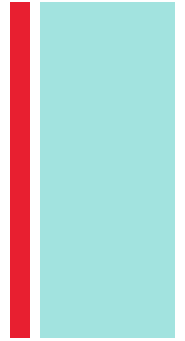


## Unit 5: Hypothesis Testing

The Practice of Statistics, 4<sup>th</sup> edition – For AP\*  
STARNES, YATES, MOORE

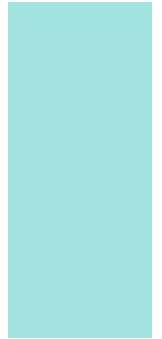
# + Unit 5: Hypothesis Testing



- 9.1 Significance Tests: The Basics
- 9.2 Tests about a Population Proportion
- **9.3 Tests about a Population Mean – Day 2**
- 9.1 & 9.2 Errors and the Power of a Test



## Section 9.3 Tests About a Population Mean



### Learning Objectives

After this section, you should be able to...

- ✓ PERFORM significance tests for paired data.



## ■ Inference for Means: Paired Data

Comparative studies are more convincing than single-sample investigations. For that reason, one-sample inference is less common than comparative inference. Study designs that involve making two observations on the same individual, or one observation on each of two similar individuals, result in **paired data**.

When paired data result from measuring the same quantitative variable twice, as in the job satisfaction study, we can make comparisons by analyzing the differences in each pair. If the conditions for inference are met, we can use one-sample  $t$  procedures to perform inference about the mean difference  $\mu_d$ .

These methods are sometimes called **paired  $t$  procedures**.

## ■ Paired $t$ Test

Researchers designed an experiment to study the effects of caffeine withdrawal. They recruited 11 volunteers who were diagnosed as being caffeine dependent to serve as subjects. Each subject was barred from coffee, colas, and other substances with caffeine for the duration of the experiment. During one two-day period, subjects took capsules containing their normal caffeine intake. During another two-day period, they took placebo capsules. The order in which subjects took caffeine and the placebo was randomized. At the end of each two-day period, a test for depression was given to all 11 subjects. Researchers wanted to know whether being deprived of caffeine would lead to an increase in depression.

Results of a caffeine deprivation study			
Subject	Depression (caffeine)	Depression (placebo)	Difference (placebo – caffeine)
1	5	16	11
2	5	23	18
3	4	5	1
4	3	7	4
5	8	14	6
6	5	24	19
7	0	6	6
8	0	3	3
9	2	15	13
10	11	12	1
11	1	0	- 1

**State:** If caffeine deprivation has no effect on depression, then we would expect the actual mean difference in depression scores to be 0. We want to test the hypotheses

$$H_0: \mu_d = 0$$

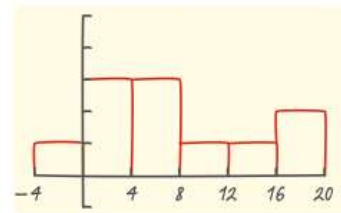
$$H_a: \mu_d > 0$$

where  $\mu_d$  = the true mean difference (placebo – caffeine) in depression score. Since no significance level is given, we'll use  $\alpha = 0.05$ .

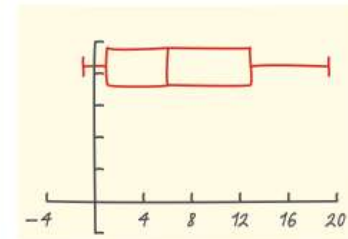
## ■ Paired $t$ Test

**Plan:** If conditions are met, we should do a paired  $t$  test for  $\mu_d$ .

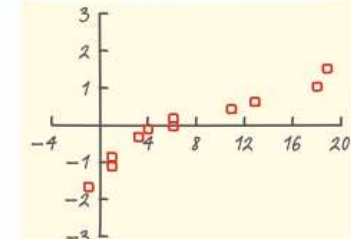
- ✓ *Random* researchers randomly assigned the treatment order—placebo then caffeine, caffeine then placebo—to the subjects.
- ✓ *Normal* We don't know whether the actual distribution of difference in depression scores (placebo - caffeine) is Normal. With such a small sample size ( $n = 11$ ), we need to examine the data to see if it's safe to use  $t$  procedures.



Change in depression  
(placebo - caffeine)



Change in depression  
(placebo - caffeine)



Change in depression  
(placebo - caffeine)

The histogram has an irregular shape with so few values; the boxplot shows some right-skewness but not outliers; and the Normal probability plot looks fairly linear. With no outliers or strong skewness, the  $t$  procedures should be pretty accurate.

- ✓ *Independent* We aren't sampling, so it isn't necessary to check the *10% condition*. We will assume that the changes in depression scores for individual subjects are independent. This is reasonable if the experiment is conducted properly.

## ■ Paired $t$ Test

**Do:** The sample mean and standard deviation are  $\bar{x}_d = 7.364$  and  $s_d = 6.918$

**Test statistic** 
$$t = \frac{\bar{x}_d - \mu_0}{s_d / \sqrt{n}} = \frac{7.364 - 0}{6.918 / \sqrt{11}} = 3.53$$

**P-value** According to technology, the area to the right of  $t = 3.53$  on the  $t$  distribution curve with  $df = 11 - 1 = 10$  is 0.0027.

**Conclude:** With a  $P$ -value of 0.0027, which is much less than our chosen  $\alpha = 0.05$ , we have convincing evidence to reject  $H_0: \mu_d = 0$ . We can therefore conclude that depriving these caffeine-dependent subjects of caffeine caused an average increase in depression scores.



## ■ Using Tests Wisely

Significance tests are widely used in reporting the results of research in many fields. New drugs require significant evidence of effectiveness and safety. Courts ask about statistical significance in hearing discrimination cases. Marketers want to know whether a new ad campaign significantly outperforms the old one, and medical researchers want to know whether a new therapy performs significantly better. In all these uses, statistical significance is valued because it points to an effect that is unlikely to occur simply by chance.

Carrying out a significance test is often quite simple, especially if you use a calculator or computer. Using tests wisely is not so simple. Here are some points to keep in mind when using or interpreting significance tests.

### **Statistical Significance and Practical Importance**

When a null hypothesis (“no effect” or “no difference”) can be rejected at the usual levels ( $\alpha = 0.05$  or  $\alpha = 0.01$ ), there is good evidence of a difference. But that difference may be very small. When large samples are available, even tiny deviations from the null hypothesis will be significant.



## ■ Using Tests Wisely

### **Don't Ignore Lack of Significance**

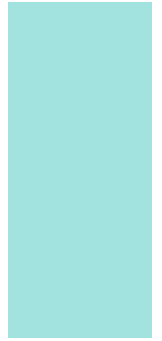
There is a tendency to infer that there is no difference whenever a  $P$ -value fails to attain the usual 5% standard. In some areas of research, small differences that are detectable only with large sample sizes can be of great practical significance. When planning a study, verify that the test you plan to use has a high probability (power) of detecting a difference of the size you hope to find.

### **Statistical Inference Is Not Valid for All Sets of Data**

Badly designed surveys or experiments often produce invalid results. Formal statistical inference cannot correct basic flaws in the design. Each test is valid only in certain circumstances, with properly produced data being particularly important.

### **Beware of Multiple Analyses**

Statistical significance ought to mean that you have found a difference that you were looking for. The reasoning behind statistical significance works well if you decide what difference you are seeking, design a study to search for it, and use a significance test to weigh the evidence you get. In other settings, significance may have little meaning.



## **Homework**

Chapter 9, #'s 86, 87a-c (write a full significance test for both)