EXAMPLE 9.5.1. A study was conducted to investigate the effect of physical training on the serum cholesterol level. Eleven subjects participated in the study. Prior to training, blood samples were taken to determine the cholesterol level of each subject. Then the subjects were put through a training program that centered on daily running and jogging. At the end of the training period, blood samples were taken again and a second reading on the serum cholesterol level was obtained. Thus, two sets of observations on the serum cholesterol level of the subjects are available. The data sets are not independent; they are based on the same subjects taken at different times and so are naturally paired by subject. These data were collected:

訓練前　　後

Subject	Pretraining level x, mg/dL	Posttraining level y, mg/dL	Difference $d = x - y$
1	182	198	−16
2	232	210	22
3	191	194	−3
4	200	220	−20
5	148	138	10
6	249	220	29
7	276	219	57
8	213	161	52
9	241	210	31
10	480	313	167
11	262	226	36

The purpose is to estimate the difference between the mean cholesterol level before and after training.

For these data,

$$\bar{d} = 33.2 \qquad s_d = 51.1$$

The partition of the $T_{n-1} = T_{10}$ curve needed is shown in Figure 9.5. The desired confidence bounds are

$$\bar{d} \pm t\frac{s_d}{\sqrt{n}} = 33.2 \pm 1.812\frac{51.1}{\sqrt{11}}$$
$$= 33.2 \pm 27.9$$

$H_0: \mu_D = 0$ vs $H_a: \mu_D > 0$ 單尾

$\bar{D} \overset{?}{>} 0 + \underset{1.812}{t_{10,0.5}} \cdot \frac{51.1}{\sqrt{11}}$

p-value $= P_r(T_{10} > t_{obs})$

$= P_r\left(T_{10} > \frac{33.2}{51.1/\sqrt{11}}\right) = P_r(T_{10} > 2.15)$

$\in (0.025, 0.5) \Rightarrow$ 拒絕

$T_{10,0.5} = 1.812 \qquad T_{10,0.025} = 2.228$

雙尾

$H_0: \mu_D = 0$ vs. $H_a: \mu_D \neq 0$

$\bar{D} \overset{?}{>} 0 + t_{10,0.025} \cdot \frac{51.1}{\sqrt{11}}$

or $\bar{D} < 0 - t_{10,0.025} \cdot \frac{51.1}{\sqrt{11}}$

p-value $= 2\,P_r(T_{10} > |t_{obs}|)$

95% C.I　$33.2 \pm 27.9 \rightarrow$ 不含 0

$\Rightarrow$ Reject H_0

Paired T Tests

Means can be compared by using the hypothesis testing approach also. The null hypothesis $\mu_X = \mu_Y$ is equivalent to the hypothesis $\mu_D = 0$. The test statistic for testing this hypothesis based on the sample of difference scores is

$$\boxed{\frac{\bar{D} - 0}{S_d/\sqrt{n}}} \qquad \text{(Paired } T \text{ Test)}$$

which follows a T distribution with $n - 1$ degrees of freedom if H_0 is true. The use of this statistic is illustrated in the following example.

EXAMPLE 9.5.3. A study is conducted of tooth emergence in Australian aborigines. The purpose is to detect differences, if they exist, in the time of emergence of left- and right-side permanent teeth. One tooth studied is the incisor. All subjects are male. The age of the subject at the time of emergence of the left incisor and his age at the time of emergence of the right incisor are determined. Thus each subject produces a pair of observations. Summary statistics for the study are as shown, where the order of subtraction is left-side age minus right-side age:

$$n = 17 \qquad \bar{d} = 1.5 \text{ yr} \qquad s_d = 4.7$$

小孩長牙

左 vs. 右

⇓

双尾

The observed value of the test statistic is

$$\frac{\bar{d} - 0}{s_d/\sqrt{n}} = \frac{1.5}{4.7/\sqrt{17}} = 1.31$$

Notice that $P[T_{16} \geq 1.31] > .10$. Since no directional preference is indicated, the test is two-tailed. The P value for the two-tailed test exceeds .20. There is not enough evidence based on this study to claim that there is a difference in the mean time of emergence of left and right incisors in male Australian aborigines.

p-value 要 double

In using these procedures, the assumption is made that the variable $D = X - Y$ is at least approximately normally distributed.

pair T test 自行練習

1. The effect of physical training on the triglyceride level was also studied by using the 11 subjects of Example 9.5.1. The following pretraining and posttraining readings (in milligrams of triglyceride per 100 milliliters of blood) were obtained:

Subject	Pretraining	Posttraining
1	68	95
2	77	90
3	94	86
4	73	58
5	37	47
6	131	121
7	77	136
8	24	65
9	99	131
10	629	630
11	116	104

D

95-68

90-77

⋮

自行求 S_D, $\bar{D}$

單尾 $H_0: \mu_D = 0$ vs. $H_a: \mu_D > 0$

双尾 $H_0: \mu_D = 0$ vs $H_a: \mu_D \neq 0$

Find a 90% confidence interval on the mean change in triglyceride level. Is there evidence that a difference exists? If so, what is the direction of the change?

考試會告知用單尾 or 雙尾

Comparing two population means

We can examine two-sample data graphically by comparing boxplots, stemplots (for small samples), or histograms (for larger samples). Now we will learn confidence intervals and tests in this setting. When both population distributions are symmetric, and especially when they are at least approximately Normal, a comparison of the mean responses in the two populations is the most common goal of inference. Here are the conditions for inference.

CONDITIONS FOR INFERENCE COMPARING TWO MEANS

- We have **two SRSs,** from two distinct populations. The samples are **independent.** That is, one sample has no influence on the other (matching violates independence, for example). We measure the same variable for both samples.
- Both populations are **Normally distributed.** The means and standard deviations of the populations are unknown. In practice, it is enough that the distributions have similar shapes and that the data have no strong outliers.

Call the variable we measure x_1 in the first population and x_2 in the second, because the variable may have different distributions in the two populations. Here is the notation we will use to describe the two populations:

Population	Variable	Population Mean	Population Standard deviation
1	x_1	μ_1	σ_1
2	x_2	μ_2	σ_2

Population	Sample size	Sample mean	Sample standard deviation
1	n_1	$\bar{x}_1$	s_1
2	n_2	$\bar{x}_2$	s_2

To do inference about the difference $\mu_1 - \mu_2$ between the means of the two populations, we start from the difference $\bar{x}_1 - \bar{x}_2$ between the means of the two samples.

EXAMPLE 18.2 Does polyester decay?

STATE: How quickly do synthetic fabrics such as polyester decay in landfills? A researcher buried polyester strips in the soil for different lengths of time, then dug up the strips and measured the force required to break them. Breaking strength is easy to measure and is a good indicator of decay. Lower strength means the fabric has decayed.

Part of the study buried 10 strips of polyester fabric in well-drained soil in the summer. Five of the strips, chosen at random, were dug up after 2 weeks; the other 5 were dug up after 16 weeks. Here are the breaking strengths in pounds:[1]

Sample 1 (2 weeks)	118	126	126	120	129
Sample 2 (16 weeks)	124	98	110	140	110

We suspect that decay increases over time. Do the data give good evidence that mean breaking strength is less after 16 weeks than after 2 weeks?

FORMULATE: This is a two-sample setting. We want to compare the mean breaking strengths in the entire population of polyester fabric, μ_1 for fabric buried for 2 weeks and μ_2 for fabric buried for 16 weeks. So we will test the hypotheses

$$H_0: \mu_1 = \mu_2 \text{ (that is, } \mu_1 - \mu_2 = 0)$$
$$H_a: \mu_1 > \mu_2 \text{ (that is, } \mu_1 - \mu_2 > 0)$$

SOLVE (FIRST STEPS): Are the conditions for inference met? Because of the randomization, we are willing to regard the two groups of fabric strips as two independent SRSs from large populations of fabric. Although the samples are small, we check for serious non-Normality by examining the data. Figure 18.1 is a back-to-back stemplot of the responses. The 16-week group is much more spread out. As far as we can tell from so few observations, there are no departures from Normality that violate the conditions for comparing two means.

From the data, calculate the summary statistics:

Group	Treatment	n	$\bar{x}$	s
1	2 weeks	5	123.80	4.60
2	16 weeks	5	116.40	16.09

$\bar{x}_1 - \bar{x}_2 = 123.80 - 116.40 = 7.40$ pounds

2 weeks		16 weeks
	9	8
	10	
8	11	00
9660	12	4
	13	
	14	0

FIGURE 18.1 Back-to-back stemplot of the breaking strength data from Example 18.2.

Two-sample t procedures

To assess the significance of the observed difference between the means of our two samples, we follow a familiar path. Whether an observed difference is surprising depends on the spread of the observations as well as on the two means. Widely different means can arise just by chance if the individual observations vary a great deal. How much the difference $\bar{x}_1 - \bar{x}_2$ can vary from one random sampling to another is given by its sampling distribution.

When two random variables are Normally distributed, the new variable "difference" also follows a Normal distribution, centered on the difference of the two variables' means and with variance equal to the sum of the two variables' variances. We already know from Chapter 14 that the sampling distributions of $\bar{x}_1$ and $\bar{x}_2$ have standard deviations $\sigma_1/\sqrt{n_1}$ and $\sigma_2/\sqrt{n_2}$, respectively. Therefore, when we look at the difference $\bar{x}_1 - \bar{x}_2$, the standard deviation of its sampling distribution is

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

This standard deviation gets larger as either population gets more variable, that is, as σ_1 or σ_2 increases. It gets smaller as the sample sizes n_1 and n_2 increase.

Because we don't know σ_1 and σ_2, we estimate them by the sample standard deviations s_1 and s_2. The result is the **standard error,** or estimated standard deviation, of the difference in sample means:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

保守原則

When we standardize the estimate, we get

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

but, because in a typical two-sample test $\mu_1 - \mu_2 = 0$, the result is the **two-sample t statistic:**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$$

The statistic t has the same interpretation as any z or t statistic: It says how far the difference $\bar{x}_1 - \bar{x}_2$ is from 0 ($\mu_1 - \mu_2$) in standard deviation units. (Exceptionally, the null hypothesis may define $\mu_1 - \mu_2$ as a value other than zero and the t statistic would be standardized using the full standardization formula.)

The two-sample t statistic has approximately a t distribution. It does not have exactly a t distribution even if the populations are both exactly Normal. In practice, however, the approximation is very accurate. There is a catch: The degrees of freedom of the t distribution we want to use are calculated from the data by a somewhat messy formula; moreover, the degrees of freedom need not be a whole number. There are two practical options for using the two-sample t procedures:

Option 1. With software, use the statistic t with accurate critical values from the approximating t distribution.

Option 2. Without software, use the statistic t with critical values from the t distribution with degrees of freedom equal to the smaller of $n_1 - 1$ and $n_2 - 1$. These procedures are always conservative for any two Normal populations.

EXAMPLE 18.3 Does polyester decay?

We can now complete Example 18.2.

SOLVE (INFERENCE): The test statistic for the null hypothesis H_0: $\mu_1 = \mu_2$ is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{123.8 - 116.4}{\sqrt{\frac{4.60^2}{5} + \frac{16.09^2}{5}}} = \frac{7.4}{7.484} = 0.9889$$

電腦算的

Software (Option 1) gives one-sided P-value $P = 0.1857$.

Without software, use the conservative Option 2. Because $n_1 - 1 = 4$ and $n_2 - 1 = 4$, there are 4 degrees of freedom. Because H_a is one-sided on the high side, the P-value is the area to the right of $t = 0.9889$ under the $t(4)$ curve. Figure 18.2 illustrates this P-value. Table C shows that it lies between 0.15 and 0.20.

CONCLUDE: The experiment did not find convincing evidence that polyester decays more in 16 weeks than in 2 weeks ($P > 0.15$).

0.9889 在此

df = 4		
t^*	0.941	1.190
P	0.20	0.15

p-value 也在此

EXAMPLE 18.4 Brain size and autism

STATE: Is autism marked by different brain growth patterns in early life, even b the diagnosis is made? Studies have linked brain size in infants and toddlers to a nu of future ailments, including autism. One study looked at the brain sizes of 30 au boys and 12 nonautistic boys (control) who all had received an MRI scan as tod Here are their whole-brain volumes in milliliters:[3]

Autistic									
1311	1250	1292	1419	1401	1297	1202	1336	1308	1353
1515	1461	1365	1364	1362	1303	1278	1247	1333	1340
1319	1286	1223	1241	1229	1209	1171	1154	1128	1230
Control									
1040	1180	1207	1179	1115	1133	1298	1263	1194	1198
1230	1114								

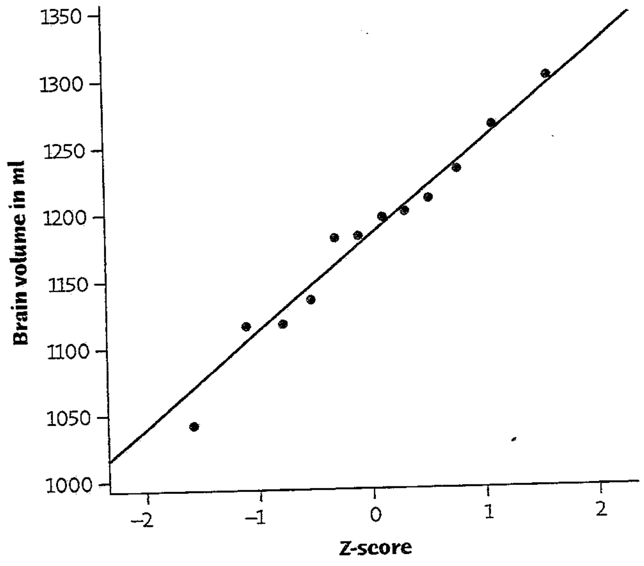
FORMULATE: We had no specific direction for the difference in brain volumes b looking at the data, so the alternative is two-sided. We will test the hypotheses

$$H_0: \mu_1 = \mu_2 \text{ (that is, } \mu_1 - \mu_2 = 0)$$
$$H_a: \mu_1 \neq \mu_2 \text{ (that is, } \mu_1 - \mu_2 \neq 0)$$

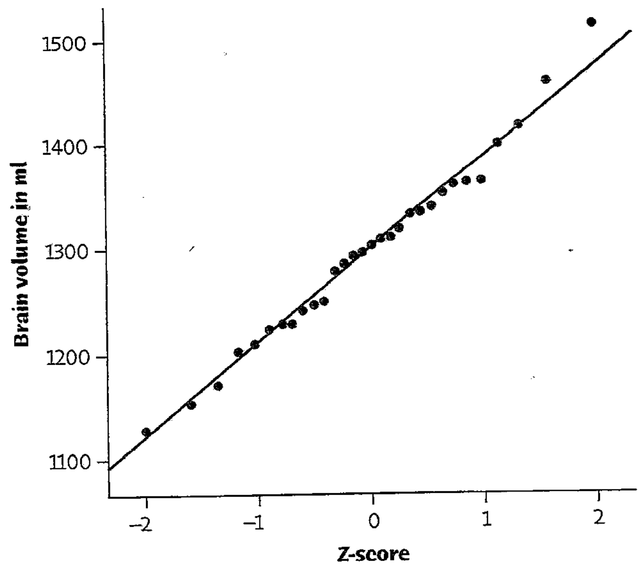
SOLVE: The two samples can be regarded as SRSs from two populations, alth there are clear limitations due to the lack of availability of young children for biome studies. Figure 18.3 shows the two data sets in a back-to-back stemplot and in Nc quantile plots.[4] The stemplot in Figure 18.3(a) shows no deviation from Normali the 30 autistic boys and a very mild possible outlier for the 12 control boys. The qua plots in Figures 18.3(b) and 18.3(c) suggest that a Normal distribution is a reaso model for both groups. Thus, the t procedures should be valid. Here are the calcula leading to the two-sample t-test:

Group	Condition	n	$\bar{x}$	s
1	Autistic	30	1297.6	88.4
2	Control	12	1179.3	70.7

目的: 檢驗兩者假設



(b)



(c)

4	10	
	10	
3 1 1	11	2
9 9 8 7	11	5 7
3 0	12	0 0 2 2 3 4 4
9 6	12	5 7 8 9 9
	13	0 0 1 1 3 3 4
	13	5 6 6 6
	14	0 1
	14	6
	15	1

(a)

The two-sample t statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{1297.6 - 1179.3}{\sqrt{\frac{88.4^2}{30} + \frac{70.7^2}{12}}} = \frac{118.3}{26.02} = 4.55$$

Software (Option 1) says that the two-sided P-value is $P = 0.0001$.

Without software, use Option 2 to find a conservative P-value. There are 11 degrees of freedom, the smaller of

$$n_1 - 1 = 30 - 1 = 29 \quad \text{and} \quad n_2 - 1 = 12 - 1 = 11$$

Figure 18.4 illustrates the P-value. Find it by comparing 4.55 with the two-sided critical values for the $t(11)$ distribution.

Notice that our t statistic is larger than the largest t critical in Table C for degrees of freedom 11. When this happens, simply conclude that your P-value is smaller than the smallest P-value provided by the table. Here we conclude that the P-value for our test is less than 0.001 (a 2-sided P).

CONCLUDE: The data give very strong evidence ($P < 0.001$) that autistic boys have larger brains on average than nonautistic boys during the toddler years.

EXAMPLE 18.5 The autistic brain: how much larger?

FORMULATE: Give a 90% confidence interval for $\mu_1 - \mu_2$, the difference in mean brain size during the toddler years between all autistic and nonautistic boys.

SOLVE AND CONCLUDE: As in Example 18.4, the conservative Option 2 uses 11 degrees of freedom. Table C shows that the $t(11)$ critical value is $t^* = 1.796$. We are 90% confident that $\mu_1 - \mu_2$ lies in the interval

$$\begin{aligned}(\bar{x}_1 - \bar{x}_2) \pm t^*\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} &= (1297.6 - 1179.3) \pm 1.796\sqrt{\frac{88.4^2}{30} + \frac{70.7^2}{12}} \\ &= 118.3 \pm 46.7 \\ &= 71.6 \text{ to } 165.0\end{aligned}$$

Notice that, because 0 lies outside the 90% confidence interval, we can reject H_0: $\mu_1 = \mu_2$ in favor of the two-sided alternative at the $\alpha = 0.10$ level of significance. Remember that a confidence interval can also help you test hypotheses.

結論（科學證據）

The authors concluded that abnormal brain development in autism may occur prior to 2 or 3 years of age and that future research should explore ways to limit or prevent the full expression of these abnormalities. However, because this is an observational study, it is not possible to conclude that autism is indeed the cause of the observed difference in brain volumes. Confounding variables might explain

.e difference, especially since the two groups were recruited separately. The re-archer honestly disclosed this and other limitations of the study. Disclosing the nitations of a study design or of a statistical procedure is an important part of the hical conduct of research.

避免[illegible] 了解研究限制

Robustness again

假設不符合時

The two-sample t procedures are more robust than the one-sample t methods, particularly when the distributions are not symmetric. When the sizes of the two samples are equal and the two populations being compared have distributions with similar shapes, probability values from the t table are quite accurate for a broad range of distributions when the sample sizes are as small as $n_1 = n_2 = 5$.[8] When the two population distributions have different shapes, larger samples are needed.

As a guide to practice, adapt the guidelines given on page 446 for the use of one-sample t procedures to two-sample procedures by replacing "sample size" with "sum of the sample sizes," $n_1 + n_2$. These guidelines err on the side of safety, especially when the two samples are of equal size. *In planning a two-sample study, choose equal sample sizes whenever possible. The two-sample t procedures are most robust against non-Normality in this case.*

ex 1

Weeds among the corn. Lamb's quarter is a common weed that interferes with the growth of corn. An agriculture researcher planted corn with the same density in identical small plots of ground. The plots were then weeded by hand to allow a fixed density of lamb's quarter plant. No other weed was allowed to grow. Here are the corn yields (in bushels per acre) for experimental plots controlled to have 1 weed per meter of planted corn (corn is always planted in rows) and 3 weeds per meter.[9]

1 weed/meter	166.2	157.3	166.7	161.1
3 weeds/meter	158.6	176.4	153.1	156.0

Explain carefully why a two-sample t confidence interval for the difference in mean yields may not be accurate.

Logging in the rain forest. "Conservationists have despaired over destruction of tropical rain forest by logging, clearing, and burning." These words begin a report on a statistical study of the effects of logging in Borneo.[7] Here are data on the number of tree species in 12 unlogged forest plots and 9 similar plots logged 8 years earlier:

Unlogged	22	18	22	20	15	21	13	13	19	13	19	15
Logged	17	4	18	14	18	15	15	10	12			

(a) The study report says, "Loggers were unaware that the effects of logging would be assessed." Why is this important? The study report also explains why the plots can be considered to be randomly assigned.

(b) Does logging significantly reduce the mean number of species in a plot after 8 years? Follow the four-step process as illustrated in Examples 18.2 and 18.3.

Logging in the rain forest, continued. Use the data in the previous exercise to give a 90% confidence interval for the difference in mean number of species between unlogged and logged plots.

Echinacea for the common cold? Echinacea is widely used as an herbal remedy for the common cold, but does it work? In a double-blind experiment, healthy volunteers agreed to be exposed to common-cold-causing rhinovirus type 39 and have their symptoms monitored. The volunteers were randomly assigned to take either a placebo or an echinacea supplement daily from 7 days before till 5 days after viral exposure. A symptom score was recorded for each subject over the 5 days following exposure, with higher scores indicating more severe symptoms. The published results reported the mean ± SEM for both groups as 13.21 ± 1.91 (echinacea) and 15.05 ± 1.43 (placebo).[6]

$\bar{x}_1$

$\bar{x}_2$

(a) The two-sample t statistic for $\bar{x}_1 - \bar{x}_2$ was $t = -0.771$. You can draw a conclusion from this t without using a table and even without knowing the sizes of the samples (remember to specify your null and alternative hypotheses first). What is your conclusion? Why don't you need the sample sizes and a table?

$\to n_1$ $\to n_2$

(b) In fact, 52 subjects were assigned to the echinacea treatment and 103 to the placebo. Fill in the values in this summary table:

Group	Treatment	n	$\bar{x}$	s
1	Echinacea	?	?	?
2	Placebo	?	?	?

$\to df = 52-1 = 51$

What degrees of freedom would you use in the conservative two-sample t procedures recommended for use without software? What P-value would you get from Table C?

group: Echinacea $\Rightarrow 1.91 = t_{51} \cdot \frac{s_1}{\sqrt{n_1}} \Rightarrow$ 求 s_1

placebo $\Rightarrow 1.43 = t_{102} \cdot \frac{s_2}{\sqrt{n_2}} \Rightarrow$ 求 s_2

用 $t_{obs} = -0.771$ 在 T_{51} 的位置求 p-value