Chapter 1

Probability

1.1 Logic

Set theory language corresponds to logical language as follows. Let Ω be a set. Consider subsets A, B.

The *intersection* is $A \cap B$. The property $\omega \in A \cap B$ is equivalent to the property $\omega \in A$ *and* $\omega \in B$.

The *intersection* is $A \cup B$. The property $\omega \in A \cup B$ is equivalent to the property $\omega \in A$ *or* $\omega \in B$.

The *complement* is A^c. The property $\omega \in A^c$ is equivalent to the property $\omega \in \Omega$ but *not* $\omega \in A$.

When A is a *subset* of B, we write $A \subset B$. Then $A \subset B$ is equivalent to the condition that for every $\omega \in \Omega$ we have $\omega \in A$ *implies* $\omega \in B$.

Two subsets A, B are *exclusive* if $A \cap B = \emptyset$. This is the same as saying that for every $\omega \in \Omega$ the property that $\omega \in A$ and $\omega \in B$ is *impossible*.

1.2 Probability axioms

There is a fixed experiment. The *sample space* Ω is the set of all possible *outcomes* of the experiment.

An *event* is a subset of the sample space. Thus an event is a set of outcomes.

If the sample space is finite with r possible outcomes, then there are 2^r events.

A *probability measure* P assigns to each event A a probability $P[A]$ with $0 \leq P[A] \leq 1$. It must satisfy the following axioms:

Impossible event: $P[\emptyset] = 0$.

Sure event: $P[\Omega] = 1$.

Additivity: $A \cap B = \emptyset$ implies $P[A \cup B] = P[A] + P[B]$.

The additivity property generalizes to finite or countable sequences of incompatible events.

Some simple consequences:

$P[A^c] = 1 - P[A]$.
$P[A] = P[A \cap B] + P[A \cap B^c]$
$P[A \cup B] = P[A] + P[B] - P[A \cap B]$.
If $A \subset B]$, then $P[A] \subset P[B]$.

An event is a *singleton* if it has precisely one outcome in it. A probability measure is *discrete* if the probability of each event is the sum of the probabilities of its singleton subsets.

If Ω is finite or countable, then every probability measure must be discrete.

If the sample space Ω is finite with r points, then the *uniform* probability measure is the discrete probability measure that assigns to each singleton subset the probability $1/r$.

An example of a uniform probability measure is the one appropriate for tossing a pair of dice. There are 36 outcomes. The probability of an event is the number of points divided by 36.

This example gives rise to another example of a discrete probability measure that is not uniform. Say that one is only interested in the sum of the numbers on the two dice. The experiment discards all other information. Then one can take as the probability space the set $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. The probabilities of the corresponding singleton sets are 1/36, 1/18, 1/12, 1/9, 5/36, 1/6, 5/36, 1/9, 1/12, 1/18, 1/36.

An example where the sample space is infinite but countable is when a coin is tossed until a head appears. The sample space consists of the number of tosses, or ∞ if a head never appears. The probability of exactly k tosses until the first head appears is $1/2^k$. That is, $P[\{k\}] = 1/2^k$ for $k = 1, 2, 3, \ldots$. Also in this example $P[\{\infty\}] = 0$.

An example where the sample space is infinite and uncountable is the time to wait for a phone call when the average waiting time is 10 minutes. The sample space consists of all positive real numbers. The probability of each singleton set is zero. This just means that the probability that a call comes at an exactly specified time is zero. However not every set has probability zero. For instance, the set consisting of all real numbers greater than t has probability $P[(t, +\infty)] = e^{-\frac{1}{10}t}$. For instance, the probability of waiting more than 20 minutes is $e^{-2} = 0.135$.

In some counting problems one encounters numbers that are very large or very close to zero. These numbers are traditionally compared using *scientific notation*. This is the notation where a number $x > 0$ is expressed as $x = c \cdot 10^k$, where $1 \leq c < 10$, and k is an integer. For instance, Avogadro's number is $6.02 \cdot 10^{23}$. How can such a huge number arise? If one thinks of a cubic centimeter of atoms with 10^8 atoms on each side, then the total is 10^{24} atoms. It is illuminating to compare large numbers to such a reference number.

1.3 Conditional probability

Consider an experiment described by a particular probability measure P. Supposed that B is an event such that $P[B] > 0$. The *conditional probability* of A

given B is

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]}. \tag{1.1}$$

The intuitive meaning of conditional probability is that these probabilities describe a new experiment, which is the original experiment but with all outcomes that do not belong to B discarded.

A typical example is a situation where there are men and women who are either healthy or sick. Say that the experiment is to pick a person at random and check the person's health. Let B be the event that the person chosen is a man, and let B^c be the event that the person chosen is a woman. Let A be the event that the person chosen is healthy, and A^c be the event that the person chosen is sick. The conditioned experiment is to pick a person at random and discard the result unless the person is chosen is a man. If the person chosen is a man, then one checks whether the person is healthy or not. Thus $P[A]$ is the chance that a person is healthy, while $P[A \mid B]$ is the chance that a man is healthy.

The formula is often most useful in the form

$$P[A \cap B] = P[A \mid B]P[B]. \tag{1.2}$$

This says that to predict the chance that A and B both happen, first compute the probability that B happens, then the probability that A also happens, given that B happened.

Another useful formula is the *law of total probability*

$$P[A] = P[A \mid B]P[B] + P[A \mid B^c]P[B^c]. \tag{1.3}$$

This says that the probability of an event may be computed by conditioning on disjoint possibilities that exhaust all possibilities. Thus the probability that someone is healthy may be computed by conditioning on sex.

Example: Ordered sampling without replacement, sample size 2. Suppose that a rather small population has n members. Suppose that m of these are successes and $n - m$ of these are failures. Take an ordered sample of size two without replacement. Let B be the event that the first chosen is a success. Let A be the event that the second chosen is a success. Then $P[B] = m/n$, while $P[B^c] = (n - m)/n$. Furthermore, $P[A \mid B] = (m - 1)/(n - 1)$, while $P[A \mid B^c] = m/(n - 1)$. The law of total probability gives

$$P[A] = \frac{m-1}{n-1}\frac{m}{n} + \frac{m}{n-1}\frac{n-m}{n} = \frac{m}{n}. \tag{1.4}$$

This result seems quite simple; could one see why it might be true without doing the calculation?

1.4 The Bayes theorem

Let H be an event (the hypothesis), and let E be another event (the evidence). The prior probability of H is $P[H]$. The posterior probability of H given E

is $P[H \mid E]$. Bayes theorem gives a way of updating the prior probability to get the posterior probability. To do this one needs the conditional probabilities $P[E \mid H]$ and $P[E \mid H^c]$. The theorem is

$$P[H \mid E] = \frac{P[E \mid H]P[H]}{P[E \mid H]P[H] + P[E \mid H^c]P[H^c]}. \tag{1.5}$$

Example: Let H be the event that a randomly chosen person has a certain rare disease. For instance, it might be the $P[H] = 0.01$. Let E be the event that a test gives a positive result. Say that $P[E \mid H] = 0.9$, while $P[E \mid H^c] = 0.2]$. So the test seems reasonably reliable, while not perfect. Say that you get a positive result. Should you panic? Compute $P[H \mid E]$ to get $(0.9)(0.01)/((0.9)(0.01) + (0.2)(0.99)) = 0.043$. Probably not. After all, it's a rare disease.

1.5 Independence

Let $\binom{n}{k}$ be the number of subsets of an n element set that have exactly k elements. This is called the *binomial coefficient*. One can calculate $\binom{n}{k}$ by Pascal's triangle. This is because $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$ and

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}. \tag{1.6}$$

The proof of this *Pascal triangle* identity is left as an exercise. To get some intuition as to why it is so, look at the special case $\binom{5}{3} = \binom{4}{2} + \binom{4}{3}$, that is, $10 = 6+4$. Start with a five element set $\{a, b, c, d, e\}$. Divide the three element subsets into two collections. The first consists of the sets $\{abc\}, \{abd\}, \{abe\}, \{acd\}, \{ace\}, \{ade\}$. The second consists of the sets $\{bcd\}, \{bce\}, \{bde\}, \{cde\}$. The first consists of subsets that have a as a member; the second group consists of subsets that do not have a as a member. Take a away from every set in the first collections. Then you get the 6 two element subsets of $\{b, c, d, e\}$. The other collection consists of the 4 three element subsets of $\{b, c, d, e\}$. Of course such an example is not a general proof, but it can be an inspiration for one.

Here are some other important identities. The first is

$$\binom{n}{k} = \binom{n}{n-k}. \tag{1.7}$$

This is proved by observing that for every subset with k elements there is a complementary set with $n - k$ elements.

Another important identity is the identity for one marked point:

$$k\binom{n}{k} = n\binom{n-1}{k-1}. \tag{1.8}$$

This is because for every k element subset together with a marked point in it there is a corresponding marked point in the original set together with a $k - 1$ element subset of the set with this point removed.

Here is the proof in more detail. Consider a set S with n elements. Look at ordered pairs consisting of a subset A of S and a point p in S. The number of such ordered pairs is $\binom{n}{k}k$. This is because there are $\binom{n}{k}$ subsets A, and for each subset there is a choice of which of the k elements to pick to be p. Now instead look at ordered pairs consisting of a point p in S and a subset B with $k-1$ elements that does not have p in it. The number of such pairs is $n\binom{n-1}{k-1}$. This is because there are n points in S, and for each point there is an choice of which subset B to take of the set S with p removed.

To finish the proof, note that there is a perfect matching between the two collections. If A, p is a pair of the first kind, then let B be the set with the same elements as A, except with p removed. Then p, B is a pair of the second kind. One can also go backward. If p, B is a set of the second kind, then let A, p be the pair of the first kind, where A is obtained from B by introducing the additional point p. Since these two collections match up perfectly, it must be that they have the same number of elements. This proves the assertion.

Another identity of this type is the identity for two marked points:

$$k(k-1)\binom{n}{k} = n(n-1)\binom{n-2}{k-2}. \tag{1.9}$$

This is also left as an exercise.

Two events A, B are *independent* if $P[A \cap B] = P[A]P[B]$. So for independent events, the probabilities multiply.

Three events A, B, C are *independent* if $P[A \cap B] = P[A]P[B]$ and $P[B \cap C] = P[B]P[C]$ and $P[C \cap A] = P[C]P[A]$ and $P[A \cap B \cap C] = P[A]P[B]P[C]$. There is a similar definition for four or more events.

A *binomial* experiment consists of n independent success-failure trials, each of which has probability of success p on each trial. There are 2^n possible outcomes. The event that a particular outcome occurs has probability $p^k(1-p)^{n-k}$. This is because the trials are independent, and so one multiplies the probabilities for the k successes and the $n-k$ failures.

In the binomial experiment, the event that there are precisely k successes is

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{1.10}$$

This is because the set $\{1, 2, 3, \ldots, n\}$ of trials has $\binom{n}{k}$ subsets with exactly k elements. Each such subset corresponds to an outcome where there are k successes in the trials corresponding to the subset and $n-k$ failures in the trials corresponding to the complement of the subset. So one adds the probability $p^k(1-p)^{n-k}$ corresponding to each such pattern $\binom{n}{k}$ times to get $p(k)$.

Consider the example of a binomial experiment with three trials. The sample space of all outcomes consists of the 8 sequences in

$$\Omega = \{(FFF), (SFF), (FSF), (FFS), (FSS), (SFS), (SSF), (SSS)\}. \tag{1.11}$$

The collection of functions from a domain to a target with just two elements is in natural correspondence with the collection of subsets of the domain. In this

case the domain is the set $\{1, 2, 3\}$, and the target is the set $\{S, F\}$. So the corresponding subsets in this case are $\emptyset, \{1\}, \{2\}, \{3\}, \{2, 3\}, \{1, 3\}, \{1, 2\}, \{1, 2, 3\}$. This explains the 1 3 3 1 Pascal triangle pattern.

The collection of all 256 subsets of the sample space is the collection of all events. For example, the events of success on the first trial, on the second trial, and on the third trial are $\{(SSS), (SSF), (SFS), (SFF)\}$ and $\{(SSS), (SSF), (FSS), (FSF)\}$ and $\{(SSS), (SFS), (FSS), (FFS)\}$. Each of these three events has probability p. The intersection of these three events is the event $\{(SSS)\}$. By independence this has probability p^3. It is one of eight events that consists of a singleton set (a set with precisely one outcome in in it).

Sometimes we get careless and refer to the probability of the outcome (SSS), but officially we should be talking about the probability of the event $\{(SSS)\}$, the singleton set determined by the outcome. Remember that for a discrete probability measure the probability of each event is obtained by summing over the probability of the singleton sets that contribute to it. This is not true for continuous probability measures.

The event of exactly two successes in the three trials is the event $\{(SSF), (SFS), (FSS)\}$. By additivity it has probability $p^2(1-p) + p^2(1-p) + p^2(1-p) = 3p^2(1-p)$. This is a typical binomial probability computation. Notice that the 3 terms come from counting the three subsets $\{2, 3\}, \{1, 3\}, \{1, 2\}$.

1.6 Counting

In this section we count functions, injective functions, and subsets.

Let D be a set with d elements, and let S be a set with s elements. The set D is the *domain*; the set S is the *target* (or co-domain). A *function* assigns to each element of D an element of S. Notice that the *range* of a function is a subset of the target, and it need not be the entire target. The number of functions from D to s is s^d. Sometimes one wants to think of S as a population and D is an index set such as $D = \{1, 2, 3, \ldots, d\}$. In this case such a function is the same as an *ordered sample with replacement* of size d from the population S. It is also the same as a *sequence* $(s_1, s_2, s_3, \ldots, s_d)$ of d elements from the set S. Alternatively, one can focus on D and regard the elements of S as categories to describe the elements of D. In this case a function is a *classification* D into s categories.

Let D be a set with d elements, and let S be a set with s elements. An *injective function* is a function that sends elements that are not equal into elements that are not equal. The number of injective functions (one-to-one functions) from D to S is $(s)_d$. Here

$$(s)_d = s(s-1)\cdots(s-d+1) = \frac{s!}{(s-d)!}. \tag{1.12}$$

Note that $0! = 1$. Sometimes one wants to think of S as a population and D is an index set such as $D = \{1, 2, 3, \ldots, d\}$. Then an injective function is the same as *ordered sample without replacement* of size d from the population S. It

is also the same as a *sequence without repetition* $(s_1, s_2, s_3, \ldots, s_d)$ of d elements from the set S. (An injective function is also called a d-permutation of S; when $d = s$ it is called a permutation.) Alternatively, one can focus on D and regard the elements of S as physical tags that can be attached to the elements of D. Since a tag can be attached to at most one element of D, an injective function is the same as a *tagging* of D from a supply of s tags.

Let S be a set with s elements. A *subset* is a set A such that every element of A is an element of S. The number of subsets of S with d elements is

$$\binom{s}{d} = \frac{(s)_d}{d!}. \tag{1.13}$$

This is because for each subset with d elements there are $d!$ injective functions with this subset as its range. A subset is the same as an *unordered sample without replacement* of size d from the population S. (Sometimes it is called a d-combination from S.)

The *ordered sampling with replacement experiment* has outcomes consisting of all functions from the index set D to the population S. The probability measure is the uniform probability measure.

Suppose that the population has a successes. Then the probability that the number of successes in the sample is k is

$$b(k) = \binom{d}{k} \frac{a^k (s-a)^{d-k}}{s^d} = \binom{d}{k} \left(\frac{a}{s}\right)^k \left(1 - \frac{a}{s}\right)^{d-k}. \tag{1.14}$$

These are binomial probabilities.

The reasoning is the following. There are s^d ordered samples with replacement (functions from $\{1, 2, \ldots, d\}$ to the population). The uniform probability thus assigns probability $1/s^d$ to the singleton event corresponding to each sample. So all we have to do is to count the functions that have exactly k successes in their range. Look at the k element subset of the domain $\{1, 2, \ldots, d\}$ corresponding to the k successes. There are $\binom{d}{k}$ way of choosing this subset. Now look at the functions from this subset to the set of successes in the population. There are a^k such functions. Then look at the functions from the complement of this subset to the set of failures in the population. There are $(s-a)^{d-k}$ of these. So these three choices determine an injective function with k successes. Therefore the numerator is determined by multiplying these three numbers.

The *ordered sampling without replacement experiment* has outcomes consisting of all injective functions from the index set D to the population S. The probability measure is the uniform probability measure.

Suppose that the population has a successes. Then the probability that the number of successes in the sample is k is

$$h(k) = \binom{d}{k} \frac{(a)_k (s-a)_{d-k}}{(s)_d} = \frac{\binom{a}{k}\binom{s-a}{d-k}}{\binom{s}{d}}. \tag{1.15}$$

These are *hypergeometric* probabilities.

The reasoning is the following. There are $(s)_d$ ordered samples without replacement (injective functions from $\{1, 2, \ldots, d\}$ to the population). The uniform probability thus assigns probability $1/(s)_d$ to the singleton event corresponding to each sample. So all we have to do is to count the injective functions that have exactly k successes in their range. Look at the k element subset of the domain $\{1, 2, \ldots, d\}$ corresponding to the k successes. There are $\binom{d}{k}$ way of choosing this subset. Now look at the injective functions from this subset to the set of successes in the population. There are $(a)_k$ such injective functions. Then look at the injective functions from the complement of this subset to the set of failures in the population. There are $(s-a)_{d-k}$ of these. So these three choices determine an arbitrary injective function with k successes. Therefore the numerator is determined by multiplying these three numbers.

The expression involving only binomial coefficients may be obtained by doing some algebra. However there is also a conceptual proof, which is left as an exercise. The idea is to repeat the above reasoning, but with subsets instead of injective functions.

1.7 Sets versus functions

The fundamental mathematics structure of probability is a set of outcomes, the sample space. Subsets of this space are events. Each event has a probability.

However each outcome may itself have a mathematical structure. In many common situations, each outcome is a function with some domain and some target. For instance, in the dice experiment for throwing d dice (or for throwing a die d times) the domain is the set of dice (or the domain is $\{1, 2, \ldots, d\}$), the target is $\{1, 2, 3, 4, 5, 6\}$, and the sample space consists of all functions from the domain to the target. In the ordered sampling with replacement experiment the domain is $\{1, 2, \ldots, d\}$, the target is the population S, and and the sample space consists of all functions from the domain to the target. In the ordered sampling without replacement experiment the domain is $\{1, 2, \ldots, d\}$, the target is the population S, and and the sample space consists of all injective functions from the domain to the target.

It is important to distinguish functions from subsets. However there are special circumstances where they are closely related. There is a $d!$ to one correspondence between injective functions from D to S and subsets of the target S. If S has precisely two points, then there is a one to one correspondence between functions from D to S and subsets of the domain D. For instance, in a success failure experiment, the sample space can be thought of either as functions from $\{1, 2, \ldots, d\}$ to $\{S, F\}$ or as subsets of $\{1, 2, \ldots, d\}$. In this case the function picture is usually more convenient.

1.8 Frequency interpretation

The *frequency interpretation* of probability says that a probability is a mathematically computed number that is used to predict an experimental sample proportion in a large sample.

More explicitly, consider an experiment and an event A. Let p be its probability. Now consider a super-experiment, consisting of n independent repetitions of the original experiment. The *sample size* n is assumed to be large. (This notion needs to be made quantitative, but this will have to wait until the next chapter.) Perform the super-experiment and get an outcome of the super-experiment (that is, n outcomes of the original experiment). Let S_n be the experimental number of repetitions in the super-experiment for which A happens. Let

$$M_n = \frac{S_n}{n} \tag{1.16}$$

be the *sample proportion*. Then the experimental number M_n is supposed to be near the theoretical number p. At least one hopes.

In statistics a special notation is used. The probability p of success is sometimes called the population proportion. The relative frequency of successes is called the sample proportion and is written with a hat notation like $\hat{p}_n$. The sample proportion $\hat{p}_n$ estimates the population proportion p.

Example. Consider the experiment of throwing a die. Let A be the event of getting a 5 or 6. Then the probability of A is $p = 1/3$. Yesterday I performed this experiment 900 times. The experimental sample proportion was $\hat{p}_n = 312/900$.

1.9 Problems

1. Express De Morgan's laws in logical language.

2. Consider an experiment with sample space $\Omega = \{a, b, c, d, e\}$. How many events are there? Write them explicitly. How many events are there with exactly 3 elements. Indicate them explicitly.

3. Consider an experiment consisting of tossing a pair of dice. This experiment has 36 outcomes. List them. How many events are there? Express the result in scientific notation.

4. Say that Ω is the sample space for an experiment. Consider a new super-experiment with sample space Ω^n. Each outcome in Ω^n is a sequence of n outcomes in Ω. The intuition is that the super-experiment is the original experiment repeated n times. Suppose that Ω is finite with r outcomes. How how many outcomes are there in Ω^n? How many events are associated with the super-experiment?

5. As a concrete example of a super-experiment that is often performed, consider throwing a pair of dice 100 times. How many outcomes are there?

Express in scientific notation. How many events? Express in scientific notation.

6. In the dice tossing example, what is the probability that the sum of the numbers on the two dice is odd?

7. In the coin tossing example, what is the probability of needing j or more tosses to get the first head?

8. In the phone call waiting time example, what is the probability that a call will come in the first half hour?

9. Recall that in the super-experiment each outcome ω is a sequence $(\omega_1, \omega_2, \ldots, \omega_n)$ of outcomes of the individual experiment. Let $A_1, \ldots, A_n$ be events associated with the individual experiment. A probability measure is a *product probability measure* if $P[\{\omega \mid \omega_1 \in A_1, \omega_2 \in A_2, \ldots, \omega_n \in A_n\}] = P[A_1]P[A_2] \cdots P[A_n]$. That is, the probability in the super-experiment that for each j the jth repetition of the original experiment has outcome in A_j is the product from $j = 1$ to n of the probabilities of the events A_j. What does this formula say if each event $A_j = A$ is the same event?

10. Consider the experiment of throwing a pair of dice. What is the probability of the event A of getting box-cars or snake-eyes? Consider the super-experiment of throwing a pair of dice 100 times. Use the product probability measure. What is the probability that A happens each time. Express in scientific notation.

11. A biased coin has a probability of p of heads. The coin is flipped twice. The probability of heads followed by heads is p^2, heads followed by tails is $p(1-p)$, tails followed by heads is $p(1-p)$, and tails followed by tails $(1-p)^2$. What is the conditional probability of two heads, given that there is at least one head?

12. Prove the law of total probability

$$\begin{aligned} P[A] &= P[A \mid B \cap C]P[B \mid C]P[C] + P[A \mid B^c \cap C]P[B^c \mid C]P[C] \\ &\quad + P[A \mid B \cap C^c]P[B \mid C^c]P[C^c] + P[A \mid B^c \cap C^c]P[B^c \mid C^c]P[C^c]. \end{aligned} \tag{1.17}$$

.

13. Use this law to analyze ordered samples without replacement of size 3. Take as in the example a population of size n with m successes and $n-m$ failures. In particular, use it to compute the probability that the third choice results in a success.

14. The prior probability of life on planet X is 1/100. Given that there is life, the probability of a periodic radio signal is 1/2. Given that there is no life, the probability of a periodic radio signal is 1/1000. A periodic radio signal is observed. What is the revised probability of life on planet X?

15. Say that there are n independent pieces of evidence, and the event that all are found is E_n. Suppose $P[E_n \mid H] = \alpha^n$ and $P[E_n \mid H^c] = \beta^n$, with $0 < \beta < \alpha < 1$. Find the limit as $n \to \infty$ of $P[E_n \mid H]$. Find the limit as $n \to \infty$ of $P[H \mid E_n]$.

16. Prove the Pascal's triangle identity. Hint: Consider a set with n elements. Call one of the elements the special element. Count separately the subsets that do not have the special element in it and the subsets that do have the special element in it.

17. Prove the identity for two marked points:

$$k(k-1)\binom{n}{k} = n(n-1)\binom{n-2}{k-2}. \tag{1.18}$$

Do this with the same method as for one marked point.

18. Prove that if A, B are independent, then A, B^c are independent.

19. Prove that if A, B, C are independent, then $P[A \cup B \cup C] = 1 - (1 - P[A])(1 - P[B])(1 - P[C])$.

20. A business student makes three independent investments in the hope of getting rich. The first has a 5 percent chance making him rich. The second has a 10 percent chance of making him rich. The third has a 20 percent chance of making him rich. What is the chance that the student will get rich from at least one of these investments?

21. What is the probability that the student will get rich from precisely one of the investments (and not the other two)?

22. A rocket has 12 independent subsystems, each of which works with probability 0.98. What is the probability that they all work? Express numerically.

23. Consider the experiment of tossing a pair of dice. Let A be the event that the first number is odd. Let B be the event that the second number is odd. Let C be the event that the sum of the two numbers is odd. Show that A, B are independent, A, C are independent, and B, C are independent. Are A, B, C independent?

24. Consider a binomial experiment with $n = 5$ independent success-failure trials. The probability of success on each trial is $1/3$. List all the outcomes of the experiment. How many are there. Find the probability of each of the singleton sets corresponding to an outcome.

25. In the same binomial experiment ($p = 1/3$ and $n = 5$), write explicitly the event that there are precisely three successes. That is, exhibit this as a set of outcomes. Calculate its probability.

26. Fix $\mu \geq 0$. Consider a binomial experiment with probability $p = \mu/n$ of success on each of n trials. The probability of a failure on each trial is $(1 - \mu/n)^n$. Find the limit of this probability as $n \to \infty$.

27. List all functions from $\{1, 2\}$ to $\{a, b, c\}$. How many are there? List all injective functions from $\{1, 2\}$ to $\{a, b, c\}$. How many are there? List all subsets of $\{a, b, c\}$ with exactly two elements. How many are there?

28. List all functions from $\{1, 2, 3\}$ to $\{a, b\}$. How many are there? List all injective functions from $\{1, 2, 3\}$ to $\{a, b\}$. How many are there? List all subsets of $\{a, b\}$ with exactly three elements. How many are there?

29. Consider ordered samples of size 5 from a population of size 10. There are fewer samples without replacement than with replacement. Less than a tenth as many? Compute the number of samples to answer this question. Answer the analogous question for ordered samples of size 10 from a population of size 20.

30. Prove in detail that for each k the limit

$$\lim_{n \to \infty} \frac{(n)_k}{n^k} = 1. \tag{1.19}$$

31. Show that the hypergeometric probabilities (defined above for ordered samples without replacement) may be also be obtained from unordered samples without replacement. Thus consider a population of size s. There are a successes and $s - a$ failures in the population. Consider the uniform probability measure on the set of all subsets of the population of fixed sample size d. The problem is to compute the probability of exactly k successes in the sample. That is, compute the numerator (the number of subsets of the population with precisely k successes and $d - k$ failures) and the denominator (the number of subsets of the population with d elements), and then divide.

Chapter 2

Discrete random variables

A *random variable* is a real function defined on the sample space. That is, it is a rule that assigns to every outcome a real number.

How does one find the value of a random variable? Answer: Do the experiment and get an outcome. Calculate the appropriate number from the outcome.

The event that a random variable X has the value x is the set of all outcomes ω such that the value $X(\omega) = x$. Most often we just abbreviate this event as $X = x$. Thus this event has a probability $P[X = x]$. If

$$\sum_x P[X = x] = 1, \tag{2.1}$$

then the random variable is said to be a *discrete* random variable. In this equation the sum is over all possible values of the random variable. Henceforth in this chapter we consider only discrete random variables.

If X is a discrete random variable, then its *probability mass function* is the function p_X defined by $p_X(x) = P[X = x]$. It is a real function defined on the set of possible values of the random variable.

For each random variable X there is an associated probability measure P_X defined as follows. The sample space is the set of possible values of the random variable. An event I is a set of such numbers. The probability is defined by

$$P_X[I] = P[X \in I] = \sum_{x \in I} p_X(x). \tag{2.2}$$

This probability measure is called referred to as the *distribution* of X. Another term is *probability law* of X.

Notice the general scheme. There is an original probability measure P that defines a probability for each subset of the original set of outcomes. If X is a random variable, then from P we can compute a new probability measure P_X that defines a probability for each subset of the values of the random variable X. Going from P to the law P_X usually involves a considerable loss of information. If Y is another random variable defined on the same set of outcomes, then it has its own probability law P_Y.

Example: Consider the set of outcomes for n success-failure trials with probability $1/2$ of success on each trial. The set of outcomes has 2^n points. The probability measure assigns to each subset of outcomes a certain probability. This is the uniform probability measure on the set of all outcomes. Now consider a random variable such as S_n, the total number of successes. The possible values of this random variable are $0, 1, 2, \ldots, n$. The probability measure P_{S_n} is defined by

$$P_{S_n}[I] = P[S_n \in I] = \sum_{k \in I} \binom{n}{k} \frac{1}{2^n}. \tag{2.3}$$

This is not a uniform probability measure. It is defined on a sample space of only $n + 1$ points. It is quite a different object.

Example: Consider the random variable $W = (S_n - n/2)^2$. For simplicity take n even, so that $n/2$ is a natural number. The random variable has possible values $j = 0, 1, 4, 9, 16, \ldots, (n/2)^2$. The probability that W is in some set J of perfect squares is equal to

$$P[W \in J] = \sum_{j \in J} p_W(j), \tag{2.4}$$

where

$$p_W(0) = P[S_n = n/2] \tag{2.5}$$

and for $j \geq 1$

$$p_W(j) = P[S_n = n/2 + \sqrt{j}] + P[S_n = n/2 - \sqrt{j}]. \tag{2.6}$$

This is because $W = (S_n - n/2)^2 = j$ precisely when $S_n = n/2 \pm \sqrt{j}$.

2.1 Examples of random variables

A random variable S_n is *binomial* (for n independent trials with probability p of success on each trial) if its probability mass function is

$$p_{S_n}(k) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.7}$$

A binomial random variable counts the number of successes in n independent trials, where the probability of success on each trial is p. The possible values are $0, 1, 2, \ldots, n$. In other words, it is the number of heads when a biased coin is flipped n times.

A random variable T is *geometric* (for probability p of success on each trial) if its probability mass function is

$$p_T(k) = (1-p)^{k-1} p. \tag{2.8}$$

A geometric random variable counts the number of independent trials (with probability p of success on each trial) until the first success. The possible values

are $1, 2, 3, \ldots$. In other words, it is the number of flips of a biased coin until the first head is achieved.

Warning: Some authorities define the geometric random variable to be one less, so that the possible values are $0, 1, 2, \ldots$. This amounts to counting the number of failures before the first success.

A random variable N is *Poisson* with mean μ if its probability mass function is

$$p_N(k) = \frac{\mu^k}{k!} e^{-\mu}. \tag{2.9}$$

It typically counts the number of events during some interval of time or in some region of space. For instance, it might count the number of strong earthquakes per year in Iceland, or it might count the number of big craters per square kilometer on the surface of the moon.

Consider the binomial probability mass function. Write $\mu = np$. The function may be written in the form

$$p_{S_n}(k) = \frac{(n)_k}{n^k} \frac{\mu^k}{k!} (1 - \frac{\mu}{n})^n (1 - \frac{\mu}{n})^{-k}. \tag{2.10}$$

There are four factors. The first is a combinatorial ratio. The second is the same $\mu^k/k!$ as in the Poisson probability. The third is the probability of n failures in n trials, when the probability of success on one trial is μ/n. The last one is a fixed power of a number near one.

Take the limit as the number of trials $n \to \infty$ and the probability of success on each trial $p \to 0$ with the mean number $\mu = np$ fixed. It is not difficult to prove that the limit is the Poisson probability. In fact the first and fourth terms approach one, while the two middle terms give the Poisson probabilities. Thus the Poisson probability has the interpretation of the probabilities associated with a huge number of trials with a tiny probability of success on each trial. The Poisson random variable is the number of successes.

2.2 The empirical distribution

Let X be a discrete random variable. In this section its probability mass function or distribution is denoted $p(x)$. Thus $p(x)$ is the mathematically computed probability that the outcome of the experiment is such that the random variable X takes the value x. Notice that the sum of $p(x)$ over all x adds to one.

Consider a super-experiment consisting of n independent repetitions of the original experiment. Let $M_n(x)$ be the sample proportion that is the number of times that X takes the value x, divided by the sample size n. This could also be called $\hat{p}_n(x)$. Notice that the sum of $M_n(x)$ over all x adds to one. This set of experimental numbers is known as the *empirical distribution*. If n is large, it should be rather close to the mathematically computed distribution.

Example: For a Poisson random variable N with mean $\mu = 1$ the probabilities of values 0,1,2,3,4,5 are $0.368, 0.368, 0.184, 0.061, 0.015, 0.003$. This morning I performed did the Poisson experiment 50 times. The empirical distribution

was 21/50, 16/50, 10/50, 1/50, 2/50, 0/50, or $0.42, 0.32, 0.20, 0.02, 0.04, 0.00$. The sum is one, but that is because I was lucky enough not to get any values 6 or higher.

2.3 Functions of a random variables

Say that X is a discrete random variable, and $Y = g(X)$. This means that after the experimental number X is found, a further calculation is done to produce Y. The relation between the two probability mass functions is that

$$p_Y(y) = \sum_{\{x|g(x)=y\}} p_X(x). \tag{2.11}$$

2.4 Expectation

Say that X is a discrete random variable. Its *expectation* is

$$E[X] = \sum_x x p_X(x). \tag{2.12}$$

Sometimes the expectation of a random variable is called the *mean* of the random variable. This is often written as

$$\mu_X = E[X]. \tag{2.13}$$

Here are two important properties of expectation involving constant random variables. One is

$$E[c] = c. \tag{2.14}$$

The other is

$$E[cX] = cE[X]. \tag{2.15}$$

There is a theorem that gives a simple way of computing the expectation of a random variable $g(X)$ that is a function of a random variable X. The formula is

$$E[g(X)] = \sum_x g(x) p_X(x). \tag{2.16}$$

The *variance* of a random variable is

$$\mathrm{var}[X] = E[(X - E[X])^2]. \tag{2.17}$$

This formula can be made totally obscure by writing it in the alternate form

$$\mathrm{var}[X] = E[X^2] - E[X]^2. \tag{2.18}$$

This is common. Sometimes the variance of a random variable is written as

$$\sigma_X^2 = \mathrm{var}[X]. \tag{2.19}$$

Another useful quantity is the *standard deviation*

$$\sigma_X = \sqrt{\mathrm{var}[X]}. \tag{2.20}$$

For a binomial random variable the mean is

$$E[S_n] = np \tag{2.21}$$

and the variance is

$$\mathrm{var}[S_n] = E[(S_n - np)^2] = np(1-p). \tag{2.22}$$

Notice that the standard deviation only grows like a multiple of $\sqrt{n}$. For n large this is considerably smaller than n.

The calculation of the expectation of S_n is not difficult. We calculate

$$E[S_n] = \sum_{k=1}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=1}^{n} n\binom{n-1}{k-1}p^k(1-p)^{n-k} = np\sum_{j=0}^{n-1}\binom{n-1}{j}p^j(1-p)^{n-1-j} = np. \tag{2.23}$$

This uses the binomial coefficient identity for one marked point followed by the substitution $j = k - 1$. The calculation of the variance is a problem that goes along the same lines.

For a geometric random variable the mean is

$$E[T] = \frac{1}{p} \tag{2.24}$$

and the variance is

$$\mathrm{var}[T] = E[(T - \frac{1}{p})^2] = \frac{1}{p}\left(\frac{1}{p} - 1\right). \tag{2.25}$$

Here the standard deviation is about the same size as the mean, at least when the mean is fairly large.

For a Poisson random variable the mean is

$$E[N] = \mu \tag{2.26}$$

and the variance is

$$\mathrm{var}[N] = E[(N-\mu)^2] = \mu. \tag{2.27}$$

Notice that the standard is the square root of the mean. Thus when the mean is large, the standard deviation is much smaller.

2.5 The weak law of large numbers for sample proportions

Let S_n be the number of successes in n independent trials, with probability p of success on each trial. Define the *sample proportion*

$$M_n = \frac{M_n}{n}. \tag{2.28}$$

The *weak law of large numbers* for sample proportions says that

$$E[M_n] = p \tag{2.29}$$

and

$$\text{var}[M_n] = E[(M_n - p)^2] = \frac{p(1-p)}{n}. \tag{2.30}$$

It is more meaningful to think of the weak law of large numbers in terms of the mean and standard deviation. For variety, use the notation $\hat{p}_n$ for the sample proportion. The weak law says

$$\mu_{\hat{p}_n} = p \tag{2.31}$$

and

$$\sigma_{\hat{p}_n} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}. \tag{2.32}$$

In statistics the estimate

$$\sigma_{\hat{p}_n} \leq \frac{1}{2}\frac{1}{\sqrt{n}} \tag{2.33}$$

is very illuminating. It says that one can bound the standard deviation of the sample proportion even if one does not know the value of p.

The weak law of large numbers gives a kind of internal consistency to the frequency interpretation of probability. It says that the experimental sample proportion $\hat{p}_n$ in the super-experiment is predicted by the super-experiment to be close to the probability p in the original experiment.

2.6 Problems

1. A basketball player has a probability 0.6 of success in a certain play. The play is repeated 10 times (independent repetitions). What is the probability of 6 or more successes?

2. The expected number of major earthquakes on a certain island in Iceland is 3.4 per year. An earthquake researcher needs several observations to complete a study. What is the probability of 2 or less earthquakes in a given year?

3. A gambler plays a game where the probability of a win on each trial is 1/12. What is the probability that the gambler wins for the first time on the 13th try?

4. A stock market speculator makes a very large number of independent investments. The probability of a success on each given investment is very small. However economic theory predicts that the expected total number of successes is 2. What is the probability of 3 successes?

5. The average number of childhood leukemia cases per year in a medium sized community is 2. What is the probability that in a given year and in twenty such communities there is at least one community with 5 or more cases?

6. A random variable S is binomial with $p = 1/2$ and $n = 4$. What is the probability mass function of $(S-2)^2$?

7. In the preceding problem, find the expectation of $(S-2)^2$ from its probability mass function. Then find the expectation of $(S-2)^2$ from the binomial probability mass function. (If they are not the same, take a study break.)

8. If X is a random variable with values that are natural numbers, then $f_X = E[X(X-1)]$ is its *second factorial moment*. Find a formula for σ_X^2 in terms of f_X and μ_X.

9. Calculate the second factorial moment of the binomial random variable S_n. Hint: Follow the pattern of the calculation of the expectation. Use the binomial coefficient identity involving two marked points.

10. Use the preceding result to calculate the variance of the binomial random variable S_n.

11. We know the geometric series sum

$$\sum_{n=0}^{\infty} q^n = \frac{1}{1-q}. \tag{2.34}$$

Prove that

$$\sum_{n=1}^{\infty} n q^{n-1} = \frac{1}{(1-q)^2} \tag{2.35}$$

and

$$\sum_{n=2}^{\infty} n(n-1) q^{n-2} = \frac{2}{(1-q)^3}. \tag{2.36}$$

12. Use the preceding problem to calculate the mean of the geometric random variable T.

13. Say the number of gas particles in a room is a Poisson random variable with mean 10^{30}. What is the ratio of the standard deviation to the expectation?

14. Consider a Poisson random variable. What is the ratio of the standard deviation to the mean in the limit when the expectation goes to infinity?

15. Consider a geometric random variable. Find the limit as $p \to 0$ of the ration of the standard deviation to the expectation. What does this say about the numbers that you will encounter in a typical measurements of a geometric random variable with large expectation?

16. A public opinion survey is an ordered sample without replacement from a population. In the typical situation when the population is very large compared to the sample, it makes very little difference if the calculation is done instead by considering an ordered sample with replacement. Thus one may deal with this as independent success-failure trials. Say that the yes portion of the population is 6/10. Consider a sample of size 1600. What is the standard deviation of the sample proportion?

17. As explained in the previous problem, a public opinion survey may be thought of as an ordered sample with replacement from a population. Say that the yes portion of the population is completely unknown to the statistician. Consider a sample of size 1600. What can the statistician give as an upper bound for the standard deviation of the sample proportion?

18. As explained in a previous problem, a public opinion survey may be thought of as an ordered sample with replacement from a population. Say that the yes portion of the population is completely unknown to the statistician. The statistician wants to choose a sample size so that the standard deviation of the sample proportion is guaranteed to be at most one percent. What sample size will work?

19. A statistician wants to estimate the unknown proportion p of people in Arizona who can identify the President of France. So the statistician takes a sample of size n and uses the experimental sample proportion $\hat{p}$ as the estimate. In addition, the statistician wants to know how good the sampling procedure is, as measured by $\sqrt{p(1-p)}/\sqrt{n}$. (The smaller the better.) At first it seems reasonable to estimate this by the experimental quantity $\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}$. For instance, if $\hat{p} = 0.17$ and $n = 625$, then the estimate is 0.015, which is considerably smaller than the upper bound 0.02. On the other hand, it seems circular to use the estimate of the unknown quantity to figure how good the procedure that gave the estimate is. Is this a scam? Or is it a reasonable idea? Explain. If you wish, do a few computer experiments to get a feel for the situation.

Chapter 3

The Bernoulli process

3.1 Functions of several random variables

Let X, Y be two discrete random variables. Their *joint probability mass function* is defined by

$$p_{X,Y}(x, y) = P[X = x, Y = y]. \tag{3.1}$$

That is, the value of the joint probability mass function at x, y is the probability that $X = x$ and $Y = y$.

The *marginal* distributions of the joint probability mass function are

$$p_X(x) = \sum_y p_{X,Y}(x, y) \tag{3.2}$$

and

$$p_Y(y) = \sum_x p_{X,Y}(x, y) \tag{3.3}$$

Example. Consider the following random variables. Let X, Y be independent binomial random variables, each with $n = 2$ and $p = 1/2$. Let $Z = 2 - X$. Let $W = g(X + Y)$, where $g(1) = 0$, $g(0) = g(2) = g(4) = 1$, $g(3) = 2$. Then each of X, Y, Z, W has the same binomial probability mass function for the possible values 0,1,2. The variables X, Y are independent, and so their joint probability mass function is obtained by multiplication:

$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$
$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$
$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{16}$

. (3.4)

The variables X, Z are highly dependent, and so their joint probability mass function is

0	0	$\frac{1}{4}$
0	$\frac{1}{2}$	0
$\frac{1}{4}$	0	0

. (3.5)

The random variables X, W are dependent, but not as strongly as in the previous case. Their probability mass function is computed as follows. First, $W = 0$ if and only if $X + Y = 1$. So the probabilities for $X = 0, W = 0$ are the same as for $X = 0, Y = 1$, and the probabilities for $X = 1, W = 0$ are the same as for $X = 1, Y = 0$. Similarly, $W = 2$ if and only if $X+Y = 3$. So $X = 1, W = 2$ is the same as $X = 1, Y = 2$, and $X = 2, W = 2$ is the same as $X = 2, Y = 1$. Finally, $W = 1$ when $X + Y = 0, 2, 4$, so $X = 0, W = 1$ is the same as $X = 0, Y = 0$ or $X = 0, Y = 2$. Also $X = 1, W = 1$ is the same as $X = 1, Y = 1$. And $X = 2, W = 1$ is the same as $X = 2, Y = 0$ or $X = 2, Y = 2$. Thus the joint probability mass function is

$$\begin{array}{|c|c|c|} \hline 0 & \frac{1}{8} & \frac{1}{8} \\ \hline \frac{1}{8} & \frac{1}{4} & \frac{1}{8} \\ \hline \frac{1}{8} & \frac{1}{8} & 0 \\ \hline \end{array}. \tag{3.6}$$

Notice that in all three examples the column sums and the row sums give the correct marginal distributions.

Say that a random variable $g(X, Y)$ is a function of random variables X, Y. Then its expectation may be computed from

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) p_{X,Y}(x, y). \tag{3.7}$$

A very important special case of this is the relation

$$E[X + Y] = E[X] + E[Y]. \tag{3.8}$$

This comes from taking $g(x, y) = x + y$. The derivation is as follows. We have

$$\sum_x \sum_y (x+y) p_{X,Y}(x, y) = \sum_x \sum_y x p_{X,Y}(x, y) + \sum_x \sum_y y p_{X,Y}(x, y) = \sum_x x p_X(x) + \sum_y y p_Y(y). \tag{3.9}$$

Example: Consider the previous example and the sum random variables $X + Y$, $X + Z$, and $X + W$. They each have expectation 2. On the other hand, their variances are 1, 0, and 3/2. It is worth working this out.

3.2 Independence

Two discrete random variables X, Y are *independent* provided that for each x, y we have $p_{X,Y}(x, y) = p_X(x) p_Y(y)$.

A crucially important result is that if X, Y are independent, then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]. \tag{3.10}$$

The proof of this is as follows. We have

$$\sum_x \sum_y g(x)h(y) p_{X,Y}(x, y) = \sum_x \sum_y g(x)h(y) p_X(x) p_Y(y) = \sum_x g(x) p_X(x) \sum_y g(y) p_Y(y). \tag{3.11}$$

A simple corollary of this is when $g(x) = x$ and $h(y) = y$. The if X, Y are independent, then it follows that

$$E[XY] = E[X]E[Y]. \tag{3.12}$$

An extremely important corollary is that if X, Y are independent, then

$$\text{var}[X + Y] = \text{var}[X] + \text{var}[Y]. \tag{3.13}$$

The proof of this is simple. We have

$$\text{var}[X+Y] = E[(X+Y-E[X+Y])^2] = E[(X-E[X])^2+2E[(X-E[X])(Y-E[Y])]+E[(Y-E[Y])^2]. \tag{3.14}$$

This holds even if the random variables are not independent. But if they are independent, then $X - E[X]$ and $Y - E[Y]$ are also independent. Therefore the cross term is

$$E[(X - E[X])(Y - E[Y])] = E[(X - E[X])]E[(Y - E[Y])] = 0 \cdot 0 = 0. \tag{3.15}$$

In terms of standard deviation, the result says that for independent random variables X, Y we have

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}. \tag{3.16}$$

3.3 The chi squared statistic for the empirical distribution

Consider a discrete random variable X with finitely many values $x_1, \ldots, x_k$. Let $p_j = p(x_j)$ be the probability mass function of X. Repeat the experiment of measuring this random variable n times. The empirical distribution is the random probability mass function $\hat{p}_j = \hat{p}(x_j)$, where $\hat{p}_j = N_j/n$ is the number N_j of times that the measurement of X gives the value x_j, divided by n.

Each random variable N_j is binomial with parameters n and p_j. Therefore it has mean np_j and variance $np_j(1 - p_j)$. The random variables are not independent, since they must satisfy $\sum_{j=1}^k N_j = n$. This shows that only $k - 1$ of the variables N_j can vary independently.

The situation is analogous for the variables $\hat{p}_j$. The random variable $\hat{p}_j$ has mean np_j and variance $(1/n)p_j(1 - p_j)$. The random variables are not independent, since they must satisfy $\sum_{j=1}^k \hat{p}_j = 1$. This shows that only $n - 1$ of the variables $\hat{p}_j$ can vary independently. In the jargon of statistics, one says that there are $k - 1$ degrees of freedom.

The χ^2 statistic for $k - 1$ degrees of freedom is defined in this situation by

$$\chi^2 = \sum_{j=1}^k \frac{(N_j - np_j)^2}{np_j} = \sum_{j=1}^k \frac{(\hat{p}_j - p_j)^2}{(p_j/n)}. \tag{3.17}$$

It is a measure of how close the empirical $\hat{p}_j$ values are to the probabilities p_j. In other words, it measures how well the experimental empirical distribution matches the mathematically computed probability distribution.

The expectation of the χ^2 statistic is

$$\sum_{j=1}^{k} \frac{np_j(1-p_j)}{np_j} = \sum_{j=1}^{k}(1-p_j) = k-1. \tag{3.18}$$

This shows that the typical value of the χ^2 statistic is roughly $k-1$. When n is large there is a simplification that makes possible more detailed probability calculations with the χ^2 distribution. (One gets the more usual χ^2 statistic with $k-1$ degrees of freedom, which is a sum of $k-1$ squares of independent normal random variables.) In this situation, a statistician would look up the properties of the χ^2 statistic with $k-1$ degrees of freedom in tables.

3.4 The weak law of large numbers for sample means

Let $X_1, \ldots, X_n$ be independent random variables, each with mean $\mu = E[X_j]$ and finite standard deviation $\sigma = \sqrt{\text{var}[X_j]}$. Let

$$S_n = X_1 + X_2 + \cdots + X_{n-1} + X_n \tag{3.19}$$

be their sum. Then the expectation of the sum is

$$\mu_{S_n} = n\mu \tag{3.20}$$

and the standard deviation is

$$\sigma_{S_n} = \sqrt{n}\sigma. \tag{3.21}$$

The *weak law of large numbers* for independent random variables mean μ and finite standard deviation σ states that the *sample mean*

$$M_n = \frac{S_n}{n} \tag{3.22}$$

has

$$\mu_{M_n} = \mu \tag{3.23}$$

and the standard deviation is

$$\sigma_{M_n} = \frac{\sigma}{\sqrt{n}}. \tag{3.24}$$

In statistics the expectation μ is often called the population mean. The sample mean for a sample of size n is denoted $\bar{X}_n$. Thus the sample mean (a random variable) estimates the population mean (a number).

In order to see how variable the sample mean is, it would be helpful to know the population variance σ^2. Typically the statistician does not know either μ or σ. So a common device is to use the *sample standard deviation*

$$\hat{\sigma}_n^2 = \frac{(X_1 - \bar{X}_n)^2 + (X_2 - \bar{X}_n)^2 + \cdots + (X_n - \bar{X}_n)^2}{n-1}. \tag{3.25}$$

The reason that many statisticians prefer the somewhat peculiar $n = 1$ in the denominator is that the numerator is actually computed from $n - 1$ numbers, the deviations from the sample mean $X_j - \bar{X}_n$. These numbers satisfy

$$(X_1 - \bar{X}_n) + (X_2 - \bar{X}_n) + \cdots + (X_n - \bar{X}_n) = 0 \tag{3.26}$$

no matter what the outcome of the experiment. So if you know any $n - 1$ of these deviations, then you can figure out the remaining one. This is simply because the sample mean is defined to make it so.

In any case, it is common to use the sample variance $\hat{\sigma}_n^2$ to estimate the population variance σ^2. It is also common to the sample standard deviation $\hat{\sigma}_n$ to estimate the population standard deviation σ. Then $\hat{\sigma}_n/\sqrt{n}$ estimates the standard deviation of the sample mean, which is $\sigma/\sqrt{n}$.

3.5 The Bernoulli process

Each day a lottery prize may or may not be awarded, and this decision is made by a random device. There is an average of one prize per week. So we can model this as independent success-failure trials with a probability $p = 1/7$ of success each day.

This sort of situation is described by the Bernoulli process. This is a sequence $X_1, X_2, X_3, X_4, X_5, \ldots$ of independent *Bernoulli* random variables. That is, these are independent random variables such that each X_j has the value 1 with probability $p > 0$ and the value 0 with probability $1 - p < 1$. In other words,

$$P[X_j = k] = p^k (1-p)^{1-k} \tag{3.27}$$

for $k = 0, 1$.

There are two associated sequences of random variables. Let $S_0 = 0, S_1 = X_2, S_2 = X_1 + X_2, X_3 = X_1 + X_2 + X_3$, and so on. That is,

$$S_n = X_1 + \cdots + X_n. \tag{3.28}$$

In the example, S_n is the number of prizes awarded in the first n days. In general we say that S_n is the number of success in the first n trials. It is a *binomial* random variable with mean np and variance $np(1-p)$. The distribution is given by

$$P[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k} \tag{3.29}$$

for $k = 0, 1, \ldots, n$.

Let Y_k be defined as the first n such that $S_n = k$. In the example, Y_k is the day when the kth prize is awarded. Set $Y_0 = 0$, and define $T_k = Y_k - Y_{k-1}$ for $k \geq 1$. Then Thus T_k is the waiting time from the $k-1$st success to the kth success. Each T_k is a *geometric* random variable with

$$P[T_k = n] = (1-p)^{n-1}p \tag{3.30}$$

for $n = 1, 2, 3, \ldots$.

Furthermore, the random variables $T_1, T_2, T_3, \ldots$ are independent. The time Y_k for the kth success is

$$Y_k = T_1 + \ldots + T_k. \tag{3.31}$$

That is, it is the sum of the waiting times.

The random variable Y_k has a *Pascal* distribution. That is,

$$P[Y_k = n] = P[S_{n-1} = k-1], X_n = 1] = \binom{n-1}{k-1} p^{k-1}(1-p)^{n-k}p \tag{3.32}$$

for $n = k, k+1, k+2, \ldots$. This is usually written more simply as

$$P[Y_k = n] = \binom{n-1}{k-1} p^{k}(1-p)^{n-k}. \tag{3.33}$$

Since Y_k is a sum of k independent geometric random variables, its mean is k/p and its variance is $k(1/p)(1/p-1)$.

The situation just described is called the *Bernoulli process*. There are two increasing random functions S_n (which counts successes up to a certain time) and Y_k (which counts the time of a certain success). Their relation is simple but subtle: the event that $Y_k > n$ is equivalent to the event that $S_n < k$. That is, the condition that at time n the kth success has not occurred is the same as the condition that at time n the number of successes is less than k.

It is instructive to do the experiment of running a Bernoulli process and plotting S_n as as function of n and Y_k as a function of k. The average slope of the S_n function is p, while the average slope of the Y_k function is $1/p$. They are not quite inverse functions to each other, though it is true that $S_{Y_k} = k$.

3.6 Problems

1. Say that X has values $1, 2, 3$ with probabilities $1/2, 1/3/1/6$. Say that Y also has values $1, 2, 3$ with probabilities $1/2, 1/3, 1/6$. Say that X, Y are independent. Find the joint probability mass function of X, Y.

2. In the previous problem, find the expectation of XY.

3. Say that X has values $1, 2, 3$ with probabilities $1/2, 1/3/1/6$. Say that Y also has values $1, 2, 3$ with probabilities $1/2, 1/3, 1/6$. Say that $X = Y$. Find the joint probability mass function of X, Y.

4. In the previous problem, find the expectation of XY.

5. Say that X, Y have joint probability mass function given by $p(1,1) = p(1,3) = p(3,1) = p(3,3) = 1/12$ and $p(1,2) = p(3,2) = p(2,1) = p(2,3) = 1/6$ and $p(2,2) = 0$. Find the probability mass functions of X, Y, $X+Y$, and XY. Find the expectations of these random variables.

6. In the proceeding problem, find the probability mass functions of $(X-2)^2$, $(Y-2)^2$, and $(X+Y-4)^2$. Find the expectations of these random variables.

7. In the preceding problem, is $E[XY] = E[X]E[Y]$? Are X, Y independent?

8. A two-stage industrial process takes a random amount of time equal with mean 8 and standard deviation 4. The next stage of the process takes a time with mean 2 and standard deviation 3. The two random times are independent. What is the mean and standard deviation of the sum of the two times? What is the mean and standard deviation of the average of these two times?

9. An ecologist wants to find the population mean weight of a certain plant by taking a sample of size 225 and computing the experimental sample mean. Unknown to the statistician, the population mean is 45 grams and the population standard deviation is 10 grams. What is the standard deviation of the sample means in this kind of experiment?

10. An ecologist wants to find the population mean weight of a certain plant by taking a sample of size 225 and computing the experimental sample mean. The experimental sample mean is 44 grams, and the experimental sample standard deviation is 4 grams. What is the statistician's estimate of the standard deviation of the sample means in this kind of experiment?

11. Prove the algebraic identity

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \sum_{i=1}^{n}(\bar{X}_n - \mu)^2. \tag{3.34}$$

    Hint: $(X_i - \mu) = (X_i - \bar{X}_n) + (\bar{X}_n - \mu)$.

12. Suppose that X_i in the preceding problem are independent random variables, each with mean μ and variance σ^2. Prove the identity

$$n\sigma^2 = E[(n-1)\hat{\sigma}_n^2] + \sigma^2. \tag{3.35}$$

    Also, prove that $E[\hat{\sigma}_n^2] = \sigma^2$.

13. Consider the waiting time random variable T in the Bernoulli process. Find the conditional probability $P[T > m+n \mid T > m]$.

14. Consider the Bernoulli process. Show that $Y_{S_n} \leq n$ and $n < Y_{S_n+1}$.

Chapter 4

Continuous random variables

A *random variable* X is a real function defined on the sample space. That is, it is a rule that assigns to every outcome a real number. A *continuous* random variable has the property that for each real number x the event $X = x$ has probability zero.

If X is a continuous random variable, then its *probability density function* is the function f_X defined by

$$P[a < X \leq b] = \int_a^b f_X(x)\,dx. \tag{4.1}$$

It is a real function defined on the real line. The values of the function f_X are not probabilities. The function instead represents a probability density. Thus $f_X(x)$ represents the probability per unit change in x at the point x.

It is always true that the integral

$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1. \tag{4.2}$$

Notice that the quantity

$$F_X(t) = P[-\infty < X \leq t] = \int_{-\infty}^{t} f_X(x)\,dx. \tag{4.3}$$

is a probability. Furthermore,

$$\frac{d}{dt}F_X(t) = f_X(t) \tag{4.4}$$

So this shows that the probability density function f_X is the rate of change of probability.

The function $F_X(t) = P[-\infty < X \leq t]$ is called the *cumulative distribution function* of the random variable X. Its values are probabilities.

4.1 Examples of continuous random variables

A random variable T is *exponential* with decay rate $\lambda > 0$ if its probability density function is

$$f_T(t) = \lambda e^{-\lambda t} \tag{4.5}$$

for $t \geq 0$ and $f_T(t) = 0$ for $t < 0$. Such a random variable is used to describe a random waiting time. Thus T is the random time (in seconds) until something happens. Thus λ is measured in inverse seconds. It is easy to compute that

$$P[0 \leq T < +\infty] = \int_0^\infty \lambda e^{-\lambda t}\, dt = 1. \tag{4.6}$$

In fact, for $0 \leq a < b$ we can even compute

$$P[a < T \leq b] = e^{-\lambda a} - e^{-\lambda b}. \tag{4.7}$$

Notice that the cumulative distribution function is

$$F_T(s) = P[0 \leq T \leq s] = 1 - e^{-\lambda s}. \tag{4.8}$$

A random variable X is *normal* with mean μ and standard deviation σ if its probability density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{4.9}$$

It is not obvious but true that

$$P[-\infty < X < +\infty] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = 1. \tag{4.10}$$

The normal distribution has a density given by the famous bell curve. The maximum is at μ, and it is symmetric about μ. It falls down on each side rather rapidly. By the point where x is 3σ above μ or below 3σ below μ it is rather close to zero.

The following facts about the normal distribution should be memorized:

$$P[\mu - \sigma < X < \mu + \sigma] = \int_{\mu-\sigma}^{\mu+\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = 0.68 \tag{4.11}$$

approximately, and

$$P[\mu - 2\sigma < X < \mu + 2\sigma] = \int_{\mu-2\sigma}^{\mu+2\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = 0.95 \tag{4.12}$$

approximately. The last approximation is a bit low, but it gives a reasonable first idea of where the probability is located.

There is no simple formula for the cumulative distribution function $F_X(t)$ for the normal distribution. However we see from the above that to a very rough approximation $F_X(\mu - 2\sigma) = 0.025$, $F(\mu - \sigma) = 0.16$, $F(\mu) = 0.5$, $F(\mu + \sigma) = 0.84$, and $F(\mu + 2\sigma) = 0.975$. A table or computer calculation will give more accurate values.

4.2 The empirical distribution

Let X be a continuous random variable. In this section its probability density function is denoted $f(x)$. Thus probabilities are computed by integrating $f(x)$. Thus

$$F_X(t) = P[-\infty < X \leq t] = \int_{-\infty}^{t} f(x)\,dx. \tag{4.13}$$

Consider a super-experiment consisting of n independent repetitions of the original experiment. Call the experimental values in this super-experiment $X_1, X_2, X_3, \ldots, X_n$. With probability one these are all distinct numbers. The *empirical distribution* is the discrete probability distribution that is uniform on these n points. In other words, the probability associated with each of the points X_j is $1/n$.

The number of observations up to t is the random variable

$$N(t) = \#\{i \mid X_i \leq t\}. \tag{4.14}$$

For each t this is a binomial random variable with parameters n and $F(t)$. The cumulative distribution function of the empirical distribution is the random function

$$\hat{F}(t) = \frac{N(t)}{n}. \tag{4.15}$$

Notice that $\hat{F}(t)$ jumps by $1/n$ at each observation. For n large the experimental cumulative distribution function $\hat{F}(t)$ should be close to the mathematically computed cumulative distribution function $F(t)$.

4.3 Functions of a random variables

Say that X is a discrete random variable, and $Y = g(X)$. This means that after the experimental number X is found, a further calculation is done to produce Y. The relation between the two probability density functions is that

$$f_Y(y) = \sum_{\{x \mid g(x)=y\}} f_X(x) \frac{1}{|g'(x)|}. \tag{4.16}$$

This comes from the formula for change of variable in an integral.

For example, if $Y = X^2$, then $g(x) = x^2$ and so $g'(x) = 2x$. Thus for fixed $y > 0$ there are two solutions for x, namely $x = \sqrt{y}$ and $x = -\sqrt{y}$. Thus

$$f_{X^2}(y) = f_X(\sqrt{y}) \frac{1}{2\sqrt{y}} + f_X(-\sqrt{y}) \frac{1}{2\sqrt{y}} \tag{4.17}$$

for $y > 0$. (The density of $Y = X^2$ is zero for $y < 0$.)

4.4 Expectation

Say that X is a discrete random variable. Its *expectation* is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx. \tag{4.18}$$

Sometimes the expectation of a random variable is called the *mean* of the random variable. This is often written as

$$\mu_X = E[X]. \tag{4.19}$$

Here are two important properties of expectation involving constant random variables. One is

$$E[c] = c. \tag{4.20}$$

The other is

$$E[cX] = cE[X]. \tag{4.21}$$

There is a theorem that gives a simple way of computing the expectation of a random variable $g(X)$ that is a function of a random variable X. The formula is

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx. \tag{4.22}$$

The *variance* of a random variable is

$$\mathrm{var}[X] = E[(X - E[X])^2]. \tag{4.23}$$

This formula can be made totally obscure by writing it in the alternate form

$$\mathrm{var}[X] = E[X^2] - E[X]^2. \tag{4.24}$$

This is common. Sometimes the variance of a random variable is written as

$$\sigma_X^2 = \mathrm{var}[X]. \tag{4.25}$$

Another useful quantity is the *standard deviation*

$$\sigma_X = \sqrt{\mathrm{var}[X]}. \tag{4.26}$$

For an exponential random variable the mean is

$$E[T] = \int_0^{\infty} t\lambda e^{-\lambda t}\, dt = \frac{1}{\lambda} \tag{4.27}$$

and the variance is

$$\mathrm{var}[T] = \int_0^{\infty} (t - \frac{1}{\lambda})^2 \lambda e^{-\lambda t}\, dt = \frac{1}{\lambda^2}. \tag{4.28}$$

Here the standard deviation is the same as the mean.

For a normal random variable the mean is

$$E[X] = \int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = \mu \tag{4.29}$$

and the variance is

$$\text{var}[X] = \int_{-\infty}^{\infty} (x-\mu)^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx = \sigma^2. \tag{4.30}$$

4.5 Problems

1. The average time from a given moment to the next large asteroid impact is 80,000 years. What is the probability of waiting 200,000 years and having no asteroid impact?

2. Certain survey results are normally distributed and have a mean of 400 kilometers and a standard deviation of 1/2 kilometer. What is the probability that the actual experimental result is off by one kilometer or more?

3. The prices for a certain commodity are uniformly distributed on the interval from 200 to 212. What is the probability of finding a price above 208?

4. A random variable X is normal with mean zero and variance 1. What is the probability density function of X^2?

5. A random variable U is uniform on the interval from 0 to 1. What is the probability density function of $(U - 1/2)^2$?

6. The standard normal random variable has density

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{4.31}$$

   This has integral one. The mean is obviously zero (by symmetry). Use integration by parts to calculate the variance

$$\sigma^2 = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \, dz. \tag{4.32}$$

   Show all steps in detail.

7. We know that for $\lambda > 0$ we have

$$\int_0^{\infty} \lambda e^{-\lambda t} \, dt = 1. \tag{4.33}$$

   Differentiate to prove that

$$\mu = \int_0^{\infty} t \, \lambda e^{-\lambda t} \, dt = \frac{1}{\lambda}. \tag{4.34}$$

8. We know from the preceding problem that

$$\int_0^\infty \lambda^2 t e^{-\lambda t}\, dt = 1. \tag{4.35}$$

Differentiate to find the second moment

$$m_2 = \int_0^\infty t^2\, \lambda e^{-\lambda t}\, dt \tag{4.36}$$

Use this to calculate the variance $\sigma^2 = m_2 - \mu^2$.

9. Let U be uniform on the interval from 0 to 1. Its mean is $1/2$. Calculate its variance

$$\sigma^2 = \int_0^1 (u - 1/2)^2\, du. \tag{4.37}$$

Another way of computing the same variance is

$$\sigma^2 = \int_0^{\frac{1}{4}} w \frac{1}{\sqrt{w}}\, dw. \tag{4.38}$$

Carry out the computation. Give a clear and complete explanation of why this gives the same answer.

10. Let Y be uniform on the interval from a to b. What is its probability density function? Use the previous problem to effortlessly compute the mean and variance of Y. Hint: $Y = a + (b - a)U$.

Chapter 5

The Poisson process

5.1 Functions of several random variables

Let X, Y be two continuous random variables. Their *joint probability density function* is defined by

$$\int\int_{\{(x,y)\in B\}} f_{X,Y}(x,y)\,dx\,dy = P[(X,Y)\in B]. \tag{5.1}$$

That is, the integral of the joint probability density function over the set B is the probability that the random point (X,Y) is in the set B.

The *marginal* distributions of the joint probability mass function are

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dy \tag{5.2}$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx. \tag{5.3}$$

Say that a random variable $g(X,Y)$ is a function of random variables X, Y. Then its expectation may be computed from

$$E[g(X,Y)] = \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\,dx\,dy. \tag{5.4}$$

A very important special case of this is the relation

$$E[X+Y] = E[X] + E[Y]. \tag{5.5}$$

This comes from taking $g(x,y) = x+y$. The derivation is as follows. We have

$$\begin{aligned}\int_{-\infty}^{\infty}(x+y)f_{X,Y}(x,y)\,dx\,dy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\,dx\,dy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} y f_{X,Y}(x,y)\,dx\,dy \\ &= \int_{-\infty}^{\infty} x f_X(x)\,dx + \int_{-\infty}^{\infty} y f_Y(y)\,dy. \end{aligned}\tag{5.6}$$

5.2 Independence

Two continuous random variables X, Y are *independent* provided that for each x, y we have $f_{X,Y}(x, y) = f_X(x) f_Y(y)$.

A crucially important result is that if X, Y are independent, then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]. \tag{5.7}$$

The proof of this is as follows. Since $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ the left hand side factors to become the right hand side:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x)h(y) f_X(x) f_Y(y)\, dx\, dy = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx \int_{-\infty}^{\infty} g(y) f_Y(y)\, dy. \tag{5.8}$$

A simple corollary of this is when $g(x) = x$ and $h(y) = y$. The if X, Y are independent, then it follows that

$$E[XY] = E[X]E[Y]. \tag{5.9}$$

An extremely important corollary is that if X, Y are independent, then

$$\text{var}[X + Y] = \text{var}([X] + \text{var}[Y]. \tag{5.10}$$

The proof of this is simple. We have

$$\text{var}[X+Y] = E[(X+Y-E[X+Y])^2] = E[(X-E[X])^2+2E[(X-E[X])(Y-E[Y])]+E[(Y-E[Y])^2]. \tag{5.11}$$

This holds even if the random variables are not independent. But if they are independent, then $X - E[X]$ and $Y - E[Y]$ are also independent. Therefore the cross term is

$$E[(X - E[X])(Y - E[Y])] = E[(X - E[X])]E[(Y - E[Y])] = 0 \cdot 0 = 0. \tag{5.12}$$

In terms of standard deviation, the result says that for independent random variables X, Y we have

$$\sigma_{X+Y} = \sqrt{\sigma_X^2 + \sigma_Y^2}. \tag{5.13}$$

The consequences for the weak law of large numbers is the same as before.

5.3 Sums of independent normal random variables

The normal distribution has a remarkable property. Let X, Y be independent normal random variables. Then $X + Y$ is also normal!

Of course one can be more specific. If X, Y have means μ_X and μ_Y and variances σ_X^2 and σ_Y^2 and are independent, then $X + Y$ has mean $\mu_X + \mu_Y$ and variance $\sigma_X^2 + \sigma_Y^2$. Since for a normal random variable the mean and variance

determine the normal distribution, we know everything about the distribution of $X + Y$.

This situation arises often in statistics. Let $X_1, \ldots, X_n$ be independent normal random variables, each with mean μ and variance σ^2. Then the sum $X_1 + \cdots + X_n$ has mean $n\mu$ and variance $n\sigma^2$, and it is also normal. Also the sample mean $\bar{X}_n = (X_1 + \cdots + X_n)/n$ has mean μ and variance σ^2/n, and it is also normal. So the process of adding or of averaging preserves the property of normality.

5.4 The Poisson process

At any time a major earthquake (or a cluster of such earthquakes) may occur some place in the world. The average rate of such earthquakes per year is λ. A radioactive substance has a decay event every now and then. The rate at which such events happen on the average is λ per second. Or a taxi comes around the corner. This happens at a certain average rate per hour. When everything is as random as possible, except for the constraint of there being a certain long term average rate, then the situation is described by the Poisson process.

This is a random increasing function $N(t)$ that counts the number of events that have occurred up to and including time t. For simplicity we take the starting time to be zero, so $N(0) = 0$ and assume that nothing happens at time zero. Thus for $0 \leq s < t$ the difference $N(t) - N(s)$ is the number of events that occur in the interval $(s, t]$. This difference random variable is required to be Poisson with mean $\lambda(t - s)$. Thus

$$P[N(t) - N(s) = k] = \frac{(\lambda(t-s))^k}{k!} e^{-\lambda(t-s)}. \tag{5.14}$$

If two such intervals are disjoint, then the corresponding random variables are required to be independent.

Let Y_k be defined as the first t such that $N(t) = k$. In the example, Y_k is the time of the kth major quake, or the kth radioactive decay. Set $Y_0 = 0$, and define $T_k = Y_k - Y_{k-1}$ for $k \geq 1$. Thus T_k is the waiting time from the $k - 1$st success to the kth success. Each T_k is a *exponential* random variable with probability density function

$$f_{T_k}(t) = \lambda e^{-\lambda t} \tag{5.15}$$

for $t \geq 0$. The intuition is that the probability that the wait is between t and $t + dt$ is the probability that the wait has been futile up to time t (which is $e^{-\lambda t}$) times the probability that something then happens very soon (which is $\lambda\, dt$).

Furthermore, the random variables $T_1, T_2, T_3, \ldots$ are independent. The time Y_k for the rth success is

$$Y_k = T_1 + \ldots + T_k. \tag{5.16}$$

That is, it is the sum of the waiting times.

The random variable Y_k has a *Erlang* distribution. That is, its probability density function is

$$f_{Y_k}(t) = \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} \lambda \tag{5.17}$$

for $t \geq 0$. This is a Poisson probability times an extra factor of λ. The intuition is that the probability that the kth event is between t and $t+dt$ is the probability that $k-1$ events have occurred up to t (which is the Poisson part) times the probability that something then happens very soon (which is $\lambda\, dt$). The formula is usually written in a less intuitive form as

$$f_{Y_k}(t) = \frac{\lambda^k t^{k-1}}{(k-1)!} e^{-\lambda t}. \tag{5.18}$$

Since Y_k is a sum of k independent geometric random variables, its mean is k/λ and its variance is k/λ^2.

The situation just described is called the *Poisson process*. There are two increasing random functions $N(t)$ (which counts successes up to time t) and Y_k (which counts the time of a certain success). Their relation is simple but subtle: the event that $Y_k > t$ is equivalent to the event that $N(t) < k$. That is, the condition that at time n the kth success has not occurred is the same as the condition that at time n the number of successes is less than k. This fact leads to an easy derivation of the Erlang density, by differentiating $P[Y_k > t] = P[N(t) < k]$.

It is instructive to do the experiment of running a Poisson process and plotting $N(t)$ as as function of t and Y_k as a function of k. The average slope of the $N(t)$ function is λ, while the average slope of the Y_k function is $1/\lambda$. They are not quite inverse functions to each other, though it is true that $N(Y_k) = k$.

5.5 Problems

1. Let U_1, U_2 be independent random variables, uniformly distributed on the interval from 0 to 1. Their joint distribution is then uniform on the square of side one. Use elementary geometry to find the cumulative probability distribution function $P[U_1 + U_2 \leq x]$ for $0 \leq x \leq 2$. The idea is to find the area common to the unit square and the region $u_1 + u_2 \leq x$. Differentiate to find the probability density function for the random variable $U_1 + U_2$.

2. Say that T_1 and T_2 are independent exponential random variables with expectations $1/\lambda_1$ and $1/\lambda_2$. Find the joint probability density function as a function of t_1, t_2.

3. Let M_1 be the smallest of T_1, T_2. This tells how long one has to wait for the first of the two things to happen. Find the probability density function of M_1. Is it exponential? Hint: Begin with $P[M_1 > s] = P[T_1 > s, T_2 > s]$. Then use independence.

4. Let M_2 be the largest of T_1, T_2. This tells how long one has to wait for both of the two things to happen. Find the probability density function of M_2. Is it exponential? Hint: Calculate $P[M_2 \leq t] = P[T_1 \leq t, T_2 \leq t]$.

5. Find the expectations of M_1 and of M_2. Hint: First find the expectation of M_1. Then use $M_1 + M_2 = T_1 + T_2$.

6. Here is a general fact. Let X_1 and X_2 be random variables with joint probability density function $f(x_1, x_2)$. Let $Y_1 = \min(X_1, X_2)$ and $Y_2 = \max(X_1, X_2)$ be the smallest and the largest of the two observations. Then the joint probability density function of Y_1, Y_2 is the function $f^*(y_1, y_2)$ equal to $f(y_1, y_2) + f(y_2, y_1)$ if $y_1 < y_2$ and equal to 0 if $y_1 > y_2$.

   Use this general fact to find the joint probability density function of M_1, M_2 in the preceding problems. Are the random variables M_1, M_2 independent?

7. The Ask Marilyn column in Parade magazine for March 13, 2005 had a letter from Patricia Ann Waddell of Spokane. "I work at a waste treatment plant, and we do assessments of the time-to-failure and time-to-repair of the equipment, then input these figures into a computer model to make plans. But when I need to explain the process to people in other departments, I find it difficult. Say a component has two failure modes. One occurs every 5 years, and the other occurs every 10 years. People usually say that the time-to-failure is 7.5 years, but this is incorrect. It's between 3 and 4 years. Do you know of a way to explain this that people will accept?" Discuss. In particular, calculate the correct time-to-failure.

8. Say that $u = az + bw$ and $v = -bz + aw$ and $a^2 + b^2 = 1$. Show that $u^2 + v^2 = z^2 + w^2$.

9. A standard normal random variable has mean zero and variance one. Say that Z and W are independent standard normal random variables. Suppose that a and b are real numbers with $a^2 + b^2 = 1$. Show that $U = aZ + bW$ and $V = -bZ + aW$ are independent standard normal random variables. Hint: Compute their joint density. Show by explicit computation that the product $\exp(-z^2/2)\exp(-w^2/2)$ is equal to the product $\exp(-u^2/2)\exp(-v^2/2)$. Be explicit about what algebraic properties of the exponential function you use.

10. In the last problem we say that if Z, W are independent standard normal random variables, and $a^2 + b^2 = 1$, then $aZ + bW$ is standard normal. Let X, Y be mean zero normal random variables with variances σ^2 and τ^2. Prove that $X + Y$ is a mean zero normal random variable with variance $\sigma^2 + \tau^2$. Hint: Use the identity

$$X + Y = \sqrt{\sigma^2 + \tau^2} \left[\frac{\sigma}{\sqrt{\sigma^2 + \tau^2}} \frac{X}{\sigma} + \frac{\tau}{\sqrt{\sigma^2 + \tau^2}} \frac{Y}{\tau} \right]. \tag{5.19}$$

11. The errors in a certain measurement are of two kinds. The first kind X is normal with mean zero and standard deviation 3 millimeters. The second one Y is normal with mean zero and standard deviation 4 millimeters. The total error is the sum $X + Y$. What is the probability that this total error has absolute value less than 5 millimeters?

12. Consider a normal population with mean 80 and standard deviation 10. An independent sample of size 400 is taken. What is the probability that the sample mean is between 79 and 81?

13. Consider the waiting time random variable T in the Poisson process. Suppose $a > 0$ and $b > 0$. Find the conditional probability $P[T > a + b \mid T > a]$.

14. Consider the Poisson process. Show that $Y_{N(t)} \leq t$ and $t < Y_{N(t)+1}$.

Chapter 6

The Central Limit Theorem

6.1 The central limit theorem

The *central limit theorem* is the following remarkable result. Let $X_1, \ldots, X_n$ be independent random variables, each with the same distribution. Let μ be the mean, and let σ be the standard deviation. Suppose that σ is finite. Write

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}. \tag{6.1}$$

Then as $n \to \infty$ the cumulative distribution function of Z_n converges to the standard normal cumulative distribution function.

The central limit theorem says that even if the population distribution is not normal, for a large enough sample the distribution of the sample mean is normal. In fact, it is normal with mean μ equal to the population mean and with standard deviation $\sigma/\sqrt{n}$, where σ is the population standard deviation.

6.2 The central limit theorem in statistics

In statistics the population mean μ and the population standard deviation σ are unknown. In fact, the population distribution is unknown. But if one is interested in finding useful information about the population mean μ, then there is a method. Use the sample mean $\hat{\mu} = \bar{X}_n$ to estimate μ. Use the sample standard deviation $\hat{\sigma} = s$ to estimate σ. The standard deviation of the sample means in this type of sampling experiment is $\sigma/\sqrt{n}$. According to the central limit theorem, for a reasonably large sample the distribution of the sample means $\bar{X}_n$ in this sort of experiment is normal with mean μ and standard deviation $\sigma/\sqrt{n}$. So the probability that $\bar{X}_n$ and μ differ by more than $2\sigma/\sqrt{n}$ is only about five percent.

So a statistician might claim that the particular experimental sample mean is off from the true population mean by about $2s/\sqrt{n}$, or less. Only about five percent of such statisticians will be wrong.

6.3 The central limit theorem for Bernoulli random variables

The central limit for independent Bernoulli random variables is the special case when each X_i is 0 or 1. As we know, then $\mu = p$ and $\sigma = \sqrt{p(1-p)}$.

So in this case $S_n = X_1 + \cdots + X_n$ is the number of successes in n independent trials. This is a binomial random variable. So the central limit theorem in this case is stated in terms of

$$Z_n = \frac{S_n - np}{\sqrt{n}\sqrt{p(1-p}} = \frac{\hat{p}_n - \mu}{\sqrt{p(1-p)}/\sqrt{n}}. \tag{6.2}$$

As $n \to \infty$ the cumulative distribution function of Z_n converges to the standard normal cumulative distribution function.

6.4 Correlation

The *covariance* of two random variables X, Y is

$$\text{cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)]. \tag{6.3}$$

Sometimes people use the alternate definition

$$\text{cov}[X, Y] = E[XY] - E[X]E[Y], \tag{6.4}$$

but this is not intuitive. Note, by the way, that

$$\sigma_X^2 = \text{var}[X] = \text{cov}[X, X]. \tag{6.5}$$

The *correlation* of two random variables X, Y is

$$\rho_{X,Y} = \frac{\text{cov}[X, Y]}{\sigma_X \sigma_Y}. \tag{6.6}$$

It may be proved that it is always true that

$$-1 \leq \rho_{X,Y} \leq 1. \tag{6.7}$$

Random variables X, Y are said to be *uncorrelated* if $\rho_{X,Y} = 0$. Of course this is equivalent to $\text{cov}[X, Y] = 0$ or to the identity $E[XY] = E[X]E[Y]$.

If two random variables X, Y are independent, then they are uncorrelated. Notice that the weak law of large numbers is true for uncorrelated random variables.

6.5 The bivariate normal distribution

The *bivariante normal* distribution (or bivariate Gaussian distribution) is determined by five parameters: $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho_{X,Y}$. It has density given by

$$f_{X,Y}(x,y) = c\exp(-\frac{1}{2(1-\rho_{X,Y}^2)}[\frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho_{X,Y}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}]). \tag{6.8}$$

The normalization constant c is there to make the total probability equal to one. It works out to be

$$c = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{X,Y}^2}}. \tag{6.9}$$

When the correlation $\rho_{X,Y} = 0$, the bivariate normal distribution factors:

$$f_{X,Y}(x,y) = c\exp(-\frac{1}{2}[\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}]) = c\exp(-\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2})\exp(-\frac{1}{2}\frac{(y-\mu_Y)^2}{\sigma_Y^2}). \tag{6.10}$$

Thus in this case the variables are independent.

If two bivariate normal random variables are uncorrelated, then they are independent.

6.6 The sample correlation

Consider an independent sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from some population of paired random variables, all with the same distribution. The *sample correlation coefficient* is the random variable

$$r = \frac{\sum_{j=1}^n (X_j - \bar{X}_n)(Y_j - \bar{Y}_n)}{\sqrt{\sum_{j=1}^n (X_j - \bar{X}_n)^2}\sqrt{\sum_{j=1}^n (Y_j - \bar{Y}_n)^2}}. \tag{6.11}$$

In statistics one sometimes hopes that the distribution of the variables is joint normal. Then there are just the five parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho_{X,Y}$. They are then estimated by the five sample statistics $\bar{X}_n, \bar{Y}_n, s_X, s_Y, r$. When n is large, these estimates should be close to the parameters, with high probability.

In particular, the sample correlation coefficient r for an independent sample from a bivariate normal distribution has a standard deviation that is given approximately by

$$\sigma_r \approx \frac{1-\rho^2}{\sqrt{n}}. \tag{6.12}$$

So if n is reasonably large, then the sample correlation r is likely to be somewhat close to the population correlation ρ. However even if $\rho = 0$, the sample quantity r will be non-zero—just not very large.

6.7 Correlated events

Suppose A, B are events. Their covariance is

$$\text{cov}[A, B] = P[A, B] - P[A]P[B]. \tag{6.13}$$

The events are independent precisely when their covariance is zero.

An event A has variance

$$\text{var}[A] = \text{cov}[A, A] = P[A] - P[A]^2 = P[A](1 - P[A]) = P[A]P[A^c]. \tag{6.14}$$

The correlation between two events is thus

$$\rho_{A,B} = \frac{P[A, B] - P[A]P[B]}{\sqrt{P[A]P[A^c]P[B]P[B^c]}}. \tag{6.15}$$

Say that the the experiment is repeated n times. Let N_A, N_B be the number of times that A, B occur, while N_{AB} is the number of times A and B occur at on the same repetition. The sample correlation for two events is

$$r = \frac{nN_{AB} - N_A N_B}{\sqrt{N_A N_{A^c} N_B N_{B^c}}}. \tag{6.16}$$

This can also be written

$$r = \frac{N_{AB}N_{A^cB^c} - N_{AB^c}N_{A^cB}}{\sqrt{N_A N_{A^c} N_B N_{B^c}}}. \tag{6.17}$$

If the sample size n is large, then r should be a good estimate of ρ.

This subject is often presented in the context of a 2 by 2 *contingency table*. The data is given as

$$\begin{array}{|c|c|} \hline N_{AB} & N_{AB^c} \\ \hline N_{A^cB} & N_{A^cB^c} \\ \hline \end{array}. \tag{6.18}$$

The sum of the four numbers is the sample size n. The numerator in the r statistic is the determinant of this 2 by 2 matrix.

6.8 Tests for independence

Sometimes statisticians want to test two variables for independence. A popular device is to use the statistic

$$Z = \frac{\hat{p}_{A|B} - \hat{p}_{A|B^c}}{\sqrt{\hat{p}_A \hat{p}_{A^c}}\sqrt{\frac{1}{N_B} + \frac{1}{N_{B^c}}}}. \tag{6.19}$$

Here $\hat{p}_{A|B} = N_{AB}/N_B$ and $\hat{p}_{A|B_c} = N_{AB^c}/N_{B_c}$ are the two proportions being compared.

If the two variables are independent, then the conditional probabilities $P[A \mid B] = P[A \mid B^c]$ are both equal to $P[A]$. Then it is reasonable to use the pooled

$\hat{p}_A = N_A/n$ to estimate $P[A]$. The denominator of Z is thus a reasonable estimate of the standard deviation of the difference in the numerator, at least in the situation when the variables are independent. So it is plausible (and true) that in this case Z has an approximate standard normal distribution.

It is a remarkable fact that this Z is essentially the same statistic as the sample correlation coefficient r. In fact, their relation is

$$r = \frac{Z}{\sqrt{n}} \tag{6.20}$$

In the case of independence, when the correlation $\rho = 0$, the sample correlation r should also be close to zero. In fact the Z statistic is approximately a standard normal random variable. This is often used in statistics to test whether $\rho = 0$ or not. So, for instance, if in the experiment $|Z| \leq 2$, then one would regard that as consistent with independence. But if $|Z|$ were much larger than that, then one would suspect a lack of independence.

Sometimes yet another test for independence is used, the chi-squared test involving

$$\chi^2 = \frac{(N_{AB} - N_A N_B/n)^2}{N_A N_B/n} + \frac{(N_{AB^c} - N_A N_{B^c}/n)^2}{N_A N_{B^c}/n} + \frac{(N_{A^c B} - N_{A^c} N_B/n)^2}{N_{A^c} N_B/n} + \frac{(N_{A^c B^c} - N_{A^c} N_{B^c}/n)^2}{N_{A^c} N_{B^c}/n} \tag{6.21}$$

However this is nothing new, in fact for the 2 by 2 table

$$\chi^2 = Z^2. \tag{6.22}$$

6.9 Correlation and cause

If A, B are positively correlated, then it might be tempting to say that B helps to cause A. However this is too simple.

It may be that there is another event C such that $P[A, B \mid C] = P[A \mid C]P[B \mid C]$, that is, A, B are *conditionally independent* given C. This is equivalent to $P[A \mid B, C] = P[A \mid C]$. Similarly, we can say that A, B are *conditionally positively correlated* given C if $P[A, B \mid C] > P[A \mid C]P[B \mid C]$. This is equivalent to $P[A \mid B, C] > P[A \mid C]$. Of course there is a similarly definition for being conditionally negative correlated.

For instance, B might be having yellow-stained fingers, and A might be having lung cancer. Then C might be smoking cigarettes. So even if A, B are positively correlated, we would not want to say that B causes A. It might well be that A, B are independent given C. Intuitively, one would expect that A, C are positively correlated, and it is C that causes A.

Here is another example. Suppose that A represents having heart disease, while B is smoking and C is exercising. Then it is possible that $P[A \mid B, C] > P[A \mid C]$, $P[A \mid B, C^c] > P[a \mid C^c]$, yet $P[A \mid B] < P[A]$. Smokers who exercise are more likely to get heart disease, Smokers who do not exercise are more likely to get heard disease, smokers are less likely to get heart disease. This could well

be true in a world where smokers like to exercise. This is called *Simpson's paradox*.

Another example of Simpson's paradox comes from admission rates at a graduate school in California. About 45% of men applicants were admitted; about 30% of women applicants were admitted. However the explanation was that women applied in large numbers to schools such as Law that admitted fewer than 10%. Men applied to schools such as Engineering that had admission rates above 50%.

Here is a definition of causation that has been proposed. Say that B causes A if, for every T that causes A but is not caused by B, the events A, B are conditionally positively correlated given T. For example, let A be lung cancer and B be smoking. Say that T is exercise.

However, notice that one could not take T to be something like tar in the lungs. In this case A, B may be conditionally independent given T. This choice of T violates the condition that T is not caused by B.

One problem with the proposed definition is that it is circular.

6.10 Problems

1. Prove that

$$\text{cov}[A, B] = (P[A \mid B] - P[A \mid B^c])P[B]P[B^c]. \tag{6.23}$$

2. Prove that

$$\rho_{A,B} = \frac{P[A \mid B] - P[A \mid B^c]}{\sqrt{P[A]P[A^c]}\sqrt{\frac{1}{P[B]} + \frac{1}{P[B^c]}}}. \tag{6.24}$$

3. Prove that the formula for the correlation of two events is a special case of the usual formula for correlation.

4. Prove that the formula for the sample correlation for two events is a special case of the usual formula for sample correlation.

5. Prove that the two formula for the sample correlation for events are the same.

6. Prove that $Z = r\sqrt{n}$, where Z is the statistic involving the difference of sample proportions and r is the sample correlation for events.