Design and analysis of experiments
Lecture 3

Jakob G. Rasmussen
Department of Mathematics
Aalborg University
Denmark

Type I and II errors

- When accepting or rejecting H_0, there are two types of errors:

	Accept H_0	Reject H_0
H_0 true	OK	Type I error
H_0 false	Type II error	OK

- α (the significance level) is the probability of making a type I error - we choose this value.
- β is the probability of making a type II error - if we had known the true parameter (we don't!), we could calculate this.
- $1 - \beta$ is the power of the test, i.e. how good is a test at rejecting a false hypothesis.
- The smaller we choose α, the larger β will be, so both error probabilities cannot be made smaller for a given dataset and model.

Paired t-test

- The paired t-test is a variant of the t-test for equal means between two samples where we assume observations y_{1i} and y_{2i} are paired.
- Observations: $y_{11}, \ldots, y_{1n}$ and $y_{21}, \ldots, y_{2n}$ (note: same n)
- Assumptions:
  - Independence between differences $y_{1i} - y_{2i}$
  - Normally distributed data:

$$y_{i1} \sim N(\mu_1 + \beta_i, \sigma^2), \quad y_{i2} \sim N(\mu_2 + \beta_i, \sigma^2)$$

  (μ is effect from group, β is effect from individual)

Hypothesis and test statistic

- Calculating differences: $d_i = y_{i1} - y_{i2}$
- Mean:

$$\mu_d = \mathbb{E}[d_i] = \mathbb{E}[y_{i1} - y_{i2}] = (\mu_1 - \beta_i) + (\mu_2 - \beta_i) = \mu_1 - \mu_2$$

- Hypothesis:

$$H_0 : \mu_d = 0 \quad (\text{i.e.}\mu_1 = \mu_2)$$
$$H_1 : \mu_d \neq 0$$

- Test statistic:

$$t_0 = \frac{\bar{d}}{s_d/\sqrt{n}} \sim t_{n-1}$$

where $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ and $s_d^2 = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$

Paired or unpaired design?

- Advantage of pairing: The effect of individuals cancel out in the paired experiment - this means a smaller variance
- Disadvantage: the sample size has been halved, since in the unpaired case we have $n_1 + n_2 = 2n$ measurements while in the paired case we have n differences.
- The choice between two-sample t-test and paired t-test is usually decided by practical reasons, not statistical:
  - Paired example: we could test the effect of a drug by measuring the same patients before and after administering the drug - this is a paired design.
  - Unpaired example: we could compare two types of medicine by administering them to two groups of patients - this is an unpaired design.
- In R:
  - Unpaired: `t.test(x1,x2)`
  - Paired: `t.test(x1,x2,paired=TRUE)` or `t.test(x1-x2)`

R

- Exercise 1

ANOVA (analysis of variance)

- When we need to compare the mean of many samples a pairwise comparison (paired or unpaired) is wrong!
- Instead we use ANOVA
- In ANOVA we have measurements from a number of groups (treatments) and want to compare the means of these groups
- Example of data at p. 67-68 Plasma etching experiment
- Despite the name analysis of variance, we are not comparing variances - we are comparing mean values!

ANOVA

- Observations: $y_{i1}, \dots, y_{in_i}$ for $i = 1, \dots, a$
- Assumptions:
  - Independence within samples
  - Independence between samples
  - Normally distributed: $y_{ij} \sim N(\mu_i, \sigma^2)$
- Typical question: Does the mean value of one or more groups differ from the others?

The model - one formulation

- Means model:

$$y_{ij} = \mu_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2),$$
$$i = 1, \ldots, a, \quad j = 1, \ldots, n_i, \quad N = \sum_{i=1}^{a} n_i$$

- Hypothesis:

$$H_0 : \mu_1 = \cdots = \mu_a$$
$$H_1 : \mu_i \text{ not all equal}$$

The model - another formulation

- Effects model:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}, \qquad \epsilon_{ij} \sim N(0, \sigma^2)$$

- Constraints:
  - Type 1: $\bar{\tau} = \sum_i (n_i \tau_i)/N = 0$
  - Type 2: $\tau_1 = 0$ (reference group/placebo - used in R)
- Hypothesis:

$$H_0 : \tau_1 = \cdots = \tau_a = 0$$
$$H_1 : \tau_i \text{ not all } 0$$

A bit of calculations...

- Data and averages:
  - Data: $y_{ij} = \mu + \tau_i + \epsilon_{ij}$
  - Group means: $\bar{y}_{i\bullet} = \mu + \tau_i + \bar{\epsilon}_{i\bullet}$
  - Overall mean: $\bar{y}_{\bullet\bullet} = \mu + \bar{\epsilon}_{\bullet\bullet}$
- We will calculate sums of squares (SS) to obtain a test statistic for testing whether all groups have the same mean.
- Error sum of squares (variation within groups):

$$SS_E = \sum_i \sum_j (y_{ij} - \bar{y}_{i\bullet})^2 = \sum_i \sum_j (\epsilon_{ij} - \bar{\epsilon}_{i\bullet})^2 \sim \sigma^2 \chi^2_{\nu_E},$$

  where d.f. is given by $\nu_E = N - a$
- Treatment sum of squares (variation between groups):

$$SS_A = \sum_i \sum_j (\bar{y}_{i\bullet} - \bar{y}_{\bullet\bullet})^2 = \sum_i \sum_j (\tau_i + \bar{\epsilon}_{i\bullet} - \bar{\epsilon}_{\bullet\bullet})^2 \sim \sigma^2 \chi^2_{\nu_A},$$

  where d.f. is given by $\nu_A = a - 1$

Test

- Test statistic:

$$F_0 = \frac{MS_A}{MS_E} = \frac{SS_A/\nu_A}{SS_E/\nu_E} \sim \frac{\sigma^2\chi^2_A/\nu_A}{\sigma^2\chi^2_E/\nu_E} = F_{\nu_A,\nu_E}$$

- If $F_0 < F_{\nu_A,\nu_E;\alpha}$ we accept H_0 - note the one-sided test, if F_0 is very small we do not reject H_0 since this just means that the group means are very much like each other.
- In practice acceptance of H_0 means that the treatments we are comparing are not visibly different, and none of them are superior.
- Rejection means that there are significant differences, but not necessarily between all pairs of groups - we will get back to this in the next lecture.

The ANOVA table

- It is standard to represent an analysis of variance as a table:

Source of variation	SS	df	MS	F	p
Treatment	SS_A	ν_A	MS_A	F_0	p-value
Error	SS_E	ν_E	MS_E		
Total	SS_{Tot}	ν_{Tot}			

- Here $SS_{Tot} = SS_A + SS_E$ and $\nu_{Tot} = \nu_A + \nu_E$

Test for equal variances

- We have assumed that the variances are equal, so we need a test for checking this, when we have two or more groups - Bartlett's test is one such test.
- Hypothesis:

$$H_0 : \sigma_1^2 = \cdots = \sigma_a^2$$
$$H_1 : \sigma_i^2 \text{ not all equal}$$

- Such a test should be conducted before an ANOVA to check the assumption of equal variances.
- It can also be used on its own to check whether various methods for doing something have the same variances - for example, when comparing a couple of machines for producing some equipment, a relevant question is often whether some machines are more precise than others.

Test for equal variances

- Test statistic:

$$\chi_0^2 = \frac{(N-a)\ln(s_p^2) - \sum_i (n_i - 1)\ln(s_i^2)}{1 + \frac{1}{3(a-1)}\left(\sum_i (n_i-1)^{-1} - (N-a)^{-1}\right)}$$

  where $s_p^2 = \frac{1}{N-a}\sum_i (n_i - 1)s_i^2$
- Under H_0 this is approximately χ_{a-1}^2-distributed, and we reject H_0 if $\chi_0^2 > \chi_{a-1;\alpha}^2$
- Note that this test is sensitive to non-normality.

R

- R-demo, part 2
- Exercise 2

An alternative to ANOVA - Kruskal-Wallis test

- If the normality assumption fails completely, ANOVA is inappropriate.
- We can then use a non-parametric test (i.e. it does not assume normality).
- Note that this tests the hypothesis that the groups have equal medians, not means.
- We do the following:
  - Sort all the data y_{ij}
  - Let R_{ij} be the rank of y_{ij} (take average in case of ties)
  - Let $R_{i\bullet} = R_{i1} + \cdots + R_{in_i}$
- For example:

$$y_{11} = 10, y_{12} = 12, y_{21} = 12, y_{22} = 14$$
$$R_{11} = 1, R_{12} = 2.5, R_{21} = 2.5, R_{22} = 4$$
$$R_{1\bullet} = 3.5, R_{2\bullet} = 6.5$$

Kruskal-Wallis test

- Test statistic:

$$H = \frac{1}{S^2}\left(\sum_{i=1}^{a}\frac{R_{i\bullet}^2}{n_i} - \frac{N(N+1)^2}{4}\right)$$

  where $S^2 = \frac{1}{N-1}\left(\sum_{i=1}^{a}\sum_{j=1}^{n_i} R_{ij}^2 - \frac{N(N+1)^2}{4}\right)$
- Under H_0 the test statistic H is approximately χ^2_{a-1}-distributed, and we reject H_0 if $\chi^2_0 > \chi^2_{a-1;\alpha}$
- The approximation is good if $n_i > 5$ for all i.

Non-parametric tests

- Many parametric tests have non-parametric alternatives when the normality assumptions fail.
- Examples:
  - Wald-Wolfovitz test and Mann-Whitney U test are alternatives to the independent (i.e. unpaired) two-sample t-test
  - Wilcoxon signed-rank test is an alternative to paired t-test
  - Kruskal-Wallis test is an alternative to ANOVA
- As a rule of thumb, the parametric test is superior (it has the best power) when the normality assumptions are fulfilled; otherwise the non-parametric alternative is the best.
- We do not have time to go through them, but if you have understood Kruskal-Wallis then the other ones should not be hard.
- Be aware: the non-parametric tests have their own subtleties, and do not correspond completely to the parametric tests. Be sure you know exactly what you are testing!

R

- R-demo, part 3
- Exercise 3