Math 217 Probability and Statistics

Prof. D. Joyce, Clark University

Friday, 30 Nov 2007

Due Today. From 7.1: 3, 6; from 7.2: 2a, 3a, 5a.

Last time. The Chebyshev inequality. Sample statistics and estimators, sample variance.

Today. The chi-squared distriubution, sample variance for a normal distribution, introduction to the central limit theorem.

The chi-squared distribution. A χ^2 random variable with ν "degrees of freedom" is defined as

$$Z_1^2 + Z_2^2 + \cdots + Z_\nu^2,$$

the sum of the squares of ν independent standard normal random variables. (The letter ν is the Greek letter nu.) It is used in statistics to estimate variances of distributions using samples of large size ν.

It can be shown that a χ^2-distribution is actually a special case of a gamma distribution with a fractional value for r.

$$\textsc{ChiSquared}(\nu) = \textsc{Gamma}(\lambda, r)$$

where $\lambda = \frac{1}{2}$ and $r = \frac{\nu}{2}$ The mean of a χ^2 distribution with ν degrees of freedom is $\mu = \nu$, while its variance is $\sigma^2 = 2\nu$.

Just as there are tables for the standard normal distribution, there are tables for the χ^2 distribution.

The sample variance for a normal distribution. Many natural phenomena have normal distributions, so statistics relating to the normal distribution are particularly important. Here's one of the important theorems about normal distributions.

Theorem. Let $X_1, X_2, \ldots, X_n$ be a sample from a normal distribution with mean μ and variance σ^2. Then the sample statistic

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{(n-1)S^2}{\sigma^2}$$

has a χ^2 distribution with $(n-1)$ degrees of freedom. Also, the sample mean $\overline{X}$ and the sample variance S^2 are independent random variables.

It's a lot of work to prove this theorem, even when n is only 2, and we'll skip the proof. It is used to develop tests about the variance σ^2 of a normal distribution, and since for large n, sample means approach normal distributions, that means it can be used to develop tests about variances for any distribution, although n has to be large in those tests.

The central limit theorem. We begin a serious study of this theorem which says that for any distribution with a finite mean and variance, the sample sum (and also the sample mean) approaches a normal distribution. Before we get to it, we should look at the normal distribution in a little more depth.

The standard normal distribution. We've denoted a standard normal random variable as Z, and, since we'll be studying so much in this last chapter, we'll denote its density function $\phi(x)$. It is given by the formula

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

One thing that's interesting about this function is its integral, that is, the cumulative distribution function for Z, is not an elementary function.

There is no function built from algebraic functions and the usual transcendental functions—the trig, log, and exponential functions—whose derivative is $\phi(x)$. Still, the integral exists since it's a continuous function, and every continuous function has an integral.

If it doesn't have an elementary integral, how do we know its integral from $-\infty$ to ∞ equals 1. In other words, how do we know ϕ is an density function. It so happens that particular integral can be computed, but there are two tricks needed to compute it. One is, rather than compute the integral, we'll compute its square. The other is converting to polar coordinates. If you've already studied multivariate calculus, then you'll be able to follow this argument, otherwise treat it as a preview to a course that I hope you'll take soon.

We'll show that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = 1$$

by squaring the left side, and showing it's 1. The square is that integral times itself, but when we write it the second time, we'll use the bound variable y instead of the bound variable x.

$$\begin{aligned} & \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-y^2/2}\, dy \\ = \;& \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\, dy\, dx \end{aligned}$$

Now we'll change to polar coordinates. Then $x^2 + y^2 = r^2$, and $dy\, dx = r\, dr\, d\theta$. (The extra r in the last equation is the Jacobian.) The new limits of integration will be $0 \le \theta \le 2\pi$ and $0 \le r \le \infty$. The double integral becomes

$$\frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r\, dr\, d\theta$$

Let's evaluate the inner integral. A substitution $u = r^2/2$, $du = r\, dr$ makes it easy.

$$\begin{aligned} \int_{r=0}^{\infty} e^{-r^2/2} r\, dr &= \int_{u=0}^{\infty} e^{-u}\, du \\ &= -e^{-u} \Big|_0^{\infty} = 1 \end{aligned}$$

Putting that in the outer integral gives

$$\frac{1}{2\pi} \int_0^{2\pi} 1\, d\theta = \frac{1}{2\pi} \theta \Big|_0^{2\pi} = 1.$$

Thus, ϕ actually is a density function.