LECTURE NOTES 4

Last class:

- Transformations of RVs
- Expectations and properties of expectations
- Moments, variances, co-variances.

This class: varances, for expectations, conditional expectations and moment generating functions.

1 A quick note

A very important special case of the rule of the lazy statistician 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:

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

2 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:

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

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:

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

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

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

2.1 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:

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

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 independent measurements are useful. The extreme case of non-independence is when $X_1 = X_2 = \ldots = X_n$, in this case we would have that:

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

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

3 Inequalities for Expectations

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:

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

Observe this is almost exactly the definition of convexity.

4 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:

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

4.1 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:

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

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.

5 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.

6 Some common distributions

This section will go over some common univariate distributions, and describe where they arise. In your homework you will do several exercises computing means, variances, MGFs etc. We will cover a few multivariate distributions when the need arises.

6.1 Univariate distributions

1. **Poisson(λ):** The Poisson distribution often arises in modelling various types of arrival processes. If we have independent arrivals (buses arriving at a bus stop), a fixed rate of arrivals, and if we have that the number of arrivals in any interval of time is proportional to the length of the interval, then the number of arrivals in an interval will have a Poisson distribution.

   The Poisson distribution is discrete with pmf:

   $$\mathbb{P}(X = k) = \exp(-\lambda)\frac{\lambda^k}{k!}.$$

   Both the mean and variance of the Poisson are equal to λ.

2. **Gaussian $N(\mu, \sigma^2)$:** This is the classic bell-curve. The density function is given by:

   $$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

   The mean of the Gaussian is μ and its variance is σ^2. We will see this distribution many times throughout this course so we will not say more about it for now.

   A standard Gaussian is one that has mean 0 and variance 1.

3. **Chi-square:** If you square and add k independent standard Gaussian RVs, you obtain a chi-squared random variable with k degrees of freedom. They often arise in

statistical hypothesis tests. The chi-squared distribution has a complicated pdf, so for now lets just compute its mean:

$$\begin{aligned}\mathbb{E}[\chi_k^2] &= \mathbb{E}\left[\left(\sum_{i=1}^{k} X_i\right)^2\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{k}\sum_{j=1}^{k} X_i X_j\right] \\ &= \mathbb{E}\left[\sum_{i=1}^{k} X_i^2\right] = k.\end{aligned}$$

4. **Cauchy:** We saw this distribution as an example of one that had no expectation. The density is given by:

$$f_X(x) = \frac{1}{\pi(1+x^2)}.$$

The Cauchy distribution is the distribution of the ratio of two independent standard Gaussian random variables, i.e., if I draw $X \sim N(0,1)$ and $Y \sim N(0,1)$ independently and compute X/Y it will have a Cauchy distribution.