# Contents



# Day 20

## Two Independent Samples t-Test

When it's used: randomized with unrelated treatments & control groups OR observational study with comparing two unrelated groups.

Very common: compute mean if everyone got "new" treatment to mean if everyone got "control" treatment. Example; two groups, one smoking and the other group not smoking.

Data looks like: one numerical response variable measured in two independent groups

**Example**

| Lung Cap (Smoke) | | Lung Cap (No Smoke) | |
|---|---|---|---|
| # | | # | |
| # | | # | |
| # | | # | |
| # | | # | |
| # | | # | |
| | | # | |
| | | # | |
| | | # | |
| | | # | |
| | | # | |
| | | | |
| | | | |
| TOTAL | | TOTAL | |
| 5 | | 10 | |

Figure 1: Numbers to numbers

| Lung Cap | | Group |
|---|---|---|
| # | | Smoker |
| # | | Smoker |
| # | | Smoker |
| # | | No smoker |

Figure 2: Numbers to groups

# Theory & Concepts

Let $\mu_1$ = population mean in group 1. $\mu_2$ = population mean in group 2.

Example: $\mu_1$ = population mean cholesterol level if everyone got new drug. $\mu_2$ = population mean cholesterol level if everyone current drug.

We want to do inference on $\mu_1 - \mu_2$, the difference of population means.

Recall:

$$t = \frac{STATISTIC - PARAMETER}{STANDARD\ ERROR}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_1)}{STANDARD\ ERROR}$$

Now we have two sample, so we have:

- $\bar{x}_1$
- $S_1$
- $n_1$
- $\bar{x}_2$
- $S_2$
- $n_2$

By CLT,

$$\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1}{\sqrt{n_1}})$$

$$\bar{X}_2 \sim N(\mu_2, \frac{\sigma_2}{\sqrt{n_2}})$$

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{(\sigma_1)^2}{n_1} + \frac{(\sigma_2)^2}{n_2}}]$$

$$Var(\bar{X}_1 - \bar{X}_2) = Var(\bar{X}_1) + Var(\bar{X}_2)$$

$$(\sigma_1)^2 + (\sigma_2)^2$$

Serious problem: we need to estimate two standard deviations!

## Solution 1 : Assume $\sigma_1 = \sigma_2$

Estimate using pooled standard deviation $S_{\text{p}}$

$$\sqrt{weighted\,average\,of\,variance}$$

**this is not needed**

## Satterthwaite Approximation

$$t \sim t(df)$$

Where `df` is calculated by Satterthwaite approximation.

In general, df is NOT an integer

## Welch's t-Test

Almost always, $H_0$:

- $\mu_1 - \mu_2 = 0$
- $(\mu_1 = \mu_2)$

Under $H_0$,

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

Almost always: use a two-sided test

Alternative hypothesis: difference between treatment & control.

**Neyman-Pearson:**

$H_1$:

- $\mu_1 - \mu_2 = \Delta$
- $\Delta$ = desired effects size in original units
- Two sided critical region:

Compute $t_{observed}$

Accept $H_1$ if $t_{observed}$ in critical region. Else accept $H_0$.

## NHST Approach

$H_a$: $\mu_1 - \mu_2 \neq 0$

NOTE: can use $\mu_1 - \mu_2 > 0$ or $\mu_1 - \mu_2 < 0$ but be very careful

Under $H_0$:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}} \sim t(K)$$

**K is given by software**

Two-sided p-value = $P(|t| \geq |t_{\text{observed}}||H_0 istrue)$

If p-value $\leq$ significance level, reject $H_0$ & accept $H_a$. Else fail to reject $H_0$.

## Example

Study comparing fat consumption of early vs late eaters. We want to know whether there is a difference in fat consumption.

Use NHST approach

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

Let $\mu_1$ = population mean fat consumption in early eaters

Let $\mu_2$ = population mean fat consumption in later eaters

### Data

Early eaters (n=202)

- $\bar{x} = 23.1g$
- $S = 12.5g$

Late eaters (n=200)

- $\bar{x} = 21.4g$
- $S = 8.2g$

$\mu_1 - \mu_2 > 0 \implies$ always above

$\mu_1 - \mu_2 < 0 \implies$ always below

Under $H_0$:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{(s_1)^2}{n_1} + \frac{(s_2)^2}{n_2}}}$$

The answer is about 1.614.

Satterthwaite approximation gives df = 347.41

```
pt(1.614, df = 347.41, lower.tail = FALSE)
```

p-value = 2(0.054) = 0.108

At 5% significance level, we fail to reject $H_0$ because $0.108 > 0.05$. We failed to find a statistically significant difference in fat consumption.