Social Robotics

Albert-Ludwigs-Universität Freiburg

Felix Lindner, Laura Wächter, Bernhard Nebel

SoSe 2019

t-Test

Outline

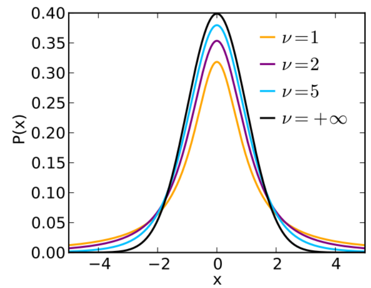
- The t-Distribution and the t-Statistics
- One-Sample t-Test
  - Based on a sample of time intervals: Do children spend more time with their toy robot than the target play time of half an hour per day?
- Paired t-Test
  - Based on two dependent samples of time intervals: Do children spend more time with their toy robot after they were told about the robot's capabilities?
- Two-Sample t-Test
  - Based on two independent samples of time intervals: Do older people spend more time with the robot compared to younger people?

Student's t distribution

If $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2(\nu)$ are independent random variables, then the variable T follows a t-distribution with ν degrees of freedom:

$$T = \frac{Z}{\sqrt{\frac{U}{\nu}}} \sim t_\nu$$

mean: 0, variance: $\nu/(\nu-2)$



The t statistics

- Given the mean $\overline{X}$ of a sample of size N drawn from a population with mean μ and standard deviation σ, we already know that $z = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{N}}}$ follows a normal standard distribution, $z \sim \mathcal{N}(0,1)$.
- It is also known that, if the population is normally distributed, then $u = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$
  - see proof https://onlinecourses.science.psu.edu/stat414/node/174
- By definition $\frac{z}{\sqrt{\frac{u}{v}}} \sim t_{n-1}$, and therefore, also
  - $t = \frac{z}{\sqrt{\frac{u}{v}}} = \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)s^2}{\sigma^2}}{(n-1)}}} = \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\frac{s}{\sigma}} = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \times \frac{\sigma}{s} = \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$
- Compare this to z! If we estimate σ by s, we obtain a t-distributed test statistics.

One-Sample t-Test I

- Using t-Test as compared to z-Test we can lift the assumption that σ is known, because t-Test empowers us to rely merely on s.
- That is, we can test how likely a given sample stems from a population with mean μ (fullstop).

Example

The robotic toy company assumes that children will play $\mu_0 = 100$ minutes per day with the robot on average (H_0). The researchers hypothesize that things will turn out different $\mu \neq \mu_0$ (H_1). Their six-day sample is: 110, 107, 100, 101, 104, 105, $\overline{X} = 104.5$, $s = 3.73$, $t = \frac{104.5-100}{\frac{3.73}{\sqrt{6}}} = 2.95$.

One-Sample t-Test II

1. As usual, to make a decision whether or not to reject H_0, we fix an α.
2. Then, we check if the calculated t exeeds some critical value:
   - H_1 Undirectional: $t \leq t_{n-1;\frac{\alpha}{2}}$ or $t \geq t_{n-1;1-\frac{\alpha}{2}}$
   - H_1 Less: $t \leq t_{n-1;\alpha}$
   - H_1 Greater: $t \geq t_{n-1;1-\alpha}$

Example continued

We want to test with significance level 5% (i.e., $\alpha = 0.05$). We reject H_0 if $t \leq t_{6-1;\frac{0.05}{2}} = -2.57$ or $t \geq t_{6-1;1-\frac{0.05}{2}} = 2.57$. Because our t-value is 2.95, H_0 can be rejected in support of H_1.

Confidence Interval

- We can also compute the p%-confidence interval for the sample mean, this time using t instead of z:

$$[\overline{X} - \frac{s}{\sqrt{N}} \times |t_{\frac{(100-p)/100}{2},df}|, \overline{X} + \frac{s}{\sqrt{N}} \times |t_{\frac{(100-p)/100}{2},df}|]$$

Example continued

For the mean $\overline{X} = 104.5$ with $s = 3.73$ and $N = 6$, the 95% confidence interval is $[100.59, 108.41]$.

One-Sample t-Test III

- Finally, the p-value can be computed:
  - H_1 Undirectional: $P(x \leq -|t|) + 1 - P(x \leq |t|)$
  - H_1 Less: $P(x \leq -|t|)$
  - H_1 Greater: $1 - P(x \leq |t|)$

Example continued

The t-Value was 2.95. The probability of some value at least as extreme as 2.95, is $P(x \leq -2.95) + 1 - P(x \leq 2.95) = 0.032$. In R: `p.value = pt(-2.95, df=5) + 1-pt(2.95, df=5)`.

Cohen's d

- A significant difference need not necessarily be a big difference.
- Cohen's d can be used to compute the effect size: $d = \frac{|\overline{X} - \mu|}{s}$
- According to Cohen, d between 0.2 and 0.5 is a small effect, a medium effect is between 0.5 and 0.8, and a d above 0.8 counts as a big effect.

Example

Given $\mu = 100$, for the mean $\overline{X} = 104.5$ and $s = 3.73$, the Cohen's d is $d = 4.5/3.73 = 1.2$.

Reporting t-Tests

Report

The time children spend with their robotic toy differs significantly from 100 minutes per day ($\overline{X} = 104.5$, $s = 3.73$, $t(5) = 2.95$, $p = 0.032$, 95% CI $[100.59, 108.41]$, $d = 1.2$).

Paired t-Test: Motivation

The t-Test statistics can be used for something more practical than the rather artificial test against a fixed μ: Testing for the difference of paired data. Consider the following setting:

Example

Five children Child-1 to Child-5 are tested for play time change w.r.t. to their robotic toy after they have been told about the robot's capabilities.

	Child-1	Child-2	Child-3	Child-4	Child-5
Before	10	17	17	15	19
After	11	25	20	18	22

Paired t-Test: Procedure by Example

	Child-1	Child-2	Child-3	Child-4	Child-5
Before	10	17	17	15	19
After	11	25	20	18	22

- H_1: Before and After differ $\mu_B \neq \mu_A$, H_0: There is no difference between Before and After $\mu_B = \mu_A$.
- H_0 can also be written as $\mu_B - \mu_A = 0$
- Hence: The data set we actually analyze is $D_i = B_i - A_i$:
  - -1, -8, -3, -3, -3, $\overline{D} = -3.6, s_D = 2.61$
  - $t = \frac{\overline{D}-0}{\frac{s_D}{\sqrt{N}}} = \frac{-3.6}{\frac{2.61}{\sqrt{5}}} = -3.084$
  - $-3.084 \leq 2.776 = t_{4;2.5\%}$
  - $p = P(x \leq -3.084) + 1 - P(x \leq 3.084) = 0.0367$

Cohen's d

- Cohen's d for paired t-Test: $d = \frac{|\overline{D}|}{s_D}$

Example

For the mean $\overline{D} = -3.6$ and $s = 2.61$, the Cohen's d is $d = 3.6/2.61 = 1.38$.

Note on Assumptions

- The t-Test can be used in case various assumptions are fulfilled:
  1. The data is interval-scaled.
     - So that computing $\overline{X}$ and s makes sense.
  2. The population is normally distributed.
     - Only in this case, $\frac{(n-1)s^2}{\sigma^2}$ is χ^2_{n-1}-distributed. Also check this useful video: `https://www.youtube.com/watch?v=V4Rm4UQHij0`
  3. For two-sample t-Tests (next), the homogeneity of the variances is additionally assumed.
- It is reported, though, that in simulation studies the t-Test proves very robust against violations of these assumptions.

Two-Sample t-Test: Motivation

Example

The robot has been deployed to older people and to younger people. The alternative hypothesis is that there will be a mean difference in time spent with the robot ($H_1 : \mu_1 \neq \mu_2$, $H_0 : \mu_1 = \mu_2$). The two samples look like this:

- Younger: 101, 100, 99, 93, 120, 89, 102, $\overline{X}_1 = 100.57, s_1 = 9.78$
- Older: 88, 90, 90, 87, 86, 90, 100, $\overline{X}_2 = 90.14, s_2 = 4.63$

- This time, we cannot proceed like in the paired test, because the scores are independent, and we also allow for different sample sizes.

Two-Sample t-Test: Idea

- Intuitively, if the two means μ_1, μ_2 are equal, as stated by H_0, then $\overline{X}_1 - \overline{X}_2$ should be close to zero.
- Therefore, the t-statistics looks like this: $t = \frac{(\overline{X}_1 - \overline{X}_2) - 0}{s_{\overline{X}_1 - \overline{X}_2}}$, where $s_{\overline{X}_1 - \overline{X}_2}$ is the standard deviation of the sample mean difference.

Two-Sample t-Test: Standard Error of Difference

- How to compute $s_{\overline{X}_1 - \overline{X}_2}$?
- From the central limit theorem, we already know that the variances of $\overline{X}_i$ are $\frac{\sigma_i^2}{N_i}$.
- Because the samples are drawn independently, it holds that $\sigma^2_{\overline{X}_1 - \overline{X}_2} = \frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}$.
- Under the assumption that $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we get $\sigma^2_{\overline{X}_1 - \overline{X}_2} = \left(\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}\right)$, hence $\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}\right)}$ (Standard Error of the Difference)
- Next, we estimate σ^2 by some estimate of the common variance s_p^2, and we get $s_{\overline{X}_1 - \overline{X}_2} = \sqrt{\left(\frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}\right)}$.

Two-Sample t-Test: Estimate of s_p^2

- How to compute s_p^2?
- We weigh the s_i^2 according to the sample sizes

$$s_p^2 = \frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{(N_1-1)+(N_2-1)}$$

- The t-statistics then reads

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\left(\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{(N_1-1)+(N_2-1)}\right)}{N_1} + \frac{\left(\frac{(N_1-1)s_1^2+(N_2-1)s_2^2}{(N_1-1)+(N_2-1)}\right)}{N_2}}} \sim t_{N_1+N_2-2}$$

Note on the case $N_1 = N_2$

- As can be easily verified, the t-statistics simplifies to $t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\frac{s_1^2+s_2^2}{2} + \frac{s_1^2+s_2^2}{2}}{N}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\frac{2(s_1^2+s_2^2)}{2}}{N}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2+s_2^2}{N}}} \sim t_{2N-2}$ in case $N_1 = N_2$

Example

- Younger: 101, 100, 99, 93, 120, 89, 102, $\overline{X}_1 = 100.57, s_1 = 9.78, s_1^2 = 95.65$
- Older: 88, 90, 90, 87, 86, 90, 100, $\overline{X}_2 = 90.14, s_2 = 4.63, s_2^2 = 21.43$
- $N_1 = N_2 = N = 7, df = 12$
- $t = 2.55 \geq t_{12;97.5\%} = 2.18$, H_0 rejected with $\alpha = 0.05$.

Cohen's d

- Cohen's d for two-samples t-Test: $d = \frac{|\overline{X}_1 - \overline{X}_2|}{\sqrt{\frac{s_1^2 + s_2^2}{2}}}$

Example

For the means $\overline{X}_1 = 100.57$, $\overline{X}_2 = 90.14$ and $s_1^2 = 95.65$, $s_2^2 = 21.42$, the Cohen's d is $d = 10.43/7.65 = 1.36$.

Note on Welch-Test

- In the derivation of the t-statistics, we assumed that the samples stem from distributions of equal variance. Before R runs a t-Test, this assumption is tested. In case the variances are too different, the Welch-Test is run.
- In Welch test, we have $s^2_{\overline{X}_1 - \overline{X}_2} = \frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}$ instead of $s^2_{\overline{X}_1 - \overline{X}_2} = \frac{s_p^2}{N_1} + \frac{s_p^2}{N_2}$. However, in this case, t does not follow a t-distribution anymore.
- It turns out that t can still be used as a test statistics if the degree of freedom is adapted:

$$df_{corr} = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2(N_1-1)} + \frac{s_2^4}{N_2^2(N_2-1)}}$$, just for your information :-)

Discussion: The Problem with Multiple Tests

- What if we want to compare more than two samples?

Sketches

Intentionally left blank :-)