CS 70 Discrete Mathematics and Probability Theory
Spring 2017 Course Notes JEAN WALRAND - Probability Review

The objective of these notes is to enable you to check your understanding and knowledge of the probability concepts studied in this course. These notes are not meant to be self-contained. They are mostly a check-list. The lecture slides and lecture notes contain the details. You should try to invent problems and see if you can solve them.

1 Probability Space

Review

The *sample space* Ω is the set of possible *outcomes* of some random experiment. The experiment selects *one* outcome (only one!) $\omega \in \Omega$. The *probability* that it selects ω is $Pr[\omega]$. For now (until Section 8), Ω is either finite or countable (i.e., its elements can be numbered $\omega_1, \omega_2, \ldots$). For each $\omega \in \Omega$, one has $Pr[\omega] \geq 0$ and $\sum_\omega Pr[\omega] = 1$. A finite probability space is said to be *uniform* when all the outcomes have the same probability $1/|\Omega|$.

An *event* is a subset A of Ω, i.e., a set of outcomes. One defines the *probability of the event* A as $Pr[A] = \sum_{\omega \in A} Pr[\omega]$. Note that if $A \cap B = \emptyset$, then $Pr[A \cup B] = Pr[A] + Pr[B]$. By induction, if the events A_m are pairwise disjoint, then $Pr[A_1 \cup \cdots \cup A_n] = Pr[A_1] + \cdots + Pr[A_n]$ for all finite n. One can also show that this identity holds even for infinite n, i.e., $Pr[\cup_{m \geq 1} A_m] = \sum_{m \geq 1} Pr[A_m]$ if the events A_m are pairwise disjoint. One says that probability is *countably additive*.

By induction, one also sees that for arbitrary events A_m one has $Pr[A_1 \cup \cdots \cup A_n] \leq Pr[A_1] + \cdots + Pr[A_n]$. This inequality is called the *union bound*. Also, for any two events A and B one has $Pr[A \cup B] = Pr[A] + Pr[B] - Pr[A \cap B]$. This is called the *inclusion-exclusion* identity.

Symmetry is a powerful argument to determine that two events have the same probability. For instance, say that there is a bag with 100 red balls and 200 blue balls. You pick five balls without replacement. The probability that the fifth ball is red is $1/3$. Indeed, this is the same as the probability that the first ball is red. To see this, think of arranging the 300 balls in a random order (permutation), so that all the permutations are equally likely. They remain so if you interchange the first and fifth ball. As another example, you deal five cards from a well-shuffled 52-card deck. The probability that the third card is red is $1/2$. The probability that the fourth card is an ace is $4/52 = 1/13$. Of course, the likelihood that the fourth card is an ace depends on the first three cards. Say that there is only one specific chocolate that you like in a box that you share with some friends by everyone in turn picking one chocolate randomly uniformly and without replacement. If you are the last one to pick, you might think that you are less likely to get your favorite chocolate than if you are first. Not so, by symmetry.

Examples

1. $\Omega = \{1,2,3,4\}$ be a uniform probability space. For $A = \{1,2,3\}$ and $B = \{3,4\}$, one finds that

   $$Pr[A] = 3/4, Pr[B] = 1/2, Pr[A \cap B] = 1/4, Pr[A \setminus B] = 1/2, Pr[\bar{A}] = 1/4, Pr[A \cup B] = 1, Pr[A \Delta B] = 3/4.$$

2. $\Omega = \{HH, HT, TH, TT\}$ be a uniform probability space. This probability space describes the random

experiment "flipping a fair coin twice." Note that the outcome of the experiment is one of the elements of Ω. It would be fundamentally wrong to say that this experiment is described by $\Omega = \{H, T\}$ where one picks two elements of Ω when one performs the experiment. It is wrong because we want to describe how the coin flips are related. For instance, if we glue the two coins together so that they both land on the same face, then the probability space becomes the uniform space $\{HH, TT\}$. We lose this relationship if we think of $\Omega = \{H, T\}$ and we select two outcomes. Hence, it is essential that ω describes the full outcome of the complete experiment, i.e., each ω describes the two coin flips. *Make sure you understand this point.* What would the probability space be if you flipped the fair coin 100 times?

3. $\Omega = \{H, T\}$ with $Pr[H] = 0.6$ and $Pr[T] = 0.4$. This probability space describes the random experiment "flipping a biased coin once."

4. $\Omega = \{1, 2, \ldots\}$ with $Pr[n] = (1-p)^{n-1}p$ for $n \geq 1$. This probability space describes the random experiment "flipping a biased coin until it yields the first H."

2 Conditional Probability and Independence

Review

For two events A and B in Ω, one defines the *conditional probability of A given B* as $Pr[A|B] := Pr[A \cap B]/Pr[B]$. We say that A and B are independent if $Pr[A \cap B] = Pr[A]Pr[B]$. Three events A, B, C are mutually independent if they are independent two by two and if, in addition, $Pr[A \cap B \cap C] = Pr[A]Pr[B]Pr[C]$. More generally, the events $\{A_i, i \in I\}$ are mutually independent if

$$Pr[A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_n}] = Pr[A_{i_1}] \times \cdots \times Pr[A_{i_n}]$$

for all $n \geq 2$ and any $\{i_1, \ldots, i_n\} \subset I$. We say that the events A and B are positively (resp., negatively) correlated if $Pr[A \cap B] > Pr[A]Pr[B]$ (resp., $Pr[A \cap B] < Pr[A]Pr[B]$). Thus, if A and B are positively correlated, then $Pr[B|A] > Pr[B]$. Note that this does not imply that A causes B. For instance, one also has $Pr[A|B] > Pr[A]$, so is it A that causes B or B that causes A? In fact, it may be neither: the events could have a common cause, like people who drive a Tesla being more likely to own a vacation home, and vice-versa.

Examples

1. Let $\Omega = \{1, 2, 3, 4\}$ be a uniform probability space. Let also $A = \{1, 2, 3\}$ and $B = \{3, 4\}$. Then $Pr[A|B] = 1/2, Pr[B|A] = 1/3$. Here, A and B are negatively correlated.

2. Roll a balanced six-sided die. The probability that the outcome is 6 given that it is larger than 4 is $1/2$.

3. A couple has two kids, at least one of which is a girl. The probability that they have two girls is $Pr[GG]/Pr[GB, BG, GG] = 1/3$. Perhaps surprisingly, it is not $1/2$.

4. Roll a balanced six-sided die twice. Let A be the event that the first roll yields m for some $m \in \{1, \ldots, 6\}$ and B the event that the second roll yields n for some n in $\{1, \ldots, 6\}$. Then A and B are independent. We say that the two rolls are independent. (See Section 4.)

5. Let $\Omega = \{1, \ldots, 8\}$ be a uniform probability space. Let $A = \{1, 2, 3, 4\}, B = \{4, 5\}, C = \{1, 2, 5, 6\}$. Then A, B, C are pairwise independent, not mutually. Note that A, B, C would not be pairwise independent if the uniform probability space were $\{1, \ldots, 6\}$.

6. Let $\Omega = \{1,\ldots,8\}$ be a uniform probability space. Then $A = \{1,2,3,4\}, B = \{2,4,6,8\}, C = \{2,3,6,7\}$ are mutually independent.

7. Let $\{A_i, i \in I\}$ be mutually independent. Let also J and K be disjoint finite subsets of I. Define E to be an event obtained by performing set operations on the events $\{A_i, i \in J\}$ and F an event obtained by performing set operations on the events $\{A_i, i \in K\}$. For instance, $E = (A_1 \cup A_2) \setminus (A_3 \cap A_4)$ and $F = (A_5 \Delta A_6) \setminus A_7$. Then E and F are independent. Here is a general proof, for arbitrary E and F. First look at the case of three events A_1, A_2, A_3. When you draw the Venn diagram, you find that there 8 disjoint sets that make up Ω: $B_1 = A_1 \cap A_2 \cap A_3, B_2 = A_1 \cap A_2^c \cap A_3^c, \ldots, B_8 = A_1^c \cap A_2^c \cap A_3^c$. These events are of the form $D_1 \cap D_2 \cap D_3$ where each D_i is either A_i or A_i^c. Any event that you can create by performing set operations on A_1, A_2, A_3 is a union of some of the eight sets B_j. More generally, define the *atoms* B_n of $\{A_i, i \in J\}$ to be all the events of the form $B_n = \cap_{i \in J} D_i$ where each D_i is either A_i or A_i^c. Similarly, define the atoms C_m of $\{A_i, i \in K\}$. These atoms B_n are pairwise disjoint. Similarly, the atoms C_m are pairwise disjoint. Also, each B_n is independent of every C_m. Moreover, E is a union of atoms B_n and F is a union of atoms C_m. Thus, $Pr[E \cap F] = Pr[\cup_{m,n}(B_n \cap C_m)] = \sum_{m,n} Pr[B_n \cap C_m] = \sum_{m,n} Pr[B_n]Pr[C_m] = (\sum_n Pr[B_n])(\sum_m Pr[C_m]) = Pr[E]Pr[F]$.

3 Bayes' Rule

Review

The setup is that Ω is a probability space, $\{A_1,\ldots,A_n\}$ are a partition of Ω and B is some event. Then

$$Pr[A_m|B] = \frac{Pr[A_m]Pr[B|A_m]}{\sum_{k=1}^n Pr[A_k]Pr[B|A_k]}, m = 1,\ldots,n.$$

The point of this formula is as follows. We think of the A_m as possible pairwise exclusive *circumstances* under which the *symptom* B can occur. One knows the *prior* probability $Pr[A_k]$ of every circumstance. One also knows the conditional probability $Pr[B|A_k]$ that B occurs under circumstance A_k . Bayes' Rule tells us how to compute the likelihood $Pr[A_k|B]$ that circumstance A_k is in effect given that B occurs. Think of A_k as a disease and B as a symptom such as a fever, or a suspicious X-ray reading, or some abnormal value for a test. One calls $Pr[A_m|B]$ the *posterior* probability of A_m given B. Thus, Bayes' Rule computes the posterior probabilities $Pr[A_m|B]$ given the prior probabilities $Pr[A_m]$ and the conditional probabilities $Pr[B|A_m]$.

The event A_m with the maximum value of $Pr[A_m|B]$ is said to be the *Maximum A Posteriori* (MAP) estimate of the circumstance given the symptom, i.e., the most likely circumstance A_m given that B occurs. The event A_m with the maximum value of $Pr[B|A_m]$ is said to be the *Maximum Likelihood Estimate* (MLE) of the circumstance given the symptom. The MLE and the MAP coincide if the priors are uniform, i.e., if $Pr[A_m] = 1/n$ for $m = 1,\ldots,n$. Otherwise, they generally differ. Thus, the MAP is the most likely A_m given B while the MLE is the A_m that makes B most likely. For instance, Ebola is the MLE circumstance of diarrhea whereas food poisoning may be the MAP.

Examples

1. A coin is equally likely to be fair or biased with $Pr[H] = 0.7$. You flip it once and get H. The posterior probability that the coin is fair is

$$\frac{0.5 \times 0.5}{0.5 \times 0.5 + 0.5 \times 0.7} \approx 0.42.$$

Thus, the coin is a bit more likely to be biased since it produced one H. The probability that the next coin flip yields H is then

$$0.42 \times 0.5 + 0.58 \times 0.7 \approx 0.62.$$

If the first flip yields T, then you can verify that the likelihood that the second flip yields H is 0.575.

2. A coin is fair with probability 0.7, otherwise it is such that $Pr[H] = 0.6$. You flip the coin ten times and get 8 heads. The posterior probability that the coin is fair is

$$\frac{0.7\binom{10}{8}(0.5)^8(0.5)^2}{0.7\binom{10}{8}(0.5)^8(0.5)^2 + 0.3\binom{10}{8}(0.6)^8(0.4)^2} \approx 0.1.$$

Thus, it is much more likely that the coin is biased since it produced so many heads. Consequently, the conditional probability that flip 11 yields heads is approximately $0.1 \times 0.5 + 0.9 \times 0.6 \approx 0.59$.

3. A coin is fair with probability 0.5, otherwise it is such that $Pr[H] = 0.6$. You flip the coin repeatedly and it takes 10 flips until the first H. The posterior probability that it is fair is

$$\frac{0.5 \times (0.5)^9(0.5)}{0.5 \times (0.5)^9(0.5) + 0.5 \times (0.4)^9(0.6)} \approx 0.86.$$

Thus, the conditional probability that the coin is fair is much larger than the prior probability 0.5 because it took so long to get the first H. Consequently, the probability that the next flip, i.e., flip 11, yields heads is approximately $0.86 \times 0.5 + 0.14 \times 0.6 \approx 0.51$.

4. There are two envelopes. The first one contains five checks in the amounts of $\{1,2,5,5,6\}$ and the second contains four checks in the amounts $\{2,5,6,8\}$. You pick an envelope at random and then pick a check at random. Given that the check you got is in the amount of 5, the posterior probability that you picked it from the first envelope is

$$\frac{0.5 \times (2/5)}{0.5 \times (2/5) + 0.5 \times (1/4)} \approx 0.62.$$

Thus, if you are offered to either keep the envelope you selected or to switch to the other envelope, you should switch because the second envelope contains more money.

4 Random Variables and Expectation

Review

A *random variable* is a *real-valued function of the outcome of a random experiment*. Thus, there is a probability space Ω with $Pr[\omega]$ defined for each $\omega \in \Omega$. This probability space defines the random experiment. A random variable X is a function $X : \Omega \to \Re$ that assigns a real number $X(\omega)$ to each outcome $\omega \in \Omega$. For $A \subset \Re$, one defines $Pr[X \in A] := Pr[X^{-1}(A)]$ where $X^{-1}(A) := \{\omega \in \Omega \mid X(\omega) \in A\}$ is the *inverse image* of A under the function X. Similarly, $Pr[X = a] := Pr[X^{-1}(\{a\})]$. The *distribution* of a random variable X is the set of its possible values and their probability. This may seem abstract, so here is a concrete example. Say that a bag contains 100 balls. Among those, 23 are marked with the value 5 and the others are marked with different values. Let ω be the ball you pick and $X(\omega)$ the value marked on the ball. Then $Pr[X = 5] = Pr[X^{-1}(5)] = 23/100$. Here, $X^{-1}(5)$ is the set of 23 balls marked with the value 5.

If X, Y are two random variables on Ω and $g : \Re^2 \to \Re$ is a function, then $g(X,Y)$ is a new random variable that assigns the real number $g(X(\omega), Y(\omega))$ to $\omega \in \Omega$.

The expected value of X is $E[X] := \sum_x xPr[X=x] = \sum_\omega X(\omega)Pr[\omega]$. Thus, $E[g(X,Y)] = \sum_\omega g(X(\omega),Y(\omega))Pr[\omega]$. Expectation is linear: $E[aX+bY] = aE[X]+bE[Y]$.

Given an event A, one defines the *conditional expectation* of X given A as $E[X|A] := \sum_x xPr[X=x|A] = \sum_\omega X(\omega)Pr[\omega|A]$. Assume that $\{A_1,\ldots,A_n\}$ is a partition of Ω. Then $E[X] = \sum_m E[X|A_m]Pr[A_m]$. In particular, $E[X] = \sum_y E[X|Y=y]Pr[Y=y]$. If we define the *conditional expectation* of X given Y as $E[X|Y] := g(Y)$ where $g(y) := E[X|Y=y]$, then we see that $E[X] = E[E[X|Y]]$.

One also finds that $E[Xg(Y)|Y] = g(Y)E[X|Y]$, so that $E[(X-E[X|Y])g(Y)] = 0$, an identity that shows that the *estimation error* $X-E[X|Y]$ is orthogonal to any function $g(Y)$. This *projection property* implies that $E[(X-h(Y))^2]$ is minimized by choosing $h(Y) = E[X|Y]$. We say that the conditional expectation $E[X|Y]$ is the *Minimum Mean Squares Estimate* (MMSE) of X given Y. Another way to appreciate this property is to note that $E[(X-h(Y))^2] = \sum_y Pr[Y=y]E[(X-h(y))^2|Y=y]$ and that, for each y, $E[(X-h(y))^2|Y=y]$ is minimized by choosing $h(y) = E[X|Y=y]$, as we see by setting to zero the derivative with respect to a of $E[(X-a)^2|Y=y] = E[X^2|Y=y] - 2aE[X|Y=y] + 2a^2$.

We say that the random variables X and Y are independent if (and only if) $Pr[X=x,Y=y] = Pr[X=x]Pr[Y=y]$ for all x,y. Equivalently, they are independent if and only if $Pr[X\in A, Y\in B] = Pr[X\in A]Pr[Y\in B]$ for all $A,B\subset\Re$. If X and Y are independent, then $g(X)$ and $h(Y)$ are independent for any functions g and h. More generally, the random variables $\{X_i, i\in I\}$ are *mutually independent* if $Pr[X_{i_1}\in A_1,\ldots,X_{i_n}\in A_n] = Pr[X_{i_1}\in A_1]\times\cdots\times Pr[X_{i_n}\in A_n]$ for all finite n, all $i_1,\ldots,i_n\in I$ and all sets $A_1,\ldots,A_n$ in $\Re$. If the random variables X_i are mutually independent, then $E[X_{i_1}X_{i_2}\cdots X_{i_n}] = E[X_{i_1}]\cdots E[X_{i_n}]$. The converse is not true.

Examples

1. Flip a biased coin with $Pr[H] = p$. Let $X=1$ if the outcome is H and 0 otherwise. Then the distribution of X is called Bernoulli with parameter p and one writes $X = B(p)$. Thus, $E[X] = p$ and $E[X^2] = E[X] = p$.

2. Let X_m for $m = 1,\ldots,n$ be i.i.d. $B(p)$. Then $X = X_1+\ldots+X_n = B(n,p)$. Thus, $E[X] = np$.

3. Roll a balanced six-sided die once and let X be the number of pips. Then $E[X] = \sum_{m=1}^6 m(1/6) = 3.5$.

4. Roll a balanced six-sided die twice. Then $\Omega = \{1,\ldots,6\}^2$ with $Pr[\omega] = 1/36$ for all $\omega\in\Omega$. Let $X((a,b)) = a+b$ for $(a,b)\in\Omega$. Then $E[X] = 2\times 3.5 = 7$, by linearity of expectation.

5. Same experiment as above. Let $Y((a,b)) = \min\{a,b\}$ and $Z((a,b)) = \max\{a,b\}$ for $(a,b)\in\Omega$. Then (it may be useful to draw a picture) $Pr[Y=1] = 11/36, Pr[Y=2] = 9/36, Pr[Y=3] = 7/36, Pr[Y=4] = 5/36, Pr[Y=5] = 3/36, Pr[Y=6] = 1/36$. Also, $Pr[Z=m] = Pr[Y=7-m]$, by symmetry. Hence, $E[Y] = 1(11/36)+2(9/36)+\cdots+6(1/36) = 91/36 \approx 2.52$. Also, $Y+Z$ is the total number of pips on the two rolls, so that $E[Y+Z] = 7$ and it follows that $E[Z] = 7-E[Y] \approx 4.48$.

6. Pick a number uniformly in $\{1,\ldots,n\}$. Then $\Omega = \{1,\ldots,n\}$ with $Pr[m] = 1/n$ for $m\in\Omega$. Define $X(m) = 2+3m+4m^2$.

7. Let $A\subset\Omega$ and define $X(\omega) = 1\{\omega\in A\}$ for $\omega\in\Omega$. The random variable X is called the *indicator* of the event A. Sometimes we write it as $1_A(\omega)$.

8. Flip n times a biased coin with $Pr[H] = p$. Then $\Omega = \{H,T\}^n$ and $Pr[\omega] = p^m(1-p)^{n-m}$ if ω has m heads. Define $X(\omega)$ to be the number of heads in ω. Then $Pr[X=m] = \binom{n}{m}p^m(1-p)^{n-m}$. This is the *binomial distribution* and we write it as $B(n,p)$.

9. Flip a biased coin with $Pr[H] = p$ until you get a first H. Then $\Omega = \{1, 2, \ldots\}$ and $Pr[n] = (1-p)^{n-1}p$ for $n \geq 1$. Let $X(n) = n$. The distribution of X is called the *Geometric distribution with parameter p*. We designate it by $G(p)$.

10. A monkey is typing away on a keyboard that has only the 26 letters. After he types 10^{12} letters, let X be the number of times that the word 'walrand' appears. Then $E[X] = (10^{12} - 7)(26)^{-7} \approx 124$. Indeed, $X = X_1 + \cdots + X_n$ with $n = 10^{12} - 7$ where X_m is the indicator of the event that the word 'walrand' appears starting with the m-th letter. One has $E[X_m] = (26)^{-7}$ and $E[X] = nE[X_1]$, by linearity of expectation.

11. An elevator loads n people on the ground floor of a building with $k+1$ floors (including the ground floor 0). Each person chooses one of the k other floors independently, with equal probabilities. The probability that no one chooses a particular floor is then $(1-1/k)^n$. Consequently, if X_m is the indicator that someone chose floor m, one sees that $E[X_m] = 1 - (1-1/k)^n$. Let $X = X_1 + \cdots + X_k$ be the number of different floors that people picked. Then, by linearity of expectation, $E[X] = kE[X_1] = k[1-(1-1/k)^n]$.

12. Let $X = G(p)$, so that $Pr[X = n] = (1-p)^{n-1}p$ for $n \geq 1$. Then $E[X] = 1/p$.

13. Let $X = G(p)$ and $Y = G(q)$ be independent. Then $\min\{X, Y\} = G(r)$ with $r = 1 - (1-p)(1-q)$.

14. We say that X is Poisson with parameter $\lambda > 0$ if $Pr[X = n] = (\lambda^n)/(n!)\exp\{-\lambda\}$ for $n \geq 0$. We write $X = P(\lambda)$. Then $E[X] = \lambda$.

15. Let $X = P(\lambda)$ and $Y = P(\mu)$ be independent. Then $X + Y = P(\lambda + \mu)$.

16. Flip a coin. With probability p, the outcome is H and one defines $Z = X$; otherwise, one defines $Z = Y$. Here, X and Y are two random variables that are independent of the coin flip. Find $E[Z]$ and $var(Z)$ in terms of p and the mean and variance of X and Y.

17. Assume that $E[X_{n+1}|X_n] = a + bX_n$ for $n \geq 1$. Taking expectation, we get $E[X_{n+1}] = a + bE[X_n]$. Hence, $E[X_2] = a + bE[X_1], E[X_3] = a + bE[X_2] = a + b(a + bE[X_1]) = a(1+b) + b^2E[X_1]$. Continuing in this way, we find that $E[X_n] = a(1 + b + \cdots + b^{n-2}) + b^{n-1}E[X_1] = a(1-b^{n-1})/(1-b) + b^{n-1}E[X_1]$.

18. Diluting. There is a bin with 100 red balls. At step 1, we pick a ball from the bin and replace it with a blue ball. We now have $X_1 = 99$ red balls and 1 blue ball. We continue in this way and let X_n be the number of red balls in the bin after n steps. At step $n+1$, the likelihood that we pick a red ball is $X_n/100$. Hence, one has $E[X_{n+1}|X_n] = X_n - X_n/100 = 0.99X_n$. Hence, $E[X_n] = 100(0.99)^n$ for $n \geq 1$.

19. Mixing. There is a bin with 100 red balls and one with 100 blue balls. At each step, one picks a ball from each bin and puts it in the other bin. Let X_n be the number of red balls in the first bin at the end of n steps. Then $E[X_{n+1}|X_n] = X_n - (X_n/100)(X_n/100) + (1 - X_n/100)(1 - X_n/100) = 1 + bX_n$ with $b = 49/50$. Hence, $E[X_n] = (1-b^{n-1})/(1-b) + b^{n-1}99$.

20. Going Viral. Assume everyone on Tweeter has a random number of friends that has mean μ. You tweet a rumor to those friends. Each of them retweets independently with probability p to each of their friends, and so on. Let X_n be the number of people who retweet the rumor at step n. Given $X_n = k$ and the number of friends $Y_1, \ldots, Y_k$ of these k people, we see that $X_{n+1} = B(Y_1 + \cdots + Y_k, p)$. Thus, $E[Xn+1|X_n = k, Y_1, \ldots, Y_k] = p(Y_1 + \cdots + Y_k)$. Hence, $E[X_{n+1}|X_n = k] = E[p(Y_1 + \cdots + Y_k)] = pk\mu = p\mu X_n$. Consequently, $E[X_{n+1}] = p\mu E[X_n]$ and we conclude that $E[X_n] = (p\mu)^{n-1}$ for $n \geq 1$ (because $X_1 = 1$). If $X = X_1 + X_2 + \cdots$, we see that $E[X] < \infty$ if and only if $p\mu < 1$.

5 Covariance, Variance, Linear Regression and Estimation

Review

The *covariance* of two random variables X and Y is defined as $cov(X,Y) = E[XY] - E[X]E[Y]$. The random variables X and Y are said to be *uncorrelated* if $cov(X,Y) = 0$, *positively correlated* if $cov(X,Y) > 0$ and *negatively correlated* if $cov(X,Y) < 0$. Note that if X and Y are independent, then $cov(X,Y) = 0$. The converse is not true. The *variance* of a random variable is defined as $var(X) = E[X^2] - E[X]^2$. One has $var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$. In particular, if X and Y are independent, then $var(X+Y) = var(X) + var(Y)$. Also, $var(a+bX) = b^2 var(X)$ and $cov(aX+bY, cV+dW) = ac.cov(X,V) + ad.cov(X,W) + bc.cov(Y,V) + bd.cov(Y,W)$.

If $f : \Re \to [0,\infty)$ is nondecreasing and $f(a) > 0$, then $Pr[X \geq a] \leq E[f(X)]/f(a)$. This is Markov's inequality. Chebyshev's inequality states that $Pr[|X - E[X]| \geq a] \leq var[X]/a^2$.

The linear function of X that minimizes $E[(Y - a - bX)^2]$ is $L[Y|X] := E[Y] + [cov(X,Y)/var(X)](X - E[X])$. We call this function the LLSE (linear least squares estimate) of Y given X. In particular, if (X,Y) is distributed uniformly in the set $\{(X_m, Y_m), m = 1, \ldots, n\}$, then $L[Y|X]$ is called the *linear regression* of Y over X. Note that this LR is non-Bayesian: it does not assume a prior distribution of (X,Y) but uses only the observed sample values.

Let $\{X_m, m \geq 1\}$ be i.i.d. with mean μ and variance σ^2 and $A_n := (X_1 + \cdots + X_n)/n$. Then, Chebyshev implies that $Pr[|A_n - \mu| \geq \varepsilon] \leq \sigma^2/(n\varepsilon^2) = \sigma^2/(n\varepsilon^2)$. Thus, for all $\varepsilon > 0$ one has $Pr[|A_n - \mu| \geq \varepsilon] \to 0$ as $n \to \infty$. This result is called the *weak law of large numbers* (WLLN). Using the same inequality and choosing ε so that $n\varepsilon^2 = 20\sigma^2$, i.e., $\varepsilon = 4.5\sigma/\sqrt{n}$, we find that $Pr[|A_n - \mu| \geq 4.5\sigma/\sqrt{n}] \leq 5\%$, which shows that $[A_n - 4.5\sigma/\sqrt{n}, A_n + 4.5\sigma/\sqrt{n}]$ is a 95%-confidence interval for μ. Using the *Central Limit Theorem*, we will be able to replace 4.5 by 2 in the confidence interval (see Section 9).

Examples

1. Let $X = B(p)$. Then $var(X) = E[X^2] - E[X]^2 = E[X] - E[X]^2 = p(1-p) \leq 1/4$.
2. Let $X = B(n,p)$. Then we can write $X = X_1 + \cdots + X_n$ where the X_m are i.i.d. $B(p)$, so that $var(X) = np(1-p)$.
3. Let Ω be the uniform probability space $\{1, \ldots, 6\}$ and let X takes the values $\{0,0,1,1,2,2\}$ and Y the values $\{0,3,6,12,3,0\}$, respectively for the different values of ω. Thus, $X(1) = X(2) = 0$ and $Y(1) = 0, Y(1) = 3$, etc. Then $E[X] = (0+0+1+1+2+2)/6 = 1, E[Y] = (0+3+6+12+3+0)/6 = 4, E[XY] = (0+0+6+12+6+0)/6 = 4$. Hence, $cov(X,Y) = 0$ and $L[Y|X] = E[Y] = 4$. Also, $Pr[Y = 0|X = 0] = 1/2$ and $Pr[Y = 3|X = 0] = 1/2$, so that $E[Y|X = 0] = 3/2$. Similarly, $E[Y|X = 1] = 9$ and $E[Y|X = 2] = 3/2$. We can write $E[Y|X] = (3/4)(X-1)(X-2) - 9(X-0)(X-2) + (3/4)(X-0)(X-1) = -(15/2)X^2 + 15X + 3/2$ since a polynomial that goes through the point (x_m, y_m) for $m = 1, \ldots, n$ can be written as $\sum_{m=1}^{n} y_m \Pi_{k \neq m}[(x - x_k)/(x_m - x_k)]$.
4. Let X, Y, Z be i.i.d. with mean 0 and variance 1. Then $L[aX + bY + cZ | dX + eY + fZ] = [(ad + be + cf)/(d^2 + e^2 + f^2)](dX + eY + fZ)$.
5. Let $\Omega = \{1,2,3,4\}$ be a uniform probability space. Let also $A = \{1,2\}, B = \{1,3\}, C = \{1,4\}$ and $X = 1_A, Y = 1_B, Z = 1_C$. Then, X, Y, Z are pairwise independent, but not mutually independent. They are also identically distributed like $B(0.5)$. Note that $E[XYZ] = 1/4 \neq E[X]E[Y]E[Z]$. Thus, when we write 'i.i.d.', we mean *mutually independent* and identically distributed, like three coin flips, for instance.

6. Let $\{X_1,\ldots,X_n\}$ be pairwise independent. Then $var(X_1+\cdots+X_n)=var(X_1)+\cdots+var(X_n)$.

7. Let $X=B(n,p)$. Then $X=X_1+\cdots+X_n$ where the X_m are i.i.d. $B(p)$. Hence, $var(X)=n.var(X_1)=np(1-p)$.

8. Let $X=P(\lambda)$. Then $E[X(X-1)]=\sum_{n\geq 0}n(n-1)\lambda^n\exp\{-\lambda\}/(n!)=\lambda^2\sum_{n\geq 2}\lambda^{n-2}\exp\{-\lambda\}/[(n-2)!]=\lambda^2\sum_{n\geq 0}\lambda^n\exp\{-\lambda\}/(n!)=\lambda^2$. Thus, $E[X^2]-E[X]=\lambda^2$, so that $E[X^2]=\lambda^2+E[X]=\lambda^2+\lambda$ and $var(X)=E[X^2]-E[X]^2=\lambda$.

9. Let $X=G(p)$. Then $X=1+ZY$ where $Y=G(p)$ and $Z=B(1-p)$ are independent. Thus, $E[X^2]=E[1+2ZY+Z^2Y^2]=1+2(1-p)E[Y]+(1-p)E[Y^2]=1+2(1-p)/p+(1-p)E[X^2]$. Hence, $E[X^2]=[1+2(1-p)/p]/p=(2-p)/p^2$. Consequently, $var(X)=E[X^2]-E[X]^2=(2-p)/p^2-1/p^2=(1-p)/p^2$.

6 Collisions and Collecting

Review

Say that there are M different coupons. When you buy a box of cereal, you get coupon m with probability $1/M$ for $m=1,\ldots,M$. It takes $X_1=1$ box to get the first coupon. Then $X_2=G((M-1)/M)$ to get a new one, then $X_3=G((M-2)/M)$ to get a new one, and so on. Thus, the average number of boxes required to get all the coupons is $M/M+M/(M-1)+M/(M-2)+\cdots+M=MH(M)$ where $H(M):=1+1/2+\cdots+1/M\approx ln(M)+0.58$. The probability that coupon m has not been seen in n boxes is $[(M-1)/M]^n=(1-1/M)^n\approx\exp\{-n/M\}$. Thus, the average number of coupons seen after n steps is $M[1-(1-1/M)^n]\approx M(1-\exp\{-n/M\})$.

One throws m balls into $n>m$ bins, independently and uniformly at random. The probability that there is no collision after 1 ball is 1, after 2 balls is $(n-1)/n$, after three balls it is $[(n-1)/n]\times[(n-2)/n]$, after four balls it is $[(n-1)/n]\times[(n-2)/n]\times[(n-3)/n]$, and so on. The probability of no collision after m balls is then $\Pi_{k=1}^{m-1}[(n-k)/n]\approx\Pi_{k=1}^{m-1}\exp\{-k/n\}=\exp\{-\sum_{k=1}^{m-1}k/n\}\approx\exp\{-m^2/(2n)\}$. Bin i is still empty after k steps with probability $(1-1/n)^k$. Thus, the expected number of empty bins is $m(1-1/n)^k\approx m\exp\{-k/n\}$. By Markov's inequality, the probability that there is at least one empty bin is at most $m\exp\{-k/n\}$.

Examples

1. Let's use c bits as a CRC for m files. This means that the m files are thrown into $n=2^c$ bins. The probability of no collision is approximately $\exp\{-m^2/(2n)\}$, so that the probability of collision is approximately $1-\exp\{-m^2/(2n)\}\approx m^2/(2n)$. If we want the probability to be less than ε, we need $m^2/(2n)\leq\varepsilon$, i.e., $m^2/(2^{c+1})\leq\varepsilon$, or $c\geq 2\log_2(m)-log_2(\varepsilon)-1$. For instance, if $m=10^6\approx 2^{20}$ and $\varepsilon=10^{-9}$, we need $c\geq 69$. The implication is that a 7-byte CRC suffices to sign messages or to detect errors in most applications.

7 Markov Chains

Review

A *Markov chain* on the finite state space $\mathscr{X}=\{1,2,\ldots,K\}$ is the random sequence $\{X_n,n\geq 0\}$ defined by

$$Pr[X_0=i_0,X_1=i_1,\ldots,X_n=i_n]=\pi_0(i_0)P(i_0,i_1)\cdots P(i_{n-1},i_n)$$

where π_0 is a probability distribution called the *initial distribution* of the Markov chain and P is a $K \times K$ nonnegative matrix whose rows sum to one. In particular, $\pi_n(i) := Pr[X_n = i]$ is such that $\pi_n = \pi_0 P^n$ and $Pr[X_n = j | X_0 = i] = P^n(i, j)$.

A MC is *irreducible* if it can go from any state i to any other state j, possibly in more than one step. It is then *aperiodic* if the return times to some state are coprime. (The return times to any state are then also coprime.)

The distribution π is *invariant* if π solves the *balance equations* $\pi P = \pi$. There is a unique invariant distribution if the Markov chain is irreducible. Moreover, the fraction of time in state i then converges to $\pi(i)$ for all i. The distribution π_n converges to π if the Markov chain is also aperiodic.

The average time $\beta(i)$ to enter a set A of states when starting from state i satisfies the *first step equations*

$$\begin{aligned}\beta(i) &= 1 + \sum_j P(i,j)\beta(j), \forall i \notin A \\ \beta(i) &= 0, \forall i \in A.\end{aligned}$$

The probability $\alpha(i)$ of entering a set A of states before a disjoint set B of states when staring from state i satisfies the *first step equations*

$$\begin{aligned}\alpha(i) &= \sum_j P(i,j)\alpha(j), \forall i \notin A \cup B \\ \alpha(i) &= 1, \forall i \in A \\ \alpha(i) &= 0, \forall i \in B.\end{aligned}$$

Note that we call the equations both for α and β first step equations even though they are not the same. The justification for the terminology is that in both cases one conditions on what happens in the first step of the MC.

Examples

1. The MC on $\{0,1\}$ with $P(0,1) = P(1,0) = a$ and $P(0,0) = P(1,1) = 1-a$ with $a \in [0,1]$ is irreducible if $a > 0$ and aperiodic if $0 < a < 1$. The invariant distribution is $\pi = [0.5, 0.5]$ when $a > 0$.

2. The MC on $\{0,1\}$ with $P(0,1) = a$ and $P(1,0) = b$ with $a, b \in [0,1]$ is irreducible if $a, b > 0$ and is then aperiodic if $a \neq 1$ or $b \neq 1$. The invariant distribution is $\pi = [b/(a+b), a/(a+b)]$ when $a, b > 0$.

3. You flip a fair coin until you get two heads in a row. The average number of flips is 6.

4. You roll a balanced six-sided die until the sum of the last two rolls is equal to 8. The average number of rolls is about 8.4.

5. In each step, you get up one rung of a ladder with probability p; otherwise, you fall back to the ground. The average number of step to reach rung 20 is $(p^{-20} - 1)/(1-p)$.

6. At each step, you win one dollar with probability p; otherwise you lose it. Starting with n dollars, the probability your fortune reaches $M > n$ before it reaches 0 is $(1-\rho^n)/(1-\rho^M)$ where $\rho = (1-p)p^{-1}$.

8 Continuous Probability

Review

Imagine choosing a real number X uniformly in $[0,1]$. Clearly $Pr[X=x]=0$ for all $x\in[0,1]$. This would be the same if we chose the number uniformly in $[0,2]$. Thus, the probability of individual outcomes does not describe properly the random experiment. Instead, one starts by defining the probability of events. For choosing uniformly in $[0,1]$, one defines $Pr[[a,b]]=b-a$, for $0\le a\le b\le 1$. One extents that definition by additivity to any (countable) union of intervals. For instance, $Pr[[0,0.3]\cup[0.7,0.9]]=0.5$. Thus, for a finite or countable sample space Ω, one starts by defining the probability of each outcome ω and one then defines the probability of an event as the sum of the probabilities of the outcomes it contains. For an uncountable sample space Ω, one starts by defining the probability of its events.

Let $f:\Re\to[0,\infty)$ be a function such that $\int_{-\infty}^{\infty}f(x)dx=1$. Define a random variable X such that $Pr[x<X<x+\varepsilon]\approx f(x)\varepsilon$ for $\varepsilon\ll 1$. Then $F(x):=Pr[X\le x]=\int_{-\infty}^{x}f(y)dy$. Thus, $f(x)$ is the derivative of $F(x)$. One calls $f(x)$ the *probability density function* (pdf) of X and $F(x)$ the *cumulative distribution function* (cdf) of X. Sometimes one writes $f_X(x):=f(x)$ and $F_X(x):=F(x)$ to specify that the pdf and cdf are those of the random variable X. This is convenient when one deals with more than one random variable. Then $E[X]=\int xf_X(x)dx$ and $E[h(X)]=\int h(x)f_X(x)dx$.

Examples

1. $X=U[a,b]$ iff $f_X(x)=(b-a)^{-1}1\{a<x<b\}$. Consequently, $E[X]=\int_a^b x(b-a)^{-1}dx=(a+b)/2$.
2. $X=Expo(\lambda)$ iff $f_X(x)=\lambda\exp\{-\lambda x\}1\{x\ge 0\}$. Consequently, $E[X]=\int_0^\infty x\lambda\exp\{-\lambda x\}dx=-\int_0^\infty d\exp\{-\lambda x\}dx$ λ^{-1}.
3. One shoots a dart uniformly in a circle with radius r. Let X be the distance of the dart to the center. Then $F_X(x)=(\pi x^2)/(\pi r^2)=x^2/r^2$ for $0\le x\le r$. Hence, $f_X(x)=2x/r^2 1\{0\le x\le r\}$ and $E[X]=2r/3$.
4. Define $Y=a+bX$, for some a and some $b>0$. Then $f_Y(y)\varepsilon=Pr[y<a+bX<y+\varepsilon]=Pr[\frac{y-a}{b}<X<\frac{y-a}{b}+\frac{\varepsilon}{b}]=f_X(\frac{y-a}{b})\frac{\varepsilon}{b}$. Thus, $f_Y(y)=\frac{1}{b}f_X(\frac{y-a}{b})$.
5. Let X and Y be two independent random variables and $W=\max\{X,Y\}$. Then $F_W(w)=Pr[W\le w]=Pr[X\le w]Pr[Y\le w]=F_X(w)F_Y(w)$.
6. Let X and Y be two independent random variables and $V=\min\{X,Y\}$. Then $1-F_V(v)=Pr[V>v]=Pr[X>v]Pr[Y>v]=(1-F_X(v))[1-F_Y(v)]$.
7. Let X and Y be two independent random variables and $Z=X+Y$. Then, $f_Z(z)\varepsilon=Pr[z<Z<z+\varepsilon]=\int_{-\infty}^{\infty}Pr[x<X<x+dx,z-x<Y<z-x+\varepsilon]=\int_{-\infty}^{\infty}f_X(x)f_Y(z-x)\varepsilon dx$. Hence, $f_Z(z)=\int_{-\infty}^{\infty}f_X(x)f_Y(z-x)dx$.
8. Assume that with probability p the random variable X has pdf $f_0(x)$ and it has pdf $f_1(x)$ otherwise. Given $X=x$, the probability that it has pdf f_0 is $[pf_0(x)]/[pf_0(x)+(1-p)f_1(x)]$.

9 Gaussian and CLT

Review

A random variable X is $\mathcal{N}(0,1)$ iff its pdf is $f_X(x) = (1/\sqrt{2\pi})\exp\{-x^2/2\}$. One can verify that $E[X] = 0$ and $var(X) = 1$. By definition, the random variable $Y = \mu + \sigma X$ is then $\mathcal{N}(\mu, \sigma^2)$. Then $E[Y] = \mu + \sigma E[X] = \mu$ and $var(Y) = \sigma^2 var(X) = \sigma^2$. Also, one can check that $f_Y(y) = (1/\sqrt{2\pi\sigma^2})\exp\{-(y-\mu)^2/(2\sigma^2)\}$ by Example 4 of the previous section. This is the *bell-shaped* pdf.

If $X = \mathcal{N}(\mu, \sigma^2)$, then $Pr[X > \mu + 2\sigma] \approx 2.5\%$ and $Pr[|X - \mu| > 2\sigma] \approx 5\%$. Also, $Pr[X > \mu + 1.65\sigma] \approx 5\%$ and $Pr[|X - \mu| > 1.65\sigma] \approx 10\%$. Note that Chebyshev shows that $Pr[|X - \mu| > 2\sigma] \leq 1/4 = 25\%$, which is a very loose bound. This is why the CLT (see below) provides smaller confidence intervals than Chebyshev. To get 5% using Chebyshev, we have to write $Pr[|X - \mu| \geq 4.5\sigma] \leq 1/(4.5)^2 \approx 5\%$.

The *Central Limit Theorem* (CLT) states that if the X_n are i.i.d. with mean μ and variance σ^2 and if $A_n = (X_1 + \cdots + X_n)/n$, then $[A_n - \mu]\sqrt{n} \approx \mathcal{N}(0, \sigma^2)$ when $n \gg 1$. Thus, we see that $Pr[|A_n - \mu|\sqrt{n} > 2\sigma] \approx 5\%$, so that $Pr[A_n - 2\sigma/\sqrt{n} \leq \mu \leq A_n + 2\sigma/\sqrt{n}] \approx 95\%$. Hence, $[A_n - 2\sigma/\sqrt{n}, A_n + 2\sigma/\sqrt{n}]$ is a 95%-confidence interval for μ.

One can show that if $X = \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y = \mathcal{N}(\mu_2, \sigma_2^2)$ are independent, then $X + Y = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$. Intuitively, X and Y are both sums of many small independent random variables, so that $X + Y$ is also, and should therefore be Gaussian. The mean and variance are the sum of those of X and Y.

The Law of Large Numbers implies that if the X_m are i,i.d. with mean μ and variance σ^2, then $A_n = (1/n)\sum_{m=1}^n X_m \approx \mu$ and $s_n^2 := (1/n)\sum_{m=1}^n (X_m - A_n)^2 \approx \sigma^2$. Thus, replacing the standard deviation σ by s_n in the confidence intervals, we see that $[A_n - 2s_n/\sqrt{n}, A_n + 2s_n/\sqrt{n}]$ is a 95%-confidence interval for μ when n is large enough. In practice, one has to be a bit careful because s_n may be a poor estimate of σ when n is not large enough. Use with care and at your own risk!

Examples

1. Let $X = X_1 + \cdots + X_n$ where the X_m are i.i.d. $B(p)$. Let also $A_n = (X_1 + \cdots + X_n)/n$. We saw that $[A_n - 2\sigma/\sqrt{n}, A_n + 2\sigma/\sqrt{n}]$ is a 95%-confidence interval for $E[X_1] = p$. Since $\sigma^2 = var(X_1) = p(1-p) \leq 1/4$, one has $\sigma \leq 1/2$, and it follows that $[A_n - 1/\sqrt{n}, A_n + 1/\sqrt{n}]$ is a 95%-confidence interval for p.

10 Some Important Distributions

1. **Bernoulli with parameter** $p : B(p)$

   $Pr[X = 1] = p; Pr[X = 0] = 1 - p$. Mean $= p$; variance $= p(1-p)$.

2. **Uniform in** $\{1, 2, \ldots, n\} : U[1, \ldots, n]$

   $Pr[X = m] = 1/n, m = 1, \ldots, n$. Mean $= (n+1)/2$; variance $= (n^2 - 1)/12$.

3. **Binomial with parameters** $n, p : B(n, p)$

   $Pr[X = m] = \binom{n}{m} p^m (1-p)^{n-m}, m = 0, \ldots, n$. Mean $= np$; variance $= np(1-p)$.

4. **Geometric with parameter** $p : G(p)$

   $Pr[X = n] = (1-p)^{n-1} p, n = 1, 2, \ldots$. Mean $= 1/p$, variance = $(1-p)/p^2$.

5. **Poisson with parameter** $\lambda : P(\lambda)$

   $Pr[X = n] = (\lambda^n/n!)e^{-\lambda}, n = 0, 1, 2, \ldots$. Mean = λ, variance $= \lambda$.

6. **Uniform in** $[0,1] : U[0,1]$**.**

   $f_X(x) = 1\{0 \le x \le 1\}$. Mean $= 1/2$, variance $= 1/12$.

7. **Exponential with parameter** $\lambda : Expo(\lambda)$.

   $f_X(x) = \lambda e^{-\lambda x} 1\{x > 0\}, F_X(x) = [1 - e^{-\lambda x}] 1\{x \ge 0\}$. Mean $= \lambda^{-1}$, variance $= \lambda^{-2}$.

8. **Gaussian with parameters** $\mu, \sigma^2 : \mathcal{N}(\mu, \sigma^2)$

   $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(x-\mu)^2/(2\sigma^2)\}$. Mean $= \mu$, variance $= \sigma^2$.

11 Appendix: Some Mathematical Facts

The following definitions and facts are used repeatedly in this course.

1. The following set notation is assumed to be familiar: $\emptyset, A \cap B, A \cup B, A \Delta B, A \setminus B, \bar{A} = A^c$.
2. $(A \cap B)^c = A^c \cup B^c$ and $(A \cup B)^c = A^c \cap B^c$.
3. $A \times B := \{(a,b) \mid a \in A, b \in B\}$.
4. $A^2 := A \times A$; $A^n := A^{n-1} \times A = \{(a_1, \ldots, a_n) \mid a_k \in A, k = 1, \ldots, n\}$.
5. A^* is the set of finite strings with elements in A. Thus $A^* := \cup_{n=0}^{\infty} A^n$ where A^0 is the set that contains one string of length 0.
6. For $a \ne 1$ and $n \ge 0$, one has $1 + a + \cdots + a^n = (1 - a^{n+1})/(1-a)$.
7. For $|a| < 1$, one has $1 + a + a^2 + \cdots = 1/(1-a)$.
8. $1 + 2a + 3a^2 + 4a^3 + \cdots = (d/da)[1 + a + a^2 + \cdots] = (d/da)(1-a)^{-1} = (1-a)^{-2}$.
9. For $0 \le m \le n$, one defines $\binom{n}{m} := \frac{n!}{m!(n-m)!}$.
10. For $n \ge 0$, one has $(a+b)^n = \sum_{m=0}^{n} \binom{n}{m} a^m b^{n-m}$.
11. $\ln(1+\varepsilon) \approx \varepsilon$ for $|\varepsilon| \ll 1$.
12. $\exp\{a\} = \sum_{n=0}^{\infty} \frac{a^n}{n!}$.
13. $\exp\{\varepsilon\} \approx 1 + \varepsilon$ for $|\varepsilon| \ll 1$.
14. $\exp\{a+b\} = \exp\{a\}\exp\{b\}$.
15. $\log_b(x) = \log_a(x) \log_b(a)$.
16. $1 + 2 + 3 + \cdots + n = n(n+1)/2$.
17. $\int_a^b f(x) dg(x) = f(b)g(b) - f(a)g(a) - \int_a^b g(x) df(x)$.
18. $\int_0^{\infty} \exp\{-ax\} dx = 1/a$ for $a > 0$.
19. $\int_{-\infty}^{\infty} \exp\{-\frac{x^2}{2}\} dx = \sqrt{2\pi}$.
20. $\int_{-\infty}^{\infty} x^2 \exp\{-\frac{x^2}{2}\} dx = \sqrt{2\pi}$.