

David S. Moore

The Basic Practice of Statistics

Third Edition

Chapter 17:

Two-Sample Problems



Will Cover

- Two-sample problems
- Comparing two population means
- Two-sample t procedures
- Examples of the two-sample t procedures
- Using technology
- Robustness again
- Details of the t approximation
- Avoid the pooled two-sample t procedures
- Avoid inference about standard deviations
- The F test for comparing two standard deviations

Two sample problems

- The goal of inference is to compare the responses to two treatments or to compare the characteristics of two populations.
- We have a separate sample from each treatment or each population.
 - We may wish to compare either the centers or the spreads of the two groups in a two-sample setting.

Eg.17.1 Two-sample problems

- A medical researcher is interested in the effect on blood pressure of added calcium in our diet. She conducts a randomized comparative experiment in which one group of subjects receives a calcium supplement and a control group gets a placebo.
- A psychologist develops a test that measures social insight. He compares the social insight of male college students with that of female college students by giving the test to a sample of students of each gender.
- A bank wants to know which of two incentive plans will most increase that use of its credit cards. It offers each incentive to a random sample of credit card customers and compares the amount charged during the following six months.

Sounds good-but no comparison

- Most women have mammograms to check for breast cancer once they reach middle age. Could a fancier test do a better job of finding cancers early? PET scans are a fancier (and more expensive) test. Doctors used PET scans on 14 women with tumors and got the detailed diagnosis right in 12 cases. That's promising. But there were no controls, and 14 cases are not statistically significant. Medical standards require randomized comparative experiments and statistically significant results. Only then can we be confident that the fancy test really is better.



Conditions for 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.

Comparing two population means

- 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	Mean	Standard deviation
1	x_1	μ_1	σ_1
2	x_2	μ_2	σ_2

Inferences

- The inference is giving a confidence interval for their difference $\mu_1 - \mu_2$ or testing the hypothesis of no difference, $H_0: \mu_1 = \mu_2$.

Populatio n	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 the inference, we start from the difference $\bar{x}_1 - \bar{x}_2$ between the means of the tow samples.

Eg.17.2 Does polyester decay?

- 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. Lower strength means the fabric have decayed.
 - Part of the study buried 10 polyester strips 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 the breaking strengths in pounds:

	2 weeks	118	126	126	120	129
–	16 weeks	124	98	110	140	110

Eg.17.2 Does polyester decay?

- 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

- The observed difference in mean strengths is
 $\bar{x}_1 - \bar{x}_2 = 123.80 - 116.40 = 7.40$ *pounds*
- Is this good evidence that polyester decays more in 16 weeks than in 2 weeks?

Check the condition

- Let μ_1 be the mean breaking in strengths in the entire population of polyester fabric buried for 2 weeks, and μ_2 for fabric buried for 16 weeks.

- The hypotheses are $H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 > \mu_2$
 - Because of the randomization, we are willing to regard the two groups of fabric strips as two independent SRSs.
- The back-to back stemplot of the responses in the right.
 - There are no departures from Normality that prevent the use of t procedures.

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

Two-sample t procedures

- To take the variation into account, we would like to standardize the observed difference $\bar{x}_1 - \bar{x}_2$ by dividing by its standard deviation.
 - The standard deviation is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
 - The estimated standard deviation as known as standard error, in short as SE:

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

Two-sample t statistic

- Two-sample t statistic $t = \frac{\bar{x}_1 - \bar{x}_2}{SE}$
 - It says how far $\bar{x}_1 - \bar{x}_2$ is from 0 in standard deviation units.
- The Two-sample t statistic has approximately a t distribution. The approximation is very accurate, but not easy to use. 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.

The two-sample t procedures

- Draw an SRSs of size n_1 from a Normal populations with unknown mean μ_1 and draw an SRSs of size n_2 from another Normal populations with unknown mean μ_2 .
 - Confidence interval
 - Testing

The two-sample t confidence interval

- A **confidence interval** for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here t^* is the $t(k)$ critical value for the density curve with area C between $-t^*$ and t^* . The degrees of freedom k are equal to the smaller of $n_1 - 1$ and $n_2 - 1$. This interval has confidence level at least C no matter what the population standard deviations may be.

- The two-sample t confidence interval again has the form
estimate $\pm t^* \text{SE}_{\text{estimate}}$

The two-sample t test

- To test the hypothesis $H_0: \mu_1 = \mu_2$, calculate the two-sample t statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

And use P-values or critical values for the $t(k)$ distribution. The true P-values or fixed significance level will always be *equal to or less than* the value calculated from $t(k)$ no matter what value the unknown population standard deviations have.

Eg.17.3 Does polyester decay?

- The test statistic for the null hypothesis $H_0: \mu_1 = \mu_2$ is

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

- Use the conservative Option 2. That is, use the t table with 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 17.1 illustrates this P-value. Table C shows that it lies between 0.15 and 0.20.
- Conclusion: the experiment did not find convincing evidence that polyester decays more in 16 weeks than in 2 weeks.

df = 4

p	.20	.15
t*	0.941	1.190

Figure 17.1

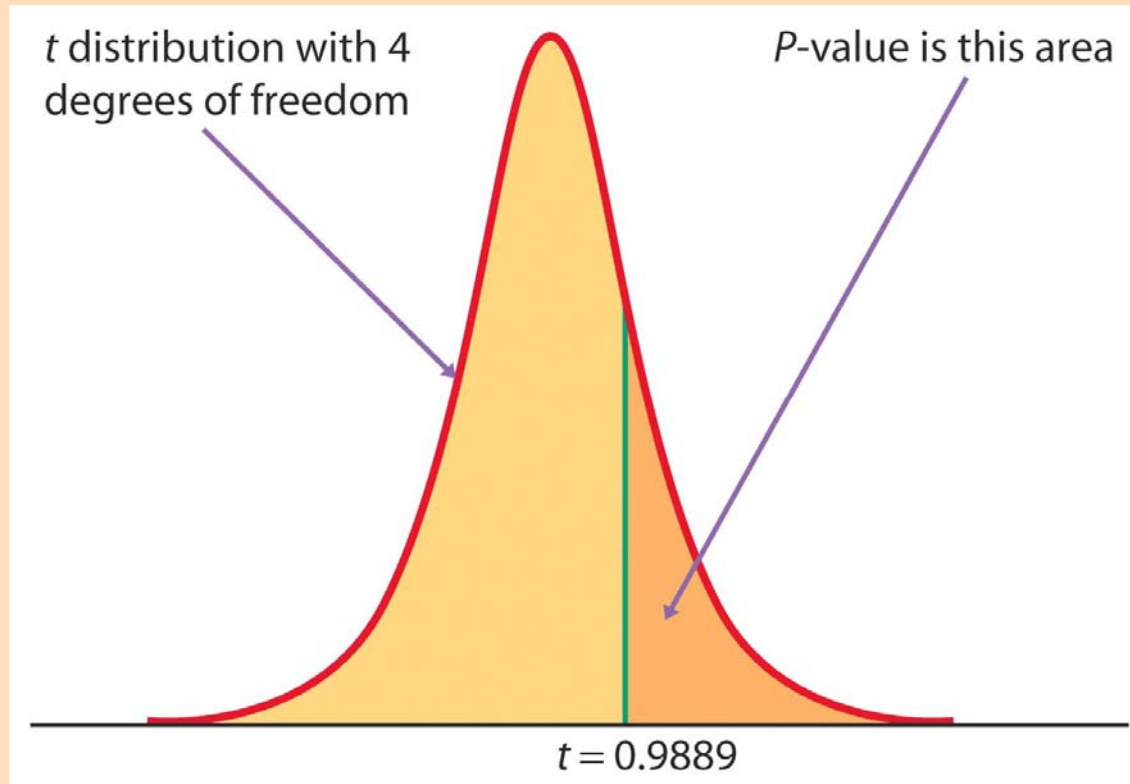


Figure 17.1 The P -value in Eg.17.3. This example uses the conservative Option2, which leads to the t distribution with 4 degrees of freedom.

Confidence interval

- For a 90% confidence interval, Table C shows that the $t(4)$ critical value is $t^* = 2.132$. We are 90% confident that the mean strength change between 2 and 16 weeks, $\mu_1 - \mu_2$, lies in the interval

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= (123.8 - 116.4) \pm 2.132 \sqrt{\frac{4.60^2}{5} + \frac{16.09^2}{5}}$$

$$= 7.40 \pm 15.96 = -8.56 \text{ to } 23.36$$

- The 90% confidence interval covers 0 tells us that we cannot reject $H_0: \mu_1 - \mu_2$ in favor of the two-sided alternative at the $\alpha = 0.10$ level of significance.

Meta-analysis

- Small samples have large margins of error. Large samples are expensive. Often we can find several studies of the same issue; if we could combine their results, we would have a large sample with a small margin of error. That is the idea of “meta-analysis.” Of course, we can’t just lump the studies together, because of differences in design and quality. Statisticians have more sophisticated ways of combining the results. Meta-analysis has been applied to issues ranging from the effect of secondhand smoke to whether coaching improves SAT scores.



Eg.17.4 Community service and attachment to friends

- Do college students who have volunteered for community service work differ from those who have not? A study obtained data from 57 students who had done service work and 17 who had not. One of the response variables was a measure of attachment to friends. (roughly, secure relationships). Measured by the Inventory of Parent and Peer Attachment. Here are the results:

Group	Condition	n	\bar{x}	s
1	Service	57	105.32	14.68
2	No service	17	96.82	14.26

- The paper reporting the study results applied a two-sample t test.

Testing procedure

- Step 1. Hypotheses. The hypotheses are

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_1: \mu_1 \neq \mu_2$$

- Step 2. Test statistics. The two-sample t statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{105.32 - 96.82}{\sqrt{\frac{14.68^2}{57} + \frac{14.26^2}{17}}} = \frac{8.5}{3.9677} = 2.142$$

- Step 3. P-value. Without software, use option 2. There are 16 degrees of freedom, the smaller of $n_1 - 1 = 56$ and $n_2 - 1 = 16$
 - Figure 17.2 illustrates the P-value. Find it by comparing 2.142 with critical values for the $t(16)$ distribution and then doubling p because the alternative is two-sided. Table C shows that $t = 2.142$ lies between the 0.025 and 0.02 critical values. The P-value is therefore between 0.05 and 0.04. The data give moderately strong evidence that students who have engaged in community service are on the average more attached to their friends.

df=16

p	.025	.02
t*	2.120	2.235

Figure 17.2

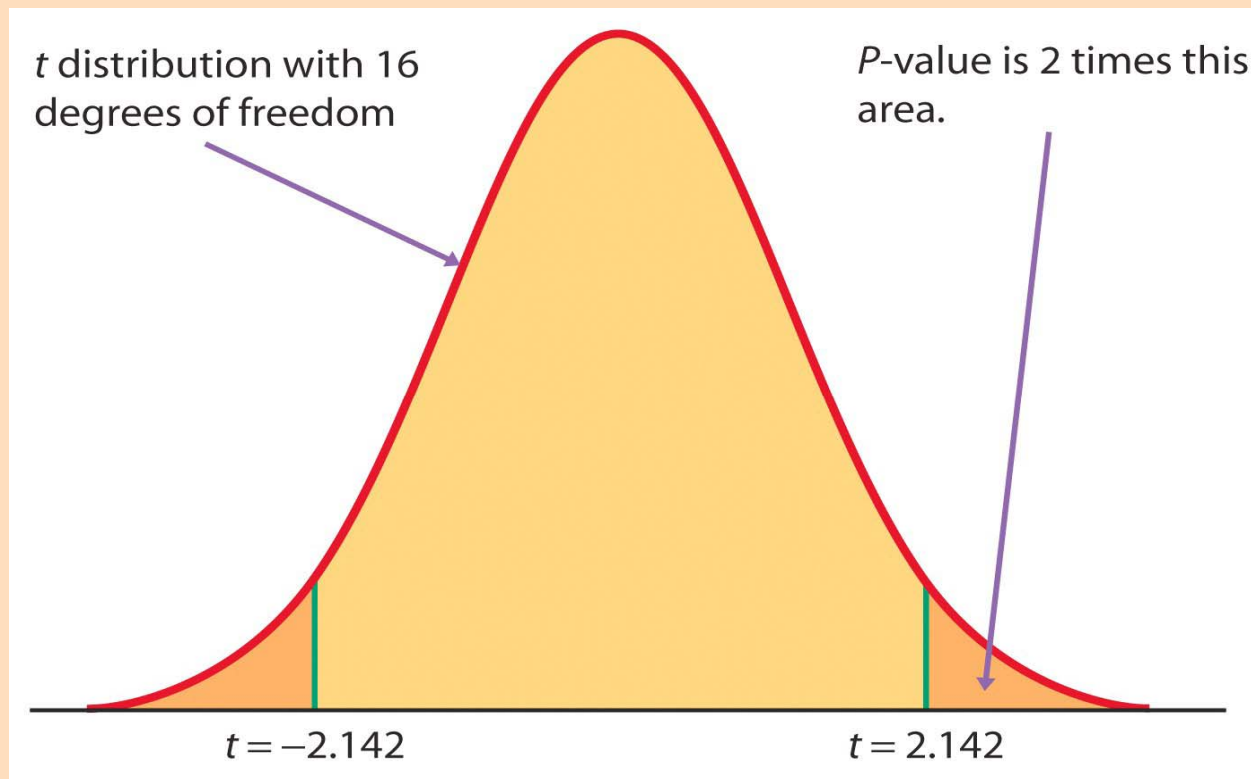


Figure 17.2 The P -value in Eg.17.4. To find P , find the area above $t = 2.142$ and double it because the alternative is two-sided.

Using technology

- Software shall use option 1 to give more accurate confidence intervals and P-values.
 - However, there is variation in how well software implements the t distribution used in Option 1.
 - The accurate approximation uses the t distribution with 4.65 degrees of freedom.
 - The p-value for $t = 0.9889$ is $P = 0.1857$.
 - The 90% confidence interval for $\mu_1 - \mu_2$ is -7.933 to 22.733 .

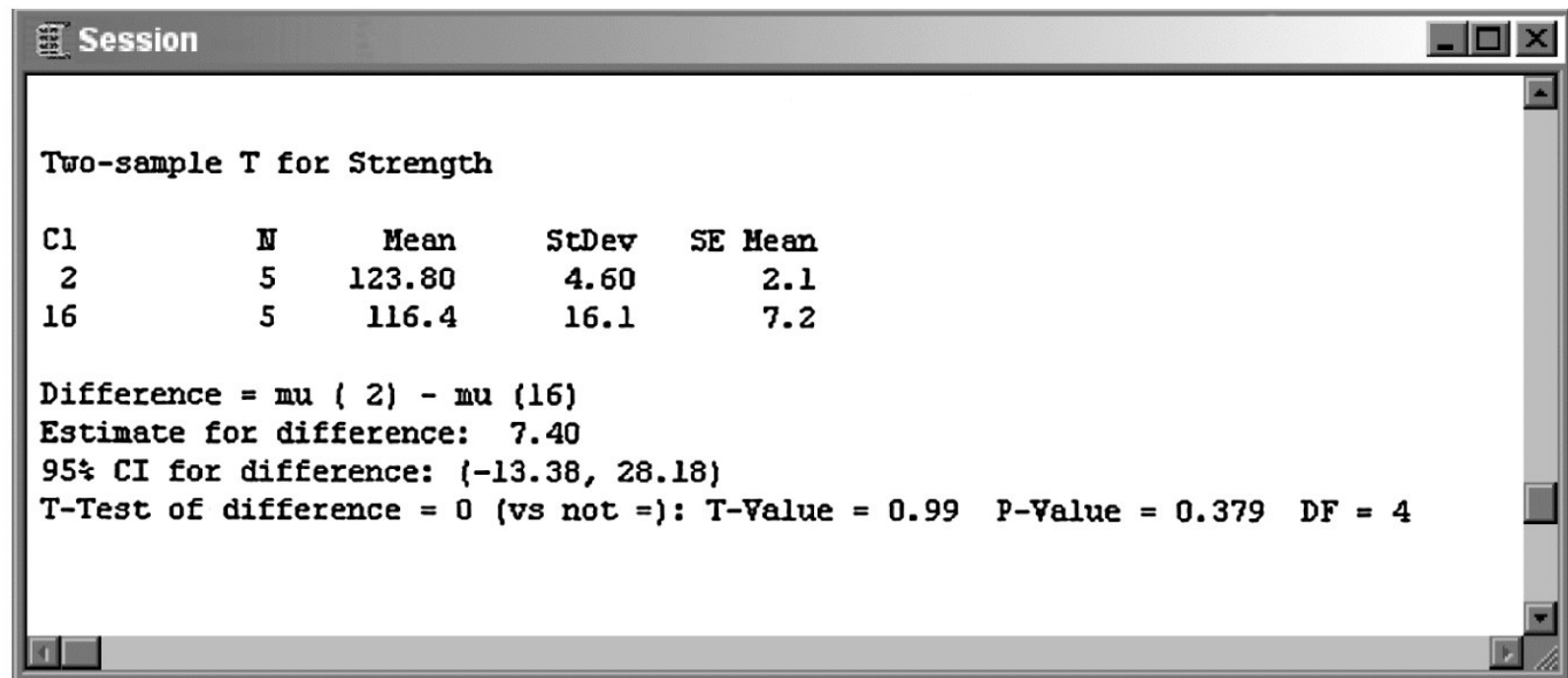
Using technology (cont.)

- Minitab uses option 1, but it truncates the exact degrees of freedom to the next smaller whole number to get critical values and P-values. ▶
- $Df = 4.65$ truncate to 4, agree with option 2.
- Excel rounds the exact degrees of freedom to the nearest whole number. Excel gives only the test, not the confidence interval. ▶
- The $df = 4.65$ becomes $df = 5$. P-value $p = 0.1841$ is therefore slightly smaller than is correct. The evidence against H_0 is stronger than is actually the case.
- Excel label for the test, “Two-Sample Assuming Unequal Variance,” is seriously misleading.
- TI-83 gets Option 1 completely right. ▶
- The two-sample t procedures we have described work whether or not the two populations have the same variance.



Minitab output

Minitab



The image shows a screenshot of the Minitab 'Session' window. The window title is 'Session'. The content displays the results of a 'Two-sample T for Strength' test. The results are presented in a table with columns: C1, N, Mean, StDev, and SE Mean. Below the table, the difference between the means is calculated, followed by the estimate for the difference, the 95% confidence interval (CI) for the difference, and the T-Test results (T-Value, P-Value, and DF).

C1	N	Mean	StDev	SE Mean
2	5	123.80	4.60	2.1
16	5	116.4	16.1	7.2

Difference = $\mu(2) - \mu(16)$
Estimate for difference: 7.40
95% CI for difference: (-13.38, 28.18)
T-Test of difference = 0 (vs not =): T-Value = 0.99 P-Value = 0.379 DF = 4



Excel output

Excel

Microsoft Excel - Book1			
	A	B	C
1	t-Test: Two-Sample Assuming Unequal Variances		
2			
3		2 weeks	16 weeks
4	Mean	123.8	116.4
5	Variance	21.2	258.8
6	Observations	5	5
7	Hypothesized Mean Difference	0	
8	df	5	
9	t Stat	0.9889	
10	P(T<=t) one-tail	0.1841	
11	t Critical one-tail	2.0150	
12	P(T<=t) Two-tail	0.3681	
13	t Critical two-tail	2.5706	
14			
15			



TI-83 output

Texas Instruments TI-83 Plus

2-SampTInt

(-7.933, 22.733)

df=4.6510

$\bar{x}_1=123.8000$

$\bar{x}_2=116.4000$

$S_{x1}=4.6043$

$\downarrow S_{x2}=16.0873$

2-SampTTest

$\mu_1 > \mu_2$

t=.9889

P=.1857

df=4.6510

$\bar{x}_1=123.8000$

$\downarrow \bar{x}_2=116.4000$



Robustness again

- 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 $n_1 = n_2 = 5$. When the two populations distributions have different shapes. Larger sample are needed.
- The two-sample t procedures are most robust against non-Normality in this case, and the conservative probability values are most accurate.

Details of the t approximation

- The exact distribution of the two-sample t statistic is not a t distribution. Moreover, the distribution changes as the unknown population standard deviations σ_1 and σ_2 change. However, an excellent approximation is available. We called this Option 1 for t procedures.

Approximate distribution of the two-sample t statistic

- The distribution of the two sample t statistic is very close to the t distribution with degrees of freedom df given by

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

- The approximation is accurate when both sample sizes n_1 and n_2 are 5 or larger.

Eg.17.5 Does polyester decay?

- In the experiment of Examples 17.2 and 17.3, the data on buried polyester fabric gave

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

- The two-sample t test statistic calculated from these values is $t = 0.9889$.

- Option 1 finds a vary accurate P-value by using the t distribution with degrees of freedom df given by

$$df = \frac{\left(\frac{4.60^2}{5} + \frac{16.09^2}{5} \right)^2}{\frac{1}{4} \left(\frac{4.60^2}{5} \right)^2 + \frac{1}{4} \left(\frac{16.09^2}{5} \right)^2} = \frac{3137.08}{674.71} = 4.65$$

- The degrees of freedom df is generally not a whole number. It is always at least as large as the smaller of $n_1 - 2$ and $n_2 - 1$. The larger degrees of freedom that results from Option 1 give slightly shorter confidence intervals and slightly smaller P-values than the conservative Option2 produces.

Avoid the pooled two-sample t procedures

- Most software, including Minitab, Excel, and the TI-83, offers two two-sample t statistics. One is often labeled for “unequal” variances, the other for “equal” variances.
 - The “unequal” variance procedure is our two-sample t . This test is valid whether or not the population variances are equal.
 - The other choice is a special version of the two-sample t statistic that assumes that the two populations have the same variance. This procedure “pooled” the two sample variances to estimate the common population variance. The resulting statistic is called the *pooled two-sample t statistic*,
- In the real world, distributions are not exactly Normal and population variances are not exactly equal. In practice, the Option2 t procedures are almost always more accurate than the pooled procedures.

Avoid inference about standard deviations

- There are methods for inference about the standard deviations of Normal populations.
 - The F test for comparing the spread of two Normal populations.
- Unlike the t procedures for means, the F test and other procedures for standard deviations are extremely sensitive to non-Normal distribution. This lack of robustness does not improve in large samples.
- It is difficult in practice to tell whether a significant F -value is evidence of unequal population spreads or simple a sign that the populations are not Normal.

The F test for comparing two standard deviations

- Suppose we have independent SRSs from two Normal populations, a sample of size n_1 from $N(\mu_1, \sigma_1)$ and a sample of size n_2 from $N(\mu_2, \sigma_2)$. The population means and standard deviations are all unknown. The two-sample t test examines whether the means are equal in this setting.
- To test the hypothesis of equal spread,
$$H_0: \sigma_1 = \sigma_2 \quad \text{vs.} \quad H_1: \sigma_1 \neq \sigma_2$$

we use the ratio of sample variances. This is the F statistic.

F procedures

- When s_1^2 and s_2^2 are sample variances from independent SRS of size n_1 and n_2 drawn from Normal populations, the **F statistic**

$$F = \frac{s_1^2}{s_2^2}$$

has the **F distribution** with $n_1 - 1$ and $n_2 - 1$ degrees of freedom when $H_0: \sigma_1 = \sigma_2$ is true.

Figure 17.5

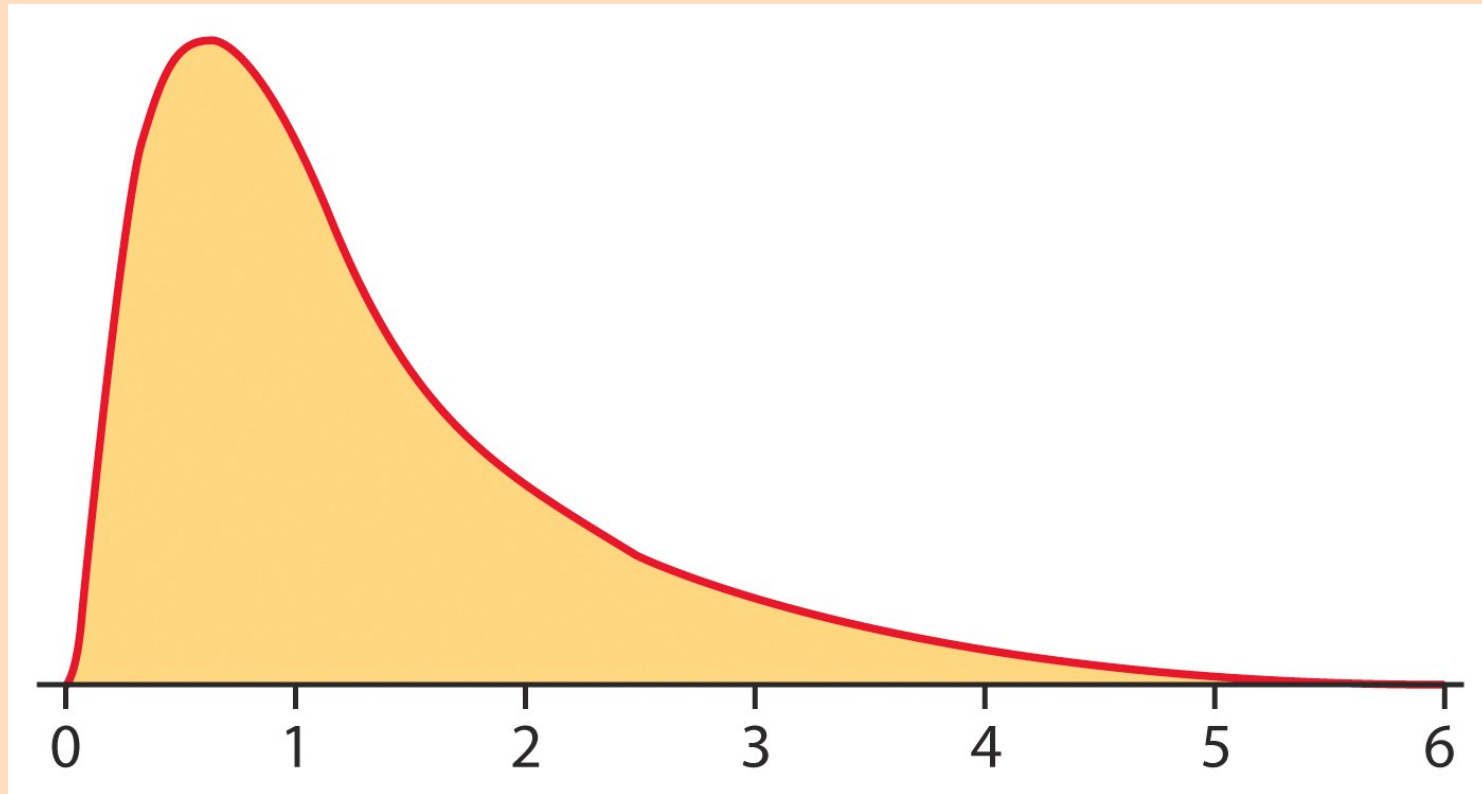


Figure 17.5 The density curve for the $F(9,10)$ distribution. The F distributions are skewed to the right.

Tables of F critical points

- We need a separate table for every pair of degrees of freedom j and k . Table D in the back of the book gives upper p critical points of the F distributions for $p = 0.10, 0.05, 0.025, 0.01$ and 0.001 .
- For example, these critical points for the $F(9, 10)$ distribution shown in Figure 17.5 are

p	.10	.05	.025	.01	.001
F^*	2.35	3.02	3.78	4.94	8.96

Carry out the F test

- Step 1. Take the test statistic to be

$$F = \frac{\text{larger } s^2}{\text{smaller } s^2}$$

This amounts to naming the populations so that Population 1 has the larger of the observed sample variances. The resulting F is always 1 or greater.

- Step 2. Compare the value of F with critical values from Table D. Then double the significance levels from the table to obtain the significance level for the two-sided F test.

Eg.17.6 Comparing variability

- Here are data summaries from Example 17.2:

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

- We might also compare the standard deviations to see whether strength loss is more or less variable after 16 weeks. We want to test

$$H_0: \sigma_1 = \sigma_2 \text{ vs. } H_1: \sigma_1 \neq \sigma_2$$

Eg.17.6 Cont.

- Note that we relabeled the groups so that Group 1 (16 weeks) has the larger standard deviation. The F test statistic is

$$F = \frac{\text{larger } s^2}{\text{smaller } s^2} = \frac{16.09^2}{4.60^2} = 12.23$$

- Compare the calculated value $F = 12.23$ with critical points for the $F(4, 4)$ distribution. Table D shows that 12.23 lies between the 0.025 and 0.01 critical values of the $F(4, 4)$ distribution. So the two-sided P-value lies between 0.05 and 0.02. The data show significantly unequal spreads at the 5% level. The P-value depends heavily on the assumption that both samples come from Normally distributed populations.

Summary

- The data in a two-sample problem are two independent SRSs, each drawn from a separate population.
- Draw independent SRSs of sizes n_1 and n_2 from two Normal populations with parameters μ_1, σ_1 and μ_2, σ_2 . The two-sample t statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

The statistic t has approximately a t distribution.

- For conservative inference procedures to compare use the two-sample t statistic with the $t(k)$ distribution. The degrees of freedom k is the smaller of $n_1 - 1$ and $n_2 - 1$. Software produces more accurate probability values from the $t(df)$ distribution with degrees of freedom df calculated from the data.

Summary II

- The **confidence interval** for $\mu_1 - \mu_2$ is

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

the confidence level is very close to C if t^* is the $t(\text{df})$ critical value. It is guaranteed to be at least C if t^* is the $t(k)$ critical value.

- **Significance tests** for $H_0: \mu_1 = \mu_2$ are based on

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- P-values calculated from the $t(\text{df})$ distribution are very accurate. A P-value calculated from $t(k)$ is slightly larger than the true P.

Summary III

- Inference procedures for comparing the standard deviations for two Normal populations are based on the ***F* statistic**, which is the ratio of sample variances

$$F = \frac{s_1^2}{s_2^2}$$

- If an SRS of size n_1 is drawn from Population1 and an independent SRS of size n_2 is drawn from Population2, the F statistic has the ***F* distribution** $F(n_1 - 1, n_2 - 1)$ if the two population standard deviation σ_1 and σ_2 are in fact equal.
- The *F* test for $H_0: \sigma_1 = \sigma_2$ and other procedures for inference on the spread of one or more Normal distributions are so strongly affected by lack of Normality that we do not recommend them for regular use.