LECTURE NOTES 5

1 Review and A Few Other Things

1. Recall that $p(y|x) = p(x,y)/p(x)$. For discrete random variables, this can be written as
$$P(Y = y|X = x) = p(y|x) = \frac{p(x,y)}{p(x)} = \frac{P(Y = y, X = x)}{P(X = x)}.$$
Note that
$$p(x,y) = p(x)p(y|x) = p(y)p(x|y).$$
More generally (chain rule):
$$p(x_1, \ldots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_n|x_1, \ldots, x_{n_1}).$$

2. The lazy statistician rule holds for multivariate distributions too:
$$\mathbb{E}[r(X,Y)] = \int\int r(x,y)p(x,y)\ dx\, dy.$$
Example: Let $r(x,y) = xy$. Then
$$\mathbb{E}[r(X,Y)] = \int\int xyp(x,y)\ dx\, dy.$$

3. Transformations for multivariate distributions. Use the cdf method. Let $Z = r(X,Y)$. Then
$$F_Z(z) = P(Z \leq z) = P(r(X,Y) \leq z) = \int\int_{A_z} r(x,y)p(x,y)\, dx\, dy$$
where $A_z = \{(x,y) :\ r(x,y) \leq z\}$. and $p_Z(z) = F'_Z(z)$.

4. Law of total expectation:
$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]].$$
This is true for functions of random variables too:
$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]].$$

2 Concentration

We saw in the last class that averages of many independent random variables tend to have much smaller variance than the individual random variables. In some sense this is comforting, but what we'd really like are quantitative bounds. We'd like to say that, for appropriate

functions T, $T(X_1, \ldots, X_n)$ is close to some fixed number θ with high probability. For example, $T(X_1, \ldots, X_n) = \overline{X}_n$ is close to μ with high probability.

Example. Suppose I toss a fair coin n times, and denote the outcome of the i$^{\text{th}}$ toss X_i where $X_i = -1$ if tails and $X_i = +1$ if heads. Note that $\mu \equiv \mathbb{E}[X_i] = 0$ and $\sigma^2 \equiv \text{Var}[Y_i] = 1$. We call μ the population mean and σ^2 the population variance.

Let

$$Y_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

be the sample mean. First, let us compute some basic properties:

1. **The mean:**

$$\mathbb{E}[Y_n] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}X_i = 0$$

   Hence, $\mathbb{E}[Y_n] = \mu$.

2. **The variance:**

$$\text{Var}(Y_n) = \mathbb{E}[Y_n^2] = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{1}{n}.$$

   Hence, $\text{Var}[Y_n] = \sigma^2/n$.

3. **The worst case:** What if we got unlucky and every coin toss landed heads. Can this even happen? How often will this happen? In this case we would have that, $Y_n = \frac{1}{n} \sum_{i=1}^{n} X_i = 1$.

4. **The best case:** We might get really lucky and have exactly $n/2$ heads and $n/2$ tails (assuming n is even). In this case we would have that, $Y_n = \frac{1}{n} \sum_{i=1}^{n} X_i = 0$.

So to summarise, in the good case the sample mean of the random variables will coincide exactly with the population mean, and in the worst-case it can be quite far from it. What we would really like to understand is the *typical case.* In this lecture, we will build the tools to show that in this example (as well as many others) most of the time the sample mean is within $c/\sqrt{n}$ from the mean, for some small constant c.

This is called the concentration of measure phenomenon. It is trying to quantify our intuition that it is very unlikely that *independent* random variables will conspire against us to cause the sample mean to be very far from the population mean.

3 Markov's Inequality

Our first inequality will not really help us with our case study problem but will help us prove more useful inequalities.

Markov's inequality: Suppose X is a non-negative random variable and that $\mathbb{E}[X]$ exists. Then for any $t \geq 0$:

$$\mathbb{P}(X > t) \leq \frac{\mathbb{E}[X]}{t}.$$

A slightly different way of writing this is that:

$$\mathbb{P}(X > t(\mathbb{E}[X])) \leq \frac{1}{t}.$$

In words: a positive random variable is unlikely to be much larger than its mean.

Proof: Since $X > 0$,

$$\begin{aligned}\mathbb{E}[X] &= \int_0^\infty x f_X(x)dx = \int_0^t x f_X(x)dx + \int_t^\infty x f_X(x)dx \\ &\geq \int_t^\infty x f_X(x)dx \geq t \int_t^\infty f_X(x)dx = t\ \mathbb{P}(X \geq t).\end{aligned}$$

Unfortunately for us the random variable Y from the previous section is not a positive random variable. So we need to derive some more general inequalities.

4 Chebyshev's Inequality

Chebyshev's inequality is a simple consequence of Markov's inequality but one that will already lead us to some very useful conclusions. Let $\mathbb{E}[X] := \mu_X$ and $\text{Var}(X) := \sigma_X^2$. We will assume that both exist.

Chebyshev's inequality: For any $k \geq 0$:

$$\mathbb{P}(|X - \mu_X| > k\sigma_X) \leq \frac{1}{k^2}.$$

In words: The deviation from the mean of any random variable is on the order of the standard deviation (hence the name), and more concretely the probability that a random variable is more than k standard deviations away from its mean is at most $1/k^2$.

An application: Before we prove Chebyshev's inequality lets apply it to the case study problem. Recall, that we had in that example for the average Y:

$$\mu_Y = 0 \qquad \text{and} \qquad \sigma_Y = \frac{1}{\sqrt{n}},$$

so that by Chebyshev's inequality we have that:

$$\mathbb{P}\left(|Y| > \frac{k}{\sqrt{n}}\right) \leq \frac{1}{k^2}.$$

This is precisely the type of quantitative bound we were hoping for. The average of n independent random variables is typically roughly $1/\sqrt{n}$ distance from its expectation.

Proof: To prove this inequality we are just going to use Markov's inequality, i.e.

$$\begin{aligned}\mathbb{P}(|X - \mu_X| \geq k\sigma_X) &= \mathbb{P}((X-\mu_X)^2 \geq k^2\sigma_X^2) \\ &\leq \frac{\mathbb{E}(X-\mu_X)^2}{k^2\sigma_X^2} = \frac{1}{k^2}.\end{aligned}$$

As a quick note, Markov's inequality only applied to positive random variables but only needed that the random variable have a finite mean, Chebyshev's inequality requires a finite mean and a finite variance. More generally, if we can bound higher moments (or the MGF) of a random variable we can get even tighter inequalities.

There are two aspects to Chebyshev's inequality that are worth noting: one of them is the typical deviation which is on the order of σ_X and the probability $\frac{1}{k^2}$. In general, by assuming more things we can improve the probability part but the deviation term will always be on the order of σ_X. For instance, you might here the terminology *exponential* concentration inequalities: these will use more refined techniques to improve the probability to be something like $\exp(-k^2)$.

5 Exponential Concentration Inequalities

5.1 Mill's inequality

Suppose that $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. An important property of Gaussians is that linear combinations of Gaussians are also Gaussian distributed so the sample mean:

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is also normally distributed. In particular, $Y_n \sim N(\mu, \sigma^2/n)$.

Exercise: Prove the above fact. The standard way is to use moment generating functions. Particularly, the moment generating function of any X_i is given by

$$M_X(t) = \exp(\mu t + \sigma^2 t^2/2).$$

Use this and independence to prove the required result.

Mill's inequality is an inequality about standard normals, i.e. $N(0,1)$ random variables. We can use this and some basic properties about Gaussians to get inequalities for general Gaussian random variables.

Mill's inequality: Let $Z \sim N(0,1)$ then:

$$\mathbb{P}(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{\exp(-t^2/2)}{t} \leq \frac{\exp(-t^2/2)}{t}.$$

Once again notice that the deviation is on the order of the standard deviation but the probability is exponential. This is much sharper than what we would get from Chebyshev's inequality. How do we use this for the average we are interested in?

First observe that if $Y_n \sim N(\mu, \sigma^2/n)$ then

$$Z := \frac{Y - \mu}{\sigma/\sqrt{n}},$$

has $N(0,1)$ distribution. We can apply Mill's inequality to Z and do some simple algebra. This gives us that:

$$\mathbb{P}\left(|Y_n| > \frac{t\sigma}{\sqrt{n}}\right) \leq \sqrt{\frac{2}{\pi}} \frac{\exp(-t^2/2)}{t}.$$

To compare, applying Chebyshev's inequality would give us:

$$\mathbb{P}\left(|Y_n| > \frac{t\sigma}{\sqrt{n}}\right) \leq \frac{1}{t^2},$$

which is much weaker for large values of t.

Proof: First we need the standard Gaussian density:

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

So that,

$$\begin{aligned}
\mathbb{P}(|Z| > t) &= \frac{2}{\sqrt{2\pi}} \int_t^\infty \exp\left(-\frac{x^2}{2}\right) dx \\
&= \sqrt{\frac{2}{\pi}} \frac{1}{t} \int_t^\infty t \exp\left(-\frac{x^2}{2}\right) dx \\
&\leq \sqrt{\frac{2}{\pi}} \frac{1}{t} \int_t^\infty x \exp\left(-\frac{x^2}{2}\right) dx \\
&= \sqrt{\frac{2}{\pi}} \frac{1}{t} \int_{t^2/2}^\infty \exp(-u) du,
\end{aligned}$$

by taking $u = x^2/2$. Integrating gives:

$$\mathbb{P}(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{\exp(-t^2/2)}{t},$$

as desired.

5.2 Hoeffding's inequality

The main drawback is that Mill's inequality only applies to Gaussian random variables. Another commonly useful exponential concentration applies to bounded random variables. This is called Hoeffding's inequality.

Hoeffding's inequality: Suppose that $X_1, \ldots, X_n$ are independent and that, $a_i \leq X_i \leq b_i$, and $\mathbb{E}[X_i] = 0$. Then for any $t > 0$,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

When $b_i = b$ and $a_i = a$ this becomes

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

We also have

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right).$$

We will not prove this one, but Wasserman's book has a proof if you are curious.

Hoeffding's inequality looks a bit different from the other inequalities we have seen today, but let us rearrange it a bit. Equivalently,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \geq t\sqrt{\frac{\sum_{i=1}^{n}(b_i-a_i)^2}{n^2}}\right) \leq 2\exp\left(-2t^2\right).$$

This is more like the earlier inequalities, but notice that we don't really have the standard deviation any more. This is really because if $a_i \leq X_i \leq b_i$ then $\mathrm{Var}(X_i) \leq (b_i - a_i)^2$.

Exercise: Prove the above fact.

So that:

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \leq \frac{\sum_{i=1}^{n}(b_i-a_i)^2}{n^2},$$

and this will allow us to interpret Hoeffding's inequality in a more familiar way. Roughly, it says that the probability that the sample average is more than t standard deviations from its expectation is at most $\exp(-2t^2)$.

Let us now use Hoeffding's inequality in our case study example of coin tosses. There each random variable is between -1 and 1 so we have that by Hoeffding's inequality:

$$\mathbb{P}\left(|Y_n| \geq t\right) \leq 2\exp(-2nt^2).$$

Observe once again this inequality is similar to Chebyshev's inequality on the left hand side, i.e. the deviation is on the order of $1/\sqrt{n}$ but the right hand side is much smaller $\exp(-t^2)$ instead of $1/t^2$.