14

Comparing Two Means

LEARNING OBJECTIVES

In this chapter we show you how to construct confidence intervals and perform hypothesis tests on the difference between the means of two populations. After reading and studying this chapter, you should be able to:

1. Perform a *t*-test on the difference between two means
2. Calculate a confidence interval for the difference between two means using the *t*-distribution
3. Construct confidence intervals and perform hypothesis tests on the difference between means of paired data based on the *t*-distribution

Visa Canada

There were 72 million credit cards in circulation in Canada in 2009, a large number of them issued by Visa. Visa operates the world's largest retail electronic payments network, capable of handling over 10,000 transactions per second. Although many people associate Visa only with credit cards, it also offers debit, prepaid, and commercial cards.

Visa's origins go back to 1958, when Bank of America issued a credit card program called BankAmericard in Fresno, California. During the 1960s it expanded to other U.S. states and to Canada, where Toronto-Dominion Bank, Canadian Imperial Bank of Commerce, Royal Bank of Canada, Banque Canadienne Nationale, and Bank of Nova Scotia issued credit cards under the Chargex name. Other names were used in other countries, but in 1975, they united under the name "Visa."

Although Visa employs 6000 people worldwide, it did not become a publicly traded company until 2008. At that time, Visa Canada, Visa International, and Visa USA merged to form Visa Inc., which had the largest IPO in U.S. history, raising $17.9 billion.

Today, Visa cards are issued in Canada by a number of banks, including CIBC, Desjardins, Laurentian Bank, Royal Bank of Canada, Scotiabank, and TD Canada Trust.

Visa supports the Olympic and Paralympic games, and in return is the only form of electronic payment accepted at Olympic venues. It provides financial support to the Canadian bobsleigh and skeleton teams, enabling them to compete at major international events (they won gold medals at the Nagano and Torino Olympic Winter Games). Visa also supports individual athletes, including 11 Canadians who competed in the 2010 Vancouver Winter Games. It pairs up young athletes with Olympic veterans, who mentor them on how to prepare mentally and physically to perform at their best.

Roadmap for Statistical Inference

Number of Variables	Objective	Large Sample or Normal Population		Small Sample and Non-normal Population or Non-numeric Data	
		Chapter	Parametric Method	Chapter	Nonparametric Method
1	Calculate confidence interval for a proportion	11			
1	Compare a proportion with a given value	12	*z*-test		
1	Calculate a confidence interval for a mean and compare it with a given value	13	*t*-test	17.2	Wilcoxon Signed-Rank Test
2	Compare two proportions	12.8	*z*-test		
2	Compare two means for independent samples	14.1–14.5	*t*-test	17.4, 17.5	Wilcoxon Rank-Sum (Mann-Whitney) Test Tukey's Quick Test
2	Compare two means for paired samples	14.6, 14.7	Paired *t*-test	17.2	Wilcoxon Signed-Rank Test
≥3	Compare multiple means	15	ANOVA: ANalysis Of VAriance	17.3 17.6	Friedman Test Kruskal-Wallis Test
≥3	Compare multiple counts (proportions)	16	χ^2 test		
2	Investigate the relationship between two variables	18	Correlation Regression	17.7, 17.8	Kendall's tau Spearman's rho
≥3	Investigate the relationship between multiple variables	20	Multiple Regression		

[11]**Sources:** Based on Visa. Retrieved from www.visa.ca and www.visa.com; and Credit Cards Canada. Retrieved from http://Canada.creditcards.com; Canadian Bankers Association. (2012). Credit cards: Statistics and facts.

In 2011, over 60% of Canadians paid off their credit card balance each month, and this percentage was independent of income level. The average Canadian household had two credit cards, the balance on which accounted for 5% of household debt. As of December 2010, over 40 credit cards in Canada had an interest rate of under 12%, making the credit card business intensely competitive.

Rival banks and lending agencies are constantly trying to create new products and offers to win new customers, keep current customers, and provide incentives for current customers to charge more on their cards.

Are some credit card promotions more effective than others? For example, do customers spend more using their credit card if they know they'll be given "double miles" or "double points" toward flights, hotel stays, or store purchases? To answer questions such as this, credit card issuers often perform experiments on a sample of customers, making them an *offer* of an incentive, while other customers receive no offer. Promotions cost the company money, so the company needs to estimate the size of any increased revenue to judge whether it's sufficient to cover expenses. By comparing the performance of the offer on the sample, the company can decide whether the new offer would provide enough potential profit if it were to be "rolled out" and offered to the entire customer base.

Experiments that compare two groups are common throughout both science and industry. Other applications include comparing the effects of a new drug with the traditional therapy, the fuel efficiency of two car engine designs, or the sales of new products on two different customer segments. Usually the experiment is carried out on a subset of the population, often a much smaller subset. Using statistics, we can make statements about whether the means of the two groups differ in the population at large, and how large that difference might be.

LO ❶

14.1 Comparing Two Means

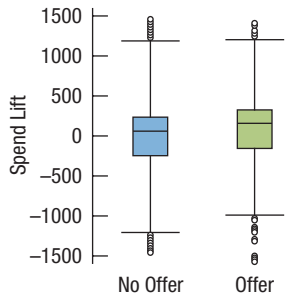


Figure 14.1 Side-by-side boxplots show a small increase in spending for the group that received the promotion.

The natural display for comparing the means of two groups is side-by-side boxplots (see Figure 14.1). For the credit card promotion, the company judges performance by comparing the *mean* spend lift (the change in spending from before receiving the promotion to after receiving it) for the two samples. If the difference in spend lift between the group that received the promotion and the group that didn't is high enough, this will be viewed as evidence that the promotion worked. Looking at the two boxplots, it's not obvious that there's much of a difference. Can we conclude that the slight increase seen for those who received the promotion is more than just random fluctuation? We'll need statistical inference.

For two groups, the statistic of interest is the difference in the observed means of the offer and no offer groups: $\bar{y}_1 - \bar{y}_2$. We've offered the promotion to a random sample of cardholders, and used another sample of cardholders who got no special offer as a control group. We know what happened in our samples, but what we'd really like to know is the difference of the means in the population at large, $\mu_1 - \mu_2$.

We compare two means in much the same way as we compared a single mean to a hypothesized value. But now the population model parameter of interest is the *difference* between the means. In our example, it's the true difference between the mean spend lift for customers offered the promotion and for customers for whom no offer was made. We estimate the difference with $\bar{y}_1 - \bar{y}_2$. How can we tell if a difference we observe in the sample means indicates a real difference in the underlying population means? We'll need to know the sampling distribution model and standard deviation of the difference. Once we know those, we can build a confidence interval and test a hypothesis just as we did for a single mean.

We have data on 500 randomly selected customers who were offered the promotion and another randomly selected 500 who were not. It's easy to find the mean and standard deviation of the spend lift for each of these groups. From these, we can find the standard deviations of the means, but that's not what we want. We need the standard deviation of the *difference* in their means. For that, we can use a simple rule: *If the sample means come from independent samples, the variance of their sum or difference is the sum of their variances.*

- **Variances Add for Sums and Differences** At first, it may seem that this can't be true for differences as well as for sums. Here's some intuition about why variation increases even when we subtract two random quantities. Grab a full box of cereal. The label claims that it contains 500 grams of cereal. We know that's not exact. There's a random quantity of cereal in the box with a mean (presumably) of 500 grams and some variation from box to box. Now pour a 50-gram serving of cereal into a bowl. Of course, your serving isn't exactly 50 grams. There's some variation there, too. How much cereal would you guess was left in the box? Can you guess as accurately as you could for the full box? The mean should be 450 grams. But does the amount left in the box have *less* variation than it did before you poured your serving? Almost certainly not! *After* you pour your bowl, the amount of cereal in the box is still a random quantity (with a smaller mean than before), but you've made it *more variable* because of the uncertainty in the amount you poured. However, notice that we don't add the *standard deviations* of these two random quantities. As we'll see, it's the *variance* of the amount of cereal left in the box that's the sum of the two variances.

As long as the two groups are independent, we find the standard deviation of the *difference* between the two sample means by adding their variances and then taking the square root:

$$\begin{aligned} SD(\bar{y}_1 - \bar{y}_2) &= \sqrt{Var(\bar{y}_1) + Var(\bar{y}_2)} \\ &= \sqrt{\left(\frac{\sigma_1}{\sqrt{n_1}}\right)^2 + \left(\frac{\sigma_2}{\sqrt{n_2}}\right)^2} \\ &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}. \end{aligned}$$

Of course, usually we don't know the true standard deviations of the two groups, σ_1 and σ_2, so we substitute the estimates, s_1 and s_2, and find a *standard error:*

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Just as we did for one mean, we'll use the standard error to see how big the difference really is. You shouldn't be surprised that, just as for a single mean, the ratio of the difference in the means to the standard error of that difference has a sampling model that follows a Student's t distribution.

An Easier Rule?

The formula for the degrees of freedom of the sampling distribution of the difference between two means is complicated. So some books teach an easier rule: The number of degrees of freedom is always at *least* the smaller of $n_1 - 1$ and $n_2 - 1$ and at most $n_1 + n_2 - 2$. The problem is that if you need to perform a two-sample t-test and don't have the formula at hand to find the correct degrees of freedom, you have to be conservative and use the lower value. And *that* approximation can be a poor choice because it can give less than *half* the degrees of freedom you're entitled to from the correct formula.

A Sampling Distribution for the Difference Between Two Means

When the conditions are met (see Section 14.3), the standardized sample difference between the means of two independent groups,

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - (\mu_1 - \mu_2)}{SE(\bar{y}_1 - \bar{y}_2)},$$

can be modelled by a Student's t-model with a number of degrees of freedom found with a special formula. We estimate the standard error with

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

What else do we need? Only the degrees of freedom for the Student's *t*-model. Unfortunately, *that* formula isn't as simple as $n - 1$. The problem is that the sampling model isn't *really* Student's *t*, but something close. The reason is that we estimated two different variances (s_1^2 and s_2^2), and they may be different. That extra variability makes the distribution even more variable than the Student's *t* for either of the means. But by using a special, adjusted degrees of freedom value, we can find a Student's *t*-model that is so close to the right sampling distribution model that nobody can tell the difference. The adjustment formula is straightforward but doesn't help our understanding much, so we leave it to the computer or calculator. (If you're curious and really want to see the formula, look in the footnote.[2])

For Example Sampling distribution of the difference of two means

The owner of a large car dealership wants to understand the negotiation process for buying a new car. Cars are given a "sticker price," but a potential buyer may negotiate a better price. The owner wonders if there's a difference in how men and women negotiate and who, if either, obtains the larger discount.

He takes a random sample of 100 customers from the last six months' sales and finds that 54 were men and 46 were women. On average, the 54 men received a discount of \$962.96 with a standard deviation of \$458.95; the 46 women received an average discount of \$1262.61 with a standard deviation of \$399.70.

Question: What is the mean difference of the discounts received by men and women? What is its standard error? If there is no difference between them, does this seem like an unusually large value?

Answer: The mean difference is \$1262.61 − \$962.96 = \$263.65. The women received, on average, a discount that was larger by \$263.65. The standard error is

$$SE(\bar{y}_{Women} - \bar{y}_{Men}) = \sqrt{\frac{s_{Women}^2}{n_{Women}} + \frac{s_{Men}^2}{n_{Men}}} = \sqrt{\frac{(399.70)^2}{46} + \frac{(458.95)^2}{54}} = \$85.87.$$

So, the difference is \$263.65/85.87 = 3.07 standard errors away from 0. That sounds like a reasonably large number of standard errors for a Student's *t* statistic with 97.94 degrees of freedom.

LO ❶

14.2 The Two-Sample *t*-Test

Notation Alert:

Δ_0 (pronounced "delta zero") isn't so standard that you can assume everyone will understand it. We use it because it's the capital Greek letter "D" for "difference."

Now we've got everything we need to construct the hypothesis test, and you already know how to do it. It's the same idea we used when testing one mean against a hypothesized value. Here, we start by hypothesizing a value for the true difference of the means. We'll call that hypothesized difference Δ_0. (It's so common for that hypothesized difference to be zero that we often just assume $\Delta_0 = 0$.) We then take the ratio of the difference in the means from our

[2]The result is due to Satterthwaite and Welch.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110–114.

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika*, *34*, 28–35.

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

This approximation formula usually doesn't even give a whole number. If you're using a table, you'll need a whole number, so round down to be safe. If you're using technology, the approximation formulas that computers and calculators use for the Student's *t*-distribution can deal with fractional degrees of freedom.

samples to its standard error and compare that ratio with a critical value from a Student's *t*-model. The test is called the **two-sample *t*-test.**

Two-Sample *t*-Test

When the appropriate assumptions and conditions are met, we test the hypothesis

$$H_0: \mu_1 - \mu_2 = \Delta_0,$$

where the hypothesized difference Δ_0 is almost always 0. We use the statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{\mathrm{SE}(\bar{y}_1 - \bar{y}_2)}.$$

The standard error of $\bar{y}_1 - \bar{y}_2$ is

$$\mathrm{SE}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

When the null hypothesis is true, the statistic can be closely modelled by a Student's *t*-model with a number of degrees of freedom given by a special formula. We use that model to compare our *t*-ratio with a critical value for *t* or to obtain a P-value.

For Example The *t*-test for the difference of two means

Question: We saw (on page xxx) that the difference between the average discount obtained by men and women appeared to be large if we assume that there is no true difference. Test the hypothesis, find the P-value, and state your conclusions.

Answer: The null hypothesis is: H_0: $\mu_{Women} - \mu_{Men} = 0$ *vs.* H_A: $\mu_{Women} - \mu_{Men} \neq 0$. The difference $\bar{y}_{Women} - \bar{y}_{Men}$ is \$263.65 with a standard error of \$84.26. The *t*-statistic is the difference divided by the standard error:

$t = \dfrac{\bar{y}_{Women} - \bar{y}_{Men}}{SE(\bar{y}_{Women} - \bar{y}_{Men})} = \dfrac{263.65}{85.87} = 3.07$. The approximation formula gives 97.94 degrees of freedom (which is close to the maximum possible of $n_1 + n_2 - 2 = 98$). The P-value (from technology) for $t = 3.07$ with 97.94 df is 0.0028. We reject the null hypoythesis. There is strong evidence to suggest that the difference in mean discount received by men and women is not 0.

LO ❶

14.3 Assumptions and Conditions

Before we can perform a two-sample *t*-test, we have to check the assumptions and conditions.

Independence Assumption

The data in each group must be drawn independently and at random from each group's own homogeneous population or generated by a randomized comparative experiment. We can't expect that the data, taken as one big group, come from a homogeneous population because that's what we're trying to test. But without randomization of some sort, there are no sampling distribution models and no inference. We should think about whether the Independence Assumption is reasonable. We can also check two conditions:

Randomization Condition: Were the data collected with suitable randomization? For surveys, are they a representative random sample? For experiments, was the experiment randomized?

10% Condition: We usually don't check this condition for differences of means. We'll check it only if we have a very small population or an extremely large sample. We needn't worry about it at all for randomized experiments.

Normal Population Assumption

As we did before with Student's *t*-models, we need the assumption that the underlying populations are *each* Normally distributed. So we check one condition.

Nearly Normal Condition: We must check this for *both* groups; a violation by either one violates the condition. As we saw for single sample means, the Normality assumption matters most when sample sizes are small. When either group is small ($n < 15$), you should not use these methods if the histogram or Normal probability plot shows skewness. For n's closer to 40, a mildly skewed histogram is okay, but you should remark on any outliers you find and not work with severely skewed data. When both groups are bigger than that, the Central Limit Theorem starts to work unless the data are severely skewed or there are extreme outliers, so the Nearly Normal Condition for the data matters less. Even in large samples, however, you should still be on the lookout for outliers, extreme skewness, and multiple modes.

Independent Groups Assumption

To use the two-sample *t*-methods, the two groups we're comparing must be independent of each other. In fact, the test is sometimes called the two *independent samples t*-test. No statistical test can verify that the groups are independent. You have to think about how the data were collected. The assumption would be violated, for example, if one group comprised husbands and the other their wives. Whatever we measure on one might naturally be related to the other. Similarly, if we compared subjects' performances before some treatment with their performances afterward, we'd expect a relationship of each "before" measurement with its corresponding "after" measurement. Measurements taken for two groups over time when the observations are taken at the same time may be related—especially if they share, for example, the chance that they were influenced by the overall economy or world events. In cases such as these, where the observational units in the two groups are related or matched, *the two-sample methods of this chapter can't be applied*. When this happens, we need a different procedure.

Guided Example Scotiabank Credit Card Promotion

Suppose Scotiabank wants to evaluate the effectiveness of offering an incentive on one of its Visa cards. The preliminary market research has suggested that a new incentive may increase customer spending. However, before the bank invests in this promotion on the entire population of cardholders, it tests it for six months on a sample of 1000, and obtains the data you'll find in the file **ch14_GE_Credit_Card_Promo**. We are hired as statistical consultants to analyze the results. To judge whether the incentive works, we will examine the change in spending (called the *spend lift*) over a six-month period. We'll see whether the spend lift for the group that received the offer was greater than that for the group that received no offer. If we observe differences, how will we know whether these differences are important (or real) enough to justify our costs?

(*continued*)

PLAN

Setup State what we want to know.

Identify the parameter we wish to estimate. Here our parameter is the difference in the means, not the individual group means.

Identify the population(s) about which we wish to make statements.

Identify the variables and context.

Make a graph to compare the two groups and check the distribution of each group. For completeness, we should report any outliers. If any outliers are extreme enough, we should consider performing the test both with and without the outliers and reporting the difference.

Model Check the assumptions and conditions.

For large samples like these with quantitative data, we often don't worry about the 10% Condition.

State the sampling distribution model for the statistic. Here the degrees of freedom will come from the approximation formula in footnote 2.

Specify your method.

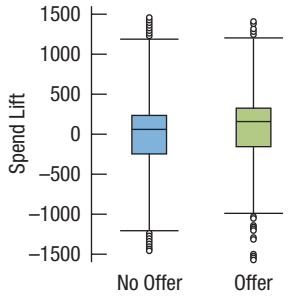
We want to know if cardholders who are offered a promotion spend more on their credit card. We have the spend lift (in $) for a random sample of 500 cardholders who were offered the promotion and for a random sample of 500 customers who were not.

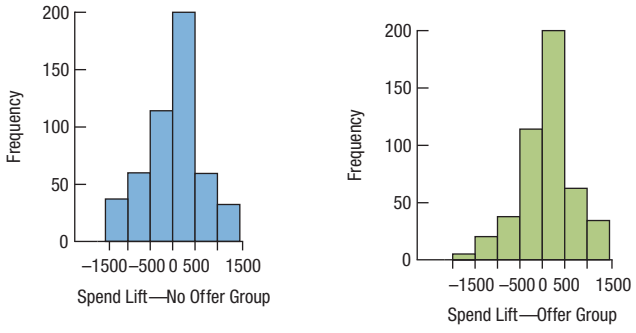
H_O: The mean spend lift for the group who received the offer is the same as for the group who did not:

H_O: $\mu_{Offer} = \mu_{No\ Offer}$
or H_O: $\mu_{Offer} - \mu_{No\ Offer} = 0$

H_A: The mean spend lift for the group who received the offer is higher:

H_A: $\mu_{Offer} > \mu_{No\ Offer}$
or H_A: $\mu_{Offer} - \mu_{No\ Offer} > 0$





The boxplots and histograms show the distribution of both groups. It looks like the distribution for each group is fairly symmetric. The boxplots indicate several outliers in each group, but we have no reason to delete them and their impact is minimal.

✓ **Independence Assumption.** We have no reason to believe that the spending behaviour of one customer would influence the spending behaviour of another customer in the same group. The data report the "spend lift" for each customer for the same time period.

✓ **Randomization Condition.** The customers who were offered the promotion were selected at random.

✓ **Nearly Normal Condition.** The samples are large, so we're not overly concerned with this condition, and the boxplots and histograms show symmetric distributions for both groups.

✓ **Independent Groups Assumption.** Customers were assigned to groups at random. There's no reason to think that those in one group can affect the spending behaviour of those in the other group. Under these conditions, it's appropriate to use a Student's *t*-model.

We will use a two-sample *t*-test.

DO

Mechanics List the summary statistics. Be sure to include the units along with the statistics. Use meaningful subscripts to identify the groups.

Use the sample standard deviations to find the standard error of the sampling distribution.

The best alternative is to let the computer use the approximation formula for the degrees of freedom and find the P-value.

We know $n_{\text{No Offer}} = 500$ and $n_{\text{Offer}} = 500$.

From technology, we find:

$$\bar{y}_{\text{No Offer}} = \$7.69 \quad \bar{y}_{\text{Offer}} = \$127.61$$
$$s_{\text{No Offer}} = \$611.62$$
$$s_{\text{Offer}} = \$566.05$$

The observed difference in the two means is

$$\bar{y}_{\text{Offer}} - \bar{y}_{\text{No Offer}} = \$127.61 - \$7.69 = \$119.92.$$

The groups are independent, so

$$SE(\bar{y}_{\text{Offer}} - \bar{y}_{\text{No Offer}}) = \sqrt{\frac{(611.62)^2}{500} + \frac{(566.05)^2}{500}}$$
$$= \$37.27.$$

The observed ***t***-value is

$$t = 119.92/37.27 = 3.218.$$

The degrees of freedom, df, comes from technology or from the formula in footnote 2:

$$df = \frac{\left(\frac{611.62^2}{500} + \frac{566.05^2}{500}\right)^2}{\frac{1}{499}\left(\frac{611.62^2}{500}\right)^2 + \frac{1}{499}\left(\frac{566.05^2}{500}\right)^2}$$
$$= 992$$

(To use critical values, we could find that the one-sided 0.05 critical value for a *t* with 992 df is $t^* = 1.646$.

Our observed *t*-value is larger than this, so we could reject the null hypothesis at the 0.05 level. In fact, since our *t*-value of 3.218 is a lot higher than 1.646, we can reject the null hypothesis at an even-higher significance level—for example, for P = 0.01, the critical value is $t^* = 2.33$.)

Using software to obtain the P-value, we get:

Promotional Group	N	Mean	StDev
No	500	7.69	611.62
Yes	500	127.61	566.05

Difference = mu (1) − mu (0)
Estimate for difference: 119.9231
$t = 3.2178$, df = 992
One-sided P-value = 0.0006669

(*continued*)

REPORT

Conclusion Interpret the test results in the proper context.

MEMO:

Re: Credit Card Spending

Our analysis of the credit card promotion experiment found that customers offered the promotion spent more than those not offered the promotion. The difference was statistically significant, with a P−value<0.001. So we conclude that this promotion will increase spending. The difference in spend lift averaged $119.92, but our analyses so far haven't determined how much income this will generate for the company and thus whether the estimated increase in spending is worth the cost of the offer.

Just Checking

Many office "coffee stations" collect voluntary payments for the food consumed. Researchers at the University of Newcastle upon Tyne performed an experiment to see whether the image of eyes watching would change employee behaviour.[3] They alternated pictures of eyes looking at the viewer with pictures of flowers each week on the cupboard behind the "honesty box." The researchers then measured the consumption of milk to approximate the amount of food consumed and recorded the contributions (in £) each week per litre of milk. The table summarizes their results.

	Eyes	Flowers
n(# weeks)	5	5
$\bar{y}$	0.417£/*litre*	0.151£/*litre*
s	0.1811	0.067

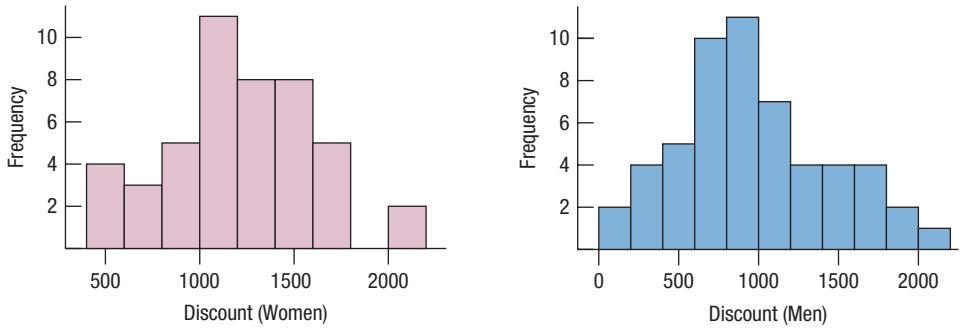
1 What null hypothesis were the researchers testing?

2 Check the assumptions and conditions needed to test whether there really is a difference in behaviour due to the difference in pictures.

For Example Checking assumptions and conditions for a two-sample *t*-test

Question: In the previous example on page xxx, we rejected the null hypothesis that the mean discount received by men and women is the same. Here are the histograms of the discounts for both women and men. Check the assumptions and conditions and state any concerns you might have about the conclusion we reached.

[3]Bateson, M., Nettle, D., & Roberts, G. (2006.) Cues of being watched enhance cooperation in a real-world setting. *Biology Letters*, *2*, 412–14. doi:10.1098/rsbl.2006.0509



Answer: We were told that the sample was random, so the Randomiazation Condition is satisfied. There's no reason to think that the men's and women's responses are related (as they might be if they were husband and wife pairs), so the Independent Group Assumption is plausible. Because both groups have more than 40 observations, the discounts can be mildly skewed, which is the case. There are no obvious outliers (there's a small gap in the women's distribution, but the observations aren't far from the centre), so all the assumptions and conditions seem to be satisfied. We have no real concerns about the conclusion we reached that the mean difference is not 0.

LO ❷

14.4 A Confidence Interval for the Difference Between Two Means

We rejected the null hypothesis that customers' mean spending wouldn't change when offered a promotion. Because the company took a random sample of customers for each group, and our P-value was convincingly small, we concluded that this difference is not zero for the population. Does this mean we should offer the promotion to all customers?

A hypothesis test really says nothing about the size of the difference. All it says is that the observed difference is large enough that we can be confident it isn't zero. That's what the term "statistically significant" means. It doesn't say that the difference is important, financially significant, or interesting. Rejecting a null hypothesis simply says that the observed statistic is unlikely to have been observed if the null hypothesis were true.

So, what recommendations can we make to the company? Almost every business decision will depend on looking at a range of likely scenarios—precisely the kind of information a confidence interval gives. We construct the confidence interval for the difference in means in the usual way, starting with our observed statistic, in this case $(\bar{y}_1 - \bar{y}_2)$. We then add and subtract a multiple of the standard error $\text{SE}(\bar{y}_1 - \bar{y}_2)$ where the multiple is based on the Student's t-distribution with the same df formula we saw before.

Confidence Interval for the Difference Between Two Means

When the conditions are met, we're ready to find a **two-sample *t*-interval** for the difference between means of two independent groups, $\mu_1 - \mu_2$. The confidence interval is

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{\text{df}} \times \text{SE}(\bar{y}_1 - \bar{y}_2),$$

where the standard error of the difference of the means is

$$\text{SE}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The critical value t^*_{df} depends on the particular confidence level and on the number of degrees of freedom.

Guided Example — Scotiabank Credit Card Spending

Let's assume that Scotiabank accepts our advice that the group of customers who received the offer increased their credit card spending by more than the group who didn't receive any offer. In statistical terms, we rejected the null hypothesis that the mean spending in the two groups was equal. From Scotiabank's perspective, it now knows that there's an increase in spending. But is the increase large enough and reliable enough to cover the costs of the promotion and roll out the promotion nationwide? We need to estimate the magnitude and variability of the spend lift.

PLAN

Setup State what we want to know.

Identify the *parameter* we wish to estimate. Here our parameter is the difference in the means, not the individual group means.

Identify the *population(s)* about which we wish to make statements.

Identify the variables and context.

Specify the method.

We want to find a 95% confidence interval for the mean difference in spending between those who are offered a promotion and those who aren't.

We looked at the boxplots and histograms of the groups and checked the conditions before. The same assumptions and conditions are appropriate here, so we can proceed directly to the confidence interval.

We will use a two-sample *t*-interval.

DO

Mechanics Construct the confidence interval. Be sure to include the units along with the statistics. Use meaningful subscripts to identify the groups.

Use the sample standard deviations to find the standard error of the sampling distribution.

The best alternative is to let the computer use the approximation formula for the degrees of freedom and find the confidence interval.

Ordinarily, we rely on technology for the calculations. In our hand calculations, we rounded values at intermediate steps to show the steps more clearly. The computer keeps full precision and is the calculation you should report. The difference between the hand and computer calculations is about $0.08.

In our previous analysis, we found:

$$\bar{y}_{\text{No Offer}} = \$7.69 \quad \bar{y}_{\text{Offer}} = \$127.61$$
$$s_{\text{No Offer}} = \$611.62 \quad s_{\text{Offer}} = \$566.05$$

The observed difference in the two means is

$$\bar{y}_{\text{Offer}} - \bar{y}_{\text{No Offer}} = \$127.61 - \$7.69 = \$119.92,$$

and the standard error is

$$SE(\bar{y}_{\text{Offer}} - \bar{y}_{\text{No Offer}}) = \$37.27.$$

From technology, the *df* is 992.007, and the 0.025 critical value for *t* with 992.007 *df* is 1.96. So the 95% confidence interval is

$$119.92 \pm 1.96(37.27) = (\$46.87, \$192.97).$$

Using software to obtain these computations, we get:

95 percent confidence interval:

46.78784, 193.05837

sample means:

No Offer	*Offer*
7.690882	*127.613987*

REPORT

Conclusion Interpret the test results in the proper context.

MEMO:

Re: Credit Card Promotion Experiment

In our experiment, the promotion resulted in an increased spend lift of $119.92 on average. Further analysis gives a 95% confidence interval of ($46.79, $193.06). In other words, we expect with 95% confidence that under similar conditions, the mean spend lift we achieve when we roll out the offer to all similar customers will be in this interval. We recommend that the company consider whether the values in this interval will justify the cost of the promotion program.

For Example A confidence interval for the difference between two means

Question: We concluded on page xxx that, on average, women receive a larger discount than men at the car dealership. How big is the difference, on average? Find a 95% confidence interval for the difference.

Answer: We've seen that the difference from our sample is $263.65 with a standard error of $85.84 and that it has 97.94 degrees of freedom. The 95% critical value for a t with 97.94 degrees of freedom is 1.984.

$$\bar{y}_{Women} - \bar{y}_{Men} \pm t^*_{97.97} SE(\bar{y}_{Women} - \bar{y}_{Men}) = 263.65 \pm 1.984 \times 85.87 = (\$93.28, \$434.02)$$

We are 95% confident that, on average, women received a discount that's between $93.28 and $434.02 larger than the men at this dealership.

14.5 The Pooled *t*-Test

If you bought a used camera in good condition from a friend, would you pay the same as you would if you bought the same item from a stranger? A researcher at Cornell University[4] wanted to know how friendship might affect simple sales such as this. She randomly divided subjects into two groups and gave each group descriptions of items they might want to buy. One group was told to imagine buying from a friend whom they expected to see again. The other group was told to imagine buying from a stranger.

Table 14.1 gives the prices they offered for a used camera in good condition.

The researcher who designed the friendship study was interested in testing the impact of friendship on negotiations. Previous theories had doubted that friendship had a measurable effect on pricing, but she hoped to find such an effect. The usual null hypothesis is that there's no difference in means, and that's what we'll use for the camera purchase prices.

[4]Halpern, J. J. (1997). The transaction index: A method for standardizing comparisons of transaction characteristics across different contexts. *Group Decision and Negotiation* 6(6), 557–572.

Price Offered for a Used Camera ($)	
Buying from a Friend	Buying from a Stranger
275	260
300	250
260	175
300	130
255	200
275	225
290	240
300	

Table 14.1 Prices offered for a used camera.

WHO	University students
WHAT	Prices offered for a used camera ($)
WHEN	1990s
WHERE	Cornell University
WHY	To study the effects of friendship on transactions

When we performed the *t*-test earlier in the chapter, we used an approximation formula that adjusts the degrees of freedom to a lower value. When $n_1 + n_2$ is only 15, as it is here, we don't really want to lose any degrees of freedom. Because this is an experiment, we might be willing to make another assumption. The null hypothesis says that whether you buy from a friend or a stranger should have no effect on the mean amount you're willing to pay for a camera. If it has no effect on the means, should it affect the variance of the transactions?

If we're willing to *assume* that the variances of the groups are equal (at least when the null hypothesis is true), then we can save some degrees of freedom. To do that, we have to pool the two variances that we estimate from the groups into one common, or *pooled*, estimate of the variance:

$$s^2_{\text{pooled}} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}$$

(If the two sample sizes are equal, this is just the average of the two variances.)

Now we just substitute this pooled variance for each of the variances in the standard error formula. Remember, the standard error formula for the difference of two independent means is

$$\text{SE}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

We substitute the common pooled variance for each of the two variances in this formula, making the pooled standard error formula simpler:

$$\text{SE}_{\text{pooled}}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s^2_{\text{pooled}}}{n_1} + \frac{s^2_{\text{pooled}}}{n_2}} = s_{\text{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The formula for degrees of freedom for the Student's *t*-model is simpler, too. It was so complicated for the two-sample *t* that we stuck it in a footnote. Now it's just $\text{df} = (n_1 - 1) + (n_2 - 1)$.

Substitute the pooled *t* estimate of the standard error and its degrees of freedom into the steps of the confidence interval or hypothesis test and you'll be using pooled *t*-methods. Of course, if you decide to use a pooled *t*-method, you must defend your assumption that the variances of the two groups are equal.

To use the pooled *t*-methods, you'll need to add the equal variance assumption that the variances of the two populations from which the samples have been drawn are equal. That is, $\sigma_1^2 = \sigma_2^2$. (Of course, we can think about the standard deviations being equal instead.)

Pooled *t*-Test and Confidence Interval for the Difference Between Means

The conditions for the **pooled *t*-test** for the difference between the means of two independent groups are the same as for the two-sample *t*-test, with the additional assumption that the variances of the two groups are the same. We test the hypothesis

$$H_0: \mu_1 - \mu_2 = \Delta_0,$$

where the hypothesized difference Δ_0 is almost always 0, using the statistic

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE_{pooled}(\bar{y}_1 - \bar{y}_2)}.$$

The standard error of $\bar{y}_1 - \bar{y}_2$ is

$$SE_{pooled}(\bar{y}_1 - \bar{y}_2) = s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where the pooled variance is

$$s^2_{pooled} = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

When the conditions are met and the null hypothesis is true, we can model this statistic's sampling distribution with a Student's *t*-model with $(n_1 - 1) + (n_2 - 1)$ degrees of freedom. We use that model to obtain a P-value for a test or a margin of error for a confidence interval.

The corresponding **pooled *t* confidence interval** is

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE_{pooled}(\bar{y}_1 - \bar{y}_2),$$

where the critical value t^* depends on the confidence level and is found with $(n_1 - 1) + (n_2 - 1)$ degrees of freedom.

Guided Example — Role of Friendship in Negotiations

We have data on the prices two randomly selected groups of people would offer for a used camera. The first group think they're buying from a friend and the second think they're buying from a stranger. Our samples are quite small, but we wish to see whether they're large enough to establish whether friendship has an impact on the price people are prepared to offer.

PLAN

Setup State what we want to know.

Identify the *parameter* we wish to estimate. Here our parameter is the difference in the means, not the individual group means. Identify the variables and context.

Hypotheses State the null and alternative hypotheses.

We want to know whether people are likely to offer a different amount for a used camera when buying from a friend than when buying from a stranger. We wonder whether the difference between mean amounts is zero. We have bid prices from eight subjects buying from a friend and seven subjects buying from a stranger, found in a randomized experiment.

(*continued*)

The research claim is that friendship changes what people are willing to pay.[5] The natural null hypothesis is that friendship makes no difference.

We didn't start with any knowledge of whether friendship might increase or decrease the price, so we choose a two-sided alternative.

Make a graph. Boxplots are the display of choice for comparing groups. We'll also want to check the distribution of each group. Histograms may do a better job.

Looks like the prices are higher if you buy from a friend. The two ranges barely overlap, so we'll be pretty surprised if we don't reject the null hypothesis.

Model Think about the assumptions and check the conditions. (Because this is a randomized experiment, we haven't sampled at all, so the 10% Condition doesn't apply.)

State the sampling distribution model.

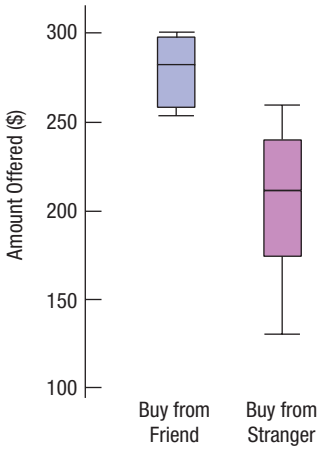
Specify the method.

H_0: The difference in mean price offered to friends and the mean price offered to strangers is zero:

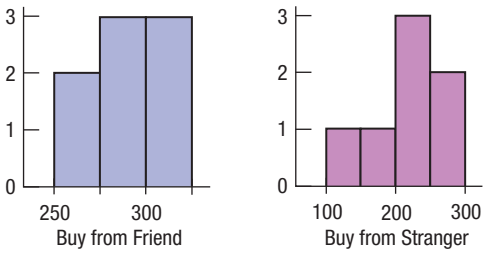
$$\mu_F - \mu_S = 0$$

H_A: The difference in mean prices is not zero:

$$\mu_F - \mu_S \neq 0$$



- ✓ **Independence Assumption**. There is no reason to think that the behaviour of one person will influence the behaviour of another.
- ✓ **Randomization Condition**. The experiment was randomized. Subjects were assigned to treatment groups at random.
- ✓ **Independent Groups Assumption**. Randomizing the experiment gives independent groups.
- ✓ **Nearly Normal Condition**. Histograms of the two sets of prices show no evidence of skewness or extreme outliers.



Because this is a randomized experiment with a null hypothesis of no difference in means, we can make the equal variance assumption. If, as we're assuming from the null hypothesis, the treatment doesn't change the means, then it's reasonable to assume that it also doesn't change the variances. Under these assumptions and conditions, we can use a Student's *t*-model to perform a pooled *t*-test.

[5]This claim is a good example of what is called a "research hypothesis" in many social sciences. The only way to check it is to deny that it's true and see where the resulting null hypothesis leads us.

DO

Mechanics List the summary statistics. Be sure to use proper notation.

From the data:

$$n_F = 8 \qquad n_S = 7$$
$$\bar{y}_F = \$281.88 \qquad \bar{y}_S = \$211.43$$
$$s_F = \$18.31 \qquad s_S = \$46.43$$

Use the null model to find the P-value. First determine the standard error of the difference between sample means.

The pooled variance estimate is

$$s_p^2 = \frac{(n_F - 1)s_F^2 + (n_S - 1)s_S^2}{n_F + n_S - 2}$$
$$= \frac{(8 - 1)(18.31)^2 + (7 - 1)(46.43)^2}{8 + 7 - 2}$$
$$= 1175.48.$$

The standard error of the difference becomes

$$SE_{pooled}(\bar{y}_F - \bar{y}_S) = \sqrt{\frac{s_p^2}{n_F} + \frac{s_p^2}{n_S}}$$
$$= 17.744.$$

Make a graph. Sketch the *t*-model centred at the hypothesized difference of zero. Because this is a two-tailed test, shade the region to the right of the observed difference and the corresponding region in the other tail.

The observed difference in means is

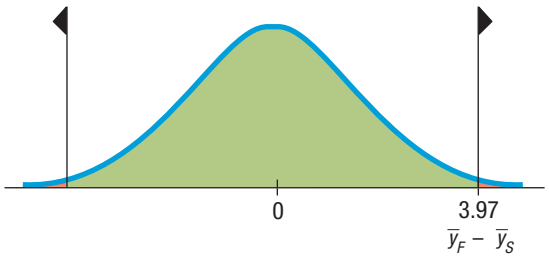
$$(\bar{y}_F - \bar{y}_S) = 281.88 - 211.43 = \$70.45,$$

which results in a *t*-ratio:

Find the *t*-value.

$$t = \frac{(\bar{y}_F - \bar{y}_S) - (0)}{SE_{pooled}(\bar{y}_F - \bar{y}_S)} = \frac{70.45}{17.744} = 3.97$$

A statistics program can find the P-value.



The computer output for a pooled *t*-test appears here.

Pooled T Test for friend vs. stranger

	N	Mean	StDev	SE Mean
Friend	8	281.9	18.3	6.5
Stranger	7	211.4	46.4	18

$t = 3.9699, df = 13, P\text{-value} = 0.001600$

Alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

32.11047 108.78238

(*continued*)

REPORT

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

Be cautious about generalizing to items whose prices are outside the range of those in this study. The confidence interval can reveal more detailed information about the size of the difference. In the original article (referenced in footnote 4 in this chapter), the researcher tested several items and proposed a model relating the size of the difference to the price of the items.

MEMO:

Re: Role of Friendship in Negotiations

Results of a small experiment show that people buying from a friend are likely to offer a different amount for a used camera than people buying from a stranger. The difference in mean offers was statistically significant ($P = .0016$).

The confidence interval suggests that people buying from a friend tend to offer more than people buying from a stranger. For the camera, the 95% confidence interval for the mean difference in price was $32.11 to $108.78, but we suspect that the actual difference may vary with the price of the item purchased.

Just Checking

Recall the experiment to see whether pictures of eyes would change compliance in voluntary contributions at an office coffee station.

3 What alternative hypothesis would you test?

4 The P-value of the test was less than 0.05. State a brief conclusion.

When Should You Use the Pooled *t*-Test?

When the variances of the two groups are in fact equal, the two methods give pretty much the same result. Pooled methods have a small advantage (slightly narrower confidence intervals, slightly more powerful tests) mostly because they usually have a few more degrees of freedom, but the advantage is slight. When the variances are *not* equal, the pooled methods are just not valid and can give poor results. You have to use the two-sample methods instead.

Because the advantages of pooling are small, and you're allowed to pool only rarely (when the equal variances assumption is met), our advice is: *Don't*.

It's never wrong *not* to pool.

As the sample sizes get bigger, the advantages that come from a few more degrees of freedom make less and less difference. So the advantage (such as it is) of the pooled method is greatest when the samples are small—just when it's hardest to check the conditions. And the difference in the degrees of freedom is greatest when the variances are not equal—just when you can't use the pooled method anyway. Our advice is to use the two-sample *t* methods to compare means.

Why did we devote a whole section to a method that we don't recommend using? That's a good question. The answer is that pooled methods are actually very important in Statistics, especially in the case of designed experiments, where we start by assigning subjects to treatments at random. We know that at the start of the experiment each treatment group is a random sample from the same population,[6] so each treatment group begins with the same population variance. In this case, assuming the variances are equal after we apply the treatment is the same as assuming that the treatment doesn't change the variance. When we test whether the true means are equal, we may be willing to go a bit further and say that the treatments made no difference *at all*. That's what we did in the friendship and bargaining experiment. Then it's not much of a stretch to assume that the variances have remained equal.

[6]That is, the population of experimental subjects. Remember that to be valid, experiments don't need a representative sample drawn from a population because we're not trying to estimate a population model parameter.

The other reason to discuss the pooled *t*-test is historical. Until recently, many software packages offered the pooled *t*-test as the default for comparing means of two groups and required you to specifically request the *two-sample* t-*test* (or sometimes the misleadingly named "unequal variance *t*-test") as an option. That's changing, but be careful to specify the right test when using software.

There's also a hypothesis test that you could use to test the assumption of equal variances. However, it's sensitive to failures of the assumptions and works poorly for small sample sizes—just the situation in which we might care about a difference in the methods. When the choice between two-sample *t* and pooled *t*-methods makes a difference (i.e., when the sample sizes are small), the test for whether the variances are equal hardly works at all.

Even though pooled methods are important in Statistics, the ones for comparing two means have good alternatives that don't require the extra assumption. The two-sample methods apply to more situations and are safer to use.

For Example The pooled *t*-test

Question: Would the owner of the car dealership on page xxx have reached a different conclusion had he used the pooled *t*-test for the difference in mean discounts?

Answer: The difference in the pooled *t*-test is that it assumes that the variances of the two groups are equal, and pools those two estimates to find the standard error of the difference. The pooled estimate of the common standard deviation is

$$s_{pooled} = \sqrt{\frac{(n_{Women} - 1)s^2_{Women} + (n_{Men} - 1)s^2_{Men}}{n_{Women} + n_{Men} - 2}} = \sqrt{\frac{(46 - 1)(399.70)^2 + (54 - 1)(458.95)^2}{46 + 54 - 2}}.$$
$$= 432.75$$

We use that to find the *SE* of the difference:

$$SE_{pooled}(\bar{y}_{Women} - \bar{y}_{Men}) = s_{pooled}\sqrt{\frac{1}{n_{Women}} + \frac{1}{n_{Men}}} = 432.75\sqrt{\frac{1}{46} + \frac{1}{54}} = \$86.83$$

(Without pooling, our estimate was \$85.87.) This pooled *t* has 46 + 54 – 2 = 98 degrees of freedom.

$$t_{98} = \frac{\bar{y}_{Women} - \bar{y}_{Men}}{SE_{pooled}(\bar{y}_{Women} - \bar{y}_{Men})} = \frac{263.65}{86.83} = 3.04$$

The P-value for a *t* of 3.04 with 98 degrees of freedom is (from technology) 0.0030. The two-sample *t*-test value was 0.0028.

There is no practical difference between these, and we reach the same conclusion that the difference is not 0. The assumption of equal variances did not affect the conclusion.

LO ❸

14.6 Paired Data

The two-sample *t*-test depends crucially on the assumption that the cases in the two groups are independent of each other. When is that assumption violated? Most commonly, it's when we have data on the *same* cases in two different circumstances. For example, we might want to compare *the same* customers' spending at our website last January with this January, or we might have *each participant* in a focus group rate two different product designs. Data such as these are called **paired data** because we have two items of data from the same subject**.**

Pairing isn't a problem; it's an opportunity. If you know the data are paired, you can take advantage of the pairing—in fact, you *must* take advantage it. You *should not* use the two-sample (or pooled two-sample) method when the data are paired. Be careful. There is no statistical test to determine whether the data are paired. You must decide whether the data are paired from understanding how they were collected and what they mean (check the W's).

Once we recognize that our data are matched pairs, it makes sense to concentrate on the *difference* between the two measurements for each individual. That

is, we look at the collection of pairwise differences in the measured variable. For example, if studying customer spending, we would analyze the *difference* between this January's and last January's spending for each customer. Because it's the *differences* we care about, we can treat them as if there was a single variable of interest holding those differences. With only one variable to consider, we can use a simple one-sample *t*-test. A **paired *t*-test** is just a one-sample *t*-test for the mean of the pairwise differences. The sample size is the number of pairs.

Paired Data Assumption

The data must actually be paired. You can't just decide to pair data from independent groups. When you have two groups with the same number of observations, it may be tempting to match them up, but that's not valid. You can't pair data just because they "seem to go together." To use paired methods, you must determine from knowing how the data were collected whether the two groups are paired or independent. Usually the context will make it clear.

Be sure to recognize paired data when you have it. Remember, two-sample *t* methods aren't valid unless the groups are independent, and paired groups aren't independent.

Independence Assumption

For these methods, it's the *differences* that must be independent of each other. This is just the one-sample *t*-test assumption of independence, now applied to the differences. In our example, one rater's opinion shouldn't affect how another person rated the colas. As always, randomization helps to ensure independence.

Randomization Condition

Randomness can arise in many ways. The *pairs* may be a random sample. For example, we may be comparing opinions of husbands and wives from a random selection of couples. 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 like this one, 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. What we want to know usually focuses our attention on where the randomness should be.

10% Condition

When we sample from a finite population, we should be careful not to sample more than 10% of that population. Sampling too large a fraction of the population calls the Independence Assumption into question. Here, we can regard our sample customers as representative of the (potentially very large) population of all customers. As with other quantitative data situations, we don't usually explicitly check the 10% Condition, but make sure to think about it.

Normal Population Assumption

We need to assume that the population of *differences* follows a Normal model. We don't need to check the data in each of the two individual groups. In fact, the data from each group can be quite skewed, but the differences can still be unimodal and symmetric.

- **Nearly Normal Condition** This condition can be checked with a histogram of the differences. As with the one-sample *t* methods, this assumption matters less as we have more pairs 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 Paired data

Question: The owner of the car dealership is curious about the maximum discount his salespeople are willing to give to customers. In particular, two of his salespeople, Frank and Nikita, seem to have very different ideas about how much discount to allow. To test his suspicion, he selects 30 cars from the lot and asks Frank and Nikita to each say how much discount they would allow a customer. If we want to test whether the mean discount they each give is the same, what test would he use? Explain.

Answer: A paired *t*-test. The responses (maximum discount in $) aren't independent between the two salespeople because they're evaluating the same 30 cars.

LO ❸

14.7 The Paired *t*-Test

The paired *t*-test is mechanically a one-sample *t*-test. We treat the differences as our variable. We simply compare the mean difference to its standard error. If the *t*-statistic is large enough, we reject the null hypothesis.

Paired *t*-Test

When the conditions are met, we're ready to test whether the mean paired difference is significantly different from a hypothesized value (called Δ_0). 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

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

where s_d is the mean of the pairwise differences, n is the number of pairs, and

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

where s_d is the standard deviation of the pairwise differences.

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

Similarly, we can construct a confidence interval for the true difference. As in a one-sample *t*-interval, we centre our estimate at the mean difference in our data. The margin of error on either side is the standard error multiplied by a critical *t*-value (based on our confidence level and the number of pairs we have).

Paired *t*-Confidence Interval

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

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

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

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

When data are paired, the *t*-test can sometimes see differences that a two-sample *t*-test can't. A paired design has roughly half the degrees the freedom of the two-sample *t*-test. Typically, we'd want *more* degrees of freedom, but usually pairing more than compensates for this by reducing the variation.

Unfortunately, you can't take the benefit of pairing unless the data are actually paired. If you're designing the study, you may be able to arrange for the data to be paired (before vs. after; using the same people to test two different methods; tracking the same customers in two different months, etc.). But be careful. If the data from the two groups are independent, you may not pair them just because the groups have the same number of observations. There must be a link that you can identify and justify between the pairs. Data collected before and after some event on the same people, companies, or subjects are naturally paired.

Just Checking

Think about each of the following situations. Would you use a two-sample t or paired t method (or neither)? Why?

5 Random samples of 50 men and 50 women are surveyed on the amount they invest on average in the stock market on an annual basis. We want to estimate any gender difference in how much they invest.

6 Random samples of students were surveyed on their perception of ethical and community service issues in both their first year and their fourth year at a university. The university wants to know whether its required programs in ethical decision making and service learning change student perceptions.

7 A random sample of work groups within a company was identified. Within each work group, one male and one female worker were selected at random. Each was asked to rate the secretarial support that their work group received. When rating the same support staff, do men and women rate them equally, on average?

8 A total of 50 companies are surveyed about business practices. They are categorized by industry, and we wish to investigate differences across industries.

9 These same 50 companies are surveyed again one year later to see if their perceptions, business practices, and R&D investment have changed.

Guided Example — Seasonal Spending in Canada

In Canada,[7] during the period 2001–2007, retail sales for December were 16%–21% higher than the average for the rest of the year. In 2008 the recession hit and December sales were only 7% higher. In each of those years, recession or not, there was a sharp drop from December to January of 23%–29%.

We have a sample of cardholders from a particular market segment and the amount they charged on their credit cards in December 2010 and in January 2011. (There were 1000 cardholders in the original sample, but 89 of them had at least one month missing, leaving a sample of $n = 911$.) We can create a paired t-confidence interval to estimate the true mean difference in spending between the two months.

Because credit card companies receive a percentage of each transaction, they need to forecast how much the average spending will increase or decrease from month to month. How much less do people tend to spend in January than in December? For any particular segment of cardholders, a credit card company could select two random samples—one for each month—and simply compare the average amount spent in January with that in December. A more sensible approach might be to select a single random sample and compare the spending between the two months for *each cardholder*. Designing the study in this way and examining the paired differences give a more precise estimate of the actual change in spending.

WHO	Cardholders in a particular market segment of a major credit cards issuer
WHAT	Amount charged on their credit cards in December and January
WHEN	2010–2011
WHERE	Canada
WHY	To estimate the amount of decrease in spending one could expect after the holiday shopping season

[7] **Source:** Based on Statistics Canada. (2009). Canada—Total retail trade, all stores, unadjusted. V43973511.

PLAN

Setup State what we want to know.

Identify the *parameter* we wish to estimate and the sample size.

We want to know how much we can expect credit card charges to change, on average, from December to January for this market segment. We have the total amount charged in December 2010 and in January 2011 for $n = 911$ cardholders in this segment. We want to find a confidence interval for the true mean difference in charges between these two months for all cardholders in this segment. Because we know that people tend to spend more in December, we'll look at the difference: *December spend—January spend*. A positive difference will mean a decrease in spending.

Model Check the conditions.

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

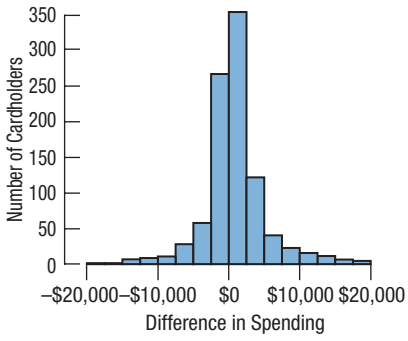
- ✓ **Paired Data Assumption:** The data are paired because they are measurements on the same cardholders in two different months.

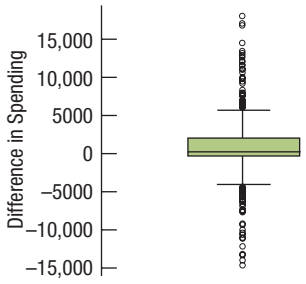
Think about what we hope to learn and where the randomization comes from.

- ✓ **Independence Assumption:** The behaviour of any individual is independent of the behaviour of the others, so the differences are mutually independent.
- ✓ **Randomization Condition:** This was a random sample from a large market segment.

Make a picture of the differences. Don't plot separate distributions of the two groups—that would entirely miss 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.

- ✓ **Nearly Normal Condition:** The distribution of the differences is unimodal and symmetric. Although there are many observations nominated by the boxplot as outliers, the distributions are symmetric. (This is typical of the behaviour of credit card spending.) There are no isolated cases that would unduly dominate the mean difference, so we'll leave all observations in the study.





Specify the sampling distribution model.

Choose the method.

The conditions are met, so we'll use a Student's *t*-model with $(n - 1) = 910$ degrees of freedom and find a **paired *t*-confidence interval.**

DO

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

$\bar{d}$ is the mean difference.

s_d is the standard deviation of the differences.

Make a picture. Sketch a t-model centred at the observed mean of 788.18.

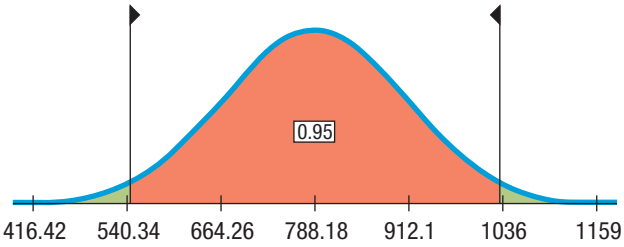
Find the standard error and the t-score of the observed mean difference. There is nothing new in the mechanics of the paired t methods. These are the mechanics of the t-interval for a mean applied to the differences.

The computer output tells us:

$n = 911$ pairs

$\bar{d} = \$788.18$

$s_d = 3740.22$



We estimate the standard error of $\bar{d}$ using:

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{3740.22}{\sqrt{911}} = \$123.919$$

$$t^*_{910} = 1.96$$

The margin of error is $ME = t^*_{910} \times SE(\bar{d})$

$= 1.96 \times 123.919 = 242.88$

So a 95% CI is $\bar{d} \pm ME = (\$545.30, \$1031.06)$.

REPORT

Conclusion Link the results of the confidence interval to the context of the problem.

MEMO:

Re: Credit Card Expenditure Changes

In the sample of cardholders studied, the change in expenditures between December and January averaged $788.18, which means that, on average, cardholders spend $788.18 *less* in January than the month before. Although we didn't measure the change for all cardholders in the segment, we can be 95% confident that the true *mean* decrease in spending is between $545.30 and $1031.06.

For Example The paired t-test

Here are some summary statistics from the study undertaken by the car dealership owner (see page xxx):

$\bar{y}_{Frank} = \$414.48$ $\bar{y}_{Nikita} = \$478.88$

$SD_{Frank} = \$87.33$ $SD_{Nikita} = \$175.12$

$\bar{y}_{Diff} = \$64.40$ $SD_{Diff} = \$146.74$

Question: Test the hypothesis that the mean maximum discount that Frank and Nikita would give is the same. Give a 95% confidence interval for the mean difference.

Answer: We use a paired *t*-test because Frank and Nikita were asked to give opinions about the same 30 cars:

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

$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = \frac{146.74}{\sqrt{30}} = \$26.79$$

$$t_{29} = \frac{64.40}{26.79} = 2.404$$

which has a (two-side) P-value of 0.0228. We reject the null hypothesis that the mean difference is 0 and conclude that there's strong evidence to suggest that they're not the same.

A 95% confidence interval,

$$t^*_{29} = 2.045 \text{ at } 95\% \text{ confidence}$$

$$\bar{d} \pm t^*_{29} \times SE(\bar{d}) = 64.40 \pm 2.045 \times 26.79$$

$$= (\$9.61, \$119.19),$$

shows that Nikita gives, on average, somewhere between $9.61 and $119.19 more for her maximum discount than Frank does.

What Can Go Wrong?

- **Watch out for paired data.** The Independent Groups Assumption deserves special attention. Some researchers deliberately violate this assumption. For example, suppose you wanted to test a diet program. You select 10 people at random to take part in your diet. You measure their weights at the beginning of the diet and after 10 weeks of the diet. So you have two columns of weights, one for *before* and one for *after*. Can you use these methods to test whether the mean has gone down? No! The data are related; each "*after*" weight goes naturally with the "before" weight for the *same* person. If the samples are *not* independent, you can't use two-sample methods. This is probably the main thing that can go wrong when using these two-sample methods. Certainly, someone's weight before and after the 10 weeks will be related (whether the diet works or not). The methods of this chapter can be used *only* if the observations in the two groups are *independent*.

- **Don't use individual confidence intervals for each group to test the difference between their means.** If you make 95% confidence intervals for the means of two groups separately and you find that the intervals don't overlap, you can reject the hypothesis that the means are equal (at the corresponding α level). But if the intervals do overlap, it doesn't mean that you *can't* reject the null hypothesis. The margin of error for the difference between the means is smaller than the sum of the individual confidence interval margins of error. Comparing the individual confidence intervals is like adding the standard deviations. But we know that it's the variances that we add, and when we do it right, we actually get a more powerful test. So don't test the difference between group means by looking at separate confidence intervals. Always make a two-sample *t*-interval or perform a two-sample *t*-test.

- **Look at the plots.** The usual (by now) cautions about checking for outliers and non-Normal distributions apply. The simple defence is to make and examine boxplots. You may be surprised at how often this simple step saves you from the

wrong or even absurd conclusions that can be generated by a single undetected outlier. You don't want to conclude that two methods have very different means just because one observation is atypical.

- **Don't use a paired *t*-method when the samples aren't paired.** When two groups don't have the same number of values, it's easy to see that they can't be paired. But just because two groups have the same number of observations doesn't mean they can be paired, even if they're 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 can't *choose* which method to use based on your preferences. Only if the data are from an experiment or study in which observations were paired can you use a paired method.
- **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).

Ethics In Action

Canada's Best Diversity Employers 2011
PRESENTED BY BMO Financial Group

The Canada's Best Diversity Employers program, funded by BMO Financial Group, recognizes employers that have exceptional diversity and inclusiveness programs covering (a) women, (b) members of visible minorities, (c) persons with disabilities, (d) Aboriginal peoples, and (e) lesbian, gay, bisexual, and transgendered/transsexual (LGBT) peoples. The evaluation of employers follows an "Equity Continuum" developed by the Toronto-based consulting firm TWI Inc.

Advocacy groups for equity and diversity in the workplace often cite middle managers as an obstacle to many organizations' efforts to be more inclusive. In response to this concern, Li Xue, the CEO for a mining company, asked Mohammed Mehri, VP of Human Resources, to look into the possibility of instituting some type of diversity training for the company's middle managers.[8] One option under consideration was an online education program that focused on cultural diversity, gender equity, and disability awareness. Mohammed suspected that an online program, although cost-effective, would not be as effective as traditional training for middle managers. In order to evaluate the program under consideration, 20 middle managers were selected to participate. Before beginning, they were given a test to assess their knowledge of and sensitivity to various diversity and equity issues. Out of a possible perfect score of 100, the test average was 63.65. Each of the 20 managers then completed the six-week online diversity education program and was retested. The average on the test after completing the online program was 69.15. Although the group achieved a higher mean test score after completing the program, the two-sample ***t***-test revealed that this average test score wasn't significantly higher than the average prior to completing the online program ($t = -0.94$, P-value = 0.176). Mohammed wasn't surprised, and began to explore more traditional diversity education programs.

ETHICAL ISSUE: *The pre-test and post-test designs violate independence and therefore the two-sample* t*-test is not appropriate (related to Item A, ASA Ethical Guidelines; see Appendix C, the American Statistical Association's Ethical Guidelines for Statistical Practice, also available online at www.amstat.org/about/ethicalguidelines.cfm).*

ETHICAL SOLUTION: *Use the correct test. The two-sample* t*-test is not appropriate for these data. Using the correct test shows that the online diversity education program was effective.*

[8]**Sources:** Based on TWI Inc. Retrieved from www.twiinc.com/about-twi.html; and Mediacorp Canada Inc. (2011). *Canada's Best Diversity Employers*. Retrieved from www.canadastop100.com/diversity

What Have We Learned?

Learning Objectives

❶ Know how to test whether the difference in the means of two independent groups is equal to some hypothesized value.

- The two-sample *t*-test is appropriate for independent groups. It uses a special formula for degrees of freedom.
- The assumptions and conditions are the same as for one-sample inferences for means, with the addition of assuming that the groups are independent of each other.
- The most common null hypothesis is that the means are equal.

❷ Be able to construct and interpret a confidence interval for the difference between the means of two independent groups.

- The confidence interval inverts the *t*-test in the natural way.

❷ Know how and when to use pooled *t* inference methods.

- There is an additional assumption that the variances of the two groups are equal.
- This may be a plausible assumption in a randomized experiment.

❸ Recognize when you have paired or matched samples and use an appropriate inference method.

- Paired *t* methods are the same as one-sample *t* methods applied to the pairwise differences.
- If data are paired they cannot be independent, so two-sample *t* and pooled-*t* methods would not be applicable.

Terms

Paired data

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. Pairing in experiments is a form of blocking and arises in other contexts. Pairing in observational and survey data is a form of matching.

Paired *t*-test

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 $t = \dfrac{\bar{d} - \Delta_0}{SE(\bar{d})}$ with $n - 1$ degrees of freedom, where $SE(\bar{d}) = \dfrac{S_d}{\sqrt{n}}$ and n is the number of pairs.

Paired *t*-confidence interval

A confidence interval for the mean of the pairwise differences between paired groups found as $\bar{d} \pm t^*_{n-1} \times SE(\bar{d})$, where $SE(\bar{d}) = \dfrac{S_d}{\sqrt{n}}$ and n is the number of pairs.

***Pooled *t*-confidence interval**

A confidence interval for the difference in the means of two independent groups used when we are willing and able to make the additional assumption that the variances of the groups are equal. It is found as:

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE_{pooled}(\bar{y}_1 - \bar{y}_2),$$

$$\text{where } SE_{pooled}(\bar{y}_1 - \bar{y}_2) = S_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

*Pooling data from two or more populations may sometimes be combined, or *pooled*, to estimate a statistic (typically a pooled variance) when we're willing to assume that the estimated value is the same in both populations. The resulting larger sample size may lead to an estimate with lower sample variance. However, pooled estimates are appropriate only when the required assumptions are true.

and the pooled variance is

$$S^2_{pooled} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)}.$$

The number of degress of freedom is $(n_1 - 1) + (n_2 - 1)$.

***Pooled *t*-test**

A hypothesis test for the difference in the means of two independent groups when we are willing and able to assume that the variances of the groups are equal. It tests the null hypothesis

$$H_0 : \mu_1 - \mu_2 = \Delta_0$$

where the hypothesized difference Δ_0 is almost always 0, using the statistic

$$t_{df} = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE_{pooled}(\bar{y}_1 - \bar{y}_2)},$$

where the pooled standard error is defined as for the pooled interval and the degrees of freedom is $(n_1 - 1) + (n_2 - 1)$.

Two-sample *t*-interval

A confidence interval for the difference in the means of two independent groups found as

$$(\bar{y}_1 - \bar{y}_2) \pm t^*_{df} \times SE(\bar{y}_1 - \bar{y}_2), \text{ where}$$

$$SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

and the number of degrees of freedom is given by the approximation formula in footnote 2 of this chapter, or with technology.

Two-sample *t*-test

A hypothesis test for the difference in the means of two independent groups. It tests the null hypothesis

$$H_0 : \mu_1 - \mu_2 = \Delta_0,$$

where the hypothesized difference Δ_0 is almost always 0, using the statistic

$$t_{df} = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)},$$

with the number of degrees of freedom given by the approximation formula in footnote 2 of this chapter, or with technology.

Skills

Plan

- Be able to recognize situations in which we want to do inference on the difference between the means of two independent groups.
- Know how to examine your data for violations of conditions that would make inference about the difference between two population means unwise or invalid.
- Be able to recognize when a pooled *t* procedure might be appropriate, and be able to explain your reasons if you decide to use a two-sample method anyway.
- Recognize whether a design that compares two groups is paired or not.

Do

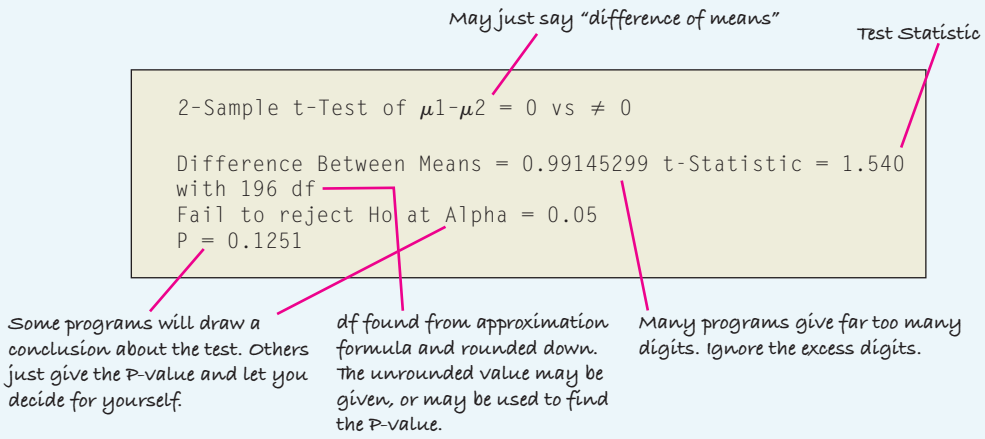
- Be able to perform a two-sample *t*-test using a statistics package or calculator (at least for finding the degrees of freedom).
- Know how to find a paired confidence interval, recognizing that it is mechanically equivalent to doing a one-sample *t*-interval applied to the differences.
- Be able to perform a paired *t*-test, recognizing that it is mechanically equivalent to a one-sample *t*-test applied to the differences.

Report

- Be able to interpret a test of the null hypothesis that the means of two independent groups are equal. (If the test is a pooled *t*-test, your interpretation should include a defence of your assumption of equal variances.)
- Know how 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.
- Know how to interpret a paired *t*-interval, recognizing that it gives an interval for the mean difference in the pairs.

Technology Help: Comparing Two Means

Here's some typical computer package output with comments:



Most statistics packages compute the test statistic for you and report a P-value corresponding to that statistic. And statistics packages make it easy to examine the boxplots of the two groups, so you have no excuse for skipping the important check of the Nearly Normal Condition.

Some statistics software packages automatically try to test whether the variances of the two groups are equal, and some software packages automatically offer both the two-sample *t* and pooled *t* results. Ignore the test for the variances; it has little power in any situation in which its results could matter. If the pooled and two-sample methods differ in any important way, you should stick with the two-sample method. Most likely, the Equal Variance Assumption needed for the pooled method has failed.

The degrees of freedom approximation usually gives a fractional value. Most packages seem to round the approximate value down to the next smallest integer (although they may actually compute the P-value with the fractional value, gaining a tiny amount of power).

There are two ways to organize data when we want to compare two independent groups. The first, called *unstacked data*, lists the data in two columns, one for each group. Each list can be thought of as a variable. In this method, the variables in the credit card example would be "Offer" and "No Offer." Graphing calculators usually prefer this form, and some computer programs can use it as well.

The alternative way to organize the data is as *stacked data*. What is the response variable for the credit card experiment? It's the "Spend Lift"—the amount by which customers increased their spending. But the values of this variable in the unstacked lists are in both columns, and actually there's an experiment factor here, too—namely, whether the customer was offered the promotion or not. So we could put the data into two different columns, one with the "Spend Lifts" in it and one with a "Yes" for those who were offered the promotion and a "No" for those who weren't. The stacked data would look like this:

Spend Lift	Offer
969.74	Yes
915.04	Yes
197.57	No
77.31	No
196.27	Yes
. . .	. . .

This way of organizing the data makes sense as well. Now the factor and the response variables are clearly visible. You'll have to see which method your program requires. Some packages even allow you to structure the data either way.

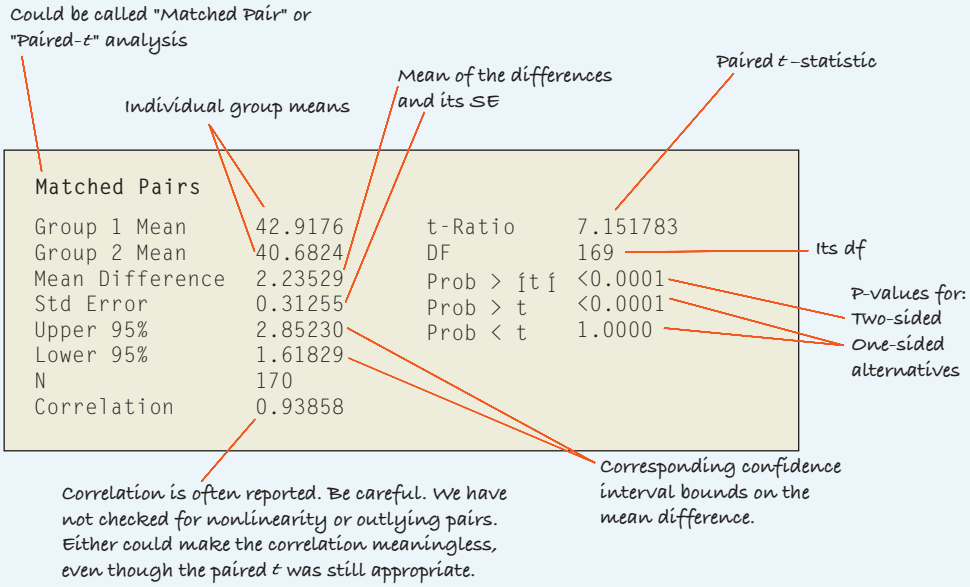
The commands to do inference for two independent groups on common statistics technology are not always found in obvious places. Here are some starting guidelines.

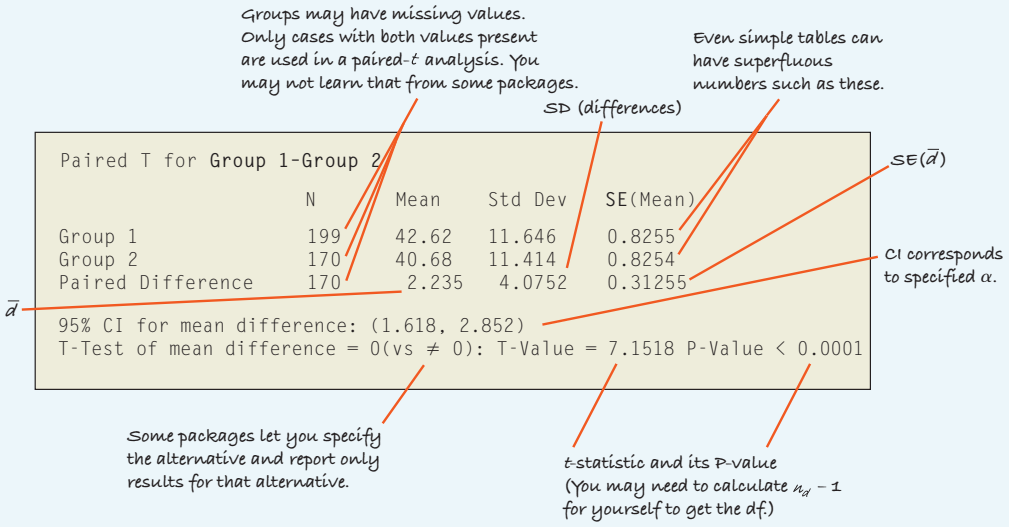
Most statistics programs can compute paired *t* analyses. Some may want you to find the differences yourself and use the one-sample *t* methods. Those that perform the entire procedure will need to know the two variables to compare. The computer, of course, can't verify that the variables are naturally 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's missing a value for either variable. 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. Below is a generic example with comments.

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.

Computers make it easy to examine the boxplots 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 that pairing says nothing about whether the scatterplot should be straight. That doesn't matter for our t methods.)





EXCEL

From the **Data** tab, **Analysis Group,** choose **Data Analysis.** Alternatively (if the Data Analysis Tool Pack isn't installed), in the **Formulas** tab, choose **More functions** > **Statistical** > **T.TEST,** and specify Type = 1, 2 or 3 in the resulting dialogue.

Fill in the cell ranges for the two groups, the hypothesized difference, and the alpha level.

Comments

Excel expects the two groups to be in separate cell ranges. Notice that, contrary to Excel's wording, we don't need to assume that the variances are *not* equal; we simply choose not to assume that they *are* equal.

MINITAB

From the **Stat** menu, choose the **Basic Statistics** submenu. From that menu, choose **2-sample *t* or Paired *t*...** Then fill in the dialogue boxes for the two paired samples, or fill in the summary data for the differences.

Comments

Minitab takes "First sample" minus "Second sample."

SPSS

From the **Analyze** menu, choose the **Compare Means** submenu. From that, choose the **Independent-Samples *t*-test or Paired-Samples *t*-test** command. Select pairs of variables to compare, and click the arrow to add them to the selection box.

Comments

You can compare several pairs of variables at once. Options include the choice to exclude cases missing in any pair from all tests.

JMP

From the **Analyze menu,** select **Fit *y* by *x*.** Select variables: a **Y, Response** variable that holds the data and an ***X*, Factor** variable that holds the group names. JMP will make a dotplot. Click the **red triangle** in the dotplot title, and choose **Unequal variances**. The *t*-test is at the bottom of the resulting table. Find the P-value from the Prob > F section of the table (they're the same).

From the **Analyze** menu, select **Matched Pairs**. Specify the columns holding the two groups in the **Y Paired Response** dialogue. Click **OK**.

Comments

JMP expects data in one variable and category names in the other. Don't be misled: There's no need for the variances to be unequal in order to use two-sample *t* methods.

Mini case studies

Real Estate

Larger homes generally fetch a higher price than smaller homes. How much can we learn about a house from the fact that it has a fireplace or more than the average number of bedrooms? Data for a random sample of 1047 homes can be found in the file **ch14_MCSP_Real_Estate.** There are six quantitative variables: *Price* ($), *Living Area* (sq. ft.), *Bathrooms* (#), *Bedrooms* (#), *Lot Size* (acres), and *Age* (years), and one categorical variable, *Fireplace*? (1 = Yes; 0 = No), denoting whether the house has at least one fireplace. We can use *t* methods to see, for example, whether homes with fireplaces sell for more, on average, and by how much. For the quantitative variables, create new categorical variables by splitting them at the median or some other splitting point of your choice, and compare home prices above and below this value. For example, the median number of *Bedrooms* of these homes is two. You might compare the prices of homes with one or two bedrooms to those with more than two. Write up a short report summarizing the differences in mean price based on the categorical variables you created.

Rate Your Professors in Canada

Rate My Professors[9] records the opinions of students about their professors according to three criteria: clarity, easiness, and helpfulness. You and your friend want to choose between two professors for your Statistics class, and you see that the average ratings of these professors are as follows, together with the number of students who have provided those ratings.

[9]www.RateMyProfessors.com/SearchSchool.jsp?Country=1

	Number of Students	Clarity	Easiness	Helpfulness
Professor X	12	4.2	4.5	3.3
Professor Y	47	4.0	4.1	3.6

Like many organizations providing ratings and evaluations, Rate My Professors provides an average rating without any measure of the extent of variation among the students from which that average was calculated. You estimate the standard deviations in the ratings as 10% of the averages in the table (e.g., you estimate the standard deviation of Prof X's Clarity rating to be $0.1 \times 4.2 = 0.42$), whereas your friend estimates them as 20% of the average. (a) Do these different assumptions result in you and your friend coming to different conclusions (at the 0.05 significance level) about which professor is better on the three criteria given? (Give three answers, one for each criterion.) (b) Does it make any difference if you pool your estimates of the standard deviations? (Give three answers, one for each criterion.) (c) What percentage would the standard deviation need to be in relation to the mean for you to be just unable to distinguish between the professors at the 90% level? (Give six answers, with two significant figures of accuracy: one for each criterion in the pooled and not pooled situations. *Hint:* Use a trial-and-error approach with software.)

Consumer Spending Patterns

You're on the financial planning team for monitoring a high-spending segment of a credit card. You know that customers tend to spend more during December before the holidays, but you're not sure about the pattern of spending in the months after the holidays. Look at the data set **ch14_MCSP_Consumer_Spending**. It contains the monthly credit card spending of 1200 customers during the months December, January, February, and March. Report on the spending differences between the months. If you'd failed to realize that these are paired data, what difference would that have made in your reported confidence intervals and tests?

Exercises

SECTION 14.1

1. A developer wants to know if the houses in two different neighbourhoods were built at roughly the same time. She takes a random sample of six houses from each neighbourhood and finds their ages from local records. The table shows the data for each sample (in years).

Neighbourhood 1	Neighbourhood 2
57	50
60	43
46	35
62	53
67	46
56	55

a) Find the sample means for each neighbourhood.
b) Find the estimated difference of the mean ages of the two neighbourhoods.
c) Find the sample variances for each neighbourhood.
d) Find the sample standard deviations for each neighbourhood.
e) Find the standard error of the difference of the two sample means. **LO ❶**

2. A market analyst wants to know if the new website he designed is showing increased page views per visit. A customer is randomly sent to one of two different websites, which offer the same products but with different designs.

Here are the page views of five randomly chosen customers for each website:

Website A	Website B
9	10
4	13
14	2
7	3
2	7

a) Find the sample mean page views for each website.
b) Find the estimated difference of the sample mean page views of the two websites.
c) Find the sample variances for each website.
d) Find the sample standard deviations for each website.
e) Find the standard error of the difference of the sample means. **LO 1**

3. The developer in Exercise 1 hires an assistant to collect a random sample of houses from each neighbourhood and finds that the summary statistics for the two neighbourhoods look as follows:

Neighbourhood 1	Neighbourhood 2
$n_1 = 30$	$n_2 = 35$
$\bar{y}_1 = 57.2$ yrs	$\bar{y}_2 = 47.6$ yrs
$S_1 = 7.51$ yrs	$S_2 = 7.85$ yrs

a) Find the estimated mean age difference between the two neighbourhoods.
b) Find the standard error of the estimated mean difference.
c) Calculate the *t*-statistic for the observed difference in mean ages, assuming that the true mean difference is 0. **LO 1**

4. Not happy with the previous results, the analyst in Exercise 2 takes a much larger random sample of customers from each website and records their page views. Here are the data:

Website 1	Website 2
$n_1 = 80$	$n_2 = 95$
$\bar{y}_1 = 7.7$ yrs	$\bar{y}_2 = 7.3$ yrs
$S_1 = 4.6$ yrs	$S_2 = 4.3$ yrs

a) Find the estimated mean difference in page visits between the two websites.
b) Find the standard error of the estimated mean difference.
c) Calculate the *t*-statistic for the observed difference in mean page visits, assuming that the true mean difference is 0. **LO 1**

SECTIONS 14.2 AND 14.3

5. For the data in Exercise 1, we want to test the null hypothesis that the mean age of houses in the two neighbourhoods is the same. Assume that the data come from a population that is Normally distributed.

a) Using the values you found in Exercise 1, find the value of the *t*-statistic for the difference in mean ages for the null hypothesis.
b) Calculate the degrees of freedom from the formula in footnote 2 on page xxx.
c) Calculate the degrees of freedom using the rule that df = min $(n_1 - 1, n_2 - 1)$.
d) Find the P-value using the degrees of freedom from part b) (You can either round the number of df and use a table or use technology or a website).
e) Find the P-value using the degrees of freedom from part d)
f) What do you conclude at $\alpha = 0.05$? **LO 1**

6. For the data in Exercise 2, we want to test the null hypothesis that the mean number of page visits is the same for the two websites. Assume that the data come from a population that is Normally distributed.

a) Using the values you found in Exercise 2, find the value of the *t*-statistic for the difference in mean ages for the null hypothesis.
b) Calculate the degrees of freedom from the formula in the footnote on page xxx.
c) Calculate the degrees of freedom using the rule that df = min $(n_1 - 1, n_2 - 1)$.
d) Find the P-value using the degrees of freedom from part b) (You can either round the number of df and use a table or use technology or a website).
e) Find the P-value using the degrees of freedom from part c)
f) What do you conclude at $\alpha = 0.05$? **LO 1**

7. Using the data in Exercise 3, test the hypothesis that the mean ages of houses in the two neighbourhoods is the same. You may assume that the ages of houses in each neighbourhood follow a Normal distribution.

a) Calculate the P-value of the statistic, knowing that the approximation formula gives 53.4 df (you'll need to round or to use technology).
b) Calculate the P-value of the statistic using the rule that df is at least min$(n_1 - 1, n_2 - 1)$.
c) What do you conclude at $\alpha = 0.05$? **LO 1**

8. Using the data in Exercise 4, test the hypothesis that the mean number of page views from the two websites is the same. You may assume that the number of page views from each website follow a Normal distribution.

a) Calculate the P-value of the statistic, knowing that the approximation formula gives 148.0 df.
b) Calculate the P-value of the statistic using the rule that df is at least min $(n_1 - 1, n_2 - 1)$.
c) What do you conclude at $\alpha = 0.05$? **LO 1**

SECTION 14.4

9. Using the data in Exercise 1, and assuming that the data come from a distribution that is Normally distributed,

a) Find a 95% confidence interval for the mean difference in ages of houses in the two neighbourhoods.
b) Is 0 within the confidence interval?
c) What does it say about the null hypothesis that the mean difference is 0? **LO ❷**

10. Using the data in Exercise 2, and assuming that the data come from a distribution that is Normally distributed,

a) Find a 95% confidence interval for the mean difference in page views from the two websites.
b) Is 0 within the confidence interval?
c) What does it say about the null hypothesis that the mean difference is 0? **LO ❷**

11. Using the summary statistics in Exercise 3, and assuming that the data come from a distribution that is Normally distributed,

a) Find a 95% confidence interval for the mean difference in ages of houses in the two neighbourhoods using the df given in Exercise 7.
b) Why is the confidence interval narrower than the one you found in Exercise 9?
c) Is 0 within the confidence interval?
d) What does it say about the null hypothesis that the mean difference is 0? **LO ❷**

12. Using the summary statistics in Exercise 4, and assuming that the data come from a distribution that's Normally distributed,

a) Find a 95% confidence interval for the mean difference in page views from the two websites.
b) Why is the confidence interval narrower than the one you found in Exercise 10?
c) Is 0 within the confidence interval?
d) What does it say about the null hypothesis that the mean difference is 0? **LO ❷**

SECTION 14.5

13. For the data in Exercise 1,

a) Test the null hypothesis at $\alpha = 0.05$ using the pooled t-test. (Show the t-statistic, P-value, and conclusion.)
b) Find a 95% confidence interval using the pooled degrees of freedom.
c) Are your answers different from what you previously found in Exercise 9? Explain briefly why or why not. **LO ❷**

14. For the data in Exercise 2,

a) Test the null hypothesis $\alpha = 0.05$ using the pooled t-test. (Show the t-statistic, P-value, and conclusion.)
b) Find a 95% confidence interval using the pooled degrees of freedom.
c) Are your answers different from what you previously found in Exercise 10? Explain briefly why or why not. **LO ❷**

15. For the data in Exercise 3,

a) Test the null hypothesis $\alpha = 0.05$ using the pooled t-test. (Show the t-statistic, P-value, and conclusion.)
b) Find a 95% confidence interval using the pooled degrees of freedom.
c) Are your answers different from what you previously found in Exercise 11? Explain briefly why or why not. **LO ❷**

16. For the data in Exercise 4,

a) Test the null hypothesis $\alpha = 0.05$ using the pooled t-test. (Show the t-statistic, P-value, and conclusion.)
b) Find a 95% confidence interval using the pooled degrees of freedom.
c) Are your answers different from what you previously found in Exercise 12? Explain briefly why or why not. **LO ❷**

SECTION 14.6

17. For each of the following scenarios, say whether the data should be treated as independent or paired samples. Explain briefly. If paired, explain what the pairing involves.

a) An efficiency expert claims that a new ergonomic desk chair makes typing at a computer terminal easier and faster. To test it, 15 volunteers are selected. Using both the new chair and their old chair, each volunteer types a randomly selected passage for two minutes and the number of correct words typed is recorded.
b) A developer wants to know if the houses in two different neighbourhoods have the same mean price. She selects 10 houses from each neighbourhood at random and tests the null hypothesis that the means are equal.
c) A manager wants to know if the mean productivity of two workers is the same. For a random selection of 30 hours in the past month, he compares the number of items produced by each worker in that hour. **LO ❸**

18. For each of the following scenarios, say whether the data should be treated as independent or paired samples. Explain briefly. If paired, explain what the pairing involves.

a) An efficiency expert claims that a new ergonomic desk chair makes typing at a computer terminal easier and faster. To test it, 30 volunteers are selected. Half of the volunteers will use the new chair and half will use their old chairs. Each volunteer types a randomly selected passage for two minutes and the number of correct words typed is recorded.
b) A real estate agent wants to know how much extra a fireplace adds to the price of a house. She selects 25 city blocks. In each block she randomly chooses a house with a fireplace and one without and records the assessment value.
c) A manager wants to know if the mean productivity of two workers is the same. For each worker he randomly selects 30 hours in the past month and compares the number of items produced. **LO ❸**

SECTION 14.7

19. A supermarket chain wants to know if its "buy one, get one free" campaign increases customer traffic enough to justify the cost of the program. For each of 10 stores it selects two days at random to run the test. For one of those days (selected by a coin flip), the program will be in effect. The chain wants to test the hypothesis that there is no mean difference in traffic against the alternative that the program increases the mean traffic. Here are the results in number of customer visits to the 10 stores:

Store #	With Program	Without Program
1	140	136
2	233	235
3	110	108
4	42	35
5	332	328
6	135	135
7	151	144
8	33	39
9	178	170
10	147	141

a) Are the data paired? Explain.
b) Compute the mean difference.
c) Compute the standard deviation of the differences.
d) Compute the standard error of the mean difference.
e) Find the value of the t-statistic.
f) How many degrees of freedom does the t-statistic have?
g) Is the alternative one- or two-sided? Explain.
h) What is the P-value associated with this t-statistic? (Assume that the other assumptions and conditions for inference are met.)
i) At $\alpha = 0.05$, what do you conclude? **LO ❸**

20. A city wants to know if a new advertising campaign to make citizens aware of the dangers of driving after drinking has been effective. It counts the number of drivers who have been stopped with more alcohol in their systems than the law allows for each day of the week in the week before and a month after the campaign starts. Here are the results:

Day of Week	Before	After
M	5	2
T	4	0
W	2	2
Th	4	1
F	6	8
S	14	7
Su	6	7

a) Are the data paired? Explain.
b) Compute the mean difference.
c) Compute the standard deviation of the differences.
d) Compute the standard error of the mean difference.
e) Find the value of the t-statistic.
f) How many degrees of freedom does the t-statistic have?
g) Is the alternative one- or two-sided? Explain.
h) What is the P-value associated with this t-statistic? (Assume that the other assumptions and conditions for inference are met.)
i) At $\alpha = 0.05$, what do you conclude? **LO ❸**

21. In order to judge whether the program is successful, the manager of the supermarket chain in Exercise 19 wants to know the plausible range of values for the mean increase in customers using the program. Construct a 90% confidence interval. **LO ❸**

22. A new operating system is installed in every workstation at a large company. The claim of the operating system manufacturer is that the time to shut down and turn on the machine will be much faster. To test it, an employee selects 36 machines and tests the combined shutdown and restart time of each machine before and after the new operating system has been installed. The mean and standard deviation of the differences (before – after) is 23.5 seconds with a standard deviation of 40 seconds.

a) What is the standard error of the mean difference?
b) How many degrees of freedom does the t-statistic have?
c) What is the 90% confidence interval for the mean difference?
d) What do you conclude at $\alpha = 0.05$? **LO ❸**

CHAPTER EXERCISES

23. Hot dogs and calories. Consumers increasingly make food purchases based on nutrition values. In its July 2007 issue, *Consumer Reports* examined the calorie content of two kinds of hot dogs: meat (usually a mixture of pork, turkey, and chicken) and all beef. The researchers purchased samples of several different brands. The meat hot dogs averaged 111.7 calories, compared with 135.4 for the beef hot dogs. A test of the null hypothesis that there's no difference in mean calorie content yields a P-value of 0.124. State the hypotheses and what you conclude. **LO ❶**

24. Hot dogs and sodium. The *Consumer Reports* article described in Exercise 23 also listed the sodium content (in mg) for the various hot dogs tested. A test of the null hypothesis that beef hot dogs and meat hot dogs don't differ in the mean amounts of sodium yields a P-value of 0.110. State the hypotheses and what you conclude. **LO ❶**

25. Learning math. The Core Plus Mathematics Project (CPMP) is an innovative approach to teaching mathematics that engages students in group investigations and mathematical modelling. After field tests in 36 high schools over a three-year period, researchers compared the performances of CPMP students with those taught

using a traditional curriculum. In one test, students had to solve applied algebra problems using calculators. Scores for 320 CPMP students were compared with those of a control group of 273 students in a traditional math program. Computer software was used to create a confidence interval for the difference in mean scores. (**Source:** Based on *Journal for Research in Mathematics Education*, *31*[3], 2000)

Conf. level: 95%

Variable: μ(CPMP) − μ(Ctrl)

Interval: (5.573, 11.427)

a) What is the margin of error for this confidence interval?
b) If we had created a 98% confidence interval, would the margin of error be larger or smaller?
c) Explain what the calculated interval means in this context.
d) Does this result suggest that students who learn mathematics with CPMP will have significantly higher mean scores in applied algebra than those in traditional programs? Explain. **LO** ❷

26. Sales performance. A chain that specializes in healthy and organic food would like to compare the sales performance of two of its primary stores in urban, residential areas with similar demographics. A comparison of the weekly sales randomly sampled over a period of nearly two years for these two stores yields the following information:

Store	*N*	Mean	StDev	Minimum	Median	Maximum
Store #1	9	242170	23937	211225	232901	292381
Store #2	9	235338	29690	187475	232070	287838

a) Create a 95% confidence interval for the difference in the mean weekly store sales.
b) Interpret your interval in context.
c) Does it appear that one store sells more on average than the other store?
d) What is the margin of error for this interval?
e) Would you expect a 99% confidence interval to be wider or narrower? Explain.
f) If you computed a 99% confidence interval, would your conclusion in part c) change? Explain. **LO** ❷

27. CPMP, again. During the study described in Exercise 25, students in both CPMP and traditional classes took another algebra test that did not allow them to use calculators. The table shows the results. Are the mean scores of the two groups significantly different? Assume that the assumptions for inference are satisfied.

Math Program	*n*	Mean	SD
CPMP	312	29.0	18.8
Traditional	265	38.4	16.2

a) Write an appropriate hypothesis.
b) Here is computer output for this hypothesis test. Explain what the P-value means in this context.

2-Sample *t*-Test of $\mu 1 - \mu 2 \neq 0$

t-Statistic = 6.451 w/574.8761 df

$P < 0.0001$

c) State a conclusion about the CPMP program. **LO** ❶

28. IT training costs. An accounting firm is trying to decide between IT training conducted in-house and the use of third-party consultants. To get some preliminary cost data, each type of training was implemented at two of the firm's offices located in different cities. The table shows the average annual training cost per employee at each location. Are the mean costs significantly different? Assume that the assumptions for inference are satisfied.

IT Training	*n*	Mean	SD
In-House	210	$490.00	$32.00
Consultants	180	$500.00	$48.00

a) Write the appropriate hypotheses.
b) Below is computer output for this hypothesis test. Explain what the P-value means in this context.

2-Sample *t*-Test of $\mu 1 - \mu 2 \neq 0$

t-Statistic = 2.38 w/303df

P = .018

c) State a conclusion about IT training costs. **LO** ❶

29. CPMP and word problems. The study of the new CPMP mathematics methodology described in Exercise 25 also tested students' abilities to solve word problems. This table shows how the CPMP and traditional groups performed. What do you conclude? (Assume that the assumptions for inference are met.) **LO** ❶

Math Program	*n*	Mean	SD
CPMP	320	57.4	32.1
Traditional	273	53.9	28.5

30. Statistical training. The accounting firm described in Exercise 28 is interested in providing opportunities for its auditors to gain more expertise in statistical sampling methods. It wishes to compare traditional classroom instruction with online self-paced tutorials. Auditors were assigned at random to one type of instruction, and the auditors were then given an exam. The table shows how the two groups performed. What do you conclude? (Assume the assumptions for inference are met.) **LO** ❶

Program	*n*	Mean	SD
Traditional	296	74.5	11.2
Online	275	72.9	12.3

31. Trucking company. A trucking company would like to compare two different routes for efficiency. Truckers are randomly assigned to two different routes. Twenty truckers following Route A report an average of 40 minutes, with a standard deviation of 3 minutes. Twenty truckers following Route B report an average of 43 minutes, with a standard deviation of 2 minutes. Histograms of travel times for the routes are roughly symmetric and show no outliers.

a) Find a 95% confidence interval for the difference in average time for the two routes.
b) Will the company save time by always driving one of the routes? Explain. **LO ❷**

32. Change in sales. Suppose the specialty food chain from Exercise 26 now wants to compare the change in sales across different regions. An examination of the difference in sales over a 37-week period in a recent year for 8 stores in Alberta compared with 12 stores in British Columbia reveals the following descriptive statistics for relative increase in sales. (If these means are multiplied by 100, they show % increase in sales.)

State	*N*	Mean	StDev
Alberta	8	0.0738	0.0666
B.C.	12	0.0559	0.0503

a) Find the 90% confidence interval for the difference in relative increase in sales over this time period.
b) Is there a significant difference in increase in sales between these two groups of stores? Explain.
c) What would you like to see to check the conditions? **LO ❷**

33. Cereal company. A food company is concerned about recent criticism that the sugar content of its children's cereals is high compared with that of adults' cereals. The data show the sugar content (as a percentage of weight) of several national brands of children's and adults' cereals.

Children's cereals: 40.3, 55, 45.7, 43.3, 50.3, 45.9, 53.5, 43, 44.2, 44, 47.4, 44, 33.6, 55.1, 48.8, 50.4, 37.8, 60.3, 46.6

Adults' cereals: 20, 30.2, 2.2, 7.5, 4.4, 22.2, 16.6, 14.5, 21.4, 3.3, 6.6, 7.8, 10.6, 16.2, 14.5, 4.1, 15.8, 4.1, 2.4, 3.5, 8.5, 10, 1, 4.4, 1.3, 8.1, 4.7, 18.4

a) Write the null and alternative hypotheses.
b) Check the conditions.
c) Find the 95% confidence interval for the difference in means.
d) Is the sugar content of children's cereals significantly higher than that of adults' cereals? **LO ❷**

34. Foreclosure rates. According to reports, home foreclosures were up 47% in March 2008 compared with the previous year (http://realestate.msn.com, April 2008). The data show home foreclosure rates (as % change from the previous year) for a sample of cities in two regions of the United States, the Northeast and the Southwest.

Northeast: 2.99, −2.36, 3.03, 1.01, 5.77, 9.95, −3.52, 7.16, −3.34, 4.75, 5.25, 6.21, 1.67, −2.45, −0.55, 3.45, 4.50, 1.87, −2.15, −0.75

Southwest: 10.15, 23.05, 18.95, 21.16, 17.45, 12.67, 13.75, 29.42, 11.45, 16.77, 12.67, 13.69, 25.81, 21.16, 19.67, 11.88, 13.67, 18.00, 12.88

a) Write the null and alternative hypotheses.
b) Check the conditions.
c) Test the hypothesis and find the P-value.
d) Is there a significant difference in the mean home foreclosure rates between these two regions of the United States? Explain. **LO ❶**

35. Investment. Investment style plays a role in constructing a mutual fund. Each individual stock is classified into two distinct groups: "Growth" and "Value." A growth stock is one with high earning potential and often pays little or no dividends to shareholders. Conversely, value stocks are commonly viewed as steady, or more conservative, with a lower earning potential. You're trying to decide what type of funds to invest in. Because you're saving toward your retirement, if you invest in a value fund, you hope that the fund remains conservative. We would call such a fund "consistent." If the fund didn't remain consistent and became higher-risk, that could impact your retirement savings. The funds in this data set have been identified as either "style-consistent" or "style drifters." Portfolio managers wonder whether consistency provides the optimal chance for successful retirement, so they believe that style-consistent funds outperform style drifters. Out of a sample of 140 funds, 66 were identified as style-consistent, while 74 were identified as style drifters. Their statistics for their average return over five years are as follows:

Type	N	Mean	StDev	Minimum	Q1	Q3	Maximum 5-yr Return
Consistent	66	9.382	2.675	1.750	7.675	11.110	15.920
Drifter	74	8.563	3.719	−0.870	**5.928**	**11.288**	**17.870**

a) Write the null and alternative hypotheses.
b) Find the 95% confidence interval of the difference in mean return between style-consistent and style drifter funds.
c) Is there a significant difference in the five-year return for these two types of funds? Explain. **LO ❶**

36. Technology adoption. The Pew Internet & American Life Project (www.pewinternet.org) conducts surveys to gauge how the Internet and technology impact the daily life of individuals, families, and communities. In a recent survey, Pew asked respondents if they thought computers and technology give people more or less control over their lives. Companies involved in innovative technologies use

the survey results to better understand their target market. One might suspect that younger and older respondents differ in their opinions on whether computers and technology give them more control over their lives. A subset of the data from this survey (*February–March 2007 Tracking Data Set*) shows the mean ages of two groups of respondents: those who reported that they believed computers and technology give them "more" control and those who reported "less" control.

Group	*N*	Mean	StDev	Min	Q1	Med	Q3	Max
More	74	54.42	19.65	18	41.5	53.5	68.5	99.0
Less	29	54.34	18.57	20	41.0	58.0	70.0	84.0

a) Write the null and alternative hypotheses.
b) Find the 95% confidence interval for the difference in mean age between the two groups of respondents.
c) Is there a significant difference in the mean ages between these two groups? Explain. **LO** ❶

37. Product testing. A company is producing and marketing new reading activities for elementary school children that it believes will improve reading comprehension scores. A researcher randomly assigns grade three students to an eight-week program in which some will use these activities and others will experience traditional teaching methods. At the end of the experiment, both groups take a reading comprehension exam. Their scores are shown in the back-to-back stem-and-leaf display. Do these results suggest that the new activities are better? Test an appropriate hypothesis and state your conclusion. **LO** ❶

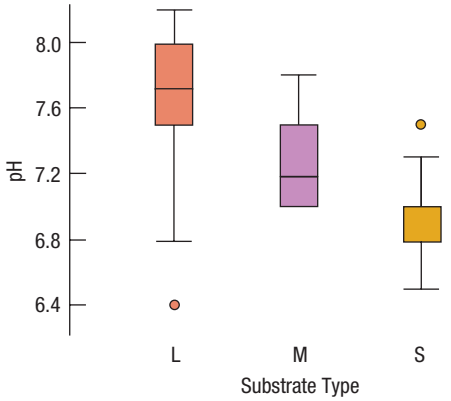
New Activities		Control
	1	07
4	2	068
3	3	377
96333	4	12222238
9876432	5	355
721	6	02
1	7	
	8	5

38. Product placement. The owner of a small organic food store was concerned about her sales of a specialty yogurt manufactured in Greece. As a result of increasing fuel costs, she recently had to increase its price. To help boost sales, she decided to place the product on a different shelf (near eye level for most consumers) and in a location near other popular international products. She kept track of sales (number of containers sold per week) for six months after she made the change. These values are shown below, along with the sales numbers for the six months prior to making the change, in stem-and-leaf displays.

After Change		Before Change	
3	2	2	0
3	9	2	899
4	23	3	224
4	589	3	7789
5	0012	4	000022
5	55558		3
6	00123	4	5567
6	67	5	0
7	0	5	6

Do these results suggest that sales are better after the change in product placement? Test an appropriate hypothesis and state your conclusion. Be sure to check assumptions and conditions. **LO** ❶

39. Acid streams. Researchers collected samples of water from mountain streams to investigate the effects of acid rain. They measured the pH (acidity) of the water and classified the streams with respect to the kind of substrate (type of rock over which they flow). A lower pH means the water is more acidic. Here's a plot of the pH of the streams by substrate (limestone, mixed, or shale):



Here are selected parts of a software analysis comparing the pH of streams with limestone and shale substrates:

Difference Between Means = 0.735

t-Statistic = 16.30 w/133 df

$P \leq 0.0001$

2-Sample *t*-Test of $\mu 1 - \mu 2 = 0$

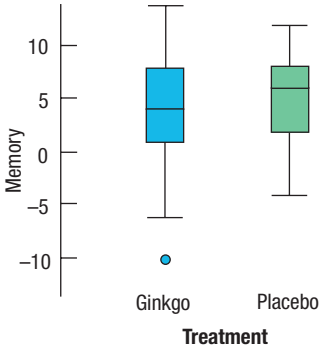
a) State the null and alternative hypotheses for this test.
b) From the information you have, do the assumptions and conditions appear to be met?
c) What conclusion would you draw? **LO** ❶

40. Hurricanes. It has been suggested that global warming may increase the frequency of hurricanes. The data show the number of hurricanes recorded annually before and after 1970. Has the frequency of hurricanes increased since 1970?

Before (1944–1969)	After (1970–2000)
3, 3, 1, 2, 4, 3, 8, 5, 3, 4, 2,	2, 1, 0, 1, 2, 3, 2, 1, 2, 2, 2,
6, 2, 2, 5, 2, 2, 7, 1, 2, 6, 1,	3, 1, 1, 1, 3, 0, 1, 3, 2, 1, 2,
3, 1, 0, 5	1, 1, 0, 5, 6, 1, 3, 5, 3, 4, 2, 3, 6, 7, 2

a) Write the null and alternative hypotheses.
b) Are the conditions for hypothesis testing satisfied?
c) If so, test the hypothesis. **LO ❶**

41. Ginkgo biloba. A pharmaceutical company is producing and marketing a ginkgo biloba supplement to enhance memory. In an experiment to test the product, subjects were assigned randomly to take ginkgo biloba supplements or a placebo. Their memory was tested to see whether it improved. Here are boxplots comparing the two groups and some computer output from a two-sample *t*-test computed for the data.



2-Sample *t*-Test of $\mu G - \mu P > 0$
Difference Between Means = −0.9914
t-Statistic = −1.540 w/196 df
P = 0.9374

a) Explain in this context what the P-value means.
b) State your conclusion about the effectiveness of ginkgo biloba.
c) Proponents of ginkgo biloba continue to insist that it works. What type of error do they claim your conclusion makes? **LO ❶**

42. Baseball. American League Baseball teams play their games with the designated hitter rule, meaning that pitchers don't bat. The league believes that replacing the pitcher, traditionally a weak hitter, with another player in the batting order produces more runs and generates more interest among fans. The data provided in the file includes the average numbers of runs scored per game (*Runs per game*) by American League and National League teams for almost the complete first half of the 2009 season (http://espn.go.com/mlb/stats/team/stat/batting/year/2009).

a) Create an appropriate display of these data. What do you see?
b) With a 95% confidence interval, estimate the mean number of runs scored by American League teams.
c) With a 95% confidence interval, estimate the mean number of runs scored by National League teams.
d) Explain why you shouldn't use two separate confidence intervals to decide whether the two leagues differ in average number of runs scored. **LO ❷**

43. Productivity. A factory hiring people to work on an assembly line gives job applicants a test of manual agility. This test counts how many strangely shaped pegs the applicant can fit into matching holes in a one-minute period. The table summarizes the data by gender of the job applicant. Assume that all conditions necessary for inference are met.

	Male	Female
Number of Subjects	50	50
Pegs Placed	19.39	17.91
Mean	2.52	3.39
SD		

a) Find 95% confidence intervals for the average number of pegs that males and females can each place.
b) Those intervals overlap. What does this suggest about any gender-based difference in manual agility?
c) Find a 95% confidence interval for the difference in the mean number of pegs that could be placed by men and by women.
d) What does this interval suggest about any gender-based difference in manual agility?
e) The two results seem contradictory. Which method is correct: doing two-sample inference, or doing one-sample inference twice?
f) Why don't the results agree? **LO ❷**

44. Online shopping. Online shopping statistics are routinely reported by www.shop.org. Of interest to many online retailers are gender-based differences in shopping preferences and behaviours. The average monthly online expenditures are reported for males and females:

		Male	Female
Group	*n*	45	45
	Mean	$352	$310
	StDev	$95	$80

a) Find 95% confidence intervals for the average monthly online expenditures for males and females.
b) These intervals overlap. What does this suggest about any gender-based difference in monthly online expenditures?
c) Find a 95% confidence interval for the *difference* in average monthly online expenditures between males and females.
d) The two results seem contradictory. Which method is correct: doing two-sample inference, or doing one-sample inference twice? **LO ❷**

45. Baseball, part 2 Do the data in Exercise 42 suggest that the American League's designated hitter rule may lead to more runs?

a) Write the null and alternative hypotheses.
b) Find a 95% confidence interval for the difference in mean runs per game, and interpret your interval.
c) Test the hypotheses stated in part a) and find the P-value.
d) Interpret the P-value and state your conclusion. Does the test suggest that the American League scores more runs on average? LO ❶, ❷

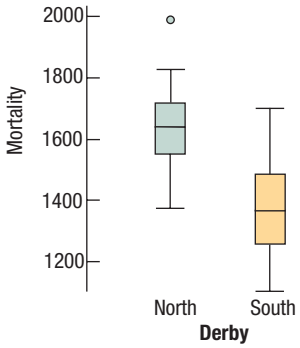
46. Online shopping again. In 2004, it was reported that the average male spends more money shopping online per month than the average female: $204 compared with $186 (www.shop.org). Do the data reported in Exercise 44 indicate that this is still true?

a) Write the null and alternative hypotheses.
b) Test the hypotheses stated in part a) and find the P-value.
c) Interpret the P-value and state your conclusion. Does the test suggest that males continue to spend more online on average than females? LO ❶

47. Drinking water. In an investigation of environmental causes of disease, data were collected on the annual mortality rate (deaths per 100,000) for males in 61 large towns in England and Wales. In addition, the water hardness was recorded as the calcium concentration (parts per million, ppm) in the drinking water. The data set also notes for each town whether it was south or north of Derby. Is there a significant difference in mortality rates in the two regions? Here are the summary statistics.

Summary of:	mortality			
For categories in:	Derby			
Group	**Count**	**Mean**	**Median**	**StdDev**
North	34	1631.59	1631	138.470
South	27	1388.85	1369	151.114

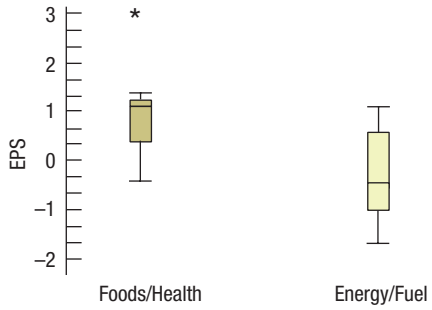
a) Test appropriate hypotheses and state your conclusion.
b) The boxplots of the two distributions show an outlier among the data north of Derby. What effect might that have had on your test? LO ❶



48. Sustainable stocks. The earnings per share ratio (EPS) is one of several important indicators of a company's profitability. There are several categories of "sustainable" stocks, including natural foods/health and green energy/bio fuels. Below are earnings per share for a sample of stocks from both of these categories (**Source:** Yahoo Financial, April 6, 2008). Is there a significant difference in earnings per share values for these two groups of sustainable stocks?

Group	Count	Mean	Median	StDev
Foods/Health	15	0.862	1.140	0.745
Energy/Fuel	16	−0.320	−0.320	0.918

a) Test appropriate hypotheses and state your conclusion.
b) Based upon the boxplots of the two distributions shown below, what might you suspect about your test? Explain. LO ❶



49. Job satisfaction. A company institutes an exercise break for its workers to see if this will improve job satisfaction, as measured by a questionnaire that assesses workers' satisfaction. Scores for 10 randomly selected workers before and after implementation of the exercise program are shown. The company wants to assess the effectiveness of the exercise program. Which type of *t*-test is appropriate for this data? You're not asked to actually conduct the *t*-test, but rather to just say which type you would use. LO ❶, ❸

Worker Number	Job Satisfaction Index	
	Before	After
1	34	33
2	28	36
3	29	50
4	45	41
5	26	37
6	27	41
7	24	39
8	15	21
9	15	20
10	27	37

50. ERP effectiveness. When implementing a packaged Enterprise Resource Planning (ERP) system, many companies report that the module they first install is Financial Accounting. Among the measures used to gauge the effectiveness of their ERP system implementation is

acceleration of the financial close process. Below is a sample of eight companies that report their average time (in weeks) to financial close before and after the implementation of their ERP system.

Company	Before	After
1	6.5	4.2
2	7.0	5.9
3	8.0	8.0
4	4.5	4.0
5	5.2	3.8
6	4.9	4.1
7	5.2	6.0
8	6.5	4.2

Which type of *t*-test is appropriate for this data? You're not asked to actually conduct the *t*-test, but rather to just say which type you would use. **LO ❶, ❸**

51. Delivery time. A small appliance company is interested in comparing delivery times of its products during two months. The company is concerned that the summer slowdowns in August cause delivery times to be different during this month. Given the following delivery times (in days) of appliances to customers for a random sample of six orders each month, test if delivery times differ across these two months. **LO ❶**

June	54	49	68	66	62	62
August	50	65	74	64	68	72

52. Branding. In June 2002, the *Journal of Applied Psychology* reported on a study that examined whether the content of TV shows influenced the ability of viewers to recall brand names of items featured in the commercials. The researchers randomly assigned volunteers to watch one of three programs, each containing the same nine commercials. One of the programs had violent content, another sexual content, and the third neutral content. After the shows ended, the subjects were asked to recall the brands of products that were advertised.

		Program Type		
		Violent	Sexual	Neutral
Brands Recalled	No. of Subjects	108	108	108
	Mean	2.08	1.71	3.77
	SD	1.87	1.76	1.77

a) Do these results indicate that viewer memory for ads may differ depending on program content? Test the hypothesis that there is no difference in ad memory between programs with sexual content and those with violent content. State your conclusion.
b) Is there evidence that viewer memory for ads may differ between programs with sexual content and those with neutral content? Test an appropriate hypothesis and state your conclusion. **LO ❶**

53. Ad campaign. You're a consultant for the marketing department of a business preparing to launch an ad campaign for a new product. The company can afford to run ads during one TV show, and has decided not to sponsor a show with sexual content. You read the study described in Exercise 52 and then use a computer to create a confidence interval for the difference in mean number of brand names remembered between the groups watching violent shows and those watching neutral shows.

Two-Sample *t*

95% CI for $\mu_{viol} - \mu_{neut}$: (−1.578, −0.602)

a) At the meeting of the marketing staff, you have to explain what this output means. What will you say?
b) What advice would you give the company about the upcoming ad campaign? **LO ❶**

54. Branding, part 2. In the study described in Exercise 52, the researchers contacted the subjects again 24 hours later and asked them to recall the brands advertised. Results for the number of brands recalled are summarized in the table.

	Program Type		
	Violent	Sexual	Neutral
No. of Subjects	101	106	103
Mean	3.02	2.72	4.65
SD	1.61	1.85	1.62

a) Is there a significant difference in viewers' abilities to remember brands advertised in shows with violent vs. neutral content?
b) Find a 95% confidence interval for the difference in mean number of brand names remembered between the groups watching shows with sexual content and those watching neutral shows. Interpret your interval in this context. **LO ❶, ❷**

55. Canadian education. Canada's "Youth in Transition Survey" follows the path taken by thousands of Canadians who were 15 years old in the year 2000 and who participated in standardized educational testing at that age. The subjects are interviewed every two years to track "where they are at." Of those at university in 2006 (i.e., at age 21), 73% went straight from secondary school to university and 10% were working in 2004 (at age 19) and then went on to university. The standardized reading score at age 15 of the two groups in the survey is different. The former group had an average score of 597 with a standard error of 8, and the latter group had an average score of 561 with a standard error of 17. (**Source:** Based on Organisation for Economic Co-operation and Development [OECD]. [2010]. Pathways to success: How knowledge and skills at age 15 shape future lives in Canada. Table 3.2.)

a) Does the average reading score at age 15 affect the path Canadians take to university?
b) Is the average reading score at age 15 of Canadians going straight to university higher than for those who work at age 19?

Assume that the number of students in each group is large enough that we can use the Normal approximation to the *t* distribution because the number of degrees of freedom is very large. **LO ❶**

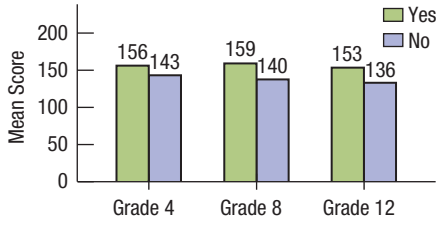
56. Science scores. Newspaper headlines recently announced a decline in science scores among U.S. high school seniors. In 2000, 15,109 seniors tested by the National Assessment in Education Program (NAEP) scored a mean of 147 points. Four years earlier, 7537 seniors had averaged 150 points. The standard error of the difference in the mean scores for the two groups was 1.22.

a) Calculate a 95% confidence interval for the difference in the mean science scores between the two groups.
b) Use your confidence interval to test the hypothesis that the mean score has decreased. What is the significance level of your result?
c) The sample size in 2000 was almost double that in 1996. Does this make the results more convincing or less? Explain. **LO ❶, ❷**

57. Credit card debt. In 2008, the average credit card debt for U.S. college students was reported to be $2200 based on 12,500 responses. A year earlier it was reported to be $2190 based on a survey of 8200 students. The standard error of the difference in mean credit card balances was $1.75.

a) Calculate a 95% confidence interval for the difference in the mean credit card balances between the two years.
b) Use your confidence interval to test the hypothesis that the mean balance has increased. What is the significance level of your result?
c) The sample size in 2008 was one and a half times that in 2007. Does this make the results more or less convincing? Explain. **LO ❶, ❷**

58. The Internet. The NAEP report described in Exercise 56 compared science scores for students who had home Internet access with the scores of those who did not, as shown in the graph. Researchers report that the differences are statistically significant.



a) Explain what "statistically significant" means in this context.
b) If their conclusion is incorrect, which type of error did the researchers commit?
c) Does this prove that using the Internet at home can improve a student's performance in science? **LO ❶**

59. Credit card debt, public or private. The average credit card debt carried by students was compared at public vs. private universities. It was reported that a significant difference existed between the two types of institutions and that students at private universities carried higher credit card debt.

a) Explain what "statistically significant" means in this context.
b) If this conclusion is incorrect, which type of error was committed?
c) Does this prove that students who choose to attend public institutions will carry lower credit card debt? **LO ❶**

60. Pizza sales. A national food product company believes that it sells more frozen pizza during the winter months than during the summer months. Average weekly sales for a sample of stores over a three-year period provided the following data for sales volume (in kilograms) during the two seasons.

Season	*n*	Mean	StDev	Minimum	Maximum
Winter	38	31,234	13,500	15,312	73,841
Summer	40	22,475	8442	12,743	54,706

a) How much difference is there between the mean amount of this brand of frozen pizza sold (in kilograms) between the two seasons? (Assume that this time frame represents typical sales in the area.)
b) Construct and interpret a 95% confidence interval for the difference between weekly sales during the winter and summer months.
c) Suggest factors that might have influenced sales of the frozen pizza during the winter months. **LO ❷**

61. More pizza sales. Here's some additional information about the pizza sales data presented in Exercise 60. It is generally thought that sales spike during the weeks leading up to football championship games, as well as during the weeks leading up to the Super Bowl at the end of January each year. If we omit those six weeks of sales from this three-year period of weekly sales, the summary statistics look like this. Do sales appear to be higher during the winter months after omitting those weeks most influenced by football championship games?

Season	*n*	Mean	StDev	Minimum	Maximum
Winter	32	28995	9913	15312	48354
Summer	40	22475	8442	12743	54706

a) Write the null and alternative hypotheses.
b) Test the null hypothesis and state your conclusion.
c) Suggest additional factors that may influence pizza sales. **LO ❶**

62. Olympic heats. In Olympic running events, preliminary heats are determined by random draw, so we should expect the ability level of runners in the various heats to be about the same, on average. Here are the times (in seconds) for the 400-metre women's run in the 2004 Olympics in Athens for preliminary heats two and five. Is there any evidence that the mean time to finish is different for randomized heats? Explain. Be sure to include a discussion of assumptions and conditions for your analysis. **LO 1**

Country	Name	Heat	Time
USA	HENNAGAN Monique	2	51.02
BUL	DIMITROVA Mariyana	2	51.29
CHA	NADJINA Kaltouma	2	51.50
JAM	DAVY Nadia	2	52.04
BRA	ALMIRAO Maria Laura	2	52.10
FIN	MYKKANEN Kirsi	2	52.53
CHN	BO Fanfang	2	56.01
BAH	WILLIAMS-DARLING Tonique	5	51.20
BLR	USOVICH Svetlana	5	51.37
UKR	YEFRMOVA Antonina	5	51.53
CMR	NGUIMGO Mireille	5	51.90
JAM	BECKFORD Allison	5	52.85
TOG	THIEBAUD-KANGNI Sandrine	5	52.87
SRI	DHARSHA K V Damayanthi	5	54.58

63. Swimming heats. In Exercise 62 we looked at the times in two different heats for the 400-metre women's run from the 2004 Olympics. Unlike track events, swimming heats are not determined at random. Instead, swimmers are seeded so that better swimmers are placed in later heats. Here are the times (in seconds) for the women's 400-metre freestyle from heats two and five. Do these results suggest that the mean times of seeded heats are not equal? Explain. Include a discussion of assumptions and conditions for your analysis. **LO 1**

Country	Name	Heat	Time
ARG	BIAGIOLI Cecilia Elizabeth	2	256.42
SLO	CARMAN Anja	2	257.79
CHI	KOBRICH Kristel	2	258.68
MKD	STOJANOVSKA Vesna	2	259.39
JAM	ATKINSON Janelle	2	260.00
NZL	LINTON Rebecca	2	261.58
KOR	HA Eun-Ju	2	261.65
UKR	BERESNYEVA Olga	2	266.30
FRA	MANAUDOU Laure	5	246.76
JPN	YAMADA Sachiko	5	249.10
ROM	PADURARU Simona	5	250.39
GER	STOCKBAUER Hannah	5	250.46
AUS	GRAHAM Elka	5	251.67
CHN	PANG Jiaying	5	251.81
CAN	REIMER Brittany	5	252.33
BRA	FERREIRA Monique	5	253.75

64. Tee tests. Does it matter what kind of tee a golfer places the ball on? The company that manufactures "Stinger" tees claims that the thinner shaft and smaller head lessen resistance and drag, reducing spin and allowing the ball to travel farther. In August 2003, Golf Laboratories, Inc. compared the distance travelled by golf balls hit off regular wooden tees with those hit off Stinger tees. All the balls were struck by the same golf club using a robotic device set to swing the club head at approximately 95 miles per hour. Summary statistics from the test are shown in the table. Assume that six balls were hit off each tee and that the data were suitable for inference.

		Total Distance (yards)	Ball Velocity (mph)	Club Velocity (mph)
Regular Tee	Avg.	227.17	127.00	96.17
	SD	2.14	0.89	0.41
Stinger Tee	Avg.	241.00	128.83	96.17
	SD	2.76	0.41	0.52

Is there evidence that balls hit off the Stinger tees would have a higher initial velocity? **LO 1**

65. Tee tests, again. Given the test results on golf tees described in Exercise 64, is there evidence that balls hit off Stinger tees would travel farther? Again assume that six balls were hit off each tee and that the data were suitable for inference. **LO 1**

66. Marketing slogan. A company is considering marketing its classical music as "music to study by." Is this a valid slogan? In a study conducted by some Statistics students, 62 people were randomly assigned to listen to rap music, music by Mozart, or no music while attempting to memorize objects pictured on a page. They were then asked to list all the objects they could remember. Here are summary statistics for each group.

	Rap	Mozart	No Music
Count	29	20	13
Mean	10.72	10.00	12.77
SD	3.99	3.19	4.73

a) Does it appear that it's better to study while listening to Mozart than to rap music? Test an appropriate hypothesis and state your conclusion.

b) Create a 90% confidence interval for the mean difference in memory score between students who study to Mozart and those who listen to no music at all. Interpret your interval. **LO 1**

66. Marketing slogan, part 2. Using the results of the experiment described in Exercise 65, is it better to study without music at all than to listen to rap music?

a) Test an appropriate hypothesis and state your conclusion at the 95% significance level.

b) If you concluded that there is a difference, estimate the size of that difference with a 90% confidence interval and explain what your interval means. LO ❶

68. Mutual funds. You've heard that if you leave your money in mutual funds for a longer period of time, you'll see a greater return. So you'd like to compare the three-year and five-year returns of a random sample of mutual funds to see whether your return is indeed expected to be greater if you leave your money in the funds for five years.

a) Using the data file, check the conditions for this test.
b) Write the null and alternative hypotheses for this test.
c) Which type of *t*-test is appropriate for this data? You're not asked to actually conduct the *t*-test, but rather to just say which type you would use. LO ❶, ❸

69. Mutual funds, part 2. An investor now tells you that if you leave your money in as long as 10 years, you'll see an even greater return, so you'd like to compare the 5-year and 10-year returns of a random sample of mutual funds to see whether your return is expected to be greater if you leave your money in the funds for 10 years.

a) Using the data provided, check the conditions for this test.
b) Write the null and alternative hypotheses for this test.
c) Which type of *t*-test is appropriate for this data? You're not asked to actually conduct the *t*-test, but rather to just say which type you would use. LO ❶, ❸

70. Real estate. Residents of neighbouring towns have an ongoing disagreement over who lays claim to the higher average price of a single-family home. Since you live in one of these towns, you decide to obtain a random sample of homes listed for sale with a major local realtor to investigate if there's actually any difference in the average home price.

a) Using the data provided, check the conditions for this test.
b) Write the null and alternative hypotheses for this test.
c) Test the hypotheses and find the P-value.
d) What is your conclusion? LO ❶

71. Real estate, part 2. Residents of one of the towns discussed in Exercise 70 claim that since their town is much smaller, the sample size should be increased. Instead of random-sampling 30 homes, you decide to sample 42 homes from the database to test the difference in the mean price of single-family homes in these two towns.

a) Using the data provided, check the conditions for this test.
b) Write the null and alternative hypotheses for this test.
c) Test the hypotheses and find the P-value.
d) What is your conclusion? Did the sample size make a difference? LO ❶

72. Home run. For the same reasons identified in Exercise 42, a friend of yours claims that the average number of home runs hit per game is higher in the American League than in the National League. Using the same 2009 data as in Exercises 42 and 45, you decide to test your friend's theory.

a) Using the data provided, check the conditions for this test.
b) Write the null and alternative hypotheses for this test.
c) Test the hypotheses and find the P-value.
d) What is your conclusion? LO ❶

73. Statistics journals. When a professional statistician has information to share with colleagues, he or she will submit an article to one of several Statistics journals for publication. This can be a lengthy process; typically, the article must be circulated for "peer review" and perhaps edited before being accepted for publication. Then the article must wait in line with other articles before actually appearing in print. In the Winter 1998 issue of *Chance* magazine, Eric Bradlow and Howard Wainer reported on this delay for several journals between 1990 and 1994. For 288 articles published in *The American Statistician*, the mean length of time between initial submission and publication was 21 months, with a standard deviation of 8 months. For 209 *Applied Statistics* articles, the mean time to publication was 31 months, with a standard deviation of 12 months. Create and interpret a 90% confidence interval for the difference in mean delay, and comment on the assumptions that underlie your analysis. LO ❷

74. Canadian education, part 2. Canada's "Youth in Transition Survey," described in Exercise 55, is based on standardized educational test scores taken in the year 2000 at age 15. Investigate whether there's a difference between provinces, assuming the sample size is large enough to use the Normal approximation to the *t* distribution because the number of degrees of freedom is very large. (**Source:** Organisation for Economic Co-operation and Development [OECD]. [2010]. Pathways to success: How knowledge and skills at age 15 shape future lives in Canada. Table 3.2.)

a) Is there a difference between the reading scores of Alberta (sample mean 550, SE 3.3) and B.C. students (sample mean 538, SE 2.9)?
b) Is there a difference between the math scores of Manitoba (sample mean 533, SE 3.7) and Saskatchewan students (sample mean 525, SE 3.0)? LO ❶

75. Egg production. Can a food additive increase egg production? Egg producers 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 one month, recording the number of eggs each hen produces.

a) Design an experiment that will require a two-sample *t* procedure to analyze the results.
b) Design an experiment that will require a paired *t* procedure to analyze the results.
c) Which experiment would you consider the stronger design? Why? LO ❶, ❸

76. Productivity and music. Some offices pipe in background music. The vendor claims this improves productivity, but might it cause more distraction? A firm's HR department wants to learn whether productivity is affected by background music. The company hires a research firm to conduct an experiment. 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, background music will be piped in.

a) Design an experiment that will require a two-sample *t* procedure to analyze the results.
b) Design an experiment that will require a paired *t* procedure to analyze the results.
c) Which experiment would you consider the stronger design? Why? **LO ❶, ❸**

77. Advertisements. 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 being advertised? A company wants to design an experiment to see if the presence of sexual images in an advertisement affects people's ability to remember the product.

a) Describe an experimental design that would require a paired *t* procedure to analyze the results.
b) Describe an experimental design that would require an independent sample procedure to analyze the results. **LO ❶, ❸**

78. All you can eat. Some sports arenas and ballparks are offering "all you can eat" sections where, for a higher ticket price, fans can feast on all the hot dogs and popcorn they want. (Alcohol and desserts are extra.) But, of course, the teams want to price those tickets appropriately. They want to design an experiment to determine how much fans are likely to eat in an "all you can eat" section and whether it's more or less than they might ordinarily eat in similar regular seats.

a) Design an experiment that would require a two-sample *t* procedure for analysis.
b) Design an experiment that would require a paired *t* procedure for analysis. **LO ❶, ❸**

79. Labour force. Values for the labour force participation rate (proportion) of women (LFPR) are published by the U.S. Bureau of Labor Statistics. We're 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. Here is software output for two possible tests.

Paired *t*-Test 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*-Test of $\mu 1 - \mu 2$
Ho : $\mu 1 - \mu 2 = 0$ Ha : $\mu 1 - \mu 2 \neq 0$
Test Ho : $\mu(1972) - \mu(1968) \neq 0$ vs
Ha : $\mu(1972) - \mu(1968) \neq 0$
Difference Between Means $= 0.0337$
t-Statistic $= 1.496$ w/35 df
P $= 0.1434$

a) Which of these tests is appropriate for these data? Explain.
b) Using the test you selected, state your conclusion. **LO ❶, ❸**

80. Cloud seeding. It has long been a dream of farmers to summon rain when it's needed for their crops. Crop losses to drought have significant economic impact. One possibility is cloud seeding, in which chemicals are dropped into clouds in an attempt to induce rain. Simpson, Alsen, and Eden (*Technometrics*, 1975) reported the results of trials in which clouds were seeded and the amount of rainfall recorded. The authors reported on 26 seeded (Group 2) and 26 unseeded (Group 1) clouds. Each group was sorted in order of the amount of rainfall, largest amount first. Here are two possible tests to study the question of whether cloud seeding works.

Paired *t*-Test of $\mu(1 - 2)$
Mean of Paired Differences $= -277.4$
t-Statistic $= -3.641$w/25 df p $= 0.0012$
2-Sample *t*-Test of $\mu 1 - \mu 2$
Difference Between Means $= -277.4$
t-Statistic $= -1.998$w/33 df p $= 0.0538$

a) Which of these tests is appropriate for these data? Explain.
b) Using the test you selected, state your conclusion. **LO ❶, ❸**

81. Friday the 13th. 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 behaviour. One question was whether people tend to stay at home more on such a date. The data show the number of cars passing Junctions 9 and 10 on the

M25 motorway for consecutive Fridays (6th and 13th) for five different time periods.

Year	Month	6th	13th
1990	July	134,012	132,908
1991	September	133,732	131,843
1991	December	121,139	118,723
1992	March	124,631	120,249
1992	November	117,584	117,263

Here are summaries of two possible analyses.

Paired *t*-Test of $\mu 1 = \mu 2$ vs. $\mu 1 > \mu 2$
Mean of Paired Differences: 2022.4
t-Statistic = 2.9377w/4 df
P = 0.0212

2-Sample *t*-Test of $\mu 1 = \mu 2$ vs. $\mu 1 > \mu 2$
Difference Between Means:2022.4
t-Statistic = 0.4237 w/7.998
P = 0.3402

a) Which of the tests is appropriate for these data? Explain.
b) Using the test you selected, state your conclusion.
c) Are the assumptions and conditions for inference met?
LO ❶, ❸

82. Friday the 13th, part 2. The researchers in Exercise 81 also examined the number of people admitted to emergency rooms for vehicular accidents on 12 Friday evenings (six each on the 6th and 13th).

Year	Month	6th Group 1	13th Group 2
1989	October	9	13
1990	July	6	12
1991	September	11	14
1991	December	11	10
1992	March	3	4
1992	November	5	12

Based on these data, is there evidence that more people are admitted on average on Friday the 13th? Here are two possible analyses of the data.

Paired *t*-Test of $\mu 1 = \mu 2$ vs. $\mu 1 < \mu 2$
Mean of Paired Differences = 3.333
t-Statistic = 2.7116w/5 df
P = 0.0211

2-Sample *t*-Test of $\mu 1 = \mu 2$ vs. $\mu 1 < \mu 2$
Difference Between Means = 3.333
t-Statistic = 1.6644w/9.940 df
P = 0.0636

a) Which of these tests is appropriate for these data? Explain.
b) Using the test you selected, state your conclusion.
c) Are the assumptions and conditions for inference met?
LO ❶, ❸

83. Online insurance. After seeing countless commercials claiming one can get cheaper car insurance from an online company, a local insurance agent was concerned that he might lose some customers. To investigate, he randomly selected profiles (type of car, coverage, driving record, etc.) for 10 of his clients and checked online price quotes for their policies. The comparisons are shown in the table. His statistical software produced the following summaries (*where price Diff = Local − Online*):

Variable	Count	Mean	StdDev
Local	10	799.200	229.281
Online	10	753.300	256.267
PriceDiff	10	45.9000	175.663

Local	Online	PriceDiff
568	391	177
872	602	270
451	488	−37
1229	903	326
605	677	−72
1021	1270	−249
783	703	80
844	789	55
907	1008	−101
712	702	10

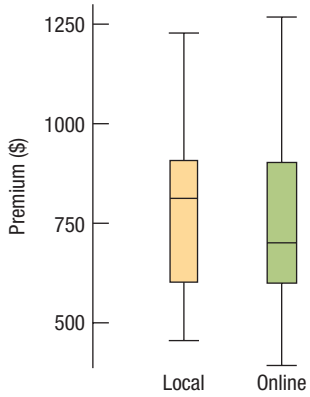
At first, the insurance agent wondered whether there was some kind of mistake in this output. He thought the Pythagorean Theorem of Statistics should work for finding the standard deviation of the price differences—in other words, that $SD(Local - Online) = \sqrt{SD^2(Local) + SD^2(Online)}$. But when he checked, he found that $\sqrt{(229.281)^2 + (256.267)^2} = 343.864$, not 175.663, as given by the software. Tell him where his mistake is. **LO ❸**

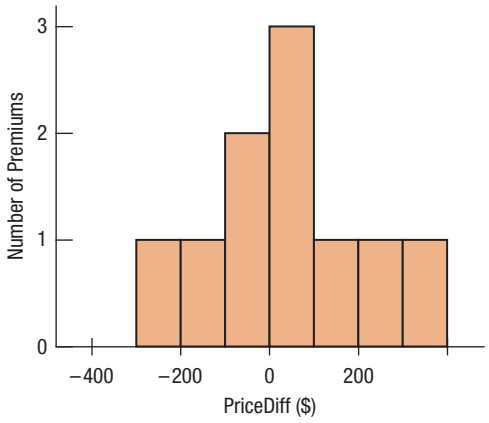
84. Wind energy. Alternative sources of energy are of increasing interest throughout the energy industry. Wind energy has great potential, but appropriate sites must be found for the turbines. To select the site for an electricity-generating wind turbine, wind speeds were recorded at several potential sites every six hours for a year. Two sites not far from each other looked good. Each had a mean wind speed high enough to qualify, but we should choose the site with a higher average daily wind speed. Because the sites are near each other and the wind speeds were recorded at the same times, we should view the speeds as paired. Here are the summaries of the speeds (in kilometres per hour):

Variable	Count	Mean	StdDev
site2	1114	7.452	3.586
site4	1114	7.248	3.421
site2 − site4	1114	0.204	2.551

Is there a mistake in this output? Why doesn't the Pythagorean Theorem of Statistics work here? In other words, shouldn't SD(*site2* − *site4*) = $\sqrt{SD^2(site2) + SD^2(site4)}$? But $\sqrt{(3.586)^2 + (3.421)^2} = 4.956$, not 2.551 as given by the software. Explain why this happened. **LO ❸**

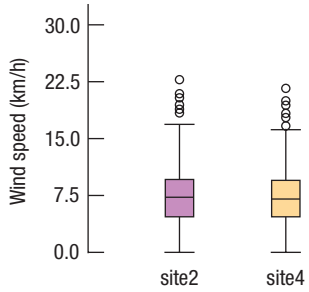
85. Online insurance, part 2. In Exercise 83, we saw summary statistics for 10 drivers' car insurance premiums quoted by a local agent and an online company. Here are displays for each company's quotes and for the difference (*Local* – *Online*):

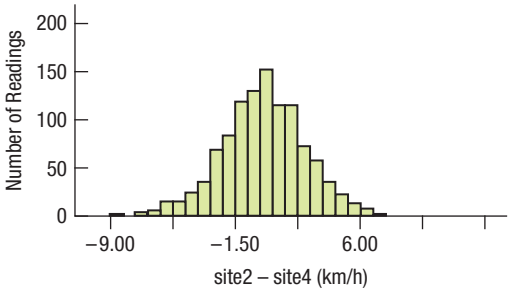




a) Which of the summaries would help you decide whether the online company offers cheaper insurance? Why?
b) The standard deviation of *PriceDiff* is quite a bit smaller than the standard deviation of prices quoted by either the local or the online companies. Discuss why.
c) Using the information you have, discuss the assumptions and conditions for inference with these data. **LO ❸**

86. Wind energy, part 2. In Exercise 84, we saw summary statistics for wind speeds at two sites near each other, both being considered as locations for an electricity-generating wind turbine. The data, recorded every six hours for a year, showed that each of the sites had a mean wind speed high enough to qualify, but how can we tell which site is best? Here are some displays:





a) The boxplots show outliers for each site, yet the histogram shows none. Discuss why.
b) Which of the summaries would you use to select between these sites? Why?
c) Using the information you have, discuss the assumptions and conditions for paired *t* inference for these data. (*Hint:* Think hard about the Independence Assumption in particular.) **LO ❸**

87. Online insurance, part 3. Exercises 83 and 85 give summaries and displays for car insurance premiums quoted by a local agent and an online company. Test an appropriate hypothesis to see if there's evidence that drivers might save money by switching to the online company. **LO ❶**

88. Wind energy, part 3. Exercises 84 and 86 give summaries and displays for two potential sites for a wind turbine. Test an appropriate hypothesis to see if there's evidence that either of these sites has a higher average wind speed. **LO ❶**

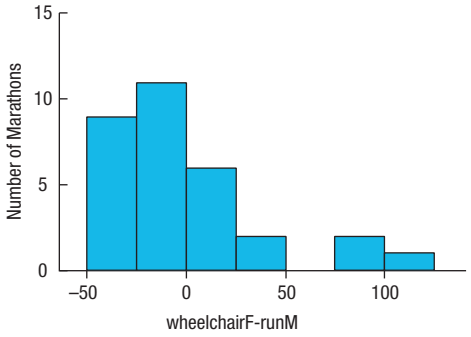
89. Wheelchair marathon. The Boston Marathon has had a wheelchair division since 1977. Who do you think is typically faster, the men's marathon winner on foot or the women's wheelchair marathon winner? Because the conditions differ year to year and speeds have improved over the years, it seems best to treat these as paired measurements. Here are summary statistics for the pairwise differences in finishing time (in minutes).

Summary of wheelchair F – run M
N = 31
Mean= −2.12097
SD = 33.4434

a) Comment on the assumptions and conditions.

b) Assuming that these times are representative of such races and the differences appeared acceptable for inference, construct and interpret a 95% confidence interval for the mean difference in finishing times.
c) Would a hypothesis test at $\alpha = 0.05$ reject the null hypothesis of no difference? What conclusion would you draw? **LO ❸**

90. Boston startup years. When we considered the Boston Marathon in Exercise 89, we were unable to check the Nearly Normal Condition. Here's a histogram of the differences.



Those three large differences are the first three years of wheelchair competition, 1977, 1978, and 1979. Often the startup years of new events are different; later on more athletes train and compete. If we omit those three years, the summary statistics change as follows.

Summary of wheelchair F − run M

$n = 28$

Mean = 12.1780

SD = 19.5116

a) Comment on the assumptions and conditions.
b) Assuming that these times are representative of such races, construct and interpret a 95% confidence interval for the mean difference in finishing time.
c) Would a hypothesis test at $\alpha = 0.05$ reject the null hypothesis of no difference? What conclusion would you draw? **LO ❸**

91. Employee athletes. An ergonomics consultant is engaged by a large consumer products company to see what it can do to increase productivity. The consultant recommends an "employee athlete" program, encouraging every employee to devote five minutes an hour to physical activity. The company worries that the gains in productivity will be offset by the loss in time on the job. Management would like to know if the program increases or decreases productivity. To measure it, the company monitors a random sample of 145 employees who word-process, measuring their hourly keystrokes both before and after the program is instituted. Here are the data:

Keystrokes per Hour			
	Before	After	Difference (AfterBefore)
Mean	1534.2	1556.9	22.7
SD	168.5	149.5	113.6
N	145	145	145

a) What are the null and alternative hypotheses?
b) What can you conclude? Explain.
c) Give a 95% confidence interval for the mean change in productivity (as measured by keystrokes per hour). **LO ❸**

92. Employee athletes, part 2. A small company, on hearing about the employee athlete program (see Exercise 91) at the large company down the street, decides to try it as well. To measure the difference in productivity, the company measures the average number of keystrokes per hour of 23 employees before and after the five-minutes-per-hour program is instituted. The data follow:

Keystrokes per Hour			
	Before	After	Difference (AfterBefore)
Mean	1497.3	1544.8	47.5
SD	155.4	136.7	122.8
N	23	23	23

a) Is there evidence to suggest that the program increases productivity? State your hypotheses clearly.
b) Give a 95% confidence interval for the mean change in productivity (as measured by keystrokes per hour).
c) Explain the difference between the results of part a) and part b). **LO ❸**

93. Productivity. A national fitness firm claims that a company may increase employee productivity by implementing one of the firm's fitness programs at the job site. As evidence of this, the fitness firm reports that one company was able to increase job productivity of a random sample of 30 employees from 57 to 70 (on a scale of 100). The standard deviation of the increases was 7.9. The national fitness firm wants to estimate the mean increase a company could expect after implementing the fitness program.

a) Check the assumptions and conditions for inference.
b) Find a 95% confidence interval.
c) Explain what your interval means in this context. **LO ❸**

94. Productivity, part 2. After implementing the fitness program described in Exercise 93, another company found that a random sample of 48 employees increased their productivity score from 49 to 56, with a standard deviation of 6.2. This company believes that the fitness firm may have exaggerated the potential results of its program. Is there evidence that

the mean improvement seen by this company is less than the one claimed by the fitness company? Be sure to check the assumptions and conditions for inference. LO ❸

95. BST. Many dairy cows now receive injections of BST, a hormone intended to spur greater milk production. After the first injection, a test herd of 60 Ayrshire cows increased their mean daily production from 47 pounds to 61 pounds of milk. The standard deviation of the increases was 5.2 pounds. We want to estimate the mean increase a farmer could expect in his own cows.

a) Check the assumptions and conditions for inference.
b) Write a 95% confidence interval.
c) Explain what your interval means in this context.
d) Given the cost of BST, a farmer believes he can't afford to use it unless he's sure of attaining at least a 25% increase in milk production. Based on your confidence interval, what advice would you give him? LO ❸

96. BST, part 2. In the experiment about hormone injections in cows described in Exercise 95, a group of 52 Jersey cows increased average milk production from 43 pounds to 52 pounds per day, with a standard deviation of 4.8 pounds. Is this evidence that the hormone may be more effective in one breed than in the other? Test an appropriate hypothesis and state your conclusion. Be sure to discuss any assumptions you make. LO ❸

97. European temperatures. The following table gives the average high temperatures in January and July for several European cities. Find a 90% confidence interval for the mean temperature difference between summer and winter in Europe. Be sure to check conditions for inference, and clearly explain what your interval means within the context of the situation. LO ❸

Mean High Temperatures (°C)

City	January	July
Vienna	1.1	23.9
Copenhagen	2.2	22.2
Paris	5.6	24.4
Berlin	1.7	23.3
Athens	12.2	32.2
Rome	12.2	31.1
Amsterdam	4.4	20.6
Madrid	8.3	30.6
London	6.7	22.8
Edinburgh	6.1	18.3
Moscow	−6.1	24.4
Belgrade	2.8	28.9

98. Marathons 2007. Shown are the winning times (in minutes) for men and women in the New York City Marathon between 1978 and 2007 (www.nycmarathon.org). Assuming that performances in the Big Apple resemble performances elsewhere, we can think of these data as a sample of performance in marathon competitions. Create a 90% confidence interval for the mean difference in winning times for male and female marathon competitors. LO ❷

99. Exercise equipment. A leading manufacturer of exercise equipment wanted to collect data on the effectiveness of its equipment. An August 2001 article in the journal *Medicine and Science in Sports and Exercise* compared how long it would take men and women to burn 200 calories during light or heavy workouts on various kinds of exercise equipment. The results summarized in the following table are the average times for a group of physically active young men and women whose performances were measured on a representative sample of exercise equipment.

Average Minutes to Burn 200 Calories

Machine Type	Hard Exertion		Light Exertion	
	Men	Women	Men	Women
Treadmill	12	17	14	22
X-C Skier	12	16	16	23
Stair Climber	13	18	20	37
Rowing Machine	14	16	21	25
Exercise Rider	22	24	27	36
Exercise Bike	16	20	29	44

a) On average, how many minutes longer than a man must a woman exercise at a light exertion rate in order to burn 200 calories? Find a 95% confidence interval.
b) Estimate the average number of minutes longer a woman must work out at light exertion than at heavy exertion to get the same benefit. Find a 95% confidence interval.
c) These data are actually averages rather than individual times. How might this affect the margins of error in these confidence intervals? LO ❸

100. Market value. Real estate agents want to set the price of a house that's about to go on the real estate market correctly. They must choose a price that strikes a balance between one that's so high that the house takes too long to sell and one that's so low that not enough value will go to the homeowner. One appraisal method is the "Comparative Market Analysis" approach, by which the market value of a house is based on recent sales of similar homes in the neighbourhood. Because no two houses are exactly the same, appraisers have to adjust comparable homes for such features as extra square footage, bedrooms, fireplaces, upgrading, parking facilities, swimming pool, lot size, location, and so on. Here are the appraised market values and the selling prices of 45 homes from the same region.

a) Test the hypothesis that on average, the market value and the sale price of homes from this region are the same.

b) Find a 95% confidence interval for the mean difference.
c) Explain your findings in a sentence or two. LO ❸

101. Job satisfaction. A company institutes an exercise break for its workers to see if this will improve job satisfaction, as measured by a questionnaire that assesses workers' satisfaction. Scores for 10 randomly selected workers before and after the implementation of the exercise program are shown in the following table.

Worker Number	Job Satisfaction Index	
	Before	After
1	34	33
2	28	36
3	29	50
4	45	41
5	26	37
6	27	41
7	24	39
8	15	21
9	15	20
10	27	37

a) Identify the procedure you would use to assess the effectiveness of the exercise program and check to see if the conditions allow for the use of that procedure.
b) Test an appropriate hypothesis and state your conclusion. LO ❸

102. Summer school. Having done poorly on their final math exams in June, six students repeat the course in summer school and take another exam in August. If we consider these students to be representative of all students who might attend this summer school in other years, do these results provide evidence that the program is worthwhile?

June	54	49	68	66	62	62
August	50	65	74	64	68	72

a) Identify the procedure you would use to assess whether this program is worthwhile, and check to see if the conditions allow for the use of that procedure.
b) Test an appropriate hypothesis and state your conclusion. LO ❸

103. Efficiency. Many drivers of cars that can run on regular gas actually buy premium gas in the belief that they'll get better gas mileage. To test that belief, a consumer research group evaluated the use of 10 cars in a company fleet in which all the cars run on regular gas. Each car was filled first with either regular or premium gasoline, decided by a coin toss, and the mileage for that tankful was recorded. Then the mileage was recorded again for the same cars with a tankful of the other kind of gasoline. The consumer research group did not let the drivers know about this experiment. Here are the results (miles per gallon):

Car No.	1	2	3	4	5	6	7	8	9	10
Regular	16	20	21	22	23	22	27	25	27	28
Premium	19	22	24	24	25	25	26	26	28	32

a) Is there evidence that cars get better gas mileage on average with premium gasoline?
b) How big might that difference be? Check a 90% confidence interval.
c) Even if the difference is significant, why might the car fleet company choose to stick with regular gasoline?
d) Suppose you had mistakenly treated these data as two independent samples instead of matched pairs. What would the significance test have found? Carefully explain why the results are so different. LO ❸

104. Advertising. A company developing an ad campaign for its cola is investigating the impact of caffeine on studying in hopes of finding evidence of its claim that caffeine helps memory. The firm asked 30 subjects, randomly divided into two groups, to take a memory test. The subjects then each drank two cups of regular (caffeinated) cola or caffeine-free cola. Thirty minutes later they each took another version of the memory test, and the changes in their scores were noted. Among the 15 subjects who drank caffeine, scores fell an average of –0.933 points with a standard deviation of 2.988 points. Among the no-caffeine group, scores went up an average of 1.429 points with a standard deviation of 2.441 points. Assumptions of Normality were deemed reasonable based on histograms of differences in scores.

a) Did scores change significantly for the group who drank caffeine? Test an appropriate hypothesis and state your conclusion.
b) Did scores change significantly for the no-caffeine group? Test an appropriate hypothesis and state your conclusion.
c) Does this indicate that some mystery substance in non-caffeinated soda may aid memory? What other explanation is plausible? LO ❸

105. Quality control. In an experiment on braking performance, a tire manufacturer measured the stopping distance for one of its tire models. On a test track, a car made repeated stops from 60 miles per hour. Twenty tests were run, 10 each on both dry and wet pavement, with results shown in the following table. (Note that actual *braking distance*, which takes into account the driver's reaction time, is much longer, typically nearly 300 feet at 60 miles per hour!)

a) Find a 95% confidence interval for the mean dry pavement stopping distance. Be sure to check the appropriate assumptions and conditions, and explain what your interval means.
b) Find a 95% confidence interval for the mean increase in stopping distance on wet pavement. Be sure to check the

appropriate assumptions and conditions, and explain what your interval means. LO ❷

Stopping Distance (ft.)	
Dry Pavement	Wet Pavement
145	211
152	191
141	220
143	207
131	198
148	208
126	206
140	177
135	183
133	223

106. Quality control, part 2. For another test of the tires in Exercise 105, the company tried them on 10 different cars, recording the stopping distance for each car on both wet and dry pavement. Results are shown in the following table.

Stopping Distance (ft.)		
Car #	Dry Pavement	Wet Pavement
1	150	201
2	147	220
3	136	192
4	134	146
5	130	182
6	134	173
7	134	202
8	128	180
9	136	192
10	158	206

a) Find a 95% confidence interval for the mean dry pavement stopping distance. Be sure to check the appropriate assumptions and conditions, and explain what your interval means.
b) Find a 95% confidence interval for the mean increase in stopping distance on wet pavement. Be sure to check the appropriate assumptions and conditions, and explain what your interval means. LO ❷, ❸

107. Environment. One major impact on the environment is the emission of CO_2 by power plants. Two states that produce the most CO_2 emissions are Texas and California. Both states claim that their power plants are improving. A random sample of power plants in the state of Texas allows us to compare its CO_2 emissions (in tonnes) between the years 2000 and 2007. Using the data provided in the file, test if there has been a significant change in CO_2 emissions in these power plants. LO ❸

108. Student satisfaction. Student surveys are often used to evaluate student satisfaction at the *end* of a course. In a recent paper in the *Journal of the Academy of Business Education* by C. Comm and D. Mathaisel, the authors suggested using "Gap Analysis," as used in marketing methodology to measure the expectation (or importance) and subsequent perception of a customer with regard to a specific product. If we regard the delivery of a university course as a "product," then we can measure the expectation of a student before the course begins and compare it with the perception of the student after the course has ended. The student survey consisted of 26 statements. A five-point Likert scale was used for each statement, where 1 = strongly agree, 2 = agree, 3 = neutral, 4 = disagree, and 5 = strongly disagree. The data in the file include a subsample of a larger data set and represent the responses of 30 students in a quantitative course at a private institution to a question gauging "interest in the subject." Based on these data, assess any gap in the average student's interest before and after the course. Assuming that gap is calculated as the prescore minus postscore, what does a positive gap, or difference, suggest about the course? LO ❸

109. Advertising claims. Advertisements for an instructional video claim that the techniques will improve the ability of Little League pitchers to throw strikes and that, after undergoing the training, players will be able to throw strikes on more than 60% of their pitches. To test this claim, we have 20 Little Leaguers throw 50 pitches each, and we record the number of strikes. After the players participate in the training program, we repeat the test. The following table shows the number of strikes each player threw before and after the training.

a) Is there evidence that after training players can throw strikes more than 60% of the time?
b) Is there evidence that the training is effective in improving a player's ability to throw strikes? LO ❶, ❸

Number of Strikes (out of 50)			
Before	After	Before	After
28	35	33	33
29	36	33	35
30	32	34	32
32	28	34	30
32	30	34	33
32	31	35	34
32	32	36	37
32	34	36	33
32	35	37	35
33	36	37	32

110. Drug costs. In a full-page ad that ran in many U.S. newspapers in August 2002, a Canadian discount pharmacy listed costs of drugs that could be ordered from a website in Canada. The following table compares prices (in U.S. dollars) for commonly prescribed drugs.

Drug Name	Cost per 100 Pills		
	United States	Canada	Percent Savings
Cardizem	131	83	37
Celebrex	136	72	47
Cipro	374	219	41
Pravachol	370	166	55
Premarin	61	17	72
Prevacid	252	214	15
Prozac	263	112	57
Tamoxifen	349	50	86
Vioxx	243	134	45
Zantac	166	42	75
Zocor	365	200	45
Zoloft	216	105	51

a) Find a 95% confidence interval for the average savings in dollars.
b) Find a 95% confidence interval for the average savings in percent.
c) Which analysis do you think is more appropriate? Why?
d) In small print, the newspaper ad says, "Complete list of all 1500 drugs available on request." How does this comment affect your conclusions above? LO ❸

111. Advertisements, part 2. In Exercise 77 you considered the question of whether sexual images in ads affect people's abilities to remember the item being advertised. To investigate, a group of Statistics students cut ads out of magazines. They were careful to find two ads for each of 10 similar items, one with a sexual image and one without. They arranged the ads in random order and had 39 subjects look at them for one minute. Then they asked the subjects to list as many of the products as they could remember. Their data are shown in the table. Is there evidence that the sexual images mattered? LO ❸

Subject Number	Ads Remembered		Subject Number	Ads Remembered	
	Sexual Image	No Sex		Sexual Image	No Sex
1	2	2	21	2	3
2	6	7	22	4	2
3	3	1	23	3	3
4	6	5	24	5	3
5	1	0	25	4	5
6	3	3	26	2	4
7	3	5	27	2	2
8	7	4	28	2	4
9	3	7	29	7	6
10	5	4	30	6	7
11	1	3	31	4	3
12	3	2	32	4	5
13	6	3	33	3	0
14	7	4	34	4	3
15	3	2	35	2	3
16	7	4	36	3	3
17	4	4	37	5	5
18	1	3	38	3	4
19	5	5	39	4	3
20	2	2			#

112. Freshman 15. Cornell Professor of Nutrition David Levitsky recruited students from two large sections of an introductory health course to test the validity of the "Freshman 15" theory that first-year students gain 15 pounds their first year. Although they were volunteers, they appeared to match the rest of the freshman class in terms of demographic variables such as sex and ethnicity. The students were weighed during the first week of the semester, then again 12 weeks later. Based on Professor Levitsky's data, estimate the mean weight gain in first-semester freshmen and comment on the "Freshman 15." (Weights are in pounds.) LO ❸

Subject Number	Initial Weight	Terminal Weight	Subject Number	Initial Weight	Terminal Weight
1	171	168	35	148	150
2	110	111	36	164	165
3	134	136	37	137	138
4	115	119	38	198	201
5	150	155	39	122	124
6	104	106	40	146	146
7	142	148	41	150	151
8	120	124	42	187	192
9	144	148	43	94	96
10	156	154	44	105	105
11	114	114	45	127	130
12	121	123	46	142	144
13	122	126	47	140	143
14	120	115	48	107	107
15	115	118	49	104	105
16	110	113	50	111	112
17	142	146	51	160	162
18	127	127	52	134	134
19	102	105	53	151	151
20	125	125	54	127	130
21	157	158	55	106	108
22	119	126	56	185	188
23	113	114	57	125	128
24	120	128	58	125	126
25	135	139	59	155	158
26	148	150	60	118	120
27	110	112	61	149	150
28	160	163	62	149	149
29	220	224	63	122	121
30	132	133	64	155	158
31	145	147	65	160	161
32	141	141	66	115	119
33	158	160	67	167	170
34	135	134	68	131	131

113. Store sales. A company that owns a chain of specialty food stores would like to see if its sales have increased over the same time period from the previous year. A random

sample of stores produced the average weekly sales for the current quarter compared with the average weekly sales for the same quarter one year ago, for a sample of 15 stores. Using the data provided in the file, determine if the average weekly sales increased over the past year for stores in this chain. **LO ❸**

114. Store profits. The store managers for the sample of stores in Exercise 113 maintain that their stores are doing better this year, despite relatively flat sales. Their argument is that they've been able to reduce costs through more efficient staffing and inventory management. Using the data provided in the file, determine if the average weekly profits for one quarter increased for these stores over the past year (from year 1 to year 2). Do your results support the claim of the store managers? **LO ❸**

115. Yogurt. Do the data in the table suggest that there is a significant difference in calories between servings of strawberry and vanilla yogurt? Test an appropriate hypothesis and state your conclusion, including a check of assumptions and conditions. **LO ❸**

Brand	Calories per Serving	
	Strawberry	Vanilla
America's Choice	210	200
Breyer's Lowfat	220	220
Columbo	220	180
Dannon Light'n Fit	120	120
Dannon Lowfat	210	230
Dannon la Crème	140	140
Great Value	180	80
La Yogurt	170	160
Mountain High	200	170
Stonyfield Farm	100	120
Yoplait Custard	190	190
Yoplait Light	100	100

116. Auto repair shops. Certain businesses and professions have reputations for being somewhat dishonest when dealing with customers. One area of concern is the honesty of auto repair shops. Many provinces require emissions checks; a vehicle that doesn't pass the check must be repaired. In one province, the Department of Transport (DT) has been receiving numerous complaints about a particular auto repair chain. The province decided to check the shops to determine whether they were unlawfully issuing "no pass" reports in order to charge customers unnecessary repair fees. The province procured eight vehicles. Each was first tested on department emissions equipment, and then the eight vehicles were randomly sent to auto repair shops for testing on emissions. As part of the check for accuracy, the hydrocarbon (HC) emissions in parts per million (ppm) were compared:

Vehicle	1	2	3	4	5	6	7	8
DT HC Level	7	10	3	1	5	8	30	7
Auto Shop HC Level	20	11	5	10	5	7	42	15

a) Is there a difference between the measured HC levels taken from the auto shop and the DT measurements? Find a suitable confidence interval.
b) Do you think the DT has evidence that the auto shop readings differ from the department readings? Perform the appropriate test.
c) If you found the test results to be significant, can the DT automatically assume the auto shop is cheating its customers? What other possible explanations could cause the differences in readings? **LO ❸**

117. Airlines. The airline industry has been severely criticized for a variety of service-related issues, including poor on-time performance, cancelled flights, and lost luggage. Some believe airline service is declining while the price of airline fares is increasing. A sample of 10 third-quarter changes in U.S. airfares is shown below.

Origin	Third Quarter 2006	Third Quarter 2007	Percent Change from 3rd Qtr 2006
Cincinnati, OH	511.11	575.67	12.6
Salt Lake City, UT	319.29	344.48	7.9
Dallas Love, TX	185.12	198.74	7.4
New York JFK, NY	324.75	345.97	6.5
Hartford, CT	341.05	363.17	6.5
Charleston, SC	475.10	367.08	−22.7
Columbus, OH	322.60	277.24	−14.1
Kona, HI	206.50	180.40	−12.6
Memphis, TN	418.70	382.29	−8.7
Greensboro/High Point, NC	411.95	377.41	−8.4

Source: Bureau of Transportation Statistics. Top five third quarter U.S. domestic average itinerary fare increases and decreases, 3rd Qtr 2006–3rd Qtr 2007—Top 100 airports based on 2006 U.S. originating domestic passengers. Fares based on 2006 U.S. domestic itinerary fares, round-trip or one-way for which no return is purchased. Averages do not include frequent flyer fares. Retrieved from www.bts.gov/press_releases/2008

a) Does the percent change in airfare from the third-quarter 2006 column represent paired data? Why or why not?
b) Was there an actual change, on average, in airline fares between the two quarters? Perform the test on both the actual and the percentage differences. Discuss the results of the tests and explain how you chose between the fares and the percentage differences as the data to test. **LO ❸**

118. Grocery prices. WinCo Foods, a large discount grocery retailer in the western United States, promotes itself as the lowest-priced grocery retailer. In newspaper ads printed and distributed during January 2008, WinCo Foods published a price comparison for products between WinCo

and several competing grocery retailers. One of the retailers compared against WinCo was Walmart, also known as a low-price competitor. WinCo selected a variety of products, listed the prices of the product charges at each retailer, and showed the sales receipt to prove that the prices at WinCo were the lowest in the area. A sample of the products and their price comparison at both WinCo and Walmart are shown in the following table:

Item	WinCo Price	Walmart Price
Bananas (lb)	0.42	0.56
Red Onions (lb)	0.58	0.98
Mini Peeled Carrots (1 lb bag)	0.98	1.48
Roma Tomatoes (lb)	0.98	2.67
Deli Tater Wedges (lb)	1.18	1.78
Beef Cube Steak (lb)	3.83	4.118
Beef Top Round London Broil (lb)	3.48	4.12
Pillsbury Devils Food Cake Mix (18.25 oz)	0.88	0.88
Lipton Rice and Sauce Mix (5.6 oz)	0.88	1.06
Sierra Nevada Pale Ale (12 – 12 oz bottles)	12.68	12.84
GM Cheerios Oat Clusters (11.3 oz)	1.98	2.74
Charmin Bathroom Tissue (12 roll)	5.98	7.48
Bumble Bee Pink Salmon (14.75 oz)	1.58	1.98
Pace Thick & Chunky Salsa, Mild (24 oz)	2.28	2.78
Nalley Chili, Regular w/Beans (15 oz)	0.78	0.78
Challenge Butter (lb quarters)	2.18	2.58
Kraft American Singles (12 oz)	2.27	2.27
Yuban Coffee FAC (36 oz)	5.98	7.56
Totino's Pizza Rolls, Pepperoni (19.8 oz)	2.38	2.42
Rosarita Refried Beans, Original (16 oz)	0.68	0.73
Barilla Spaghetti (16 oz)	0.78	1.23
Sun-Maid Mini Raisins (14 – .5 oz)	1.18	1.36
Jif Peanut Butter, Creamy (28 oz)	2.54	2.72
Dole Fruit Bowl, Mixed Fruit (4 – 4 oz)	1.68	1.98
Progresso Chicken Noodle Soup (19 oz)	1.28	1.38
Precious Mozzarella Ball, Part Skim (16 oz)	3.28	4.23
Mrs. Cubbison Seasoned Croutons (6 oz)	0.88	1.12
Kellogg's Raisin Bran (20 oz)	1.98	2.50
Campbell's Soup at Hand, Cream of Tomato (10.75 oz)	1.18	1.26

a) Do the prices listed indicate that, on average, prices at WinCo are lower than prices at Walmart?
b) At the bottom of the price list, the following statement appears: "Though this list is not intended to represent a typical weekly grocery order or a random list of grocery items, WinCo continues to be the area's low price leader." Why do you think WinCo added this statement?
c) What other comments could be made about the statistical validity of the test on price comparisons given in the ad? **LO ❸**

119. Forsee and Amazon. Each year Forsee surveys 8500 consumers about customer satisfaction with retailers, and publishes the average score out of 100. In 2010, the highest score any retailer achieved was 86. In 2011, Amazon achieved a score of 88. Assume that the standard deviations of the scores are 30% of the means. Are customers in 2011 more satisfied with Amazon than with the retailer that achieved the score of 86 in 2010 at the 0.05 significance level? Answer this question

a) assuming that Forsee surveyed different consumers in 2010 and in 2011;
b) as in a) but with a pooled estimate of the standard deviation; and
c) assuming that Forsee surveyed the same consumers in 2010 and in 2011 and that the standard deviation of the difference in the scores was six times the mean difference.
d) Suppose Forsee had surveyed only 100 consumers each year instead of 8500. Would it make a difference to your answers to a), b), and c) at the 0.05 significance level? **LO ❶, ❸**

120. Forsee and Netflix. The two surveys described in the previous question resulted in a drop in the average rating of Netflix from 86 in 2010 to 79 in 2011. Assume that the standard deviations of the scores are 30% of the means. Are consumers less satisfied with Netflix in 2011 than in 2010 at the 0.05 significance level? Answer this question

a) assuming that Forsee surveyed different consumers in 2010 and in 2011;
b) as in a) but with a pooled estimate of the standard deviation; and
c) assuming that Forsee surveyed the same consumers in 2010 and in 2011 and that the standard deviation of the difference in the scores was six times the mean difference.
d) Suppose Forsee had surveyed only 85 consumers in 2010 and 75 in 2011 instead of 8500. Would it make a difference to your answers to a), b) and c) at the 0.01 significance level? **LO ❶, ❸**

121. Canadian house sizes. The Organisation for Economic Co-operation and Development (OECD) surveyed 1000 homes at random in each of its member countries in 2011, and found that the number of rooms per person in Canadian homes was 2.5 on average, whereas in the U.S. it was 2.3. Assume that the standard deviation is 35% of these average values. Do Canadian homes in general have more rooms per person than homes in the U.S.? Answer this question

a) without pooling the estimates of the standard deviation; and
b) with a pooled estimate of the standard deviation.

c) Why is it not possible to use the "paired samples" approach for this question?
d) At which significance level (0.01, 0.05, or 0.1) would your answers to a) and b) be different if the OECD had surveyed only 100 homes in each country instead of 1000?
LO ❶, ❸

122. Canadian hours of work. The Organisation for Economic Co-operation and Development (OECD) surveyed 1000 people at random in each of its member countries in 2011, and found that the number of hours worked by Canadians was 1699 hours per year on average, whereas in the U.S. it was 1768. Assume that the standard deviation is 30% of these average values. In Canada, do people work fewer hours per year than people in the U.S.? Answer this question

a) without pooling the estimates of the standard deviation; and
b) with a pooled estimate of the standard deviation.
c) Why is it not possible to use the "paired samples" approach for this question?
d) At which significance level (0.01, 0.05, or 0.1) would your answers to a) and b) be different if the OECD had surveyed only 100 people in each country instead of 1000?
LO ❶, ❸

123. Female births in India. The National Sample Survey, which monitors demographic characteristics in India, found the 2005 gender ratio of live births to be 924 females per thousand males. In 2011 the number had increased to 977 females per thousand males. We're interested in whether these data show that the gender ratio increased between 2005 and 2011 at the 0.05 significance level, assuming that the size of the sample is 1000 in 2005 and in 2011. Which if any of the methods described in this chapter can be used to address this question

a) without pooling the estimates of the standard deviation?
b) with a pooled estimate of the standard deviation?
c) using "paired samples"?

Note: You're not asked to use the method; you're asked only whether the method could be used. **LO ❶, ❸**

Just Checking Answers

1 Ho: $\mu_{eyes} - \mu_{flowers} = 0$
- ✓ **Independence Assumption:** The amount paid by one person should be independent of the amount paid by others.
- ✓ **Randomization Condition:** This study was observational. Treatments alternated a week at a time and were applied to the same group of office workers.
- ✓ **Nearly Normal Condition:** We don't have the data to check, but it seems unlikely there would be outliers in either group.
- ✓ **Independent Groups Assumption:** The same workers were recorded each week, but week-to-week independence is plausible.
- ✓ **Independent Groups Assumption:** The same workers were recorded each week, but week-to-week independence is plausible.

2 $H_A: \mu_{eyes} - \mu_{flowers} \neq 0$.

3 The average amount of money that office workers left to pay for food at an office coffee station was different when a picture of eyes was placed behind the "honesty box" than when the picture was one of flowers.

4 These are independent groups sampled at random, so use a two-sample *t*-confidence interval to estimate the size of the difference.

5 If the same random sample of students was sampled both in the first year and again in the fourth year of their university experience, then this would be a paired *t*-test.

6 A male and a female are selected from each work group. The question calls for a paired *t*-test.

7 Since the sample of companies is different in each of the industries, this would be a two-sample test. There is no "pairing."

8 Since the same 50 companies are surveyed twice to examine a change in variables over time, this would be a paired *t*-test.