

# Homework 1

## Stat 215A, Fall 2022

**Due:** push a `homework1.pdf` file to your `stat-215-a` GitHub repo by **Friday, September 23 11:59pm**

### 1 Hypothesis testing, the $t$ -distribution

Imagine we observe  $(x_1, y_1), \dots, (x_n, y_n)$  where  $(x_i, y_i)$  are multivariate normal with mean  $(\mu_x, \mu_y)$ ,  $\text{Var}(x_i) = \text{Var}(y_i) = \sigma^2$  and correlation  $\rho$ . 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)
2. What is the expectation of  $s_{pooled}^2$ ?
3. The statement above (on the  $t$ -statistic) isn’t quite right. Are any additional assumptions needed?

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)
5. What distribution does  $t_{paired}$  have?
6. What is the expectation of  $s_{diff}^2$ ?
7. Compare  $s_{diff}^2 \frac{1}{n}$  to  $s_{pooled}^2 \frac{2}{n}$ . When is  $s_{diff}^2 \frac{1}{n} < s_{pooled}^2 \frac{2}{n}$ ? When is  $E(s_{diff}^2 \frac{1}{n}) < E(s_{pooled}^2 \frac{2}{n})$ ?
8. From these computations, what do you learn?

### 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.

Please look into the following map: [http://www.ph.ucla.edu/epi/snow/snowmap1\\_1854\\_lge.htm](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?

### 3 Miscellaneous

- Contrast the data science life cycle outlined by the PCS framework to an unstructured data science process (this can be either abstractly or with a fixed data example). What is an application area where you believe the PCS framework can have impact?
- Give an example of a methodological problem that Tukey names as a significant question in data analysis and you believe is still an open problem. Explain where and when this problem arises in real data.
- What are some dangers of an open feedback loop between theory and practice in science/data science?