Chapter 4

Statistical inference

Note:
This chapter includes several code examples in R.
Feel free to skip them as they're not essential for understanding the content.
If you're feeling adventurous, feel free to skim them and try to reproduce them in Python.

Statistical inference can be formulated as a set of operations on data that yield estimates and uncertainty statements about predictions and parameters of some underlying process or population. From a mathematical standpoint, these probabilistic uncertainty statements are derived based on some assumed probability model for observed data. In this chapter, we sketch the basics of probability modeling, estimation, bias and variance, and the interpretation of statistical inferences and statistical errors in applied work. We introduce the theme of uncertainty in statistical inference and discuss how it is a mistake to use hypothesis tests or statistical significance to attribute certainty from noisy data.

4.1 Sampling distributions and generative models

Sampling, measurement error, and model error

Statistical inference is used to learn from incomplete or imperfect data. There are three standard paradigms for thinking about the role of 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 coefficients a and b in the model $y_i = a + bx_i$), but the data are measured with error (most simply, $y_i = a + bx_i + \epsilon_i$, although one can also consider models with measurement error in x). Measurement error need not be additive: multiplicative models can make sense for positive data, and discrete distributions are needed for modeling discrete data.
- *Model error* refers to the inevitable imperfections of the models that we apply to real data.

These three paradigms are different: the sampling model makes no reference to measurements, the measurement model can apply even when complete data are observed, and model error can arise even with perfectly precise observations. In practice, we often consider all three issues when constructing and working with 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, and also model error because any assumed functional form can only be approximate. In addition, any student's ability will vary by time and by circumstance; this variation can be thought of as measurement or model error.

This book follows the usual approach of setting up regression models in the measurement-error framework ($y_i = a + bx_i + \epsilon_i$), with the ϵ_i's also interpretable as model error, and with the sampling interpretation implicit in that the errors $\epsilon_1, \ldots, \epsilon_n$ can be considered as a random sample from a distribution (for example, normal with mean 0 and standard deviation σ) that represents a hypothetical "superpopulation." We raise this issue only to clarify the connection between probability distributions (which are typically modeled as draws from an urn, or distribution) and the measurement or model errors used in regression.

The sampling distribution

The *sampling distribution* is the set of possible datasets that could have been observed if the data collection process had been re-done, along with the probabilities of these possible values. The sampling distribution is determined by the data collection process, or the model being used to represent that process, which can include random sampling, treatment assignment, measurement error, model error, or some combination of all of these. The term "sampling distribution" is somewhat misleading, as this variation need not come from or be modeled by any sampling process—a more accurate term might be "probabilistic data model"—but for consistency with traditional terminology in statistics, we call it the sampling distribution even when no sampling is involved.

The simplest example of a sampling distribution is the pure random sampling model: if the data are a simple random sample of size n from a population of size N, then the sampling distribution is the set of all samples of size n, all with equal probabilities. The next simplest example is pure measurement error: if observations y_i, $i = 1, \dots, n$, are generated from the model $y_i = a + bx_i + \epsilon_i$, with fixed coefficients a and b, pre-specified values of the predictor x_i, and a specified distribution for the errors ϵ_i (for example, normal with mean 0 and standard deviation σ), then the sampling distribution is the set of possible datasets obtained from these values of x_i, drawing new errors ϵ_i from their assigned distribution.

As both these examples illustrate, the sampling distribution in general will not typically be known, as it depends on aspects of the population, not merely on the observed data. In the case of the simple random sample of size n, the sampling distribution depends on all N datapoints. In practice, then, we will not *know* the sampling distribution; we can only *estimate* it. Similarly, for the measurement-error model, the sampling distribution depends on the parameters a, b, and σ, which in general will be estimated from the data, not known.

Even if data have not been collected by any random process, for statistical inference it is helpful to assume some probability model for the data. For example, in Section 7.1 we fit and interpret a regression predicting presidential election outcomes from the national economy. The 16 elections in our dataset are not a random sample from any larger set of elections, nor are elections the result of some random process. Nonetheless, we assume the model $y_i = a + bx_i + \epsilon_i$ and work out the sampling distribution implied from that.

The sampling distribution is said to be a *generative model* in that it represents a random process which, if known, could generate a new dataset. Next we discuss how to use the sampling distribution to define the statistical properties of estimates.

4.2 Estimates, standard errors, and confidence intervals

Parameters, estimands, and estimates

In statistics jargon, *parameters* are the unknown numbers that determine a statistical model. For example, consider the model $y_i = a + bx_i + \epsilon_i$, in which the errors ϵ_i are normally distributed with mean 0 and standard deviation σ. The parameters in this model are a, b, and σ. The parameters a and b are called *coefficients*, and σ is a called a *scale* or *variance parameter*. One way to think about estimated parameters is that they can be used to simulate new (hypothetical) data from the model.

An *estimand*, or *quantity of interest*, is some summary of parameters or data that somebody is interested in estimating. For example, in the regression model, $y = a + bx +$ error, the parameters a and b might be of interest—a is the intercept of the model, the predicted value of y when $x = 0$; and b is the slope, the predicted difference in y, comparing two data points that differ by 1 in x. Other quantities of interest could be predicted outcomes for particular new data points, or combinations of predicted values such as sums, differences, averages, and ratios.

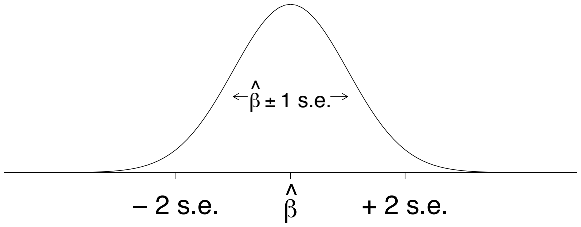

Figure 4.1 *Distribution representing uncertainty in the estimate of a quantity of interest β. The range of this distribution corresponds to the possible values of β that are consistent with the data. In this case, we have assumed a normal distribution for this sampling distribution and therefore we have assigned an approximate 68% chance that β will lie within 1 standard error (s.e.) of the point estimate, $\hat{\beta}$, and an approximate 95% chance that β will lie within 2 standard errors. Assuming the model is correct, it should happen only about 5% of the time that the estimate, $\hat{\beta}$, falls more than 2 standard errors away from the true β.*

We use the data to construct *estimates* of parameters and other quantities of interest. The sampling distribution of an estimate is a byproduct of the sampling distribution of the data used to construct it. We evaluate the statistical properties of estimates analytically or by repeatedly simulating from the random sampling distribution on the computer, as discussed in Chapter 5.

Standard errors, inferential uncertainty, and confidence intervals

The *standard error* is the estimated standard deviation of an estimate and can give us a sense of our uncertainty about the quantity of interest. Figure 4.1 illustrates in the context of a normal (bell-shaped) sampling distribution, which for mathematical reasons (the Central Limit Theorem; see page 41) can be expected to arise in many statistical contexts.

As discussed in Sections 5.3 and 9.1, in our current practice we usually summarize uncertainty using simulation, and we give the term "standard error" a looser meaning to cover any measure of uncertainty that is comparable to the posterior standard deviation.

However defined, the standard error is a measure of the variation in an estimate and gets smaller as sample size gets larger, converging on zero as the sample increases in size.

Example: Coverage of confidence intervals

The *confidence interval* represents a range of values of a parameter or quantity of interest that are roughly consistent with the data, given the assumed sampling distribution. If the model is correct, then in repeated applications the 50% and 95% confidence intervals will include the true value 50% and 95% of the time; see Figure 4.2 and Exercise 5.7.[1]

The usual 95% confidence interval for large samples, based on an assumption that the sampling distribution follows the normal distribution, is to take an estimate ±2 standard errors; see Figure 4.1. Also from the normal distribution, an estimate ±1 standard error is a 68% interval, and an estimate $\pm\frac{2}{3}$ 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 based on the normal distribution is about three times as wide as a 50% interval.

Standard errors and confidence intervals for averages and proportions

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. This property holds regardless of any assumption about the shape of the sampling distribution, but the standard error might be less informative for sampling distributions that are far from normal.

A proportion is a special case of an average in which the data are 0's and 1's. 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 $\sqrt{\hat{p}(1-\hat{p})/n}$. If

[1] Code for this example is in folder `Coverage`.

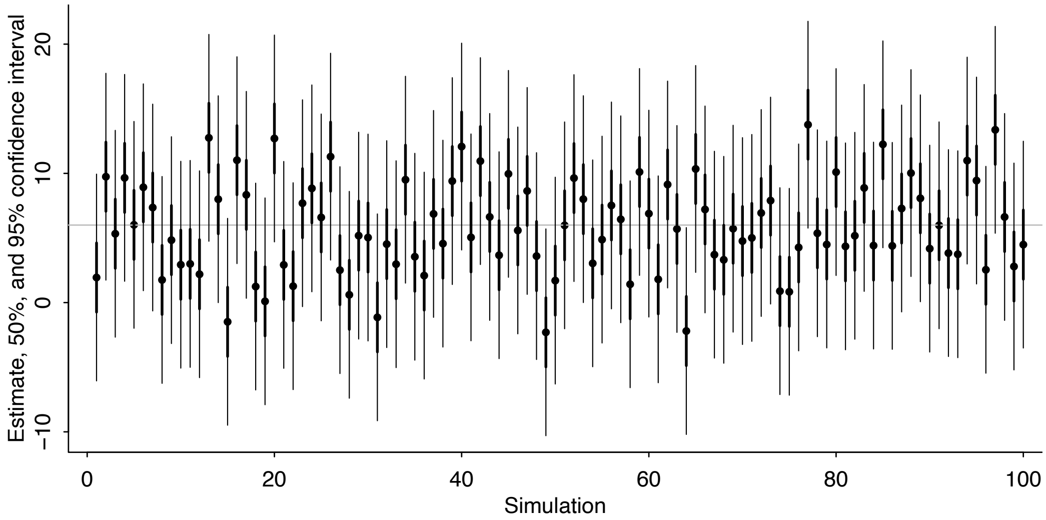

Figure 4.2 *Simulation of coverage of confidence intervals: the horizontal line shows the true parameter value, and dots and vertical lines show estimates and confidence intervals obtained from 100 random simulations from the sampling distribution. If the model is correct, 50% of the 50% intervals and 95% of the 95% intervals should contain the true parameter value, in the long run.*

p is near 0.5, we can approximate this by $0.5/\sqrt{n}$. Consider that $\sqrt{0.5 * 0.5} = 0.5$, $\sqrt{0.4 * 0.6} = 0.49$, and $\sqrt{0.3 * 0.7} = 0.46$.)

Confidence intervals for proportions come directly from the standard-error formula. If 700 people 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 $[0.7 \pm 2\sqrt{0.7 * 0.3/1000}] = [0.67, 0.73]$ or, in R,

```
estimate <- y/n
se <- sqrt(estimate*(1-estimate)/n)
int_95 <- estimate + qnorm(c(0.025, 0.975))*se
```

Standard error and confidence interval for a proportion when $y = 0$ or $y = n$

The above estimate and standard error are usually reasonable unless the number of Yes or the number of No responses is close to zero. Conventionally, the approximation is considered acceptable if y and $n - y$ are both at least 5. In the extreme case in which $y = 0$ or $n - y = 0$, there is an obvious problem in that the formula yields a standard error estimate of zero, and thus a zero-width confidence interval.

A standard and reasonable quick correction for constructing a 95% interval when y or $n - y$ is near zero is to use the estimate $\hat{p} = \frac{y+2}{n+4}$ with standard error $\sqrt{\hat{p}(1-\hat{p})/(n+4)}$. For example, if $y = 0$ and $n = 75$, the 95% interval is $[\hat{p} \pm 2\,\text{s.e.}]$, where $\hat{p} = \frac{2}{79} = 0.025$ and $\text{s.e.} = \sqrt{(0.025)(1-0.025)/79} = 0.018$; this comes to $[-0.01, 0.06]$. It makes no sense for the interval for a proportion to contain negative values, so we truncate the interval to obtain $[0, 0.06]$. If $y = n$, we perform the same procedure but set the upper bound of the interval to 1.

Standard error for a comparison

The standard error of the difference of two independent quantities is computed as,

$$\text{standard error of the difference} = \sqrt{\text{se}_1^2 + \text{se}_2^2}. \tag{4.1}$$

Consider a survey of 1000 people—400 men and 600 women—who are asked how they plan to vote in an upcoming election. Suppose that 57% of the men and 45% of the women plan to vote for the Republican candidate. The standard errors for these proportions are $\text{se}_{\text{men}} = \sqrt{0.57 * 0.43/400}$ and $\text{se}_{\text{women}} = \sqrt{0.45 * 0.55/600}$. The estimated gender gap in support for the Republican is $0.57 - 0.45 = 0.12$, with standard error $\sqrt{\text{se}^2_{\text{men}} + \text{se}^2_{\text{women}}} = 0.032$.

Sampling distribution of the sample mean and standard deviation; normal and χ^2 distributions

Suppose you draw n data points $y_1, \dots, y_n$ from a normal distribution with mean μ and standard deviation σ, and then compute the sample mean, $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, and standard deviation, $s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2}$. These two statistics have a sampling distribution that can be derived mathematically from the properties of independent samples from the normal. The sample mean, $\bar{y}$, is normally distributed with mean μ and standard deviation $\sigma/\sqrt{n}$. The sample standard deviation has a distribution defined as follows: $s_y^2 * (n-1)/\sigma^2$ has a χ^2 distribution with $n-1$ degrees of freedom. We give neither the formulas nor the derivations of these distributions here, as they are not necessary for the applied methods in this book. But it is good to know the names of these distributions, as we can compute their quantiles and simulate from them in R to get a sense of what can be expected from data summaries. In addition, these distributions reappear in regression modeling when performing inference for coefficients and residual variation.

Degrees of freedom

The concept of *degrees of freedom* arises with the χ^2 distribution and several other places in probability and statistics. Without going into the technical details, we can briefly say that degrees of freedom relate to the need to correct for overfitting when estimating the error of future predictions from a fitted model. If we simply calculate predictive error on the same data that were used to fit the model, we will tend to be optimistic unless we adjust for the parameters estimated in the fitting. Roughly speaking, we can think of observed data as supplying n "degrees of freedom" that can be used for parameter estimation, and a regression with k coefficients is said to use up k of these degrees of freedom. The fewer degrees of freedom that remain at the end, the larger the required adjustment for overfitting, as we discuss more carefully in Section 11.8. The degrees of freedom in the above-mentioned χ^2 distribution corresponds to uncertainty in the estimation of the residual error in the model, which in turn can be shown to map to overfitting adjustments in prediction.

Confidence intervals from the *t* distribution

The t distribution is a family of symmetric distributions with heavier tails (that is, a greater frequency of extreme values) compared to the normal distribution. The t is characterized by a center, a scale, and a degrees of freedom parameter that can range from 1 to ∞. Distributions in the t family with low degrees of freedom have very heavy tails; in the other direction, in the limit as degrees of freedom approach infinity, the t distribution approaches the normal.

When a standard error is estimated from n data points, we can account for uncertainty using the t distribution with $n-1$ degrees of freedom, calculated as n data points minus 1 because the mean is being estimated from the data. 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:

```
n <- length(y)
estimate <- mean(y)
se <- sd(y)/sqrt(n)
int_50 <- estimate + qt(c(0.25, 0.75), n-1)*se
int_95 <- estimate + qt(c(0.025, 0.975), n-1)*se
```

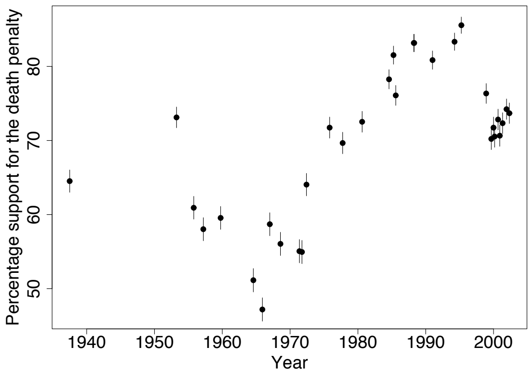

Figure 4.3 *Illustration of visual comparison of confidence or uncertainty intervals. Graph displays the proportion of respondents supporting the death penalty (estimates ±1 standard error—that is, 68% confidence intervals—under the simplifying assumption that each poll was a simple random sample of size 1000), from Gallup polls over time.*

The difference between the normal and t intervals will only be apparent with low degrees of freedom.

Inference for discrete data

For nonbinary discrete data, we can simply use the continuous formula for the standard error. 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 create the data as a vector of length 1000 in R:

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

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

Linear transformations

To get confidence intervals for a linearly transformed parameter, simply transform the intervals. In the above example, the 95% interval for the number 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 $[520\,000, 620\,000]$.

Comparisons, visual and numerical

Example: Death penalty opinions

Uncertainties can often be compared visually, as in Figure 4.3, 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.[2] 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 $\sqrt{(1.4\%)^2 + (1.3\%)^2} = 1.9\%$.

Weighted averages

Confidence intervals for other derived quantities can be determined by appropriately combining the separate means and variances. Suppose that separate surveys conducted in France, Germany, Italy, and

[2] Data and code for this example are in the folder `Death`.

other countries yield estimates of 0.55 ± 0.02, 0.61 ± 0.03, $0.38 \pm 0.03, \ldots$, for some opinion question. The estimated proportion for all adults in the European Union is $\frac{N_1}{N_{\text{tot}}}0.55 + \frac{N_2}{N_{\text{tot}}}0.61 + \frac{N_3}{N_{\text{tot}}}0.38 + \cdots$, where $N_1, N_2, N_3, \ldots$ are the total number of adults in France, Germany, Italy, …, and N_{tot} is the total number in the European Union. Put this all together, and the standard error of this weighted average becomes $\sqrt{(\frac{N_1}{N_{\text{tot}}}0.02)^2 + (\frac{N_2}{N_{\text{tot}}}0.03)^2 + (\frac{N_3}{N_{\text{tot}}}0.03)^2 + \cdots}$.

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

```
W <- N/sum(N)
weighted_avg <- sum(W*p)
se_weighted_avg <- sqrt(sum((W*se)^2))
interval_95 <- weighted_avg + c(-2,2)*se_weighted_avg
```

4.3 Bias and unmodeled uncertainty

The inferences discussed above are all contingent on the model being true, with unbiased measurements, random samples, and randomized experiments. But real data collection is imperfect, and where possible we should include the possibility of model error in our inferences and predictions.

Bias in estimation

Roughly speaking, we say that an estimate is *unbiased* if it is correct on average. For a simple example, consider a survey, a simple random sample of adults in the United States, in which each respondent is asked the number of hours he or she spends watching television each day. Assuming responses are complete and accurate, the average response in the *sample* is an unbiased estimate of the average number of hours watched in the *population*. Now suppose that women are more likely than men to answer the survey, with nonresponse depending only on sex. In that case, the sample will, on average, overrepresent women, and women on average watch less television than men; hence, the average number of hours watched in the sample is now a *biased* estimate of the proportion in the population. It is possible to correct for this bias by reweighting the sample as in Section 3.1; recognizing the existence of the bias is the first step in fixing it.

In practice, it is typically impossible to construct estimates that are truly unbiased, because any bias correction will itself only be approximate. For example, we can correct the bias in a sample that overrepresents women and underrepresents men, but there will always be other biases: the sample might overrepresent white people, more educated people, older people, and so on.

To put it another way, bias depends on the sampling distribution of the data, which is almost never exactly known: random samples and randomized experiments are imperfect in reality, and any approximations become even more tenuous when applied to observational data. Nonetheless, a theoretically defined sampling distribution can still be a helpful reference point, and so we speak of the bias and variation of our estimates, recognizing that these are defined relative to some assumptions.

Adjusting inferences to account for bias and unmodeled uncertainty

Consider the following example: a poll is conducted with 600 respondents to estimate the proportion of people who support some political candidate, and your estimate then has a standard error of approximately $0.5/\sqrt{600} = 0.02$, or 2 percentage points. Now suppose you could redo the survey with 60 000 respondents. With 100 times the sample size, the standard error will be divided by 10, thus the estimated proportion will have a standard error of 0.002, or 0.2 percentage points, much too low to use as a measure of uncertainty for the quantity of interest.

What is wrong with taking a survey with $n = 60\,000$ and saying that the support for the candidate is 52.5% ± 0.2%, or reporting a confidence interval of [52.1, 52.9]? The problem here is that 0.002 is the scale of the sampling error, but there are many other sources of uncertainty, including nonsampling error (coming from the sample not being representative, because different people can choose to answer the survey), systematic differences between survey respondents and voters, variation in opinion over time, and inaccurate survey responses. All these other sources of error represent unknown levels of bias and variance. From past election campaigns, we know that opinions can easily shift by a percentage point or two from day to day, so 0.2% would represent meaningless precision.

Survey respondents are not balls drawn from an urn, and the probabilities in the "urn" are changing over time. In other examples, there are problems with measurement or with the assumption that treatments are randomly assigned in an experiment or observational study. How can we account for sources of error that are not in our statistical model? In general, there are three ways to go: improve data collection, expand the model, and increase stated uncertainty.

Data collection can be improved using more careful measurement and sampling: in the polling example, this could be done by collecting data at different places and times: instead of a single massive survey of 60 000 people, perform a series of 600-person polls, which will allow you to estimate sources of variability other than sampling error.

Models can be expanded in many ways, often using regression modeling as described later in this book. For example, instead of assuming that survey respondents are a simple random sample, we can divide the population into demographic and geographic categories and assume a simple random sample within each category; see Section 17.1. This model is still far from perfect, but it allows us to reduce bias in estimation by adjusting for certain known differences between sample and population. Similar modeling can be performed to adjust for differences between treatment and control groups in a causal analysis, as discussed in Chapter 20.

Finally, when all else fails—which it will—you can increase your uncertainty to account for unmodeled sources of error. We typically assume errors are independent, and so we capture additional uncertainty by adding variances. The variance is the square of the standard deviation; see Section 3.5. For example, in polling for U.S. elections, nonsampling error has been estimated to have a standard error of approximately 2.5 percentage points. So if we want to account for total uncertainty in our survey of 600 people, we would use a standard error of $\sqrt{2^2 + 2.5^2} = 3.2$ percentage points, and for the survey of 60 000 people, the standard error would be $\sqrt{0.2^2 + 2.5^2} = 2.51$ percentage points. This formula shows that very little would be gained by increasing the sample size in this way.

More generally, we can think of total uncertainty as $\sqrt{S_1^2 + S_2^2}$, where S_1 is the standard error (that is, the standard deviation of an estimate) from the model, and S_2 represents the standard deviation of independent unmodeled uncertainty. The mathematics of this expression implies that it will typically be most effective to reduce the *larger* of these two quantities. For example, suppose that $S_1 = 2$ and $S_2 = 5$, so we start with a total uncertainty of $\sqrt{2^2 + 5^2} = 5.5$. If we reduce the first source of error from 2 to 1, that takes us down to $\sqrt{1^2 + 5^2 = 5.1.}$ But if we reduce the second source of error from 5 to 4, that reduces the total uncertainty to $\sqrt{1^2 + 4^2} = 4.1$. In this case, a 20% reduction in the larger error was more effective than a 50% reduction in the smaller error.

As noted above, unmodeled error can be decisive in many real problems. Why, then, in this book do we focus on quantification of error within models? The simple answer is that this is what we can do: modeled error is what statistical methods can handle most easily. In practice, we should be aware of sources of errors that are not in our models, we should design our data collection to minimize such errors, and we should set up suitably complex models to capture as much uncertainty and variation as we can. Indeed, this is a key role of regression: adding information to a model to improve prediction should also allow us to better capture uncertainty in generalization to new data. Beyond this, we recognize that our inferences depend on assumptions such as representativeness and balance (after adjusting for predictors) and accurate measurement, and it should always be possible to increase standard errors and widen interval estimates to account for additional sources of uncertainty.

4.4 Statistical significance, hypothesis testing, and statistical errors

One concern when performing data analysis is the possibility of mistakenly coming to strong conclusions that do not replicate or do not reflect real patterns in the underlying population. Statistical theories of hypothesis testing and error analysis have been developed to quantify these possibilities in the context of inference and decision making.

Statistical significance

A commonly used decision rule that we do *not* recommend is to consider a result as stable or real if it is "statistically significant" and to take "non-significant" results to be noisy and to be treated with skepticism. For reasons discussed in this section and the next, we prefer not to focus on statistical significance, but the concept is important enough in applied statistics that we define it here.

Statistical significance is conventionally defined as a p-value less than 0.05, relative to some *null hypothesis* or prespecified value that would indicate no effect present, as discussed below in the context of hypothesis testing. For fitted regressions, this roughly corresponds to coefficient estimates being labeled as statistically significant if they are at least two standard errors from zero, or not statistically significant otherwise.

Speaking more generally, an estimate is said to be not statistically significant if the observed value could reasonably be explained by simple chance variation, much in the way that a sequence of 20 coin tosses might happen to come up 8 heads and 12 tails; we would say that this result is not statistically significantly different from chance. In that example, the observed proportion of heads is 0.40 but with a standard error of 0.11—thus, the data are less than two standard errors away from the null hypothesis of 50%.

Hypothesis testing for simple comparisons

We review the key concepts of conventional hypothesis testing with a simple hypothetical example. A randomized experiment is performed to compare the effectiveness of two drugs for lowering cholesterol. The mean and standard deviation of the post-treatment cholesterol levels are $\bar{y}_T$ and s_T for the n_T people in the treatment group, and $\bar{y}_C$ and s_C for the n_C people in the control group.

Estimate, standard error, and degrees of freedom. The parameter of interest here is $\theta = \theta_T - \theta_C$, the expectation of the post-test difference in cholesterol between the two groups. Assuming the experiment has been done correctly, the estimate is $\hat{\theta} = \bar{y}_T - \bar{y}_C$ and the standard error is $\mathrm{se}(\hat{\theta}) = \sqrt{s_C^2/n_C + s_T^2/n_T}$. The approximate 95% interval is then $[\hat{\theta} \pm t^{0.975}_{n_C+n_T-2} * \mathrm{se}(\hat{\theta})]$, where $t^{0.975}_{df}$ is the 97.5th percentile of the unit t distribution with df degrees of freedom. In the limit as $df \rightarrow \infty$, this quantile approaches 1.96, corresponding to the normal-distribution 95% interval of ±2 standard errors.

Null and alternative hypotheses. To frame the above problem as a hypothesis test problem, one must define *null* and *alternative* hypotheses. The null hypothesis is $\theta = 0$, that is, $\theta_T = \theta_C$, and the alternative is $\theta \neq 0$, that is, $\theta_T \neq \theta_C$.

The hypothesis test is based on a *test statistic* that summarizes the deviation of the data from what would be expected under the null hypothesis. The conventional test statistic in this sort of problem is the absolute value of the t-score, $t = |\hat{\theta}|/\mathrm{se}(\hat{\theta})$, with the absolute value representing a "two-sided test," so called because either positive or negative deviations from zero would be noteworthy.

p-value. In a hypothesis test, the deviation of the data from the null hypothesis is summarized by the *p-value*, the probability of observing something at least as extreme as the observed test statistic. For this problem, under the null hypothesis the test statistic has a unit t distribution with ν degrees of freedom. In R, we can compute the p-value by doing this: `2*(1 - pt(abs(theta_hat)/se_theta, n_C+n_T-2))`. If the standard deviation of θ is known, or if the sample size is large, we can use the normal distribution (also called the z-test) instead. The factor of 2 in the previous expression corresponds to a *two-sided test* in which the hypothesis is rejected if the observed difference is too much higher or too much lower than the comparison point of 0. In common practice, the null hypothesis is said to be "rejected" if the p-value is less than 0.05—that is, if the 95% confidence interval for the parameter excludes zero.

Hypothesis testing: general formulation

In the simplest form of hypothesis testing, the null hypothesis H_0 represents a particular probability model, $p(y)$, with potential replication data y^{rep}. To perform a hypothesis test, we must define a test statistic T, which is a function of the data. For any given data y, the p-value is then $\Pr(T(y^{\text{rep}}) \geq T(y))$: the probability of observing, under the model, something as or more extreme than the data.

In regression modeling, testing is more complicated. The model to be fit can be written as $p(y|x, \theta)$, where θ represents a set of parameters including coefficients, residual standard deviation, and possibly other parameters, and the null hypothesis might be that some particular coefficient of interest equals zero. To fix ideas, consider the model $y_i = a + bx_i + \text{error}_i$, where the errors are normally distributed with mean 0 and standard deviation σ. Then θ is the vector (a, b, σ). In such settings, one might test the hypothesis that $b = 0$; thus, the null hypothesis is *composite* and corresponds to the regression model with parameters $(a, 0, \sigma)$ for any values of a and σ. The p-value, $\Pr(T(y^{\text{rep}}) \geq T(y))$, then depends on a and σ, and what is typically done is to choose the maximum (that is, most conservative) p-value in this set. To put it another way, the hypothesis test is performed on the null distribution that is closest to the data.

Much more can be said about hypothesis testing. For our purposes here, all that is relevant is how to interpret and compute p-values for simple comparisons, and how to understand the general connections between statistical significance, p-values, and the null hypothesis.

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) can be directly tested by fitting the model that includes the parameter in question and examining the corresponding 95% interval. If the interval excludes zero (or the specified fixed value), then the hypothesis is said to be rejected at the 5% level.

Testing whether two parameters are equal is equivalent to testing whether their difference equals zero. We can 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 ± 2 percentage points, then the magnitude of this estimated difference is probably as important as that the confidence interval for the change excludes zero.

The hypothesis of whether a parameter is positive is directly assessed via its confidence interval. 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.

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. This wording may feel cumbersome but we need to be careful, as it is a common mistake for researchers to act as if an effect is negligible or zero, just because this hypothesis cannot be rejected from data at hand.

Type 1 and type 2 errors and why we don't like talking about them

Statistical tests are typically understood based on *type 1 error*—the probability of falsely rejecting a null hypothesis, if it is in fact true–and *type 2 error*—the probability of *not* rejecting a null hypothesis that is in fact false. But this paradigm does not match up well with much of social science, or science more generally.

A fundamental problem with type 1 and type 2 errors is that in many problems we do not think the null hypothesis can be true. For example, a change in law will produce *some* changes in behavior; the question is how these changes vary across people and situations. Similarly, a medical intervention will work differently for different people, and a political advertisement will change the opinions of some people but not others. In all these settings, one can imagine an average effect that is positive or negative, depending on whom is being averaged over, but there is no particular interest in a null hypothesis of no effect. The second concern is that, in practice, when a hypothesis test is rejected (that is, when a study is a success), researchers and practitioners report, and make decisions based on, the point estimate of the magnitude and sign of the underlying effect. So, in evaluating a statistical test, we should be interested in the properties of the associated effect-size estimate, conditional on it being statistically significantly different from zero.

The type 1 and 2 error framework is based on a deterministic approach to science that might be appropriate in the context of large effects (including, perhaps, some of the domains in which significance testing was developed in the early part of the last century), but it is much less relevant in modern social and biological sciences with highly variable effects.

Type M (magnitude) and type S (sign) errors

With these concerns in mind, we prefer the concepts of type S ("sign") and type M ("magnitude") errors, both of which can occur when a researcher makes a *claim with confidence* (traditionally, a p-value of less than 0.05 or a confidence interval that excludes zero, but more generally any statement that is taken as strong evidence of a positive effect). A *type S error* occurs when the sign of the estimated effect is of the opposite direction as the true effect. A *type M error* occurs when the magnitude of the estimated effect is much different from the true effect. A statistical procedure can be characterized by its type S error rate—the probability of an estimate being of the opposite sign of the true effect, conditional on the estimate being statistically significant—and its expected exaggeration factor—the expected ratio of the magnitude of the estimated effect divided by the magnitude of the underlying effect.

When a statistical procedure is noisy, the type S error rate and the exaggeration factor can be large. In Section 16.1 we give an example where the type S error rate is 24% (so that a statistically significant estimate has a one-quarter chance of being in the wrong direction) and the expected exaggeration factor is 9.5.

In quantitative research we are particularly concerned with type M errors, or exaggeration factors, which can be understood in light of the "statistical significance filter." Consider any statistical estimate. For it to be *statistically significant*, it has to be at least two standard errors from zero: if an estimate has a standard error of S, any publishable estimate must be at least $2S$ in absolute value. Thus, the larger the standard error, the higher the estimate *must* be, if it is to be published and taken as serious evidence. No matter how large or small the underlying effect, the minimum statistically significant effect size *estimate* has this lower bound. This selection bias induces type M error.

Hypothesis testing and statistical practice

We do not generally use null hypothesis significance testing in our own work. In the fields in which we work, we do not generally think null hypotheses can be true: in social science and public health, just about every treatment one might consider will have *some* effect, and no comparisons or regression coefficient of interest will be *exactly* zero. We do not find it particularly helpful to formulate and test null hypotheses that we know ahead of time cannot be true. Testing null hypotheses is just a matter of data collection: with sufficient sample size, any hypothesis can be rejected, and there is no real point to gathering a mountain of data just to reject a hypothesis that we did not believe in the first place.

That said, not all effects and comparisons are detectable from any given study. So, even though we do not ever have the research goal of rejecting a null hypothesis, we do see the value of checking the consistency of a particular dataset with a specified null model. The idea is that *non-rejection* tells us that there is not enough information in the data to move beyond the null hypothesis. We give an example in Section 4.6. Conversely, the point of *rejection* is not to disprove the null—in general, we disbelieve the null hypothesis even before the start of any study—but rather to indicate that there is information in the data to allow a more complex model to be fit.

A use of hypothesis testing that bothers us is when a researcher starts with hypothesis A (for example, that a certain treatment has a generally positive effect), then as a way of confirming hypothesis A, the researcher comes up with null hypothesis B (for example, that there is a zero correlation between treatment assignment and outcome). Data are found that reject B, and this is taken as evidence in support of A. The problem here is that a *statistical* hypothesis (for example, $\beta = 0$ or $\beta_1 = \beta_2$) is much more specific than a *scientific* hypothesis (for example, that a certain comparison averages to zero in the population, or that any net effects are too small to be detected). A rejection of the former does not necessarily tell you anything useful about the latter, because violations of technical assumptions of the statistical model can lead to high probability of rejection of the null hypothesis even in the absence of any real effect. What the rejection *can* do is motivate the next step of modeling the comparisons of interest.

4.5 Problems with the concept of 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 five pitfalls: two that are obvious and three that are less well understood.

Statistical significance is not the same as practical importance

A result can be statistically significant—not easily explainable by chance alone—but without being large enough to be important in practice. For example, if a treatment is estimated to increase earnings by \$10 per year with a standard error of \$2, this would be statistically but not practically significant (in the U.S. context). Conversely, an estimate of \$10 000 with a standard error of \$10 000 would not be statistically significant, but it has the possibility of being important in practice (and is also consistent with zero or negative effects).

Non-significance is not the same as zero

In Section 3.5 we discussed a study of the effectiveness of arterial stents for heart patients. For the primary outcome of interest, the treated group outperformed the control, but not statistically significantly so: the observed average difference in treadmill time was 16.6 seconds with a standard error of 9.8, corresponding to a 95% confidence interval that included zero and a *p*-value of 0.20. A fair summary here is that the results are uncertain: it is unclear whether the net treatment effect is positive or negative in the general population. It would be inappropriate to say that stents have no effect.

The difference between "significant" and "not significant" is not itself statistically significant

Changes in statistical significance do not themselves necessarily achieve statistical significance. 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 $\sqrt{10^2 + 10^2} = 14$.

Researcher degrees of freedom, p-hacking, and forking paths

Another problem with statistical significance is that it can be attained by multiple comparisons, or multiple potential comparisons. When there are many ways that data can be selected, excluded, and analyzed in a study, it is not difficult to attain a low p-value even in the absence of any true underlying pattern. The problem here is *not* just the "file-drawer effect" of leaving non-significant findings unpublished, but also that any given study can involve a large number of "degrees of freedom" available to the researcher when coding data, deciding which variables to include in the analysis, and deciding how to perform and summarize the statistical modeling. Even if a published article shows just a single regression table, there could well be thousands of possible alternative analyses of the same data that are equally consistent with the posited theory. Researchers can use this freedom to "p-hack" and achieve a low p-value (and thus statistical significance) from otherwise unpromising data. Indeed, with sufficient effort, statistically significant patterns can be found from just about any data at all, as researchers have demonstrated by finding patterns in pure noise, or in one memorable case by finding statistically significant results from a medical imaging study performed on a dead salmon.

Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of potential comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values.

Consider the following testing procedures:

1. Simple classical test based on a unique test statistic, T, which when applied to the observed data yields $T(y)$.
2. Classical test pre-chosen from a set of possible tests: thus, $T(y; \phi)$, with preregistered ϕ. Here, ϕ does *not* represent parameters in the model; rather, it represents choices in the analysis. For example, ϕ might correspond to choices of control variables in a regression, transformations, and data coding and excluding rules, as well as deciding on which main effect or interaction to focus.
3. Researcher degrees of freedom without fishing: computing a single test based on the data, but in an environment where a different test would have been performed given different data; thus $T(y; \phi(y))$, where the function $\phi(\cdot)$ is observed in the observed case.
4. "Fishing": computing $T(y; \phi_j(y))$, for $j = 1, \dots, J$: that is, performing J tests and then reporting the best result given the data, thus $T(y; \phi^{\text{best}}(y))$.

We believe that researchers are commonly doing #3, but the confusion is that, when this problem is pointed out to them, researchers think they are being accused of doing #4. To put it another way, researchers assert that they are not doing #4 and the implication is that they are doing #2. The problem with #3 is that, even without explicit fishing, a researcher can induce a huge number of researcher degrees of freedom and thus obtain statistical significance from noisy data, leading to apparently strong conclusions that do not truly represent the underlying population or target of study and that fail to reproduce in future controlled studies.

Our recommended solution to this problem of "forking paths" is not to compute adjusted p-values but rather to directly model the variation that is otherwise hidden in all these possible data coding and analysis choices, and to accept uncertainty and not demand statistical significance in our results.

The statistical significance filter

A final concern is that statistically significant estimates tend to be overestimates. This is the type M, or magnitude, error problem discussed in Section 4.4. Any estimate with $p < 0.05$ is by necessity at least two standard errors from zero. If a study has a high noise level, standard errors will be high, and so statistically significant estimates will automatically be large, no matter how small the underlying effect. Thus, routine reliance on published, statistically significant results will lead to systematic overestimation of effect sizes and a distorted view of the world.

All the problems discussed above have led to what has been called a replication crisis, in which studies published in leading scientific journals and conducted by researchers at respected universities have failed to replicate. Many different problems in statistics and the culture of science have led to the replication crisis; for our purposes here, what is relevant is to understand how to avoid some statistical misconceptions associated with overcertainty.

Example: A flawed study of ovulation and political attitudes

Example: Ovulation and voting

We demonstrate the last two problems mentioned above—multiple potential comparisons and the statistical significance filter—using the example of a research article published in a leading journal of psychology. The article begins:

> Each month many women experience an ovulatory cycle that regulates fertility. Whereas research finds that this cycle influences women's mating preferences, we propose that it might also change women's political and religious views. Building on theory suggesting that political and religious orientation are linked to reproductive goals, we tested how fertility influenced women's politics, religiosity, and voting in the 2012 U.S. presidential election. In two studies with large and diverse samples, ovulation had drastically different effects on single versus married women. Ovulation led single women to become more liberal, less religious, and more likely to vote for Barack Obama. In contrast, ovulation led married women to become more conservative, more religious, and more likely to vote for Mitt Romney. In addition, ovulatory-induced changes in political orientation mediated women's voting behavior. Overall, the ovulatory cycle not only influences women's politics, but appears to do so differently for single versus married women.

One problem here is that there are so many different things that could be compared, but all we see is some subset of the comparisons. Some of the choices available in this analysis include the days of the month characterized as peak fertility, the dividing line between single and married (in this particular study, unmarried but partnered women were counted as married), data exclusion rules based on reports of menstrual cycle length and timing, and the decision of which interactions to study. Given all these possibilities, it is no surprise at all that statistically significant comparisons turned up; this would be expected even were the data generated purely by noise.

In addition, relative to our understanding of the vast literature on voting behavior, the claimed effects seem implausibly large—a type M error. For example, the paper reports that, among women in relationships, 40% in the ovulation period supported Romney, compared to 23% in the non-fertile part of their cycle. Given that opinion polls find very few people switching their vote preferences during the campaign for any reason, these numbers seem unrealistic. The authors might respond that they don't care about the magnitude of the difference, just the sign, but (a) with a magnitude of this size, we are talking noise (not just sampling error but also errors in measurement), and (b) one could just as easily explain this as a differential nonresponse pattern: maybe liberal or conservative women in different parts of their cycle are more or less likely to participate in a survey. It would be easy enough to come up with a story about that.

As researchers and as evaluators of the research of others, we need to avoid the trap of considering

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 4.4 *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.)*

this sort of small study as providing definitive evidence—even if certain comparisons happen to be statistically significant.

4.6 Example of hypothesis testing: 55,000 residents need your help!

Example: Co-op election

We illustrate the application of statistical hypothesis testing with a story. One day several 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 4.4 shows a subset of the data.[3] 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 4.5. 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 4.6 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 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 during each of the six recorded stages of the count. (For example, from

[3] Data and code for this example are in the folder Coop.

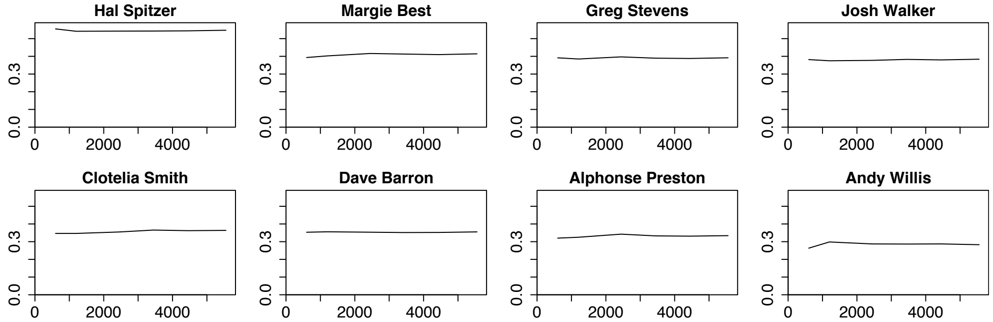

Figure 4.5 *Proportion of votes received by each candidate in the cooperative board election, after each stage of counting: 600, 1200, 2444, . . . , 5553 votes. There were 27 candidates in total; for brevity we display just the leading 8 vote-getters here. The vote proportions appear to be extremely stable over time; this might be misleading, however, since the vote at any time point includes all the previous vote tallies. See Figure 4.6.*

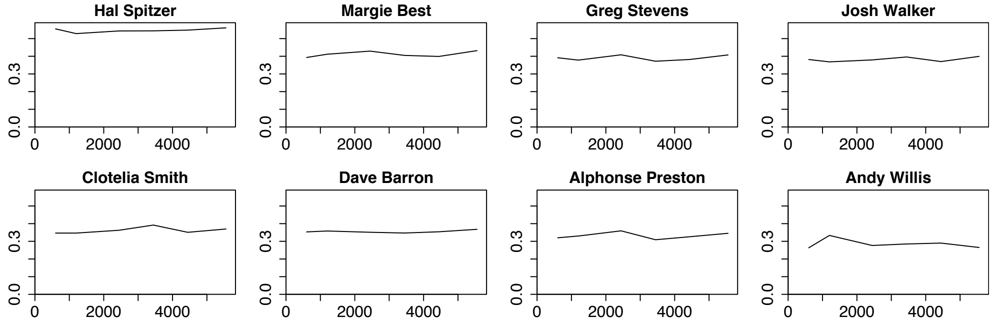

Figure 4.6 *Proportion of votes received by each of the 8 leading candidates in the cooperative board election, at each disjoint stage of voting: the first 600 votes, the next 600, the next 1244, then next 1000, then next 1000, and the final 1109, with the total representing all 5553 votes. The plots here and in Figure 4.5 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 4.4, 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 during each stage. The test statistic for candidate i is then the sample standard deviation of these six values $p_{i1}, \dots, p_{i6}$,

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

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

3. *The theoretical distribution of the data 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. If π_i is the total proportion of voters who would vote for candidate i, then the proportion who vote for candidate i during time period t, p_{it}, follows a distribution with mean π_i and a variance of $\pi_i(1-\pi_i)/n_t$. Under the null hypothesis, the variance of the p_{it}'s across time should on average equal the average of the six corresponding theoretical variances. Therefore, the variance of the p_{it}'s—whose square root is our test statistic—should equal, on average, the theoretical value $\mathrm{avg}_{t=1}^{6}\pi_i(1-\pi_i)/n_t$. 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

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

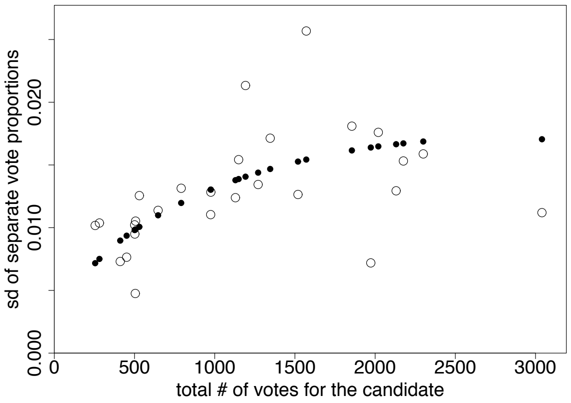

Figure 4.7 *The open circles show, for each of the 27 candidates in the cooperative board election, the standard deviation of the proportions of the vote received by the candidate across the six time points plotted versus the total number of votes received by the candidate. The solid dots show the expected 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.*

4. *Comparing the test statistic to its theoretical distribution.* Figure 4.7 plots the observed and theoretical values of the test statistic for each of the 27 candidates, 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 4.7 for an example. We can then compute for each candidate a summary, called a χ^2 statistic, $\sum_{j=1}^{2} \sum_{t=1}^{6} (\text{observed}_{jt} - \text{expected}_{jt})^2 / \text{expected}_{jt}$, and compare it to its theoretical distribution: under the null hypothesis, this statistic has what is called a χ^2 distribution with $(6-1) * (2-1) = 5$ degrees of freedom.

   Unlike the usual applications of χ^2 testing in statistics, 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 4.4, 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 we can still perform the calculation. The value of the χ^2 statistic is 115. Under the null hypothesis, this would be compared to a χ^2 distribution with $(27 - 1) * (6 - 1) = 130$ degrees of freedom, which has a mean of 130 and standard deviation $\sqrt{2 * 130} = 16.1$. We would not be particularly surprised to see a χ^2 statistic of 115 from this distribution.

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.

4.7 Moving beyond hypothesis testing

Null hypothesis significance testing has all sorts of problems, but it addresses a real concern in quantitative research: we want to be able to make conclusions without being misled by noisy data, and hypothesis testing provides a check on overinterpretation of noise. How can we get this benefit of statistical reasoning while avoiding the overconfidence and exaggerations that are associated with conventional reasoning based on statistical significance?

Here is some advice, presented in the context of the study of ovulation and political attitudes discussed on page 62 but applicable to any study requiring statistical analysis, following the principle noted above that the most important aspect of a statistical method is its ability to incorporate more information into the analysis:

- Analyze *all* your data. For most of their analyses, the authors threw out all the data from participants who were premenstrual or having their period. ("We also did not include women at the beginning of the ovulatory cycle (cycle days 1–6) or at the very end of the ovulatory cycle (cycle days 26–28) to avoid potential confounds due to premenstrual or menstrual symptoms.") That was a mistake. Instead of discarding one-third of their data, they should have included that other category in their analysis. This is true of any study in management or elsewhere: use all of your data to provide yourself and your readers with all the relevant information. Better to anticipate potential criticisms than to hide your data and fear for the eventual exposure.
- Present *all* your comparisons. The paper quoted on page 62 leads us through various comparisons and p-values that represent somewhat arbitrary decisions throughout of what to look for. It would be better to display and analyze more data, for example a comparison of respondents in different parts of their cycle on variables such as birth year, party identification, and marital status, along with seeing the distribution of reported days of the menstrual cycle. In this particular case we would not expect to find anything interesting, as any real underlying patterns will be much less than the variation, but speaking generally we recommend displaying more of your data rather than focusing on comparisons that happen to reach statistical significance. The point here is not to get an improved p-value via a multiple comparisons correction but rather to see the big picture of the data. We recognize that, compared to the usual deterministically framed summary, this might represent a larger burden of effort for the consumer of the research as well as the author of the paper.
- Make your data public (subject to any confidentiality restrictions). If the topic is worth studying, you should want others to be able to make rapid progress.

As we discuss starting in Chapter 10, regression models are helpful in allowing us to model varying treatment effects and situation-dependent phenomena. At the same time, good analysis is no substitute for good data collection. In small-sample studies of small effects, often all that a careful analysis will do is reveal the inability to learn much from the data at hand. In addition, we must move beyond the idea that effects are "there" or not, and the idea that the goal of a study is to reject a null hypothesis. As many observers have noted, these attitudes lead to trouble because they deny the variation inherent in the topics we study, and they deny the uncertainty inherent in statistical inference.

It is fine to design a narrow study to isolate some particular features of the world, but you should think about variation when generalizing your findings to other situations. Does $p < 0.05$ represent eternal truth or even a local truth? Quite possibly not, for two reasons. First, uncertainty: when studying small effects, it is very possible for a large proportion of statistically significant findings to be in the wrong direction as well as be gross overestimates of the magnitude of the underlying effect. Second, variation: even if a finding is "real" in the sense of having the same sign as the corresponding comparison in the population, things can easily be different in other populations and other scenarios. In short, an estimated large effect size is typically too good to be true, whereas a small effect could disappear in the noise.

Does this mean that quantitative research is hopeless? Not at all. We can study large differences, we can gather large samples, and we can design studies to isolate real and persistent effects. In such settings, regression modeling can help us estimate interactions and make predictions that more fully account for uncertainty. In settings with weaker data and smaller samples that may be required to study rare but important phenomena, Bayesian methods can reduce the now-common pattern of researchers getting jerked around by noise patterns that happen to exceed the statistical significance threshold. We can move forward by accepting uncertainty and embracing variation.

4.8 Bibliographic note

Agresti and Coull (1998) consider the effectiveness of various quick methods of inference for binomial proportions and propose the confidence interval based on the estimate $(y + 2)/(n + 4)$.

The death penalty example comes from Shirley and Gelman (2015). See Krantz-Kent (2018) for evidence that women watch less television than men. Shirani-Mehr et al. (2018) estimate the variation arising from nonsampling errors in political polls. The voting example in Section 4.6 comes from Gelman (2004b).

For further discussion and references on the problems with statistical significance discussed in Sections 4.4 and 4.5, see de Groot (1956), Meehl (1967, 1978, 1990), Browner and Newman (1987), Krantz (1999), Gelman and Stern (2006), McShane and Gal (2017), and McShane et al. (2019). Type M and type S errors were introduced by Gelman and Tuerlinckx (2000) and further discussed by Gelman and Carlin (2014). Gelman (2018) discusses failures of null hypothesis significance testing in many modern research settings. Simmons, Nelson, and Simonsohn (2011) introduced the terms "researcher degrees of freedom" and "*p*-hacking." Our discussion of forking paths is taken from Gelman and Loken (2014). Orben and Przybylski (2019) dissect a series of controversial research articles and find 604 million forking paths. The salmon scan example comes from Bennett et al. (2009). The discussion of the study of ovulation and political attitudes comes from Gelman (2015a).

Various discussions of the replication crisis in science include Vul et al. (2009), Nosek, Spies, and Motyl (2012), Button et al. (2013), Francis (2013), Open Science Collaboration (2015), and Gelman (2016c).

4.9 Exercises

4.1 *Comparison of proportions*: A randomized experiment is performed within a survey. 1000 people are contacted. Half the people contacted are promised a \$5 incentive to participate, and half are not promised an incentive. The result is a 50% response rate among the treated group and 40% response rate among the control group. Give an estimate and standard error of the average treatment effect.

4.2 *Choosing sample size*: You are designing a survey to estimate the gender gap: the difference in support for a candidate among men and women. Assuming the respondents are a simple random sample of the voting population, how many people do you need to poll so that the standard error is less than 5 percentage points?

4.3 *Comparison of proportions*: You want to gather data to determine which of two students is a better basketball shooter. One of them shoots with 30% accuracy and the other is a 40% shooter. Each student takes 20 shots and you then compare their shooting percentages. What is the probability that the better shooter makes more shots in this small experiment?

4.4 *Designing an experiment*: You want to gather data to determine which of two students is a better basketball shooter. You plan to have each student take N shots and then compare their shooting percentages. Roughly how large does N have to be for you to have a good chance of distinguishing a 30% shooter from a 40% shooter?

4.5 *Sampling distribution*: Download a data file on a topic of interest to you. Read the file into R and order the data by one of the variables.

(a) Use the `sample` function in R to draw a simple random sample of size 20 from this population. What is the average value of the variable of interest in your sample?

(b) Repeat this exercise several times to get a sense of the sampling distribution of the sample mean for this example.

Example: Girl births

4.6 *Hypothesis testing*: 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`. These proportions were used by von Mises (1957) to support a claim that that the sex ratios were less variable than would be expected under the binomial distribution. We think von Mises was mistaken in that he did not account for the possibility that this discrepancy could arise just 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 observed standard deviation of the 24 proportions will not be identical to its theoretical expectation. In this case, is this difference small enough to be explained by random variation? 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 random variable with 23 degrees of freedom; see page 53.

4.7 *Inference from a proportion with $y = 0$*: Out of a random sample of 50 Americans, zero report having ever held political office. From this information, give a 95% confidence interval for the proportion of Americans who have ever held political office.

4.8 *Transformation of confidence or uncertainty intervals*: On page 15 there is a discussion of an experimental study of an education-related intervention in Jamaica. The point estimate of the multiplicative effect is 1.42 with a 95% confidence interval of [1.02, 1.98], on a scale for which 1.0 corresponds to a multiplying by 1, or no effect. Reconstruct the reasoning by which this is a symmetric interval on the log scale:

(a) What is the point estimate on the logarithmic scale? That is, what is the point estimate of the treatment effect on log earnings?

(b) What is the standard error on the logarithmic scale?

4.9 *Inference for a probability*: A multiple-choice test item has four options. Assume that a student taking this question either knows the answer or does a pure guess. A random sample of 100 students take the item, and 60% get it correct. Give an estimate and 95% confidence interval for the percentage in the population who know the answer.

4.10 *Survey weighting*: Compare two options for a national opinion survey: (a) a simple random sample of 1000 Americans, or (b) a survey that oversamples Latinos, with 300 randomly sampled Latinos and 700 others randomly sampled from the non-Latino population. One of these options will give more accurate comparisons between Latinos and others; the other will give more accurate estimates for the total population average.

(a) Which option gives more accurate comparisons and which option gives more accurate population estimates?

(b) Explain your answer above by computing standard errors for the Latino/other comparison and the national average under each design. Assume that the national population is 15% Latino, that the items of interest are yes/no questions with approximately equal proportions of each response, and (unrealistically) that the surveys have no problems with nonresponse.

4.11 *Working through your own example*: Continuing the example from the final exercises of the earlier chapters, perform some basic comparisons, confidence intervals, and hypothesis tests and discuss the relevance of these to your substantive questions of interest.