Homework 1
Stat 215A, Fall 2020

Due: push a `homework1.pdf` file to your `stat-215-a` GitHub repo by **Thursday, September 17 11:59pm**

1 Hypothesis testing, the t-distribution

Imagine we observe $(x_1, y_1), \ldots, (x_n, y_n)$ where (x_i, y_i) are multivariate normal with mean (μ_x, μ_y), $Var(x_i) = Var(y_i) = \sigma^2$ and correlation ρ. We are interested in testing the null hypothesis that $\mu_x = \mu_y$.
Under the null hypothesis we know

$$t = \frac{(\bar{x} - \bar{y})}{s_{pooled}\sqrt{2/n}}$$

is distributed as a Student's t with $2n - 2$ degrees of freedom, where s_{pooled} is the pooled sample standard deviation. See any undergraduate text (or Wikipedia page "Student's t-test") if you are unfamiliar with the t distribution.

1. Write s_{pooled} in terms of $x_i, y_i, \bar{x}$ and $\bar{y}$ (this is a standard definition)

$$s_{\text{pooled}} = \sqrt{\frac{(n_x - 1) \cdot s_x^2 + (n_y - 1) \cdot s_y^2}{n_x + n_y - 2}}$$

   Plugging in the definition of s^2 we get:

$$s_{\text{pooled}} = \sqrt{\frac{\sum_{i=1}^{n_x} (x_i - \overline{x})^2 + \sum_{i=1}^{n_y} (y_i - \overline{y})^2}{n_x + n_y - 2}}$$

2. What is the expectation of s^2_{pooled}?

   It is a well known result in statistics that s^2 is an unbiased estimator for the sample variance (assuming normal i.i.d sample).
   Using that result:

$$\begin{aligned}
\mathbb{E}s^2_{\text{pooled}} &= \mathbb{E}\frac{(n_x - 1) \cdot s_x^2 + (n_y - 1) \cdot s_y^2}{n_x + n_y - 2} \\
&= \frac{(n_x - 1) \cdot \sigma^2 + (n_y - 1) \cdot \sigma^2}{n_x + n_y - 2} \\
&= \sigma^2
\end{aligned}$$

3. The statement above (on the t-statistic) isn't quite right. Are any additional assumptions needed?
   We need to assume $\rho = 0$ otherwise the sum of the estimated variances isn't distributed χ^2

Consider doing a paired t-test with the same data. The test statistic here is

$$t_{paired} = \frac{(\bar{x} - \bar{y})}{s_{diff}\sqrt{1/n}}.$$

4. Write s_{diff} in terms of $x_i, y_i, \bar{x}$ and $\bar{y}$. (another standard definition)
   Let $d_i = x_i - y_i$

$$\begin{aligned} s_{\text{diff}} &= \sqrt{\frac{\sum_{i=1}^{n} \left(d_i - \overline{d}\right)^2}{n-1}} \\ &= \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - y_i - \overline{x} + \overline{y}\right)^2}{n-1}} \end{aligned}$$

5. What distribution does t_{paired} have?
   t_{paired} has a student t distribution with $n-1$ degrees of freedom this is because we simply take the diffs and from that point on it is a regular normal i.i.d sample

6. What is the expectation of s^2_{diff}?
   s^2_{diff} is an unbiased estimator for $\text{Var}(x_i - y_i) = 2\sigma^2 - 2\rho$, hence:

$$\mathbb{E}s^2_{\text{diff}} = 2\sigma^2 - 2\rho$$

7. Compare $s^2_{diff}\frac{1}{n}$ to $s^2_{pooled}\frac{2}{n}$. When is $s^2_{diff}\frac{1}{n} < s^2_{pooled}\frac{2}{n}$?
   First let us note that

$$\begin{aligned} s^2_{diff}\frac{1}{n} &= \frac{\sum_{i=1}^{n} \left(x_i - y_i - \overline{x} + \overline{y}\right)^2}{(n-1)\cdot n} \\ &= \frac{\sum_{i=1}^{n} \left((x_i - \overline{x}) - (y_i - \overline{y})\right)^2}{(n-1)\cdot n} \\ &= \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 + (y_i - \overline{y})^2 - 2\cdot((x_i - \overline{x})(y_i - \overline{y}))}{(n-1)\cdot n} \\ &= \frac{\sum_{i=1}^{n} \left((x_i - \overline{x})^2 + (y_i - \overline{y})^2\right) - 2\cdot \hat{\text{Cov}}\left(X, Y\right)}{(n-1)\cdot n} \end{aligned}$$

   Using this identity we conclude that:

$$s^2_{diff}\frac{1}{n} < s^2_{pooled}\frac{2}{n} \Leftrightarrow \hat{\text{Cov}}\left(X, Y\right) > 0$$

   When is $E(s^2_{diff}\frac{1}{n}) < E(s^2_{pooled}\frac{2}{n})$?
   Using the calculation above we immediately get that:

$$\mathbb{E}(s^2_{diff}\frac{1}{n}) < \mathbb{E}(s^2_{pooled}\frac{2}{n}) \Leftrightarrow \rho > 0$$

8. From these computations, what do you learn?
   We would prefer using the the pooled statistic if we know that $\rho = 0$ since the variance estimator is less noisy (doesn't have the estimated covariance component). In other cases we should pick the paired statistic, as the pooled statistics doesn't have the correct distribution.

2 Questions from Freedman

In Freedman, do questions 1 - 5 and 9 starting on page 13. This may sound like a lot of work. However, once you do the reading, each question should have a straight-forward answer.

- In the HIP trial (table 1), what is the evidence confirming that treatment has no effect on death from other causes?
  The similarity between control and total treatment all other death rate. If there was an effect we would expect to see on the total death rate from all other causes (Given that most of the treatment group we in fact treated).

- Someone wants to analyze the HIP data by comparing the women who accept screening to the controls. Is this a good idea?
  This would not be a good idea, as it was explained that the women who accepted the treatment were on average richer and more educated.

- Was Snow's study of the epidemic of (table 2) a randomized controlled experiment or a natural experiment? Why does it matter that the Lambeth company moved its intake point in 1852? Explain briefly.
  It was a natural experiment, as the researcher did not decide what house would get it's water supply from what company. As the point of the experiment was to check the effect of contaminated water of cholera it was important to have a company with clean water - Lambeth from 1852. Otherwise we wouldn't have a proper control/treatment groups.

- Was Yule's study a randomized controlled experiment or an observational study?
  It was an observational study

- In equation (2), suppose the coefficient of ΔOut had been -0.755. What would Yule have had to conclude? If the coefficient had been +0.005?
  Following Yule's logic, had the coefficient been -0.755 he would come to the conclusion that an increase in out-relief ration would decrease poverty. If the coefficient had been 0.005 he would probably think that there's no effect.

- Keefe et al (2001) summarize their data as follows:
  "Thirty-five patients with rheumatoid arthritis kept a diary for 30 days. The participants reported having spiritual experiences, such as a desire to be in union with God, on a frequent basis. On days that participants rated their ability to control pain using religious coping methods as high, they were much less likely to have joint pain."
  Does the study show that religious coping methods are effective at controlling joint pain? If not, how would you explain the data?
  This study does not show that religious coping methods are effective at controlling joint pain, as this finding can also be explained by saying that there's actually a reverse causality meaning that they reported that their ability as high because they were having less pain.

Please look into the following map: `http://www.ph.ucla.edu/epi/snow/snowmap1_1854_lge.htm`. This is the map made by John Snow regarding the Broad Street pump. Each small block marks a cholera patient.

- How would you transform this display into numerical measures?

- What would be gained quantifying the effects? What would be lost?
  We are interested in a numerical measure that would be able to distinguish houses/people that consumed water from Broad street pump to those who did not, since the actual problem we are interested in is the connection between consuming water from the pump to cholera.
  The most naive way that I can think of is to simply take the distance to the pump with the basic logic that the closer you are the more likely you are to consume water from the pump, using the distance we can try to formulate a statistical test to determine whether being closer is correlated with cholera.

We would obviously miss all the exceptions i.e people who live far/close and use/do not use the pump which will make it harder on us to establish reliable results.