This course was developed as a part of several VLIR-UOS projects:

- Cross-cutting Statistics: 2011-2016, 2017.
- Cross-cutting Statistics: 2017.
- Statistics for development : 2018-2022.
- The >rR-BioStat platform ITP project: 2024-2026.

The >eR-Biostat initiative

Making R based education materials in statistics accessible for all

Introduction to Statistical inference and estimation using R: Inference for numerical data (one population & two populations)

Developed by

Ziv Shkedy (Hasselt Univesrsity) and Tadesse Awoke (Gondar University)

LAST UPDATE: 03/2024

Visit us on Facebook ER-BioStat

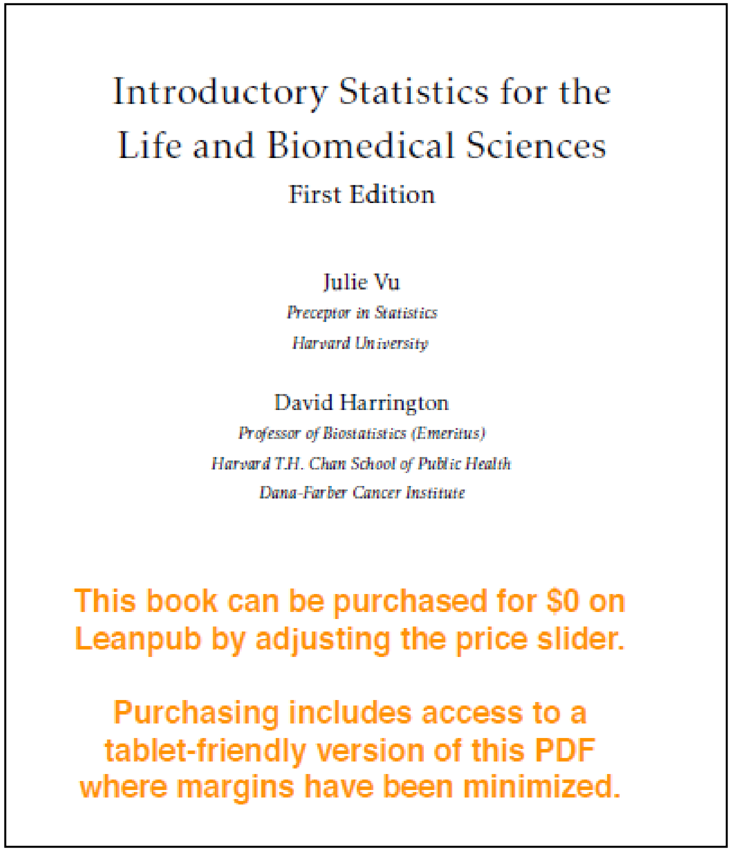
GitHub https://github.com/eR-Biostat

twitter @erbiostat

Development team

- Tadele Worku Mengesha (Gondar University).
- Abdisa Gurmessa (Jmma University).
- Ziv Shkedy (Hasselt Univesrsity).
- Tadesse Awoke (Gondar University).

Recommended reading



Introductory Statistics for the Life and Biomedical Sciences

First Edition

Julie Vu

Preceptor in Statistics

Harvard University

David Harrington

Professor of Biostatistics (Emeritus)

Harvard T.H. Chan School of Public Health

Dana-Farber Cancer Institute

This book can be purchased for $0 on Leanpub by adjusting the price slider.

Purchasing includes access to a tablet-friendly version of this PDF where margins have been minimized.

Chapter 5: Inference for numerical data

- We cover mainly Chapter 5.
- The examples that are used for illustration are not the same as the examples in the book.

Software

- R functions:
  - t.test().

YouTube tutorials

- YouTube tutorials are available for:
  - Two-Sample t Test in R: Independent Groups (R Tutorial 4.2)(host: MarinStatsLectures-R Programming & Statistics): https://www.youtube.com/watch?v=RlhnNbPZC0A&t=70s.
  - Statistics with R - Two sample t-test with R (t.test) (host: Dragonfly Statistics): https://www.youtube.com/watch?v=e5JJxBb80CQ.

Datasets

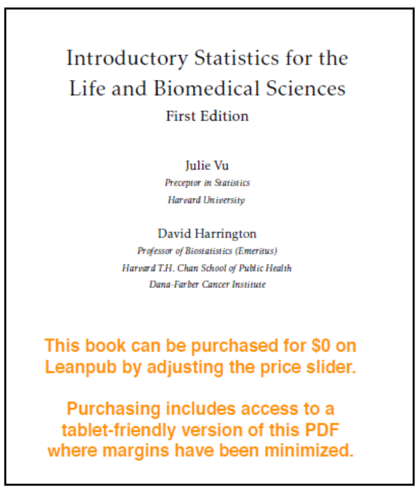
- Data are given as a part of R programs for the course.

- External datasets (which are not given as a part of the R code) and used for illustration are available online:

Topics

1. Inference for one sample mean with t-distribution.
2. One-sample t test.
3. Two-samples test:
    1. Independent samples.
    2. Paired samples.

Inference for one-sample mean with the t distribution


Introductory Statistics for the
Life and Biomedical Sciences
First Edition

Julie Vu
Preceptor in Statistics
Harvard University

David Harrington
Professor of Biostatistics (Emeritus)
Harvard T.H. Chan School of Public Health
Dana-Farber Cancer Institute

This book can be purchased for $0 on Leanpub by adjusting the price slider.

Purchasing includes access to a tablet-friendly version of this PDF where margins have been minimized.

Section 5.1

This part (slides 9 – slide 60) was covered also in the first slides set.

5.1: Inference for one-sample means with the t distribution

t-test for a population

- We assume that X~N(μ,σ^2) & n is small
- For this test, we used the Student t distribution.

as $$X \sim N(\mu,\sigma^2)$$

than: $$\cancel{\bar{X} \sim N(\mu,\frac{S^2}{n})}$$

and $$T_{\bar{X}} = \frac{\bar{X}-\mu}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)$$

X has a normal distribution with unknown μ and σ^2. n is small

$$E(S^2) = \sigma^2$$

example

- A researcher would like the following hypotheses :

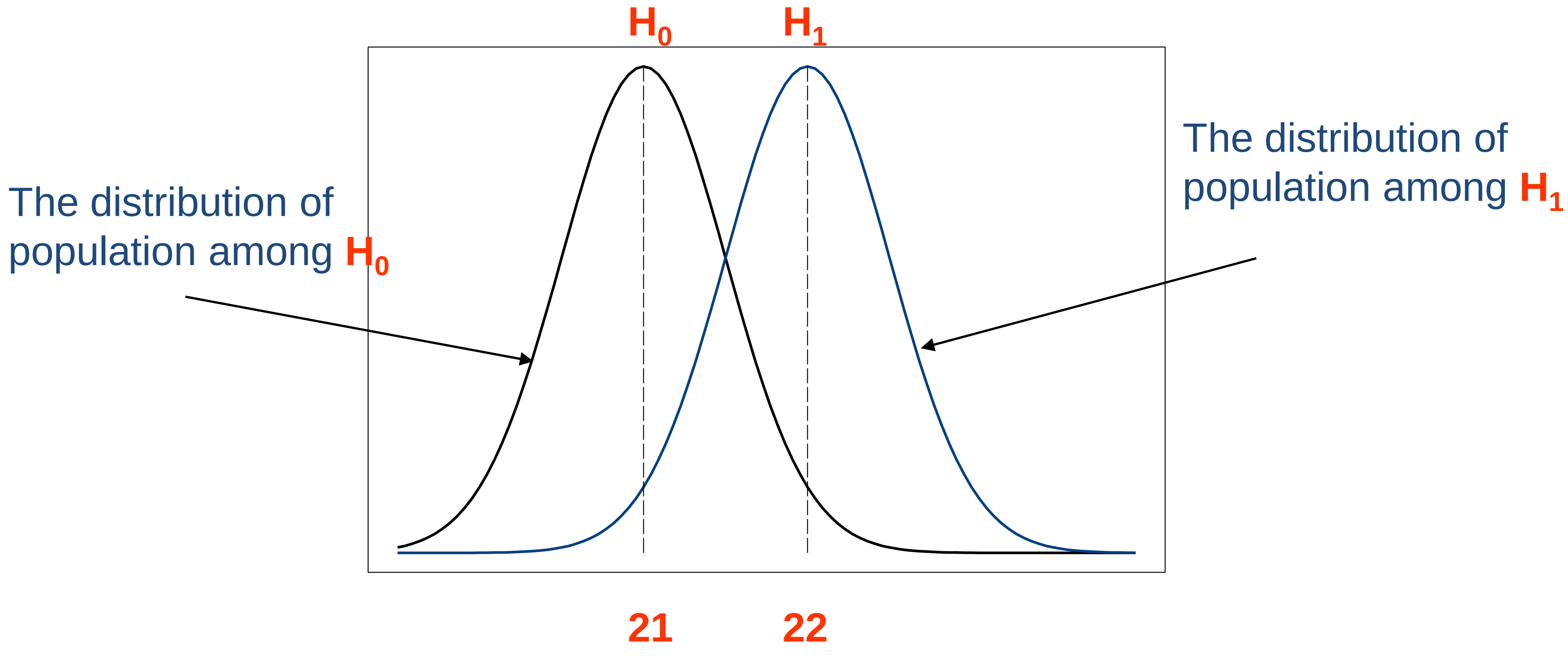
$$H_0 : \mu = 21$$
$$H_1 : \mu = 22$$

- We assume that

$$X \sim N(\mu, \sigma^2)$$

The distribution of the population

$$X \sim N(21, \sigma^2) \qquad under \quad H_0$$
$$X \sim N(22, \sigma^2) \qquad under \quad H_1$$



the sample

- To test the hypotheses, we draw a sample of size 9 (n = 9) from the population.
- X has a normal distribution with unknown μ and σ^2.

n is small

$$\left.\begin{aligned} &X_i \sim N(\mu, \sigma^2) \\ &n = 9 \qquad (small) \\ &\sigma^2 : unknown \end{aligned}\right\}$$

$$\boxed{\frac{\overline{X} - 21}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)}$$

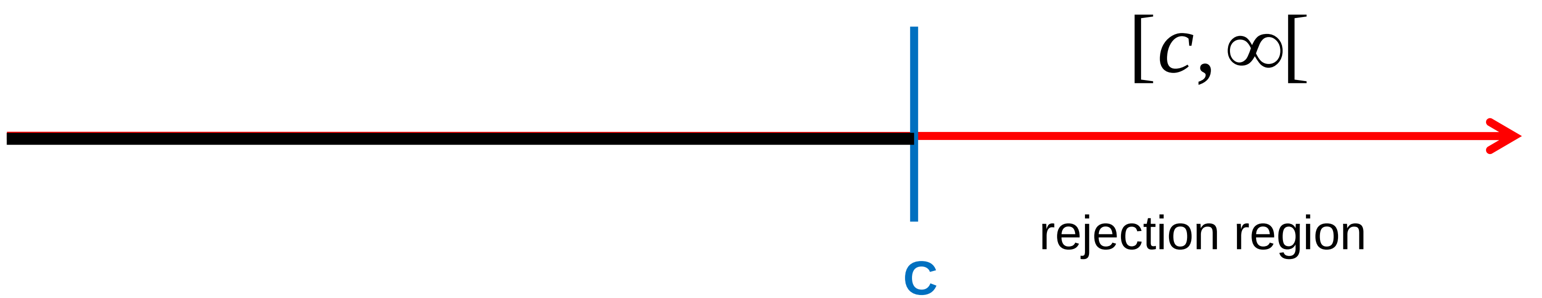
The distribution of the test statistic population under $\mathbf{H_0}$

The rejection region

$$H_0 : \mu = 21$$
$$H_1 : \mu = 22$$

$$\mu_0 < \mu_1$$

when the value of $\overline{X}$ is greater than c then we reject the null hypothesis



when the value of T is larger than c then we reject the null hypothesis

$[t, \infty[$

The choice of c

- We choose c so that Type I error 0.05.

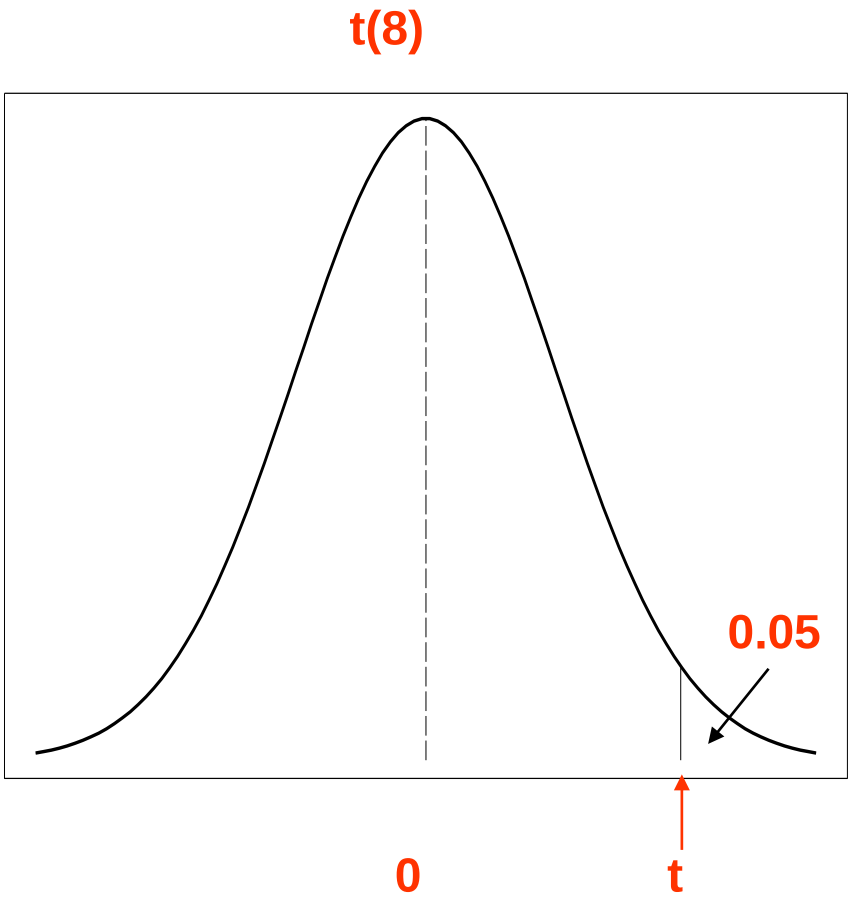
$$\alpha = 0.05$$

$$\bar{X} > c \Rightarrow H_0 \quad reject \quad \Longleftrightarrow \quad T > t \Rightarrow H_0 \quad reject$$

$$P(\bar{X} > c) = P(T > t) = 0.05$$

if the null hypothesis is correct

The critical point



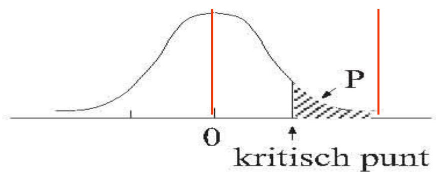
$$P(T > t) = \alpha$$

The distribution of the test statistic under H_0

$$\boxed{\frac{\bar{X} - 21}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)}$$

Student's t-distribution and critical point

Tabel 4 : Kritische punten student t verdeling



P / v.g.	.25	.10	.05	.025	.010	.005	.001
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.765	1.638	2.353	3.182	4.541	5.841	10.213
4	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	.700	1.372	1.812	2.228	2.764	3.169	4.144
11	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.688	1.330	1.734	2.101	2.552	2.878	3.610
19	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.685	1.319	1.714	2.069	2.500	2.807	3.485
24	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	.684	1.316	1.708	2.060	2.485	2.787	3.450
26	.684	1.315	1.706	2.056	2.479	2.779	3.435
27	.684	1.314	1.703	2.052	2.473	2.771	3.421
28	.683	1.313	1.701	2.048	2.467	2.763	3.408
29	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	.674	1.282	1.645	1.960	2.326	2.576	3.090

$$n = 9(small)$$

$$df. = 8$$

$$\alpha = 0.05$$

$$P(T > t) = 0.05$$

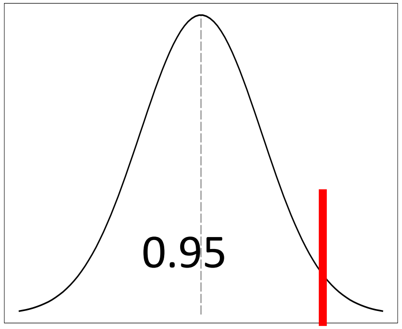
$$P(T > 1.86) = 0.05$$

Student's t-distribution and critical point in R

```
> df<-8
> alpha<-0.05
> crit.val<-qt(1-alpha,df)
> crit.val
[1] 1.859548

> pt(crit.val,df)
[1] 0.95
```

$$P(T \leq 1.86) = 0.95$$



$$n = 9(small)$$

$$df. = 8$$

$$\alpha = 0.05$$

$$P(T > t) = 0.05$$

$$P(T > 1.86) = 0.05$$

The sample

subject	X_i
1	22
2	19
3	17
4	26
5	21
6	20
7	29
8	27
9	22

n=9

$$\bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i = 22.556$$

$$s^2 = \frac{1}{9-1}\sum_{i=1}^{9}\left(x_i - \bar{x}\right) = 3.972^2$$

The estimators for the unknown parameters (μ and σ ²) in the population

The rejection region & statistic

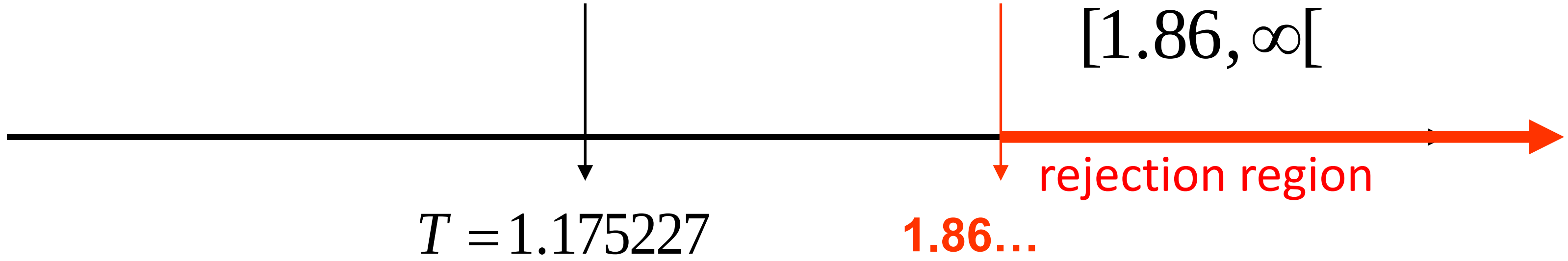
$s^2 = 3.972$

$\bar{x} = 22.556$

$n = 9$

$$\frac{\bar{x} - 21}{\sqrt{\frac{3.972^2}{9}}} = 1.175227$$

$T < t \Rightarrow$ We do not reject H_0



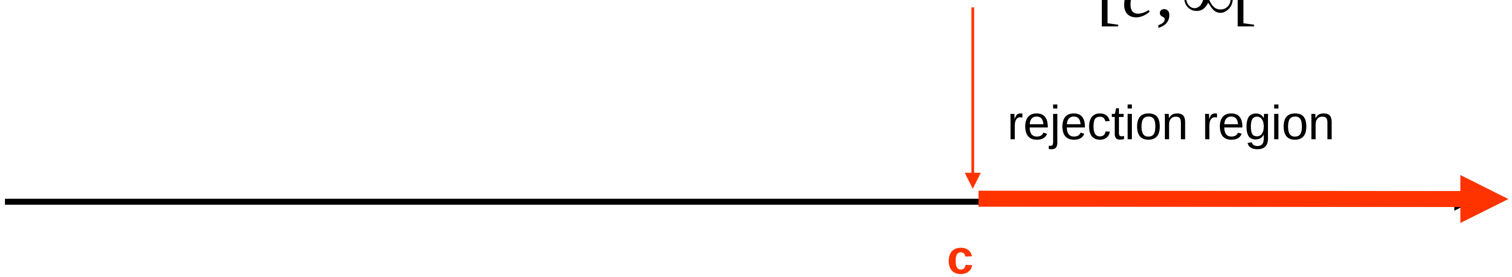
The rejection region

$$P\left(T > \frac{c-21}{\sqrt{\frac{S^2}{n}}}\right) = 0.05$$

$$P(T > 1.86) = 0.05$$

$$\frac{c-21}{\sqrt{\frac{S^2}{n}}} = 1.86 \quad \Rightarrow \quad c = 1.86 \times \sqrt{\frac{S^2}{n}} + 21$$

$$[c, \infty[$$



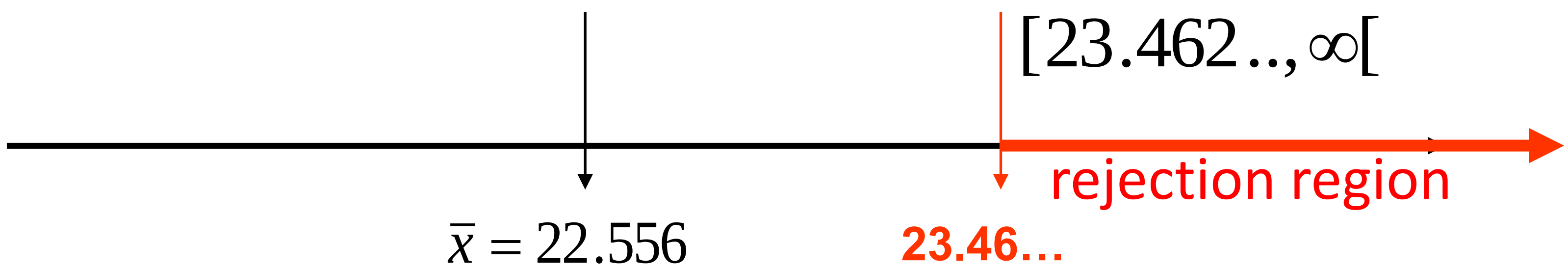
$$c = t \times \sqrt{\frac{S^2}{n}} + \mu_0$$

The rejection region

$$s^2 = 3.972$$
$$\bar{x} = 22.556$$
$$n = 9$$

$$c = 1.86 \times \sqrt{\frac{3.972^2}{9}} + 21 = 23.46264$$

$\bar{x} < c \Rightarrow$ We do not reject H_0



the checklist

Step	information	example
1	The hypotheses (the qualifying problem)	$H_0 : \mu = 21$ $H_1 : \mu = 22$
2	The distribution in the population and σ2	$X \sim N(\mu, \sigma^2)$ σ^2 not known
3	sample size	$n = 9 < 30$
4	The distribution of the sample mean	Unknown
5	The level of significance	$\alpha = 0.05$
6	The test statistic	$\boxed{\dfrac{\overline{X} - 21}{\sqrt{\dfrac{S^2}{n}}} \sim t(8)}$
7	The distribution of the test statistic	
8	The critical point (or points)	1.86 t(8)

R code

```
> x=c(22,19,17,26,21,20,29,27,22)
> xbar=mean(x)
> mu = 21
>  s = sd(x)
>  n = length(x)
>  t = (xbar-mu)/(s/sqrt(n))
>  t                      # test statistic
[1] 1.174854
> crit.val = qt(1-alpha, n-1, lower.tail = TRUE)
> crit.val     # critical value
[1] 1.859548
```

The test statistic 1.174854 is greater than the critical value of 1.859548. Hence, at 0.05 significance level, we can reject the null hypothesis.

Example:

heights and weights for American women aged 30–39.

```
> womenheight=women$height
> t.test(womenheight,mu=60,conf.level=0.90)

        One Sample t-test

data:  womenheight
t = 4.3301, df = 14, p-value = 0.000692
alternative hypothesis: true mean is not equal to 60
90 percent confidence interval:
 62.96621 67.03379
sample estimates:
mean of x
       65
```

Testing a hypothesis about a Population parameter (2)

One sided and two-sided testing problems

The hypothesis and the alternative hypothesis

- In the previous example, we tested the hypothesis that the mean of a normal distribution with unknown variance equal to a certain value (21).
- As an alternative hypothesis we mean that it was equal to another specified
- value (22).

$$H_0 : \mu = 21$$
$$H_1 : \mu = 22$$

- In practice, the researcher usually do not know the exact details of the alternative hypothesis.

Case (a)

The average under H_1 is smaller than the average under H_0

$$\boxed{\begin{aligned} H_0 &: \mu = \mu_{H_0} \\ H_1 &: \mu < \mu_{H_0} \end{aligned}}$$

null hypothesis

alternative hypothesis

One sided test problem

case (b)

The average under H_1 is greater than the average under H_0 :

$$\boxed{\begin{aligned} H_0 &: \mu = \mu_{H_0} \\ H_1 &: \mu > \mu_{H_0} \end{aligned}}$$

null hypothesis

alternative hypothesis

One sided test problem

Case (c)

The average under H_0 is not equal to the mean under H_1 :

$$\boxed{\begin{aligned} H_0 &: \mu = \mu_{H_0} \\ H_1 &: \mu \neq \mu_{H_0} \end{aligned}}$$

null hypothesis

alternative hypothesis

two sided test problem

example (case a)

$$\boxed{\begin{aligned} H_0 &: \mu = \mu_{H_0} \\ H_1 &: \mu < \mu_{H_0} \end{aligned}}$$

One sided test

Example: one-tailed test

- A gynecologist says that girls at birth, averaging less than 51 cm.
- His colleague Judge reproach him that his claim is based on a prejudice, and that the average length is 51 cm indeed.
- They draw a sample of 100 girls.
- In the sample:

$$\bar{x} = 50.8 \qquad \& \qquad s^2 = 1.6$$

The testing problem

The choice of H_1 reflects here the assertion of the first gynecologist

$$H_0 : \mu = 51$$ null hypothesis

$$H_1 : \mu < 51$$ alternative hypothesis

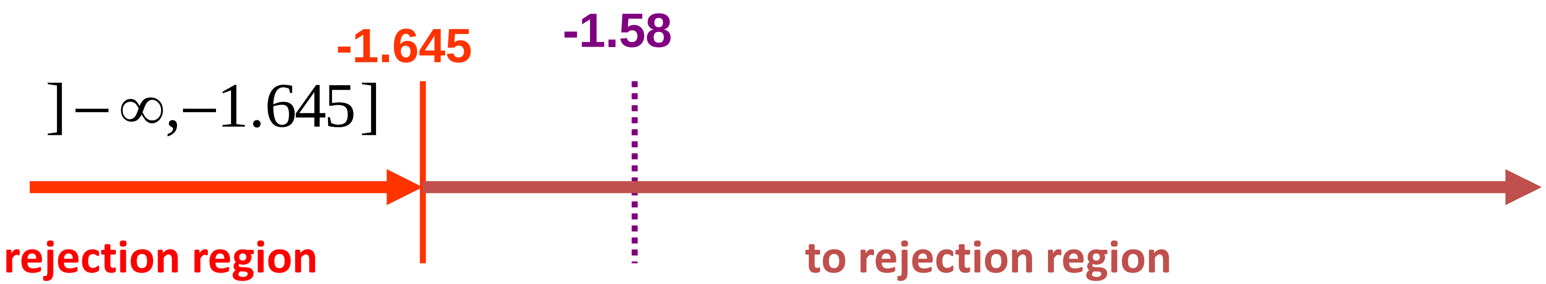
One-sided test

The test statistic

- We now supplement the sample values and find:

$$\frac{\bar{x}-51}{\sqrt{\frac{s^2}{n}}}=\frac{50.8-51}{\sqrt{\frac{1.6}{100}}}=-1.58$$

- Conclusion: at significance level of 5%, the length of girls at birth 51 cm.



The checklist

Stap	information	example
1	The hypotheses (the qualifying problem)	One-sided test
2	The distribution in the population & σ^2	
3	sample size	
4	The distribution of the sample mean under H_0	
5	The level of significance	
6	The test statistic	$\boxed{\dfrac{\overline{X}-51}{\sqrt{\dfrac{S^2}{n}}} \sim ?}$
7	The distribution of the review greatness	
8	The critical point (or points)	

R code

```
> xbar=50.8;s=sqrt(1.6);n=100;H0=51
> test.statgy=(xbar-H0)/(s/sqrt(n))
> test.statgy
[1] -1.581139
> crit.point1=qnorm(0.95,lower.tail=TRUE)#p=0.05 one tailed
> -crit.point1
[1] -1.644854
```

Example (case c)

$$\boxed{\begin{aligned} H_0 &: \mu = \mu_{H_0} \\ H_1 &: \mu \neq \mu_{H_0} \end{aligned}}$$

two-tailed test

Example: two-tailed test

- At a certain university one takes many years an intelligence-off normally distributed with mean score results yields 115 (115 = μ under H_0).
- An administrator wants now for the new class to test the hypothesis that the mean is the same as in previous years.
- He takes a sample of size 50, and: mean 118 and variance 98.

n $\bar{X}$ S^2

The testing problem

The choice of H_1 is to be determined by the consideration that the administrator has no idea whether the new crop is better or worse than the previous

$$H_0 : \mu = 115$$ Null hypothesis

$$H_1 : \mu \neq 115$$ Alternative hypothesis

two-tailed test

The rejection region (sided test)

$$\frac{\bar{x}-115}{\sqrt{\frac{s^2}{n}}} = \frac{118-115}{\sqrt{\frac{98}{50}}} = 2.14$$

The test statistic

2.14 > 1.96 ⇒ the administrator rejects H_0 at significance level 0.05.

The checklist

Stap	information	Example
1	The hypotheses (the testing problem)	One-sided test
2	The distribution in the population& σ^2	
3	sample size	
4	The distribution of the sample mean under H_0	
5	The level of significance	
6	The test statistic	$\boxed{\frac{\bar{X}-115}{\sqrt{\frac{S^2}{n}}} \sim ?}$
7	The distribution of the test statistic	
8	The critical point (or points)	

R code

```
> xbar=118;s=sqrt(98);n=50;H0=115
> test.statcrop=(xbar-H0)/(s/sqrt(n))
> test.statcrop
[1] 2.142857
> alpha = 0.05
> crit.pointcrop = qnorm(1-alpha/2)
> crit.pointcrop
[1] 1.959964
> -crit.pointcrop
[1] -1.959964
> c(-crit.pointcrop,crit.pointcrop)
[1] -1.959964  1.959964
```

population, n & σ^2

- from above examples show that we are always three things to keep in mind:

1. which assumption we make about the distribution of the population?
2. is the variance σ^2 is known or should they be estimated using S^2?
3. how big is the sample (which is the value of n)?

1: n large

For n large

$$\boxed{\frac{\bar{X}-\mu_{H_0}}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)}$$

If σ² is known

$$\boxed{\frac{\bar{X}-\mu_{H_0}}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)}$$

If σ² is unknown

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

2: n small & normal distribution

2: n Small & normal distribution

$$\boxed{\frac{\bar{X}-\mu_{H_0}}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)}$$

If σ^2 is known

$$\boxed{\frac{\bar{X}-\mu_{H_0}}{\sqrt{\frac{S^2}{n}}} \sim t_{(n-1)}}$$

If σ^2 is unknown

n	σ^2	statistics	distribution of the population	Distribution for statistical
large	known	$\boxed{\dfrac{\overline{X}-\mu_{H_0}}{\sqrt{\dfrac{\sigma^2}{n}}}}$	normal distribution or not known	Z(0,1)
large	not known	$\boxed{\dfrac{\overline{X}-\mu_{H_0}}{\sqrt{\dfrac{S^2}{n}}}}$	normal distribution or not known	Z(0,1)
small	known	$\boxed{\dfrac{\overline{X}-\mu_{H_0}}{\sqrt{\dfrac{\sigma^2}{n}}}}$	normaal verdeling	Z(0,1)
small	not known	$\boxed{\dfrac{\overline{X}-\mu_{H_0}}{\sqrt{\dfrac{S^2}{n}}}}$	normal distribution	t(n-1)
small	not known		normal distribution	Not for this course

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Part 7:
The P-value

The significance level and the critical point

- In the examples on hypothesis testing, we have until now always pre specified significance level α (usually α = 0.05).
- We determine the rejection region so:

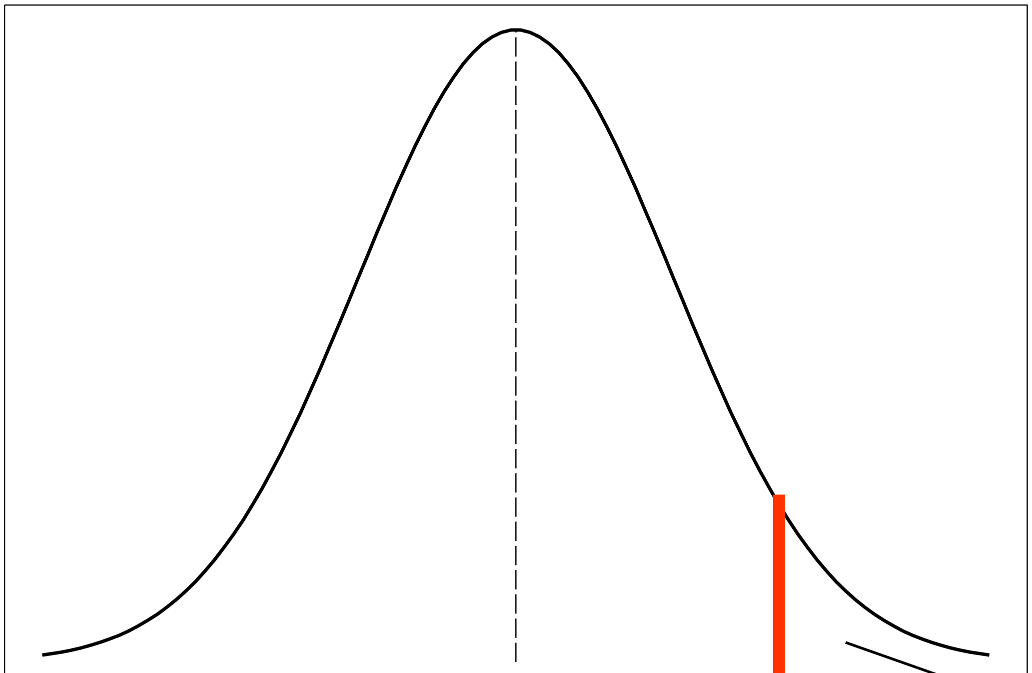
$$P_{H_0}\left(\bar{x} \in [c, \infty[\right) = \alpha$$

The level of significance and the p-value

- The relationship between the p-value and the level of significance is clear:

H_0 rejected on
significance level α if and only if
the p-value $<\alpha$

Right-sided test $\mu_0 < \mu_1$



$$[c, \infty[$$

The distribution of the test statistic under H_0

observed value of $\frac{\bar{x} - \mu_{H_0}}{\sqrt{\frac{\sigma^2}{n}}}$ of $\frac{\bar{x} - \mu_{H_0}}{\sqrt{\frac{S^2}{n}}}$

p-value

Example 1 (p-value): right-tailed test

population

$$X_i \sim N(\mu, \sigma^2)$$
$$n = 9 \qquad (small)$$
$$\sigma^2 : unknown$$

sample

$$\bar{x} = 22.556$$
$$s^2 = 3.972^2$$

$$H_0 : \mu = 21$$
$$H_1 : \mu > 21$$

$$\frac{\bar{x} - 21}{\sqrt{\frac{3.972^2}{9}}} = 1.175227$$

We look at student t distribution with 8 df

$$p - value = P(T > 1.175227) = 0.1481026$$

P-value = 0.1481026 > 0.05, we can not reject H_0.

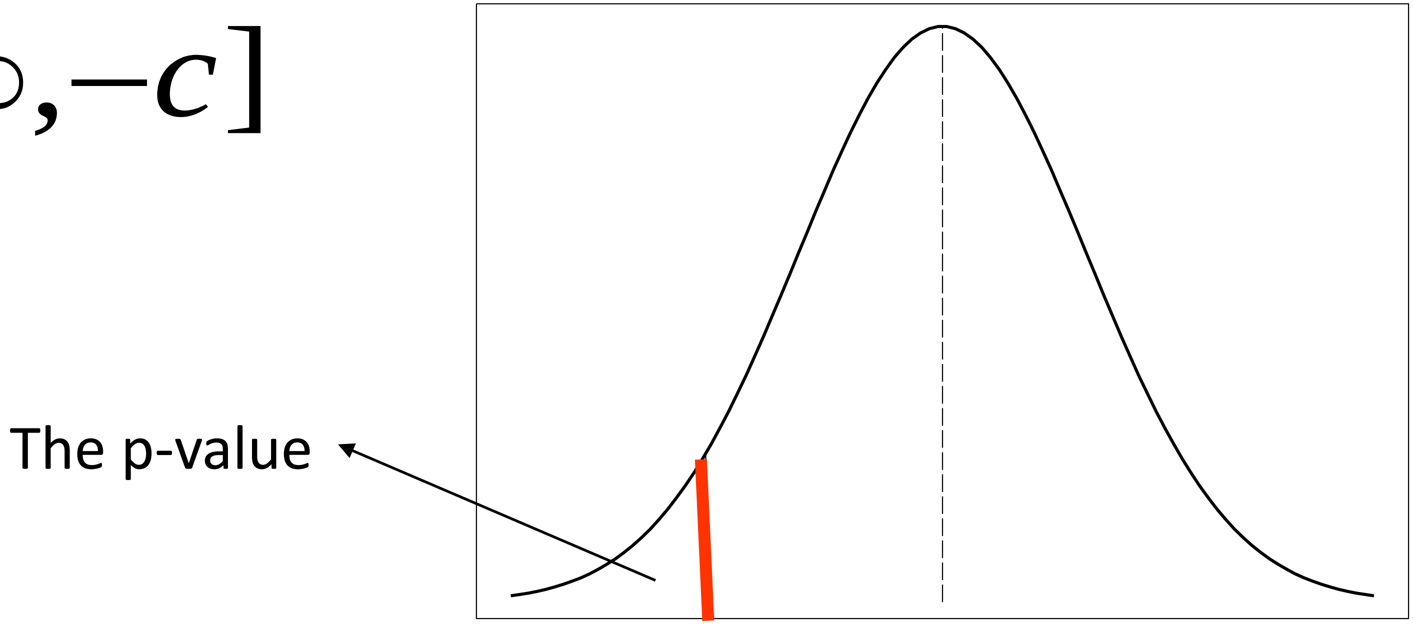
R code

```
> x=c(22,19,17,26,21,20,29,27,22)
> xbar=mean(x)
> mu = 21
>  s = sd(x)
>  n = length(x)
>  t = (xbar-mu)/(s/sqrt(n))
> alpha = .05
>  pval = pt(t,df=n-1, lower.tail=FALSE)
>  pval
[1] 0.1369174
```

P value

Left-sided test $\mu_0 > \mu_1$

$$]-\infty, -c]$$



observed value of $\dfrac{\vec{x}-\mu_{H_0}}{\sqrt{\dfrac{\sigma^2}{n}}}$ of $\dfrac{\vec{x}-\mu_{H_0}}{\sqrt{\dfrac{S^2}{n}}}$

Example 2 (p-value): left-tailed test

$$H_0 : \mu = 51$$
$$H_1 : \mu < 51$$

$$\boxed{\frac{50.8 - 51}{\sqrt{\frac{1.6}{100}}} = -1.58}$$

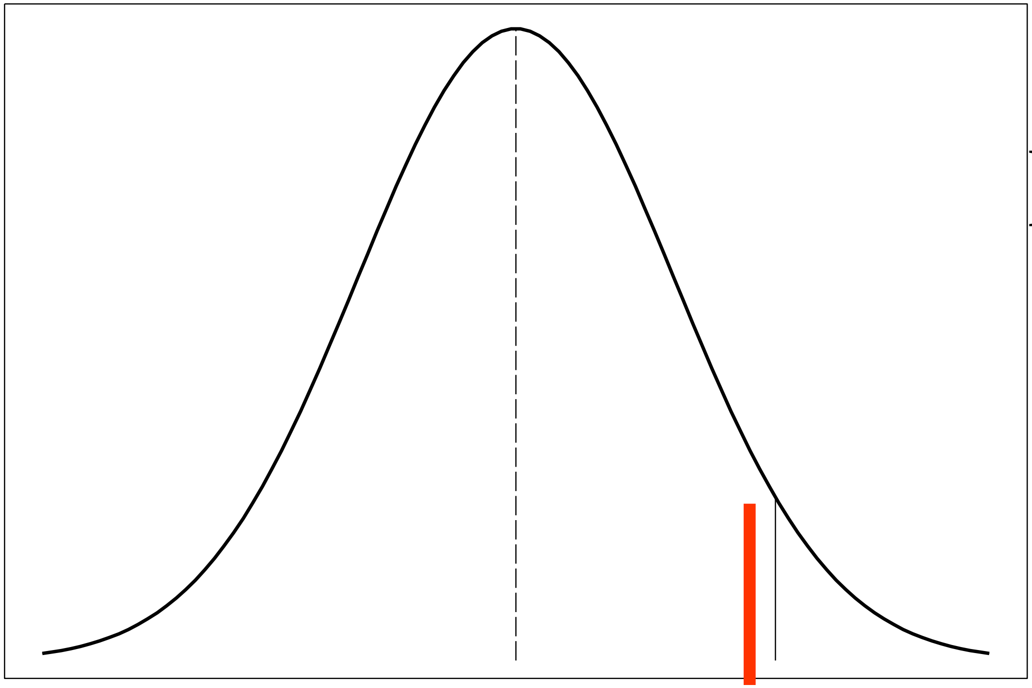
$$p - value = P(Z < -1.58) = 0.0571$$

for each significance level > 0.0571 H_0 will be rejected

R code

```
>  xbar=50.8;s=sqrt(1.6);n=100;H0=51
> test.statgy=(xbar-H0)/(s/sqrt(n))
> test.statgy
[1] -1.581139
> pval1 = pt(test.statgy,df=n-1,
+ lower.tail=TRUE)
> pval1
[1] 0.05851802
```

two-tailed test



$$]-\infty,-c]\cup[c,\infty[$$

observed value of $\dfrac{\bar{x}-\mu_{H_0}}{\sqrt{\dfrac{\sigma^2}{n}}}$ of $\dfrac{\bar{x}-\mu_{H_0}}{\sqrt{\dfrac{S^2}{n}}}$

$$p-value = 2\times P\left(Z > \frac{\bar{x}-\mu_{H_0}}{\sqrt{\dfrac{s^2}{n}}} \right)$$

Example 3 (p-value)

$$H_0 : \mu = 115$$
$$H_1 : \mu \neq 115$$

$$\boxed{\frac{98-115}{\sqrt{\frac{98}{50}}} = 2.14}$$

$$p - value = 2 \times P\left(Z > 2.14\right) = 2 \times [1 - \Phi(2.14)] = 0.0324$$

for each significance level> 0.0324 H_0 will be rejected

R code

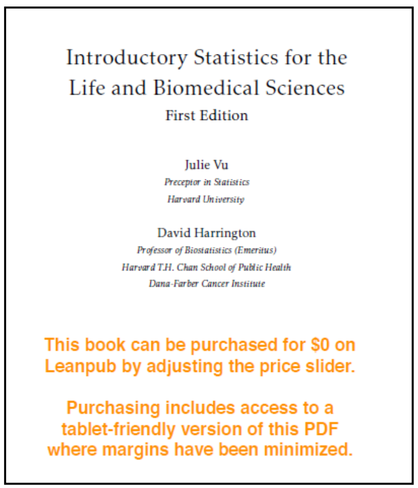
```
> bar=118;s=sqrt(98);n=50;H0=115
> test.statcrop=(xbar-H0)/(s/sqrt(n))
> test.statcrop
[1] 2.142857
>  2*(1-pnorm(test.statcrop))
[1] 0.03212457
```

The level of significance and the p-value

- Statistical computer packages give as output a hypothesis test the p-value.
- A generally accepted criterion (e. g in scientific publications) is as follows

1. If the P-value <0.05, then H0 is rejected, and then the results are significant.
2. if the p-value> 0.05, then H0 is not rejected, and then the results are not significant.

Hypotheses tests and Confidence intervals for two populations



Introductory Statistics for the Life and Biomedical Sciences
First Edition

Julie Vu
Preceptor in Statistics
Harvard University

David Harrington
Professor of Biostatistics (Emeritus)
Harvard T.H. Chan School of Public Health
Dana-Farber Cancer Institute

This book can be purchased for $0 on Leanpub by adjusting the price slider.

Purchasing includes access to a tablet-friendly version of this PDF where margins have been minimized.

- Section 5.2: two sample test for paired data
- Section 5.3: two sample test for independent data

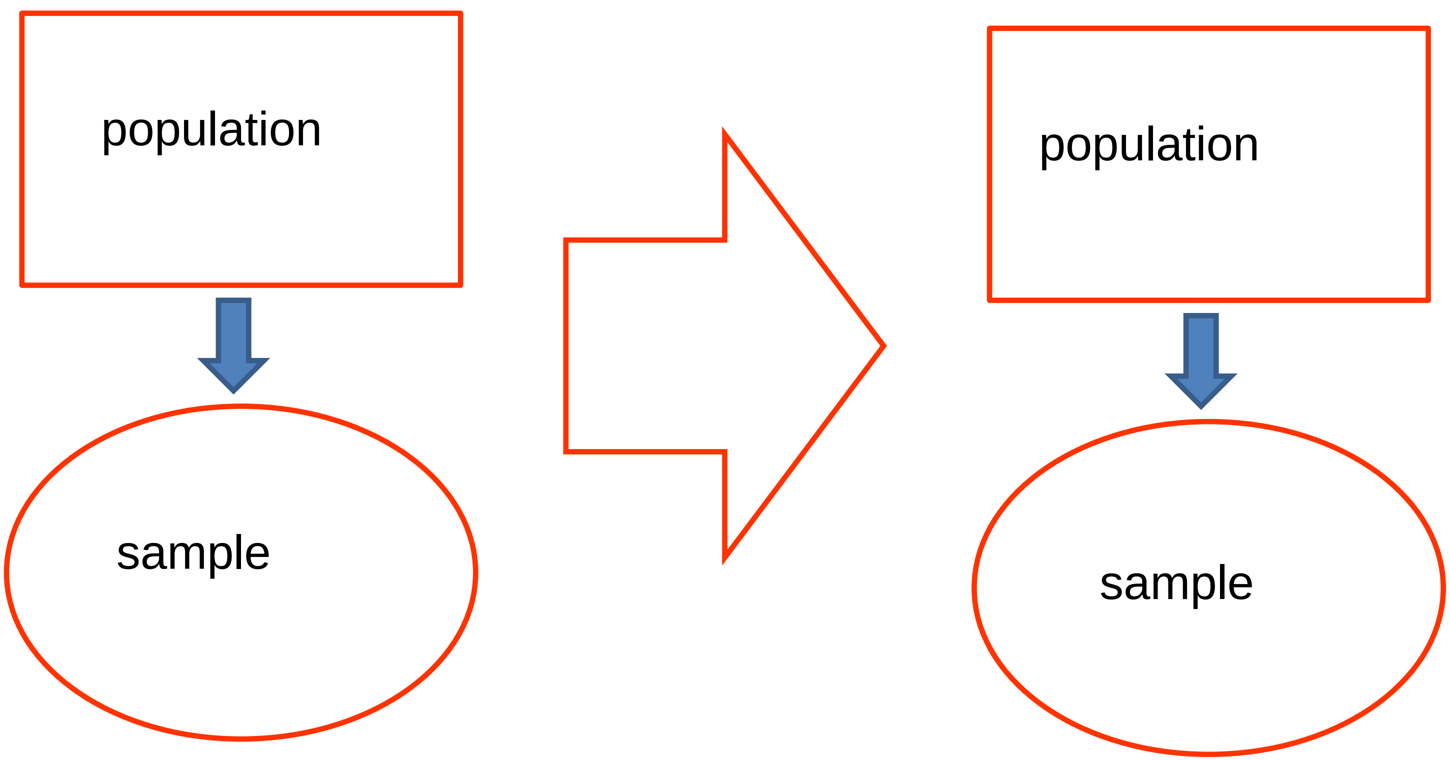
Objectives

- To distinguish between a problem associated with measurements and a two-sample problem using example.
- To perform a test of hypothesis about the difference of two population means and two population proportions.
- To calculate a confidence interval for the difference of two population means and the difference of two population proportions.
- The tests and confidence intervals can perform and interpret using R.

Section 5.1 → Section 5.2

Section 5.1

Section 5.2



One measurement per individual

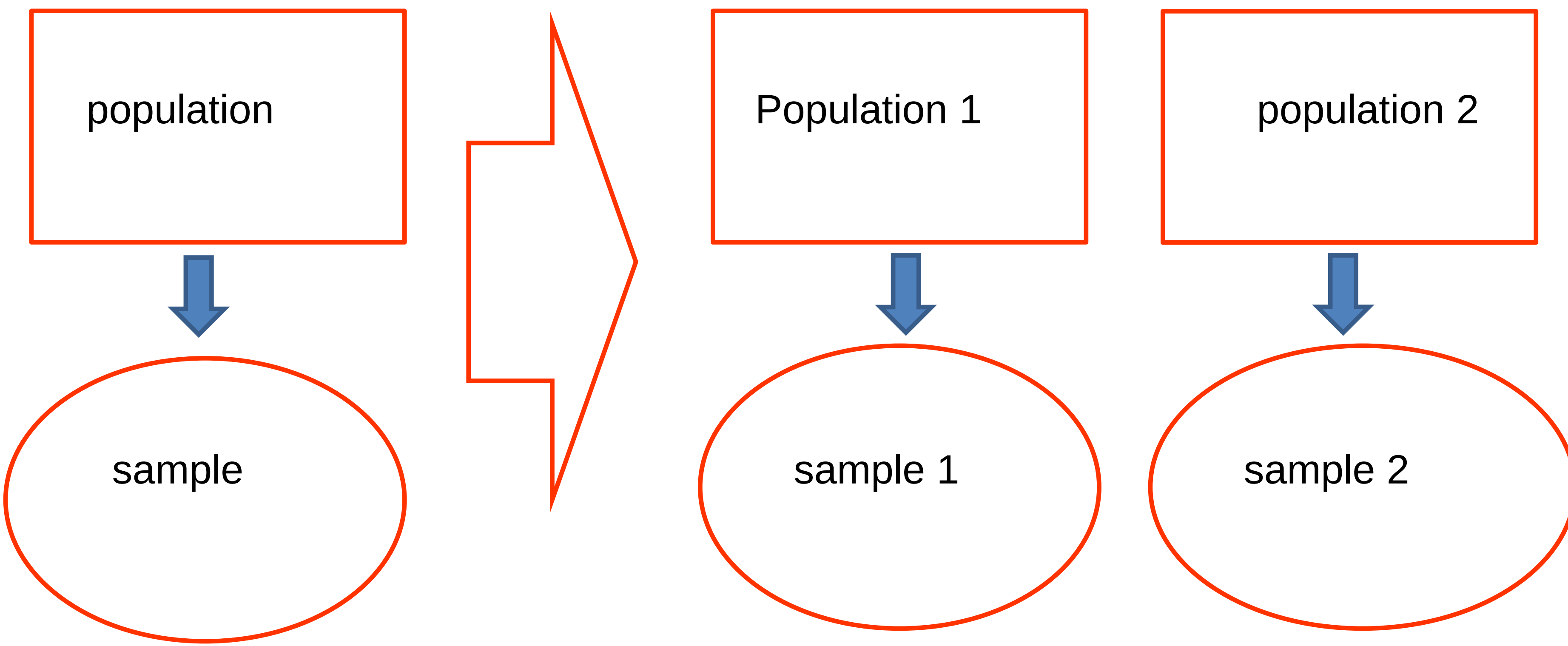
Two measurements per individual

Section 5.1 → Section 5.3

Section 5.1

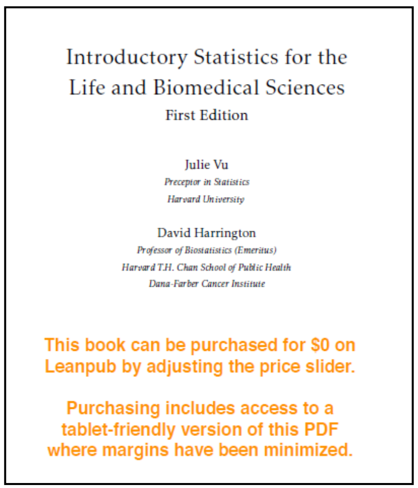
Section 5.3

Twee populaties



Two independent samples

Two-samples test for paired data


Introductory Statistics for the
Life and Biomedical Sciences
First Edition

Julie Vu
Preceptor in Statistics
Harvard University

David Harrington
Professor of Biostatistics (Emeritus)
Harvard T.H. Chan School of Public Health
Dana-Farber Cancer Institute

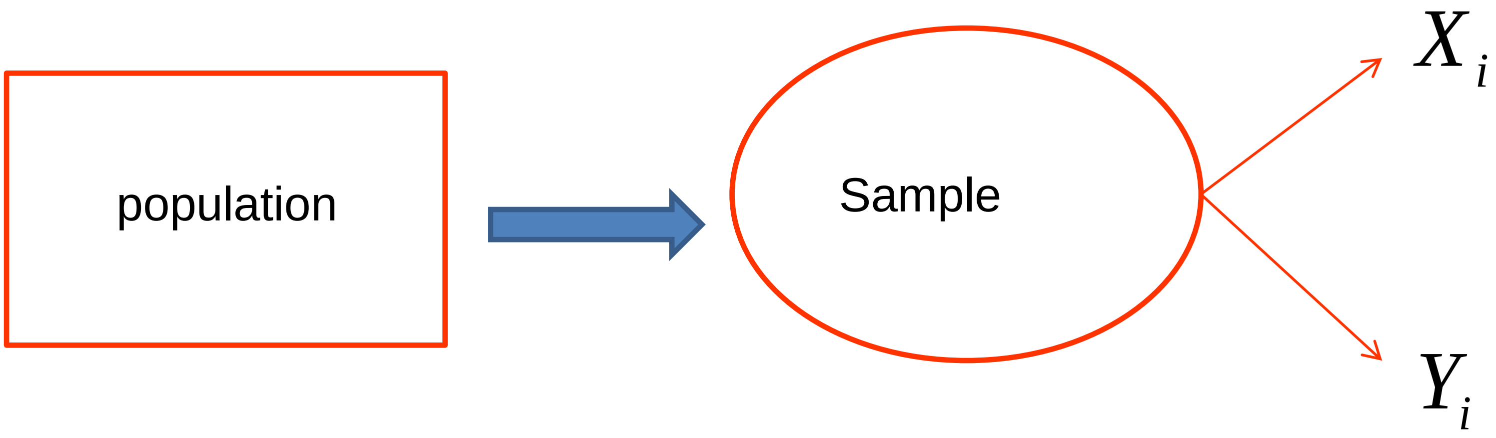
This book can be purchased for $0 on Leanpub by adjusting the price slider.

Purchasing includes access to a tablet-friendly version of this PDF where margins have been minimized.

Section 5.2

Paired measurements

- Paired measurements: two surveys done (same characteristic) per individual.
- Usually happen that two measurements with a certain interval.
- The question then asked is: is there a difference between the two measurements?



Example 1: Paired measurements

- 10 women participating in a clinical trial.
- We have two measurements of systolic blood pressure of 10 women: before and during treatment by hormone therapy.
- The question of the researcher: <u>is there a difference between the systolic blood pressure before and during treatment?</u>

The data

systolic blood pressure

	before X_i	during Y_i
1	115	128
2	112	115
3	107	106
4	119	128
5	115	122
6	138	145
7	126	132
8	108	109
9	104	102
10	115	117

- The average of the population before treatment: μ_1

$$E(X_i) = \mu_1 \qquad Var(X_i) = \sigma_1^2$$

- The average of the population during the treatment: μ_2.

$$E(Y_i) = \mu_2 \qquad Var(Y_i) = \sigma_2^2$$

Solution

- The null hypothesis: there is no difference between the systolic blood pressure before and during treatment

$$H_0 : \mu_2 = \mu_1$$

- We can also write this as H_0

$$H_0 : \mu_2 - \mu_1 = 0$$

during

before

The difference (after - before)

- We are interested in the difference of two means.
- We have each woman associate number (the difference between the measurements)

$$D_i = Y_i - X_i$$

$$D_i = SBP : during - SBP : before$$

The difference

- The population average of the difference between the first measurement X and the second measurement Y is equal to:

$$E(D_i) = E(Y_i) - E(X_i) = \mu_D$$

The test hypothesis

- Define: $\mu_2 - \mu_1 = \mu_{D:}$

$$H_0 : \mu_D = 0$$ Null hypothesis

$$H_a : \mu_D \neq 0$$ Alternative hypothesis

Two sided test of hypothesis

Alternatieve hypothese

$$\left.\begin{array}{ll}(a) & H_a : \mu_D > 0 \\ (b) & H_a : \mu_D < 0\end{array}\right\} \quad \text{One sided}$$

$$(c) \qquad H_a : \mu_D \neq 0 \qquad \text{Two sided}$$

Solution

- We assume that the differences are normally distributed.
- The sample size (n = 10) is small and σ^2 is not known.

$$\left.\begin{aligned} & D_i = Y_i - X_i \sim N(\mu_D, \sigma_D^2) \\ & n : small \\ & \sigma_D^2 : unknown \end{aligned}\right\}$$

Case 3

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D \neq 0$$

t test for a population

The test statistic

- The sample :

$$D_1, D_2, \ldots\ldots D_{10} \qquad\qquad \overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

- The distribution of the test statistic under H_0 :

$$\boxed{\frac{\overline{D} - \mu_D}{\sqrt{\frac{S_D^2}{n}}} \sim t_{(n-1)}} \xrightarrow{\text{under } H_0} \boxed{\frac{\overline{D} - 0}{\sqrt{\frac{S_D^2}{n}}} \sim t_{(n-1)}}$$

The rejection region (1)
(Two-sided test)

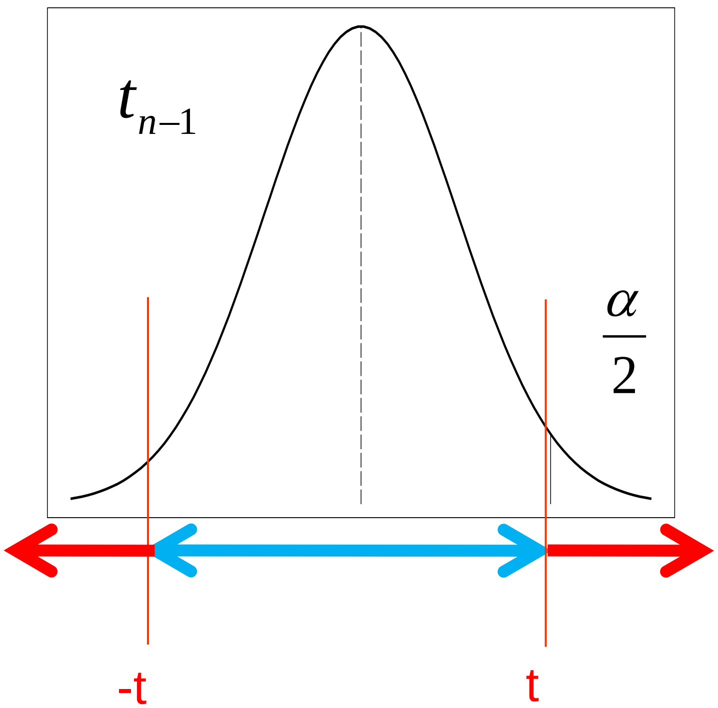
$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D \neq 0$$

- We reject the null hypothesis if the value of $\bar{d}$ is large or small

$$\left.\begin{array}{l} \bar{d} > c_2 \\ \bar{d} < c_1 \end{array}\right\} \text{we reject the null hypothesis}$$

The rejection region (2)
(Two-sided test)

For two-sided test problem



$$]-\infty, -t_{n-1,1-\frac{\alpha}{2}}] \cup [t_{n-1,1-\frac{\alpha}{2}}, \infty[$$

$$P(H_0 \; reject \; while \; it \; is \; true) = \alpha$$

$$P(T \leq -t) = \frac{\alpha}{2} \qquad en \qquad P(T \geq t) = \frac{\alpha}{2}$$

$$\left.\begin{array}{l} \dfrac{\bar{d}-0}{\sqrt{\dfrac{s_D^2}{n}}} > t \\[2em] \dfrac{\bar{d}-0}{\sqrt{\dfrac{s_D^2}{n}}} < -t \end{array}\right\}$$

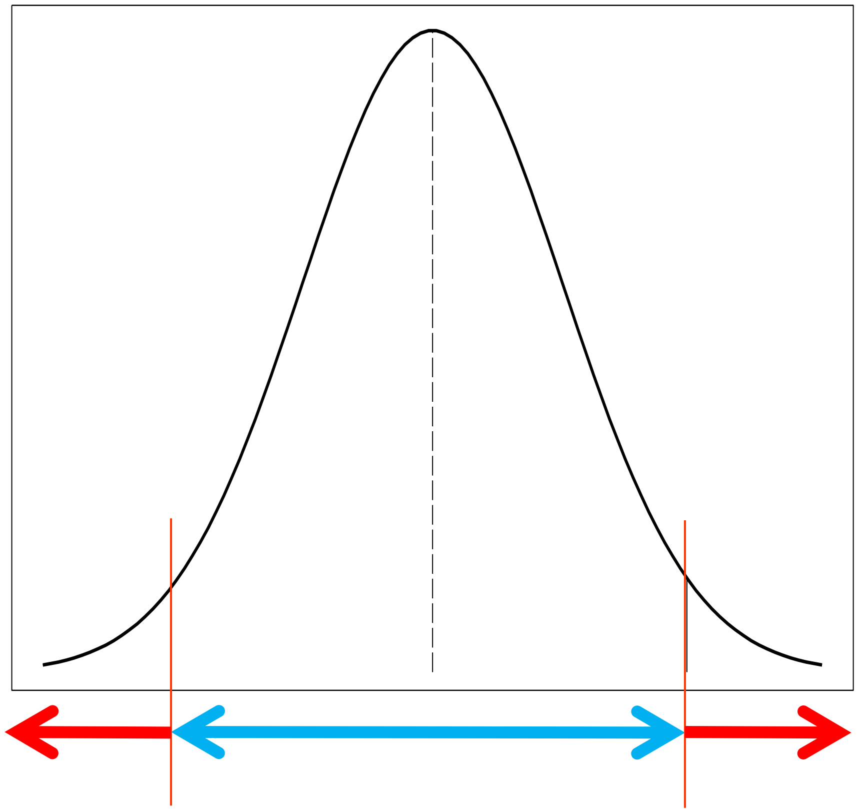
we reject the null hypothesis

The rejection region

For significance level of 5% (and 9 degrees of freedom)

$$T = \frac{\overline{D} - d}{\sqrt{\frac{S_D^2}{9}}} \sim t_{(9)}$$

$$]-\infty, -2.262] \cup [2.262, \infty[$$



-t=-2.262 t=2.262

Decision

difference = systolic blood pressure during a treatment - systolic blood pressure before a treatment

$$\frac{\bar{d}-0}{\sqrt{\frac{s_D^2}{10}}} = \frac{-4.5}{\sqrt{\frac{22.27778}{10}}} = -3.0149 < -2.262$$

$$\boxed{d_i = y_i - x_i}$$

$$\bar{d} = -4.8$$

$$s_D^2 = 20.8444$$

We reject the null hypothesis at 5% significance level.

Test a difference between paired measurement using R

```
> Before<-c(115, 112, 107, 119, 115, 138, 126, 108, 104, 115)
> During<-c(128, 115, 106, 128, 122, 145, 132, 109, 102, 117)
> library(MASS)
> t.test(Before, During, paired=TRUE)

Paired t-test

data:  Before and During
t = -3.0149, df = 9, p-value = 0.0146
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -7.876437 -1.123563
sample estimates:
mean of the differences
                   -4.5
```

The checklist

Step	Information	Example
1	The hypotheses (the testing problem)	$H_0 : \mu_D = 0$ $H_1 : \mu_D \neq 0$ two-tailed test
2	Determine the Case	$D_i \sim N(0, \sigma_D^2)$ σ^2 not known $n = 10 < 30$
3	The test statistic The distribution of the test statistic under the null hypothesis	$\boxed{\dfrac{\overline{D} - 0}{\sqrt{\dfrac{S^2}{10}}} \sim t(9)}$
4	The level of significance	$\alpha = 0.05$
5	The critical point (or points)	-2.262 & 2.262 t(9) two-tailed test
6	Calculate the test statistic	-3.322
7	The rejection region & Conclusion	Reject the null hypothesis

Notes

- We can also test for

$$H_0 : \mu_D = d$$ Null hypothese

$$H_1 : \mu_D \neq d$$ Alternatieve hypothese

Where d is a specified number (not necessarily 0).

Confidence interval

- It is also possible to give a confidence interval for the mean difference μ_D.

- A 95% confidence interval for μ_D

$$\boxed{\left[\overline{D} - a \times \sqrt{\frac{S_D^2}{n}}, \overline{D} + a \times \sqrt{\frac{S_D^2}{n}}\right]}$$

$$a = t_{n-1,1-\frac{\alpha}{2}}$$

Confidence interval

- A 95% confidence interval for μ_D

$$\overline{D} = -4.5, n = 10, t_{9,0.975} = 2.262, S_D^2 = 22.27778$$

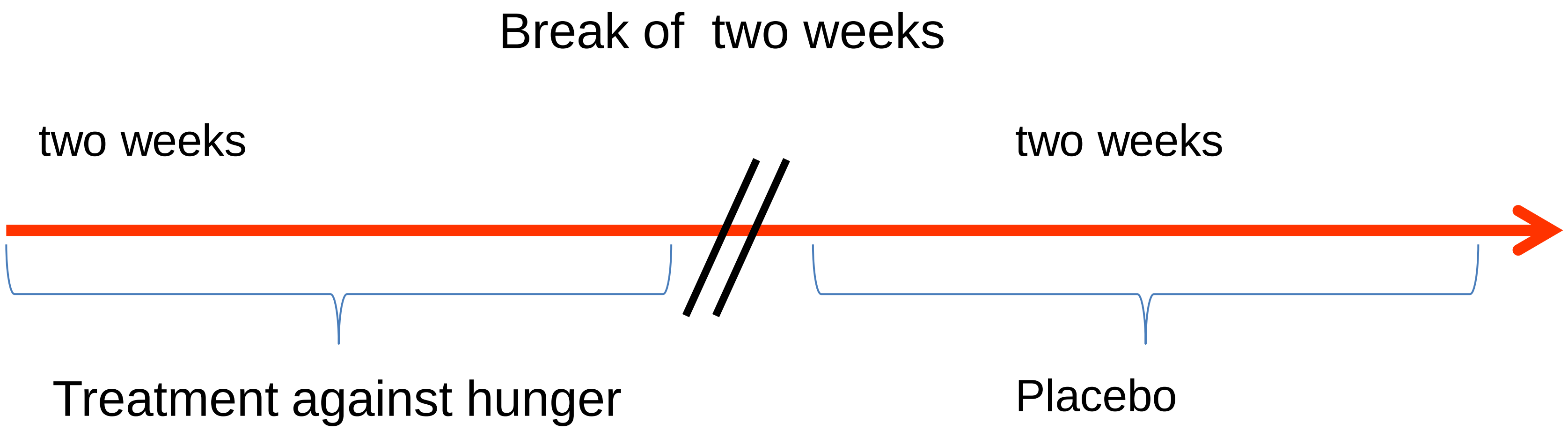
$$\left[\underbrace{-4.5 - 2.262 \times \sqrt{\frac{22.2778}{10}}}, \underbrace{-4.5 + 2.262 \times \sqrt{\frac{22.27778}{10}}}\right]$$

[-7.876, -1.124]

Example 2
One sided test problem

- A study on treatment against hunger.
- A group of 9 people have medication hungry for two weeks.
- During this period, record the people's hunger-score in a scale of 1 to 150 (1 = not hungry, 150 = very hungry).
- After the treatment period, people have a break of two weeks without medication.
- After two weeks, the men got a placebo for two weeks.
- During the placebo period, scoring the people their hunger score.

Example 2



First measurement: hunger score under treatment

Second measurement: hunger score under placebo

Notation

X_i hunger score under placebo

Y_i hunger score under treatment

$\Bigg\}$ Two engineering methods for each individual

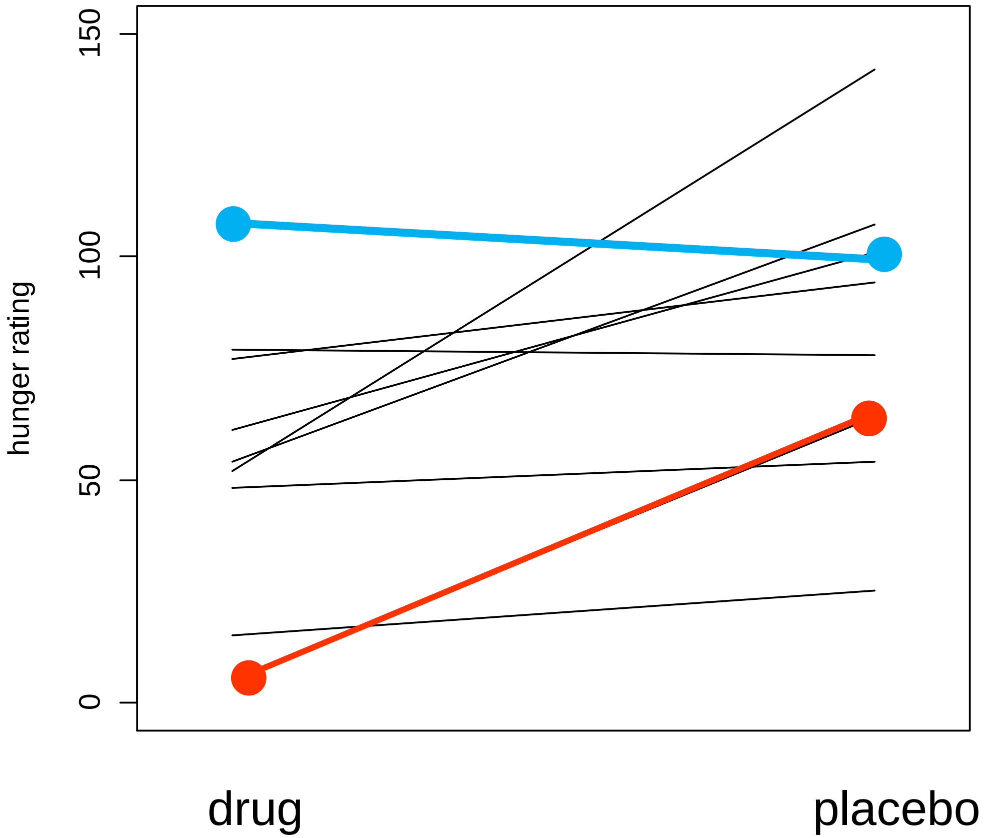
$$D_i = Y_i - X_i$$

$$X_i \sim N\left(\mu_1, \sigma_1^2\right) \quad \& \quad Y_i \sim N\left(\mu_2, \sigma_2^2\right)$$

The data (1)

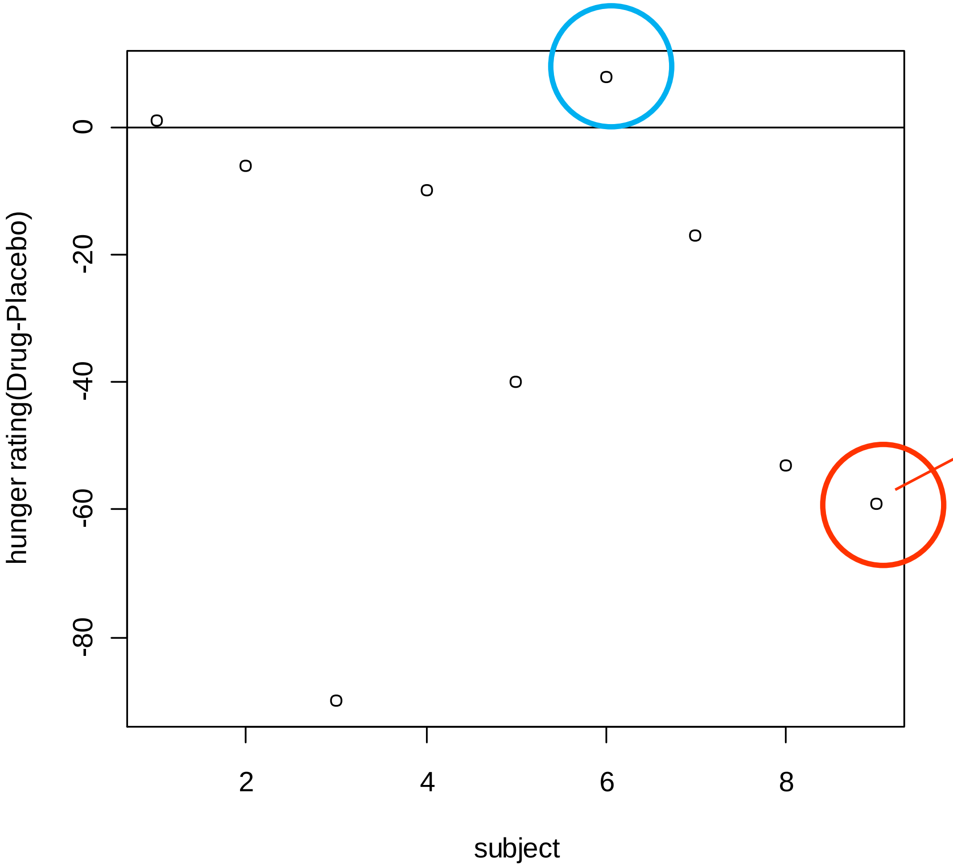
hunger score

	treatment	Placebo
1	79	78
2	48	54
3	52	142
4	15	25
5	61	101
6	107	99
7	77	94
8	54	107
9	5	64



The data (2)

$$D_6 = Y_6 - X_6 = 107 - 99 = 8$$



$$D_9 = Y_9 - X_9 = 5 - 64$$

The testing problem

If the drug works, we expect that the hunger score under treatment will be lower than the hunger score under placebo.

$$E(D_i) = E(Y_i) - E(X_i) = \mu_D$$

If the drug works, we expect

$$\mu_D < 0$$

One sided test problem

$$H_0 : \mu_D = 0$$
$$H_1 : \mu_D < 0$$

The distribution of the test statistic

- n = 9 (small)
- σ_1^2 and σ_2^2 unknown but equal
- populations are normally distributed

$$\boxed{\frac{\overline{D}-0}{\sqrt{\frac{S_D^2}{9}}} \sim t(8)}$$

We reject H_0 if:

$$\overline{D} < c \quad \Leftrightarrow \quad \frac{\overline{D}-0}{\sqrt{\frac{S_D^2}{9}}} < t_{(n-1,\alpha)}$$

$]-\infty,-t]$

The data

hunger score

	Treatment	Placebo
1	79	78
2	48	54
3	52	142
4	15	25
5	61	101
6	107	99
7	77	94
8	54	107
9	5	64

$$\bar{d} = -29.5556$$

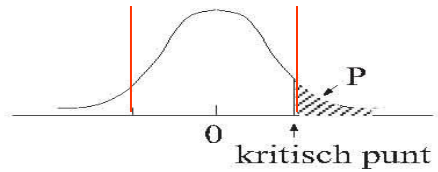
$$S_D^2 = 32.82^2$$

the test statistic

$$\frac{-29.5556 - 0}{\sqrt{\frac{32.82^2}{9}}} = -2.7014$$

The critical point in the table of Student t-distribution

Tabel 4 : Kritische punten student t verdeling



P / v.g.	.25	.10	.05	.025	.010	.005	.001
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	.816	1.886	2.920	4.303	6.965	9.925	22.326
3	.765	1.638	2.353	3.182	4.541	5.841	10.213
4	.741	1.533	2.132	2.776	3.747	4.604	7.173
5	.727	1.476	2.015	2.571	3.365	4.032	5.893
6	.718	1.440	1.943	2.447	3.143	3.707	5.208
7	.711	1.415	1.895	2.365	2.998	3.499	4.785
8	.706	1.397	1.860	2.306	2.896	3.355	4.501
9	.703	1.383	1.833	2.262	2.821	3.250	4.297
10	.700	1.372	1.812	2.228	2.764	3.169	4.144
11	.697	1.363	1.796	2.201	2.718	3.106	4.025
12	.695	1.356	1.782	2.179	2.681	3.055	3.930
13	.694	1.350	1.771	2.160	2.650	3.012	3.852
14	.692	1.345	1.761	2.145	2.624	2.977	3.787
15	.691	1.341	1.753	2.131	2.602	2.947	3.733
16	.690	1.337	1.746	2.120	2.583	2.921	3.686
17	.689	1.333	1.740	2.110	2.567	2.898	3.646
18	.688	1.330	1.734	2.101	2.552	2.878	3.610
19	.688	1.328	1.729	2.093	2.539	2.861	3.579
20	.687	1.325	1.725	2.086	2.528	2.845	3.552
21	.686	1.323	1.721	2.080	2.518	2.831	3.527
22	.686	1.321	1.717	2.074	2.508	2.819	3.505
23	.685	1.319	1.714	2.069	2.500	2.807	3.485
24	.685	1.318	1.711	2.064	2.492	2.797	3.467
25	.684	1.316	1.708	2.060	2.485	2.787	3.450
26	.684	1.315	1.706	2.056	2.479	2.779	3.435
27	.684	1.314	1.703	2.052	2.473	2.771	3.421
28	.683	1.313	1.701	2.048	2.467	2.763	3.408
29	.683	1.311	1.699	2.045	2.462	2.756	3.396
30	.683	1.310	1.697	2.042	2.457	2.750	3.385
40	.681	1.303	1.684	2.021	2.423	2.704	3.307
60	.679	1.296	1.671	2.000	2.390	2.660	3.232
120	.677	1.289	1.658	1.980	2.358	2.617	3.160
∞	.674	1.282	1.645	1.960	2.326	2.576	3.090

$$P(T > 1.86) = 0.05$$

$$P(T < -1.86) = 0.05$$

$$\frac{-30-0}{\sqrt{\frac{33^2}{9}}} = -2.72 < -1.86$$

⇨ we reject H_0

Test a difference between paired measurement using R

```
> placebo <-c(78, 54, 142, 25, 101, 99, 94, 107, 64)
> treatment <-c(79, 48, 52, 15, 61, 107, 77, 54, 5)
> library(MASS)
> t.test(treatment, placebo, paired=TRUE)

Paired t-test

data:  treatment and placebo
t = -2.7014, df = 8, p-value = 0.02701
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -54.784709  -4.326402
sample estimates:
mean of the differences
              -29.55556
```

In the R code: a two-sided test !!!

$$H_0: \mu_D = 0$$
$$H_1: \mu_D \neq 0$$

The checklist

Step	Information	Example
1	The hypotheses (the qualifying problem)	$H_0 : \mu_D = 0$ $H_1 : \mu_D < 0$ One - sided key
2	Detarmine the case	$D_i \sim N(0, \sigma_D^2)$ σ^2 not known $n = 9 < 30$
3	The test statistic The distribution of the test statistic under the null hypothesis	$\boxed{\dfrac{\overline{D} - 0}{\sqrt{\dfrac{S^2}{10}}} \sim t(8)}$
4	The level of significance	$\alpha = 0.05$
5	The critical point (or points)	-1.86 t(8)
6	Calculate the test statistic	-2.7014
7	Conclusion	Reject the null hypothsis

Example: the sleep data

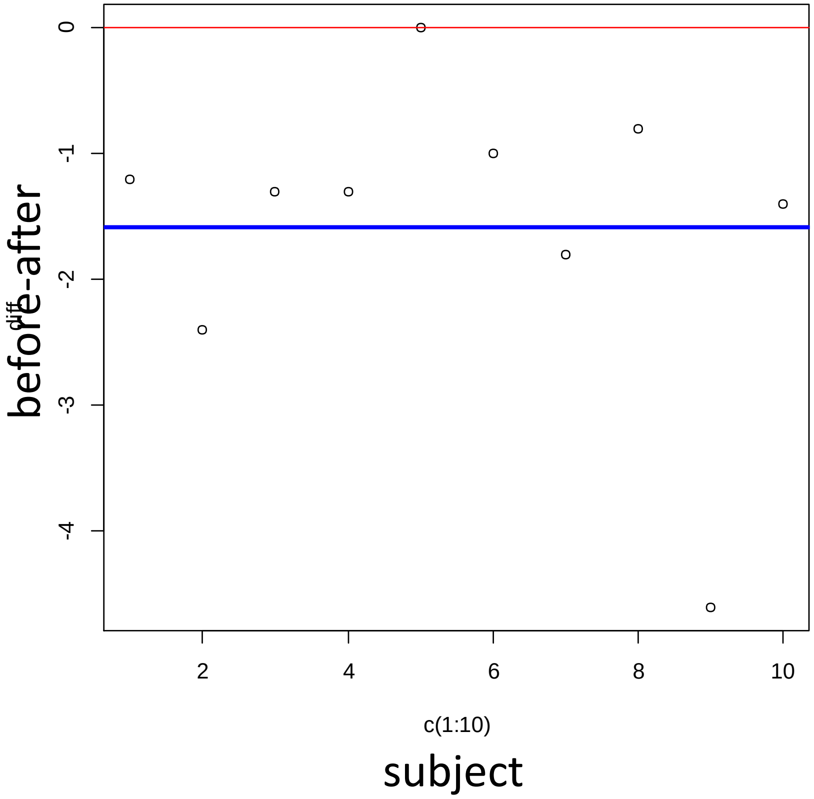
- Data which show the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients.
- For each patient: data before and after the treatment.
- Response: increase hours in sleep before and after the treatment.
- Research question: does the treatment increase the sleeping hours ?
- More about the data: use `help(sleep)` in R.

The sleep data: before - after

```
> before<-sleep$extra[sleep$group == 1]
> after<-sleep$extra[sleep$group == 2]
> diff<-before-after
> plot(c(1:10),diff)
> abline(0,0,col=2)

> mean(before)
[1] 0.75
> mean(after)
[1] 2.33

> mean(before)-mean(after)
[1] -1.58
```



$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n} D_i$$

The hypothesis test

- The null hypothesis: increase of sleeping hours in the same before and after the treatment.

$$\left.\begin{array}{l} H_0 : \mu_D = 0 \\ H_a : \mu_D \neq 0 \end{array}\right\} \text{Two sided test of hypothsis}$$

The sleep data: paired t test in R

```
> totes(before,after, paired = TRUE)

    Paired t-test

data:  before and after
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -2.4598858 -0.7001142
sample estimates:
mean of the differences
                  -1.58
```

$$H_0 : \mu_D = 0$$
$$H_a : \mu_D \neq 0$$

The C.I does not cover the value of zero.

Two sample test for independent data

Introductory Statistics for the
Life and Biomedical Sciences
First Edition

Julie Vu
Preceptor in Statistics
Harvard University

David Harrington
Professor of Biostatistics (Emeritus)
Harvard T.H. Chan School of Public Health
Dana-Farber Cancer Institute

This book can be purchased for $0 on
Leanpub by adjusting the price slider.

Purchasing includes access to a
tablet-friendly version of this PDF
where margins have been minimized.

Section 5.3

Two populations and two independent samples

Population 1

$$\mu_1, \sigma_1^2$$

Population 2

$$\mu_2, \sigma_2^2$$

we draw two samples independently

sample 1

Sample 2

$$X_1, X_2, \ldots, X_{n_1}$$

$$Y_1, Y_2, \ldots, Y_{n_2}$$

we draw two samples independently

We're back interested in the difference between the two averages μ_1 and μ_2 and set as nulhypothes

$$H_0 : \mu_2 - \mu_1 = (\mu_2 - \mu_1)_{H_0}$$

If the two populations is no difference in mean
Then $(\mu_2-\mu_1)_{H0}=0$

$$H_0 : \mu_2 - \mu_1 = 0$$

The sample means

- The quantity that we get to test the hypothesis, initially from our samples will be as expected:

$$\overline{Y} - \overline{X}$$

(the difference of the sample means)

The sample means

$$E(\bar{X}) = \mu_1$$

$$E(\bar{Y}) = \mu_2$$

$$Var(\bar{X}) = \frac{\sigma_1^{\ 2}}{n_1}$$

$$Var(\bar{Y}) = \frac{\sigma_2^{\ 2}}{n_2}$$

The distribution of the difference

$$E(\overline{Y} - \overline{X}) = E(\overline{Y}) - E(\overline{X}) = \mu_2 - \mu_1$$

$$\underbrace{Var(\overline{X} - \overline{Y}) = \frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}$$

For two independent samples

$$\overline{Y} - \overline{X} \sim ?$$

Case 1:

1. σ_1^2 and σ_2^2 known
2. Both populations are normally distributed.
3. $\sigma_1^2 = \sigma_2^2$

$$Var(\bar{Y} - \bar{X}) = \frac{\sigma_2^{\ 2}}{n_2} + \frac{\sigma_1^{\ 2}}{n_1} = \frac{\sigma^{\ 2}}{n_2} + \frac{\sigma^{\ 2}}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

$$\boxed{\frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)_{H_0}}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)}$$

Case 2: (two-sample t-test)

1. n_1 or n_2 small.
2. both populations are normally distributed.
3. σ_1^2 and σ_2^2 unnown but $\sigma_1^2 = \sigma_2^2$

$$\boxed{\frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)_{H_0}}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)}}$$

$$\boxed{S_p^{\ 2} = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}\left(X_i - \bar{X}\right)^2 + \sum_{i=1}^{n_2}\left(Y_i - \bar{Y}\right)^2\right] = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)S_1^{\ 2} + (n_2 - 1)S_2^{\ 2}\right]}$$

pooled sample variance

Case 3:

1. n_1 and n_2 large (>30).
2. the populations are not normally distributed.
3. σ_1^2 and σ_2^2 are known ($\sigma_1^{\ 2} \neq \sigma_2^{\ 2}$).

$$\boxed{\frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)_{H_0}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)}$$

Case 4a:

1. n_1 and n_2 large (>30).
2. the populations are not normally distributed.
3. σ_1^2 and σ_2^2 are not known $\quad \sigma_1^2 \neq \sigma_2^2$

$$\boxed{\frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)_{H_0}}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0,1)}$$

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2 \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

Situatie 4b:

1. n_1 and n_2 large (>30).
2. the populations are not normally distributed
3. 2. σ_1^2 and σ_2^2 are not unknown $\sigma_1^{\,2} = \sigma_2^{\,2}$

$$\boxed{\frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)_{H_0}}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)}$$

$$\boxed{S_p^{\,2} = \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}\left(X_i - \bar{X}\right)^2 + \sum_{i=1}^{n_2}\left(Y_i - \bar{Y}\right)^2\right] = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)S_1^{\,2} + (n_2 - 1)S_2^{\,2}\right]}$$

pooled sample variance

Example 3

One sided test problem

When an experiment is compared to the results of two treatments, A and B.
Treatment A was applied to a group of 6 randomly selected animals and treatment B in a group of 5 randomly selected animals.
The results were:

A	17,19,15,18,21,18
B	18,15,13,16,13

Here: n_1=6 and n_2=5.

The testing problem

- The researcher claims that treatment A better average results than treatment B.
- The average treatment A is greater than the average of treatment B (which type of test is this?)
- We assume that both populations are normally distributed and the same variance.

The testing problem

The results of treatment A, and this is what we call X_i

treatment of B, Y_i .

$$E(X_i) = \mu_1 \qquad\qquad E(Y_i) = \mu_2$$

We formulate the null and alternative hypothesis :

$$H_0 : \mu_2 - \mu_1 = 0$$
$$H_1 : \mu_2 - \mu_1 < 0$$

one sided test problem

Two independent samples

– $n_1 = 6$ and $n_2 = 5$ (small)
– σ_1^2 and σ_2^2 unknown but equal
– populations are normally distributed

case 2:

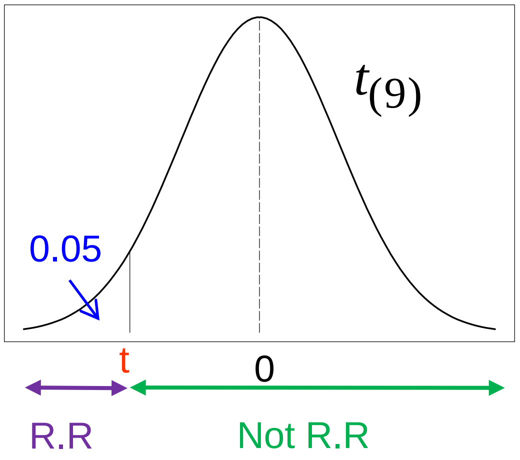
$$\Rightarrow \quad \frac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)_{H_0}}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\bar{Y} - \bar{X} - 0}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{(n_1+n_2-2)} = t_{(9)}$$

The rejection region

$$H_0 : \mu_2 - \mu_1 = 0$$
$$H_1 : \mu_2 - \mu_1 < 0$$

if $\dfrac{\bar{Y} - \bar{X} - 0}{\sqrt{S_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} < t$ We reject the null hypothesis

$H_1 \Rightarrow$ one sided hypothesis, Significance level = 0.05, Critical point from the t-distribution table: t = -1.833



Solution: the sample mean and variance

$$\bar{x} = 18 \qquad s_1^2 = \frac{20}{5} = 4$$

$$\bar{y} = 15 \qquad s_2^2 = \frac{18}{4} = 4.5$$

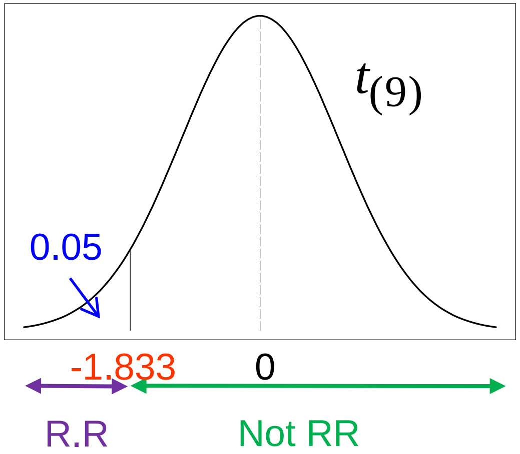
$$s_p^2 = \frac{1}{9}\left[5 \times s_1^2 + 4 \times s_2^2\right] = 4.22$$

pooled sample variance

$$\boxed{S_p{}^2 = \frac{1}{n_1 + n_2 - 2}\left[(n_1 - 1)S_1{}^2 + (n_2 - 1)S_2{}^2\right]}$$

Two independent samples

$$t = \frac{\bar{y} - \bar{x}}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{15 - 18}{\sqrt{4.22\left(\frac{1}{6} + \frac{1}{5}\right)}} = -2.41 < -1.833$$



$$\Rightarrow H_0 \text{ rejected at 5\% significance level}$$

Based on the sample results, we conclude that the average treatment A better results than treatment B (at 5%significance level).

One sided test for two independent sample using R

```
> A <- c(17, 19, 15, 18, 21, 18)
> B <- c(18, 15, 13, 16, 13)
> library(MASS)
>  t.test(B, A,var.equal=T, alternative="less")

Two Sample t-test
```

$$H_0 : \mu_2 - \mu_1 = 0$$
$$H_1 : \mu_2 - \mu_1 < 0$$

A one sided test.

```
data:  B and A
t = -2.4111, df = 9, p-value = 0.01959
alternative hypothesis: true difference in means is less than
0
95 percent confidence interval:
      -Inf -0.7191565
sample estimates:
mean of x mean of y
       15        18
```

The checklist

Step	information	Example
1	The Hypothesis test	$H_0: \mu_2 - \mu_1 = 0$ $H_1: \mu_2 - \mu_1 < 0$ One-sided test
2	Determine the case ➢ Case 2	$X_i \sim N(\mu_1, \sigma_1^2)$ σ² unknown $Y_i \sim N(\mu_2, \sigma_2^2)$ $n_1 = 6 < 30$ $n_2 = 5 < 30$
3	The test statistic The distribution of the test statistic under the null hypothesis	$\dfrac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)_{H_0}}{\sqrt{S_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim t(9)$
4	The level of significance	$\alpha = 0.05$
5	The critical point (or points) & R.R	-1.833 t(9)
6	Calculate the test statistic	-2.41
7	Conclusion	Reject the null hypothesis

Remark

- Also here one can calculate a confidence interval for the difference between the population means.

- A 95% confidence interval for the difference $\mu_2 - \mu_1$ from the example given by

$$\boxed{\left[\bar{Y} - \bar{X} - a \times \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \bar{Y} - \bar{X} + a \times \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right]}$$

$$a = t_{n_1 + n_2 - 2, 1 - \frac{\alpha}{2}}$$

Confidence interval

$$\boxed{\left[15-18-2.262\times\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)},15-18+2.262\times\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\right]}$$

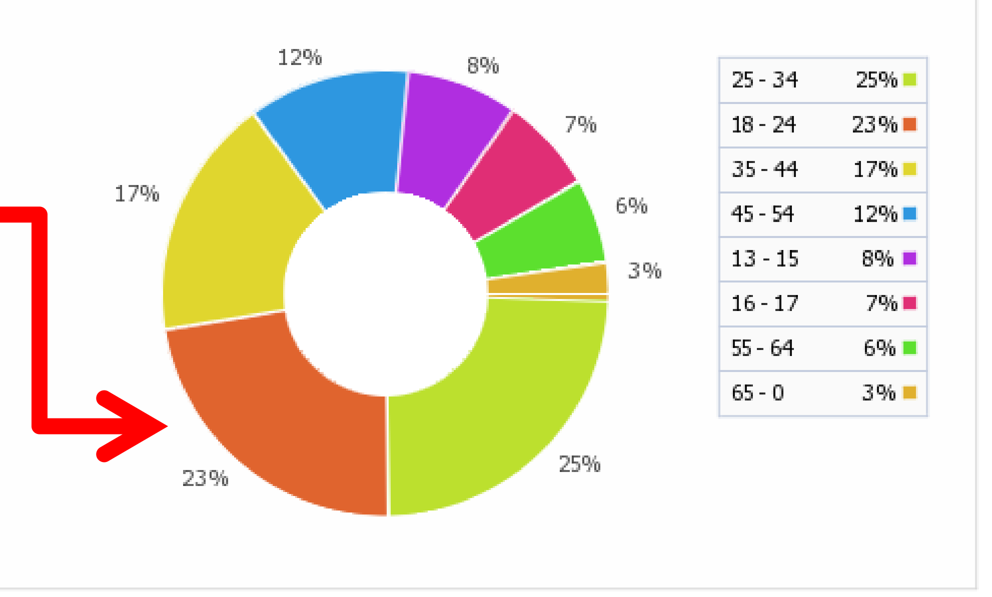
$$\boxed{\left[15-18-2.262\times 1.244,15-18+2.262\times 1.244\right]}$$

[-5,8,-0.19]

Example 4: number FACEBOOK friends
Two sided test problem

According FACEBOOK statistics, 23% of the population aged 18-24 in Belgium have a FACEBOOK account (2011).

The researcher would like to know how-many FACEBOOK friends, men and women in this age group.



http://www.socialbakers.com/facebook-statistics/belgium

Multiplying the FACEBOOK friends

- A researcher wants the number of facebook friends male and female patients, aged 18 to 24 compared.
- The researcher assumes that in this age, there is no difference between the number of men and women friends in facebook.

X_i number of facebook friends for a woman $$E(X_i) = \mu_1, Var(X_i) = \sigma_1^2$$

Y_i number of facebook friends for a man $$E(Y_i) = \mu_2, Var(Y_i) = \sigma_2^2$$

Information on the population and sample

– n_1 = 35 and n_2 = 40 (>30)
– σ_1^2 and σ_2^2 unknown but equal
– distribution of the population: not known

Case 4b:

$$\boxed{\frac{\bar{Y}-\bar{X}-(\mu_2-\mu_1)_{H_0}}{\sqrt{S_P^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim N(0,1)}$$

The testing problem

We formulate the null and alternative hypothesis :

$$\left.\begin{aligned} H_0 &: \mu_2 - \mu_1 = 0 \\ H_1 &: \mu_2 - \mu_1 \neq 0 \end{aligned}\right\}$$

Under H_0, in this age, there is no difference between the number of face book friends of men and women.

The sample

- The researcher draws a sample of 35 men and 40 women in the age group 18-24.

$$\bar{x}_M = 107.8$$

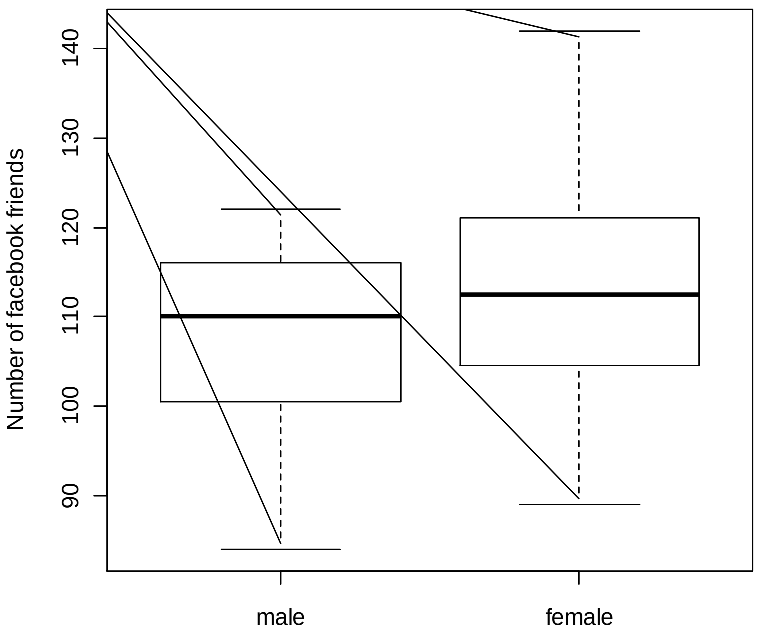
$$S_M^2 = 102.635$$

$$\mathrm{n}_{\mathrm{M}} = 35$$

$$\bar{x}_W = 112.575$$

$$S_W^2 = 147.019$$

$$n_W = 40$$



The rejection region (two sided test)

$$\boxed{\frac{\bar{y}-\bar{x}-0}{\sqrt{s_P^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = -1.8354}$$

The test statistic

-1.8354> -1.96 we can not reject H_0 at significance level 0.05.

The rejection region (for significance level of 10%)

-1.8354 < -1.645 ⇒ we reject the null hypothesis at significance level 0.1.

The checklist

Step	informatie	Example
1	Test of Hypothesis	$H_0: \mu_2 - \mu_1 = 0$ $H_1: \mu_2 - \mu_1 \neq 0$ **Two sided test problem**
2	Determine case	$X_i \sim unknown$ $Y_i \sim unknown$ $n_1 = 35 > 30$ $n_2 = 40 > 30$ σ_1^2 and σ_2^2 are unknown
3	The test statistic The distribution of the test statistic under the null hypothesis	$\dfrac{\bar{Y} - \bar{X} - (\mu_2 - \mu_1)_{H_0}}{\sqrt{S_p^2\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} \sim N(0,1)$
4	The level of significance	$\alpha = 0.05$
5	The critical point (or points) & R.R	-1.96 & 1.96 N(0,1)
6	Calculate the test statistic	-1.8354
7	Conclusion	Do not reject at 5% level of significance

Example: chicken weights by feed type

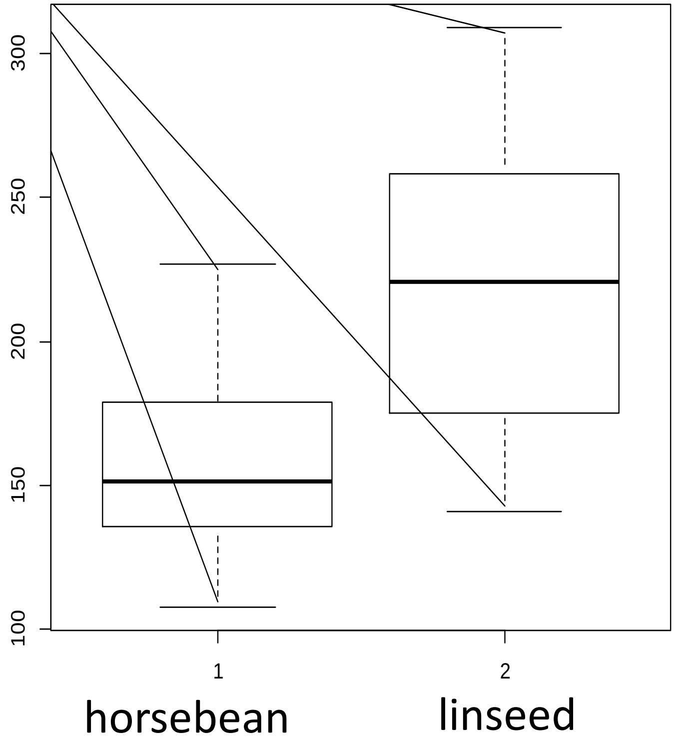
- An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens.
- Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement.
- Their weights in grams after six weeks are given along with feed types.

Example: chicken weights by feed type

- Main interest: the weight of two feed supplements groups: horsebean & linseed.
- Research question: does the feed type (horsebean or linseed) influence the chick weight ?

Example: Chicken weights by feed type

```
> x<-chickwts$weight[chickwts$feed=="horsebean"]
> y<-chickwts$weight[chickwts$feed=="linseed"]
> mean(x)
[1] 160.2
> mean(y)
[1] 218.75
> boxplot(x,y)
```



The testing problem

We formulate the null and alternative hypothesis :

$$H_0 : \mu_2 - \mu_1 = 0$$
$$H_1 : \mu_2 - \mu_1 \neq 0$$

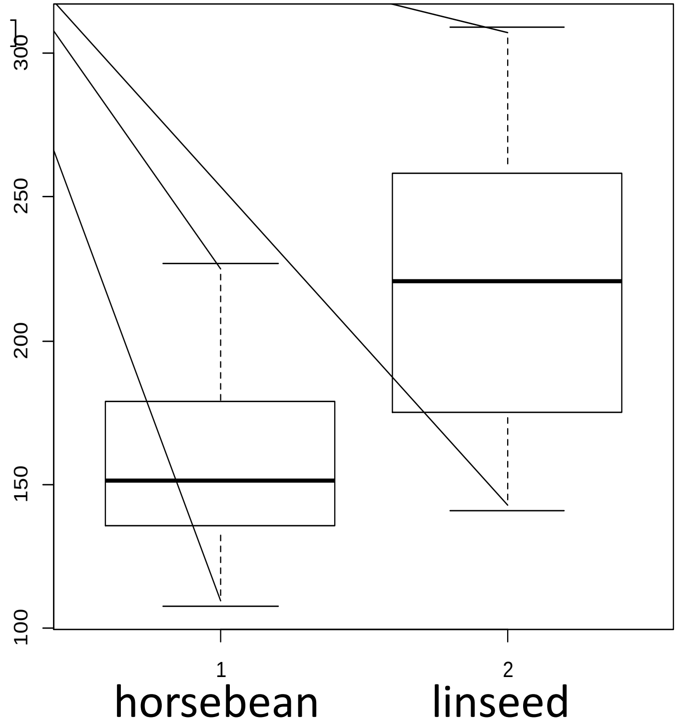
Under H_0, there is no difference between the chicks' weight under the two diets.

The population mean of the linseed feed

The population mean of the horsebean feed

Example: Chicken weights by feed type

```
> x<-chickwts$weight[chickwts$feed=="horsebean"]
> y<-chickwts$weight[chickwts$feed=="linseed"]
> mean(x)
[1] 160.2
> mean(y)
[1] 218.75
> boxplot(x,y)
```



Chicken weights by feed type: two sample test for independent data in R

```
> t.test(y,x,var.equal = TRUE)

        Two Sample t-test

data:  y and x
t = 2.934, df = 20, p-value = 0.008205
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  16.92382 100.17618
sample estimates:
mean of x mean of y
   218.75    160.20
```

$$H_0 : \mu_2 - \mu_1 = 0$$

$$H_1 : \mu_2 - \mu_1 \neq 0$$