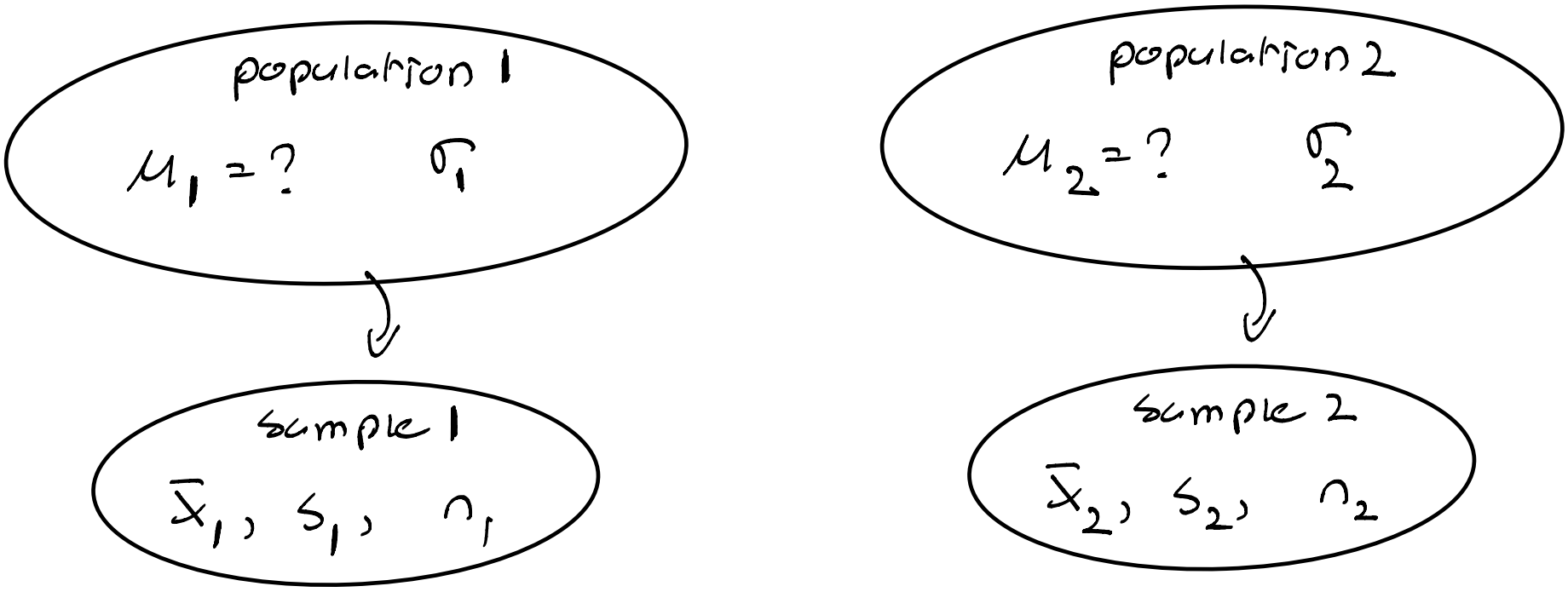
STA258H5
University of Toronto Mississauga

Al Nosedal, Asal Aslemand and Omid Jazi

Winter 2023

CONFIDENCE INTERVALS

- For two means
- For two proportions
- [illegible]



Interested in the difference of population means

$$\mu_1 - \mu_2 \quad (\text{or } \mu_2 - \mu_1)$$

This difference gives information on whether one population has a larger/smaller mean or whether they are equal

Comparing Means with Independent Samples

Introduction to the Independent-Measures Design

Consider research studies that compare two separate groups. For example:

- A social psychologist may want to compare men and women in terms of their political attitudes.
- An educational psychologist may want to compare two methods (regular, intervention) for teaching statistics.

Independent-measures research design or a Between-subject design:

- A research design that uses a separate group of participants for each treatment condition, or for each population.
- The goal is to compare population quantitative means between two completely separate groups.
- The researchers collect data for each group and estimate the sample means.

Introduction to the Independent-Measures Design (cont.)

There are two possible explanations for the difference between the two groups (their sample means):

- It is possible that there really is a difference between the two groups so that one group mean is different from another group mean.
- It is possible that there is no difference between the two groups and the mean difference obtained in the study is simply the result of sampling error.

Comparing Means of Independent Samples (Independence Assumptions)

1) Independent Response Assumption:
Within each group, we need independent responses from the cases. We cannot check that for certain, but randomization provides evidence of independence. So, we need to check the following: Randomization Condition:

- The data in each group should be drawn independently and at random from a population or generated by a completely randomized designed experiment.
- 10% Condition: We usually don't check this condition for differences of means. We'll check it only if we have a very small population or an extremely large sample.

Comparing Means of Independent Samples (Independence Assumptions)

2) Independent Groups Assumption:
To use the independent samples t methods, the two groups we are comparing must be independent of each other (that is, the two groups should be unrelated to each other). This is often referred to as independent samples t-test.

Comparing Means of Independent Samples (Normal Population Assumptions)

We should check the assumption that the underlying populations of individual responses are each Normally distributed. Nearly Normal Condition:

- We must check this for both groups; a violation for either one violates the condition.
- The Normality assumption matters most when sample sizes are very small.
- For $n < 10$ in either group, this method should not be used if the histogram or Normality plots show clear skewness.
- For n's of 10 or so, a moderately skewed histogram is okay. But, for strongly skewed data or data containing outliers this method should be avoided.
- For larger samples $n \geq 20$, data skewness is less of an issue - but, we still need to check if there are any outliers in the data, extreme skewness, and multiple modes.

Conditions for inference comparing two means

- We have two SRSs, from two distinct populations. The samples are independent. That is, one sample has no influence on the other. Matching violates independence, for example. We measure the same response variable for both samples.
- Both populations are Normally distributed. The means and standard deviations of the populations are unknown. In practice, it is enough that the distributions have similar shapes and that the data have no strong outliers.

1/ When σ_1 and σ_2 are both known

CI for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

← Standard error

margin of error

Very rare to use this because in reality we often don't know σ_1 and σ_2

2/ When σ_1 and σ_2 are both unknown

2.1 case 1 : $\sigma_1 = \sigma_2$ (equal St. devs)

$(\sigma_1^2 = \sigma_2^2)$ (equal variances)

CI for $\mu_1 - \mu_2$

pooled method

pooled sample St. dev

df

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2,\ \alpha/2)} \cdot \underbrace{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}_{\text{standard error}}$$

$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad}_{\text{margin of error}}$

where

pooled sample variance

$$s_p^2 = \frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}$$

variance from the aggregated sample (weighted average accomodating for sample sizes)

pooled sample St. dev

$$s_p = \sqrt{s_p^2}$$

$n_1 - 1 + n_2 - 1 = n_1 + n_2 - 2$

Comparing Two Populations Means: Independent Sampling (Equal Variances Assumed)

Consider two independent populations with unknown means μ_1 and μ_2, and unknown standard deviations σ_1 and σ_2 ($\sigma_1 = \sigma_2$), respectively. We can make an inference about their mean difference $\mu_1 - \mu_2$ by using the difference between their point estimates (sample means): $\bar{Y}_1 - \bar{Y}_2$. When the assumptions and conditions are met,

$$\frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}},$$

can be modelled by a $t(\nu)$ distribution; where $\nu = n_1 + n_2 - 2$ and $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$.

Comparing Two Populations Means: Independent Sampling (Equal Variances Assumed) (cont.)

Conditions Required for Valid Inference about $\mu_1 - \mu_2$:

1. The two samples are randomly selected in an independent manner from the two target populations.
2. Both sampled populations have distributions that are approximately Normal.
3. The population variances are equal (e.g., $(\sigma_1 = \sigma_2)$).

Confidence Intervals for μ1−μ2 (with equal variances)

Parameter : $\mu_1 - \mu_2$.

Confidence interval ($\nu =$ df) :

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{\alpha/2}(\nu) S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where $\nu = n_1 + n_2 - 2$ and $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$
(requires that Normal samples are independent and the assumption that $\sigma_1^2 = \sigma_2^2$). The critical value $t_{\alpha/2}(\nu)$ depends on particular confidence level and the number of degrees of freedom.

Example

Comparing Two Population Means Managerial Success Indexes for Two Groups (With Equal Variances Assumed)
Behavioural researchers have developed an index designed to measure managerial success. The index (measured on a 100-point scale) is based on the manager's length of time in the organization and their level within the term; the higher the index, the more successful the manager. Suppose a researcher wants to compare the average index for the two groups of managers at a large manufacturing plant. Managers in group 1 engage in high volume of interactions with people outside the managers' work unit (such interaction include phone and face-to-face meetings with customers and suppliers, outside meetings, and public relation work). Managers in group 2 rarely interact with people outside their work unit. Independent random samples of 12 and 15 managers are selected from groups 1 and 2, respectively, and success index of each is recorded.

Example

Comparing Two Population Means Managerial Success Indexes for Two Group (With Equal Variances Assumed)
Note: The response variable is "Managerial Success Indexes".

- Managerial success indexes is a continuous quantitative variable, measured on 100-point scale.

The explanatory variable is "Type of group".

- Type of group (Group 1: Interaction with outsiders, Group 2: Fewer interactions) is a nominal categorical variable.

R Code

```r
# Importing data file into R;

success=read.csv(file="success.csv",header=TRUE);

# Getting names of variables;

names(success);

# Seeing first few observations;

head(success);

# Attaching data file;
attach(success);
```

R Code

```
## [1] "Success_Index" "Group"
##   Success_Index Group
## 1            65     1
## 2            66     1
## 3            58     1
## 4            70     1
## 5            78     1
## 6            53     1
```

R Code (Descriptive Statistics)

```r
# loading library mosaic;

library(mosaic);

favstats(Success_Index~Group);
```

Note. Group 1 = "interaction with outsiders" and Group 2 = "fewer interaction".

R Code (Descriptive Statistics)

```
##   .group min    Q1 median    Q3 max     mean       sd  n
## 1      1  53 62.25   65.5 69.25  78 65.33333 6.610368 12
## 2      2  34 42.50   50.0 54.50  68 49.46667 9.334014 15
##   missing
## 1       0
## 2       0
```

Note. Group 1 = "interaction with outsiders" and Group 2 = "fewer interactions".

Assume $\sigma_1 = \sigma_2$ ←

Example (Slide 13)

① high volume	② low volume
$\bar{x}_1 = 65.333$	$\bar{x}_2 = 49.467$
$s_1 = 6.610$	$s_2 = 9.334$
$n_1 = 12$	$n_2 = 15$

95% CI (Slide 38) where $\sigma_1 = \sigma_2$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2,\ \alpha/2)} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$t_{(n_1+n_2-2,\ \alpha/2)}$: ?

s_p: ?

pooled var

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(12-1)\cdot 6.61^2 + (15-1)\,9.334^2}{12+15-2} = 67.97$$

pooled St. dev

$$s_p = \sqrt{67.97} = 8.243$$

$t_{(n_1+n_2-2,\ \alpha/2)}$ for a 95% CI

$df = n_1 + n_2 - 2$
$= 12 + 15 - 2$
$= 25$

$1 - \alpha = 0.95$
$\alpha = 0.05$
$\alpha/2 = 0.025$

$t_{(n_1+n_2-2,\ \alpha/2)} = t_{(25,\ 0.025)} = 2.060$ (ttable)

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2,\ \alpha/2)} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

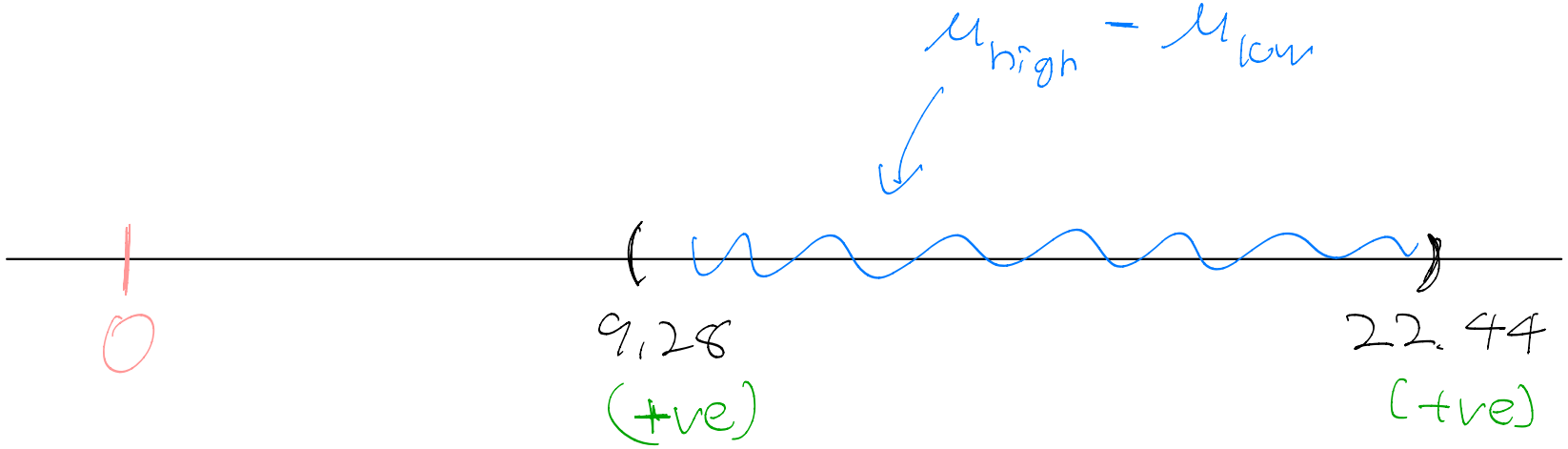
$$= (65.333 - 49.462) \pm (2.060)(8.243)\sqrt{\frac{1}{12} + \frac{1}{15}}$$

$= 15.68 \pm 6.58$

$= (9.28,\ 22.44)$

Interpretation

We are 95% confident the difference in mean index scores between high and low interaction styles is between 9.28 and 22.44



The CI suggests

$$\mu_{high} - \mu_{low} > 0$$

$$\mu_{high} > \mu_{low}$$

Suppose we labelled low interaction as group 1
high " " " 2

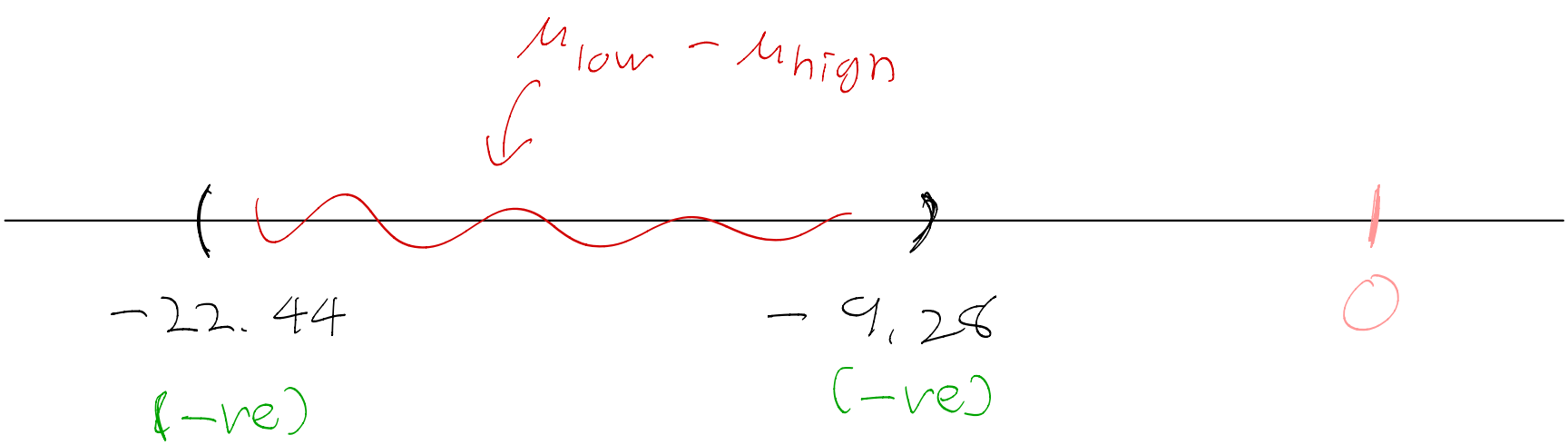
	high volume		low volume
$\bar{x}_2$	$\bar{x}_1 = 65.333$	$\bar{x}_1$	$\bar{x}_2 = 49.467$
s_2	$s_1 = 6.610$	s_1	$s_2 = 9.334$
n_2	$n_1 = 12$	n_1	$n_2 = 15$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(n_1+n_2-2,\ \alpha/2)} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= (49.467 - 65.333) \pm (2.060)(8.243)\sqrt{\frac{1}{12} + \frac{1}{15}}$$

$$= -15.68 \pm 6.58$$

$$= (-22.44, -9.28)$$



CI suggests

$$\mu_{low} - \mu_{high} < 0$$

$$\mu_{low} < \mu_{high}$$

Note:
labelling of groups should not matter, interpretation and results are consistent.

R Code (Descriptive Statistics)

```
# WITHOUT library ;

summary(Success_Index[Group==1]);
length(Success_Index[Group==1]);
sd(Success_Index[Group==1]);

summary(Success_Index[Group==2]);
length(Success_Index[Group==2]);
sd(Success_Index[Group==2]);
```

R Code (Descriptive Statistics)

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   53.00   62.25   65.50   65.33   69.25   78.00
## [1] 12
## [1] 6.610368
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   34.00   42.50   50.00   49.47   54.50   68.00
## [1] 15
## [1] 9.334014
```

Nearly Normal Condition (Group 1: "interaction with outsiders"):

```
stem(Success_Index[Group==1]);
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   5 | 38
##   6 | 0335689
##   7 | 018
```

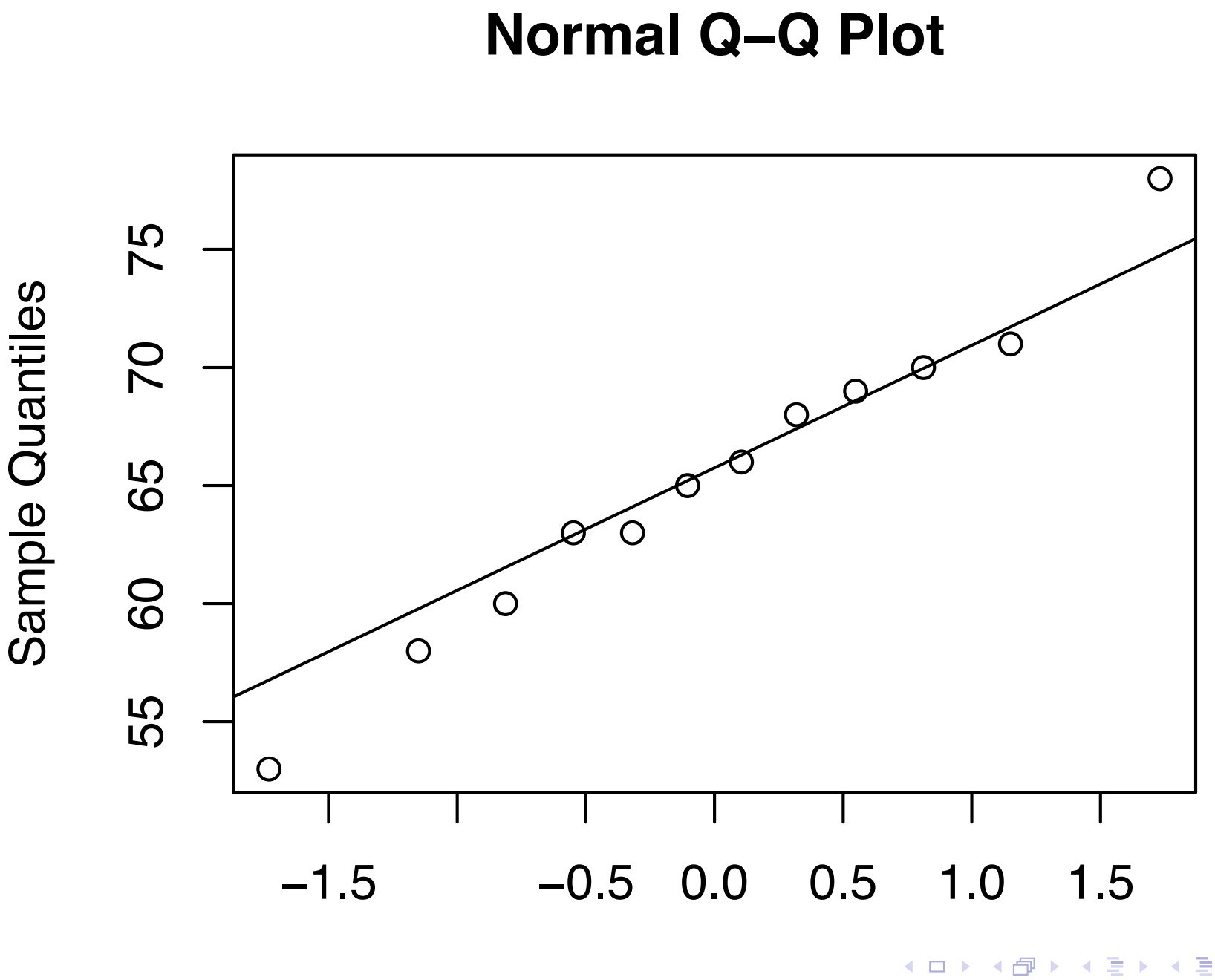
Nearly Normal Condition (Group 2: "fewer interactions"):

```
stem(Success_Index[Group==2]);
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   3 | 46
##   4 | 22368
##   5 | 023367
##   6 | 28
```

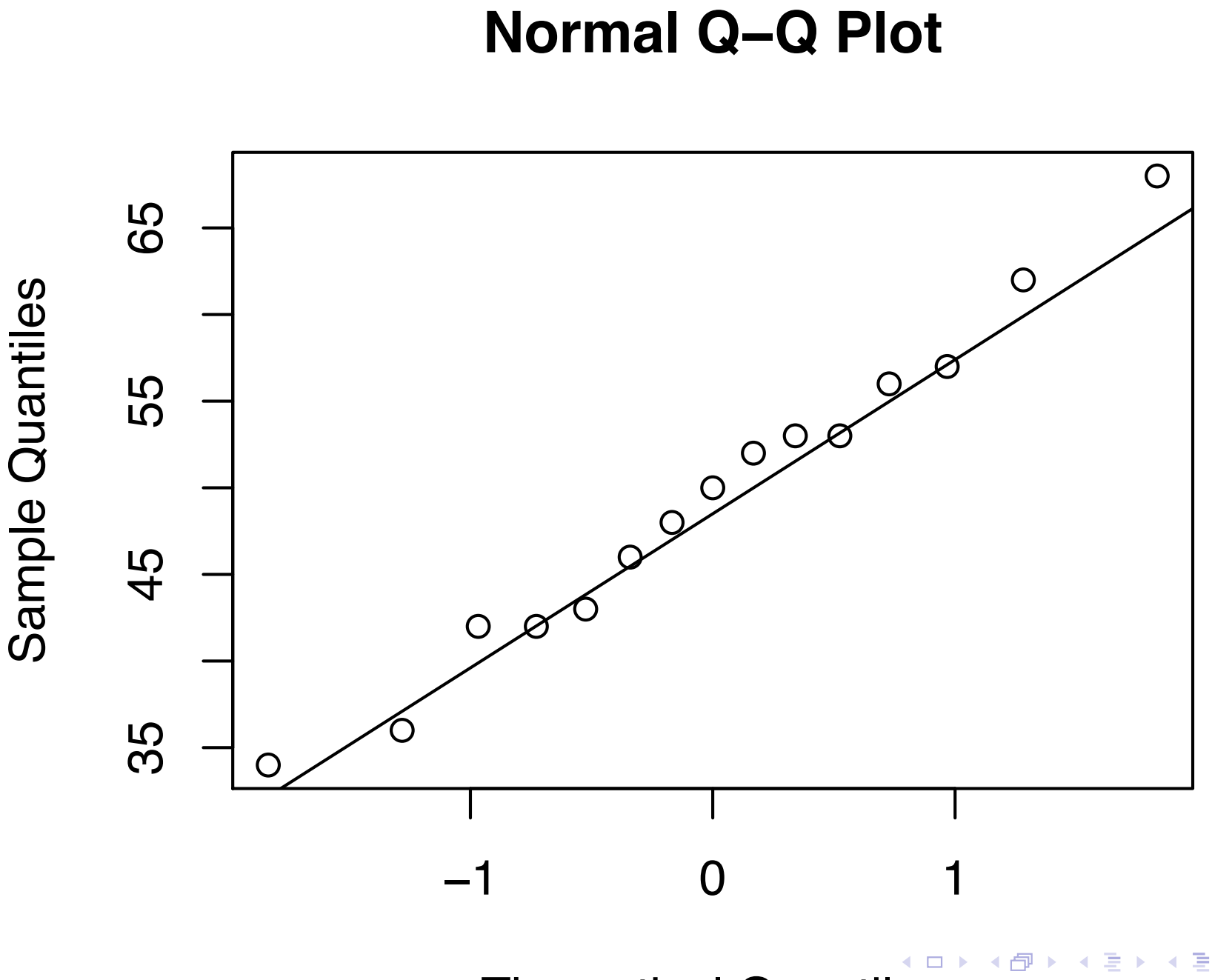
Nearly Normal Condition (Group 1: "interaction with outsiders"):

```
qqnorm(Success_Index[Group==1]);
qqline(Success_Index[Group==1]);
```


Normal Q–Q Plot
Sample Quantiles
55
60
65
70
75
–1.5
–0.5
0.0
0.5
1.0
1.5

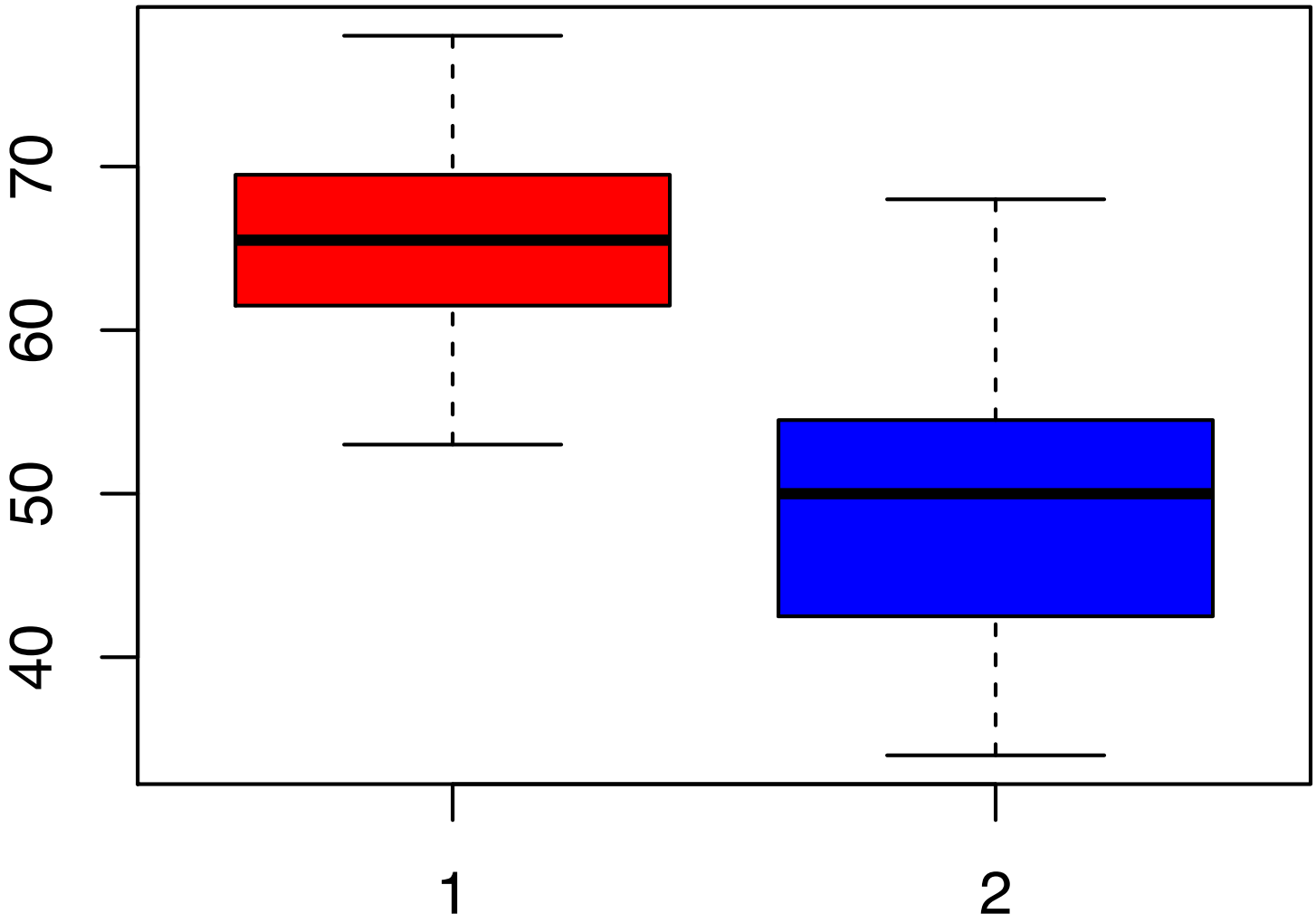
Nearly Normal Condition (Group 2: "fewer interactions"):

```
qqnorm(Success_Index[Group==2]);
qqline(Success_Index[Group==2]);
```


Normal Q–Q Plot
Sample Quantiles
35
45
55
65
−1
0
1

Nearly Normal Condition:

```
boxplot(Success_Index~Group,col=c("red","blue") )
```


70
60
50
40
1
2

Boxplot with ggplot2

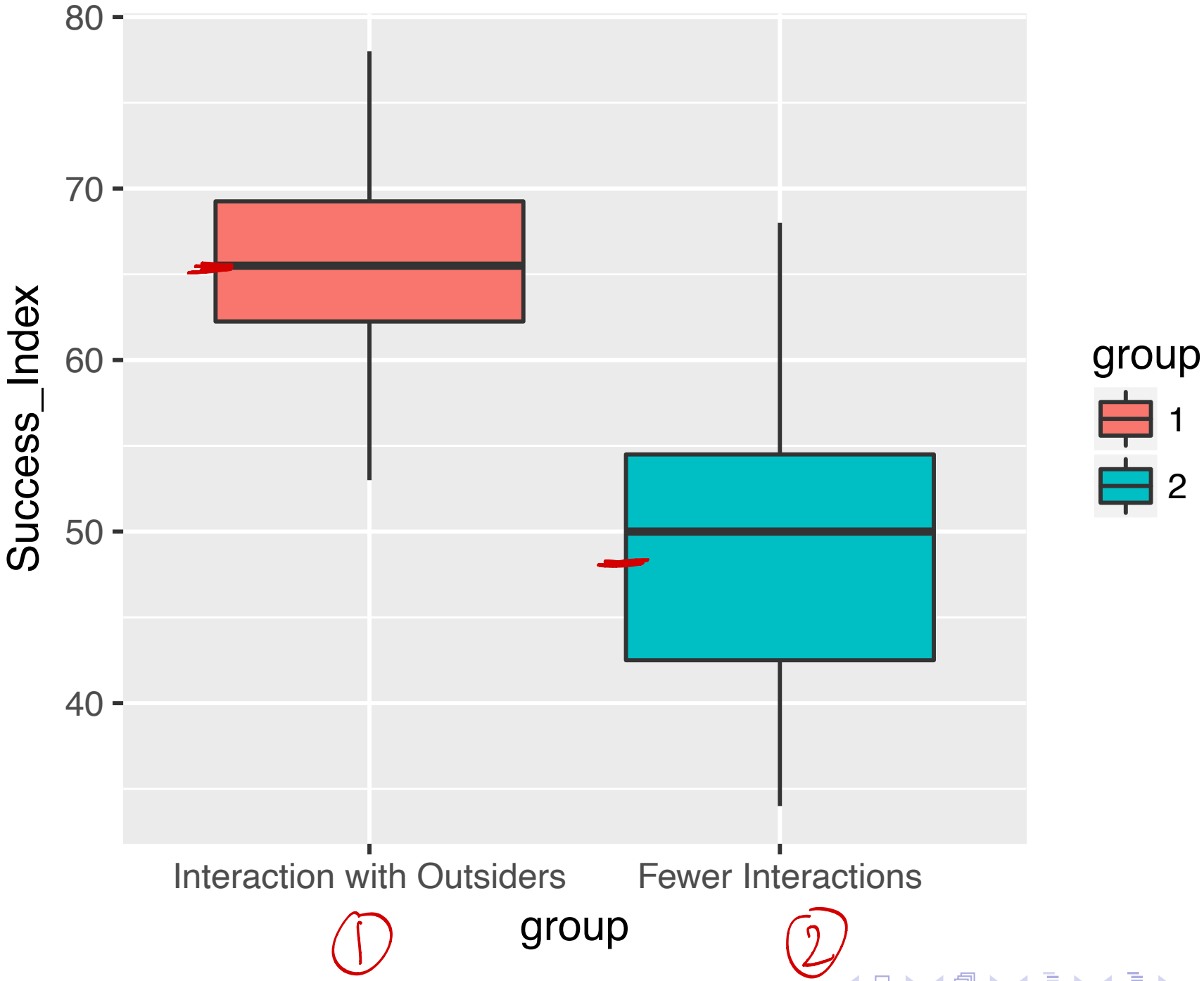
```
# loading library;
library(ggplot2);

# converting a numeric variable into factor (categorical data,
group<-factor(Group);

# bp: just a name (not code) to store boxplots;
bp<-ggplot(success,
aes(x=group,y=Success_Index, fill = group) );

our.labs=c("Interaction with Outsiders","Fewer Interactions");

bp +
geom_boxplot( )+
scale_x_discrete(labels = our.labs);
```


80
70
60
50
40
Success_Index
Interaction with Outsiders
Fewer Interactions
group
1
2
group
1
2

Checking the Assumptions and Conditions

Independent Group Assumption: The success index in group 1 is unrelated to success index in group 2. **Randomization Condition:** The 27 managers were randomly and independently selected (12 for group 1, and 15 for group 2).
Nearly Normal Condition: The two boxplots of success indexes do not show skewness; the two stemplots/histograms of success indexes are unimodal, fairly symmetric and approx. bell-shaped. Q-Q plots also suggest Normality assumption is reasonable.
Equal variances Assumptions: The two boxplots of success indexes appear to have the same spread; thus, the samples appear to have come from populations with approximately same variance.
Since the conditions are satisfied, it is appropriate to construct t CI with $df = 12 + 15 - 2 = 25$.

Example (cont.)

From the data, the following statistics were calculated:

$n_1 = 12$

$n_2 = 15$

$\bar{x}_1 = 65.33$

$\bar{x}_2 = 49.47$

$s_1^2 = 6.61^2$

$s_2^2 = 9.33^2$

Example (cont.)

The pooled variance estimator is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(12 - 1)6.61^2 + (15 - 1)9.33^2}{12 + 15 - 2} = 67.97$$

Example (cont.)

The number of degrees of freedom of the test statistic is

$$\nu = n_1 + n_2 - 2 = 12 + 15 - 2 = 25$$

Example (cont.)

The confidence interval estimator of the difference between two means with equal population variance is

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

or

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

Example (cont.)

The 95% confidence interval estimate of the difference between the return for directly purchased mutual funds and the mean return for broker-purchased mutual funds is

$$(65.33 - 49.47) \pm 2.059539\sqrt{67.97\left(\frac{1}{12} + \frac{1}{15}\right)}.$$

$$15.86 \pm 6.58.$$

The lower and upper limits are 9.28 and 22.44.

95% CI for $\mu_1 - \mu_2$ using R

```
# 95% CI for the difference between means;
# equal variances is assumed;

t.test(Success_Index~Group,
var.equal=TRUE, conf.level=0.95)$conf.int;
```

95% CI for $\mu_1 - \mu_2$ using R

```
## [1]  9.288254 22.445079
## attr(,"conf.level")
## [1] 0.95
```

Intepretation

We are 95% confident that the mean success index is between 9.28 and 22.44 points higher for group 1 than group 2.

Example: Direct and Broker-Purchased Mutual Funds

Millions of investors buy mutual funds, choosing from thousands of possibilities. Some funds can be purchased directly from banks or other financial institutions whereas others must be purchased through brokers, who charge a fee for this service. This raises the question, Can investors do better by buying mutual funds directly than by purchasing mutual funds through brokers? To help answer this question, a group of researchers randomly sampled the annual returns from mutual funds that can be acquired directly and mutual funds that are bought through brokers and recorded the net annual returns, which are the returns on investment after deducting all relevant fees.

Example: Direct and Broker-Purchased Mutual Funds (cont.)

Find a 95% CI for $\mu_1 - \mu_2$, where μ_1 = mean net annual return from directly purchased mutual funds and μ_2 = mean of broker-purchased funds. Assume annual returns are Normally distributed and $\sigma_1^2 = \sigma_2^2$.
From the data, the following statistics were calculated:
$n_1 = 50$
$n_2 = 50$
$\bar{x}_1 = 6.63$
$\bar{x}_2 = 3.72$
$s_1^2 = 37.49$
$s_2^2 = 43.34$

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The pooled variance estimator is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(49)37.49 + (49)43.34}{50 + 50 - 2} = 40.42$$

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The number of degrees of freedom of the test statistic is

$$\nu = n_1 + n_2 - 2 = 50 + 50 - 2 = 98$$

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The confidence interval estimator of the difference between two means with equal population variance is

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

or

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}.$$

pooled

equivalent

Example: Direct and Broker-Purchased Mutual Funds (cont.)

The 95% confidence interval estimate of the difference between the return for directly purchased mutual funds and the mean return for broker-purchased mutual funds is

$$(6.63 - 3.72) \pm 1.984\sqrt{40.42\left(\frac{1}{50} + \frac{1}{50}\right)}.$$

$$2.91 \pm 2.52.$$

The lower and upper limits are 0.39 and 5.43.

2.2 case 2: $\sigma_1 \neq \sigma_2$ unequal St. devs

$(\sigma_1^2 \neq \sigma_2^2)$ (unequal variances)

Welch's method

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(df, \alpha/2)} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

← Standard error

margin of error

Where

$df = \min(n_1 - 1, n_2 - 1)$ ← smaller value between $n_1 - 1$ and $n_2 - 1$

(by hand)

Note: Software (R) uses a more sophisticated formula to estimate df (slide 50)

↳ result in discrepancies between calculation by hand and R.

Example slide 52 → self study

Example: Direct and Broker-Purchased Mutual Funds (cont.)

With 95% confidence, we estimate that the return on directly purchased mutual funds is on average between 0.38 and 5.43 percentage points larger than broker-purchased mutual funds.

Confidence Intervals for μ1−μ2 (with unequal variances)

Draw an SRS of size n_1 from a Normal population with unknown mean μ_1, and draw and independent SRS of size n_2 from another Normal population with unknown mean μ_2. A confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here t^* is the critical value for the $t(k)$ density curve with area C between $-t^*$ and t^*. The degrees of freedom k are equal to the smaller of $n_1 - 1$ and $n_2 - 1$.

Degrees of freedom (Option 1)

Option 1. With software, use the statistic t with accurate critical values from the approximating t distribution.
The distribution of the two-sample t statistic is very close to the t distribution with degrees of freedom df given by

NEVER be asked to calculate by hand

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{s_2^2}{n_2}\right)^2}$$

This approximation is accurate when both sample sizes n_1 and n_2 are 5 or larger.

Degrees of freedom (Option2)

Option 2. Without software, use the statistic t with critical values from the t distribution with degrees of freedom equal to the smaller of $n_1 - 1$ and $n_2 - 1$. These procedures are always conservative for any two Normal populations.

Example

A company that sells educational materials reports statistical studies to convince customers that its materials improve learning. One new product supplies "directed reading activities" for classroom use. These activities should improve the reading ability of elementary school pupils.
A consultant arranges for a third-grade class of 21 students to take part in these activities for an eight-week period. A control classroom of 23 third-graders follows the same curriculum without the activities. At the end of the eight weeks, all students are given a Degree of Reading Power (DRP) test, which measures the aspects of reading ability that the treatment is designed to improve. The data appear in the following table.

Assume $\sigma_1 \neq \sigma_2$

Data

Treatment				Control			
24	61	59	46	42	33	46	37
43	44	52	43	43	41	10	42
58	67	62	57	55	19	17	55
71	49	54		26	54	60	28
43	53	57		62	20	53	48
49	56	33		37	85	42	

Find a 95% confidence interval for the mean improvement in the entire population of third-graders.

```r
# Step 1. Entering data;

treatment=c(24, 61, 59, 46, 43, 44, 52, 43, 58, 67,
62, 57, 71, 49, 54, 43, 53, 57, 49, 56, 33);

control=c(42, 33, 46, 37, 43, 41, 10, 42, 55, 19, 17,
55, 26, 54, 60, 28, 62, 20, 53, 48, 37, 85, 42);
```

Checking the assumptions

Nearly Normal Condition (treatment):

```
# Making stemplot;

stem(treatment);
```

Checking the assumptions

Nearly Normal Condition (treatment):

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   2 | 4
##   3 | 3
##   4 | 3334699
##   5 | 23467789
##   6 | 127
##   7 | 1
```

Checking the assumptions

Nearly Normal Condition (control):

```
# Making stemplot;

stem(control);
```

Checking the assumptions

Nearly Normal Condition (control):

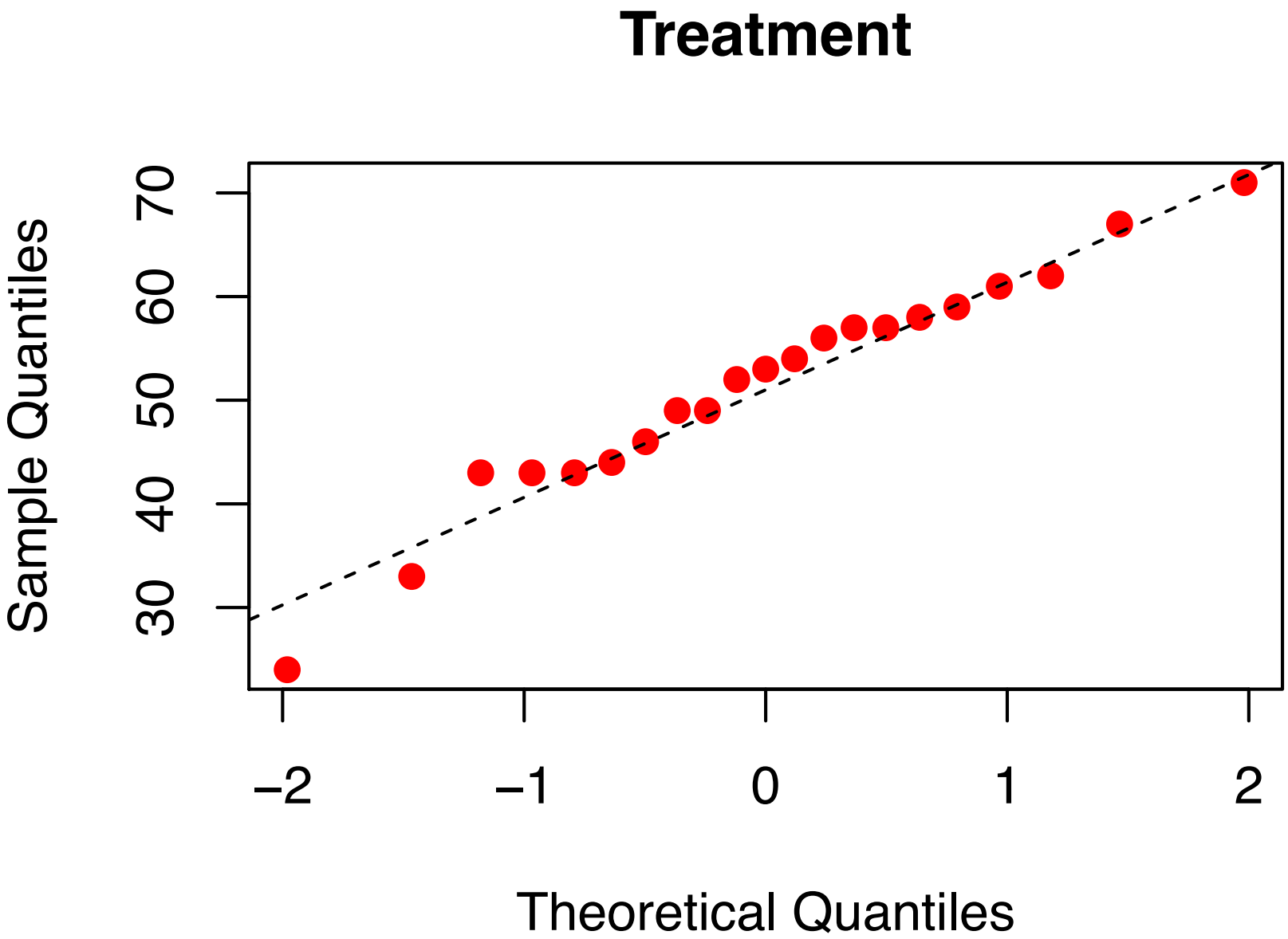
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 079
##   2 | 068377
##   4 | 12223683455
##   6 | 02
##   8 | 5
```

Checking the assumptions

Nearly Normal Condition (treatment):

```
# Making Q-Q plot;
qqnorm(treatment,pch=19,col="red",main="Treatment");
qqline(treatment,lty=2);
```

Nearly Normal Condition (treatment):

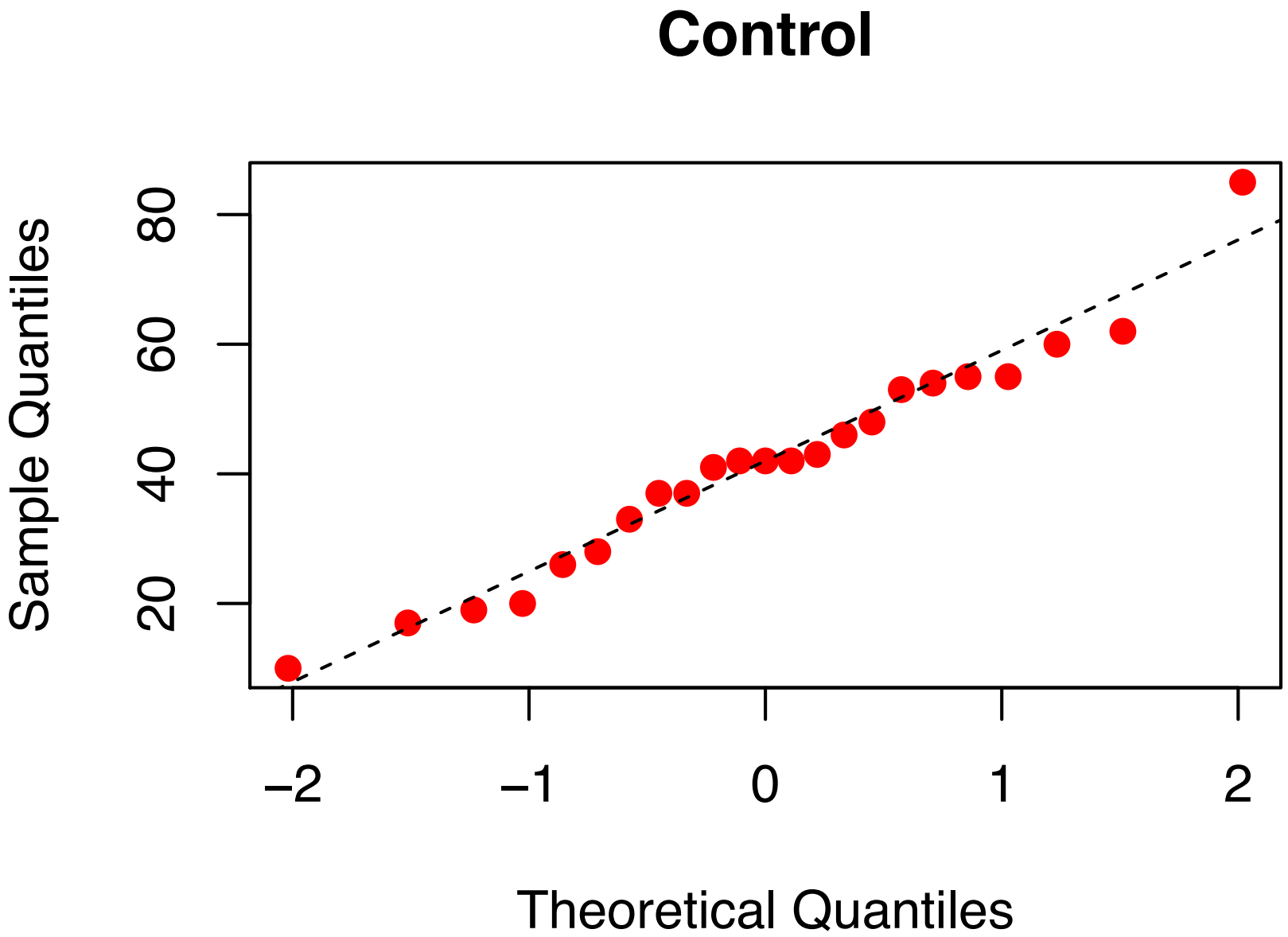


Checking the assumptions

Nearly Normal Condition (control):

```
# Making Q-Q plot;
qqnorm(control,pch=19,col="red",main="Control");
qqline(control,lty=2);
```

Nearly Normal Condition (control):



Stemplots suggest that there is a mild outlier in the control group but no deviation from Normality serious enough to prevent us from using t procedures. Normal Q-Q plots for both groups confirm that both are roughly Normal. The summary statistics are:

```
summary(treatment);

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   24.00   44.00   53.00   51.48   58.00   71.00

summary(control);

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   30.50   42.00   41.52   53.50   85.00
```

```r
# Step 1. Entering data;

treatment=c(24, 61, 59, 46, 43, 44, 52, 43, 58, 67,
62, 57, 71, 49, 54, 43, 53, 57, 49, 56, 33);

control=c(42, 33, 46, 37, 43, 41, 10, 42, 55, 19, 17,
55, 26, 54, 60, 28, 62, 20, 53, 48, 37, 85, 42);

# Step 2. Confidence Interval;

t.test(treatment,control,conf.level=0.95);
```

95% CI (using R)

```
##
##  Welch Two Sample t-test
##
## data:  x and control
## t = 2.3109, df = 37.855, p-value = 0.02638
## alternative hypothesis: true difference in means is not equ
## 95 percent confidence interval:
##   1.23302 18.67588
## sample estimates:
## mean of x mean of y
##  51.47619  41.52174
```

$\sigma_1 = \sigma_2$

95% CI (using table)

Summary statistics

```
round(mean(treatment),2);

## [1] 51.48

round(sd(treatment),2);

## [1] 11.01

round(mean(control),2);

## [1] 41.52

round(sd(control),2);

## [1] 17.15
```

95% CI (using Table)

The conservative approach uses the $t(20)$ distribution. Table 5 gives $t^* = 2.086$. With this approximation we have

$$(\bar{x}_T - \bar{x}_C) \pm t^* \sqrt{\frac{s_T^2}{n_T} + \frac{s_C^2}{n_C}}$$

$$(51.48 - 41.52) \pm (2.086 \times 4.31)$$

$$(0.97, 18.95)$$

Interpretation

We estimate, with 95% confidence, that the mean improvement in DRP scores is between 1 and 19 points. Although we have good evidence of some improvement, the data do not allow a very accurate estimate of the size of the average improvement.

Additional info about the DRP study

The design of the DRP study is not ideal. Random assignment of students was not possible in a school environment, so existing third-grade classes were used. The effect of the reading programs is therefore confounded with any other differences between the two classes. The classes were chosen to be as similar as possible in variables such as the social and economic status of the students. Pretesting showed that the two classes were on the average quite similar in reading ability at the beginning of the experiment. To avoid the effect of two different teachers, the same teacher taught reading in both classes during the eight-week period of the experiment. We can therefore be somewhat confident that our two-sample procedure is detecting the effect of the treatment and not some other difference between the classes.

Logging in the rain forest

"Conservationists have despaired over destruction of tropical rain forest by logging, clearing, and burning". These words begin a report on a statistical study of the effects of logging in Borneo. Here are data on the number of tree species in 12 unlogged forest plots and 9 similar plots logged 8 years earlier:
Unlogged: 22 18 22 20 15 21 13 13 19 13 19 15
Logged : 17 4 18 14 18 15 15 10 12
Use the data to give a 99% confidence interval for the difference in mean number of species between unlogged and logged plots.
Assume that the random variable number of species is Normally distributed.

Assume $\sigma_1 \neq \sigma_2$

Example (slide 70)

① unlogged	② logged
$n_1 = 12$	$n_2 = 9$
$\bar{x}_1 = 17.5$	$\bar{x}_2 = 13.667$
$s_1 = 3.529$	$s_2 = 4.50$

use software or calculator to obtain from data

99% CI for $\mu_1 - \mu_2$ (assuming $\sigma_1 \neq \sigma_2$)

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(df, \alpha/2)} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$t_{(df, \alpha/2)}$?

$t_{(df, \alpha/2)}$ for a 99% CI

(by hand)

$df = \min(n_1 - 1, n_2 - 1)$

$= \min(12 - 1, 9 - 1)$

$= \min(11, 8)$

$= 8$

$1 - \alpha = 0.99$

$\alpha = 0.01$

$\alpha/2 = 0.005$

$t_{(df, \alpha/2)} = t_{(8, 0.005)} = 3.355$ (t table)

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(df,\alpha/2)} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

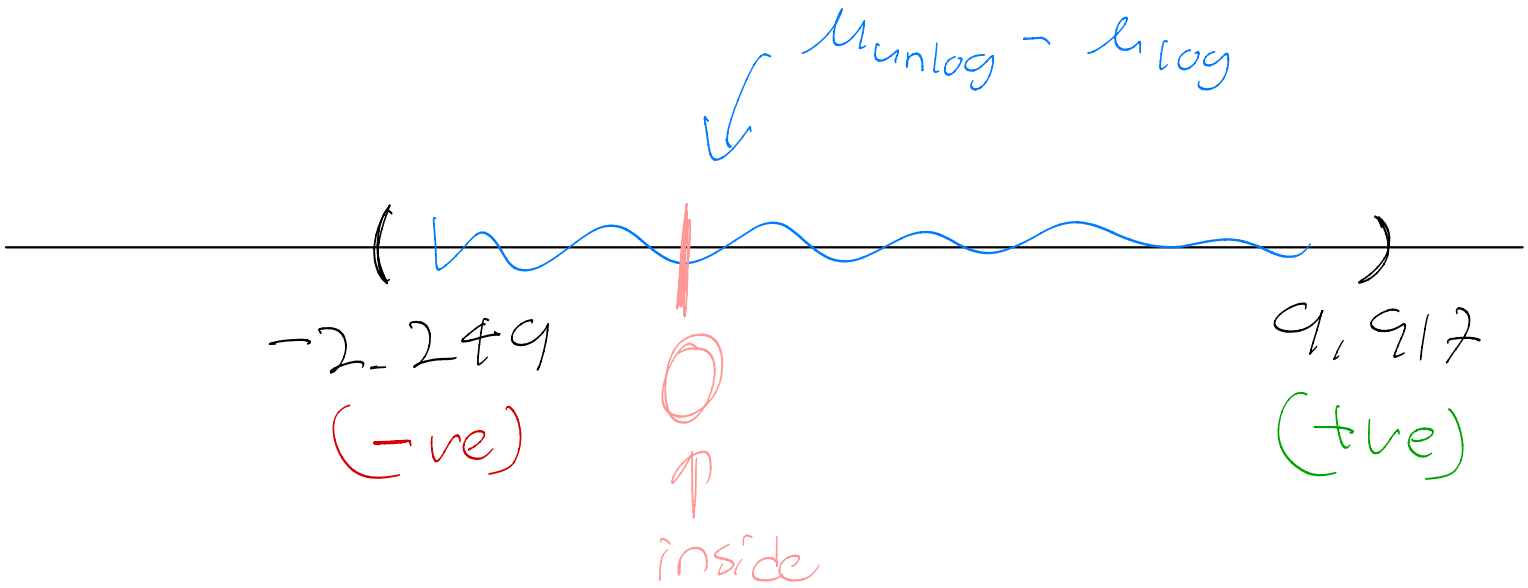
$$= (17.5 - 13.667) \pm 3.355 \sqrt{\frac{3.529^2}{12} + \frac{4.5^2}{9}}$$

$$= 3.834 \pm 6.0834$$

$$= (-2.249, 9.917)$$

Interpretation

We are 99% confident the difference in the mean number of tree species between unlogged and logged plots is between −2.249 and 9.917.



plausible that $\mu_{unlog} - \mu_{log} = 0$

$\mu_{unlog} = \mu_{log}$

Conditions for all 2 sample CIs on $\mu_1 - \mu_2$

1/ Independence (between & within samples)

2/ Random samples

3/ If Both samples sizes are small ($n_1 < 30$ AND $n_2 < 30$) populations should be normal

4/ If one sample is small ($n_1 < 30$ XOR $n_2 < 30$) the population this sample was taken from should be normal

Note: If samples are large ($n_1 \geq 30$ AND $n_2 \geq 30$) normality assump of populations not required (CLT)

Solution

1. Find $\bar{x}_1 - \bar{x}_2$; $\bar{x}_1 - \bar{x}_2 = 17.5 - 13.6666 = 3.8334$
2. Find SE = Standard Error. We have that: $s_1 = 3.5290$, $s_2 = 4.5$, $n_1 = 12$ and $n_2 = 9$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{3.5290^2}{12} + \frac{4.5^2}{9}} = 1.8132$$

3. Find $m = t^* SE$. From Table, we have $df = 8$ and 99% confidence level, then $t^* = 3.355$. Hence, $m = (3.355)(1.8132) = 6.0832$.
4. Find Confidence Interval.

$\bar{x}_1 - \bar{x}_2 \pm t^* SE = 3.8334 \pm 6.0832 = -2.2498$ to 9.9166.

R gives a 99% confidence interval of -1.520400 to 9.187067 species, using $df = 14.793$.

```r
# Step 1. Entering data;

unlogged=c(22,18,22,20,15,21,13,13,19,13,19,15);

logged=c(17,4,18,14,18,15,15,10,12);

# Step 2. Confidence Interval;

t.test(unlogged,logged,conf.level=0.99);
```

```
##
##  Welch Two Sample t-test
##
## data:  x and logged
## t = 2.1141, df = 14.793, p-value = 0.05192
## alternative hypothesis: true difference in means is not equ
## 99 percent confidence interval:
##  -1.520400  9.187067
## sample estimates:
## mean of x mean of y
##  17.50000  13.66667
```

Comparing Means with Paired Samples

Comparing Means with Paired Samples

- When observations in sample 1 matches with an observation in sample 2.
- Observations in sample 1 are, usually, highly, correlated with observations in sample 2.
- The data are often called matched pairs.
- For each pair (the same cases), we form: Difference = observation in sample 2 - observation in sample 1.
- Thus, we have one single sample of differences scores.
- For example, in longitudinal studies: Pre- and post-survey of attitudes towards statistics (Same student is measured twice: Time 1 (pre) and Time 2 (post). We measure change in the attitudes: Post - Pre (for each student).
- Often these types of studies are called, *repeated measures* .

Paired Samples: Assumptions and Conditions

Paired Data Condition:

- The data must be quantitative and paired.

Independence Assumption:

- If the data are paired, the groups are not independent. For this methods, it is the differences that must be independent of each other.
- The pairs may be a random sample.
- In experimental design, the order of the two treatments may be randomly assigned, or the treatments may be randomly assigned to one member of each pair.
- In a before-and-after study, we may believe that the observed differences are representative sample of a population of interest. If there is any doubt, we need to include a control group to be able to draw conclusions.

Review

2 samples (Diff in means)

1/ σ_1, σ_2 both known

2/ " " " unknown

2.1: $\sigma_1 = \sigma_2$ (pooled)

2.2: $\sigma_1 \neq \sigma_2$ (welch's)

CI's on Paired Data

2 measurements on <u>each</u> unit.

(ie measurement 1 and 2 are both dependent on each unit)

both Depend on unit

Sample units	Measurement 1 (M_1)	Measurement 2 (M_2)	Difference $M_2 - M_1$ (or $M_1 - M_2$)
1	X_{11}	X_{12}	$X_{d1} = X_{12} - X_{11}$ (or $X_{11} - X_{12}$)
2	X_{21}	X_{22}	$X_{d2} = X_{22} - X_{21}$
⋮	⋮	⋮	⋮
n	X_{n1}	X_{n2}	$X_{dn} = X_{n2} - X_{n1}$
		mean	$\bar{X}_d$
		Sh. dev	S_d

CI on paired differences

$$\bar{x}_d \pm t_{(n-1,\,\alpha/2)} \cdot \frac{s_d}{\sqrt{n}}$$

(similar to one sample CI when σ unknown)

Paired Samples: Assumptions and Conditions

Independence Assumption (cont.):

- If samples are bigger than 10% of the target population, we need to acknowledge this and note in our report. When we sample from a finite population, we should be careful not to sample more than 10% of that population. Sampling too large a fraction of the population calls the independence assumption into question.

Paired Samples: Assumptions and Conditions

Normal Population Assumption
We assume that the population of differences follows a Normal model. We need to check:
Nearly Normal Condition:

- This condition can be checked with a histogram or Boxplot of differences - but not of the individual groups.
- As with the case of the one-sample t-methods, robustness against departure form normality increases with sample size; in other words, the Normality assumptions matter less the more pairs we have to consider.

Note: When the conditions are met, we can model the sampling distribution of difference in sample means with a Student's t-model with n- 1 degrees of freedom.

Confidence Interval: Paired t-Interval

When the assumptions and conditions are met, the confidence interval for the mean of paired difference $\mu_1 - \mu_2$ is:
Point Estimate $\pm$ Margin of Error of the Point Estimate

$$\bar{X}_d \pm t^*_{df} SE(\bar{X}_d)$$

Where the standard error of the mean difference is estimated as

$$SE(\bar{X}_d) = \frac{s_d}{\sqrt{n}}$$

The critical value t^*_{df} depends on the particular confidence level and the number of df = n - 1, which is based on the number of pairs, n.

Example

Comparing 2016 and 2017 Voter Turnout % in OCED Countries
The "Better Life Index" program (BLI, 2017 and 2016) includes set of indicators regarding social protection and well-being in OECD countries. A quantitative variable named "Voter Turnout" is a sub-component of the component in this BLI program, which is defined as the ratio between the number of individuals that cast a ballot during an election (whether this vote is valid or not) to the population registered to vote. Thus, the unit of measurement is population percentage of voter turnout in OECD countries. As institutional features of voting systems vary a lot across countries and across types of elections, the indicator refers to the elections (parliamentary or presidential) that have attracted the largest number of voters in each country. OECD indicates that they obtained this information from International Institute for Democracy and Electoral Assistance (IDEA); Comparative Studies of Electoral System for inequality data (self-reported voter turnout).

R Code

```r
# Importing data file into R;

voterT=read.csv(file="VoterTurnout.csv",header=TRUE);

# Getting names of variables;

names(voterT);

# Seeing first few observations;

head(voterT);

# Attaching data file;
attach(voterT);
```

R Code

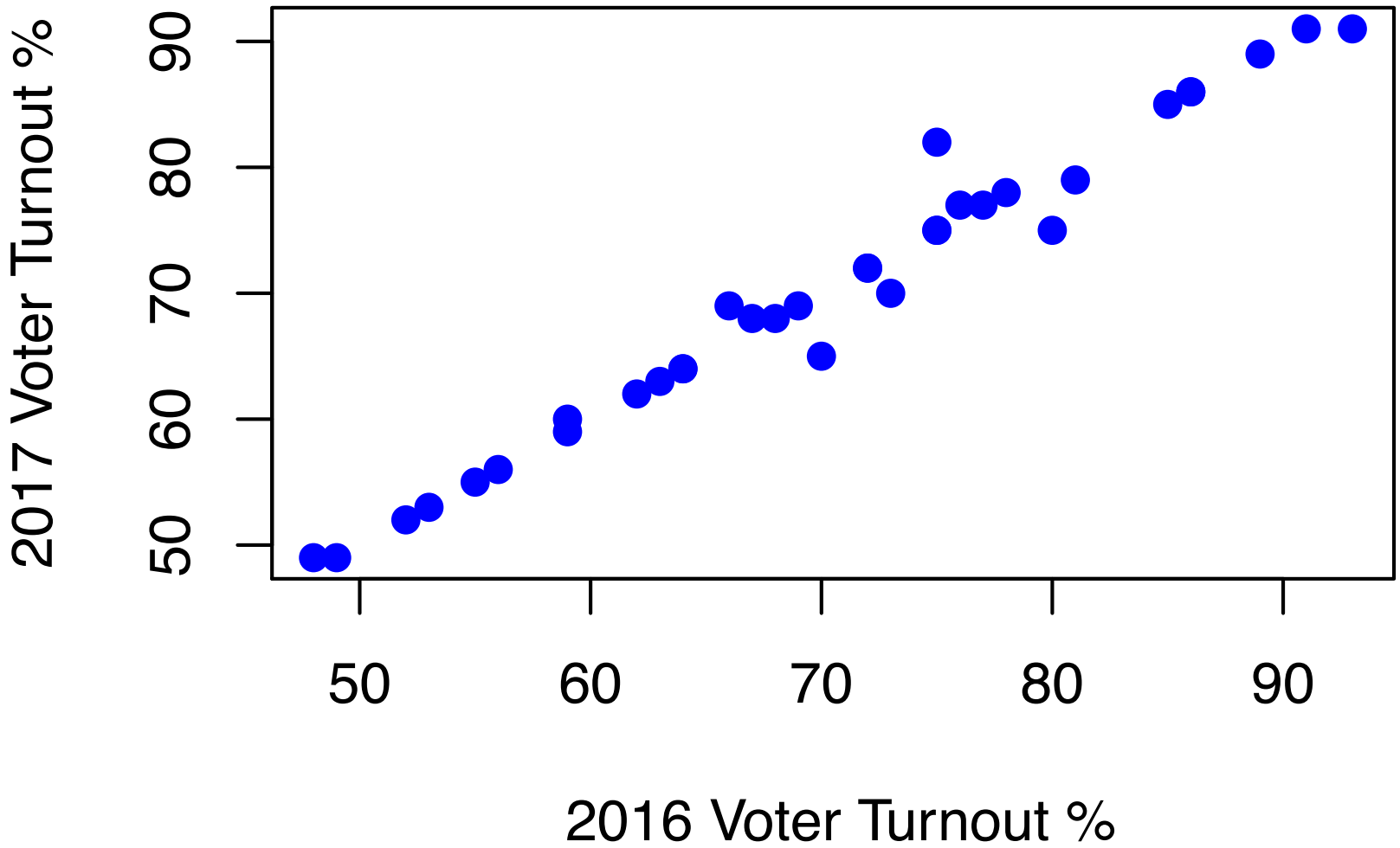
```
## [1] "Country"                "Voter_turnout_2017"
## [3] "Voter_turnout_2016"
##          Country Voter_turnout_2017 Voter_turnout_2016
## 1      Australia                 91                  93
## 2        Austria                 75                  75
## 3        Belgium                 89                  89
## 4         Canada                 68                  68
## 5          Chile                 49                  49
## 6 Czech Republic                 59                  59
```

Association 2016 and 2017 Voter Turnout % in OECD Countries

```
# Scatterplot of data;
plot(Voter_turnout_2016, Voter_turnout_2017,
xlab =" 2016 Voter Turnout %",
ylab="2017 Voter Turnout %", pch=19, col="blue");

# Sample Correlation, r;
cor(Voter_turnout_2016, Voter_turnout_2017);
```

Association 2016 and 2017 Voter Turnout % in OECD Countries



Association 2016 and 2017 Voter Turnout % in OECD Countries (Sample Correlation)

```
## [1] 0.9868955
```

Plot Interpretation: As percentage of voter turnout in 2016 increases, percentage of voter turnout in 2017 tend to increase.
Estimated Correlation Interpretation: If an OECD country's 2016 voter turnout is 1 standard deviation above the mean, its 2017 voter turnout percentage is approx. 0.987 standard deviation above the mean 2017 voter turnout %.

Working with Summary Statistics

Let μ_1 denote the population mean voter turnout percentages in 2017. Let $\bar{X}_1$ denote the sample mean voter turnout percentages in 2017, that estimates μ_1.
Let μ_2 denote the population mean voter turnout percentages in year 2016. Let $\bar{X}_2$ denote the sample mean voter turnout percentages in 2016, that estimates μ_2.
Let μ_d denote the population mean difference in voter turnout percentages ($\mu_d = \mu_1 - \mu_2$). Let $\bar{X}_d$ denote the sample mean of the difference in voter turnout percentages, that estimates μ_d.
Note: The mean of the estimated differences $\bar{X}_d$ equals the differences between the estimated means $\bar{X}_1 - \bar{X}_2$.

Summary statistics using R

```
# Obtaining summary statistics;

# loading library;
library(mosaic);
favstats(Voter_turnout_2017);

favstats(Voter_turnout_2016);

Diff_Voter_turnout = Voter_turnout_2017 - Voter_turnout_2016;
favstats(Diff_Voter_turnout);
```

Summary statistics using R

```
##  min    Q1 median    Q3 max     mean       sd  n missing
##   49 62.25   69.5 77.75  91 70.14706 12.06832 34       0
##  min    Q1 median    Q3 max     mean       sd  n missing
##   48 62.25     71 77.75  93 70.23529 12.24017 34       0
##  min Q1 median Q3 max        mean       sd  n missing
##   -5  0      0  0   7 -0.08823529 1.975113 34       0
```

Summary statistics using R

```r
# Obtaining summary statistics;

# WITHOUT loading library;

Diff_Voter_turnout = Voter_turnout_2017 - Voter_turnout_2016;
mean(Diff_Voter_turnout);
sd(Diff_Voter_turnout);
```

Summary statistics using R

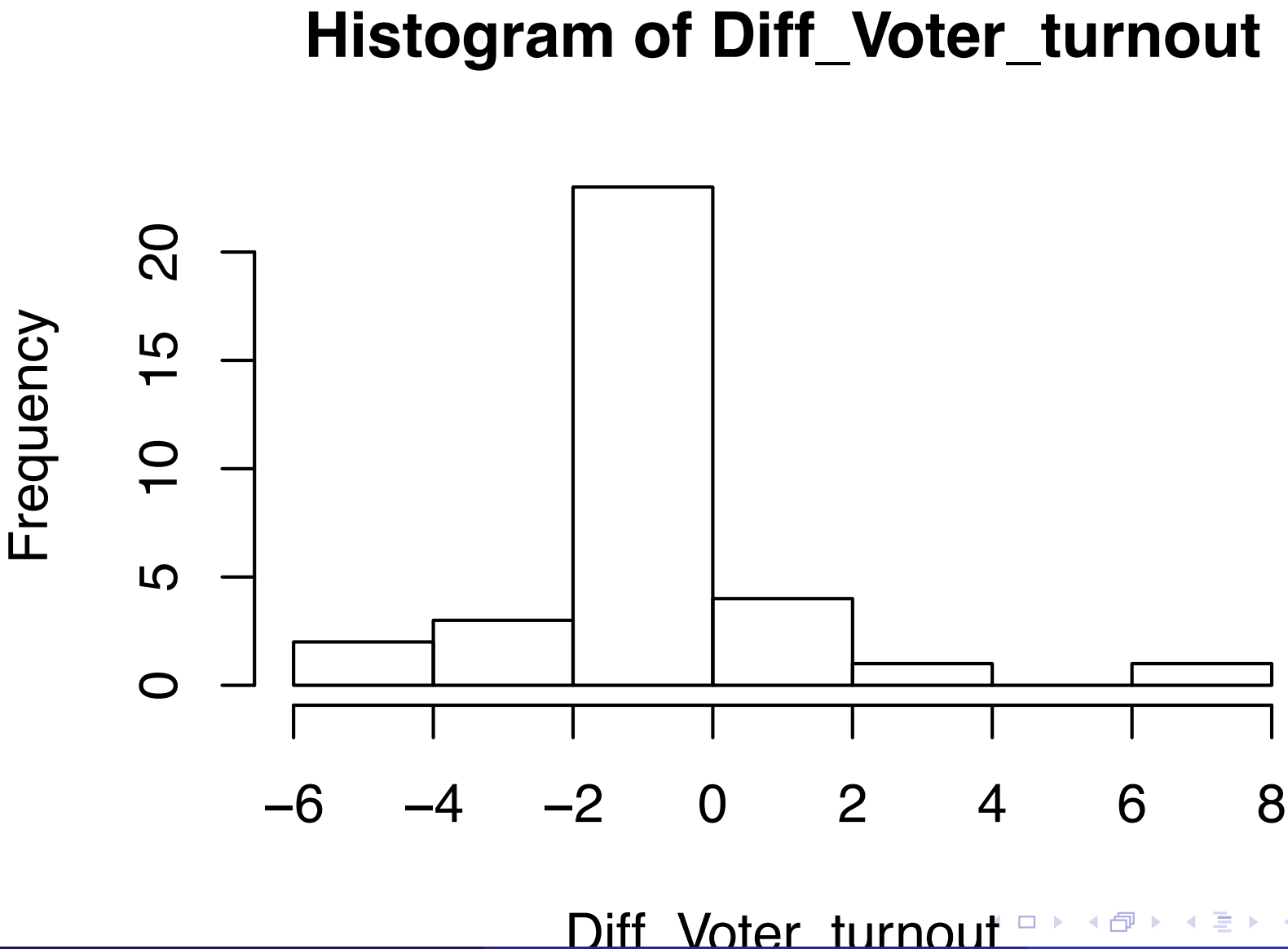
```
## [1] -0.08823529
## [1] 1.975113
```

Checking Assumptions and Conditions

```
Diff_Voter_turnout = Voter_turnout_2017 - Voter_turnout_2016;

# Making histogram;
hist(Diff_Voter_turnout);
```

Checking Assumptions and Conditions



Checking Assumptions and Conditions

Paired Data Condition:
The data are paired because we are interested in difference voter turnout percentages in 2016 and 2017 within each OECD country.
Independence/Randomization:
Each pair (2016 voter turnout %, 2017 voter turnout %) is independent of the others, so the differences are independent.
Nearly Normal Condition:
The histogram of the differences is approx. bell-shaped and symmetric.
The conditions are met; so, we can use a Student's t-model with n-1 degrees of freedom to construct confidence intervals.

95% Confidence Interval (Paired t- Interval) in R

```
t.test(Voter_turnout_2017, Voter_turnout_2016, paired=TRUE);
```

95% Confidence Interval (Paired t- Interval) in R

```
##
##  Paired t-test
##
## data:  x and Voter_turnout_2016
## t = -0.26049, df = 33, p-value = 0.7961
## alternative hypothesis: true difference in means is not equ
## 95 percent confidence interval:
##  -0.7773846  0.6009140
## sample estimates:
## mean of the differences
##              -0.08823529
```

Our Example: 95% CI (paired t-interval)

95% CI for μ_d: $\bar{X}_d \pm t^*_{df} SE(\bar{X}_d)$.
$\bar{X}_d = -0.088$
$SE(\bar{X}_d) = \frac{s_d}{\sqrt{n}} = \frac{1.975}{\sqrt{34}} = 0.3387$
Confidence Level: 0.95; $\alpha = 1 - 0.95 = 0.05$; $\alpha/2 = 0.025$
$df = n - 1 = 34 - 1 = 33$, $t^*_{df} = t^*_{0.025;33} \approx 2.0345$.

```
qt(0.975, 33);

## [1] 2.034515
```

Finally,

$$-0.088 \pm 2.034515(0.3387) = -0.088 \pm 0.689$$

$$(-0.777, 0.6009)$$

Interpretation

The value "0" is in this CI. We have **no** evidence to indicate that there is a difference in 2017 and 2016' mean voter turnout percentages in OECD countries.

Example

A manufacturer wanted to compare the wearing qualities of two different types of automobile tires, A and B. In the comparison, a tire of type A and one of type B were randomly assigned and mounted on the rear wheels of each of five automobiles. The automobiles were then operated for a specified number of miles, and the amount of wear was recorded for each tire. These measurements appear in a table below. Do the data provide sufficient evidence to indicate a difference in mean wear for tire types A and B?

Auto	1	2	3	4	5
Tire A	10.6	9.8	12.3	9.7	8.8
Tire B	10.2	9.4	11.8	9.1	8.3

Example

Find a 95% confidence interval for $(\mu_A - \mu_B) = \mu_d$ using the data given above. Assume differences are Normally distributed.

vehicle	Tire A	Tire B	Differences (A − B)
1	10.6	10.2	$x_{1d} = 10.6 - 10.2 = 0.4$
2	9.8	9.4	$x_{2d} = 9.8 - 9.4 = 0.4$
3	12.3	11.8	$x_{3d} = $ $= 0.5$
4	9.7	9.1	$x_{4d} = $ $= 0.6$
5	8.6	8.3	$x_{5d} = $ $= 0.5$

$n = 5$

mean $\bar{x}_d = 0.46$

Sh. dev $s_d = 0.0837$

95% CI on paired differences

$$\bar{x}_d \pm t_{(n-1,\alpha/2)} \cdot \frac{s_d}{\sqrt{n}}$$

$t_{(n-1,\alpha/2)}$

$df = n - 1$

$= 4$

$1 - \alpha = 0.95$

$\alpha = 0.05$

$\alpha/2 = 0.025$

$t_{(n-1,\alpha/2)} = t_{(4,0.025)} = 2.776$

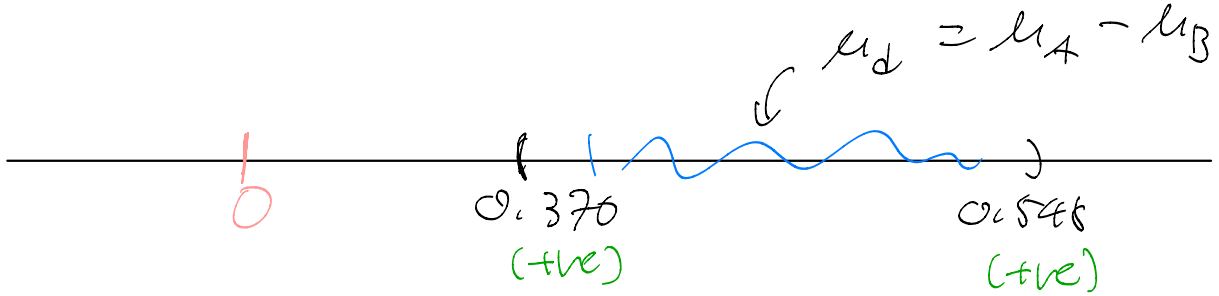
$$\bar{x}_d \pm t_{(n-1,\alpha/2)} \cdot \frac{s_d}{\sqrt{n}} = 0.46 \pm (2.776)\left(\frac{0.0837}{\sqrt{5}}\right)$$

$$= 0.46 \pm 0.1034$$

$$= (0.376, 0.548)$$

Interpretation

We are 95% confident the mean difference is between 0.376 (units) and 0.548 (units)



CI suggests $\mu_d > 0$

$\mu_A - \mu_B > 0$

$\mu_A > \mu_B$

average wear of tire A is greater than average for B.

Look out for 2 measurements which depend on each unit!

Solution

You can verify that the mean and standard deviation of the five **difference** measurements are $\bar{d} = 0.48$ and $s_d = 0.0837$.
A 95% confidence interval for the difference between the mean wear is

$$\bar{d} \pm t^* \frac{s_d}{\sqrt{n}}$$

$$0.48 \pm (2.776)\frac{0.0837}{\sqrt{5}}$$

$$0.48 \pm 0.1039$$

95% Confidence Interval (Paired t- Interval) in R

```
A=c(10.6 , 9.8 ,12.3 , 9.7 , 8.8);
B=c(10.2 , 9.4 , 11.8 , 9.1, 8.3);

t.test(A, B, paired=TRUE);
```

95% Confidence Interval (Paired t- Interval) in R

```
##
##  Paired t-test
##
## data:  x and B
## t = 12.829, df = 4, p-value = 0.0002128
## alternative hypothesis: true difference in means is not equ
## 95 percent confidence interval:
##  0.3761149 0.5838851
## sample estimates:
## mean of the differences
##                    0.48
```

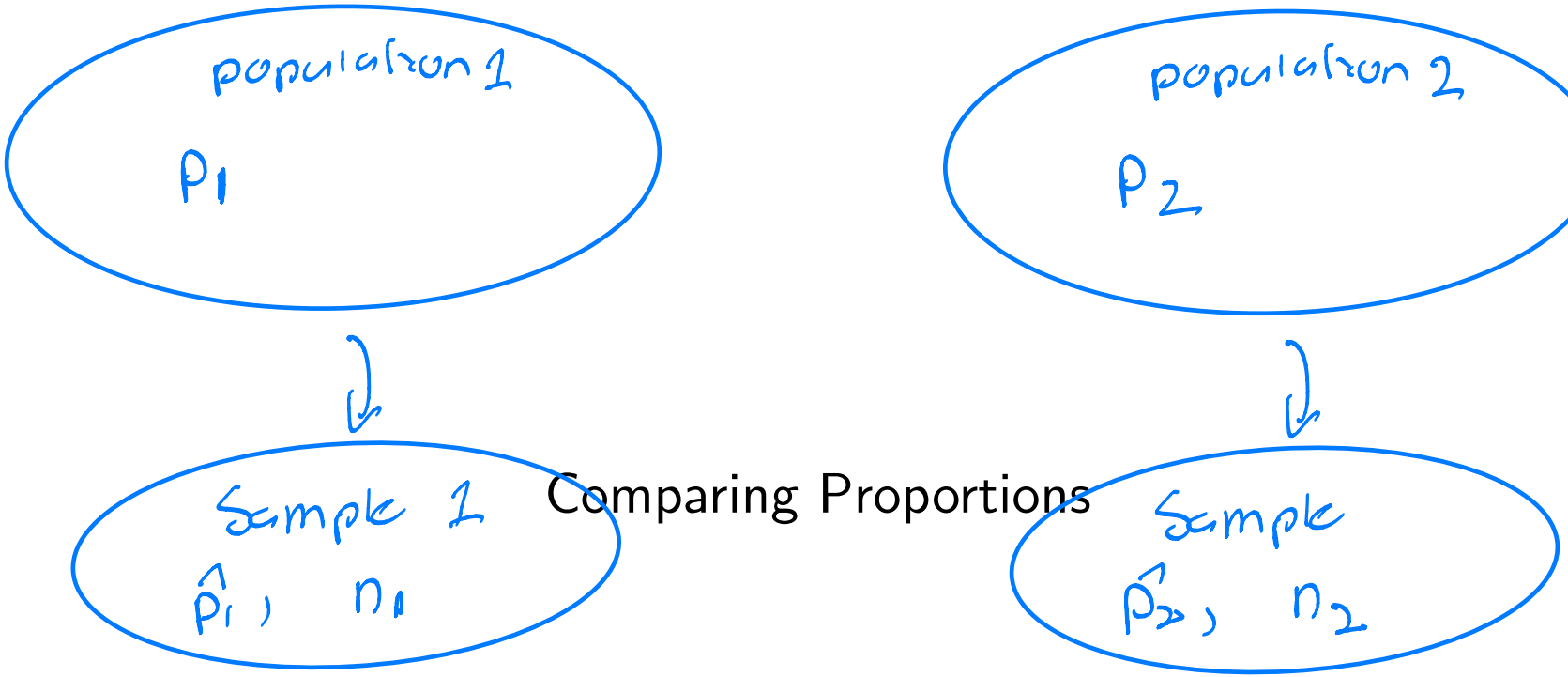
Assumptions

1/ units are independent (the measurements are not, they depend on each unit)

2/ units are randomly sampled.

3/ If we have a small sample ($n < 30$), then population of differences should be normal

(slides 77 - 78)



Comparing Proportions

Interested in: $p_1 - p_2$ (or $p_2 - p_1$)

CI on a difference of proportions

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

standard error

margin of error

Note:

$\hat{p}_1$ and $\hat{p}_2$ may be provided directly or indirectly

$$\hat{p}_1 = \frac{\#\text{ successes in sample 1}}{\text{size of sample 1}} = \frac{x_1}{n_1}$$

$$\hat{p}_2 = \frac{\#\text{ successes in sample}}{\text{size of sample}} = \frac{x_2}{n_2}$$

Large-sample confidence interval for comparing two proportions

Draw an SRS of size n_1 from a population having proportion p_1 of successes and draw an independent SRS of size n_2 from another population having proportion p_2 of successes. When n_1 and n_2 are large, an approximate level C confidence interval for $p_1 - p_2$ is

$$(\hat{p}_1 - \hat{p}_2) \pm z^* SE$$

In this formula the standard error SE of $\hat{p}_1 - \hat{p}_2$ is

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and z^* is the critical value for the standard Normal density curve with area C between $-z^*$ and z^*.

Assumptions and Conditions When Comparing Proportions

Independence Assumptions
Independent Response Assumption:
Within each group, we need independent responses from the cases. We cannot check that for certain, but randomization provides evidence of independence. So, we need to check the following:

- Randomization Condition: The data in each group should be drawn independently and at random from a population or generated by a completely randomized designed experiment.
- The 10% Condition: If the data are sampled without replacement, the sample should not exceed 10% of the population. If samples are bigger than 10% of the target population, random draws are no longer approximately independent.
- Independent Groups Assumption: The two groups we are comparing must be independent from each other.

Assumptions and Conditions When Comparing Proportions

Sample Size Condition

Each of the groups must be big enough. As with individual proportions, we need larger groups to estimate proportions that are near 0% and 100%. We check the success/failure condition for each group.

- Success/ Failure Condition: Both groups are big enough that at least 10 successes and at least 10 failures have been observed in each group or will be expected in each (when testing hypothesis).

Note: Two-sided significance tests (later we will discuss this concept) are robust against violations of this condition. In this case, we can conduct significance tests with smaller sample sizes. In practice, the two-sided significance test works well if there are at least five successes and five failures in each sample.

Comparing Two Proportions: Epidural and Nursing At Six Months

There is some concern that if a woman has an epidural to reduce pain during childbirth, the drug can get into the baby's bloodstream, making the baby sleepier and less willing to breastfeed. In 2006, the International Breastfeeding Journal published results of a study conducted at Sydney University. Researchers followed up on 1178 births, noting whether the mother had an epidural and whether the baby was still nursing after six months. The results are summarized in a contingency table.
Do breastfeeding proportions differ between mothers who had epidural and those who did not? Let p_1 denote the proportion among mothers that had epidural who are breastfeeding at 6 months. Let p_2 denote the proportion among mothers that did not have epidural who are breastfeeding at 6 months.

Data

Epidural	Breasfeeding at 6 Months	
	Yes	No
Yes	206	190
No	498	284

Checking Assumptions and Conditions

Randomization Condition:
We do not know whether mother were randomly selected, but we can view them as representative of a larger collection of mothers under similar conditions.
Independent Groups Assumption:
It is reasonable to believe that mothers who had epidural and mother who did not have epidural to reduce pain during birth are independent of each other. **10% Condition:** We can imagine many more mothers under similar conditions.
Success/Failure Condition: For mothers who had epidural, the count for successes was 206, and for failure was 190; For mothers who did not have epidural the count for successes was 498, for failure was 284; The observed numbers of both success and failures are more than 10 for both groups.
Since these conditions are met, we can use a two-proportion Z-CI.

Data (again)

Conditional Percentages are also displayed in each cell.

Epidural	Breasfeeding at 6 Months Yes	No
Yes	206	190
	52%	48%
No	498	284
	64%	36%

Is there a difference in proportion of breastfeeding mothers who had epidural and those who did not?
Let p_1 denote the proportion among mothers that had epidural who are breastfeeding at 6 months. Let p_2 denote the proportion among mothers that did not have epidural who are breastfeeding at 6 months.
$\hat{p}_1 \approx \frac{206}{396} = 0.52$ and $\hat{p}_2 \approx \frac{498}{782} = 0.64$.

Confidence Interval for $p_1 - p_2$

$\hat{p}_1 = \frac{206}{396} \approx 0.52$ and $\hat{p}_2 = \frac{498}{782} \approx 0.64$.

The standard error is $SE = \sqrt{\frac{(0.52)(1-0.52)}{396} + \frac{(0.64)(1-0.64)}{782}} \approx 0.0304$.

The 95% confidence interval is:

$$(0.52 - 0.64) \pm 1.96(0.0304)$$

Lower Confidence Limit $= -0.12 - 0.0596 = -0.1796$

Upper Confidence Limit $= -0.12 + 0.0596 = -0.0604$

R Code

```
successes=c(206, 498);

totals=c(396, 782);

prop.test(successes,totals, conf.level=0.95,
correct=FALSE);
```

R Code

```
##
##  2-sample test for equality of proportions without
##  continuity correction
##
## data:  x and n
## X-squared = 14.869, df = 1, p-value = 0.0001152
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.17626992 -0.05698333
## sample estimates:
##    prop 1    prop 2
## 0.5202020 0.6368286
```

Interpretation

We are 95% confident the percent breastfeeding mothers at 6 months for those who had epidural are between 6.04% and 17.96% less than those who did not have epidural.

Example: How to quit smoking

Nicotine patches are often used to help smokers quit. Does giving medicine to fight depression help? A randomized double-blind experiment assigned 244 smokers who wanted to stop to receive nicotine patches and another 245 to receive both a patch and the antidepression drug bupropion. Results: After a year, 40 subjects in the nicotine patch group had abstained from smoking, as had 87 in the patch-plus-drug group. Give a 99% confidence interval for the difference (treatment minus control) in the proportion of smokers who quit.

(1) (baseline) patch only

$n_1 = 244$

$x_1 = 40$

$\hat{p}_1 = \frac{40}{244}$ % stopped

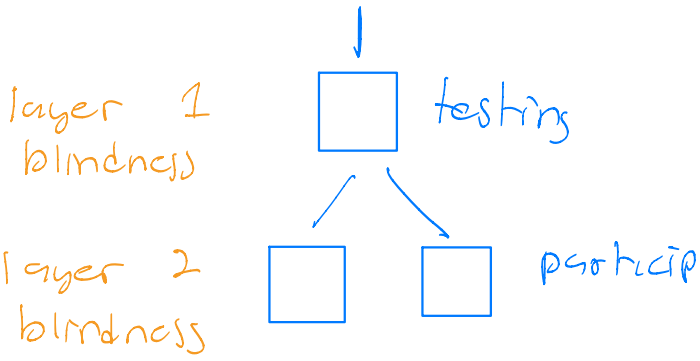
(2) patch + antidepressant

$n_2 = 245$

$x_2 = 87$

$\hat{p}_2 = \frac{87}{245}$

manufacturer (know)



R: qnorm(0.975)

99% CI for difference in proportions

↳ $z_{\alpha/2} = 2.575$ (or 2.57 or 2.58) (exercise)

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$= \left(\frac{40}{244} - \frac{87}{245}\right) \pm (2.575) \sqrt{\frac{\frac{40}{244}\left(1-\frac{40}{244}\right)}{244} + \frac{\frac{87}{245}\left(1-\frac{87}{245}\right)}{245}}$$

$= -0.1912 \pm 0.0996$

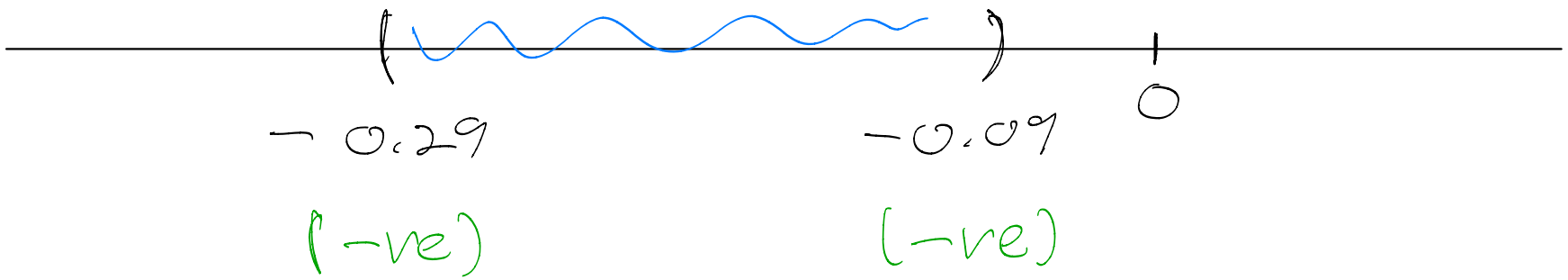
$= (-0.2908, -0.0916)$

Interp:

We are 99% confident the difference in proportions between taking only the patch and taking the patch and anti-depressant is between -0.2908 and -0.0916

(-29%) (-9%)

$p_1 - p_2 = p_{patch} - p_{patch + anti}$



CI suggests $p_1 - p_2 < 0$

$p_{patch} = p_{patch+anti} < 0$

$p_{patch} < p_{patch+anti}$

↑

larger prop stopped with patch

Assumptions

1/ Independent samples

2/ Random samples

3/ $n_1 > 30$ AND $n_2 > 30$

Solution

$\hat{p}_1 = \frac{40}{244} \approx 0.1639$ and $\hat{p}_2 = \frac{87}{245} \approx 0.3551$.
The standard error is
$SE = \sqrt{\frac{(0.1639)(1-0.1639)}{244} + \frac{(0.3551)(1-0.3551)}{245}} \approx 0.0387.$
The 99% confidence interval is:

$$(0.3551 - 0.1639) \pm 2.576(0.0387)$$

Lower Confidence Limit $= 0.1912 - 0.0996 = 0.0915$
Upper Confidence Limit $= 0.1912 + 0.0996 = 0.2908$

R Code

```
successes=c(87, 40);

totals=c(245, 244);

prop.test(successes,totals, conf.level=0.99,
correct=FALSE);
```

R Code

```
##
##  2-sample test for equality of proportions without
##  continuity correction
##
## data:  x and n
## X-squared = 23.237, df = 1, p-value = 1.432e-06
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  0.09152484 0.29081039
## sample estimates:
##    prop 1    prop 2
## 0.3551020 0.1639344
```

R Code

```
successes=c(40,87);

totals=c(244,245);

prop.test(successes,totals, conf.level=0.99,
correct=FALSE);
```

R Code

```
##
##  2-sample test for equality of proportions without
##  continuity correction
##
## data:  x and n
## X-squared = 23.237, df = 1, p-value = 1.432e-06
## alternative hypothesis: two.sided
## 99 percent confidence interval:
##  -0.29081039 -0.09152484
## sample estimates:
##    prop 1    prop 2
## 0.1639344 0.3551020
```

Comparing Variances

Not covered

Comparing Two Population Variances: Independent Sampling

How do you know whether the homogeneity of variance assumption is satisfied?
One simple method involves just looking at two sample variances. Logically, if two population variances are equal, then the two sample variances should be very similar. When the two sample variances are reasonably close, you can be reasonably confident that the homogeneity assumption is satisfied and proceed with, for example, Student t-interval. However, when one sample variance is three or four times larger than the other, then there is reason for a concern. The common statistical procedure for comparing population variances σ_1^2 and σ_2^2 makes an inference about the ratio of $\frac{\sigma_1^2}{\sigma_2^2}$.

Making An Inference for Ratio of Population Variances

To make an inference about the ratio of $\frac{\sigma_1^2}{\sigma_2^2}$ we collect sample data and use the ratio of the sample variances $\frac{s_1^2}{s_2^2}$.

The sampling distribution of $\frac{s_1^2}{s_2^2}$ is based on the two of the assumptions already required for the t procedure:

1. The two sampled populations are Normally distributed.
2. The samples are Normally and independently selected from their respective populations.

When these assumptions are satisfied, the sampling distribution of $\frac{s_1^2}{s_2^2}$ is an F-distribution with $(n_1 - 1)$ numerator degrees of freedom and $(n_2 - 1)$ denominator degrees of freedom.

$100(1-\alpha)$ % CI for $\frac{\sigma_1^2}{\sigma_2^2}$

We know that $\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2} \sim F_{n_1-1,n_2-1}$. So,

$$P\left[F_{n_1-1,n_2-1,1-\alpha/2} < \frac{\sigma_2^2 s_1^2}{\sigma_1^2 s_2^2} < F_{n_1-1,n_2-1,\alpha/2}\right] = 1-\alpha$$

where $F_{n_1-1,n_2-1,1-\alpha/2}$ and $F_{n_1-1,n_2-1,\alpha/2}$ are the values of the F-distribution with n_1-1 and n_2-1 degrees of freedom, leaving areas of $1-\alpha/2$ and $\alpha/2$, respectively, to the right. Rearranging gives

$$P\left[\frac{s_1^2}{s_2^2}\frac{1}{F_{n_1-1,n_2-1,\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}\frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}}\right] = 1-\alpha$$

$100(1-\alpha)$ % CI for $\frac{\sigma_1^2}{\sigma_2^2}$

Using the fact that $F_{n_1-1,n_2-1,1-\alpha/2} = \frac{1}{F_{n_2-1,n_1-1,\alpha/2}}$, we have

$$P\left[\frac{s_1^2}{s_2^2}\frac{1}{F_{n_1-1,n_2-1,\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2}F_{n_2-1,n_1-1,\alpha/2}\right] = 1-\alpha$$

Example

Comparing Two Population Variances Managerial Success Indexes for Two Groups.
Behavioural researchers have developed an index designed to measure managerial success. The index (measured on a 100-point scale) is based on the manager's length of time in the organization and their level within the term; the higher the index, the more successful the manager. Suppose a researcher wants to compare the average index for the two groups of managers at a large manufacturing plant. Managers in group 1 engage in high volume of interactions with people outside the managers' work unit (such interaction include phone and face-to-face meetings with customers and suppliers, outside meetings, and public relation work). Managers in group 2 rarely interact with people outside their work unit.

Example

Independent random samples of 12 and 15 managers are selected from groups 1 and 2, respectively, and success index of each is recorded. Note: The response variable is "Managerial Success Indexes".

- Managerial success indexes is a continuous quantitative variable, measured on 100-point scale.

The explanatory variable is "Type of group".

- Type of group (Group 1: Interaction with outsiders, Group 2: Fewer interactions) is a nominal categorical variable.

R Code

```r
# Importing data file into R;

success=read.csv(file="success.csv",header=TRUE);

# Getting names of variables;

names(success);

# Attaching data file;
attach(success);
```

R Code

```
## [1] "Success_Index" "Group"
```

R Code (Descriptive Statistics)

```r
# loading library mosaic;

library(mosaic);

favstats(Success_Index~Group);
```

Note. Group 1 = "interaction with outsiders" and Group 2 = "fewer interaction".

R Code (Descriptive Statistics)

```
##   .group min    Q1 median    Q3 max     mean       sd  n
## 1      1  53 62.25   65.5 69.25  78 65.33333 6.610368 12
## 2      2  34 42.50   50.0 54.50  68 49.46667 9.334014 15
##   missing
## 1       0
## 2       0
```

Note. Group 1 = "interaction with outsiders" and Group 2 = "fewer interactions".

R Code (Critical values)

```r
# Finding F-critical value with R
# alpha = 0.05;
# alpha/2 = 0.025;

qf(0.025, df1=11, df2=14);

## [1] 0.2977245

qf(0.975, df1=11,df2=14);

## [1] 3.09459
```

95% Confidence Interval for $\frac{\sigma_1^2}{\sigma_2^2}$

$$\left(\frac{6.610368^2}{9.334014^2(3.0945898)}, \frac{6.610368^2}{9.334014^2(0.2977245)} \right)$$

$$(0.1621, 1.6846)$$

Since "1" is in this 95% CI, we have no evidence that the population variances of managerial success indexes for the two groups differ.

R Code (Critical values)

```
# 95% CI for the ratio of two variances;

var.test(Success_Index~Group);

##
##  F test to compare two variances
##
## data:  Success_Index by Group
## F = 0.50155, num df = 11, denom df = 14, p-value =
## 0.2554
## alternative hypothesis: true ratio of variances is not equa
## 95 percent confidence interval:
##  0.1620733 1.6846122
## sample estimates:
## ratio of variances
##          0.5015503
```