Sampling, Sampling methods and re-sampling

Statistics is the science of collecting, organizing, analyzing and interpreting data in order to make decisions.

Population is a collection of all outcomes, responses, measurement or counts that are of interest.

A **sample** is a subset of non-overlapping elements from the population.

Terms used in sampling

① Population mean (μ) :- average of a group characteristic

$$\mu = \frac{\sum x}{N}$$

$x \rightarrow$ all items in group
$N \rightarrow$ # items in group.

Sample mean ($\bar{x}$) :- average value of a sample

$$\bar{x} = \frac{\sum x_i}{n}$$

$x_i \rightarrow$ items in sample
$n \rightarrow$ sample size.

② Population variance (σ^2) :- measure of the spread of the population data (each pt)
It is calculated as the average of the distances of each data point into population to the mean squared.

$$\sigma^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}$$

$x_i \rightarrow$ Observations in population
$\mu \rightarrow$ mean of populatn
$N \rightarrow$ # observations

Sample variance (s^2) :-

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

$\bar{x} \rightarrow$ sample mean
$n \rightarrow$ sample size.

(3) Standard deviation of population (σ): spread of a group of numbers from the mean.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{N}}$$

Std deviation of sample (s):

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x - \bar{x})^2}{n-1}}$$

(4) Sample proportion

p = Count of successes in sample / size of sample.

Parameter :- variable of interest

Statistic :- information obtained from the sample about the parameter

We measure the sample using statistics in order to draw inferences about the population and its parameters.

Sampling Error

→ error caused by the act of taking a sample

→ It is the difference between the result taken out from a sample and the result taken out from the population.

→ sampling error depends on the size of the sample.

→ smaller the sample size, larger the error.

Classification of sampling methods

- Probability sampling
  - → systematic
  - → simple Random
  - → cluster
  - → stratified
  - → multistage
- Non probability sampling
  - → convenience
  - → judgment
  - → quota
  - → snowball.

Probability sampling

→ each and every unit of the population has an equal chance for selection as a sampling unit.

1) Simple Random sampling :-

- each element has an equal chance of being included in the sample
- selected using random number generators
- types 1) with replacement :- a unit once selected has the chance for again selection.
  2) without replacement :- a unit once selected cannot be selected again.

Stratified sampling :-

population is divided into two or more groups called strata, according to some criterion, such as geographic location, grade level, age or income and subsamples are randomly selected from each strata.

Elements within each strata are homogeneous, but heterogeneous across strata.

3) cluster sampling :-

A population is divided into subgroups (clusters). A simple random sample is taken of the subgroups and then all members of the cluster selected are surveyed.

4) Systematic random sampling

order all units based on some variable and then every nth number on the list is selected.

5) Multistage random sampling :-

- perform cluster sampling and then select a few from each cluster using some other sampling technique.

Non-Probability sampling

→ probability of each case being selected from the total population is not known.

1) Judgement sampling :- an experienced researcher selects the sample based on some appropriate characteristics of sample members.

quota sampling

- population is divided into cells on the basis of relevant control characteristics.
- a quota of sample units is established for each cell.

(3) Snowball sampling

- research starts with a key case.
- this case introduces/identifies the next case.
- stop when no new cases are given, or the sample is large enough.

(4) Convenience sampling :-

- selecting haphazardly those cases that are the easiest to obtain.

Problm 1 :- On a challenging math exam, with 15 pts possible, the scores of the eight test-takers were 6, 3, 8, 2, 5, 11, 6, 7. Calculate the population std deviation.

Soln

$$\mu = \frac{6+3+8+2+5+11+6+7}{8} = \frac{48}{8} = 6$$

$(x-\mu)^2$

$(6-6)^2 = 0^2 = 0$

$(3-6)^2 = 9$

$(8-6)^2 = 4$

$(2-6)^2 = 16$

$(5-6)^2 = 1$

$(11-6)^2 = 25$

$(6-6)^2 = 0$

$(7-6)^2 = 1$

$$\sigma^2 = \frac{0+9+4+16+1+25+0+1}{8}$$

$$= 7$$

$$\sigma = \sqrt{7} = 2.65 \text{ pts}.$$

Problm 2

The weights, in grams, of a population of 10 lab rats are 238, 257, 265, 267, 268, 273, 275, 276, 280, 281.

$\mu = 268$

$\sigma^2 = 205.7$

$\sigma = 14.34$ gms

Chapter

8

Descriptive Statistics

There are some important measures which help to know the data better. These measures give the idea of over all distribution of the observations in the data-set. These measures together can be called as descriptive statistics. In the first section of this chapter we discuss the measures of central tendency; in second section, measures of dispersion; measures of asymmetry in third section and measures of relationship in fourth section.

8.1 MEASURES OF CENTRAL TENDENCY

These measures are also called as statistical averages or averages. A measure of central tendency is a value around which all the observations have a tendency to cluster. Such a value is considered as the most representative figure of the entire data-set. Three most popular and important measures of central tendency are mean, median, and mode.

8.1.1 Mean

It is the most common measure of central tendency and may be defined as the value which we get by dividing the total of the values of various given items in a series by the total number of items. we can work it out as under:

$$\text{Mean (or } \bar{X}) = \frac{\Sigma X_i}{n} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

where, $\bar{X}$ = The symbol we use for mean (pronounced as X bar)

Σ = Symbol for summation

X_i = Value of the ith item X, $i = 1, 2, \ldots, n$

n = Total number of items

In case of frequency distribution when the whole data set of size n is summarized in k class intervals, mean is calculated as

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i} = \frac{f_1 X_1 + f_2 X_2 + \ldots + f_k X_k}{f_1 + f_2 + \ldots + f_k}$$

where, X_i = Mid-point of ith class interval and $f_1 + f_2 + \ldots + f_k = n$.

Sometimes, instead of calculating the simple mean, as stated above, we may workout the weighted mean for a realistic average. The weighted mean can be worked out as follows:

$$\bar{X}_w = \frac{\sum w_i X_i}{\sum w_i}$$

where, $\bar{X}_w$ = Weighted item

w_i = Weight of ith item X, assigned by the researcher using the some prior knowledge

X_i = Value of the ith item X

Mean is the simplest and most widely used measure of central tendency. People use mean in daily life so much that it has become a synonym of average. For example, Ravi consumes 4 cigarettes, on an average daily; Suresh drinks approximately 2 litres of water, on an average daily. However, the mean suffers from some limitations. When the data-set has one or more extreme values, the magnitude of mean is affected and it provides a wrong impression of the other values in the data-set, when used to represent the whole data-set.

8.1.2 Median

When the data-set has outliers, mean becomes flowed as a representative of the data-set. In such a case, median is used as a measure of central tendency. Median divides the data-set into two equal parts. Half of the items are less than the median and remaining half of the items are larger than the median. In order to obtain the median, we first arrange the data-set into ascending or descending order. If number of observations in the data set is n, then the

$$\text{Median} = \left(\frac{n+1}{2}\right) \text{the observation, when } n \text{ is odd;}$$

$$= \frac{1}{2}\left[\left(\frac{n}{2}\right)\text{th observation} + \left(\frac{n}{2}+1\right)\text{th observation}\right], \text{ when } n \text{ is even.}$$

For example, median of the data-set 2, 5, 6, 11, 59 is 6, while median of the data-set 5, 6, 11, 59 is $\frac{1}{2}[6+11]$ or 8.5.

Median is a positional average and is used only in the context of qualitative phenomena, for example, in estimating intelligence, etc., which are often encountered in sociological fields. Median is not useful where items need to be assigned relative importance and weights. It is not frequently used in sampling statistics.

8.1.3 Mode

The most frequently occuring observation in the data-set is mode. Mode is a french word having the meaning fashion. It is particularly useful in the study of popular sizes. For example, a manufacturer of shoes is usually interested in finding out the size most in demand so that he may manufacture a larger quantity of that size. Like median, mode is also a positional average and is not affected by extreme values. Mode is not amenable to algebraic treatment. A data-set may not have any mode or there may be more than one modes in a data-set.

8.1.4 Other Averages

Geometric mean is also useful under certain conditions. It is defined as the nth root of the product of the values of n times in a given series. Symbolically, we can put it thus:

$$\text{Geometric mean (or G.M.)} = \sqrt[n]{\pi X_i}$$

$$= \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \ldots X_n}$$

where, G.M. = Geometric mean

n = Number of items

X_i = ith value of the variable X

π = Conventional product notation

For instance, the geometric mean of the numbers, 4, 6, and 9 is worked out as

$$\text{G.M.} = \sqrt[3]{4 \cdot 6 \cdot 9}$$

$$= 6$$

The most frequently used application of this average is in the determination of average per cent of change i.e., it is often used in the preparation of index numbers or when we deal in ratios.

Harmonic mean is defined as the reciprocal of the average of reciprocals of the values of items of a series. Symbolically, we can express it as under:

$$\text{Harmonic mean (H.M.)} = \text{Rec.}\frac{\sum \text{Rec} X_i}{n}$$

$$= \text{Rec.}\frac{\text{Rec.}\, X_1 + \text{Rec.}\, X_2 + \ldots + \text{Rec.}\, X_n}{n}$$

where, H.M. = Harmonic mean

Rec. = Reciprocal

X_i = ith value of the variable X

n = Number of items.

For instance, the harmonic mean of the numbers 4, 5, and 10 is worked out as

8.2.1 Range

Range is the simplest possible measure of dispersion and is defined as the difference between the values of the extreme items of a series. Thus,

$$\text{Range} = \begin{pmatrix}\text{Highest value of an}\\ \text{item in a series}\end{pmatrix} - \begin{pmatrix}\text{Lowest value of an}\\ \text{item in a series}\end{pmatrix}$$

The utility of range is that it gives an idea of the variability very quickly, but the drawback is that range is affected very greatly by fluctuations of sampling. Its value is never stable, being based on only two values of the variable. As such, range is mostly used as a rough measure of variability and is not considered as an appropriate measure in serious research studies.

8.2.2 Mean Deviation

Mean deviation is the average of difference of the values of items from some average of the series. Such a difference is technically described as deviation. In calculating mean deviation we ignore the

8.2.3 Standard Deviation

Standard deviation is most widely used measure of dispersion of a series and is commonly denoted by the symbol ' σ ' (pronounced as sigma). Standard deviation is defined as the square-root of the average of squares of deviations, when such deviations for the values of individual items in a series are obtained from the arithmetic average. It is worked out as under:

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\Sigma (X_i - \bar{X})^2}{n}}$$

Or

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\Sigma f_i (X_i - X)^2}{\Sigma f_i}}$$

in case of frequency distribution where f_i means the frequency of the ith item.

When we divide the standard deviation by the arithmetic average of the series, the resulting quantity is known as *coefficient of standard deviation which* happens to be a relative measure and is often used for comparing with similar measure of other series. When this coefficient of standard deviation is multiplied by 100, the resulting figure is known as *coefficient of variation*. Sometimes, we work out the square of standard deviation, known as *variance*, which is frequently used in the context of analysis of variation.

The standard deviation (along with several related measures like variance, coefficient of variation, etc.) is used mostly in research studies and is regarded as a very satisfactory measure of dispersion in a series. It is amenable to mathematical manipulation because the algebraic signs are not ignored in its calculation (as we ignore in case of mean deviation). It is less affected by fluctuations of sampling. These advantages make standard deviation and its coefficient a very popular measure of the scatteredness of a series. It is popularly used in the context of estimation and testing of hypotheses.

8.3 MEASURES OF SKEWNESS

ampling Distribution

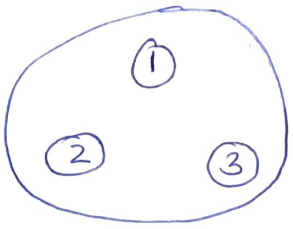
A sampling distribution is
- a probability distribution of a statistic
- obtained from a large ~~population~~ number of samples drawn from a specific population.

It is arrived out through repeated sampling from a larger population.

process: 1) collect a sample
- record the mean
- throw it back

2) collect another sample
- record the mean
- throw it back.
- and repeat the process.

e.g.



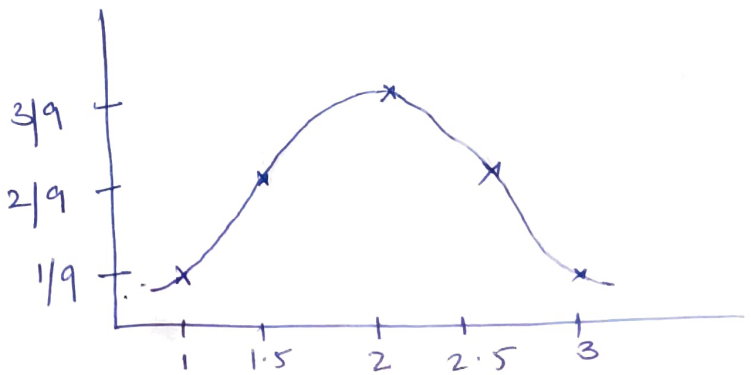
parameter ⇒ mean

$$\mu = \frac{1+2+3}{3} = 2$$

step 1) Take samples of size 2, replace ball after picking.

# picked	$\bar{x}$
1, 1	1
1, 2	1.5
1, 3	2
2, 1	1.5
2, 2	2
2, 3	2.2
3, 1	2
3, 2	2.5
3, 3	3.

step 2 plot the distribution of sample means.



ns, sampling distribution is the frequency with which you get different values for the statistic that is trying to estimate the population parameter.

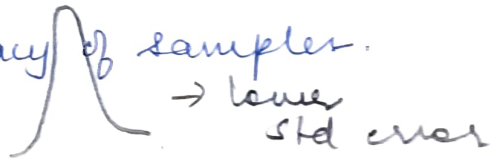
Standard Error

The standard deviation of a normal sampling distribution is called standard error.

The standard error is directly affected by

1) the number of cases in the sample
2) the variability of the population distribution.

It is used to measure the accuracy of sampler.



① Standard error of sample mean

The difference between $\bar{y}$ and $\bar{x}$ is the std error of sample mean.

a) with replacement

$$SE_{(\bar{x})} = \frac{\sigma}{\sqrt{n}}$$

b) without replacement

$$SE_{(\bar{x})} = \frac{\sigma}{\sqrt{n}} \left(\sqrt{\frac{N-n}{N-1}} \right)$$

Note :-

If σ is not available, then it can be obtained by using the formula

$$\sigma = \frac{S\sqrt{n}}{\sqrt{n-1}}$$

Standard Error of proportion

- estimate the proportion of individuals in a population with a certain characteristics.

e.g. we might want to estimate the proportion of residents in a certain city who support a new law.

Sample proportion

$$\hat{p} = \frac{x}{n}$$

→ x :- count of individuals in the sample with a certain characteristic

n :- total number of individuals in the sample.

We consider the value of $\hat{p}$ to be the best estimate for proportion of residents supporting new law.

Then if 47 of 300 residents in sample supported the new law, sample proportion $= \frac{47}{300} = 0.157$.

How do we know if this estimate will match exactly the true population proportion?

∴ we calculate standard error of proportion.

① with replacement

$$SE(\hat{p}) = \sqrt{\frac{pq}{n}} \qquad q = 1 - p$$

② without replacement

$$SE(\hat{p}) = \sqrt{\frac{pq}{n}} \left(\sqrt{\frac{N-n}{N-1}} \right)$$

.) Consider a random sample of 250 people among which 180 persons agreed on smoking.

Sol. In this problem, the number of success is 180, that is $x = 180$ and the number of total observation is 250. $n = 250$

$$p = \frac{180}{250} = 0.72$$

$$SE_{(\hat{p})} = \sqrt{\frac{0.72(1-0.72)}{250}}$$

$$= \sqrt{\frac{0.2016}{250}}$$

$$= 0.028$$

02) A random sample of 200 articles taken from a large batch of articles contain 15 defective articles. What is the estimate of standard error of the sample proportion of defective articles?

$n = 200$

$p = 15/200 = 0.075$

$q = 1 - 0.075 = 0.925$

$$SE_{(\hat{p})} = \sqrt{\frac{pq}{n}} = \sqrt{\frac{0.75 \times 0.925}{200}} = 0.0186$$

A factory produces 60000 pair of shoes on a daily basis. From a sample of 600 pairs 3% are found to be defective.
Estimate std. error of proportion.

Sol:

$N = 60000$

$n = 600$

$p = 3\% = 0.03$

$q = 1 - p = 0.97$

$$SE_{(\hat{p})} = \sqrt{\frac{pq}{n}} \left(\sqrt{\frac{N-n}{N-1}} \right)$$

$$= \sqrt{\frac{0.03 \times 0.97}{600}} \left(\sqrt{\frac{60000 - 600}{59999}} \right)$$

$$= 0.006964 \times 0.994995$$

$$= 0.0069$$

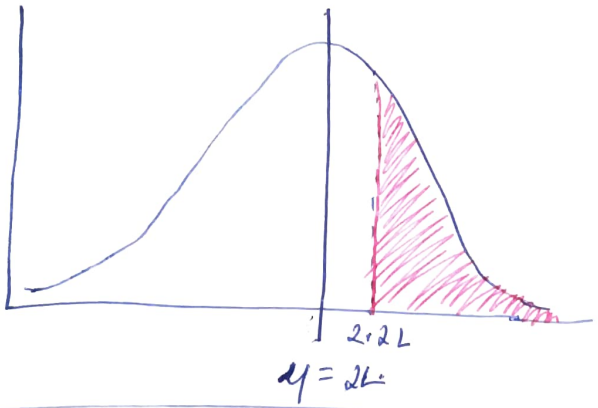
1) An average male drinks 2L of water when active outdoors (with a std dev. of 0.7L).
You are planning a full day nature trip with 50 men and will bring 110L of water.
What is the probability that you will run out of water?

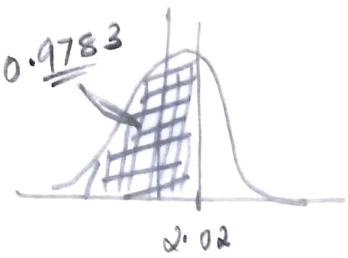
Soln.

$\left(\frac{110}{50}\right) = 2.2$

$P(\text{run out}) = P(\text{use} > 110\text{L})$

$= P(\text{avg water per man} > 2.2\text{L/man})$





Note: when you have multiple samples and want to describe the standard deviation of those sample means, you would use the z score formula :-

$$z = \frac{(x-\mu)}{(\sigma/\sqrt{n})}$$

(or when using sampling distr of means)

$$z = \frac{(2.2-2)}{(0.7/\sqrt{50})} = \frac{0.2}{0.099} = 2.02$$

$P(\bar{x} > 2.02) = 1 - 0.9783$

$= 0.0217$

Prob that you will run out of water is 2.17%

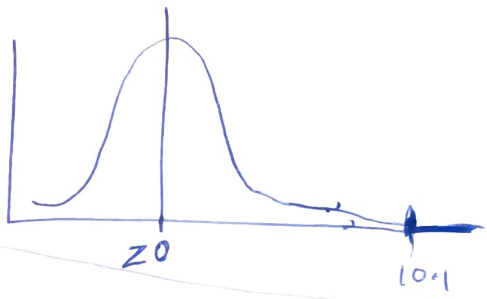
2) In general, the mean height of women is 65" with a std deviation of 3.5". What is the probability of finding a random sample of 50 women with a mean height of 70", assuming the heights are normally distributed.

Sol:

$$z = \left(\frac{x-\mu}{\sigma/\sqrt{n}}\right) = \frac{(70-65)}{(3.5/\sqrt{50})} = \frac{5}{0.495} = 10.1$$

$$P(\bar{x} \geq 10.1) = 1$$

∴ Prob of finding woman with height of 70" is 1%



Note:- 99% of values fall within 3 std dev. from the mean in normal distribut?
∴ less than 1% probability of values that fall after 3 std dev's.

0.48745 ⇒ (0-2.24)

2.24 → 0.0125

Covariance

- determines whether 2 variables are directly proportional or inversely proportional to one another.

- Covariance can have both positive and negative values. 1) Positive covariance
  2) Negative covariance.

- Positive Covariance:- both variables move in the same direction.

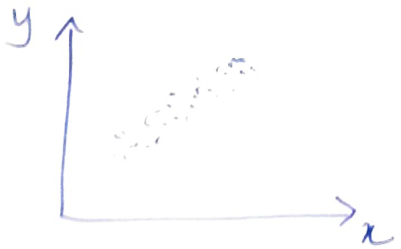
- Negative Covariance:- both the variables move in opposite direction.

- Let x and y be two variables whose relationship has to be calculated.

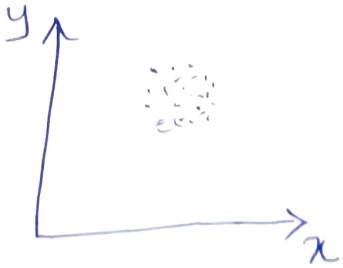
$$\boxed{Cov(x,y) = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1}}$$

x_i, y_i : data value of x and y respectively.

$\bar{x}, \bar{y}$: mean of x and y resp.



$cov(x,y) > 0$
(positive)



$Cov(x,y) \approx 0$
(no relation)

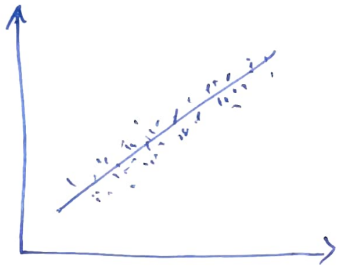
$Cov(x,y) < 0$
(negative)

Correlation Coefficient

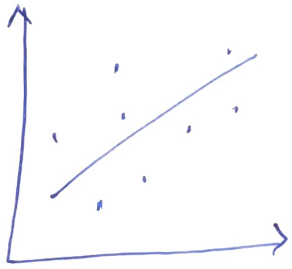
- ~~evaluation of changes between~~
- depth of relationship between variables.
- it is the estimated measure of covariance.

Correlation coefficient

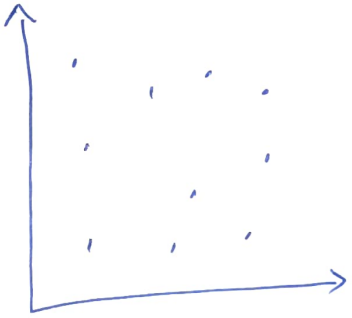
$$\sigma(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$



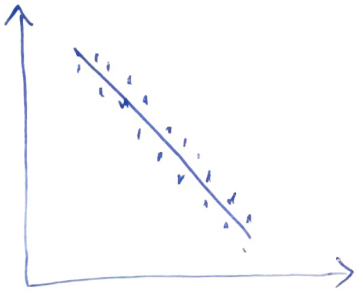
Strong +ve corr.



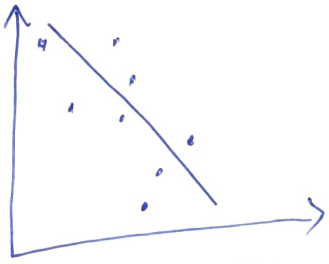
weak +ve cor.



No. Corr.



Strong -ve corr.



weak -ve corr.

Covariance	correlation
- Measure to show the extent to which given two random variables change with respect to each other. - value lies betn $-\infty$ and $+\infty$ - indicates direction of the linear relationship betn the 2 variables	- measure to describe how strongly the given 2 random variables are related to each other. value lies betn -1 and $+1$. - measures the strength of the linear relationship between the 2 variables.

Variance :- measure of spread of data around its mean value

Covariance :- measure of relatn betn two random variables

1) Calculate the coefficient of covariance for the following data: (population)

x	2	8	18	20	28	30
y	5	12	18	23	45	50

$\bar{x} = 17.67$

$\bar{y} = 25.5$