Comparing Means and Proportions
Exercise

1. Confidence intervals tell us how often our method would produce an interval that captures the true population parameter if we were to use the method many times. `cidemo2` is a user-written Stata program that allows you to draw many simple random samples to see how confidence intervals behave. From the Stata command window, type `-findit cidemo2-`. In the viewer window that appears, follow the links to install `cidemo2`.

a. By default, `cidemo2` draws 100 random samples from a normal population with mean = 50 and sd = 10. The 95% confidence intervals for each sample are displayed as an error bar graph. You specify the sample size and can change the confidence level and the number of samples. Draw 100 random samples of size $n = 50$ from a normal distribution, and display 80% confidence intervals. Type `-cidemo2 50, level(80)-`. What percent of the 100 confidence intervals capture the true mean of 50?

b. Now draw 1000 random samples size $n = 50$ and display 80% confidence intervals. What percent of the 1000 confidence intervals capture the population mean of 50? If the result differs from (a), why?

c. Set the confidence level to 95% and draw 100 random samples of size $n = 50$. Are the intervals longer or shorter than the 80% intervals? What percent of the intervals captured the true mean?

d. Increase the sample size to 500 and draw 100 random samples. What happens to the 95% confidence intervals?

2. An anthropologist suspects that color blindness is less common in societies that live by hunting and gathering than in settled agricultural societies. He tests a number of adults in two populations in Africa, one of each type. The proportion of color-blind people is significantly lower ($p < 0.05$) in the hunter-gatherer population. What additional information would you want to help you decide whether you accept the claim about color blindness?

3. Table 20.1 in Bernard (2002:563) shows data that Penn Handwerker collected on the number of children college students in the U.S. and in Liberia said they wanted to have. Use Stata to replicate and extend Bernard's analysis. The data are in bernard20-1.dta.

a. Reproduce the descriptive statistics and confidence intervals in Bernard's Table 20.2.

b. Use the `-ttest-` command to test the null hypothesis, H_0, that the sample means from students in the U.S. and in Liberia come from populations with equivalent true means. Interpret the results.

c. Plot the confidence intervals for mean number of children wanted in the two samples, and interpret the graph. What does this graph add to your interpretation (and to Bernard's in the text)?

4. Figure 11.1 in Drennan (1996:150) shows a stem-and-leaf plot for the area (in meters) of house floors for two periods (Formative and Classic).

a. Create a Stata dataset from Drennan's stem-and-leaf plot. There are 84 cases and two variables.

b. Produce the descriptive statistics in Drennan's Table 11.1.

c. Create a boxplot and a confidence interval plot of house floor area in the Formative and Classic periods. (If you want to combine the two plots into a single graph, as Drennan does, see Stata's help on `-graph combine-`). Based on these data, what can we infer about the change in house floor area between the Formative and Classic periods? State your interpretation in one clear sentence.

d. Test the null hypothesis that there is no difference in mean house floor area between the two periods. Interpret the result.

e. How would our results be affected if only half as many houses had been recovered archaeologically? Take a random sample of the data (using `-sample-` or `-swor-`) and repeat the analysis in parts (b)-(d). Interpret your results in one or two clear sentences.

5. A central question Franz Boas asked in his classic study of head form was whether descendants of immigrants who were born in the U.S. differed from their foreign-born counterparts, as a result of the change in environment. Use the Boas family data set to answer that question for the Sicilian group (hint: `-if-`).

a. Produce a table that reports the sample size, mean, standard deviation, standard error of the mean, and 95% confidence intervals for cephalic index (a ratio of head width to head length) in the U.S.-born and foreign-born samples.

b. Test the null hypothesis of no difference in the mean cephalic index of U.S.-born and foreign-born descendants of immigrants. Is there an important difference between the two subsamples?

c. What additional information would you want to have before you would accept a conclusion based on this analysis?

6. James Carey (1993) studied the distribution of culture-bound illnesses in the Nuñoa District of the southern Peruvian Andes. Carey wanted to know whether the risk of having a culture-bound illness was distributed evenly within and between communities. He conducted a household survey in three communities that were chosen to represent the range of community types in the District. The following table shows the number of cases of *manchariska* ("fright" or "soul loss") that people reported to Carey in the survey.

	Nuñoa town	Sincata ayllu	Chillihua cooperative
Manchariska	12	11	3
No *manchariska*	199	73	86

Use the `-tabchii-` command (the immediate form of `tabchi`; see Stata help) to perform a chi-square test of the differences between communities. Is there evidence that the prevalence of manchariska varies in the three communities from which the sample was drawn?