Hypothesis Testing

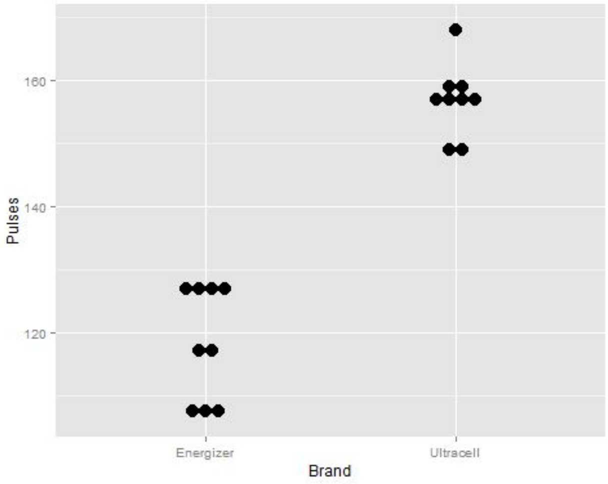
A Demonstration of the Two-sample and paired samples t-test

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

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 - Overview

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

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

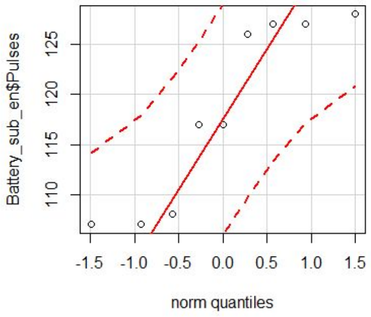
$$H_A: u_{Energizer} - u_{Ultracell} \neq 0$$

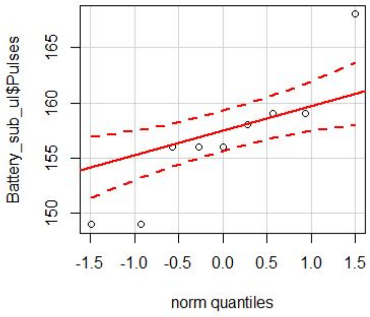
- **Assumptions**:
  - Comparing two independent population means with unknown population variance.
  - Population data are normally distributed or large sample used ($n > 30$ for both groups)
  - Population homogeneity of variance
- **Decision Rules**:
  - Reject H_0:
    - If p-value < 0.05 (α significance 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.

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

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: $\sigma^2_{Energizer} = \sigma^2_{Ultracell}$
  - H_A: The data are drawn from two populations that have UNEQUAL variance: $\sigma^2_{Energizer} \neq \sigma^2_{Ultracell}$
- 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...

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

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.*

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`

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

Paired-samples *t*-test- Overview

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

$$H_0: u_\Delta = 0$$
$$H_A: u_\Delta \neq 0$$

- **Assumptions**:
  - Comparing the population average difference or change, u_Δ, between two matched measurements, $d_i = x_{i2} - x_{i1}$.
  - 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.

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...