Research Methods

William G. Zikmund

Determination of Sample Size

What does Statistics Mean?

- Descriptive statistics
  - Number of people
  - Trends in employment
  - Data
- Inferential statistics
  - Make an inference about a population from a sample

Population Parameter Versus Sample Statistics

Population Parameter

- Variables in a population
- Measured characteristics of a population
- Greek lower-case letters as notation

Sample Statistics

- Variables in a sample
- Measures computed from data
- English letters for notation

Making Data Usable

- Frequency distributions
- Proportions
- Central tendency
  - Mean
  - Median
  - Mode
- Measures of dispersion

Frequency Distribution of Deposits

Amount	Frequency (number of people making deposits in each range)
less than $3,000	499
$3,000 - $4,999	530
$5,000 - $9,999	562
$10,000 - $14,999	718
$15,000 or more	811
	3,120

Percentage Distribution of Amounts of Deposits

Amount	Percent
less than $3,000	16
$3,000 - $4,999	17
$5,000 - $9,999	18
$10,000 - $14,999	23
$15,000 or more	26
	100

Probability Distribution of Amounts of Deposits

Amount	Probability
less than $3,000	.16
$3,000 - $4,999	.17
$5,000 - $9,999	.18
$10,000 - $14,999	.23
$15,000 or more	.26
	1.00

Population Mean

$$\mu = \frac{\Sigma X_i}{N}$$

Measures of Central Tendency

- **Mean** - arithmetic average
  - μ, Population; $\overline{X}$, sample
- **Median** - midpoint of the distribution
- **Mode** - the value that occurs most often

Sample Mean

$$\overline{X} = \frac{\sum X_i}{n}$$

Number of Sales Calls Per Day by Salespersons

Salesperson	Number of Sales calls
Mike	4
Patty	3
Billie	2
Bob	5
John	3
Frank	3
Chuck	1
Samantha	5
	26

Sales for Products A and B, Both Average 200

Product A	Product B
196	150
198	160
199	176
199	181
200	192
200	200
200	201
201	202
201	213
201	224
202	240
202	261

Measures of Dispersion

- The range
- Standard deviation

Measures of Dispersion or Spread

- Range
- Mean absolute deviation
- Variance
- Standard deviation

The Range as a Measure of Spread

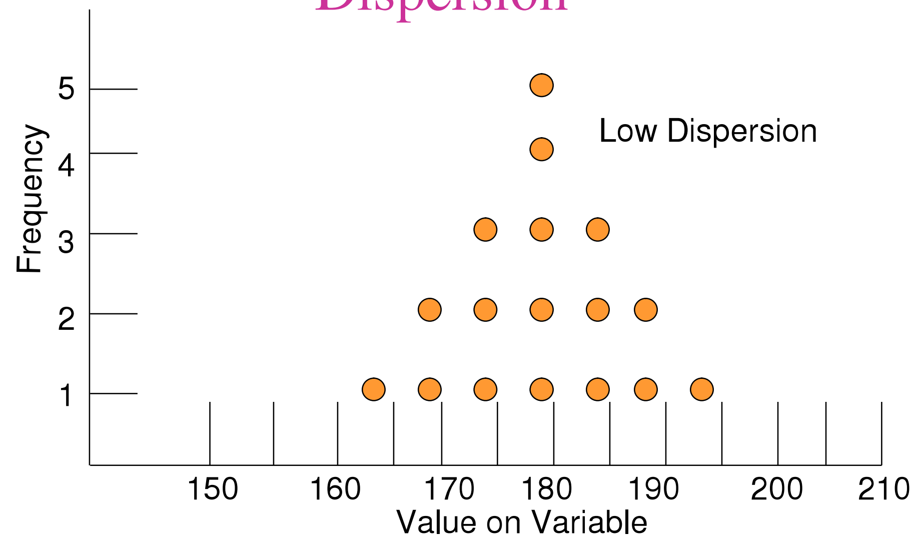
- The range is the **distance** between the smallest and the largest value in the set.
- Range = largest value – smallest value

Deviation Scores

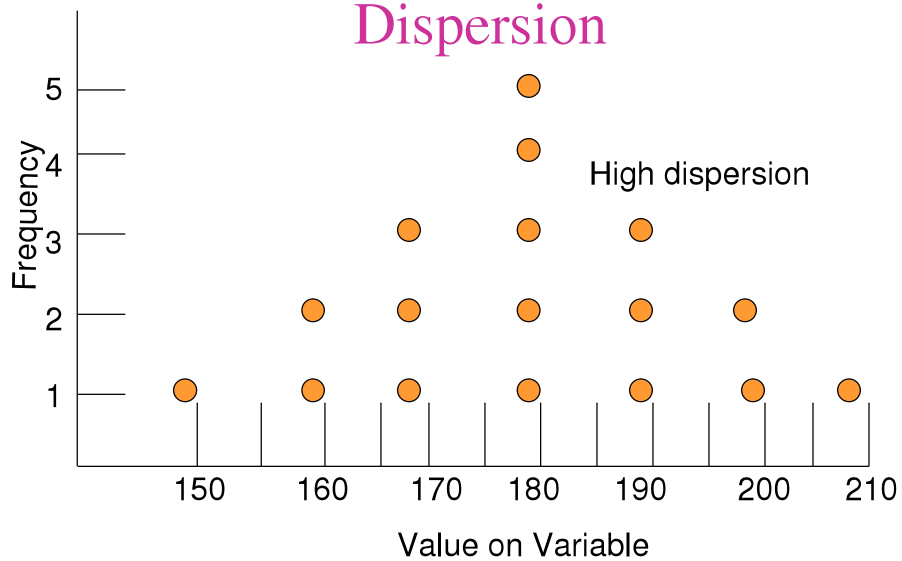
- The differences between each observation value and the mean:

$$d_i = x_i - \bar{x}$$

Low Dispersion Verses High Dispersion



Low Dispersion Verses High Dispersion



Average Deviation

$$\frac{\sum(X_i - \overline{X})}{n} = 0$$

Mean Squared Deviation

$$\frac{\sum(X_i - \overline{X})^2}{n}$$

The Variance

Population

σ^2

Sample

S^2

Variance

$$S^2 = \frac{\Sigma(X - \overline{X})^2}{n - 1}$$

Variance

- The variance is given in squared units
- The standard deviation is the square root of variance:

Population Standard Deviation

$$\sigma = \sqrt{\sigma^2}$$

Sample Standard Deviation

$$S = \sqrt{\frac{\Sigma\left(X_i - \overline{X}\right)^2}{n-1}}$$

Sample Standard Deviation

$$S = \sqrt{S^2}$$

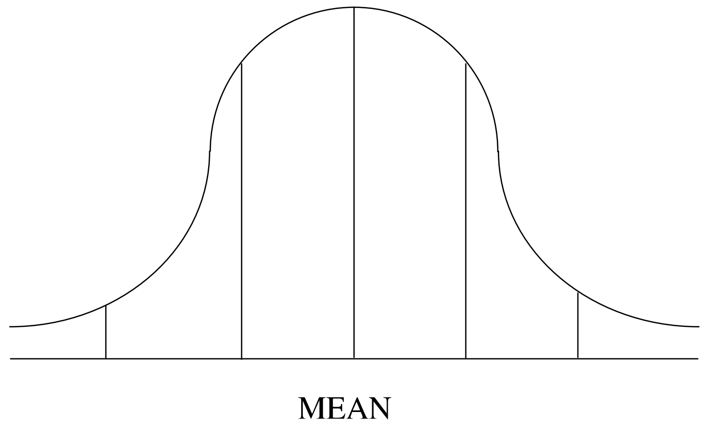
Sample Standard Deviation

$$S = \sqrt{\frac{\Sigma\left(X_i - \overline{X}\right)^2}{n-1}}$$

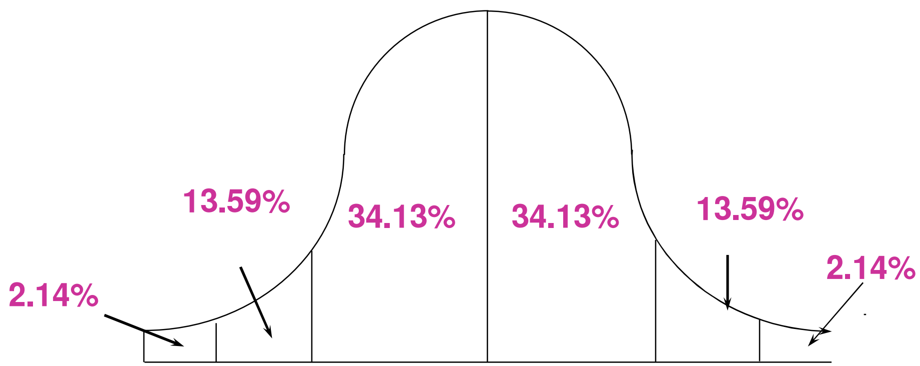
The Normal Distribution

- Normal curve
- Bell shaped
- Almost all of its values are within plus or minus 3 standard deviations
- I.Q. is an example

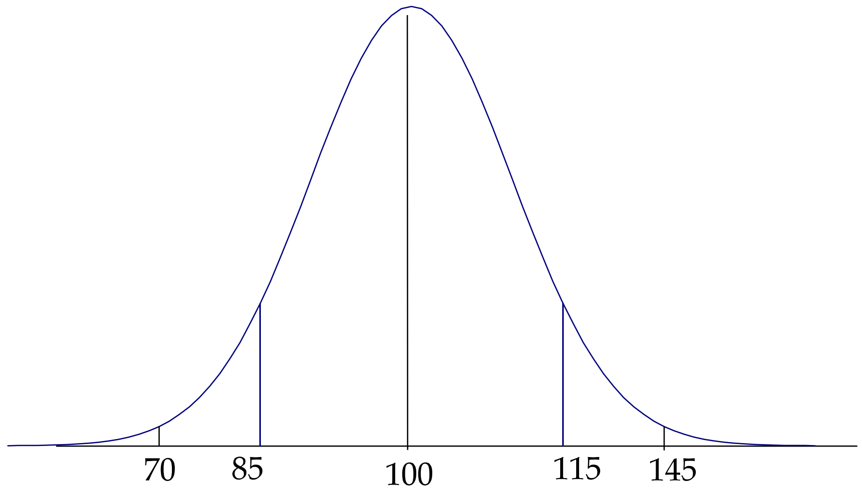
Normal Distribution



Normal Distribution



Normal Curve: IQ Example



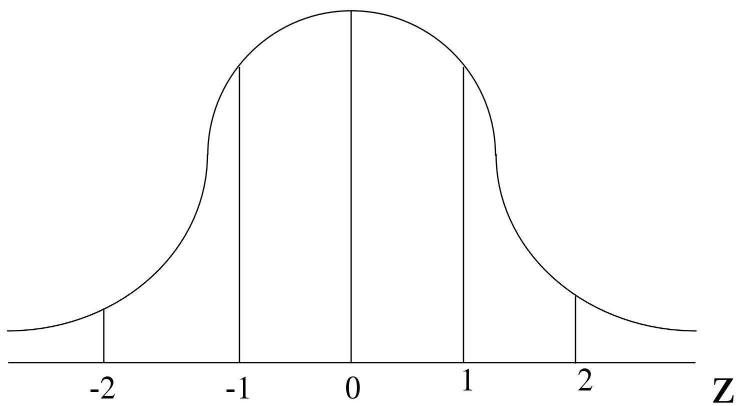
Standardized Normal Distribution

- Symetrical about its mean
- Mean identifies highest point
- Infinite number of cases - a continuous distribution
- Area under curve has a probability density = 1.0
- Mean of zero, standard deviation of 1

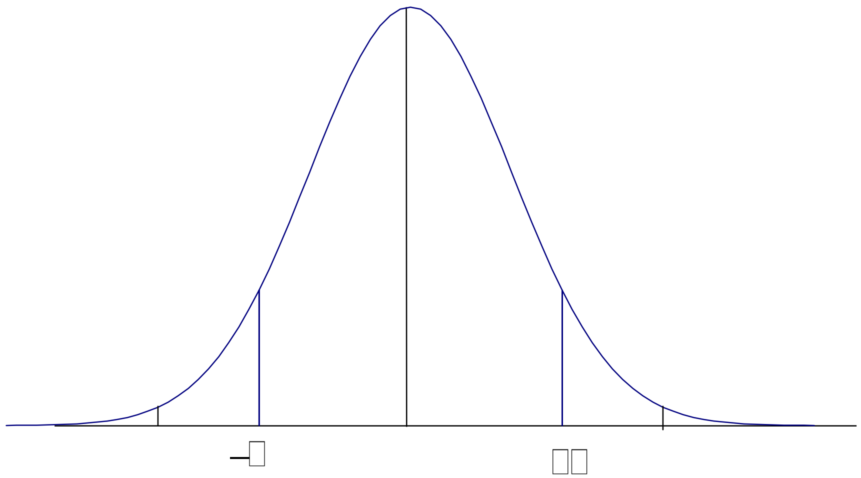
Standard Normal Curve

- The curve is bell-shaped or symmetrical
- About 68% of the observations will fall within 1 standard deviation of the mean
- About 95% of the observations will fall within approximately 2 (1.96) standard deviations of the mean
- Almost all of the observations will fall within 3 standard deviations of the mean

A Standardized Normal Curve



The Standardized Normal is the Distribution of Z



Standardized Scores

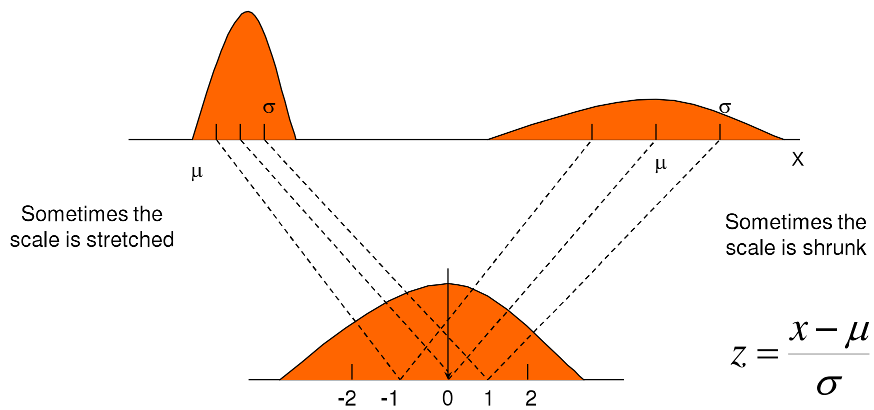
$$z = \frac{x - \mu}{\sigma}$$

Standardized Values

- Used to compare an individual value to the population mean in units of the standard deviation

$$z = \frac{x - \mu}{\sigma}$$

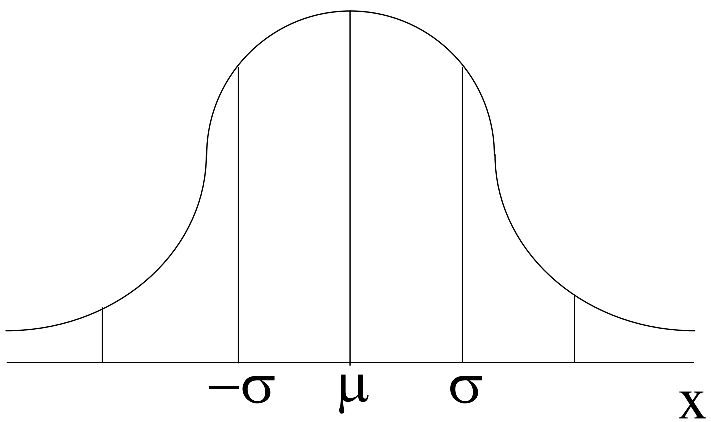
Linear Transformation of Any Normal Variable Into a Standardized Normal Variable



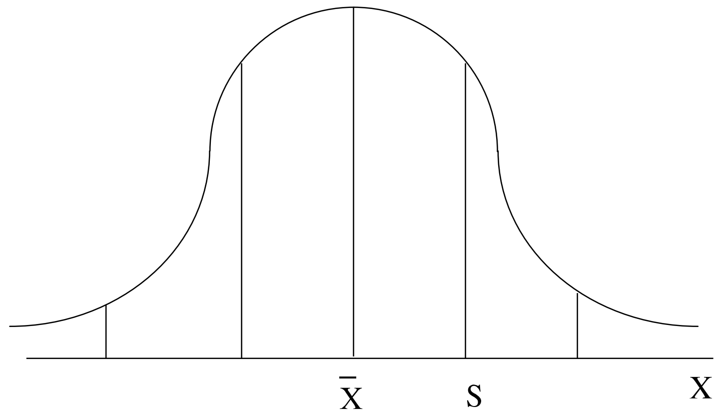
$$z = \frac{x - \mu}{\sigma}$$

- **Population distribution**
- **Sample distribution**
- **Sampling distribution**

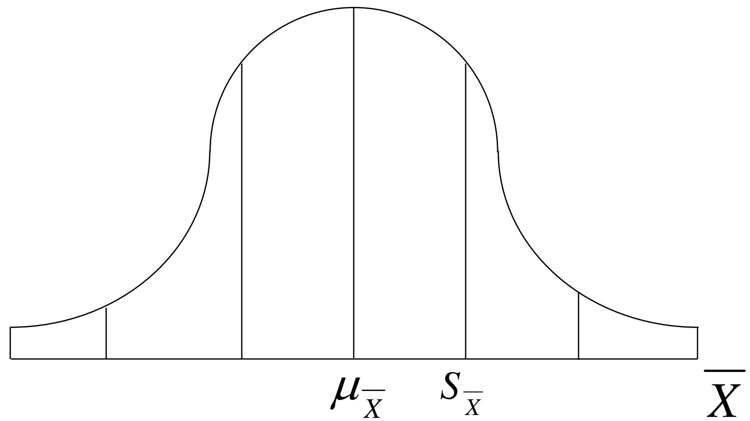
Population Distribution



Sample Distribution



Sampling Distribution



Standard Error of the Mean

- Standard deviation of the sampling distribution

Central Limit Theorem

Standard Error of the Mean

$$S_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Distribution	Mean	Standard Deviation
Population	μ	σ
Sample	$\overline{X}$	S
Sampling	$\mu_{\bar{X}}$	$S_{\bar{X}}$

Parameter Estimates

- Point estimates
- Confidence interval estimates

Confidence Interval

$$\mu = \overline{X} \pm a \text{ small sampling error}$$

$$\text{SMALL SAMPLING ERROR} = Z_{cl} S_{\overline{X}}$$

$$E = Z_{cl} S_{\overline{X}}$$

$$\mu = \overline{X} \pm \mathrm{E}$$

Estimating the Standard Error of the Mean

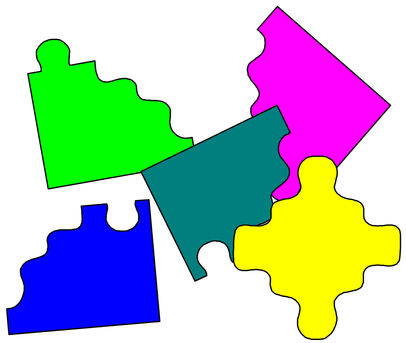
$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

$$\mu = \overline{X} \pm Z_{cl} \frac{S}{\sqrt{n}}$$

Random Sampling Error and Sample Size are Related

Sample Size

- Variance (standard deviation)
- Magnitude of error
- Confidence level



Sample Size Formula

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

Sample Size Formula - Example

Suppose a survey researcher, studying expenditures on lipstick, wishes to have a 95 percent confident level (Z) and a range of error (E) of less than $2.00. The estimate of the standard deviation is $29.00.

Sample Size Formula - Example

$$n = \left(\frac{zs}{E}\right)^2 = \left[\frac{(1.96)(29.00)}{2.00}\right]^2$$

$$= \left[\frac{56.84}{2.00}\right]^2 = (28.42)^2 = 808$$

Sample Size Formula - Example

Suppose, in the same example as the one before, the range of error (E) is acceptable at $4.00, sample size is reduced.

Calculating Sample Size

99% Confidence

$$n = \left[\frac{(2.57)(29)}{2}\right]^2 = \left[\frac{74.53}{2}\right]^2 = [37.265]^2 = 1389$$

$$n = \left[\frac{(2.57)(29)}{4}\right]^2 = \left[\frac{74.53}{4}\right]^2 = [18.6325]^2 = 347$$

Sample Size Formula - Example

$$n = \left(\frac{zs}{E}\right)^2 = \left[\frac{(1.96)(29.00)}{4.00}\right]^2$$

$$= \left[\frac{56.84}{4.00}\right]^2 = (14.21)^2 = 202$$

Standard Error of the Proportion

$$S_p = \sqrt{\frac{pq}{n}}$$

or

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

Confidence Interval for a Proportion

$$p \pm Z_{cl} S_p$$

Sample Size for a Proportion

$$n = \frac{Z^2 pq}{E^2}$$

$$n = \frac{z^2 pq}{E^2}$$

Where:

n = Number of items in samples

Z^2 = The square of the confidence interval in standard error units.

p = Estimated proportion of success

q = (1-p) or estimated the proportion of failures

E^2 = The square of the maximum allowance for error between the true proportion and sample proportion or zs_p squared.

Calculating Sample Size at the 95% Confidence Level

p = .6
q = .4

$$n = \frac{(1.96)^2(.6)(.4)}{(.035)^2} = \frac{(3.8416)(.24)}{001225} = \frac{.922}{.001225} = 753$$

UF UNIVERSITY of FLORIDA
IFAS Extension

Determining Sample Size[1]

Glenn D. Israel[2]

Perhaps the most frequently asked question concerning sampling is, "What size sample do I need?" The answer to this question is influenced by a number of factors, including the purpose of the study, population size, the risk of selecting a "bad" sample, and the allowable sampling error. Interested readers may obtain a more detailed discussion of the purpose of the study and population size in *Sampling The Evidence Of Extension Program Impact*, PEOD-5 (Israel, 1992). This paper reviews criteria for specifying a sample size and presents several strategies for determining the sample size.

SAMPLE SIZE CRITERIA

In addition to the purpose of the study and population size, three criteria usually will need to be specified to determine the appropriate sample size: the level of precision, the level of confidence or risk, and the degree of variability in the attributes being measured (Miaoulis and Michener, 1976). Each of these is reviewed below.

The Level Of Precision

The *level of precision*, sometimes called *sampling error*, is the range in which the true value of the population is estimated to be. This range is often expressed in percentage points, (e.g., ±5 percent), in the same way that results for political campaign polls are reported by the media. Thus, if a researcher finds that 60% of farmers in the sample have adopted a recommended practice with a precision rate of ±5%, then he or she can conclude that between 55% and 65% of farmers in the population have adopted the practice.

The Confidence Level

The *confidence* or *risk level* is based on ideas encompassed under the Central Limit Theorem. The key idea encompassed in the Central Limit Theorem is that when a population is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the true population value. Furthermore, the values obtained by these samples are distributed normally about the true value, with some samples having a higher value and some obtaining a lower score than the true population value. In a normal distribution, approximately 95% of the sample values are within two standard deviations of the true population value (e.g., mean).

In other words, this means that, if a 95% confidence level is selected, 95 out of 100 samples

1. This document is PEOD6, one of a series of the Agricultural Education and Communication Department, Florida Cooperative Extension Service, Institute of Food and Agricultural Sciences, University of Florida. Original publication date November 1992. Reviewed April 2009. Visit the EDIS Web Site at http://edis.ifas.ufl.edu.
2. Glenn D. Israel, associate professor, Department of Agricultural Education and Communication, and extension specialist, Program Evaluation and Organizational Development, Institute of Food and Agricultural Sciences (IFAS), University of Florida, Gainesville 32611.

will have the true population value within the range of precision specified earlier (Figure 1). There is always a chance that the sample you obtain does not represent the true population value. Such samples with extreme values are represented by the shaded areas in Figure 1. This risk is reduced for 99% confidence levels and increased for 90% (or lower) confidence levels.

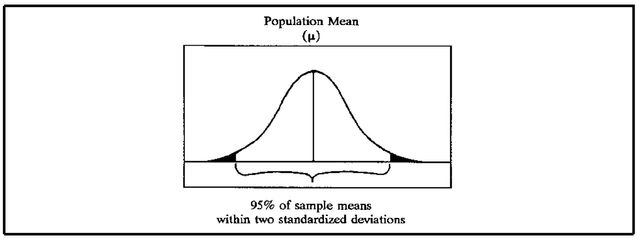


Figure 1.

Degree Of Variability

The third criterion, the *degree of variability* in the attributes being measured refers to the distribution of attributes in the population. The more heterogeneous a population, the larger the sample size required to obtain a given level of precision. The less variable (more homogeneous) a population, the smaller the sample size. Note that a proportion of 50% indicates a greater level of variability than either 20% or 80%. This is because 20% and 80% indicate that a large majority do not or do, respectively, have the attribute of interest. Because a proportion of .5 indicates the maximum variability in a population, it is often used in determining a more conservative sample size, that is, the sample size may be larger than if the true variability of the population attribute were used.

STRATEGIES FOR DETERMINING SAMPLE SIZE

There are several approaches to determining the sample size. These include using a census for small populations, imitating a sample size of similar studies, using published tables, and applying formulas to calculate a sample size. Each strategy is discussed below.

Using A Census For Small Populations

One approach is to use the entire population as the sample. Although cost considerations make this impossible for large populations, a census is attractive for small populations (e.g., 200 or less). A census eliminates sampling error and provides data on all the individuals in the population. In addition, some costs such as questionnaire design and developing the sampling frame are "fixed," that is, they will be the same for samples of 50 or 200. Finally, virtually the entire population would have to be sampled in small populations to achieve a desirable level of precision.

Using A Sample Size Of A Similar Study

Another approach is to use the same sample size as those of studies similar to the one you plan. Without reviewing the procedures employed in these studies you may run the risk of repeating errors that were made in determining the sample size for another study. However, a review of the literature in your discipline can provide guidance about "typical" sample sizes which are used.

Using Published Tables

A third way to determine sample size is to rely on published tables which provide the sample size for a given set of criteria. Table 1 and Table 2 present sample sizes that would be necessary for given combinations of precision, confidence levels, and variability. Please note two things. First, these sample sizes reflect the number of *obtained* responses, and not necessarily the number of surveys mailed or interviews planned (this number is often increased to compensate for nonresponse). Second, the sample sizes in Table 2 presume that the attributes being measured are distributed normally or nearly so. If this assumption cannot be met, then the entire population may need to be surveyed.

Using Formulas To Calculate A Sample Size

Although tables can provide a useful guide for determining the sample size, you may need to calculate the necessary sample size for a different combination of levels of precision, confidence, and variability. The fourth approach to determining sample size is the application of one of several formulas (Equation 5 was used to calculate the sample sizes in Table 1 and Table 2).

$$\textbf{Equation 3:} \quad n = \frac{N}{1 + N(e)^2}$$

Equation 5.

Formula For Calculating A Sample For Proportions

For populations that are large, Cochran (1963:75) developed the Equation 1 to yield a representative sample for proportions.

$$\textbf{Equation 1:} \quad n_0 = \frac{Z^2pq}{e^2}$$

Equation 1.

Which is valid where n_0 is the sample size, Z^2 is the abscissa of the normal curve that cuts off an area at the tails (1 - equals the desired confidence level, e.g., 95%)[1], e is the desired level of precision, p is the estimated proportion of an attribute that is present in the population, and q is 1-p. The value for Z is found in statistical tables which contain the area under the normal curve.

To illustrate, suppose we wish to evaluate a state-wide Extension program in which farmers were encouraged to adopt a new practice. Assume there is a large population but that we do not know the variability in the proportion that will adopt the practice; therefore, assume p=.5 (maximum variability). Furthermore, suppose we desire a 95% confidence level and ±5% precision. The resulting sample size is demonstrated in Equation 2.

$$n_0 = \frac{Z^2pq}{e^2} = \frac{(1.96)^2(.5)(.5)}{(.05)^2} = 385 \text{ farmers}$$

Equation 2.

Finite Population Correction For Proportions

If the population is small then the sample size can be reduced slightly. This is because a given sample size provides proportionately more information for a small population than for a large population. The sample size (n_0) can be adjusted using Equation 3 .

$$\textbf{Equation 2:} \quad n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

Equation 3.

Where n is the sample size and N is the population size.

Suppose our evaluation of farmers' adoption of the new practice only affected 2,000 farmers. The sample size that would now be necessary is shown in Equation 4.

$$n = \frac{n_0}{1+\frac{(n_0 - 1)}{N}} = \frac{385}{1+\frac{(385 - 1)}{2000}} = 323 \text{ farmers}$$

Equation 4.

As you can see, this adjustment (called the finite population correction) can substantially reduce the necessary sample size for small populations.

A Simplified Formula For Proportions

Yamane (1967:886) provides a simplified formula to calculate sample sizes. This formula was used to calculate the sample sizes in Tables 2 and 3 and is shown below. A 95% confidence level and P = .5 are assumed for Equation 5 .

$$n = \frac{N}{1 + N(e)^2} = \frac{2000}{1 + 2000(.05)^2} = 333 \text{ farmers}$$

Equation 5.

Where n is the sample size, N is the population size, and e is the level of precision. When this formula is applied to the above sample, we get Equation 6.

Formula For Sample Size For The Mean

The use of tables and formulas to determine sample size in the above discussion employed proportions that assume a dichotomous response for the attributes being measured. There are two methods to determine sample size for variables that are polytomous or continuous. One method is to combine responses into two categories and then use a sample size based on proportion (Smith, 1983). The second method is to use the formula for the sample size for the mean. The formula of the sample size for the mean is similar to that of the proportion, except for the measure of variability. The formula for the mean employs ^{2} instead of (p x q), as shown in Equation 7.

Equation 4: $$n_0 = \frac{Z^2 \quad \sigma^2}{e^2}$$

Equation 7.

Where n_0 is the sample size, z is the abscissa of the normal curve that cuts off an area at the tails, e is the desired level of precision (in the same unit of measure as the variance), and ^{2} is the variance of an attribute in the population.

The disadvantage of the sample size based on the mean is that a "good" estimate of the population variance is necessary. Often, an estimate is not available. Furthermore, the sample size can vary widely from one attribute to another because each is likely to have a different variance. Because of these problems, the sample size for the proportion is frequently preferred[2].

OTHER CONSIDERATIONS

In completing this discussion of determining sample size, there are three additional issues. First, the above approaches to determining sample size have assumed that a simple random sample is the sampling design. More complex designs, e.g., stratified random samples, must take into account the variances of subpopulations, strata, or clusters before an estimate of the variability in the population as a whole can be made.

Another consideration with sample size is the number needed for the data analysis. If descriptive statistics are to be used, e.g., mean, frequencies, then nearly any sample size will suffice. On the other hand, a good size sample, e.g., 200-500, is needed for multiple regression, analysis of covariance, or log-linear analysis, which might be performed for more rigorous state impact evaluations. The sample size should be appropriate for the analysis that is planned.

In addition, an adjustment in the sample size may be needed to accommodate a comparative analysis of subgroups (e.g., such as an evaluation of program participants with nonparticipants). Sudman (1976) suggests that a minimum of 100 elements is needed for each major group or subgroup in the sample and for each minor subgroup, a sample of 20 to 50 elements is necessary. Similarly, Kish (1965) says that 30 to 200 elements are sufficient when the attribute is present 20 to 80 percent of the time (i.e., the distribution approaches normality). On the other hand, skewed distributions can result in serious departures from normality even for moderate size samples (Kish, 1965:17). Then a larger sample or a census is required.

Finally, the sample size formulas provide the number of responses that need to be obtained. Many researchers commonly add 10% to the sample size to compensate for persons that the researcher is unable to contact. The sample size also is often increased by 30% to compensate for nonresponse. Thus, the number of mailed surveys or planned interviews can be substantially larger than the number required for a desired level of confidence and precision.

ENDNOTES

1. The area corresponds to the shaded areas in the sampling distribution shown in Figure 1.

2. The use of the level of maximum variability (P=.5) in the calculation of the sample size for the proportion generally will produce a more conservative sample size (i.e., a larger one) than will be calculated by the sample size of the mean.

REFERENCES

Cochran, W. G. 1963. *Sampling Techniques*, 2nd Ed., New York: John Wiley and Sons, Inc.

Israel, Glenn D. 1992. *Sampling The Evidence Of Extension Program Impact*. Program Evaluation and Organizational Development, IFAS, University of Florida. PEOD-5. October.

Kish, Leslie. 1965. *Survey Sampling*. New York: John Wiley and Sons, Inc.

Miaoulis, George, and R. D. Michener. 1976. *An Introduction to Sampling*. Dubuque, Iowa: Kendall/Hunt Publishing Company.

Smith, M. F. 1983. *Sampling Considerations In Evaluating Cooperative Extension Programs*. Florida Cooperative Extension Service Bulletin PE-1. Institute of Food and Agricultural Sciences. University of Florida.

Sudman, Seymour. 1976. *Applied Sampling*. New York: Academic Press.

Yamane, Taro. 1967. *Statistics, An Introductory Analysis*, 2nd Ed., New York: Harper and Row.

Table 1.

Table 1. Sample size for ±3%, ±5%, ±7% and ±10% Precision Levels Where Confidence Level is 95% and P=.5.				
Size of	Sample Size (n) for Precision (e) of:			
Population	±3%	±5%	±7%	±10%
500	a	222	145	83
600	a	240	152	86
700	a	255	158	88
800	a	267	163	89
900	a	277	166	90
1,000	a	286	169	91
2,000	714	333	185	95
3,000	811	353	191	97
4,000	870	364	194	98
5,000	909	370	196	98
6,000	938	375	197	98
7,000	959	378	198	99
8,000	976	381	199	99
9,000	989	383	200	99
10,000	1,000	385	200	99
15,000	1,034	390	201	99
20,000	1,053	392	204	100
25,000	1,064	394	204	100
50,000	1,087	397	204	100
100,000	1,099	398	204	100

Table 1.

>100,000	1,111	400	204	100
a = Assumption of normal population is poor (Yamane, 1967). The entire population should be sampled.				

Table 2.

Table 2. Sample size for ±5%, ±7% and ±10% Precision Levels Where Confidence Level is 95% and P=.5.			
Size of	Sample Size (n) for Precision (e) of:		
Population	±5%	±7%	±10%
100	81	67	51
125	96	78	56
150	110	86	61
175	122	94	64
200	134	101	67
225	144	107	70
250	154	112	72
275	163	117	74
300	172	121	76
325	180	125	77
350	187	129	78
375	194	132	80
400	201	135	81
425	207	138	82
450	212	140	82