Weekly Review 3

This week we had three lectures, during which we discussed conditional probability and finished Chapter 4, and then started Chapter 5, which introduces a very important notion called *random variable*. We defined the expression like $\Pr(X = x)$ and explained the idea of the mean of a random variable. In the example class we solved some problems in counting.

From the definition of conditional probability, the famous and important *Bayes' Theorem* was introduced and applied to the following clinical situation. For a screening test of a certain disease, let D denote the event that a person has the disease, D^{c} the event that a person does not have the disease, T the event that the screening test shows a positive result and T^{c} the event that the screening test result is negative. If we know the values of

$$\begin{aligned}\Pr(T|D) &= \textit{sensitivity},\\ \Pr(T^{\mathrm{c}}|D^{\mathrm{c}}) &= \textit{specificity},\\ \Pr(D) &= \textit{rate (prevalence) of the disease in the relevant population},\end{aligned}$$

then by Bayes' theorem (which we derived straightforwardly from the definition of conditional probability), we can obtain the values of

$$\textit{predictive value positive} = \Pr(D|T) = \frac{\Pr(T|D)\Pr(D)}{\Pr(T|D)\Pr(D) + \Pr(T|D^{\mathrm{c}})\Pr(D^{\mathrm{c}})},$$

$$\textit{predictive value negative} = \Pr(D^{\mathrm{c}}|T^{\mathrm{c}}) = \frac{\Pr(T^{\mathrm{c}}|D^{\mathrm{c}})\Pr(D^{\mathrm{c}})}{\Pr(T^{\mathrm{c}}|D^{\mathrm{c}})\Pr(D^{\mathrm{c}}) + \Pr(T^{\mathrm{c}}|D)\Pr(D)}.$$

The *predictive value positive* is also known as the *predictive value of a positive test* and the *positive predictive value*. *Tree diagrams* (which I used to explain in class why, in a group of say 100,000 persons, for a disease of prevalence 1/1000, the predictive value positive is only about 2% even when the sensitivity and specificity are 99% and 95% respectively) are quite helpful for our calculations when we apply Bayes' theorem.

A more general statement of *Bayes' theorem* is

$$\Pr(B_i|A) = \frac{\Pr(A|B_i)\Pr(B_i)}{\Pr(A|B_1)\Pr(B_1) + \cdots + \Pr(A|B_n)\Pr(B_n)},$$

where $B_i \cap B_j = \emptyset$ whenever $i \neq j$ and $\Pr(B_1 \cup \cdots \cup B_n) = 1$. The difference between this general statement and the one we used for calculating the predictive value positive is in the denominator on the right-hand side. The whole sample space Ω is now divided not only into two categories (e.g. disease and not-disease) but into $n \geq 2$ mutually exclusive (i.e. non-overlapping) categories. To apply the general statement, we have to know all the conditional probabilities $\Pr(A|B_i)$ (c.f. sensitivity and specificity) and the proportions $\Pr(B_i)$ of individual B_i (c.f. the rate of disease).

We also discussed the famous Monty Hall problem, which is a puzzle involving (conditional) probability loosely based on the American game show *Let's Make a Deal*. The name comes from the show's host, Monty Hall. A widely known statement of the problem appeared in a letter to Marilyn vos Savant's *Ask Marilyn* column in *Parade* (Marilyn vos Savant is an American magazine columnist, author, lecturer and playwright who rose to fame through her listing in the *Guinness Book of World Records* under "Highest IQ". Since 1986 she has written *Ask Marilyn*, a Sunday column in *Parade* magazine, distributed as a Sunday supplement in hundreds of newspapers in the United States, in which she answers questions from readers on a variety of subjects):

> `Suppose you are on a game show, and you are given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say Number 1, and the host, who knows what is behind the doors, opens another door, say Number 3, which has a goat. He then says to you, ''Do you want to pick door Number 2?'' Is it to your advantage to switch your choice?`

If you want to learn more about this Monty Hall problem, you may read the book written by Jason Rosenhouse (2009, *The Monty Hall Problem. The Remarkable story of Math's Most Contentious Brain Teaser*, Oxford University Press), which is available in our library.

This concludes our discussion on Chapter 4. (Chapter 4 in the book does not cover everything we have discussed. Thus, we can be proud to say that we are studying at a more advanced level than the textbook!)

Chapter 5 is devoted to some more advanced notions in probability. As a motivating example, we consider a game with three dice in a casino. For example, using \$1 to bet, if you win \$30 (the actual betting odds depends on which casino you are visiting, in many casinos in Macau you can win, I believe, only \$24) whenever the outcome is a triple (the outcomes of the three dice are the same) but lose \$1 whenever it is not, in the long term, your gain *per game* is

$$\frac{1}{36} \times 30 + \frac{35}{36} \times (-1) = -\frac{5}{36},$$

i.e. on the average you lose \$5/36 in each game. On the other hand, betting on *Big* (some casinos call it *Large*), you win \$1 whenever the sum of the outcome is eleven or above and is not a triple but lose \$1 whenever it is less than eleven or it is a triple, in the long term your gain per game is

$$\frac{35}{72} \times 1 + \frac{35}{72} \times (-1) + \frac{1}{36} \times (-1) = -\frac{1}{36}.$$

Thus, if you have to bet either on triples or *Big*'s repeatedly, even you can win more by betting on triples, it is more sensible to bet on *Big*'s, because in the long term your loss will then be smaller. Of course, since your long-term gain is always negative when playing games of chance in a casino, it is even more sensible not to gamble, unless you are the host of casino! (There is an advanced book entitled *Inequalities for Stochastic Processes: How to Gamble if You Must*, but do not think that it could tell you how to get rich in a casino.)

Now, if we do not restrict ourselves to gambling situation, does the above calculation have any meaning in general? An abstraction and generalisation of the above discussion will lead us to the notions such as *random variables* and their *probability distributions* and *expectations*.

The return of your investment in a game sometimes is positive (when we win) and sometimes is negative (when we lose) and it is an example of what we call a random variable.

Consider another random variable in the dice game: let X denote the sum of the outcomes of three dice, what is the meaning (not the numerical value) of $\Pr(X = 10)$? We know that probability is a function defined for events, and events are subsets of the sample space Ω. Thus, we know what the meaning of $\Pr(A)$ for $A \subset \Omega$ is. However, what is the meaning of the probability of an equality?

As I said, the notion of random variables is the most fundamental concept in advanced probability theory. Unfortunately it is a misleading name because, as I have explained, a random variable is neither random nor a variable but a function defined for individual outcomes, also called *sample points*, in the sample space. (Just like an *alligator pear* is neither an alligator nor a pear, but an avocado!) Outcomes of the experiment are random, but a random variable is something like the hat of a magician and its job is that you input an outcome into it, and it will transform (in mathematical terms, it will *map*) your input to a number. For the same input, we will definitely get the same number. Nothing random! However, *most beginners find it easier to think of random variables simply as quantities that can take on different values depending on chance*. (That is the interpretation adopted in our textbook. Again, we can be proud to say that we are more advanced than our textbook!)

For a random variable X, we can talk about its *distribution*, i.e. $\Pr(X = x)$. Such a notation $\Pr(X = x)$, although intuitively clear, is not a proper notation because $\Pr(\cdot)$ is function defined for events (subsets of the sample space) and $X = x$ is not an event! Thus, we should keep in mind that X is in fact a function defined for individual outcomes in the sample space Ω so that for an outcome ω we get a corresponding value $X(\omega)$ depending on what ω is, and

$$\Pr(X = x) = \Pr(\{\omega \in \Omega : X(\omega) = x\}).$$

The event on the right-hand side refers to the collection of outcomes (as we have already seen before, a collection is denoted by $\{\cdots\}$) that each outcome ω, which belongs to Ω (the symbol $\in$ means "belongs to" or "in"), satisfies the requirement that $X(\omega) = x$. It is clear from the definition that the uppercase X is the random variable of interest and the lowercase x is just an unspecified value, which can be replaced by k, y, a, b, 4, -1.23 or any other symbols.

The *mean of a distribution* or the *mean of a random variable* is the same as the *expected value* or *expectation*. More precisely, the mean μ of a random variable X is defined by

$$\mu_X \equiv \mathbb{E}(X) \equiv \mathbf{E}(X) := \sum_{\text{all possible } x} x \Pr(X = x).$$

If no confusion is possible, we may drop the subscript and simply write μ instead of μ_X. This is a very important and fundamental concept in probability theory.

If I roll a die, I win \$$x$ when the outcome is x, then my *expected return* is

$$1 \times \Pr(X = 1) + 2 \times \Pr(X = 2) + \cdots + 6 \times \Pr(X = 6) = 3.5.$$

The correct interpretation of the value 3.5 in this context is that if repeating the experiment infinitely many times (remind you that the interpretation of probability requires repeating the experiment infinitely many times), then on the average I will obtain \$3.5 in each experiment. Note that although the mean of X is also called the expected value of X, it does NOT imply that we expect to win \$3.5 in any single experiment. (In this example, the expected value 3.5 will never be seen in any single experiment, which is rolling a die!) The expected value is just the long term (or theoretical) average.

We then defined the *variance* σ_X^2 *of a random variable* X by

$$\mathbf{var}(X) \equiv \sigma_X^2 := \mathbf{E}\{(X-\mu)^2\} = \sum (x-\mu)^2 \Pr(X = x).$$

Again, if no confusion is possible, we drop the subscript and write σ^2 instead of σ_X^2. The variance σ^2, similar to the mean μ of X, is the variance of the many data that can be (but not necessarily have been) generated repeatedly by X.

A computationally simpler version of the above formula (try to work it out by yourselves, but it does not matter if you do not understand it at all for this course) is

$$\begin{aligned}\sigma^2 &= \sum (x-\mu)^2 \Pr(X = x) = \sum x^2 \Pr(X = x) - 2\mu \sum x \Pr(X = x) + \mu^2 \sum \Pr(X = x) \\ &= \sum x^2 \Pr(X = x) - \mu^2 = \mathbf{E}(X^2) - \mu^2.\end{aligned}$$

Then we considered important examples of distributions that arise in various applications. The first example comes from a generalisation of rolling a fair die: if $\Pr(X = x)$ is a constant for every possible x and zero otherwise, we say that X follows the *uniform distribution*. It arises whenever all possible outcomes are equally likely, e.g. choosing a number randomly between 1, 2, ..., 10. However, life is not always so simple.

Suppose we perform an experiment that consists of a sequence of trials, and these trials satisfy the following three criteria.

1. Each trial results in either a *success* or a *failure* (where *success* is just a generic term and is not necessarily something favourable in reality).
2. The probability of a *success* is p for each trial.
3. The trials are independent.

Let X be the number of successes out of n such trials (which are called *Bernoulli trials*). Using an elementary argument, we have derived in class that

$$\Pr(X = x) = \begin{cases} \dbinom{n}{x} p^x (1-p)^{n-x}, & x = 0, \ldots, n, \\ \\ 0, & \text{otherwise}, \end{cases}$$

and we say X follows (or has) a *binomial distribution* and write $X \sim \mathrm{B}(n, p)$. (Note that I will omit "$= 0$, otherwise" hereafter, but you have to keep this in mind and have to write it down explicitly, whenever and wherever applicable, in the mid-term and the final exam.)

Using the fact (which is an application of the *binomial theorem*) that

$$\sum_{k=0}^{m} \binom{m}{k} p^k (1-p)^{m-k} = \{p + (1-p)\}^m = 1,$$

we derived that the mean of X, where $X \sim \mathrm{B}(n, p)$, is

$$\mathbf{E}(X) = \mu = np,$$

which agrees with our intuitive interpretation of the mean. We also got the variance of this binomial X:

$$\mathbf{var}(X) = \sigma^2 = np(1-p),$$

by considering $\sum x^2 \Pr(X = x) = \sum x(x - 1 + 1) \Pr(X = x)$ in some intermediate steps in the derivation. I hope you have found the technique inspiring.

Although the mathematical techniques that I demonstrated in recent lectures are not needed for the understanding of the second (and more important) half of this course, they are themselves basic knowledge and fundamental techniques in probability and mathematical statistics (and would be important for MATH2216 Statistical Methods and Theory and other statistics courses).

Next week we will discuss some important distributions (which result from some frequently encountered situations). We will be able to finish Chapter 5 and perhaps even be able to start Chapter 6.

Wish you a very Happy New Year of Rooster!

Cheers,
Heng Peng