Learning Goals:

- Under appropriate conditions, conduct a hypothesis test about a mean for a matched pairs design. State a conclusion in context.
- Interpret the P-value as a conditional probability.

Introduction:

In this activity we will again be testing a hypothesis about a population mean, but this time we will have data from a matched pairs design. For our purposes, *matched pair* means that we have two measurements for each individual, such as a pre- and post-exam score for a set of students.

It can also mean that two groups of subjects are matched based on demographic characteristics. One person in each pair is exposed to a treatment and the other person is not. We will not work with this type of matched pairs design in this course, but it is handled the same way.

A matched pairs model is really a straightforward extension of what we already have done. The new population is simply the difference in measurements for each individual. For example, $x_{post} - x_{pre}$ in the pre- and post-exam scenario. And the mean of the population of differences is indicated by the new notation μ_d, read "mu sub d." Here, *d* stands for difference. The analysis itself just uses the differences as the raw data.

Example:

The following post was pulled from allnurses.com

"Hey everyone,
So I am in midst of an argument at work between a CNA and a new grad RN over which temp is more accurate Oral or Tympanic [ear] now I know different variables come into play using both i.e. with oral the patient may have drank something cold or hot or with tympanic laying on one side on pillow can cause false highs etc. etc., so lets assume no such variables are in play - which temp would be most accurate?"

Source: http://allnurses.com/general-nursing-discussion/which-temp-is-615783.html

Suppose that we have 12 randomly selected patients and a nurse measures each patient's oral temperature and then tympanic (ear) temperature.

Here are the (hypothetical) results: Complete the table.

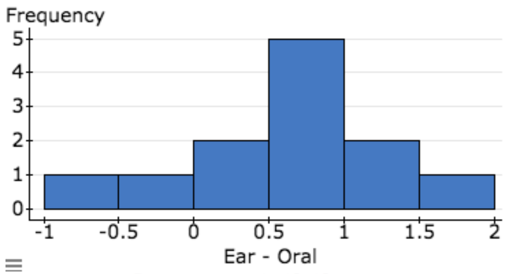
	1	2	3	4	5	6	7	8	9	10	11	12
Oral (°F)	98.4	97	96.5	97.1	98.9	98.2	96.8	99	97.9	100	99.6	98.7
Ear (°F)	98.9	98.2	96	97.8	99	98.7	98.6	98	98.6	101	100	99.6
Difference (Ear-Oral)	0.5	1.2	-0.5	0.7	0.1	0.5	1.8	-1	0.7			

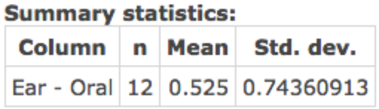
Let’s test the claim that oral and tympanic temperatures are different using a 5% significance level.

State the hypotheses:

Collect the data:

Here is a histogram of the 12 differences.




Summary statistics:

Column	n	Mean	Std. dev.
Ear - Oral	12	0.525	0.74360913

- Draw an arrow to locate Patient #2 in the histogram.

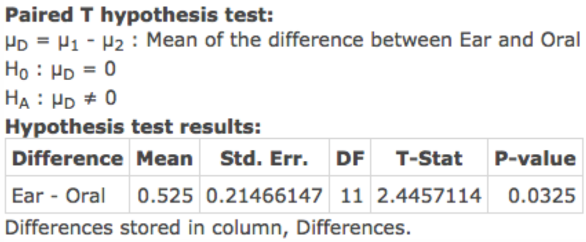
- What does a positive difference indicate?

- What does the mean of 0.525 represent?

Assess the data:

- Verify that the conditions for use of the T-model are met.

- Verify that the T-score is about 2.45 as shown in the StatCrunch print-out.


Paired T hypothesis test:

$\mu_D = \mu_1 - \mu_2$: Mean of the difference between Ear and Oral

$H_0 : \mu_D = 0$

$H_A : \mu_D \neq 0$

Hypothesis test results:

Difference	Mean	Std. Err.	DF	T-Stat	P-value
Ear - Oral	0.525	0.21466147	11	2.4457114	0.0325

Differences stored in column, Differences.

- What is the P-value?

State a conclusion:

Group work:

1) Determining the alternative hypothesis requires some careful thinking with a matched pairs design. For each of the following situations, state the alternative hypothesis as $\mu_d > 0, \mu_d < 0, or\ \mu_d \neq 0$. Briefly explain your reasoning.

   a) In a study of a drug designed to reduce cholesterol levels, patients take the drug for six weeks. Researchers record the cholesterol levels of patients before and after taking the drug. The difference (before minus after) is used to test the claim that the drug reduces cholesterol levels.

   b) In a study of a treatment designed to increase red blood cell counts, researchers use the difference in red blood cell count (before minus after) to test the claim that the treatment has the desired effect.

   c) In an automobile safety study, researchers use a matched pairs design to examine the effect of cruise control on reaction time to brake. They test the claim that the use of cruise control is less safe because it increases reaction time to brake. For the T-test the difference in reaction times is defined as "cruise control" minus "no cruise control."

2) Students from a statistics class perform an experiment to assess the effect of music on concentration. They measure "concentration" by recording the time it takes a student to complete a simple word puzzle. Each student does a word puzzle while listening to classical music and in silence. A coin flip determines the order of treatments.

They test the claim that listening to classical music will improve concentration; therefore, the time for completing a word puzzle will be faster with music (a shorter completion time) and the mean of the difference in completion times ("no music" minus "music") will be positive. $(H_0\colon \mu_d = 0,\ H_a\colon \mu_d > 0)$.

They use a sample of 15 student volunteers; therefore, the sample is not randomly chosen, but the treatments are randomly ordered, which allows them to conduct a hypothesis test if the conditions are met. The distribution of the difference in puzzle completion times is not strongly skewed and there are no outliers, so the conditions are met for use of a T-test.

For a sample mean of 3 seconds and a standard deviation of 5 seconds, the P-value is approximately 0.024.

What can we conclude? Explain why you chose, or do not choose, each option.

a) For the population of college students, this study suggests that listening to classical music produces a 3-second improvement in puzzle completion time.

b) Classical music is associated with statistically significant improvements in concentration as measured by time to complete a simple word puzzle.

c) College students listening to classical music completed word puzzles significantly faster than students not listening to music.

d) For the population of college students, this study provides fairly strong evidence that listening to classical music improves word puzzle completion time, which in turn suggests that listening to classical music may improve concentration.

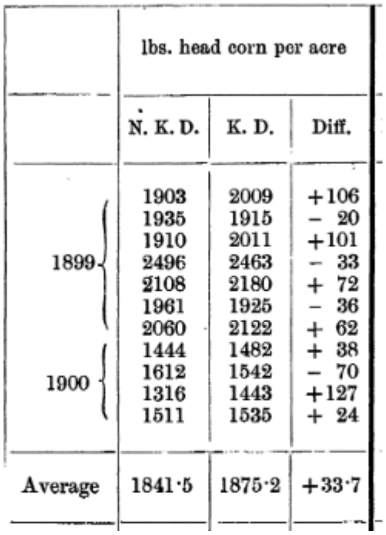
3) William S. Gosset invented the T-model when he was an employee of the Guinness Brewing Company in Dublin. The Guinness Co. was interested in increasing grain yields to lower the cost of brewing beer. Samples from these experiments were often small and statistical methods of the day were inadequate for analyzing small data sets.

In his famous 1908 paper titled "The Probable Error of a Mean," Gossett illustrates his T-model using the results of an experiment published by Dr. Voelcker in the *Journal of Agricultural Society*. Voelcker's experiment was designed to determine if kiln-dried seed produced greater yields. In this experiment 11 different plots of land were planted with two different types of seed: regular or kiln-dried.

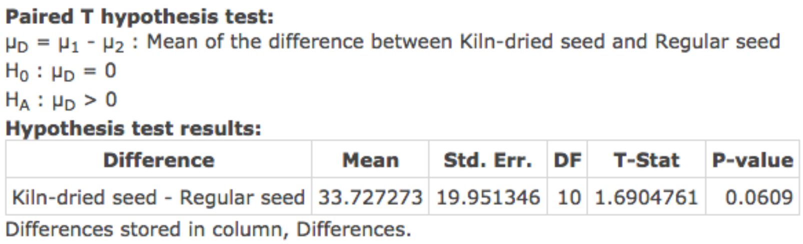
Grain grown in different plots may experience different growing conditions, such as differences in soil fertility or amount of light. To reduce the potential influence of these confounding variables, the experiment used a matched pairs design and both types of seed were grown in each of the 11 plots.

The table shows the data from this experiment as reported in Gossett's paper. The variable is corn yield in pounds per acre.

We used the data from Gossett's paper to run a matched pairs T-test.



	lbs. head corn per acre		
	N. K. D.	K. D.	Diff.
1899	1903	2009	+106
	1935	1915	− 20
	1910	2011	+101
	2496	2463	− 33
	2108	2180	+ 72
	1961	1925	− 36
	2060	2122	+ 62
1900	1444	1482	+ 38
	1612	1542	− 70
	1316	1443	+127
	1511	1535	+ 24
Average	1841·5	1875·2	+33·7



Paired T hypothesis test:

$\mu_D = \mu_1 - \mu_2$: Mean of the difference between Kiln-dried seed and Regular seed

$H_0 : \mu_D = 0$

$H_A : \mu_D > 0$

Hypothesis test results:

Difference	Mean	Std. Err.	DF	T-Stat	P-value
Kiln-dried seed - Regular seed	33.727273	19.951346	10	1.6904761	0.0609

Differences stored in column, Differences.

a) What does H_0: $\mu_d = 0$ mean in the context of corn yields and seed types?

b) Explain why the alternative hypothesis is $\mu_d > 0$ instead of $\mu_d < 0$.

c) The standard error is about 20. What are the units for the standard error? What does this number tell us?

d) What does the T-stat of 1.69 tell us in this context?

e) In his paper Gossett quotes Voelcker's conclusion, "In such seasons as 1899 and 1900 there is no particular advantage to kiln-drying before sowing." He adds that his own examination justifies this conclusion. Our T-test is consistent with this conclusion at a 5% level of significance.

Of course, Voelcker did not perform any statistical inference procedures. After all, statistical inference procedures for analyzing small data sets did not exist, which is why Gossett wrote his paper.

Voelcker just observed that the kiln-dried seeds produced 33.7 more pounds of corn per acre than the regular seeds. Why do statisticians want to run a hypothesis test instead of just relying on descriptive summaries of the data?

Historical note excerpted from Wikipedia: Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery. To prevent further disclosure of confidential information, Guinness prohibited its employees from publishing any papers regardless of the contained information. However, after pleading with the brewery and explaining that his mathematical and philosophical conclusions were of no possible practical use to competing brewers, he was allowed to publish them, but under a pseudonym ("Student"), to avoid difficulties with the rest of the staff. Thus his most noteworthy achievement is now called Student's, rather than Gosset's, t-distribution. Source: https://en.wikipedia.org/wiki/William_Sealy_Gosset