Name: _____

1 Objective: One Way ANOVA

The question that ANOVA answering is if the variations between the means due to true differences about the populations means or just due to sampling variability. To answer this question, ANOVA calculates a parameter called F-statistics, which compares the variation among sample means (among different groups) to the variation within groups.

F-statistics = Variation among sample means / Variation within groups

Through the F-statistics we can see if the variation among sample means dominates over the variation within groups, or not. In the first case we will have strong evidence against the null hypothesis (means are all equals), while in the second case we would have little evidence against the null hypothesis.

One way anova considers a population that is subdivided into several groups, either because of different applied treatments or naturally.

The first question asked is: are the means of the various groups the same?

If they are not judged the same, the means of the selected groups are contrasted and various estimations are performed.

2 Key Words

`proc univariate, proc glm`, one way ANOVA.

3 Brief Description

In this lab we begin looking at the one way anova technique through an example which uses data appeared in the New England Journal of Medicine (NEJM), vol. 332, pp. 720-23, 1980.

The population is grouped by smoking habits for past 20 years:
NS = non smoker,
PS = passive smoker (work smoke, no home smoke),
NI = non inhaler,
LS = light smoker ($<$ 10 per day),
MS = moderate smoker (between 10 and 40 per day),
HS = heavy smoker (more than 40 per day).

The variable measured for each person:

FEF = forced mid-expiratory flow [the ability to blow air]

The data reported in the article is the sample mean and standard deviation of the FEF values for each of

the 6 smoking groups, as well as the sample size from each group.

Such data are enough to plug into formulas and do statistical calculations, but I'd like to have real data. So I've written a SAS data step to make some that should match that reported in NEJM. You can find that data step on Blackboard under in the file `lab20_1st`.

Copy the file `lab20_1st` into the SAS program editor and observe its structure: a seed is specified, as well an sample sizes for the various smoking groups. Then the data step creates samples of these sizes, using the means and standard deviations reported in the article to build appropriate values for FEF. At the same time it creates a class variable, AMT = amount, which specifies the represented group.

In this lab, we'll generate FEF data with various sample sizes and use `PROC GLM` to perform ANOVA, CONTRASTS and ESTIMATION. Think again of margin of error in CI: as sample size increases the margin of error decreases, since it involves the factor $1/\sqrt{n}$ where n is the sample size, Taking this as a cue, we'd expect it to be harder to conclude that group means differ when little information (small sample size) is available.

4 Lab Steps

[A] Use the data step to generate data sets.

[1] Create data set `smoke1` with all sample sizes = 4.

[2] Create data set `smoke2` with all sample sizes = 100.

[3] Create data set `smoke3` with all sample sizes = 200 except that for NI which is 50 (These are the sample sizes reported in the article).

[B] Use `proc means; class amt; var fef;` to find the class means of the each of the 3-data sets.

[1] What is the order of the class variables? Can you make it otherwise?

[2] Do the means and std devs appear right? Explain.

[3] Do the means all seem the same? That is: is there a difference in means of `fef` between groups?

[C] Use `proc plot` on each data set to `plot fef*amt`.

[1] What is the order of the class variables? Can you make it otherwise?

[2] Do the means all seem the same? That is: is there a difference in `fef` between groups? Do the plots overlap across groups, or not?

[D] Use histograms to compare class distributions.

```
title 'Comparing class distributions';
      PROC univariate data=smoke3 noprint;
        class amt;
        var fef;
        histogram fef / vscale=count normal(noprint);
        inset normal(mu sigma);
      run;
title;
```

[E] We'd like to use `proc univariate` to get box plots, but that demands sorting the data by `amt` so that the order of the class variables is the same as the box plots.

[1] Use `proc sort` on each of the 3 data sets to get new sets `ssmo1`, `ssmo2`, and `ssmo3`, each sorted by `amt`.

[2] Use

```
PROC univariate data = ssmo1 plot;
        var fef;
        by amt;
run;
```

on each of the 2 sorted data sets to get box plots.

[3] Sketch the box plots for `ssmo2`. Do they all appear to overlap?

[4] Now find the overall standard deviation, SD2, of `ssmo2` (without regard to class `amt`) and find a margin of error, `SD2/sqrt(100)`. *Optional*: Vertically, around each box plot mean (is it the + or the middle — ?) sketch a one s.e. margin of error (m.e.) window. These m.e. windows reflect better the effect of sample size on our guess of the group means. Do these m.e. windows overlap?

[5] Repeat [3 & 4] for data sets `ssmo1` and `ssmo3`.

[F] On each of the three data sets use `PROC glm`, see below; note the original ordering by class.

```
PROC glm data=smoke3 order=data;
class amt;
model fef = amt;
means amt / lsd;  /* lsd  performs pirwise t-test or Fishers Least Significant Difference test */
run;
```

[1] For each of the 3 data sets list the p-value of the anova F-statastic on your sketch. Recall: small p-value supports the conclusion that NOT ALL group means are the same.

[2] Look at the LSD = least significant difference output to see which means are judged the same. Sketch the output for `smoke1` and `smoke2` on previous sketch and notice what you get for `smoke3`.

[G] Now run contrasts and estimation on `smoke3` only. Follow this code:

```
PROC glm data = smoke3 order = data;
        class amt;
        model fef = amt;
        contrast 'ns␣vs␣ps' amt 1 -1 0 0 0 0;
        contrast 'ns␣vs␣ave␣inhale' amt 1 0 0 -.1 -.7 -.2;
        contrast '#␣of␣smo' amt 0 0 0 -14 -4 18;
        estimate 'ns␣vs␣ave␣inh' amt 1 0 0 -.1 -.7 -.2;
        estimate 'smo␣mn' intercept 70 amt 0 0 0 10 20 40 / divisor = 70;
        estimate 'nonsmo␣mn␣vs␣smo␣mn' amt 35 35 0 -10 -20 -40 / divisor = 70;
run;
quit;
```

[1] The order = *** option lets you select various orderings for the class variable. Try commenting it out to see what difference it makes.

[2] The contrast statement has two parts. The first, in quotes, is the label part, where you chose some description of the contrast you are running. The second is an exact specification of the contrast you want. NOTE: [a] the contrast coefficients must add to 0. [b] the order of the group variables is important here.

[3] MEANING OF CONTRAST: To explain this write MNS for the mean FEF of the non smoking group, etc.

[a] The first contrast performs an F-test of the hypothesis H0: MNS - MPS = 0.

[b] The second tests the hypothesis H0: MNS - (.1 MLS + .7 MMS + .2 MHS) = 0.

[c] The third began by choosing a weighted average of the 3 smoking groups based on the "average" number smoked per day, say 10, 20, and 42, and then turning that into a contrast by subtracting (10 + 20 + 42) / 3 = 24 from each (i.e., $LS = 10 - 24 = -14, MS = 20 - 24 = -4, HS = 42 - 24 = 18$.

[d] You run a contrast between MPS and MMS.

[4] MEANING OF ESTIMATION:

[a] The only thing that can be estimated is linear combinations of group means.

[b] Because of the way calculations are done, if that combination does not add to zero, as in a contrast, then you must include in the estimate specification the term `intercept T` where `T` is the TOTAL of the linear coefficients for the group terms.

[c] The first estimate is for MNS - (.1 MLS + .7 MMS + .2 MHS). From the output you can build a 95% CI of that. Notice that the output also includes a t-test.

[d] The second gives an estimate of the weighted average of the smoking group means.

[e] Estimate the difference between MPS and MMS.