Appendix A

Elementary Probability

In mathematics, statistics is a subset of probability theory which, in turn, is a subset of set theory. Hence, an understanding of statistics requires some understanding of probability, the topic of this chapter.

A.1 Terminology

Just as parametric statistics has the concepts of a population and a sample, probability has the concepts of a universe (analogous to a population) and events (analogous to items sampled from the universe). Once a universe has been defined, an event may be a concrete object (e.g., a female rat, a person with Alzheimer's disease), an abstract object or class of concrete objects (e.g., a genotype), a number meeting defined rules (e.g., systolic blood pressure greater than 140), and even a statistic (e.g., a mean or a variance). For the purpose of learning probability, however, it is useful to view the universe under discussion as a big hat filled with discrete objects. We take that approach here in defining the elementary rules of probability. Later, we will generalize these rules to more abstract problems.

Definition 1: The probability of an event equals the number of those events divided by the total number of events. This is also called the *unconditional probability* of an event. For example, suppose an introductory class in statistics has 213 students, 87 of whom are psychology majors. Then the probability of randomly selecting a psychology major equals $87/213 = .408$. Because of this definition, probabilities will always be positive numbers that are greater than or equal to 0 and less than or equal to 1.0. The unconditional probability of an event, say, A, is usually denoted as $p(A)$ or p_A.

Definition 2: Two events, A and B, are *mutually exclusive* when event A and event B cannot exist in the same object. Examples: (1) In a coin toss it is not possible to observe a heads and a tails at the same time. Hence, heads and tails are mututally exclusive. (2) In a deck of cards, a spade and a heart are mutually exclusive because no card can be both a spade and a heart at the same time.

One the other hand, a diamond and a jack are not mutually exclusive because the Jack of diamonds exists.

Definition 3: The *joint* probability of A and B is the probability of the intersection of A and B. For example, assume that the universe is an ordinary deck of playing cards. The joint probability of a face card and a heart refers to all face cards that are also hearts, namely, the jack, queen, and king of hearts. If A and B are mutually exclisuve, then their joint probability must be 0.

Definition 4: The *conditional probability* of B given A is the probability of event B when the universe is restricted to all objects where A is present. Imagine a big hat of objects. Remove all objects in which A is present and toss them into another hat. The *unconditional probability* of B is the probability of picking an object with B from the *first* hat. The *conditional probability* of B given A is the probability of picking an object with B from the *second* hat. The notation for conditional probability is $p(B|A)$. Read, "the probability of B given A.

For example, recall the stat class with 213 students and 87 psych majors. The conditional probability that a random student is an upper division student given that the student is a psych major equals the frequency of upper division students among all psych majors. It is not the frequency of uppder division students among all 213 students in the class.

Definition 5: Two events, A and B, are *independent* when $p(B|A) = p(B)$ and $p(A|B) = p(A)$.

Definition 6: Two events, A and B, are *dependent* when $p(B|A) \neq p(B)$ and $p(A|B) \neq p(A)$. Note that the term "dependent" does not necessarily imply a causal relationship.

A.2 Rules of elementary probability

A.2.1 Rule 1: The probability of an event

This is the same as Definition 1. It is repeated here for convenience.

The probability of event A is the frequency of all objects with A divided by the total number of objects in the universe of discourse (the "hat").

A.2.2 Rule 2: Summation rule

This rules states

The sum of all mututally exclusive events is 1.0.

For example, let the universe be a deck of cards. Suits are mutually exclusive. Hence,

$$p(\text{Club}) + p(\text{Diamond}) + p(\text{Heart}) + p(\text{Spade}) = 1.0 \tag{A.1}$$

A.2.3 Rule 3: The "and" rule

Technically, this is the probability of joint events, $p(A \cap B)$, but because it is talked of and thought of using the word "and," it is called the "and rule.

$$p(A \text{ and } B) = p(B|A)p(A) = p(A|B)p(B) \tag{A.2}$$

Note that when A and B are independent, then Definition 5 implies that

$$p(A \text{ and } B) = p(B|A)p(A) = p(A|B)p(B) = p(A)p(B) \tag{A.3}$$

A.2.4 Rule 4: The "or" rule

Technically, this is the probability of the union of two events, $p(A \cup B)$, but it is called the "or" rule because the word "or" is used in talking and think of such problems.

$$p(A \text{ or } B) = p(A) + p(B) - p(A \text{ and } B) \tag{A.4}$$

A.3 Important probability distributions

A.3.1 Binomial distribution

The binomial distribution applies to a a Bernoulli variable and is used is such problems as "if you tossed a coin 6 times, what is the probability of getting two heads?" Let p denote the probability of the outcome of interest, Hence, the probability of the other outcome must be $(1 - p)$. Let n denote the toral number of events. Usually, n is expressed as the "total number of trials," where a "trial" is a single random sampling of the Bernoulli variable. In the coin-toss question, $n = 6$. Finally, let r denote the number of the events of interest. In the example, the events of interest are heads, so $r = 2$. Often r is termed the "number of successes" where a "success" is defined as an event of interest.

The formula for the binomial is

$$P(r \text{ of } n) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r} \tag{A.5}$$

In terms of the example where $p = 0.5$, $n = 6$, and $r = 2$,

$$P(2 \text{ heads of 6 coin tosses}) = \frac{6!}{2!(6-2)!} 0.5^2 (1-0.5)^{6-2} = \tag{A.6}$$

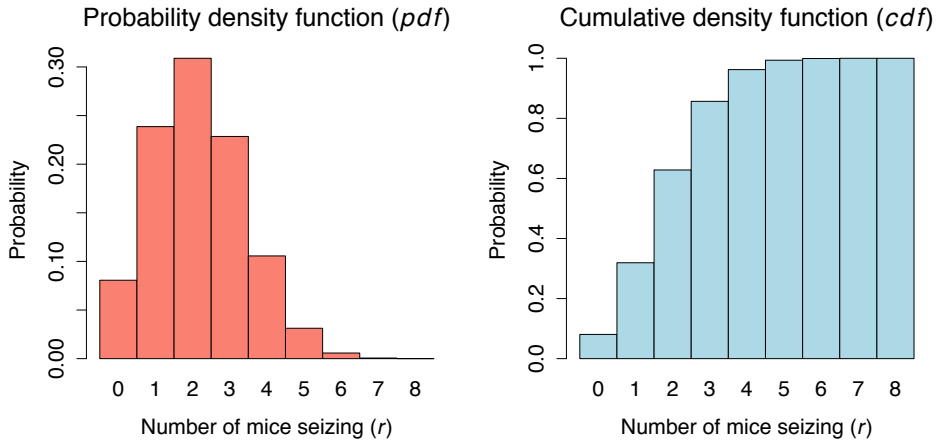
Let's take a specific example of the binomial and use it to introduce some new concepts in probability. Assume that previous research indicates that the probability of a mouse having a seizure under a specific experimental condition is .27. If a group has 8 mice, that are all the pobbile outcomes in terms of the number of mice having seiquzres.

Here $p = .27$, $n = 8$, and we want to solve Equation A.5 for $r = 0, 1, \ldots 8$. Table A.1 presents the results of these calculations under the column "*pdf*," and the left panel of Figure A.1 gives them as a graph. Here, r is the number of mice who are expected to have a seizure. Hence, the probability that no mice have a seizure is 0.08, and the probability that one mouse seizes and the other seven

Table A.1: Probability of all possible outcomes of seizures.

	pdf	*cdf*
r	$r =$	$r \leq$
0	0.080646	0.081
1	0.238624	0.319
2	0.308903	0.628
3	0.228504	0.857
4	0.105644	0.962
5	0.031259	0.994
6	0.005781	0.999
7	0.000611	1.000
8	0.000028	1.000

Figure A.1: Graphical representation of expected outcomes in the mouse seizure example.



do not is 0.24. Note that the probabilities for all mice seizing is very remote, less than 5 in 100,000.

The technical name of the function given algebraically in Equation A.5 and enumerated in column "*pdf*" in Table A.1 and displayed in the left panel of A.1 is the binomial *probability density function*. A probability density function or pdf gives the probability of observing a value X for a function. Hence, Equation A.5 gives the probability of observing any value of X (which is written as r in the equation) in a binomial pdf.

The right hand column of Table A.1 and the right panel in Figure A.1 give the *cumulative density function* (*cdf*) aka the *cumulative distribution function* or simply *distribution function*. The cdf is a mathematical function that returrns the probability of observing a value less than or equal to X. For example, in Table A.1, the probability of $X \leq 3$, i.e., three or fewer mice having seizures, is 0.857. In mathematical terms, the cdf is the integral of the pdf from the lowest possible value of X to the value of X that is of interest.

A.4 The binomial and the logic of hypothesis testing

You have probably heard of the term "null hypothesis." If you stretch your neurons, you might recall something about statistics and rejecting the null hypothesis. In this section, we will develop the concept of the null hypothesis and rejecting the null hypothesis using induction and the binomial distribution.

You have just won several million dollars in Las Vegas betting on the flips of a silver dollar. As you are about to cash in your earnings, the casino officials along with the Las Vegas police, the Nevada state police, some officials from the Nevada gaming commission, a cadre of FBI and Interpol agents, and a very scary bunch of gumbas carrying baseball bats surround you and demand that you prove the silver dollar is a fair coin. No rocket science here. You cash in, give it all to the gumbas, and run.[1]

Seriously, how would you demonstrate that the coin is fair? Everyone's intuitive response is to flip the coin a large number of times, tabulate the number of heads and tails, and see if the frequency of, say, heads is 50% after all the flips. That is a great strategy, but to turn it into real science, one must convert the philiosophy into a concrete, empirical study. "Flip the coin a large number of times" must be translated into a real, honest to goodness number for a scientific study. Should it be 50, 500, 5,000 or 50 million? "Frequency of ... head is 50%" must also be translated into the anticipated a real, honest to goodness number for the outcomes from the number of flips. If the coin is flipped 500 times and 49.7% of the outcomes are heads, does that mean that the coin is biased? After all, 49.7% is not 50%. But then again, 49.7% is pretty close to 50%.

The mathematical solution to the outcomes is the binomial pdf. To calculate

[1] Loosely adapted from a stand-up comedy routine at Danny's Upstairs Comedy Club at Tutta Pasta in Hoboken, NJ. I regret that I forget the name of the comedian.

the distribution of outcomes for heads, we must have actual numbers for two quantities. They are: n, the total number of coin flips, and p, the probability of heads. The number for n is something that the designers of the empirical study must agree to beforehand. So, the distribution of outcomes really depends on the probability of a head, p.

There are four possible hypotheses about p: (1) the coin is fair, so $p = 0.5$; (2) the coin in biased in favor of heads, so $p > 0.5$; (3) the coin is biased in favor of tails so $p < 0.5$; and (4) the coin is biased with no commitment on whether the bias favors heads or tails, so $p \neq 0.5$. A sadistic statistics professor would now give the following homeworks assignment: Set n to 50 and compute Table A.1 and Figure A.1 for each of the four hypotheses.

You can compute the pdf and the cdf and draw the graphs for the first hypothesis because you can substitute the number 0.5 for p in Equation A.5. You cannot, however, perform computations for hypotheses (2), (3), and (4). Why? Because they are *mathematically too imprecise*. Hypothesis 2, for example, states that p is a value greater than 0.5, but does not say whether it is 0.62138, 0.501, or 0.7. Note carefully that this is *not* a criticism of the hypothesis. In fact, it would be downright stupid to randomly pick a number greater than 0.5 just for the sake of having a number. If you picked, say, 0.59, you might fail to detect the bias if the true p were, say, 0.53. The fact that hypothesis (2) is mathematically imprecise is a statement of reality. Many times in science, we have good, testable hypotheses that are mathematically imprecise. We must arrange our math and statistics to enable us to test those hypotheses. We should not change the hypotheses for the sake of mathematical precision.

So what do we do? Simple. We compute the outcomes for the only hypothesis that we can, namely hypothesis (1) that states the coin was fair so $p = 0.5$. Enter the null hypothesis.

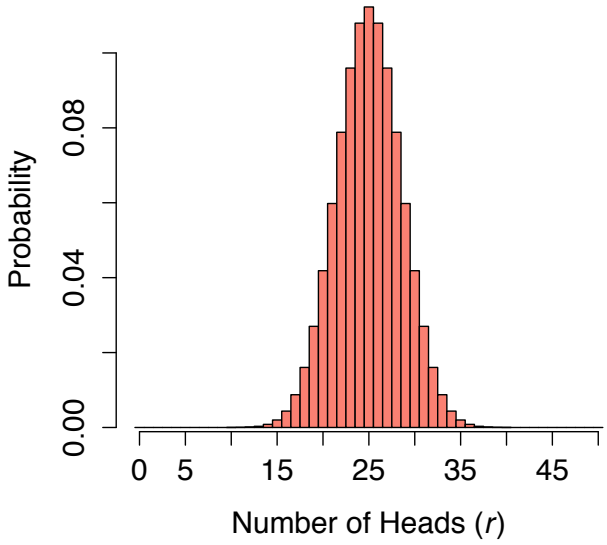
A.4.1 The null hypothesis

The null hypothesis is used in science because it is usually the only hypothesis that has sufficient mathematical precision to allow calculations of the probabilities of the different outcomes of the study. Usually, the null hypothesis is of little interest in itself. It plays an important role only because of its mathematical precision. For example, recall the seizure example. If we were testing a drug purported to reduce seizures, we would have a substantive hypothesis that $p < 0.27$. We cannot, however, use the binomial pdf to calculate the outcomes. Instead, we calulate the outcomes from the null hypothesis that $p = 0.27$, giving Table A.1 and Figure A.1.

> A null hypothesis is established because it is the only hypothesis with sufficient mathematical precision to permit calculation of the probabilities of the various outcomes of a study.

Figure A.2 plots the probabilities of the number of heads for 50 tosses of a fair coin. Common sense tells us that if we performed the study and observed 10 heads (or at the upper end, 40 heads) the coin is obviously not fair. But what

Figure A.2: Probability of the outcomes from 50 tosses of a fair coin.



if we observed 17 heads? Should we call the coin biased then? Scientifically speaking, what is the best cut off for concluding that the coin is biased?

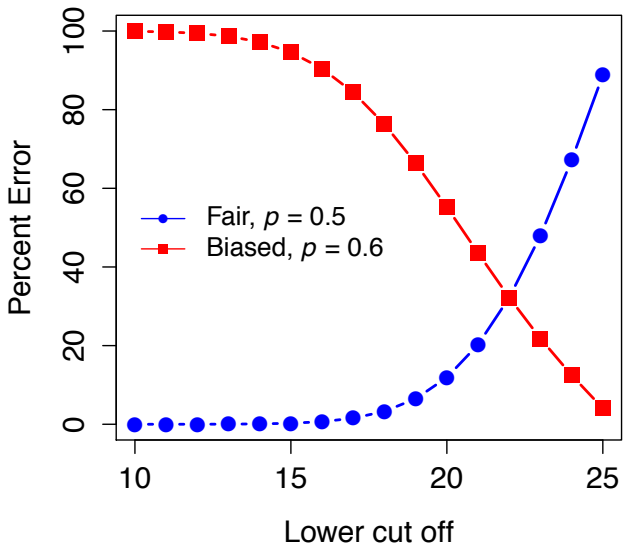
A.4.2 The 5% solution

You may be surprised to learn that there is no "scientifically best" answer for the cut offs. Some cut offs, however, are better than others. For argument's sake, assume that we call the coin fair if and only if we observed exactly 25 heads. The probability of observing this when the coin is really fair is 0.112. Hence, if the coin is really fair, the probability that we call it biased would be (1 - .112) = .888. In other words, 89% of the time we would make an erroneous decision.

We could extend the criterion outwards. Let's say that we call the coin fair if we observe between 24 and 26 heads. From the pdf, the probability of observing 24, 25, or 126 heads when the coin is fair is .328. Here, 67% of your judgements would be false positives (calling the coin biased when it is, in fact, fair). If we continue adjusting the cut offs and calculating the error, we arrive at the blue line in Figure A.3.

If we examine the blue line, then we could minimize the false positive rate by picking a low lower cut off and a high upper cut off. That strategy suffers, however, from another type of error–false negatives. A false negative error occurs when the coin is indeed biased but we call it fair. To calculate the false

Figure A.3: Error rates for different cut offs for a fair coin and a biased coin where the probability of a head is 0.60.



negative rate we must have a pdf for a binomial in which p is something other than 0.5. The red line in Figure A.3 plot the false negative error rate for a biased coin in which the probability of a heads is 0.6. Note that as we make the cut off lower and lower to avoid false positives, we increase the rate of false negatives.

If we had other types of information available (e.g., what is the probability that a coin will be biased to begin with? what is the relative seriousness of a false positive versus a false negative error? etc.) and with some assumptions, we could arrive at a optimal cut point for this study. The problem, of course, is that in the real word, we would not know many of these types of information. In particular, we would no know p, the probability of a head for a biased coin.

In real life problems in neuroscience, we also do not know the distribution of potential outcomes for a substantive hypothesis. Hence, we cannot compute a graph like the one in Figure A.3. Consequently, we must deal in generalities.

In most areas of neuroscience and in many other fields, scientists have arrived at a consensus for an acceptable false positive rate. It is one false positive in 20 decisions or a 5% solution. Is this close to optimum? If we knew the results of experiments before we conduct them, then we could answer that. But we don't. The 5% solution is an arbitrary convention but one that seems to have served us well.[2]

Let's recap. We use the null hypothesis to calculate the distribution of the outcomes of a study. The 5% solution entails tabulating the 5% most unlikely of these outcomes. If the actual outcome of the experiment (or study) falls into this 5% most unlikely group, then we reject the null hypothesis.

Back to the coin toss problem and Figure A.2. To get the 5% most unlikly outcomes from the figure, we must first realize that we can observe either too few heads or too many heads. Hence, the 5% will be split evenly between the lower tail and the upper tail of the distribution in Figure A.2. Half of 5% is 2.5%. Hence, we can use the cdf (cumulative distribution) to find the cut off so that 2.5% of the distribution falls below it and 97.5% above it. The closest value to this is 20 heads. Here, 3.2% of the distribution includes 20 or fewer heads. At the upper end, 3.2% of the distribution include 30 or more heads.

We fill no flip the coin 50 times. If we observe 20 or fewer heads or 30 or more heads, we will call the coin biased. Otherwise, we call the coin fair.

[2]In other sciences and for other problems, the false positive rate can be set to a much more stringent level. In particle physics and some genomic problems it can be of the order of 10^{-7}or 10^{-8}.