CHAPTER 16

Hypothesis testing with the bootstrap

16.1 Introduction

In Chapter **??** we describe the permutation test, a useful tool for hypothesis testing. At the end of that chapter we relate hypothesis tests to confidence intervals, and in particular showed how a bootstrap confidence interval could be used to provide a significance level for a hypothesis test. In this chapter we describe bootstrap methods that are designed directly for hypothesis testing. We will see that the bootstrap tests give similar results to permutation tests when both are available. The bootstrap tests are more widely applicable though less accurate.

16.2 The two-sample problem

We begin with the two-sample problem as described in the last chapter. We have samples **z** and **y** from possibly different probability distributions F and G, and we wish to test the null hypothesis $H_0 :\ F = G$. A bootstrap hypothesis test, like a permutation test, is based on a test statistic. In the previous chapter this was denoted by $\hat{\theta}$. To emphasize that a test statistic need not be an estimate of a parameter, we denote it here by $t(\mathbf{x})$. In the mouse data example, $t(\mathbf{x}) = \bar{z} - \bar{y}$, the difference of means with observed value 30.63. We seek an achieved significance level

$$\text{ASL} = \text{Prob}_{H_0}\{t(\mathbf{x}^*) \geq t(\mathbf{x})\} \tag{16.1}$$

as in (**??**). The quantity $t(\mathbf{x})$ is fixed at its observed value and the random variable $\mathbf{x}^*$ has a distribution specified by the null hypothesis H_0. Call this distribution F_0. Now the question is, what is F_0? In the permutation test of the previous chapter, we fixed the order

Algorithm 16.1

Computation of the bootstrap test statistic for testing $F = G$

1. Draw B samples of size $n + m$ with replacement from $\mathbf{x}$. Call the first n observations $\mathbf{z}^{*b}$ and the remaining m observations $\mathbf{y}^{*b}$, for $b = 1, 2, \ldots B$.
2. Evaluate $t(\cdot)$ on each sample,
$$t(\mathbf{x}^{*b}) = \bar{\mathbf{z}}^{*b} - \bar{\mathbf{y}}^{*b}, \quad b = 1, 2, \cdots B. \tag{16.2}$$
3. Approximate ASL_{boot} by
$$\widehat{\text{ASL}}_{\text{boot}} = \#\{t(\mathbf{x}^{*b}) \geq t_{obs}\}/B, \tag{16.3}$$
where $t_{obs} = t(\mathbf{x})$ the observed value of the statistic.

statistics $\mathbf{v}$ and defined F_0 to be the distribution of possible orderings of the ranks $\mathbf{g}$. Bootstrap hypothesis testing, on the other hand, uses a "plug-in" style estimate for F_0. Denote the combined sample by $\mathbf{x}$ and let its empirical distribution be $\hat{F}_0$, putting probability $1/(n + m)$ on each member of $\mathbf{x}$. Under H_0, $\hat{F}_0$ provides a nonparametric estimate of the common population that gave rise to both $\mathbf{z}$ and $\mathbf{y}$. Algorithm 16.1 shows how $\widehat{\text{ASL}}_{\text{boot}}$ is computed.

Notice that the only difference between this algorithm and the permutation algorithm in equations (**??**) and (**??**) is that samples are drawn with replacement rather than without replacement. It is not surprising that it gives very similar results (left panel of Figure 16.1). One thousand bootstrap samples were generated, and 120 had $t(\mathbf{x}^*) > 30.63$. The value of $\widehat{\text{ASL}}_{\text{boot}}$ is $120/1000 = .120$ as compared to .152 from the permutation test.

More accurate testing can be obtained through the use of a studentized statistic. In the above test, instead of $t(\mathbf{x}) = \bar{z} - \bar{y}$ we could use

$$t(\mathbf{x}) = \frac{\bar{z} - \bar{y}}{\bar{\sigma}\sqrt{1/n + 1/m}}, \tag{16.4}$$

where $\bar{\sigma} = \{[\sum_{i=1}^{n}(z_i - \bar{z})^2 + \sum_{j=1}^{m}(y_j - \bar{y})^2]/[n + m - 2]\}^{1/2}$. This is the two-sample t statistic described in Chapter **??**. The observed value of $t(\mathbf{x})$ was 1.12. Repeating the above bootstrap algorithm,

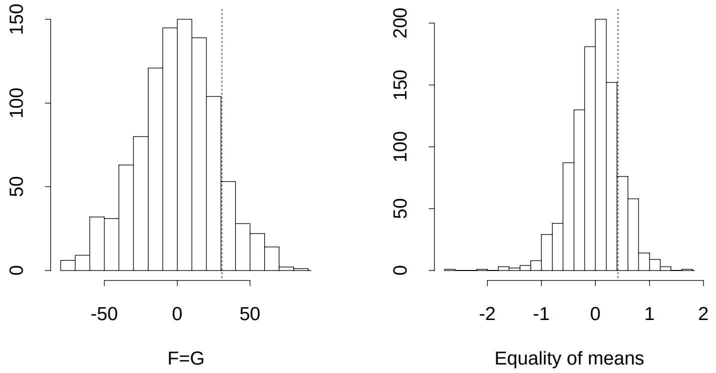

Figure 16.1. *Histograms of bootstrap replications for the mouse data example. The left panel is a histogram of bootstrap replications of $\bar{z}-\bar{y}$ for the test of $H_0 : F = G$, while the right panel is a histogram of bootstrap replications of the studentized statistic (16.5) for the test of equality of means. The dotted lines are drawn at the observed values (30.63 on the left, .416 on the right). In the left panel, $\widehat{ASL}_{\rm boot}$ (the bootstrap estimate of the achieved significance level) equals .120, the proportion of values greater than 30.63. In the right panel, $\widehat{ASL}_{\rm boot}$ equals .152.*

using $t(\mathbf{x}^*)$ defined by (16.4), produced 134 values out of 1000 larger than 1.12 and hence $\widehat{\rm ASL}_{\rm boot}$=.134. In this calculation we used exactly the same set of bootstrap samples that gave the value .120 for $\widehat{\rm ASL}_{\rm boot}$ based on $t(\mathbf{x}) = \bar{z}-\bar{y}$. Unlike in the permutation test, where we showed in Problem **??**.9 that studentization does not affect the answer, studentization does produce a different value for $\widehat{\rm ASL}_{\rm boot}$. However, in this particular approach to bootstrapping the two-sample problem, the difference is typically quite small.

Algorithm 16.1 tests the null hypothesis that the two populations are identical, that is, $F = G$. What if we wanted to test only whether their means were equal? One approach would be to use the two-sample t statistic (16.4). Under the null hypothesis and assuming normal populations with equal variances, this has a Student's t distribution with $n + m - 2$ degrees of freedom. It uses the pooled estimate of standard error $\bar{\sigma}$. If we are not willing to assume that the variances in the two populations are equal, we could base the test on

$$t(\mathbf{x}) = \frac{\bar{z} - \bar{y}}{\sqrt{\bar{\sigma}_1^2/n + \bar{\sigma}_2^2/m}}, \tag{16.5}$$

where $\bar{\sigma}_1^2 = \sum_1^n (z_i - \bar{z})^2/(n-1)$, $\bar{\sigma}_2^2 = \sum_1^m (y_i - \bar{y})^2/(m-1)$. With normal populations, the quantity (16.5) no longer has a Student's t distribution and a number of approximate solutions have therefore been proposed. In the literature this is known as the Behrens-Fisher problem.

The equal variance assumption is attractive for the t-test because it simplifies the form of the resulting distribution. In considering a bootstrap hypothesis test for comparing the two means, there is no compelling reason to assume equal variances and hence we don't make that assumption. To proceed we need estimates of F and G that use only the assumption of a common mean. Letting $\bar{x}$ be the mean of the combined sample, we can translate both samples so that they have mean $\bar{x}$, and then resample each population separately. The procedure is shown in detail in Algorithm 16.2.

The results of this are shown in the right panel of Figure 16.1. The value of $\widehat{\text{ASL}}_{\text{boot}}$ was $152/1000 = .152$.

16.3 Relationship between the permutation test and the bootstrap

The preceding example illustrates some important differences between the permutation test and the bootstrap hypothesis test. A permutation test exploits special symmetry that exists under the null hypothesis to create a permutation distribution of the test statistic. For example, in the two-sample problem when testing $F = G$, all permutations of the order statistic of the combined sample are equally probable. As a result of this symmetry, the ASL from a permutation test is exact: in the two-sample problem, ASL_{perm} is the exact probability of obtaining a test statistic as extreme as the one observed, having fixed the data values of the combined sample.

In contrast, the bootstrap explicitly estimates the probability mechanism under the null hypothesis, and then samples from it to estimate the ASL. The estimate $\widehat{\text{ASL}}_{\text{boot}}$ has no interpretation as an exact probability, but like all bootstrap estimates is only guaranteed to be accurate as the sample size goes to infinity. On the other hand, the bootstrap hypothesis test does not require the

Algorithm 16.2

Computation of the bootstrap test statistic for testing equality of means

1. Let $\hat{F}$ put equal probability on the points $\tilde{z}_i = z_i - \bar{z} + \bar{x}, i = 1, 2, \dots n$, and $\hat{G}$ put equal probability on the points $\tilde{y}_i = y_i - \bar{y} + \bar{x}$, $i = 1, 2, \dots m$, where $\bar{z}$ and $\bar{y}$ are the group means and $\bar{x}$ is the mean of the combined sample.
2. Form B bootstrap data sets $(\mathbf{z}^{*b}, \mathbf{y}^{*b})$ where $\mathbf{z}^{*b}$ is sampled with replacement from $\tilde{z}_1, \tilde{z}_2, \cdots \tilde{z}_n$ and $\mathbf{y}^{*b}$ is sampled with replacement from $\tilde{y}_1, \tilde{y}_2, \dots \tilde{y}_m$.
3. Evaluate $t(\cdot)$ defined by (16.5) on each data set,
$$t(\mathbf{x}^{*b}) = \frac{\bar{z}^{*b} - \bar{y}^{*b}}{\sqrt{\bar{\sigma}_1^{2*b}/n + \bar{\sigma}_2^{2*b}/m}}, \quad b = 1, 2, \cdots B. \tag{16.6}$$
4. Approximate $\mathrm{ASL}_{\mathrm{boot}}$ by
$$\widehat{\mathrm{ASL}}_{\mathrm{boot}} = \#\{t(\mathbf{x}^{*b}) \geq t_{obs}\}/B, \tag{16.7}$$
where $t_{obs} = t(\mathbf{x})$ is the observed value of the statistic.

special symmetry that is needed for a permutation test, and so can be applied much more generally. For instance in the two-sample problem, a permutation test can only test the null hypothesis $F = G$, while the bootstrap can test equal means and equal variances, or equal means with possibly unequal variances.

16.4 The one-sample problem

As our second example, consider a one-sample problem involving only the treated mice. Suppose that other investigators have run experiments similar to ours but with many more mice, and they observed a mean lifetime of 129.0 days for treated mice. We might want to test whether the mean of the treatment group in Table **??** was 129.0 as well:

$$H_0 : \mu_z = 129.0. \tag{16.8}$$

A one sample version of the normal test could be used. Assuming a normal population, under the null hypothesis

$$\bar{z} \sim N(129.0, \sigma^2/n), \tag{16.9}$$

where σ is the standard deviation of the treatment times. Having observed $\bar{z} = 86.9$, the ASL is the probability that a random variable $\bar{z}^*$ distributed accordingly to (16.9) is less than the observed value 86.9

$$\text{ASL} = \Phi(\frac{86.9 - 129.0}{\sigma/\sqrt{n}}), \tag{16.10}$$

where Φ is the cumulative distribution function of the standard normal.

Since σ is unknown, we insert the estimate

$$\bar{\sigma} = \{\sum_1^n (z_i - \bar{z})^2/(n-1)\}^{1/2} = 66.8 \tag{16.11}$$

into (16.10) giving

$$\text{ASL} = \Phi(\frac{-42.1}{66.8/\sqrt{7}}) = 0.05. \tag{16.12}$$

Student's t-test gives a somewhat larger ASL

$$\text{ASL} = \text{Prob}\{t_6 < \frac{-42.1}{66.8/\sqrt{7}}\} = 0.07. \tag{16.13}$$

So there is marginal evidence that the treated mice in our study have a mean survival time of less than 129.0 days. The two-sided ASLs are .10 and .14, respectively.

Notice that a two-sample permutation test cannot be used for this problem. If we had available all of the times for the treated mice (rather than just their mean of 129.0), we could carry out a two-sample permutation test of the equivalence of the two populations. However we do not have available all of the times but know only their mean; we wish to test $H_0 : \mu_z = 129.0$.

In contrast, the bootstrap can be used. We base the bootstrap hypothesis test on the distribution of the test statistic

$$t(\mathbf{z}) = \frac{\bar{z} - 129.0}{\bar{\sigma}/\sqrt{7}} \tag{16.14}$$

under the null hypothesis $\mu_z = 129.0$. The observed value is

$$\frac{86.9 - 129.0}{66.8/\sqrt{7}} = -1.67. \tag{16.15}$$

But what is the appropriate null distribution? We need a distribution $\hat{F}$ that estimates the population of treatment times *under* H_0. Note first that the empirical distribution $\hat{F}$ is not an appropriate estimate for F because it *does not obey* H_0. That is, the mean of $\hat{F}$ is not equal to the null value of 129.0. Somehow we need to obtain an estimate of the population that has mean 129.0. A simple way is to translate the empirical distribution $\hat{F}$ so that it has the desired mean. [1] In other words, we use as our estimated null distribution the empirical distribution on the values

$$\begin{aligned} \tilde{z}_i &= z_i - \bar{z} + 129.0 \\ &= z_i + 42.1 \end{aligned} \tag{16.16}$$

for $i = 1, 2, \cdots 7$. We sample $\tilde{z}_1^*, \ldots \tilde{z}_7^*$ with replacement from $\tilde{z}_1, \ldots \tilde{z}_7$, and for each bootstrap sample compute the statistic

$$t(\tilde{\mathbf{z}}^*) = \frac{\bar{\tilde{z}}^* - 129.0}{\bar{\tilde{\sigma}}^*/\sqrt{7}}, \tag{16.17}$$

where $\bar{\tilde{\sigma}}^*$ is the standard deviation of the bootstrap sample. A total of 100 out of 1000 samples had $t(\tilde{\mathbf{z}}^*)$ less than -1.67, and therefore the achieved significance level is $100/1000 = .10$, as compared to .05 and .07 for the normal and t tests, respectively.

Notice that our choice of null distribution assumes that the possible distributions for the treatment times, as the mean times vary, are just translated versions of one another. Such a family of distributions is called a *translation family.* This assumption is also present in the normal and t tests; but in those tests we assume further that the populations are normal. In either case, it might be sensible to take logarithms of the survival times before carrying out the analysis, because the logged lifetimes are more likely to satisfy a translation or normal family assumption (Problem 16.1).

There is a different but equivalent way of bootstrapping the one-sample problem. We draw with replacement from the (untranslated) data values $z_1, z_2, \ldots z_7$ and compute the statistic

$$t(\mathbf{z}^*) = \frac{\bar{z}^* - \bar{z}}{\bar{\sigma}^*/\sqrt{7}}, \tag{16.18}$$

[1] A different method is discussed in Problem 16.5.

where $\bar{\sigma}^*$ is the standard deviation of the bootstrap sample. This statistic is the same as (16.17) since

$$\bar{\bar{z}}^* - 129.0 = (\bar{z}^* - \bar{z} + 129.0) - 129.0 = \bar{z}^* - \bar{z}$$

and the standard deviations are equal as well. This also shows the equivalence between the one-sample bootstrap hypothesis test and the bootstrap-t confidence interval described in Chapter **??**. That interval is based on the percentiles of the statistic (16.18) under bootstrap sampling from $z_1, z_2, \dots z_7$, exactly as above. Therefore the bootstrap-t confidence interval consists of those values μ_0 that are not rejected by the bootstrap hypothesis test described above. This general connection between confidence intervals and hypothesis tests is given in more detail in Section **??**.

16.5 Testing multimodality of a population

Our second example is a much more exotic one. It is a case where a simple normal theory test does not exist and a permutation test cannot be used, but the bootstrap can be used effectively. The data are the thicknesses in millimeters of 485 stamps, printed in 1872. The stamp issue of that year was thought to be a "philatelic mixture", that is, printed on more than one type of paper. It is of historical interest to determine how many different types of paper were used.

A histogram of the data is shown in the top left panel of Figure 16.2. This sample is part of a large population of stamps from 1872, and we can imagine the distribution of thickness measurements for this population. We pose the statistical question: how many modes does this population have? A mode is defined to be a local maximum or "bump" of the population density. The number of modes is suggestive of the number of distinct types of paper used in the printing.

From the histogram in Figure 16.2, it appears that the population might have 2 or more modes. It is difficult to tell, however, because the histogram is not smooth. To obtain a smoother estimate, we can use a *Gaussian kernel density estimate*. Denoting the data by $x_1, \dots x_n$, a Gaussian kernel density estimate is defined by

$$\hat{f}(t;h) = \frac{1}{nh}\sum_{1}^{n}\phi\Big(\frac{t - x_i}{h}\Big), \tag{16.19}$$

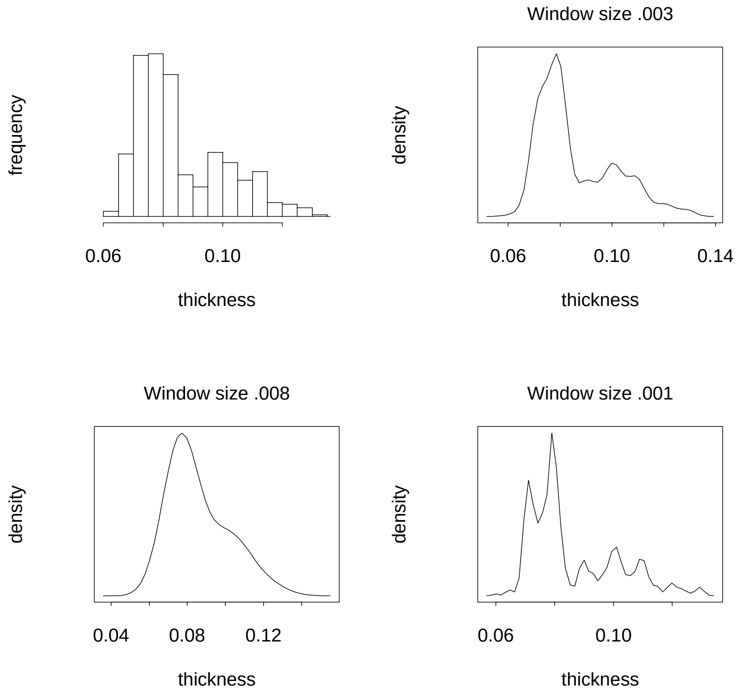

Figure 16.2. *Top left panel shows histogram of thicknesses of 485 stamps. Top right and bottom panels are Gaussian kernel density estimates for the same sample, using window size .003 (top right), .008 (bottom left) and .001 (bottom right).*

where $\phi(t)$ is the standard normal density $(1/\sqrt{2\pi})\exp(-t^2/2)$. The parameter h is called the *window size* and determines the amount of *smoothing* that is applied to the data. Larger values of h produce a smoother density estimate.

We can think of (16.19) as adding up n little Gaussian density curves centered at each point x_i, each having standard deviation h; Figure 16.3 illustrates this.

The top right panel of Figure 16.2 shows the resulting density estimate using $h = .003$; there are 2 or 3 modes. However by varying h, we can produce a greater or lesser number of modes. The

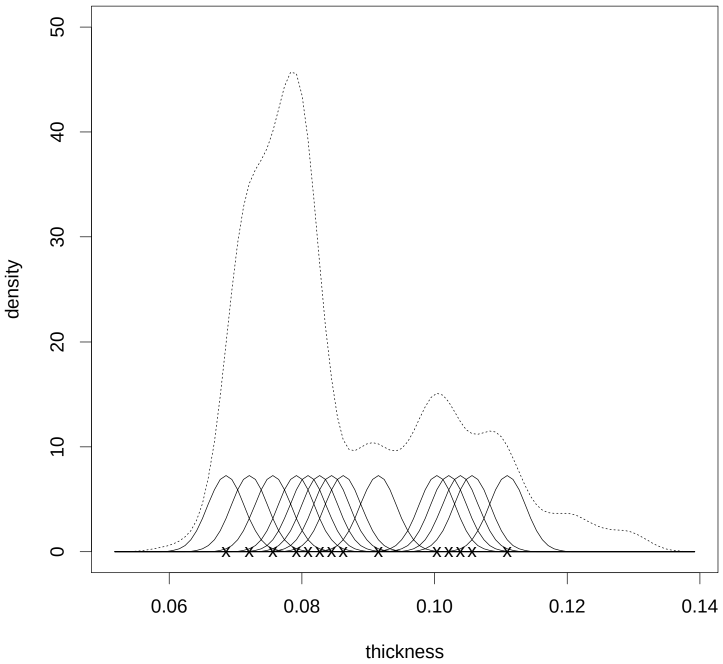

Figure 16.3. *Illustration of a Gaussian kernel density estimate. A small Gaussian density is centered at each data value (marked with an "x") and the density estimate (broken line) at each value is determined by adding up the values of all the Gaussian densities at that point. For the stamp data there are actually 485 little Gaussian densities used (one for each point); for clarity we have shown only a few.*

bottom left and right show the estimates obtained using $h = .008$ and $h = .001$, respectively. The former has one mode, while the latter has at least 7 modes! Clearly the inference that we draw from our data depends strongly on the value of h that we choose.

If we approach the problem in terms of hypothesis testing, there is a natural way to choose h. We will need the following important result, which we state without proof: as h increases, the number

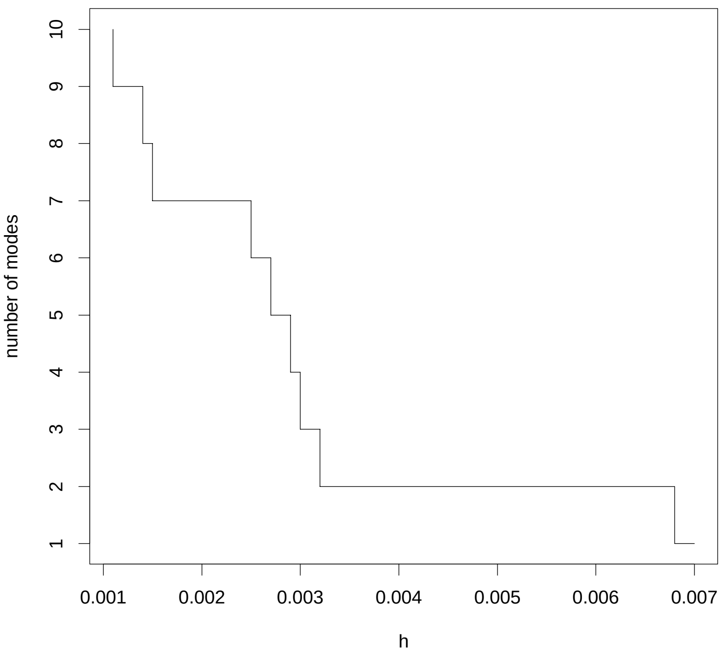

Figure 16.4. *Stamp data: number of modes in the Gaussian kernel density estimate as a function of the window size h.*

of modes in a Gaussian kernel density estimate is non-increasing. This is illustrated for the stamp data in Figure 16.4.

Now consider testing

$$H_0 : \text{number of modes} = 1 \tag{16.20}$$

versus number of modes > 1. Since the number of modes decreases as h increases, there is a smallest value of h such that $\hat{f}(t;\ h)$ has one mode. Call this $\hat{h}_1$. Looking at Figure 16.4, $\hat{h}_1 \approx .0068$.

It seems reasonable to use $\hat{f}(t;\ \hat{h}_1)$ as the estimated null distribution for our test of H_0. In a sense, it is the density estimate closest to our data that is consistent with H_0. By "closest", we mean that it uses the least amount of smoothing (smallest value of h) among all estimates with one mode.

There is one small adjustment that we make to $\hat{f}(\cdot;\ \hat{h}_1)$. Formula (16.19) artificially increases the variance of the estimate (Problem 16.2), so we rescale it to have variance equal to the sample variance. Denote the rescaled estimate by $\hat{g}(\cdot; \hat{h}_1)$.

Finally, we need to select a test statistic. A natural choice is $\hat{h}_1$, the smallest window size producing a density estimate with one mode. A large value of $\hat{h}_1$ indicates that a great deal of smoothing must be done to create an estimate with one mode and is therefore evidence against H_0.

Putting all of this together, the bootstrap hypothesis test for H_0 : number of modes = 1 is based on the achieved significance level

$$\mathrm{ASL}_{\mathrm{boot}} = \mathrm{Prob}_{\hat{g}(\cdot;\ \hat{h}_1)}\{\hat{h}_1^* > \hat{h}_1\}. \tag{16.21}$$

Here $\hat{h}_1$ is fixed at its observed value of .0068; the bootstrap sample $x_1^*, x_2^* \dots x_n^*$ is drawn from $\hat{g}(\cdot;\ \hat{h}_1)$ and $\hat{h}_1^*$ is the smallest value of h producing a density estimate with one mode from the bootstrap data $x_1^*, x_2^* \dots x_n^*$.

To approximate $\mathrm{ASL}_{\mathrm{boot}}$ we need to draw bootstrap samples from the rescaled density estimate $\hat{g}(\cdot;\ \hat{h}_1)$. That is, rather than sampling with replacement from the data, we sample from a smooth estimate of the population. This is called *the smooth bootstrap.* Because of the convenient form of the Gaussian kernel estimate, drawing samples from $\hat{g}(\cdot;\ \hat{h}_1)$ is easy. We sample $y_1^*, y_2^*, \dots y_n^*$ with replacement from $x_1, x_2, \dots x_n$ and set

$$x_i^* = \bar{y}^* + (1 + \hat{h}_1^2/\hat{\sigma}^2)^{-1/2}(y_i^* - \bar{y}^* + \hat{h}_1\epsilon_i);\ \ i = 1, 2, \dots n, \tag{16.22}$$

where $\bar{y}^*$ is the mean of $y_1^*, y_2^*, \dots y_n^*$, $\hat{\sigma}^2$ is the plug estimate of variance of the data and ϵ_i are standard normal random variables. The factor $(1 + \hat{h}_1^2/\hat{\sigma}^2)^{-1/2}$ scales the estimate so that its variance is approximately $\hat{\sigma}^2$ (Problem 16.3.) A summary of the steps is shown in Algorithm 16.3. (Actually a computational shortcut is possible for step 2; see Problem 16.3.)

We carried out this process with $B = 500$. Out of 500 bootstrap samples, none had $\hat{h}_1^* > .0068$, so $\widehat{\mathrm{ASL}}_{\mathrm{boot}} = 0$. We repeated this for H_0 : number of modes $= 2, 3, \dots$, and Table 16.1 shows the resulting P-values. Interpreting these results in a sequential manner, starting with number of modes = 1, we reject the unimodal hypothesis but do not reject the hypothesis of 2 modes. This is

Algorithm 16.3

Computation of the bootstrap test statistic for multimodality

1. Draw B bootstrap samples of size n from $\hat{g}(\cdot;\ \hat{h}_1)$ using (16.22).
2. For each bootstrap sample compute $\hat{h}_1^*$ the smallest window width that produces a density estimate with one mode. Denote the B values of $\hat{h}_1^*$ by $\hat{h}_1^*(1), \ldots \hat{h}_1^*(B)$.
3. Approximate $\mathrm{ASL}_{\mathrm{boot}}$ by

$$\widehat{\mathrm{ASL}}_{\mathrm{boot}} = \#\{\hat{h}_1^*(b) \geq \hat{h}_1\}/B. \qquad (16.23)$$

where the inference process would end in many instances. If we were willing to entertain more exotic hypotheses, then from Table 16.1 there is also a suggestion that the population might have 7 modes.

16.6 Discussion

As the examples in this chapter illustrate, the two quantities that we must choose when carrying out a bootstrap hypothesis test are:

(a) A test statistic $t(\mathbf{x})$.

(b) A null distribution $\hat{F}_0$ for the data under H_0.

Given these, we generate B bootstrap values of $t(\mathbf{x}^*)$ under $\hat{F}_0$ and estimate the achieved significance level by

$$\widehat{\mathrm{ASL}}_{\mathrm{boot}} = \#\{t(\mathbf{x}^{*b}) \geq t(\mathbf{x})\}/B. \qquad (16.24)$$

As the stamp example shows, sometimes the choice of $t(\mathbf{x})$ and $\hat{F}_0$ are not obvious. The difficulty in choosing $\hat{F}_0$ is that, in most instances, H_0 is a composite hypothesis. In the stamp example, H_0 refers to all possible densities with one mode. A good choice for $\hat{F}_0$ is the distribution that obeys H_0 and is most reasonable for our data; this choice makes the test conservative, that is, the test is less likely to falsely reject the null hypothesis. In the stamp example, we tested for unimodality by generating samples from the unimodal distribution that is mostly nearly bimodal. In other

Table 16.1. *P-values for stamp example.*

number of modes(m)	$\hat{h}_m$	P-value
1	.0068	.00
2	0032	.29
3	.0030	.06
4	.0029	.00
5	.0027	.00
6	.0025	.00
7	.0015	.46
8	.0014	.17
9	.0011	.17

words, we used the smallest possible value for $\hat{h}_1$ and this makes the probability in (16.21) as large as possible.

The choice of test statistic $t(\mathbf{x})$ will determine the power of the test, that is, the chance that we reject H_0 when it is false. In the stamp example, if the actual population density is bimodal but the Gaussian kernel density does not approximate it accurately, then the test based on the window width $\hat{h}_1$ will not have high power.

Bootstrap tests are useful in situations where the alternative hypothesis is not well-specified. In cases where there is a parametric alternative hypothesis, likelihood or Bayesian methods might be preferable.

16.7 Bibliographic notes

Monte Carlo tests, related to the tests in this chapter, are proposed in Barnard (1963), Hope (1968), and Marriott (1979); see also Hall and Titterington (1989). Some theory of bootstrap hypothesis testing, and its relation to randomization tests, is given by Romano (1988, 1989). A discussion of practical issues appears in Hinkley (1988, 1989), Young (1988b), Noreen (1989), Fisher and Hall (1990), and Hall and Wilson (1991). See also Tibshirani (1992) for a comment on Hall and Wilson (1991). Young (1986) describe simulation-based hypothesis testing in the context of geometric statistics. Beran and Millar (1987) develop general asymptotic theory for stochastic minimum distance tests. In this work, the test statistic is the distance to a composite null hypothesis and a stochastic search procedure is used to approximate it. Besag and Clifford (1989) propose methods based on Markov chains for significance testing with dependent data. The two-sample problem with unequal variance has a long history: see, for example, Behrens (1929) and Welch (1947); Cox and Hinkley (1974) and Robinson (1982) give a more modern account. The use of the bootstrap for testing multimodality is proposed in Silverman (1981, 1983). It is applied to the stamp data in Izenman and Sommer (1988). Density estimation is described in many books, including Silverman (1986) and Scott (1992). The smooth bootstrap is studied by Silverman and Young (1987) and Hall, DiCiccio and Romano (1989).

16.8 Problems

16.1 Explain why the logarithm of survival times are more likely to be normally distributed than the times themselves.

16.2 (a) If y_i is sampled with replacement from $x_1, x_2, \dots x_n$, ϵ_i has a standard normal distribution and $\hat{h}_1$ is considered fixed, show that

$$r_i = y_i^* + \hat{h}_1 \epsilon_i \tag{16.25}$$

is distributed according to $\hat{f}(\cdot;\ \hat{h}_1)$, the Gaussian kernel density estimate defined by (16.19).

(b) Show that x_i^* given by (16.22) has the same mean as r_i^* but has variance approximately equal to $\hat{\sigma}^2$ rather than $\hat{\sigma}^2 + \hat{h}_1^2$ (the variance of r_i^*).

16.3 Denote by $\hat{h}_k$ the smallest window width producing a density estimate with k modes from our original data, and let $\hat{h}_k^*$ be the corresponding quantity for a bootstrap sample $\mathbf{x}^*$. Show that event

$$\{\hat{h}_k^* > \hat{h}_k\} \tag{16.26}$$

is the same as the event

$$\{\hat{f}^*(\cdot;\ \hat{h}_k) \text{ has more than } k \text{ modes}\}, \tag{16.27}$$

where $\hat{f}^*(\cdot;\ \hat{h}_k)$ is the Gaussian kernel density estimate based on the bootstrap sample $\mathbf{x}^*$. Hence it is not necessary to find $\hat{h}_k^*$ for each bootstrap sample; one need only check whether $\hat{f}(\cdot;\ \hat{h}_k)$ has more than k modes.

16.4 In the second example of this chapter, we tested whether the mean of the treatment group was equal to 129.0. We argued that one should not use the empirical distribution as the null distribution but rather should first translate it to have mean 129.0. In this problem we carry out a small simulation study to investigate this issue.

(a) Generate 100 samples $\mathbf{z}$ of size 7 from a normal population with mean 129.0 and standard deviation 66.8. For each sample, perform a bootstrap hypothesis test of $\mu_z = 129.0$ using the test statistic $\bar{z} - 129.0$ and using as the estimated null distribution 1) the empirical distribution, and 2) the empirical distribution translated to have mean 129.0.

Compute the average of ASL for each test, averaged over the 100 simulations.

(b) Repeat (a), but simulate from a normal population with a mean of 170. Discuss the results.

16.5 Suppose we have a sample $z_1, z_2, \ldots z_n$, and we want an estimate of the underlying population F restricted to have mean μ. One approach, used in Section 16.4, is to use the empirical distribution on the translated data values $z_i - \bar{z} + \mu$. A different approach is to leave the data values fixed, and instead change the probability p_i on each data value. Let $\mathbf{p} = (p_1, p_2, \ldots p_n)$ and let F_p be the distribution putting probability p_i on x_i for each i. Then it is reasonable to choose $\mathbf{p}$ so that the mean of $F_p = \sum p_i x_i = \mu$, and F_p is as close as possible to the empirical distribution $\hat{F}$. A convenient measure of closeness is the Kullback-Leibler distance

$$d_{F_p}(F_p, \hat{F}) = \sum_1^n p_i \log\Big(\frac{1}{np_i}\Big). \tag{16.28}$$

(a) Using Lagrange multipliers, show that the probabilities that minimize expression (16.28) subject to $\sum p_i x_i = \mu$, $\sum p_i = 1$ are given by

$$p_i = \frac{\exp(tx_i)}{\sum_{i=1}^n \exp(tx_i)} \tag{16.29}$$

where t is chosen so that $\sum p_i x_i = \mu$. This is sometimes called an *exponentially tilted* version of $\hat{F}$.

(b) Use this approach to carry out a test of $\mu = 129.0$ in the mouse data example of Section 16.4 and compare the results to those in that section.