SM-1402 Basic Statistics

Chapter 6: Comparing groups *[handout version]*

Dr. Haziq Jamil
Mathematical Sciences, Faculty of Science, UBD
https://haziqj.ml

Semester II 2022/23

Learning outcomes

- Perform the paired t-test to compare group means in a before/after situation.
- Perform the two sample t-test (unequal variances) to compare means of two groups.
- Perform the ANOVA test to compare means of two or more groups.
- Understand the assumptions and limitations of the t-tests/ANOVA, and know when to apply each depending on the situation.

Required reading

- Madsen (2016) Chapter 8.

Introduction

In data analysis, there will be a time where you would like to compare two sets of data. Several situations:

1. In a planned experiment, you would like to see whether there is a *treatment effect* as compared to a control group.
2. Also in a planned experiment, you would like to see whether there is a "before/after" effect.
3. In a sample survey, you would like to see whether two or more demographic groups have similar traits.

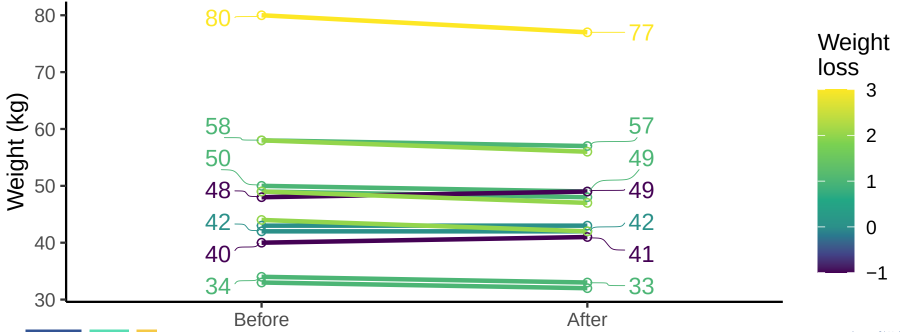
Depending on the situation at hand, there is a specific kind of statistical test to use:

i. Paired t-test
ii. Two-sample t-test (unequal variances)
iii. Analysis of variance

Before vs after effect

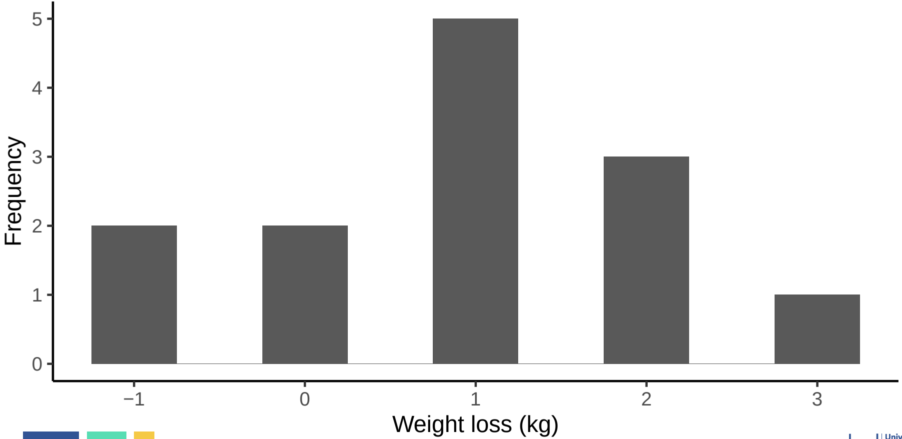
Example 1

The girls in the Fitness Club survey are selected at random ($n = 13$), and told to exercise at least 1 hour daily over a period of 4 weeks. (No other lifestyle changes are asked to be made.) Their weight before and after the exercise program were made. Can we conclude that there is a weight loss among the girls?



Before vs after effect (cont.)

Before	42	58	58	40	49	80	50	48	49	34	33	43	44
After	42	57	56	41	48	77	49	49	47	33	32	43	42
Difference	0	1	2	-1	1	3	1	-1	2	1	1	0	2



Paired t-test

If we computed the mean of the difference, we would get

$$\bar{d} = \frac{0+1+2-1+1+3+1-1+2+1+1+0+2}{13} = \frac{12}{13} = 0.92$$

There seems to be an average weight loss of 0.92 kg, but is this significantly different from zero?

Proposition 2 (Paired t-test)

Assume that H_0 is true, that *the mean difference is 0*. Then

$$T = \frac{\bar{d}}{s/\sqrt{n}}$$

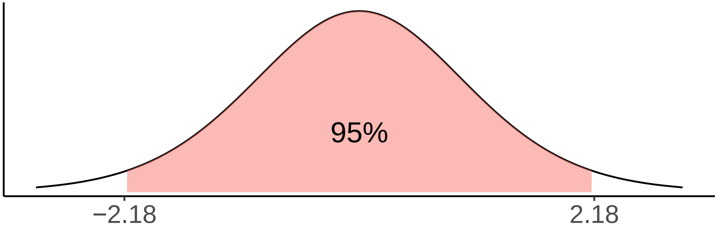
follows a t_{n-1} distribution, where s is the standard deviation of the differences.

Paired t-test (cont.)

For our current example, we calculate s to be $s = 1.19$. Hence,

$$T = \frac{0.92}{1.19/\sqrt{13-1}} = 2.68.$$

For context, the top and bottom 2.5% percentiles of the t_{12} distribution are

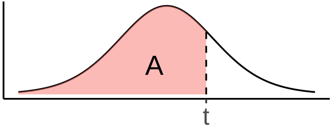


Since $2.68 > 2.18$, we reject the null hypothesis (at the 95% significance level) and conclude that there is a statistically significant difference in the before and after weights.

Why 'top and bottom' tails?

- If all differences are 0, then $\bar{d} = 0$ and $T = 0$. Values that are close to 0 supports our null hypothesis.
- Values of T that are very far away from 0 is evidence against the null hypothesis.
    - Large positive values of d makes T far on the right side of 0.
    - Large negative values of d also makes T far from 0, but on the left side.
    - Hence both 'top and bottom' are considered "bad" for our hypothesis.
- Testing the top and bottom tails is called a *two-sided test*. While one-sided tests exist, they are less stringent so two-sided tests are more commonly used.
- Other important fractiles of the t_{12} are as follows:

Sig. level	0.10	0.05	0.01	0.001
A	0.95	0.975	0.995	0.9995
t	1.782	2.179	3.055	4.318



A confidence interval for the difference

Proposition 3

A 95% confidence interval for the differences is given by

$$\bar{d} \pm t_{n-1}(2.5\%) \times s/\sqrt{n},$$

where n is the sample size and s is the SD of the differences.

Continuing our example, we can compute the 95% confidence interval for the differences as

$$0.92 \pm 2.18 \times 1.19/\sqrt{12} = 0.92 \pm 0.748 = [0.17, 1.67]$$

So a lower estimate for the weight loss is 0.17 kg, while an upper estimate for the weight loss is 1.67 kg. Notice how 0 is not included in this interval. This is another way of determining whether or not the weight loss was significant.

Comparing two group means

Example 4

We wish to examine differences in physical fitness for boys and girls in the Fitness Club survey. The body mass index (BMI) can be calculated from the sample, and used as a measure of fitness. Are there differences in BMI between boys and girls?

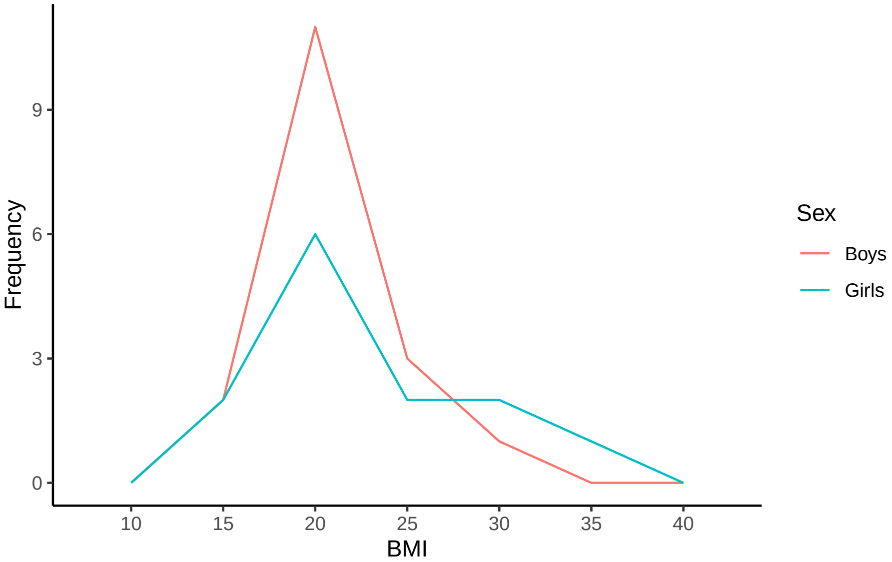
Boys BMI

16.33	16.96	17.78	19.16	19.43	19.64	19.80	19.88	20.57
21.22	21.56	21.93	22.15	23.64	26.78	27.22	28.72	

Girls BMI

15.23	15.82	17.78	17.98	19.48	21.12	21.41
21.49	22.98	23.67	28.06	30.38	33.48	

Frequency polygons



Summary statistics

	Boys ($j = 1$)	Girls ($j = 2$)
Mean ($\bar{x}_j$)	21.34	22.22
SD (s_j)	3.52	5.53
Sample size (n_j)	17	13
df ($n_j - 1$)	16	12

We have two groups of data values. We are interested in whether there is a difference between the **mean of the two groups** (and if there is, we want to estimate the mean difference).

Remark

Comparing two groups can refer to

- Two different groups of individuals for comparison's sake.
- Two groups of subjects subjected to different "treatments".

Two sample t-test

Proposition 5 (Two sample t-test)

Assume that H_0 is true, that *the mean difference between the two groups is 0*. Then

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

follows a t distribution.

The degrees of freedom for the test statistic follows a rather complicated formula:

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

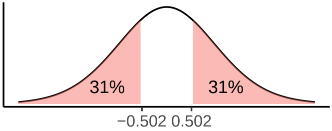
Back to example

For the example at hand, we calculate the test statistic

$$T = \frac{21.34 - 22.22}{\sqrt{\frac{3.52^2}{17} + \frac{5.53^2}{13}}} = -0.502$$

If we plug in all the numbers to the degrees of freedom formula, we get $k = 19.2$ (rounded off to $k = 19$). The relevant quantiles of the t_{19} distribution are tabulated below:

Sig. level	0.10	0.05	0.01	0.001
A	0.95	0.975	0.995	0.9995
t	1.729	2.093	2.861	3.883



We then conclude that $\Pr(|T| > 0.502) \not< 0.05$, so we have insufficient evidence to reject our hypothesis.

Confidence interval

In the same way, we can construct a 95% confidence interval for the difference in the means.

Proposition 6

A 95% confidence interval for the differences in means of the two groups is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n-1}(2.5\%) \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Thus for our current example, the 95% interval is

$$(21.34 - 22.22) \pm 2.093 \times \sqrt{\frac{3.52^2}{17} + \frac{5.53^2}{13}} = [-4.56, 2.79]$$

In this case, 0 <u>is</u> included in the interval.

Remarks

- All statistical tests rely on the assumption that the data are normally distributed. So check this before you perform the test; otherwise the results are invalid.
- For the two sample t-test, it is best to assume that the variances in each group are different.
    - There is in fact a t-test which assumes equality of variances (and an accompanying variance F-test aka Levene's test), but in practice, this *pooled-variance* version is rarely useful.
    - Theory suggests that the unequal variance version (aka Welch test) is more powerful.
- While there is no requirement for $n_1 = n_2$, the more unequal the sample sizes are, the less powerful the test becomes.
- The larger the sample size, the better!

Quick and dirty way to determine sample size

Suppose that we make some extreme assumptions:

i. Use the normal quantile instead of t (so we use $1.96 \approx 2$)
ii. Both variances are equal to σ^2
iii. Both sample sizes are equal to n

Then the 95% interval becomes

$$(\bar{x}_1 - \bar{x}_2) \pm 2 \times \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = (\bar{x}_1 - \bar{x}_2) \pm \overbrace{2\sqrt{\frac{2\sigma^2}{n}}}^{u}.$$

The quantity u is known as the statistical uncertainty. If we fix this quantity, then we can invert the formula to get n:

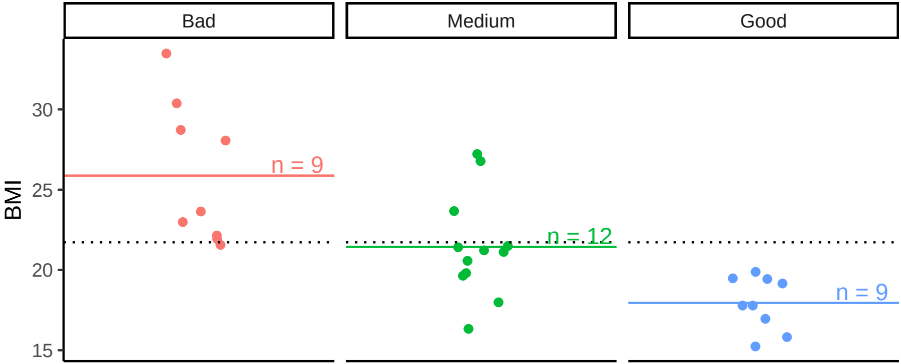
$$u = 2\sqrt{\frac{2\sigma^2}{n}} \;\Rightarrow\; n = 8\sigma^2/u^2$$

So this gives an easy way to plan *how many samples should be in each group*.

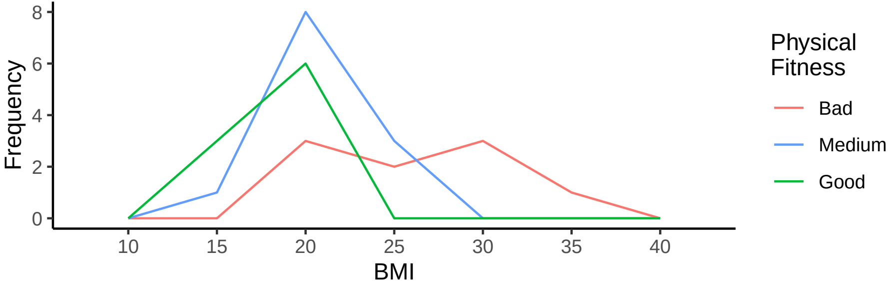
Comparing means of more than two groups

Example 7

In the Fitness Club survey, respondents indicated whether they had "Good", "Medium" or "Bad" physical fitness. Is there any association between these responses and BMI?



Exploratory analysis



	Overall	Bad ($j = 1$)	Medium ($j = 2$)	Good ($j = 3$)
Mean ($\bar{x}_j$)	21.72	25.88	21.44	17.95
SD (s_j)	4.44	4.37	3.19	1.68
Sample size (n_j)	30	9	12	9

ANOVA

The analysis of variance (ANOVA), despite its name, is a method to compare group means two or more groups[1].

Proposition 8

Assume that H_0 is true, that *all group means are equal (to the grand mean $\bar{x}$)*. Compute the ANOVA Sum of Squares (SS) table below.

Source	SS	df	MSS	F-statistic
Between	$\sum_j n_j(\bar{x}_j - \bar{x})^2$	$m-1$	$\frac{\sum_j n_j(\bar{x}_j-\bar{x})^2}{m-1}$	$\frac{\sum_j n_j(\bar{x}_j-\bar{x})^2/(m-1)}{\sum_{i,j}(x_{ij}-\bar{x}_j)^2/(n-m)}$
Within	$\sum_{i,j}(x_{ij} - \bar{x}_j)^2$	$n-m$	$\frac{\sum_{i,j}(x_{ij}-\bar{x}_j)^2}{n-m}$	
Total	$\sum_{i,j}(x_{ij} - \bar{x})^2$	$n-1$		

Then the test statistic is distributed according to an $F_{m-1,n-m}$ distribution.

[1] If comparing two groups, it's identical to the two-sample t-test with **equal variance**.

ANOVA manual calculation

"Bad" BMI: 28.1, 23.0, 30.4, 33.5, 23.6, 21.9, 22.1, 21.6, 28.7
"Medium" BMI: 21.5, 21.4, 18.0, 23.7, 21.1, 26.8, 27.2, 16.3, 21.2, 20.6, 19.6, 19.8
"Good" BMI: 15.2, 17.8, 15.8, 19.5, 19.2, 17.8, 19.4, 19.9, 17.0

- Between variation:

$$BSS = 9(25.9 - 21.7)^2 + 12(21.4 - 21.7)^2 + 9(18.0 - 21.7)^2 = 284.695$$

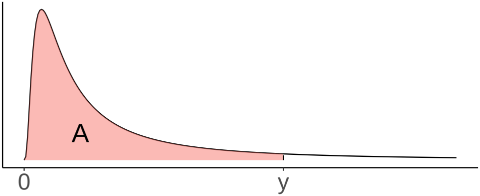
- Within variation
  - "Bad": $(28.1 - 25.9)^2 + (23.0 - 25.9)^2 + \cdots + (28.7 - 25.9)^2 = 152.5$
  - "Medium": $(21.5 - 21.4)^2 + (21.4 - 21.4)^2 + \cdots + (19.8 - 21.4)^2 = 111.8$
  - "Good": $(15.2 - 18.0)^2 + (17.8 - 18.0)^2 + \cdots + (17.0 - 18.0)^2 = 22.7$
  - "Bad" + "Medium" + "Good" = 286.934
- F statistic:

$$F = \frac{BSS/(m-1)}{WSS/(n-m)} = \frac{284.695/(3-1)}{286.934/(30-3)} = 13.3947$$

- p-value: $\Pr(F_{2,27} > 13.3947) \ll 0.05$. [Hence, reject H_0]

The F-distribution

Each table entry is y, where $\int_{-\infty}^{y} f(x)\,dx = A$ with $X \sim F_{k_1,k_2}$.



	k_1							
A	1	2	3	4	5	6	7	8
$k_2 = 27$								
0.95	4.210	3.354	2.960	2.728	2.572	2.459	2.373	2.305
0.975	5.633	4.242	3.647	3.307	3.083	2.923	2.802	2.707
0.995	9.342	6.489	5.361	4.740	4.340	4.059	3.850	3.687
0.9995	15.633	10.206	8.163	7.064	6.369	5.887	5.530	5.255

The F-distribution (cont.)

- The F tables can be a bit tricky to read, since there are 2 degrees of freedom. Familiarise yourselves with the table.
- If you are unable to find the exact degrees of freedom, you may approximate or round off, e.g. $F_{2,27} \approx F_{2,30}$.
- Another useful result is the following: If $X \sim F_{k_1,k_2}$ and $\Pr(X < x) = A$, then $\Pr(Y < 1/x) = 1 - A$ where $Y \sim F_{k_2,k_1}$

Example 9

- The top 2.5% quantile of $F_{5,10}$ is 4.236, i.e. $\Pr(X < 4.236) = 0.975$.
- Then, the *bottom* 2.5% quantile of $F_{10,5}$ is $1/4.236 = 0.236$.

(You can't actually check it from the tables, so try this in Excel)

Final remarks

- The ANOVA assumes that data (within the groups)
  i. are normally distributed
  ii. have equal variances
- Again, sample sizes need not be equal (balanced), but it is best if it can be equal.
- Other (non-parametric) tests not discussed here:
  1. Wilcoxon signed-rank test (a test of medians) instead of one-sample or paired t-test.
  2. Mann-Whitney U-test (a test of medians) instead of two-sample t-test.
  3. Kruskal-Wallis test instead of one-way ANOVA.