## (Hypothesis testing)

- **Hypothesis test for the mean $\mu$**
  **Case 1:** X has a normal distribution with known variance $\sigma^2$.
  **Case 2:** X has a normal distribution with unknown variance $\sigma^2$.
  **Case 3:** X has a general distribution, but we have a large sample size.

  **Case 1**:

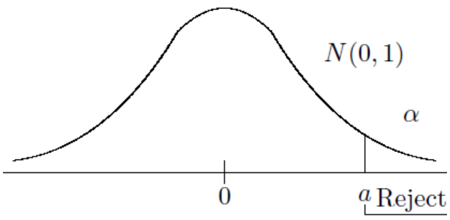
$$\begin{cases} H_0: & \mu = \mu_0 \\ H_a: & \mu > \mu_0 \end{cases}$$

Test statistics is

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

The rejection region is given by

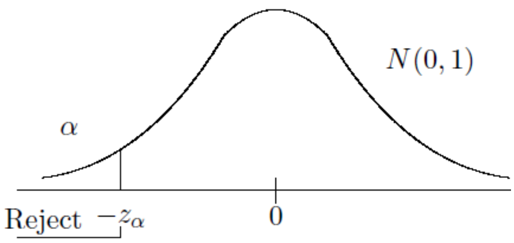
$$[z_\alpha, +\infty[$$




$$\begin{cases} H_0: & \mu = \mu_0 \\ H_a: & \mu < \mu_0 \end{cases}$$

We have that the rejection region is

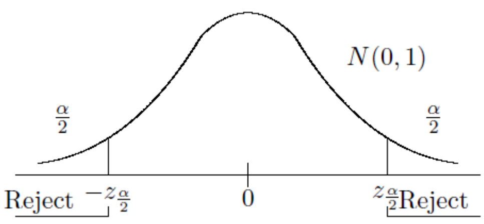
$$]-\infty, -z_\alpha].$$



$$\begin{cases} H_0: & \mu = \mu_0 \\ H_a: & \mu \neq \mu_0 \end{cases}$$

We reject $H_0$ if $\bar{X}$ is far from $\mu_0$ in either direction. The rejection region is

$$]-\infty, -z_{\frac{\alpha}{2}}] \cup [z_{\frac{\alpha}{2}}, +\infty[.$$




## Example:

From a long term experience a factory owner knows that a worker can produce a product in an average time of 89 min. However on Monday morning, there is the impression that it takes longer.

To test whether this impression is correct a sample ($n = 12$) is taken with $\bar{x} = 92.2$. We assume that the production time is normal with $\sigma^2 = 144$. Verify whether this impression is correct at significance level 5%.

## Solution:

$$\begin{cases} H_0: & \mu = 89 \\ H_a: & \mu > 89 \end{cases}$$

In the example, we reject $H_0$ if the test statistic is within the interval $[z_{0.05}, +\infty[= [1.645, +\infty[$.

We have that $n = 12$, $\bar{x} = 92.2$ and $\sigma^2 = 144$ such that

$$t = \frac{92.2 - 89}{12/\sqrt{12}} = 0.9237 < 1.645.$$

We can not reject $H_0$ at significance level 5%.

There is insufficient evidence to show that it takes longer to produce on Monday morning.

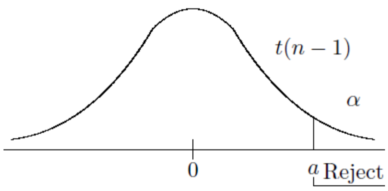
## Case 2:

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_a: & \mu > \mu_0 \end{cases}$$

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

The rejection region is given by

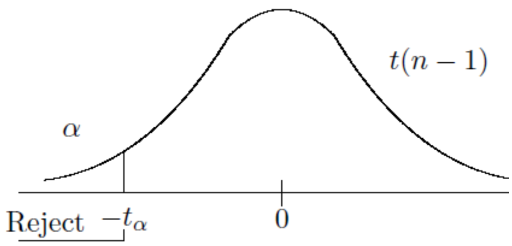
$$[t_\alpha, +\infty[.$$




$$\begin{cases} H_0: & \mu = \mu_0 \\ H_a: & \mu < \mu_0 \end{cases}$$

We have that the rejection region is

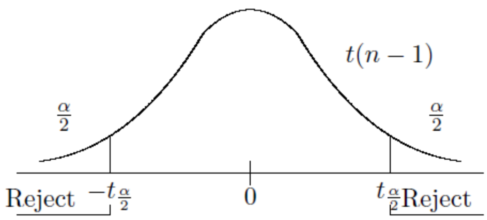
$$]-\infty, -t_\alpha].$$




$$\begin{cases} H_0: & \mu = \mu_0 \\ H_a: & \mu \neq \mu_0 \end{cases}$$

We reject $H_0$ if $\bar{X}$ is far from $\mu_0$ in either direction. The rejection region is

$$]-\infty, -t_{\frac{\alpha}{2}}] \cup [t_{\frac{\alpha}{2}}, +\infty[.$$



Case 3: For large sample size

the test statistic

$$T = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \approx N(0,1) \text{ under } H_0$$

$$T = \frac{\bar{X} - \mu_0}{\sqrt{S^2/n}} \approx N(0,1) \text{ under } H_0$$

- **Matched pairs:**

$$H_0 \ : \ \mu_D = d \quad , d \in I\!R.$$

versus

$$\begin{aligned} H_a &: \quad \mu_D &> d \\ H_a &: \quad \mu_D &< d \\ H_a &: \quad \mu_D &\neq d \end{aligned}$$

We define the test statistic, under $H_0$,

$$T = \frac{\bar{D} - d}{\sqrt{S_D^2/n}} \sim t(n-1)$$

**Example:**
In 10 women the systolic blood pressure (mm Hg) is measured at the beginning of a clinical trial. Afterwards they have a fertility treatment with hormones. During this treatment they are again measured.

| Id | before | during | Id | before | during |
|---|---|---|---|---|---|
| 1 | 115 | 128 | 6 | 138 | 145 |
| 2 | 112 | 115 | 7 | 126 | 132 |
| 3 | 107 | 106 | 8 | 105 | 109 |
| 4 | 119 | 128 | 9 | 104 | 102 |
| 5 | 115 | 122 | 10 | 115 | 117 |

In the example, we have that

$$\begin{aligned} H_0 &: \quad \mu_D &= 0 \\ H_a &: \quad \mu_D &\neq 0 \end{aligned}$$

We get the test statistic

$$T = \frac{\bar{D} - 0}{\sqrt{S_D^2/10}} \sim t(9).$$

With $\alpha = 5\%$, the rejection region is

$$]-\infty, -t_{0.025}] \cup [t_{0.025}, +\infty[ \; = \; ]-\infty, -2.262] \cup [2.262, +\infty[$$

From the data, we get

| individu $i$ | $D_i = Y_i - X_i$ |
|---|---|
| 1 | 13 |
| 2 | 3 |
| 3 | -1 |
| 4 | 9 |
| 5 | 7 |
| 6 | 7 |
| 7 | 6 |
| 8 | 4 |
| 9 | -2 |
| 10 | 2 |

and

$$\begin{aligned} \bar{D} &= 4.8 \\ S_D^2 &= \frac{187.6}{9} = 20.8444. \end{aligned}$$

we have that

$$t = \frac{4.8 - 0}{\sqrt{20.8444/10}} = 3.3247$$

We reject $H_0$ at significance level 5%.

The hormones have a significant effect on the systolic blood pressure.

**Remarque:**

We can find a $(1-\alpha)100\%$ confidence interval for the mean difference $\mu_D$,

$$\left[ \bar{D} - t_{\frac{\alpha}{2}} \sqrt{\frac{S_D^2}{n}}, \bar{D} + t_{\frac{\alpha}{2}} \sqrt{\frac{S_D^2}{n}} \right].$$

- **Independent samples:**

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

versus

$$H_a : \mu_1 - \mu_2 > d$$
$$H_a : \mu_1 - \mu_2 < d$$
$$H_a : \mu_1 - \mu_2 \neq d$$

if $\sigma^2 = \sigma_1^2 = \sigma_2^2$ unknown,

$$T = \frac{\bar{Y} - \bar{X} - [\mu_2 - \mu_1]_0}{\sqrt{S_P^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t(n_1 + n_2 - 2)$$

with $S_P^2$ a **pooled variance**

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

if $\sigma_1^2 \neq \sigma_2^2$ and $n_1, n_2 \geq 30$,

$$T = \frac{\bar{Y} - \bar{X} - [\mu_2 - \mu_1]_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim N(0, 1).$$

if $\sigma_1^2 \neq \sigma_2^2$,

$$T = \frac{\bar{Y} - \bar{X} - [\mu_2 - \mu_1]_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t(\nu)$$

with

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

**Example:**

In an experiment, we compare the results of treatments $A$ and $B$.

| | | | | | | |
|---|---|---|---|---|---|---|
| Treatment $A$ : | 17 | 19 | 15 | 18 | 21 | 18 |
| Treatment $B$ : | 18 | 15 | 13 | 16 | 13 | |

Investigate whether treatment $A$ gives better results?

We assume that both populations are normal and have equal variances.

$$\begin{aligned} H_0 &: \mu_2 - \mu_1 &= 0 \\ H_a &: \mu_2 - \mu_1 &< 0 \end{aligned}$$

The test statistic is given by, under $H_0$,

$$T = \frac{\bar{Y} - \bar{X} - 0}{\sqrt{S_P^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t(n_1 + n_2 - 2).$$

We get for $\alpha = 5\%$ the rejection region

$$]-\infty, -t_{0.05}] = ]-\infty, -1.833].$$

From the data, we have

$$\begin{array}{lll} n_1 = 6 & \bar{x} = 18 & s_1^2 = \frac{20}{5} = 4 \\ n_2 = 5 & \bar{y} = 15 & s_2^2 = \frac{18}{4} = 4.5. \end{array}$$

The pooled variance $s_P^2$ is given by

$$s_P^2 = \frac{5 \times 4 + 4 \times 4.5}{5 + 6 - 2} = 4.22$$

and we have

$$t = \frac{15 - 18}{\sqrt{4.22 \left( \frac{1}{6} + \frac{1}{5} \right)}} = -2.41 < -1.833.$$

We reject $H_0$ at significance level 5%. This means that treatment $A$ has significant better results than treatment $B$

**Remarque :**
We can derive a confidence interval for $\mu_2 - \mu_1$.

For example when $\sigma^2 = \sigma_1^2 = \sigma_2^2$,

$$\bar{Y} - \bar{X} \pm t_{n_1+n_2-2, \frac{\alpha}{2}} \sqrt{S_P^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}.$$