"Introduction to Criminal Justice Statistics".

- In Section 1 of this course you will cover these topics:
  - Why The Social Researcher Uses Statistics
  - Organizing The Data
  - Measures Of Central Tendency

Topic Objective:

After reading this topic, student would be able to:

- Discuss the nature of social research
- Learn about the stages of social research
- Explain the use of use of series of numbers to do social research
- Understands testing of hypothesis
- Understand the functions of statistics

Definition/Overview:

Social research refers to research conducted by social scientists (primarily within sociology and social psychology), but also within other disciplines such as social policy, human geography, political science, social anthropology and education. Sociologists and other social scientists study diverse things: from census data on hundreds of thousands of human beings, through the

in-depth analysis of the life of a single important person to monitoring what is happening on a street today - or what was happening a few hundred years ago.

Key Points:

1. Nature of Social Research

Social scientists use many different methods in order to describe, explore and understand social life. Social methods can generally be subdivided into two broad categories. Quantitative methods are concerned with attempts to quantify social phenomena and collect and analyze numerical data, and focus on the links among a smaller number of attributes across many cases. Qualitative methods, on the other hand, emphasize personal experiences and interpretation over quantification, are more concerned with understanding the meaning of social phenomena and focus on links among a larger number of attributes across relatively few cases. While very different in many aspects, both qualitative and quantitative approaches involve a systematic interaction between theories and data.

Common tools of quantitative researchers include surveys, questionnaires, and secondary analysis of statistical data that has been gathered for other purposes (for example, censuses or the results of social attitudes surveys). Commonly used qualitative methods include focus groups, participant observation, and other techniques.

Before the advent of sociology and application of the scientific method to social research, human inquiry was mostly based on personal experiences, and received wisdom in the form of tradition and authority. Such approaches often led to errors such as inaccurate observations, over generalization, selective observations, subjectivity and lack of logic.

Social research (and social science in general) is based on logic and empirical observations. Charles C. Ragin writes in his Constructing Social Research book that "Social research involved the interaction between ideas and evidence. Ideas help social researchers make sense of

evidence, and researchers use evidence to extend, revise and test ideas". Social research thus attempts to create or validate theories through data collection and data analysis, and its goal is exploration, description and explanation. It should never lead or be mistaken with philosophy or belief. Social research aims to find social patterns of regularity in social life and usually deals with social groups (aggregates of individuals), not individuals themselves (although science of psychology is an exception here). Research can also be divided into pure research and applied research. Pure research has no application on real life, whereas applied research attempts to influence the real world.

There are no laws in social science that parallel the laws in the natural science. A law in social science is a universal generalization about a class of facts. A fact is an observed phenomenon, and observation means it has been seen, heard or otherwise experienced by researcher. A theory is a systematic explanation for the observations that relate to a particular aspect of social life. Concepts are the basic building blocks of theory and are abstract elements representing classes of phenomena. Axioms or postulates are basic assertions assumed to be true. Propositions are conclusions drawn about the relationships among concepts, based on analysis of axioms. Hypotheses are specified expectations about empirical reality which are derived from propositions. Social research involves testing these hypotheses to see if they are true.

2. Stages of Social Research

Social research involves creating a theory, operationalization (measurement of variables) and observation (actual collection of data to test hypothesized relationship). Social theories are written in the language of variables, in other words, theories describe logical relationships between variables. Variables are logical sets of attributes, with people being the 'carriers' of those variables (for example, gender can be a variable with two attributes: male and female). Variables are also divided into independent variables (data) that influences the dependent variables (which scientists are trying to explain). For example, in a study of how different dosages of a drug are related to the severity of symptoms of a disease, a measure of the severity of the symptoms of the disease is a dependent variable and the administration of the drug in specified doses is the

independent variable. Researchers will compare the different values of the dependent variable (severity of the symptoms) and attempt to draw conclusions.

3. Functions of Statistics

Statistics work in a number of ways in social research. Social research can be deductive or inductive. The inductive inquiry (also known as grounded research) is a model in which general principles (theories) are developed from specific observations. In deductive inquiry specific expectations of hypothesis are developed on the basis of general principles (i.e. social scientists start from an existing theory, and then search for proof). For example, in inductive research, if a scientist finds that some specific religious minorities tend to favor a specific political view, he may then extrapolate this to the hypothesis that all religious minorities tend to have the same political view. In deductive research, a scientist would start from a hypothesis that religious affiliation influenced political views and then begin observations to prove or disprove this hypothesis.

There is usually a trade off between the number of cases and the number of their variables that social research can study. Qualitative research usually involves few cases with many variables, while quantitative involves many phenomena with few variables.

Qualitative methods can be used in order to develop quantitative research tools. For example, focus groups could be used to explore an issue with a small number of people and the data gathered using this method could then be used to develop a quantitative survey questionnaire that could be administered to a far greater number of people allowing results to be generalized.

Qualitative methods can be used to explore and facilitate the interpretation of relationships between variables. For example researchers may inductively hypothesize that there would be a positive relationship between positive attitudes of sales staff and the amount of sales of a store. However, quantitative, deductive, structured observation of 576 convenience stores could reveal that this was not the case, and in order to understand why the relationship between the variables was negative the researchers may undertake qualitative case studies of four stores including participant observation. This might abductively confirm that the relationship was negative, but

that it was not the positive attitude of sales staff that led to low sales, but rather that high sales led to busy staff who were less likely to be express positive emotions at work.

4. Testing of Hypothesis

A statistical hypothesis test is a method of making statistical decisions using experimental data. It is sometimes called confirmatory data analysis to emphasise the comparison with exploratory data analysis. The decisions are almost always made using null-hypothesis tests; a null-hypothesis test just answers the question What is the probability that the same data would be found due random chance alone?

The test described here is more fully the null-hypothesis statistical significance test. The null hypothesis is a conjecture that exists solely to be falsified by the sample. Statistical significance is a possible finding of the test that the sample is unlikely to have occurred by chance given the truth of the null hypothesis. The name of the test describes its formulation and its possible outcome. One characteristic of the test is its crisp decision: reject or do not reject (which is not the same as accept). A calculated value is compared to a threshold, which is determined from the tolerable risk of error.

Statistics is a mathematical science pertaining to the collection, analysis, interpretation or explanation, and presentation of data. It is applicable to a wide variety of academic disciplines, from the natural and social sciences to the humanities, and to government and business.

Statistical methods can be used to summarize or describe a collection of data; this is called descriptive statistics. In addition, patterns in the data may be modeled in a way that accounts for randomness and uncertainty in the observations, and then used to draw inferences about the process or population being studied; this is called inferential statistics. Both descriptive and inferential statistics comprise applied statistics. There is also a discipline called mathematical statistics, which is concerned with the theoretical basis of the subject.

Topic Objective:

After reading this topic, student would be able to:

- Discusses frequency distributions of nominal data
- Explains proportions and percentages
- Discuss simple frequency distributions of ordinal and interval data
- Explain cumulative distributions
- Discuss graphic presentations

Definition/Overview:

Data refer to a collection of facts usually collected as the result of experience, observation or experiment, or processes within a computer system, or a set of premises. This may consist of numbers, words, or images, particularly as measurements or observations of a set of variables. Data are often viewed as a lowest level of abstraction from which information and knowledge are derived. After being collected and processed, data need to be organized to produce useful information. When organizing data, it helps to be familiar with some of the definitions. This topic outlines those definitions and provides some simple techniques for organizing and presenting data.

Key Points:

1. Frequency Distributions of Nominal Data

In statistics, a frequency distribution is a list of the values that a variable takes in a sample. It is usually a list, ordered by quantity, showing the number of times each value appears. In computer science, data is anything in a form suitable for use with a computer. Data is often distinguished from programs. A program is a set of instructions that detail a task for the computer to perform.

In this sense, data is thus everything that is not program code. In an alternate usage, binary files (which are not human-readable) are sometimes called "data" as distinguished from human-readable "text". The total amount of digital data in 2007 was estimated to be 281 billion gigabytes. Managing and operating on frequency tabulated data is much simpler than operation on raw data. There are simple algorithms to calculate median, mean, standard deviation etc. from these tables. Statistical hypothesis testing is founded on the assessment of differences and similarities between frequency distributions. This assessment involves measures of central tendency or averages, such as the mean and median, and measures of variability or statistical dispersion, such as the standard deviation or variance.

2. Proportions and Percentages

In mathematics, a percentage is a way of expressing a number as a fraction of 100 (per cent meaning "per hundred"). It is often denoted using the percent sign, "%". For example, 45% (read as "forty-five percent") is equal to 45 / 100, or 0.45. Percentages are used to express how large one quantity is, relative to another quantity. The first quantity usually represents a part of, or a change in, the second quantity, which should be greater than zero. For example, an increase of $ 0.15 on a price of $ 2.50 is an increase by a fraction of 0.15 / 2.50 = 0.06. Expressed as a percentage, this is therefore a 6% increase. Although percentages are usually used to express numbers between zero and one, any dimensionless proportionality can be expressed as a percentage. For instance, 111% is 1.11 and −0.35% is −0.0035. A frequency distribution is said to be skewed when its mean and median are different. The kurtosis of a frequency distribution is the concentration of scores at the mean, or how peaked the distribution appears if depicted graphicallyfor example, in a histogram. If the distribution is more peaked than the normal distribution it is said to be leptokurtic; if less peaked it is said to be platykurtic. Frequency distributions are also used in frequency analysis to crack codes and refer to the relative frequency of letters in different languages.

Although percentages are usually used to express numbers between zero and one, any dimensionless proportionality can be expressed as a percentage. For instance, 111% is 1.11 and

−0.35% is −0.0035. Percentages are correctly used to express fractions of the total. For example, 25% means 25 / 100, or one quarter, of some total.

Percentages larger than 100%, such as 101% and 110%, may be used as a literary paradox to express motivation and exceeding of expectations. For example, "We expect you to give 110% [of your ability]"; however, there are cases when percentages over 100 can be meant literally (such as "a family must earn at least 125% over the poverty line to sponsor a spouse visa").

3. Simple Frequency Distributions of Ordinal and Interval Data

The normal distribution, also called the Gaussian distribution, is an important family of continuous probability distributions, applicable in many fields. Each member of the family may be defined by two parameters, location and scale: the mean ("average", μ) and variance (standard deviation squared) σ2, respectively. The standard normal distribution is the normal distribution with a mean of zero and a variance of one (the red curves in the plots to the right). Carl Friedrich Gauss became associated with this set of distributions when he analyzed astronomical data using them, and defined the equation of its probability density function. It is often called the bell curve because the graph of its probability density resembles a bell. The importance of the normal distribution as a model of quantitative phenomena in the natural and behavioral sciences is due to the central limit theorem. Many measurements, ranging from psychological to physical phenomena (in particular, thermal noise) can be approximated, to varying degrees, by the normal distribution. While the mechanisms underlying these phenomena are often unknown, the use of the normal model can be theoretically justified by assuming that many small, independent effects are additively contributing to each observation.

The normal distribution also arises in many areas of statistics. For example, the sampling distribution of the sample mean is approximately normal, even if the distribution of the population from which the sample is taken is not normal. In addition, the normal distribution maximizes information entropy among all distributions with known mean and variance, which makes it the natural choice of underlying distribution for data summarized in terms of sample mean and variance. The normal distribution is the most widely used family of distributions in statistics and many statistical tests are based on the assumption of normality. In probability

theory, normal distributions arise as the limiting distributions of several continuous and discrete families of distributions.

4. Cumulative Distributions

In probability theory and statistics, the cumulative distribution function (CDF), also called cumulative density function, probability distribution function or just distribution function, completely describes the probability distribution of a real-valued random variable X. For every real number x, the CDF of X is given by

where the right-hand side represents the probability that the random variable X takes on a value less than or equal to x. The probability that X lies in the interval (a, b] is therefore $F_X(b) - F_X(a)$ if $a < b$. If treating several random variables X,Y,...etc. the corresponding letters are used as subscripts while, if treating only one, the subscript is omitted. It is conventional to use a capital F for a cumulative distribution function, in contrast to the lower-case f used for probability density functions and probability mass functions. This applies when discussing general distributions: some specific distributions have their own conventional notation, for example the Normal Distribution. The CDF of X can be defined in terms of the probability density function f as follows:

Note that in the definition above, the "less or equal" sign, '≤' is a convention, but it is a universally used one, and is important for discrete distributions. The proper use of tables of the Binomial and Poisson distributions depend upon this convention. Moreover, important formulas like Levy's inversion formula for the characteristic function also rely on the "less or equal" formulation. A cross tabulation (often abbreviated as cross tab) displays the joint distribution of two or more variables. They are usually presented as a contingency table in a matrix format. Whereas a frequency distribution provides the distribution of one variable, a contingency table describes the distribution of two or more variables simultaneously. Each cell shows the number of respondents who gave a specific combination of responses, that is, each cell contains a single cross tabulation. Cross tabs are frequently used because:

- They are easy to understand. They appeal to people who do not want to use more sophisticated measures.
- They can be used with any level of data: nominal, ordinal, interval, or ratio - cross tabs treat all data as if it is nominal
- A table can provide greater insight than single statistics
- It solves the problem of empty or sparse cells
- They are simple to conduct

5. Graphic Presentations

Advanced Function Presentation (AFP) is a presentation architecture and family of associated printer software and hardware that provides for document and information presentation independent of specific applications and devices. Using AFP, users can control formatting, the form of paper output, whether a document is to be printed or viewed online, and manage document storage and access in a distributed network across multiple operating system platforms. AFP is primarily used in large enterprises for production variable data printing (VDP). AFP applications allow users or print room operators to distribute print jobs among a group of printers and to designate backup printers when one fails. AFP is considered to be a "cornerstone" of electronic document management (EDM) applications such as print-and-view, archive and retrieval, and Computer Output to Laser Disk (COLD). AFP was originally developed as the general purpose document and information presentation architecture of IBM. The first specifications and products go back to 1984. The major concepts of object-driven structures, print integrity, resource management, and support for high print speeds have been preserved ever since. In October 2004 IBM initiated the formation of the AFP Color Consortium (AFPCC). The purpose was to collaboratively develop color management support in the AFP architecture. This resulted in the creation of the new AFP CMOCA (Color Management Object Content Architecture) specification, which was first published in 2006.

Topic Objective:

After reading this topic, student would be able to:

- Discuss the mode
- Discuss the median
- Explain the mean
- Learn about step-by-step illustration: mode, median and mean

Definition/Overview:

In mathematics, an average, or central tendency of a data set refers to a measure of the "middle" or "expected" value of the data set. There are many different descriptive statistics that can be chosen as a measurement of the central tendency of the data items.

The most common method is the arithmetic mean, but there are many other types of averages. The average is calculated by combining the measurements related to a group of people or objects, to compute a number as being the average of the group.

The term central tendency refers to the "middle" value or perhaps a typical value of the data, and is measured using the mean, median, or mode. Each of these measures is calculated differently, and the one that is best to use depends upon the situation.

Key Points:

1. Mean

The mean is the most commonly-used measure of central tendency. When we talk about an "average", we usually are referring to the mean. The mean is simply the sum of the values divided by the total number of items in the set. The result is referred to as the arithmetic mean.

Sometimes it is useful to give more weighting to certain data points, in which case the result is called the weighted arithmetic mean.

The notation used to express the mean depends on whether we are talking about the population mean or the sample mean:

= population mean

= sample mean

The population mean then is defined as:

=

[Fig 1]

where

= number of data points in the population

= value of each data point i.

The mean is valid only for interval data or ratio data. Since it uses the values of all of the data points in the population or sample, the mean is influenced by outliers that may be at the extremes of the data set.

2. Median

The median is determined by sorting the data set from lowest to highest values and taking the data point in the middle of the sequence. There is an equal number of points above and below the median. For example, in the data set {1,2,3,4,5} the median is 3; there are two data points greater than this value and two data points less than this value. In this case, the median is equal to the mean. But consider the data set {1,2,3,4,10}. In this dataset, the median still is three, but the mean is equal to 4. If there is an even number of data points in the set, then there is no single point at the middle and the median is calculated by taking the mean of the two middle points. The median can be determined for ordinal data as well as interval and ratio data. Unlike the mean, the median is not influenced by outliers at the extremes of the data set. For this reason, the median often is used when there are a few extreme values that could greatly influence the mean and distort what might be considered typical. This often is the case with home prices and with

income data for a group of people, which often is very skewed. For such data, the median often is reported instead of the mean. For example, in a group of people, if the salary of one person is 10 times the mean, the mean salary of the group will be higher because of the unusually large salary. In this case, the median may better represent the typical salary level of the group.

3. Mode

The mode is the most frequently occurring value in the data set. For example, in the data set {1,2,3,4,4}, the mode is equal to 4. A data set can have more than a single mode, in which case it is multimodal. In the data set {1,1,2,3,3} there are two modes: 1 and 3.

The mode can be very useful for dealing with categorical data. For example, if a sandwich shop sells 10 different types of sandwiches, the mode would represent the most popular sandwich. The mode also can be used with ordinal, interval, and ratio data. However, in interval and ratio scales, the data may be spread thinly with no data points having the same value. In such cases, the mode may not exist or may not be very meaningful. When to use Mean, Median, and Mode. The following table summarizes the appropriate methods of determining the middle or typical value of a data set based on the measurement scale of the data.

MEASUREMENT SCALE	BEST MEASURE OF THE "MIDDLE"
Nominal (Categorical)	Mode
Ordinal	Median
Interval	Symmetrical data: Mean Skewed data: Median
Ratio	Symmetrical data: Mean Skewed data: Median

[Table 1]

4. Illustration

In probability theory and statistics, a median is described as the number separating the higher half of a sample, a population, or a probability distribution, from the lower half. The median of a finite list of numbers can be found by arranging all the observations from lowest value to highest value and picking the middle one. If there is an even number of observations, the median is not unique, so one often takes the mean of the two middle values. At most half the population has values less than the median and at most half have values greater than the median. If both groups contain less than half the population, then some of the population is exactly equal to the median. For example, if a < b < c, then the median of the list {a, b, c} is b, and if a < b < c < d, then the median of the list {a, b, c, d} is the mean of b and c, i.e. it is (b + c)/2. The median can be used when a distribution is skewed or when end values are not known. A disadvantage is the difficulty of handling it theoretically. In statistics, the mode is the value that occurs the most frequently in a data set or a probability distribution. In some fields, notably education, sample data are often called scores, and the sample mode is known as the modal score. Like the statistical mean and the median, the mode is a way of capturing important information about a random variable or a population in a single quantity. The mode is in general different from mean and median, and may be very different for strongly skewed distributions. The mode is not necessarily unique, since the same maximum frequency may be attained at different values. The most ambiguous case occurs in uniform distributions, wherein all values are equally likely

Test Introduction to Criminal Justice Statistics

- In Section 2 of this course you will cover these topics:
  - Measures Of Variability
  - Probability And The Normal Curve
  - Samples And Populations

Topic Objective:

After reading this topic, student would be able to:

- Discuss the range
- Explain the variance and standard deviation
- Learn about step-by-step illustration: standard deviation
- Understands the meaning of the standard deviation
- Discuss looking at the larger picture: describing data

Definition/Overview:

In statistics, (statistical) dispersion (also called statistical variability or variation) is variability or spread in a variable or a probability distribution. Common examples of measures of statistical dispersion are the variance, standard deviation and inter-quartile range. In the physical sciences, such variability may result only from random measurement errors: instrument measurements are often not perfectly precise, i.e., reproducible. One may assume that the quantity being measured is unchanging and stable, and that the variation between measurements is due to observational error. In the biological sciences, this assumption is false: the variation observed might be intrinsic to the phenomenon: distinct members of a population differ greatly. This is also seen in the arena of manufactured products; even there, the meticulous scientist finds variation. The simple model of a stable quantity is preferred when it is tenable. Each phenomenon must be examined to see if it warrants such a simplification.

Key Points:

1. Range

In descriptive statistics, the range is the length of the smallest interval which contains all the data. It is calculated by subtracting the smallest observation (sample minimum) from the greatest (sample maximum) and provides an indication of statistical dispersion. It is measured in the same units as the data. Since it only depends on two of the observations, it is a poor and weak measure of dispersion except when the sample size is large. For a population, the range is greater than or equal to twice the standard deviation, which equality only for the coin toss (Bernoulli distribution with p =). The range, in the sense of the difference between the highest and lowest scores, is also called the crude range. When a new scale for measurement is developed, then a potential maximum or minimum will emanate from this scale. This is called the potential (crude) range. Of course this range should not be chosen too small, in order to avoid a ceiling effect. When the measurement is obtained, the resulting smallest or greatest observation, will provide the observed (crude) range. The midrange point, i.e. the point halfway between the two extremes, is an indicator of the central tendency of the data. Again it is not particularly robust for small samples.

2. Standard Deviation

In statistics, standard deviation is a simple measure of the variability or dispersion of a data set. A low standard deviation indicates that all of the data points are very close to the same value (the mean), while high standard deviation indicates that the data is spread out over a large range of values. For example, the average height for adult men in the United States is about 70 inches, with a standard deviation of around 3 inches. This means that most men (about 68%, assuming a normal distribution) have a height within 3 inches of the mean (67 inches 73 inches), while almost all men (about 95%) have a height within 6 inches of the mean (64 inches 76 inches). If the standard deviation were zero, then all men would be exactly 70 inches high. If the standard deviation were 20 inches, then men would have much more variable heights, with a typical range of about 50 to 90 inches. In addition to expressing the variability of a population, standard deviation is commonly used to measure confidence in statistical conclusions.

3. Illustration: Standard Deviation

For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be conducted multiple times. (Typically the reported margin of error is about twice the standard deviation, the radius of a 95% confidence interval.) In science, researchers commonly report the standard deviation of experimental data, and only effects that fall far outside the range of standard deviation are considered statistically significant. Standard deviation is also important in finance, where the standard deviation on the rate of return on an investment is a measure of the risk. Formulated by Francis Galton in the late 1860s, the standard deviation remains the most common measure of statistical dispersion. A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data. When only a sample of data from a population is available, the population standard deviation can be estimated by a modified standard deviation of the sample, explained below.

4. Meaning of the Standard Deviation

In statistics, standard deviation is a simple measure of the variability or dispersion of a data set. A low standard deviation indicates that all of the data points are very close to the same value (the mean), while high standard deviation indicates that the data is spread out over a large range of values. A measure of statistical dispersion is a real number that is zero if all the data are identical, and increases as the data becomes more diverse. It cannot be less than zero. Most measures of dispersion have the same scale as the quantity being measured. In other words, if the measurements have units, such as meters or seconds, the measure of dispersion has the same units. In descriptive statistics, the range is the length of the smallest interval which contains all the data. It is calculated by subtracting the smallest observations from the greatest and provides an indication of statistical dispersion. It is measured in the same units as the data. Since it only depends on two of the observations, it is a poor and weak measure of dispersion except when the sample size is large. For a population, the range is more than twice the standard deviation. The range, in the sense of the difference between the highest and the lowest scores is also called the crude range. When a new scale for measurement is developed, then a potential maximum or minimum will emanate from this scale. This is called the potential (crude) range. Of course this

range should not be chosen too small, in order to avoid a ceiling effect. When the measurement is obtained, the resulting smallest or greatest observation, will provide the observed (crude) range. The midrange point, i.e. the point halfway between the two extremes, is an indicator of the central tendency of the data. Again it is not particularly robust for small samples. In probability theory and statistics, the variance of a random variable, probability distribution, or sample is one measure of statistical dispersion, averaging the squared distance of its possible values from the expected value (mean). Whereas the mean is a way to describe the location of a distribution, the variance is a way to capture its scale or degree of being spread out. The unit of variance is the square of the unit of the original variable. The positive square root of the variance, called the standard deviation, has the same units as the original variable and can be easier to interpret for this reason.

The variance of a real-valued random variable is its second central moment, and it also happens to be its second cumulant. Just as some distributions do not have a mean, some do not have a variance. The mean exists whenever the variance exists, but not vice versa. In probability and statistics, the standard deviation is a measure of the dispersion of a set of values. It can apply to a probability distribution, a random variable, a population or a multi-set. The standard deviation is usually denoted with the letter σ (lower case sigma). It is defined as the root-mean-square (RMS) deviation of the values from their mean, or as the square root of the variance. Formulated by Galton in the late 1860s, the standard deviation remains the most common measure of statistical dispersion, measuring how widely spread the values in a data set are. If many data points are close to the mean, then the standard deviation is small; if many data points are far from the mean, then the standard deviation is large. If all data values are equal, then the standard deviation is zero. A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data

Topic Objective:

After reading this topic, student would be able to:

- Explain probability
- Discuss probability distributions
- Learn about characteristics of the normal curve
- Discuss the model and the reality of the normal curve
- Learn about the area under the normal curve
- Step-by-Step Illustration: Probability under the Normal Curve.

Definition/Overview:

The word probability has been used in a variety of ways since it was first coined in relation to games of chance. Does probability measure the real, physical tendency of something to occur, or is it just a measure of how strongly one believes it will occur? In answering such questions, we interpret the probability values of probability theory. There are two broad categories of probability interpretations which can be called 'physical' and 'evidential' probabilities. Physical probabilities, which are also called objective or frequency probabilities, are associated with random physical systems such as roulette wheels, rolling dice and radioactive atoms. In such systems, a given type of event (such as the dice yielding a six) tends to occur at a persistent rate, or 'relative frequency', in a long run of trials. Physical probabilities either explain, or are invoked to explain, these stable frequencies. Thus talk about physical probability makes sense only when dealing with well defined random experiments. The two main kinds of theory of physical probability are frequentist accounts (such as those of Venn, Reichenbach and von Mises) and propensity accounts (such as those of Popper, Miller, Giere and Fetzer). Evidential probability, also called Bayesian probability, can be assigned to any statement whatsoever, even when no random process is involved, as a way to represent its subjective plausibility, or the degree to which the statement is supported by the available evidence. On most accounts, evidential probabilities are considered to be degrees of belief, defined in terms of dispositions to gamble at certain odds. The four main evidential interpretations are the classical (e.g. Laplace's)

interpretation, the subjective interpretation (de Finetti and Savage), the epistemic or inductive interpretation (Ramsey, Cox) and the logical interpretation (Keynes and Carnap). Probability is the likelihood or chance that something is the case or will happen. Probability theory is used extensively in areas such as statistics, mathematics, science and philosophy to draw conclusions about the likelihood of potential events and the underlying mechanics of complex systems.

Key Points:

1. Probability

The word probability does not have a consistent direct definition. In fact, there are two broad categories of probability interpretations:

- Frequentists talk about probabilities only when dealing with well defined random experiments. The probability of a random event denotes the relative frequency of occurrence of an experiment's outcome, when repeating the experiment.
- Bayesians, however, assign probabilities to any statement whatsoever, even when no random process is involved. Probability, for a Bayesian, is a way to represent an individual's degree of belief in a statement, given the evidence.

Like other theories, the theory of probability is a representation of probabilistic concepts in formal termsthat is, in terms that can be considered separately from their meaning. These formal terms are manipulated by the rules of mathematics and logic, and any results are then interpreted or translated back into the problem domain. Two major applications of probability theory in everyday life are in risk assessment and in trade on commodity markets. Governments typically apply probabilistic methods in environmental regulation where it is called "pathway analysis", often measuring well-being using methods that are stochastic in nature, and choosing projects to undertake based on statistical analyses of their probable effect on the population as a whole. It is not correct to say that statistics are involved in the modelling itself, as typically the assessments of risk are one-time and thus require more fundamental probability models,

2. The Normal Curve

The graph of the normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow. All normal distributions look like a symmetric, bell-shaped curve, as shown below.

The curve on the left is shorter and wider than the curve on the right, because the curve on the left has a bigger standard deviation.

3. Probability and the Normal Curve

The normal distribution is a continuous probability distribution. This has several implications for probability.

- The total area under the normal curve is equal to 1.
- The probability that a normal random variable X equals any particular value is 0.
- The probability that X is greater than a equals the area under the normal curve bounded by a and plus infinity (as indicated by the non-shaded area in the figure below).
- The probability that X is less than a equals the area under the normal curve bounded by a and minus infinity (as indicated by the shaded area in the figure below).

Additionally, every normal curve (regardless of its mean or standard deviation) conforms to the following "rule".

- About 68% of the area under the curve falls within 1 standard deviation of the mean.
- About 95% of the area under the curve falls within 2 standard deviations of the mean.
- About 99.7% of the area under the curve falls within 3 standard deviations of the mean.

4. Normal Distribution

The normal distribution, also called the Gaussian distribution, is an important family of continuous probability distributions, applicable in many fields. Each member of the family may be defined by two parameters, location and scale: the mean ("average", μ) and variance (standard deviation squared, σ2) respectively. The standard normal distribution is the normal distribution with a mean of zero and a variance of one (the red curves in the plots to the right). Carl Friedrich Gauss became associated with this set of distributions when he analyzed astronomical data using them, and defined the equation of its probability density function. It is often called the bell curve because the graph of its probability density resembles a bell. The importance of the normal distribution as a model of quantitative phenomena in the natural and behavioral sciences is due in part to the central limit theorem. Many measurements, ranging from psychological to physical phenomena (in particular, thermal noise) can be approximated, to varying degrees, by the normal distribution. While the mechanisms underlying these phenomena are often unknown, the use of the normal model can be theoretically justified by assuming that many small, independent effects are additively contributing to each observation. The normal distribution is also important for its relationship to least-squares estimation, one of the simplest and oldest methods of statistical estimation. The normal distribution also arises in many areas of statistics. For example, the sampling distribution of the sample mean is approximately normal, even if the distribution of the population from which the sample is taken is not normal. In addition, the normal distribution maximizes information entropy among all distributions with known mean and variance, which makes it the natural choice of underlying distribution for data summarized in terms of sample mean and variance. The normal distribution is the most widely used family of distributions in statistics and many statistical tests are based on the assumption of normality. In probability theory, normal distributions arise as the limiting distributions of several continuous and discrete families of distributions

Topic Objective:

After reading this topic, student would be able to:

- Discuss random sampling
- Explain sampling error
- Discuss sampling distribution of means
- Learn about confidence intervals
- Learn about step-by-step illustration: confidence interval

Definition/Overview:

In statistics, a statistical population is a set of entities concerning which statistical inferences are to be drawn, often based on a random sample taken from the population. For example, if we were interested in generalizations about crows, then we would describe the set of crows that is of interest. Notice that if we choose a population like all crows, we will be limited to observing crows that exist now or will exist in the future. Probably, geography will also constitute a limitation in that our resources for studying crows are also limited. Sampling is that part of statistical practice concerned with the selection of individual observations intended to yield some knowledge about a population of concern, especially for the purposes of statistical inference. Each observation measures one or more properties (weight, location, etc.) of an observable entity enumerated to distinguish objects or individuals. Survey weights often need to be applied to the data to adjust for the sample design. Results from probability theory and statistical theory are employed to guide practice.

Key Points:

1. Random Sampling

A sample is a subject chosen from a population for investigation. A random sample is one chosen by a method involving an unpredictable component. Random sampling can also refer to taking a number of independent observations from the same probability distribution, without involving any real population. A probability sample is one in which each item has a known probability of being in the sample. The sample usually will not be completely representative of the population from which it was drawn this random variation in the results is known as sampling error. In the case of random samples, mathematical theory is available to assess the sampling error. Thus, estimates obtained from random samples can be accompanied by measures of the uncertainty associated with the estimate. This can take the form of a standard error, or if the sample is large enough for the central limit theorem to take effect, confidence intervals may be calculated.

Random sampling- all members of the population have an equal chance of being selected as part of the sample. You might think this means just standing in the street and asking passers-by to answer your questions. However, there would be many members of the population who would not be in the street at the time you are there, therefore, they do not stand any chance of being part of your sample. To pick a random sample, it is necessary to take all the names on the electoral register(a list of all the people who live in a particular area) and pick out, for example, every fiftieth name. This particular person needs to be interviewed to make the sample truly random. Random sampling is very expensive and time consuming, but gives a true sample of the population.

2. Sampling Error

Population is also used to refer to a set of potential measurements or values, including not only cases actually observed but those that are potentially observable. Suppose, for example, we are interested in the set of all adult crows now alive in the county of Cambridgeshire, and we want to

know the mean weight of these birds. For each bird in the population of crows there is a weight, and the set of these weights is called the population of weights.

The sampling process comprises several stages:

- Defining the population of concern
- Specifying a sampling frame, a set of items or events possible to measure
- Specifying a sampling method for selecting items or events from the frame
- Determining the sample size
- Implementing the sampling plan
- Sampling and data collecting
- Reviewing the sampling process

Within any of the types of frame identified above, a variety of sampling methods can be employed, individually or in combination. sampling is divided in two categories 1. Probability Sampling 2. Nonprobability Sampling. Probability sampling includes: Simple Random Method, Systamatic Sampling, Stratified Sampling and Cluster or Multistage Sampling Nonprobability Sampling includes: Accidental Sampling, Quota Sampling and Purposive Sampling

3. Step-By-Step Illustration

In statistics, sampling error or estimation error is the error caused by observing a sample instead of the whole population. An estimate of a quantity of interest, such as an average or percentage, will generally be subject to sample-to-sample variation. These variations in the possible sample values of a statistic can theoretically be expressed as sampling errors, although in practice the exact sampling error is typically unknown. Sampling error also refers more broadly to this phenomenon of random sampling variation. The likely size of the sampling error can generally be controlled by taking a large enough random sample from the population, although the cost of doing this may be prohibitive;. If the observations are collected from a random sample, statistical theory provides probabilistic estimates of the likely size of the sampling error for a particular statistic or estimator. These are often expressed in terms of its standard error. Sampling error can

be contrasted with non-sampling error. Non-sampling error is a catch-all term for the deviations from the true value that are not a function of the sample chosen, including various systematic errors and any random errors that are not due to sampling. Non-sampling errors are much harder to quantify than sampling error.

In statistics, a sampling distribution is the probability distribution, under repeated sampling of the population, of a given statistic (a numerical quantity calculated from the data values in a sample). The formula for the sampling distribution depends on the distribution of the population, the statistic being considered, and the sample size used. A more precise formulation would speak of the distribution of the statistic as that for all possible samples of a given size, not just "under repeated sampling". For example, consider a very large normal population (one that follows the so-called bell curve). Assume we repeatedly take samples of a given size from the population and calculate the sample mean (\bar x, the arithmetic mean of the data values) for each sample. Different samples will lead to different sample means. The distribution of these means is the "sampling distribution of the sample mean" (for the given sample size). This distribution will be normal since the population was normal. (According to the central limit theorem, if the population is not normal but "sufficiently well behaved", the sampling distribution of the sample mean will still be approximately normal provided the sample size is sufficiently large.)

4. Standard Error

The standard error of a method of measurement or estimation is the estimated standard deviation of the error in that method. Specifically, it estimates the standard deviation of the difference between the measured or estimated values and the true values. Notice that the true value of the standard deviation is usually unknown and the use of the term standard error carries with it the idea that an estimate of this unknown quantity is being used. It also carries with it the idea that it measures, not the standard deviation of the estimate itself, but the standard deviation of the error in the estimate, and these can be very different. In applications where a standard error is used, it would be good to be able to take proper account of the fact that the standard error is only an estimate. Unfortunately this is not often possible and it may then be better to use an approach

that avoids using a standard error, for example by using maximum likelihood or a more formal approach to deriving confidence intervals. One well-known case where a proper allowance can be made arises where the Student's t-distribution is used to provide a confidence interval for an estimated mean or difference of means. In other cases, the standard error may usefully be used to provide an indication of the size of the uncertainty, but its formal or semi-formal use to provide confidence intervals or tests should be avoided unless the sample size is at least moderately large. Here "large enough" would depend on the particular quantities being analysed.

5. Sampling Distribution of Means and Confidence Interval

In statistics, a confidence interval (CI) is an interval estimate of a population parameter. Instead of estimating the parameter by a single value, an interval likely to include the parameter is given. Thus, confidence intervals are used to indicate the reliability of an estimate. How likely the interval is to contain the parameter is determined by the confidence level or confidence coefficient. Increasing the desired confidence level will widen the confidence interval. For example, a CI can be used to describe how reliable survey results are. In a poll of election voting-intentions, the result might be that 40% of respondents intend to vote for a certain party. A 95% confidence interval for the proportion in the whole population having the same intention on the survey date might be 36% to 44%. All other things being equal, a survey result with a small CI is more reliable than a result with a large CI and one of the main things controlling this width in the case of population surveys is the size of the sample questioned. Confidence intervals and interval estimates more generally have applications across the whole range of quantitative studies.

- In Section 3 of this course you will cover these topics:
  - Testing Differences Between Means
  - Analysis Of Variance
  - Nonparametric Tests Of Significance

Topic Objective:

After reading this topic, student would be able to:

- Discuss the null hypothesis: no difference between means
- Learn about the research hypothesis: a difference between means
- Explain the sampling distribution of differences between means.
- Explain the levels of significance
- Learn about the standard error of the difference between means
- Discuss the testing the difference between means
- Step-by-Step Illustration

Definition/Overview:

The mean difference is a measure of statistical dispersion equal to the average absolute difference of two independent values drawn from a probability distribution. A related statistic is the relative mean difference, which is the mean difference divided by the arithmetic mean. An important relationship is that the relative mean difference is equal to twice the Gini coefficient, which is defined in terms of the Lorenz curve.

The mean difference is also known as the absolute mean difference and the Gini mean difference. The mean difference is sometimes denoted by Δ or as MD. The mean deviation is a different measure of dispersion. Both the standard deviation and the mean difference measure dispersion -- how spread out are the values of a population or the probabilities of a distribution. The mean difference is not defined in terms of a specific measure of central tendency, whereas the standard deviation is defined in terms of the deviation from the arithmetic mean. Because the standard deviation squares its differences, it tends to give more weight to larger differences and less weight to smaller differences compared to the mean difference. When the arithmetic mean is

finite, the mean difference will also be finite, even when the standard deviation is infinite. See the examples for some specific comparisons.

Key Points:

1. Null Hypothesis: No Difference between Means

In statistics, a null hypothesis (H0) is a hypothesis set up to be nullified or refuted in order to support an alternative hypothesis. When used, the null hypothesis is presumed true until statistical evidence, in the form of a hypothesis test, indicates otherwise that is, when the researcher has a certain degree of confidence, usually 95% to 99%, that the data does not support the null hypothesis. It is possible for an experiment to fail to reject the null hypothesis. It is also possible that both the null hypothesis and the alternate hypothesis are rejected if there are more than those two possibilities.

In scientific and medical applications, the null hypothesis plays a major role in testing the significance of differences in treatment and control groups. The assumption at the outset of the experiment is that no difference exists between the two groups (for the variable being compared): this is the null hypothesis in this instance. Other types of null hypothesis may be, for example, that:

- Values in samples from a given population can be modelled using a certain family of statistical distributions.
- The variability of data in different groups is the same, although they may be centered around different values.
- There is no difference, hence null
- Assumption: mean of sample 1 = mean of sample 2.
- The 2 samples have been drawn from equivalent populations, and the differences between them could result from chance alone.

- If the results we actually get are very unlikely (less than 5 in 100), we reject the null hypothesis, and confirm that there is a statistically significant difference between the populations from which these 2 samples are drawn.
- We often carry out experiments, where our research hypothesis is that there is a difference between the means. It is what we want to find.
- But the statistical hypothesis is still that they are not different. We test to see if we have enough evidence to reject that hypothesis and say the differences are unlikely to be due to chance.

2. Sample Mean Differences

- Mean differences among samples from a population are themselves normally distributed
- So, if we know the population variance, we could calculate a z-score in the same fashion.
- How often do we know σ ?
- But, since we rarely know σ, we usually use the t distribution instead of the normal (z) distribution.
- Thus we calculate t = X1 - X2
- sx1 - x2

3. Levels of Significance

In statistics, a result is called statistically significant if it is unlikely to have occurred by chance. "A statistically significant difference" simply means there is statistical evidence that there is a difference; it does not mean the difference is necessarily large, important, or significant in the common meaning of the word. The significance level of a test is a traditional frequentist statistical hypothesis testing concept. In simple cases, it is defined as the probability of making a decision to reject the null hypothesis when the null hypothesis is actually trues (a decision known as a Type I error, or "false positive determination"). The decision is often made using the p-value: if the p-value is less than the significance level, then the null hypothesis is rejected. The smaller the p-value, the more significant the result is said to be. In more complicated, but practically important cases, the significance level of a test is a probability such that the

probability of making a decision to reject the null hypothesis when the null hypothesis is actually true is no more than the stated probability.

4. Step-by-Step Illustration

This allows for those applications where the probability of deciding to reject may be much smaller than the significance level for some sets of assumptions encompassed within the null hypothesis.

- Alpha is the level of probability at which the null hypothesis will be rejected with confidence. Usually we set $\alpha = .05$
- p is the probability that the null hypothesis is true, in light of the sample data. We usually reject Ho if $p < .05$
- Clearly, they are related in our usage.

5. Type I and Type II Errors

- Type I: Rejecting the null hypothesis when it is true.
- Type II: Retaining the null hypothesis when it is false.
- To decrease Type I, decrease alpha.
- To decrease Type II, increase N.
- As one goes up the other goes down.

6. Test of Difference between Means

- H0: $\mu 1 = \mu 2$
- Find the sample means.
- Find the sample variances.
- Compute the standard error of the difference between means.
- Compute t.
- Compare to critical value of t from the table. (df = N1+ N2 - 2)

- Compare your calculated t to the table t. If calculated t is greater than table t, reject the null hypothesis.
- If calculated t is smaller, retain the null hypothesis.
- This is the most commonly used test of difference between means, and the one we will emphasize in this course. Be able to do this one in your sleep!

Topic Objective:

After reading this topic, student would be able to:

- Discuss the logic of analysis of variance
- Learn about the sum of squares
- Learn about mean square
- Explain step-by-step illustration: analysis of variance
- Learn about the requirements for using the F ratio

Definition/Overview:

In statistics, analysis of variance (ANOVA) is a collection of statistical models, and their associated procedures, in which the observed variance is partitioned into components due to different explanatory variables. The initial techniques of the analysis of variance were developed by the statistician and geneticist R. A. Fisher in the 1920s and 1930s, and is sometimes known as Fisher's ANOVA or Fisher's analysis of variance, due to the use of Fisher's F-distribution as part of the test of statistical significance. A statistically significant effect in ANOVA is often followed up with one or more different follow-up tests. This can be done in order to assess which groups are different from which other groups or to test various other focused hypotheses. Follow up tests are often distinguished in terms of whether they are planned (a priori) or post hoc. Planned tests are determined before looking at the data and post hoc tests are performed after looking at the data. Post hoc tests such as Tukey's test most commonly compare every group mean with every other group mean and typically incorporate some method of controlling of Type

I errors. Comparisons, which are most commonly planned, can be either simple or compound. Simple comparisons compare one group mean with one other group mean. Compound comparisons typically compare two sets of groups means where one set has at two or more groups (e.g., compare average group means of group A, B and C with group D). Comparisons can also look at tests of trend, such as linear and quadratic relationships, when the independent variable involves ordered levels.

Key Points:

1. Analysis Of Variance

In statistics, analysis of variance (ANOVA) is a collection of statistical models, and their associated procedures, in which the observed variance is partitioned into components due to different explanatory variables. The initial techniques of the analysis of variance were developed by the statistician and geneticist R. A. Fisher in the 1920s and 1930s, and is sometimes known as Fisher's ANOVA or Fisher's analysis of variance, due to the use of Fisher's F-distribution as part of the test of statistical significance. In practice, there are several types of ANOVA depending on the number of treatments and the way they are applied to the subjects in the experiment:

- One-way ANOVA is used to test for differences among two or more independent groups. Typically, however, the One-way ANOVA is used to test for differences among at least three groups, since the two-group case can be covered by a T-test. When there are only two means to compare, the T-test and the F-test are equivalent; the relation between ANOVA and t is given by $F = t2$.
- One-way ANOVA for repeated measures is used when the subjects are subjected to repeated measures; this means that the same subjects are used for each treatment. Note that this method can be subject to carryover effects.
- Factorial ANOVA is used when the experimenter wants to study the effects of two or more treatment variables. The most commonly used type of factorial ANOVA is the 22 (read: two by two) design, where there are two independent variables and each variable has two levels or

distinct values. Factorial ANOVA can also be multi-level such as 33, etc. or higher order such as 222, etc. but analyses with higher numbers of factors are rarely done by hand because the calculations are lengthy and the results are hard to interpret. However, since the introduction of data analytic software, the utilization of higher order designs and analyses has become quite common.

- When one wishes to test two or more independent groups subjecting the subjects to repeated measures, one may perform a factorial mixed-design ANOVA, in which one factor is a between-subjects variable and the other is within-subjects variable. This is a type of mixed-effect model.

2. Logic of ANOVA

Partitioning of the sum of squares. The fundamental technique is a partitioning of the total sum of squares into components related to the effects used in the model. For example, we show the model for a simplified ANOVA with one type of treatment at different levels.

The number of degrees of freedom (abbreviated df) can be partitioned in a similar way and specifies the chi-square distribution which describes the associated sums of squares.

3. The F-test

An F-test is any statistical test in which the test statistic has an F-distribution if the null hypothesis is true. The name was coined by George W. Snedecor, in honour of Sir Ronald A. Fisher. Fisher initially developed the statistic as the variance ratio in the 1920s. Examples include:

- The hypothesis that the means of multiple normally distributed populations, all having the same standard deviation, are equal. This is perhaps the most well-known of hypotheses tested by means of an F-test, and the simplest problem in the analysis of variance (ANOVA).
- The hypothesis that a proposed regression model fits well. See Lack-of-fit sum of squares.
- The hypothesis that the standard deviations of two normally distributed populations are equal, and thus that they are of comparable origin.

Note that if it is equality of variances (or standard deviations) that is being tested, the F-test is extremely non-robust to non-normality. That is, even if the data displays only modest departures from the normal distribution, the test is unreliable and should not be used. The F-test is used for comparisons of the components of the total deviation. For example, in one-way, or single-factor ANOVA, statistical significance is tested for by comparing the F test statistic

where:

, I = number of treatments

and

, n_T = total number of cases

to the F-distribution with I-1,n_T degrees of freedom. Using the F-distribution is a natural candidate because the test statistic is the quotient of two mean sums of squares which have a chi-square distribution.

4. ANOVA on Ranks

As first suggested by Conover and Iman in 1981, in many cases when the data do not meet the assumptions of ANOVA, one can replace each original data value by its rank from 1 for the smallest to N for the largest, then run a standard ANOVA calculation on the rank-transformed data. "Where no equivalent nonparametric methods have yet been developed such as for the two-way design, rank transformation results in tests which are more robust to non-normality, and resistant to outliers and non-constant variance, than is ANOVA without the transformation." . However Seaman et al. (1994) noticed that the rank transformation of Conover and Iman (1981) is not appropriate for testing interactions among effects in a factorial design as it can cause an increase in Type I error (alpha error). Furthermore, if both main factors are significant there is little power to detect interactions. A variant of rank-transformation is 'quantile normalization' in which a further transformation is applied to the ranks such that the resulting values have some defined distribution (often a normal distribution with a specified mean and variance). Further

analyses of quantile-normalized data may then assume that distribution to compute significance values.

Topic Objective:

After reading this topic, student would be able to:

- Discuss the Chi-Square Test
- Learn about step-by-step illustration: chi-square test of significance
- Explain step-by-step illustration: comparing several groups
- Explain the median test
- Learn about step-by-step illustration: median test
- Explain testing differences

Definition/Overview:

Non-parametric statistics uses distribution free methods which do not rely on assumptions that the data are drawn from a given probability distribution. As such it is the opposite of parametric statistics. It includes non-parametric statistical models, inference and statistical tests. The term non-parametric statistic can also refer to a statistic (a function on a sample) whose interpretation does not depend on the population fitting any parametrized distributions. Order statistics are one example of such a statistic that plays a central role in many non-parametric approaches. Non-parametric methods are widely used for studying populations that take on a ranked order (such as movie reviews receiving one to four stars). The use of non-parametric methods may be necessary when data has a ranking but no clear numerical interpretation, such as when assessing preferences. As non-parametric methods make fewer assumptions, their applicability is much wider than the corresponding parametric methods. In particular, they may be applied in situations where less is known about the application in question. Also, due to the reliance on fewer assumptions, non-parametric methods are more robust. Another justification for the use of non-

parametric methods is simplicity. In certain cases, even when the use of parametric methods is justified, non-parametric methods may be easier to use. Due both to this simplicity and to their greater robustness, non-parametric methods are seen by some statisticians as leaving less room for improper use and misunderstanding.

Key Points:

1. Chi-Square Test

A chi-square test (also chi-squared or $\chi 2$ test) is any statistical hypothesis test in which the test statistic has a chi-square distribution when the null hypothesis is true, or any in which the probability distribution of the test statistic (assuming the null hypothesis is true) can be made to approximate a chi-square distribution as closely as desired by making the sample size large enough. Some examples of chi-squared tests where the chi-square distribution is only approximately valid:

- Chi-square test, also known as the chi-square goodness-of-fit test or chi-square test for independence. When mentioned without any modifiers or without other precluding context, this test is usually understood.
- Yates' chi-square test, also known as Yates' correction for continuity.
- Mantel-Haenszel chi-square test.
- Linear-by-linear association chi-square test.
- The portmanteau test in time-series analysis, testing for the presence of autocorrelation
- Likelihood-ratio tests in general statistical modelling, for testing whether there is evidence of the need to move from a simple model to a more complicated one (where the simple model is nested within the complicated one).

  One case where the distribution of the test statistic is an exact chi-square distribution is the test that the variance of a normally-distributed population has a given value based on a sample variance. Such a test is uncommon in practice because values of variances to test against are seldom known exactly.

2. Step-By-Step Illustration: Chi-Square Test of Significance

If a sample of size x is taken from a population having a normal distribution, then there is a well-known result (see distribution of the sample variance) which allows a test to be made of whether the variance of the population has a pre-determined value. For example, a manufacturing process might have been in stable condition for a long period, allowing a value for the variance to be determined essentially without error. Suppose that a variant of the process is being tested, giving rise to a small sample of product items whose variation is to be tested. The test statistic T in this instance could be set to be the sum of squares about the sample mean, divided by the nominal value for the variance (ie. the value to be tested as holding). Then T has a chi-square distribution with n1 degrees of freedom. For example if the sample size is 21, the acceptance region for T for a significance level of 5% is the interval 9.59 to 34.17.

3. Chi Square Statistic

- fo = observed frequency
- fe = expected frequency

4. Two-way Chi Square Example

- Null hypothesis: The relative frequency [or percentage] of liberals who are permissive is the same as the relative frequency of conservatives who are permissive.
- Categories (independent variable) are liberals and conservatives. Dependent variable being measured is permissiveness.

- Because we had 20 respondents in each column and each row, our expected values in this cross-tabulation would be 10 cases per cell.
- Note that both rows and columns are nominal data -- which could not be handled by t test or ANOVA. Here the numbers are frequencies, not an interval variable.

- Unfortunately, most examples do not have equal row and column totals, so it is harder to figure out the expected frequencies.
- What frequencies would we see if there were no difference between groups (if the null hypothesis were true)?
- If 25 out of 40 respondents(62.5%) were permissive, and there were no difference between liberals and conservatives, 62.5% of each would be permissive.
- We get the expected frequencies for each cell by multiplying the row marginal total by the column marginal total and dividing the result by N.
- Well put the expected values in parentheses.

- So the chi square statistic, from this data is
- (15-12.5)squared / 12.5 PLUS the same values for all the other cells
- = .5 + .5 + .83 + .83 = 2.66
- df = (r-1) (c-1) , where r = rows, c =columns so df = (2-1)(2-1) = 1
- From Table C, $\alpha = .05$, chi-sq = 3.84
- Compare: Calculate 2.66 is less than table value, so we retain the null hypothesis.

5. Median Test

In statistics, Mood's median test is a special case of chi-square test. It is a nonparametric test that tests the null hypothesis that the medians of the populations from which two samples are drawn are identical. The data in each sample are assigned to two groups, one consisting of data whose values are higher than the median value in the two groups combined, and the other consisting of data whose values are at the median or below. A chi-square test is then used to determine whether the observed frequencies in each group differ from expected frequencies derived from a

distribution combining the two groups. The test has low power (efficiency) for moderate to large sample sizes, and is largely regarded as obsolete. The Wilcoxon-Mann-Whitney U two-sample test should be considered instead. Siegel & Castellan suggest that there is no alternative to the median test when one or more observations are "off the scale." The relevant difference between the two tests is that the median test only considers the position of each observation relative to the overall median, whereas the Wilcoxon-Mann-Whitney test takes the ranks of each observation into account. Thus the latter test is usually the more powerful of the two.

- In Section 4 of this course you will cover these topics:
  - Correlation
  - Regression Analysis

Topic Objective:

After reading this topic, student would be able to:

- Discuss the strength of correlation
- Explain direction of correlation
- Learn about curvilinear correlation
- Explain the correlation coefficient
- Step-by-Step Illustration: Correlation Coefficient.

Definition/Overview:

In probability theory and statistics, correlation, (often measured as a correlation coefficient), indicates the strength and direction of a linear relationship between two random variables. In

general statistical usage, correlation or co-relation refers to the departure of two variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of data. A number of different coefficients are used for different situations. The best known is the product-moment correlation coefficient, which is obtained by dividing the covariance of the two variables by the product of their standard deviations.

Key Points:

1. Strength of Correlation

In probability theory and statistics, correlation (often measured as a correlation coefficient) indicates the strength and direction of a linear relationship between two random variables. That is in contrast with the usage of the term in colloquial speech, denoting any relationship, not necessarily linear. In general statistical usage, correlation or co-relation refers to the departure of two random variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of the data. A number of different coefficients are used for different situations. The best known is the product-moment correlation coefficient, which is obtained by dividing the covariance of the two variables by the product of their standard deviations. Despite its name, it was first introduced by Francis Galton

2. Correlation: Testing Significance

- The calculated r is compared to the table F value.
- As usual, if the calculated r is greater than the tabled r, we reject the null hypothesis.

3. Cross-Correlation

In signal processing, cross-correlation is a measure of similarity of two waveforms as a function of a time-lag applied to one of them. This is also known as a sliding dot product or inner-product. It is commonly used to search a long duration signal for a shorter, known feature. It also has applications in pattern recognition, single particle analysis, electron tomographic averaging, and cryptanalysis. For example, consider two real valued functions f and g that differ only by a shift along the x-axis. One can calculate the cross-correlation to figure out how much g must be shifted along the x-axis to make it identical to f. The formula essentially slides the g function along the x-axis, calculating the integral of their product for each possible amount of sliding. When the functions match, the value of is maximized. The reason for this is that when lumps (positives areas) are aligned, they contribute to making the integral larger. Also, when the troughs (negative areas) align, they also make a positive contribution to the integral because the product of two negative numbers is positive. With complex-valued functions f and g, taking the conjugate of f ensures that aligned lumps (or aligned troughs) with imaginary components will contribute positively to the integral.

4. Coefficient

- Coefficient of Thermal Expansion (thermodynamics) (dimensionless) - Relates the change in temperature to the change in a material's dimensions.
- Partition Coefficient (KD) (chemistry) - The ratio of concentrations of a compound in two phases of a mixture of two immiscible solvents at equilibrium.
- Hall coefficient (electrical physics) - Relates a magnetic field applied to an element to the voltage created, the amount of current and the element thickness. It is a characteristic of the material from which the conductor is made.
- Lift Coefficient (CL or CZ) (Aerodynamics) (dimensionless) - Relates the lift generated by an airfoil with the dynamic pressure of the fluid flow around the airfoil, and the planform area of the airfoil.
- Ballistic coefficient (BC) (Aerodynamics) (units of kg/m2) - A measure of a body's ability to overcome air resistance in flight. BC is a function of mass, diameter, and drag coefficient.

- Transmission Coefficient (quantum mechanics) (dimensionless) - Represents the probability flux of a transmitted wave relative to that of an incident wave. It is often used to describe the probability of a particle tunnelling through a barrier.
- Damping Factor a.k.a. viscous damping coefficient (Physical Engineering) (units of Newton-seconds per meter) - relates a damping force with the velocity of the object whose motion is being dampened.

  In science, a coefficient is the number in front of the chemical formulas. Also, the coefficient indicates how many molecules take part in the reaction.

5. Correlation Coefficient

In probability theory and statistics, correlation (often measured as a correlation coefficient) indicates the strength and direction of a linear relationship between two random variables. That is in contrast with the usage of the term in colloquial speech, denoting any relationship, not necessarily linear. In general statistical usage, correlation or co-relation refers to the departure of two random variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of the data. A number of different coefficients are used for different situations. The best known is the product-moment correlation coefficient, which is obtained by dividing the covariance of the two variables by the product of their standard deviations. Despite its name, it was first introduced by Francis Galton.

Topic Objective:

After reading this topic, student would be able to:

- Discuss the regression model
- Learn about interpreting the linear egression
- Explain regression and correlation
- Learn about step-by-step illustration: step-by-step illustration: regression analysis

Definition/Overview:

In statistics, regression analysis is a collective name for techniques for the modeling and analysis of numerical data consisting of values of a dependent variable (also called response variable or measurement) and of one or more independent variables (also known as explanatory variables or predictors). The dependent variable in the regression equation is modeled as a function of the independent variables, corresponding parameters ("constants"), and an error term. The error term is treated as a random variable. It represents unexplained variation in the dependent variable. The parameters are estimated so as to give a "best fit" of the data. Most commonly the best fit is evaluated by using the least squares method, but other criteria have also been used. Regression can be used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships. These uses of regression rely heavily on the underlying assumptions being satisfied. Regression analysis has been criticized as being misused for these purposes in many cases where the appropriate assumptions cannot be verified to hold. One factor contributing to the misuse of regression is that it can take considerably more skill to critique a model than to fit a model.

Key Points:

1. Regression Model

Regression analysis is a technique used for the modeling and analysis of numerical data consisting of values of a dependent variable (response variable) and of one or more independent variables (explanatory variables). The dependent variable in the regression equation is modeled as a function of the independent variables, corresponding parameters ("constants"), and an error term. The error term is treated as a random variable. It represents unexplained variation in the dependent variable. The parameters are estimated so as to give a "best fit" of the data. Most commonly the best fit is evaluated by using the least squares method, but other criteria have also been used. The earliest form of regression was the method of least squares, which was published

by Legendre in 1805, and by Gauss in 1809. The term least squares is from Legendres term, moindres carrs. However, Gauss claimed that he had known the method since 1795. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun. Euler had worked on the same problem (1748) without success.[citation needed] Gauss published a further development of the theory of least squares in 1821, including a version of the GaussMarkov theorem. The term "regression" was coined by Francis Galton, a cousin of Charles Darwin, in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average. For Galton, regression had only this biological meaning, but his work was later extended by Udny Yule and Karl to a more general statistical context. At the present time, the term "regression" is often synonymous with "least squares curve fitting".

2. Linear Egression

In statistics, linear regression is a form of regression analysis in which the relationship between one or more independent variables and another variable, called dependent variable, is modeled by a least squares function, called linear regression equation. This function is a linear combination of one or more model parameters, called regression coefficients. A linear regression equation with one independent variable represents a straight line. The results are subject to statistical analysis. Classical assumptions for linear regression include the assumptions that the sample is selected at random from the population of interest, that the dependent variable is continuous on the real line, and that the error terms follow identical and independent normal distributions, that is, that the errors are i.i.d. and Gaussian. Note that these assumptions imply that the error term does not statistically depend on the values of the independent variables, that is, that is statistically independent of the predictor variables. This article adopts these assumptions unless otherwise stated. Note that in more advanced treatments all of these assumptions may be relaxed. In particular note that the assumption that the error terms are normally distributed is of no consequence unless the sample is very small because central limit theorems imply that, so long as the error terms have finite variance and are not too strongly

correlated, the parameter estimates will be approximately normally distributed even when the underlying errors are not.

3. Correlation Coefficient

In probability theory and statistics, correlation (often measured as a correlation coefficient) indicates the strength and direction of a linear relationship between two random variables. That is in contrast with the usage of the term in colloquial speech, denoting any relationship, not necessarily linear. In general statistical usage, correlation or co-relation refers to the departure of two random variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of the data. A number of different coefficients are used for different situations. The best known is the product-moment correlation coefficient, which is obtained by dividing the covariance of the two variables by the product of their standard deviations. Despite its name, it was first introduced by Francis Galton.

4. Step-By-Step Illustration: Regression Analysis

In statistics, regression analysis is a collective name for techniques for the modeling and analysis of numerical data consisting of values of a dependent variable (also called response variable or measurement) and of one or more independent variables (also known as explanatory variables or predictors). The dependent variable in the regression equation is modeled as a function of the independent variables, corresponding parameters ("constants"), and an error term. The error term is treated as a random variable. It represents unexplained variation in the dependent variable. The parameters are estimated so as to give a "best fit" of the data. Most commonly the best fit is evaluated by using the least squares method, but other criteria have also been used. Regression can be used for prediction (including forecasting of time-series data), inference, hypothesis testing, and modeling of causal relationships. These uses of regression rely heavily on the underlying assumptions being satisfied. Regression analysis has been criticized as being misused for these purposes in many cases where the appropriate assumptions cannot be verified to hold.

One factor contributing to the misuse of regression is that it can take considerably more skill to critique a model than to fit a model

Test Introduction to Criminal Justice Sta

- In Section 5 of this course you will cover these topics:
  - Nonparametric Measures Of Correlation
  - Applying Statistical Procedures To Research Problems

Topic Objective:

After reading this topic, student would be able to:

- Discuss Spearman's rank-order correlation coefficient
- Learn about step-by-step illustration: Spearman's rank-order correlation coefficient
- Explain Goodman's and Kruskal's Gamma.
- Understand step-by-step illustration: Goodman's and Kruskal's Gamma.
- Discuss correlation coefficient
- Discuss looking at the larger picture: measuring association

Definition/Overview:

Non-parametric statistics uses distribution free methods which do not rely on assumptions that the data are drawn from a given probability distribution. As such it is the opposite of parametric statistics. It includes non-parametric statistical models, inference and statistical tests. The term non-parametric statistic can also refer to a statistic (a function on a sample) whose interpretation does not depend on the population fitting any parametrized distributions. Order statistics are one

example of such a statistic that plays a central role in many non-parametric approaches. Non-parametric methods are widely used for studying populations that take on a ranked order (such as movie reviews receiving one to four stars). The use of non-parametric methods may be necessary when data has a ranking but no clear numerical interpretation, such as when assessing preferences. As non-parametric methods make fewer assumptions, their applicability is much wider than the corresponding parametric methods. In particular, they may be applied in situations where less is known about the application in question. Also, due to the reliance on fewer assumptions, non-parametric methods are more robust. Another justification for the use of non-parametric methods is simplicity. In certain cases, even when the use of parametric methods is justified, non-parametric methods may be easier to use. Due both to this simplicity and to their greater robustness, non-parametric methods are seen by some statisticians as leaving less room for improper use and misunderstanding.

Key Points:

1. Spearman's rank-order correlation coefficient

In statistics, Spearman's rank correlation coefficient or Spearman's rho, named after Charles Spearman and often denoted by the Greek letter ρ (rho) or as rs, is a non-parametric measure of correlation that is, it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables. Basically, there is at least one **nonparametric** equivalent for each parametric general type of test. In general, these tests fall into the following categories:

- Tests of differences between groups (independent samples);
- Tests of differences between variables (dependent samples);
- Tests of relationships between variables.

2. Correlation Coefficient

Correlation coefficient is a parametric statistic and when distributions are not normal it may be less useful than non-parametric correlation methods, such as Chi-square, Point bi-serial correlation, Spearman's ρ and Kendall's τ. They are a little less powerful than parametric methods if the assumptions underlying the latter are met, but are less likely to give distorted results when the assumptions fail.

3. Nonparametric Correlations

The following are three types of commonly used nonparametric correlation coefficients (Spearman R, Kendall Tau, and Gamma coefficients). Note that the chi-square statistic computed for two-way frequency tables, also provides a careful measure of a relation between the two (tabulated) variables, and unlike the correlation measures listed below, it can be used for variables that are measured on a simple nominal scale. Spearman R assumes that the variables under consideration were measured on at least an ordinal (rank order) scale, that is, that the individual observations can be ranked into two ordered series. Spearman R can be thought of as the regular product moment correlation coefficient, that is, in terms of proportion of variability accounted for, except that Spearman R is computed from ranks.

4. Goodman's and Kruskal's Gamma

A gamma test tests the strength of association of the cross tabulated data when both variables are measured at the ordinal level. It makes no adjustment for either table size or ties. Values range from -1 (100% negative association, or perfect inversion) to +1 (100% positive association, or perfect agreement). A value of zero indicates the absence of association. This test statistic is also known as Goodman and Kruskal's gamma (which is distinct from Goodman and Kruskal's lambda). Kendall tau. Kendall tau is equivalent to Spearman R with regard to the underlying assumptions. It is also comparable in terms of its statistical power. However, Spearman R and Kendall tau are usually not identical in magnitude because their underlying logic as well as their computational formulas are very different. More importantly, Kendall tau and Spearman R imply different interpretations: Spearman R can be thought of as the regular product moment

correlation coefficient, that is, in terms of proportion of variability accounted for, except that Spearman R is computed from ranks. Kendall tau, on the other hand, represents a probability, that is, it is the difference between the probability that in the observed data the two variables are in the same order versus the probability that the two variables are in different orders. Gamma. The Gamma statistic is preferable to Spearman R or Kendall tau when the data contain many tied observations. In terms of the underlying assumptions, Gamma is equivalent to Spearman R or Kendall tau; in terms of its interpretation and computation it is more similar to Kendall tau than Spearman R. In short, Gamma is also a probability; specifically, it is computed as the difference between the probability that the rank ordering of the two variables agree minus the probability that they disagree, divided by 1 minus the probability of ties. Thus, Gamma is basically equivalent to Kendall tau, except that ties are explicitly taken into account.

5. Correlation Coefficient

In probability theory and statistics, correlation (often measured as a correlation coefficient) indicates the strength and direction of a linear relationship between two random variables. That is in contrast with the usage of the term in colloquial speech, denoting any relationship, not necessarily linear. In general statistical usage, correlation or co-relation refers to the departure of two random variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of the data. A number of different coefficients are used for different situations. The best known is the product-moment correlation coefficient, which is obtained by dividing the covariance of the two variables by the product of their standard deviations. Despite its name, it was first introduced by Francis Galton.

Several authors have offered guidelines for the interpretation of a correlation coefficient. Cohen (1988), has observed, however, that all such criteria are in some ways arbitrary and should not be observed too strictly. This is because the interpretation of a correlation coefficient depends on the context and purposes. A correlation of 0.9 may be very low if one is verifying a physical law using high-quality instruments, but may be regarded as very high in the social sciences where there may be a greater contribution from complicating factors. Along this vein, it is important to remember that "large" and "small" should not be taken as synonyms for "good" and "bad" in

terms of determining that a correlation is of a certain size. For example, a correlation of 1.0 or −1.0 indicates that the two variables analyzed are equivalent modulo scaling. Scientifically, this more frequently indicates a trivial result than a profound one. For example, consider discovering a correlation of 1.0 between how many feet tall a group of people are and the number of inches from the bottom of their feet to the top of their heads.

Topic Objective:

After reading this topic, student would be able to:

- Discuss research situations
- Explain research solutions

Definition/Overview:

Research is defined as human activity based on intellectual application in the investigation of matter. The primary purpose for applied research is discovering, interpreting, and the development of methods and systems for the advancement of human knowledge on a wide variety of scientific matters of our world and the universe. Research can use the scientific method, but need not do so. Scientific research relies on the application of the scientific method, a harnessing of curiosity. This research provides scientific information and theories for the explanation of the nature and the properties of the world around us. It makes practical applications possible. Scientific research is funded by public authorities, by charitable organizations and by private groups, including many companies. Scientific research can be subdivided into different classifications according to their academic and application disciplines. In statistics, efficiency is a term used in the comparison of various statistical procedures and, in particular, it refers to a measure of the desirability of an estimator or of an experimental design. The relative efficiency of two procedures is the ratio their efficiencies, although often this term is used where the comparison is made between a given procedure and a notional "best possible" procedure. The efficiencies and the relative efficiency of two procedures theoretically depend on

the sample size available for the given procedure, but it is often possible to use the asymptotic relative efficiency (defined as the limit of the relative efficiencies as the sample size grows) as the principal comparison measure. Efficiencies are often defined using the variance or mean square error as the measure of desirability. However, for comparing significance tests, a meaningful measure can be defined based on the sample size required for the test to achieve a given power. A statistical parameter is a parameter that indexes a family of probability distributions. It can be regarded as a numerical characteristic of a population or a model. Among parameterized families of distributions are the normal distributions, the Poisson distributions, the binomial distributions, and the exponential distributions. The family of normal distributions has two parameters, the mean and the variance: if these are specified, the distribution is known exactly. The family of chi-squared distributions, on the other hand, has only one parameter, the number of degrees of freedom. In statistical inference, parameters are sometimes taken to be unobservable, and in this case the statistician's task is to infer what he or she can about the parameter based on observations of random variables distributed according to the probability distribution in question, or, more concretely stated, based on a random sample taken from the population of interest. In other situations, parameters may be fixed by the nature of the sampling procedure used or the kind of statistical procedure being carried out (for example, the number of degrees of freedom in a chi-squared test). Even if a family of distributions is not specified, quantities such as the mean and variance can still be regarded as parameters of the distribution of the population from which a sample is drawn. Statistical procedures can still attempt to make inferences about such population parameters.

Key Points:

1. Research Situations

Basic research (also called fundamental or pure research) has as its primary objective the advancement of knowledge and the theoretical understanding of the relations among variables. It is exploratory and often driven by the researchers curiosity, interest, and intuition. Therefore, it is sometimes conducted without any practical end in mind, although it may have unexpected results

pointing to practical applications. The terms basic or fundamental indicate that, through theory generation, basic research provides the foundation for further, sometimes applied research. As there is no guarantee of short-term practical gain, researchers may find it difficult to obtain funding for basic research. Examples of questions asked in basic research:

- Does string theory provide physics with a grand unification theory?
- Which aspects of genomes explain organismal complexity?
- Is it possible to prove or disprove Goldbach's conjecture? (i.e., that every even integer greater than 2 can be written as the sum of two, not necessarily distinct primes)

Traditionally, basic research was considered as an activity that preceded applied research, which in turn preceded development into practical applications. Recently, these distinctions have become much less clear-cut, and it is sometimes the case that all stages will intermix. This is particularly the case in fields such as biotechnology and electronics, where fundamental discoveries may be made alongside work intended to develop new products, and in areas where public and private sector partners collaborate in order to develop greater insight into key areas of interest. For this reason, some now prefer the term frontier research.

Scientific method refers to bodies of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge. To be termed scientific, a method of inquiry must be based on gathering observable, empirical and measurable evidence subject to specific principles of reasoning. A scientific method consists of the collection of data through observation and experimentation, and the formulation and testing of hypotheses. Although procedures vary from one field of inquiry to another, identifiable features distinguish scientific inquiry from other methodologies of knowledge. Scientific researchers propose hypotheses as explanations of phenomena, and design experimental studies to test these hypotheses. These steps must be repeatable in order to dependably predict any future results. Theories that encompass wider domains of inquiry may bind many hypotheses together in a coherent structure. This in turn may help form new hypotheses or place groups of hypotheses into context. Among other facets shared by the various fields of inquiry is the conviction that the process be objective to reduce a biased interpretation of the results. Another basic expectation is to document, archive and share all data and methodology so they are available for careful scrutiny by other scientists, thereby allowing other researchers the opportunity to verify results by attempting to reproduce

them. This practice, called full disclosure, also allows statistical measures of the reliability of these data to be established.

2. Research Solutions

A research uses different methods to provide the intended solutions. Basic research (also called fundamental or pure research) has as its primary objective the advancement of knowledge and the theoretical understanding of the relations among variables. It is exploratory and often driven by the researchers curiosity, interest, and intuition. Therefore, it is sometimes conducted without any practical end in mind, although it may have unexpected results pointing to practical applications. The terms basic or fundamental indicate that, through theory generation, basic research provides the foundation for further, sometimes applied research. As there is no guarantee of short-term practical gain, researchers may find it difficult to obtain funding for basic research. Examples of questions asked in basic research:

- Does string theory provide physics with a grand unification theory?
- Which aspects of genomes explain organismal complexity?
- Is it possible to prove or disprove Goldbach's conjecture? (i.e., that every even integer greater than 2 can be written as the sum of two, not necessarily distinct primes) Traditionally, basic research was considered as an activity that preceded applied research, which in turn preceded development into practical applications.

Recently, these distinctions have become much less clear-cut, and it is sometimes the case that all stages will intermix. This is particularly the case in fields such as biotechnology and electronics, where fundamental discoveries may be made alongside work intended to develop new products, and in areas where public and private sector partners collaborate in order to develop greater insight into key areas of interest. For this reason, some now prefer the term frontier research.

The historical method comprises the techniques and guidelines by which historians use historical sources and other evidence to research and then to write history. There are various history

guidelines commonly used by historians in their work, under the headings of external criticism, internal criticism, and synthesis. This includes higher criticism and textual criticism.

Most funding for scientific research comes from two major sources, corporations (through research and development departments) and government (primarily through universities and in some cases through military contractors). Many senior researchers (such as group leaders) spend more than a trivial amount of their time applying for grants for research funds. These grants are necessary not only for researchers to carry out their research, but also as a source of merit. Some faculty positions require that the holder has received grants from certain institutions, such as the US National Institutes of Health (NIH). Government-sponsored grants (e.g. from the NIH, the National Health Service in Britain or any of the European research councils) generally have a high status.

Test Introduction to Criminal Justice Statistics