Mathematics 376 – Probability and Statistics II
Lab Project 2 and Problem Set 4 – More on Confidence Intervals
February 22 and 23 – *due:* February 25

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).

In this week's lab we will begin by applying some of this to an applied problem. Then, we will consider the problem of finding *shortest confidence intervals* for standard deviations, given the confidence level α. Finally, we will consider the problem of constructing a confidence interval for a *ratio* of subpopulation variances.

Lab Problems

A) A lab technician tests blood samples from 25 men and computes 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	400	228	205	338	252	

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

1) To compute confidence intervals, we need to know values like the $z_{\alpha/2}$, or $t_{\alpha/2}(\nu)$, or $\chi^2_{\alpha/2}$ in our general formulas. Our book's tables are good, but they do not contain every possible case we might need. Any needed values from the standard normal, t, χ^2, or F distributions can also be obtained via the `NormalCDF`, `TCDF`, `ChiSquareCDF`, and `FCDF` functions in the Maple package. (Recall there's on-line documentation available on the course homepage.) For instance, by "trial and error" determine an approximation of $z_{.02}$ that seems to be good to at least 4 decimal places. Compare with the best estimate you can get using the book's standard normal table.
2) Similarly, determine an approximation to $t_{.02}(10)$, and compare with the values in the text.

3) Determine an approximation to $\chi^2_{.02}(10)$ and compare with the values in the book's tables.

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:

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

4) Determine a 95% confidence interval for the population standard deviation σ for the cholesterol data.

This method is something of a *compromise* – it gives a definite procedure to get the confidence intervals, but it does not always give the *shortest possible confidence intervals.*

5) *Why* might it be of interest to get the shortest possible confidence interval $[\widehat{\theta_L}, \widehat{\theta_U}]$ where $P(\widehat{\theta_L} \leq \sigma < \widehat{\theta_U}) = 1-\alpha$?

We will next see how to select a, b (different from $b = \chi^2_{\alpha/2}$ and $a = \chi^2_{1-\alpha/2}$ as above) in order to *minimize* the length of the interval

$$\left[\sqrt{\frac{(n-1)S^2}{b}}, \sqrt{\frac{(n-1)S^2}{a}}\right]$$

subject to the constraint that

$$\int_a^b f(y)\, dy = 1-\alpha,$$

where $f(y)$ is the $\chi^2(n-1)$ pdf. It turns out that finding the appropriate a, b is equivalent to solving the two equations:

$$\begin{aligned} \int_a^b f(y)\, dy &= 1-\alpha \\ a^{n/2}e^{-a/2} - b^{n/2}e^{-b/2} &= 0 \end{aligned} \tag{2}$$

The following Maple procedure does this computation:

```
ShortestSigmaCI:=proc(alpha,n)
local a,b,eq1,eq2,g,x,s;
g:=x^((n-1)/2-1)*exp(-x/2)/(2^((n-1)/2)*GAMMA((n-1)/2));
eq1:=int(g,x=a..b)=1-alpha;
eq2:=evalf(a^(n/2)*exp(-a/2)=b^(n/2)*exp(-b/2));
s:=fsolve({eq1,eq2},{a,b},{a=0..n-1,b=n-1..infinity});
return s;
end:
```

6) Enter this procedure into your worksheet and use it to determine the a, b for the shortest confidence 95% intervals for the standard deviation with $n = 3, 4, 5, 6, 7, 8, 9, 10$. Compare with the "compromise" intervals for this α and n. How much are we gaining? (Note: n is the number of samples here – the corresponding χ^2 distribution has $n - 1$ degrees of freedom.)
7) (For after the lab) Show that the solution of (2) above does give the minimum length interval. (Hint: This is a constrained optimization problem. What method from MATH 241 applies to these?)

B) (This one will require some thought before you start plugging numbers into Maple!) 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*.
3) 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:

- Chapter 8/88,89,96,100;

Additional Problem

(A follow-up problem to part 3 of Lab Problem B above.) To determine the endpoints of the confidence interval for the ratio of variances σ_1^2/σ_2^2, you needed to work with a formula like this

$$P\left(\frac{S_1^2/S_2^2}{f_{\alpha/2}} \leq \sigma_1^2/\sigma_2^2 \leq \frac{S_1^2/S_2^2}{f_{1-\alpha/2}}\right) = 1 - \alpha.$$

In the problem from the Lab Project, you were looking for a 95% confidence interval, so you needed $f_{.025}, f_{.975}$. The F-table in the back of the book gives the values f_α for $\alpha = .1, .05, .025, .01, .005$, but not the complementary values $\alpha = .9, .95, .975, .99, .995$. As you might have wondered, *why aren't those other entries there??* Fortunately, with the `FCDF` function from our Maple package, you could compute $f_{.975}$, etc., by the same method we used for χ^2 and t distributions in the lab, so this wasn't a problem. In this question you will show that the other values could be derived from the table entries given with a little cleverness, too!

1) Let the numbers $f_\alpha(n,m)$ be defined by

$$P(Y \geq f_\alpha(n,m)) = \alpha$$

if Y has an F distribution with n numerator degrees of freedom and m denominator degrees of freedom. Show that for all $0 < \alpha < 1$, these numbers satisfy:

$$f_{1-\alpha}(n,m) = \frac{1}{f_\alpha(m,n)}$$

(note the reversal in the numbers of degrees of freedom on the right.) **Hint:** What is the standard way to get a random variable that has an F-distribution?
2) Using the table in the text and part 1, determine the number $f_{.99}(7,10)$.