Stat 135, Fall 2006 A. Adhikari

HOMEWORK 2 SOLUTIONS

1. In the 0-1 case, the sample proportion **is** the sample mean. So by (i) the variance of the sample proportion is $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$. By (iii), an unbiased estimate of this quantity is

$$\frac{s_X^2}{n} \cdot (1 - \frac{1}{N}) \cdot \frac{N-n}{N-1} \quad = \quad \frac{s_X^2}{n} \cdot \frac{N-n}{N}$$

By (ii), this is

$$\frac{\hat{\sigma}_X^2}{n} \cdot \frac{n}{n-1} \cdot \frac{N-n}{N} \quad = \quad \frac{\hat{\sigma}_X^2}{n-1} \cdot \frac{N-n}{N}$$

and by (iv) this is

$$\frac{\hat{p}(1-\hat{p})}{n-1} \cdot \frac{N-n}{N}$$

2. The estimate of the total is the $393\hat{p}$. The sampling distribution of this estimate is approximately normal (the problem allows you to appeal to the CLT) with expectation $393 \times 0.652 = 257.02$ and variance $393^2 \cdot \frac{p(1-p)}{25} \cdot \frac{393-25}{393-1}$. You don't have to estimate p (and hence don't need the result of Problem 1) because it's given to be 0.654. So draw a normal curve with mean 257.02 and SE 36.23.

I know that many of you are just going to estimate the population proportion rather than the total. That's OK: the normal curve will be centered at 0.654 and have an SE of 0.092.

3a. Plug into the formula in Problem 1 to see that the standard error is 0.027. The confidence interval is $0.18 \pm 1.645 \times 0.027 = (0.1356, 0.2244)$.

b. Since $\hat{p}_1$ and $\hat{p}_2$ are independent, $Var(\hat{d}) = Var(\hat{p}_1) + Var(\hat{p}_2)$. By the result in Example D, $Var(\hat{p}_1) = 0.03^2$, so $Var(\hat{d}) = 0.03^2 + 0.027^2 = 0.001629$ and $SE(\hat{d}) = 0.04$.

I expect that many of you will re-compute the variance of $\hat{p}_1$ using the size of the new sample. But $\hat{p}_1 = 0.12$ came from the old sample, not the new one.

c. 99%: $-0.06 \pm 2.57 \times 0.04 = (-0.1628, 0.0428)$.

95%: $-0.06 \pm 1.96 \times 0.04 = (-0.1384, 0.0184)$.

90%: $-0.06 \pm 1.65 \times 0.04 = (-0.126, 0.006)$.

All these intervals contain 0, so there is no clear evidence that the two population proportions are different.

4a. $\binom{4}{2} = 6$.

b. The expectation of the sample mean is

$$\frac{1}{4}(\frac{x_1+x_2}{2} \ + \ \frac{x_2+x_3}{2} \ + \ \frac{x_3+x_4}{2} \ + \ \frac{x_1+x_4}{2}) \quad = \quad \frac{1}{4}(x_1+x_2+x_3+x_4)$$

That's the population mean. So yes, it's unbiased.

Now assume the population is the list 1, 2, 3, 4. The histogram is uniform on those 4 numbers (four equal bars centered on 1, 2, 3 and 4; no, I'm not going to draw it for you). The mean is 2.5 and the SD (don't you love the computational formula?) is 1.12, roughly.

The sample mean has the values 1.5, 2.5, 3.5, and 2.5 respectively for the four samples. So its histogram has 3 bars, centered over 1.5, 2.5, and 3.5. The one in the middle is twice as tall as the ones on either end. The expectation of this distribution is 2.5 and the standard error is 0.707.

5. $Cov(aX + bY, cX + dY) = acCov(X, X) + adCov(X, Y) + bcCov(Y, X) + bdCov(Y, Y) = ac\sigma^2 + (ad + bc)\gamma + bd\tau^2$.

6a. Use linearity of expectation and the fact that $E(X_i) = \mu$ for all i.

$$E(\bar{X}_c) \;=\; E(\sum_{i=1}^{n} c_i X_i) \;=\; \sum_{i=1}^{n} c_i E(X_i) \;=\; (\sum_{i=1}^{n} c_i)\mu$$

If this is to be equal to μ, then $\sum_i c_i = 1$.

Notice that this problem shows that you can construct any number of unbiased estimates of μ. Each X_i by itself is such an estimate, as is $17X_1 - 20X_2 + 4X_7$, as is ... go on, make as many as you like. They're all unbiased. The difference lies in the variance. Hence:

b. Use the fact that the variance of a sum is the sum of the variances plus the sum of all the covariances. The method of Problem 5 helps too.

$$\begin{aligned} Var(\bar{X}_c) &= \sum_i c_i^2 Var(X_i) \;+\; \sum\sum_{i\neq j} c_i c_j Cov(X_i, X_j) \\ &= \sigma^2 \sum_i c_i^2 \;+\; Cov(X_1, X_2) \sum\sum_{i\neq j} c_i c_j \end{aligned}$$

To minimize this subject to the constraint in part **a**, use the method of Lagrange multipliers. The function to differentiate is

$$\sigma^2 \sum_i c_i^2 \;+\; Cov(X_1, X_2) \sum\sum_{i\neq j} c_i c_j \;-\; \lambda(\sum_i c_i - 1)$$

This function is entirely symmetric in the c_i's. Differentiate with respect to each c_i and set equal to 0. Notice that the symmetry of the resulting equations means that all the minimizing c_i's must satisfy the same conditions. Hence the Lagrange equations will be satisfied if they are all equal. Since they sum to 1, they must all be $1/n$.

7a. Since $E(\bar{X}_1) = E(\bar{X}_2) = \mu$, we have $E(X) = \mu(\alpha + \beta)$. So the condition for unbiasedness is $\alpha + \beta = 1$.

b. Assume and $\alpha + \beta = 1$. Then $Var(X) = \alpha^2\sigma_1^2 + (1 - \alpha)^2\sigma_2^2$. (Yes, I dropped the $\bar{X}$ in the notation for the variances.) Differentiate with respect to α, set equal to 0 and solve for α to get the minimizing values

$$\begin{aligned} \alpha^* &= \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} \\ \beta^* &= \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2}. \end{aligned}$$

8a. $P(\text{answer yes}) = P(\text{get Statement 1 and answer yes}) + P(\text{get Statement 2 and answer yes})$
$= pq + (1 - p)(1 - q) = 2pq - q + 1 - p = (2p - 1)q + (1 - p)$.

Notice that if $p = 1/2$ you've lost all information about q in the probability of a "yes" answer. So assume $p \neq 1/2$.

b. Solve for q in the expression in **a**: $q = \frac{r-1+p}{2p-1}$. This is a linear function of r.

c. R is the sample proportion of those answering "yes", that is, it is the average of n indicators each of which is 1 with chance r. So $E(R) = r$. So to estimate q, use Q which is obtained by replacing r by R in the expression for q in **b**:

$$Q = \frac{R-1+p}{2p-1}$$

That's nicely linear in R so it's easy to find its expectation:

$$E(Q) \quad = \quad \frac{E(R)-1+p}{2p-1} \quad = \quad \frac{r-1+p}{2p-1} \quad = \quad q$$

d. "Ignoring the finite population correction" means that we're assuming that the answers are independent across people. So R is a binomial (n, r) proportion, hence the familiar formula for its variance.

e. This is just the variance of a linear function of R.

$$Var(Q) \quad = \quad Var(\frac{R-1+p}{2p-1}) \quad = \quad \frac{Var(R)}{(2p-1)^2} \quad = \quad \frac{r(1-r)}{n(2p-1)^2}$$

9a. Now P(answer yes)
$= P$(get Statement 1 and answer yes) $+ P$(get unrelated question and answer yes).
This gives $r = pq + (1-p)z$ where z is the known proportion of "yes" answers to the unrelated question.

So $q = \frac{r-(1-p)z}{p}$, and the same reasoning as before gives the estimate

$$Q \quad = \quad \frac{R-(1-p)z}{p}$$

b. Just as before. Since $E(R) = r$,

$$E(Q) \quad = \quad \frac{E(R)-(1-p)z}{p} \quad = \quad \frac{r-(1-p)z}{p} \quad = \quad q$$

c. As before, treat the answers as independent across people. So $Var(R) = r(1-r)/n$, and

$$Var(Q) \quad = \quad Var\left(\frac{R-z(1-p)}{p}\right) \quad = \quad \frac{Var(R)}{p^2} \quad = \quad \frac{r(1-r)}{np^2}$$

10. Your answers will all be different, and I will accept anything halfway to reasonable. I'd like you to notice that when p is very small the two schemes are quite different. In the first scheme the spinner will most likely land on the statement, "I don't have the characteristic," while in the second scheme it will most likely land on the statement, "Were you born in June?" So in the second scheme, most of the people will be telling you about their birth month, not about the characteristic of interest.