

The Central Limit Theorem

August 18, 2009

The Normal Distribution

If X is normally distributed with mean μ and variance σ^2 (we will write this as $X \sim N(\mu, \sigma^2)$), then its probability density function (pdf) is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}.$$

The graph of this function is bell-shaped, with a maximum at $x = \mu$, and an approximate “width” of 4σ . If $X \sim N(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim N(0, 1)$, and the pdf of this *standard* normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

Suppose that $X \sim N(0, 1)$. What is $\mathbb{P}(X \leq x)$, the cumulative distribution function of X ? There is no formula for it, so the values of

$$\mathbb{P}(X \leq x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-t^2/2} dt$$

are listed in tables at the end of most statistics books.

The Central Limit Theorem

The importance of the normal distribution in mathematics and statistics stems from the following theorem. Suppose I have n **independent, identically distributed** random variables X_1, \dots, X_n , each with finite mean μ and finite nonzero variance σ^2 . Suppose also that n is large (at least 30). Then, *regardless of the distribution of each X_i* , the sum

$$S_n = X_1 + \dots + X_n$$

is approximately normally distributed with mean (exactly) $n\mu$ and variance (exactly) $n\sigma^2$. Consequently, the *sample mean*

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

is approximately normally distributed with mean μ and variance σ^2/n .

The Probabilistic Viewpoint

Here are three examples of the use of the central limit theorem to estimate probabilities. Suppose I toss a fair coin 100 times. The mean and variance of the number of heads in a single coin toss are $1/2$ and $1/4$ respectively. Consequently, the total number of heads is approximately normally distributed with mean 50, variance 25, and standard deviation 5. From this information, and a table of values of $\Phi(x)$, I can estimate that I will get between 45 and 55 heads with probability about 0.7, and between 40 and 60 heads with probability about 0.95, without having to sum lots of binomial probabilities.

Next, suppose that I use the random number generator on my calculator to generate 100 random numbers between 0 and 1, each having the uniform distribution. Since the mean and variance of each random number are $1/2$ and $1/12$ respectively, I know that the sum of the numbers is approximately normally distributed with mean 50 and variance $25/3$. With this information, I can answer such questions as: What is the probability that the sum is at least 60?

Finally, suppose buses arrive according to a Poisson process with rate 5 buses/hour. How long do I have to wait for the 100th bus? Although it is possible to calculate the exact probability distribution, it is much simpler to observe that the waiting time is the sum of 100 independent, exponentially distributed random variables, each with mean $1/5$ and variance $1/25$. Consequently, the waiting time to the 100th bus is approximately normally distributed with mean 20 and variance 4. Thus, for instance, with probability about 0.975, I will not have to wait more than 24 hours for the 100th bus.

The Statistical Viewpoint

In statistics, we are often trying to estimate the mean of a certain quantity in a population. For example, we might want to know the mean weight of men aged between 18 and 65 in Wisconsin. Note that it is not at all clear that this quantity is normally distributed - probably it isn't. However, if we weigh 100 randomly chosen men in that age range, then the central limit theorem allows us to assume that the sample mean (i.e. the average of the 100 weights) *is* approximately normally distributed. In this situation, we are treating each weight as a random variable, whose (unknown) pdf has the same shape as the (unknown) histogram of all the weights of all the men aged 18 to 65 in Wisconsin. Knowing that the sample mean has an approximately normal distribution enables us to calculate confidence intervals or test hypotheses.