9.07 Introduction to Probability and Statistics for Brain and Cognitive Sciences
Emery N. Brown

Lecture 1: Introduction to Probability Theory

I. Objectives

Introduce the basic concepts of probability theory

Introduce the basic axioms and rules of probability theory

Learn to perform probability computations using counting methods

Understand conditional probability and Bayes' rule

Understand the concept of independence

II. Concepts of Probability (DeGroot and Schervish, 2002)

Probability Theory is the branch of mathematics that is concerned with the analysis of random phenomena or chance. **Statistics** is the science of making decisions under uncertainty. Probability models are used to formulate our understanding of uncertainty, stochastic behavior and or noise in data analyses. Therefore, we study some basic probability theory that will be useful for developing our statistical models and carrying out statistical analyses. Probability theory is what we will study in the first seven lectures of the course.

There are three interpretations of probability. The mathematical theory of probability which we outline here applies to all three.

A. The Frequency Interpretation of Probability. In this interpretation of probability, the probability of a specific outcome of a process means the *relative frequency* with which this outcome would occur if the process were repeated a large number of times under similar conditions. The classic example is tossing a fair coin. We would expect the relative frequency of either heads or tails to be $1/2$. A few issues to consider with this definition are listed below.

How large is large?

How do you insure identical conditions?

Is a coin toss really random? See the work of Persi Diaconis on this problem.

How far away from ½ can the tosses be and have the relative frequency still be interpretable as ½?

In principle, the relative frequency definition applies only to problems in which the process can be repeated a large number of times. Would this condition apply in neuroscience experiments?

B. The Classical Interpretation of Probability. This concept of probability is based on the notion of equally likely outcomes. For example, in a coin toss there are two outcomes: a heads or a tail. If the outcomes are equally likely to occur then each must have probability of ½. Similarly if there are n events each of which is equally likely, then the probability of each must be $1/n$ because the total probability must sum to 1. The difficulty here is that:

The concept of equally likely outcomes is based on a notion of probability which is what the definition seeks to define.

There is no systematic approach to assigning probabilities to outcomes that are not assumed to be equally likely.

C. The Subjective Interpretation of Probability. Under the subjective interpretation of probability, the probability that a person assigns to a possible outcome of some process represents his or her own judgment of the likelihood that the outcome will occur. The judgment will be based on each person's beliefs and understanding of what is known about the process. This belief must be expressed numerically. For example, a person with no special knowledge about a coin toss may assign a probability of ½ to the occurrence of either a heads or a tails. Similarly, a person who knows that a box contains 5 pennies, 4 copper and 1 steel when asked, what is the probability of choosing a copper penny would state 4/5. For this concept of probability we know that:

The subjective interpretation can be formalized.

This formalism requires subjective assignments of probabilities to a possibly infinite set of outcomes to be logically consistent. While possible it is often difficult to do.

The subjective interpretation appears to contain no objective way for two or more scientists to combine objectively their beliefs about the probability of a particular outcome.

The subjective notion of probability reflects the subjective notion of science. Remember experts in the same field can make drastically different predictions based an analysis of the same data. Also science is not objective. Scientists decide which problems to study, which experiments to do, which data to collect and also, how to interpret the findings.

In this course we will make use of all three interpretations of probability.

III. Axioms and Rules of Probability Theory

A. Basic Concepts of Set Theory

We require some basic concepts from set theory to formulate our concepts of probability theory. Probability theory begins by defining for any problem a triplet $(\Omega, \Im, \mathrm{P}),$ where Ω is the collection of all possible outcomes also termed the sample or **outcome space**. The elements of Ω are called **events**. $\Im$ is the **family** of objects of Ω or the non-empty collection of subsets of $\Omega.$ We define the specific properties of $\Im$ below in **Definition 1.1**. P is the **probability measure** or **probability rule** which is a function from $\Im$ into $[0,1]$ that assigns to any event $A \in \Im$ $P(A),$ the probability of the event $A.$

Example 1.1 Single Trial of a Learning Experiment. In a learning experiment a subject is given multiple trials to learn a task. On any trial the subject either responds correctly or incorrectly. Lets call a 1 a correct response and 0 an incorrect response. Hence,

$$\Omega = \{0,1\}.$$

In this example there are only two events $A = \{0\}$ and $B = \{1\}.$

Example 1.2 Nucleotides. Deoxyribonucleic acid (DNA) is composed of sequences of nucleotides. The nucleotides are adenine (A), guanine (G), cytosine (C) and thiamine (T). If we are interested in the possible nucleotide types that could appear at a given location in DNA then the outcome space is

$$\Omega = \{A, C, G, T\}.$$

Example 1.3 Interspike Intervals. Neurons transmit information through the nervous system by discharging and propagating electrical impulses called action potentials. Neurophysiologists record the action potentials or spiking events and keep track of the spike event times or equivalently the time between spike events termed the interspike intervals. If we disregard for the moment the absolute refractory period of a neuron, then the interspike intervals could be any positive number. In this case, the outcome space would be

$$\Omega = \{t \mid t > 0\}.$$

The **union** of two events A and B is the event C that either A occurs, B occurs or both occur. We write this as $C = A \bigcup B.$ In **Example 1.2**, A is the event that the nucleotide at the given DNA location is A or G and B is the event that the nucleotide is either G or T then $C = A \bigcup B = \{A, G, T\}.$ The **intersection** of two events $C = A \bigcap B$ is the event that both A and B occur. For the case just cited $C = A \bigcap B = \{G\}.$ In **Example 1.3**, suppose that the neuron is a hippocampal pyramidal neuron that discharges in bursts, i.e. ISI's of between 3 to 20 msec, and has modulation by the theta rhythm ISI's of 110 to 140 msec. If $A = \{t \mid t \in [3,20]\}$ and $B = \{t \mid t \in [110,140]\}$ then $A \bigcap B = \{t \mid t \in [3,20] \text{ and } t \in [110,140]\}$ and $A \bigcap B = \varnothing$ where $\varnothing$ denotes the empty set. If two events have no elements in their intersection then they are said to be **disjoint.** The **complement** of an event A, A^c is the event that A does not occur. Stated otherwise, A^c are all of the events except A. The complement of $A = \{t \mid t \in [3,20]\}$ is $A^c = \{t \mid 0 < t < 3 \bigcup t > 20\}.$

The family of events of Ω has important special properties which we define below.

Definition 1.1. A non-empty collection of subsets of $\Im$ is called a **family** of subsets of Ω provided that the following three properties hold:

i) If $A \in \Im$ then $A^c \in \Im$

ii) If $A_n \in \Im \quad n = 1, 2, \ldots,$ then $\bigcup_{n=1}^{\infty} A_n$ and $\bigcap_{n=1}^{\infty} A_n$ are both in $\Im$ (1.1)

iii) $\Omega \in \Im$

Example 1.2 (continued). How big is $\Im$? In this example, we can easily list all of the subsets. They are

$$\begin{array}{cccc}
\{A\} & \{C\} & \{G\} & \{T\} \\
\{AC\} & \{AG\} & \{AT\} & \{CT\} \\
\{CG\} & \{CT\} & \{ACG\} & \{ACT\} \\
\{AGT\} & \{CGT\} & \{\varnothing\} & \{\Omega\}
\end{array}$$

In this case we see that there are 16 subsets. We have $2^4 = 16.$ This example illustrates the general result that if a set has a finite number of elements $n,$ then the number of subsets is $2^n.$

Venn diagrams are often useful tools for visualizing set theoretic operations. We summarize here some elementary laws of set theory.

Commutative Laws

$$A \bigcup B = B \bigcup A$$

$$A \bigcap B = B \bigcap A$$

Associative Laws

$$(A \bigcup B) \bigcup C = A \bigcup (B \bigcup C)$$

$$(A \bigcap B) \bigcap C = A \bigcap (B \bigcap C)$$

Distributive Laws

$$(A \bigcup B) \bigcap C = (A \bigcap C) \bigcup (B \bigcap C)$$

$$(A \bigcap B) \bigcup C = (A \bigcup C) \bigcap (B \bigcup C)$$

B. Axioms of Probability Theory

Definition 1.2 (Axioms of Probability Theory). A probability law or rule P on a family of subsets of $\Im$ is a real-valued function having domain $\Im$ and satisfying the following properties:

i) $\Pr(\Omega) = 1$

ii) $\Pr(A) \geq 0$ for all $A \in \Im$

iii) If $A_n, n = 1, 2, ...$ are mutually disjoint events in $\Im,$ then (1.2)

$$\Pr(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \Pr(A_n)$$

C. Elementary Rules of Probability Theory

We can easily derive the following properties from **Definitions 1.1** and **1.2.**

Proposition 1.1. $\Pr(A^c) = 1 - \Pr(A)$

Proof: $A \bigcup A^c = \Omega$. $\Pr(A \bigcup A^c) = \Pr(A) + \Pr(A^c) = 1$ or $\Pr(A^c) = 1 - \Pr(A)$.

Proposition 1.2. $\mathrm{P}(\varnothing) = 0$

Proof: $\Pr(\varnothing) = 1 - \Pr(\Omega) = 1 - 1 = 0$.

Proposition 1.3. If $A \subset B$ then $\Pr(A) \leq \Pr(B)$

Proof: $B = A \bigcup (B \bigcap A^c) = (B \bigcap A) \bigcup (B \bigcap A^c)$. $\Pr(A) = \Pr(B) - \Pr(B \bigcap A^c) \leq \Pr(B)$.

Proposition 1.4. (Addition Law). $\Pr(A \bigcup B) = \Pr(A) + \Pr(B) - \Pr(A \bigcap B)$.

Proof: Let $C = A \bigcap B^c, D = A \bigcap B, E = A^c \bigcap B$

Hence,

$$\Pr(A \bigcup B) = \Pr(C) + \Pr(D) + \Pr(E) \tag{1.3}$$

$A = C \bigcup D$ and C and D are disjoint, so

$$\Pr(A) = \Pr(C) + \Pr(D). \tag{1.4}$$

Similarly, $\Pr(B) = \Pr(D) + \Pr(E)$. Therefore, we have

$$\begin{aligned} \Pr(A) + \Pr(B) &= \Pr(C) + \Pr(E) + 2\Pr(D) \\ &= \Pr(A \bigcup B) + \Pr(D) \\ &= \Pr(A \bigcup B) + \Pr(A \bigcap B) \end{aligned} \tag{1.5}$$

or

$$\Pr(A \bigcup B) = \Pr(A) + \Pr(B) - \Pr(A \bigcap B). \tag{1.6}$$

Example 1.1. (continued). Suppose we consider two trials in this experiment. Let A be the event of a correct response on the first trial and B be the event of an incorrect response on the second trial. The sample space is

$$\Omega = \{00, 01, 10, 11\}$$

We assume that there is no learning so that each outcome is equally likely and the probability of each outcome is thus, ¼. $C = A \bigcup B$ is the event that the response is correct on the first trial or incorrect on the second trial. We have $\Pr(C) \neq \Pr(A) + \Pr(B) = 1$. We have that $A \bigcap B$ is the event that a correct response is given on the first trial and an incorrect response is given on the second trial. These events are $A = \{10, 11\}$, $B = \{00, 10\}$ and $A \bigcap B = \{10\}$ and we have

$$\Pr(C) = \Pr(A) + \Pr(B) - \Pr(A \bigcap B) = 0.5 + 0.5 - 0.25 = 0.75. \tag{1.7}$$

IV. Computing Probabilities Using Counting Methods
For finite sample spaces probabilities are easy to compute. If $\Omega = \{\omega_1, ..., \omega_n\}$, a set of disjoint events, and $\Pr(\omega_i) = p_i$, then to find the probability of an event A in Ω, it suffices to compute

$$\Pr(A) = \sum_{\omega_j \in A} \Pr(\omega_j).$$

Example 1.1 (continued). If the subject executes two trials of the experiment then the outcome space is $\Omega = \{00, 01, 10, 11\}$. If A is the event that the subject gives at least one correct response then $A = \{01, 10, 11\}$ and $\Pr(A) = 0.75$. This is an example of a common situation. If Ω has n outcomes and each is equally likely and **mutually exclusive** (disjoint), and if A consists of k mutually exclusive events then

$$\Pr(A) = \frac{k}{n} = \frac{\text{Number of ways } A \text{ can occur}}{\text{Total number of outcomes}}.$$

Definition 1.3 (The Multiplication Principle). If there are p experiments and the first has n_1 outcomes, the second $n_2, \ldots,$ and the p^{th} has n_p outcomes, then the total number of outcomes is $n_1 \times n_2, \ldots, \times n_p = \prod_{i=1}^{p} n_i$.

Proof: By induction, if $p = 2$ then there are n_1 outcomes for the first experiment and n_2 for the second. The n_1 choices for the first experiment and each can be paired with the n_2 possibilities from the second experiment. Hence, there are $n_1 \times n_2$ possible outcomes. If we assume the result is true for a study with $p-1$ experiments, then there are $\prod_{i-1}^{p-1} n_i$ outcomes for the first $p-1$ experiments and each one of these outcomes can be paired with n_p outcomes of experiment $p.$ Therefore, there are $(\prod_{i=1}^{p-1} n_i) \times n_p = \prod_{i=1}^{p} n_i.$

Example 1.1 (continued). If there are p trials in this learning experiment, then the total number of outcomes is $\prod_{i=1}^{p} n_i = \prod_{i=1}^{p} 2 = 2^p.$ How many subsets are there in this experiment?

Example 1.2 (Continued). The genetic code or DNA sequence for an amino acid consists of a triplet of three nucleotides. How many possible amino acids could there be in theory? The actual number is 20. Because there are four nucleotides A, C, G and T and three positions to fill to make a triplet then there are $4 \times 4 \times 4 = 4^3 = 64$ possible amino acids. How many elements are there in $\Im$ for this problem?

A. Permutations (Sampling without replacement with regard to order)

To understand approaches to counting possible outcomes it is important to understand two important concepts of sampling. These are sampling **with replacement** and sampling **without replacement**. In sampling without replacement there are two cases: **with regard to order** or **without regard to order**.

Definition 1.4 Given a population with n elements, a sample of size $r \le n$ is drawn **without replacement** if the element selected is not returned to the population after each draw. The number of possible samples is $n(n-1)(n-2)...(n-r+1)$. If $n=r$ then the number of possible samples is $n(n-1)(n-2)...2 \cdot 1 = n!$, termed n **factorial**.

Definition 1.5 Each ordered arrangement of objects is called a **permutation**.

Example 1.2 (continued). If we sample the possible nucleotides with replacement how many amino acids can be coded for in theory if a sequence of three nucleotides is required to code for an amino acid? If we take $n=4$ and $r=3$ then as we showed above, we have the number of amino acids is $4^3 = 64$. If we sample the nucleotides without replacement, we construct nucleotide sequences with no repeats. In this case, the number of amino acids that have a distinct sequence of nucleotides is $4 \times 3 \times 2 = 24$.

Notice that we can write $n(n-1)...(n-r+1)$ as

$$P_{n,r} = \frac{n(n-1)...(n-r+1)(n-r)...2 \cdot 1}{(n-r)...2 \cdot 1} = \frac{n!}{(n-r)!}. \tag{1.8}$$

Equation 1.8 defines the number of permutations of n objects taken r at a time for $r \le n$.

Example 1.4 (The Birthday Problem). Suppose there are n people in a room. What is the probability that at least two have a common birthday? We assume that every day is equally likely and that there are no leap years or wars (why do we make this assumption?). Let A be the event that at least two people have a common birthday. Instead of $\Pr(A)$, we consider $\Pr(A^c)$. The outcome space Ω has $(365)^n$ outcomes because the n people could have been born on any one of the 365 days. The event A^c occurs if each person is born on a different day. This event has $\frac{365!}{(365-n)!}$ outcomes. Therefore,

$$\Pr(A^c) = \frac{365!}{(365-n)!(365)^n} \tag{1.9}$$

and

$$\Pr(A) = 1 - \Pr(A^c) = 1 - \frac{365!}{(365-n)!(365)^n}. \tag{1.10}$$

As a function of n the probability is

n	$\Pr(A)$
4	0.016
16	0.284
23	0.507
32	0.753
40	0.891
56	0.988

This result is easy to understand if you think of this problem as the problem of throwing n balls into 365 urns and requiring that no urn has two or more balls. The event A^c is the event that every urn has at most one ball and A is the complement of this event.

B. Combinations (Sampling without replacement without regard to order)

Example 1.2 (continued). Suppose the order of the nucleotides does not matter to code for an amino acid. How many amino acids would there be? For example, if we disregard order the nucleotide sequences $\{A,C,T\},\{C,A,T\}\{T,A,C\}\{C,T,A\}\{A,T,C\}\{T,C,A\}$ are equivalent and since there are 3 nucleotides the number of ways to order them is $3!$ Hence, we have the number of acids, disregarding order is $\frac{4!}{1!3!}=4.$ Notice that this is less than the 24 amino acids we realized we could obtain if we sampled without replacement with regard to order and the 64 we could obtain if we sampled with replacement.

Proposition 1.5. The number of unordered samples of r objects selected without replacement is $\binom{n}{r}$. The number $C_{n,r}=\binom{n}{r}$, is read "n choose r" and is called the **binomial coefficient.** It was the number of ways of choosing exactly r objects from a group of n without replacement and without regard to order. The binomial coefficient comes from the binomial expansion

$$(a+b)^n=\sum_{k=0}^{n}\binom{n}{k}a^k b^{n-k}. \tag{1.11}$$

We have the special case

$$2^n=\sum_{k=0}^{n}\binom{n}{k}, \tag{1.12}$$

which, as we mentioned above, defines the number of subsets of a set of n objects. We use the convention that $0!=1.$

Proof: By the multiplication principle, the number of ordered samples equals the number of unordered samples multiplied by the number of ways to order the samples. The number of ordered samples is $\frac{n!}{(n-r)!}$ and because a sample of size r can be ordered in $r!$ ways, the number of unordered samples is $\frac{n!}{(n-r)!r!}$.

V. Conditional Probability and Bayes' Rule

A. Conditional Probability

Conditional probability allows us to assess how likely one event is given that another has happened. If A and $B \in \Omega$, then

$$\Pr(A \mid B) = \frac{\Pr(A \cap B)}{\Pr(B)}. \tag{1.13}$$

We read this as the probability of A given B. Heuristically, we can think of this as coming from

$$\begin{aligned}\Pr(A \mid B) &= \frac{(\text{Area of } A \cap B)}{(\text{Area } B)} = \frac{(\text{Area of } A \cap B) / \text{Area of } (\Omega)}{(\text{Area of } B) / \text{Area of } (\Omega)} \\ &= \frac{\Pr(A \cap B)}{\Pr(B)}\end{aligned} \tag{1.14}$$

By a similar argument, we have

$$\Pr(B \mid A) = \frac{\Pr(A \cap B)}{\Pr(A)} \tag{1.15}$$

or we have

$$\Pr(B \mid A) = \frac{\Pr(A)\Pr(B \mid A)}{\Pr(A)} \tag{1.16}$$

If we write $\Pr(A)\Pr(B \mid A) = \Pr(B)\Pr(A \mid B)$ we have the **Multiplication Rule of Probability**. Given an event A and a disjoint partition of $\Omega = B = \bigcup_{i=1}^{n} B_i$, we have that

$$\Pr(\Omega) = \Pr(B) = \sum_{i=1}^{n} \Pr(B_i) \tag{1.17}$$

then

$$\Pr(A) = \Pr(A \cap B) = \sum_{i=1}^{n} \Pr(B_i \cap A) = \sum_{i=1}^{n} \Pr(B_i)\Pr(A \mid B_i). \tag{1.18}$$

The above result is sometimes referred to as the **Law of Total Probability**. Now for $j = 1, \ldots, n$ we may write

$$\Pr(B_j \mid A) = \frac{\Pr(A \cap B_j)}{\Pr(A)} = \frac{\Pr(B_j)\Pr(A \mid B_j)}{\sum_{i=1}^{n} \Pr(B_i)\Pr(A \mid B_i)} \qquad (1.19)$$

This last expression is **Bayes' Rule**. In its simplest form, it is merely a re-statement of the **Multiplication Rule of Probability.**

B. Bayes' Rule and Screening Tests

An important application of Bayes' Rule as stated in Eq. 1.19 is to conduct screening for disease processes given a symptom or a test. We consider an example to illustrate this point.

Example 1.5 Screening for Multiple Sclerosis (Dangood, 2005). Multiple sclerosis (MS) is a chronic, inflammatory disease that affects the central nervous system. MS can cause a variety of symptoms, including changes in sensation, visual problems, muscle weakness, depression, difficulties with coordination and speech, severe fatigue, and pain. MS will cause impaired mobility and disability in more severe cases. Multiple sclerosis affects neurons. Surrounding and protecting some of these neurons is a phospholipid layer known as the myelin sheath, which helps neurons transmit their electrical impulses with less degradation. MS causes gradual destruction of myelin (demyelination) and transection of axons in patches throughout the brain and spinal cord. The name *multiple sclerosis* refers to the multiple scars (or scleroses) on the myelin sheaths. This scarring causes symptoms which vary widely depending upon which signals are interrupted. The main theory today is that MS results from attacks by an individual's immune system on the nervous system. For this reason MS is usually considered an autoimmune disease. There is also a view that MS is not an autoimmune disease, but rather a metabolically dependent neurodegenerative disease.

Let $n = 2$ in Eq. 1.19 and suppose that A is the event of a positive result from a new genetic screening test for multiple sclerosis. Let B_1 be the event that the patient has multiple sclerosis (MS) and let B_2 be the event that he or she does not. We can use this problem to make some useful definitions for screening in terms of Bayes' rule.

The **sensitivity** of the test for MS is the probability of observing a positive test result given that the patient has MS $\Pr(A \mid B_1)$. The **specificity** of the screening test for MS is the probability of having a negative test given that the patient does not have MS and is defined as $\Pr(A^c \mid B_1^c) = \Pr(A^c \mid B_2)$. The **predictive value positive** of the test for MS is the probability the patient has MS given that he or she has a positive test and is defined as $\Pr(B_1 \mid A)$. The **predictive value negative** of the test for MS is the probability the patient does not have MS given that he or she has a negative test and is defined as $\Pr(B_1{}^c \mid A^c) = \Pr(B_2 \mid A^c)$. The **prevalence** of MS is the probability of observing the disease in the population and is given as $\Pr(B_1)$. If the probability of cases of MS is considered in a specific time interval such as a year, then $\Pr(B_1)$ is the **incidence** of MS.

Let us assume that the properties of the test may be summarized as follows: In persons with MS, the test will be positive in 98% of them, whereas if in persons that do not have MS the test will be positive in 5% of them. At present, the prevalence of MS in the US is 0.14%. Let $B = B_1$ and $B^c = B_2$. These data suggest that

$$\begin{aligned}\Pr(A \mid B) &= 0.98 \\ \Pr(A \quad \mid B^c) &= 0.05 \\ \Pr(B) &= 0.0014\end{aligned} \tag{1.20}$$

Suppose we want to determine the probability $\Pr(A)$ that an arbitrary person tests positive. The tested person either has MS or does not. The event A occurs in combinations with B and B^c. There are no other possibilities. In terms of events we have

$$A = (A \cap B) \cup (A \cap B^c) \tag{1.21}$$

and

$$\Pr(A) = \Pr(A \cap B) + \Pr(A \cap B^c) \tag{1.22}$$

because $A \cap B$ and $A \cap B^c$ are disjoint events. We apply the multiplication rule in such a way that the known probabilities appear.

These definitions are important to know particularly if your work involves the clinical and epidemiological neuroscience literature. Bayes' Rule and our Venn diagrams make explicit the relationships among these terms. Hence, we have

$$\Pr(A \cap B) = \Pr(A \mid B)\Pr(B) \tag{1.23}$$

$$\Pr(A \cap B^c) = \Pr(A \mid B^c)\Pr(B^c) \tag{1.24}$$

and we have by applying the **Law of Total Probability** (Eq. 1.18)

$$\Pr(A) = \Pr(A \mid B)\Pr(B) + \Pr(A \mid B^c)\Pr(B^c) \tag{1.25}$$

From the data above it follows that $\Pr(B^c) = 1 - \Pr(B) = 1 - 0.0014 = 0.9986$ and that the probability of a positive test is

$$\Pr(A) = 0.98 \times 0.0014 + 0.05 \times 0.9986 = 0.051302. \tag{1.26}$$

The more important question to ask about this new test and MS is: Suppose that we have a positive test, what is the probability that the person has MS? To answer this question we compute

$$\begin{aligned}\Pr(B \mid A) &= \frac{\Pr(A \cap B)}{\mathrm{P}(A)} = \frac{\Pr(B)\Pr(A \mid B)}{\Pr(A)} \\ &= \frac{\Pr(B)\Pr(A \mid B)}{\Pr(A \mid B)P(B) + \Pr(A \mid B^c)P(B^c)} \\ &= \frac{0.98 \times 0.0014}{0.051032} = 0.026\end{aligned} \tag{1.27}$$

Hence, the predictive value positive for this test is 0.026.

Remark 1.1. If A is the event of a positive test and B is the event that a person has MS, a perfect test would have $\Pr(B \mid A) = 1$ (**predictive value positive**) and $\Pr(B^c \mid A^c) = 1$ (**predictive value negative**). The first statement says whenever (or given that) there is a positive test what is the probability the person has MS. The second statement says that whenever (or given that) there is a negative test what is the probability the person does not have MS. It would also be ideal to have $\Pr(B \mid A) = 1$ (**sensitivity**) and $\Pr(B^c \mid A^c) = 1$ (**specificity**). The sensitivity says given that the person has MS what is the probability that the test is positive. The specificity says given that the person does not have MS what is the probability that the test is negative. We should really have all of these characteristics to make a perfect test.

As a second application of Eq.1.19, let us consider the problem of decoding neural spiking activity from primary motor cortex.

Example 1.6. Reach Direction Given an Observed Neural Firing Pattern (Simplest Decoding Problem). Suppose that a monkey is making reaching movements with a manipulandum in 8 directions while spiking activity is being recorded from a set of single neurons in primary motor cortex. If A is an observed ensemble firing pattern, and B_j is the j^{th} direction, then $\Pr(B_j \mid A)$ above represents the probability that the observed firing pattern A encodes direction B_j. This is the simple model for neural spike train decoding that appeared in Sanger (1996) using a Poisson model.

VI. Independence

Two events E_1 and E_2 are independent if

$$\Pr(E_1 \cap E_2) = \Pr(E_1)\Pr(E_2) \qquad (1.28)$$

This implies that

$$\Pr(E_2 \mid E_1) = \Pr(E_1 \cap E_2)/\Pr(E_1) + \Pr(E_1)\Pr(E_2)/\Pr(E_1) = \Pr(E_2) \qquad (1.29)$$

Intuitively, this statement says that knowledge about E_1 gives no information about E_2. In general a set of n events $E_1, \ldots, E_n$ is **independent** if

$$\Pr(E_1 \cap E_2 \ldots \cap E_n) = \prod_{i=1}^{n} \Pr(E_i). \qquad (1.30)$$

Example 1.1 (continued). If we suppose that we record perform on three trials of the learning experiment, that performance the trials are independent and the probability of a correct response is $\frac{1}{2}$, then what is the probability of three correct responses? From Eq. 1.30, if we let $E_i = \{\text{correct response on trial } i\}$ for $i = 1,2,3,$ then we have

$$\Pr(E_1 \cap E_2 \cap E_3) = \prod_{i=1}^{3} \Pr(E_i) = \tfrac{1}{2} \times \tfrac{1}{2} \times \tfrac{1}{2} = \tfrac{1}{8}.$$

Independence is a very strong condition. To see this we consider the following example.

Example 1.7 (Pairwise Independent Events That Are Not Independent)

Let a ball be drawn from an urn containing 4 balls, numbered 1, 2, 3 and 4. Define the events $E_1 = \{1, 2\}$, $E_2 = \{1, 3\}$ and $E_3 = \{1, 4\}$

$$\Pr(E_1 \cap E_2) = \Pr(E_1)\Pr(E_2) = \frac{1}{4}$$

$$\Pr(E_2 \cap E_3) = \Pr(E_2)\Pr(E_3) = \frac{1}{4} \tag{1.31}$$

$$\Pr(E_1 \cap E_3) = \Pr(E_1)\Pr(E_3) = \frac{1}{4}$$

$$\Pr(E_1 \cap E_2 \cap E_3) = \Pr(\{1\}) = \frac{1}{4}$$

$$\Pr(E_1)\Pr(E_2)\Pr(E_3) = \frac{1}{8}$$

Hence $\Pr(E_1 \cap E_2 \cap E_3) \neq \Pr(E_1)\Pr(E_2)\Pr(E_3)$ and the events are pair-wise independent but not independent.

Proposition 1.6 We can now write a general statement about the probability of the intersection of events. Given events $E_1, E_2, \ldots\ , E_n$

i) $\Pr(E_1 \cap E_2 \cap E_3 \cap \ldots \cap E_n) = \prod_{i=2}^{n} \Pr(E_i |E_1 \cap E_2 \cap \ldots \cap E_{i-1})\Pr(E_1)$ (1.32)

if $E_1, E_2, \ldots, E_n$ are **independent** then

ii) $\Pr(E_1 \cap E_2 \cap E_3 \cap \ldots \cap E_n) = \prod_{i=1}^{n} \Pr(E_i),$ (1.33)

and if the events $E_1, \ldots, E_n$ have **Markov dependence**, i.e. $\Pr(E_i \mid E_1 \cap E_2 \cap \ldots \cap E_{i-1}) = \Pr(E_i \mid E_{i-1})$ then

iii) $\Pr(E_1 \cap E_2 \cap E_3 \cap \ldots \cap E_n = \prod_{i=2}^{n} \Pr(E_i \mid E_{i-1})\Pr(E_1)$ (1.34)

Proof: Let $B = E_n$ and $A = E_1 \cap E_2 \cap \ldots \cap E_{n-1}$. Then

$$\begin{aligned} \Pr(E_1 \cap E_2 \cap \ldots \cap E_n) &= \Pr(B \cap A) \\ &= \Pr(B \mid A)\Pr(A) \\ &= \Pr(E_n \mid E_1 \cap E_2 \cap \ldots \cap E_{n-1})\Pr(E_1 \cap E_2 \cap \ldots \cap E_{n-1}) \end{aligned} \tag{1.35}$$

Now let $B = E_{n-1}$ and $A = E_1 \cap E_2 \cap \ldots \cap E_{n-2}$. Then we get

$$\begin{aligned} \Pr(E_1 \cap E_2 \cap \ldots \cap E_n) &= \Pr(E_n \mid E_1 \cap E_2 \cap \ldots \cap E_{n-1})\Pr(B \cap A) \\ &= \Pr(E_n \mid E_1 \cap E_2 \cap \ldots \cap E_{n-1})\Pr(B \mid A)\Pr(A) \\ &= \Pr(E_n \mid E_1 \cap E_2 \cap \ldots \cap E_{n-1})\Pr(E_{n-1} \mid E_1 \cap E_2 \cap \ldots \cap E_{n-2})\Pr(E_1 \cap E_2 \cap \ldots \cap E_{n-2}) \end{aligned} \tag{1.36}$$

The final result in i) follows by repeated application of the multiplication rule of probability. To establish ii) we note that if the E_i are independent, then

$$\Pr(E_i \mid E_1 \cap E_2 \cap \ldots \cap E_{i-1}) = \Pr(E_i). \tag{1.37}$$

The result in iii) follows by substituting the right side of Eq. 1.37 into Eq. 1.32. Similarly, to establish iii) it suffices to substitute $\Pr(E_i \mid E_{i-1})$ into Eq. 1.32.

The result in Eq. 1.32 is a general statement about a general way to factor the joint probability of n events. It will be useful when we derive the joint distribution of point processes. Eq. 1.33 will be useful when we need to formulate the joint probability density of n independent random variables to carry out our likelihood analyses. Eq. 1.34 will be useful for our state-space analyses.

VII. Summary

In this lecture we have introduced the basic concepts, definitions, axioms and rules of probability theory along with standard counting methods for enumerating outcomes and computing probabilities when sampling with and without replacement. In addition, we introduced the basic concepts of conditional probabilities, Bayes' rule and independence. This material will be the building blocks for the work we do in probability theory and statistics.

VIII. References

Text References

DeGroot MH, Schervish MJ. *Probability and Statistics*, 3rd edition. Boston, MA: Addison Wesley, 2002.

Rice, JA, Mathematical Statistics and Data Analysis, 3rd edition. Boston, MA, 2007.

Rosner B. *Fundamentals of Biostatistics*, 6th edition. Boston, MA: Duxbury Press, 2006.

Literature References

Dangond, F.Multiple sclerosis. *eMedicine Neurology*.Updated 2005 Apr 25.

Sanger TD. Probability density estimation for the interpretation of neural population codes. *J Neurophys* 1996, 76: 2790-2793.

9.07 Statistics for Brain and Cognitive Science
Fall 2016