Probability and Biology

- Probability comes up in everyday life — predicting the weather, lotteries or sports betting, strategies for card games, understanding risks of passing genetic diseases to children, assessing your own risks of diseases associated in part with genetic causes.

abilityLarget

of Statistics

consin - Madison

Random Sampling

- Most of the formal methods of statistical inference we will use in this class are based on the assumption that the individual units in the sample are sampled at random from the population of interest.
- (Ignore for the present that in practice, individuals are almost never sampled at random, in a very formal sense, from the population of interest.)
- Taking a simple random sample of size n is equivalent to the process of:
  1. representing every individual from a population with a single ticket;
  2. putting the tickets into large box;
  3. mixing the tickets thoroughly;
  4. drawing out n tickets without replacement.
- Stratified random sampling and cluster sampling are examples of random sampling processes that are not simple. Data analysis for these types of sampling strategies go beyond the scope of this course.

Probability and Biology

Why should we know something about probability?

- Some biological processes seem to be directly affected by chance outcomes. Examples include formation of gametes and occurrence of genetic mutations.
- Formal statistical analysis of biological data assumes that variation not explained by measured variables is caused by chance.
- Chance might be used in the design of an experiment, such as the random allocation of treatments or random sampling of individuals.
- Probability is the language with which we express and interpret assessment of uncertainty in a formal statistical analysis.
- Formal statistical analysis depends on modeling observed data as the realization of a random process.

Inference from Samples to Populations

- Statistical inference involves making statements about populations on the basis of analysis of sampled data.
- The Simple random sampling model is useful because it allows precise mathematical description of the random distribution of the discrepancy between statistical estimates and population parameters. This is known as chance error due to random sampling.
- When using the random sampling model, it is important to ask what is the population to which the results will be generalized? The use statistical methods that assume random sampling on data that is not collected as a random sample is prone to sampling bias, in which individuals do not have the same chance of being sampled.
- Sampling bias can lead to incorrect statistical inferences because the sample is unrepresentative of the population in important ways.

Simple Random Sampling

- The defining characteristic of the process of simple random sampling is that every possible sample of size n has the same chance of being selected.
- In particular, this means that (a) every individual has the same chance of being included in the sample; and that (b) members of the sample are chosen independently of each other.
- Note that point (a) above is insufficient to define a simple random sample. As an example, consider sampling one couple at random from a set of ten couples. Each person would have a one in ten chance of being in the sample, but the sampling is not independent. Possible samples of two people from the population who are not in a couple have no chance of being sampled while each couple has a one in ten chance of being sampled.

Probability

- Probability is a numerical measure of the likelihood of an event.
- Probabilities are always between 0 and 1, inclusive.
- **Notation:** The probability of an event E is written $\Pr\{E\}$.

Examples:

If a fair coin is tossed, the probability of a head is

$$\Pr\{\text{Heads}\} = 0.5$$

If bucket contains 34 white balls and 66 red balls and a ball is drawn at random, the probability that the drawn ball is white is

$$\Pr\{\text{white}\} = 34/100 = 0.34$$

Using R to Take a Random Sample

Suppose that you have a numbered set of individuals, numbered from 1 to 98, and that I wanted to sample ten of these. Here is some R code that will do just that.

```
> sample(1:98, 10)
 [1] 19 74  3 51 70 75 14 31 76 86
```

In the `sample` function, the first argument is the set from which to sample (in this case the integers from 1 to 98) and the second argument is the sample size.

In the output, the [1] is R's way of saying that that row of output begins with the first element.

The same code executed again results in a different random sample.

Examples of Interpretations of Probability

- **Coin-tossing** — it is reasonable to consider tossing a coin many times where each coin toss can be thought of as a repetition of the same basic chance operation. The probability of heads can be thought of a the long-run relative frequency of heads.
- **Packer Football** — the outcome of the next Packer game is uncertain, but it is less reasonable to think about the outcome (Packers win, lose, or tie) as something that could be repeated indefinitely. The long-run relative frequency interpretation of probability does not allow for an interpretation of the probability of an event that will occur only once.
- **Evolution** — the statement "molluscs form a monophyletic group" means that all living individuals classified as molluscs have a common ancestor that is not an ancestor of any non-molluscs. It is uncertain whether or not this statement is true.

Interpretations of Probability

- The frequency interpretation of probability defines the probability of an event E as the relative frequency with which event E would occur in an indefinitely long sequence of independent repetitions of a chance operation.
- A subjective interpretation of probability defines probability as an individual's degree of belief in the likelihood of an outcome. This school of thought allows the use of probability to discuss events that are not hypothetically repeatable.
- The textbook follows a frequency interpretation of probability.
- Statistical methods based on subjective probability are called Bayesian, named after the Reverend Thomas Bayes who first proved a mathematical theorem we will encounter later. In the Bayesian approach to statistics, everything unknown is described with a probability distribution. Bayes' Theorem describes the proper way to modify a probability distribution in light of new information.

Comparing Bayesian and Frequentist Approaches

- A Bayesian approach to statistical inference allows one to quantify uncertainty in a statement with a probability and describes how to update the probability in light of new data.
- A frequency approach to statistical inference does not allow direct quantification of uncertainty with probabilities for events that happen only once.
- A frequentist approach would ask instead, if I assume that the event is true, how likely is an observed outcome? If the probability of the observed outcome is low enough relative to some alternative, this would be seen as evidence against the hypothesis.

Interpretations of Probability

- In particular, Bayesian methods treat population parameters as random variables, requiring a probability distribution based on prior knowledge and not on data.
- Frequency methods treat population parameters as fixed, but unknown.
- Methods of statistical analysis based on the frequency interpretation of probability are in most common use in the biological science, but Bayesian approaches are becoming more accepted and more prevalent.
- It is my desire to teach you the frequentist approach to statistical inference while leaving you open-minded about learning Bayesian statistics at a future encounter with statistics.
- This requires education in the calculus of probability.

Conditional Probability and Probability Trees

It is a common setting in biological probability problems for an event to consist of the outcomes from a sequence of possibly dependent chance occurrences. In this case, a probability tree is a very useful device for guiding the appropriate calculations.

We have already discussed definitions of probability and events. The following example will illustrate definitions of conditional probability, independence of events and several rules for calculating probabilities of complex events.

Example

The following relative frequencies are known from review of literature on the subject of strokes and high blood pressure in the elderly.

- 1. Ten percent of people aged 70 will suffer a stroke within five years;
- 2. Of those individuals who had their first stroke within five years after turning 70, forty percent had high blood pressure at age 70;
- 3. Of those individuals who did not have a stroke by age 75, twenty percent had high blood pressure at age 70.

What is the probability that a 70 year-old patient with high blood pressure will have a stroke within five years?

To answer this question, it is useful to introduce a notation to define the relevant events and their probabilities.

Example (cont.)

Here are two events of relevance.

$S = \{\text{stroke before age 75}\}$ $H = \{\text{high blood pressure at age 70}\}$

With this notation, the statement 1. Ten percent of people aged 70 will suffer a stroke within five years; becomes

$$\Pr\{S\} = 0.10$$

The second statement 2. Of those individuals who had their first stroke within five years after turning 70, forty percent had high blood pressure at age 70; becomes

$$\Pr\{H \mid S\} = 0.40$$

The symbol $\mid$ is read "given" and indicates that the value 0.40 is a conditional probability. It may not be true that 40 percent of all 70-year-olds have high blood pressure. The statement is conditional on having had a stroke between ages 70 and 75.

Example (cont.)

The third statement 3. Of those individuals who did not have a stroke by age 75, twenty percent had high blood pressure at age 70; becomes

$$\Pr\{H \mid S^c\} = 0.20$$

The question of interest, What is the probability that a 70 year-old patient with high blood pressure will have a stroke within five years? becomes

$$\text{What is } \Pr\{S \mid H\}?$$

Notice that in this question, the order of conditioning is reversed. It is precisely this situation where Bayes' Theorem is useful.

Formal Probability Rules

- **Non-negativity:** For any event E, $0 \leq \Pr\{E\} \leq 1$.
- **Outcome space:** The probability of the event of all possible outcomes is 1.
- **Complements:** $\Pr\{E^c\} = 1 - \Pr\{E\}$.
- **Disjoint events:** If events E_1 and E_2 are disjoint or mutually exclusive, meaning that it is impossible for both events to occur in a single realization, then

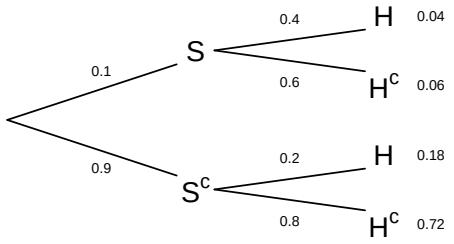
$$\Pr\{E_1 \text{ or } E_2\} = \Pr\{E_1\} + \Pr\{E_2\}.$$

- **Inclusion-exclusion:** For any two events E_1 and E_2,

$$\Pr\{E_1 \text{ or } E_2\} = \Pr\{E_1\} + \Pr\{E_2\} - \Pr\{E_1 \text{ and } E_2\}$$

A Probability Tree for the Example

- Exactly one path through the tree is realized.
- The probability of a path through the tree is the product of the edge probabilities.
- Probabilities out of a point must sum to one.
- $\Pr\{S\} = 0.10$ implies that $\Pr\{S^c\} = 1 - 0.10 = 0.90$. These unconditional probabilities appear at the first branching point.
- Conditional probabilities appear at the other branching points.



Conditional Probability and Independence

Definition: The conditional probability of E_2 given E_1 is defined to be

$$\Pr\{E_2 \mid E_1\} = \frac{\Pr\{E_2 \text{ and } E_1\}}{\Pr\{E_1\}}$$

provided that $\Pr\{E_1\} > 0$.

Definition: Two events are independent if one event does not affect the probability of the other event. Specifically, events E_1 and E_2 are independent if

$$\Pr\{E_2 \mid E_1\} = \Pr\{E_2\}$$

An equivalent definition is events E_1 and E_2 are independent if

$$\Pr\{E_1 \text{ and } E_2\} = \Pr\{E_1\} \times \Pr\{E_2\}$$

- **Multiplication:** For any events E_1 and E_2,

$$\Pr\{E_1 \text{ and } E_2\} = \Pr\{E_1\} \times \Pr\{E_2 \mid E_1\}$$

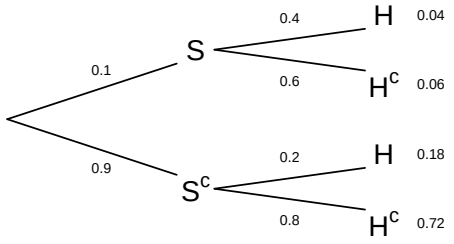
Example (cont.)

- We can find the unconditional probability of high blood pressure.

$$\Pr\{H\} = \Pr\{S \text{ and } H\} + \Pr\{S^c \text{ and } H\} = 0.04 + 0.18 = 0.22$$

- Of those with high blood pressure, the proportion who had a stroke is computed as follows.

$$\Pr\{S \mid H\} = 0.04/0.22 \doteq 0.182$$



Derivation of Bayes' Theorem

$$\begin{aligned}
\Pr\{E_i \mid A\} &= \frac{\Pr\{E_i \text{ and } A\}}{\Pr\{A\}} && \text{def. of conditional prob.} \\
&= \frac{\Pr\{E_i\}\Pr\{A \mid E_i\}}{\Pr\{A\}} && \text{multiplication rule} \\
&= \frac{\Pr\{E_i\}\Pr\{A \mid E_i\}}{\sum_{j=1}^{n} \Pr\{A \mid E_j\}\Pr\{E_j\}} && \text{law of total probability}
\end{aligned}$$

For specific problems, it is often easier to apply the definition of conditional probability and then use the multiplication rule and the law of total probability separately than to just pull out Bayes' Theorem in all of its glory.

Law of Total Probability

Suppose that we want to find $\Pr\{A\}$, but we only know conditional probabilities of A given list of events that encompasses all possibilities. We can find $\Pr\{A\}$ by conditioning on which of these events occurred.

Law of Total Probability: Suppose that events $E_1, E_2, \ldots, E_n$ form a partition of the space of possible outcomes. This means that exactly one of the events must occur. Suppose we also know all of the probabilities $\Pr\{E_i\}$ and all of the conditional probabilities $\Pr\{A \mid E_i\}$. Then,

$$\Pr\{A\} = \sum_{i=1}^{n} \Pr\{A \mid E_i\}\Pr\{E_i\}$$

This is equivalent to adding up some of the probabilities at the end of a probability tree.

Probability Rules for Conditional Probabilities

The rules extend to conditional probabilities. Let A be another event on which we condition.

- **Non-negativity:** For any event E, $0 \le \Pr\{E \mid A\} \le 1$.
- **Outcome space:** $\Pr\{A \mid A\} = 1$.
- **Complements:** $\Pr\{E^c \mid A\} = 1 - \Pr\{E \mid A\}$.
- **Disjoint events:** If events E_1 and E_2 are disjoint,

  $$\Pr\{E_1 \text{ or } E_2 \mid A\} = \Pr\{E_1 \mid A\} + \Pr\{E_2 \mid A\}$$

- **Inclusion-exclusion:** For any two events E_1 and E_2,

  $$\Pr\{E_1 \text{ or } E_2 \mid A\} = \Pr\{E_1 \mid A\} + \Pr\{E_2 \mid A\} - \Pr\{E_1 \text{ and } E_2 \mid A\}$$

- **Multiplication:** For any events E_1 and E_2,

  $$\Pr\{E_1 \text{ and } E_2 \mid A\} = \Pr\{E_1 \mid A\} \times \Pr\{E_2 \mid E_1 \text{ and } A\}$$

- **Conditional probability:** $\Pr\{E_2 \mid E_1 \text{ and } A\} = \frac{\Pr\{E_2 \text{ and } E_1 \mid A\}}{\Pr\{E_1 \mid A\}}$.
- **Law of total probability:** If $E_1, E_2, \ldots, E_n$ are a partition of A, then

  $$\Pr\{B \mid A\} = \sum_{i=1}^{n} \Pr\{B \mid E_i \text{ and } A\}\Pr\{E_i \mid A\}$$

Bayes' Theorem

Bayes' Theorem is a statement of a generalization of the calculation we carried out with the probability tree.

Bayes' Theorem follows from the previous formal definitions.

Bayes' Theorem: Suppose that events $E_1, E_2, \ldots, E_n$ form a partition of the space of possible outcomes. Then,

$$\Pr\{E_i \mid A\} = \frac{\Pr\{A \mid E_i\}\Pr\{E_i\}}{\sum_{j=1}^{n} \Pr\{A \mid E_j\}\Pr\{E_j\}}$$

An interpretation of Bayes' Theorem is the following. Before observing any data, we have prior probabilities $\Pr\{E_i\}$ for each of these events. After observing an event A, we calculate the posterior probabilities $\Pr\{E_i \mid A\}$ in response to the new information that event A occurred.

Example (cont.)

The father is part of the F2 generation, which implies that $\Pr\{F\} = 3/4$ and $\Pr\{H\} = 1/2$. Knowing that the father has the dominant trait affects the probability that he is heterozygous.

$$\Pr\{H \mid F\} = \frac{\Pr\{H \text{ and } F\}}{\Pr\{F\}} = \frac{\Pr\{H\}}{\Pr\{F\}} = \frac{1/2}{3/4} = \frac{2}{3}$$

Notice that in this example, $\Pr\{H \text{ and } F\} = \Pr\{H\}$ because every heterozygote exhibits the dominant trait. (H is a subset of F, so the event H and F is the same event as H.)

Genetics Example

Problem:

A single gene has a dominant allele A and recessive allele a. A cross of AA versus aa leads to F1 offspring of type Aa. Two of these mice are crossed to get the F2 generation, some of which are AA, some of which are Aa, and some of which are aa. A male with the dominant trait from the F2 generation is randomly selected. He is either homozygous dominant (AA) or heterozygous (Aa). He is mated with a homozygous recessive (aa) female. They have one offspring with the dominant trait.

Given the other information in the problem, what is the probability that the father is heterozygous?

Example (cont.)

The original question was to find $\Pr\{H \mid D \text{ and } F\}$. First, we can use a variation on the definition of conditional probability by continuing to condition on event F that the father has the dominant trait.

$$\Pr\{H \mid D \text{ and } F\} = \frac{\Pr\{H \text{ and } D \mid F\}}{\Pr\{D \mid F\}}$$

A variation of the multiplication rule applies to the numerator.

$$\Pr\{H \text{ and } D \mid F\} = \Pr\{H \mid F\} \Pr\{D \mid H \text{ and } F\} = \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}$$

A variation of the law of total probability applies to the denominator.

$$\begin{aligned} \Pr\{D \mid F\} &= \Pr\{H \mid F\} \Pr\{D \mid H \text{ and } F\} + \Pr\{H^c \mid F\} \Pr\{D \mid H^c \text{ and } F\} \\ &= \left(\frac{2}{3} \times \frac{1}{2}\right) + \left(\frac{1}{3} \times 1\right) = \frac{2}{3} \end{aligned}$$

So the final answer is $\frac{1/3}{2/3} = \frac{1}{2}$.

Example (cont.)

Begin by defining several events.

$$\begin{aligned} D &= \{\text{offspring is dominant}\} \\ F &= \{\text{father is dominant}\} \\ H &= \{\text{father is heterozygous for the trait}\} \end{aligned}$$

With this notation, we are asked to find $\Pr\{H \mid D \text{ and } F\}$.

It is also useful to write down the probabilities we do know from the problem setting in terms of these events. For this problem, this assumes some background knowledge of genetics.

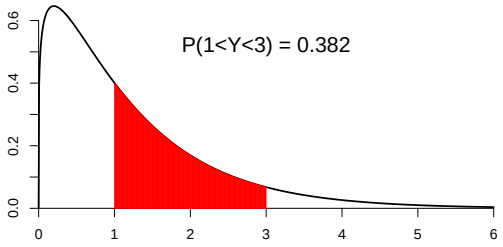
We know the genotype of the mother (aa). We can't compute $\Pr\{D\}$ directly, but we could if we knew the father's genotype as well. If the father's genotype is (AA), the offspring is certain to have the dominant trait, $\Pr\{D \mid H^c\} = 1$, while if the father's genotype is heterozygous (Aa), then the offspring is equally likely to be dominant or recessive, $\Pr\{D \mid H\} = 0.5$.

Random Variables

- A random variable is a numerical variable whose value depends on a chance outcome.
- While we can imagine other cases, most random variables we will see in practice are either discrete or continuous.
- A probability distribution answers the question, how likely is it that a random variable's realized value will be in some set?
- There are multiple ways to describe probability distributions.
- The distributions of continuous random variables are most often described by probability density curves.
- The distributions of discrete random variables are most often described by a table or formula that specifies the probability associated with each possible value. This is called a probability mass function.

Probability Density Curves

- You can think of a probability density as an idealized histogram, scaled so that the total area under the curve is one.
- For any two numbers a and b (with $a < b$), and a continuous random variable Y, we have that $\Pr\{a < Y < b\}$ is the area under the curve between a and b.
- Density curves must be non-negative and the total area under each curve must be exactly one.



Probability Mass Functions

- A probability mass function is a list of the possible values of the random variable and the probability associated with each possible value.
- The sum of the probabilities over all possible values must be one, and the probabilities must be non-negative.

Example: Two different ways to specify the same discrete probability distribution.

y	1	2	3	4
$\Pr\{Y = y\}$	0.1	0.2	0.3	0.4

$\Pr\{Y = y\} = y/10$ for $y = 1, 2, 3, 4$.

Cumulative Distribution Functions

- The cumulative distribution function (cdf) answers the question, how much probability is less than or equal to y?
- The cdf F is defined to be $F(y) = \Pr\{Y \leq y\}$.
- CDFs of continuous random variables are continuous curves that never decrease, moving from 0 up to 1.
- CDFs of discrete random variables are step functions that never decrease while moving from 0 up to 1 in discrete jumps.
- For discrete random variables, I think of probability as one unit of stuff that has been broken into chunks. The chunks are spread out on a number line.
- For continuous random variables, I think of probability as one unit of stuff that has been ground into a fine dust and spread out on a number line.

The Binomial Distribution

The binomial distribution is a discrete probability distribution that arises in many common situations. The canonical example of the binomial distribution is counting the number of heads in a fixed number of independent coin tosses.

Independent-Trials Model

In a series of n independent trials, each trial results in a success (the outcome we are counting) or a failure. Each trial has the same probability p of success. A binomial random variable counts the number of successes in a fixed number of trials.

There are five key characteristics to look for when deciding if a random variable has a binomial distribution.

Means of Random Variables

The mean of a probability distribution is the location where the probability balances.

For discrete random variable Y the mean μ_Y, also known as the expected value $E(Y)$, is defined as a sum.

$$E(Y) = \mu_Y = \sum_i y_i \Pr\{Y = y_i\}$$

where the sum is understood to go over all possible values. The sum is a weighted average of the possible values of Y where the weights are the probabilities. The mean is a measure of the center of a distribution.

The mean of a continuous random variable assumes knowledge of calculus.

The Binomial Distribution

- 1. Each trial has two possible outcomes. (It is also okay for there to be multiple outcomes that are grouped to two classes of outcomes.)
- 2. Trials are independent.
- 3. The number of trials, n, is fixed in advance.
- 4. Each trial has the same success probability, p.
- 5. The random variable counts the number of successes in the n trials.

Parameters: The binomial distribution is completely determined by two parameters. These are n, the number of trials, and p the success probability common to each trial.

Variance of Random Variables

The variance of a probability distribution is a measure of its spread, namely the expected value of the squared deviation from the mean.

For discrete random variable Y the variance σ_Y^2, also known as $\mathrm{Var}(Y)$, is defined as

$$E((Y - \mu_Y)^2) = \sigma_Y^2 = \sum_i (y_i - \mu_Y)^2 \Pr\{Y = y_i\}$$

where the sum is understood to go over all possible values.

The standard deviation is the square root of the variance. The standard deviation may be interpreted as a typical distance for the random variable to be from the mean of the distribution.

Binomial Distribution (cont.)

The probability mass function for the binomial distribution is

$$\Pr\{Y = j\} = {}_nC_j p^j (1-p)^{n-j} \qquad \text{for } j = 0, 1, \ldots, n$$

where

$${}_nC_j = \frac{n!}{j!(n-j)!}$$

This is the probability of exactly j successes.

The formula arises because $p^j(1-p)^{n-j}$ is the probability of each sequence of exactly j successes and $n-j$ failures and there are ${}_nC_j$ different such sequences.

The mean of a binomial distribution is $\mu = np$.

The variance of a binomial distribution is $\sigma^2 = np(1-p)$ so that the standard deviation is $\sigma = \sqrt{np(1-p)}$.

Binomial Distribution (cont.)

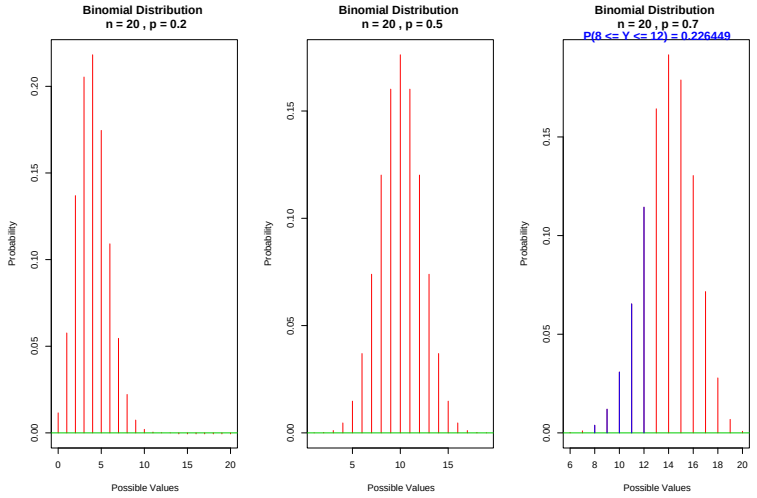
You should be able to calculate binomial probabilities using your calculator.

We will teach you to do so using R.

It is even more important for you to recognize from a problem description when a binomial random variable is lurking and when the random variable has a different distribution.

Random sampling is a setting that does not fit the binomial setting exactly, because the individuals in a sample are not independent — (the same individual cannot be drawn twice). However, if the sample size n is much smaller than the population size, the binomial distribution is an excellent approximation to the genuine distribution.

Some Binomial Graphs



Some Binomial Graphs