Lecture 10: Sample distributions, Law of Large Numbers, the Central Limit Theorem

3 October 2005

Very beginning of the course: samples, and summary statistics of samples, like sample mean, sample variance, etc.

If the population has a certain distribution, and we take a sample/collect data, we are drawing multiple random variables. Summaries are functions of samples. In general, we call a function of the sample **a statistic**.

We try to generate samples so that each measurement is independent, because this maximizes the information we get about the population.

Sampling distribution = distribution of the summary statistic, when the observations are drawn independently from a fixed distribution. All of the machinery of probability, random variables, etc., we have developed so far is to let us model this mathematically.

For instance, $X_1, X_2, \ldots X_n$ is a sample of size n.

Now let's consider $\overline{X}_n = \frac{1}{n}\sum X_i$. $\overline{X}_n$ is the random variable which represents the sample mean.

We want to know what happens to the sampling distributions with large samples.

There are two major results here: the law of large numbers, and the central limit theorem. Together, these are the Law and the Prophets of probability and statistics.

The Law of Large Numbers

The law of large numbers says that large random samples are almost always representative of the population.

Specifically,

$$\Pr\left(|\overline{X}_n - \mu| > 0\right) \rightarrow 0$$

or

$$\Pr\left(\overline{X}_n \rightarrow \mu\right) = 1$$

Let's prove this. Assume μ and σ are both finite.

$$\mathbf{E}\left[\overline{X}_n\right] = \mathbf{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$$

$$\begin{aligned} &= \frac{1}{n}\mathbf{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left[X_i\right] \\ &= \frac{1}{n}n\mu \\ &= \mu \\ \mathrm{Var}\left(\overline{X}_n\right) &= \frac{1}{n^2}n\sigma^2 \\ &= \frac{\sigma}{n} \end{aligned}$$

Now let's use Chebyshev: For any $\epsilon > 0$,

$$\begin{aligned} \Pr\left(|\overline{X}_n - \mathbf{E}\left[\overline{X}_n\right]| > \epsilon\right) &\leq \frac{\mathrm{Var}\left(\overline{X}_n\right)}{\epsilon^2} \\ \Pr\left(|\overline{X}_n - \mu| > \epsilon\right) &\leq \frac{\sigma^2}{n\epsilon^2} \\ \Pr\left(|\overline{X}_n - \mu| > \epsilon\right) &\rightarrow 0 \end{aligned}$$

So we can make sure that the sample average $\overline{X}_n$ is probably a good approximation to the population average μ just by taking enough samples. We choose a degree of approximation (ϵ) and a probability level, and that sets the number of samples:

$$n = \frac{\sigma}{\delta\epsilon^2}$$

guarantees that the probability of being off by more than ϵ is at most δ.

Central Limit Theorem

The central limit theorem (CLT) is the most important result in statistics.

$X_1, \ldots X_n$ are independent, again, with mean μ and variance σ^2. Then for large n

$$\begin{aligned} \sum_{i=1}^{n} X_i &\sim \mathcal{N}(n\mu, n\sigma^2) \\ \overline{X}_n &\sim \mathcal{N}(\mu, \frac{\sigma^2}{n}) \end{aligned}$$

and more exactly,

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

Moral: the sampling distribution of the mean will be Gaussian, if you just take enough independent samples.

Why? Basically: the Gaussian distribution is "stable":

- the sum of two independent Gaussians is another Gaussian;
- the product of a Gaussian and a constant is a Gaussian

So averaging — which is adding independent variables, and dividing by constants — leaves Gaussians alone. So non-Gaussian distributions tend to be "attracted" to Gaussians, when you add up enough of them.[1] Notice that, because of the two facts I just mentioned, the central limit theorem always holds *exactly* if X itself is Gaussian, rather than just being a large-sample approximation.

Rule of thumb: the CLT ought to hold pretty well by $n = 30$, or at the worst $n = 100$. Something's probably wrong if it's not working by that point.

Application: the Gaussian approximation to the binomial, which we saw before, is an instance of the CLT. The binomial, remember, is a sum of n independent Bernoulli variables, each with mean p and variance $p(1-p)$. So if $B \sim \mathrm{Bin}(n, p)$, by the central limit theorem, if n is large

$$\begin{aligned} \frac{B}{n} &\sim \mathcal{N}(p, p(1-p)/n) \\ B &\sim \mathcal{N}(np, np(1-p)) \\ B &\sim \mathcal{N}(\mathbf{E}[B], \mathrm{Var}(B)) \end{aligned}$$

If you're measuring the result of many independent random variables, each of which makes some small, but not necessarily *equal* contribution to the outcome, you should expect the result to be Gaussian. We'll see more about this next time, under the heading of "propagation of errors".

In the bad old days before computers, you had to work out the sampling distribution by hand, using clever math. A lot of old-fashioned statistics assumes things are Gaussian, or Poisson, etc., because in those special cases, you can compute the sampling distribution of important statistics, like the median or the variance. Part of the importance of the central limit theorem was that it gave people a way around this, by providing a *general* mathematical result about the sampling distribution of an especially important statistic, namely the sample mean. You could then devote your time to turning the statistic you cared about into some kind of sample mean. These days, however, it's comparatively simple to just *simulate* sampling from whatever distribution we like, and calculate the statistic for each of the simulated samples. We can make the simulated distribution come arbitrarily close to the real sampling distribution just by simulating often enough, which is generally fast. We'll see some examples of this later on. Even with simulation, however, we'll still see that the central limit theorem is incredibly important to statistics.

[1] There are stable non-Gaussian distributions, but they have infinite variance, so you won't see them very often.