Mathematics 376 – Probability and Statistics II
Group Lab Project 2 – More on Confidence Intervals
February 23, 2004

Background

We have now discussed constructing:

- Large-sample confidence intervals for means, differences of subpopulation means, fractions (or proportions) and differences of subpopulation fractions (using pivotal quantities that are standard normal)
- 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 today'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 variances, 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:

164	272	261	248	235	192	203	278	268
230	242	305	286	310	345	289	326	
335	297	328	400	228	194	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) Compute the sample mean and sample variance for these data using the appropriate procedures in the Maple package (recall there's on-line documentation available on the course homepage if you need it).
2) To compute confidence intervals, we need to know values like the $z_{\alpha/2}$ or $t_{\alpha/2}(\nu)$ 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 and t distribution can also be obtained via the `NormalCDF` and `TCDF` functions in the Maple package. For instance, by "trial and error" determine an approximation of $z_{.075}$ 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. Similarly, determine an approximation to $t_{.075}(10)$.

3) Compute 85% and 95% confidence intervals for the population mean cholesterol level using both the large-sample and small-sample formulas in each case. Discuss your results. (Include at least answers to the following questions: For each confidence level, which interval is larger and why? How do the intervals for different confidence levels compare?)
4) Next, determine a 85% confidence interval for the variance σ^2 for the cholesterol data.

B) 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) $\alpha/2$. That is in the general form

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

to get a $(1-\alpha)\times 100\%$ confidence interval, we used: $b = \chi^2_{\alpha/2}(n-1)$ and $a = \chi^2_{1-\alpha/2}(n-1)$, where, as in the notation of the text book's χ^2 tables,

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

if Y has a χ^2 distribution with some fixed number of degrees of freedom. This 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.* That is, we can often get tighter bounds by selecting different a, b in order to *minimize* the length of the interval

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

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:

```
ShortestSigma2CI:=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:
```

1) Enter this procedure into your worksheet and use it to determine the a, b for the shortest confidence 95% intervals for the variance 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.)
2) **Extra Credit** Show that the solution of (2) above does give the minimum length interval. *Hint:* The constraint $\int_a^b f(y)\, dy = 1 - \alpha$ implies that you can think of b as a function of a.

C) (This one will require quite a bit of 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.
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?