STAT2201

Analysis of Engineering & Scientific Data

Unit 8

Slava Vaisman

The University of Queensland
School of Mathematics and Physics

Two Sample Inference

- This time, we consider two different samples.

$$x_1, \dots, x_{n_1}, \quad y_1, \dots, y_{n_2}.$$

- These samples are modeled as an i.i.d. sequence of random variables

$$X_1, \dots, X_{n_1}, \quad Y_1, \dots, Y_{n_2}.$$

- The n_1 is not necessarily equal to the n_2.
- We model $\{X_i\}_{1 \le i \le n_1}$ and $\{Y_i\}_{1 \le i \le n_2}$ with

$$X_i \sim \mathrm{N}\left(\mu_1, \sigma_1^2\right), \quad Y_i \sim \mathrm{N}\left(\mu_2, \sigma_2^2\right),$$

- and distinguish between the following cases:

$$\begin{cases} \text{equal variances:} & \sigma_1^2 = \sigma_2^2 = \sigma^2, \\ \text{unequal variances:} & \sigma_1^2 \neq \sigma_2^2. \end{cases}$$

Medical treatment

Recall experimental medical treatment example, in which 14 subjects were randomly assigned to control or treatment group. The survival times (in days) are shown in the table below.

	Data	Mean
Treatment group	91, 140, 16, 32, 101, 138, 24	77.428
Control group	3, 115, 8, 45, 102, 12	47.5

We asked:

- Did the treatment prolong the survival?
- Is the observed result *significant*, or due to a *chance*?

Note that we are dealing with two samples: $x_1, \ldots, x_7$ and $y_1, \ldots, y_6$. Note that $n_1 = 7$ and $n_2 = 6$.

Inference

	Data	Mean
Treatment group	91, 140, 16, 32, 101, 138, 24	77.428
Control group	3, 115, 8, 45, 102, 12	47.5

- We could carry single sample inference for each population separately. Namely, for:

$$\mu_1 = \mathbb{E}[X_i], \text{ and } \mu_2 = \mathbb{E}[Y_i].$$

- However, we are generally more interested to know if the treatment helps (prolongs the survival time).
- Specifically, we focus on the difference in means:

$$\Delta_\mu = \mu_1 - \mu_2 = \mathbb{E}[X_i] - \mathbb{E}[Y_i].$$

Inference

- For $\Delta_\mu = \mu_1 - \mu_2 = \mathbb{E}[X_i] - \mathbb{E}[Y_i]$, we can carry out inference jointly.

- Specifically, it is common to examine:

  1. $\Delta_\mu > 0 \Rightarrow \mu_1 > \mu_2$, or
  2. $\Delta_\mu < 0 \Rightarrow \mu_1 < \mu_2$, or
  3. $\Delta_\mu = 0 \Rightarrow \mu_1 = \mu_2$.

- We can also replace the zero with some Δ_0 to get:

  1. $\Delta_\mu > \Delta_0 \Rightarrow \mu_1 - \mu_2 > \Delta_0$, or
  2. $\Delta_\mu < \Delta_0 \Rightarrow \mu_1 - \mu_2 < \Delta_0$, or
  3. $\Delta_\mu = \Delta_0 \Rightarrow \mu_1 - \mu_2 = \Delta_0$.

A point estimator for Δ_μ

- A point estimator for Δ_μ is given by:

  $$\overline{X} - \overline{Y},$$

  where $\overline{X}$ and $\overline{Y}$ are sample means.

- The estimate from the data is given by $\overline{x} - \overline{y}$, where

  $$\overline{x} = \frac{1}{n_1} \sum_{i=1}^{n_1} x_i,$$

  and

  $$\overline{y} = \frac{1}{n_2} \sum_{i=1}^{n_2} y_i.$$

Estimating the variances

Point estimates for σ_1^2 and σ_2^2 are the individual sample variances:

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \overline{x})^2 , \quad s_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \overline{y})^2 . \qquad (1)$$

1. **Equal variances**: note that both s_1^2 and s_2^2 estimate σ^2. The so called pooled variance estimator can be obtained via:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

2. **Unequal variances**: just use (1) to obtain point estimates for σ_1^2 and σ_2^2.

The test statistic

Note that:

- $\mathbb{E}\left[\overline{X} - \overline{Y}\right] = \mathbb{E}\left[\overline{X}\right] - \mathbb{E}\left[\overline{Y}\right] = \Delta_0$
- The variance is:

$$\begin{aligned}
\mathrm{Var}\left(\overline{X} - \overline{Y}\right) &= \mathrm{Var}\left(\overline{X} + (-1)\overline{Y}\right) = \mathrm{Var}\left(\overline{X}\right) + (-1)^2\mathrm{Var}\left(\overline{Y}\right) \\
&= \mathrm{Var}\left(\overline{X}\right) + \mathrm{Var}\left(\overline{Y}\right).
\end{aligned}$$

This leads to the following test statistic T defined via (note the similarity to the one-sample tests we discussed):

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

The test statistic

We consider the statistic

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

under equal/unequal variance setting.

- **Equal variances**:

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\overline{X} - \overline{Y} - \Delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

- **Unequal variances**:

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

Equal variances

In the equal variance case, under H_0 it holds (approximately):

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2).$$

That is, the T test statistic follows a t-distribution with $n_1 + n_2 - 2$ degrees of freedom.

Unequal variances

In the unequal variance case, under H_0 it holds (approximately):

$$T = \frac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(\nu),$$

where

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

If ν is not an integer, may round down to the nearest integer (if we would like to use the table).

That is, the T test statistic follows a t-distribution with ν degrees of freedom.

Two sample t-test with equal variance

Testing Hypotheses on Differences of Mean, Variance Unknown and Assumed Equal (two sample T-Tests with equal variance)

Model: $X_i \overset{i.i.d.}{\sim} N(\mu_1, \sigma^2), \qquad Y_i \overset{i.i.d.}{\sim} N(\mu_2, \sigma^2).$

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0.$

Test statistic: $$t = \frac{\bar{x} - \bar{y} - \Delta_0}{s_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}, \qquad T = \frac{\overline{X} - \overline{Y} - \Delta_0}{S_p\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}.$$

Alternative Hypotheses	**P-value**	**Rejection Criterion for Fixed-Level Tests**
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$P = 2\left[1 - F_{n_1+n_2-2}\left(\lvert t\rvert\right)\right]$	$t > t_{1-\alpha/2,n_1+n_2-2}$ or $t < t_{\alpha/2,n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$P = 1 - F_{n_1+n_2-2}(t)$	$t > t_{1-\alpha,n_1+n_2-2}$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$P = F_{n_1+n_2-2}(t)$	$t < t_{\alpha,n_1+n_2-2}$

Two sample t-test with unequal variance

Testing Hypotheses on Differences of Mean, Variance Unknown and NOT Equal (two sample T-Tests with unequal variance)

Model: $X_i \overset{i.i.d.}{\sim} N(\mu_1, \sigma_1^2), \qquad Y_i \overset{i.i.d.}{\sim} N(\mu_2, \sigma_2^2).$

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0.$

Test statistic: $t = \dfrac{\bar{x} - \bar{y} - \Delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}, \qquad T = \dfrac{\overline{X} - \overline{Y} - \Delta_0}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}.$

Alternative Hypotheses	P-value	Rejection Criterion for Fixed-Level Tests
$H_1 : \mu_1 - \mu_2 \neq \Delta_0$	$P = 2\left[1 - F_v(\lvert t \rvert)\right]$	$t > t_{1-\alpha/2,v}$ or $t < t_{\alpha/2,v}$
$H_1 : \mu_1 - \mu_2 > \Delta_0$	$P = 1 - F_v(t)$	$t > t_{1-\alpha,v}$
$H_1 : \mu_1 - \mu_2 < \Delta_0$	$P = F_v(t)$	$t < t_{\alpha,v}$

$1 - \alpha$ Confidence Intervals

1. Equal variance case:

$$\mu_1 - \mu_2 \in \left(\overline{x} - \overline{y} \pm t_{1-\alpha/2, n_1+n_2-2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right)$$

2. Unequal variance case:

$$\mu_1 - \mu_2 \in \left(\overline{x} - \overline{y} \pm t_{1-\alpha/2, \nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right)$$

t-test example

```
Treatment = [91, 140, 16, 32, 101, 138, 24]
Control   = [3, 115, 8, 45, 102, 12  ]
UnequalVarianceTTest(Treatment,Control)
```

```
Output:

Two sample t-test (unequal variance)
------------------------------------
Population details:
    parameter of interest:   Mean difference
    value under h_0:         0
    point estimate:          29.928571428571443
    95% confidence interval: (-33.0286, 92.8857)
Test summary:
    outcome with 95% confidence: fail to reject h_0
    two-sided p-value:           0.3175326630084628
Details:
    number of observations:   [7,6]
    t-statistic:              1.0475473589407192
    degrees of freedom:       10.89399347312799
    empirical standard error: 28.570136875563534
```