Chapter 5
Introductory Inference

In this chapter we discuss selected topics and issues dealing with statistical inferences from samples to populations, building upon the brief introduction to these ideas in Chapter 3. The discussion here is at an intermediate technical level and at a speed appropriate for review of material learned in the prerequisite course.

We provide procedures for constructing confidence intervals and conducting hypothesis tests for several frequently encountered situations.

5.1 Normal (z) Intervals and Tests

A confidence interval and test concerning a population mean were briefly described in Chapter 3. This is a more extensive presentation.

The confidence interval on the mean μ of a normal population when the standard deviation is known was given in Equation (3.18). The development there assumed that the population was normal. However, since the Central Limit Theorem discussed in Section 3.4.2 guarantees that $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ is approximately normally distributed if n is "sufficiently large", the interval

$$\left(\bar{y} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}},\ \bar{y} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) \tag{5.1}$$

is an approximate two-sided $100(1-\alpha)\%$ confidence interval when the population is not normal. The closer the population is to a normal population, the closer will be this interval's coverage probability to $1-\alpha$. Thus, in the nonnormal case, this interval is an approximate CI for μ.

© Springer Science+Business Media New York 2015

R.M. Heiberger, B. Holland, *Statistical Analysis and Data Display*,
Springer Texts in Statistics, DOI 10.1007/978-1-4939-2122-5_5

Also shown in the rightmost column of Table 5.1 are one-sided confidence intervals for μ. These are less commonly used than two-sided intervals because they have infinite width. But they are sometimes encountered in contexts where an upper or lower bound for μ is required.

5.1.1 Test of a Hypothesis Concerning the Mean of a Population Having Known Standard Deviation

We consider three pairs of null and alternative hypotheses in Table 5.1 and Figure 5.1.

Table 5.1 Confidence intervals and tests with known standard deviation σ, where $\sigma_{\bar{y}} = \dfrac{\sigma}{\sqrt{n}}$ and $z_{\text{calc}} = \dfrac{\bar{y} - \mu_0}{\sigma_{\bar{y}}}$. The six situations are shown graphically in Figure 5.1

		Tests				
H_0	H_1	Rejection Region		p-value	Confidence Interval	
		z-scale	y-scale		Lower	Upper
$\mu \le \mu_0$	$\mu > \mu_0$	$z_{\text{calc}} > \ z_\alpha$	$\bar{y} > \mu_0 + z_\alpha \sigma_{\bar{y}}$	$P(Z > z_{\text{calc}})$	$(\ \bar{y} - z_\alpha \sigma_{\bar{y}}\ ,$	$\infty \qquad)$
$\mu \ge \mu_0$	$\mu < \mu_0$	$z_{\text{calc}} < -z_\alpha$	$\bar{y} < \mu_0 - z_\alpha \sigma_{\bar{y}}$	$P(Z < z_{\text{calc}})$	$(\qquad -\infty\ ,$	$\bar{y} + z_\alpha \sigma_{\bar{y}}\ )$
$\mu = \mu_0$	$\mu \ne \mu_0$	$\lvert z_{\text{calc}} \rvert > \ z_{\frac{\alpha}{2}}$	$\lvert \bar{y} - \mu_0 \rvert > z_{\frac{\alpha}{2}} \sigma_{\bar{y}}$	$2P(Z > \lvert z_{\text{calc}} \rvert)$	$(\bar{y} - z_{\frac{\alpha}{2}} \sigma_{\bar{y}}\ ,$	$\bar{y} + z_{\frac{\alpha}{2}} \sigma_{\bar{y}})$

The first two pairs are called *one-tailed* or *one-sided* tests because their rejection regions lie on one side of the normal distribution. The third pair has a two-sided rejection region and hence is termed a two-tailed or two-sided test. In any given problem, only one of these three is applicable. For expository purposes, it is convenient to discuss them together.

Some authors formulate the one-sided tests with sharp null hypotheses

H_0	H_1
$\mu = \mu_0$	$\mu > \mu_0$
$\mu = \mu_0$	$\mu < \mu_0$

However, with the sharp formulation it can happen that neither the null nor alternative hypothesis is true, in which case the action of rejecting the null hypothesis has an uncertain interpretation.

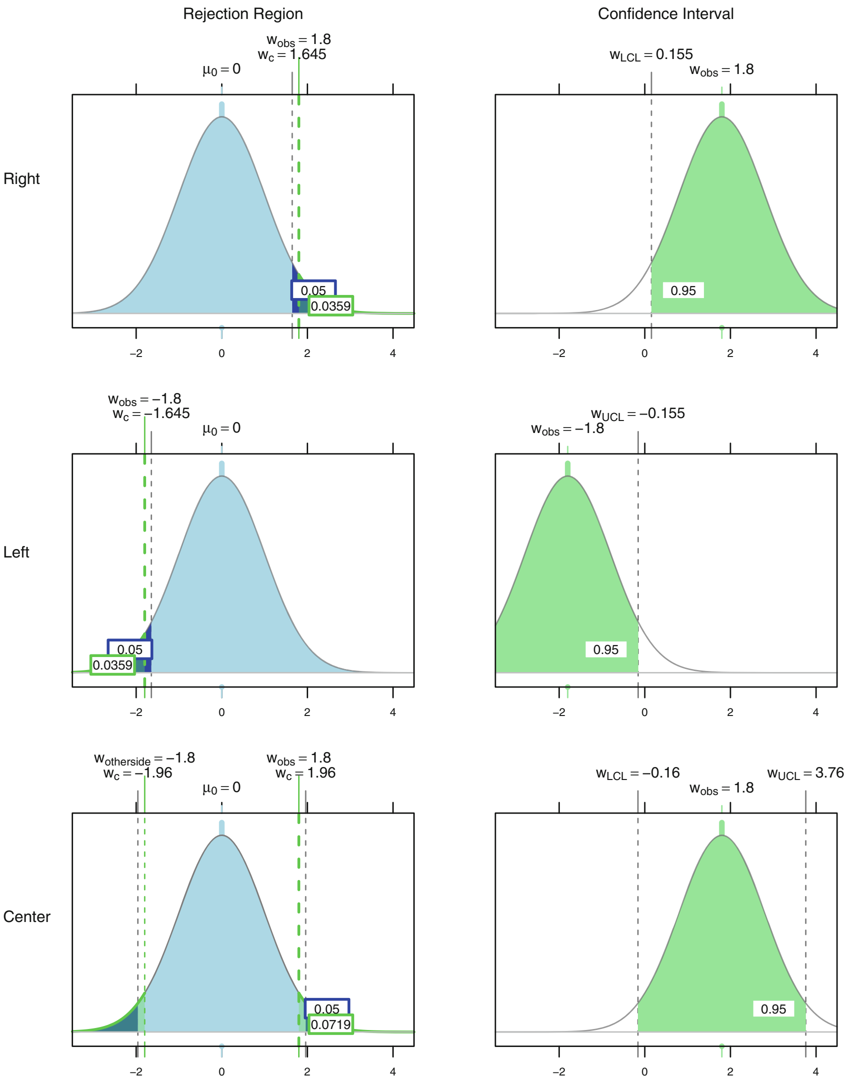


Fig. 5.1 Graphical display of the six situations described in Table 5.1: Confidence intervals and tests with known standard deviation σ. See Section 5.1.1 for full discussion.

For the first pair of hypotheses, we reject H_0 if the sample mean is sufficiently greater than μ_0, specifically, if $\bar{y} > (\mu_0 + z_\alpha \sigma / \sqrt{n})$. Otherwise, H_0 is retained. Equivalently, if we define the calculated Z statistic under the null hypothesis,

$$z_{\text{calc}} = \frac{\bar{y} - \mu_0}{\sigma / \sqrt{n}} \tag{5.2}$$

then we reject H_0 if $z_{\text{calc}} > z_\alpha$; otherwise H_0 is retained. The p-value of this test is $P(Z > z_{\text{calc}})$.

The testing procedure for the second pair of hypotheses is the mirror image of the first pair. H_0 is rejected if $\bar{y} < (\mu_0 - z_\alpha \sigma / \sqrt{n})$ and retained otherwise. Equivalently, we reject H_0 if $z_{\text{calc}} < -z_\alpha$. The p-value of this test is $P(Z < z_{\text{calc}})$.

For the third pair, the two-sided test, we reject H_0 if either

$$\bar{y} < (\mu_0 - z_{\frac{\alpha}{2}} \sigma / \sqrt{n}) \quad \text{or} \quad \bar{y} > (\mu_0 + z_{\frac{\alpha}{2}} \sigma / \sqrt{n});$$

equivalently, if $|z_{\text{calc}}| > z_{\frac{\alpha}{2}}$. The p-value of this two-sided test is $2P(Z > |z_{\text{calc}}|)$. Hence H_0 is rejected if $\bar{y}$ is sufficiently above or sufficiently below μ_0. Another equivalent rule is to reject H_0 if and only if μ_0 falls outside the $100(1-\alpha)\%$ confidence interval for μ.

The rejection region for all three pairs is included in Table 5.1.

5.1.2 Confidence Intervals for Unknown Population Proportion p

We consider a confidence interval on the unknown proportion p of successes in a population consisting of items or people labeled as successes and failures. Such populations are very frequently encountered in practice. For example, we might wish to estimate the proportion p of voters who will ultimately vote for a particular candidate, based on a random sample from a population of likely voters. Inspectors of industrial output may wish to estimate the proportion p of a day's output that is defective based on a random sampling of this output.

Suppose the sample size is n, of which Y items are successes and that $\hat{p} = \frac{Y}{n}$, a point estimator of p, is the proportion of sampled items that fall into the success category. Until recently, the usual $100(1-\alpha)\%$ confidence interval for p suggested in the statistics literature was

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

This interval is satisfactory when $n \geq 100$ unless p is close to either 0 or 1. The large sample is needed for the Central Limit Theorem to assure us that the discrete probability distribution of $\hat{p}$ is adequately approximated by the continuous normal distribution.

Agresti and Caffo (2000) suggest the following alternative confidence interval for p, where $\tilde{p} = \frac{Y+2}{n+4}$ and $\tilde{n} = n + 4$:

$$\tilde{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}\,(1-\tilde{p})}{\tilde{n}}} \tag{5.3}$$

Agresti and Caffo show that their interval has coverage probability that typically is much closer to the nominal $1 - \alpha$ than the usual confidence interval. It differs from the usual interval in that we artificially add two successes and two failures to the original sample. For p near 0 or 1, the usual interval, which is symmetric about $\hat{p}$, may extend beyond one of these extremes and hence not make sense, while the alternative interval is likely to remain entirely between 0 and 1.

Conventional one-sided confidence intervals for p are shown in Table 5.2. Comparable to Agresti and Caffo's proposed two-sided interval, Cai (2003) proposes improved one-sided confidence intervals for p having coverage probabilities closer to $1 - \alpha$ than the conventional intervals. These lower and upper intervals, respectively, are

$$\left[0, \mathcal{F}_{\text{Be}}^{-1}(1 - \alpha \mid Y + .5, n - Y + .5)\right] \tag{5.4}$$

and

$$\left[\mathcal{F}_{\text{Be}}^{-1}(\alpha \mid Y + .5, n - Y + .5), 1\right] \tag{5.5}$$

where $\mathcal{F}_{\text{Be}}^{-1}(\alpha \mid a, b)$ denotes the value x of a random variable corresponding to the 100α percentile of the beta distribution with parameters a and b. See Section J.1.1 for a brief discussion of the beta distribution.

5.1.3 Tests on an Unknown Population Proportion p

Assume we have a sample of $n \geq 100$ items from a population of successes and failures, and we wish to test a hypothesis about the proportion p of successes. Paralleling the previous discussion of tests on a population mean, there are two one-tailed tests and one two-tailed test as detailed in Table 5.2 and Figure 5.2. As in the discussion of the confidence interval on p, the normal approximation to the distribution of $\hat{p}$ requires that n not be too small. Note that the confidence intervals are based on densities centered on the observed proportion $\hat{p} = x/n$. They therefore have a different standard deviation $\sqrt{\hat{p}(1 - \hat{p})/n}$, and therefore height at the center of the density, than the densities centered at the null hypothesis p_0 with standard deviation $\sqrt{p_0(1 - p_0)/n}$.

5.1.4 Example—One-Sided Hypothesis Test Concerning a Population Proportion

As an illustration, suppose a pollster wishes to test the hypothesis that at least 50% of a city's voting population favors a certain bond issue. The pollster observed only

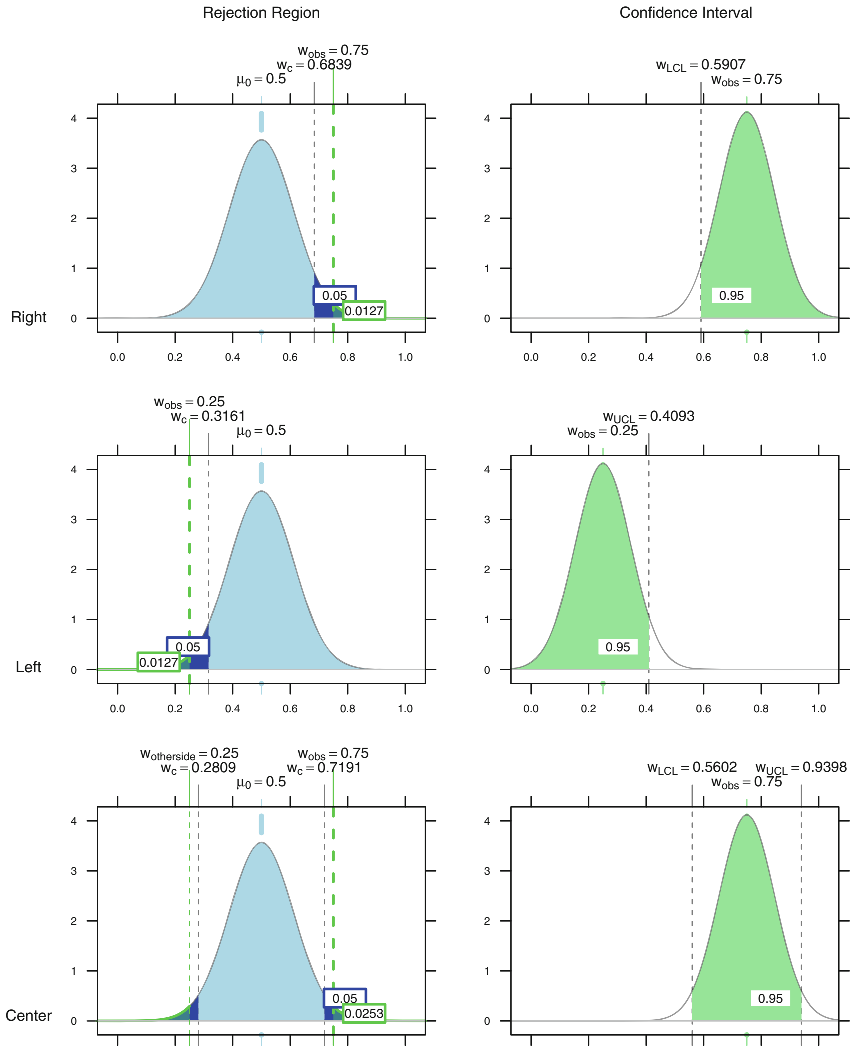


Fig. 5.2 Graphical display of the six situations described in Table 5.2: Confidence intervals and tests for population proportions. Note that the confidence intervals, centered on the observed $\hat{p}$, have differently scaled density functions than the null hypothesis distributions, centered on the hypothesized p_0. For the tests, the standard deviation is $\sigma_{p_0} = \sqrt{\big(p_0/(1 - p_0)\big)/n}$. For the confidence intervals, the standard deviation is $s_{\hat{p}} = \sqrt{\big(\hat{p}/(1 - \hat{p})\big)/n}$. In this example the densities for the confidence interval are taller and narrower. See Section 5.1.3 for full discussion.

Table 5.2 Conventional confidence intervals and tests with unknown population proportion p, where

$\sigma_{p_0} = \sqrt{\dfrac{p_0\,(1-p_0)}{n}}$ and $z_{\text{calc}} = \dfrac{\hat{p} - p_0}{\sigma_{p_0}}$ for tests, and $s_{\hat{p}} = \sqrt{\dfrac{\hat{p}\,(1-\hat{p})}{n}}$ for confidence intervals.

		Tests				
H_0	H_1	Rejection Region		p-value	Confidence Interval	
		z-scale	p-scale		Lower	Upper
$p \le p_0$	$p > p_0$	$z_{\text{calc}} > z_\alpha$	$\hat{p} > p_0 + z_\alpha \sigma_{p_0}$	$P(Z > z_{\text{calc}})$	$(\hat{p} - z_\alpha s_{\hat{p}}\,,$	$1 \quad)$
$p \ge p_0$	$p < p_0$	$z_{\text{calc}} < -z_\alpha$	$\hat{p} < p_0 - z_\alpha \sigma_{p_0}$	$P(Z < z_{\text{calc}})$	$(\quad 0\,,$	$\hat{p} + z_\alpha s_{\hat{p}})$
$p = p_0$	$p \ne p_0$	$\lvert z_{\text{calc}}\rvert > z_{\frac{\alpha}{2}}$	$\lvert\hat{p} - p_0\rvert > z_{\frac{\alpha}{2}} \sigma_{p_0}$	$2P(Z > \lvert z_{\text{calc}}\rvert)$	$(\hat{p} - z_{\frac{\alpha}{2}} s_{\hat{p}}\,,$	$\hat{p} + z_{\frac{\alpha}{2}} s_{\hat{p}})$

222 of a random sample of 500 persons in the population favors this bond issue. Let us conduct this test at $\alpha = 0.01$.

Here H_1 is of the form H_1: $p < .50$. We reject H_0 if

$$\hat{p} < p_0 - z_{.01}\sqrt{\frac{p_0(1-p_0)}{n}} \tag{5.6}$$

With $p_0 = .50$, $\hat{p} = 222/500 = 0.444$, $z_{.01} = 2.326$, and

$$\sqrt{p_0(1-p_0)/n} = .0224 \tag{5.7}$$

we find that the right side of (5.6) is 0.448 so that H_0 is (barely) rejected. In this example, $z_{\text{calc}} = -2.500$ so that the p-value $= P(Z < -2.500) = .0062$. Hence we reject H_0 because $\alpha = .01 > p = .0062$.

5.2 *t*-Intervals and Tests for the Mean of a Population Having Unknown Standard Deviation

When we wish to construct a confidence interval or test a hypothesis about an unknown population mean μ, more often than not the population standard deviation σ is also unknown. Then we must use the sample standard deviation $s = \sum\left((x - \bar{x})^2\right)/(n-1)$ from Equation 3.9 in place of σ when standardizing $\bar{y}$. But while $(\bar{y} - \mu)/(\sigma/\sqrt{n})$ has an approximate normal distribution if n is sufficiently large, $(\bar{y}-\mu)/(s/\sqrt{n})$ has an approximate t distribution with $n-1$ degrees-of-freedom. The latter standardization with s in the denominator has more variability than the former standardization with σ in the denominator. The t distribution reflects this increased variability because it has less probability concentrated near zero than does the standard normal distribution.

The confidence interval and tests for μ using the t distribution are similar to those using the normal (Z) distribution (that is, Table 5.1 is applicable), with t_{calc} replacing z_{calc} and t_α replacing z_α. For this problem, the degrees-of-freedom parameter for the t distribution is always $n - 1$.

For example, to test $H_0\colon \mu \geq \mu_0$ vs $H_1\colon \mu < \mu_0$, we reject H_0 if

$$t_{\text{calc}} = \frac{\bar{y} - \mu}{s/\sqrt{n}} < -t_\alpha \tag{5.8}$$

Here the p-value $= P(t < t_{\text{calc}})$ is calculated from the t distribution with $n-1$ degrees of freedom.

Calculating the power associated with t-tests is more difficult than for the normal tests because the alternative distribution is not the same as the null distribution. With the normal tests, both distributions have the same shape. With the t-tests, the alternative distribution has the *noncentral t* distribution with noncentrality parameter $(\mu_1 - \mu_0)/(\sigma/\sqrt{n})$. We postpone further discussion of the noncentral t distribution to Section 5.6.2 and Figure 5.10 in the context of sample size calculations. Also see the illustration in Section J.2.2.

The approximate confidence interval on μ is $\bar{y} \pm t_{\frac{\alpha}{2}} \dfrac{s}{\sqrt{n}}$.

5.2.1 Example—Inference on a Population Mean μ

Hand et al. (1994) presents a data set, reproduced in `data(vocab)`, containing the scores on a vocabulary test of a sample of 54 students from a study population. Assume that the test was constructed to have a mean score of 10 in the general population. We desire to assess whether the mean score of the study population is also $\mu = 10$. Assuming that standard deviation for the study population is not known, we wish to calculate a 95% confidence interval for μ and to test $H_0\colon \mu = 10$ vs $H_1\colon \mu \neq 10$.

We begin by looking at a stem-and-leaf display of the sample data to see if the underlying assumption of normality is tenable. We observe in Figure 5.3 that the sample is slightly positively skewed with one high value that may be considered an outlier. Based on the Central Limit Theorem, the t-based procedures in Figure 5.4 are justified here. The small p-value ($p \approx 3_{10}{}^{-14}$) is a strong evidence that μ is not 10. The 95% confidence interval (12.30, 13.44) suggests that the mean score is close to 12.9 in the study population.

We examine a nonparametric approach to this problem in Section 16.2.

```
> stem(vocab$score, scale=2)

  The decimal point is at the |

   9 | 0
  10 | 0000
  11 | 0000000000000
  12 | 0000000
  13 | 000000000
  14 | 000000000
  15 | 0000
  16 | 00000
  17 | 0
  18 |
  19 | 0
```

Fig. 5.3 Stem-and-leaf display of vocabulary scores.

5.3 Confidence Interval on the Variance or Standard Deviation of a Normal Population

Let the (unbiased) estimator of σ^2 based on a sample of size n be denoted s^2. Then $(n-1)s^2/\sigma^2$ has a χ^2 distribution with df $= n - 1$. Thus

$$P\left(\chi^2_{\frac{\alpha}{2},n-1} < (n-1)\, s^2/\sigma^2 < \chi^2_{1-\frac{\alpha}{2},n-1}\right) = 1 - \alpha$$

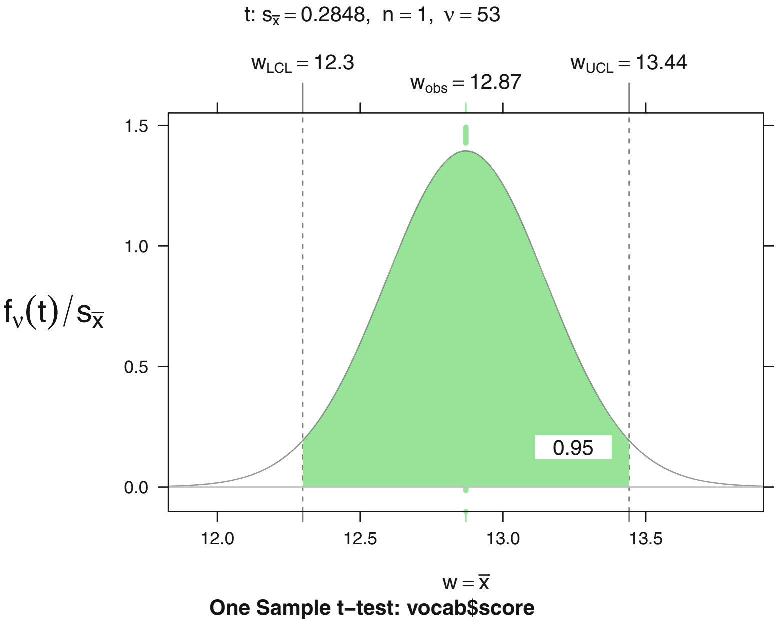
Inverting this statement leads to the $100(1-\alpha)\%$ confidence interval for σ^2:

$$\left(\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}},\ \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right)$$

If instead a CI on σ is desired, take the square roots of both the lower and upper limits in the above. We graph the estimation of a confidence interval in Figure 5.5.

The distribution of $(n-1)s^2/\sigma^2$ can also be used to conduct a test about σ^2 (or σ). For example, to test H_0: $\sigma^2 \le \sigma_0^2$ vs H_1: $\sigma^2 > \sigma_0^2$, the p-value is $1 - \mathcal{F}_{\chi^2_{n-1}}\left((n-1)s^2/\sigma_0^2\right)$. Tests of the equality of two or more variances are addressed in Section 6.10.

```
> vocab.t <- t.test(vocab$score, mu=10)

> vocab.t

One Sample t-test

data:  vocab$score
t = 10.08, df = 53, p-value = 6.372e-14
alternative hypothesis: true mean is not equal to 10
95 percent confidence interval:
 12.30 13.44
sample estimates:
mean of x
    12.87
```



	w_{obs}	w_{LCL}	w_{UCL}
$\bar{x}$	12.87	12.3	13.44
t	0	−2.005	2.009

	Left	Confidence	Right
Probability	0.025	0.95	0.025

Fig. 5.4 *t*-test and *t*-based confidence interval of vocabulary scores.

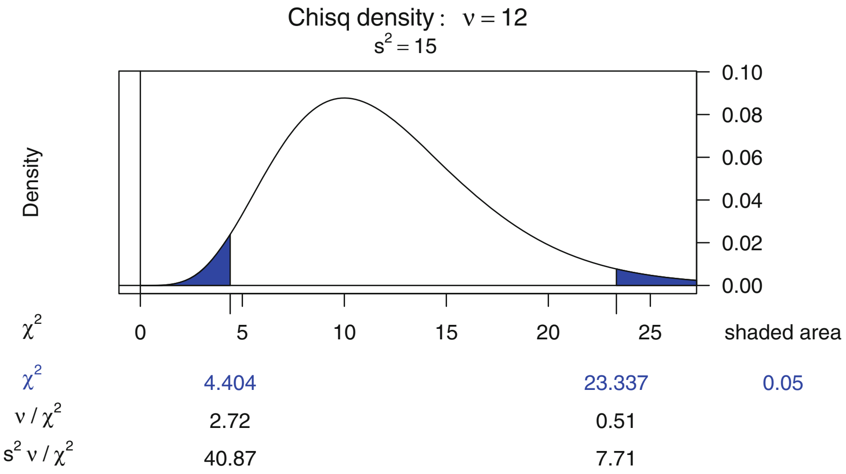


Fig. 5.5 Confidence interval for variance assuming a chi-square distribution with $\nu = 12$ degrees of freedom and an observed $s^2 = 15$. The estimated 95% confidence interval on σ^2 is $(7.71, 40.87)$. By taking the square root, we find the estimated 95% confidence interval on σ is $(2.777, 6.393)$.

5.4 Comparisons of Two Populations Based on Independent Samples

Two populations are often compared by constructing confidence intervals on the difference of the population means or proportions. In this discussion it is assumed that random samples are independently selected from each population.

5.4.1 Confidence Intervals on the Difference Between Two Population Proportions

The need for confidence intervals on the difference of two proportions is frequently encountered. We might wish to estimate the difference in the proportions of voters in two populations who favor a particular candidate, or the difference in the proportions of defectives produced by two company locations.

Labeling the populations as 1 and 2, the traditional confidence interval, assuming that both populations are large and that neither proportion is close to either 0 or 1, is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \tag{5.9}$$

Agresti and Caffo (2000) also provided an improved confidence interval for this situation, which again provides confidence closer to $100(1-\alpha)\%$ than the preceding interval. For $i = 1, 2$, let $\tilde{p}_i = \frac{Y_i+1}{n_i+2}$, i.e., revise the estimate of p_i by adding one success and one failure to both samples. Then the improved interval is

$$(\tilde{p}_1 - \tilde{p}_2) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}} \tag{5.10}$$

To test the null hypothesis H_0: $p_1 - p_2$ the appropriate statistic is

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{5.11}$$

where $\hat{p} = \dfrac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$.

Notice the distinction between the standard error portions of Equations 5.9 and 5.10. The standard error in the test statistic 5.11 is calculated under the assumption that the null hypothesis is true. The larger standard error in 5.9 cannot utilize this assumption.

5.4.2 Confidence Interval on the Difference Between Two Means

For a CI on a difference of two means under the assumption that the population variances are unknown, there are two cases. If the variances can be assumed to be equal, their common value is estimated as a weighted average of the two individual sample variances. In general, the process of calculating such meaningfully weighted averages is referred to as *pooling*, and the result in this context is called a pooled variance:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \tag{5.12}$$

The pooled estimator s_p^2 has more degrees of freedom (uses more information) than either s_1^2 or s_2^2 for the estimation of the common population variance. When the pooled variance is used as the denominator of F-tests it provides a more powerful test than either of the components, and therefore it is preferred for this purpose. Then the CI is

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\frac{\alpha}{2}, n_1+n_2-2}\, s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

In the case where the variances cannot be assumed equal, there are two procedures. The Satterthwaite option is

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\frac{\alpha}{2},\mathrm{df}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where df is the integer part of

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

The Satterthwaite option is sometimes referred to as the Welch option.

The Cochran option is

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

where $t = \dfrac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$, $w_i = s_i^2/n_i$, and t_i is $t_{\frac{\alpha}{2},(n_i-1)}$.

The Satterthwaite option is more commonly used than the Cochran option. In practice, they lead to similar results.

5.4.3 *Tests Comparing Two Population Means When the Samples Are Independent*

There are two situations to consider with independent samples. When the populations may be assumed to have a common unknown variance σ, the calculated t statistic is

$$t_{\mathrm{calc}} = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{5.13}$$

where s_p was defined in Equation (5.12) and t_{calc} has $n_1 + n_2 - 2$ degrees of freedom.

When the two samples might have different unknown variances, then the test is based on

$$s_{(\bar{y}_1-\bar{y}_2)} = \sqrt{\text{var}\,(\bar{y}_1-\bar{y}_2)} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \quad \text{and} \quad t_{\text{calc}} = \frac{\bar{y}_1-\bar{y}_2}{s_{(\bar{y}_1-\bar{y}_2)}} \tag{5.14}$$

In either case, we consider one of the three tests in Table 5.3.

Table 5.3 Confidence intervals and tests for two population means. When the samples are independent and we can assume a common unknown variance, use $s_{\Delta\bar{y}} = s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ and t_{calc} as given by Equation (5.13). When the samples are independent and we assume different unknown variances, use $s_{\Delta\bar{y}} = s_{(\bar{y}_1-\bar{y}_2)}$ and t_{calc} as given by Equation (5.14). When the samples are paired, use $s_{\Delta\bar{y}} = s_{\bar{d}}$ and t_{calc} as given by Equation (5.15).

		Tests		Confidence Interval	
		Rejection			
H_0	H_1	Region	p-value	Lower	Upper
$\mu_1 \le \mu_2$	$\mu_1 > \mu_2$	$t_{\text{calc}} > t_\alpha$	$P(t > t_{\text{calc}})$	$\big((\bar{y}_1-\bar{y}_2) - t_\alpha s_{\Delta\bar{y}}\,,$	$\infty\big)$
$\mu_1 \ge \mu_2$	$\mu_1 < \mu_2$	$t_{\text{calc}} < -t_\alpha$	$P(t < t_{\text{calc}})$	$\big(-\infty\,,$	$(\bar{y}_1-\bar{y}_2) + t_\alpha s_{\Delta\bar{y}}\big)$
$\mu_1 = \mu_2$	$\mu_1 \ne \mu_2$	$\lvert t_{\text{calc}}\rvert > t_{\frac{\alpha}{2}}$	$2P(t > \lvert t_{\text{calc}}\rvert)$	$\big((\bar{y}_1-\bar{y}_2) - t_{\frac{\alpha}{2}} s_{\Delta\bar{y}}\,,$	$(\bar{y}_1-\bar{y}_2) + t_{\frac{\alpha}{2}} s_{\Delta\bar{y}}\big)$

R uses the `t.test` function which calculates a one-sample, two-sample, or paired t-test, or a Welch modified two-sample t-test. The Welch modification is synonymous with the Satterthwaite method.

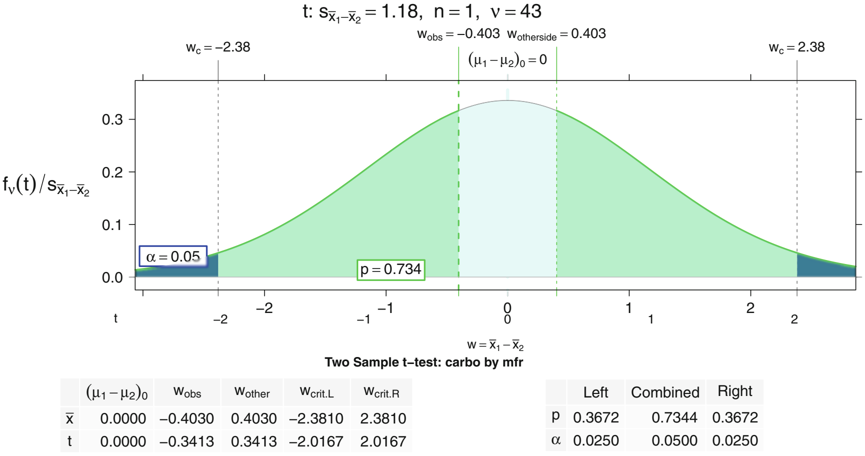
The example in Tables 5.4 and 5.5 and Figure 5.6 compares two means where the samples are independent and assumed to have a common unknown variance. Table 5.4 shows the t-test calculated with the `t.test` function. Table 5.5 calculates the t-value manually using the definitions in Equations 5.12 and 5.13. Figure 5.6 plots the result of the `t.test` with the `NTplot` function.

Table 5.4 Select the subset of the `cereals` dataset for “Cold cereal” and manufacturers “G” and “K”. Use `t.test` to compare their mean carbohydrate values assuming independent samples with a common unknown variance. The result from the `t.test` is plotted in Figure 5.6.

```
> data(cereals)

> table(cereals[,c("mfr","type")])
   type
mfr  C  H
  A  0  1
  G 22  0
  K 23  0
  N  5  1
  P  9  0
  Q  7  1
  R  8  0

> C.KG <- cereals$type=="C" & cereals$mfr %in% c("K","G")

> cerealsC <- cereals[C.KG, c("mfr", "carbo") ]

> cerealsC$mfr <- factor(cerealsC$mfr)

> bwplot(carbo ~ mfr, data=cerealsC) +
+ dotplot(carbo ~ mfr, data=cerealsC)

> t.t <- t.test(carbo ~ mfr, data=cerealsC, var.equal=TRUE)

> t.t

Two Sample t-test

data:  carbo by mfr
t = -0.3415, df = 43, p-value = 0.7344
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.784  1.978
sample estimates:
mean in group G mean in group K
          14.73           15.13
```

Table 5.5 The t-value $-.3415$ in Table 5.4 is calculated manually.

```
> mm <- tapply(cerealsC$carbo, cerealsC$mfr, mean)

> vv <- tapply(cerealsC$carbo, cerealsC$mfr, var)

> ll <- tapply(cerealsC$carbo, cerealsC$mfr, length)

> s2p <- ((ll-1) %*% vv) / sum(ll-1)

> tt <- -diff(mm) / (sqrt(s2p) * sqrt(sum(1/ll)))

> tt
         [,1]
[1,] -0.3415
```



	$(\mu_1-\mu_2)_0$	w_{obs}	w_{other}	$w_{crit.L}$	$w_{crit.R}$
$\bar{x}$	0.0000	−0.4030	0.4030	−2.3810	2.3810
t	0.0000	−0.3413	0.3413	−2.0167	2.0167

	Left	Combined	Right
p	0.3672	0.7344	0.3672
α	0.0250	0.0500	0.0250

Fig. 5.6 Show the `NTplot(t.t, zaxis=TRUE)` of the t-test in Table 5.4. There are two horizontal scales on the bottom axis of the plot. The $w = \bar{x}_1 - \bar{x}_2$ scale is the top scale and the t scale is the bottom scale. Specific interesting values in the w scale are identified on the top axis. $w_{\text{obs}} = -.403$ and its symmetrically placed $w_{\text{otherside}} = .403$ are very close to the center of the graph, illustrating that the observation is not anywhere near the rejection region $|W| > 2.38$.

5.4.4 Comparing the Variances of Two Normal Populations

We assume here that independent random samples are available from both populations. The F distribution is used to compare the variances σ_1^2 and σ_2^2 of two normal populations. Let s_1^2 and s_2^2 be the variances of independent random samples of size $n_i, i = 1, 2$ from these populations.

To test

$$H_0\colon \sigma_1^2 \le \sigma_2^2$$

vs

$$H_1\colon \sigma_1^2 > \sigma_2^2$$

define $F = s_1^2/s_2^2$ and reject H_0 if F is sufficiently large. The p-value of the test is $1 - \mathcal{F}_{F_{(n_1-1,n_2-1)}}(F)$. The power of this and other F-tests is sensitive to the second (denominator) df parameter and is usually not adequate unless this df ≥ 20.

A $100(1-\alpha)\%$ confidence interval for a ratio of variances of two normal populations, σ_1^2/σ_2^2, is

$$\left(\frac{s_1^2}{s_2^2}\frac{1}{F_{\text{low}}},\ \frac{s_1^2}{s_2^2}F_{\text{high}}\right)$$

where

F_{low} is $F_{1-\frac{\alpha}{2},n_1-1,n_2-1}$, the upper $100(1-\frac{\alpha}{2})$ percentage point of an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, and

F_{high} is $F_{1-\frac{\alpha}{2},n_2-1,n_1-1}$, the upper $100(1-\frac{\alpha}{2})$ percentage point of an F distribution with $n_2 - 1$ and $n_1 - 1$ degrees of freedom.

An extension to testing the homogeneity of more than two population variances will be presented in Section 6.10.

5.5 Paired Data

Sometimes we wish to compare the mean change in a measurement observed on an experimental unit under two different conditions. For example:

1. Compare the subject knowledge of students before and after they receive instruction on the subject.
2. Compare the yield per acre of a population of farms for a crop grown with two different fertilizers.
3. Compare the responses of patients to both an active drug and a placebo, when they are administered each of them in sequential random order with a suitable "washout" period between the administrations.

This “matched pairs” design is superior to a design of the same total size using independent samples because (in illustrations 1 and 3 above) the person to person variation is removed from the comparison of the two administrations, thereby improving the precision of this comparison. The principles of designing experiments to account for and remove extraneous sources of variation are discussed in more detail in Chapter 13.

It is assumed that the populations have a common variance and are approximately normal. Let $y_{11}, y_{12}, \ldots, y_{1n}$ be the sample of n items from the population under the first condition, having mean μ_1, and similarly let $y_{21}, y_{22}, \ldots, y_{2n}$ be the sample from the population under the second condition, having mean μ_2.

Define the n differences $d_1 = y_{11} - y_{21},\ d_2 = y_{12} - y_{22}, \ldots, d_n = y_{1n} - y_{2n}$. Let $\bar{d}$ and s_d be the mean and standard deviation, respectively, of the sample of n d_i's. Then an approximate $100(1-\alpha)\%$ confidence interval on the mean difference $\mu_1 - \mu_2$ is $\bar{d} \pm t_{\frac{\alpha}{2},n-1}\, s_{\bar{d}}$ where $s_{\bar{d}} = s_d / \sqrt{n}$. Tests of hypotheses proceed similarly to t-tests for two independent samples. Table 5.3 can still be used, but with

$$s_{\bar{d}} = s_d / \sqrt{n}, \quad \text{and} \quad t_{\text{calc}} = \frac{\bar{d}}{s_{\bar{d}}} \tag{5.15}$$

with degrees of freedom $n - 1$.

5.5.1 Example—t-test on Matched Pairs of Means

Woods et al. (1986), later in Hand et al. (1994), investigate whether native English speakers find it easier to learn Greek than native Greek speakers learning English. Thirty-two sentences are written in both languages. Each sentence is scored according to the quantity of errors made by an English speaker learning Greek and by a Greek speaker learning English. It is desired to compare the mean scores of the two groups. The data are available as `data(teachers)`; the first column is the error score on the English version of the sentence and the second column is the error score on the Greek version of the sentence.

These are 32 pairs of observations because the same sentence is evaluated in both languages. It would be incorrect to regard these as independent samples. The dotplot in Figure 5.7 reveals that for most sentences the English version shows fewer errors. The stem-and-leaf of the differences in Figure 5.8a shows the difference variable is positively skewed so that a transformation, as discussed in Section 4.8, is required. Care must be used with a power transformation because many of the differences are negative. The smallest difference is −16. Therefore, we investigate a

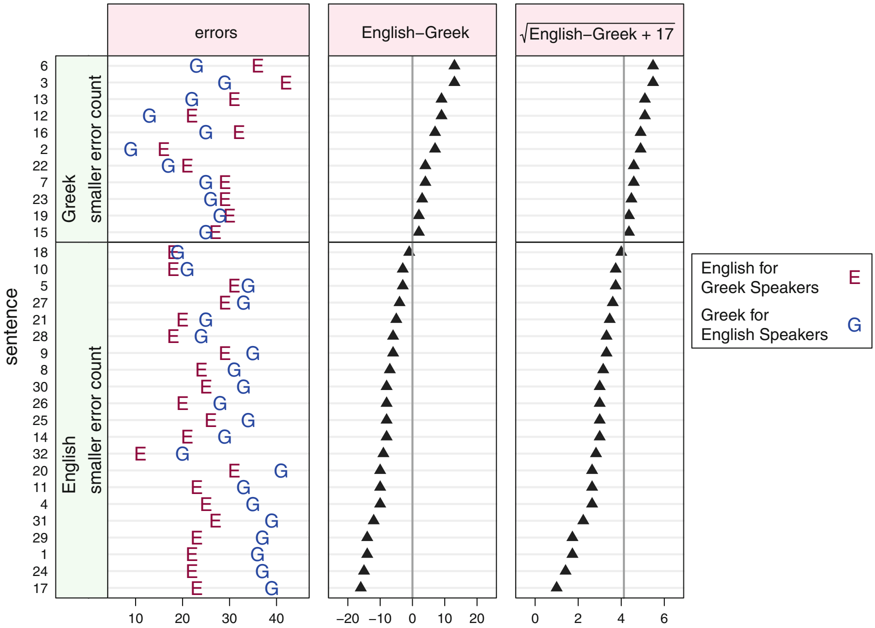


Fig. 5.7 Dotplot of language difficulty scores. The difficulty in learning each of 32 sentences written in English for Greek speakers (marked English) and written in Greek for English speakers (marked Greek) is noted. The panels are defined by placing the sentences in which the English version showed fewer errors on the bottom and the sentences in which the Greek version showed fewer errors on the top. The sentences have been ordered by the difference in the English and Greek error scores. The left panels show the observed error scores. The center panels show the differences, English−Greek, of the error scores. The right panels show the square root transformed differences, $\sqrt{\text{English}-\text{Greek}+17}$. The t-tests in Table 5.8 will be based on the differences and the transformed differences.

square root transformation following the addition of 17 to each value. The second stem-and-leaf in Figure 5.8b illustrates that this transformation succeeds in bringing the data closer to symmetry. Since a difference of zero in the original scale corresponds to a transformed difference of $\sqrt{17} \approx 4.123$, the null hypothesis of equal difficulty corresponds to a comparison of the sample means in the transformed scale to 4.123, not to 0. The observed p-value is .0073, showing a very clear difference in difficulty of learning the two languages. For comparison, the t-test on the untransformed differences show a p-value of only .0346.

```
> stem(teachers$"English-Greek")

  The decimal point is 1 digit(s) to the
  right of the |

  -1 | 65
  -1 | 442000
  -0 | 988887665
  -0 | 4331
   0 | 22344
   0 | 7799
   1 | 33


> t.test(teachers$"English-Greek")


One Sample t-test

data:  teachers$"English-Greek"
t = -2.211, df = 31, p-value = 0.03457
alternative hypothesis:
     true mean is not equal to 0
95 percent confidence interval:
 -6.2484 -0.2516
sample estimates:
mean of x
    -3.25
```

a. Original Scale

```
> stem(sqrt(teachers$"English-Greek" + 17),
+      scale=.5)

  The decimal point is at the |

  1 | 0477
  2 | 26668
  3 | 00002335677
  4 | 04456699
  5 | 1155


> t.test(sqrt(teachers$"English-Greek" + 17),
+        mu=sqrt(17))

One Sample t-test

data:  sqrt(teachers$"English-Greek" + 17)
t = -2.871, df = 31, p-value = 0.00731
alternative hypothesis:
     true mean is not equal to 4.123
95 percent confidence interval:
 3.086 3.947
sample estimates:
mean of x
    3.517
```

b. Transformed Scale

Fig. 5.8 Stem-and-leaf display and *t*-test of sentence difference scores from Figure 5.7 in the original scale and in the offset square-root transformed scale.

5.6 Sample Size Determination

Deciding on an appropriate sample size is a fundamental aspect of experimental design. In this section we provide discussions of the minimum required sample size for some situations of inference about population means:

- A confidence interval on μ with specified width W and confidence coefficient $100(1-\alpha)\%$.
- A test about μ having specified Type I error α, and power $1-\beta$ at a specified distance δ from the null hypothesized parameter.

These are key design objectives for many experiments with modest inferential goals. Specialized software exists for the purpose of determining sample sizes in a vast array of inferential situations. But our discussion here is limited to a few commonly encountered situations for which the formulas are sometimes mentioned in elementary statistics texts.

We assume throughout this discussion that the sample size will be large enough to guarantee that the standardized test statistic is approximately normally distributed. If, as is usual, a sample size calculation does not yield an integer, it is conservative to take n as the next-higher integer. The sample size formulas here are all the result of explicitly solving a certain equation for n. In situations not discussed here, an explicit solution for n may not exist, and the software may produce an iterative solution for n.

5.6.1 Sample Size for Estimation

Since the width of a confidence interval can be expressed as a function of the sample size, the solution of the problem of sample size for a confidence interval is straightforward in the case of a single sample.

For a CI on a single mean, assuming a known population variance σ^2,

$$n = \frac{4\sigma^2\left(\Phi^{-1}(1-\frac{\alpha}{2})\right)^2}{W^2} \tag{5.16}$$

where Φ^{-1} is the inverse cumulative distribution of a standard normal distribution defined in Section J.1.9. Equation 5.16 is found by solving Equation 5.1 for n when we want the width of the confidence interval to be $W = 2\,z_{\frac{\alpha}{2}}\,\frac{\sigma}{\sqrt{n}}$. If σ^2 is unknown, a reasonable guess may be made in its place. (Note that the sample variance is not known prior to selecting the sample.) If we are unable to make a reasonable guess, an ad hoc strategy would be to take a small pilot sample of n_0 items and replace σ in the formula with the standard deviation of the pilot sample. Then if the calculation results in a recommended n greater than n_0, one samples $n - n_0$ additional items.

The required sample size for the Agresti and Caffo CI on a single proportion, Equation (5.3), is

$$n = \frac{\left(\Phi^{-1}(1-\frac{\alpha}{2})\right)^2}{W^2} - 4 \tag{5.17}$$

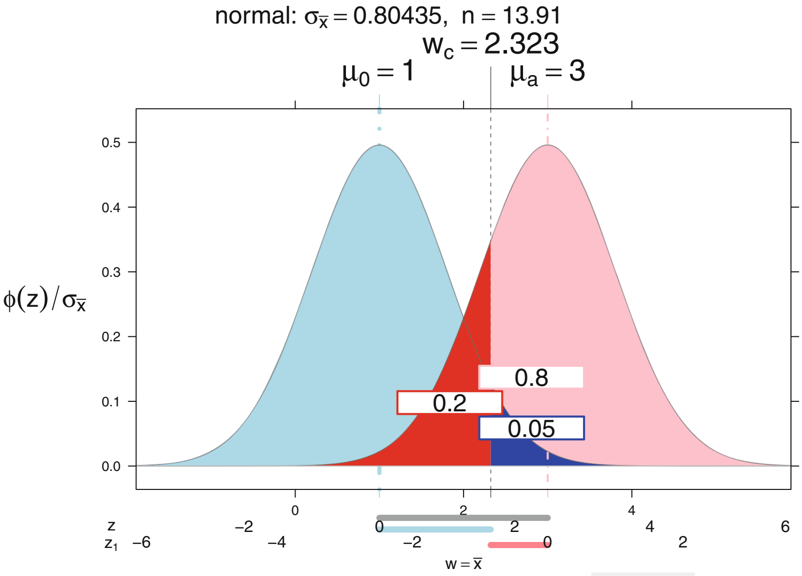
This formula is based on the normal approximation to the binomial distribution. Many statistics texts contain charts for finding the required sample size based on the exact binomial distribution.

5.6.2 Sample Size for Hypothesis Testing

For hypothesis testing we are interested in controlling the specified Type II error probability β when the unknown parameter being tested is a distance δ from the null hypothesized value. For a one-tailed test on the mean of a population with known variance σ^2, use

$$n = \sigma^2 \left(\Phi^{-1}(1-\alpha) + \Phi^{-1}(1-\beta)\right)^2 / \delta^2 \tag{5.18}$$

We illustrate Equation 5.18 in Figure 5.9.



	μ_0	μ_a	$w_{crit.R}$
$\bar{x}$	1	3	2.323
z	0	2.4865	1.6448
z_1	−2.4865	0	−0.84167

	Probability
α	0.05
power	0.8
β	0.2

Fig. 5.9 Sample size and power for the one-sample, one-sided normal test. This figure illustrates Equation 5.18. Both the null and alternative distributions are normal with the same standard error $\sigma = 3$. There are three colored line segments in the horizontal axis region. The top line segment (light gray) on the $w = \bar{x}$ scale is $\delta = \mu_a - \mu_0 = 2$ w-units wide, going from $w = \mu_0 = 1$ to $w = \mu_a = 3$. The middle line segment (light blue) and the bottom line segment (pink) together also are $\delta = 2$ w units wide. The middle segment is $2.323 - 1 = 1.323$ w units wide which is equal to 1.6448 z units. The bottom segment is $3 - 2.323 = 0.677$ w units wide which is equal to .84167 z units We know $\sigma = 3$ and we know that $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. We need to solve for $n = 3^2(1.6488 + .84167)^2/2^2 = 13.96$. We round up to use $n = 14$.

For a two-tailed test, use

$$n = \sigma^2 \left(\Phi^{-1}(1 - \tfrac{\alpha}{2}) + \Phi^{-1}(1 - \beta)\right)^2 / \delta^2 \tag{5.19}$$

For testing the equality of the means of two populations with a common variance, with δ now equal to the mean difference under the alternative hypothesis, use

$$n = 2\,\sigma^2 \left(\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)\right)^2 / \delta^2 \tag{5.20}$$

for the one-tailed test, and

$$n = 2\,\sigma^2 \left(\Phi^{-1}(1 - \tfrac{\alpha}{2}) + \Phi^{-1}(1 - \beta)\right)^2 / \delta^2 \tag{5.21}$$

for the two-tailed test.

When the variance σ^2 is unknown and has to be estimated with s^2 from the sample, the formulas are more difficult because the inverse t cumulative function for the alternative depends on the standard deviation through the noncentrality parameter $(\mu_1 - \mu_0)/(\sigma/\sqrt{n})$. The t formulas might require several iterations as the degrees of freedom, hence the critical values are a function of the sample size.

Tables 5.6, 5.7, and 5.8 show sample size calculations for the one-sample, one-sided test. The example is done three times. Table 5.6 shows the calculation using Equation 5.18 when σ^2 is assumed and the normal equations apply. Table 5.7 uses the R function `power.t.test` which solves the t equations efficiently. Table 5.8 iterates the definitions for the t distribution. Figure 5.10 shows the power plot for the n value in Table 5.7 and one of the n values in Table 5.8.

Lastly, consider attempting to detect a difference between a proportion p_1 and a proportion p_2. The required common sample size for the one-tailed test is

$$n = \frac{\left(p_1\,(1 - p_1) + p_2\,(1 - p_2)\right)\left(\Phi^{-1}(1 - \alpha) + \Phi^{-1}(1 - \beta)\right)^2}{(p_1 - p_2)^2} \tag{5.22}$$

From the preceding pattern, you should be able to deduce the modification for the two-tailed test (see Exercise 5.16).

Table 5.6 We calculate the sample size for a one-sided, one-sample test for the normal distribution with the assumption that the variance is known. In Tables 5.7 and 5.8 we show the same calculations for the *t*-test under the assumption that the variance has been estimated from the sample.

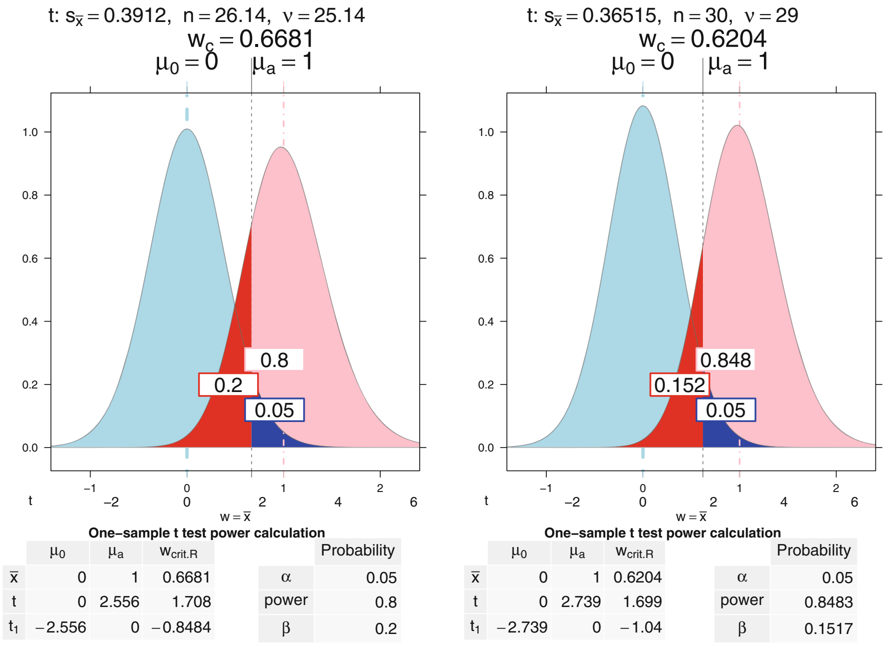
```
> ## one sided
> alpha <- .05

> power <- .80

> beta <- 1-power

> delta <- 1

> sd <- 2

> ## Approximation using formula assuming normal is appropriate
> sd^2*(qnorm(1-alpha) + qnorm(1-beta))^2 / delta^2
[1] 24.73

> ## [1] 24.73
> ## n is slightly smaller with the normal assumption.
>
```

Table 5.7 We calculate the sample size for a one-sided, one-sample *t*-test using the `power.t.test` function. We show the same calculation manually in Table 5.8. We show a static plot of the result in the left column of Figure 5.10. We also show the **shiny** code to specify a dynamic plot.

```
> ## solve using power.t.test
> PTT <-
+ power.t.test(delta=delta, sd=sd, sig.level=alpha, power=power,
+              type="one.sample", alternative="one.sided")

> PTT

     One-sample t test power calculation

              n = 26.14
          delta = 1
             sd = 2
      sig.level = 0.05
          power = 0.8
    alternative = one.sided


> NTplot(PTT, zaxis=TRUE)  ## static plot

> ## NTplot(PTT, zaxis=TRUE, shiny=TRUE)  ## dynamic plot
>
```

Table 5.8 We manually calculate the sample size for a one-sided, one-sample t-test to illustrate the iterative process directly. The `power.t.test` function does this much more efficiently (see Table 5.7). The iterative process starts with an initial sample size n_0 and calculates the critical value $t_{c,0}$ using the central t distribution for that sample size. The second step in the process is to evaluate the power associated with that critical value assuming fixed δ and a series of sample sizes and their associated df and ncp. For the next iterate choose as the new sample size n_1 the sample size whose power is closest to the target power. Calculate a new critical value $t_{c,1}$ and then a new set of powers associated with that critical value. Continue until convergence, meaning the new sample size is the same as the previous one.

```
> ## solve manually with t distribution.  Use ncp for alternative.
> n0 <- 30 ## pick an n0 for starting value

> t.critical <- qt(1-alpha, df=n0-1)

> t.critical
[1] 1.699

> ## [1] 1.699
>
> ## a series of n values
> nn <- 23:30

> names(nn) <- nn

> nn
23 24 25 26 27 28 29 30
23 24 25 26 27 28 29 30

> ## find the power for a series of n values for the specified critical value
> pt(t.critical, df=nn-1, ncp=delta/(sd/sqrt(nn)), lower=FALSE)
    23     24     25     26     27     28     29     30
0.7568 0.7722 0.7868 0.8006 0.8136 0.8258 0.8374 0.8483

> ##     23     24     25     26     27     28     29     30
> ## 0.7568 0.7722 0.7868 0.8006 0.8136 0.8258 0.8374 0.8483
>
> ## recalculate critical value with new n=26
> t.critical <- qt(1-alpha, df=26-1)

> t.critical
[1] 1.708

> ## find the power for a series of n values for the new critical value
> pt(t.critical, df=nn-1, ncp=delta/(sd/sqrt(nn)), lower=FALSE)
    23     24     25     26     27     28     29     30
0.7540 0.7695 0.7842 0.7981 0.8112 0.8235 0.8352 0.8461

> ##     23     24     25     26     27     28     29     30
> ## 0.7540 0.7695 0.7842 0.7981 0.8112 0.8235 0.8352 0.8461
> ## conclude n between 26 and 27
>
```



	μ_0	μ_a	$w_{crit.R}$
$\bar{x}$	0	1	0.6681
t	0	2.556	1.708
t_1	−2.556	0	−0.8484

	Probability
α	0.05
power	0.8
β	0.2

	μ_0	μ_a	$w_{crit.R}$
$\bar{x}$	0	1	0.6204
t	0	2.739	1.699
t_1	−2.739	0	−1.04

	Probability
α	0.05
power	0.8483
β	0.1517

Fig. 5.10 Sample size and power figures for the one-sample, one-sided *t*-test. The left figure shows the sample size n=26.14 calculated in Table 5.7. The right figure shows the starting position with n=30 from Table 5.8. When the sample size n is larger (on the right), the df goes up, the height of the densities (both null and alternative) go up, the densities become thinner, the critical value in the t scale and in the $\bar{x}$ scale goes down. The alternative distribution is noncentral t, has a different maximum height, and is not symmetric.

5.7 Goodness of Fit

Goodness-of-fit tests are used to assess whether a dataset is consistent with having been sampled from a designated hypothesized distribution. In this section we discuss two general goodness-of-fit tests, the Chi-Square Goodness-of-Fit Test and the Kolmogorov–Smirnov Goodness-of-Fit Test. For testing goodness of fit to specific distributions, there may be better (more powerful) specialized tests than these. For example, the Shapiro–Wilk test of normality (`shapiro.test`) is more powerful than either general test.

Since many statistics procedures assume an underlying normal distribution, a test of goodness of fit to normal, either before or after transformation, is frequently performed. Occasionally, analysts need to check for fit to other distributions. For example, it is often the case that the distribution of a test statistics is known asymptotically (i.e., if the sample is "large"), but not if the sample is of modest size. It is therefore of interest to investigate how large a sample is needed for the asymptotic distribution to be an adequate approximation. This requires a series of goodness-of-fit tests to the asymptotic distribution. In Chapter 15, we will learn in our discussion of the analysis of contingency table data that the distribution of $\chi^2 = \sum \frac{(O-E)^2}{E}$ is approximately chi-square provided that no cell sizes are too small. A determination of the ground rule for "too small" required tests of goodness of fit to chi-square distributions with appropriate degrees of freedom.

This class of tests assesses whether a sample may be assumed to be taken from a null hypothesized distribution.

5.7.1 Chi-Square Goodness-of-Fit Test

The chi-square distribution may be used to conduct goodness-of-fit tests, i.e., ones of the form

$$\begin{array}{ll} & H_0\text{: the data are from a [specified population]} \\ \text{vs} & \\ & H_1\text{: the data are from some other population} \end{array}$$

For certain specific populations, including normal ones, other specialized tests are more powerful.

The test begins by partitioning the population into k classes or categories. For a discrete population the categories are the possible values; for a continuous population the choice of a decomposition is rather arbitrary, and the ultimate conclusion may well depend on the selected size of k and the selected partition.

The test statistic is the same as that used for contingency tables. For each category, calculate from the probability distribution the theoretical or expected frequency E. If over all k categories, there is a substantial discrepancy between the k observed frequencies O and the k E's, then H_0 is rejected. The measure of discrepancy is the test statistic $\chi^2 = \sum \frac{(O-E)^2}{E}$. A "large" value of χ^2 is evidence against H_0. If the total sample size, $n = \sum O = \sum E$, is sufficiently "large", χ^2 is approximately chi-square distributed and the p-value is approximately the chi-square tail probability associated with χ^2 with $k - 1$ degrees of freedom.

For adequacy of the chi-square approximation it is suggested that all expected frequencies be at least 5. If this is not the case, the analyst may consider combining adjacent categories after which this condition is met. Then k represents the number of categories following such combining.

Sometimes, the statement of the null hypothesis is so vague that calculation of expected frequencies requires that some parameters be estimated from the data. In such instances, the df is further reduced by the number of such parameters estimated. This possibility is illustrated in Example 5.7.3.

5.7.2 Example—Test of Goodness-of-Fit to a Discrete Uniform Distribution

A six-sided die (singular of the word *dice*) is rolled 30 times with the following outcomes: 1, 3 times; 2, 7 times; 3, 5 times; 4, 8 times; 5, 1 time; and 6, 6 times. Test whether the die is fair.

A fair die is one that has a discrete uniform distribution on 1, 2, 3, 4, 5, 6. Each of these six possibilities has $\frac{1}{6}$ chance of occurring, and all six E's are $30(\frac{1}{6}) = 5$. Then

$$\chi^2 = \frac{(3-5)^2}{5} + \ldots + \frac{(6-5)^2}{5} = 6.8$$

and the p-value from χ^2_5 is 0.236. Hence these 30 observations do not provide evidence to refute the fairness of the die. We show the calculations in Table 5.9 and the plot of the test in Figure 5.11.

Table 5.9 Test of Goodness-of-Fit to a Discrete Uniform Distribution. The test is plotted in Figure 5.11.

```
> dice <- sample(rep(1:6, c(3,7,5,8,1,6)))

> dice
 [1] 4 6 4 2 3 2 4 4 6 3 6 4 3 2 3 4 6 2 6 2 1 4 3 5 1 2 1 6
[29] 4 2

> table(dice)
dice
1 2 3 4 5 6
3 7 5 8 1 6

> chisq.test(table(dice))

Chi-squared test for given probabilities

data:  table(dice)
X-squared = 6.8, df = 5, p-value = 0.2359
```

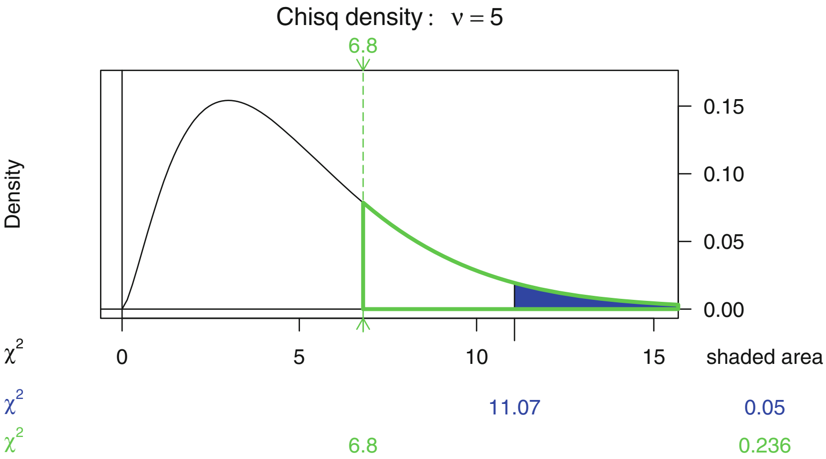


Fig. 5.11 Plot of the hypothesis test of Table 5.9. The observed value $\chi^2 = 6.8$ shows $p = 0.236$ and is in the middle of the do-not-reject region,

Table 5.10 Observed and expected frequencies for the goodness-of-fit example in Section 5.7.3.

Y	O	E	$\frac{(O-E)^2}{E}$
0	13	6.221	7.388
1	18	20.736	0.361
2	20	27.648	2.116
3	18	18.432	0.010
4	6	6.144	0.003
5	5	0.819	21.337
			31.215

5.7.3 Example—Test of Goodness-of-Fit to a Binomial Distribution

In a certain community, there were 80 families containing exactly five children. It was noticed that there was an excess of boys among these. It was desired to test whether Y = "number of girls in family" is a binomial r.v. with $n = 5$ and $p = .4$. The expected frequencies calculated from this binomial distribution are shown in Table 5.10 along with the observed frequencies and the calculated χ^2_5 statistic. Then the p-value is, 8.510^{-6}, calculated as the tail probability at 31.215 for a chi-square distribution with 5 df. We conclude that the sample data contain more dispersion than does binomial(5, .4). The excess dispersion is visible in the left panel of Figure 5.12.

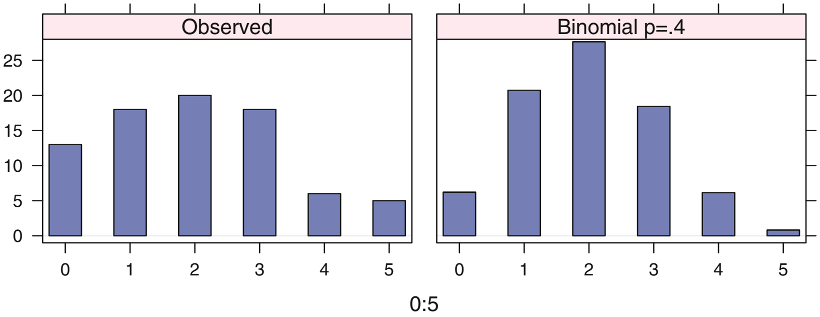


Fig. 5.12 Plot of family size data from Table 5.11. The Observed data is more spread out than the Expected (binomial) data. The sample variance for the Observed is `var(rep(0:5, times=Observed)) == 1.987` and the sample variance for the Expected is `var(rep(0:5, times=Expected)) == 1.131`.

In this example, the value of the binomial proportion parameter, p, was specified. If instead it had to be estimated, the df would decrease from 5 to 4. We illustrate the calculation of both tests in R in Table 5.11.

5.8 Normal Probability Plots and Quantile Plots

Quantile plots (Q-Q plots) are visual diagnostics used to assess whether (a) a dataset may reasonably be treated as if it were a sample from a designated probability distribution, or (b) whether two datasets show evidence of coming from a common unspecified distribution.

The normal probability plot, an important special case of the more general quantile plot, is used to assess whether data are consistent with a normal distribution. The normal probability plot is a standard diagnostic plot in regression analysis (Chapters 8–11) used to check the assumption of normally distributed residuals. This condition is required for the validity of many of the usual inferences in a regression analysis. If the normality assumption appears to be violated, it is often possible to retain a simple analysis by transforming the data scale, for example by a power transformation, and then reanalyzing and replotting to see if the residuals from the transformed data are close to normal. The choice of transformation may be guided by the interpretation of the normal probability plot.

In R, a normal probability plot is produced with the `qqmath` function (in **lattice**) or the `qqnorm` function (in base graphics). function. Normal probability plots are included in the default plots for the results of linear model analyses.

A quantile plot to assess consistency of observed data y_i with a designated distribution is easily constructed. We sort the observed data to get $y_{[i]}$, find the quantiles of the distribution by looking up the fractions $(i - \frac{1}{2})/n$ in the inverse cumulative

Table 5.11 Calculation of p-value for chi-square test with known p and with estimated $\hat{p}$. The Observed and Expected frequencies are plotted in Figure 5.12.

```
> Observed <- c(13, 18, 20, 18, 6, 5)

> names(Observed) <- 0:5

> ## binomial proportion p=.4 is specified
> Expected <- dbinom(0:5, size=5, p=.4)*80

> names(Expected) <- 0:5

> chisq.test(Observed, p=Expected, rescale.p=TRUE)

Chi-squared test for given probabilities

data:  Observed
X-squared = 31.21, df = 5, p-value = 8.496e-06

Warning message:
In chisq.test(Observed, p = Expected, rescale.p = TRUE) :
  Chi-squared approximation may be incorrect

> ## binomial proportion p is calculated from the observations
> p <- sum(Observed * (0:5)/5)/sum(Observed)

> p
[1] 0.4025

> Expected <- dbinom(0:5, size=5, p=p)*80

> names(Expected) <- 0:5

> WrongDF <- chisq.test(Observed, p=Expected, rescale.p=TRUE)
Warning message:
In chisq.test(Observed, p = Expected, rescale.p = TRUE) :
  Chi-squared approximation may be incorrect

> WrongDF

Chi-squared test for given probabilities

data:  Observed
X-squared = 30.72, df = 5, p-value = 1.066e-05


> c(WrongDF$statistic, WrongDF$parameter)
X-squared        df
    30.72      5.00

> ## correct df and p-value
> pchisq(WrongDF$statistic, df=WrongDF$parameter - 1, lower=FALSE)
X-squared
3.498e-06
```

distribution function to get $q_i = F^{-1}((i - \frac{1}{2})/n)$, and then plotting the sorted data $y_{[i]}$ against the quantiles q_i. Consistency is suggested if the points tend to fall along a straight line. A pattern of a departure from a straight-line quantile plot usually suggests the nature of the departure from the assumed distribution. The R one-sample quantile plots (both the **lattice** `qqmath` and the base graphics `qqnorm`) default to the usual convention of plotting the data against the theoretical values. Other software and a number of references reverse the axes. Readers of presentations containing quantile plots should be alert to which convention is used, and writers must be sure to label the axes to indicate the convention, because the choice matters considerably for interpretation of departures from compatibility.

A general Q-Q (or quantile-quantile) plot is invoked in R with the base graphics command `qqplot(x, y, plot=TRUE)`, whereby the quantiles of two samples, `x` and `y`, are compared. As with a normal probability case, the straightness of the Q-Q plot indicates the degree of agreement of the distributions of `x` and `y`, and departure from a well-fitting straight line on an end of the plot indicates the presence of outlier(s). Quoting from the S-Plus online help for `qqplot`:

> A Q-Q plot with a “U” shape means that one distribution is skewed relative to the other. An “S” shape implies that one distribution has longer tails than the other. In the default configuration (data on the *y*-axis) a plot from `qqnorm` that is bent down on the left and bent up on the right means that the data have longer tails than the Gaussian [normal].

For a normal probability plot with default configuration, a plot that is bent up on the left and bent down on the right indicates that the data have shorter tails than the normal. A curved plot that opens upward suggests positive skewness and curvature opening downward suggests negative skewness.

It is possible to construct a Q-Q plot comparing a sample with any designated distribution, not just the normal distribution. In R and S-Plus this is accomplished with the function `ppoints(y)`, which returns a vector of n=`length(y)` fractions uniformly spaced between 0 and 1 which will be used as input to the quantile (inverse cumulative distribution) function. For example, all three R statements

```
plot(sort(y) ~ qlnorm(ppoints(y)))
qqplot(qlnorm(ppoints(y)), y)
qqmath(y, distribution=qlnorm)
```

produce a lognormal Q-Q plot of the data in `y`. See Appendix J for the lognormal distribution.

If it is unclear from a normal probability plot whether the data are in fact normal, the issue may be further addressed by a specialized goodness-of-fit test to the normal distribution, the Shapiro–Wilk test. This test works by comparing

$S(y)$ the empirical distribution function of the data, the fraction of the data that is less than or equal to y

with

$\Phi\big((y - \bar{y})/s\big)$ the probability that a normal r.v. Y (with mean $\bar{y}$ and s.d. s) is less than or equal to y

Over the observed sample, $S(y)$ and $\Phi\big((y - \bar{y})/s\big)$ should be highly correlated if the data are normal, but not otherwise. The Shapiro–Wilk statistic W is closely related to the square of this correlation. If the normal probability plot is nearly a straight line, W will be close to 1. A small value of W is evidence of nonnormality. The Shapiro–Wilk test is available in R with the `shapiro.test` function. For this specific purpose the Shapiro–Wilk test is more powerful than a general goodness-of-fit test such as the Kolmogorov–Smirnov procedure discussed in Section 5.9.

5.8.1 Normal Probability Plots

Figure 5.13 contrasts the appearance of normal probability plots for the normal distribution and various departures from normality. Typically, the plot has these appearances:

- An "S" shape for distributions with thinner tails than the normal.
- An inverted "S" shape for distribution with heavier tails than the normal.

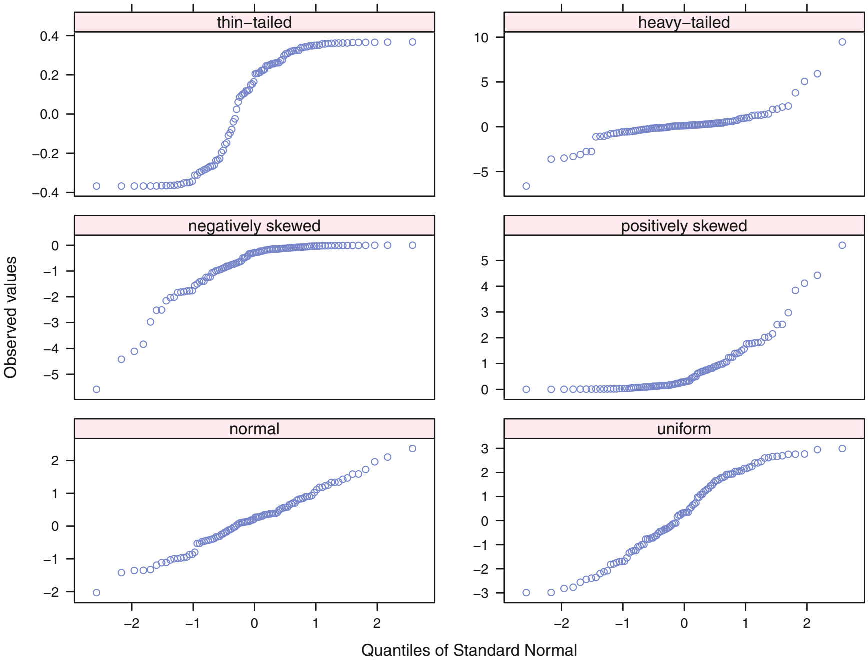


Fig. 5.13 Normal probability plots of data randomly selected from normal and other distributions. The density plots of these variables are in Figure 5.14.

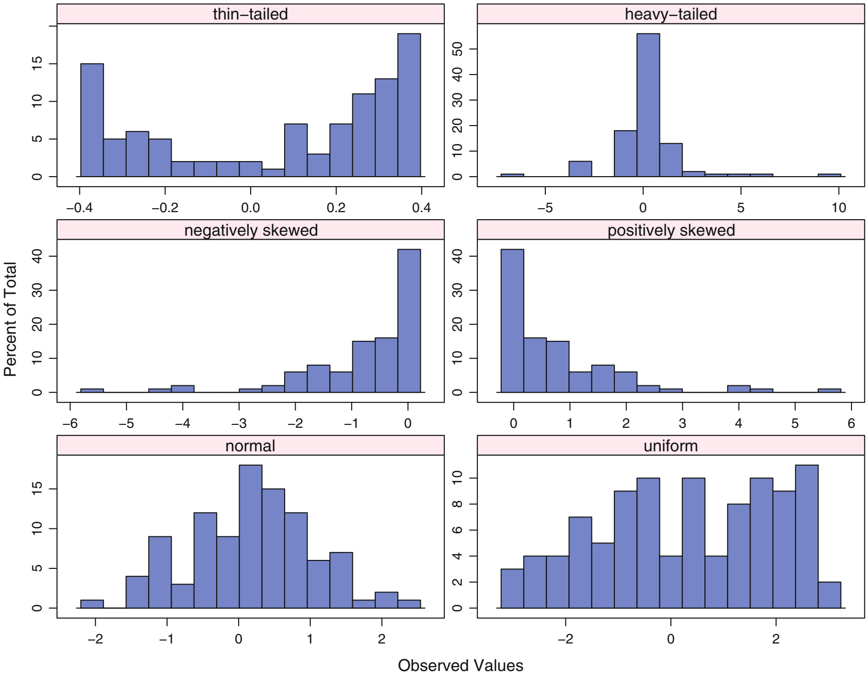


Fig. 5.14 Density plots of the data randomly selected from normal and other distributions. This is the same data whose normal probability plots are shown in Figure 5.13

.

- A “J” shape for positively skewed distributions.
- An inverted “J” shape for negatively skewed distributions.
- Isolated points at the extremes of a plot for distributions having outliers.

5.8.2 Example—Comparing t-Distributions

We compare a random sample of 100 from a t distribution with 5 df to quantiles from a longer-tailed t_3 distribution and from shorter-tailed t_7 and normal distributions. The four superimposed Q-Q plots and a reference 45° line are shown in Figure 5.15.

Note that the picture we get will vary according to the particular random sample selected. In this example the plot against the quantiles of t_5, the same distribution from which the sample was drawn, is close to the 45° line. The longer-tailed t_3 quantiles show a reflected “S” shape. The shorter-tailed t_7 and normal distributions show an “S” shape.

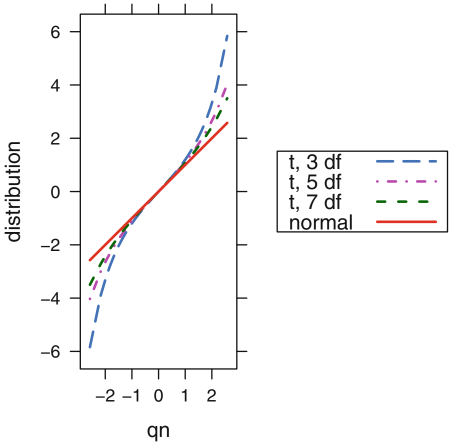


Fig. 5.15 Q-Q plots for the *t* distribution with several different degrees of freedom. The normal is the same as the *t* with infinite degrees of freedom. The scaling here is isometric, the same number of inches per unit on the *x* and *y* scales. The aspect ratio is chosen to place the normal QQ plot exactly on the 45° line.

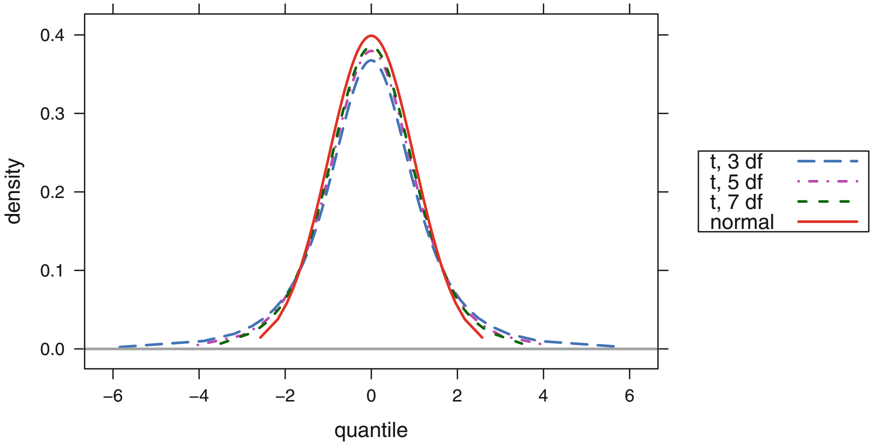


Fig. 5.16 *t* densities. The normal (*t* with infinite degrees of freedom) is the tallest and thinnest. As the degrees of freedom decrease, the center gets less high and the tails have noticeable weight farther away from the center.

Long and short tails refer to the appearance of plots of the density functions. Note that the normal has almost no probability (area) outside of ±2.5. The *t* distributions have more and more probability in the tails of the distribution (larger $|q|$) as the degrees of freedom decrease. The superimposed densities are displayed in Figure 5.16.

5.9 Kolmogorov–Smirnov Goodness-of-Fit Tests

The Kolmogorov–Smirnov goodness-of-fit tests are used to formally assess hypothesis statements concerning probability distributions. The K–S one-sample test tests whether a random sample comes from a hypothesized null distribution. The K–S two-sample test tests whether two independent random samples are coming from the same but unspecified probability distribution. The alternative hypothesis can be either one-sided or two-sided.

The K–S one-sample test involves comparing the maximum discrepancy between the empirical cumulative distribution of the data, defined as $S(y)$ = fraction of the data that is less than or equal to y, and the cumulative distribution function of the hypothesized population being sampled. The K–S two-sample test statistic is the maximum discrepancy between the empirical distribution functions of the two samples.

The Shapiro–Wilk test of normality is more powerful than K–S for assessing normality. The Shapiro–Wilk test statistic W more fully uses the sample than does K–S. If we have data that are close to normal except for one very unusual point, K–S will be more sensitive to this point than W. In general, the K–S procedure focuses on the most extreme departure from the hypothesized distribution while Shapiro–Wilk's assessment based on Q-Q focuses on the average departure.

The K–S tests are performed in R with the function `ks.test`. See the R help file for `ks.test` for details. This function can handle both one- and two-sample tests. For the one-sample test, a long list of probability distributions can be specified as the null hypothesis. The parameters of the null distribution can be estimated from the data or left unspecified. With some exceptions, the alternative hypothesis can be `"greater"` or `"less"` as well as `"two-sided"`. The interpretation of a one-sided hypothesis is that one c.d.f. is uniformly and appreciably shifted to one side of the other c.d.f.

5.9.1 Example—Kolmogorov–Smirnov Goodness-of-Fit Test

We illustrate the One-Sample Kolmogorov–Smirnov Test in Table 5.12 and Figure 5.17. We illustrate the Two-Sample Kolmogorov–Smirnov Test in Table 5.13 and Figure 5.18.

We selected two random samples of 300 items, the first from a t distribution with 5 df, and the second from a standard normal distribution. Table 5.12 shows the K–S tests and Figure 5.17 the plot of the tests.

Table 5.12 Kolmogorov–Smirnov One-Sample Test. The first test corresponds to the left panels of Figure 5.17. We see a p-value of 0.2982 and do not reject the null. The second test corresponds to the right panels of Figure 5.17. We see a p-value of 0.003808 and reject the null.

```
> rt5 <- rt(300, df=5)

> rnn <- rnorm(300)

> ks.test(rt5, function(x)pt(x, df=2))

One-sample Kolmogorov-Smirnov test

data:  rt5
D = 0.0563, p-value = 0.2982
alternative hypothesis: two-sided


> ks.test(rnn, function(x)pt(x, df=2))

One-sample Kolmogorov-Smirnov test

data:  rnn
D = 0.1022, p-value = 0.003808
alternative hypothesis: two-sided
```

Table 5.13 Kolmogorov–Smirnov Two-Sample Test. The test corresponds to Figure 5.18. We see a p-value of 0.09956 and we do not reject the null.

```
> ks.test(rt5, rnn)

Two-sample Kolmogorov-Smirnov test

data:  rt5 and rnn
D = 0.1, p-value = 0.09956
alternative hypothesis: two-sided
```

In the table we test to see if these sample datasets are consistent with a t distribution with 2 df. The 5-df dataset is consistent with the 2-df null distribution. The normal dataset is not. The top panel in both columns of Figure 5.17 shows the distribution for the hypothesized t_2 distribution, and the vertical deviations of

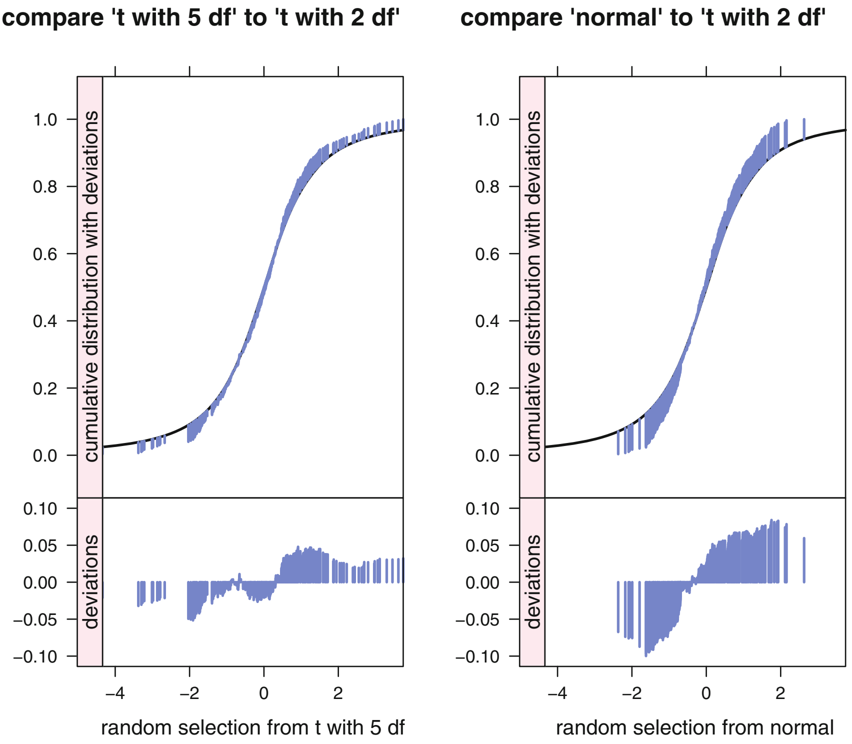


Fig. 5.17 Kolmogorov–Smirnov plots. Kolmogorov–Smirnov One-Sample Test. On the left we compare a random selection from the *t* distribution with 5 df to a null hypothesis distribution of *t* with 2 df. The `ks.test` in Table 5.12 shows a *p*-value of 0.2982 and does not reject the null. On the right we compare a random selection from the standard normal distribution to a null hypothesis distribution of *t* with 2 df. The `ks.test` in Table 5.12 shows a *p*-value of 0.003808 and rejects the null. The solid line in the top panels is the CDF for null distribution, in this example the *t* with 2 df. The deviation lines connect the observed *y*-values from the dataset under test to the hypothesized *y*-values from the null distribution. The deviation lines are magnified and centered in the bottom panels. The largest |vertical deviation| is the value of the K–S statistic in Table 5.12.

the data from the hypothesized distribution. The largest absolute value of these vertical deviations is the Kolmogorov–Smirnov statistic. The lower panel shows the deviations.

In Table 5.13 and Figure 5.18 we directly compare two different samples to see if the Two-Sample `ks.test` can distinguish between them. In this example the null hypothesis is retained. The plot shows both empirical distribution functions.

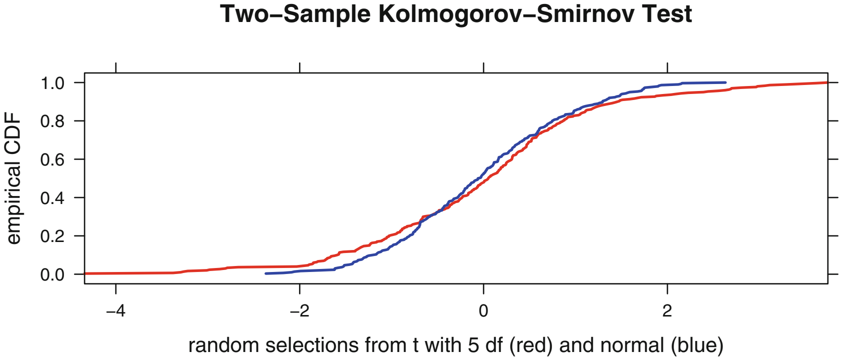


Fig. 5.18 Kolmogorov–Smirnov two-sample plot. We plotted the two empirical CDF on the same axes. The largest absolute vertical deviation is the value of the K–S statistic. Interpolation to calculate the vertical deviations is messier in the two-sample case, therefore we didn't do it for the figure. The `ks.test` function in Table 5.13 does do the interpolation.

5.10 Maximum Likelihood

Maximum likelihood is a general method of constructing "good" point estimators. *Likelihood ratio* is a general method of constructing tests with favorable properties. We briefly consider both of these ideas.

5.10.1 Maximum Likelihood Estimation

We start from the joint distribution of the sample statistics. The maximum likelihood estimator (MLE) is the value of the parameter that maximizes this expression of the joint distribution, which is called the likelihood function L. In practice it is usually easier to solve the equivalent problem of maximizing $\ln(L)$, equivalent since $\ln(\cdot)$ is an increasing function.

As a simple example, we derive the MLE of the mean μ of a normal population with known standard deviation σ, based on a random sample of n from this population.

The likelihood function $L(\mu)$ is a function of the parameter μ. $L(\mu)$ is constructed as the product of the individual density functions for the observed data values y_i.

$$L(\mu) = \prod \phi\left(\frac{y_i - \mu}{\sqrt{2}\,\sigma}\right) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{\sum(y_i - \mu)^2}{2\,\sigma^2}\right) \tag{5.23}$$

Apart from an additive constant that does not depend on μ, we find

$$\ln(L) = -\frac{\sum(y_i - \mu)^2}{2\sigma^2}$$

The value of μ that maximizes this expression is the value of μ that minimizes

$$\sum(y_i - \mu)^2 = \sum(y_i - \bar{y})^2 + n(\bar{y} - \mu)^2$$

The answer, $\hat{\mu} = \bar{y}$, is both the "least-squares" and maximum likelihood estimator of μ. (The least-squares and maximum likelihood estimators do not necessarily coincide for other estimands than μ.)

5.10.2 Likelihood Ratio Tests

Let $y_1, y_2, \ldots, y_n$ denote a random sample of some population and let $L = L(y_1, y_2, ..., y_n)$ denote the *likelihood* of this sample, i.e., the joint probability distribution of the sample values. Let H_0 be a null hypothesis about the parameter(s) of this population. A *likelihood ratio (LR) test* of H_0 uses the likelihood ratio

$$\lambda = \frac{\text{maximum of } L \text{ over only those parameter values for which } H_0 \text{ is true}}{\text{maximum of } L \text{ over all possible parameter values}} \tag{5.24}$$

or some random variable that is a strictly increasing or strictly decreasing function of only λ. H_0 is rejected if λ is sufficiently small, where "sufficiently small" depends on α.

While likelihood ratio tests do not, in general, have optimal properties, experience has taught that they frequently are competitive. One reason for their popularity is that they have a known asymptotic (i.e., large sample size n) distribution: $-2\ \ln(\lambda)$ is approximately a χ^2 r.v. with d.f. equal to the number of parameters constrained by H_0. This fact can be used to construct a large sample test.

For example, to test $H_0{:}\,\mu = 0$ vs $H_1{:}\,\mu \neq 0$, where μ is the mean of a normal population with unknown variance, it is not difficult to show that the likelihood ratio test procedure gives $\lambda = \frac{1}{(1+t^2)^{n/2}}$, where $|t|$ is the usual absolute t statistic used for this purpose. Here $|t|$ arises as the appropriate test statistic because it is a strictly decreasing function of λ.

5.11 Exercises

5.1. Suppose that hourly wages in the petroleum industry in Texas are normally distributed with a mean of \$17.60 and a standard deviation of \$1.30. A large company in this industry randomly sampled 50 of its workers, determining that their hourly wage was \$17.30. Stating your assumptions, can we conclude that this company's average hourly wage is below that of the entire industry?

5.2. The mean age of accounts payable has been 22 days. During the past several months, the firm has tried a new method to reduce this mean age. A simple random sample of 200 accounts payable last month had mean age 20.2 days and standard deviation 7.2 days. Use a confidence interval to determine if the new method has made a difference.

5.3. The Security and Exchange Commission (SEC) requires companies to file annual reports concerning their financial status. Firms cannot audit every account receivable, so the SEC allows firms to estimate the true mean. They require that a reported mean must be within \$5 of the true mean with 98% confidence. In a small sample of 20 from firm Y, the sample standard deviation was \$40. What must the total sample size be so that the audit meets the standard of the SEC?

5.4. The Kansas City division of a company produced 982 units last week. Of these, 135 were defective. During this same time period, the Detroit division produced 104 defectives out of 1,088 units. Test whether the two divisions differed significantly in their tendency to produce defectives.

5.5. A human resources manager is interested in the proportion of firms in the United States having on-site day-care facilities. What is the required sample size to be 90% certain that the sample proportion will be within 5% of the unknown population proportion?

5.6. A health insurance company now offers a discount on group policies to companies having a sufficiently high percentage of nonsmoking employees. Suppose a company with several thousand workers randomly samples 200 workers and finds that 186 are nonsmokers. Find a 95% confidence interval for the proportion of this company's employees who do not smoke.

5.7. Out of 750 people chosen at random, 150 were unable to identify your product. Find a 90% confidence interval for the proportion of all people in the population who will be unable to identify your product.

5.8. A national poll, based on interviews with a random sample of 1,000 voters, gave one candidate 56% of the vote. Set up a 98% confidence interval for the proportion of voters supporting this candidate in the population. You need not complete the calculations.

5.9. Two hundred people were randomly selected from the adult population of each of two cities. Fifty percent of the city #1 sample and 40% of the city #2 sample were opposed to legalization of marijuana. Test the two-sided hypothesis that the two cities have equal proportions of citizens who favor legalization of marijuana. (Calculate and interpret the p-value.)

5.10. A random sample of 200 people revealed that 80 oppose a certain bond issue. Find a 90% confidence interval for the proportion in the population who oppose this bond issue. Work the arithmetic down to a final numerical answer.

5.11. The confidence interval answer to the previous question is rather wide. How large a sample would have been required to reduce the confidence interval error margin to 0.02?

5.12. Random samples of 400 voters were selected in both New Jersey and Pennsylvania. There were 210 New Jersey respondents and 190 Pennsylvania respondents who stated that they were leaning toward supporting the Democratic nominee for President. Test the claim (alternative hypothesis) that the proportion of all New Jersey voters who lean Democratic exceeds the proportion of all Pennsylvania voters who lean Democratic.

a. Set up H_0 and H_1.

b. Calculate $\hat{p}_1$, $\hat{p}_2$, and $\hat{p}$.

c. Calculate z_{calc}.

d. Approximate the p-value.

e. State your conclusion concerning the claim.

5.13. The relative rotation angle between the L2 and L3 lumbar vertebrae is defined as the acute angle between posterior tangents drawn to each vertebra on a spinal X-ray. See Figure 7.20 for an illustration with different vertebrae. When this angle is too large the patient experiences discomfort or pain. Chiropractic treatment of this condition involves decreasing this angle by applying (nonsurgical) manipulation or pressure. Harrison et al. (2002) propose a particular such treatment. They measured the angle on both pre- and post-treatment X-rays from a random sample of 48 patients. The data are available as `data(har1)`.

a. Test whether the mean post-treatment angle is less than the mean angle prior to treatment.

b. Construct a quantile plot to assess whether the post-treatment sample is compatible with a t distribution with 5 degrees of freedom.

5.14. The Harrison et al. (2002) study also measured the weights in pounds of the sample of 48 treated patients and a random sample of 30 untreated volunteer controls.

a. Use the data available as `data(har2)` to compare the mean weights of the treatment and control populations.

b. Use these data to compare the standard deviation of weights of the treatment and control populations.

c. Construct and interpret a normal probability plot for the weights of the treated patients.

5.15. The *Poisson* probability distribution is defined on the set of nonnegative integers. The Poisson is often used to model the number of occurrences of some event per unit time or unit space. Examples are the number of phone calls reaching a switchboard in a given minute (with the implication that the number of operators scheduled to answer the phones will be determined from the model) or the number of amoeba counted in a 1 ml. specimen of pond water. The probability that a Poisson r.v. Y has a particular (nonnegative integer) value y is given by

$$P(Y = y) = \frac{e^{-\mu}\mu^{y}}{y!}, \quad y = 0, 1, 2, \ldots$$

(While the value of y may be arbitrarily large, the probability of obtaining a very large y is infinitesimally small.) The parameter μ is the mean number of occurrences per unit. The mean μ of the Poisson distribution is either known in advance or must be estimated from the data. Poisson probabilities may be calculated with R as noted in Section J.3.6.

You are asked to perform a chi-square goodness-of-fit test of the Poisson distribution to the following data, which concern the number of specimens per microscope field in a sample of lake water.

y:	0	1	2	3	4	5	6	7
O:	21	30	54	26	11	3	3	2

The observed value O_y is the number of fields in which exactly y specimens were observed. In this example, $\sum O_y = 150$ fields were examined and, for example, exactly $O_2 = 54$ of the fields showed $y = 2$ specimens. The Poisson parameter μ is unknown and should be estimated as a weighted average of the possible values y, i.e.,

$$\hat{\mu} = \frac{\sum\limits_{y=0}^{7} y\,O_y}{\sum\limits_{y=0}^{7} O_y}$$

5.16. Extend the one-tailed sample size formula for comparing two proportions, Equation (5.22), to the two-tailed case.