Next Up Previous
Next: R Scripts **Up:** Assignments **Previous:** Assignments

Project 1

1. During each election year we are innundated with the results of political preference polls. The media presents the results of these polls with little or no discussion of their accuracy. TV news reports typically present a set of poll percentages as if they were the actual population percentages, while newspapers may include a small-font footnote stating that the results have an error of $\pm 3-5\%$. Such statements are based on the conservative bound on the s.d. of a sample proportion and ignore several issues that can impact the results significantly. This project is intended to examine some of those issues.

We will consider a situation in which 3 candidates are running for a political office and a poll is conducted which randomly selects a sample of size n from the population of *eligible* voters and then obtains the sample percentages for each candidate. Note that the percentages for each candidate are not independent, and so confidence intervals for each percentage would not be independent. Suppose that the poll asks respondents how likely they are to vote and then labels each respondent as either *likely to vote* or *not likely to vote*. The final poll report is based only on those who are labeled as *likely to vote*.

There are two types of populations to be examined.

1. The preferences of those who claim they will vote but don't vote and those who actually vote are the same. Let (p_1, p_2, p_3) denote the population proportions who prefer candidates $1, 2, 3$, respectively. Examine each of the following situations.
   All groups:

   a) $(p_1, p_2, p_3) = (0.50, 0.30, 0.20)$.

   b) $(p_1, p_2, p_3) = (0.50, 0.48, 0.02)$.

2. The preferences of those who claim they will vote but don't vote and those who actually vote are not the same. Let (p_1, p_2, p_3) denote the population proportions of those who actually vote, and let (r_1, r_2, r_3) denote the population proportions of those who claim they will vote when contacted, but don't actually vote during the election. Let *q* denote the proportion of those labeled as likely to vote who do not vote. Then the simulated sample will be a mixture from these two subpopulations. Use the following sets of proportions.

   Group 1: $(p_1, p_2, p_3) = (0.40, 0.30, 0.30)$, $(r_1, r_2, r_3) = (0.34, 0.33, 0.33)$.

   Group 2: $(p_1, p_2, p_3) = (0.46, 0.44, 0.10)$, $(r_1, r_2, r_3) = (0.34, 0.33, 0.33)$.

   Group 3: $(p_1, p_2, p_3) = (0.50, 0.25, 0.25)$, $(r_1, r_2, r_3) = (0.10, 0.45, 0.45)$.

   Group 4: $(p_1, p_2, p_3) = (0.40, 0.30, 0.30)$, $(r_1, r_2, r_3) = (0.25, 0.25, 0.50)$.

For each type of population, simulate sample sizes of *N = 500, 1000, 1500* for the polls and use 2000 replications for the simulations. For part 2, use each of the mixing probabilities, $q = 0.10, 0.20, 0.30, 0.40, 0.50$. Assume that all individuals in these samples responded that they intended to vote. Construct simultaneous 95% confidence intervals for $(p_1 - p_2, p_1 - p_3, p_2 - p_3)$ (see below) and estimate the probability that a confidence interval contains the respective difference for each situation. For the second part, generate the appropriate mixture of the two subpopulations, but the confidence intervals should be computed the same way as for the first part. In all cases, use both confidence interval methods given below on the same set of simulated data. Summarize your results, including a comparison of the two methods.

Simulation details. The first part involves using *rmultinom()* to generate random samples with the appropriate probabilities. Note that

$$M = \text{rmultinom}(nrep, N, p)$$

returns a matrix with *m* rows and *nrep* columns, where *m* is the length of *p* (number of groups), *N* is the sample size, and *nrep* is the number of samples that are simulated. Each column of that matrix corresponds to a simulated sample of counts for the *m* groups, the sum of each column equals *N*, and the sample proportions are obtained by dividing this matrix by *N*.

The second part requires the following steps.
1. Generate *nrep* binomial r.v.'s with success probability *q* and sample size *N*. This vector gives sample sizes for those who don't vote and have preferences given by (r_1, r_2, r_3). Subtract the binomial vector from *N* to obtain sample sizes for those who do vote and so have preferences given by (p_1, p_2, p_3).

2. Since each multinomial sample will have different sample sizes, you will need to setup two blank matrices the same size as in the first part, one matrix (*V*) for those who vote and one matrix (*NV*) for those who don't vote. Then use a for-loop to generate the individual samples (columns of the respective matrices) like was done in part 1 but with the appropriate sample sizes and preferences. The final matrix that represents what would be the poll data is the sum of these two matrices, *M = V + NV*.

Simultaneous Confidence Intervals for Differences Between Multinomial Proportions

Suppose a population contains m types of individuals, and suppose p_i, $1 \le i \le m$ represents the proportion of the population that is type i. Let *N* denote the size of a randomly selected sample from this population, let $\Delta_{i,j} = p_i - p_j$, $\delta_{i,j} = \hat{p}_i - \hat{p}_j$. Note that there are $M = m(m-1)/2$ pairwise differences between proportions.

Method 1. A set of approximate $1 - \alpha$ confidence intervals for $\Delta_{i,j}$, $1 \le i < j \le m$, is

$$\delta_{i,j} \pm \sqrt{\frac{Ad_{i,j}}{N}},$$

where

$$\begin{aligned} A &= \chi^2_{M-1}(\alpha/M), \\ d_{i,j} &= \hat{p}_i + \hat{p}_j - (\delta_{i,j})^2, \end{aligned}$$

and $\chi^2_n(\alpha)$ denotes the $1-\alpha$ quantile from the chi-square distribution with n d.f. This critical value can be obtained in **R** by

```
cv = qchisq(1-alpha,n)
```

Method 2. The confidence intervals are given by

$$\delta_{i,j} \pm \frac{a}{\sqrt{N}},$$

where a is the solution to the equation

$$1 - 2[1 - \Phi(a)] - 4[m-2][1 - \Phi(a\sqrt{2})] = 1 - \alpha,$$

and Φ is the standard normal d.f. This equation can be solved in **R** by

```
m = 3
alpha = .05
x = seq(1,4,length=1000)
y = 1 - 2*(1 - pnorm(x)) - 4*(m-2)*(1-pnorm(x*sqrt(2)))
a = min(x[y >= 1 - alpha])
#plot to show this
plot(x,y,type="l")
abline(h = 1-alpha,col="red")
abline(v = a,col="red")
```

Note. Since we wish to obtain simultaneous confidence intervals for each pairwise difference, then *alpha* represents the probability that at least one of the confidence intervals does not contain the true value of the corresponding population difference. Specifically, let *CI[i,j]* denote the confidence interval for

$$\Delta_{i,j} = p_i - p_j, \ 1 \le i < j \le m.$$

Then the goal is:

$$P(\Delta_{1,2} \in CI[1,2] \cap \cdots \cap \Delta_{m-1,m} \in CI[m-1,m]) = 1 - \alpha.$$

Reference

Piegorsch, W.W. and Richwine, K.A. (2001) *Large-Sample Pairwise Comparisons Among Multinomial Proportions with an Application to Analysis of Mutant Species*. **J. Agricultural, Biological, and Environmental Statistics 6**, 3, pp 305-325.

Next Up Previous

Next: R Scripts **Up:** Assignments **Previous:** Assignments

ammann
2018-09-28