Lecture 7: ANOVA

Sandy Eckel
seckel@jhsph.edu

30 April 2008

Statistical methods for comparing multiple groups

Continuous data: comparing multiple means

- Analysis of variance

Binary data: comparing multiple proportions

- Chi-square tests for $r \times 2$ tables
  - Independence
  - Goodness of Fit
  - Homogeneity

Categorical data: comparing multiple sets of categorical responses

- Similar chi-square tests for $r \times c$ tables

ANOVA: Definition

- Statistical technique for comparing means for multiple (usually ≥ 3) independent populations
  - To compare the means in 2 groups, just use the methods we learned to conduct a hypothesis test for the equality of two population means (called a t-test)
- Partition the total variation in a response variable into
  - Variability within groups
  - Variability between groups
- ANOVA = **AN**alysis **O**f **VA**riance

ANOVA: Concepts

- Estimate group means
- Assess magnitude of variation attributable to specific sources
- Partition the variation according to source
- Extension of 2-sample t-test to multiple groups
- Population model
- Sample model: sample estimates, standard errors (sd of sampling distribution)

Types of ANOVA

One-way ANOVA

- **One factor — e.g. smoking status (never, former, current)**

Two-way ANOVA

- Two factors — e.g. gender and smoking status

Three-way ANOVA

- Three factors — e.g. gender, smoking and beer consumption

Emphasis

One-way ANOVA is an extension of the t-test to 3 or more samples

- **focus analysis on group differences**

Two-way ANOVA (and higher) focuses on the interaction of factors

- Does the effect due to one factor change as the level of another factor changes?

Recall: sample estimator of variance

Sample variance:

$$s^2 = \frac{\sum_i^n (X_i - \bar{X})^2}{n-1}$$

The distance from any data point to the mean is the **deviation** from this point to the mean:

$$(X_i - \bar{X})$$

The **sum of squares** is the sum of all squared deviations:

$$SS = \sum_i^n (X_i - \bar{X})^2$$

ANOVA partitioning of the variance I

Variation in all observations = **Variation between each observation and its group mean** + **Variation between each group mean and the overall mean**

In other words,

Total sum of squares = **Within group sum of squares** + **Between groups sum of squares**

In shorthand:

SST = SSW + SSB

ANOVA partitioning of the variance II

SST This is the sum of the squared deviations between each observation and the overall mean:

$$SST = \sum_{i}^{n} (X_i - \bar{X})^2$$

SSW This is the sum of the squared deviations between each observation the mean of the group to which it belongs:

$$SSW = \sum_{i}^{n} (X_i - \bar{X}_{group(i)})^2$$

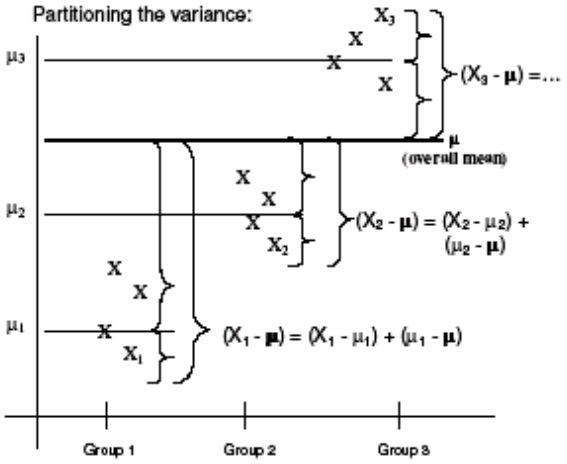
ANOVA partitioning of the variance III

SSB This is the sum of the squared deviations between each group mean and the overall mean, weighted by the sample size of each group (n_{group}):

$$SSB = \sum_{group} n_{group} (\bar{X}_{group} - \bar{X})^2$$

If the group means are not very different, the variation between them and the overall mean (SSB) will not be much more than the variation between the observations within a group (SSW)

One-way ANOVA: the picture



Within groups mean square: MSW

- We assume the variance σ^2 is the same for each of the group's populations
- We can pool (combine) the estimates of σ^2 across groups and use an overall estimate for the common population variance:

$$\begin{aligned} \text{Variation within a group} &= \hat{\sigma}^2_W \\ &= \frac{SSW}{N-k} = \text{MSW} \end{aligned}$$

- MSW is called the "within groups mean square"

Between groups mean square: MSB

- We can also look at systematic variation among groups

$$\begin{aligned} \text{Variation between groups} &= \hat{\sigma}_B^2 \\ &= \frac{SSB}{k-1} = \text{MSB} \end{aligned}$$

An ANOVA table

- Suppose there are k groups (e.g. if smoking status has categories current, former or never, then k=3)
- We calculate a test statistic for a hypothesis test using the sum of square values as follows:

ANOVA Table

Source of Variation	Sum of Squares	Df	Mean Square	F
Between Groups	SSB	k - 1	$\frac{SSB}{k-1}$ = MSB	$\frac{MSB}{MSW}$
Within Groups	SSW	N - k	$\frac{SSW}{N-k}$ = MSW	
Total	SST	N - 1		

Hypothesis testing with ANOVA

- In ANOVA we ask:
  is there truly a difference in means across groups?
- Formally, we can specify the hypotheses:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_k \\ H_a &: \text{at least one of the } \mu_i\text{'s is different} \end{aligned}$$

- The null hypothesis specifies a global relationship between the means
- Get your test statistic (more next slide...)
- If the result of the test is significant (p-value $\leq \alpha$),
  then perform individual comparisons between pairs of groups

A test statistic

- **Goal:** Compare the two sources of variability: MSW and MSB
- Our test statistic is

$$F_{obs} = \frac{\text{MSB}}{\text{MSW}} = \frac{\hat{\sigma}_B^2}{\hat{\sigma}_W^2} = \frac{\text{variance between the groups}}{\text{variance within the groups}}$$

- If F_{obs} is small (close to 1), then variability between groups is negligible compared to variation within groups
  $\Rightarrow$ The grouping does not explain much variation in the data
- If F_{obs} is large, then variability between groups is large compared to variation within groups
  $\Rightarrow$ The grouping explains a lot of the variation in the data

The F-statistic

- For our observations, we assume $X \sim N(\mu_{gp}, \sigma^2)$, where μ_{gp} is the (possibly different) mean for each group's population.
  - Note: under H_0, we assume all the group means are the same
- We have assumed the same variance σ^2 for all groups — important to check this assumption
- Under these assumptions, we know the null distribution of the statistic $F = \frac{MSB}{MSW}$
- The distribution is called an F-distribution

The F-distribution

- Remember that a χ^2 distribution is always specified by its degrees of freedom
- An F-distribution is any distribution obtained by taking the quotient of two χ^2 distributions divided by their respective degrees of freedom
- When we specify an F-distribution, we must state two parameters, which correspond to the degrees of freedom for the two χ^2 distributions
- If $X_1 \sim \chi^2_{df_1}$ and $X_2 \sim \chi^2_{df_2}$ we write:

$$\frac{X_1/df_1}{X_2/df_2} \sim F_{df_1, df_2}$$

Back to the hypothesis test . . .

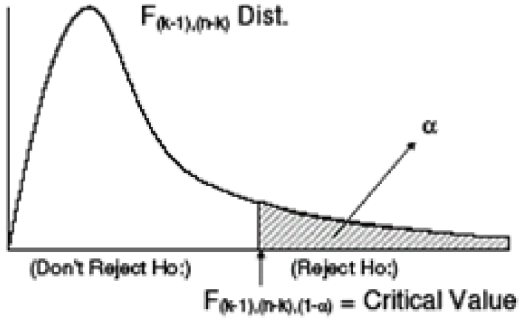
Knowing the null distribution of $\frac{\text{MSB}}{\text{MSW}}$, we can define a decision rule for the ANOVA test statistic (F_{obs}):

- Reject H_0 if $F_{obs} \geq F_{\alpha;k-1,N-k}$
- Fail to reject H_0 if $F_{obs} < F_{\alpha;k-1,N-k}$

- $F_{\alpha;k-1,N-k}$ is the value on the $F_{k-1,N-k}$ distribution that, used as a cutoff, gives an area in the uppper tail $= \alpha$
- We are using a one-sided rejection region

ANOVA: F-tests I

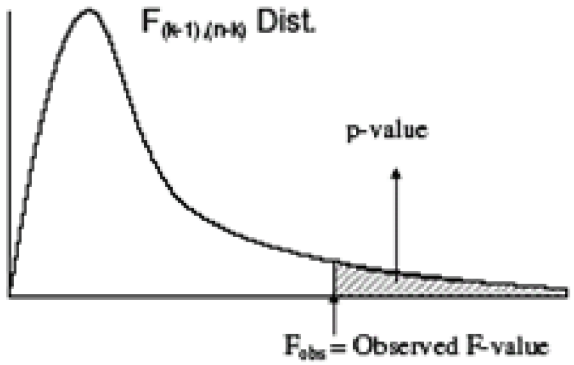
How to find the value of $F_{\alpha;k-1,N-k}$



Use the `R` function `qf(alpha, df1, df2, lower.tail = F)`

ANOVA: F-tests II

How to find the p-value of an ANOVA test statistic



Use the R function `pf(Fobs, df1, df2, lower.tail = F)`

Example: ANOVA for HDL

Study design: Randomized controlled trial

- 132 men randomized to one of
  - Diet + exericse
  - Diet
  - Control

Follow-up one year later:

- 119 men remaining in study

Outcome: mean change in plasma levels of HDL cholesterol from baseline to one-year follow-up in the three groups

Model for HDL outcomes

We model the means for each group as follows:

$$\begin{aligned} \mu_c = E(HDL|gp = c) &= \text{mean change in control group} \\ \mu_d = E(HDL|gp = d) &= \text{mean change in diet group} \\ \mu_{de} = E(HDL|gp = de) &= \text{mean change in diet and exercise group} \end{aligned}$$

We could also write the model as

$$E(HDL|gp) = \beta_0 + \beta_1 I(gp = d) + \beta_2 I(gp = de)$$

Recall that I(gp=D), I(gp=DE) are 0/1 group indicators

HDL ANOVA Table

We obtain the following results from the HDL experiment:

Source of Variation	Sum of Squares	Df	Mean Square	F
Between Groups	0.728	2	0.364	13.0
Within Groups	3.236	116	0.028	
Total	3.964	118		

HDL ANOVA results & conclusions

F-test

- $H_0 : \mu_c = \mu_d = \mu_{de}$ (or $H_0 : \beta_1 = \beta_2 = 0$)
- H_a : at least one mean is different from the others

Test statistic

- $F_{obs} = 13.0$
- $df_1 = k - 1 = 3 - 1 = 2$
- $df_2 = N - k = 116$
- Rejection region: $F > F_{0.05;2,116} = 3.07$

Conclusions

- Since $F_{obs} = 13.0 > 3.07$, we reject H_0
- We conclude that at least one of the group means is different from the others (p$<$ 0.05)

Which groups are different?

- We might proceed to make individual comparisons
- Conduct two-sample t-tests to test for a difference in means for each pair of groups (assuming equal variance):

$$t_{obs} = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

- Recall: s_p^2 is the 'pooled' sample estimate of the common variance
- df $= n_1 + n_2 - 2$

Multiple Comparisons

- Performing individual comparisons require multiple hypothesis tests
- If $\alpha = 0.05$ for each comparison, there is a 5% chance that each comparison will falsely be called significant
- Overall, the probability of Type I error is elevated above 5%
- For n independent tests, the probability of making a Type I error at least once is $1 - 0.95^n$
  - Example: For $n = 10$ tests, the probability of at least one Type I error is 0.40 !
- Question: How can we address this multiple comparisons issue?

Bonferroni adjustment

- A possible correction for multiple comparisons
- Test each hypothesis at level $\alpha^* = (\alpha/3) = 0.0167$
- Adjustment ensures overall Type I error rate does not exceed $\alpha = 0.05$
- However, this adjustment may be too conservative

Multiple comparisons α

Hypothesis	$\alpha^* = \alpha/3$
$H_0 : \mu_c = \mu_d$ (or $\beta_1 = 0$)	0.0167
$H_0 : \mu_c = \mu_{de}$ (or $\beta_2 = 0$)	0.0167
$H_0 : \mu_d = \mu_{de}$ (or $\beta_1 - \beta_2 = 0$)	0.0167
	Overall $\alpha = 0.05$

HDL: Pairwise comparisons I

Control and Diet groups

- $H_0 : \mu_c = \mu_d$ (or $\beta_1 = 0$)
- $t = \frac{-0.05-0.02}{\sqrt{\frac{0.028}{40}+\frac{0.028}{40}}} = -1.87$
- p-value $= 0.06$

HDL: Pairwise comparisons II

Control and Diet + exercise groups

- $H_0 : \mu_c = \mu_{de}$ (or $\beta_2 = 0$)
- $t = \frac{-0.05-0.14}{\sqrt{\frac{0.028}{40}+\frac{0.028}{39}}} = 5.05$
- p-value $= 4.4 \times 10^{-7}$

HDL: Pairwise comparisons III

Diet and Diet + exercise groups

- $H_0 : \mu_d = \mu_{de}$ (or $\beta_1 - \beta_2 = 0$)
- $t = \frac{-0.02-0.14}{\sqrt{\frac{0.028}{40}+\frac{0.028}{39}}} = -3.19$
- p-value $= 0.0014$

Bonferroni corrected p-values

Hypothesis	p-value	adjusted α
$H_0 : \mu_c = \mu_d$	0.06	0.0167
$H_0 : \mu_c = \mu_{de}$	4.4×10^{-7}	0.0167
$H_0 : \mu_d = \mu_{de}$	0.0014	0.0167
	Overall $\alpha = 0.05$	

The p-value must be less than the adjusted α to reject H_0

Conclusion: Significant difference in HDL change for DE group compared to other groups

Summary of Lecture 7

- Sample variance, sums of squares
- ANOVA
  - partition of the sums of squares
  - ANOVA table
  - hypothesis test setup
  - test statistic
- F distribution
- Multiple hypothesis testing, Bonferroni correction
- ANOVA "dull hypothesis" joke: http://www.phdcomics.com/comics/archive.php?comicid=905