Prof. Ahlstrom
MATH 2310 Sec. 01

Probability and Statistics for the Sciences and Engineering (MATH 2310)
Lab 7: Hypothesis testing
20 points

Credit to Dr. McLean Sloughter and Dr. Galen Egan for designing significant portions of this assignment

---

In this lab, you will write code to calculate summary statistics, such as the sample mean and standard deviation. You will also write code using these values for hypothesis testing.

In previous labs, we looked at a research paper from 1983 by Chambers, Cleveland, Kleiner, and Tukey examining the effectiveness of cloud seeding using silver nitrate. Total rainfall (in acre-feet) was measured for 26 seeded clouds and 26 unseeded clouds. We will be examining this data set further. This file (`clouds.xlsx`) can be accessed on Canvas. **Submit code and outputs** for each of the following questions.

1. **Cloud seeding**

   (a) Compute the mean and standard deviation of the unseeded cloud data. Also compute the mean and standard deviation of the seeded cloud data.

   (b) Our data sets in this case are not large enough to justify applying the central limit theorem directly and using a z-test, **but** we'll do it anyways to start out with, as an academic exercise and for the sake of comparison.

   The large-sample test statistic used to consider whether the means of two distributions are equal to each other is of the following form:

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

   where Sample 1 of size n_1 has mean $\overline{x}_1$ and standard deviation s_1, and Sample 2 of size n_2 has mean $\overline{x}_2$ and standard deviation s_2.

   Let's have Sample 1 be our "seeded" cloud data, and Sample 2 be our "unseeded" cloud data. We want to consider the null hypothesis H_0 that the population mean for seeded data μ_s is approximately equal to the mean for unseeded data μ_u. It is believed that seeding clouds will actually increase rainfall, so we will consider the alternative hypothesis H_a that $\mu_s > \mu_u$.

Compute the value of this test statistic z based on the values you calculated in part (a).

(c) Let's consider a significance level $\alpha = 0.05$ with regard to this problem. **Compute the value** of the associated threshold z_α out of a standard normal distribution. The relevant rejection region will then be $z \geq z_\alpha$. Does the test statistic that you calculated in part (b) satisfy this rejection region inequality? What does that mean in the context of this problem?

(d) Compute the p-value associated with the test statistic you found in part (b). Recall that for an upper-tailed problem like this, the p-value can be calculated as $P = 1 - \Phi(z)$, where Φ represents the cumulative distribution function (cdf) of the standard normal distribution. What does this p-value mean in the context of this problem?

(e) As mentioned previously, our data sets for seeded and unseeded clouds are probably not actually large enough to justify applying the assumptions of the central limit theorem. We could, however, use a small-sample Student's t hypothesis test in order to approach the problem. Recall that this would require an initial assumption that the underlying population is normally distributed.

For a normal distribution, we would presume that the source distribution is unimodal and symmetric. In a previous lab assignment, however, we found that the cloud data sets were rather skewed. Taking the log transform, however, seemed to result in a more symmetric sample distribution. Perhaps the lognormal distribution, then, is more well suited than just the normal distribution to model expected rainfall amounts.

Hence, **take a log transform** of the cloud data, so that we can more plausibly assert the assumption that the underlying population is normally distributed.

(f) Compute the mean and the standard deviation of the **log-transformed** data for **seeded** clouds. Also compute the mean and the standard deviation of the **log-transformed** data for **unseeded** clouds.

(g) Similarly to part (b), we can calculate a test statistic to consider whether the means of the two distributions are approximately equal to each other. This test statistic takes the following form:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

Compute this test statistic t using the **log transformed** summary data from part (f).

(h) We can now perform a Student's t-test comparing our calculated test statistic to a threshold based on a Student's t distribution. Let's again use $\alpha = 0.05$ as our significance level. The t distribution requires an additional parameter, called the "degrees

of freedom". The type of problem we're looking at here, using hypothesis testing on two separate samples with potentially unequal variances, is sometimes called a "Welch's t-test", and provides the following formula for the degrees of freedom to be used for the hypothesis test:

$$d_f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

Using this formula, **compute the value** of the threshold $t_{\alpha,\, d_f}$ for this test. The rejection region for this test statistic (calculated in part (g)) will then be $t \geq t_{\alpha,\, d_f}$. Does the test statistic that you calculated in part (b) satisfy this rejection region inequality? What does that mean in the context of this problem?

(i) Compute the p-value associated with the test statistic you found in part (g). Recall that for an upper-tailed problem like this, the p-value can be calculated as $P = 1 - F_{d_f}(t)$, where F_{d_f} represents the cumulative distribution function (cdf) of the Student's t distribution with d_f degrees of freedom. What does this p-value mean in the context of this problem?