LECTURE: CHEBYSHEV'S INEQUALITY

1. Chebyshev's Inequality

Markov's Inequality:

If $Y \geq 0$ and $a > 0$ is given, then

$$P(Y \geq a) \leq \frac{E(Y)}{a}$$

This estimates $P(Y \geq a)$ assuming nothing about Y except for its mean $E(Y)$. Intuitively, this says that it is unlikely that Y takes on large values.

IF you happen to also know $\text{Var}(Y)$, you get a better inequality, called

Chebyshev's Inequality:

If Y is **any** random variable and $a > 0$, then

$$P(|Y - E(Y)| \geq a) \leq \frac{\text{Var}(Y)}{a^2}$$

This says that if the variance of Y is small, then it is unlikely that Y is far from its average $E(Y)$. For example, if on an exam the average is 50 and the variance is 10, then it's not likely that you have a 90. But if the variance is 40, then it's more likely that you have a 90.

Why?

Use Markov with $Z = (Y - E(Y))^2 \geq 0$ and a^2 instead of a to get

$$P([Y - E(Y)]^2 \geq a^2) \overset{\text{DEF}}{=} P(Z \geq a^2) \leq \frac{E(Z)}{a^2} \overset{\text{DEF}}{=} \frac{E[(Y - E(Y))^2]}{a^2} = \frac{\text{Var}(Y)}{a^2}$$

$$P(|Y - E(Y)| \geq a) = P([Y - E(Y)]^2 \geq a^2) \leq \frac{\text{Var}(Y)}{a^2} \checkmark$$

Example 1:

Suppose we randomly select an article from a journal article whose length is distributed with a mean of 1000 words and a standard deviation of 150 words. Find an upper bound on the probability that an article is outside of the range 600-1400 words

Let Y be the length of an article in this journal.

Here we know both the mean and the standard deviation of Y, so we can use Chebyshev's Inequality.

$$P((Y \geq 1400) \cup (Y \leq 600)) = P(|Y - 1000| \geq 400) \leq \frac{\text{Var}(Y)}{400^2} = \frac{150^2}{400^2} \approx 0.14$$

Note: In particular, this implies

$$P(Y \geq 1400) \leq P((Y \geq 1400) \cup (Y \leq 600)) \leq 0.14$$

This is a much better bound than we got using Markov's Inequality, which was 0.74

Note: **IF** the distribution of Y is symmetric about the mean (think normal) we can divide $P((Y \geq 1400) \cup (Y \leq 600))$ by 2 to get:

$$P(Y \geq 1400) \leq \frac{P((Y \geq 1400) \cup (Y \leq 600))}{2} \leq \frac{0.14}{2} = 0.07$$

Which is even better! This is **NOT** true in general, only if you have symmetry!

Comparison: For comparison purposes, let's see how much better this bound is if we knew the exact distribution of Y

Suppose $Y \sim \mathrm{N}(1000, 150)$. Then, using shifting,

$$\begin{aligned} P(600 \leq Y \leq 1400) &= P\left(\frac{600-1000}{150} \leq Z \leq \frac{1400-1000}{150}\right) \\ &= P(-2.67 \leq Z \leq 2.67) \\ &= F(2.67) - F(-2.67) \\ &= 0.9962 - 0.0038 = 0.9924 \end{aligned}$$

Hence $P((Y \geq 1400) \cup (Y \leq 600)) = 1 - P(600 \leq Y \leq 1400) = 0.0076$

Which is waaaay better than the 0.14 and 0.07 bounds we got before!

Moral: Although Chebyshev's Inequality gives a decent bound on the probability of outliers, there is no substitute for knowing the actual probability distribution!

Standard Deviations: Sometimes we like to measure deviation from the mean in terms of "numbers of standard deviations". For the normal distribution, this is encapsulated in the 68-95-99 rule. We can state Chebyshev's Inequality in these terms if we like:

Fact:

Let Y be a random variable with mean μ and variance σ^2.

Then the probability of deviating at least k standard deviations from the mean is bounded by

$$P(|Y - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

Why? Take $a = k\sigma$ in Chebyshev's Inequality above, then

$$P(|Y - E(Y)| \geq k\sigma) \leq \frac{\operatorname{Var}(Y)}{(k\sigma)^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2}$$

2. Multivariate Distributions

In practice, we are often interested in the distribution of many quantities at the same time, such as the height **and** the weight of chimpanzees.

Since the distribution involves several quantities, we call it a **multivariate distribution**. One question, for instance, might be whether or not these quantities are independent. Here they probably aren't, since taller chimpanzees usually weigh more.

In this section, we will primarily be interested in **bivariate distributions**, that is the probability distribution of **two** random variables.

As before, we start with the discrete case and then consider the continuous case. First, let's define the joint probability distribution for a pair of discrete random variables:

Definition:

If Y_1 and Y_2 are two discrete random variables, then the **joint distribution** of Y_1 and Y_2 (joint pmf) is

$$p(y_1, y_2) = P(Y_1 = y_1, Y_2 = y_2) \quad \text{for all possible pairs } (y_1, y_2)$$

Here $P(Y_1 = y_1, Y_2 = y_2)$ is short-hand for $P((Y_1 = y_1) \cap (Y_2 = y_2))$

Example 2:

Suppose you roll two dice. Let X_1 be the roll of the first die and X_2 the roll of the second die.

Then the joint distribution of X_1 and X_2 is given by:

$$p(x_1, x_2) = \frac{1}{36}$$

Where $x_1 = 1, 2, 3, 4, 5, 6$ and $x_2 = 1, 2, 3, 4, 5, 6$

Just as in the case for a single discrete random variable, all the possible probabilities are non-negative and they sum to 1.

Fact:

Let Y_1 and Y_2 be discrete random variables with joint pmf $p(y_1, y_2)$. Then

$$0 \leq p(y_1, y_2) \leq 1 \text{ for all } y_1, y_2$$
$$\sum_{\text{all } (y_1, y_2)} p(y_1, y_2) = 1$$

Just like for single discrete random variables, we can construct a joint probability distribution by assigning probabilities that add up to 1:

Example 3:

Suppose you survey undergraduates and ask them 2 questions:

(1) Do you have an exam this week?

(2) How many cups of coffee did you drink today?

Let X_1 be the discrete random variable with values $\{\texttt{yes}, \texttt{no}\}$ indicating whether or not a student has an exam this week.

Let X_2 be the number of cups of coffee a student drank today. For simplicity, we will let X_2 take only the values $\{0, 1, 2\}$

We can display the joint probability distribution for the pair (X_1, X_2) in a 2×3 table. We can choose any probabilities for the six pairs as long as they sum to 1. One possible choice is shown in the table below.

		X_2: 0	1	2
	`yes`	2/20	3/20	3/20
X_1	`no`	6/20	4/20	2/20

You can check that the sum of the entries in the table is indeed 1

3. Marginal Distribution

Consider again a joint distribution (Y_1, Y_2) of two discrete random variables with pmf $p(y_1, y_2)$ like the exam-coffee example above.

Y_1 and Y_2 are themselves discrete random variables. What are their distributions?

Suppose we wish to find the distribution for Y_1. To do that, just sum over all the possible values of Y_2

Definition:

If Y_1 and Y_2 are discrete random variables with joint pmf $p(y_1, y_2)$

The **marginal distribution** of Y_1 is given by

$$p_1(y_1) = \sum_{\text{all } y_2} p(y_1, y_2)$$

And the marginal distribution of Y_2 is given by

$$p_2(y_2) = \sum_{\text{all } y_1} p(y_1, y_2)$$

Here we just sum over all the possibilities of the *other* random variable.

Example 4:

In the exam-coffee example above, calculate the marginal distributions for X_1 and X_2

In this case, for the marginal distribution of X_2, we sum the values in each column. Then the bottom row called "Total" is the marginal distribution of X_2. Similarly, we can find the marginal distribution for X_1 by summing each row. Then the right column also labeled "Total" is the marginal distribution for X_1.

		X_2 = 0	1	2	Total
X_1	`yes`	2/20	3/20	3/20	8/20
	`no`	6/20	4/20	2/20	12/20
	Total	8/20	7/20	5/20	

Note: In fact, the marginal distribution is called "marginal" because its values lie in the margins of the joint distribution table.

You can check that the two marginal distributions sum to 1 and are thus valid probability distributions for discrete random variables.