Chapter 1

Reasoning with Quantitative Evidence

Statistics provides a logical framework and a set of tools that allow us to weigh the evidence contained in quantitative information. This is the primary contribution that statistics has made to human progress. Statistical analyses can be a part of, but are not a substitute for, a reasoned process of drawing conclusions from information. If you find that working through the logic behind a hypothesis test or the construction of a confidence interval, take solace in the fact that it took humans a very long time to figure this stuff out. Mathematicians mastered Fourier series, complex analysis, differential equations, and a host of other difficult problems before they figured out what the average of a sample says about the world. Part of what makes statistical analyses challenging is that most of the inferences we make from quantitative information are inductive rather than deductive; rarely are we able to say that we proved something to be true by collecting data. Most often we talk about evidence supporting one claim rather than another.

The following is a simplified workflow describing how statisticians are trained to use data to answer questions.

1. Formulate a question
2. Design a study and collect data
3. Choose a statistical model for the data
4. Use data to estimate and make inferences about model parameters
5. Make a judgment about the answer to the question based on the evidence

These tasks are all interrelated; the data collected should be relevant to the question, the model should be appropriate for the data and contain parameters

(i.e. unknown components) relevant to the question, and the estimates and inferences are informed by the model assumptions and the data. A detailed discussion of this framework is laid out in. In this course, we will focus mostly on the third and fourth aspects: picking models and estimating parameters, though we will work in some discussion about how to design studies that give us the best chance of answering our questions.

1.1 A Question, Study, and Decision Rule

The Cornell University "Facts" website (`www.cornell.edu/about/facts.cfm`) states that 26% of Cornell students call New York State their region of origin, defined as their home at the time of matriculation. This number represents an overall percentage considering all undergraduate, graduate, and professional students. Probably if we broke the total population of students down into subpopulations, the percentage of NY students would differ. We might ask the **question** *do 26% of students in CALS call NY their region of origin?*. Since it would be difficult for us to commission a census of all CALS students, we might **design a small study** to survey a sample of students, such as the students in this class. Then we could calculate the percentage of NY students in this class, and compare that to the reported overall percentage of 26%. A first question to ask is whether the data we obtain are relevant to the question. Maybe. It depends on whether the population of students in this class is representative of the total population of CALS students, with respect to their region of origin. A surefire way to make sure a sample is representative is to select a random sample, but that is not feasible here, so we are left to grapple with our potentially non-representative sample. Already things are getting a little messy, but we can try to reason about whether the sample is representative enough, and proceed with the study, keeping in mind this potential flaw and an idea of how much it might affect the results.

Let p be the proportion of NY students in CALS. Then we can reframe our question more succinctly as, does $p = 0.26$? Suppose there are n students in the class. Our data consist of $y_1, \ldots, y_n$, where $y_i = 1$ if student i is from NY, and $y_i = 0$ otherwise. Then the proportion of NY students in this class is

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

In this course, we typically denote data with lowercase letters, although there will be exceptions to this convention. Almost certainly, $\widehat{p}$ will not exactly equal 0.26. Even if we did get exactly $\widehat{p} = 0.26$, we still would not be able to claim for sure that $p = 0.26$. At best, we can make statements about whether $p = 0.26$ is plausible or not plausible.

OK, so if we got exactly $\widehat{p} = 0.26$, the conclusion should clearly be that $p = 0.26$ is plausible. But how far away from 0.26 does $\widehat{p}$ need to be in order to claim

that $p = 0.26$ is not plausible? For now, let's leave the precise answer to that question alone, and simply define the *form* that an answer will take. The form of the answer is that $p = 0.26$ is plausible if $|\widehat{p} - 0.26| < c$, and not plausible if $|\widehat{p} - 0.26| \geq c$. Statistical modeling will show its usefulness in helping us decide the threshold c.

1.2 Statistical Model

We all have some understanding of what a model is: a simplified representation of how the world works. Climate models represent the earth system with differential equations and use supercomputers to solve the equations; economic models incorporate the preferences and behaviors of various actors in an economy; species-prey models posit dynamic relationships among animal populations. None are meant to be taken literally as a statement about how the world works. Nonetheless, models are useful tools in the pursuit of understanding how the world works. But what makes a model statistical? This is not an easy question; in 2001, 86 pages in the Annals of Statistics were devoted to an article, discussion, and rejoinder trying to answer it [1]. Thus it is difficult and maybe misleading to give a one sentence definition, but we can try. A statistical model is a family of probability distributions meant to represent an assumption about how data are generated. Statistical models serve to formalize our assumptions about data-generating mechanisms using probability, and probability gives us a mathematics and a language for handling uncertainty. Like other types of models, they are not meant to be taken literally, but they serve as useful tools for evaluating quantitative evidence.

For our region of origin study, let Y_i be a random variable (RV) that takes value 1 with probability p and value 0 with probability $1 - p$, that is

$$P(Y_i = 1) = p, \quad P(Y_i = 0) = 1 - p,$$

and let the Y_i's be identically distributed, i.e. each $Y_1, \ldots, Y_n$ has the same probability distribution. This is a "named" probability distribution known as the Binomial$(1, p)$ distribution. Random variables are different from regular variables in that the "value" of a random variable is represented as a probability distribution rather than a single number, like our data y_i. We typically use uppercase letters for RVs, although again there will be exceptions. We think that Y_i is a reasonable model for y_i because they both can take on only the values 0 and 1, and the model contains a parameter p that, when estimated, can help us answer our question of interest. We say that this is family of probability distributions because each different p gives a different distribution. The statistical model has not been fully specified, however, until we say how the collection of random variables $Y_1, \ldots, Y_n$ are related to one another. This is often a subtly tricky part. For example, students that come from the same region may be more likely to be friends, and since they're friends they may

be more likely to enroll in similar courses. So if we knew that student i and student j were friends, and we knew that $Y_i = 1$, then we might believe that $P(Y_j = 1|Y_j = 1) > p$. For now, we assume that the random variables are independent. Independence has a formal mathematical definition, but here just take it to mean that knowing the realized value of Y_i does not change our beliefs about any other Y_j. The two assumptions of independence and identical distributions are abbreviated as *i.i.d.*, and so our specification of the statistical model can be stated as

$$Y_i \overset{iid}{\sim} \text{Binomial}(1, p) \quad \text{for } i = 1, \ldots, n.$$

1.3 Sampling Distributions

Once we have our statistical model, we can do calculations with it. Of particular use is to calculate the probability distribution of

$$\widehat{p} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Since each Y_i is a random variable, and $\widehat{p}$ is simply a function of the Y_i's, then $\widehat{p}$ is also a random variable. Note that we have just committed a notational crime by using the same letter for $\widehat{p} = 1/n \sum y_i$ and $\widehat{p} = 1/n \sum Y_i$, which are two different things! The first is a function of the data, and simply a number, while the second is a random variable. We have to learn to live with this kind of offensive notational inconsistency because everybody else does it. It is especially bad here because students tend to struggle when thinking about the distinction between the data version and the random version. Unfortunately, the notation makes this harder for us, but luckily, we can usually tell from context whether we are referring to the data version or the random variable version. By the way, we call the data version the estimate, and we call the random variable version the estimator. At least there are two different words for them.

Since the estimator $\widehat{p}$ is a random variable, we can calculate its probability distribution. It's actually easier to calculate the distribution of $S = \sum Y_i$; we'll divide by n later if we need to. The first thing to figure out is what values S can take. It has to be an integer, since it is the sum of integers, and it has to fall between 0 (in the case that no students are from NY) and n (in the case that all students are from NY). Because of the *i.i.d.* assumption, S has a Binomial(n, p) distribution, which has probability mass function (pmf)

$$P(S = k) = \frac{n!}{k!\,(n-k)!} p^k\,(1-p)^{n-k} \quad \text{for } k = 0, \ldots, n.$$

This allows us to calculate the probability that S takes on any value k, as a function of the probability p. Since $P(S = k) = P(\widehat{p} = k/n)$, we can use the pmf to evaluate probabilities involving $\widehat{p}$ as well (from context, we are talking about

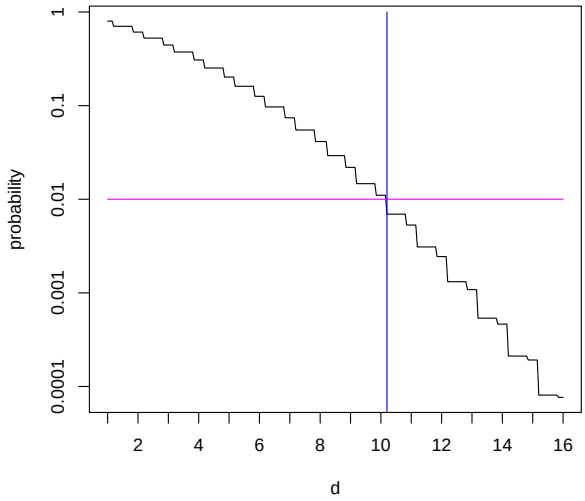

Figure 1.1: Plot of $P(|S - 20.8| > d)$ when $p = 0.26$ as a function of d. The smallest value of d for which the probability is less than 0.01 gives us our plausibility (significance) threshold. Here, that value is 10.2.

the random estimator $\widehat{p}$). *In particular*, we can plug in the hypothesized value $p = 0.26$, pick a value c, and evaluate $P(|\widehat{p} - 0.26| \geq c)$, which is the probability that we deem a true hypothesis to be implausible. Deeming $p = 0.26$ to be implausible when it is true would be a mistake, and so we would like to ensure that this mistake is unlikely to happen.

The larger we make c, the less likely we are to deem a true hypothesis implausible. The standard way of picking c is to say that we would like to make this mistake with probability α or less, where α is a small number–like 0.005, 0.01, or 0.05–and then figure out what c must be to ensure this. Stated mathematically, we find the smallest c such that

$$P(|\widehat{p} - 0.26| \geq c) \leq \alpha.$$

As an example, suppose there are $n = 80$ students in the course. Then the event that $|\widehat{p} - 0.26| \geq c$ is equivalent to the event that $|S - 20.8| \geq 80c := d$. We'll do the calculation in terms of d and then convert back to c by dividing d by 80. Figure 1.1 shows the probabilities as a function of d. We can see that $d = 10.2$ is the smallest value for which the probability drops below 0.01. This means that we deem the hypothesis to be implausible if $|\widehat{p} - 0.26| \geq 10.2/80 = 0.1275$. In other words, if $\widehat{p}$ is between 0.1325 and 0.3875, we deem the hypothesis plausible, which might seem to be a surprisingly large range. This is because we picked a pretty stringent threshold probability of 0.01. A weaker threshold of 0.1 gives the range 0.1825 to 0.3375.

1.4 Null Hypothesis Significance Testing

The previous sections describe an example of null hypothesis significance testing (NHST). In this section, we give NHST a more formal treatment, including a more general view of decision rules and a discussion of statistical power. Once data $y_1, \ldots, y_n$ have been collected and a statistical model P_θ has been chosen, we define the null hypothesis as a statement about the unknown parameter θ, as in

$$H_0 : \theta = \theta_0.$$

It is possible to define more complicated null hypotheses, such as $\theta \geq \theta_0$, but we stick with the simple null hypothesis here. In our example above, the parameter was $\theta = p$, and the null hypothesis was $H_0 : p = 0.26$. The next step is to define a statistic t, which is a function of the data (hence the lowercase letter), and a decision rule, which says

$$D(t) = \begin{cases} \text{Fail to Reject } H_0 & \text{if } t \in A \\ \text{Reject } H_0 & \text{if } t \notin A. \end{cases}$$

Rejecting H_0 is another way of saying that we deem H_0 to be implausible given the data. Failing to reject H_0 means that we deem it to be plausible given the data. In our example above, the statistic was $t = \widehat{p}$, and $A = (0.1325, 0.3375)$. Our decision rule is more general in that t need not represent an estimate of θ, and the set A need not be a single interval. For example, t could be a likelihood ratio or an F statistic. The choice of the set A depends on the significance level α of the test, and A is chosen so that

$$P_{\theta_0}(T \notin A) \leq \alpha.$$

Note that this probability is calculated assuming the null-hypothesized value θ_0, and we plug in the random variable version of the statistic T when we compute probabilities with the decision rule.

It should be noted that not all decision rules are created equally. While they all have the property that the probability of rejecting a true null hypothesis is less than α, some decision rules are better than others in the sense that their probability of rejecting a false null hypothesis could differ. Let θ_1 be a particular value of the parameter, and define

$$\text{power}(\theta_1) = P_{\theta_1}(T \notin A),$$

which is the probability that we reject the null hypothesis when it is false. Keep in mind that power calculations assume that the null is false but are calculated for decision rules determined under an assumption that the null is true. Read the previous sentence again and make sure you understand it.

Author's opinion: NHST has come under fire recently as one of many potential culprits in the reproducibility crisis, in which attempts to reproduce published scientific findings have often either failed or given results that are weaker than originally claimed. My intention here has been to give an explanation of the logic that undergirds NHST. My personal view on the matter is that the logical framework of NHST is sound but that any attempt to define hard "objective" thresholds on claims of discovery will invite people to game the system. This is why I think that, while statistical analyses are an important component of the scientific process, they should not be relied on as the sole determinant of whether data constitute a discovery. Statistical evidence is important but should be weighed against other forms of evidence.

1.5 Confidence Intervals

Hypothesis testing is useful in cases where the truthfulness of a particular claim is the primary goal of the analysis. Is the speed of light in a vacuum equal to 2.99792458×10^8 m/s? Does a drug have zero effect on patients' recovery time? However, in other cases, we simply want the analysis to return a range of plausible values of the parameter. Confidence intervals provide a logical framework for doing this.

Giving the textbook definition, a $(1-\alpha)$ confidence interval is a realization of a random interval that had probability $(1-\alpha)$ of containing the actual value of the parameter under the assumed statistical model for the data, regardless of the actual value of the parameter. The reason for defining the interval in terms of $(1-\alpha)$ will become clear soon.

Let's start with an absurd example to make sure we understand the definition. Suppose we have data $y_1, \ldots, y_{20}$, which we model as

$$Y_i \overset{iid}{\sim} N(\mu, \sigma^2),$$

a normal distribution with mean μ and variance σ^2. We construct a 0.95 confidence interval for μ as follows:

$$ci = \begin{cases} \emptyset & \text{if } y_1 \text{ is smallest} \\ (-\infty, \infty) & \text{otherwise,} \end{cases}$$

where $\emptyset$ is the empty set. No matter what the value of μ, it is not in the confidence interval if y_1 is smallest, and in the confidence interval if not. We use the statistical model to evaluate

$$P(\mu \in CI) = P(Y_1 \text{ is not smallest}) = 1 - P(Y_1 \text{ is smallest}) = 1 - 1/20 = 0.95.$$

This is true no matter what μ is, so ci is a valid confidence interval. In our dataset, suppose that y_8 is smallest. Then $ci = (-\infty, \infty)$, which gives absolutely no information about μ, but technically it is a 0.95 confidence interval.

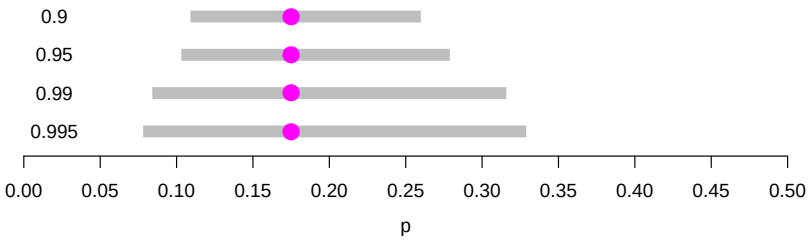

Figure 1.2: Confidence intervals for p in the region of origin example for four different confidence levels.

We can use hypothesis tests to give more useful confidence intervals. Consider the set

$$ci = \{\mu^* | \text{we fail to reject } H_0 : \mu = \mu^* \text{ at level } \alpha\},$$

which should be read as the set of all parameter values μ^* such that we fail to reject $H_0 : \mu = \mu^*$. We could imagine doing a whole bunch of hypothesis tests at level α for different values of μ^*, and if we fail to reject, put that value into the confidence interval. In other words, the confidence interval (a set of plausible values) consists of all the parameter values that we can't rule out.

The derivation to show that ci is a $(1-\alpha)$ confidence interval is confusingly simple. Let μ_0 represent the true value of the parameter. Remember we need to show that $P(\mu_0 \in CI) = 0.95$.

$$\begin{aligned} P(\mu_0 \in CI) &= P(\text{fail to reject } H_0 : \mu = \mu_0 \text{ at level } \alpha) \\ &= 1 - P(\text{reject } H_0 : \mu = \mu_0 \text{ at level } \alpha) \\ &= 1 - \alpha. \end{aligned}$$

The last equality is true because α is the precisely the probability that we reject a true null hypothesis! This method of constructing confidence intervals is sometimes called *inverting* the hypothesis test.

Suppose in the region of origin example that we get $s = 14$ when $n = 80$, and our decision rule is again to reject $H_0 : p = p_0$ when $|s - 80p_0| > d$. It takes a bit of computing, but we can do the hypothesis test for a grid of values for p_0 between 0 and 1 separated by 0.001. Collecting the values for which we fail to reject the null results in the confidence intervals in Figure 1.2. Note that intervals of higher confidence are longer, which happens because smaller α requires more evidence (larger d) to reject. Also, the confidence intervals are not symmetric around the estimate $\widehat{p} = 0.175$. This is because the variance of the sample proportion is $p(1-p)/n$, so values of p near $1/2$ have higher variance, so $p = 0.275$ is more likely to produce $\widehat{p} = 0.175$ than $p = 0.075$ is.

1.6 p-values

The p-value is closely related to the hypothesis test but has a slightly more stringent treatment of the decision rule. One must define an ordering of the statistics, that is, for any two values t_1 and t_2 of the statistic, we need to be able to say whether t_1 or t_2 is "more extreme." The decision rule in our region-of-origin example fits the mold because the decision rule was to reject $H_0 : p = 0.26$ if $t = |\widehat{p} - 0.26| \geq c$. The statistic is a positive number, and larger values correspond to larger deviations from the hypothesized proportion, and thus larger values are more extreme. The p-value is simply

$$\text{p-value} = P_{\theta_0}(T \geq t),$$

the probability–under the null hypothesis–of observing a statistic as extreme or more extreme than the one we did observe. This may look slightly confusing because both T, the random version of the statistic, and t, the actual nonrandom version, appear inside the probability. The probability calculation comes from the probability distribution of T.

Suppose we constructed a hypothesis test with significance α for a decision rule of the form above with $t_0 = c$. Conducting the test is equivalent to checking whether than p-value is greater or smaller than α. To see why this is true, suppose that a and b are two possible values of the statistic and that b is more extreme than a. Then the following inequality is true,

$$P(T \text{ more extreme than } b) \leq P(T \text{ more extreme than } a),$$

because if T is more extreme than b, it also has to be more extreme than a because b is more extreme than a. Suppose that the observed statistic t is less extreme than c. Then of course $P(T \geq t) \geq P(T \geq c) = \alpha$. On the other hand, if t is more extreme than c, $P(T \geq t) \leq P(T \geq c) = \alpha$. This means that if the p-value is less than α we reject, and if the p-value is greater than α we fail to reject.

Bibliography

[1] Peter McCullagh. What is a statistical model? *Annals of statistics*, pages 1225–1267, 2002.