Statistics

Week 11: Single Factor Experiments (Chapters 12)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Information

Tuesday second half: last guest lecture.

Thursday 1pm: homework 3 due. You can submit on *eDimension*, or in the homework box labeled 'Statistics' on level 7, building 1.

Outline

Introduction

Independent samples design allows us to compare two groups. Now we look at techniques for comparing *more than two* groups.

More formally, we look at an experiment which measures a response from more than two groups (or treatments). The treatments are levels of a single treatment factor.

The available experimental units are randomly assigned to each treatment (no matching).

Example: we might want to measure the compression during a crash for small, medium, and large cars.

Set up

Group (or treatment)			
1	2	$\cdots$	k
y_{11}	y_{21}	$\cdots$	y_{k1}
y_{12}	y_{22}	$\cdots$	y_{k2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_{1n_1}	y_{2n_2}	$\cdots$	y_{kn_k}

The group sizes n_i do *not* necessarily equal.

Total sample size: $N = \sum_{i=1}^{k} n_i$.

Sample mean for group i: $\bar{y}_i$.

Sample standard deviation for group i: s_i.

Grand mean: $\bar{\bar{y}} = \frac{1}{N} \sum_{i,j} y_{ij}$. Note: this is a double sum.

Set up, continued

We assume that for each group, the response is *normally* distributed, and that all the groups have the same variance σ^2 but not necessarily the same mean μ_i.

Exercise

(1) Show that

$$\sum_{i=1}^{k} n_i \left(\bar{y}_i - \bar{\bar{y}}\right) = 0.$$

(2) If all the group sizes are the same, give an interpretation of $\bar{\bar{y}}$.

Confidence interval

Since each s_i^2 is an estimator of σ^2, we can pool them together to get a better estimate for σ^2:

$$s^2 := \frac{\sum_{i,j}(y_{ij} - \bar{y}_i)^2}{N - k} = \frac{\sum_i (n_i - 1) s_i^2}{\sum_i (n_i - 1)}.$$

Using s, we can write down the $(1-\alpha)$-level confidence interval for μ_i (the true mean of group i):

$$\bar{y}_i - t_{N-k,\,1-\alpha/2} \frac{s}{\sqrt{n_i}} \le \mu_i \le \bar{y}_i + t_{N-k,\,1-\alpha/2} \frac{s}{\sqrt{n_i}}.$$

The null hypothesis

Our primary interest is in comparing whether the μ_i's are actually different. Set up $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$.

A preliminary test can be carried out using side-by-side box plots.

Can we use confidence intervals to test for H_0? It turns out that even if all the CI's contain a number in common, it is not obvious how strongly this supports H_0.

The tabular set up of the data values y_{ij} suggests that we can use a tool encountered before: *analysis of variance*.

The idea behind ANOVA

The idea is to compare the variation *between* the groups to the variation *within* each group.

The total variance can (once again) be decomposed into the above two terms.

SST $:= \sum_{i,j}(y_{ij} - \bar{\bar{y}})^2$, df $= N - 1$ (total),

SSE $:= \sum_{i,j}(y_{ij} - \bar{y}_i)^2$, df. $= N - k$ (within),

SSA $:= \sum_{i=1}^{k} n_i\,(\bar{y}_i - \bar{\bar{y}})^2$, df $= k - 1$ (between).

The ANOVA identity

SSA is the weighted sum of squared errors between all treatments, and can also be written as $\sum_{i,j}(\bar{y}_i - \bar{\bar{y}})^2$. Its degree of freedom is $(k-1)$ due to the relation in the previous exercise.

A 'large enough' value of SSA would indicate that H_0 is false.

Let **MSA** $=$ SSA$/(k-1)$, **MSE** $=$ SSE$/(N-k) = s^2$, and $F =$ MSA/MSE.

The ANOVA identity

SST = SSA + SSE.

The set up here is just like regression (in fact, it is because the data can be written as a regression model).

Proof of the ANOVA identity

$$
\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{\bar{y}})^2 &= \sum_{i,j}(y_{ij}-\bar{y}_i+\bar{y}_i-\bar{\bar{y}})^2 \\
&= \sum_{i,j}(\bar{y}_i-\bar{\bar{y}})^2+\sum_{i,j}(y_{ij}-\bar{y}_i)^2+2\sum_{i}(\bar{y}_i-\bar{\bar{y}})\sum_{j}(y_{ij}-\bar{y}_i) \\
&= \sum_{i} n_i(\bar{y}_i-\bar{\bar{y}})^2+\sum_{i,j}(y_{ij}-\bar{y}_i)^2+0.
\end{aligned}
$$

As in regression, F satisfies a $F_{k-1,\,N-k}$ distribution if H_0 is true.

We can reject H_0 with $(1-\alpha)$ confidence if $F > f_{k-1,\,N-k,\,1-\alpha}$.

Exercises

Use $\alpha = 0.05$ throughout.

(1) Complete all the calculations using the formulas, in the spreadsheet '*cars*'.

Check your answers against *Excel*'s `Anova: Single Factor` function.

(2) Complete all the calculations in the spreadsheet '*sugar*'; think about how to find SSE and SST.

(3) Construct the ANOVA tables in the spreadsheet '*anorexia*', and answer the question.