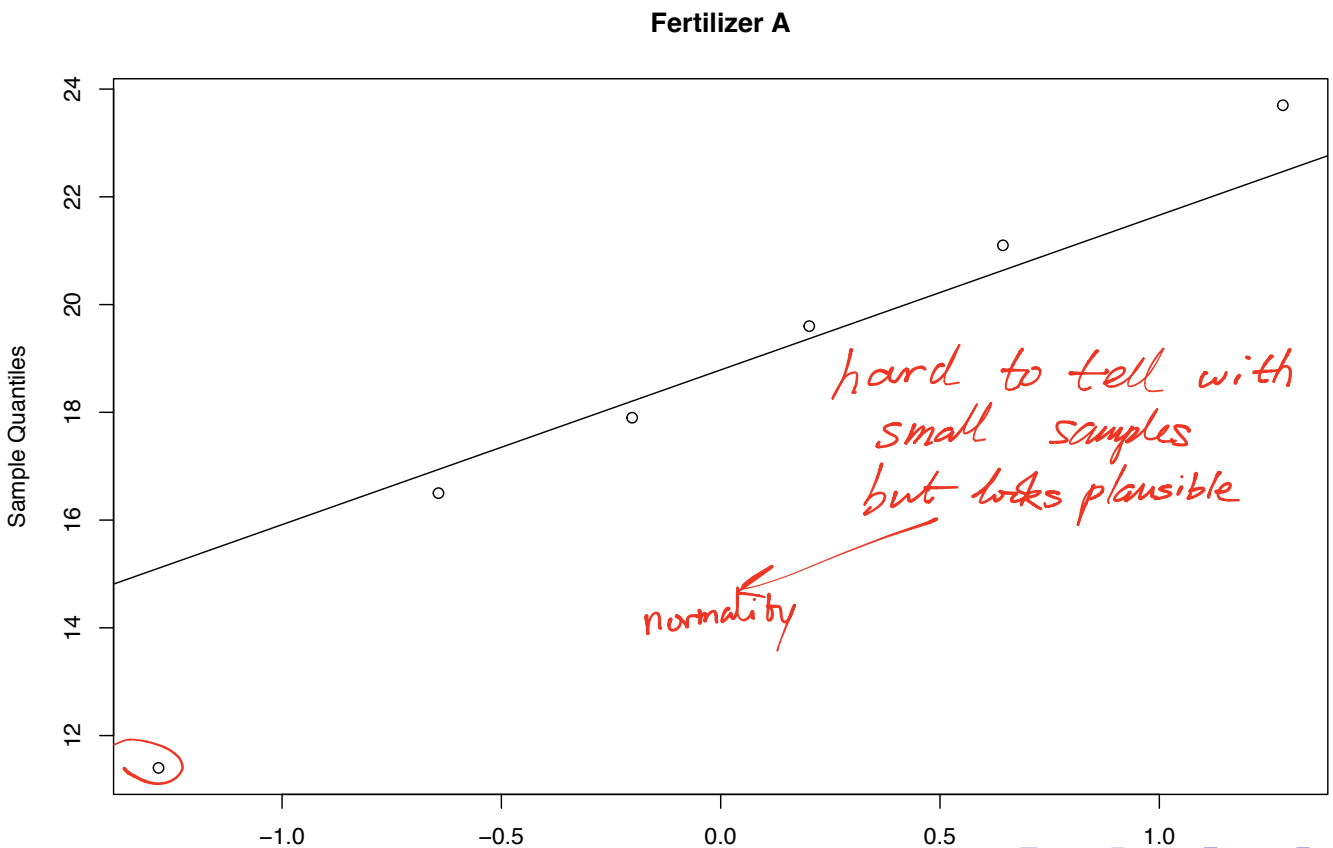
```
##
##  Two Sample t-test
##
## data:  yA and yB
## t = -2.1793, df = 10, p-value = 0.02715
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.9987621
## sample estimates:
## mean of x mean of y
##  18.36667  24.30000
```

very similar to randomization test

The two-sample t-test

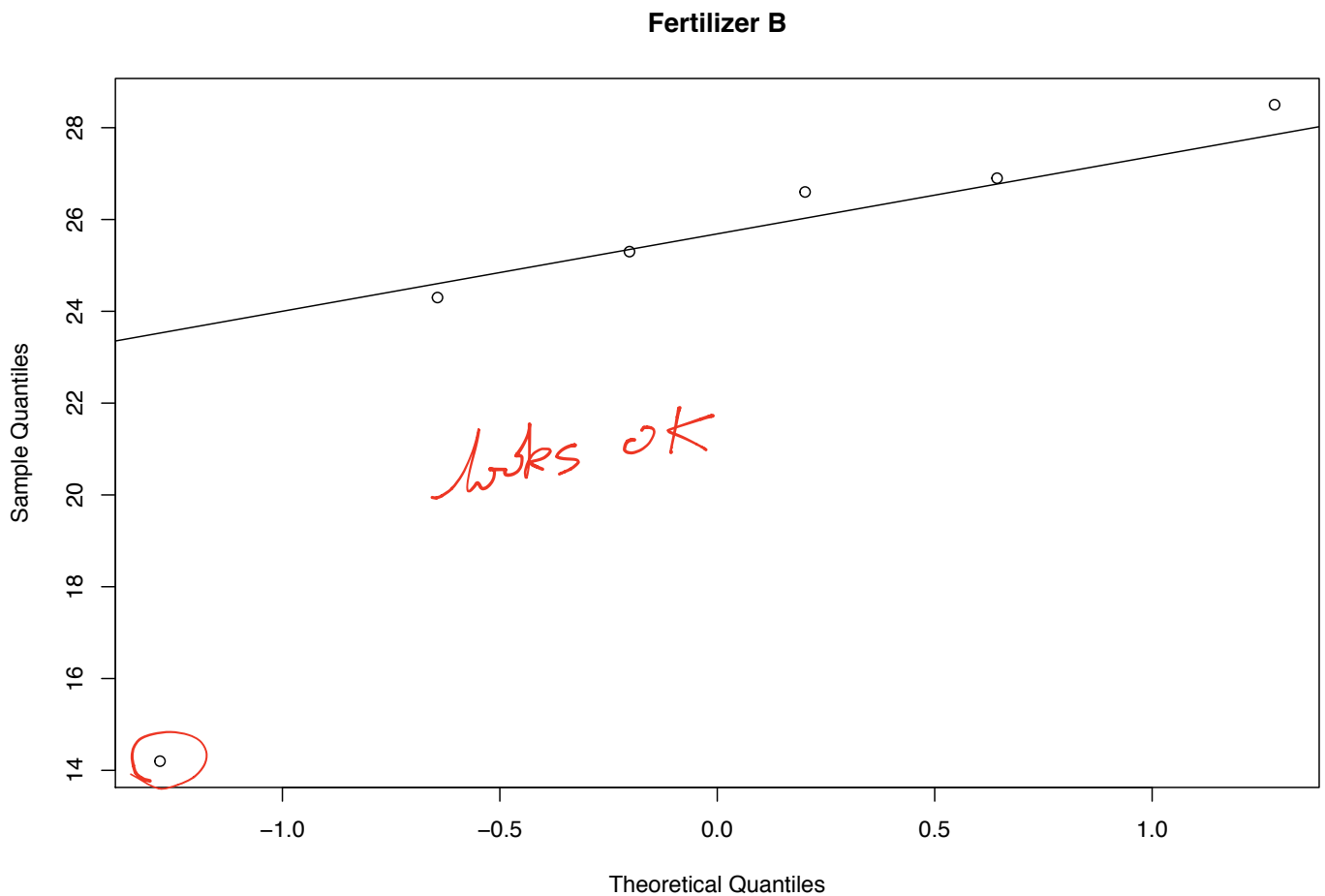
The assumption of normality can be checked using normal quantile plots, although the t-test is robust against non-normality.

```
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**.

basically you can swap the labels

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
```

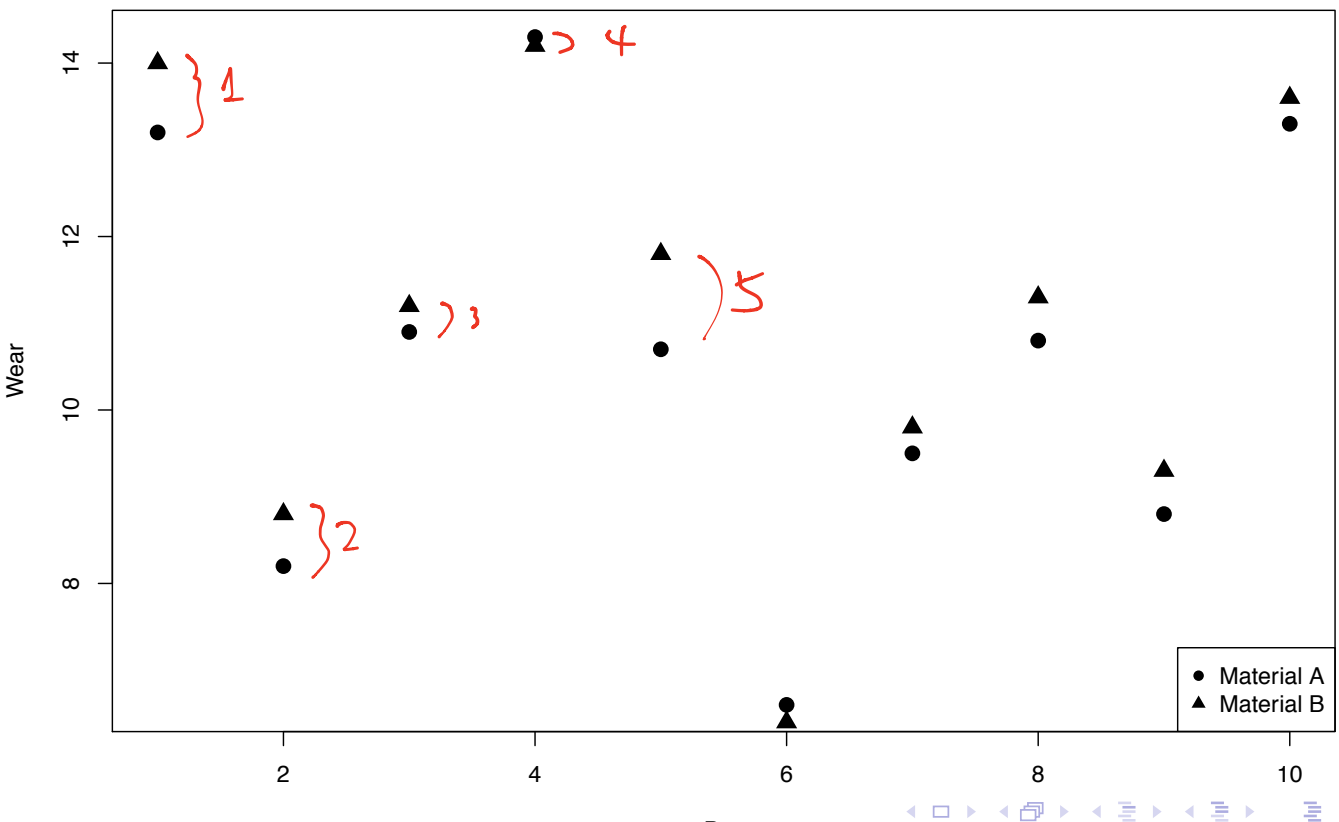
```
##    boy matA sideA matB sideB
## 1    1 13.2     L 14.0     R
## 2    2  8.2     L  8.8     R
## 3    3 10.9     R 11.2     L
## 4    4 14.3     L 14.2     R
## 5    5 10.7     R 11.8     L
## 6    6  6.6     L  6.4     R
## 7    7  9.5     L  9.8     R
## 8    8 10.8     L 11.3     R
## 9    9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

what side got A is decided by flipping a coin

- experimental units left/Right shoe
- how to assign treatments to expt units
- flip a coin < H → R
  T → 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

```
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, Δ

```
##    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
```

mean of diffs

$$\frac{-0.8 + -0.6 + \cdots + -0.3}{10} = -0.41$$

BOY'S SHOE EXPERIMENT

in the diff

$2 \times 2 \times 2 \times \cdots \times 2 = 2^{10} = 1024$ arrangements of +/−

- 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:

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

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)
}
```

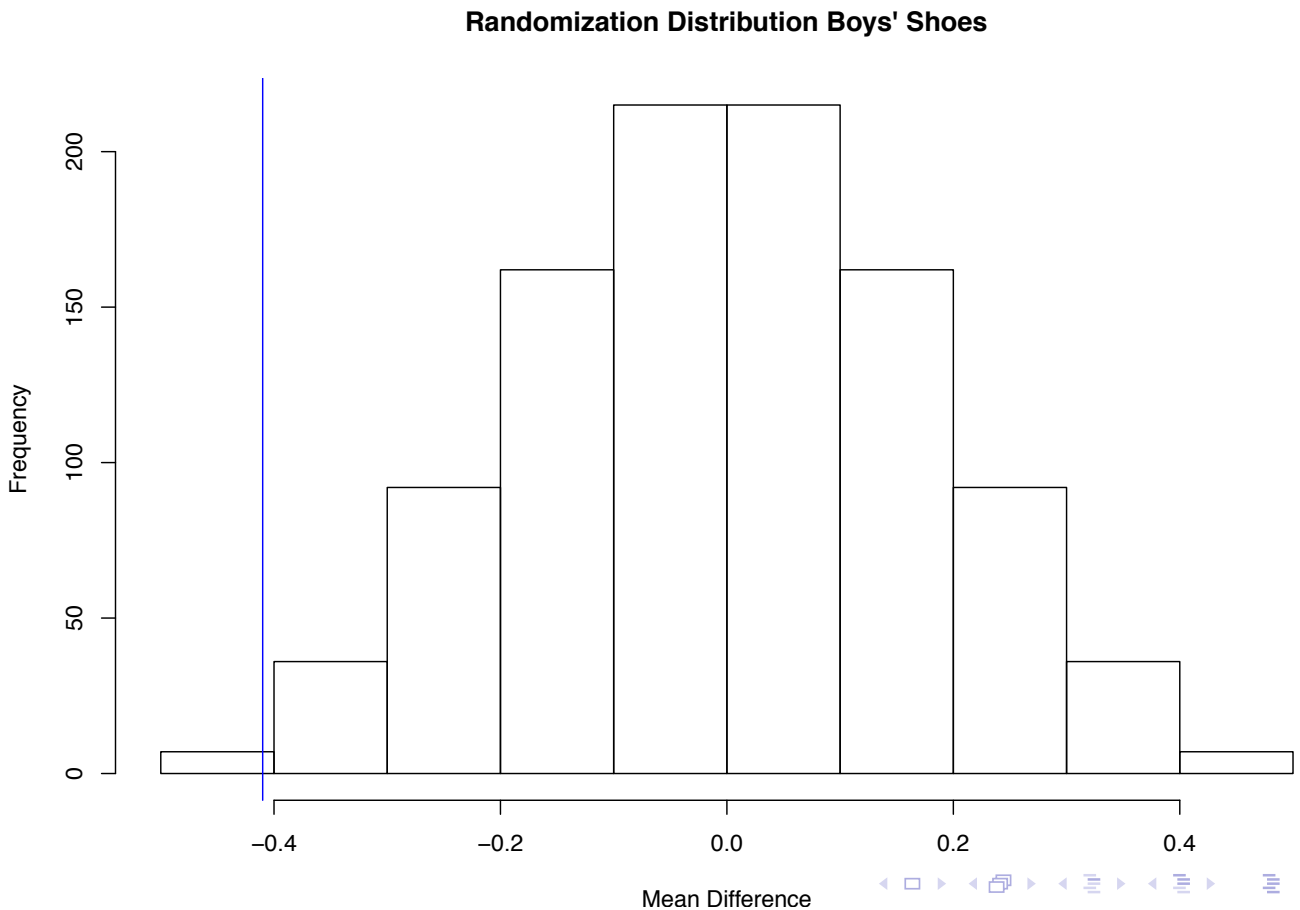
1024 values diff in rand dist'n

multiplies by +/-1

$$\frac{\pm 0.8 \pm 0.6 \pm \cdots \pm 0.3}{1024}$$

Randomized paired comparison

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



Randomized paired comparison

-0.41

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

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

1024

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

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

$= \frac{7}{1024}$

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.

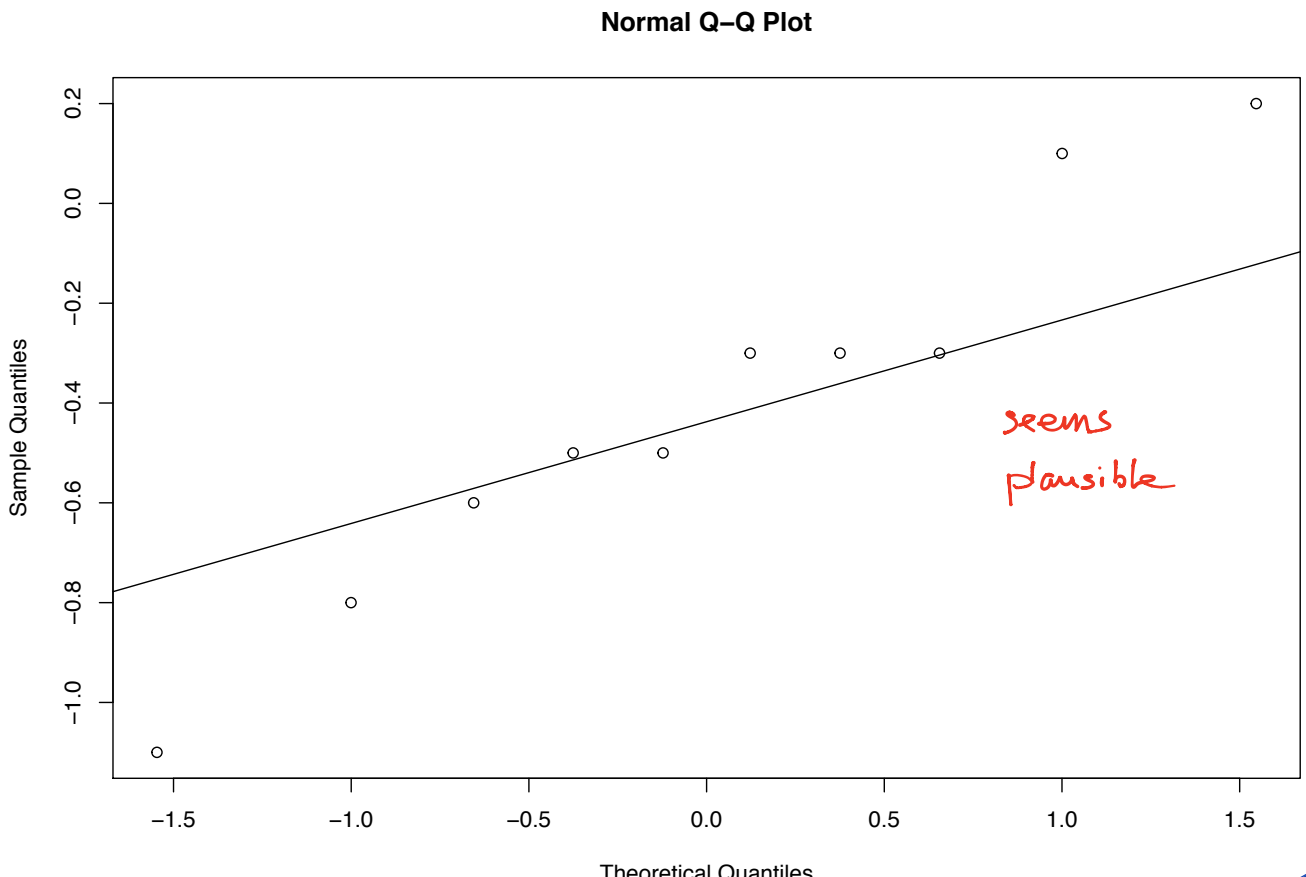
```
# same as a one-sample t-test on the diff
t.test(diff,alternative = "less")
```

data=diff

```
##
##  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)
```



END