

Statistics 403 Problem Set 8

Due in lab on Friday, November 6th

1. Suppose we are collecting independent binary data Y_1, \dots, Y_n , where each Y_i is equal to either zero or one, with $P(Y_i = 1) = p$. For a given value of p , determine the sample size such that $\text{SD}(\bar{Y}) = 0.1$.

Solution: We have

$$\text{var}(\bar{Y}) = p(1 - p)/n = 1/100.$$

Thus, $n = 100 \cdot p(1 - p)$.

2. The variance of \bar{Y} under cluster sampling is

$$\sigma^2/n + \tau^2 \sum_{i=1}^q (n_i/n)^2,$$

where n_1, \dots, n_q are the sizes of the clusters, $n = n_1 + \dots + n_q$ is the total sample size, and σ^2 and τ^2 are as defined in the notes. This is the same formula as given on slide 39. (i) Simplify this formula in the special case where all the clusters have the same size. (ii) Explain whether the variance of \bar{Y} given by your result from part (i) is strictly increasing in q , strictly decreasing in q , or neither strictly increasing nor strictly decreasing.

Solution: If the cluster sizes are equal, $n_i = n/q$, so $n_i/n = 1/q$, and $\sum_{i=1}^q (n_i/n)^2 = \sum_{i=1}^q 1/q^2 = 1/q$. Thus the variance formula simplifies to $\sigma^2/n + \tau^2/q$, which is strictly decreasing in q . This means that the variance is always lower if there are more, smaller clusters (however this may be more expensive to carry out).

3. If we have data observed in clusters, the “intracluster correlation coefficient” (ICC) is $\tau^2/(\sigma^2 + \tau^2)$. Suppose we observe data in clusters of equal size, and the ICC is 0.4. Derive an expression for the relative difference

$$\frac{\text{var}(\bar{X}^{\text{srs}}) - \text{var}(\bar{X}^{\text{clust}})}{\text{var}(\bar{X}^{\text{srs}})}$$

as a function of the ICC (denote it r), the number of clusters q , and the sample size n . You should be able to get a simple expression that involves only n , r , and q . Explain in simple terms why the formula makes sense in the special cases where $r = 0$ and where $r = 1$.

Solution:

$$\begin{aligned}
\frac{(\sigma^2 + \tau^2)/n - (\sigma^2/n + \tau^2/q)}{(\sigma^2 + \tau^2)/n} &= \frac{(\sigma^2 + \tau^2) - (\sigma^2 + n\tau^2/q)}{\sigma^2 + \tau^2} \\
&= \frac{\tau^2(1 - n/q)}{\sigma^2 + \tau^2} \\
&= r(1 - n/q).
\end{aligned}$$

When $r = 0$, the relative difference becomes 0, meaning that cluster sampling and iid sampling have the same variance. This makes sense because $r = 0$ means that $\tau^2 = 0$, hence the data are independent both within and between clusters.

When $r = 1$, the values within a cluster are perfectly correlated. Thus there is only one meaningful observation per cluster, and our data are equivalent to an independent sample of size q . Letting $v = \sigma^2 + \tau^2$ be the variance, the variance of \bar{X}^{SRS} is v/n , the variance of \bar{X}^{clust} is v/q , and the relative difference is $(v/n - v/q)/(v/n) = 1 - n/q$, just as is given by the formula above when $r = 1$.

4. Suppose we are doing stratified sampling with respect to gender, in a situation where the genders are equally represented in the population. Suppose also that $p_f = 1/2 + d$ and $p_m = 1/2 - d$ for some number d . Suppose that our budget only allows collecting data on 20 subjects, and we would like to achieve a precision of $\text{SD}(\bar{Y}) = 0.1$. (i) How big must d be in order for stratified sampling to achieve this precision? (ii) Explain why your answer to part (i) would be the same even if the genders were not equally represented in the population.

Solution: From the formula on slide 30, the variance of \bar{Y} is

$$(q_f(0.5 + d)(0.5 - d) + q_m(0.5 + d)(0.5 - d))/20 = (1/4 - d^2)/20.$$

Note that this doesn't depend on q_f and q_m . Setting this equal to $0.1^2 = 0.01$ yields $d = .22$.

5. Suppose we are comparing study designs for estimating the expected value of a binary random variable Y . All study designs collect $n = 20$ independent values from a population which is 20% female and 80% male. The probability that $Y = 1$ for females is 0.8 and the probability that $Y = 1$ for males is 0.4. We will compare the study designs based on the widths of 95% confidence intervals for EY . The three study designs are: (i) an iid sample, (ii) a stratified sample consisting of 10 females and 10 males, and (iii) a stratified sample using the optimal sampling fractions given on slide 32.

Solution: Since $p = q_f p_f + q_m p_m = 0.48$, and the variance of \bar{Y} from an iid sample is $p(1 - p)/n$, it follows that the width for study design (i) is $4\sqrt{0.48 \cdot 0.52/n} = 0.45$.

The widths for study designs (ii) and (iii) can be calculated from the formula on slide 32. The widths are 0.52 for (ii) and 0.42 for (iii). Note that (ii) is worse than for an iid sample since we are doing stratified sampling with the wrong sampling fractions. If optimal stratified sampling is done correctly (iii), the interval is slightly shorter than for an independent sample (i).