

Paired Samples and Blocks



WHO Olympic speed-skaters

WHAT Time for women's 1500 m

UNITS Seconds

WHEN 2006

WHERE Torino, Italy

WHY To see whether one lane is faster than the other

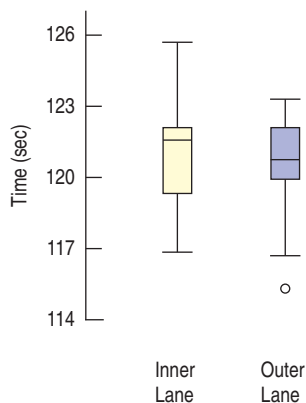


FIGURE 25.1

Using boxplots to compare times in the inner and outer lanes shows little because it ignores the fact that the skaters raced in pairs.

Speed-skating races are run in pairs. Two skaters start at the same time, one on the inner lane and one on the outer lane. Halfway through the race, they cross over, switching lanes so that each will skate the same distance in each lane. Even though this seems fair, at the 2006 Olympics some fans thought there might have been an advantage to starting on the outside. After all, the winner, Cindy Klassen, started on the outside and skated a remarkable 1.47 seconds faster than the silver medalist.

Here are the data for the women's 1500-m race:

Inner Lane		Outer Lane	
Name	Time	Name	Time
OLTEAN Daniela	129.24	(no competitor)	
ZHANG Xiaolei	125.75	NEMOTO Nami	122.34
ABRAMOVA Yekaterina	121.63	LAMB Maria	122.12
REMPER Shannon	122.24	NOH Seon Yeong	123.35
LEE Ju-Youn	120.85	TIMMER Marianne	120.45
ROKITA Anna Natalia	122.19	MARRA Adelia	123.07
YAKSHINA Valentina	122.15	OPITZ Lucille	122.75
BJELKEVIK Hedvig	122.16	HAUGLI Maren	121.22
ISHINO Eriko	121.85	WOJCICKA Katarzyna	119.96
RANEY Catherine	121.17	BJELKEVIK Annette	121.03
OTSU Hiromi	124.77	LOBYSHEVA Yekaterina	118.87
SIMIONATO Chiara	118.76	JI Jia	121.85
ANSCHUETZ THOMS Daniela	119.74	WANG Fei	120.13
BARYSHEVA Varvara	121.60	van DEUTEKOM Paulien	120.15
GROENEWOLD Renate	119.33	GROVES Kristina	116.74
RODRIGUEZ Jennifer	119.30	NESBITT Christine	119.15
FRIESINGER Anni	117.31	KLASSEN Cindy	115.27
WUST Ireen	116.90	TABATA Maki	120.77

We can view this skating event as an experiment testing whether the lanes were equally fast. Skaters were assigned to lanes randomly. The boxplots of times recorded in the inner and outer lanes (look back a page) don't show much difference. But that's not the right way to compare these times. Conditions can change during the day. The data are recorded for races run two at a time, so the two groups are not independent.

Paired Data

Data such as these are called **paired**. We have the times for skaters in each lane for each race. The races are run in pairs, so they can't be independent. And since they're not independent, we can't use the two-sample t methods. Instead, we can focus on the *differences* in times for each racing pair.

Paired data arise in a number of ways. Perhaps the most common way is to compare subjects with themselves before and after a treatment. When pairs arise from an experiment, the pairing is a type of *blocking*. When they arise from an observational study, it is a form of *matching*.

FOR EXAMPLE

Identifying paired data

Do flexible schedules reduce the demand for resources? The Lake County, Illinois, Health Department experimented with a flexible four-day workweek. For a year, the department recorded the mileage driven by 11 field workers on an ordinary five-day workweek. Then it changed to a flexible four-day workweek and recorded mileage for another year.¹ The data are shown.

Question: Why are these data paired?

The mileage data are paired because each driver's mileage is measured before and after the change in schedule. I'd expect drivers who drove more than others before the schedule change to continue to drive more afterwards, so the two sets of mileages can't be considered independent.

Name	5-Day mileage	4-Day mileage
Jeff	2798	2914
Betty	7724	6112
Roger	7505	6177
Tom	838	1102
Aimee	4592	3281
Greg	8107	4997
Larry G.	1228	1695
Tad	8718	6606
Larry M.	1097	1063
Leslie	8089	6392
Lee	3807	3362

Pairing isn't a problem; it's an opportunity. If you know the data are paired, you can take advantage of that fact—in fact, you *must* take advantage of it. You *may not* use the two-sample and pooled methods of the previous chapter when the data are paired. Remember: Those methods rely on the Pythagorean Theorem of Statistics, and that requires the two samples be independent. Paired data aren't. There is no test to determine whether the data are paired. You must determine that from understanding how they were collected and what they mean (check the *W*'s).

Once we recognize that the speed-skating data are matched pairs, it makes sense to consider the difference in times for each two-skater race. So we look at the *pairwise* differences:

¹ Charles S. Catlin, "Four-day Work Week Improves Environment," *Journal of Environmental Health*, Denver, 59:7.

AS **Activity: Differences in Means of Paired Groups.** Are married couples typically the same age, or do wives tend to be younger than their husbands, on average?

Skating Pair	Inner Time	Outer Time	Inner – Outer
1	129.24		.
2	125.75	122.34	3.41
3	121.63	122.12	-0.49
4	122.24	123.35	-1.11
5	120.85	120.45	0.40
6	122.19	123.07	-0.88
7	122.15	122.75	-0.60
8	122.16	121.22	0.94
9	121.85	119.96	1.89
10	121.17	121.03	0.14
11	124.77	118.87	5.90
12	118.76	121.85	-3.09
13	119.74	120.13	-0.39
14	121.60	120.15	1.45
15	119.33	116.74	2.59
16	119.30	119.15	0.15
17	117.31	115.27	2.04
18	116.90	120.77	-3.87

The first skater raced alone, so we'll omit that race. Because it is the *differences* we care about, we'll treat them as if *they* were the data, ignoring the original two columns. Now that we have only one column of values to consider, we can use a simple one-sample *t*-test. Mechanically, a **paired *t*-test** is just a one-sample *t*-test for the means of these pairwise differences. The sample size is the number of pairs.

So you've already seen the *Show*.

Assumptions and Conditions



PAIRED DATA ASSUMPTION

Paired Data Assumption: The data must be paired. You can't just decide to pair data when in fact the samples are independent. When you have two groups with the same number of observations, it may be tempting to match them up.

Don't, unless you are prepared to justify your claim that the data are paired.

On the other hand, be sure to recognize paired data when you have them. Remember, two-sample *t* methods aren't valid without independent groups, and paired groups aren't independent. Although this is a strictly required assumption, it is one that can be easy to check if you understand how the data were collected.

INDEPENDENCE ASSUMPTION

Independence Assumption: If the data are paired, the *groups* are not independent. For these methods, it's the *differences* that must be independent of each other. There's no reason to believe that the difference in speeds of one pair of races could affect the difference in speeds for another pair.

Randomization Condition: Randomness can arise in many ways. The pairs may be a random sample. In an experiment, the order of the two treatments may be randomly assigned, or the treatments may be randomly assigned to one member of each pair. In a before-and-after study, we may believe that the observed differences are a representative sample from a population of interest. If we have any doubts, we'll need to include a control group to be able to draw conclusions.

10% of what?

A fringe benefit of checking the 10% Condition is that it forces us to think about what population we're hoping to make inferences about.

What we want to know usually focuses our attention on where the randomness should be.

In our example, skaters were assigned to the lanes at random.

10% Condition: We're thinking of the speed-skating data as an experiment testing the difference between lanes. The 10% Condition doesn't apply to randomized experiments, where no sampling takes place.

NORMAL POPULATION ASSUMPTION

We need to assume that the population of *differences* follows a Normal model. We don't need to check the individual groups.

Nearly Normal Condition: This condition can be checked with a histogram or Normal probability plot of the *differences*—but not of the individual groups. As with the one-sample *t*-methods, this assumption matters less the more pairs we have to consider. You may be pleasantly surprised when you check this condition. Even if your original measurements are skewed or bimodal, the *differences* may be nearly Normal. After all, the individual who was way out in the tail on an initial measurement is likely to still be out there on the second one, giving a perfectly ordinary difference.

FOR EXAMPLE**Checking assumptions and conditions**

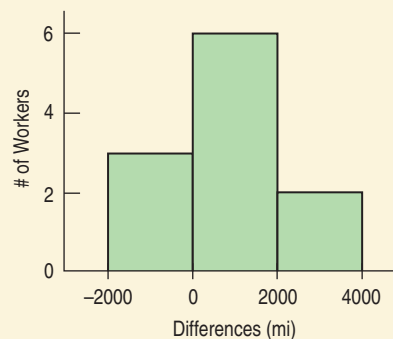
Recap: Field workers for a health department compared driving mileage on a five-day work schedule with mileage on a new four-day schedule. To see if the new schedule changed the amount of driving they did, we'll look at paired differences in mileages before and after.

Question: Is it okay to use these data to test whether the new schedule changed the amount of driving?

- ✓ **Paired Data Assumption:** The data are paired because each value is the mileage driven by the same person before and after a change in work schedule.
- ✓ **Independence Assumption:** The driving behavior of any individual worker is independent of the others, so the differences are mutually independent.
- ✓ **Randomization Condition:** The mileages are the sums of many individual trips, each of which experienced random events that arose while driving. Repeating the experiment in two new years would give randomly different values.
- ✓ **Nearly Normal Condition:** The histogram of the mileage differences is unimodal and symmetric:

Since the assumptions and conditions are satisfied, it's okay to use paired-*t* methods for these data.

Name	5-Day mileage	4-Day mileage	Difference
Jeff	2798	2914	-116
Betty	7724	6112	1612
Roger	7505	6177	1328
Tom	838	1102	-264
Aimee	4592	3281	1311
Greg	8107	4997	3110
Larry G.	1228	1695	-467
Tad	8718	6606	2112
Larry M.	1097	1063	34
Leslie	8089	6392	1697
Lee	3807	3362	445



The steps in testing a hypothesis for paired differences are very much like the steps for a one-sample *t*-test for a mean.

THE PAIRED t -TEST

When the conditions are met, we are ready to test whether the mean of paired differences is significantly different from zero. We test the hypothesis

$$H_0: \mu_d = \Delta_0,$$

where the d 's are the pairwise differences and Δ_0 is almost always 0.

We use the statistic

$$t_{n-1} = \frac{\bar{d} - \Delta_0}{SE(\bar{d})},$$

where \bar{d} is the mean of the pairwise differences, n is the number of *pairs*, and

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}}.$$

$SE(\bar{d})$ is the ordinary standard error for the mean, applied to the differences.

When the conditions are met and the null hypothesis is true, we can model the sampling distribution of this statistic with a Student's t -model with $n - 1$ degrees of freedom, and use that model to obtain a P-value.

STEP-BY-STEP EXAMPLE

A Paired t -Test

Question: Was there a difference in speeds between the inner and outer speed-skating lanes at the 2006 Winter Olympics?

THINK

Plan State what we want to know.

Identify the *parameter* we wish to estimate. Here our parameter is the mean difference in race times.

Identify the variables and check the W's.

Hypotheses State the null and alternative hypotheses.

Although fans suspected one lane was faster, we can't use the data we have to specify the direction of a test. We (and Olympic officials) would be interested in a difference in either direction, so we'd better test a two-sided alternative.

REALITY CHECK

The individual differences are all in seconds. We should expect the mean difference to be comparable in magnitude.

Model Think about the assumptions and check the conditions.

I want to know whether there really was a difference in the *speeds* of the two lanes for speed skating at the 2006 Olympics. I have data for the women's 1500-m race.

H_0 : Neither lane offered an advantage:

$$\mu_d = 0.$$

H_A : The mean difference is different from zero:

$$\mu_d \neq 0.$$

✓ **Independence Assumption:** Each race is independent of the others, so the differences are mutually independent.

State why you think the data are paired. Simply having the same number of individuals in each group and displaying them in side-by-side columns doesn't make them paired.

Think about what we hope to learn and where the randomization comes from. Here, the randomization comes from the racer pairings and lane assignments.

Make a picture—just one. Don't plot separate distributions of the two groups—that entirely misses the pairing. For paired data, it's the Normality of the *differences* that we care about. Treat those paired differences as you would a single variable, and check the Nearly Normal Condition with a histogram or a Normal probability plot.

Specify the sampling distribution model.
Choose the method.



Mechanics

n is the number of *pairs*—in this case, the number of races.

\bar{d} is the mean difference.

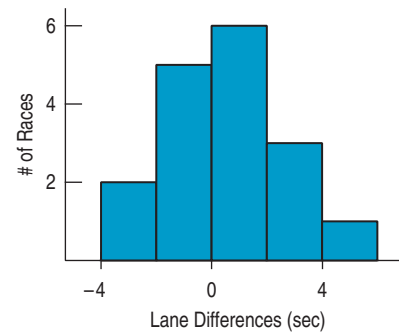
s_d is the standard deviation of the differences.

Find the standard error and the t -score of the observed mean difference. There is nothing new in the mechanics of the paired- t methods. These are the mechanics of the t -test for a mean applied to the differences.

Make a picture. Sketch a t -model centered at the hypothesized mean of 0. Because this is a two-tail test, shade both the region to the right of the observed mean difference of 0.499 seconds and the corresponding region in the lower tail.

Find the P-value, using technology.

- ✓ **Paired Data Assumption:** The data are paired because racers compete in pairs.
- ✓ **Randomization Condition:** Skaters are assigned to lanes at random. Repeating the experiment with different pairings and lane assignments would give randomly different values.
- ✓ **Nearly Normal Condition:** The histogram of the differences is unimodal and symmetric:



The conditions are met, so I'll use a Student's t -model with $(n - 1) = 16$ degrees of freedom, and perform a **paired t -test**.

The data give

$$n = 17 \text{ pairs}$$

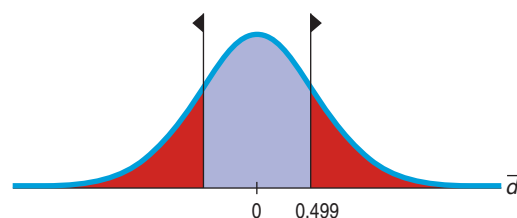
$$\bar{d} = 0.499 \text{ seconds}$$

$$s_d = 2.333 \text{ seconds.}$$

I estimate the standard deviation of \bar{d} using

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{2.333}{\sqrt{17}} = 0.5658$$

$$\text{So } t_{16} = \frac{\bar{d} - 0}{SE(\bar{d})} = \frac{0.499}{0.5658} = 0.882$$



$$P\text{-value} = 2P(t_{16} > 0.882) = 0.39$$

REALITY CHECK

The mean difference is 0.499 seconds. That may not seem like much, but a smaller difference determined the Silver and Bronze medals. The standard error is about this big, so a t -value less than 1.0 isn't surprising. Nor is a large P -value.



Conclusion Link the P -value to your decision about H_0 , and state your conclusion in context.

The P -value is large. Events that happen more than a third of the time are not remarkable. So, even though there is an observed difference between the lanes, I can't conclude that it isn't due simply to random chance. It appears the fans may have interpreted a random fluctuation in the data as favoring one lane. There's insufficient evidence to declare any lack of fairness.

FOR EXAMPLE

Doing a paired t -test

Recap: We want to test whether a change from a five-day workweek to a four-day workweek could change the amount driven by field workers of a health department. We've already confirmed that the assumptions and conditions for a paired t -test are met.

Question: Is there evidence that a four-day workweek would change how many miles workers drive?

H_0 : The change in the health department workers' schedules didn't change the mean mileage driven; the mean difference is zero:

$$\mu_d = 0.$$

H_A : The mean difference is different from zero:

$$\mu_d \neq 0.$$

The conditions are met, so I'll use a Student's t -model with $(n - 1) = 10$ degrees of freedom and perform a **paired t -test**.

The data give

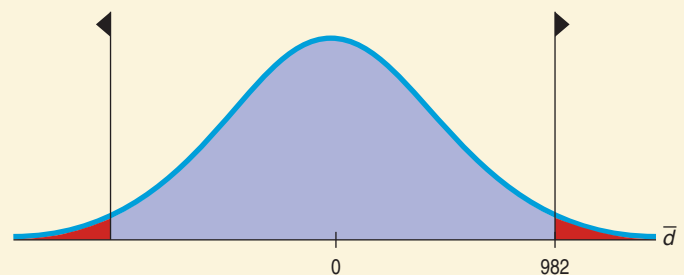
$$n = 11 \text{ pairs}$$

$$\bar{d} = 982 \text{ miles}$$

$$s_d = 1139.6 \text{ miles.}$$

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{1139.6}{\sqrt{11}} = 343.6$$

$$\text{So } t_{10} = \frac{\bar{d} - 0}{SE(\bar{d})} = \frac{982.0}{343.6} = 2.86$$



$$P\text{-value} = 2P(t_{10} > 2.86) = 0.017$$

The P -value is small, so I reject the null hypothesis and conclude that the change in workweek did lead to a change in average driving mileage. It appears that changing the work schedule may reduce the mileage driven by workers.

Note: We should propose a course of action, but it's hard to tell from the hypothesis test whether the reduction matters. Is the difference in mileage important in the sense of reducing air pollution or costs, or is it merely statistically significant? To help make that decision, we should look at a confidence interval. If the difference in mileage proves to be large in a practical sense, then we might recommend a change in schedule for the rest of the department.

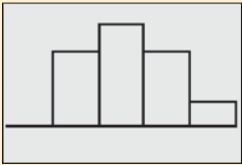
TI Tips

Testing a hypothesis with paired data

```
L1-L2→L3
(-116 1612 1328...
```

L1	L2	L3	3
2798	2914	116	
7724	6112	1612	
7505	6177	1328	
838	1102	-264	
4552	2281	1311	
8107	4997	3110	
1228	1695	-467	

L3(1) = -116



```
T-Test
Inpt: Data Stats
μ₀: 0
List: L3
Freq: 1
μ: F00 <μ₀ >μ₀
Calculate Draw
```

```
T-Test
μ≠0
t=2.85899122
P=.0169862463
x̄=982.8181818
Sx=1140.136116
n=11
```

Since the inference procedures for matched data are essentially just the one-sample t procedures, you already know what to do . . . once you have the list of paired differences, that is. That list is not hard to create.

Test a hypothesis about the mean of paired differences.

- Think: Are the samples independent or paired. Independent? Go back to the last chapter! Paired? Read on.
- Enter the driving data from page 588 into two lists, say *5-Day mileage* in **L1**, *4-Day mileage* in **L2**.
- Create a list of the differences. We want to take each value in **L1**, subtract the corresponding value in **L2**, and store the paired difference in **L3**. The command is **L1-L2 → L3**. (The arrow is the **STO** button.) Now take a look at **L3**. See—it worked!
- Make a histogram of the differences, **L3**, to check the nearly Normal condition. Notice that we do not look at the histograms of the *5-day mileage* or the *4-day mileage*. Those are not the data that we care about now that we are using a paired procedure. Note also that the calculator's first histogram is not close to Normal. More work to do . . .
- As you have seen before, small samples often produce ragged histograms, and these may look very different after a change in bar width. Reset the **WINDOW** to **Xmin=-3000**, **Xmax=4500**, and **Xsc1=1500**. The new histogram looks okay.
- Under **STAT TESTS** simply use **2:T-Test**, as you've done before for hypothesis tests about a mean.
- Specify that the hypothesized difference is 0, you're using the **Data** in **L3**, and it's a two-tailed test.
- **Calculate**.

The small P-value shows strong evidence that on average the change in the workweek reduces the number of miles workers drive.

Confidence Intervals for Matched Pairs

In developed countries, the average age of women is generally higher than that of men. After all, women tend to live longer. But if we look at *married couples*, husbands tend to be slightly older than wives. How much older, on average, are husbands? We have data from a random sample of 200 British couples, the first 7 of which are shown below. Only 170 couples provided ages for both husband and wife, so we can work only with that many pairs. Let's form a confidence interval for the mean difference of husband's and wife's ages for these 170 couples. Here are the first 7 pairs:

WHO 170 randomly sampled couples
WHAT Ages
UNITS Years
WHEN Recently
WHERE Britain

Wife's Age	Husband's Age	Difference (husband - wife)
43	49	6
28	25	-3
30	40	10
57	52	-5
52	58	6
27	32	5
52	43	-9
⋮	⋮	⋮

Clearly, these data are paired. The survey selected *couples* at random, not individuals. We're interested in the mean age difference within couples. How would we construct a confidence interval for the true mean difference in ages?

PAIRED t -INTERVAL

When the conditions are met, we are ready to find the confidence interval for the mean of the paired differences. The confidence interval is

$$\bar{d} \pm t_{n-1}^* \times SE(\bar{d}),$$

where the standard error of the mean difference is $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$.

The critical value t^* from the Student's t -model depends on the particular confidence level, C , that you specify and on the degrees of freedom, $n - 1$, which is based on the number of pairs, n .

Making confidence intervals for matched pairs follows exactly the steps for a one-sample t -interval.

STEP-BY-STEP EXAMPLE

A Paired t -Interval

Question: How big a difference is there, on average, between the ages of husbands and wives?

THINK

Plan State what we want to know.

Identify the variables and check the W's.

Identify the parameter you wish to estimate. For a paired analysis, the parameter of interest is the mean of the differences. The population of interest is the population of differences.

Model Think about the assumptions and check the conditions.

I want to estimate the mean difference in age between husbands and wives. I have a random sample of 200 British couples, 170 of whom provided both ages.

- ✓ **Paired Data Assumption:** The data are paired because they are on members of married couples.
- ✓ **Independence Assumption:** The data are from a randomized survey, so couples should be independent of each other.
- ✓ **Randomization Condition:** These couples were randomly sampled.
- ✓ **10% Condition:** The sample is less than 10% of the population of married couples in Britain.

Make a picture. We focus on the differences, so a histogram or Normal probability plot is best here.

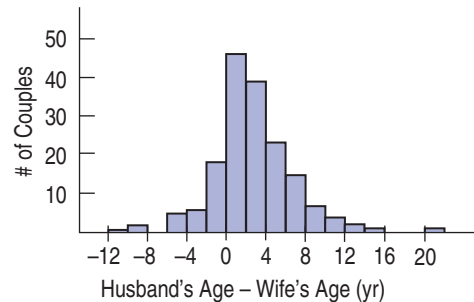
REALITY CHECK

The histogram shows husbands are often older than wives (because most of the differences are greater than 0). The mean difference seen here of about 2 years is reasonable.

State the sampling distribution model.

Choose your method.

✓ **Nearly Normal Condition:** The histogram of the husband – wife differences is unimodal and symmetric:



The conditions are met, so I can use a Student's t -model with $(n - 1) = 169$ degrees of freedom and find a **paired t -interval**.

SHOW

Mechanics

n is the number of *pairs*, here, the number of couples.

\bar{d} is the mean difference.

s_d is the standard deviation of the differences.

Be sure to include the units along with the statistics.

The critical value we need to make a 95% interval comes from a Student's t table, a computer program, or a calculator.

REALITY CHECK

This result makes sense. Our everyday experience confirms that an average age difference of about 2 years is reasonable.

$$n = 170 \text{ couples}$$

$$\bar{d} = 2.2 \text{ years}$$

$$s_d = 4.1 \text{ years}$$

I estimate the standard error of \bar{d} as

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{4.1}{\sqrt{170}} = 0.31 \text{ years.}$$

The df for the t -model is $n - 1 = 169$.

The 95% critical value for t_{169} (from the table) is 1.97.

The margin of error is

$$ME = t_{169}^* \times SE(\bar{d}) = 1.97(0.31) = 0.61$$

So the 95% confidence interval is

$$2.2 \pm 0.6 \text{ years,}$$

or an interval of (1.6, 2.8) years.

TELL

Conclusion Interpret the confidence interval in context.

I am 95% confident that British husbands are, on average, 1.6 to 2.8 years older than their wives.

TI Tips

Creating a confidence interval

```

TInterval
Inpt:Data Stats
x:2.2
Sx:4.1
n:170
C-Level:.95
Calculate

```

```

TInterval
(1.5792,2.8208)
x:2.2
Sx:4.1
n:170

```

Now let's get the TI to create a confidence interval for the mean of paired differences.

We'll demonstrate by using the statistics about the ages of the British married couples. (If we had all the data, we could enter that, of course. All 170 couples? Um, no thanks.) The husband in the sample were an average of 2.2 years older than their wives, with a standard deviation of 4.1 years. We've already seen that the data are paired and that a histogram of the differences satisfies the Nearly Normal Condition. (With a sample this large, we could proceed with inference even if we didn't have the actual data and were unable to make the histogram.)

- Once again, we treat the paired differences just like data from one sample. A confidence interval for the mean difference, then, like that for a mean, uses the **STAT TESTS** one-sample procedure **8:TInterval**.
- Specify **Inpt:Stats**, and enter the statistics for the paired differences.
- **Calculate**.

Done. Finding the interval was the easy part. Now it's time for you to *Tell* what it means. Don't forget to talk about married couples in Britain.

Effect Size

When we examined the speed-skating times, we failed to reject the null hypothesis, so we couldn't be certain whether there really was a difference between the lanes. Maybe there wasn't any difference, or maybe whatever difference there might have been was just too small to matter at all. Were the fans right to be concerned?

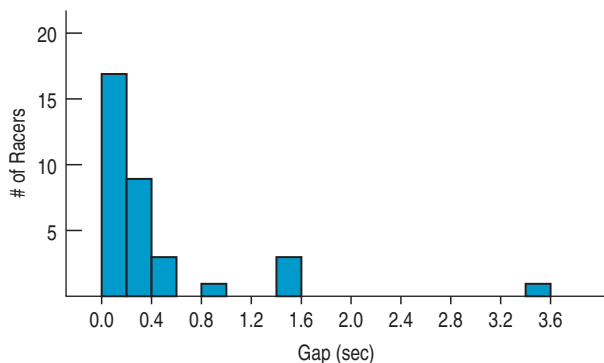
We can't tell from the hypothesis test, but using the same summary statistics, we can find that the corresponding 95% confidence interval for the mean difference is $(-0.70 < \mu_d < 1.70)$ seconds.

A confidence interval is a good way to get a sense for the size of the effect we're trying to understand. That gives us a plausible range of values for the true mean difference in lane times. If differences of 1.7 seconds were too small to matter

in 1500-m Olympic speed skating, we'd be pretty sure there was no need for concern.

But in fact, except for the Gold – Silver gap, the successive gaps between each skater and the next-faster one were *all* less than the high end of this interval, and most were right around the middle of the interval.

So even though we were unable to discern a real difference, the confidence interval shows that the effects we're considering may be big enough to be important. We may want to continue this investigation by checking out other races on this ice and being alert for possible differences at other venues.



FOR EXAMPLE

Looking at effect size with a paired- t confidence interval

Recap: We know that, on average, the switch from a five-day workweek to a four-day workweek reduced the amount driven by field workers in that Illinois health department. However, finding that there is a significant difference doesn't necessarily mean that difference is meaningful or worthwhile. To assess the size of the effect, we need a confidence interval. We already know the assumptions and conditions are met.

Question: By how much, on average, might a change in workweek schedule reduce the amount driven by workers?

$$\begin{aligned}\bar{d} &= 982 \text{ mi} & SE(\bar{d}) &= 343.6 & t_{10}^* &= 2.228 \text{ (for 95\%)} \\ ME &= t_{10}^* \times SE(\bar{d}) & &= 2.228(343.6) & &= 765.54\end{aligned}$$

So the 95% confidence interval for μ_d is 982 ± 765.54 or $(216.46, 1747.54)$ fewer miles.

With 95% confidence, I estimate that by switching to a four-day workweek employees would drive an average of between 216 and 1748 fewer miles per year. With high gas prices, this could save a lot of money.

Blocking

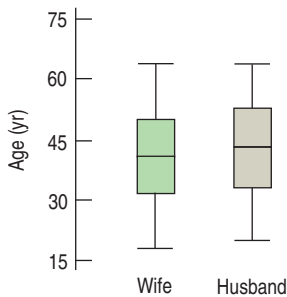


FIGURE 25.2

This display is worthless. It does no good to compare all the wives as a group with all the husbands. We care about the paired differences.

Because the sample of British husbands and wives includes both older and younger couples, there's a lot of variation in the ages of the men and in the ages of the women. In fact, that variation is so great that a boxplot of the two groups would show little difference. But that would be the wrong plot. It's the *difference* we care about. Pairing isolates the extra variation and allows us to focus on the individual differences. In Chapter 13 we saw how we could design an experiment with blocking to isolate the variability between identifiable groups of subjects, allowing us to better see variability among treatment groups due to their response to the treatment. A paired design is an example of blocking.

When we pair, we have roughly half the degrees of freedom of a two-sample test. You may see discussions that suggest that in "choosing" a paired analysis we "give up" these degrees of freedom. This isn't really true, though. If the data are paired, then there never were additional degrees of freedom, and we have no "choice." The fact of the pairing determines how many degrees of freedom are available.

Matching pairs generally removes so much extra variation that it more than compensates for having only half the degrees of freedom. Of course, inappropriate matching when the groups are in fact independent (say, by matching on the first letter of the last name of subjects) would cost degrees of freedom without the benefit of reducing the variance. When you design a study or experiment, you should consider using a paired design if possible.




JUST CHECKING

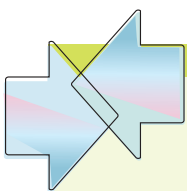
Think about each of the situations described below.

- ▶ Would you use a two-sample t or paired- t method (or neither)? Why?
 - ▶ Would you perform a hypothesis test or find a confidence interval?
1. Random samples of 50 men and 50 women are asked to imagine buying a birthday present for their best friend. We want to estimate the difference in how much they are willing to spend.
 2. Mothers of twins were surveyed and asked how often in the past month strangers had asked whether the twins were identical.

3. Are parents equally strict with boys and girls? In a random sample of families, researchers asked a brother and sister from each family to rate how strict their parents were.
4. Forty-eight overweight subjects are randomly assigned to either aerobic or stretching exercise programs. They are weighed at the beginning and at the end of the experiment to see how much weight they lost.
 - a) We want to estimate the mean amount of weight lost by those doing aerobic exercise.
 - b) We want to know which program is more effective at reducing weight.
5. Couples at a dance club were separated and each person was asked to rate the band. Do men or women like this band more?

WHAT CAN GO WRONG?

- ▶ **Don't use a two-sample t -test when you have paired data.** See the What Can Go Wrong? discussion in Chapter 24.
- ▶ **Don't use a paired- t method when the samples aren't paired.** Just because two groups have the same number of observations doesn't mean they can be paired, even if they are shown side by side in a table. We might have 25 men and 25 women in our study, but they might be completely independent of one another. If they were siblings or spouses, we might consider them paired. Remember that you cannot *choose* which method to use based on your preferences. If the data are from two independent samples, use two-sample t methods. If the data are from an experiment in which observations were paired, you must use a paired method. If the data are from an observational study, you must be able to defend your decision to use matched pairs or independent groups.
- ▶ **Don't forget outliers.** The outliers we care about now are in the differences. A subject who is extraordinary both before and after a treatment may still have a perfectly typical difference. But one outlying difference can completely distort your conclusions. Be sure to plot the differences (even if you also plot the data).
- ▶ **Don't look for the difference between the means of paired groups with side-by-side boxplots.** The point of the paired analysis is to remove extra variation. The boxplots of each group still contain that variation. Comparing them is likely to be misleading. 



CONNECTIONS

The most important connection is to the concept of blocking that we first discussed when we considered designed experiments in Chapter 13. Pairing is a basic and very effective form of blocking.

Of course, the details of the mechanics for paired t -tests and intervals are identical to those for the one-sample t -methods. Everything we know about those methods applies here.

The connection to the two-sample and pooled methods of the previous chapter is that when the data are naturally paired, those methods are not appropriate because paired data fail the required condition of independence.



WHAT HAVE WE LEARNED?

When we looked at various ways to design experiments, back in Chapter 13, we saw that pairing can be a very effective strategy. Because pairing can help control variability between individual subjects, paired methods are usually more powerful than methods that compare independent groups. Now we've learned that analyzing data from matched pairs requires different inference procedures.

- ▶ We've learned that paired t -methods look at pairwise differences. Based on these differences, we test hypotheses and generate confidence intervals. These procedures are mechanically identical to the one-sample t -methods we saw in Chapter 23.
- ▶ We've also learned to *Think* about the design of the study that collected the data before we proceed with inference. We must be careful to recognize pairing when it is present but not assume it when it is not. Making the correct decision about whether to use independent t -procedures or paired t -methods is the first critical step in analyzing the data.

Terms

Paired data

588. Data are paired when the observations are collected in pairs or the observations in one group are naturally related to observations in the other. The simplest form of pairing is to measure each subject twice—often before and after a treatment is applied. More sophisticated forms of pairing in experiments are a form of blocking and arise in other contexts. Pairing in observational and survey data is a form of matching.

Paired t -test

591. A hypothesis test for the mean of the pairwise differences of two groups. It tests the null hypothesis

$$H_0: \mu_d = \Delta_0,$$

where the hypothesized difference is almost always 0, using the statistic

$$t = \frac{\bar{d} - \Delta_0}{SE(\bar{d})}$$

with $n - 1$ degrees of freedom, where $SE(\bar{d}) = \frac{s_d}{\sqrt{n}}$, and n is the number of pairs.

Paired- t confidence interval

595. A confidence interval for the mean of the pairwise differences between paired groups found as

$$\bar{d} \pm t_{n-1}^* \times SE(\bar{d}), \text{ where } SE(\bar{d}) = \frac{s_d}{\sqrt{n}} \text{ and } n \text{ is the number of pairs.}$$

Skills

THINK

- ▶ Be able to recognize whether a design that compares two groups is paired.

SHOW

- ▶ Be able to find a paired confidence interval, recognizing that it is mechanically equivalent to doing a one-sample t -interval applied to the differences.
- ▶ Be able to perform a paired t -test, recognizing that it is mechanically equivalent to a one-sample t -test applied to the differences.

TELL

- ▶ Be able to interpret a paired t -test, recognizing that the hypothesis tested is about the mean of the differences between paired values rather than about the differences between the means of two independent groups.
- ▶ Be able to interpret a paired t -interval, recognizing that it gives an interval for the mean difference in the pairs.

PAIRED *t* ON THE COMPUTER

Most statistics programs can compute paired-*t* analyses. Some may want you to find the differences yourself and use the one-sample *t* methods. Those that perform the entire procedure will need to know the two variables to compare. The computer, of course, cannot verify that the variables are naturally paired. Most programs will check whether the two variables have the same number of observations, but some stop there, and that can cause trouble. Most programs will automatically omit any pair that is missing a value for either variable (as we did with the British couples). You must look carefully to see whether that has happened.

As we've seen with other inference results, some packages pack a lot of information into a simple table, but you must locate what you want for yourself. Here's a generic example with comments:

Could be called "Matched Pair" or "Paired-*t*" analysis

Individual group means

Mean of the differences and its SE

Paired *t*-statistic

Matched Pairs		t-Ratio	
Group 1 Mean	42.9176	t-Ratio	7.151783
Group 2 Mean	40.6824	DF	169
Mean Difference	2.23529	Prob > t	<0.0001
Std Error	0.31255	Prob > t	<0.0001
Upper 95%	2.85230	Prob < t	1.0000
Lower 95%	1.61829		
N	170		
Correlation	0.93858		

its df

P-values for:
Two-sided
One-sided alternatives

Corresponding confidence interval bounds on the mean difference.

Correlation is often reported. Be careful. We have not checked for nonlinearity or outlying pairs. Either could make the correlation meaningless, even though the paired *t* was still appropriate.

Other packages try to be more descriptive. It may be easier to find the results, but you may get less information from the output table.

Groups may have missing values. Only cases with both values present are used in a paired-*t* analysis. You may not learn that from some packages.

Even simple tables can have superfluous numbers such as these.

SD (differences)

SE(\bar{d})

CI corresponds to specified α .

Paired T for hAge-wAge				
	N	Mean	Std Dev	SE(Mean)
hAge	199	42.62	11.646	0.8255
wAge	170	40.68	11.414	0.8254
Paired Difference	170	2.235	4.0752	0.31255

\bar{d}

95% CI for mean difference: (1.618, 2.852)

T-Test of mean difference = 0 (vs \neq 0): T-Value = 7.1518 P-Value < 0.0001

Some packages let you specify the alternative and report only results for that alternative.

t-statistic and its P-value (You may need to calculate $n_d - 1$ for yourself to get the df.)

Computers make it easy to examine the boxplots of the two groups and the histogram of the differences—both important steps. Some programs offer a scatterplot of the two variables. That can be helpful. In terms of the scatterplot, a paired t -test is about whether the points tend to be above or below the 45° line $y = x$. (Note, though, that pairing says nothing about whether the scatterplot should be straight. That doesn't matter for our t -methods.)

EXERCISES

- More eggs?** Can a food additive increase egg production? Agricultural researchers want to design an experiment to find out. They have 100 hens available. They have two kinds of feed: the regular feed and the new feed with the additive. They plan to run their experiment for a month, recording the number of eggs each hen produces.
 - Design an experiment that will require a two-sample t procedure to analyze the results.
 - Design an experiment that will require a matched-pairs t procedure to analyze the results.
 - Which experiment would you consider the stronger design? Why?
- MTV.** Some students do homework with the TV on. (Anyone come to mind?) Some researchers want to see if people can work as effectively with as without distraction. The researchers will time some volunteers to see how long it takes them to complete some relatively easy crossword puzzles. During some of the trials, the room will be quiet; during other trials in the same room, a TV will be on, tuned to MTV.
 - Design an experiment that will require a two-sample t procedure to analyze the results.
 - Design an experiment that will require a matched-pairs t procedure to analyze the results.
 - Which experiment would you consider the stronger design? Why?
- Sex sells?** Ads for many products use sexual images to try to attract attention to the product. But do these ads bring people's attention to the item that was being advertised? We want to design an experiment to see if the presence of sexual images in an advertisement affects people's ability to remember the product.
 - Describe an experimental design requiring a matched-pairs t procedure to analyze the results.
 - Describe an experimental design requiring an independent sample procedure to analyze the results.
- Freshman 15?** Many people believe that students gain weight as freshmen. Suppose we plan to conduct a study to see if this is true.
 - Describe a study design that would require a matched-pairs t procedure to analyze the results.
 - Describe a study design that would require a two-sample t procedure to analyze the results.
- Women.** Values for the labor force participation rate of women (LFPR) are published by the U.S. Bureau of Labor Statistics. We are interested in whether there was a

difference between female participation in 1968 and 1972, a time of rapid change for women. We check LFPR values for 19 randomly selected cities for 1968 and 1972. Shown below is software output for two possible tests:

Paired t-T est of $\mu(1 - 2)$
 Test Ho: $\mu(1972-1968) = 0$ vs Ha: $\mu(1972-1968) \neq 0$
 Mean of Paired Differences = 0.0337
 t-Statistic = 2.458 w/ 18 df
 p = 0.0244

2-Sample t-T est of $\mu_1 - \mu_2$
 Ho: $\mu_1 - \mu_2 = 0$ Ha: $\mu_1 - \mu_2 \neq 0$
 Test Ho: $\mu(1972) - \mu(1968) = 0$ vs
 Ha: $\mu(1972) - \mu(1968) \neq 0$
 Difference Between Means = 0.0337
 t-Statistic = 1.496 w/ 35 df
 p = 0.1434

- Which of these tests is appropriate for these data? Explain.
- Using the test you selected, state your conclusion.

- T** 6. **Rain.** Simpson, Alsen, and Eden (*Technometrics* 1975) report the results of trials in which clouds were seeded and the amount of rainfall recorded. The authors report on 26 seeded and 26 unseeded clouds in order of the amount of rainfall, largest amount first. Here are two possible tests to study the question of whether cloud seeding works. Which test is appropriate for these data? Explain your choice. Using the test you select, state your conclusion.

Paired t-T est of $\mu(1 - 2)$
 Mean of Paired Differences = -277.39615
 t-Statistic = -3.641 w/ 25 df
 p = 0.0012

2-Sample t-T est of $\mu_1 - \mu_2$
 Difference Between Means = -277.4
 t-Statistic = -1.998 w/ 33 df
 p = 0.0538

- Which of these tests is appropriate for these data? Explain.
 - Using the test you selected, state your conclusion.
- T** 7. **Friday the 13th, I.** In 1993 the *British Medical Journal* published an article titled, "Is Friday the 13th Bad for Your Health?" Researchers in Britain examined how Friday the 13th affects human behavior. One question was