

7. Theory of Probability

7.1 Introduction to Probability

One commonality found in the problems we've examined thus far is that each problem involves a large number of possibly true situations and we are interested in narrowing down the set to one definitely true situation. In the weighing problem for example, at the beginning any one of the objects might be the "bad" one (heavy or light) and we seek to gather enough information to determine that one is definitely bad, and the others are definitely not.

Probability theory is studied to provide the groundwork for modeling such problems and provide information about what to expect in various situations. Interestingly enough, it was originally developed by Blaise Pascal and Pierre de Fermat to study gambling schemes.

We can define a set of all the possible outcomes in a problem, called a **sample space**, in which each outcome has a **probability** equal to the proportion of cases it is expected to be true. In the weighing problem, any of the objects could be the bad one and there is an outcome in the sample space for each such scenario. The probability of each is equal to 1 divided by the number of objects, assuming they all have equal likelihood of being bad.

We can determine our expected outcome to some situation by representing the outcome as some function of the sample space points, and averaging the function over the entire sample space. This is often much simpler than it sounds, as the function referred to frequently outputs the point itself, as long as it can be represented by a number and can be averaged. For example, we could determine someone's expected bowling score by summing from 0 to 300 $i \cdot p(i)$, where $p(i)$ is the probability that he will score i . We are simply averaging $f(i) = i$ over the sample space. More on this in the next section.

Commonly we can represent the potential situations by strings of binary digits (bits). In the remainder of these notes we assume that the sample space consists of points that are represented as such.

7.2 Definitions and Concepts

Event

An event is any subset of the sample space. For example, having all 1s in the string is an event, as is having the last bit be 0. Any other property of the strings in the sample space corresponds to an event. The subset for event E consists of the points in the sample space having the property that defines E .

Random Variable

A random variable is a function defined on the points of the sample space. For example, the function could be the number of 1s in the string, the position of the first 1, how much one could win in a betting scheme if the given bit string were

true, or any other (usually numeric) function on the strings. An **indicator** random variable for an event E is a special kind of random variable that is either 1 or 0. It is 1 if the string is in the event subset and 0 otherwise. We shall call the indicator variable I(E).

Given any random variable f, let E(f=x) be the event that f(p) = x, where p is a bit string in the sample space. Thus, I(E(f=x)) is 1 iff f(p) = x. Therefore, we can write:

$$\sum_{all\ x} x[I(E(f = x))]$$

Probability

Every point q in the sample space is true a certain proportion of the time, p(q), that we call the probability. Clearly, these probabilities cannot be 0. Also, the probabilities must sum 1 over the entire sample space. This is true since the sample space includes all possibilities, and thus the proportion of time that one of outcomes in the sample space occurs is 1.

Uniform Sample Space

This is a sample space in which all the outcomes have the same probability. Since they must sum to 1, the probability must be the reciprocal of the number of points in the sample space. In a uniform sample space of N outcomes, the probability of each outcome is $1/N$.

Event Probability

The probability of event E, p(E), is the sum of p(q) over all the strings q in the subset corresponding to E.

Expected Value

The expected value of a random variable f, denoted as \bar{f} or $\langle f \rangle$, is the sum over all points q in the sample space of p(q)*f(q). This is the same as the mean or average of the variable. Notice that the expectation of an indicator variable for an event is the probability of that event.

Expectation is a linear property of random variables. That is, for random variables f and g, with constants a and b:

Linearity of Random Variables

$$\langle af + bg \rangle = a\langle f \rangle + b\langle g \rangle$$

Deviation

The deviation of f from its mean value is f - $\langle f \rangle$.

Variance

The variance of a random variable is the mean square deviation of f , or $\langle (f - \langle f \rangle)^2 \rangle$. This simplifies to $\langle f^2 \rangle - \langle f \rangle^2$ (proof deferred to exercises), or the difference between the expectation of its square and the square of its expectation. Variance is denoted as $\sigma^2(f)$. Thus, $\sigma(f)$ is the root mean square deviation of f .

The variance of a function f is quadratic in f , thus $\sigma^2(af) = a^2\sigma^2(f)$ (prove in exercises).

Conditional Probability

Recall that an event E corresponds to a subset of the sample space. The conditional probability of q given E , written $p(q|E)$, is the probability that q is true, given the reduced sample space of event E . That is, given that we are restricted to E 's subset of points, the probability that q occurs is the conditional probability.

We can write the probability of q as the sum of $p(s)*I(s)$ for all points s in the sample space divided by the sum of all $p(s)$ in the sample space, where I is the indicator variable for q occurring. In a full sample space, the sum of all $p(s)$ is 1 and the sum of all $p(s)*I(s)$ is $p(q)$.

In our reduced sample space, the sum of all $p(s)$ is simply $p(E)$, since the sample space is reduced to the subset for event E . Also, the indicator variable for q given E is the product of the indicator variables for q and E , since the points we are interested in are those for which q is satisfied and E is satisfied. When they are both satisfied, I_q*I_E is 1, otherwise, it equals 0. This is appropriate behavior for our indicator variable. Thus, the $p(q|E)$ is the sum of all $p(s)*I_q(s)*I_E(s)$ divided by the sum of $I_E(s)*p(s)$ for all s . This is the same as saying $P(q \text{ and } E)$ (oft written as $P(q \cap E)$), divided by $p(E)$.

To summarize this:

$p(q E) = p(E \cap q) / p(E)$

Of course, by the same formula, we know that $p(E|q) = p(q \cap E) / p(Q)$. Since $p(q \cap E)$ is the same as $p(E \cap q)$, we can solve for this quantity, and substitute it into the boxed equation above to get $p(q|E)$ in terms of $p(E|q)$. Specifically, $p(q|E) = p(E|q)*p(q)/p(E)$. This tells us that $p(q|E)$ differs from $p(q)$ by a factor of $p(E|q)/p(E)$.

Mutually Exclusive and $p(A \cap B)$

Two events are mutually exclusive if they never occur at the same time. A simple example is the event that the bit string has more 1s than 0s, and the event that the bit string has more 0s than 1s. The probability that either of the two mutually exclusive events occurs is simply the sum of the probabilities of the two events.

For two events that can occur at the same time, notice that adding the probabilities of the two events “double counts” certain situations. Suppose our sample space contains all 2-bit strings (00 01 10 11). If one event is that the string starts with 1, and the other is that the string has at least one 1 in it, the bit string 10 is counted in both events. Thus, to get the probability that either of the two events occurs, we sum the probabilities and then subtract out the probability that both occur, so that nothing is counted twice.

<p>For all events A,B</p> $P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$
--

Similarly for three events, to find the probability that one or more occurs, we sum up the individual probabilities and subtract out the probability that any combination of two occurs. That is, we subtract out $P(A \cap B)$, $P(A \cap C)$ and $P(B \cap C)$. However, for any scenario that satisfies all three events, we have added it in three times (adding the individual probabilities) and subtracted it out three times leaving nothing, so we must add in again. Thus, we add $P(A \cap B \cap C)$.

<p>For events A_1, A_2, A_3</p> $P(A_1 \text{ or } A_2 \text{ or } A_3) = \sum_{\text{all } j} P(A_j) - \sum_{j < k} P(A_j \cap A_k) + P(A_1 \cap A_2 \cap A_3)$

You will have to find a similar statement for four events in the exercises.

Independence

Two events are said to be independent if the probability of A given B is the same as the probability of A alone. That is, two events are independent if knowing that one occurs does not affect the probability of the other occurring. Knowing the formula for A given B, and equating it to the probability of A, it is easy to see that independence implies that the probability of A and B is equal to the product of their probabilities.

<p>For independent events A and B</p> $P(A \cap B) = P(A)P(B)$
--

Random variables said to be independent if knowing the value of one does not affect the probabilities of the possible values of the other. For two variables f and g, we know $p(f=a) = p(f=a|g=b) = p(f=a \ \& \ g=b)/p(g=b)$. Thus, $p(f=a \ \& \ g=b) = p(f=a)*p(g=b)$. Multiplying both sides of this by $a*b$ and summing over all possible a and b, we obtain a relation for the expectation of the product of f and g.

<p>For independent random variables</p> $\langle f*g \rangle = \langle f \rangle * \langle g \rangle$

It follows that the indicator variable relation $I(f=a, g=b) = I(f=a) * I(g=b)$ holds as well. It should be obvious that the expectation of the sum of two random variables is the sum of their expectations.

Covariance

Now let's examine the variance of the sum of two random variables. We start knowing that variance in this case is $\langle (f+g)^2 \rangle - \langle f+g \rangle^2$.

$$\begin{aligned} \text{Var}(f+g) &= \langle (f+g)^2 \rangle - \langle f+g \rangle^2 \\ &= \langle f^2 + 2f * g + g^2 \rangle - \langle f \rangle^2 - 2\langle f \rangle \langle g \rangle - \langle g \rangle^2 \\ &= \langle f^2 \rangle - \langle f \rangle^2 + \langle g^2 \rangle - \langle g \rangle^2 + 2(\langle f * g \rangle - \langle f \rangle \langle g \rangle) \\ &= \text{Var}(f) + \text{Var}(g) + 2(\langle f * g \rangle - \langle f \rangle \langle g \rangle) \end{aligned}$$

We can define the quantity $(\langle f * g \rangle - \langle f \rangle \langle g \rangle)$ as the covariance of two random variables. Thus:

$$\text{Var}(f+g) = \text{Var}(f) + \text{Var}(g) + 2\text{Cov}(f, g)$$

As we determined above, the $\langle f * g \rangle = \langle f \rangle \langle g \rangle$ for independent variables, and thus the covariance is 0.

$$\begin{aligned} &\text{For independent random variables} \\ &\text{Var}(f+g) = \text{Var}(f) + \text{Var}(g) \end{aligned}$$

This innocent looking statement is actually very curious and important. We have here a linear type property of a quadratic function of f and g . It is the basis of the results we will need, which concern sums and averages of many independent random variables.

7.3 Comments and Clarifications

At first, it may seem as though we have made a lot of definitions but gotten essentially nowhere. However, now that we have these concepts fleshed out, we can examine some important results.

One thing to note before moving on is that the field of probability contains a lot of loaded words that do not necessarily mean what they sound like. The expectation of a random variable, for example, does not necessarily mean that we expect a certain value for that variable. Rather, the expectation of a variable is the average over its sample space, weighted according to probability. In the case of indicator variables, the expectation is equal to the probability of the event (between 0 and 1), even though the variable can only be 0 or 1. Certainly we do not actually expect that the value of an indicator variable is an intermediate number between 0 and 1.

However, if the variance is very small, we actually do expect that the value of a random variable will be near its expectation. Thus, to draw accurate conclusions from

probability theory, we need to be able to learn something about the expectation of f , based on its variance.

Let's take a look at situations that can be modeled as long sequences of independent random variables. Specifically we will look at variables that can only take on 0 or 1 as their value. We call these **Boolean random variables**. Several Boolean variables with the same probability of being 1 are said to be **identically distributed**. This is because, if we continuously assign and reassign values to two such variables based on their probability of being 1, the distribution of 1s and 0s for the two variables should be the same. If we are concerned with a sequence of length N of independent and identically distributed Boolean random variables then we can make very strong statements about their sum and their average. We can determine the probability that their sum or average takes on a given value.

As we relax the conditions here, so that the variables do not have to all be identical, can take on other values, and can have limited dependence on one another, we can say somewhat less, but for a wide variety of circumstances there is a similar conclusion, called the **central limit theorem**. It has many variations based on varying assumptions about the situation modeled, usually with roughly the same conclusion. We will not go into this much here.

There is a weaker statement called the **law of large numbers**, which is sufficient for our purposes here. It states that if you choose a value from a set of independent identically distributed random variables many times, the proportion of the time that you get any particular value will be close to the probability that your random variable takes on that value, as your number of picks increases. There are actually several laws of large numbers on this topic.

This justifies the interpretation of the probability of an event as the proportion of the time the event would occur if you kept putting yourself in the situation modeled by your sample space over and over again, independently.

The rest of the notes consists of a description of our last tool and how variance is involved, and then some of the consequences that can be deduced from all this about sequences of random variables.

7.4 Chebyshev's Inequality

As mentioned, the definition of the variance of a random variable f is the mean square of f 's deviation from its mean. We can use this concept of variance to put an upper limit on the probability the variable takes on values with deviations from the mean that are greater than $x\sigma$, where σ^2 is variance, and x is any value greater than 1.

One such bound is Chebyshev's inequality, which answers the question: What is the maximum proportion of a variable's sample space that can correspond to a deviation greater than $x\sigma$?

Let's call that proportion of the sample space z . That is, the product of z and the number of elements in the sample space is the number that deviate from the mean by at least $x\sigma$. The minimum variance in this scenario would result when all of the points with deviation greater than $x\sigma$ had deviance that was only infinitesimally greater and all the points with deviation less than $x\sigma$ have deviance of 0. The square deviation of the samples with deviation $> x\sigma$ would be $(x\sigma)^2$, while the square deviation of the other samples would be 0. Since a proportion of z of the samples have the non-zero deviation, then the mean square deviation (variance) is equal to $z(x\sigma)^2$. This of course is at most σ^2 , since the optimal possible variance must be less than the actual variance of the variable. Thus, $z \leq \frac{1}{x^2}$. Recall that z is the proportion of the sample space with deviation greater than $x\sigma$. Thus Chebyshev's Inequality:

$$P(|F - f| \geq k) \leq \sigma^2/k^2, \text{ where } k = x\sigma \text{ and } F \text{ is a value taken on by } f.$$

This implies that when σ is very small, the portion of the sample space over which the value of the random variable is close to the mean is very large. This is a law of large numbers, as σ goes to 0.

7.5 Conclusions

Suppose that we have N independent Boolean variables (having values 0 or 1), having a probability p of being 1. When k of the variables have value 1, the sum will be k . In this case, $N-k$ have 0 as their value.

For each of the many ways of choosing k of the N variables to be 1 and $N-k$ to be 0, the probability of the specific arrangement is $p^k(1-p)^{N-k}$ since the variables are independent. Using basic combinatorics, there are $C(N,k)$ mutually exclusive ways of this occurring. Thus the probability that the sum is k is:

$$p^k(1-p)^{N-k}C(N,k)$$

The distribution of sums of the variables or of the when each is assigned a 1 or 0 based on a fixed p and large N is called a **binomial distribution**. It forms a bell curve shape with each different sum along the x -axis and its corresponding probability (calculated above) on the y -axis.

We know that the mean of the distribution, or the expected sum, is the expected value of each variable summed together. This is simply pN , since there are N variables, each with expected value of p . The mean of this average then is p itself. The variance of each variable is $p(1-p)$ (prove in exercises) and thus the variance of the sum is $Np(1-p)$. The variance of their average is $p(1-p)/N$, since variance is a quadratic property and thus we must divide by N^2 when averaging it.

By the Chebyshev inequality, we now know that the average of the N variables approaches p as long as p does not approach 0. This is the **Weak Law of Large Numbers**.

This statement provides a meaning of sorts, to the probability p of a single variable. It is also exactly the statement we will need in section 8 of the course notes.

Similar results to all of those above hold when assumptions are weakened. Here all variables are assumed to be independent, and identically distributed.

Thus, as long as the variables are independent and each of their variances are small compared to the sum of the variances, it is possible to prove that the resulting probability for the average of the variables, is not unlike a binomial distribution, except it will have mean given by the average of the p 's and variance given by the sum of the individual variances divided by N^2 .

This holds true for sums of random variables which take on more values than just 0 and 1, and even (with some changes) when there is a limited amount of dependence among the variables, as well.

Results of this kind are called central limit theorems.

When we add a sum of a large number of independent and identically distributed random variables, and the sum of their means stays finite, we get a special case of a binomial distribution, called a Poisson distribution.

When the mean of the Poisson distribution approaches λ , the distribution has the form:

$$p(k) = e^{-\lambda} \lambda^k / k!.$$

Again, distributions can tend to this one with much weakened assumptions about them.

You should also realize that the facts about probability that we have noticed above, that the mean of a sum is the sum of the means in all cases, that the variance of the sum of independent variables is the sum of the variances, etc. are extremely useful tools for solving probabilistic questions, as you will see in the exercises below.

Exercises

- Exercise 1* Prove that the variance of a random variable $\langle (f - \langle f \rangle)^2 \rangle$ is equal to $\langle f^2 \rangle - \langle f \rangle^2$.
- Exercise 2* Prove that the variance of a random variable f is quadratic. That is, prove that $\sigma^2(af) = a^2 \sigma^2(f)$.
- Exercise 3* What is the general statement of the probability that one or more of four events will occur?

Exercise 4 Prove that the variance of a Boolean variable, having probability of p that it has a value of 1, is $p(1-p)$.

Exercise 5 Use a spreadsheet to plot (chart x y scatter) $p(k/N)$ vs k for binomial distributions having $p=1/2$ and $p=1/3$ for values of N from 5 to however high you can, on one sheet for each p .

Exercise 6 Derive the formula for a Poisson distribution from the binomial distribution with $p=m/N$, for finite k as N gets large.

Exercise 7 Given a class of N students with independently chosen birthdays, deduce the probability that no two have the same birthday. This can be done by deducing a formula for it and taking logs and expanding them to approximate the answer.

Another way to estimate the answer is by assuming you have a Poisson distribution of the number of pairs of students with the same birthday. How do these answers compare?

Exercise 8 a. Suppose you pick 2 numbers between 0 and 1 at random (that is, the probability of picking a number in any interval of length d is $1/d$) for each). What is the expected value of their difference? Hint: what is the answer if you pick 3 points at random on a unit circle?

b. Suppose you pick 2 distinct numbers at random between 1 and N . What is the expected value of their difference? Hint: compare with part a.

c. Suppose you pick k distinct numbers at random out of N ? Same question.

Exercise 9 Suppose k persons get on an elevator in the basement of an office building one mornings and they get out each with equal probability of emerging at each floor there being m floors at which they might get out, and suppose they get out independently. Find the expected number of floors at which the elevator does not stop, because nobody gets out.

Evaluate for $m=11$ and $k=10$ and vice versa.

-Steven Kannan