Chapter 9
Model Checking

CHAPTER OUTLINE

The statistical inference methods developed in Chapters 6 through 8 all depend on various assumptions. For example, in Chapter 6 we assumed that the data s were generated from a distribution in the statistical model $\{P_\theta : \theta \in \Omega\}$. In Chapter 7, we also assumed that our uncertainty concerning the true value of the model parameter θ could be described by a prior probability distribution Π. As such, any inferences drawn are of questionable validity if these assumptions do not make sense in a particular application.

In fact, all statistical methodology is based on assumptions or choices made by the statistical analyst, and these must be checked if we want to feel confident that our inferences are relevant. We refer to the process of checking these assumptions as *model checking*, the topic of this chapter. Obviously, this is of enormous importance in applications of statistics, and good statistical practice demands that effective model checking be carried out. Methods range from fairly informal graphical methods to more elaborate hypothesis assessment, and we will discuss a number of these.

9.1 Checking the Sampling Model

Frequency-based inference methods start with a statistical model $\{f_\theta : \theta \in \Omega\}$, for the true distribution that generated the data s. This means we are assuming that the true distribution for the observed data is in this set. If this assumption is not true, then it seems reasonable to question the relevance of any subsequent inferences we make about θ.

Except in relatively rare circumstances, we can never know categorically that a model is correct. The most we can hope for is that we can assess whether or not the observed data s could plausibly have arisen from the model.

If the observed data are surprising for each distribution in the model, then we have evidence that the model is incorrect. This leads us to think in terms of computing a P-value to check the correctness of the model. Of course, in this situation the null hypothesis is that the model is correct; the alternative is that the model could be any of the other possible models for the type of data we are dealing with.

We recall now our discussion of P-values in Chapter 6, where we distinguished between practical significance and statistical significance. It was noted that, while a P-value may indicate that a null hypothesis is false, in practical terms the deviation from the null hypothesis may be so small as to be immaterial for the application. When the sample size gets large, it is inevitable that any reasonable approach via P-values will detect such a deviation and indicate that the null hypothesis is false. This is also true when we are carrying out model checking using P-values. The resolution of this is to estimate, in some fashion, the size of the deviation of the model from correctness, and so determine whether or not the model will be adequate for the application. Even if we ultimately accept the use of the model, it is still valuable to know, however, that we have detected evidence of model incorrectness when this is the case.

One P-value approach to model checking entails specifying a *discrepancy statistic* $D : S \to R^1$ that measures deviations from the model under consideration. Typically, large values of D are meant to indicate that a deviation has occurred. The actual value $D(s)$ is, of course, not necessarily an indication of this. The relevant issue is whether or not the observed value $D(s)$ is surprising under the assumption that the model is correct. Therefore, we must assess whether or not $D(s)$ lies in a region of low probability for the distribution of this quantity when the model is correct. For example, consider the density of a potential D statistic plotted in Figure 9.1.1. Here a value $D(s)$ in the left tail (near 0), right tail (out past 15), or between the two modes (in the interval from about 7 to 9) all would indicate that the model is incorrect, because such values have a low probability of occurrence when the model is correct.

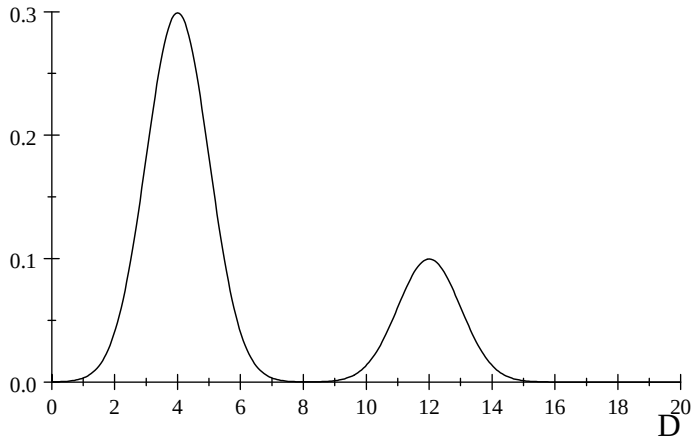

Figure 9.1.1: Plot of a density for a discrepancy statistic D.

The above discussion places the restriction that, when the model is correct, D must have a single distribution, i.e., the distribution cannot depend on θ. For many commonly used discrepancy statistics, this distribution is unimodal. A value in the right tail then indicates a lack of fit, or *underfitting*, by the model (the discrepancies are unnaturally large); a value in the left tail then indicates *overfitting* by the model (the discrepancies are unnaturally small).

There are two general methods available for obtaining a single distribution for the computation of P-values. One method requires that D be ancillary.

Definition 9.1.1 A statistic D whose distribution under the model does not depend upon θ is called *ancillary*, i.e., if $s \sim P_\theta$, then $D(s)$ has the same distribution for every $\theta \in \Omega$.

If D is ancillary, then it has a single distribution specified by the model. If $D(s)$ is a surprising value for this distribution, then we have evidence against the model being true.

It is not the case that any ancillary D will serve as a useful discrepancy statistic. For example, if D is a constant, then it is ancillary, but it is obviously not useful for model checking. So we have to be careful in choosing D.

Quite often we can find useful ancillary statistics for a model by looking at *residuals*. Loosely speaking, residuals are based on the information in the data that is left over after we have fit the model. If we have used all the relevant information in the data for fitting, then the residuals should contain no useful information for inference about the parameter θ. Example 9.1.1 will illustrate more clearly what we mean by residuals. Residuals play a major role in model checking.

The second method works with any discrepancy statistic D. For this, we use the conditional distribution of D, given the value of a sufficient statistic T. By Theorem 8.1.2, this conditional distribution is the same for every value of θ. If $D(s)$ is a surprising value for this distribution, then we have evidence against the model being true.

Sometimes the two approaches we have just described agree, but not always. Consider some examples.

EXAMPLE 9.1.1 *Location Normal*
Suppose we assume that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where $\mu \in R^1$ is unknown and σ_0^2 is known. We know that $\bar{x}$ is a minimal sufficient statistic for this problem (see Example 6.1.7). Also, $\bar{x}$ represents the fitting of the model to the data, as it is the estimate of the unknown parameter value μ.

Now consider

$$r = r\,(x_1, \ldots, x_n) = (r_1, \ldots, r_n) = (x_1 - \bar{x}, \ldots, x_n - \bar{x})$$

as one possible definition of the residual. Note that we can reconstruct the original data from the values of $\bar{x}$ and r.

It turns out that $R = \left(X_1 - \bar{X}, \ldots, X_n - \bar{X}\right)$ has a distribution that is independent of μ, with $E(R_i) = 0$ and $\mathrm{Cov}(R_i, R_j) = \sigma_0^2(\delta_{ij} - 1/n)$ for every i, j ($\delta_{ij} = 1$ when $i = j$ and 0 otherwise). Moreover, R is independent of $\bar{X}$ and $R_i \sim N(0, \sigma_0^2\,(1 - 1/n))$ (see Problems 9.1.19 and 9.1.20).

Accordingly, we have that r is ancillary and so is any discrepancy statistic D that depends on the data only through r. Furthermore, the conditional distribution of $D(R)$ given $\bar{X} = \bar{x}$ is the same as the marginal distribution of $D(R)$ because they are independent. Therefore, the two approaches to obtaining a P-value agree here, whenever the discrepancy statistic depends on the data only through r.

By Theorem 4.6.6, we have that

$$D(R) = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} R_i^2 = \frac{1}{\sigma_0^2} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

is distributed $\chi^2\,(n-1)$, so this is a possible discrepancy statistic. Therefore, the P-value

$$P(D > D(r)), \tag{9.1.1}$$

where $D \sim \chi^2\,(n-1)$, provides an assessment of whether or not the model is correct.

Note that values of (9.1.1) near 0 or near 1 are both evidence against the model, as both indicate that $D(r)$ is in a region of low probability when assuming the model is correct. A value near 0 indicates that $D(r)$ is in the right tail, whereas a value near 1 indicates that $D(r)$ is in the left tail.

The necessity of examining the left tail of the distribution of $D(r)$, as well as the right, is seen as follows. Consider the situation where we are in fact sampling from an $N(\mu, \sigma^2)$ distribution where σ^2 is much smaller than σ_0^2. In this case, we expect $D(r)$ to be a value in the left tail, because $E\,(D(R)) = (n-1)\sigma^2/\sigma_0^2$.

There are obviously many other choices that could be made for the D statistic. At present, there is not a theory that prescribes one choice over another. One caution should be noted, however. The choice of a statistic D cannot be based upon looking at the data first. Doing so invalidates the computation of the P-value as described above, as then we must condition on the data feature that led us to choose that particular D. ∎

EXAMPLE 9.1.2 *Location-Scale Normal*

Suppose we assume that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution, where $\left(\mu, \sigma^2\right) \in R^1 \times (0, \infty)$ is unknown. We know that $\left(\bar{x}, s^2\right)$ is a minimal sufficient statistic for this model (Example 6.1.8). Consider

$$r = r\,(x_1, \ldots, x_n) = (r_1, \ldots, r_n) = \left(\frac{x_1 - \bar{x}}{s}, \ldots, \frac{x_n - \bar{x}}{s}\right)$$

as one possible definition of the residual. Note that we can reconstruct the data from the values of $(\bar{x}, s^2)$ and r.

It turns out R has a distribution that is independent of (μ, σ^2) (and hence is ancillary — see Challenge 9.1.28) as well as independent of $(\bar{X}, S^2)$. So again, the two approaches to obtaining a P-value agree here, as long as the discrepancy statistic depends on the data only through r.

One possible discrepancy statistic is given by

$$D(r) = -\frac{1}{n} \sum_{i=1}^{n} \ln\left(\frac{r_i^2}{n-1}\right).$$

To use this statistic for model checking, we need to obtain its distribution when the model is correct. Then we compare the observed value $D(r)$ with this distribution, to see if it is surprising.

We can do this via simulation. Because the distribution of $D(R)$ is independent of (μ, σ^2), we can generate N samples of size n from the $N(0, 1)$ distribution (or any other normal distribution) and calculate $D(R)$ for each sample. Then we look at histograms of the simulated values to see if $D(r)$, from the original sample, is a surprising value, i.e., if it lies in a region of low probability like a left or right tail.

For example, suppose we observed the sample

-2.08	-0.28	2.01	-1.37	40.08

obtaining the value $D(r) = 4.93$. Then, simulating 10^4 values from the distribution of D, under the assumption of model correctness, we obtained the density histogram given in Figure 9.1.2. See Appendix B for some code used to carry out this simulation. The value $D(r) = 4.93$ is out in the right tail and thus indicates that the sample is not from a normal distribution. In fact, only 0.0057 of the simulated values are larger, so this is definite evidence against the model being correct.

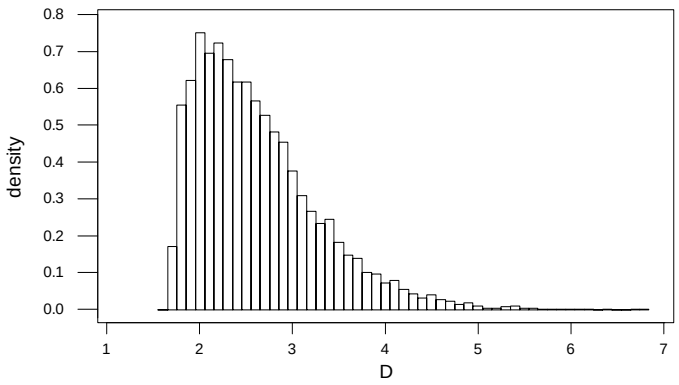

Figure 9.1.2: A density histogram for a simulation of 10^4 values of D in Example 9.1.2.

Obviously, there are other possible functions of r that we could use for model checking here. In particular, $D_{\text{skew}}(r) = (n-1)^{-3/2} \sum_{i=1}^{n} r_i^3$, the *skewness statistic*, and $D_{\text{kurtosis}}(r) = (n-1)^{-2} \sum_{i=1}^{n} r_i^4$, the *kurtosis statistic*, are commonly used. The skewness statistic measures the symmetry in the data, while the kurtosis statistic measures the "peakedness" in the data. As just described, we can simulate the distribution of these statistics under the normality assumption and then compare the observed values with these distributions to see if we have any evidence against the model (see Computer Problem 9.1.26). ∎

The following examples present contexts in which the two approaches to computing a P-value for model checking are not the same.

EXAMPLE 9.1.3 *Location-Scale Cauchy*
Suppose we assume that $(x_1, \ldots, x_n)$ is a sample from the distribution given by $\mu + \sigma Z$, where $Z \sim t(1)$ and $(\mu, \sigma^2) \in R^1 \times (0, \infty)$ is unknown. This time, $(\bar{x}, s^2)$ is not a minimal sufficient statistic, but the statistic r defined in Example 9.1.2 is still ancillary (Challenge 9.1.28). We can again simulate values from the distribution of R (just generate samples from the $t(1)$ distribution and compute r for each sample) to estimate P-values for any discrepancy statistic such as the $D(r)$ statistics discussed in Example 9.1.2. ∎

EXAMPLE 9.1.4 *Fisher's Exact Test*
Suppose we take a sample of n from a population of students and observe the values $(a_1, b_1), \ldots, (a_n, b_n)$, where a_i is gender ($A = 1$ indicating male, $A = 2$ indicating female) and b_i is a categorical variable for part-time employment status ($B = 1$ indicating employed, $B = 2$ indicating unemployed). So each individual is being categorized into one of four categories, namely,

$$\begin{aligned}
&\text{Category 1, when } A = 1,\ B = 1,\\
&\text{Category 2, when } A = 1,\ B = 2,\\
&\text{Category 3, when } A = 2,\ B = 1,\\
&\text{Category 4, when } A = 2,\ B = 2.
\end{aligned}$$

Suppose our model for this situation is that A and B are independent with $P(A = 1) = \alpha_1$, $P(B = 1) = \beta_1$ where $\alpha_1 \in [0, 1]$ and $\beta_1 \in [0, 1]$ are completely unknown. Then letting X_{ij} denote the count for the category, where $A = i$, $B = j$, Example 2.8.5 gives that

$$(X_{11}, X_{12}, X_{21}, X_{22}) \sim \text{Multinomial}(n, \alpha_1\beta_1, \alpha_1\beta_2, \alpha_2\beta_1, \alpha_2\beta_2).$$

As we will see in Chapter 10, this model is equivalent to saying that there is no relationship between gender and employment status.

Denoting the observed cell counts by $(x_{11}, x_{12}, x_{21}, x_{22})$, the likelihood function is given by

$$\begin{aligned}
&\left(\alpha_1\beta_1\right)^{x_{11}} \left(\alpha_1\beta_2\right)^{x_{12}} \left(\alpha_2\beta_1\right)^{x_{21}} \left(\alpha_2\beta_2\right)^{x_{22}}\\
&= \alpha_1^{x_{11}+x_{12}} \left(1-\alpha_1\right)^{n-x_{11}-x_{12}} \beta_1^{x_{11}+x_{21}} \left(1-\beta_1\right)^{n-x_{11}-x_{21}}\\
&= \alpha_1^{x_{1\cdot}} \left(1-\alpha_1\right)^{n-x_{1\cdot}} \beta_1^{x_{\cdot 1}} \left(1-\beta_1\right)^{n-x_{\cdot 1}},
\end{aligned}$$

where $(x_{1\cdot}, x_{\cdot 1}) = (x_{11} + x_{12}, x_{11} + x_{21})$. Therefore, the MLE (Problem 9.1.14) is given by

$$\left(\hat{\alpha}_1, \hat{\beta}_1\right) = \left(\frac{x_{1\cdot}}{n}, \frac{x_{\cdot 1}}{n}\right).$$

Note that $\hat{\alpha}_1$ is the proportion of males in the sample and $\hat{\beta}_1$ is the proportion of all employed in the sample. Because $(x_{1\cdot}, x_{\cdot 1})$ determines the likelihood function and can be calculated from the likelihood function, we have that $(x_{1\cdot}, x_{\cdot 1})$ is a minimal sufficient statistic.

In this example, a natural definition of residual does not seem readily apparent. So we consider looking at the conditional distribution of the data, given the minimal sufficient statistic. The conditional distribution of the sample $(A_1, B_1), \ldots, (A_n, B_n)$, given the values $(x_{1\cdot}, x_{\cdot 1})$, is the uniform distribution on the set of all samples where the restrictions

$$\begin{aligned} x_{11} + x_{12} &= x_{1\cdot}, \\ x_{11} + x_{21} &= x_{\cdot 1}, \\ x_{11} + x_{12} + x_{21} + x_{22} &= n \end{aligned} \tag{9.1.2}$$

are satisfied. Notice that, given $(x_{1\cdot}, x_{\cdot 1})$, all the other values in (9.1.2) are determined when we specify a value for x_{11}.

It can be shown that the number of such samples is equal to (see Problem 9.1.21)

$$\binom{n}{x_{1\cdot}}\binom{n}{x_{\cdot 1}}.$$

Now the number of samples with prescribed values for $x_{1\cdot}$, $x_{\cdot 1}$, and $x_{11} = i$ is given by

$$\binom{n}{x_{1\cdot}}\binom{x_{1\cdot}}{i}\binom{n - x_{1\cdot}}{x_{\cdot 1} - i}.$$

Therefore, the conditional probability function of x_{11}, given $(x_{1\cdot}, x_{\cdot 1})$, is

$$P(x_{11} = i \mid x_{1\cdot}, x_{\cdot 1}) = \frac{\binom{n}{x_{1\cdot}}\binom{x_{1\cdot}}{i}\binom{n-x_{1\cdot}}{x_{\cdot 1}-i}}{\binom{n}{x_{1\cdot}}\binom{n}{x_{\cdot 1}}} = \frac{\binom{x_{1\cdot}}{i}\binom{n-x_{1\cdot}}{x_{\cdot 1}-i}}{\binom{n}{x_{\cdot 1}}}.$$

This is the Hypergeometric$(n, x_{\cdot 1}, x_{1\cdot})$ probability function.

So we have evidence against the model holding whenever x_{11} is out in the tails of this distribution. Assessing this requires a tabulation of this distribution or the use of a statistical package with the hypergeometric distribution function built in.

As a simple numerical example, suppose that we took a sample of $n = 20$ students, obtaining $x_{\cdot 1} = 12$ unemployed, $x_{1\cdot} = 6$ males, and $x_{11} = 2$ employed males. Then the Hypergeometric(20, 12, 6) probability function is given by the following table.

i	0	1	2	3	4	5	6
$p(i)$	0.001	0.017	0.119	0.318	0.358	0.163	0.024

The probability of getting a value as far, or farther, out in the tails than $x_{11} = 2$ is equal to the probability of observing a value of x_{11} with probability of occurrence as small as or smaller, than $x_{11} = 2$. This P-value equals

$$(0.119 + 0.017 + 0.001) + 0.024 = 0.161.$$

Therefore, we have no evidence against the model of independence between A and B. Of course, the sample size is quite small here.

There is another approach here to testing the independence of A and B. In particular, we could only assume the independence of the initial unclassified sample, and then we always have

$$(X_{11}, X_{12}, X_{21}, X_{22}) \sim \text{Multinomial}(n, \alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}),$$

where the α_{ij} comprise an unknown probability distribution. Given this model, we could then test for the independence of A and B. We will discuss this in Section 10.2. ∎

Another approach to model checking proceeds as follows. We enlarge the model to include more distributions and then test the null hypothesis that the true model is the submodel we initially started with. If we can apply the methods of Section 8.2 to come up with a uniformly most powerful (UMP) test of this null hypothesis, then we will have a check of departures from the model of interest — at least as expressed by the possible alternatives in the enlarged model. If the model passes such a check, however, we are still required to check the validity of the enlarged model. This can be viewed as a technique for generating relevant discrepancy statistics D.

9.1.1 Residual and Probability Plots

There is another, more informal approach to checking model correctness that is often used when we have residuals available. These methods involve various plots of the residuals that should exhibit specific characteristics if the model is correct. While this approach lacks the rigor of the P-value approach, it is good at demonstrating gross deviations from model assumptions. We illustrate this via some examples.

EXAMPLE 9.1.5 *Location and Location-Scale Normal Models*
Using the residuals for the location normal model discussed in Example 9.1.1, we have that $E(R_i) = 0$ and $\text{Var}(R_i) = \sigma_0^2(1 - 1/n)$. We standardize these values so that they also have variance 1, and so obtain the standardized residuals $(r_1^*, \ldots, r_n^*)$ given by

$$r_i^* = \sqrt{\frac{n}{\sigma_0^2\,(n-1)}}(x_i - \bar{x}). \tag{9.1.3}$$

The standardized residuals are distributed $N(0, 1)$, and, assuming that n is reasonably large, it can be shown that they are approximately independent. Accordingly, we can think of $r_1^*, \ldots, r_n^*$ as an approximate sample from the $N(0, 1)$ distribution.

Therefore, a plot of the points $\left(i, r_i^*\right)$ should not exhibit any discernible pattern. Furthermore, all the values in the y-direction should lie in $(-3, 3)$, unless of course n is very large, in which case we might expect a few values outside this interval. A discernible pattern, or several extreme values, can be taken as some evidence that the model assumption is not correct. Always keep in mind, however, that any observed pattern could have arisen simply from sampling variability when the true model is correct. Simulating a few of these residual plots (just generating several samples of n from the $N(0, 1)$ distribution and obtaining a residual plot for each sample) will give us some idea of whether or not the observed pattern is unusual.

Figure 9.1.3 shows a plot of the standardized residuals (9.1.3) for a sample of 100 from the $N(0,1)$ distribution. Figure 9.1.4 shows a plot of the standardized residuals for a sample of 100 from the distribution given by $3^{-1/2}Z$, where $Z \sim t\,(3)$. Note that a $t\,(3)$ distribution has mean 0 and variance equal to 3, so $\text{Var}(3^{-1/2}Z) = 1$ (Problem 4.6.16). Figure 9.1.5 shows the standardized residuals for a sample of 100 from an Exponential(1) distribution.

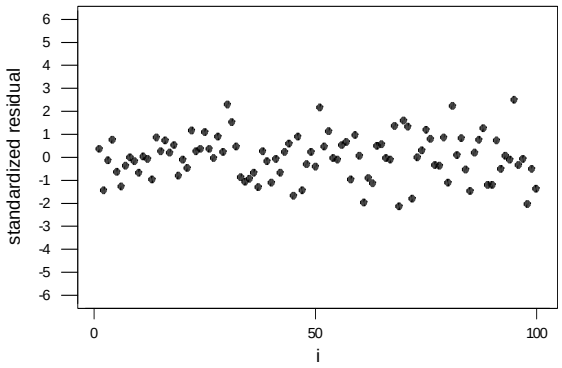

Figure 9.1.3: A plot of the standardized residuals for a sample of 100 from an $N(0,1)$ distribution.

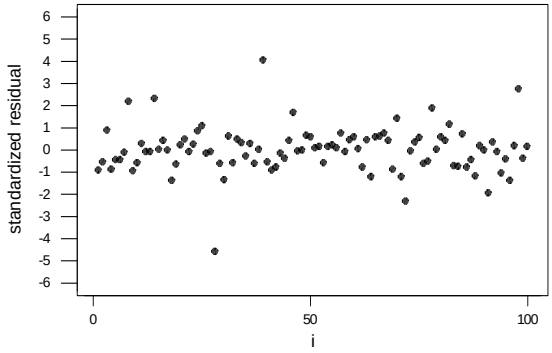

Figure 9.1.4: A plot of the standardized residuals for a sample of 100 from $X = 3^{-1/2}Z$ where $Z \sim \text{t}(3)$.

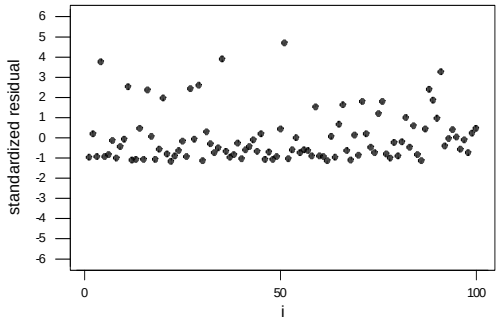

Figure 9.1.5: A plot of the standardized residuals for a sample of 100 from an Exponential(1) distribution.

Note that the distributions of the standardized residuals for all these samples have mean 0 and variance equal to 1. The difference in Figures 9.1.3 and 9.1.4 is due to the fact that the t distribution has much longer tails. This is reflected in the fact that a few of the standardized residuals are outside $(-3, 3)$ in Figure 9.1.4 but not in Figure 9.1.3. Even though the two distributions are quite different — e.g., the $N(0, 1)$ distribution has all of its moments whereas the $3^{-1/2}\, t\,(3)$ distribution has only two moments — the plots of the standardized residuals are otherwise very similar. The difference in Figures 9.1.3 and 9.1.5 is due to the asymmetry in the Exponential(1) distribution, as it is skewed to the right.

Using the residuals for the location-scale normal model discussed in Example 9.1.2, we define the standardized residuals $r_1^*, \ldots, r_n^*$ by

$$r_i^* = \sqrt{\frac{n}{s^2\,(n-1)}}(x_i - \bar{x}). \tag{9.1.4}$$

Here, the unknown variance is estimated by s^2. Again, it can be shown that when n is large, then $\left(r_1^*, \ldots, r_n^*\right)$ is an approximate sample from the $N(0, 1)$ distribution. So we plot the values $\left(i, r_i^*\right)$ and interpret the plot just as we described for the location normal model. ∎

It is very common in statistical applications to assume some basic form for the distribution of the data, e.g., we might assume we are sampling from a normal distribution with some mean and variance. To assess such an assumption, the use of a *probability plot* has proven to be very useful.

To illustrate, suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma^2)$ distribution. Then it can be shown that when n is large, the expectation of the i-th order statistic satisfies

$$E(X_{(i)}) \approx \mu + \sigma\,\Phi^{-1}(i/(n+1)). \tag{9.1.5}$$

If the data value x_j corresponds to order statistic $x_{(i)}$ (i.e., $x_{(i)} = x_j$), then we call $\Phi^{-1}\,(i/\,(n+1))$ the *normal score* of x_j in the sample. Then (9.1.5) indicates that if

we plot the points $(x_{(i)}, \Phi^{-1}(i/(n+1)))$, these should lie approximately on a line with intercept μ and slope σ. We call such a plot a *normal probability plot* or *normal quantile plot*. Similar plots can be obtained for other distributions.

EXAMPLE 9.1.6 *Location-Scale Normal*
Suppose we want to assess whether or not the following data set can be considered a sample of size $n = 10$ from some normal distribution.

2.00	0.28	0.47	3.33	1.66	8.17	1.18	4.15	6.43	1.77

The order statistics and associated normal scores for this sample are given in the following table.

i	1	2	3	4	5
$x_{(i)}$	0.28	0.47	1.18	1.66	1.77
$\Phi^{-1}(i/(n+1))$	-1.34	-0.91	-0.61	-0.35	-0.12
i	6	7	8	9	10
$x_{(i)}$	2.00	3.33	4.15	6.43	8.17
$\Phi^{-1}(i/(n+1))$	0.11	0.34	0.60	0.90	1.33

The values

$$(x_{(i)}, \Phi^{-1}(i/(n+1)))$$

are then plotted in Figure 9.1.6. There is some definite deviation from a straight line here, but note that it is difficult to tell whether this is unexpected in a sample of this size from a normal distribution. Again, simulating a few samples of the same size (say, from an $N(0, 1)$ distribution) and looking at their normal probability plots is recommended. In this case, we conclude that the plot in Figure 9.1.6 looks reasonable. ∎

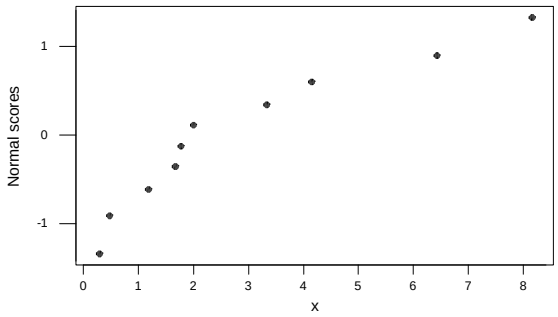

Figure 9.1.6: Normal probability plot of the data in Example 9.1.6.

We will see in Chapter 10 that the use of normal probability plots of standardized residuals is an important part of model checking for more complicated models. So, while they are not really needed here, we consider some of the characteristics of such plots when assessing whether or not a sample is from a location normal or location-scale normal model.

Assume that n is large so that we can consider the standardized residuals, given by (9.1.3) or (9.1.4) as an approximate sample from the $N(0, 1)$ distribution. Then a normal probability plot of the standardized residuals should be approximately linear, with y-intercept approximately equal to 0 and slope approximately equal to 1. If we get a substantial deviation from this, then we have evidence that the assumed model is incorrect.

In Figure 9.1.7, we have plotted a normal probability plot of the standardized residuals for a sample of $n = 25$ from an $N(0, 1)$ distribution. In Figure 9.1.8, we have plotted a normal probability plot of the standardized residuals for a sample of $n = 25$ from the distribution given by $X = 3^{-1/2}Z$, where $Z \sim t(3)$. Both distributions have mean 0 and variance 1, so the difference in the normal probability plots is due to other distributional differences.

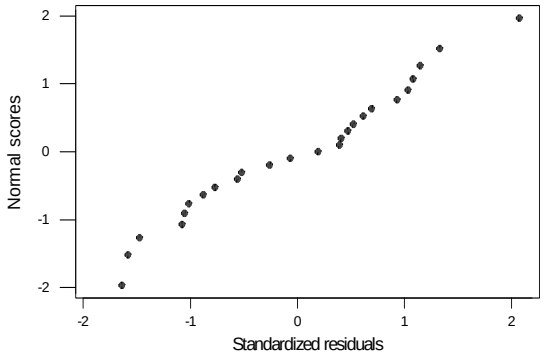

Figure 9.1.7: Normal probability plot of the standardized residuals of a sample of 25 from an $N(0, 1)$ distribution.

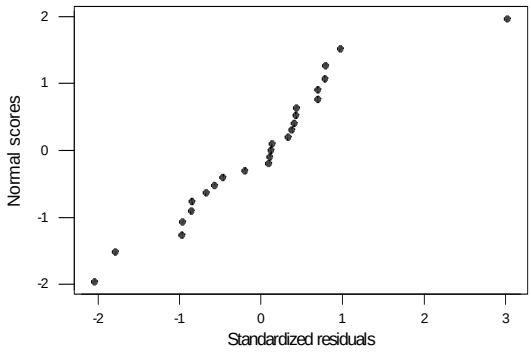

Figure 9.1.8: Normal probability plot of the standardized residuals of a sample of 25 from $X = 3^{-1/2}Z$ where $Z \sim \mathrm{t}(3)$.

9.1.2 The Chi-Squared Goodness of Fit Test

The chi-squared goodness of fit test has an important historical place in any discussion of assessing model correctness. We use this test to assess whether or not a categorical random variable W, which takes its values in the finite sample space $\{1, 2, \ldots, k\}$, has a specified probability measure P, after having observed a sample $(\omega_1, \ldots, \omega_n)$. When we have a random variable that is discrete and takes infinitely many values, then we partition the possible values into k categories and let W simply indicate which category has occurred. If we have a random variable that is quantitative, then we partition R^1 into k subintervals and let W indicate in which interval the response occurred. In effect, we want to check whether or not a specific probability model, as given by P, is correct for W based on an observed sample.

Let $(X_1, \ldots, X_k)$ be the observed counts or frequencies of $1, \ldots, k$, respectively. If P is correct, then, from Example 2.8.5,

$$(X_1, \ldots, X_k) \sim \text{Multinomial}(n, p_1, \ldots, p_k)$$

where $p_i = P(\{i\})$. This implies that $E(X_i) = np_i$ and $\text{Var}(X_i) = np_i(1 - p_i)$ (recall that $X_i \sim \text{Binomial}(n, p_i)$). From this, we deduce that

$$R_i = \frac{X_i - np_i}{\sqrt{np_i(1 - p_i)}} \xrightarrow{D} N(0, 1) \tag{9.1.6}$$

as $n \to \infty$ (see Example 4.4.9).

For finite n, the distribution of R_i, when the model is correct, is dependent on P, but the limiting distribution is not. Thus we can think of the R_i as standardized residuals when n is large. Therefore, it would seem that a reasonable discrepancy statistic is given by the sum of the squares of the standardized residuals with $\sum_{i=1}^{k} R_i^2$ approximately distributed $\chi^2(k)$. The restriction $x_1 + \cdots + x_k = n$ holds, however, so the R_i are not independent and the limiting distribution is not $\chi^2(k)$. We do, however, have the following result, which provides a similar discrepancy statistic.

Theorem 9.1.1 If $(X_1, \ldots, X_k) \sim \text{Multinomial}(n, p_1, \ldots, p_k)$, then

$$X^2 = \sum_{i=1}^{k} (1 - p_i) R_i^2 = \sum_{i=1}^{k} \frac{(X_i - np_i)^2}{np_i} \xrightarrow{D} \chi^2(k - 1)$$

as $n \to \infty$.

The proof of this result is a little too involved for this text, so see, for example, Theorem 17.2 of *Asymptotic Statistics* by A. W. van der Vaart (Cambridge University Press, Cambridge, 1998), which we will use here.

We refer to X^2 as the *chi-squared statistic*. The process of assessing the correctness of the model by computing the P-value $P(X^2 \geq X_0^2)$, where $X^2 \sim \chi^2(k - 1)$ and X_0^2 is the observed value of the chi-squared statistic, is referred to as the *chi-squared goodness of fit test*. Small P-values near 0 provide evidence of the incorrectness of the probability model. Small P-values indicate that some of the residuals are too large.

Note that the ith term of the chi-squared statistic can be written as

$$\frac{(X_i - np_i)^2}{np_i} = \frac{(\text{number in the } i\text{th cell} - \text{expected number in the } i\text{th cell})^2}{\text{expected number in the } i\text{th cell}}.$$

It is recommended, for example, in *Statistical Methods*, by G. Snedecor and W. Cochran (Iowa State Press, 6th ed., Ames, 1967) that grouping (combining cells) be employed to ensure that $E(X_i) = np_i \geq 1$ for every i, as simulations have shown that this improves the accuracy of the approximation to the P-value.

We consider an important application.

EXAMPLE 9.1.7 *Testing the Accuracy of a Random Number Generator*
In effect, every Monte Carlo simulation can be considered to be a set of mathematical operations applied to a stream of numbers $U_1, U_2, \ldots$ in $[0, 1]$ that are supposed to be i.i.d. Uniform$[0, 1]$. Of course, they cannot satisfy this requirement exactly because they are generated according to some deterministic function. Typically, a function $f : [0, 1]^m \to [0, 1]$ is chosen and is applied iteratively to obtain the sequence. So we select $U_1, \ldots, U_m$ as initial *seed values* and then $U_{m+1} = f(U_1, \ldots, U_m)$, $U_{m+2} = f(U_2, \ldots, U_{m+1})$, etc. There are many possibilities for f, and a great deal of research and study have gone into selecting functions that will produce sequences that adequately mimic the properties of an i.i.d. Uniform$[0, 1]$ sequence.

Of course, it is always possible that the underlying f used in a particular statistical package or other piece of software is very poor. In such a case, the results of the simulations can be grossly in error. How do we assess whether a particular f is good or not? One approach is to run a battery of statistical tests to see whether the sequence is behaving as we know an ideal sequence would.

For example, if the sequence $U_1, U_2, \ldots$ is i.i.d. Uniform$[0, 1]$, then

$$\lceil 10U_1 \rceil, \lceil 10U_2 \rceil, \ldots$$

is i.i.d. Uniform$\{1, 2, \ldots, 10\}$ ($\lceil x \rceil$ denotes the smallest integer greater than x, e.g., $\lceil 3.2 \rceil = 4$). So we can test the adequacy of the underlying function f by generating $U_1, \ldots, U_n$ for large n, putting $x_i = \lceil 10U_i \rceil$, and then carrying out a chi-squared goodness of fit test with the 10 categories $\{1, \ldots, 10\}$ with each cell probability equal to 1/10.

Doing this using a popular statistical package (with $n = 10^4$) gave the following table of counts x_i and standardized residuals r_i as specified in (9.1.6).

i	x_i	r_i
1	993	−0.23333
2	1044	1.46667
3	1061	2.03333
4	1021	0.70000
5	1017	0.56667
6	973	−0.90000
7	975	−0.83333
8	965	−1.16667
9	996	−0.13333
10	955	−1.50000

All the standardized residuals look reasonable as possible values from an $N(0, 1)$ distribution. Furthermore,

$$\begin{aligned} X_0^2 &= (1-0.1)\left\{\begin{array}{l}(-0.23333)^2+(1.46667)^2+(2.03333)^2\\ +(0.70000)^2+(0.56667)^2+(-0.90000)^2\\ +(-0.83333)^2+(-1.16667)^2+(-0.13333)^2\\ +(-1.50000)^2\end{array}\right\} \\ &= 11.0560 \end{aligned}$$

gives the P-value $P(X^2 \geq 11.0560) = 0.27190$ when $X^2 \sim \chi^2(9)$. This indicates that we have no evidence that the random number generator is defective.

Of course, the story does not end with a single test like this. Many other features of the sequence should be tested. For example, we might want to investigate the independence properties of the sequence and so test if each possible combination of (i, j) occurs with probability 1/100, etc. ∎

More generally, we will not have a prescribed probability distribution P for X but rather a statistical model $\{P_\theta : \theta \in \Omega\}$, where each P_θ is a probability measure on the finite set $\{1, 2, \ldots, k\}$. Then, based on the sample from the model, we have that

$$(X_1, \ldots, X_k) \sim \text{Multinomial}(n, p_1(\theta), \ldots, p_k(\theta))$$

where $p_i(\theta) = P_\theta(\{i\})$.

Perhaps a natural way to assess whether or not this model fits the data is to find the MLE $\hat{\theta}$ from the likelihood function

$$L(\theta \mid x_1, \ldots, x_k) = (p_1(\theta))^{x_1} \cdots (p_k(\theta))^{x_k}$$

and then look at the standardized residuals

$$r_i(\hat{\theta}) = \frac{x_i - np_i(\hat{\theta})}{\sqrt{np_i(\hat{\theta})(1-p_i(\hat{\theta}))}}.$$

We have the following result, which we state without proof.

Theorem 9.1.2 Under conditions (similar to those discussed in Section 6.5), we have that $R_i(\hat{\theta}) \xrightarrow{D} N(0, 1)$ and

$$X^2 = \sum_{i=1}^{k}(1-p_i(\hat{\theta}))R_i^2(\hat{\theta}) = \sum_{i=1}^{k}\frac{(X_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \xrightarrow{D} \chi^2\,(k-1-\dim\Omega)$$

as $n \to \infty$.

By $\dim\Omega$, we mean the dimension of the set Ω. Loosely speaking, this is the minimum number of coordinates required to specify a point in the set, e.g., a line requires one coordinate (positive or negative distance from a fixed point), a circle requires one coordinate, a plane in R^3 requires two coordinates, etc. Of course, this result implies that the number of cells must satisfy $k > 1 + \dim\Omega$.

Consider an example.

EXAMPLE 9.1.8 *Testing for Exponentiality*
Suppose that a sample of lifelengths of light bulbs (measured in thousands of hours) is supposed to be from an Exponential(θ) distribution, where $\theta \in \Omega = (0, \infty)$ is unknown. So here $\dim \Omega = 1$, and we require at least two cells for the chi-squared test. The manufacturer expects that most bulbs will last at least 1000 hours, 50% will last less than 2000 hours, and most will have failed by 3000 hours. So based on this, we partition the sample space as

$$(0, \infty) = (0, 1] \cup (1, 2] \cup (2, 3] \cup (3, \infty).$$

Suppose that a sample of $n = 30$ light bulbs was taken and that the counts $x_1 = 5$, $x_2 = 16$, $x_3 = 8$, and $x_4 = 1$ were obtained for the four intervals, respectively. Then the likelihood function based on these counts is given by

$$L(\theta \mid x_1, \ldots, x_4 0 = (1 - e^{-\theta})^5 (e^{-\theta} - e^{-2\theta})^{16} (e^{-2\theta} - e^{-3\theta})^8 (e^{-3\theta})^1,$$

because, for example, the probability of a value falling in (1, 2] is $e^{-\theta} - e^{-2\theta}$, and we have $x_2 = 16$ observations in this interval. Figure 9.1.9 is a plot of the log-likelihood.

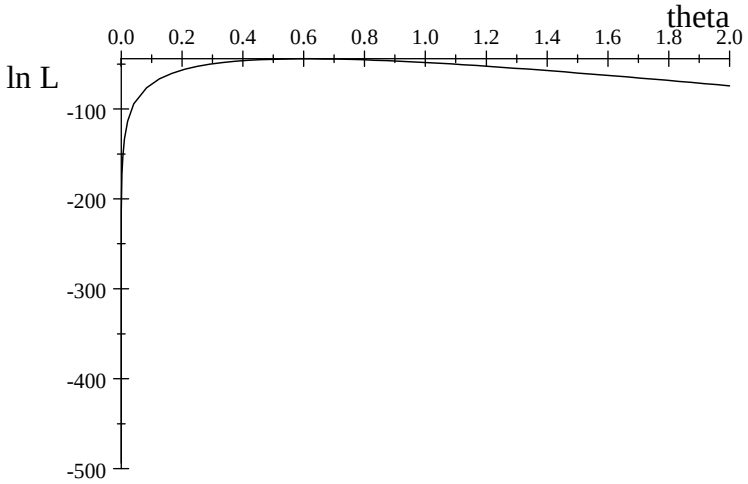

Figure 9.1.9: Plot of the log-likelihood function in Example 9.1.8.

By successively plotting the likelihood on shorter and shorter intervals, the MLE was determined to be $\hat{\theta} = 0.603535$. This value leads to the probabilities

$$\begin{aligned} p_1(\hat{\theta}) &= 1 - e^{-0.603535} = 0.453125, \\ p_2(\hat{\theta}) &= e^{-0.603535} - e^{-2(0.603535)} = 0.247803, \\ p_3(\hat{\theta}) &= e^{-2(0.603535)} - e^{-3(0.603535)} = 0.135517, \\ p_4(\hat{\theta}) &= e^{-3(0.603535)} = 0.163555, \end{aligned}$$

the fitted values

$$\begin{aligned} 30p_1(\hat{\theta}) &= 13.59375, \\ 30p_2(\hat{\theta}) &= 7.43409, \\ 30p_3(\hat{\theta}) &= 4.06551, \\ 30p_4(\hat{\theta}) &= 4.90665, \end{aligned}$$

and the standardized residuals

$$\begin{aligned} r_1(\hat{\theta}) &= (5 - 13.59375) / \sqrt{30\,(0.453125)\,(1 - 0.453125)} = -3.151875, \\ r_2(\hat{\theta}) &= (16 - 7.43409) / \sqrt{30\,(0.247803)\,(1 - 0.247803)} = 3.622378, \\ r_3(\hat{\theta}) &= (8 - 4.06551) / \sqrt{30\,(0.135517)\,(1 - 0.135517)} = 2.098711, \\ r_4(\hat{\theta}) &= (1 - 4.90665) / \sqrt{30\,(0.163555)\,(1 - 0.163555)} = -1.928382. \end{aligned}$$

Note that two of the standardized residuals look large. Finally, we compute

$$\begin{aligned} X_0^2 &= (1 - 0.453125)\,(-3.151875)^2 + (1 - 0.247803)\,(3.622378)^2 \\ &\quad + (1 - 0.135517)\,(2.098711)^2 + (1 - 0.163555)\,(-1.928382)^2 \\ &= 22.221018 \end{aligned}$$

and

$$P\left(X^2 \geq 22.221018\right) = 0.0000$$

when $X^2 \sim \chi^2\,(2)$. Therefore, we have strong evidence that the Exponential(θ) model is not correct for these data and we would not use this model to make inference about θ.

Note that we used the MLE of θ based on the count data and not the original sample! If instead we were to use the MLE for θ based on the original sample (in this case, equal to $\bar{x}$ and so much easier to compute), then Theorem 9.1.2 would no longer be valid. ∎

The chi-squared goodness of fit test is but one of many discrepancy statistics that have been proposed for model checking in the statistical literature. The general approach is to select a discrepancy statistic D, like X^2, such that the exact or asymptotic distribution of D is independent of θ and known. We then compute a P-value based on D. The *Kolmogorov–Smirnov test* and the *Cramer–von Mises test* are further examples of such discrepancy statistics, but we do not discuss these here.

9.1.3 Prediction and Cross-Validation

Perhaps the most rigorous test that a scientific model or theory can be subjected to is assessing how well it predicts new data after it has been fit to an independent data set. In fact, this is a crucial step in the acceptance of any new empirically developed scientific theory — to be accepted, it must predict new results beyond the data that led to its formulation.

If a model does not do a good job at predicting new data, then it is reasonable to say that we have evidence against the model being correct. If the model is too simple, then

the fitted model will underfit the observed data and also the future data. If the model is too complicated, then the model will overfit the original data, and this will be detected when we consider the new data in light of this fitted model.

In statistical applications, we typically cannot wait until new data are generated to check the model. So statisticians use a technique called *cross-validation*. For this, we split an original data set $x_1, \ldots, x_n$ into two parts: the *training set* T, comprising k of the data values and used to fit the model; and the *validation set* V, which comprises the remaining $n - k$ data values. Based on the training data, we construct predictors of various aspects of the validation data. Using the discrepancies between the predicted and actual values, we then assess whether or not the validation set V is surprising as a possible future sample from the model.

Of course, there are

$$\binom{n}{k}$$

possible such splits of the data and we would not want to make a decision based on just one of these. So a cross-validational analysis will have to take this into account. Furthermore, we will have to decide how to measure the discrepancies between T and V and choose a value for k. We do not pursue this topic any further in this text.

9.1.4 What Do We Do When a Model Fails?

So far we have been concerned with determining whether or not an assumed model is appropriate given observed data. Suppose the result of our model checking is that we decide a particular model is *inappropriate*. What do we do now?

Perhaps the obvious response is to say that we have to come up with a more appropriate model — one that will pass our model checking. It is not obvious how we should go about this, but statisticians have devised some techniques.

One of the simplest techniques is the *method of transformations*. For example, suppose that we observe a sample $y_1, \ldots, y_n$ from the distribution given by $Y = \exp(X)$ with $X \sim N(\mu, \sigma^2)$. A normal probability plot based on the y_i, as in Figure 9.1.10, will detect evidence of the nonnormality of the distribution. Transforming these y_i values to $\ln y_i$ will, however, yield a reasonable looking normal probability plot, as in Figure 9.1.11.

So in this case, a simple transformation of the sample yields a data set that passes this check. In fact, this is something statisticians commonly do. Several transformations from the family of *power transformations* given by Y^p for $p \neq 0$, or the logarithm transformation $\ln Y$, are tried to see if a distributional assumption can be satisfied by a transformed sample. We will see some applications of this in Chapter 10. Surprisingly, this simple technique often works, although there are no guarantees that it always will.

Perhaps the most commonly applied transformation is the logarithm when our data values are positive (note that this is a necessity for this transformation). Another very common transformation is the square root transformation, i.e., $p = 1/2$, when we have count data. Of course, it is not correct to try many, many transformations until we find one that makes the probability plots or residual plots look acceptable. Rather, we try a few simple transformations.

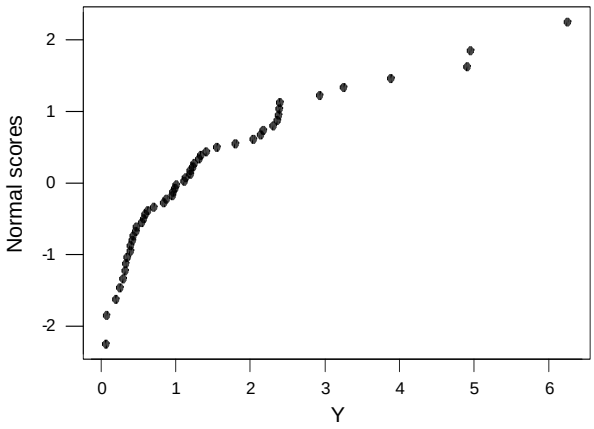

Figure 9.1.10: A normal probability plot of a sample of $n = 50$ from the distribution given by $Y = \exp(X)$ with $X \sim N(0, 1)$.

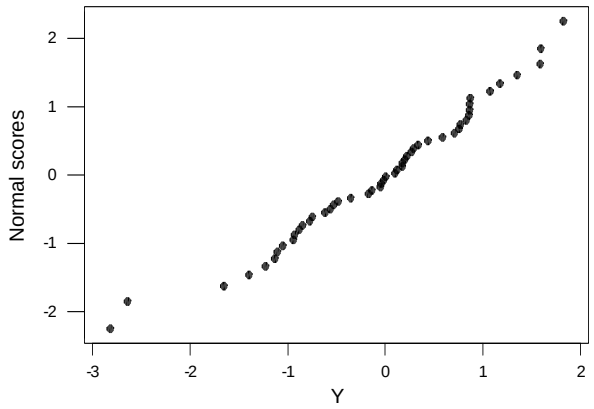

Figure 9.1.11: A normal probability plot of a sample of $n = 50$ from the distribution given by $\ln Y$, where $Y = \exp(X)$ and $X \sim N(0, 1)$.

Summary of Section 9.1

- Model checking is a key component of the practical application of statistics.
- One approach to model checking involves choosing a discrepancy statistic D and then assessing whether or not the observed value of D is surprising by computing a P-value.

- Computation of the P-value requires that the distribution of D be known under the assumption that the model is correct. There are two approaches to accomplishing this. One involves choosing D to be ancillary, and the other involves computing the P-value using the conditional distribution of the data given the minimal sufficient statistic.
- The chi-squared goodness of fit statistic is a commonly used discrepancy statistic. For large samples, with the expected cell counts determined by the MLE based on the multinomial likelihood, the chi-squared goodness of fit statistic is approximately ancillary.
- There are also many informal methods of model checking based on various plots of residuals.
- If a model is rejected, then there are several techniques for modifying the model. These typically involve transformations of the data. Also, a model that fails a model-checking procedure may still be useful, if the deviation from correctness is small.

EXERCISES

9.1.1 Suppose the following sample is assumed to be from an $N(\theta, 4)$ distribution with $\theta \in R^1$ unknown.

1.8	2.1	−3.8	−1.7	−1.3	1.1	1.0	0.0	3.3	1.0
−0.4	−0.1	2.3	−1.6	1.1	−1.3	3.3	−4.9	−1.1	1.9

Check this model using the discrepancy statistic of Example 9.1.1.

9.1.2 Suppose the following sample is assumed to be from an $N(\theta, 2)$ distribution with θ unknown.

−0.4	1.9	−0.3	−0.2	0.0	0.0	−0.1	−1.1	2.0	0.4

(a) Plot the standardized residuals.

(b) Construct a normal probability plot of the standardized residuals.

(c) What conclusions do you draw based on the results of parts (a) and (b)?

9.1.3 Suppose the following sample is assumed to be from an $N(\mu, \sigma^2)$ distribution, where $\mu \in R^1$ and $\sigma^2 > 0$ are unknown.

14.0	9.4	12.1	13.4	6.3	8.5	7.1	12.4	13.3	9.1

(a) Plot the standardized residuals.

(b) Construct a normal probability plot of the standardized residuals.

(c) What conclusions do you draw based on the results of parts (a) and (b)?

9.1.4 Suppose the following table was obtained from classifying members of a sample of $n = 10$ from a student population according to the classification variables A and B, where $A = 0, 1$ indicates male, female and $B = 0, 1$ indicates conservative, liberal.

	$B = 0$	$B = 1$
$A = 0$	2	1
$A = 1$	3	4

Check the model that says gender and political orientation are independent, using Fisher's exact test.

9.1.5 The following sample of $n = 20$ is supposed to be from a Uniform[0, 1] distribution.

0.11	0.56	0.72	0.18	0.26	0.32	0.42	0.22	0.96	0.04
0.45	0.22	0.08	0.65	0.32	0.88	0.76	0.32	0.21	0.80

After grouping the data, using a partition of five equal-length intervals, carry out the chi-squared goodness of fit test to assess whether or not we have evidence against this assumption. Record the standardized residuals.

9.1.6 Suppose a die is tossed 1000 times, and the following frequencies are obtained for the number of pips up when the die comes to a rest.

x_1	x_2	x_3	x_4	x_5	x_6
163	178	142	150	183	184

Using the chi-squared goodness of fit test, assess whether we have evidence that this is not a symmetrical die. Record the standardized residuals.

9.1.7 Suppose the sample space for a response is given by $S = \{0, 1, 2, 3, \ldots\}$.

(a) Suppose that a statistician believes that in fact the response will lie in the set $S = \{10, 11, 12, 13, \ldots\}$ and so chooses a probability measure P that reflects this. When the data are collected, however, the value $s = 3$ is observed. What is an appropriate P-value to quote as a measure of how surprising this value is as a potential value from P?

(b) Suppose instead P is taken to be a Geometric(0.1) distribution. Determine an appropriate P-value to quote as a measure of how surprising $s = 3$ is as a potential value from P.

9.1.8 Suppose we observe $s = 3$ heads in $n = 10$ independent tosses of a purportedly fair coin. Compute a P-value that assesses how surprising this value is as a potential value from the distribution prescribed. Do not use the chi-squared test.

9.1.9 Suppose you check a model by computing a P-value based on some discrepancy statistic and conclude that there is no evidence against the model. Does this mean the model is correct? Explain your answer.

9.1.10 Suppose you are told that standardized scores on a test are distributed $N(0, 1)$. A student's standardized score is -4. Compute an appropriate P-value to assess whether or not the statement is correct.

9.1.11 It is asserted that a coin is being tossed in independent tosses. You are somewhat skeptical about the independence of the tosses because you know that some people practice tossing coins so that they can increase the frequency of getting a head. So you wish to assess the independence of $(x_1, \ldots, x_n)$ from a Bernoulli(θ) distribution.

(a) Show that the conditional distribution of $(x_1, \ldots, x_n)$ given $\bar{x}$ is uniform on the set of all sequences of length n with entries from $\{0, 1\}$.

(b) Using this conditional distribution, determine the distribution of the number of 1's observed in the first $\lfloor n/2 \rfloor$ observations. (Hint: The hypergeometric distribution.)

(c) Suppose you observe (1, 1, 1, 1, 1, 0, 0, 0, 0, 1). Compute an appropriate P-value to assess the independence of these tosses using (b).

COMPUTER EXERCISES

9.1.12 For the data of Exercise 9.1.1, present a normal probability plot of the standardized residuals and comment on it.

9.1.13 Generate 25 samples from the $N(0, 1)$ distribution with $n = 10$ and look at their normal probability plots. Draw any general conclusions.

9.1.14 Suppose the following table was obtained from classifying members of a sample on $n = 100$ from a student population according to the classification variables A and B, where $A = 0, 1$ indicates male, female and $B = 0, 1$ indicates conservative, liberal.

	$B = 0$	$B = 1$
$A = 0$	20	15
$A = 1$	36	29

Check the model that gender and political orientation are independent using Fisher's exact test.

9.1.15 Using software, generate a sample of $n = 1000$ from the Binomial$(10, 0.2)$ distribution. Then, using the chi-squared goodness of fit test, check that this sample is indeed from this distribution. Use grouping to ensure $E(X_i) = np_i \geq 1$. What would you conclude if you got a P-value close to 0?

9.1.16 Using a statistical package, generate a sample of $n = 1000$ from the Poisson(5) distribution. Then, using the chi-squared goodness of fit test based on grouping the observations into five cells chosen to ensure $E(X_i) = np_i \geq 1$, check that this sample is indeed from this distribution. What would you conclude if you got a P-value close to 0?

9.1.17 Using a statistical package, generate a sample of $n = 1000$ from the $N(0, 1)$ distribution. Then, using the chi-squared goodness of fit test based on grouping the observations into five cells chosen to ensure $E(X_i) = np_i \geq 1$, check that this sample is indeed from this distribution. What would you conclude if you got a P-value close to 0?

PROBLEMS

9.1.18 (*Multivariate normal distribution*) A random vector $Y = (Y_1, \ldots, Y_k)$ is said to have a multivariate normal distribution with mean vector $\mu \in R^k$ and variance matrix $\Sigma = (\sigma_{ij}) \in R^{k \times k}$ if

$$a_1 Y_1 + \cdots + a_k Y_k \sim N\left(\sum_{i=1}^{k} a_i \mu_i, \sum_{i=1}^{k}\sum_{j=1}^{k} a_i a_j \sigma_{ij}\right)$$

for every choice of $a_1, \ldots, a_k \in R^1$. We write $Y \sim N_k(\mu, \Sigma)$. Prove that $E(Y_i) = \mu_i$, $\text{Cov}(Y_i, Y_j) = \sigma_{ij}$ and that $Y_i \sim N(\mu_i, \sigma_{ii})$. (Hint: Theorem 3.3.4.)

9.1.19 In Example 9.1.1, prove that the residual $R = (R_1, \ldots, R_n)$ is distributed multivariate normal (see Problem 9.1.18) with mean vector $\mu = (0, \ldots, 0)$ and variance matrix $\Sigma = (\sigma_{ij}) \in R^{k\times k}$, where $\sigma_{ij} = -\sigma_0^2/n$ when $i \neq j$ and $\sigma_{ii} = \sigma_0^2(1 - 1/n)$. (Hint: Theorem 4.6.1.)

9.1.20 If $Y = (Y_1, \ldots, Y_k)$ is distributed multivariate normal with mean vector $\mu \in R^k$ and variance matrix $\Sigma = (\sigma_{ij}) \in R^{k\times k}$, and if $X = (X_1, \ldots, X_l)$ is distributed multivariate normal with mean vector $\nu \in R^l$ and variance matrix $\Upsilon = (\tau_{ij}) \in R^{l\times l}$, then it can be shown that Y and X are independent whenever $\sum_{i=1}^k a_i Y_i$ and $\sum_{i=1}^l b_i X_i$ are independent for every choice of $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_l)$. Use this fact to show that, in Example 9.1.1, $\bar{X}$ and R are independent. (Hint: Theorem 4.6.2 and Problem 9.1.19.)

9.1.21 In Example 9.1.4, prove that $(\hat{\alpha}_1, \hat{\beta}_1) = (x_{1\cdot}/n, x_{\cdot 1}/n)$ is the MLE.

9.1.22 In Example 9.1.4, prove that the number of samples satisfying the constraints (9.1.2) equals

$$\binom{n}{x_{1\cdot}}\binom{n}{x_{\cdot 1}}.$$

(Hint: Using i for the count x_{11}, show that the number of such samples equals

$$\binom{n}{x_{1\cdot}} \sum_{i=\max\{0, x_{1\cdot}+x_{\cdot 1}-n\}}^{\min\{x_{1\cdot}, x_{\cdot 1}\}} \binom{x_{1\cdot}}{i}\binom{n-x_{1\cdot}}{x_{\cdot 1}-i}$$

and sum this using the fact that the sum of Hypergeometric$(n, x_{\cdot 1}, x_{1\cdot})$ probabilities equals 1.)

COMPUTER PROBLEMS

9.1.23 For the data of Exercise 9.1.3, carry out a simulation to estimate the P-value for the discrepancy statistic of Example 9.1.2. Plot a density histogram of the simulated values. (Hint: See Appendix B for appropriate code.)

9.1.24 When $n = 10$, generate 10^4 values of the discrepancy statistic in Example 9.1.2 when we have a sample from an $N(0, 1)$ distribution. Plot these in a density histogram. Repeat this, but now generate from a Cauchy distribution. Compare the histograms (do not forget to make sure both plots have the same scales).

9.1.25 The following data are supposed to have come from an Exponential(θ) distribution, where $\theta > 0$ is unknown.

1.5	1.6	1.4	9.7	12.1	2.7	2.2	1.6	6.8	0.1
0.8	1.7	8.0	0.2	12.3	2.2	0.2	0.6	10.1	4.9

Check this model using a chi-squared goodness of fit test based on the intervals

$$(-\infty, 2.0], (2.0, 4.0], (4.0, 6.0], (6.0, 8.0], (8.0, 10.0], (10.0, \infty).$$

(Hint: Calculate the MLE by plotting the log-likelihood over successively smaller intervals.)

9.1.26 The following table, taken from *Introduction to the Practice of Statistics*, by D. Moore and G. McCabe (W. H. Freeman, New York, 1999), gives the measurements in milligrams of daily calcium intake for 38 women between the ages of 18 and 24 years.

808	882	1062	970	909	802	374	416	784	997
651	716	438	1420	1425	948	1050	976	572	403
626	774	1253	549	1325	446	465	1269	671	696
1156	684	1933	748	1203	2433	1255	110		

(a) Suppose that the model specifies a location normal model for these data with $\sigma_0^2 = (500)^2$. Carry out a chi-squared goodness of fit test on these data using the intervals $(-\infty, 600]$, $(600, 1200]$, $(1200, 1800]$, $(1800, \infty)$. (Hint: Plot the log-likelihood over successively smaller intervals to determine the MLE to about one decimal place. To determine the initial range for plotting, use the overall MLE of μ minus three standard errors to the overall MLE plus three standard errors.)

(b) Compare the MLE of μ obtained in part (a) with the ungrouped MLE.

(c) It would be more realistic to assume that the variance σ^2 is unknown as well. Record the log-likelihood for the grouped data. (More sophisticated numerical methods are needed to find the MLE of (μ, σ^2) in this case.)

9.1.27 Generate 10^4 values of the discrepancy statistics D_{skew} and D_{kurtosis} in Example 9.1.2 when we have a sample of $n = 10$ from an $N(0, 1)$ distribution. Plot these in density histograms. Indicate how you would use these histograms to assess the normality assumption when we had an actual sample of size 10. Repeat this for $n = 20$ and compare the distributions.

CHALLENGES

9.1.28 (MV) Prove that when $(x_1, \ldots, x_n)$ is a sample from the distribution given by $\mu + \sigma Z$, where Z has a known distribution and $(\mu, \sigma^2) \in R^1 \times (0, \infty)$ is unknown, then the statistic

$$r(x_1, \ldots, x_n) = \left(\frac{x_1 - \bar{x}}{s}, \ldots, \frac{x_n - \bar{x}}{s} \right)$$

is ancillary. (Hint: Write a sample element as $x_i = \mu + \sigma z_i$ and then show that $r(x_1, \ldots, x_n)$ can be written as a function of the z_i.)

9.2 Checking for Prior–Data Conflict

Bayesian methodology adds the prior probability measure Π to the statistical model $\{P_\theta : \theta \in \Omega\}$, for the subsequent statistical analysis. The methods of Section 9.1 are designed to check that the observed data can realistically be assumed to have come from a distribution in $\{P_\theta : \theta \in \Omega\}$. When we add the prior, we are in effect saying that our knowledge about the true distribution leads us to assign the prior predictive probability M, given by $M(A) = E_\Pi(P_\theta(A))$ for $A \subset \Omega$, to describe the process generating the data. So it would seem, then, that a sensible Bayesian model-checking

approach would be to compare the observed data s with the distribution given by M, to see if it is surprising or not.

Suppose that we were to conclude that the Bayesian model was incorrect after deciding that s is a surprising value from M. This only tells us, however, that the probability measure M is unlikely to have produced the data and not that the model $\{P_\theta : \theta \in \Omega\}$ was wrong. Consider the following example.

EXAMPLE 9.2.1 *Prior–Data Conflict*
Suppose we obtain a sample consisting of $n = 20$ values of $s = 1$ from the model with $\Omega = \{1, 2\}$ and probability functions for the basic response given by the following table.

	$s = 0$	$s = 1$
$f_1(s)$	0.9	0.1
$f_2(s)$	0.1	0.9

Then the probability of obtaining this sample from f_2 is given by $(0.9)^{20} = 0.12158$, which is a reasonable value, so we have no evidence against the model $\{f_1, f_2\}$.

Suppose we place a prior on Ω given by $\Pi(\{1\}) = 0.9999$, so that we are virtually certain that $\theta = 1$. Then the probability of getting these data from the prior predictive M is

$$(0.9999)(0.1)^{20} + (0.0001)(0.9)^{20} = 1.2158 \times 10^{-5}.$$

The prior probability of observing a sample of 20, whose prior predictive probability is no greater than 1.2158×10^{-5}, can be calculated (using statistical software to tabulate the prior predictive) to be approximately 0.04. This tells us that the observed data are "in the tails" of the prior predictive and thus are surprising, which leads us to conclude that we have evidence that M is incorrect.

So in this example, checking the model $\{f_\theta : \theta \in \Omega\}$ leads us to conclude that it is plausible for the data observed. On the other hand, checking the model given by M leads us to the conclusion that the Bayesian model is implausible. ∎

The lesson of Example 9.2.1 is that we can have model failure in the Bayesian context in two ways. First, the data s may be surprising in light of the model $\{f_\theta : \theta \in \Omega\}$. Second, even when the data are plausibly from this model, the prior and the data may conflict. This conflict will occur whenever the prior assigns most of its probability to distributions in the model for which the data are surprising. In either situation, inferences drawn from the Bayesian model may be flawed.

If, however, the prior assigns positive probability (or density) to every possible value of θ, then the consistency results for Bayesian inference mentioned in Chapter 7 indicate that a large amount of data will overcome a prior–data conflict (see Example 9.2.4). This is because the effect of the prior decreases with increasing amounts of data. So the existence of a prior–data conflict does not necessarily mean that our inferences are in error. Still, it is useful to know whether or not this conflict exists, as it is often difficult to detect whether or not we have sufficient data to avoid the problem.

Therefore, we should first use the checks discussed in Section 9.1 to ensure that the data s is plausibly from the model $\{f_\theta : \theta \in \Omega\}$. If we accept the model, then we look for any prior–data conflict. We now consider how to go about this.

The prior predictive distribution of any ancillary statistic is the same as its distribution under the sampling model, i.e., its prior predictive distribution is not affected by the choice of the prior. So the observed value of any ancillary statistic cannot tell us anything about the existence of a prior–data conflict. We conclude from this that, if we are going to use some function of the data to assess whether or not there is prior–data conflict, then its marginal distribution has to depend on θ.

We now show that the prior predictive conditional distribution of the data given a minimal sufficient statistic T is independent of the prior.

Theorem 9.2.1 Suppose T is a sufficient statistic for the model $\{f_\theta : \theta \in \Omega\}$ for data s. Then the conditional prior predictive distribution of the data s given T is independent of the prior π.

PROOF We will prove this in the case that each sample distribution f_θ and the prior π are discrete. A similar argument can be developed for the more general case.

By Theorem 6.1.1 (factorization theorem) we have that

$$f_\theta(s) = h(s)g_\theta(T(s))$$

for some functions g_θ and h. Therefore the prior predictive probability function of s is given by

$$m(s) = h(s) \sum_{\theta \in \Omega} g_\theta(T(s))\pi(\theta).$$

The prior predictive probability function of T at t is given by

$$m^*(t) = \sum_{\{s:T(s)=t\}} h(s) \sum_{\theta \in \Omega} g_\theta(t)\pi(\theta).$$

Therefore, the conditional prior predictive probability function of the data s given $T(s) = t$ is

$$m(s \mid T = t) = \frac{h(s) \sum_{\theta \in \Omega} g_\theta(t)\pi(\theta)}{\sum_{\{s':T(s')=t\}} h(s') \sum_{\theta \in \Omega} g_\theta(t)\pi(\theta)} = \frac{h(s)}{\sum_{\{s':t(s')=t\}} h(s')},$$

which is independent of π. ∎

So, from Theorem 9.2.1, we conclude that any aspects of the data, beyond the value of a minimal sufficient statistic, can tell us nothing about the existence of a prior–data conflict. Therefore, if we want to base our check for a prior–data conflict on the prior predictive, then we must use the prior predictive for a minimal sufficient statistic. Consider the following examples.

EXAMPLE 9.2.2 *Checking a Beta Prior for a Bernoulli Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from a Bernoulli(θ) model, where $\theta \in [0, 1]$ is unknown, and θ is given a Beta(α, β) prior distribution. Then we have that the sample count $y = \sum_{i=1}^{n} x_i$ is a minimal sufficient statistic and is distributed Binomial(n, θ).

Therefore, the prior predictive probability function for y is given by

$$\begin{aligned} m(y) &= \binom{n}{y} \int_0^1 \theta^y (1-\theta)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \, d\theta \\ &= \frac{\Gamma(n+1)}{\Gamma(y+1)\Gamma(n-y+1)} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)} \\ &\propto \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(y+1)\Gamma(n-y+1)}. \end{aligned}$$

Now observe that when $\alpha = \beta = 1$, then $m(y) = 1/(n+1)$, i.e., the prior predictive of y is Uniform$\{0, 1, \ldots, n\}$, and no values of y are surprising. This is not unexpected, as with the uniform prior on θ, we are implicitly saying that any count y is reasonable.

On the other hand, when $\alpha = \beta = 2$, the prior puts more weight around 1/2. The prior predictive is then proportional to $(y+1)(n-y+1)$. This prior predictive is plotted in Figure 9.2.1 when $n = 20$. Note that counts near 0 or 20 lead to evidence that there is a conflict between the data and the prior. For example, if we obtain the count $y = 3$, we can assess how surprising this value is by computing the probability of obtaining a value with a lower probability of occurrence. Using the symmetry of the prior predictive, we have that this probability equals (using statistical software for the computation) $m(0) + m(2) + m(19) + m(20) = 0.0688876$. Therefore, the observation $y = 3$ is not surprising at the 5% level.

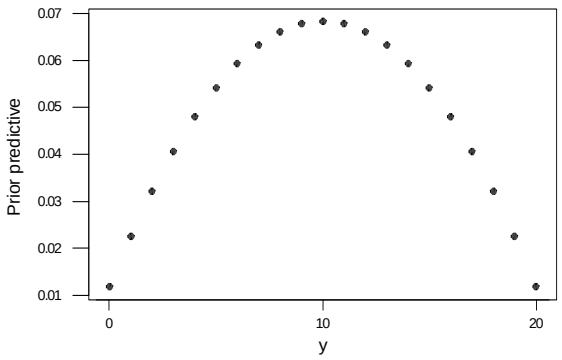

Figure 9.2.1: Plot of the prior predictive of the sample count y in Example 9.2.2 when $\alpha = \beta = 2$ and $n = 20$.

Suppose now that $n = 50$ and $\alpha = 2, \beta = 4$. The mean of this prior is $2/(2+4) = 1/3$ and the prior is right-skewed. The prior predictive is plotted in Figure 9.2.2. Clearly, values of y near 50 give evidence against the model in this case. For example, if we observe $y = 35$, then the probability of getting a count with smaller probability of occurrence is given by (using statistical software for the computation) $m(36) + \cdots + m(50) = 0.0500457$. Only values more extreme than this would provide evidence against the model at the 5% level.

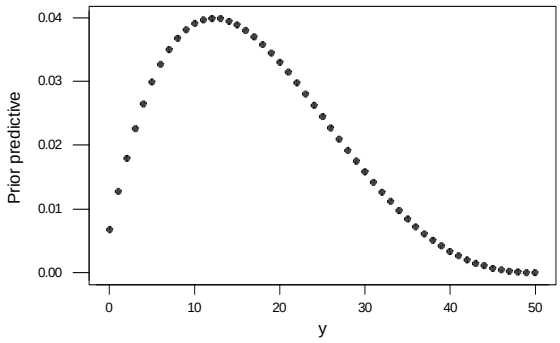

Figure 9.2.2: Plot of the prior predictive of the sample count y in Example 9.2.2 when $\alpha = 2$, $\beta = 4$ and $n = 50$.

∎

EXAMPLE 9.2.3 *Checking a Normal Prior for a Location Normal Model*
Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where $\mu \in R^1$ is unknown and σ_0^2 is known. Suppose we take the prior distribution of μ to be an $N(\mu_0, \tau_0^2)$ for some specified choice of μ_0 and τ_0^2. Note that $\bar{x}$ is a minimal sufficient statistic for this model, so we need to compare the observed of this statistic to its prior predictive distribution to assess whether or not there is prior–data conflict.

Now we can write $\bar{x} = \mu + z$, where $\mu \sim N(\mu_0, \tau_0^2)$ independent of $z \sim N(0, \sigma_0^2/n)$. From this, we immediately deduce (see Exercise 9.2.3) that the prior predictive distribution of $\bar{x}$ is $N(\mu_0, \tau_0^2 + \sigma_0^2/n)$. From the symmetry of the prior predictive density about μ_0, we immediately see that the appropriate P-value is

$$M(|\bar{X} - \mu_0| \leq |\bar{x} - \mu_0|) = 2(1 - \Phi(|\bar{x} - \mu_0|/(\tau_0^2 + \sigma_0^2/n)^{1/2})). \tag{9.2.1}$$

So a small value of (9.2.1) is evidence that there is a conflict between the observed data and the prior, i.e., the prior is putting most of its mass on values of μ for which the observed data are surprising. ∎

Another possibility for model checking in this context is to look at the posterior predictive distribution of the data. Consider, however, the following example.

EXAMPLE 9.2.4 (*Example 9.2.1 continued*)
Recall that, in Example 9.2.1, we concluded that a prior–data conflict existed. Note, however, that the posterior probability of $\theta = 2$ is

$$\frac{(0.0001)\,(0.9)^{20}}{(0.9999)\,(0.1)^{20} + (0.0001)\,(0.9)^{20}} \approx 1.$$

Therefore, the posterior predictive probability of the observed sequence of 20 values of 1 is 0.12158, which does not indicate any prior–data conflict. We note, however, that in this example, the amount of data are sufficient to overwhelm the prior; thus we are led to a sensible inference about θ. ∎

The problem with using the posterior predictive to assess whether or not a prior–data conflict exists is that we have an instance of the so-called *double use of the data*. For we have fit the model, i.e., constructed the posterior predictive, using the observed data, and then we tried to use this posterior predictive to assess whether or not a prior–data conflict exists. The double use of the data results in overly optimistic assessments of the validity of the Bayesian model and will often not detect discrepancies. We will not discuss posterior model checking further in this text.

We have only touched on the basics of checking for prior–data conflict here. With more complicated models, the possibility exists of checking individual components of a prior, e.g., the components of the prior specified in Example 7.1.4 for the location-scale normal model, to ascertain more precisely where a prior–data conflict is arising. Also, ancillary statistics play a role in checking for prior–data conflict as we must remove any ancillary variation when computing the P-value because this variation does not depend on the prior. Furthermore, when the prior predictive distribution of a minimal sufficient statistic is continuous, then issues concerning exactly how P-values are to be computed must be addressed. These are all topics for a further course in statistics.

Summary of Section 9.2

- In Bayesian inference, there are two potential sources of model incorrectness. First, the sampling model for the data may be incorrect. Second, even if the sampling model is correct, the prior may conflict with the data in the sense that most of the prior probability is assigned to distributions in the model for which the data are surprising.
- We first check for the correctness of the sampling model using the methods of Section 9.1. If we do not find evidence against the sampling model, we next check for prior–data conflict by seeing if the observed value of a minimal sufficient statistic is surprising or not, with respect to the prior predictive distribution of this quantity.
- Even if a prior–data conflict exists, posterior inferences may still be valid if we have enough data.

EXERCISES

9.2.1 Suppose we observe the value $s = 2$ from the model, given by the following table.

	$s = 1$	$s = 2$	$s = 3$
$f_1(s)$	$1/3$	$1/3$	$1/3$
$f_2(s)$	$1/3$	0	$2/3$

(a) Do the observed data lead us to doubt the validity of the model? Explain why or why not.

(b) Suppose the prior, given by $\pi(1) = 0.3$, is placed on the parameter $\theta \in \{1, 2\}$. Is there any evidence of a prior–data conflict? (Hint: Compute the prior predictive for each possible data set and assess whether or not the observed data set is surprising.)

(c) Repeat part (b) using the prior given by $\pi(1) = 0.01$.

9.2.2 Suppose a sample of $n = 6$ is taken from a Bernoulli(θ) distribution, where θ has a Beta(3, 3) prior distribution. If the value $n\bar{x} = 2$ is obtained, then determine whether there is any prior–data conflict.

9.2.3 In Example 9.2.3, establish that the prior predictive distribution of $\bar{x}$ is given by the $N(\mu_0, \tau_0^2 + \sigma_0^2/n)$ distribution.

9.2.4 Suppose we have a sample of $n = 5$ from an $N(\mu, 2)$ distribution where μ is unknown and the value $\bar{x} = 7.3$ is observed. An $N(0, 1)$ prior is placed on μ. Compute the appropriate P-value to check for prior–data conflict.

9.2.5 Suppose that $x \sim$ Uniform$[0, \theta]$ and $\theta \sim$Uniform[0, 1]. If the value $x = 2.2$ is observed, then determine an appropriate P-value for checking for prior–data conflict.

COMPUTER EXERCISES

9.2.6 Suppose a sample of $n = 20$ is taken from a Bernoulli(θ) distribution, where θ has a Beta(3, 3) prior distribution. If the value $n\bar{x} = 6$ is obtained, then determine whether there is any prior–data conflict.

PROBLEMS

9.2.7 Suppose that $(x_1, \ldots, x_n)$ is a sample from an $N(\mu, \sigma_0^2)$ distribution, where $\mu \sim N(\mu_0, \tau_0^2)$. Determine the prior predictive distribution of $\bar{x}$.

9.2.8 Suppose that $(x_1, \ldots, x_n)$ is a sample from an Exponential(θ) distribution where $\theta \sim$ Gamma(α_0, β_0). Determine the prior predictive distribution of $\bar{x}$.

9.2.9 Suppose that $(s_1, \ldots, s_n)$ is a sample from a Multinomial$(1, \theta_1, \ldots, \theta_k)$ distribution, where $(\theta_1, \ldots, \theta_{k-1}) \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_k)$. Determine the prior predictive distribution of $(x_1, \ldots, x_k)$, where x_i is the count in the ith category.

9.2.10 Suppose that $(x_1, \ldots, x_n)$ is a sample from a Uniform$[0, \theta]$ distribution, where θ has prior density given by $\pi_{\alpha,\beta}(\theta) = \theta^{-\alpha} I_{[\beta,\infty)}(\theta)/(\alpha - 1)\beta^{\alpha-1}$, where $\alpha > 1$, $\beta > 0$. Determine the prior predictive distribution of $x_{(n)}$.

9.2.11 Suppose we have the context of Example 9.2.3. Determine the limiting P-value for checking for prior–data conflict as $n \to \infty$. Interpret the meaning of this P-value in terms of the prior and the true value of μ.

9.2.12 Suppose that $x \sim$ Geometric(θ) distribution and $\theta \sim$ Uniform[0, 1].

(a) Determine the appropriate P-value for checking for prior–data conflict.

(b) Based on the P-value determined in part (a), describe the circumstances under which evidence of prior–data conflict will exist.

(c) If we use a continuous prior that is positive at a point, then this an assertion that the point is possible. In light of this, discuss whether or not a continuous prior that is positive at $\theta = 0$ makes sense for the Geometric(θ) distribution.

CHALLENGES

9.2.13 Suppose that $X_1, \ldots, X_n$ is a sample from an $N(\mu, \sigma^2)$ distribution where $\mu \mid \sigma^2 \sim N(\mu_0, \tau_0^2\sigma^2)$ and $1/\sigma^2 \sim$ Gamma(α_0, β_0). Then determine a form for the

prior predictive density of $(\bar{X}, S^2)$ that you could evaluate without integrating. (Hint: Use the algebraic manipulations found in Section 7.5.)

9.3 The Problem with Multiple Checks

As we have mentioned throughout this text, model checking is a part of good statistical practice. In other words, one should always be wary of the value of statistical work in which the investigators have not engaged in, and reported the results of, reasonably rigorous model checking. It is really the job of those who report statistical results to convince us that their models are reasonable for the data collected, bearing in mind the effects of both underfitting and overfitting.

In this chapter, we have reported *some* of the possible model-checking approaches available. We have focused on the main categories of procedures and perhaps the most often used methods from within these. There are many others. At this point, we cannot say that any one approach is the best possible method. Perhaps greater insight along these lines will come with further research into the topic, and then a clearer recommendation could be made.

One recommendation that can be made now, however, is that it is not reasonable to go about model checking by implementing every possible model-checking procedure you can. A simple example illustrates the folly of such an approach.

EXAMPLE 9.3.1
Suppose that $(x_1, \ldots, x_n)$ is supposed to be a sample from the $N(0, 1)$ distribution. Suppose we decide to check this model by computing the P-values

$$P_i = P(X_i^2 \geq x_i^2)$$

for $i = 1, \ldots, n$, where $X_i^2 \sim \chi^2(1)$. Furthermore, we will decide that the model is incorrect if the minimum of these P-values is less than 0.05.

Now consider the repeated sampling behavior of this method when the model is correct. We have that

$$\min\{P_1, \ldots, P_n\} < 0.05$$

if and only if

$$\max\{x_1^2, \ldots, x_n^2\} \geq \chi^2_{0.95}(1),$$

and so

$$\begin{aligned} &P\left(\min\{P_1, \ldots, P_n\} < 0.05\right) \\ &= P(\max\{X_1^2, \ldots, X_n^2\} \geq \chi^2_{0.95}(1)) = 1 - P(\max\{X_1^2, \ldots, X_n^2\} \leq \chi^2_{0.05}(1)) \\ &= 1 - \prod_{i=1}^{n} P(X_i^2 \leq \chi^2_{0.95}(1)) = 1 - (0.95)^n \to 1 \end{aligned}$$

as $n \to \infty$. This tells us that if n is large enough, we will reject the model with virtual certainty even though it is correct! Note that n does not have to be very large for there to be an appreciable probability of making an error. For example, when $n = 10$, the

probability of making an error is 0.40; when $n = 20$ the probability of making an error is 0.64; and when $n = 100$, the probability of making an error is 0.99. ∎

We can learn an important lesson from Example 9.3.1, for, if we carry out too many model-checking procedures, we are almost certain to find something wrong — even if the model is correct. The cure for this is that before actually observing the data (so that our choices are not determined by the actual data obtained), we decide on a few relevant model-checking procedures to be carried out and implement only these.

The problem we have been discussing here is sometimes referred to as the problem of *multiple comparisons*, which comes up in other situations as well — e.g., see Section 10.4.1, where multiple means are compared via pairwise tests for differences in the means. One approach for avoiding the multiple-comparisons problem is to simply lower the cutoff for the P-value so that the probability of making a mistake is appropriately small. For example, if we decided in Example 9.3.1 that evidence against the model is only warranted when an individual P-value is smaller than 0.0001, then the probability of making a mistake is 0.01 when $n = 100$. A difficulty with this approach generally is that our model-checking procedures will not be independent, and it does not always seem possible to determine an appropriate cutoff for the individual P-values. More advanced methods are needed to deal with this problem.

Summary of Section 9.3

- Carrying out too many model checks is not a good idea, as we will invariably find something that leads us to conclude that the model is incorrect. Rather than engaging in a "fishing expedition," where we just keep on checking the model, it is better to choose a few procedures before we see the data, and use these, and only these, for the model checking.