

Chapter 14: Fundamental Concepts of Surveying

- Element, Universe, Survey, and Census
- Probability and Nonprobability Surveys
- Problems common to all surveys (SKIP)
- Simple Random Sampling

Chapter 15: Survey Designs

- Simple Random Samples
- Confidence Intervals for Means and Proportions
- Effect of Sample Size on Confidence Intervals

The **elements** in a study are the basic units (individuals or things) about which information is sought. (page 371)

The **universe** is the collection of elements about which we wish to be informed. (page 372)

The set of all measurements on a variable in a universe is called a **population**. (page 406)

Populations and Parameters

Element	y (0 or 1) measurements	x measurements
I_1	Y_1	X_1
I_2	Y_2	X_2
I_3	Y_3	X_3
\vdots	\vdots	\vdots
I_N	Y_N	X_N
Total	τ_y	τ_x
Mean	μ_y	μ_x
Standard Deviation	S_y	S_x

Here

$$\tau_x = \sum_{i=1}^N X_i \quad \mu_x = \frac{\tau_x}{N}$$

$$\tau_y = \sum_{i=1}^N Y_i \quad \mu_y = \frac{\tau_y}{N}$$

and

$$\sigma_x = \sqrt{\frac{1}{N} \sum_{i=1}^N (X_i - \mu_x)^2}$$

$$\sigma_y = \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - \mu_y)^2}$$

Any numerical characteristic of a population is called a **parameter**. (page 406)

Most universes will contain many different populations.

Proportions

When a variable is binary ($Y = 0$ or 1), then the population total, τ_y , is just the number of ones or number of "successes."

The population mean, μ_y , is just the proportion of ones or successes otherwise written as $\mu_y = \pi$.

Algebra shows that, in this binary case,

$$\begin{aligned}\sigma_y &= \sqrt{\frac{1}{N} \sum_{i=1}^N (Y_i - \mu_y)^2} \\ &= \sqrt{\pi(1 - \pi)}\end{aligned}$$

This is the same as the standard deviation for a Bernoulli process.

Sampling Error: the difference between the value of a sample estimate and the corresponding value in the population that is due only to the sampling process.

Precision means statements about the width of intervals within which predictions about characteristics of the universe are made.

The narrower the limits the more precise the predictions or estimates.

The Characteristics of Probability Surveys (page 380)

What do we get from a probability survey? Clearly, we do not get complete information about the characteristics of the universe.

Instead we get **estimates** or guesses about the characteristics of the population.

Because the elements in the sample are selected by randomization, we can also provide quantitative statements about the precision of our estimates.

We can quantify the margin of sampling error.

Simple Random Sampling

In general, if a universe has N elements, there are

$$\binom{N}{n} = \frac{N!}{n!(N-n)!}$$

different samples of size n .

For example: with a class of 480 students how many different samples of 6 could be selected?

$$\begin{aligned}\frac{N!}{n!(N-n)!} &= \frac{480!}{6!(480-6)!} \\ &= \frac{480!}{(6!)(474!)} \approx 1.646 \times 10^{13}\end{aligned}$$

An extremely large number!

Therefore we illustrate with small examples.

Example: (page 382)

A simple universe of size 9 (Exhibit 14.6A)

Consider all possible samples of size 2.
(Exhibit 14.6B)

List them and the associated values of several variables.

Now look at the sample means and their distributions and characteristics. (Exhibits 14.6B and C)

Notation for Data and Statistics

Element in sample	y (0 or 1) measurements	x measurements
i_1	y_1	x_1
i_2	y_2	x_2
i_3	y_3	x_3
\vdots	\vdots	\vdots
i_n	y_n	x_n
Total	$\sum_{i=1}^n y_i$	$\sum_{i=1}^n x_i$
Mean	$\bar{y} = p$	\bar{x}
Standard Deviation	$s_y = \sqrt{p(1-p)}$	s_x

Sampling Distribution

The distribution of a sample statistic over all possible samples is called a **sampling distribution**.

Numerical characteristics of samples are called **statistics**.

(Compare with population parameters.)

All statistics are subject to sampling variation, that is, they will vary from sample to sample.

The mean \bar{y} and standard deviation s_y are statistics.

The sample mean \bar{y} is an estimate of the population mean μ_y

The sample standard deviation s_y is an estimate of the population standard deviation σ_y

All estimates are subject to sampling variation, that is, they will vary from sample to sample.

Facts About the Sampling Distribution of the Mean

First, we have the (simple) result

$$\mu_{\bar{y}} = \mu_y$$

The standard deviation of a statistic is also called the **standard error** of the statistic.

Here we need to investigate the standard error of the mean.

How are the sampling fraction and finite population correction factor related?

$f = n/N$	$fpc = \sqrt{1 - f}$
0.05	0.9747
0.10	0.9487
0.20	0.8944
0.30	0.8337
0.50	0.7071
0.75	0.5000
0.90	0.3162

If $N = 2,000,000$ and $n = 1500$, then
 $f = 1500/2,000,000 = 0.00075$ and
 $fpc = \sqrt{1 - f} = \sqrt{1 - 0.00075} = 0.99962493 \approx 1$

Unless we say otherwise, we will assume for all of our work that N is much larger than n so that we take $fpc = 1$.

Standard Error of the Mean

For the (theoretical) sampling distribution of the sample mean \bar{y} , it may be shown that the standard error is given by (page 411)

$$SE_{\bar{y}} = \frac{\sigma_y}{\sqrt{n}} fpc$$

where the **finite population correction** factor is

$$fpc = \sqrt{\frac{N-n}{N}} = \sqrt{1-f}$$

and $f = n/N$ is called the **sample fraction**.

Finally:

- if the sample size n is moderately large and
- the sampling fraction f is small to moderate,

then the distribution of \bar{y} is approximately normal. (This is the **Central Limit Effect** once more! page 411)

The parameters of the normal distribution are of course:

$$\mu_{\bar{y}} = \mu_y$$

and

$$SE_{\bar{y}} = \frac{\sigma_y}{\sqrt{n}} fpc$$

In practice, since s_y and hence $SE_{\bar{y}}$ is unknown, we must use the estimated standard error

$$se_{\bar{y}} = \frac{s_y}{\sqrt{n}} fpc$$

For **proportions** $\sigma = \sqrt{\pi(1-\pi)}$ and we use an estimated standard error of

$$se_p = \sqrt{\frac{p(1-p)}{n}}$$

Example:

New York Times/CBS News Poll reported Feb. 28 1995. "Are police searches without a warrant a good idea or a bad idea?"

69% said "bad idea" (20% said "good idea")

What is the margin of error in that 69%?

What is the 95% confidence interval for the "true" proportion in the population (π) that think warrantless searches are a bad idea?

Interval Estimation (page 410)

Proportions

The sample proportion, p , is an estimate of the population (or process) proportion π .

The Confidence Interval for π is

$$p \pm z_c \sqrt{\frac{p(1-p)}{n}}$$

where z_c is chosen from a standard normal distribution to produce the desired confidence level.

Typically, $z_c = 2$ for the usual 95% confidence.

This confidence interval is based on the Central Limit Effect for proportions and assumes that n is reasonably large.

Here $n = 1190$ and

$$\begin{aligned} se_p &= \sqrt{\frac{p(1-p)}{n}} \\ &= \sqrt{\frac{0.69(1-0.69)}{1190}} = \sqrt{\frac{0.69(0.31)}{1190}} \\ &= 0.013 \end{aligned}$$

(Since N is about 200,000,000 we set the fpc to 1.)

So with 95% confidence the margin of error is $\pm 2(0.013) = \pm 0.026$

or about ± 3 percentage points (Don't say 3%)

The 95% confidence interval for π is

$$0.69 \pm 0.026$$

or 0.664 to 0.716

If we want 99.7% confidence, we have

$$0.69 \pm 3(0.013) = 0.69 \pm 0.039$$

or 0.651 to 0.729

We have more confidence but the interval is wider.

For 99% confidence we look up the z-value multiplier (or 99.5 percentile) and get $z = 2.575$

So the margin of error is $\pm 2.575(0.013) = 0.033$ with 99% confidence.

The 99% confidence interval is

$$0.69 \pm 0.033 \quad \text{or} \quad 0.657 \quad \text{to} \quad 0.723$$

with 99% confidence.

Notice that both of these confidence intervals are of the form

$$\text{parameter estimate} \pm z_c \text{se}_{\text{estimate}}$$

where $\text{se}_{\text{estimate}}$ is the standard error of the parameter estimate, that is, the estimate of the standard deviation of the parameter estimate.

Means

The sample mean, \bar{y} , is an estimate of the population (or process) mean μ .

Confidence interval for μ

$$\bar{y} \pm z_c \frac{s}{\sqrt{n}}$$

where z_c is chosen from a standard normal distribution to produce the desired confidence level.

Typically, $z_c = 2$ for the usual 95% confidence.

This confidence interval is based on the Central Limit Effect for means and assumes that n is reasonably large.

Confidence Levels (page 412)

The choice of the factor 3 in $\pm 3\text{se}_{\bar{y}}$ was quite arbitrary but produced 99.7% confidence.

Other multipliers can be used to yield different confidence levels.

For example, using a multiplier of 2 would give 95% confidence in the interval.

In general, the intervals are of the form

$$\bar{y} \pm z \text{se}_{\bar{y}}$$

where z is chosen to achieve a desired confidence level.

The Trade-Off Between High Confidence and Narrow Confidence Intervals

(page 412)

To make precise statements about μ , we would like narrow confidence intervals and high confidence.

However, with n (and N) fixed, the width of the confidence interval increases as the confidence increases.

Chap. 14 - page 25

Unfortunately, p is not known at this point—we have no data! A conservative approach is to set $p = 1/2$ as this is the “worst” case, i.e., the one with the most variability. Doing this and solving for n gives

$$n = \frac{1}{B^2}$$

For example, if we want a margin of error of about plus or minus 3 percentage points, i.e., $B=0.03$, then we need a sample of size

$$n = 1/(0.03)^2 = 1111.111 \text{ or about } 1111.$$

Chap. 14 - page 27

Choosing Sample Sizes (page 417)

When discussing parameter estimation the part $z_c se_{\text{estimate}}$ is often called the **margin of error** (or, more correctly, margin of sampling error.) It's the plus or minus part of the confidence interval.

In choosing a sample size for a study we might require that the margin of error be of a certain size, say B . Once B is specified we could attempt to choose a sample size n that will achieve the required margin of error.

For proportions and 95% confidence level, this means we want to solve for n in the equation

$$B = 2 \sqrt{\frac{p(1-p)}{n}}$$

Chap. 14 - page 26

For means μ , we have a somewhat more difficult situation. Here we want

$$B = 2 \frac{s}{\sqrt{n}}$$

but s is unknown!

We must have some idea of the variability within the population (or process) if we want to specify the margin of error when estimating the mean.

Sometimes we have some previous experience with a similar situation that will give us some guidance.

We may have to carry out a (small) pilot study to get a “ballpark figure” for the variability before we choose our sample size for the full study.

Chap. 14 - page 28