



Estimation Sampling Training

Data Quality & Integrity



Topics:

- ❑ Estimation Sampling Overview
- ❑ Terminology
- ❑ Concepts of Estimation Sampling
- ❑ Designing and Selecting a Sample
- ❑ Evaluating a Sample
- ❑ Tools Demonstration
- ❑ Case Study

Reference Material:

- ❑ Appendix A: Design Quality Assurance Checklist
 - ❑ Appendix B: Evaluation Quality Assurance Checklist
 - ❑ Appendix C: Mathematical Formulas
-



Estimation Sampling Overview



General Approaches to Sampling:

- Non-Statistical Sampling (i.e. judgmental)
 - *Items are not selected randomly*
 - *Results may not be extrapolated to the entire population*



General Approaches to Sampling (cont'd):

- ❑ Statistical Sampling:
 - *Uses statistical theory and laws of probability to select and evaluate the sample.*
 - *Enables the auditor to quantify the sampling risk.*



Statistical sampling still requires professional judgment at the planning, performance and evaluation stages!



General Approaches to Sampling (cont'd):

- ❑ Statistical Sampling (cont'd):
 - *Each item in the population has a known probability of selection*
 - *Items are selected randomly*
 - *Results from a statistical sample can be extrapolated to the entire population with some degree of confidence*



Statistical Sampling Methods

- ❑ Attribute Sampling
 - *Statistical sampling method that reaches a conclusion about a population in terms of a rate of occurrence.*

- ❑ Variables Sampling
 - *Statistical sampling method that reaches a conclusion on the continuous amounts of a population, either average, ratio or total.*

- ❑ Cumulative Monetary Amounts (CMA) Sampling, also known as Systematic Selection
 - *A “hybrid” statistical sampling method that uses attribute theory to express a conclusion in dollar amounts rather than as a rate of occurrence.*



Key Concepts

- ❑ Sampling is the selection and implementation of statistical observations in order to estimate properties of an underlying population.

- ❑ A sample is a subset of a population that is obtained through some process, possibly random selection or selection based on a certain set of criteria, for the purposes of investigating the properties of the underlying parent population.

- ❑ An estimate is a value obtained by applying the estimation rule to a particular sample.

- ❑ An estimator is a rule that tells how to calculate an estimate based on the information in the sample.



Estimation Sampling should be considered when:

- ❑ The examination of the entire population requires detailed, time-consuming, or extensive manual procedures to be performed on every item.

- ❑ Reasonable estimates are an acceptable alternative to an examination of the entire population.



Estimation Sampling can be used to:

- ❑ Confirm that a population recorded value is correct within a specific range.

- ❑ Estimate a population value with the intention of comparing the estimated range to the recorded amount .

- ❑ Estimate a population value within a range where the sampling units have no recorded amount.



Benefits of Estimation Sampling

- ❑ Reduced Cost

- ❑ Increased Efficiency



Benefits of Estimation Sampling

□ Improved Accuracy

- *Sampling Risk: The risk that the auditor's conclusions based on the results of a sample may be different from the conclusions that would have been reached if the tests were applied to the entire population.*
- *Non-Sampling Risk: All other detection risks not related to sampling. Although not measurable, it can be minimized by adequate planning and supervision, training, QA, etc.*
- *Reduces cost, labor and time associated with a review.*
- *Produces objective.*



Methodology of Estimation Sampling

- ❑ Determine the sampling application objective
- ❑ Define the population, sampling unit, and characteristic to be tested
- ❑ Determine the sampling method
- ❑ Define the sampling parameters
 - *including planned confidence level and precision*
- ❑ Determine the sample size
 - *sample should be representative of the original population*
- ❑ Select the sample
- ❑ Examine the sample items
- ❑ Evaluate the sample results
- ❑ Report the results



Terminology



Key Statistical Terms

□ Population

- *The collection of items from which the sample is selected, and hence, to which any conclusions drawn will be applicable.*

□ Sampling Unit

- *The items that comprise the population and are individually available for selection and testing.*

□ Bias

- *The term that refers to how far the average statistic lies from the parameter it is estimating, that is, the error which arises when estimating a quantity. Errors from chance will cancel each other out in the long run, those from bias will not.*



Key Statistical Terms (cont'd)

□ Standard Deviation

- *A statistical measure of the extent to which individual values of a variable differ from their arithmetic mean.*

□ Expected Rate of Deviations

- *Error rate anticipated before sampling begins*

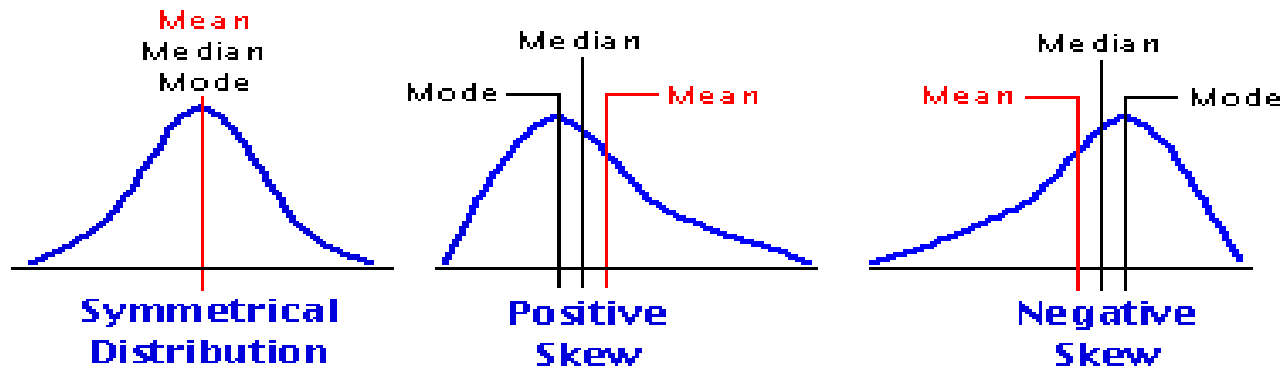
□ Standard Error

- *The standard deviation of a sample divided by \sqrt{n} (where n is the sample size) . It estimates the standard deviation of the sample mean based on the population mean. (Note that while this definition makes no reference to a normal distribution, many uses of this quantity implicitly assume such a distribution.)*

Key Statistical Terms (cont'd)

□ Skewness

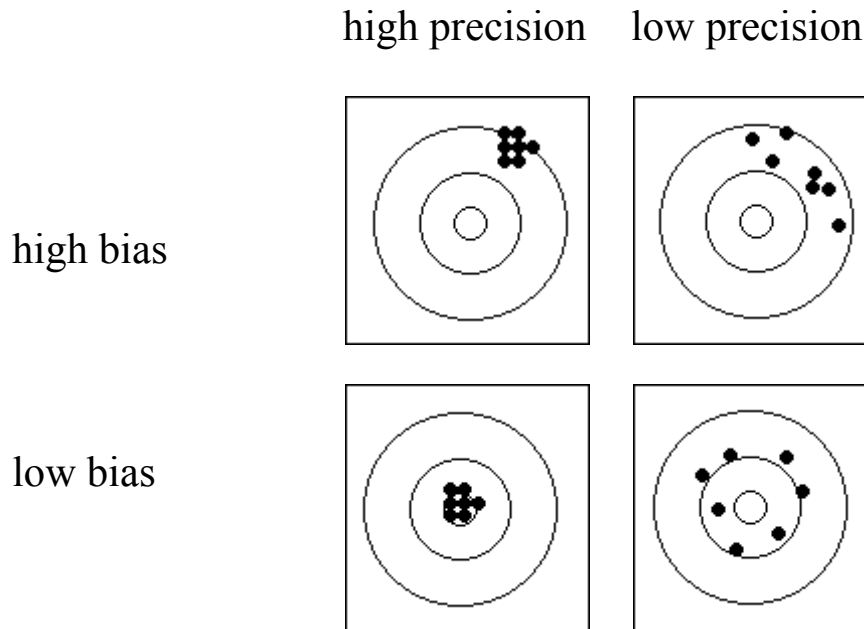
- *A measure of the degree of asymmetry of a distribution. If the left tail of the distribution curve is more pronounced than the right tail then the function is said to have negative skewness; if the reverse is true then the function has positive skewness; and if the two tails are equal then the curve has zero skewness (normal distribution).*



Key Statistical Terms (cont'd)

□ Precision

- A measure of the probable closeness between a sample point estimate and the value of the corresponding unknown population characteristic.*





Key Statistical Terms (cont'd)

- ❑ Upper Precision Limit (also called “maximum tolerable error”)
 - *Estimate of the maximum occurrence rate of the attribute in the population at the stated confidence level*

- ❑ Consistent Estimate
 - *Estimate that is guaranteed to exactly equal the true population value when sample size consists of the entire population*



Key Statistical Terms (cont'd)

□ Confidence Level

- *The probability that the associated confidence interval of a point estimate (i.e., the estimated precision range) will include the true population value.*
- *The width of the interval*
 - *increases as the confidence coefficient increases*
 - *decreases as the sample size increases*
- *90% of the estimates are within 1.65 standard errors of the true population value; 95% of the estimates are within 1.96 standard errors of the true population value, etc. (Usually, the 95% confidence level is a good rule of thumb.)*



Key Statistical Terms (cont'd)

☐ Confidence Level (cont'd)

Confidence Interval	d	Result
90%	1.65	$\bar{x} \pm 1.65s$
95%	1.96	$\bar{x} \pm 1.96s$
99%	2.58	$\bar{x} \pm 2.58s$

☐ Confidence Interval

- *The range of values, estimated at the specified confidence level, between which you expect the true population value to fall. The range is measured in terms of **precision** about a point estimate.*



Key Statistical Terms (cont'd)

❑ Point Estimate

- *The mid point of the confidence interval.*

❑ PPS Variables Estimate

- *An estimate derived using the average of each sample item's ratio of target value to source value.*

❑ Significance Test

- *Is a test for determining the probability that a given result could not have occurred by chance (its significance).*

The -score associated with the the observation of a random variable is given by



Key Statistical Terms (cont'd)

$$z_i \equiv \frac{x_i - \bar{x}}{\sigma}$$

where \bar{x} is the mean and σ the standard deviation of all observations

x_1, \dots, x_n



Key Statistical Terms (cont'd)

□ Statistical Test

is a test used to determine the statistical significance of an observation.

A significance test is performed to determine if an observed value of a statistic differs enough from a hypothesized value of a parameter to draw the inference that the hypothesized value of the parameter is not the true value. The hypothesized value of the parameter is called the "null hypothesis." A significance test consists of calculating the probability of obtaining a statistic as different or more different from the null hypothesis (given that the null hypothesis is correct) than the statistic obtained in the sample. If this probability is sufficiently low, then the difference between the parameter and the statistic is said to be "statistically significant."

Just how low is sufficiently low? The choice is somewhat arbitrary but by convention levels of .05 and .01 are most commonly used.

For instance, an experimenter may hypothesize that the size of a food reward does not affect the speed a rat runs down an alley. One group of rats receives a large reward and another receives a small reward for running the alley. Suppose the mean running time for the large reward were 1.5 seconds and the mean running time for the small reward were 2.1 seconds.

Two main types of error can occur:

- 1. A type I error occurs when a false negative result is obtained in terms of the null hypothesis by obtaining a false positive measurement.*
 - 2. A type II error occurs when a false positive result is obtained in terms of the null hypothesis by obtaining a false negative measurement.*
-



Key Statistical Terms (cont'd)

- ❑ Source Variable
 - *The population variable that is the basis for selecting the sample*

- ❑ Target Variable
 - *The population variable containing the value of sampling interest. This is the variable that will be estimated.*

- ❑ Stratification
 - *a process of grouping members of the population into relatively homogeneous subgroups (strata) before sampling.*



Key Statistical Terms (cont'd)

□ Statistical Analysis

- *Analysis that provides methods for Design: planning and carrying out research studies, Description: Summarizing and exploring data and Inference: making predictions or generalizing about phenomena represented by the data*
- *Can be classified as either descriptive or inferential, according to whether its main purpose is to describe data or to make predictions.*

□ Statistical Inference

- *Predictions made using data*



Key Statistical Terms (cont'd)

- Variable sampling
 - *Tests items which can take any value within a continuous range*
 - *Applies to populations made up of dollars, pounds, days, etc.*
 - *Population items have variable characteristic (e.g. the test indicates that the population value is between \$5,100,000 and \$5,300,000 at a 90% confidence level)*



Key Statistical Terms (cont'd)

- Variable sampling (cont'd)
 - *Includes mean per unit, difference, ratio, regression and PPS variables*
 - *Estimate average or total value of the population (e.g. population total value, total misstatement, total allowance, average ratio)*
 - *Two-sided estimates are typical (i.e. minimum and maximum values at specified confidence level)*
 - *Sample is designed to estimate an average ratio or total.*



Key Statistical Terms (cont'd)

□ Attribute Sampling

- *Estimate occurrence rate of a characteristic*
- *Characteristic is present or not*
 - *Correct or incorrect*
- *Random selection*
- *Statements such as:*
 - *Test indicates population error rate is no more than 10% at a 90% confidence level*
- *One-sided estimates typical (e.g. maximum error rate)*



Key Statistical Terms (cont'd)

- CMA Sampling (also known as Dollar Unit Sampling)
 - *Tests items which can take any value within a continuous range*
 - *Estimate maximum overstatement amount*
 - *Few monetary errors expected*
 - *PPS selection*



Key Statistical Terms (cont'd)

□ CMA Sampling (also known as Dollar Unit Sampling) (cont'd)

- *Statements such as:*
 - *Test indicates population is overstated by no more than \$500,000 at a 90% confidence level*
- *Usually one-sided estimate of maximum overstatement*
- *Estimate is conservative*
 - *Not useful if many errors*
 - *No statistical basis for audit adjustment*



Key Statistical Terms (cont'd)

□ Degrees of Freedom

- *a measure of the number of independent pieces of information on which a parameter estimate is based, equal to the number of observations (values) minus the number of additional parameters estimated for that calculation. As the number of parameters to be estimated increases, the degrees of freedom available decreases. It can be thought of two ways: in terms of sample size and in terms of dimensions and parameters.*

□ T Distribution

- *a probability distribution that arises in the problem of estimating the mean of a normally distributed population when the sample size is small. It is the basis of the t-tests for the statistical significance of the difference between two sample means, and for confidence intervals for the difference between two population means.*



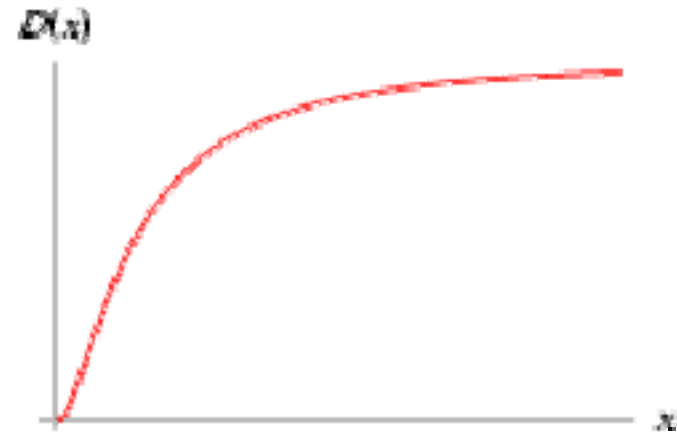
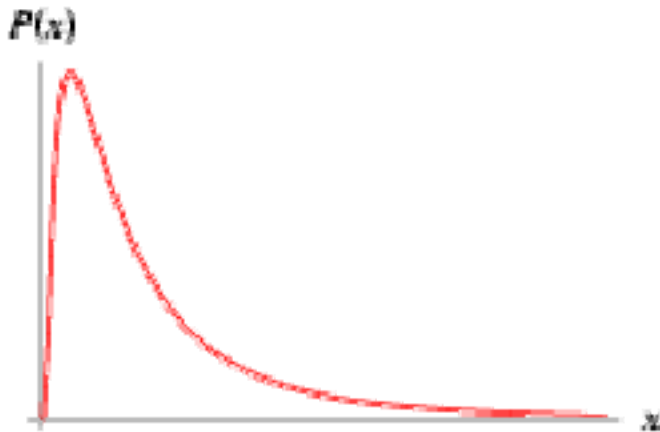
Key Statistical Concepts



Continuous Distributions

The "law of large numbers" is one of several theorems expressing the idea that as the number of trials of a random process increases, the percentage difference between the expected and actual values goes to zero.

Log Normal Distribution





Continuous Distributions (Cont'd)

A continuous distribution in which the logarithm of a variable has a normal distribution. It is a general case of Gibrat's distribution, to which the log normal distribution reduces with and . A log normal distribution results if the variable is the product of a large number of independent, identically-distributed variables in the same way that a normal distribution results if the variable is the sum of a large number of independent, identically-distributed variables.

The probability density and cumulative distribution functions for the log normal distribution are

$$P(x) = \frac{1}{S\sqrt{2\pi}x} e^{-(\ln x - M)^2 / (2S^2)} \quad 1)$$

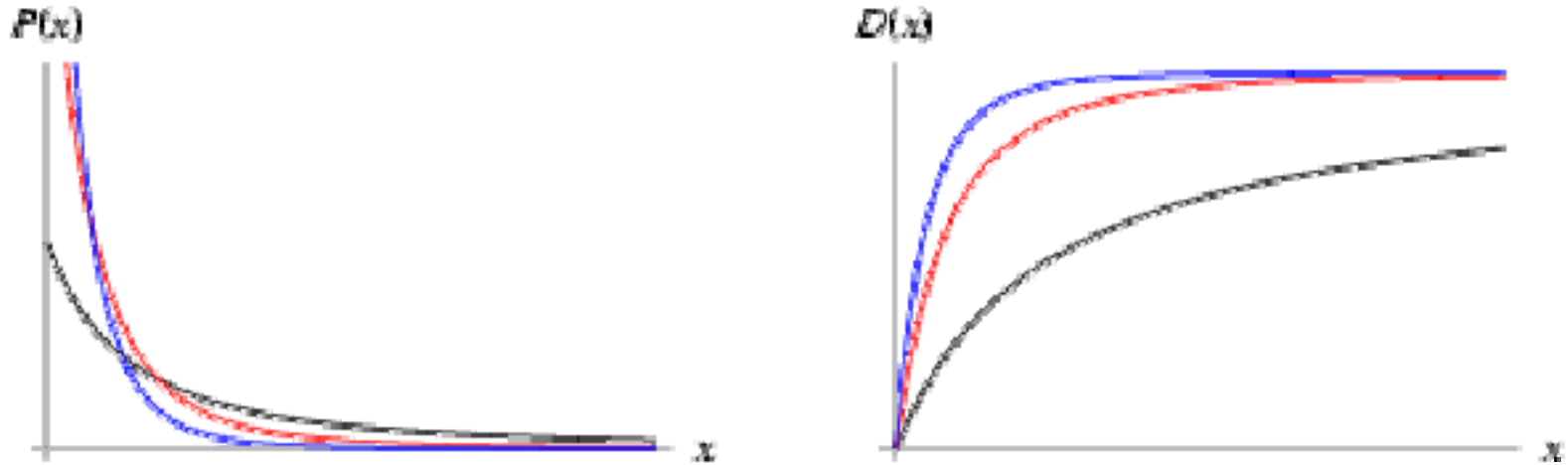
$$D(x) = \frac{1}{2} \left[1 + \operatorname{erf} \left(\frac{\ln x - M}{S\sqrt{2}} \right) \right] \quad 2)$$

where $\operatorname{erf}(x)$ is the *erf* function



Continuous Distributions (Cont'd)

Pareto Distribution



The distribution with probability density function and distribution function



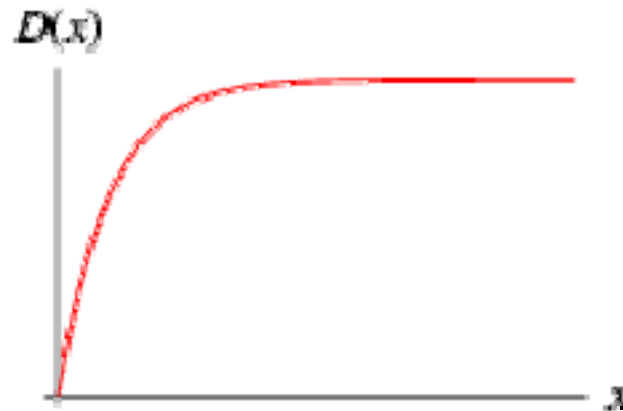
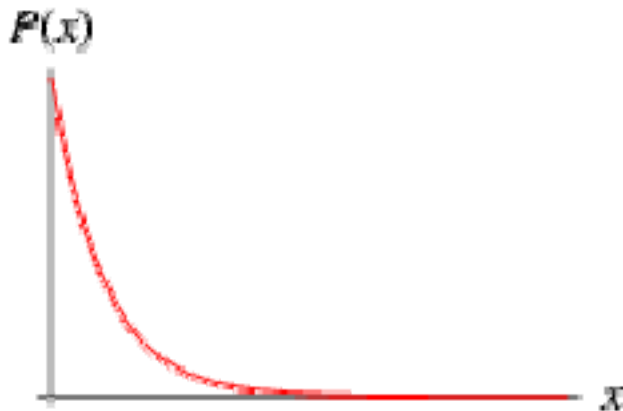
Continuous Distributions (Cont'd)

$$F(x) = \frac{\alpha b^\alpha}{x^{\alpha+1}} \quad 1)$$

$$D(x) = 1 - \left(\frac{b}{x}\right)^\alpha \quad 2)$$

defined over the interval $x \geq b$

Exponential Distribution





Continuous Distributions (Cont'd)

Given a Poisson distribution with rate of change λ , the distribution of waiting times between successive changes (with $t = 0$) is:

$$D(x) = P(X \leq x) \quad 1)$$

$$= 1 - P(X > x) \quad 2)$$

$$= 1 - e^{-\lambda x} \quad 3)$$

and the probability distribution function is

$$P(x) = D'(x) = \lambda e^{-\lambda x} \quad 4)$$

It is implemented in Mathematica as `ExponentialDistribution[lambda]` in the Mathematica add-on package

`Statistics`ContinuousDistributions`` (which can be loaded with the command `<<Statistics``).



Continuous Distributions (Cont'd)

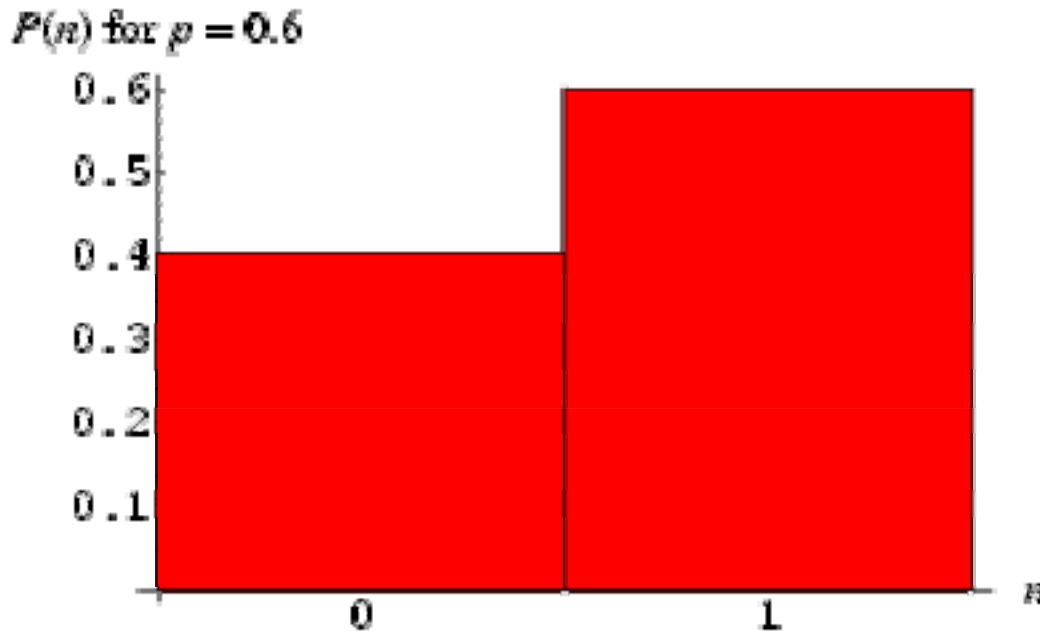
The exponential distribution is the only continuous memoryless random distribution. It is a continuous analog of the geometric distribution.

This distribution is properly normalized since

$$\int_0^{\infty} P(x) dx = \lambda \int_0^{\infty} e^{-\lambda x} dx = 1.$$



Discrete Distributions



The Bernoulli distribution is a discrete distribution having two possible outcomes labelled by $n=0$ and $n=1$ in which ("success") occurs with probability p and n ("failure") occurs with probability $q=1-p$, where $0 < p < 1$. It therefore has probability function



Discrete Distributions (Cont'd)

$$P(x) = \begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1, \end{cases} \quad 1)$$

which can also be written

$$P(x) = p^x (1 - p)^{1-x}. \quad 2)$$

The corresponding distribution function is

$$D(x) = \begin{cases} 1 - p & \text{for } x = 0 \\ 1 & \text{for } x = 1. \end{cases} \quad 3)$$

The Bernoulli distribution is implemented in Mathematica as `BernoulliDistribution[p]` in the Mathematica add-on package `Statistics`DiscreteDistributions`` (which can be loaded with the command `<<Statistics``).



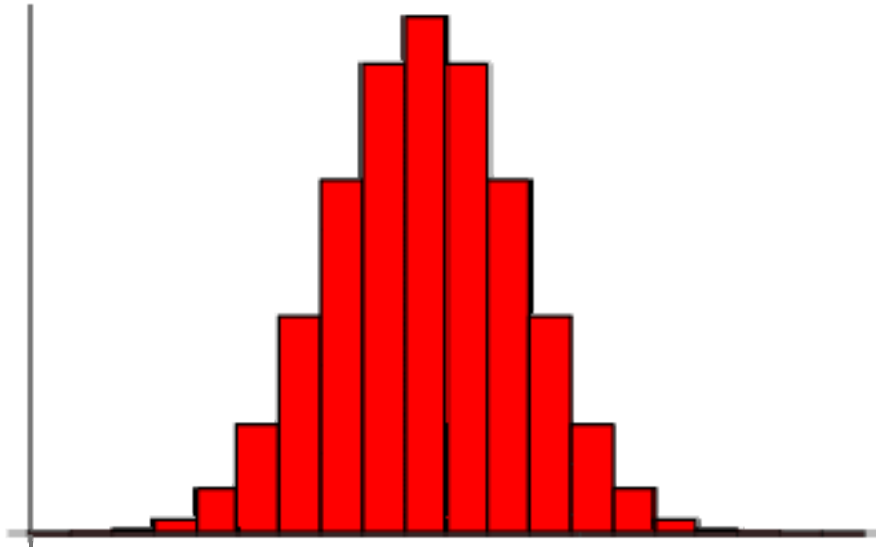
Discrete Distributions (Cont'd)

The performance of a fixed number of trials with fixed probability of success on each trial is known as a Bernoulli trial.

The distribution of heads and tails in coin tossing is an example of a Bernoulli distribution with $p=q=1/2$. The Bernoulli distribution is the simplest discrete distribution, and it the building block for other more complicated discrete distributions. The distributions of a number of variate types defined based on sequences of independent Bernoulli trials that are curtailed in some way are summarized in the following table (Evans et al. 2000, p. 32).

distribution	definition
binomial distribution	number of successes in n trials
geometric distribution	number of failures before the first success
negative binomial distribution	number of failures before the x -th success

Discrete Distributions (Cont'd)



The binomial distribution gives the discrete probability distribution of obtaining exactly successes out of Bernoulli trials (where the result of each Bernoulli trial is true with probability and false with probability). The binomial distribution is therefore given by



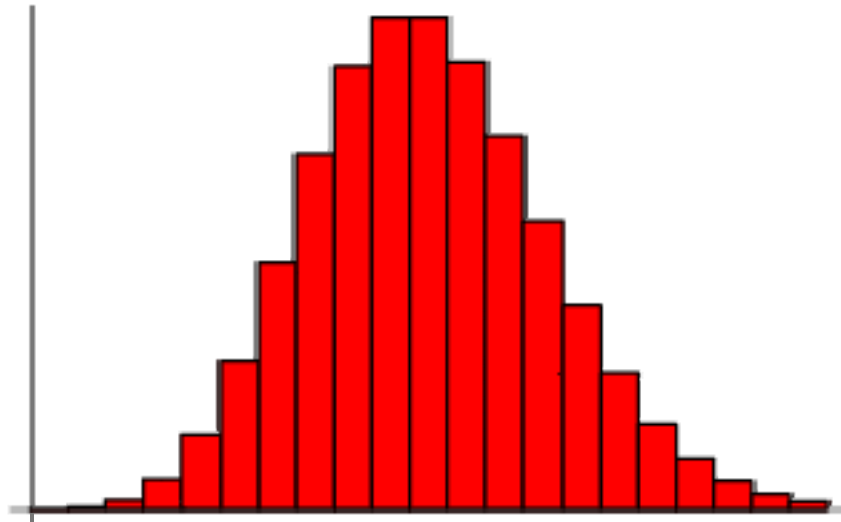
Discrete Distributions (Cont'd)

$$P_p (n | N) = \binom{N}{n} p^n q^{N-n} \quad 1)$$

$$= \frac{N!}{n! (N-n)!} p^n (1-p)^{N-n}, \quad 2)$$

where $\binom{N}{n}$ is a binomial coefficient. The above plot shows the distribution of n successes out of $N=20$ trials with $p=q=1/2$.

Discrete Distributions (Cont'd)



Given a Poisson process, the probability of obtaining exactly successes in trials is given by the limit of a binomial distribution

$$P_p (n | N) = \frac{N!}{n! (N - n)!} p^n (1 - p)^{N-n}. \quad 1)$$



Discrete Distributions (Cont'd)

Viewing the distribution as a function of the expected number of successes

$$v \equiv N p \quad 2)$$

instead of the sample size N for fixed p , equation (\diamond) then becomes

$$P_{v/N}(n | N) = \frac{N!}{n! (N - n)!} \left(\frac{v}{N}\right)^n \left(1 - \frac{v}{N}\right)^{N-n}, \quad 3)$$

Letting the sample size become large, the distribution then approaches

$$\begin{aligned} P_v(n) &= \lim_{N \rightarrow \infty} P_{v/N}(n) \\ &= \frac{v^n e^{-v}}{n!}, \end{aligned}$$



Discrete Distributions (Cont'd)

which is known as the Poisson distribution (Papoulis 1984, pp. 101 and 554; Pfeiffer and Schum 1973, p. 200). Note that the sample size N has completely dropped out of the probability function, which has the same functional form for all values of ν



Significance Tests

The probability that a statistical test will be positive for a true statistic is sometimes called the test's sensitivity, and the probability that a test will be negative for a negative statistic is sometimes called the specificity. The following table summarizes the names given to the various combinations of the actual state of affairs and observed test results.

result	name
true positive result	sensitivity
false negative result	1- sensitivity
true negative result	specificity
false negative result	1- specificity



Significance Tests (Cont'd)

Multiple-comparison corrections to statistical tests are used when several statistical tests are being performed simultaneously. For example, let's suppose you were measuring leg length in eight different lizard species and wanted to see whether the means of any pair were different. Now, there are $8!/2!6! = 28$ pairwise comparisons possible, so even if all of the population means are equal, it's quite likely that at least one pair of sample means would differ significantly at the 5% level. An alpha value of 0.05 is therefore appropriate for each individual comparison, but not for the set of all comparisons.

In order to avoid a lot of spurious positives, the alpha value therefore needs to be lowered to account for the number of comparisons being performed. This is a correction for multiple comparisons. There are many different ways to do this. The simplest, and the most conservative, is the Bonferroni correction. In practice, more people are more willing to accept false positives (false rejection of null hypothesis) than false negatives (false acceptance of null hypothesis), so less conservative comparisons are usually used.



Significance Tests (Cont'd)

Once sample data has been gathered through an observational study or experiment, statistical inference allows analysts to assess evidence in favor or some claim about the population from which the sample has been drawn. The methods of inference used to support or reject claims based on sample data are known as tests of significance.

Every test of significance begins with a null hypothesis H_0 . H_0 represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug. We would write H_0 : there is no difference between the two drugs on average.

The alternative hypothesis, H_a , is a statement of what a statistical hypothesis test is set up to establish. For example, in a clinical trial of a new drug, the alternative hypothesis might be that the new drug has a different effect, on average, compared to that of the current drug. We would write H_a : the two drugs have different effects, on average. The alternative hypothesis might also be that the new drug is better, on average, than the current drug. In this case we would write H_a : the new drug is better than the current drug, on average.



Significance Tests (Cont'd)

The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either "reject H_0 in favor of H_a " or "do not reject H_0 "; we never conclude "reject H_a ", or even "accept H_a ".

If we conclude "do not reject H_0 ", this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against H_0 in favor of H_a ; rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.

Hypotheses are always stated in terms of population parameter, such as the mean μ . An alternative hypothesis may be one-sided or two-sided. A one-sided hypothesis claims that a parameter is either larger or smaller than the value given by the null hypothesis. A two-sided hypothesis claims that a parameter is simply not equal to the value given by the null hypothesis -- the direction does not matter.



Significance Tests (Cont'd)

Hypotheses for a one-sided test for a population mean take the following form:

$$H_0: \mu = k$$

$$H_a: \mu > k$$

or

$$H_0: \mu = k$$

$$H_a: \mu < k.$$

Hypotheses for a two-sided test for a population mean take the following form:

$$H_0: \mu = k$$

$$H_a: \mu \neq k.$$

A confidence interval gives an estimated range of values which is likely to include an unknown population parameter, the estimated range being calculated from a given set of sample data. (Definition taken from Valerie J. Easton and John H. McColl's Statistics Glossary v1.1)



Example

Suppose a test has been given to all high school students in a certain state. The mean test score for the entire state is 70, with standard deviation equal to 10. Members of the school board suspect that female students have a higher mean score on the test than male students, because the mean score \bar{x} from a random sample of 64 female students is equal to 73. Does this provide strong evidence that the overall mean for female students is higher?

The null hypothesis H_0 claims that there is no difference between the mean score for female students and the mean for the entire population, so that $\mu = 70$. The alternative hypothesis claims that the mean for female students is higher than the entire student population mean, so that $\mu > 70$.



Significance Tests for Unknown Mean and Unknown Standard Deviation

In most practical research, the standard deviation for the population of interest is not known. In this case, the standard deviation σ is replaced by the estimated standard deviation s , also known as the standard error. Since the standard error is an estimate for the true value of the standard deviation, the distribution of the sample mean \bar{x} is no longer normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Instead, the sample mean follows the t distribution with mean μ and standard deviation $\frac{s}{\sqrt{n}}$.

The t distribution is also described by its degrees of freedom. For a sample of size n , the t distribution will have $n-1$ degrees of freedom. The notation for a t distribution with k degrees of freedom is $t(k)$. As the sample size n increases, the t distribution becomes closer to the normal distribution, since the standard error approaches the true standard deviation σ for large n .



Significance Tests for Unknown Mean and Unknown Standard Deviation (Cont'd)

For claims about a population mean from a population with a normal distribution or for any sample with large sample size n (for which the sample mean will follow a normal distribution by the Central Limit Theorem) with unknown standard deviation, the appropriate significance test is known as the t -test, where the test statistic is defined as

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

The test statistic follows the t distribution with $n-1$ degrees of freedom. The test statistic z is used to compute the P -value for the t distribution, the probability that a value at least as extreme as the test statistic would be observed under the null hypothesis.



Example

The dataset "Normal Body Temperature, Gender, and Heart Rate" contains 130 observations of body temperature, along with the gender of each individual and his or her heart rate. Given the table below, please test the one-sided hypothesis that the normal body temperature is statistically equal to 98.6.

Descriptive Statistics

Variable	N	Mean	Median	Tr Mean	StDev	SE Mean
TEMP	130	98.249	98.300	98.253	0.733	0.064

Variable	Min	Max	Q1	Q3
TEMP	96.300	100.800	97.800	98.700



Answer:

Since the normal body temperature is generally assumed to be 98.6 degrees Fahrenheit, one can use the data to test the following one-sided hypothesis:

$$H_0: \mu = 98.6 \text{ vs}$$

$$H_a: \mu < 98.6.$$

The t test statistic is equal to $(98.249 - 98.6)/0.064 = -0.351/0.064 = -5.48$. $P(t < -5.48) = P(t > 5.48)$. The t distribution with 129 degrees of freedom may be approximated by the t distribution with 100 degrees of freedom (found in Table E in Moore and McCabe), where $P(t > 5.48)$ is less than 0.0005. This result is significant at the 0.01 level and beyond, indicating that the null hypotheses can be rejected with confidence.



Concepts of Estimation Sampling



To Develop Statistical Sampling Methodology:

- ❑ Clearly state the objective;
- ❑ Define the population to be sampled;
- ❑ Choose an appropriate sample design;
- ❑ Determine the sample size;
- ❑ Extrapolate;
- ❑ Present the results.



While Stating the Objectives of the Sample:

- ❑ Specify the information to be gathered
 - *What information will be collected for each sample item?*
 - *Examples: Claim number, paid amount, error amount etc.*
- ❑ Determine if sampling is appropriate
 - *Is the function so important that no sampling uncertainty can be tolerated?*



While Stating the Objectives of the Sample (cont'd):

- ❑ Establish what is to be inferred from the sample
 - *What will the outcome of the sample be?*
 - *Examples: Accuracy rate, total overpayment amount etc.*
- ❑ Decide how much uncertainty is tolerable
 - *Confidence*
 - *Precision*



Overview

- ❑ Precision and Confidence

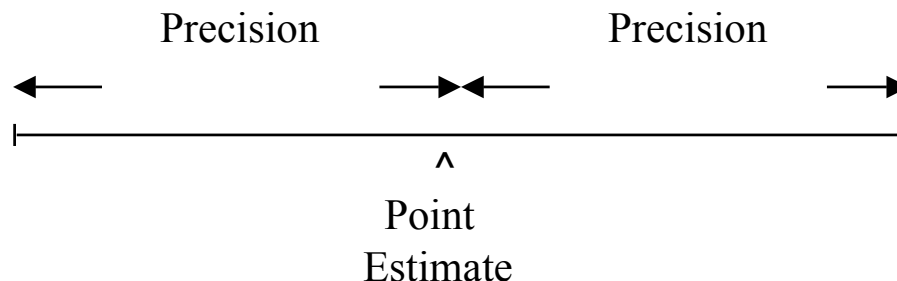
- ❑ Population Distribution

- ❑ Other concepts



Precision and Confidence

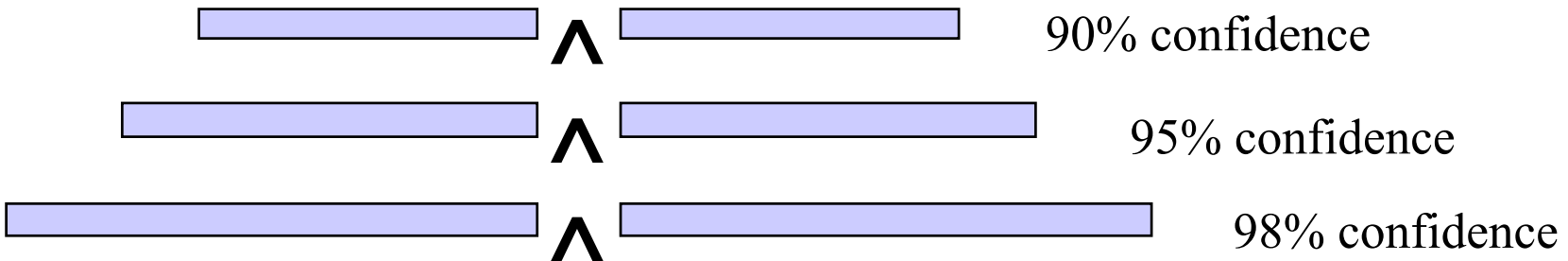
- ❑ Statistical samples are evaluated in terms of precision and confidence level, which includes a range of values (plus and/or minus) around the point estimate of the population.
- ❑ The confidence level is the range and the precision is a measure of this range around the point estimate (mid-point of range).





Precision and Confidence (cont'd)

- ❑ Precision and confidence level are statistically inseparable in that neither should be expressed without the other. (Precision is a function of confidence.)
- ❑ For example, for a given sample:





DISCUSSION: Precision and Confidence

- ❑ Which is a wider confidence interval: 95% or 99%?

- ❑ When you construct a 95% confidence interval, what are you 95% confident about?

- ❑ What is the effect of sample size on the width of a confidence interval?



Defining the Population:

- ❑ Establish the sampling unit:
 - *What should the sampling unit be?*
 - *Some examples include account, invoice, claim and patient.*
- ❑ Determine relevant characteristics as they pertain to the objective:
 - *Period of time*
 - *Geography*
 - *Market*
 - *Payer*
 - *Inpatient/ Outpatient*



Defining the Population (cont'd):

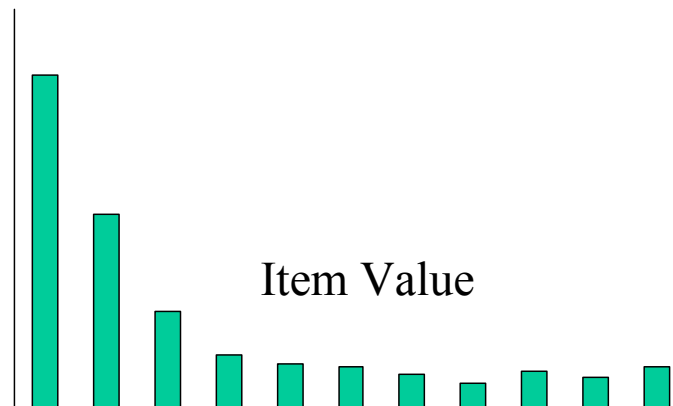
- Determine how the population will be available:
 - *Is the population contained in paper file?*
 - *Will the entire population be available prior to selecting the sample?*
 - *What data elements will be available for the population?*
 - *Some examples include account, invoice, claim and patient.*
- Determine the sampling frame:
 - *The sample frame is the list of all elements in the population*



Population Distribution

- Value
- Normal Distribution
- Point Estimate

Frequency

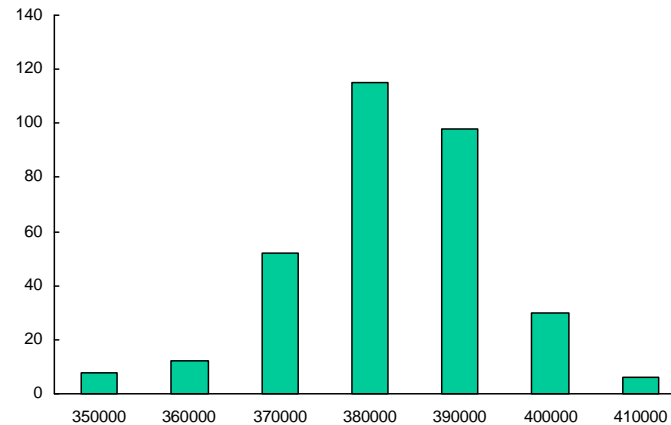




Population Distribution (cont'd)

- ❑ 300 Estimates
- ❑ 300 samples of 20 items

Frequency



Point Estimate Value



Population Distribution (cont'd)

Normal Probability Curve:

- ❑ True Population Value at Mid-Point

- ❑ 68% of Estimates within 1 Standard Error

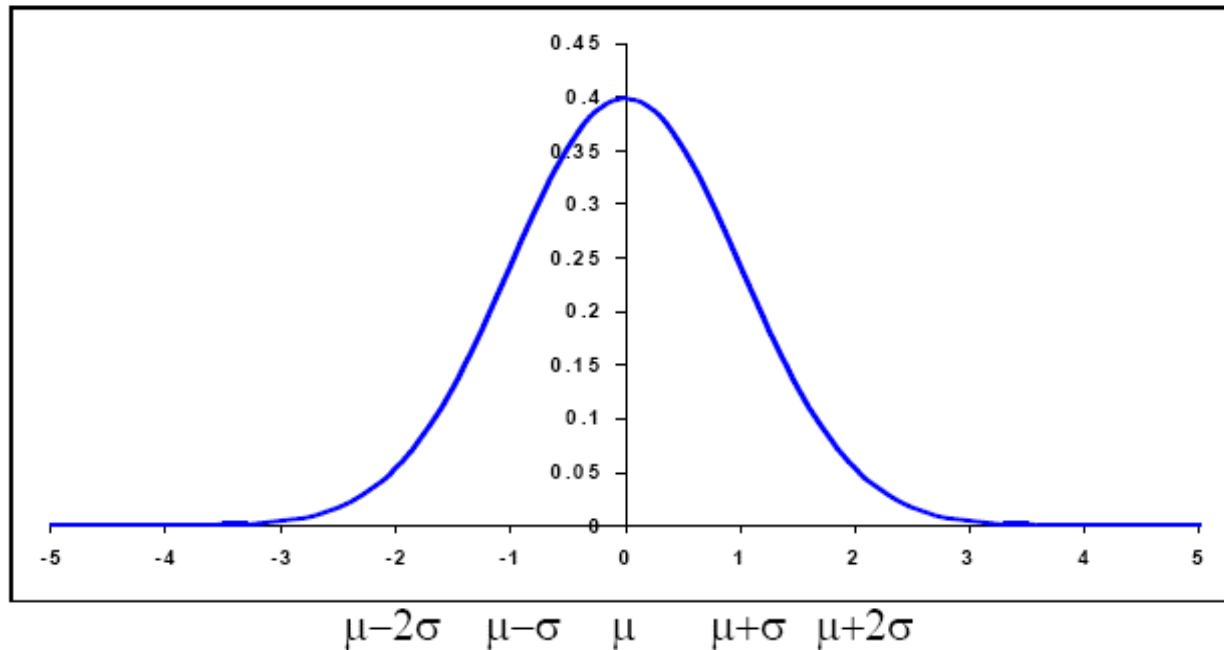
- ❑ 95% of Estimates within 1.96 Standard Errors

- ❑ Any one estimate has a 95% chance of being within 1.96 standard errors of the true value



Population Distribution (cont'd)

Normal Probability Curve

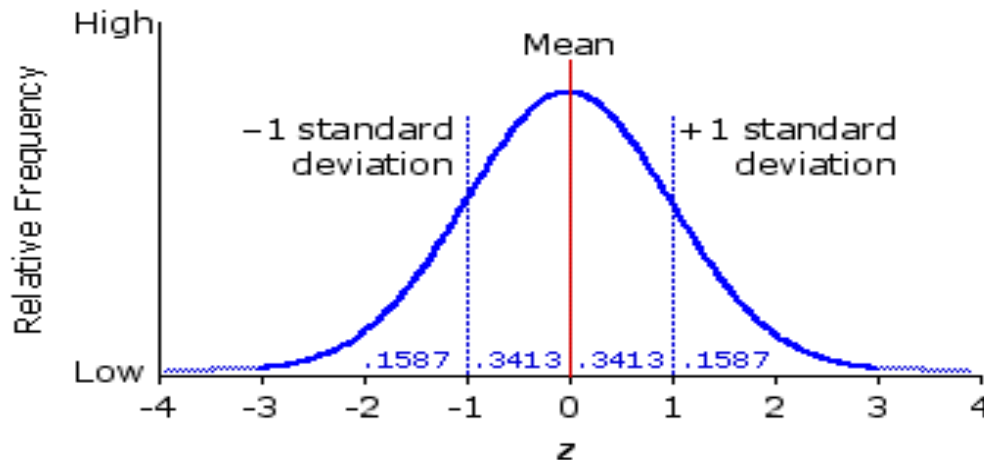




Population Distribution (cont'd)

Normal Probability Curve

- *In a normal distribution, 68% of the point estimates are within one standard deviation from the mean, 95% of the estimates are within two standard deviations from the mean and 99.7% of point estimates are within three standard deviations.*





DISCUSSION: Population Distribution

- ❑ An observation is .50 standard deviations below the mean on a normally distributed variable. What proportion of the data fall below that observation? Above it?

- ❑ True or False: As the sample size increases, the standard error of the sampling distribution of “Y bar” increases.

- ❑ If numerous samples of sample size $N = 15$ are taken from a uniform distribution and a relative frequency distribution of the means drawn, what would the shape of the frequency distribution look like?

- ❑ How does the t distribution compare with the normal distribution? How does the difference affect the size of confidence intervals constructed using z relative to those constructed using t? Does sample size make a difference?



Other Concepts

Central Limit Theorem

- *This theorem states that for random sampling, as sample size n grows, the sampling distribution of the mean approaches a normal distribution.*

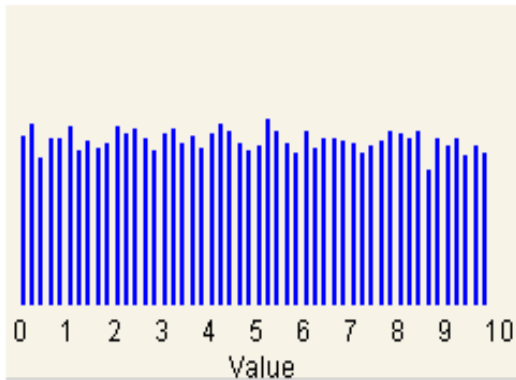


Fig. 1) Histogram of Population - Uniform Distribution: population = 10,000; mean = 5.013; std dev 2.897

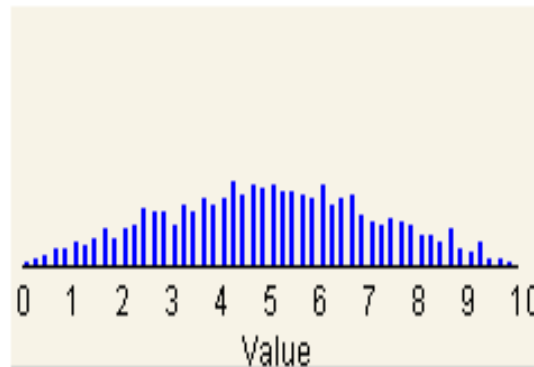


Fig. 2) Sampling Distribution $n = 2$: number of samples = 2010; mean = 4.995; std dev 2.011; std error = 2.048

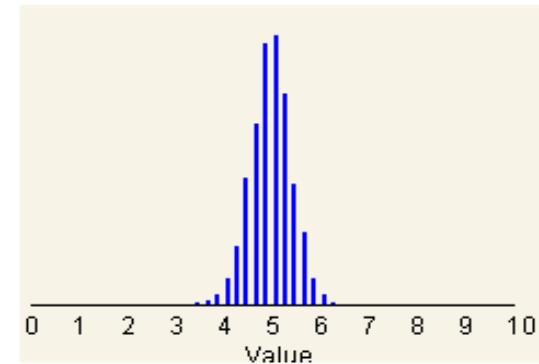


Fig. 4) Sampling Distribution $n = 50$: number of samples = 2010; mean = 4.999; std dev 0.411; std error = 0.410



Other Concepts (cont'd)

Central Limit Theorem (cont'd)

- For large enough sample size, repeated sample estimates will be normally distributed.
 - *If this is true, we can use our formulas to obtain precision level at 95% confidence.*
 - *Since our formulas don't work well when the design is "bad," and we cannot always give perfect rules for what makes a design "good" or "bad," the best bet is to use judgment and try to abide by guidelines.*



Other Concepts (cont'd)

Skewness

- ❑ Many accounting populations are highly skewed, with relatively small number of population items representing a significant portion of the total value. If the sample design does not consider this characteristic, selected samples may not be consistent with the application objectives, and the confidence interval may be statistically inefficient.
- ❑ To address the skewness of the original distribution, use:
 - *stratification by value design (e.g. the actual distribution design option in the ES Program)*
 - *Probability Proportional-to-Size (PPS) selection techniques*



Other Concepts (cont'd)

T Distribution

- ❑ The t-distribution is a probability distribution that arises in the problem of estimating the mean of a normally distributed population when the sample size is small. It is the basis of the t-tests for the statistical significance of the difference between two sample means, and for confidence intervals for the difference between two population means.
- ❑ It has relatively more scores in its tails than does the normal distribution, and its shape depends on the degrees of freedom. With 100 or more degrees of freedom, the t distribution is almost indistinguishable from the normal distribution.



Other Concepts (cont'd)

T Distribution (cont'd)

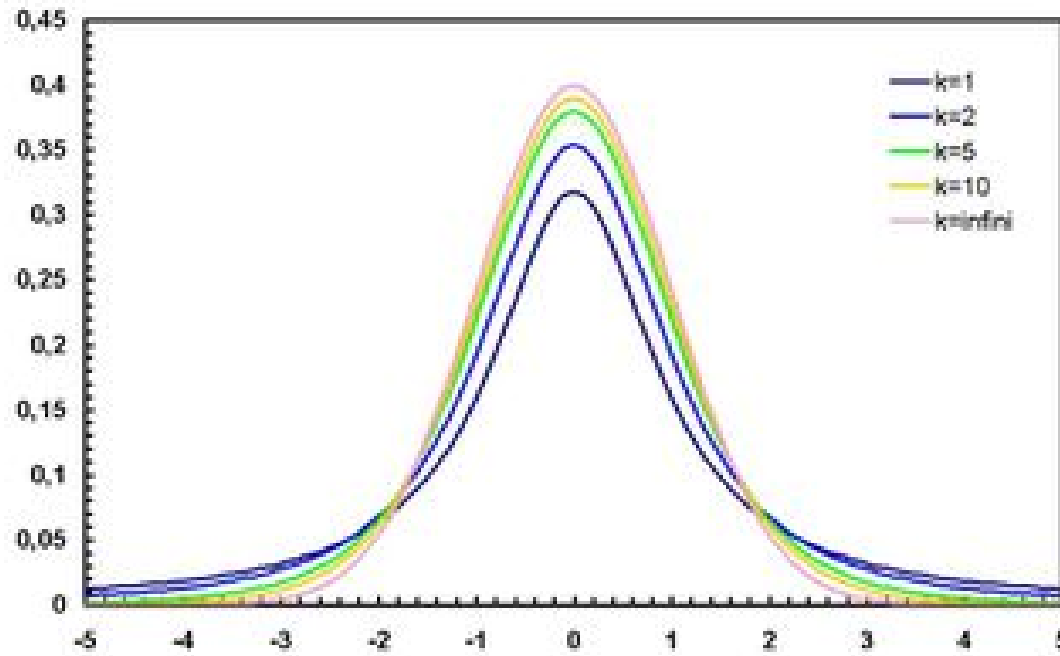
- ❑ As the degrees of freedom increases, the t distribution approaches the normal distribution.
- ❑ The larger the tails of the distribution, the farther out you have to go from the mean in order to contain a given percentage of the scores. For example, to contain 95% of the t distribution with 4 degrees of freedom, the interval must extend 2.78 estimated standard deviations from the mean in both directions. Compare this to the normal distribution for which the interval need only extend 1.96 standard deviations in both directions.



Other Concepts (cont'd)

T Distribution (cont'd)

Probability density function





Designing and Selecting
a Sample



Overview

- ❑ Stratifying a Sample

- ❑ Designing a Sample

- ❑ Selecting a Sample

- ❑ Tools



Designing the Sample:

- ❑ There are many different types of sample designs. There is no “correct” way to design a sample. Some designs may be more efficient (smaller sample size). In addition, the design may be dictated by characteristics of the population (such as multiple locations).
- ❑ In this industry, we most often see simple random samples and stratified random samples.
- ❑ Simple Random Samples:
 - *A simple random sample is the most basic sample design*
 - *Each item in the population has an equal probability of being selected*



Designing the Sample (cont'd):

❑ Stratified Random Samples

- *Stratification involves grouping items together into strata and then selecting a random sample within each stratum.*
 - *Methods for determining stratum breaks include professional judgment and cumulative square root*

❑ Stratification can be used to decrease sample size

❑ Stratification can be used when population items are naturally grouped. For example, population items can be naturally grouped by dollar amount.

❑ Stratification may be used if insight is required for a sub-group of the population



Stratification

- ❑ *When sub-populations vary considerably from each other, it is advantageous to sample each subpopulation (stratum) independently.*
- ❑ *The strata should be mutually exclusive: every element in the population must be assigned to only one stratum.*
- ❑ *The strata should also be collectively exhaustive: no population element can be excluded.*
- ❑ *Once each stratum has been identified, random sampling is then applied within each stratum to capture the number of population items. This often improves the representativity of the sample by reducing sampling error.*



Stratification (cont'd)

- Stratification is intended to group “like” items together (i.e., items expected to demonstrate similar behavior with respect to the sampling objectives)
 - *The standard error is reduced when “like” items are grouped together, and narrower precision ranges are produced as a result.*

- Stratification occurs by item value (low dollar items may be expected to have different error rates than high dollar items)
 - *Other characteristics (which may be related to expected error rate) may be location, category, financial classification, or product type for example.*

- Concept of efficiency: ability to reduce overall sample size and achieve similar precision results
 - *For example, a sample with 4 strata and 30 items in each may achieve similar precision as an unstratified sample of 400 items.*



Stratification (cont'd)

- ❑ Stratified random samples are particularly useful in skewed distributions.
- ❑ A stratified random sample will typically require a smaller sample size to meet specified precision and confidence levels than a simple random sample
- ❑ Theoretically, the type of sample design employed does not affect the sample results. A simple random sample should give the same results as a stratified random sample. However, it is important that the sample design be considered when determining sample results.



Stratification (cont'd)

□ Setting up strata

- *Setting the number of strata is always a tradeoff*
 - *If there are too many strata, there is not sufficient coverage within each stratum and the design can be exceedingly complex.*
 - *If there are too few strata, you may not be grouping “like” items together.*
- *Setting up top strata*
 - *No extrapolation impact*
 - *Ideal for very high dollar amounts and high-risk populations*

Efficiency Tip: Ideally, effort expended on top stratum should not exceed the coverage of the population dollars that is achieved.



Stratification (cont'd)

Applications of Stratified Sampling:

- ❑ Consider using a sample for a US political survey. Suppose we want the respondents to reflect the diversity of the population of the United States. Accordingly, the researcher specifically seeks to include participants of various minority groups such as race or religion, based on their proportionality to the total population as mentioned above. A stratified survey would thus be more representative of the US population than a survey of random estimation sampling.



DISCUSSION: Stratification

- ❑ Why would we use stratification on a sample design?

- ❑ What is the ideal number of strata to use?

- ❑ What is the recommended minimum sample size per strata? For the sample overall?



DISCUSSION: Stratification (cont'd)

- From previous experience, a CPA is aware of the fact that cash disbursements contain a few unusually large disbursements. In using statistical sampling, the CPA's best course of action is to:
 - *Eliminate any unusually large disbursements that appear in the sample.*
 - *Continue to draw new samples until no unusually large disbursements appear in the sample.*
 - *Stratify the cash disbursements population so that the large disbursements are reviewed separately.*
 - *Increase the sample size to lessen the effect of the unusually large disbursements.*



Sample Design Alternatives:

□ Probability-Proportional-to-Size (PPS) Sampling:

- *Selection technique that takes into account the population value skewness.*
- *Every dollar has an equal probability of selection (it is a monetary interval selection technique in the audit approach).*
- *Selection interval is determined by the total population amount, MP and R.*
- *Alternative selection methods that approximate PPS (e.g. Two Strata (TS)).*

Note: PPS selection is used for CMA sampling, also known as Dollar Unit Sampling.

- *Sample items are selected according to size*
- *Allows large items to have a higher probability of selection*



Sample Design Alternatives (cont'd):

Cluster Sampling

- ❑ Involves selecting a collection of units or clusters
- ❑ Single stage-A cluster is randomly selected, then secondary units within the cluster are sampled
- ❑ Multi stage-A cluster is randomly selected, then a sample of secondary units within the cluster are sampled
 - *Example: A company operates many facilities. A cluster may be a facility, and individual claims may be sampled from each sampled facility.*



Designing a Sample

How do you determine what the “right” sample size is?

- ❑ Sample size determination always relies on judgment
 - *There is no one way of knowing if a sample size is sufficient to achieve confidence and precision goals until you receive the sample results back*
 - *Consider coverage of different kinds of population items*

- ❑ Trade-off between sample size and confidence level/precision
 - *Generally, the larger the sample size the narrower the precision range, however there is a point of “diminishing returns”*
 - *Often, the key driver in sample size determination is the number of items that the client wants to and/or can reasonably review, given time/budget restraints*



Designing a Sample

- Sample size depends on:
 - *Population size*
 - *Variance to be estimated*
 - *Desired precision and confidence*



Designing a Sample (cont'd):

Additional information that may be used, when available:

- Have a source field, and target field is expected to be very close
 - *Actual distribution functionality in ES Design works very well in this case*

- Have results from a previous sample
 - *If similar results are expected from the current population, the previous results can be used as a proxy for the expected standard deviation*



Designing a Sample (cont'd):

Additional information that may be used, when available (cont'd):

- ❑ Suggest performing a probe sample (sometimes referred to as a pilot sample)
 - *Review a relatively small sample and use the standard deviation from the results as a proxy for the standard deviation expected from a full sample*
- ❑ An assumption (best guess)



Designing a Sample (cont'd)

Guidance has been developed for specific applications:

- Guidance provides rules of thumb and cannot be “proved”
 - Design Guidance:
 - *Sampling strata should each contain a minimum of 20 sample items or more.*
 - *The total sample size determined for the sampling strata should not be less than 100 items unless the population is very small (e.g., less than 400 items)*
 - Evaluation Guidance:
 - *There should be at least 20 variances / errors overall in the sampling strata*
 - *There should be at least 5 variances / errors in each of the sampling strata, or at least in a majority of them*



Selecting a Sample

□ Random Sampling

- *Selection technique where every combination of the same number of sampling units in the population has an equal (or known) probability of selection.*
- *Most likely to result in a representative sample.*
- *Does not consider population value*
- *Common techniques include random number selection and interval (or systematic) selection*

- In order to apply the random selection method you must set up some process or procedure that assures that the different units in your population have equal probabilities of being chosen.



Selecting a Sample (cont'd)

Systematic Samples:

- ❑ Systematic samples are not preferred, but in practice, they typically result in a sample that is sufficiently random.
- ❑ Systematic samples are often used when it is impractical to assign a number to each item in a population (for example, paper files). The need for systematic samples has been decreasing with the amount of data that are captured electronically.



Selecting a Sample (cont'd)

Systematic Samples (cont'd):

- ❑ In order to select a systematic sample, a skip interval must be determined. The skip interval is calculated by dividing the population size by the sample size. Then a random number between 1 and the skip interval is generated. Starting with the randomly selected starting place, sample items are selected according to the skip interval.
- ❑ Example: Suppose we have a population of 100 and wish to choose a sample of 10. The skip interval is $100/10=10$. So, we generate a random number between 1 and 10. The random number serves as the first sampled item and then we select every 10th item.



DISCUSSION: Designing/Selecting a Sample

- An auditor plans to examine a sample of 20 checks for countersignatures as prescribed by the client's internal control procedures. One of the checks in the chosen sample of 20 cannot be found. The auditor should consider the reasons for this limitation and:
- *Evaluate the results as if the sample size had been 19.*
 - *Treat the missing check as a deviation when evaluating the sample.*
 - *Treat the missing check in the same manner as the majority of the other 19 checks (countersigned or not).*
 - *Choose another check to replace the missing check in the sample.*



DISCUSSION: Designing/Selecting a Sample (cont'd)

- An underlying feature of random-based selection of items is that each:
 - *Stratum of the accounting population be given equal representation in the sample.*
 - *Item in the accounting population be randomly ordered.*
 - *Item in the accounting population should have an opportunity to be selected.*
 - *Item must be systematically selected using replacement.*



Tools Used for Designing and Selecting a Sample

□ ES Design

- *Software internally developed by DQI to replace the existing WinES statistical software*
- *Based off of Microsoft Access, the ES Software was written in a manner to allow continuous improvements to be made by DQI*
- *New regulations (e.g. Sarbanes-Oxley) have resulted in certain restrictions related to use of this software at attest clients*

□ ACL

- *External vendor software with extended data analysis capabilities*
 - *Dollar stratification and sample selection capabilities can be used to facilitate sample design/selection process by educated users*
-



Choosing Extrapolation Methodology

- The choice of an extrapolation methodology depends on:
 - *What type of sampling has been performed (attribute or variable)*
 - *What type of result is required (ratio, total amount etc.)*
- When there is a choice of extrapolation methodologies, one method may provide a more precise estimate than another



Choosing Extrapolation Methodology (cont'd)

- ❑ It is important that samples be evaluated to determine achieved precision. Even though a sample is designed to meet specific precision and confidence requirements, there is no guarantee that those criteria have been met.
 - For example, in variables sampling, if the assumptions used to develop the sample size were incorrect, then the achieved precision may not satisfy precision requirements.
- ❑ Evaluation of the sample results will allow the user to either enlarge the sample to meet precision and confidence requirements and/or re-design the process on a go-forward basis.



Evaluating **a Sample**



Overview

- ❑ Source and Target

- ❑ Statistical Estimators

- ❑ Evaluation Guidance

- ❑ Tools



Source and Target

- Source: known for the entire population
 - *In formulas, source is notated as “x” or “X”*

- Target: known only for the sample; estimated for the entire population
 - *In formulas, target is notated as “y” or “Y”*

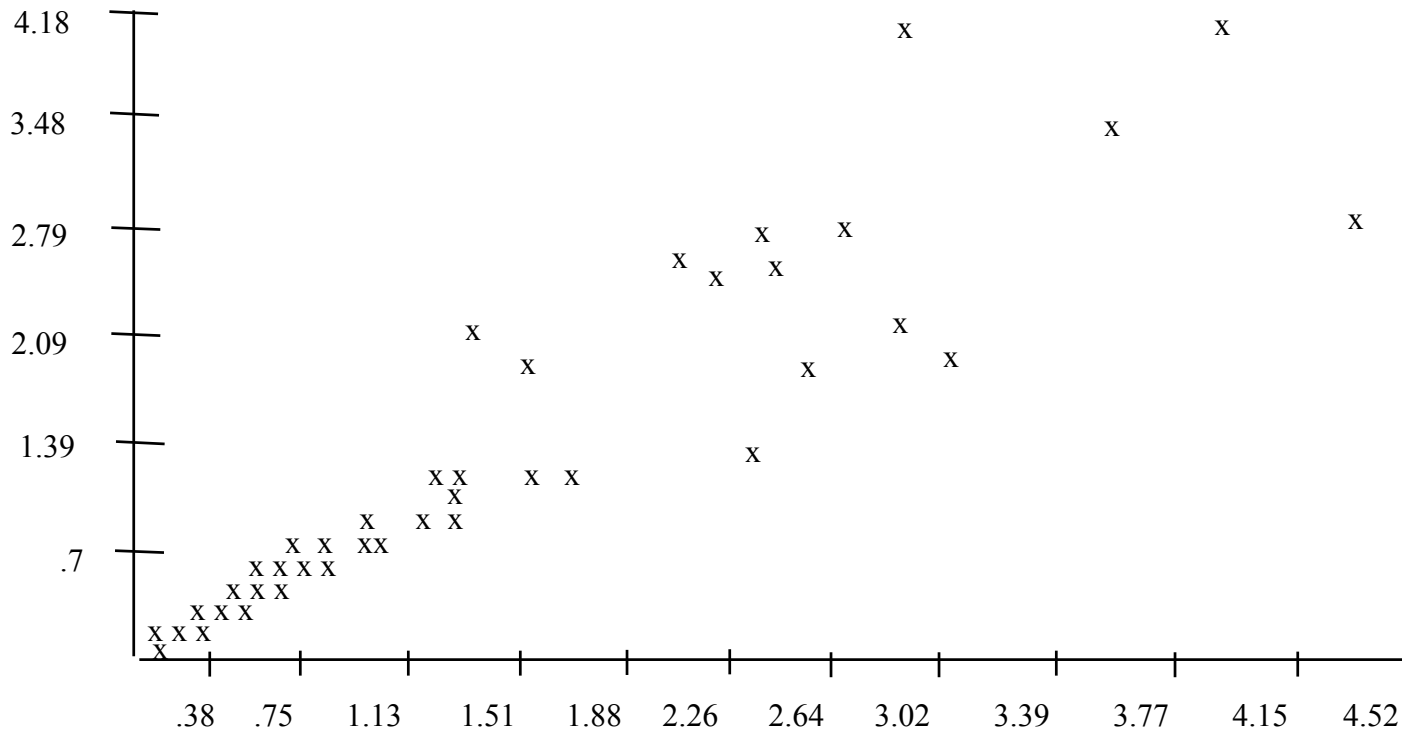
- Other common notations:
 - *N = number of items in population*
 - *n = number of items in sample*



Source and Target (cont'd)

Plot of source vs. target values: What line is the “best fit”?

Scatter diagram of 80-item sample





DISCUSSION: Source and Target

□ **What is the source and what is the target?**

- *Sampling inventory to determine the total value of inventory*
- *Sampling claims to determine the total dollar value of reimbursements that should not have been made*
- *Sampling business meal expenses to determine the dollar amount that should be reclassified*



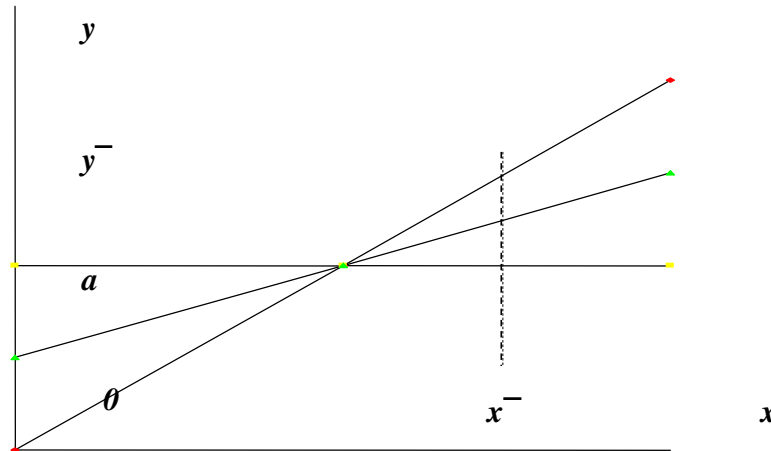
Statistical Estimators

Estimator	Approach to Determine Point Estimate of Total Target Dollar Amount in Population	Selection Technique
Mean-per-unit (MPU)	Computes average target value in sample and multiplies by total number of items in population	Random
Difference	Computes average difference between source and target values in the sample and multiplies by total number of items in population (same technique as MPU but applied to the difference as opposed to the target). The result is added to total population dollars to determine point estimate of total target dollar amount in population.	Random
Ratio	Computes ratio of average target value to average source value in sample and multiplies by total population dollars	Random
Regression	Performs simple linear regression model to estimate target value based on source value ($y\text{-hat} = a + bx$)	Random
PPS Variables	Computes average of the proportions of target value to source value for each item in the sample and multiples by total population dollars	PPS



Statistical Estimators (cont'd)

Mean Per Unit vs. Ratio vs. Regression lines



Ratio Line $\hat{y} = bx$

Regression Line $\hat{y} = a + bx$

Mean-per-unit Line $\hat{y} = \bar{y}$



Statistical Estimators (cont'd)

MPU Evaluation

□ Point Estimate = $\left(\frac{\sum_1 y_1}{n} \right) N = \bar{y}N$

□ Precision = $t * \text{Standard Error (SE)}$
(t depends on confidence level)

□ Standard Error is primarily impacted by the sum of $(y_1 - \bar{y})^2$

□ Standard Error = $N \sqrt{\frac{N-n}{N}} \frac{1}{\sqrt{n}} \text{StdDev}$

□ Standard Deviation = $\sqrt{\frac{\sum_1 (y_i - \bar{y})^2}{n-1}}$



Statistical Estimators (cont'd)

Difference Estimation

- Difference Estimate

$$\hat{Y} = N\bar{d} + X$$

N = Total number of population items

X = Population total

- Use difference d_i between recorded and actual values (i.e. source and target) $\bar{d} = \frac{\sum d_i}{n}$
- Calculate the average difference between sample item recorded and actual values (sum sample item differences, and divide by sample size)
- Precision = $t * SE$
- Standard error is primarily affected by the sum of $(d_i - \bar{d})^2$
- More precise than MPU, ratio or regression estimate if the size of the differences is independent of the recorded values



Statistical Estimators (cont'd)

Ratio Evaluation

□ Use relationship between recorded and actual values (i.e. source and target)

□ Calculate line of best fit through the origin

□
$$\text{Ratio} = \frac{\text{Actual}}{\text{Recorded}} = \frac{\sum y}{\sum x} = \frac{\bar{y}}{\bar{x}}$$

□ Point Estimate = $\left(\frac{\bar{y}}{\bar{x}}\right) * X$ (X = Total population dollar value)

□ Precision = $t * SE$

□ Standard Error (SE) is primarily affected by $(y_i - rx_i)^2$

□ Tends to be a more precise estimate than MPU or difference estimation if there is a reasonable correlation between the size of the recorded and actual values



Statistical Estimators (cont'd)

Regression Evaluation

- ❑ Use relationship between recorded and actual values (i.e. source and target)
- ❑ Method of least squares used to calculate a line that best fits the sample (not usually through origin)
- ❑ Regression line = $a + bx_i$
- ❑ Regression line passes through (\bar{x}, \bar{y})
- ❑ Point Estimate = $N[\bar{y} + b(\bar{X} - \bar{x})]$
- ❑ Precision = $t * SE$
- ❑ Standard Error is primarily affected by $(y_i - \bar{y})^2$ and $(x_i - \bar{x})^2$



Statistical Estimators (cont'd)

When stratification has been used:

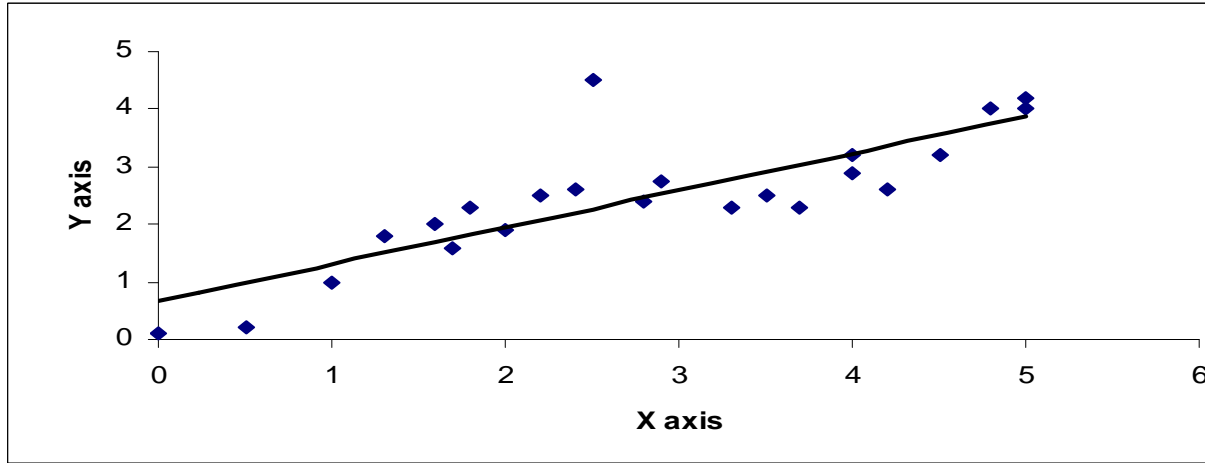
Formulas described on previous pages are used on each stratum individually

- ❑ Overall Point Estimate = sum of Strata Point Estimates
- ❑ Overall Standard Error = square the standard error for each strata; sum; take square root
- ❑ For ratio and regression estimators, a “Combined” approach is also available in addition to the “Separate” approach. For example, consider the Ratio estimator:
 - *In the “Separate” approach, a separate ratio is computed and applied for each stratum independent of the other strata.*
 - *In the “Combined” approach, one overall ratio is computed by combining the weighted average stratum mean values of source and target. This same ratio is then applied to each individual stratum.*

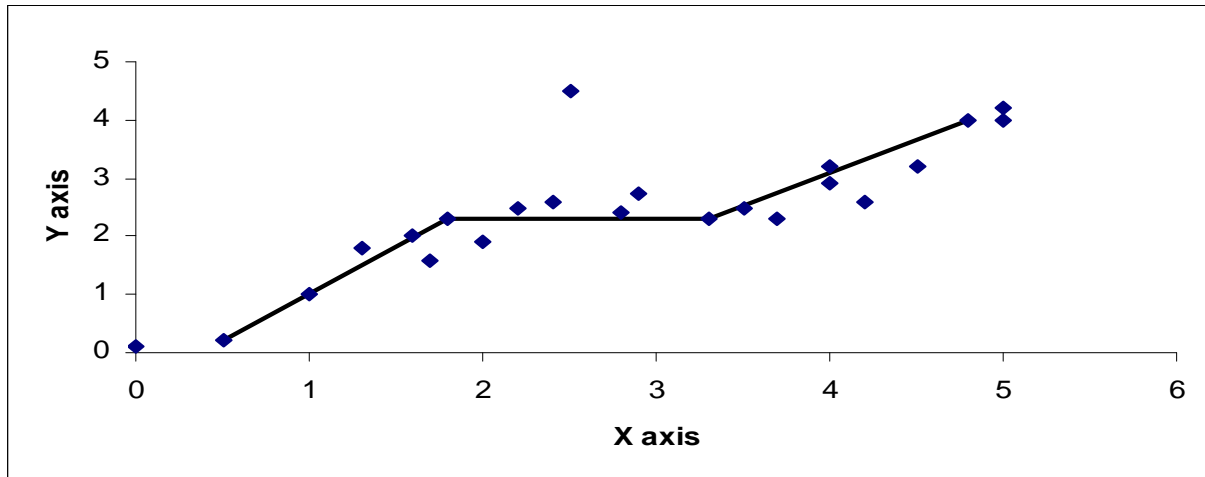


Statistical Estimators (cont'd)

Combined



Separate





Statistical Estimators (cont'd)

PPS Variables

- ❑ PPS Variables Estimate

$$y^{\epsilon} = \bar{p} X \quad X = \text{Population recorded total value}$$

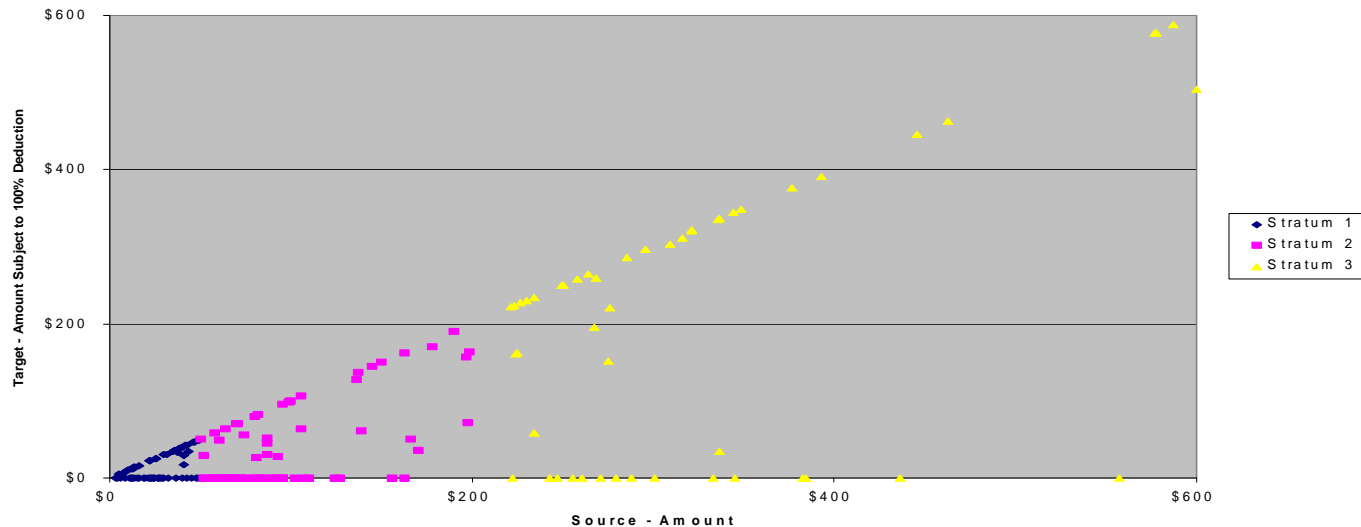
- ❑ Use proportion p_i for each sample item of actual to recorded values $\left(\frac{y_i}{x_i}\right)$
- ❑ Calculate the average proportion between sample item actual and recorded values (sum sample item proportions and divide by sample size) $\bar{p} = \frac{\sum_1 p_i}{n}$
- ❑ Precision = $t * SE$
- ❑ Standard error is primarily affected by the sum of $(p_i - \bar{p})^2$
- ❑ The only estimation method for a PPS selection



Statistical Estimators (cont'd)

- ❑ How do you choose the “right” estimator?
 - *When both source and target values are available, plotting source vs. target can provide insights into relationship/correlation*
- ❑ Which estimators would be expected to work for this sample?

Sample Results Correlation Analysis





Evaluation Guidance

- ❑ Review design & selection (especially if you were not involved)
- ❑ Perform qualitative evaluation
- ❑ Perform quantitative evaluation
 - *Number of errors per stratum*
 - *Look at standard errors by stratum*
 - *Look at relationship of source to target by stratum (plotting source vs. target)*
 - *Are there any items with a significantly large difference between source and target?*
- ❑ Look at different estimators
 - *Are any point estimates “out of bounds”?*
 - *Do any have significantly lower standard errors?*



Appendix A:

Design Quality Assurance Checklist



SAMPLING POPULATION DEFINITION

Check that the following factors regarding the sampling population and the representation of that population used for selection were properly considered:

- ❑ The optimal number of locations and/or subpopulations to be included in one application
- ❑ Whether there is more than one type of sampling unit for which separate applications are preferred or required (e.g., raw materials inventory parts versus finished goods)
- ❑ Whether there are items to be excluded from the sampling population (e.g., on-consignment inventory)
- ❑ Whether 100% verification of a portion of the population is required (e.g., high value items, theft sensitive inventory parts)
- ❑ Completeness of the population representation from which selections are to be made



SAMPLE DESIGN PARAMETERS

NUMBER OF STRATA:

In general, use 3 to 10 sampling strata (plus top and bottom strata). If large variances are expected, use fewer sampling strata (and larger sample sizes per stratum).

For small populations of less than 400 items, use 1 or 2 strata. For large populations, (e.g., more than 20,000 items and exceeding 25 million dollars in total), up to 12 strata may be efficient. More than 12 strata is rarely appropriate unless the population exceeds 100,000 items.

For designs that are mathematically determined based on the population item value distribution (e.g., using the ES software), see the table below for guidance. For designs that are *not* mathematically determined based on the population item value distribution, fewer sampling strata are more practical (typically, 3 to 5 strata).

The maximum number of strata typically recommended for sample design based on the population item value distribution (e.g., using ES software):

<u>\$ Population value</u>	<u>< 10,000 items</u>	<u>> 10,000 items</u>
< 5,000,000	6	8
5,000,000 - 25,000,000	8	10
> 25,000,000	10	12



MINIMUM SAMPLE SIZE: The minimum sample size for a stratum should be set to 20, unless many variances are expected when it might be set to 30.

TOP CUTOFF:

Typically the top stratum should include:

- Not more than 200 items
- Approximately 10 to 30 percent of the population value, provided the population distribution is such that this is efficient.

To check efficiency:

- Compute ratio of number of top stratum items to total sample size (1)
- Compute ratio of top stratum total value to population total value (2)

If $(1) > (2)$, consider increasing the top stratum cutoff

If $(1) < (2)$, consider lowering the top stratum cutoff (unless top stratum includes over 75% of population value)

For example, suppose a sample of 300 items is selected from a population of 10,000 items and total value \$4 million. If the top stratum includes 90 items with total value \$1 million: $(1) = 30\%$, $(2) = 25\%$, $(1) > (2)$ and therefore the top stratum cutoff should be increased to include fewer items in the top stratum.



BOTTOM CUTOFF:

If items with zero quantity are to be sampled, as opposed to checking them all, a cutoff of -1 should be used, to include any items with negative values in the bottom stratum but not zeros. If all zero quantity items are to be checked, use a cutoff of 0, to include items with zero and negative values in the bottom stratum.

DESIGN PRECISION:

A significantly lower precision than the evaluation precision requirement of P% should be used for *design precision*.

For initial designs that are mathematically determined based on the population item value distribution (e.g., using ES software), a design monetary precision of approximately one half of evaluation precision is generally recommended. Designs following the first application may use a lower or higher design precision, depending on the results achieved and the expectation that similar results will be obtained in the future.

CONFIDENCE LEVEL:

Applications for financial purposes should generally use a statistical confidence level of 95%



SAMPLE SIZES

OVERALL SAMPLE SIZE:

Sample sizes for financial purposes, for populations that are relatively well controlled, are typically in the range 300 to 750. Check the sample size for the following:

- The total sample size determined for the sampling strata (i.e., all strata except the reject and top strata) should not be less than 100 items unless the population is very small (e.g., less than 400 items).

For small populations with fewer than 400 items, sample size should ordinarily not be more than 25% of the total items in the sampling population but should not be less than 50. (For populations with less than 200 items, estimation sampling may not be practical - consult.)



INDIVIDUAL STRATUM SAMPLE SIZES:

The sample size determined for each stratum should be checked as follows:

- Sampling strata should each contain a minimum of 20 sample items or more. If many variances/errors are expected, a larger stratum sample size is preferable, say 30.
- Check the allocation of the population value and the sample size by strata for reasonableness. In general there should be a consistent relationship between the amount of population value in each strata and its allocated sample size (the exception may be strata 1 & 2, if there are many items of low value in the population)
- Check higher value sampling strata for sample sizes equal to or close to the population sizes in the strata. This may indicate that a lower top cutoff is preferable.



Appendix B:

Evaluation Quality Assurance Checklist



SAMPLING POPULATION DEFINITION

Check that the following factors regarding the sampling population and the representation of that population used for selection were properly considered:

- ❑ The optimal number of locations and/or subpopulations to be included in one application
- ❑ Whether there is more than one type of sampling unit for which separate applications are preferred or required (e.g., raw materials inventory parts versus finished goods)
- ❑ Whether there are items to be excluded from the sampling population (e.g., on-consignment inventory)
- ❑ Whether 100% verification of a portion of the population is required (e.g., theft sensitive inventory parts)
- ❑ Completeness of the population representation from which selections are to be made.



SAMPLE EXAMINATION

Check that the following conditions were complied with:

- ❑ All selected items were examined and a target value (e.g., audit value) established for each. Items may not be discarded unless the sampling objective is clearly inapplicable.
- ❑ All top and bottom stratum items selected were examined.
- ❑ Items with large variances / errors were rechecked.
- ❑ Variances / errors have not been excluded from the evaluation on judgmental grounds. By definition, a representative sample cannot include an isolated instance of error.

SAMPLE EVALUATION

The evaluation should be reviewed to determine if the evaluation precision at XX% confidence (usually 95%) is less than P% (and less than \$ pp million), and if the evaluation is reasonable and statistically valid.

- ❑ If the precision exceeds P% (or \$ pp million), the application is not acceptable.
- ❑ If the precision is less than P% (and \$ pp million), the application is likely to be acceptable, but should still be reviewed for reasonableness and statistical validity.



REASONABLENESS:

- Check the evaluation for the following:
- Does the difference between the recorded total population value and the sampling point estimate of that value exceed (2xP) percent (or (2xpp) million)? If so, is sampling appropriate?
- Does the difference between the recorded total population value and the sampling point estimate of that value make sense in light of your knowledge of the population and variances / errors found?
- Has the population subject to sampling been reconciled to the general ledger?
- Do the types and number of variances / errors found have any control implications?
- Have rechecks of large variances / errors been done at an appropriate level irrespective of the achieved precision?
- Have large variances / errors been investigated and explained?



STATISTICAL VALIDITY

Check the evaluation report for the following:

- ❑ Population sizes and values, both in total and by stratum, should agree with the population sizes and values reported as a result of the selection process (e.g., the ES selection report).
- ❑ Sample sizes, both in total and by stratum, should agree with the sample sizes reported as a result of the selection process (e.g., the ES selection report). Rechallenge the adequacy of the sample sizes in total and by stratum
 - The sample size from the sampling strata should exceed 100, unless the population size is less than 400, in which case a minimum of 50 applies.
 - The sample size in each sampling stratum should exceed a minimum of 20.
- ❑ If the t-distribution value is larger than 2.2 (for 95% confidence), reconsider the adequacy of the sample sizes and whether fewer strata should be used.
- ❑ If the estimation method is other than mean-per-unit, check whether there is a sufficient number of variances / errors in the sampling strata, in total and by stratum, to ensure that the estimate is reliable



STATISTICAL VALIDITY (CONT'D)

- There should be at least 20 variances / errors overall in the sampling strata, (i.e., not including variances in the top stratum).
- There should be at least 5 variances / errors in each of the sampling strata, or at least in a majority of them
- Check whether there are individual strata with standard errors larger or smaller than the norm for this application. Larger standard errors are generally caused by large variance(s) / error(s) in the stratum. Consider rechecks of such items. Smaller standard errors are often the result of few variances / errors. Reconsider whether the number of variances / errors in such strata is sufficient.

If there are too few variances / errors to be certain of the statistical validity (less than 20 overall or less than 5 per stratum in a majority of strata):

- Reevaluate the sample using stratified mean-per-unit,
- Check that the point estimate determined under the original method falls within the mean-per-unit confidence interval at XX% (e.g. 95%) confidence level
 - If it does, accept the estimate as statistically valid,
 - If it does not fall within the mean-per-unit confidence interval, consult.



Appendix C:

Mathematical Formulas



SYMBOLS USED

The population is partitioned into L strata, numbered $1, 2, \dots, L$. The strata contain N_1, N_2, \dots, N_L items respectively. The total size of the population is N . For an unstratified estimate, the calculations in this section should be considered as if for one stratum ($L=1, h=1$).

Because all population items in the top stratum and bottom stratum are examined, no statistical evaluation is performed on those strata. The top and bottom strata can therefore be disregarded for the purposes of this chapter, and it will be convenient to do so. Thus, we assume that the L strata do not include a top or a bottom stratum and the total size of the population is:

$$N = \sum_{h=1}^L N_h$$

Independent random samples are drawn from each stratum. Their sizes are n_1, n_2, \dots, n_L respectively. Throughout, we will use h to denote the stratum number and i to denote the item number within the stratum. Thus the i -th target value in stratum number h is denoted by y_{hi} .



SYMBOLS USED (CONT'D)

The following symbols and factors refer to stratum h :

N_h	Total number of items in the stratum population.
n_h	Number of items in the sample.
y_{hi}	Target value of i -th item.
x_{hi}	Source value of i -th item (when available).
X_h	Total source value of items in the stratum population.

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{n_h} \quad \text{Sample mean of target variable}$$



SYMBOLS USED (CONT'D)

$$s_{yh} = \sqrt{\frac{\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2}{n_h - 1}}$$

Sample standard deviation of target variable

$$\sum = \sum_{i=1}^{n_h}$$

Summation over sample items in stratum h



Mathematics of MPU Estimation

Stratum MPU Estimate and Standard Error

The MPU estimate for stratum h is:

$$\hat{Y}_h = N_h \bar{y}_h$$

The standard error of the estimate in stratum h is:

$$S(\hat{Y}_h) = N_h \frac{S_{yh}}{\sqrt{n_h}} \sqrt{1 - n_h/N_h}$$



Total Estimate, Standard Error and Precision

The estimates and standard errors for the L strata are now combined into a total estimate and standard error. The total estimate is:

$$\hat{Y} = \sum_{h=1}^L (\hat{Y}_h)$$

The total standard error is the square root of the sum of the squared stratum standard errors.

$$S(\hat{Y}) = \sqrt{\sum_{h=1}^L [S(\hat{Y}_h)]^2}$$

The next thing that must be determined is an appropriate t value, which is a functional confidence level, and the number of degrees of freedom, df_e . A formula for computing effective degrees of freedom is given in *Sampling Techniques* by Cochran. It is the formula used in the ES Program.

Total Estimate, Standard Error and Precision (cont'd)

$$df_e = \frac{\left(\sum_{h=1}^L g_h s_{yh}^2 \right)^2}{\sum_{h=1}^L g_h^2 s_{yh}^4 / (n_h - 1)}$$

where

$$g_h = \frac{N_h (N_h - n_h)}{n_h}$$

When a regression estimate is used $(n_h - 2)$ is used instead of $(n_h - 1)$.

The value of df_e always lies between the smallest of the values $(n_h - 1)$ and their sum.

The formula requires the assumption that the means of the y_{hi} are normally distributed. If they are not, df_e might overstate the effective number of degrees of freedom. In practice, the formula appears to work satisfactorily.



Total Estimate, Standard Error and Precision (cont'd)

Once the effective degrees of freedom are known, a t value can be determined from tables or, as the ES Program does it, by computation. If $C\%$ is the confidence level and $t(C, df_e)$ denotes the t value, then the precision of the estimate is:

$$\textit{Precision} = t(C, df_e) S(\hat{Y})$$



Mathematics of Difference Estimation

The difference estimation formulas make use of a few symbols and factors for stratum h in addition to those given in 1:

$$d_{hi} = y_{hi} - x_{hi} \quad \text{Difference for item } i \text{ between source and target values}$$

The mathematics of difference estimation is similar to the mathematics of MPU estimation, with the differences d_{hi} replacing the target values y_{hi} . The difference estimate for stratum h is:

$$\bar{d}_h = \frac{\sum d_{hi}}{n_h} \quad \text{Sample mean of differences}$$

The mathematics of difference estimation is similar to the mathematics of MPU estimation, with the differences d_{hi} replacing the target values y_{hi} . The difference estimate for stratum h is:

$$\hat{Y}_h = N_h \bar{d}_h + X_h$$



Mathematics of Difference Estimation

The sample difference standard deviation for stratum h is:

$$s_h = \frac{\sum (d_{hi} - \bar{d}_h)^2}{n_h - 1}$$

The standard error of the estimate in stratum h is:

$$S(Y_h) = N_h \frac{s_{dh}}{\sqrt{n_h}} \sqrt{1 - n_h / N_h}$$

The remaining calculations of total estimate, total standard error and precision follow the MPU calculations in section 2.

Note: For those interested in the alternative presentations of and referred to in section 7.2, these use the additional factors needed for ratio and regression estimates.

Mathematics of Difference Estimation (cont'd)

$$\sum d_{hi} = \sum y_{hi} - \sum x_{hi}$$

$$s_{dh} = \frac{\sum (x_{hi} - \bar{x}_h)^2 + \sum (y_{hi} - \bar{y}_h)^2 - 2 \sum (x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h)}{n_h - 1}$$

When using a difference estimator to calculate a difference rather than an amount, the formulas used in the calculation change. The difference for a difference estimate is:

$$\mathcal{D} = N\bar{d}$$

The standard error of the difference for a difference estimate is:

$$S(\mathcal{D}) = N \frac{s_d}{\sqrt{n}} \sqrt{1 - n/N}$$



Mathematics of Ratio Estimation

In stratified ratio estimation, the following symbols and factors are needed for each stratum in addition to those that were detailed in section 1.

$$\bar{x}_h = \frac{\sum x_{hi}}{n_h}$$

Sample mean of source variable

$$(x_{hi} - \bar{x}_h)$$

Difference between item source value and sample mean of source values

$$(y_{hi} - \bar{y}_h)$$

Difference between item target value and sample mean of target values

The total value of the source variable across all strata is:

$$X = \sum_{h=1}^L X_h$$



Separate Ratio Estimate

The ratio for stratum h is:

$$r_h = \frac{\bar{y}_h}{\bar{x}_h}$$

The ratio estimate for stratum h is:

$$\hat{Y}_h = r_h X_h$$

The sample separate ratio standard deviation for stratum h is:

$$s_{rh} = \sqrt{\frac{\sum (y_{hi} - \bar{y}_h)^2 - 2r_h \sum (x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h) + r_h^2 \sum (x_{hi} - \bar{x}_h)^2}{n_h - 1}}$$

The standard error of the estimate in stratum h is:

$$S(Y_h) = N_h \frac{s_{rh}}{\sqrt{n_h}} \sqrt{1 - n_h/N_h}$$

The remaining calculations of total estimate, total standard error, and precision follow the MPU calculations in section 2.



Combined Ratio Estimate

The ratio for all strata is the ratio of the weighted stratum means across all strata.

$$r = \frac{\sum_{h=1}^L N_h \bar{y}_h}{\sum_{h=1}^L N_h \bar{x}_h}$$

The combined ratio estimate for stratum h is:

$$\hat{Y}_h = rX_h$$

The sample combined ratio standard deviation for stratum h is:

$$s_{rh} = \sqrt{\frac{\sum (y_{hi} - \bar{y}_h)^2 - 2r \sum (x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h) + r^2 \sum (x_{hi} - \bar{x}_h)^2}{n_h - 1}}$$



Combined Ratio Estimate (cont'd)

The standard error of the estimate in stratum h is:

$$S(Y_h) = N_h \frac{S_{rh}}{\sqrt{n_h}} \sqrt{1 - n_h/N_h}$$

The remaining calculations of total estimate, total standard error, and precision follow the MPU calculations.

When using a ratio estimator to calculate a ratio rather than an amount, the formulas used in the calculation changes. The ratio for a ratio estimate is:

$$R = \frac{\bar{y}}{\bar{x}}$$

The standard error of the ratio for a ratio estimate can be calculated in two different ways, depending upon how much information is available. If the population total (X) and the population size (N) are both entered, then:

$$S(R) = \frac{S_r}{\sqrt{n}} \frac{1}{X} \sqrt{1 - n/N}$$



If the population total (X) and the population size (N) are *not* both entered, then:

$$S(R) = N \frac{s_r}{\sqrt{n}} \frac{1}{\bar{x}} \sqrt{1 - \frac{n}{N}}$$

where s_r is the sample ratio standard deviation.

Coefficient of Correlation

The degree of correlation between the source variable and the target variable is measured by the coefficient of correlation. This is used to determine the optimality of using ratio and regression estimation and may be used to improve mathematical sample designs based on source variable.

$$CC = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2} \sqrt{\sum (y - \bar{y})^2}}$$



Mathematics of Regression Estimation

Regression estimation follows very much the same pattern as ratio estimation. The only additional factors that are required in the calculations are:

$$\bar{X}_h = \frac{X_h}{N_h} \quad \text{Population mean of the source variable in each stratum}$$

$$W_h = \frac{N_h}{N} \quad \text{Stratum weight}$$

Separate Regression Estimate

The regression coefficient for stratum h is:

$$b_h = \frac{\sum (x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h)}{\sum (x_{hi} - \bar{x}_h)^2}$$



Separate Regression Estimate (cont'd)

The regression estimate for stratum h is:

$$\hat{Y}_h = N_h \left[\bar{y}_h + b_h \left(\bar{X}_h - \bar{x}_h \right) \right]$$

The sample separate regression standard deviation for stratum h is:

$$S_{gh} = \sqrt{\frac{\sum (y_{hi} - \bar{y}_h)^2 - b_h^2 \sum (x_{hi} - \bar{x}_h)^2}{n_h - 2}}$$

The standard error of the estimate in stratum h is:

$$S(Y_h) = N_h \frac{S_{gh}}{\sqrt{n_h}} \sqrt{1 - n_h/N_h}$$

Notice that the divisor for the standard deviation is (n_h-2) . With one exception, the formulas for total estimate, total standard error and precision are the same as those used for MPU estimates. The exception relates to the calculation of effective degrees of freedom. In that formula, the divisor of (n_h-1) must be replaced by (n_h-2) .



Combined Regression Estimate

The combined regression coefficient is :

$$b = \frac{\sum_{h=1}^L G_h \sum (x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h)}{\sum_{h=1}^L G_h \sum (x_{hi} - \bar{x}_h)^2}$$

where:

$$G_h = \frac{W_h^2 (1 - n_h / N_h)}{n_h (n_h - 1)}$$

is the factor used to weight the sum of cross products and the sum of squares for stratum h .

The regression estimate for stratum h is:

$$\hat{Y} = N_h [\bar{y}_h + b (\bar{X}_h - \bar{x}_h)]$$



Combined Regression Estimate (cont'd)

The sample combined regression standard deviation for stratum h is:

$$S_{gh} = \sqrt{\frac{\sum (y_{hi} - \bar{y}_h)^2 - 2b \sum (x_{hi} - \bar{x}_h)(y_{hi} - \bar{y}_h) + b^2 \sum (x_{hi} - \bar{x}_h)^2}{n_h - 2}}$$

The standard error of the estimate in stratum h is:

$$S(Y_h) = N_h \frac{S_{gh}}{\sqrt{n_h}} \sqrt{1 - n_h/N_h}$$

As for separate regression, the divisor for the standard deviation is $(n_h - 2)$

With one exception, the formulas for total estimate, total standard error and precision are the same as those used for MPU estimates. The exception relates to the calculation of effective degrees of freedom. In that formula, the divisor of (n_h-1) must be replaced by (n_h-2) .



Mathematics of PPS Estimation

The PPS estimation formulas make use of a few symbols and factors for stratum h that are not used in the other estimation formulas:

$$p_{hi} = \frac{y_{hi}}{x_{hi}} \quad \text{Proportion for item } i \text{ in stratum } h$$

$$\bar{p}_h = \sum \frac{p_{hi}}{n_h} \quad \text{Sample mean of proportions}$$

The mathematics of PPS estimation is similar to the mathematics of MPU estimation, with the proportions p_{hi} replacing the target values y_{hi} , and the source totals X_h replacing the stratum sizes N_h in most places.

The PPS estimate for stratum h is: $\hat{Y}_h = X_h \bar{p}_h$

The sample PPS standard deviation for stratum h is:

$$s_{ph} = \sqrt{\frac{\sum (p_{hi} - \bar{p}_h)^2}{n_h - 1}}$$



Mathematics of PPS Estimation

The standard error of the estimate in stratum h is:

$$S(\hat{Y}_h) = X_h \frac{S_{ph}}{\sqrt{n_h}} \sqrt{1 - n_h/N_h}$$

The remaining calculations of total estimate, total standard error, and precision follow the MPU calculations in section 2.

When using a PPS estimator to calculate an error proportion rather than an amount, the formulas used in the calculation changes. The point estimate for PPS Variables for an Error Proportion is:

$$\hat{p} = \frac{\sum p_i}{n} \qquad p_i = \frac{y_i - x_i}{x_i}$$

The standard deviation for PPS Variables for an Error Proportion is:

$$S_{ph} = \frac{\sum (p_{hi} - \bar{p}_h)^2}{n_h - 1}$$



Mathematics of PPS Estimation (cont'd)

The standard error for PPS Variables for an Error Proportion is:

$$S(\bar{Y}) = \frac{s_p}{\sqrt{n}} \sqrt{1-n / N}$$

Mathematics of Stratified-by-Value Sample Design

In this section, we give a brief account of the mathematics behind the methods used by the ES Program to calculate preliminary sample designs using stratification by source value. A more complete account can be found in Chapter 5 of *Sampling Techniques* by Cochran.



The Construction of Strata

Given a top cut-off and a bottom cut-off, and given the required number of strata, L , the problem is to find stratum cut-offs u_1, u_2, \dots, u_L , such that the standard error of the estimate is a minimum.

$$s(\hat{Y}) = \sqrt{\sum S^2(\hat{Y}_h)}$$

The optimal solution is not easy to find, even if using a computer. It requires a time-consuming, iterative approach.

Cochran presents a quick approximate method for solving the minimization problem that was first published by *Dalenius and Hodges* in 1959. In this method, a simplifying assumption is made that the population is distributed uniformly within strata. The narrower the strata, the more valid the assumption. The general idea behind the method is to compile a frequency distribution of the population and to set stratum boundaries so that they divide up equally the cumulative square root of that distribution. This is how it is done:

The Construction of Strata

1. Compile a frequency distribution of the population. This can be done by partitioning the interval between the bottom and the top cut-offs into a number of cells of equal length, and computing the population frequency within the cells. Let T denote the number of cells. The frequency in cell k can be denoted by $f(y_k)$, where y_k is the boundary of cell number k .
2. Calculate $\sqrt{f(y_k)}$ the square root of the frequency, for each cell.

Calculate the cumulative square root frequency for each cell. The cumulative square root for cell number k is:

$$CUM_k = \sum_{j=1}^k \sqrt{f(y_j)}$$

The cumulative square root for the last cell is CUM_T where:

$$CUM_T = \sum_{j=1}^T \sqrt{f(y_j)}$$



The Construction of Strata

Set the L stratum boundaries so that they create approximately equal intervals on the CUM_k scale. The length of each interval on that scale should then be approximately CUM_T/L

Table 1 shows how the method could be applied to the Gizmo data. In this example, the frequency distribution has been created by partitioning the interval 0 to 5000 into 50 cells. When this method is applied, care must be taken not to have too many cells relative to the size of the population, or too few relative to the number of strata. The ES Program creates a number of cells equal to 10 times L , the number of strata.



The Construction of Strata (cont'd)

Table 1 Gizmo. Stratification Using the $CUM \sqrt{f(y)}$

Cell No.	Cell Boundary	$f(y)$	$CUM \sqrt{f(y)}$	$CUM \sqrt{f(y)}$ Boundaries	Upper Cut- Offs
Bot.	0	0	0.00	0.00	0
1	100	65	8.06		
2	200	78	16.89		
3	300	55	24.31		
4	400	62	32.18		
5	500	52	39.40	38.70	500
6	600	50	46.47		
7	700	55	53.88		
8	800	45	60.59		
9	900	34	66.42		
10	1000	40	72.75		



The Construction of Strata (cont'd)

Cell No.	Cell Boundary	$f(y)$	$CUM \sqrt{f(y)}$	$CUM \sqrt{f(y)}$ Boundaries	Upper Cut- Offs
11	1100	35	78.66	77.41	1100
12	1200	29	84.05		
13	1300	21	88.63		
14	1400	35	94.55		
15	1500	30	100.02		
16	1600	23	104.82		
17	1700	21	109.40		
18	1800	12	112.87		
19	1900	20	117.34	116.11	1900
20	2000	14	121.08		
21	2100	9	124.08		
22	2200	15	127.95		
23	2300	15	131.83		
24	2400	11	135.14		
25	2500	12	138.61		
26	2600	12	142.07		
27	2700	14	145.91		
28	2800	9	148.81		
29	2900	10	151.97		
30	3000	8	154.80		
31	3100	8	157.63	154.82	3100
32	3200	6	160.08		
33	3300	4	162.08		
34	3400	6	164.53		
35	3500	3	166.26		
36	3600	2	167.68		



The Construction of Strata (cont'd)

Cell No.	Cell Boundary	$f(y)$	$CUM \sqrt{f(y)}$	$CUM \sqrt{f(y)}$ Boundaries	Upper Cut- Offs
37	3700	5	169.91		
38	3800	3	171.64		
39	3900	5	173.88		
40	4000	1	174.88		
41	4100	4	176.88		
42	4200	6	179.33		
43	4300	6	181.78		
44	4400	3	183.51		
45	4500	1	184.51		
46	4600	6	186.96		
47	4700	4	188.96		
48	4800	3	190.69		
49	4900	2	192.11		
50	5000	2	193.52	193.52	5000
Top		29			
		1000			



Estimation Sampling Training