POL 345: Quantitative Analysis and Politics

Precept Handout 7

Week 8 (Verzani Chapter 8: 8.1-8.5)

Remember to complete the entire handout and submit the precept questions to the Blackboard 24 hours before precept. In this handout, we cover the following new materials:

- Using `qnorm()` to calculate critical values ($z_{p/2}$)
- Using `cbind()` to combine multiple vectors into a matrix

1 Bias and Efficiency of Estimators

A good estimator of a parameter has a sampling distribution that (1) is centered around the true value of the parameter (unbiasedness) and (2) has a small standard deviation (efficiency).

2008 Presidential Election Polls. Given the media's constant reporting of polling numbers during campaigns and policy debates, and our reliance on them to predict elections, we would like to know whether these estimates are unbiased. We begin by considering only the most recent polls (those conducted within 60 days of the election). For each state, we compute the estimation error as the difference between the poll estimated margin of victory and the actual margin of victory. The *estimated* bias is the sample average of these estimation errors across the states.

We also care about how far off, on average, these polls are from the truth. So for each state we compute the squared differences between the poll estimated margin of victory and the actual margin. Recall from lecture the formula for *mean squared error* or MSE: $\mathbb{E}[(\hat{\theta} - \theta)^2]$.

```
> load("e08.RData")
> e08 <- e08[e08$DaysToElection <= 60, ]
> e08$margin <- e08$Dem - e08$GOP
> polls <- e08[e08$DaysToElection != 0, ]
> results <- e08[e08$DaysToElection == 0, ]
> O.margin.polls <- tapply(polls$margin, polls$State, mean)
> O.margin.actual <- results$margin

> error <- O.margin.polls - O.margin.actual
> bias <- mean(error)
> bias

[1] -3.07728
```

The negative bias indicates that our method, on average, produced estimates of Obama's margin of victory that were too low. This bias was relatively small given the scale, but might have led to our errant prediction from two years ago that Indiana and Missouri would likely go for McCain.

2 Confidence Intervals Using the Normal Distribution

The formula for the $(1 - p)\%$ confidence interval is given by,

$$[\hat{\theta} - \text{s.e.} \times z_{p/2},\ \hat{\theta} + \text{s.e.} \times z_{p/2}]$$

where $z_{p/2}$ is the critical value and can be computed using the `qnorm()` function. Recall that we are relying on the approximate (asymptotic) distribution of the sample mean, based on the central limit theorem.

- The `qnorm(p, mean, sd)` function returns the 100p-th percentile of the normally distributed random variable, with the mean and standard deviation equal to `mean` and `sd`, respectively. Setting `lower.tail` to `FALSE` will return the $100(1 - \mathtt{p})$-th percentile instead.

- `cbind(x,y,z,...)` will put multiple vectors of the same length side by side and combine them into one large matrix.

- Here we will look at the polling data again and find the resulting confidence interval using a normal distribution. See page 7 of the Statistical Inference lecture slides.

  ```
  > e08$Dem2p <- e08$Dem/(e08$Dem + e08$GOP)
  > polls <- e08[e08$DaysToElection != 0, ]
  > results <- e08[e08$DaysToElection == 0, ]
  > ## Obama Support by State
  > M <- summary(polls$State) ## number of polls per state
  > Obama.2p <- tapply(polls$Dem2p, polls$State, mean) ## sample means
  > st.e <- sqrt(Obama.2p*(1-Obama.2p)/M*1000) ## standard deviation
  > MoE <- qnorm(0.975)*st.e ## margin of error
  > lower <- Obama.2p - MoE
  > upper <- Obama.2p + MoE
  > cbind(lower, upper)[1:10,] ## Confidence intervals
  ```

  ```
                  lower    upper
  Alabama      -9.61201 10.35527
  Alaska       -9.21855 10.03016
  Arizona      -8.85481  9.77917
  Arkansas    -12.12494 13.00980
  California   -8.17234  9.36926
  Colorado     -5.03549  6.08351
  Connecticut -11.84740 13.03003
  D.C.        -20.43799 22.16431
  Delaware    -10.90383 12.09413
  Florida      -3.91873  4.93440
  ```

3 Coverage Probability of Confidence Intervals

In repeated sampling, we expect our confidence intervals to contain the true value of the parameter a set proportion of the time. For example, ninety-five percent of the 95% CIs should contain the true value. Biased confidence intervals (based on biased standard error and/or biased estimate), however, will in repeated sampling contain the true value less often than expected given the size of the intervals.

We will conduct a series of national surveys to gauge support for Obama. One set of surveys will poll people from all states while a second set of polls will skip over residents from low-population states. Note that while we use the command `sd()`, the results yield the same results as using the standard deviation formula for proportions.

```
> load("election.RData")
> obama <- election$obama  ## proportion of support in each state
> pop.obama <- election$pop  ## voters in each state
> ## Excluding low-population states
> biasedobama <- election$obama[election$pop>=1000000]
> pop.biasedobama <- election$pop[election$pop>=1000000]
```

```
> ## Calculate national mean support for Obama
> obama.mean <- weighted.mean(election$obama, election$pop)
> n.sims <- 5000
> in.interval.ub <- rep(NA, n.sims)
> in.interval.b <- rep(NA, n.sims)
> survey.size <- 1000
> for (i in 1:n.sims){
+   ub.obama <- sample(obama, survey.size, replace=TRUE, prob=pop.obama)
+   low <- mean(ub.obama) - qnorm(0.975)*sd(ub.obama)/sqrt(survey.size)
+   hi <- mean(ub.obama) + qnorm(0.975)*sd(ub.obama)/sqrt(survey.size)
+   in.interval.ub[i] <- ifelse(low <= obama.mean & obama.mean <= hi, 1, 0)
+
+   ## Subset of states with population above 1 million
+   b.obama <- sample(biasedobama, survey.size, replace=TRUE, prob=pop.biasedobama)
+   low2 <- mean(b.obama) - qnorm(0.975)*sd(b.obama)/sqrt(survey.size)
+   hi2 <- mean(b.obama) + qnorm(0.975)*sd(b.obama)/sqrt(survey.size)
+   in.interval.b[i] <- ifelse(low2 <= obama.mean & obama.mean <= hi2, 1, 0)
+ }
> ## Proportion of times confidence interval contains truth
> mean(in.interval.ub) ## unbiased estimator

[1] 0.95

> mean(in.interval.b) ## biased estimator

[1] 0.9074
```

4 Precept Questions

In preparation for precept, please answer the following questions, which are based on the 2008 election example given above. Submit your code following the instructions given in the syllabus.

1. Create a scatter plot where the point estimate of the margin of Obama's victory for each state is plotted against the actual margin of victory. Add the 45 degree line (as a dashed line) to the plot. Also, add the 95% confidence intervals to each point in the plot using a vertical line. **Hint:** Using a loop will be helpful for adding the confidence intervals.

2. In the above example, we calculated the bias of our poll-based estimates for Obama's margin of victory. Begin by calculating the proportion of 95% confidence intervals (centered around the biased poll-based estimates) that contain the actual election result. Next, shift each confidence interval by the estimated bias to produce bias-adjusted confidence intervals (i.e.: centered around the point estimate free of overall bias). Compute the proportion of these new confidence intervals that contain the election result. Give a brief interpretation.