Probability: Problem Set 7

Fall 2009
Instructor: W. D. Gillam
Due Nov. 20, start of class

Instructions. Print your name in the upper right corner of the paper and write "Problem Set 7" on the first line on the left. Skip a few lines. When you finish this, indicate on the second line on the left the amount of time you spent on this assignment and rate its difficulty on a scale of $1-5$ (1 = easy, 5 = hard). When you are asked to plot graphs, please use a computer and reasonable plotting software (Mathematica, Matlab, etc.)

IIDRV := independent identically distributed random variables. Remember it.

(1) Calculate the moment generating function $g_X(t) = E(e^{tX})$ when X has a gamma distribution with parameters α, β.

(2) Use the above result to prove that when $X_1, \dots, X_n$ are IIDRV with standard normal distribution, then $X_1^2 + \cdots + X_n^2$ has a chi squared distribution with n degrees of freedom. Do this by showing that the moment generating functions coincide. Recall that we proved this in class by another method (more-or-less explicit calculation of density functions...).

(3) Calculate the moment generating function $g_X(t)$ when X has a normal distribution with expected value μ and variance σ^2. Use this to prove that a sum of independent random variables with (possibly different!) normal distributions again has a normal distribution. As in the previous problem, do this by equating moment generating functions instead of explicitly calculating the convolution of densities as we did in class.

(4) A random variable X is said to have a t-distribution with n degrees of freedom if X has the same distribution as

$$\frac{X_0}{\sqrt{\frac{1}{n}(X_1^2 + \cdots + X_n^2)}} \quad = \quad \sqrt{n}\frac{X_0}{\sqrt{X_1^2 + \cdots + X_n^2}},$$

where $X_0, \dots, X_n$ are IIDRV with standard normal distribution. This just differs by a factor of $\sqrt{n}$ from Student's distribution, but it is sometimes more natural. Calculate the variance of X and observe that it is > 1 (assuming that $n > 2$ so that it exists). Plot the standard normal density and the t-densities with $n = 1, 2, 3$ on the same axes. Observe that the t-density is similar to the standard normal bell curve, but slightly more spread out, accounting for its slightly larger variance.

(5) When we sketched the proof of Stirling's Formula in class, recall that we approximated $n!$ by writing

$$n! \quad = \quad \left(\frac{n}{e}\right)^n \sqrt{2\pi n} + \epsilon_1 + \epsilon_2,$$

where the "errors" ϵ_i were supposed to be *small* (compared to the leading term) in the sense that

$$\lim_{n\to\infty} \left(\frac{n}{e}\right)^{-n} \frac{\epsilon_i}{\sqrt{2\pi n}} \quad = \quad 0$$

for $i = 1, 2$. The error ϵ_1 arose when we approximated the integral

$$n! \quad = \quad \int_0^\infty x^n e^{-x} dx$$

by using the series expansion

$$\ln(x^n e^{-x}) \quad = \quad n \ln n - n - \frac{(x-n)^2}{2n} + O((x-n)^3)$$

to approximate the integrand; I will not ask you to prove this error is small, but try to prove it if you want. The second error

$$\epsilon_2 \quad = \quad \left(\frac{n}{e}\right)^n \int_{-\infty}^0 \exp\left(\frac{-(x-n)^2}{2n}\right) dx$$

arose when we wanted to change limits of integration. Calculate this integral explicitly and verify that

$$\lim_{n\to\infty} \frac{1}{\sqrt{n}} \int_{-\infty}^0 \exp\left(\frac{-(x-n)^2}{2n}\right) dx \quad = \quad 0.$$

(The following problems are taken from a statistics textbook and are intended to give you some “real world” examples. Use a computer to evaluate the normal probabilities. You can easily find web-based software that will do integrals of the normal density functions.)

(6) A forester studying the effects of fertilization on certain pine forests in the Southeast is interested in estimating the average basal area of pine trees. In studying the basal areas of similar pine trees for many years, he has discovered these measurements (in square inches) to be normally distributed with standard deviation[1] approximately 4 square inches. If the forester samples $n = 9$ trees at random, find the probability that their average basal area will be within 2 square inches of the population mean.

(7) Suppose the forester wants to be at least 90 percent certain that the sample mean will be within 2 square inches of the population mean. How many pine trees must he measure?

(8) Suppose this forester is transfered to Maine and is interested in estimating the basal area of pine trees there. He assumes the basal areas of pine trees in Maine will have a normal distribution, but he isn't sure about the average basal area of pine trees in Maine, or its standard deviation. He measure the basal areas of some randomly selected pine trees and obtains:

$$162 \text{ in}^2,\ 171 \text{ in}^2,\ 155 \text{ in}^2,\ 189 \text{ in}^2,\ 148 \text{ in}^2,\ 195 \text{ in}^2,\ 165 \text{ in}^2,\ 189 \text{ in}^2,\ 150 \text{ in}^2.$$

What is his best guess for the expected value and standard deviation of the basal area of Maine pine trees?

(9) The EPA is concerned with the problem of setting criteria for the amounts of certain toxic chemicals to be allowed in freshwater lakes and rivers. A common measure of toxicity for a pollutant is the *LC50*: the concentration of the pollutant that will kill half of the test species in a given amount of time (usually 96 hours

[1]Standard deviation is the square root of the variance. Note that the units make sense. Variance would be measured in inches^4 since it is the sum of the expected value of the square of a quantity measured in square inches and the square of the expected value of a quantity whose expected value is measured in square inches.

for fish species).[2] In many studies, the values contained in the (natural) log of LC50 measurements are normally distributed and hence the analysis is based on ln LC50 data.

Studies of the effects of copper on trout show the variance of log LC50 measurements to be around .4 with concentrations measured in mg / L. If $n = 10$ studies on LC50 for copper are to be carried out, find the probability that the sample mean of log LC50 will differ from the true population mean by no more than .5. If the EPA wants to be 95 percent sure that the sample mean of log LC50 will differ from the population mean by no more than .5, how many tests should they carry out?

(10) Suppose $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are independent random variables, and the X_i are normally distributed with expected value μ_1, variance σ_1^2, while the Y_i are normally distributed with expected value μ_2 and variance σ_2^2. Calculate the distribution of $Z = X_1 + \cdots + X_n + Y_1 + \cdots + Y_m$.

(11) Suppose the effects of copper on bass show the variance of the log LC50 measurements to be .8. If the population means of the log LC50 are the same for bass and trout, find the probability that, with random samples of ten log LC50 measurements for each species, the sample mean for trout exceeds the sample mean for bass by at least 1.

[2] You are probably familiar with the LD50 measure of poisons.