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

Lecture 3: Transformation, Expectation, and Variance

Lecturer: Jing Lei

3.1 Review and Outline

Last class we saw:

- Random variables
- The distribution function, i.e., $\mathbb{P}(X \leq x)$
- The probability mass/density functions.

My notes will mainly follow the Wasserman book and cover the following topics: transformations of random variables (Section 2.11), expectation (3.1), properties of expectation (3.2), variance and co-variance (3.3). These are covered in Chapter 2 of Casella and Berger.

3.2 A couple of quick notes

We defined the pdf for a continuous random variable as the function that satisfies:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du.$$

A consequence of this definition and the fundamental theorem of calculus is that we can always calculate the pdf from the CDF as:

$$f_X(x) = F_X'(x).$$

This is sometimes used as the definition of the density function.

Finally, we often use the symbol $\sim$ to denote "distributed as", i.e. $X \sim \text{Ber}(p)$ means that the random variable X has a Bernoulli distribution with parameter p.

3.3 Transformations of Random Variables

The basic question here is: suppose I have a random variable X with pdf/pmf f_X and CDF F_X, and I consider $Y = r(X)$, for some function r. For instance, $r(X)$ might be something like X^2 or $\exp(X)$. How do I compute the pdf/pmf or CDF of Y?

3.3.1 The discrete case

In the discrete case Y is also a discrete random variable and its pmf is given by:

$$\begin{aligned}\mathbb{P}(Y = y) &= \mathbb{P}(r(X) = y) \\ &= \mathbb{P}(\{x : r(x) = y\}) = \mathbb{P}(X \in r^{-1}(y)).\end{aligned}$$

This somewhat opaque formula is perhaps clarified via an example: suppose $X \in \{-1, 0, 1\}$, with probabilities $1/4, 1/2$ and $1/4$, and consider the random variable $Y = X^2$.

The way to proceed is to compute the different values that Y can take and then adding up the necessary probabilities. In this case Y can take the values $\{0, 1\}$ and does so with probability $1/2$ each.

3.3.2 The continuous case

This is substantially more involved. I will sketch the basic ideas here.

There is one case when things simplify: suppose that the transformation r is invertible, and we have $s = r^{-1}$ then we have the formula:

$$f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right|.$$

As a basic example, suppose we consider $X = U[0, 1]$ and $Y = X^2$. Before we apply the formula, we should try to guess what we think the pdf of Y would be. Squaring a number between $[0, 1]$ only makes it smaller, so we should expect that the density of Y would be peaked near 0, i.e., the uniform density will get "squeezed" towards 0.

Now, we can apply the formula. Observe that $r(X) = X^2$, and since $X \geq 0$, $s(Y) = \sqrt{Y}$, so we have that:

$$f_Y(y) = \frac{1}{2\sqrt{y}}.$$

Graphing this density will show that it matches our intuition. Notice again, a density is not a probability! This density $\to \infty$ near 0.

The book also describes a method to compute the density of transforms which are not invertible. We will cover this if necessary later on.

Some intuition for the formula: In general, dealing with densities takes some getting used to. Roughly, for a continuous random variable one should think of multiplying the density by the length of a small interval to approximate the probability of the random variable falling in this interval.

When you transform the random variable then the transformation affects *both* the value of the density and also the length of the interval. The term $\left|\frac{ds(y)}{dy}\right|$, is called the Jacobian of the transformation and accounts for the stretching of the interval caused by the transformation.

Here is a very rough mathematical calculation. For a very small Δ:

$$\mathbb{P}(Y \in [y_0 - \Delta, y_0 + \Delta]) \approx f_Y(y_0) 2\Delta.$$

If Y falls in the interval this is the same as saying that X falls in a slightly different interval, i.e.,

$$\begin{aligned}
\mathbb{P}(Y \in [y_0 - \Delta, y_0 + \Delta]) &= \mathbb{P}(X \in [s(y_0 - \Delta), s(y_0 + \Delta)]) \\
&\approx \mathbb{P}\left(X \in [s(y_0) - \Delta \frac{ds(y_0)}{dy}, s(y_0) + \Delta \frac{ds(y_0)}{dy}]\right) \\
&\approx f_X(s(y_0)) 2\Delta \left|\frac{ds(y_0)}{dy}\right|,
\end{aligned}$$

where in the second line we assumed the derivative was positive (otherwise we would flip the end points of the interval). Equating these gives the result.

3.4 Expectation

A common goal is to understand or summarize the behaviour of a random variable. One way to do this is by trying to understand some type of "typical behaviour" of a random variable. The expectation, or mean, or average, or first moment of a random variable is defined as:

$$\mathbb{E}[X] = \int x dF_X(x) = \int x f_X(x) dx \quad \text{or} \quad \sum_x x f_X(x).$$

In general, we say that the expectation does not exist if $\mathbb{E}|X| = \infty$.

One way to think of the expectation of a random variable is by supposing you could repeat the underlying experiment many times in order to obtain new copies of the random variable $X_1, \ldots, X_n$. In this case, the expectation $\mathbb{E}(X) \approx \frac{1}{n} \sum_{i=1}^n X_i$. This is called the law of large numbers and we will re-visit this in a few lectures.

Some examples:

1. Suppose $X \sim \text{Ber}(p)$ then,

$$\mathbb{E}[X] = p * 1 + (1 - p) * 0 = p.$$

2. Suppose that $X \sim U[-2, 4]$, then

$$\mathbb{E}[X] = \int_{-2}^{4} \frac{x}{6} dx = \frac{1}{12}(16 - 4) = 1.$$

3. The most famous example of a distribution whose mean doesn't exist is a Cauchy distribution. It has density $f_X(x) = \frac{1}{\pi(1+x^2)}$ for $x \in (-\infty, \infty)$.

$$\mathbb{E}|X| = \int_{-\infty}^{\infty} \frac{|x|}{\pi(1+x^2)} dx = \frac{2}{\pi} \int_{0}^{\infty} \frac{x}{1+x^2} dx.$$

You can check that the anti-derivative of $\frac{x}{1+x^2}$ is $\frac{\log(1+x^2)}{2}$. So the integral above is:

$$\mathbb{E}|X| = \frac{1}{\pi} \lim_{M \to \infty} \log(1 + M^2) = \infty.$$

3.5 Properties of Expectations

3.5.1 Expectation of transformation

For a random variable $Y = r(X)$ the expectation is given by

$$\mathbb{E}[Y] = \mathbb{E}[r(X)] = \int_x r(x) dF_X(x).$$

This is often called the rule of the lazy statistician.

A very important special case of this is when $Y = \mathbb{I}_A(X)$, i.e, $Y = 1$ if $X \in A$ and 0 otherwise. This is called an indicator random variable. Then we have:

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{I}_A(X)] = \int_x \mathbb{I}_A(x) dF_X(x) = \int_{x \in A} f_X(x) dx = \mathbb{P}(X \in A).$$

3.5.2 Linearity of expectations

One of the most commonly used properties of expectations is that they are linear, i.e., if you have a collection of RVs $X_1, \ldots, X_n$ and some constants $a_1, \ldots, a_n$, then

$$\mathbb{E}[\sum_i a_i X_i] = \sum_i a_i \mathbb{E}[X_i].$$

There are many nice applications of this fact, but a basic one is that we can now calculate the mean of binomial (the C&B book does this directly), but we can use the fact that a $\text{Bin}(n, p)$ random variable is just a sum of n independent Bernoulli's.

So, if $X \sim \text{Bin}(n, p)$ and $Y_i \sim \text{Ber}(p)$ then:

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

3.5.3 Moments and Central Moments

For a random variable X we define its k^{th} moment to be:

$$\mu_k := \mathbb{E}[X^k].$$

Usually we denote the expectation as μ instead of μ_1. The central moments are then defined as:

$$\alpha_k := \mathbb{E}[(X - \mu)^k].$$

3.6 Bivariate Distributions and Independence of Random Variables

3.6.1 Bivariate Distributions

Suppose we have a pair of discrete random variables X, Y then we can define their **joint** pmf by:

$$f_{XY}(x, y) = \mathbb{P}(X = x \text{ and } Y = y).$$

In the continuous case the joint density function is the one that integrates to give us probabilities, i.e., it is the non-negative function that for any set $A \subset \mathbb{R} \times \mathbb{R}$ satisfies the property that:

$$\int_{(x,y)\in A} f_{XY}(x, y)\, dx\, dy = \mathbb{P}((X, Y) \in A).$$

An example: Suppose (X, Y) are jointly uniform over the unit square. Then it has density:

$$f_{XY}(x, y) = 1,$$

if $0 \leq x \leq 1$ and $0 \leq y \leq 1$. Use this to calculate the probability $\mathbb{P}(X \leq 1/4, Y \leq 1/2)$. To do this we integrate the joint density:

$$\mathbb{P}(X \leq 1/4, Y \leq 1/2) = \int_0^{1/4} \int_0^{1/2} 1 \; dx \; dy = \frac{1}{8}.$$

3.6.2 Independence

Formally, X and Y are independent if for every pair of sets A, B we have that:

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B).$$

In words, we say that two random variables are independent if their joint probability is equal to the product of their marginal probabilities. This seems to suggest that we need to check this condition for every possible pair of sets A and B. It turns out that we can instead check that the joint pmf/pdf factorizes, i.e., that

$$f_{XY}(x, y) = f_X(x) f_Y(y),$$

for all (x, y). Equivalently, if you can write the joint probability of two random variables as the product of a function of just the first one, and just the second one then the two random variables are independent, i.e., if

$$f_{XY}(x, y) = h(x)g(y),$$

for some pair of functions h, g then the two random variables are independent.

Example: Suppose X and Y have density:

$$f_{XY}(x, y) = 2\exp(-(x + 2y)),$$

for $x \geq 0, y \geq 0$. Are X and Y independent?

We see that we can write their joint density as:

$$f_{XY}(x, y) = 2\exp(-(x + 2y)) = 2\exp(-x)\exp(-2y) = h(x)g(y),$$

so we can conclude that these random variables are independent.

An important fact about independent random variables X and Y, is that for any functions f and g,

$$\mathbb{E}(f(X)g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y)).$$

More generally, we can define joint independence for a set of RVs $X_1, \ldots, X_n$ by requiring that their joint factorizes as:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathbb{P}(X_1 = x_1) \ldots \mathbb{P}(X_n = x_n).$$

For jointly independent random variables:

$$\mathbb{E}(g_1(X_1)g_2(X_2) \ldots g_k(X_k)) = \mathbb{E}(g_1(X_1)) \ldots \mathbb{E}(g_k(X_k)).$$

3.7 Variance and Covariance

The second central moment of a random variable is called its variance. The variance of a distribution measures its spread – roughly how far it is on average from its mean. We use σ_X^2 to denote the variance of X. Its square root, i.e., σ_X is the standard deviation.

A basic fact is that:

$$\sigma_X^2 = \mathbb{E}(X-\mu)^2 = \mathbb{E}[X^2 + \mu^2 - 2\mu X] = \mathbb{E}(X^2) - \mu^2.$$

For constants a, b, we have

$$\sigma_{aX+b}^2 = a^2\sigma_X^2.$$

For two random variables X, Y we define their covariance as:

$$\mathrm{Cov}(X,Y) = \mathbb{E}((X-\mu_X)(Y-\mu_Y)).$$

The covariance is a measure of association. We can re-write it as:

$$\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

One can also think of it as a measure of a type of (linear) deviation from independence. For independent random variables the covariance is 0. We often work with a standardized form of the covariance, known as the correlation:

$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

We will prove this either in an assignment or during a later lecture but the correlation is always between -1 and 1, i.e.,

$$-1 \leq \mathrm{Cor}(X,Y) \leq 1.$$

The covariance of a random variable and itself is just its variance. In general, for a collection of random variables:

$$\text{Variance}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathrm{Cov}(X_i, X_j).$$

Exercise: Prove the above fact. You can use the following result: for a set of numbers $x_1, \ldots, x_n$,

$$\left(\sum_{i=1}^n x_i\right)^2 = \sum_{i=1}^n \sum_{j=1}^n x_i x_j.$$

3.7.1 Variance of averages of independent random variables

We will cover this in much more detail when talking about inequalities so this is just a teaser. Suppose I take the average of n independent and identically distributed random variables $X_1, \ldots, X_n$ and compute the variance of the average. We can use the above formula to see that:

$$\text{Variance}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Variance}(X_i) = \frac{\sigma_X^2}{n}.$$

There are two important points to notice:

1. The variance of the average is much smaller than the variance of the individual random variables: this is one of the core principles of statistics and helps us estimate various quantities reliably by making repeated measurements.

2. It is also worth trying to understand why we need independent measurements. The extreme case of non-independence is when $X_1 = X_2 = \ldots = X_n$, in this case we would have that:

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

   There is no reduction of variance by taking repeated measurements if they strongly influence each other.