CHAPTER 25

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

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
REMPEL 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

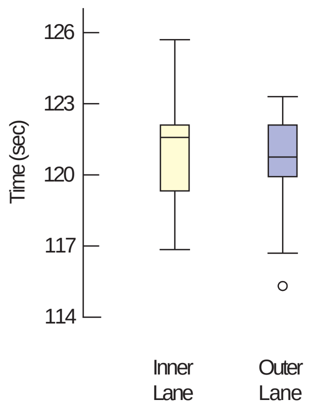


FIGURE 25.1

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

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.

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

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.

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 Theo- rem 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 deter- mine 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.

A S *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 2 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.

Name	5-Day mileage	4-Day mileage	Difference
Jeff	2798	2914	2116
Betty	7724	6112	1612
Roger	7505	6177	1328
Tom	838	1102	2264
Aimee	4592	3281	1311
Greg	8107	4997	3110
Larry G.	1228	1695	2467
Tad	8718	6606	2112
Larry M.	1097	1063	34
Leslie	8089	6392	1697
Lee	3807	3362	445

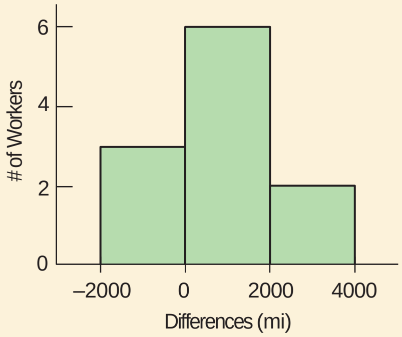
Question: Is it okay to use these data to test whether the new schedule changed the amount of driving?

Ç **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.

Ç **Independence Assumption:** The driving behavior of any individual worker is independent of the others, so the differences are mutually independent.

Ç **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.

Ç **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.

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.

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.

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.

H_0: Neither lane offered an advantage:

$$\mu_d = 0.$$

H_A: The mean difference is different from zero:

$$\mu_d \neq 0.$$

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.

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