ESMA 6600 - Theory of Probability

Dr. Wolfgang Rolke

May 21, 2019

Contents

1 Probability

1.1 Fundamentals

The probability of rain tomorrow is 30% (or 0.3)

What does that mean?

It will likely come at a surprise that the answer to that question is not a simple one. In fact, experts (mathematicians, statisticians, philosophers etc) have been thinking about this question for centuries, and yet there is no single universally accepted answer to it, even today.

In this course we will eventually do the "math thing" and start with a set of axioms, from which one can derive the whole theory of probability.

But rather than simply state those axioms I want to show you that they are at least "reasonable".

The following is in large part taken from the book

Principles of Uncertainty by Joseph B. Kadane

1.1.1 The Idea of the Sure Loser

Let's go back to the question of rain tomorrow, and let's also add the question: Is the temperature tomorrow going to be over 90 degrees Fahrenheit? If we take the two together we get the four possibilities:

A_1 : Rain and High above 90 degrees F tomorrow
A_2 : Rain and High at or below 90 degrees F tomorrow
A_3 : No Rain and High above 90 degrees F tomorrow
A_4 : No Rain and High at or below 90 degrees F tomorrow

Tomorrow, one and only one of these events will occur. In mathematical language, the events are *exhaustive* (at least one must occur) and *disjoint* (no more than one can occur). (Sometimes we also say they form a *partition*).

Let's assume there are four tickets, one for each of the four possibilities. You can offer to either buy one of these tickets from me or sell me one. Let's say you decide to sell me ticket A_1 , for which I will pay you \$p. Then if it rains tomorrow and the high is above ninety you will pay me \$1, otherwise you owe me nothing.

Now an important idea is that in a deal that is fair for both sides it doesn't matter who is the seller and who is the buyer. The price p reflects what you think the ticket is worth either way. It is (in your opinion, which might be different from mine) a fair price for this ticket. In fact I will decide whether to buy or sell a ticket after you named the price.

This is like the old idea of how to devide a piece of cake between two people: one divides it into two pieces, the other gets to choose. Clearly it is in the first persons best interest to divide the piece as evenly as possible.

The intuition behind this is that if you are willing to buy or sell a ticket on A_1 for \$0.70, you consider A_1 more likely than if you were willing to buy or sell it for only \$0.10. The price p is in your opinion the likelihood of A_1 happening.

Let us suppose that in general your price for a \$1 ticket on A_1 is $\Pr(A_1)$ (pronounced "price of A_1"), and in particular you name 30 cents. This means that I can sell you such a ticket for \$0.30 (or buy such a ticket from you for \$0.30). If I sell the ticket to you and it rains tomorrow and the temperature is above 90 degrees Fahrenheit, I would have to pay you \$1. If it does not rain or if the temperature does not rise to be above 90 degrees Fahrenheit, I

would not pay you anything. Thus in the first case, you come out \$0.70 ahead, while in the second case I am ahead by \$0.30.

Similarly you name prices $\Pr(A_2)$, $\Pr(A_3)$ and $\Pr(A_4)$.

Now of course you would like to win money in this game, but there is no way to make sure of that. On the other hand it would clearly be stupid for you to name prices that would assure the I win (and you loose). What can we say about prices that would make you a **sure looser**?

To take the simplest requirement first, suppose you make the mistake of offering a negative price for an event, for example

$\Pr(A_1)$ = - \$0.05

This would mean that you offer to sell me ticket A_1 for the price of -\$0.05, (i.e., you will give me the ticket and 5 cents). If event A_1 happens, that is, if it rains and the high temperature is more than 90 degrees Fahrenheit, you owe me \$1, so your total loss is \$1.05. On the other hand, if event A_1 does not happen, you still lose \$0.05. Hence in this case, no matter what happens, you are a sure loser. To avoid this kind of error, your prices cannot be negative, that is, for every event A, you must specify prices satisfying

$\Pr(A) \geq 0$ (Rule 1)

Now consider the sure event S. In the example we are discussing, S is the same as the event of either A_1 or A_2 or A_3 or A_4 , which is a formal mathematical way of saying either it will rain tomorrow or it will not, and either the high temperature will be above 90 degrees Fahrenheit or not.

What price should you give to the sure event S? If you give a price below \$1, say \$0.75, I can buy that ticket from you for \$0.75. Since the sure event is sure to happen, tomorrow you will owe me \$1, and you will have lost \$0.25, whatever the weather will be. So you are sure to lose if you name any price below \$1. Similarly, if you offer a price above \$1 for the sure event S, say \$1.25, I can sell you the ticket for \$1.25. Tomorrow, I will certainly owe you \$1, but I come out ahead by \$0.25 whatever happens. So you can see that the only way to avoid being a sure loser is to have a price of exactly \$1 for S. This isthe second requirement to avoid a sure loss, namely,

$\Pr(S) = 1$ (Rule 2)

Next, let's consider the relationship of the price you would give to each of two disjoint sets A and B to the price you would give to the event that at least one of them happens, which is called the union of the events A and B, and is written $A \cup B$.

To be specific, let A be the event A_1 above, and B be the event A_2 above. These events are disjoint, that is, they cannot both occur, because it is impossible that the high temperature for the day is both above and below 90 degrees Fahrenheit. The union of A and B in this case is the event that it rains tomorrow.

Suppose, that your prices are \$0.20 for A_1 , \$0.25 for A_2 and \$0.40 for the union of A_1 and A_2 . Then I can sell you a ticket on A_1 for \$0.20, and a ticket on A_2 for \$0.25, and buy from you a ticket on the union for \$0.40.

Let's see what happens. Suppose first that it does not rain. Then none of the tickets have to be settled by payment. But you gave me \$0.20 + \$0.25 = \$0.45 for the two tickets you bought, and I gave you \$0.40 forthe ticket I bought, so I come out \$0.05 ahead.

Now suppose that it does rain. Then one of A_1 and A_2 occurs (but only one. Remember that they are disjoint). So I have to pay you \$1. But the union also occurred, so you have to pay me \$1 as well. In addition I still have the \$0.05 that I gained from the sale and purchase of the tickets to begin with. So in every case, I come out ahead by \$0.05, and you are a sure loser.

The problem seems to be that you named too low a price for the ticket on the union. Indeed, any price less than \$0.45 leads to sure loss, with the same argument as above.

How about charging more than \$0.45, say \$0.60? Now if I do the exact opposite, namely sell you the union and buy from you A_1 and A_2 it easy to see that you are a sure looser again. The only way for you not be a sure looser is if you choose the prices such that

$\Pr(A \cup B) = \Pr(\text{A}) + \Pr(\text{B})$ (Rule 3)

Now we have seen three requirements for any prices so you are not a sure looser. In a bit we will show that in fact these are all the requirements!

Prices satisfying these equations are said to be **coherent**. The derivations of equations are constructive, in the sense that I reveal exactly which of your offers I accept to make you a sure loser. Also note that because it is always you who is naming the prices my beliefs (prices) are irrelevant to making you a sure loser.

The equations are of course the equations that define Pr() to be a probability (with the possible strengthening of Rule 3 to be taken up later). To emphasize that, we will now assume that you have decided not to be a sure loser, and hence to have your prices satisfy equations 1-3. I will write P() instead of Pr(), and think of P(A) as your probability of event A.

Although the approach here is called subjective, there are both subjective and objective aspects of it. It is an objective fact, that is a mathematical theorem, that you cannot be made a sure loser if and only if your prices satisfy equations 1-3. However, the prices that you assign to tickets on any given set of events are personal, or subjective, in that the theorems do not specify those values. Different people can have different probabilities without violating coherence.

To see why this is natural, consider the following example: Imagine I have a coin that we both regard as fair, that is, it has probability 1/2 of coming up heads. I flip it, but I don't look at it, nor do I show it to you. Reasonably, our probabilities are still 1/2 for heads.

Now I look at it, and observe heads, but I don't show it to you. My probability is now 1. Perhaps yours is still 1/2. But perhaps you saw that I raised my left eyebrow when I looked at the coin, and you think I would be more likely to do so if the coin came up heads than tails, and so your probability is now 60%. Finally I show you the coin, and your probability now rises to 1.

The point of this thought-experiment is that probability is a function not only of the coin, but also of the information available to the person whose probability it is. Thus subjectivity occurs, even in the single flip of a fair coin, because each person can have different information and beliefs.

There are a number of subtleties here which we will not discuss in detail. As one example, let's say you named a price p for A this morning. Now in the the afternoon you change your mind, not because there is any new information but just because you feel like it. Now if your new price is higher, and I bought the ticket for A from you this morning, I can now sell it back to you for the new higher price, and you are a sure looser again! One might call this a dynamic sure loss. It is important to remember that coherence is a minimal set of requirements on probabilistic opinions. The most extraordinary nonsense can be expressed coherently, such as that the moon is made of green cheese. Moreover, there is a substantial body of psychological research dedicated to finding systematic ways in which the prices that people actually offer for tickets or the equivalent fail to be coherent.

There is a special issue about whether personal probabilities can be zero or one. The implication is that you would bet your entire fortune present and future against a penny on the outcome, which is surely extreme. In the example above, I propose that when I see that the coin came up heads, my probability is one that it is heads. But could I have seen wrong? For the sake of the argument I am willing to set that possibility aside, but I must concede that sometimes I do get things wrong, so I can't really mean probability one.

Is there an event you would be willing to bet your live on?

The approach to probability described here is sometimes referred to as **behavioristic**. It is not the only one. Two other common approaches are

- **Frequentist**

Here a probability is the long run frequency of an event happening. Take our event A_1 : Rain and High above 90 degrees F tomorrow. We could go to some website (or meteorological office) and find out in how many of the last 10000 days it rained and the high was over 90, and then use the ratio as our probability.

There are two main problems with this approach. One is that we only know the "true" probability after observing infinitely many experiments, clearly impossible. The other one is what to do with an event like: The universe will explode in less than 10^{10} days. There is only one universe, so how can we run this experiment more than once?

- **Axiomatic**

Here a probability is simply defined as any assignment of numbers to events that satisfy Rules 1-3, without any regard to their meaning. Of course, probability theory is supposed to help us with real live events, so this is also not very satisfying.

These are not all the apporaches either, there are many others, all of them with some strenghts and some weaknesses.

1.1.2 Sufficiency of Rules 1-3

Let us now show that if your prices satisfy equations 1-3, you can not be made a sure loser. To do so we will have to use some concepts and results of probability theory which we will get to later in the course but which you likely have heard of before!

Suppose first that you announce price p for a ticket on event A. If you buy such a ticket it will cost you p, but you will gain \$1 if A occurs, and nothing otherwise. Thus your gain from

the transaction is exactly $I_A - p$, where I_A is the indicator function of A, that is $I_A = 1$ if A happened and 0 if not. If you sell such a ticket, your gain is $p - I_A$. Both of these can be represented by saying that your gain is $\alpha(I_A - p)$ where α is the number oftickets you buy. If α is negative, you sell α tickets.

With many such offers your total gain (or loss) is

$$W = \sum_{i=1}^{n} \alpha_i (I_{A_i} - p_i)$$

where your price on event A_i is p_i. The numbers α_i may be positive or negative, but are not in your control. But whatever choices of α's I make, positive or negative, W is the random variable that represents your gain, and it takes a finite number of values. Let's compute the expectation of W:

$$E[W] = E[\sum_{i=1}^{n} \alpha_i (I_{A_i} - p_i)] =$$

$$\sum_{i=1}^{n} \alpha_i (E[I_{A_i}] - p_i) =$$

$$\sum_{i=1}^{n} \alpha_i (P(A_i) - p_i) = 0$$

Now if you could be made a sure loser we would have $P(W < 0) = 1$ (there is no chance of you winning), but then $E[W] < 0$ as well, and we have a contradiction.

1.2 Basic Theorems

1.2.1 Basic Definitions

An **experiment** is a well-defined procedure that produces a set of outcomes. For example, "roll a die", "randomly select a card from a standard 52-card deck", "flip a coin" and "pick any moment in time between 10am and 12 am" are experiments.

A **sample space** is the set of outcomes from an experiment. Thus, for "flip a coin" the sample space is {H, T}, for "roll a die" the sample space is {1, 2, 3, 4, 5, 6} and for "pick

any moment in time between 10am and 12 am" the sample space is [10, 12].

An **event** is a subset, say A, of a sample space S. For the experiment "roll a die", an event is "obtain a number less than 3". Here, the event is {1, 2}.

1.2.2 Kolmogorov's Axioms of Probability

For any probability P we have

$$\text{Axiom 1} \quad 0 \leq P(A) \leq 1$$

$$\text{Axiom 2} \quad P(S) = 1$$

$$\text{Axiom 3} \quad P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

if $A_1, A_2, ..$ are mutually exclusive.

Note one difference between the axioms here and our discussion on coherence before. There we showed that in order to avoid being a sure looser we have to have

$P(A \cup B) = P(A) + P(B)$

if A and B are disjoint. The extension to a finite collection of disjoint sets is straightforward (via induction) but in the axioms we also allow an infinite collection. This is called **countable additivity**, and is an extension to the requirements of coherence as discussed before. It can be shown that without this extension there is another type of dynamic sure loss.

1.2.2.1 Example

say we have a sample space $S = \{e_1, ..., e_n\}$ and an event $A = \{e_{k_1}, ..., e_{k_m}\}$. Let's assume that all the events are **equally likely**. Then:

$$1 = P(S) = \qquad \text{Axiom 2}$$

$$P(\{e_1, e_2, .., e_n\}) = P(\bigcup_{i=1}^{n} \{e_i\} = \qquad \text{Axiom 3}$$

$$\sum_{i=1}^{n} P(\{e_i\}) = \sum_{i=1}^{n} p = np$$

$$\text{so } p = 1/n$$

$$P(A) = P(\{e_{i_1}, e_{i_2}, .., e_{i_k}\}) = P(\bigcup_{j=1}^{k} \{e_{i_j}\} = \qquad \text{Axiom 3}$$

$$\sum_{j=1}^{k} P(\{e_{i_j}\}) = \sum_{j=1}^{k} \frac{1}{n} = \frac{k}{n}$$

and so in this (very special) case finding a probability becomes a counting problem. We will discuss some formulas for this soon.

1.2.3 Set Theory

Recall the following formulas for sets:

Commutativity:

$A \cup B = B \cup A$ and $A \cap B = B \cap A$

Associativity

$A \cup (B \cup C) = (A \cup B) \cup C$
$A \cap (B \cap C) = (A \cap B) \cap C$

Distributive Law

$A \cup (B \cap C) = (A \cup B)) \cap (A \cup C)$
$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

DeMorgan's Law

$(A \cup B)^c = A^c \cap B^c$
$(A \cap B)^c = A^c \cup B^c$

1.2.4 Basic Theorems and Inequalities

Theorem (Addition Formula)

Let A and B be two sets, then

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof: first note that

$A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$

and that all of these are disjoint. Therefore by the third axiom we have

$P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B)$

but

$$
\begin{aligned}
&P(A \cap B^c) + P(A \cap B) + P(A^c \cap B) = \\
&\{P(A \cap B^c) + P(A \cap B)\} + P(A^c \cap B) + P(A \cap B) - P(A \cap B) = \\
&P((A \cap B^c) \cup (A \cap B)) + P((A^c \cap B) \cup (A \cap B)) - P(A \cap B) = \\
&P(A \cap (B^c \cup B)) + P((A^c \cup A) \cap B)) - P(A \cap B) = \\
&P(A \cap S) + P(S \cap B) - P(A \cap B) = \\
&P(A) + P(B) - P(A \cap B)
\end{aligned}
$$

Theorem (Bonferroni's Inequality)

Let A and B be two sets, then

$P(A \cap B) \geq P(A) + P(B) - 1$

proof follows directly from the addition formula and $P(A \cup B) \leq 1$.

Theorem (Complement)

Let A be any set, then

$P(A) = 1 - P(A^c)$

Proof: $S = A \cup A^c$, so

$1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$

Theorem (Subset)

Let A and B be two sets such that

$A \in B$ then $P(A) \leq P(B)$

proof:

$$
\begin{aligned}
&B = B \cap S = B \cap (A \cup A^c) = \\
&(B \cap A) \cup (B \cap A^c) = \\
&A \cup (B \cap A^c)
\end{aligned}
$$

so

$$
\begin{aligned}
&P(B) = P(A \cup (B \cap A^c) = \\
&P(A) + P(B \cap A^c) \geq P(B)
\end{aligned}
$$

Theorem (Boole's Inequality)

Let $A_1, .., A_n$ be a (finite) collection of sets, then

$P(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i)$

proof follows by induction from the above

1.2.5 Borel-Cantelli lemmas

Definition

Let $\{A_n, n \geq 1\}$ be a sequence of events. Then

a. the sequence is called increasing if $A_n \in A_{n+1}$

If $\{A_n, n \geq 1\}$ is an increasing sequence of events we define the new event $\lim A_n$ by

$\lim A_n = \cup_{n=1}^{\infty} A_n$

b) the sequence is called decreasing if $A_{n+1} \in A_n$

If $\{A_n, n \geq 1\}$ is a decreasing sequence of events we define the event $\lim A_n$ by

$\lim A_n = \cap_{n=1}^{\infty} A_n$

Theorem

If $\{A_n, n \geq 1\}$ is either an increasing or a decreasing sequence of events then

$\lim P(A_n) = P(\lim A_n)$

proof: Suppose first that $\{A_n, n \geq 1\}$ is an increasing sequence. Define the events F_n by

$F_1 = A_1$

$F_{n+1} = A_{n+1} \cap (\cup_{i=1}^{n} A_i)^c = A_{n+1} \cap A_n^c$

That is, F_n consists of those points that were not in any earlier $A_i, i = 1, .., n-1$

By their definition the F_n are mutually exclusive, so

$$P(\cup_{i=1}^{n} A_i) = P(\cup_{i=1}^{n} F_i) = \sum_{i=1}^{n} P(F_i) \ \forall n$$

so

$$P(\lim_{n\to\infty} A_n) =$$

$$P(\cup_{i=1}^{\infty} A_i) =$$

$$\sum_{i=1}^{\infty} P(F_i) =$$

$$\lim_{n\to\infty} \sum_{i=1}^{n} P(F_i) =$$

$$\lim_{n\to\infty} P(\cup_{i=1}^{n} F_i) =$$

$$\lim_{n\to\infty} P(\cup_{i=1}^{n} A_i) =$$

$$\lim_{n\to\infty} P(A_n)$$

The proof for a decreasing sequence $\{A_n, n \geq \}\$$ follows directly from the fact that then $\{A_n^c, n \geq 1\}$ is an increasing sequence.

1.2.5.1 Example

Consider a population consisting of individuals able to produce offspring of the same kind. The number of individuals initially present, denoted by X_0, is called the size of the zero'th generation. All offspring of the zero'th generation constitute the first generation and their number is denoted by X_1. In general, let X_n denote the size of the n^{th} generation.

Let $A_n = \{X_n = 0\}$. Now since $X_n = 0$ implies $X_{n+1} = 0$, it follows that $\{A_k, k \geq n\}$ is an increasing sequence and thus $\lim P(A_n)$ exists. What is the meaning of this probability? We

have

$$\lim P(X_n = 0) = \lim P(A_n) =$$
$$P(\lim A_n) =$$
$$P(\cup A_n) =$$
$$P(\cup\{X_n = 0\}) =$$
$$P(\text{the population eventually dies out})$$

Definition

Let $\{A_n\}$ be an infinite sequence of events. Then

$\{A_n$ i.o.$\}$ ("A_n infinitely often")

is the event that for any m there exists an n>m such that $P(A_n) > 0$.

Note: $\{A_n \text{ i.o.}\} = \cap_m \cup_{n \geq m} A_n$

Theorem (Borel-Cantelli lemma)

Let $A_1, A_2, ..$ be sequence of events. If $\sum P(A_i) < \infty$ then

$P(\{A_n \text{ i.o.}\}) = 0$.

proof Let's call the event that an infinite number of the A_i's occur $\limsup A_i$. Then

$$\limsup A_i = \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i$$

This is because if an infinite number of A_i's occur, then for any n there exists an m >n such that A_m occurs, therefore

$\cup_{i=n}^{\infty} A_i$ occurs, and then the intersection occurs as well.

Now $\cup_{i=n}^{\infty} A_i; n \geq 1\}$ is a decreasing sequence and so it follows that

$$P(\limsup A_i) =$$
$$P(\cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i) =$$
$$lim_{n\to\infty} P(\cup_{i=n}^{\infty} A_i) \leq$$
$$lim_{n\to\infty} \sum_{i=n}^{\infty} P(A_i) = 0$$

1.2.5.2 Example

Let $X_1, X_2, ..$ be such that $P(X_n = 0) = 1/n^2 = 1 - P(X_n = 1)$

Let $A_n = \{X_n = 0\}$, then

$\sum P(A_n) = \sum 1/n^2 < \infty$.

so it follows that the probability that X_n equals 0 for an infinite number of n is also 0. Hence, for an n sufficiently large X_n must equal 1.

1.2.6 Independence

Definition

a. Two events A and B are said to be **independent** if

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

b. A set of events $\{A_n, n \geq 1\}$ is said to be **pairwise independent** if for any i and j A_i and A_j are independent.

c. A set of events $\{A_n, n \geq 1\}$ is said to be **independent** if for any set of indices {i_1 ,..,i_n) we have

$$P(\cap_{i=1}^{n} A_i) = \prod_{i=1}^{n} P(A_i)$$

Theorem

Pairwise independence does not imply independence.

proof Consider the sample space $S = \{1, 2, 3, 4\}$ where all outcomes are equally likely. Define the events

$A = \{1, 2\}, B = \{1, 3\}$ and $C = \{1, 4\}$

then we have

P(A) = P(B) = P(C) = 1/2

and

$P(A \cap B) = P(B \cap C) = P(A \cap C) = 1/4$

so we have pairwise independence, but

$1/8 = P(A)P(B)P(C) \neq P(A \cap B \cap C) = 1/4$

Theorem (Second Borel-Cantelli lemma)

If $A_1, A_2, ..$ are independent events such that

$\sum P(A_i) = \infty$ then

$P(\{A_n \text{i.o.}\}) = 1$

proof: If an infinite number of the A_i's occur, then by the same reasoning as in the first Borel-Cantelli lemma we have $\limsup A_n$ occur, so we need to show that

$P(\limsup A_n) = 1$

or equally

$1 - P(\limsup A_n) = 0$

Now

$$1 - P(\limsup_{n\to\infty} A_n) =$$

$$1 - P(\{A_n \; i.o.\}) \;\; =$$

$$P(\{A_n \; i.o.\}^c) =$$

$$P(\{\cap_{m=1}^{\infty} \; \cup_{n=m}^{\infty} A_n\}^c) =$$

$$P(\cup_{m=1}^{\infty} \; \{\cup_{n=m}^{\infty} A_n\}^c) =$$

$$P(\cup_{m=1}^{\infty} \; \cap_{n=m}^{\infty} A_n^c) =$$

$$\lim_{m\to\infty} P(\cap_{n=m}^{\infty} A_n^c)$$

$$\lim_{m\to\infty} P(\cap_{n=m}^{\infty} A_n^c) = \prod_{n=m}^{\infty} P(A_n^c) =$$

$$\lim_{m\to\infty} \prod_{n=m}^{\infty} [1 - P(A_n)] \leq$$

$$\lim_{m\to\infty} \prod_{n=m}^{\infty} \left[1 - P(A_n) + \frac{P(A_n)^2}{2!} - \frac{P(A_n)^3}{3!} \pm . . \right] =$$

$$\lim_{m\to\infty} \prod_{n=m}^{\infty} \left[\sum_{k=0}^{\infty} (-1)^k \frac{P(A_n)^k}{k!} \right] =$$

$$\lim_{m\to\infty} \prod_{n=m}^{\infty} \exp\{-P(A_n)\} \;\; =$$

$$\lim_{m\to\infty} \exp\{-\sum_{n=m}^{\infty} P(A_n)\} \;\; =$$

$$\lim_{m\to\infty} \exp\{-\infty\} \; = \; 0$$

1.2.6.1 Example

consider the following game we start with an urn with one white ball and one black ball. Let

A_1 ={"white ball drawn"}

next we add a black ball and let A_2 again be the event {"white ball drawn"}. We continue on like that.

Now $P(A_n) = 1/(n+1)$. Clearly the A_i's are independent and

$\sum P(A_i) = \sum 1/(n+1) = \infty$

and therefore

$P(\{A_n \text{i.o.}\}) = 1$

So no matter how many balls are already in the urn, at some point in the future we will again pick a white one.

Say in each round we add k black balls, then the same

$\sum P(A_i) = 1/(1+1) + 1/(2+k) + 1/(2+2k) + .. = \sum 1/(2+nk) = \infty$

for any k>0, so the same conclusion holds!

If we take the two Borel Cantelli lemmas together we have the following: Let $\{A_n\}$ be a sequence of independent events, then

$P(\{A_n \text{i.o.}\}) = 0$ or 1

This is an example of a so called 0-1 law, of which there are quite a few in Probability Theory. Here is one of the famous ones:

Definition

Let $\{A_n\}$ be an infinite sequence of events. An event B is called a *tail event* if knowing whether or not A_n occurred for each n determines B, but B is independent of any finite collection of A_n's.

1.2.6.2 Example

$P({A_n i.o.}) is a tail event.

Theorem (Kolmogorov's 0-1 law)

Let B be a tail event of the sequence $\{A_n\}$. Then P(B) is either 0 or 1.

1.3 Conditional Probability and Independence

1.3.0.1 Example

A bag contains slips of paper. Each paper has a number and a letter written on them. They are: A5, A7, B1, B2, C2, C4, D2, D4, E1, E3. A slip is chosen at random, what is the probability it has a 2 on it?

Easy: 3/10

Now say somebody picks a slip and tells you it has the letter B on it. Now what is the probability it also has a 2 on it?

Again easy: 1/2

This is an example of a conditional probability. we write

P(#2 | Letter B)

(" probability of #2 **given** letter B)

Above we found the conditional probability by changing the sample space. First it was S={A5, A7, B1, B2, C2, C4, D2, D4, E1, E3} but once we knew the slip of paper had the letter B it changed to S={B1, B2}.

In general this changing of the sample space is too difficult, but we can find conditional probabilities using the formula

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

Note: this only works if P(B)>0.

Note: this formula can also be derived using the idea of coherence and the concept of the sure looser discussed earlier.

1.3.0.2 Example (cont.)

P(#2 | Letter B) =P(#2 $\cap$ Letter B) /P(Letter B) = (1/10) / (2/10) = 1/2.

It is important to notice that conditional probabilities are just like regular ones, for example they obey the axioms of Kolmogorov:

Axiom 1: $P(A|B) = P(A \cap B)/P(B)$, but $P(A \cap B)$ and P(B) are both regular probabilities, so $P(A \cap B) \geq 0$, $P(B) > 0$, so $P(A|B) = P(A \cap B)/P(B) \geq 0$.

Also $A \cap B \in B$, so $P(A|B) = P(A \cap B)/P(B) \leq P(B)/P(B) = 1$.

Axiom 2: $P(S|B) = P(S \cap B)/P(B) = P(B)/P(B) = 1$.

Axiom 3: say $A_1, .., A_n$ are mutually exclusive, then

$$P(\cup_{i=1}^{\infty} A_i|B) = \frac{P((\cup_{i=1}^{\infty} A_i)\cap B)}{P(B)} =$$

$$\frac{P(\cup_{i=1}^{\infty} (A_i \cap B))}{P(B)} = \frac{\sum_{i=1}^{\infty} P(A_i \cap B)}{P(B)} =$$

$$\sum_{i=1}^{\infty} \frac{P(A_i \cap B)}{P(B)} = \sum_{i=1}^{\infty} P(A_i|B)$$

1.3.1 Multiplication Rule

A simple manipulation of the equation above yields

$$P(A \cap B) = P(A|B)P(B)$$

1.3.1.1 Example

You draw two cards from a standard 52-card deck. What is the probability to draw 2 Aces?

Solution:

Let A = "First card drawn is an ace" Let B = "Second card drawn is an ace"

Then

$$\mathbf{P(\text{both cards drawn are aces}) =}$$

$$\mathbf{P(\text{first card drawn is an ace and second card drawn is an ace}) =}$$

$$\mathbf{P(A \cap B) = P(A) \cdot P(B|A) = \frac{4}{52} \cdot \frac{3}{51}}$$

It's easy to extend this to more than two events: What is the probability of drawing 4 aces when drawing 4 cards?

Let A_i = "i$^{\text{th}}$ card drawn is an ace"

Then

$$\mathbf{P(\text{all 4 cards drawn are aces}) =}$$

$$\mathbf{P(\text{first card drawn is an ace and second card drawn is an ace and third card drawn is an ace and fourth card drawn is an ace}) =}$$

$$\mathbf{P(A_1 \cap A_2 \cap A_3 \cap A_4) =}$$

$$\mathbf{P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdot P(A_4 \mid A_1 \cap A_2 \cap A_3) =}$$

$$\mathbf{= \frac{4}{52} \cdot \frac{3}{51} \frac{2}{50} \cdot \frac{1}{49}}$$

even a little more complicated: In most Poker games you get in the first round 5 cards (Later you can exchange some you don't like but we leave that out). What is the probability that you get 4 aces?

$$\text{P(4 of the 5 cards drawn are aces)} =$$
$$\text{P}\left(\left(A_1 \cap A_2 \cap A_3 \cap A_4 \cap \overline{A}_5\right) \cup \left(A_1 \cap A_2 \cap A_3 \cap \overline{A}_4 \cap A_5\right) \cup \left(A_1 \cap A_2 \cap \overline{A}_3 \cap A_4 \cap A_5\right)\right.$$
$$\left. \cup \left(A_1 \cap \overline{A}_2 \cap A_3 \cap A_4 \cap A_5\right) \cup \left(\overline{A}_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5\right)\right) =$$
$$\frac{4}{52} \cdot \frac{3}{51} \frac{2}{50} \cdot \frac{1}{49} \cdot \frac{48}{48} + \frac{4}{52} \cdot \frac{3}{51} \frac{2}{50} \cdot \frac{48}{49} \cdot \frac{1}{48} + \frac{4}{52} \cdot \frac{3}{51} \frac{48}{50} \cdot \frac{2}{49} \cdot \frac{1}{48} + \frac{4}{52} \cdot \frac{48}{51} \frac{3}{50} \cdot \frac{2}{49} \cdot \frac{1}{48} + \frac{48}{52} \cdot \frac{4}{51} \frac{3}{50} \cdot \frac{2}{49} \cdot \frac{1}{48}$$
$$= 5 \cdot \frac{4}{52} \cdot \frac{3}{51} \frac{2}{50} \cdot \frac{1}{49}$$

1.3.2 Law of Total Probability and Bayes Rule

Definition

A set of events $\{A_i\}$ is called a **partition** of the sample space S if

$$A_i \cap A_j) = \emptyset \text{ for all } i \neq j$$
$$\cup_{i=1}^{n} A_i = S$$

1.3.2.1 Example

a student is selected at random from all the undergraduate students at the Colegio

A_1 = "Student is female", A_2 = "Student is male"

or maybe

A_1 = "Student is freshman", .., A_4 = "Student is senior"

Theorem (Law of Total Probability)

Let the set of events {A_i } be a partition, and let B be any event, then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Proof

$$P(B) = P(B \cap S) =$$
$$P(B \cap (\cup_i^{\infty} = 1A_i) =$$
$$P(\cup_{i=1}^{\infty}(B \cap A_i) =$$
$$\sum_{i=1}^{\infty} P(B \cap A_i) =$$
$$\sum_{i=1}^{\infty} P(B|A_i)P(A_i)$$

1.3.2.2 Example

A company has 452 employees, 210 men and 242 women. 15% of the men and 10% of the women have a managerial position. What is the probability that a randomly selected person in this company has a managerial position?

Let A_1 = "person is female", A_2 = "person is male".

Let B = "person has a managerial position"

Then

$$P(A_1) = \frac{242}{452}$$
$$P(A_2) = \frac{210}{452}$$
$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) =$$
$$0.1\frac{242}{452} + 0.15\frac{210}{452} = 0.123$$

1.3.2.3 Example

Say you roll a fair die. If you roll an even number you roll the die again, otherwise you keep the result of the first roll. What are the probabilities of rolling a 1, or a 2 or..,..,6?

$$P(1) = \sum_{i=1}^{6} P(1|\text{first roll is } i)P(\text{first roll is } i) =$$
$$1\frac{1}{6} + \frac{1}{6}\frac{1}{6} + 0\frac{1}{6} + \frac{1}{6}\frac{1}{6} + 0\frac{1}{6} + \frac{1}{6}\frac{1}{6} = \frac{1}{6}(1 + \frac{1}{2}) = \frac{1}{4}$$

$$P(2) = \sum_{i=1}^{6} P(2|\text{first roll is } i)P(\text{first roll is } i) =$$
$$0\frac{1}{6} + \frac{1}{6}\frac{1}{6} + 0\frac{1}{6} + \frac{1}{6}\frac{1}{6} + 0\frac{1}{6} + \frac{1}{6}\frac{1}{6} = \frac{1}{6}(\frac{1}{2}) = \frac{1}{12}$$

and the same for 3-6

Theorem (Bayes' Formula)

Let the set of events $\{A_i\}$ be a partition, and let B be any event, then

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{i=1}^{n} P(B|A_i)P(A_i)}$$

Notice that the denominator is just the law of total probability, so we could have written the formula also in this way

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{P(B}$$

only usually the first form is the one that is needed because of the available information.

proof:

$$\begin{aligned}
&P(A_k|B) = \\
&P(A_k \cap B)/P(B) = \\
&P(B \cap A_k)/P(B) = \\
&P(B|A_k)P(A_k)/P(B)
\end{aligned}$$

1.3.2.4 Example

In the company above a person is randomly selected, and that person is in a managerial position. What is the probability the person is female?

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

$$\frac{0.1 \times 242/452}{0.123} = 0.434$$

Bayes formula sometimes results in strange answers:

1.3.2.5 Example

As part of the attempt to avoid further terrorist attacks on the US some people have proposed face-recognition technics for airports. Basically each person entering the security checkpoint of the airport is photographed and the digital picture is then compared to a list of pictures of known terrorist suspects. Such systems are never 100% correct, they do make an occasional mistake. Say that the system classifies an actual terrorist as ok 50% of the time (many terrorists won't be in the database because they have never been investigated). This is called a false-negative. Also say that the system wrongly classifies an ok person as a terrorist 0.1% of the time (false-positive). Say at some large airport there are 10 million passengers per year, 20 of whom are actually terrorists. What is the probability that a person classified as a terrorist by the face-recognition system actually is not a terrorist?

Let's use the following notation:

Let A_1 = "person is not a terrorist", A_2 = "person is a terrorist"

B = "person is classified as a terrorist"

Now

$$P(A_1) = \frac{9,999,980}{10,000,000} \quad P(A_2) = \frac{20}{10,000,000}$$

$$P(B|A_1) = 0.001 \quad P(B|A_2) = 0.5$$

$$P(B) = P(B|A_1) \cdot P(A_1) + P(B|A_2) \cdot P(A_2) =$$

$$0.001 \cdot \frac{9,999,980}{10,000,000} + 0.5 \cdot \frac{20}{10,000,000} = 0.001000998$$

$$P(A_1|B) = \frac{P(B|A_1) \cdot P(A_1)}{P(B)} = \frac{0.001 \cdot \frac{9,999,980}{10,000,000}}{0.001000998} = 0.9990$$

So only 1 in 1000 people "accused" by the system actually is a terrorist!

Note: in this calculation you need to carry along a lot of digits until the final answer.

Bayes' Rule plays a very important role in Statistics and in Science in general. It provides a natural method for updating your knowledge based on data.

1.3.3 Independence

Sometimes knowing that one event occurred does not effect the probability of another event. For example if you through a red and a blue die, knowing that the red die shows a "6" will not change the probability that the blue die shows a "2".

Formally we have

$$P(A|B) = P(A)$$

or using the multiplication rule we get the previous formula for two independent events

$$P(A \cap B) = P(A|B)P(B) = P(A)P(B)$$

1.3.3.1 Example

Say you flip a fair coin 5 times. What is the probability of 5 "heads"?

Let A_i = "i^{th} flip is heads"

Now it is reasonable to assume that the A_i's are independent and so

$$P(A_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5) =$$
$$P(A_1) \cdot P(A_2) \cdot P(A_3) \cdot P(A_4) \cdot P(A_5)$$
$$= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 2^{-5} = 0.03125$$

1.4 Combinatorics

We have previously seen that if we have a finite sample space S and all the outcomes are equally likely, then P(A)=n(A)/n(S), so finding probabilities means counting the number of outcomes in A and in S. The mathematical theory that deals with counting is called *combinatorics*. Here we will consider some special cases and their formulas.

1.4.1 Fundamental Theorem of Counting

Theorem (Fundamental Theorem of Counting)**

If a job consists of k separate tasks, the i^{th} of which can be carried out in n_i ways, the entire job can be done in $n_1 \times n_2 \times .. \times n_k$ ways

proof

say k=2, say task 1 has outcomes a_1 ,..,a_n and task 2 has outcomes b_1 ,..,b_m , then clearly there are $n \times m$ combinations. The general case follows from mathematical induction.

1.4.1.1 Example

You possess 5 pairs of shoes, 10 pairs of socks, 4 pairs of trousers and 9 shirts. How many combinations of outfits are there?

$5 \times 10 \times 4 \times 9$

1.4.2 Basic Counting Formulas

Many of the problems in combinatorics are variations of the following: say we have a box with n balls, numbered 1 to n, and we select k of them. In how many ways can this be done? In order to answer this question we need to be more specific on how the draws are done:

Case 1: with order and with repetition

Balls are drawn as follows: pick one ball, write down the number, replace the ball in the box, draw a second ball etc. In this case we have the order in which the balls are drawn, and we have repetition , that is the same ball can be chosen more than once.

Say n=10 and k=3, then some possible outcomes are: (7,2,3), (7,6,7), (6,7,7)

According to the Fundamental Theorem of counting this each task (drawing a ball) can be done in n ways, and there are k tasks, so the total number of ways is

$n \times n \times n \times ... \times n = n^k$

Case 2: with order but without repetition

Balls are drawn as follows: pick one ball, write down the number, put the ball aside, not back in the box, draw a second ball etc. In this case we have the order in which the balls are drawn but each ball can be drawn only once, so there is no repetition

Say n=10 and k=3, then some possible outcomes are: (7,2,3), (7,6,4) but not (6,7,7)

According to the Fundamental Theorem of counting the first task (drawing the first ball) can be done in n ways, the second task can be done in n-1 ways and so on until the kˆth task which can be done in n-k+1 ways, so the total number of ways is

$n \times (n-1) \times (n-2) \times ... \times (n-k+1)$

This is often call the number of *permutations* of n objects, k at a time, and we use the notation P_k^n.

An important special case is k=n, which is just called the permutations of n objects.

Definition

n!=n(n-1)(n-2)..1 is called "n factorial"

Note by definition 0!=1

with this definition we have $P_k^n = n!/(n-k)!$.

Case 3: without order and without repetition

Balls are drawn as follows: pick all the balls simultaneously. In this case we have no order and no repetition

Say n=10 and k=3, then some possible outcomes are: (7,2,3), (7,6,4) but (6,4,7) is the same as (7,6,4)

This is called the number of combinations of n objects, k at a a time, and is denoted by C_k^n.

To do this think in terms of a two tasks: first select without order and without repetition (which can be done in C_k^n ways) and then order the k selected objects (in k! ways) The result is a selection with order but without repetition, but this is P_k^n!. So we find:

$P_k^n = C_k^n \times k!$, or $C_k^n = n!/(n-k)!/k!$

Definition (of binomial coefficients)

$$C_k^n = \frac{n!}{(n-k)!k!} = \binom{n}{k}$$

We say "n choose k"

Case 4: without order but repetition

as is case 1, but the order is now irrelevant. This is somewhat more complicated, but the answer is

$$\binom{n+k-1}{k}$$

1.4.2.1 Example

How many different license plates can there be in PR? A license plate has three letters and three numbers, order is important and there is repetition, there are

$26 \times 26 \times 26 \times 10 \times 10 \times 10 = 17,576,000$

possible plates

1.4.2.2 Example: Poker

Poker is played in a large number of different ways. Here we will keep it simple: we have a deck of 52 cards. Each card has a suit (Hearts, Diamonds, Clubs and Spades) and a denomination (2-10, Jack, Queen, King and Ace). A "hand" is any selection of 5 cards. The order is not important, and each combination of suit-denomination appears only once, so selection is done without order and without repetition.

How many 5-card hands are there?

$C^5 2_5 = 52!/47!/5! = 2598960$

What is the probability of a "four of a kind", that is four cards of the same denomination?

First choose the denomination (13 ways), net select all those cards (1 way), finally choose a card from the rest of the deck (48 ways) so

$$P(\text{four of a kind}) = 13 \times 1 \times 48/2598960 = 0.00024$$

.

What is the probability of a "full house", that is three cards of one denomination and two cards of a second denomination?

First choose the denomination for the three cards (13 ways) next pick the three cards (Cˆ4_3 =4 ways) then pick the denominations for the two cards (12 ways) and finally pick those two cards (Cˆ4_2 =6 ways), so

$$P(\text{full house}) = 13 \times 4 \times 12 \times 6/2598960 = 0.00144$$

Notice that the probability of a "four of a kind" is smaller than the one for "full house", and there fore a "four of a kind" beats a "full house"

1.4.3 Pidgeon Hole Principle

The Pidgeon Hole Principle states that if there are n pidgeons, m pidgeon holes and n>m, there must be at least one hole with more than one pidgeon.

This completely obvious principle has many, often surprising, consequences. One of the most famous is this one:

There must be at least two people in Puerto Rico with the exact same number of hairs on their heads!

We can demonstrate thisas follows. Since a typical human head has an average of around 150,000 hairs, it is reasonable to assume (as an upper bound) that no one has more than 1,000,000 hairs on their head (m = 1 million holes). There are more than 1,000,000 people in San Juan (n is bigger than 1 million items). Assigning a pigeon hole to each number of hairs on a person's head, and assigning people to pigeon holes according to the number of hairs on their head, there must be at least two people assigned to the same pigeon hole by the 1,000,001st assignment.

Perhaps the first written reference to the pigeon hole principle appears in 1622 in a short sentence of the Latin work *Selectae Propositiones*, by the French Jesuit Jean Leurechon, where he wrote "It is necessary that two men have the same number of hairs, or other things, as each other".

Combinatorics is a huge area of mathematics, with applications in almost all fields of science.

1.5 Random Variables and Distribution Functions

Definition

A *random variable* (r.v.) X is set-valued function from the sample space into $\mathbb{R}$. For any set of real numbers $A \in \mathbb{R}$ we define the probability

$$P(X \in A) = P(X^{-1}(A))$$

where $X^{-1}(A)$ is the set of all points in S such that X maps the points into A.

1.5.0.1 Example

Say we flip a fair coin three times. Let X be the number of "heads" in these three flips.

Now S=({H,H,H}, (H,H,T), .., (T,T,T)}.

X maps S into $\mathbb{R}$, for example X({(H,H,H)})=3 and X({(H,H,T)})=2.

What is P(X=2)?

P(X=2) =
$P(X^{-1}(2))$ =

P(all the outcomes in S that are mapped onto 2) =
P({(H,H,T), (H,T,H), (T,H,H)} = 3/8

There are two basic types of r.v.'s:

- If X takes countably many values, X is called a **discrete** r.v.
- If X takes uncountably many values, X is called a **continuous** r.v.

1.5.0.2 Example

Consider the following experiment: we randomly select a point in the interval [A,B] for some A<B. We allow all points in [A,B], so X takes uncountably many values, and therefore is a continuous random variable. By "randomly" we mean that the probability for a chosen point to be in some interval depends only on the length of the interval. Let X be the point chosen. Clearly

1 = P(A<X<B)

Let A<a<b<B. Now the interval (a,b) has length b-a, the interval (A,B) has length B-A, and we have

a-b = [(b-a)/(B-A)]*(B-A) = c(B-A)

where c=[(b-a)/(B-A)]

therefore

P(a<X<b) = cP(A<X<B) = c*1 = (b-a)/(B-A)

This is a standard random variable. We often use the following notation:

$X \sim U[A, B]$

There are some technical difficulties when defining a r.v. on a sample space like $\mathbb{R}$, it turns out to be impossible to define it for every subset of $\mathbb{R}$ without getting logical contradictions. The solution is to define a *σ-algebra* on the sample space and then define X only on that σ-algebra.

The most commonly used σ-algebra is the Borel σ-algebra, which is the union and intersection of all intervals of the type (a, b), (a, b], [a, b) where a and b can be $\pm\infty$. All of this belongs to the brach of mathematics called *measure theory*. In what follows we will ignore these technical difficulties.

1.5.0.3 Example

In the example of the uniform random variable above we defined probabilities only for intervals. It turns out that this is all that is needed. In fact the set of all unions and intersections of intervals forms a σ-algebra on the real line. There are however also sets on the real line that can not be expressed as the union and intersection of intervals. For those probabilities are then not defined.

Almost everything to do with r.v.'s has to be done twice, once for discrete and once for continuous r.v.'s. This separation is only artificial, it goes away once a more general definition of "integral" is used (Riemann-Stilties or Lebesgue).

1.5.1 (Cumulative) Distribution Function

Definition

The (cumulative) distribution function (cdf) of a r.v. X is defined by

$$F(x) = P(X \leq x) \text{ for all } x \in \mathbb{R}$$

1.5.1.1 Example

$X \sim U[A, B]$

x<A: F(x) = P(X<x) = 0
A<x<B: F(x) = P(X<x) = P(A<X<x) = (x-A)/(B-A)
x>B: F(x) = P(X<x) = 1

Theorem

Let F be the cdf of some random variable X. Then

1. cdf's are standard functions on $\mathbb{R}$
2. $0 \leq F(x) \leq 1$
3. cdf's are non-decreasing
4. cdf's are right-continuous
5. $$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to \infty} F(x) = 1$$

proof

1. probabilities are unique so $F(x) = P(X \leq x) = P(\{\omega \in S : X(\omega) \leq x\})$ is unique
2. $0 \leq P(X \leq x) = P(\{\omega \in S : X(\omega) \leq x\}) \leq 1$ from axiom 1

3) say x<y, then

$\{\omega \in S : X(\omega) \leq x\} \subset \{\omega \in S : X(\omega) \leq y\}$

and so

$F(x) = P(X \leq x) \leq P(X \leq y) = F(y)$

4) a function F is right continuous if

$$\lim_{h\downarrow 0} F(x+h) = \lim_{h\downarrow 0} P(X \le x+h) =$$

$$\lim_{h\downarrow 0} P(\{\omega : X(\omega) \le x+h\}) =$$

$$\lim_{h\downarrow 0} P(\{\omega : X(\omega) \le x\} \cup \{\omega : x < X(\omega) \le x+h\}) =$$

$$\lim_{h\downarrow 0} [P(\{\omega : X(\omega) \le x\}) + P(\{\omega : x < X(\omega) \le x+h\})] =$$

$$P(\{\omega : X(\omega) \le x\}) + \lim_{h\downarrow 0} P(\{\omega : x < X(\omega) \le x+h\}) =$$

$$F(x) + \lim_{h\downarrow 0} P(x < X \le x+h)$$

$$\text{Let } A_{x,h} = \{\omega : x < X(\omega) \le x+h\}$$

$$\text{Then } \lim_{h\downarrow 0} A_{x,h} = \cap_h A_{x,h} = \varnothing$$

the limit is the same as the intersection because if h<k $A_{x,h} \subset A_{x,k}$, and the intersection is the empty set otherwise there exists y such y $\in A_{x,h}$ for all h, that is $x < y \le x+h$, a contradiction.

5) similar to 4.

Theorem

Let F be function that is increasing, right-continuous and has $0 \le F(x) \le 1$, then there exists a random variable that has F as its cdf.

proof: too deep for us

Theorem

Let F be the cdf of a rv X. Then F has at most countably many points of discontinuity.

proof

F is increasing so any point of discontinuity is a jump point up. Let A_n be the set of all point where F jumps up by more then 1/n. Then $|A_n|$<n because 0<F<1. Let A be the set of all jump points of F, then

$A = \cup A_n$

and therefore A is countable.

Note another consequence of this proof: for any $\epsilon > 0$ there are at most finitely many points where F jumps up by more then ϵ.

1.5.1.1.1 Example

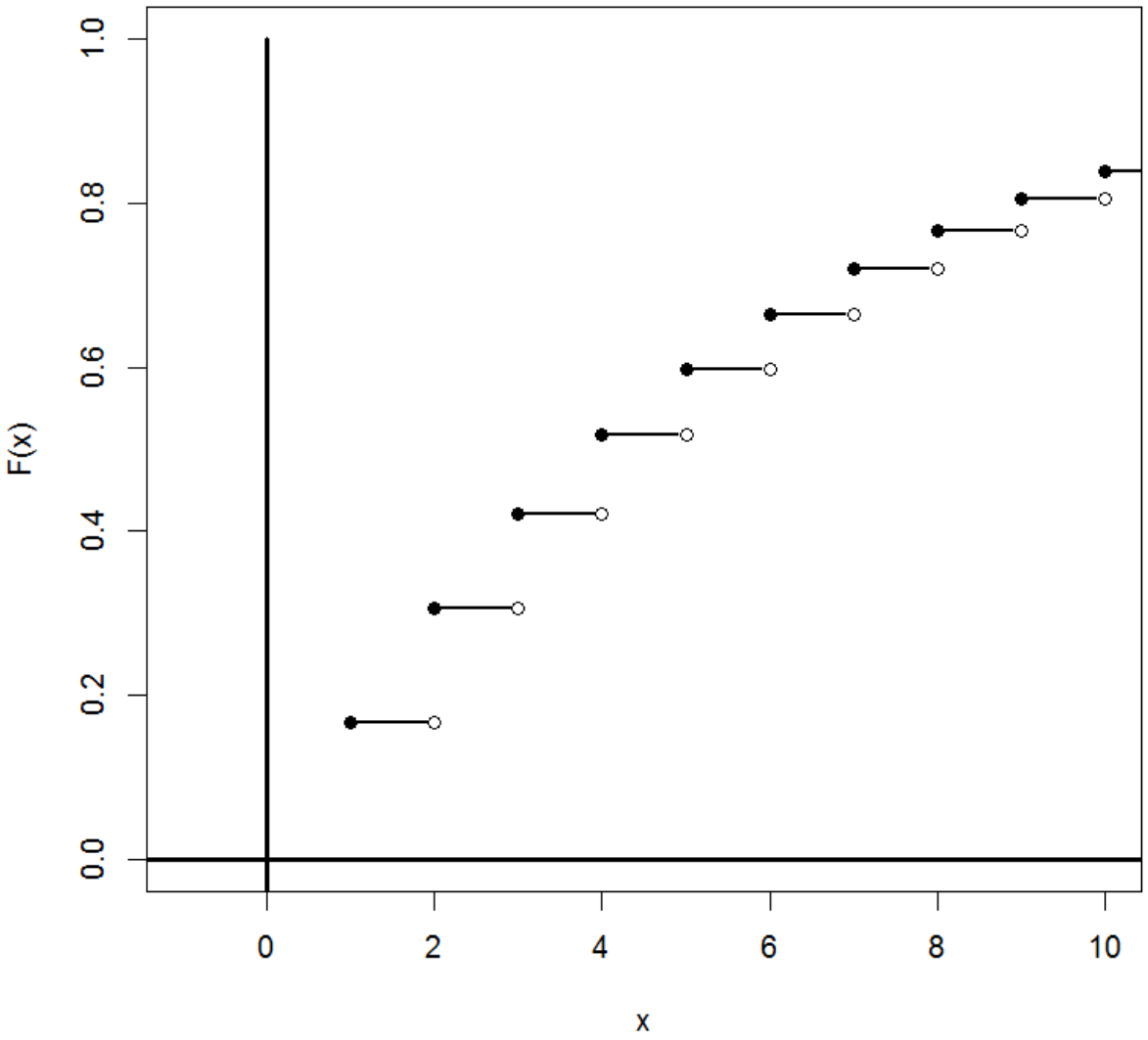
We roll a fair die until the first "6". Let the rv X be the number of rolls. Find the cdf F of X.

Solution: note $X \in \{1, 2, 3, ...\}$

let A_i be the event "a six on the i^{th} roll", i=1,2,3, ... Then

$$P(X = k) = P(A_1^c \cap A_2^c \cap .. \cap A_{k-1}^c \cap A_k) =$$
$$P(A_1^c)P(A_2^c)..P(A_{k-1}^c)P(A_k) = (\frac{5}{6})^{k-1}\frac{1}{6}$$
$$P(X \leq k) = \sum_{i=1}^{k} P(X = k) =$$
$$\sum_{i=1}^{k}(\frac{5}{6})^{i-1}\frac{1}{6} =$$
$$\sum_{j=0}^{k-1}(\frac{5}{6})^{j}\frac{1}{6} =$$
$$\frac{1}{6}\frac{1-(5/6)^{(k-1)+1}}{1-5/6} = 1-(5/6)^k$$

so for $k \leq x < k+1$ we have F(x)=1-$(5/6)^k$



Notice that the cdf is a step function. This is always the case for a discrete random variable.

1.5.2 Probability Density Function

The probability density function (pdf) of a discrete r.v. X is defined by
$f(x) = P(X = x)$

1.5.2.1 Example

the pdf of X in the example above is given by f(x) = $1/6*(5/6)^{x-1}$ if $x \in \{1, 2, ..\}$, 0 otherwise.
Note that it follows from the definition and the axioms that for any pdf f we have

$$f(x) \geq 0$$
$$\sum_x f(x) = 1$$

1.5.2.2 Example

Say f(x)=c/x^2, x=1,2,3,... is a pdf. Find c

$1 = \sum_x f(x) = \sum_{i=1}^{\infty} c/i^2 = c\pi^2/6$

so $c = 6/\pi^2$.

1.5.2.3 Example

say we have a coin that comes up heads with probability p. Let X be the number of heads in n flips of the coin. Find the pdf of X.

Let's start with a couple of small n's:

n=1:

S={H,T}
P(X=0)=P(T)=1-p P(X=1)=P(H)=p

n=2:

S={(H,H),(H,T),(T,H),(T,T)}
P(X=0) = P((T,T))=P(T)P(T) = (1-p)^2
P(X=1) = P({(H,T),(T,H)}) = P(H)P(T)+P(T)P(H) = 2p(1-p)
P(X=2)=P((H,H))=P(H)P(H)=p^2

n=3:

S={(H,H,H),(H,H,T),(H,T,H),(H,T,T),(T,H,H),(T,H,T),(T,T,H),(T,T,T)}
P(X=0) = P((T,T,T))=P(T)P(T)P(T)=(1-p)^3
P(X=1) = P({(H,T,T),(T,H,T),(T,T,H)}) = 3P(H)P(T)P(T) = 3p(1-p)^2
P({(H,T,T),(H,T,H),(T,H,H)}) = 3P(H)P(H)P(T) = 3p^2(1-p)
P(X=3)=P((H,H,H))=P(H)P(H)P(H)=p^3

apparently for some n we have something like

$P(X = k) = c_{n,k} p^k (1-p)^{n-k}$

what is c_n,k ? From the first few cases one might guess

$c_{n,k} = \binom{n}{k}$

Let's verify this using an induction proof. For this we will use the law of total probability, conditioning on whether the last flip was a heads or a tails. Also, Let X_n be X if we flip the coin n times. Now

$$P(X_n = k) =$$

$$P(X_{n-1} = k-1|H \text{ in last flip})P(H) + P(X_{n-1} = k|T \text{ in last flip})P(T) =$$

$$\binom{n-1}{k-1}p^{k-1}(1-p)^{n-k}p + \binom{n-1}{k}p^k(1-p)^{n-1-k}(1-p) =$$

$$\left[\frac{(n-1)!}{(n-k)!(k-1))!} + \frac{(n-1)!}{(n-1-k)!k!}\right]p^k(1-p)^{n-k} =$$

$$\left[\frac{k}{n}\frac{n!}{(n-k)!k!} + \frac{n-k}{n}\frac{n!}{(n-k)!k!}\right]p^k(1-p)^{n-k} =$$

$$\left[\frac{k}{n} + 1 - \frac{k}{n}\right]\frac{n!}{(n-k)!k!}p^k(1-p)^{n-k} = \frac{n!}{(n-k)!k!}p^k(1-p)^{n-k}$$

Definition

f is the pdf of the continuous random variable X iff

$F(x) = \int_{-\infty}^{x} f(tdt)$

or (if the cdf is differentiable at x)

$f(x) = F'(x)$

Again it follows from the definition and the axioms that for any pdf f we have

$$f(x) \geq 0$$
$$\int_{-\infty}^{\infty} f(x)dx = 1$$

1.5.2.4 Example

$X \sim U[A, B]$

$x < A$ or $x > B$: f(x)=F'(x) = 0
$A < x < B$: f(x)=F'(x) = d/dx[(x-A)/(B-A)] = 1/(B-A)
at x=A and x=B f is not defined because F is not differentiable

1.5.2.5 Example:

Show that $f(x) = \lambda \exp(-\lambda x)$ if x>0, 0 otherwise defines a pdf, where $\lambda > 0$.

Solution: clearly $f(x) \geq 0$ for all x.

$$\int_{-\infty}^{\infty} f(x)dx =$$
$$\int_{0}^{\infty} \lambda \exp(-\lambda x)dx =$$
$$-\exp(-\lambda x)|_0^{\infty} = 0 - (-1) = 1$$

This r.v. X is called an *exponential* r.v. with rate λ. We often write $X \sim \text{Exp}(\lambda)$.

1.5.2.6 Example

Say f(x)=c/x^2, x>1 is a pdf. Find c.

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_{1}^{\infty} \frac{c}{x^2}dx = (-c\frac{1}{x}|_1^{\infty} = c$$

1.5.2.7 Example

Say $f(x) = cx\sin(\pi x)$, $0 \le x \le 1$, 0 otherwise, is a pdf. Find c.

$$1 = \int_{-\infty}^{\infty} f(x)dx = \int_0^1 cx\sin(\pi x)dx =$$

$$c\left(x \cdot (-\tfrac{1}{\pi}\cos(\pi x))|_0^1 - \int_0^1 -\tfrac{1}{\pi}\cos(\pi x)dx\right) =$$

$$c\left(\tfrac{1}{\pi} + \tfrac{1}{\pi}\int_0^1 \cos(\pi x)dx\right) =$$

$$c\left(\tfrac{1}{\pi} + \tfrac{1}{\pi^2}\sin(\pi x)|_0^1\right) = \tfrac{c}{\pi}$$

$$\text{SO } c = \pi$$

1.5.2.8 Example

Say f(x)=cexp{-x^2} is a pdf. Find c.

Unfortunately f does not have an anti-derivative, so this is tricky problem. Using Numerical integration one can show that c=0.8547.

Although we usually deal with random variables that are either discrete or continuous, in real life they can be mixed:

1.5.2.9 Example

Consider the following experiment: first we flip a fair coin. If the coin comes up heads we roll a fair die and X is the number on the die, otherwise we choose $X \sim U[1,6]$. Find the cdf of X

$$F(x) = 0 \text{ if } x < 1$$

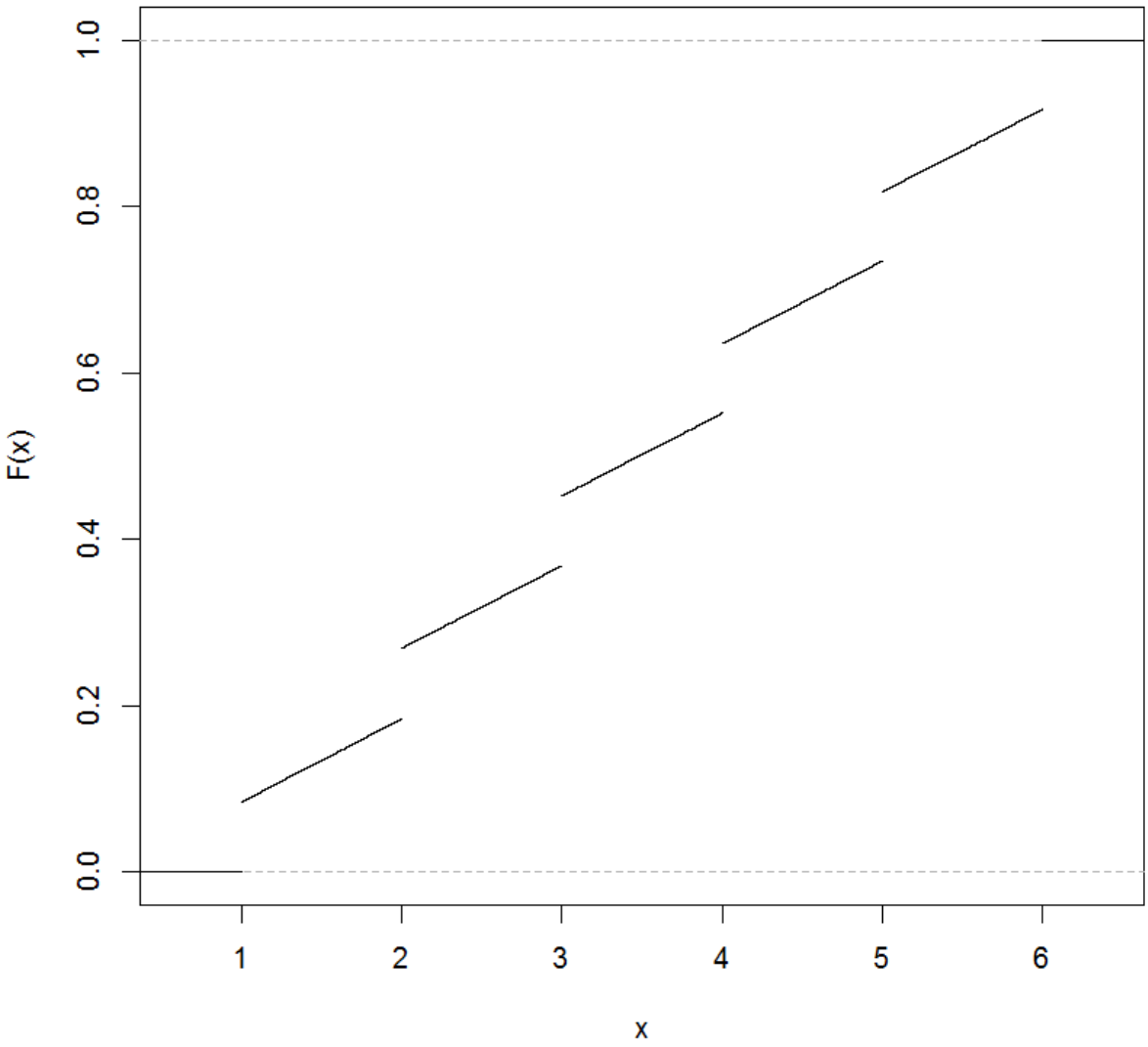
$$F(x) = 1 \text{ if } x > 6$$

$$\text{let } 1 \leq x \leq 6$$

$$F(x) = P(X \leq x) =$$

$$P(X \leq x|H)P(H) + P(X \leq x|T)P(T) =$$

$$\begin{cases} 0 & x < 1 \\ \frac{1}{6}\frac{1}{2} + \frac{x-1}{5}\frac{1}{2} & 1 \leq x < 2 \\ \frac{2}{6}\frac{1}{2} + \frac{x-1}{5}\frac{1}{2} & 2 \leq x < 3 \\ \frac{3}{6}\frac{1}{2} + \frac{x-1}{5}\frac{1}{2} & 3 \leq x < 4 \\ \frac{4}{6}\frac{1}{2} + \frac{x-1}{5}\frac{1}{2} & 4 \leq x < 5 \\ \frac{5}{6}\frac{1}{2} + \frac{x-1}{5}\frac{1}{2} & 5 \leq x < 6 \\ 1 & x \geq 6 \end{cases} = \begin{cases} 0 & x < 1 \\ (\lfloor x \rfloor + \frac{6}{5}(x-1))\frac{1}{12} & 1 \leq x < 6 \\ 1 & x \geq 6 \end{cases}$$



1.5.2.10 Example (Simple Random Walk)

Let X_0 =0, and P(X_n =1)=p, P(X_n =-1)=q=1-p

let $S_n = \sum_{k=0}^{n} X_k$ and define the event $A_n = \{S_n = 0\}$.

We want to find $P(\{A_n \text{ i.o. }\})$. By Kolmogorov's 0-1 law we know that it is either 0 or 1. But which is it?

Let $P_{00}^{(n)} = P(A_n)$, then

$P_{00}^{(2n+1)} = 0, n = 1, 2, ..$

$P_{00}^{2n} = P$(n jumps left and n jumps right)=

$\binom{2n}{n} p^n q^n = \frac{(2n)!}{n!^2} p^n q^n$

Sterling's formula:

$n! \sim n^{n+0.5} e^{-n} \sqrt{2\pi}$

$P_{00}^{2n} = \frac{(2n)^{2n+0.5} e^{-2n} \sqrt{2\pi}}{(n^{n+0.5} e^{-n} \sqrt{2\pi})^2} (pq)^n =$

$\frac{(2n)^{2n+0.5}}{n^{2n+1} \sqrt{2\pi}} (pq)^n =$

$\frac{(2n)^{2n+1}}{n^{2n+1} \sqrt{2\pi} \sqrt{2n}} (pq)^n =$

$\frac{2^{2n}}{\sqrt{\pi n}} (pq)^n = \frac{(4pq)^n}{\sqrt{\pi n}}$

Now

$pq = p(1-p) \leq \frac{1}{4}$

$4pq \leq 1$

and

$4pq = 1$ iff $p = \frac{1}{2}$

so

$\sum_{n=1}^{\infty} P_{00}^{(n)} = \sum_{n=1}^{\infty} \frac{(4pq)^n}{\sqrt{\pi n}}$

$\begin{cases} = \infty & \text{if } p = \frac{1}{2} \\ < \infty & \text{if } p \neq \frac{1}{2} \end{cases}$

and by the Borel-Cantelli lemmas we see that $P(\{A_n \text{ i.o. }\}) = 1$ if p=1/2 and 0 otherwise.

1.6 Random Vectors

A random vector is a multi-dimensional random variable.

1.6.0.1 Example 1

we roll a fair die twice. Let X be the sum of the rolls and let Y be the absolute difference between the two rolls. Then (X,Y) is a 2-dimensional random vector. The joint pdf of (X,Y) is given by:

	0	1	2	3	4	5
2	1	0	0	0	0	0
3	0	2	0	0	0	0
4	1	0	2	0	0	0
5	0	2	0	2	0	0
6	1	0	2	0	2	0
7	0	2	0	2	0	2
8	1	0	2	0	2	0
9	0	2	0	2	0	0
10	1	0	2	0	0	0
11	0	2	0	0	0	0
12	1	0	0	0	0	0

where every number is divided by 36.

---

all definitions are straightforward extensions of the one-dimensional case.

1.6.0.2 Example

for a discrete random vector we have the pdf f(x,y) = P(X=x,Y=y)

Say f(4,0) = P(X=4, Y=0) = P({(2,2)}) = 1/36 or f(7,1) = P(X=7,Y=1) = P({(3,4),(4,3)}) = 1/18

Example

Say f(x,y)=cxy is a pdf with $x \in \{1,2,3\}$ and $y \in \{0,2\}$. Find c.

$1 = \sum_{x,y} f(x,y) =$
$f(1,0) + f(1,2) + f(2,0) + f(2,2) + f(3,0) + f(3,2) =$
$c(1 \times 0 + 1 \times 2 + 2 \times 0 + 2 \times 2 + 3 \times 0 + 3 \times 2) = 12c$ so $c = 1/12$

1.6.0.3 Example

Say $f(x,y) = cxy, 0 \leq x, y \leq 1$ is a pdf. Find c.

$$1 = \int\int_{R^2} f(x,y)d(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = \int_0^1\int_0^1 xydydx =$$
$$\int_0^1 x\left(\int_0^1 ydy\right)dx = \int_0^1 x(\tfrac{1}{2}y^2|_0^1 dx = \int_0^1 x\tfrac{1}{2}dx = \tfrac{1}{2}\cdot(\tfrac{1}{2}x^2|_0^1 = \tfrac{1}{4}$$

so c=4.

1.6.0.4 Example

Say $f(x,y) = cxy, 0 \leq x < y \leq 1$ is a pdf. Find c.

$$1 = \int\int_{R^2} f(x,y)d(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dydx = \int_0^1\int_x^1 xydydx =$$
$$\int_0^1 x\left(\int_x^1 ydy\right)dx = \int_0^1 x(\tfrac{1}{2}y^2|_x^1 dx = \int_0^1 x\tfrac{1}{2}(1-x^2)dx = \tfrac{1}{2}\cdot\int_0^1 x(1-x^2)dx =$$
$$\tfrac{1}{2}\cdot\int_0^1 x - x^3 dx = \tfrac{1}{2}\cdot(\tfrac{1}{2}x^2 - \tfrac{1}{4}x^4|_0^1 = \tfrac{1}{2}(\tfrac{1}{2} - \tfrac{1}{4} - 0) = \tfrac{1}{2}\tfrac{1}{4} = \tfrac{1}{8}$$

so c=8.

1.6.0.5 Example

Say (X,Y) is a discrete rv with joint pdf $f(x,y) = cp^x, x, y \in \{0,1,..\}, y \leq x, 0 < p < 1$. Find c

$$1 = \sum_{x,y} f(x,y) = \sum_{y=0}^{\infty}\sum_{x=y}^{\infty} cp^x = c\sum_{y=0}^{\infty}\left(\sum_{x=0}^{\infty} p^x - \sum_{x=0}^{y-1} p^x\right) =$$
$$c\sum_{y=0}^{\infty}\left(\frac{1}{1-p} - \frac{1-p^y}{1-p}\right) = \frac{c}{1-p}\sum_{y=0}^{\infty} p^y = \frac{c}{(1-p)^2} \Rightarrow c = (1-p)^2$$

1.6.0.6 Example

Say (X,Y,Z) is a continuous rv with f(x,y,z) = c(x+y)z if $0 < x, y, z < 1$ and 0 otherwise. Find c

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y,z)d(x,y,z) = \int_0^1\int_0^1\int_0^1 c\ (x+y)zdxdydz =$$
$$c\int_0^1 z(\int_0^1 xdx + \int_0^1 ydy)dz = c\int_0^1 z(\tfrac{x^2}{2}|_0^1 + \tfrac{y^2}{2}|_0^1)dz =$$
$$c\tfrac{z^2}{2}|_0^1 = \tfrac{c}{2} = 1$$

so c=2

1.6.0.7 Example

Let's extend the idea of a uniform random variable to two dimensions. To begin, let's start with the unit square $[0,1]_2$.

Again, the idea of uniform is taken to mean that the probability of a point (X,Y) being in some area is proportional to the size of the area. Therefore if A is some area in $[0,1]_2$ we have

$P((X,Y) \in A) = \text{area}(A)$

say $0 < x, y < 1$, then

$F(x,y) = P(X < x, Y < y) = \text{area}([0,x] * [0,y]) = xy$

$f(x,y) = d^2/dxdyF(x,y) = d/dx[d/dy(xy)] = d/dx[x] = 1$

Now say (X,Y) is uniform on $\{(x,y\} : 0 < x < y^\alpha < 1\}$ for some $\alpha > 0$. Find the joint pdf of (X,Y).

First we need the total area:

$$\int_0^1\int_0^{y^\alpha} dxdy = \int_0^1 y^\alpha dy = \tfrac{1}{\alpha+1}y^{\alpha+1}|_0^1 = \tfrac{1}{\alpha+1}$$

so $f(x,y) = \alpha + 1$ if $0 < x < y^\alpha < 1$.

1.6.1 Marginal Distributions

Definition

Say (X,Y) is a discrete (continuous) r.v. with joint pdf (pdf) f. Then the *marginal* pdf f_X is given by

$$f_X(x) = \sum_y f(x,y) \text{ if } X \text{ is discrete}$$

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \text{ if } X \text{ is continuous}$$

1.6.1.1 Example

Say X is the sum and Y is the absolute difference of two dice. If we add the row and column totals to the table above we get

	0	1	2	3	4	5	X
2	1	0	0	0	0	0	1
3	0	2	0	0	0	0	2
4	1	0	2	0	0	0	3
5	0	2	0	2	0	0	4
6	1	0	2	0	2	0	5
7	0	2	0	2	0	2	6
8	1	0	2	0	2	0	5
9	0	2	0	2	0	0	4
10	1	0	2	0	0	0	3
11	0	2	0	0	0	0	2
12	1	0	0	0	0	0	1
Y	6	10	8	6	4	2	36

and these are the marginals. For example we find $f_X\ (2) = 1/36$ or $f_Y\ (3) = 6/36$.

1.6.1.2 Example

Say (X,Y) is a rv with joint pdf f(x,y)=xy/12 with $x \in \{1,2,3\}$ and $y \in \{0,2\}$ Now

$f_X\ (3)$ = f(3,0) + f(3,2) = 3x0x1/12 + 3x2x1/12 = 6/12 = 1/2

$f_Y\ (0)$ = f(1,0) + f(2,0) + f(3,0) = 0

1.6.1.3 Example

Say (X,Y) is a rv with joint pdf f(x,y)=8xy, $0 \leq x < y \leq 1$. Find f_Y (y)

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_0^y 8xydx =$$

$$4y \int_0^y 2xdx = 4y \cdot x^2|_0^y = 4y^3, \quad 0 \leq y \leq 1$$

Note that f_Y (y) is s proper pdf: $f_Y(y) \geq 0$ for all y and

$$\int_{-\infty}^{\infty} f_Y(y)dy = \int_0^1 4y^3 dy = y^4|_0^1 = 1$$

1.6.1.4 Example

Say (X,Y,Z) is a continuous rv with f(x,y,z) = 2(x+y)z if $0 < x, y, z < 1$ and 0 otherwise.

$$f_{XY}(x,y) = \int_{-\infty}^{\infty} f(x,y,z)dz = \int_0^1 2(x+y)zdz = (x+y)z^2|_0^1 = x+y, 0 < x,y < 1$$

$$f_{XZ}(x,z) = \int_{-\infty}^{\infty} f(x,y,z)dy = \int_0^1 2(x+y)zdz = 2xzy + zy^2|_0^1 = 2xz + z, 0 < x,z < 1$$

$$f_{YZ}(y,z) = (2y+1)z, 0 < y,z < 1$$

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)dy = \int_0^1 x + ydy = xy + \tfrac{1}{2}y^2|_0^1 = x + \tfrac{1}{2}, 0 < x < 1$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)dx = \int_0^1 x + ydx = \tfrac{1}{2}x^2 + xy|_0^1 = y + \tfrac{1}{2}, 0 < y < 1$$

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XZ}(x,z)dx = \int_0^1 (2x+1)zdx = (x^2+x)z|_0^1 = 2z, 0 < z < 1$$

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_0^1 x + 1/2dx = \tfrac{x^2}{2} + \tfrac{x}{2}|_0^1 = 1$$

$$\int_{-\infty}^{\infty} f_Z(z)dz = \int_0^1 2zdz = z^2|_0^1 = 1$$

1.6.2 Conditional Random Variables

Definition

let (X,Y) be a r.v. with joint pdf f(x,y) and marginal f_Y . For any y such that f_Y (y)>0 the *conditional pdf* of X|Y=y is defined by

$$f_{X|Y=y}(x|y) = \frac{f(x,y)}{f_Y(y)}$$

Note that a conditional pdf requires a specification for a value of the random variable on which we condition, something like f_X|Y=y . An expression like f_X|Y is not defined!

Note that is is exactly the same as the definition for conditional probabilities of events. For example if (X,Y) is a discrete rv, then

$$P(X = k|Y = j) = \frac{P(X = k, Y = j)}{P(Y = j)} = \frac{f(k,j)}{f_Y(j)} = f_{X|Y=j}(k|j)$$

1.6.2.1 Example

Say X is the sum and Y is the absolute difference of two dice. Find $f_{X|Y=5}$ (7|5) and $f_{Y|X=3}$ (7|3)

	0	1	2	3	4	5	X
2	1	0	0	0	0	0	1
3	0	2	0	0	0	0	2
4	1	0	2	0	0	0	3
5	0	2	0	2	0	0	4
6	1	0	2	0	2	0	5
7	0	2	0	2	0	2	6
8	1	0	2	0	2	0	5
9	0	2	0	2	0	0	4
10	1	0	2	0	0	0	3
11	0	2	0	0	0	0	2
12	1	0	0	0	0	0	1
Y	6	10	8	6	4	2	36

$$f_{X|Y=5}(7|5) = \frac{f(7,5)}{f_Y(5)} = \frac{2/36}{2/36} = 1$$

$$f_{Y|X=7}(3|7) = \frac{f(7,3)}{f_X(7)} = \frac{2/36}{6/36} = \frac{1}{3}$$

1.6.2.2 Example

f(x,y)=8xy, $0 \leq x < y \leq 1$. Find $f_{X|Y=y}(x|y)$

$f_{X|Y=y}(x|y) = f(x,y)/f_Y(y) = 8xy/4y^3 = 2x/y^2$,

for x, y with $\mathbf{0 \leq x \leq y}$.

Here y is a fixed number!

Again, note that a conditional pdf is a proper pdf:

$$\int_{-\infty}^{\infty} f_{X|Y=y}(x|y)dx = \int_0^y 2x/y^2 dx = x^2/y^2|_0^y = 1$$

1.6.2.3 Example

Say (X,Y,Z) is a continuous rv with f(x,y,z) = 2(x+y)z if $0 < x, y, z < 1$ and 0 otherwise. Then all the marginals are:

$$f_{XY|Z=z}(x,y|z) = \frac{f(x,y,z)}{f_Z(z)} = \frac{2(x+y)z}{2z} = x+y, 0 < x,y < 1$$

$$f_{XZ|Y=y}(x,z|y) = \frac{f(x,y,z)}{f_Y(y)} = \frac{2(x+y)z}{y+1/2}, 0 < x,z < 1$$

$$f_{YZ|X=x}(y,z|x) = \frac{f(x,y,z)}{f_X(x)} = \frac{2(x+y)z}{x+1/2}, 0 < y,z < 1$$

$$f_{X|Y=y,Z=z}(x|y,z) = \frac{f(x,y,z)}{f_{YZ}(y,z)} = \frac{2(x+y)z}{(2y+1)z} = \frac{2(x+y)}{2y+1}, 0 < x < 1$$

$$f_{X|Y=y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{x+y}{y+1/2}, 0 < x < 1$$

...

$$\int_{-\infty}^{\infty} f_{X|Y=y}(x|y) = \int_0^1 \frac{x+y}{y+1/2} dx = \frac{1}{y+1/2}\left(\frac{x^2}{2} + xy\right|_0^1 = 1$$

1.6.2.4 Example

say $f(x,y) = \alpha + 1$ if $0 < x < y^\alpha < 1$. Find the marginals and the conditional pdf's. Verify that they are proper pdf's.

$$f_Y(y) = \int_0^{y^\alpha} \alpha + 1 dx = (\alpha + 1)y^\alpha, \quad 0 < y < 1$$

$$\int_{-\infty}^{\infty} f_Y(y)dy = \int_0^1 (\alpha + 1)y^\alpha dy = y^{\alpha+1}|_0^1 = 1$$

$$f_{X|Y=y}(x,y) = \frac{f(x,y)}{f_Y(y)} = \frac{\alpha+1}{(\alpha+1)y^\alpha} = \frac{1}{y^\alpha}, \quad 0 < x < y^\alpha$$

$$X|Y = y \sim U[0, y^\alpha]$$

$$f_X(x) = \int_{x^{1/\alpha}}^1 \alpha + 1 dy = (\alpha + 1)(1 - x^{1/\alpha}), \; 0 < x < 1$$

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_0^1 (\alpha + 1)(1 - x^{1/\alpha})dx = (\alpha + 1)(x - \frac{1}{1+1/\alpha}x^{1+1/\alpha})|_0^1 =$$

$$(\alpha + 1)(1 - \frac{1}{1+1/\alpha}) = (\alpha + 1)(1 - \frac{\alpha}{\alpha+1}) = 1$$

$$f_{Y|X=x}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\alpha+1}{(\alpha+1)(1-x^{1/\alpha})} = \frac{1}{1-x^{1/\alpha}}, \quad x^{1/\alpha} < y < 1$$

$$Y|X = x \sim U[x^{1/\alpha}, 1]$$

1.6.3 Independence

Definition

Two r.v. X and Y are said to be independent iff

$$f_{X,Y}(x, y) = f_X(x)f_Y(y)$$

for all x,y

Notation: we will use the notation $X \perp Y$ if X and Y are independent.

1.6.3.1 Example

Say X is the sum and Y is the absolute difference of two dice. Previously we found

$f_{X,Y}$ (7,1) = 1/18

but

f_X (7)f_Y (1) = 1/6 x 10/36 = 5/108

so X and Y are not independent

Theorem

say f(x,y) is the joint pdf of a random vector (X,Y). Then X and Y are independent if there exist functions g and h such that

f(x,y)=g(x)h(y)

proof

the only difference between the definition and the theorem is that g and h need not be proper densities. But first of all we can assume that g and h are non-negative, otherwise just take |g| and |h|.

Moreover

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{-\infty}^{\infty} g(x)h(y)dy = g(x)\int_{-\infty}^{\infty} h(y)dy = cg(x)$$

$$1 = \int_{-\infty}^{\infty} f_X(x)dx = \int_{-\infty}^{\infty} cg(x)dx = c\int_{-\infty}^{\infty} g(x)dx$$

where $0 < c < \infty$, so g/c = f_X , and similarly h/d=f_Y

1.6.3.2 Example

say f(x,y)=exp(-x-y), x,y>0. Then

f(x,y) = exp(-x-y) = exp(-x)exp(-y) = g(x)h(y)

so X and Y are independent.

Mostly the concept of independence is used in reverse: we assume X and Y are independent (based on good reason!) and then make use of the formula:

1.6.3.3 Example

Say we use the computer to generate 10 independent exponential r.v's with rate λ. What is the probability density function of this random vector?

We have $f_{X_i}(x_i) = \lambda \exp(-\lambda x_i)$ for i=1,2,..,10 so

$$f_{(X_1,..,X_{10})}(x_1,..,x_10) =$$
$$\lambda \exp(-\lambda x_1) \times .. \times \lambda \exp(-\lambda x_10) =$$
$$\lambda^{10} \exp(-\lambda(x_1 + .. + x_{10}))$$

1.6.3.4 Example

Say (X,Y) is a discrete random vector with

```
##       1    2
## 1 1/10 1/10
## 2 1/10  1/2
## 3 1/10 1/10
```

	1	2
1	1/10	1/10
2	1/10	1/2
3	1/10	1/10

Find the conditional pdf of X|Y=y

$f_{X|Y=y}$ (x|y)=f(x,y)/f_Y (y)

$f_Y(y) = \sum_x f(x, y)$

f_Y (1) = f(1,1)+f(2,1)+f(3,1)=3/10
f_Y (2) = f(1,2)+f(2,2)+f(3,2)=7/10

so

$f_{X|Y=1}$ (1|1)=f(1,1)/f_Y (1) = (1/10)/(3/10) = 1/3
$f_{X|Y=1}$ (2|1)=f(2,1)/f_Y (1) = (1/10)/(3/10) = 1/3
$f_{X|Y=1}$ (3|1)=f(3,1)/f_Y (1) = (1/10)/(3/10) = 1/3

so

x	$f_{X\|Y=1}(x\|1)$
1	1/3
2	1/3
3	1/3

$f_{X|Y=2}$ (1|2)=f(1,2)/f_Y (2) = (1/10)/(7/10) = 1/7
$f_{X|Y=2}$ (2|2)=f(2,2)/f_Y (2) = (1/2)/(7/10) = 5/7
$f_{X|Y=2}$ (3|2)=f(3,2)/f_Y (2) = (1/10)/(7/10) = 1/7

so

x	$f_{X\|Y=2}(x\|2)$
1	1/7
2	5/7
3	1/7

1.6.3.5 Example

Let the continuous random vector (X,Y) have joint density $f(x,y) = e^{-y}, 0 < x < y < \infty$
Show that f is indeed a proper density

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_{0}^{\infty}\int_{0}^{y} e^{-y}dxdy = \int_{0}^{\infty} ye^{-y}dy =$$

$$y \cdot (-e^{-y}|_0^{\infty} - \int_0^{\infty} -e^{-y}dy = \int_0^{\infty} e^{-y}dy = (-e^{-y}|_0^{\infty} = 1$$

Find $f_{Y|X=x}$ (y|x)
$f_{Y|X=x}$ (y|x) = f(x,y)/f_X (x)

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_x^{\infty} e^{-y}dy = (-e^{-y}|_x^{\infty} = (0 - (-e^{-x}) = e^{-x}, x > 0$$

so

$$f_{Y|X=x}(y|x) = f(x,y)/f_X(x) = e^{-y}/e^{-x} = e^{x-y}, x < y < \infty$$

Show that $f_{Y|X=x}$ (y|x) is also a proper density

$$\int_{-\infty}^{\infty} f_{Y|X=x}(y|x)dy = \int_x^{\infty} e^{x-y}dy = -e^{x-y}|_x^{\infty} = 1$$

1.6.3.6 Example

We have a "device" which generates a random number Y according to an exponential distribution with rate λ. We don't know exactly what λ is, but we do know that $\lambda = x$ with probability 0.5^x where x=1,2,3,... Find the pdf of Y. Verify that your answer is a proper pdf.

We have a discrete r.v X with pdf

$f_X(x)=0.5^x$, x=1,2,..

and a conditional rv Y with pdf

$f_{Y|X=x}(y|x)=x\exp(-xy)$, y>0

We want $f_Y(y)$. It turns out that if we are dealing with a continuous rv. it is often better to first find the cdf $F_Y(y) = P(Y \leq y)$. Now first we have

$$f_{Y|X=x}(x|y) = \frac{f(x,y)}{f_X(x)} \Rightarrow f(x,y) = f_{Y|X=x}(y|x)f_X(x) = x\exp(-xy)0.5^x, y \geq 0, x = 1,2\ldots$$

so

$$F_Y(y) = P(Y \leq y) = \int_{-\infty}^{y} f_Y(t)dt = \int_{-\infty}^{y}\left[\sum_{x=1}^{\infty} f(x,t)\right]dt =$$

$$\sum_{x=1}^{\infty}\left[\int_{-\infty}^{y} f(x,t)dt\right] = \sum_{x=1}^{\infty}\left[\int_{0}^{y} x\exp(-xt)0.5^x dt\right] =$$

$$\sum_{x=1}^{\infty}\left(\frac{1}{2}\right)^x\left[\int_{0}^{y} x\exp(-xt)dt\right] = \sum_{x=1}^{\infty}\left(\frac{1}{2}\right)^x(-\exp(-xt)|_0^y =$$

$$\sum_{x=1}^{\infty}\left(\frac{1}{2}\right)^x(1-e^{-xy}) = \sum_{x=1}^{\infty}\left(\frac{1}{2}\right)^x - \sum_{x=1}^{\infty}\left(\frac{1}{2e^y}\right)^x =$$

$$\left(\sum_{x=0}^{\infty}\left(\frac{1}{2}\right)^x - 1\right) - \left(\sum_{x=0}^{\infty}\left(\frac{1}{2e^y}\right)^x - 1\right) = \frac{1}{1-1/2} - \frac{1}{1-1/(2e^y)} =$$

$$2 - \frac{2e^y}{2e^y-1}, y > 0$$

a little bit of care: the geometric series $\sum q^k$ only converges if |q|<1. Here y>0, so e^y>1 so $1/2e^y < 0.5 < 1$, we are save.

Now

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\left(2 - \frac{2e^y}{2e^y-1}\right) =$$
$$-\frac{2e^y(2e^y-1)-2e^y 2e^y}{(2e^y-1)^2} = \frac{2e^y}{(2e^y-1)^2}, \; y > 0$$
$$f_Y(y) = \frac{2e^y}{(2e^y-1)^2} \geq 0$$
$$\int_{-\infty}^{\infty} f_Y(y)dy = F_Y(y)\Big|_0^{\infty} = \lim_{t\to\infty} F_Y(y)\Big|_0^t$$
$$\lim_{t\to\infty}\left[2 - \frac{2e^t}{2e^t-1} - \left(2 - \frac{2e^0}{2e^0-1}\right)\right] =$$
$$\lim_{t\to\infty}\left[2 - \frac{2e^t}{2e^t-1}\right] = 1$$

This type of model is called a *hierarchical model*, with one rv defined conditional on another. This way of describing a model is very useful in real live.

1.6.4 Law of Total Probability

We have previously seen the law of total probability for events. There are corresponding versions for random variables:

- **Discrete-Discrete**

Say X and Y are discrete rv's with pdf's f_X and f_Y , respectively. Let B={X=x} and A_y ={Y=y}.

Then $\{A_y, y \in S\}$ forms a partition and we have

$$f_X(x) = P(X = x) = P(B) =$$
$$\sum_y P(B|A_y)P(A_y) =$$
$$\sum_y f_{X|Y=y}(x|y)f_Y(y)$$

- **Discrete-Continuous**

Say X is a discrete rv with pdf f_X and Y is a continuous rv with pdf f_Y.

Here we need to be careful: for a discrete rv f_X (x)=P(X=x) makes sense, but for a continuous one we have

$$P(Y = y) = \lim_{h\to 0} P(y \le Y \le y + h) =$$
$$\lim_{h\to 0} \int_y^{y+h} f_Y(t)dt =$$
$$\lim_{h\to 0}(F_Y(y + h) - F_Y(y)) = 0$$

for all y!

first we condition on the discrete rv.: Now the event B={Y=y} does not work because P(B)=0 for all y. Let's instead consider the event $B = \{Y \le y\}$:

$$F_Y(y) = P(Y \le y) = \sum_x P(Y \le y|X = x)P(X = x) = \sum_x \int_{-\infty}^{y} f_{Y|X=x}(y|x)f_X(x)\,dx$$

so

$$f_y(y) = \frac{d}{dy}F_Y(y) = \frac{d}{dy}\left[\sum_x \int_{-\infty}^{y} f_{Y|X=x}(y|x)f_X(x)\ dx\ \right] =$$
$$\sum_x \frac{d}{dy}\left[\int_{-\infty}^{y} f_{Y|X=x}(y|x)\,dx\right]f_X(x) = \sum_x f_{Y|X=x}(y|x)f_X(x)$$

For conditioning on the continuous rv we need to define a new discrete rv Y' with

$Y' = ih$ if $ih \le Y < (i + 1)h$

Then

$$f_{Y'}(ih) = P(Y' = ih) =$$
$$P(ih \leq Y < (i+1)h) =$$
$$\int_{ih}^{(i+1)h} f_Y(y)dy \simeq f_Y(ih)((i+1)h - ih)$$

$$f_{X|Y'=ih}(x|ih) = \frac{f(x,ih)}{f_{Y'}(ih)} \rightarrow \frac{f(x,y)}{f_Y(y)} = f_{X|Y=y}(x|y)$$

$$f_X(x) = \sum_{i=-\infty}^{\infty} f_{X|Y'=ih}(x|ih) f_{Y'}(ih) =$$
$$\sum_{i=-\infty}^{\infty} f_{X|Y'=ih}(x|ih) f_Y(ih)((i+1)h - ih)$$
$$\rightarrow \int_{-\infty}^{\infty} f_{X|Y=y}(x|y) f_Y(y)dy$$

because this is a Riemann sum, so it converges to the corresponding integral.

- **Continuous-Continuous**

Actually, same as above, the same proof works for this case as well!

1.6.4.1 Example

back to the example above with the "device". Now we have the following solution:

$$f_Y(y) = \sum_{x=1}^{\infty} f_{Y|X=x}(y|x) f_X(x) =$$

$$\sum_{x=1}^{\infty} x e^{-xy} \frac{1}{2^x} = \sum_{x=1}^{\infty} x \left(\frac{1}{2e^y}\right)^x$$

$$\text{Let } p = \frac{1}{2e^y}$$

$$\sum_{n=1}^{\infty} np^n = \sum_{n=1}^{\infty} \left(\sum_{k=1}^{n} 1\right) p^n = \sum_{n=1}^{\infty} \sum_{k=1}^{n} p^n =$$

$$\sum_{k=1}^{\infty} \sum_{n=k}^{\infty} p^n = \sum_{k=1}^{\infty} \left(\sum_{n=0}^{\infty} p^n - \sum_{n=0}^{k-1} p^n\right) =$$

$$\sum_{k=1}^{\infty} \left(\frac{1}{1-p} - \frac{1-p^k}{1-p}\right) = \frac{1}{1-p} \sum_{k=1}^{\infty} p^k =$$

$$\frac{1}{1-p} \left(\sum_{k=0}^{\infty} p^k - 1\right) = \frac{1}{1-p} \left(\frac{1}{1-p} - 1\right) = \frac{p}{(1-p)^2} =$$

$$\frac{\frac{1}{2e^y}}{(1-\frac{1}{2e^y})^2} = \frac{2e^y}{(2e^y-1)^2}, y > 0$$

1.6.4.2 Example

again the example above with the "device", but now the rate X has a uniform distribution on [0,1], that is f_X (x)=1 if $0 < x < 1$. Then:

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

$$\int_0^1 x e^{-xy} \cdot 1 dx = (x \cdot -\frac{1}{y} e^{-xy}|_0^1 - \int_0^1 -\frac{1}{y} e^{-xy} dx =$$

$$-\frac{e^{-y}}{y} + \frac{1}{y} \int_0^1 e^{-xy} dx = -\frac{e^{-y}}{y} + \frac{1}{y^2} (-e^{-xy}|_0^1 =$$

$$-\frac{e^{-y}}{y} + \frac{1-e^{-y}}{y^2} = \frac{1}{y^2}(1 - e^{-y} - ye^{-y}), y > 0$$

1.7 Expectation and Correlation

1.7.1 Expectation and Variance

Definition

The **expectation** (or **expected value**) of a function g of a random variable X is defined by

$$Eg(X) = \begin{cases} \sum_k g(k)f(k) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x)f(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

if $E[|g(X)|] < \infty$.

Let $\mu = E[X]$, then μ is called the *mean* of X.

$\sigma^2 = Var(X) = E[(X-\mu)^2]$ is called the *variance* of X. The square root of the variance σ is called the *standard deviation*.

1.7.1.1 Example

Say X is the sum of two dice. Find E[X], E[X^2] and E[1/X].

we have

```
x <- 2:12
y <- c(1:5, 6, 5:1)
df <- data.frame(x=x, y=paste0(y, "/36"))
colnames(df) <- c("x", "P(X=x)")
kable.nice(df)
```

	x	P(X=x)
1	2	1/36
2	3	2/36
3	4	3/36
4	5	4/36
5	6	5/36
6	7	6/36
7	8	5/36
8	9	4/36
9	10	3/36
10	11	2/36
11	12	1/36

so $E[X] = 2\times 1/36+3\times 2/36+4\times 3/36+ 5\times 4/36+6\times 5/36+7\times 6/36+8\times 5/36+9\times 4/36+ 10\times 3/36+11\times 2/36+12\times 3/36 = 7$

$E[X^2] = 2^2\times 1/36+3^2\times 2/36+4^2\times 3/36+ 5^2\times 4/36+6^2\times 5/36+7^2\times 6/36+8^2\times 5/36+ 9^2\times 4/36+10^2\times 3/36+11^2\times 2/36+12^2\times 3/36 = 54.83$

$E[1/X] = 1/2\times 1/36+1/3\times 2/36+1/4\times 3/36+ 1/5\times 4/36+1/6\times 5/36+1/7\times 6/36+1/8\times 5/36+ 1/9\times 4/36+1/10\times 3/36+1/11\times 2/36+1/12\times 3/36 = 0.172$

1.7.1.2 Example

we roll fair die until the first time we get a six. What is the expected number of rolls?

We saw that f(x) = 1/6*(5/6)$^{x-1}$ if $x \in \{1, 2, ..\}$. Here we just have g(x)=x, so

$$EX = \sum_k g(k)f(k) = \sum_{k=1}^{\infty} k\frac{1}{6}(\frac{5}{6})^{k-1} = \frac{1}{6}\sum_{k=1}^{\infty} k(\frac{5}{6})^{k-1}$$

How do we compute this sum? Here is a "standard" trick:

$$\sum_{k=1}^{\infty} kt^{k-1} = \sum_{k=1}^{\infty}(\frac{d}{dt}t^k) = \frac{d}{dt}\sum_{k=1}^{\infty} t^k =$$

$$\frac{d}{dt}\left(\sum_{k=0}^{\infty} t^k - 1\right) = \frac{d}{dt}(\frac{1}{1-t} - 1) = \frac{1}{(1-t)^2}$$

and so we find

$$EX = \frac{1}{6}\frac{1}{(1-\frac{5}{6})^2} = 6$$

This is a special example of a *geometric rv*, that is a discrete rv X with pdf f(x)=p(1-p)$^{x-1}$, x=1,2,..

Note that if we replace 1/6 above with p, we can show that

$$E[X] = \sum_{x=1}^{\infty} xp(1-p)^{x-1} = \frac{1}{p}$$

we write $X \sim$ Geom(p)

Theorem

let X,Y be some random variables, and let a,b be some real numbers. Then

$$E[aX+b] = aEX+b$$

$$E[X+Y] = EX+EY$$

$$Var(aX+b) = a^2 VarX$$

$$VarX = EX^2 - (EX)^2$$

proof (all for X discrete)

$$E[aX+b] = \sum_x (ax+b) f_X(x) = a\sum_x x f_X(x) + b\sum_x f_X(x) = aE[X]+b$$

$$E[X+Y] = \sum_{x,y}(x+y)f(x,y) = \sum_{x,y} xf(x,y) + \sum_{x,y} yf(x,y) =$$
$$\sum_x x\sum_y f(x,y) + \sum_y y\sum_x f(x,y) = \sum_x x f_X(x) + \sum_y y f_Y(y) = E[X]+E[Y]$$

$$Var(aX+b) = E[(aX+b-E[aX+b])^2 = E[(aX+b-aEX-b)^2 =$$
$$E[a^2(X-EX)^2] = a^2 Var(X)$$

$$VarX = E[(X-\mu)^2] = E[X^2 - 2X\mu + \mu^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - E[X]^2$$

1.7.1.3 Example

Say X is the sum of two dice. What is Var(X)?

Var(X) = E[X^2]-(E[X])^2 = 54.83-7^2 = 5.83

1.7.1.4 Example

find the mean and the standard deviation of a uniform [A,B] r.v.

We will use a little trick for this: say $X \sim U[0,1]$, and let Y=(B-A)X+A, then for $A < y < B$

$$F_Y(y) = P(Y < y) = P(((B-A)X + A < y) =$$
$$P(X < (y-A)/(B-A)) = (y-A)/(B-A)$$

$$f_Y(y) = 1/(B-A) \text{ for } A < y < B$$

so $Y \sim U[A,B]$

$$EX^k = \frac{1^{k+1}-0^{k+1}}{(k+1)(1-0)} = \frac{1}{k+1}$$

$$\mu = EX = \frac{1}{2}$$

$$VarX = EX^2 - (EX)^2 = \frac{1}{2+1} - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

$$EY = E[(B-A)X + A] = (B-A)EX + A = (B-A)\frac{1}{2} + A = \frac{A+B}{2}$$

$$VarY = Var[(B-A)X + A] = (B-A)^2 VarX = \frac{(B-A)^2}{12}$$

$$\sigma = \frac{B-A}{2\sqrt{3}}$$

Definition

$\mu_k = E[X^k]$ is called the k$^{\text{th}}$ *moment* of X.

$\kappa_k = E[(X-\mu)^k]$ is called the k$^{\text{th}}$ *central moment* of X.

$\gamma_1 = \kappa_3/(\kappa_2)^(3/2)$ is called the *skewness* of X.

$\gamma_2 = \kappa_4/\kappa_2^2 (= \mu_4/\sigma^4 - 3)$ is called the *kurtosis* of X.

The kurtosis of a distribution measure its "peakness", that is how sharp its maximum is. A distribution with $\gamma_2 = 0$ is called *mesokurtic*. This is the case for example for a standard normal (see later), which is then a kind of baseline example. If $\gamma_2 < 0$ it is called *platykurtic* and has a broader peak and thinner tails. If $\gamma_2 > 0$ it is called *leptokurtic* meaning it has a sharper peak than the standard normal and heavier tails.

1.7.1.5 Example

say X has pdf

$$f(x) = \frac{1}{\sqrt{2\pi\tau}} \exp\left\{-\frac{x^2}{2\tau}\right\}$$

for $x \in \mathbb{R}$ and $\tau > 0$

Then $E[X^k]=0$ for all odd numbers k because then $x^k f(x)$ is an odd function. For even moments we find

$$f(x) = \frac{1}{\sqrt{2\pi\tau}} \exp\{-\frac{x^2}{2\tau}\}$$

$$E[X^{2k}] = \int_{-\infty}^{\infty} x^{2k} \frac{1}{\sqrt{2\pi\tau}} \exp\{-\frac{x^2}{2\tau}\}dx =$$

$$\tau^k \int_{-\infty}^{\infty} \left(\frac{x}{\sqrt{\tau}}\right)^{2k} \frac{1}{\sqrt{2\pi\tau}} \exp\{-\frac{1}{2}\left(\frac{x}{\sqrt{\tau}}\right)^2\}(\frac{1}{\sqrt{\tau}}dx) = (t = x/\sqrt{\tau})$$

$$\tau^k \int_{-\infty}^{\infty} x^{2k} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}dx =$$

$$\tau^k \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^{2k-1} \cdot x \exp\{-\frac{x^2}{2}\}dx =$$

$$\tau^k \frac{1}{\sqrt{2\pi}} \left[x^{2k-1}(-\exp\{-\frac{x^2}{2}\}|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} (2k-1)x^{2k-2} \cdot (-\exp\{-\frac{x^2}{2}\})dx \right] =$$

$$\tau^k \left[(2k-1) \int_{-\infty}^{\infty} x^{2k-2} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{x^2}{2}\}dx \right] = \tau^k (2k-1) E[X^{2k-2}]$$

$$E[X^2] = \tau(2 \cdot 1 - 1)E[X^{2-2}] = \tau$$

$$E[X^4] = \tau(2 \cdot 2 - 1)E[X^{2\cdot 2-2}] = 3\tau^2$$

$$\kappa_2 = E[(X-\mu)^2] = E[X^2] = \tau$$

$$\kappa_4 = E[(X-\mu)^4] = E[X^4] = 3\tau^2$$

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} - 3 = \frac{3\tau^2}{\tau^2} - 3 = 0$$

Theorem

Say X is a non-negative rv, that is $P(X \geq 0) = 1$. Then

$$E[X] = \begin{cases} \sum_{k=1}^{\infty} P(X \geq k) & \text{if X is discrete} \\ \int_0^{\infty} P(X > x)dx & \text{if X is continuous} \end{cases}$$

proof

$$E[X] = \sum_{k=1}^{\infty} kP(X = k) = \sum_{k=1}^{\infty} \left(\sum_{j=1}^{k} 1 \right) P(X = k) =$$

$$\sum_{k=1}^{\infty} \sum_{j=1}^{k} P(X = k) = \sum_{j=1}^{\infty} \sum_{k=j}^{\infty} P(X = k) =$$

$$\sum_{k=1}^{\infty} P(X \geq k)$$

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_0^{\infty} xf(x)dx = \int_0^{\infty} \left(\int_0^{x} 1dt \right) f(x)dx =$$

$$\int_0^{\infty} \int_0^{x} f(x)dtdx = \int_0^{\infty} \int_t^{\infty} f(x)dxdt = \int_0^{\infty} P(X > x)dx$$

1.7.1.6 Example

say $X \sim$ Geom(p), then

$$P(X > k) = 1 - P(X \le k) = 1 - \sum_{i=1}^{k} pq^{i-1} =$$
$$1 - p\sum_{j=0}^{k-1} q^j = 1 - p\frac{1-q^k}{1-q} = q^k$$

$$E[X] = \sum_{k=1}^{\infty} P(X \ge k) = \sum_{k=0}^{\infty} P(X > k) =$$
$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q} = \frac{1}{p}$$

1.7.1.7 Example

Say (X,Y) is a discrete rv with joint pdf f(x,y)=cp^x, x,y in {0,1,..}, $y \le x$, and $0 < p < 1$. Find c.

We already did that before by summing first over y and then over x. We can use the above for an even simpler proof:

$$1 = \sum_{x,y} f(x,y) = \sum_{x=0}^{\infty}\sum_{y=0}^{x} cp^x = c\sum_{x=0}^{\infty} p^x\left(\sum_{y=0}^{x} 1\right) =$$
$$c\sum_{x=0}^{\infty}(x+1)p^x = c\sum_{x=1}^{\infty} xp^{x-1} = \frac{c}{1-p}\sum_{x=1}^{\infty} x(1-p)p^{x-1} =$$
$$\frac{c}{1-p}E[G] = \frac{c}{(1-p)^2}$$

where G is a geometric rv with rate 1-p

1.7.1.8 Example

Say $X \sim$ U[A,B]. Find E[X^k]. For the case A=0, B=1 find the kurtosis.

$$EX^k = \int_{-\infty}^{\infty} x^k f(x)dx = \int_A^B x^k \frac{1}{B-A}dx = \frac{x^{k+1}}{(k+1)(B-A)}\Big|_A^B = \frac{B^{k+1}-A^{k+1}}{(k+1)(B-A)}$$

$$\mu = E[X] = \frac{1}{2}$$

$$E[X^2] = \frac{1}{3} \quad E[X^3] = \frac{1}{4} \quad E[X^4] = \frac{1}{5}$$

$$\kappa_2 = E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] =$$

$$\frac{1}{3} - 2\frac{1}{2}\frac{1}{2} + \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

$$\kappa_4 = E[(X-\mu)^4] =$$

$$E[X^4 - 4X^3\mu + 6X^2\mu^2 - 4X\mu^3 + \mu^4] =$$

$$\frac{1}{5} - 4\frac{1}{4}\frac{1}{2} + 6\frac{1}{3}\left(\frac{1}{2}\right)^2 - 4\frac{1}{2}\left(\frac{1}{2}\right)^3 + \left(\frac{1}{2}\right)^4 =$$

$$\frac{1}{5} - \frac{1}{2} + \frac{1}{2} - \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 = \frac{1}{5} - \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^4 =$$

$$\frac{1}{16 \cdot 5}[16 - 4 \cdot 5 + 5] = \frac{1}{80}$$

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} - 3 = \frac{1/80}{(1/12)^2} - 3 = -1.2$$

so a U[0,1] is platykurtic

1.7.1.9 Example

Say x is a rv with f(x)=c/(1+x^2). (X is called a *Cauchy* random variable). Find c and show that EX does not exist.

$$\int_{-\infty}^{\infty} f(x)dx = \int_{-\infty}^{\infty} \frac{c}{1+x^2}dx = c\lim_{y\to\infty}\int_{-y}^{y} \frac{1}{1+x^2}dx =$$

$$c\lim_{y\to\infty}(a\tan x|_{-y}^{y} = c\lim_{y\to\infty}(a\tan y - a\tan(y)) =$$

$$c(\tfrac{\pi}{2} - (-\tfrac{\pi}{2})) = c\pi = 1$$

so $c = \frac{1}{\pi}$

$$E|X| = \int_{-\infty}^{\infty} |x|f(x)dx = \frac{2}{\pi}\lim_{y\to\infty}\int_{0}^{y} \frac{x}{1+x^2}dx =$$

$$\frac{1}{\pi}\lim_{y\to\infty}\log(1+y^2) = \infty$$

1.7.1.10 Example

Find the mean and the standard deviation of an exponential rv with rate λ.

$$\mu = EX = \int_{-\infty}^{\infty} xf(x)dx = \int_{0}^{\infty} x\lambda e^{-\lambda x}dx =$$

$$(x \cdot (-e^{-\lambda x})|_0^{\infty} - \int_0^{\infty} -e^{-\lambda x}dx = -\frac{1}{\lambda}e^{-\lambda x}|_0^{\infty} = \frac{1}{\lambda}$$

$$EX^2 = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{0}^{\infty} x^2\lambda e^{-\lambda x}dx =$$

$$(x^2 \cdot (-e^{-\lambda x})|_0^{\infty} - \int_0^{\infty} -2xe^{-\lambda x}dx = 2\int_0^{\infty} xe^{-\lambda x}dx =$$

$$\frac{2}{\lambda}\int_0^{\infty} x\lambda e^{-\lambda x}dx = \frac{2}{\lambda}\mu = \frac{2}{\lambda^2}$$

and so

$$\sigma = \sqrt{V(X)} = \sqrt{EX^2 - \mu^2} = \sqrt{\frac{2}{\lambda^2} - \frac{1}{\lambda^2}} = \frac{1}{\lambda}$$

1.7.1.11 Example

Let X be a rv with pdf f(x)=(a+1)x^a, $0 < x < 1, a > 0$. For what values of a is X mesokurtic, platykurtic or leptokurtic?

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

$$\sigma^4 = E[(X-\mu)^2]^2 \qquad \mu_4 = E[(X-\mu)^4]$$

$$E[X^k] = \int_{-\infty}^{\infty} x^k f(x)dx = \int_0^1 x^k (a+1)x^a dx =$$

$$(a+1)\int_0^1 x^{a+k}dx = \frac{a+1}{a+k+1} x^{a+k+1}\Big|_0^1 = \frac{a+1}{a+k+1}$$

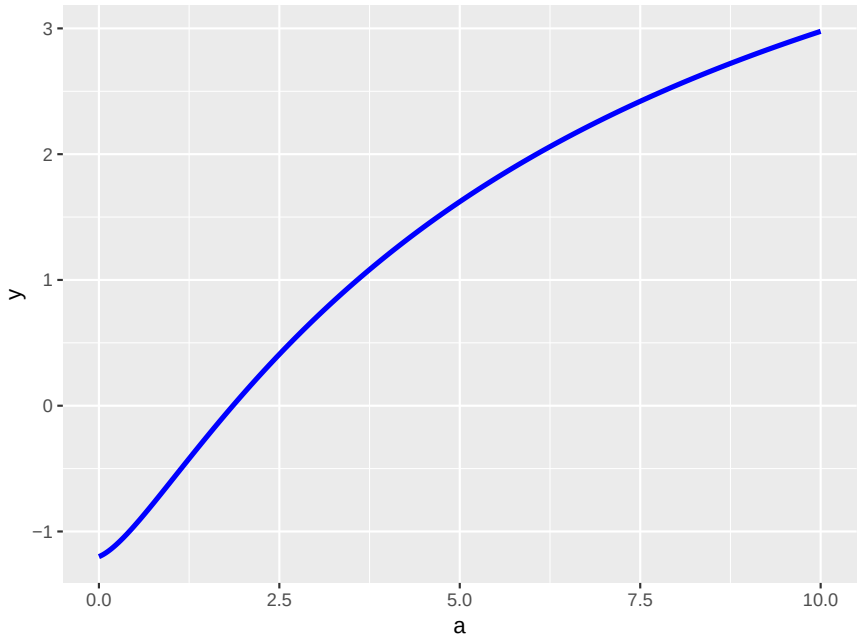
$$\mu = \frac{a+1}{a+2}$$

$$\sigma^2 = E[X^2] - \mu^2 = \frac{a+1}{a+3} - \left(\frac{a+1}{a+2}\right)^2$$

$$\mu_k = E[(X-\mu)^4] = E[X^4] - 4E[X^3]\mu + 6E[X^2]\mu^2 - 4E[X]\mu^3 + \mu^4 =$$
$$\frac{a+1}{a+5} - 4\frac{a+1}{a+4}\frac{a+1}{a+2} + 6\frac{a+1}{a+3}\left(\frac{a+1}{a+2}\right)^2 - 3\left(\frac{a+1}{a+2}\right)^4$$

This is a rather complicated function of a, so it is best to use a computer to do a graph:

```
f <- function(a) {
  muk <- function(a, k) (a+1)/(a+k+1)
  mu <- muk(a, 1)
  sig2 <- muk(a, 2) - mu^2
  mu4 <-   muk(a, 4) -
         4*muk(a, 3)*mu +
         6*muk(a, 2)*mu^2 -
         4*muk(a, 1)*mu^3 + mu^4
  mu4/sig2^2-3
}
a <- seq(0, 10, length=1000)
y <- f(a)
ggplot(data=data.frame(a=a, y=y), aes(a, y)) +
  geom_line(color="blue", size=1.2)
```




```
max(a[y<0])
```

```
## [1] 1.851852
```

therefore X is platykurtic for $a < 1.85$ and leptokurtic for all other a.

There is a way to "link" probabilities and expectations is via the indicator function I_A defined as

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

because with this we have for a (continuous) r.v. X with density f:

$$EI_A(X) = \int_{-\infty}^{\infty} I_A(x) f(x) dx = \int_A f(x) dx = P(X \in A)$$

Theorem

say we have a nonnegative rv X, that is $P(X \geq 0) = 1$. Then P(X=0)=1 iff E[X]=0.

proof

say P(X=0)=1, then X is a discrete rv with pdf f(0)=1 and so $E[X] = 0 \times 1 = 0$

say E[X]=0. Assume P(X=0)<1, therefore P(X>0) = 1-P(X=0) > 1-1 = 0, so there exists $\delta > 0$ and $\epsilon > 0$ such that $P(X > \delta) > \epsilon$. Then

if X is discrete

$$EX = \sum_x xP(X = x) \geq \sum_{x>\delta} xP(X = x) \geq$$

$$\sum_{x>\delta} \delta P(X = x) = \delta \sum_{x>\delta} P(X = x) = \delta P(X > \delta) \geq \delta\varepsilon > 0$$

if X is continuous

$$EX = \int_0^\infty xf(x)dx \geq \int_\delta^\infty xf(x)dx \geq \int_\delta^\infty \delta f(x)dx$$

$$= \delta \int_\delta^\infty f(x)dx = \delta P(X > \delta) \geq \delta\varepsilon > 0$$

in either case we have a contradiction with EX=0.

1.7.2 Expectations of Random Vectors

The definition of expectation easily generalizes to random vectors:

1.7.2.1 Example

say (X,Y) is a discrete random vector with joint pdf

```
df <- data.frame(x=c(0.1, 0, 0.1), y=c(0.1, 0.5, 0.2))
dimnames(df) <- list(0:2, 1:2)
kable.nice(df)
```

	1	2
0	0.1	0.1
1	0.0	0.5
2	0.1	0.2

Find E[XY]

$E[XY] = 0 \times 1 \times 0.1 + 0 \times 2 \times 0.1 + 1 \times 1 \times 0 + 1 \times 2 \times 0.5 + 2 \times 1 \times 0.1 + 2 \times 2 \times 0.2 = 2.0$

1.7.2.2 Example

Let (X,Y) be a discrete random vector with f(x,y) = $(1/2)^{x+y}$, $x \geq 1, y \geq 1$. Find E[XY^2]

$$E[XY^2] = \sum_{x=1}^{\infty}\sum_{y=1}^{\infty} xy^2\left(\frac{1}{2}\right)^{x+y} =$$
$$\sum_{x=1}^{\infty}\sum_{y=1}^{\infty} x\left(\frac{1}{2}\right)^{x}\cdot y^2\left(\frac{1}{2}\right)^{y} =$$
$$\sum_{x=1}^{\infty} x\left(\frac{1}{2}\right)^{x}\cdot\sum_{y=1}^{\infty} y^2\left(\frac{1}{2}\right)^{y}$$

First we have

$$\sum_{x=1}^{\infty} x\left(\frac{1}{2}\right)^{x} = \sum_{x=1}^{\infty} x\frac{1}{2}\left(1-\frac{1}{2}\right)^{x-1} = \frac{1}{1/2} = 2$$

because this is the mean of a geometric rv with p=1/2. Next

$$\sum_{x=1}^{\infty} x(x-1)\left(\frac{1}{2}\right)^{x} = \left(\frac{1}{2}\right)^{2}\sum_{x=0}^{\infty} x(x-1)\left(\frac{1}{2}\right)^{x-2} = \frac{1}{4}\sum_{x=0}^{\infty}\frac{d^2}{dt^2}t^x\Big|_{t=1/2} =$$
$$\frac{1}{4}\frac{d^2}{dt^2}\sum_{x=0}^{\infty} t^x\Big|_{t=1/2} = \frac{1}{4}\frac{d^2}{dt^2}\sum_{x=0}^{\infty} t^x\Big|_{t=1/2} = \frac{1}{4}\frac{d^2}{dt^2}\frac{1}{1-t}\Big|_{t=1/2} =$$
$$\frac{1}{4}\frac{2}{(1-t)^3}\Big|_{t=1/2} = \frac{1}{4}\frac{2}{(1-1/2)^3} = 4$$

but

$$\sum_{x=1}^{\infty} x(x-1)\left(\frac{1}{2}\right)^{x} = \sum_{x=1}^{\infty} x^2\left(\frac{1}{2}\right)^{x} - \sum_{x=1}^{\infty} x\left(\frac{1}{2}\right)^{x}$$

and so

$$\sum_{x=1}^{\infty} x^2\left(\frac{1}{2}\right)^{x} = \sum_{x=1}^{\infty} x(x-1)\left(\frac{1}{2}\right)^{x} + \sum_{x=1}^{\infty} x\left(\frac{1}{2}\right)^{x} = 4+2 = 6$$

finally

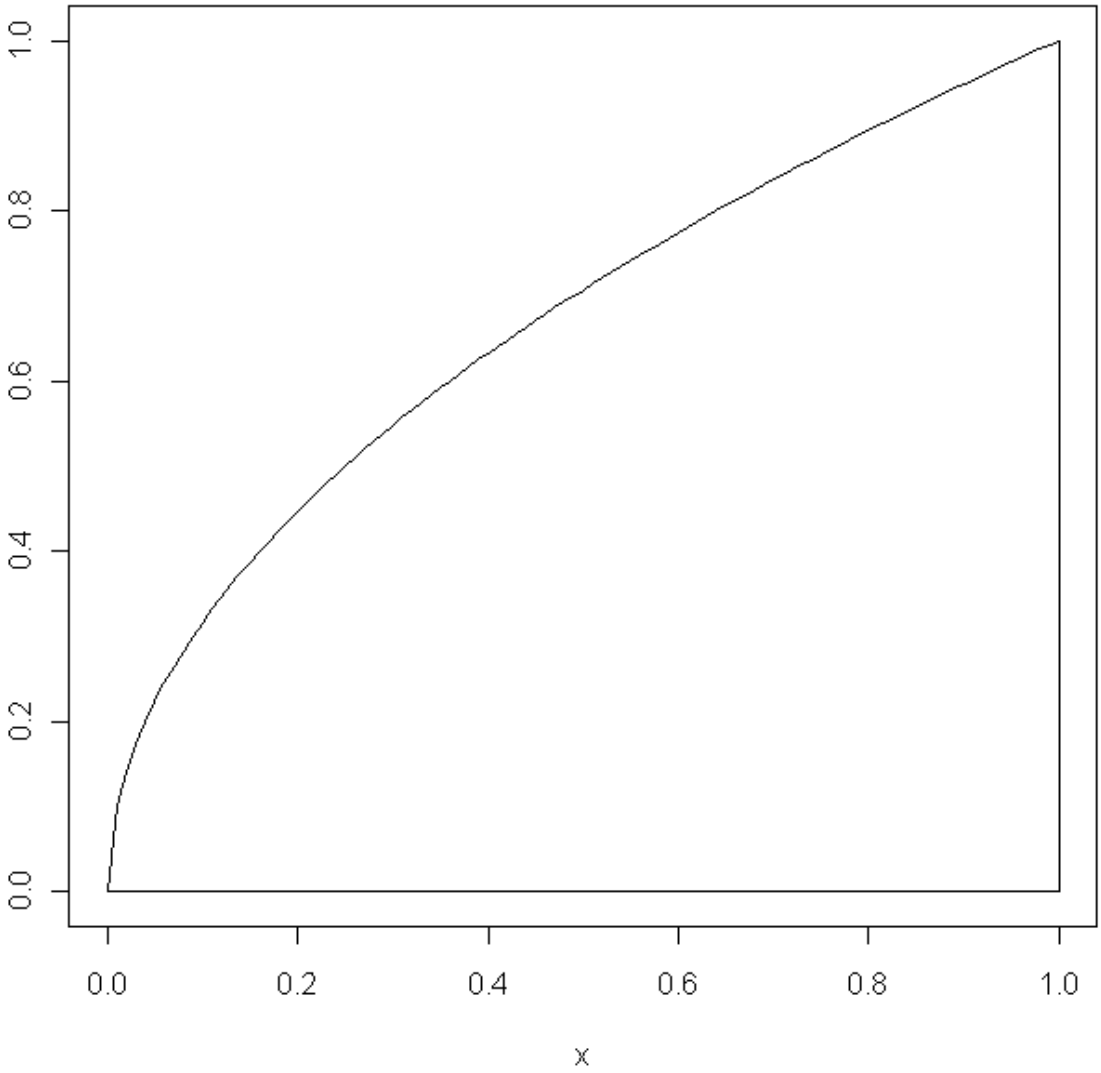
$$E[XY^2] = 2\cdot 6 = 12$$

Note that if we replace 1/2 with p we have just shown that E[X]=1/p and Var(X)=(1-p)/p^2 for $X \sim$ Geom(p).

1.7.2.3 Example

say (X,Y) is a continuous rv with f(x,y)=c if $0 < y < x^a < 1$ for some a>0. Find E[XY].

What we have here is a uniform rv on the area described by $0 < y < x^a < 1$, shown here for a=1/2:


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
x

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = \int_0^1\int_0^{x^a} cdydx =$$

$$c\int_0^1 x^a dx = \frac{c}{a+1}x^{a+1}\Big|_0^1 = \frac{c}{a+1} = 1$$

$$\text{so } c = a+1$$

$$EXY = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy =$$

$$\int_0^1\int_0^{x^a} xy(a+1)dydx =$$

$$(a+1)\int_0^1 x\left(\int_0^{x^a} ydy\right)dx =$$

$$(a+1)\int_0^1 x\left(\frac{1}{2}y^2\Big|_0^{x^a}\right)dx =$$

$$(a+1)\frac{1}{2}\int_0^1 x^{2a+1}dx =$$

$$\frac{a+1}{2(2a+2)}x^{2a+2}\Big|_0^1 = \frac{1}{4}$$

So

1.7.3 Covariance and Correlation

Definition

Say X and Y are two random variables. Then the *covariance* is defined by

$cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)]$

and the *correlation* of X and Y is defined by

$\rho_{XY} = cor(X,Y) = cov(X,Y)/(\sigma_X\sigma_Y)$

Note cov(X,X) = Var(X)

As with the variance we have a simpler formula for actual calculations:

cov(X,Y) = E(XY) - (EX)(EY)

Obviously, if cov(X,Y)=0, then

$\rho_{XY} = cor(X,Y) = cov(X,Y)/(\sigma_X \sigma_Y) = 0$

as well

1.7.3.1 Example

Let X and Y be the sum and absolute value of the difference of two rolls of a die. What is the covariance of X and Y?

So we have

$\mu_X = E[X] = 2 * 1/36 + 3 * 2/36 + ... + 12 * 1/36 = 7.0$

$\mu_Y = E[Y] = 0 * 6/36 + 1 * 12/36 + ... + 5 * 2/36 = 70/36$

$E[XY] = 0 * 2 * 1/36 + 1 * 2 * 0/36 + .2 * 2 * 0/36.. + 5 * 12 * 0/36 = 490/36$

and so

$cov(X,Y) = E[XY] - E[X]E[Y] = 490/36 - 7.0 * 70/36 = 0$

Note that in the example above we previously saw that X and Y are **not** independent, so we here have an example that a covariance of 0 does **not** imply independence! It does work the other way around, though:

Theorem

If X and Y are independent, then cov(X,Y) = 0

proof (in the case of X and Y continuous):

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dxdy =$$

$$\int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy = E[X]E[Y]$$

so

$$cov(X,Y) = E[XY] - E[X]E[Y] =$$

$$E[X]E[Y] - E[X]E[Y] = 0$$

We saw above that E[X+Y] = E[X] + E[Y]. How about Var(X+Y)?

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

$$E(X^2 + 2XY + Y^2) - [(EX)^2 + 2EXEY + (EY)^2] =$$

$$[EX^2 - (EX)^2] + [EY^2 - (EY)^2] + 2[EXY - EXEY] =$$

$$VarX + VarY + 2Cov(X,Y)$$

and if $X \perp Y$ we have Var(X+Y) = Var(X) + Var(Y)

1.7.3.2 Example

Consider again the example from before: we have continuous rv's X and Y with joint density f(x,y)=8xy, $0 \leq x < y \leq 1$. Find the covariance and the correlation of X and Y.

cov(X,Y)=E[XY]-E[X]E[Y].

We have seen before that f_Y (y)=$4y^3$, $0 < y < 1$, so

$E[Y] = \int_{-\infty}^{\infty} yf_Y(y)dy = \int_0^1 y4y^3dy = 4/5y^5|_0^1 = 4/5$

Now

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_x^1 8xydy = 4xy^2|_x^1 = 4(x-x^3), 0 < x < 1$$

so

$$E[X] = \int_0^1 x4(x-x^3)dx = \frac{4}{3}x^3 - \frac{4}{5}x^5|_0^1 = \frac{4}{3} - \frac{4}{5} = \frac{8}{15}$$

and

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dydx = \int_0^1\int_x^1 8x^2y^2dydx = \int_0^1 \frac{8}{3}x^2y^3|_x^1 dx =$$
$$\int_0^1 \frac{8}{3}(x^2-x^5)dx = \frac{8}{9}x^3 - \frac{4}{9}x^6|_0^1 = \frac{4}{9}$$

and so $cov(X,Y) = 4/9 - 8/15 \times 4/5 = 12/675$

Also

$$EX^2 = \int_0^1 x^2 4(x-x^3)dx = x^4 - \frac{2}{3}x^6|_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$$

$$\text{so } \sigma_X^2 = EX^2 - \mu_x^2 = \frac{1}{3} - \left(\frac{8}{15}\right)^2 = \frac{11}{225}$$

$$EY^2 = \int_0^1 y^2 4y^3 dy = \frac{2}{3}y^6|_0^1 = \frac{2}{3}$$

$$\text{so } \sigma_Y^2 = EY^2 - \mu_Y^2 = \frac{2}{3} - \left(\frac{4}{5}\right)^2 = \frac{2}{75}$$

$$cor(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{12}{675}}{\sqrt{\frac{11}{225}\frac{2}{75}}} = 0.492$$

1.7.3.3 Example

say (X,Y) is a discrete rv with joint pdf f given by

X \ Y	0	1
0	a	b
1	c	d

where a,b,c and d are numbers such that f is a pdf, that is a,b,c,d$\geq$ 0 and a+b+c+d=1. Note that this is the most general case of a discrete random vector where X and Y just take two values.

What can be said in this generality?

Now the marginals of X and Y are given by

f_X (0)=a+b, f_X (1)=c+d

f_Y (0)=a+c, f~Y‘ (1)=b+d

so

$EX = 0 \times (a+b) + 1 \times (c+d) = c+d$

$EY = 0 \times (a+c) + 1 \times (b+d) = b+d$

also

$EXY = 0 \times 0 \times a + 1 \times 0 \times b + 0 \times 1 \times c + 1 \times 1 \times d = d$

and so

cov(X,Y) =
d-(c+d)(b+d) =
d-cb-cd-bd-dˆ2 =
d-bc-(c+b)d-dˆ2 =
d-bc-(1-a-d)d-dˆ2 = d-bc-d+ad+d^{2-d}2 = ad-bc

so X and Y are uncorrelated iff ad-bc=0

Of course

$$ad - bc = \begin{vmatrix} a & b \\ c & d \end{vmatrix}$$

is the determinant of this matrix.

When are X and Y independent? For that we need f(x,y)=f_X (x)f_Y (y) for all x and y, so we need

a=(a+b)(a+c)
b=(a+b)(b+d)
c=(a+b)(b+d)
d=(c+d)(b+d)

but

a = (a+b)(a+c) =
aˆ2+(c+b)a+bc =
aˆ2+(1-a-d)a+bc =
a-ad+bc

or

ad-bc=0

Similarly we find that each of the other three equations holds iff ad-bc=0. So

$X \perp Y$ iff $ad - bc = 0$

and here we have a case where $X \perp Y$ iff cov(X,Y)=0.

Notice that if $X \perp Y$ then $rX + s \perp Y$ for any r,s with $r \neq 0$, so the above does not depend on the fact that X and Y take values 0 and 1, although the proof is much easier this way.

If you know cov(X,Y)=2.37, what does this tell you? Not much, really, except X and Y are not independent. But if I tell you cor(X,Y)=0.89, that tells you more:

Theorem

1. $|\rho_{XY}| \leq 1$

2. $\rho_{XY} = \pm 1$ iff there exist $a \neq 0$ and b such that P(X=aY+b)=1

proof

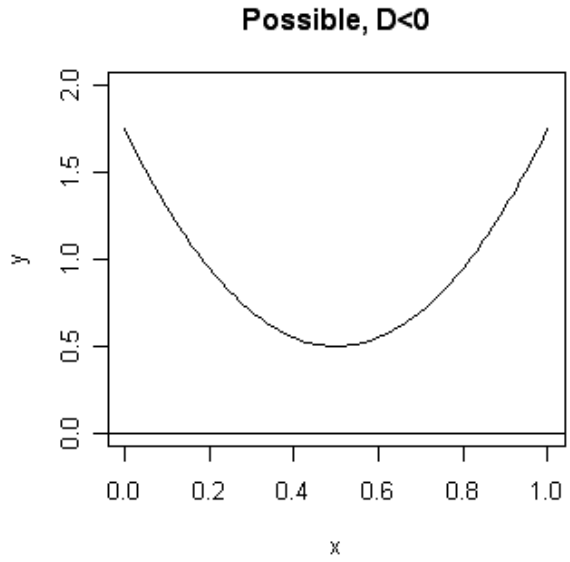
1. Consider the function

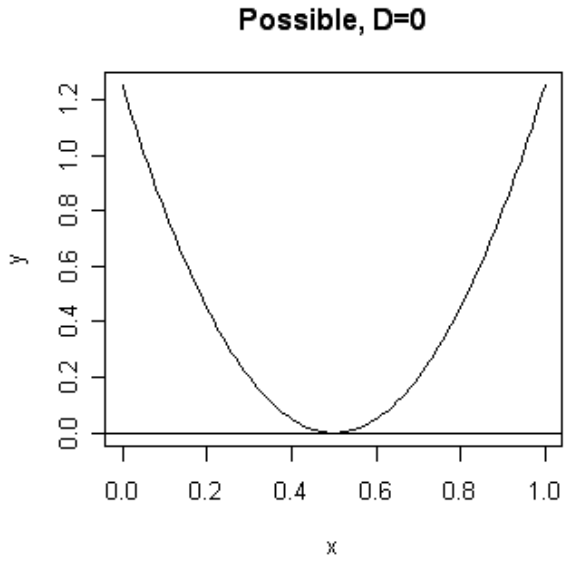
$h(t) = E[(X - \mu_X)t + (Y - \mu_Y)]^2$

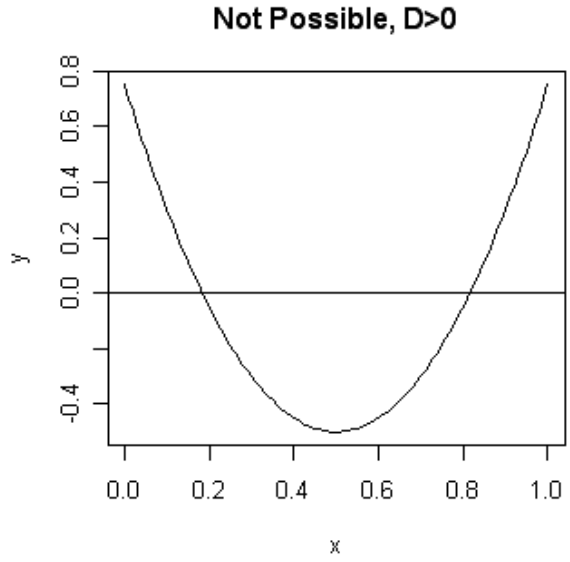
Now h(t) is the expectation of a non-negative function, so $h(t) \geq 0$ for all t. Also

$$h(t) = t^2 E(X-\mu_X)^2 + 2tE(X-\mu_X)(Y-\mu_Y) + E(Y-\mu_Y)^2 =$$
$$t^2\sigma_X^2 + 2tCov(X,Y) + \sigma_Y^2$$
$$D = b^2 - 4ac = (2Cov(X,Y))^2 - 4\sigma_X^2\sigma_Y^2 \leq 0$$
$$\Longrightarrow -\sigma_X\sigma_Y \leq Cov(X,Y) \leq_X \sigma_Y$$
$$\Longrightarrow -1 \leq \rho_{XY} \leq 1$$

because the quadratic function $h(t) \geq 0$, so it has at most one real root and so the discriminant has to be less or equal to 0.







2. Continuing with the argument above we see that $|\rho_{XY}| = 1$ iff D=0, that is if h(t) has a single root. But

$[(X - \mu_X)t + (Y - \mu_Y)]^2 \geq 0$

for all t and we have

h(t)=0 iff $P([(X - \mu_X)t + (Y - \mu_Y)]^2 = 0) = 1$

This is the same as

$P((X - \mu_X))t + (Y - \mu_Y) = 0) = 1$

so P(X=aY+b)=1 with a=-t and b=$\mu_X t + \mu_Y$, where t is the single root of h(t).

This theorem is also a direct consequence of a very famous inequality in mathematics. To

state it in some generality we need the following

Definition

Let V be some vector space V. A mapping $<.,.>:V^2 \to \mathbb{R}$ is an *inner product* on V if for $x, y, z \in V$ and $a \in \mathbb{R}$

$$< x,y >= \overline{< y,x >}$$

$$< ax+z,y >= a < x,y > + < z,y >$$

$$< x,x > \geq 0$$

where the line denotes complex complement.

A vector space with an inner product is called an *inner product space*.

Often we also write <x,x>=||x||^2 and then ||x|| is called the *norm*.

1.7.3.4 Example

1. R^n with <x,y>=$\sum x_i y_i$
2. the space of continuous functions C with

$< f,g >= \int f(x)g(x)dx$

Note that in an inner product space we have a version of the Pythagorean theorem: if x and y are such that

<x,y>=0

they are said to be orthogonal, and then we have

||x+y||^2 =
<x+y,x+y> =
<x,x>+<x,y>+<y,x>+<y,y> =
||x||$^{2+||y||}$2

Theorem (Cauchy-Schwartz)

say x and y are any two vectors of an inner product space, then

$< x,y >^2 \leq ||x||||y||$

and "=" holds iff x=ay+b for some $a, b \in \mathbb{R}$.

The Cauchy-Schwartz inequality is one of the most important results in Mathematics. It has a great many consequences, for example the general formulation of the Heisenberg uncertainty principle in Quantum Mechanics is derived using the Cauchy–Schwarz inequality in the Hilbert space of quantum observables.

1.7.3.5 Example

Let X and Y be some rv, and define <X,Y>=E[XY]. Then <X,Y> is an inner product. Moreover if E[X]=μ and E[Y]=ν by Cauchy-Schwartz

$$
\begin{aligned}
&Cov(X,Y)^2 = (E[(X-\mu)(Y-\nu)])^2 = \\
&< X-\mu, Y-\nu >^2 \leq \\
&< X-\mu, X-\mu >< Y-\nu, Y-\nu >= \\
&E(X-\mu)^2 E(Y-\nu)^2 = Var(X)Var(Y) =
\end{aligned}
$$

and so

$$
|\rho_{XY}| = |Cor(X,Y)| = \left| \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \right| =
$$

$$
\frac{1}{\sqrt{Var(X)Var(Y)}} \sqrt{Cov(X,Y)^2} \leq
$$

$$
\frac{1}{\sqrt{Var(X)Var(Y)}} \sqrt{Var(X)Var(Y)} = 1
$$

and we have "=" iff P(aX+b=Y)=1

It is one of the fascinating features in Mathematics that a theorem is sometimes easier to prove in greater generality:

proof (Cauchy-Schwartz)

Let x and y be two vectors in an inner product space. If y=0 the inequality is true (and an equation), so assume $y \neq 0$. Let

$$z = x - \frac{\langle x,y\rangle}{\langle y,y\rangle} y$$

$$\langle z,y\rangle = \left\langle x - \frac{\langle x,y\rangle}{\langle y,y\rangle} y, y \right\rangle = \langle x,y\rangle - \frac{\langle x,y\rangle}{\langle y,y\rangle}\langle y,y\rangle = 0$$

so z and y are orthogonal, therefore by Pythagoras

$$\|x\|^2 = \| \frac{\langle x,y\rangle}{\langle y,y\rangle} y + z\|^2 = \| \frac{\langle x,y\rangle}{\langle y,y\rangle} y\|^2 + \|z\|^2 =$$

$$| \frac{\langle x,y\rangle}{\langle y,y\rangle}|^2 \cdot \|y\|^2 + \|z\|^2 \geq \frac{|\langle x,y\rangle|^2}{\|y\|^4}\|y\|^2 = \frac{|\langle x,y\rangle|^2}{\|y\|^2}$$

and "=" iff z=0, in which case x=ay

A little bit of care with covariance and correlation: they are designed to measure **linear** relationships. Consider the following:

1.7.3.6 Example

let $X \sim$ U[-1,1], and let $Y=X^2$. Then E[X]=0 and

$E[Y] = EX^2 =$
$Var(X)+(E[X])^2 =$
$Var(X) = (1-(-1))^2/12 = 4/12 = 1/3.$

Also

$E[XY] = E[X^3] = (1^4-(-1)^4)/4/(1-(-1)) = 0$

so $cov(X,Y)=0 - 0 \times 1/3 = 0.$

So here is a case of two uncorrelated rv's, but if we know X we know exactly what Y is! Correlation is only a sensible measure of *linear relationships, not any others.*

So as we said above, if you know cov(X,Y)=2.37, that does not tell you much. But if you know cor(X,Y)=0.89 and if there is a linear relationship between X and Y, we know that it is a strong positive one.

Theorem

The correlation is *scale-invariant*, that is if $a \neq 0$ and b are any numbers, then

cor(aX+b,Y)=cor(X,Y)

proof

$$cov(aX+b,Y) = E[(aX+b)Y] - E[aX+b]E[Y] =$$
$$E[aXY+bY] - (aE[X]+b)E[Y] =$$
$$aE[XY] + bE[Y] - aE[X]E[Y] - bE[Y] =$$
$$a(E[XY] - E[X]E[Y]) = a \cdot cov(X,Y)$$

also

$$\sigma^2_{aX+b} = E[aX+b]^2 - (E[aX+b])^2 = a^2\sigma^2_X$$

so

$$\rho_{(aX+b)Y} = \frac{cov(aX+b,Y)}{\sigma_{aX+b}\sigma_Y} = \frac{acov(X,Y)}{a\sigma_X\sigma_Y} = \rho_{XY}$$

so for example the correlation between the ocean temperature and the windspeed of a hurricane is the same whether the temperature is measured in Fahrenheit or Centigrade.

1.7.4 Conditional Expectation and Variance

Definition

Say X|Y=y is a conditional r.v. with pdf f. Then the conditional expectation of g(X)|Y=y is defined by

$$E[g(X)|Y=y] = \sum_x g(x) f_{X|Y=y}(x|y) \text{ if X is discrete}$$
$$E[g(X)|Y=y] = \int_{-\infty}^{\infty} g(x) f_{X|Y=y}(x|y) \text{ if X is continuous}$$

1.7.4.1 Example

Say (X,Y) is a discrete rv with joint pdf f(x,y)=(1-p)^{2p}x, x,y in {0,1,..}, $y \leq x$, and $0 < p < 1$. Find E[Y|X=x]

first we need $f_Y|X=x$ (y|x), and for that we need f_X (x):

$$f_X(x) = \sum_y f(x,y) = \sum_{y=0}^{x} (1-p)^2 p^x = (1-p)^2(x+1)p^x \quad x = 0,1,2,..$$

so

$f_Y|X=x\ (y|x) = f(x,y)/f_X\ (x) = (1\text{-}p)^{2p}x/((1\text{-}p)^{2(x+1)p}x) = 1/(x+1)$

so Y|X=x has a discrete uniform distribution on {0,1,..,x}.

Therefore

$$E[Y|X = x] = \sum_{y=0}^{x} y \frac{1}{x+1} = \frac{1}{x+1} \frac{x(x+1)}{2} = \frac{x}{2}$$

1.7.4.2 Example

Say X and Y have a joint density f(x,y)=8xy, $0 \le x < y \le 1$.

We previously found $f_Y\ (y) = 4y^3$, 0<y<1, and $f_X|Y=y\ (x|y) = 2x/y^2$, $0 \le x \le y$. So

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y=y}(x|y)dx = \int_0^y x \frac{2x}{y^2} dx = \frac{2}{3y^2} x^3 \Big|_0^y = \frac{2y^3}{3y^2} = \frac{2}{3}y$$

Throughout this calculation we treated y as a constant. Now, though, we can change our point of view and consider E[X|Y=y] = 2y/3 as a function of y:

g(y)=E[X|Y=y]=2y/3

What are the values of y? Well, they are the observations we might get from the rv. Y, so we can also write

g(Y)=E[X|Y=Y]=2Y/3

but Y is a rv, then so is 2Y/3, and we see that we can define a rv Z=g(Y)=E[X|Y]

Recall that the expression $f_X|Y$ does not make sense. Now we see that on the other hand the expression E[X|Y] makes perfectly good sense!

Let's continue this example and find the conditional variance of X|Y=y:

$$V[X|Y=y] = E[X^2|Y=y] - (E[X|Y=y])^2$$

$$E[X^2|Y=y] = \int_{-\infty}^{\infty} x^2 f_{X|Y=y}(x|y) =$$

$$\int_0^y x^2 \frac{2x}{y^2} dx = \frac{x^4}{2y^2}\Big|_0^y = \frac{y^2}{2}$$

$$V[X|Y=y] = \frac{y^2}{2} - \left(\frac{2}{3}y\right)^2 = \left(\frac{1}{2} - \frac{4}{9}\right)y^2 = \frac{1}{18}y^2$$

and again we can consider the conditional variance of X|Y:

Var(X|Y)=$Y^2 2/18$

1.7.4.3 Example

An urn contains 2 white and 3 black balls. We pick two balls from the urn. Let X be denote the number of white balls chosen. An additional ball is drawn from the remaining three. Let Y equal 1 if the ball is white and 0 otherwise.

For example

$f(0,0) = P(X=0, Y=0) = 3/5 * 2/4 * 1/3 = 1/10$

(choose black-black-black)

The complete pdf is given by:

	0	1	2
0	1/10	2/5	1/10
1	1/5	1/5	0

Now for the marginals we have, for example

f_X (0)=1/10+1/5=3/10

or in general:

x	0	1	2
P(X=x)	3/10	3/5	1/10

for Y we have

y	0	1
P(Y=y)	3/5	2/5

The conditional density of X|Y=0 is

x	0	1	2
P(X=x|Y=0)	1/6	2/3	1/6

and so

$E[X|Y = 0] = 0 \times 1/6 + 1 \times 2/3 + 2$ů$1/6 = 1.0$

The conditional distribution of X|Y=1 is

x	0	1	2
P(X=x|Y=1)	1/2	1/2	0

and so

$E[X|Y = 1] = 0 \times 1/2 + 1 \times 1/2 + 2$ů$0 = 1/2$

Finally the conditional r.v. Z = E[X|Y] has pmf

z	1	1/2
P(Z=z)	3/5	2/5

with this we can find

$E[Z] = E[E[X|Y]] = 1 \times 3/5 + 1/2 \times 2/5 = 4/5$

Theorem say X and Y are random variables. Then

E[X] = E{E[X|Y]}

and

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

(There is a simple explanation for this seemingly complicated formula!)

proof (for continuous X and Y)

$$E\{E[X|Y]\} = \int_{-\infty}^{\infty} E[X|Y = y] f_Y(y) dy =$$
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{X|Y=y}(x|y) dx f_Y(y) dy =$$
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x \frac{f(x,y)}{f_Y(y)} f_Y(y) dx dy =$$
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x,y) dx dy =$$
$$\int_{-\infty}^{\infty} x \left(\int_{-\infty}^{\infty} f(x,y) dy \right) dx =$$
$$\int_{-\infty}^{\infty} x f_X(x) dx = E[X]$$

$$Var\{E[X|Y]\} = E\{E[X|Y]^2\} - \{E[X|Y]\}^2 = E\{E[X|Y]^2\} - E[X]^2$$
$$E\{Var[X|Y]\} = E\{E[X^2|Y] - E[X|Y]^2\} = E\{E[X^2|Y]\} - E\{E[X|Y]^2\} =$$
$$E[X^2] - E\{E[X|Y]^2\}$$

$$E\{Var[X|Y]\} + Var\{E[X|Y]\} = E[X^2] - E\{E[X|Y]^2\} + E\{E[X|Y]^2\} - E[X]^2 =$$
$$E[X^2] - E[X]^2 = Var(X)$$

1.7.4.4 Example

above we found E[E[X|Y]] = 4/5. Now

$E[X] = 0 \times 3/10 + 1 \times 3/5 + 2 \times 1/10 = 4/5$

1.7.4.5 Example

let's say we have a continuous bivariate random vector with the joint pdf f(x,y) = c(x+2y) if $0 < x < 2, 0 < y < 1$, 0 otherwise.

Now

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)d(x,y) = \int_0^1\int_0^2 c(x+2y)dxdy =$$

$$c\int_0^1 \tfrac{1}{2}x^2 + 2yx|_0^2 dy = c\int_0^1 2+4ydy =$$

$$c(2y+2y^2|_0^1 = 4c = 1.....c = 1/4$$

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^1 \tfrac{1}{4}(x+2y)dy =$$

$$\tfrac{1}{4}(xy+y^2|_0^1 = \tfrac{x+1}{4}....0 < x < 2$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_0^2 \tfrac{1}{4}(x+2y)dx =$$

$$\tfrac{1}{4}(\tfrac{1}{2}x^2 + 2yx|_0^2 = \tfrac{1}{4}(2+4y) = \tfrac{1}{2}+y \quad 0 < y < 1$$

$$f_{Y|X=x}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{\frac{1}{4}(x+2y)}{\frac{x+1}{4}} = \frac{x+2y}{x+1} \quad 0 < y < 1$$

$$E[Y|X=x] = \int_{-\infty}^{\infty} y f_{Y|X=x}(y|x)dy = \int_0^1 y\frac{x+2y}{x+1}dy =$$

$$\tfrac{1}{x+1}(\tfrac{x}{2}y^2 + \tfrac{2}{3}y^3|_0^1 = \frac{x/2+2/3}{x+1}....0 < x < 2$$

$$Z = E[Y|X]$$

$$EZ = E(E[Y|X]) = EY = \int_{-\infty}^{\infty} y f_Y(y)dy = \int_0^1 y(\tfrac{1}{2}+y)dy =$$

$$\tfrac{1}{4}y^2 + \tfrac{1}{3}y^3|_0^1 = \tfrac{1}{4} + \tfrac{1}{3} = \tfrac{7}{12}$$

or

$$EZ = \int_{-\infty}^{\infty} E[Y|X=x]f_X(x)dx = \int_0^2 \frac{x/2+2/3}{x+1}\frac{x+1}{4}dx =$$

$$\tfrac{1}{4}\int_0^2 x/2 + 2/3dx = \tfrac{1}{4}(x^2/4 + 2x/3|_0^2 =$$

$$\tfrac{1}{4}(1+\tfrac{4}{3}) = 7/12$$

1.7.4.6 Example

say X has a density f_X (x) = (a+1)x^a, $0 < x < 1, a > 1$. $Y|X = x \sim$ Exp(x). Find E[Y] and Var(Y).

To find E[Y] we first need the density of Y:

$$f(x,y) = f_X(x)f_{Y|X=x}(y|x) = (a+1)x^a \cdot xe^{-xy} = (a+1)x^{a+1}e^{-xy}$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx = \int_0^1 (a+1)x^{a+1}e^{-xy}dx$$

and this integral can not be found explicitely, so this won't work.

But

$$Y|X = x \sim Exp(x) \text{ so } E[Y|X = x] = \frac{1}{x} \text{ and } Var(Y|X = x) = \frac{1}{x^2}$$

$$E[Y] = E\{E[Y|X]\} = E[\frac{1}{X}] = \int_0^1 \frac{1}{x}(a+1)x^a dx = \int_0^1 (a+1)x^{a-1}dx = \frac{a+1}{a}x^a|_0^1 = \frac{a+1}{a}$$

$$Var(Y) = E\{Var(Y|X)\} + Var\{E[Y|X]\} = E\{\frac{1}{X^2}\} + Var(\frac{1}{X}) =$$
$$E\{\frac{1}{X^2}\} + E\{(\frac{1}{X})^2\} - E\{\frac{1}{X}\}^2 =$$
$$2\int_0^1 \frac{1}{x^2}(a+1)x^a dx - (\frac{a+1}{a})^2 =$$
$$2\int_0^1 (a+1)x^{a-2}dx - (\frac{a+1}{a})^2 =$$
$$2\frac{a+1}{a-1}x^{a-1}|_0^1 - (\frac{a+1}{a})^2 = 2\frac{a+1}{a-1} - (\frac{a+1}{a})^2$$

1.7.4.7 Example

Let's have another look at the example of the "device" which generates a random number Y according to an exponential distribution with rate λ where $\lambda = x$ with probability 0.5^x where x=1,2,3,... We previously found that f_Y (y) = $2e^y/(2e^y-1)\hat{}2$, $y \geq 0$.

Let's find E[X|Y]

Note E[Y|X] would be easy (=1/X because $Y \sim$ Exp(X)), E[Y] would be a simple calculus problem ($\int y2e^y/(2e^y-1)^2 dy$) and E[X] would be the easiest (=2 because $X \sim$ Geom(1/2)), just E[X|Y=y] needs a little work:

$$E[X|Y=y] = \sum_{x=1}^{\infty} x f_{X|Y=y}(x|y)$$

$$f_{X|Y=y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{f_{Y|X=x}(y|x) f_X(x)}{f_Y(y)} = \frac{xe^{-xy}0.5^x}{2e^y/(2e^y-1)^2} =$$

$$xe^{-xy}0.5^x(2e^y-1)^2/2e^y,\ y \geq 0,\ x = 1,2,\ldots$$

$$E[X|Y=y] = \sum_{x=1}^{\infty} x \cdot xe^{-xy}0.5^x(2e^y-1)^2/2e^y =$$

$$(2e^y-1)^2/2e^y \sum_{x=1}^{\infty} x^2 \left(\frac{1}{2e^y}\right)^x$$

now let $Z \sim Geom(p)$, then

$$\sum_{x=1}^{\infty} x^2 pq^{x-1} = E[Z^2] = V[Z] + E[Z]^2 =$$

$$\frac{q}{p^2} + \left(\frac{1}{p}\right)^2 = \frac{q+1}{p^2} = \frac{2-p}{p^2},\ \text{so with } p = 1 - \frac{1}{2e^y}$$

$$\sum_{x=1}^{\infty} x^2 \left(\frac{1}{2e^y}\right)^x = \frac{1/2e^y}{1-\frac{1}{2e^y}} \sum_{x=1}^{\infty} x^2 \left(1 - \frac{1}{2e^y}\right)\left(\frac{1}{2e^y}\right)^{x-1} =$$

$$\frac{1}{2e^y-1} \frac{2-(1-1/2e^y)}{(1-1/2e^y)^2} = \frac{1}{2e^y-1} \frac{2-\frac{2e^y-1}{2e^y}}{\left(\frac{2e^y-1}{2e^y}\right)^2} = \frac{1}{2e^y-1} \frac{\frac{4e^y-2e^y+1}{2e^y}}{\left(\frac{2e^y-1}{2e^y}\right)^2} =$$

$$\frac{1}{2e^y-1} \frac{2e^y(4e^y-2e^y+1)}{(2e^y-1)^2} = \frac{2e^y(4e^y-2e^y+1)}{(2e^y-1)^3}$$

and

$$E[X|Y=y] = (2e^y-1)^2/2e^y \frac{2e^y(4e^y-2e^y+1)}{(2e^y-1)^3} =$$

$$\frac{4e^y-2e^y+1}{2e^y-1} = \frac{4e^y}{2e^y-1} - 1$$

$$E[X|Y] = \frac{4e^Y}{2e^Y-1} - 1$$

we said above that E[X]=2.

Let's check the formula E[X]=E{E[X|Y]}:

$$E[E[X|Y]] = E\left[\frac{4e^Y}{2e^Y-1} - 1\right] = E\left[\frac{4e^Y}{2e^Y-1}\right] - 1$$

$$\int_0^\infty \frac{4e^y}{2e^y-1}\frac{2e^y}{(2e^y-1)^2}dy - 1 = \int_0^\infty \frac{8e^{2y}}{(2e^y-1)^3}dy - 1 =$$

$$\int_2^\infty \frac{2x}{(x-1)^3}dx - 1 = \quad \text{(with change of variables } x = 2e^y\text{)}$$

$$-\frac{2x}{2(x-1)^2}\Big|_2^\infty - \int_2^\infty -\frac{2}{2(x-1)^2}dx - 1 = 2 + \int_2^\infty \frac{1}{(x-1)^2}dx - 1 =$$

$$1 - \frac{1}{x-1}\Big|_2^\infty = 1 + 1 = 2$$

1.7.5 Moment Generating and Characteristic Functions

Definition

The moment generating function of a rv X is defined by

$$\Phi(t) = E[\exp(tX)]$$

The characteristic function of a rv X is defined by

$$\Phi(t) = E[\exp(itX)]$$

In general characteristic functions are much more useful in Probability Theory, but they require some knowledge of complex analysis, and so we will just consider moment generating functions.

1.7.5.1 Example

Let $X \sim \text{Exp}(\lambda)$, find the mgf Φ.

$$\phi(t) = E[e^{tX}] = \int_0^\infty e^{tx}\lambda e^{-\lambda x}dx =$$

$$\frac{\lambda}{\lambda-t}\int_0^\infty (\lambda - t)e^{-(\lambda-t)x}dx = \begin{cases} \frac{\lambda}{\lambda-t} & \text{if } t < \lambda \\ \infty & \text{if } t \geq \lambda \end{cases}$$

The name comes from the following theorem

Theorem

Say $\Phi(t)$ is the mgf of a rv X. Say there exists an $\epsilon > 0$ such that $|\Phi(t)| < \infty$ for all t in $(-\epsilon, \epsilon)$. Then

$\Phi^k(0) = E[X^k]$ for all k.

proof

say X is a discrete rv with pdf f(x) and X takes finitely many values. Then

$$\Phi^k(0) = \frac{d^k}{dt^k}\Phi(x)|_{t=0} =$$

$$\frac{d^k}{dt^k}E[e^{tX}]|_{t=0} =$$

$$\frac{d^k}{dt^k}\sum_x e^{tx}f(x)|_{t=0} =$$

$$\sum_x \frac{d^k}{dt^k}e^{tx}f(x)|_{t=0} =$$

$$\sum_x x^k e^{tx}f(x)|_{t=0} =$$

$$\sum_x x^k f(x) = E[X^k]$$

The extension to an infinite sample space and to a continuous rv requires some real analysis theorems.

1.7.5.2 Example

For the exponential rv we have

$$\phi(t) = \frac{\lambda}{\lambda - t}, t < \lambda$$

$$\phi'(0) = \left(\frac{d}{dt}\frac{\lambda}{\lambda - t}\right|_{t=0} = \frac{\lambda}{(\lambda - t)^2}\Big|_{t=0} = \frac{1}{\lambda} = EX$$

$$\phi''(0) = \left(\frac{d^2}{dt^2}\frac{\lambda}{\lambda - t}\right|_{t=0} = \left(\frac{d}{dt}\frac{\lambda}{(\lambda - t)^2}\right|_{t=0} =$$

$$\frac{2\lambda}{(\lambda - t)^3}\Big|_{t=0} = \frac{2}{\lambda^2} = EX^2$$

Warning nobody uses the moment generating function to generates moments! It has other uses:

Theorem

let $X_1, \ldots, X_n$ be a sequence of independent rv.s with mgf's Φ_i , and let $Z = \sum X_i$, then

$$\Phi_Z(t) = \prod \Phi_i(t)$$

if the distributions of the X_i are the same as well, then $\Phi_i = \Phi_X$ for all i and

$$\Phi_Z = (\Phi(t))^n$$

proof

$$\phi_Z(t) = E[\exp(tZ)] = E[\exp(t(X_1 + \ldots + X_n))] =$$

$$E[\exp(tX_1 + \ldots + tX_n)] = E[\exp(tX_1) \cdot \ldots \cdot \exp(tX_n)] =$$

$$E[\exp(tX_1)] \cdot \ldots \cdot E[\exp(tX_n)] = \phi_1(t) \cdot \ldots \cdot \phi_n(t)$$

here is a very deep theorem, without proof:

Theorem

let X and Y be rv.s with mgf's Φ_X and Φ_Y , respectively. If both mgf's are finite in an open neighborhood of 0 and if $\Phi_X(t) = \Phi_Y(t)$ for all t in this neighborhood, then $F_X(u) = F_Y(u)$ for all u.

In other words, the cdf determines the mgf and **vice versa**. This means that one way to show that two random variables have the same distribution is to show that they have the same mgf.

1.7.5.3 Example

show that the sum of two independent exponential rv. is **not** an exponential rv.

say $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\rho)$, then

$\Phi_X(t) = \lambda/(\lambda - t)$ and $\Phi_Y(t) = \rho/(\rho - t)$, so

$\Phi_{X+Y}(t) = \lambda/(\lambda - t)\rho/(\rho - t) \neq a/(a - t)$

for any a and all t.

1.7.5.4 Example

Consider the two pdfs given by

$$f_1(x) = \frac{1}{\sqrt{2\pi}\,x} e^{-(\log x)^2/2}, \qquad 0 \leq x < \infty$$

$$f_2(x) = f_1(x)[1 + \sin(2\pi \log x), \qquad 0 \leq x < \infty$$

(f_1 is called a log-normal distribution).

Now it turns out that if X_1 has density f_1 , then

$$EX_1^r = \int_0^\infty x^r \frac{1}{\sqrt{2\pi}\,x} e^{-(\log x)^2/2} dx =$$

$$\int_{-\infty}^{\infty} e^{rt} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}[t^2 - 2rt]) dt =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}[t^2 - 2rt + r^2 - r^2]) dt =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}[t - r]^2) \exp(r^2/2) dt = e^{r^2/2}$$

because

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}[t - r]^2) dt = 1 \text{ for all } r$$

where we use the change of variables t=log(x)

but

$$EX_2^r = \int_0^\infty x^r f_1(x)[1 + \sin(2\pi \log x)dx =$$

$$\int_0^\infty x^r f_1(x)dx + \int_0^\infty x^r f_1(x)\sin(2\pi \log x)dx =$$

$$e^{r^2/2} + \int_{-\infty}^\infty e^{r(t+r)} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(t+r)^2} \sin(2\pi(t+r))dx = e^{r^2/2}$$

because

$e^{r(t+r)} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(t+r)^2} \sin(2\pi(t+r))$ is an odd function.

(use change of variables t=log(x)-r)

and so here is an example that shows that the condition of the theorem above is also necessary, without it you can have two rv's with all their moments equal but different distributions.

1.8 Functions of a Random Variable - Transformations

1.8.0.1 Example

say $X \sim$ U[0,1] and $\lambda > 0$. What is the pdf of the random variable $Y = -\lambda \log(X)$?

Solution: we first find the cdf and then the pdf as follows:

$$F_Y(y) = P(Y \leq y) = P(-\lambda \log X \leq y) =$$

$$P(\log X \geq -\frac{y}{\lambda}) = P(X \geq e^{-\frac{y}{\lambda}}) =$$

$$1 - P(X \leq e^{-\frac{y}{\lambda}}) = 1 - e^{-\frac{y}{\lambda}}$$

$$f_y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{\lambda} e^{-\frac{y}{\lambda}}$$

if y>0. For $y < 0$ note that $P(-\log X < y) = 0$ because $0 < X < 1$, so $\log X < 0$, so $-\log X > 0$ always.

This is an example of a function (or *transformation*) of a random variable. These transformations play a major role in probability and statistics. We will see how to find their pdf's on a few examples.

1.8.0.2 Example

say X is the number of roles of a fair die until the first six. We have already seen that P(X=x) = $1/6*(5/6)^{x-1}$, x=1,2,.. Let Y be 1 if X is even, 0 otherwise. Find the pdf of Y.

Note: here both X and Y are discrete.

Solution:

$$P(Y=1) = P(X \in \{2,4,6,..\}) = \sum_{k=1}^{\infty} P(X=2k) =$$

$$\sum_{k=1}^{\infty} \frac{1}{6} \cdot \left(\frac{5}{6}\right)^{2k-1} = \frac{1}{6} \sum_{k=0}^{\infty} \left(\left(\frac{5}{6}\right)^2\right)^k =$$

$$\frac{1}{6} \cdot \frac{1}{1-\left(\frac{5}{6}\right)^2} = \frac{6}{11}$$

and P(Y=0) = 1 - P(Y=1) = 5/11

1.8.0.3 Example

say X is a continuous r.v with pdf

$f_X(x) = \frac{1}{2}\exp(-|x|)$, $x \in \mathbb{R}$

(this is called a double exponential)

Let Y=I_[-1,1] (X). Find the pdf of Y.

Note: here X is continuous and Y is discrete.

$$P(Y=1) = P(-1 \le X \le 1) = \int_{-1}^{1} \frac{1}{2} e^{-|x|} dx =$$

$$\int_0^1 e^{-x} dx = -e^{-x}\Big|_0^1 = -(e^{-1}-1) = 0.632$$

$$P(Y=0) = 1 - 0.632 = 0.368$$

1.8.0.4 Example

again let X have pdf $f_X(x) = \frac{1}{2}\exp(-|x|)$. Let Y =$X^2$. Then for $y < 0$ we have $P(Y \leq y) = 0$. So let y>0. Then

$$F_Y(y) = P(Y \leq y) = P(X^2 \leq y) = P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right) =$$

$$\int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2}e^{-|x|}dx = \int_0^{\sqrt{y}} e^{-x}dx = -e^{-x}|_0^{\sqrt{y}} = -\left(e^{-\sqrt{y}} - 1\right) = 1 - e^{-\sqrt{y}}$$

$$\text{and so } f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{1}{2\sqrt{y}}e^{-\sqrt{y}}, y \geq 0$$

1.8.0.5 Example

Let $X \sim$ U[0,2], and let $Y = \sin(2\pi X)$. Find f_Y (y).

First of course we always have $-1 \leq \sin(x) \leq 1$ and therefore F_Y (y)=0 if y<1 and F~Y(y)=1 if y>1.

Now if -1<y<1 we have

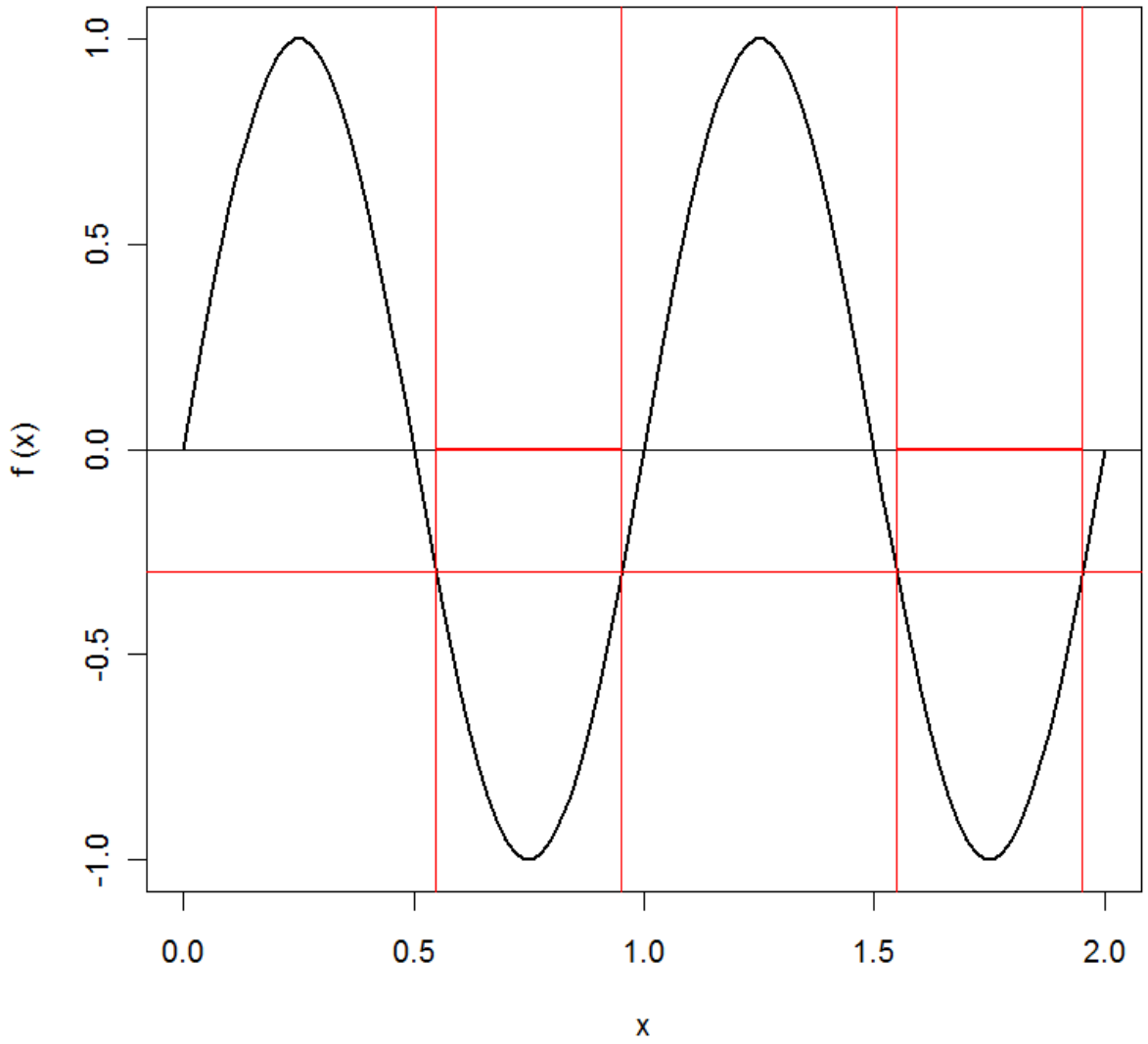
$P(Y \leq y) = P(\sin(2\pi X) \leq y)$

the hard part is solving the inequality

$\sin(2\pi X) \leq y$

The points were we have $\sin(2\pi X) = y$ are of course $y = \arcsin(y)/2/\pi$. Let $a = \arcsin(y)/2/\pi$ and note that arcsin(-y)=-arcsin(y).

Consider the case y=-0.3, then $\arcsin(-0.3)/2/\pi = -0.0485$



and therefore

$$P(Y \le y) = P(\sin(2\pi X) \le y) =$$
$$P(-a + 1/2 \le X \le a + 1) +$$
$$P(-a + 3/2 \le X \le a + 2) =$$
$$(a + 1 - (-a + 1/2) + a + 2 - (-a + 3/2))/2 =$$
$$2a + 1/2$$

Similarly for 0<y<1 we get

$$P(Y \le y) = P(\sin(2\pi X) \le y) =$$
$$P(0 \le X \le a) +$$
$$P(-a + 1/2 \le X \le a + 1) +$$
$$P(-a + 3/2 \le X \le 2) =$$
$$(a + a + 1 - (-a + 1/2) + (2 - (-a + 3/2))/2 =$$
$$2a + 1/2$$
$$f_Y(y) = \frac{d}{dy}(a\sin(y)/\pi + 1/2) = (a\sin)'(y)/\pi$$

if |y|<1

Notice that arcsin is a strictly increasing function, so its derivative is positive. Also

$$\lim_{y\to-\infty} F_Y(y) = \lim_{y\to-\infty} [\arcsin(y)/\pi + 1/2] = 0$$
$$\lim_{y\to\infty} F_Y(y) = \lim_{y\to\infty} [\arcsin(y)/\pi + 1/2] = 1$$

Next up some examples of functions of random vectors:

1.8.0.6 Example

Say (X,Y) is a discrete rv with joint pdf $f_{X,Y}(x,y)=(1-p)^2p^x$, x,y in $\{0,1,..\}$, $y \le x$, and $0<p<1$. Let U=X and V=X-Y. Find $f_{U,V}$ (u,v).

First what are the possible values of (U,V)? We have $u = x \in \{0, 1, ..\}$ and $y \le x$ or $0 \le x - y = v$ and so $v \in \{0, 1, ..\}$.

Finally $v = x - y = u - y \le u$ because $y \ge 0$.

Now for any $(u, v) \in \{0, 1, ..\}$ with $v \le u$ we have

$f_{U,V}$ (u,v) =
P(U=u,V=v) =
P(X=u,X-Y=v) =
P(X=u,u-Y=v) =
P(X=u,Y=u-v) =
$(1-p)^2p^u$

So we see that $f_{U,V}$ (u,v)=$f_{X,Y}$ (u,v), or (X,Y) has the same distribution as (U,V)!

Before we go on let's revisit the first example above, where we had $X \sim$ U[0,1], $\lambda > 0$ and $Y = -\lambda\log(X)$. We found $f_Y(y) = 1/\lambda\exp(-y/\lambda)$ Now let $g(x) = -\lambda\log(x)$ and notice that g is strictly decreasing. Then

$$g(x) = -\lambda \log(x)$$
$$g^{-1}(y) = e^{-y/\lambda}$$
$$(g^{-1})'(y) = -\frac{1}{\lambda} e^{-y/\lambda}$$
$$F_Y(y) = P(Y \le y) = P(g(X) \le y) =$$
$$P(X \ge g^{-1}(y)) = \int_{g^{-1}(y)}^{\infty} f_X(t)dt$$
$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} \int_{g^{-1}(y)}^{\infty} f_X(t)dt =$$
$$-f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| =$$
$$1 \cdot \left| -\frac{1}{\lambda} e^{-y/\lambda} \right| = \frac{1}{\lambda} e^{-y/\lambda}$$

This can be generalized as follows:

Lemma

Let X be a continuous rv with cdf F, and F is strictly increasing. Then $F(X) \sim$ U[0,1].

proof

F is strictly increasing, therefore F^{-1} exists, and so

$P(F(X) \le x) = P(X \le F^{-1}(x)) = F(F^{-1}(x)) = x$

and even a bit more:

Theorem (Probability Integral Transform)

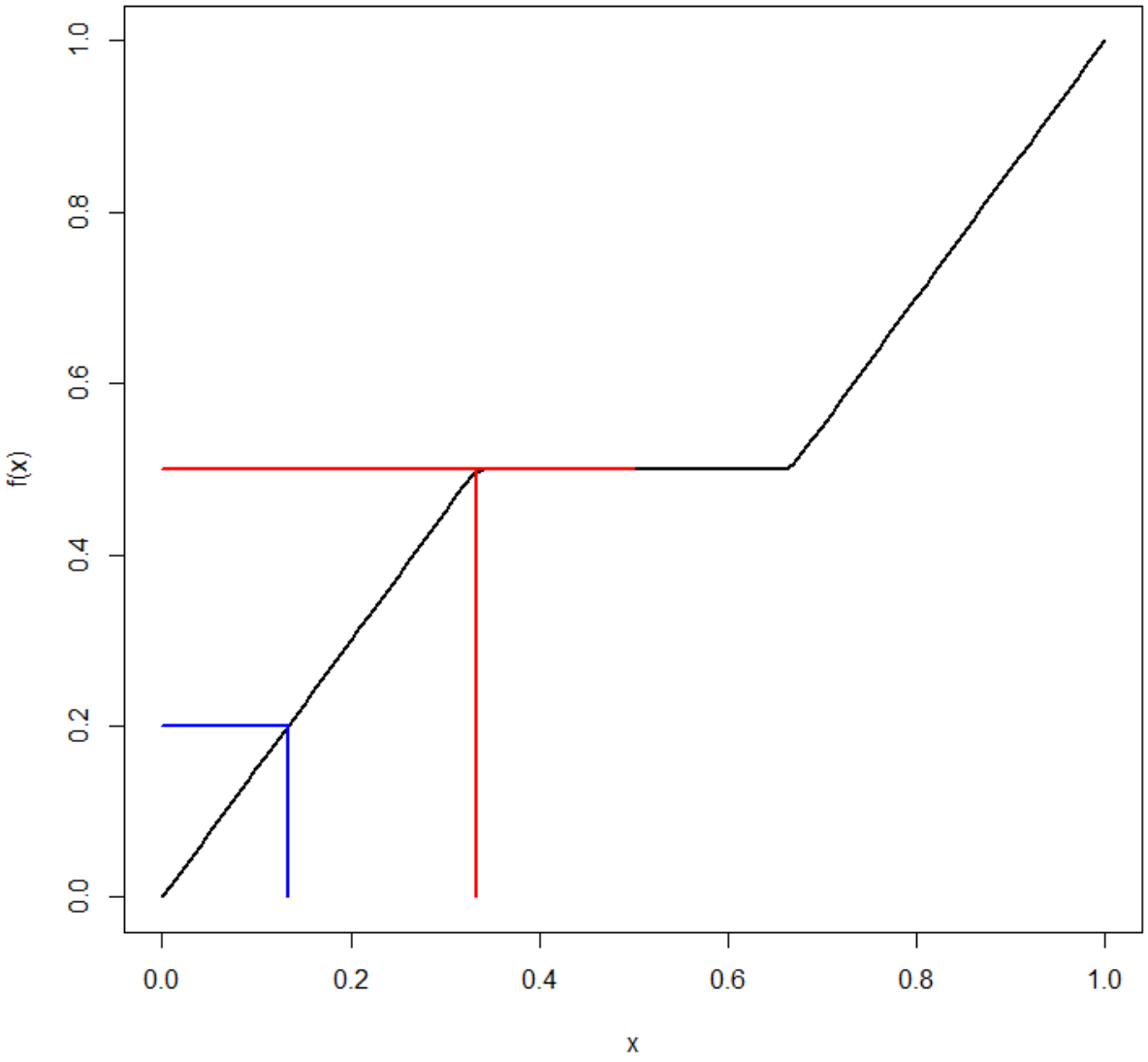
Let X be a continuous rv with cdf F. Then $F(X) \sim$ U[0,1]

proof let F be the cdf and define the generalized inverse function F^* by

$$F^*(x) = \min\{t : F(t) \ge x\}$$

First note that if F is strictly increasing we have $F^* = F^{-1}$.

Moreover we have F(F*(x))=x. This is easiest to see whith a graph:



Now the proof is the same as the one for the lemma!

In one dimension this is rarely useful, it is usually easier to just do the problem directly as above. It does become useful in higher dimensions.

1.8.0.7 Example

say (X,Y) is a bivariate standard normal r.v, that is it has joint density given by

$$f(x,y) = \frac{1}{2\pi}\exp\left\{-\frac{1}{2}(x^2+y^2)\right\}$$

for $(x, y) \in \mathbb{R}^2$

Let the r.v. (U,V) be defined by U=X+Y and V=X-Y. Find the joint pdf of (U,V).

To start let's define the functions g_1 (x,y) = x+y and g_2 (x,y) = x-y, so that U=g_1 (X,Y) and V = g_2 (X,Y).

For what values of u and v is f_(U,V) (u,v) positive? Well, for any values for which the system of 2 linear equations in two unknowns u=x+y and v=x-y has a solution. These solutions are

x = h_1 (u,v) = (u + v)/2 y = h_2 (u,v) = (u - v)/2

From this we find that for any $(u, v) \in \mathbb{R}^2$ there is a unique $(x, y) \in \mathbb{R}^2$ such that u=x+y and v=x-y. So the transformation $(x, y) \to (u, v)$ is one-to-one and therefore has a Jacobian given by

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial y} \end{vmatrix} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

Now from multivariable calculus we have the following:

$$f_{U,V}(u,v) = f_{X,Y}(h_1(x,y), h_2(x,y)) \cdot |J| =$$
$$\frac{1}{2\pi} \exp\left(-\frac{1}{2}\left(\left(\frac{u+v}{2}\right)^2 + \left(\frac{u-v}{2}\right)^2\right)\right) \cdot \left|-\frac{1}{2}\right| =$$
$$\frac{1}{4\pi} \exp\left(-\frac{1}{2}\left(\frac{u^2+2uv+v^2+u^2-2uv+v^2}{4}\right)\right) = \frac{1}{4\pi} \exp\left(-\frac{1}{2}\left(\frac{u^2+v^2}{2}\right)\right) =$$
$$\left(\frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{1}{2}\frac{u^2}{2}}\right) \cdot \left(\frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{1}{2}\frac{v^2}{2}}\right)$$

Note that the density factors into a function of u and a function of v. As we saw beforethis means that U and V are independent.

1.8.0.8 Example

Say X_1 , .., X~n‘ are iid U[0,1] Let $Y_1 = X_1$, $Y_2 = X_2 - X_1$,..,$Y_n = X_n - X_{n-1}$. Now

$$y_1 = x_1, \text{so } x_1 = y_1$$

$$y_2 = x_2 - x_1 = x_2 - y_1, \text{ so } x_2 = y_2 + y_1$$

$$y_k = x_k - x_{k-1} = x_k - y_{k-1}, \text{ so } x_k = y_k + y_{k-1}$$

$$\begin{vmatrix} 1 & 0 & 0 & .. & 0 \\ 1 & 1 & 0 & .. & 0 \\ 0 & 1 & 1 & 0 & \\ .. & & & & \\ 0 & .. & 0 & 1 & 1 \end{vmatrix} = 1$$

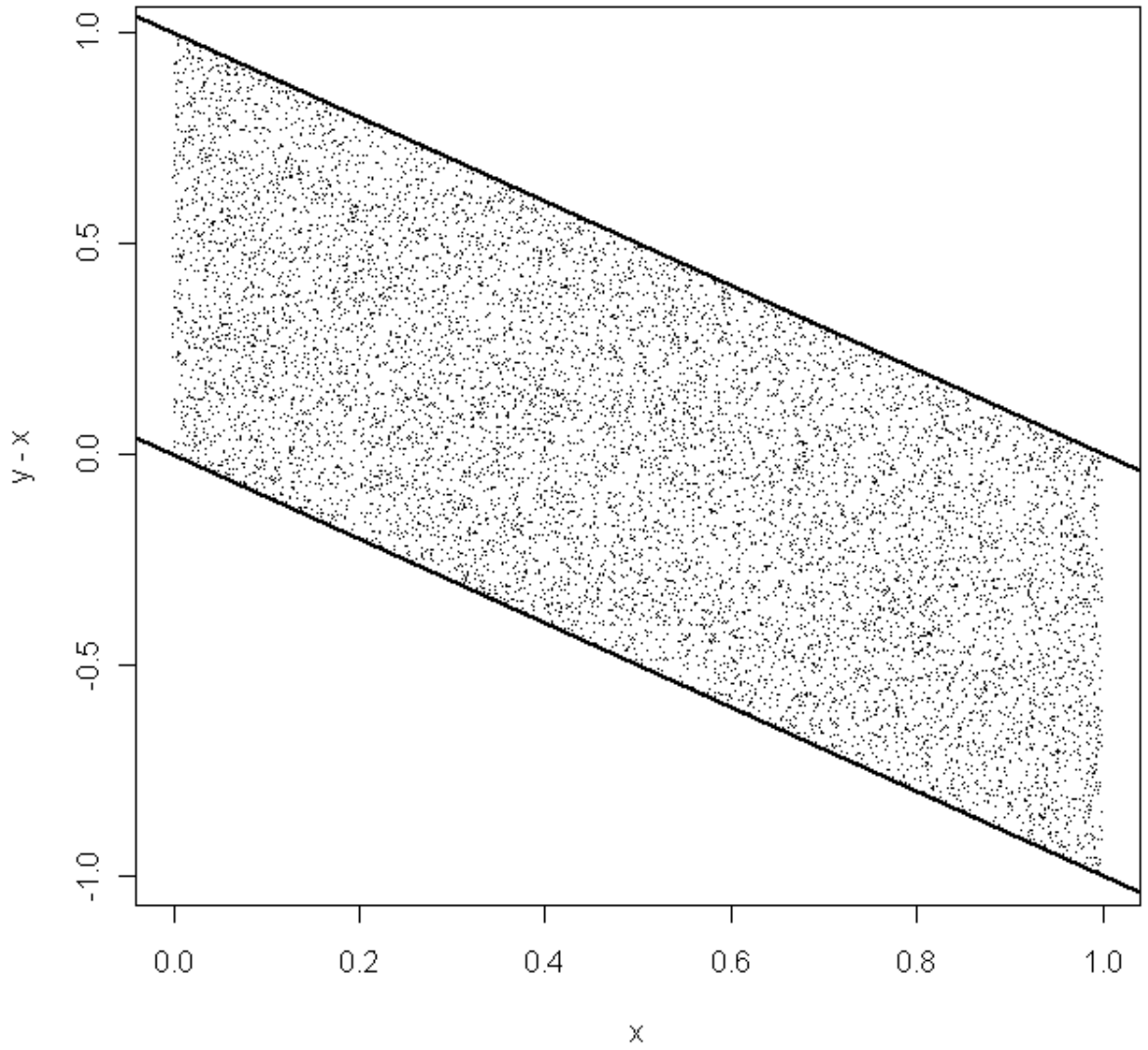
$$f_{Y_1,..,Y_n}(y_1,..,y_n) = f_{X_1,..,X_n}(y_1, y_1 + y_2, .., y_{n-1} + y_n)|1| = 1$$

so (Y_1 , .., Y_n) is uniform. But careful, uniform on what set? $y_2 = x_2 - x_1$, $0 \leq x_i \leq 1$, therefore $-1 \leq y_2 \leq 1$.

We have

$0 \leq y_1 \leq 1$
$-y_k - 1 \leq y_k \leq 1 - y_k - 1$, k=2,..,n

For n=2 the set is shown here:



1.8.0.9 Example

A rv X is called a normal (or Gaussian) rv with mean μ and standard deviation σ if it had density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

a special case is a** standard normal rv**, which has $\mu = 0$ and $\sigma = 1$.

Say X and Y are independent standard normal rv's. Let Z = X + Y. Find the pdf of Z.

Note: now we have a transformation from $\mathbb{R}^2 \to \mathbb{R}$.

Z = X + Y = U in the example above, so the pdf of Z is just the marginal of U and we find

$$f_Z(z) = f_U(z) = \int_{-\infty}^{\infty} f_{U,V}(z,v)dv =$$

$$\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{1}{2}\frac{z^2}{2}} \right) \cdot \left(\frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{1}{2}\frac{v^2}{2}} \right) dv =$$

$$\left(\frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{1}{2}\frac{z^2}{2}} \right) \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{1}{2}\frac{v^2}{2}} dv =$$

$$\frac{1}{\sqrt{2\pi}\sqrt{2}} e^{-\frac{1}{2}\frac{z^2}{2}}$$

Say X and Y are two continuous independent r.v with pdf f's f_X and f~Y‘ , and let Z = X+Y. If we repeat the above calculations we can show that in general the pdf of Z is given by

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

This is called the *convolution formula*.

There is a second method for deriving the convolution formula which is useful, using the law of total probability for rv's:

$$f_Z(z) = f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{X+Y|Y=y}(z|y) f_Y(y) dy$$

$$F_{X+Y|Y=y}(z) = P(X+Y < z|Y = y)$$

$$= P(X+y < z) = P(X < z-y) = F_X(z-y)$$

$$f_{X+Y|Y=y}(z) = f_X(z-y)$$

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y) dy$$

1.8.0.10 Example

say X and Y are independent exponential rv's with rate λ. Find the pdf of Z=X+Y

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-x) f_Y(x) dx =$$

$$\int_0^z \lambda e^{-\lambda(z-x)} \lambda e^{-\lambda x} dx = \lambda^2 e^{-\lambda z} \int_0^z dx = \lambda^2 z e^{-\lambda z}, z \geq 0$$

1.8.0.11 Example

Say (X,Y) is a discrete rv with joint pdf $f_{X,Y}$ (x,y)=(1-p)^{2}p^x, $x, y \in \{0, 1, ..\}, y \leq x, 0 < p < 1$. Let U=I(X=Y). Find f_U (u)

$$f_U(1) = P(U = 1) = P(X = Y) = \sum_{x=0}^{\infty} P(X = x, Y = x) =$$

$$\sum_{x=0}^{\infty} (1-p)^2 p^x = (1-p)^2 \frac{1}{1-p} = 1-p$$

1.8.0.12 Example

Say X_1 , .., X~n' are iid U[0,1]. Let M=max{X_1 , .., X_n }.

Find E[M] and Var(M).

First we find f_M

$$F_M(x) = P(M \le x) = P(\max\{X_1,..,X_n\} \le x) =$$

$$P(X_1 \le x,..,X_n \le x) = \prod_{i=1}^{n} P(X_i \le x) = [P(X_1 \le x)]^n = x^n$$

$$f_M(x) = \frac{\partial}{\partial x} F_M(x) = \frac{\partial}{\partial x} x^n = nx^{n-1}$$

Now

$$EM = \int_{-\infty}^{\infty} x f_M(x)dx = \int_0^1 x n x^{n-1} dx = \frac{n}{n+1} x^{n+1}\Big|_0^1 = \frac{n}{n+1}$$

$$EM^2 = \int_{-\infty}^{\infty} x^2 f_M(x)dx = \int_0^1 x^2 n x^{n-1} dx = \frac{n}{n+2} x^{n+2}\Big|_0^1 = \frac{n}{n+2}$$

$$VM = \frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2$$

This is a special case of what are called *Order Statistics*. Many statistical methods, for example the median and the range, are based on an ordered data set.

One of the difficulties when dealing with order statistics are ties, that is the same observation appearing more than once. This should only occur for discrete data because for continuous data the probabiltity of a tie is zero. They may happen anyway because of rounding, but we will ignore them in what follows.

Say X_1, .., X_n are iid with density f. Then $X_{(i)}$ is the i^{th} order statistics if $X_{(1)} < \ldots < X_{(i)} < \ldots < X_{(n)}$.

Note $X_{(1)} = \min\{X_i\}$ and $X_{(n)} = \max\{X_i\}$.

Let's find the pdf of $X_{(i)}$. For this let Y be a r.v. that counts the number of $X_j \le x$ for some fixed number x. We will see shortly that if p=F(x)

$$P(Y = j) = \binom{n}{j} p^j (1-p)^{n-j}$$

Note also that the event $\{Y \ge i\}$ means that more than i observations are less or equal to x, so the i^{th} largest is less or equal to x. Therefore

[](graphs/prob623.png" >

with that we find

$$\frac{d}{dx}F_{X_{(i)}}(x) = \sum_{k=i}^{n} \binom{n}{k} F(x)^k (1-F(x))^{n-k} =$$

$$\sum_{k=i}^{n} \binom{n}{k} \frac{d}{dx}[F(x)^k (1-F(x))^{n-k}] =$$

$$\sum_{k=i}^{n} \binom{n}{k} [kF(x)^{k-1} f(x)(1-F(x))^{n-k} + F(x)^k (n-k)(1-F(x))^{n-k-1}(-f(x))] =$$

$$\sum_{k=i}^{n} \binom{n}{k} [kF(x)^{k-1}(1-F(x))^{n-k} - F(x)^k (n-k)(1-F(x))^{n-k-1}] f(x)$$

let $t = F(x)$

$$\sum_{k=i}^{n} \binom{n}{k} [kt^{k-1}(1-t)^{n-k} - t^k (n-k)(1-t)^{n-k-1}] =$$

$$\sum_{k=i}^{n} \binom{n}{k} kt^{k-1}(1-t)^{n-k} - \sum_{k=i}^{n} \binom{n}{k} t^k (n-k)(1-t)^{n-k-1} =$$

$$\binom{n}{i} it^{i-1}(1-t)^{n-i} + \sum_{k=i+1}^{n} \binom{n}{k} kt^{k-1}(1-t)^{n-k} -$$

$$\sum_{k=i}^{n-1} \binom{n}{k} t^k (n-k)(1-t)^{n-k-1} = \quad (n = k \text{ term is } 0)$$

$$\frac{n!}{i!(n-i)!} it^{i-1}(1-t)^{n-i} + \sum_{k=i+1}^{n} \binom{n}{k} kt^{k-1}(1-t)^{n-k} -$$

$$\sum_{k=i}^{n-1} \binom{n}{k} t^k (n-k)(1-t)^{n-k-1} =$$

$$\frac{n!}{(i-1)!(n-i)!} t^{i-1}(1-t)^{n-i} + \sum_{k=i}^{n-1} \binom{n}{k+1} (k+1) t^k (1-t)^{n-k-1} -$$

$$\sum_{k=i}^{n-1} \binom{n}{k} t^k (n-k)(1-t)^{n-k-1} = *$$

Note $\binom{n}{k+1}(k+1) = \frac{n!}{(k+1)!(n-k-1)!}(k+1) = \frac{n!}{k!(n-k-1)!}$

and $\binom{n}{k}(n-k) = \frac{n!}{k!(n-k)!}(n-k) = \frac{n!}{k!(n-k-1)!}$, so

$$* = \frac{n!}{(i-1)!(n-i)!} t^{i-1}(1-t)^{n-i} + \sum_{k=i}^{n-1} \frac{n!}{k!(n-k-1)!} t^k (1-t)^{n-k-1} -$$

$$\sum_{k=i}^{n-1} \frac{n!}{k!(n-k-1)!} t^k (n-k)(1-t)^{n-k-1} = \frac{n!}{(i-1)!(n-i)!} t^{i-1}(1-t)^{n-i}$$

so

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!} F(x)^{i-1}(1-F(x))^{n-i} f(x)$$

1.8.0.13 Example

Say X_1, .., X_n are iid U[0,1]. Then for 0<x<1 we have f(x)=1 and F(x)=x. Therefore

$$f_{X_{(i)}}(x) = \frac{n!}{(i-1)!(n-i)!}F(x)^{i-1}(1-F(x))^{n-i}f(x) =$$

$$\frac{n!}{(i-1)!(n-i)!}x^{i-1}(1-x)^{n-i}, \quad 0 < x < 1$$

or

$$f_{Min}(x) = f_{X_{(1)}}(x) = \frac{n!}{(1-1)!(n-1)!}x^{1-1}(1-x)^{n-1} = n(1-x)^{n-1}$$

$$f_{Max}(x) = f_{X_{(n)}}(x) = \frac{n!}{(n-1)!(n-n)!}x^{n-1}(1-x)^{n-n} = nx^{n-1}$$

1.8.0.14 Example

Say X_1, .., X_n are iid U[0,1]. Let g be the density of the order statistic ($X_{(1)}$, .., $X_{(n)}$). Then $g(x_{(1)}, ..,x_{(n)})=n!$ for $0<x_{(1)}< ..<x_{(n)}<1$

The simple "proof" is as follows: for any set of n distinct numbers there are n! permutations, exactly one of which has $0<x_{(1)}< ..<x_{(n)}<1$.

A "formal" proof can be done using a generalization of the change of variables formula. The problem is that the inverse transform is not unique, in fact there are n! of them because the ordered set of numbers could have come from any of the n! permutations. Once the inverse transform is fixed, though, the Jacobian is just the identity matrix with the rows rearranged, and therefore has determinant 1. Then

$g(x_{(1)}, ..,x_{(n)}) = n!f(x_1, ..,x_n)|J| = n!$

2 Standard Probability Distributions

2.1 Discrete Distributions

2.1.1 Discrete Uniform

Let $N \geq 2$ be an integer and consider the rv X that chooses a number from 1 to N with equal probability, that is

$P(X = k) = 1/N$ for $1 \leq k \leq N$

Then

$$E[X] = \sum_{k=1}^{N} k\frac{1}{N} = \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2}$$

$$E[X^2] = \sum_{k=1}^{N} k^2\frac{1}{N} = \frac{1}{N}\frac{N(N+1)(2N+1)}{6} = \frac{(N+1)(2N+1)}{6}$$

$$Var(X) = \frac{(N+1)(2N+1)}{6} - \left(\frac{N+1}{2}\right)^2 =$$

$$\frac{1}{12}[2(N+1)(2N+1) - 3(N+1)^2] =$$

$$\frac{1}{12}[4N^2 + 6N + 2 - 3N^2 - 6N - 3] =$$

$$\frac{1}{12}[N^2 - 1] = \frac{(N+1)(N-1)}{12}$$

2.1.2 General Discrete RV on a Finite Set

Let $N \geq 2$ be an integer and consider the rv X with

$P(X = k) = p_k$ for $1 \leq k \leq N$

Nothing more can be said until the p_k are specified.

2.1.2.1 Example

say $p_k = 2p_k\text{-}1$

$$p_k = 2p_{k-1} = 4p_{k-2} = 2^k p_1, k = 2, .., N$$

$$\sum_{k=1}^{N} p_k = \sum_{k=1}^{N} 2^k p_1 = p_1 \sum_{k=0}^{N-1} 2^{k+1} = 2p_1 \sum_{k=0}^{N-1} 2^k =$$

$$2p_1\frac{1-2^N}{1-2} = 2p_1(2^N - 1) = 1$$

$$p_1 = \frac{1}{2^{N+1}-2}$$

$$p_k = \frac{2^k}{2^{N+1}-2} \quad k = 1, .., N$$

2.1.3 Bernoulli Distribution

A r.v. X is said have a *Bernoulli* distribution with success parameter p if

P(X=0)=1-p and P(X=1)=p.

Sometimes we call the outcomes "success"(=1) and "failure"(=0)

Note: often we use q = 1-p

Shorthand: $X \sim$ Ber(p)

$$
\begin{aligned}
&EX = 0 * q + 1 * p = p \\
&EX^2 = 0^2 * q + 1^2 * p = p \\
&Var(X) = EX^2 - (EX)^2 = p - p^2 = pq \\
&\psi(t) = E[\exp(tX)] = \exp(t0)q + \exp(t1)p = q + e^t p
\end{aligned}
$$

2.1.4 Binomial Distribution

Say Y_1 , ... , Y_n are iid Ber(p) and let X=Y_1 +..Y_n, then X is said to have a *binomial* distribution with parameters n and p. ($X \sim$ Bin(n,p)).

We have

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

because "X=k" means k successes and n-k failures. Any specific sequence of k successes and n-k failures has probability $p^k q^{n-k}$, and there are n choose k such sequences.

It is easy to see that this defines a proper pdf:

$$\sum_{i=0}^{n} P(X = i) =$$

$$\sum_{i=0}^{n} \binom{n}{i} p^i q^{n-i} =$$

$$(p + q)^i = 1$$

which also explains the name binomial.

For the mean and variance we have

$$E[X] = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} = \sum_{k=1}^{n} k\frac{n!}{(n-k)!k!}p^k q^{n-k} =$$

$$\sum_{k=1}^{n} \frac{n!}{(n-k)!(k-1)!}p^k q^{n-k} =$$

$$np\sum_{k=1}^{n} \frac{(n-1)!}{(n-1-(k-1))!(k-1)!}p^{k-1} q^{n-1-(k-1)} =$$

$$np\sum_{j=0}^{n-1} \frac{n!}{(n-1-j)!j!}p^j q^{n-1-j} = np$$

or easier:

$$E[X] = E[\sum_{k=1}^{n} Y_k] = \sum_{k=1}^{n} E[Y_k] = \sum_{k=1}^{n} p = np$$

$$Var(X) = E[(X-np)^2] = E[\left(\sum_{k=1}^{n} Y_k - np\right)^2] =$$

$$E[\left(\sum_{k=1}^{n} (Y_k - p)\right)^2] =$$

$$E[\sum_{k=1}^{n}\sum_{j=1}^{n} (Y_k - p)(Y_j - p)] =$$

$$\sum_{k=1}^{n}\sum_{j=1}^{n} E[(Y_k - p)(Y_j - p)] =$$

$$\sum_{k=1}^{n} E[(Y_k - p)(Y_k - p)] + \sum_{j=1, j\neq k}^{n} E[(Y_k - p)(Y_j - p)] =$$

$$\sum_{k=1}^{n} E[(Y_k - p)^2] + \sum_{j=1, j\neq k}^{n} E[(Y_k - p)]E[(Y_j - p)] =$$

$$\sum_{k=1}^{n} pq + \sum_{j=1, j\neq k}^{n} 0 = npq$$

and

$$E[e^{tX}] = \sum_{k=0}^{n} e^{tk}\binom{n}{k}p^k q^{n-k} =$$

$$\sum_{k=0}^{n} \binom{n}{k}(e^t p)^k q^{n-k} = (q + p\ e^t)^n$$

2.1.4.1 Example

A company wants to hire 5 new employees. From previous experience they know that about 1 in 10 applicants are suitable for the jobs. What is the probability that if they interview 20 applicants they will be able to fill those 5 positions?

Consider each interview a "trial" with the only two possible outcomes: "success" (can be hired) or "failure" (not suitable). Assumptions:

1. "success probability" is the same for all applicants (as long as we know nothing else about them this is ok.)
2. trials are independent (depends somewhat on the setup of the interviews but should be ok)

then if we let X = "#number of suitable applicants in the group of 20" we have $X \sim$ Bin(20,0.1) and we find

$$P(X \geq 5) = 1 - P(X \leq 4) =$$
$$1 - \sum_{i=0}^{4} \binom{20}{i} 0.1^i 0.9^{20-i} =$$
$$1 - (0.9^{20} + 20 \cdot 0.1 \cdot 0.9^{19} + 190 \cdot 0.1^2 \cdot 0.9^{18} +$$
$$1140 \cdot 0.1^2 \cdot 0.9^{18} + 4845 \cdot 0.1^2 \cdot 0.9^{18}) =$$
$$1 - 0.9568 = 0.0432$$

2.1.5 Geometric Distribution

Say Y_1 , Y_2 , .. are iid Ber(p) and let X be the number of trials needed until the first success. Then X is said to have a *geometric* distribution with rate p ($X \sim G(p)$), and we have

$$P(X = k) = pq^{k-1},\ k = 1, 2, ..$$

Note: sometime the geometric is defined as the number of failures before the first success. Clearly this is then X-1.

This defines a proper pdf:

$$\sum_{i=1}^{\infty} P(X = i) =$$

$$\sum_{i=1}^{\infty} pq^{i-1} =$$

$$p \sum_{i=0}^{\infty} q^{i} = p \frac{1}{1-q} = 1$$

For the mean and variance we have

$$E[X] = \sum_{k=1}^{\infty} kpq^{k-1} = p\sum_{k=1}^{\infty} \frac{d}{dx}x^k|_{x=q} =$$
$$p\frac{d}{dx}\left(\sum_{k=1}^{\infty} x^k\right)|_{x=q} = p\frac{d}{dx}\left(\sum_{k=0}^{\infty} x^k - 1\right)|_{x=q} =$$
$$p\frac{d}{dx}\left(\frac{1}{1-x} - 1\right)|_{x=q} = p\frac{1}{(1-x)^2}|_{x=q} = \frac{p}{(1-q)^2} = \frac{1}{p}$$

$$E[X(X-1)] = \sum_{k=1}^{\infty} k(k-1)pq^{k-1} = pq\sum_{k=2}^{\infty} \frac{d^2}{dx^2}x^k|_{x=q} =$$
$$pq\frac{d^2}{dx^2}\left(\sum_{k=0}^{\infty} x^k - 1 - x\right)|_{x=q} = pq\frac{d^2}{dx^2}\left(\frac{1}{1-x} - 1 - x\right)|_{x=q} =$$
$$pq\frac{d}{dx}\left(\frac{1}{(1-x)^2} - 1\right)|_{x=q} = pq\left(\frac{2}{(1-x)^3}\right)|_{x=q} =$$
$$pq\frac{2}{(1-q)^3} = \frac{2q}{p^2}$$
$$E[X(X-1)] = E[X^2 - X)] = E[X^2] - E[X] = E[X^2] - \frac{1}{p}$$
$$Var(X) = E[X^2] - E[X]^2 = E[X(X-1)] + \frac{1}{p} - (\frac{1}{p})^2 =$$
$$\frac{2q}{p^2} + \frac{1}{p} - (\frac{1}{p})^2 = \frac{2(1-p)+p-1}{p^2} = \frac{q}{p^2}$$

$$\psi(t) = E[e^{tX}] = \sum_{k=1}^{\infty} e^{tk}pq^{k-1} = e^t p\sum_{k=0}^{\infty}(e^t q)^k = \frac{e^{tp}}{1-e^{tq}}$$

2.1.5.1 ** Example**

(same as above) How many applicants will the company need to interview to be 90% sure to be able to fill at least one of the five positions?

If we let Y be the number of trials until the first success (= an applicant is suitable) we have $Y \sim G(0.1)$. Then

$$0.9 = P(Y \leq n) = \sum_{k=1}^{n} pq^{k-1} =$$

$$p\sum_{k=0}^{n-1} q^k = p\frac{1-q^{(n-1)+1}}{1-q} = 1 - q^n$$

$$q^n = 0.1$$

$$\log q^n = n\log q = \log 0.1$$

$$n = \frac{\log 0.1}{\log 0.9} = 21.8 \sim 22$$

In general the geometric rv. is a model for "lifetimes" or "times until failure" of components, that is for the number of time periods until a component fails. But how do we know in real live whether the geometric might be a good model for a specific case? The next theorem helps:

Theorem

Say X is a discrete rv. on {1,2,3,..} Then

P(X>k)=P(X>k+j|X>j) for all k and j iff $X \sim G(p)$

Note P(X>k)=P(X>k+j|X>j) for all k and j is called the **memoryless property**, and the theorem states that for discrete rv.s on the positive integers this property is unique to the geometric rv.

proof

Say $X \sim G(p)$, then

$$P(X > k) = 1 - P(X \leq k) =$$
$$1 - \sum_{i=1}^{k} pq^{i-1} = 1 - p\sum_{i=0}^{k-1} q^i =$$
$$1 - p(1 - q^{(k-1)+1})/(1-q) = q^k$$

and then

$$P(X > k+j|X > j) = \frac{P(X > k+j, X > j)}{P(X > j)} =$$

$$\frac{P(X > k+j)}{P(X > j)} = \frac{q^{k+j}}{q^j} = q^k = P(X > k)$$

now assume $X \in \{1, 2, ..\}$ has the memoryless property. Let the event A={X>1}, then

$$P(X > k+1) = P(X > k+1|A)P(A) + P(X > k+1|A^c)P(A^c) =$$
$$P(X > k+1|X > 1)P(X > 1) + P(X > k+1|X = 1)P(X = 1) =$$
$$P(X > k+1|X > 1)P(X > 1) + 0 =$$
$$P(X > k)P(X > 1)$$

by the memoryless property. Now set $q = P(X > 1)$ then for $k > 1$

$$P(X > k) = qP(X > k-1) = q^2P(X > k-2) = .. = q^{k-1}P(X > 1) = q^k$$

So the geometric is a reasonable model if it is reasonable to assume an experiment has the memoryless property.

2.1.5.2 Example

Say we want to model the number of days until a light bulb burns out. Is the geometric a good model for this? The question is whether the number of days has the memoryless property?

2.1.5.3 Example

Say we want to model the number of years until a person dies. Is the geometric a good model for this? The question is whether the number of years has the memoryless property?

2.1.6 Negative Binomial Distribution

Despite the different name this is actually a generalization of the geometric, namely where X is the number of trial needed until the r$^{\text{th}}$ success. ($X \sim NB(p, r)$).

The pdf is given by

$$P(X = k) = \binom{k-1}{r-1} p^r q^{k-r}$$

because we have to have r successes (with probability p each) and k-r failures (with probability q) Moreover the r -1 success (before the last) one can be at any point during the k-1 trials (before the last one).

As an alternative definition one often uses the number of failures until the first success, Y=X-r. With this we find

$$Y = X - r$$

$$P(Y = k) = P(X = k + r) =$$

$$\binom{k+r-1}{r-1} p^r q^{k+r-r} = \binom{k+r-1}{r-1} p^r q^k; \quad k = 0, 1, ..$$

$$\binom{k+r-1}{r-1} = \frac{(k+r-1)!}{(k+r-1-(r-1))!(r-1)!} = \frac{(k+r-1)!}{(r-1)!k!} =$$

$$\frac{(k+r-1)!}{(k+r-1-k)!k!} = \binom{k+r-1}{k}$$

$$P(Y = k) = \binom{k+r-1}{k} p^r q^k; \quad k = 0, 1, ..$$

Does this define a proper pdf? To show this takes a bit of work. First need an extension of binomial coefficients:

$$\binom{n}{k} = \frac{n(n-1)(n-2)..(n-k+1)}{k!}$$

$$\binom{\alpha}{k} := \frac{\alpha(\alpha-1)(\alpha-2)..(\alpha-k+1)}{k!}$$

for any $\alpha \geq k$ and 0 otherwise.

Next we need the Taylor series expansion of $(1 + x)^\alpha$ at x=0:

$$f(x) = (1+x)^{\alpha}$$

$$\frac{d}{dx}f(x)|_{x=0} = \alpha(1+x)^{\alpha-1}|_{x=0} = \alpha$$

$$\frac{d^2}{dx^2}f(x)|_{x=0} = \alpha(\alpha-1)(1+x)^{\alpha-2}|_{x=0} = \alpha(\alpha-1)$$

$$\frac{d^k}{dx^k}f(x)|_{x=0} = \prod_{r=0}^{k-1}(\alpha-r)$$

$$(1+x)^{\alpha} = \sum_{k=0}^{\infty}\left[\prod_{r=0}^{k-1}(\alpha-r)\right]\frac{x^k}{k!} =$$

$$\sum_{k=0}^{\infty}\frac{\alpha(\alpha-1)..(\alpha-k+1)}{k!}x^k = \sum_{k=0}^{\infty}\binom{\alpha}{k}x^k$$

This is called the *Binomial series* and is correct for any (even complex) number α. Notice that it is of course a generalization of the Binomial formula for non-integer powers:

$$(1+x)^n = \sum_{k=0}^{\infty}\binom{n}{k}x^k = \sum_{k=0}^{n}\binom{n}{k}x^k$$

Finally

$$\binom{k+r-1}{k} = \frac{(k+r-1)(k+r-2)..r}{k!} =$$

$$(-1)^k \frac{(-r)(-r-1)..(-r-k+1)}{k!} = (-1)^k \binom{-r}{k}$$

$$p^{-r} = (1-q)^{-r} =$$

$$\sum_{k=0}^{\infty} \binom{-r}{k} (-q)^k$$

$$\sum_{k=0}^{\infty} \binom{k+r-1}{k} q^k$$

$$\sum_{k=0}^{\infty} \binom{k+r-1}{k} p^r q^k = 1$$

and here we also have an explanation for the name "negative binomial"!

Theorem

Let $Y_1, .., Y_r \sim G(p)$ and independent, the $X = Y_1 + .. + Y_r \sim NB(p, r)$.

proof (by induction)

if r=1 X=Y_1 , so

$$P(X = k) = P(Y_1 = k) = pq^{k-1} = \binom{k-1}{1-1} p^1 q^{k-1}$$

say assertion is true for r, then

$$P(X_{r+1} = k) = P(\sum_{j=1}^{r+1} Y_j = k) =$$

$$\sum_{j=1}^{k-1} P(\sum_{j=1}^{r} Y_j + Y_{r+1} = k | Y_{r+1} = j) P(Y_{r+1} = j) =$$

$$\sum_{j=1}^{k-1} P(X_r = k-j) P(Y_{r+1} = j) =$$

$$\sum_{j=1}^{k-1} \binom{k-j-1}{r-1} p^r q^{k-j-r} p q^{j-1} =$$

$$p^{r+1} q^{k-(r+1)} \sum_{j=1}^{k-1} \binom{k-j-1}{r-1} =$$

$$\binom{k-1}{r} p^{r+1} q^{k-(r+1)}$$

Note that we don't need to worry about the constant term, it has to be what it needs to be to make this a proper random variable!

We also find

$$E[X] = E[\sum_{j=1}^{r} Y_j] = \sum_{j=1}^{r} E[Y_j] = \frac{r}{p}$$

$$Var[X] = Var(\sum_{j=1}^{r} Y_j) = \sum_{j=1}^{r} Var(Y_j) = \frac{rq}{p^2}$$

$$\psi_X(t) = \psi_{Y_1}(t)^r = \left(\frac{e^{tp}}{1-e^{tq}} \right)^r$$

Theorem:

Say $X \sim$ Bin(n,p) and $Y \sim$ NB(p,r). Then F_X (r-1)=1-F_Y (n)

proof (probabilistic)

F_X (r-1) = P(X<r) is the probability of less than r successes in n trials

1-F_Y (n) = P(Y>n) is the probability of not having r successes in n trials

same thing!

2.1.7 Hypergeometric Distribution

One of the problems with the use of the Binomial distribution in real life is that most sampling is done in such a way that the same object can not be selected a second time.

2.1.7.1 Example

In a survey of 100 likely voters, 45 said they would vote for party AA.

Obviously this "selecting" would not allow the same person to be chosen twice.

If the selection is done *without repetition*, we get to the hypergeometric distribution. Of course, if the sample size is small compared to the population size the probabilities are almost the same.

In general the hypergeometric can be described as follows:

Consider an urn containing N+M balls, of which N are white and M are black. If a sample of size n is chosen at random and if X is the number of white balls chosen, then X has a hypergeometric distribution with parameters (n,N,M).

$X \sim HG(n, N, M)$

We have

$$P(X = k) = \frac{\binom{N}{k}\binom{M}{n-k}}{\binom{N+M}{n}}$$

because there are $\binom{N}{k}$ ways to select k objects from N without repetition and without order. Likewise there are $\binom{M}{n-k}$ selections of n-k out of M and $\binom{N+M}{n}$ selections of n objects from N+M.

Does this define a proper pdf? This is a consequence of

Theorem (Vandermonde)

For any $N, M, n \geq 0$ we have

$$\binom{N+M}{n} = \sum_{k=0}^{n} \binom{N}{k}\binom{M}{n-k}$$

We will give two very different proofs of this identity:

proof (Combinatorial)

Consider an urn with N white and M black balls. We randomly select n of these without order and without repetition. How many different arrangements are there? Clearly the answer is $\binom{N+M}{n}$

How many arrangements are there if we want k white balls (and therefore also have to have n-k black balls)? Again, clearly the answer is $\binom{N}{k}\binom{M}{n-k}$ but the first selection is the the same as the second, where we allow k to be any number between 0 and n, and the identity follows.

proof (Algebraic)

First note that

$$\left(\sum_{i=0}^{N} a_i x^i\right)\left(\sum_{j=0}^{M} b_j x^j\right) =$$

$$\sum_{i=0}^{N}\sum_{j=0}^{M} a_i b_j x^{i+j} =$$

$$a_0 b_0 + (a_0 b_1 + a_1 b_0)x +$$

$$(a_0 b_2 + a_1 b_1 + a_2 b_0)x^2 + \ldots +$$

$$\left(\sum_{k=0}^{r} a_k b_{r-k}\right)x^r + \ldots + \left(\sum_{k=0}^{N+M} a_k b_{r-k}\right)x^{N+M} =$$

$$\sum_{r=0}^{N+M}\left(\sum_{k=0}^{r} a_k b_{r-k}\right)x^{N+M}$$

where we define $a_i = 0$ for i=N+1,..,N+M and $b_j = 0$ for j=M+1,..,N+M.

Now using the binomial formula we have

$$\sum_{r=0}^{N+M} \binom{N+M}{r} x^r =$$

$$(1+x)^{N+M} =$$

$$(1+x)^N (1+x)^M =$$

$$\left(\sum_{i=0}^{N} \binom{N}{i} x^i\right)\left(\sum_{j=0}^{M} \binom{M}{j} x^j\right) =$$

$$\sum_{r=0}^{N+M} \left(\sum_{k=0}^{r} \binom{N}{i}\binom{M}{r-i}\right) x^r$$

and the identity follows from the fact that if two polynomials are equal for all x they have to have the same coefficients.

This identity is named after the French mathematician [Alexandre Theofile Vandermonde](https://en.wikipedia.org/wiki/Alexandre-Th%C3%A9ophile_Vandermonde (1772), famous mostly for his matrix. It really should be named after Zhu_Shijie who invented it much earlier in 1303.

To find the expected value of X we need the following identity:

$$\binom{n}{k} = \frac{n}{k}\frac{(n-1)!}{(n-k)!(k-1)!} = \frac{n}{k}\binom{n-1}{k-1}$$

now

$$E[X] = \sum_{k=0}^{N} k\frac{\binom{N}{k}\binom{M}{n-k}}{\binom{N+M}{n}} = \sum_{k=1}^{N} k\frac{\binom{N}{k}\binom{M}{n-k}}{\binom{N+M}{n}} =$$

$$\sum_{k=1}^{N} k\frac{\frac{N}{k}\binom{N-1}{k-1}\binom{M}{n-1-(k-1)}}{\frac{N+M}{n}\binom{N+M-1}{n-1}} =$$

$$\frac{nN}{N+M}\sum_{k=1}^{N}\frac{\binom{N-1}{k-1}\binom{M-1}{n-1-(k-1)}}{\binom{N+M-1}{n-1}} =$$

$$\frac{nN}{N+M}\sum_{j=0}^{N-1}\frac{\binom{N-1}{j}\binom{M-1}{n-1-j)}}{\binom{N-1+M}{n-1}} = \frac{nN}{N+M}$$

similarly we find

$$Var(X) = \frac{nNM}{(N+M)^2\left(1-\frac{n-1}{N+M-1}\right)}$$

Note that as the population size gets large we find if

$$\text{say } \frac{N}{N+M} \to p, \text{ then}$$

$$E[X] = \frac{nN}{N+M} \to np$$

$$Var(X) = \frac{nNM}{(N+M)^2(1-\frac{n-1}{N+M-1})} =$$

$$n\frac{N}{N+M}\frac{M}{N+M}\frac{N+M-1}{N+M-1-n+1} =$$

$$n\frac{N}{N+M}(1-\frac{N}{N+M})(\frac{N-1}{N+M-n}+1-\frac{N}{N+M-n})$$

$$\to np(1-p)(p+1-p) = npq$$

so as one would expect the mean and the variance approach those of the Binomial distribution as the population size gets large. As a ballpark one uses the hypergeometric if the sample size is more than 10% of the population size.

2.1.7.2 Example

say our company has a pool of 100 candidates for the job, 10 of whom are suitable for hiring. If they interview 50 of the 100, what is the probability that they will fill the 5 positions?

Here $X \sim HG(50, 10, 90)$ and so

$P(X \geq 5) = 1 - P(X \leq 4) = 1 - 0.3703 = 0.6297$

using the binomial distribution for our example we would have found $P(X \geq 5) = 0.5688$, quite different from the hypergeometric. On the other hand if our candidate pool had 1000 applicants, 100 of whom are suitable we would have found $P(X \geq 5) = 0.5731$.

2.1.8 Poisson Distribution

A random variable X is said to have a Poisson distribution with rate λ, ($X \sim P(\lambda)$) if

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda},\ k = 0, 1, 2, ..$$

this defines a proper pdf:

$$\sum_{k=0}^{\infty} P(X = k) =$$
$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} = e^{\lambda} e^{-\lambda} = 1$$

for the mean and variance we have

$$E[X] = \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$$

$$E[X(X-1)] = \sum_{k=0}^{\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} = \lambda^2 e^{-\lambda} \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} = \lambda^2$$
$$Var(X) = E[X(X-1)] + E[X] - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

$$\psi(t) = E[e^{tX}] = \sum_{k=0}^{\infty} e^{tk} \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=0}^{\infty} \frac{(e^t \lambda)^k}{k!} e^{-\lambda} = e^{(e^t - 1)\lambda}$$

One way to visualize the Poisson distribution is as follows:

Theorem

$X \sim$ Bin(n,p) such that n is large and p is small. That is, the number of trials is large but the success probability is small. Then X is approximately Poisson with rate $\lambda = np$.

proof 1

let $B \sim$ Bin(n,p) and $M \sim Pois(\lambda)$. A well-known result in calculus is

if $a_n \to a$ then $(1 + \frac{a_n}{n})^n \to e^a$

now

$$\psi_B(t) = (q + pe^t)^n = (1 - \frac{np}{n} + \frac{npe^t}{n})^n =$$

$$(1 + \frac{(e^t - 1)np}{n})^n \to e^{(e^t-1)\lambda} = \psi_M(t)$$

proof 2

$$P(M = x) = \frac{\lambda^x}{x!}e^{-\lambda} = \frac{\lambda}{x}\frac{\lambda^{x-1}}{(x-1)!}e^{-\lambda} = \frac{\lambda}{x}P(M = x - 1)$$

$$P(B = x) = \binom{n}{x}p^x q^{n-x} = \frac{n!}{(n-x)!x!}p^x q^{n-x} = \frac{n-x+1}{x}\frac{n!}{(n-x+1)!(x-1)!}\frac{p}{q}p^{x-1}q^{n-(x-1)}$$

$$= \frac{n-x+1}{x}\binom{n}{x-1}\frac{p}{q}p^{x-1}q^{n-(x-1)} = \frac{n-x+1}{x}\frac{p}{q}P(B = x - 1)$$

Now if $np \to \lambda$ we have

$$\frac{n-x+1}{x}\frac{p}{q} = \frac{np-p(x-1)}{x-px} \simeq \frac{\lambda}{x}$$

because p has to be small. So

$$P(B = x) = \frac{\lambda}{x}P(B = x - 1)$$

Also

$$P(B = 0) = (1-p)^n = (1 - \frac{np}{n})^n \to e^{\lambda} = P(M = 0)$$

so the approximation works for x=0, and then the recursion relationship assures that it works for all x as well.

2.1.8.1 Example

say you drive from Mayaguez to San Juan. Assume that the probability that on one kilometer of highway there is a police car checking the speed is 0.04. What is the probability that you will encounter at least 3 police cars on your drive?

If we assume that the police cars appear independently (?) then X = # of police cars $\sim Bin(180, 0.04)$, so

$P(X \geq 3) = 1 - \text{pbinom}(2, 180, 0.04) =$
$1 - 0.0234 = 0.9766$

One the other hand X is also approximately P(180*0.04) = P(7.2) and so

$P(X \geq 3) = 1 - \text{ppois}(2, 7.2) = 1 - 0.0254 = 0.9746$

The main questions with approximations are always:

1) how good is it?

2) when does it work?

Here is another connection between the Poisson and the Binomial Distribution. To proof it we first need

Theorem

Let $X \sim Pois(\lambda)$, $Y \sim Pois(\mu)$ and X and Y independent. Then $X + Y \sim Pois(\lambda + \mu)$.

proof

using moment generating functions we have

$$\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t) =$$

$$exp\{(e^t - 1)\lambda\}exp\{(e^t - 1)\mu\} =$$

$$exp\{(e^t - 1)(\lambda + \mu)\}$$

Now:

Theorem Let $X \sim Pois(\lambda), Y \sim Pois(\mu)$ and X and Y independent. Then

$X|X + Y = n \sim Bin(n, \lambda/(\lambda + \mu))$

proof:

$$P(X = k|X + Y = n) =$$

$$\frac{P(X=k,X+Y=n)}{P(X+Y=n)} =$$

$$\frac{P(X=k,Y=n-k)}{P(X+Y=n)} =$$

$$\frac{P(X=k)P(Y=n-k)}{P(X+Y=n)} =$$

$$\frac{\frac{\lambda^k}{k!}e^{-\lambda}\frac{\mu^{n-k}}{(n-k)!}e^{-\mu}}{\frac{(\lambda+\mu)^n}{n!}e^{-(\lambda+\mu)}} =$$

$$\frac{n!}{(n-k)!k!}\frac{\lambda^k\mu^{n-k}}{(\lambda+\mu)^n} =$$

$$\binom{n}{k}\frac{\lambda^k\mu^{n-k}}{(\lambda+\mu)^k(\lambda+\mu)^{n-k}} =$$

$$\binom{n}{k}\left(\frac{\lambda}{\lambda+\mu}\right)^k\left(\frac{\mu}{\lambda+\mu}\right)^{n-k} =$$

$$\binom{n}{k}\left(\frac{\lambda}{\lambda+\mu}\right)^k\left(1-\frac{\lambda}{\lambda+\mu}\right)^{n-k}$$

2.1.8.2 Example

say the number of men and women that come into a store during one day have Poisson distributions with rates 20 and 50, respectively. If a total of 100 came to the store today, what is the probability that at most 25 were men?

$$P(X \leq 25|X+Y=100) = \sum_{i=0}^{25} \binom{100}{i} \left(\frac{20}{70}\right)^i \left(1-\frac{20}{70}\right)^{100-i} = 0.2511$$

Of course the assumption that men and women come to the store independently is probably questionable!

2.1.9 Multinomial Distribution

Let p_1 ,..,p_n be numbers with $0 \leq p_i \leq 1$ and $\sum p_i = 1$. Then the rv (X_1 ,..,X_n) has a multinomial distribution with m trials if

$$P(X_1 = x_1,..,X_n = x_n) = m! \prod_{i=1}^{n} \frac{p_i^{x_i}}{x_i!}$$

$$\text{if } x_i \in \{0,1,2,..\} \text{ and } \sum_i^n x_i = m$$

We write $(X_1,..,X_n) \sim M(m,p_1,..,p_n)$

That this defines a proper pdf follows from the multinomial theorem for $1 = (p_1 + .. + p_n)^m$

2.1.9.1 Example

we roll a fair die 100 times. Let X_1 be the number of "1"s, X_2 be the number of "2"s,.., X_6 be the number of "6"s. Then

$(X_1,..,X_6) \sim M(100, 1/6,..,1/6)$

Note: if n=2 we have $x_1 + x_2 = m$, or $x_2 = m - x_1$ and $p_1 + p_2 = 1$, so

$$P(X_1 = x_1, X_2 = x_2) = m! \prod_{i=1}^{2} \frac{p_i^{x_i}}{x_i!} =$$

$$m! \frac{p_1^{x_1}}{x_1!} \frac{p_2^{x_2}}{x_2!} = m! \frac{p_1^{x_1}}{x_1!} \frac{(1-p_1)^{m-x_1}}{(m-x_1)!} = \binom{m}{x_1} p_1^{x_1} (1-p_1)^{m-x_1}$$

and so $X_1 \sim Bin(m, p_1)$. The multinomial distribution is therefore a generalization of the binomial distribution where each trial has n possible outcomes.

Theorem

Let $(X_1, .., X_n) \sim M(m, p_1, .., p_n)$. Then the marginal distribution of X~k‘ is Bin(m,p_k)

proof:

let's denote by

$B_x = \{(x_1, .., x_{k-1}, x_{k+1}, .., x_n) : x_1 + .. + x_{k-1} + x_{k+1} + .. + x_n = m - x\}$

then:

$$f_{X_k}(x) = \sum_{B_x} f(x_1, .., x_{k-1}, x, x_{k+1}, .., x_n) = \sum_{B_x} m! \frac{p_k^x}{x!} \prod_{i=1, i\neq k}^{n} \frac{p_i^{x_i}}{x_i!} =$$

$$\sum_{B_x} m! \frac{p_k^x}{x!} \prod_{i=1, i\neq k}^{n} \frac{p_i^{x_i}}{x_i!} \frac{(m-x)!}{(m-x)!} \frac{(1-p_k)^{m-x}}{(1-p_k)^{m-x}} =$$

$$\frac{m!}{x!(m-x)!} p_k^x (1-p_k)^{m-x} \sum_{B_x} \frac{(m-x)!}{x_1! .. x_{k-1}! x_{k+1}! .. x_n!} \left(\frac{p_i}{1-p_k} \right)^{x_i} = \binom{m}{x} p_k^x (1-p_k)^{m-x}$$

where the sum is 1 because we summing over all possible values of a multinomial rv $(Y_1, .., Y_{n-1}) \sim M(m - x, p_1/(1 - p_k), .., p_n/(1 - p_k)$, or because we use the multinomial theorem from calculus.

From this it follows that E[X_k]=mp_k and Var(X_k)=mp_k (1-p_k)

Next we will find the moment generating function of a multinomial. For this we need a generalization of the mgf for random vectors:

$\psi(t_1, .., t_n) = E[\exp(t_1 X_1 + .. + t_n X_n)]$

then

$$\psi(t_1,..,t_n) = E[e^{\sum_{i=1}^{n} t_i X_i}] =$$

$$\sum_I e^{\sum_{i=1}^{n} t_i X_i} m! \prod_{i=1}^{n} p_i^{x_i}/x_i! =$$

$$\sum_I m! \prod_{i=1}^{n} (e^{t_i} p_i)^{x_i}/x_i! = \left(\sum_{i=1}^{n} p_i e^{t_i}\right)^n$$

Theorem

Let $(X_1,..,X_n) \sim M(m, p_1,..,p_n)$. Then the conditional distribution of $(X_1,..,X_n)|X_k = x \sim M(m-x, p_1/(1-p_k),..,p_n/(1-p_k))$

proof

$$P(X_i = x_i, i = 1,..,n|X_k = x) = \frac{P(X_i=x_i;i=1,..,n)}{P(X_k=x)} =$$

$$\frac{m! \prod_{i=1}^{n} p_i^{x_i}/x_i!}{\binom{m}{x} p_k^x (1-p_k)^{n-x}} = \frac{m! \prod_{i=1}^{n} p_i^{x_i}/x_i!}{\frac{m!}{(m-x)!x!} p_k^x (1-p_k)^{n-x}} =$$

$$(m-x)! \left(\prod_{i=1, i\neq k}^{n-x} p_i^{x_i}/x_i!\right) \Big/ \left(\prod_{i=1, i\neq k}^{n-x} (1-p_k)^{x_i}\right) =$$

$$(m-x)! \prod_{i=1, i\neq k}^{n-x} \left(\frac{p_i}{1-p_k}\right)^{x_i}/x_i!$$

Theorem

Let $(X_1, .., X_n) \sim M(m, p_1, .., p_n)$. Then

$Cov(X_i, X_j) = -mp_ip_j$

proof

$$E[X_iX_j] = \frac{d^2}{dx_idx_j}\psi(t$$

$$\frac{d^2}{dx_idx_j}\left(\sum_{i=1}^{n} p_ie^{t_i}\right)^n$$

$$\frac{d}{dx_i}n\left(\sum_{i=1}^{n} p_ie^{t_i}\right)^{n-}$$

$$n(n-1)\left(\sum_{i=1}^{n} p_ie^{t_i}\right.$$

$$n(n-1)p_ip_j$$

$$Cov(X_i, X_j) = E[X_i$$

$$n(n-1)p_ip_j - np_in$$

we will use the mgf. Note that one can show for a random vector that

the fact that the covariance is always negative makes sense because if X_1 is larger X_j is likely to be smaller as the sum of the X_k 's has to be n.

Calculating the correlation we get

$$\rho = Cor(X_i, X_j) = \frac{Cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}} =$$

$$\frac{-np_ip_j}{\sqrt{np_i(1-p_i)np_j(1-p_j)}} = -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}$$

and the somewhat surprising fact that the correlation does not depend on n!

2.2 Continuous Distributions

2.2.1 Uniform Distribution

X is said to have a uniform distribution on the interval [A,B] if

$$X \sim U[A, B]$$

$$f(x) = \frac{1}{B-A}, \quad A < x < B, \quad 0 \text{ otherwise}$$

$$E[X] = \frac{A+B}{2}, \quad Var[X] = \frac{(B-A)^2}{12}$$

2.2.2 Exponential Distribution

X is said to have an exponential distribution rate λ if

$f(x) = \lambda e^{-\lambda x}, x > 0, \lambda > 0$

we write $X \sim Exp(\lambda)$

Note

$$F_X(x) = \int_0^x \lambda e^{-\lambda t} dt = \int_0^x e^{-\lambda t}(\lambda dt) = \int_0^{x/\lambda} e^{-y} dy = F_Y(x/\lambda)$$

where $Y \sim Exp(1)$

so if $Y \sim Exp(1)$ then $X = Y/\lambda \sim Exp(\lambda)$

$$E[Y^k] = \int_0^\infty y^k e^{-y} dy =$$

$$y^k(-e^{-y})|_0^\infty - \int_0^\infty k y^{k-1}(-e^{-y}) dy =$$

$$k\int_0^\infty y^{k-1} e^{-y} dy = kE[Y^{k-1}]$$

$$E[Y] = 1E[Y^0] = 1$$

$$E[Y^2] = 2E[Y^1] = 2 \cdot 1 = 2$$

$$Var(Y) = E[Y^2] - (E[Y])^2 = 2 - 1^2 = 1$$

$$E[X] = E[\tfrac{Y}{\lambda}] = \tfrac{1}{\lambda}$$

$$Var(X) = Var(\tfrac{Y}{\lambda}) = \tfrac{1}{\lambda^2}$$

The "trick" we used here, namely showing a result for a special case ($\lambda = 1$) and then doing the general case, is often a good idea! Let's use it again to find the moment generating function:

$$\psi_Y(t) = E[e^{tY}] = \int_0^\infty e^{ty}e^{-y}dy = \int_0^\infty e^{-(1-t)y}dy =$$

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

if $1-t > 0$ or $t < 1$

$$\psi_X(t) = E[e^{tX}] = E[e^{tY/\lambda}] = \psi_Y(t/\lambda) =$$

$$\frac{1}{1-t/\lambda} = \frac{\lambda}{\lambda-t} \text{ if } t < \lambda$$

We have previously talked about the memoryless property, and the fact that among discrete distributions on $\mathbb{N}$ it is unique to the geometric rv. Now we have

Theorem

X has an exponential distribution iff X is a positive continuous r.v. and

$P(X > s+t|X > s) = P(X > t) forall s, t > 0$

proof:

Assume $X \sim Exp(\lambda)$. Then

$$P(X > x) = 1 - P(X < x) = 1 - \int_0^x \lambda e^{-\lambda t}dt =$$

$$1 - (-e^{\lambda t}|_0^x = 1 - (-e^{-\lambda x} - (-1)) = e^{-\lambda x}$$

Now $P(X > s+t|X > s) = \frac{P(X>s+t \cap X>s)}{P(X>s)} = \frac{P(X>s+t)}{P(X>s)} =$

$$\frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X > t)$$

because $\{X > s+t\} \subset \{X > s\}$

on the other hand assume X is continuous with density f and

$P(X > s + t|X > s) = P(X > t) for all s, t > 0$

Above we saw that this implies

$P(X > s + t) = P(X > s) * P(X > t)$

Let h(x) = P(X>x) and let $\epsilon > 0$. Note h(0) = P(X>0) = 1 because X is positive.

$$h'(0) = \lim_{\varepsilon \to 0}\left(\frac{h(\varepsilon)-h(0)}{\varepsilon}\right) = \lim_{\varepsilon \to 0}\left(\frac{P(X>\varepsilon)-1}{\varepsilon}\right) =$$

$$\lim_{\varepsilon \to 0}\left(\frac{1-P(X<\varepsilon)-1}{\varepsilon}\right) = \lim_{\varepsilon \to 0}\left(\frac{-P(X<\varepsilon)}{\varepsilon}\right) =$$

$$-\lim_{\varepsilon \to 0}\left(\frac{1}{\varepsilon}\int_0^{\varepsilon} f(x)dx\right) = -f(0) \doteq -\beta$$

$$h(x+\varepsilon) - h(x) = P(X > x+\varepsilon) - P(X > x) =$$

$$P(X > x)P(X > \varepsilon) - P(X > x) = P(X > x)(P(X > \varepsilon) - 1)$$

$$h'(x) = \lim_{\varepsilon \to 0}\left(\frac{h(x+\varepsilon)-h(x)}{\varepsilon}\right) = \lim_{\varepsilon \to 0}\left(\frac{h(x)(h(\varepsilon)-1))}{\varepsilon}\right) =$$

$$h(x)h'(0) = -\beta h(x)$$

$$\frac{h'(x)}{h(x)} = -\beta, \quad \int \frac{h'(x)}{h(x)}dx = \int -\beta dx = -\beta x + c$$

$$\log h(x) = -\beta x + c$$

$$\log h(0) = 0 = -\beta 0 + c$$

$$h(x) = e^{-\beta x} = P(X > x) = 1 - P(X < x)$$

$$F(x) = 1 - e^{-\beta x}$$

and so we see $X \sim Exp(\beta)$

2.2.3 The Gamma Distribution

Recall the Gamma function:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt, \;\; \alpha > 0$$

The Gamma function is famous for many things, among them the relationship

$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$

which follows from:

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt = t^{\alpha-1}(-e^{-t})|_0^\infty - \int_0^\infty (\alpha-1) t^{\alpha-2}(-e^{-t}) dt = (\alpha-1)\int_0^\infty t^{\alpha-2} e^{-t} dt = (\alpha-1)\Gamma(\alpha-1)$$

This implies

$\Gamma(n) = (n-1)!$

so the Gamma function is a continuous version of the factorial. It has many other interesting properties, for example

$\Gamma(1/2) = \sqrt{\pi}$

Now X is said have a *Gamma distribution* ($X \sim \Gamma(\alpha, \beta)$) with parameters (α, β) if

$$f(x) = \frac{1}{\Gamma(a)\beta^a} x^{\alpha-1} e^{-x/\beta}, x, \alpha, \beta > 0$$

$$F_X(x) = P(X \le x) = \int_0^x \frac{1}{\Gamma(a)\beta^a} t^{\alpha-1} e^{-t/\beta} dt =$$

$$\int_0^x \frac{1}{\Gamma(a)} \left(\frac{t}{\beta}\right)^{\alpha-1} e^{-t/\beta} \left(\frac{1}{\beta} dt\right) = \int_0^{\beta x} \frac{1}{\Gamma(a)} y^{\alpha-1} e^{-y} dy = P(Y \le \beta x)$$

where $Y \sim \Gamma(\alpha, 1)$

so if $Y \sim \Gamma(\alpha, 1)$ $X = \beta Y \sim \Gamma(\alpha, \beta)$

$$E[Y^k] = \int_0^\infty y^k \frac{1}{\Gamma(a)} y^{\alpha-1} e^{-y} dy =$$

$$\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} \int_0^\infty \frac{1}{\Gamma(a+k)} y^{\alpha+k-1} e^{-y} dy = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)} = \frac{(\alpha+k-1)\cdot\ldots\cdot\alpha\Gamma(\alpha)}{\Gamma(\alpha)} = (\alpha + k - 1) \cdot \ldots \cdot \alpha$$

$$E[Y] = \alpha$$

$$Var(Y) = E[Y^2] - E[Y]^2 = (\alpha + 1)\alpha - \alpha^2 = \alpha$$

$$E[X] = E[\beta Y] = \alpha\beta$$

$$Var(X) = Var(\beta Y) = \alpha\beta^2$$

$$\psi_Y(t) = \int_0^\infty e^{ty} \frac{1}{\Gamma(a)} y^{\alpha-1} e^{-y} dy = \int_0^\infty \frac{1}{\Gamma(a)} y^{\alpha-1} e^{-(1-t)y} dy =$$

$$\left(\frac{1}{1-t}\right)^\alpha \int_0^\infty \frac{1}{\Gamma(a)\left(\frac{1}{1-t}\right)^\alpha} y^{\alpha-1} e^{-y/\left(\frac{1}{1-t}\right)} dy = \left(\frac{1}{1-t}\right)^\alpha \text{ if } t < 1$$

$$\psi_X(t) = E[e^{tX}] = E[e^{t\beta Y}] = \psi_Y(\beta t) = \left(\frac{1}{1-\beta t}\right)^\alpha \text{ if } t < \frac{1}{\beta}$$

By definition we have X>0, and so the Gamma is the basic example of a r.v. on $(0, \infty)$, or a little more general (using a change of variables) on any open half interval.

Note if $X \sim \Gamma(1, \beta)$ then $X \sim E(1/\beta)$.

Another important special case is if $X \sim \Gamma(n/2, 2)$, then X is called a Chi-square r.v. with n degrees of freedom, denoted by $X \sim \chi^2(n)$.

2.2.3.1 Example

Find the kurtosis of X where $X \sim Exp(\lambda)$

$$E[X^k] = \int_0^\infty x^k \lambda e^{-\lambda x} dx =$$

$$\lambda\Gamma(k+1)(1/\lambda)^{k+1} \int_0^\infty \frac{1}{\Gamma(k+1)(1/\lambda)^{k+1}} x^{(k+1)-1} e^{-x/(1/\lambda)} dx =$$

$$\lambda\Gamma(k+1)(1/\lambda)^{k+1} = \frac{k!}{\lambda^k}$$

$$\mu = E[X] = \frac{1}{\lambda}$$

$$E[X^2] = \frac{2}{\lambda^2} \quad E[X^3] = \frac{6}{\lambda^3} \quad E[X^4] = \frac{24}{\lambda^4}$$

$$\kappa_2 = E[(X-\mu)^2] = E[X^2] - \mu^2 =$$

$$\frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

$$\kappa_4 = E[(X-\mu)^4] =$$

$$E[X^4 - 4X^3\mu + 6X^2\mu^2 - 4X\mu^3 + \mu^4] =$$

$$\frac{24}{\lambda^4} - 4\frac{6}{\lambda^3}\frac{1}{\lambda} + 6\frac{2}{\lambda^2}\frac{1}{\lambda^2} - 3\frac{1}{\lambda^4}$$

$$\frac{1}{\lambda^4}[24 - 24 + 12 - 3] = \frac{9}{\lambda^4}$$

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} - 3 = \frac{9/\lambda^4}{(1/\lambda^2)^2} - 3 = 6$$

so the kurtosis is greater than 0, therefore an exponential is leptocurtic.

There is an important connection between the Gamma and the Poisson distributions:

Theorem

if $X \sim \Gamma(n, \beta)$ and $Y \sim Pois(x/\beta)$ then

$P(X \leq x) = P(Y \geq n)$

proof (by induction)

$$n = 1 : X_1 \sim \Gamma(1, \beta) = Exp(1/\beta)$$

$$P(X_1 \leq x) = \int_0^x \frac{1}{\beta} e^{-t/\beta} dt =$$

$$-e^{-t/\beta}|_0^x = 1 - e^{-x/\beta} = 1 - \frac{(x/\beta)^0}{0!} e^{-x/\beta} =$$

$$1 - P(Y = 0) = P(Y \geq 1)$$

assume statment is true for n

$$X_{n+1} \sim \Gamma(n + 1, \beta)$$

$$P(X_{n+1} \leq x) = \int_0^x \frac{1}{\Gamma(n+1)\beta^{n+1}} t^{(n+1)-1} e^{-t/\beta} dt =$$

$$\frac{1}{n!\beta^{n+1}} \int_0^x t^n e^{-t/\beta} dt =$$

$$\frac{1}{n!\beta^{n+1}} \left[t^n(-\beta)e^{-t/\beta}|_0^x - \int_0^x nt^{n-1}(-\beta)e^{-t/\beta} dt \right] =$$

$$\frac{1}{n!\beta^{n+1}} \left[-x^n \beta e^{-x/\beta} + n\beta \int_0^x t^{n-1} e^{-t/\beta} dt \right] =$$

$$-\frac{(x/\beta)^n}{n!} e^{-x/\beta} + \int_0^x \frac{1}{(n-1)!\beta^n} t^{n-1} e^{-t/\beta} dt =$$

$$-P(Y = n) + P(X_n \leq x) =$$

$$-P(Y = n) + P(Y \geq n) =$$

$$P(Y \geq n + 1)$$

2.2.4 The Beta Distribution

X is said to have a Beta distribution with parameters α and β ($X \sim Beta(\alpha, \beta)$) if

$$f(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad 0 < x < 1$$

$$\Gamma(x)\Gamma(y) =$$

$$\left(\int_0^\infty u^{x-1}e^{-u}du\right)\left(\int_0^\infty v^{y-1}e^{-v}dv\right) =$$

$$\int_0^\infty\int_0^\infty u^{x-1}v^{y-1}e^{-(u+v)}dudv$$

$$z = u+v,\ \ t = \frac{u}{u+v}$$

$$u = zt,\ \ v = z(1-t)$$

$$J = \begin{vmatrix} t & z \\ 1-t & -z \end{vmatrix} = -zt - z(1-t) = z$$

$$\Gamma(x)\Gamma(y) =$$

$$\int_0^\infty\int_0^1 (zt)^{x-1}(z(1-t))^{y-1}e^{-z}zdtdz =$$

$$\int_0^\infty\left(\int_0^1 t^{x-1}(1-t)^{y-1}dt\right)z^{x+y-1}e^{-z}dz =$$

$$\left(\int_0^1 t^{x-1}(1-t)^{y-1}dt\right)\left(\int_0^\infty z^{x+y-1}e^{-z}dz\right) =$$

$$\left(\int_0^1 t^{x-1}(1-t)^{y-1}dt\right)\Gamma(x+y)$$

$$\int_0^1 t^{x-1}(1-t)^{y-1}dt = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$$

it is easy to calculate the moments of a Beta distribution:

$$E[X^k] = \int_0^1 x^k \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\,dx =$$

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 x^{\alpha+k-1}(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha+\beta+k)} \int_0^1 \frac{\Gamma(\alpha+k+\beta)}{\Gamma(\alpha+k)\Gamma(\beta)} x^{\alpha+k-1}(1-x)^{\beta-1}dx =$$

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha+\beta+k)}$$

so

$$E[X] = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{\Gamma(\alpha+\beta+1)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\alpha\Gamma(\alpha)}{(\alpha+\beta)\Gamma(\alpha+\beta)} = \frac{\alpha}{\alpha+\beta}$$

$$E[X^2] = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+2)}{\Gamma(\alpha+\beta+2)} = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{(\alpha+1)\alpha\Gamma(\alpha)}{(\alpha+\beta+1)(\alpha+\beta))\Gamma(\alpha+\beta)} = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)}$$

$$V[X] = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2 = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

The mgf is given by

$$E[X^k] = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+k)}{\Gamma(\alpha+\beta+k)} =$$

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)} \frac{(\alpha+k-1)..\alpha\Gamma(\alpha)}{(\alpha+\beta+k-1)..(\alpha+\beta)\Gamma(\alpha+\beta)}$$

$$\frac{(\alpha+k-1)..\alpha}{(\alpha+\beta+k-1)..(\alpha+\beta)} = \prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r}$$

$$\psi(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f(x)dx =$$

$$\int_{-\infty}^{\infty} \left(\sum_{k=0}^{\infty} \frac{(tx)^k}{k!}\right) f(x)dx =$$

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} \int_{-\infty}^{\infty} x^k f(x)dx =$$

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} E[X^k] =$$

$$1 + \sum_{k=1}^{\infty} \prod_{r=0}^{k-1} \frac{\alpha+r}{\alpha+\beta+r} \frac{t^k}{k!}$$

Note that the Taylor expansion of the moment generating function is completely general.

By definition we have 0<X<1, and so the Beta is the basic example of a r.v. on [0,1], or a little more general (using a change of variables) on any open finite interval.

Special cases:

1. Beta(1,1) = U[0,1]
2. $X \sim Beta(p, 1)$ then

$f(x) = cx^{p-1}(1-p)^1 = cx^{p-1} = px^{p-1}, 0 < x < 1, p > 0$

and for this pdf we have E[X]=p/(p+1), Var[X]=p/[(p+1)^2(p+2)]

Let's go back to the Gamma distribution for a moment. Say X and Y are independent $\Gamma(\alpha, \beta)$ and let Z=X+Y. Then

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

if $t > 0$ and $z - t > 0 \Rightarrow t < z$, so

$$f_Z(z) = \int_0^z \frac{1}{\Gamma(\alpha)\beta^\alpha} t^{\alpha-1} e^{-t/\beta} \frac{1}{\Gamma(\alpha)\beta^\alpha} (z-t)^{\alpha-1} e^{-(z-t)/\beta} dt =$$

$$\left(\frac{1}{\Gamma(\alpha)\beta^\alpha}\right)^2 e^{-z/\beta} \int_0^z t^{\alpha-1}(z-t)^{\alpha-1} dt = \quad \text{change of variables } x = t/z$$

$$\left(\frac{1}{\Gamma(\alpha)\beta^\alpha}\right)^2 e^{-z/\beta} \int_0^1 (zx)^{\alpha-1}(z-(zx))^{\alpha-1}(z dx) =$$

$$\left(\frac{1}{\Gamma(\alpha)\beta^\alpha}\right)^2 z^{2\alpha-1} e^{-z/\beta} \int_0^1 x^{\alpha-1}(1-x)^{\alpha-1} dx =$$

$$\left(\frac{1}{\Gamma(\alpha)\beta^\alpha}\right)^2 z^{2\alpha-1} e^{-z/\beta} \frac{\Gamma(\alpha)^2}{\Gamma(2\alpha)} \cdot \int_0^1 \frac{\Gamma(\alpha+\alpha)}{\Gamma(\alpha)\Gamma(\alpha)} x^{\alpha-1}(1-x)^{\alpha-1} dx =$$

$$\left(\frac{1}{\Gamma(\alpha)\beta^\alpha}\right)^2 z^{2\alpha-1} e^{-z/\beta} \frac{\Gamma(\alpha)^2}{\Gamma(2\alpha)} = \frac{1}{\Gamma(2\alpha)\beta^{2\alpha}} z^{2\alpha-1} e^{-z/\beta}$$

for $z > 0$

so we see that $Z \sim \Gamma(2\alpha, \beta)$. In other words, the sum of independent Gamma r.v.'s is again Gamma.

Some special cases:

1. X,Y iid Exp(λ) then $X + Y \sim \Gamma(2, \lambda)$ (and **not** exponential)
2. $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$ and $X \perp Y$, then $X + Y \sim \chi^2(n + m)$.

Previously we found a curious relationship between the Poisson and the gamma distributions. There is a similar one between the Beta and the Binomial:

Theorem

if $X \sim Beta(n, m)$ and $Y \sim Bin(n + m - 1, 1 - x)$ then

$P(X \leq x) = P(Y < m)$

$$X_m \sim Beta(n, m),\ Y \sim Bin(n + m - 1, 1 - x)$$

$$P(X_m \leq x) = P(Y < m)$$

$m = 1$:

$$P(X_1 \leq x) = \int_0^x \frac{\Gamma(n+1)}{\Gamma(n)\Gamma(1)} t^{n-1}(1-t)^{1-1}dt = \int_0^x nt^{n-}$$

$$t^n|_0^x = x^n = \binom{n}{0}(1-x)^0(1-(1-x))^n =$$

$$P(Y = 0) = P(Y < 1)$$

assume statement true for m

$$P(X_{m+1} \leq x) = \int_0^x \frac{\Gamma(n+m+1)}{\Gamma(n)\Gamma(m+1)} t^{n-1}(1-t)^{m+1-1}dt =$$

$$\frac{(n+m)!}{(n-1)!m!}\left[\frac{1}{n}t^n(1-t)^m|_0^x - \int_0^x \frac{1}{n}t^n m(1-t)^{m-1}(-1)d\right.$$

$$\frac{(n+m)!}{n!m!}x^n(1-x)^m + \int_0^x \frac{(n+m)!}{n!(m-1)!}t^n(1-t)^{m-1}dt =$$

$$\binom{n+m}{m}(1-x)^m(1-(1-x))^{(n+m)-m} + P(X_m \leq x)$$

$$P(Y = m) + P(X_m \leq x) =$$

$$P(Y = m) + P(Y < m) = P(Y \leq m) = P(Y < m$$

proof (by induction on m)

2.2.5 Cauchy Distribution

A rv. X has a Cauchy distribution if

$$f(x;\theta) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}, \quad -\infty < x < \infty$$

As we saw before the Cauchy has one interesting property:

$E[|X|] = \infty$

so the Cauchy has no mean (and therefore no moments at all). The reason is that it has thick "tails", that is the probability of observing a large value (+ or -) is large.

2.3 The Normal (Gaussian) Distribution

2.3.1 Normal Distribution

X is said to have a normal distribution with mean μ and variance σˆ2 ($X \sim N(\mu, \sigma)$) if it has density

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

If $\mu = 0$ and $\sigma = 1$ it is called a standard normal, and often denoted by Z instead of X.

Careful: some papers and textbooks define the normal as $X \sim N(\mu, \sigma^2)$, that is they use the variance instead of the standard deviation.

Theorem

a. $Z \sim N(0,1)$ then $X = \mu + \sigma Z \sim N(\mu, \sigma)$

b. $X \sim N(\mu, \sigma)$, then $Z = (x-\mu)/\sigma \sim N(0,1)$

proof

a) $Z \sim N(0,1), \quad X = \mu + \sigma Z$

$$P(X \le x) = P(\mu + \sigma Z \le x) = P(Z \le \tfrac{x-\mu}{\sigma}) =$$

$$\int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}t^2\right) dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\tfrac{1}{2}\frac{(u-\mu)^2}{\sigma^2}\right) du$$

b) $X \sim N(\mu, \sigma), \quad Z = \frac{X-\mu}{\sigma}$

$$P(Z \le x) = P(X \le \mu + \sigma x) = \int_{-\infty}^{\mu+\sigma x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\tfrac{1}{2}\frac{(t-\mu)^2}{\sigma^2}\right) dt =$$

$$\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\tfrac{1}{2}u^2\right) du$$

one consequence of this theorem is that we can often do a proof for the standard normal, and then quickly generalize it to all normals.

Theorem

show that the function above is indeed a pdf for all μ and $\sigma > 0$.

proof

- $f(x) \geq 0$ for all x
- first we show this for a standard normal:

First note $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right)dz = 1$ iff $\int_0^{\infty} \exp\left(-\frac{1}{2}z^2\right)dz = \sqrt{\frac{\pi}{2}}$

now $\left[\int_0^{\infty} \exp\left(-\frac{1}{2}z^2\right)dz\right]^2 = \left(\int_0^{\infty} \exp\left(-\frac{1}{2}t^2\right)dt\right) \cdot \left(\int_0^{\infty} \exp\left(-\frac{1}{2}u^2\right)du\right) =$
$\int_0^{\infty}\int_0^{\infty} \exp\left(-\frac{1}{2}(t^2+u^2\right)dtdu$

change of variables: $t = r\cos\theta$ and $u = r\sin\theta$

then $t^2 + u^2 = r^2$ and $dtdu = rd\theta dr$, so

$\left[\int_0^{\infty} \exp\left(-\frac{1}{2}z^2\right)dz\right]^2 = \int_0^{\infty}\int_0^{\pi/2} re^{-r^2/2}d\theta dr = \int_0^{\infty} \frac{\pi}{2} re^{-r^2/2}d\theta dr =$
$\frac{\pi}{2} - e^{-r^2/2}|_0^{\infty} = \frac{\pi}{2}$

the change of variables above is of course called the change to *polar coordinates*.

the general case now follows easily:

$$P(-\infty < X < \infty) =$$
$$P(-\infty < (X-\mu)/\sigma < \infty) =$$
$$P(-\infty < Z < \infty) = 1$$

2.3.1.1 Example

we previously said that $\Gamma(1/2) = \sqrt{\pi}$. Here is a proof that uses the standard normal distribution:

$$\Gamma(\tfrac{1}{2}) = \int_0^\infty t^{1/2-1}e^{-t}dt = \int_0^\infty \frac{1}{\sqrt{t}}e^{-t}dt =$$

$$\sqrt{2}\int_0^\infty e^{-t}(\frac{1}{\sqrt{2t}}dt) = (x = \sqrt{2t}, dx = \frac{1}{\sqrt{2t}}dt)$$

$$\sqrt{2}\int_0^\infty e^{-x^2/2}dx = \sqrt{2}\,\frac{1}{2}\int_{-\infty}^\infty e^{-x^2/2}dx =$$

$$\sqrt{\pi}\int_{-\infty}^\infty \frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = \sqrt{\pi}$$

Theorem

Say $X \sim N(\mu, \sigma)$ then

1. $E[X] = \mu$ and $Var(X) = \sigma^2$
2. $\psi(t) = \exp(\mu t + \sigma^2 t^2/2)$
3) $P(X > \mu) = P(X < \mu) = 1/2$ and $P(X > \mu + x) = P(X < \mu - x)$
4)

$$P(|X-\mu| < \sigma) = 0.6826$$

$$P(|X-\mu| < 2\sigma) = 0.9544$$

$$P(|X-\mu| < 3\sigma) = 0.9974$$

proof

1.

$E[Z] = \int_{-\infty}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = -\frac{1}{\sqrt{2\pi}} e^{-z^2/2}|_{-\infty}^{\infty} = 0$

$V[Z] = E[Z^2] - E[Z]^2 = E[Z^2] = \int_{-\infty}^{\infty} z^2 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = (z \cdot -\frac{1}{\sqrt{2\pi}} e^{-z^2/2}|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} -\frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1$

and the general case:

$E[X] = E[\mu + \sigma Z] = \mu$ and $V[X] = V[\mu + \sigma Z] = \sigma^2$

2.

Let Z be a standard normal, then

$$\psi_Z(t) = E[e^{tZ}] = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(x^2 - 2tx)\} dx =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(x^2 - 2tx + t^2 - t^2)\} dx =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(x - t)^2\} \exp\{t^2/2\} dx =$$

$$\exp\{t^2/2\}$$

now let $X \sim N(\mu, \sigma)$, then

$$\begin{aligned}
&\psi_X(t) = \\
&E[\exp(tX)] = \\
&E[\exp(t(\mu + \sigma Z)] = \\
&E[\exp(\mu t + \sigma t Z)] = \\
&E[\exp(\mu t) \exp(\sigma t Z)] = \\
&\exp(\mu t) E[\exp(\sigma t Z)] = \\
&\exp(\mu t) \psi_Z(\sigma t) = \\
&\exp(\mu t) \exp((\sigma t)^2/2) = \\
&\exp(\mu t + \sigma^2 t^2/2)
\end{aligned}$$

3. first we show $P(X > \mu + x) = P(X < \mu - x)$:

$$P(X > \mu + x) =$$

$$\int_{\mu+x}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\} dt = \quad (s = 2\mu - t; ds = -ds)$$

$$\int_{\mu-x}^{-\infty} \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(\mu-s)^2}{2\sigma^2}\right\} (-ds) =$$

$$\int_{-\infty}^{\mu-x} \frac{1}{\sqrt{2\pi\sigma}} \exp\left\{-\frac{(s-\mu)^2}{2\sigma^2}\right\} ds =$$

$$P(X < \mu - x)$$

now with x=0 it follows that

$P(X > \mu) = P(X < \mu) = 1 - P(X > \mu)$, and so $P(X > \mu) = 1/2$.

4.

$$P(|X-\mu|< x\sigma) =$$

$$P(-x < \frac{X-\mu}{\sigma} < x) =$$

$$P(-x < Z < x) =$$

$$P(Z < x) - P(Z < -x) =$$

$$P(Z < x) - P(Z > x) =$$

$$P(Z < x) - (1 - P(Z < x)) \quad =$$

$$2P(Z < x) - 1$$

$$P(Z < x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}t^2\}dt$$

and this last integral requires numerical integration because the cdf of a normal does not exist explicitly. At least we have shown that the probabilities are independent of μ and σ.

2.3.1.2 Example

We have seen before that the Cauchy rv. has very thick tails, that is the probabilities P(X>t) are large. On the other hand the normal distribution has very thin tails. There is also a distribution that is somewhat in between, called the t distribution with n degrees of freedom. It has density

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\Gamma(\frac{n}{2})} \frac{1}{\sqrt{\pi n}} \frac{1}{(1 + t^2/n)^{(n+1)/2}}$$

For n=1 this is the Cauchy, as $n \to \infty$ it approaches the standard normal

Theorem

Say $X \sim N(\mu, \sigma), Y \sim N(\nu, \tau)$ and X and Y are independent. Then X+Y and X-Y are also normal.

proof

$$\begin{aligned}
&\psi_{X+Y}(t) = \\
&\psi_X(t)\psi_Y(t) = \\
&\exp(\mu t + \sigma^2 t^2/2)\exp(\nu t + \tau^2 t^2/2) = \\
&\exp((\mu + \nu)t + (\sigma^2 + \tau^2)t^2/2)
\end{aligned}$$

and so $X + Y \sim N(\mu + \nu, \sigma^2 + \tau^2)$

$$\begin{aligned}
&\psi_{-Y}(t) = \\
&E[\exp(t[-Y])] = \\
&E[\exp((-t)Y)] = \\
&\exp(\nu(-t) + \tau^2(-t)^2/2) = \\
&\exp(-\nu t + \tau^2 t^2/2) =
\end{aligned}$$

and so $-Y \sim N(-\nu, \tau^2)$

finally $X - Y \sim N(\mu - \nu, \sigma^2 + \tau^2)$

‘

Because of the importance of the normal distribution a number of theorems have been found to characterize it. Here is one such result:

Theorem (Bernstein)

If $X \perp Y$ and $X + Y \perp X - Y$, then X and Y are normal.

proof

We will do this proof through a couple of lemmas:

Lemma

If X and Y are iid normal, then X+Y and X-Y are also independent normal we have just shown that X+Y and X-Y are normal. Now

Cov(X-Y,X+Y) =
Cov(X,X) + Cov(X,Y) + Cov(-Y,X) + Cov(-Y,Y) =
Var(X) + Cov(X,Y) - Cov(Y,X) -Var(Y) = 0

and as we shall see in a little bit this implies X-Y and X+Y are independent.

Lemma

If X and Z are independent such that Z and X+Z are normal, then X is normal as well because

$$\psi_{X+Z}(t) = \psi_X(t)\psi_Z(t)$$

$$\psi_X(t) = \psi_{X+Z}(t)/\psi_Z(t) =$$

$$\exp\{at + \tfrac{1}{2}b^2t^2\}/\exp\{ct + \tfrac{1}{2}d^2t^2\} =$$

$$\exp\{(a-c)t + \tfrac{1}{2}(b^2 - d^2)t^2\}$$

$$b^2 - d^2 = Var(X+Z) - Var(Z) = Var(X) > 0$$

so the mgf of X is of the form

$\psi(t) = \exp\{\mu t + \frac{1}{2}\tau t^2\}$

so $X \sim N(\mu, \sqrt{\tau})$

Lemma

If X, Z are independent random variables and Z is normal, then X+Z has a non-vanishing probability density function which has derivatives of all orders.

wlog assume $Z \sim N(0, 1/\sqrt{2})$. Consider

$f(x) = E[\exp(-(x-X)^2]$

Then $f(x) \neq 0$ for each x because exp(-(x-X)^2>0. Moreover all derivatives exist and are bounded uniformly ($x^k\exp(-x^2)$ has a finite minimum and maximum for all k), and therefore f has derivatives of all orders.

Now

$$P(X+Z \le t) = \int_{-\infty}^{\infty} P(X+Z \le t | Z = z) f_Z(z) dz =$$

$$\int_{-\infty}^{\infty} f_Z(z) P(X \le t - z) dz =$$

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\, 1/\sqrt{2}} \exp\{-\frac{1}{2} \frac{z^2}{1/2}\} P(X \le t - z) dz =$$

$$\pi^{-1/2} \int_{-\infty}^{\infty} \exp\{-z^2\} P(X \le t - z) dz =$$

$$\pi^{-1/2} \int_{-\infty}^{\infty} \exp\{-z^2\} \int_{-\infty}^{t-z} f_X(y) dy dz =$$

$$\pi^{-1/2} \int_{-\infty}^{\infty} \exp\{-z^2\} \int_{-\infty}^{t} f_X(x - z) dx dz =$$

$$\pi^{-1/2} \int_{-\infty}^{\infty} \int_{-\infty}^{t} \exp\{-z^2\} f_X(x - z) dx dz =$$

$$\pi^{-1/2} \int_{-\infty}^{\infty} \int_{-\infty}^{t} \exp\{-(x - y)^2\} f_X(y) dx dy =$$

$$\pi^{-1/2} \int_{-\infty}^{\infty} \int_{-\infty}^{t} \exp\{-(y - x)^2\} f_X(y) dy dx =$$

$$\pi^{-1/2} \int_{-\infty}^{t} E[\exp\{-y - X)^2\}] dy$$

so f is the density of X+Z.

Now for the finish of Bernstein's theorem: First we change notation and use rv's X_1 and X_2 . So we know X_1 and X_2 are independent and so are $X_1 + X_2$ and $X_1 - X$~2'.

Let Z_1 and Z_2 be iid normal rv's, independent of X_1 and X_2. Then define rv's

$Y_k = X_k + Z_k$

By the third lemma each of the Y_k's have a smooth non-zero pdf.

The joint density of the $(Y_1 + Y_2, Y_1 - Y_2)$ is

$$u = x + y, \quad v = x - y$$

$$x = \frac{u+v}{2}, \quad y = \frac{u-v}{2}$$

$$J = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}$$

$$f_{Y_1+Y_2,Y_1-Y_2}(u,v) = f_{Y_1}\left(\frac{u+v}{2}\right) f_{Y_2}\left(\frac{u-v}{2}\right)|J| =$$

$$\frac{1}{2} f_{Y_1}\left(\frac{u+v}{2}\right) f_{Y_2}\left(\frac{u-v}{2}\right)$$

and because $Y_1 + Y_2$ and $Y_1 - Y_2$ are independent by assumption this factors into two functions, one of x and the other of y, say a(x) and b(y).

Consider the functions

$Q_k\ (x) = \log(f_k\ (x))$

then the Q's are twice differentiable and we have

$$\begin{aligned} &Q_1(x+y) + Q_2(x-y) = \\ &\log(f_1(x+y)) + log(f_2(x-y)) = \\ &\log(f_1(x+y) f_2(x-y)) = \\ &\log(2a(2x)b(2y)) = \\ &\log 2 + \log a(2x) + \log b(2y) \end{aligned}$$

so

$$\frac{d}{dx}\{Q_1(x+y)+Q_2(x-y)\}=$$
$$\frac{d}{dx}\{\log 2+\log a(2x)+\log b(2y)\}=$$
$$\frac{d}{dx}\{\log a(2x)\}$$
$$\frac{d^2}{dxdy}\{Q_1(x+y)+Q_2(x-y)\}=$$
$$\frac{d}{dy}\frac{d}{dx}\{\log a(2x)\})=0$$

but also

$\frac{d^2}{dxdy}\{Q_1(x+y)+Q_2(x-y)\}=Q_1''(x+y)-Q_2''(x-y)$

and so

$Q_1''(x+y)=Q_2''(x-y)$

taking x=y we have

$Q_1''(2x)=Q_2''(0)=\text{const}$

and taking x=-y we have

$Q_2''(2y)=Q_1''(0)=\text{const}$

so

$Q_k(x)=a_kx^2+b_kx+c_k$

and

$f_k(x)=\exp(a_kx^2+b_kx+c_k)$

as a pdf f_k has to be integrable, so $a_k<0$, and by integrating f_k over R we find $c_k=-\frac{1}{2}\log(2\pi a_k)$. Therefore f_k is a normal density, and so Y_1 and Y_2 are normal.

Now Y_1 and Y_2 are iid normal and the independence of Y_1+Y_2 and Y_1 -Y~2‘ follows from the first lemma.

The theorem then follows from the second lemma.

2.3.2 Bivariate Normal RV

Definition

Let $\mu_1,\mu_2\in\mathbb{R}$, $\sigma_1,\sigma_2\in\mathbb{R}^+$ and $\rho\in[-1,1]$, then the random vector (X,Y) is said to have a bivariate normal distribution if

$$f(x,y)=\frac{1}{2\pi\sigma_1^2\sigma_2^2\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2}-2\frac{(x-\mu_1)}{\sigma_1}\frac{(y-\mu_2)\rho}{\sigma_2}+\frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right)$$

Theorem

a. Let (X,Y) be a bivariate normal. Let $U = (X - \mu_1)/\sigma_1$ and $V = (Y - \mu_2)/\sigma_2$. Then (U,V) is a bivariate normal random vector with

$\mu_U = \mu_V = 0, \sigma_U = \sigma_V = 1$ and $\rho_{UV} = \rho$

b. Let (U,V) be a bivariate normal random vector with $\mu_U = \mu_V = 0, \sigma_U = \sigma_V = 1$. Let $X = \mu_1 + \sigma_1 U$ and $Y = \mu_2 + \sigma_2 V$. Then (X,Y) a bivariate normal with parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ

proof

follows from a simple application of the transformation theorem

Theorem

Let (X,Y) be a bivariate normal. Then $X \sim N(\mu_1, \sigma_1)$.

proof

we will do this under the assumption $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$. The proof of the general case is exactly the same, only the arithmetic is a bit ugly.

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy =$$

$$\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}[x^2 - 2xy\rho + y^2]\right)dy$$

$$\int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}[x^2 - 2xy\rho + y^2 + x^2\rho^2 - x^2\rho^2]\right)dy$$

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}[y - x\rho]^2\right)dy =$$

$$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Theorem

Let (X,Y) be a bivariate normal. Then Cor(x,y) = ρ.

proof

again we will do this under the assumption $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$.

We already have E[X]=E[Y]=0. Now

$$E[XY] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf(x,y)dxdy$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}[x^2-2xy\rho+y^2]\right)dy =$$

$$\int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\left(\int_{-\infty}^{\infty} y\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}[y-x\rho]^2\right)dy\right)dx =$$

$$\int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi}}e^{-x^2/2}x\rho dx = \rho E[X^2] = \rho$$

and so Cor(X,Y) = Cov(X,Y)/[sd(X)sd(Y)] = E[XY] = ρ.

Theorem

let X and Y be two normal rv's, then

$X \perp Y$ iff Cor(X,Y)=0

proof

one direction is always true. For the other we have if $\rho = 0$

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-0^2}}\exp\left(-\frac{1}{2(1-0^2)}\left(\frac{(x-\mu_X)^2}{\sigma_X^2}-2\frac{(x-\mu_X)(y-\mu_Y)0}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right)\right) =$$

$$\frac{1}{\sqrt{2\pi}\sigma_X}\exp\left(\frac{1}{2}\frac{(x-\mu_X)^2}{\sigma_X^2}\right)\cdot\frac{1}{\sqrt{2\pi}\sigma_Y}\exp\left(\frac{1}{2}\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right) = f_X(x)\cdot f_Y(y)$$

2.3.2.1 Example

the joint distribution of two normal rv's need not be bivariate normal

Say $X \sim N(0,1)$ and let Y=-X is |X|>1 and Y=X if |X| < 1, then

$$\text{say } y < -1\text{: } P(Y \le y) = P(-X \le y) = P(X \ge -y) = P(X \le y)$$

$$\text{say } -1 < y < 1\text{: } P(Y \le y) = P(Y < -1) + P(-1 < Y < y) =$$

$$P(-X \le -1) + P(-1 < X < y) = P(X \ge 1) + P(-1 < X < y) =$$

$$P(X \le -1) + P(-1 < X < y) = P(X \le y)$$

$$\text{say } y > 1\text{: } P(Y \le y) = P(Y < -1) + P(-1 < Y < 1) + P(1 < Y < y) =$$

$$P(-X < -1) + P(-1 < X < 1) + P(1 < -X < y) =$$

$$P(X < -1) + P(-1 < X < 1) + P(-y < X < -1) = P(X < y)$$

so $Y \sim N(0,1)$ as well, but for example f(-2,-2)=0.

Theorem

say (X,Y) is a bivariate normal rv, then

a. $Z = X + Y \sim N(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_1\rho}))$

b. $Z = X|Y = y \sim N(\mu_1 - \rho(\sigma_1/\sigma_1)(y - \mu_2), \sigma_1\sqrt{1-\rho^2}))$

proof

a. is obvious: if (X,Y) is a bivariate normal rv, then $X \sim N(\mu_1, \sigma_1), Y \sim N(\mu_2, \sigma_2)$ and Cor(x,y)=ρ. Therefore

X+Y has a normal distribution, E[X+Y] = μ_1+μ_2 and Var(X+Y) = Var(X)+Var(Y)+2Cov(X,Y) = $\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2\rho$.

b. assume $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$, then

$$f_{X|Y=y}(x|y) = \frac{f(x,y)}{f_y(y)} =$$

$$\frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}[x^2-2xy\rho+y^2]\right)}{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}y^2\right)} =$$

$$\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}[x^2 - 2xy\rho + y^2 - (1-\rho^2)y^2]\right) =$$

$$\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}[x^2 - 2xy\rho + \rho^2y^2]\right) =$$

$$\frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}[x - \rho y]^2\right)$$

and so $X|Y = y \sim N(\rho y, \sqrt{1-\rho^2})$

2.3.3 Multivariate Normal RV

Let $\mu = (\mu_1, .., \mu_n)^T$ be a vector and $\Sigma = [\sigma_{ij}]$ be a symmetric positive semi-definite matrix (ie $x^T\Sigma x \geq 0$ for all x), then the random vector

$\mathbf{X} = (X_1, .., X_n)^T$

has a multivariate normal distribution if it has joint density

$$f\mathbf{X}(\mathbf{x}) = (2\pi)^{-n/2}|\Sigma|^{-1/2}\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}))$$

where $|\Sigma|$ is the determinant of Σ.

2.3.3.1 Example

n=1

$\Sigma = [a]$, $x^T\Sigma x = ax^2 \geq 0$ iff $a \geq 0$

$|\Sigma| = a, \Sigma^{-1} = 1/a$, and

$$f(x) = (2\pi)^{-1/2}|\Sigma|^{-1/2}\exp(-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{a}}\exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{a}\right)$$

so a is the variance of X.

2.3.3.2 Example

n=2: we have a symmetric 2x2 matrix Σ

$$\Sigma = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$$

$$(x\ y)\Sigma\binom{x}{y} = (x\ y)\begin{pmatrix} a & b \\ b & c \end{pmatrix}\binom{x}{y} =$$

$$(x\ y)\begin{pmatrix} ax + by \\ bx + cy \end{pmatrix} =$$

$$ax^2 + bxy + bxy + cy^2 =$$

$$ax^2 + 2bxy + cy^2 \geq 0$$

$$y = 0 \rightarrow a \geq 0, \quad x = 0 \rightarrow c \geq 0$$

$$\text{Let } u = \sqrt{a}x \quad v = \sqrt{c}y, \ d = \frac{b}{\sqrt{ac}}$$

then we want

$$u^2 + 2duv + v^2 \geq 0$$

$$u^2 + 2duv + v^2 =$$

$$u^2 + 2duv + d^2v^2 - d^2v^2 + v^2 =$$

$$(u + dv)^2 + (1 - d^2)v^2 \geq 0$$

iff

$$|d| \leq 1$$

so in order for Σ to be positive semidefinite we need $a, c \geq 0$ and $|d| = |b/\sqrt{ac}| \leq 1$ or $|b| \leq \sqrt{ac}$.

Inspired by the above calculation let's write Σ as follows:

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

Note that this is just as general as before, with $a = \sigma_x^2$, $c = \sigma_y^2$ and $b = \rho\sigma_x\sigma_y$. Then

$$|\Sigma| = \sigma_x^2\sigma_y^2(1-\rho^2)$$

$$\Sigma^{-1} = \frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)}\begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}$$

$$(x\ y)\Sigma^{-1}\begin{pmatrix} x \\ y \end{pmatrix} =$$

$$(x\ y)\frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)}\begin{pmatrix} \sigma_y^2 & -\rho\sigma_x\sigma_y \\ -\rho\sigma_x\sigma_y & \sigma_x^2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} =$$

$$\frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)}(x\ y)\begin{pmatrix} \sigma_y^2x - \rho\sigma_x\sigma_y y \\ \sigma_x^2y - \rho\sigma_x\sigma_y x \end{pmatrix} =$$

$$\frac{1}{\sigma_x^2\sigma_y^2(1-\rho^2)}(\sigma_y^2x^2 - \rho\sigma_x\sigma_y yx + \sigma_x^2y^2 - \rho\sigma_x\sigma_y xy) =$$

$$\frac{1}{1-\rho^2}\left(\frac{x^2}{\sigma_x^2} - 2\rho\frac{yx}{\sigma_x\sigma_y} + \frac{y^2}{\sigma_y^2}\right)$$

so we have a bivariate normal.

Theorem

Say **X** has a multivariate normal distribution. Then

$\mathbf{Z} = ((X_1 - \mu_1)/\sigma_{11}, .., (X_n - \mu_n)/\sigma_{nn})^T$

has a multivariate normal distribution with mean vector μ=(0,..,0)$^{\mathrm{T}}$ and variance-covariance matrix diag[σ_{ij}]. Then

$X_i \sim N(\mu_i, \sigma_{ii})$

b. $cov(X_i, X_j) = \sigma_{ij}$

without proof

We have the following characterization of a multivariate normal distribution, in some ways a generalization of Bernstein's theorem:

Theorem

Let $\mathbf{X}=(X_1, .., X_n)^T$. Then $\mathbf{X}$ has a multivariate normal distribution if and only if every linear combination $t_1X_1 + .. + t_nX_n$ has a normal distribution.

proof

one direction is obvious because the marginals of a multivariate rv are normal and the sum of normals is normal. The other direction can be shown using mgf's, where the mgf M of $\mathbf{X}$ is given by

$$M_{\mathbf{X}}(\mathbf{t}) = \exp(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t})$$

2.3.4 Theory of Errors

In real life almost any measuring device makes some errors. Some instruments are lousy and make big ones, other instruments are excellent and make small ones. Example

You want to measure the length a certain streetlight is red. You ask 10 friends to go with you and everyone makes a guess. Example

You want to measure the length a certain streetlight is red. You ask 10 friends to go with you. You have a stopwatch that you give to each friend.

Clearly in the second case we expect to get much smaller errors.

Around 1800 Karl Friedrich Gauss was thinking about what one could say in great generality about such measurement errors. He came up with the following rules that (almost) all measurement errors should follow, no matter what the instrument:

- Small errors are more likely than large errors.
- an error of ϵ is just as likely as an error of $-\epsilon$
- In the presence of several measurements of the same quantity, the most likely value of the quantity being measured is their average.

Now it is quite astonishing that JUST FROM THESE THREE rules he was able to derive the normal distibution!

3 Inequalities and Limit Theorems

3.1 Inequalities

Inequalities are very important in probability theory, both for the theory and for practical applications.

We start with a lemma that has nothing to do with probability:

Lemma

let a and b be any positive numbers, and let p and q be any positive numbers with 1/p+1/q=1. Then

$$\frac{a^p}{p} + \frac{b^q}{q} \geq ab$$

with "=" iff $a^p=b^q$

proof

fix b, and consider the function g with

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab$$

Now

$$\frac{d}{da}g(a) = a^{p-1} - b = 0 \Rightarrow a = b^{1/(p-1)}$$

$$\frac{d^2}{da^2}g(a)\Big|_{b^{1/(p-1)}} = (p-1)a^{p-2}\Big|_{b^{1/(p-1)}} = (p-1)b^{(p-2)/(p-1)} > 0$$

so g has a minimum at $a^p-1=b$. So

$$\frac{1}{p}a^p + \frac{1}{q}a^{(p-1)q} - aa^{p-1} = \frac{1}{p}a^p + \frac{1}{q}a^q - a^p = 0$$

because $(p-1)q = p$. This follows from

1/p+1/q=1 implies 1+p/q=p implies p/q=p-1.

Moreover the minimum of g is unique because g is convex for all a, so "=" holds iff $a^p-1=b$, which is the same as $a^p=b^q$.

Theorem (Holder's Inequality)

Let X and Y be any two rvs, and let p and q be as above. Then

$$|E[XY]| \leq E[|XY|] \leq (E[|X]^p)^{1/p}(E[|Y]^q)^{1/q}$$

proof

The first "$\leq$" follows from

$-|XY| \leq XY \leq |XY|$. For the second $\leq$ define

$$a = \frac{|X|}{(E|X|^p)^{1/p}} \text{ and } b = \frac{|Y|}{(E|Y|^q)^{1/q}}$$

using the lemma we get

$$\frac{1}{p}\frac{|X|^p}{E|X|^p} + \frac{1}{q}\frac{|Y|^q}{E|Y|^q} \geq \frac{|XY|}{(E|X|^p)^{1/p}(E|Y|^q)^{1/q}}$$

$$\text{but } E\left[\frac{1}{p}\frac{|X|^p}{E|X|^p} + \frac{1}{q}\frac{|Y|^q}{E|Y|^q}\right] = \frac{1}{p}\frac{E|X|^p}{E|X|^p} + \frac{1}{q}\frac{E|Y|^q}{E|Y|^q} = \frac{1}{p} + \frac{1}{q} = 1$$

This is actually a generalization of the Cauchy-Schwartz inequality we discussed earlier: if p=q=1/2 we get

$$|E[XY]| \leq E[|XY|] \leq \sqrt{E[X^2]}\sqrt{E[Y^2]}$$

These inequalities are stated here in terms of expectations, but they hold in general for sums and integrals as well.

Some other useful cases are:

If we set Y=1 we get

$E|X| \leq E|X|^{p\,1/p}, 1 < p < \infty$

For 1<r<p, if we replace |X| by $|X|^r$, we get

$E|X|^r \leq \{E|X|^{pr}\}^{1/p}$

and writing s=pr (which implies s>r) we get

Liapunov's Inequality

$(E|X|^r)^{1/r} \leq (E|X|^s)^{1/s}$ for $1 < r < s < \infty$

3.1.0.1 Example

If a rv X has a finite k^{th} moment, it has a finite j^{th} moment for all $j \leq k$.

By Liapunovs inequality

$E[|X|^j] \leq (E[|X|^k])^{j/k} < \infty$

The inequalities above are not really probability theory but are inequalities from real analysis. Next we consider a new type of inequality true specifically in probability theory:

Theorem (Markov's Inequality)

If X takes on only non negative values, then for any a>0

$$P(X \geq a) \leq \frac{E[X]}{a}$$

proof

$$P(X \geq a) = \int_a^\infty f(x)dx \leq$$

$$\int_a^\infty \frac{x}{a} f(x)dx \leq \qquad \text{because } x > a \text{ for all } x \in (a, \infty)$$

$$\frac{1}{a}\int_0^\infty xf(x)dx = \qquad \text{because} \int_0^a xf(x)dx \geq 0$$

$$\frac{1}{a}\int_{-\infty}^\infty xf(x)dx = \qquad \text{because } f(x) = 0 \text{ for all } x < 0$$

$$\frac{E[X]}{a}$$

Markov's inequality implies what is perhaps the most famous inequality in probability:

Theorem (Chebyshev's Inequality)

If X is a r.v. with mean μ and variance σ^2, then for any k>0:

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

proof

$$P(|X - \mu| \geq k\sigma) =$$

$$P((X - \mu)^2 \geq k^2\sigma^2) \leq \text{ because of Markov's Inequality}$$

$$\frac{E[(X-\mu)^2]}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

Theorem (Chernoff Bounds)

Let X be a rv with moment generating function $\psi(t) = E[e^{tX}]$. Then for any a>0

$P(X \geq a) \leq e^{-ta}\psi(t)$ for all t>0
$P(X \leq a) \leq e^{-ta}\psi(t)$ for all t<0

proof

For t>0

$$P(X \geq a) =$$
$$P(e^{tX} \geq e^{ta}) \leq$$
$$\frac{E[e^{tX}]}{e^{ta}} = e^{-ta}\psi(t)$$

The proof for t<0 is similar.

As we know a random variable that has a moment generating function that is finite in an open neighborhood of 0 has all its moments, that $E[|X|^k] < \infty$ for all k. So this is a rather strong condition, and therefore leads to very good bounds.

3.1.0.2 Example

say $Z \sim N(0,1)$, then

$$P(Z \geq a) \leq e^{-ta}e^{t^2/2} = \exp\{t^2/2 - ta\} \text{ for all } t > 0$$
$$\log P(Z \geq a) \leq \frac{t^2}{2} - ta \text{ for all } t > 0$$
$$\log P(Z \geq a) \leq \min\{\frac{t^2}{2} - ta | t > 0\}$$
$$\frac{d}{dt}\{\frac{t^2}{2} - ta\} = t - a = 0 \;\; t = a$$
$$\log P(Z \geq a) \leq \frac{a^2}{2} - a * a = -\frac{a^2}{2}$$
$$P(Z \geq a) \leq e^{-a^2/2}$$

which is a very useful upper bound on the tail probabilities of a standard normal.

For the last inequality first recall

Definition

A function g is said to be *convex* if for all x and y and $0 < \lambda < 1$

$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$

Theorem (Jensen's Inequality)

For any rv X, if g is a convex function we have

$Eg(X) \geq g(EX)$

proof

let l(x) be a tangent line to g(x) at the point g(EX). Write l(x)=a+bx for some a and b. Now by the convexity of g we have

$g(x) \geq a + bx$

and so

$Eg(X) \geq E(a + bx) = a + bEX = l(EX) = g(EX)$

Of course if g is a concave function, the -g is convex and we have

$E[g(X)] = -E[-g(X)] \leq -(-g(E[X]) = g(E[X])$

3.1.0.3 Example

g(x)=x^2 is convex, and so

$EX^2 \geq (EX)^2$

which implies

$Var(X) = EX^2 - E[X]^2 \geq 0$

3.1.0.4 Example

If x>0 g(x)=1/x is convex, so $E(1/X) \geq 1/EX$

3.2 Limit Theorems

3.2.1 Convergence Concepts

Say we have a sequence of numbers a_n . Then there is just one definition of a "limit", namely $a_n \to a$ iff for every $\epsilon > 0$ there exists an n_ϵ such that $|a_n - a| < \epsilon$ for all $n > n_\epsilon$.

3.2.1.1 Example

say $a_n = (1 + 1/n)^n$. Show that $a_n \to e$

Fix n, and let t be such that $1 \leq t \leq 1 + 1/n$. Then

$$\frac{1}{1+\frac{1}{n}} \le \frac{1}{t} \le 1$$

$$\int_1^{1+1/n} \frac{1}{1+\frac{1}{n}}dt \le \int_1^{1+1/n} \frac{1}{t}dt \le \int_1^{1+1/n} 1dt$$

$$\frac{1}{1+\frac{1}{n}}\frac{1}{n} \le \ln(t)\Big|_1^{1+1/n} \le \frac{1}{n}$$

$$\frac{1}{1+n} \le \ln(1+\frac{1}{n}) \le \frac{1}{n}$$

$$e^{\frac{1}{1+n}} \le 1+\frac{1}{n} \le e^{\frac{1}{n}}$$

$$e \le (1+\frac{1}{n})^{n+1} \text{ and } (1+\frac{1}{n})^{n} \le e$$

$$\frac{e}{1+\frac{1}{n}} \le (1+\frac{1}{n})^{n} \le e$$

$$\frac{n}{n+1}e \le (1+\frac{1}{n})^{n} \le e$$

$$|a_n - a| = |(1+\frac{1}{n})^{n} - e| < \varepsilon$$

$$-\varepsilon < (1+\frac{1}{n})^{n} - e < \varepsilon$$

$$e-\varepsilon < (1+\frac{1}{n})^{n} < e+\varepsilon$$

$$e-\varepsilon < \frac{n}{n+1}e$$

$$(n+1)e-(n+1)\varepsilon - ne < 0$$

$$e-(n+1)\varepsilon < 0$$

$$n > \frac{e}{\varepsilon} - 1$$

so fix an $\epsilon > 0$. Then if $n > e/\epsilon - 1$ we have

$|(1 + 1/n)^n - e| < \epsilon$

and therefore

$(1 + 1/n)^n \to e$

Things already get a little more complicated if we go to sequences of functions. Here there are two ways in which they can converge:

- **Pointwise Convergence:** $f_n(x) \to f(x)$ pointwise iff for every x in S and every $\epsilon > 0$ there exists an $n_{\epsilon,x}$ such that $|f_n(x) - f(x)| < \epsilon$ for all $n > n_{\epsilon,x}$.
- **Uniform Convergence:** $f_n(x) \to f(x)$ pointwise iff for every x in S and every $\epsilon > 0$ there exists an n_ϵ such that $|f_n(x) - f(x)| < \epsilon$ for all $n > n_\epsilon$.

and there is a simple hierarchy: uniform convergence implies pointwise convergence but not vice versa.

3.2.1.2 Example

say $f_n(x) = 1 + x/n, x \in S = [A, B]$ where A<B, f(x)=1, then $f_n(x) \to f(x)$ uniformly.

$|f_n(x) - f(x)| = |1 + x/n - 1| =$
$|x/n| \leq \max(|A|, |B|)/n < \epsilon$

if $n \geq \max(|A|, |B|)/\epsilon$

3.2.1.3 Example

say $f_n(x) = x^n, S = [0, 1], f(x) = I_1(x)$, then $f_n(x) \to f(x)$ pointwise but not uniformly.

say x<1 then $|f_n(x) - f(x)| = x^n < \epsilon$ for all $n > n_{\epsilon,x} = \log(\epsilon)/\log(x)$.

say x=1 then $|f_n(x) - f(x)| = 0 < \epsilon$ for all $n > n_{\epsilon,x} = 1$

but

fix $\varepsilon < \frac{1}{2e}$ and assume

$|f_n(x) - f(x)| = x^n < \varepsilon$ for all $x \in [0, 1)$ and all $n > n_\varepsilon$

$x = 1 - \frac{1}{n}$

$x^n = (1 - \frac{1}{n})^n > \frac{1}{2e}$

Now when we go to probabilities it gets a bit more complicated still. Say we have a sequence of rv's X_n with means μ_n and cdf's F_n , and a rv X with mean μ and cdf F. Then we have:

Definition

- **Convergence in Mean:** $X_n \to X$ in mean iff $\mu_n \to \mu$.
- **Convergence in Quadratic Mean:** $X_n \to \mu$ in quadratic mean iff $E[X_n] \to \mu$ and $Var(X_n) \to 0$.
- **Convergence in L^p:** $X_n \to X$ in L^p iff $E[|X_n - X|^p] \to 0$.
- **Convergence in Distribution (Law)** $X_n \to X$ in law iff $F_n(x) \to F(x)$ pointwise for all x where F is continuous.
- **Convergence in Probability:** $X_n \to X$ in probability iff for every $\epsilon > 0 \lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$.
- **Almost Sure Convergence:** $X_n \to X$ almost surely iff for every $\epsilon > 0$ $P(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1$.

3.2.1.4 Example

Let X_n have density $f_n(x) = nx^{n-1}, 0 < x < 1$ and let X be such that P(X=1)=1. Then

$$f_n(x) = nx^{n-1},\ 0 \leq x \leq 1$$

$$f_X(x) = \begin{cases} 1 & x = 1 \\ 0 & x \neq 1 \end{cases} = \delta_1(x)$$

$$EX = 1 \cdot 1 = 1$$

$$EX_n = \int_{-\infty}^{\infty} x f_n(x) dx = \int_0^1 x n x^{n-1} dx = n \int_0^1 x^n dx =$$

$$\frac{n}{n+1} x^{n+1} \Big|_0^1 = \frac{n}{n+1} \to 1 = EX$$

$$EX_n^2 = \int_{-\infty}^{\infty} x^2 f_n(x)dx = \int_0^1 x^2 n x^{n-1} dx =$$

$$\frac{n}{n+2} x^{n+2}|_0^1 = \frac{n}{n+2}$$

$$VarX_n = \frac{n}{n+2} - \left(\frac{n}{n+1}\right)^2 =$$

$$\frac{n(n+1)^2 - n^2(n+2)}{(n+1)^2(n+2)} = \frac{n}{(n+1)^2(n+2)} \to 0 \text{ as } n \to \infty$$

$$F_{X_n}(x) = x^n$$

$$F_X(x) = \begin{cases} 0 & x < 1 \\ 1 & x \geq 1 \end{cases}$$

$$F_{X_n}(x) = x^n \to F_X(x) \text{ for all } x$$

$$P(|X_n - X| \geq \varepsilon) = 1 - P(|X_n - 1| < \varepsilon) =$$

$$1 - P(1 - \varepsilon < X_n < 1 + \varepsilon) =$$

$$1 - P(1 - \varepsilon < X_n) = P(X_n < 1 - \varepsilon) =$$

$$(1 - \varepsilon)^n \to 0$$

$$A_{n,\varepsilon} = \{\omega : |X_n(\omega) - 1| < \varepsilon\} =$$

$$\{\omega : 1 - \varepsilon < X_n(\omega) < 1 + \varepsilon\} =$$

$$\{\omega : X_n(\omega) > 1 - \varepsilon\} \to S$$

$$P(\lim_{n\to\infty} |X_n - X| < \varepsilon) = P(\lim_{n\to\infty} A_{n,\varepsilon}) = P(S) = 1$$

$$E[|X_n - X|^p] = E[|X_n - 1|^p] =$$

$$\int_0^1 |x-1|^p n x^{n-1} dx = \int_0^1 (1-x)^p n x^{n-1} dx =$$

$$\frac{\Gamma(n)\Gamma(p+1)}{\Gamma(n+p+1)} n \int_0^1 \frac{\Gamma(n+p+1)}{\Gamma(n)\Gamma(p+1)} x^{n-1}(1-x)^{(p+1)-1} dx =$$

$$\frac{(n-1)!\Gamma(p+1)}{(n+p)\cdot(n+p-1)\cdot\ldots\cdot(p+1)\Gamma(p+1)} n = \frac{n!}{(n+p)\cdot(n+p-1)\cdot\ldots\cdot(p+1)} =$$

$$\prod_{k=1}^{n} \frac{k}{k+p} = \exp(\sum_{k=1}^{n} \log(\frac{k}{k+p})) =$$

$$\exp(\sum_{k=1}^{n} \log(1 - \frac{p}{k+p})) \simeq$$

$$\exp(-\sum_{k=1}^{n} \frac{p}{k+p}) \to 0 \text{ as } n \to \infty$$

so $X_n \to X$ in all the ways possible.

3.2.1.5 Example

Say X_n has density $f(x) = n/(n+1)x^{1/n}, 0 < x < 1$ and $X \sim$ U[0,1]. Then

$$EX_n = \int_{-\infty}^{\infty} x f_n(x)dx = \int_0^1 x \frac{n}{n+1} x^{1/n} dx =$$

$$\frac{n}{n+1}\frac{1}{2+1/n} x^{2+1/n}\Big|_0^1 = \frac{n}{n+1}\frac{1}{2+1/n} \to \frac{1}{2} = EX$$

$$EX_n^2 = \int_{-\infty}^{\infty} x^2 f_n(x)dx = \int_0^1 x^2 \frac{n}{n+1} x^{1/n} dx =$$

$$\frac{n}{n+1}\frac{1}{3+1/n} x^{3+1/n}\Big|_0^1 = \frac{n}{n+1}\frac{1}{3+1/n} \to \frac{1}{3}$$

so $Var[X_n] \to 1/3 - (1/2)^2 = 1/12 \neq 0$, so we don't have convergence in quadratic mean. We do have convergence in distribution, though:

$$F_n(x) = x^{1+1/n} \to x = F(x) \; 0 < x < 1$$

How about convergence in probability? As stated here that can not be decided because we would need the joint density of X_n and X. If they are independent X_n will not converge to X in probability:

$$P(|X_n - X| > \varepsilon) = 1 - P(-\varepsilon < X_n - X < \varepsilon) =$$

$$1 - \int_0^1 P(-\varepsilon < x - X < \varepsilon) f_n(x)dx =$$

$$1 - \int_0^1 P(x - \varepsilon < X < x + \varepsilon) f_n(x)dx \simeq$$

$$1 - \int_0^1 2\varepsilon f_n(x)dx = 1 - 2\varepsilon > 0$$

3.2.2 Relationships between Convergences

Unfortunately there is no simple hierarchy between the different modes of convergence. Here are some relationships:

Theorem

a. convergence in quadratic mean implies convergence in probability.

b. convergence in probability implies convergence in distribution. The reverse is true if the limit is a constant.

c. almost sure convergence implies convergence in probability, but not vice versa.

proof:

a. convergence in quadratic mean implies that the limit is a constant, say μ. Then

$$P(|X_n - X| \geq \varepsilon) = 1 - P(\mu - \varepsilon < X_n < \mu + \varepsilon) =$$

$$1 - P(\mu - \mu_n - \varepsilon < X_n - \mu_n < \mu - \mu_n + \varepsilon) \leq$$

$$1 - P(-(m_n + \varepsilon) < X_n - \mu_n < m_n + \varepsilon) =$$

$$P(|X_n - \mu_n| \geq m_n + \varepsilon) =$$

$$P(|X_n - \mu_n| \geq \frac{m_n + \varepsilon}{sd(X_n)} sd(X_n)) \leq \frac{1}{\left(\frac{m_n + \varepsilon}{sd(X_n)}\right)^2} =$$

$$\frac{VarX_n}{(m_n + \varepsilon)^2} \to 0$$

where

$$m_n = \min\{\pm|\mu_n - \mu|\}$$

b. say $X_n \to X$ in probability. We will do the proof in the case where say X_n and X are continuous r.vs (so we need not worry about terms of the form P(X=x)). Now

$$P(X \le x - \varepsilon) =$$

$$P(X \le x - \varepsilon, |X_n - X| < \varepsilon \cup X \le x - \varepsilon, |X_n - X| > \varepsilon) =$$

$$P(X \le x - \varepsilon, |X_n - X| < \varepsilon) + P(X \le x - \varepsilon, |X_n - X| > \varepsilon) \le$$

$$P(X \le x - \varepsilon, X_n - \varepsilon < X < X_n + \varepsilon) + P(|X_n - X| > \varepsilon) \le$$

$$P(X_n - \varepsilon \le x - \varepsilon, X_n - \varepsilon < X < X_n + \varepsilon) + P(|X_n - X| > \varepsilon) \le$$

$$P(X_n \le x) + P(|X_n - X| > \varepsilon)$$

$$F(x - \varepsilon) \le F_n(x) + P(|X_n - X| > \varepsilon)$$

similarly

$$F(x + \varepsilon) \ge F_n(x) - P(|X_n - X| > \varepsilon)$$

so

$$F(x - \varepsilon) - F(x) - P(|X_n - X| > \varepsilon) \le$$

$$F_n(x) - F(x) \le$$

$$F(x + \varepsilon) - F(x) + P(|X_n - X| > \varepsilon)$$

say $X_n \to c$ in law, then

$$F_n(x) \to F(x) = I_{[c,\infty)}(x)$$

$$P(|X_n - c| \ge \varepsilon) = P(X_n < c - \varepsilon \quad \text{or} \quad X_n > c + \varepsilon) =$$

$$P(X_n < c - \varepsilon) + P(X_n > c + \varepsilon) =$$

$$F_n(c - \varepsilon) + (1 - F_n(c + \varepsilon)) \to 0 + (1 - 1) = 0$$

c. is done with a counter example: Let $X \sim$ U[0,1] and define

$$A_1 = [0,1]$$
$$A_2 = [0,1/2] \quad A_3 = [1/2,1]$$
$$A_4 = [0,1/4] \quad A_5 = [1/4,1/2] \quad A_6 = [1/2,3/4] \quad A_7 = [3/4,1]$$
$$\text{etc}$$
$$\text{so } A_{2^m+i} = [\tfrac{i}{2^m}, \tfrac{i+1}{2^m}], m = 0,1,2,\ldots i = 0,1,..,2^m - 1$$

$$\text{Let } Y_n = I_{A_n}(X) \text{ and } Y = \delta_0$$

$$P(|Y_n - Y| \geq \varepsilon) = P(|Y_n| \geq \varepsilon) = P(Y_n = 1) = P(I_{A_n}(X) = 1)$$
$$P(X \in A_n) = length(A_n) \to 0$$

and so $Y_n \to Y$ in probability.

However, for every x in [0,1] Y_n (x)=0 infinitely often and Y_n (x)=1 infinitely often, therefore Y_n does not converge to Y almost surely.

Theorem

If $X_n \to 0$ in L^p then $X_n \to 0$ in probability.

Note: $X_n \to X$ iff $X_n - X \to 0$.

proof

Say $X_n \to 0$ in L^p, and let $g(x)=|x|^p$, then by Chebyshev's inequality

$P(|X_n| \geq \epsilon) \leq E[|X_n|^p)/\epsilon^p \to 0$

and so $X_n \to 0$ in probability.

Theorem (Slutsky)

Say $X_n \to X$ in distribution and g is any continuous function, then $g(X_n) \to g(X)$ in distribution.

Lemma

From the above results, it is easy to show that if $X_n \to X$ in distribution and $Y_n \to c$ in distribution, then

- $X_n + Y_n \to X + c$ in distribution.

- $X_n Y_n \to cX$ in distribution

which is the most common version of Slutsky's theorem.

In loose terms, the theorem states that if a rv converges to a constant, then it essentially behaves as a constant for addition and multiplication.

Note that the condition that Y_n converge to a constant is necessary.

3.2.3 Almost Sure Convergence

Consider a number ω in [0,1], and let's write the number using its decimal expansion as

$\omega = 0.x_1 x_2 x_3 ...$

Let T_n be the set of ω's with $x_n > 0$ and $x_m = 0$ for m>n. Then T_n has finitely many numbers. Let T be the set of ω's with a terminating expansion, then

$T = \cup_n T_n$

and so T is countable. If we consider the game: pick a number in [0,1] at random, we therefore have P(T)=0 because the total number of real numbers in [0,1] is uncountable!

Now let k be such that $0 \leq k \leq 9$, and define $\nu(\omega, n, k)$ to be the number of x_i's among the first n digits in the expansion of ω with $x_i = k$. Define (if it exists)

$$\lim_{n\to\infty} \frac{\nu(\omega, n, k)}{n} = \varphi_k(\omega)$$

so $\phi_k(\omega)$ is the long run relative frequency of the digit k in the expansion of ω.

Intuitively it seems obvious that $\phi_k(\omega) = 1/10$ for all k for "most" ω's, but is actually very simple to write down numbers for which this is not true: $0.1111\ldots$, $0.121212\ldots$ etc.

Numbers for which $\phi_k(\omega) = 1/10$ for all k are called *simply normal*, and how many such numbers there are was an open question in Number Theory for a long time.

Theorem (Borel)

Let N be the set of simply normal numbers. Then P(N)=1

proof (outline)

Let ζ_i be a random variable with $P(\zeta_i = k) = 1/10$. Let the sequence of ζ_i's be independent and set

$\omega = 0.\zeta_1 \zeta_2 \zeta_3 ...$

Let $S_n = \zeta_1 + ... + \zeta_n$, then by the strong law of large numbers

$S_n/n \to 1/10$ almost surely

therefore $P(\phi_k = 1/10) = 1$ for all k, and finally

$$P(\bigcap_{k=0}^{9} [\varphi_k = 1/10]) = 1$$

Considered as a problem in probability theory this seems almost trivial. 100 years ago, though, probability theory was considered a branch of mathematics hardly worthy of the name, and indeed before the advent of modern probability theory in the 1930's it was a strange field of strange results. When Borel published his proof it came as a shock to many mathematicians that this strange field could yield interesting results in the purest of fields, Number Theory.

Note that although N has probability 1, N^c is still uncountable. For example, take all the numbers in N and remove all the 0's, then all these numbers are in N^c, but there are clearly still uncountably many of these!

So, is $\pi - e = 0.423310825131..$ simply normal? Actually there does not exist a mathematical theory even today that would allow us to answer that question easily.

Let's consider the following generalization: Let $k_1..k_r$ be a "block" of r consecutive numbers in the expansion of ω, let $\nu(\omega, n, k_1..k_r)$ be the number of times the block occurs in ω up to n and let

$\phi(\omega, k_1..k_r) = lim_{n\to\infty}\nu(\omega, n, k_1..k_r)/n$

Consider the set $A \in [0, 1]$ such that for $\omega \in A$ $\phi(\omega, k_1..k_r) = 1/10^r$ for all possible blocks $k_1..k_r$.

Now the same reasoning as above shows that P(A)=1.

A number for which this is true not only for any r but also if we change from base 10 to any other base is called *normal*.

Consider what this to imply: Take your name, "code" it into a string of numbers. Pick a real number ω at random, then with probability 1 somewhere in the extension of that number you will find your name!

It gets weirder:

- you will not find it it just once but over and over again.
- let x be any real number and let $\epsilon > 0$, then this not only true for the interval [0,1] but also for the interval $[x, x + \epsilon]$
- It does not matter how large r is! Take the complete works of Shakespeare, they are in there too!

Is this really true? Our math seems to say yes, but so far no direct prove has been done.

To read more about normal numbers go here: Normal Number

3.2.4 Laws of Large Numbers

Theorem (Weak Law of Large Numbers WLLN)

Let $X_1, X_2, ...$ be a sequence of independent and identically distributed (iid) r.v.'s having mean μ. Then $\bar{X}$ converges to μ in probability.

proof (assuming in addition that $V(X_i) = \sigma^2 < \infty$)

$$E[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} EX_i = \mu$$

$$Var[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n^2}\sum_{i=1}^{n} VarX_i = \frac{\sigma^2}{n} \to 0$$

so $\bar{X} \to \mu$ in quadratic mean and therefore in probability.

It is best to think of this (and other) limits theorems not as one theorem but as a family of theorems, all with the same conclusion but with different conditions. For example there are weak laws even if the X_n's are not independent, don't have the same mean and don't even have finite standard deviations.

This theorem forms the bases of (almost) all simulation studies: say we want to find a parameter θ of a population. We can generate data from a random variable X with pdf $f(x|\theta)$ such that $Eh(X) = \theta$. Then by the law of large numbers

$$\overline{h(X)} = \frac{h(X_1) + \ldots + h(X_n)}{n} \to \theta \text{ as } n \to \infty$$

Theorem (Strong Law of Large Numbers SLLN)

Let $X_1, X_2, ...$ be a sequence of independent and identically distributed (iid) r.v.'s having mean μ. Then $\bar{X}$ converges to μ almost surely.

proof needs some measure theory, can be found in most standard textbooks

3.2.4.1 Example

in a game a player rolls 5 fair dice. He then moves his game piece along k fields on a board, where k is the smallest number on the dice + largest number on the dice. For example if his dice show 2, 2, 3, 5, 5 he moves 2+5 = 7 fields. What is the mean number of fields θ a player will move?

To do this analytically would be quite an exercise. To do it via simulation is easy. Using R we can this as follows:

```r
x <- matrix(sample(1:6,
                   size=5*10^5,
                   replace=TRUE), nrow=5)
z <- apply(x, 2, min) + apply(x, 2, max)
round(mean(z), 1)
```

```
## [1] 7
```

Theorem (Weierstrass Approximation Theorem)

Let f be a continuous function on [0,1], and let $\epsilon > 0$. Then there exists a polynomial p such that

$|f(x) - p(x)| < \epsilon$ for all $x \in [0, 1]$

Note that this theorem is from real analysis, it has nothing what so ever to do with probability theory. It is actually a special (early) version of one of the most famous theorems in Real Analysis, the Stone-Weierstrass theorem.

proof

We will consider polynomials of the form

$$p_n(x) = \sum_{k=0}^{n} f(\tfrac{k}{n}) \binom{n}{k} x^k (1-x)^{n-k}$$

which are called *Bernstein* polynomials after Sergei Bernstein. Note their connection to the binomial distribution as well as the Beta distribution.

We will show that for any $\epsilon > 0$ there exists an $n(\epsilon)$ such that

$|f(x) - p_{n(\epsilon)}(x)| < \epsilon$ for all $x \in [0, 1]$.

For each x, consider a sequence of Bernoulli trials $\{X_n\}$ with success probability x, and let $S_n = \sum_{k=1}^{n} X_k$. We know that $S_n \sim Bin(n, x)$, so that

$$P(S_n = k) = \binom{n}{k} x^k (1-x)^{n-k}$$

$$E[f(\tfrac{S_n}{n})] = \sum_{k=0}^{n} f(\tfrac{k}{n}) \binom{n}{k} x^k (1-x)^{n-k} = p_n(x)$$

S_n is a sum of independent random variables with a finite variance, so by the weak law of large numbers

$S_n/n \to E[f(x)] = f(x)$

in probability, and we have shown pointwise convergence of p_n (x) to f(x). It remains to show uniform convergence. Let $\delta > 0$, then

$$
\begin{aligned}
&|p_n(x) - f(x)| = \\
&\left|\sum_{k=0}^{n} f\left(\tfrac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k} - f(x)\right| = \\
&\left|\sum_{k=0}^{n} f\left(\tfrac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k} - \sum_{k=0}^{n} f(x)\binom{n}{k}x^k(1-x)^{n-k}\right| = \\
&\left|\sum_{k=0}^{n} \left[f\left(\tfrac{k}{n}\right) - f(x)\right]\binom{n}{k}x^k(1-x)^{n-k}\right| \leq \\
&\sum_{k=0}^{n} \left|f\left(\tfrac{k}{n}\right) - f(x)\right|\binom{n}{k}x^k(1-x)^{n-k} = \qquad (*) \\
&E\left[\left|f\left(\tfrac{S_n}{n}\right) - f(x)\right|\right] = \\
&E\left[\left|f\left(\tfrac{S_n}{n}\right) - f(x)\right|; \left|\tfrac{S_n}{n} - x\right| > \delta\right] + E\left[\left|f\left(\tfrac{S_n}{n}\right) - f(x)\right|; \left|\tfrac{S_n}{n} - x\right| \leq \delta\right] =
\end{aligned}
$$

where (*) follows from the triangle inequality $|a+b| \leq |a| + |b|$.

Now f is continuous on a compact interval, so f is uniformly continuous. Therefore for any $\epsilon > 0$ there exists a $\delta > 0$ (independent of x and y) such if $|x - y| < \delta$ we have $|(f(x) - f(y)| < \epsilon/2$.

With this choice of δ the second term above is bounded by $\epsilon/2$.

On the other hand again using the triangle inequality and the fact that [0,1] is a compact set we have

$|f(x) - f(y)| \leq |f(x)| + |-f(y)| \leq$
$2 \max |f|, 0 \leq x \leq 1 =: M < \infty$

and so the first term is bounded above by $MP(|S_n/n - x| > \delta)$

Finally we have E[S~n‘]=nx and $Var(S_n) = nx(1-x) \leq n/4$ for $0 \leq x \leq 1$.

By Chebyshev's inequality we have

$$
\begin{aligned}
&P(|S_n/n - x| > \delta) \leq \\
&Var(S_n)/\delta^2 = \\
&nx(1-x)/(\delta^2 n^2) \leq \\
&1/(4\delta^2 n)
\end{aligned}
$$

and so we have bounded the first term above as well.

3.2.5 Convergence of Series

Let $\{X_n; n = 1, 2, ..\}$ be a sequence of random variables. Let $S_n = X_1 + .. + X_n$. What can be said about the convergence of such a series? The most famous result is

Theorem (Kolmogorov's Three Series 1929)

Let $\{X_n\}$ be independent and define for some fixed A>0 a random variable Y_n as Y_n=X_n if $|X\ n\ | \leq A$ and 0 otherwise. (Essentially, Y_n is X_n truncated at $\pm A$).

Then the series $\sum X_n$ converges almost surely if and only if the following three series converge:

$$1) \sum_n P(|X_n| > A) = \sum_n P(X_n \neq Y_n)$$

$$2) \sum_n E[Y_n]$$

$$3) \sum_n \sigma^2(Y_n)$$

3.2.5.1 Example

As we know

$\sum 1/n = 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + ... = \infty$

and

$\sum (-1)^{n+1}/n = 1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + ... = \log(2) < \infty.$

How about something in between, namely if in each term of the sum we choose -1 with probability p and 1 with probability 1-p? That is let $Z_n \sim$ Ber(p), and $X_n = (-1)^{Z_n}/n$.

In the three-series theorem let's take A=1, then $Y_n = X_n$, and the sum in 1) is 0.

For 2) we have

$$E[Y_n] = E[X_n] = E[(-1)^{Z_n}/n] = \frac{1}{n}(1 \cdot p + (-1)(1-p)) = \frac{2p-1}{n}$$

$$\sum_{n=1}^{\infty} E[Y_n] = \sum_{n=1}^{\infty} \frac{2p-1}{n} = \begin{cases} 0 & \text{if } p = 1/2 \\ \infty & \text{if } p \neq 1/2 \end{cases}$$

and so the sum can only converge if p=1/2. To see whether it

really does we need to check 3):

$$\sigma^2(Y_n) = E[Y_n^2] = E[X_n^2] = E[((-1)^{Z_n}/n)^2] = \frac{1}{n^2}$$
$$\sum_{n=1}^{\infty} \sigma^2(Y_n) = \sum_{n=1}^{\infty} \frac{1}{n^2} < \infty$$

and so indeed the sum converges almost surely

3.2.5.2 Example

$$Z_n \sim Ber(1/2) \quad X_n = (-1)^{Z_n}/\sqrt{n}$$

$$A = 1$$
$$\sum_{n=1}^{\infty} P(X_n \neq Y_n) = 0$$
$$E[Y_n] = 0, \sum_{n=1}^{\infty} E[Y_n] = 0$$

$$\sigma^2(Y_n) = E[Y_n^2] = E[X_n^2] = E[((-1)^{Z_n}/\sqrt{n})^2] = \frac{1}{n}$$
$$\sum_{n=1}^{\infty} \sigma^2(Y_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

and so this random sum does not converge almost surely, although the 'fixed-sign" sum $\sum(-1)^n/\sqrt{n}$ does.

3.3 Central Limit Theorems

Recall: a random variable X is said to be normally distributed with mean μ and variance σ^2 if it has density:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right) \text{ for all } x \in \Re$$

We use the symbol Φ for the distribution function of a standard normal r.v., so

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

Let $X_1, X_2, ..$ be a sequence of r.v.'s with means $E[X_i] = \mu_i$ and $\text{sd}(X_i) = \sigma_i$. Let $\bar{X}_n$ be the sample mean of the first n observations. Then a *central limit theorem* would assert that

$$P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_i}{\sigma_i} \leq x \right) \to \Phi(x)$$

for all x, or that this standardized sum converges to a standard normal in distribution.

Note that plural "s" in the title. As with the laws of large number there are many central limit theorems, all with different conditions on

a. dependence between the X_i 's

b. μ_i's

c. σ_i's

as a rough guide we have to have some combination of

a. not to strong a dependence

b. $\mu_i \to \mu$ finite

c. σ_i goes neither to 0 nor to ∞ to fast

The version of the CLT for Bernoulli rv's is famous all by itself, it is called the *DeMoivre-Laplace* theorem. It was the first CLT with a rigorous proof.

Theorem (DeMoivre-Laplace)

let $\{X_n\}$ be independent rv's with $X_i \sim Ber(p)$. Let $S_n = \sum_{i=1}^{n} X_i$ and let $Z \sim N(0,1)$ then

$$\frac{S_n - np}{\sqrt{np(1-p)}} \to Z$$

in distribution.

The theorem was proven by Abraham de Moivre in 1738 for the case p=1/2, and generalized to $p \neq 1/2$ by Pierre-Simon Laplace in his famous book Theorie Analytique des Probabilites, published in 1812.

proof

We begin by showing that if for a large n we let k be in the neighborhood of np, then

$$\binom{n}{k} p^k q^{n-k} \simeq \frac{1}{\sqrt{2npq}} \exp\left\{-\frac{1}{2npq}(k-np)^2\right\}$$

We will make use Sterling's formula for n!:

$$n! \simeq n^n e^{-n} \sqrt{2\pi n}$$

Now

$$\binom{n}{k} p^k q^{n-k} = \frac{n!}{(n-k)k!} p^k q^{n-k} =$$

$$\frac{n^n e^{-n}\sqrt{2\pi n}}{(n-k)^{n-k} e^{-(n-k)}\sqrt{2\pi(n-k)}\, k^k e^{-k}\sqrt{2\pi k}} p^k q^{n-k} =$$

$$\frac{1}{\sqrt{2\pi}} \frac{n^n \sqrt{n}}{(n-k)^{n-k}\sqrt{n-k}\, k^k \sqrt{k}} p^k q^{n-k} =$$

$$\sqrt{\frac{n}{2\pi(n-k)k}} \frac{n^n}{(n-k)^{n-k} k^k} p^k q^{n-k} =$$

$$\sqrt{\frac{n}{2\pi(n-k)k}} \frac{n^{n-k} n^k}{(n-k)^{n-k} k^k} p^k q^{n-k} =$$

$$\sqrt{\frac{n}{2\pi(n-k)k}} \left(\frac{np}{k}\right)^k \left(\frac{nq}{n-k}\right)^{n-k} =$$

$$\sqrt{\frac{n}{2\pi(n-k)k}} \left(\frac{k}{np}\right)^{-k} \left(\frac{n-k}{nq}\right)^{-(n-k)}$$

Define

$$x = \frac{k-np}{\sqrt{npq}}$$

then

$$1 + x\sqrt{\frac{q}{np}} = 1 + \frac{k-np}{\sqrt{npq}}\sqrt{\frac{q}{np}} = 1 + \frac{k-np}{np} = \frac{k}{np}$$

$$1 - x\sqrt{\frac{p}{nq}} = 1 - \frac{k-np}{\sqrt{npq}}\sqrt{\frac{p}{nq}} = 1 - \frac{k-np}{nq} =$$

$$\frac{nq-(k-np)}{nq} = \frac{nq-k+n(1-q)}{nq} = \frac{n-k}{nq}$$

and then

$$\binom{n}{k} p^k q^{n-k} \simeq$$

$$\sqrt{\frac{n}{2\pi(n-k)k}} \left(1 + x\sqrt{\frac{q}{np}}\right)^{-k} \left(1 - x\sqrt{\frac{p}{nq}}\right)^{-(n-k)} =$$

$$\sqrt{\frac{n}{2\pi(n-k)k}} \sqrt{\frac{n^{-2}}{n^{-2}}} \left(1 + x\sqrt{\frac{q}{np}}\right)^{-k} \left(1 - x\sqrt{\frac{p}{nq}}\right)^{-(n-k)} =$$

$$\sqrt{\frac{n^{-1}}{2\pi(n-k)kn^{-2}}} \left(1 + x\sqrt{\frac{q}{np}}\right)^{-k} \left(1 - x\sqrt{\frac{p}{nq}}\right)^{-(n-k)} =$$

$$\sqrt{\frac{1}{2\pi n \frac{n-k}{n} \frac{k}{n}}} \left(1 + x\sqrt{\frac{q}{np}}\right)^{-k} \left(1 - x\sqrt{\frac{p}{nq}}\right)^{-(n-k)} =$$

$$\sqrt{\frac{1}{2\pi n(1-\frac{k}{n})\frac{k}{n}}} \left(1 + x\sqrt{\frac{q}{np}}\right)^{-k} \left(1 - x\sqrt{\frac{p}{nq}}\right)^{-(n-k)} \simeq$$

$$\sqrt{\frac{1}{2\pi npq}} \left(1 + x\sqrt{\frac{q}{np}}\right)^{-k} \left(1 - x\sqrt{\frac{p}{nq}}\right)^{-(n-k)}$$

and so we have the first term in the expression in the theorem.

Next

$$\left(1 + x\sqrt{\frac{q}{np}}\right)^{-k} \left(1 - x\sqrt{\frac{p}{nq}}\right)^{-(n-k)} =$$

$$\exp\left\{\log\left[\left(1 + x\sqrt{\frac{q}{np}}\right)^{-k} \left(1 - x\sqrt{\frac{p}{nq}}\right)^{-(n-k)}\right]\right\} =$$

$$\exp\left\{(-k)\log\left(1 + x\sqrt{\frac{q}{np}}\right) - (n-k)\log\left(1 - x\sqrt{\frac{p}{nq}}\right)\right\} =$$

$$\exp\left\{(-np + x\sqrt{npq})\log\left(1 + x\sqrt{\frac{q}{np}}\right) - (nq - x\sqrt{npq})\log\left(1 - x\sqrt{\frac{p}{nq}}\right)\right\}$$

where the last expression follows from the definition of x.

Next we will use the Taylor expansion of $\log(1 \pm x)$, which says if x is close to 0 then $\log(1 \pm x) \approx \pm x - x^2/2$

so

$$\exp\left\{(-np + x\sqrt{npq})\log\left(1 + x\sqrt{\tfrac{q}{np}}\right) - (nq - x\sqrt{npq})\log\left(1 - x\sqrt{\tfrac{p}{nq}}\right)\right\} \simeq$$

$$\exp\left\{(-np + x\sqrt{npq})\left(x\sqrt{\tfrac{q}{np}} - \tfrac{x^2 q}{2np}\right) - (nq - x\sqrt{npq})\left(-x\sqrt{\tfrac{p}{nq}} - \tfrac{x^2 p}{2nq}\right)\right\} = *$$

Now

$$(-np + x\sqrt{npq})\left(x\sqrt{\tfrac{q}{np}} - \tfrac{x^2 q}{2np}\right) = -npx\sqrt{\tfrac{q}{np}} + np\tfrac{x^2 q}{2np} + x\sqrt{npq}\,x\sqrt{\tfrac{q}{np}} - x\sqrt{npq}\,\tfrac{x^2 q}{2np} =$$

$$-\sqrt{npq}\,x + \tfrac{1}{2}x^2 q + x^2 q - \tfrac{x^3 q}{2\sqrt{np}} = -\sqrt{npq}\,x - \tfrac{1}{2}x^2 q - \tfrac{x^3 q}{2\sqrt{np}}$$

and

$$(nq - x\sqrt{npq})\left(-x\sqrt{\tfrac{p}{nq}} - \tfrac{x^2 p}{2nq}\right) = -nqx\sqrt{\tfrac{p}{nq}} - nq\tfrac{x^2 p}{2nq} + x\sqrt{npq}\,x\sqrt{\tfrac{p}{nq}} + x\sqrt{npq}\,\tfrac{x^2 p}{2nq} =$$

$$-\sqrt{npq}\,x - \tfrac{1}{2}x^2 p + x^2 p + \tfrac{x^3 p}{2\sqrt{nq}} = -\sqrt{npq}\,x + \tfrac{1}{2}x^2 p + \tfrac{x^3 p}{2\sqrt{nq}}$$

so

$$* = \exp\left\{-\sqrt{npq}\,x - \tfrac{1}{2}x^2 q - \tfrac{x^3 q}{2\sqrt{np}} + \sqrt{npq}\,x - \tfrac{1}{2}x^2 p - \tfrac{x^3 p}{2\sqrt{nq}}\right\} =$$

$$\exp\left\{-\tfrac{1}{2}x^2 q - \tfrac{x^3 q}{2\sqrt{np}} - \tfrac{1}{2}x^2 p - \tfrac{x^3 p}{2\sqrt{nq}}\right\} =$$

$$\exp\left\{-\tfrac{1}{2}x^2(q+p) - \tfrac{x^3 q}{2\sqrt{np}} - \tfrac{x^3 p}{2\sqrt{nq}}\right\} =$$

$$\exp\left\{-\tfrac{1}{2}x^2 - \tfrac{x^3 q}{2\sqrt{np}} - \tfrac{x^3 p}{2\sqrt{nq}}\right\} \simeq$$

$$\exp\{-\tfrac{1}{2}x^2\} = \exp\left\{-\tfrac{1}{2}\left(\tfrac{k-np}{\sqrt{npq}}\right)^2\right\} = \exp\left\{-\tfrac{1}{2}\tfrac{(k-np)^2}{npq}\right\}$$

where we use the fact that as $n \to \infty$ $x \to 0$.

Finally we can show that

$$\frac{S_n - np}{\sqrt{np(1-p)}} \to Z$$

in distribution. For this let

$$y \in R, n \geq 1 \quad k(y) = \lfloor np + y\sqrt{npq} \rfloor$$

$$\sum_{j=0}^{k(y)} \binom{n}{j} p^j q^{n-j} = \sum_{j=0}^{k(y)} \frac{1}{\sqrt{2\pi npq}} \exp\left\{-\frac{1}{2}\frac{(j-np)^2}{npq}\right\} =$$

$$\sum_{j=0}^{k(y)} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(j-np)^2}{npq}\right\} \frac{1}{\sqrt{npq}} =$$

$$\sum_{j=0}^{k(y)} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(j-np)^2}{npq}\right\} \left(\frac{j+1}{\sqrt{npq}} - \frac{j}{\sqrt{npq}}\right) \simeq$$

$$\int_{-\infty}^{k(y)} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(x-np)^2}{npq}\right\} dx =$$

$$\int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx$$

and we are done!

Notice that the proof here is still incomplete in two ways: first we did not discuss the remainder term of the Taylor polynomial and second we should have been more precise about the as $n \to \infty$ $x \to 0$ part. For a truly rigorous proof see for example the following derivation by Steven Dunbar

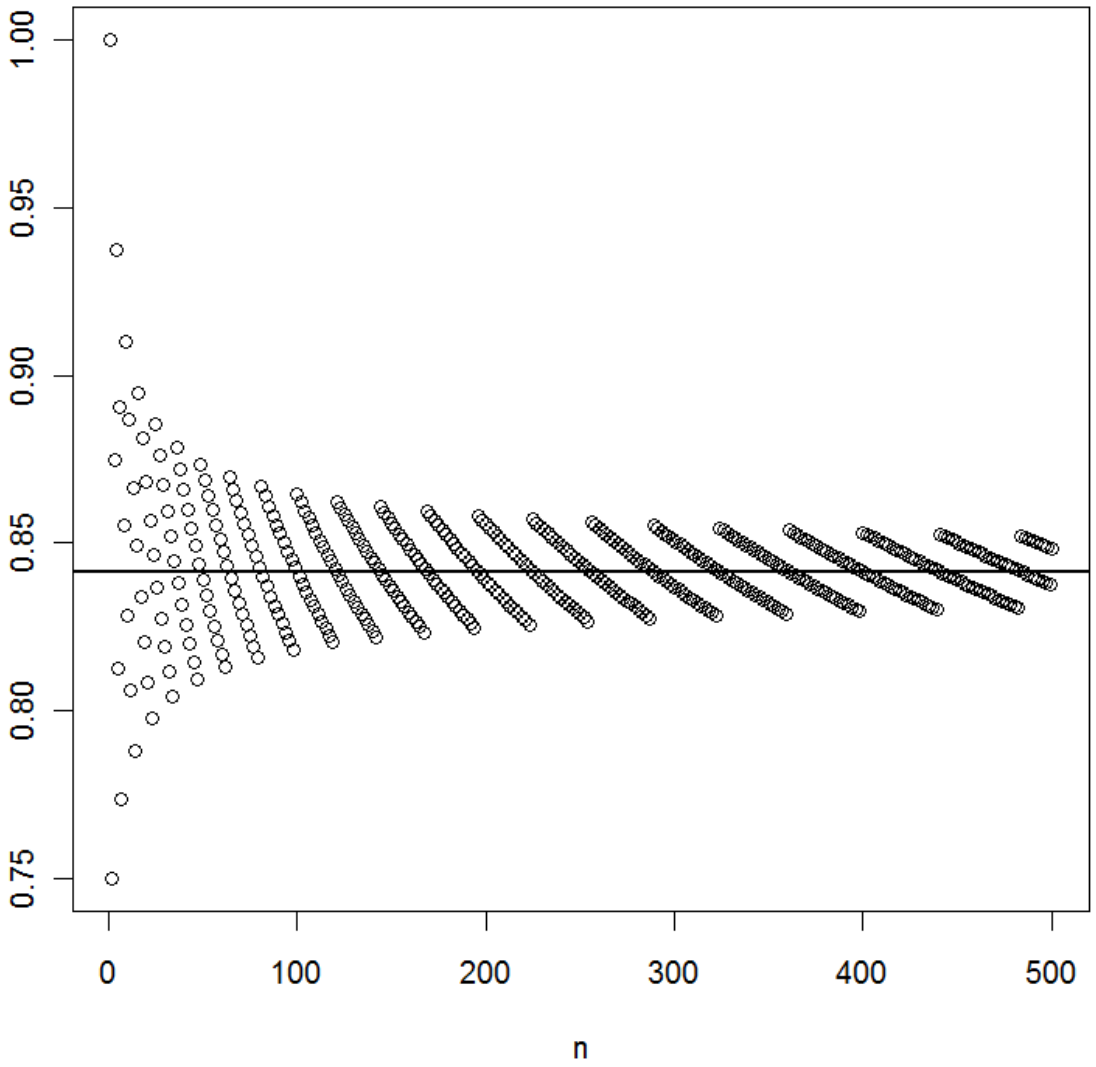
3.3.0.1 Example

For example say p=0.5 and x=1.0, then

$$P\left(\frac{\sum_{i=1}^{n} X - np}{\sqrt{np(1-p)}} \leq x\right) = P\left(\frac{\sum_{i=1}^{n} X - n/2}{\sqrt{n}/2} \leq x\right) =$$

$$P\left(\sum_{i=1}^{n} X \leq \frac{\sqrt{n}}{2}x + \frac{n}{2}\right) = \sum_{k=0}^{\frac{\sqrt{n}}{2}x + \frac{n}{2}} \binom{n}{k} 2^{-n}$$

the following graph shows these probabilities together with $\Phi(1)$ for n=1:1:500:



Here is the most basic version of a general CLT:

Theorem (Liapunov 1901)

$\{X_n\}$ are independent and identically distributed with mean μ and standard deviation σ. Moreover the mgf of X_n exists in an open neighborhood of 0. Then

$$P(\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \leq x) \to \Phi(x) \text{ as } n \to \infty$$

for all $x \in \mathbb{R}$.

proof

We will show that the mgf's of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converge to the mgf of a standard normal rv.

Let $Y_n = (X_n - \mu)/\sigma$, then

$\sqrt{n}(\bar{X}_n - \mu)/\sigma = 1/\sqrt{n}\sum Y_i$

so

$$\psi_{\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma})}(t) = E[e^{t\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma})}] = E[\exp(t\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i)] =$$
$$\prod_{i=1}^{n} E[\exp(\frac{t}{\sqrt{n}}Y_i)] = \prod_{i=1}^{n}\psi_{Y_i}(\frac{t}{\sqrt{n}}) = \left(\psi_Y(\frac{t}{\sqrt{n}})\right)^n$$

We now expand this into a Taylor series:

$$\psi_Y(t/\sqrt{n}) = \sum_{i=0}^{\infty}\psi_Y^{(k)}(0)\frac{(t/\sqrt{n})^k}{k!} =$$
$$1 + 0 \cdot (t/\sqrt{n})/1! + 1 \cdot (t/\sqrt{n})^2/2! + R(t/\sqrt{n}) = 1 + \frac{t^2}{2n} + R(t/\sqrt{n})$$

because EY^0=E1=1, EY^1=0 and EY^2=1.

An application of Taylor's theorem shows the remainder term

$nR(t/\sqrt{n})$ goes to 0 as $n \to \infty$. So

$$(\psi_Y(t/\sqrt{n}))^n = \left(1 + \frac{t^2}{2n} + R(t/\sqrt{n})\right)^n = \left(1 + \frac{t^2/2 + nR(t/\sqrt{n})}{n}\right)^n \to e^{t^2/2}$$

where we use a well-known lemma from real analysis: if $a_n \to a$, then $(1 + a_n/n)^n \to e^a$.

3.3.0.2 Example

Maybe the most important quantity in Statistics is the **sample mean** $\bar{X} = 1/n\sum X_i$. Here is an example: say the ages of people in a town have some distribution with mean 31.37 and standard deviation 12.34. If we randomly select a person, what is the probability that person is over 35 years old?

We have a rv X with $\mu = 31.37$ and $\sigma = 12.34$. We want P(X>35.0) but we don't know the density of X, so there is no way to do this.

Let's say we could sample 25 people, what is the probability that their mean age is over 35? Now we want

$P(\bar{X} > 35.0)$

and we have

$$P(\overline{X} > 35) = P(\sqrt{25}\,\frac{\overline{X}-31.37}{12.34} > \sqrt{25}\,\frac{35-31.37}{12.34}) \simeq$$
$$P(Z > 1.47) = 1 - \Phi(1.47) = 1 - 0.93 = 0.07$$

3.3.0.3 Example

Say we want to do a mail survey, that is we send letters with questionnaires to randomly selected people and hope they fill it out and send it back. From long experience it is known that such surveys have a "return rate" of about 25%, that is only 1 in 4 people send their survey back. How many surveys do we need to send out to be 99% sure to get more than 100 back?

Say we send out n questionnaires. Let the rv X be the number of questionnaires we get back, then $X \sim Bin(n, 0.25)$. We need to solve the equation P(X>100) = 0.99.

How do we find n? Note that

$\mu_X = np = 0.25n$ and
$\sigma_X = \sqrt{npq} = \sqrt{n0.25 \times 0.75} = 0.433\sqrt{n}$

and so $X \sim N(0.25n, 0.433\sqrt{n})$

We need n such that

$0.99 = P(X > 100) = 1 - P(X \leq 100)$

or

$P(X \leq 100) = 0.01$

so

$$0.01 = P(X < 100) = P(\frac{X - 0.25n}{0.433\sqrt{n}} < \frac{100 - 0.25n}{0.433\sqrt{n}}) = P(Z < \frac{100 - 0.25n}{0.433\sqrt{n}})$$

and so

$(100 - 0.25n)/(0.433\sqrt{n}) = \Phi^{-1}(0.01) = -2.326$

now:

$$\frac{100-0.25n}{0.433\sqrt{n}} = -2.326$$

$$100 - 0.25n = -2.326 \cdot 0.433\sqrt{n} = -1.0072\sqrt{n}$$

$$(100 - 0.25n)^2 = (-1.0072\sqrt{n})^2 = 1.0144n$$

$$100^2 - 2 \cdot 100 \cdot 0.25n + (0.25n)^2 = 1.0144n$$

$$10000 - 50n + 0.0625n^2 - 1.0144n = 0$$

$$0.0625n^2 - 51.0144n + 10000 = 0$$

$$n = \frac{-b \pm \sqrt{b^2-4ac}}{2a} = \frac{-(-51.0144) \pm \sqrt{(-51.0144)^2-4 \cdot 0.0625 \cdot 10000}}{2 \cdot 0.0625} = \frac{51.0144 \pm 10.12}{0.125}$$

which gives either n=(51.0144-10.12)/0.125=327 or n=(51.0144+10.12)/0.125=489.

So the quadratic equation gives us two possible solutions, so let's check which one is right. We find

$\Phi((327 - 100)/0.25) = 0.9906$
$\Phi((489 - 100)/0.25) = 0.0103$

so we see n=489 is the correct answer.

This solution is quite general. Say this company sends out questionnaires all the time, but with different return rates p, different desired number of returns m and a different probability p_m of at least m returns. Repeating the above calculation for this general case we find

$$n = \frac{2mp + k^2pq + \sqrt{(2mp + k^2pq)^2 - 4p^2m^2}}{2p^2}$$

As we saw above, the CLT is really a family of theorems, all with the same conclusion but with different assumptions. In fact, there are probably a 1000 different CLT's! Here is what is probably the most famous of them:

Theorem (Lindeberg-Feller 1922)

let X_n be independent random variables with $E[X_n]=0$ and $Var(X_n)=\sigma_n^2 < \infty$. Let

$$S_n = \sum_{i=1}^{n} X_i$$

let

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2$$

and

$$\Lambda_n(\varepsilon) = \sum_{i=1}^{n} E\left[\left(\frac{X_i}{s_n} \right)^2 I(|X_i| \geq \varepsilon s_n) \right]$$

then if

$\Lambda_n(\epsilon) \to 0$ as $n \to \infty$ for any $\epsilon > 0$, S_n/s_n converges to a standard normal in distribution.

Note: The condition on $\Lambda_n(\epsilon)$ of the theorem is known as the *Lindeberg condition*. Feller showed that it is in some sense not only necessary but also sufficient. In that sense it is the ultimate CLT for independent rv's.

3.3.0.4 Example

Say Y_1 , Y_2 ,..iid with mean μ and sd σ. Set $X_i = Y_i - \mu$. Now

$$s_n^2 = \sum_{i=1}^n \sigma_i^2 = \sum_{i=1}^n \sigma^2 = n\sigma^2$$

$$E\left[\left(\frac{X_i}{s_n}\right)^2 I(|X_i| \geq \varepsilon s_n) \right] = E\left[\left(\frac{X_i}{\sqrt{n}\,\sigma}\right)^2 I(|X_i| \geq \varepsilon \sqrt{n}\,\sigma) \right] =$$

$$\frac{1}{n\sigma^2} E[X_i^2 I(|X_i| \geq \varepsilon \sqrt{n}\,\sigma)]$$

$$\Lambda_n(\varepsilon) = \sum_{i=1}^n \frac{1}{n\sigma^2} E[X_i^2 I(|X_i| \geq \varepsilon \sqrt{n}\,\sigma)] = \frac{1}{\sigma^2} E[X_1^2 I(|X_1| \geq \varepsilon \sqrt{n}\,\sigma)]$$

$$\sigma^2 = E[X_1^2] = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_{|x|<\varepsilon\sigma\sqrt{n}} x^2 f(x)dx + \int_{|x|\geq\varepsilon\sigma\sqrt{n}} x^2 f(x)dx$$

$$= E[X_1^2 I(|X_1| < \varepsilon \sqrt{n}\,\sigma)] + E[X_1^2 I(|X_1| \geq \varepsilon \sqrt{n}\,\sigma)]$$

but the left term converges to σ^2, so the right term has to converge to 0.

3.3.0.5 Example

The CLT has found applications in just about any field of mathematics or science. Here is an application in number theory:

Erdos-Kac CLT

Say we pick an integer at random from {1,2,..,n}. Then the integer has about $\log\log(n) + \Phi(\sqrt{\log\log(n)})$ prime divisors.

In all approximation theorems like the central limit theorem a major issue is always how good the approximation is for finite n, that is in a specific case how far we still are from the limit. The following theorem gives some answers:

Theorem (Berry-Esseen)

Let $X_1, X_2, ..$ be iid rv with $E[X_1] = 0, Var(X_1) = \sigma^2$ and $E[|X_1|^3] = \rho < \infty$, then if F_n is the cdf of $\sqrt{n}S_n/\sigma$ we have

$$\sup_x |F_n(x) - \Phi(x)| \leq \frac{C\rho}{\sigma^3\sqrt{n}}$$

Calculated values of the constant C have decreased markedly over the years, from the original value of 7.59 by Esseen (1942), to 0.7882 by van Beek (1972), then 0.7655 by Shiganov (1986), then 0.7056 by Shevtsova (2007), then 0.7005 by Shevtsova (2008), then 0.5894 by Tyurin (2009), then 0.5129 by Korolev & Shevtsova (2009), then 0.4785 by Tyurin (2010). The detailed review can be found in the papers Korolev & Shevtsova (2009), Korolev & Shevtsova (2010). The best estimate as of 2012 is C=0.4748.

3.3.0.6 Example

say $Z_i \sim Ber(p)$, and let $X_i = Z_i/p - 1$, then

$$E[X_i] = E[Z_i/p - 1] = E[Z_i]/p - 1 = p/p - 1 = 0$$

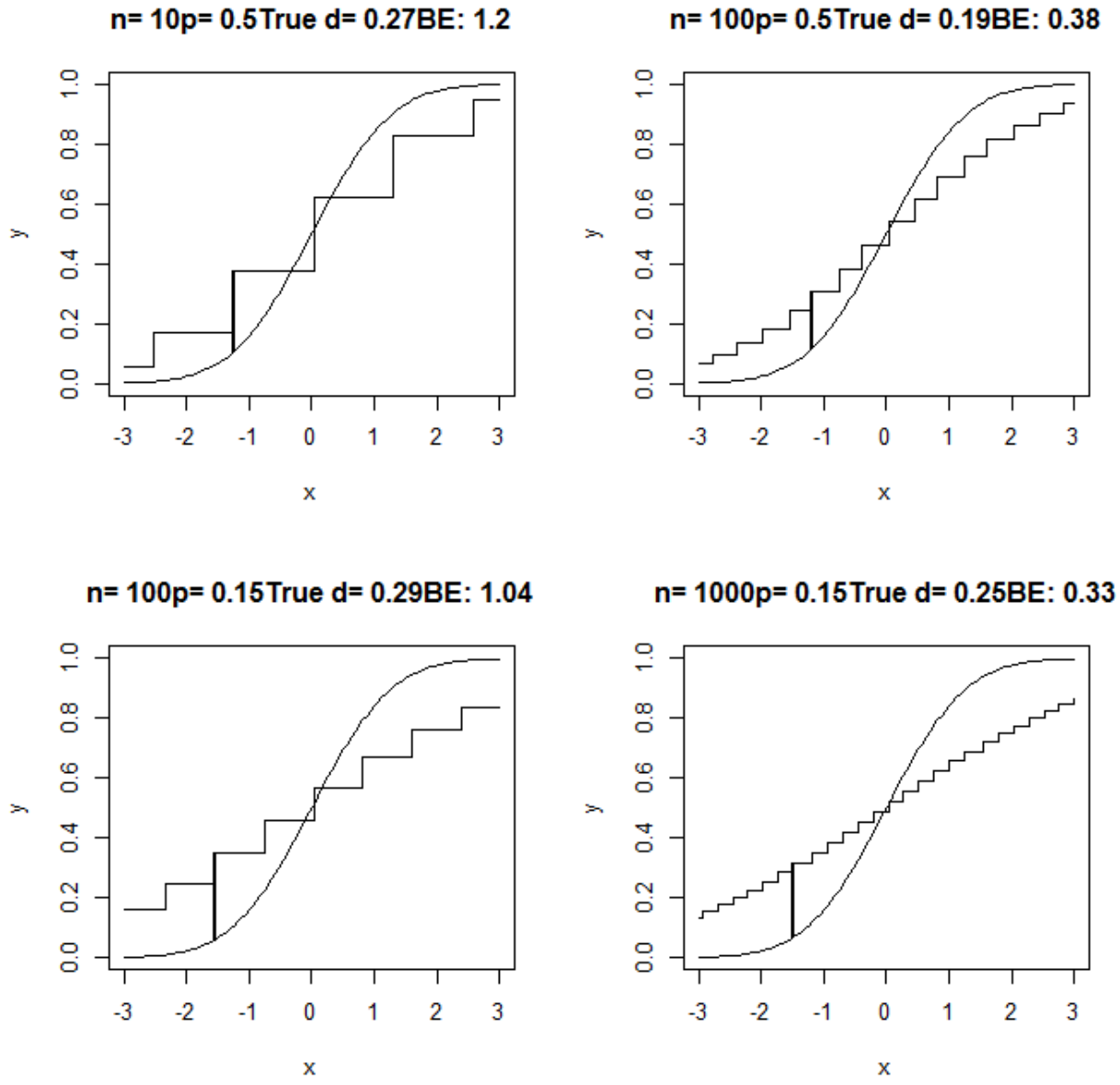
$$E[X_i^2] = E[(Z_i/p - 1)^2] = \frac{1}{p^2}E[(Z_i - p)^2] =$$
$$\frac{1}{p^2}[(0-p)^2(1-p) + (1-p)^2 p] = \frac{1-p}{p}[p + 1 - p] = \frac{1-p}{p}$$

$$E[|X_i|^3] = E[|Z_i/p - 1|^3] = \frac{1}{p^3}E[|Z_i - p|^3] =$$
$$\frac{1}{p^3}[|0-p|^3(1-p) + |1-p|^3 p] = \frac{1-p}{p^2}[p^2 + (1-p)^2]$$

$$\frac{C\rho}{\sigma^3} = \frac{0.4748\frac{1-p}{p^2}[p^2+(1-p)^2]}{\left(\frac{1-p}{p}\right)^{3/2}} = \frac{0.4748(p^2+(1-p)^2)}{\sqrt{p(1-p)}}$$

If p=1/2 the bound is 0.4748. As p gets close to 0, or 1 the bound goes to ∞.

Here are four examples:



3.4 Law of the Iterated Logarithm

Let $X_1, X_2, ..$ be iid rv's with mean 0 and std 1. Let $S_n = X_1 + .. + X_n$. As we saw before, by the strong law of large numbers we have $S_n/n \to 0$ almost surely, and in the last section we saw that by the central limit theorem we have $S_n/\sqrt{n}$ converges in distribution to a standard normal random variable. So in some sense S_n/n "squeezes" down to one point whereas $S_n/\sqrt{n}$ "spreads out", roughly between -3 and 3. It is a reasonable question then whether there is an in-between case, namely a sequence $\{a_n\}$ such that

$\sqrt{n} < a_n < n$

and S_n/a_n converges to something between a constant and a distribution. The answer is

given by

Theorem (Law of the Iterated Logarithm, Kolmogorov 1929)

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{n \log\log n}} = \sqrt{2} \text{ a.s.}$$

by symmetry the corresponding liminf is $-\sqrt{2}$, so this sequence oscillates between $\pm\sqrt{2}$.

proof (outline)

If the X_i's are Bernoulli rvs, we have a random walk and one can show that $P(S_n = 0) = 1/\sqrt{\pi n}$.

A similar argument (again starting with Sterling's formula) can be used to show that if n+k is even

$$P(S_n = k) \approx \sqrt{\frac{2}{\pi n}} \exp\{-\frac{k^2}{2n}\}$$

next with some arithmetic one can show that there exists a c>0 such that

$$P(S_n \geq k) \geq c\frac{\sqrt{n}}{k} \exp\{-\frac{k^2}{2n}\}$$

and finally on applies the Borel-Cantelli lemma to show that for any $\epsilon > 0$

$$S_n \leq \sqrt{2n \log\log n}\,(1+\varepsilon) \text{ a.s.}$$

3.4.0.1 Example

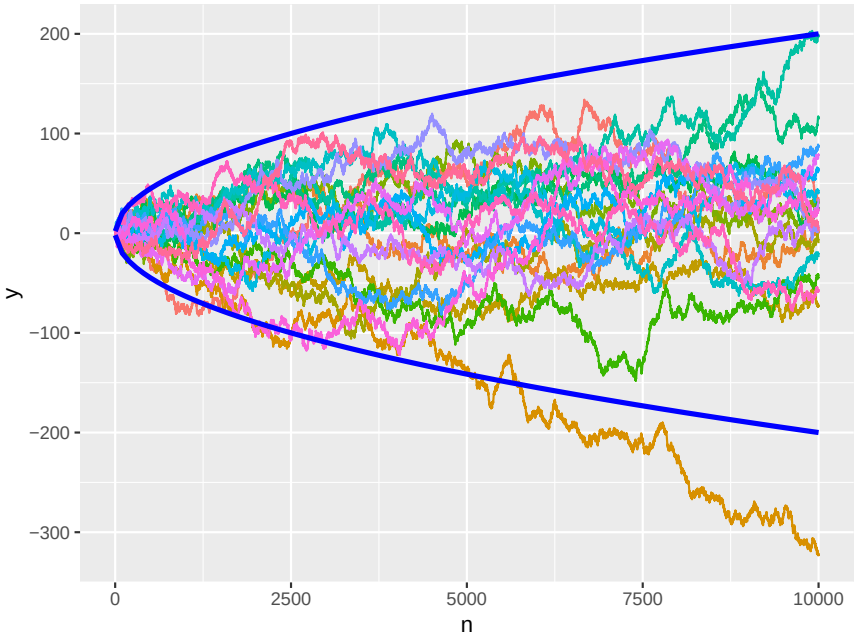
let ~Y_1 , Y_2 , ..~ be iid Ber(1/2), then E[Y_i]=1/2 and Var(Y_i)=1/4. Let X_i = (Y~i -1/2)/(1/2) = 2Y_i -1, then E[X~i']=0 and Var(X_i)=1. Let $S_n = X_1 + .. + X_n$.

The following graph has 100 simulated sequences with n=10000 and the four different "normalizations"

```
K <- 20
df <- data.frame(n=rep(1:1e4, K),
                 y=0*K*1e4,
                 z=rep("1", K*1e4))
for(i in 1:K) {
  df$y[((i-1)*1e4+1):(i*1e4)] <-
    cumsum(2*sample(0:1, size=1e4, replace = TRUE)-1)
  df$z[((i-1)*1e4+1):(i*1e4)] <- rep(paste(i), 1e4)
}
df$z <- factor(df$z)
```

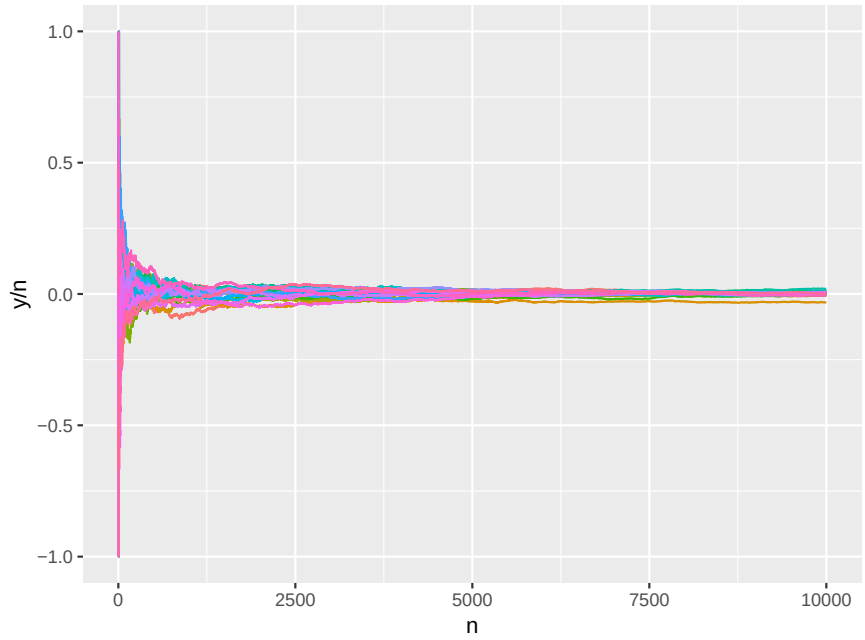
1. S_n

```
ggplot(data=df, aes(n, y, color=z)) +
  geom_line(size=0.5) +
  theme(legend.position = "none") +
    stat_function(fun = function(x) 2*sqrt(x),
                  size=1.2, color="blue") +
    stat_function(fun = function(x) -2*sqrt(x),
                  size=1.2, color="blue")
```



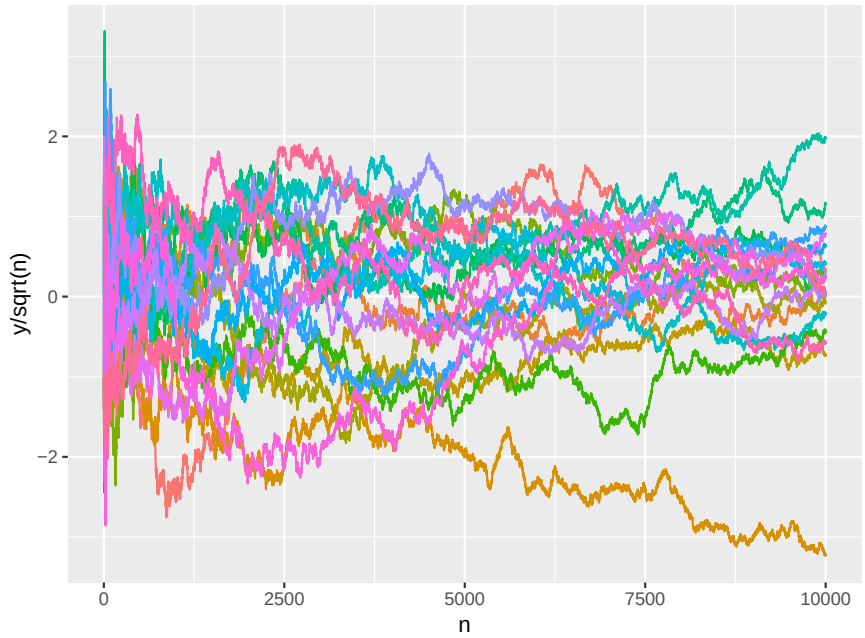
2. S_n /n

```
ggplot(data=df, aes(n, y/n, color=z)) +
  geom_line() + theme(legend.position = "none")
```



3. $S_n/\sqrt{n}$

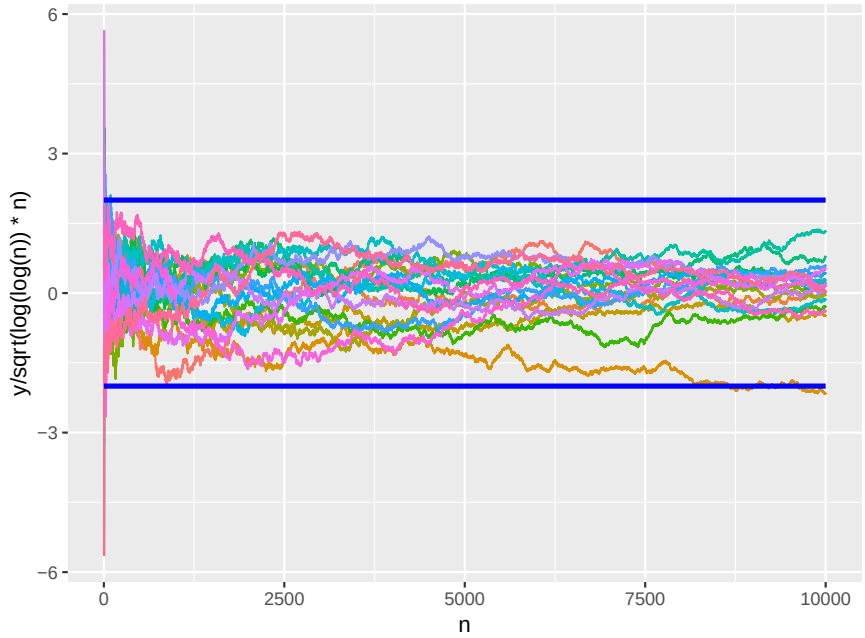
```
ggplot(data=df, aes(n, y/sqrt(n), color=z))  +
  geom_line(size=0.5) +
  theme(legend.position = "none")
```



4. $S_n/\sqrt{n \log\log n}$

```
ggplot(data=df, aes(n, y/sqrt(log(log(n))*n), color=z)) +
  geom_line(size=0.5) +
  theme(legend.position = "none") +
```

```
    stat_function(fun = function(x) 2,
                  size=1.2, color="blue") +
    stat_function(fun = function(x) -2,
                  size=1.2, color="blue")
```



3.5 Approximation Methods

3.5.1 Approximations based on Taylor's Theorem

Say we have a r.v. X with density f, a function h and we want to know Var(h(X)). Of course by definition we have

$$Eh(X) = \int_{-\infty}^{\infty} h(x)f(x)dx$$

$$Eh^2(X) = \int_{-\infty}^{\infty} h^2(x)f(x)dx$$

$$Vh(X) = Eh^2(X) - (Eh(X))^2$$

but sometimes these integrals (sums) are very difficult to evaluate. In this section we discuss some methods for approximating the variance.

Recall: If a function h(x) has derivatives of order r, that is if $g^{(r)}(x)$ exists, then for any

constant a the *Taylor polynomial* of order r is defined by

[](graphs/prob516.png>

One of the most famous theorems in mathematics called Taylor's theorem states that the remainder of the approximation h(x)-T_r(x) goes to 0 faster than the highest order term:

Taylor's theorem

$$\lim_{x \to a} \frac{h(x) - T_r(x)}{(x-a)^r} = 0$$

There are various formulas for the remainder term, but we won't need them here.

3.5.1.1 Example

say h(x) = log(x+1) and we want to approximate h at x=0. Then we have

$$h(x)|_{x=0} = h(0) = 0 \qquad h'(x)|_{x=0} = \frac{1}{x+1}|_{x=0} = 1$$

$$h''(x)|_{x=0} = -\frac{1}{(x+1)^2}|_{x=0} = -1 \qquad h^{(3)}(x)|_{x=0} = \frac{2}{(x+1)^3}|_{x=0} = 2$$

$$h^{(r)}(x)|_{x=0} = (-1)^{r+1}\frac{(r-1)!}{(x+1)^r}|_{x=0} = (-1)^{r+1}(r-1)!$$

$$T_0(x) = h(0) = 0$$

$$T_1(x) = T_0(x) + h'(0) \cdot (x-0) = 0 + 1 \cdot (x-0) = x$$

$$T_2(x) = T_1(x) + \frac{h''(0)}{2}(x-0)^2 = x + \frac{-1}{2} \cdot (x-0)^2 = x - x^2/2$$

$$T_3(x) = T_2(x) + \frac{h^{(3)}(0)}{6}(x-0)^3 = x - x^2/2 + \frac{2}{6}(x-0)^3 =$$

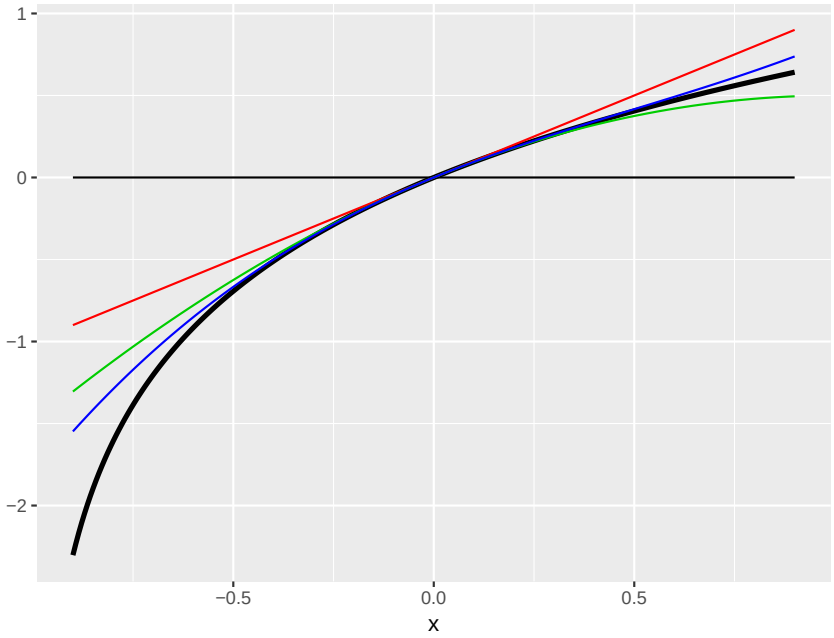
$$x - x^2/2 + x^3/3$$

The approximation is illustrated here:

```
a <- 0; r <- 3
x <- seq(-0.9, 0.9, length = 250)
h <- rep(0, r+1)
h[1] <- log(a+1)
for (n in 1:r) h[n+1] <- (-1)^(n+1)/n/(a+1)^n
y <- matrix(0, 250, r+1)
y[, 1] <- rep(log(a+1), 250)
for (k in 1:r) y[ , k+1] <- y[, k] + h[k+1]*(x-a)^k
```

```
df <- data.frame(x=x, ly=log(x+1))
plt <- ggplot(df, aes(x, ly)) +
        geom_line(size=1.2) + ylab("")

for (k in 1:(r + 1))
  plt <- plt +
    geom_line(data=data.frame(x=x, y=y[, k]), aes(x, y), color=k)
plt
```



One application of this is the

3.5.2 Delta Method

Let Y_n be a sequence of rv's that satisfies

$$\sqrt{n}(Y_n - \theta) \to N(0, \sigma)$$

in distribution. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exist and and is not 0. Then

$$\sqrt{n}\left(g(Y_n) - g(\theta)\right) \to N(0, \sigma g'(\theta))$$

proof

the Taylor expansion of $g(Y_n)$ around $Y_n = \theta$ is

$g(Y_n) = g(\theta) + g'(\theta)(Y_n - \theta) + R$

where $R \to 0$ as $Y_n \to \theta$. Now

$\sqrt{n}[g(Y_n) - g(\theta)] =$

$\sqrt{n}[g(\theta) + g'(\theta)(Y_n - \theta) + R - g(\theta)] =$

$g'(\theta)\sqrt{n}(Y_n - \theta) + \sqrt{n}R \to g'(\theta)X$

where $X \sim N(0, \sigma)$

3.5.2.1 Example

say $X_1, .., X_n$ iid Exp(1), so EX=VarX=1, then by the CLT

$\sqrt{n}(\bar{X} - 1) \to N(0, 1)$

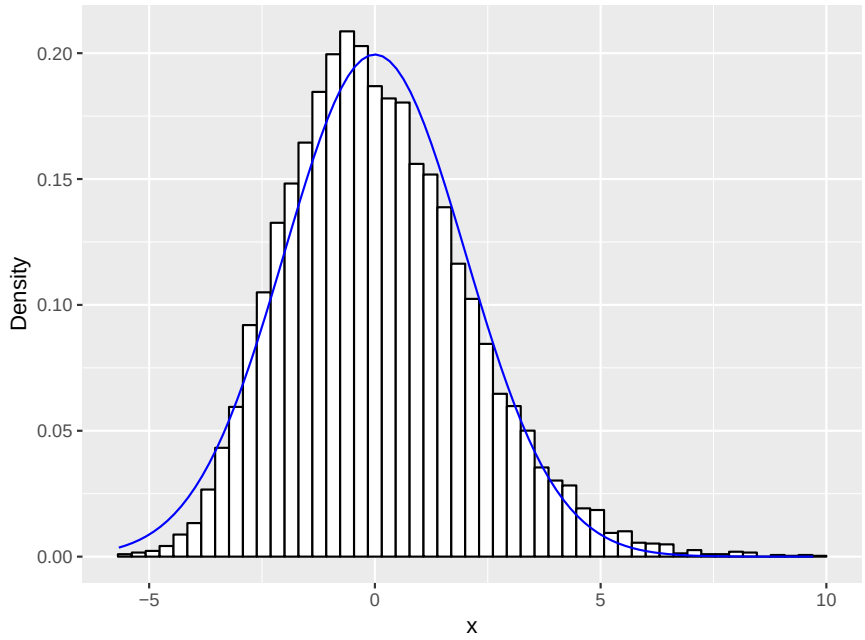
Let $g(x)=x^p$, so $g'(x)=px^{p-1}$ and by the delta method

$\sqrt{n}(\bar{X}^p - 1) \to N(0, p)$

```
n <- 100; p <- 2; B <- 10000
x <-  matrix(rexp(n*B, 1), ncol=n)
xbar <- apply(x, 1, mean)
print(mean(xbar))
```

```
## [1] 0.9986222
```

```
y <- sqrt(n)*(xbar^p-1)
bw <- diff(range(y))/50
df <- data.frame(x=y)
ggplot(df, aes(x)) +
  geom_histogram(aes(y = ..density..),
    color = "black",
    fill = "white",
    binwidth = bw) +
    labs(x = "x", y = "Density") +
  stat_function(fun = dnorm,
                colour = "blue",
                args=list(mean=0, sd=p))
```



say we have a sequence of iid rv's $X_1, .., X_n$, each with mean $\mu \neq 0$ and standard deviation σ. We know from the law of large numbers that

$\sqrt{n}(\bar{X} - \mu) \to N(0, \sigma)$

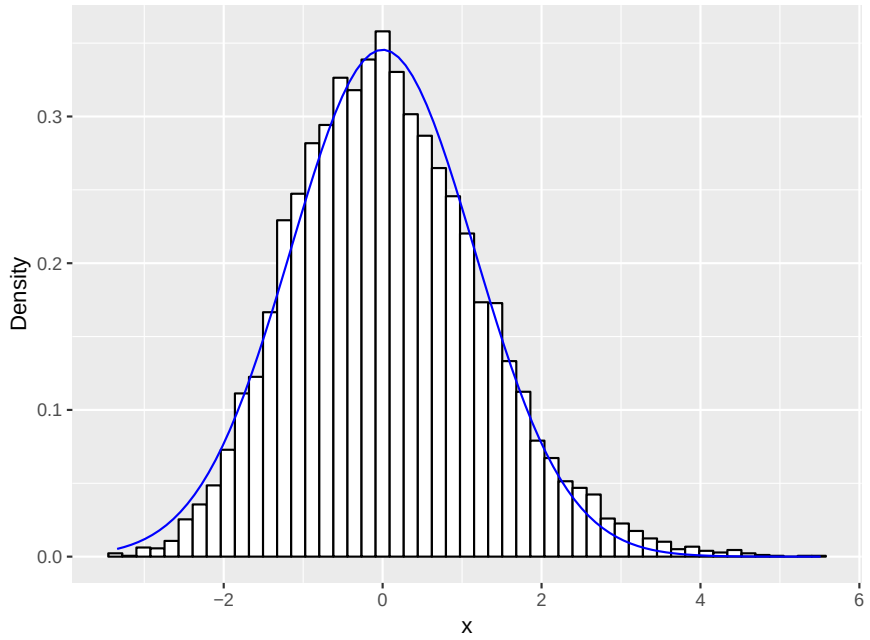
Now let g(x)=1/x, then g'(x)=-1/x^2 and we get

$\sqrt{n}(1/\bar{X} - 1/\mu) \to N(0, \sigma/\mu^2)$

say for example $X_i \sim U[0, 1]$, then $\mu = 1/2$ and $\sigma = 1/\sqrt{12}$, so according to the delta method

$\sqrt{n}(1/\bar{X} - 1/\mu) \sim N(0, \sigma/\mu^2) = N(0, 2/\sqrt{3})$

```
n <- 100; B <- 10000
x <-  matrix(runif(n*B), ncol=n)
xbar <- apply(x, 1, mean)
y <- sqrt(n)*(1/xbar - 1/0.5)
bw <- diff(range(y))/50
df <- data.frame(x=y)
ggplot(df, aes(x)) +
  geom_histogram(aes(y = ..density..),
    color = "black",
    fill = "white",
    binwidth = bw) +
    labs(x = "x", y = "Density") +
  stat_function(fun = dnorm,
                colour = "blue",
                args=list(mean=0, sd=2/sqrt(3)))
```



For our purposes we will need only first-order approximations (that is using the first derivative) but we will need a multivariate extension as follows: say X_1, ..,X_n are r.v. with means $\mu_1, .., \mu_n$ and define $\mathbf{X}$=(X_1, ..,X_n) and $\mu = (\mu_1, .., \mu_n)$. Suppose there is a differentiable function h($\mathbf{X}$) for which we want an approximate estimate of the variance. Define

$$h_i'(\boldsymbol{\mu}) = \frac{\partial}{\partial t_i} h(\mathbf{t})\Big|_{t_1=\mu_1,..,t_n=\mu_n}$$

The first order Taylor expansion of h about μ is

$$h(\mathbf{t}) = h(\boldsymbol{\mu}) + \sum_{i=1}^{n} h_i'(\boldsymbol{\mu})(t_i - \mu_i) + \text{Remainder}$$

Forgetting about the remainder we have

$$Eh(\mathbf{X}) \approx E\left[h(\boldsymbol{\mu}) + \sum_{i=1}^{n} h_i'(\boldsymbol{\mu})(\mathbf{X}_i - \mu_i) \right] = h(\boldsymbol{\mu}) + \sum_{i=1}^{n} h_i'(\boldsymbol{\mu})(E\mathbf{X}_i - \mu_i) = h(\boldsymbol{\mu})$$

and

$$Vh(\mathbf{X}) \approx E\Big[(h(\mathbf{X}) - h(\boldsymbol{\mu}))^2\Big] \approx E\Big[\Big(\sum_{i=1}^n h_i'(\boldsymbol{\mu})(\mathbf{X}_i - \mu_i)\Big)^2\Big]$$
$$= E\Big[\sum_{i,j=1}^n \Big(h_i'(\boldsymbol{\mu})(\mathbf{X}_i - \mu_i) \cdot h_j'(\boldsymbol{\mu})(\mathbf{X}_j - \mu_j)\Big)\Big] =$$
$$E\Big[\sum_{i=1}^n (h_i'(\boldsymbol{\mu}))^2(\mathbf{X}_i - \mu_i)^2\Big] + 2 \cdot E\Big[\sum_{i<j}^n \Big(h_i'(\boldsymbol{\mu})(\mathbf{X}_i - \mu_i) \cdot h_j'(\boldsymbol{\mu})(\mathbf{X}_j - \mu_j)\Big)\Big] =$$
$$\sum_{i=1}^n (h_i'(\boldsymbol{\mu}))^2 E(\mathbf{X}_i - \mu_i)^2 + 2 \cdot \sum_{i<j}^n \Big(h_i'(\boldsymbol{\mu})h_j'(\boldsymbol{\mu})E[(\mathbf{X}_i - \mu_i)(\mathbf{X}_j - \mu_j)]\Big) =$$
$$\sum_{i=1}^n (h_i'(\boldsymbol{\mu}))^2 V\mathbf{X}_i + 2 \cdot \sum_{i<j}^n h_i'(\boldsymbol{\mu})h_j'(\boldsymbol{\mu})Cov(\mathbf{X}_i, \mathbf{X}_j)$$

3.5.2.2 Example

Say we have just one rv X, then the formula simplifies to

$$V[h(X)] \approx [h'(\mu)]^2 V[X]$$

say $X \sim N(\mu, 1)$ with μ large enough so that P(X>0)=1. We want to find Var[log(X)]. Set h(x)=log(x), then h'(x)=1/x and

$$V[\log(X)] \approx [\frac{1}{\mu}]^2 1 = \frac{1}{\mu^2}$$

check with

```
var(log(rnorm(10000, 10)))
```

```
## [1] 0.01037173
```

3.5.2.3 Example

Say we have two rv's X and Y and $X \perp Y$, then the formula simplifies to

$$V[h(X,Y)] \approx h_x^2(\mu_x, \mu_Y)V(X) + h_y^2(\mu_x, \mu_Y)V(Y)$$

say X and Y have a geometric distribution with parameters p and r, respectively. We want to approximate the variance of $\sqrt{X^2 + Y^2}$

Now $\mu_X = 1/p, V[X] = (1-p)/p^2, \mu_Y = 1/r, V[Y] = (1-r)/r^2$

let $h(x,y) = \sqrt{x^2 + y^2}$, then

$$\frac{d}{dx}h(x,y) = h_x(x,y) = \frac{x}{\sqrt{x^2+y^2}}$$

$$\frac{d}{dy}h(x,y) = h_y(x,y) = \frac{y}{\sqrt{x^2+y^2}}$$

and so

$$V[\sqrt{X^2+Y^2}] \approx \left(\frac{\mu_x}{\sqrt{\mu_x^2+\mu_y^2}}\right)^2 V[X] + \left(\frac{\mu_Y}{\sqrt{\mu_x^2+\mu_y^2}}\right)^2 V[Y] =$$

$$\frac{\frac{1}{p^2}}{\frac{1}{p^2}+\frac{1}{r^2}}\frac{1-p}{p^2} + \frac{\frac{1}{r^2}}{\frac{1}{p^2}+\frac{1}{r^2}}\frac{1-r}{r^2} = \frac{1-p}{p^2(1+(p/r)^2)} + \frac{1-r}{r^2(1+(r/p)^2}$$

```
p <- 0.2; r <- 0.3; B <- 10000
x <- rgeom(B, p)+1
y <- rgeom(B, r)+1
round(c(var(sqrt(x^2+y^2)),
  (1-p)/(p^2*(1+(p/r)^2))+(1-r)/(r^2*(1+(r/p)^2))), 3)
```

```
## [1] 19.284 16.239
```

3.5.2.4 Example

say we have a sample X_1, ..,X_n from a Bernoulli r.v. with success parameter p, that is P(X=1)=p=1-P(X=0). One popular measure of the probability of winning a game is the odds p/(1-p). For example when you roll a fair die the odds of getting a six are (1/6)/(1-(1/6) = 1:5.

An obvious estimator for p is $\hat{p}$, the sample mean, or here the proportion of "successes" in the n trials. Then an obvious estimator for the odds is $\hat{p}/(1-\hat{p})$. The question is, what is the variance of this estimator?

First note that

$$\begin{aligned}
&Var[\hat{p}] == \\
&Var[1/n\sum X_i] = \\
&1/n^2\sum V[X_i] = \\
&1/nV[X_1] = \\
&p(1-p)/n
\end{aligned}$$

Using the above approximation we get the following: let h(p)=p/(1-p), so h'(p)=1/(1-p)^2 and

$$V\left(\frac{\hat{p}}{1-\hat{p}}\right) \approx [h'(p)]^2 V(\hat{p}) =$$

$$\left[\frac{1}{(1-p)^2}\right]^2 \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3}$$

```
p <- 0.25; n <- 25; B <- 10000
x <- matrix(rbinom(B, 1, p), ncol=n)
phat <- apply(x, 1, mean)
odds <- phat/(1 - phat)
round(c(var(odds), p/n/(1 - p)^3), 4)
```

```
## [1] 0.0301 0.0237
```

3.5.2.5 Example

let's consider the random vector with joint pdf

f(x,y) = 1, 0<x,y< 1

Say we want to find Var(X/Y). Of course $X, Y \sim U[0,1]$ and independent, so

E[X]=E[Y]=1/2
Var[X]=Var[Y]=1/12 and
Cov(X,Y)=0.

Then if we consider the function h(x,y) = x/y we have

$$\frac{\partial}{\partial x}h(x,y) = \frac{1}{y} \text{ and } \frac{\partial}{\partial y}h(x,y) = -\frac{x}{y^2}$$

$$Eh(X,Y) = E(X/Y) \approx \frac{\mu_X}{\mu_Y}$$

$$Vh(X,Y) = V(X/Y) \approx \left(\frac{1}{\mu_Y}\right)^2 V(X) + \left(-\frac{\mu_X}{\mu_Y^2}\right)^2 V(Y) + 2 \cdot \left(\frac{1}{\mu_Y}\right)\left(-\frac{\mu_X}{\mu_Y^2}\right) Cov(X,Y) =$$

$$\frac{1}{\mu_Y^2}V(X) + \frac{\mu_X^2}{\mu_Y^4}V(Y) - 2 \cdot \frac{\mu_X}{\mu_Y^3}Cov(X,Y)$$

and so

$$V(X/Y) = \frac{1}{\mu_Y^2}V(X) + \frac{\mu_X^2}{\mu_Y^4}V(Y) - 2\frac{\mu_X}{\mu_Y^3}Cov(X,Y) =$$

$$\frac{1}{(1/2)^2}\frac{1}{12} + \frac{(1/2)^2}{(1/2)^4}\frac{1}{12} - 0 = \frac{2}{3}$$

How good is this approximation?

```
var(runif(10000)/runif(10000))
```

```
## [1] 5022.664
```

shows that it is actually very bad! The reason is that occasionally the denominator is very small, so the ratio is very big. Let's change the problem a little: now

f(x,y) = 1, 1<x,y<2

that is $X, Y \sim U[1,2]$, so E[X]=E[Y]=3/2, Var[X]=Var[Y]=1/12 and Cov(X,Y)=0. Now

$$V(X/Y) = \frac{1}{\mu_Y^2}V(X) + \frac{\mu_X^2}{\mu_Y^4}V(Y) - 2\frac{\mu_X}{\mu_Y^3}Cov(X,Y) =$$

$$\frac{1}{(3/2)^2}\frac{1}{12} + \frac{(3/2)^2}{(3/2)^4}\frac{1}{12} - 0 = \frac{2}{27}$$

and this is actually quite good:

```
round(c(2/27, var(runif(10000,1,2)/runif(10000,1,2))), 4)
```

```
## [1] 0.0741 0.0848
```

Generally ratios are often trouble!

3.5.2.6 Example

let's consider the random vector with joint pdf

f(x,y) = 6x, 0<x<y< 1

Say we want to find Var(X/Y)

First we have

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_x^1 6xdy = 6xy|_x^1 = 6x(1-x) \quad 0 < x < 1$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dy = \int_0^y 6xdx = 3x^2|_0^y = 3y^2 \quad 0 < y < 1$$

so

$$EX = \int_0^1 x6x(1-x)dx = 2x^3 - \frac{3}{2}x^4|_0^1 = \frac{1}{2}$$

$$EY = \int_0^1 y3y^2dy = \frac{3}{4}y^3|_0^1 = \frac{3}{4}$$

$$EX^2 = \int_0^1 x^2 6x(1-x)dx = \frac{3}{2}x^4 - \frac{6}{5}x^5|_0^1 = \frac{3}{10} \text{ and so } V(X) = \frac{3}{10} - \left(\frac{1}{2}\right)^2 = \frac{1}{20}$$

$$EY^2 = \int_0^1 y^2 3y^2dy = \frac{3}{5}y^5|_0^1 = \frac{3}{5} \text{ and so } V(Y) = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{3}{80}$$

$$EXY = \int_0^1 \int_0^y xy6xdxdy = \int_0^1 (2x^3y|_0^y dy = \int_0^1 2y^4dy = \frac{2}{5}y^5|_0^1 = \frac{2}{5}$$

$$\text{and so } Cov(X,Y) = EXY - EX \cdot EY = \frac{2}{5} - \frac{1}{2} \cdot \frac{3}{4} = \frac{1}{40}$$

so

$$V(X/Y) = \frac{1}{(3/4)^2}\frac{1}{20} + \frac{(1/2)^2}{(3/4)^4}\frac{3}{80} - 2\frac{1/2}{(3/4)^3}\frac{1}{40} = 0.058$$

and this is quite good:

```
x <- rbeta(10000, 2, 2)
y <- runif(10000, x, 1)
round(var(x/y), 4)
```

```
## [1] 0.0555
```

3.5.3 Importance Sampling

3.5.3.1 Example

say we have a rv X geometric with p=0.5. We want to find P(log(X!)>50).

Let's try to solve this problem analytically. First, log(x!) is an increasing function of x, so there exists x_{50} such that

log(x!)>50 iff x>x_{50}

so that $P(\log(X!) > 50) = P(X \geq x_{50})$. Finding x_{50} analytically is hopeless, though. We can do it with R by trial and error, using *log(factorial(n))* for different values of n. We find n=22.5, so

$$P(X \geq 23) = 1 - P(X \leq 22) = 1 - \sum_{i=1}^{22} \left(\frac{1}{2}\right)^i = 1 - \left(\sum_{i=0}^{22} \left(\frac{1}{2}\right)^i - 1\right) = 2 - \frac{1-0.5^{23}}{1-0.5}$$

or about 2.38×10^{-7}

How about an R check? The problem with this is that the probability p we want to find is very small, so in a simple simulation we can expect the outcome of interest only about every 1 in 4.2 million runs. In order to get some reasonably good estimate we probably need to run the simulation with n=10^9.

Here is a strange idea: the problem is that our event of interest, log(X!)>50, is very rare, it almost never happens. Let's instead sample from a distribution Y which has large values much more often, so that log(Y!)>50 happens more often. For example, let's try Y geometric with p=0.05:

```
B <- 100000
y <- rgeom(B, 0.05)+1
sum(log(factorial(y))>50)/B
```

```
## [1] 0.3199
```

It seems P(log(Y!)>50)=0.32, but there is a problem, we get a warning from R: **value out of range in 'gammafn'**. The reason is that R calculates log(y!) by first calculating y! and then taking log, and it finds y! via the gamma function, but for y bigger than about 170 y! can no longer be found that way. But log(175!)=732.33, not so big at all. The routine logfac is based on on Stirling's' approximation:

$$n! \simeq \sqrt{2\pi}\, n^{n+\frac{1}{2}} e^{-n}$$

```
logfac <- function(n)
  ifelse(n<20, log(factorial(n)),
         0.918938533+(n+0.5)*log(n)-n+(1/12-1/(360*n^2))/n)
```

So P(log(Y!)>50)=0.35. But what good is that? I want X! Well:

$$P(\log(X!) > 50) = E[I_{[\log(X!)>50]}(X)] =$$

$$\sum_{x=1}^{\infty} I_{[\log(x!)>50]}(x) f_X(x) =$$

$$\sum_{x=1}^{\infty} I_{[\log(x!)>50]}(x) f_X(x) \cdot \frac{f_Y(x)}{f_Y(x)} =$$

$$\sum_{x=1}^{\infty} \left[I_{[\log(x!)>50]}(x) \frac{f_X(x)}{f_Y(x)} \right] f_Y(x) =$$

$$E\left[I_{[\log(x!)>50]}(Y) \frac{f_X(Y)}{f_Y(Y)} \right] \simeq$$

$$\frac{1}{n} \sum_{i=1}^{n} \left[I_{[\log(x!)>50]}(Y_i) \frac{f_X(Y_i)}{f_Y(Y_i)} \right]$$

so if we sample from Y and find the sum here we still get an estimate of the probability for X. This is done here:

```
y <- y[logfac(y) >= 50]
w <- dgeom(y-1, 0.5)/dgeom(y-1, 0.05)
sum(w)/B
```

```
## [1] 2.358689e-07
```

In general we have the following: Let X be a rv' with pdf f and and Y a rv' with pdf g. Say we want to find E[h(X)]. Then

$$E[h(X)] = \sum_{i=1}^{\infty} h(x_i) f(x_i) =$$

$$\sum_{i=1}^{\infty} h(x_i) f(x_i) \frac{g(x_i)}{g(x_i)} =$$

$$\sum_{i=1}^{\infty} \left[h(x_i) \frac{f(x_i)}{g(x_i)} \right] g(x_i) =$$

$$E\left[h(Y) \frac{f(Y)}{g(Y)} \right] = E[h(Y) w(Y)]$$

where

$$w(y) = \frac{f(y)}{g(y)}$$

are called the "weights"

Note this was done for discrete rv's but it works just as well for continuous ones.

Note how to choose Y? Obviously we need Y such that it can't happen that P(Y=x)>0 and P(X=x)=0. In general we should choose a Y with the same "support" as X, that is P(X=x)>0 iff P(Y=x)>0. It is not necessary to have a Y that "looks like" X. For example in the case above we could have choosen Y with pdf

$f_Y(x) = 6/(\pi^2 x^2)$

It is also a good idea to choose Y such that the event of interest (log(Y!)>50) happens about 50% of the time.

3.5.3.2 Example

say X, Y and Z have a standard normal distribution. Find P(|XYZ|>K), for example K=10.

Now there is no way to do this analytically, and again the probability is very small. So we will use IS with X', Y' and Z' generated from normal distributions with mean 0 and standard deviation s. For our case of K=10 s=3 works good. In general, for some K play around a bit to find a good s.

```
B <- 10000; s=3
x <- rnorm(B, 0, s)
y <- rnorm(B, 0, s)
```

```r
z <- rnorm(B, 0, s)
T <- abs(x*y*z)
I <- c(1:B)[T > 10]
print(length(I)/B)
```

```
## [1] 0.3497
```

```r
w <- dnorm(x[I])/dnorm(x[I], 0, s) *
     dnorm(y[I])/dnorm(y[I], 0, s) *
     dnorm(z[I])/dnorm(z[I], 0, s)
sum(w)/B
```

```
## [1] 0.0003887312
```

4 Statistics

4.1 Basic Concepts of Statistics

Statistics is of course a whole branch of Science all by itself. Here we will just introduce a few basic ideas.

Probability theory and Statistics have in common that both start with a probability model. Typically such models have *parameters*, for example the success probability p in a Bernoulli rv or the rate λ in an exponential distribution. In probability theory we then have problems such as: if p=0.2, what is the mean of the Bernoulli rv? In other words we assume we know the parameters and then ask questions about possible outcomes.

In Statistics it is exactly the other way around: we already have observed outcomes from a rv and we are asking what the parameters might be.

4.1.0.1 Example

Say a Bernoulli trial has been carried out 1000 times, and resulted in 345 successes and 655 failures. What can be said about p?

The answer seems obvious: a good guess for p should be

$\hat{p} = 345/1000 = 0.345$

but of course it may not be obvious in other problems. Moreover, even in this most simple of problems there are issues. For example, say we strongly suspect that p=0.4. Is this compatible with our experiment?

Let's do a little probability: Each Bernoulli trial is a rv $Y_i \sim$ Ber(p). Moreover we can assume (here) that the Bernoulli trials are independent, so the number of success is

$X = Y_1 + .. + Y_n \sim$ Bin(n,p)

with n=1000. So our probability model is

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

notice the notation here: f(x|p), reminiscent of our notation for conditional pdf's. This is intentional, because we want to think of this as

"Probability of some outcome given some value of the parameter"

In Statistics, though, the "unknown" is the parameter p, and we already know x=345, so we formally we can turn this around and write

$$L(p|x) = \binom{n}{x} p^x (1-p)^{n-x}$$

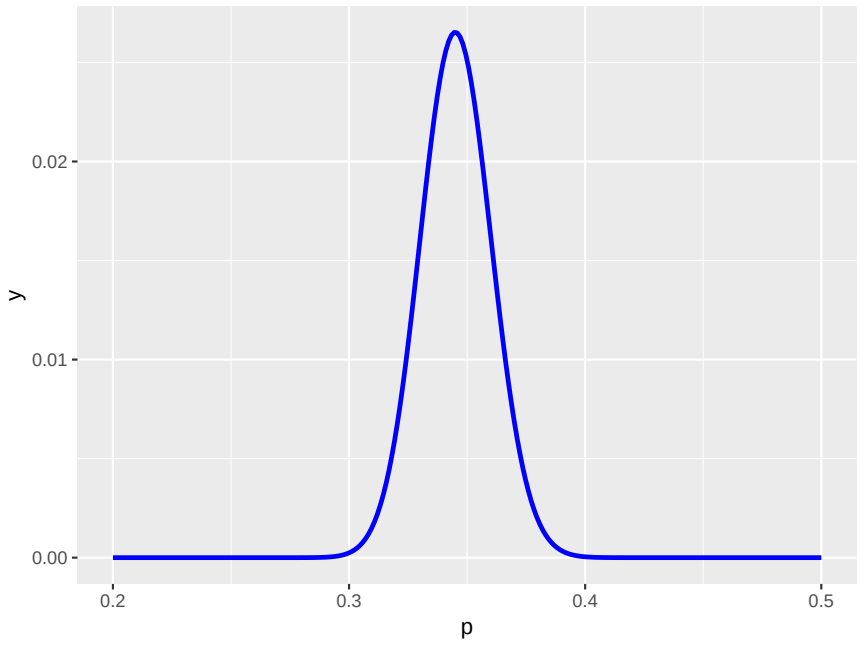
This is then called the *likelihood function*.

Notice that the right side is exactly the same, but we have shifted our point of view: instead of p being fixed and x being a variable, now x is fixed and p is a variable!

The likelihood function is the most fundamental entity in (almost) any statistical analysis.

What does it look like? Here is its graph:

```
p <- seq(0.2, 0.5, length=250)
y <- dbinom(345, 1000, p)
df <- data.frame(p=p, y=y)
ggplot(data=df, aes(p, y)) +
  geom_line(size=1.1, color="blue")
```



We see that L(p|345) has a unique maximum. This of course is the value of p that is *most likely* given the data. Let's find it analytically.

Clearly we want to find

argmax{L(p|x)}

so we could find

d/dp{L(p|x)}

It turns out that often it is easier to find

d/dp{log(L(p|x))}

but this is of course the same because

d/dp{log(L(p|x))} =
d/dp{L(p|x)}/L(p|x) = 0

iff d/dp{L(p|x)}=0

So now

$$\log(L(p|x)) = \log(\binom{n}{x}p^x(1-p)^{n-x}) =$$

$$\log\binom{n}{x} + x\log p + (n-x)\log(1-p)$$

$$\frac{d}{dp}\log(L(p|x)) = \frac{x}{p} - \frac{n-x}{1-p} = 0$$

$$x(1-p) = (n-x)p$$

$$p = \frac{x}{n} = \frac{345}{1000} = 0.345$$

and we found the same "obvious" answer!

The estimator $\hat{p}$ of p arrived at in this way is called the *maximum likelihood estimator*.

Here is a completely different solution to the same problem: We will now think of p as an unknown quantity. As with all things unknown we might already have some idea what p might be, just not exactly. It then makes sense to treat p as a rv, with some probability distribution. Of course p is a parameter, a fixed quantity and therefore not a random variable. It is our uncertainty regarding its value that makes it appear "random".

So, how should we model p? One thing is clear: $p \in [0,1]$, so maybe a Beta distribution might work. Now we have two random variables, and their distributions:

- $X \sim Bin(n,p)$
- $p \sim Beta(\alpha,\beta)$

Don't worry about the introduction of yet other parameters - α and β. We will talk about them soon.

Assuming that X and p are independent we can the find the joint pdf of (X,p):

$$f(x,p) = f(x|p)\pi(p) = \binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(a+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

The logic is this: π is what we knew about p before our experiment, f(x|p) is the outcome of the experiment. An obvious question then is: how has the experiment changed what we know about p? This should be "encoded" in the posterior distribution f(p|x):

$$f(p|x) = \frac{f(x,p)}{f_X(x)}$$

$$f_X(x) = \int_{-\infty}^{\infty} f(x,p)dp =$$

$$\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(a+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp =$$

$$\binom{n}{x} \frac{\Gamma(a+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^1 p^{x+\alpha-1}(1-p)^{n-x+\beta-1} dp =$$

$$\binom{n}{x} \frac{\Gamma(a+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+a+\beta)} \int_0^1 \frac{\Gamma(n+a+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} p^{x+\alpha-1}(1-p)^{n-x+\beta-1} dp =$$

$$\binom{n}{x} \frac{\Gamma(a+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+a+\beta)}$$

$$f(p|x) = \frac{f(x,p)}{f_X(x)} = \frac{\binom{n}{x} p^x (1-p)^{n-x} \frac{\Gamma(a+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}}{\binom{n}{x} \frac{\Gamma(a+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\alpha)\Gamma(n-x+\beta)}{\Gamma(n+a+\beta)}} =$$

$$\frac{\Gamma(n+a+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} p^{x+\alpha-1}(1-p)^{n-x+\beta-1}$$

so $p|X = x \sim Beta(x+\alpha, n-x+\beta)$.

Finally, to get an estimate of p we need to "extract" one value from this posterior distribution. One choice would be to use the mean:

$$\hat{p} = E[p|X = x] = \frac{x + \alpha}{n + \alpha + \beta}$$

Now, though, we need to say what α and β are. These are meant to "encode" what we knew about p before the experiment. Here are two examples:

- we knew absolutely nothing about p, any value between 0 and 1 was just as likely as any other. Then we might use

$p \sim U[0, 1] = Beta(1, 1)$

This type of prior is called *non-informative* because it does not make any one value of p more likely than any other. Now we find p = (x+1)(n+2) = 346/1002 = 0.3453.

- we were quite certain that p is somewhere between 0.3 and 0.5, with 0.4 the most likely and the pdf of p symmetric around 0.4.

This type of prior is called *subjective* because it encodes our subjective belief about p. Now E[p]=0.4 and sd(p)=0.2/4 = 0.05 (using the ballpark estimate range=4sd) and therefore

$$E[p] = \frac{\alpha}{\alpha+\beta} = 0.4$$

$$\alpha = 0.04(\alpha + \beta)$$

$$\alpha = \frac{0.4}{0.6}\beta = \frac{2\beta}{3}$$

$$Var(p) = \frac{\alpha\beta}{(\alpha+b)^2(\alpha+\beta+1)} = 0.05^2$$

$$\frac{\alpha\beta}{(\alpha+b)^2(\alpha+\beta+1)} = \frac{\frac{2\beta}{3}\beta}{(\frac{2\beta}{3}+b)^2(\frac{2\beta}{3}+\beta+1)} =$$

$$\frac{2\beta^2/3}{(5\beta/3)^2(5\beta+3)/3} = \frac{18}{25(5\beta+3)} = 0.05^2$$

$$18 = 0.05^2 \cdot 25(5\beta + 3)$$

$$5\beta + 3 = 288$$

$$\beta = 57, \ \alpha = 38$$

and then we get

$p = (x + \alpha)/(n + \alpha + \beta) = (345 + 38)/(1000 + 38 + 57) = 0.3498$

The two solutions outlined above are examples of the two fundamentally different approaches to Statistics we have today: the first one is based on a definition of probability as the long run relative frequency of an event, in fact in this example that is exactly what the maximum likelihood estimator is:

x/n = relative frequency of successes

This type of statistical analysis is called *Frequentist*. The second solution uses Bayes' theorem to combine a *prior distribution* on the parameter with the data likelihood to calculate the posterior distribution, and this approach is called *Bayesian Statistics*.

The essential difference between the two is not the use of Bayes' theorem but the use of a prior in Bayesian statistics.

4.1.0.2 Example

let's continue the discussion of the experiment above. So far we have found a *point estimates* for p, that is a single number we think is a good guess. But of course it is highly unlikely that we got the true value of p exactly right, for example if we were to repeat the experiment maybe next time we would see 376 successes, and our (frequentist) estimate would then be 0.376. Instead of a single best guess, maybe we should give a range of values (almost) certain to include the true p. This is done by quoting an *interval estimate*:

Frequentist Solution

Definition

A $100(1-\alpha)\%$ *confidence interval* for a parameter θ is an interval (L(x),U(x)) such that

$P(L(X) < \theta < U(X)) \geq 1 - \alpha$ for all θ

To find a confidence interval for our example we can make use of the central limit theorem:

$$X \sim Bin(n,p)$$

$$\frac{X-np}{\sqrt{npq}} \sim N(0,1)$$

$$1-\alpha = P(-z_{\alpha/2} < \frac{X-np}{\sqrt{npq}} < z_{\alpha/2}) \simeq$$

$$P(-z_{\alpha/2}\sqrt{npq} < X - np < z_{\alpha/2}\sqrt{npq}) =$$

$$P(X - z_{\alpha/2}\sqrt{npq} < np < X + z_{\alpha/2}\sqrt{npq}) =$$

$$P(\frac{X}{n} - z_{\alpha/2}\sqrt{\frac{pq}{n}} < p < \frac{X}{n} + z_{\alpha/2}\sqrt{\frac{pq}{n}}) \simeq$$

$$P(\frac{X}{n} - z_{\alpha/2}\sqrt{\frac{X/n(1-X/n)}{n}} < p < \frac{X}{n} + z_{\alpha/2}\sqrt{\frac{X/n(1-X/n)}{n}})$$

$$L(x) = \frac{x}{n} - z_{\alpha/2}\sqrt{\frac{x/n(1-x/n)}{n}}$$

$$U(x) = \frac{x}{n} + z_{\alpha/2}\sqrt{\frac{x/n(1-x/n)}{n}})$$

$$(1-a)100\% = 90\% \ \alpha = 0.1 \ z_{\alpha/2} = 1.645$$

$$L(x) = \frac{345}{1000} - 1.645\sqrt{\frac{\frac{345}{1000}(1-\frac{345}{1000})}{1000}} = 0.345 - 0.025 = 0.32$$

$$U(x) = \frac{345}{1000} + 1.645\sqrt{\frac{\frac{345}{1000}(1-\frac{345}{1000})}{1000}} = 0.345 + 0.025 = 0.37$$

and so a 90% confidence interval for p is (0.32,0.37)

Bayesian Solution

Definition

A $100(1-\alpha)\%$ **credible interval** for a parameter θ is an interval (L,U) such that

$P(L < \theta < U|X = x) = 1 - \alpha$

so again the answer comes from the posterior distribution p|X=x. In our case (using again

the Beta prior) we have found the posterior distribution to be

$p|X = x \sim Beta(x + \alpha, n - x + \beta)$

and so we need L and U such that

$P(L < p < U|X = x) = 1 - \alpha$

there are many such intervals, for example we could take L=0 and find U accordingly. One idea often used is to split up the probability α into $\alpha/2$ on the left and $\alpha/2$ on the right (actually that is just what we did in the frequentist solution as well) Then we get

$$\alpha/2 = P(p < L|X = x) = \int_0^L \frac{\Gamma(n+\alpha+\beta)}{\Gamma(x+\alpha)\Gamma(n-x+\beta)} t^{x+\alpha-1}(1-t)^{n-x+\beta-1} dt$$

and the corresponding equation for U. These are equations that have to be solved numerically:

```
x <- 345; n <- 1000; alpha <- 1; beta <- 1
round(qbeta(c(0.025, 0.975), x+alpha, n-x+beta), 3)
```

```
## [1] 0.316 0.375
```

```
alpha <- 38; beta <- 57
round(qbeta(c(0.025, 0.975), x+alpha, n-x+beta), 3)
```

```
## [1] 0.322 0.378
```

Notice that the credible interval using the Beta(1,1) prior is the same as the frequentist confidence interval. This happens quite often (but not always) when we use a non-informative prior. Also notice that although the Beta(1,1) prior is very different from the Beta(38,57), the resulting intervals are almost the same. This is what we expect (and hope) to see if there is a lot of data. To what degree the result of an analysis depends on the chosen prior is always an important consideration in Bayesian analysis, and is studied in what is called a sensitivity analysis.

5 Stochastic Processes

5.1 Introduction

Up to now when we had a sequence of random variables X_1, X_2, .. we assumed them to be independent. we have gone about as far as one can with this assumption. So from now on we will consider situations were the rv's are dependent.

There is exactly one way in which a collection of rv's can be independent, but there are infinitely many ways in which they can depend on one another. In order to make any progress we then need to say something about the dependence structure.

Definition

Any collection of random vectors $\{\mathbf{X_t}, t \in T\}$ is called a *stochastic process*. All the values that the random variables X_t can take on are called the *state space*. Because X_t is a regular

random variable (or vector) we again differentiate between continuous and discrete state space cases. Moreover T can be discrete (1,2,..) or continuous (t>0) as well.

5.1.0.1 Example

Let $X_i \sim U[0,1], i = 1, 2, ..; X_i \perp X_j$, then {X_i, i=1,2,..} is a continuous state space discrete time process.

5.1.0.2 Example

Let $X_i \in \{0, .., 39\}$ the position on the board of your token after i roles of the dice in a game of Monopoly. Then $\{X_i, i = 1, 2, ..\}$ is a discrete state space discrete time process.

5.1.0.3 Example

Let $P(Z_i = -1) = p = 1 - P(Z_i = 1), Z_i \perp Z_j$ if $i \neq j$. Let

$X_n = \sum_{i=1}^{n} Z_i$

then $X_n \in \{0, \pm 1, \pm 2, ..\}$ and so $\{X_n, n = 1, 2, ..\}$ is a discrete state space discrete time process. This is a very famous stochastic process called a *random walk.*

Here is a list of things we often want to know about a stochastic process: - what is the distribution of X_n, especially in the limit?
- what is EX_n, especially in the limit?
- what is cor(X~n, X_{n+k})?
- Do certain events ever occur, and if so with what probability? For example in the random walk, if we start at 0 what is the probability to reach 100?

5.2 Poisson Process

Definition

A stochastic process $\{N(t), t > 0\}$ is called a *counting process* if N(t) is the number of times an event occurred up to time t.

5.2.0.1 Example

Consider an ATM machine and let N(t) be the number of customers served by the ATM machine at time t.

Because of the way it is defined every counting process has the following properties:

1. $N(t) \geq 0$
2. N(t) is an integer
3. If s<t then $N(s) \leq N(t)$

4. If s<t then N(t)-N(s) is the number of events that have occurred in the interval (s,t].

Definition

Let $\{N(t); t \geq 0\}$ be a counting process. Then

a. $\{N(t); t \geq 0\}$ is said to have *independent increments* if the number of events that occur in disjoint intervals are independent.

b. $\{N(t); t \geq 0\}$ is said to have *stationary increments* if the distribution of the number of events that occur in any interval of time depend only on the length of the interval.

5.2.0.2 Example

The process of our ATM machine probably has independent increments but not stationary increments. Why?

The most important example of a counting process is the Poisson process. To define it we need the following notation, called *Landau's o symbol*:

Definition (Landau's o symbol)

a function f is said to be o(h) if

$$\lim_{h \to 0} \frac{f(h)}{h} = 0$$

5.2.0.3 Example

f(x)=x^2 is o(h) but f(x)=x is not.

Definition

A counting process $\{N(t); t \geq 0\}$ is said to be a Poisson process with rate $\lambda > 0$ if

1. N(0)=0
2. N(t) has stationary and independent increments
3. $P(N(h) = 1) = \lambda h + o(h)$
4. $P(N(h) \geq 2) = o(h)$

Notice that this implies that during a short time period the probability of an event occurring is proportional to the length of the interval and the probability of a second (or even more) events occurring is very small.

Theorem

Let $\{N(t); t \geq 0\}$ be a Poisson process, then

$$N(t + s) - N(s) \sim Pois(\lambda t)$$

for all $s \geq 0$.

proof

Let $p_n(t) = P(N(t)=n)$. Then

$$p_0(t+h) = P(N(t+h) = 0) =$$
$$P(N(t) = 0, N(t+h) - N(t) = 0) =$$
$$P(N(t) = 0) \cdot P(N(t+h) - N(t) = 0) =$$
$$p_0(t) \cdot (1 - \lambda h + o(h))$$

where the last equation follows from

$P(N(h) = 0) =$
$1 - P(N(h) \geq 1) = 1 - P(N(h) = 1) - P(N(h) \geq 2)$

Now

$$\frac{p_0(t+h)-p_0(t)}{h} = -\lambda p_0(t) + \frac{o(h)}{h}$$
$$\text{so } p_0'(t) = \lim_{h \to 0} \frac{p_0(t+h)-p_0(t)}{h} = -\lambda p_0(t) + \lim_{h \to 0} \frac{o(h)}{h} = -\lambda p_0(t)$$
$$\text{or } \frac{p_0'(t)}{p_0(t)} = -\lambda \text{ so } \int \frac{p_0'(t)}{p_0(t)} dt = \int -\lambda dt + c = -\lambda t + c$$
$$\text{but } \int \frac{p_0'(t)}{p_0(t)} dt = \log p_0(t) + c$$
$$\text{and so } \log p_0(t) = -\lambda t + c$$
$$0 = \log p_0(0) = -\lambda 0 + c = c \Rightarrow c = 0$$
$$\text{and so } p_0(t) = e^{-\lambda t}$$

The same basic idea works for the case p_n (t) as well to finish the proof.

Remark It is intuitively clear why the definition above should lead to the Poisson distribution. Take the interval (0,t] and subdivide it into k equal size intervals (0,t/k], (t/k, 2t/k] .. ((k-1)t/k,t]. The probability of 2 or more events in any one interval goes to 0 as k goes to ∞ because

P(2 or more events in any subinterval)
$\leq \sum$ P(2 or more events in the kˆth subinterval)
= ko(t/k)

$t \times o(t/k)/(t/k) \to 0$ as $k \to \infty$.

Hence N(t) will (with probability going to 1) just equal the number of subintervals in which an event occurs. However, by independent and stationary increments this number will have a binomial distribution with parameters k and $p = \lambda t + o(t/k)$. Hence by the Poisson approximation to the binomial we see that N(t) will have a Poisson distribution with rate λt.

5.2.0.4 Example

Suppose that N points are uniformly distributed over the interval (0, N). Let X be the number of points in (0,1). Find the pdf of X if N is large.

Let's try this directly first:

$$P(X = 0) = P(U_i > 1,\ i = 1,..,N) =$$

$$P(U_1 > 1)^N = \left[1 - \frac{1}{N}\right]^N \approx e^{-1}$$

$$P(X = 1) = P(U_n < 1, U_i > 1,\ i = 1,..,N, i \neq n) =$$

$$\sum_{n=1}^{N} P(U_n < 1, U_i > 1,\ i \neq n) =$$

$$\sum_{n=1}^{N} \frac{1}{N}(1 - \frac{1}{N})^{N-1} = (1 - \frac{1}{N})^{N-1} \approx \frac{e^{-1}}{1-\frac{1}{N}} = \frac{Ne^{-1}}{N-1}$$

$$P(X = 2)?$$

and this get's ugly fast. Instead consider the following: Let N(t) be the points in (0,t), then for t small (relative to N) $\{(N(t), t \geq 0\}$ will be a Poisson process with rate λ. Now

P(N(1)=0) = P(X=0) = e^{-1}, so $\lambda = 1$ and so

P(X=n) = P(N(1)=n) = $e^{-1/n}$!

Definition

Let T_1 be the time when the first event occurs, T_2 the time from the first event until the second event etc. The sequence T_1, T_2, .. is called the sequence of *interarrival times*.

Theorem

Let $\{N(t); t \geq 0\}$ be a Poisson process, and $\{T_i; i \geq 1\}$ be the interarrival times. Then

$T_1, T_2, .. \sim Exp(\lambda)$ and $T_i \perp T_j$

proof

Note that {T_1 >t} is equivalent to {no events occurred in (0,t]} and so

$$P(T_1 > t) = P(N(t) = 0) = e^{-\lambda t}$$
$$F_{T_1}(t) = 1 - e^{-\lambda t}$$
$$f_{T_1}(t) = \lambda e^{-\lambda t}$$

and we see that $T_1 \sim Exp(\lambda)$. But

$$P(T_2 > t | T_1 = s) = P(0 \text{ events in } (s, s+t] | T_1 = s) =$$
$$P(0 \text{ events in } (s, s+t]) = P(0 \text{ events in } (0, t]) = e^{-\lambda t}$$

because of independent and stationary increments. So we find that $T_2 \sim Exp(\lambda)$ and that $T_1 \perp T_2$. By induction it is clear that the sequence $\{T_n, n = 1, 2, ..\}$ is an iid sequence of exponential r.v. with mean $1/\lambda$.

Remark This result should not come as a surprise because the assumption of independent and stationary increments means that the process from any moment on is independent of all that occurred before and also has the same distribution as the process started at 0. In other words the process is memoryless, and we have previously shown that any continuous rv on $(0, \infty)$ with the memoryless property has to have an exponential distribution.

Definition

Let S_n be the arrival time of the n^{th} event. (This is also often called the *waiting time* until the n^{th} event).

Theorem

Let $\{N(t); t \geq 0\}$ be a Poisson process, and $\{S_n; n \geq 1\}$ be the waiting times. Then

$S_n \sim \Gamma(n, \lambda)$

proof

Clearly $S_n = \sum_{i=1}^{n} T_i$, and so we find

$S_n \sim \Gamma(n, \lambda)$

5.2.0.5 Example

Up to yesterday a store has 999,856 customers. They are planning to hold a little party when the 1,000,000$^{\text{th}}$ customer comes into the store. From experience they know that a customer arrives about every 4 minutes, and the store is open from 9am to 6pm. What is the probability that they will have the party today?

They will have the party if at least 144 customers come into the store today. Let's assume that the customers arrive according to a Poisson process with rate 4min (?), then we want the probability P(S_144 < 9*60).

Now $S_{144} \sim \Gamma(144, 4)$

and so

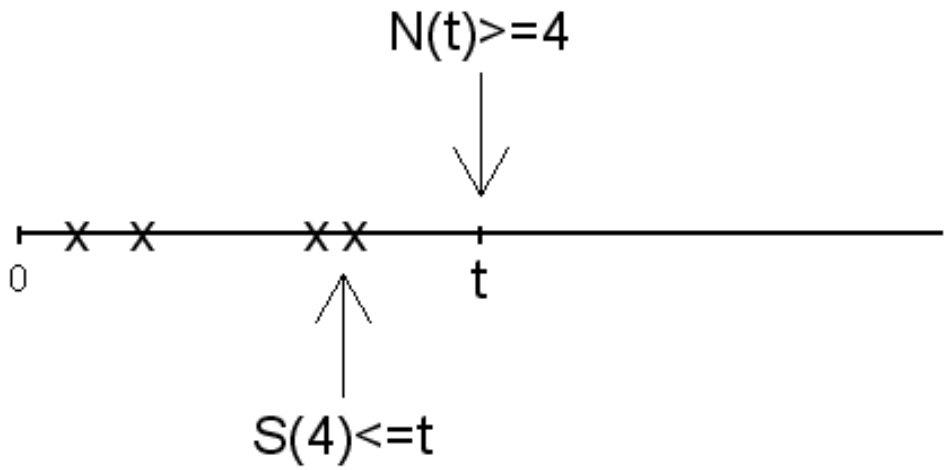
```
round(pgamma(9*60, 144, 1/4), 4)
```

```
## [1] 0.2302
```

Here is another proof of the last theorem. We use the fact that the n$^{\text{th}}$ event occurs at or before time t if and only if the number of events occurring by time t is at least n. So

$$N(t) \geq n \text{ iff } S_n \leq t$$

This is a very useful equivalence, and much more general than just for the Poisson process, so



here is an illustration:

With this we find

$$F_{S_n}(t) = P(S_n \leq t) = P(N(t) \geq n)$$

$$= \sum_{k=\mathbf{n}}^{\infty} P(N(t) = k) = \sum_{k=\mathbf{n}}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

$$\text{and so } f_{S_n}(t) = F'_{S_n}(t) = \sum_{k=\mathbf{n}}^{\infty} \left(\frac{\lambda^k k t^{k-1}}{k!} e^{-\lambda t} + \frac{(\lambda t)^k}{k!} (-\lambda) e^{-\lambda t} \right) =$$

$$\sum_{k=\mathbf{n}}^{\infty} \lambda \frac{(\lambda t)^{k-1}}{(k-1)!} e^{-\lambda t} - \lambda \sum_{k=\mathbf{n}}^{\infty} \frac{(\lambda t)^k}{k!} e^{-\lambda t} = \lambda \frac{(\lambda t)^{n-1}}{(n-1)!} e^{-\lambda t}$$

5.2.0.6 Example

Say that on any given day hundreds of cars pass through a certain intersection. Any one of them has a (hopefully small) probability of having an accident at that intersection. Let X(t) be the number of accidents in the t days, then is X(t) a Poisson process?

There are two problems with the assumptions of the Poisson process here:

- different days might have different numbers of cars going through (weekdays vs. weekends?)
- the probability of having an accident is probably very different for different cars.

The first problem might be handled by considering a different time-scale (accidents per week?), the second problem actually is not a problem at all:

let Z_1, Z_2, .. be independent Bernoulli rv's with $P(Z_i = 1) = p_i$. Let $S_n = Z_1 + .. + Z_n$. Then if $\lambda = p_1 + .. + p_n$ it can be shown that

$$\left| P(S_n = k) - \frac{\lambda^k}{k!} e^{-\lambda} \right| \leq \sum_{i=1}^{n} p_i^2$$

In the "classic" case where $p_1 = .. = p_n = p = \lambda/n$ we have

$$\left| P(S_n = k) - \frac{\lambda^k}{k!} e^{-\lambda} \right| \leq \frac{\lambda^2}{n}$$

and we see that this theorem not only gives us reason to think that the Poisson approximation works in the example above, it also provides a useful estimate of the error in the Poisson approximation to the Binomial.

We have seen previously that if

$U_1, .., U_n \sim U[0, 1]$

and independent, then $(U_{(1)},..,U_{(n)})$ has joint density $f(u_{(1)},..,u_{(n)})=n!$, $0<u_{(1)}< ..<u_{(n)}<1$.

Clearly if $U_1, .., U_n \sim U[0, t]$ and independent, then $(U_{(1)},..,U_{(n)})$ has joint density

$f(u_{(1)},..,u_{(n)})=n!/t^n$, $0<u_{(1)}< ..<u_{(n)}<t$

Let $W_i = U_{(i)}$. Now

Theorem

let W_1, W_2,.. be the arrival times in a Poisson process with rate λ. Then

$$f_{(W_1,..,W_n)|N(t)=n}(w_1, . . , w_n) = \frac{n!}{t^n} \quad 0 < w_1 <. . < w_n \leq t$$

In other words, conditional on the total number of arrivals the arrival times have the same distribution as the order statistic of a uniform.

5.2.0.7 Example

say $\{N(t), t \geq 0\}$ is a Poisson process with rate λ. Find the mean time of the first event, given that $N(1) = n, n \geq 1$.

$$E[W_1|N(1) = n]$$

$$\{(W_1, .., W_n)|N(1) = n\} =^d (U_1, .., U_n)$$

$$U_i \sim U[0, 1], \text{ indep}$$

$$P(W_1 > t|N(1) = n) =$$

$$P(W_1 > t, , .., W_n > t|N(1) = n) =$$

$$P(U_1 > t, .., U_n > t) = P(U_1 > t)^n =$$

$$[1 - P(U_1 < t)]^n = (1 - t)^n$$

$$E[W_1|N(1) = n] = \int_0^1 (1 - t)^n dt =$$

$$-\frac{1}{n+1}(1 - t)^{n+1}|_0^1 = \frac{1}{n+1}$$

5.2.0.8 Example

Customers arrive at a store according to a Poisson process of rate λ. Each customer pays \$1 on arrival, and we want to evaluate the expected value of the total sum collected during (0,t] discounted back to time 0. If the discount (inflation) rate is β, then this is given by

$$M = E\left[\sum_{k=1}^{N(t)} e^{-\beta W_k}\right]$$

Now

$$M = E\Big[\sum_{k=1}^{N(t)} e^{-\beta W_k}\Big] =$$

$$E\Big[\sum_{k=1}^{N(t)} e^{-\beta W_k} | N(t)\Big] =$$

$$\sum_{n=1}^{\infty} E\Big[\sum_{k=1}^{n} e^{-\beta W_k} | N(t) = n\Big] P(N(t) = n)$$

$$E\Big[\sum_{k=1}^{n} e^{-\beta W_k} | N(t) = n\Big] = E\Big[\sum_{k=1}^{n} e^{-\beta U_k}\Big] =$$

$$nE\big[e^{-\beta U_1}\big] = n\int_0^t e^{-\beta u}\frac{1}{t}du = \frac{n}{t}\big(-\frac{1}{\beta}e^{-\beta u}\big|_0^t =$$

$$\frac{n}{\beta t}(1 - e^{-\beta t})$$

$$M = \sum_{n=1}^{\infty} E\Big[\sum_{k=1}^{n} e^{-\beta W_k} | N(t) = n\Big] P(N(t) = n) =$$

$$\sum_{n=1}^{\infty} \frac{n}{\beta t}(1 - e^{-\beta t}) P(N(t) = n) =$$

$$\frac{1}{\beta t}(1 - e^{-\beta t}) \sum_{n=1}^{\infty} nP(N(t) = n) =$$

$$\frac{1}{\beta t}(1 - e^{-\beta t}) E[N(t)] = \frac{\lambda}{\beta}(1 - e^{-\beta t})$$

Theorem

Consider a Poisson process $\{(N(t), t \geq 0\}$ with rate λ, and suppose each time time an event occurs it is classified as either type I or II, with probabilities p and q=1-p, respectively, independent of anything else. Let $N_1(t)$ and $N_2(t)$ be the respective number of type I and II arrivals by time t, then $\{N_1(t), t \geq 0\}$ and $\{N_1(t), t \geq 0\}$ are both Poisson process with resp. rate $p\lambda$ and $(1-p)\lambda$. Furthermore the processes are independent.

proof

$$P(N_1(t) = n, N_2(t) = m) =$$
$$\sum_{k=0}^{\infty} P(N_1(t) = n, N_2(t) = m | N(t) = k) P(N(t) = k) =$$
$$P(N_1(t) = n, N_2(t) = m | N(t) = n+m) P(N(t) = n+m) =$$
$$P(N_1(t) = n, N_2(t) = m | N(t) = n+m) \frac{(\lambda t)^{n+m}}{(n+m)!} e^{-\lambda t} =$$
$$\binom{n+m}{m} p^n q^m \frac{(\lambda t)^{n+m}}{(n+m)!} e^{-\lambda t} =$$
$$\left(\frac{(\lambda p t)^{n+m}}{(n+m)!} e^{-\lambda p t} \right) \left(\frac{(\lambda q t)^{n+m}}{(n+m)!} e^{-\lambda q t} \right) =$$
$$P(N_1(t) = n) P(N_2(t) = m)$$

where the equation in the middle follows from the fact that if there were a total of n+m events the probability that n where of type I and m were of type II is just the binomial probability.

The theorem easily generalizes to r types.

5.2.0.9 Example

Customers arrive at a store according to a Poisson process with rate of 2 per hour. Each customer is a "Buyer" with probability 0.3 or a "Window-Shopper" with probability q=0.7. What is the probability of at least 1 sale during a 2 hour period?

P(at least 1 sales) =
$P(N_1(t) \geq 1) = 1 - P(N_1(t) = 0) = 1 - \exp(-2 * 2 * 0.3) = 1 - e^{-1.2} = 0.7$

5.2.0.10 Example (Coupon Collection Problem)

There are m different coupons. Each time a person collects a coupon it is, independently of those previously obtained, of type j with probability p_j. Let N denote the number of coupons one needs in order to have a complete collection of at least one of each type. Find E[N].

Let N_j be the number of coupons needed until we have one of type j, then $N = \max\{N_j; 1 \leq j \leq m\}$.

It is easy to see that $N_j \sim G(p_j)$, but they are not independent and so finding the distribution of their maximum is very difficult.

Let's assume that that coupons are collected according to a Poisson process with rate 1, and say an event is of type j if the coupon collected was of type j. If we let N_j (t) denote

the number of type j coupons collected by time t, then it follows that $\{N_j(t), t \geq 0\}$ are independent Poisson processes with rates p_j.

Let X_j denote the time of the first event of type j, and let $X = \max\{X_j; 1 \leq j \leq m\}$ be the time when we have all the coupons. Now the X_j are the waiting times of independent Poisson processes, so they have an exponential distributions and are independent, so

$$P(X < t) = P(\max_{i \leq j \leq m} X_j < t) =$$

$$P(X_1 < t, . . , X_m < t) = \prod_{j=1}^{m} P(X_1 < t) =$$

$$\prod_{j=1}^{m} (1 - e^{-p_j t})$$

$$E[X] = \int_0^\infty x f_X(x) dx = \int_0^\infty \int_0^x dt f_X(x) dx =$$

$$\int_0^\infty \int_x^\infty f_X(x) dx dt = \int_0^\infty P(X > t) dt =$$

$$\int_0^\infty \left[1 - \prod_{j=1}^{m} (1 - e^{-p_j t}) \right] dt$$

Now let T_i be the i^{th} interarrival time, that is the time between finding the $(i-1)^{st}$ and the i^{th} coupon. $X = \sum T_i$, but $T_i \sim Exp(1)$, and they are independent, so

$E[X|N] = E[\sum T_i|N] = NE[T_1|N] = N$

so

E[X]=E{E[X|N]}=E[N]

For example, say $p_1 = .. = p \sim n = p = 1/m$, then

$$E[N] = \int_0^\infty \left[1 - \prod_{i=1}^{m} (1 - e^{-t/m}) \right] dt =$$

$$\int_0^\infty [1 - (1 - e^{-t/m})^m] dt =$$

$$m \int_0^\infty [1 - (1 - e^{-x})^m] dx =$$

$$m \int_0^\infty \left[1 - \sum_{k=0}^{m} \binom{m}{k} (-e^{-x})^{m-k} \right] dx =$$

$$m \int_0^\infty \left[1 - \sum_{k=0}^{m-1} \binom{m}{k} (-e^{-x})^{m-k} - 1 \right] dx =$$

$$m \int_0^\infty \sum_{k=0}^{m-1} \binom{m}{k} (-e^{-x})^{m-k} dx =$$

$$m \sum_{k=0}^{m-1} \binom{m}{k} \int_0^\infty (-1)^{m-k} e^{-x(m-k)} dx =$$

$$m \sum_{k=0}^{m-1} \binom{m}{k} (-1)^{m-k} \int_0^\infty e^{-x(m-k)} dx =$$

$$m \sum_{k=0}^{m-1} \binom{m}{k} (-1)^{m-k} \frac{1}{-(m-k)} e^{-x(m-k)} \Big|_0^\infty =$$

$$m \sum_{k=0}^{m-1} \binom{m}{k} (-1)^{m-k} \frac{1}{-(m-k)} =$$

$$m \sum_{k=0}^{m-1} \binom{m}{k} \frac{(-1)^{m-k-1}}{m-k}$$

m	E[N]
10	3.00
10	5.50
10	8.33
10	11.42
10	14.70
10	18.15
10	21.74
10	25.46
10	29.29

What if m-1 have the same probability, but one is rarer, say only half of the probability of the others? So (wlog) $2p_1 == p_2 = .. = p_m = 1$, then p~i $= 1/(m-1/2)$ for $i \leq 2 \leq m$ and $p_1 = 1/[2(m-1/2)]$

$$E[N] = \int_0^\infty \left[1 - \prod_{i=1}^{m} (1 - e^{-p_i t}) \right] dt =$$

$$\int_0^\infty [1 - (1 - e^{-t/[2(m-0.5)]})(1 - e^{-t/(m-0.5)})^{m-1}] dt$$

this integral has to be found numerically, using some numerical integration method. We find

m	E[N]
10	3.00
10	5.50
10	8.33
10	11.42
10	14.70
10	18.15
10	21.74
10	25.46
10	29.29

Proposition: If $\{N_i(t); t \geq 0\}$, i=1,..,k represent the number of type i events occurring in (0,t] and if $P_i(t)$ is the probability that an event occurring at time t is of type i, then

$$E[N_1(t)] = \lambda \int_0^t P_i(s)ds$$

5.2.0.11 Example (HIV-Aids)

one of the difficulties in tracking the number of HIV infected people is its long incubation time, that is an infected person does not show any symptoms for a number of years, but is capable of infecting others.

Let us suppose that individuals contract HIV according to a Poisson process with unknown rate λ. Suppose that the incubation time until symptoms appear is a rv with cdf G, which is known, and suppose incubation times are independent. Let $N_1(t)$ be the number of individuals that have shown symptoms at time t, and let $N_2(t)$ be the number that have contracted HIV at time t but not yet shown symptoms. An individual that contracts HIV at time s will show symptoms at time t with probability G(t-s), so it follows from the above proposition that

$$E[N_1(t)] = \lambda \int_0^t G(t-s)ds = \lambda \int_0^t G(x)dx$$

$$E[N_2(t)] = \lambda \int_0^t 1 - G(t-s)ds = \lambda \int_0^t 1 - G(x)dx$$

say we know the number of individuals with system as time t is n_1 , then

$$E[N_1(t)] = \lambda \int_0^t G(x)dx \simeq n_1$$

$$\lambda \simeq \frac{n_1}{\int_0^t G(x)dx}$$

$$n_2 \simeq E[N_2(t)] = \lambda \int_0^t 1 - G(x)dx \simeq$$

$$n_1 \frac{\int_0^t 1-G(x)dx}{\int_0^t G(x)dx}$$

say G exponential rate μ then

$$G(x) = 1 - e^{-x/\mu}$$

$$n_2 \simeq n_1 \frac{\int_0^t 1-G(x)dx}{\int_0^t G(x)dx} = n_1 \frac{\int_0^t e^{-x/\mu}dx}{\int_0^t 1-e^{-x/\mu}dx} =$$

$$n_1 \frac{-\mu e^{-x/\mu}|_0^t}{x-\mu e^{-x/\mu}|_0^t} = n_1 \frac{\mu\left(1-e^{-t/\mu}\right)}{t-\mu(1-e^{-t/\mu})}$$

for example if t=16 years, $\mu = 10$ years and n_1 =220,000, then n_2 =219,00.

5.3 Markov Chains

5.3.1 Basic Definition

Definition

The sequence of r.v. X_1, X_2, .. is said to be a *Markov chain* if for any event A we have

$P(X_n \in A | X_1 = x_1, .., X_{n-1} = x_{n-1}) = P(X\ n \in A | X_{n-1} = x_{n-1})$

that is X_n depends only on X_{n-1} but not on any of the r.v. before it.

Clearly a Markov chain is a discrete-time space stochastic process. It can have either a discrete or continuous state space.

If we think of the index n as a time variable, then all that matters for the state of the system at time n is where it was at time n-1, but not on how it got to that state.

5.3.1.1 Example (Random Walk)

Say we flip a coin repeatedly. Let the random variable Y_i be 1 if the i^{th} flip is heads, -1 otherwise. Now let $X_n = \sum_{i=1}^{n} Y_i$.

Clearly we have

$$P(X_n = x_n | X_1 = x_1, .., X_{n-1} = x_{n-1}) = \begin{cases} 0.5 \text{ if } x_n = x_{n-1} - 1 \\ 0.5 \text{ if } x_n = x_{n-1} + 1 \\ 0 \text{ otherwise} \end{cases}$$

For a Markov chain all the relevant (probability) information is contained in the probability to get from state i to state j in k steps. For k=1 this is contained in the *transition matrix* P $= (p_{ij})$, and in fact as we shall see P is all we need.

5.3.1.2 Example (Random Walk, cont)

Here we have $p_{ij} = 1/2$ if |i-j|=1, 0 otherwise.

5.3.1.3 ** Example (Asymmetric Random Walk)**

As above the state space are the integers but now we go from i to i+1 with probability p, to i-1 with probability q and stay at i with probability 1-p-q.

5.3.1.4 Example (Ehrenfest chain)

Say we have two boxes, box 1 with k balls and box 2 with r-k balls. We pick one of the balls at random and move it to the other box. Let X_n be the number of balls in box 1 at time n.

First note that we have $X_n \in \{0, 1, .., r\}$. Now say X_n =k, so there are k balls in urn 1, therefore r-k balls in urn 2. In the next step we either move a ball from urn 1 to urn 2, or vice versa, so X_{n+1} =k+1 or X_{n+1} =k-1. Now

$p_{k,k+1}$ = P(X_{n+1} =k+1|X_n =k) =
P(move ball from urn 2 to urn 1 | k balls in urn 1) =
P(pick one of the r-k balls in urn 2) = (r-k)/r

also

$p_{k,k-1}$ = 1-$p_{k,k+1}$ = k/r and $p_{i,j}$ = 0 otherwise.

Ehrenfest used this model to study the exchange of air molecules in two chambers connected by a small hole.

5.3.1.5 Example (Umbrella)

Say you own r umbrellas, which are either at home or in your office. In the morning if it rains you take an umbrella, if there is one at home, equally in the evening in the office. Say it rains in the morning or in the evening independently with probability p. Analyze this as a Markov chain and find the transition matrix.

Solution 1: Say X_n is the number of umbrellas at home in the morning of the n^{th} day, then $X_n \in \{0, 1, .., r\}$. Now

P(X_n =i|X_{n-1} =i) =
P(it is raining in the morning and evening or it is not raining in the morning and evening) =
$p^2 + q^2, 1 \leq i \leq r$

P(X_n =i-1|X_{n-1} =i) =
P(it is raining in the morning but not in the evening) =
$pq, 1 \leq i \leq r$

P(X_n =i+1|X_{n-1} =i) =
P(it is not raining in the morning but it is raining in the evening) =
$qp, 1 \leq i \leq r - 1$

P(X_n =0|X_{n-1} =0) =
P(it is not raining in the evening) = q

P(X_n =1|X_{n-1} =0) =
P(it is raining in the evening) = p

P(X_n =r|X_{n-1} =r) =
P(it is not raining in the morning or it is raining both times) = q+p^2

so

	0	1	2	..	$r-1$	r
0	q	p	0			0
1	pq	p^2+q^2	pq			0
2	0	pq	p^2+q^2			0
..						
$r-1$					p^2+q^2	pq
r	0	0	0		pq	$q+p^2$

Solution 2: Say X_n is the number of umbrellas at your present location (home or work), then $X_n \in \{0, 1, .., r\}$. Now

$P(X_n = r | X_{n-1} = 0)$ = P(no umbrellas where you were last) = 1
$P(X_n = r\text{-}i | X_{n-1} = i)$ = P(it is not raining) = q, $1 \leq i \leq r$
$P(X_n = r\text{-}i+1 | X_{n-1} = i)$ = P(it is raining) = p, $1 \leq i \leq r$

	0	1	2	..	$r-1$	r
0	0	0	0		0	1
1	0	0	0		q	p
2	0	0	0		p	0
..						
$r-1$	0	q	p		0	0
r	q	p	0		0	0

both of these describe the “experiment”.

Say we have a Markov chain X_n, n=1,2,.. with transition matrix P. Define the n-step transition matrix

$P^{(n)} = (p_{ij}^{(n)})$

by

$p_{ij}^{(n)} = P(X_n = j | X_0 = i)$

Of course $P^{(1)} = P$. Now

$$p_{ij}^{(2)} = P(X_{n+2} = j | X_n = i) =$$
$$\sum_{k \in S} P(X_{n+2} = j | X_n = i, X_{n+1} = k) P(X_{n+1} = k | X_n = i) =$$
$$\sum_{k \in S} P(X_{n+2} = j | X_{n+1} = k) P(X_{n+1} = k | X_n = i) =$$
$$\sum_{k \in S} p_{ik} p_{kj} = (PP)_{ij}$$

In general $p_{ij}^{(n)} = \left(\prod_{i=1}^{n} P \right)_{ij}$

5.3.1.6 Example (Ehrenfest chain)

Let’s find the 2-step transition matrix for the Ehrenfest chain with r=3. The transition matrix is given by

$$P = \begin{array}{cc|cccc} & & \multicolumn{4}{c}{\mathbf{j}} \\ & & \mathbf{0} & \mathbf{1} & \mathbf{2} & \mathbf{3} \\ \hline & \mathbf{0} & 0 & 1 & 0 & 0 \\ \mathbf{i} & \mathbf{1} & 1/3 & 0 & 2/3 & 0 \\ & \mathbf{2} & 0 & 2/3 & 0 & 1/3 \\ & \mathbf{3} & 0 & 0 & 1 & 0 \end{array}$$

and so the 2-step transition matrix is

[](graphs/mark14.png" >

For example , P(X_3=2|X_1=2) = 7/9 because if $X_1 = 2$ the chain could have gone to 1 (with probability 2/3) and then back to 2 (with probability 2/3) or it could have gone to 3 (with probability 1/3) and then back to 2 (with probability 1), so we have

$P(X_3 = 2|X_1 = 2) = 2/3 * 2/3 + 1/3 * 1 = 7/9$

5.3.1.7 Example (Umbrellas)

Finding P^2 for solution 1 (in this generality) is difficult, but for solution 2 we have

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & q & p \\ 0 & 0 & 0 & p & 0 \\ 0 & q & p & 0 & 0 \\ q & p & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & q & p \\ 0 & 0 & 0 & p & 0 \\ 0 & q & p & 0 & 0 \\ q & p & 0 & 0 & 0 \end{pmatrix} =$$

$$\begin{pmatrix} q & p & 0 & 0 & 0 \\ pq & q^2+p^2 & pq & 0 & 0 \\ 0 & pq & q^2+p^2 & 0 & 0 \\ 0 & 0 & 0 & q^2+p^2 & pq \\ 0 & 0 & 0 & pq & q+p^2 \end{pmatrix}$$

5.3.2 Eigenvalues and Eigenvectors

In order to find $P^{(n)}$ we could just find PPP..P n-times. With a little linear algebra this becomes easier: For many matrices P there exists a matrix U and a diagonal matrix D such that $P=UDU^{-1}$. Here is how to find U and D:

First we need to find the eigenvalues of the matrix P, that is we need to find the solutions of the equations $Px = \lambda x$. This is equivalent to $(P - \lambda I)x = 0$ or to $\det(P - \lambda I) = 0$. So we have:

$$\det(P-\lambda I)=\begin{vmatrix} -\lambda & 1 & 0 & 0 \\ 1/3 & -\lambda & 2/3 & 0 \\ 0 & 2/3 & -\lambda & 1/3 \\ 0 & 0 & 1 & -\lambda \end{vmatrix} =$$

$$-\lambda\cdot\begin{vmatrix} -\lambda & 2/3 & 0 \\ 2/3 & -\lambda & 1/3 \\ 0 & 1 & -\lambda \end{vmatrix} - 1\cdot\begin{vmatrix} 1/3 & 2/3 & 0 \\ 0 & -\lambda & 1/3 \\ 0 & 1 & -\lambda \end{vmatrix} =$$

$$-\lambda\cdot[-\lambda(\lambda^2-1/3)-2/3(-2/3\lambda)+0]-[1/3(\lambda^2-1/3)]=$$

$$-\lambda[-\lambda^3+1/3\lambda+4/9\lambda]-1/3\lambda^2+1/9=\lambda^4-10/9\lambda^2+1/9=0$$

$$\lambda^2=\frac{10/9\pm\sqrt{(10/9)^2-4/9}}{2}=\frac{1}{2}\left(\frac{10}{9}\pm\sqrt{\frac{100-4\cdot 9}{9^2}}\right)=\frac{1}{2}\left(\frac{10\pm 8}{9}\right)=1 \text{ or } 1/9$$

so the eigenvalues are

$\lambda_1=-1,\ \lambda_2=-1/3,\ \lambda_3=1/3$ and $\lambda_4=1$

The D above now is the matrix with the eigenvalues on the diagonal. The columns of the matrix U are the corresponding eigenvectors (with Euclidean length 1), so for example

$$(P-\lambda_1 I)x=\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1/3 & 1 & 2/3 & 0 \\ 0 & 2/3 & 1 & 1/3 \\ 0 & 0 & 1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}=0$$

$$\begin{aligned} x_1+x_2 &= 0 \\ 1/3x_1+x_2+2/3x_3 &= 0 \\ 2/3x_2+x_3+1/3x_4 &= 0 \\ x_3+x_4 &= 0 \end{aligned}$$

Of course we have $\det(P-\lambda I)=0$, so this system is does not have a unique solution. Setting

$x_1 = 1$ we can then easily find a solution x=(1,-1,1,-1).

This vector has Euclidean length $\sqrt{(1^2 + (-1)^2 + 1^2 + (-1)^2)} = 2$, so the normalized eigenvector is x=(1/2,-1/2,1/2,-1/2)

Similarly we can find eigenvectors for the other eigenvalues.

Why does this help in computing P^(n)? It turns out that we have

$P^{(2)} = PP = UDU^{-1}UDU^{-1} = UDDU^{-1} = UD^2U^{-1}$ and

$$D^2 = \begin{pmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & & \lambda_3 & 0 \\ 0 & 0 & 0 & \lambda_4 \end{pmatrix}^2 = \begin{pmatrix} \lambda_1^2 & 0 & 0 & 0 \\ 0 & \lambda_2^2 & 0 & 0 \\ 0 & & \lambda_3^2 & 0 \\ 0 & 0 & 0 & \lambda_4^2 \end{pmatrix}$$

and in general we have $P^{(n)} = UD^nU^{-1}$.

Note

$$\sum_{k \in S} p_{ik} = P(X_n \in S | X_{n-1} = i) = 1$$

$$P\begin{pmatrix} 1 \\ 1 \\ .. \end{pmatrix} = \begin{pmatrix} \sum_{k \in S} p_{1k} \\ \sum_{k \in S} p_{2k} \\ .. \end{pmatrix} = 1\begin{pmatrix} 1 \\ 1 \\ .. \end{pmatrix}$$

so $\lambda = 1$ is always an eigenvalue of a transition matrix P, with (unnormalized) eigenvector $(1,1,..,1)^T$.

5.3.2.1 Example (Umbrella)

solution 2 and r=2, then

$$\det(P-\lambda I) = \begin{vmatrix} -\lambda & 0 & 1 \\ 0 & q-\lambda & p \\ q & p & -\lambda \end{vmatrix} =$$

$$(-\lambda)\begin{vmatrix} q-\lambda & p \\ p & -\lambda \end{vmatrix} + \begin{vmatrix} 0 & q-\lambda \\ q & p \end{vmatrix} =$$

$$(-\lambda)((q-\lambda)(-\lambda)-p^2)-(q-\lambda)q =$$

$$(-\lambda)(-\lambda q+\lambda^2-p^2)-q^2+\lambda q =$$

$$\lambda^3-\lambda^2 q-\lambda(q+p^2)+q^2 =$$

$$(\lambda-1)(\lambda^2+(1-q)\lambda-q^2) =$$

$$(\lambda-1)(\lambda^2+\lambda p-q^2) = 0$$

$$\lambda_{2,3} = \frac{-p\pm\sqrt{p^2+4q^2}}{2}$$

and in this generality that's about it

An important consequence of the Markov property is the fact that given the present the past and the future are independent. This is formalized in the

Theorem (Chapman-Kolmogorov equation)

Let $\{X_n, n \geq 0\}$ be a Markov chain. Let x,y,z be in the state space, then

$$P(X_{m+n} = z \rfloor X_1 = x) = \sum_y P(X_m = y \rfloor X_1 = x) P(X_n = z \rfloor X_1 = y)$$

proof

is an immediate consequence of the law of total probability and the Markov property.

5.3.3 Classification of States

There are a number of properties a Markov chains may or may not have. Here are some:

Definition

A Markov chain is said to be *irreducible* if for each pair of states i and j there is a positive probability that starting in state i the chain will eventually move to state j.

5.3.3.1 Example

Both the two random walks, the Ehrenfest chain and the Umbrella chains are irreducible.

5.3.3.2 ** Example (Casino)**

Consider the following chain: you go to the Casino with \$10. You play Blackjack, always betting \$5. Let X_n be your "wealth" after the n^{th} game. Then X_n is in {0,5,10,15,..} and

$P(X_{n+k} = j|X_k = 0) = 0$ for all n > 1.

("0" is called a coffin or absorbing state). So this chain is not irreducible.

Definition

A Markov chain is said to be *aperiodic* if for some $n \geq 0$ and some state j we have

$P(X_n = j|X_0 = j) > 0$ and $P(X_n + 1 = j|X_0 = j) > 0$

In other words there should be a chance to return to state j in either n steps or in n+1 steps.

5.3.3.3 Example

Random walk I, the Ehrenfest chain and the Umbrella chain are not aperiodic because it is only possible to return to the same state in an even number of steps, but not an odd number. Random Walk II is aperiodic.

Definition

A state x of a Markov chain is said to be *recurrent* if P(the chain returns to x infinitely often) = 1. A Markov chain is called recurrent if all its states are recurrent. A state that is not recurrent is called *transient*.

A recurrent state i is said to be *positive recurrent* if starting at i the expected time until the return to i is finite, otherwise it is called *null recurrent*.

Theorem

In a finite-state chain all recurrent states are positive recurrent. **proof**

Say S={x_1 ,..,x_m } and assume (wlog) {x_1 ,..,x_k } are recurrent. Then for any $i \neq j, i, j \leq m$

$P(X_n = x_i|X_0 = x_j) > 0$

because x_i is recurrent, so we have to be able to get there infinitely often no matter where we start. Therefore any recurrent state is reachable from any other recurrent state.

Assume that there is no positive recurrent state. Then all states are either transient or null recurrent. So the expected return time to all the states is infinite. But there are only finitely

many states, so this is impossible. Therefore there has to be at least one positive-recurrent state.

Say x is a positive-recurrent state, and y is recurrent. Then

$$P(X_{n+1} = y|X_0 = x) =$$
$$\sum_{j=1}^{m} P(X_{n+1} = y|X_0 = x, X_n = x_j)P(X_n = x_j|X_0 = x) \geq$$
$$P(X_{n+1} = y|X_n = x)P(X_n = x|X_0 = x) = P_{xy}P(X_n = x|X_0 = x)$$

$$P(X_{n+1} = y|X_0 = x)$$
$$\sum_{j=1}^{m} P(X_{n+1} = y|X_0 = x, X_1 = x_j)P(X_1 = x_j|X_0 = x) \geq$$
$$P(X_n = y|X_0 = y)P_{xy}$$

$$E[\text{expected time until the return to x}] =$$
$$\sum_{n=1}^{\infty} nP(X_n = x|X_0 = x) =$$
$$\sum_{n=1}^{\infty} nP(X_n = x|X_0 = x) \leq$$
$$\sum_{n=1}^{\infty} n\frac{1}{P_{xy}}P(X_{n+1} = y|X_0 = x) \leq$$
$$\sum_{n=1}^{\infty} n\frac{1}{P_{xy}^2}P(X_n = y|X_0 = y) =$$
$$\frac{1}{P_{xy}^2}E[\text{expected time until the return to y}] < \infty$$

and so y is positive recurrent

5.3.3.4 Example

The Ehrenfest chain and the Umbrella chains are clearly poitive recurrent. In the Casino example "0" is a recurrent state, the others are not.

Are the random walks recurrent? Good question! It seems clear that the asymmetric r.v. is not (if $p \neq 0.5$), because eventually one expects the walk to run off to infinity (or - infinity). How about Random Walk I? Actually let's consider a more general problem:

5.3.3.5 Example (Random Walk III)

let the state space be the lattice of integers on $\mathbb{R}^d$, that is $X_n = (i_1, .., i_d)$ for i_k any integer. Then the chain goes from one point on the lattice to any of the 2d points that differ by one in one coordinate with probability 1/2d.

One of the great results in probability states:

Theorem

The simple random walk is recurrent if $d \leq 2$, transient otherwise

or as Kakutani once said "A drunk man will find his way home but a drunk bird may get lost".

Definition

Positive recurrent aperiodic states are called *ergodic*.

5.3.4 Stationary Distribution

Until now we started the chain at time 0 in some specified state j. Let's consider what happens if we choose that state according to some distribution π:

$$P_\pi(X_1 = k) =$$

$$\sum_j P(X_1 = k | X_0 = j) P(X_0 = j) =$$

$$\sum_j P_{jk} \pi_j = (\pi^T P)_k$$

and clearly an interesting case is if the probability to be in a certain state does not change, that is if

$\pi^T P = \pi^T$

Note:

$\pi^T P^{(n)} = \pi^T P P^{(n-1)} = \pi^T P^{(n-1)} = ... = \pi^T$

so this immediately implies that the probability for the chain to be in state k is always π_k.

With this idea we have the

Definition

Let S be the state space of a Markov chain X with transition matrix P. Let π be a "measure" on S. Then π is called a *stationary measure* of X if $\pi^T P = \pi^T$.

We won't discuss exactly what it means for π to be a "measure". You can think of it in the same way as a probability distribution, only that we don't have $\sum \pi_i = 1$.

Note:

$\pi^T P = \pi^T$ iff $(P^T \pi)^T = \pi^T$ iff
$P^T \pi = \pi$ iff
$(P^T - I)\pi = 0$

so again this leads to a system of equations is singular. Often we can get a unique solution by requiring that π be a proper probability distribution, that is that $\sum \pi_i = 1$.

The interpretation is the following: Say we choose the initial state X_0 according to π, that is $P(X_0 = i) = \pi_i$. Then π_i is the long-run proportion of time the chain visits state i, that is

$\pi_i = \lim \frac{1}{N} \sum_{k=1}^{N} I[X_i = i]$

There is an extension of the WLLN to Markov chains. That is, say h is a function on the state space, then

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} h(X_i) = Eh(Z)$$

where Z is a r.v. with pdf π.

One of the main results for Markov chains is the following:

###Theorem

If the Markov chain $\{X_n\}$ is irreducible and ergodic, then

$$\lim_{n \to \infty} P(X_n = j) = \pi_j$$

proof omitted

5.3.4.1 Example (Ehrenfest chain)

To find a (?) stationary measure we have to solve the system of equations

$\pi^T P = \pi^T$

Let's start with the case r=3:

$$\pi^T = (\pi_0\ \pi_1\ \pi_2\ \pi_3) =$$

$$(\pi_0\ \pi_1\ \pi_2\ \pi_3)\begin{pmatrix} 0 & 1 & 0 & 0 \\ \frac{1}{3} & 0 & \frac{2}{3} & 0 \\ 0 & \frac{2}{3} & 0 & \frac{1}{3} \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Here this means the system

$\pi_0 = 1/3\pi_1$
$\pi_1 = \pi_0 + 2/3\pi_2$
$\pi_2 = 2/3\pi_1 + \pi_3$
$\pi_3 = 1/3\pi_2$
$\pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$

and so $\pi = (1/8, 3/8, 3/8, 1/8)$

Before doing the general case it is often a good idea to do a specific case that has all the "parts" (ie equations), so let's do next r=5:

First the transition matrix:

$$P_5 = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & 0 & \frac{4}{5} & 0 & 0 & 0 \\ 0 & \frac{2}{5} & 0 & \frac{3}{5} & 0 & 0 \\ 0 & 0 & \frac{3}{5} & 0 & \frac{2}{5} & 0 \\ 0 & 0 & 0 & \frac{4}{5} & 0 & \frac{1}{5} \\ 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

and now the equations:

$$\pi_0 = \tfrac{1}{5}\pi_1$$

$$\pi_1 = \pi_0 + \tfrac{2}{5}\pi_2$$

$$\pi_2 = \tfrac{4}{5}\pi_1 + \tfrac{3}{5}\pi_3$$

$$\pi_3 = \tfrac{3}{5}\pi_2 + \tfrac{4}{5}\pi_4$$

$$\pi_4 = \tfrac{2}{5}\pi_3 + \pi_5$$

$$\pi_5 = \tfrac{1}{5}\pi_4$$

$$\pi_1 = 5\pi_0$$

$$\pi_2 = \tfrac{5}{2}(\pi_1 - \pi_0) = \tfrac{5}{2}(5\pi_0 - \pi_0) = 10\pi_0$$

$$\pi_3 = \tfrac{5}{3}(\pi_2 - \tfrac{4}{5}\pi_1) = \tfrac{5}{3}(10\pi_0 - \tfrac{4}{5}5\pi_0) = 10\pi_0$$

$$\pi_4 = \tfrac{5}{4}(\pi_3 - \tfrac{3}{5}\pi_2) = \tfrac{5}{4}(10\pi_0 - \tfrac{3}{5}10\pi_0) = 5\pi_0$$

$$\pi_5 = \tfrac{1}{5}\pi_4 = \pi_0$$

$$\sum \pi_i = (1 + 5 + 10 + 10 + 5 + 1)\pi_0 = 32\pi_0$$

$$\pi^T = \begin{pmatrix} 1 & 5 & 10 & 10 & 5 & 1 \end{pmatrix}/32$$

$$\text{or } \pi_j = \binom{5}{j}/2^5$$

Finally, for the general case:

$$P_{j,j-1} = \frac{j}{r} \quad P_{j,j+1} = \frac{r-j}{r} \quad 1 \le j \le r-1$$
$$P_{01} = P_{r,r-1} = 1$$

$$\pi^T = \pi P^T$$
$$\pi_j = \sum_{k=0}^{r} \pi_k P_{kj}$$

$$\pi_0 = \sum_{k=0}^{r} \pi_k P_{k0} = \frac{1}{r}\pi_1$$
$$\pi_1 = \pi_0 P_{01} + \pi_2 P_{21} = \pi_0 + \frac{2}{r}\pi_2$$
$$\pi_j = \pi_{j-1} P_{j-1,j} + \pi_{j+1} P_{j+1,j} = \frac{r-j+1}{r}\pi_{j-1} + \frac{j+1}{r}\pi_{j+1}$$
$$\pi_{r-1} = \pi_{r-2} P_{r-2,r-1} + \pi_r P_{r-1,r} = \frac{2}{r}\pi_{r-2} + \pi_r$$
$$\pi_r = P_{r-1.r}\pi_{r-1} = \frac{1}{r}\pi_{r-1}$$

$$\pi_1 = r\pi_0 = \binom{r}{1}$$
$$\pi_2 = \frac{r}{2}(\pi_1 - \pi_0) = \frac{r}{2}(r-1)\pi_0 = \binom{r}{2}\pi_0$$

$$\pi_{j+1} = \frac{r}{j+1}\left(\pi_j - \frac{r-j+1}{r}\pi_{j-1}\right) =$$
$$\frac{r}{j+1}\left(\binom{r}{j} - \frac{r-j+1}{r}\binom{r}{j-1}\right)\pi_0 =$$
$$\frac{r}{j+1}\left(\frac{r!}{(r-j)!j!} - \frac{r-j+1}{r}\frac{r!}{(r-j+1)!(j-1)!}\right)\pi_0 =$$
$$\frac{r}{j+1}\frac{r!}{(r-j)!j!}\left(1 - \frac{j}{r}\right)\pi_0 =$$
$$\frac{r*r!}{(r-j)!(j+1)!}\left(\frac{r-j}{r}\right)\pi_0 = \binom{r}{j+1}\pi_0$$
$$\pi_r = \frac{1}{r}\pi_{r-1} = \frac{1}{r}\binom{r}{r-1}\pi_0 = \pi_0$$
$$\sum_{i=0}^{r}\pi_i = \sum_{i=0}^{r}\binom{r}{i} = 2^r$$
$$\pi_j = \binom{r}{j}/2^r$$

5.3.4.2 Example (Umbrellas)

For solution 1 the system of equations is

$$qx_0 + pqx_1 = x_0$$

$$px_{i-1} + (p^2 + q^2)x_i + pqx_{i+1} = x_i \quad i = 1, .., r$$

$$pqx_{r-1} + (q + p^2)x_r = x_r$$

so

$$-(1-q)x_0 + pqx_1 = 0 \text{ or } -px_0 + pqx_1 = 0 \text{ or } x_0 = qx_1$$

$$px_0 + (p^2 + q^2)x_1 + pqx_2 = x_1$$

$$(pq + p^2 + q^2 - 1)x_1 + pqx_2 = 0$$

$$((p+q)^2 - pq - 1)x_1 + pqx_2 = 0$$

$$-pqx_1 + pqx_2 = 0 \text{ or } x_2 = x_1$$

and so on shows x=c(q,1,..,1) solves the system. Now q+1+..+1=q+r, so the stationary distribution is $\pi_0 = q/(q+r), \pi_i = 1/(q+r) i = 1, .., r$.

For solution 2 we have

$$qx_r = x_0$$

$$qx_{r-i} + px_{r-i+1} = x_i \quad i = 1, .., r-1$$

$$x_0 + px_1 = x_r$$

so

$$px_1 = x_r - x_0 = x_r - qx_r = px_r \text{ or } x_1 = x_r$$

$$qx_{r-1} + px_r = x_1 = x_r \text{ or } x_{r-1} = x_r$$

and we see that we get the same stationary distribution as in solution 1.

So, what percentage of times do you get wet? Clearly this is

P(no umbrella and it rains) = $q\pi_0 = q^2/(q+r)$

5.3.4.3 Example (Random Walk)

Let S be the integers and define a Markov chain by $p_{i,i+1}$ = p and $p_{i,i-1}$ = q = 1-p. A stationary measure is given by $\pi_i = 1$ for all i because $(\pi P)_i = 1p + 1q = 1$.

Now assume $p \neq q$ and define $\pi_i = (p/q)^i$. Then

$$\sum_i \pi_i P_{ij} = \pi_{j+1} P_{j+1,j} + \pi_{j-1} P_{j-1,j} =$$

$$\left(\frac{p}{q}\right)^{j+1} q + \left(\frac{p}{q}\right)^{j-1} p = \left(\frac{p}{q}\right)^{j} (p+q) = \left(\frac{p}{q}\right)^{j} = \pi_j$$

Note that this shows that stationary measure are not unique.

Here is another property of Markov chains: A Markov chain is said to be *time-reversible* if

$\pi_i P_{ij} = \pi_j P_{ji}$

for all $i \neq j$. It can be shown that for a time reversible Markov chain if the chain is started from π and run backwards in time it again has transition matrix P.

5.3.4.4 Example

The Ehrenfest chain is time-reversible. We will show this for the case i=k, j=k+1:

$$\pi_k P_{k,k+1} = \binom{r}{k} / 2^r \, \frac{r-k}{r} = \frac{r!(r-k)}{(r-k)!k!r2^r} =$$

$$\frac{r!}{(r-k-1)!k!r2^r} = \frac{r!(k+1)}{(r-k-1)!(k+1)!r2^r} =$$

$$\binom{r}{k+1} / 2^r \, \frac{k+1}{r} = \pi_{k+1} P_{k+1,k}$$

5.3.5 The Gambler's Ruin Problem

Suppose you go to the casino and repeatedly play a game where you win and double your "bet" with probability p and loose with probability q=1-p. For example, if you play roulette and always place your bet on "red" we have p=18/38.

Suppose you go in with the following plan: you have $i to start, you always bet $1 in each round and you stay until you either lost all your money or until you have reached $N. What is the probability of reaching $N before going broke?

If we let X_n denote your "fortune" after n rounds $\{X_n\}$ is a Markov chain on {0,1,..,N} with transition probabilities

$p_{00} = p_{NN} = 1$
$p_{i,i+1} = p$
$p_{i,i-1} = q$

for i in {1,..,N-1}

Also we $X_0 = i$.

Let P_i denote the probability that, starting at i the fortune will eventually reach N. We have:

$P_i = P(X_n = N \text{ for some } n \geq 1 | X_0 = i) =$

$P(X_n = N \text{ for some } n \geq 1 | X_0 = i, X_1 = i+1)P(X_1 = i+1 | X_0 = i) +$

$P(X_n = N \text{ for some } n \geq 1 | X_0 = i, X_1 = i-1)P(X_1 = i-1 | X_0 = i) =$

$P_{i+1}p + P_{i-1}q$

so $P_i = P_{i+1}p + P_{i-1}q$ for $i = 1, 2.., N-1$

Now $P_{i+1}p + P_{i-1}q = P_i = (p+q)P_i = pP_i + qP_i$

or $P_{i+1} - P_i = \frac{q}{p}(P_i - P_{i-1})$

Note that $P_0 = 0$ and so

$P_2 - P_1 = \frac{q}{p}(P_1 - P_0) = \frac{q}{p}P_1$

$P_3 - P_2 = \frac{q}{p}(P_2 - P_1) = \left(\frac{q}{p}\right)^2 P_1$

$...P_i - P_{i-1} = \frac{q}{p}(P_{i-1} - P_{i-2}) = \left(\frac{q}{p}\right)^{i-1} P_1$

adding the first $i-1$ of these equations together we find

$\sum_{j=2}^{i}(P_j - P_{j-1}) = \sum_{j=2}^{i}\left(\frac{q}{p}\right)^{j-1} P_1$

$P_i = \sum_{j=1}^{i-1}\left(\frac{q}{p}\right)^{j} P_1 + P_1 = \sum_{j=0}^{i-1}\left(\frac{q}{p}\right)^{j} P_1$

and so $P_i = \begin{cases} \frac{1-(q/p)^i}{1-q/p}P_1 & \text{if } p \neq q \\ iP_1 & \text{if } p = q \end{cases}$

Note that $P_N = 1$ and that the formula above also holds for i=N, so we have

$$1 = P_N = \begin{cases} \frac{1-(q/p)^N}{1-q/p} P_1 & \text{if } p \neq q \\ NP_1 & \text{if } p = q \end{cases}$$

$$\text{so } P_1 = \begin{cases} \frac{1-q/p}{1-(q/p)^N} & \text{if } p \neq q \\ \frac{1}{N} & \text{if } p = q \end{cases}$$

and finally

$$P_i = \begin{cases} \frac{1-(q/p)^i}{1-(q/p)^N} & \text{if } p \neq 1/2 \\ \frac{i}{N} & \text{if } p = 1/2 \end{cases}$$

$$\text{Note that } \lim_{N\to\infty} P_i = \begin{cases} 1-(q/p)^i & \text{if } p \neq 1/2 \\ 0 & \text{if } p = 1/2 \end{cases}$$

The main "trick" in this calculation was to condition on the "right" event (here X_1). This is often the case when doing math with Markov chains.

Say in our example playing roulette you start with \$100. What is the probability of reaching N before going broke? We find

N	P_N
101	0.9
105	0.59
110	0.35
120	0.12
130	0.04

Is it the same probability to start with \$100 and reach \$110 or to start with \$200 and reach \$220? The answer is no, P_{220} =0.12 for i=100.

5.3.5.1 Example (two-state process)

Here X_n takes only two possible states, coded as 0 and 1. Therefore the transition matrix is given by

$$P = \begin{pmatrix} \alpha & 1-a \\ 1-\beta & \beta \end{pmatrix}$$

$$\det(P-\lambda) = \begin{vmatrix} \alpha-\lambda & 1-a \\ 1-\beta & \beta-\lambda \end{vmatrix} =$$

$$(\alpha-\lambda)(\beta-\lambda)-(1-\alpha)(1-\beta) =$$

$$\lambda^2-(\alpha+\beta)\lambda+\alpha\beta-1+\alpha+\beta-\alpha\beta =$$

$$(\lambda-1)(\lambda+1-\alpha-\beta) = 0$$

$$\lambda_1 = 1,\ \lambda_2 = \alpha+\beta-1$$

$$(P-\lambda_1 I)x = \begin{pmatrix} \alpha-1 & 1-a \\ 1-\beta & \beta-1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$(\alpha-1)x_1+(1-\alpha)x_2 = 0 \text{ so } x_1 = x_2$$

$$u_1 = (1/\sqrt{2}, 1/\sqrt{2})^T$$

$$(P-\lambda_2 I)x = \begin{pmatrix} \alpha-(\alpha+\beta-1) & 1-a \\ 1-\beta & \beta-(\alpha+\beta-1) \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} =$$

$$\begin{pmatrix} 1-\beta & 1-a \\ 1-\beta & 1-\alpha) \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$(1-\beta)x_1+(1-\alpha)x_2 = 0 \text{ so if } x_2 = 1\ x_1 = -\tfrac{1-\alpha}{1-\beta}$$

$$\sqrt{x_1^2+x_2^2} = \sqrt{(-\tfrac{1-\alpha}{1-\beta})^2+1} = \frac{\sqrt{(1-\alpha)^2+(1-\beta)^2}}{1-\beta}$$

$$\text{Let } z = \sqrt{(1-\alpha)^2+(1-\beta)^2}$$

$$u_2 = \tfrac{1-\beta}{z}(-\tfrac{1-\alpha}{1-\beta}, 1)^T = \tfrac{1}{z}(\alpha-1, 1-\beta)^T$$

Now

$$P^n = UD^nU^{-1}$$

$$U^{-1} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{\alpha-1}{z} \\ \frac{1}{\sqrt{2}} & \frac{1-\beta}{z} \end{pmatrix}^{-1} =$$

$$\frac{1}{\frac{1}{\sqrt{2}}\frac{1-\beta}{z} - \frac{\alpha-1}{z}\frac{1}{\sqrt{2}}} \begin{pmatrix} \frac{1-\beta}{z} & -\frac{\alpha-1}{z} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} =$$

$$\frac{\sqrt{2}z}{2-\alpha-\beta} \begin{pmatrix} \frac{1-\beta}{z} & \frac{1-\alpha}{z} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

$$P^n = UD^nU^{-1} =$$

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{\alpha-1}{z} \\ \frac{1}{\sqrt{2}} & \frac{1-\beta}{z} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & (\alpha+\beta-1)^n \end{pmatrix} \frac{\sqrt{2}z}{2-\alpha-\beta} \begin{pmatrix} \frac{1-\beta}{z} & \frac{1-\alpha}{z} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} =$$

$$\frac{\sqrt{2}z}{2-\alpha-\beta} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{\alpha-1}{z}(\alpha+\beta-1)^n \\ \frac{1}{\sqrt{2}} & \frac{1-\beta}{z}(\alpha+\beta-1)^n \end{pmatrix} \begin{pmatrix} \frac{1-\beta}{z} & \frac{1-\alpha}{z} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} =$$

$$\frac{\sqrt{2}z}{2-\alpha-\beta} \begin{pmatrix} \frac{1}{\sqrt{2}}\frac{1-\beta}{z} - \frac{1}{\sqrt{2}}\frac{\alpha-1}{z}(\alpha+\beta-1)^n & \frac{1}{\sqrt{2}}\frac{1-\alpha}{z} + \frac{1}{\sqrt{2}}\frac{\alpha-1}{z}(\alpha+\beta-1)^n \\ \frac{1}{\sqrt{2}}\frac{1-\beta}{z} - \frac{1}{\sqrt{2}}\frac{1-\beta}{z}(\alpha+\beta-1)^n & \frac{1}{\sqrt{2}}\frac{1-\alpha}{z} + \frac{1}{\sqrt{2}}\frac{1-\beta}{z}(\alpha+\beta-1)^n \end{pmatrix} =$$

$$\frac{1}{2-\alpha-\beta} \begin{pmatrix} 1-\beta+(1-\alpha)(\alpha+\beta-1)^n & 1-\alpha-(1-\alpha)(\alpha+\beta-1)^n \\ 1-\beta-(1-\beta)(\alpha+\beta-1)^n & 1-\alpha+(1-\beta)(\alpha+\beta-1)^n \end{pmatrix} =$$

For the stationary distribution we find

$$(\pi_1 \ \pi_2)P = \begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix}$$

$$\alpha\pi_1 + (1-\beta)\pi_2 = \pi_1$$

$$\pi_1 + \pi_2 = 1$$

$$(\alpha - 1)\pi_1 + (1-\beta)(1-\pi_1) = 0$$

$$(\alpha + \beta - 2)\pi_1 + 1 - \beta = 0$$

$$\pi_1 = \frac{1-\beta}{2-\alpha-\beta}, \quad \pi_2 = \frac{1-\alpha}{2-\alpha-\beta}$$

Finally

$0 < \alpha, \beta < 1$ so $0 < \alpha + \beta < 2$ so $-1 < \alpha + \beta - 1 < 1$

so $(\alpha + \beta - 1)^n \to 0$ as $n \to \infty$

$\lim_{n\to\infty} P^n =$

$$\lim_{n\to\infty} \frac{1}{2-\alpha-\beta} \begin{pmatrix} 1-\beta+(1-\alpha)(\alpha+\beta-1)^n & 1-\alpha+(1-\alpha)(\alpha+\beta-1)^n \\ 1-\beta+(1-\beta)(\alpha+\beta-1)^n & 1-\alpha+(1-\beta)(\alpha+\beta-1)^n \end{pmatrix} =$$

$$\frac{1}{2-\alpha-\beta} \begin{pmatrix} 1-\beta & 1-\alpha \\ 1-\beta & 1-\alpha \end{pmatrix} = \begin{pmatrix} \frac{1-\beta}{2-\alpha-\beta} & \frac{1-\alpha}{2-\alpha-\beta} \\ \frac{1-\beta}{2-\alpha-\beta} & \frac{1-\alpha}{2-\alpha-\beta} \end{pmatrix} = \begin{pmatrix} \pi_1 & \pi_2 \\ \pi_1 & \pi_2 \end{pmatrix}$$

5.4 Continuous-time Markov Chains

Definition

Say $\{X(t), t \geq 0\}$ is a continuous-time stochastic process taking values on the nonnegative integers. Then X(t) is a Markov chain if for all $s, t \geq 0$, and nonnegative integers i,j, x(u),

$0 \leq u < s$ we have

$$P(X(t+s) = j | X(s) = i, X(u) = x(u), 0 \leq u < s) = P(X(t+s) = j | X(s) = i)$$

The main result for such chains is

Theorem

let $\{X(t), t \geq 0\}$ be a continuous-time Markov chain with X(0)=i. Let T_i be the time the process stays in state i. Then T_i has an exponential distribution.

proof

$$P(T_i > s + t | T_i > s, X_0 = i) =$$
$$P(X(u) = i, 0 < u < s + t | X(u) = i, 0 < u < s, X_0 = i) =$$
$$P(X(u) = i, s < u < s + t | X(s) = i) =$$
$$P(X(u) = i, 0 < u < t | X(0) = i) =$$
$$P(T_i > t)$$

But this means that T_i is *memoryless*! Of course T_i is also non-negative and continuous, and therefore T_i has to have an exponential distribution.

With this we have the following characterization of a continuous-time Markov chain:

1. the amount of time spent in state i is an exponential distribution with mean v_i.
2. when the process leaves state i it next enters state j with some probability, say P_{ij}.

So a continuous-time Markov chain is a process that moves from state to state in accordance with a discrete-space Markov chain, but also spends an exponentially distributed amount of time in each state.

Let's consider a finite- statespace continuous-time Markov chain, that is $X(t) \in \{0, .., N\}$. Let

$P_{ij}(t) = P(X(t) = j | X(0) = i)$

then the the Markov property asserts that $\{X(t), t \geq 0\}$ satisfies

$$a)\ p_{ij}(t) \geq 0$$

$$b)\ \sum_{j=0}^{N} p_{ij}(t) = 1$$

$$c)\ p_{ik}(s+t) = \sum_{j=0}^{N} p_{ij}(s) p_{jk}(t)$$

$$d^*)\ \lim_{h \downarrow 0} p_{ij}(h) = I(i=j)$$

where c) follows from the Chapman-Kolmogorov equations.

d*) is not strictly a consequence of the Markov property but is usually a sensible additional condition.

Let P(t)=(p_{ij}) denote the matrix of transition probabilities at time t, so P is a matrix whose entries are functions of t.

Now c) can be written as

$P(s+t) = P(s)P(t)$ for all $t, s \geq 0$

and d*) as

$\lim_{h \to 0} P(h) = I$

this implies that P(t) is (right)-continuous at time 0, meaning each entry is continuous at t=0. Now

$\lim_{h\downarrow 0} P(t+h) = \lim_{h\downarrow 0} P(t)P(h) =$

$P(t)\lim_{h\downarrow 0} P(h) = P(t)$

let $0 < h < t$, then

$P(t) = P(t-h+h) = P(t-h)P(h)$

by d^* if h is small $P(h) \approx I$, so $P(h)^{-1}$exists

and $P(h)^{-1} \approx I$

so $P(t-h) = P(t)P(h)^{-1}$ and

$\lim_{h\uparrow 0} P(t-h) = P(t)$

and so P(t) is continuous for all $t \geq 0$. Actually, we have even more:

$q_i =: \lim_{h\downarrow 0} \frac{1-P_{ii}(h)}{h} < \infty$

$q_{ij} =: \lim_{h\downarrow 0} \frac{P_{ii}(h)}{h} < \infty,\ i \neq j$

now

$\sum_{j=0}^{N} P_{ij}(h) = 1$, so $\sum_{j=0, j\neq i}^{N} P_{ij}(h) = 1 - P_{ii}(0)$

so $q_i = \sum_{j=0, j\neq i}^{N} q_{ij}$

which shows that P(t) is even differentiable

The rates q_i and q_{ij} give as a second way to describe a Markov chain, called the *infinitesimal description*:

$$P(X(t+h) = j | X(t) = i) = q_{ij}h + o(h)$$
$$P(X(t+h) = i | X(t) = i) = 1 - q_{ii}h + o(h)$$

Let

$$A = \begin{pmatrix} -q_0 & q_{01} & .. & q_{0N} \\ q_{10} & -q_{11} & .. & q_{1N} \\ \cdots & \cdots & \cdots & \cdots \\ q_{N0} & q_{N1} & & -q_{NN} \end{pmatrix}$$

$$\text{then } \lim_{h \downarrow 0} \frac{P(h)-I}{h} = A = P'(0)$$

$$\frac{P(t+h)-P(t)}{h} = \frac{P(t)P(h)-P(t)}{h} = \frac{P(h)-I}{h}P(t)$$

$$\text{so } \lim_{h \downarrow 0} \frac{P(t+h)-P(t)}{h} = AP(t) = P'(t)$$

$$\text{so } P(t) = \exp\{At\} = I + \sum_{n=1}^{\infty} \frac{A^n t^n}{n!}$$

5.4.0.1 Example (Two-state Chain)

say $\{X(t), t \geq 0\}$ is a Markov chain with $X(t) \in \{0, 1\}$ and

$$A = \begin{pmatrix} -a & a \\ b & -b \end{pmatrix}$$

We need A^n, so

$$det(A-\lambda I)=\begin{vmatrix} -a-\lambda & a \\ b & -b-\lambda \end{vmatrix}=$$

$$(a+\lambda)(b+\lambda)-ab=$$

$$ab+(a+b)\lambda+\lambda^2-ab=$$

$$(a+b+\lambda)\lambda=0$$

$$\lambda_1=0,\ \ \lambda_2=-(a+b)$$

$\lambda_1=0$, so A is singular, so no diagonalization, but

$$A^2=\begin{pmatrix} -a & a \\ b & -b \end{pmatrix}\begin{pmatrix} -a & a \\ b & -b \end{pmatrix}=$$

$$\begin{pmatrix} a^2+ab & -a^2-ab \\ -ab-b^2 & b^2+ab \end{pmatrix}=$$

$$-(a+b)\begin{pmatrix} -a & a \\ b & -b \end{pmatrix}=-(a+b)A$$

$$A^n=[-(a+b)]^{n-1}A$$

$$P(t)=I+\sum_{n=1}^{\infty}\frac{A^n t^n}{n!}=I+\sum_{n=1}^{\infty}\frac{[-(a+b)]^{n-1}t^n}{n!}A=$$

$$I+\frac{1}{-(a+b)}\sum_{n=1}^{\infty}\frac{[-(a+b)]^n t^n}{n!}A=$$

$$I-\frac{1}{a+b}\left[\sum_{n=0}^{\infty}\frac{[-(a+b)]^n t^n}{n!}-1\right]A=$$

$$I-\frac{1}{a+b}[\exp\{-(a+b)t\}-1]A=$$

$$I+\frac{1}{a+b}A-\frac{1}{a+b}Ae^{-(a+b)t}$$

it is easy to find the stationary distribution of a continuous-time discrete-space Markov chain in terms of the infinitesimal matrix. If all states communicate, that is if P_{ij} (t)>0 for all i,j and some t>0, then

$\lim_{t\to\infty} P_{ij}(t) = \pi_j > 0$

exists, and

$$\lim_{t\to\infty} P'(t) = \lim_{t\to\infty} AP(t) = A\pi$$
$$\text{but } \lim_{t\to\infty} P'(t) = 0$$

otherwise the chain would never leave i, and so we have $\pi A = 0$ or

$\pi_j q_j = \sum_{i\neq j} \pi_i q_{ij}$

j=0,..,N

5.4.0.2 Example (Redundancy)

A company has a computer for its website. If the computer is down they can't sell anything, so they have a backup, which takes over if the first computer is down. The operating computer fails after an exponentially distributed time (with rate μ). Repair times are also exponentially distributed (with rate λ). Let's assume that μ is fixed but we have a choice of λ (by hiring more technicians). We want to make sure that in the long run at most 1% of the time both computers are down. How should we choose λ?

Let X(t) be the number of computers in operating condition at time t, so X(t) is 0, 1 or2. Then X(t) is a Markov chain with infinitesimal matrix

$$A = \begin{array}{c|ccc} & 0 & 1 & 2 \\ 0 & -\lambda & \lambda & 0 \\ 1 & \mu & -(\mu+\lambda) & \lambda \\ 2 & & \mu & -\mu \end{array}$$

What is the average "total downtime", that is the time when neither computer is working? The system of equations for the stationary distribution is

$$\lambda\pi_0 = \mu\pi_1$$

$$(\mu+\lambda)\pi_1 = \mu\pi_0 + \lambda\pi_2$$

$$\mu\pi_2 = \lambda\pi_1$$

$$\pi_0 + \pi_1 + \pi_2 = 1$$

$$\frac{\mu}{\lambda}\pi_1 + \pi_1 + \frac{\lambda}{\mu}\pi_1 = 1$$

$$(\frac{\mu}{\lambda} + 1 + \frac{\lambda}{\mu})\pi_1 = 1$$

$$\pi_1 = \frac{1}{\frac{\mu}{\lambda}+1+\frac{\lambda}{\mu}} = \frac{\lambda/\mu}{1+\lambda/\mu+(\lambda/\mu)^2}$$

$$\pi_0 = \frac{1}{1+\lambda/\mu+(\lambda/\mu)^2}$$

$$\pi_2 = \frac{(\lambda/\mu)^2}{1+\lambda/\mu+(\lambda/\mu)^2}$$

$$\lim_{t\to\infty} P(X(t) > 0) = 1 - \pi_0 =$$

$$\frac{\lambda/\mu+(\lambda/\mu)^2}{1+\lambda/\mu+(\lambda/\mu)^2}$$

and we see that the probability only depends on the ratio λ/μ. Set $x = \lambda/\mu$, then

$$\frac{x+x^2}{1+x+x^2} = p$$

$$x + x^2 = (1 + x + x^2)p$$

$$(1-p)x^2 + (1-p)x - p = 0$$

$$x = \frac{-(1-p) \pm \sqrt{(1-p)^2 - 4(1-p)(-p)}}{2(1-p)} =$$

$$\frac{-(1-p) \pm \sqrt{(1-p)(1-p+4p)}}{2(1-p)} =$$

$$\frac{-(1-p) \pm \sqrt{(1-p)(1+3p)}}{2(1-p)} = \frac{1}{2}\left(-1 \pm \sqrt{\frac{1+3p}{1-p}}\right)$$

$$p = 0.01, \; x = -1.01 \text{ or } 0.010$$

$$\text{so } \lambda = 0.01\mu$$

5.4.1 Continuous-time Markov Chains with Infinite Statespace

5.4.1.1 Example (Pure Birth process)

Many webpages have a counter that keeps track of the number of people who have visited the site. We can model such a counter as a Markov Chain called a "Pure Birth" process. At time 0 there have been 0 visitors. Say at time t there have been X(t)=n. The counter stays at n for time T that has an exponential distribution with rate λ.

5.4.1.2 Example (Birth and Death Processes)

Consider a system whose state at any time is the number of "people" in the system. Suppose if there are n people in the system then

(i) new arrivals enter the system at an exponential rate λ_n ("births")

(ii) people leave the system at and exponential rate μ_n ("deaths")

(iii) births and deaths occur independently of each other

Thus a birth and death process is a Markov chain with state-space {0,1,..} and

1. $v_0 = \lambda_0$
2. $v_i = \lambda_i + \mu_i$
3. $\mathrm{P}_{01} = 1$
4. $P_{i,i+1} = \lambda_i/(\lambda_i + \mu_i)$
5. $P_{i,i-1} = \mu_i/(\lambda_i + \mu_i)$

where 4) is because we go from i to i+1 if there is a birth before a death. Let $X \sim Exp(\lambda), Y \sim Exp(\mu)$ and $X \perp Y$. Now

[](graphs/mark117.png" >

5.4.1.3 Example (A simple epidemic model)

Consider a population of m individuals that at time 0 consists of 1 "infected" and m-1 "susceptible" (individuals that might get infected, maybe because they have not been immunized. Once infected an individual remains so forever and we suppose that in any time interval h any given infected person will cause, with probability $\alpha h + o(h)$ any given susceptible to become infected. If we let X(t) denote the number of infected people in the population at time t, X(t) is a pure birth process with

$\lambda_n = (m-n)n\alpha, n = 1, .., m-1$

because if there are n infected people the m-n uninfected ones get infected at a rate of $n\alpha$.

Let $\mathrm{T_i}$ be the time to go from i to i+1 infected, and let T be the time until the total population is infected, then

$$T = \sum_{i=1}^{m-1} T_i$$

$$T_i \sim Exp((m-i)i\alpha), \;\; T_i \perp T_j$$

$$E[T] = \sum_{i=1}^{m-1} E[T_i] =$$

$$\sum_{i=1}^{m-1} \frac{1}{(m-i)i\alpha} = \frac{1}{\alpha} \sum_{i=1}^{m-1} \frac{1}{i(m-i)} =$$

$$\frac{1}{m\alpha} \sum_{i=1}^{m-1} \left(\frac{1}{m-i} + \frac{1}{i} \right) =$$

$$\frac{2}{m\alpha} \sum_{i=1}^{m-1} \frac{1}{i} = \frac{2}{m\alpha} \sum_{i=1}^{m-1} \frac{1}{i}(i+1-i) \simeq$$

$$\frac{2}{m\alpha} \int_1^{m-1} \frac{1}{t} dt = \frac{2}{m\alpha} \log t \Big|_1^{m-1} = \frac{2\log(m-1)}{m\alpha}$$

5.5 Martingales

A *martingale* is a stochastic process that formalizes the idea of a fair game.

Definition

A stochastic process $\{Z_n, n \geq 1\}$ is said to be a martingale process if

$E[||Z_n||] \leq \infty$ for all n

and

$E[Z_{n+1}|Z_1, .., Z_n] = Z_n$

so if we think of Z_n as the fortune of a gambler than for a martingale process the expected fortune stays constant. Note

$E[Z_{n+1}]=E\{E[Z_{\_n+1}|Z_1,..,Z_n]\}=E[Z_n]=..=E[Z_1]$

5.5.0.1 Example

let X_1, X_2,.. be independent rv's with mean 0 and let $Z_n = X_1 + .. + X_n$, then

$$E[Z_n|Z_1,..,Z_{n-1}] = E[Z_{n-1} + X_n|Z_1,..,Z_{n-1}] =$$
$$E[Z_{n-1}|Z_1,..,Z_{n-1}] + E[X_n|Z_1,..,Z_{n-1}] =$$
$$Z_{n-1} + E\{X_n\} = Z_{n-1}$$

5.5.0.2 Example

let X_1, X_2,.. be independent rv's with mean 0 and let $Z_n = \prod_{i=1}^{n} = X_i$, then

$$E[Z_n|Z_1,..,Z_{n-1}] = E[Z_{n-1}X_n|Z_1,..,Z_{n-1}] =$$
$$Z_{n-1}E[X_n|Z_1,..,Z_{n-1}] =$$
$$Z_{n-1}E\{X_n\} = Z_{n-1}$$

Definition

A positive integer-valued, possibly infinite, rv N is said to be a *random time* for the process Z_n if the event {N=n} is determined by the random variables Z_1,.., Z_n. That is, knowing Z_1,..,Z_n tells us whether or not N=n. If $P(N < \infty) = 1$, then N is called a *stopping time.*

5.5.0.3 Example

say a gambler plays roulette. He starts with \$100 and bets \$1 in each round. He decides to stop if he reaches \$200 (or goes broke). Then if N is the number of games he plays N is stopping time.

Definition

Let N be a random time for the process $\{Z_n, n \geq 1\}$, then

$$\overline{Z}_n = \begin{cases} Z_n & \text{if } n \leq N \\ Z_N & \text{if } n > N \end{cases}$$

is called the *stopped process.*

Proposition

If N is a random time for the martingale $\{Z_n, n \geq 1\}$, then the stopped process is also a martingale.

without proof

Here is the main result for martingales:

Theorem (The Martingale Stopping Theorem}

If either

i. the stopped process is uniformly bounded
ii. N is bounded, or
iii. $E[N] < \infty$

and there is an $M < \infty$ such that

$E[|Z_{n+1} - Z_n||Z_1, .., Z_n] < M$

then $E[Z_n] = E[Z_1]$

In other words in a fair game if a gambler uses a stopping time to decide when to quit, then his expected final fortune is equal to his expected initial fortune. Thus in the sense of expected value, no successful gambling strategy is possible if one of the conditions of the theorem are satisfied.

5.5.0.4 Example

There are many supposedly "guaranteed" strategies on how to win in a casino. A popular one is this: bet \$1 on red in roullette. if you loose double your bet and so on. Say you loose 3 times and then win, then your net-win is -1+(-2)+(-4)+8=+1, so you win \$1. In fact as soon as you win a "sequence" of n losses followed by a win always ends with an overall win of \$1!

Great!

Unfortunately according to the martingale stopping theorem, even if roulette were a fair game this would still not work! Why not?

By the way, strategies of this type have a name, the St. Petersburg strategy.

Corollary (Wald's equation)

If $X_i, i \geq 1$ are iid with $E[|X|] < \infty$ and if N is a stopping time for X_1, X_2, .. with $E[N] < \infty$, then

$$E\left[\sum_{i=1}^{N} X_i\right] = E[N]E[X_1]$$

5.5.0.5 Example

suppose a computer randomly generates integers. Let N be the number of integers it has to generate before we see a predetermined sequence, for example say 0 0 0 0.

To compute E[N] imagine a sequence of gambles, each initially having 1 unit, playing a fair game. Gambler i begins playing at the beginning of the i^{th} day bets 1 unit that the value on that day is equal to 0. If he wins (and so has 10 units) he bets those 10 units on the second day, again to get 0. If he wins again he will have 100 units and so on. If 0 0 0 0 happens he wins \$10000-\$1 = \$9999, if any of his bets fail he looses \$1.

At the beginning of each day another gambler starts to play. If we let X_n denote the total winings of the casino after the n^{th} day, then since all bets are fair X_n is a martingale with mean 0.

Let N denote the time until 0 0 0 0 happens. Now at the end of day N each of the gamblers 1, .., N-4 would have lost 41, gambler N-3 would have won \$9999, gambler N-2 would have won \$999, gambler N-1 \$99 and gambler N \$9. So

$X_N = N - 4 - 9999 - 999 - 99 - 9 = N - 11110 = 0$

so E[N]=11110

5.6 Brownian Motion and Stationary Processes

In 1827 the English botanist Robert Brown observed that microscopic pollen grains suspended in water perform a continual swarming motion.

This phenomenon was first explained by Einstein in 1905 who said the motion comes from the pollen being hit by the molecules in the surrounding water. The mathematical derivation of the Brownian motion process was first done by Wiener in 1918, and in his honor it is often called the Wiener process.

Brownian motion is a continuous-time continuous state-space stochastic process defined as follows: the process $\{X(t), t \geq 0\}$ is a Brownian motion process iff

1. X(0)=0
2. $\{X(t), t \geq 0\}$ has stationary and independent increments
3. for all t>0 $X(t) \sim N(0, \sigma\sqrt{t})$

One way to visualize a Brownian motion process is as the limit of symmetric random walks: Let $\{Z_n, n \geq 1\}$ be the symmetric random walk on the integers. If we now speed the process up and scale the jumps accordingly we get a Brownian motion process in the limit. More precisely, suppose we jump every δt and make a jump of size δx. If we let Z(t) denote the position of the process at time t then

$$Z(t) = \Delta x Z_1 + \Delta x Z_2 + .. + \Delta x Z_{[t/\Delta t]}$$

Now $EZ(t) = 0$ and $VarZ(t) = \sum_{i=1}^{[t/\Delta t]} (\Delta x)^2 VarZ_k = VarZ(t) = \sum_{i=1}^{[t/\Delta t]} (\Delta x)^2 1$

$$VarZ(t) = \sum_{i=1}^{[t/\Delta t]} (\Delta x)^2 VarZ_k = VarZ(t) = \sum_{i=1}^{[t/\Delta t]} (\Delta x)^2 1 = (\Delta x)^2 [t/\Delta t]$$

Now set $\Delta x = \sigma \sqrt{\Delta t}$ and let $\Delta t \to 0$ then

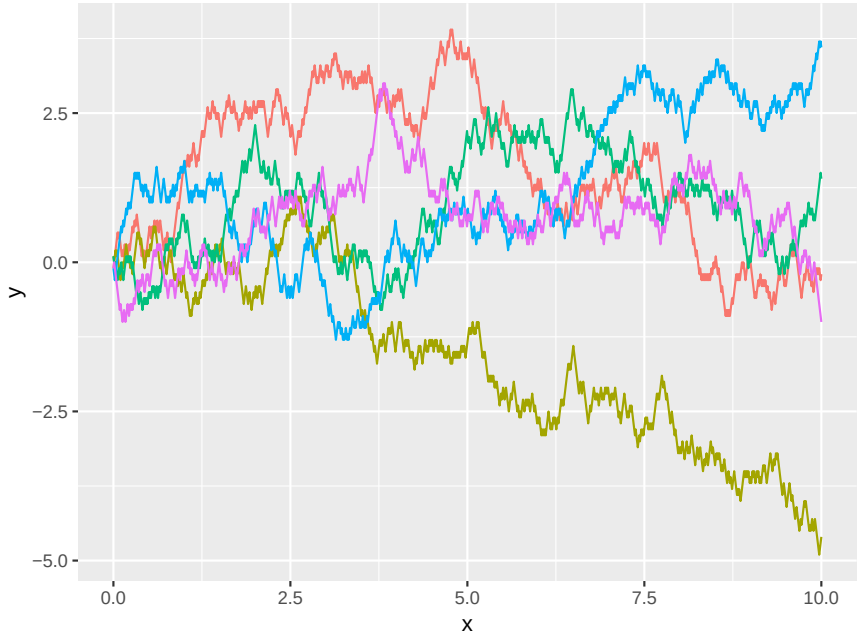
$VarZ(t) \to s^2 t$

and by the CLT $Z(t) \to N$ in distribution

The Brownian motion process plays a role in the theory of stochastic processes similar to the role of the normal distribution in the theory of random variables.

If $\sigma = 1$ the process is called standard Brownian motion.

Next we draw sample paths of a standard Brownian motion process.



Here are some properties of Brownian motion:

1. BM will eventually hit any and every real value, no matter how large or how negative! It may be a million units above the axis, but it will (with probability 1) be back down again to 0, by some later time.
2. Once BM hits zero (or any particular value), it immediately hits it again infinitely often, and then again from time to time in the future.
3. Spatial Homogeneity: B(t) + x for any $x \in \mathbb{R}$ is a BM started at x.
4. Symmetry: -B(t) is a Brownian motion

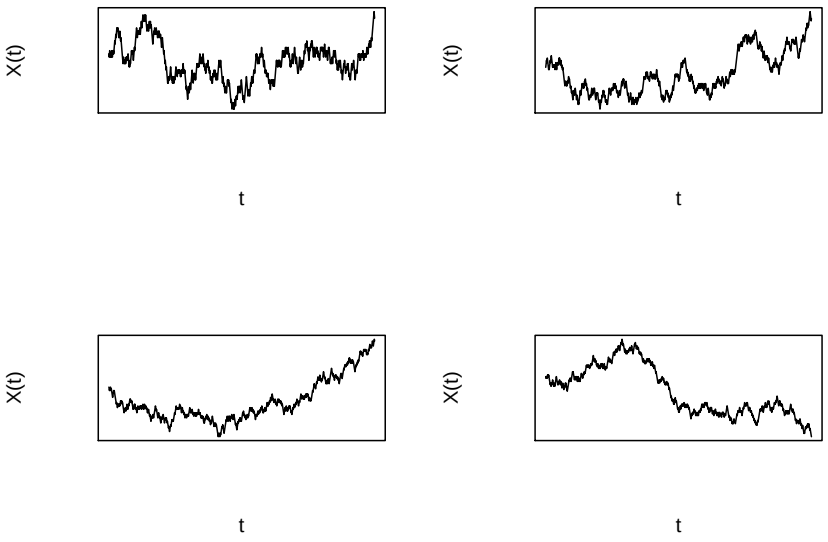
5. Scaling: $\sqrt{c}B(t/c)$ for any c > 0 is a BM
6. Time inversion:

$$Z(t) = \begin{cases} 0 & \text{if } t = 0 \\ tB(1/t) & \text{if } t \neq 0 \end{cases}$$

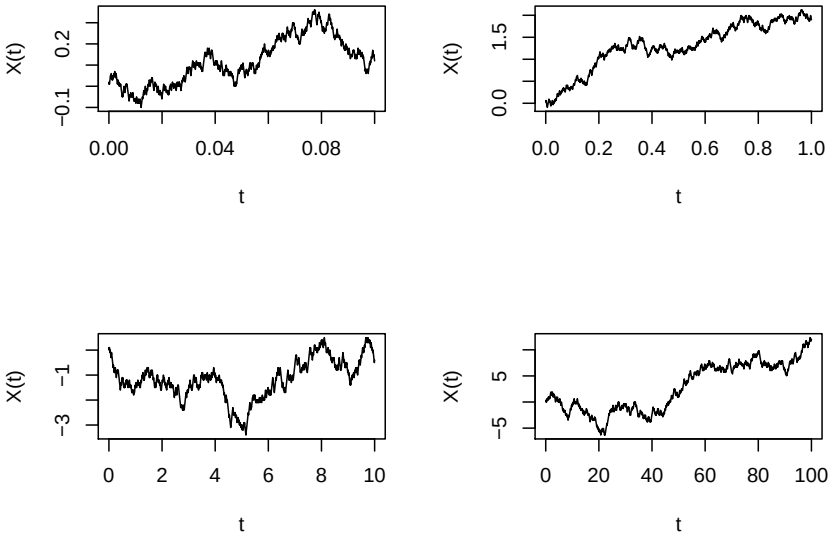
is a BM.

7. BM is time reversible
8. BM is self-similar (that is its paths are fractals):

Consider the four graphs of BM paths drawn here:



They are drawn without labeling on the axis. They appear completely the same, but if we add the tick marks



we see that the scales are completely different. This phenomena is called *self-similarity*.

Brownian Motion is an example of a process that has a fractal dimension of 2. One of its occurrences is in microscopic particles and is the result of random jostling by water molecules (if water is the medium). So in moving from a given location in space to any other, the path taken by the particle is almost certain to fill the whole space before it reaches the exact point that is the 'destination' (hence the fractal dimension of 2).

9. The last property of BM is so much fun we will give it its own section:

5.6.1 Continuity and Differentiability

When studying a continuous-time stochastic process it is often useful to think of any particular realization of the process as a function. Say S is the sample space of the process, that is the set of all possible paths $\{X(t), t \geq 0\}$, and let $\omega \in S$.

Then $f(t) = X(t, \omega)$ is a function. (Usually we suppress ω, though).

In the case of BM, what are the properties of a typical realization B(t)? First let's look at continuity:

Now by the definition we have that

$B(t+h) - B(t) \sim N(0, \sqrt{h})$

therefore E[(B(t+h)-B(t))^2] = h, and so the size of an increment of |B(t+h)-B(t)| is about $\sqrt{h}$. So as $h \to 0$, $\sqrt{h} \to 0$ which implies continuity.

How about differentiability? Now we have

$$\frac{\partial B}{\partial t} = \lim_{h \to 0} \frac{B(t+h) - B(t)}{h} \approx \lim_{h \to 0} \frac{\sqrt{h}}{h} \to \infty$$

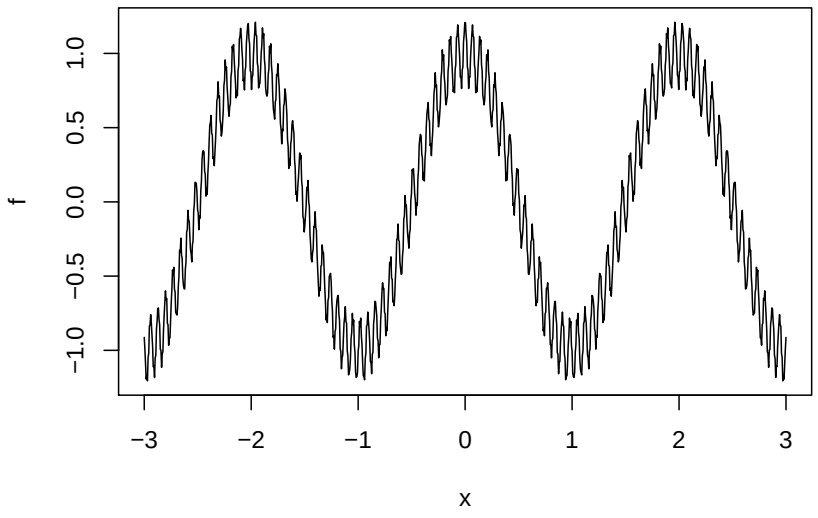
and we see that BM is nowhere differentiable!

(Of course this is rather heuristic but it can be made rigorous).

The idea of functions that are continuous but nowhere differentiable has a very interesting history. It was first discussed in 1806 by André Marie Ampère and trying to show that such a function exists was one of the main open problems during the 19$^{\text{th}}$ century. More than fifty years later it was Karl Theodor Wilhelm Weierstrass who finally succeeded in constructing such a function as follows:

$$W(x) = \sum_{k=0}^{\infty} b^k \cos(a^k \pi x)$$

Here is what this looks like for b = .2 and a = 5 + 7.5π(and a finite sum!)



The hard part here was not the construction but to show that the function existed! For the proof he developed what is now known as the *Stone-Weierstrass* theorem.

Shorty after that a new branch of mathematics called *functional analysis* was developed. It studies the properties of real-valued functions on function space. Here are some examples of such functionals:

$$\Phi(f) = f(t) \text{ for some } t$$
$$\Phi(f) = \int_{-\infty}^{\infty} f(t)dt$$
$$\Phi(f) = \int_{-\infty}^{\infty} f^2(t)dt$$
$$\Phi(f) = ||f||$$

Of course one needs to specify the space of functions for which a certain functional applies. Standard "function spaces" are C, the space of all continuous functions and C^1, the space of all continuous functions with a continuous derivative.

One of the results of functional analysis is that C is much larger than C^1, actually of a higher order of infinity, shown with the *Baire category theorem*.

So consider the following "experiment": pick any continuous function in C. Then the probability that it has a continuous derivative anywhere is 0! So functions such as Weierstrass (or the paths of BM) are not the exception, they are the rule. Or, all the functions we study in mathematics are completely irrelevant in nature!

5.6.2 Stochastic Differential Equations

Theorem

let g(x) be a continuous function and let $\{B(t), t\ ge0\}$ be a standard Brownian motion. For each fixed t>0, there exists a random variable

$$\Psi(g) = \int_0^t g(x)dB(x)$$

which is the limit of the approximating sums

$$\Psi_n(g) = \sum_{k=1}^{2^n} g(\frac{k}{2^n}t)\Big[B(\frac{k}{2^n}t) - B(\frac{k-1}{2^n}t) \Big]$$

as $n \to \infty$. The random variable $\Psi(g)$ is normally distributed with mean 0 and variance

$$Var[\Psi(g)] = \int_0^r g^2(x)dx$$

If f(x) is another continuous function of x, then $\Psi(f)$ and $\Psi(g)$ have a joint normal distribution with covariance

$$E[\Psi(f)\Psi(g)] = \int_0^t f(x)g(x)dx$$

There is a version of the integration by parts formula:

$$\int_0^t g(x)dB(x) = g(t)B(t) - \int_0^t B(x)g'(x)dx$$

so

$$\int_0^t 1dB(x) = 1B(t) - \int_0^t B(x)g'(x)dx = B(t)$$

integrated Brownian motion:

$$\int_0^t (t-x)dB(x) = (t-t)B(t) - \int_0^t B(x)(-1)dx =$$

$$\int_0^t B(x)dx$$

$$E\left[\int_0^t B(x)dx\right] = 0$$

$$Var\left[\int_0^t B(x)dx\right] = \int_0^t (t-x)^2 dx =$$

$$-\frac{1}{3}(t-x)^3\Big|_0^t = \frac{t^3}{3}$$