Sampling for Success

Dr. Jim Mirabella
President, Mirabella Research Services, Inc.
Professor of Research & Statistics

Session Objectives

- Upon completion of this workshop, participants will be able to:
  - Compare and contrast different approaches to sampling
  - Compute minimum sample sizes
  - Evaluate potential threats to reliability and validity

Sampling Terminology

- Observation – A single piece of data
- Population – A collection of all possible observations sharing some common set of characteristics
- Census – An investigation of all the individual observations making up a population

Sampling Terminology

- Sample – A subset or some part of a larger population; a sample can be the entire population
- Sampling – The process of using a small number of items from a larger population to draw conclusions about the whole population

Why Sample?

- Lower cost
- Greater speed of data collection
- Greater accuracy of results
- Availability of population elements
- Destructiveness of observations

Stages In Sample Selection

1. Define the target population
2. Select a sampling frame
3. Choose probability or non-probability sampling method
4. Determine sample size
5. Choose a data collection technique
6. Select sample

Defining the Target Population

Target Population

- Target Population – The specific group relevant to the research project; the group that the sample truly represents.
- Consider the following questions:
  - "What is the relevant population?"
  - "To whom do we want to talk?"
- Tangible characteristics must be used to define the population
  - Demographic
  - Behavioral

Selecting the Sampling Frame

Sampling Frame

- Sampling Frame – The list of elements from which a sample may be drawn
  - A complete and correct list of population members
  - Ideally, the source should be representative of the population
  - The source should not bias the results

Choosing a Probability or Non-probability Sample

Probability vs. Non-probability Samples

- Probability Sample – A sample in which each items are selected on the basis of known probabilities
- Non-probability Sample – A sample in which items are selected without regard to their probability of occurrence

Non-probability Samples

- Convenience Sample – A sample of items most readily available
  - Mall surveys
- Judgment Sample – A sample selected by an experienced researcher based upon some appropriate characteristic
  - The Dow Jones Industrial Average

Non-probability Samples

- Quota Sample – A sample that ensures that certain characteristics of the population will be represented to the exact intended extent
  - Used often in polling
- Snowball Sample – Initial respondents are selected randomly, and they help obtain additional respondents that are hard to find
  - Surveying 10 shoppers selected at random and asking each of them for the names of five friends

Non-probability Sampling

<u>Advantages</u>

- Lower cost
- Less time
- May be the only feasible alternative

<u>Disadvantages</u>

- Greater opportunity for bias
- Results not generalizable
- Lack of objectivity

Probability Samples

- Simple Random Sample – Each element of the population has an equal chance of being included in the sample (e.g., Lotto)
- Systematic Random Sample – A sample in which every *k*th number is selected from a comprehensive list

Probability Samples

- Stratified Random Sample – A subsample drawn from samples within different strata to mirror the population
- Cluster Random Sample – Randomly select entire clusters of observations and all members of the cluster are included in the sample

Probability Sampling

<u>Advantages</u>

- Minimization of bias
- Generalizability of results

<u>Disadvantages</u>

- More costly
- More time consuming

Determining Sample Size

Sample Size

- A sample does not have to be large to be useful, as long as it's representative
- What is the "right" sample size?
  - Is it a percentage of the population?
  - Is population size a factor?
  - Is there a magic minimum?

Sample Size

- When should samples be large?
  - Serious or costly decisions
  - Time and resources readily available
- When should you permit a small sample?
  - Few major decisions based on results
  - Only rough estimates needed
  - High data collection costs
  - Time constraints

Sample Size

- According to Dr. George Gallup:

"You do not need a large sampling proportion to do a good job if you first stir the pot well."

  - If you have two pots of soup on the stove (one large, one small), you don't need to take more spoonfuls from the larger pot to sample the tastes accurately.

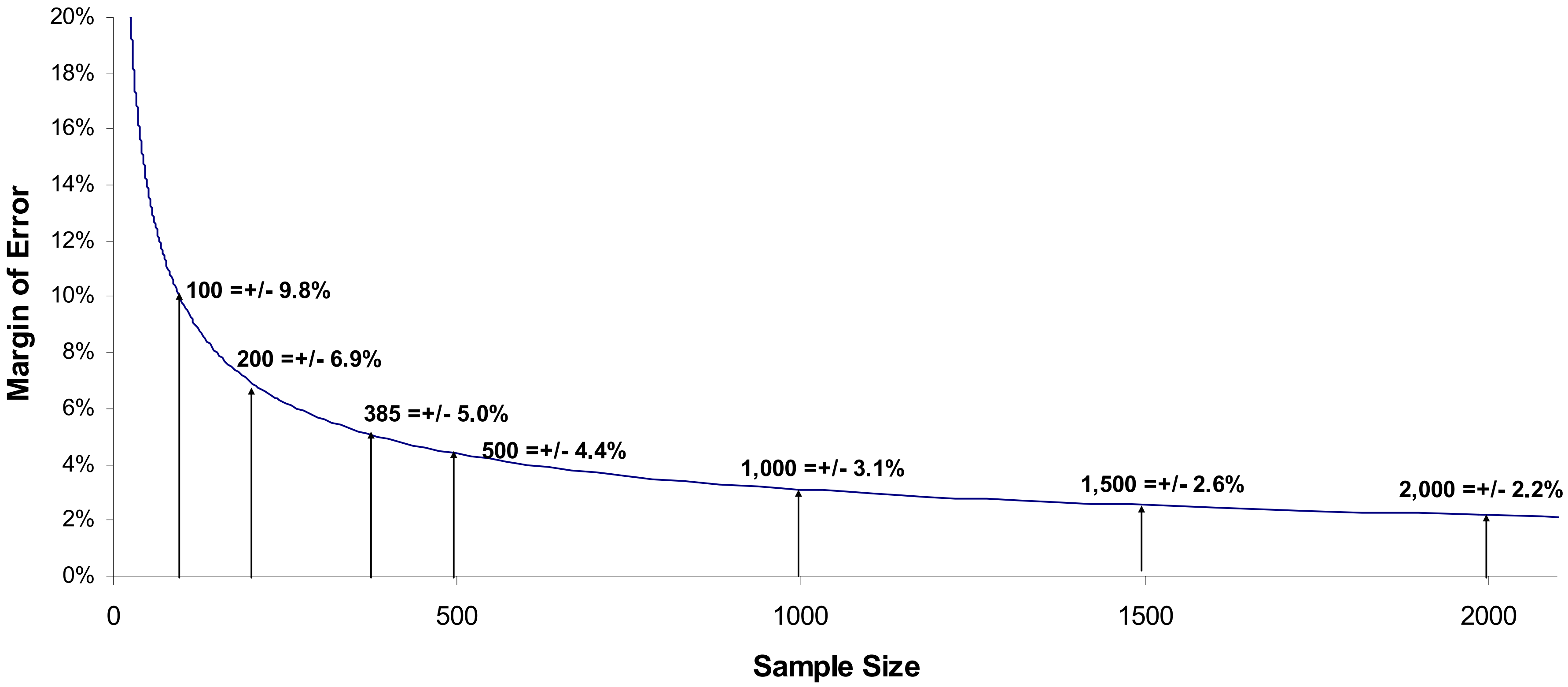
Sample Size

- How much sampling error will you allow?
  - Sampling error is the difference between the actual population value and the results of the sample
  - As tolerable error decreases, sample size increases exponentially
    - Increasing sample size from 1,000 to 2,000 reduces sampling error by 1%
    - Increasing sample size from 2,000 to 4,000 reduces sampling error by ½%

Sample Size

- How confident must you be?
  - Confidence is the likelihood that the actual population value is within the tolerable error range
  - As confidence level increases, sample size increases

Sample Sizes vs. Error (95% Confidence)



Estimating Sample Size (means)

$$n = \left(\frac{ZS}{E} \right)^2$$

Estimating Sample Size (proportions)

$$n = .25\left(\frac{Z}{E}\right)^2$$

Estimating Sample Size

- Estimate the standard deviation of the population (S)
  - Prior information
  - Pilot study
  - Rule-of-thumb estimate that standard deviation is 1/6 of the response range

Estimating Sample Size

- Determine a confidence level (Z_{CI})
  - For 90% confidence level, Z = 1.645
  - For 95% confidence level, Z = 1.96
  - For 99% confidence level, Z = 2.576

Estimating Sample Size

- Make a judgment about the desired magnitude of error (E)
  - How much “wiggle” room do you want?
  - For situations involving means (averages), the error is expressed in the same units as the average
  - For situations involving proportions (such as polls), what percent are you willing to be from the truth?

Estimating Sample Size (means)

If you wanted to learn the average age of employees within 3 years at a 95% confidence level:

- S =5 years (educated guess or prior study)
- Z = 1.96
- E = 3

$$n = \left(\frac{ZS}{E} \right)^2$$

$$n = \left(\frac{1.96 \times 5}{3} \right)^2$$

$$n = 10.67 \approx 11$$

Estimating Sample Size (means)

If you wanted to learn the average age of employees within 1 year at a 95% confidence level:

- S =5 years (educated guess or prior study)
- Z = 1.96
- E = 1

$$n = \left(\frac{ZS}{E}\right)^2$$

$$n = \left(\frac{1.96 \times 5}{1}\right)^2$$

$$n = 96.04 \approx 97$$

Estimating Sample Size (proportions)

If you wanted to learn the proportion of employees happy at work within 3% at a 95% confidence level:

- Z = 1.96
- E = .03

$$n = .25\left(\frac{Z}{E}\right)^2$$

$$n = .25\left(\frac{1.96}{.03}\right)^2$$

$$n = 1067.1 \approx 1068$$

Estimating Sample Size (proportions)

If you wanted to learn the proportion of employees happy at work within 5% at a 95% confidence level:

- Z = 1.96
- E = .05

$$n = .25\left(\frac{Z}{E}\right)^2$$

$$n = .25\left(\frac{1.96}{.05}\right)^2$$

$$n = 384.16 \approx 385$$

A Big, Big Warning!

- The *n* calculated in the previous formulas is the number of questionnaires you need to get back completed
- You will probably <u>not</u> (ever) get back all of the questionnaires you distribute
- Adjust sample size <u>upward</u> to compensate for anticipated response rate
- For a typical mailed survey, a 10-20% response is likely, so you will need to distribute 5 to 10 times more surveys than needed

Using Excel to Compute Sample Sizes

Sample Size Determination for Proportion			
Estimate of True Proportion* =	**0.50**		
Sampling Error =	**0.05**		
Confidence Level =	**95.00%**	Sample Size Needed =	**385**
Finite Population Size =	**500**	Sample Size Needed =	**218**
**Use .50 for the "estimate of true proportion" when unknown.*			

Sample Size Determination for Mean			
Population Standard Deviation =	**5**		
Sampling Error =	**1**		
Confidence Level =	**95.00%**	Sample Size Needed =	**97**
Finite Population Size =	**500**	Sample Size Needed =	**81**

Excel file – Confidence Intervals & Sample Sizes located at http://home.attbi.com/~jimandkaren/jim/resources

Choose a Data Collection Technique

Data Collection Technique

	Personal	**Phone**	**Mail**	**E-Mail**
Costs	*High*	*Med*	*Low*	*Very low*
Time required	*Medium*	*Low*	*High*	*Low*
Data quantity per respondent	*High*	*Med*	*Low*	*Med*
Reaches widespread sample	*No*	*Maybe*	*Yes*	*Yes*
Reaches special locations	*Yes*	*Maybe*	*No*	*Yes*
Interaction with respondents	*Yes*	*Maybe*	*No*	*No*
Degree of interviewer bias	*High*	*Med*	*Low*	*Low*
Severity of non-response bias	*Low*	*Low*	*High*	*High*
Presentation of visual stimuli	*Yes*	*No*	*Maybe*	*Yes*

Selecting the Sample

Select the Sample

- Random number table OR rand() function in Excel
- Use last digits of SSN for systematic sample
- Randomly select departments and survey all members (cluster sampling)
- ???

Threats to Reliability and Validity

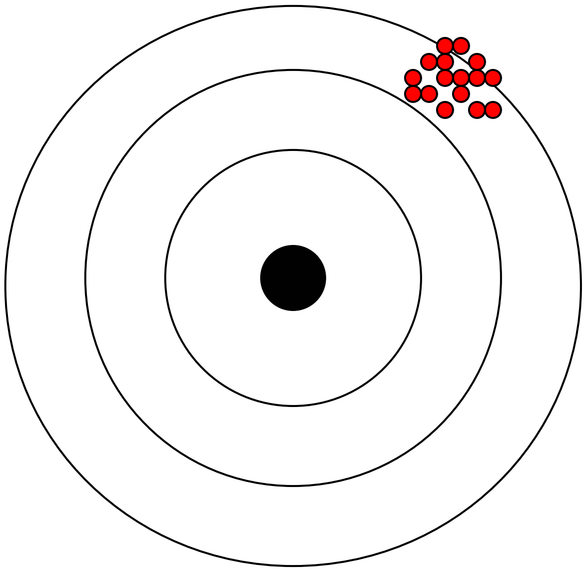
Some Definitions

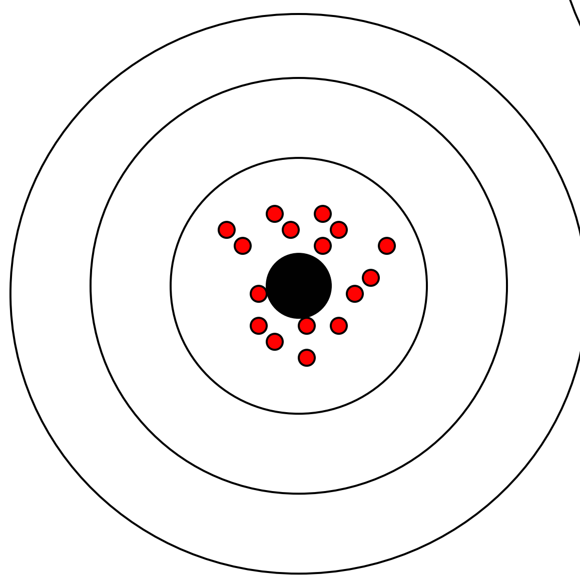
- Reliability – The degree to which measures are free from error and yield consistent results
  - How well a test consistently measures its intent
  - Can outcomes be replicated?
  - Survey scales and their effect on reliability
  - Test-retest method
  - Split-half method

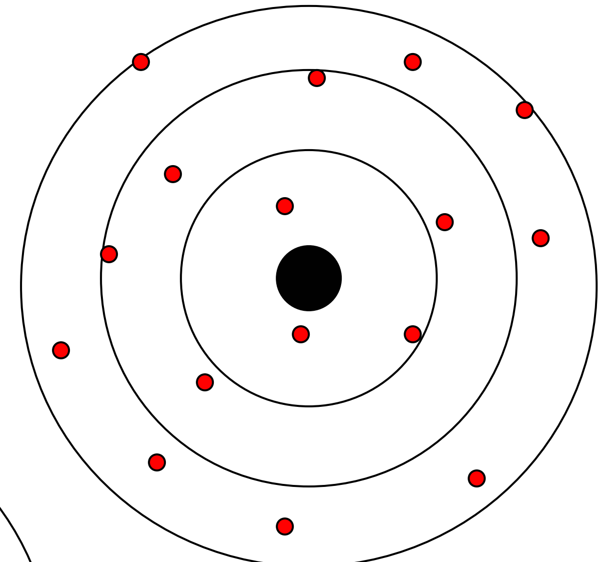
Some Definitions

- Validity – The ability of a scale or measuring instrument to measure what it is intended to measure
  - How well test measures what it is supposed to measure
  - Content validity = does the measure cover the range of content included within the intended concept?
  - Face validity = is the measure reflecting what it should?
  - Predictive validity = is the measure a valid predictor as expected?

Analogy of Validity vs. Reliability


Reliable but not valid


Valid and reliable


Valid but not reliable

Pitfalls Throughout the Sampling Hierarchy

1. Start with Total Population
2. Select Sampling Frame
   - Sampling Frame Error – Certain elements of the population are not included in the sampling frame
     - Include unwanted units OR exclude desired units
     - Using a telephone book to define the sample frame for residents of a particular neighborhood
     - Predicting elections with non-voters

Pitfalls (continued)

3. Select Sample
   - Random Sampling Error – The difference between the result of a sample and the result of a census due solely to observations chosen
     - 75% of a selected sample might be male when only 40% of the population is male
     - Caused by bad luck
     - Caused by sampling bias (i.e., *tendency to favor selection of certain data*)
     - "Alf Landon wins by Landslide"

Pitfalls (continued)

4. Gather Responses
   - Non-Response Error – Errors that cause the sample to be less than representative of the population
     - A disproportionately large group of males responds to a questionnaire
     - Respondents unavailable OR refuse to cooperate
     - Responses may represent extreme views
     - Most serious limitation of surveys
     - Don't confuse *response rate* with *sample size*
     - How can this be improved?

Beware of Voluntary Samples

- 900 / 800 number surveys
- "Opinions" site at malls
- News / sports polls on Web
- Talk shows
- Magazine votes
- Voluntary surveys may bring large response totals *(not the same as response rate),* but don't be satisfied with a large sample size. If it is not representative of the population, size will not compensate.

Business Card

Also an adjunct professor at

Webster University, Capella University, Jacksonville University, Nova Southeastern University, Florida Community College at Jacksonville, Tiffin University & Central Michigan University