STOR 435.001: Lecture 18

Limit Theorems

Jan Hannig

UNC Chapel Hill

Motivation

- We know various intutive things about the real world.
- Example: suppose you wanted to measure obesity levels or diabetes rates in a country e.g. the US.
- The larger the sample you take, in most cases the better the idea you will have of your population.
- How large a sample is enough for studies of interest?
- Why is it enough to take a sample of around 1100 people to get a good idea of things like election poll results? Why should one even trust such studies?
- In CS: Lots of algorithms called MCMC (Markov Chain Monte Carlo) to simulate sophisticated systems. How long do these algorithms need to run to give good results?

Key to all of these: Limit theory.

Central limit theorem (CLT)

If $X_1, X_2, \ldots$ are independent, identically distributed random variables with mean μ and variance σ^2, then the distribution of

$$Z_n := \frac{X_1 + \ldots + X_n - n\mu}{\sigma\sqrt{n}}$$

tends to the standard normal as $n \to \infty$. That is, for any $-\infty < a < \infty$, as $n \to \infty$,

$$F_{Z_n}(a) = P(Z_n \leq a) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx = F_Z(a).$$

This is useful to compute approximate probabilities, and explains why the normal distribution is so wide-spread.

Punchline

Under the above assumptions, **when n is large**, the sum $S_n \approx N(n\mu, \sigma_n^2 = n\sigma^2)$ and the sample average $\bar{S}_n = S_n/n \approx N(\mu, \sigma^2/n)$.

CLT: Example

X_i are independent such that $X_i = \pm 1$ with probability $1/2$. This is an example of a **random walk** where at each stage, the process (e.g. stock price) moves up by 1 or down by 1. Using central limit theorem approximately find the chance that in 50 days, $S_{50} = X_1 + X_2 + \cdots X_{50} > 10$?

☼

Random walk hypothesis of Wall Street: `http://www.wsj.com/articles/SB10001424052702303376904579135634030221864`

Idea of proof of CLT:

☼

Example 3b: The number of students who enroll in a psychology course is a Poisson random variable with mean 100. The professor in charge of the course has decided that if the number enrolling is 120 or more, he will teach the course in two separate sections, whereas if fewer than 120 students enroll, he will teach all of the students together in a single section. What is the probability that the professor will have to teach two sections?

Example 3a: An astronomer is interested in measuring the distance, in light-years, from his observatory to a distant star. Although the astronomer has a measuring technique, he knows that, because of changing atmospheric conditions and normal error, each time a measurement is made it will not yield the exact distance, but merely an estimate. As a result, the astronomer plans to make a series of measurements and then use the average value of these measurements as his estimated value of the actual distance. If the astronomer believes that the values of the measurements are independent and identically distributed random variables having a common mean d (the actual distance) and a common variance of 4 (light-years), how many measurements need he make to be reasonably sure that his estimated distance is accurate to within $\pm .5$ light-year?

☼

Example 3a cont'ed:

Limit theorems

Weak law of large numbers (LLN)

If $X_1, X_2, \ldots$ are independent, identically distributed random variables with mean μ, then, for any $\epsilon > 0$,

$$P\left(\left|\frac{X_1 + \ldots + X_n}{n} - \mu\right| \geq \epsilon\right) \to 0,$$

as $n \to \infty$.

Punchline: Sample mean gets closer and closer to population mean as your sample size becomes larger.

The proof of this result (also assuming finite variance) is based on the so-called Chebyshev's inequality, which follows from Markov's inequality.

⚠

Markov's inequality

If X is a non-negative random variable, then, for any $a > 0$,

$$P(X \geq a) \leq \frac{EX}{a}.$$

☼

Chebyshev's inequality

If X is a random variable with mean μ and variance σ^2, then, for any $a > 0$,

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}.$$

Example 2a: Suppose that it is known that the number of items produced in a factory during a week is a random variable with mean 50. (a) What can be said about the probability that this week's production will exceed 75? (b) If the variance of a week's production is known to equal 25, then what can be said about the probability that this week's production will be between 40 and 60?

Solution contd

Proof of weak LLN: We will now use Chebyshev's inequality to show Weak laws of large numbers.

☼