T-Test of Unpaired data-Independent samples

R-Training workshop

17th-26TH February 2016

Pwani University

Dr. David Mburu (PhD)

Objectives

- Describe the sampling distribution of a difference between means
- Estimate the difference between means
- Calculate the standard error of the difference
  - For large samples (using normal distribution)
  - For smaller samples (T distr. requires similar standard deviation)
- Use T-Test in R to compare the difference of two means

- Calculate a confidence interval for the difference between two means

Assumptions & notation

- The underlying distribution of both groups is approximately normal

	Population	Sample	Population	Sample
Mean	μ_1	$\bar{x}_1$	μ_2	$\bar{x}_2$
SD	σ_1	s_1	σ_2	s_2

The difference is written as ($\bar{x}_1 - \bar{x}_2$).
This is used to estimate $\mu_1 - \mu_2$, the difference of the population means.

Two sample (unpaired) t-test

- Previously for matched pairs analysis

- - each observation was matched to another,
- - used the difference between the observations as the variable for analysis,
- - did not need to worry about the variability within each sample

For unpaired data

- There is no connection the two samples

- We need to account for the variability of both samples

- The variance will be larger, and hence the standard error of the mean will be larger than for paired analysis

Unpaired T-test

For unpaired data

There is no connection between the two samples

We need to account for the variability of **both samples**

- Get means, std dev & SE(mean) for each sample
- Then take the difference in means
- Calculate the standard error of the difference

Assumptions:

Each of two sample means is normally distributed

The sampling distribution of the difference ($\bar{x}_1 - \bar{x}_2$) is also normally distributed

The expected value of the mean of the difference is $\mu_1 - \mu_2$, the difference of the population means.

Standard error of a difference in means

The variances of the means pooled

$$\mathrm{Var}\,(\bar{x}_1 - \bar{x}_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2$$

1. For large samples

When both groups are large (sample size >30)
The standard error of the difference becomes:

$$\mathrm{SE}\,(\bar{x}_1 - \bar{x}_2) = \sqrt{(s_1^2/n_1 + s_2^2/n_2)}$$

This can be used in the same way as any other Standard Error:
Estimating the 95% CI around the difference in means.
Testing a null hypothesis H_0 for the difference in the means.

Calculating the standard error of the difference in means
- for small samples ($n<30$)

1. First look at the distributions of the groups. Are the standard deviations similar?

If so, we can calculate the pooled standard deviation

2. Calculate the common standard deviation

$$s = \sqrt{\left(\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{(n_1 - 1) + (n_2 - 1)}\right)}$$

3. Use the common s, to estimate the standard error by:

$$SE = s \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Calculating a 95% CI for the difference.

- The 95% CI of the difference in the means is calculated in the same way as any other 95% CI

  - Use the Z value from normal distribution for large samples.

- 95% CI = ($\bar{x}_1$ - $\bar{x}_2$) +/– 1.96 x SE (diff)

- For samples that have smaller sample size, then use the T values to generate the 95% CI

- 95% CI = ($\bar{x}_1$ - $\bar{x}_2$) +/– T $(_{\alpha/2,\mu})$ x SE (diff)

Hypothesis testing - the difference in the means.

- Hypothesis testing follows the same rules:

  - Define the null hypothesis. $H_0: (\mu_1 - \mu_2) = 0$
  - Use the difference in means and the standard error of the difference.
  - Obtain the z-value, and look up in tables of the normal distribution.

$$Z = (\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2) / SE\ (diff)$$

- For groups with adequate sample size, we use the normal distribution.

Unpaired T-test

Two sided test.
Implies that equal importance given to differences on both sides of the null hypothesis
Usual test, as do not know which treatment is better before the trial starts.

One sided test.
Only used if explicitly stated as the objective in the trial or study.
Greater significance but only for one comparison. Other comparison not significant, no matter what value of T is seen.

Summary of the comparison of two means.

Calculate the difference in the means, and the SE of the difference.

The null hypothesis is the diff = 0.

Use difference and SE(diff) to calculate 95% CI and to test the null hypothesis using ***ttest***.

Any other tips or ideas to assist in the comparison of means?

Practical. Analysing PCV in maltreat.

- Label the variable PCV
- Histogram to check its shape
- Get the means, SE and the 95% CI
- Test the hypothesis that the mean PCV of these children is 35%
- Test for differences between exposures, sexes, +/- fever, +/- gametrocytes, enrolled or not.