

Lecture 4

Simple Random Sampling

Simple random sampling without replacement of size n

Some necessary combinatorial notation:

- (n factorial) $n! = n \times (n-1) \times (n-2) \times \cdots \times 2 \times 1$. This is the number of unique arrangements or orderings (or permutations) of n distinct items.
- (N choose n) $\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N(N-1)\cdots(N-n+1)}{n!}$. This is the number of combinations of n items selected from N distinct items (and the order of selection doesn't matter).
- There are $\binom{N}{n}$ possible SRSs of size n from a population of size N . Each has $P(\mathcal{S}) = \frac{1}{\binom{N}{n}}$

In general, we will assume sampling is without replacement.

Estimation of \bar{y}_U and t

Using the resampling approach from Lab 2, we could calculate variability or construct confidence intervals of our estimators from the approximate sampling distribution created in R.

A natural **estimator** for the population mean \bar{y}_U is the sample mean \bar{y} .

$$\hat{y}_U = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

\hat{y}_U and \hat{t} are **design unbiased**.

Does design unbiased imply anything about the variation of the estimator?

The next problem is to

You have probably been taught that the variance of the sample mean $V(\bar{Y}) = S^2/n$ and this is appropriate for samples from very large or even infinite populations.

Ex. Consider an extreme case where $n = N$, that is all units are sampled. What is the variability in \bar{y} ?

When dealing with **finite populations** that are not very large, we adjust the variance formulas using the

Then with the use of FPC the variance formulas are:

$$*V(\hat{y}_U) = \frac{S^2}{n} \left(1 - \frac{n}{N}\right) = \frac{S^2}{n} \left(\frac{N-n}{N}\right) \qquad V(\hat{t}) = N^2 V(\hat{y}_U) = N(N-n) \frac{S^2}{n} *$$

Using the estimated variance s^2 in place of S^2 in the above formulas and taking the square root gives the *standard error (SE)* of these estimators.

Notes about the FPC:

- If N is large relative to n , then the FPC will be
- If N is not large relative to n , then omitting the FPC from the variance formulas can
- As $n \rightarrow N$, $\frac{N-n}{N}$
- Thinking about this deeper, where does the uncertainty in \bar{y} come from?

The **Coefficient of Variation (CV)** of an estimator \bar{y} measures

$$CV(\bar{y}) = \frac{\sqrt{V(\bar{y})}}{E(\bar{y})} = \sqrt{1 - \frac{n}{N}} \frac{S}{\sqrt{n}} \times \bar{y}_U$$

Note the CV does not have a unit of measurement.

The estimated CV uses the standard error divided by the sample mean

Confidence Intervals for \bar{y}_U and t

Similar to confidence intervals you have seen in other statistics courses, we will use asymptotics and the central limit theorem to construct confidence intervals.

In an introductory statistics course, you were given confidence interval formulas

$$\bar{y} \pm \quad \text{and} \quad \bar{y} \pm$$

These formulas are applicable if a sample was to be taken from an infinitely or extremely large population. But when we are dealing with finite populations,

In the finite population version of the Central Limit Theorem, we assume the estimators $\hat{\bar{y}}_U = \bar{y}$ and $\hat{t} = N\bar{y}$ have sampling distributions that are approximately normal. That is,

For large samples, approximate 100(1- α)% confidence intervals for \bar{y}_U and t are:

For \bar{y}_U

$$\bar{y} \pm z^* \sqrt{\left(\frac{N-n}{N}\right) \frac{s^2}{n}}$$

$$\bar{y} \pm z^* s \sqrt{\left(\frac{N-n}{N}\right) / n}$$

For t :

$$N\bar{y} \pm z^* \sqrt{N(N-n) \frac{s^2}{n}}$$

$$N\bar{y} \pm z^* s \sqrt{N(N-n)/n},$$

where z^* is the upper $\alpha/2$ critical value from the standard normal distribution

For smaller samples, approximate $100(1-\alpha)\%$ confidence intervals for \bar{y}_U and t are:

For \bar{y}_U

$$\bar{y} \pm t^* \sqrt{\left(\frac{N-n}{N}\right) \frac{s^2}{n}}$$

$$\bar{y} \pm t^* s \sqrt{\left(\frac{N-n}{N}\right) / n}$$

For t :

$$N\bar{y} \pm t^* \sqrt{N(N-n) \frac{s^2}{n}}$$

$$N\bar{y} \pm t^* s \sqrt{N(N-n)/n},$$

where t^* is the upper $\alpha/2$ critical values from the t distribution with $n - 1$ degrees of freedom.

So what constitutes a large sample?

The quantity being added and subtracted from $\hat{\bar{y}}_U = \bar{y}$ or $\hat{t} = N\bar{y}$ in the confidence interval is known as the

One-sided Confidence Intervals for \bar{y}_U and t

Occasionally, a researcher may want a one-sided confidence interval.

Approximate *upper* and *lower* $100(1-\alpha)\%$ confidence intervals for \bar{y}_U and t are:

	For \bar{y}_U	For t :
upper	$\left(\bar{y} - t^* s \sqrt{\left(\frac{N-n}{N}\right) / n}, \infty \right)$	$\left(N\bar{y} - t^* s \sqrt{N(N-n)/n}, \infty \right)$
lower	$\left(-\infty, \bar{y} + t^* s \sqrt{\left(\frac{N-n}{N}\right) / n} \right)$	$\left(-\infty, N\bar{y} + t^* s \sqrt{N(N-n)/n}, \infty \right),$

where t^* is the upper α critical value from a $t(n - 1)$ distribution. For larger samples, t^* can be replaced with z^{**}

If the response (y-values) cannot be negative, replace $-\infty$ with 0 in the lower confidence interval formulas. If the response cannot be positive, replace ∞ with 0 in the upper confidence formulas.

Later, we will discuss another method of generating a confidence interval

Confidence Intervals in R

There are R packages available, but writing code (and functions) in R to compute confidence intervals is straightforward. The code below generates a population of data from a Poisson distribution and samples 10 observations.

```
N <- 100
mu <- 15
n <- 10
population <- rpois(N,mu)
sample.responses <- sample(population,n)
y.bar <- mean(sample.responses)
s <- sd(sample.responses)
alpha <- .05
conf.int <- c(y.bar - qt(1-alpha/2, df = n-1)*s*sqrt(((N-n)/N)/n),
              y.bar + qt(1-alpha/2, df = n-1)*s*sqrt(((N-n)/N)/n))
```

The confidence interval for the population mean is (12.29, 16.11).

Another way to do this is to write a function in R. Here is the same code placed into a function.

```
FPC_ci <- function(sample,N,alpha=.05){
  # FUNCTION TO CALCULATE FPC BASED CONFIDENCE INTERVAL
  # ARGUMENTS:
  #   sample: observed responses
  #   N: total population size
  #   alpha = significance level
  n <- length(sample)
  y.bar <- mean(sample)
  s <- sd(sample)
  conf.int <- c(y.bar - qt(1-alpha/2, df = n-1)*s*sqrt(((N-n)/N)/n),
                y.bar + qt(1-alpha/2, df = n-1)*s*sqrt(((N-n)/N)/n))
  return(conf.int)
}

FPC_ci(sample.responses, 100)

## [1] 12.28584 16.11416
```

Calculate the confidence interval again for varying sizes of n and N (this will require resimulating the data).

- How do your results vary?
- What are the implications of changing n or N ?

Proportion Estimation

Suppose we are interested in an attribute associated with the sampling units.

Statistically the goal is to estimate proportion p . We use an indicator function that assigns a y_i value to unit i as follows:

$$y_i = \begin{cases} 1 & \text{if unit } i \text{ possess the attribute} \\ 0 & \text{otherwise} \end{cases}$$

Then $t = \sum_{i=1}^N y_i$ and $\bar{y}_U = \frac{1}{N} \sum_{i=1}^N y_i = p$.

By taking a SRS of size n , we can estimate p with the **sample proportion** \hat{p} of units that possess that attribute:

$$\hat{p} = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$$

The sampled proportion \hat{p} is unbiased for p .

For a finite population of 0 and 1 values, the variance of \hat{p} is

$$V(\hat{p}) = \left(\frac{N-n}{N} \right) \frac{S^2}{n} = \left(\frac{N-n}{N} \right) \left(\frac{N}{N-1} \right) \frac{p(1-p)}{n} = \left(\frac{N-n}{N-1} \right) \frac{p(1-p)}{n}$$

Because S^2 is unknown, we estimate it with $s^2 = \frac{n}{n-1} \hat{p}(1-\hat{p})$.

Note: the effects of omitting the FPC from the formulas in large and small samples are the same as earlier.

Confidence intervals for p

Let the random variable Y be the number of units in a SRS of size n that possess the attribute of interest.

Hypergeometric distribution for SRS: the probability that a SRS of size n will have exactly j sampling units that possess the attribute is:

$$Pr(Y = j) = \frac{\binom{t}{j} \binom{N-t}{n-j}}{\binom{N}{n}}.$$

This is the probability that SRS will consist of j ones and $n - j$ zeroes selected from the population containing t ones and $N - t$ zeros.

Remember there are t ones and $N-t$ zeros in the population. However, t is unknown.

Although the problem no longer depends on t , it still depends on the unknown proportion parameter p . What is commonly done is to apply the normal approximation to the binomial distribution:

$$\hat{p} \sim N(p, V(\hat{p}))$$

Thus, if the sample size n is large enough, we use $\hat{V}(\hat{p})$ to estimate $V(\hat{p})$. An approximate $100(1-\alpha)\%$ confidence interval for p is:

$$\hat{p} \pm z^* \sqrt{\hat{V}(\hat{p})} \quad \text{OR} \quad \hat{p} \pm z^* \sqrt{\left(\frac{N-n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}},$$

where z^* is the upper $\alpha/2$ critical value from the standard normal distribution. Sample sizes are typically large enough to use z^* instead of t^* .

The normal approximation will be reasonable given

Sample Size Determination with Simple Random Sampling

It is well known that an increase in sample size n will lead to a more precise estimator of \bar{y}_U or t . It is also obvious that an increase in the sample size n will make the sample more expensive to collect. There will, however, be a limited amount of resources available for data collection.

When designing a sampling plan, the researcher wants to achieve a desired degree of reliability at the lowest possible cost while satisfying the resource limitations for data collection.

To do this, the researcher tries to achieve a balance to avoid the following mistakes:

To determine a sample size n when estimating a parameter θ , we do the following:

Sample Size for \bar{y}_U

Goal: Estimate the SRS size required so the probability that the difference between the estimator $\hat{\bar{y}}_U$ and the population mean \bar{y}_U does not exceed a maximum allowable difference d is at most α .

Mathematically, find n such that

Assuming \bar{y} is approximately normally distributed, this is equivalent to finding n so that the margin of error

$$z_{\alpha/2} \sqrt{\left(\frac{N-n}{N}\right) \frac{S^2}{n}} \leq d.$$

Solving this inequality for n yields

$$n = \frac{\left(\frac{zs}{d}\right)^2}{1 + \left(\frac{zs}{d}\right)^2/N} = \frac{n_0}{1 + \frac{n_0}{N}}$$

where $n_0 = \left(\frac{zs}{d}\right)^2$ and is the sample size calculation for an infinite population, z is the critical value from a $N(0, 1)$ distribution.

Rounding up the value of n yields the desired sample size.

If the population size N is very large, then $1/N \approx 0$. In this case, $n \approx n_0$. This is the typical formula for infinite sample settings.

There remains one major problem.

Sample size for t

Goal: Estimate the SRS size required so the probability that the difference between the estimator $\hat{t} = N\bar{y}$ and the population total t does not exceed a maximum allowable difference d is at most α .

Mathematically, find n such that $Pr(|\hat{t} - t| > d) < \alpha$ for a specified maximum allowable difference d .

Assuming $N\bar{y}$ is approximately normally distributed, this is equivalent to finding n so that the margin of error is $z_{\alpha/2} \sqrt{N(N-n) \frac{S^2}{n}} \leq d$ Solving this inequality for n yields

$$n = \frac{n_0 N^2}{1 + N n_0} = \frac{1}{\frac{1}{N^2 n_0} + \frac{1}{N}}.$$

Sample size for p

Goal: Estimate the SRS size required so the probability that the difference between the sample proportion \hat{p} and the population proportion p does not exceed a maximum allowable difference d is at most α .

Mathematically, find n such that $Pr(|\hat{p} - p| > d) \leq \alpha$ for a specified maximum allowable difference d .

Assuming \hat{p} is approximately normally distributed, this is equivalent to finding n so that the margin of error $z_{\alpha/2} \sqrt{\left(\frac{N-n}{N-1}\right) \frac{p(1-p)}{n}} \leq d$.

Solving this inequality for n yields

$$n \approx \frac{1}{\frac{1}{n_0} + \frac{1}{N}}$$

where $n_0 = \frac{z_{\alpha/2}^2 p(1-p)}{d^2}$

Unfortunately, the sample size formulas assume you know the population proportion p , the quantity you are trying to estimate.

Note that the standard deviation of \hat{p} is largest when $p = 1/2$. Thus it is conservative to use $p = 1/2$ if there is no prior reasonable estimate.

Sample Size Calculations in R

```
#### Sample Size Calculations for Proportions
d <- .1 # maximum allowable difference
alpha <- .05 # 95% confidence interval
N <- 500 # population size
p <- .5 # use p= 1/2 for conservative estimate
n.0 <- (qnorm(1-alpha/2)^2 * p *(1-p)) / d^2
n = round(1 / (1/n.0 + 1/N))
```

For an example where the population size is 500 and the goal is to estimate the proportion with a maximum allowable difference of 0.1, the required sample size is 81.