The >eR-Biostat initative

Making R based education materials in statistics accessible for all

Basic concepts in statistical inference using R: Single comparison (Part 1)

Developed by Ziv Shkedy (Hasselt University, Belgium, July 2017)

based on an online course developed by

Marc Lavielle

Inria Saclay (Xpop) & Ecole Polytechnique (CMAP)
March, 2017

Visit us on Facebook ER-BioStat

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

Email: erbiostat@gmail.com

twitter @erbiostat

The course was developed as a part of the >eR-BioStat initiative.

External datasets are available in the GitHub page of the course.

R code is available online in

http://sia.webpopix.org/statisticalTests1.html

E-learning system using R

Biostatistics

Contents

- Introduction
- Student's t-test
  - One sample t-test
  - One sided test
  - Two sided test
  - Confidence interval for the mean
  - Two samples t-test
  - What should we test?
  - Assuming equal variances
  - Assuming different variances
  - Power of a t-test
- Mann-Whitney-Wilcoxon test

Both slides and online materials are available.

- The limited role of the p-value
- Equivalence tests
  - Introduction
  - Two samples test
  - The TOST procedure
  - Difference testing versus equivalence testing
  - One sample test

These topics are not covered in the slides but online materials and R code are available.

http://sia.webpopix.org/statisticalTests1.html#the-limited-role-of-the-p-value
http://sia.webpopix.org/statisticalTests1.html#equivalence-tests

The course materials

Online materials can be found in

http://sia.webpopix.org/statisticalTests1.html

E-learning system using R

Biostatistics

YouTube tutorials

- YouTube tutorials related to the topics covered in the slides are available for:
  - One sample t-test (host by Mike Marin): https://www.youtube.com/watch?v=kvmSAXhX9Hs
  - Two samples t-test(host by Mike Marin): https://www.youtube.com/watch?v=RlhnNbPZC0A
  - Wilcoxon test for two independent samples (host by Clarie Reed): https://www.youtube.com/watch?v=jkpRGUkzFn4
  - Wilcoxon test for two independent samples using R (host by Mike Marin): https://www.youtube.com/watch?v=KroKhtCD9eE

E-learning system using R

Biostatistics

Contents

t-test for a population

- We assume that X~N(μ,σ²)
- For this test, we used the Student t distribution.

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

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

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

X has a normal distribution with unknown μ and σ².

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

One sided t-test

- One sided alternative:

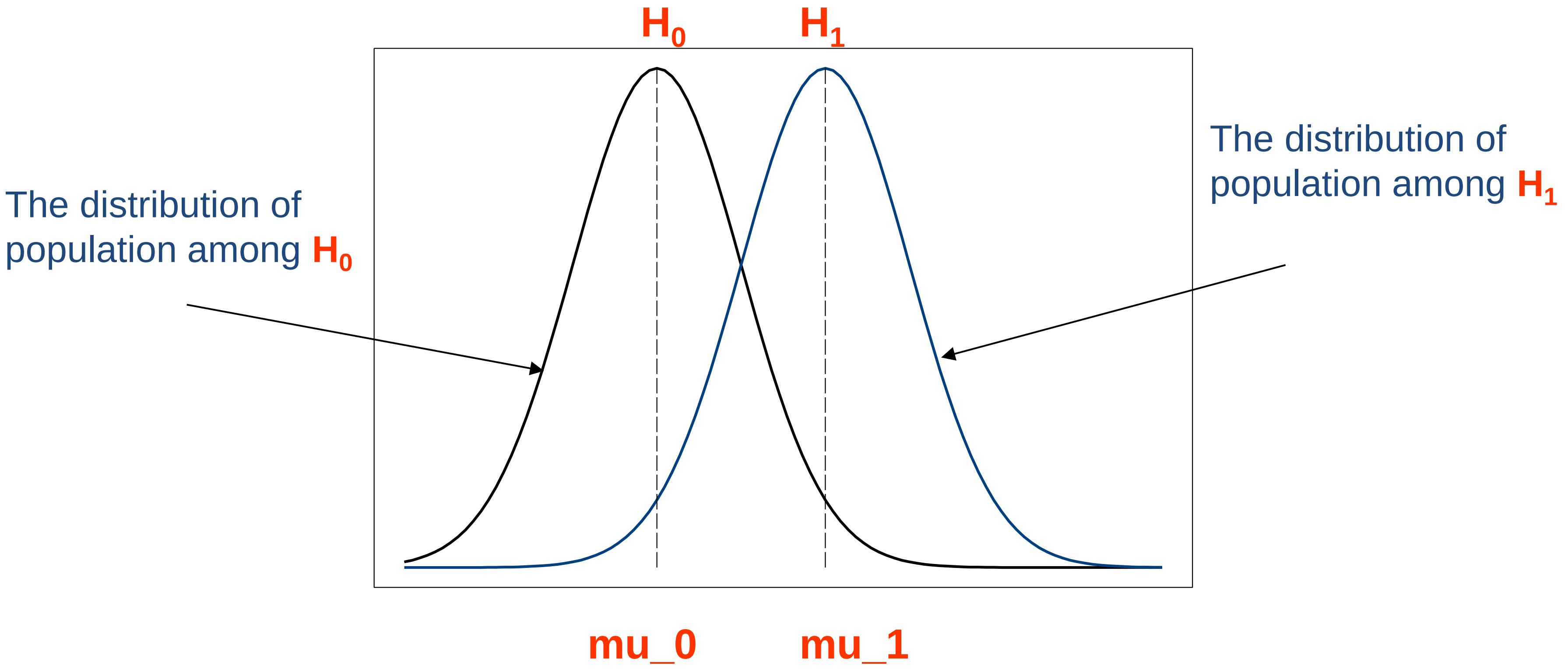
$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

- We assume that

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

The distribution of the population

$$X \sim N(\mu_0, \sigma^2) \qquad under \quad H_0$$
$$X \sim N(\mu_1, \sigma^2) \qquad under \quad H_1$$



The sample & test statistic

- To test the hypotheses, we draw a sample from the population.
- X follows a normal distribution with unknown μ and σ^2.

$$\left.\begin{aligned} &X_i \sim N(\mu, \sigma^2) \\ &n: \quad ? \\ &\sigma^2 : unknown \end{aligned}\right\} \quad \boxed{\frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)}$$

The distribution of the test statistic population under **H_0**

One sided t-test using R

The hypotheses:

$$H_0 : \mu = 500$$
$$H_1 : \mu > 500$$

The data (weight of male):

```
> x <- data[data$gender=="Male","weight"]
```

The mean under the null hypothesis

```
> mu0 <- 500
```

R code:

```
> t.test(x, alternative="greater", mu=mu0)
```

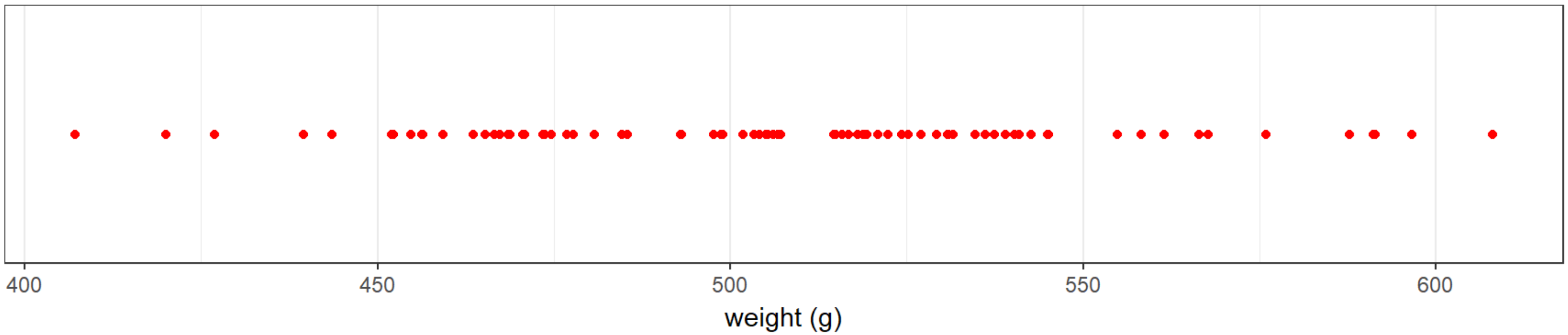
$H_1 : \mu > \mu_0$

data

μ_0

The data

Weight of 78 rat male.




```
> x <- data[data$gender=="Male","weight"]
> length(x)
[1] 78
> min(x)
[1] 407.2
> max(x)
[1] 608
> mean(x)
[1] 506.2218
```

R output for a one sided t-test

```
> t.test(x, alternative="greater", mu=mu0)

        One Sample t-test

data:  x
t = 1.2708, df = 77, p-value = 0.1038
alternative hypothesis: true mean is greater than
500
95 percent confidence interval:
 498.0706      Inf
sample estimates:
mean of x
 506.2218
```

$\hat{\mu}$

The rejection region

```
> t.test(x, alternative="greater", mu=mu0)

t = 1.2708, df = 77, p-value = 0.1038
```

$\underbrace{t = 1.2708}_{\text{Test statistic}}$

Test statistic

$$\frac{\overline{X}-\mu_0}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)$$

```
> alpha <- 0.05
> n <- length(x)
> df <- n-1
> df
[1] 77
> qt(1-alpha, df)
[1] 1.664885
```

$t_{(77)}$

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

C=1.664885

$[c, \infty[$

The choice of c

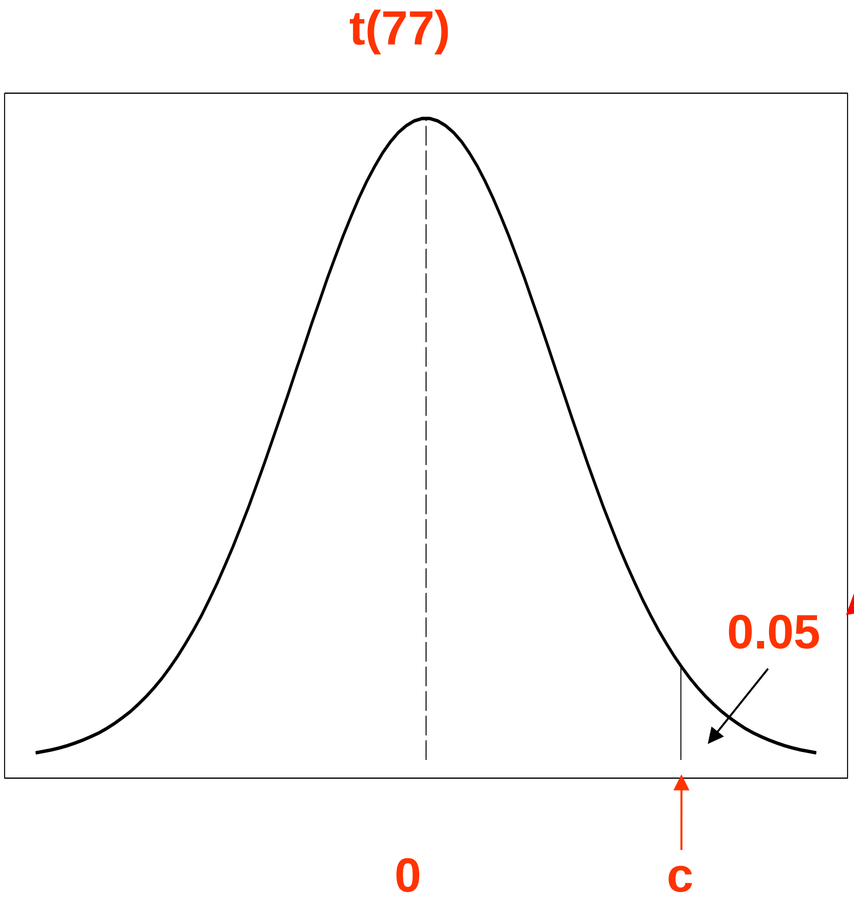
Determine c so that Type I error =5%

$$P(\overline{X} > k) = P(T > c) = 0.05$$

$$P(\overline{X} > c) = P\left(\frac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}} > \frac{k - \mu_0}{\sqrt{\frac{S^2}{n}}} \right) = 0.05 \quad \Rightarrow \quad P\left(T > \frac{k - \mu_0}{\sqrt{\frac{S^2}{n}}} \right) = 0.05$$

c

The critical point in R



$$P(T > c) = P(T > 1.664885) = \alpha$$

The distribution of the test statistic under H_0

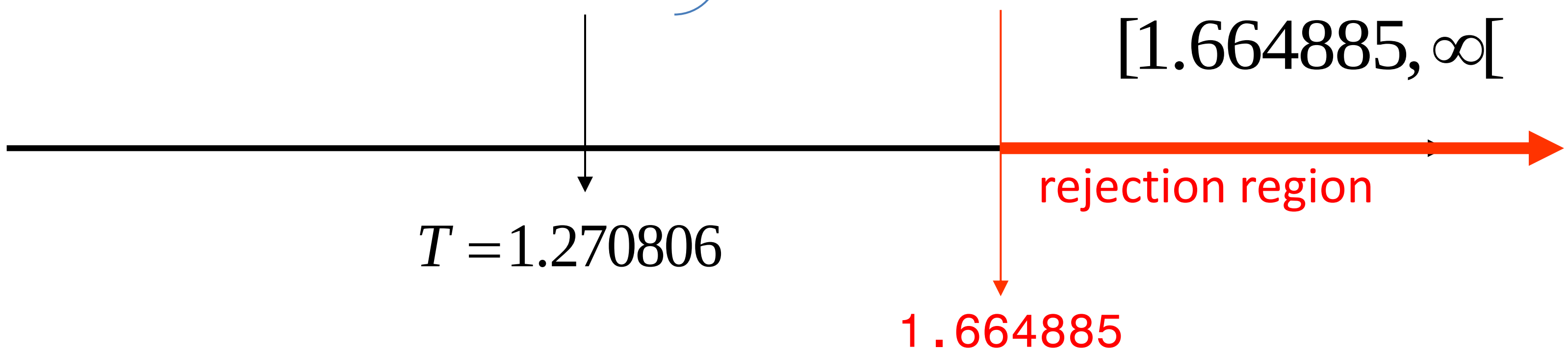
$$\boxed{\frac{\bar{X} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)}$$

```
> alpha <- 0.05
> n <- length(x)
> df <- n-1
> df
[1] 77
> qt(1-alpha, df)
[1] 1.664885
```

The rejection region & statistic in R

```
> x.mean <- mean(x)
> x.mean
[1] 506.2218
> x.sd <- sd(x)
> x.sd
[1] 43.23981
> n <- length(x)
> n
[1] 78
> t.stat <- sqrt(n)*(x.mean-mu0)/x.sd
> t.stat
[1] 1.270806
```

$$\frac{506.2218-500}{\sqrt{\frac{43.23981^2}{78}}}=1.270806$$



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

The p value in R

```
> t.test(x, alternative="greater", mu=mu0)

        One Sample t-test

data:  x
t = 1.2708, df = 77, p-value = 0.1038
alternative hypothesis: true mean is greater than 500
95 percent confidence interval:
 498.0706      Inf
sample estimates:
mean of x
 506.2218
```

$t_{(77)}$

$$P_{H_0}(T_{stat} > T_{stat}^{observed})$$

```
> p.value <- 1 - pt(t.stat,df)
> p.value
[1] 0.1038119
```

Example 2: one sided alternative

The hypotheses:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu < \mu_0$$

t-test in R:

```
> mu0 <- 515
> t.test(x, alternative="less", mu=mu0)
```

R output for a one-sided t-test

```
> t.test(x, alternative="less", mu=mu0)

        One Sample t-test

data:  x
t = -1.793, df = 77, p-value = 0.03845
alternative hypothesis: true mean is less than
515
95 percent confidence interval:
    -Inf 514.373
sample estimates:
mean of x
 506.2218
```

$\mu_0 = 515$

$H_0 : \mu = 115$

$H_1 : \mu < 115$

The test statistic in R

```
> x.mean <- mean(x)
> x.sd <- sd(x)
> n <- length(x)
> t.stat <- sqrt(n)*(x.mean-mu0)/x.sd
> t.stat
[1] -1.792954
```

$$\frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{n}}} = -1.792954$$

```
> alpha <- 0.05
> n <- length(x)
> df <- n-1
> df
[1] 77
> qt(alpha, df)
[1] -1.664885
```

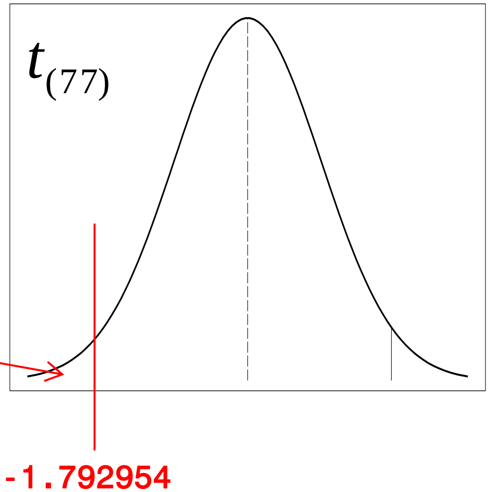
The critical point in R $\Leftrightarrow$ $H_0 : \mu = 115$, $H_1 : \mu < 115$ $\Leftrightarrow$ $P\left(T \le \frac{k-\mu_0}{\sqrt{\frac{S^2}{n}}} \right) = 0.05$

rejection region $]-\infty, -1.664885]$

The p value in R

```
> t.stat
[1] -1.792954
> df
[1] 77
> p.value <- pt(t.stat,df)
> p.value
[1] 0.03845364
```

$$P_{H_0}(T_{stat} < T_{stat}^{observed}) = P_{H_0}(T_{stat} < -1.792954)$$



Contents

YouTube tutorial:

`https://www.youtube.com/watch?v=kvmSAXhX9Hs`

A two-sided test in R

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

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

Twp sided test in R

```
> mu0 = 500
> t.test(x, alternative="two.sided", mu=mu0)
```

$$H_0: \mu = 500$$

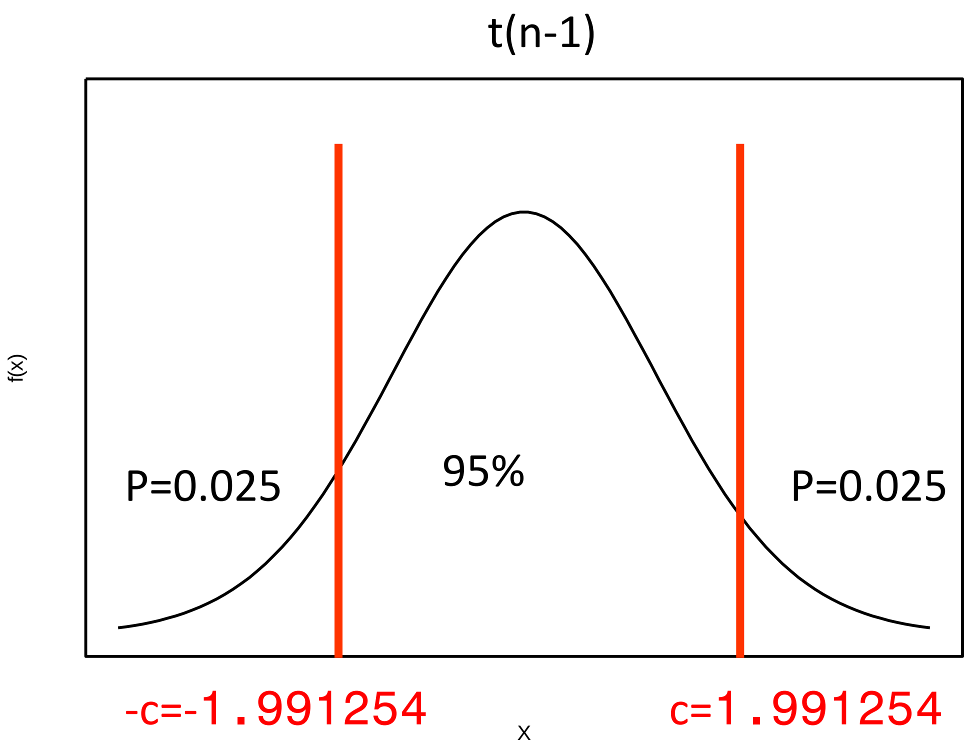
$$H_1: \mu \neq 500$$

R output for a two sided t-test

```
> t.test(x, alternative="two.sided", mu=mu0)

        One Sample t-test

data:  x
t = 1.2708, df = 77, p-value = 0.2076
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 496.4727 515.9709
sample estimates:
mean of x
 506.2218
```

$\hat{\mu}$

Critical values in R



n=78 and α=0.05:

```
> alpha <- 0.05
> n <- length(x)
> df <- n-1
> df
[1] 77
> qt(1-alpha/2, df)
[1] 1.991254
```

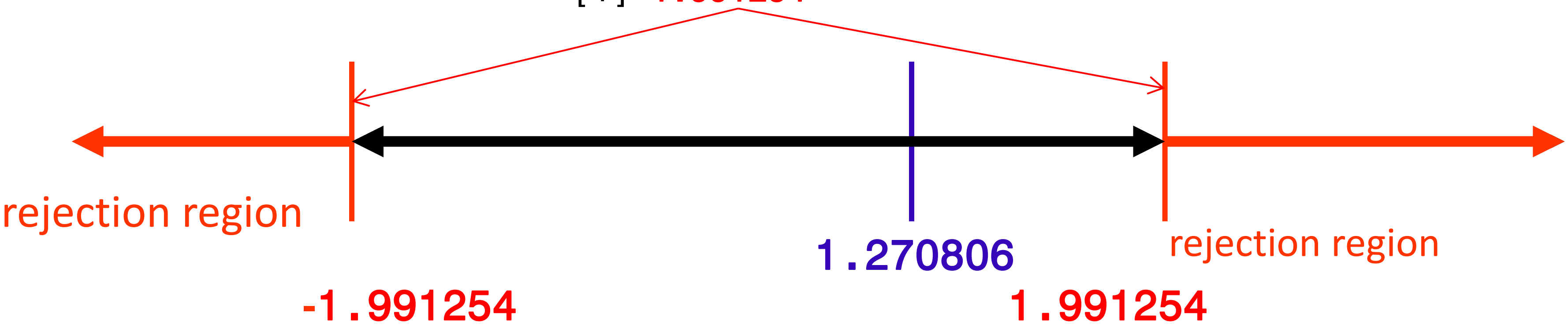
$$P\left(-1.991254 \leq \frac{\overline{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} \leq 1.991254\right) = 0.95$$

The rejection region

```
> x.mean <- mean(x)
> x.sd <- sd(x)
> n <- length(x)
> t.stat <- sqrt(n)*(x.mean-mu0)/x.sd
> t.stat
[1] 1.270806
```

$$\frac{\bar{x}-\mu_0}{\sqrt{\frac{s^2}{n}}} = 1.270806$$

```
> alpha <- 0.05
> n <- length(x)
> df <- n-1
> qt(1-alpha/2, df)
[1] 1.991254
```



The p value in R

```
> t.test(x, alternative="two.sided", mu=mu0)

        One Sample t-test

data:  x
t = 1.2708, df = 77, p-value = 0.2076
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 496.4727 515.9709
sample estimates:
mean of x
 506.2218
```

$t_{(77)}$

$$2 \times P_{H_0}(T_{stat} \le -T_{stat}^{observed}) = 2 \times P_{H_0}(T_{stat} \le -1.2708)$$

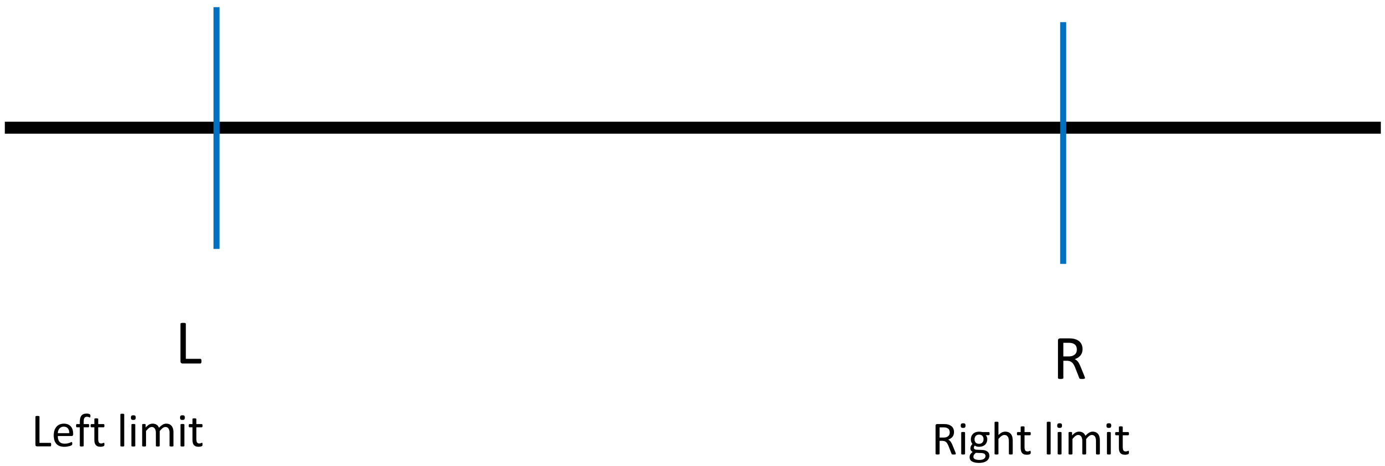
```
> p.value <- 2*pt(-t.stat,df)
> p.value
[1] 0.2076238
```

Contents

Confidence interval

On the basis of the sample, we look for an interval [L, R] so :

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

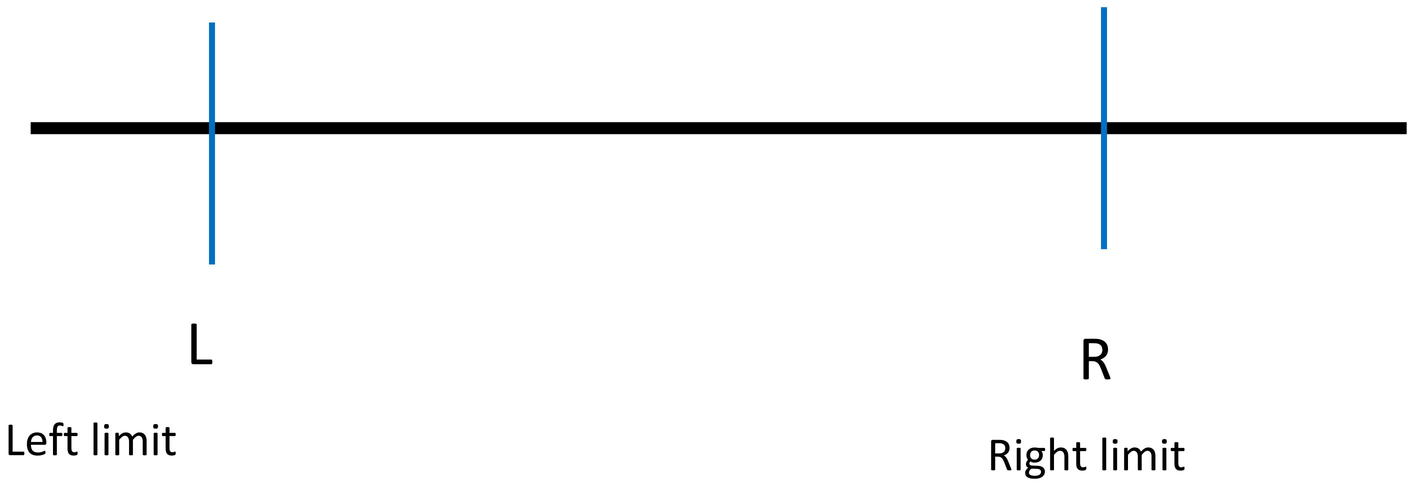


We find an interval [L, R] that contains the value of the population mean (μ) with "high probability"

Confidence interval

Large → 1 - α

$$\boxed{P\left(L \le \mu \le R\right) = 1 - \alpha}$$

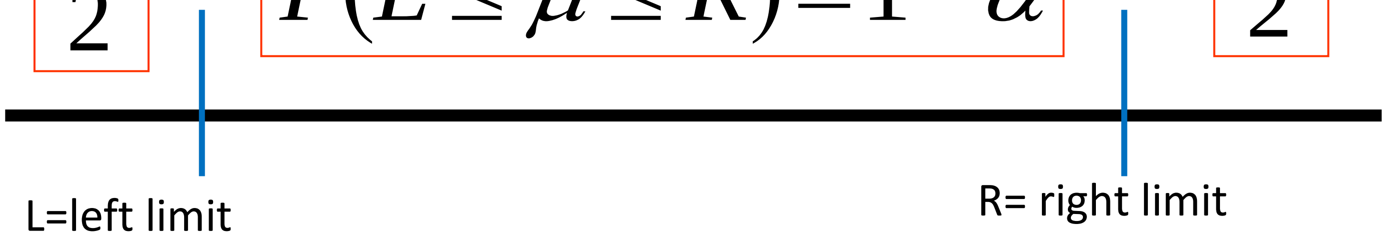


Example: $\alpha = 0.05 \Rightarrow 1 - \alpha = 0.95$

Confidence interval

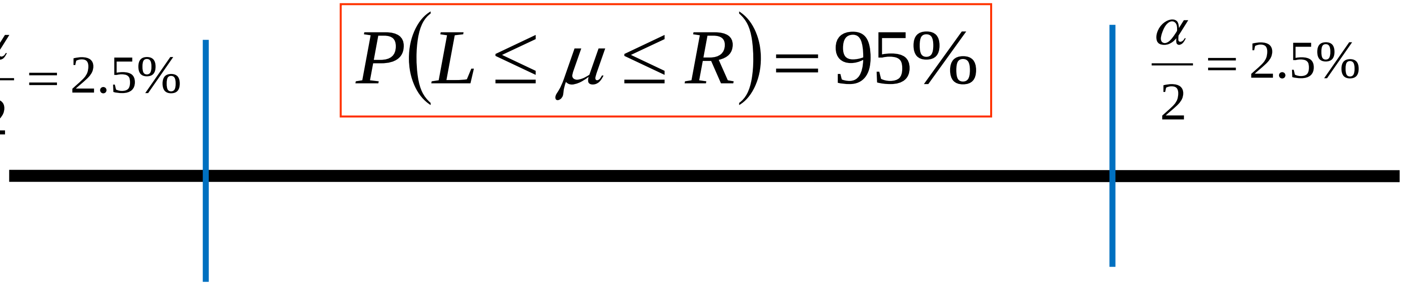
We are looking for an interval so

$$\boxed{\frac{\alpha}{2}} \qquad \boxed{P(L \le \mu \le R) = 1 - \alpha} \qquad \boxed{\frac{\alpha}{2}}$$



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

$$\frac{\alpha}{2} = 2.5\% \qquad \boxed{P(L \le \mu \le R) = 95\%} \qquad \frac{\alpha}{2} = 2.5\%$$



Confidence interval

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

$$\boxed{P\left(L\leq\mu\leq U\right)=1-\alpha}$$

A (1-α) CI for μ is :

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

A 95% confidence interval in R

```
> t.test(x, alternative="two.sided", mu=mu0)

        One Sample t-test

data:  x
t = 1.2708, df = 77, p-value = 0.2076
alternative hypothesis: true mean is not equal to 500
95 percent confidence interval:
 496.4727 515.9709
sample estimates:
mean of x
 506.2218
```

$$\left[\bar{X} - t \times \sqrt{\frac{S^2}{n}}, \bar{X} + t \times \sqrt{\frac{S^2}{n}}\right] = [496.4727, 515.9709]$$

```
> c.val<-qt(1-alpha/2, df)
> c.val
[1] 1.991254
```

```
> alpha <- 0.05
> n <- length(x)
> df <- n-1
> c.val<-qt(1-alpha/2, df)
> c.val
[1] 1.991254
> x.mean <- mean(x)
> x.mean
[1] 506.2218
> x.mean+c.val*(x.sd/sqrt(n))
[1] 515.9709
```

A 90% confidence interval in R

```
> t.test(x, alternative="two.sided", mu=mu0,conf.level = 0.90)

        One Sample t-test

data:  x
t = 1.2708, df = 77, p-value = 0.2076
alternative hypothesis: true mean is not equal to 500
90 percent confidence interval:
 498.0706 514.3730
sample estimates:
mean of x
 506.2218
```

Contents

YouTube tutorial:

`https://www.youtube.com/watch?v=RlhnNbPZCOA`

Two populations and two independent samples

Population 1

μ_1, σ_1^2

Population 2

μ_2, σ_2^2

We draw two samples independently

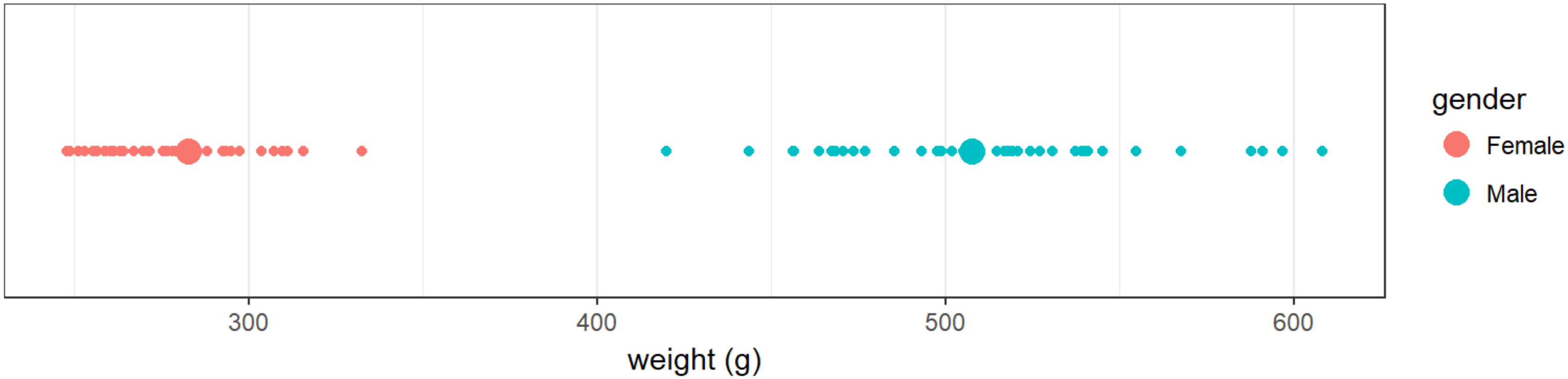
sample 1

Sample 2

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

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

Example 1: weight of female vs. male - the data




```
> y <- data[data$gender=="Female" ,"weight"]
> length(y)
[1] 80
> mean(y)
[1] 282.8025
> x <- data[data$gender=="Male","weight"]
> length(x)
[1] 78
> mean(x)
[1] 506.2218
```

female

male

Two independent samples

We are interested in the difference between the two means μ_1 and μ_2 and set the null hypothesis:

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

If the means of the two populations are equal then $(\mu_2-\mu_1)_{H0}=0$

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

Two sided alternative

The hypotheses:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

t-test in R:

```
> t.test(x, y)
```

The two samples:

```
y <- data[data$gender=="Female" ,"weight"]
x <- data[data$gender=="Male","weight"]
```

Two-sided t-test assuming unequal variance in the two populations (see later).

R output for a two samples t-test

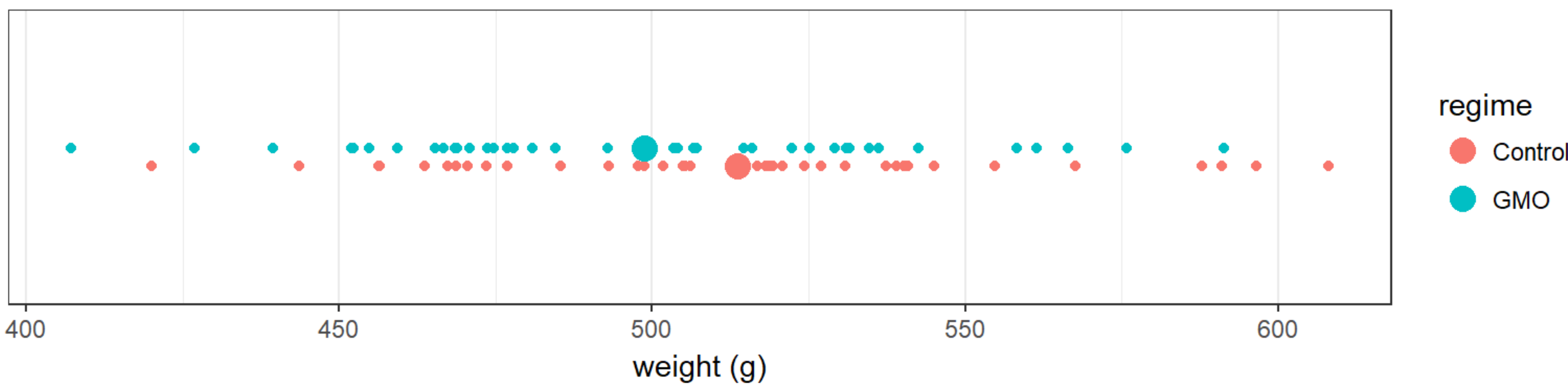
```
> t.test(x, y)

        Welch Two Sample t-test

data:  x and y
t = 40.35, df = 117.08, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 212.4535 234.3851
sample estimates:
mean of x mean of y
 506.2218  282.8025
```

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Two-sided test

Example 2: weight control vs. GMO for male




```
> x <- data[data$gender=="Male" & data$regime=="Control","weight"]
> length(x)
[1] 39
> mean(x)
[1] 513.7077
> y <- data[data$gender=="Male" & data$regime=="GMO","weight"]
> length(y)
[1] 39
> mean(y)
[1] 498.7359
```

Treatment groups

Two sample t-test in R: equal variance

The hypotheses:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

t-test in R:

```
> alpha <- 0.05
> t.test(x, y, conf.level=1-alpha, var.equal=TRUE)
```

$\underbrace{\texttt{conf.level=1-alpha}}_{\text{95\% C.I}}$

$\underbrace{\texttt{var.equal=TRUE}}_{\sigma_1^2 \text{ and } \sigma_2^2 \text{ unnown but } \sigma_1^2 = \sigma_2^2}$

Two-sided t-test assuming equal variance in the two populations

Assuming equal variances

1. both populations are normally distributed.
2. σ_1^2 and σ_2^2 unnown but $\sigma_1^2 = \sigma_2^2$

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

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

pooled sample variance

R output for a two samples t-test (equal variance)

```
> t.test(x, y, conf.level=1-alpha, var.equal=TRUE)

        Two Sample t-test

data:  x and y
t = 1.5426, df = 76, p-value = 0.1271
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.358031 34.301621
sample estimates:
mean of x mean of y
 513.7077  498.7359
```

$\hat{\mu}_1$ $\hat{\mu}_2$

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

$$n_1+n_2-2=39+39-2$$

```
> length(x)
[1] 39
> length(y)
[1] 39
```

Confidence interval

- A (1-alpha)*100% confidence interval for the difference μ_2 - μ_1 is given by

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

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

```
> n1<-length(x)
> n2<-length(y)
> df<-n1+n2-2
> df
[1] 76
> alpha<-0.05
> qt(1-alpha/2,df)
[1] 1.991673
```

$$n_1=n_2=39$$

$$df=78-2$$

$$\alpha=0.05$$

Assuming unequal variances

1. the populations are normally distributed.
2. σ_1^2 and σ_2^2 are unknown but $\sigma_1{}^2 \neq \sigma_2{}^2$

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

$$S_1{}^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_i - \bar{X})^2$$

$$S_2{}^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (Y_i - \bar{Y})^2$$

Two sample t-test in R: unequal variance

The hypotheses:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

t-test in R:

```
> t.test(x, y, conf.level=1-alpha)
> t.test(x, y, conf.level=1-alpha, var.equal=FALSE)
```

95% C.I

σ_1^2 and σ_2^2

unonwn but $\sigma_1^2 \neq \sigma_2^2$

R output for a two samples t-test (unequal variance)

```
> t.test(x, y, conf.level=1-alpha, var.equal=FALSE)

        Welch Two Sample t-test

data:  x and y
t = 1.5426, df = 75.976, p-value = 0.1271
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.358129 34.301719
sample estimates:
mean of x mean of y
 513.7077  498.7359
```

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

$\hat{\mu}_1$ $\hat{\mu}_2$

Contents

Two sample t –test: one sided alternative

$$H_0 : \mu_1 - \mu_2 = \Delta$$
$$H_1 : \mu_1 - \mu_2 > \Delta$$

Δ: the true difference between the population means under the null hypothesis.

The two errors

population

Test results

	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

Type I error and Power

Hypotheses:

$$H_0 : \mu_1 - \mu_2 = \Delta$$
$$H_1 : \mu_1 - \mu_2 > \Delta$$

Test statistic:

$$t = \frac{\overline{Y} - \overline{X} - (\mu_2 - \mu_1)_{H_0}}{\sqrt{S_P^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

For a given value of c, Type I error:

$$P_{H_0}(t > c) = \alpha$$

We reject H_0 when it is correct

For a given value of c, power

$$P_{H_1}(t \le c) = 1 - \beta = power$$

Power calculation in R

Hypotheses:

$$H_0 : \mu_1 - \mu_2 = \Delta$$
$$H_1 : \mu_1 - \mu_2 > \Delta$$

Test statistic:

$$t = \frac{(\mu_2 - \mu_1)_{H_0}}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

To calculate the power we need to know:

H_0:

H_1:

One sided/two sided ?

$\Delta = (\mu_2 - \mu_1)_{H_0}$ True difference in the populations

σ^2 variance

n Sample size

Power calculation in R

To calculate the power we need to know:

$$\left.\begin{array}{l} H_0: \\ \\ H_1: \end{array}\right\} \text{One sided/two sided ?}$$

$$\Delta = (\mu_2 - \mu_1)_{H_0} = 10$$

$$\sigma^2 = 30^2$$

$$n = 80$$

$$\alpha = 0.05$$

In R:

```
> alpha=0.05
> nx.new <- ny.new <- 80
> delta.mu <- 10
> x.sd <- 30
> df <- nx.new+ny.new-2
> dt <-
delta.mu/x.sd/sqrt(1/nx.new
+1/ny.new)
> dt
[1] 2.108185
```

$$dt = \frac{(\mu_2 - \mu_1)_{H_0}}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{10}{\sqrt{30^2\left(\frac{1}{80} + \frac{1}{80}\right)}} = 2.108..$$

Power calculation in R

- Use the R package `pwr`.
- Basic Functions for Power Analysis in R (CRAN).
  - Power calculation for a given sample size.
  - Sample size calculation for a given power.
- YouTube tutorial (Calculatingg Power in R):

  https://www.youtube.com/watch?v=7xghHcmQC50

Power calculation in R

In R:

```
> alpha=0.05
> nx.new <- ny.new <- 80
> delta.mu <- 10
> x.sd <- 30
> df <- nx.new+ny.new-2
> dt <-
delta.mu/x.sd/sqrt(1/nx.new+1/n
y.new)
> dt
[1] 2.108185
```

The `pwr.t.test()` function in R:

```
>pwr.t.test(n=nx.new,
            d=delta.mu/x.sd,
            type="two.sample",
            alternative="two.sided",
            sig.level=alpha)
```

$n=nx.new \longleftrightarrow n$

$d=delta.mu/x.sd \longleftrightarrow \frac{\Delta}{\sigma}$

type="two.sample", alternative="two.sided" $\longleftrightarrow$

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

sig.level=alpha $\longleftrightarrow \alpha$

Power calculation in R: output

```
> pwr.t.test(n=nx.new, d=delta.mu/x.sd, type="two.sample",
alternative="two.sided", sig.level=alpha)

     Two-sample t test power calculation

              n = 80
              d = 0.3333333
      sig.level = 0.05
          power = 0.5538758
    alternative = two.sided

NOTE: n is number in *each* group
```

Power calculation in R: output

```
> pwr.t.test(n=nx.new, d=delta.mu/x.sd, type="two.sample",
alternative="two.sided", sig.level=alpha)

     Two-sample t test power calculation

              n = 80
              d = 0.3333333
      sig.level = 0.05
          power = 0.5538758
    alternative = two.sided

NOTE: n is number in *each* group
```

Power calculation in R: output

```
> alpha=0.05
> nx.new <- ny.new <- 150
> delta.mu <- 10
> x.sd <- 30
> pwr.t.test(n=nx.new, d=delta.mu/x.sd,
type="two.sample", alternative="two.sided",
sig.level=alpha)
```

Power for n=150 per group.

```
     Two-sample t test power calculation

              n = 150
              d = 0.3333333
      sig.level = 0.05
          power = 0.820553
    alternative = two.sided

NOTE: n is number in *each* group
```

$$\Delta = (\mu_2 - \mu_1)_{H_0} = 10$$

$$\sigma^2 = 30^2$$

$$n = 150$$

$$\alpha = 0.05$$

Sample size calculation for a given power and Type I error

```
> alpha=0.05
> delta.mu <- 10
> x.sd <- 30
> pwr.t.test(power=0.8, d=delta.mu/x.sd, sig.level=alpha)
```

What is the sample size, per group for power of 80% and Type I error of 5% ?

```
     Two-sample t test power calculation

              n = 142.2462
              d = 0.3333333
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

$\Delta = (\mu_2 - \mu_1)_{H_0} = 10$

$\sigma^2 = 30^2$

$n = ?$

$\alpha = 0.05$

$1 - \beta = 0.8$

Contents

Wilcoxon test for two independent samples

- This part of the course does not has slides. Instead, we suggest YouTube tutorials.

- General explanation about Wilcoxon test for two independent samples:

  https://www.youtube.com/watch?v=jkpRGUkzFn4

- Wilcoxon test for two independent samples in R:

  https://www.youtube.com/watch?v=KroKhtCD9eE

Two-samples Wilcoxon text in R

A general call of the R function wilcox.test() has the form of

```
wilcox.test(sample 1, sample 2,
            alternative="two.sided",
            conf.level=1-alpha)
```

For different alternatives use:

```
alternative = c("two.sided", "less", "greater")
```

R output for a two-sided Wilcoxon test for independent samples

```
> wilcox.test(x, y, alternative="two.sided",
conf.level=1-alpha)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 904.5, p-value = 0.1516
alternative hypothesis: true location shift is not equal
to 0

Warning message:
In wilcox.test.default(x, y, alternative = "two.sided",
conf.level = 1 -  :cannot compute exact p-value with ties
```