STAT 401A - Statistical Methods for Research Workers

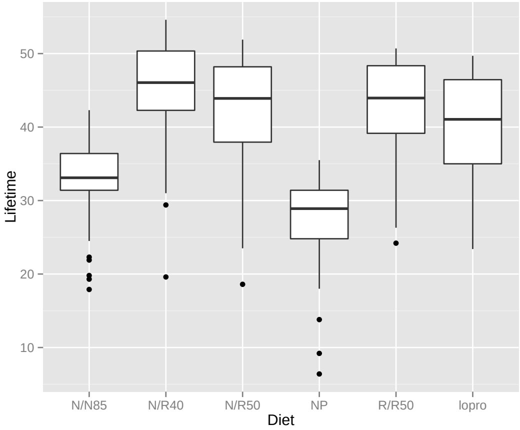
One-way ANOVA

Jarad Niemi (Dr. J)

Iowa State University

last updated: September 29, 2014

Lifetime (months) of mice on different diets



One-way ANOVA model/assumptions

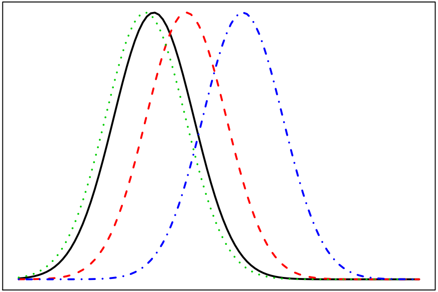
$$Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right)$$

for $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$.
(n_j means there can be different # of observations in each group)

Assumptions:

- Normality
  - Not skewed
  - Not heavy-tailed
- Common variance for all groups
- Independence
  - No cluster effects
  - No serial effects
  - No spatial effects

ANOVA assumptions graphically



What if you want to compare two groups?

We may still be interested in comparing two groups.

Statistical hypothesis: Is there a difference in mean lifetimes between the mice in two groups, e.g. NP and N/N85?

Statistical question: What is the difference in mean lifetimes between the mice in two groups, e.g. NP and N/N85?

Two-group analysis

Begin with the two group (equal variance) model:

$$Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right)$$

but now $j = 1, 2$ and $i = 1, \ldots, n_j$
To perform a hypothesis test or a CI for the difference in means, the relevant quantities are:

- $\overline{Y}_2 - \overline{Y}_1$
- $SE(\overline{Y}_2 - \overline{Y}_1) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- t distribution with $n_1 + n_2 - 2$ degrees of freedom

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n1 + n_2 - 2}$$

What if you have more than two groups?

Multi-group analysis

The multi-group (equal variance) model:

$$Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right)$$

but now $j = 1, \ldots, J$ and $i = 1, \ldots, n_j$
(n_j means there can be different # of observations in each group)
To perform a hypothesis test or a CI for the difference in means, the relevant quantities are:

- $\overline{Y}_2 - \overline{Y}_1$
- $SE(\overline{Y}_2 - \overline{Y}_1) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- t distribution with $n_1 + n_2 + \cdots + n_J - J$ degrees of freedom

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_J - 1)s_J^2}{n1 + n_2 + \cdots + n_J - J}$$

Hypothesis test for comparison of two means (in multi-group data)

If $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ for $j = 1, \ldots, J$ and we want to test the hypothesis

- $H_0 : \mu_1 = \mu_2$
- $H_1 : \mu_1 \neq \mu_2$

then we compute:

$$t = \frac{\overline{Y}_1 - \overline{Y}_2}{SE(\overline{Y}_1 - \overline{Y}_2)}$$

where

$$SE(\overline{Y}_1 - \overline{Y}_2) = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_J - 1)s_J^2}{n_1 + n_2 + \cdots + n_J - J}.$$

Then we compare t to a t distribution with $n_1 + n_2 + \cdots + n_J - J$ degrees of freedom.

Diet effect on mice lifetime

Table : Summary statistics for mice lifetime (months) on different diets

	Diet	n	mean	sd
1	N/N85	57	32.7	5.1
2	N/R40	60	45.1	6.7
3	N/R50	71	42.3	7.8
4	NP	49	27.4	6.1
5	R/R50	56	42.9	6.7
6	lopro	56	39.7	7.0

Test for difference in mean lifetime between NP and N/N85, i.e.

$$H_0 : \mu_4 = \mu_1 \text{ vs } H_a : \mu_4 \neq \mu_1.$$

Showing work

$$
\begin{aligned}
\overline{Y}_1 - \overline{Y}_4 &= 32.7 - 27.4 = 5.3 \\
df &= 57 + 60 + 71 + 49 + 56 + 56 - 6 = 343 \\
s_p^2 &= \frac{(57-1)5.1^2+(60-1)6.7^2+(71-1)7.8^2+(49-1)6.1^2+(56-1)6.7^2+(56-1)7.0^2}{57+60+71+49+56+56-6} \\
&= \frac{15314}{343} = 44.6 \\
s_p &= \sqrt{s_p^2} = \sqrt{44.6} = 6.7 \\
SE(\overline{Y}_1 - \overline{Y}_4) &= s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_4}} = 6.7\sqrt{\frac{1}{57} + \frac{1}{49}} = 1.3 \\
t &= \frac{\overline{Y}_1 - \overline{Y}_4}{SE(\overline{Y}_1 - \overline{Y}_4)} = \frac{5.3}{1.2} = 4.1 \\
p &= 2P(t_{343} < -|t|) = 2P(t_{343} < -4.1) = 0.000052
\end{aligned}
$$

So we reject the null hypothesis that there is no difference between mean lifetime of mice on the NP and N/N85 diets.

Confidence interval for the difference of two means (in multi-group data)

If $Y_{ij} \stackrel{ind}{\sim} N(\mu_j, \sigma^2)$ for $j = 1, \ldots, J$, a $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$\overline{Y}_1 - \overline{Y}_4 \pm t_{df}(1-\alpha/2) SE(\overline{Y}_1 - \overline{Y}_4)$$

where the t critical value, $t_{n_1+n_2+\cdots+n_J-J}(1-\alpha/2)$, needs to be calculated using a statistical software.

A 95% confidence interval for the difference in mean lifetime for N/N85 minus NP $(\mu_1 - \mu_4)$ is

$$5.3 \pm 1.96 \times 1.3 = (2.8, 7.8).$$

The statistical conclusion would be

> *In this study, mice on the N/N85 diet lived an average of 5.3 months longer than mice on the NP diet (95% CI (2.8,7.8)).*

One-way ANOVA F-test

Are any of the means different?

Hypotheses in English:

H_0: all the means are the same

H_a: at least one of the means is different

Statistical hypotheses:

$$\begin{array}{lll} H_0: & \mu_j = \mu \text{ for all } i & \qquad Y_{ij} \stackrel{iid}{\sim} N(\mu, \sigma^2) \\ H_a: & \mu_j \neq \mu_{j'} \text{ for some } j \text{ and } j' & \qquad Y_{ij} \stackrel{ind}{\sim} N\left(\mu_j, \sigma^2\right) \end{array}$$

An ANOVA table organizes the relevant quantities for this test and computes the pvalue.

ANOVA table

A start of an ANOVA table:

Source of variation	Sum of squares	d.f.	Mean square
Factor A (Between groups)	$SSA = \sum_{j=1}^{J} n_j \left(\overline{Y}_j - \overline{Y}\right)^2$	$J-1$	$\frac{SSA}{J-1}$
Error (Within groups)	$SSE = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}_j\right)^2$	$n-J$	$\frac{SSE}{n-J}$ $(= s_p^2)$
Total	$SST = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left(Y_{ij} - \overline{Y}\right)^2$	$n-1$	

where

- J is the number of groups,
- n_j is the number of observations in group j,
- $n = \sum_{j=1}^{J} n_j$ (total observations),
- $\overline{Y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} Y_{ij}$ (average in group j),
- and $\overline{Y} = \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n_j} Y_{ij}$ (overall average).

ANOVA table

An easier to remember ANOVA table:

Source of variation	Sum of squares	df	Mean square	F-statistic	p-value
Factor A (between groups)	SSA	$J-1$	MSA $=$ SSA$/J-1$	MSA/MSE	(see below)
Error (within groups)	SSE	$n-J$	MSE $=$ SSE$/n-J$		
Total	SST=SSA+SSE	$n-1$			

Under H_0,

- the quantity MSA/MSE has an F-distribution with $J-1$ numerator and $n-J$ denominator degrees of freedom,
- larger values of MSA/MSE indicate evidence against H_0, and
- the p-value is determined by $P(F_{J-1,n-J} > MSA/MSE)$.

F-distribution