Cambridge Series on Statistical and Probabilistic Mathematics

This series of high quality upper-division textbooks and expository monographs covers all areas of stochastic applicable mathematics. The topics range from pure and applied statistics to probability theory, operations research, mathematical programming, and optimzation. The books contain clear presentations of new developments in the field and also of the state of the art in classical methods. While emphasizing rigorous treatment of theoretical methods, the books contain important applications and discussions of new techniques made possible be advances in computational methods.

Bootstrap methods and their application

A. C. Davison
Professor of Statistics, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne

D. V. Hinkley
Professor of Statistics, Department of Statistics and Applied Probability, University of California, Santa Barbara

CAMBRIDGE
UNIVERSITY PRESS

4

Tests

4.1 Introduction

Many statistical applications involve significance tests to assess the plausibility of scientific hypotheses. Resampling methods are not new to significance testing, since randomization tests and permutation tests have long been used to provide nonparametric tests. Also Monte Carlo tests, which use simulated datasets, are quite commonly used in certain areas of application. In this chapter we describe how resampling methods can be used to produce significance tests, in both parametric and nonparametric settings. The range of ideas is somewhat wider than the direct bootstrap approach introduced in the preceding two chapters. To begin with, we summarize some of the key ideas of significance testing.

The simplest situation involves a *simple null hypothesis* H_0 which completely specifies the probability distribution of the data. Thus, if we are dealing with a single sample $y_1, \ldots, y_n$ from a population with CDF F, then H_0 specifies that $F = F_0$, where F_0 contains no unknown parameters. An example would be "exponential with mean 1". The more usual situation in practice is that H_0 is a *composite null hypothesis*, which means that some aspects of F are not determined and remain unknown when H_0 is true. An example would be "normal with mean 1", the variance of the normal distribution being unspecified.

P-values

A statistical test is based on a *test statistic* T which measures the discrepancy between the data and the null hypothesis. In general discussion we shall follow the convention that large values of T are evidence against H_0. Suppose for the moment that this null hypothesis is simple. If the observed value of the test statistic is denoted by t then the level of evidence against H_0 is measured by

the *significance probability*

$$p = \Pr(T \geq t \mid H_0), \tag{4.1}$$

often called the *P-value*. A corresponding notion is that of a critical value t_p for t, associated with testing at level p: if $t \geq t_p$ then H_0 is rejected at level p, or $100p\%$. Necessarily t_p is defined by $\Pr(T \geq t_p \mid H_0) = p$. The level p is also called the *error rate* or the *size* of the test, and $\{(y_1, \ldots, y_n) : t \geq t_p\}$ is called the *level p critical region* of the test. The distribution of T under H_0 is called the *null distribution* of T.

Under H_0 the P-value (4.1) has a uniform distribution on [0,1], if T is continuous, so that the corresponding random variable P has distribution

$$\Pr(P \leq p \mid H_0) = p. \tag{4.2}$$

This yields the error rate interpretation of the P-value, namely that if the observed test statistic were regarded as just decisive against H_0, then this is equivalent to following a procedure which rejects H_0 with error rate p. The same is not exactly true if T is discrete, and for this reason modifications to (4.1) are sometimes suggested for discrete data problems: we shall not worry about the distinction here.

It is important in applications to give a clear idea of the degree of discrepancy between data and null hypothesis, if not giving the P-value itself then at least indicating how it compares to several levels, say $p = 0.10, 0.05, 0.01$, rather than just testing at the 0.05 level.

Choice of test statistic

In the parametric setting, we have an explicit form for the sampling distribution of the data with a finite number of unknown parameters. Often the null hypothesis specifies numerical values for, or relationships between, some or all of these parameters. There is also an *alternative hypothesis* H_A which describes what alternatives to H_0 it is most important to detect, or what is thought likely to be true if H_0 is not. This alternative hypothesis guides the specific choice of T, usually through use of the likelihood function

$$L(\theta) = f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n \mid \theta),$$

i.e. the joint density of the observations. For example, when H_0 and H_A are both simple, say $H_0 : \theta = \theta_0$ and $H_A : \theta = \theta_A$, then the best test statistic is the likelihood ratio

$$T = L(\theta_A)/L(\theta_0). \tag{4.3}$$

A rather different situation is where we wish to test the goodness of fit of the parametric model. Sometimes this can be done by embedding the model into a larger model, with one or a few additional parameters corresponding

to departure from the original model. We would then test those additional parameters. Otherwise general purpose goodness of fit tests will be used, for example chi-squared tests.

In the nonparametric setting, no particular forms are specified for the distributions. Then the appropriate choice of T is less clear, but it should be based on at least a qualitative notion of what is of concern should H_0 not be true. Usually T would be based on a statistical function $s(\hat{F})$ that reflects the characteristic of physical interest and for which the null hypothesis specifies a value. For example, suppose that we wish to test the null hypothesis H_0 that X and Y are independent, given the random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$. The correlation $s(F) = \text{corr}(X, Y) = \rho$ is a convenient measure of dependence, and $\rho = 0$ under H_0. If the alternative hypothesis is positive dependence, then a natural test statistic is $T = s(\hat{F})$, the raw sample correlation; if the alternative hypothesis is just "dependence", then the two-sided test statistic $T = s^2(\hat{F})$ could be used.

Conditional tests

In most parametric problems and all nonparametric problems, the null hypothesis H_0 is composite, that is it leaves some parameters unknown and therefore does not completely specify F. Therefore P-value (4.1) is not generally well-defined, because $\Pr(T \geq t \mid F)$ may depend upon which F satisfying H_0 is taken. There are two clean solutions to this difficulty. One is to choose T carefully so that its distribution is the same for all F satisfying H_0: examples include the Student-t test for a normal mean with unknown variance, and rank tests for nonparametric problems. The second and more widely applicable solution is to eliminate the parameters which remain unknown when H_0 is true by conditioning on the sufficient statistic under H_0. If this sufficient statistic is denoted by S, then we define the conditional P-value by

$$p = \Pr(T \geq t \mid S = s, H_0). \tag{4.4}$$

Familiar examples include the Fisher exact test for a 2×2 table and the Student-t test mentioned earlier. Other examples will be given in the next two sections.

A less satisfactory approach, which can nevertheless give good approximations, is to estimate F by a CDF $\hat{F}_0$ which satisfies H_0 and then calculate

$$p = \Pr(T \geq t \mid \hat{F}_0). \tag{4.5}$$

Typically this value will not satisfy (4.2) exactly, but will deviate by an amount which may be practically negligible.

Pivot tests

When the null hypothesis concerns a particular parameter value, the equivalence between significance tests and confidence sets can be used. This equiv-

alence is that if the value θ_0 is outside a $1 - \alpha$ confidence set for θ, then θ differs from θ_0 with P-value less than α. The particular alternative hypothesis for which this applies is determined by the type of confidence set: for example, if the confidence set is all values to the right of a lower confidence limit, then the implied alternative is $H_A : \theta > \theta_0$. A specific form of test based on this equivalence is the *pivot test*. For example, suppose that T is an estimator for scalar θ, with estimated variance V. Suppose further that the studentized form $Z = (T - \theta)/V^{1/2}$ is a pivot, meaning that its distribution is the same for all relevant F, and in particular for all θ. The Student-t statistic is a familiar instance of this. For the one-sided test of $H_0 : \theta = \theta_0$ versus $H_A : \theta > \theta_0$, the P-value attached to the observed studentized test statistic $z_0 = (t - \theta_0)/v^{1/2}$ is

$$p = \Pr\{(T - \theta_0)/V^{1/2} \geq (t - \theta_0)/v^{1/2} \mid H_0\}.$$

But because Z is a pivot,

$$\Pr\{Z \geq (t - \theta_0)/v^{1/2} \mid H_0\} = \Pr\{Z \geq (t - \theta_0)/v^{1/2} \mid F\},$$

and therefore

$$p = \Pr\left(Z \geq z_0 \mid F\right). \tag{4.6}$$

The particular advantage of this, in the resampling context, is that we do not have to construct a special null hypothesis sampling distribution.

In parametric problems it is usually possible to express the model in terms of the parameter of interest ψ and other (nuisance) parameters λ, so that the null hypothesis concerns only ψ. In the above discussion of conditional tests, (4.4) would be independent of λ. One general approach to construction of a test statistic T is to generalize the simple likelihood ratio (4.3), and to define

$$LR = \frac{\max_{H_A} L(\psi, \lambda)}{\max_{H_0} L(\psi, \lambda)}.$$

For testing $H_0 : \psi = \psi_0$ versus $H_A : \psi \neq \psi_0$, this generalized likelihood ratio is equivalent to the more convenient expression

$$LR = \frac{L(\hat{\psi}, \hat{\lambda})}{L(\psi_0, \hat{\lambda}_0)} = \frac{\max_{\psi,\lambda} L(\psi, \lambda)}{\max_{\lambda} L(\psi_0, \lambda)}. \tag{4.7}$$

Of course this also applies when there is no nuisance parameter. For many models it is possible to show that $T = 2 \log LR$ has approximately the χ^2_d distribution under H_0, where d is the dimension of ψ, so that

$$p \doteq \Pr(\chi^2_d \geq t), \tag{4.8}$$

independently of λ. Thus the likelihood ratio LR is an approximate pivot.

There is a variety of related statistics, including the score statistic, and the signed likelihood ratio for one-parameter problems. With each likelihood-based

statistic there is a simple approximation to the null distribution, and modifications to improve approximation in moderate-sized samples. The likelihood ratio method appears limited to parametric problems, but as we shall see in Chapter 10 it is possible to define analogues in the nonparametric case.

With all of the P-value calculations introduced thus far, simple approximations for p exist in many cases by appealing to limiting results as n increases. Part of the purpose of this chapter is to provide resampling alternatives to such approximations when they either fail to give appropriate accuracy or do not exist at all. Section 4.2 discusses ways in which resampling and simulation can help with parametric tests, starting with exact Monte Carlo tests. Section 4.3 briefly reviews permutation and randomization tests. This leads on to the wider topic of nonparametric bootstrap tests in Section 4.4. Section 4.5 describes a simple method for improving P-values when these are biased. Most of the examples in this chapter involve relatively simple applications. Chapters 6 and beyond contain more substantial applications.

4.2 Resampling for Parametric Tests

Broadly speaking, parametric resampling may be useful in any testing problem where either standard approximations do not apply or where the accuracy of such approximations is suspect. There is a wide range of such problems, including hypotheses with order constraints, hypotheses involving separate models, and graphical tests. In all of these problems, the basic method is to use a parametric resampling scheme as outlined in Section 2.2 except that here the simulation model must satisfy the relevant null hypothesis.

4.2.1 Monte Carlo tests

One special situation is when the null hypothesis distribution of T does not involve any nuisance parameters. Occasionally this happens directly, but more often it is induced, either by standardizing some initial statistic, or by conditioning on a sufficient statistic, as explained earlier. In the latter case the exact P-value is given by (4.4) rather than (4.1). In practice the exact P-value may be difficult or impossible to calculate, and Monte Carlo tests provide convenient approximations to the full tests. As we shall see, Monte Carlo tests are exact in their own right, and among bootstrap tests are special in this way.

The basic Monte Carlo test compares the observed statistic t to R independent values of T which are obtained from corresponding samples independently simulated under the null hypothesis model. If these simulated values are denoted by $t_1^*,\ldots,t_R^*$, then under H_0 all $R+1$ values $t, t_1^*,\ldots,t_R^*$ are equally likely values of T. That is, assuming T is continuous,

$$\Pr(T < T^*_{(r)} \mid H_0) = \frac{r}{R+1}, \tag{4.9}$$

where as usual $T^*_{(r)}$ denotes the rth ordered value. If exactly k of the simulated t^* values exceed t and none equal it, then

$$p = \Pr(T \geq t \mid H_0) \doteq p_{\text{mc}} = \frac{k+1}{R+1}. \tag{4.10}$$

The right-hand side is referred to as the Monte Carlo P-value. If T is continuous, then it follows from (4.9) that under H_0 the distribution of the corresponding random variable P_{mc} is uniform on $(\frac{1}{R+1},\ldots,\frac{R}{R+1},1)$. This result is the discrete analogue of (4.2), and guarantees that P_{mc} has the error rate interpretation. In this sense the Monte Carlo test is exact. It differs from the full test, which corresponds to $R=\infty$, by blurring the critical region of the full test for any attainable level.

If T is discrete, then repeat values of t^* can occur. If exactly l of the t^* values equal t, then it is sometimes advocated that one bounds the significance probability,

$$\frac{k+1}{R+1} \leq p_{\text{mc}} \leq \frac{k+l+1}{R+1}.$$

Our strict interpretation of (4.1) would have us use the upper bound, and so we adopt the general definition

#(A) means the number of times the event A occurs.

$$p_{\text{mc}} = \frac{1+\#\{t_r^* \geq t\}}{R+1}. \tag{4.11}$$

Example 4.1 (Logistic regression) Suppose that $y_1,\ldots,y_n$ are independent binary outcomes, with corresponding scalar covariate values $x_1,\ldots,x_n$, and that we wish to test whether or not x influences y. If our chosen model is the logistic regression model

$$\log\frac{\Pr(Y_j=1 \mid x_j)}{\Pr(Y_j=0 \mid x_j)} = \lambda + \psi x_j, \quad j=1,\ldots,n,$$

then the null hypothesis is $H_0: \psi = 0$. Under H_0 the sufficient statistic for λ is $S=\sum Y_j$ and $T=\sum x_j Y_j$ is the natural test statistic; T is in fact optimal for the logistic model, but is also effective for monotone transformations of the odds ratio other than logarithm. The significance is to be calculated according to (4.4).

The null distribution of $Y_1,\ldots,Y_n$ given $S=s$ is uniform over all $\binom{n}{s}$ permutations of $y_1,\ldots,y_n$. Rather than generate all of these permutations to compute (4.4) exactly, we can generate R random permutations and apply (4.11). A simulated sample will then be $(x_1,y_1^*),\ldots,(x_n,y_n^*)$, where $y_1^*,\ldots,y_n^*$ is a random permutation of $y_1,\ldots,y_n$, and the associated test statistic will be $t^*=\sum x_j y_j^*$.

0	1	2	3	4	3	4	2	2	1
0	2	0	2	4	2	3	3	4	2
1	1	1	1	4	1	5	2	2	3
4	1	2	5	2	0	3	2	1	1
3	1	4	3	1	0	0	2	7	0

Table 4.1 $n = 50$ counts of balsam-fir seedlings in five feet square quadrats.

In some applications there will be repeats among the x values, or equivalently m_i binomial trials with a_i occurrences of $y = 1$ at the ith distinct value of x. If the data are expressed in the latter form, then the same random permutation procedure can be applied to the original expanded form of data with $n = \sum m_i$ individual ys. ■

Example 4.2 (Overdispersed counts) The data in Table 4.1 are $n = 50$ counts of fir seedlings in small quadrats, part of a larger dataset. The actual spatial layout is preserved, although we are not concerned with this here. Rather we wish to test the null hypothesis that these data are a random sample from a Poisson distribution with unknown mean. The concern is that the data are overdispersed relative to the Poisson distribution, which strongly suggests that we take as test statistic the dispersion index $T = \sum(Y_j - \bar{Y})^2/\bar{Y}$. Under the Poisson model $S = \sum Y_j$ is sufficient for the common mean, so we carry out a conditional test and apply (4.4). For the data, $t = 55.15$ and $s = 107$.

Now under the null hypothesis Poisson model, the conditional distribution of $Y_1, \ldots, Y_n$ given $\sum Y_j = s$ is multinomial with denominator s and n categories each having probability n^{-1}. It is easy to simulate from this distribution. In the first $R = 99$ simulated values t^*, 24 are larger than $t = 55.15$. So the Monte Carlo P-value (4.11) is equal to 0.25, and we conclude that the data dispersion is consistent with Poisson dispersion. Increasing R to 999 makes little difference, giving $p = 0.235$. The left panel of Figure 4.1 shows a histogram of all 999 values of $t^* - t$: the unshaded part of the histogram corresponds to values $t^* \geq t$ which count toward significance.

For this simple problem the null distribution of T given $S = s$ is approximately χ^2_{n-1}. That this approximation is accurate for our data is illustrated in the right panel of Figure 4.1, which plots the ordered values of t^* against quantiles of the χ^2_{49} distribution. The P-value obtained with this approximation is 0.253, close to the exact value. There are two points to make about this. First, the simulation results enable us to check on the accuracy of the theoretical approximation: if the approximation is good, then we can use it; but if it isn't, then we have the Monte Carlo P-value. Secondly, the Monte Carlo method does not require knowledge of a theoretical approximation, which may not even exist in more complicated problems, such as spatial analysis of these data. The Monte Carlo method applies very generally. ■

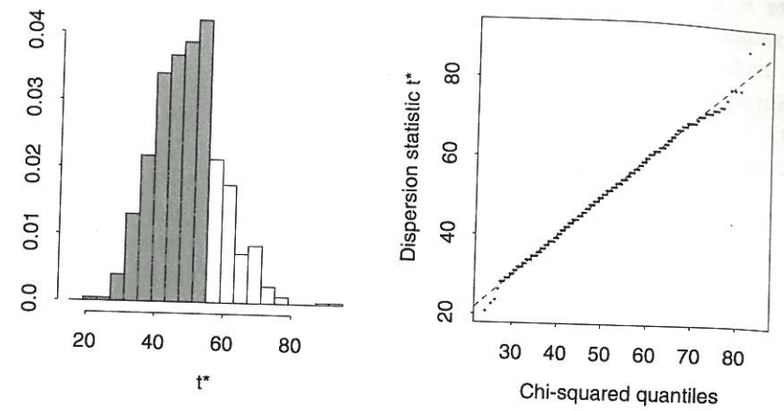

Figure 4.1 Simulation results for dispersion test. Left panel: histogram of $R = 999$ values of the dispersion statistic t^* obtained under multinomial sampling: the data value is $t = 55.15$ and $p_{mc} = 0.235$. Right panel: chi-squared plot of ordered values of t^*, dashed line corresponding to χ^2_{49} approximation to null conditional distribution.

It seems intuitively clear that the sensitivity of the Monte Carlo test increases with R. We shall discuss this issue later, but for now we note that it is advisable to take R to be at least 99.

There are two important aspects of the Monte Carlo test which make it widely useful. The first is that we only need to be able to simulate data under the null hypothesis, this being relatively simple even in some very complicated problems, such as those involving spatial processes (Chapter 8). Secondly, $t, t^*_1, \ldots, t^*_R$ do not need to be independent outcomes: the method remains valid so long as they are *exchangeable* outcomes, which is to say that the joint density of $T, T^*_1, \ldots, T^*_R$ under H_0 is invariant under permutation of its arguments. This allows us to apply Monte Carlo tests to quite complicated problems, as we see next.

4.2.2 Markov chain Monte Carlo tests

In some applications of the exact conditional test, with P-value given by (4.4), the conditional probability calculation is difficult or impossible to do directly. The Monte Carlo test is in principle appropriate here, since the null distribution (given s) does not depend upon unknown parameters. A practical obstacle is that in complicated problems it may be difficult to simulate independent samples directly from that conditional null distribution. However, as we observed before, the Monte Carlo test only requires exchangeable samples. This opens up a new possibility, the use of Markov chain Monte Carlo simulation, in which only the unconditional null distribution is needed.

The basic idea is to represent data $y = (y_1, \ldots, y_n)$ as the result of N steps of a Markov chain with some initial state $x = (x_1, \ldots, x_n)$, and to

generate each y^* by an independent simulation of N steps with the same initial state x. If the Markov chain has equilibrium distribution equal to the null hypothesis distribution of $Y = (Y_1, \ldots, Y_n)$, then y and the R replicates of y^* are exchangeable outcomes under H_0 and (4.11) applies.

Suppose that under H_0 the data have joint density $f_0(y)$ for $y \in \mathscr{B}$, where both f_0 and $\mathscr{B}$ are conditioned on sufficient statistic s if we are dealing with a conditional test. For simplicity suppose that $\mathscr{B}$ has $|\mathscr{B}|$ elements, which we now regard as possible states labelled $(1, 2, \ldots, |\mathscr{B}|)$ of a Markov chain $\{Z_t,\ t = \ldots, -1, 0, 1, \ldots\}$ in discrete time. Consider the data y to be one realization of Z_N. We then have to fix an appropriate value or state for Z_0, and with this initial state simulate the R independent values of Z_N which are the R values of Y^*. The Markov chain is defined so that f_0 is the equilibrium distribution, which can be enforced by appropriate choice of the one-step forward transition probability matrix Q, say, with elements

$$q_{uv} = \Pr(Z_{t+1} = v \mid Z_t = u), \quad u, v \in \mathscr{B}.$$

For the moment suppose that Q is already known.

The first part of the simulation is to produce a value for Z_0. Starting from state y at time N, we simulate N backward steps of the Markov chain using the one-step backward transition probabilities

$$\Pr(Z_t = u \mid Z_{t+1} = v) = f_0(u) q_{uv} / f_0(v). \tag{4.12}$$

Let the final state, the realized value of Z_0, be x. Note that if H_0 is true, so that y was indeed sampled from f_0, then $\Pr(Z_0 = x) = f_0(x)$. In the second part of the simulation, which we repeat independently R times, we simulate N forward steps of the Markov chain, starting in state x and ending up in state $y^* = (y_1^*, \ldots, y_n^*)$. Since under H_0 the chain starts in equilibrium,

$$\Pr(Y^* = y^* \mid H_0) = \Pr(Z_N = y^*) = f_0(y^*).$$

That is, if H_0 is true, then the R replicates $y_1^*, \ldots, y_R^*$ and data y are all sampled from f_0, as we require. Moreover, the R replicates of y^* are jointly exchangeable with the data under H_0. To see this, we have first that

$$f(y, y_1^*, \ldots, y_R^* \mid H_0) = f_0(y) \sum_x \Pr(Z_0 = x \mid Z_N = y) \prod_{r=1}^{R} \Pr(Z_N = y_r^* \mid Z_0 = x),$$

using the independence of the replicate simulations from x. But by the definition of the first part of the simulation, where (4.12) applies,

$$f_0(y) \Pr(Z_0 = x \mid Z_N = y) = f_0(x) \Pr(Z_N = y \mid Z_0 = x),$$

and so

$$f(y, y_1^*, \ldots, y_R^* \mid H_0) = \sum_x f_0(x) \left\{ \Pr(Z_N = y \mid Z_0 = x) \prod_{r=1}^{R} \Pr(Z_N = y_r^* \mid Z_0 = x) \right\},$$

which is a symmetric function of $y, y_1^*, \ldots, y_R^*$ as required. Given that the data vector and simulated data vectors are exchangeable under H_0, the associated test statistic values $(t, t_1^*, \ldots, t_R^*)$ are also exchangeable outcomes under H_0. Therefore (4.11) applies for the P-value calculation.

To complete the description of the method, it remains to define the transition probability matrix Q so that the chain is irreducible with equilibrium distribution $f_0(y)$. There are several ways to do this, all of which use ratios $f_0(v)/f_0(u)$. For example, the Metropolis algorithm starts with a carrier Markov chain on state space $\mathscr{B}$ having any symmetric one-step forward transition probability matrix M, and defines one-step forward transition from state u in the desired Markov chain as follows:

- given we are in state u, select state v with probability m_{uv};
- accept the transition to v with probability $\min\{1, f_0(v)/f_0(u)\}$, otherwise reject it and stay in state u.

It is easy to check that the induced Markov chain has transition probabilities

$$q_{uv} = \min\{1, f_0(v)/f_0(u)\} m_{uv}, \quad u \neq v,$$

and

$$q_{uu} = m_{uu} + \sum_{v \neq u} \max\{0, 1 - f_0(v)/f_0(u)\} m_{uv},$$

and from this it follows that f_0 is indeed the equilibrium distribution of the Markov chain, as required. In applications it is not necessary to calculate the probabilities m_{uv} explicitly, although the symmetry and irreducibility of the carrier chain must be checked. If the matrix M is not symmetric, then the acceptance probability in the Metropolis algorithm must be modified to $\min\left[1, f_0(v) m_{vu} / \{f_0(u) m_{uv}\}\right]$.

The crucial feature of the Markov chain method is that f_0 itself is not needed, only ratios $f_0(v)/f_0(u)$ being involved. This means that for conditional tests, where f_0 is the conditional density for Y given $S = s$, only ratios of the unconditional null density for Y are needed:

$$\frac{f_0(v)}{f_0(u)} = \frac{\Pr(Y = v \mid S = s, H_0)}{\Pr(Y = u \mid S = s, H_0)} = \frac{\Pr(Y = v \mid H_0)}{\Pr(Y = u \mid H_0)}.$$

This greatly simplifies many applications.

The realizations of the Markov chain are symmetrically tied to the artificial starting value x, and this induces a symmetric correlation among $(t, t_1^*, \ldots, t_R^*)$.

This correlation depends upon the particular construction of Q, and reduces to zero at a rate which depends upon Q as m increases. While the correlation does not affect the validity of the P-value calculation, it does affect the power of the test: the higher the correlation, the lower the power.

Example 4.3 (Logistic regression) We return to the problem of Example 4.1, which provides a very simple if artificial illustration. The data y are a binary sequence of length n with s ones, and calculations are to be conditional on $\sum Y_j = s$. Recall that direct Monte Carlo simulation is possible, since all $\binom{n}{s}$ possible data sequences are equally likely under the null hypothesis of constant probability of a unit response.

One simple Markov chain has one-step transitions which select a pair of subscripts i, j at random, and switch y_i and y_j. Clearly the chain is irreducible, since one can progress from any one binary sequence with s ones to any other. All ratios of null probabilities $f_0(v)/f_0(u)$ are equal to one, since all binary sequences with s ones are equally probable. Therefore if we run the Metropolis algorithm, all switches are accepted. But note that this Markov chain, while simple to implement, is inefficient and will require a large number of steps to induce approximate independence of the t_r^*s. The most effective Markov chain would have one-step transitions which are random permutations, and for this only one step would be required. ■

Example 4.4 (AML data) For data such as those in Example 3.9, consider testing the null hypothesis of proportional hazard functions. Denote the failure times by $z_1 < z_2 < \cdots < z_n$, assuming no ties for the moment, and define r_{ij} to be the number in group i who were at risk just prior to z_j. Further, let y_j be 0 or 1 according as the failure at z_j is in group 1 or 2, and denote the hazard function at time z for group i by $h_i(z)$. Then

$$\Pr(Y_j = 1) = \frac{r_{2j}h_2(z_j)}{r_{1j}h_1(z_j) + r_{2j}h_2(z_j)} = \frac{\theta_j}{a_j + \theta_j},$$

where $a_j = r_{1j}/r_{2j}$ and $\theta_j = h_2(z_j)/h_1(z_j)$ for $j = 1, \ldots, n$. The null hypothesis of proportional hazards implies the hypothesis $H_0 : \theta_1 = \cdots = \theta_n$.

For the data of Example 3.9, where $n = 18$, the values of y and a are given in Table 4.2; one tie has been randomly split. Note that censored data contribute only to the rs: the times are not used.

Of course the Y_js are not independent, because a_j depends upon the outcomes of $Y_1, \ldots, Y_{j-1}$. However, for the purposes of illustration here we shall pretend that the a_js are fixed, as well as the survival times and censoring times. That is, we shall treat the Y_js as independent Bernoulli variables with probabilities as given above. Under this pretence the conditional likelihood for

z	5	5	8	8	9	12	13	18	23	23	27	30	31	33	34	43	45	48
r_1	11	11	11	11	11	10	10	8	7	7	6	5	5	4	4	3	3	2
r_2	12	11	10	9	8	8	7	6	6	5	5	4	3	3	2	2	1	0
a	$\frac{11}{12}$	1	$\frac{11}{10}$	$\frac{11}{9}$	$\frac{11}{8}$	$\frac{10}{8}$	$\frac{10}{7}$	$\frac{8}{6}$	$\frac{7}{6}$	$\frac{7}{5}$	$\frac{6}{5}$	$\frac{5}{4}$	$\frac{5}{3}$	$\frac{4}{3}$	2	$\frac{3}{2}$	3	∞
y	1	1	1	1	0	1	0	0	1	0	1	1	0	1	0	1	1	0

Table 4.2 Ingredients of the conditional test for proportional hazards. Failure times as in Table 3.4; at time $z = 23$ the failure in group 2 is taken to occur first.

$\theta_1, \ldots, \theta_{18}$ is simply

$$\prod_{j=1}^{18} \left(\frac{\theta_j}{a_j + \theta_j}\right)^{y_j} \left(\frac{a_j}{a_j + \theta_j}\right)^{1-y_j}.$$

Note that because $a_{18} = \infty$, Y_{18} must be 0 whatever the value of θ_{18}, and so this final response is uninformative. We therefore drop y_{18} from the analysis. Having done this, we see that under H_0 the sufficient statistic for the common hazard ratio θ is $S = \sum_{j=1}^{17} Y_j$, whose observed value is $s = 11$.

Whatever the test statistic T, the exact conditional P-value (4.4) must be approximated. Direct simulation appears impossible, but a simple Markov chain simulation is possible. First, the state space of the chain is $\mathscr{B} = \{x = (x_1, \ldots, x_{17}) : \sum x_j = s\}$, that is all permutations of $y_1, \ldots, y_{17}$. For any two vectors x and $\tilde{x}$ in the state-space, the ratio of null conditional joint probabilities is

$$\frac{p(\tilde{x} \mid s, \theta_1 = \cdots = \theta_{17})}{p(x \mid s, \theta_1 = \cdots = \theta_{17})} = \prod_{j=1}^{17} a_j^{x_j - \tilde{x}_j}.$$

We take the carrier Markov chain to have one-step transitions which are random permutations: this guarantees fast movement over the state space. A step which moves from x to $\tilde{x}$ is then accepted with probability $\min\left(1, \prod_{j=1}^{17} a_j^{x_j - \tilde{x}_j}\right)$. By symmetry the reverse chain is defined in exactly the same way.

The test statistic must be chosen to match the particular alternative hypothesis thought relevant. Here we suppose that the alternative is a monotone ratio of hazards, for which $T = \sum_{j=1}^{17} Y_j \log(Z_j)$ seems to be a reasonable choice. The Markov chain simulation is applied with $N = 100$ steps back to give the initial state x and 100 steps forward to state y^*, the latter repeated $R = 99$ times. Of the resulting t^* values, 48 are less than or equal to the observed value $t = 17.75$, so the P-value is $(1 + 48)/(1 + 99) = 0.49$. Thus there appears to be no evidence against the proportional hazards model.

Average acceptance probability in the Metropolis algorithm is approximately 0.7, and results for $N = 10$ and $N = 1000$ appear indistinguishable from those for $N = 100$. This indicates unusually fast convergence for applications of the Markov chain method. ■

The use of R conditionally independent realizations of the Markov chain is sometimes referred to as the *parallel method*. In contrast is the *series method*, where only one realization is used. Since the successive states of the chain are dependent, a randomization device is needed to induce exchangeability. For details see Problem 4.2.

4.2.3 Parametric bootstrap tests

In many problems of course the distribution of T under H_0 will depend upon nuisance parameters which cannot be conditioned away, so that the Monte Carlo test method does not apply exactly. Then the natural approach is to fit the null model $\hat{F}_0$ and use (4.5) to compute the P-value, i.e. $p = \Pr(T \geq t \mid \hat{F}_0)$. For example, for the parametric model where we are testing $H_0 : \psi = \psi_0$ with λ a nuisance parameter, $\hat{F}_0$ would be the CDF of $f(y \mid \psi_0, \hat{\lambda}_0)$ with $\hat{\lambda}_0$ the maximum likelihood estimator (MLE) of the nuisance parameter when ψ is fixed equal to ψ_0. Calculation of the P-value by (4.5) is referred to as a bootstrap test.

If (4.5) cannot be computed exactly, or if there is no satisfactory approximation (normal or otherwise), then we proceed by simulation. That is, R independent replicate samples $y_1^*, \ldots, y_n^*$ are drawn from $\hat{F}_0$, and for the rth such sample the test statistic value t_r^* is calculated. Then the significance probability (4.5) will be approximated by

$$p_{\text{boot}} = \frac{1 + \#\{t_r^* \geq t\}}{R+1}. \tag{4.13}$$

Ordinarily one would use a simple proportion here, but we have chosen to make the definition match that for the Monte Carlo test in (4.11).

Example 4.5 (Separate families test) Suppose that we wish to choose between the alternative model forms $f_0(y \mid \eta)$ and $f_1(y \mid \zeta)$ for the PDF of the random sample $y_1, \ldots, y_n$. In some circumstances it may make sense to take one model, say f_0, as a null hypothesis, and to test this against the other model as alternative hypothesis. In the notation of Section 4.1, the nuisance parameter is $\lambda = (\eta, \zeta)$ and ψ is the binary indicator of model, with null value $\psi_0 = 0$ and alternative value $\psi_A = 1$. The likelihood ratio statistic (4.7) is equivalent to the more convenient form

$$T = n^{-1} \log \frac{L_1(\hat{\zeta})}{L_0(\hat{\eta})} = n^{-1} \sum_{j=1}^{n} \log \frac{f_1(y_j \mid \hat{\zeta})}{f_0(y_j \mid \hat{\eta})}, \tag{4.14}$$

where $\hat{\eta}$ and $\hat{\zeta}$ are the MLEs and L_0 and L_1 the likelihoods under f_0 and f_1 respectively. If the two families are strictly separate, then the chi-squared approximation (4.8) does not apply. There is a normal approximation for the

null distribution of T, but this is often quite unreliable except for very large n. The parametric bootstrap provides a more reliable and simple option.

The parametric bootstrap works as follows. We generate R samples of size n by random sampling from the fitted null model $f_0(y \mid \hat{\eta})$. For each sample we calculate estimates $\hat{\eta}^*$ and $\hat{\zeta}^*$ by maximizing the simulated log likelihoods

$$\ell_1^*(\zeta) = \sum \log f_1(y_j^* \mid \zeta), \quad \ell_0^*(\eta) = \sum \log f_0(y_j^* \mid \eta),$$

and compute the simulated log likelihood ratio statistic

$$t^* = n^{-1}\{\ell_1^*(\hat{\zeta}^*) - \ell_0^*(\hat{\eta}^*)\}.$$

Then we calculate p using (4.13).

As a particular illustration, consider the failure-time data in Table 1.2. Two plausible models for this type of data are gamma and lognormal, that is

$$f_0(y \mid \eta) = \frac{\kappa(\kappa y)^{\kappa-1} \exp(-\kappa y/\mu)}{\mu^\kappa \Gamma(\kappa)}, \quad f_1(y \mid \zeta) = (\beta y)^{-1} \phi\left(\frac{\log y - \alpha}{\beta}\right), \quad y > 0.$$

For these data the MLEs of the gamma mean and index are $\hat{\mu} = \bar{y} = 108.083$ and $\hat{\kappa} = 0.707$, the latter being the solution to

$$\log(\kappa) - h(\kappa) = \log(\bar{y}) - \overline{\log y}$$

$\overline{\log y}$ and $s^2_{\log y}$ are the average and sample variance for the $\log y_j$.

with $h(\kappa) = d \log \Gamma(\kappa)/d\kappa$, the digamma function. The MLEs of the mean and variance of the normal distribution for $\log Y$ are $\hat{\alpha} = \overline{\log y} = 3.829$ and $\hat{\beta}^2 = (n-1)s^2_{\log y}/n = 2.339$. The test statistic (4.14) is

$$t = -\hat{\kappa} \log(\hat{\kappa}/\bar{y}) - \hat{\kappa}\hat{\alpha} + \hat{\kappa} + \log \Gamma(\hat{\kappa}) - \tfrac{1}{2} \log(2\pi\hat{\beta}^2) - \tfrac{1}{2},$$

whose value for the data is $t = -0.465$. The left panel of Figure 4.2 shows a histogram of $R = 999$ values of t^* under sampling from the fitted gamma model: of these, 619 are greater than t and so $p = 0.62$.

Note that the histogram has a fairly non-normal shape in this case, suggesting that a normal approximation will not be very accurate. This is true also for the (rather complicated) studentized version Z of T: the right panel of Figure 4.2 shows the normal plot of bootstrap values z^*. The observed value of z is 0.4954, for which the bootstrap P-value is 0.34, somewhat smaller than that computed for t, but not changing the conclusion that there is no evidence to change from a gamma to a lognormal model for these data. There are good general reasons to studentize test statistics; see Section 4.4.1.

It should perhaps be mentioned that significance tests of this kind are not always helpful in distinguishing between models, in the sense that we could find evidence against either both or neither of them. This is especially true with small samples such as we have here. In this case the reverse test shows no evidence against the lognormal model. ■

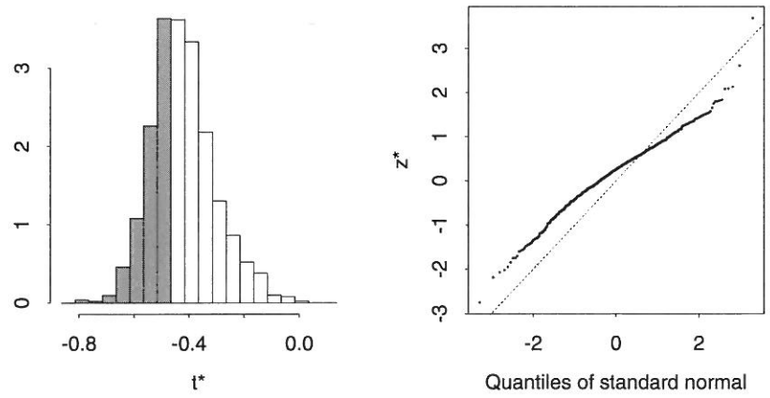

Figure 4.2 Null hypothesis resampling for failure data. Left panel shows histogram of t^* under gamma sampling. Right panel shows normal plot of z^*; $R = 999$ and gamma parameters $\hat{\mu} = 108.0833, \hat{\kappa} = 0.7065$; dotted line is theoretical $N(0,1)$ approximation.

4.2.4 Graphical tests

Graphical methods are popular in model checking: examples include normal and half-normal plots of residuals in regression, plots of Cook distance in regression, plots of nonparametric hazard function estimates, and plots of intensity functions in spatial analysis (Section 8.3). In many cases the nominal shape of the plot is a straight line, which aids the detection of deviation from a null model. Whatever the situation, informed interpretation of the plot requires some notion of its probable variation under the model being checked, unless the sample size is so large that deviation is obvious (c.f. the plot of resampling results in Figure 4.2). The simplest and most common approach is to superimpose a "probable envelope", to which the original data plot is compared. This probable envelope is obtained by Monte Carlo or parametric resampling methods.

Graphical tests are not usually appropriate when a single specific alternative model is of interest. Rather they are used to suggest alternative models, depending upon the manner in which such a plot deviates from its null expected behaviour, or to find suspect data. (Indeed graphical tests are not tests in the usual sense, because there is usually no simple notion of "rejectable" behaviour: we comment more fully on this below.)

Suppose that the graph plots $T(a)$ versus a for $a \in \mathscr{A}$, a bounded set. The observed plot is $\{t(a) : a \in \mathscr{A}\}$. For example, in a normal plot $\mathscr{A}$ is a set of normal quantiles and the values of $t(a)$ are the ordered values of a sample, possibly studentized. The idea of the plot is to compare $t(a)$ with the probable behaviour of $T(a)$ for all $a \in \mathscr{A}$ when H_0 is true.

Example 4.6 (Normal plot) Consider the data in Table 3.1, and suppose in

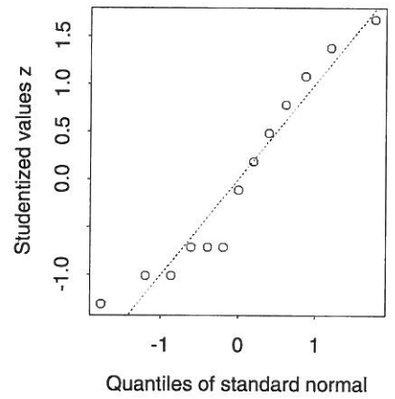

Figure 4.3 Normal plot of $n = 13$ studentized values for final sample in Table 3.1.

particular that we want to assess whether or not the last sample of $n = 13$ measurements can be assumed normal. A normal plot of the data is shown in Figure 4.3, which plots the ordered studentized values $z_{(i)} = (y_{(i)} - \bar{y})/s$ against the quantiles $a_i = \Phi^{-1}(\frac{i}{14})$ of the $N(0,1)$ distribution. In the general notation $\mathscr{A}$ is the set of normal quantiles, and $t(a_i) = z_{(i)}$. The dotted line is the expected pattern, approximately, and the question is whether or not the points deviate sufficiently from this to suggest that the sample is non-normal. ■

Assume for the moment that the null hypothesis joint distribution of $\{T(a) : a \in \mathscr{A}\}$ involves no unknown nuisance parameters. This is true for a normal plot if we use studentized sample values z_i as in the previous example. Then for any fixed a we can subject $t(a)$ to a Monte Carlo test. For each of R independent sets of data $y_1^*, \ldots, y_n^*$, which are obtained by sampling from the null model, we compute the simulated plot

$$t^*(a), \quad a \in \mathscr{A}.$$

Under the null hypothesis, $T(a), T_1^*(a), \ldots, T_R^*(a)$ are independent and identically distributed for any fixed a, so that (4.9) applies with $T = T(a)$. That is,

$$\Pr\left(T(a) < T_{(j)}^*(a) \mid H_0\right) = \frac{j}{R+1}. \tag{4.15}$$

This leads to (4.11) as the one-sided P-value at the given value of a, if large values of $t(a)$ are evidence against the null model. There are obvious