Lecture 25

Wednesday, November 13, 2024 1:43 PM

You may ask why we always talk about the normal distribution. How about other kind of distribution? Is the normal distribution well-loved simply because it makes computations easier compared to other distributions? Or does it have any true significance? The answer to this question lies in the *Central Limit Theorem*.

Suppose you want to find out the proportion of a population that has a certain property A. For example, the unemployment rate among all American, the percentage of married students among all college students, the percentage of people who own more than one house, the death rate among all newborn babies in the world. Let p be this proportion, which is to be found. It is a population parameter.

You sample n individuals from the population. To each individual, you assign a value 1 if it has property A, and 0 if it doesn't. In other words, you get a list of numbers $X_1, X_2, \dots, X_n$, each equal to 0 or 1, with the understanding that $X_k = 1$ if the individual k has property A, and $X_k = 0$ if it doesn't. The proportion of individuals in the sample that has property A is

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Assume that you select the individuals randomly and put an individual back to the pool after you have examined it. That way, each time you examine an individual, you have the same chance of get 1 or 0. Therefore, the "random variables" $X_1, X_2, \dots, X_n$ have exactly the same distribution.

The Central Limit Theorem says that for large n, the random variable $\bar{X}_n$ is approximately normally distributed with mean EX_1 and standard deviation $\sigma = \frac{\sigma_{X_1}}{\sqrt{n}}$. This is surprising because the sample mean $\bar{X}_n$ has a normal distribution (approximately) regardless of the distribution of X_1. Thus, the normal distribution comes into the picture quite naturally. Let's do some more detailed calculation:

$$P(X_k = 1) = p, \;\; P(X_k = 0) = 1 - p.$$

The mean of each X_k is $EX_k = p$ and the variation is

$$Var(X_k) = EX_k^2 - (EX_k)^2 = p - p^2 = p(1-p)$$

The standard deviation of X_k is

$$\sigma_{X_1} = \sqrt{Var(X_k)} = \sqrt{p(1-p)}$$

Thus, the standard deviation of $\bar{X}_n$ is

$$\sigma = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

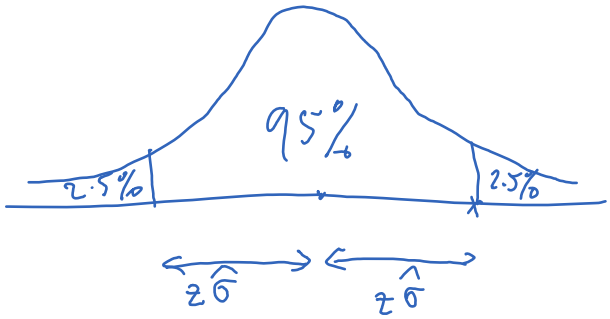
By the 68-95-99.7% rule, there are about 95% of all data $\bar{X}_n$ that stay within two standard deviations from the mean. That means, out of all possible samples of size n, about 95% of them will be in the interval $[p-2\sigma, p+2\sigma]$. Because p is unknown, we will approximate it by $p \approx \hat{p}$, where $\hat{p}$ is the sample proportion (a sample statistics). With this approximation, we also approximate

$$\sigma \approx \hat{\sigma} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The interval $[\hat{p}-2\hat{\sigma},\ \hat{p}+2\hat{\sigma}]$ is called the 95% confidence interval. Approximately 95% of all possible samples of size n have a sample mean lying in this interval. The margin of error of the 95% confidence interval is

$$2\hat{\sigma} = 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

If $\hat{p} \approx 1/2$ then $2\hat{\sigma} \approx \frac{1}{\sqrt{n}}$. This is the formula used in the textbook (Section 6D).
To be a little more precise, the 95% of all data lying in the interval $[\hat{p}-z\hat{\sigma}, \hat{p}+z\hat{\sigma}]$ where z is the z-score of the percentile $95\% + 2.5\% = 97.5\%$ (see the picture below), which is about 1.96 (more precise than 2).



Therefore, a more precise formula for the margin of error of the 95% confidence interval is

$$1.96\hat{\sigma} = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This is a formula commonly used in various textbooks.