One-way ANOVA F-test

STAT 401 - Statistical Methods for Research Workers

Jarad Niemi

Iowa State University

19 September 2013

Two-sample assumptions

Previously, we studied a number of two-sample tests of location, e.g. two means (medians) are the same:

	Samples	
	Independent	Paired
Parametric	t-test, Welch's t-test (unequal variances)	paired t-test
Non-parametric	rank sum test	sign test, signed-rank test

Except permutation test, all made assumptions about the distributions:

- non-parametric tests assume shifted distributions and
- parametric tests assumed normality.

t-tests assume $Y_{ij} \stackrel{ind}{\sim} N(\mu_i, \sigma_i^2)$ for individual j of group i with possibly additional structure:

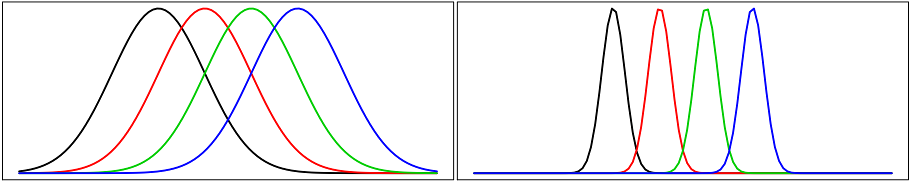
- two-sample t-test: $\sigma_i = \sigma$ for all i

t-tests are robust to departures from these assumptions

ANOVA assumptions

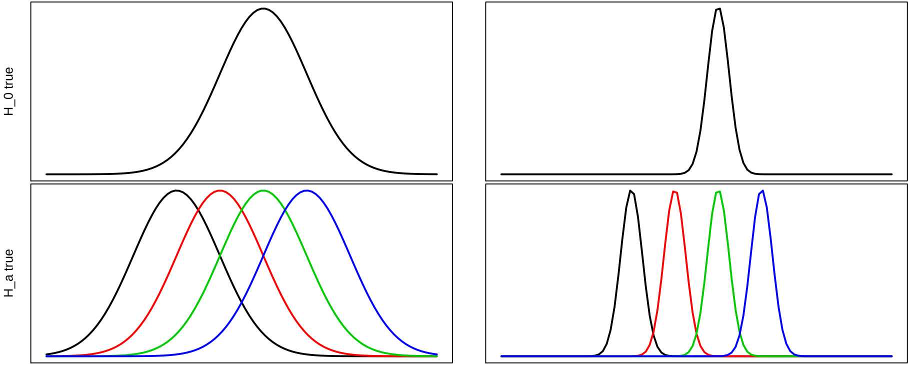
$$Y_{ij} \stackrel{ind}{\sim} N(\mu_i, \sigma^2)$$

$j = 1, \ldots, n_i$ and $i = 1, \ldots, \mathrm{I}$
(n_i means there can be different # of observations in each group)



One-way ANOVA F-test

$$H_0: \quad \mu_i = \mu \text{ for all } i \qquad\qquad Y_{ij} \stackrel{iid}{\sim} N(\mu, \sigma^2)$$
$$H_a: \quad \mu_i \neq \mu_{i'} \text{ for some } i \text{ and } i' \qquad\qquad Y_{ij} \stackrel{ind}{\sim} N(\mu_i, \sigma^2)$$



Full vs reduced model

Model	*Full*	*Reduced*
Assumption	$H_a : Y_{ij} \stackrel{ind}{\sim} N(\mu_i, \sigma^2)$	$H_0 : Y_{ij} \stackrel{iid}{\sim} N(\mu, \sigma^2)$
Mean	$\hat{\mu}_i = \overline{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$	$\hat{\mu} = \overline{Y} = \frac{1}{n} \sum_{i=1}^{I} \sum_{j=1}^{n_i} Y_{ij}$
Residual	$r_{ij} = Y_{ij} - \overline{Y}_i$	$r_{ij} = Y_{ij} - \overline{Y}$
Sum of squares	$SSE = \sum_{i=1}^{\mathrm{I}} \sum_{j=1}^{n_i} r_{ij}^2$	$SST = \sum_{i=1}^{\mathrm{I}} \sum_{j=1}^{n_i} r_{ij}^2$

ANOVA table

A start of an ANOVA table:

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

where

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

ANOVA table

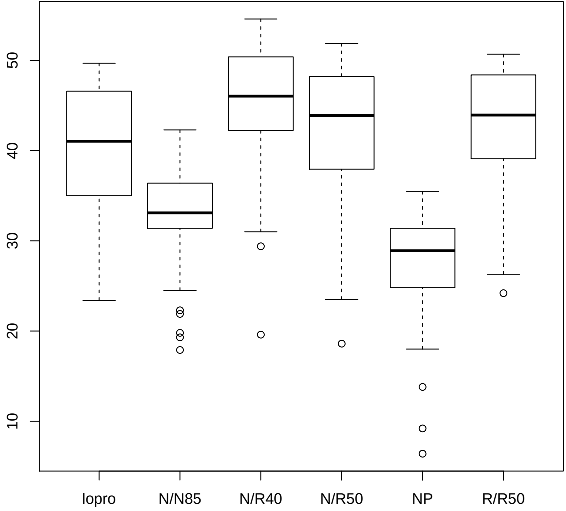
An easier to remember ANOVA table:

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

Under H_0,

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

Lifetimes of mice on different diets



Diet	Count
lopro	56
N/N85	57
N/R40	60
N/R50	71
NP	49
R/R50	56

SAS code and output for one-way ANOVA

```
DATA mice;
  INFILE 'case0501.csv' DSD FIRSTOBS=2;
  INPUT lifetime diet $;

PROC GLM DATA=mice;
  CLASS diet;
  MODEL lifetime = diet;
  RUN;
```

```
                                            The GLM Procedure

Dependent Variable: lifetime

                                                  Sum of
       Source                       DF           Squares     Mean Square    F Value    Pr > F
       Model                         5       12733.94181      2546.78836      57.10    <.0001
       Error                       343       15297.41532        44.59888
       Corrected Total             348       28031.35713
```

Summary

One-way ANOVA F-test model:

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

The test:

$$\begin{aligned} H_0 &: \quad \mu_i = \mu \text{ for all } i \\ H_a &: \quad \mu_i \neq \mu_{i'} \text{ for some } i \text{ and } i' \end{aligned}$$

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