big question: we have several groups of observations, do these groups differ in terms of some outcome variable of interest?

As an example, let's say we have a dataset of results for a clinical trial. 18 participants are trying different combinations of 2 kinds of therapy (9 in CBT, 9 none) and 3 kinds of drugs (6 joyzepam, 6 anxifree, 6 placebo). **How can we tell whether an improvement in mood is significant or just a random coincidence?**

ANOVA hypotheses

$$H_0 : \text{it is true that } \mu_p = \mu_A = \mu_J$$

$$H_1 : \text{it is NOT true that } \mu_p = \mu_A = \mu_J$$

Some formulas

Sample variance of Y

$$\text{Var}(Y) = \frac{1}{N} \sum_{k=1}^{G} \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

Where:

- N = number of samples (18)
- G = number of groups (3 - one for each drug)
- N_k = number of people in kth group (6 people each)

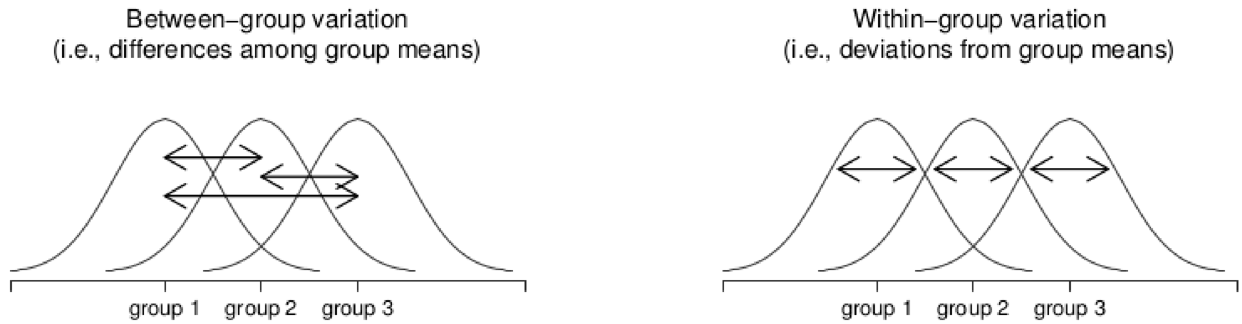
Now look at sum of squares, just calculate variance but don't divide by N

$$\text{SS}_{\text{tot}} = \sum_{k=1}^{G} \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y})^2$$

Now we break up the sum of squares into two different kinds of variation:

- **between-group SS** = differences between the means of two classes

- **wthin-group SS** = differences from group means within each group



(a) (b)

$$\text{within group SS} = \text{SS}_w = \sum_{k=1}^{G} \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$$

In other words, we're isolating one group, then calculating the squared difference between each Y_i and the mean of that group $\bar{Y}_k$, and adding that up for all groups

$$\text{between group SS} = SS_b = \sum_{k=1}^{G} N_k (\bar{Y}_k - \bar{Y})^2$$

Here, take the mean of every group ($\bar{Y}_k$)/and compare it against the mean of the whole dataset ($\bar{Y}$)
and now:

$$\text{SS}_w + \text{SS}_b = \text{SS}_{tot}$$

What does this mean? *The variability associated with the outcome variable can be split into two parts: variation due to sample means for different groups (SS_b) and the rest of the variation (SS_w)*

What does this mean for ANOVA? If the means were the same (that is, the null hypothesis is True), then all the sample means would be small and SS_b would be very small. If the alternative hypothesis is True, then the between-groups differences would be larger (i.e. most of the variation in Y can be explained by taking separate groups) and the within-group SS would be smaller.

The F-test

to compare SS_w to SS_b, we need to run an F-test

Degrees of freedom

$$\text{df}_b = G - 1$$
$$\text{df}_w = N - G$$

Remember: N is the number of samples (18 for the sample dataset desribed above), G is the number of groups (3 drugs being tested)

Convert sum of squares to a mean squares:

$$\mathrm{MS}_b = \frac{\mathrm{SS}_b}{\mathrm{df}_b}$$
$$\mathrm{MS}_w = \frac{\mathrm{SS}_w}{\mathrm{df}_w}$$

Now our F-ratio is:

$$F = \frac{\mathrm{MS}_b}{\mathrm{MS}_w}$$

Summary

	df	sum of squares	mean squares	F-statistic	p-value
between groups	$\mathrm{df}_b = G - 1$	$\mathrm{SS}_b = \sum_{k=1}^{G} N_k (\bar{Y}_k - \bar{Y})^2$	$\mathrm{MS}_b = \frac{\mathrm{SS}_b}{\mathrm{df}_b}$	$F = \frac{\mathrm{MS}_b}{\mathrm{MS}_w}$	[complicated]
within groups	$\mathrm{df}_w = N - G$	$\mathrm{SS}_w = \sum_{k=1}^{G} \sum_{i=1}^{N_k} (Y_{ik} - \bar{Y}_k)^2$	$\mathrm{MS}_w = \frac{\mathrm{SS}_w}{\mathrm{df}_w}$	-	-

Assumptions

- Residuals are normally distributed (use QQ plots or Shapiro-Wilk test)
- Homogeneity of variance / homoscedasicity (every group has the same standard deviation)
- independence (knowing one residual tells you nothing about any other residual)

References

- learning statistics with R chapter 14 (see both the textbook and the assocated summary)