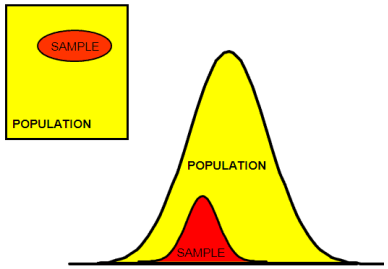
STATISTICS

Unit 3

1. Sampling theory in Statistics
1.1 Type of samples
1.2 Method of obtaining random samples
1.3 Central Limit Theorem
1.4 Distribution of random samples
1.5 Other sampling distributions
1.6 Sampling from finite populations
1.7 Student's t distribution
1.8 Degrees of freedom
1.9 Probable Error
1.10 Table of t – distribution

1 Sampling theory in Statistics

- Sampling theory is the field of statistics that is involved with the collection, analysis and the interpretation of data gathered from the random samples of a population under study.



- Population is a complete set of elements (persons or objects) that posses some common characteristic defined by the sampling criteria established by the researcher.
- Sample is the selected elements (people or objects) chosen for participation in a study; people are referred to as subjects or participants.
- Parameter is a numerical value or measure of a characteristic of the population like mean and variance.

Read more from :
http://www.course-notes.org/Statistics/Sampling_Theory
http://www.umsl.edu/~lindquists/sample.html

1 Sampling theory in Statistics- continued

- The application of sampling theory is concerned with

a. *proper selection of observations from a population that constitutes a sample*

b. *use of probability theory, along with prior knowledge about the population parameters, to analyze the data from the random sample and*

c. *develop conclusions from the analysis*

- The normal distribution, along with related probability distributions is most heavily utilized in developing the theoretical background for sampling theory.
- Samples play a very important part in statistical work because it is often impossible or too expensive to analyze the whole population.
- Information obtained from a sample or a set of samples is useful in the estimation of the unknown population parameters. This is called statistical inference or estimation.
- We often wish to compare two samples from the same population to determine the hypothesis that the differences are significant or not. This is part of decision theory.

1 Sampling theory in Statistics- continued

- *Power of a statistical test is the probability that it correctly rejects the null hypothesis (H_0), when it is false; alternatively, it is the probability of correctly accepting the alternative hypothesis (H_1) when it is true.*
- *The sample size is an important feature of any empirical study in which the goal is to make inferences about a population from a sample.*
- *In practice, the sample size used in a study is determined based on the expense of data collection; and the need to have sufficient statistical power.*
- Small sample (size < 30) ; Large sample (size >= 30)
- The limitations created by a small sample size can have profound effects on the outcome and worth of a study. A small sample size may have detrimental effects.
- The assumptions we make in case of large samples do not hold good for small samples; ie.,

  a. The random sampling distribution of a statistic is approximately normal
  b. Values given by the samples are sufficiently close to the population value and can be used in its place for calculating the standard error of the estimate.

1.1 Types of Samples

- *Bessel's correction for small sample: Use n − 1 instead of n in the formula for the sample variance and sample standard deviation, where n is the number of observations in a sample.*

Read more: http://www.ehow.com/info_8545371_effects-small-sample-size-limitation.html

Types of samples

1. *Random Sampling* – *Each member of the population has an equal chance of being chosen.*
2. *Stratified sampling* – *A heterogeneous population may be divided into homogeneous subgroups, and the sample is then drawn from each subgroup in a random manner.*
3. *Judgment sampling* – *This is the deliberate selection of a ample by the statistician, to obtain a representative cross section of the population.*

- A number of other terms are used to represent variants of these three major divisions, such as systematic, double, sequential, area, cluster, quota, and proportional.

1.1 Types of Samples – continued

1 ***Systematic*** *sample is a sample design in which a list of the population is used as a sampling frame and cases are selected by skipping through the list at regular intervals.*

2 ***Double*** *sampling is a standard form of sample design for industrial inspection purposes. In accordance with the characteristics of a particular plan, two samples are drawn, n1 and n2, and the first sample inspected. The batch then can be accepted or rejected upon the results of this inspection or the second sample be inspected and the decision made upon the combined result.*

3 ***Sequential*** *sample is in accordance with a sampling plan in which an undermined number of samples are tested one by one, accumulating the results until a decision can be made.*

1.1 Types of Samples – continued

4 ***Cluster*** *sampling is often clustered by geography, or by the time periods*

5 ***Area*** *sampling is a form of sampling in which the clusters that are selected are drawn from maps rather than listings of individuals, groups, institutions. Area sampling is a special form of cluster sampling in which the sample items are clustered are clustered on a geographic area basis*.

6 *In* ***quota*** *sampling, a population is first segmented into mutually exclusive sub-groups, just as stratified sampling. Then judgment is used to select the units from each segment based on a specified proportion.*

7 *In* ***proportional*** *sampling, the population is divided into sub-populations (strata) and random samples are taken of each stratum.*

1.2 Method of obtaining random samples

Method of obtaining random samples

1. Very often, all the units in the population to be sampled, are assigned or can be assigned a number. For example, automobiles have VIN, license numbers.

2. Sample can be chosen from the population using a table of random numbers. Tables of random numbers are usually tabulated in blocks of 5 digits.

3. A start may be made with any block on any page. If the serial numbers of the units to be sampled ran from 1 to 900, the last three digits of each block of five digits would be used. Any random number greater than 900 would be ignored. 1 would be treated as 001, 97 as 097 etc.

1.3 Central Limit Theorem

Central Limit Theorem

The central limit theorem states that the sampling distribution of the mean, for any set of independent and identically distributed random variables, will tend towards the normal distribution as the sample size gets larger.

1.4 Distribution of sample means

Distribution of sample means

- A number of samples, all of size N, are taken from a certain population, and the mean of each sample is calculated.
- We then have a new distribution – the distribution of the means of the samples.
- These sample means have a normal distribution, provided the sample size (N) is large, even though the population may not have a normal distribution. The mean of the distribution is μ_p, the mean of the population; and the standard deviation is σ.
- This standard deviation is called the Standard error of the sampling distribution of the means.

1.5 Other sampling distributions

Other sampling distributions

- Consider a proportion p and a large population, obtained by rolling a dice or other means, based on the proportion. If samples are taken from this population, the sampling distribution of proportion of successes will be p and the standard deviation (standard error) of the distribution will be

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

- Although the population is a binomial distribution, the sampling distribution of the proportion is close to normal.
- If two independent sets of samples are taken from two separate populations with means μ_1 and μ_1 and the standard deviation σ_1 and $\sigma_{2,}$
- Mean of the sum of the means $= \mu_1 + \mu_1$

1.6 Sampling from finite populations

Sampling from finite populations

- The central limit theorem and the standard errors of the mean and of the proportion are based on the premise that the samples selected are chosen with replacement.
- However, in virtually all survey research, sampling is conducted without replacement from populations that are of a finite size, N.
- In these cases, particularly when the sample size is at least 5% of the population size (i.e., sample size is not small), we need to apply a correction factor (i.e., finite population correction factor – fpc) that is used to define both the standard error of the mean and the standard error of the proportion.

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

where n is the sample size and N is the population size

1.6 Sampling from finite populations- continued

Sampling from finite populations

STANDARD ERROR OF THE MEAN FOR FINITE POPULATIONS

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

STANDARD ERROR OF THE PROPORTION FOR THE FINITE POPULATION

$$\sigma_{p_s} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Example: What is the correction factor to be applied to the standard deviation for a finite population where the population size is 100 and the sample size is 10?

fpc = sq. root [(100 – 10) / (100 –1)]

= sq. root [90 / 99] = $\sqrt{0.91}$

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

The factor to be applied to the standard deviation is sq. root of 0.91 = 0.95

1.7 Student's t distribution

Student's t distribution

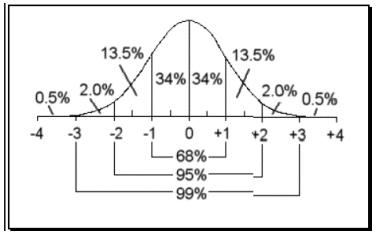
- The t distribution (aka, Student's t distribution) is a probability distribution that is used to estimate population parameters when the sample size is small and / or when the population variance is unknown.
- According to the central limit theorem, the sampling distribution of a statistic (e.g. sample mean) will follow a normal distribution, as long as the sample size is sufficiently large.
- But sample sizes are small, and often we do not know the standard deviation of the population. When either of the problems occur, statisticians rely on the distribution of the t-statistic (or t-score), whose values are given by

$t = [\bar{x} - \mu] / [s / \sqrt{n}]$ where $\bar{x}$ is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size.

1.7 Student's t distribution – continued

Student's t distribution

$$z = \frac{\bar{x} - \mu}{\sigma / N^{1/2}}$$



- z is a standard normal curve, where μ and σ refer to the population.
- In most cases σ is unknown and we must substitute

where s is the standard deviation of the sample.

$$\sigma_{est} = \sqrt{\frac{N}{N-1}} s$$

$$t = \frac{\bar{x} - \mu}{s / \sqrt{N-1}}$$

This equation is called student's t-distribution.

This t distribution approximates to normal distribution when N is large.

- The t-distribution allows us to conduct statistical analyses on certain data sets that are not appropriate for analysis, using the normal distribution.

1.8 Degrees of freedom

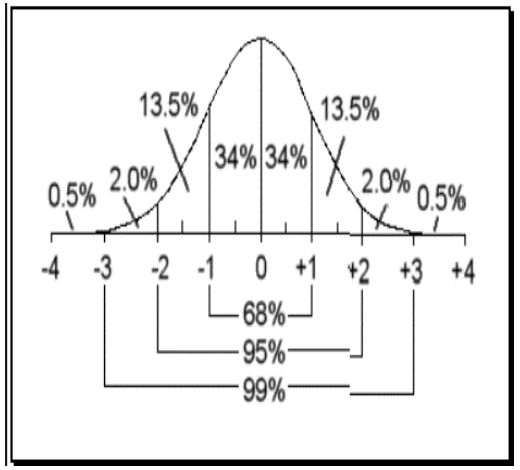
Degrees of freedom

- There are actually many different t distributions. The particular form of the t distribution is determined by its degrees of freedom.
- The degrees of freedom refers to the number of independent observations in a sets of data.
- When estimating a mean score or a proportion from a single sample, the number of independent observations is equal to the sample size minus one.
- Hence, the distribution of the t statistic from sample of size 8 would be described by a t distribution having 8 – 1 or 7 degrees of freedom.

1.9 Probable Error

Probable Error

- *The table for the areas under the normal curve enables us to determine the probability of values being within any particular range outside the mean.*
- *Thus, for from the range –σ to +σ, the probability is 68%; from the range –2σ to +2σ, the probability is 95.5%; from the range –3σ to +3σ, the probability is 99.7%. The range corresponding to 50% is called the probable error., since values are equally likely to be inside or outside this range, For the normal curve, this range is –.6745σ to +.6745σ*



- For the t–distribution, this range is larger. For 10 degrees of freedom the range is –.700σ to +.700σ and for 5 degrees of freedom the range is –.727σ to +.727σ.

1.10 Table of t distribution

Table of the t distribution

- In using the t distribution, we are normally concerned with the probability that a given value will be outside the range $-x\sigma$ to $+x\sigma$.

Degrees of freedom	*Probability*			
	0.5	*0.10*	*0.05*	*0.01*
1	1.000	6.31	12.71	63.66
2	0.816	2.92	4.30	9.92
3	0.765	2.35	3.18	5.84
4	0.741	2.13	2.78	4.60
5	0.727	2.02	2.57	4.03
10	0.700	1.81	2.23	3.17
20	0.687	1.72	2.09	2.84
∝	0.674	1.64	1.96	2.58

1.10 Table of t distribution – continued

Example

If the number of degrees of freedom is 10, what range of values will include 90% of the total number of means recorded in a large number of sample tests?

Solution:

- *If 90% of the values are within the range,10% will be outside the range.*
- *Entering the table with probability of 0.10 and 10 degrees of freedom, we obtain a value of 1.81.*
- *Hence the required range is -1.81σ to 1.81σ*

Degrees of freedom	Probability			
	0.5	0.10	0.05	0.01
1	1.000	6.31	12.71	63.66
2	0.816	2.92	4.30	9.92
3	0.765	2.35	3.18	5.84
4	0.741	2.13	2.78	4.60
5	0.727	2.02	2.57	4.03
10	0.700	1.81	2.23	3.17
20	0.687	1.72	2.09	2.84
∞	0.674	1.64	1.96	2.58