***Statistics***

Statistics is the science of information which may either be qualitative or quantitative. The price of an apartment is the quantitative variable whereas its location is a qualitative variable.

A quantitative variable can be described by a number for which arithmetic operations such as averaging makes sense. A qualitative variable simply records the quality.

**SAMPLES AND POPULATIONS**

The population consists of all the measurements in which the investigator is interested whereas a sample is a subset of measurements selected from the population and representative of the population.

A simple random sample is when the sample is chosen from the population randomly so that every possible sample of n elements has an equal chance of being selected.

**DATA AND DATA COLLECTION**

A set of measurements obtained on some variable is called a data set.

Data is collected by various methods. Sometimes our data set consists of the entire population in other cases data may constitute a sample from some population.

A conclusion drawn about a population based on the information derived from the sample of this population is called **Statistical Inference.**  To ensure the accuracy of statistical inference, data must be drawn randomly from the population of interest, and we must make sure that every segment of the population is adequately and proportionally represented in the sample.

**Non-Responsive Bias** is the biasing of results which occurs when we do not consider the fact that some people will simply not respond to any survey.

The Pth **percentile** of a group of numbers is the value below which lie P percent of the numbers in the group. The position of the Pth percentile is given by (n+1)P/100 where n is the number of data points.

1st Quartile -> It is the 25th percentile. Similarly 2nd and 3rd quartile.

**Median** is the point below which lie half the data. It is the 50th percentile.

**Interquartile** range is the Difference of 3rd and 1st quartiles.

**Measure of Central Tendency**

Three common measures of central tendency are **Mean**, **Median** and **Mode.**

Mode if the data set is the value that occurs most frequently.

Mean of a set of observations is the average. Mean of sample is referred by X bar and mean of the entire population is denoted by the Greek letter mu.

**Characteristics of the three measures of centrality**

**Mean** summarizes all the information in the data. It is the single point that can be viewed as the point where all the mass –the weight- of the observations is concentrated.

The **Median**, on the other hand, is an observation in the center of the data set. ½ the data lies above this observation and the other ½ below it.

The **mode** tells us our data set’s most frequently occurring value. There can be several modes.

If a data set or population is symmetric and if the distribution of the observations has only one mode , then the mode, the median, and the mean are all equal.

**MEASURE OF VARIABILITY**

Set I: 1,2,3,44,5,6,6,7,8,9,10,11 Mean=Median=Mode=6

Set II:4,5,5,5,6,6,6,6,7,7,7,8 Mean=Median=Mode=6

The 2 data sets have the same central tendency but different variability/dispersion.

There are several measure of variability/dispersion. One being the interquartile range. Another such measure is the range.

The **range** of a set of observations is the difference between the largest and the smallest observation.

Two other more commonly used measures of dispersion are **variance** and **standard deviation.** These 2 are more useful as compared to the range and interquartile range because they use the information contained in all the observations in the data set or the population.

The **VARIANCE** of a set of observations is the average squared deviation of the data points from their mean.

The **STANDARD** **DEVIATION** of a sample is the square root of the sample variance.

*Why would we use the S.D. when we already have its square, the variance?*

The S.D. is a more meaningful measure. The variance is the average squared deviation from the mean. It is squared because if we just compute the deviations from the mean and then average them we will get Zero. Therefore, when seeking a measure of the variance we square the observation which removes the negative signs and thus the measure is not equal to Zero. But the variance obtained is still a squared quantity. By taking the root we un-square the units and get a quantity denoted in the original units of the problem. Statisticians work with variance because its mathematical properties simplify computations whereas people applying statistics use S.D. since it is more easily interpreted.

**Grouped Data and Histogram**

We define a group of data values within specified group boundaries as a class.

When data are grouped into classes, we may also plot a frequency distribution of the data. Such a frequency plot is called a histogram.

A **histogram** is a chart made of bars of different heights. Each bar height represents the **frequency** of values in the class represented by the bar.

The **Relative Frequency** of a class is the count of data points in the class divided by total number of points.

**SKEWNESS AND KURTOSIS**

**Skewness** is a measure of the degree of asymmetry of a frequency distribution. When the distribution stretches more to the right than it does to the left then the distribution is right skewed. Similarly a left skewed distribution is on which tends to stretch more towards the left.

Zero skewness implies a symmetric distribution.

**+skew** --> right skew

**-skew** --> left skew

Two distributions that have the same mean, variance and skewness could still be significantly different in shape. We may then look at their **kurtosis.**

**Kurtosis** is the measure of peakedness of a distribution**.**

The larger the kurtosis the more peaked will be the distribution. Kurtosis is calculated and reported either as an absolute or relative value. Absolute kurtosis is always a positive number. The absolute kurtosis of the **Normal Distribution** is 3. This value 3 is taken as the datum to calculate the relative kurtosis.

Relative kurtosis can be negative and we will always work with relative kurtosis.

A negative kurtosis implies a flatter distribution than the normal distribution, and is called ***platykurtic***. A positive kurtosis means the vice-versa and is called ***Leptokurtic.***

***Relationship between the Mean and S.D.***

The mean is the measure of centrality of a set of observations whereas the S.D. is the measure of their spread. Two rules establish a relation between these two.

1. **CHEBYSHEV’S THEOREM:**

\*At least ¾ of the observations will lie within 2 S.D. of the mean

\*At least 8/9th of the observations in a set will lie within 3 S.D. of the mean.

In general the rule states that at least (1- 1/k^2) observations will lie within k S.D. of the man

1. **The EMPIRICAL RULE**

If the distribution of the data is more or less symmetric with a single mode or high point then the following tighter rules will apply.

\*Approx. 68% of the data will lie within 1 S.D. of the mean.

\*Approx. 95% of the data will lie within 2 S.D. of the mean.

\*Approx. 99% of the data will lie within 3 S.D. of the mean.

***PROBABILITY***

Probability is the Quantitative measure of Uncertainty – A number that conveys the strength of our belief in the occurrence of an uncertain event.

**Objective/Classical Probability:** is probability based on symmetry of games and chance or similar situations. It is based on the idea that some probabilities are equally likely like the throwing of a die.

**Subjective Probability:** It involves personal judgment, information ,intuition and other subjective evaluation criteria.

A **Set** is a collection of elements. The empty set is a set containing no elements and is denoted by Ø

**Universal** **Set** is a set containing everything in the given context and Is denoted by S.

An **experiment** is a process that leads to one of the several possible outcomes.

An **outcome** of an experiment is some observation or measurement.

The **sample space** is the universal set S pertinent to the given experiment which is the set of all possible outcomes of an experiment.

An **event** is a subset of the sample space which is a set of basic outcomes.

**Mutually Exclusive events**

When the sets corresponding to 2 events are disjoint (i.e. have no intersection), the 2 events are called mutually exclusive.

For mutually exclusive events:-

**Conditional Probability**

We may define the probability of event A conditional upon the occurrence of Event B.

The C.P. of event A given the Event B is

Assuming P(B)≠0

**Independence of Events**

Two events are said to be independent of each other if and only if the following three conditions are met:-

P(A|B)=P(A)

P(B|A)=P(B)

And, most important

P(A=P(A)P(B)

The 3rd equation tells us that when A and B are independent of each other then we can obtain the probability of the joint occurrence of A and B simply by multiplying both the probabilities together. This rule is thus called the **Product rule** of independent events.

**Product Rules for Independent Events**

The rules for the union and the intersection of 2 independent events extend nicely to sequences of more than 2 events and are quite useful in random sampling.

*Random sampling from a large population implies independence.*

**Union Rule:** P(A1

**Intersection Rule:**

***Note:-*** When 2 events are mutually exclusive, they are not independent. In fact they are dependent events in the sense that if one happens the other cannot happen. The probability of the intersection of 2 mutually exclusive events is 0 whereas the probability of the intersection of two independent events is not 0, it is equal to the product of the probabilities of separate events.

**Baye’s Theorem**

**The law of total probability**

From the law of total probability using conditional probability, we get a new equation:-

Now P(B|A)=P(A|B)P(B)/P(A)

Replacing the value of P(A) from (1)

….. **Baye’s Theorem**

***RANDOM VARIABLE***

A random variable is an uncertain quantity whose value depends on chance.

A random variable has a probability law – a rule that assigns probabilities to the different values of the random variable. The probability law, the probability assignment, is called the probability distribution of the random variable.

Random variable(X) and Probability Distribution P(X)

**Sample space** **Random Variable(X=number of girls)**

BBBB X=0

GBBB

BGBB X=1

BBGB  
BBBG

GGBB  
GBGB  
GBBG X=2  
BGGB  
BGBG  
BBGG  
BGGG  
GBGG  
GGBG  
GGGB X=3  
GGGG X=4

Hence A **random** **variable** is a function of the **sample** **space**.

P(X=0) = 1/16=0.0625

P(X=1) = 4/16=0.25

P(X=2) = 6/16=0.375

P(X=3) = 4/16=0.25

P(X=4) = 1/16=0.375

**Probability Bar chart**

See From Business statistics Book(Pg No. 95)

***Discrete and Continous Random Variable***

A **discrete** **random** variable can assume at most a countable number of values whereas a **continuous random variable** may take on any value in an interval of numbers.

The probability distribution of a discrete random variable X must satisfy the following 2 conditions.

* P(X)>= for all values of X
* ∑P(X) = 1

**Cumulative Distribution Function,** F(X), of a discrete random variable X is

**EXPECTED VALUES OF DISCRETE RANDOM VARIABLES**

The mean of a probability distribution of a random variable is a measure of the centrality of the probability distribution. It is a measure that considers both the values of the random variable and the probability. The mean is a weighted average of the possible values of the random variable – the weights being the probabilities.

This mean of the probability distribution of the random variable is called the **Expected Value** of the R.V. We use both µ and E(X) as the symbol for expected value.

|  |  |  |
| --- | --- | --- |
| X | P(X) | xP(X) |
| 0 | .1 | 0 |
| 1 | .2 | .2 |
| 2 | .3 | .6 |
| 3 | .2 | .6 |
| 4 | .1 | .4 |
| 5 | .1 | .5 |
|  | 1.00 | 2.3🡨Mean, E(X) |

The Expected value of H(X), a function of the random variable X, is

The function H(X) could be X^2, X^3, log X etcetera.

**VARIANCE AND STANDARD DEVIATION OF A RANDOM VARIABLE**

**BERNOULLI RANDOM VARIABLE(BRV)**

If the outcome of a trial can only be either success or failure, then the trial is a Bernoulli trial. The number of successes X in one Bernoulli trial, which can be 1 or 0 , is a **BERNOULLI RANDOM VARIABLE.**

Bernoulli random variable is too simple to be of practical use, hence we use Binomial random variable whose basic building block is the BRV.

**BINOMIAL RANDOM VARIABLE**

Conditions to be met for Binomial Random Variable

* The trials must be Bernoulli trials in that the outcomes can only be either success or failure.
* The outcomes of the trials must be independent.
* The probability of success in each trial must be a **CONSTANT**.

If X ~ B(n,p), then

E(X) =np V(X) =np (1-p)

**NEGATIVE BINOMIAL DISTRIBUTION**

If we need 2 pins and the probability of making a pin successfully is 0.6 then we would stop as soon as we get 2 good pins. It may be produced in the 1st 2 attempts or 3,4,5 attempts. The number of trials could be 2, 3, 4, 5…..etc. (In contrast in binomial distribution the number of trials is fixed)

The number of trials in this scenario is said to follow **Negative Binomial Distribution.** Let s denote the number of successes and p the probability of success in each trial. Let X denotes the number of trials made until the desire number of Successes is achieved.

If X~NB(s,p), then

Where, x=s,s+1,s+2…..

E(X)=S/p V(X)=s(1-p)/p^2

**THE GEOMETRIC DISTRIBUTION**

The geometric distribution is a special case of the negative binomial distribution where s=1.

If X~G(p), then

E(X)=1/p V(X)=(1-p)/p^2

**THE POISSON DISTRIBUTION**

As long as the expected value µ=np is neither too large nor too small i.e. between 0.01 and 50, the binomial formula for P(X=x) can be approximated as,

Where x=0, 1, 2, 3…..

E(X)=np

V(X)=np(1-p) but since p is very small 1-p~1 hence, V(X)= np

If we count the number of times a rare event will occur during a fixed interval, then the number would follow a Poisson distribution.

**CONTINOUS RANDOM VARIABLE**

A continuous random variable is a random variable that can take on any value in an interval of numbers.

The probabilities associated with a continuous random variable X are determined by the probability distribution function of the R.V. The function f(x) has the following properties.

* F(x)>=0 for all x
* The probability that X will be between two number A and B is equal to the area under f(x) between A and B.
* The total area under f(x) =1.0

**THE EXPONENTIAL DISTRIBUTION(a memory-less process)**

Some examples of exponential distribution:

* The time between 2 successive breakdowns of a machine will be exponentially distributed. The mean in this case is known as **MTBF(mean time between failures)**
* The time gap between 2 successive arrivals of a waiting line, known as the inter-arrival time, will be exponentially distributed.

x>=0

**THE NORMAL DISTRIBUTION**

The normal distribution is an important continuous distribution because a good number of random variables occurring in practice can be approximated to it. *If a random variable is affected by many independent causes, and the effect of each cause is not overwhelmingly large compared to the other effects, then the random variable will closely follow a normal distribution.*

For a normal distribution with mean µ and S.D. the probability density function f(x) is given by the formula:

If mean is 100 and variance Is 9 then

**Properties of the normal distribution**

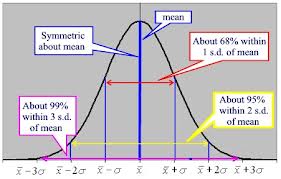
*If several independent random variables are normally distributed, then their sum will also be normally distributed. The mean of the sum will be the sum of the individual means, and by virtue of independence, the variance of the sum will be the sum of all the individual variances.*

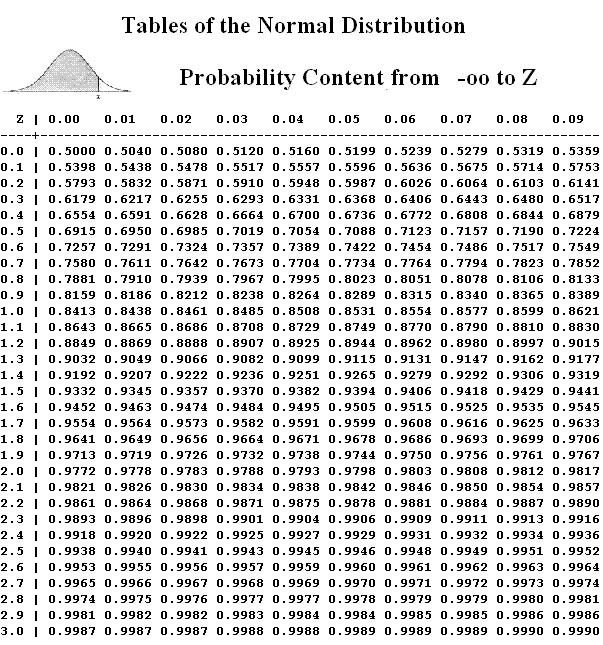
Note that the variances can be added but not the standard deviations. We can never add the S.D.’s. Calculate the combined variance and then square root it to get the combined S.D.

Another interesting property of the normal distribution is that if X is normally distributed then aX+b will also be normally distributed with **mean** aE(X)+b and **variance**

**The standard Normal distribution**

We define the standard normal distribution variable Z as the normal random variable with mean=0 and S.D.=1



**

The transformation of X to Z:

***SAMPLING AND SAMPLING DISTRIBUTIONS***

Statistics is a science of inference, of generalization from a part (the random sample) to the whole (the population). Hence the random sample of n elements should be a sample selected from the population in such a way that every set of n elements is likely to be selected as any other set of n elements so that prediction based on the sample should also be true for the population.

A numerical measure of the population like mean and S.D. is called a **population** **parameter**, or simply a **parameter**.

A numerical measure of the sample is called a **sample statistic**, or simply **statistic.**

**Population parameters** are estimated by **sample statistics**, when a sample statistic is used to estimate a population parameter, the statistic is called the **estimator** of the parameter.

Like sample mean x bar is the sample statistic used as an estimator of the population mean *mu.*

The **population proportion P** is equal to the number of elements in the population belonging to the category of interest, divided by the total number of elements in the population.

|  |  |  |
| --- | --- | --- |
| Estimator |  | Population Parameter |
| X Bar | Estimates | µ |
|  | Estimates |  |
| P(carrot) | Estimates | P |

**Types of Sampling**

There are 2 methods of selecting samples from populations. *Nonrandom or Judgment Sampling , and random or probability sampling.* In judgment sampling, personal knowledge and opinion is used to identify the items from the population whereas in random sampling all the items in the population have a chance of being chosen.

**RANDOM SAMPLING**

1. **Simple Random Sampling:** It Selects samples by methods that allow each possible sample to have an equal probability of being picked and each item in the entire population to have an equal chance of being included in the sample.
2. **Systematic Sampling:** Elements are selected from the population at a uniform interval that is measured in time, order or space. If we want to interview every 20th student in a college campus we would choose a random starting point in the 1st 20 names and then pick every 20th name after that. Systematic sampling differs from simple random sampling in that each element has an equal chance of being selected but each sample does not have an equal chance of being selected. Even though systematic sampling maybe inappropriate when the elements lie in a sequential pattern, this method may require less times and sometimes results in lower costs than simple random-sample method.
3. **Stratified Sampling:** Here we divide the population into relatively homogenous groups called strata. Stratified sampling is appropriate when the population is already divided into groups of different sizes and we wish to acknowledge this fact.
4. **Cluster Sampling:** Here we divide the population into groups, or clusters, and then select a random sample of these clusters. We assume that these individual clusters are representative of the entire population as a whole. A well designed cluster sampling procedure can produce a more precise sample at a considerably less cost than that of simple random sampling.

**NOTE:** With both stratified and cluster sampling, the population is divided into well-defined groups. We use stratified sampling when each group has a small variation within itself but there is a wide variation between these groups. We use cluster sampling when these is considerable variation within groups but the groups are essentially similar to each other.

**NON RANDOM SAMPLING**

Non random sampling designs do not provide unit in the population a known chance of being selected in the sample. The selection procedure is partially subjective.

1. **Convenience Sampling:** Based on the convenience of the researcher who selects the sample most convenient to him/her.
2. **Judgment Sampling:** Researcher excises his/her judgment to draw a sample which he/she might think is representative of the population.
3. **Quota Sampling:** It consists of fixation of certain quotas on the basis of certain parameters so as to make the sample representative of the entire population.
4. **Shopping Mall Intercept Sampling:** It involves drawing samples in market places, malls fairs in different socioeconomic locations.
5. **Snowball Sampling:** Initial respondents are selected randomly then additional respondents are selected by their referral and so on. Also known as **Multiplicity Sampling.**

**SAMPLING DISTRIBUTIONS**

A probability distribution of all the possible means of the samples is a distribution of the sample means. Statisticians call this a *Sampling distribution of the mean.*

***Standard Error of the mean*** is the standard deviation of the distribution of sample means. Similarly the standard deviation of the distribution of sample proportions is shortened to the ***standard error of the proportion.***

Suppose we wish to learn about the height of freshmen at a university. We could take a series of samples and calculate the mean height for each sample. It is highly unlikely for all the mean heights to be same, we expect to see some variability in our means. This variability results from *sampling error* due to chance i.e. there are differences between each sample and the population.

The **standard deviation of the distribution of sample means** measures the extent to which we expect the means from the different samples to vary because of this chance error in the sampling process. Thus, **the standard deviation of the distribution of a sample statistic is known as the *standard error of the statistic.***

The Standard error indicates not only the size of the chance error, but also the accuracy we are likely to get if we use a sample statistic to estimate a population parameter. A distribution of sample means that is less spread(has a small standard error) is a better estimator of the population mean than a distribution of sample means that is widely dispersed and have a large standard error.

|  |  |
| --- | --- |
| **When we wish to refer to the** | **We use the conventional term** |
| Standard deviation of the distribution of the sample means | Standard error of the mean |
| Standard deviation of the dist. of the sample proportions | Standard error of the proportion |
| Standard deviation of the distribution of the sample medians | Standard error of the median |
| Standard deviation of the distribution of the sample ranges | Standard error of the range |

**Deriving the Sampling Distribution of the Mean**

µ

The Population Distribution:

µ is the mean of this distribution

Sigma is the standard deviation

If somehow we were able to take all the possible samples of a given size from this population distribution, they would be graphically represented by the below 4 samples.

X bar X bar X bar X bar

Above is the sample frequency distribution which represents the enormous number of sample distributions possible. Each sample distribution is a discrete distribution and has its own mean called x bar and its own standard deviation(s)

Now if we were able to take the means from all these sample distributions and produce a distribution of these sample means, it would look like this:

Above is the sampling distribution of the mean which is the distribution of all the sample means and has:

Mean of the sampling distribution of the means called ‘mu sub x bar’

Standard error of the mean called ‘sigma sub x bar’

If we increase our sample size from 5 to 20, it would not change the standard deviation of the items in the original population. But with samples of size 20 we would have increased the effect of averaging in each sample and would expect even less dispersion among the sample means.

**The sampling distribution has a mean equal to the population mean:**

*Mu sub x bar() = µ*

**The sampling distribution has a S.D.(standard error) equal to the population S.D. divided by the square root of the sample size:**

*Sigma sub x bar =*

***EXAMPLE:***A bank calculates that its individual savings accounts are normally distributed with a mean of 2000$ and a S.D. of 600$. If the bank takes a random sample of 100 accounts, what is the probability that the sample mean will lie between $1900 and $2200?

**Solution:**

*=*

*=600/10*

*=60*

***Standardizing of the sample mean:***

For x bar=1900$: Z=(1900-2000)/60=-1.67

For x bar=2200$: Z=(2200-2000)/60=+0.83

From table we get the combined area as 0.7492 which would be the probability that the sample mean will lie between 1900 and 2200$.

**SAMPLING FROM NON NORMAL POPULATIONS**

***CENTRAL LIMIT THEOREM***

* The mean of the sampling distribution of the mean will be equal to the population mean regardless of the sample size, even if the population is not normal.
* As the sample size increases, the sampling distribution of the mean will approach normality, regardless of the shape of the population distribution.
* CLT assures us that the sampling distribution of the mean approaches normal as the sample size increases. Statisticians use the normal distribution as an approximation to the sampling distribution whenever the **sample size is at least 30.**
* The Significance of the CLT is that it permits us to use sample statistics to make inferences about the population parameters without knowing anything about the shape of the frequency distribution of that population other than what we can get from the sample.

**AN OPERATIONAL CONSIDERATION IN SAMPLING:** THE RELATIONSHIP BETWEEN SAMPLE SIZE AND ERROR

*()*  is a measure of dispersion of the sample means around the population mean. If the dispersion decreases, then the value taken by the sample mean tends to cluster more closely around µ. Conversely, if the dispersion increases, then the value tends to cluster less closely around µ. Hence, **as the standard error decreases, the value of any sample mean will probably be closer to the value of the population mean.** As the standard error decreases, the precision with which the sample mean can be used to estimate the population mean increases.

*As n increases, decreases.* Hence if we increase the sample size from 10 to 100(10 folds drop), the S.E. drops from 31.63 to 10(1/3rd drop). Because varies inversely with the square root of n, there is **Diminishing return in sampling. Although sampling more items will decrease the S.E. but the increased precision is not worth the additional cost of sampling.** In statistical sense, it seldom pays to take excessive large samples. Statisticians should focus on the concept of right sample size.

**THE FINITE POPULATION MULTIPLIER**

Till now we were assuming an infinite population or in which we sample from a finite population with replacement.

**S.E. of the mean for finite population:**

N: Size of the population

n: Size of the sample

Finite population multiplier:

Note: It is the absolute size of the sample that determines sampling precision, not the fraction of the population sampled.

***ESTIMATION***

This chapter introduces methods that enable us to estimate with reasonable accuracy the population proportion (the proportion of the population that possesses a given characteristic) and the population mean.

**Types of Estimate:**

* Point Estimate: A point estimate is a single number that is used to estimate an unknown population parameter.
* Interval Estimate: A range of values used to estimate a population parameter. It indicates the error in two ways – by the extent of its range and by the probability of the true population parameter lying within that range.

**Estimator and Estimates**

An **estimator** is a sample statistic used to estimate a population parameter. The sample mean x bar can be an estimator of the population mean µ.

When we have observed a specific numerical value of our estimator, we call that value an **estimate.** An estimate is a specific observed value of a statistic.

Criteria of a good Estimator:

* **Un-Biasedness:** A statistic is an unbiased estimator if, on average, it tends to assume values that are above the population parameter being estimated as frequently and to the same extent as it tend to assume values that are below the population parameter being estimated.
* **Efficiency:** Efficiency refers to the size of the standard error of the statistic. If we compare 2 statistics from a sample of the same size and try to decide which one is the more efficient estimator, we would pick up the statistic which has the smaller standard of error, or S.D. of the sampling distribution
* **Consistency:** A statistic is a consistent estimator of a population parameter if as the sample size increases; it becomes almost certain that the value of the statistic comes very close to the value of the population parameter.
* **Sufficiency:** An estimator is sufficient if it makes use of the information in the sample that no other estimator could extract from the sample additional information about the population parameter being estimated.

**POINT ESTIMATES**

The sample mean x bar is the best estimator of the population mean µ. It is unbiased, consistent, the most efficient estimator, and, as long as the sample is sufficiently large, its sampling distribution can be approximated by the normal distribution. If we know x bar, we can make statements about any estimate we make from sampling information.

**POINT ESTIMATE OF THE POPULATION VARIANCE AND THE STANDARD DEVIATION**

The most frequently used estimator of the population standard deviation sigma is the sample standard deviation ‘s’(refer Levin and Rubin book to see manual method of calculation. Pg333).

The reason we study estimators is so we can learn about the populations by sampling, without counting every item in the population.

**INTERVAL ESTIMATES:BASIC CONCEPTS**

To find the range we need to find the standard error of the mean by applying CLT and then the range can be calculated.

Our sample of 200 users has a mean battery life of 36 months. The director wants to know about the range within which the battery life will lie. S.D.=10.

Sigma sub x bar=sigma/under root of n

=10/under root (200)

=10/14.14=.707 month

Range = 36-.707 to 36+.707= 35.293 to 36.707(36+/-1 sigma sub x bar)

Hence we are 68.3% sure that it will lie within the above range

Range=36+/- 2S.D.=34.586-37.414= we are 95.5% confident that it will lie within this range.

In statistics, the probability that we associate with an interval estimate is called the **CONFIDENCE LEVEL**.

From above example:-

**CONFIDENCE INTERVAL:-**35.293 to 36.707

**CONFIDENCE LEVEL:-**68.3%

High confidence levels produce large confidence intervals which are not precise and give fuzzy estimates.

Note: Suppose we calculate from one sample the following confidence interval and level.

“We are 95% sure that the mean battery life of the population lies within 30 to 42 months”

*“****This statement does not mean that the chance is .95 that the mean life will fall within the interval establishes in one sample. Instead, it means that if we select many random samples of the same size and calculate a confidence interval for each of these samples, then in about 95.5% of the cases, the population mean will lie within that interval.***”

**INTERVAL ESTIMATES USING THE T-DISTRIBUTION**

Use of the T-distribution is required whenever the sample size is less than 30 and the population S.D. is not known. Furthermore in using the T-distribution we assume that the population is normal or approximately normal.

T-distribution is flatter than the normal distribution, and there is a different T-distribution for every possible sample size. As the sample gets larger, the shape of the T-distribution loses its flatness and becomes approximately equal to the normal distribution. A **t-distribution is lower at the mean and higher at the tails than a normal distribution**.

**Degree of freedom = (Sample Size – 1)**

**TESTING HYPOTHESIS: ONE SAMPLE TESTS**

Hypothesis testing begins with an assumption, called a hypothesis that we make about a population parameter. Then we collect sample data, produce sample statistics, and use this information to decide how likely it is that our hypothesized population parameter is correct.

Unfortunately, the difference between the hypothesized population parameter and the actual statistic is more often neither so large that we automatically reject our hypothesis nor so small that we just as quickly accept it.

Example: If the Hypothesized population mean is 0.04inch and the population S.D. is .004 inch, what are the chances of getting a sample mean (0.0408) that differs from 0.04 inch by 0.0008 inch?

To determine whether the population mean is actually 0.04 inch, we must calculate the probability that a random sample with a mean of 0.0408 inch will be selected from a population with a µ of 0.04 inch and a sigma of 0.004 inch. The probability will indicate whether it is reasonable to observe a sample like this if the population mean is actually 0.04 inch.

Calculate S.E. of mean:

Sigma sub x bar =

=0.004 in./ = 0.0004 inch.

Now we find Z value to discover how far is the mean of our sample

Z= (x bar – mu)/sigma sub x bar

= (0.0408-0.04/0.0004 = 2 standard errors of the mean.

Use the table to find out that 4.5 % is the total chance that our sample mean would be differing from the population mean by 2 standard errors. With this low a chance we cannot conclude that a population with a true mean of 0.4 would be likely to produce a sample like this.

The minimum standard for an acceptable probability, 4.5 %, is also the risk we take of rejecting a hypothesis that could have been true; there can be no risk free trade off.

**TESTING HYPOTHESIS**

In hypothesis testing, we must state the assumed or hypothesized value of the population parameter before we begin sampling. The assumption we wish to test is called the *null hypothesis* and is symbolized by or “H sub-zero”.

Whenever we reject the *null hypothesis*, the conclusion we do accept is called the *alternative hypothesis* and is symbolized by or “H sub-one”.

If we use a hypothesized value of a population mean in a problem, we would represent it as:

We will consider 3 possible alternative hypotheses:



The purpose of hypothesis testing is not to question the computed value of the sample statistic but to make a judgment about the difference between the sample statistic and a hypothesized population parameter.

If we test a hypothesis at 5% *significance level*, then it means that we will reject the null hypothesis if the difference between the sample statistic and the hypothesized population parameter is so large that it or a larger difference would occur, on the average only 5 or fewer times out of 100 samples. If we assume the hypothesis is correct, then the *significance level* will indicate the % of sample means outside certain limits (*Confidence level)*

**Whenever we say that we accept the null hypothesis, we actually mean that there is not sufficient statistical evidence to reject it. Use of the term ‘accept’, instead of ‘do not reject’, has become a standard. It means simply that when sample data do not cause us to reject a null hypothesis, we behave as if the hypothesis is true.**

**Note: The higher the significance level, the higher the probability of rejecting a null hypothesis.**

**TYPE I AND TYPE II ERRORS**

Type I Error (α): rejecting a null hypothesis when it’s true

Type II Error (β): accepting a null hypothesis when it’s false

The probability of making one type of error can be reduced only if we are willing to increase the probability of making the other type of error.

**DECIDING WHICH DISTRIBUTION TO USE IN HYPOTHESIS TESTING**

|  |  |  |
| --- | --- | --- |
|  | When Population S.D. is known | When Population S.D. is not known |
| Sample size n is greater than 30 | Normal Distribution  Z table | Normal Distribution  Z table |
| Sample size n is less than 30 and we assume the population is normal or approximately so | Normal Distribution  Z table | t distribution  t table |

As in estimation, use the finite population multiplies whenever the population is finite in size, sampling is done without replacement, and the sample is more than 5% of the population.

**TWO TAILED AND ONE TAILED TEST OF HYPOTHESES**

A *two-tailed test of hypothesis will reject the null hypothesis if the sample is significantly higher than or lower than the hypothesized population mean.* Thus in a two-tailed test there are two rejection regions.

However there are situations in which a two-tailed test is not appropriate and we must use a *one-tailed test*.

**Left tailed test-**

**Right tailed test-**

**Finally it should be reminded again that in each example of hypothesis testing when we accept a null hypothesis on the basis of sample information, we are really saying that there is no statistical evidence to reject it. We are not saying that the null hypothesis is true.** The only way to prove a null hypothesis is to know the population parameter, and that is not possible with sampling. Thus, we accept the null hypothesis and behave as if it’s true simply because we can find no way to reject it.

**SUMMARY OF THE FIVE STEP PROCESS OF HYPOTHESIS TESTING**

|  |  |
| --- | --- |
| STEP | ACTION |
| 1 | Decide whether this is a two-tailed test or a one tailed test. State your hypothesis. Select a level of significance appropriate for this decision |
| 2 | Decide which distribution (t or z) is appropriate and find the critical value(s) for the chosen level of significance. |
| 3 | Calculate the *standard error of the sample statistic*. Use the standard error to convert the observed value of the sample statistic to the standardized value. |
| 4 | Sketch the distribution and mark the position of the standardized sample value and the critical value(s) for the test |
| 5 | Compare the value of the standardized sample statistic with the critical value(s) for this test and interpret the result. |

**NOTE: FOR MANUAL EXAMPLES GO TO PAGE 399 OF LEVIN AND RUBIN.**

**MEASURING THE POWER OF A HYPOTHESIS TEST**

Because rejecting a null hypothesis when it is false is exactly what a good test should do, a high value of (something near 1.0) means that the test is working fine; a low value of means that the test is working very poorly.  **Because the value of**  **is the measure of how well the test is working, it is known as the *power of the test.***

**NOTE: FOR MANUAL EXPLANATION OF T AND Z TESTS REFER TO THE BOOK**

**If the sample size is 30 or less and sigma is not known, we should use the t-distribution with n-1 degrees of freedom.**

**TESTING HYPOTHESIS: TWO SAMPLE TESTS**

Do female employees earn less than male employees? Did one group of experimental animals react differently from that at another?   
In each of these examples, decision makers are concerned with the parameters of two populations. In these situations, they are not as interested in the actual value of the parameters as they are in the relation between the values of the 2 parameters i.e. how these parameters differ.

Because we now wish to study 2 populations, not just one, the sampling distribution of interest is the *sampling distribution of the difference between sample means.*

Suppose population1 and population2 has standard deviations and means. Now we take two samples from population 1 and 2 and then we subtract the sample mean 🡨 *Difference between sample means.*

The difference will be either positive or negative. By constructing a distribution of all possible sample differences we end up with the sampling distribution of the difference between sample means.

*The mean of the sampling distribution of the difference between sample means* is symbolized by and is equal to and is the same as

(-). The standard deviation of the distribution of the difference between the sample means is called the *standard error of the difference between the 2 means* and is calculated by using the below formula:

**NOTE: SEE EXAMPLES FROM LEVIN AND RUBIN BOOK (Pg No:445)**

**TESTING DIFFERENCES BETWEEN MEANS WITH DEPENDENT SAMPLES**

Earlier our samples were independent of each other. Sometimes, however it makes sense to take samples that are not independent of each other. Often the use of such dependent (or paired) samples enables us to perform a more precise analysis, because these allow us to control for extraneous factors. With dependent samples, we still follow the same basic procedure that we have followed in all our hypothesis testing. The only differences are that we will use a different formula for the estimated standard error of the sample differences and that we will require that both samples be of the same size.

Ex: A health spa has a program which guarantees more than 17 kg of weight loss. The spa provides records of 10 participants with before and after weight. **So conceptually what we have is not 2 samples of before and after weights, rather one sample of weight losses.**

**CHI-SQUARE AND ANOVA**

Suppose we have proportions from 5 populations instead of only two. In this case, the methods for comparing proportions described earlier do not apply. **Chi**-**square** tests enable us to test whether more than 2 population proportions can be considered equal. Actually it allows us to do a lot more than just test for the equality of several proportions. If we classify a proportion into several categories with respect to two attributes, we can use a chi-square test to determine whether the 2 attributes are independent of each other.

The **ANOVA** enables us to test whether more than 2 population means van be considered equal.

**CHI SQUARE TEST AS A TEST OF INDEPENDENCE**

Many times, managers need to know whether the differences they observe among several sample proportions are significant or only due to chance.

**CONTINGENCY TABLES**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **North-East** | **South-East** | **Central** | **West coast** | **Total** |
| **Number who prefer present method** | **68** | **75** | **57** | **79** | **279** |
| **Number who prefer new method** | **32** | **45** | **33** | **31** | **141** |
| **Total employees sampled in each region** | **100** | **120** | **90** | **110** | **420** |

Suppose that in the four regions, national healthcare company samples its employees attitude towards job performance reviews. The above table which illustrates the response is called a **Contingency table.**  Notice that the 4 columns provide one basis of information – geographical region and the 2 rows classify the information other way: preference for review methods. The above table is called a 2X4 contingency table because it contains 2 rows and 4 columns.

**OBSERVED AND EXPECTED FREQUENCIES:**

Suppose we now symbolize the true proportions of the total population of employees who prefer the present plan as

* : Proportion in north-east who prefer present plan
* : Proportion in south-east who prefer present plan
* : Proportion in central region who prefer present plan
* : Proportion in west coast who prefer present plan

Using these symbols we can calculate the null and alternate hypothesis as follows:

If the null hypothesis is true then we can combine the data from the 4 samples and then estimate the proportion of the total workforce that prefers the present method:

Combined proportion = (68+75+57+79)/(100+120+90+110)

= 279/420=0.6643

**PROPORTION OD SAMPLED EMPLOYEES IN EACH REGION EXPECTED TO PREFER THE 2 REVIEW METHODS:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **North-east** | **South-east** | **Central** | **West-coast** |
| **Total number sampled estimated proportion who prefer present method** | 100  X0.6643 | 120  X0.6643 | 90  X0.6643 | 110  X0.6643 |
| **Number expected to prefer present method** | **66.43** | **79.72** | **59.79** | **73.07** |
| **Total number sampled estimated proportion who prefer new method** | 100  X0.3357 | 120  X0.3357 | 90  X0.3357 | 110  X0.3357 |
| **Number expected to prefer new method** | **33.57** | **40.28** | **30.21** | **36.93** |

COMPARISON OF EXPECTED AND OBSERVED FREQUENCIES OF SAMPLED EMPLOYEES.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | North-East | South-East | Central | West-Coast |
| FREWUENCY PREFERRING PRESENT METHOD |  |  |  |  |
| Observed(actual) frequency | 68 | 75 | 57 | 79 |
| Expected(theoretical) frequency | 66.43 | 79.72 | 59.79 | 73.07 |
| FREQUENCY PREFERRING NEW METHOD |  |  |  |  |
| Observed(actual) frequency | 32 | 45 | 33 | 31 |
| Expected(theoretical) frequency | 33.57 | 40.28 | 30.21 | 36.93 |

**THE CHI-SQUARE STATISTIC**

Fo - Observed Frequency

Fe – Expected Frequency

Now when we calculate our value for chi-square (Levin and Rubin book, page no. 536) we get the value 2.764. If this value was as large as 20, it would indicate a substantial difference between our observed and expected values. A chi-square of 0, on the other hand, indicates that the observed frequencies exactly match the expected frequencies. The value of chi-square can never be negative because the differences between the expected and observed values are always squared.

**THE CHI-SQUARE DISTRIBUTION**

If the null hypothesis is true, then the sampling distribution of the chi-square statistic can be closely approximated by a continuous curve known as a chi-square distribution. As in the case of the t-distribution, there is a different chi square distribution for each different number of degree of freedom.

The chi-square distribution is a probability distribution, hence the total area under each curve of chi-square is 1

DETERMINING THE DEGREES OF FREEDOM:

Number of degrees of freedom = (number of rows-1)(number of columns-1)

**USING THE CHI-SQUARE TEST**

Returning to our example

α = 0.10 and DOF = (2-1)X(4-1) = 3

From table we get the value of chi–square as 6.251(using DOF and α)

**Sample chi-square value=2.764**

**Acceptable Chi-square value=6.251**

Therefore, we accept the null hypothesis that there is no difference between the attitudes about job interviews.

**CONTINGENCY TABLES WITH MORE THAN 2 ROWS**

RT – Row total for the row containing that cell

CT – Column total for the column containing that cell

N – Total number of observations

– Expected frequency in a given cell

NOTE: To use a chi-square hypothesis test, we must have a sample size large enough to guarantee the similarity between the theoretically correct distribution and our sampling distribution of , the chi square statistic. When the expected frequencies are too small, the value of chi-square will result in too many rejections of the null hypothesis. **To avoid making incorrect inferences from chi-square hypothesis test, follow the general rule that an expected frequency of less than 5 in one cell of a contingency table is too small.**

**CHI-SQUARE AS A TEST OF GOODNESS OF FIT: TESTING THE APPROPRIATENESS OF A DISTRIBUTION**

The chi-square test can also be used to decide whether a particular probability distribution, such as the binomial, Poisson or normal is the appropriate distribution. This is an important ability since as decision makers using statistic, we will need to choose a probability distribution to represent the distribution of the data. **The chi-square test enables us to ask this question and to test whether there is a significant difference between an observed frequency distribution and the theoretical frequency distribution.**

**CALCULATING OBSERVED AND EXPECTED FREQUENCIES**

Suppose that the Gordon company requires that college seniors who are seeking positions with it to be interviewed by 3 different executives. This enables the company to obtain a consensus evaluation of each candidate. Each executive gives the candidate either a positive or a negative rating.

|  |  |
| --- | --- |
| Possible +ve ratings from the 3 interviews | Number of candidates receiving these ratings |
| 0 | 18 |
| 1 | 47 |
| 2 | 24 |
| 3 | 11 |
| Total-> | 100 |

For staffing purpose a director of recruitment thinks that the interview process can be approximated by a binomial distribution with p = 0.40 i.e. with a 40% chance of any candidate receiving a positive rating on any one interview. If the director wants to test this hypothesis at 0.20 level of significance, how should he proceed?

α = 0.20

To solve this, we must determine whether the discrepancies between the observed and expected frequencies should be ascribed to chance. We begin by determining the binomial probabilities for this interview situation.

|  |  |
| --- | --- |
| Possible +ve ratings from the 3 interviews | Binomial probabilities of these outcomes(p=0.40) |
| 0 | 0.2160 |
| 1 | 0.4320 |
| 2 | 0.2880 |
| 3 | 0.0640 |
| Total🡪 | 1.0000 |

OBSERVED FREQUENCIES, APPROPRIATE BINOMIAL PROBABILITIES, AND EXPECTED FREQUENCIES FOR THE PROBLEM

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Possible +ve ratings | Observed freq of candidates receiving the ratings | Binomial probabilities of possible outcomes | Number of candidates interviewed | Expected freq of candidates receiving the ratings |
| 0 | 18 | 0.2160 | X 100 | 21.6 |
| 1 | 47 | 0.4320 | X 100 | 43.2 |
| 2 | 24 | 0.2880 | X 100 | 28.8 |
| 3 | 11 | 0.0640 | X 100 | 6.4 |
| TOTAL🡪 | 100 | 1.0000 |  | 100.0 |

Now calculating the chi-square using the formula we get chi-square = 5.0406 (see in Levin and Rubin book, page 551)

**DETERMINING THE DEGREES OF FREEDOM IN A GOODNESS-OF-FIT TEST**

We must count the number of classes (k) first which in the above example is 0,1,2,3=4. **First employee the k-1 rule and then subtract an additional DOF for each population parameter that has to be estimated from the sample data.**

In the interview example we have 4 classes and we are not required to estimate any population parameter hence DOF = 3.

Now calculating chi-square value from table with alpha = 0.20 and DOF = 3 we get the value = 4.642. Hence we reject the null hypothesis and conclude that the binomial fails to provide a good description of our frequencies.

**ANALYSIS OF VARIANCE(ANOVA)**

ANOVA enables us to test for the significance of the differences among more than two sample means. Using ANOVA we will be able to make inferences about whether our samples are drawn from populations having the same mean.

It is useful in such situations as comparing the mileage achieved by 5 different brands of gasoline, testing which of four different training methods produces the fastest learning record.

Problem: The training director wants to evaluate 3 different training methods in order to determine whether there were any differences in effectiveness.

|  |  |  |
| --- | --- | --- |
| METHOD1 | METHOD2 | METHOD3 |
| - | - | 18 |
| 15 | 22 | 24 |
| 18 | 27 | 19 |
| 19 | 18 | 16 |
| 22 | 21 | 22 |
| 11 | 17 | 15 |
| 85 | 105 | 114 |
| ÷5 | ÷5 | ÷6 |
| 17= | 21= | 19= |
|  |  | 🡨Sample Sizes |

Grand Mean ( ) = (15+18+19+22+11+22+27+18+21+17+18+24+19+16+22+15)/16

= 19

**STATEMENT OF HYPOTHESIS:**

In this case, our reason for using ANOVA is to decide whether these 3 samples were drawn from populations having the same means. Because we are testing the effectiveness of the three training methods, we must determine whether the 3 samples, represented by the sample means could have been drawn from population having the same mean.

In order to use ANOVA we must assume that each of the samples is drawn from a normal population and that each of these population has the same variance, . However if the sample sizes are large enough we do not need the assumption of normality.

**ANOVA is based on a comparison of two different estimates of the variance, , of overall population.** In this case, we can calculate one of these estimates by examining the variance among the three sample means, which are 17, 21 and 19. The other estimate of the population variance is determined by the variation within the 3 samples themselves. Then we compare these 2 estimates of the population variance, because both the estimates are of variance, they should be approximately equal in value when the *null hypothesis is true*. If the null hypothesis is not true, these 2 estimates will differ considerably.

* **Determine one estimate of the population variance from the variance *among the sample means.***
* **Determine a 2nd estimate of the variance from the variance within the samples**
* **Compare these 2 estimates. If they are approx. equal, accept the null hypothesis.**

**CALCULATING THE VARIANCE AMONG THE SAMPLE MEANS**

Variance among the sample means:

Now from earlier we know that Standard error of the mean=

Squaring the above equation:

Now we know the variance among the sample mean,. So we substitute it in place of . Now the above equation becomes:

**ESTIMATE OF BETWEEN-COLUMN VARIANCE**

Where,

K= number of samples.

**CALCULATING THE VARIANCE WITHIN SAMPLES**

Estimates of within column variance:

Where,

K=number of samples

= sample variance of the jth sample

**THE F HYPOTHESIS TEST: Computing and Interpreting the F statistic**

Step 3 in ANOVA compares these 2 estimates of the population variance by computing there ratio, called F, as follows:

In the above example, F=20/14.769=1.354🡨F ratio.

**When populations are not the same, the between-column variance tends to be larger than the within-column variance, and the F value tends to be large. This leads to the rejection of null hypothesis.**

THE F DISTRIBUTION

Like other statistics we studies, if the null hypothesis is true, then the F statistic has a particular sampling distribution. The F distribution is identified by a *pair of DOF unlike the T and chi-square distribution which have just 1 DOF*. **The first number is the DOF in the numerator of the F ratio and the 2nd is the DOF in the denominator.**

**Numerator degrees of freedom=**

**Denominator degrees of freedom=**

For the above example: F=1.354

DOF of numerator=3-1=2

DOF of denominator = (5-1) + (5-1) + (6-1) =13

From F table we find the value of 3.81. Because the calculated value is within the acceptance region we would accept the null hypothesis.

Our problem examined the effect of the type of training method on employee productivity, nothing else. Had we wished to measure the effect of 2 factors, such as training program and the age of employee, we would need the ability to use 2-ways ANOVA, a statistical method which is very advanced.

SIMPLE REGRESSION AND CORRELATION

This chapter tells us how to determine the *relationship between variables*

We used chi-square tests of independence to determine whether a statistical relationship existed between 2 variables. The chi-square test tells us whether there is such a relationship, but it does not tell us what the relationship is. **Regression and Correlation analysis show us how to determine both the nature and the strength of a relationship between two variables.**

In regression analysis, we shall develop an *estimating equation* – i.e. a mathematical formula that relates the known variable to the unknown variable. Then after we have learned the pattern of this relationship, we can apply correlation analysis to determine the degree to which the variables are related. Correlation analysis then tells us how well the estimating equation described the relationship.

**TYPES OF RELATIONSHIP**

Regression and Correlation analysis are based on the relationship or association between 2(or more) variables. The known variable(s) is called the *independent variable(s).* The variable we are trying to predict is called the *dependent variable.*

For example there is a relationship between annual sales of aerosol spray cans and the quantity of fluorocarbons emitted into the atmosphere each year. Here the number of spray cans is the independent variable and the fluorocarbons emission is dependent. There can be direct (one increases the other increases) relationship as well as indirect (one increases other decreases) relationship between the independent and the dependent variable.

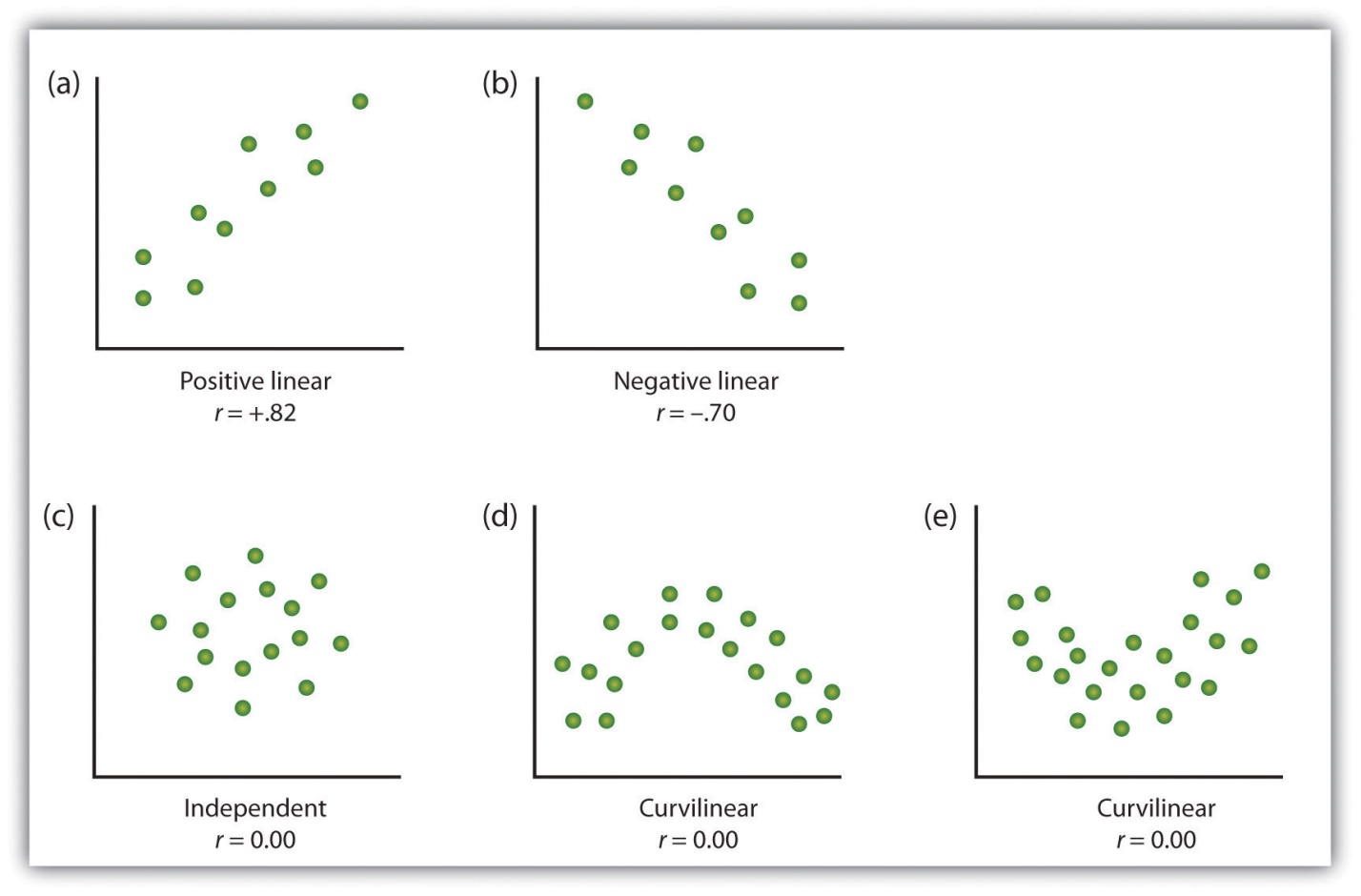
In regression we have only one dependent variable in our estimating equation. However we can use more than one independent variable.

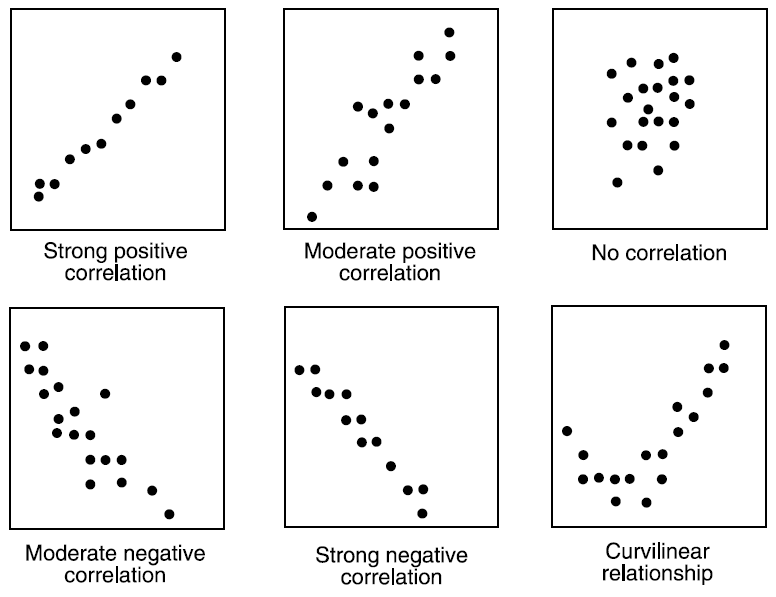
**We often find a causal relationship between variables i.e. the independent variable causes the dependent variable to change. While it can be the case for some cases but in many cases, other factors also cause the changes in both dependent and independent variable.**

**For this reason, it is important that we consider the relationship found by regression to be relationships of association but not necessarily of cause and effect. Unless you have specific reasons for believing that the values of the dependent variable are caused by the values of the independent variable(s), do not infer causality from the relationship you find by regression.**

**SCATTER DIAGRAMS**

The first step in determining whether there is a relationship between two variables is to examine the graph of the observed (known) data, known as a *scatter diagram.*

A scatter diagram can give us 2 types of information. Visually we can look for patterns that indicate that the variables are related. Then if the variable are related we can see what kind of line or estimating equation describes the relationship.**

**

***ESTIMATION USING THE REGRESSION LINE***

Equation of line: Y = a + bX

Y: Dependent variable

A: y intercept

B: slope

X: Independent variable.

Slope of a straight line

**THE METHOD OF LEAST SQUARES**

Now that we have seen how to determine the equation of a straight line, let see how we can calculate the equation of a line that is drawn through the middle of a set of points in a scatter diagram. To a statistician, the line will have a “good fit” if it minimizes the error between the estimated points on the line and the actual observed points that were used to draw it.

**We want to find a way to penalize large absolute errors so that we can avoid them. We can accomplish this if we square the individual errors before we add them.** Squaring each term accomplishes 2 goals:

* Magnifies/ penalizes the large errors
* Its cancels the effect of the positive and negative values.

(See example in book, page 623)

**SLOPE OF THE BEST FITTING REGRESSION LINE**

B: slope of the best fitting line

X: values of the independent variable

Y: values of the dependent variable

X bar: mean of the values of the independent variable

Y bar: mean of the values of the dependent variable

N: number of data points (i.e. the number of pair values for the independent and dependent variable)

**Y-INTERCEPT OF THE BEST FITTING REGRESSION LINE**

Suppose the director is interested in the relationship between the age of a garbage truck and the annual expense she should expect to incur.

|  |  |  |
| --- | --- | --- |
| Truck Number | Age of Truck | Repair expense |
| 101 | 5 | 7 |
| 102 | 3 | 7 |
| 103 | 3 | 6 |
| 104 | 1 | 4 |

Now ,

Now for a truck which is 4 years old,

Thus the manager might expect to spend about 675 $ annually in repairs of a 4 year old truck.

**THE STANDARD ERROR OF THE ESTIMATE**

To measure the reliability of the estimating equation, statisticians have developed the *standard error of estimate*. Standard deviation was used to measure the dispersion of a set of observations about the mean. **The standard error of estimate measures the variability or scatter of the observed values around the regression line.**

**Standard error of the estimate:**

Y: values of the dependent variable

Y cap: estimated values from the estimating equation that correspond to each value of Y

N: number of data points used to fit the regression line.

In the previous example:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X | Y | (3.75+0.75X) |  |  |
| 5 | 7 | 7.5 | -0.5 | 0.25 |
| 3 | 7 | 6.0 | 1.0 | 1.00 |
| 3 | 6 | 6.0 | 0.0 | .00 |
| 1 | 4 | 4.5 | -0.5 | 0.25 |
|  |  |  | TOTAL-> | 1.50 |

**SHORTCUT METHOD FOR FINDING THE STANDARD ERROR OF THE ESTIMATE**

Assuming that the observed points are normally distributed around the regression line, we can expect to find **68% of the points within +/- 1, 95.5% within +/- 2and 99.7% within +/- 3.**

At this point we should state the assumptions we are making,

* The observed values of Y are normally distributed around each estimated value of Y cap
* The variance of the distributions around each possible value of Y is the same.

**CORRELATION ANALYSIS**

**Correlation analysis is the statistical tool we can use to describe the degree to which one variable is linearly related to another.** Often correlation analysis is used in conjunction with regression analysis to measure how well the regression line explains the variation of the dependent variable, Y. Statisticians have developed 2 measures for describing the correlation between two variables: *the coefficient of determination* and the *coefficient of correlation.*

**THE COEFFICIENT OF DETERMINATION**

The coefficient of determination is the primary way we can measure the extent, or strength of the association that exists between 2 variables X and Y. Because we have used a sample of points to develop the regression lines, we refer to this measure as the *sample coefficient of determination* which is developed from the relationship between 2 kinds of variation: the variation of the Y values in a data set around:

* The fitted regression line
* Their own mean

**Variation of Y values around the regression line:**

**Variation of Y values around their own mean:**

**Sample coefficient of determination:**

**When , then there is perfect correlation.**

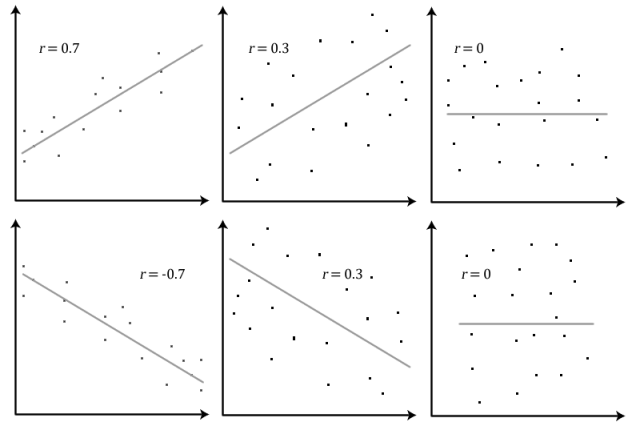
**When then there is no correlation.**

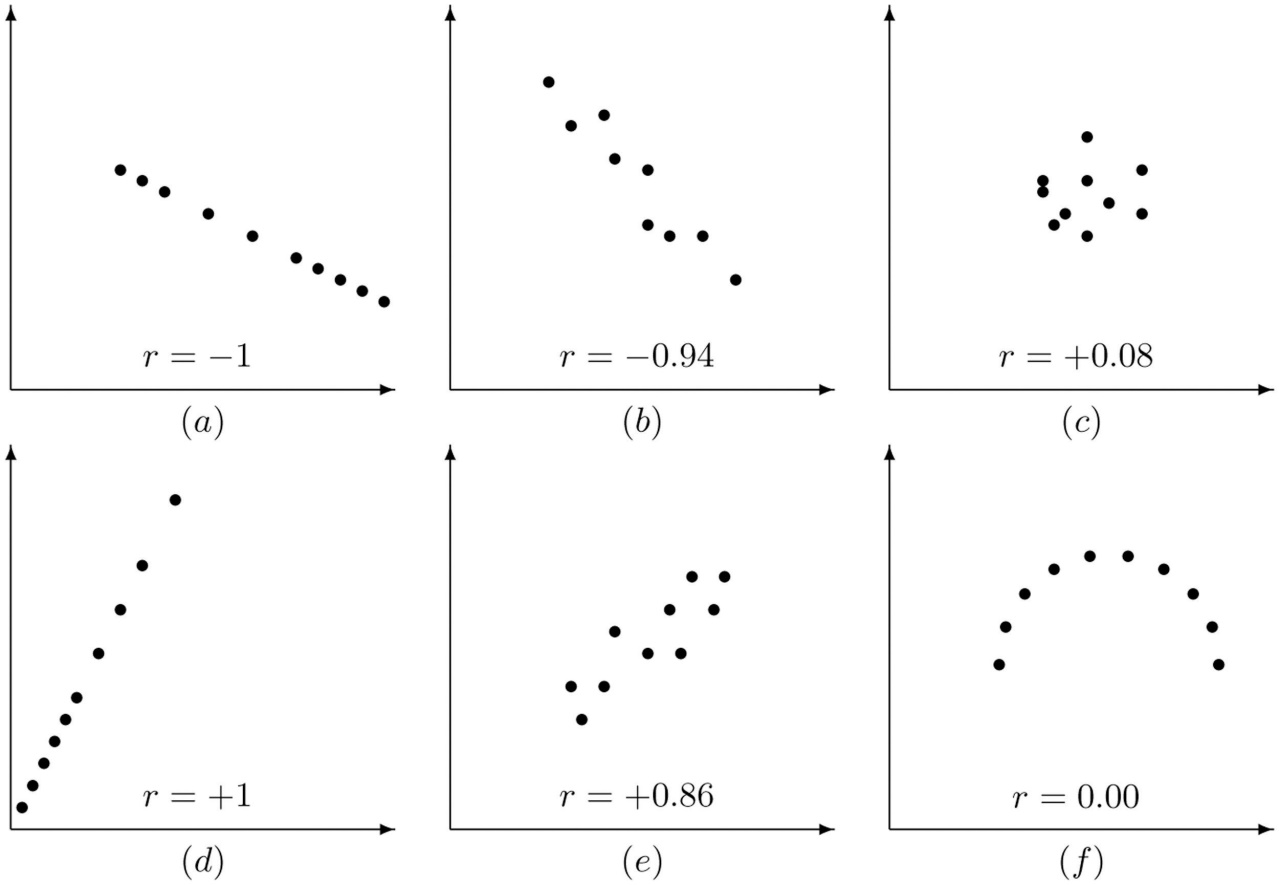
**One point that must be emphasized strongly is that measure only the strength of a linear relationship between 2 variables.**

**SHORTCUT METHOD OF FINDING**

**THE COEFFICIENT OF CORRELATION**

When the slope of the estimating equation is positive, r is positive square root, but if b is negative, r is negative square root. Thus, **the sign of r indicates the direction of relationship between the 2 variables X and Y.**

****

****

**USING REGRESSION AND CORRELATION ANALYSIS: LIMITATIONS, ERRORS AND CAVEATS**

* **Extrapolation beyond the Range of the Observed Data:** A common mistake is to assume that the estimating line can be applied over any range of values. An estimating equation is only valid over the same range as the one from which the sample was initially taken.
* **Cause and Effect:** Assuming that a change in one variable is caused by a change in the other variable. Regression and Correlation analysis can in no way determine cause and effect.
* **Using past trends to estimate future trends:** The historical data used to estimate the regression line must be reappraised. Conditions can change and violate one or more of the assumptions on which our regression analysis depends. Another error than can arise from the use of historical data is the dependence of some variables on time.
* **Misinterpreting the coefficients of Correlation and Determination:** The coefficient of correlation is often misinterpreted as percentage. If r=0.6 it is incorrect to state that the regression equation explains 60% of the total variation in Y. Instead if r=0.6 then **=0.6X0.6=0.36**, Only 36% of the total variation is explained by the regression line.
* **Finding relationships when they do not exist.**