

Lab 14

Confidence Intervals and Local Baseball

This lab will use the first half of a baseball season to form proportion confidence intervals for a player's true batting ability and compare our results to batting averages from the second half of the season.

About the Data

The dataset for today's lab came from the web pages of the Anaheim Angels and the Los Angeles Dodgers. Using the "stats" sections of the web pages, the players for both teams were sorted based on number of at-bats and the top 22 players from each team were selected. The records reflect the performances of players in the regular 2002 baseball season up until the all-star break.

Today's Lab

Begin by loading the data for today's lab into *Stata* with the `use` command and get a feel for what is contained in the dataset by using the `describe` and `browse` commands.

```
. use http://www.stat.ucla.edu/labs/datasets/baseball.dta
. describe
. browse
```

You'll notice that the most common baseball statistic, batting average, is missing from our variable list. We begin by adding this variable to the dataset using the `generate` command that we've seen in earlier labs.

```
. generate bavg = hits/atbats
```

One other variable that will become useful later in this lab, is the observation number for each player, i.e. their location in the data set. We can add that variable as well.

```
. generate obnum = _n
```

To find out which players have the highest batting averages, we can sort the data and then list the top ten players:

```
. sort bavg
. list in 34/44
```

Question 1: What is the highest batting average? What about the players with these high batting averages do you think enabled them to have such an incredibly high average?

What's important to realize here, is that these batting averages do not reflect all of the at bats these players will ever have, they are merely the batting average for the sample of at bats that occurred in the first half of the 2002 season. We can use this sample information to make inferences about the population.

Often in statistics, we use confidence intervals when trying to estimate unknown population parameters by using the sample statistics we know. The reasoning behind confidence intervals is fairly intuitive. Say you were asked to estimate the height of your statistics professor. What is the probability you'd be right if you gave one precise answer? Now, what is the probability you'd be right if you answered that your professor is somewhere between five

feet and seven feet tall? Much higher likelihood of you being right when you give a range rather than one value, isn't there? Since statisticians, like most everyone else, like being right, we give an interval estimate rather than just a point estimate.

In this case, the population parameter that we're interested in estimating is the true season batting average of the 42 players in our data set. The batting average is the proportion of at-bats in which a player achieves a hit, so the population parameter we're interested in is a proportion.

Question 2: What is the equation for a confidence interval of a proportion?

Question 3: What are the lowest 10 batting averages? How might they create problems when calculating the confidence intervals for these players?

Since we have such a small sample for some of the players, we will drop all players out of the dataset that have fewer than 10 at bats and focus only on those players for whom we have more data.

```
. drop if atbats < 10
```

Using the `generate` and `invttail` commands, we will have *Stata* calculate the upper and lower bounds on the confidence intervals for each of the 44 players.

The `invttail` function in *Stata* reads in two arguments, the first is the degrees of freedom, and the second is the cumulative area under the curve from the right. The function takes in these two arguments, and returns the corresponding *t*-test statistic. Since we do not have the population standard deviation, we need to use the *t* distribution, rather than the normal *z* curve, in our confidence interval calculations. The degrees of freedom will vary for each player since it is based on the number of at bats. We want to create 95% confidence intervals, thus the area under the curve that we enter into the `invttail` function is .025. The remaining parts of the commands should look familiar as this is the same confidence interval formula you can find in

your text.

```
. generate lower = bavg - invttail(atbats-1, .025)*  
(sqrt((bavg*(1-bavg))/atbats))  
. generate upper = bavg + invttail(atbats-1, .025)*  
(sqrt((bavg*(1-bavg))/atbats))
```

To create a nice graphical display of the confidence intervals, enter the following command:

```
. twoway (rcap lower upper obnum) (scatter bavg obnum)
```

This command tells *Stata* to graph *bavg*, *lower*, and *upper* on the *y*-axis with *obnum* on the *x*-axis. The `rcap` command lets *Stata* know to connect the *lower* value to the *upper* value with an “I” shape.

Question 4: Why are some confidence intervals longer than others?

To look at the effect a larger sample has on the size of a confidence interval, sort the data by number of at bats, then create an *obnum2* variable so this order can be maintained in our plot.

```
. sort atbats  
. generate obnum2=_n
```

Question 5: How do you expect our confidence interval graph to change when we use the obnum2 variable on the x-axis instead of obnum? Why?

Check and see if the graph is as you expected.

```
. twoway (rcap lower upper obnum2) (scatter bavg obnum2)
```

To single out an individual player, use the `list` command to find their *obnum*, and then `list` again to get all information on the player. We will use D. Roberts of the Dodgers is used as an example:

```
. list obnum name  
. list if obnum==8
```

Question 6: Select a different player and state and interpret the confidence interval for that player.

Open Internet Explorer from the dock on the bottom on the screen.

If the player you selected plays for the LA Dodgers, go to:

http://losangeles.dodgers.mlb.com/NASApp/mlb/la/stats/la_sortable_player_stats.jsp

If the player you selected plays for the Anaheim Angels, go to:

http://anaheim.angels.mlb.com/NASApp/mlb/ana/stats/ana_sortable_player_stats.jsp

Click on “Historical Stats” located in the left-hand column towards the bottom of the page.

Select the 2002 season and check your player’s batting average.

Question 7: Was your player’s season batting average within the confidence interval you created earlier?

Question 8: At the end of the 2002 season, how many players do we expect to have a batting average outside of the confidence intervals we just created? Why?

Assignment

The slugging percentage in baseball is defined as the average number of bases achieved per at bat. Generate this statistic for each player and perform analysis similar that used earlier, to answer the following questions.

Question 9: What is the highest slugging percentage in the dataset? Who is the player who achieved this average?

Question 10: What position does the Dodger player with the highest slugging percentage play? What position does the Angel with the highest slugging percentage play?

Question 11: Print out the confidence intervals for slugging percentage. For which players is the interval widest? Which player has the highest lower bound on his confidence interval?

Question 12: What is the average width of the confidence intervals for slugging percentage? Which team has the higher average interval width and why? (Hint: You'll need to use the `generate`, `sort`, and the `summarize` commands.)