# Introduction to Hypothesis Testing

1. One-sample Student t-test
2. Test for paired data
3. Two-sample Student t-test

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MF9130E – Introductory Course in Statistics
10.04.2024

# One-sample Student t-test

## One-sample Student t-test (shortly, one sample t-test)

- The one-sample Student t-test is one of the most frequently applied tests in statistics. It is used to **test a certain hypothesis about the unknown population mean $\mu$**

## Background

- The t-test was devised by William Sealy Gosset, working for **Guinness brewery** in Dublin, to cheaply monitor the quality of stout
- Published in Biometrika in 1908 under the pen name **"Student"** as Guinness regarded the fact that they used statistics a trade secret

# P-value

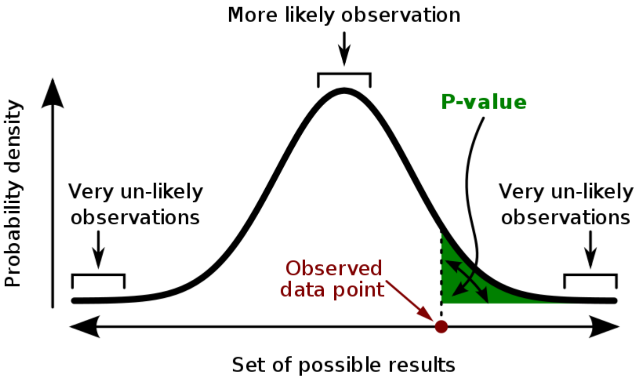
## Definition

The probability that the observed result, or a result more extreme, is true, given $H_0$ is true.

Then:

- The p-value is a measure of how likely our observed result is, under the $H_0$ assumption.
- If the p-value is small, then what we have observed is rare under $H_0$, which means we have *evidence against it.*
- p-values are used to *evaluate* the hypothesis test result, in terms of the *strength of the evidence* that the test provides.

# P-value


More likely observation
Probability density
P-value
Very un-likely observations
Very un-likely observations
Observed data point
Set of possible results

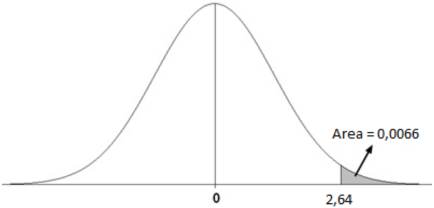
## The one sample t-test: an example

- 30 measures of lactate dehydrogenase (LD)
- **Question:** $\mu = 105$?
- **Test:** $H_0$: $\mu = 105$, $H_a$: $\mu > 105$
- We know that if $H_0$ is true, then

$$T_0 = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \to t(n-1),$$

  $T_0$ is called **test statistic**
- For our example: $T_0 = \frac{108.8 - 105}{7.88/\sqrt{30}} = 2.64$
- When would you reject the null hypothesis? Two options:
  - when $T_0$ is large, meaning when $T_0 > t_{n-1,\alpha}$, **OR**
  - when $p < \alpha$
- In our example: $T_0 = 2.64 > t_{29,0.05} = 1.699 \to$ Rejection
- When the test is **two-sided** ($H_a$: $\mu \neq 105$), use $t_{29,0.025} = 2.04$

## How to get the P-value




- If two-sided test: $p = 2P_{H_0}(t > |T_0|)$
- R or other statistical softwares produce the p-value automatically

# Paired data

## Paired measurements

- In medical settings we often deal with **paired measurements**, which is two outcomes measured on
  - the *same* individual under different exposure (or treatment) circumstances
  - *two* individuals matched by certain key characteristics
- The pairing in the data is taking into account by considering the *differences* between each pair of outcome observations. In that way the **data are turned into a single sample** of differences

## Paired measurements

2 measures of each individual (for example before/after treatment)

| Individual | Measure 1 | Measure 2 |
|---|---|---|
| 1 | $X_{11}$ | $X_{12}$ |
| 2 | $X_{21}$ | $X_{22}$ |
| 3 | $X_{31}$ | $X_{32}$ |
| ... | ... | ... |

## Example: 7.3 in Kirkwood & Sterne

We consider the results of a **clinical trial** to test the effectiveness of a sleeping drug. The sleep of ten patients was observed during one night with the **drug** and one night with **placebo**. For each patient a *pair* of sleep times, was recorded and the *difference* between these calculated

| | Hours of sleep | | |
|---|---|---|---|
| **Patient** | **Drug** | **Placebo** | **Difference** |
| 1 | 6.1 | 5.2 | 0.9 |
| 2 | 6.0 | 7.9 | -1.9 |
| 3 | 8.2 | 3.9 | 4.3 |
| 4 | 7.6 | 4.7 | 2.9 |
| 5 | 6.5 | 5.3 | 1.2 |
| 6 | 5.4 | 7.4 | -2.0 |
| 7 | 6.9 | 4.2 | 2.7 |
| 8 | 6.7 | 6.1 | 0.6 |
| 9 | 7.4 | 3.8 | 3.6 |
| 10 | 5.8 | 7.3 | -1.5 |
| **Mean** | $\bar{X}_1 = 6.66$ | $\bar{X}_0 = 5.58$ | $\bar{X} = 1.08$ |

The observed **mean difference** in sleep time was $\bar{X} = 1.08$ hours, and the empirical **standard deviation** of the differences was $s = 2.31$. The estimated **standard error** of the differences is $s/\sqrt{n} = 2.31/\sqrt{10} = 0.73$ hours

A 95% **confidence interval** for the mean difference in sleep time in the population is given by

$$(1.08 - 2.26 \times 0.73, 1.08 + 2.26 \times 0.73) = (-0.57, 2.73),$$

where 2.26 is the two-sided **5% point** of the $t$ distribution with $(n-1) = 9$ degrees of freedom

The **mean difference** in sleep time was $\bar{X} = 1.08$ hours, and the estimated **standard error** was $s/\sqrt{n} = 0.73$ hours. The **test statistic** is given by

$$t = 1.08/0.73 = 1.48,$$

which is $t$ distributed with $(n-1) = 9$ degrees of freedom when the null hypothesis of no effect is true. The corresponding $P$-**value**, which is the probability of getting a $t$ value with a size as large as this or larger in a $t$ distribution with 9 degrees of freedom, is

$$p = 0.17$$

So, *there is no evidence against the null hypothesis that the drug does not affect sleep time*

# Two sample t-test

So far...

- Tests and confidence intervals for
  - Single sample
  - Paired samples
- We know how to test (the procedure)

Now:

- Test for the difference in the mean of two independent samples

## The data: two different settings. Now focus on situation 2

1. Paired data: 2 measures of each individual (for example before/after treatment)

| Individual | Measure 1 | Measure 2 |
|---|---|---|
| 1 | $X_{11}$ | $X_{12}$ |
| 2 | $X_{21}$ | $X_{22}$ |
| 3 | $X_{31}$ | $X_{32}$ |
| … | … | … |

2. 2 groups: 1 measure of each individual, each which corresponds to a group (for example sick/healthy people)

| Group 1 | | Group 2 | |
|---|---|---|---|
| Ind. | Measure | Ind. | Measure |
| 1 | $X_{11}$ | 1 | $X_{12}$ |
| 2 | $X_{21}$ | 2 | $X_{22}$ |
| … | | … | |
| | | 14 | $X_{14\ 2}$ |
| 15 | $X_{15\ 1}$ | | |

## The two sample t-test

- The **null hypothesis** is given by

$$\mathrm{H}_0 : \mu_1 = \mu_0 \quad \text{or} \quad \mathrm{H}_0 : \mu_1 - \mu_0 = 0,$$

  i.e. there is *no difference between the population means* in the two groups

- The **test statistic** is given by

$$t = \frac{\bar{X}_1 - \bar{X}_0}{s\sqrt{(1/n_1 + 1/n_0)}},$$

  which follow a $t$ distribution with $(n_1 + n_0 - 2)$ degrees of freedom. Here, $s$ is the common estimate of the population standard deviation:

$$s = \sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2}\right]}$$

## Example: 7.2 in Kirkwood & Sterne

We return to the data of **birth weights**. The **test statistic** is given by

$$t = \frac{3.1743 - 3.6267}{0.4121\sqrt{(1/14 + 1/15)}} = -\frac{0.4524}{0.1531} = -2.95$$

The corresponding ***P*-value** calculated from the $t$ distribution with $(14 + 15 - 2) = 27$ degrees of freedom is given as:

$$p = 0.006$$

Therefore, *the data suggest that smoking during pregnancy reduces the birthweight of the baby*

## Test statistic

- In the one sample t-test we had

  $T = \frac{\bar{X}_n - \mu_0}{s}\sqrt{n}$

- Now $T = \frac{\bar{X}_1 - \bar{X}_2}{s_p}\sqrt{n_p}$

$S_p$ is the pooled standard deviation $\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$ and $n_p = \frac{n_1 \cdot n_2}{n_1+n_2}$.

## Test statistic (cont.)

- $T \sim t_{n_1+n_2-2}$ under the null hypothesis $H_0$
- Rejection and conclusion:

| | |
|---|---|
| $H_0$ | $\mu_1 = \mu_2$ |
| Rejection, if | $\|T_0\|$ large |
| Rejection, if | $P = 2\,P_{H_0}(t > \|T_0\|) < \alpha$ |
| Conclusion | $\mu_1 \neq \mu_2$ |

## Small samples, unequal standard deviations

- When the population standard deviations, $\sigma_1$ and $\sigma_0$, of the two groups are different, and the sample size, $n$, is not large, the main possibilities are:
  - Use a **transformation** on the data which makes the standard deviations similar so that methods based on the $t$ distribution can be used
  - Use **non-parametric** methods based on ranks
  - Use either the **Fisher-Behrens** or the **Welch** tests, which allow for unequal deviations
  - Estimate the difference between the means using the original measurements, but use **bootstrap** methods to derive confidence intervals

## How to check for normal distribution

- Box-plot
- Histograms
- Q-Q plot

## What if the data does not look normal?

- Try to find a meaningful transformation
- Use a test which does not assume normally distributed data

$\rightarrow$ Lecture on transformations and non-parametric methods in day 1 of week 2