

## Statistics 406 Problem Set 8

Due in lab, Tuesday November 21st

1. Calculate the poll averages and standard deviations for the House races, using all polls occurring after day 290. Then calculate a “Z-score” for each poll average. To do this, subtract the actual margin (D-R) from the poll margin, and divide by the standard deviation of the poll margin. Describe any similarities or differences between the Z-scores and an actual standard normal distribution. Then repeat the analysis using only a single poll for each race (selected from those occurring after day 290). Compare the two groups of polls based on the variances of their Z-scores.

You can use the following code to get started.

```
## Actual house results.
R <- read.table('house_results.dat')

## House polls.
P <- read.table('house_polls.dat')

## Convert the actual results to percentages.
T <- R[,2] + R[,3]
R[,2] <- 100 * R[,2] / T
R[,3] <- 100 * R[,3] / T

## The actual margin.
R <- cbind(R, R[,2]-R[,3])

## Get a list of the districts with polling data.
D <- unique(as.vector(P[,2]))

## A table of results.
Z <- NULL
D1 <- NULL

## Loop through the races.
for (d in D)
{
  ## Get the polls for race d.
  i1 <- which(P[,2] == d)

  ## Get the recent polls (roughly three weeks back).
  i1 <- i1[P[i1,1] > 290]
```

```

    if (length(i1) == 0) { next }

    ## The row of R for the race d.
    i2 <- which(R[,1] == d)
}

```

### Solution:

```

## Actual house results.
R <- read.table('house_results.dat')

## House polls.
P <- read.table('house_polls.dat')

## Convert the actual results to percentages.
T <- R[,2] + R[,3]
R[,2] <- 100 * R[,2] / T
R[,3] <- 100 * R[,3] / T

## The actual margin.
R <- cbind(R, R[,2]-R[,3])

## Get a list of the districts with polling data.
D <- unique(as.vector(P[,2]))

## A table of results.
Z <- NULL
D1 <- NULL

## Loop through the races.
for (d in D)
{
    ## Get the polls for race d.
    i1 <- which(P[,2] == d)

    ## Get the recent polls (roughly three weeks back).
    i1 <- i1[P[i1,1] > 290]
    if (length(i1) == 0) { next }

    ## Uncomment this line to work with a single poll, rather than
    ## the poll average.
    i1 <- i1[1]
}

```

```

## The row of R for the race d.
i2 <- which(R[,1] == d)

## The difference of poll averages for the two candidates.
mp <- mean(P[i1,3]) - mean(P[i1,4])

## The standard deviation for the poll average difference.
sd <- 3.5/sqrt(length(i1))

Z <- c(Z, (mp - R[i2,4])/sd)
D1 <- c(D1, d)
}

```

I got  $SD(Z) = 1.93$  for poll averages, and  $SD(Z) = 1.71$  for single polls. The value would be 1 for perfectly calibrated polls. The difference between the single poll and poll average Z-score variance is probably not meaningful. It seems that both the poll averages and the individual polls are somewhat mis-calibrated. The biases are -0.26 (poll averages), and 0.52 (single polls) which are small and inconsistent. This indicates that the inflated Z-score variance is due to roughly equal errors in both directions. Also, I tried deleting the most extreme Z-scores (both positive and negative). In doing this, the variance does not approach 1 until roughly 8 values (20% of the total) are removed. This suggests that the inflated Z-score variance is not due to a small number of outliers, but rather is due to a more systematic under-reporting of uncertainty across the polls.

2. Suppose as a simple model that a person's political views are summarized by two "factors"  $X$  and  $Y$ . For example,  $X$  could represent views on social issues while  $Y$  represents views on economic issues. Suppose that these are only mildly correlated, say  $\text{cor}(X, Y) = 0.2$ , and for simplicity assume that both  $X$  and  $Y$  follow standard normal distributions. Suppose that a voter makes a decision to vote for candidate  $A$  with probability

$$\frac{\exp(c_a X + d_a Y + E)}{1 + \exp(c_a X + d_a Y + E)},$$

and votes for candidate  $B$  otherwise. Here the term  $E$  represents political attitudes that are not related to either factor  $X$  or factor  $Y$ , and  $c_a$  and  $d_a$  are constants that reflect how support for candidate  $A$  is related to the two factors. For example, if  $c_a = 0$  then attitudes toward factor  $X$  are not associated with support for candidate  $X$ , whereas if  $c_a > 0$  then people with higher scores on factor  $X$  tend to favor candidate  $A$ .

Suppose that due to a subtle bias in polling, people with high  $X$  values are under-represented in the poll. To model this, assume that anyone who is sampled for inclusion in the poll

but that has  $X > 1$  is randomly excluded with probability  $\theta$ . If a person is excluded, then generate another person as a replacement (if the replacement has  $X > 1$  then repeat the random selection, possibly leading to a third candidate being considered, and so on).

Carry out a simulation study to assess how the coverage probability of a 95% CI for the support level for a single candidate (in a two-candidate election) is affected by this type of biased sampling. Assume that  $c_a = 0.5$ ,  $d_a = 0.3$ ,  $E$  is standard normal, and that the poll sample size is 1000. Also use  $\hat{\sigma} = 1/2$  in forming the CI rather than the actual sample SD. Compare your results for  $\theta = 0, 0.2$ , and  $0.5$ .

The population structure used here reflects an exact tie in the population. Therefore the coverage probability will be the proportion of the time that the CI contains  $1/2$ .

Use the following code to simulate  $X$  and  $Y$  so that the correlation between them is 0.2:

```
X <- rnorm(1)
Y <- 0.2*X + sqrt(1 - 0.2^2)*rnorm(1)
```

### Solution:

```
ca <- 0.5
cb <- 0.3

V <- NULL
P <- array(0, 1000)

## Exclusion probability for people with X>1.
theta <- 0.2

for (rep in 1:500)
{
  ## Generate the election data for one election.
  for (k in 1:1000)
  {
    while (1)
    {
      X <- rnorm(1)
      Y <- 0.2*X + sqrt(1 - 0.2^2)*rnorm(1)
      E <- rnorm(1)
      if (X < 1) { break }
      else if (runif(1) < 1-theta) { break }
    }
  }
}
```

```

    }

    ## The voter-level probabilities for the first candidate.
    P[k] <- exp(X + Y + E) / (1 + exp(X + Y + E))
  }

  ## The actual election result (proportion favoring the first candidate).
  V[rep] <- mean(runif(1000) < P)
}

## Check the proportion of the CI's that cover the true value (1/2).
w <- 1.96/sqrt(4000)
C <- mean( (V-w<0.5) & (V+w>0.5) )

```

I get coverage probabilities of 0.95, 0.89, and 0.84 for  $\theta = 0, 0.2, 0.5$ , respectively. Since people with high  $X$  values are more likely to support the first candidate, excluding some of them from polls creates biased poll results. The resulting CI's have the correct width, but are centered at the wrong value. Therefore their coverage is less than the nominal 95%.