Assignment 9 — Due November 14, 2003

1. Consider two different instruments used to measure the concentration of lead in water. One instrument is quite inexpensive whereas the other is expensive. A large tub of water contaminated with lead was prepared. The lead concentration throughout the tub was made as constant as possible by vigorous mixing. Eight randomly selected samples were obtained from the tub and subsequently analyzed with the inexpensive instrument; thirteen randomly selected samples were obtained from the tub and subsequently analyzed with the expensive instrument. The data are given below with the concentrations in parts per million.

   inexpensive instrument: 4.5, 8.1, 5.9, 3.7, 7.2, 5.4, 5.7, 6.6

   expensive instrument: 6.7, 7.0, 6.4, 7.3, 6.8, 6.9, 7.0, 6.6, 7.1, 6.8, 6.7, 6.8, 6.6

   (a) Determine using an appropriate normal-based test whether or not the mean lead concentrations are the same for both instruments. Perform (without using the computer) the tests assuming (i) the variances are equal and (ii) assuming that the variances are not equal. For part (ii) of this problem, you should use the approach described in the course notes in Section 10.3.2.

   (b) Repeat part (a) using R. (See R Appendix 10.9 for instructions.)

   (c) Perform a test for the equality of variances of the two instruments. What do you conclude?

   (d) Which of the two tests in part (a) (and in (b)) are you more likely to trust? Why? Explain what causes the difference you observe in the results of the two tests. (Hint: Consider the sample sizes.)

2. Breeding work is being done to incorporate good fruiting characteristics into a compact line of cucumbers. One difficulty with the compact line is a seed abnormality called "transverse" seed type. A question of interest is to determine whether the proportion of abnormal seeds produced will change (either increase or decrease) during the growing season. Consider the data from a particular plant. Early in the season 66 randomly selected seeds were examined of which 25 were transverse. Later in the season, 115 randomly selected seeds were examined of which 32 were transverse.

   (a) Perform a test of the claim that the proportion of abnormal (transverse) seeds is the same early and late in the growing season.

   (b) Find a 99% CI for the difference in proportions for early and late in the growing season.

   (c) State the assumptions necessary for this problem and evaluate them where possible.

3. Lakes in Southern Ontario are classified as small if their area is smaller than 20 hectares (ha), medium for areas between 20 and 100 ha, and large for areas above 100 ha. Given below are the measured alkalinity values (in $CaCO_3$ mg/L) for 12 randomly selected medium lakes and 15 randomly selected large lakes.

Medium:	3	4.5	8	13	18	21.5	22	24	26.5	29	31	34.5			
Large:	5	9	14.5	19	24.5	26	26.5	28	30.5	31.5	33	35	36.5	38	39

   (a) Make some type of display for these data that will enable you to determine if a normal model is reasonable for these data. Comment on your display.

   (b) Assuming that a normal model is not appropriate, carry out a test of the null hypothesis that the alkalinity of the two sizes of lakes is the same versus the two-sided alternative.

4. **Combining independent estimates of σ^2.** Suppose that the sample variances (s^2) computed from random samples from $k = 4$ (independent) populations are as follows:

population	1	2	3	4
sample variance	5.4	10.9	4.7	2.8

If we assume that each population has the same (true) population variance σ^2, then each separate sample variance s^2 is an estimate of σ^2. To combine these estimates into a single "pooled" estimate of σ^2, we use a weighted mean of the sample variances. The weights to use are proportional to the degrees of freedom associated with each sample variance (df $= n_i - 1$ where n_i is the sample size for the ith group). Thus the pooled estimator is:

$$s_p^2 = \frac{\sum_{i=1}^{4}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i\cdot})^2}{\sum_{i=1}^{4}(n_i - 1)} = \frac{\sum_{i=1}^{4}(n_i - 1)s_i^2}{(N-4)}$$

where $N = n_1 + n_2 + n_3 + n_4$. If there were k groups instead of 4, the final expression would read:

$$\frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{(N-k)}$$

Note carefully that we combine the *variances* in this way, not the standard deviations. If we are given standard deviations to combine, we must square them first to obtain the variances.

(a) Suppose $n_1 = n_2 = n_3 = n_4 = 8$. Calculate the pooled estimate s_p^2 and find 95% confidence limits for σ^2.

(b) Repeat part (a) for the following three sets of sample sizes: (n_1, n_2, n_3, n_4) = (2,2,2,26); (6,22,2,2); (4,12,4,12). Note that all 3 sets (and the set in (a)) have 32 total observations each. Comment on how the pooled estimates change with the different sets considered.

Remark: When variances are pooled in this way, the degrees of freedom associated with s_p^2 is the sum of the individual degrees of freedom: $\sum_{i=1}^{k}(n_i - 1) = N - k$. This equals 28 in all cases here. If the populations are normal (which we are to assume here in computing the CI's) then a chi-squared distribution on $N - k$ degrees of freedom (in this example, a chi-squared distribution on 28 df) is used for inference. Of course the assumption that all population variances are the same is important in the entire procedure.

5. The following is a subset from a larger experiment. The dry sheer strength (in pounds per square inch) of birch plywood bonded with 4 different resin glues (A, B, C, and D) were compared. Three pieces of plywood were tested with each glue type.

A	504	499	471
B	455	483	474
C	500	505	480
D	520	509	562

(a) Compute an ANOVA table for these data (using a hand calculator), including all relevant sums of squares, mean squares, and degrees of freedom.

(b) State the statistical model underlying the procedures in the analysis of variance as applied to these data. Define symbols used and make clear all distributional assumptions.

(c) State in words and symbols the null hypothesis and alternative hypothesis appropriate to this problem. Compute the relevant F-test and find the p-value for the test.

6. You are planning an independent-sample comparison of the effects of two diets on the weight gains of juvenile rats. With μ_1 and μ_2 as the population mean weight gains for the two diets, you are interested in testing $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 > \mu_2$. You know that weight gains are distributed approximately as a normal distribution. Furthermore, you know that the variance of the weight gain is 25 gm^2 for both diets. You have chosen a sample size of 9 rats for the first diet and a sample size of 16 rats for the second. Let $\bar{X}_1$ be the sample mean weight gain from diet 1 and let $\bar{X}_2$ be the sample mean from diet 2. Suppose that you choose to reject H_0 if $\bar{X}_1 - \bar{X}_2 > 6$.

Suppose that $\mu_1 = 28$ gm and $\mu_2 = 20$ gm. Find the power of the given test. (Hint: You know that $\bar{X}_1 - \bar{X}_2$ is normally distributed with a mean that depends on the hypothesized μ's and variance that you can calculate easily.)

Readings: Week 10: Course Notes Chapter 11