greatlearning

Reference Material for PGPBABI-SMDM Residency II

A. Statistical Inference Problem

As we have discussed before, sample is collected from a population to understand the population well. Descriptive statistics (descriptive analytics, managerial report etc) are all devices to summarize the information available in the sample. However, the goal of studying the sample is to get a good idea about the population, of which the sample is a representative.

The sample is a good representative of the population if the sample is a random sample.

The techniques to understand the population through a random sample comes under the umbrella of statistical inference. Recall that a population is known by its probability distribution, which, in turn, is completely defined by the population parameters. Statistical inference problems deal with 'knowing' the population parameters.

Two main divisions of inference problem are

- Estimation – Point estimation and Interval estimation
- Hypothesis testing

At this stage it is worth reiterating that a sample statistic is also a random variable, since it has an associated probability distribution. It is completely based on the observed sample and completely known. It does not depend on any unknown population parameter. Every statistic will also have a standard deviation.

A problem in point estimation involves finding appropriate statistic to estimate population parameter and formulating its standard deviation.

Example 1: A simple random sample of 5 months of sales data shows the number of items sold as follows: 94, 100, 85, 94, 92. A point estimate of the population mean number of units sold per month is $\bar{x} = 93$. Point estimate of population standard deviation is $s = 5.38$.

Example 2: A sample of 50 Fortune 500 companies showed 5 were based in New York, 6 in California, 2 in Minnesota and 1 in Wisconsin. A point estimate of the proportion of Fortune 500 companies based in NY is p = 5/50 = 10%.

Point estimates are intuitively appealing, but their numerical values change from sample to sample. Hence it is often more useful to consider a range of values for the population parameter under consideration, which is given by interval estimates.

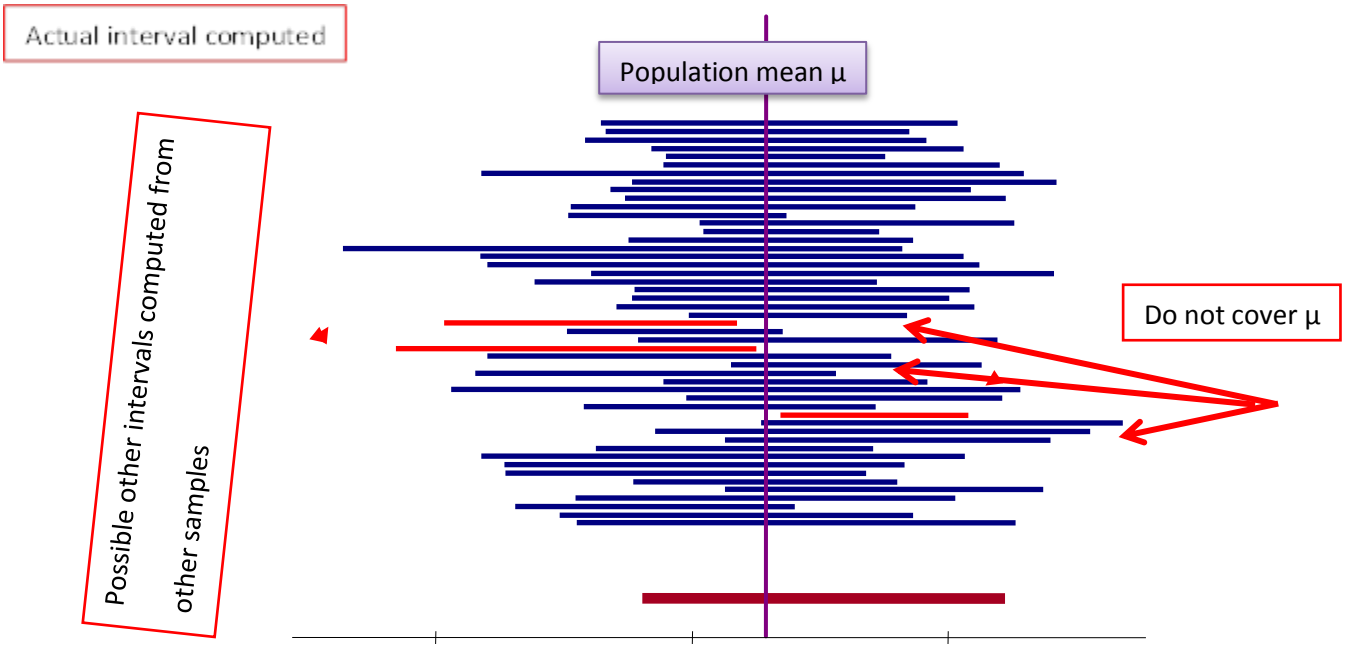
B. Confidence Interval of Population Mean

A confidence interval (interval estimate) is a range of values constructed around the point estimate. Both the point estimate and its standard deviation (standard error) are used in constructing the interval estimate. Since distribution of the point estimate is known, the confidence in the estimate, or the probability that the interval will include the population parameter is known. If samples are taken from the population repeatedly, then in 100P% of cases the interval will cover the population parameter, if 100P% is the associated confidence coefficient.

If repeated samples are taken from the same population, a different set of observations are available. Sample statistics computed on the basis of different set of observations are likely to be different. Hence, confidence intervals constructed on the basis of different samples will be different. Some of these confidence intervals will cover the true unknown population mean, but some of the intervals will not.

In a given situation, i.e. for a given sample, whether the parameter is included in the interval or not will remain unknown. Since the population parameter is unknown, which interval covers the parameter and which does not, will never be known.

Following diagram shows graphically that different samples will provide different interval estimates for the population mean. Depending on the confidence coefficient, a known proportion of the intervals computed are expected not to cover the true population mean.



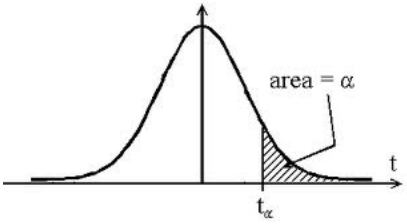
Formula for Confidence Interval for Population Mean

If a random variable X ~ N(μ, σ) distribution, then $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ ~ t with n – 1 degrees of freedom.

Parameter for this distribution is the degrees of freedom (DF). For very large DF t distribution becomes identical to N(0, 1) distribution. For small to moderate n, t distribution has heavier tails compared to N(0, 1) distribution. This distribution is used to construct confidence interval of population mean. When σ, population parameter, is unknown

100(1 – α)% CI: $\bar{x} \pm t(\alpha/2;n-1)\frac{s}{\sqrt{n}}$

α : Given constant, usually 5% or 1% and t(α/2;n – 1): Upper 100(α/2)-the percentile point of a N(0, 1) distribution



Example 1: Construct a 95% CI for population mean μ if $\bar{x}$ = 75, s = 12 and n = 49.

100(1 – α)% CI: $\bar{x} \pm t(\alpha/2;\ n\text{-}1)\frac{s}{\sqrt{n}}$

Given α = 0.05 t(α/2; 48) = 2.01 => CI: 75 ± 2.01(12/7) = (71.55, 78.45)

Example 2: An office manager would like to reduce the mean time for handling a customer complaint. She has looked at the record for 38 customers and noted that average time for handling customer complaints is 28.7 minutes with a stdev of 3.8 minutes.

a) What is the point estimate of the mean time required to handle a customer complaint?

Sample average is the point estimate of population mean: 28.7 min

b) What is the stdev of the point estimate?

Stdev of $\bar{x} = s/\sqrt{n} = 3.8 / \sqrt{38} = 0.6164$

c) Construct a 98% confidence interval for the mean time to handle customer complaints

98% CI : $\bar{x} \pm t(\alpha/2;\ n\text{-}1)\frac{s}{\sqrt{n}}$ = 28.7 ± t(0.01;37) 0.6164 = (27.20, 30.20)

INTERPRETATION OF CONFIDENCE INTERVAL:

If repeated samples are taken from the underlying population, then in 100(1-α)% cases the CI will cover the true population parameter. In the given problem: If the office manager selects a large number of samples of size n = 38 from records of handling customer complaint and for each one of them constructs a 98% CI, then she can expect 98% of the constructed intervals will cover the true population mean time of handling customer complaints.

Probability associated with a confidence interval is known as the Confidence Coefficient. Width of a CI is the difference between its Upper Bound (UB) and Lower Bound (LB). Usually a CI is symmetric about the sample average (but it does not have to be!). Width of a CI depends on

- Sample stdev: If s increases CI becomes wider
- Sample size: if n increases CI becomes shorter
- Confidence coefficient: If α increases CI becomes wider

C. Introduction to Hypothesis Testing

Hypothesis is a conjecture about a parameter of a population. Eg an HR manager may be interested knowing that a certain programme is 60% effective; an Operations manager may be interested in knowing that every cereal box contains 300g cornflakes; a Marketing manager may be interested in knowing that for every Rs 10,000 increase in advertising campaign, 50% increase in sales is expected. Each of the above involves statements regarding population parameters. Recall that parameter has an UNKNOWABLE constant value.

Based on belief or previous knowledge (process / domain / observation) a hypothesis is formulated.

Object of the hypothesis testing procedure is to SET a value for the parameter and perform a statistical TEST to see whether that value is tenable in the light of the evidence gathered from the sample.

Null hypothesis is the presumed current state of the matter (prevalent opinion / previous knowledge /basic assumption /prevailing theory)

H_0: Training is 60% effective

Alternative hypothesis is rival opinion / research hypothesis / improvement target. Three different alternatives are possible:

H_A: Training effectiveness is less than 60%

H_A: Training effectiveness is more than 60%

H_A: Training effectiveness is not equal to 60%

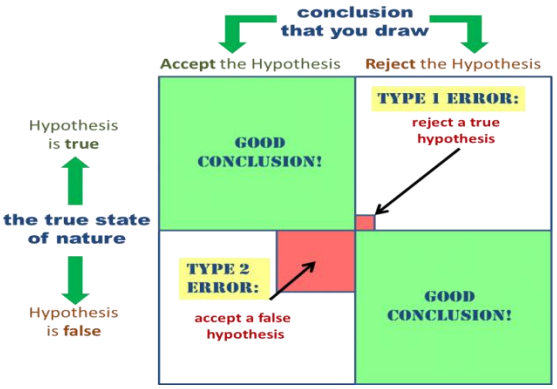
Null hypothesis is assumed to be true till reasonably strong evidence to the contrary is found. Based on a random sample a decision is made whether there exists enough evidence against H_0. Evidence is tested against a pre-determined Decision Rule. If H_0 is not supported (automatically) H_A comes into force.

A Test Statistic is a random variable based on H_0 and sample observations. It usually follows a standard distribution, eg normal, t, F, chi-square, etc and is used to make a choice between H_0 and H_A. Since hypothesis testing is done on the basis of sampling distribution, the decisions made are probabilistic. Hence it is very important to understand the errors associated with hypothesis testing.

Errors are NOT mistakes.

Type I error: If H_0 is TRUE but based on observed sample and decision rule we Reject H_0

Type II error: If H_0 is FALSE but based on observed sample and decision rule we Do Not Reject H_0



To make a valid conclusion regarding a null hypothesis, the following steps are in order:

i) Set up Null and Alternative (Research) hypotheses: H_0 and H_A
ii) Construct proper test statistic
iii) Consider distribution of test statistic ASSUMING H_0 is true
iv) Set a REJECTION RULE a priori
v) Evaluate test statistic, compare and conclude
vi) For univariate data, i.e. when data on a single variable is studied, to summarize the information, univariate statistics, e.g. mean, median, stdev, quartiles etc are considered and graphically histograms and box-plots are looked at. For bivariate data, in addition to all of the above for each of the component variables, scatterplot and correlation are considered. Scatterplot and correlation measures the strength of relationship between the two components.