

## 1. Basics of ANOVA

- (a) There are four populations. These are the collections of hospitals in the four regions of the country. For each population we are interested in the mean infection rate of all hospitals in that region. Note we are not interested in the sample means — we already know these. We are using the sample means to estimate/compare/infer the population means.
- (b)  $SS(Tr) = \sum_i n_i (\bar{Y}_i - \bar{Y}_{..})^2 = 28*(4.87-4.357)^2 + 32*(4.39-4.357)^2 + 37*(3.93-4.357)^2 + 16*(4.38-4.357)^2 = 14.16$  and  $SS(Error) = \sum_i (n_i - 1)s_i^2 = 27 * 1.27^2 + 31 * 1.34^2 + 36 * 1.46^2 + 15 * 0.88^2 = 189.36$ . These can then be put into the ANOVA table:

Source	df	SS	MS
Groups	3	14.16	4.72
Error	109	189.36	1.74
Total	112	104.45	

- (c) The question of whether the four regions (populations) differ in their mean infection rates is a test of the hypothesis  $H_o : \mu_{ne} = \mu_{nc} = \mu_s = \mu_w$  vs the alternative that  $H_o$  is not true. This is tested by comparing  $F = 4.72/1.74 = 2.71$  to the  $F_{3,109}$  distribution. You end up with  $p = .048$  which suggests strong (but not very strong) evidence that the regions differ.
- (d) The p-value obtained from the usual two-sample t-test assumes that this is a single hypothesis generated in advance of data collection. Here we examine the data to choose one pairwise comparison (from among the six possible comparisons) that seems most likely to be significant. We need to take this selection process into account. Put another way it is not proper to compare  $(\bar{Y}_{ne} - \bar{Y}_s)/s.e.$  with the usual  $t$  distribution, instead we should recognize that we are actually evaluating  $\max_i \bar{Y}_i - \min_i \bar{Y}_i / s.e.$  and use an appropriate reference distribution.

## 2. ANOVA and contrasts

- (a) The violent behavior scores can't possibly be normally distributed for two reasons. First, it is a discrete score from 0 to 15. Second, the s.d. in each group is nearly equal to the mean but the score is nonnegative. This suggests a skewed distribution. One way to see this is to note that it is impossible for an individual score to be two or more s.d.'s below the mean here but of course this is quite possible in normal data. Fortunately the sample size in each group is quite large which means that even though the individual observations don't follow a normal distribution the sample means will be approximately normal.
- (b) Linear contrast
- The contrast weights identify the specific pattern in the means to which the contrast will be sensitive. In this case the weights suggest increases in violence as more tv is watched.
  - If the hypothesis that the contrast equals zero is rejected, then the group means are consistent with the hypothesized pattern. In this case rejecting zero means a linear pattern is present. Make sure to get this point.
  - If the hypothesis that the contrast equals zero is not rejected, then the group means are not consistent with the hypothesized pattern. In essence the sample means are orthogonal to the contrast.
  - The estimated contrast is  $\hat{\gamma} = 4 * 4.03 + 2 * 3.10 + 0 * 2.87 - 2 * 2.59 - 4 * 2.94 = 5.38$  and  $s.e.(\hat{\gamma}) = \sqrt{MSE * (16/227 + 4/526 + 0 + 4/310 + 16/488)}$  where  $MSE = 226 * 2.73^2 + 525 * 2.44^2 + 665 * 2.36^2 + 309 * 2.45^2 + 487 * 2.81^2 / 2212 = 6.4258$  and thus  $s.e.(\hat{\gamma}) = 0.89$ . The test statistic is  $5.38/0.89 = 6.04$  which when compared to  $t_{2212}$  distribution yields  $p$ -value  $< .0001$  and thus there is definitely a linear pattern with more tv associated with more violence.
- (c)  $c = (-2, -2, -2, 3, 3)$  or its equivalent form  $c = (-1/3, -1/3, -1/3, 1/2, 1/2)$  can be used to compare the average violence score in the first three groups (less than five hours) to the average violence score in the last two groups (five or more hours).
- (d) This is an observational study and thus can't establish a causal relationship. There can always be confounding factors in such a study. Perhaps non-attentive parenting causes both more tv watching and more violence.

## 3. SAS: ANOVA with paired comparisons and diagnostics

- (a) Examining the results of PROC UNIVARIATE or PROC MEANS or plotting the PSA scores against Gleason score demonstrates non-constant variance. (The latter two approaches are used in the SAS program below.) The Gleason = 8 group has much higher variance than the others. Normal probability plots show that none of the groups appear to follow a normal distribution.

- (b) The square root transformation helps with both of the identified problems but doesn't solve them. The logarithm transformation seems to produce data for which constant variance is plausible. The log data looks fairly normal as well. Some people did formal tests of normality and concluded the log transformation didn't work for all Gleason groups. Clearly though the log data is much closer to what we need for ANOVA than the untransformed data so it makes sense to proceed with the log transformed data.
- (c) Doing an ANOVA on the log PSA scores suggests that there are differences in the mean log PSA scores among the three groups. This is clear from the F-test of the hypothesis that the three groups (Gleason = 6,7,8) have the same mean log PSA score. This hypothesis is clearly rejected ( $F_{2,94} = 21.56, p < .0001$ ).
- (d) Using the multiple comparisons approach (even though there are only three paired comparisons) with the Bonferroni correction shows that the group with Gleason score 8 has significantly higher mean log PSA score than the other two groups. The other two groups are not significantly different at the .05 level but the difference there is nearly significant. Note that because the groups have unequal sample size the pairwise comparison results are given line-by-line separately for each pairwise comparison. (When the sample sizes are the same then the s.e. for the difference in means is  $\sqrt{MSE(2/n)}$  which is the same for each pairwise comparison. This is what allows for just listing the means and identifying by letter which ones are different. When the sample sizes are different in the different groups, then the s.e. for the difference in means varies across pairwise comparisons and you get the kind of output you see here.)
- (e) For the linear trend use  $c = (-1, 0, 1)$  which yields a highly significant result ( $F = 42.41, p < .0001$ ). This means we reject the null hypothesis and find in favor of a linear trend. For the quadratic trend I used  $c = (-1, 2, 1)$  which yields a marginally significant result ( $F = 3.28, p = .0734$ ). This makes us suspect non-linearity (a quadratic trend but it is not very strong evidence. If both were deemed significant then it would suggest the means differ in ways that are consistent with both the presence of linear trend and the presence of a quadratic trend. A couple of important comments here. First, several people tried other "quadratic" contrasts. Any contrast on three groups is quadratic because you can fit a quadratic to any three points! The quadratic I used above is known as the pure quadratic because it is orthogonal to the linear contrast. Second, if you want to understand what is happening ... look at the sample means. For log PSA these are 1.87 when Gleason = 6, 2.39 when Gleason = 7 and 3.62 when Gleason = 8. These increase but there seems to be some curvature (the change from 6 to 7 is less than the change from 7 to 8).
- (f) The paragraphs were better this week. The key is to try and connect to the science rather than just report on statistical methods. Here's mine: An observational study of 97 men with prostate cancer is used to analyze the association between Gleason scores (measuring cancer severity) and PSA levels (blood test results). Preliminary analysis of the data suggests that analysis of log PSA levels was more appropriate because of the wide range of such measurements. Our analysis shows significant differences among the mean log PSA scores in the three groups defined by Gleason scores ( $p < .0001$  using ANOVA F-test). Higher log PSA scores are associated with higher Gleason scores (mean log PSA = 3.62 when Gleason = 8, 2.39 when Gleason = 7 and 1.87 when Gleason = 6). These data suggest a strong relationship but we note that this sample covers only a very narrow range of Gleason scores.

#### Sample SAS Program

```
filename prost 'h:\HAL\Courses\Stat210\prostate.txt';
data psa;
    infile prost firstobs=2;
    input id psa cancvol weight age bph sem capspen gleason;
    logpsa = log(psa);
    sqrtpsa = sqrt(psa);
proc sort;
    by gleason;
proc rank normal=blom out=norm;
    var psa;
    by gleason;
    ranks nrm;
proc gplot data=norm;
    plot psa*nrm sqrtpsa*nrm logpsa*nrm;
    by gleason;
proc gplot;
    plot psa*gleason sqrtpsa*gleason logpsa*gleason;
proc means mean stddev min max n;
    var psa sqrtpsa logpsa;
    by gleason;
proc glm order=data;
    class gleason;
    model logpsa = gleason;
```

```

means gleason / lsd bon tukey scheffe;
contrast 'lin' gleason -1 0 1;
contrast 'quad' gleason -1 2 -1;
run;

```

#### 4. Multiple comparisons

- (a) Seven treatment study:
- i. With 7 treatments there are  $(7 \text{ choose } 2) = 7*6/2 = 21$  pairwise comparisons.
  - ii. The pairwise comparisons are not independent. This can be answered formally from the contrast perspective where  $(1,-1,0,0,0,0,0)$  and  $(1,0,-1,0,0,0,0)$  are not orthogonal. Or you can just note that  $\bar{Y}_1 - \bar{Y}_2$  and  $\bar{Y}_1 - \bar{Y}_3$  are not independent because their covariance would include the variance of  $\bar{Y}_1$ .
  - iii. The expected number of false positives is  $21*.05 = 1.05$ . This does not require independent comparisons.
- (b) Take  $A_i$  to be the event that we “correctly” accept the  $i$ th null hypothesis, i.e., the  $i$ th null is true and we don't reject it. Then we want to show that  $P(\bigcap_i A_i) \geq 1 - \alpha$  or equivalently that  $P((\bigcap_i A_i)^c) = 1 - P(\bigcap_i A_i) \leq \alpha$  (where  $\bigcap$  denotes the intersection of these events and  $c$  denotes the complementary event). The first step is DeMorgan's Law which tells us that  $(\bigcap_i A_i)^c = (\bigcup_i A_i^c)$  where  $\bigcup$  denotes the union of events. Then note that  $P(\bigcup_i A_i^c) \leq \sum_i P(A_i^c)$  (think about Venn diagrams where the “intersections” are double counted in the sum on the right). Putting all this together shows that  $P((\bigcap_i A_i)^c) = P(\bigcup_i A_i^c) \leq \sum_i P(A_i^c)$ . Then if we choose our threshold such that  $P(A_i^c)$  (the probability of rejecting the  $i$ th null when it is true) is less than or equal to  $\alpha/M$  we will have experimentwise error rate  $\alpha$ .

#### 5. ANOVA theory

- (a) The  $F$  statistic for two groups has  $MS(\text{Groups}) = SS(\text{Groups}) / (n_1 + n_2 - 2) = \frac{\sum_{i=1}^2 n_i (\bar{Y}_i - \bar{Y}_{..})^2}{n_1 + n_2 - 2} = \frac{(n_1 n_2^2 + n_2 n_1^2) (\bar{Y}_1 - \bar{Y}_2)^2 / (n_1 + n_2)^2}{n_1 + n_2 - 2} = \frac{n_1 n_2 (\bar{Y}_1 - \bar{Y}_2)^2 / (n_1 + n_2)}{n_1 + n_2 - 2}$  (using  $\bar{Y}_{..} = n_1 \bar{Y}_1 / (n_1 + n_2) + n_2 \bar{Y}_2 / (n_1 + n_2)$ ) and  $MS(\text{Error}) = SS(\text{Error}) / (n_1 + n_2 - 2) = s_p^2$ . Then  $F = MS(\text{Groups}) / MS(\text{Error}) = (\bar{Y}_1 - \bar{Y}_2)^2 / (s_p^2 (1/n_1 + 1/n_2)) = t^2$ .
- (b) Missing data
- i. The 95% confidence interval for  $\gamma$  is  $\hat{\gamma} \pm t_{N-r, .975} \sqrt{MSE \sum_i c_i^2 / n}$  where  $\hat{\gamma} = \sum_i c_i \bar{Y}_i$ .
  - ii. Because individuals are dropping out randomly this should not have a major effect on the means and s.d.s in each sample; this means we can assume that  $\hat{\gamma}$  and the  $MSE$  will remain about the same. The confidence interval is almost certainly wider with missing data. If you replace  $n$  by  $n_i \leq n$  in the contrast standard error then the standard error will increase in magnitude. Also, as individuals drop we have fewer d.f. on the t-percentile which makes this increase. Thus these two factors lead to a wider confidence interval.
  - iii. The most important point here is that if subjects are dropping out for a reason related to the response (e.g., low response) then we will not be able to get reliable inference for  $\gamma$  (or more generally for the population means). Suppose there is some variation in the  $\mu_i$ 's (some treatments work better than others) — well, if all people whose treatments don't work drop out then we will be comparing the effectiveness of treatments when they work!! It's quite possible we will conclude that all treatments are equally effective because of the missing data. It turns out that this is the key issue in dealing with missing data. We need to determine (or at least think about) whether the data are missing for reasons that are likely to bias our analysis!