Chapter 2

Randomness

We are beset with randomness. Much of our measurements or observations cannot be predicted exactly. Indeed, why go to the trouble of collecting data if we know already what will result?

At the heart of statistics is a way of thinking about randomness. We presume that, for any observation subject to randomness, there are possible values for the observation, and that those possible values have probabilities[1]. That is, they are more or less likely, or they have a greater or lesser chance or risk of occurring... We may not know all the possible values, and we may not know the probabilities of all the possible values, but we assume that the possible values and their probabilities exist - and we leavfe aside all the philosophical questions about what precisely that might mean.

Please note that we spoke above of “an observation subject to randomness.” This would seem to refer to an observation that could be recorded as a single number. A height, say, or whether or not death occurred, or the date of a default of some sort, for example. But generally, we make observations that yield whole data sets filled with numbers. In this situation, the possible values for the observation are the possibilities for the whole data set. Likewise, the probabilities quantify how likely it is that one might observe a particular possible data set.

For example, suppose that in a clinical trial of patients suffering from some malady, we provide some number n of subjects with a new treatment, and then, for each subject, record *cure* or *not cure*, according to whether or not the patients are cured of the malady or not. Then the possible values of

[1] Strictly speaking, when it comes to the mathematical theory of probability, we say that subsets of possible values have probabilities, rather than that the individual values have probabilities.

our observation would be the sequences of length n with entries that are *cure* or *not cure*. Without knowing more about the nature of the malady and the treatment, and how the patients were obtained, we would be hard pressed to quantify how likely any particular sequence might be. Nevertheless, we hold fast to our assumption that every possible sequence has an associated probability.

Some of this might feel quite natural. If an observation is subject to randomness, well, then certainly there must be more than one possible value for the observation. The idea that there are probabilities associated with each of the possible outcomes is more of a stretch. Certainly, nobody has ever seen a probability. (You have!? What did it look like?) Certainly, only one of the possible outcomes will actually come to pass, too. Why, then, should we believe that there are probabilities associated with all of the various outcomes? The philosophers might argue over these questions, but for our purposes, there really is only one relevant answer: please, if you don't want to, don't believe in probabilities; but also, please, do come over for poker night... just don't forget to stop at the ATM on your way!

Where does randomness comes from? At first blush, our answer will seem terrifically unsatisfactory: randomness comes from sources of variability. This is unsatisfactory in that it begs the question of where the randomness in the sources of variability comes from. Nevertheless, it is a useful answer: it tells us that we should enquire into the various possible sources of variability, and into what might be their effect on the possible values and probabilities for our observations. Such enquiries are key to doing statistical analyses properly.

What would be the sources of variability in the example clinical trial (where *cure or* not cure is recorded for each patient)? Subjects might differ in the severity of their malady. Subjects might differ in their genetic make-up. Their environments would differ, too, in ways that influence their prognosis. There may be errors in detecting or recording whether or not cure has occurred. Also, there may be variation in the patients' health more generally. Subjects may not all be treated in the same way: in many studies, not all patients would receive the treatment; rather, the patients would be randomly assigned to a medication or placebo. All these possible sources could influence what would actually be observed in the sequence of recorded patient outcomes.

One advantage to thinking of randomness as stemming from sources of variability is that it allows us to think in a practical way about whether, when our observations have several components, the several components are independent of each other. When observations are affected by shared sources of variability, then they generally[2] will not be independent. That is, the value of

[2] There is a technical subtlety here behind the choice to use the work "generally." It is possible for components of an observation to share sources of variability, and yet still be independent, in the sense that the value of one component will have no implications for how likely are the various possible values for the other component.

one will have implications for how likely are the various possible values for the other.

In our example clinical trial, some subjects may be siblings, and so share the source of their genetic material and environment. These shared sources of variability might have a common influence on the siblings' outcomes. If whatever source of variability helped the first sibling to achieve cure, that source of variability might very well have the same effect on the second sibling. If we see that one sibling is cured, then we might expect that the other sibling would be cured, too - or at least we might expect that the other sibling would be more likely to be cured. Cure in one sibling would thus not be independent of cure in the other.

Another advantage to thinking of randomness as stemming from sources of variability is that it allows us to think in a practical way about sources of variability that, if known, would help to predict outcomes. Without knowing anything about a patient other than they are participating in our example clinical trial, we would suppose that their probability of cure is the same as that of anyone else in the population from which the patient was drawn. If we knew, however, that they had only a mild version of the malady, had no genetic or environmental risk factors known to interfere with the effect of the treatment, and received the treatment rather than placebo, then we might presume that they would have a higher probability of cure. Knowledge of the values of the patient's sources of variability could thus help in predicting patients' outcomes.

It should probably go without saying that probabilities, by convention, range from zero to one. A probability of zero corresponds to impossibility[3], a probability of one corresponds to certainty. A probability of a half corresponds to as likely a not. A probability of one third, corresponds to half as likely as not. If there are n possible values for an outcome, and all outcomes are equally likely, the probability of any one outcome would be $1/n$.

[3] There is another technicality here. Although our observations are only ever recorded as one of a finite set of discrete possible values, in the mathematical theory of probability, we sometimes think of the underlying phenomenon as taking on values in a continuum, with our observations only constrained to a finite set by the limitations of our measurement and recording technology. In these situations, the probability of any particular outcome among the uncountably infinite set of possible values in the continuum might very well be zero, even though it would not make sense to view every outcome as impossible.

Problem Set

1. In the text, there was mention of the distinction between probabilities for a single measurement and probabilities for a whole data set of measurements. There was also mention of the concept of independence. What can be said about the relationship between these two?

2. Consider three variations on an experimental design. A researcher uses a ruler to measure the width of a desk several times, and each time records the width; the researcher randomly chooses from an array of rulers each time measuring and recording the width of the desk; the researcher randomly chooses from the array of labelled rulers, and records the measurement and the label on the ruler. How do you think about the sources of variability in these experiments? What if the various rulers had been studied extensively to the point where the various imperfections in the rulers were well known and available to the researcher? Back when someone was studying each of the labelled rulers to understand their various imperfections, would they have thought differently about the sources of variability in their measurements?

3. Hearken back to your introductory probability and statistics course to describe, in your own words, the distinction between *accuracy* and *precision.* Try to include in your answer some mention of what kinds of things have accuracy or precision. You might find it useful to make use of the concepts of *expectation*, *bias*, and *variability.*

4. Formulate a short checklist for thinking about all of the sources of variability that might influence the data available for a statistical analysis.

Chapter 3

Distributions

This chapter is mostly a glossary of sorts. Recall that where there is randomness, observations have a probabilistic behavior. That is, there must be more than one possible outcome, and the various possible outcomes have probabilities. When discussing the probabilistic behavior of an observation, it will prove helpful to have a kind of short-hand terminology. This chapter provides a very small glossary of such terminology. It also affords an opportunity to review some of the concepts from elementary statistics that are requisite for an understanding of regression models and methods.

When an observation can take on only two possible values, we say that the observation is *dichotomous*. When those two values are 0 and 1, we say that the observation has a *Bernoulli* distribution. When an observation has a Bernoulli distribution, there is a probability associated with the outcome 1. To completely characterize the the probabilistic behavior of a Bernoulli observation, we need to specify that probability. Thus, we might, for example, speak of a Bernoulli outcome with success probability 0.3. Sometimes it is useful to have notation for the success probability without specifying the value of the probability. In such cases, it is common to refer to the probability as p, and we would say that our observation is "Bernoulli p."

If in our example of the clinical trial, cure and absence of cure were coded as 1 and 0 in in a data set, then the recorded outcome for each separate individual would be a Bernoulli observation. We would hope that the p for the patients' Bernoulli distributions were close to 1. That is, we would hope for the probability of cure to be high.

When we make several Bernoulli observations, say n, all of them independent (that is, with no sources of variability in common), all with the same success probability (that is, all with the same p, and we count the number of positive outcomes, the resulting count observation is said to have a *binomial* distribution. To completely specify a binomial distribution we have the specify the success probability for the underlying variables. We also have to specify the number of observations. That is, we would have to specify p and n.

In our example clinical trial, the sum of the observations - that is, the number of observations with outcome value 1 - might have a binomial distribution. It would have a binomial distribution if each patient had the same p and cure or not cure in the patients were all independent.

When Binomial distribution has a large n and correspondingly small p, the probabilities of the various counts can be closely approximated by the so-called Poisson distribution. When the data follow a Poisson distribution, the probability that the observed count is a value k takes the form

$$\frac{e^{-\lambda}\lambda^k}{k!},$$

where λ is given by np.

The Poisson probabilities are a good approximation to a count formed by adding up Bernoulli variables even if the Bernoulli variables do not all have the same p. As long as every p is small and the number of Bernoulli observations is large, and the Bernoulli variables are independent, then the probabilities of the various possible counts are well approximated by the Poisson probabilities.

Poisson distributions are relevant where a count is the number of events occurring in some interval of time or some interval, or area, or volume of space. Not every such count has an approximate Poisson distribution, however. Here are conditions that ensure that the Poisson distribution is applicable: first, it should be possible that the interval or area or volume may be thought of as many, many, non-overlapping, tiny sub-intervals or sub-areas, or sub-volumes; second, whether or not there is an event in any sub-interval or sub-area or sub-volume is only negligibly dependent on what happens in any of the others; third, the probability of an event in any sub-interval or sub-area or sub-volume is proportional to the size of the sub-interval or sub-area or sub-volume; and fourth, the probability of more than one event in any one sub-interval is negligible. It is not hard to see why counts formed under these conditions should have a Poisson distribution: the Bernoulli observations recording whether or not there is an event in the sub-intervals or the sub-areas or the sub-volumes are many, independent, and the p for each, being proportional to the size of the sub-interval or sub-area or sub-volume, is small.

Now consider the situation in which our observation may be thought of as a sum of many components, each component independent of all of the others, each component more or less commensurate[1]. Suppose also that the

[1] What is meant by commensurate here? Roughly, for the approximation that is about to

number of such components is very large. The distribution of the sum takes as its sources of variability all of the components. Thus the distribution of the sum depends on the distributions of the components. It is a remarkable and important result in probability theory that under these conditions (large n and independent commensurate components), the distribution of the sum of the has a distribution that depends on the distributions of the various components through only two numbers.

To explain this more fully, we need to digress to the concepts of the expectation and variance of a distribution. First, the expectation of a distribution of an observation. Imagine a number line. Imagine placing bits of mass along the number line, one bit at each of the possible values for the observation, the magnitude of each mass equal to the probability of the corresponding value. The balancing point of that number line is the so-called expectation of the distribution of the observation[2].

Now, the variance of the distribution of an observation. Imagine that you take your observation, subtract from it the expectation of the observation, and then square the result. That square is random: it inherits randomness from the original observation. Therefore, the square also has a distribution. The distribution of the square therefore has an expectation. We call the expectation of the squared deviation of the observation from its expectation the variance[3]. (A further bit of nomenclature: we call the square root of the variance the standard deviation.)

It is useful to spend a moment on what the expectation and variance of an observation tells about the probabilistic behavior of the variable. The expectation maybe thought of as a kind of middle value for the observation. It may not be a possible value, but the possible values, weighted by their probabilities, balance out around the expectation. In this sense, the expectation provides a crude summary of the distribution, one that indicates whether, on the whole, we should think of the observation as, in a general way, positive or negative, large or small.

The variance is the expectation of the distribution of the squared deviation of the observation from its expectation. If an observation is likely far from its expectation, then the squared deviation is likely to be large. If the probabilistic behavior of an observation is such that it is likely close to its expectation, then the squared deviation is likely to be small. In the former case, the expectation of the squared deviation would tend to be large (the possible values of the squared deviation, weighted by their probabilities, would tend to be large).

be discussed here, the so-called normal or Gaussian approximation, it is sufficient that the possible values of the components all are bounded above and below by some fixed positive constant. This prerequisite may be relaxed, as may be the prerequisite that all the components are independent.

[2] We sometimes say that it is the expectation of the observation itself.

[3] The variance of the distribution, or or the variance of the observation.

In the later case, the expectation of the squared deviation would tend to be small. Thus the variance indicates how likely it is that an observation would deviate substantially from its expectation. If one is observing a highly labile phenomenon, the variance of one's observation would tend to be large. If one is observing a largely static phenomenon, the variance of one's observation would tend to be small.

While we are at it, we might as well mention the covariance and correlation. The covariance between two observations subject to randomness is the expectation of the product of the difference between each observation and their expectation. If the probability model governing two observations is such that the two observations tend to move together - that is, if when one is larger than its expectation, the other tends to be so as well, and when one is smaller than it's expectation, so, also, does the other, then the product would have a higher probability of being positive, and the covariance would be positive. If instead the two observations tended to move against each other, then the covariance would be negative. In this way, the covariance quantifies the degree to which the two observations are positively or negatively associated.

The magnitude of the covariance reflects not just how associated two observations are, but also the extent to which they tend to vary around their expectations. Highly variable observations would tend to have higher covariances. This means that a change in the units in which one records one or the other observations would have an effect on the magnitude of the covariance. If height is measured in inches, say, instead of miles, the probability model for height would have substantial probabilities ranging from zero to seventy or eighty... while in miles, the variability in measured height would be restricted to little more than 0.001. The change in units would have a proportionate effect on the covariance. A measure of association that is invariant to choice of units is the correlation. The correlation is the ratio of the covariance to the product of the standard deviations.

We are now ready to return to the approximate distribution of a sum. The expectation of the sum is equal to the sum of the expectations of the components. (This is true for any sum.) The variance of the sum is equal to the sum of the variances. (This is true for any sum in which the components are all independent.) Let μ denote the expectation of the sum. Let σ^2 denote the variance of the sum. Here is the approximation to the distribution of the sum when the number of components is large: the probability that the sum lies in any interval[4] is well approximated by the area above the interval on a number line and below a curve plotted above the number line, the curve corresponding

[4] Strictly speaking, the result is that for a fixed interval, the probability that a normalized version of the sum (the sum is normalized by subtracting μ and dividing by σ), as the number of summands increases, the probability that the sum falls into the interval becomes closer and closer to the area under the curve of the standard normal curve, the curve with $\mu = 0$ and $\sigma^2 = 1$.

to the function

$$f(y) \ = \ \frac{e^{-(y-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}.$$

That sums with many summands have distributions that are well approximated by these so-called bell-shaped curves is known as the central limit theorem. The central limit theorem is useful for approximating the distribution of observations from nature that are formed as an aggregate of many underlying independent commensurate sources of variability. They are also useful for approximating the distribution of variables that a statistician might create in the course of a data analysis - as long as those variables are created as a sum of many independent commensurate observations.

While we are on the topic of limit theorems, we might also mention the law of large numbers. We said that sum of independent commensurate observations has, for large n, approximately a bell-shaped density function. We might divide the sum by the number of summands - that is, we might form the sample average of the independent observations. What can be said about the distribution of the sample average? It turns out that, with large sample sizes, the average has very little variability: with high probability, the random sample average will be very close to the expectation of the sample average. When the observations in the sample all have the same expectation, then the expectation of the sample average is the common expectation of the observations. For this reason, with large n, a sample average is generally a good surrogate for the common expectation of the observations in the sample average.

The last kind of distribution in our small glossary is the exponential. Exponential distributions correspond to the probabilistic behavior of time-to-event observations with the memoryless property. A time-to-event observation is said to have the memoryless property if for any passage of time t, for any additional time a, the probability that the event will not occur within time a, given that it has not yet occurred by time t does not depend on t. If our observation were one's time-to-death, say, the observation would have the memoryless property if the probability that one would survive for some specific additional length of time does not depend on how old one is.

The probability that an observation with an exponential density falls within an interval of time is given by the area above the interval and below a curve of the form

$$f(t) \ = \ \lambda e^{-\lambda t}.$$

The value λ determines whether the time to event tends to be long or short. When λ is large, the time-to-event is likely to be short, and when λ is small, the time-to-event is likely to be long.

Computation Questions

1. Form a data set that contains the probabilities for each of the possible values of a Binomial observation with n=20 and p=0.5. Your data set should have two columns, one containing the binomial count, and one containing the corresponding probability. Print out the data set.

2. Modify your data set to add another column containing the product of the count and the probability. Compute the sum of the probabilities, and the sum of the products. Comment on what you observed.

3. Do the same thing for for a Poisson observation with $\lambda = 2$. (You can't make a data set that contains probabilities for each of the possible values of a Poisson variable - make sure you understand why not - so for this exercise, just include the values 0 through 5.

4. Try the previous problem again, but with 40 instead of 5. Comment on what you observe.

5. For a data set with two columns, one for X values and one for Y values[5]. You may create X in whatever manner you choose. Let n, the number of rows in your data set, be 15. Create Y as follows,

$$Y_i \;=\; 3 \;+\; 2X_i \;+\; \epsilon_i,$$

   where ϵ_i has a normal distribution with expectation zero and variance 1, and all of the different normal variables are independent. Find the sample average of the X values, $\bar{X}$, and the sample average of the Y values, $\bar{Y}$. Now compute the sums

$$\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) \text{ and } \sum_{i=1}^{n}(X_i - \bar{X})^2.$$

   Report their ratio. Redo the experiment, this time with $n = 150$. And with $n = 5000$. Comment on what you see.

6. Form a data set with 10,000 independent random exponential variables with parameter 1. Compute the sample average of your variables. Now restrict your data set to those observations that are greater than 1. Subtract 1 from each observation. Compute the sample average of the resulting data set. Do it again. And again. And one more time. Comment on your results. Now do the same thing, except this time use normal variables with expectation 0 and variance 1. Compare and contrast the results for normal and for exponential observations.

[5] We will use the notation Y_i for the i^{th} Y value - that is, the value on the i^{th} row of the data set. We will use analogous notation for X values.