Question: Why should biologists know about probability?

Answer (1): Some biological processes seem to be directly affected by chance outcomes. Examples:
- formation of gametes;
- recombination events;
- occurrence of genetic mutations;

Answer (2): Formal statistical analysis of biological data models unexplained variation as caused by chance.

Answer (3): In many designed experiments, probability is used for:
- the random allocation of treatments; or
- the random sampling of individuals.
Question: Why should biologists know about probability?

Answer (4): Probability is the language with which we express and interpret assessment of uncertainty in a formal statistical analysis.

Answer (5): Probability comes up in everyday life
- predicting the weather;
- gambling;
- strategies for games;
- understanding risks of passing genetic diseases to children;
- assessing your own risk of disease associated in part with genetic causes.

Random Sampling

The formal methods of statistical inference taught in this course assume random sampling 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.)

Simple Random Samples

The process of taking a simple random sample of size n is equivalent to:
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.
Other Random Sampling Strategies

- 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.

Simple Random Sampling

Definition
A simple random sample of size $n$ is a random sample taken so that every possible sample of size $n$ has the same chance of being selected.

In a simple random sample:
- every individual has the same chance of being included in the sample;
- every pair of individuals has the same chance of being included in the sample;
- in fact, every set of $k$ individuals has the same chance of being included in the sample.

Insufficient criterion for SRS

- The condition that every individual has the same chance of being included in the sample is insufficient to imply a simple random sample.
- For example, consider sampling one couple at random from a set of ten couples.
  1. Each person would have a one in ten chance of being in the sample;
  2. However, each possible set of two people does not have the same chance of being sampled.
  3. Pairs of people from the population who are not coupled have no chance of being sampled;
  4. while each pair of people in a couple has a one in ten chance of being sampled.

Using R to Take a Random Sample

Suppose that you have a numbered set of individuals, numbered from 1 to 104, and that I wanted to sample ten of these. Here is some R code that will do just that.

```R
> sample(1:104, 10)
[1] 9 11 55 100 67 62 68 25 19 54
```
- the first argument is the set from which to sample (in this case the integers from 1 to 104)
- the second argument is the sample size;
- the [1] is R’s way of saying that that row of output begins with the first element.
- executing the same R code again results in a different random sample.
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?

Sampling Bias

Using methods based on random sampling on data 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.

Random Experiments

Definition
A random experiment is a process with outcomes that are uncertain.

Example: Rolling a single six-sided die once. The outcome (which number lands on top) is uncertain before the die roll.

Outcome Space

Definition
The outcome space is the set of possible simple outcomes from a random experiment.

Example: In a single die roll, the set of possible outcomes is:

$$\Omega = \{1, 2, 3, 4, 5, 6\}$$
Events

Definition
An event is a set of possible outcomes.

Example: In a single die roll, possible events include:
- $A =$ “the die roll is even”;
- $B =$ “the die roll is a ‘6’”;
- $C =$ “the die roll is ‘4 or less’”.

Examples

- If a fair coin is tossed, the probability of a head is
  \[ P\{\text{Heads}\} = 0.5 \]

- If bucket contains 34 white balls and 66 red balls and a ball is drawn uniformly at random, the probability that the drawn ball is white is
  \[ P\{\text{white}\} = \frac{34}{100} = 0.34 \]

Probability

Definition
The probability of an event $E$, denoted $P\{E\}$, is a numerical measure between 0 and 1 that represents the likelihood of the event $E$ in some probability model. Probabilities assigned to events must follow a number of rules.

Example: The probability $P\{\text{the die roll is a ‘6’}\}$ equals 1/6 under a probability model that gives equal probability to each possible result, but could be different under a different model.

Frequentist Interpretation of Probability

The frequentist 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.
Subjective Interpretation of Probability

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.

Frequentist Statistics

- The textbook follows a frequency interpretation of probability.
- Frequentist methods treat population parameters as *fixed, but unknown*.
- Methods of statistical analysis based on the frequentist interpretation of probability are in most common use in the biological sciences.

Bayesian Statistics

- 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.
- In particular, Bayesian methods treat population parameters as random variables, requiring a probability distribution based on *prior knowledge* and not on *data*.

Modern Statistics in Biology

- Bayesian approaches are becoming more accepted and more prevalent in general, and in biological applications in particular.
- 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.
- Preparing you for Bayesian statistics requires education in the calculus of probability.
Examples of Interpretations of Probability

- **Coin-tossing** — tossing a coin many times can be thought of as a repetition of the same basic chance operation. The probability of heads can be thought of as the long-run relative frequency of heads.

- **Packer Football** — the outcome (win, loss, or tie) of the next Packer game is uncertain, but it is unreasonable to think of each game as being infinitely repeatable. The long-run relative frequency interpretation of probability does not apply.

- **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, but there is no long-run frequency interpretation.

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 frequentist 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.

Conditional Probability and Probability Trees

- In biological applications, events often consist of a sequence of possibly dependent chance occurrences.

- A **probability tree** is a very useful device for guiding the appropriate calculations in this case.

- 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?
Here are two events of relevance.\[\begin{align*}
S &= \{\text{stroke before age 75}\} \\
H &= \{\text{high blood pressure at age 70}\}
\end{align*}\]

With this notation, the statement

1. Ten percent of people aged 70 will suffer a stroke within five years;

becomes \( P(S) = 0.10 \).

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 \( P(H \mid S^c) = 0.20 \).

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 \( P(H \mid S) = 0.40 \).

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.

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 What is \( P(S \mid H) \)?

In this problem, we want to find \( P(S \mid H) \), but we know \( P(H \mid S) \) and other things.

To solve, we need Bayes' Theorem.
Probability Tree Rules

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

Example (cont.)

- We can find the unconditional probability of high blood pressure.
  
  $\mathbb{P}(H) = \mathbb{P}(S \text{ and } H) + \mathbb{P}(S^c \text{ and } H) = 0.04 + 0.18 = 0.22$

- Of those with high blood pressure, the proportion who had a stroke is computed as follows.
  
  $\mathbb{P}(S \mid H) = 0.04/0.22 \div 0.182$

A Probability Tree for the Example

```
S  0.1
   /  \  0.4  0.6
  /    \  S  ^c H
S^c  0.9
     /  \  0.2  0.8
    /    \  H  ^c H
```

Example (cont.)

```
S  0.4
   /  \  0.04  0.06
  /    \  S  ^c H
S^c  0.9
     /  \  0.1  0.6
    /    \  H  ^c H
```

```
S  0.1
   /  \  0.4  0.6
  /    \  S  ^c H
S^c  0.9
     /  \  0.2  0.8
    /    \  H  ^c H
```
Probability Rules

**Non-negativity Rule**
For any event $E$, $0 \leq P(E) \leq 1$.

**Outcome Space Rule**
The probability of the event of all possible outcomes is 1, or $P(\Omega) = 1$.

**Complements Rule**
$P(E^c) = 1 - P(E)$.

Disjoint Events

**Definition**
Events $E_1$ and $E_2$ are called disjoint, or mutually exclusive, if they have no outcomes in common. Mathematically, $E_1$ and $E_2$ are called disjoint if and only if $E_1 \cap E_2 = \emptyset$.

Addition Rule

**Addition Rule**
If events $E_1$ and $E_2$ are disjoint, then
$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2).$$

**Inclusion-exclusion rule**
The Inclusion-exclusion rule is that for any two events $E_1$ and $E_2$,
$$P(E_1 \text{ or } E_2) = P(E_1) + P(E_2) - P(E_1 \text{ and } E_2).$$

Conditional Probability

**Definition**
The conditional probability of $E_2$ given $E_1$ is defined to be
$$P(E_2 \mid E_1) = \frac{P(E_2 \text{ and } E_1)}{P(E_1)}$$
provided that $P(E_1) > 0$. 
Independence of Events

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

\[ P(E_2 \mid E_1) = P(E_2) \]

provided $P(E_1) > 0$.

Definition
Events $E_1$ and $E_2$ are also independent if

\[ P(E_1 \text{ and } E_2) = P(E_1) \times P(E_2) \]

Law of Total Probability

Definition
Events $E_1, E_2, \ldots, E_n$ form a partition of $\Omega$ if they are disjoint and if $E_1 \cup E_2 \cup \cdots \cup E_n = \Omega$.

- Exactly one event of a partition must occur.
- We can find $P(A)$ by conditioning on the events $E_i$.

Law of Total Probability

If events $E_1, E_2, \ldots, E_n$ form a partition of the outcome space, then,

\[ P(A) = \sum_{i=1}^{n} P(A \mid E_i)P(E_i) \]

- The law of total probability is equivalent to summing the probabilities of some paths through a probability tree.

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

If events $E_1, E_2, \ldots, E_n$ form a partition of the space of possible outcomes, then,

\[ P(E_i \mid A) = \frac{P(A \mid E_i)P(E_i)}{\sum_{j=1}^{n} P(A \mid E_j)P(E_j)} \]
Interpretation

- Before observing any data, we have prior probabilities \( P\{E_i\} \) for each of these events.
- After observing an event \( A \), we calculate the posterior probabilities \( P\{E_i \mid A\} \) in response to the new information that event \( A \) occurred.

Derivation of Bayes' Theorem

\[
P\{E_i \mid A\} = \frac{P\{E_i \text{ and } A\}}{P\{A\}} \quad \text{def. of conditional prob.}
\]

\[
= \frac{P\{E_i\}P\{A \mid E_i\}}{P\{A\}} \quad \text{multiplication rule}
\]

\[
= \frac{n \sum P\{A \mid E_j\}P\{E_j\}}{P\{A\}} \quad \text{law of total probability}
\]

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.

Probability Rules for Conditional Probabilities

- Probability rules extend to conditional probabilities.
  - Non-negativity
    - For any event \( E \), \( 0 \leq P\{E \mid A\} \leq 1 \).
  - Outcome space
    - \( P\{A \mid A\} = 1 \).
  - Complements
    - \( P\{E^c \mid A\} = 1 - P\{E \mid A\} \).

Rules (cont.)

- Disjoint events
  - If events \( E_1 \) and \( E_2 \) are disjoint,
    - \( P\{E_1 \text{ or } E_2 \mid A\} = P\{E_1 \mid A\} + P\{E_2 \mid A\} \)

- Inclusion-exclusion
  - For any two events \( E_1 \) and \( E_2 \),
    - \( P\{E_1 \text{ or } E_2 \mid A\} = P\{E_1 \mid A\} + P\{E_2 \mid A\} - P\{E_1 \text{ and } E_2 \mid A\} \)

- Multiplication Rule
  - For any events \( E_1 \) and \( E_2 \),
    - \( P\{E_1 \text{ and } E_2 \mid A\} = P\{E_1 \mid A\} \times P\{E_2 \mid E_1 \text{ and } A\} \)
Conditional probability

\[ \mathbb{P}\{E_2 \mid E_1 \text{ and } A\} = \frac{\mathbb{P}\{E_2 \text{ and } E_1 \mid A\}}{\mathbb{P}\{E_1 \mid A\}} \]

Law of total probability

If \( E_1, E_2, \ldots, E_n \) are a partition of \( A \), then

\[ \mathbb{P}\{B \mid A\} = \sum_{i=1}^{n} \mathbb{P}\{B \mid E_i \text{ and } A\} \mathbb{P}\{E_i \mid A\} \]

Genetics Example

**Problem:**
A single gene has a dominant allele \( A \) and recessive allele \( a \). A cross of \( AA \) versus \( aa \) leads to \( F_1 \) offspring of type \( Aa \). Two of these mice are crossed to get the \( F_2 \) 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 \( F_2 \) 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.)

- Begin by defining several events.
  
  \( D = \{\text{offspring is dominant}\} \)
  
  \( F = \{\text{father is dominant}\} \)
  
  \( H = \{\text{father is heterozygous for the trait}\} \)

- With this notation, we are asked to find \( \mathbb{P}\{H \mid D \text{ and } F\} \).

- Write down the given probabilities in terms of these events.

- For this problem, this assumes some background knowledge of genetics.

Example (cont.)

- We know the genotype of the mother \( (aa) \).

- We can’t compute \( \mathbb{P}\{D\} \) 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, \( \mathbb{P}\{D \mid H^c\} = 1 \).

- If the father’s genotype is heterozygous \( (Aa) \), then the offspring is equally likely to be dominant or recessive, \( \mathbb{P}\{D \mid H\} = 0.5 \).
Example (cont.)

- The father is part of the F2 generation.
- This implies that $P\{F\} = 3/4$ and $P\{H\} = 1/2$.
- Given that the father has the dominant trait affects the probability that he is heterozygous.
  
  \[
  P\{H \mid F\} = \frac{P\{H \text{ and } F\}}{P\{F\}} = \frac{P\{H\}}{3/4} = \frac{2}{3/4} = \frac{2}{3}
  \]

- Here, $P\{H \text{ and } F\} = P\{H\}$ because every heterozygote exhibits the dominant trait.

Example (cont.)

- A variation of the law of total probability applies to the denominator.
  
  \[
  P\{D \mid F\} = P\{H \mid F\}P\{D \mid H \text{ and } F\} + P\{H^c \mid F\}P\{D \mid H^c \text{ and } F\}
  \]
  
  \[
  = \left( \frac{2}{3} \times \frac{1}{2} \right) + \left( \frac{1}{3} \times 1 \right) = \frac{2}{3}
  \]

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

Example (cont.)

- The original question is to find $P\{H \mid D \text{ and } F\}$.
- Use a variation on the definition of conditional probability.
  
  \[
  P\{H \mid D \text{ and } F\} = \frac{P\{H \text{ and } D \mid F\}}{P\{D \mid F\}}
  \]

- A variation of the multiplication rule applies to the numerator.
  
  \[
  P\{H \text{ and } D \mid F\} = P\{H \mid F\}P\{D \mid H \text{ and } F\} = \frac{2}{3} \times \frac{1}{2} = \frac{1}{3}
  \]

Random Variables

Definition
A random variable is a numerical variable whose value depends on a chance outcome.

Definition
A discrete random variable has a discrete (possibly infinite) list of possible values.

Definition
A continuous random variable has a continuum of possible values. This course will not deal with random variables whose distributions are mixed.
Distributions of Random Variables

Definition
The probability distribution of a random variable specifies the probability that the random variable is in each possible 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 a probability mass function.
- A probability mass function is a table or formula that specifies the probability associated with each possible value.

Probability Density Curves

- A probability density is like an idealized histogram, scaled so that the total area under the curve is one.
- If \( a < b \) and \( Y \) is a continuous random variable, \( \mathbb{P}\{a < Y < b\} \) 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.

\[ \mathbb{P}\{1 < Y < 3\} = 0.382 \]

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.

A table display of a probability mass function:

<table>
<thead>
<tr>
<th>( y )</th>
<th>( \mathbb{P}{Y = y} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
</tr>
</tbody>
</table>

A formula specification of a probability mass function:

\[ \mathbb{P}\{Y = y\} = y/10 \text{ for } y = 1, 2, 3, 4 \]

Cumulative Distribution Functions

The cumulative distribution function (CDF) answers the question, how much probability is less than or equal to \( y \)?

Definition
The cumulative distribution function \( F \) is the function

\[ F(y) = \mathbb{P}\{Y \leq y\} \]
CDFs (cont.)

- 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.

Means of Random Variables

- The **mean** of a probability distribution is the location where the probability balances.
- The mean of a random variable is known as its **expected value**.

Expected Value

**Definition**

For discrete random variable $Y$ the mean $\mu_Y$ or expected value $E(Y)$ is

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

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.

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.

**Definition**

The variance of a discrete random variable $Y$ $\sigma_Y^2$ or $\text{Var}(Y)$ is defined as

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

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

**Definition**
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.

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.
- 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.

**Characteristics of Binomials**

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.

**Binomial Distribution (cont.)**

The probability mass function for the binomial distribution is

\[ P(Y = j) = \binom{n}{j} p^j (1 - p)^{n-j} \quad \text{for } j = 0, 1, \ldots, n \]

where

\[ \binom{n}{j} = \frac{n!}{j!(n-j)!} \]

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 \( \binom{n}{j} \) different such sequences.
Moments of Binomial Distributions

- The mean of a binomial distribution is $\mu = np$.
- The variance of a binomial distribution is $\sigma^2 = np(1 - p)$.
- 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.
- You must be able to recognize binomial random variables in problem descriptions.
- 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.