STAT 131 - Intro to Probability Theory

Lecture 4: Expectation

Suzana Șerboi, Ph.D.
Mathematics and Statistics, UCSC

Definition of expectation

One of the most important concepts in probability theory is that of the expectation of a random variable.

The **expected value** (also called the **expectation** or **mean**) of of a discrete r.v. X whose distinct possible values are $x_1, x_2, \ldots$ is defined by:

$$E(X) = \sum_{i=1}^{\infty} x_j P(X = x_i).$$

If the support is finite, then the formula can be replaced by a finite sum. We can also write $E(X) = \sum_x \underbrace{x}_{\text{value}} \underbrace{P(X = x)}_{\text{PMF at } x}$.

In words, the expected value of X is a weighted average of the possible values that X can take on, each value being weighted by the probability that X assumes it.

Example Let X be the result of rolling a fair 6-sided die, so X takes on the values $1, 2, 3, 4, 5, 6$, with equal probabilities.

Intuitively, we should be able to get the average by adding up these values and dividing by 6. Using the definition, the expected value is

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = \frac{1}{6}(1 + 2 + \cdots + 6) = 3.5.$$

Note that X never equals its mean in this example. This is similar to the fact that the average number of children per household in some country could be 1.8 , but that doesn't mean that a typical household has 1.8 children!

Example Recall that if $X \sim \mathrm{Bern}(p)$ then X has PMF $p_X(1) = P(X = 1) = p$ and $p_X(0) = P(X = 0) = 1 - p$.

Then

$$E(X) = 1 \cdot p + 0 \cdot (1 - p) = p.$$

Intuitively, this makes sense since it is between the two possible values of X, compromising between 0 and 1 based on how likely each is.

Expectation - Discrete Uniform distribution

Let $X \sim \mathrm{DUnif}(\{1, \ldots, n\})$. That is, X takes the values $1, \ldots, n$ and

$$p_X(x) = \begin{cases} \frac{1}{n} & x = 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbb{E}(X) = \sum_{x=1}^{n} x \frac{1}{n} = \frac{1}{n} \sum_{x=1}^{n} x = \frac{1}{n}(1 + 2 + \cdots + n) = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$$

If $n = 6$ this corresponds to the expected value of a dice roll: $7/2$.

> If X and Y are discrete r.v.s with the same distribution, then
>
> $$E(X) = E(Y)$$
>
> (if either side exists).

Proof. In the definition of $E(X)$, we only need to know the PMF of X.

The converse of the above proposition is false since the expected value is just a one-number summary, not nearly enough to specify the entire distribution.

Expectation - Binomial distribution

Let $X \sim \text{Bin}(n, p)$, let's find $E(X)$:

$$E(X) = \sum_{k=0}^{n} k P(X = k) = \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k}.$$

We will use: $k\binom{n}{k} = n\binom{n-1}{k-1}$. This is easy to check algebraically (using the fact that $m! = m(m-1)!$ for any positive integer m).

$$\sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = \sum_{k=0}^{n} n \binom{n-1}{k-1} p^k q^{n-k}$$

$$= \sum_{k=0}^{n} np \binom{n-1}{k-1} p^{k-1} q^{n-k}$$

$$= np \underbrace{\sum_{j=0}^{n} \binom{n-1}{j} p^j q^{n-1-j}}_{Bin(n-1,p) \text{ PMF}} = np \cdot 1 = np$$

Independent and identically distributed (i.i.d)

We will often work with random variables that are independent and have the same distribution. We call such r.v.s **independent and identically distributed**, or **i.i.d.** for short.

Random variables are independent if they provide no information about each other; they are identically distributed if they have the same PMF (or equivalently, the same CDF).

> If $X \sim \mathrm{Bin}(n, p)$, viewed as the number of successes in n independent Bernoulli trials with success probability p, then we can write $X = X_1 + \ldots + X_n$ where the X_i are i.i.d. $\mathrm{Bern}(p)$.

Proof. Let $X_i = 1$ if the ith trial was a success, and 0 if the ith trial was a failure. It's as though we have a person assigned to each trial, and we ask each person to raise their hand if their trial was a success. If we count the number of raised hands (which is the same as adding up the X_i), we get the total number of successes.

Linearity of expectation

The most important property of expectation is linearity. The expected value of a sum of r.v.s is the sum of the individual expected values and we can take out constant factors from an expectation:

For any r.v.s X, Y and any constant c,

$$E(X+Y) = E(X) + E(Y)$$

$$E(cX) = cE(X).$$

Linearity is an extremely handy tool for calculating expected values, often allowing us to bypass the definition of expected value altogether.

Expectation - Binomial Distribution Let $X \sim \mathrm{Bin}(n, p)$. Using linearity of expectation, we obtain a much shorter path to the result $E(X) = np$.
We write X as the sum of n independent $\mathrm{Bern}(p)$ r.v.s: $X = I_1 + \ldots + I_n$, where each I_j has expectation $E(I_j) = p$.
By linearity, $E(X) = E(I_1) + \ldots + E(I_n) = np$.

Expectation - Hypergeometric distribution

Let $X \sim \text{HGeom}(w, b, n)$, interpreted as the number of white balls in a sample of size n drawn without replacement from an urn with w white and b black balls.

As in the Binomial case, we can write X as a sum of Bernoulli random variables,

$$X = I_1 + \ldots + I_n,$$

where I_j equals 1 if the jth ball in the sample is white and 0 otherwise.

By symmetry, $I_j \sim \text{Bern}(p)$ with $p = w/(w+b)$, since unconditionally the j th ball drawn is equally likely to be any of the balls.

Unlike in the Binomial case, the I_j are not independent, since the sampling is without replacement: given that a ball in the sample is white, there is a lower chance that another ball in the sample is white. However, linearity still holds for dependent random variables! Thus,

$$E(X) = n \cdot \frac{w}{w+b}.$$

Geometric distribution $X \sim \text{Geom}(p)$

Consider a sequence of independent Bernoulli trials, each with the same success probability $p \in (0, 1)$, with trials performed until a success occurs. Let X **be the number of failures before the first successful trial**. Then X has the Geometric distribution with parameter p. We write $X \sim \text{Geom}(p)$.

Example: If we flip a fair coin until it lands Heads for the first time, then the number of Tails before the first occurrence of Heads is distributed as $\text{Geom}(1/2)$.

Typical application: how many defective products in a line do I need to find before finding a non-defective product.

Geometric PMF: If $X \sim \text{Geom}(p)$, then the PMF of X is $P(X = k) = q^k p$ for $k = 0, 1, 2, \ldots$, where $q = 1 - p$.

To get the Geometric PMF, imagine the Bernoulli trials as a string of 0's (failures) ending in a single 1 (success). Each 0 has probability $q = 1 - p$ and the final 1 has probability p, so a string of k failures followed by one success has probability $q^k p$.

Expectation - Geometric distribution

Let $X \sim \text{Geom}(p)$. By definition,

$$E(X) = \sum_{k=0}^{\infty} kq^k p, \text{ where } q = 1 - p.$$

The geometric series

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q} \text{ converges when } 0 < q < 1.$$

But the above sum it's not a geometric series because of the extra k multiplying each term. But we notice that each term looks similar to kq^{k-1}, the derivative of q^k (with respect to q), so we differentiate both sides with respect to q, and get $\sum_{k=0}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}$.

$$\text{Thus } E(X) = \sum_{k=0}^{\infty} kq^k p = pq \sum_{k=0}^{\infty} kq^{k-1} = pq \frac{1}{(1-q)^2} = \frac{q}{p}.$$

Negative Binomial $X \sim \text{NBin}(r, p)$

The Negative Binomial distribution generalizes the Geometric distribution: instead of waiting for just one success, we can wait for any predetermined number r of successes.

Sequence of independent Bernoulli trials, each with the same success probability $p \in (0, 1)$, **X is the number of failures before the rth success**.

Typical application: how many defective products in a line do I need to find before finding the rth non-defective product.

Negative Binomial PMF: If $X \sim \text{NBin}(r, p)$, then the PMF of X is

$$P(X = n) = \binom{n + r - 1}{r - 1} p^r q^n \text{ for } n = 0, 1, 2 \ldots, \text{ where } q = 1 - p.$$

To get the Negative Binomial PMF, imagine a string of 0's and 1's, with 1's representing successes. The probability of any specific string of $n 0$'s and $r 1$'s is $p^r q^n$. How many such strings are there? Because we stop as soon as we hit the r th success, the string must terminate in a 1. Among the other $n + r - 1$ positions, we choose $r - 1$ places for the remaining 1's to go.

Expectation - Negative Binomial

Let $X \sim \mathrm{NBin}(r, p)$, viewed as the number of failures before the r th success in a sequence of independent Bernoulli trials with success probability p. Then we can write $X = X_1 + \cdots + X_r$ where the X_i are i.i.d. $\mathrm{Geom}(p)$.

Proof. See Theorem 4.3.10. page 161.

Using linearity, the expectation of the Negative Binomial now follows without any additional calculations.

Let $X \sim \mathrm{NBin}(r, p)$. We write $X = X_1 + \cdots + X_r$, where the X_i are i.i.d. $\mathrm{Geom}(p)$. By linearity,

$$E(X) = E\left(X_1\right) + \cdots + E\left(X_r\right) = r \cdot \frac{q}{p}.$$

Indicator random variable

The indicator r.v. I_A (or $I(A)$) for an event A is defined to be 1 if A occurs and 0 otherwise. So I_A is a Bernoulli random variable, where success is defined as "event A occurs" and failure is defined as "event A does not occur".

Some useful properties of indicator r.v.s are summarized below.

Let A and B be events. Then the following properties hold:

- $(I_A)^k = I_A$ for any positive integer k.
- $I_{A^c} = 1 - I_A$.
- $I_{A\cap B} = I_A I_B$.
- $I_{A\cup B} = I_A + I_B - I_A I_B$.

Indicator r.v.s are important as they provide a link between probability and expectation.

Fundamental bridge between probability and expectation

There is a one-to-one correspondence between events and indicator r.v.s, and the probability of an event A is the expected value of its indicator r.v. I_A:

$$P(A) = E(I_A)$$

Proof. For any event A, we have an indicator r.v. I_A. This is a one-to-one correspondence since A uniquely determines I_A and vice versa. Since $I_A \sim Bern(p)$ with $p = P(A)$, we have $E(I_A) = P(A)$.

Note The fundamental bridge is useful in many expected value problems. We can often express a complicated discrete r.v. whose distribution we don't know as a sum of indicator r.v.s, which are extremely simple. The fundamental bridge lets us find the expectation of the indicators; then, using linearity, we obtain the expectation of our original r.v.

Example (Matching Cards)

We have a well-shuffled deck of n cards, labeled 1 through n. A card is a match if the card's position in the deck matches the card's label. Let X be the number of matches; find $E(X)$.

Solution Let's write $X = I_1 + I_2 + \cdots + I_n$, where

$$I_j = \begin{cases} 1 & \text{if the } j \text{ th card in the deck is a match,} \\ 0 & \text{otherwise.} \end{cases}$$

In other words, I_j is the indicator for A_j, the event that the j th card in the deck is a match.

By the fundamental bridge,

$$E\left(I_j\right) = P\left(A_j\right) = \frac{1}{n} \text{ for all } j.$$

By linearity,

$$E(X) = E\left(I_1\right) + \cdots + E\left(I_n\right) = n \cdot \frac{1}{n} = 1.$$

The expected number of matched cards is 1, regardless of n.

Law of unconscious statistician (LOTUS)

A function of a random variable is a random variable. That is, if X is a random variable, then X^2, e^X, and $\sin(X)$ are also random variables, as is $g(X)$ for any function $g : \mathbb{R} \to \mathbb{R}$. See Section 3.7 in the textbook for more details.

It turns out that it is possible to find $E(g(X))$ directly using the distribution of X, without first having to find the distribution of $g(X)$.

If X is a discrete r.v. and g is a function from $\mathbb{R}$ to $\mathbb{R}$, then

$$E(g(X)) = \sum_x g(x)P(X = x),$$

where the sum is taken over all possible values of X.

Example Let X denote a random variable that takes on any of the values $-1, 0$, and 1 with respective probabilities $P(X = -1) = .2$, $P(X = 0) = .5$ and $P(X = 1) = .3$. Then $E[X^2] = (-1)^2(.2) + 0^2(.5) + 1^2(.3) = .5$.

Variance and standard deviation

The **variance** of an r.v. X is:

$$Var(X) = E(X - E(X))^2.$$

The square root of the variance is called **standard deviation (SD)**:

$$SD(X) = \sqrt{Var(X)}.$$

For any r.v. X,

$$Var(X) = E(X^2) - (E(X))^2$$

Proof. Let $\mu = E(X)$. Using linearity of expectation,

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

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

Example Consider a fair die with $X = i$ is "number i rolled" then we have seen that $E(X) = \frac{7}{2}$ and

$$E\left(X^2\right) = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6}$$
$$= \frac{1+4+9+16+25+36}{6} = \frac{91}{6}$$

$$\text{Then } \operatorname{Var}(X) = E(X^2) - (E(X))^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

Some properties of variance:

- $Var(X + c) = Var(X)$ for any constant c.
- $Var(cX) = c^2 Var(X)$ for any constant c. Variance is not linear!
- If X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$
- $Var(X) \geq 0$, with equality if and only if $P(X = a) = 1$ for some constant a.

Variance - Geometric distribution

Let $X \sim Geom(p)$. We know $E(X) = q/p$. By LOTUS:

$$E(X^2) = \sum_{k=0}^{\infty} k^2 P(X = k) = \sum_{k=0}^{\infty} k^2 pq^k = \sum_{k=1}^{\infty} k^2 pq^k$$

Taking derivative the geometric series $\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$ we get $\sum_{k=0}^{\infty} kq^{k-1} = \sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}$.

Multiplying both sides by q and taking derivative again we have:

$$\sum_{k=1}^{\infty} kq^k = \frac{q}{(1-q)^2} \Rightarrow \sum_{k=1}^{\infty} k^2 q^{k-1} = \frac{1+q}{(1-q)^3}$$

$$Var(X) = E(X^2) - (E(X))^2 = pq\frac{(1+q)}{(1-q)^3} - \left(\frac{q}{p}\right)^2$$

$$= \frac{q(1+q)}{p^2} - \left(\frac{q}{p}\right)^2 = \frac{q}{p^2}.$$

Variance - Negative Binomial distribution

Since $NBin(r, p)$ r.v. can be represented as a sum of r i.i.d $Geom(p)$ r.v.s, and since variance is additive for independent random variables, it follows that the variance of $NBin(r, p)$ is $r \cdot \frac{q}{p^2}$.

Variance - Binomial distribution

Let $X \sim Bin(n, p)$ and represent $X = I_1 + I_2 + \cdots + I_n$ where I_j is the indicator of the jth trial being a success. Each I_j has variance:

$$Var(I_j) = E(I_j^2) - (E(I_j))^2 = p - p^2 = p(1 - p).$$

Note that $I_j^2 = I_j$. Then, since I_j are independent, we can add their variances:

$$Var(X) = Var(I_1) + Var(I_2) + \cdots + Var(I_n) = np(1 - p).$$

Poisson Distribution $X \sim Pois(\lambda)$

An r.v. X has the **Poisson distribution with parameter λ**, where $\lambda > 0$, if the PMF of X is: $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$, for $k = 0, 1, 2, \ldots$. We write this as $X \sim Pois(\lambda)$.

You can show that this is a valid PMF using the Taylor series: $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$

The Poisson distribution is often used in situations where we are counting the number of success in a particular region or interval of time, and there are a **large number of trials, each with a small probability of success**. Some examples of r.v.s that could follow a distribution that is approx Poisson:

- Number of emails your receive in an hour.
- Number of chips in a chocolate chip cookie.
- Number of earthquakes in a year in some region of the world.

The parameter λ can be interpreted as the rate of occurrence of these rare events. For example $\lambda = 20$ emails per hour, $\lambda = 10$ chips per cookie, $\lambda = 2$ earthquakes per year.

Expectation - Poisson distribution

Let $X \sim Pois(\lambda)$. Then $E(X) = \sum_{k=0}^{\infty} kP(x = k) = \sum_{k=0}^{\infty} e^{-\lambda} k \frac{\lambda^k}{k!}$

$$= e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda.$$

Variance - Poisson distribution

For variance, we first need $E(X^2) = \sum_{k=0}^{\infty} k^2 P(X = k) = \sum_{k=0}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!}$

If we differentiate the series: $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$ with respect to λ, and multiply by λ in both sides: $\sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} = \lambda e^{\lambda}$.

Repeat the procedure (differentiate and multiply by λ):

$\sum_{k=1}^{\infty} k^2 \frac{\lambda^{k-1}}{k!} = e^{\lambda} + \lambda e^{\lambda} = e^{\lambda}(1 + \lambda) \Rightarrow \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} = \lambda e^{\lambda}(1 + \lambda)$

$Var(X) = E(X^2) - (E(X))^2 = e^{-\lambda} \lambda e^{\lambda}(1 + \lambda) - \lambda^2 = \lambda(1 + \lambda) - \lambda^2 = \lambda.$