Hypothesis Testing

A Demonstration of the Two-sample and paired samples t-test

Dr James Baglin (james.baglin@rmit.edu.au)

Comparing the mean of two populations

To compare the mean of two independent populations, we select a sample from each population (not necessary with the same size) and use them to decide if the means, μ_1 and μ_2 are equal. This is called two sample test. We must have the following assumptions':

1- Both populations are normal or both sample sizes are greater than 30 (so we can apply central limit theorem).

2- either Variance of both populations are unknown but equal or both variances are unknown and we do not know if they are equal (latter is more common in real world).

We can test the homogeneity of variance using Levene’s test.

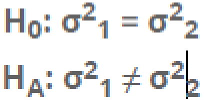
Hypothesis for Comparing the mean of two populations
Hypothesis for homogeneity test of variances of two populations

Hypothesis for testing mean of two populations:

Null hypothesis, H_0: $\mu_1 - \mu_2 = 0.0$

Alternative hypothesis, H_a : $\mu_1 - \mu_2 \neq 0.0$ or $\mu_1 - \mu_2 > 0.0$ or $\mu_1 - \mu_2 < 0.0$

Hypothesis for homogeneity test of variances of two populations:



H_0: $\sigma^2_1 = \sigma^2_2$

H_A: $\sigma^2_1 \neq \sigma^2_2$

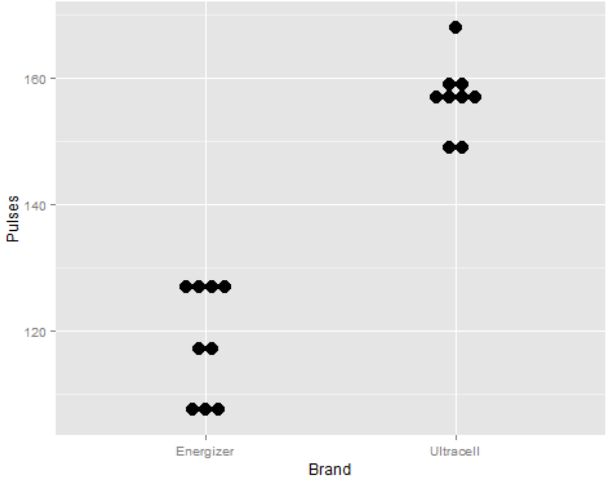
R code for homogeneity Levene's test :

```
 > leveneTest(Body_temp ~ Gender, data = Body_temp)
 Levene's Test for Homogeneity of Variance (center = median)
           Df     F value  Pr(>F)
group      1      0.04     0.84
           128
```

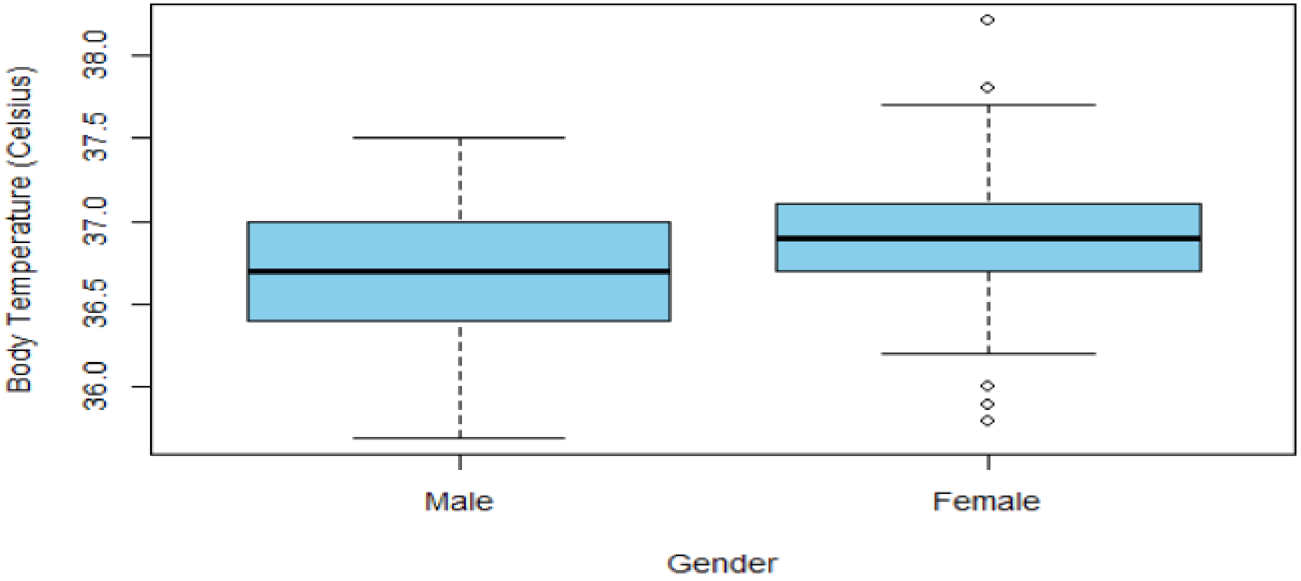
Two-sample *t*-test - Example

- Investigators compare the average number of pulses (e.g. camera flashes) to deplete a 1.5 V battery to 0.8 V (very flat) using a random sample of 9 Energizer and 9 Ultracell (Aldi) batteries (Approx. 50% cheaper).
- The data are contained in the `Battery.csv` dataset on the Data Repository
- The estimated difference between means = 118.22 - 156.67 = -38.45 pulses (Energizer - Ultracell)

```
> Battery_sub <- subset(Battery, subset = Voltage == 0.8)
> favstats(~Pulses | Brand, data = Battery_sub)
      Brand min  Q1 median  Q3 max     mean       sd n missing
1 Energizer 107 108    117 127 128 118.2222 9.148467 9       0
2 Ultracell 149 156    156 159 168 156.6667 5.700877 9       0
```



Two-sample *t*-test - Example



Two-sample *t*-test - Overview

- **Hypotheses for the two-sample (independent samples) *t*-test**:

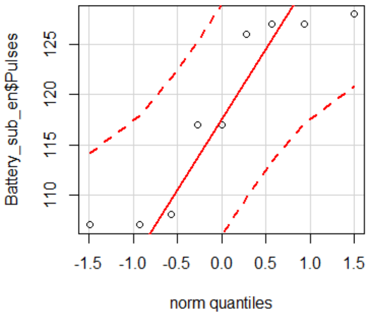
$$H_0: u_{Energizer} - u_{Ultracell} = 0$$

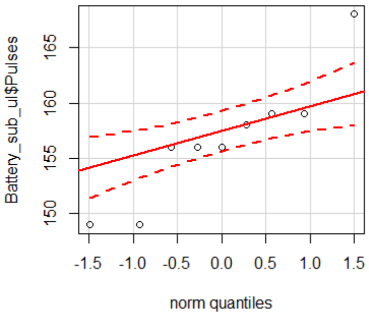
$$H_A: u_{Energizer} - u_{Ultracell} \neq 0$$

- **Assumptions**:
  - Comparing two independent population means with unknown population variance.
  - Population data are normally distributed or large sample used ($n > 30$ for both groups)
  - Population homogeneity of variance
- **Decision Rules**:
  - Reject H_0:
    - If *p*-value < 0.05 (α significance level)
    - If 95% *CI* of the difference between means does not capture $H_0: u_{Energizer} - u_{Ultracell} = 0$
  - Otherwise, fail to reject H_0.
- **Conclusion**:
  - Test will be statistically significant if we reject H_0
  - Otherwise, the test is not statistically significant.

Two-sample *t*-test - Assumptions - Normality

- Normality is only a problem in small samples (generally samples sizes less than 30) due to the CLT
- However, when we need to test normality the most (i.e, n < 30 in one group), there is no good method.
- Visual inspections might help…





- But, often they don’t because there is insufficient information…
- Sometime we just need to make an assumption or maybe look for alternative methods - e.g. nonparametric methods, e.g. randomisation test.

Two-sample *t*-test - Homogeneity of Variance

- You can default to the Welch two-sample *t*-test in R which does not assume Homogeneity of variance. Or...
- Check using the Levene's test:
  - H_0: The data are drawn from two populations that have EQUAL variance: $\sigma^2_{\text{Energizer}} = \sigma^2_{\text{Ultracell}}$
  - H_A: The data are drawn from two populations that have UNEQUAL variance: $\sigma^2_{\text{Energizer}} \neq \sigma^2_{\text{Ultracell}}$
- Look at the *p*-value produced by the Levene's test
  - Assume equal variance if you *Fail to reject* H_0, $p > .05$ (Assumption not violated)
  - Otherwise, do not assume equal variance, $p < .05$ (Assumption violated)
- Assumption violated: Use Welch two-sample t-test in R - `var.equal=FALSE`
- Assumption not violated: Use the standard two-sample t-test in R - `var.equal=TRUE`

```
> leveneTest(Pulses ~ Brand, data = Battery_sub)

Levene's Test for Homogeneity of Variance (center = median)

      Df F value  Pr(>F)

group  1  3.7606 0.07032 .

      16

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

R reports `var.equal=FALSE` by default.

p > 0.05. The Levene's test tells us it is safe to assume homogeneity of variance...

Two-sample *t*-test - R

```
> t.test(~Pulses | Brand, data = Battery_sub)

        Welch Two Sample t-test

data:  Pulses by Brand
t = -10.699, df = 13.399, p-value = 6.13e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -46.18347 -30.70542
sample estimates:
mean in group Energizer mean in group Ultracell
               118.2222                156.6667
```

I think it is always better to NOT assume equal variance.

The p-value is really small, *p* < .001

The 95% CI of the difference between means does not capture H_0: $u_{Energizer}$- $u_{Ultracell}$= 0

Two-sample *t*-test - Interpretation

Two-sample *t*-test result summary:

- We assumed normality, but there might be some doubt.
- We defaulted to not assuming equal variance, despite the Levene's test indicating it was safe to assume.
- Estimated difference between means: 118.22 - 156.67 = -38.45 pulses (Energizer - Ultracell)
- 95% *CI* of difference between means [-46.18, -30.71]
- *p*-value < .001

Decision:

- *Reject* H_0

What do we conclude?

The results of the study found a statistically significant mean difference between Energizer and Ultracell pulse counts, t(df = 13.40) = -10.7, p < .001, difference between means = -38.45 pulses, 95% CI [-46.18, -30.71]*. Ultracell batteries performed significantly better on average than the more expensive Energiser batteries.*

Test statistics and its degrees of freedom for two sample test

The test Statistics “t” for two sample test is defined by)(for equal variances)

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

Where

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

Where t follow a T-distribution with the following degrees of freedom

Df= n1 + n2 -2

Test statistics and its degrees of freedom for two sample test

The test Statistics “t” for two sample test is defined by)(for unequal variances)

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

Where t follow a T-distribution with the following degrees of freedom

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

Two sample test using R for equal and not equal variances

- If you don't specify var.equal in the t.test() function for R, the two-sample t-test not assuming equal variance is reported by default. This test is also known as the Welch two-sample t-test.

```
> t.test(Body_temp ~ Gender, data = Body_temp,
        var.equal=FALSE,
        alternative="two.sided")
Welch Two Sample t-test

data:  Body_temp by Gender
t = -2.32, df = 128, p-value = 0.022
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.302145     -0.024009
sample estimates:
  mean in group Male mean in group Female

              36.726            36.889
```

Two sample test using R for equal and not equal variances

The 95% CI of the difference between the means (- 0.163) was calculated using the following formula in R:

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2, 1-\frac{\alpha}{2}}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2, 1-\frac{\alpha}{2}}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}\right]$$

Paired Samples to test the differences between two different procedures, conditions, environments

Paired samples refer to situation where the measurements are taken on the same source, subjects or objects under two different procedures, conditions, environments to decide whether the two conditions are the same or not.

Therefore, the samples are dependent, e.g, one group of passengers using two different airlines to test the quality of the airlines, one group of engineers using two different procedures to identify the faults before take off to decide which procedure is more effective in identifying the major faults in a timely manner, the measure of life quality by moving the same group of people to live in two different cities for the same period of time to see if the two cities have the same impact on being happy.

Steps to carry out hypothesis testing for paired samples

1- for each subject we have two measurements (one under each condition), we find the differences for all subjects in the sample, i.e., we get n differences.

2- We test the mean of differences u_{Δ}, i.e., we are dealing with population of differences only so it is a one sample test.

3- If the two conditions are the same then mean of differences is zero. i.e.,

Paired-samples *t*-test- Overview

- **Hypotheses for the paired (dependent) samples *t*-test**:

$$H_0\text{: } u_\Delta = 0$$

$$H_A\text{: } u_\Delta \neq 0$$

- **Assumptions**:
    - Comparing the population average difference or change, u_Δ, between two matched measurements, $d_i = x_{i2} - x_{i1}$.
    - Differences, Δ are normally distributed or large sample used ($n > 30$)
- **Decision Rules**:
    - Reject H_0:
        - If *p*-value < 0.05 (α significance level)
        - If 95% *CI* of the mean difference does not capture H_0: $u_\Delta = 0$
    - Otherwise, fail to reject H_0.
- **Conclusion**:
    - Test will be statistically significant if we reject H_0
    - Otherwise, the test is not statistically significant.

Test statistics and Confidence interval for paired samples

Test statistics following T- distribution with df = n-1 is:

$$t = \frac{\bar{d}}{\frac{s_\Delta}{\sqrt{n_\Delta}}}$$

The (1-α) % CI is :

$$\left[\bar{d} - t_{n_\Delta,1-\frac{\alpha}{2}}\frac{s_\Delta}{\sqrt{n_\Delta}}, \bar{d} + t_{n_\Delta,1-\frac{\alpha}{2}}\frac{s_\Delta}{\sqrt{n_\Delta}}\right]$$

Graphical interpretation of paired samples

- **Dependent Sample Assessment Plots Using granova and R**

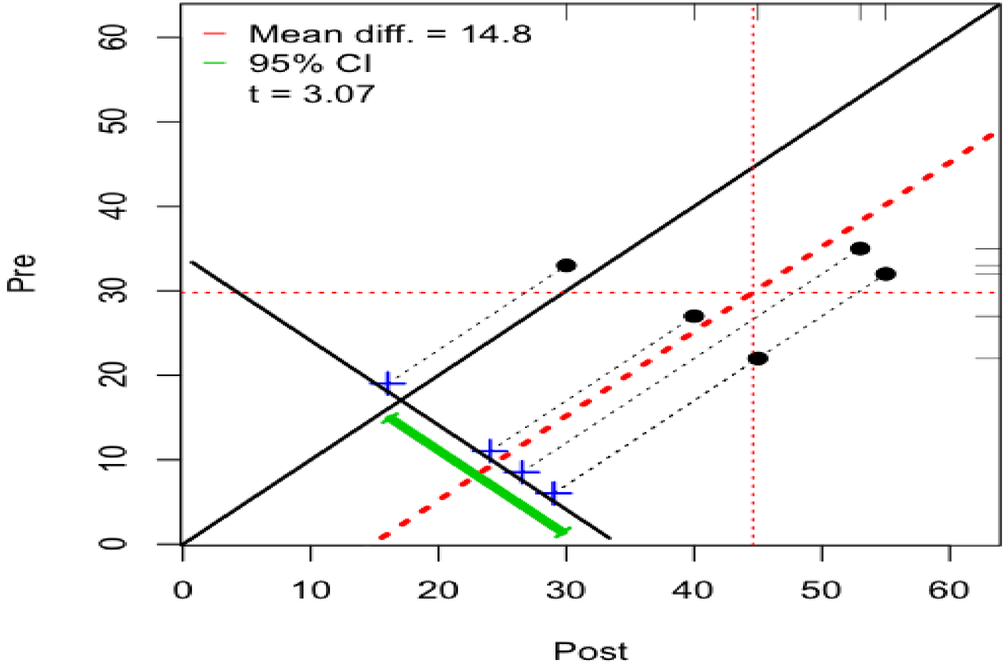
Think of pre-intervention and post-intervention response data scores, when studying the effects of intervention. For example Suppose you’re an educator and you administer an assessment to students at the beginning of a unit asking about their level of confidence or understanding of a topic. You then teach a lesson that spans some period of time. At the end you collect responses to the same questions again. You now have a dependent sample: two responses that related to the same individual for some number of individuals.

Graphical interpretation of paired samples

	Pre	*Post*
Adam	22	45
Beth	33	30
Cindy	35	53
David	32	55
Elisabeth	27	40

For such a small sample, you can quickly eye-ball the raw data and see that there seems to have been an upward shift in scores, but is it significant (in the statistical sense)? Could you so easily eye-ball the results for a class of 20, 30, or 100 students? Probably not.

Graphical interpretation of paired samples



Graphical interpretation of paired samples

- The plot has several features worth mentioning:
- The x-axis and y-axis use the same range (they're on the same scale), so the plot is square.
- We have plotted post-assessment scores along the x-axis so that the mean difference will be positive for increases in post-scores and negative for decreases in post-scores.
- The solid black line running from the lower-left to the upper-right represents x and y values that are the same (10, 10), (20, 20), and so on; this is called the identity line. Therefore, if there was no change between the pre- and post-assessment, we would expect the points to appear along this 45 degree line.
- Any points below this line represents a positive change (scores increased from the pre- to post-assessment).
- Any points above this line represents a negative change (scores decreased from the pre- to post-assessment).
- The horizontal, thinly-dashed red line represents the pre-assessment mean; here, about 29.
- The vertical, thinly-dashed red line represents the post-assessment mean; here, about 44.
- The thick, dashed red line running diagonally is the mean of the difference between pre- and post-assessment scores (the difference mean); here, 14.8, i.e., post-assessment scores were 14.8 points higher than pre-assessment scores, on average.
- The green bar indicates the 95% confidence interval: the range of values for the population mean difference that are reasonable, in light of these data.
- If the green bar overlaps the identity line, then any observed difference is not statistically significant.
- Conversely, if the green bar does not overlap the identity line, then any observed difference is statistically significant. (It's up to the analyst to decide whether it's of practical significance!)

Paired-samples *t*-test - Example

- Does reaction time improve with practice?
- We will test this claim by measuring your average reaction times twice to determine if you improve on your second try.
  1. Measure your average RT (out of five tries) **twice** using the following online test - http://www.humanbenchmark.com/tests/reactiontime
  2. Upload your results to the Google form (no trolling!) - http://goo.gl/forms/FY8vr5Fsb6 (login required)
  3. When instructed, download results from the Data Repository - `Reaction Time Practice.csv`
- Import the data into RStudio and name the data object `Reaction.Time.Practice`

Reaction Time
Seconds
(1 second = 1.00)
START
STOP
TO DO
• Press start
• When the tone sounds press stop as quick as you can
• Try again.
This is one test you can improve with practice, as your brain and muscles learn the activity. Reaction time is the interval between perception of a stimulus and the beginning of a response to it. During this interval the brain receives and processes a message from your eyes and ears, then sends a signal along the nerves telling the muscles to respond. The speed of this transmission and processing of information determines your reaction time.

Paired-samples *t*-test - Differences

Example…

RT First x_{i1}	RT Second x_{i2}	$d = x_{i2} - x_{i1}$
285	271	-14
210	232	22
278	224	-54
Average	$\bar{d} = \frac{\sum x_{i2} - x_{i1}}{n}$	-15.33

- We will finish up this example for our first class exercise...