Mathematics 376 – Mathematical Statistics
Computer Lab Day 1 in R
February 6, 2012

Goals

The goals of today's lab are:

- To gain some more intuitive understanding of the χ^2, t, F distributions from §7.2 – their densities, distributions, etc. and computing with them in R.
- To solidify understanding of how those distributions relate to sampling from normal populations.

Basic R Mechanics – from Fall Lab Day 1

Refer to the assignment sheet for the first lab day if you need to refresh your memory about running R in Haberlin 136, about how to save and print your results, etc. The Lab 1 assignment is posted on the course homepage, so you can open it and look at the instructions there as needed.

Background on Probability Distributions in R – from Fall Lab Day 2

The R package contains built-in functions for computing probability functions for many of the discrete random variables we have discussed, plus functions for the probability densities and cumulative distributions of many of the standard types of continuous random variables, including the χ^2, t, F types that we have just introduced. Here's how it works. (Also look at table on 332 of Dalgaard *Introductory Statistics with R*.) Each type of random variable is covered by a *family* of 4 functions distinguished by a *prefix letter* `d, p, q`, or `r`:

- `d` – the probability function in the discrete case, or the density function continuous case,
- `p` – the cumulative distribution
- `q` – the quantile function (essentially the *inverse function* of the cumulative distribution, but of course the cumulative distribution can fail to be injective, so this needs to be taken with a grain of salt)
- `r` – random number generator

Following the prefix letter comes the rest of the function name and the inputs needed to compute the corresponding values. For instance the family of functions for χ^2 random variables looks like this, where *df* represents the number of degrees of freedom:

- `dchisq(y,d)` – the probability density function for a $\chi^2(d)$ random variable (d = degrees of freedom):

$$f(y) = \begin{cases} \frac{y^{(d/2)-1}e^{-y/2}}{\Gamma(d/2)2^{d/2}} & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

- `pchisq(y,df)` – the cumulative distribution $P(Y \leq y)$
- `qchisq(q,df)` – computes the y such that $P(Y \leq y) = q$
- `rchisq(N,df)` – generates a list of N random numbers following a $\chi^2(n)$ distribution.

The t-distribution family of functions are called `dt, pt, qt, rt`. The F-distribution family of functions have the names `df, pf, qf, rf`.

Today's Lab

I. Basic computations.

1. If $Y \sim \chi^2(12)$, find $P(Y > 10)$.
2. Find a real value a such that $P(Y > a) = .01$ if $Y \sim \chi^2(10)$. Check your answer using the tables in our text.
3. If $Y \sim t(23)$, find $P(-2.2 < Y < 2.2)$. Compare with the corresponding probability for the the standard normal Z – $P(-2.2 < Z < 2.2)$. Explain how this matches with our intuitive picture of the t-densities and their dependence on n.
4. If $Y \sim F(10, 14)$, find $P(5 < Y < 16)$.
5. If $Y \sim F(9, 6)$, determine a real number a such that $P(Y \leq a) = .05$

II. Overall characteristics of the χ^2 family of densities.

1. Produce a plot showing the χ^2 densities for $n = 4, 6, 8, 10, 12$ degrees of freedom together on the interval from $y = 0$ to $y = 30$. (Notes: The `R curve` command will be useful for this. Recall that you can superimpose plots using the `add = TRUE` option. The plotting interval can also be controlled using the `from =` and `to =` options.)
2. Describe in a short paragraph how the value of n = number of degrees of freedom affects the shape of the $\chi^2(n)$ density graph. (Look closely at the shape at $y = 0$ and at the location and height of the maximum as n changes.)

III. Overall characteristics of the F family of densities.

1. Produce a plot showing the F densities for numerator degrees of freedom $n = 2, 3, 4, 5, 6, 7, 8, 9$ and denominator degrees of freedom $d = 6$ together on the interval from $y = 0$ to $y = 6$.
2. Produce a plot showing the F densities for numerator degrees of freedom $n = 6$ and denominator degrees of freedom $d = 2, 3, 4, 5, 6, 7, 8, 9$ together on the interval from $y = 0$ to $y = 6$.

IV. A special property of the F-distributions. The motivation for this question is the fact (which you may have noted) that our book's F-tables only include the percentage points F_α for "small" values of α – namely $\alpha = .005, .01, .025, .05, .1$. What about the complementary "large" values $\alpha = .995, .99, .975, .95, .9$? Why aren't they there? This problem will show you the answer to those questions and how to generate the $\alpha = .995, .99, .975, .95, .9$ percentage points of the F-distributions if those are needed.

1. For instance, our book's tables show that $F_{.1}(10, 8) = .254$ (see page 855). Check this using `R`. (Note: The book's table shows the value $F_{.1}$ such that $P(Y \geq F_{.1}) = .1$ (i.e. the .1 is the area in the *upper tail*. This means that $P(Y < F_{.1}) = .9$.)

2. What happens if you enter

   `qf(.9,10,8)*qf(.1,8,10)`?

   What happens if you enter $qf(\alpha, n, d) * qf(1 - \alpha, d, n)$ for any particular values of α, n, and d? Try as many examples as you need to see the answer to the next part.

3. If we know the percentage point $F_\alpha(n, d)$ (say for one of the "small values" given in the book's table), how can we find the complementary percentage point $F_{1-\alpha}(n, d)$? State a general formula for this, and use it to compute $F_{.95}(15, 4)$ using the information in the book's tables. Check with `R`.

4. (for after the lab) Prove your formula in the previous part using the definition of F-distributed random variables.

V. The T-statistic. Recall that if $Y_1, \ldots, Y_n$ are i.i.d. samples from a $N(\mu, \sigma^2)$ population, then the closest thing to a standard normal that we can compute without knowledge of σ is the T-statistic:

$$T = \frac{\sqrt{n}(\overline{Y} - \mu)}{S}.$$

Here S is the sample standard deviation:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2}.$$

In this problem, you will study the empirical distribution of the T-statistics in a case where $n = 8$ is relatively small, along the lines of something we did on Lab 3 from the fall.

1. Our main goal is to understand what happens if we generate *lots* of such collections of $n = 8$ samples, find their sample means $\overline{Y}$ and sample variances S^2, then consider *the distribution of the T-statistics of the samples*

   $$T = \frac{\sqrt{8}(\overline{Y} - \mu)}{S}.$$

   Do these have an approximately normal distribution, or is it different?

2. Here is one way to do this in R using a simple *for loop*. First we create space to store the values of T computed from 1000 different random samples:

```
ts <- array(1:1000)
```

   Next, we generate 1000 different random samples of size $n = 8$ from a $N(10, 25)$ distribution, find the T-statistic values of each of them, and store them in the array created above:

```
for (i in 1:1000)
   {
   sample <- rnorm(8,10,25)
   ts[i] <- sqrt(8)*(mean(sample) - 10)/sd(sample)
   }
```

   (Note: This could all be entered as one input line provided you leave spaces in the appropriate places. However, I think it is more readable (and it is easier to identify typos if you happen to make one) if you press ENTER at the end of each of the lines as you are typing this. If you do it that way, you should note that R generates a new input prompt + each time, indicating that you are still in the body of the for loop. The final } will close off the loop and take you back to the > prompt.)

3. What does the distribution of the sample means look like? We can see that by generating a density histogram for the data in the ts array, together with a standard normal density:

```
curve(dnorm(x,0,1),from=-5,to=5)
 hist(ts,breaks=20,freq=FALSE)
```

   What do you see from this display? Is the standard normal density a good match for the distribution of the T-statistic? Is there another density that should fit better? Add that to your plot. (Another Question: Why did I put the first two plotting commands in this order and not the other way around?)

Assignment

Lab reports containing input, output, and answers to the question posed above are due no later than Monday, February 13. If you do not finish during the hour today, you can return to HA 136 any time it is not in use by another class.