CHAPTER 2

Concepts and methods from basic probability and statistics

Simple methods from introductory statistics have three important roles in regression and multilevel modeling. First, simple probability distributions are the building blocks for elaborate models. Second, multilevel models are generalizations of classical complete-pooling and no-pooling estimates, and so it is important to understand where these classical estimates come from. Third, it is often useful in practice to construct quick confidence intervals and hypothesis tests for small parts of a problem—before fitting an elaborate model, or in understanding the output from such a model.

This chapter provides a quick review of some of these methods.

2.1 Probability distributions

A probability distribution corresponds to an urn with a potentially infinite number of balls inside. When a ball is drawn at random, the "random variable" is what is written on this ball.

Areas of application of probability distributions include:

- Distributions of data (for example, heights of men, heights of women, heights of adults), for which we use the notation $y_i,\ i = 1, \ldots, n$.
- Distributions of parameter values, for which we use the notation $\theta_j,\ j = 1, \ldots, J$, or other Greek letters such as α, β, γ. We shall see many of these with the multilevel models in Part 2 of the book. For now, consider a regression model (for example, predicting students' grades from pre-test scores) fit separately in each of several schools. The coefficients of the separate regressions can be modeled as following a distribution, which can be estimated from data.
- Distributions of error terms, which we write as $\epsilon_i,\ i = 1, \ldots, n$—or, for group-level errors, $\eta_j,\ j = 1, \ldots, J$.

A "distribution" is how we describe a set of objects that are not identified, or when the identification gives no information. For example, the heights of a set of unnamed persons have a distribution, as contrasted with the heights of a particular set of your friends.

The basic way that distributions are used in statistical modeling is to start by fitting a distribution to data y, then get predictors X and model y given X with errors ϵ. Further information in X can change the distribution of the ϵ's (typically, by reducing their variance). Distributions are often thought of as data summaries, but in the regression context they are more commonly applied to ϵ's.

Normal distribution; means and variances

The Central Limit Theorem of probability states that the sum of many small independent random variables will be a random variable with an approximate normal

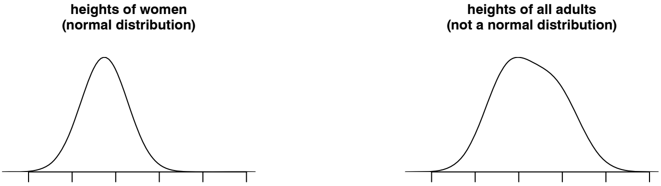

Figure 2.1 *(a) Heights of women (which approximately follow a normal distribution, as predicted from the Central Limit Theorem), and (b) heights of all adults in the United States (which have the form of a mixture of two normal distributions, one for each sex).*

distribution. If we write this summation of independent components as $z = \sum_{i=1}^{n} z_i$, then the mean and variance of z are the sums of the means and variances of the z_i's: $\mu_z = \sum_{i=1}^{n} \mu_{z_i}$ and $\sigma_z = \sqrt{\sum_{i=1}^{n} \sigma_{z_i}^2}$. We write this as $z \sim \mathrm{N}(\mu_z, \sigma_z^2)$.

The Central Limit Theorem holds in practice—that is, $\sum_{i=1}^{n} z_i$ actually follows an approximate normal distribution—if the individual $\sigma_{z_i}^2$'s are small compared to the total variance σ_z^2.

For example, the heights of women in the United States follow an approximate normal distribution. The Central Limit Theorem applies here because height is affected by many small additive factors. In contrast, the distribution of heights of all adults in the United States is not so close to normality. The Central Limit Theorem does not apply here because there is a single large factor—sex—that represents much of the total variation. See Figure 2.1.

Linear transformations. Linearly transformed normal distributions are still normal. For example, if y are men's heights in inches (with mean 69.1 and standard deviation 2.9), then $2.54y$ are their heights in centimeters (with mean $2.54 \cdot 69 = 175$ and standard deviation $2.54 \cdot 2.9 = 7.4$).

For an example of a slightly more complicated calculation, suppose we take independent samples of 100 men and 100 women and compute the difference between the average heights of the men and the average heights of the women. This difference will be normally distributed with mean $69.1 - 63.7 = 5.4$ and standard deviation $\sqrt{2.9^2/100 + 2.7^2/100} = 0.4$ (see Exercise 2.4).

Means and variances of sums of correlated random variables. If x and y are random variables with means μ_x, μ_y, standard deviations σ_x, σ_y, and correlation ρ, then $x + y$ has mean $\mu_x + \mu_y$ and standard deviation $\sqrt{\sigma_x^2 + \sigma_y^2 + 2\rho\sigma_x\sigma_y}$. More generally, the weighted sum $ax + by$ has mean $a\mu_x + b\mu_y$, and its standard deviation is $\sqrt{a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\rho\sigma_x\sigma_y}$. From this we can derive, for example, that $x - y$ has mean $\mu_x - \mu_y$ and standard deviation $\sqrt{\sigma_x^2 + \sigma_y^2 - 2\rho\sigma_x\sigma_y}$.

Estimated regression coefficients. Estimated regression coefficients are themselves linear combinations of data (formally, the estimate $(X^tX)^{-1}X^ty$ is a linear combination of the data values y), and so the Central Limit Theorem again applies, in this case implying that, for large samples, estimated regression coefficients are approximately normally distributed. Similar arguments apply to estimates from logistic regression and other generalized linear models, and for maximum likelihood

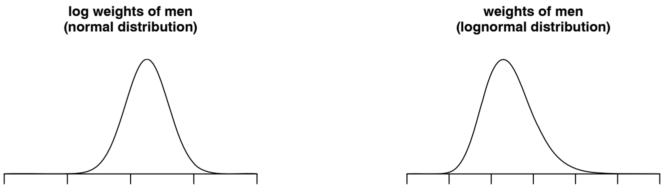

Figure 2.2 *Weights of men (which approximately follow a lognormal distribution, as predicted from the Central Limit Theorem from combining several small multiplicative factors), plotted on the logarithmic and original scales.*

estimation in general (see Section 18.1), for well-behaved models with large sample sizes.

Multivariate normal distribution

More generally, a random vector $z = (z_1, \dots, z_K)$ with a K-dimensional *multivariate normal distribution* with a vector mean μ and a covariance matrix Σ is written as $z \sim \mathrm{N}(\mu, \Sigma)$. The diagonal elements of Σ are the variances of the K individual random variables z_k; thus, we can write $z_k \sim \mathrm{N}(\mu_k, \Sigma_{kk})$. The off-diagonal elements of Σ are the covariances between different elements of z, defined so that the correlation between z_j and z_k is $\Sigma_{jk}/\sqrt{\Sigma_{jj}\Sigma_{kk}}$. The multivariate normal distribution sometimes arises when modeling data, but in this book we encounter it in models for vectors of regression coefficients.

Approximate normal distribution of regression coefficients and other parameter estimates. The least squares estimate of a vector of linear regression coefficients β is $\hat{\beta} = (X^t X)^{-1} X^t y$ (see Section 3.4), which, when viewed as a function of data y (considering the predictors X as constants), is a linear combination of the data. Using the Central Limit Theorem, it can be shown that the distribution of $\hat{\beta}$ will be approximately multivariate normal if the sample size is large. We describe in Chapter 7 how we use this distribution to summarize uncertainty in regression inferences.

Lognormal distribution

It is often helpful to model all-positive random variables on the logarithmic scale. For example, the logarithms of men's weights (in pounds) have an approximate normal distribution with mean 5.13 and standard deviation 0.17. Figure 2.2 shows the distributions of log weights and weights among men in the United States. The exponential of the mean and standard deviations of log weights are called the *geometric mean* and *geometric standard deviation* of the weights; in this example, they are 169 pounds and 1.18, respectively. When working with this *lognormal distribution*, we sometimes want to compute the mean and standard deviation on the original scale; these are $\exp(\mu + \frac{1}{2}\sigma^2)$ and $\exp(\mu + \frac{1}{2}\sigma^2)\sqrt{\exp(\sigma^2) - 1}$, respectively. For the men's weights example, these come to 171 pounds and 29 pounds.

Binomial distribution

If you take 20 shots in basketball, and each has 0.3 probability of succeeding, and if these shots are independent of each other (that is, success in one shot does not increase or decrease the probability of success in any other shot), then the number of shots that succeed is said to have a *binomial distribution* with $n = 20$ and $p = 0.3$, for which we use the notation $y \sim \text{Binomial}(n, p)$. As can be seen even in this simple example, the binomial model is typically only an approximation with real data, where in multiple trials, the probability p of success can vary, and for which outcomes can be correlated. Nonetheless, the binomial model is a useful starting point for modeling such data. And in some settings—most notably, independent sampling with Yes/No responses—the binomial model generally is appropriate, or very close to appropriate.

Poisson distribution

The *Poisson distribution* is used for count data such as the number of cases of cancer in a county, or the number of hits to a website during a particular hour, or the number of persons named Michael whom you know:

- If a county has a population of 100,000, and the average rate of a particular cancer is 45.2 per million persons per year, then the number of cancers in this county could be modeled as Poisson with expectation 4.52.
- If hits are coming at random, with an average rate of 380 per hour, then the number of hits in any particular hour could be modeled as Poisson with expectation 380.
- If you know approximately 1000 persons, and 1% of all persons in the population are named Michael, and you are as likely to know Michaels as anyone else, then the number of Michaels you know could be modeled as Poisson with expectation 10.

As with the binomial distribution, the Poisson model is almost always an idealization, with the first example ignoring systematic differences among counties, the second ignoring clustering or burstiness of the hits, and the third ignoring factors such as sex and age that distinguish Michaels, on average, from the general population.

Again, however, the Poisson distribution is a starting point—as long as its fit to data is checked. The model can be expanded to account for "overdispersion" in data, as we discuss in the context of Figure 2.5 on page 21.

2.2 Statistical inference

Sampling and measurement error models

Statistical inference is used to learn from incomplete or imperfect data. There are two standard paradigms for inference:

- In the *sampling model*, we are interested in learning some characteristics of a population (for example, the mean and standard deviation of the heights of all women in the United States), which we must estimate from a sample, or subset, of that population.
- In the *measurement error model*, we are interested in learning aspects of some underlying pattern or law (for example, the parameters a and b in the model

$y = a + bx$), but the data are measured with error (most simply, $y = a + bx + \epsilon$, although one can also consider models with measurement error in x).

These two paradigms are different: the sampling model makes no reference to measurements, and the measurement model can apply even when complete data are observed. In practice, however, we often combine the two approaches when creating a statistical model.

For example, consider a regression model predicting students' grades from pre-test scores and other background variables. There is typically a sampling aspect to such a study, which is performed on some set of students with the goal of generalizing to a larger population. The model also includes measurement error, at least implicitly, because a student's test score is only an imperfect measure of his or her abilities.

This book follows the usual approach of setting up regression models in the measurement-error framework ($y = a + bx + \epsilon$), with the sampling interpretation implicit in that the errors $\epsilon_i, \ldots, \epsilon_n$ can be considered as a random sample from a distribution (for example, $\mathrm{N}(0, \sigma^2)$) that represents a hypothetical "superpopulation." We consider these issues in more detail in Chapter 21; at this point, we raise this issue only to clarify the connection between probability distributions (which are typically modeled as draws from an urn, or distribution, as described at the beginning of Section 2.1) and the measurement error models used in regression.

Parameters and estimation

The goal of statistical inference for the sorts of *parametric models* that we use is to estimate underlying parameters and summarize our uncertainty in these estimates. We discuss inference more formally in Chapter 18; here it is enough to say that we typically understand a fitted model by plugging in estimates of its parameters, and then we consider the uncertainty in the parameter estimates when assessing how much we actually have learned from a given dataset.

Standard errors

The standard error is the standard deviation of the parameter estimate and gives us a sense of our uncertainty about a parameter and can be used in constructing confidence intervals, as we discuss in the next section. When estimating the mean of an infinite population, given a simple random sample of size n, the standard error is $\sigma/\sqrt{n}$, where σ is the standard deviation of the measurements in the population.

Standard errors for proportions

Consider a survey of size n with y Yes responses and $n-y$ No responses. The estimated proportion of the population who would answer Yes to this survey is $\hat{p} = y/n$, and the standard error of this estimate is $\sqrt{\hat{p}(1-\hat{p})/n}$. This estimate and standard error are usually reasonable unless $y = 0$ or $n-y = 0$, in which case the resulting standard error estimate of zero is misleading.[1]

[1] A reasonable quick correction when y or $n-y$ is near zero is to use the estimate $\hat{p} = (y+1)/(n+2)$ with standard error $\sqrt{\hat{p}(1-\hat{p})/n}$; see Agresti and Coull (1998).

2.3 Classical confidence intervals

Confidence intervals from the normal and t distributions

The usual 95% confidence interval for large samples based on the normal distribution is an estimate ± 2 standard errors. Also from the normal distribution, an estimate ± 1 standard error is a 68% interval, and an estimate $\pm$ 2/3 of a standard error is a 50% interval. A 50% interval is particularly easy to interpret since the true value should be as likely to be inside as outside the interval. A 95% interval is about three times as wide as a 50% interval. The t distribution can be used to correct for uncertainty in the estimation of the standard error.

Continuous data. For example, suppose an object is weighed five times, with measurements $y = 35, 34, 38, 35, 37$, which have an average value of 35.8 and a standard deviation of 1.6. In R, we can create the 50% and 95% t intervals (based on 4 degrees of freedom) as follows:

R code

```
n <- length(y)
estimate <- mean(y)
se <- sd(y)/sqrt(n)
int.50 <- estimate + qt(c(.25,.75),n-1)*se
int.95 <- estimate + qt(c(.025,.975),n-1)*se
```

Proportions. Confidence intervals for proportions come directly from the standard-error formula. For example, if 700 persons in a random sample support the death penalty and 300 oppose it, then a 95% interval for the proportion of supporters in the population is simply $[0.7 \pm 2\sqrt{0.7 \cdot 0.3/1000}] = [0.67, 0.73]$ or, in R,

R code

```
estimate <- y/n
se <- sqrt (estimate*(1-estimate)/n)
int.95 <- estimate + qnorm(c(.025,.975))*se
```

Discrete data. For nonbinary discrete data, we can simply use the continuous formula for the standard error. For example, consider a hypothetical survey that asks 1000 randomly selected adults how many dogs they own, and suppose 600 have no dog, 300 have 1 dog, 50 have 2 dogs, 30 have 3 dogs, and 20 have 4 dogs. What is a 95% confidence interval for the average number of dogs in the population? If the data are not already specified in a file, we can quickly code the data vector R:

R code

```
y <- rep (c(0,1,2,3,4), c(600,300,50,30,20))
```

We can then continue by computing the mean, standard deviation, and standard error, as shown with continuous data above.

Comparisons, visual and numerical

Confidence intervals can often be compared visually, as in Figure 2.3, which displays 68% confidence intervals for the proportion of American adults supporting the death penalty (among those with an opinion on the question), from a series of Gallup polls. For an example of a formal comparison, consider a change in the estimated support for the death penalty from $80\% \pm 1.4\%$ to $74\% \pm 1.3\%$. The estimated difference is 6%, with a standard error of $\sqrt{(1.4\%)^2 + (1.3\%)^2} = 1.9\%$.

Linear transformations

To get confidence intervals for a linear transformed parameter, simply transform the intervals. For example, in the example on page 18, the 95% interval for the number

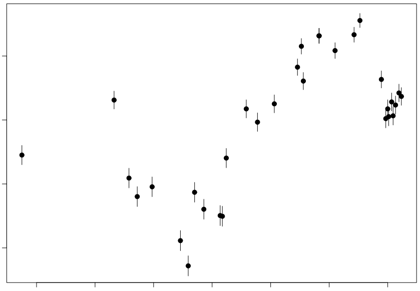

Figure 2.3 *Illustration of visual comparison of confidence intervals. Graph displays the proportion of respondents supporting the death penalty (estimates ±1 standard error—that is, 68% confidence intervals—under the simplifying assumption that each poll was a simple random sample of size 1000), from Gallup polls over time.*

of dogs per person is $[0.52, 0.62]$. Suppose this (hypothetical) random sample were taken in a city of 1 million adults. The confidence interval for the total number of pet dogs in the city is then [520,000, 620,000].

Weighted averages

Confidence intervals for other derived quantities can be determined by appropriately combining the separate means and variances. For example, suppose that separate surveys conducted in France, Germany, Italy, and other countries yield estimates of $55\% \pm 2\%$, $61\% \pm 3\%$, $38\% \pm 3\%$, $\ldots$, for some opinion question. The estimated proportion for all adults in the European Union is $\frac{N_1}{N_{\text{tot}}}55\% + \frac{N_2}{N_{\text{tot}}}61\% + \frac{N_3}{N_{\text{tot}}}38\% + \cdots$, where $N_1, N_2, N_3, \ldots$ are the total number of adults in France, Germany, Italy, $\ldots$, and N_{tot} is the total number in the European Union. The standard error of this weighted average is $\sqrt{(\frac{N_1}{N_{\text{tot}}}2\%)^2 + (\frac{N_2}{N_{\text{tot}}}3\%)^2 + (\frac{N_3}{N_{\text{tot}}}3\%)^2 + \cdots}$.

Given `N`, `p`, `se`—the vectors of population sizes, estimated proportions of Yes responses, and standard errors—we can compute the weighted average and its 95% confidence interval in R:

```
w.avg <- sum(N*p)/sum(N)
se.w.avg <- sqrt (sum ((N*se/sum(N))^2))
int.95 <- w.avg + c(-2,2)*se.w.avg
```

R code

Using simulation to compute confidence intervals for ratios, logarithms, odds ratios, logits, and other functions of estimated parameters

For quantities more complicated than linear transformations, sums, and averages, we can compute standard errors and approximate confidence intervals using simulation. Section 7.2 discusses this in detail; here we illustrate with a quick example.

Consider a survey of 1100 persons, of whom 700 support the death penalty, 300 oppose, and 100 express no opinion. An estimate of the proportion in the population who support the death penalty, among those with an opinion, is 0.7, with a 95% confidence interval is $[0.67, 0.73]$ (see page 18).

Now suppose these 1000 respondents include 500 men and 500 women, and suppose that the death penalty was supported by 75% of the men in the sample and only 65% of the women. We would like to estimate the *ratio* of support for the death penalty among men to that among women. The estimate is easily seen to be $0.75/0.65 = 1.15$—men support it 15% more than women—but computing the standard error is more challenging. The most direct approach, which we recommend, uses simulation.

In R we create 10,000 simulation draws of the inference for men and for women, compute the ratio for each draw, and then determine a 95% interval based on the central 95% of these simulations:

R code

```
n.men <- 500
p.hat.men <- 0.75
se.men <- sqrt (p.hat.men*(1-p.hat.men)/n.men)

n.women <- 500
p.hat.women <- 0.65
se.women <- sqrt (p.hat.women*(1-p.hat.women)/n.women)

n.sims <- 10000
p.men <- rnorm (n.sims, p.hat.men, se.men)
p.women <- rnorm (n.sims, p.hat.women, se.women)
ratio <- p.men/p.women
int.95 <- quantile (ratio, c(.025,.975))
```

which yields a 95% interval of $[1.06, 1.25]$.

2.4 Classical hypothesis testing

The possible outcomes of a hypothesis test are "reject" or "not reject." It is never possible to "accept" a statistical hypothesis, only to find that the data are not sufficient to reject it.

Comparisons of parameters to fixed values and each other: interpreting confidence intervals as hypothesis tests

The hypothesis that a parameter equals zero (or any other fixed value) is directly tested by fitting the model that includes the parameter in question and examining its 95% interval. If the interval excludes zero (or the specified fixed value), then the hypothesis is rejected at the 5% level.

Testing whether two parameters are equal is equivalent to testing whether their difference equals zero. We do this by including both parameters in the model and then examining the 95% interval for their difference. As with inference for a single parameter, the confidence interval is commonly of more interest than the hypothesis test. For example, if support for the death penalty has decreased by $6\% \pm 2.1\%$, then the magnitude of this estimated difference is probably as important as that the change is statistically significantly different from zero.

The hypothesis of whether a parameter is positive is directly assessed via its confidence interval. If both ends of the 95% confidence interval exceed zero, then we are at least 95% sure (under the assumptions of the model) that the parameter is positive. Testing whether one parameter is greater than the other is equivalent to examining the confidence interval for their difference and testing for whether it is entirely positive.

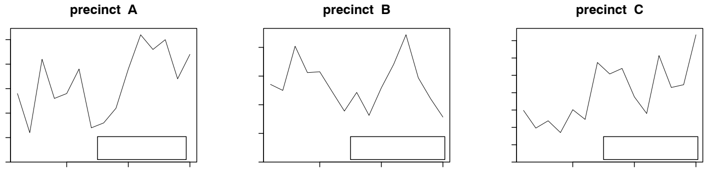

Figure 2.4 *Number of stops by the New York City police for each month over a 15-month period, for three different precincts (chosen to show different patterns in the data).*

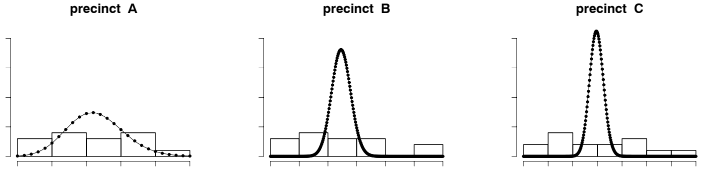

Figure 2.5 *Histograms of monthly counts of stops for the three precincts displayed in 2.4, with fitted Poisson distributions overlain. The data are much more variable than the fitted distributions, indicating overdispersion that is mild in precinct A and huge in precincts B and C.*

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

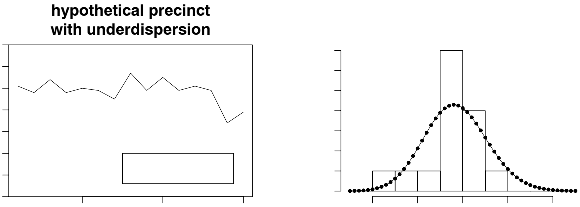

Figure 2.6 *(a) Time series and (b) histogram of number of stops by month for a hypothetical precinct with underdispersed counts. The theoretical Poisson distribution (with parameter set to the mean of the data) is overlain on the histogram.*

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 significance levels can correspond to small, nonsignificant changes in the underlying variables.

For example, consider two independent studies with effect estimates and standard errors of 25 ± 10 and 10 ± 10. The first study is statistically significant at the 1% level, and the second is not at all significant at 1 standard error away from zero. Thus it would be tempting to conclude that there is a large difference between the two studies. In fact, however, the difference is not even close to being statistically significant: the estimated difference is 15, with a standard error of $\sqrt{10^2 + 10^2} = 14$.

Section 21.8 gives a practical example of the pitfalls of using statistical significance to classify studies, along with a discussion of how these comparisons can be better summarized using a multilevel model.

2.6 55,000 residents desperately need your help!

We illustrate the application of basic statistical methods with a story. One day a couple of years ago, we received a fax, entitled HELP!, from a member of a residential organization:

> Last week we had an election for the Board of Directors. Many residents believe, as I do, that the election was rigged and what was supposed to be votes being cast by 5,553 of the 15,372 voting households is instead a fixed vote with fixed percentages being assigned to each and every candidate making it impossible to participate in an honest election.
>
> The unofficial election results I have faxed along with this letter represent the tallies. Tallies were given after 600 were counted. Then again at 1200, 2444, 3444, 4444, and final count at 5553.
>
> After close inspection we believe that there was nothing random about the count and tallies each time and that specific unnatural percentages or rigged percentages were being assigned to each and every candidate.
>
> Are we crazy? In a community this diverse and large, can candidates running on separate and opposite slates as well as independents receive similar vote percentage increases tally after tally, plus or minus three or four percent? Does this appear random to you? What do you think? HELP!

Figure 2.7 shows a subset of the data. These vote tallies were deemed suspicious because the proportion of the votes received by each candidate barely changed throughout the tallying. For example, Clotelia Smith's vote share never went below 34.6% or above 36.6%. How can we HELP these people and test their hypothesis?

We start by plotting the data: for each candidate, the proportion of vote received after 600, 1200, . . . votes; see Figure 2.8. These graphs are difficult to interpret, however, since the data points are not in any sense independent: the vote at any time point includes all the votes that came before. We handle this problem by subtraction to obtain the number of votes for each candidate in the intervals between the vote tallies: the first 600 votes, the next 600, the next 1244, then next 1000, then next 1000, and the final 1109, with the total representing all 5553 votes.

Figure 2.9 displays the results. Even after taking differences, these graphs are fairly stable—but how does this variation compare to what would be expected if votes were actually coming in at random? We formulate this as a hypothesis test and carry it out in five steps:

1. *The null hypothesis* is that the voters are coming to the polls at random. The fax writer believed the data contradicted the null hypothesis; this is what we want to check.
2. *The test statistic* is some summary of the data used to check the hypothesis. Because the concern was that the votes were unexpectedly stable as the count

Clotelia Smith	208	416	867	1259	1610	2020
Earl Coppin	55	106	215	313	401	505
Clarissa Montes	133	250	505	716	902	1129
...	...	...	...	...	...	...

Figure 2.7 *Subset of results from the cooperative board election, with votes for each candidate (names altered for anonymity) tallied after 600, 1200, 2444, 3444, 4444, and 5553 votes. These data were viewed as suspicious because the proportion of votes for each candidate barely changed as the vote counting went on. (There were 27 candidates in total, and each voter was allowed to choose 6 candidates.)*

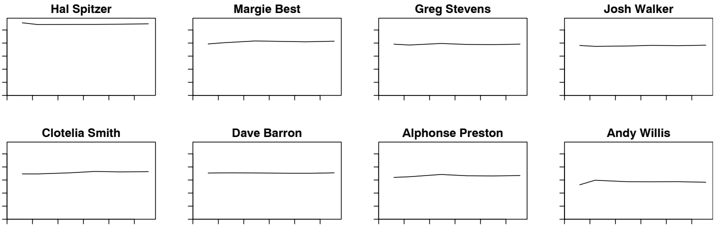

Figure 2.8 *Proportion of votes received by each candidate in the cooperative board election, after each stage of counting: 600, 1200, 2444, ..., 5553 votes. There were 27 candidates in total; for brevity we display just the leading 8 vote-getters here. The vote proportions appear to be extremely stable over time; this might be misleading, however, since the vote at any time point includes all the previous vote tallies. See Figure 2.9.*

proceeded, we define a test statistic to summarize that variability. For each candidate i, we label $y_{i1}, \dots, y_{i6}$ to be the numbers of votes received by the candidates in each of the six recorded stages of the count. (For example, from Figure 2.7, the values of $y_{i1}, y_{i2}, \dots, y_{i6}$ for Earl Coppin are $55, 51, \dots, 104$.) We then compute $p_{it} = y_{it}/n_t$ for $t = 1, \dots, 6$, the proportion of the votes received by candidate i at each stage. The test statistic for candidate i is then the sample standard deviation of these six values $p_{i1}, \dots, p_{i6}$,

$$T_i = \text{sd}_{t=1}^{6} p_{it},$$

a measure of the variation in his or her votes over time.

3. *The theoretical distribution of the test statistic if the null hypothesis were true.* Under the null hypothesis, the six subsets of the election are simply six different random samples of the voters, with a proportion π_i who would vote for candidate i. From the binomial distribution, the proportion p_{it} then has a mean of π_i and a variance of $\pi_i(1-\pi_i)/n_t$. On average, the variance of the six p_{it}'s will equal the average of the six theoretical variances, and so the variance of the p_{it}'s—whose square root is our test statistic—should equal, on average, the theoretical value $\text{avg}_{t=1}^{6} \pi_i(1-\pi_i)/n_t$. The probabilities π_i are not known, so we follow standard practice and insert the empirical probabilities, p_i, so that the expected value of the test statistic, for each candidate i, is

$$T_i^{\text{theory}} = \sqrt{p_i(1-p_i)\text{avg}_{t=1}^{6}(1/n_t)}.$$

4. *Comparing the test statistic to its theoretical distribution.* Figure 2.10 plots the observed and theoretical values of the test statistic for each of the 27 candidates,

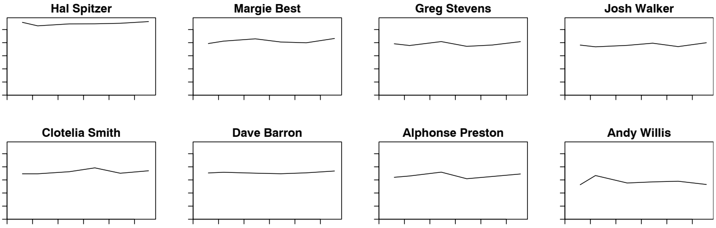

Figure 2.9 *Proportion of votes received by each of the 8 leading candidates in the cooperative board election, at each disjoint stage of voting: the first 600 votes, the next 600, the next 1244, then next 1000, then next 1000, and the final 1109, with the total representing all 5553 votes. The plots here and in Figure 2.8 have been put on a common scale which allows easy comparison of candidates, although at the cost of making it difficult to see details in the individual time series.*

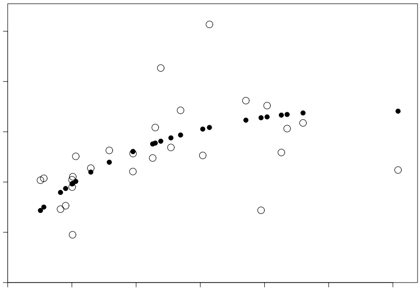

Figure 2.10 *The open circles show, for each of the 27 candidates in the cooperative board election, the standard deviation of the proportions of the vote received by the candidate in the first 600, next 600, next 1244, . . . , and the final 1109 votes, plotted versus the total number of votes received by the candidate. The solid dots show the expected standard deviation of the separate vote proportions for each candidate, based on the binomial model that would be appropriate if voters were coming to the polls at random. The actual standard deviations appear consistent with the theoretical model.*

as a function of the total number of votes received by the candidate. The theoretical values follow a simple curve (which makes sense, since the total number of votes determines the empirical probabilities p_i, which determine T_i^{theory}), and the actual values appear to fit the theory fairly well, with some above and some below.

5. *Summary comparisons using χ^2 tests.* We can also express the hypothesis tests numerically. Under the null hypothesis, the probability of a candidate receiving votes is independent of the time of each vote, and thus the 2×6 table of votes including or excluding each candidate would be consistent with the model of independence. (See Figure 2.10 for an example.) We can then compute for each candidate a χ^2 statistic, $\sum_{j=1}^{2} \sum_{t=1}^{6} (\text{observed}_{jt} - \text{expected}_{jt})^2/\text{expected}_{jt}$, and compare to a χ^2 distribution with $(6-1) \times (2-1) = 5$ degrees of freedom.

Unlike the usual application of χ^2 testing, in this case we are looking for unexpectedly *low* values of the χ^2 statistic (and thus p-values close to 1), which would indicate vote proportions that have suspiciously little variation over time. In fact, however, the χ^2 tests for the 27 candidates show no suspicious patterns: the p-values range from 0 to 1, with about 10% below 0.1, about 10% above 0.9, and no extreme p-values at either end.

Another approach would be to perform a χ^2 test on the entire 27×6 table of votes over time (that is, the table whose first row is the top row of the left table on Figure 2.7, then continues with the data from Earl Coppin, Clarissa Montes, and so forth). This test is somewhat suspect since it ignores that the votes come in batches (each voter can choose up to 6 candidates) but is a convenient summary test. The value of the χ^2 statistic is 115, which, when compared to a χ^2 distribution with $(27-1) \times (6-1) = 130$ degrees of freedom, has a p-value of 0.83—indicating slightly less variation than expected, but not statistically significant. That is, if the null hypothesis were true, we would not be particularly surprised to see a χ^2 statistic of 115.

We thus conclude that the intermediate vote tallies are consistent with random voting. As we explained to the writer of the fax, opinion polls of 1000 people are typically accurate to within 2%, and so, if voters really are arriving at random, it makes sense that batches of 1000 votes are highly stable. This does not rule out the possibility of fraud, but it shows that this aspect of the voting is consistent with the null hypothesis.

2.7 Bibliographic note

De Veaux, Velleman, and Bock (2006) is a good introductory statistics textbook, and Ramsey and Schafer (2001) and Snedecor and Cochran (1989) are also good sources for classical statistical methods. A quick summary of probability distributions appears in appendix A of Gelman et al. (2003).

Agresti and Coull (1998) consider the effectiveness of various quick methods of inference for binomial proportions. Gilovich, Vallone, and Tversky (1985) discuss the applicability of the binomial model to basketball shooting, along with psychological difficulties in interpreting binomial data.

See Browner and Newman (1987), Krantz (1999), and Gelman and Stern (2006) for further discussion and references on the problems with statistical significance.

The data on heights and weights of Americans come from Brainard and Burmaster (1992). The voting example in Section 2.6 comes from Gelman (2004c).

2.8 Exercises

The data for the assignments in this and other chapters are at `www.stat.columbia.edu/~gelman/arm/examples/`. See Appendix C for further details.

1. A test is graded from 0 to 50, with an average score of 35 and a standard deviation of 10. For comparison to other tests, it would be convenient to rescale to a mean of 100 and standard deviation of 15.

   (a) How can the scores be linearly transformed to have this new mean and standard deviation?

   (b) There is another linear transformation that also rescales the scores to have mean 100 and standard deviation 15. What is it, and why would you *not* want to use it for this purpose?

2. The following are the proportions of girl births in Vienna for each month in 1908 and 1909 (out of an average of 3900 births per month):

   .4777 .4875 .4859 .4754 .4874 .4864 .4813 .4787 .4895 .4797 .4876 .4859
   .4857 .4907 .5010 .4903 .4860 .4911 .4871 .4725 .4822 .4870 .4823 .4973

   The data are in the folder `girls`. von Mises (1957) used these proportions to claim that the sex ratios were less variable than would be expected by chance.

   (a) Compute the standard deviation of these proportions and compare to the standard deviation that would be expected if the sexes of babies were independently decided with a constant probability over the 24-month period.

   (b) The actual and theoretical standard deviations from (a) differ, of course. Is this difference statistically significant? (Hint: under the randomness model, the actual variance should have a distribution with expected value equal to the theoretical variance, and proportional to a χ^2 with 23 degrees of freedom.)

3. Demonstration of the Central Limit Theorem: let $x = x_1 + \cdots + x_{20}$, the sum of 20 independent Uniform(0,1) random variables. In R, create 1000 simulations of x and plot their histogram. On the histogram, overlay a graph of the normal density function. Comment on any differences between the histogram and the curve.

4. Distribution of averages and differences: the heights of men in the United States are approximately normally distributed with mean 69.1 inches and standard deviation 2.9 inches. The heights of women are approximately normally distributed with mean 63.7 inches and standard deviation 2.7 inches. Let x be the average height of 100 randomly sampled men, and y be the average height of 100 randomly sampled women. In R, create 1000 simulations of $x - y$ and plot their histogram. Using the simulations, compute the mean and standard deviation of the distribution of $x - y$ and compare to their exact values.

5. Correlated random variables: suppose that the heights of husbands and wives have a correlation of 0.3. Let x and y be the heights of a married couple chosen at random. What are the mean and standard deviation of the average height, $(x + y)/2$?