# e-PGPathshala
# Subject : Computer Science
# Paper: Data Analytics
# Module: Statistical Inferenceing
# Module No: CS/DA/7
# Quadrant 1 – e-text

## 1.1 Introduction

The major truth of statistics is that results vary with samples. Understanding this concept of variability between all possible samples helps to determine how typical or atypical may be our particular result. Sampling distributions provide a fundamental to answer these problems. In statistical inference experimental or observational data are modeled as the observed values of random variables, which give a framework from which useful conclusions may be drawn. To perform a hypotheses test in the statistical procedure is to state two hypotheses and use an appropriate statistical test to reject one of the hypotheses and accept the other.

## 1.2 Learning Outcomes

- To know the fundamentals statistical inference and reasoning

- To understand the basics of hypothesis testing for analytics

## 1.3 Sampling Distribution

- A **sampling distribution** is the probability distribution for all possible values of the sample statistic.
- The **sampling distribution** of a statistic is the **distribution** of that statistic, considered as a random variable, when derived from a random **sample** of size n. It may be considered as the **distribution** of the statistic for all possible samples from the same population of a given size.
- Each sample contains different elements so the value of the sample statistic differs for each sample selected. These statistics provide different estimates of the parameter. The sampling distribution describes how these different values are distributed.
- With the sampling distribution of $\overline{x}$ , we can “make probability statements about how close the sample mean is to the population mean μ”. Alternatively, it provides a way of determining the probability of various levels of sampling error.

### 1.3.1 Central Limit Theorem

If all possible random samples of size N are drawn from a population with mean $m_y$ and a standard deviation $\sigma_y$ , then as N becomes larger, the sampling distribution of sample means becomes approximately normal, with mean $m_y$ and standard deviation $\sigma_y/N$.

In probability theory, the **central limit theorem** (**CLT**) states that, given certain conditions, the arithmetic mean of a sufficiently large number of iterates of independent random variables, each with a well-defined expected value and well-defined variance, will be approximately normally distributed, regardless of the underlying distribution. To illustrate what this means, suppose that a sample is obtained containing a large number of observations, each observation being randomly generated in a way that does not depend on the values of the other observations, and that the arithmetic average of the observed values is computed. If this procedure is performed many times, the central limit theorem says that the computed values of the average will be distributed according to the normal distribution (commonly known as a "bell curve"). A simple example of this is that if one flips a coin many times, the probability of getting a given number of heads should follow a normal curve, with mean equal to half the total number of flips.

### 1.3.2 Sampling Distribution of mean

- When a sample is selected, the sampling method may allow the researcher to determine the sampling distribution of the sample mean $\overline{x}$ . By assuming that the mean of the sampling distribution will be μ, ie, the mean of the population. If this occurs , then the expected value of the statistic $\overline{x}$ is μ. This characteristic of the sample mean is that of being an unbiased estimator of μ. In this case, $E(\overline{x})=\mu$
- If the variance of the sampling distribution can be determined , then the researcher is able to determine how variable $\overline{x}$ is when there is a repeated sample. The researcher hopes to have a small variability for the sample means, so most estimates of μ are close to μ.
- A fair **die** is thrown infinitely many times, with the random variable X = # of spots on any throw.
- The probability distribution of X is:

| X | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| P(x) | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

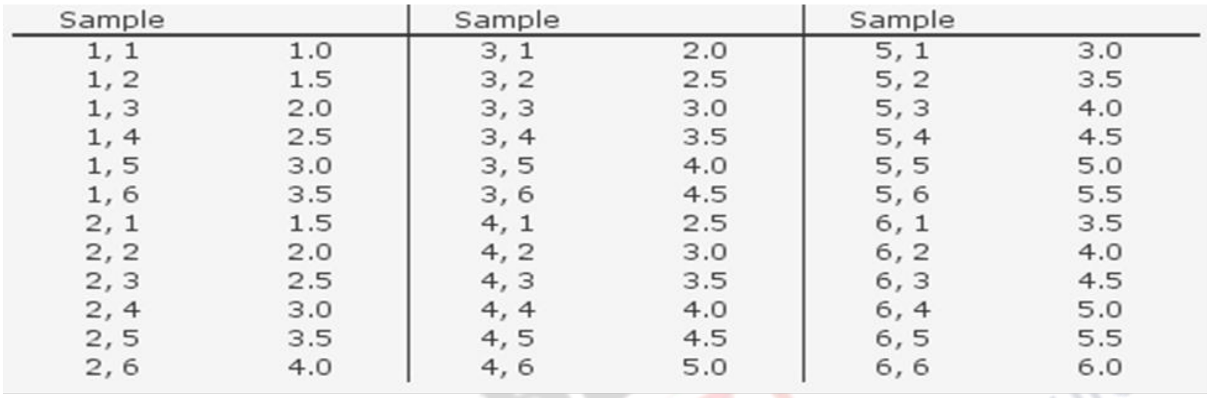
and the mean and variance are calculated as well:

$$\mu = \sum xP(x) = 1(\tfrac{1}{6}) + 2(\tfrac{1}{6}) + \ldots + 6(\tfrac{1}{6}) = 3.5$$

$$\sigma^2 = \sum (x-\mu)^2 P(x) = (1-3.5)^2(\tfrac{1}{6}) + \ldots + (6-3.5)^2(\tfrac{1}{6}) = 2.92$$
$$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$$

### 1.3.3 Sampling Distribution of Two Dice

A sampling distribution is created by looking at all samples of size n=2 (i.e. two dice) and their mean….



| Sample | | Sample | | Sample | |
|---|---|---|---|---|---|
| 1, 1 | 1.0 | 3, 1 | 2.0 | 5, 1 | 3.0 |
| 1, 2 | 1.5 | 3, 2 | 2.5 | 5, 2 | 3.5 |
| 1, 3 | 2.0 | 3, 3 | 3.0 | 5, 3 | 4.0 |
| 1, 4 | 2.5 | 3, 4 | 3.5 | 5, 4 | 4.5 |
| 1, 5 | 3.0 | 3, 5 | 4.0 | 5, 5 | 5.0 |
| 1, 6 | 3.5 | 3, 6 | 4.5 | 5, 6 | 5.5 |
| 2, 1 | 1.5 | 4, 1 | 2.5 | 6, 1 | 3.5 |
| 2, 2 | 2.0 | 4, 2 | 3.0 | 6, 2 | 4.0 |
| 2, 3 | 2.5 | 4, 3 | 3.5 | 6, 3 | 4.5 |
| 2, 4 | 3.0 | 4, 4 | 4.0 | 6, 4 | 5.0 |
| 2, 5 | 3.5 | 4, 5 | 4.5 | 6, 5 | 5.5 |
| 2, 6 | 4.0 | 4, 6 | 5.0 | 6, 6 | 6.0 |

While there are 36 possible samples of size 2, there are only 11 values for $\bar{x}$ , and some (e.g. $\bar{x}$ =3.5) occur more frequently than others (e.g. $\bar{x}$ =1).
The ***sampling distribution*** of $\bar{x}$ is shown below:

$$\mu_{\bar{x}} = \sum \bar{x} P(\bar{x}) = 1.0(\tfrac{1}{36}) + 1.5(\tfrac{2}{36}) + \ldots + 6.0(\tfrac{1}{36}) = 3.5$$

$$\sigma^2_{\bar{x}} = \sum (\bar{x}-\mu_{\bar{x}})^2 P(\bar{x}) = (1.0-3.5)^2(\tfrac{1}{36}) + \ldots + (6.0-3.5)^2(\tfrac{1}{36}) = 1.46$$
$$\sigma_{\bar{x}} = \sqrt{\sigma^2_{\bar{x}}} = \sqrt{1.46} = 1.21$$

| $\bar{x}$ | P($\bar{x}$) |
|---|---|
| 1.0 | 1/36 |
| 1.5 | 2/36 |
| 2.0 | 3/36 |
| 2.5 | 4/36 |
| 3.0 | 5/36 |
| 3.5 | 6/36 |
| 4.0 | 5/36 |
| 4.5 | 4/36 |
| 5.0 | 3/36 |
| 5.5 | 2/36 |
| 6.0 | 1/36 |

### 1.3.4 Resampling

The resampling method concerns a key problem in statistics that how to conclude the 'truth' from a sample of data that is incomplete or drawn from an unclear population.**I**n statistics **resampling** is used for :

1. Estimating the precision of sample statistics (mean, variance, percentiles) by using subsets of available data or drawing randomly with replacement from a set of data points.
2. Exchanging labels on data points when performing significance tests (**permutation tests**, also called exact tests, randomization tests, or re-randomization tests)
3. Validating models by using random subsets.

Common resampling techniques include bootstrapping, Jack knifing and permutation tests.

**Resampling procedures** are based on the assumption that the underlying population distribution is the same as a given sample. The approach is to create a large number of samples from this pseudo-population using the techniques described in sampling and then draw some conclusions from some statistic (mean, median, etc.) of the sample.

**Types of Re-sampling Methods**

I. **Monte Carlo Simulation –** This is a method that derives data from a mechanism (such as a proportion) that models the process to understand (the population). This will produce new samples of simulated data, which can be examined as possible results. After doing many repetitions, Monte Carlo tests produce exact p-values that can be interpreted as an error rate; to allow the number of repeats sharpens the critical region.

II. **Randomization (Permutation) Test –** this is a type of statistical significance test, in which a reference distribution is obtained by calculating all possible values of the test statistic by rearranging the labels on the observed data points. Like other above mentioned approach, permutation methods for significance testing also produce exact p-values. These tests are the oldest, simplest, and most common form of resampling tests and are suitable whenever the null hypothesis makes all permutations of the observed data equally likely. In this method, data is reassigned randomly without replacement. They are usually based on the Student t and Fisher’s F test. Most non-parametric tests are based on permutations of rank orderings of the data. This method has become practical because of computers. It may be impossible to derive all the possible permutations without computers .This method can be used when dealing with an unknown distribution.

III. **Bootstrapping –** This approach is based on the fact that all we know about the underlying population is what we derived in our samples. Becoming the most widely used resampling method, it estimates the sampling distribution of an estimator by sampling with replacement from the original estimate, most often with the purpose of deriving robust estimates of standard errors and confidence intervals of a population parameter. Like all Monte Carlo based methods, this approach can be used to define confidence Intervals and in hypothesis testing. This method is beneficial to side step problems with non-normality or if the distribution parameters are unknown. This method can be used to calculate an appropriate sample size for experimental design.

IV. **Jackknife** – This method is used in statistical inference to estimate the bias and standard error in a statistic, when a random sample of observations is used to calculate it. This method provides a systematic method of resampling with a mild amount of calculations. It offers “improved” estimate of the sample parameter to create less sampling bias. The basic idea behind the jackknife estimator lies in systematically re-computing the statistic estimate leaving out one observation at a time from the sample set. From this new “improved" sample statistic can be used to estimate the bias can be variance of the statistic.

## 1.4 Statistical inference

Statistical inference means making conclusions based on extracting data. One context for inference is the parametric model, in which data are supposed to come from a certain distribution family, the members of which are distinguished by differing parameter values. The normal distribution family is an example. One tool of statistical inference is the likelihood ratio, in which a parameter value is considered “consistent with the data” if the ratio of its likelihood to the maximum likelihood is at least some threshold value, such as 10% or 1%. While more sophisticated inferential tools exist, this one may be the most straight forward and obvious.

At the core of statistics lie the ideas of statistical inference. Statistical inference methods enables the investigator to argue from the particular observations in a sample to the general case. In contrast to logical deductions made from the general case to the specific case, a statistical inference can sometimes be incorrect. One of the great intellectual advances of the this century is the realization that strong scientific evidence can be developed on the basis of observations.

The subject of statistical inference extends beyond statistics' historical purposes of describing and displaying data. It deals with collecting informative data, interpreting and making conclusions of these data. Statistical inference includes all processes of acquiring knowledge that involve fact finding through the collection and examination of data. These processes are as diverse as opinion polls, agricultural field trials, clinical trials of new medicines, and the studying of properties of exotic new materials. As a result, statistical inference has spread all fields of human endeavor in which the evaluation of information must be grounded in data based evidence.

A few characteristics are common to all studies involving fact finding through the collection and interpretation of data. First, in order to acquire new knowledge, relevant data must be collected. Second, some variability is unavoidable when observations are made under the same or similar conditions. The third, which sets the stage for statistical inference, is that access to a complete set of data is either not feasible from a practical standpoint or is physically impossible to obtain.

To more fully describe statistical inference, it is necessary to introduce several key terminologies and concepts. The first step in making a statistical inference is to model the population by a *probability distribution* which has a numerical feature of interest called a *parameter*. The problem of statistical inference arises once we want to make generalizations about the *population* when only a *sample* is available. Based on a sample, a statistics, must serve as the source of information about a parameter. Three salient points guide the development of procedures for statistical inference:

1. Because a sample is only part of the population, the numerical value of the statistic will not be the exact value of the parameter.
2. The observed value of the statistic depends on the particular sample selected.
3. Some variability in the values of a statistic, over different samples, is unavoidable.

The two main classes of inference problems are *estimation* of parameter and *testing hypotheses* about the value of the parameter. The first class consists of point estimators, a single number estimate of the value of the parameter, and interval estimates. Typically, the interval estimate specifies an interval of reasonable values for the parameter but the subclass also includes prediction intervals for future observations. A test of hypotheses provides a yes or no answer as to whether the parameter lies in a specified region of values.

Because statistical inferences are based on a sample, they will sometimes be in error. Because the actual value of the parameter is unknown, a test of hypotheses

may yield the wrong yes or no answer and the interval of observable values may not contain the true value of the parameter.
Statistical inferences, or generalizations from the sample to the population, are founded on an understanding of the manner in which variation in the population is transmitted, via sampling, to variation in a statistic.

There are two main approaches, *frequentist* and *Bayesian*, for making statistical inferences. They are based on the *likelihood* but their frameworks are entirely different.

The frequentist treats parameters as fixed but unknown quantities in the distribution which organize variation in the sample. Then, the frequentist tries to protect against errors in inference by controlling the probabilities of these errors. The long-run relative frequency interpretation of probability then guarantees that if the experiment is repeated many times only a small proportion of times will produce incorrect inferences. Using this approach different problem keeps the overall proportion of errors less.

## 1.5 Hypothesis Testing

A statistical hypothesis **is** an assumption about a population parameter. This assumption may or may not be true. Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses.

### Statistical Hypotheses

The best way to determine whether a statistical hypothesis is true would be to examine the entire population. Since that is often impractical, researchers typically examine a random sample from the population. If sample data are not consistent with the statistical hypothesis, the hypothesis is rejected. There are two types of statistical hypotheses.

- **Null hypothesis**. The null hypothesis, denoted by $H_0$, is usually the hypothesis that sample observations result purely from chance.
- **Alternative hypothesis**. The alternative hypothesis, denoted by $H_1$ or $H_a$, is the hypothesis that sample observations are influenced by some non-random cause.

For example, suppose we wanted to determine whether a coin was fair and balanced. A null hypothesis might be that half the flips would result in Heads and half, in Tails. The alternative hypothesis might be that the number of Heads and Tails would be very different. Symbolically, these hypotheses would be expressed as:

$$H_0: P = 0.5$$
$$H_a: P \neq 0.5$$

Suppose we flipped the coin 50 times, resulting in 40 Heads and 10 Tails. Given this result, we would be inclined to reject the null hypothesis. We would conclude, based on the evidence, that the coin was probably not fair and balanced.

### 1.5.1 Directional and Nondirectional hypotheses

A directional alternative hypothesis states that the null hypothesis is wrong, and also specifies whether the true value of the parameter is greater than or less than the reference value specified in null hypothesis.

A non-directional alternative hypothesis states that the null hypothesis is wrong. A non-directional alternative hypothesis does not predict whether the parameter of interest is larger or smaller than the reference value specified in the null hypothesis.

The advantage of using a directional hypothesis is increased power to detect the specific effect you are interested in. The disadvantage is that there is no power to detect an effect in the opposite direction.

**Example of directional and non- directional hypotheses**

**Non-directional**

A researcher has results for a sample of students who took a national exam at a high school. The researcher wants to know if the scores at that school differ from the national average of 850. A non-directional alternative hypothesis is appropriate because the researcher is interested in determining whether the scores are either less than or greater than the national average. ($H_0$: μ = 850 vs. $H_1$: μ≠ 850)

**Directional**

A researcher has exam results for a sample of students who took a training course for a national exam. The researcher wants to know if trained students score above the national average of 850. A directional alternative hypothesis can be used because the researcher is specifically hypothesizing that scores for trained students are greater than the national average. ($H_0$: μ = 850 vs. $H_1$: μ > 850

### 1.6 Inferential reasoning

Inferential reasoning is an important component of statistics. Researchers have suggested that students should develop an informal understanding of the ideas that underlie inference before learning the concepts formally.

Statistical inference is the process of making conclusions about populations or scientific truths from data.There are many modes of performing inference including statistical modeling, data oriented strategies and explicit use of designs and randomization in analyses. Furthermore, there are broad theories (frequentists, Bayesian, likelihood, design based) and numerous complexities (missing data, observed and unobserved confounding, biases) for performing inference. Since everyday life involves making decisions based on data, making inferences is an important skill to have. However, a number of studies on assessments of students' understanding statistical inference suggest that students have difficulties in reasoning about inference.

Given the importance of reasoning about statistical inference and difficulties that students have with this type of reasoning, statistics educators and researchers have been exploring alternative approaches towards teaching statistical inference. Recent research suggests that students have some sound intuitions about data and these intuitions can be refined and nudged towards prescriptive theory of inferential reasoning. More of an informal and conceptual approach that builds on the previous big ideas and makes connection between foundational concepts is therefore favorable.

Recently, informal inferential reasoning has been the focus of research and discussion among researchers and educators in statistics education as it is seen as having a potential to help build fundamental concepts that underlie formal statistical inference. Many advocate that underlying concepts and skills of inference should be introduced as they can help make the formal statistical inference more accessible.

According to Statistical Reasoning, three essential principles to informal inference are :

1. generalizations (including predictions, parameter estimates, and conclusions) that go beyond describing the given data;
2. the use of data as evidence for those generalizations; and
3. conclusions that express a degree of uncertainty, whether or not quantified, accounting for the variability or uncertainty that is unavoidable when generalizing beyond the immediate data to a population or a process.

Informal inferential reasoning involved the following related ideas

- *Properties of aggregates*. This includes the ideas of distributions, signal (a stable component of population/process such as averages) and noise (a variable component of population/process such as the deviations of individual value around an average) and types of 'noise' or variability (measurement variability, natural variability, sampling variability).
- *Sample size*. Bigger samples are better because they provide a more accurate estimate of the population/process signals.

- *Controlling for bias*. The use of random sampling to be sure not to introduce bias in the sampling process and thus increase the chance that the sample we get is representative of the population
- *Tendency*. Distinguish between claims that are always true and that are often or sometimes true.

### 1.6.1 Type I and Type II Errors(α and β errors)

When doing hypothesis testing, two types of mistakes may be occured and we call them Type I error and Type II error.

| *Decision* | *Reality* | |
|---|---|---|
| | H0H0 is true | H0H0 is false |
| Reject $H_o$ | Type I error | Correct |
| Accept $H_o$ | Correct | Type II error |

If we reject $H_0$ when $H_0$ is true, we commit a Type I error. The probability of Type I error is denoted by: α.
If we accept $H_0$ when $H_0$ is false, we commit a Type II error. The probability of Type II error is denoted by: β

Our convention is to set up the hypotheses so that Type I error is the more serious error. Type I error is known as a false positive or error of the first kind. It is the error of rejecting a null hypothesis when it is actually true. In other words, this is the error of accepting an alternative hypothesis (the real hypothesis of interest) when the results can be attributed to chance i.e., it occurs when we are observing a difference when in truth there is none. So the probability of making a type I error in a test with rejection region R is 0.

Type II error is known as a false negative error of the second kind. It is the error of not rejecting a null hypothesis when the alternative hypothesis is the true state of nature. In other words, this is the error of failing to accept an alternative hypothesis when you don't have adequate power. ie, it occurs when we are failing to observe a difference when in truth there is one. So the probability of making a type II error in a test with rejection region R is 1. Understanding Type I and Type II Errors Hypothesis testing is the art of testing if variation between two sample distributions can just be explained through random chance or not. If we have to conclude that two distributions vary in a meaningful way, we must take enough precaution to see that the differences are not just through random chance. The main point of Type I error is that we don't want to make an unwarranted hypothesis so we exercise a lot of care by minimizing the chance of its occurrence. Traditionally we try to set Type I error as .05 or .01 - as in there is only a 5 or 1 in 100 chance that the variation that we are seeing is due to chance. This is called the 'level of significance'. Again, there is no guarantee that 5 in 100 is rare enough so significance levels need to be chosen carefully.

The null hypothesis is either true or false, and represents the default claim for a treatment or procedure. For example, when examining the effectiveness of a drug on a disease, the null hypothesis would be that the drug has no effect on a disease. Type I errors are equivalent to false positives. If we reject the null hypothesis in this situation, then our claim is that the drug does in fact have some effect on a disease. But if the null hypothesis is true, then in reality the drug does not combat the disease at all. The drug is falsely claimed to have a positive effect on a disease. Type II errors are equivalent to false negatives. If we think back again to the scenario in which we are testing a drug, how will be the type II error. A type II error would occur if we accepted that the drug had no effect on a disease, but in reality it did.

**Case Studies**

A) Hypothesis testing - Royal Mint in England

Source: http://www.dummies.com/how-to/content/applying-hypothesis-testing-in-business-statistics.html

One particularly interesting application of hypothesis testing comes from the Royal Mint in England. The Royal Mint has been producing coins for more than 1,100 years. It currently produces coins for circulation in the United Kingdom, as well as commemorative coins. It also produces coins and medals for foreign governments.

Throughout the Royal Mint's history, producing coins of consistent quality has been a constant challenge. Testing the quality of each coin would be a huge burden, so during the late 13th century, the Royal Mint developed a process for testing the quality of its coins based on randomly chosen samples. This process is known as the "Trial of the Pyx." (The Pyx refers to wooden chests that contain the random samples of coins; the name derives from the Pyx Chamber in Westminster Abbey where the chests were once kept.)

As part of the trial, samples of coins produced throughout the year are set aside to determine whether they have the proper weight, diameter, and chemical composition. The coins are then presented to a jury for testing. (Members of the jury were originally goldsmiths, who had expertise in assessing the weight and purity of gold and silver.)

The Trial of the Pyx can be seen as a hypothesis test, where the null and alternative hypotheses are as follows:

Null hypothesis: The coins conform to the required weight, diameter, and composition.

Alternative hypothesis: The coins don't conform to the required weight, diameter, and composition.

The coins being tested are compared with a standard coin, known as a Trial Plate. If the coins are significantly different from the Trial Plate, the null hypothesis is rejected. This outcome indicates that the production process is flawed and needs to be fixed. Otherwise, the null hypothesis is not rejected, and the coins are acceptable.

Note: Although the Trial of the Pyx predates the formal development of hypothesis testing by hundreds of years, it's an excellent example of how you can use hypothesis testing to draw conclusions from sample data with a high level of confidence. In fact, the Royal Mint was able to establish a reputation for producing coins of consistent quality; clearly, its testing procedures were a major source of its success.

**Summary**

- Re-sampling techniques eliminate the drawbacks of sampling
- Hypothesis testing is one commonly used technique for statistical inference
- Statistical inference and reasoning are very important for analytics