Probability Theory for CS

November 2, 2017

1 Expected Value and Variance

1.1 Expected Value

If $X : \Omega \to \mathbb{R}$ is a discrete random variable, then its *expected value* is

$$E[X] = \sum_{x \in \text{range}(X)} x \; p_X(x). \tag{1}$$

Example 1. Let $A \subseteq \Omega$ be an event. The *indicator random variable* of A, denoted I_A, is a random variable, defined by

$$I_A(y) = \begin{cases} 1 & \text{if } y \in A \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

The expected value of the indicator random variable is

$$E[I_A] = 1 \times p_{I_A}(1) + 0 = P\{y \in A | I_A(y) = 1\} = P(A).$$

$\Diamond$

The following lemma states how to compute the expectation of a function of a random variable.

Lemma 1. *Let $X : \Omega \to \mathbb{R}$ be a random variable, and $g : \mathbb{R} \to \mathbb{R}$ be a function. Then*

$$E[g(X)] = \sum_{x \in \text{range}(x)} g(x) \; p_X(x) \tag{3}$$

An important property of the expected value of a random variable is that it is linear.

Lemma 2 (Linearity of Expectation)**.** *Let $X : \Omega \to \mathbb{R}$ be a random variable, and $a, b \in \mathbb{R}$. Then $E[aX + b] = aE[X] + b$.*

Proof.

$$\begin{aligned} E[aX + b] &= \sum_{x \in \text{range}(X)} [ax + b]p(x) \\ &= \sum_{x \in \text{range}(X)} axp(x) + \sum_{x \in \text{range}(X)} bp(x) \\ &= aE[X] + b. \end{aligned}$$

$\square$

1.2 Variance

The variance of a random variable X, denoted $\mathrm{Var}(X)$, is defined by

$$\mathrm{Var}(X) = E[\ (X - E(X))^2\ ]. \tag{4}$$

The following provides an equivalent definition of variance.

Lemma 3.

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2. \tag{5}$$

Proof.

$$\begin{aligned}
\mathrm{Var}(X) &= E[\ (X - E(X))^2\ ] \\
&= E[\ X^2 - 2XE[X] + E[X]^2\ ] \\
&= E[X^2] - 2E[XE[X]] + E[X]^2 \qquad \text{[Linearity of Expectation]} \\
&= E[X^2] - 2E[X]E[X] + E[X]^2 \\
&= E[X^2] - (E[X])^2.
\end{aligned}$$

□

1.3 Expected Value and Variance of Sums of Random Variables

Suppose $X_1, X_2, \ldots, X_n$ are random variables. By linearity of expectation, we have

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E[X_i]. \tag{6}$$

To deal with the variance of sums of random variables, we need to introduce the notion of *covariance*.

Definition 4. The covariance of random variables X and Y is defined by

$$\mathrm{Cov}(X, Y) = E\left[(X - E(X))(Y - E(Y))\right] \tag{7}$$

The following properties of covariance are easy to prove.

Lemma 5. *For any random variables X, Y and Z and $c \in \mathbb{R}$, we have*

1. *$Cov(X, Y) = E[XY] - E[X]E[Y]$.*
2. *$Cov(X, X) = Var(X, X)$.*
3. *$Cov(X, Y) = Cov(Y, X)$.*
4. *$Cov(cX, Y) = c\ Cov(X, Y)$.*
5. *$Cov(X, Y + Z) = Cov(X,Y) + Cov(X, Z)$.*

The last property can be generalized, and we have

$$\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, Y_j). \tag{8}$$

From this, we get an expression for the variance of the sum of random variables.

$$\begin{aligned}
Var\left(\sum_{i=1}^{n} X_i\right) &= Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) && \text{[Property 2]} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} \text{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \text{Cov}(X_i, X_i) + \sum_{i=1}^{n}\sum_{j\neq i} \text{Cov}(X_i, X_j) \\
&= \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{i=1}^{n}\sum_{j\neq i} \text{Cov}(X_i, X_j).
\end{aligned}$$

1.4 Independent Random Variables: Expectation and Variance

For independent X and Y, we have

$$E[XY] = E[X]E[Y].$$

This can be shown as follows.

$$\begin{aligned}
E[XY] &= \sum_{x\in\text{range}(X)} \sum_{y\in\text{range}(Y)} xyp_XY(x,y) \\
&= \sum_{x\in\text{range}(X)} \sum_{y\in\text{range}(Y)} xyp_X(x)p_{Y|X}(y|x) \\
&= \sum_{x\in\text{range}(X)} \sum_{y\in\text{range}(Y)} xyp_X(x)p_Y(y) && \text{[Independence]} \\
&= \left[\sum_{x\in\text{range}(X)} xp_X(x)\right]\left[\sum_{y\in\text{range}(Y)} yp_Y(y)\right] \\
&= E[X]E[Y]
\end{aligned}$$

Homework Show that for independent random variables $X_1, \ldots, X_n$,

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i).$$

2 Discrete Distributions - Examples

3 Geometric Distribution

Suppose we have a biased coin with the probability of heads being p and the probability of tails being $1-p$. Then, how many coin tosses do we have to make on average before we find a head? This question is related to the *geometric distribution*. We shall explain the origin of the name in a short while.

Let X be the random variable denoting the position of the first head in the sequence of coin tosses. Then the probability that X is equal to i is

$$p_X(i) = (1-p)^{i-1}p.$$

Thus

$$E(X) = \sum_{i=1}^{\infty} i(1-p)^{i-1}p.$$

4 Binomial Distribution

Question: If you flip a biased coin with probability of heads equals to p for a total of n times, then on average, how many heads will you see?

The answer to this is related to the Binomial distribution.

Let us find the expected value of the number of heads. We perform the calculation in two ways, directly first, and then in a simpler manner using indicator random variables.

First, we compute the expectation by considering the probability of having a certain number of heads. Let X denote the random variable representing the number of heads observed in n coin tosses. Then

$$\text{Prob}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$

The expected value of the number of heads is then $E[X]$. We have

$$\begin{aligned}
E[X] &= \sum_{k=0}^{n} k\text{Prob}[X = k] \\
&= \sum_{k=0}^{n} k\binom{n}{k} p^k (1-p)^{n-k} \\
&= np \sum_{k=0}^{n-1} \binom{n-1}{k-1} p^{k-1} (1-p)^{(n-1)-(k-1)} \\
&= np(p + (1-p))^{n-1} \\
&= np,
\end{aligned}$$

where we have used the identity that

$$k\binom{n}{k} \;=\; nk\frac{(n-1)!}{(k-1)!(n-k)!} \;=\; nk\binom{n-1}{k-1}.$$

We can calculate the same value using indicator random variables and linearity of expectation. To do this, we express X as the sum of simpler random variables. Let X_i be 1 if the i^{th} toss is a head, and 0 otherwise. Then $X = \sum_{i=1}^{n} X_i$.

We now find $E[X_i]$. We have

$$E[X_i] = 1 \times \text{Prob}(X_i = 1) + 0 \times \text{Prob}(X_i = 0) = p.$$

By Linearity of Expectation,

$$E[X] = \sum_{i=1}^{n} E[X_i] = np.$$

5 Applications of Concepts

Example 2. (A Coupon Collecting Problem) Suppose there are m different types of coupons, and suppose that all kinds are equally likely. What is the expected number of coupons one needs to collect to ensure that we have one of each type?

Let X be the random variable denoting the number of coupons we need to collect to have one of each type.

Let X_i, $i = 1, \ldots, m$ denote the number of additional coupons we need to find a new type after i types have been collected. We can verify that

$$X = \sum_{i=0}^{m-1} X_i.$$

It is important to find such an appropriate decomposition of the problem into simpler problems, and verify that the original problem can be expressed using the subproblems. [1] By Linearity of Expectation, $E[X] = \sum_{i=0}^{m-1} E[X_i]$, so it suffices to find $E[X_i]$ for each i.

When i types have been collected, the probability that a new coupon is of a new type is $\frac{m-i}{m}$. Hence X_i is a geometric random variable with parameter $\frac{m-i}{m}$.

$$E[X_i] = \frac{m}{m-i}.$$

Thus we have

$$E[X] = \sum_{i=0}^{m-1} \frac{m}{m-i} = m \sum_{i=1}^{m} \frac{1}{i}.$$

Moreover,

$$Var(X) = \sum_{i=0}^{m-1} Var(X_i) = m \sum_{i=0}^{m-1} \frac{i}{(m-i)^2}.$$

$\Diamond$

6 Tail Inequalities

Suppose X_1 is a random trial taking one of n values in $\mathbb{R}$. A single trial results in the observation of a single value only. If we independently repeat the experiment, denote it by the random variable X_2, then it represents another observation of the experiment. In this manner, we consider X_1, X_2, $\ldots$, a possibly infinite sequence of random variables, all with the same probability distribution (*i.e* $\text{Prob}[X_i = j] = \text{Prob}[X_1 = j]$ for all $1 \leq i$ and $1 \leq j \leq n$), and mutually independent.

[1] What if we define Y_i, $i = 1, \ldots, m$ to be the number of coupons we need to collect to get the i^{th} type? Is it true that $X = \sum_{i=1}^{n} Y_i$?

Since the probability distributions of the random variables is identical to each other, we have $E[X_i] = E[X_1]$ for all i.

By Linearity of expectation,

$$\frac{E\left[\sum_{i=1}^{m} X_i\right]}{m} = \frac{E[X_1] + \cdots + E[X_m]}{m} = \frac{E[X_1]m}{m} = E[X_1].$$

Thus the expectation of any sum of these random variables is $E[X_1]$.

Intuition tells us that as we take many independent observations of an experiment and compute their average, this quantity is a more reliable estimate of the actual average than a calculation based on very few observations. The following inequalities formalize this intuition. All of them express the fact that as the number of samples (*i.e.* independent repetitions of an experiment) grows large, the probability of deviating from $E[X_1]$ is very small.

They are called tail inequalities because ...

6.1 Markov Inequality

The simplest tail bound is the following.

Theorem 6. *If X is a non-negative random variable, then for any $c > 0$.*

$$P(X \geq c) \leq \frac{E[X]}{c}.$$

Proof. Define the random variable I as 1 if $X \geq c$ and 0 otherwise. Since $X \geq 0$, we see that

$$X \geq cI = \begin{cases} c & \text{if } X \geq c \\ 0 & \text{if } 0 < X < c. \end{cases}$$

Taking Expectations on both sides of the inequality, we get

$$E[X] \geq cE[I] = c\text{Prob}[X \geq i],$$

which establishes the result. □

6.2 Chebyshev Inequality

The Chebyshev inequality is similar to the Markov inequality, except that the bound is expressed in terms of the variance of the random variable.

Theorem 7. *Let X be a random variable. Then for any constant $c > 0$,*

$$P(X \geq c) \leq \frac{Var(X)}{c^2}.$$

Proof. We have

$$\begin{aligned} E[X^2] &= \sum_{x \in \text{range}(X)} x^2 p(x) \\ &= \sum_{x<c} x^2 p(x) + \sum_{x \geq c} x^2 p(x) \\ &\geq \sum_{x \geq c} x^2 p(x) \\ &\geq c^2 \sum_{x \geq c} p(x) \\ &\geq c^2 \text{Prob}(X \geq c), \end{aligned}$$

which proves the result. □

6.3 Moment-Generating Functions

In order to proceed to the most powerful tail bound for independent random variables, we need to take a detour into a particular function defined on a random variable.

The values $E[X]$, $E[X^2]$, ... are sometimes called the *moments* of a random variable.

Definition 8. The *moment generating function* of a random variable X is a function $\phi : \mathbb{R} \to \mathbb{R}$ defined by

$$\phi(t) = E[e^{tX}].$$

The series expansion of the exponential function gives us

$$\phi[t] = E\left[\frac{tX}{1!} + \frac{t^2X^2}{2!} + \dots\right].$$

Differentiating with respect to t,

$$\begin{aligned}\phi'(t) &= E\left[X + \frac{2tX^2}{2!} + \frac{3t^2X^3}{3!} + \dots\right] \\ &= E\left[Xe^{tX}\right].\end{aligned}$$

Differentiating again,

$$\phi''(t) = E\left[X^2e^{tX}\right],$$

and so on.

Evaluating at $t = 0$ gives

$$\begin{aligned}\phi'(0) &= E[X] \\ \phi''(0) &= E[X^2]\end{aligned}$$

and so on. This is the reason the function above is called the moment generating function of X.

Example 3. The moment generating function of a geometric random variable with

$$P[X = n] = pq^{n-1}, \quad n = 1, 2, \dots$$

is

$$\begin{aligned}\phi(t) &= E\left[e^{tX}\right] \\ &= \sum_{n=1}^{\infty} e^{tn}pq^{n-1} \\ &= pe^t \sum_{n-1}^{\infty} (qe^t)^{n-1} \\ &= pe^t \frac{1}{1 - qe^t},\end{aligned}$$

if $qe^t < 1$, *i.e.* $0 \le t < -logq$.

Differentiating with respect to t gives

$$\phi'(t) = \frac{pe^t}{(1-qe^t)^2}$$

Evaluating at $t = 0$ gives

$$E[X] = \frac{p}{(1-q)^2} = \frac{1}{p}$$

$\Diamond$

Note. If X and Y are independent random variables, then

$$E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}].$$

The first equality is true for all random variables X and Y. The last equality does not hold in general for dependent random variables.

6.4 Chernoff Bounds

The following bound is a very powerful tail bound and is widely used in computer science. Whereas the previous inequalities based on reciprocals of polynomials in c, the next bound gives a bound based on the reciprocal of an exponential in c. This is a much more powerful result.

Another powerful feature of the following result is that it gives a two-sided bound.

Theorem 9 (Chernoff Bound)**.** *Let X have a moment generating function $\phi(t) = E[e^{tX}]$. Then for any $c > 0$,*

$$\text{Prob}[X \geq c] \leq e^{-tc}\phi(t).$$
$$\text{Prob}[X \leq c] \leq e^{-tc}\phi(t).$$

Before we prove the theorem, note that the bound holds for any t. Hence the tightest inequality can be obtained by picking the t for which the right-hand side is the smallest. This t will vary depending on the random variable X and its moment generating function.

Proof. For $t > 0$,

$$\begin{aligned} P(X \geq c) &= P(e^{tx} \geq e^{tc}) \\ &\leq E[e^t X]e^{tc} \qquad \text{[by Markov inequality.]} \end{aligned}$$

$\square$

Instead of choosing the best t, it is sometimes useful to give bounds which are independent of t. The following theorem is a common version of Chernoff bound. We give a useful lemma first, without proof.

Lemma 10 (Hoeffding Lemma)**.** *For $0 \leq p \leq 1$*

$$pe^{t(1-p)} + (1-p)e^{-tp} \leq e^{-t^2/8}.$$

This is applied to bounding tail probabilities of the sum of independent Bernoulli random variables.

Theorem 11 (Chernoff Bound for Independent Bernoulli Random Variables)**.** *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables, and S their sum. Then for any $a \geq 0$*

$$P[S - E[S] \geq a] \leq e^{-2a^2/n} P[S - E[S] \leq -a] \qquad \leq e^{-2a^2/n}.$$

Proof. For any $a > 0$, $t > 0$,

$$\begin{aligned}
P[S - E[S] \geq a] &= P[e^{t(S-E[S])} \geq e^{ta}] \\
&\leq e^{-ta} E[e^{t(S-E[S])}] && \text{[by Markov Inequality]} \\
&= e^{-ta} E\left[e^{\sum_{i=1}^n t(X_i - E[X_i])}\right] \\
&= e^{-ta} E\left[\prod_{i=1}^n e^{t(X_i - E[X_i]}\right] \\
&= e^{-ta} \prod_{i=1}^n E[e^{t(X_i - E[X_i])}] && \text{[by independence]} \\
&= e^{-ta} \prod_{i=1}^n p e^{t(1-p)} + (1-p)e^{-tp} \\
&\leq e^{-ta} \prod_{i=1}^n e^{t^2/8} \\
&= e^{-ta} e^{nt^2/8}.
\end{aligned}$$

Setting $t = 4a/n$, we get the required bound. □

7 Applications of Tail Bounds in Computer Science - Examples

Example 4. Union Bound Let A_1, A_2, $\ldots$ A_n be events. Then

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

This is called the *union bound.* We show this using Markov Inequality as follows.

Define the indicator random variables $I_{A_i} = 1$ if $A_i = 1$ and 0 otherwise, for $1 \leq i \leq n$. Then, consider $N = \sum_{i=1}^n I_{A_i}$. This random variable is greater than or equal to 1 if at least one of the events A_i occurs. *i.e.*

$$P\left(\cup_{i=1}^n A_i\right) = \text{Prob}[N \geq 1].$$

By Markov Inequality,

$$\text{Prob}[N \geq 1] \leq E[N].$$

We have

$$E[N] = E\left[\sum_{i=1}^n I_{A_i}\right] = \sum_{i=1}^n E\left[I_{A_i}\right] = \sum_{i=1}^n P(A_i).$$

Thus

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

$\Diamond$

Example 5. Suppose you have a decision problem where the answer to a question is either 1 or 0. Suppose a randomized algorithm has probability p of giving the correct answer, where $p > 0.5$. You run the algorithm n times, and output the answer that appeared the most number of times. We can get a useful upper bound on the probability that this majority answer is wrong, as follows.

First, suppose for an input x, the actual answer is 1. Let X_i denote the answer of the i^{th} run of the algorithm. Then the majority answer is correct if

$$\sum_{i=1}^{n} X_i \geq \frac{n}{2}.$$

The majority answer is incorrect if

$$\sum_{i=1}^{n} X_i < \frac{n}{2}.$$

By Chernoff Bounds, we have

$$\begin{aligned} P\left[\sum_{i=1}^{n} X_i < \frac{n}{2}\right] &= P\left[\sum_{i=1}^{n} X_i - np < \frac{n}{6}\right] \\ &\leq e^{-n/18}. \end{aligned}$$

The argument for when the actual answer for x is 0, is similar. Thus, the probability that the majority answer is wrong decreases exponentially in n, assuming $p > 0.5$. $\Diamond$

8 Central Limit Theorem and the Normal Distribution

9 Optional - Bayes' Rule

Theorem 12. *If A and B are two events, where $P(A) > 0$ and $P(B) > 0$,*

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B \cap A^c) + P(B \cap A)} \\ &= \frac{P(B|A)P(A)}{P(B|A^c)P(A^c) + P(B|A)P(A)}. \end{aligned}$$

The proof of this theorem directly follows from the definition of conditional probability.

Bayes' theorem simplifies arguments involving conditional probability. We give an illustration.

Example 6. We have two coins, a fair coin, and a biased coin where the chance of a head is 0.3. We choose a coin randomly with equal chance of it being the first coin and the second.

Now, we throw the chosen coin twice, and observe that both outcomes are heads.

What is the probability that the chosen coin was the biased coin? $\Diamond$