36-700: Probability and Mathematical Statistics **Spring 2019**

Lecture 4: Bivariate Distributions, Independence

Lecturer: Jing Lei

4.1 Review and Outline

Last time we saw:

- Transformations of RVs
- Expectations and properties of expectations
- Variances, covariances
- Bivariate distribution, independence

In this lecture we'll discuss inequalities for expectations, conditional expectations (3.5 of Wasserman) and moment generating functions (3.6 of Wasserman).

4.2 Inequalities for Expectations

Broadly, very often we want to upper bound certain expectations. Two inequalities are very useful in this context:

1. **Cauchy-Schwarz inequality:** The Cauchy-Schwarz inequality says that:

   $$\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}.$$

   Exercise: Use the Cauchy-Schwarz inequality to verify that the correlation between two random variables is bounded between -1 and 1.

2. **Jensen's inequality:** First, we need to recall what convex functions are: a function g is convex if for every x, y and $\alpha \in [0, 1]$,

   $$g(\alpha x + (1-\alpha)y) \leq \alpha g(x) + (1-\alpha)g(y).$$

   Pictorially, convex functions are ones for which the line joining any two points on the curve lies entirely above the curve.

   Jensen's inequality says that for a convex g:

   $$g(\mathbb{E}[X]) \leq \mathbb{E}g(X).$$

4.3 Conditional Expectation

If we have two random variables X and Y and we would like to compute the average value of Y amongst all the times that $X = x$.

As a quick detour, why might we want to do this? One reason is, if we are trying to predict Y from X (this is sometimes called regression). Intuitively, our best prediction would be the average of Y values for all the points where $X = x$. We will re-visit this idea later on in the course.

The conditional expectation of a random variable is just the average with respect to the conditional distribution, i.e.,

$$\mathbb{E}[Y|X = x] = \sum_y y f_{Y|X}(y|x) \quad \text{or} \quad = \int_y y f_{Y|X}(y|x) dy.$$

An important point about the conditional expectation is that it is a function of X, unlike the expectation of a random variable (which is just a number). Usually, we use $\mathbb{E}[Y|X]$ to denote the random variable whose value is $\mathbb{E}[Y|X = x]$, when $X = x$. This is something that you should pause to digest.

Example: Suppose I draw $X \sim U[0, 1]$ and then I draw $Y|X = x \sim U[x, 1]$. What is the conditional expectation $\mathbb{E}[Y|X]$?

A reasonable guess would be that $\mathbb{E}[Y|X = x] = (1 + x)/2$, but lets do this from first principles. We first compute the conditional density of $Y|X$.

$$f_{Y|X}(y|x) = \frac{1}{1 - x}, \quad \text{for} \quad x < y < 1.$$

Now using the formula for the conditional expectation we obtain,

$$\mathbb{E}[Y|X = x] = \int_x^1 \frac{1}{1 - x} y dy = \frac{1 + x}{2}.$$

Independence: If two random variables X and Y are independent, then

$$\mathbb{E}[Y|X = x] = \mathbb{E}[Y].$$

In general, this does not go both ways, i.e., dependent random variables might also satisfy this expression. As an exercise think of an example of this.

The law of total expectation: This is also called the tower property.

$$\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y].$$

It is worth trying to parse this formula more carefully and adding some subscripts:

$$\mathbb{E}_X[\mathbb{E}_{Y|X}[Y|X]] = \mathbb{E}[Y].$$

Intuitively, this expression has a divide and conquer flavour, i.e. what it says is that to compute the average of a random variable Y, you can first compute its average over a bunch of partitions of the sample space (where some other random variable X is fixed to different values), and then average the resulting averages.

It is quite simple to prove this (by interchanging the order of the two expectations). So we will instead see an example.

Example: Suppose I had a population of people, 47% of whom were men and the remaining 53% were women. Suppose that the average height of the men was 70 inches, and the women was 71 inches. What is the average height of the entire population?

By the law of total expectation:

$$\begin{aligned}\mathbb{E}[H] &= \mathbb{E}[\mathbb{E}[H|S]] \\ &= \mathbb{E}[H|S=m]\mathbb{P}(S=m) + \mathbb{E}[H|S=f]\mathbb{P}(S=f) \\ &= 70 * 0.47 + 71 * 0.53 = 70.53.\end{aligned}$$

4.3.1 Conditional Variance

One can similarly define the conditional variance as:

$$\mathbb{V}(Y|X=x) = \mathbb{E}[(Y - \mathbb{E}[Y|X=x])^2|X=x].$$

There is an analogous law of total variance that says that:

$$\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X=x)) + \mathbb{V}(\mathbb{E}(Y|X=x)).$$

Again, just intuitively, this is a divide and conquer way to compute the variance. We first compute the variance on each partition where X is held fixed and average those (this is the first term) but we now also need to account for the fact that each variance was computed around a different mean, i.e., we need to account for the variance of the mean across the partitions. This is the second term.

4.4 The moment generating function

The moment generating function (MGF) of a random variable X is given by:

$$M_X(t) = \mathbb{E}\exp(tX).$$

In general, the MGF need not exist (just like the expectation), and sometimes will not exist for large values of t. We will just ignore this for now.

This function is called the moment generating function because its derivatives evaluated at 0 gives us the moments of X, i.e.,

$$M'_X(0) = \left[\frac{d}{dt}\mathbb{E}\exp(tX)\right]_{t=0} = \mathbb{E}\left[\frac{d}{dt}\exp(tX)\right]_{t=0} = \mathbb{E}[X\exp(tX)]_{t=0} = \mathbb{E}[X].$$

In a similar fashion:

$$M_X^{(k)}(0) = \mathbb{E}[X^k].$$

Lets do a couple of examples:

Example 1: Compute the MGF of a Bernoulli random variable, and use it to compute the mean of the random variable.

A direct computation gives us that:

$$M_X(t) = \mathbb{E}\exp(tX) = (p\exp(t) + 1 - p).$$

To compute the mean we take the first derivative and evaluate it at 0, i.e.

$$M'_X(0) = (p\exp(t))_{t=0} = p.$$

Example 2: Compute the MGF of an Exponential RV, with mean 1.

The exponential RV with mean λ has pdf:

$$f_X(x) = \lambda\exp(-\lambda x),$$

for $x \geq 0$, so the MGF is given by:

$$M_X(t) = \int_0^\infty \exp(-x)\exp(tx)dx = \int_0^\infty \exp((t-1)x) = \frac{1}{1-t},$$

when $t < 1$. The MGF does not exist for $t > 1$ since the integral above diverges.

There are two important properties of MGFs that to an extent explain their ubiquity in Statistics:

1. **Sums of independent RVs:** If we have random variables $X_1, \ldots, X_n$ which are independent and $Y = \sum_{i=1}^n X_i$ then

$$M_Y(t) = \prod_{i=1}^n M_{X_i}(t).$$

   Basically, this gives us a very easy way to calculate effectively every moment of a sum of independent random variables. We will use this repeatedly in the next lecture.

2. **Equality of MGFs:** We have seen that the MGF can give us a lot of information about a random variable. A basic question is whether the MGF completely determines a random variable. The answer (somewhat surprisingly) turns out to be yes:

   If the MGF of X and Y exist, in a neighbourhood around 0, and are equal then X and Y have the same distribution.

Exercise: Find the expectation, variance, and mgf for all the "named" univariate distributions (e.g., Ber(p), Bin(n, p), Geom(p), Poi(λ), U(a, b), $N(\mu, \sigma^2)$).