Stat 265 - HW 1 Solutions/Comments (Spring 2018)

1. **Completely randomized study** – Fisher's test

   (a) This question was asking about experimental design. Good experimental design requires that we try to identify the effect of the manipulation of interest while keeping other things the same. If we did not fly the plane on non-seeding days, then we would not know whether observed differences were due to the plane flying through or the treatment. Thus having the plane on both days eliminates a possible confounding factor. Several students suggested that a failure to have the plane fly would violate the unconfoundedness assumption. This is not quite right ... the unconfoundedness assumption concerns whether we believe randomization conditional upon a specified set of covariates (there are none here). The question also asks about why the pilots were kept blind to the assigned treatment. This is to avoid any potential bias – the pilots might (intentionally or not) fly differently on treatment days. This could effect the outcome of the study.

   (b) Everyone was able to do this simulation using the code provided. The randomization p-value (based on an extremely large sample) is 0.044. As we discover in part (c) the standard error associated with 1000 simulated random assignments is .0065. This means you may have gotten a result anywhere betwee .031 and .057.

   (c) The aim here was to have you recognize and evaluate the role of Monte Carlo variability in the simulations. The standard error of the estimated p-value is $\sqrt{p(1-p)/1000}$; with $p = .044$ the standard error is approx .0065. I asked about the "expected variability in a sample proportion" and several people reported the binomial variance. We almost always report the standard deviation (standard error) because it is in the same units as the original measurement.

   (d) Best here to use the two-sample t-test allowing for unequal variances. I generally use two-sided alternatives. Here that yields p-value .054. (It was OK to assume equal variances and/or use one-sided alternatives.) The t-test and Fisher procedure would produce similar results in large samples.

   (e) Many people used the ratio of standard deviations as a test statistic. This yielded simulated p-values around .075.

   (f) For the test statistic comparing the mean of the logarithms of the amounts the p-value turns out to be .015.

   (g) I was hoping that folks would actually plot the data or at least examine it closely. If you do so, then you see that the data are very highly skewed and that the variability in the two groups is very different. Taking the logarithm of the data is extremely natural in this case for both statistical (it makes the data better fit the assumptions behind traditional statistical techniques) and scientific (the data are volumes and so you would not expect a treatment to have an additive effect) reasons. Thus my conclusion is that cloud seeding increases mean log rainfall by 1.14 or equivalently increases median rainfall by a factor of 3.1 (exp(1.14)) or so.

2. **Completely randomized study** – Neyman's test

   (a) The original design is superior because the manipulation is being done at the level of the classroom. All students in the classroom are being effected by the same "implementation". We would expect the outcomes for individual students in the class to be dependent because they have a common teacher and environment and they impact each other. It is appropriate to indicated this likely violated SUTVA (though I'm not sure this is the most precise way to describe the issue). It is not true that this violates the assumption of individualistic assignment; the marginal probability for each student is the same even though there are restrictions that insure individuals in the same room receive the same treatment. I agree this is a little confusing; it has to do with the way we use the term individualistic.

   (b) The estimated average treatment effect is 14.5. The estimated (Neyman) variance is 60.6. An approximate 95% CI is $14.5 + / - 1.96\sqrt(60.6) = (-0.76, 29.76)$. We would not reject the null hypothesis at the .05 level. Note that it is not strictly speaking appropriate to use a t critical value here; that relies on a normal distribution assumption that the Neyman approach is not making. Of course, it's also not appropriate to use a large sample approximation with six observations in each group!

(c) Everyone noticed that the pre-test scores in the two groups had very similar means. Thus we would believe that the treatments were randomly assigned. Note that a significance test can not **prove** that treatment was randomly assigned. A rejection of the null hypothesis of equal means would lead us to question random assignment but the failure to reject just means that random assignment is plausible.

(d) The estimated average treatment effect is 15.28 and the estimated variance is 28.97. A 95% CI is $(4.73, 25.83)$. The treatment effect estimate changes a bit but the most noticable change is that the confidence interval is considerably narrower (our estimate is more precise). We would now reject the hypothesis of no effect at the .05 level.

(e) The regression model leads to an estimated treatment effect of 15.07 and standard error of 5.11. Using 1.96 as we did for Neyman approach leads to a CI of (5.05, 25.09). The CI is a bit wider if you use the traditional regression t-critical value. As in (d) we reject the hypothesis of no treatment effect.

(f) Note that the last two analysis are both appropriate for a randomized study. The advantage of the change score approach is that it does not make any specific assumption about there being a linear relationship between post-test and pre-test. The advantage of the regression model is that it the change score implicitly assumes a slope of one and there's no reason to believe this is true. Thus we might expect the regression to provide more precise inference. Not much difference in this example.

3. **Theory**

(a) There are several ways to write the test statistic. One natural way is $\hat{\tau} = \frac{1}{N_T}\sum_i W_i Y_i(1) - \frac{1}{N_C}\sum_i (1 - W_i)Y_i(0)$ with $N_T = N_C = 6$.

(b) Under complete randomization only W_i's are random. Then $E(\hat{\tau}) = \frac{1}{N_T}\sum_i E(W_i)Y_i(1) - \frac{1}{N_C}\sum_i (1 - E(W_i))Y_i(0)$. Moreover under randomization it is clear that $E(W_i) = Pr(W_i = 1) = 6/12 = 1/2$ which yields $E(\hat{\tau}) = \frac{1}{2N_T}\sum_i Y_i(1) - \frac{1}{2N_C}\sum_i Y_i(0) = \frac{1}{12}\sum_i (Y_i(1) - Y_i(0)) = \tau$ (the finite population causal estimand). Thus our test statistic is an unbiased estimate of the average treatment effect.

(c) We went over this class. You can write the change score estimate in the same manner as part (a) with $Y_i(w)$ replaced by $Y_i(w) - X_i$. The same argument then leads to the conclusion that the statistic is unbiased.

(d) The two statistics are both unbiased which means they have the same expected value over the set of possible randomizations. In problem 2, we observe one particular randomization – there is no guarantee that the two statistics will be the same for a given randomization (just on average).

4. **Projects**

Thanks for the proposals.