Ch 9.3 Significant Testing with Paired Data

Inference for Means: Paired Data

Study designs that involve making two observations on the same individual, or one observation on each of two similar individuals, yield **paired data.** When paired data result from measuring the same quantitative variable twice, 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 μ_d. (These methods are sometimes called **paired *t* procedures**).

Jan 27-12:19 PM

Example 7: 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 2-day period, subjects took capsules containing their normal caffeine intake. During another 2-day period, they took placebo capsules. The order in which subjects took caffeine and the placebo was randomized. At the end of each 2-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. The table on the next slide contains data on the subjects' scores on the depression test. Higher scores show more symptoms of depression. For each subject, we calculated the difference in test scores following each of the two treatments (placebo – caffeine). We chose this order of subtraction to get mostly positive values.

Jan 27-12:27 PM

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

Jan 27-12:26 PM

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 μ_d = the true mean difference (placebo – caffeine) in depression score for subjects like these. Because no significance level is given, we'll use $\alpha = 0.05$.

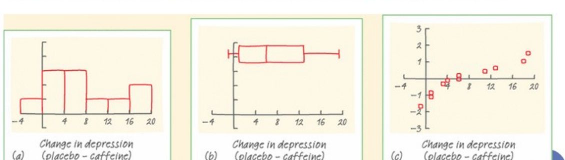
Plan: If conditions are met, we should do a paired *t* test for μ_d.

✓**Random:** Researchers randomly assigned the treatment order—placebo then caffeine, caffeine then placebo—to the subjects.

✓**10%:** We aren't sampling, so it isn't necessary to check the 10% condition.

Jan 27-12:20 PM

✓**Normal:** We don't know whether the actual distribution of difference in depression scores (placebo–caffeine) for subjects like these is Normal. With such a small sample size ($n = 11$), we need to graph the data to see if it's safe to use t procedures. The figure below shows hand sketches of a calculator histogram, boxplot, and Normal probability plot for these data. The histogram has an irregular shape with so few values; **the boxplot shows some right skewness but no outliers**; and the Normal probability plot is slightly curved, indicating mild skewness. **With no outliers or strong skewness, the *t* procedures should be fairly accurate.**

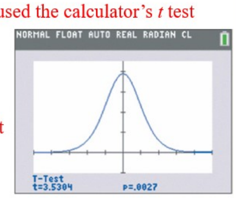


Jan 27-12:23 PM

Calc a T-test for List 3

Do: Test Statistic: $t = \dfrac{\bar{x}_d - \mu_0}{\frac{s_d}{\sqrt{n}}} = \dfrac{7.364 - 0}{\frac{6.918}{\sqrt{11}}} = 3.53$

We entered the differences in list1 and then used the calculator's *t* test command with the "Draw" option.



Test statistic $t = 3.53$

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

Note: The calculator doesn't report the degrees of freedom, but you should.

Conclude: Reject H_0. Since the *P*-value of 0.0027, is less than our chosen $\alpha = 0.05$, we have convincing evidence to suggest that the true mean difference (placebo − caffeine) in depression score is positive for subjects like these.

Jan 27-12:23 PM

Be sure to report the degrees of freedom with any t procedure, even if technology doesn't.

2. The subjects in this experiment were *not* chosen at random from the population of caffeine-dependent individuals. As a result, we can't generalize our findings to *all* caffeine-dependent people—only to people like the ones who took part in this experiment.

3. Because researchers randomly assigned the treatments, they can make an inference about cause and effect. The data from this experiment provide convincing evidence that depriving caffeine-dependent subjects like these of caffeine causes an average increase in depression scores.

Jan 27-12:27 PM

Try:

air - nitrogen = +#
difference > 0

Refer to the Data Exploration from Chapter 4 on page 257. Do the data give convincing evidence at the $\alpha = 0.05$ significance level that filling tires with nitrogen instead of air decreases pressure loss?

DATA EXPLORATION Nitrogen in tires—a lot of hot air?

Most automobile tires are inflated with compressed air, which consists of about 78% nitrogen. Aircraft tires are filled with pure nitrogen, which is safer than air in case of fire. Could filling automobile tires with nitrogen improve safety, performance, or both?

Consumers Union designed a study to test whether nitrogen-filled tires would maintain pressure better than air-filled tires. They obtained two tires from each of several brands and then filled one tire in each pair with air and one with nitrogen. All tires were inflated to a pressure of 30 pounds per square inch and then placed outside for a year. At the end of the year, Consumers Union measured the pressure in each tire. The amount of pressure lost (in pounds per square inch) during the year for the air-filled and nitrogen-filled tires of each brand is shown in the table below.[30]

Brand	Air	Nitrogen	Brand	Air	Nitrogen
BF Goodrich Traction T/A HR	7.6	7.2	Pirelli P6 Four Seasons	4.4	4.2
Bridgestone HP50 (Sears)	3.8	2.5	Sumitomo HTR H4	1.4	2.1
Bridgestone Potenza G009	3.7	1.6	Yokohama Avid H4S	4.3	3.0
Bridgestone Potenza RE950	4.7	1.5	BF Goodrich Traction T/A V	5.5	3.4
Bridgestone Potenza EL400	2.1	1.0	Bridgestone Potenza RE950	4.1	2.8
Continental Premier Contact H	4.9	3.1	Continental ContiExtreme Contact	5.0	3.4
Cooper Lifeliner Touring SLE	5.2	3.5	Continental ContiProContact	4.8	3.3
Dayton Daytona HR	3.4	3.2	Cooper Lifeliner Touring SLE	3.2	2.5
Falken Ziex ZE-512	4.1	3.3	General Exclaim UHP	6.8	2.7
Fuzion Hrl	2.7	2.2	Hankook Ventus V4 H105	3.1	1.4
General Exclaim	3.1	3.4	Michelin Energy MXV4 Plus	2.5	1.5
Goodyear Assurance Tripletred	3.8	3.2	Michelin Pilot Exalto A/S	6.6	2.2
Hankook Optimo H418	3.0	0.9	Michelin Pilot HX MXM4	2.2	2.0
Kumho Solus KH16	6.2	3.4	Pirelli P6 Four Seasons	2.5	2.7
Michelin Energy MXV4 Plus	2.0	1.8	Sumitomo HTR+	4.4	3.7
Michelin Pilot XGT H4	1.1	0.7			

Jan 27-12:58 PM

Correct Answer
S: $H_0: \mu_d = 0$ versus $H_a: \mu_d > 0$, where μ_d is the true mean difference (air – nitrogen) in pressure lost. *P*: Paired *t* test for μ_d. Random: Treatments were assigned at random to each pair of tires. Normal/Large Sample: $n = 31 \geq 30$. *D*: $\bar{x} = 1.252$ and $s_x = 1.202$. $t = 5.80$ and *P*-value ≈ 0. *C*: Because the *P*-value of approximately $0 < \alpha = 0.05$, we reject H_0. We have convincing evidence that the true mean difference in pressure (air – nitrogen) > 0. In other words, we have convincing evidence that tires lose less pressure when filled with nitrogen than when filled with air, on average.

Jan 27-1:04 PM

Using Tests Wisely

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 convincing 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.

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.

Jan 27-12:55 PM

Example 10: Might the radiation from cell phones be harmful to users? Many studies have found little or no connection between using cell phones and various illnesses. Here is part of a news account of one study:

A hospital study that compared brain cancer patients and a similar group without brain cancer found no statistically significant difference between brain cancer rates for the two groups. But when 20 distinct types of brain cancer were considered separately, a significant difference in brain cancer rates was found for one rare type. Puzzlingly, however, this risk appeared to decrease rather than increase with greater mobile phone use.

Think for a moment. Suppose that the 20 null hypotheses for these 20 significance tests are all true. Then each test has a 5% chance of being significant at the 5% level. That's what $\alpha = 0.05$ means: results this extreme occur only 5% of the time just by chance when the null hypothesis is true. We expect about 1 of 20 tests to give a significant result just by chance. Running one test and reaching the $\alpha = 0.05$ level is reasonably good evidence that you have found something; running 20 tests and reaching that level only once is not.

Jan 27-12:31 PM