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

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

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 12. That interval is based on the percentiles of the statistic (16.18) under bootstrap sampling from $z_1, z_2, \ldots z_7$, 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 12.3.

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 $x_1, \ldots x_n$, a Gaussian kernel density estimate is defined by

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

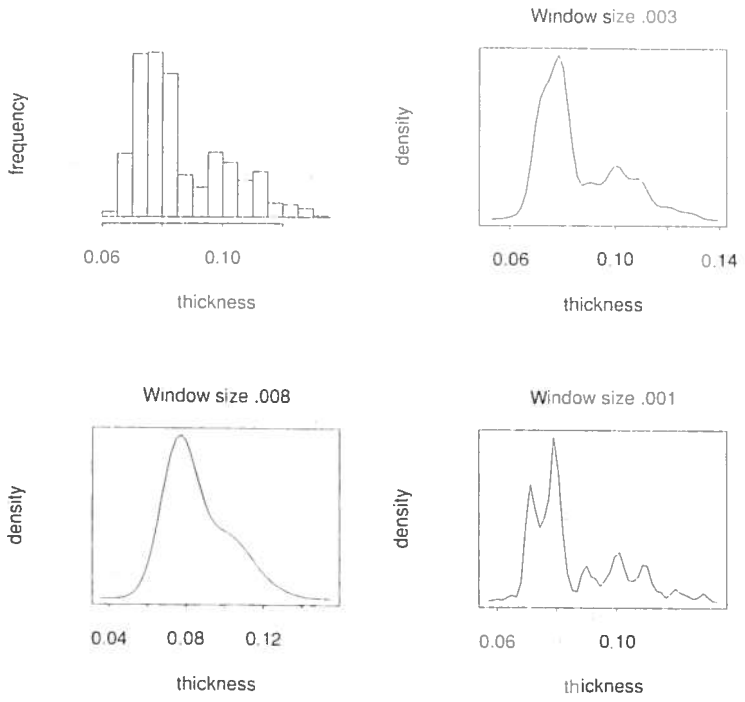

Figure 16.2. *Top left panel shows histogram of thicknesses of 485 stamps. Top right and bottom panels are Gaussian kernel density estimates for the same sample, using window size .003 (top right), .008 (bottom left) and .001 (bottom right).*

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

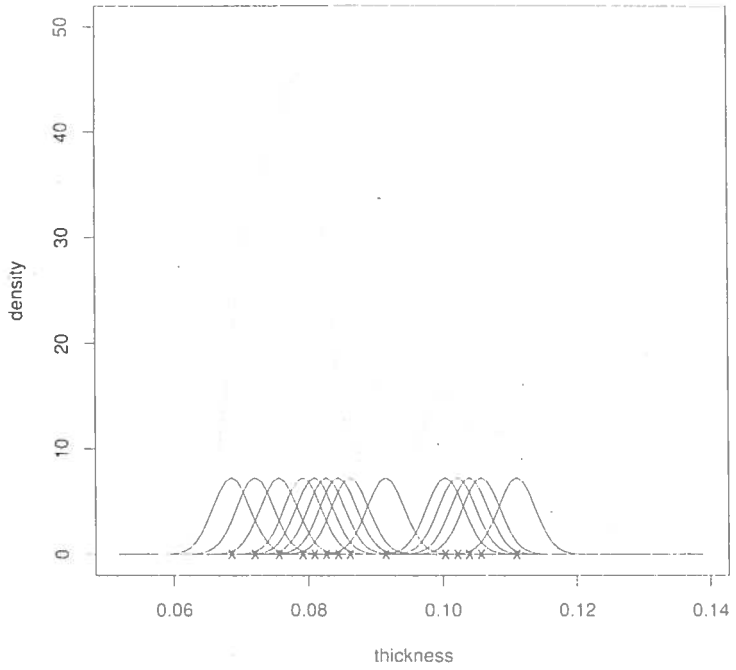

Figure 16.3. *Illustration of a Gaussian kernel density estimate. A small Gaussian density is centered at each data value (marked with an "x") and the density estimate (broken line) at each value is determined by adding up the values of all the Gaussian densities at that point. For the stamp data there 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

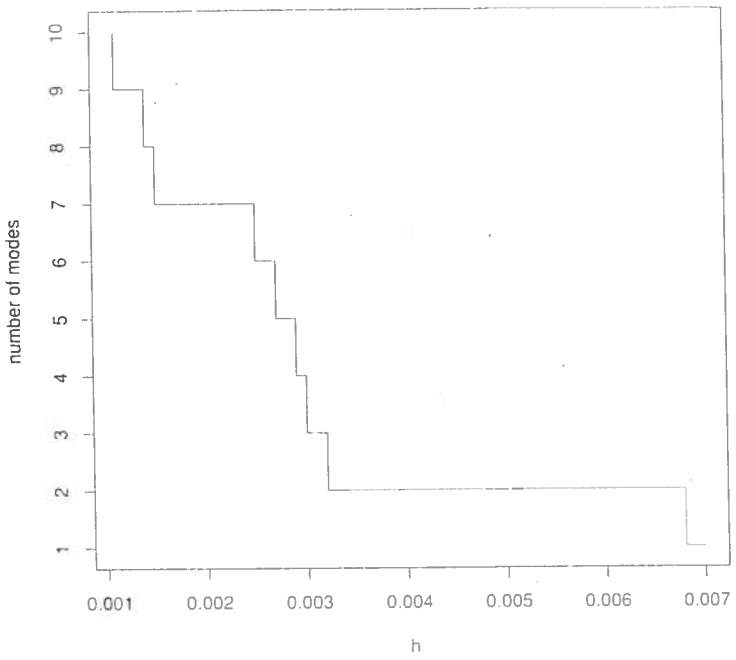

Figure 16.4. *Stamp data: number of modes in the Gaussian kernel density estimate as a function of the window size h.*

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

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

Here $\hat{h}_1$ is fixed at its observed value of .0068; the bootstrap sample $x_1^*, x_2^* \ldots x_n^*$ 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 $x_1^*, x_2^* \ldots x_n^*$.

To approximate ASL_{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 $y_1^*, y_2^*, \ldots y_n^*$ with replacement from $x_1, x_2, \ldots x_n$ and set

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

where $\bar{y}^*$ is the mean of $y_1^*, y_2^*, \ldots y_n^*$, $\hat{\sigma}^2$ 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 $\widehat{\text{ASL}}_{\text{boot}} = 0$. We repeated this for H_0 : number of modes = $2, 3, \ldots$, 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 ASL_{boot} by

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

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

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

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)	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	.47
9	.0011	.47

words, we used the smallest possible value for h_1 and this makes the probability in (16.21) as large as possible.

The choice of test statistic $t(x)$ 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 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