Announcements

Make sure to sign in on the google form (I send a list of which section questions are useful for which pset questions afterwards)

Pset 9 due Friday 4/16

Optimal Polling

When conducting political polls, a reasonable goal is to construct the most accurate and precise estimate of public opinion while contacting as few people as possible. Each person you contact will require some amount of time and labor, so minimizing this count is preferable. In this problem, we will be looking at binary voter support of some candidate where voter i has the indicator y_i which is 1 if the voter supports the candidate and 0 otherwise. We assume that everyone in the size N population has an opinion about the candidate, so the y_i are not random, but which of these we actually sample is random. We want to estimate $\mu = \frac{1}{N}\sum_{i=1}^{N} y_i$, the average support for the candidate.

In the United States, 31 states, the District of Columbia, and the U.S. Virgin Islands allow voters to indicate their partisan affiliations on voter registration forms and also report these total registration numbers publicly. Suppose for this problem that someone can only be a registered Democrat, registered Republican, or independent, and suppose we have information on which of these groups each person is. Note that someone's voter registration does not constrain whom the voter can vote for. For example, a registered Democrat can vote for a Republican.

1. Suppose we are in a state that makes this voter registration public so we can look up someone's party affiliation if the person has one. Explain intuitively why it might be suboptimal to take a simple random sample of voters, contact them, and average their opinions on the candidate.

If we know in advance the person's party affiliation, we might be able to guess the voter's opinion on the candidate and not bother contacting him or her.

2. Let μ be the support for the candidate in the whole state. Let $\mu_1 = \frac{1}{N_1}\sum_{i:\text{Voter i is a Democrat}} y_i$ be the support for the candidate among all registered Democrats, μ_2 be the support among all registered Republicans, and μ_3 be the support among all independents. Let p_1 be the proportion of Democrats in the state, p_2 be the proportion of Republicans, and p_3 be the proportion of independents. We treat these p as known since we can look up the registered proportions. Suppose a random sample of n people is taken without replacement from the state and the people are contacted about their opinions on the candidate. Let $Y_1, ..., Y_n$ be the opinions reported in the sample. Find the bias and variance of

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

in terms of the μs and ps. Recall that the variance of the sample average with a finite population correction is $\frac{\text{Var}(Y_1)}{n}\frac{N-n}{N-1}$.

The bias is

$$E(\hat{\mu}) - \mu = \left(\frac{1}{n}\sum_{i=1}^{n} E(Y_i)\right) - \mu = \left(\frac{1}{n}\sum_{i=1}^{n} \mu\right) - \mu = 0$$

Letting $X_{i,k}$ be the indicator of voter i being registered for party k, the variance is

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}) &= \frac{\mathrm{Var}(Y_1)}{n}\frac{N-n}{N-1}\\
&= \frac{E(\mathrm{Var}(Y_1|X_{1,1},X_{1,2},X_{1,3}))+\mathrm{Var}(E(Y_1|X_{1,1},X_{1,2},X_{1,3}))}{n}\frac{N-n}{N-1}\\
&= \frac{E(\mu_1(1-\mu_1)X_{1,1}+\mu_2(1-\mu_2)X_{1,2}+\mu_3(1-\mu_3)X_{1,3})+\mathrm{Var}(\mu_1X_{1,1}+\mu_2X_{1,2}+\mu_3X_{1,3})}{n}\frac{N-n}{N-1}\\
&= \frac{p_1\mu_1(1-\mu_1)+p_2\mu_2(1-\mu_2)+p_3\mu_3(1-\mu_3)+\mu_1^2p_1+\mu_2^2p_2+\mu_3^2p_3-(\mu_1p_1+\mu_2p_2+\mu_3p_3)^2}{n}\frac{N-n}{N-1}\\
&= \frac{p_1\mu_1+p_2\mu_2+p_3\mu_3-(\mu_1p_1+\mu_2p_2+\mu_3p_3)^2}{n}\frac{N-n}{N-1}\\
&= \frac{\mu(1-\mu)}{n}\frac{N-n}{N-1}
\end{aligned}$$

where the second equality is by Eve's law and the variance in the fourth equality is solved by noting that $X_jX_k=0$ for $j\neq k$. You could also just get the last result immediately.

3. Now, consider the stratified estimator
$$\tilde{\mu}=\sum_{l=1}^{3}p_l\bar{Y}_l$$
with sample size n_l for strata l. Find its bias and variance in terms of the μs, ps, and ns.

$$E(\tilde{\mu})-\mu=\left(\sum_{l=1}^{3}p_l\mu_l\right)-\mu=\left(\sum_{l=1}^{3}\frac{N_l}{N}\mu_l\right)-\mu=0$$

Since the stratified means are independent,

$$\begin{aligned}
\mathrm{Var}(\tilde{\mu}) &= \sum_{l=1}^{3}p_l^2\mathrm{Var}(\bar{Y}_l)\\
&= \sum_{l=1}^{3}p_l^2\frac{\sigma_l^2}{n_l}\frac{N_l-n_l}{N_l-1}\\
&= \sum_{l=1}^{3}p_l^2\frac{\sigma_l^2}{n_l}\frac{Np_l-n_l}{Np_l-1}\\
&= \sum_{l=1}^{3}p_l^2\frac{\mu_l(1-\mu_l)}{n_l}\frac{Np_l-n_l}{Np_l-1}
\end{aligned}$$

4. Using the fact that the optimal subsample size is $n_l/n\propto N_l\sigma_l$, find n_l as a function of n, the N_ls, and the μ_ls.

We need a k to satisfy

$$n_l=knN_l\sqrt{\mu_l(1-\mu_l)}$$
$$\sum_{l=1}^{3}n_l=n$$

Therefore,

$$\sum_{l=1}^{3}knN_l\sqrt{\mu_l(1-\mu_l)}=n\implies k=\frac{1}{\sum_{l=1}^{3}N_l\sqrt{\mu_l(1-\mu_l)}}$$

and our sample sizes will be

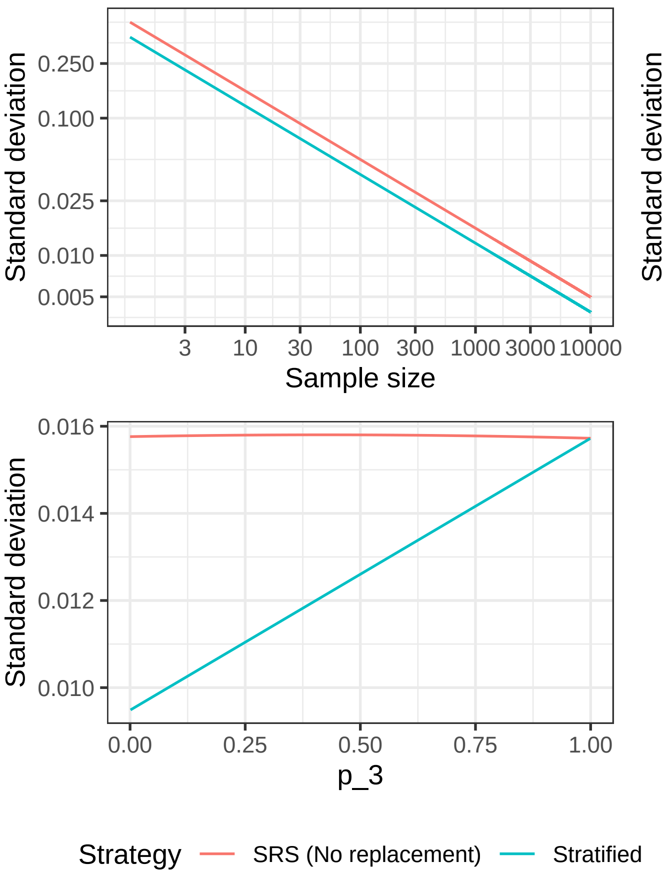
$$n_l = \frac{nN_l\sqrt{\mu_l(1-\mu_l)}}{\sum_{j=1}^{3} N_j\sqrt{\mu_j(1-\mu_j)}}$$

You could also just note that the answer should be n times $N_l\sigma_l$ divided by the sum of $N_j\sigma_j$s.

5. Explain why the answer above is still useful even though we have μ_ls in it.

Normally, we do not want estimands in estimators or values that should be functions of only the data. However, if we had a pilot survey to estimate μ_l for each political group or if we had historical data on a similar candidate, we could guess a μ_l and use it throughout. If we guess wrong, we don't induce any bias; we just give up some precision.

6. For $N = 10^6, n = 1000, p_1 = 0.25, p_2 = 0.3, p_3 = 0.45, \mu_1 = 0.9, \mu_2 = 0.1, \mu_3 = 0.55$, the following plot shows the standard deviation of each estimator. One variable at a time is varied in each plot. Explain why these results make sense.



As we include more people in our sample, the standard deviation should continue to fall. When the independents are most split ($\mu_3 = 0.5$), we have the highest standard error because we have the most variability in responses. When the proportion of independents is near 1, we are essentially only randomly sampling from the independents in either scheme. When the proportion of independents is near 0, we sample optimally from the Democrats and Independents and get a low variability since they are more consistent in their voting.

7. Create a Horvitz-Thompson estimator for the SRS case and stratified sampling case. Are either equivalent to the corresponding estimators above?

The Horvitz-Thompson estimator is $\hat{\tau} = \sum_{i:y_i \in S} \frac{y_i}{\pi_i}$ where π_i is the probability of inclusion. In the SRS case, this is $\hat{\tau} = \sum_{i:y_i \in S} \frac{y_i}{n/N} = N\bar{Y}$. Dividing by N to estimate the mean rather than the sum gives $\bar{Y}$, which is equivalent to at the beginning.

In the stratified case, this is $\hat{\tau} = \sum_{i:y_i \in S} \frac{y_i}{n_{l,i}/N_{l,i}}$ where $n_{l,i}$ is the n_l that corresponds to the party affiliation of the i^{th} person. Since there are n_l in each group, we have

$$\hat{\tau} = \sum_{l=1}^{3} \frac{N_l}{n_l} \sum_{i:y_i \in S, \text{Voter i in group l}} y_i = \sum_{l=1}^{3} \frac{N_l}{n_l} \bar{Y}_l n_l = \sum_{l=1}^{3} N_l \bar{Y}_l = N \sum_{l=1}^{3} p_l \bar{Y}_l$$

so dividing by N to estimate the mean rather than the total gives the same $\tilde{\mu} = \sum_{l=1}^{3} p_l \bar{Y}_l$ from above.

Realistic conjugacy

In Normal-Normal conjugacies, we often assume both the group mean's variance and the observation's variance are known. However, there are very few examples where this is actually the case. To ameliorate this, we can either switch to a more complicated conjugate prior (e.g. Normal-Gamma) or use asymptotics. Here, we will do the latter. Suppose we have $\mu_i \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $Y_{i,j} \sim [\mu_i, \sigma_i^2]$ i.i.d. conditional on μ_i. This notation means $Y_{i,j}$ has mean μ_i and variance σ_i^2 conditional on μ_i and σ_i^2, but it does not place distributional assumptions on $Y_{i,j}$.

1. Conditional on μ_j, show that

$$\sqrt{n_i}\left(\frac{\bar{Y}_i - \mu_i}{\hat{\sigma}_i}\right) \to \mathcal{N}(0,1)$$

where $\bar{Y}_i = \frac{1}{n_i}\sum_{i=1}^{n_i} Y_{i,j}$ and $\hat{\sigma}_i^2$ is the sample variance of the $Y_{i,j}$.

Since the sample variance is a consistent estimator for σ_i^2, $\hat{\sigma}_i^2 \xrightarrow{p} \sigma_i^2$, so by the continuous mapping theorem $\frac{\hat{\sigma}_i}{\sigma_i} \xrightarrow{p} 1$. By the Central Limit Theorem,

$$\sqrt{n_i}\left(\frac{\bar{Y}_i - \mu_i}{\sigma_i}\right) \to \mathcal{N}(0,1)$$

and

$$\sqrt{n_i}\left(\frac{\bar{Y}_i - \mu_i}{\sigma_i}\right) = \sqrt{n_i}\left(\frac{\bar{Y}_i - \mu_i}{\hat{\sigma}_i}\right)\frac{\hat{\sigma}_i}{\sigma_i} \xrightarrow{p} \sqrt{n_i}\left(\frac{\bar{Y}_i - \mu_i}{\hat{\sigma}_i}\right)$$

by Slutsky's theorem, so

$$\sqrt{n_i}\left(\frac{\bar{Y}_i - \mu_i}{\hat{\sigma}_i}\right) \to \mathcal{N}(0,1)$$

2. Use the approximate distribution of $\bar{Y}_i$ to write an approximate Normal-Normal conjugacy.

The conjugacy is:

$$\bar{Y}_i \dot{\sim} \mathcal{N}\left(\mu_i, \frac{\hat{\sigma}_i^2}{n_i}\right)$$
$$\mu_i \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

3. Suppose we observed $Y_{i,1}, ..., Y_{i,n_i}$. Find the posterior distribution for μ_i.

Using the Normal-Normal conjugacy results,

$$\mu_i | \vec{Y}_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$$
$$\sigma_1^2 = \frac{1}{\frac{1}{\sigma_0^2} + \frac{n_i}{\hat{\sigma}_i^2}}$$
$$\mu_1 = \sigma_1^2\left(\frac{\mu_0}{\sigma_0^2} + \frac{n_i \bar{Y}_i}{\hat{\sigma}_i^2}\right)$$

4. Recall Stein's theorem that for $Y_j \sim \mathcal{N}(\mu_j, \sigma^2)$ independent for $j = 1, ..., K$ with $K \geq 3$, the μ_j unknown, and σ^2 known, for squared loss $\sum_{j=1}^K (\mu_j - \hat{\mu_j})^2$, the MLE $\vec{Y}$ is inadmissible. The James Stein estimator
$$\hat{\mu}_{JS,j} = \left(1 - \frac{(K-2)\sigma^2}{\sum_{j=1}^K Y_j^2}\right) Y_j$$
has strictly lower risk. However, as before, we run into the complication that σ^2 is almost never known, and there's no reason to think the variances would be equal. Using the approximate distribution from above, find the James Stein estimator for $\vec{\mu}$. The estimator should make no assumptions about the variances or sample sizes being equal.

Using the approximation from 2,

$$\frac{\bar{Y}_i - \mu_i}{\sqrt{\hat{\sigma}_i^2/n_i}} \dot{\sim} \mathcal{N}(0,1) \implies \frac{\bar{Y}_i}{\sqrt{\hat{\sigma}_i^2/n_i}} \dot{\sim} \mathcal{N}\left(\frac{\mu_i}{\sqrt{\sigma_i^2/n_i}}, 1\right)$$

Matching the formula above, we get

$$\frac{\hat{\mu}_{JS,i}}{\sqrt{\sigma_i^2/n_i}} = \left(1 - \frac{(K-2)}{\sum_{j=1}^K \left(\frac{\bar{Y}_i}{\sqrt{\hat{\sigma}_i^2/n_i}}\right)^2}\right) \frac{\bar{Y}_i}{\sqrt{\hat{\sigma}_i^2/n_i}} \implies \hat{\mu}_{JS,i} \approx \left(1 - \frac{(K-2)}{\sum_{j=1}^K \left(\frac{\bar{Y}_i}{\sqrt{\hat{\sigma}_i^2/n_i}}\right)^2}\right) \bar{Y}_i$$

5. The following simulation compares the two estimators on a small set of exponential random variables with the hierarchical mean added. Note that the only assumption made here was that the mean parameter is distributed Normally, not that the individual observations are also Normal, that they have the same variance, or that they have the same sample size. Remarkably, this works even with all the n less than 10, quite far from $n = \infty$.

```r
mu_0 <- 5
sigma_0 <- 3
nsims <- 10^5
K <- 10
ns <- c(8,9,7,9,8,9,7,8,7,9)

naive_loss <- vector(length = nsims)
js_loss <- vector(length = nsims)
for (i in 1:nsims) {
  mus <- rnorm(K, mu_0, sigma_0)
  mu_hat_naive <- vector(length = K)
  ybars <- vector(length = K)
  sigma_sqs <- vector(length = K)
  for (j in 1:K) {
    ys <- mus[j] + rexp(ns[j], 1/j) - j # Mean mu_j
    mu_hat_naive[j] <- mean(ys)
    ybars[j] <- mean(ys)
    sigma_sqs[j] <- var(ys)
  }
  # Compute JS
  mu_hat_js <- (1 - (K-2) / sum((ybars/(sqrt(sigma_sqs/ns)))^2)) * ybars

  # Get squared losses for both
  naive_loss[i] <- sum((mu_hat_naive - mus)^2)
  js_loss[i] <- sum((mu_hat_js - mus)^2)
}
```

```
c("Naive Risk" = mean(naive_loss), "James Stein Risk" = mean(js_loss))
```

```
##       Naive Risk James Stein Risk
##         48.51810         46.82261
```