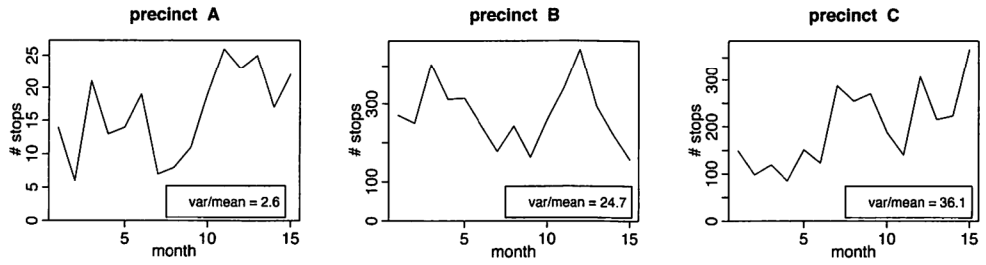

Figure 2.4 *Number of stops by the New York City police for each month over a 15-month period, for three different precincts (chosen to show different patterns in the data).*

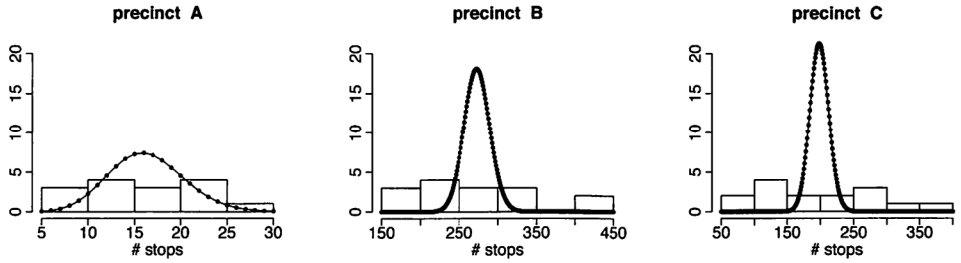

Figure 2.5 *Histograms of monthly counts of stops for the three precincts displayed in 2.4, with fitted Poisson distributions overlain. The data are much more variable than the fitted distributions, indicating overdispersion that is mild in precinct A and huge in precincts B and C.*

Testing for the existence of a variance component

We illustrate with the example of overdispersion in the binomial or Poisson model. For example, the police stop-and-frisk study (see Sections 1.2, 6.2, and 15.1) includes data from a 15-month period. We can examine the data within each precinct to see if the month-to-month variation is greater than would be expected by chance.

Figure 2.4 shows the number of police stops by month, in each of three different precincts. If the data in any precinct really came from a Poisson distribution, we would expect the variance among the counts, $\text{var}_{t=1}^{15} y_t$, to be approximately equal to their mean, $\text{avg}_{t=1}^{15} y_t$. The ratio of variance/mean is thus a measure of dispersion, with var/mean = 1 indicating that the Poisson model is appropriate, and var/mean > 1 indicating overdispersion (and var/mean < 1 indicating underdispersion, but in practice this is much less common). In this example, all three precincts are overdispersed, with variance/mean ratios well over 1.

To give a sense of what this overdispersion implies, Figure 2.5 plots histograms of the monthly counts in each precinct, with the best-fitting Poisson distributions superimposed. The observed counts are much more variable than the model in each case.

Underdispersion

Count data with variance less than the mean would indicate *underdispersion*, but this is rare in actual data. In the police example, underdispersion could possibly result from a "quota" policy in which officers are encouraged to make approximately the same number of stops each month. Figure 2.6 illustrates with hypothetical data in which the number of stops is constrained to be close to 50 each month. In this

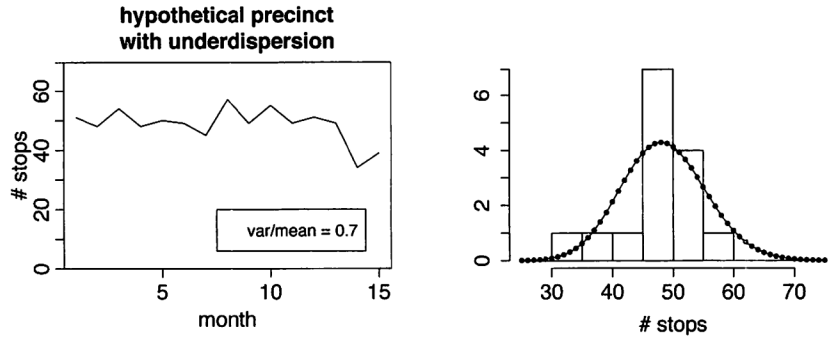

Figure 2.6 *(a) Time series and (b) histogram of number of stops by month for a hypothetical precinct with underdispersed counts. The theoretical Poisson distribution (with parameter set to the mean of the data) is overlain on the histogram.*

particular dataset, the mean is 49 and the variance is 34, and the underdispersion is clear in the histogram.

Multiple hypothesis testing and why we do not worry about it

A concern is sometimes expressed that if you test a large number of hypotheses, then you're bound to reject some. For example, with 100 different hypothesis tests, you would expect about 5 to be statistically significant at the 5% level—even if all the hypotheses were true. This concern is sometimes allayed by *multiple comparisons* procedures, which adjust significance levels to account for the multiplicity of tests.

From our data analysis perspective, however, we are not concerned about multiple comparisons. For one thing, we almost never expect any of our "point null hypotheses" (that is, hypotheses that a parameter equals zero, or that two parameters are equal) to be true, and so we are not particularly worried about the possibility of rejecting them too often. If we examine 100 parameters or comparisons, we expect about half the 50% intervals and about 5% of the 95% intervals to exclude the true values. There is no need to correct for the multiplicity of tests if we accept that they will be mistaken on occasion.

2.5 Problems with statistical significance

A common statistical error is to summarize comparisons by statistical significance and to draw a sharp distinction between significant and nonsignificant results. The approach of summarizing by statistical significance has two pitfalls, one that is obvious and one that is less well known.

First, statistical significance does not equal practical significance. For example, if the estimated predictive effect of height on earnings were \$10 per inch with a standard error of \$2, this would be statistically but not practically significant. Conversely, an estimate of \$10,000 per inch with a standard error of \$10,000 would not be statistically significant, but it has the possibility of being practically significant (and also the possibility of being zero; that is what "not statistically significant" means).

The second problem is that changes in statistical significance are not themselves significant. By this, we are not merely making the commonplace observation that any particular threshold is arbitrary—for example, only a small change is required to move an estimate from a 5.1% significance level to 4.9%, thus moving it into statistical significance. Rather, we are pointing out that even large changes in sig-