1. Two-sample tests and assumption violations

(a) There are two two-sample t-test procedures that we have learned. The first uses a pooled s.d. and the second uses separate s.d.’s (and the Satterthwaite approximation for the degrees of freedom). 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 unpoled test. The unpoled test statistic is \( t = \frac{29.23 - 18.80}{\sqrt{3.05^2/10 + 7.39^2/10}} = 3.73 \). The estimated d.f. is 12.0 and the resulting P-value is .0028 (two-sided) which provides 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 \( t = \frac{29.23 - 18.80}{5.65\sqrt{2/16}} = 3.73 \) (remember the test statistics are the same if the two groups have the same sample size) where 5.65 is the pooled standard deviation. When compared to the \( \text{t} \)-distribution with 18 degrees of freedom this yields \( P = .0015 \) (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, 11, 12, 13. Note there is a tie for 8th and 9th so both observations get rank 8.5. The sum of the ranks is 65.5 which should be compared (assuming large enough samples) to \( N(105, 13.2^2) \) reference distribution. This yields two-sided \( P = .0028 \) which is similar to the result 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. Just changing that one value has no impact on the rank test in this case because that value is already the largest. The impact on the t-test may surprise you. The trauma patients already had higher metabolic expenditures and now we’ve just made the difference more extreme — but the outlier changes both the mean and the s.d.. The \( t \)-test may surprise you. The trauma patients already had significant higher metabolic expenditures and now we’ve just made the difference more extreme but the outlier changes both the mean and the s.d.. The new t-test statistic is 1.18 and the p-value is 0.27 so one data entry error can completely change the conclusion.

2. Theory: dependence

(a) Within a sample

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

ii. 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 \( Var(\bar{Y}) = \frac{1}{n^2}(n \sigma^2 + \sum_{i} \sum_{j \neq i} \sigma_{ij}^2) = \frac{\sigma^2}{n}(n + n(n-1)p) \) where the last term is obtained by counting the number of correlated pairs. This can be simplified a bit to yield \( Var(\bar{Y}) = \frac{\sigma^2}{n(1+(n-1)p)}. \)

iii. The variance is \( \sigma^2(1 + 99 \times .05)/100 = \sigma^2(.0595) = \sigma^2/16.81. \)

iv. The result in (b) shows that \( Var(\bar{Y}) = \frac{\sigma^2}{n(1+(n-1)p)}. \) If \( p \) equal zero we find that the variance for \( n \) independent observations is \( \sigma^2/n. \) If there is correlation, then the result in (b) shows that the "effective sample size" is reduced from \( n \) to \( n/(1+(n-1)p). \) The effective sample size in our case is \( (100/5.95) = 16.81. \)

(b) Between samples

i. Note that \( E(Y_{1j}) = \mu_1 \) and \( E(Y_{2j}) = \mu_2. \) Also note that \( Var(Y_{1j}) = Var(Y_{2j}) = \sigma_\beta^2 + \sigma_\epsilon^2. \) We compute \( Cov(Y_{1j}, Y_{2j}) = E((Y_{1j} - E(Y_{1j}))(Y_{2j} - E(Y_{2j}))) = E((\beta_j + \epsilon_{1j})(\beta_j + \epsilon_{2j})) = \sigma_\beta^2. \) Then \( Corr(Y_{1j}, Y_{2j}) = \frac{\sigma_\beta^2}{(\sigma_\beta^2 + \sigma_\epsilon^2)}. \)

ii. The pooled and unpoled t-tests statistics both assume independence between samples. They will use the incorrect variance. Usually the dependence is positive; in this case the estimated variance under independence \( (2\sigma^2/n) \) is an overestimate of the true variance \( (2\sigma^2(1 - \rho)/n) \) (where \( \rho \) is the correlation between pairs of observations on the same subject). The overestimate will make the procedures conservative (confidence intervals too wide, test-statistics too small).

iii. Note that \( d_j = Y_{1j} - Y_{2j} \) is the difference between two normal random variables and thus has a normal distribution. It has mean \( E(d_j) = \mu_1 - \mu_2 \) and \( Var(d_j) = Var(\epsilon_{1j} + \epsilon_{2j}) = 2\sigma_\epsilon^2. \) It thus has constant variance. Finally the \( d_j \)'s are independent because the different individuals are assumed independent. Thus \( d_j \)'s satisfy usual assumptions.

iv. \( d \pm t_{n-1,1-\alpha/2}s_d/\sqrt{n} \) is a 100(1 - \( \alpha \))\% confidence interval for \( \mu_1 - \mu_2. \)

3. 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. Some people mentioned that pairing eliminates possible confounders. This is true but it is not the primary motivation for pairing. Why do I say this? 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. The most direct evidence that pairing worked is obtained by comparing the standard error for the difference in means under the paired analysis \((230.5/\sqrt{10} = 72.89)\) 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 \(\rho = .48\).

(b) Using the paired analysis we find the 95% CI for the difference in means is 195.6 ± \(t_{9.975}230.5/\sqrt{10} = 195.6 ± 2.262 \times 72.89 = (30.7, 360.5)\). Alcohol reduces the amount of time of appropriate 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. Note: This problem statement suggests a one-sided alternative. Thus one could argue for a one-sided confidence interval (i.e., a lower confidence limit); I generally prefer two-sided confidence intervals. This does however change the relationship between the confidence interval and the p-value.

(c) The type I error (rejecting \(H_o\) and banning alcohol use when it doesn’t effect performance) 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. Recognizing this, it can be argued that we should therefore use a different threshold for declaring significance in this case, perhaps requiring p-value less than .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 \(H_o\) 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 lead us to want to reject \(H_o\)!

(d) If there is some effect of going into the simulator second, then this will be confounded with the treatment. For example, if pilots learn by being in the simulator, then their second experience (with alcohol) might be better than it would be in the absence of learning. Here that would minimize the difference between the no alcohol and alcohol conditions. If that is true then the population and the population of interest. One could argue that the research group is interested in the population describing how effective the participant thought the product was likely to be with higher scores meaning more than it would be in the absence of learning. Here that would minimize the difference between the no alcohol and alcohol conditions. It would be better to randomize the order so half of the pilots did the alcohol condition first.

4. ANOVA and contrasts

(a) First note that the question is not very clear about what the outcome measure is. It is a score on a 7-point scale describing how effective the participant thought the product was likely to be with higher scores meaning more effective. As we have discussed in class questions like this draw our attention to the difference between the study population and the population of interest. One could argue that the research group is interested in the population of all potential consumers or more likely all potential Hispanic consumers. If that is true then the population mean is the mean effectiveness score for all consumers shown a given advertisement. It seems more appropriate here to define the population as students at this university. I don’t worry too much about the answer you give as long as you give some thought to the topic! (Of course there are wrong answers .... It would be wrong to say the population is the 200 students in the study!)

(b) We can ignore the missing data if the individuals who did not reply are assumed to be like those that did. Thus if we could we might like to compare age, gender, GPAs for the students with missing replies and compare them to the students who did reply. If they differed in some way that might impact whether the sample was in fact representative of the population.

(c) The ANOVA table entries are as follows:

\[
\begin{align*}
SS(\text{Within}) &= SS(\text{Residuals}) = \sum_{i} (n_i - 1)s_i^2 = 473.78 \\
SS(\text{Between}) &= SS(\text{Groups}) = \sum_{i} n_i(Y_i - \bar{Y})^2 = 18.69 \\
SS(\text{Total}) &= SS(\text{Within}) + SS(\text{Between}) = 492.47.
\end{align*}
\]

The d.f. are respectively 170, 3, 173. The mean square within (error) is 2.79 and the mean square between (groups) is 6.23. We can test the null hypothesis of equal population means across the four groups with the F-test; these data yield \(F = 6.23/2.79 = 2.23\) which when compared to the \(F_{3,170}\) distribution yields a p-value of approximately .09. There is only weak evidence of a difference in the responses to the different advertisements.

(d) Contrasts

i. The natural contrast weights here are (1/3, 1/3, 1/3, -1) which compares the average of the three versions where the model stares at the camera to the one version where the model doesn’t. The contrast is \((3.19+3.72+3.86)/3 - 3.11 = 0.48\). The contrast standard error is \(\sqrt{MSE * (1/9(1/45) + 1/9(1/39) + 1/9(1/47))} = \sqrt{2.79 * 0.0309} = 0.29\). The 95% CI is then \(0.48 ± 1.974 * .29 = (-.10, 1.06)\). An important thing to note that is if you used (1, 1, 1, -3) as the weights, then both the mean and the s.e. would change by a factor of three. Thus your inferences would be similar ... but the contrast is a bit harder to interpret since it is no longer on the scale of the original measurements.

ii. The contrast weights in this case are (-1/2, 1, -1/2, 0) which leads to a 95% CI of \(.195 ± 1.974 * .32 = (-.435, .825)\).

(e) It is possible to provide a very short summary here since in truth there is not much evidence that the model’s eye color or eye gaze has any impact on the student scores. There is weak evidence for a couple of effects but
nothing that would meet the usual significance levels. A possible paragraph: Researchers used a sample of 174 college students at a large predominantly Hispanic university to test the impact of a model’s eye color or eye gaze (at the camera or down) on consumer’s attitudes toward the product. They created four versions of the same advertisement, one with the model looking down (eyes not visible) and three others with the model looking at the camera (with either blue, green or brown eyes). There was only weak evidence that eyes had any impact on consumers (p-value = .09 for a one-way analysis of variance comparing the mean scores, F(3, 170)). Closer examination suggests that there was essentially no difference based on eye color but that students did find ads with the model staring at the camera as more effective. One limitation here is that there were some missing values and no information about why they were missing.

5. ANOVA in R for prostate cancer data - see code below

(a) The qqplots show evidence of skewed distributions with long tails within each of the groups (gleason=6, gleason=7, gleason=8). The standard deviation of the measurements in the gleason=8 group is much larger than in the other groups.

(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. 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) One can use the libraries lsmeans and car that I demonstrated in discussion to carry out the paired comparison tests for part (d) and the contrast tests for part (e). (See code below). Using the multiple comparisons approach in this problem (even though there are only three paired comparisons) with the Bonferroni correction (we should use the .991667 percentile of the t distribution as our threshold which corresponds to a two-sided test with $\alpha/3 = .01667$) 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 (with the Bonferroni adjustment) but the difference there is nearly significant.

(e) For the linear trend we use $c = (-1, 0, 1)$ which yields a highly significant result ($t = 6.512, p < .0001$). 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 ($t = 1.82, p = .0734$). Note that here I did not use a multiple comparisons procedure (but could and probably should have). These results make us suspect non-linearity (a quadratic trend) but the evidence is not very strong. 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. One can propose other forms for the “quadratic” contrast. In fact, any contrast (other than the linear contrast) has a quadratic component 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) Once again, the key is to try and connect to the science rather than just report on statistical methods. Imagine you are communicating with non-statisticians (so the statistics goes in parentheses). Here’s mine:

An observational study of 97 men with prostrate 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 ($F(2,94)=21.6$)). 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).

Sample R Code

```r
# read data
datainput <- read_csv("H://HAL/Courses/Stat210//prostate.csv")
prostate <- data.frame(datainput)
# explore data
prostate$gleason <- factor(prostate$gleason)
tapply(prostate$psa, prostate$gleason, describe)
qqnorm(prostate$psa[prostate$gleason==6])
qqnorm(prostate$psa[prostate$gleason==7])
qqnorm(prostate$psa[prostate$gleason==8])
boxplot(prostate$psa ~ prostate$gleason)
# try transformations
prostate$rootpsa <- sqrt(prostate$psa)
```
```r
tapply(prostate$rootpsa, prostate$gleason, describe)
tapply(prostate$logpsa, prostate$gleason, describe)
```