Math 37600 PR - (19364) - Lectures 01

Ethan Akin
Email: eakin@ccny.cuny.edu

Fall, 2024

Introduction

- The book is *Introduction to Mathematical Statistics* by Robert V. Hogg, Joseph W. McKean and Allen T. Craig Eighth Edition (hereafter HMC).
- READ THE BOOK. Keep up with the homework. I will be collecting homework.
- Ask questions.
- The course information sheet and the term's homework assignments are posted on my site:

  $https://math.sci.ccny.cuny.edu/peoplename = EthanAkin$

- I will be posting there a pdf of the slides I am using here. The first batch for the course is already up. I will post the remaining pieces as we get to them.
- The class will meet from 2pm to 3:40 on Tuesdays and Thursdays in NAC 6/114. If due to inclement weather, I am unable to come in, or if we are switched to online mode, we will meet at the scheduled time using Blackboard Collaborate Ultra.
- Office: MR (Marshak) 325A

  Office Hours: Tuesday 11:00am-1:00pm,
  Thursday 11:00am-12:00.

  Email: eakin@ccny.cuny.edu

Sec 1.3: Probability Properties

An *event* A is a subset of the *sample space* S, the set of possible outcomes. We write A^c for the complementary event $S \setminus A$. Two events corresponding to disjoint sets are said to be *mutually exclusive* .

Each event A is assigned a probability $P(A)$. The axioms are given in HMC Def. 1.3.1:

- $P(A) \geq 0$ for every event A.
- $P(S) = 1$.
- If $A_1, A_2, \ldots$ is a sequence of mutually exclusive events ($A_i \cap A_j = \emptyset$ if $i \neq j$), then

$$P(\bigcup_i A_i) = \sum_i P(A_i).$$

Imagine S is a subset of the plane with area 1 and for $A \subset S$, think of $P(A)$ is the area of A.

We review the elementary consequences of the axioms given as HMC Thm. 1.3.1 - 1.3.6.

- $P(A^c) = 1 - P(A)$ for every event A.
  In particular, $P(\emptyset) = 0$.
- If $A \subseteq B$, then $P(A) \leq P(B)$.
  In particular, $0 \leq P(A) \leq 1$.
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
- If $A_1, A_2, \ldots$ is an increasing sequence of events, then

$$P(\bigcup_i A_i) = \lim_{i \to \infty} P(A_i) = \sup\{P(A_i)\}.$$

- If $A_1, A_2, \ldots$ is a decreasing sequence of events, then

$$P(\bigcap_i A_i) = \lim_{i \to \infty} P(A_i) = \inf\{P(A_i)\}.$$

Sec 1.3: Permutations and Combinations

We count the number of ways we can make k choices from a set of size n. Think of k blanks and we are filling them in, one blank after the other from the n set.

Sampling WITH Replacement You make k selections - in order. After each choice is recorded, the object is replaced. So for each of the k selections there are n choices. Thus, there are $n^k = n \cdot n \ldots n$ possible samples.

Sampling WITHOUT Replacement You make k selections - in order. After each choice is recorded, the object is not replaced. After the first selection with n choices is made, there are now only $n-1$ choices for the second selection. Thus, there are $n_k = n \cdot n-1 \cdot \ldots n-(k-1)$ possible samples.

In particular, the number of *permutations* of the set, that is, the number of ways of rearranging the order of the elements is $n_n = n!$ (n *factorial*). Notice that $n_k = n! \div (n-k)!$.

To count the number of subsets of size k, we first count the number of samples of size k without replacement. Each is a list of k elements where the order in which they have been chosen matters. There are n_k such lists or ordered samples. Each rearrangement of a sample yields the same set. For a list of k elements, there are $k!$ rearrangements or permutations. To count the subsets, we group together the $k!$ samples which give the same subset.

For example, from a deck of $n = 52$ we have $52_5 = 52 \cdot 51 \ldots 48$ deals of five cards. Here a deal is a list of five cards dealt in order. So there are 52_5 deals of cards.

For a hand of cards the order in which the cards were dealt does not matter. Each of the 5! rearrangements of the original deal yields the same hand. So we divide by 5! to get the number of hands.

Thus, the number of subsets of size k which can be chosen from a set of size n is

$$\binom{n}{k} = n_k \div k! = \frac{n!}{k!(n-k)!}.$$

The symbol $\binom{n}{k}$ is read n *choose* k.

A flush in hearts consists of five different cards each of which is among the 13 cards which make up the suit of hearts. We compute the probability of a flush in hearts two ways.

For the first method, our sample space is all deals of five cards. We have seen there are 52_5 such deals. The number of deals from the heart suit alone is 13_5. So the probability of a flush in hearts is $(13_5)/(52_5) = \frac{13 \cdot 12 \cdot 11 \cdot 10 \cdot 9}{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}$.

For the second method, our sample space is all hands of five cards. We have seen there are $\binom{52}{5} = 52_5 \div 5!$ such hands. The number of hands from the heart suit alone is $\binom{13}{5} = 13_5 \div 5!$. So the probability of a flush in hearts is

$$\binom{13}{5} / \binom{52}{5} \;=\; (13_5 \div 5!)/(52_5 \div 5!) \;=\; (13_5)/(52_5)$$

because the 5!'s cancel out.

So the two methods give the same answer.

Let us look at HMC Exercise 1.3.13: Three distinct integers are chosen from among the first twenty positive integers. What is the probability that (a) that the sum is even, (b) that the product is even. Note that event (b) is the complement of all three odd, while (a) is the union of the mutually exclusive events: all three even or exactly one is even. On the other hand, that the sum is odd is the union of the mutually exclusive events: all three odd or exactly one is odd. Since there are the same number of evens and odds, we see that the probability that sum is even is $\frac{1}{2}$ whether we use sampling with or without replacement. For (b) we get different answers.

With Replacement: P(All three odd) $= \frac{1}{8}$ and so the product is even with probability $= \frac{7}{8} = .875$.

Without Replacement: P(All three odd) (Just like the flush in hearts) $= \frac{10 \cdot 9 \cdot 8}{20 \cdot 19 \cdot 18}$ and so the product is even with probability .895.

In general, if we add up an odd number of numbers between 1 and 20, the probability that the sum is even equals the probability that the sum is odd. Here we use a trick. Look at the function $x \mapsto y = 21 - x$. This maps evens to odds and vice versa. Furthermore, if the sum of the x's is even, then the sum of the y's is odd and vice-versa, because the odd number of 21's has an odd sum. Thus, we see that the number of lists with an even sum is the same as the number of lists with an odd sum. So the probability of each is $\frac{1}{2}$

Notice that the sum problem becomes more complicated if we have an even number of integers picked out. It is possible to show that with replacement we still get that the probability of an even sum and the probability of an odd sum each equals $\frac{1}{2}$.

Without replacement. It is possible to prove that if $2k$ integers are added up, we get:

The probability of an odd sum is greater than the probability of an even sum if k is odd, but

The probability of an even sum is greater than the probability of an odd sum if k is even.

To be continued.

Sec 1.4: Conditional Probability

For events A and B the *conditional probability of A given B* is

$$P(A|B) \;=\; \frac{P(A \cap B)}{P(B)}.$$

In particular, if $A \subset B$ then $P(A|B) \;=\; \frac{P(A)}{P(B)}$. These require $P(B) > 0$.

Of great importance is the simple re-write of the definition:

$$P(A|B) \cdot P(B) \;=\; P(A \cap B).$$

Think of $P(A)$ as the probability that a point in the sample space is in the set A. Now suppose we know that B is true. We are given that the point is in B. In that case the point is in A exactly when it is in both A and B, and so is in $A \cap B$.

Now $P(A|B)$ is the new probability given the information that B is true and so we can exclude the points of B^c.

The definition $P(A|B)$ is $\frac{P(A \cap B)}{P(B)}$ with both the top and the bottom computed from the original sample space. Often, however, we will regard B as the new sample space and compute $P(A|B)$ directly.

Let us go back HMC Exercise 1.3.13.

Think of choosing two integers from 1 to 20 without replacement. Suppose the event A is that the second one is even and B is the probability that the first one is even. It is easy to see that $P(B) = \frac{1}{2}$. Also, $P(A) = \frac{1}{2}$. Think of putting the first integer in my left hand and the second in my right. For either one the probability is the same.

The event $A \cap B$ is that both are even.

$$P(A \cap B) = \frac{10_2}{20_2} = \frac{10 \cdot 9}{20 \cdot 19} = \frac{1}{2} \cdot \frac{9}{19}.$$

So

$$P(A|B) = P(A \cap B)/P(B) = (\frac{1}{2} \cdot \frac{9}{19}) \div (\frac{1}{2}) = \frac{9}{19}.$$

Now lets compute it the other way. Given B is true, our new sample space consists of 19 integers, 9 of which are even and 10 of which are odd. Now A is true when we choose an even integer in this sample space. So the probability of A given B is $\frac{9}{19}$.

Now let us return to problem HMC Exercise 1.3.13.

We are choosing $2k$ integers from among $\{1, 2, \ldots, 20\}$
For example, if $k = 1$ so that we have two integers, we get an even sum if we have either EE or OO each with probability $\frac{10}{20} \cdot \frac{9}{19}$. We get an odd sum if we have either EO or OE each with probability $\frac{10}{20} \cdot \frac{10}{19}$. So the difference in the probabilities is $P(odd) - P(even) = \frac{1}{19}$.

With $k = 2$ and so there are four integers we get an even sum with either EEEE or OOOO or a rearrangement of EEOO. We get an odd sum with either a rearrangement of EOOO or OEEE. From this we can compute that
$P(even) - P(odd) = \frac{9}{2 \cdot 19 \cdot 17}$.

Let us see how to compute the probability that one out of the four integers chosen is even. So the list is a rearrangement of EOOO. The probability that the list is of the type EOOO is

$$\frac{10}{20} \cdot \frac{10}{19} \cdot \frac{9}{18} \cdot \frac{8}{17}.$$

The other arrangements are OEOO, OOEO, and OOOE. To see why there are four, think of choosing one blank out of the four for the E ($\binom{4}{1} = 4$ choices) and then the rest are O's. Each arrangement has the same probability: $\frac{10_1 \cdot 10_3}{20_4}$. So the probability that one out of the four integers chosen is even equals

$$4 \cdot \frac{10_1 \cdot 10_3}{20_4}.$$

There is another, probably easier, way of computing the probability of the event that one out of the four integers chosen are even.

Our sample space consists of all subsets of 4 integers out of 20. There are $\binom{20}{4} = 20_4 \div 4!$ such sets.

For our event we choose 1 integer out of the 10 evens and 3 integers out of the 10 odds. This shows that there are

$$\binom{10}{1} \cdot \binom{10}{3} = (10_1 \div 1!) \cdot (10_3 \div 3!)$$

subsets in the event. So the probability equals

$$\binom{10}{1} \cdot \binom{10}{3} \div \binom{20}{4} = (4!/(1!3!)) \cdot \frac{10_1 \cdot 10_3}{20_4}.$$

The same as before.

The remaining cases of an odd sum are rearrangements of OEEE resulting in the same probability that we just computed. So the probability of an odd sum is

$$8 \cdot \frac{10_1 \cdot 10_3}{20_4}.$$

What about the even sums? The cases EEEE and OOOO each has probability

$$\frac{10_4}{20_4} = \binom{10}{4} \div \binom{20}{4}.$$

As an exercise you should compute the probability for the remaining cases of an even sum. These are rearrangements of EEOO with two out of the four even.

In the important equation:

$$P(A|B) \cdot P(B) \;=\; P(A \cap B),$$

notice that A and B play symmetric roles in $P(A \cap B)$ but not in $P(A|B)$.

From this we obtain Bayes' Theorem which relates $P(A|B)$ and $P(B|A)$.

$$P(B|A) \;=\; \frac{P(A \cap B)}{P(A)} \;=\; \frac{P(A|B) \cdot P(B)}{P(A)}.$$

A list of events $B_1, B_2, \ldots, B_k$ is a *partition* of S when the sets are mutually exclusive and *exhaustive*. From a partition we obtain the *Law of Total Probability* (LOTP) for an event A:

$$P(A) = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{i=1}^{k} P(A|B_i) \cdot P(B_i).$$

The most common partition consists of B and its complement B^c. Using the LOTP we get a useful version of Bayes' Theorem:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|B^c) \cdot P(B^c)}.$$

Notice that we used:

$$P(A) = P(A \cap B) + P(A \cap B^c) = \underline{P(A|B)} \cdot P(B) + \underline{P(A|B^c)} \cdot P(B^c).$$

It is not true that

$$P(A) = \underline{P(A|B)} + \underline{P(A|B^c)}.$$

In fact, because $P(B) + P(B^c) = 1$, we see that $P(A)$ is a weighted average of $P(A|B)$ and $P(A|B^c)$ and so lies between them.

Look at the following idea:

$$P(A|B) + P(A|B^c) = \frac{P(A \cap B)}{P(B)} + \frac{P(A \cap B^c)}{P(B^c)}$$

$$=??$$

$$\frac{P(A \cap B) + P(A \cap B^c)}{P(B) + P(B^c)} = \frac{P(A)}{1} = P(A)$$

But the addition of fractions at $=??$ is incorrect.

What is true is the sum for A_1 and A_2 mutually exclusive

$$P(A_1|B) + P(A_2|B) = \frac{P(A_1 \cap B)}{P(B)} + \frac{P(A_2 \cap B)}{P(B)} =$$

$$\frac{P(A_1 \cap B) + P(A_2 \cap B)}{P(B)} = \frac{P((A_1 \cup A_2) \cap B)}{P(B)} = P(A_1 \cup A_2|B).$$

As the authors put it: Conditional probability given B is a probability.

Testing for a rare disease: Suppose the disease affects 1% of the population and it has a test which is 95% accurate. Suppose that D is the event that a random person has the disease and T is the probability that a person being tested tests positive.

The assumption about the prevalence of the disease means $P(D) = .01$. About the accuracy, we assume that both the sensitivity of the test $P(T|D)$ and the specificity of the test $P(T^c|D^c)$ are .95. What we want is $P(D|T)$, the probability of the disease given a positive test. $P(D|T) =$

$$\frac{P(T|D) \cdot P(D)}{P(T|D) \cdot P(D) + P(T|D^c) \cdot P(D^c)} = \frac{.95 \times .01}{.95 \times .01 + .05 \times .99}.$$

This is approximately .16.

Suppose that in a population of $10,000$, 100 have the disease, so D contains 100 people, 1% of the population.

If the sensitivity of the test is 95% then 95 out of those 100 will test positive.

If the specificity of the test is 95%, then only 5% of the remaining population will test positive, but that is 5% of $9,900$ people which is 495 false positives.

The set T of those who test positive contains $495 + 95 = 590$ people of whom 95 actually have the disease.

So $P(D|T) = 95/590 = .16$.

Notice that $P(T|D) + P(T|D^c) = .95 + .05 = 1 > P(T)$.

If the sensitivity is 95% but the specificity is only 90%, then

$$P(T|D) = .95, \quad , P(T^c|D^c) = .90, \quad P(T|D^c) = .10.$$

In that case, $P(T|D) + P(T|D^c) = 1.05 > 1$ and so is not a probability at all.

Let us look at HMC Exercise 1.4.9: Bowl I contains 6R and 4B chips. Five are moved to Bowl II (previously empty) and then one chip is selected from BowlII. Given that the selected chip was B what is the probability that 2R and 3B were moved to Bowl II?

Look first at the following: Suppose Bowl I contains n chips and one of them is X. Move k chips to Bowl II and select one. What is the probability that the selected chip is X?

Let EX be the event that X is included among the k moved. $P(EX) = \binom{n-1}{k-1} \div \binom{n}{k} = \frac{k}{n}$.

Let SX be the event that X is the selected chip. Notice that $SX \subset EX$. $P(SX|EX) = \frac{1}{k}$.

Therefore,

$$P(SX) = P(SX|EX) \cdot P(EX) = \frac{1}{k} \cdot \frac{k}{n} = \frac{1}{n}.$$

This is just the probability that when one chip is selected from Bowl I, it is X. Applying this to each of the 4 B chips, we see that for the event SB that a B chip was selected, we have $P(SB) = \frac{4}{10} = \frac{2}{5}$.

Now let E(2R3B) be the event that two R's and three B's were moved. $P(E(2R3B)) = (\binom{6}{2} \cdot \binom{4}{3}) \div \binom{10}{5} = \frac{5}{21}$.

We want $P(E(2R3B)|SB)$ which is not easy to see directly. Instead we apply Bayes' Theorem.

$P(SB|E(2R3B)) = \frac{3}{5}$. So $P(SB\&E(2R3B)) = \frac{3}{5} \cdot \frac{5}{21} = \frac{1}{7}$ and

$$P(E(2R3B)|SB) = P(SB\&E(2R3B)) \div P(SB) = \frac{5}{14}.$$

Sec 1.4: Independence

Events A and B are *independent* when

$$P(A \cap B) = P(A) \cdot P(B).$$

This is a strong condition describing when we can multiply probabilities this way.

Notice that $P(A \cap B^c) = P(A) - P(A \cap B)$ and $P(A) \cdot P(B^c) = P(A) - P(A) \cdot P(B)$. So if A and B are independent, then A and B^c are independent.

Independence is true in an uninteresting way if either $P(A)$ or $P(B)$ equals 0.

When $P(B) > 0$, A and B are independent exactly when

$$P(A|B) \;=\; P(A).$$

To illustrate this, look back at sampling. When we sample with replacement, the outcome of the second choice is independent of the outcome of the first.

When we sample without replacement, the outcome of the second choice depends to some extent on the result of the first choice. For example, what was chosen as the first pick cannot occur as the choice of the second pick.

Going back to the case of choosing an integer from among the first twenty, if the first choice is even (event B), the probability that the second choice is even (event A) is $\frac{9}{19}$, but if the first choice was odd, then the probability that the second choice is even is $\frac{10}{19}$. Meanwhile, recall that $P(A) = P(B) = \frac{1}{2}$. That is,

$$P(A|B) = \frac{9}{19}, \quad P(A|B^c) = \frac{10}{19}, \quad P(A) = \frac{10}{20} = \frac{1}{2}.$$

Sec 1.5, 1.6, 1.7: Random Variables

A *random variable*(= rv) X is a real-valued function defined on the sample space. What is "random" is the location of the point s of the sample space S. For each location $s \in S$ $X(s)$ is the value of the function X at s.

The *range* of X is just the set of values of the function. So a function $X : S \to [a, b]$ has range contained in the interval $[a, b]$.

The *cumulative distribution function* (cdf) F_X and the *survival function* G_X of X are defined by

$$F_X(x) = P(X \leq x), \quad G_X(x) = P(X \geq x)$$

The expression $\{X \leq x\}$ is shorthand for the event $\{s \in S : X(s) \leq x\}$.

HMC Theorem 1.5.1 lists the properties of F_X. It is non-decreasing, tending to 0 and 1 as x tends to $-\infty$ and $+\infty$, respectively. While it may have jumps, it is continuous from the right.

An rv X is *discrete* when its range is finite or countably infinite. In our examples, the discrete rv's will have range contained in the set of integers. The support $\mathcal{S}_X$ of a discrete rv consists of those values which occur with positive probability.

The *probability mass function* (pmf) of a discrete rv X is defined by

$$\begin{aligned} f_X(x) &= P(X = x), \\ F_X(x) &= \sum_{y \le x} f_X(y). \end{aligned}$$

Thus, $f_X(x) > 0$ exactly at the values in the support of the rv X. The cdf F_X has jumps exactly at the values in the support of X. Between two adjacent such values it is constant.

For us, an rv X is *continuous* when the cdf F_X is continuous and is differentiable except at a finite set of points.

The *probability density function* (pdf) of a continuous rv X is defined by

$$f_X(x) \;=\; F'_X(x). \;=\; \frac{dF_X}{dx}.$$

So the cdf and pdf are related by:

$$f_X \;=\; \frac{dF_X}{dx}, \qquad F_X(x) \;=\; \int_{-\infty}^{x} f_X(t)\; dt.$$

Notice that the variable t in the definite integral is a dummy variable which is integrated away.

For X a continuous rv the support S_X is the set of points of positive density, i.e. $f_X(x) > 0$. This despite the fact that $P(X = x) = 0$ for every number x.

Notice that $f_X(x)$ is not the probability that $X = x$. Instead $f_X(x)dx$ is the probability that X is in a little interval of length dx which contains x. We label such an interval as $[x + dx]$.

Sec 1.8: Expectation

The expectation $E(X)$ of an rv X is defined in HMC Definition 1.8.1.

$$E(X) = \begin{cases} \sum_x \; x f_X(x) & \text{if } X \text{ is discrete,} \\ \\ \int_{-\infty}^{\infty} \; x f_X(x) \, dx & \text{if } X \text{ is continuous.} \end{cases}$$

That is, it is the weighted average of the values of the rv. Think of a complicated bet on a game with payoff V_i if event A_i occurs with $A_1, \ldots, A_k$ mutually exclusive. The expected value is $\sum_{i=1}^{k} \; V_i P(A_i)$.

If an rv $Y = g(X)$ with $g : \mathbb{R} \to \mathbb{R}$, then the definition of the expectation $E(Y)$ uses the pmf or pdf of Y. However, we can compute it by using the pmf or pdf of X:

$$E(Y) \;=\; E(g(X)) \;=\; \begin{cases} \sum_x \; g(x) f_X(x) & \text{if } X \text{ is discrete,} \\ \\ \int_{-\infty}^{\infty} \; g(x) f_X(x) \; dx & \text{if } X \text{ is continuous.} \end{cases}$$

These require that the sum, when the support is infinite, and the integral, when the support is unbounded so that the integral is improper, converge absolutely. We need only sum or integrate over the set of values in the support of X. While the proof uses matters from Chapter 2, we will use the *linearity* of expectation:

$$E(cX_1 + X_2) \;=\; cE(X_1) + E(X_2).$$

A list of rv's $X_1, X_2, \ldots, X_n$ is *independent* when for every list of real numbers $x_1, x_2 \ldots x_n$,

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \;=\; F_{X_1}(x_1) \cdot F_{X_2}(x_2) \cdot \ldots F_{X_n}(x_n).$$

That is, the probability of the intersection of the events $\{X_i \leq x_i\}$ is the product of the probabilities.

It then follows that $g_1(X_1), g_2(X_2), \ldots, g_n(X_n)$ is independent for any list of functions $g_1, g_2, \ldots, g_n$.

If X_1 and X_2 are independent, then

$$E(X_1 \cdot X_2) \;=\; E(X_1) \cdot E(X_2).$$

This is usually not true without independence.

Sec 1.9: Mean, Variance, Moment Generating Function

For an rv X the *mean* of X is the expectation (if it exists) $\mu_X = E(X)$.

The *variance* $Var(X)$ is given by

$$Var(X) = E((X-\mu_X)^2) = E(X^2 - 2\mu_X X + \mu_X^2)$$

$$= E(X^2) - 2\mu_X E(X) + \mu_X^2 = E(X^2) - \mu_X^2 = E(X^2) - E(X)^2.$$

$$Var(cX) \;=\; E((cX - c\mu_X)^2) \;=\; c^2 Var(X).$$

If X_1 and X_2 are independent, then

$$Var(X_1 + X_2) \;=\; E((X_1 - \mu_1 + X_2 - \mu_2)^2) \;=$$

$$E((X_1 - \mu_1)^2) + 2E((X_2 - \mu_2)(X_1 - \mu_1)) + E((X_2 - \mu_2)^2)$$

$$= \; Var(X_1) + Var(X_2).$$

Again this is usually not true without independence.

The *moment generating function* (the mgf) M_X is defined by

$$M_X(t) = E(e^{tX}).$$

$$M_X'(t) = E(Xe^{tX}), M_X''(t) = E(X^2e^{tX}), \ldots M_X^{(n)}(t) = E(X^ne^{tX}).$$

$$E(X) = M_X'(0), E(X^2) = M_X''(0), \ldots, E(X^n) = M_X^{(n)}(0).$$

HMC Theorem 1.9.2 says that the mgf characterizes the distribution of its rv.

If X is a continuous rv with support in the interval $[a, b]$, then

$$M_X(t) = \int_a^b e^{tx} f_X(x) dx,$$

where f_X is the pdf of X.

In particular, if the support is $[0, \infty)$, then M_X is the Laplace transform of the density function.

If X is a discrete rv with range in $\mathbb{Z}_+$, then

$$M_X(t) = \sum_{x=0}^{\infty} e^{tx} f_X(x),$$

where f_X is the pmf of X.

If $Y = a + mX$, then

$$M_Y(t) \;=\; E(e^{at+mtX}) \;=\; e^{at} \cdot M_X(mt).$$

If X_1 and X_2 are independent, then

$$E(e^{t(X_1+X_2)}) \;=\; E(e^{tX_1} \cdot e^{tX_2}) \;=\; E(e^{tX_1}) \cdot E(e^{tX_2}).$$

That is, $M_{X_1+X_2} = M_{X_1} \cdot M_{X_2}$.

Sec 3.1: Bernoulli,Binomial and Geometric Distributions

For an event A the indicator I_A is the rv taking the values $0, 1$ with

$$I_A(s) = \begin{cases} 1 & \text{for } s \in A, \\ 0 & \text{for } s \notin A. \end{cases}$$

Thus, $A = \{I_A = 1\}$ and $A^c = \{I_A = 0\}$.

In general, an rv I is a *Bernoulli* rv with probability p, written $I \sim Bern(p)$ when I has range in $\{0, 1\}$ and $P(I = 1) = p$. We usually write $P(I = 0) = 1 - p = q$, so that $f_I(1) = p, f_I(0) = q$. The expectation $E(I) = p$. In particular, for the indicator function I_A, $E(I_A) = P(A)$.

If $I \sim Bern(p)$, then $I^2 = I$ and so $E(I^2) = E(I) = p$. It follows that

$$Var(I) = E(I^2) - E(I)^2 = p - p^2 = pq.$$

If $I \sim Bern(p)$, then the mgf M_I is given by $M_I(t) = pe^t + q$.

A *binomial* rv $X \sim Bin(n, p)$ counts the number of successes in n independent Bernoulli trials. Thus, $X = I_1 + I_2 + \ldots I_n$, where $I_1, I_2, \ldots I_n$ are independent, identically distributed (iid) rv's with each $I_i \sim Bern(p)$.

We use this to compute the mean, variance and mgf without using the pmf of X:

$$\mu_X = np, \quad Var(X) = npq, \quad M_X(t) = (pe^t + q)^n.$$

The rv X has range $\{0, 1, \ldots, n\}$ and pmf given by $f_X(k) = \binom{n}{k} p^k q^{n-k}$. for $k = 0, 1, \ldots, n$.

The sequence $I_1, \ldots, I_n$ is the list of *outcomes* of the n trials. If $X = k$, then the list has k 1's, k successes, and $n - k$ 0's, $n - k$ failures. Since $I_i = 1$ has probability p and the I_i's are independent, the probability of such a list is $p^k q^{n-k}$.

There are 2^n such lists, which is the number of ways of filling in n slots each with either a 1 or a 0. To obtain a list with k 1's we choose k of the slots to be where the successes occur. There are $\binom{n}{k}$ ways of choosing a subset of size k from among the n slots. That is, there are $\binom{n}{k}$ lists with exactly k successes. So

$$P(X = k) = \binom{n}{k} p^k q^{n-k}.$$

If Y counts the number of failures before the first success of a sequence of independent $Bern(p)$ trials, then Y is a *geometric* rv written $Y \sim Geom(p)$.

The range of Y is the set of non-negative integers and $f_Y(k) = q^k p$ for $k = 0, 1, \ldots$.

That is, $Y = k$ when the list of outcomes begins with k 0's and then a 1.

The mgf is given by

$$M_Y(t) = \sum_{k=0}^{\infty} q^k p e^{tk} = p \sum_{k=0}^{\infty} (qe^t)^k = \frac{p}{1 - qe^t}.$$

$$E(Y) = M_Y'(0) = \frac{q}{p}, \quad E(Y^2) = M_Y''(0) = \frac{q + q^2}{p^2}.$$

So $Var(Y) = \frac{q}{p^2}$.

Because $Y \geq k$ exactly when the first k trials are failures, the survival function $G_Y(k) = P(Y \geq k) = q^k$.

Urn Examples, Hypergeometric Distribution

An urn contains r red balls and b blue balls for a total of $N = r + b$. It is assumed that when a choice is made, all of the balls are equally likely. If X is the number of red balls chosen given n independent choices, then there are two different cases:

Sampling WITH Replacement There are n independent choices each $Bern(p)$ with $p = \frac{r}{N}$. So $X \sim Bin(n, p)$ and $f_X(k) = \binom{n}{k} \cdot \frac{r^k b^{n-k}}{N^n}$.

Sampling WITHOUT Replacement Now it is necessary that $n \leq N$ and $X \leq r$. This time X has a *hypergeometric* distribution. There are $\binom{N}{n}$ ways of choosing n balls from the urn. Among these we want to count the number with exactly k red balls. There are $\binom{r}{k}$ ways of choosing k red balls and $\binom{b}{n-k}$ to choose there remaining blue balls. Therefore $f_X(k) = \binom{r}{k} \cdot \binom{b}{n-k} \div \binom{N}{n}$.

To see these results in parallel, we consider a list of the n outcomes in order which contains exactly k red balls so that the remaining $n-k$ are blue. Observe that there are $\binom{n}{k}$ such lists.

Sampling WITH Replacement The probability of a particular list is $p^k q^{n-k} = \frac{r^k b^{n-k}}{N^n}$. So

$$f_X(k) = \binom{n}{k} \cdot \frac{r^k b^{n-k}}{N^n},$$

Sampling WITHOUT Replacement This time the probability of a particular list is $\frac{r_k b_{n-k}}{N_n}$. So

$$f_X(k) = \binom{n}{k} \cdot \frac{r_k b_{n-k}}{N_n}.$$

You should check that this formula is the same as the one given above (write everything in terms of factorials).

Thus, the hypergeometric distribution is the analogue of the binomial distribution, but without replacement. Back to the urn and let Y be the number of blue balls which are drawn before the first red ball is chosen.

Sampling WITH Replacement There is a sequence of independent choices each $Bern(p)$ with $p = \frac{r}{N}$. So $Y \sim Geom(p)$ with

$$f_Y(k) = \frac{b^k r}{N^{k+1}}.$$

Sampling WITHOUT Replacement Now it is necessary that $Y \leq b$. For $Y = k$ it is first necessary that the first k choices are all blue. This has probability $\binom{b}{k} \div \binom{N}{k} = \frac{b_k}{N_k}$. This is the probability that $Y \geq k$. Assuming this, $Y = k$ when the next choice out of the remaining $N - k$ balls is red. So the conditional probability is $\frac{r}{N-k}$.

$$f_Y(k) = [\binom{b}{k} \cdot r] \div [\binom{N}{k} \cdot (N - k)] = \frac{b_k r}{N_{k+1}}$$

Again to see these results in parallel, we consider a list of the outcomes in order with exactly k blue balls first and then 1 red ball. There is one such list of length $k+1$

Sampling WITH Replacement The probability of a particular list is $q^k p = \frac{b^k r}{N^{k+1}}$. So

$$f_Y(k) = \frac{b^k r}{N^{k+1}}.$$

Sampling WITHOUT Replacement This time the probability of a particular list is $\frac{b_k r}{N_{k+1}}$. So

$$f_Y(k) = \frac{b_k r}{N_{k+1}}.$$

Consider the special case when $r = 1$ (as in Exercise 1.6.2). For Sampling WITH Replacement we have

$$f_Y(k) = \frac{(N-1)^k}{N^{k+1}},$$

But for Sampling WITHOUT Replacement we have for $k = 0, \ldots, N-1$

$$f_Y(k) = \frac{(N-1)_k}{N_{k+1}} = \frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdots \frac{N-k}{N-k+1} \frac{1}{N-k} = \frac{1}{N}.$$

To see why this uniformity holds, imagine pulling all of the balls out, one at a time and lining them up. $Y = k$ when the red ball is in slot $k+1$. There are N equally likely slots in which the red ball could land as all of the rearrangements are equally likely. So the probability that it lands in slot $k+1$, i.e. that $Y = k$ is $\frac{1}{N}$.

Sec 3.2: Poisson Distributions

An rv X is *Poisson-m*, written $X \sim Poiss(m)$, when it has range the non-negative integers and $f_X(k) = e^{-m}\frac{m^k}{k!}$ for $k = 0, 1, \ldots$.

The mgf is given by:

$$M_X(t) = \sum_{k=0}^{\infty} e^{tk} e^{-m} \frac{m^k}{k!} = e^{-m} \sum_{k=0}^{\infty} \frac{(me^t)^k}{k!} = e^{m(e^t-1)}.$$

$$E(X) = M_X'(0) = m, \quad E(X^2) = M_X''(0) = m^2 + m.$$

So that $Var(X) = m$.
From the mgf we see that if $X_1 \sim Poiss(m_1)$ and $X_2 \sim Poiss(m_2)$ and X_1 and X_2 are independent, then $X_1 + X_2 \sim Poiss(m_1 + m_2)$.

We illustrate the series methods by computing not $E(X^n)$ but $E(X_n)$ with $X_n = X \cdot (X-1) \ldots (X-(n-1))$ for some positive integer n. $E(X_n) = \sum_{k=0}^{\infty} k_n e^{-m} \frac{m^k}{k!}$. Notice that $k_n = 0$ for $k = 0, \ldots, n-1$. So

$$E(X_n) = \sum_{k=n}^{\infty} k_n e^{-m} \frac{m^k}{k!} = \sum_{k=n}^{\infty} \frac{k!}{(k-n)!} e^{-m} \frac{m^k}{k!}$$

$$= m^n \sum_{k=n}^{\infty} e^{-m} \frac{m^{k-n}}{(k-n)!} = m^n \sum_{j=0}^{\infty} e^{-m} \frac{m^j}{j!} = m^n.$$

So with $n = 1$, $E(X) = m$.

With $n = 2$, $E(X^2) - E(X) = E(X(X-1)) = m^2$.

$$Var(X) = E(X(X-1)) + E(X) - E(X)^2 = m^2 + m - m^2 = m.$$

Uniform Distributions

An rv X is *uniform* on an interval (a, b), written $X \sim Unif(a, b)$ when the pdf $f_X(x)$ is constant for $x \in (a, b)$ and is 0 elsewhere. To obtain a pdf, the constant must be the reciprocal of the length of the interval, $\frac{1}{b-a}$.

If $U \sim Unif(0, 1)$ then by direct computation, $E(U) = \frac{1}{2}$ and $E(U^2) = \frac{1}{3}$ so that $Var(U) = \frac{1}{12}$.

The mgf is given by $M_U(t) = \frac{e^t - 1}{t}$ for $t \neq 0$ (and $M_U(0) = 1$).

If $X \sim Unif(a, b)$ then $U = \frac{1}{b-a}(X - a) \sim Unif(0, 1)$. So $E(X) = \frac{a+b}{2}$, $Var(X) = \frac{(b-a)^2}{12}$.

Universality of the Uniform

If $U \sim Unif(0,1)$, then the cdf satisfies $F_U(x) = \int_0^x 1\, du = x$, for $0 < x < 1$ and this cdf characterizes the uniform distribution on $(0,1)$.

Now suppose that X is a continuous rv with support (a,b) for $-\infty \leq a < b \leq \infty$. That is, the density $f_X(x)$ is positive on (a,b) and is 0 elsewhere. So $F_X : (a,b) \to (0,1)$ is a strictly increasing function.

If we define $\hat{U} = F_X(X)$, then for $x \in (0,1)$:

$$F_{\hat{U}}(x) = P(F_X(X) \leq x) = P(X \leq F_X^{-1}(x)) = F_X(F_X^{-1}(x)) = x.$$

That is, $\hat{U} = F_X(X) \sim Unif(0,1)$.

So by applying the cdf of an arbitrary continuous rv to the rv itself we obtain a uniform rv on $(0,1)$.

On the other hand, if $F : (a, b) \to (0, 1)$ is a strictly increasing continuous function with $F^{-1} : (0, 1) \to (a, b)$ and U is a $Unif(0, 1)$ rv, then we define $\hat{X} = F^{-1}(U)$.

$$F_{\hat{X}}(x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x).$$

That is, $\hat{X}$ is a continuous rv with $F_{\hat{X}} = F$.

So by using a uniform rv we can build a continuous rv with an arbitrary increasing function $F : (a, b) \to (0, 1)$ as its cdf.

Exponential Distributions

An exponential rv $X \sim Exp(\lambda)$ has range $(0, \infty)$ and pdf $f_X(x) = \lambda e^{-\lambda x}$.

$$\int ze^{-z}dz = -(z+1)e^{-z}, \quad \int z^2e^{-z}dz = -(z^2+2z+2)e^{-z}.$$

So if $Z \sim Exp(1)$, $E(Z) = Var(Z) = 1$. The mgf is given by

$$M_Z(t) = \int_0^\infty e^{tx}e^{-x}dx = \frac{1}{1-t}$$

for $t < 1$.

If $X \sim Exp(\lambda)$, then

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}.$$

So if $X \sim Exp(\lambda)$, and $Z = \lambda X$, then

$$F_Z(x) = P(Z \leq x) = P(X \leq x/\lambda) = F_X(x/\lambda) = 1 - e^{-x}.$$

Thus, $Z = \lambda X \sim Exp(1)$. Hence, $E(X) = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$

$$M_X(t) = M_Z(\frac{t}{\lambda}) = \frac{\lambda}{\lambda - t} \quad \text{for } t < \lambda.$$

Sec. 2.1: Joint Distribution for Two Random Variables: Discrete Case

We regard a pair of rv's (X, Y) as a map from the sample space to the plane $\mathbb{R}^2$.

The *Joint PMF* of a discrete pair X, Y is given by

$$f_{X,Y}(x, y) \;=\; P((X, Y) = (x, y)) \;=\; P(X = x, Y = y),$$

with $(x, y) \in \mathbb{R}^2$. That is, $f_{X,Y}(x, y)$ is the probability of the event that, simultaneously, X is x and Y is y.

For a subset $A \subset \mathbb{R}^2$, the event $(X, Y) \in A$ has probability

$$P((X, Y) \in A) \;=\; \sum_{(x,y)\in A} f_{X,Y}(x, y).$$

Just as we usually compute double integrals as iterated integrals, we usually compute such sums by first fixing x and sum over the y's such that (x, y) is in A, obtaining a value for each x and then sum over all $x \in \mathbb{R}$.

For example, the PMF of X is obtained from the joint PMF as the *marginal PMF* $f_X(x) = P(X = x) = \sum_y f_{X,Y}(x, y)$.

Notice that $f_X(x) = 0$ if and only if $f_{X,Y}(x, y) = 0$ for all $y \in \mathbb{R}$.

The conditional PMF of Y given X is

$$f_{Y|X}(y|x) = P(Y = y|X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f_{X.Y}(x, y)}{f_X(x)}.$$

It is still a function of the pair (x, y) but we write $y|x$ as a reminder that it is the probability that $Y = y$, assuming that $X = x$. While $f_{X,Y}(x, y)$ is defined for all $(x, y) \in \mathbb{R}^2$, $f_{Y|X}(y|x)$ is only defined when x is in the support of X. That is when $f_X(x) > 0$.

Recall that Bayes' Rule says

$$P(X = x|Y = y) = \frac{P(Y = y|X = x) \cdot P(X = x)}{P(Y = y)}.$$

which says in PMF notation

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x) \cdot f_X(x)}{f_Y(y)}.$$

Sec. 2.1: Joint Distribution for Two Random Variables: Continuous Case

For a pair of continuous distributions (X, Y) we have the Joint PDF $f_{X,Y}(x, y)$. This time it is the density, probability per unit area, so that $f_{X,Y}(x, y)dxdy$ is the probability of a little $dx \times dy$ rectangle containing the point (x, y). That is, we think of

$$P(X \in [x + dx], Y \in [y + dy]) = f_{X,Y}(x, y)dxdy.$$

For a subset $A \subset \mathbb{R}^2$ we integrate to get the probability.

$$P((X, Y) \in A) = \int_A f_{X,Y}(x, y)dxdy.$$

As before we use iterated integrals and, in particular, we obtain the density of X as the marginal density by integrating away y.

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

The conditional PDF of Y given X is

$$f_{Y|X}(y|x) = \frac{f_{X.Y}(x,y)}{f_X(x)}.$$

It is helpful to use the following version of the actual conditional probability formula:

$$f_{Y|X}(y|x)dy = \frac{f_{X.Y}(x,y)dxdy}{f_X(x)dx}.$$

This indicates the usefulness of the notation $f_{Y|X}(y|x)$ as compared with $f_{X.Y}(x,y)$. The latter is a density per unit area, but $f_{Y|X}(y|x)$ like $f_Y(y)$ is a density per unit length, but the density function in general depends on the value of x.

As usual we can rewrite this as

$$f_{X.Y}(x,y)dxdy = (f_{Y|X}(y|x)dy) \cdot (f_X(x)dx).$$

In particular, we have the continuous version of the LOTP

$$f_Y(y)dy = \int_{x=-\infty}^{x=\infty} f_{X,Y}(x,y)dxdy = \int_{x=-\infty}^{x=\infty} (f_{Y|X}(y|x)dy \cdot f_X(x))dx.$$

The continuous version of Bayes Rule is best thought of as

$$f_{Y|X}(y|x)dy = \frac{(f_{X|Y}(x|y)dx) \cdot (f_Y(y)dy)}{f_X(x)dx}.$$

Sec. 2.5: Independent Random Variables

Random variables X, Y are *independent* when their joint distribution is the product of the marginals.

In the discrete case, this means $f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$ for all $(x, y) \in \mathbb{R}^2$.

This is equivalent to $f_{Y|X}(y|x) = f_Y(y)$ for all $y \in \mathbb{R}$ and all values x in the range of X (so that $f_X(x) > 0$).

In the continuous case, this means $f_{XY}(x, y)dxdy = f_X(x)dx \cdot f_Y(y)dy$ for all $(x, y) \in \mathbb{R}^2$.

This is equivalent to $f_{Y|X}(y|x)dy = f_Y(y)dy$ for all $y \in \mathbb{R}$ and all values x in the support of X (so that $f_X(x) > 0$).

Sec. 2.3: Expectation

For a function $g(x, y)$ the rv $g(X, Y)$ has expectation using the joint distribution.

$$E(g(X,Y)) = \begin{cases} \sum_{(x,y)} g(x,y) f_{X,Y}(x,y) & \text{for } (X,Y) \text{ discrete,} \\ \int\int g(x,y) f_{X,Y}(x,y) dxdy & \text{for } (X,Y) \text{ continuous.} \end{cases}$$

We can now explain results given earlier:
$E(X + Y) = E(X) + E(Y)$ and, when X and Y are independent $E(X \cdot Y) = E(X) \cdot E(Y)$. We will just give the proofs in the continuous case. It is similar for the discrete case.

$$E(X+Y) = \int\int (x+y) f_{XY}(x,y)\,dxdy$$

$$= \int\int x f_{X,Y}(x,y)\,dxdy + \int\int y f_{X,Y}(x,y)\,dxdy.$$

$$\int\int x f_{X,Y}(x,y)\,dxdy = \int_{x=-\infty}^{x=\infty} x\Big(\int_{y=-\infty}^{y=\infty} f_{X,Y}(x,y)\,dy\Big)dx$$

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

Similarly, for $\int\int y f_{X,Y}(x,y)\,dxdy$.

$$E(X \cdot Y) = \int \int xyf_{X,Y}(x,y)dxdy$$

If X and Y independent, then

$$E(X \cdot Y) = \int \int xyf_X(x)f_Y(y)dxdy = \int yf_Y(y)[\int xf_X(x)dx]dy$$
$$= E(X) \cdot \int yf_Y(y)dy = E(X) \cdot E(Y).$$

For a function $g(x, y)$ we have the conditional expectation

$$E(g(x, Y)|X = x) = \begin{cases} \sum_y g(x, y) f_{Y|X}(y|x) & \text{for } (X, Y) \text{ discrete,} \\ \int g(x, y) f_{Y|X}(y|x) dy & \text{for } (X, Y) \text{ continuous.} \end{cases}$$

In each case we are summing or integrating away the y variable and we are left with a function of x. When we compose this function of x with the rv X, we obtain the new rv $E(g(Y)|X)$ which is a function of the rv X.

Notice that for $h(x)g(x, y)$ we obtain

$$E(h(X)g(X, Y)|X) \;=\; h(X)E(g(X, Y)|X).$$

HMC Theorem 2.3.1 says

$$E(g(Y)) = E(E(g(Y)|X))).$$

That is, from $g(Y)$ we obtain $E(g(Y)|X = x)$ which is a function of x.

The rv $g(Y)$ obtained from $g(y)$ and is a function of Y while the rv $E(g(Y)|X))$ obtained from $E(g(Y)|X = x)$ and is a function of X. But they have the same expectation.
We call this the computation of $E(g(Y))$ by first conditioning on X.

We will do the computation for the continuous case.

Recall that $f_{Y|X}(y|x)f_X(x) = f_{X,Y}(x,y)$.

$$E(E(g(Y)|X))) = \int_x [\int_y g(y) f_{Y|X}(y|x)dy] f_X(x)dx$$

$$= \int_x [\int_y g(y) f_{Y|X}(y|x) f_X(x)dy]dx$$

$$= \int_x [\int_y g(y) f_{X,Y}(x,y)dy]dx$$

$$= \int_y [\int_x g(y) f_{X,Y}(x,y)dx]dy$$

$$= \int_y g(y)[\int_x f_{X,Y}(x,y)dx]dy$$

$$= \int_y g(y) f_Y(y)dy = E(g(Y)).$$

If we take the variance of Y with respect to the conditional distribution of $y|x$ we obtain the conditional variance:

$$Var(Y|X) = E([Y - E(Y|X)]^2|X) = E(Y^2|X) - E(Y|X)^2.$$

This is an rv, it is a function of X not to be confused with the number

$$\begin{aligned} Var(E(Y|X)) &= E(E(Y|X)^2) - E(E(Y|X))^2 \\ &= E(E(Y|X)^2) - E(Y)^2. \end{aligned}$$

In fact, the second part of HMC Theorem 2.3.1 shows:

$$Var(Y) \;=\; E(Var(Y|X)) \;+\; Var(E(Y|X)).$$

Let us see why this is true.

First observe: $E(Y^2|X) - E(Y)^2$ is a function of X whose expectation is $E(Y^2) - E(Y)^2 = Var(Y)$.

But

$$E(Y^2|X) - E(Y)^2 =$$
$$[E(Y^2|X) - E(Y|X)^2] + [E(Y|X)^2 - E(Y)^2]$$
$$= [Var(Y|X)] + [E(Y|X)^2 - E(Y)^2].$$

These are all function of X. When we take the expected values we get

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)].$$

Sec. 2.4: Covariance and Correlation

For an rv X, $Var(X) = E((X - \mu_X)^2)$, and the *standard deviation* is defined to be $\sigma_X = \sqrt{Var(X)}$. For a pair of rv's X, Y, the covariance and correlation are defined by:

$$Cov(X, Y) = E((X - \mu_X)(Y - \mu_Y)) = E(XY) - \mu_X \mu_Y,$$
$$Corr(X, Y) = Cov(X, Y) \div [\sigma_X \sigma_Y].$$

Thus, $Var(X) = Cov(X, X)$. If $Z = tX + sY$ then $\mu_Z = t\mu_X + s\mu_Y$ and

$$\begin{aligned} Var(Z) &= t^2 Var(X) + s^2 Var(Y) + 2stCov(X, Y) \\ &= t^2\sigma_X^2 + s^2\sigma_Y^2 + 2st\sigma_X\sigma_Y\rho, \end{aligned}$$

where $\rho = Corr(X, Y)$.

From the formula above we see that always $|\rho| \leq 1$.

With $Z = tX + Y$ we have for all real numbers t, that $t^2\sigma_X^2 + 2t\sigma_X\sigma_Y\rho + \sigma_Y^2 = Var(Z) \geq 0$. So this quadratic function of t cannot have two different real roots.

In the quadratic formula for the roots of $At^2 + Bt + C = 0$, this means that the discriminant $B^2 - 4AC \leq 0$.

So

$$4(\sigma_X\sigma_Y\rho)^2 - 4\sigma_X^2\sigma_Y^2 \leq 0.$$

This means $(\rho)^2 \leq 1$ and so $-1 \leq \rho \leq 1$.

Sec. 2.3: Extension to Several Variables

We think of $(X_1, \ldots, X_n)$ as an rv with values in $\mathbb{R}^n$.

In the discrete case there is a joint pmf and in the continuous case a joint pdf.

Again independence means that the joint pmf or pdf is the product of the marginals.

We will frequently use the case where the list $X_1, \ldots, X_n$ are iid's (= independent, identically distributed rv's). So if $f_X(x)$ is the pmf or pdf of any of them, then the joint pmf or pdf is given by

$$f_{(X_1,\ldots,X_n)}(x_1, \ldots, x_n) = f_X(x_1) f_X(x_2) \ldots f_X(x_n).$$

Sec. 1.7.2[1.7.1], 2.2, 2.7: Transformations

Suppose that $z : (a, b) \to (c, d)$ is a differentiable function and that X is a continuous rv with range in (a, b). Let $Z = z(X)$. We assume that z' is never 0. If $z' > 0$ then z is an increasing function which preserves inequalities. If $z' < 0$ then z is a decreasing function which reverses inequalities.

$$X \le x \quad \Longleftrightarrow \quad \begin{cases} Z \le z(x) & \text{if } z' > 0, \\ Z \ge z(x) & \text{if } z' < 0. \end{cases}$$

$$F_X(x) \;=\; \begin{cases} F_Z(z(x)) & \text{if } z' > 0, \\ G_Z(z(x)) & \text{if } z' < 0. \end{cases}$$

Differentiating we obtain the u substitution formula:

$$f_X(x)dx \;=\; f_Z(z(x)) \cdot |z'(x)|dx \;=\; f_Z(z)dz.$$

Linear Change of Variables

We will omit the general results about change of variable, but we will need the special case of a linear change of variables from (x, y) to (u, v)

$$\begin{aligned} u &= ax + by \\ v &= cx + dy \end{aligned}$$

In order that the change be reversible, it is necessary that the determinant $J = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$ be nonzero. Then we can solve using Cramer's rule to get

$$\begin{aligned} Jx &= du + -bv \\ Jy &= -cu + av. \end{aligned}$$

Notice that the coefficient matrix is the same as the Jacobian matrix of partial derivatives

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} \partial u/\partial x & \partial u/\partial y \\ \partial v/\partial x & \partial v/\partial y \end{pmatrix}.$$

So we will write $\frac{\partial(u,v)}{\partial(x,y)}$ for the determinant J.
By analogy with the one-dimensional change of variables we write $dudv = |\frac{\partial(u,v)}{\partial(x,y)}|dxdy$ (note the absolute value - these are comparing areas). With $(U, V) = (aX + bY, cX + dY)$

$$\begin{aligned} f_{X,Y}(x,y)dxdy &= f_{U,V}(ax+by, cx+dy) \cdot |\frac{\partial(u,v)}{\partial(x,y)}|dxdy \\ &= f_{U,V}(u,v)dudv. \end{aligned}$$

Observe that

$$tU + sV \;=\; (at + cs)X + (bt + ds)Y.$$

Exponentiate and take the expected values to get for the MGF's

$$M_{U,V}(t, s) \;=\; M_{X,Y}(at + cs, bt + ds).$$

For more variables, the formulas are similar.

Convolution

Suppose that X, Y are independent r.v.s with CDF's f_X, f_Y. Let $Z = X + Y$. We have already obtained the formula for the pdf of Z in Exercise 2.1.7[2.1.6], see also Exercise 2.2.5. If $z = x + y, t = y$. We consider the change of variables $(z, t) \to (x, y)$ by $x = z - t, y = t$. The determinant $J = 1$ and so we have

$$f_{Z,T}(z,t)dzdt = f_{X,Y}(z-t,t)1dzdt = f_X(z-t)f_Y(t)dzdt.$$

Alternatively, we can write

$$f_{Z,T}(z,t)dzdt = P(X+Y \in [z+dz], Y \in [t+dt])$$
$$= P(X \in [z-t+dz],\ Y \in [t+dt]) =$$
$$P(X \in [z-t+dz]) \cdot P(Y \in [t+dt]) = f_X(z-t)f_Y(t)dzdt.$$

and so

$$f_{Z,T}(z,t)dzdt \;=\; f_X(z-t)f_Y(t)dzdt.$$

We obtain the PDF of $Z = X + Y$ as the marginal PDF by integrating away the t variable.

$$f_Z(z)dz = \int_{t=-\infty}^{t=\infty} f_X(z-t)f_Y(t)dt.$$

The function f_Z is called the *convolution* of the functions f_X and f_Y.

Sec. 3.3: The Gamma Distribution

The Gamma function is defined by an integral for $\alpha > 0$.

$$\Gamma(\alpha) = \int_0^\infty x^\alpha e^{-x} \frac{dx}{x} = \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

In particular, we see that $\Gamma(1) = 1$.

From integration by parts we obtain the identity $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ for all $\alpha > 0$. Let $u = x^\alpha, dv = e^{-x}dx$ so that $du = \alpha x^{\alpha-1}dx, v = -e^{-x}$.

$$\int_0^\infty x^\alpha e^{-x} dx \;=\; -x^\alpha e^{-x}|_0^\infty + \alpha \int_0^\infty x^{\alpha-1} e^{-x} dx.$$

In particular, it follows that $\Gamma(n) = (n-1)!$.

Dividing by $\Gamma(\alpha)$ we obtain the $\Gamma(\alpha, 1)$ rv X with pdf

$$f_X(x)dx = \frac{1}{\Gamma(\alpha)} x^{\alpha} e^{-x} \frac{1}{x} dx.$$

If $X \sim \Gamma(\alpha, 1)$ and $Y = \beta X$ with $\beta > 0$ we say that $Y \sim \Gamma(\alpha, \beta)$. From such a scale change we have

$$f_Y(y)dy = \frac{1}{\beta} f_X(\frac{y}{\beta})dy = \frac{1}{\Gamma(\alpha)} (\frac{y}{\beta})^{\alpha} e^{-\frac{y}{\beta}} \frac{1}{y} dy.$$

We compute the MGF of Y

$$E(e^{tY}) = \frac{1}{\Gamma(\alpha)} \int_0^\infty e^{ty} (\frac{y}{\beta})^\alpha e^{-\frac{y}{\beta}} \frac{dy}{y}.$$

Let $\mu = \frac{1}{\beta} - t$ so that $\frac{1}{\beta} = \mu + t$. The integral becomes

$$\frac{1}{\Gamma(\alpha)} \int_0^\infty ((\mu + t)y)^\alpha e^{-\mu y} \frac{dy}{y}$$

$$= \frac{(\mu + t)^\alpha}{\mu^\alpha} \frac{1}{\Gamma(\alpha)} \int_0^\infty (\mu y)^\alpha e^{-\mu y} \frac{dy}{y}$$

This is just $\frac{(\mu+t)^\alpha}{\mu^\alpha} = (\frac{1}{1-\beta t})^\alpha$ since the rest is the integral of the $\Gamma(\alpha, 1/\mu)$ pdf. This requires $\mu > 0$ and so $\beta t < 1$.

Thus, if $Y \sim \Gamma(\alpha, \beta)$, then

$$M_Y(t) = E(e^{tY}) = (\frac{1}{1-\beta t})^{\alpha} \quad \text{for } t < \beta^{-1}$$

.

$$M_Y(t) = (1-\beta t)^{-\alpha},\ M'(t) = \alpha\beta(1-\beta t)^{-(\alpha+1)},$$
$$M''_Y(t) = \alpha(\alpha+1)\beta^2(1-\beta t)^{-(\alpha+2)}.$$

So $E(Y) = M'_Y(0) = \alpha\beta, E(Y^2) = M''_Y(0) = \alpha(\alpha+1)\beta^2$ and so $Var(Y) = \alpha\beta^2$.

Finally, if X and Y are independent r.v.'s

- $X \sim \Gamma(\alpha_1, \beta), Y \sim \Gamma(\alpha_2, \beta) \Rightarrow X + Y \sim \Gamma(\alpha_1 + \alpha_2, \beta)$.

Observe that the $\Gamma(1, \beta)$ distribution is just the $Expo(\lambda)$ distribution with $\lambda = \frac{1}{\beta}$. So it follows that the sum of n independent $Expo(\lambda)$ r.v.'s has distribution $\Gamma(n, 1/\lambda)$.

Sec. 3.3: The Chi-Squared Distribution

The special case of the Gamma Distribution with $\alpha = r/2$ and $\beta = 2$ with r a positive integer is called the χ^2 distribution with r degrees of freedom. So if $Y \sim \chi^2(r)$

$$f_Y(y)dy = \frac{1}{\Gamma(r/2)}(\frac{y}{2})^{r/2}e^{-\frac{y}{2}}\frac{1}{y}dy \quad \text{with } M_Y(t) = (\frac{1}{1-2t})^{r/2}.$$

So $E(Y) = r, Var(Y) = 2r$.

Sec. 3.4: The Normal Distribution

The standard *normal distribution* r.v. $Z \sim \mathcal{N}(0,1)$ has PDF on $\mathbb{R}$ given by

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

Just as we use special notation ϕ for the density function f_Z of the standard normal, we use Φ for the cdf of the standard normal, so that

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt.$$

There is a trick for computing $I = \int_{-\infty}^{\infty} e^{-x^2/2}\, dx$ which uses polar coordinates.

$$I^2 = \int_{-\infty}^{\infty} e^{-x^2/2}\, dx \cdot \int_{-\infty}^{\infty} e^{-y^2/2}\, dy =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\, dxdy = \int_0^{2\pi}\int_0^{\infty} re^{-r^2}\, dr\, d\theta$$

$$= \int_0^{2\pi}\int_0^{\infty} e^{-u}\, du\, d\theta = 2\pi.$$

So $\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = I = \sqrt{2\pi}$.

If $Z \sim \mathcal{N}(0,1)$, then

$$E(e^{tZ}) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$
$$= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(z-t)^2/2} dz = e^{t^2/2}.$$

$$M_Z(t) = e^{t^2/2}, \quad M_Z'(t) = te^{t^2/2}, \quad M_Z''(t) = e^{t^2/2} + t^2 e^{t^2/2}.$$

So $E(Z) = M_Z'(0) = 0$ and $Var(Z) = E(Z^2) = M_Z''(0) = 1$.

If $X = \mu + \sigma Z$, then $E(X) = \mu$ and $Var(X) = \sigma^2$. We write $X \sim \mathcal{N}(\mu, \sigma^2)$. So if $X \sim \mathcal{N}(\mu, \sigma^2)$ we convert to the standard normal by $Z = \frac{X-\mu}{\sigma}$ or $X = \sigma Z + \mu$ so that

$$f_X(x)dx = f_Z(z)dz = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})dx = \frac{1}{\sigma\sqrt{2\pi}}e^{-(\frac{x-\mu}{\sigma})^2/2}dx.$$

$$P(X < \mu + \sigma z) = \Phi(z),$$
$$P(|X - \mu| < \sigma z) = P(\mu - \sigma z < X < \mu + \sigma z) = \Phi(z) - \Phi(-z).$$

$$E(e^{tX}) = e^{\mu t + \sigma^2 t^2/2}.$$

From the mgf, it follows that any linear combination of independent normal rv's is again normal. See HMC Theorem 3.4.2.

The importance of the Chi-Squared distribution comes from its connection with the normal.

HMC Theorem 3.4.1: If $Z \sim \mathcal{N}(0,1)$, then $V = Z^2 \sim \chi^2(1)$.
Proof: For $v > 0$:

$$F_V(v) = P(-\sqrt{v} \leq Z \leq \sqrt{v}) = 2\int_0^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

Now use the change of variable: $t = \sqrt{u}$ to get:

$$F_V(v) = \int_0^{v} \frac{1}{\sqrt{2\pi u}} e^{-u/2} du.$$

This means that $f_V(v)dv = \frac{1}{\sqrt{\pi}}(\frac{v}{2})^{1/2} e^{-\frac{v}{2}} \frac{1}{v} dv$ for $v > 0$ and $= 0$ otherwise.
Recall that the density for a $\chi^2(1)$ variable is $\frac{1}{\Gamma(1/2)}(\frac{y}{2})^{1/2} e^{-\frac{y}{2}} \frac{1}{y} dy$.

This shows that $V \sim \chi^2(1)$, and, in passing, shows that $\Gamma(1/2) = \sqrt{\pi}$.

It follows that if $Z_1, \ldots, Z_r$ are independent $\mathcal{N}(0,1)$ rv's then

$$V_r \;=\; \sum_{i=1}^{r} Z_i^2 \;\sim \chi^2(r).$$

If $W \sim \mathcal{N}(0,1)$, $V \sim \chi^2(r)$ and W and V are independent, then $T = W \div \sqrt{V/r}$ is said to have a t-distribution.

Bivariate Normal

A random vector $(X_1, \ldots, X_n)$ is said to be a *multivariate normal* MVN when any mixture $t_1X_1 + \ldots, +t_nX_n$ is a normal r.v., including the possibility of the degenerate case of a constant with variance zero. We will restrict our attention to the case $n = 2$.

Recall that for $Z \sim \mathcal{N}(0, 1)$ the MGF $M_Z(t) = E(e^{tZ}) = e^{t^2/2}$. If $N \sim \mathcal{N}(\mu, \sigma^2)$ then $Z = (N - \mu)/\sigma \sim \mathcal{N}(0, 1)$ and so

$$M_N(t) = E(e^{t(\mu+\sigma Z)}) = e^{\mu t} e^{\sigma^2 t^2/2} = e^{E(N)t+Var(N)t^2/2}.$$

So with $t = 1$ we have

$$E(e^N) = e^{E(N)+Var(N)/2}.$$

If (X, Y) is bivariate normal, then $N = tX + sY$ is normal with $E(N) = t\mu_X + s\mu_Y$ and $Var(N) = \sigma_X^2 t^2 + \sigma_Y^2 s^2 + 2\sigma_X\sigma_Y\rho st$, where μ_X, μ_Y are the means and σ_X^2, σ_Y^2 are the variances of X and Y respectively and $\rho = Corr(X, Y)$. It follows that

$$M_{X,Y}(t, s) = E(e^{tX+sY}) = e^{t\mu_X+s\mu_Y+\frac{1}{2}(\sigma_X^2 t^2+\sigma_Y^2 s^2+2\sigma_X\sigma_Y\rho st)}.$$

The distribution of a bivariate normal (X, Y) is thus determined by $(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ correlation coefficient.

In particular, if $\rho = 0$ then $M_{X,Y}(t, s) = M_X(t) \cdot M_Y(s)$ which implies that X and Y are independent.

Independent rv's are always uncorrelated, but the converse is not usually true. For normals it is true.]

Just as we can begin with a standard normal Z and obtain an arbitrary normal $N \sim \mathcal{N}(\mu, \sigma^2)$ as $\mu + \sigma Z$, we can similarly build an arbitrary bivariate normal by starting with Z, W a pair of i.i.d standard normals.

Let $X = \sigma_1 Z$, $Y = \sigma_2(\rho Z + \bar{\rho} W)$ with $\bar{\rho} = \sqrt{1 - \rho^2}$.
X and Y have mean zero, and

$$Var(X) = \sigma_1^2, \quad Var(Y) = \sigma_2^2[\rho^2 Var(Z) + \bar{\rho}^2 Var(W)] = \sigma_2^2.$$

$$Cov(X, Y) = \sigma_1 \sigma_2 [\rho Var(Z) + \bar{\rho} Cov(Z, W)] = \sigma_1 \sigma_2 \rho.$$

By adding the constant vector (μ_1, μ_2) to (X, Y) we obtain a bivariate normal with parameters $(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

We obtain the joint PDF for X, Y by using the change of variables formula.

$$\begin{array}{rcccc} Z & = & \frac{1}{\sigma_1}X & + & 0 \\ W & = & -\frac{\rho}{\sigma_1\bar{\rho}}X & + & \frac{1}{\sigma_2\bar{\rho}}Y \end{array}$$

Since $f_{Z,W}(z,w) = \frac{1}{2\pi}exp[-\frac{1}{2}(z^2+w^2)]$ and $\frac{\partial(z,w)}{\partial(x,y)} = \frac{1}{\sigma_1\sigma_2\bar{\rho}}$, we get that $f_{X,Y}(x,y) =$

$$\frac{1}{\sigma_1\sigma_2\bar{\rho}2\pi}exp[-\frac{1}{2}((\frac{x}{\sigma_1})^2+(-\frac{\rho x}{\bar{\rho}\sigma_1}+\frac{y}{\bar{\rho}\sigma_2})^2)]$$

$$= \frac{1}{\sigma_1\sigma_2\bar{\rho}2\pi}exp[-\frac{1}{2\bar{\rho}^2}((\frac{x}{\sigma_1})^2+(\frac{y}{\sigma_2})^2-\frac{2\rho xy}{\sigma_1\sigma_2})].$$

Sec. 3.6: Sample Mean,Sample Variance, Student's Theorem

If X has mean μ and variance σ^2, then for any c:

$$E((X-c)^2) = E([(X-\mu)+(\mu-c)]^2)$$
$$= E((X-\mu)^2)+(\mu-c)^2 = \sigma^2+(c-\mu)^2.$$

Let $X_1,\ldots,X_n$ be iid's each with mean μ and variance σ^2. Define $\bar{X}=\frac{1}{n}\sum_{i=1}^n X_i$ to be the *sample mean*. $E(\bar{X})=\mu$ and $Var(\bar{X})=\frac{\sigma^2}{n}$. Because $\sum_{i=1}^n[(X_i-\bar{X})(\bar{X}-\mu)]=0$,

$$\frac{1}{n}\sum_{i=1}^n(X_i-\mu)^2=\frac{1}{n}\sum_{i=1}^n[(X_i-\bar{X})+(\bar{X}-\mu)]^2$$
$$=(\frac{1}{n}\sum_{i=1}^n(X_i-\bar{X})^2) + (\bar{X}-\mu)^2.$$

Multiply by n and take expected value. Note that $E((\bar{X}-\mu)^2) = Var(\bar{X}) = \frac{\sigma^2}{n}$

We obtain $n\sigma^2 = E(\sum_{i=1}^{n}(X_i - \bar{X})^2) + \sigma^2$.

So if we defined the *sample variance* by $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ we have $E(S^2) = \sigma^2$.

Now assume the iid's are all $\mathcal{N}(\mu, \sigma^2)$. In that case, $\bar{X}$ and S^2 are independent. We omit the proof from HMC Theorem 3.6.3.

$$\frac{(n-1)S^2}{\sigma^2} + (\frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}})^2 = \sum_{i=1}^{n}(\frac{X_i-\mu}{\sigma})^2.$$

The right side is a $\chi^2(n)$ variable and the second term on the left is $\chi^2(1)$. It follows that $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

Finally, $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with $n-1$ degrees of freedom.

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

$$V = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$

Since Z and V are independent,
$Z \div \sqrt{V/(n-1)} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ is a t-distribution with $n-1$ degrees of freedom.

Sec. 4.4: Order Statistics

Let $X_1, \ldots, X_n$ be iid continuous rv's with pdf $f(x)dx$. The possibility of a tie $X_i = X_j$ for some $i \neq j$ has probability zero and so, discarding the possibility of a tie, we can rearrange the X_i's in order: $Y_1 < Y_2 < \cdots < Y_n$. That is, Y_1 is the minimum, Y_2 is the second smallest up to Y_n the maximum of the X_i's.

For the minimum Y_1 we have, using $G_X = 1 - F_X$:

$$G_{Y_1}(y_1) = P(Y_1 > y_1) = P(X_1 > y_1, \ldots, X_n > y_1) = G_X(y_1)^n.$$

Similarly, for the maximum $F_{Y_n}(y_n) = F_X(y_n)^n$.

Differentiating we obtain the pdf's:

$$f_{Y_1}(y_1) = nf_X(y_1)(1-F_X(y_1))^{n-1}, \quad f_{Y_n}(y_n) = nf_X(y_n)F_X(y_n)^{n-1}.$$

That is, the event $Y_1 \in [y_1 + dy_1]$ occurs when some $X_i \in [y_1 + dy_1]$ and the remaining $n-1$ X_j's are greater than y_1. The factor n is because there are n choices for X_i.

For $1 < k < n$ we similarly compute $f_{Y_k}(y_k)$. The event $Y_k \in [y_k + dy_k]$ occurs when some $X_i \in [y_k + dy_k]$ (n choices) and of the remaining $n-1$ X_j's exactly $k-1$ are less than y_k ($\binom{n-1}{k-1}$ choices) while the remaining ones are greater than y_k.

$$f_{Y_k}(y_k) = n\binom{n-1}{k-1} f_X(y_k) F_X(y_k)^{k-1} (1 - F_X(y_k))^{n-k}$$

$$= \frac{n!}{(k-1)!(n-k)!} f_X(y_k) F_X(y_k)^{k-1} (1 - F_X(y_k))^{n-k}.$$

You should similarly be able to see that for $1 < k < \ell < n$ the joint distribution of Y_k, Y_ℓ is given by:

$$f_{Y_k Y_\ell}(y_k, y_\ell) \;=\; n \cdot (n-1) \cdot \binom{n-2}{k-1}\binom{n-k-1}{\ell-k-1} \cdot$$
$$f_X(y_k) f_X(y_\ell) F_X(y_k)^{k-1} [F_X(y_\ell) - F_X(y_k)]^{\ell-k-1} [1 - F_X(y_\ell)]^{n-\ell}.$$

Observe that: $n \cdot (n-1) \cdot \binom{n-2}{k-1}\binom{n-k-1}{\ell-k-1} = \frac{n!}{(k-1)!(\ell-k-1)!(n-\ell)!}$.

Finally, the joint distribution of $Y_1, Y_2, \ldots, Y_n$ is obtained via the $n!$ choices of ordering for $X_1, \ldots, X_n$:

$$f_{Y_1 \ldots Y_n}(y_1, \ldots, y_n) dy_1 \ldots dy_n = n! f_X(y_1) \cdot \ldots f_X(y_n).$$

Sec. 4.1: Samples, Realizations and Statistics

In statistics we want to use data to identify an unknown pmf or pdf $f(x)$. The data we use is a *random sample* a sequence $X_1, \ldots, X_n$ iid rv's with the unknown distribution. The actual observed values $x_1, \ldots, x_n$ are the *realizations* of the sample.

A function $T = T(X_1, \ldots, X_n)$ is called a *statistic* of the sample. Once the values $x_1, \ldots, x_n$ of the sample have been observed, we obtain $t = T(x_1, \ldots, x_n)$ the *realization of the statistic*.

We will be considering two situations.

Case 1 (Point Estimation): We know the form of the pmf or pdf lies within a known family $\{f(x;\theta) : \theta \in \Omega\}$ and we want to know the true value of the unknown *parameter* θ which determines the distribution. To estimate, that is, to guess, a value for θ we use a statistic T which we call an *estimate* or a *point estimator* for θ. The statistic T is a function of the rv's, of the data, it may not depend on θ.

Case 2 (Nonparametric Estimation): In this case, $f(x)$ is completely unknown and we use the data to estimate the distribution.

Sec. 4.1.1[4.1]: Maximum Likelihood Estimate

For point estimation of the parameter θ, we use the notation $E_\theta(X)$ for the expected value with respect to $f(x;\theta)$. An estimator $T(X_1,\ldots,X_n)$ for θ is called an *unbiased estimator* when $E_\theta(T) = \theta$.

For example, with $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then $E(\bar{X}) = \mu_X$ and so the sample mean is an unbiased estimator of the true mean.

Similarly we saw that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ satisfies $E(S^2) = Var(X)$ and so S^2 is an unbiased estimator for the variance. Thus, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ has expected value $\frac{n-1}{n}Var(X)$ and so is biased.

With parameter θ the joint distribution is given by $f_{X_1,\ldots,X_n}(x_1,\ldots,x_n) = \prod_{i=1}^{n} f_X(x_i;\theta)$.

We view this as a function of θ defining the *likelihood function* $L(\theta)$ by

$$L(\theta) = L(\theta; x_1,\ldots,x_n) = \prod_{i=1}^{n} f(x_i;\theta)$$

When the support is independent of θ it is often convenient to use the *log-likelihood function* $\ell(\theta)$ defined for $x_1,\ldots,x_n$ in the support:

$$\ell(\theta) = \ell(\theta; x_1,\ldots,x_n) = \sum_{i=1}^{n} \ln(f(x_i;\theta)).$$

If it is unique, the value $\hat{\theta}$ at which $L(\theta)$ achieves its maximum is called the *maximum likelihood estimator* (mle) for θ.

We consider Examples 4.1.1 - 4.1.4 of HMC. Notice that in all cases, we think of the realization $x_1, \ldots, x_n$ as fixed and vary θ to find the maximum of L or ℓ.

We then obtain $\hat{\theta}$ as a function of $x_1, \ldots, x_n$. The estimator, is then the statistic $\hat{\theta}(X_1, \ldots, X_n)$.

Example 4.1.1: $f(x; \theta) = \theta^{-1} e^{-x/\theta}$ and so this is the family of $\Gamma(1, \theta)$ distributions with unknown mean θ. Alternatively this is the family of $Exp(1/\theta)$ exponential distributions.

$$\ell(\theta) = \ln[\prod_{i=1}^{n} \theta^{-1} e^{-x_i/\theta}] = \ln[\theta^{-n} e^{-\theta^{-1}\sum_i x_i}]$$

$$= -n\ln(\theta) - \theta^{-1}\sum_i x_i.$$

To find the maximum we take the derivative with respect to θ:

$$\frac{\partial \ell}{\partial \theta} = -n\theta^{-1} + \theta^{-2}\sum_{i=1}^{n} x_i.$$

The critical point is $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$ which is a local maximum by the second derivative test. It is the only critical point as so is the maximum point. The MLE is $\hat{\theta} = \bar{X}$, the sample mean, which is unbiased. Notice that $\bar{X}$ has a $\Gamma(n, \theta/n)$ distribution.

Example 4.1.2: $f(x;\theta) = \theta^x(1-\theta)^{(1-x)}$ with $x = 0$ or 1. This is the family of $Bern(\theta)$ distributions with values $0, 1$. Notice that the sum $\sum_{i=1}^n X_i$ has a $Bin(n, \theta)$ distribution.

$$\ell(\theta) = \ln(\prod_i \theta^{x_i}(1-\theta)^{(1-x_i)}) = \ln[\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}]$$

$$= (\sum_{i=1}^n x_i)\ln\theta + (n - \sum_{i=1}^n x_i)\ln(1-\theta).$$

$$\frac{\partial \ell}{\partial \theta} = \theta^{-1}(\sum_{i=1}^n x_i) - (1-\theta)^{-1}(n - \sum_{i=1}^n x_i) = n[\frac{\bar{x}}{\theta} - \frac{1-\bar{x}}{1-\theta}].$$

Again the sample mean $\bar{x}$ is the estimator for θ with $\bar{x} = \frac{k}{n}$ where $k = \sum_{i=1}^n x_i$ is the number of successes in the realization.
The MLE is $\hat{\theta} = \bar{X} = n^{-1}\sum_i X_i$.

Example 4.1.3: $f(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}((x-\mu)/\sigma)^2}$. This is the family $\mathcal{N}(\mu,\sigma^2)$ with unknown parameters the pair (μ,σ).

$$\ell(\mu,\sigma) = \ln((\frac{1}{\sigma\sqrt{2\pi}})^n exp(-\frac{1}{2}\sum_i(\frac{x_i-\mu}{\sigma})^2) =$$

$$-\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{\sigma^{-2}}{2}\sum_i(x_i-\mu)^2.$$

$$\frac{\partial\ell}{\partial\mu} = \sigma^{-2}\sum_{i=1}^{n}(x_i-\mu).$$

$$\frac{\partial\ell}{\partial\sigma} = -n\sigma^{-1} + \sigma^{-3}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Setting the first equal to zero we solve to obtain $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$, i.e. the sample mean again.

Substituting $\bar{x}$ for μ in the equation $\frac{\partial \ell}{\partial \sigma} = 0$, we obtain $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$.

The estimator for μ is unbiased, but the estimator for σ^2 is biased. When we return to the theory of MLE's later we will examine the relationship between the estimate of σ^2 and that of σ.

Example 4.1.4: This is the family $Unif(0, \theta)$ and so $f(x; \theta) = \theta^{-1}$ if $0 \leq x \leq \theta$ and $= 0$ otherwise. To be precise, we write $I_{[0,\theta]}$ for the indicator function of the interval $[0, \theta]$. That is,

$$I_{[0,\theta]}(x) = \begin{cases} 1 \text{ for } 0 \leq x \leq \theta, \\ 0 \text{ otherwise} \end{cases} .$$

Then $f(x; \theta) = \theta^{-1} I_{(0,\theta)}(x)$. Here the range varies with θ and so we can't use the log likelihood function.

$$L(\theta) = \frac{1}{\theta^n} \prod_{i=1}^{n} I_{[0,\theta]}(x_i).$$

Regardless of θ we always have $x_i \geq 0$ and so the product is zero unless $x_1, \ldots, x_n \leq \theta$. That is, $L(\theta) = 0$ unless $\max_i(x_i) \leq \theta$ in which case $L(\theta)$ is θ^{-n}. So

$$L(\theta, \mathbf{x}) = \theta^{-n} I_{[0,\theta]}(\max_i(x_i)).$$

Because θ^{-n} is a decreasing function of θ we obtain the maximum of L by choosing θ as small as possible, but $L(\theta) = 0$ unless $\theta \geq \max_i(x_i)$. So the maximum value of $L(\theta)$ occurs at $\max(x_1, \ldots, x_n)$.

Thus, $\hat{\theta} = \max(X_1, \ldots, X_n)$.

Recall that with $X \sim Unif(0, \theta)$, $F_X(x) = x/\theta$ for $0 \leq x \leq \theta$ and so with $M = \max(X_1, \ldots, X_n)$, $F_M(x) = F_X(x)^n = x^n/\theta^n$. Differentiating, $f_M(x) = nx^{n-1}/\theta^n$ and so

$$E(M) = \frac{n}{\theta^n} \int_0^\theta x^n dx = \frac{n}{n+1}\theta.$$

So the estimator $\hat{\theta} = M$ is biased.
Now we consider an example which uses calculus and which exhibits end-point issues.

Example: Let $f(x;\theta) = \theta x^{\theta-1}$ for $0 < x < 1$ and $= 0$ otherwise. The parameter is assumed to satisfy $\theta \geq 1$. Notice that with $\theta = 1$ the density is uniform on $(0, 1)$.

$$\frac{1}{n}\ell(\theta) = (\theta - 1)\frac{1}{n}(\sum_{i=1}^{n} \ln(x_i)) \; + \; \ln(\theta) = (\theta - 1)\overline{\ln(x)} + \ln(\theta).$$

$$\frac{1}{n}\ell'(\theta) = \overline{\ln(x)} \; + \; 1/\theta.$$

If $\overline{\ln(x)} \leq -1$, then $\frac{1}{n}\ell(\theta)$ is a decreasing function of θ and so the maximum $\hat{\theta} = 1$.

If $\overline{\ln(x)} > -1$, then $\hat{\theta} = -(1/\overline{\ln(x)})$.

The cut-off point occurs when $\overline{\ln(x)} = -1$ or $(x_1 \cdot \ldots x_n)^{1/n} = e^{-1}$.

Sec. 4.1.2[4.1.1]: Nonparametric Estimates

Given $X_1, \ldots X_n$ iid rv's from an unknown rv X, the best estimate of X is the *empirical distribution*. This is the discrete rv with values $X_1, ..., X_n$ with each X_i equally likely. The cdf $F_X(x)$ is estimated by the sample cdf $\hat{F}(x)$ given by

$$\hat{F}(x) = \frac{1}{n}\#\{i : X_i \leq x\}.$$

In Chapter 5, we will return to this to describe the sense in which $\hat{F}$ approximates F.

For now we follow HMC, considering first the case of a discrete rv X with known values but unknown probabilities. For each value x of X we estimate $f_X(x)$ by the average number of occurrences of x.

That is, define the function

$$I_x(t) = \begin{cases} 1 & \text{if } t = x, \\ 0 & \text{if } t \neq x. \end{cases}$$

Thus, for any rv X, $I_x(X)$ is a $Bern(p)$ rv with $p = P(X = x)$. We use the estimate

$$\hat{p}(x)(X_1, \ldots, X_n) \;=\; \frac{1}{n}\sum_{i=1}^{n} I_x(X_i) \;=\; \frac{1}{n}\#\{i : X_i = x\}.$$

Notice that $E(\hat{p}(x)) = P(X = x)$.

For a continuous rv with density $f(x)$,
$P(x-h<X<x+h)=\int_{x-h}^{x+h} f(t)dt \approx f(x)2h$. Define

$$I_{x,h}(t)=\begin{cases}1 & \text{if } x-h<t<x+h,\\ 0 & \text{otherwise.}\end{cases}$$

Thus, for any rv X, $I_{x,h}(X)$ is a $Bern(p)$ rv with
$p=P(x-h<X<x+h)$. We use the estimate

$$\hat{f}(x)(X_1,\ldots,X_n) = \frac{1}{2nh}\sum_{i=1}^{n} I_{x,h}(X_i)$$

$$= \frac{1}{2nh}\#\{i : x-h<X_i<x+h\}.$$

Notice that $E(\hat{p}(x))=P(x-h<X<x+h)/2h \approx f(x)$.

Sec. 4.2: Confidence Intervals

Returning to point estimation for pdf $f(x;\theta)$ we follow HMC Def. 4.2.1 defining a confidence interval using a pair of statistics $L(X_1,\ldots,X_n) < U(X_1,\ldots,X_n)$. When for every parameter value θ, $P_\theta(L < \theta < U) \geq 1-\alpha$, we call (L,U) a $(1-\alpha)\cdot 100\%$ confidence interval for θ.

Once the sample is drawn, we have the realized confidence interval (ℓ, u). This does not mean that the probability that θ lies in (ℓ, u) is at least $1-\alpha$. This makes no sense because θ is a fixed, but unknown parameter. What it does mean is that if θ lies outside the interval (ℓ, u) then the sample was anomalous, representing an event of probability less than α.

We will obtain confidence intervals by using a *pivot variable*, a function of the estimator of θ as well as θ itself and which has a known distribution. The confidence interval is obtained by using some algebraic manipulation.

Suppose we have a known rv Z with mean 0 and variance 1. Assume that we trying to estimate the mean μ and the variance σ^2 for an rv X with $X \sim \sigma Z + \mu$. We use a sample of iid's $X_i = \sigma Z_i + \mu$. We have

$$\bar{X} \;=\; \sigma\bar{Z} + \mu, \quad \text{and so} \quad \sqrt{n}(\bar{X} - \mu) \;=\; \sigma(\sqrt{n}\bar{Z}).$$

Notice that $E(\sqrt{n}\bar{Z}) = 0$ and $Var(\sqrt{n}\bar{Z}) = 1$.

If Z is a normal rv, then all of the Z_i's and $\sqrt{n}\bar{Z}$ are standard normals.

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \;=\; \sigma^2 \sum_{i=1}^{n}(Z_i - \bar{Z})^2$$

In the normal case, *Student's Theorem*, HMC Theorem 3.6.1 says that $\sum_{i=1}^{n}(Z_i - \bar{Z})^2$ has a $\chi^2(n-1)$ distribution and is independent of $\bar{Z}$. Furthermore,

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{(\sum_{i=1}^{n}(X_i - \bar{X})^2)/(n-1)}} = \frac{\sqrt{n}\bar{Z}}{\sqrt{(\sum_{i=1}^{n}(Z_i - \bar{Z})^2)/(n-1)}}$$

has a so-called Student's t-distribution with $n-1$ degrees of freedom.

The important thing is that the expression on the left, except for the unknown μ can be computed from the data. The expression on the right has a known distribution.

In HMC Example 4.2.1, the t-distribution is used to obtain a confidence interval for μ.

With $T = \frac{\sqrt{n}\bar{Z}}{\sqrt{(\sum_{i=1}^{n}(Z_i-\bar{Z})^2)/(n-1)}}$ we use the table to choose $t_{\alpha/2,n-1}$ so that $P(T < t_{\alpha/2,n-1}) = 1-\alpha/2$ and therefore $P(T < -t_{\alpha/2,n-1}) = \alpha/2$. So

$$1-\alpha = P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) =$$

$$P(-t_{\alpha/2,n-1} < \frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{(\sum_{i=1}^{n}(X_i-\bar{X})^2)/(n-1)}} < t_{\alpha/2,n-1})$$

$$= P(\bar{X} - t_{\alpha/2,n-1}S/\sqrt{n} < \mu < \bar{X} + t_{\alpha/2,n-1}S/\sqrt{n}).$$

That is, $L = \bar{X} - t_{\alpha/2,n-1}S/\sqrt{n}$, $U = \bar{X} + t_{\alpha/2,n-1}S/\sqrt{n}$.

To get a confidence interval for σ^2 we use the previous equation.

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(Z_i - \bar{Z})^2 = W$$

Again, the expression on the left, except for the unknown σ^2 can be computed from the data. The expression on the right has a known distribution, namely $\chi^2(n-1)$.

Given α we choose $q_{\alpha/2}, Q_{\alpha/2}$ so that

$$P(W \leq q_{\alpha/2}) = P(W \geq Q_{\alpha/2}) = \alpha/2$$

and so $P(q_{\alpha/2} < W < Q_{\alpha/2})) = 1 - \alpha$.

That is,

$$P(q_{\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < Q_{\alpha/2}) =$$

$$P(q_{\alpha/2} < \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 < Q_{\alpha/2}) = 1 - \alpha.$$

Our $(1-\alpha)100\%$ confidence interval for σ^2 is

$$(\frac{(n-1)s^2}{Q_{\alpha/2}}, \frac{(n-1)s^2}{q_{\alpha/2}}).$$

For most of our examples we will use a *large sample confidence interval* which uses that *Central Limit Theorem*. We will be looking at the CLT in detail in Chapter 5. If $X_1, X_2, \ldots$ is an infinite iid sequence from a distribution with mean μ and variance σ^2, then the rv's $W_n = \sqrt{n}\bar{Z} = (\bar{X} - \mu)/(\sigma/\sqrt{n})$ all have mean 0 and variance 1. The Central Limit Theorem says that as $n \to \infty$ the cdf of W_n tends to Φ the cdf of the $\mathcal{N}(0,1)$ standard normal distribution. In addition the sample variance S^2 approaches σ^2 and so the cdf of $Y_n = (\bar{X} - \mu)/(S/\sqrt{n})$ approaches Φ as well.

We obtain as an approximation, a large sample confidence interval by pretending that $W_n \sim \mathcal{N}(0,1)$ or $Y_n \sim \mathcal{N}(0,1)$ which is approximately true when n is large.

This is illustrated by Example 4.2.2.

$Y_n = (\bar{X} - \mu)/(S/\sqrt{n})$ has cdf approximately that of the standard normal Z with cdf Φ.

Given $\alpha < 1$ we define $z_{\alpha/2}$ by $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ so that $\Phi(-z_{\alpha/2}) = \alpha/2$. That is, $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$. Hence, it is approximately true that

$$P(-z_{\alpha/2} < (\bar{X} - \mu)/(S/\sqrt{n}) < z_{\alpha/2}) = 1 - \alpha.$$

So our large sample $(1 - \alpha)100\%$ confidence interval is $(\bar{x} - z_{\alpha/2}s/\sqrt{n}, \bar{x} + z_{\alpha/2}s/\sqrt{n})$, using the realized values.

In Example 4.2.3, the X_i's are $Bern(p)$ rvs with $\mu = p$ and $\sigma^2 = p(1-p)$. So $\bar{X} = \hat{p}$ the ratio of the number of successes to the number of trials.

The Central Limit Theorem applies to

$$W_n = (\bar{X} - \mu)/(\sigma/\sqrt{n}) = (\hat{p} - p)/\sqrt{p(1-p)/n}.$$

Again we replace $\sqrt{p(1-p)}$ by the sample value $\sqrt{\hat{p}(1-\hat{p})}$. So our large sample $(1-\alpha)100\%$ confidence interval is $(\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$.

Exercise 4.2.3[4.2.4]: (a) If $X \sim \Gamma(1, \theta)$ then $\frac{2X}{\theta} \sim \Gamma(1, 2)$. By the additive property of the Gamma distribution $\frac{2}{\theta} \sum_{i=1}^{n} X_i \sim \Gamma(n, 2) = \Gamma(2n/2, 2)$ which is a χ^2 distribution with $r = 2n$.

(b) With $F_{2n}(x)$ the cdf for the χ^2 distribution with $2n$ degrees of freedom, we let $q_{\alpha/2}, Q_{\alpha/2}$ be defined by $F_{2n}(q_{\alpha/2}) = \alpha/2$ and $1 - F_{2n}(Q_{\alpha/2}) = \alpha/2$. So

$$P(q_{\alpha/2} < \frac{2n\bar{X}}{\theta} < Q_{\alpha/2}) = 1 - \alpha.$$

The $1 - \alpha$ confidence interval is then given by $(\frac{2n\bar{x}}{Q_{\alpha/2}}, \frac{2n\bar{x}}{q_{\alpha/2}})$.

(c) The large sample confidence interval is given by $(\bar{x} - z_{\alpha/2}s/\sqrt{n}, \bar{x} + z_{\alpha/2}s/\sqrt{n})$.

Complete the numerical comparison for homework by using the tables.

Example 4.4.7: Confidence Interval for the Median

We use order statistics for the median.

For a continuous rv with unknown cdf F, we want a confidence interval for the median $\xi_{.5}$ defined by $F(\xi_{.5}) = \frac{1}{2}$. Thus, the event $X < \xi_{.5}$ has probability $\frac{1}{2}$.

For a $Bin(n, \frac{1}{2})$ rv S, let $k = k_{\alpha/2}$ be the largest positive integer such that $F_S(k) = P(S \leq k) \leq \alpha/2$. By symmetry, $P(S \geq n - k) \leq \alpha/2$.

For a sample $X_1, \ldots, X_n$ the order statistics are $Y_1 < Y_2 \cdots < Y_n$.

The indicators $I_{X_i < \xi_{.5}}$ are independent $Bern(\frac{1}{2})$ rv's.

$Y_{k+1} \geq \xi_{.5}$ when at most k among the X_i's are less than $\xi_{.5}$.
$Y_{n-k} \leq \xi_{.5}$ when at most k among the X_i's are greater than or equal to $\xi_{.5}$.

$$P(Y_{k+1} \geq \xi_{.5}) \;=\; P(Y_{n-k} \leq \xi_{.5}) = F_S(k).$$

So $P(Y_{k+1} < \xi_{.5} < Y_{n-k}) \geq 1 - \alpha$.

The confidence interval is (y_{k+1}, y_{n-k}).

Sec. 4.5: Alternative Hypotheses

We consider an rv with pdf or pmf given by $f(x;\theta)$ where the parameter θ varies in a parameter space Ω. The set Ω is partitioned into two sets ω_0 and ω_1. That is, these sets are disjoint and with union all of Ω.

We wish to decide whether the true parameter θ^* lies in ω_0 or ω_1.

$\mathbf{H}_0$ - The *Null Hypothesis* is $\theta^* \in \omega_0$. Often ω_0 consists of a single value θ_0 in which case the null hypothesis is $\theta^* = \theta_0$.

$\mathbf{H}_1$ - The *Alternative Hypothesis* is $\theta^* \in \omega_1$.

We design a statistical procedure to decide between the two hypotheses by using a sample $X_1, \ldots, X_n$.

Letting $\mathcal{S}^n$ denote the space of possible sample values, we use for a *test* a subset $C \subset \mathcal{S}^n$ which we call the *critical region* and we use the decision rule:

Reject H_0 (Accept H_1) when $(X_1, \ldots, X_n) \in C$.

Retain H_0 (Reject H_1) when $(X_1, \ldots, X_n) \in C^c$.

There are two kinds of errors which can occur. I find it helpful to use medical terminology so that $\theta^* \in \omega_0$ is a *negative result* (In medicine a negative result for a test is good news).

Type I error - *False Positive* : In fact $\theta^* \in \omega_0$ but we reject the null hypothesis.

Type II error - *False Negative* : We accept the null hypothesis despite the fact that θ^* really lies in ω_1.

There is a trade-off between the two sorts of errors. For a particular critical region C we define the *size* or the *significance* of C to be

$$\alpha = \max_{\theta \in \omega_0} P_\theta((X_1, \ldots, X_n) \in C).$$

So in the case when ω_0 contains a single value θ_0, the size is $P_{\theta_0}((X_1, \ldots, X_n) \in C)$. So this is then the probability of a Type I error when we use C for our decision.

For $\theta \in \omega_1$ we want to maximize

$$1 - P_\theta(\text{Type II error}) = P_\theta((X_1, \ldots, X_n) \in C).$$

The set ω_1 usually contains more than one alternative. We define the *power function* of the critical region C by

$$\gamma_C(\theta) = P_\theta((X_1, \ldots, X_n) \in C) \qquad \theta \in \omega_1.$$

We look at HMC Examples 4.5.2, 4.5.3, and 4.5.4.

Example 4.5.2: We are considering the Bernoulli p family $Bern(p)$. The null hypothesis is $H_0 : p = p_0$ and the alternative is $H_1 : p < p_0$.

Given a sample $X_1, \ldots, X_n$, the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$ is the MLE estimate for the mean p. Intuitively, if $\bar{X}$ is small relative to p_0 then we would reject H_0 and accept otherwise.

Let $S = \sum_{i=1}^{n} X_i$. This is a $Bin(n, p)$ rv. We choose a fixed k so that we reject if $S \leq k$. That is, the critical region is

$$C_k = \{S \leq k\}.$$

Compare with a confidence interval for the mean μ of an $\mathcal{N}(\mu, \sigma^2)$ distribution with the variance known.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = Z$$

is a standard normal and so our confidence interval is of the form

$$(\bar{X} - z\sigma/\sqrt{n}, \bar{X} + z\sigma/\sqrt{n}).$$

We choose z so that, given α

$$P(-z < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z) = P(-z < Z < z) = 1 - \alpha.$$

Similarly, we choose k so that the size of the critical region C_k has size α, that is, $P_{p_0}(C_k) = \alpha$.

Since S is a $Bin(n, p_0)$ rv, we want to choose k so that $F_{n,p_0}(k) = \alpha$, where F_{n,p_0} is the cdf of a $Bin(n, p_0)$ distribution.

The power function is given by

$$\gamma_p(C_k) \;=\; P_p(C_k) \;=\; P_p(S \leq k) \;=\; F_{n,p}(k).$$

Because the distribution is discrete we choose k so that $F_{n,p_0}(k)$ is as large as possible and at most α.

For each fixed k, the power function is a decreasing function of p.

Proof: Let $p < p_1$ so that $p_1 = p + \epsilon$ with $0 < \epsilon \leq 1 - p$. That is, $0 < \epsilon/(1-p) \leq 1$.

If $X \sim Bern(p), W \sim Bern(\epsilon/(1-p))$ with X and W independent, then $Y = X + (1-X)W \sim Bern(p_1)$. To see this, observe that the range of Y is $\{0, 1\}$ and so it is Bernoulli. Since independence implies

$$E(Y) = E(X) + E(1-X) \cdot E(Y) = p + (1-p) \cdot [\epsilon/(1-p)] = p_1,$$

it follows that $Y \sim Bern(p_1)$. Notice that $Y \geq X$.

If $X_1, \ldots, X_n$ are iid $Bern(p)$ rv's and $W_1, \ldots, W_n$ are iid $Bern(\epsilon/(1-p))$ rv's with all the X_i's independent of the W_i's, then with $Y_i = X_i + (1-X_i)W_i$ we get $Y_1, \ldots, Y_n$ iid $Bern(p_1)$ rv's with $X_i \leq Y_i$ for all i. So for any k,

$$P(\sum_{i=1}^{n} Y_i \leq k) \;\leq\; P(\sum_{i=1}^{n} X_i \leq k).$$

Example 4.6.3: If n is large then $S = n\bar{X}$ is $Bin(n, p_0)$ and

$$\frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)/n}} = Z_0$$

is approximately a standard normal. If, with $\alpha < \frac{1}{2}$ and we choose z_α so that $\Phi(z_\alpha) = 1 - \alpha$ then $\Phi(-z_\alpha) = \alpha$. We can use the critical region

$$\begin{aligned} C_\alpha &= \{\frac{\bar{X} - p_0}{\sqrt{p_0(1-p_0)/n}} \leq -z_\alpha\} \\ &= \{\bar{X} \leq p_0 - z_\alpha \cdot \sqrt{p_0(1-p_0)/n}\} \\ &= \{S \leq np_0 - z_\alpha \cdot \sqrt{np_0(1-p_0)}\}. \end{aligned}$$

Example 4.5.3: This extends what we just did. We suppose that X has a finite mean μ and a finite variance σ^2. We want to test the simple hypothesis $H_0 : \mu = \mu_0$ against the composite hypothesis $H_1 : \mu > \mu_0$. As usual, we use a sample $X_1, \ldots, X_n$.

The CLT says that $\frac{\bar{X}-\mu}{S/\sqrt{n}} = Z$ is approximately a standard normal. Notice that here S^2 is the sample variance. Given α we choose z_α so that $\Phi(z_\alpha) = 1 - \alpha$. That is, the tail above z_α has area α. We use as our critical region

$$C_\alpha = \{\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \geq z_\alpha\} = \{\bar{X} \geq \mu_0 + z_\alpha S/\sqrt{n}\}.$$

If we actually know the variance σ^2 then we would use the true variance σ instead of the sample variance S. We can then compute the approximate power function, formula (4.5.12) in HMC.

$$
\begin{aligned}
\gamma(\mu) &= P_\mu(\bar{X} \geq \mu_0 + z_\alpha \sigma/\sqrt{n}) \\
&= P_\mu\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \geq \frac{\mu_0-\mu}{\sigma/\sqrt{n}} + z_\alpha\right) \\
&\approx 1 - \Phi\left(z_\alpha - \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}\right) \\
&= \Phi\left(-z_\alpha + \frac{\sqrt{n}(\mu-\mu_0)}{\sigma}\right).
\end{aligned}
$$

The function γ is increasing in μ.

Example 4.5.4: When X is known to be normal, then for any n, $\frac{\bar{X}-\mu}{S/\sqrt{n}} = T$ has a $t-$ distribution with $n-1$ degrees of freedom. We use the above critical region but with z_α replaced by $t_{\alpha,n-1}$ again chosen so that the tail above $t_{\alpha,n-1}$ has area α, but this time with the Student's t pdf instead of the standard normal pdf.

Sec. 4.6: Two-Sided Tests and p-Values

For a test with $\omega_0 = \theta_0$, against a set of alternatives ω_1, suppose we use as our test $C = \{Y \leq c\}$ for $Y = u(X_1, \ldots, X_n)$. That is, we reject if the realized value $y < c$.

For a test of size α, we choose c so that $\alpha = P_{\theta_0}(Y \leq c) = F_Y(c; \theta_0)$.

The p-value is $P_{\theta_0}(Y \leq y) = F_Y(y; \theta_0)$ where y is the observed value, i.e. the realization of Y after the sample is taken. See HMC Remark 4.6.1.

In the continuous rv case, F_Y is an increasing function. So $F_Y(y; \theta_0) \leq \alpha = F_Y(c; \theta_0)$ is equivalent to $y \leq c$. Thus, we need not compute c. We reject when the p-value is less than α.

For two sided tests we use the normal approximation as in HMC Examples 4.6.1 and 4.6.3.
Example 4.6.1: This is a variation of Example 4.5.3. We suppose that X has a finite mean μ and a finite variance σ^2. We want to test the simple hypothesis $H_0 : \mu = \mu_0$ against the composite hypothesis $H_1 : \mu \neq \mu_0$. As usual, we use a sample $X_1, \ldots, X_n$.

For the one-sided test with $H_1 : \mu > \mu_0$ we rejected when the sample mean $\bar{X}$ is above a certain level. By analogy, we choose $h < \mu_0 < k$ and reject when either $\bar{X} < h$ or $\bar{X} > k$. That is,

$$C = \{\bar{X} \leq h\} \cup \{\bar{X} \geq k\}$$
$$\alpha = P_{\mu_0}(C) = P_{\mu_0}(\bar{X} \leq h) + P_{\mu_0}(\bar{X} \geq k).$$

As before the CLT says that $\frac{\bar{X}-\mu}{S/\sqrt{n}} = Z$ is approximately a standard normal. With $z_{\alpha/2}$ chosen so that $\Phi(-z_{\alpha/2}) = 1 - \Phi(z_{\alpha/2}) = \alpha/2$ we use

$$C_\alpha = \{|\frac{\bar{X} - \mu_0}{S/\sqrt{n}}| \geq z_{\alpha/2}\} =$$
$$\{\bar{X} \leq \mu_0 - z_{\alpha/2}S/\sqrt{n}\} \cup \{\bar{X} \geq \mu_0 + z_{\alpha/2}S/\sqrt{n}\}$$

Notice that

$$C_\alpha^c = \{\bar{X} : \mu_0 - z_{\alpha/2}S/\sqrt{n} < \bar{X} < \mu_0 + z_{\alpha/2}S/\sqrt{n}\}$$
$$= \{\bar{X} : \bar{X} - z_{\alpha/2}S/\sqrt{n} < \mu_0 < \bar{X} + z_{\alpha/2}S/\sqrt{n}\}$$

That is, we accept the null-hypothesis H_0 when μ_0 lies in the $(1-\alpha)100\%$ confidence interval.

If we know what the variance σ^2 equals, or approximately equals, then we can substitute σ for S to compute the approximate power function

$$\begin{aligned}\gamma(\mu) &= P_\mu(\bar{X} \leq \mu_0 - z_{\alpha/2}\sigma/\sqrt{n}\}) + P_\mu(\bar{X} \geq \mu_0 + z_{\alpha/2}\sigma/\sqrt{n}\}) \\ &\approx \Phi(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - z_{\alpha/2}) + (1 - \Phi(\frac{\mu_0 - \mu}{\sigma/\sqrt{n}} + z_{\alpha/2})),\end{aligned}$$

because this time it is $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ which is approximately an $\mathcal{N}(0,1)$ rv.

If X is normal then assuming H_0, then $\frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ has a $t-$ distribution with $n-1$ degrees of freedom and we can use $t_{\alpha/2,n-1}$ instead of $z_{\alpha/2}$ to get a more accurate test.

Example 4.6.3 (again): We consider X to be $Bern(p)$ with parameter p. We write $\hat{p}$ for the sample mean $\bar{X}$ because it is the unbiased estimate for the mean p.

We observed that, assuming H_0

$$\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} = Z_0$$

is approximately normal for n large. So we used $Z_0 \leq -z_\alpha$ as a test for $H_0 : p = p_0$ against $H_1 : p < p_0$.

For any p,

$$\frac{\hat{p}-p}{\sqrt{\hat{p}(1-\hat{p})/n}} = Z_p$$

is approximately normal for n large and we can use $Z_{p_0} \leq -z_\alpha$ as a test for $H_0 : p = p_0$ against $H_1 : p < p_0$.

For a two sided test, we can use $C = \{|Z_{p_0}| \geq z_{\alpha/2}\}$ as a critical region to test for $H_0 : p = p_0$ against $H_1 : p \neq p_0$.

Thus, we accept H_0 using this test, when the results lie in the complementary region

$$\begin{aligned} C^c &= \{\hat{p} : p_0 - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < \hat{p} < p_0 + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\} \\ &= \{\hat{p} : \hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p_0 < \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\} \end{aligned}$$

Again, this is equivalent to p_0 lying in the $(1-\alpha)100\%$ confidence interval.

Just a quick aside about the variance estimate.
Recall that if $X \sim Bern(p)$ then $E(X) = p$ and $Var(X) = p(1-p)$. $\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the MLE estimate for p and it is unbiased. It follows that the MLE estimate for the variance is $\hat{p}(1-\hat{p})$. Notice that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2 = \bar{X} - \bar{X}^2 = \hat{p}(1-\hat{p}),$$

because $X_i^2 = X_i$ for all i.

Thus, $\hat{p}(1-\hat{p}) = \frac{n-1}{n}S^2$ where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance.

For a *randomized test* we use an additional rv Y independent of the rv of the test and with pdf or pmf $f_Y(y)$ for y in the support S_Y.

Notice that f_Y does not depend on θ and so the joint distribution is

$$f_{\mathbf{X},Y}(\mathbf{x}) = f(x_1;\theta)\cdot\ldots f(x_n;\theta)f_Y(y).$$

Usually $Y \sim Bern(p)$, i.e. it is a Bernoulli rv.

The critical region $C \subset \mathcal{S}^n \times \mathcal{S}_Y$ and we reject H_0 when $(\mathbf{x}, y) \in C$.

As before, the size is $\max_{\theta\in\omega_0} P_\theta(C)$ and for $\theta \in \omega_1$ the power is

$$\gamma_C(\theta) = P_\theta(C).$$

Sec. 8.1: Most Powerful Simple Tests

In this section we consider a simple test for a family $f(x;\theta)$ of pdf's or pmf's. That is, the simple null hypothesis $H_0 : \theta = \theta_0$ against the simple alternative $H_1 : \theta = \theta_1$.

The samples $\mathbf{X} = X_1, \ldots, X_n$ lie in $\mathcal{S}^n$ where $\mathcal{S}$ is the support of the family of rv's. We are assuming that the support does not depend on θ. A critical region is a subset of $\mathcal{S}^n$ so that we reject H_0 when $\mathbf{X} \in C$ and accept it when $\mathbf{X} \in C^c$.

For a critical region C, the size or significance $\alpha = P_{\theta_0}(C)$, the probability of a Type I error, and the power of the test is $\gamma_C(\theta_1) = P_{\theta_1}(C)$, the probability of correctly rejecting the null hypothesis. That is, that the power is 1 minus the probability of a Type II error.

HMC Definition 8.1.1: A critical region C of size α is a *best critical region of size* α if, whenever A is a critical region of size α,

$$P_{\theta_1}(C) \geq P_{\theta_1}(A).$$

That is, the power of the region C is the maximum power possible for a critical region of size α.

We use the likelihood function $L(\theta)$

$$L(\theta; \mathbf{x}) = L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

Since θ_0 and θ_1 are known values, we can compute

$$\Lambda(\theta_0, \theta_1; \mathbf{x}) = \frac{L(\theta_0; \mathbf{x})}{L(\theta_1; \mathbf{x})}$$

and use it to define a critical region.

Neyman-Pearson Theorem

HMC Theorem 8.1.1: (**Neyman-Pearson Theorem**) If a subset C of the sample space satisfies:

- $P_{\theta_0}(C) = \alpha$;
- $\Lambda(\theta_0, \theta_1; \mathbf{x}) \leq k$ for all $\mathbf{x} \in C$;
- $\Lambda(\theta_0, \theta_1; \mathbf{x}) \geq k$ for all $\mathbf{x} \in C^c$;

for some positive constant k, then C is a best critical region of size α.

Proof: We will temporarily write $L(\theta; A)$ for $\int_A L(\theta; \mathbf{x}) d\mathbf{x}$. This is just alternate notation for $P_\theta(A)$. So, for example,

$$L(\theta; C) = L(\theta; A^c \cap C) + L(\theta; A \cap C).$$
$$L(\theta; A) = L(\theta; A \cap C^c) + L(\theta; A \cap C),$$

Subtracting and cancelling the common terms we see that for any θ.

$$L(\theta; C) - L(\theta; A) \;=\; L(\theta; A^c \cap C) - L(\theta; A \cap C^c)$$

If A is any other critical region of size α we want to show

$$L(\theta_1; C) - L(\theta_1; A) \geq 0.$$

Because of the assumptions about C and C^c, we have

$$L(\theta_1; A^c \cap C) \geq k^{-1} L(\theta_0; A^c \cap C),$$
$$-L(\theta_1; A \cap C^c) \geq -k^{-1} L(\theta_0; A \cap C^c).$$

$$L(\theta_1; C) - L(\theta_1; A) = L(\theta_1; A^c \cap C) - L(\theta_1; A \cap C^c) \geq$$
$$k^{-1}[L(\theta_0; A^c \cap C) - L(\theta_0; A \cap C^c)] = k^{-1}[L(\theta_0; C) - L(\theta_0; A)].$$

Because C and A are critical regions of size α,
$L(\theta_0; C) = L(\theta_0; A) = \alpha$. So $L(\theta_1; C) - L(\theta_1; A) \geq 0$.

The Neyman-Pearson Theorem works with the same proof with C and A randomized tests. From that we get

HMC Corollary 8.1.1: If C is a best critical region of size α for $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, then $P_{\theta_1}(C) \geq \alpha$. That is, the power is greater than or equal to the size.

Proof: We compare C with the trivial randomized test which uses $Y \sim Bern(\alpha)$, and we use $A = \{Y = 1\}$. So the power equals the size for this test because $P_\theta(A) = \alpha$ for all θ.

By the Neyman-Pearson Theorem $P_{\theta_1}(C) \geq P_{\theta_1}(A) = \alpha$.

Here θ_0 and θ_1 are known parameter values. So $\Lambda(\theta_0, \theta_1; \mathbf{x})$ is a statistic and so for any k we can define the critical region $C_k = \{\Lambda(\theta_0, \theta_1; \mathbf{x}) \leq k\}$. The Neyman-Pearson Theorem then says that C_k is a best critical region of its size

If we start with size α, then we choose k so that $P_{\theta_0}(C_k) = \alpha$.

We can often express Λ in terms of a single statistic $T(\mathbf{X})$ separate from θ_0 and θ_1. In the next section we will see how this is done in general.

Example 8.1.2 and 8.2.3: Let $X \sim \mathcal{N}(\theta, 1)$ so that $f(x;\theta) = \frac{1}{\sqrt{2\pi}} exp(-\frac{(x-\theta)^2}{2})$. We test for θ_0 against $\theta_1 > \theta_0$. Notice first that

$$(-\frac{\sum_{i=1}^n (x_i - \theta_0)^2}{2}) - (-\frac{\sum_{i=1}^n (x_i - \theta_1)^2}{2}) =$$

$$-(\theta_1 - \theta_0)\sum_{i=1}^n x_i + \frac{n}{2}(\theta_1^2 - \theta_0^2).$$

So

$$\Lambda(\theta_0, \theta_1; \mathbf{X}) = exp(-(\theta_1 - \theta_0)\sum_{i=1}^n X_i + \frac{n}{2}(\theta_1^2 - \theta_0^2))$$

$$= exp(n(\theta_1 - \theta_0)[-\bar{X} + \frac{\theta_1 + \theta_0}{2}]),$$

and we use

$$C_k = \{\Lambda(\theta_0, \theta_1; \mathbf{X}) \leq k\} = \{\bar{X} \geq \frac{\theta_1 + \theta_0}{2} - \frac{\ln k}{n(\theta_1 - \theta_0)}\}.$$

For any θ, $\sqrt{n}(\bar{X} - \theta) \sim \mathcal{N}(0,1)$. So if $1 - \Phi(z_\alpha) = \alpha$, C_k has size α with $\theta = \theta_0$ when $C_k = \{\bar{X} \geq \theta_0 + \frac{z_\alpha}{\sqrt{n}}\}$. We can solve this for k.

$$k = exp[(\theta_1 - \theta_0)[\frac{(\theta_1 - \theta_0)}{2} - \frac{z_\alpha}{\sqrt{n}}].$$

However, we don't need to determine k.
If $\theta = \theta_1$, then $X - \theta_1 \sim \mathcal{N}(0,1)$ and so
$\sqrt{n}(\bar{X} - \theta_1) \sim \mathcal{N}(0,1)$. The power is given by

$$\begin{aligned}\gamma_{C_k}(\theta_1) = P_{\theta_1}(C_k) &= P(\sqrt{n}(\bar{X} - \theta_1) \geq z_\alpha - \sqrt{n}(\theta_1 - \theta_0)) \\ &= 1 - \Phi(z_\alpha - \sqrt{n}(\theta_1 - \theta_0)) \geq 1 - \Phi(z_\alpha) = \alpha.\end{aligned}$$

As HMC remark, although we have been assuming that the pdf's or pmf's come from a parameterized family, this need not so For the Neyman-Pearson result. All that is needed is that the two distributions of the two simple hypotheses have the same range.

Example 8.1.3: Here the authors test the Poiss(1) pmf $H_0 : f_0(x) = e^{-1}/x!$ against the Geom($\frac{1}{2}$) $H_1 : (\frac{1}{2})^{x+1}$ for $x = 0, 1, \ldots$.

$$\Lambda(\mathbf{X}) = (e^{-n}/x_1! \ldots x_n!) \div ((1/2)^n (1/2)^{x_1 + \ldots x_n}) = \frac{(2e^{-1})^n 2^{\sum x_i}}{\prod x_i!}.$$

For any k, $C_k = \{\Lambda \leq k\}$ defines a best critical region for $\alpha = P_0(C_k)$. However, as the book illustrates, computing what the set C is can be complicated enough that this is really only of theoretical interest.

> *In theory, theory and practice are the same thing, but in practice, they are really not.*

Remark 8.1.2: Recall that the size $\alpha = P_{\theta_0}(C)$ is the probability of a Type I error and $\beta = P_{\theta_1}(C^c) = 1 - P_{\theta_1}(C)$ is the probability of a Type II error. With $d_0, d_1 > 0$, suppose that we want to minimize $d_0\alpha + d_1\beta$. In the notation of the proof of the Neyman-Pearson Theorem, this is

$$d_0 \int_C L(\theta_0) + d_1 \int_{C^c} L(\theta_1) = d_1 + \int_C [d_0 L(\theta_0) - d_1 L(\theta_1)].$$

Clearly we minimize this by choosing

$$C = \{d_0 L(\theta_0) - d_1 L(\theta_1) < 0\} = \{\Lambda < \frac{d_1}{d_0}\}.$$

Given α this is the same as minimizing β, ie. maximizing the power $\gamma_C(\theta_1)$. That is, choosing a best critical region.