Mathematics 376 – Mathematical Statistics
Lab Project 3 and Problem Set 4 – More on Confidence Intervals
February 27 and 29, 2012 – *due:* Friday, March 2

Background

We have now discussed constructing:

- Large-sample confidence intervals for means, differences of subpopulation means, proportions and differences of subpopulation proportions (using pivotal quantities that have normal distributions)
- Small-sample confidence intervals for means and differences of means under the assumption that the population has a normal distribution (using pivotal quantities that have t distributions)
- Confidence intervals for variances or standard deviations (using pivotal quantities that have χ^2 distributions).
- Confidence intervals for a *ratio* of subpopulation variances (using pivotal quantities that have F distributions).

In this week's lab we will begin by studying how statisticians might test to see whether the assumptions underlying the small-sample cases are reasonable. Then we will apply some of this to an applied problem.

Worked Example and Background

In deriving the small-sample formulas for confidence intervals for population means and differences of population means, recall that we had to make some rather restrictive additional hypotheses in order to apply the facts we know about t-distributions:

- we need to assume that the samples are drawn from *normal* populations
- for the difference of means case, we need to assume the two population *variances are equal*.

We have seen how to derive some evidence about whether the equality of population variances is reasonable, via our confidence intervals for σ_1^2/σ_2^2 (derived using the F-distribution). But this still leaves some questions:

- Can we test whether it is reasonable to assume the samples are coming from a normal population?
- How much of a difference does it make if they are not?

Thus, it becomes important to know that there are ways–some more intuitive, some more precise–to get some indication whether it is reasonable to make the assumption that a collection of sampled data has been drawn from a normal population.

The first way we will discuss is a graphical method called the *normal quantile-quantile plot*. The idea here is to compare the distribution of the sampled data values with the distribution we would expect from a normal population in a particular way. Doing this by

hand, we would order the sample and compare the quantiles (i.e. order statistics) for the data to the expected values of the quantiles for a sample of the same size drawn from a normal population. A scatter plot of data points constructed from the two sets of quantiles is generated to give a visual comparison. To make life easier, `R` has a built-in command called `qqnorm` that does this computation and produces the plot. To see how this works, we will work with the following (simulated) net income data (in units of $1000) from a random selection of $n = 9$ tax returns. Enter in `R`:

```
incomes <- c(20.1,35.5,44.7,50.3,60.8,90.7,140.2,240.9,357.3)
qqnorm(incomes)
```

What does this plot mean? Well, for example, the sample maximum (the $Y_{(9)}$) in our incomes is the 357.3 (that is one of the *"Sample Quantiles"*). On the other hand, if we took a random sample $Z_1, \ldots, Z_9$ from a standard normal distribution, the expected value of the maximum there is $E(Z_{(9)}) \doteq 1.485$ (that is one of the *"Theoretical Quantiles"*). (This value can be computed by numerically integrating the density function for $Z_{(9)}$ using the formulas from §6.7.) The point farthest to the right in the QQ Plot is in fact $(1.485, 357.3)$(!)

What are we looking for?? Well, if the data *are* coming from a normal population then as usual $\frac{Y-\mu}{\sigma} = Z$ has a standard normal distribution. So

$$Y = \sigma Z + \mu,$$

and we would expect the scatter plot to be nearly a *straight line*, since the quantiles would also be related by this linear equation. Some "random up and down variation" from that line would also be expected because of chance errors coming from the sampling process.

In this example, though, notice that there is an apparent systematic curvature (concavity) in the scatter plot. If you see this kind of pattern, it is a strong indication that assuming normality of the underlying population is probably not justified!!

For now, we will leave this discussion at this somewhat intuitive level. We will return to these questions later on, in our discussion of regression (finding a line that "best fits a collection of data") and the correlation coefficient that we discussed briefly last semester in connection with covariances.

For your general "statistical literacy," it is also important to be aware that there are a number of other much more sophisticated statistical normality tests that are commonly used to get information in this situation. `R` has a very good one (called the Shapiro-Wilk W Test). Try entering

```
shapiro.test(incomes)
```

We will discuss the ideas involved in this sort of test (especially the exact meaning of the `p-value`) after Spring Break. For now, what this is saying is that (with a p-value less than .05 or so), there is *actually pretty strong evidence to reject the assumption that these samples were coming from a normal population*! On the other hand, *larger* p-values (say .1 or larger) could be the result of random variations in the sampling, so we don't necessarily want to reject the normality assumption in those cases.

Lab Problems

A) We can test for normality now, but does it really *matter* if we compute confidence intervals using the small sample formulas when the normality assumption is violated? In this problem you will be working with a new family of distributions in `R`, the lognormal distributions. These are defined in problem 4.128 in our book – the idea is that if Y is lognormal with parameters μ, σ then $X = \ln(Y)$ is normal with the same μ and σ. There is a family of functions in `R` for the lognormals just like the other cases we have seen:

- `dlnorm` gives the lognormal density
- `plnorm` gives the lognormal cumulative distribution
- `qlnorm` gives the lognormal quantiles
- `rlnorm` generates random samples from a lognormal

With this background, we can get down to work.

1) To get a feeling for what lognormal distributions look like, plot

   ```
   curve(dlnorm(x,0,1),from=0,to=10)
   ```

   to see the density for the lognormal with $\mu = 0$ and $\sigma = 1$. This is different from a normal density, but similar too in some ways (it's also "unimodal" for instance).
2) Generate a single random sample of size $n = 100$ from the lognormal distribution and use the `qqnorm` and `shapiro.test` commands discussed above on it. What is the conclusion you draw.
3) Following what we did in Lab 1, generate 1000 random samples of size 8 from a lognormal distribution with $\mu = 0$ and $\sigma = 1$. The expected value of this lognormal is $e^{1/2} \doteq 1.6487$. Compute the T-statistic

   $$T = \sqrt{8}\frac{\overline{Y} - e^{1/2}}{S}$$

   for each sample, and plot a histogram of the T values. Overlay a $t(7)$ density curve. Discuss your results and think about the main question posed at the start of this problem.

(*Comment:* The following statement comes from page 525 of Wackerly, Mendenhall, and Scheaffer – "... investigations [of the empirical distributions of the T-statistic] have shown that moderate departures from normality in the distribution of the population have little effect on the distribution of the test statistic." In order not to be misled by what they say, I think one would have to have a more precise notion of "moderate departures from normality!" There is actually a lot of interest in developing "robust statistics" today – ones that don't require assumptions like the ones we have used for the T-statistic, and that give reasonable results even when those assumptions don't hold.)

B) A lab technician tests blood samples from 20 men and measures the total serum cholesterol level (LDL + HDL) for each. The data obtained were as follows:

199	272	261	248	235	192	203	278	268
230	242	305	286	310	345	289	326	
335	297	328						

(Ouch!! Not a healthy-eating, regular-exercise group on the whole – current guidelines say this level should be no higher than 200 in adult males to minimize risk of heart disease and strokes.)

1) Use R to test whether it is reasonable to assume that this data comes from a normal population. Explain your conclusions.
2) Use R to compute a 90% confidence interval for the population mean serum cholesterol level based on this data. You may want to refer back to the assignment sheet from Lab 2, although that used the large sample formulas.

In deriving our formulas for the $(1-\alpha)\times 100\%$ confidence interval for the variance, we used an interval where the "tails" of the χ^2-distributions had equal probabilities (areas) $\alpha/2$. This gives the general form

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right]$$

for the $(1-\alpha)\times 100\%$ confidence interval, where, as in the notation of the text book's χ^2 tables, if Y has a χ^2 distribution,

$$P(Y \geq \chi^2_\beta) = \beta.$$

Of course, this method also gives confidence intervals for standard deviations by taking square roots:

$$\left[\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}}\right]$$

3) Determine a 95% confidence interval for the population σ for the cholesterol data.

B) (This one will require some thought before you start plugging numbers into R!) A group of pediatric nurses were interested in the effect of prenatal care on the birthweight of babies. Mothers were divided into two groups, and their babies' weights were compared. The mothers in group 1 had 5 or fewer prenatal care sessions, and their babies had birthweights:

$$49, 108, 110, 82, 93, 114, 134, 114, 96, 52, 101, 114, 120, 116$$

(all in ounces). The mothers in group 2 had 6 or more prenatal care sessions, and their babies had birthweights:

$$133, 108, 93, 119, 119, 98, 106, 87, 153, 116, 129, 97, 110, 131$$

(also in ounces). One question they were trying to address with this study was:

Was there more variation in the birthweights of the babies whose mothers had fewer prenatal care visits than in the birthweights of the babies whose mothers had more visits?

1) Explain why $\theta = \sigma_1^2/\sigma_2^2$ is an appropriate target parameter to consider, and why the statistic $\widehat{\theta} = S_1^2/S_2^2$ is a reasonable estimator (S_i^2 = sample variance for group $i = 1, 2$).
2) What is the value of this statistic here, and what does it suggest?
3) But now, the next question is: How reliable is that conclusion? How likely is it that we could get a different conclusion for different random samples from the same distributions? To answer this we can try to construct a confidence interval for the ratio $\theta = \sigma_1^2/\sigma_2^2$. To do this, explain why $(S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$ is an appropriate "pivotal quantity" to use to derive confidence intervals for σ_1^2/σ_2^2. Recall this means that:
   - σ_1^2/σ_2^2 must be the only unknown part and
   - the distribution of the pivotal quantity must be *known*.
4) Explain how to get a $(1 - \alpha) \times 100\%$ confidence interval for σ_1^2/σ_2^2, and use your method to find a 95% confidence interval for σ_1^2/σ_2^2 using the data above. What does this suggest about your conclusion from part 2?

Problems from the text – Solve using `R`

- Chapter 8/88,89,96,100