Lectures on Elementary Probability

William G. Faris

February 22, 2002

Contents

Chapter 1

Combinatorics

1.1 Factorials and binomial coefficients

The *factorial* number $n!$ is the number of ways of listing n objects in order. It is also called the number of *permutations* of the n objects. The number

$$(n)_k = \frac{n!}{(n-k)!} \tag{1.1}$$

is the number of ways of making a list of k out of the n objects in order. Sometimes this is called the number of *permutations* of size k of the n objects.

The *binomial coefficient*

$$\binom{n}{k} = \frac{(n)_k}{k!} \tag{1.2}$$

is the number of ways of choosing k of the n objects without regard to order. This is also the number of *combinations* of size k from the n objects. It is just the number of subsets of size k of a set of size n.

The basic formula for computing binomial coefficients is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}. \tag{1.3}$$

Note the important identity

$$\binom{n}{k} = \binom{n}{n-k}. \tag{1.4}$$

1.2 Sampling with replacement

Let I be a set with n elements and let M be a set with m elements. A *function* from I to M is a rule that associates to each element of I a corresponding element of M. The are various other descriptions of this concept. If I is a set of trials and M is a population, then a function from I to M is called a *ordered sample with replacement* of size n from M. If I is a set of balls and M is a set of urns, then a function from I to M is a way of placing the balls in the urns. If I is a set and M is an set with m elements, then a function from I to M is a way of partitioning the set I into m disjoint subsets (some of which may be empty) labeled by the elements of M.

Theorem 1.1 *Suppose I has n elements and M has m elements. The number of functions from I to M is m^n.*

Given a function f from I to M, there is another function χ from M to the natural numbers $\{0, 1, 2, 3, \ldots\}$. This function χ is called the *occupation number* function. It is defined so that for each y in M the value $\chi(y)$ is the number of x in I such that $f(x) = y$. Note that $\sum_{y \in M} \chi(y) = n$.

Such a function χ from M to the natural numbers could be called an *unordered sample with replacement*, since it ignores the order in which the elements of N are chosen. However note that the relation between the unordered sample with replacement and the ordered sample with replacement is rather complicated, given by the multinomial coefficients. A multinomial coefficient counts the number of ways of partitioning an n element set I into m subsets labeled by the elements of M such that the number of elements labeled by y in M is $\chi(y)$.

Theorem 1.2 *Suppose I has n elements and M has m elements. Given a function χ from M to the natural numbers with $\sum_{y \in M} \chi(y) = n$ (that is, given the occupation numbers), the number of functions f from I to M with these occupation numbers is the* multinomial coefficient

$$\binom{n}{\chi} = \frac{n!}{\prod_{y \in N} \chi(y)!}. \tag{1.5}$$

If $M = \{1, \ldots, m\}$ and if we write $n_i = \chi(i)$ for the occupation numbers, then this is

$$\binom{n}{n_1 \ldots n_m} = \frac{n!}{n_1! \cdots n_m!}. \tag{1.6}$$

Proof: Divide the set I into m categories by first picking the subset that comprises the first category, then the subset of the remaining elements that comprise the second category, and so on. This gives the multinomial coefficient as a product of binomial coefficients:

$$\binom{n}{n_1 \ldots n_m} = \binom{n}{n_1}\binom{n - n_1}{n_2}\binom{n - n_1 - n_2}{n_3} \cdots \binom{n - n_1 - n_2 - \cdots - n_{m-1}}{n_m}. \tag{1.7}$$

This may now be expressed in terms of factorials.

It is interesting to note that the total number of functions from I to M may be written in the form

$$\sum_{\chi} \binom{n}{\chi} = m^n. \tag{1.8}$$

Theorem 1.3 *The total number of occupation number functions χ on a set M with m elements such that $\sum_{y\in M}\chi(y) = n$ is given by the binomial coefficient*

$$\binom{m+n-1}{n} = \binom{m+n-1}{m-1}. \tag{1.9}$$

Proof: Consider a set of $m+n-1$ elements arranged in a row. For each way of choosing a subset of $m-1$ elements as barriers there is a partition of the remaining n elements into m subsets with sizes given by a set of occupation numbers.

Binomial coefficients are a special case of multinomial coefficients. Let I have n elements. A subset of I is the same as a specification of a function from I to $\{0,1\}$ that is 1 on the subset and 0 on its complement. Say that one specifies occupation numbers $n-k = \chi(0)$ and $k = \chi(1)$. Then the number of functions with these occupation numbers is the multinomial coefficient $\binom{n}{\chi}$ which is also the number of subsets of size k given by the binomial coefficient $\binom{n}{k}$.

1.3 Sampling without replacement

Let I be a set with n elements and let M be a set with m elements. An *injective function* f from I to M is a rule that associates to each element of I a corresponding element of M, and such that distinct elements of I are associated to distinct elements of M. Such a function is also called an *injection* or a *one-to-one* function. The are various other descriptions of this concept. If I is a set of trials and M is a population, then an injective function from I to M is called an *ordered sample without replacement* of size n from M. If I is a set of balls and M is a set of urns, then an injective function from I to M is a way of placing the balls in the urns so that no two balls go in the same urn.

Theorem 1.4 *Suppose I has n elements and M has m elements. The number of injective functions f from I to M is $(m)_n$.*

Given an injective function f from I to M, there is another function χ from M to the numbers $\{0,1\}$. This function χ is just the same occupation number function as before, only restricted to have values 0 and 1. It is defined so that for each y in M the value $\chi(y)$ is the number of x in I such that $f(x) = y$. Note that $\sum_{y\in M}\chi(y) = n$.

Such a function χ from M to $\{0,1\}$ is called an *indicator function.* (In some contexts it is also called a characteristic function.) This is because it indicates or characterizes a subset of A of M, namely the subset on which the function χ has the value 1. Such a subset is also called an *unordered sample without replacement* from M. The number of such indicator functions, that is, the number of such subsets, is 2^m.

Suppose I has n elements and M has m elements. Given an indicator function χ from M to the natural numbers with $\sum_{y\in M}\chi(y) = n$, the number of injective functions f from I to M with this indicator function is $n!$. In other words, given a subset A of M with n elements, the number of injective functions from I to M with range A is $n!$.

This theorem says that there is a simple relation between unordered samples without replacement and ordered samples without replacement. For every one of the former there are exactly $n!$ of the latter.

Theorem 1.5 *Let M be a set with m elements. The total number of indicator functions χ on M with sum n (that is, the total number of subsets A of M having n elements) is given by the binomial coefficient*

$$\binom{m}{n} = \frac{(m)_n}{n!} \tag{1.10}$$

Proof: The total number of injective functions from I to M is the product of the number of subsets that may be taken as the range times the number of injections with that given range. This is

$$\binom{m}{n} n! = (m)_n. \tag{1.11}$$

Note that the total number of subsets is the sum over n of the number of subsets of size n. This gives the identity

$$\sum_{n=0}^{m} \binom{m}{n} = 2^m. \tag{1.12}$$

Chapter 2

Probability Axioms

2.1 Logic and sets

In probability there is a set called the *sample space* S. An element of the sample space is called an *outcome* of the experiment.

An *event* is identified with a subset E of the sample space S. If the experimental outcome belongs to the subset, then the event is said to happen.

We can combine events by set theory operations. Each such operation corresponds to a logical operation.

Union corresponds to *or*: s is in $E \cup F$ if and only if s is in E or s is in F.

Intersection corresponds to *and*: s is in $E \cap F$ if and only if s is in E and s is in F.

Complement corresponds to *not*: s is in E^c if and only if s is in S but is not in E.

The whole sample space S is the *sure* event. It corresponds to *true*: s is in S is always true.

The empty set is the *impossible* event. It corresponds to *false*: s is in $\emptyset$ is always false.

Two events E, F are *exclusive* if $E \cap F = \emptyset$. This means that s in E and s in F is always false.

2.2 Probability

The probability rules correspond to the logical operations. A probability assignment P assigns to each event E a number $P[E]$ with $0 \leq P[E] \leq 1$.

The *or* rule: If E, F are *exclusive*, then

$$P[E \cup F] = P[E] + P[F]. \tag{2.1}$$

The *and* rule: If E, F are *independent*, then

$$P[E \cap F] = P[E]P[F]. \tag{2.2}$$

The *not* rule:

$$P[E^c] = 1 - P[E]. \tag{2.3}$$

The *sure* rule:

$$P[S] = 1. \tag{2.4}$$

The *impossible* rule:

$$P[\emptyset] = 0. \tag{2.5}$$

Note: All of these rules may be taken as axioms of probability except for the *and* rule. This is exceptional, since the concept of independent events is not defined in terms of set operations. In fact, the *and* rule is to be thought of as the definition of independence in terms of probability.

2.3 Countable additivity

In probability the *or* rule is extended to an infinite sequence of exclusive events. This is the *some* rule. Let $E_1, E_2, E_3, \ldots$ be an infinite sequence of events. The event $\bigcup_{i=1}^{\infty} E_i$ is the probability that some E_i occurs.

The *some* rule is that if $E_1, E_2, E_3, \ldots$ is a sequence of events such that $E_i \cap E_j = \emptyset$ for all $i \neq j$, then

$$P[\bigcup_{i=1}^{\infty} E_i] = \sum_{i=1}^{\infty} P[E_i]. \tag{2.6}$$

What is the reason for this rule? Take the example of flipping a fair coin infinitely many times. This is an example when the sample space has infinitely many points. Each point is an entire infinite sequence of heads and tails. There are a huge number of such sequences.

Let E_i be the event that the first head occurs on the ith trial. Then $P[E_i] = 1/2^i$. It follows from the *some* rule that the probability that a head occurs on some trial is $\sum_{i=1}^{\infty} 1/2^i = 1$. This seems reasonable. The probability of getting a head on some trial is the negation of the event of never getting a head. Therefore the probability of never getting a head is zero. The event of getting a head on some trial has probability one, yet it is only probabilistically certain, not logically certain. Therefore such an event with probability one is said to be *almost sure*.

2.4 Discrete uniform probabilities

One elementary situation is when the number of points in the sample space S is finite. It may be a large number, but it is not infinite. In this case, the *uniform* probabilities are defined as follows:

$$P[E] = \frac{\#E}{\#S}. \tag{2.7}$$

Here $\#E$ is the number of elements in the set E, while $\#S$ is the number of points in the set S. The uniform probabilities satisfy all the probability axioms.

We shall see in the next sections that this notion has a number of important and useful special cases.

- The sample space consists of all functions from a finite set to another finite set.
- The sample space consists of all injective functions from a finite set to another finite set.
- The sample space consists of all subsets of a finite set, such that each subset has a certain specified number of elements. (This is the same as all functions from the set to a two-point set, such that each function has certain specified occupation numbers.)
- The sample space consists of all functions from a finite set to another finite set, such that each function has certain specified occupation numbers.

2.5 Binomial probabilities

In this section we consider an n element set I of trials and an m element set M representing a population. The fundamental assumption is that every ordered sample with replacement has the same probability. In other words, this is the assumption of uniform probability on the set of all functions from I to M.

The number of ordered samples with replacement is m^n. Therefore each ordered sample with replacement has probability

$$P[f] = \frac{1}{m^n}. \tag{2.8}$$

Let A be a subset of M representing the successes. Suppose that A has a elements.

Theorem 2.1 *The probability that an ordered sample with replacement has exactly k successes is*

$$P(k) = \binom{n}{k}\frac{a^k(m-a)^{n-k}}{m^n} = \binom{n}{k}p^k(1-p)^{n-k}. \tag{2.9}$$

where $p = a/m$.

Proof: Let K be a subset of the set of trials I with k elements. The probability $P(K)$ that an ordered sample with replacement has successes that exactly belong to the set K is the number of such samples time the probability of each sample. The number of such samples is the number of functions from K to A

times the number of functions from the complement of K to the complement of A, that is, a^k times $(m-a)^{n-k}$. Therefore the probability is

$$P(K) = a^k (m-a)^{n-k} \frac{1}{m^n}. \tag{2.10}$$

There are $\binom{n}{k}$ such subsets K of I with k elements. The probability $P(k)$ is thus $\binom{n}{k}$ times $P(K)$.

The probabilities for ordered samples with replacement are examples of what are known as *binomial* probabilities. The number n is the number of trials, and the number $p = a/m$ is the probability of success on each trial.

It is also possible to consider unordered samples with replacement (occupation numbers). However these do not all have the same probability. The probability of a particular occupation number function χ that sums to n is

$$P(\chi) = \binom{n}{\chi} \frac{1}{m^n}. \tag{2.11}$$

This is called the *multinomial* probability for n trials in the case when the probability of each of the m elements being chosen on one trial is the same $1/m$.

The following result shows that if we use this probability formula we get another derivation of the result given above. This is nothing new, but it is of some interest to see that one gets consistent results.

Theorem 2.2 *Consider that each ordered sample with replacement has the same probability. The probability that an unordered sample with replacement has exactly k successes is*

$$P(k) = \binom{n}{k} a^k (m-a)^{n-k} \frac{1}{m^n}. \tag{2.12}$$

This is of course just a repetition of the formula for the case of ordered samples with replacement.

Proof: The probability $P(k)$ that a randomly chosen subset of M of size n has exactly k successes is the sum of the probabilities of all occupation numbers χ for the m element set such that the restriction σ of χ to the subset of a successes has sum k and the restriction τ of χ to the subset of $n-k$ failures has sum $n-k$. Thus this is

$$P(k) = \sum_\sigma \sum_\tau \binom{n}{\sigma\,\tau} \frac{1}{m^n}. \tag{2.13}$$

The multinomial coefficient for occupation number $\chi = \sigma, \tau$ is a binomial coefficient times the product of multinomial coefficients associated with σ and τ. Thus

$$P(k) = \binom{n}{k} \sum_\sigma \binom{k}{\sigma} \sum_\tau \binom{n-k}{\tau} \frac{1}{(m)_n} = \binom{n}{k} a^k (m-a)^{n-k} \frac{1}{m^n}. \tag{2.14}$$

Note: There are actually two possible choices of probability assignment. In physics this comes up when we think of I as a set of particles and M as a set of states. The one we use, in which the ordered samples with replacement each have the same probability, is called the Maxwell-Boltzmann probability distribution. The other possibility, in which the unordered samples with replacement each have the same probability, is called in physics the Bose-Einstein probability distribution. This possibility is appropriate in some cases when the particles are regarded as fundamentally indistinguishable. Particles for which the Bose-Einstein distribution is appropriate are called bosons.

2.6 Hypergeometric probabilities

In this section we consider an n element set I of trials and an m element set M representing a population. The fundamental assumption is that every ordered sample without replacement has the same probability. In other words, this is the assumption of uniform probability on the set of all injective functions from I to M.

The number of ordered samples without replacement is $(m)_n$. Therefore each ordered sample with replacement has probability

$$P[f] = \frac{1}{(m)_n}. \tag{2.15}$$

Let A be a subset of M representing the successes. Suppose that A has a elements.

Theorem 2.3 *The probability that an ordered sample without replacement has exactly k successes is*

$$P(k) = \binom{n}{k} \frac{(a)_k (m-a)_{n-k}}{(m)_n}. \tag{2.16}$$

Proof: Let K be a subset of the set of trials I with k elements. The probability $P(K)$ that an ordered sample with replacement has successes that exactly belong to the set K is the number of such samples time the probability of each sample. The number of such samples is the number of injective functions from K to A times the number of injective functions from the complement of K to the complement of A, that is, $(a)_k$ times $(m-a)_{n-k}$. Therefore the probability is

$$P(K) = (a)_k (m-a)_{n-k} \frac{1}{(m)_n}. \tag{2.17}$$

There are $\binom{n}{k}$ such subsets K of I with k elements. The probability $P(k)$ is thus $\binom{n}{k}$ times $P(K)$.

Since for each unordered sample without replacement (subset) there are exactly $n!$ ordered samples without replacement, the probability of each unordered sample without replacement is the same. That is, if J is an n element subset of

M representing the unordered sample without replacement, then the probability of J is

$$P(J) = n!\frac{1}{(m)_n} = \frac{1}{\binom{m}{n}}. \tag{2.18}$$

Each subset has the same probability. So we can perform the reasoning with unordered samples without replacement and get the same result. Thus it is justified to use the uniform probability on the set of all subsets of M that have n elements.

Theorem 2.4 *The probability that an unordered sample without replacement has exactly k successes is*

$$P(k) = \frac{\binom{a}{k}\binom{m-a}{n-k}}{\binom{m}{n}}. \tag{2.19}$$

This is exactly the same answer as for the case of ordered samples without replacement.

Proof: The probability $P(k)$ that a randomly chosen subset of M of size n has exactly k successes is the number of such subsets times the probability of each subset. The number of such subsets is the number of subsets of A with k elements times the number of subsets of the complement of A with $n-k$ elements, that is, $\binom{a}{k}$ times $\binom{m-a}{n-k}$. Therefore the probability is

$$P(K) = \binom{a}{k}\binom{m-a}{n-k}\frac{1}{\binom{m}{n}}. \tag{2.20}$$

The probabilities given in the two theorems about samples without replacement are known as the *hypergeometric* probabilities. It is confusing that there are two different formulas for the hypergeometric probabilities, but that is because there are two points of view (ordered and unordered) that give the same results for sampling without replacement.

Note: In physics one considers situations when I is a set of particles and M is a set of states. The probability assignment in which the unordered samples without replacement are used is called the Fermi-Dirac probability distribution. This is appropriate in cases when the particles are regarded as fundamentally indistinguishable and when they also obey the exclusion principle: no two particles can occupy the same state. Such particles are called fermions.

2.7 Divisions into subsets

Another problem is when that sample space consists of all divisions of a set I with n elements into m disjoint subsets. The subsets are labeled by another set M. The number of elements in the subset corresponding to each element of M is given by the value of a fixed occupation number function. Thus we are just counting the number of functions from I to M with fixed occupation numbers. For example, we can think of I as a set of $n = 52$ cards and M as a set of $m = 4$

players, each of whom gets 13 cards. So the occupation numbers are 13, 13, 13, 13.

There are alternative of thinking of this. We can think of taking the subsets in order and then selecting the elements of each subset in order. The number of ordered divisions into labeled subsets is

$$(n)_{n_1}(n-n_1)_{n_2}(n-n_1-n_2)_{n_3}\cdots(n-n_1-\cdots-n_{m-1})_{n_m} = n! \qquad (2.21)$$

This says that if you select in order all elements of I that go in the first subset, then those that go in the second subset, and so on, this is just the same as selecting all the elements of I in order. In the card example this is the (52)! ways of dealing the cards.

If we take the subsets in order, but then neglect the order within each subset, then we get the division into labeled subsets. The number of such divisions into labeled subsets is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n-n_1-\cdots-n_{m-1}}{n_m} = \binom{n}{n_1, n_2, \cdots, n_m}. \qquad (2.22)$$

This says that if you select the first subset, then the next subset, and so on, this generates the entire division. The number of such divisions is given by the multinomial coefficient. It is the number of functions from I to M with given occupation numbers, but thought of as ways of dividing up I. In the card example this is the $\binom{52}{13,13,13,13}$ ways choosing a subset of 13 cards for each player.

Consider the case when all the occupation numbers are equal. Then there is even a third way of thinking of this, where we forget about the labels. That is, we think of the n element set I and of m subsets, each with the given number n_1 of elements, so that $n = mn_1$. This neglects the order in which the sets are taken as well as neglecting the order within each subset. The number of such divisions into unlabeled subsets is then

$$\frac{1}{m!}\binom{n}{n_1, n_1, \cdots, n_1}. \qquad (2.23)$$

In the card example this is the $\frac{1}{4!}\binom{52}{13,13,13,13}$ ways making four piles of 13 cards without specifying an order for the piles.

Some problems can be solved with any one of the three points of view. For instance, take the problem of finding the probability that all the spades go to one of the players (the player not specified).

1. From the ordered point of view, this is 4(13)!(39)! divided by (52)!. This is because there are (13)! ways the spades can be dealt to a specified player and (39)! ways the remaining cards can be dealt to the other three players. There are 4 ways to specify the player.

2. From the division into labeled subsets point of view, this is $4\binom{39}{13,13,13}$ divided by $\binom{52}{13,13,13,13}$. This is because there is only one way to give the spades to a specified player, and there are $\binom{39}{13,13,13}$ ways the remaining cards may be

divided up among the other three players. There are 4 ways to specify the player.

2. From the division into unlabeled subsets point of view, this is $(1/3!)\binom{39}{13,13,13}$ divided by $(1/4!)\binom{52}{13,13,13,13}$. This is because to make four equal piles out of the 52, one of which consists of the spades, is the same as making three equal piles out of the 39 non-spades. There are $(1/3!)\binom{39}{13,13,13}$ ways of doing this.

Chapter 3

Discrete Random Variables

3.1 Mean

A discrete random variable X is a function from the sample space S that has a finite or countable infinite number of real numerical values. For a discrete random variable each event $X = x$ has a probability $P[X = x]$. This function of x is called the *probability mass function* of the random variable X.

The *mean* or *expectation* of X is

$$\mu_X = E[X] = \sum_x xP[X = x], \tag{3.1}$$

where the sum is over the values of X.

One special case of a discrete random variable is a random variable whose values are natural numbers. Sometimes for technical purposes the following theorem is useful. It expresses the expectation in terms of a sum of probabilities.

Theorem 3.1 *Let Y be a random variable whose values are natural numbers. Then*

$$E[Y] = \sum_{j=1}^{\infty} P[Y \geq j]. \tag{3.2}$$

Proof: We have

$$E[Y] = \sum_{k=1}^{\infty} kP[Y = k] = \sum_{k=1}^{\infty} \sum_{j=1}^{k} P[Y = k] = \sum_{j=1}^{\infty} \sum_{k=j}^{\infty} P[Y = k]. \tag{3.3}$$

On the other hand,

$$\sum_{k=j}^{\infty} P[Y = k] = P[Y \geq j]. \tag{3.4}$$

Note that if X is a discrete random variable, and g is a function defined for the values of X and with real values, then $g(X)$ is also a random variable.

Theorem 3.2 *The joint probability mass function of a random variable $Y = g(X)$ that is a function of a random variable X may be computed in terms of the joint probability mass function of X by*

$$P[g(X) = y] = \sum_{x:g(x)=y} P[X = x]. \tag{3.5}$$

By definition, the expectation of $g(X)$ is

$$E[g(X)] = \sum_y yP[g(X) = y], \tag{3.6}$$

where the sum is over all values $y = g(x)$ of the random variable $g(X)$. The following theorem is particularly important and convenient. If a random variable $Y = g(X)$ is expressed in terms of another random variable, then this theorem gives the expectation of Y in terms of the probability mass function of X.

Theorem 3.3 *The expectation of a function of a random variable X may also be expressed in terms of the probability mass function of X by*

$$E[g(X)] = \sum_x g(x)P[X = x]. \tag{3.7}$$

Proof:

$$E[g(X)] = \sum_y y \sum_{x:g(x)=y} P[X = x] = \sum_y \sum_{x:g(x)=y} g(x)P[X = x] = \sum_x g(x)P[X = x]. \tag{3.8}$$

3.2 Variance

The *variance* of X is

$$\sigma_X^2 = \mathrm{Var}(X) = E[(X - \mu_X)^2]. \tag{3.9}$$

The *standard deviation* of X is the square root of the variance.

By the theorem, the variance may be computed by

$$\sigma_X^2 = \sum_x (x - \mu_X)^2 P[X = x]. \tag{3.10}$$

It is thus a weighted mean of squared deviations from the mean of the original random variable. Sometimes this is called a “mean square”. Then the standard deviation is a “root mean square”.

The following result is useful in some computations.

Theorem 3.4 *The variance is given by the alternative formula*

$$\sigma_X^2 = E[X^2] - E[X]^2. \tag{3.11}$$

3.3 Joint probability mass functions

Consider discrete random variables X, Y. Their joint probabilities are the numbers $P[X = x, Y = y]$. Here we are using the comma to indicate *and* or intersection. Thus the event indicated by $X = x, Y = y$ is the event that $X = x$ and $Y = y$. The resulting function of two variables x, y is called the *joint probability mass function* of X, Y. In a similar way, we can talk about the joint probability mass function of three or more random variables.

Given the joint probability mass function of X, Y, it is easy to compute the probability mass function of X, and it is equally easy to compute the probability mass function of Y. The first formula is

$$P[X = x] = \sum_y P[X = x, Y = y], \tag{3.12}$$

where the sum is over all possible values of the discrete random variable Y. Similarly

$$P[Y = y] = \sum_x P[X = x, Y = y], \tag{3.13}$$

where the sum is over all possible values of the discrete random variable X.

Theorem 3.5 *The joint probability mass function of a random variable $Z = g(X, Y)$ that is a function of random variables X, Y may be computed in terms of the joint probability mass function of X, Y by*

$$P[g(X, Y) = z] = \sum_{x,y:g(x,y)=z} P[X = x, Y = y]. \tag{3.14}$$

Theorem 3.6 *The expectation of a random variable $Z = g(X, Y)$ that is a function of random variables X, Y may be computed in terms of the joint probability mass function of X, Y by*

$$E[g(X, Y)] = \sum_x \sum_y g(x, y) P[X = x, Y = y]. \tag{3.15}$$

Proof:

$$E[g(X, Y)] = \sum_z z P[g(X, Y) = z] = \sum_z z \sum_{x,y:g(x,y)=z} P[X = x, Y = y] \tag{3.16}$$

However this is equal to

$$\sum_z \sum_{x,y:g(x,y)=z} g(x, y) P[X = x, Y = y] = \sum_{x,y} g(x, y) P[X = x, Y = y]. \tag{3.17}$$

This theorem has a very important consequence for expectations. This is the property called *additivity*. It says that the expectation of a sum is the sum of the expectations. We state it for two random variables, but again the idea extends to the sum of three or more random variables.

Theorem 3.7 *If X and Y are random variables with expectations $E[X]$ and $E[Y]$, then they satisfy the addition property*

$$E[X+Y] = E[X] + E[Y]. \tag{3.18}$$

Important note: The theorem can also be stated using the other notation for mean in the form $\mu_{X+Y} = \mu_X + \mu_Y$.

Proof:

$$E[X+Y] = \sum_x \sum_y (x+y) P[X = x, Y = y]. \tag{3.19}$$

This is equal to

$$\sum_x \sum_y x P[X = x, Y = y] + \sum_x \sum_y y P[X = x, Y = y] = E[X] + E[Y]. \tag{3.20}$$

3.4 Uncorrelated random variables

The *covariance* of random variables X and Y is defined by

$$\mathrm{Cov}(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]. \tag{3.21}$$

In particular, the variance is given by

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{Cov}(X,X). \tag{3.22}$$

Theorem 3.8 *The variance of a sum is given in terms of the variances of the terms and the covariance by*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Cov}(X,Y) + \mathrm{Var}(Y). \tag{3.23}$$

The correlation of random variables X and Y is defined by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}. \tag{3.24}$$

Thus the result of the theorem may also be written in the form

$$\sigma_{X+Y}^2 = \sigma_X^2 + 2\rho(X,Y)\sigma_X\sigma_Y + \sigma_Y^2. \tag{3.25}$$

Random variables are said to be *uncorrelated* if their correlation is zero. This is the same as saying that their covariance is zero.

Theorem 3.9 *For uncorrelated random variables the variances add. That is, if $Cov(X,Y) = 0$, then*

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2. \tag{3.26}$$

3.5 Independent random variables

Consider discrete random variables X and Y. They are said to be *independent* if the events $X = x$, $Y = y$ are independent for all possible values x, y. That is, X, Y are independent if

$$P[X = x, Y = y] = P[X = x][P[Y = y] \tag{3.27}$$

for all possible values x, y of the two random variables.

There is of course an appropriate generalization of the notion of independence to more than two random variables. Again there is a multiplication rule, but this time with more than two factors.

Theorem 3.10 *If X and Y are independent random variables, and f and g are arbitrary functions, then the random variables $f(X)$ and $g(Y)$ satisfy the multiplication property*

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)]. \tag{3.28}$$

This theorem has an important consequence: Independent random variables are uncorrelated. It immediately follows that for independent random variables the variances add.

Theorem 3.11 *If X and Y are independent random variables, then X and Y are uncorrelated random variables.*

Proof: If X and Y are independent, then from the previous theorem

$$E[(X - \mu_X)(Y - \mu_Y)] = E[X - \mu_X]E[Y - \mu_Y] = 0 \cdot 0 = 0. \tag{3.29}$$

Chapter 4

The Bernoulli Process

4.1 Bernoulli random variables

A *Bernoulli random variable* X is a random variable with values 0 or 1. Consider a Bernoulli random variable with $P[X = 1] = p$ and $P[X = 0] = 1 - p$. Then

$$\mu_X = 1p + 0(1 - p) = p, \tag{4.1}$$

and

$$\sigma_X^2 = (1 - p)^2 p + (0 - p)^2 (1 - p) = p(1 - p). \tag{4.2}$$

A Bernoulli random variable is also called an *indicator function* (or sometimes a characteristic function). Let E be the event that $X = 1$. Then the random variable X indicates this event. Conversely, given an event E, there is a random variable X that has the value 1 on any outcome in E and the value 0 on any outcome in the complement of E. This constructs an indicator function for the event.

4.2 Successes and failures

Consider an infinite sequence $X_1, X_2, X_3, \ldots$ of independent Bernoulli random variables, each with $P[X_i = 1] = p$. This is called a *Bernoulli process.* It consists of an infinite sequence of independent trials, each of which can result in a success or a failure. The random variable X_i is 1 if there is a success on the ith trial, and it is 0 if there is a failure on the ith trial.

By independence, the probability of a particular sequence of successes and failures is given by a product. Thus, for instance, the probability

$$P[X_1 = 1, X_2 = 0, X_3 = 1, X_4 = 1, X_5 = 0] = p(1 - p)pp(1 - p) = p^3(1 - p)^2. \tag{4.3}$$

More generally, the probability of a particular sequence of k successes and $n - k$ failures is $p^k(1 - p)^{n-k}$.

4.3 Binomial random variables

Consider a Bernoulli process with probability of success p on each trial. Let $N_n = X_1 + \cdots + X_n$ be the number of successes in the first n trials. Then N_n is a *binomial random variable* with parameters n and p. Its distribution is given by the following theorem.

Theorem 4.1 *Let N_n be binomial random variable that counts the number of successes in n independent trials, with probability p of success on each trial. Then*

$$P[N_n = k] = \binom{n}{k} p^k (1-p)^{n-k}. \tag{4.4}$$

Proof: Each particular sequence of successes and failures with k successes and $n-k$ failures has probability $p^k(1-p)^{n-k}$. The event that there are a total of k successes results if the sequence of successes is one of the k element subsets of the set $\{1, 2, \ldots, n\}$ of trials. There are $\binom{n}{k}$ such subsets. Thus we add $p^k(1-p)^{n-k}$ this many times. This gives the formula of the theorem.

In the proof of several theorems we use the following handy combinatorial fact:

$$k\binom{n}{k} = n\binom{n-1}{k-1}. \tag{4.5}$$

This can be proved as follows. Consider a set with n elements. Consider all pairs consisting of a subset with k elements together with a particular element of the subset. The number of such pairs is obviously given by the left hand side. (Choose the subset, and then choose the element in it.) The number of such pairs is also obviously given by the right hand side. (Choose the element, and then choose a subset not containing the element to give the remaining elements.) This shows that the left hand side is equal to the right hand side.

Theorem 4.2 *The expectation of the binomial random variable N_n is*

$$E[N_n] = np. \tag{4.6}$$

Proof: Use the combinatorial fact:

$$E[N_n] = \sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} n\binom{n-1}{k-1} p^k (1-p)^{n-k} = np\sum_{k=1}^{n}\binom{n-1}{k-1} p^{k-1}(1-p)^{n-k} = np. \tag{4.7}$$

The last sum is just the probability that N_{n-1} has some value $0, 1, \ldots, n-1$, which is one.

Alternative proof: Use the fact that the mean of a Bernoulli random variable is p and the fact that the mean of a sum is the sum of the means.

Theorem 4.3 *The variance of the binomial random variable N_n is*

$$\sigma_{N_n}^2 = np(1-p). \tag{4.8}$$

Proof: Use the combinatorial fact twice:

$$E[(N_n-1)N_n] = \sum_{k=0}^{n}(k-1)k\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=2}^{n}(n-1)n\binom{n-2}{k-2}p^k(1-p)^{n-k}. \tag{4.9}$$

Factor to obtain

$$(n-1)np^2\sum_{k=2}^{n}\binom{n-2}{k-2}p^{k-2}(1-p)^{n-k} = (n-1)np^2. \tag{4.10}$$

The sum is just the probability that N_{n-2} has some value $0, 1, \ldots, n-2$, which is one. Thus

$$\sigma^2_{N_n} = E[N_n^2]-E[N_n]^2 = E[N_n^2]-E[N_n]+E[N_n]-E[N_n]^2 = n(n-1)p^2+np-n^2p^2 = np-np^2. \tag{4.11}$$

Alternative proof: Use the fact that the variance of a Bernoulli random variable is $p(1-p)$ and the fact that the variance of a sum of independent random variables is the sum of the variances.

These formulas for the mean and variance are very important. The fact that the mean is np is very intuitive. The result for the variance is more subtle. It is perhaps better to look at the standard deviation. This is

$$\sigma_{N_n} = \sqrt{n}\sqrt{p(1-p)}. \tag{4.12}$$

Note that the factor $\sqrt{p(1-p)} \leq 1/2$ for all p with $0 \leq p \leq 1$. Thus the standard deviation is always bounded above by $(1/2)\sqrt{n}$.

When $0 < p < 1$, the standard deviation is greater than zero. In fact, when $0.1 \leq p \leq 0.9$, the standard deviation factor $\sqrt{p(1-p)}$ satisfies $0.3 \leq \sqrt{p(1-p)} \leq 0.5$. Thus in many circumstance this number is reasonably well limited, even if we do not have a precise knowledge of p.

Now consider the case when $0 < p < 1$ and n is large. Then the standard deviation, which is proportional to $\sqrt{n}$, is considerably less than the mean, which is proportional to n. This means that for large n the random variable N_n is peaked near the mean, at least in a relative sense.

4.4 The Poisson approximation

Something quite different happens when n is very large but at the same time p is very small. The binomial probability may be written

$$\binom{n}{k}p^k(1-p)^{n-k} = \frac{(n)_k}{k!}p^k(1-p)^{n-k} = \frac{(np)^k}{k!}\frac{(n)_k}{n^k}(1-p)^{n-k}. \tag{4.13}$$

If we set $\lambda = np$ we can write this as

$$\frac{\lambda^k}{k!}\frac{(n)_k}{n^k}(1-\frac{\lambda}{n})^n(1-\frac{\lambda}{n})^{-k}. \tag{4.14}$$

Now take the limit as $n \to \infty$ and $p \to 0$ with $\lambda = np$ fixed. The first factor does not depend on n. The second factor goes to one. The third factor has limit $e^{-\lambda}$. The fourth factor goes to one. So the result is

$$P[N = k] = \frac{\lambda^k}{k!} e^{-\lambda}. \tag{4.15}$$

This is the *Poisson* probability formula.

Here is some more detail on the limits. The quotient $(n)_k$ divided by n^k may be computed by dividing each of the k factors in $(n)_k$ by n. Then the limit of each of the resulting k factors is one. So the limit of the product is one. This is reasonable. The $(n)_k$ is the number of ordered samples without replacement of size k from a population of size n. The n^k is the number of ordered samples with replacement of size k from a population of size n. It is reasonable that for a very large population the two methods of sampling give equivalent results.

The limit of $(1 - \lambda/n)^n$ may be computed by using properties of logarithms. The Taylor series of $1/(1 - x) = 1 + x + x^2 + x^3 + \cdots$, the geometric series. Therefore, by integrating, the Taylor series of the logarithm is $-\ln(1 - x) = x+x^2/2+x^3/3+x^4/4+\cdots$, the integral of the geometric series. The logarithm of $(1-\lambda/n)^n$ is $n \ln(1-\lambda/n)$. Expand the logarithm in a Taylor series. This gives $-n[\lambda/n + (1/2)\lambda^2/n^2 + \cdots]$. Multiply out. This gives $-[\lambda + (1/2)\lambda^2/n + \cdots]$. This shows that as $n \to \infty$ the limit of the logarithm of $(1 - \lambda/n)^n$ is $-\lambda$. Since the exponential is the inverse function of the logarithm, the limit as $n \to \infty$ of $(1 - \lambda/n)^n$ is $e^{-\lambda}$.

Note that in this chapter the parameter λ is dimensionless and has the interpretation of the mean of the Poisson probability. The variance is also λ. Therefore the standard deviation is $\sqrt{\lambda}$. When λ is large, then the standard deviation is small relative relative to the mean. Thus for large λ the Poisson distribution is peaked around the mean, at least in a relative sense.

4.5 Geometric random variables

Consider a Bernoulli process with parameter $p > 0$. Let W be the trial on which the first success occurs. Then the probability of the event that $W = n$ is the probability of a pattern of $n - 1$ failures followed by a success. This kind of random variable is called *geometric*. This proves the following theorem.

Theorem 4.4 *For a geometric random variable W with parameter p the distribution is given for $n = 1, 2, 3, \ldots$ by*

$$P[W = n] = P[N_{n-1} = 0]P[X_n = 1] = (1 - p)^{n-1}p. \tag{4.16}$$

Theorem 4.5 *The expectation of the geometric random variable is*

$$E[W] = \frac{1}{p}. \tag{4.17}$$

Proof: This is the series

$$E[W] = \sum_{n=1}^{\infty} nP[W = n] = \frac{1}{p} \sum_{n=1}^{\infty} n(1-p)^{n-1}p^2 = \frac{1}{p}. \tag{4.18}$$

This is because $n(1-p)^{n-1}p^2$ is the probability that the second success occurs on the $n+1$st trial. If we grant that there is eventually going to be a second success, then these probabilities should add to one.

Theorem 4.6 *The variance of the geometric random variable is*

$$\sigma_W^2 = \frac{1}{p^2}(1-p). \tag{4.19}$$

Proof: Consider the series

$$E[W(W+1)] = \sum_{n=1}^{\infty} n(n+1)P[W = n] = \frac{2}{p^2} \sum_{n=1}^{\infty} \frac{n(n+1)}{2} (1-p)^{n-1}p^3 = \frac{2}{p^2}. \tag{4.20}$$

This is because $n(n+1)/2\,(1-p)^{n-1}p^3$ is the probability that the third success occurs on the $n+2$nd trial. If we grant that there is eventually going to be a third success, then these probabilities should add to one. Thus

$$\sigma_W^2 = E[W^2] - E[W]^2 = E[W^2] + E[W] - E[W] - E[W]^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1}{p^2}(1-p). \tag{4.21}$$

These formula for the mean and expectation of the geometric waiting time are quite important. The formula for the mean is $1/p$, which is a quite intuitive result. The formula for the variance is also important. It is better to look at the standard deviation. This is $1/p\sqrt{1-p}$. If p is well below one, so that one has to wait a long time for a success, then the standard deviation is almost as large as the mean. This means that one not only has to wait a long time for a success, but the variability of the length of the wait is also quite long. Maybe it will be short; maybe it will be several times the mean.

4.6 Negative binomial random variables

Consider a Bernoulli process with parameter $p > 0$. Let $T_r = W_1 + \cdots + W_r$ be the trial on which the rth success occurs. Then the probability of the event that $W = n$ is the probability of a pattern of $r-1$ successes and $n-r$ failures on the first $n-1$ trials, followed by a success. This kind of random variable is called *negative binomial*. This proves the following theorem.

Theorem 4.7 *For a negative binomial random variable T_r with parameters r and p the distribution is given for $n = r, r+1, r+2, \ldots$ by*

$$P[T_r = n] = P[N_{n-1} = r-1]P[X_n = 1] = \binom{n-1}{r-1} p^r (1-p)^{n-r}. \tag{4.22}$$

Theorem 4.8 *The expectation of the negative binomial random variable is*

$$E[W] = r\frac{1}{p}. \tag{4.23}$$

Proof: This is the series

$$E[T_r] = \sum_{n=r}^{\infty} nP[T_r = n] = \frac{r}{p}\sum_{n=r}^{\infty}\binom{n}{r}(1-p)^{n-r}p^{r+1} = \frac{1}{p}. \tag{4.24}$$

This is because $P[T_{r+1} = n+1] = \binom{n}{r}(1-p)^{n-r}p^{r+1}$ is the probability that the $r+1$st success occurs on the $n+1$st trial. If we grant that there is eventually going to be a second success, then these probabilities should add to one.

Alternative proof: Use the fact that the mean of a geometric random variable is $1/p$ and the fact that the mean of a sum is the sum of the means.

Theorem 4.9 *The variance of the negative binomial random variable is*

$$\sigma_{T_r}^2 = r\frac{1}{p^2}(1-p). \tag{4.25}$$

Proof: Consider the series

$$E[T_r(T_r{+}1)] = \sum_{n=r}^{\infty} n(n{+}1)P[T_r = n] = r(r{+}1)\frac{1}{p^2}\sum_{n=r}^{\infty} n\binom{n+1}{r+1}(1{-}p)^{n-r}p^{r+2} = \frac{r(r+1)}{p^2}. \tag{4.26}$$

This is because $P[T_{n+2} = n+1] = \binom{n+1}{r+1}(1-p)^{n-r}p^{r+2}$ is the probability that the $r{+}2$nd success occurs on the $n{+}2$nd trial. If we grant that there is eventually going to be a $r+2$nd success, then these probabilities should add to one. Thus

$$\sigma_{T_r}^2 = E[T_r^2]{-}E[T_r]^2 = E[T_r^2]{+}E[T_r]{-}E[T_r]{-}E[T_r]^2 = \frac{r(r+1)}{p^2}{-}\frac{r}{p}{-}\frac{r^2}{p^2} = \frac{r}{p^2}(1{-}p). \tag{4.27}$$

Alternative proof: Use the fact that the variance of a geometric random variable is $1/p^2(1-p)$ and the fact that the variance of a sum of independent random variables is the sum of the variances.

These formula for the mean and expectation of the negative binomial waiting time are quite important. When $r = 1$ this is the geometric waiting time, so we already know the facts. So consider the case when we are waiting for the rth success and r is large. The formula for the mean is $r\,1/p$, which is a quite intuitive result. The formula for the variance is also important. It is better to look at the standard deviation. This is $\sqrt{r}\,1/p\sqrt{1-p}$. Note that the standard deviation is considerably smaller than the mean. This means that the distribution is somewhat peaked about the mean, at least in a relative sense. Very small values are unlikely, so are very large values (relative to the mean). The reason for this behavior is that the waiting time for the rth success is the sum of r independent geometric waiting times. These are individually quite variable. But the short geometric waiting times and the long geometric waiting times tend to cancel out each other. The waiting time for the rth success is comparatively stable.

4.7 Summary

Consider a sequence of independent success-failure trials, with probability p of success on each trial. The random variable N_n that counts the number of successes in the first n trials is a binomial random variable. (When $n = 1$ the random variable N_1 is a Bernoulli random variable.) The random variable T_r that gives the number of the trial on which the rth success occurs is a negative binomial random variable. (When $r = 1$ this is a geometric random variable.) The relation between N_n and T_r is that $N_n \geq r$ is the same event as the event $T_r \leq n$.

Chapter 5

Continuous Random Variables

5.1 Mean

A *continuous random variable* X with a *probability density function* f is a random variable such that for each interval B of real numbers

$$P[X \in B] = \int_B f(x)\,dx. \tag{5.1}$$

Such a continuous random variable has the property that for each real number x the event that $X = x$ has a probability zero. This does not contradict the countable additivity axiom of probability theory, since the set of real number cannot be arranged in a sequence.

Often the random variable X has units, such as seconds or meters. In such a case the probability density function f has values $f(x)$ that have inverse units: inverse seconds or inverse centimeters. This is so that the integral that defines the probability will be dimensionless. The inverse dimensions of $f(x)$ and the dimensions of dx cancel, and the final integral is a dimensionless probability.

The values of a probability density function are not probabilities. One goes from the probability density function to probability by integration:

$$P[X \leq b] = \int_{-\infty}^{b} f(x)\,dx. \tag{5.2}$$

In the other direction, one goes from probabilities to the probability density function by differentiation:

$$f(x) = \frac{d}{dx} P[X \leq x]. \tag{5.3}$$

In spite of this somewhat subtle relation, most of our thinking about continuous random variables involves their probability density functions.

The *mean* or *expectation* of X is

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f(x)\, dx. \tag{5.4}$$

Sometimes for technical purposes the following theorem is useful. It expresses the expectation in terms of an integral of probabilities. It is most often used for random variables Y that have only positive values; in that case the second term is of course zero.

Theorem 5.1 *Let Y be a continuous random variable with probability density function f. Then*

$$E[Y] = \int_0^{\infty} P[Y > y]\, dy - \int_{-\infty}^{0} P[Y < y]\, dy. \tag{5.5}$$

Proof: We have

$$E[Y] = \int_0^{\infty} x f(x)\, dx + \int_{-\infty}^{0} x f(x)\, dx = \int_0^{\infty} \int_0^{x} dy f(x)\, dx - \int_{-\infty}^{0} \int_x^{0} dy f(x)\, dx. \tag{5.6}$$

Interchange the order of integration. This gives

$$E[Y] = \int_0^{\infty} \int_y^{\infty} f(x)\, dx\, dy - \int_{-\infty}^{0} \int_{-\infty}^{y} f(x)\, dx\, dy. \tag{5.7}$$

This immediately gives the result of the theorem.

Note that if X is a continuous random variable and g is a function defined for the values of X and with real values, then $Y = g(X)$ is also a random variable. The following theorem is particularly important and convenient. If a random variable $Y = g(X)$ is expressed in terms of a continuous random variable, then this theorem gives the expectation of Y in terms of probabilities associated to X.

Theorem 5.2 *The expectation of a function $Y = g(X)$ of a continuous variable X may be expressed in terms of the probability density function f associated with X by*

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx. \tag{5.8}$$

Proof:

$$E[g(X)] = \int_0^{\infty} P[g(X) > y]\, dy - \int_{-\infty}^{0} P[g(X) < y]\, dy. \tag{5.9}$$

Write this in terms of the density. This gives

$$E[g(X)] = \int_0^{\infty} \int_{g(x)>y} f_X(x)\, dx\, dy - \int_{-\infty}^{0} \int_{g(x)<y} f_X(x)\, dx\, dy. \tag{5.10}$$

Now we can interchange the order of integration. This gives

$$E[g(X)] = \int_{g(x)>0} \int_0^{g(x)} dy f_X(x)\,dx - \int_{g(x)<0} \int_{g(x)}^0 dy f_X(x)\,dx. \tag{5.11}$$

This simplifies to

$$E[g(X)] = \int_{g(x)>0} g(x) f_X(x)\,dx + \int_{g(x)<0} g(x) f_X(x)\,dx = \int g(x) f_X(x)\,dx. \tag{5.12}$$

If $Y = g(X)$, then it is possible to compute the probability density function of Y in terms of the probability density function of X, but this is somewhat more complicated, because there is a derivative that must be taken into account.

Theorem 5.3 *Let $Y = g(X)$ be a function of a continuous random variable X, and assume that the derivative $g'(x)$ vanishes only at isolated points. Then the probability density function f_Y is expressed in terms of the probability density function f_X by*

$$f_Y(y) = \sum_{x:g(x)=y} f_X(x) \frac{1}{|g'(x)|}. \tag{5.13}$$

Note: If we think of $y = f(x)$, then this may be written in the form

$$f_Y(y) = \sum_{x:g(x)=y} f_X(x) \left| \frac{dx}{dy} \right|. \tag{5.14}$$

Proof: We have

$$f_Y(y) = \frac{d}{dy} P[g(X) \le y] = \frac{d}{dy} \int_{x:g(x)\le y} f_X(x)\,dx. \tag{5.15}$$

For each y the set of x such that $g(x) \le y$ is a union of intervals. Call a typical left hand point of such an interval $l(y)$ and a typical right hand point of the interval $r(y)$. Then $g(l(y)) = y$ and $g(r(y) = y$. The derivative of the integral associated with such an interval is

$$\frac{d}{dy} \int_{l(y)}^{r(y)} f_X(x)\,dx = f_X(r(y)) \frac{dr(y)}{dy} - f_X(l(y)) \frac{dl(y)}{dy}. \tag{5.16}$$

However $g'(l(y)) dl(y)/dy = 1$ and $g'(r(y)) dr(y)/dy = 1$. So we can write the result as

$$f_Y(y) = \sum_{x:g(x)=y, g'(x)>0} f_X(x) \frac{1}{g'(x)} - \sum_{x:g(x)=y, g'(x)<0} f_X(x) \frac{1}{g'(x)}. \tag{5.17}$$

This is equivalent to the statement in the theorem.

5.2 Variance

The *variance* of X is

$$\sigma_X^2 = \mathrm{Var}(X) = E[(X - \mu_X)^2]. \tag{5.18}$$

The *standard deviation* of X is the square root of the variance.

By the theorem, the variance may be computed by

$$\sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)\, dx. \tag{5.19}$$

It is thus a weighted mean of squared deviations from the mean of the original random variable. Sometimes this is called a "mean square". Then the standard deviation is a "root mean square".

The following result is useful in some computations.

Theorem 5.4 *The variance is given by the alternative formula*

$$\sigma_X^2 = E[X^2] - E[X]^2. \tag{5.20}$$

5.3 Joint densities

Consider continuous random variables X and Y with densities $f_X(x)$ and $f_Y(y)$. They have a *joint probability density function* $f_{X,Y}$ if for every region in the plane the probability

$$P[(X,Y) \in C] = \int \int_C f_{X,Y}(x,y)\, dx\, dy. \tag{5.21}$$

In a similar way, we can talk about the joint probability density function of three or more random variables.

Given the joint probability density function of X, Y, it is easy to compute the the probability density function of X, and it is equally easy to compute the probability density function of Y. The first formula is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dy. \tag{5.22}$$

Similarly

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx. \tag{5.23}$$

Theorem 5.5 *The expectation of a random variable $Z = g(X,Y)$ that is a function of random variables X, Y may be computed in terms of the joint probability density function of X, Y by*

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)\, dx\, dy. \tag{5.24}$$

Proof:

$$E[g(X,Y)] = \int_0^\infty P[g(X,Y) > z]\,dz - \int_{-\infty}^0 P[g(X,Y) < z]\,dz. \tag{5.25}$$

Writing this in terms of the joint density gives

$$E[g(X)] = \int_0^\infty \int\int_{g(x,y)>z} f(x,y)\,dx\,dy\,dz - \int_{-\infty}^0 \int\int_{g(x,y)<z} f(x,y)\,dx\,dy\,dz. \tag{5.26}$$

Now we can interchange the order of integration. This gives

$$E[g(X)] = \int\int_{g(x,y)>0} \int_0^{g(x,y)} dz f(x,y)\,dx\,dy - \int\int_{g(x,y)<0} \int_{g(x,y)}^0 dz f(x,y)\,dx\,dy. \tag{5.27}$$

This simplifies to

$$E[g(X)] = \int\int_{g(x,y)>0} g(x,y)f(x,y)\,dx\,dy + \int\int_{g(x,y)<0} g(x,y)f(x,y)\,dx\,dy = \int\int g(x,y)f(x,y)\,dx\,dy. \tag{5.28}$$

This theorem has a very important consequence for expectations. This is the property called *additivity.* It says that the expectation of a sum is the sum of the expectations. We state it for two random variables, but again the idea extends to the sum of three or more random variables.

Theorem 5.6 *If X and Y are random variables with expectations $E[X]$ and $E[Y]$, then they satisfy the additivity property*

$$E[X+Y] = E[X] + E[Y]. \tag{5.29}$$

Important note: The theorem can also be stated in the form $\mu_{X+Y} = \mu_X + \mu_Y$.

Proof:

$$E[X+Y] = \int_{-\infty}^\infty \int_{-\infty}^\infty (x+y)f(x,y)\,dx\,dy = \int_{-\infty}^\infty \int_{-\infty}^\infty xf(x,y)\,dx\,dy + \int_{-\infty}^\infty \int_{-\infty}^\infty yf(x,y)\,dx\,dy = E[X]+E[Y]. \tag{5.30}$$

5.4 Uncorrelated random variables

The *covariance* of random variables X and Y is defined by

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]. \tag{5.31}$$

In particular, the variance is given by

$$\sigma_X^2 = \mathrm{Var}(X) = \mathrm{Cov}(X,X). \tag{5.32}$$

Theorem 5.7 *The variance of a sum is given in terms of the variances of the terms and the covariance by*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Cov}(X,Y) + \mathrm{Var}(Y). \tag{5.33}$$

The correlation of random variables X and Y is defined by

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}. \tag{5.34}$$

Thus the result of the theorem may also be written in the form

$$\sigma^2_{X+Y} = \sigma^2_X + 2\rho(X,Y)\sigma_X\sigma_Y + \sigma^2_Y. \tag{5.35}$$

Random variables are said to be *uncorrelated* if their correlation is zero. This is the same as saying that their covariance is zero.

Theorem 5.8 *For uncorrelated random variables the variances add. That is, if $Cov(X,Y) = 0$, then*

$$\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y. \tag{5.36}$$

5.5 Independent random variables

Consider continuous random variables X and Y. They are said to be *independent* if their joint probability density functions factors as

$$f_{X,Y}(x,y) = f_X(x) f_Y(y). \tag{5.37}$$

There is of course an appropriate generalization of the notion of independence to more than two random variables. Again there is a multiplication rule, but this time with more than two factors.

Theorem 5.9 *If X and Y are independent random variables, and f and g are arbitrary functions, then the random variables $f(X)$ and $g(Y)$ satisfy the multiplicativity property then*

$$E[f(X)g(Y)] = E[f(X)]E[g(Y)]. \tag{5.38}$$

This theorem has an important consequence: Independent random variables are uncorrelated. It immediately follows that for independent random variables the variances add.

Theorem 5.10 *If X and Y are independent random variables, then X and Y are uncorrelated random variables.*

Proof: If X and Y are independent, then from the previous theorem

$$E[(X-\mu_X)(Y-\mu_Y)] = E[X-\mu_X]E[Y-\mu_Y] = 0 \cdot 0 = 0. \tag{5.39}$$

Chapter 6

The Poisson Process

6.1 Linear differential equations

In the following we shall need the solution of the homogeneous linear differential equation

$$\frac{dy}{dt} = -\lambda y. \tag{6.1}$$

The solution is given by doing the integrals

$$\int \frac{1}{y}\, dy = -\lambda \int dt. \tag{6.2}$$

The solution is

$$\ln(|y|) = -\lambda t + A, \tag{6.3}$$

or, setting $C = \pm e^A$,

$$y = Ce^{-\lambda t}. \tag{6.4}$$

Clearly $y = C$ when $t = 0$. This says that y decays exponentially with t at rate λ from the initial amount C. Summary: The solution of this equation represents *exponential decay* from the initial amount

We shall also need the solution of the inhomogeneous linear differential equation

$$\frac{dy}{dt} = -\lambda y + f(t). \tag{6.5}$$

The extra term $f(t)$ is called a *source* term. The method is to write the solution in the form

$$y = ue^{-\lambda t}. \tag{6.6}$$

If the $f(t)$ term were not present we could take u to be a constant, as above. However to get the solution with this source term present, we take u to be a function of t. This gives

$$\frac{du}{dt} e^{-\lambda t} = f(t). \tag{6.7}$$

The solution of this equation is

$$u = \int_0^t e^{\lambda s} f(s)\, ds + C. \tag{6.8}$$

The original y is thus

$$y = \int_0^t e^{-\lambda(t-s)} f(s)\, dx + Ce^{-\lambda t}. \tag{6.9}$$

Clearly $y = C$ when $t = 0$. This says that the solution at time t is given by the decay from the initial amount C over the entire time t, plus the integral of the decay from the source at time s over the remaining time interval of length $t-s$. Summary: The solution of this equation represents exponential decay from the initial amount plus exponential decay from the source from throughout its history.

6.2 Jumps

Imagine a device that emits a click sound from time to time. We consider a family of random variables $N(t)$ that count how many clicks have occurred between time zero and time t, measured in units of seconds. These clicks occur at an average rate λ in units of inverse seconds. In this chapter it will be λt that is the dimensionless number that represents the mean of $N(t)$.

As t increases the number $N(t)$ occasionally jumps by one. The jump occurs exactly at the time of the click, and the value of $N(t)$ at that time is the total number of clicks, including the one that took place at that time.

The idea is that the chance of a jump in any tiny interval of time of length dt is $\lambda\, dt$ and is independent of the number of previous jumps. This gives the fundamental differential equations. The first is

$$P[N(t + dt) = 0] = (1 - \lambda\, dt)P[N(t) = 0]. \tag{6.10}$$

This says that probability that $N(t + dt) = 0$ is the probability that $N(t) = 0$ and there is no jump between t and $t + dt$. The others for $k \geq 1$ are

$$P[N(t + dt) = k] = (1 - \lambda\, dt)\, P[N(t) = k] + \lambda\, dtP[N(t) = k - 1]. \tag{6.11}$$

They say that the probability that $N(t+dt) = k$ is the probability that $N(t) = k$ and there is no jump between t and $t+dt$ plus the probability that $N(t) = k-1$ and there is a jump between t and $t + dt$.

These equations can (and should) be written as differential equations

$$\frac{dP[N(t) = 0]}{dt} = -\lambda P[N(t) = 0] \tag{6.12}$$

with initial condition $P[N(0) = 0] = 1$ and

$$\frac{dP[N(t) = k]}{dt} = -\lambda P[N(t) = k] + \lambda P[N(t) = k - 1] \tag{6.13}$$

with initial condition $P[N(0) = k] = 0$ for $k \geq 1$.

6.3 Poisson random variables

The first equation is homogeneous and has the solution

$$P[N(t) = 0] = e^{-\lambda t}. \tag{6.14}$$

The second equation is inhomogeneous with source term $\lambda P[N(t) = k - 1]$ and has the solution

$$P[N(t) = k] = \int_0^t e^{-\lambda(t-s)} \lambda P[N(s) = k - 1]\, ds. \tag{6.15}$$

These equations may be solved to find $P[N(t) = k]$.

Theorem 6.1 *The solution of the differential equations is*

$$P[N(t) = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t}. \tag{6.16}$$

The result is that $N(t)$ is a Poisson *random variable with parameter λt.*

Proof: This is proved by noting that it is true for $k = 0$. Furthermore, we shall see that for each $k \geq 1$, if it is true for $k - 1$ it is also true for k. Therefore it must be true for all k.

To show that if it is true for $k - 1$ it is also true for k, suppose that

$$P[N(s) = k - 1] = \frac{(\lambda s)^{k-1}}{(k-1)!} e^{-\lambda s} \tag{6.17}$$

for all $s \geq 0$. Then from the solution of the differential equation

$$P[N(t) = k] = \int_0^t e^{-\lambda(t-s)} \lambda \frac{(\lambda s)^{k-1}}{(k-1)!} e^{-\lambda s}\, ds. \tag{6.18}$$

This can be written

$$P[N(t) = k] = \frac{\lambda^k}{(k-1)!} e^{-\lambda t} \int_0^t s^{k-1}\, ds = \frac{\lambda^k}{(k-1)!} e^{-\lambda t} \frac{t^k}{k}. \tag{6.19}$$

This immediately gives the result that

$$P[N(t) = k] = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \tag{6.20}$$

for all $t \geq 0$.

Theorem 6.2 *The expectation of the Poisson random variable $N(t)$ is*

$$E[N(t)] = \lambda t. \tag{6.21}$$

Proof: Compute

$$E[N(t)] = \sum_{k=0}^{\infty} kP[N(t) = k] = \lambda t \sum_{n=1}^{\infty} \frac{\lambda t^{k-1}}{(k-1)!} e^{-\lambda t} = \lambda t. \tag{6.22}$$

The last equation uses the fact that the probabilities of the events $N(t) = k - 1$ for $k = 1, 2, 3, \ldots$ add to one.

Theorem 6.3 *The variance of the Poisson random variable $N(t)$ is*

$$\sigma^2_{N(t)} = \lambda t. \tag{6.23}$$

Proof: Compute

$$E[((N(t) - 1)N(t)] = \sum_{k=0}^{\infty}(k-1)kP[N(t) = k] = (\lambda t)^2 \sum_{n=2}^{\infty} \frac{\lambda t^{k-2}}{(k-2)!} e^{-\lambda t} = \lambda t. \tag{6.24}$$

The last equation uses the fact that the probabilities of the events $N(t) = k - 2$ for $k = 2, 3, 4, \ldots$ add to one. Thus

$$\sigma^2_{N(t)} = E[N(t)^2] - E[N(t)]^2 = E[N(t)^2] - E[N(t)] + E[N(t)] - E[N(t)]^2 = (\lambda t)^2 - \lambda t - (\lambda t)^2 = \lambda t. \tag{6.25}$$

These formulas for the mean and variance are very important. The fact that the mean is λt is very intuitive. The result for the variance is more subtle. It is perhaps better to look at the standard deviation. This is

$$\sigma_{N(t)} = \sqrt{\lambda t}. \tag{6.26}$$

Now consider the case when λt is large. Then the standard deviation is considerably less than the mean. This means that for large time the random variable $N(t)$ is peaked near the mean, at least in a relative sense.

6.4 The gamma function

The *gamma function* is defined for all real numbers $t > 0$ by

$$\Gamma(t) = \int_0^{\infty} u^{t-1} e^{-u}\, du. \tag{6.27}$$

The important property of the Γ function is the identity

$$\Gamma(t+1) = t\Gamma(t). \tag{6.28}$$

This may be proved by integration by parts.

Theorem 6.4 *The value of* $\Gamma(m+1)$ *for natural numbers* $m = 0, 1, 2, 3, \ldots$ *is the factorial*

$$\Gamma(m+1) = \int_0^{\infty} u^m e^{-u}\, du = m!. \tag{6.29}$$

Such an integral is called a $\Gamma(m+1)$ integral. The result follows from the identity and the obvious fact that $\Gamma(1) = 1$.

It is worth recording that there is also an explicit formula for the Gamma function for half-integer values.

Theorem 6.5 *The value of* $\Gamma(m+\frac{1}{2})$ *for natural numbers* $m = 0, 1, 2, 3, \ldots$ *is given by*

$$\Gamma(m+\frac{1}{2}) = (m-\frac{1}{2})(m-\frac{3}{2})\cdots\frac{3}{2}\frac{1}{2}\sqrt{\pi}. \tag{6.30}$$

Such an integral is called a $\Gamma(m+\frac{1}{2})$ integral. The result follows from the identity and the non-obvious fact that $\Gamma(\frac{1}{2}) = \sqrt{\pi}$. The reason for this last fact is that

$$\Gamma(\frac{1}{2}) = \int_0^\infty t^{-\frac{1}{2}} e^{-t}\, dt = 2\int_0^\infty e^{-w^2}\, dw = \sqrt{\pi}. \tag{6.31}$$

The last equality comes from the normalization integral for a Gaussian (normal) density with variance parameter 1/2.

6.5 Exponential random variables

Consider a Poisson process. Let W be the waiting time until the first jump. This kind of random variable is called *exponential* with parameter λ. Then since the event $W \leq t$ is the same as the event $N(t) \geq 1$, we have the following result.

Theorem 6.6 *Let W be an exponential random variable with parameter λ. Then*

$$P[W \leq t] = 1 - e^{-\lambda t} \tag{6.32}$$

for $t \geq 0$.

The following theorem may be proved by differentiating. However we give another more direct proof below.

Theorem 6.7 *For an exponential random variable W with parameter λ the probability density is given by*

$$f(t) = \lambda e^{-\lambda t} \tag{6.33}$$

for $t \geq 0$.

Proof: If the waiting time for the first jump is in the interval from t to λdt, then there were no jumps up to time t, and then there was a jump in the following tiny interval. Thus

$$P[t < W < t + dt] = P[N(t) = 0]\lambda\, dt. \tag{6.34}$$

Theorem 6.8 *The expectation of an exponential random variable with parameter λ is*

$$E[W] = \frac{1}{\lambda}. \tag{6.35}$$

Proof: Compute the integral

$$E[W] = \int_0^\infty t\lambda e^{-\lambda t}\, dt. \tag{6.36}$$

After a change of variable it becomes a $\Gamma(2)$ integral.

Theorem 6.9 *The variance of the geometric random variable is*

$$\sigma_W^2 = \frac{1}{\lambda^2}. \tag{6.37}$$

Proof: Compute the integral

$$E[W^2] = \int_0^\infty t^2\lambda e^{-\lambda t}\, dt. \tag{6.38}$$

After a change of variable it becomes a $\Gamma(3)$ integral. The result is $2/\lambda^2$. It follows that $\mathrm{Var}(W) = E[W^2] - E[W]^2 = 1/\lambda^2$.

These formula for the mean and expectation of the exponential waiting time are quite important. The formula for the mean is $1/\lambda$, which is a quite intuitive result. The formula for the variance is also important. It is better to look at the standard deviation. This is also $1/\lambda$. The standard deviation is the same as the mean. The variability of the length of the wait is comparable to the average length of the wait. Maybe the jump will occur soon, maybe considerably later.

6.6 Gamma random variables

Consider a Poisson process with parameter λ. Let T_r be the time of the rth jump. We can write it as the sum of waiting times between jumps as $T_r = W_1 + W_2 + \cdots + W_r$. If $t < T_r \leq t + dt$, then up to t there were only $r-1$ jumps, and the last jump occurs in the tiny interval from t to dt. By independence, the probability is

$$P[t < T_r < t + dt] = P[N(t) = k-1]\lambda\, dt. \tag{6.39}$$

This kind of random variable is called *gamma* with parameters r and λ. This proves the following theorem.

Theorem 6.10 *For a gamma distribution with parameters r and λ the probability density is given by*

$$f(t) = \frac{(\lambda t)^{(r-1)}}{(r-1)!} e^{-\lambda t}\lambda. \tag{6.40}$$

Since this is a probability density function the integral from zero to infinity must be equal to one. This may be checked by a change of variable that reduces the integral to a $\Gamma(r)$ integral. This is the origin of the name.

Theorem 6.11 *The expectation of the gamma random variable is*

$$E[T_r] = r\frac{1}{\lambda}. \tag{6.41}$$

Proof: Compute the integral

$$E[T_r] = \int_0^\infty t\frac{(\lambda t)^{(r-1)}}{(r-1)!}e^{-\lambda t}\lambda\,dt. \tag{6.42}$$

After a change of variable it becomes a $\Gamma(r+1)$ integral.

Alternative proof: Use the formula for the mean of an exponential random variable and the fact that the mean of a sum is the sum of the means.

Theorem 6.12 *The variance of the gamma random variable is*

$$\sigma^2_{T_r} = r\frac{1}{\lambda^2}. \tag{6.43}$$

Proof: Compute the integral

$$E[T_r^2] = \int_0^\infty t^2\frac{(\lambda t)^{(r-1)}}{(r-1)!}e^{-\lambda t}\lambda\,dt. \tag{6.44}$$

After a change of variable it becomes a $\Gamma(r+2)$ integral. The result is $(r+1)r/\lambda^2$. It follows that $\mathrm{Var}(T_r) = E[T_r^2] - E[T_r]^2 = r/\lambda^2$.

Alternative proof: Use the formula for the variance of an exponential random variable and the fact that the variance of a sum of independent random variables is the sum of the variances.

These formula for the mean and expectation of the gamma waiting time are quite important. When $r = 1$ this is the exponential waiting time, so we already know the facts. So consider the case when we are waiting for the rth success and r is large. The formula for the mean is r/λ, which is a quite intuitive result. The formula for the variance is also important. It is better to look at the standard deviation. This is $\sqrt{r}/\lambda$. Note that the standard deviation is considerably smaller than the mean. This means that the distribution is somewhat peaked about the mean, at least in a relative sense. Very small values are unlikely, so are very large values (relative to the mean). The reason for this behavior is that the waiting time for the rth success is the sum of r independent exponential waiting times. These are individually quite variable. But the short exponential waiting times and the long exponential waiting times tend to cancel out each other. The waiting time for the rth success is comparatively stable.

6.7 Summary

Consider a Poisson process with rate λ. Thus the expected number of jumps in an interval of time of length t is λt. The random variable $N(t)$ that counts the number of jumps up to and including time t is a Poisson random variable. The random variable T_r that gives the time of the rth success is a gamma random variable. (When $r = 1$ this is an exponential random variable.) The relation between $N(t)$ and T_r is that $N(t) \geq r$ is the same event as the event $T_r \leq t$.

The Poisson process is a limiting situation for the Bernoulli process. The Bernoulli process takes place at discrete trials indexed by n, with probability p of success on each trial. Take a small time interval $\Delta t > 0$. Let $t = n\Delta t$ and $p = \lambda \Delta t$. Then the binomial random variable N_n is approximately Poisson with mean $np = \lambda t$, and the negative binomial random variable T_r when multiplied by Δt is approximately exponential with mean $(r/p)\Delta t = r/\lambda$.

Chapter 7

The weak law of large numbers

7.1 Independent implies uncorrelated

We begin by recalling the fact that independent random variables are uncorrelated random variables. We begin with the definitions in a particularly convenient form.

Random variables X, Y are *independent* if for every pair of functions f, g we have

$$E[f(X)g(X)] = E[f(X)]E[g(X)]. \tag{7.1}$$

Random variables X, Y are *uncorrelated* if

$$E[XY] = E[X]E[Y]. \tag{7.2}$$

This follows from the fact that the covariance of X, Y may be written in the form $\mathrm{Cov}(X, Y) = E[XY] - E[X]E[Y]$.

The following theorem comes from taking the functions $f(x) = x$ and $g(y) = y$ in the definition of independence.

Theorem 7.1 *If X and Y are independent, then X and Y are uncorrelated.*

The converse is false. Uncorrelated does not imply independent.

7.2 The sample mean

Theorem 7.2 *Let $X_1, X_2, X_3, \ldots, X_n$ be random variables, each with mean μ. Let*

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \tag{7.3}$$

be their sample mean. Then the expectation of $\bar{X}_n$ is

$$E[\bar{X}_n] = \mu. \tag{7.4}$$

Proof: The expectation of the sample mean $\bar{X}_n$ is

$$E[\bar{X}_n] = E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}E[\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} E[X_i] = \frac{1}{n}n\mu = \mu. \tag{7.5}$$

Theorem 7.3 *Let $X_1, X_2, X_3, \ldots, X_n$ be random variables, each with mean μ and standard deviation σ. Assume that each pair X_i, X_j of random variables with $i \neq j$ is uncorrelated. Let*

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \tag{7.6}$$

be their sample mean. Then the standard deviation of $\bar{X}_n$ is

$$\sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}. \tag{7.7}$$

Proof: The variance of the sample mean $\bar{X}_n$ is

$$\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\mathrm{Var}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n^2}n\sigma^2 = \frac{1}{n}\sigma^2. \tag{7.8}$$

We can think of these two results as a form of the weak law of large numbers. The law of large numbers is the "law of averages" that says that averaging uncorrelated random variable gives a result that is approximately constant. In this case the sample mean has expectation μ and standard deviation $\sigma/\sqrt{n}$. Thus if n is large enough, it is a random variable with expectation μ and with little variability.

The factor $1/\sqrt{n}$ is both the blessing and the curse of statistics. It is a wonderful fact, since it says that averaging reduces variability. The problem, of course, is that while $1/\sqrt{n}$ goes to zero as n gets larger, it does so rather slowly. So one must somehow obtain a quite large sample in order to ensure rather moderate variability.

The reason the law is called the weak law is that it gives a statement about a fixed large sample size n. There is another law called the strong law that gives a corresponding statement about what happens for all sample sizes n that are sufficiently large. Since in statistics one usually has a sample of a fixed size n and only looks at the sample mean for this n, it is the more elementary weak law that is relevant to most statistical situations.

Corollary 7.1 *Let $E_1, E_2, E_3, \ldots E_n$ be events, each with probability p. Let f_n be the proportion of events that happen. Then the expectation of f_n is*

$$E[f_n] = p. \tag{7.9}$$

Proof: Let $X_1, X_2, X_3, \ldots, X_n$ be the corresponding Bernoulli random variables, so that $X_i = 1$ for outcomes in E_i and $X_i = 0$ for outcomes in the complement of E_i. Then the expectation of X_i is p. The sample frequency f_n is just the sample mean of $X_1, \ldots, X_n$.

Corollary 7.2 *Let $E_1, E_2, E_3, \ldots E_n$ be events, each with probability p. Assume that each pair E_i, E_j of events with $i \neq j$ is independent. Let f_n be the proportion of events that happen. Then the standard deviation of f_n is*

$$\sigma_{f_n} = \frac{\sqrt{p(1-p)}}{\sqrt{n}}. \tag{7.10}$$

Proof: Let $X_1, X_2, X_3, \ldots, X_n$ be the corresponding Bernoulli random variables, so that $X_i = 1$ for outcomes in E_i and $X_i = 0$ for outcomes in the complement of E_i. Then the variance of X_i is $p(1-p)$. The sample frequency f_n is just the sample mean of $X_1, \ldots, X_n$. So its variance is obtained by dividing by n.

It is very important to note that this result gives a bound for the variance of the sample proportion that does not depend on the probability. In fact, it is clear that

$$\sigma_{f_n} \leq \frac{1}{2\sqrt{n}}. \tag{7.11}$$

This is an equality only when $p = 1/2$.

7.3 The Chebyshev inequality

Theorem 7.4 *If Y is a random variable with mean μ and variance σ, then*

$$P[|Y - \mu| \geq a] \leq \frac{\sigma^2}{a^2}. \tag{7.12}$$

Proof: Consider the random variable D that is equal to a^2 when $|Y - \mu| \geq a$ and is equal to 0 when $|Y - \mu| < a$. Clearly $D \leq (Y - \mu)^2$. It follows that the expectation of the discrete random variable D is less than or equal to the expectation of $(Y - \mu)^2$. This says that

$$a^2 P[|Y - \mu| \geq a] \leq E[(Y - \mu)^2] = \sigma^2. \tag{7.13}$$

Corollary 7.3 *Let $X_1, X_2, X_3, \ldots X_n$ be random variables, each with mean μ and standard deviation σ. Assume that each pair X_i, X_j of random variables with $i \neq j$ is uncorrelated. Let*

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \tag{7.14}$$

be their sample mean. Then

$$P[|\bar{X}_n - \mu| \geq a] \leq \frac{\sigma^2}{na^2}. \tag{7.15}$$

This gives a perhaps more intuitive way of thinking of the weak law of large numbers. It says that if the sample size n is very large, then the probability that the sample mean $\bar{X}_n$ is far from the population mean μ is moderately small.

The result is stated in terms of an arbitrary number a with the same dimensions as the X_i. It can also be stated in terms of an dimensionless constant t with $a = t\sigma/\sqrt{n}$. Then the weak law of large numbers takes the form

$$P[|\bar{X}_n - \mu| \geq t\frac{\sigma}{\sqrt{n}}] \leq \frac{1}{t^2}. \tag{7.16}$$

This is actually not a very useful result, since the probability is often much less than the bound $1/t^2$ on the right hand side. If $t = 2$, then the bound is $1/4$, but the central limit theorem shows that $1/20$ is often a much more accurate estimate.

7.4 The sample variance

The sample mean

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \tag{7.17}$$

is a random variable that may be used to estimate an unknown population mean μ. In the same way, the sample variance

$$s^2 = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n-1} \tag{7.18}$$

may be used to estimate an unknown population variance σ^2.

The $n-1$ in the denominator seems strange. However it is due to the fact that while there are n observations X_i, their deviations from the sample mean $X_i - \bar{X}_n$ sum to zero, so there are only $n-1$ quantities that can vary independently. The following theorem shows how this choice of denominator makes the calculation of the expectation give a simple answer.

Theorem 7.5 *Let $X_1, X_2, X_3, \ldots X_n$ be random variables, each with mean μ and standard deviation σ. Assume that each pair X_i, X_j of random variables with $i \neq j$ is uncorrelated. Let s^2 be the sample variance. Then the expectation of s^2 is*

$$E[s^2] = \sigma^2. \tag{7.19}$$

Proof: Compute

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n ((X_i - \bar{X}_n + \bar{X}_n - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 + \sum_{i=1}^n (\bar{X}_n - \mu)^2. \tag{7.20}$$

Notice that the cross terms sum to zero. Take expectations. This gives

$$n\sigma^2 = E[\sum_{i=1}^n (X_i - \bar{X}_n)^2)] + n\frac{\sigma^2}{n}. \tag{7.21}$$

The result then follows from elementary algebra.

Remark: It is not true that the expectation of s is σ. In fact, we have the following theorem and corollary. It is not so much a useful result as a warning not to take things for granted. It does much to explain why statisticians find it more convenient to work with the variance.

Theorem 7.6 *Let X be a random variable that is not constant. Then $E[X]^2 < E[X^2]$. In fact*

$$E[X]^2 = E[X^2] - \mathrm{Var}(X). \tag{7.22}$$

Proof: This is just the usual alternative form $\mathrm{Var}(X) = E[X^2] - E[X]^2$ of writing the variance.

Corollary 7.4 *Suppose that the sample standard deviation s is not constant. Then the expectation of the sample standard deviation s is strictly less than the population standard deviation σ. In fact,*

$$E[s]^2 = \sigma^2 - \mathrm{Var}(s). \tag{7.23}$$

7.5 Experiments and super experiments

This last section deals with the internal consistency of probability theory. Say that we think of measuring a random variable X with mean μ as part of a single experiment. Then we can think of repeating this original experiment n times under independent, identical conditions. This would give a new experiment, which we might call a super experiment. There would be n experimental numbers, the $X_1, \ldots, X_n$ from the n repetitions. These X_i would be independent random variables, all with the same distribution. So the weak law of large numbers would apply. The probability that the sample mean $\bar{X}_n$ is within a of the expectation μ is some number p' bounded below by $1 - \sigma^2/(na^2)$. This is close to one if n is large.

All this applies in the special case when the X_i are Bernoulli random variables corresponding to events that each have probability p. The sample mean in this case is the sample proportion. Then the weak law of large numbers says that the probability that the sample proportion is within a of p is some number p' bounded below by $1 - p(1-p)/(na^2)$. This is close to one if n is large.

However, the statements about the super experiment are also statements about probability. So what is their empirical interpretation? Imagine a super super experiment, in which the entire super experiment is repeated m times. Consider the event that the sample mean in a super experiment is within a of μ. This event has probability p' in the super experiment. Then the probability that the sample proportion in the super super experiment of this event is within a' of p' is some number p'' bounded below by $1 - p'(1-p')/(ma'^2)$. This is itself close to one if m is large.

This all shows that probability theory has a certain internal consistency. Does it show that probability theory must apply correctly to the real world? That is not so clear. In the end one seems to need a principle that an event with probability close to one should actually happen. The mathematical theory does not quite provide this.

If the mathematical theory of probability does not fully explain the success of probability in practice, this does not mean that probability is not a useful tool. All it means is that the ultimate justification may need to be found in some other branch of theoretical science.

Chapter 8

The central limit theorem

8.1 Centering and standardizing

Let X be a random variable with mean μ. Then the *centered* X is the random variable $X - \mu$. It measures the deviation of X from its expected value. Let X be non-constant with standard deviation $\sigma > 0$. The *standardized* X is the random variable

$$Z = \frac{X - \mu}{\sigma}. \tag{8.1}$$

It measures the deviation of X from its expected value in units of the standard deviation.

8.2 Normal random variables

A *normal* (or *Gaussian*) random variable X with mean μ and variance σ^2 is a random variable with density

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \tag{8.2}$$

The centered version of X will have density

$$f_{X-\mu}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}. \tag{8.3}$$

The standardized version Z of X will have density

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{8.4}$$

We shall most often work with the centered or standardized versions. The tables of the distributions of normal random variables are for the standardized version.

Why is the normal distribution so important? The explanation traces back to the following remarkable property of the normal distribution.

Theorem 8.1 *Let X and Y be independent centered normal random variables with variances σ^2. Let $a^2 + b^2 = 1$. Then $aX + bY$ is a centered normal random variable with variance σ^2.*

Proof: The joint density of X and Y is

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2}{2\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}. \tag{8.5}$$

We can use properties of exponentials to write this as

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{8.6}$$

Let $Z = aX + bY$ and $W = -bX + aY$. We can write the joint density of X and Y as

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(ax+by)^2+(-bx+ay)^2}{2\sigma^2}}. \tag{8.7}$$

Now we can factor this again as

$$f(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(ax+by)^2}{2\sigma^2}} e^{-\frac{(-bx+ay)^2}{2\sigma^2}}. \tag{8.8}$$

Let $Z = aX + bY$ and $W = -bX + aY$. The transformation $z = ax + by$, $w = -bx + ay$ is a rotation. Therefore it preserves areas: $dz\,dw = dx\,dy$. It follows that Z and W have joint density

$$f(z,w) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}. \tag{8.9}$$

Corollary 8.1 *Let $X_1, \ldots, X_n$ be independent random variables with centered normal distributions with variance σ^2. Then the sum $X_1 + X_2 + \cdots + X_n$ is also normal, with variance $n\sigma^2$.*

Proof: We can prove this by showing that it is true when $n = 1$ and by showing that for all n, if it is true for n, it is also true for $n + 1$.

The fact that it is true for $n = 1$ is obvious. Suppose that it is true for n. Then the sum $X_1 + \cdots + X_n$ is normal with variance $n\sigma^2$. It follows that $(X_1 + \cdots + X_n)/\sqrt{n}$ is normal with variance σ^2. The next term X_{n+1} is also normal with variance σ^2. Take $a = \sqrt{n}/\sqrt{n+1}$ and $b = 1/\sqrt{n+1}$. Then by the theorem $(X_1 + \cdots + X_n + X_{n+1})/\sqrt{n+1}$ is normal with variance σ^2. It follows that $X_1 + \cdots + X_{n+1}$ is normal with variance $(n+1)\sigma^2$.

8.3 The central limit theorem: statement

Theorem 8.2 *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each with mean μ and standard deviation σ. Let $\bar{X}_n$ be their sample mean. Then the standardized variable*

$$Z_n = \frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \tag{8.10}$$

has the property that

$$\lim_{n\to\infty} P[s < Z_n \leq t] = P[s < Z \leq t], \tag{8.11}$$

where Z is standard normal.

Sometimes the theorem is used in the version where $s = -\infty$. It is also common to see it in the two sided version where $t > 0$ and $s = -t$. For instance, it says that for large n the sample mean satisfies

$$P[|\bar{X}_n - \mu| \leq t\frac{\sigma}{\sqrt{n}}] \approx \int_{-t}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\,dz. \tag{8.12}$$

It is not only close to the population mean, but this tells exactly how close it is likely to be.

Corollary 8.2 *Let $E_1, \ldots, E_n$ be independent events, each with probability p. Let N_n be the number of events that happen, and let $f_n = N_n/n$ be the relatively frequency of the events that happen. Then the standardized variable*

$$Z_n = \frac{N_n - np}{\sqrt{n}\sqrt{p(1-p)}} = \frac{f_n - p}{\sqrt{p(1-p)}/\sqrt{n}} \tag{8.13}$$

has the property that

$$\lim_{n\to\infty} P[s < Z_n \leq t] = P[s < Z \leq t], \tag{8.14}$$

where Z is standard normal.

8.4 The central limit theorem: proof

In this section we give a proof of the central limit theorem. We actually prove a somewhat different version that deals with expectations instead of probabilities. It is shown in more advanced treatments that the different versions are equivalent.

Theorem 8.3 *Let $X_1, X_2, X_3, \ldots$ be independent identically distributed random variables with mean zero and variance σ^2. Let Z be a normal random variable with mean zero and variance σ^2. Let f be a smooth function with bounded derivatives. Then*

$$E[f(\frac{X_1 + \cdots + X_n}{\sqrt{n}})] \to E[f(Z)] \tag{8.15}$$

as $n \to \infty$. Here Z is a standard normal random variable.

Proof: Write

$$E[f(\frac{X_1 + \cdots + X_n}{\sqrt{n}})] - E[f(\frac{Z_1 + \cdots + Z_n}{\sqrt{n}})] = \sum_{i=1}^{n} (E[f(W_i + \frac{X_i}{\sqrt{n}})] - E[f(W_i + \frac{Z_i}{\sqrt{n}})]). \tag{8.16}$$

Here

$$W_i = \frac{X_1 + \cdots + X_{i-1} + Z_{i+1} + \cdots + Z_n}{\sqrt{n}}. \tag{8.17}$$

Next write

$$f(W_i + \frac{X_i}{\sqrt{n}}) = f(W_i) + f'(W_i)\frac{X_i}{\sqrt{n}} + \frac{1}{2}f''(W_i)\frac{X_i^2}{n} + R_i. \tag{8.18}$$

Here R_1 is a remainder. In the same way, write

$$f(W_i + \frac{Z_i}{\sqrt{n}}) = f(W_i) + f'(W_i)\frac{Z_i}{\sqrt{n}} + \frac{1}{2}f''(W_i)\frac{Z_i^2}{n} + S_i. \tag{8.19}$$

Again S_i is a remainder.

Now calculate the expectations. The expectations of the zeroth order terms are the same.

Use independence to calculate

$$E[f'(W_i)\frac{X_i}{\sqrt{n}}] = E[f'(W_i)]E[\frac{X_i}{\sqrt{n}}] = 0. \tag{8.20}$$

and similarly for the other case. The expectations of the first order terms are zero.

Use independence to calculate

$$E[f'(W_i)\frac{X_i^2}{n}] = E[f'(W_i)]E[\frac{X_i^2}{n}] = E[f''(W_i)]\frac{\sigma^2}{n} \tag{8.21}$$

and similarly for the other case. The expectations of the second order terms are the same.

This is the heart of the argument: up to second order everything cancels exactly. This shows that the sum of interest is a sum of remainder terms.

$$E[f(\frac{X_1 + \cdots + X_n}{\sqrt{n}})] - E[f(\frac{Z_1 + \cdots + Z_n}{\sqrt{n}})] = \sum_{i=1}^{n}(E[R_i] - E[S_i]). \tag{8.22}$$

The only remaining task is to show that these remainder terms are small. This is a somewhat technical task. Here is an outline of how it goes.

Write

$$R_i = \frac{1}{2}(f''(W_i + \alpha\frac{X_i}{\sqrt{n}}) - f''(W_i))\frac{X_i^2}{n}. \tag{8.23}$$

In the same way, write

$$S_i = \frac{1}{2}(f''(W_i + \beta\frac{Z_i}{\sqrt{n}}) - f''(W_i))\frac{Z_i^2}{n}. \tag{8.24}$$

Look at the first remainder $E[R_i]$. Break up the sample space into two parts. The first part is where $|X_i| \leq \epsilon\sqrt{n}$. The second part is where $|X_i| > \epsilon\sqrt{n}$.

Assume that the absolute value of the third derivative of f is bounded by a constant C. Then the size of the first part of the first remainder is bounded by

$$E[|R_i|; |X_i| \leq \epsilon\sqrt{n}] \leq \frac{1}{2}C\epsilon\frac{\sigma^2}{n}. \tag{8.25}$$

Assume that the absolute value of the second derivative of f is bounded by a constant M. The size of the second part of the first remainder is bounded by

$$E[|R_i|; |X_i| > \epsilon\sqrt{n}] \leq M\frac{1}{n}E[X_i^2; X_i > \epsilon\sqrt{n}]. \tag{8.26}$$

Let ϵ depend on n, so we can write it as a sequence ϵ_n. Then the first remainder is bounded by

$$E[|R_i|] \leq \frac{1}{2}C\epsilon_n\frac{\sigma^2}{n} + M\frac{1}{n}E[X_i^2; X_i > \epsilon_n\sqrt{n}]. \tag{8.27}$$

Thus the absolute value of the first sum is bounded by

$$\sum_{i=1}^{n} E[|R_i|] \leq \frac{1}{2}C\epsilon_n\sigma^2 + ME[X_i^2; X_i > \epsilon_n\sqrt{n}]. \tag{8.28}$$

Take $\epsilon_n \to 0$ with $\epsilon_n\sqrt{n} \to \infty$. Then the first sum goes to zero as $n \to \infty$. The same idea works for the other sum. This completes the proof.

Chapter 9

Estimation

9.1 Sample mean and population mean

The picture is that there is a very large (theoretically infinite) population. Each member of the population has some characteristic quantity X. The mean of this quantity for the whole population is μ. The standard deviation of this quantity over the whole population is σ.

One can think of taking a single random member of the population and measuring this quantity X_1. Then the expectation of the random variable X_1 is μ, and the standard deviation of X_1 is σ.

Now consider the experiment of taking a random sample of size n and measuring the corresponding quantities $X_1, \ldots, X_n$.

Theorem 9.1 *Let $X_1, X_2, X_3, \ldots, X_n$ be random variables, each with mean μ. Let*

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \tag{9.1}$$

be their sample mean. Then the expectation of $\bar{X}_n$ is

$$E[\bar{X}_n] = \mu. \tag{9.2}$$

Theorem 9.2 *(Weak Law of Large Numbers) Let $X_1, X_2, X_3, \ldots, X_n$ be independent random variables, each with mean μ and standard deviation σ. Let*

$$\bar{X}_n = \frac{X_1 + \cdots + X_n}{n} \tag{9.3}$$

be their sample mean. Then the standard deviation of $\bar{X}_n$ is

$$\sigma_{\bar{X}_n} = \frac{\sigma}{\sqrt{n}}. \tag{9.4}$$

For the statistician the μ and σ are unknown. The significance of the above theorems is that it is reasonable to use the experimental sample mean $\bar{X}_n$ to estimate the unknown population parameter μ, provided that n is large enough. If n is large enough, then $\bar{X}_n$ is quite likely to be near μ. The central limit theorem gives a rather precise idea of how variable these sample means are.

Theorem 9.3 *(Central Limit Theorem) Let $X_1, \ldots, X_n$ be independent and identically distributed random variables, each with mean μ and standard deviation σ. Let $\bar{X}_n$ be their sample mean. Then the standardized variable*

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \tag{9.5}$$

has the property that

$$\lim_{n\to\infty} P[s < Z_n \leq t] = P[s < Z \leq t], \tag{9.6}$$

where Z is standard normal.

Now the only trouble with this is that the σ is also an unknown population parameter. So to understand what one is doing, one has to estimate σ. The following result tells how to do this.

Define the sample variance as

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}{n-1} \tag{9.7}$$

This is an experimental quantity used to estimate the unknown population variance σ^2.

Theorem 9.4 *Let $X_1, X_2, X_3, \ldots X_n$ be independent random variables, each with mean μ and standard deviation σ. Let s^2 be the sample variance. Then the expectation of s^2 is*

$$E[s^2] = \sigma^2. \tag{9.8}$$

9.2 Sample median and population median

The picture is that there is a very large (theoretically infinite) population. Each member of the population has some characteristic quantity X. Consider a number α between zero and one. Then there is supposed to be a number t_α such that the proportion of the population for which the X is less than or equal to α is t_α.

One can think of taking a single random member of the population and measuring this quantity X_1. The assumption is that X_1 is a continuous random variables. Then the cumulative distribution function $F(t) = P[X \leq t]$ is continuous. It follows that there is a t_α such that $F(t_\alpha) = \alpha$.

There are several common examples. The most important is the value such that half the population is above this value and half the population is below this value. Thus when $\alpha = 1/2$ the corresponding $t_{\frac{1}{2}}$ is called the population median m.

Similarly, when $\alpha = 1/4$ the $t_{\frac{1}{4}}$ is called the lower population quartile. In the same way, when $\alpha = 3/4$ the $t_{\frac{3}{4}}$ is called the upper population quartile. In statistics the function F characterizing the population is unknown. Therefore all these t_α are unknown quantities associated with the population.

Now consider the experiment of taking a random sample of size n and measuring the corresponding quantities $X_1, \ldots, X_n$. Thus again we have independent random variables all with the same distribution. We are assuming that the distribution is continuous. Thus the probability is one that for all $i \neq j$ the quantities $X_i \neq X_j$ are unequal.

The order statistics $X_{(1)}, \ldots, X_{(n)}$ are the quantities obtained by arranging the random variables $X_1, \ldots, X_n$ in increasing order. Thus by definition

$$X_{(1)} < X_{(2)} < \cdots < X_{(i)} < \cdots < X_{(n-1)} < X_{(n)}. \tag{9.9}$$

The order statistics are no longer independent, as we now see.

The joint density of $X_1, \ldots, X_n$ is $f(x_1) \cdots f(x_n)$. This product structure is equivalence to the independence of the random variables. On the other hand, the joint density of the order statistics $X_{(1)}, \ldots, X_{(n)}$ is $n! f(x_1) \cdots f(x_n)$ for $x_1 < x_2 < \cdots < x_n$ and zero otherwise. There is no way to factor this. The order statistics are far from independent.

The order statistics are quite useful for estimation. Take $\alpha = i/(n+1)$. Then it seems reasonable to use the order statistics $X_{(i)}$ to estimate t_α.

Thus, for instance, if n is odd and $i = (n+1)/2$ and $\alpha = 1/2$, then $X_{(i)}$ is the sample median. This estimates the population median $m = t_{\frac{1}{2}}$.

The fundamental theorem on order statistics is the following. It shows that questions about order statistics reduce to questions about binomial random variables.

Theorem 9.5 *Let $X_1, \ldots, X_n$ be independent random variables with a common continuous distribution. Let $X_{(1)}, \ldots, X_{(n)}$ be their order statistics. For each x, let $N_n(x)$ be the number of i such that $X_i \leq x$. Then $N_n(x)$ is a binomial random variable with parameters n and $F(x)$. Furthermore,*

$$P[X_{(j)} \leq x] = P[N_n(x) \geq j]. \tag{9.10}$$

This result can be stated even more explicitly in terms of the binomial probabilities. In this form it says that if $P[X_i \leq x] = F(x)$, then

$$P[X_{(j)} \leq x] = \sum_{k=j}^{n} \binom{n}{k} F(x)^k (1 - F(x))^{n-k}. \tag{9.11}$$

This theorem is remarkable, in that it gives a rather complete description of order statistics for large sample sizes. This is because one can use the central limit theorem for the corresponding binomial random variables.

Theorem 9.6 *Let $X_1, \ldots, X_n$ be independent random variables with a common continuous distribution. Let $X_{(1)}, \ldots, X_{(n)}$ be their order statistics. Fix α and* *let $F(t_\alpha) = \alpha$. Let $n \to \infty$ and let $j \to \infty$ so that $\sqrt{n}(j/n - \alpha) \to 0$. Then the order statistics $X_{(j)}$ is approximately normally distributed with mean*

$$E[X_{(j)}] \approx t_\alpha \tag{9.12}$$

and standard deviation

$$\sigma_{X_{(j)}} \approx \frac{\sqrt{\alpha(1-\alpha)}}{f(t_\alpha)\sqrt{n}}. \tag{9.13}$$

Proof: Compute

$$P[X_{(j)} \leq t_\alpha + \frac{a}{\sqrt{n}}] = P[\frac{N_n(t_\alpha + \frac{a}{\sqrt{n}})}{n} \geq \frac{j}{n}]. \tag{9.14}$$

The random variable has mean $\alpha'_n = F(t_\alpha + a/\sqrt{n})$ and variance $\alpha'_n(1-\alpha'_n)/n$. So we can use the central limit theorem and $j/n \to \alpha$ to write this as

$$P[X_{(j)} \leq t_\alpha + \frac{a}{\sqrt{n}}] \approx P[Z \geq \frac{\alpha - \alpha'_n}{\sqrt{\alpha'_n(1-\alpha'_n)}/\sqrt{n}}]. \tag{9.15}$$

However

$$\alpha'_n - \alpha = F(t_\alpha + \frac{a}{\sqrt{n}}) - F(t_\alpha) \approx f(t_\alpha)\frac{a}{\sqrt{n}}. \tag{9.16}$$

So we use this together with $\alpha_n \to \alpha$ to get

$$P[X_{(j)} \leq t_\alpha + \frac{a}{\sqrt{n}}] \approx P[Z \geq -\frac{f(t_\alpha)a}{\sqrt{\alpha(1-\alpha)}}]. \tag{9.17}$$

In other words,

$$P[X_{(j)} \leq t_\alpha + \frac{a}{\sqrt{n}}] \approx P[-\frac{\sqrt{\alpha(1-\alpha)}}{f(t_\alpha)}Z \leq a]. \tag{9.18}$$

This gives the result.

Corollary 9.1 *Let $X_1, \ldots, X_n$ be independent random variables with a common continuous distribution. Let $X_{(1)}, \ldots, X_{(n)}$ be their order statistics. Let m be the population median. Consider sample sizes n that are odd, so that the sample median $M_n = X_{(\frac{n+1}{2})}$ is defined. Let $n \to \infty$. Then the sample median M_n is approximately normally distributed with mean*

$$E[M_n] \approx m \tag{9.19}$$

and standard deviation

$$\sigma_{M_n} \approx \frac{1}{2f(m)\sqrt{n}}. \tag{9.20}$$

9.3 Comparison of sample mean and sample median

Say that each X_i has density f with population mean

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx \tag{9.21}$$

and population variance

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx. \tag{9.22}$$

Then our tendency is to use the sample mean $\bar{X}_n$ to estimate the population mean μ.

Say that

$$F(x) = \int_{-\infty}^{x} f(t)\, dt \tag{9.23}$$

is the distribution function. Say that m is the population median, so $F(m) = \frac{1}{2}$. Then our tendency is to use the sample median M_n to estimate the population median m.

Now say that the density function is symmetric about some point. Then the population mean is the same as the sample mean. So the mean and the median are both trying to estimate the same quantity. Which is better?

The relative efficiency of the sample median to the sample mean is found by seeing which has smaller variance. The smaller the variance, the more efficient the estimation. The efficiency of the sample median with respect to the sample mean is the variance of the sample mean divided by the variance of the sample median. This ratio is

$$\frac{\sigma^2/n}{1/(4f(m)^2 n)} = 4f(m)^2\sigma^2. \tag{9.24}$$

If this ratio is less than one, then the sample mean is a better estimator. If this ration is greater than one, then the sample median is better.

For the normal distribution $f(m) = f(\mu) = 1/(\sqrt{2\pi}\sigma)$. Thus the relative efficiency of the sample median to the sample mean is $2/\pi$. This is not particularly good. If a statistician somehow knows that the population distribution is normal, then the sample mean is the better statistic to use.

However it is quite possible that the density value $f(m) > 0$ is well away from zero, but the distribution has long tails so that the σ is huge. Then the median may be much more efficient than the mean. So in many situations the median is the safer estimator to use.

Of course, maybe the statistician does not know the size of $f(m)$ or the size of σ. Some kind of preliminary analysis of the data is then necessary to establish a preference.

9.4 The Kolmogorov-Smirnov statistic

Say that one has a hypothesis that the population distribution is F and wants to check it. The following theorem gives a method.

Theorem 9.7 *Let X be a random variable with a continuous distribution F. Then $U = F(X)$ is a uniform random variable.*

Theorem 9.8 *Let $X_1, \ldots, X_n$ be independent random variables with the same continuous distribution. Let $X_{(1)}, \ldots, X_{(n)}$ be their order statistics. Let $U_1 = F(X_1), \ldots, U_n = F(X_n)$. Then these are independent random variables, each uniformly distributed in the interval* $[0, 1]$. *Their order statistics are $U_{(1)} = F(X_{(1)}), \ldots, U_{(n)} = F(X_{(n)})$.*

The method is to compute the $U_{(1)}, \ldots, U_{(i)}, \ldots, U_{(n)}$ and compare them with $1/(n+1), \ldots, i/(n+1), \ldots, n/(n+1)$. If they are close, then this is a confirmation of the hypothesis.

There is a famous Kolmogorov-Smirnov statistic that is based on this general idea. This statistics gives a quantitative measure of the degree to which the order statistics $U_{(i)} = F(X_{(i)})$ behave as anticipated by the hypothesis.

In this statistic, the comparison is between the order statistics $U_{(i)}$ and the numbers $(i - 1/2)/n$ (rather than $i/(n+1)$). The reason for using $(i - 1/2)/n$ seems technical. The following may provide some motivation. The proportion of order statistics less than the ith order statistic is $(i-1)/n$; the proportion of order statistics less than or equal to the ith order statistics is i/n. The average of these two numbers is $(i - 1/2)/n$.

Thus the Kolmogorov-Smirnov statistic is defined to be

$$D = \frac{1}{2n} + \max_{1 \leq i \leq n} \left| F(X_{(i)}) - \frac{i - \frac{1}{2}}{n} \right|. \tag{9.25}$$

The first term is there because the Kolmogorov-Smirnov statistics is usually defined by a different and perhaps more natural formula, from which this term emerges after a calculation.

A typical result about the Kolmogorov-Smirnov statistics is that if F is the distribution function of the X_i, then for moderately large n

$$P[D > \frac{1.36}{\sqrt{n}}] \leq 0.05. \tag{9.26}$$

So this provides a test of whether the data X_i really come from a population described by F. If so, then large values of D are unlikely.