

Sample Size Considerations (Chapter 4)

Up to now, we have assumed that the sample size n was known, and have studied properties of various resulting estimators of the population mean or total. Taking a step back, we now consider the more realistic question from a design point of view, namely: How large a sample do we need to attain some desired accuracy for the parameters we wish to estimate?

- One way this is considered is to specify a maximum allowable difference d between the estimate and the true value of the parameter, which is exceeded with some small probability α .
- In mathematical terms, the goal is to find the smallest sample size n which satisfies:

$$P(|\hat{\theta} - \theta| > d) < \alpha,$$

for some specified d and α , where $\theta, \hat{\theta}$ denote the population parameter and corresponding estimator respectively.

Sample Size Required to Estimate the Population Mean

Consider the estimation of the population mean. Here, we want:

$$P(|\bar{y} - \mu| > d) < \alpha.$$

Under simple random sampling, the sampling distribution of \bar{y} is at least approximately normal for large n whether sampling from an infinite or finite population by either the regular Central Limit Theorem or the finite population version (it's exactly normal if sampling from an infinite normal population); that is $\frac{\bar{y} - \mu}{\sqrt{\text{Var}(\bar{y})}} \overset{\sim}{\sim} N(0, 1)$.

- Let z be the upper $\alpha/2$ quantile from the standard normal distribution (the upper $\alpha/2$ quantile means the same thing as the $1 - \alpha/2$ quantile). Then, using the approximate normality of \bar{y} :

$$\begin{aligned} P\left[\left|\frac{\bar{y} - \mu}{\sqrt{\text{Var}(\bar{y})}}\right| > z\right] = \alpha &\implies P\left[\left|\frac{\bar{y} - \mu}{\sqrt{\frac{N-n}{N} \cdot \frac{\sigma^2}{n}}}\right| > z\right] = \alpha \\ &\implies P\left[|\bar{y} - \mu| > z\sqrt{\frac{N-n}{N} \cdot \frac{\sigma^2}{n}}\right] = \alpha, \end{aligned}$$

so that $d = z\sqrt{\frac{N-n}{N} \cdot \frac{\sigma^2}{n}}$. Solving for n gives:

$$\boxed{n = \frac{1}{\frac{1}{N} + \frac{d^2}{\sigma^2 z^2}} = \frac{1}{\frac{1}{N} + \frac{1}{n_0}}}, \text{ where } n_0 = \frac{\sigma^2 z^2}{d^2}.$$

- Note that if N is large, the $1/N$ term in the denominator can be ignored, and this formula reduces to: $n = n_0$, which is the sample size requirement calculated in standard statistics textbooks for a given d & α .
- Problem: we don't know σ so we need a preliminary guess. How?

Sample Size Required to Estimate the Population Total: A similar result can be obtained for specifying the accuracy of the estimate for the population *total*, where if:

- d = the maximum allowable difference between the population total and its estimate,
- α = the probability this difference is larger than d ,

then the sample size required to satisfy: $P(|\hat{\tau} - \tau| > d) < \alpha$ is given by:

$$n = \frac{1}{\left(\frac{1}{N} + \frac{d^2}{N^2\sigma^2z^2}\right)} = \frac{1}{\frac{1}{N} + \frac{1}{n_0}}, \text{ where } n_0 = \frac{N^2\sigma^2z^2}{d^2}.$$

Example: Returning to the deer pellet example from class, suppose that the sample with 10 transects was merely a pilot study to gain information about the variability in the number of pellet groups. In this pilot study, the following summary statistics were calculated:

$$\bar{y} = 5.55 \text{ pellet groups/150 sq.ft, } s^2 = 14.06, \quad N = 5808.$$

Suppose we want to estimate τ (the total number of pellet groups in the 20-acre area) to within 5000, with probability 0.95 ($\alpha = 0.05$). How large a sample is required?

- Since the sample variance from the pilot study, s^2 , was a “guess” for the population variance σ^2 , we might add something to the sample size determined to be conservative.

Specifying the Relative Accuracy: Instead of specifying a desired difference d as above, the sample size determination problem can equivalently be stated in terms of the relative accuracy with which we would like to estimate some parameter.

- Suppose we want to estimate the population mean to within 10% ($r = 0.10$) with probability 0.95. We want:

$$d = r\mu \implies n_0 = \frac{\sigma^2z^2}{r^2\mu^2}.$$

- Note that there are two unknown parameters here: σ^2 and μ . Let $\gamma = \sigma/\mu$ (coefficient of variation). Then n_0 can be rewritten as:

$$n_0 = \frac{\sigma^2 z^2}{r^2 \mu^2} = \frac{z^2 \gamma^2}{r^2}.$$

- Writing n_0 in this fashion leaves only *one* unknown parameter in computing the sample size, namely γ . Hence, this latter formula can be used in situations where the coefficient of variation can be specified more easily than the mean and variance individually.
- Suppose we want to estimate the population total to within 10% ($r = 0.10$) with probability 0.95. We want:

$$d = r\tau \implies n_0 = \frac{N^2 \sigma^2 z^2}{r^2 \tau^2}.$$

Example: If we had wanted to estimate the population mean number of pellet groups per 150 square feet to within 10% of the mean, we compute:

$$n_0 = \frac{\sigma^2 z^2}{r^2 \mu^2} = \frac{(14.06)(1.96)^2}{(.10)^2 (5.55)^2} = \underline{175.35 \text{ transects.}}$$

- With a finite population correction (fpc), $n = 170.2$ transects.

Sample Size Required to Estimate a Population Proportion: To obtain an estimator \hat{p} within d of the population proportion p with probability $1 - \alpha$, the sample size required is:

$$n = \frac{1}{\frac{N-1}{Nn_0} + \frac{1}{N}} \approx \frac{1}{\frac{1}{n_0} + \frac{1}{N}}, \text{ where: } n_0 = \frac{z^2 p(1-p)}{d^2}.$$

- Note that these formulas have the same basic form as those for the population mean. As with the mean, if N is sufficiently large, the fpc can be ignored, and n_0 is the desired sample size.
- Just as the analogous computation for the mean required a “guess” of the standard deviation, here we require a “guess” of the population proportion. If no such estimate is available, we could be conservative by setting p to the value which maximizes n . What value of p is that?
- The book also provides a section on determining the sample sizes necessary for estimating several proportions simultaneously (pp. 42-44).

```
# This is an example of an R script to calculate the sample size needed
# to estimate the total number of deer pellets in a 20-acre field, as done
# on page 16 of the class notes. Here, for different choices of the desired
# detectable difference d, the sample size n0 (without the fpc) and n (with
# the fpc) are computed and plotted against the d values.
```

```
N <- 5808          # Defines the population size.
s2 <- 14.06        # Defines the sample variance.
z <- qnorm(.975,0,1) # Defines the standard normal quantile.
d <- c(1000,1500,2000,2500,3000,# The next two commands are
      3500,4000,4500,5000) # alternative ways to do the same thing;
d <- seq(1000,5000,500) # the "seq" command is very useful here.
n0 <- N^2*s2*z^2/d^2   # Computes n0, the sample size
                        # without the fpc.
n <- 1/(1/n0 + 1/N)    # Computes n, the sample size with the fpc.
plot(d,n0,xlab="Difference", # Plots the sample sizes (w/o fpc) versus
     ylab="Sample size",pch=1, # the differences (d) with axis labels,
     cex=1.5)           # and plotting character 1 (open circle).
                        #
points(d,n,pch=16)      # Plots the sample sizes (w/ fpc) versus
                        # the differences (d) in overlay, with
                        # plotting character 16 (filled circle).
legend(3000,1500,c("n0","n"), # Puts a legend on the plot at (2000,6000),
      pch=c(1,16),cex=1.5) # using the plotting characters 1 & 16.
title("Sample Size Required vs. Desired Difference") # Puts a title on the plot.
```

Sample Size Required vs. Desired Difference

