Weekly Review 5

This week we had three hours. We discussed in details the normal distribution and concluded our discussion on probability by building a bridge between the world of probability and the realm of statistics, namely the central limit theorem, presented in Chapter 7. The example class shows some binomial and Poisson examples and you can see that to make such questions a bit more challenging, it is typical that the probability p in binomial or the parameter λ in Poisson is not given explicitly but is a value that you have to calculate from the information given in the question.

Suppose $X \sim \mathrm{N}(\mu, \sigma^2)$. The probability density function of X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\,\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for all } x.$$

To get the probability $\Pr(X \le x)$, we have to find the area under the curve of $f(x)$.

For those who know calculus, the area is equal to the integral

$$\Pr(X \le x) = \int_{-\infty}^{x} f(u)\,\mathrm{d}u.$$

If you try to do the integration, you will soon find that there is no closed form solution (i.e. no nice mathematical formula can be the solution).

If $X \sim \mathrm{N}(\mu, \sigma^2)$ and we want to calculate $\Pr(X \le x)$, what we have to do is to consider another random variable $Z \sim \mathrm{N}(0, 1)$ and then

$$\Pr(X \le x) = \Pr\left(Z \le \frac{x-\mu}{\sigma}\right).$$

Therefore, only one table is sufficient for all normal distributions and this important special normal distribution $\mathrm{N}(0, 1)$ is called the *standard normal distribution*.

Now, we can see that the calculation of probability for normal distribution is straightforward and routine. Let me describe it as follows. Suppose $X \sim \mathrm{N}(\mu, \sigma^2)$, we calculate $\Pr(X \le x)$ mechanically by the following steps:

1. calculate $z = (x - \mu)/\sigma$ (correct to two decimal places);
2. turn to Table IV (the very first piece of paper of the textbook) and check the entry corresponding to the numerical value of z.

For example, consider $z = 2.74$. First, find the row starting with 2.7 and then find the column with heading .04, we get 0.9969. Please, do **not** write $2.74 = 0.9969$; it is ridiculously absurd! The figure on the top right corner of the table tells you that the value 0.9969 is the area under the curve to the left of $z = 2.74$, and so a proper way to relate these two values is:

$$\Pr(Z \leq 2.74) = 0.9969.$$

Another example, consider $z = -1.37$. First, find the row starting with -1.3, and then find the column with heading 0.07, we get 0.0853.

However, the page containing the table for negative z is in fact not necessary and will **not** be provided in the examination. Consider, say, $z = 1.37$, we get

$$\Pr(Z \leq 1.37) = 0.9147,$$

and because of symmetry,

$$\Pr(Z \leq -1.37) = \Pr(Z \geq 1.37) = 1 - \Pr(Z \leq 1.37) = 1 - 0.9147 = 0.0853.$$

I would advice you to sketch the curve of the normal distribution and indicate the area you want to calculate first. Then get the value from the table for positive z and see how you can get the answer you want from this value. The fundamental relations are:

$$\begin{aligned}
\Pr(a \leq Z \leq b) &= \Pr(Z \leq b) - \Pr(Z \leq a), \\
\Pr(Z \geq a) &= \Pr(Z > a) = 1 - \Pr(Z \leq a) = \Pr(Z \leq -a), \\
\Pr(Z \geq -a) &= \Pr(Z \leq a), \\
\Pr(Z \geq 0) &= \Pr(Z \leq 0) = 0.5.
\end{aligned}$$

However, do not try to memorise them by heart. Always sketch the curve first and then do the corresponding arithmetics.

You **must** know how to use the normal table. In the exam you must use the tables provided, in which **I may intentionally change the entries a little bit** to check whether you are really using the tables provided or you just have a super-calculator with you. Marks will be deducted if you are not using the tables provided.

Sometimes we are interested in getting a cutoff value, say a, where the area corresponding to $\Pr(Z \leq a)$ is some pre-specified value given in the question. For example, to find the cutoff value for the top 2.5%, we have to check what value of a will give $\Pr(Z \leq a) = 0.975$, and from the table we get $a = 1.96$.

One interesting application of the normal distribution is that it can be used to approximate the binomial distribution when n is large and p close to 0.5. (Question: Why do we want to do so?) Practically, we may use $Y \sim \mathrm{N}\big(np, np(1-p)\big)$ to approximate $X \sim \mathrm{B}(n,p)$ when $np > 5$ and $n(1-p) > 5$. Please note that the *continuity correction* is important when we approximate a discrete distribution by a continuous distribution. More precisely, because X must be integers but Y is not, we have e.g.

$$\Pr(X \le 4) = \Pr(X \le 4.5) \approx \Pr(Y \le 4.5), \qquad \Pr(X < 4) = \Pr(X < 3.5) \approx \Pr(Y < 3.5),$$
$$\Pr(X \ge 4) = \Pr(X \ge 3.5) \approx \Pr(Y \ge 3.5), \qquad \Pr(X > 4) = \Pr(X > 4.5) \approx \Pr(Y > 4.5).$$

To get e.g. $\Pr(Y \le 4.5)$, you do exactly the same as before, i.e.

$$\Pr(Y \le 4.5) = \Pr\left(Z \le \frac{4.5 - np}{\sqrt{np(1-p)}}\right),$$

where Z has the standard normal distribution $\mathrm{N}(0,1)$. Be careful: the denominator is $\sqrt{np(1-p)}$, because $\sigma = \sqrt{np(1-p)}$. Forgetting the square root sign is a common mistake among careless students (and professors).

I mentioned (but did not work out any example) that the normal distribution can also be used to approximate Poisson when λ is very large, because Poisson itself can be used to approximate binomial for large n (and small p). Of course the continuity correction is still needed because Poisson is also discrete.

However, be careful! If the question is talking about a normal random variable itself, then you should not make the ± 0.5 continuity correction; the continuity correction is needed when and only when you are using a continuous distribution (such as the normal) to approximate a discrete distribution (such as the binomial or the Poisson). The purpose of this correction is to make a discrete distribution closer to a continuous distribution so that the approximation has a smaller error, and that is exactly the reason why this ± 0.5 adjustment is called the *continuity* correction. When a question concerns not a discrete but a continuous distribution, then we are not approximating any discrete distribution, and of course making any continuity correction is inappropriate.

That is the end of Chapter 6. Chapter 7 builds a bridge from probability to statistics.

I have repeatedly mentioned the following in the early weeks of this semester. Suppose we have a very large population and we want to know some important *parameters* such as its mean μ and its variance σ^2. However, in real life situations, because of limited resources and limited time, we are not able to study each individual of the very large population. What we can do is to take a *random sample of size* n and then study each individual in this sample. From the information obtained from this sample, we want to estimate the unknown population parameters μ and σ^2.

The first parameter we studied is the population mean μ. Naturally, if we have a random sample $\{X_1, \ldots, X_n\}$, then we can calculate the sample mean $\overline{X}$, which is an example of what we called a *statistic* (generally, a *statistic* is a number calculated from a sample, e.g. sample mean, sample median, sample variance, sample maximum, etc.). Then we use $\overline{X}$ to estimate μ. Therefore, we say that $\overline{X}$ is an *estimator* of μ. If the sample mean of a sample is, say, 151, and we simply say that my *estimate* of μ is 151. That is to say, the estimator $\overline{X}$ is a random variable, whilst a particular realisation of it, which is a number, generically denoted by the lower case $\overline{x}$, is an estimate. However, it is not clear how accurate and how precise an estimate is. Thus, we would like to tell the reader the possible size of error. When we say "possible size", obviously it has something to do with probability. To calculate probability related to $\overline{X}$, we have to know something about $\overline{X}$. What is that 'something'? We have to know its distribution. It has a distribution because the sample is taken randomly and we can take samples repeatedly. If I take another sample, I may get a different sample mean; and if I take 100 samples, I will have 100 sample means, and they are not necessarily the same value. The distribution of a statistic, such as the sample mean, is generally called a *sampling distribution*.

Since samples are random, the sample mean $\overline{X}$ is a random variable, and consequently we can talk about its distribution and in particular, its mean and its variance. It happens that the mean $\mu_{\overline{X}}$ and the variance $\sigma^2_{\overline{X}}$ of the sample mean $\overline{X}$ of a random sample of size n are related to the population mean μ and the population variance σ^2 in a very nice way:

$$\mu_{\overline{X}} = \mu, \tag{1}$$

$$\sigma^2_{\overline{X}} = \begin{cases} \dfrac{\sigma^2}{n} & \text{for an infinite population,} \\[2ex] \dfrac{\sigma^2}{n}\dfrac{N-n}{N-1} & \text{for a finite population of size } N, \end{cases} \tag{2}$$

where the standard deviation of $\overline{X}$, denoted by $\sigma_{\overline{X}}$, is also (in fact, more commonly) called the *standard error of the mean*. Though in practical applications we often have populations of finite sizes, to make life easier (well well, easier $\neq$ easy), in this course we only consider an infinite population and so $N = \infty$; consequently, the standard error is:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}.$$

Note that I did not explain mathematically why equations (1) and (2) hold, but our intuition suggests that we would be happy if they are true, and I simply told you without any proofs that, luckily, they are really true.

In particular, if $n = 1$, i.e. we just randomly take one number, denoted by X_1, from the population, the sample mean $\overline{X}$ of this sample is of course the same as X_1, whose mean is μ and variance is $\sigma^2/1 = \sigma^2$, which are the same as the population mean and the population variance. Taking a sample of size n can be interpreted as taking n samples, each of size 1, independently, and then pooling all these together to form one sample of size n. That is to say, in the sample $\{X_1, \ldots, X_n\}$, each individual X_i (no matter what the value i is) plays an equally important role as the others do; each X_i has the same mean μ and the same variance σ^2, and in fact, all X_1, ..., X_n have the same distribution, because each of them is sampled from the same population in exactly the same way. We usually say that they are *independent and identically distributed*, or *i.i.d.* for short.

Knowing the mean and the variance is not enough; we have to know its sampling distribution. Section 7.3 states the most important theorem in Statistics, the *central limit theorem*, which says that

Central Limit Theorem *The sample mean $\overline{X}$ follows approximately the normal distribution with mean $\mu_{\overline{X}} = \mu$ and variance $\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}$ if n is large, i.e.*

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

has approximately the standard normal distribution when n is large, no matter what the distribution of X_i is. However, in particular, if the population itself has a normal distribution, then $\overline{X}$ follows exactly, no matter whether n is large or small, the normal distribution.

(Often, people will consider that when $n \geq 30$, then n is large enough to guarantee a reasonably good approximation by the normal distribution.)

The central limit theorem explains why we have to study the normal distribution before we have a deeper discussion of statistical inference. This is the only topic in Chapter 7 that I would like you to understand and also is the end of our appetizer.

Now, we are ready to enjoy the main course of our dinner, namely, *statistical inference*, which includes *estimation* and *hypothesis testing*.

The first estimation problem addressed is the estimation of the population mean μ by the sample mean $\overline{X}$. Clearly, because $\overline{X}$ follows the normal distribution, $\Pr(\overline{X} = \mu) = 0$. Therefore, it is more sensible to talk about errors. Perhaps we would like to know, for a given value of acceptable error, say a, what is the probability $\Pr(|\overline{X} - \mu| \leq a)$? This is a simple question on probabilities of a normal distribution, is not particularly exciting here. In applications, when we do not know what μ is, a more natural question is: how large is the error? However, since $\overline{X}$ is normally distributed, mathematically speaking, any value is possible. We only know that extremely large or extremely small values are unlikely but still possible. Therefore, if we do not know what μ is and if we want to use $\overline{X}$ to estimate this unknown μ, it is more useful if you can tell me something like: with say 95% chance the mean μ is some value between $\overline{X} \pm a$.

Next week we will see how to get the value a above. We will then consider the estimation of the parameter p, denoting the proportion of successes in a population containing solely successes and failures (e.g. males and females, yes and no, or pass and fail, etc.).

Cheers,
Heng Peng