STATS8: Introduction to Biostatistics

Probability

Babak Shahbaba
Department of Statistics, UCI
Introduction

• We have used plots and summary statistics to learn about the distribution of variables and to investigate their relationships.

• We now want to generalize our findings to the population.

• However, we almost always remain uncertain about the true distributions and relationships in the population.

• Therefore, when we generalize our findings from a sample to the whole population, we should explicitly specify the extent of our uncertainty.

• We now discuss probability as a measure of uncertainty.

• We use some examples from genetics.
Some Commonly Used Genetic Terms

- Gene
- Single Nucleotide Polymorphisms (SNPs)
- Alleles
- Genotype
- Homozygous vs. heterozygous
- Phenotype
- Recessive vs. dominant
Random phenomena and their sample space

- A phenomenon is called *random* if its outcome (value) cannot be determined with certainty before it occurs.

- For example, coin tossing and genotypes are random phenomena.

- The collection of all possible outcomes $S$ is called the sample space.

  Coin tossing: $S = \{H, T\}$,
  Die rolling: $S = \{1, 2, 3, 4, 5, 6\}$,
  Bi-allelic gene: $S = \{A, a\}$,
  Genotype: $S = \{AA, Aa, aa\}$. 
Probability

• To each possible outcome in the sample space, we assign a probability \( P \), which represents how certain we are about the occurrence of the corresponding outcome.

• For an outcome \( o \), we denote the probability as \( P(o) \), where \( 0 \leq P(o) \leq 1 \).

• The total probability of all outcomes in the sample space is always 1.

Coin tossing: \( P(H) + P(T) = 1 \),

Die rolling: \( P(1) + P(2) + P(3) + P(4) + P(5) + P(6) = 1 \).

• Therefore, if the outcomes are equally probable, the probability of each outcome is \( 1/n_S \), where \( n_S \) is the number of possible outcomes.
Random events

• An **event** is a subset of the sample space $S$.

• A possible event for die rolling is $E = \{1, 3, 5\}$. This is the event of rolling an odd number.

• For the genotype example, $E = \{AA, aa\}$ is the event that a person is homozygous.

• An event occurs when any outcome within that event occurs.

• We denote the probability of event $E$ as $P(E)$.

• The probability of an event is the sum of the probabilities for all individual outcomes included in that event.
Example

- As a running example, we consider a bi-allelic gene $A$ with two alleles $A$ and $a$.

- We assume that allele $a$ is recessive and causes a specific disease.

- Then only people with the genotype $aa$ have the disease.
Example

- We can define four events as follows:
  
  The homozygous event:  \( HM = \{ AA, aa \} \),
  
  The heterozygous event:  \( HT = \{ Aa \} \),
  
  The no-disease event:  \( ND = \{ AA, Aa \} \),
  
  The disease event:  \( D = \{ aa \} \).

- Assume that the probabilities for different genotypes are  
  \( P(AA) = 0.49 \),  \( P(Aa) = 0.42 \), and  \( P(aa) = 0.09 \).

- Then,

  \[
  P(HM) = 0.49 + 0.09 = 0.58, \\
  P(HT) = 0.42, \\
  P(ND) = 0.49 + 0.42 = 0.91, \\
  P(D) = 0.09.
  \]
Complement

• For any event $E$, we define its complement, $E^c$, as the set of all outcomes that are in the sample space $S$ but not in $E$.

• For the gene-disease example, the complement of the homozygous event $HM = \{AA, aa\}$ is the heterozygous event $\{Aa\}$; we show this as $HM^c = HT$.

• Likewise, the complement of the disease event, $D = \{aa\}$, is the no-disease event, $ND = \{AA, Aa\}$; we show this as $D^c = ND$.

• The probability of the complement event is 1 minus the probability of the event:

$$P(E^c) = 1 - P(E).$$
Union

- For two events $E_1$ and $E_2$ in a sample space $S$, we define their union $E_1 \cup E_2$ as the set of all outcomes that are at least in one of the events.

- The union $E_1 \cup E_2$ is an event by itself, and it occurs when either $E_1$ or $E_2$ (or both) occurs.

- For example, the union of the heterozygous event, $HT$, and the disease event, $D$, is $\{Aa\} \cup \{aa\} = \{Aa, aa\}$.

- When possible, we can identify the outcomes in the union of the two events and find the probability by adding the probabilities of those outcomes.
Intersection

• For two events \( E_1 \) and \( E_2 \) in a sample space \( S \), we define their intersection \( E_1 \cap E_2 \) as the set of outcomes that are in both events.

• The intersection \( E_1 \cap E_2 \) is an event by itself, and it occurs when both \( E_1 \) and \( E_2 \) occur.

• The intersection of the heterozygous event and the no-disease event is \( HM \cap ND = \{AA\} \).

• When possible, we can identify the outcomes in the union of the two events and find the probability by adding the probabilities of those outcomes.
Joint vs. marginal probability

- We refer to the probability of the intersection of two events, \( P(E_1 \cap E_2) \), as their **joint probability**.

- In contrast, we refer to probabilities \( P(E_1) \) and \( P(E_2) \) as the **marginal probabilities** of events \( E_1 \) and \( E_2 \).

- For any two events \( E_1 \) and \( E_2 \), we have

\[
P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).
\]

- That is, the probability of the union \( P(E_1 \cup E_2) \) is the sum of their marginal probabilities minus their joint probability.

- The union of the heterozygous and the no-disease events is

\[
P(HM \cup ND) = P(HM) + P(ND) - P(HM \cap ND)
\[
= 0.58 + 0.91 - 0.49 = 1.
\]
Disjoint events

• Two events are called **disjoint** or **mutually exclusive** if they never occur together: if we know that one of them has occurred, we can conclude that the other event has not.

• Disjoint events have no elements (outcomes) in common, and their intersection is the empty set.

• For the above example, if a person is heterozygous, we know that he does not have the disease so the two events $HT$ and $ND$ are disjoint.
Disjoint events

• For two disjoint events $E_1$ and $E_2$, the probability of their intersection (i.e., their joint probability) is zero:

$$P(E_1 \cap E_2) = P(\emptyset) = 0$$

• Therefore, the probability of the union of the two disjoint events is simply the sum of their marginal probabilities:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

• In general, if we have multiple disjoint events, $E_1, E_2, \ldots, E_n$, then the probability of their union is the sum of the marginal probabilities:

$$P(E_1 \cup E_2 \cup \ldots \cup E_n) = P(E_1) + P(E_2) + \ldots + P(E_n)$$
Partition

- When two or more events are disjoint and their union is the sample space $S$, we say that the events form a partition of the sample space.

- Two complementary events $E$ and $E^c$ always form a partition of the sample space since they are disjoint and their union is the sample space.
Conditional probability

- Very often, we need to discuss possible changes in the probability of one event based on our knowledge regarding the occurrence of another event.

- The **conditional probability**, denoted $P(E_1|E_2)$, is the probability of event $E_1$ given that another event $E_2$ has occurred.

- The conditional probability of event $E_1$ given event $E_2$ can be calculated as follows: (assuming $P(E_2) \neq 0$)

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}.$$ 

- This is the joint probability of the two events divided by the marginal probability of the event on which we are conditioning.
Conditional probability

- Consider the gene-disease example. Suppose we know that a person is homozygous and are interested in the probability that this person has the disease, $P(D|HM)$.

- The probability of the intersection of $D$ and $HM$ is $P(D \cap HM) = P(\{aa\}) = 0.09$.

- Therefore, the conditional probability of having the disease knowing that the genotype is homozygous can be obtained as follows:

$$P(D|HM) = \frac{P(D \cap HM)}{P(HM)} = \frac{0.09}{0.58} = 0.16.$$  

- In this case, the probability of the disease has increased from $P(D) = 0.09$ to $P(D|HM) = 0.16$. 
The law of total probability

- By rearranging the equation for conditional probabilities, we have

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

- Now suppose that a set of \( K \) events \( B_1, B_2, \ldots, B_K \) forms a partition of the sample space.

- Using the above equation, we have

\[ P(A) = P(A \mid B_1)P(B_1) + \cdots + P(A \mid B_K)P(B_K). \]

- This is known as the **law of total probability**.
Independent events

- Two events $E_1$ and $E_2$ are independent if our knowledge of the occurrence of one event does not change the probability of occurrence of the other event.

\[
P(E_1|E_2) = P(E_1),
\]
\[
P(E_2|E_1) = P(E_2).
\]

- For example, if a disease is not genetic, knowing a person has a specific genotype (e.g., AA) does not change the probability of having that disease.
Independent events

• When two events $E_1$ and $E_2$ are independent, the probability that $E_1$ and $E_2$ occur simultaneously, i.e., their joint probability, is the product of their marginal probabilities:

$$P(E_1 \cap E_2) = P(E_1) \times P(E_2).$$

• Therefore, the probability of the union of two independent events is as follows:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1) \times P(E_2).$$
Bayes’ theorem

- Sometimes, we know the conditional probability of $E_1$ given $E_2$, but we are interested in the conditional probability of $E_2$ given $E_1$.

- For example, suppose that the probability of having lung cancer is $P(C) = 0.001$ and that the probability of being a smoker is $P(SM) = 0.25$.

- Further, suppose we know that if a person has lung cancer, the probability of being a smoker increases to $P(SM|C) = 0.40$.

- We are, however, interested in the probability of developing lung cancer if a person is a smoker, $P(C|SM)$. 
Bayes’ theorem

- In general, for two events $E_1$ and $E_2$, the following equation shows the relationship between $P(E_2|E_1)$ and $P(E_1|E_2)$:

$$P(E_2|E_1) = \frac{P(E_1|E_2)P(E_2)}{P(E_1)}.$$

- This formula is known as Bayes’ theorem or Bayes’ rule.

- For the above example,

$$P(C|SM) = \frac{P(SM|C)P(C)}{P(SM)} = \frac{0.4 \times 0.001}{0.25} = 0.0016.$$

- Therefore, the probability of lung cancer for smokers increases from 0.001 to 0.0016.