# SDS 220 - Practicing Hypothesis Testing

IMS 16, 17 , 19, 20

1. Elevated mercury concentrations are an important problem for both dolphins and other animals. We want to investigate the average mercury content in dolphin muscle using a sample of 19 Risso's dolphins from the Taiji area in Japan. Summary statistics: $\overline{x} = 4.4$, $s_x = 2.3$, $min = 1.7$, and $max = 9.2$. Test if the average mercury content for dolphins in the population is 6.

   (a) Write the hypotheses in symbols.

   (b) Check conditions.

   (c) Find the p-value.

   (d) What is the conclusion of the hypothesis test using $\alpha = .10$?

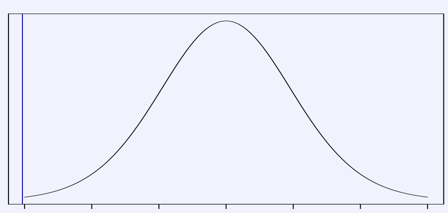
   (a) $H_0 : \mu = 6 \qquad H_A : \mu \neq 6$

   (b) **Independence** ✓: This appears to be a random sample of dolphins that are independent of each other. **Normality** ✓: The smallest and largest values indicate the data is approximately symmetric. There do not appear to be extreme outliers.

   (c) The test statistic is

   $$T = \frac{4.4 - 6}{2.3/\sqrt{19}} \approx -3.03$$

   We can find the p-value by typing the code `2*pt(-3.03, df= 19)`, which returns 0.0072.


   (d) Our p-value is approximately 0.0072 which is less than $\alpha = 0.10$. We conclude against $H_0$. It is unlikely that the average mercury content for dolphins is 6.

2. A Gallup poll surveyed Americans about their employment status and whether they have diabetes. The survey results indicate that 1.5% of the 47,774 employed (full or part time) and 2.5% of the 5,855 unemployed 18-29 year olds have diabetes.

   (a) State appropriate hypotheses to test for difference in proportions of diabetes between employed and unemployed Americans.

   (b) Conduct a hypothesis test using $\alpha = 0.05$.

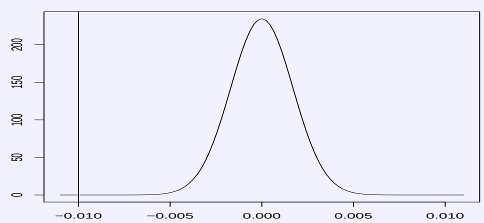
   (c) The sample difference is about 1%. The hypothesis test had a p-value that is very small (about 0), meaning the difference is statistically significant. Use this result to explain the difference between statistically significant and practically significant findings.

(a) $H_0 : p_e - p_u = 0 \qquad H_A : p_e - p_u \neq 0$

(b) First we calculate $\hat{p}_{pool} = \frac{862.985}{53629} \approx 0.0161$, so we can check conditions. **Independence** ✓: We have no reason to suspect dependence between the groups (we do not have people simultaneously in both groups), or between subjects (one subject cannot give another subject diabetes). **Success-failure** ✓:

- $47,774 \times \hat{p}_{pool} \geq 10$
- $47,774 \times (1 - \hat{p}_{pool}) \geq 10$
- $5,855 \times \hat{p}_{pool} \geq 10$
- $5,855 \times (1 - \hat{p}_{pool}) \geq 10$

We calculate the standard error as $\sigma_{\overline{x}} = \sqrt{\frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{47,774} + \frac{\hat{p}_{pool}(1-\hat{p}_{pool})}{5855}} \approx 0.0017$. The p-value is then found using `2*pnorm(0.015 - 0.025, 0, sd = 0.0017)` which returns a value of approximately 0.




(c) Statistically significant means that test concludes there is a difference between two groups. Practically significant means that there is a difference that we would care about in the real world between two groups. The rate of diabetes is still very low for both groups. Although the test detected a difference, it may not be a difference we care about.

3. The mean number of sick days an employee takes per year is believed to be about ten. Members of a personnel department do not believe this figure. They randomly survey eight employees. The number of sick days they took for the past year are as follows: 12; 4; 15; 3; 11; 8; 6; 8. Let $x$ = the number of sick days they took for the past year. Should the personnel team believe that the mean number is ten?

   (a) Write the hypotheses in symbols.

   (b) Check conditions.

   (c) Find the p-value.

   (d) Calculate a 90% CI and interpret.

   (e) What is the conclusion of the hypothesis test using $\alpha = .10$?

(a) $H_0 : \mu = 10$ $\qquad H_A : \mu \neq 10$

(b) **Independence** ✓: We have no reason to suspect dependence between subjects (subjects were not sampled more than once, and cannot force each other to take sick days). **Normality** ✓: There are no particularity outrageous values. The data seems reasonably symmetric.

(c) We use the 't.test()' function to produce the p-value and the confidence interval for parts c and d.

```
> t.test(c(12, 4, 15, 3, 11, 8, 6,8),
+        mu = 10,
+        alternative = "two.sided",
+        conf.level = .90)

        One Sample t-test

data:  c(12, 4, 15, 3, 11, 8, 6, 8)
t = -1.12, df = 7, p-value = 0.2996
alternative hypothesis: true mean is not equal to 10
90 percent confidence interval:
  5.626287 11.123713
sample estimates:
mean of x
    8.375
```

Our p-value is 0.2296.

(d) We are 90% confident that the interval [5.6263, 11.1237] captures the true mean number of sick days an employee takes a year.

(e) Our p-value=0.2296 is greater than $\alpha = 0.10$. We conclude in favor of $H_0$. It seems plausible that the true average number of sick days an employee takes a year is 10.

4. Suppose a statistics instructor believes that there is no significant difference between the mean class scores for the students that take the course during the day (day students) and night (night students) on Exam 2. She takes random samples from each of the populations. The mean and standard deviation for 35 day students were 75.86 and 16.91, respectively. The mean and standard deviation for 37 night students were 75.41 and 19.73. Assume the data from the two populations is roughly symmetric.

The hypotheses are

$$H_0 : \mu_D - \mu_N = 0$$

$$H_A : \mu_D - \mu_N \neq 0$$

. We next need to check the conditions.

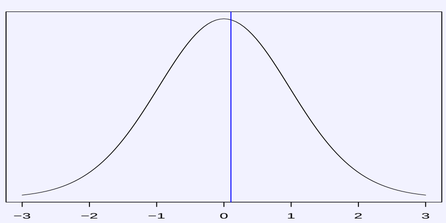
- **Independence** ✓: Within groups independence is satistfied, each student is unique and received individual scores. Between groups is satisfied, students can only belong to one class at a time.
- **Normality** ✓: The prompt told us that data from each group is approximately normal.

Now we need to calculate the test statistic, and related values.

$$SE = \sqrt{\frac{\sigma_D^2}{n_D} + \frac{\sigma_N^2}{n_N}} = \sqrt{\frac{16.91^2}{35} + \frac{19.73}{37}} \approx 4.3233$$

$$T = \frac{\overline{x}_D - \overline{x}_N}{SE} \approx \frac{75.86 - 75.41}{4.3233} \approx 0.1041$$

The degrees of freedom is $df = \min(n_D - 1, n_N - 1) = 34$. With these values we can calculate the p-value using `2*(1-pt(0.1041, df= 34))`, which would return p-value = 0.9177.




Our p-value = 0.9177 is greater than $\alpha = 0.05$. We conclude in favor of $H_0$. It seems plausible that the mean class scores between day and night students are the same, and there is no significant difference between the two groups.

5. A national survey conducted among a simple random sample of 1507 adults shows that 56% of Americans think the Civil War is still relevant to American politics and political life. Conduct a hypothesis test to determine if these data provide strong evidence that the majority of the Americans think the Civil War is still relevant.

We want to access if the data provides evidence that the majority of Americans think the Civil War is relevant. Thus, we set up our hypotheses as:

$$H_0 : p \leq 0.5$$

$$H_A : p > 0.5$$

Essentially, we are checking if it is at all possible that the claim is not true. Do we have strong evidence to reject $H_0$? We first need to check our conditions.

- **Indepdence** ✓: It is given that we have a simple random sample, which implies independence.
- **Success-Failure** ✓: $1507 \times 0.5 \geq 10$ (since $p_0 = 0.5 = 1 - p_0$ we only check one value)

We now need the standard error under $H_0$, which is $\sqrt{\frac{0.5 \times 0.5}{1507}} \approx 0.0129$. We can calculate our p-value using `1-pnorm(0.56, mean = 0.5, sd = 0.0129)`, which gives us a value approximately 0. We conclude against $H_0$. It seems likely that the majority of Americans think the Civil War is still relevant to American politics.

6. A 90% confidence interval for a population mean is (65, 77). The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 25 observations. Calculate the sample mean and the standard error. Assume that all conditions necessary for inference are satisfied. Use the $t$-distribution in any calculations.

The sample mean is the mid-point of the confidence interval, i.e., the average of the upper and lower bounds:

$$\overline{x} = \frac{65 + 77}{2} = 71$$

Since $n$=25, df=24, and the critical t-score is $t^*_{df} = 1.71$. Then, $6 = 1.71 \frac{s_x}{25} \rightarrow s_x \approx 17.54$.