MATHEMATICS E-156, SPRING 2014
MATHEMATICAL FOUNDATIONS OF STATISTICAL SOFTWARE
Module #8 (Student t distribution)

Last modified: April 4, 2014

Reading from Chihara and Hesterberg

- Section 7.1 (you have already read the first part of this)
- Appendix B, Sections B.9, B.10, and B.11

Proof of the Week

- I consolidated several from the original list. Everything in the math notes qualifies. However, a couple of these proofs may be too intricate to qualify as candidate final exam questions.

R scripts

- Script 8A-GammaChiSquare.R
  Topic 1 - properties of the gamma function(goes with math notes on page 3)
  Topic 2 - density function for the gamma distribution(special case of the mgf on page 4)
  Topic 3 - moment generating function for the gamma distribution(goes with math notes, page 4)
  Topic 4 - the connection between the chi square and gamma distribution(goes with math notes, page 5)
  Topic 5 - the connection between the chi square and normal distributions(goes with math notes, page 5)

- Script 8B-Student t.R
  Topic 1 - the sum of two standard normal random variables(goes with math notes, page 6)
  Topic 2 - the sum of three standard normal random variables
  Topic 3 - the sum of k standard normal random variables with mean μ, variance σ^2(goes with math notes, pages 7-9)
  Topic 4 - the sum of k independent random variables from a distribution that is not normal.

- Script 8C-TConfidenceIntervals.R
  Topic 1 - review of the case where we know the population variance but not the mean
  Topic 2 - the case where we have to estimate the standard deviation also.
  Topic 3 – t confidence intervals with real-world data

Mathematical notes

1. Define the gamma function for $r > 0$ by $\Gamma(r) = \int_0^\infty x^{r-1}e^{-x}dx$. Prove that

   - $\Gamma(r+1) = r\Gamma(r)$ if $r > 0$.

   - For integer $n > 0, \Gamma(n+1) = n!$

   - $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.

2. A random variable X has the gamma distribution if its probability density function is

$$f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, x \geq 0$$

Prove the following:

- The moment generating function is $M(t) = (\frac{\lambda}{\lambda - t})^r$.

- $$E[X] = \frac{r}{\lambda}; E[X^2] = \frac{(r+1)r}{\lambda^2}; \text{Var}[X] = \frac{r}{\lambda^2}$$

- If $X_1, X_2, \cdots X_n$ are independent random variables with $X_i \sim \text{Gamma}(r_i, \lambda)$, then

$$X = X_1 + X_2 + \cdots + X_n \sim \text{Gamma}(r_1 + r_2 + \cdots + r_n, \lambda).$$

3. A random variable X has the chi-square distribution with m degrees of freedom if

$$X \sim \chi^2_m \sim \text{Gamma}(\frac{m}{2}, \frac{1}{2})$$

. Using properties of the gamma distribution, prove the following:

- If $X_1, X_2, \cdots X_n$ are independent chi-square random variables with degrees of freedom $m_1, m_2, \cdots m_n$, then $X = X_1 + X_2 + \cdots + X_n$ is chi-square with $m = m_1 + m_2 + \cdots + m_n$ degrees of freedom.

- If $Z \sim N(0,1)$, then Z^2 is chi-square with one degree of freedom.

- If $Z_1, Z_2, \cdots Z_k$ are independent $N(0,1)$ random variables, then $X = Z_1^2 + Z_2^2 + \cdots Z_k^2$ has a chi-square distribution with k degrees of freedom.

- The moment generating function of X is $M(t) = (1-2t)^{-k/2}$.

4. Let X_1, X_2 be a random sample from $N(0,1)$, with sample mean $\overline{X}$ and sample variance S^2. Prove the following:

   - $X_1^2 + X_2^2 = 2\overline{X}^2 + S^2$.

   - $E[\overline{X}^2 S^2] = E[\overline{X}^2]E[S^2]$
     (Use the fact that if X is standard normal, $E[X^2] = 1$ and $E[X^4] = 3$.

- Two random variables Z and W are independent if and only if their moment generating function factors: i.e.

  $$E[e^{Zs+Wt}] = E[e^{Zs}]E[e^{Wt}].$$

  Show that this is the case if X_1 and X_2 are independent standard normal random variables and we choose
  $Z = X_1 + X_2, W = X_1 - X_2$.
  This is true only for a normal distribution; so we must use the fact that for normal X with $\mu = 0, M(t) = E[e^{Xt}] = e^{\sigma^2 t^2/2}$.

- If two random variables Z and W are independent, so are $f(Z)$ and $g(W)$. Show that $\overline{X}$ is independent of S^2.

5. Let $X_1, X_2, \cdots X_n$ be a random sample from $N(\mu, \sigma^2)$, with sample mean $\overline{X}$ and sample variance S^2. It continues to be true that $\overline{X}$ and S^2 are independent random variables. Define

$$U = \frac{1}{\sigma^2}\sum_{i=1}^{k}(X_i - \mu)^2; V = \frac{1}{\sigma^2}\sum_{i=1}^{k}(X_i - \overline{X})^2; W = \frac{1}{\sigma^2}n(\overline{X} - \mu)^2.$$

Prove the following:

- $U = V + W$.

- $(n-1)S^2/\sigma^2$ has a chi-square distribution with $n-1$ degrees of freedom.

6. Let $Z \sim N(0,1)$ and let W denote a chi-square distribution with k degrees of freedom, independent of Z. Let T be the ratio

$$T = \frac{Z}{\sqrt{W/k}}$$

Prove that T has a t distribution: i.e. its probability density function is

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

7. Prove that if $X_1, X_2, \cdots X_n$ are a random sample from $N(\mu, \sigma^2)$, then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a t distribution with $n - 1$ degrees of freedom.

8. Student t confidence interval

   Let $X_1, X_2, \cdots X_n$ be a random sample from $N(\mu, \sigma^2)$, with both μ and σ unknown. The sample mean is $\overline{X}$; the sample variance is S^2.

   Let q denote the $(1-\alpha/2)$ quantile of the Student t distribution with $n-1$ degrees of freedom. By symmetry, $-q$ is the $\alpha/2$ quantile.

   Define random variables

$$L = \overline{X} - \frac{qS}{\sqrt{n}}; U = \overline{X} + \frac{qS}{\sqrt{n}}.$$

   Prove that $P(L > \mu) = \alpha/2$ and $P(U < \mu) = \alpha/2$,
   so that $P(L \leq \mu \leq U) = 1 - \alpha$. and $[L, U]$ is a $1 - \alpha$ confidence interval.

Section problems

1. Find an explicit formula, evaluating all gamma functions, for the Student t density with $n = 3$. Confirm by integration (use R – this is hard but not impossible to do by calculus) that the variance is $k/(k-2) = 3$. Make a histogram of the sample third moment for 100, 1000, and 10000 samples. Does this look like a consistent estimator?

2. Start with a random variable Y that has the binomial distribution with $n = 20, p = 0.5$. If you subtract the mean and divide by the standard deviation, you will get a discrete random variable X whose distribution will resemble a standard normal distribution, and the approach used for Student t might work fairly well. Try drawing samples of size $k = 4$ from such a distribution, and repeat 1000 times (or more if your computer is fast) to answer the following questions.

   (a) If you multiply the square of the sample mean by n, does it have a distribution that is approximately chi-square with one degree of freedom?

   (b) Does the sum of the squares of the samples have a distribution that is approximately chi-square with $k = 4$ degrees of freedom?

   (c) If you multiply the sample variance by $k-1$, does it have a distribution that is approximately chi-square with $k - 1 = 3$ degrees of freedom?

   (d) Is the square of the sample mean uncorrrelated with the sample variance?

   (e) If you divide the sample mean by the sample standard deviation, does it have an approximate Student t distribution with $k - 1 = 3$ degees of freedom?

   You can model your calculations on the ones in script 8B-Student t. Different people can answer different parts of the question and experiment with differnt value of n and k

   (The remaining two problems will have to wait a week)

3. Page 202, exercise 9. You can check your answer on page 403.

4. Page 202, exercise 7. It is very easy to crank out a t confidence interval – the hard part is to decide whether it is meaningful or not.

Homework assignment This assignment should be submitted as a single R script. Include enough comments so that it is clear what you are doing and where each problem begins. You can upload it to the dropbox on the Class 8 page of the Web site.

It is OK to paste and edit lines from the scripts on the course Web site. It is not OK to paste lines from your classmates' solutions!

1. (a) Invent a way of creating a random variable X that has a chi-square distribution with six degrees of freedom by sampling from an exponential distribution. You get to choose the sample size and the value of lambda. Verify your result by making a histogram of the sample means and overlaying a chi-square distribution.

   (b) The Student t distribution results when you divide a standard normal random variable by the square root of an independent, suitably rescaled, chi-square random variable. Nowhere does it say the the numerator and denomiator have to come from the same source. They only have to be independent.

   Create a variable with an approximate Student t distribution with six degrees of freedom by taking the mean of 50 samples from Unif[0,1] and dividing by the chi square variable that you created for part (a). Show that a histogram of values from this distribution matches the built-in t density function on R.

2. By evaluating the gamma functions in the formula for the t density on page 389, find explicit formulas for the t density function for 4 and 5 degrees of freedom, and show that they agree with the R function $dt(x)$. Using 50000 trials, confirm that the variance is close to $k/(k-2)$ and attempt to estimate the fourth moment in each case.

3. It turns out that the only distribution for which the sample mean and the sample variance are independent is the normal distribution. For samples of size 6 from $N(0,1)$, doing 5000 experiments, create scatter plots of the square of the sample mean against the sample variance, and calculate the correlation. Once you have determined whether the correlation is positive or negative, offer an anecdotal explanation, based on an extreme case, of why this should be so.

4. According to the Internet, in 2012 Pope Benedict XVI "put a dampener on the festive period by rubbishing the idea that donkeys or any other animal have a place in the traditional nativity scene." Since animal for creches are big business in some French villages (Google "Aubagne santons"), we shall imagine that they can be replaced by cherubim who are precise ceramic replicas of North Carolina newborns from 2004 , the subject of section 5.1 of the textbook. When we looked at the birth weights of these babies in script 6A, we found a mean $\mu = 3448.26$ and a standard deviation $\sigma = 487.5$, and we noticed that a histogram of the weights looked sort of normal.

   If random variable Y is the weight of one of the these babies, it might not be unreasonable to assume that the random variable $X = (Y - \mu)/\sigma$ has a standard normal distribution.

   The mayor of Aubagne wants to know the expected weight of one of the babies and asks his assistant to hack the birth records of North Carolina hospitals to determine μ and σ. Alas, all the assistant can get is a sample of six birth weights. This is the same sort of problem that was faced by Student.

   We know the entire population distribution and can do a simulation to determine whether these samples of 6 behave as if drawn from a population with a normal distribution.

   Create samples of size $k = 6$ that have an approximate standard normal distribution by selecting six weights from the NC baby data, subtracting μ and dividing by σ. Using 1000 such samples (or more if your computer is fast), answer the following questions.

   (a) If you multiply the square of the sample mean by n, does it have a distribution that is approximately chi-square with one degree of freedom?

   (b) Does the sum of the squares of the samples have a distribution that is approximately chi-square with $k = 6$ degrees of freedom?

   (c) If you multiply the sample variance by $k-1$, does it have a distribution that is approximately chi-square with $k - 1 = 5$ degrees of freedom?

   (d) Is the square of the sample mean uncorrrelated with the sample variance?

   (e) If you divide the sample mean by the sample standard deviation, does it have an approximate Student t distribution with $k - 1 = 5$ degrees of freedom?

   As with section problem 2, you can model your calculations on the ones in script 8B-Student t.

(The last two problems will have to wait a week)

5. Page 202, exercise 6. After solving the problem, do a simulation of 1000 trials, drawing ice cream samples from a normal distribution with population mean 18.05 g and population variance 5 g, and make a plot like the one in figure 7.1 to show that the t confidence interval performs as expected.

6. Page 204, exercise 19. After solving the problem (the answer is on page 403), do a simulation where the sales tax paid is a random variable with a gamma distribution with a mean of 5 dollars and a standard deviation of \$3.50 (see page 408 for the formulas that will let you compute the parameters). Do a simulation with samples of size 500, and determine how well the 75% upper t confidence interval performs.

   This is interesting because with samples of 500, the CLT works pretty well even with a skewed distribution, but the theory behind Student t assumes that the underlying distribution is normal. This subtlety was probably wasted on the officials who drew up the policy.