Stat 210 - HW 2 Solutions/Comments (Fall 2009)

1. **Two sample tests and outliers**

   (a) There are two two-sample t-tests that we have learned. The first uses a pooled s.d. and the second uses separate s.d. (and the Satterthwaite approximation). The question asks for an "appropriate test". In this case the sample variances are quite different which suggests that we use the separate variance or unpooled test. The unpooled test statistic is $t = \frac{28.53-18.80}{\sqrt{3.05^2/10+7.12^2/10}} = 3.97$. The estimated d.f. is 12.2 and the resulting P-value is .0018 (two-sided) which provide strong evidence that trauma patients expend more energy. Please remember to specify whether the p-value you report is for a one-sided or two-sided alternative. Just for your information .... the pooled test statistic is $t = \frac{28.53-18.80}{5.48\sqrt{2/10}} = 3.97$ (remember the test statistics are the same if the two groups have the same sample size) where 5.48 is the pooled standard deviation. When compared to the t-distrubtion with 18 degrees of freedom this yields $P = .0009$ (two-sided). The pooled analysis is overly optimistic in its conclusions that there is a significant difference (but not by much).

   (b) We can use the Wilcoxon rank sum test by ordering all the observations and recording the ranks of the non-trauma sample as 1,2,3,4,5,6,8.5,10,11,12. Note there is a tie for 8th and 9th so both observations get rank 8.5. The sum of the ranks is 62.5 which should be compared (assuming large enough samples) to $N(105, 13.2^2)$ reference distribution. This yields two-sided P of .0013 which is similar to the results above in suggesting significantly more metabolic expenditures for the trauma patients.

   (c) You were only required to note that the t-test statistic is more sensitive to outliers than the rank test. What happens to the t statistic depends on the specifics of the case since both the mean and s.d. are effected. For your information the results change as follows in this example: the mean of the non-trauma group is now 39.59, the t-statistic (for both pooled and unpooled) is now -0.52 which suggests a non-significant difference in the opposite direction than before, and the Wilcoxon statistic is 70.5 which still indicates significantly less metabolism in the non-trauma group.

2. **SAS: two sample methods**

   (a) It seems reasonable to assume that results for different hospitals are independent (both within and between groups). Examining either PROC UNIVARIATE output, a scatterplot of INF vs M, or the standard deviations reported in PROC TTEST (that's what I did) suggests that the two groups have different sample s.d.'s (1.11 for medical schools and 1.34 for non-medical schools) but these are not very different. By my rule of thumb the assumption of constant variance in the populations seems OK. The normal probability plots (separate by group; see program for how to do this) suggest normality is quite reasonable. Thus the assumptions seem fine.

   (b) The pooled t-test result reported by SAS gives $t = 2.52$ on 111 d.f. for two-sided p-value of .013. It appears the mean infection rate is significantly higher in hospitals with a medical school than in hospitals without one.

   (c) The estimate for the difference in mean infection rates is 0.87 and PROC TTEST gives a 95% confidence interval for the difference in mean infection rates as (.19, 1.55).

   (d) I proposed this transformation without having first checked to see whether it works. In fact it seems to make things worse. The variances are more uneven and the infection rates are less normal (especially in the non-medical school group). Oops! For the record the t-test results are similar.

   (e) This is an observational study. We did not assign medical schools at random to hospitals. This is important because hospitals with medical schools may differ in many ways from those without (e.g., medical schools may be more likely in urban areas, bigger hospitals, etc.). The possibility of such confounding variables means that we can't assume our "significant" difference means that medical school hospitals are less safe.

   (f) I will periodically ask you to write paragraph summaries of your data analyses. My solutions will give you some idea of what I'm looking for. It is important to connect with the real problem being analyzed. Here is one possibility:
   A comparison of hospital-acquired infection rates between hospitals with a medical school and those without suggests that hospitals with a medical school have higher infection rates. The mean infection rate in medical school hosptials is 5.09 infections per 100 patients compared to a mean infection rate of 4.22 infections per 100 patients in non-medical school hospitals. The difference of 0.87 infections per 100 patients is statistically significant ($p = .013$) and a 95% confidence interval for the population difference in means is $(0.19, 1.55)$. One limitation is that this is an observational study which means that the increased rate of hospital-acquired infections in medical school hospitals may be due to causes other than the presence of a medical school.

   Sample program: Note the program below creates all the output that you need for this question. In practice you probably wouldn't do this all at once; rather you might begin with the untransformed data and answer (a)-(c) and then modify the program to do (d).

```
filename senic 'h:\HAL\Courses\Stat210\senic.txt';
options linesize=80;
data infect;
    infile senic firstobs=2;
    input id stay age inf cult xray beds m r pat nur facil;
    logitinf = log(inf/(100-inf));
proc sort data=infect;
    by m;
proc rank normal=blom data=infect out=norm;
  var inf;
  by m;
  ranks nrm;
proc gplot data=norm;
  plot inf*nrm logitinf*nrm;
  by m;
proc ttest data=norm;
  class m;
  var inf logitinf;
run;
```

3. **Theory: effect of non-independence**

   (a) This is a standard result. Briefly, $E(\bar{Y}) = E(\frac{1}{n}\sum_i Y_i) = \frac{1}{n}\sum_i E(Y_i) = \frac{n\mu}{n} = \mu$ and $Var(\bar{Y}) = Var(\frac{1}{n}\sum_i Y_i) = \frac{1}{n^2}(\sum_i Var(Y_i) + \sum_i \sum_{j\neq i} Cov(Y_i, Y_j)) = \frac{1}{n^2}(n\sigma^2 + n(n-1)0) = \sigma^2/n$.

   (b) With dependence the mean argument works exactly the same but the covariances are not all zero in the variance calculation. Here we end up with $Var(\bar{Y}) = \frac{1}{n^2}(n\sigma^2 + \sum_i \sum_{j\neq i} \sigma^2 \rho^{|i-j|}) = \frac{\sigma^2}{n^2}(n + \sum_{i=1}^{n-1} 2(n-i)\rho^i)$ where the last term is obtained by carefully counting the number of terms of each size. This can be simplified a bit but that is not necessary for this problem.

   (c) If you compute the three examples then you find the variance of $\bar{Y}$ is $(i).7276\sigma^2; (ii).2600\sigma^2; (iii).0085\sigma^2$. For ten independent samples we'd have $.1000\sigma^2$ so you can see the pattern described in the problem (higher variance or less precision with positively correlated data).

4. **Theory: non-constant variance**

   (a) We are given that $s_1^2 = \sigma_1^2\chi^2_{n_1-1}/(n_1-1)$ and that $E(\chi^2_{n_1-1}) = n_1 - 1$ and $Var(\chi^2_{n_1-1}) = 2(n_1-1)$. It follows that $E(s_1^2) = \sigma_1^2(n_1-1)/(n_1-1) = \sigma_1^2$ and $Var(s_1^2) = \sigma_1^4 2(n_1-1)/(n_1-1)^2 = 2\sigma_1^4/(n_1-1)$. It also follows that $E(s_2^2) = \sigma_2^2$ and $E(s_p^2) = ((n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2)/(n_1+n_2-2)$.

   (b) MAJOR ERROR here on my part. Sorry! It should have said to define $Z = s_1^2/n_1 + s_2^2/n_2$ and then define $W = \nu Z/E(Z)$. I will first complete the problem as written and then show you what I really wanted! As written: Define $W = \nu s_p^2/(E(s_p^2))$. Clearly $E(W) = \nu E(s_p^2)/E(s_p^2) = \nu$. Next, $Var(W) = \nu^2 Var((n_1-1)s_1^2 + (n_2-1)s_2^2)/(n_1+n_2-2)^2 E(s_p^2)^2$. We can use our result from (a) for $Var(s_i^2)$ and $E(s_p^2)$ to determine that $Var(W) = \nu^2(2\sigma_1^4(n_1-1) + 2\sigma_2^4(n_2-1))/((n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2)^2$.

   (c) Setting $Var(W) = 2\nu$ and solving for ν leads to $\nu = ((n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2)^2/((n_1-1)\sigma_1^4 + (n_2-1)\sigma_2^4)$ which DOES NOT YIELD THE SATTERTHWAITE APPROXIMATION DESPITE WHAT I SAID!!!

   Part (b) as intended: It is still true that $E(W) = \nu E(Z)/E(Z) = \nu$. But now $Var(W) = \nu^2 Var(Z)/E(Z)^2 = \nu^2(\frac{2\sigma_1^4}{(n_1-1)n_1^2} + \frac{2\sigma_2^4}{(n_2-1)n_2^2})/(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})^2$.

   Part (c) as intended: Now setting $Var(W) = 2\nu$ yields $\nu = (\frac{sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})^2/(\frac{\sigma_1^4}{(n_1-1)n_1^2} + \frac{\sigma_2^4}{(n_2-1)n_2^2})$ which does give the Satterthwaite approximation if you replace σ_i^2 by s_i^2. I apologize for the error.

5. **Paired t-procedure**

   (a) The paired design was a good idea because using each pilot as his/her own control reduces variability and allows for more precise estimation of the effect of alcohol. Many people mentioned that pairing eliminates possible confounders. This is true **but it is not the primary motivation for pairing**. Remember that an analysis with two independent samples of pilots randomized to the two conditions would also eliminate confounding but would lead to less precise comparisons.

   (b) Using the paired analysis we find the 95% CI for the difference in means is $195.6 \pm t_{9,.975}230.5/\sqrt{10} = 195.6 \pm 2.262 * 72.89 = (30.7, 360.5)$. Alcohol reduces performance by 30-360 seconds on average. We are confident that alcohol reduces performance because 0 is not in the confidence interval. Recall that this means a two-sided p-value would be less than .05.

(c) The most direct evidence that pairing worked is obtained by comparing the standard error for the difference in means under the paired analysis ($230/\sqrt{10} = 72.89$) and under an independent samples analysis (assuming the given standard deviations this would be $\sqrt{238.8^2/10 + 210.9^2/10} = 100.75$). Thus pairing leads to a more precise inference. You can also compute the correlation of the two columns (no alcohol and alcohol) as .48 which is reasonably high and thus justifies the pairing. Note that you can also infer the correlation by setting the variance of the difference (230.5^2) equal to the formula for for the variance of the difference ($238.8^2 + 210.9^2 - 2(238.8)(210.9)\rho$) and solving for $\rho = .48$.

(d) Everyone gave the correct definition of type I error and type II error. Almost everyone realized that a type I error (rejecting H_o and banning alcohol use when it's probably OK) is not a big deal in this case while a type II error (accepting H_o and allowing alcohol use when it's harmful) is a big deal. Very few people however went the next step and argued that we should therefore use a BIGGER cutoff in this case say .10 or .15. This will make it easier to reject H_o which means more type I errors but it will also make it easier to reject when H_a is true which means fewer type II errors. Here's a case where a p-value of .09 or even .12 might be enough to want to reject H_o!!