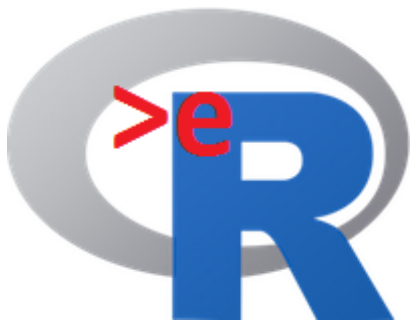




This course was developed as a part of the VLIR-UOS Cross-Cutting projects:

- Statistics: 2011-2016, 2017.
- Statistics: 2017.
- Statistics for development : 2018-2022.
- The >eR-BioStat ITP: 2024-2026.



The >eR-Biostat initiative  
Making R based education materials in  
statistics accessible for all

# Introduction to Statistical inference and estimation using R: Foundations of inference (one population)

Developed by  
Ziv Shkedy (Hasselt Univesrity, Belgium) and Tadesse Awoke  
(Gondar University, Ethiopia).

LAST UPDATE: 03/2024



ER-BioStat



<https://github.com/eR-Biostat>



@erbiostat

# Development team

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

The book is available for free  
online:

<https://www.openintro.org/book/biostat/>

## Chapter 4: Foundations for inference

# 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.

- In this part of the course, we cover mainly Chapter 4 and Section 5.1.
- The examples that are used for illustration **are not** the same as the examples in the book.

**Chapter 4 & 5.1: Foundations for inference**



# Software

- R functions: Hypothesis testing for one/two samples
  - `z.test()` .
  - `t.test()` .
- R program for the examples is available online.



# 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. Samples and populations.
2. Point estimators.
3. Variability of the sample mean.
4. Confidence intervals.
5. Hypotheses testing in one population.
6. Decision errors.
7. Hypotheses testing using  $t$  distribution.





# Part 1:

## Samples and populations



## 1.1: Notations and definitions

# Populations

- The population is the 'big picture' of which we are certain parameters want to know, but we have all sorts of reasons (practical, physical, financial, ...) can not fully observe.
- examples:
  - The birth lengths of all children as last year were recorded in all Flemish maternity hospitals.
  - The results of the parliamentary elections in Belgium if today would be voted.
  - The blood of the inhabitants of the European Union.

# Sample

- A Sample: is a portion of a population taken in such a way that it would represent the population that we would like to investigate.

# Populations and samples

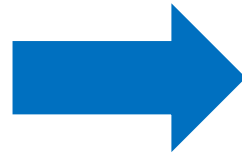
1. A population can be described by the random variable  $X$ .
2. A sample from that population of  $X$ :  $X_1, \dots, X_n$  have the same distribution as  $X$  and independent.

# Population and sample: notations

Population :  $X$

$$X_1, X_2, \dots, X_N$$

$$\mu, \sigma^2$$



sample

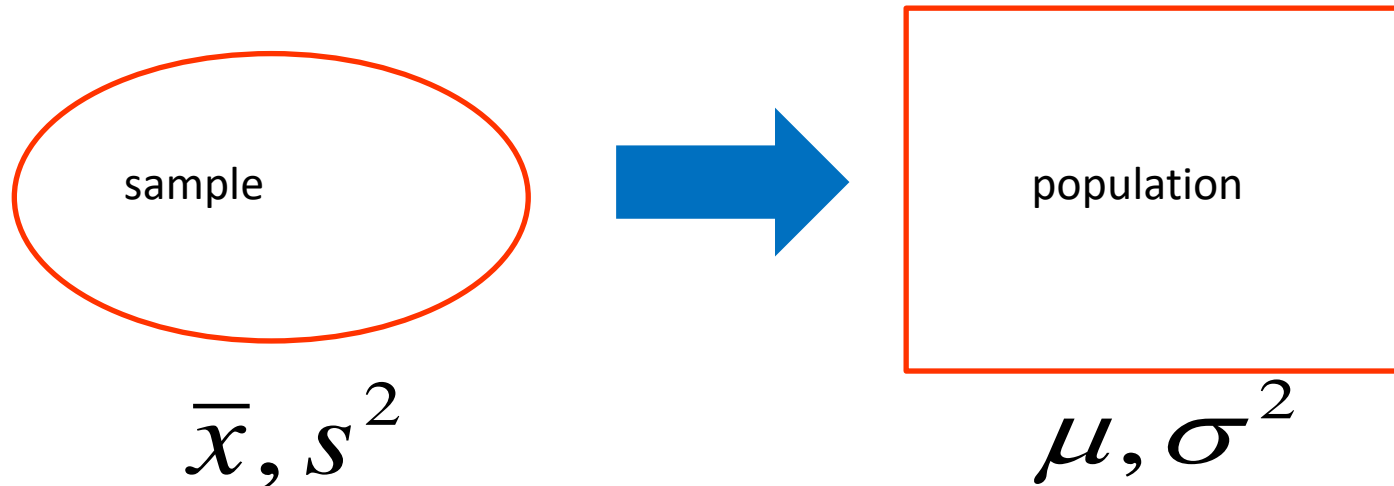
$$X_1, X_2, \dots, X_n$$

$X_i$  : Random variable from the population.

$\mu, \sigma^2$  : The unknown parameters.

# Population and sample: notations

- Based on a sample of size  $n$  from a population, we try to make a statement about the population.



# Example

- Population: ozone level in parts per billion from 1300 to 1500 hours at Roosevelt Island.

$$X_1, X_2, \dots, X_N$$

unknown parameters

$$E(X) = \mu = ?$$

$$Var(X) = \sigma^2 = ?$$

$X_i$  = ozone levels in parts per billion.



# Example: the airquality data in R

Daily air quality measurements in New York, May to September 1973.

```
> help(airquality)
```

```
> airquality$Ozone
```

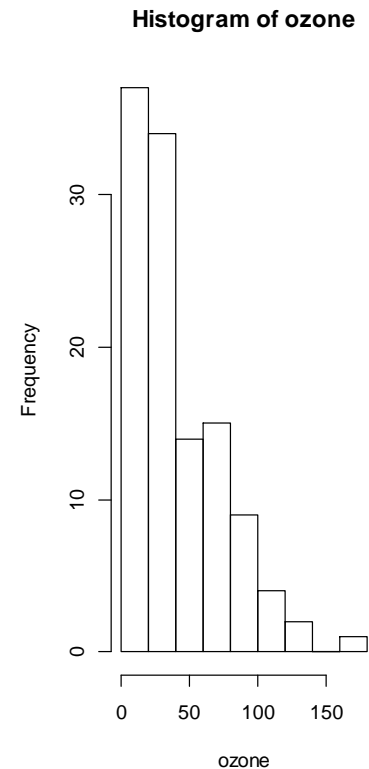
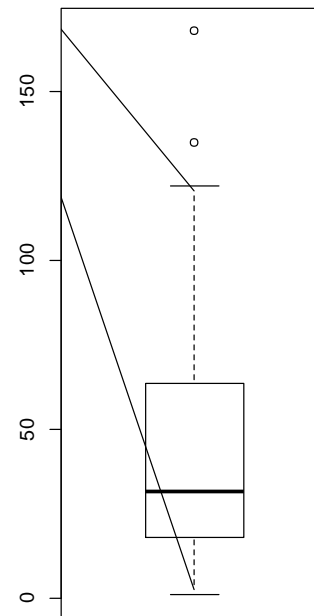
[1]	41	36	12	18	NA	28	23	19	8	NA	7	16	11	14	18	14	34	6
[19]	30	11	1	11	4	32	NA	NA	NA	23	45	115	37	NA	NA	NA	NA	NA
[37]	NA	29	NA	71	39	NA	NA	23	NA	NA	21	37	20	12	13	NA	NA	NA
[55]	NA	NA	NA	NA	NA	NA	NA	135	49	32	NA	64	40	77	97	97	85	NA
[73]	10	27	NA	7	48	35	61	79	63	16	NA	NA	80	108	20	52	82	50
[91]	64	59	39	9	16	78	35	66	122	89	110	NA	NA	44	28	65	NA	22
[109]	59	23	31	44	21	9	NA	45	168	73	NA	76	118	84	85	96	78	73
[127]	91	47	32	20	23	21	24	44	21	28	9	13	46	18	13	24	16	13
[145]	23	36	7	14	30	NA	14	18	20									

Missing values

# Graphical output

```
> ozone<-na.omit(airquality$Ozone)
> length(ozone)
[1] 116
> par(mfrow=c(1,2))
> boxplot(ozone)
> hist(ozone)
```

A skew distribution with  
“long” right tail.





## 1.2 Random samples in R

# Random sample form a normal distribution

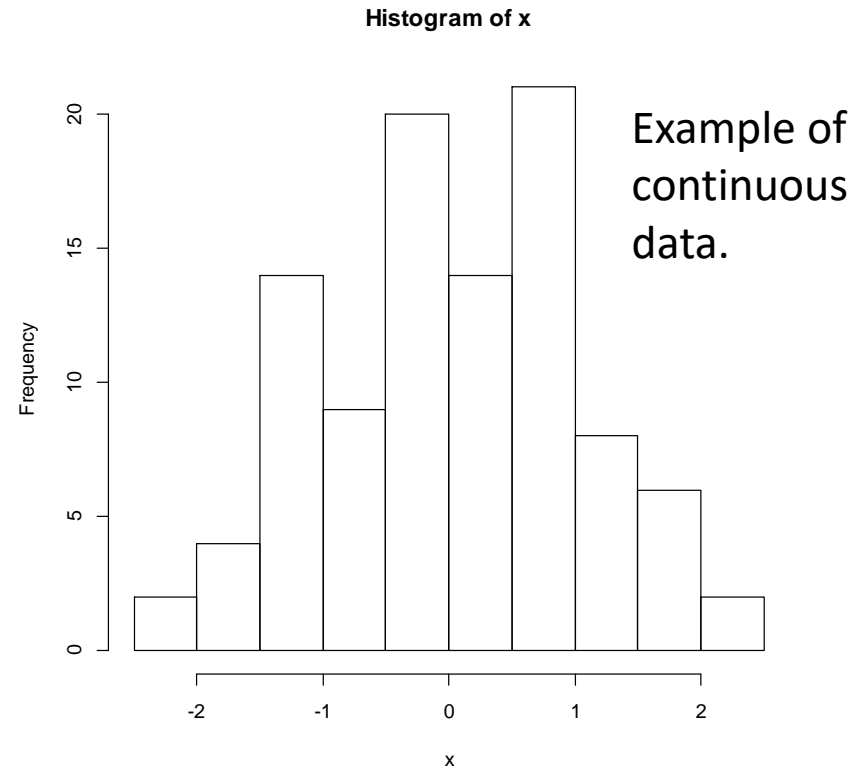
Population:

$$X_i \sim N(\mu = 0, \sigma = 1)$$

Random sample, n=100

$$X_1, X_2, \dots, X_{100}$$

```
> x<-rnorm(100,0,1)
> hist(x)
```



Histogram of 100  
observations from  $N(0,1)$ .

# Random sample form a Binomial distribution

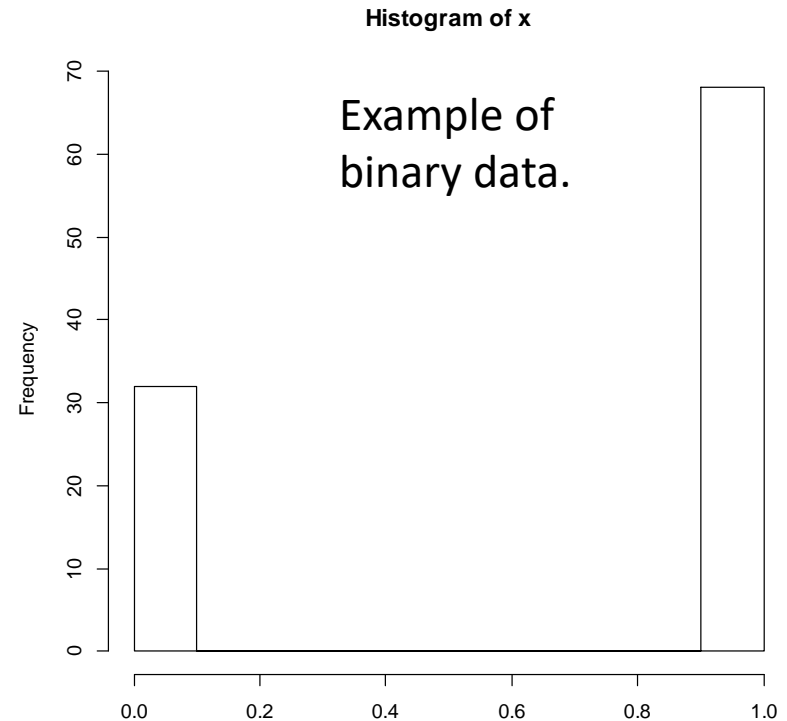
Population:

$$X_i \sim B(100, \pi = 0.7)$$

Random sample, n=100

$$X_1, X_2, \dots, X_{100}$$

```
>x<-rbinom(100,1,0.7)
>hist(x)
```



Histogram of 100  
observations from B(0.7).

# Random sample form a Poisson

Population:  $X_i \sim P(3)$

Random sample,  $n=100$   $X_1, X_2, \dots, X_{100}$

```
> x<-rpois(100,3)
```

```
> x
```

```
[1] 0 1 0 9 3 3 3 2 5 3 3 1 4 1 3 7 4 3 5 3 1 5 2 0 5 4 0 3 0 1 3 1 3 1 2 2 4  
[38] 4 5 3 5 0 3 6 3 2 3 2 0 2 1 6 4 3 3 1 1 3 5 9 4 5 2 2 3 2 3 3 2 3 3 3 0  
[75] 4 4 5 3 4 4 6 2 4 3 3 3 4 3 3 3 4 5 7 2 2 7 5 5 3 2
```

A random sample of  
100 values from  $P(3)$

```
> table(x)
```

```
x  
 0  1  2  3  4  5  6  7  9  
8 10 15 33 14 12  3  3  2
```

A frequency table

Example of  
binary data.



## Part 2:

Point estimation for population mean  $\mu$  and  
population variance  $\sigma^2$

# Sample Statistics as Estimators of Population Parameters

- A **sample statistic** is a numerical measure of a summary characteristic of a sample.
- A **population parameter** is a numerical measure of a summary characteristic of a population.
- An **estimator** of a population parameter is a sample statistic used to estimate the population parameter.



# Definition: sample mean and sample variance

If  $X_1, \dots, X_n$  is a random sample from a population  $X$  then:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{The sample mean}$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \quad \text{The sample variance}$$

# The mean and the variance

The mean and the variance of the sample values:

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

$$\begin{aligned} s^2 &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \left( \frac{1}{n-1} \sum_{i=1}^n x_i^2 \right) - \left( \frac{n}{n-1} \bar{x}^2 \right) \end{aligned}$$

# Example: the airquality data

The estimators for the unknown parameter ( $\mu, \sigma$ ) in the population

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

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

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

Point estimators of the population parameters in R:

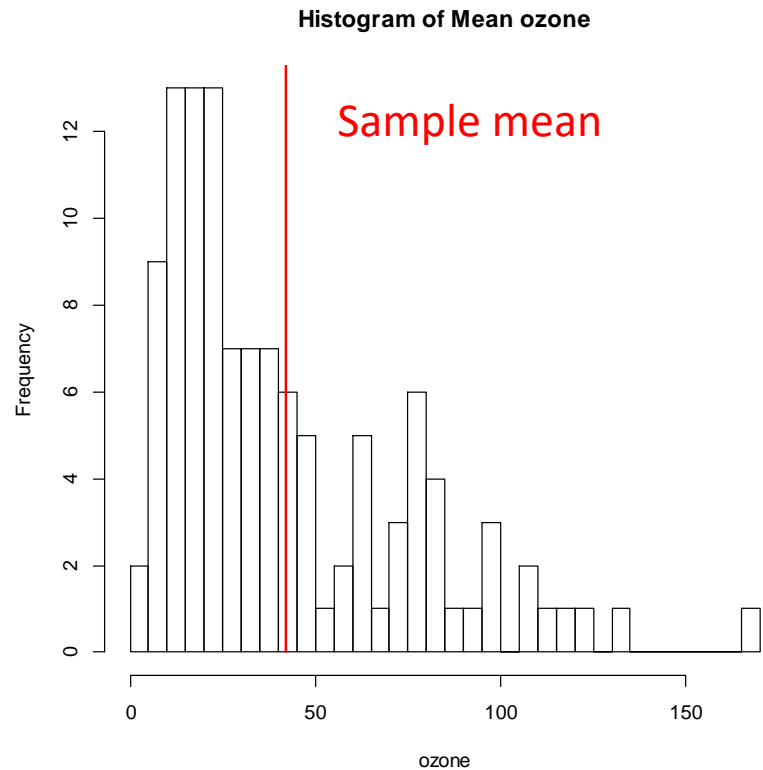
```
> ozone=airquality$Ozone
> meanozone=mean(ozone, na.rm=T)
> meanozone
[1] 42.12931

> varozone=var(ozone, na.rm=T)
> varozone
[1] 1088.201

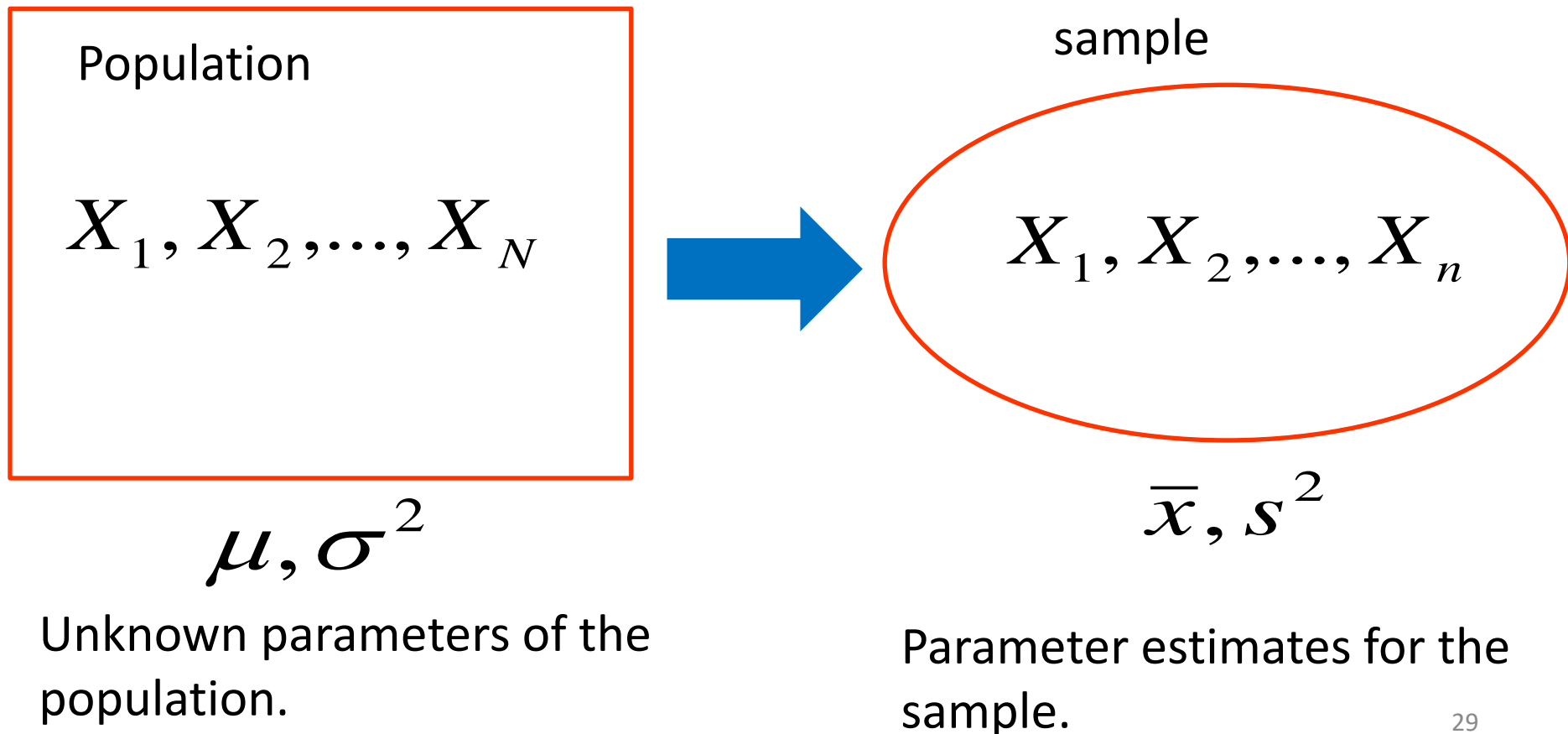
> sqrt(varozone)
[1] 32.98788
```

# Graphical output

```
> hist(ozone,breaks=25,main="Histogram of Mean ozone")  
> lines(c(meanozone,meanozone),c(0,23),col="red",lwd=2)  
> text(meanozone,25,round(meanozone,2))  
> text(100,25,"point estimate of mean ozone ",col=3)
```



# Population and sample



# Example: Average Heights and Weights for American Women

The data set gives the average heights and weights for **15** American women aged 30–39.

```
> help(women)
> women
  height weight
1     58    115
2     59    117
3     60    120
4     61    123
5     62    126
6     63    129
7     64    132
8     65    135
9     66    139
10    67    142
11    68    146
12    69    150
13    70    154
14    71    159
15    72    164
```




Sample size.

# Example: Average Heights and Weights for American Women

Parameter estimates for height:


Estimator for  $\mu$

```
> womenheight=women$height  
> meanheight=mean(womenheight,na.rm=T)  
> meanheight  
[1] 65
```


$$\bar{x} = \frac{1}{15} \sum_{i=1}^{15} x_i = 65$$

Estimator for  $\sigma^2$

```
> womenheight=women$height  
> varheight=var(womenheight,na.rm=T)  
> varheight  
[1] 20
```


$$s^2 = \frac{1}{14} \sum_{i=1}^{15} (x_i - \bar{x})^2 = 20$$



## Part 3: Variability of estimates

4.1.1: The Sampling Distributions of the sample mean

4.1.2: Standard error of the mean

Basic properties for a sample mean  $\mu$  and a sample variance  $\sigma^2$

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 4.1



# The sample mean

Let  $X_1, X_2, \dots, X_n$  is a random sample from a population  $X$  with mean  $\mu$  and variance  $\sigma^2$  :

$$E(X_i) = \mu$$

$$Var(X_i) = \sigma^2$$



Parameters of the population

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$



Random variable in the sample.

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

$$Var(\bar{X}) = ?$$

# The variance of the sample mean

If the population  $X$  have mean  $\mu$  and variance  $\sigma^2$   
i.e.,  $E(X) = \mu$  and  $\text{Var}(X) = \sigma^2$ , then the sample mean  $\bar{X}$  is a mean  
and variance given by:

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

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

# The sample variance

- The sample variance  $S^2$  is an estimator of the population variance  $\sigma^2$

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

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

# Sampling Distributions of the sample mean

- $\bar{X}$  is a random variable. Its value is determined partly by which people are randomly chosen to be in the sample.
- Many possible samples, many possible  $\bar{X}$  's.
- The variance of the sample mean:

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

- The variance of the sample mean is reduced as the sample size increases.

# Example

```
> mx1=c(1:1000)
> mx2=c(1:1000)
> mx3=c(1:1000)
> mx4=c(1:1000)
> for(i in 1:1000)
+ {
+ sample1=rnorm(2,5,1)
+ sample2=rnorm(10,5,1)
+ sample3=rnorm(50,5,1)
+ sample4=rnorm(100,5,1)
+ mx1[i]=mean(sample1)
+ mx2[i]=mean(sample2)
+ mx3[i]=mean(sample3)
+ mx4[i]=mean(sample4)
+ }
> mean(mx1)
[1] 4.967702
```

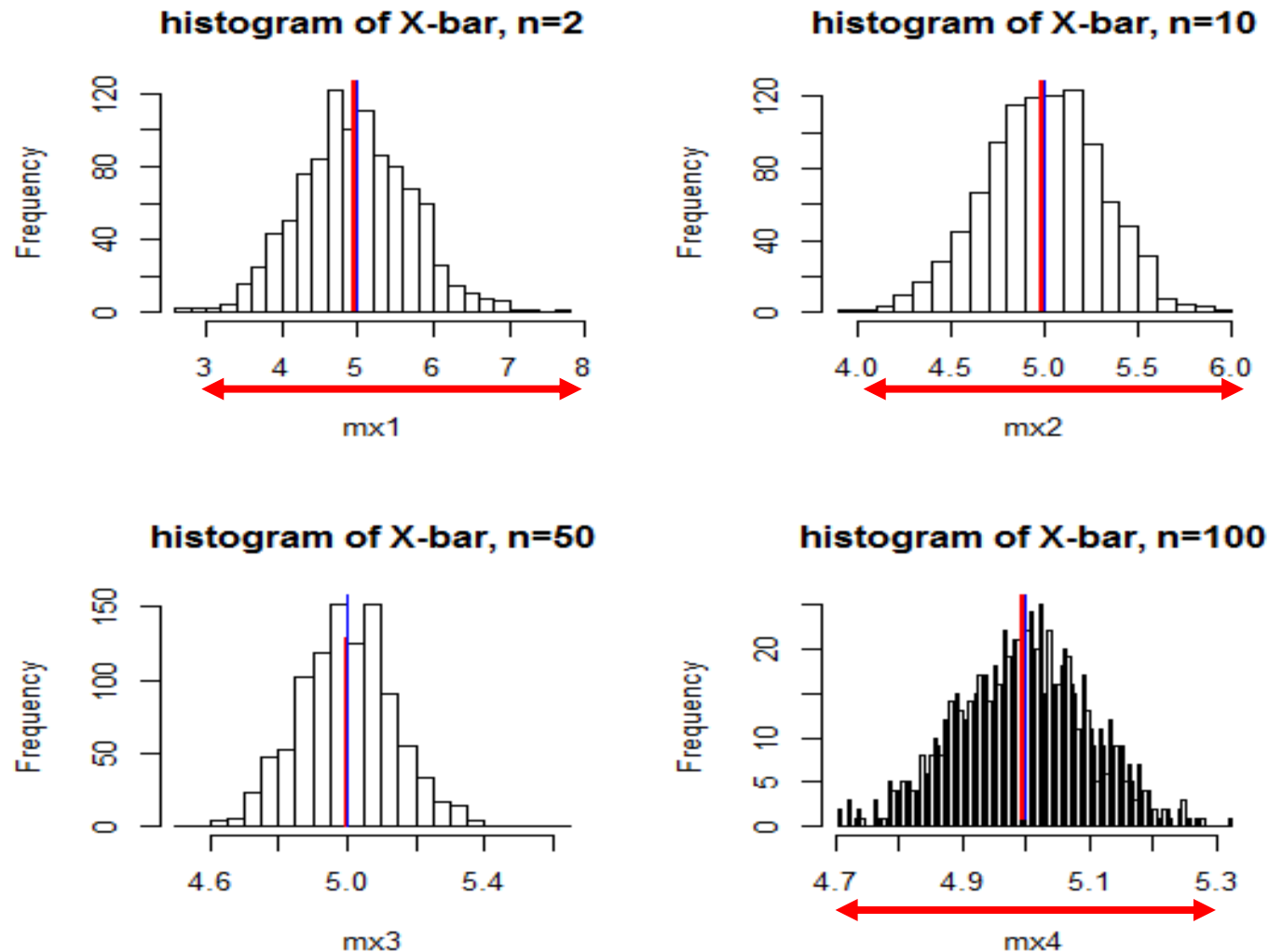
Example:

- 4 samples from normal distribution.
- Sample size=1000.



Sampling from normal distributions.

# Graphical output: distribution of the sample mean



- $n$ : sample size.
- Distribution of the sample mean for different values of  $n$ .

# Example for graphical output

```
> par(mfrow=c(2,2))
> hist(mx1,breaks=20, main="histogram of X-bar, n=2")
> lines(c(mean(mx1),mean(mx1)),c(1,129),col="red",lwd=2)
> lines(c(5,5),c(1,1000),col="blue")
> hist(mx2,breaks=20, main="histogram of X-bar, n=10")
> lines(c(mean(mx2),mean(mx2)),c(1,129),col="red",lwd=2)
> lines(c(5,5),c(1,1000),col="blue")
> hist(mx3,breaks=20, main="histogram of X-bar, n=50")
> lines(c(mean(mx3),mean(mx3)),c(1,129),col="red",lwd=2)
> lines(c(5,5),c(1,1000),col="blue")
> hist(mx4,breaks=200, main="histogram of X-bar, n=100")
> lines(c(mean(mx4),mean(mx4)),c(1,129),col="red",lwd=2)
> lines(c(5,5),c(1,1000),col="blue")
```

# Distribution of the sample mean

We distinguish between three cases :

1. X has a normal distribution with unknown  $\mu$  and  $\sigma^2$  known.
2. X has an unknown distribution, but we have a large sample.
3. X has a normal distribution with unknown  $\mu$  and  $\sigma^2$  and small sample.



# Distribution of the sample mean: case 1

When the sample comes from a normal population with known  $\sigma^2$ , i.e:

$$X_i \sim N(\mu, \sigma^2) \quad i = 1, \dots, n$$

Then for the sample mean :  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

And so:

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

## Distribution of the sample mean: case 2

If the sample comes from a population whose distribution is unknown, but the **sample size is large** (in practice  **$n > 30$** ), then approximate:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

If the variance in the population  $\sigma^2$  is unknown, use the parameter estimate, the sample variance  $s^2$ :

$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}} \sim N(0,1)$$

## $\sigma^2$ and $S^2$

The population variance  $\sigma^2$  is usually not known.  
We can use the estimator for  $\sigma^2$ : sample variance  $S^2$ .

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

The sample variance  $S^2$  is an unbiased estimator of the population variance

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n} \quad \Rightarrow \quad \text{SE}(\bar{X}) = \sqrt{\frac{S^2}{n}}$$

The standard error of the sample mean.

## Distribution of the sample mean: case 3

If the sample comes from a population whose **distribution is normal and the variance is unknown**, and the sample is small , then :

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

A Student t-distribution with (n-1) degrees of freedom is denoted by t (n-1).

# The standard error of the sample mean

An estimate for the standard error of the sample mean.

$$SE = \frac{s}{\sqrt{n}}$$

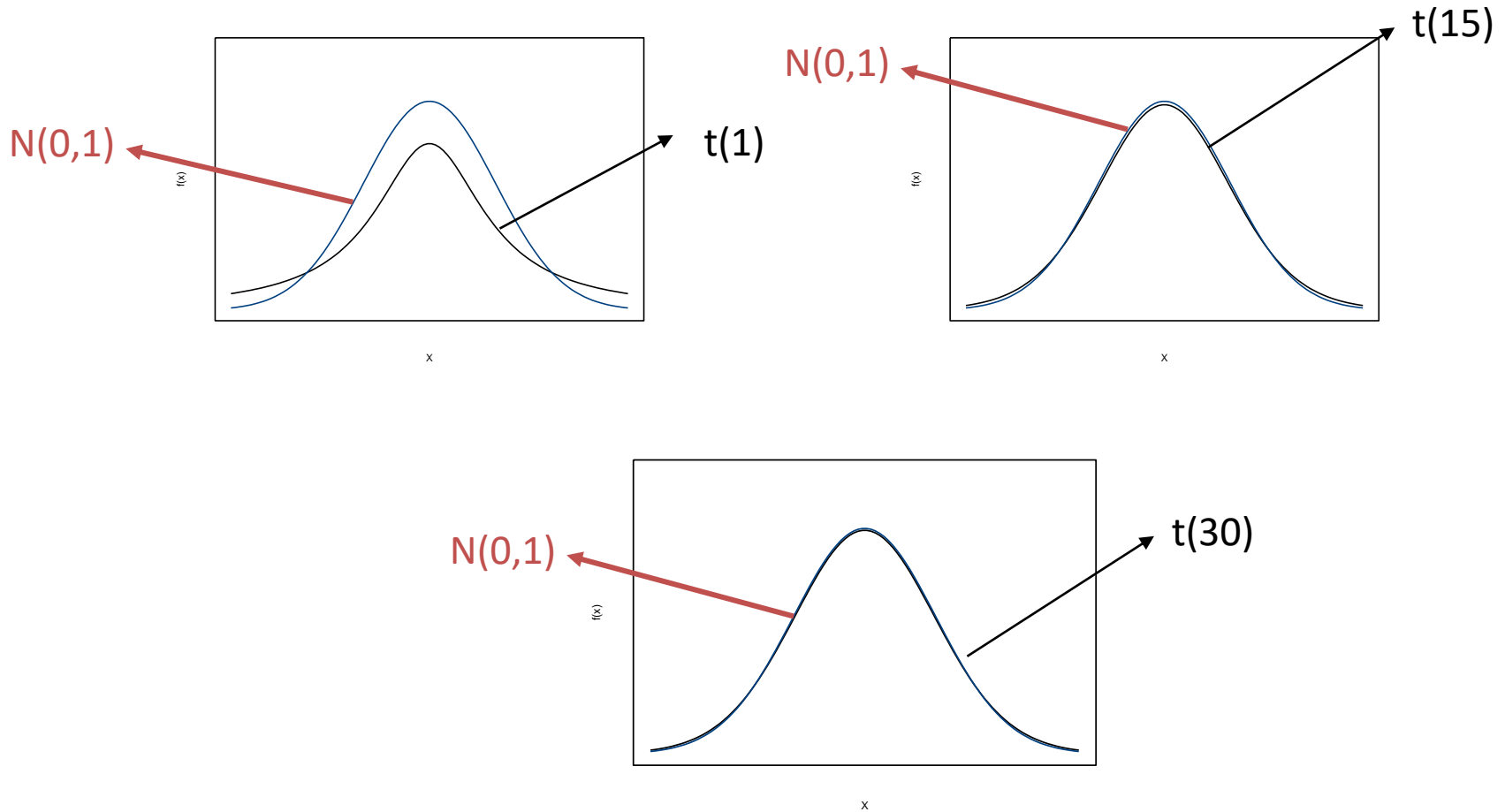
$s$  → The standard deviation.  
 $n$  → The sample size.

If the sample comes from a population whose **distribution is normal** and the **variance is unknown**, and the **sample size is small**, then :

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

A Student t-distribution with (n-1) degrees of freedom is denoted by  $t_{(n-1)}$ .

# Student's t-distributions, and $N(0,1)$



# Example

$$X \sim t_{(13)}$$

What is the value of  $a$ , so that

$$P(X > a) = 0.025 \text{ ?}$$

# Student's t-distribution

p df	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	1	3.078	6.314	12.706	31.821	63.657	318.309
2	0.817	1.886	2.92	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.215
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.44	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.86	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.25	4.297
10	0.7	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.696	1.356	1.782	2.179	2.681	3.055	3.93
13	0.694	1.35	1.771	2.16	2.65	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.69	1.337	1.746	2.12	2.583	2.921	3.686
17	0.689	1.333	1.74	2.11	2.567	2.898	3.646
18	0.688	1.33	1.734	2.101	2.552	2.878	3.61
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.08	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.5	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.06	2.485	2.787	3.45
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.31	1.697	2.042	2.457	2.75	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
50	0.679	1.299	1.676	2	2.403	2.678	3.261
60	0.679	1.31	1.671	2.009	2.39	2.66	3.232
120	0.677	1.289	1.658	1.98	2.358	2.61	3.16
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.000

Look in row labeled  
**13** and column labeled  
**.025** to find  $P(x > 2.16)$   
**= 0.025**

$$P(X > 2.16) = 0.025$$



# Student's t-distribution

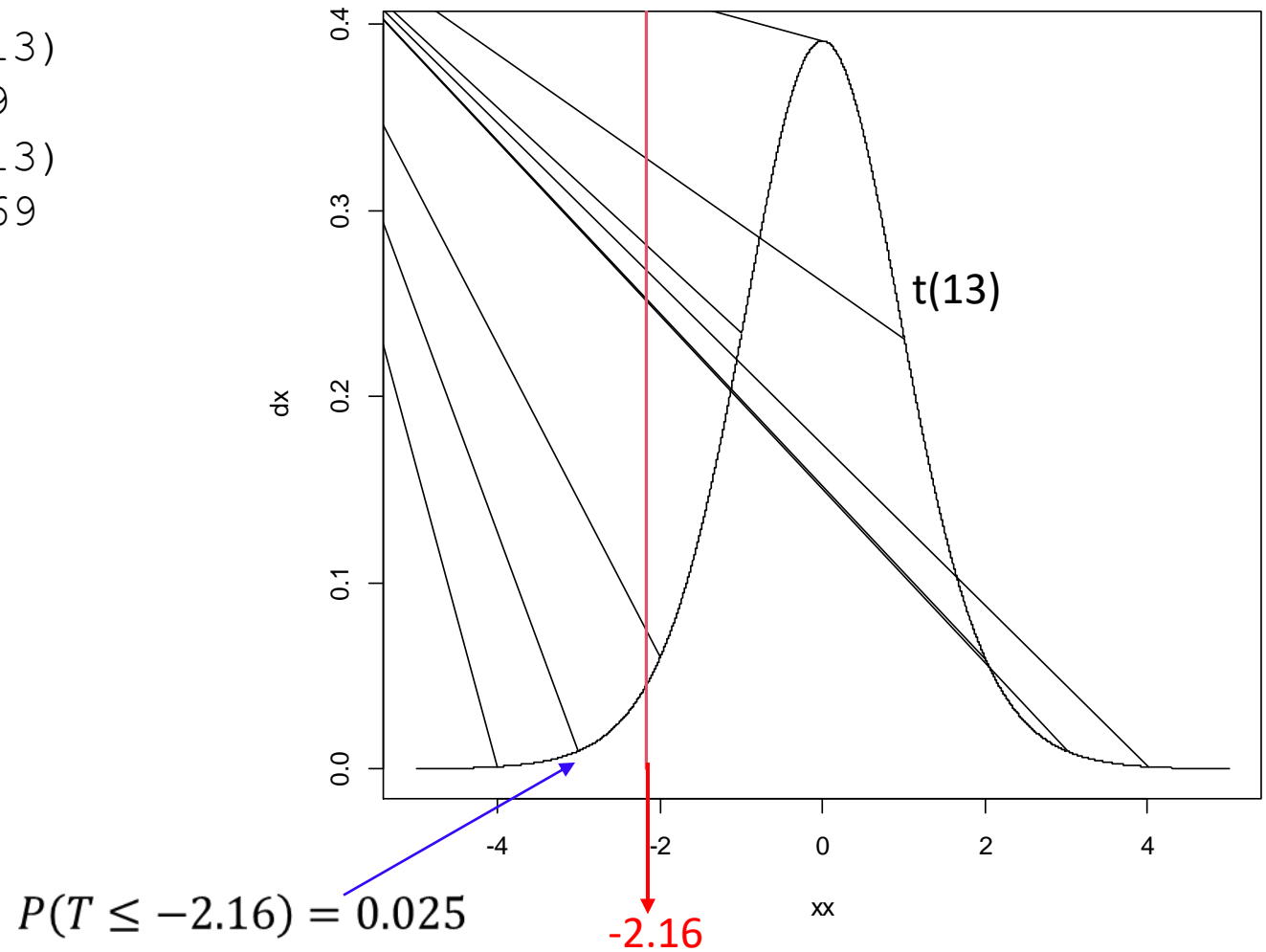
p df	0.25	0.1	0.05	0.025	0.01	0.005	0.001
1	1	3.078	6.314	12.706	31.821	63.657	318.309
2	0.817	1.886	2.92	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.215
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.44	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.86	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.25	4.297
10	0.7	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.696	1.356	1.782	2.179	2.681	3.055	3.93
13	0.694	1.35	1.771	2.16	2.65	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.69	1.337	1.746	2.12	2.583	2.921	3.686
17	0.689	1.333	1.74	2.11	2.567	2.898	3.646
18	0.688	1.33	1.734	2.101	2.552	2.878	3.61
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.08	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.5	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.06	2.485	2.787	3.45
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.31	1.697	2.042	2.457	2.75	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
50	0.679	1.299	1.676	2	2.403	2.678	3.261
60	0.679	1.31	1.671	2.009	2.39	2.66	3.232
120	0.677	1.289	1.658	1.98	2.358	2.61	3.16
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.000

$$P(T > 2.16) = 0.025$$

$$P(T < -2.16) = 0.025$$

# Student's t-distribution in R

```
> qt(0.975, 13)
[1] 2.160369
> qt(0.025, 13)
[1] -2.160369
```

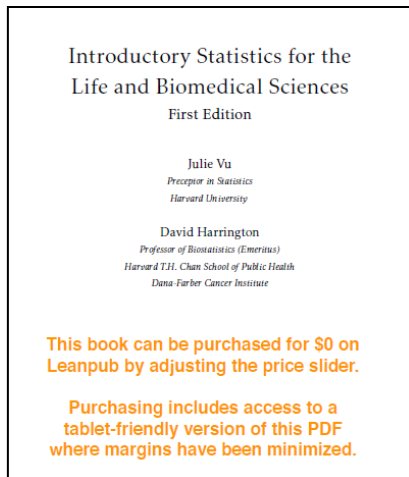


	Case 1	Case 2	Case 3
<b>Distribution of the population</b>	$X_i \sim N(\mu, \sigma^2)$	$X_i \sim \text{unknown}$	$X_i \sim N(\mu, \sigma^2)$
$\sigma^2$	Known	Known / Not known	Not known
<b>Sample size</b>	No condition	$\geq 30$	$< 30$
<b>Distribution of the sample mean</b>	$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$	$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ $\bar{X} \sim N(\mu, \frac{S^2}{n})$	
<b>Standardized distribution</b>	$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$	$\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2 \text{ of } S^2}{n}}} \sim N(0,1)$	$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim t_{(n-1)}$



## Part 4: Confidence intervals

### Interval Estimation for a population mean $\mu$



# Interval Estimation

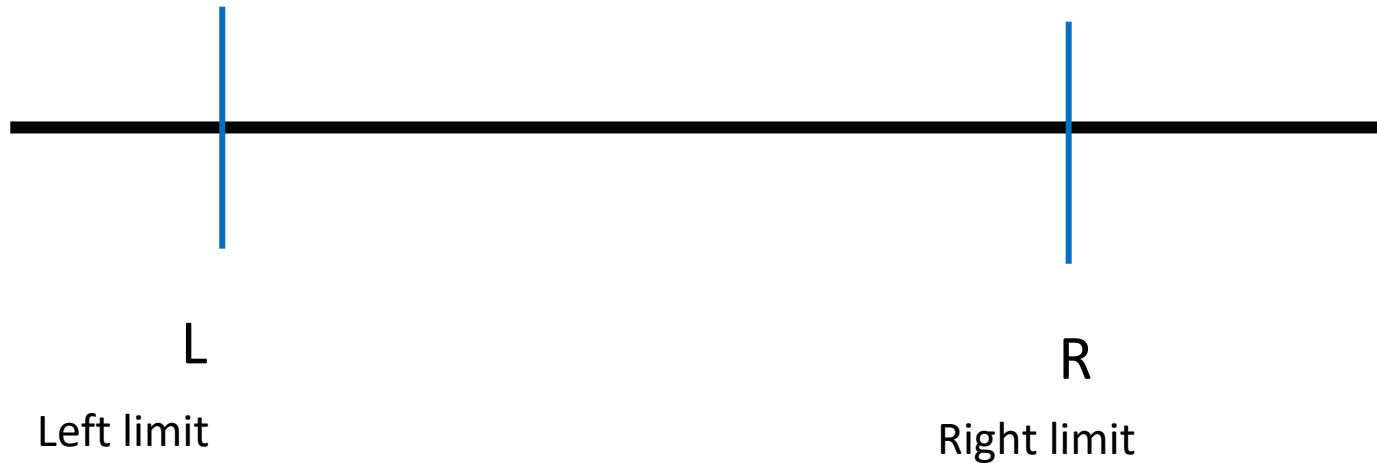


- On the basis of the sample, we find an interval in parameter space which contains this parameter  $\mu$  “almost always”
- Such an interval is called a confidence interval

# Confidence interval

On the basis of the sample, we find an interval  $[L, R]$  so :

$$P(\mu \in [L, R]) = \text{"large"}$$

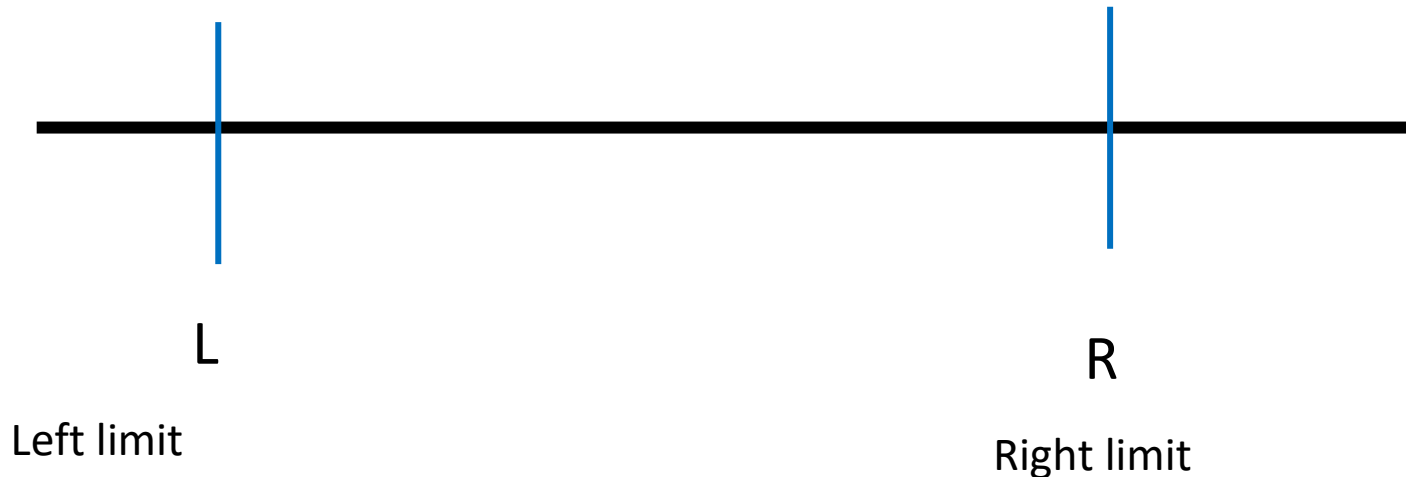


We find an interval  $[L, R]$  that contains the value of the population mean ( $\mu$ ) with "high probability"

# Confidence interval

Large  $\rightarrow 1 - \alpha$

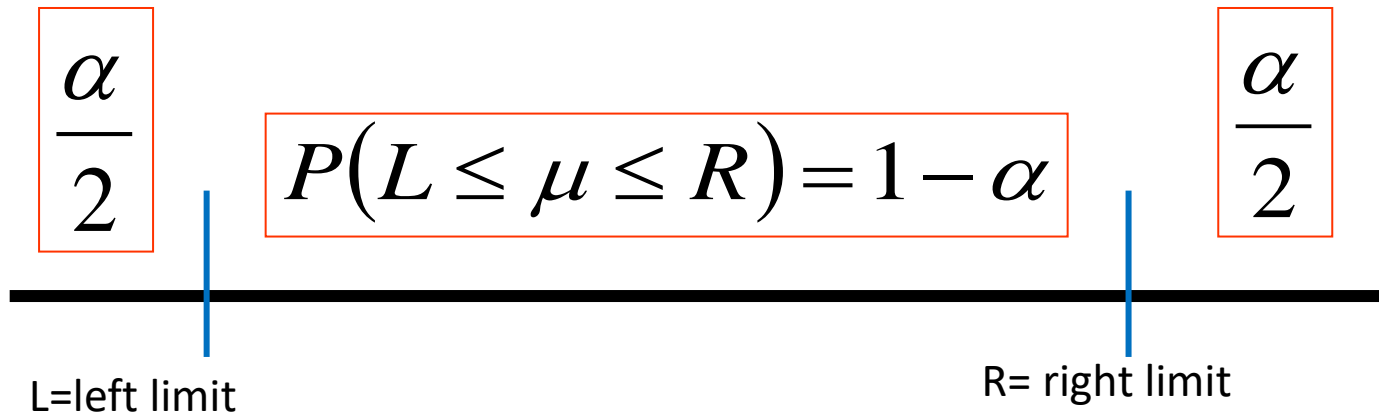
$$P(L \leq \mu \leq R) = 1 - \alpha$$



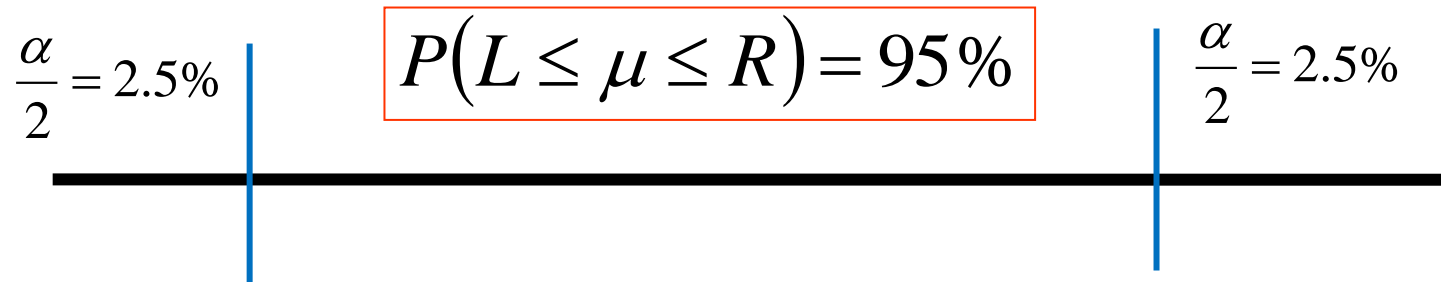
Example:  $\alpha = 0.05 \Rightarrow 1 - \alpha = 0.95$  Confidence level

# Confidence interval

We are looking for an interval that



$$\alpha = 5\% \Rightarrow 1 - \alpha = 95\%$$





# A confidence interval for a population mean

- How can we calculate L and R?
- CI calculate in 5 steps:
  - **Step 1**: choose a confidence level  $1-\alpha$   
(e.g. 90%, 95%, 99%)
  - **Step 2**: decision on the basis of the data in which case you are in (1, 2 or 3)
  - **Step 3**: Find the critical points in the correct table  
(Standard Normal or t-table)
  - **Step 4**: calculate the point estimators for  $\mu$  ( $\sigma^2$  and if not known)
  - **Step 5**: calculate the confidence interval

## Case 1

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

then:  $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

And  $Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$

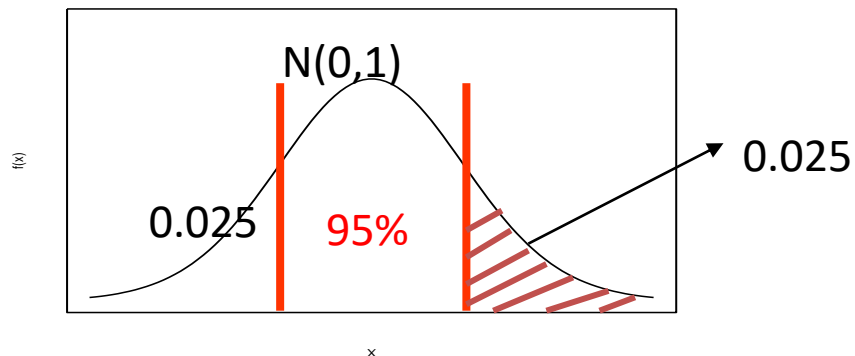
1. X has a normal distribution with unknown  $\mu$  and known  $\sigma^2$ .

# CI for case 1

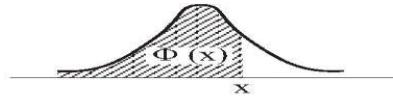
**Step 1:** example, choose  $1-\alpha = 0.95$

**Step 2:** Case 1, thus: 
$$Z_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$$

**Step 3:** critical point:



Tabel 3 : Standaard normale verdeling



	$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$									
x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998



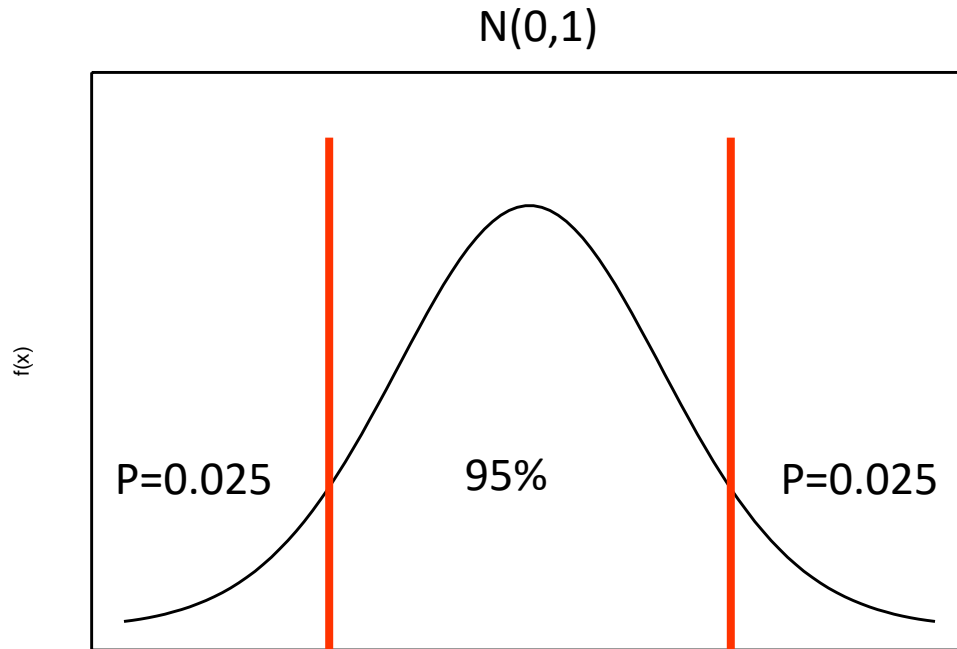
$$P(Z < 1.96) = 0.975$$

$$P(Z > 1.96) = 0.025$$

# Critical value in R

```
> qnorm(0.025,0,1)
[1] -1.959964
> qnorm(0.975,0,1)
[1] 1.959964
>
```

# CI for case 1



Critical point = -1.96    ×    Critical point = 1.96

From the table of the **standard normal distribution**, we find that :

$$P\left[-1.96 \leq \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq 1.96\right] = 0.95$$

Thus, critical points: -1.96 and 1.96

# CI for case 1

Step 4: calculate the point estimator :  $\bar{X}$

Step 5: calculate the CI

For this, we know:

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq 1.96\right) = 0.95$$

or, after the conversion of the formula:

$$P\left(\bar{X} - 1.96 \times \sqrt{\frac{\sigma^2}{n}} \leq \mu \leq \bar{X} + 1.96 \times \sqrt{\frac{\sigma^2}{n}}\right) = 0.95$$

## CI for case 1

$$P\left(\underbrace{\bar{X} - 1.96 \times \sqrt{\frac{\sigma^2}{n}}}_{L} \leq \mu \leq \underbrace{\bar{X} + 1.96 \times \sqrt{\frac{\sigma^2}{n}}}_{R}\right) = 0.95$$

L

R

$$P(L \leq \mu \leq R) = 1 - \alpha$$

So, a  $(1-\alpha)$  CI for  $\mu$  is :

$$\left[ \bar{x} - z \sqrt{\frac{\sigma^2}{n}}, \bar{x} + z \sqrt{\frac{\sigma^2}{n}} \right]$$

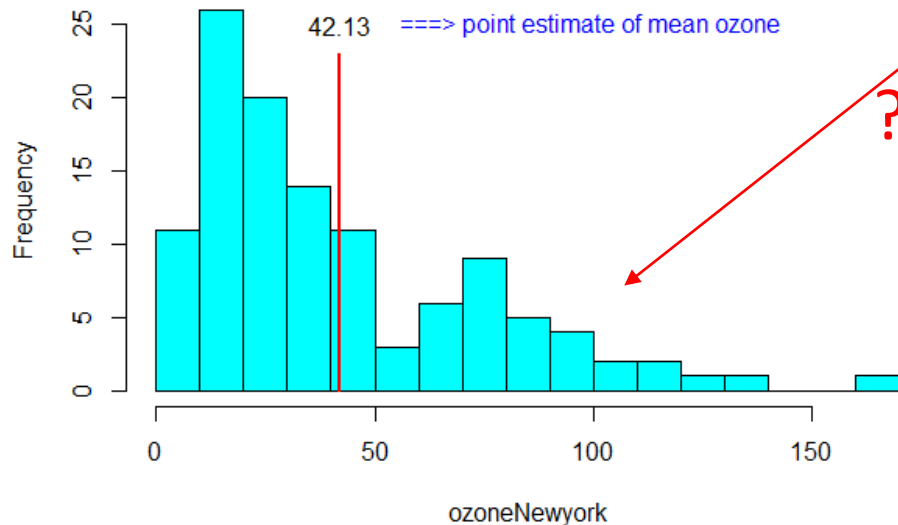


# Example for case 1

## The airquality data

- Daily reading of the mean ozone in parts per billion value at Roosevelt Island from May 1, 1973 to Sep 30, 1973 were use.
- $X$  = Daily reading of ozone value

Histogram of Mean ozone in parts per billion



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

and  $\sigma^2 = 1024$  known

???

Sample ( $n = 116$ ) from population

The sample mean  
 $\bar{x} = 42.13$

# The mean

$$X_i \sim N(\mu, \sigma^2 = 1024)$$

- The mean is unknown parameter.
1. Determine a point estimator for ozone value
  2. Determine a 95% CI for the mean of the ozone values

# Example for case 1

The 95% CI for  $\mu$  : the mean ozone value in the population

**Step 1:** *Choose a confidence level  $1-\alpha = 0.95$*

**Step 2:** *Decide on the basis of the data in which case you are in:*  
**population is normally distributed**  
 **$\sigma^2$  known**

→ **Case 1**, so normal distribution

**Step 3** Find the critical points in the appropriate table **N (0,1)**: -  
1.96 and 1.96

**Step 4:** *Calculate the point estimator for  $\mu$ :*

**Step 5:** *calculate the confidence interval for  $\bar{x} = 42.13$*

## Example for case 1

$$P\left(\underbrace{\bar{X} - 1.96 \times \sqrt{\frac{\sigma^2}{n}}}_L \leq \mu \leq \underbrace{\bar{X} + 1.96 \times \sqrt{\frac{\sigma^2}{n}}}_R\right) = 0.95$$
$$42.13 - 1.96 \times \sqrt{\frac{1024}{116}} \qquad 42.13 + 1.96 \times \sqrt{\frac{1024}{116}}$$
$$= 36.306 \qquad = 47.973$$

A 95% CI for the population mean  $\mu$  of the ozone value [36.306, 47.953]

### Interpretations:

Based on our sample, we are 95% confident that the true mean of ozone value lie in between 36.306 and 47.973

# Confidence interval using R

## For Case 1

```
> ozone=na.omit(airquality$Ozone)
> sigma=32
> sem=sigma/sqrt(n)
> E=qnorm(0.975)*sem
> xbar=mean(ozone)
> xbar+c(-E,E)
```

```
[1] 36.30601 47.95261
```

# Analysis using the R package TeachingDemos

```
> library(TeachingDemos)
> ozone=na.omit(airquality$Ozone)
> sigma=32
> z.test(ozone,sd=sigma)
```

One sample test with normal distribution  
using the function **z.test()**

One Sample z-test

data: ozone

$z = 14.1796$ ,  $n = 116.000$ , Std. Dev. =  $32.000$ , Std. Dev. of the  
sample mean =  $2.971$ , p-value  $< 2.2e-16$

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

36.30601 47.95261

sample estimates:

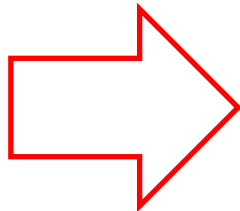
mean of ozone

42.12931

## Case study 1a:

The `airquality` data: analysis of the  
average wind speed

Confidence interval for the  
population mean



Page 187

## Case 2

If  $X \sim F$

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

and  $T_{\bar{X}} = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)$

3. X has an unknown distribution, but we have a **large sample** ( $n > 30$ )

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

The same as case 1 but we replace  $\sigma^2$  by  $S^2$ .



## CI for case 2

**Step 1:** example, choose  $1-\alpha = 0.95$

**Step 2:** case 2, so :

$$\boxed{\frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)} \quad \text{or} \quad \boxed{\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)}$$

**Step 3:** critical points: -1.96 and 1.96  
(the same as in Case 1, since we are still using the **standard normal distribution** function)

**Step 4:** Calculate the point estimator (s)  $\bar{x}$  (and possibly  $s^2$ )

## CI for case 2

**Step 5:** In the same manner as in Case 1:

The  $(1-\alpha)$  CI for  $\mu$  is :

$$\left[ \bar{x} - z \sqrt{\frac{\sigma^2}{n}}, \bar{x} + z \sqrt{\frac{\sigma^2}{n}} \right] \quad \text{or} \quad \left[ \bar{x} - z \sqrt{\frac{s^2}{n}}, \bar{x} + z \sqrt{\frac{s^2}{n}} \right]$$

## Example for case 2

### The airquality Data

- Suppose  $X$  = Daily reading of ozone value.
- $X$  has an unknown distribution with unknown variance.
- But large sample ( $n = 116 \gg 30$ ).

The 95% CI for  $\mu$  : the mean ozone value in the population

*Step 1: choose confidence level  $1-\alpha = 0.95$*

*Step 2: case 2, so :*

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

*Step 3: critical points: -1.96 and 1.96*

*(the same as in Case 1, since we are still using the  
standard normal distribution function)*

## Example for case 2

*Step 4 : Calculate the point estimators:*

$$\bar{x} = 42.13 \text{ and } s^2 = 1088.201$$

*Step 5 : In the same manner as in Case 1:*

*The  $(1-\alpha)$  CI for  $\mu$  is :*

$$\Rightarrow \left[ \bar{x} - z \sqrt{\frac{s^2}{n}}, \bar{x} + z \sqrt{\frac{s^2}{n}} \right]$$

$$\Rightarrow \left[ 42.13 - 1.96 \sqrt{\frac{1088.2}{116}}, 42.13 + 1.96 \sqrt{\frac{1088.2}{116}} \right]$$

$$\Rightarrow [36.126, 48.132]$$

## Example for case 2

A 95% CI for the population mean  $\mu$  of the ozone value [36.126, 48.132]

### **Interpretations:**

Based on our sample, we are 95% confident that the true mean of ozone value lie between 36.126 and 48.132.

# Confidence Interval using R

## For Case 2

```
> ozone=na.omit(airquality$Ozone)
> sigma=sd(ozone)
> sem=sigma/sqrt(n)
> E=qnorm(0.975)*sem
> xbar=mean(ozone)
> xbar+c(-E,E)
```

```
[1] 36.12624 48.13238
```

# Analysis using the R package TeachingDemos

```
> library(TeachingDemos)
> ozone=na.omit(airquality$Ozone)
> z.test(ozone, sd=sd(ozone),conf.level = 0.95)
```

One Sample `z-test`

```
data:  ozone
z = 13.7549, n = 116.000, Std. Dev. = 32.988, Std.
Dev. of the sample mean = 3.063, p-value <
2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 36.12624 48.13238
sample estimates:
mean of ozone
 42.12931
```

## CI for case 3

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

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

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

3. X has a normal distribution with unknown  $\mu$  and  $\sigma^2$ .  
**n is small.**

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

We use  $t(n-1)$  instead of  $N(0,1)$ .



## CI for case 3

**Step 1:** example, choose  $1-\alpha = 0.95$

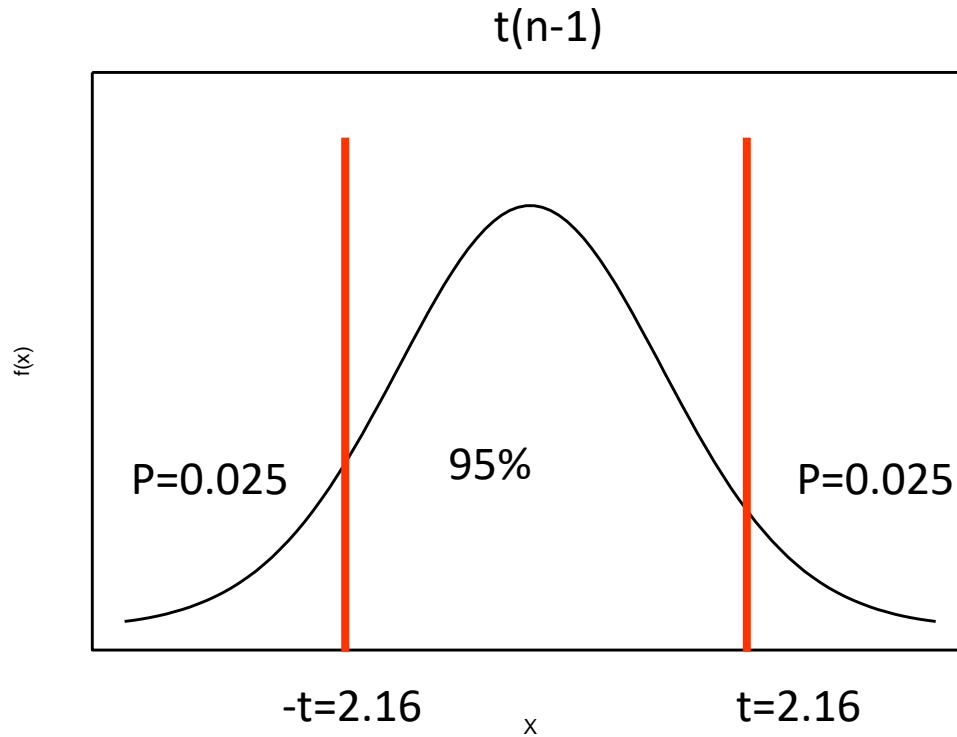
**Step 2:** case 3, so:

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

**Step 3:** critical points: gives you the **t-table**  
value will depend on the number of degrees of freedom  
(and hence on the size of the sample)

Take for example,  $n = 14$

# CI for case 3



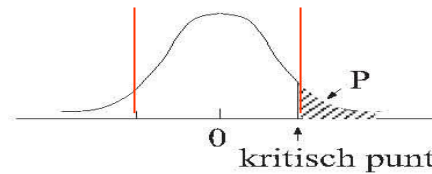
Example:  $n=14$  and  $\alpha=0.05$

From t table with 13 df, we find that

$$P\left(-2.16 \leq \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} \leq 2.16\right) = 0.95$$

# Student t-distribution

Tabel 4 : Kritische punten student t verdeling



P	.25	.10	.05	.025	.010	.005	.001
v.g.							
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

$$P(T > 2.16) = 0.025$$

$$P(T < -2.16) = 0.025$$

$$C.I : 95\% \Rightarrow \alpha = 0.05 \Rightarrow \frac{\alpha}{2} = 0.025$$

## CI for case 3

$$P\left(\underbrace{\bar{X} - t \times \sqrt{\frac{S^2}{n}}}_L \leq \mu \leq \underbrace{\bar{X} + t \times \sqrt{\frac{S^2}{n}}}_R\right) = 1 - \alpha$$

$$P(L \leq \mu \leq U) = 1 - \alpha$$

So: a  $(1-\alpha)$  CI for  $\mu$  is :

$$\left[ \bar{X} - t \times \sqrt{\frac{S^2}{n}}, \bar{X} + t \times \sqrt{\frac{S^2}{n}} \right]$$

## Example for case 3

### The women Data:

Suppose  $X$  = the height of woman aged 30 – 39.

We assume that:

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

Sample Size:  $n = 15$

Mean:  $\bar{x} = 65$

Variance:  $s^2 = 20$

## Example for case 3

The 95% CI for  $\mu$ , the average height of women aged 30 - 39 :

**Step 1:** Choose a confidence level  $1-\alpha = 0.95$

**Step 2:** Decide on the basis of the data in which case you are in:  
normal distribution, small sample ( $n = 15 < 30$ ) and unknown  $\sigma^2$   
→ Case 3 , so t distribution:  $t(15-1)$

**Step 3:** Find the critical points in the correct table : -2.145 and 2.145

**Step 4:** Calculate the point estimators for  $\mu$  and  $\sigma^2$  :

$$\bar{x} = 65 \text{ and } s^2 = 20$$

## Example for case 3

**Step 5:** *calculate the confidence interval :*

$$\left[ \bar{x} - t \sqrt{\frac{s^2}{n}}, \bar{x} + t \sqrt{\frac{s^2}{n}} \right] = \left[ 65 - 2.145 \sqrt{\frac{20}{15}}, 65 + 2.145 \sqrt{\frac{20}{15}} \right] \\ = [62.52; 67.48]$$

A 95% CI for the population mean  $\mu$  of the women height is  
[62.52, 67.48]

### **Interpretation:**

Based on our sample, we are 95% confident that the true mean of the women height aged 30 – 39 will lies in between 62.52 and 67.48

# Confidence Interval using R

## *For Case 3*

```
> Height=women$height
> n=length(Height)
> SE=s/sqrt(n)
> E=qt(0.975,df=n-1)*SE
> xbar=mean(Height)
> xbar+c(-E,E)
```

```
[1] 62.52341 67.47659
```

OR

```
> library(TeachingDemos)
> t.test(women$height)
```



# Analysis using the R function `t.test()`

```
> t.test(women$height)
```

One Sample t-test

```
data:  women$height
```

```
t = 56.2917, df = 14, p-value < 2.2e-16
```

```
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
```

```
 62.52341  67.47659
```

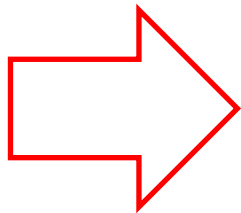
```
sample estimates:
```

```
mean of x
```

```
 65
```

Case study 2:  
The The NHANES dataset: BMI

Confidence interval for the  
population mean

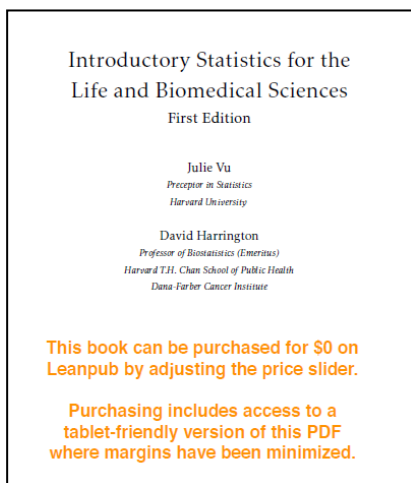


Page 198



## Part 5:

# Hypotheses testing



### Section 4.3

## 5.3.1: A Formal Approach to Hypothesis Testing

Testing a hypothesis about a  
population parameter

# Hypothesis tests

- In the testing theory we try, on the basis of a sample, a hypothesis about the population keys.
- In practice, we test a hypothesis about a parameter of a population distribution.
- Example:
  - The parameter of interest: the population mean.
  - The parameter  $\mu$  in a population with distribution  $N(\mu, \sigma^2)$ .

# Example: population and sample

- Two zoologists study the population of Horse shoe crabs.
- The variable of interest is the number of satellites, E,g the number of satellites in the horse shoe crab population.
- In the sample:

$$s^2 = 9.9119$$

$$\bar{x} = 2.91$$

$$n = 173$$

The zoologists collected data about 173 horse shoe crabs and in the sample the mean is equal to 2.91 and the variance to 9.911.



# Notation

$X_i$  : the number of satellites.

$$X_1, X_2, \dots, X_N$$

$$E(X_i) = \mu$$

$$Var(X_i) = \sigma^2$$

# The two hypotheses of zoologists

- The hypothesis of the first zoologist (about population mean) is that

$$E(X_i) = \mu = 2.5$$

- The hypothesis of the second zoologist (about population mean) is that

$$E(X_i) = \mu = 2.9$$

both zoologists know the variance in the population mean is not known.



# What we want to do in this section ?

- We look for a procedure we can use to make a decision about the parameter in the population.

Population

$$X_1, X_2, \dots, X_N$$

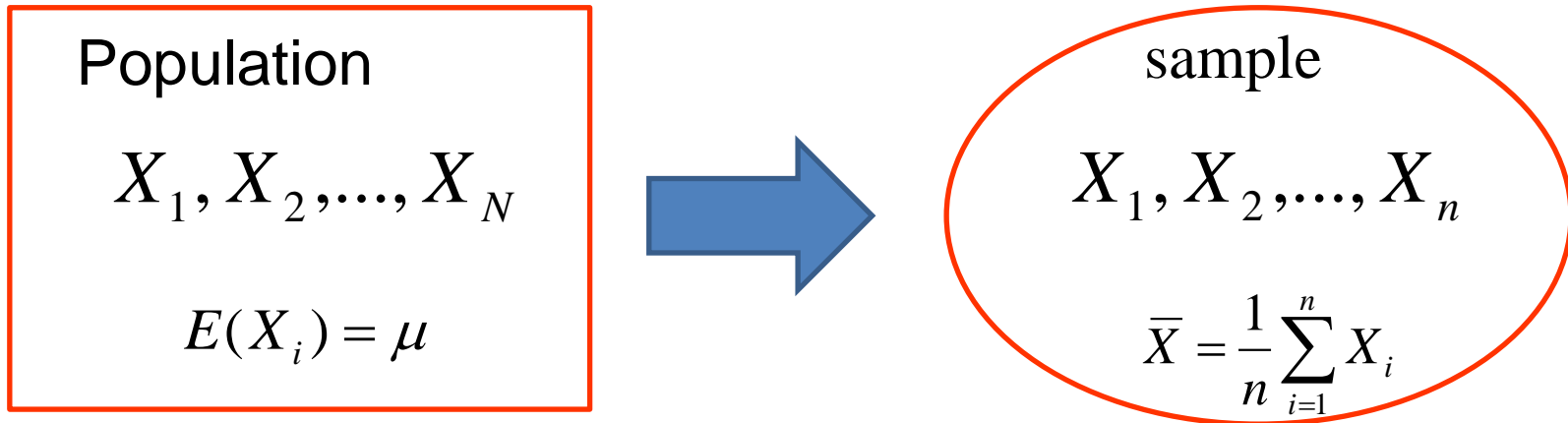
$$E(X_i) = \mu$$

# What we want to do in this section ?

We look for a procedure we can use to make a decision about the parameter in the population:

$$\mu = 2.5 \quad \text{or} \quad \mu = 2.9$$

# The decision rule



Based on the mean in the sample, we would like to make a decision about the parameter (the mean) in the population

The estimate from  
the sample of the  
unknown parameter

$\bar{x}$

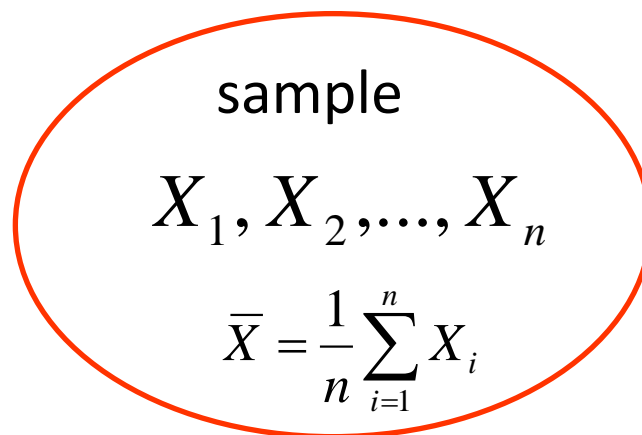
$$\mu = 2.5$$

or

$$\mu = 2.9$$

?

# The decision rule



Point estimator

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

confidence interval

$$\bar{X} \pm a \times \sqrt{\text{Var}(\bar{X})}$$

previous  
sections

this  
section

On the basis of the sample and the sample mean we want to test hypotheses about the population parameter

# The sample

- The two zoologists decide to a large sample taking the population and the number of satellites per crab to count.
- The sample size is 173.

$$X_i \sim \text{unknown}$$

$$n > 30$$

$$\sigma^2 : \text{unknown}$$

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

# The two hypotheses

- We call the hypothesis of the first zoologist: **the null hypothesis**

- We call the hypothesis of the second the zoologist: **the alternative hypothesis**

$$\begin{array}{l} H_0 : \mu = 2.5 \\ H_1 : \mu = 2.9 \end{array}$$

$$\bar{X} \sim N(2.5, \frac{S^2}{n})$$

$\mu$  under  $H_0$

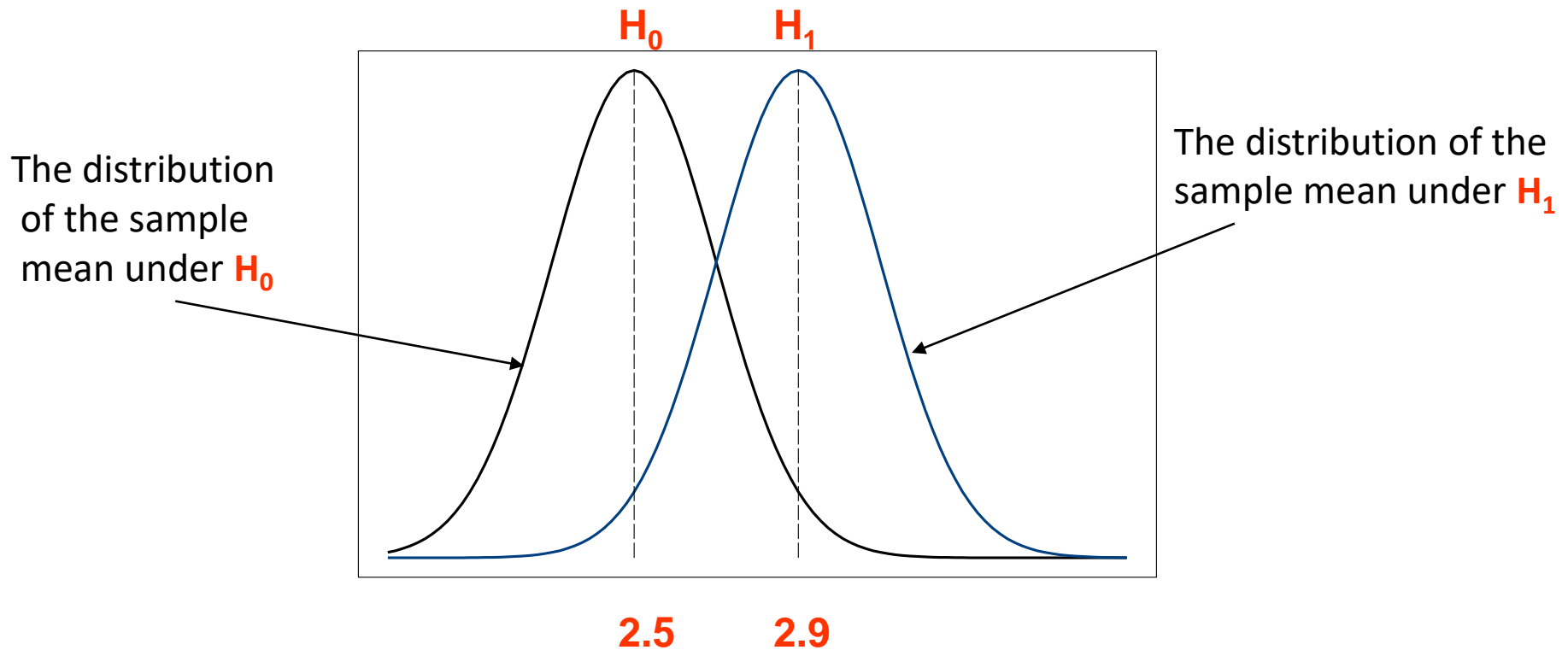
$$\bar{X} \sim N(2.9, \frac{S^2}{n})$$

$\mu$  under  $H_1$

# The distribution of the sample mean

$$\bar{X} \sim N\left(2.5, \frac{S^2}{n}\right) \quad \text{under } H_0$$

$$\bar{X} \sim N\left(2.9, \frac{S^2}{n}\right) \quad \text{under } H_1$$



# The rejection region

$$\begin{array}{l} H_0 : \mu = 2.5 \\ H_1 : \mu = 2.9 \end{array} \quad \longrightarrow \quad \mu_0 = 2.5 < \mu_1 = 2.9$$

- From the nature of the alternative hypothesis it follows that we will reject the null hypothesis if we find in our sample a value of  $\bar{x}$  which is ``too large''.
- We determine **c** so :

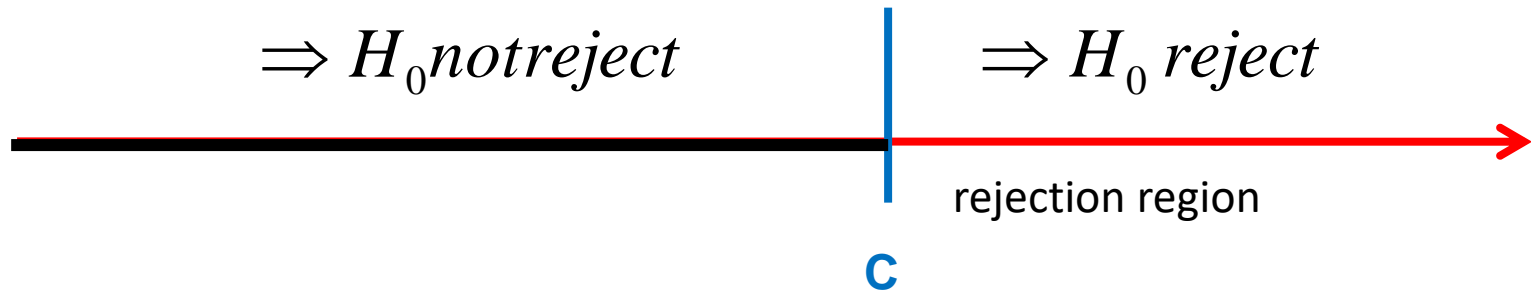
To a ``large'' value of  $\bar{x}$  (which we reject the null hypothesis)  
 $\bar{x}$  means a value that is greater than **c**.



# The rejection region: the decision rule

when we find a  $\bar{x}$  value that is greater than  $c$ , then we reject the null hypothesis.

The decision rule  $\bar{X} > c \Rightarrow H_0 \text{ reject}$



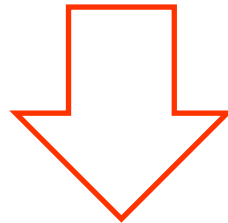
If the  $\bar{x}$  rejection region is, we reject the null hypothesis.

# “The Question”

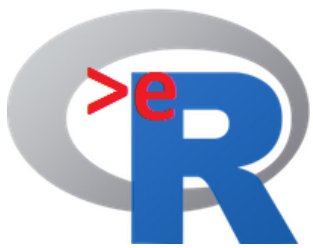
The decision rule is:

$$\bar{X} > c \Rightarrow H_0 \quad \text{reject}$$

How can we choose the value of  $c$ ????



The two types of errors



## Part 6: Decision errors

### The two types of errors

Introductory Statistics for the  
Life and Biomedical Sciences  
First Edition

Julie Vu  
*Professor 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 4.3.4: Decision errors



## 6.1 Two types of errors

# Decision errors

Definition: the two types of errors

- The two types of errors that can be made by taking a decision:
  1. if we reject the null hypothesis when it is correct then we say that we committed a type I error.
  2. if we accept the null hypothesis when it is wrong, then we say that we committed a type II error.

# Type I error

- Example: If the first zoologist is correct about the value of  $\mu$  (2.5) and we reject the null hypothesis ( $\mu = 2.5$ ) then we make a Type I error.
- General: If the null hypothesis is true and we reject the null hypothesis we make a Type I error.

# Type II error

- Example: If the second zoologist is correct about the value of  $\mu$  ( $\mu = 2.9$ ), and we do not reject the null hypothesis ( $\mu = 2.5$ ) we make a type II error.
- General: If the alternative hypothesis is correct and we do not reject the null hypothesis then we make a Type II error.

# The two errors

Test	Population	
	$H_0$ is true	$H_0$ not true
reject $H_0$	incorrect statement Type I error	correct statement
Not reject $H_0$	correct statement	incorrect statement Type II error



# The two errors

- We implement a "decision rule" for a given value of the probability of a Type I error.
- Decision rule for a given value of “small” Type I error (usually 0.01 or 0.05).
- If the researcher uses the probability of a type I error at 0.05 then there is 5% that the null hypothesis is rejected, while it is true.

# The significance level

- The Probability of a Type I error is denoted by  $\alpha$
- We call  $\alpha$  the significance level.
- We say that the null hypothesis  $H_0$  is rejected at a significance level of  $\alpha$  (or cannot be rejected at significance level  $\alpha$ ).



Choosing the value of critical point for the test

## 6.2 The decision rule

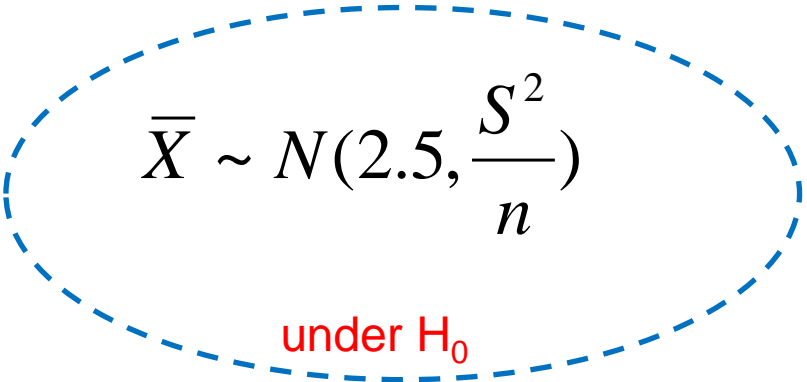
# The distribution of the sample mean

The null and alternative hypotheses

$$H_0 : \mu = 2.5$$

$$H_1 : \mu = 2.9$$

The distribution of the sample mean under the null and the alternative hypotheses


$$\bar{X} \sim N(2.5, \frac{S^2}{n})$$

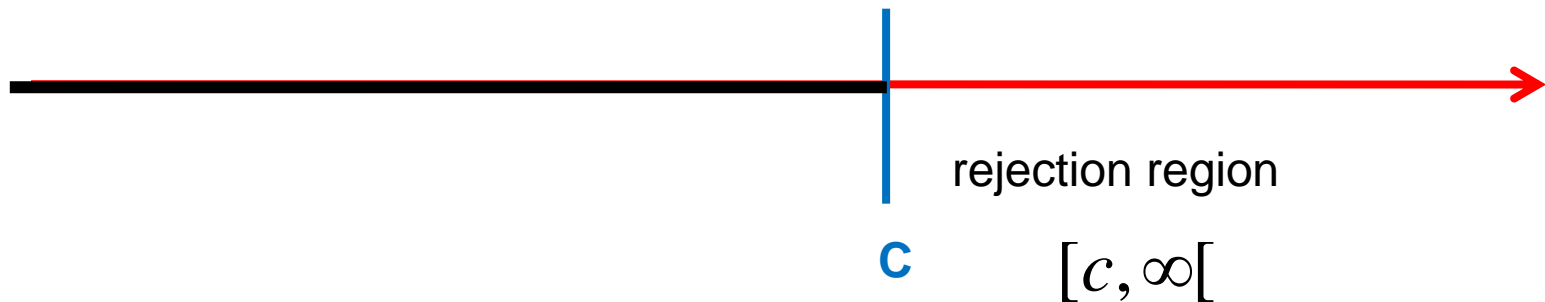
under  $H_0$

$$\bar{X} \sim N(2.9, \frac{S^2}{n})$$

under  $H_1$

# The rejection region

When the value of  $\bar{x}$  is greater than C, we reject null hypothesis



$$\bar{X} > c \Rightarrow H_0 \quad \text{reject}$$

How can we choose c ?

## Type I error & c

- If the null hypothesis is true and we reject the null hypothesis we make a type I error.

$$\alpha = 0.05$$

$$\bar{X} > c \Rightarrow H_0 \quad \textit{reject}$$

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

**if the null hypothesis is true!!**

# The critical point and the rejection region

Determine  $c$  under the null hypothesis that

$$P(\bar{X} > c) = 0.05 \quad \text{if} \quad \bar{X} \sim N\left(2.5, \frac{S^2}{n}\right)$$

$$P(\bar{X} > c) = P\left(\frac{\bar{X} - 2.5}{\sqrt{\frac{S^2}{n}}} > \frac{c - 2.5}{\sqrt{\frac{S^2}{n}}}\right) = 0.05$$

# The critical point and the rejection region

$$P\left(\frac{\bar{X} - 2.5}{\sqrt{\frac{S^2}{n}}} > \frac{c - 2.5}{\sqrt{\frac{S^2}{n}}}\right) = 0.05$$

$$P\left(Z > \frac{c - 2.5}{\sqrt{\frac{S^2}{n}}}\right) = 0.05$$

**The test statistic**

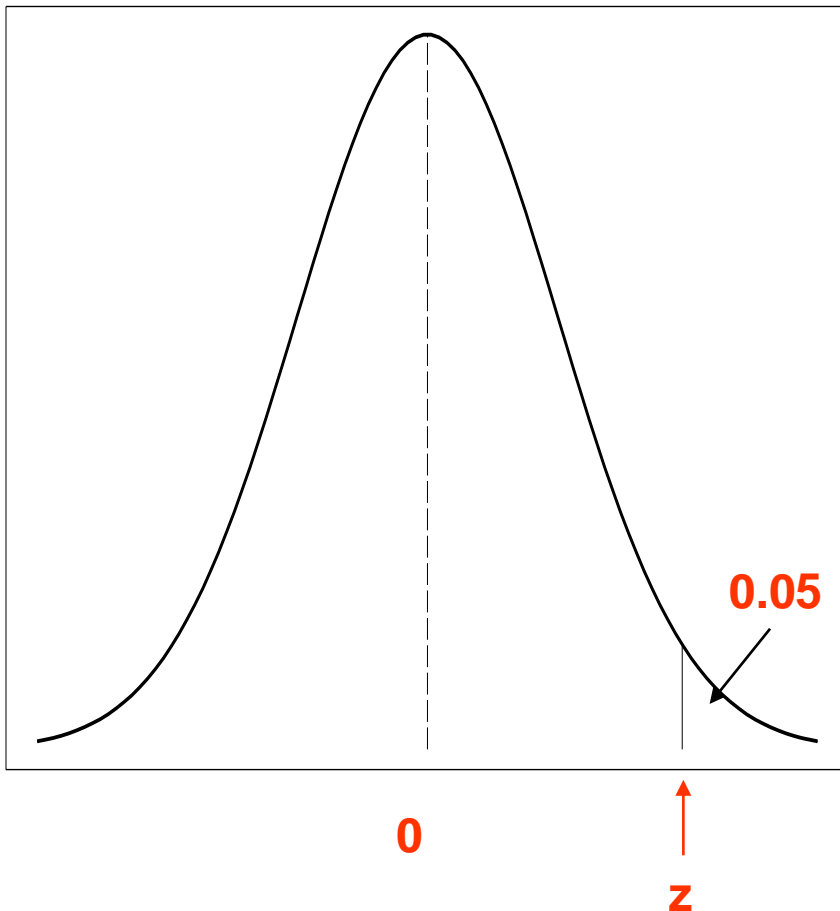
The sample mean  $\leftarrow \frac{\bar{x} - \mu_0}{\sqrt{\frac{S^2}{n}}} \rightarrow$  Population mean under  $H_0$

$\sqrt{\frac{S^2}{n}}$   $\rightarrow$  Standard error of the sample mean.



# The critical point and the rejection region

$N(0,1)$



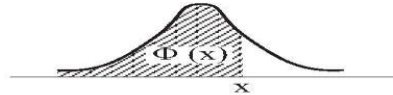
We choose the significance level  $\alpha = 0.05$ .

This means that we reject the null hypothesis with probability 0.05 if it is correct, or with a probability of 95% the null hypothesis will be acceptable (not reject) if they are indeed correct.

This leads to the determination of a critical point  $z$

$$P(Z < z) = 0.95$$

Tabel 3 : Standaard normale verdeling



	$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$									
x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

$$P(Z < 1.645) = 0.95$$

$$P(Z > 1.645) = 0.05$$

In R:

```
> qnorm(0.95,0,1)
[1] 1.644854
```

## The critical point and the rejection region

$$P\left(Z > \frac{c - 2.5}{\sqrt{\frac{S^2}{n}}}\right) = 0.05$$

$$P(Z > 1.645) = 0.05$$

$$\frac{c - 2.5}{\sqrt{\frac{S^2}{n}}} = 1.645$$

## The critical point and the rejection region

$$\frac{c - 2.5}{\sqrt{\frac{S^2}{n}}} = 1.645 \quad \longrightarrow \quad c = 1.645 \times \sqrt{\frac{S^2}{n}} + 2.5$$

From the Sample (see page 91):

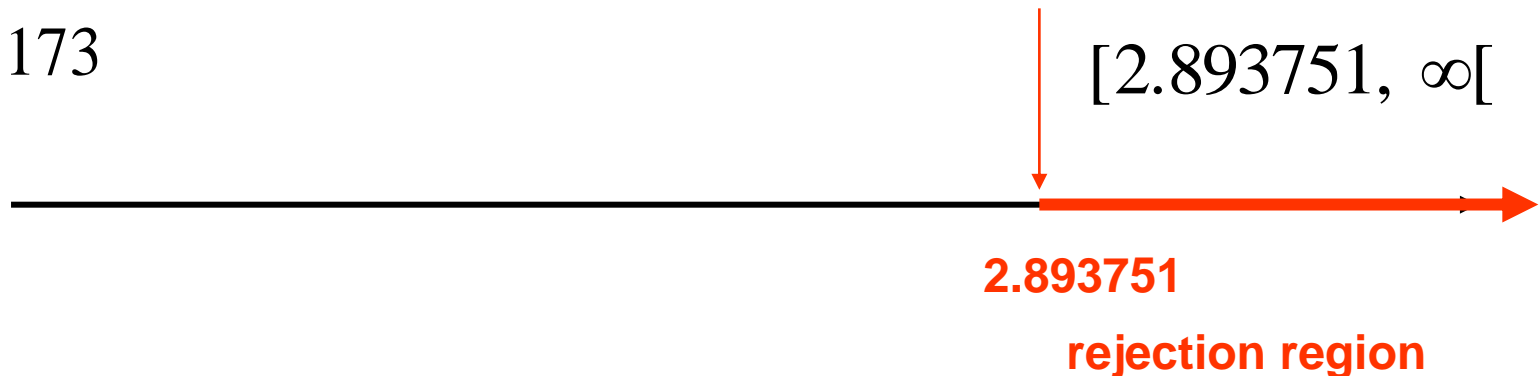
$$s^2 = 9.9119$$

$$\bar{x} = 2.91$$

$$n = 173$$

$$\longrightarrow c = 1.645 \times \sqrt{\frac{9.9119}{173}} + 2.5 = 2.893751$$

$$[2.893751, \infty[$$



# The critical point and the rejection region

$$\frac{\bar{x} - 2.5}{\sqrt{\frac{S^2}{n}}} > \frac{c - 2.5}{\sqrt{\frac{S^2}{n}}} \Rightarrow \dots$$

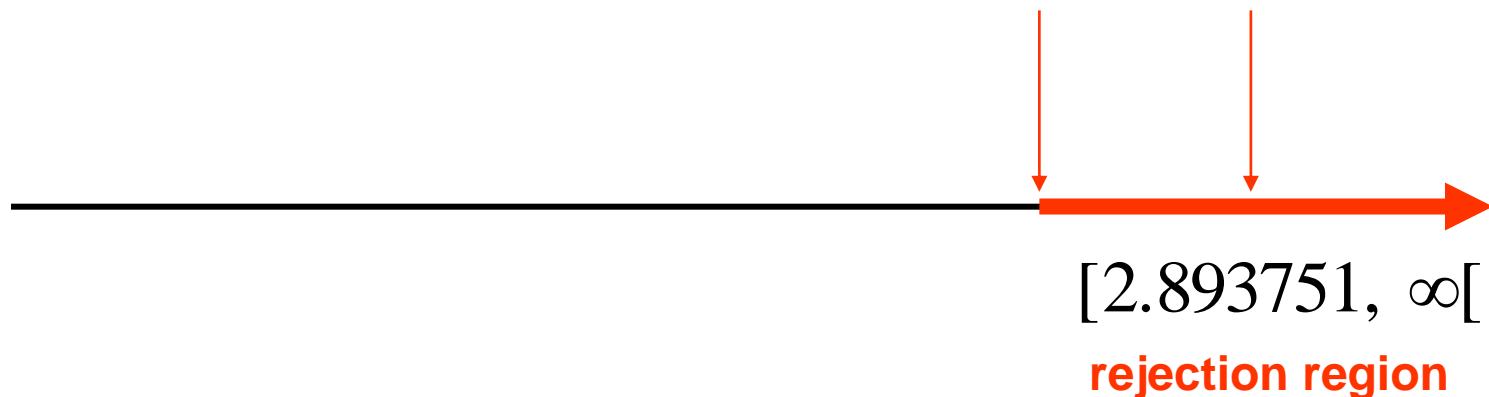
$$c = z \times \sqrt{\frac{S^2}{n}} + \mu_0$$

$$\bar{x} = 2.91 > 2.893 \dots \Rightarrow H_0 \quad \text{reject}$$

**c**

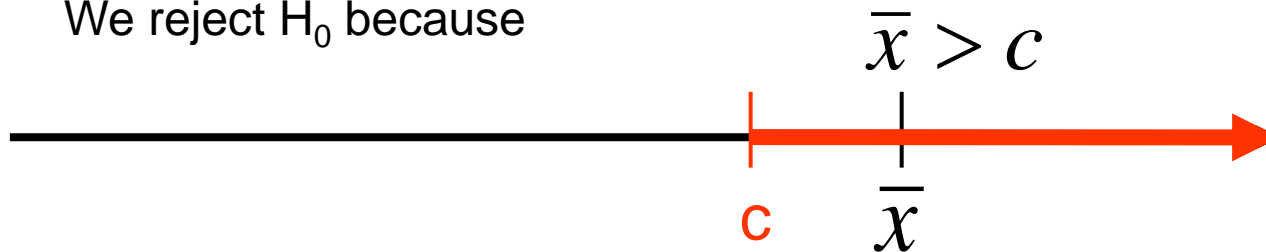
**2.893751**

$\bar{x} = 2.91$



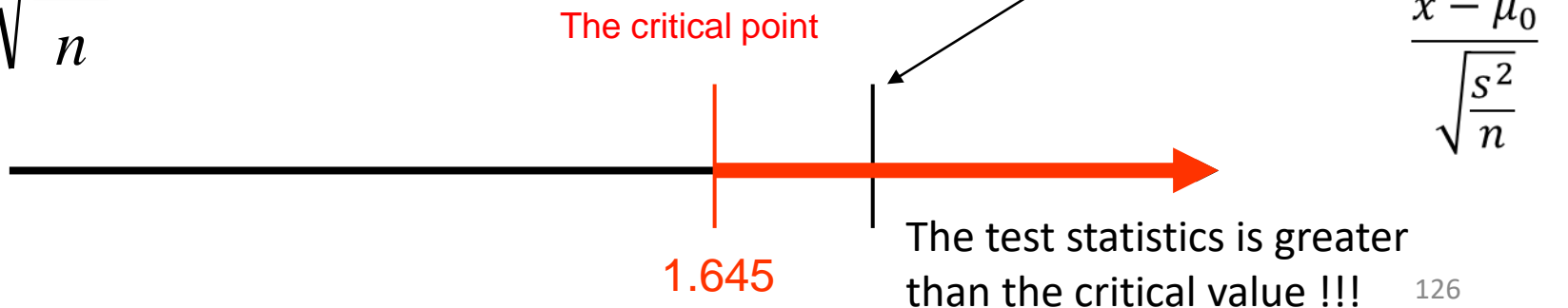
# The critical point and the test statistic

We reject  $H_0$  because




We can use the value of the **test statistic** comparing the critical point

$$\frac{\bar{x} - 2.5}{\sqrt{\frac{S^2}{n}}} > 1.645$$



# the checklist

Step	information	example
1	The hypotheses (the qualifying problem)	$H_0 : \mu = 2.5$ $H_1 : \mu = 2.9$
2	The distribution in the population & $\sigma^2$	$X \sim unknown$ $\sigma^2$ not known
3	sample size	$n = 173 > 30$ 
4	The distribution of the sample mean under the null hypothesis	$\bar{X} \sim N\left(2.5, \frac{S^2}{n}\right)$
5	The level of significance	$\alpha = 0.05$
6	The test statistic	<div style="border: 1px solid red; padding: 10px; display: inline-block;"> <math display="block">\frac{\bar{X} - 2.5}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)</math> </div>
7	The distribution of the review greatness under the null hypothesis	
8	The critical point	1.645    N(0,1)

# Examples by R

```
> xbar=2.91;s=sqrt(9.9119)
> n=173
> H0=2.5
> cc=qnorm(0.95)*(s/sqrt(n))+H0
> cc
[1] 2.893716
```

$$\left\{ \begin{array}{l} c = z \times \sqrt{\frac{S^2}{n}} + \mu_0 \\ c = 1.645 \times \sqrt{\frac{9.9119}{173}} + 2.5 = 2.893751 \end{array} \right.$$

We reject  $H_0$


$$\bar{x} > c$$

$$\bar{x} = 2.91 > 2.893 \dots \Rightarrow \text{reject } H_0$$



## By using critical point

```
> xbar=2.91;s=sqrt(9.9119);n=173;H0=2.5  
> test.stat=(xbar-H0)/(s/sqrt(n))  
> test.stat  
[1] 1.712886  
> crit.point=qnorm(0.95)  
> crit.point  
[1] 1.644854
```


$$\frac{\bar{x} - 2.5}{\sqrt{\frac{9.9119}{173}}} = 1.712886$$

We reject  $H_0$

$$t = 1.71 > 1.644.. = t_{\alpha}$$



Test statistic



Critical value

# Example: the air quality data

The estimators for the unknown parameter ( $\mu, \sigma$ ) in the population

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

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

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

Point estimators of the population parameters in R:

```
> ozone=airquality$Ozone
> meanozone=mean(ozone, na.rm=T)
> meanozone
[1] 42.12931

> varozone=var(ozone, na.rm=T)
> varozone
[1] 1088.201

> sqrt(varozone)
[1] 32.98788
```

# Example: the air quality data

- Mean ozone in parts per billion from 1300 to 1500 hours at Roosevelt Island.

```
> z.test(ozone, sd=s, mu=40)  Testing the null hypothesis that mu =40
```

One Sample z-test

$$\frac{\bar{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}$$

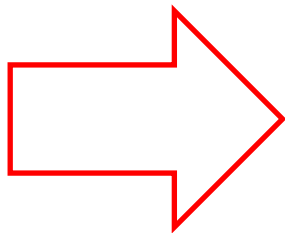
data: ozone  
**z = 0.6952**, n = 116.000, Std. Dev. = 32.988, Std. Dev. of the sample mean = 3.063, p-value = 0.4869  
alternative hypothesis: true mean is not equal to 40  
95 percent confidence interval:  
36.12624 48.13238  
sample estimates:  
mean of ozone  
42.12931

```
>  
sqrt(1088.201)  
[1] 32.98789
```

## Case study 1b:

The airquality data: analysis of the  
average wind speed

Test of hypothesis about the  
population mean (two sided test)



Page 193



## Part 7: Inference for one-sample means with the $t$ distribution

### Section 5.1: Single-sample inference 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.



## 7.1: hypothesis testing using a t distribution

# t-test for a population

- We assume that  $X \sim N(\mu, \sigma^2)$  &  $n$  is small
- For this test, we used the Student t distribution.

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

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

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

X has a normal distribution with unknown  $\mu$  and  $\sigma^2$ .  $n$  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

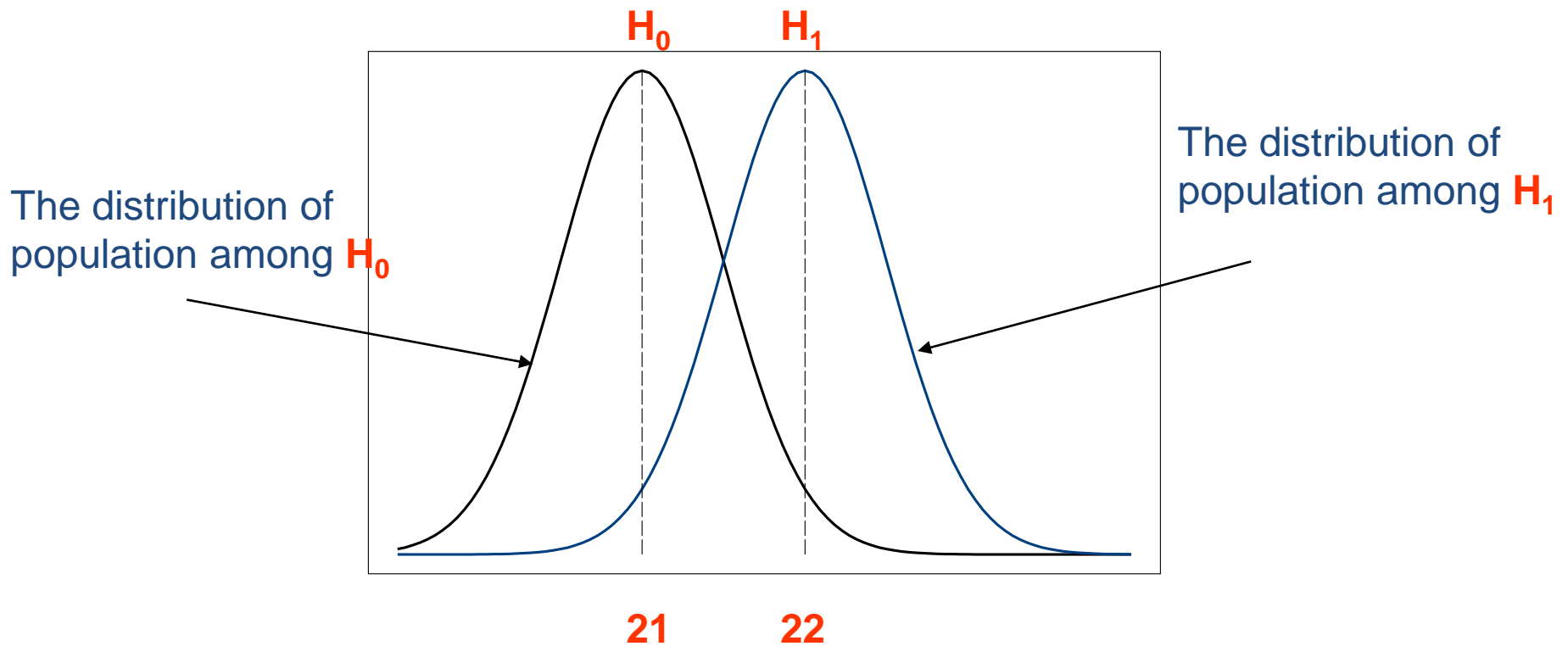
$$X \sim N(\mu, \sigma^2)$$



# The distribution of the population

$$X \sim N(21, \sigma^2) \quad \text{under } H_0$$

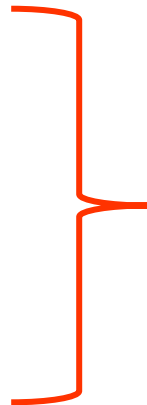
$$X \sim N(22, \sigma^2) \quad \text{under } H_1$$



# 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  $\mu$  and  $\sigma^2$ .
- $n$  is small

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



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

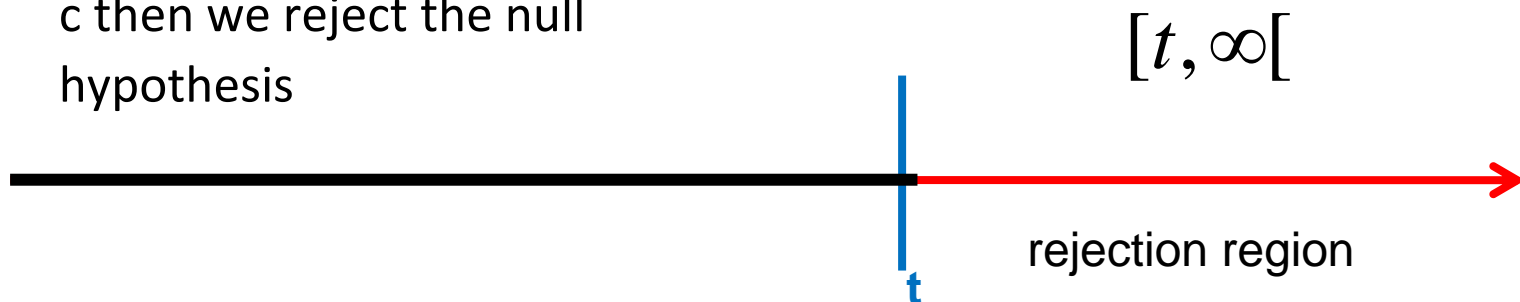
The distribution of the  
test statistic population  
under  $H_0$

# The rejection region

$H_0 : \mu = 21$   
 $H_1 : \mu = 22$   $\Rightarrow \mu_0 < \mu_1 \Rightarrow$  when the value of  $\bar{x}$  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



# The choice of c

- We choose c so that Type I error 0.05.

$$\alpha = 0.05$$

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

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

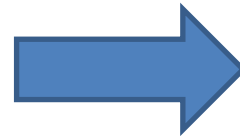
if the null hypothesis is  
correct

## The choice of c

Determine c so that Type I error =5%

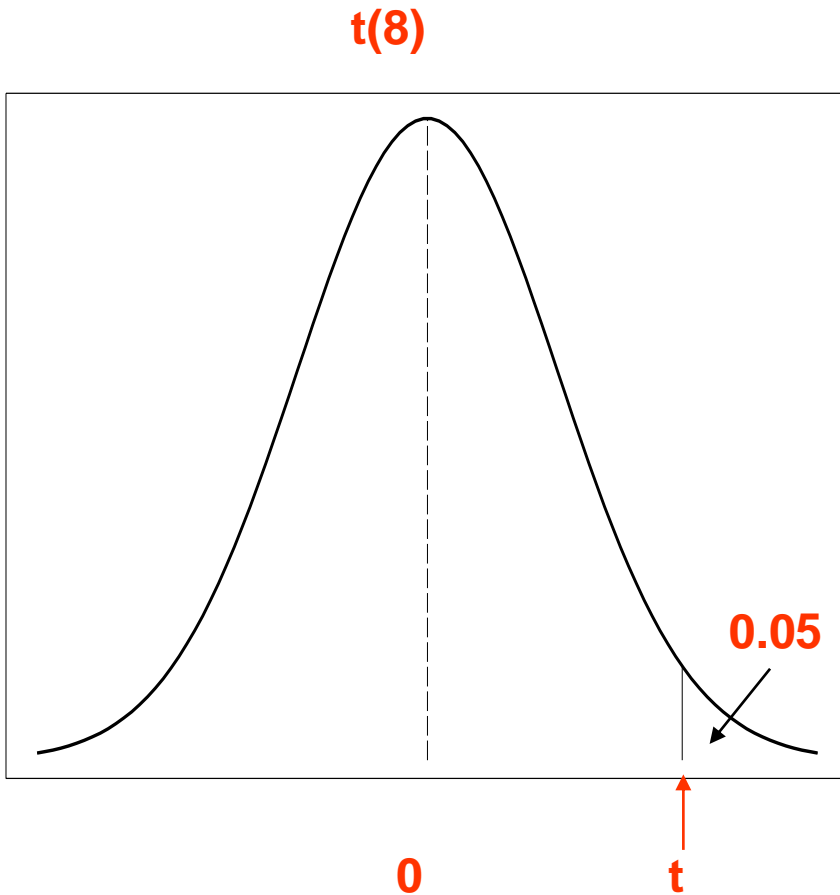
$$P(\bar{X} > c) = 0.05$$

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



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

# The critical point



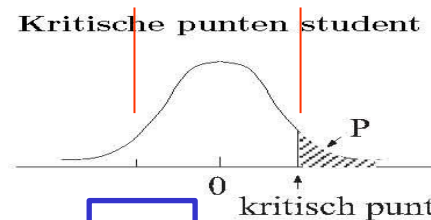
The distribution of the test statistic under  $H_0$

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

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

# Student's t-distribution and critical point

Tabel 4 : Kritische punten student t verdeling



P	.25	.10	.05	.025	.010	.005	.001
v.g.							
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

$$n = 9 (small)$$

$$df. = 8$$

$$\alpha = 0.05$$

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

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

# The sample

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

$n=9$

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

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

The estimators  
for the  
unknown  
parameters ( $\mu$   
and  $\sigma^2$ ) 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

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

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

$$\frac{c-21}{\sqrt{\frac{S^2}{n}}} = 1.86$$



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



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

# The rejection region

$$s^2 = 3.972$$

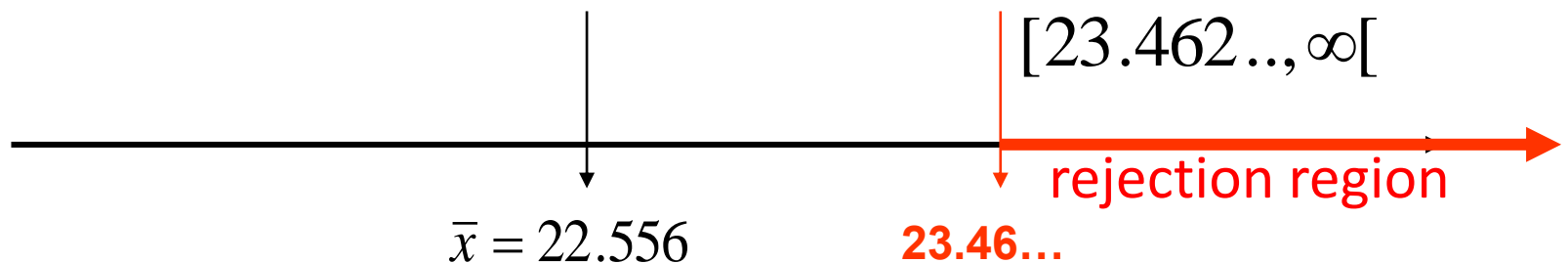
$$\bar{x} = 22.556$$

$$n = 9$$




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

$\bar{x} < c \Rightarrow$  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 $\sigma^2$	$X \sim N(\mu, \sigma^2)$ $\sigma^2$ 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	 <div style="border: 1px solid red; padding: 10px; display: inline-block;"> <math display="block">\frac{\bar{X} - 21}{\sqrt{\frac{S^2}{n}}} \sim t(8)</math> </div>
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:

## The women data:

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

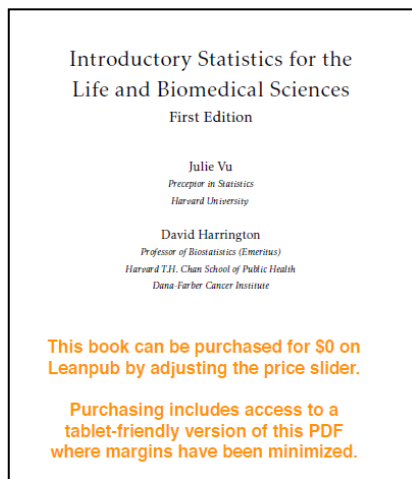
$$\begin{array}{l} H_0: \mu = 60 \\ H_1: \mu \neq 60 \end{array}$$

```
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

## 7.2: One sided and two-sided testing problems



### Section 4.3.1 The Formal Approach to Hypothesis Testing

# 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  $H_0$

$$H_0 : \mu = \mu_{H_0}$$

null hypothesis

$$H_1 : \mu < \mu_{H_0}$$

alternative  
hypothesis

One sided test problem

## Case (b)

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

$$H_0 : \mu = \mu_{H_0}$$

null hypothesis

$$H_1 : \mu > \mu_{H_0}$$

alternative  
hypothesis

One sided test problem

## Case (c)

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

$$H_0 : \mu = \mu_{H_0}$$

null hypothesis

$$H_1 : \mu \neq \mu_{H_0}$$

alternative  
hypothesis

two sided test problem

## Example (case a)

$$H_0 : \mu = \mu_{H_0}$$

$$H_1 : \mu < \mu_{H_0}$$

One sided test

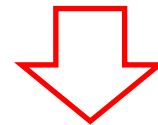
## Example: one-tailed test

- 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:

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

$$n=100$$

The variance  $\sigma^2$  is unknown but large n.



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

# 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

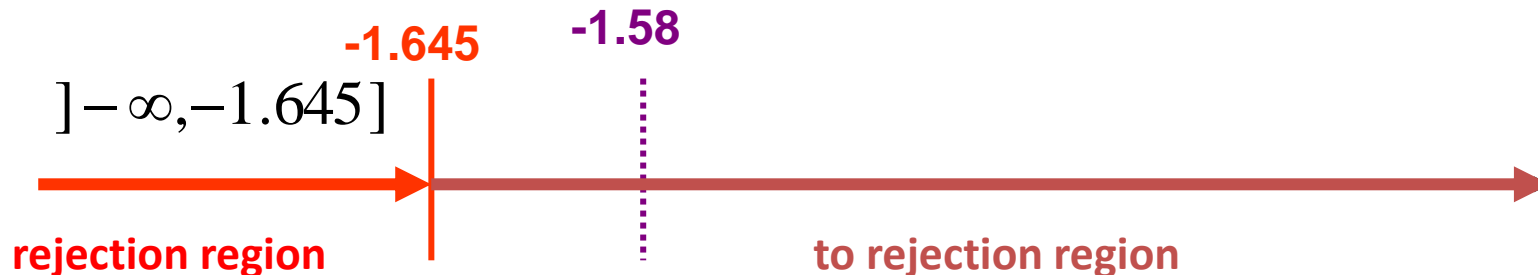
One-sided test

# 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

Step	information	example
1	The hypotheses (the qualifying problem)	One-sided test
2	The distribution in the population & $\sigma^2$	
3	sample size	
4	The distribution of the sample mean under $H_0$	
5	The level of significance	
6	The test statistic	$\frac{\bar{X} - 51}{\sqrt{\frac{S^2}{n}}} \sim ?$
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)

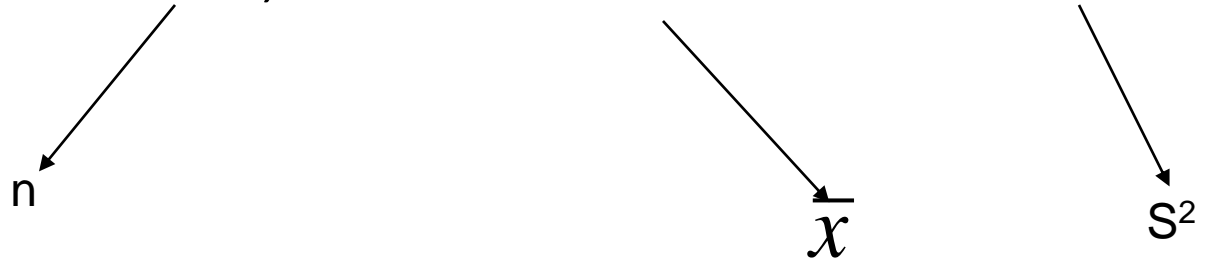
$$H_0 : \mu = \mu_{H_0}$$

$$H_1 : \mu \neq \mu_{H_0}$$

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 ( $115 = \mu$  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.



# 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

Step	information	Example
1	The hypotheses (the testing problem)	One-sided test
2	The distribution in the population & $\sigma^2$	
3	sample size	
4	The distribution of the sample mean under $H_0$	
5	The level of significance	
6	The test statistic	$\frac{\bar{X} - 115}{\sqrt{\frac{S^2}{n}}} \sim ?$
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, $n$ & $\sigma^2$

- 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  $\sigma^2$  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

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

If  $\sigma^2$  is known

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

If  $\sigma^2$  is unknown

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

## 2: n small & normal distribution

2: n Small & normal distribution

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

If  $\sigma^2$  is known

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

If  $\sigma^2$  is unknown

# The choice of the setting

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

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



## 7.3 The P-value

### Section 4.3.1: The Formal Approach to Hypothesis Testing

Introductory Statistics for the  
Life and Biomedical Sciences  
First Edition

Julie Vu  
*Professor 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.

# The significance level and the critical point

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

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

# The level of significance and the p-value

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

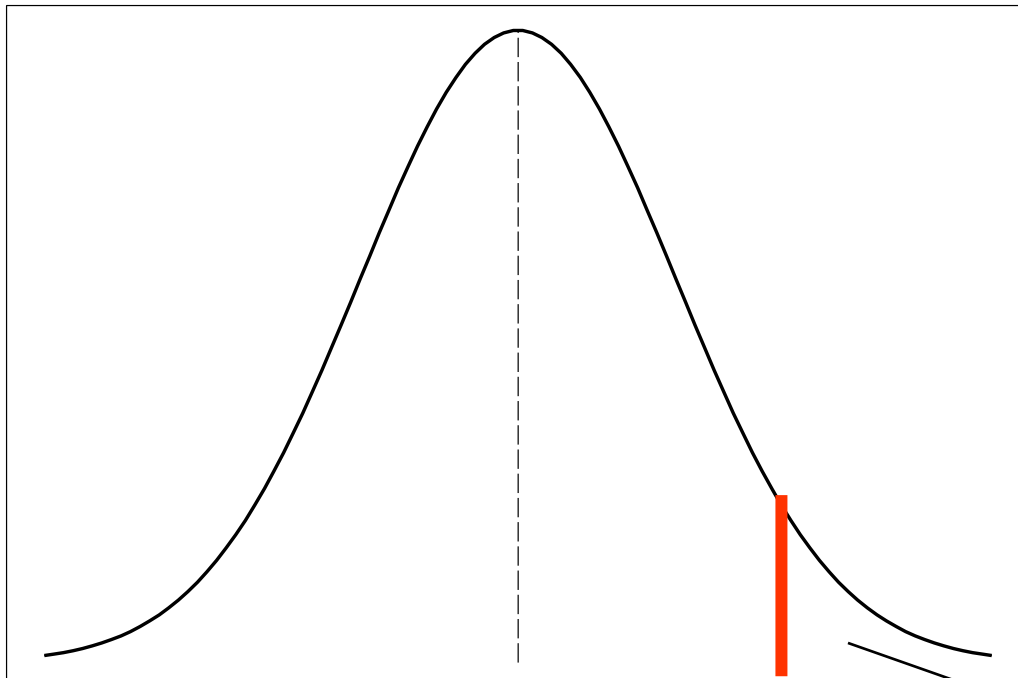
*$H_0$  is rejected on  
significance level  $\alpha$  if and only if  
the  $p$ -value  $< \alpha$*

# Right-sided test

$$\mu_0 < \mu_1$$

$$[c, \infty[$$

The distribution of  
the test statistic  
under  $H_0$



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

p-value

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

population

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

$$n = 9 \quad (\text{small})$$

$$\sigma^2 : \text{unknown}$$

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

$$p\text{-value} = P(T > 1.175227) = 0.1481026$$

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 }
```

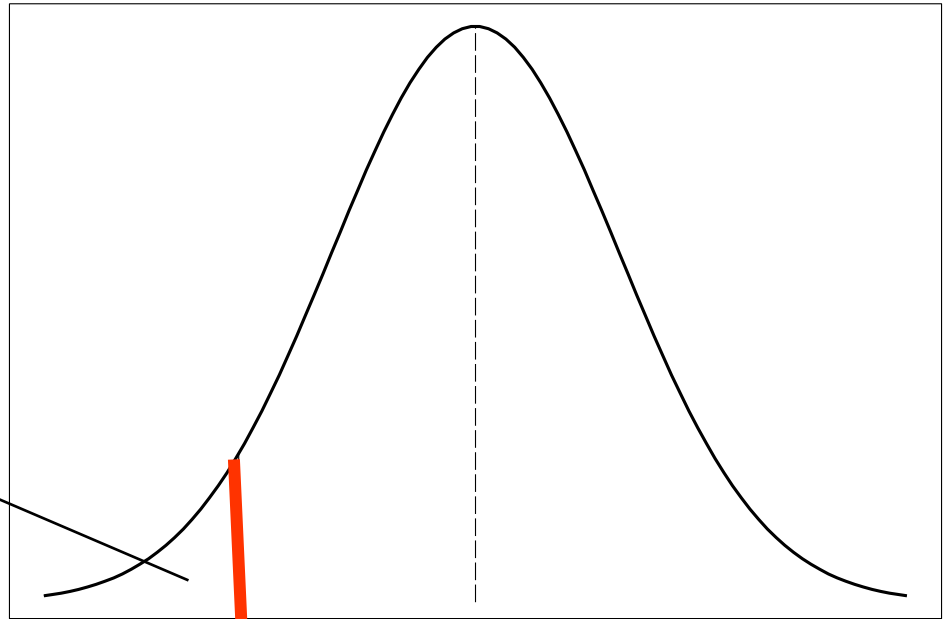


P value

Left-sided test  $\mu_0 > \mu_1$

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

The p-value



observed value of

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

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

$$H_0 : \mu = 51$$

$$H_1 : \mu < 51$$



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

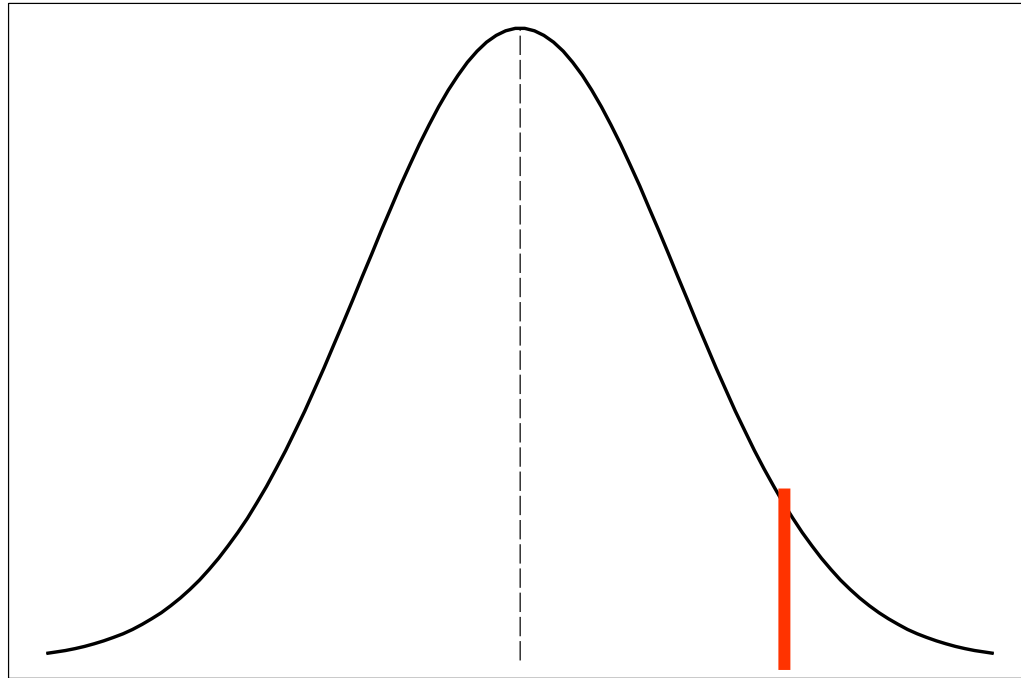
$$p\text{-value} = P(Z < -1.58) = 0.0571$$

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

$$\frac{\bar{x} - \mu_{H_0}}{\sqrt{\frac{\sigma^2}{n}}}$$

of

$$\frac{\bar{x} - \mu_{H_0}}{\sqrt{\frac{S^2}{n}}}$$

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

### Example 3 (p-value)

$$H_0 : \mu = 115$$

$$H_1 : \mu \neq 115$$



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

$$p\text{-value} = 2 \times P(Z > 2.14) = 2 \times [1 - \Phi(2.14)] = 0.0324$$

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

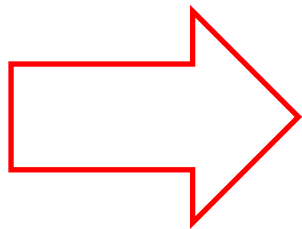
- Statistical computer packages give as output a hypothesis test the p-value.
- A generally accepted criterion (e. g in scientific publications) is as follows
  1. If the P-value  $< 0.05$ , then  $H_0$  is rejected, and then the results are significant.
  2. if the p-value  $> 0.05$ , then  $H_0$  is not rejected, and then the results are not significant.



## Case study 3:

The The NHANES dataset: number of sleep  
hours per night

Test of hypothesis about the  
population mean (one sided test)



Page 203

# Case studies

Examples from the online book

## Case study 1a:

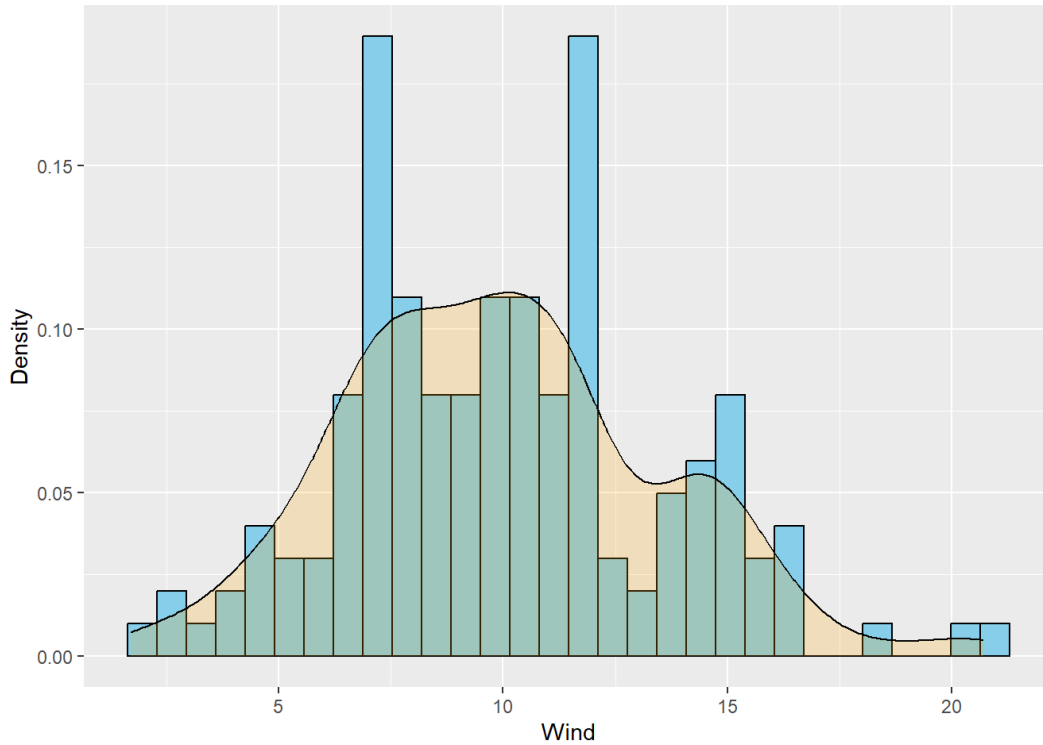
The `airquality` data: analysis of the  
average wind speed

Confidence interval for the  
population mean

# The average wind speed per day

- The `airquality` dataset gives information about 153 daily air quality measurements in New York, May to September 1973.
- The variable `Wind` is the average wind speed in miles per hour at 0700 and 1000 hours at LaGuardia Airport.

# The average wind speed per day



$$n = 153$$

$$\bar{x} = 9.55$$

$$s = 3.52$$

# Distribution of the test statistic

## Case 2

If  $X \sim F$

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

and  $T_X = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)$

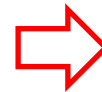
3. X has an unknown distribution, but we have a **large sample** ( $n > 30$ )

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

Our example:

$$n=153$$

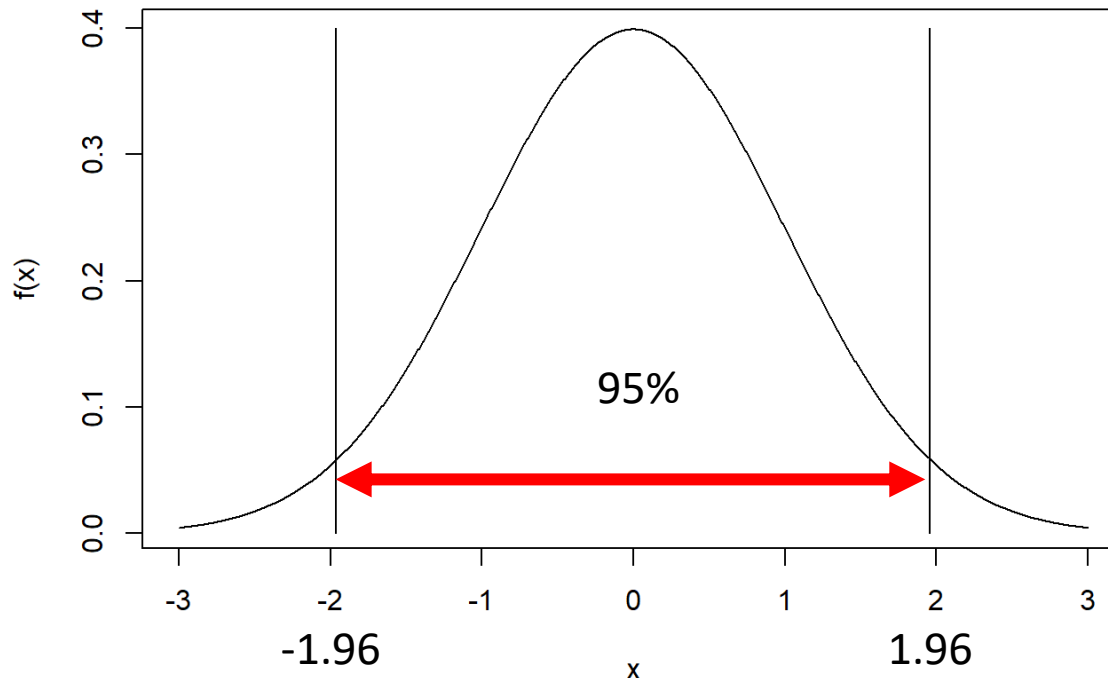


Use  $N(0,1)$  to choose the critical values for the Upper and Lower limits.

The same as case 1 but we replace  $\sigma^2$  by  $S^2$ .

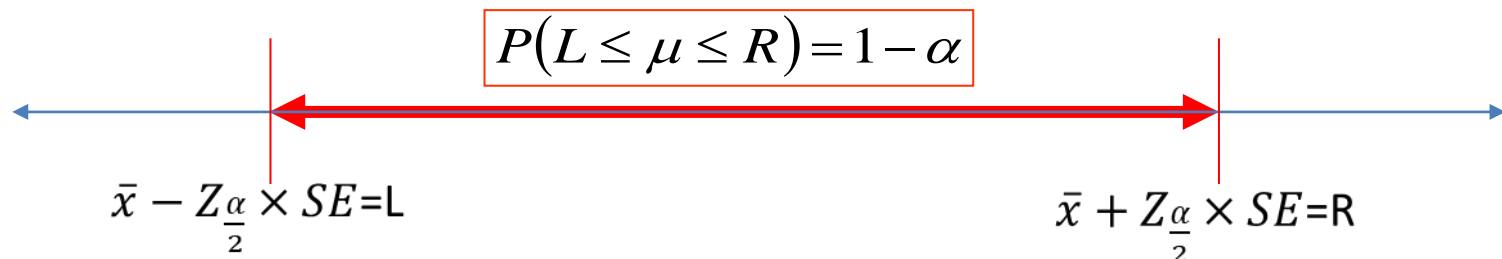
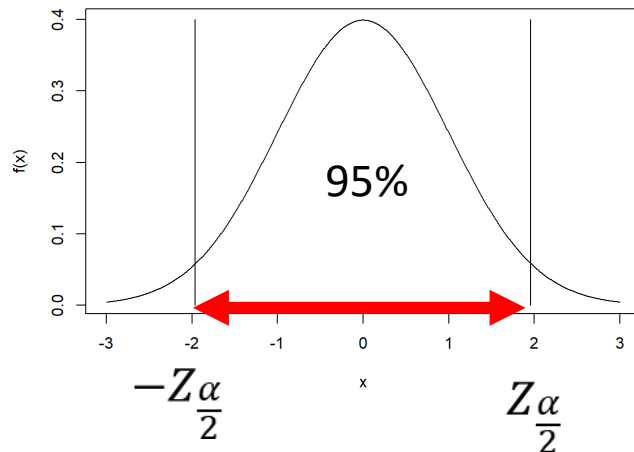
# Critical values from $N(0,1)$

- For  $\alpha=0.025$ , we are looking for two values for which the probability to be between these value is  $0.95=(1-2\alpha)$ .



# Lower (L) and upper (R) limits

- For a given value of  $\alpha$ , we are looking for two values for which the probability to be between these value is  $0.95=(1-\alpha)$ .






# The average wind speed per day

- A 95% confidence interval:

$$(\bar{x} - m, \bar{x} + m)$$

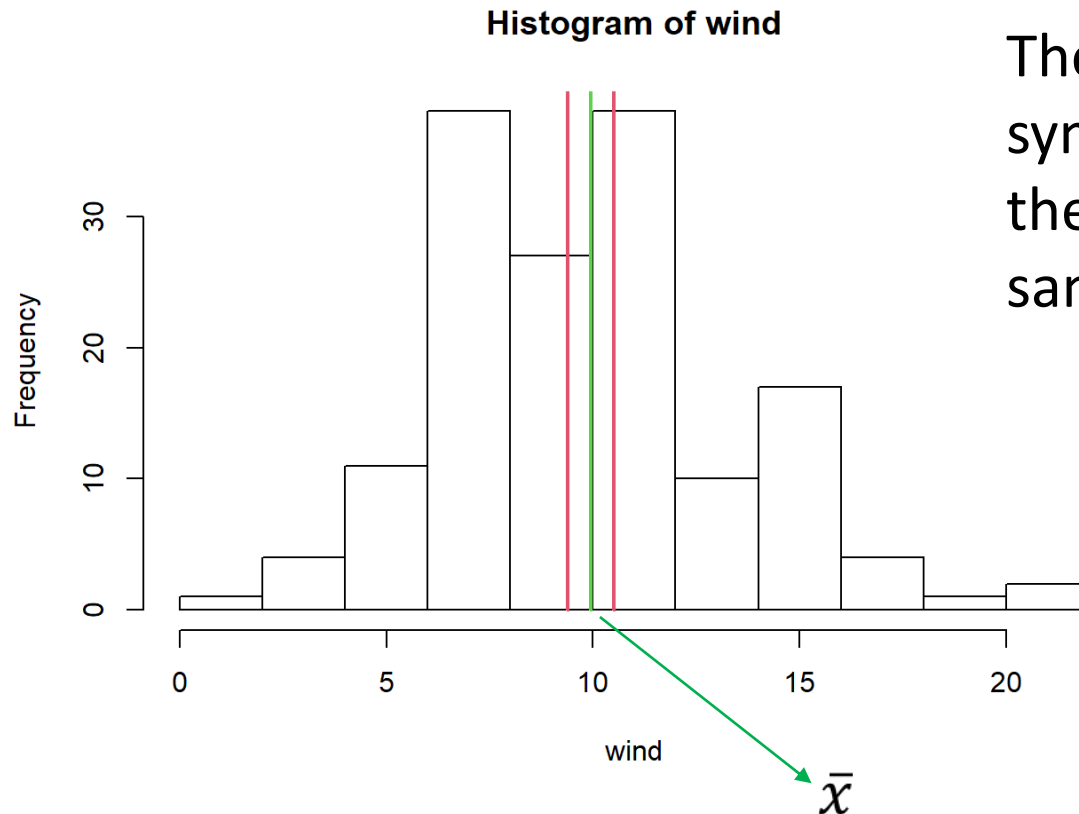
$$m = Z_{\alpha} \times SE = Z_{\alpha} \times \frac{s}{\sqrt{n}} = 1.96 \times \frac{3.52}{\sqrt{153}} = 0.5582$$

$$(\bar{x} - m, \bar{x} + m) = (9.391, 10.508)$$



SE of the  
sample mean.

# 95% confidence interval



The interval is symmetric around the value of the sample mean.

$$9.391 = \bar{x} - Z_{\frac{\alpha}{2}} \times SE = L \quad \boxed{P(L \leq \mu \leq R) = 1 - \alpha} \quad \bar{x} + Z_{\frac{\alpha}{2}} \times SE = R = 10.508$$

## Case study 1b:

The `airquality` data: analysis of the  
average wind speed

Test of hypothesis about the  
population mean (two sided test)

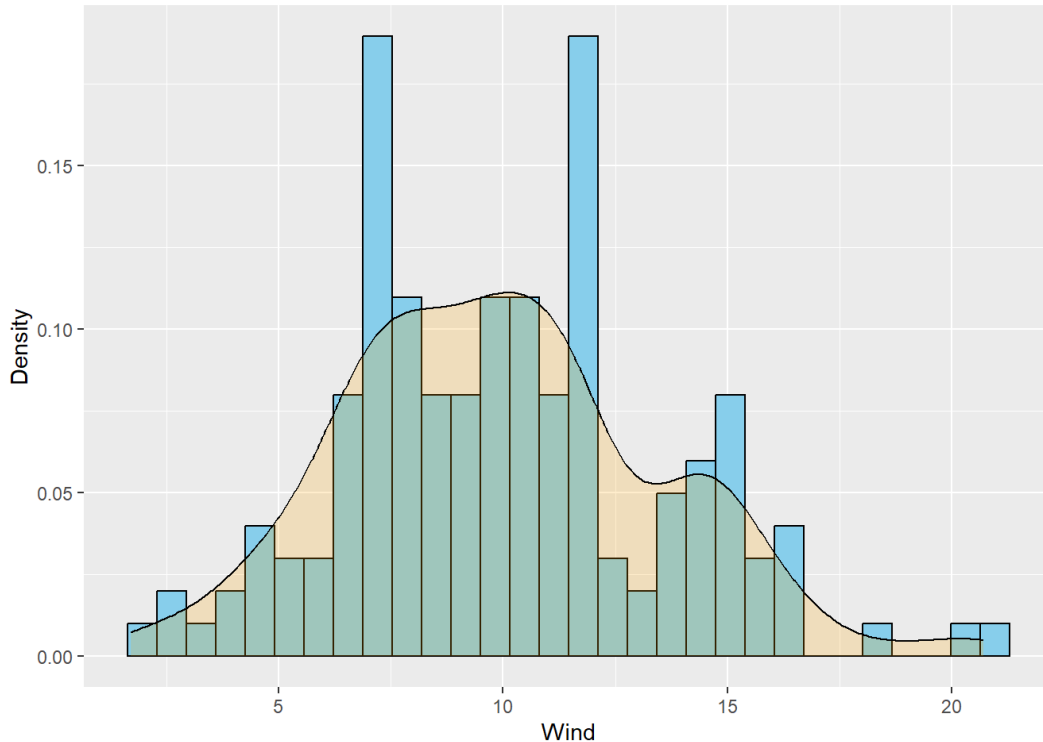
# The average wind speed per day

$$H_0: \mu = 9$$

$$H_A: \mu \neq 9$$

- We test the null hypothesis that the population mean is equal to 9.
- The alternative hypothesis: the mean is not equal to 9 (we do not specify a value).
- For the analysis, we assume that the variance in the population is known.

# The average wind speed per day



- Sample size:  
 $n = 153$
- Point estimators in the sample:  
 $\bar{x} = 9.55$   
 $s = 3.52$
- We assume that in the population:  
 $\sigma = 3.52$

# Distribution of the test statistic

## Case 2

If  $X \sim F$

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

and  $T_X = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)$

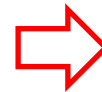
3. X has an unknown distribution, but we have a **large sample** ( $n > 30$ )

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

Our example:

$$n=153$$



Use  $N(0,1)$  to choose the critical value.

The same as case 1 but we replace  $\sigma^2$  by  $S^2$ .

# Test statistic

$n=153$ , under  $H_0$  the distribution of the test statistic:

$$t = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Test statistic:

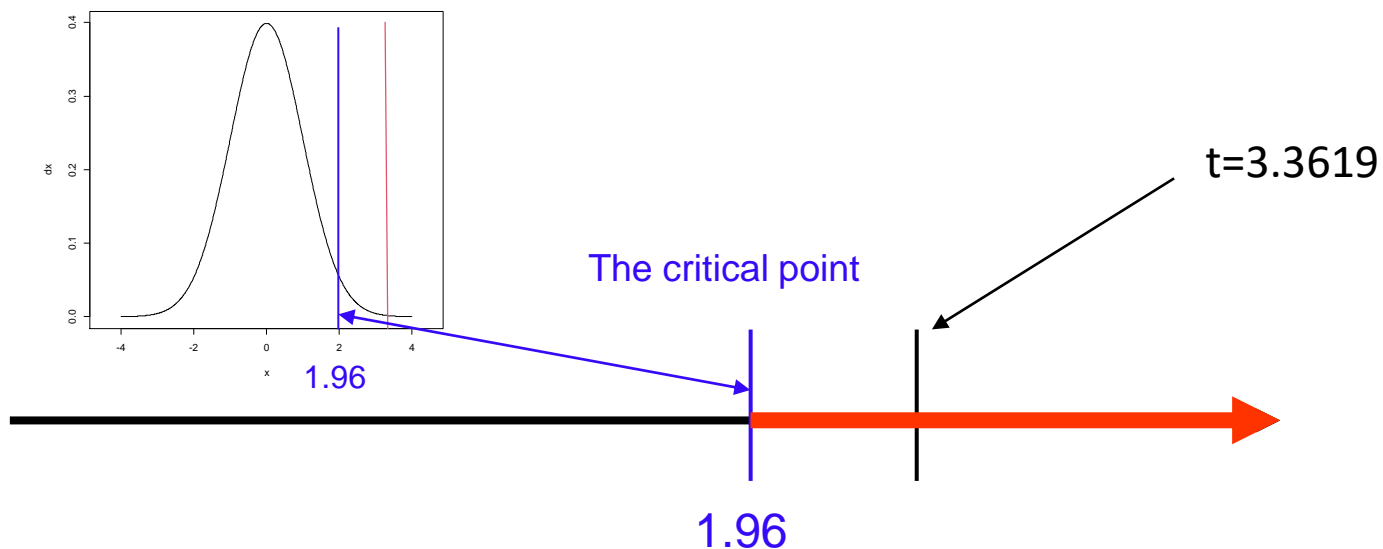
$$t = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{9.95 - 9}{\frac{3.523}{\sqrt{153}}} = 3.3619$$

# The critical point and the test statistic

We can use the value of the test statistic comparing the critical point.

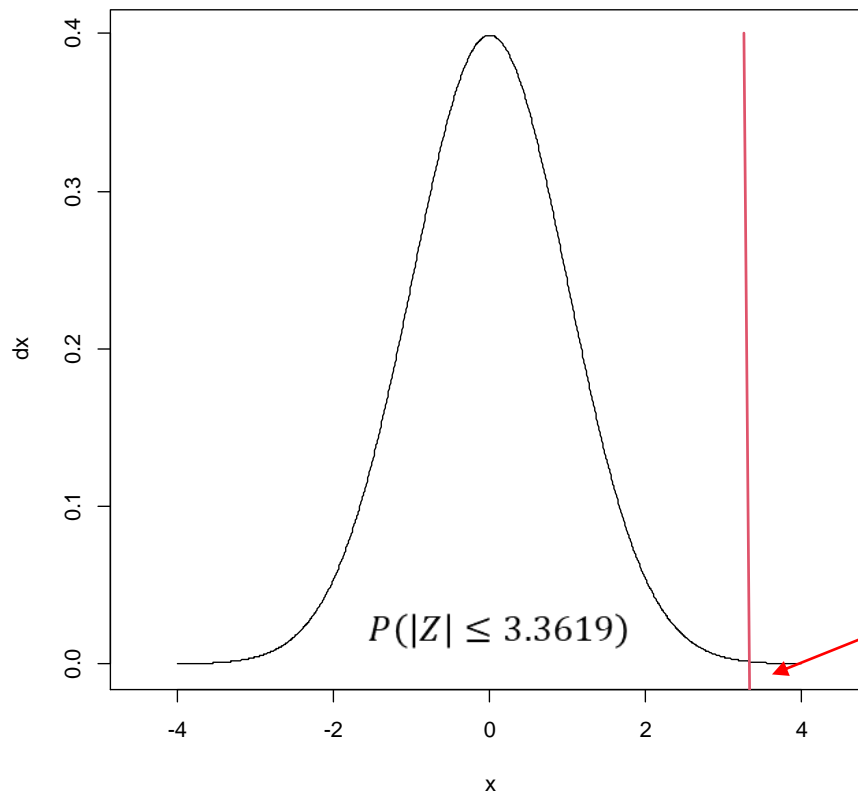
For **two sided** test and  $\alpha=0.05$ ,  $Z_{0.975}=1.96$ .

We reject  $H_0$  :  $3.3619 > 1.96$ .





# p-value



$$2 \times (1 - P(|Z| \leq 3.3619))$$

$$2 \times 0.00038 = 0.000775 < 0.05 = \alpha$$

- For  $\alpha=0.05$ , p-value  $< 0.05$ .
- We reject the null hypothesis and conclude that the mean in the population is not 9.

$$P(|Z| > 3.3619)$$

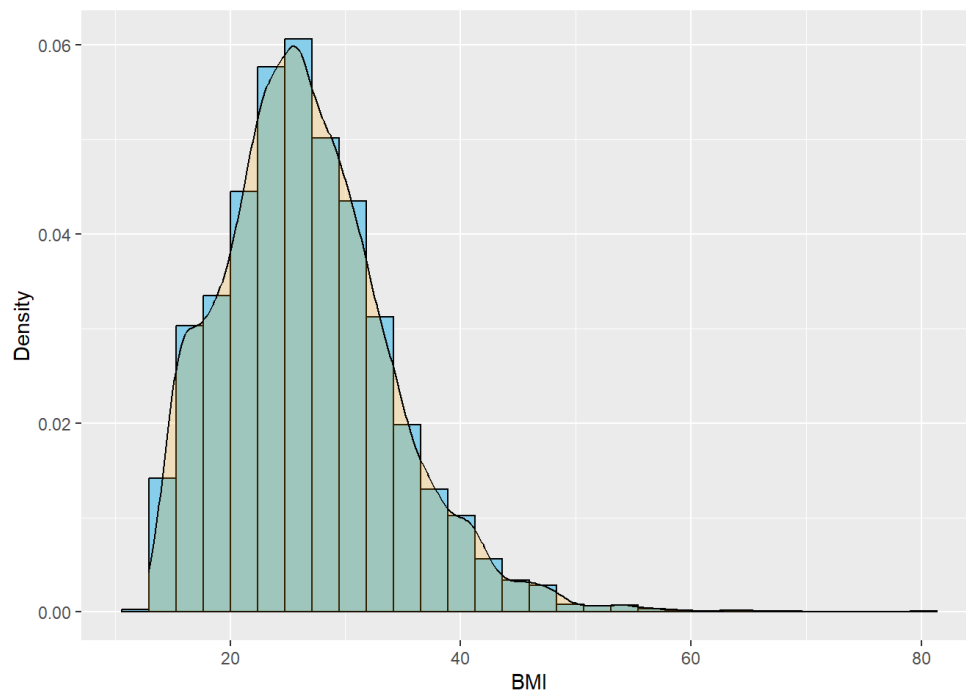
## Case study 2:

### The The NHANES dataset: BMI

# The The NHANES dataset: BMI

- The NHANES dataset consists of data from the US National Health and Nutrition Examination Study.
- Information about 76 variables is available for 10000 individuals included in the study.
- The 10000 individuals are considered as the population.
- Some individuals excluded due to missing values.
- In this part we focus on the BMI (the R object BMI).

# The The NHANES dataset: BMI

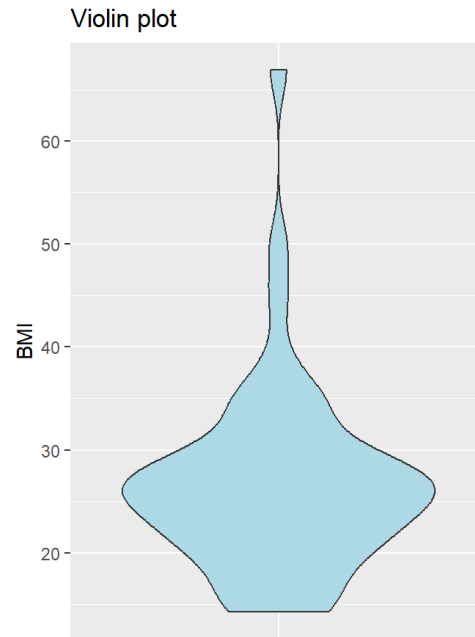
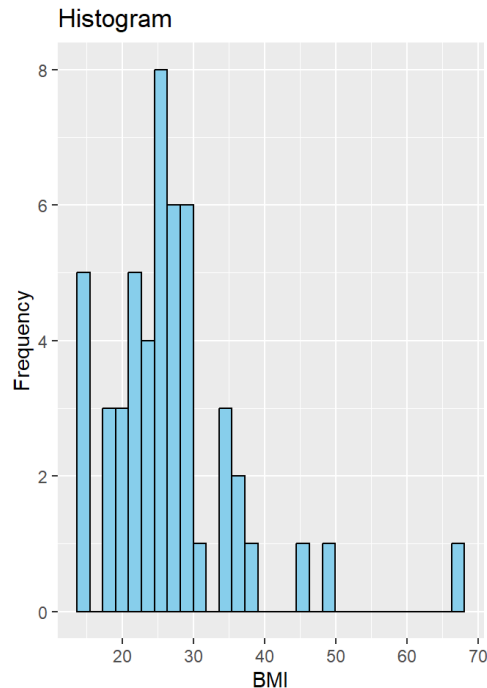


- The population:
  - N=9634.
  - Population mean:
- Population variance:

$$\mu = 26.76$$

$$\sigma^2 = 54.41$$

# The The NHANES dataset: BMI



- A sample of 50 individuals from the population.

- Sample mean:

$$\bar{x} = 26.76$$

- Sample variance:

$$s^2 = 87.67$$

# Distribution of the test statistic

## Case 2

If  $X \sim F$

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

and  $T_X = \frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)$

3. X has an unknown distribution, but we have a **large sample** ( $n > 30$ )

$$E(X) = \mu$$

$$Var(X) = \sigma^2$$

Our example:

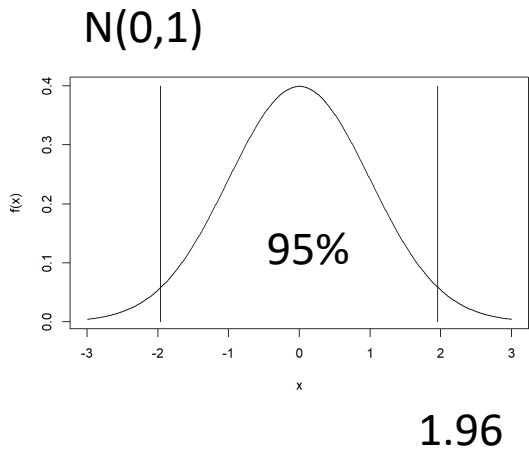
$$n=50$$



Use  $N(0,1)$  to choose the critical value for the upper and lower limits.

The same as case 1 but we replace  $\sigma^2$  by  $S^2$ .

# 95% C. I. for the mean BMI



$$(\bar{x} - m, \bar{x} + m)$$

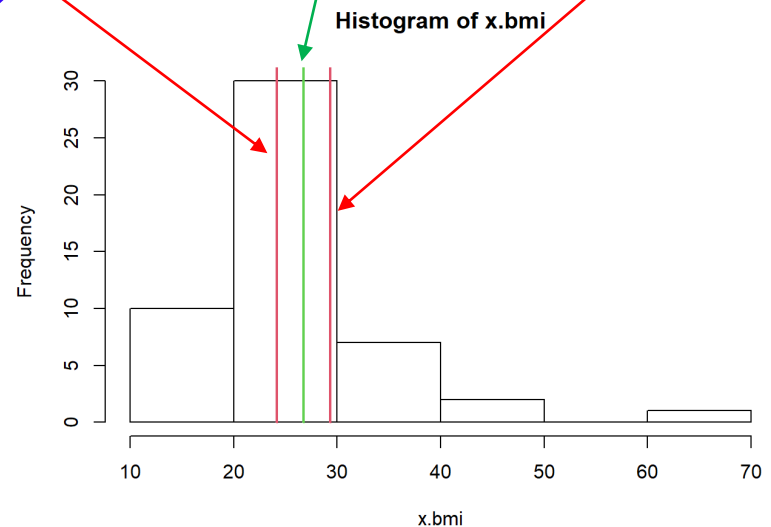
$$\left( \bar{x} - 1.96 \times \sqrt{\frac{87.67}{50}}, \bar{x} + 1.96 \times \sqrt{\frac{87.67}{50}} \right)$$

$$\alpha = 0.025$$

$$Z_{\alpha} = 1.96$$

$$(1 - 2\alpha) = 0.95$$

The standard error of the sample mean



Case study 3:  
The The NHANES dataset: number of sleep  
hours per night

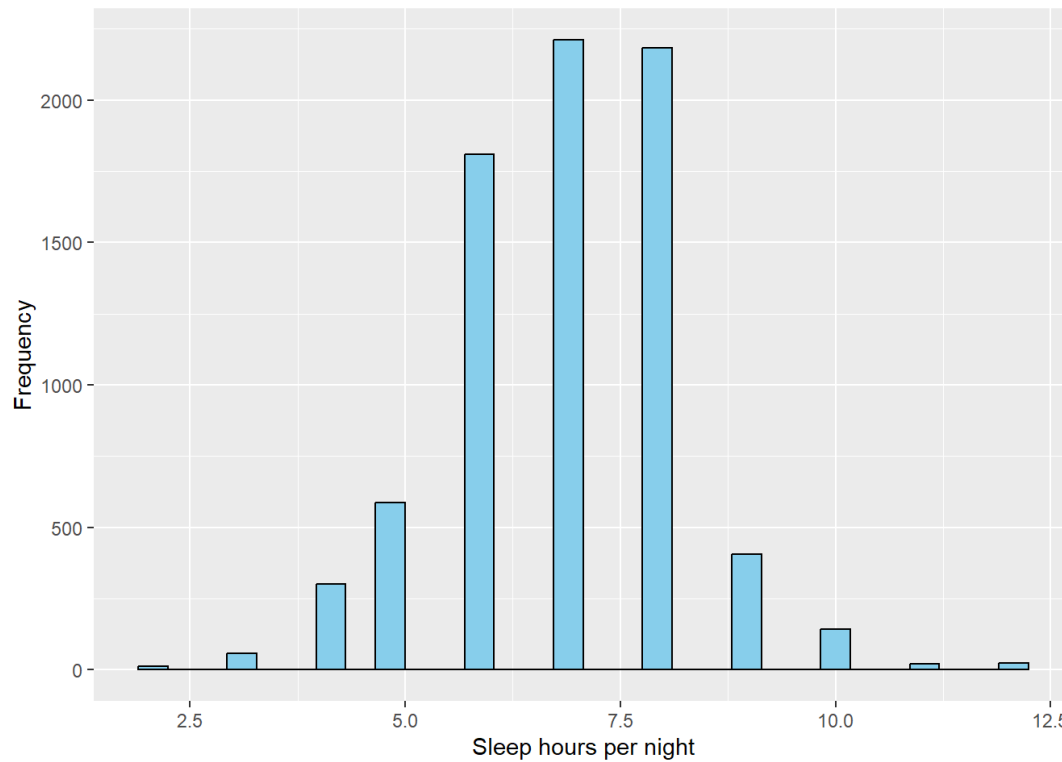
Test of hypothesis about the  
population mean (one sided test)



# The NHANES data set: analysis of the number of sleep hours per night

- In this section, the variable of interest is the number of sleeping hours per night (the variable SleepHrsNight).
- Information about the number of sleeping hours per night is available for 7755 individuals (i.e., the population).

# The number of sleep hours per night

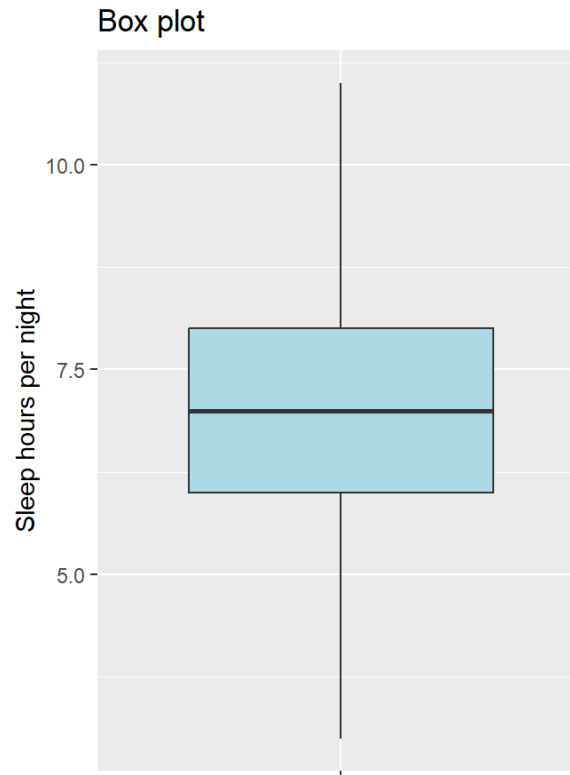
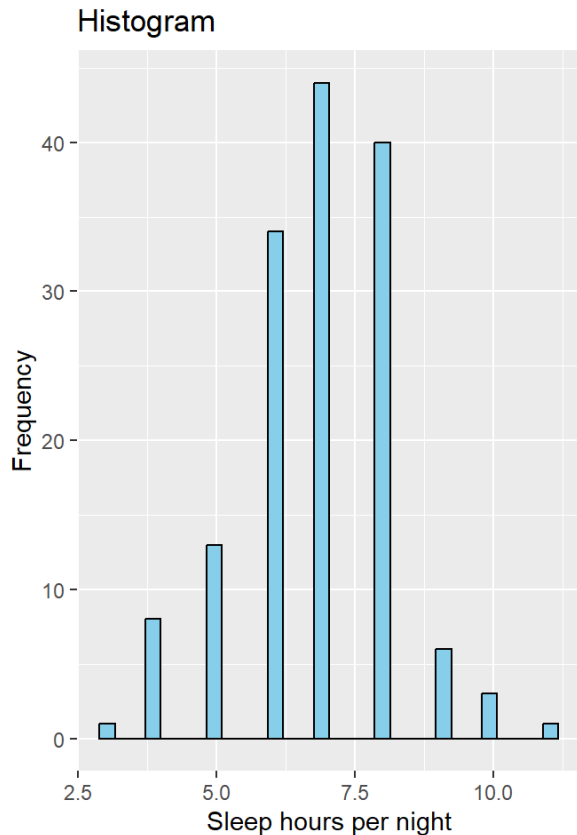


$$n = 7755$$

$$\mu = 6.927$$

$$\sigma = 1.813$$

# A random sample from the population



- A random sample from the population:

$$n = 150$$

$$\bar{x} = 6.846$$

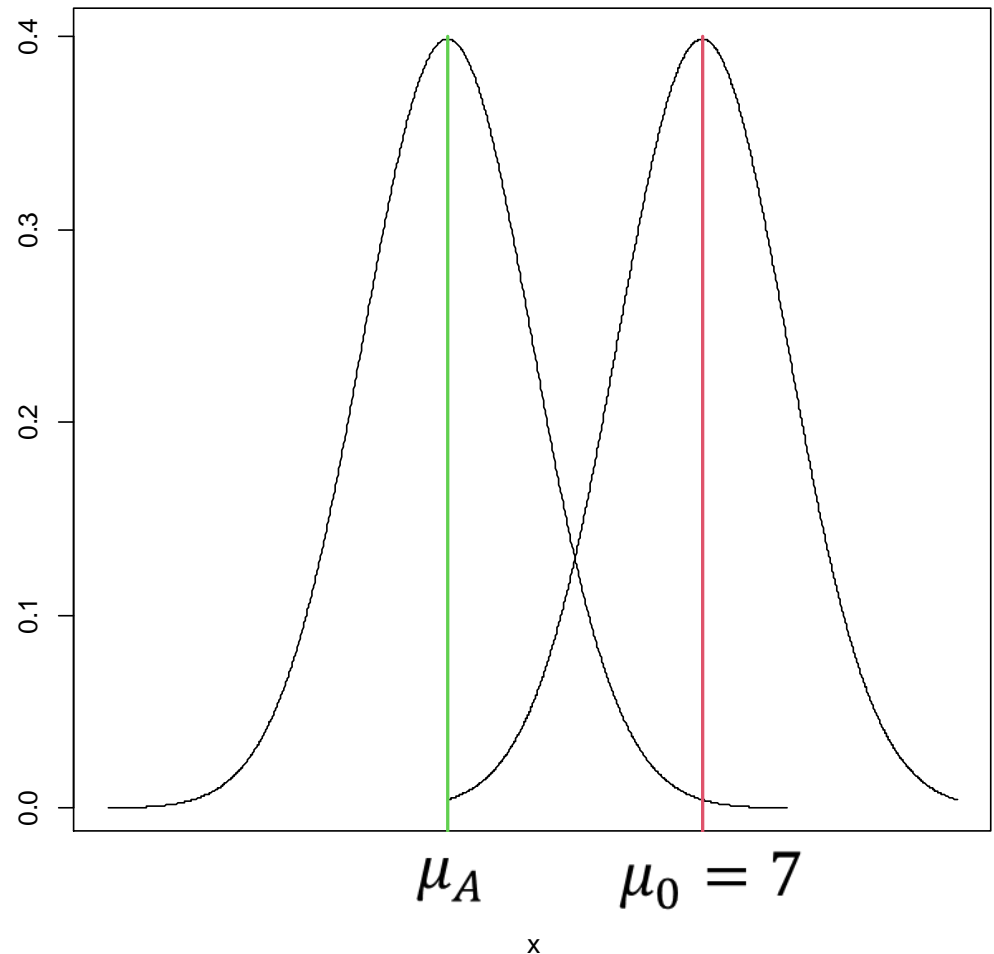
$$s^2 = 1.862$$

# Test of hypothesis: a one sided test

$$H_0: \mu = 7$$

$$H_A: \mu < 7$$

- We test the null hypothesis versus a one sided alternative.
- In our case, under the alternative the mean is smaller than 7 (but not specified).

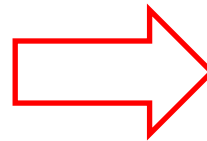


**The null hypothesis**

# Test statistic

$$t = \frac{\bar{x} - \mu_0}{\sqrt{\frac{1.3646^2}{150}}} = -1.3761$$

The population variance  $\sigma^2$  is unknown but... $n=150$ .



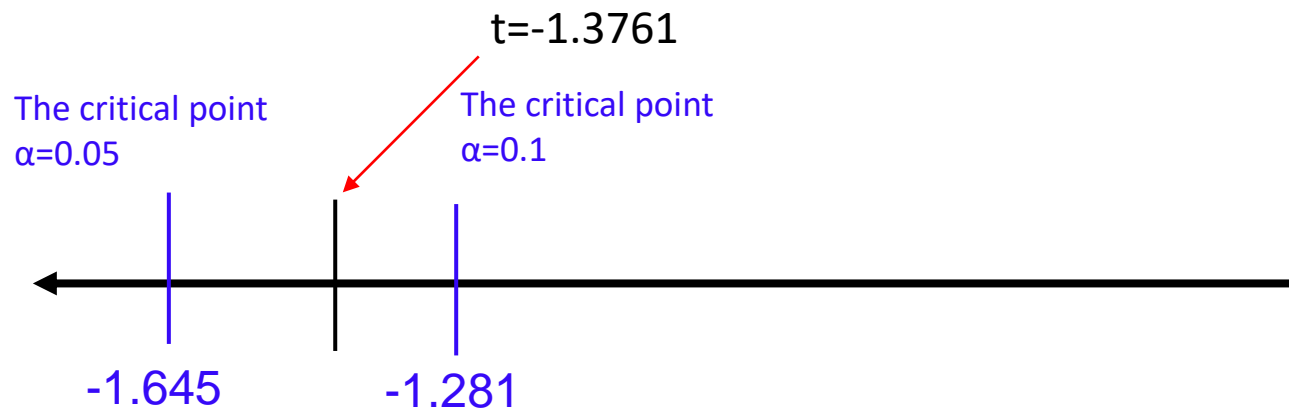
$$\frac{\bar{X} - \mu}{\sqrt{\frac{S^2}{n}}} \sim N(0,1)$$

# The critical points and the test statistic

For one sided test and  $\alpha=0.05$ ,  $Z=-1.645$ .

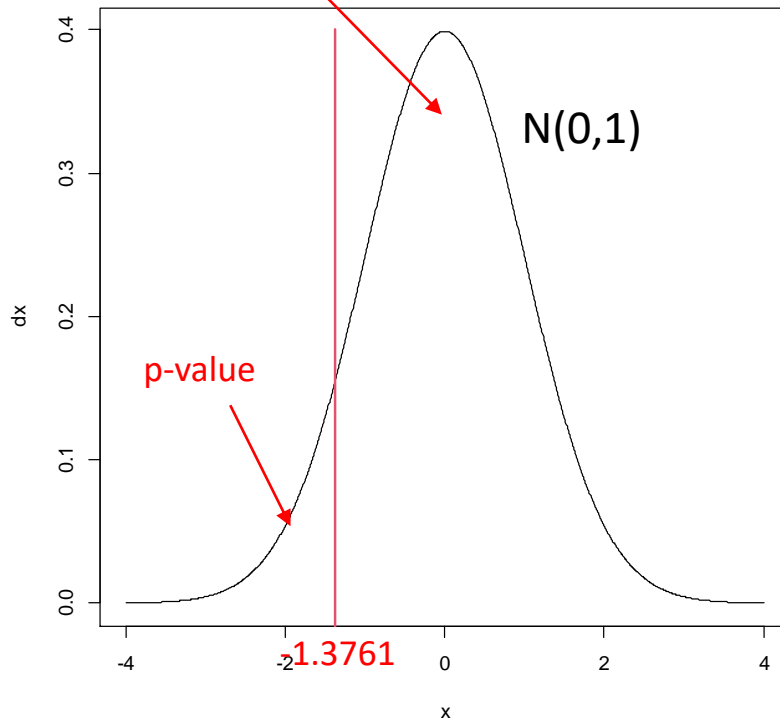
For one sided test and  $\alpha=0.1$ ,  $Z=-1.281$ .

For  $\alpha=0.1$  We reject  $H_0$  :  $-1.3761 < -1.281$ .



# p-value

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



$$H_0: \mu = 7$$

$$H_A: \mu < 7$$

$$P(Z < -1.3761) = 0.08439$$

- For  $\alpha=0.05$ , we DO NOT reject the null hypothesis.
- For  $\alpha=0.1$ , we reject the null hypothesis.

Case study 4:  
The The NHANES dataset: cholesterol level

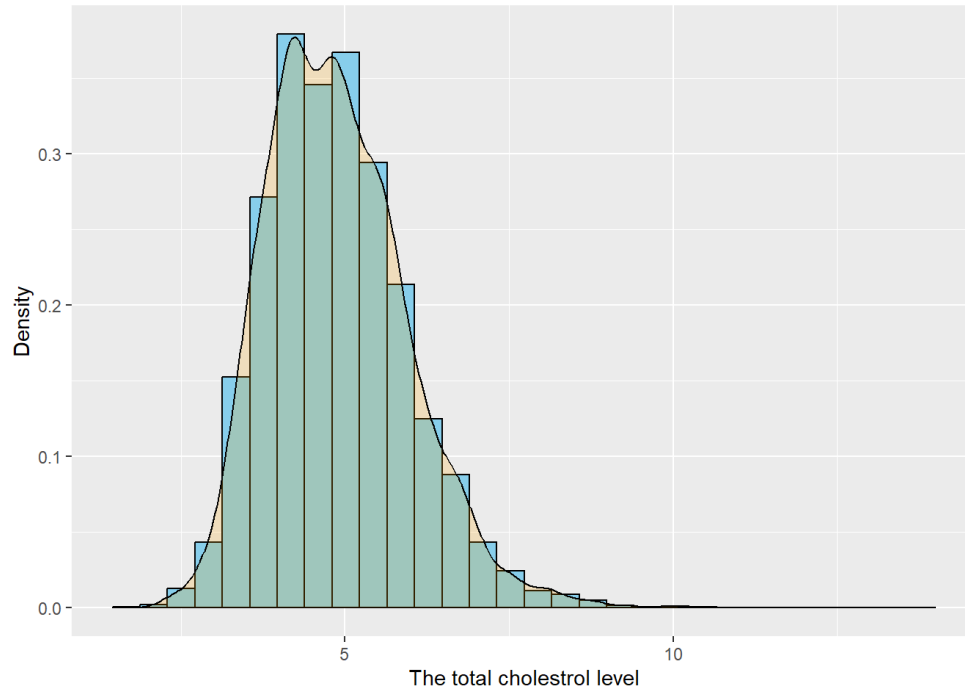
Test of hypothesis about the  
population mean (two sided test and  
confidence interval)



# The NHANES data set: analysis of the total cholesterol level

- In this section we focus on the total cholesterol level (the variable `TotChol`).
- After omitting all individuals with missing values, 8474 individuals are included in the analysis.

# Total cholesterol level



- The population and parameters:

$$n = 8474$$

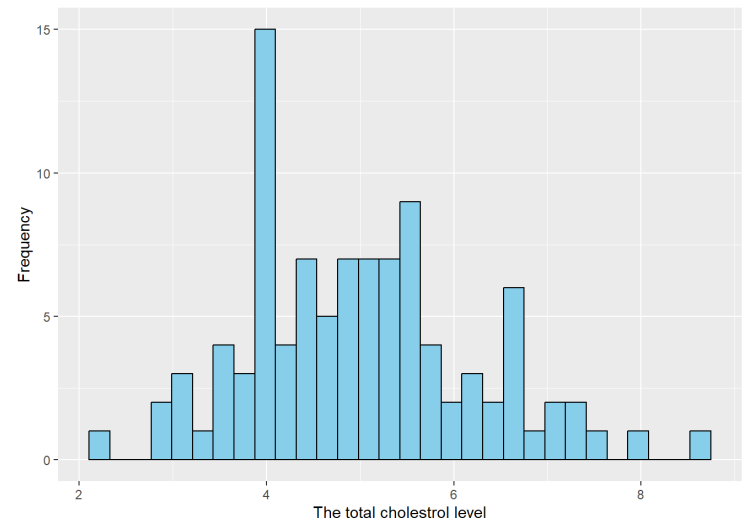
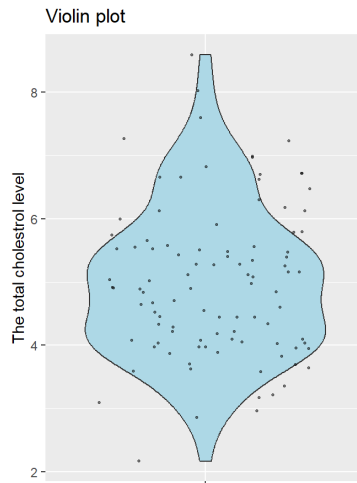
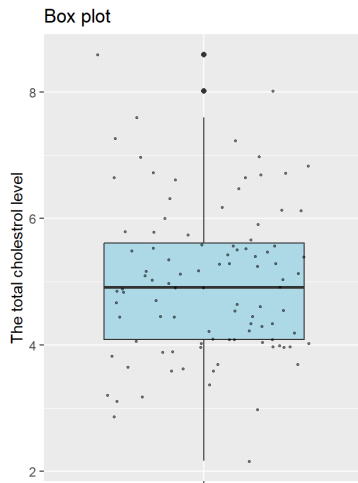
$$\mu = 4.879$$

$$\sigma^2 = 1.0755^2$$

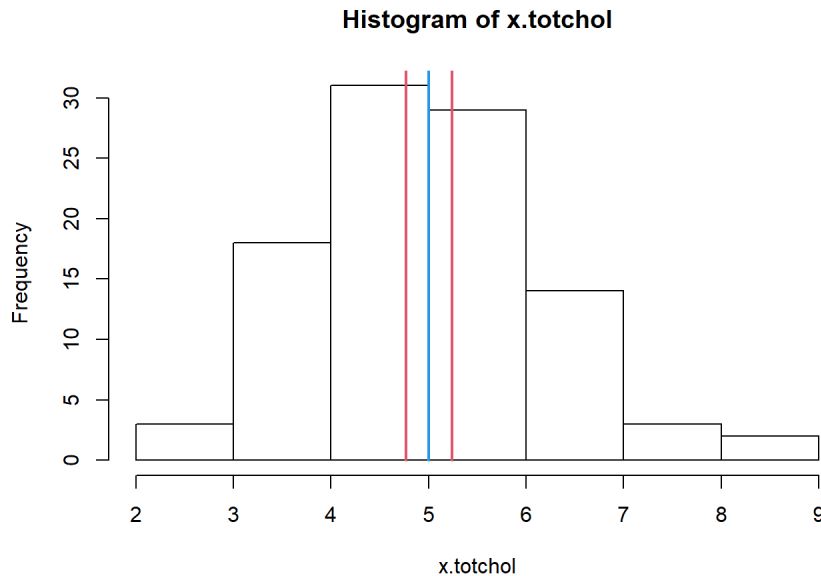
# A random sample of 100 individuals

For a sample of 100 individuals:

$$\left. \begin{array}{l} \bar{x} = 5.004 \\ s = 1.207 \end{array} \right\} \frac{s}{\sqrt{n}} = \frac{1.2072}{10} = 0.12073$$



# A 95% confidence interval for the population mean



$$\alpha = 0.025$$

$$Z_{\alpha} = 1.96$$

$$m = 1.96 \times 0.12073$$

$$\bar{x} \pm m = 5.0041 \pm 1.96 \times 0.12073$$

$$(4.7674, 5.2407)$$

# Test of hypothesis for the population mean (two-sided test)

$$H_0: \mu = 5.1$$

$$H_A: \mu \neq 5.1$$

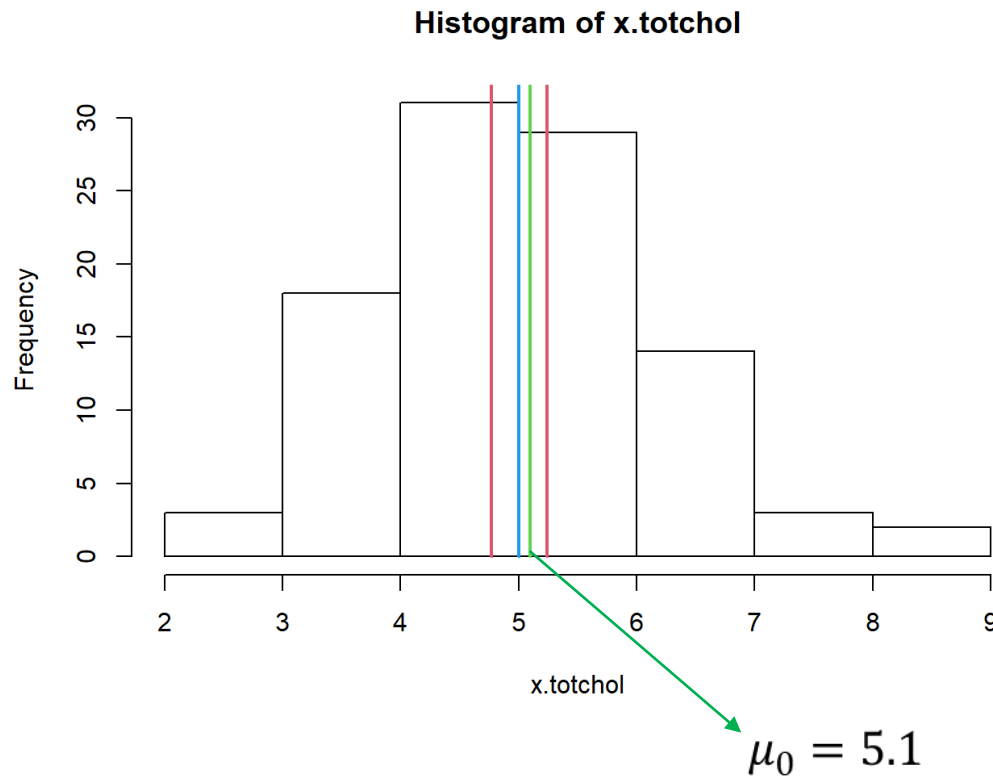
## Test statistics and p-value

$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = \frac{5.0041 - 5.1}{\frac{1.2073}{10}} = -0.7943$$

$$2 \times (1 - P(|Z| > 0.7943)) = 0.427$$

$$p - value = 0.427 > 0.05 = \alpha$$

# A two sided test and a confidence interval



- A 95% confidence interval for the population mean:  
 $(4.7674, 5.2407)$
- The value of  $\mu_0 = 5.1$  is covered by the 95% confidence interval.

