

Handed out: Wednesday October 14, 2009

Due: Wednesday October 21, 2009

Reading: Oct. 14 - Oct. 16 ANOVA: multiple comparisons, diagnostics (Chap 16-18)
 Oct. 19 Randomized block ANOVA (Chap 21)
 Oct. 21 - Oct. 28 Correlation/Simple regression (Chap 1-5)

1. **Basics of ANOVA** - We return to the SENIC data (of HW 2) to illustrate some ANOVA basics. The hospitals (which you will recall are a random sample from the U.S.) are divided into regions of the country. The average infection rate and the standard deviation of the infection rates from each region as well as the number of hospitals in each region are provided in the table below:

region	n	mean	s.d.
north-east	28	4.87	1.27
north-central	32	4.39	1.34
south	37	3.93	1.46
west	16	4.38	0.88

- (a) Identify the “population means” that we would compare in an analysis of variance based on the data above. In other words, identify the populations being compared and the relevant characteristic of these populations.
- (b) Obtain the ANOVA table for comparing the four population means using the information provided.
- (c) Is there evidence that the four types of hospitals differ in their mean infection rate? Explain.
- (d) In looking at the table of means it seems natural to compare the north-east (highest infection rate) and south (lowest infection rate). A two-sample t-test finds the difference in means for these two regions to be statistically significant with $p < .01$ (two-sided). Explain why the p-value obtained from this test could be considered misleading.
2. **ANOVA and contrasts** – A survey in Ohio asked students in grades 3-8 how often they watch television (self-reported as hours per day) and how often they engage in violent behavior (a composite score between 0 and 15 based on self-reported responses to a number of items). Data for 5 groups of children determined by the self-reported hours of television are provided in the table below.

tv hours per day	n	mean	s.d.
< 1	227	2.94	2.73
1 – 2	526	2.59	2.44
3 – 4	666	2.87	2.36
5 – 6	310	3.10	2.45
> 6	488	4.03	2.81

- (a) Explain why the analysis of variance assumption of normality does not seem plausible here and then explain why the violation of this assumption is not a major concern. (Hint: Recall the response variable is nonnegative.)
- (b) The ANOVA F-test rejects the null hypothesis of equal means in the five groups ($p < .001$). The investigators wonder whether there is trend with more television being associated with more violence. This can be tested with a contrast with weights $(-4, -2, 0, 2, 4)$, i.e., by analyzing the population contrast $\gamma = -4\mu_{<1} - 2\mu_{1-2} + 0\mu_{3-4} + 2\mu_{5-6} + 4\mu_{>6}$.
- Explain why these contrast weights are appropriate for the theory being tested.
 - Suppose that the null hypothesis that the contrast equals zero is rejected, then what would that tell us about the group means?
 - Suppose that the null hypothesis that the contrast equals zero is not rejected, then what would that tell us about the group means?
 - Carry out a significance test of the hypothesis that the contrast of means is equal to zero. State the result of your test.
- (c) An alternative hypothesis is that those who watch less than five hours are different than those that watch five or more hours. (But under this hypothesis we don't expect any difference between say 1 hour watchers or 3 hour watchers or between 5 hour watchers and 6 hour watchers.) Identify a set of contrast weights that could be used to assess this hypothesis.
- (d) Does this study establish that watching a lot of television **causes** violent behavior in children? Explain why or why not. If it does not, then what is another explanation for the observed results?

3. **ANOVA in SAS including paired comparisons and diagnostics** – We analyze the prostate cancer data from the textbook (see appendix D). A medical center urology group is studying the association between prostate-specific antigen (PSA) levels and a number of clinical measurements on 97 men with cancer about to undergo prostatectomies as treatment. We focus here on the association of PSA levels (this is a non-invasive blood test result) and Gleason scores (cancer severity score ranging from 6 to 8 with higher scores indicating worse prognosis). Note that PSA levels and Gleason scores are higher for this group than they would be for a random sample of men because all of these men have advanced cancers. The data are available as prostate.txt, prostate.xls, prostate.csv on the course website. There are 9 variables for each man:

id = identification number

PSA = prostate-specific antigen level (mg/ml)

cancvol = estimate of prostate cancer volume (cc)

weight = prostate weight (gm)

age = age of patient (yrs)

bph = amount of benign prostatic hyperplasia (cm²)

sem = presence/absence of seminal vesicle invasion

capspen = degree of capsular penetration (cm) gleason = grade of disease (6, 7, or 8 for these men)

- Before examining the results of an ANOVA we should check our model assumptions. This can be done in two ways: by examining the data in each group for normality and constant variance or by running an ANOVA and examining the residuals. Here it is sufficient to take the first approach. Comment on the assumptions for ANOVA.
- A logarithmic or square root transformation often helps with continuous positive measurements such as the PSA measurements. Reconsider the assumptions using log transformed data and square-root transformed data (create new variables by including LOGY=LOG(Y); and SQRTY=SQRT(Y); in the DATA step of your SAS program). Do the transformed data satisfy the model assumptions? Which transformation works best?
- For the remainder of the question use the log transformed data. Is there evidence that mean log PSA levels vary across groups? Explain.
- Identify the pairs of groups which differ significantly using the Bonferroni approach.
- Define and analyze contrasts to see if there is a:
 - linear trend of PSA vs Gleason scores
 - quadratic trend of PSA levels vs Gleason scores.
 What does it mean if **both** contrasts are significant?
- Describe your findings on these data in a paragraph.

SAS hints: Use SAS PROC UNIVARIATE and a normal probability plot for each group defined by the Gleason score to assess the assumptions in (a); a scatterplot of PSA vs Gleason score may also be useful. For other parts see the SAS sample program for ANOVA on the website.

4. Multiple comparisons - theory

- Consider an experiment with seven treatments.
 - How many pairwise comparisons would be required to compare each pair of treatments?
 - Are the pairwise comparisons independent? Explain.
 - Suppose that we carry out all pairwise comparisons using two-sample tests with $\alpha = .05$ and no adjustment for multiple comparisons. Find the expected number of false positives.
- Recall that the Bonferroni approach uses α/M as the significance level in each test (where M is the number of comparisons). Prove that the Bonferroni approach provides a set of tests whose familywise error rate is α . (Hints: Define relevant events A_i for each test and then show that $P(\bigcap_i A_i) > 1 - \alpha$. It will help to consider the complement of $\bigcap_i A_i$.)

5. **ANOVA theory** – The usual ANOVA model takes $Y_{ij} = \mu_i + \epsilon_{ij}$ with ϵ_{ij} 's independent identically distributed $N(0, \sigma^2)$ random variables (assume there are r groups and n_i observations per group).

- Prove that the F-statistic for ANOVA with $r = 2$ groups is the square of the pooled two-sample t -statistic.
- Missing data – Suppose that a study is designed with $n_i = n$ for each group and the goal is to develop a confidence interval for the population contrast $\gamma = \sum_i c_i \mu_i$. We briefly consider the impact of missing data.
 - Give the formula for a 95% confidence interval for the contrast.
 - Suppose some of the subjects drop out (randomly) before the response is measured so that we see $n_i \leq n$ individuals in group i for $i = 1, \dots, r$. Describe the effect of these missing data on your confidence interval for γ . (Is it wider or narrower? If so why?)
 - Suppose instead that the subjects drop out only if their response (Y) is too low (i.e., their treatment is not effective). Describe the effect of the missing data on your inference for γ now.