Course Name	**: Basic Statistics**
Course Code	**: BIRD 223**
Course Level	**: Level 4**
Credit Units	**: 4 CU**
Contact Hours	**: 60 Hrs**

Course Description

The Course encompasses different forms of statistics, the appropriate methods of calculating the central tendency, understanding how to estimate various scales in determining range, use of variations and sequences, standard deviation and statistics related to cross tabulation.

Course objectives

- To equip students with analytical skills and statistical concepts useful in decision making.
- To improve their knowledge of describing and interpreting statistical records.
- To enable them get firm exposure to data collection, presentation, and analysis and interpretation for rational decisions on crucial matters.

Course Content

Introduction to statistics

- Definition of Statistics
- Common uses of statistics
- Relevance of statistics in an economy
- Types of statistics i.e descriptive, inferential statistics

Methods of calculating the central tendency

- Mean
- Mode
- Median
- Average

Estimates of scale

- Standard deviation
- Interquartile range
- Range
- Mean difference
- Median absolute deviation

- Average absolute deviation
- Sources of statistical dispersion

Statistics related to cross tabulation

- Chi-square
- Contingency coefficient
- Cramer's V
- Lambda coefficient
- Phi coefficient
- Kendall tau

Statistical Inference

- Definition of statistical inference
- Exploratory data analysis
- Exploratory data Analysis Development(EDAD)

Variance

- Definition of variance
- Forms of variance i.e continuous case, discrete case
- Approximating the variance of a function
- Distinguish between population and variance and sample variance
- Generalizations of variances

Skewness

- Definition of Skewness
- Forms of Skewness ie Sample Skewness, kurtosis
- Sample kurtosis
- Formulas for calculating kurtosis ie mean absolute error, interquartile range,

Standard deviation

- Definition of standard deviation
- Probability distribution or random variable
- Steps in calculating standard deviation
- Simplification of the formula
- Estimating population standard deviation

Mode of delivery Face to face lectures

Assessment

Course work 40%

Exams 60%

Total Mark 100%

Descriptive statistics

STATISTICS – a body of principles and methods of extracting inform ation from numerical data. It is divided into two broad categories: inferential and descriptive statistics.

Descriptive statistics – the methods of organizing, summarizing and presenting data in convenient meaningful and easy to interpret forms e.g tables, graphs, charts, averages, variations from averages. Are used to describe the main features of a collection of data in quantitative terms. Descriptive statistics are distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aim to quantitatively summarize a data set, rather than being used to support inferential statements about the population that the data are thought to represent. Even when a data analysis draws its main conclusions using inductive statistical analysis, descriptive statistics are generally presented along with more formal analyses, to give the audience an overall sense of the data being analyzed.

Common uses

A common example of the use of descriptive statistics occurs in medical research studies. In a paper reporting on a study involving human subjects, there typically appears a table giving the overall sample size, sample sizes in important subgroups (e.g. for each treatment or exposure group), and demographic or clinical characteristics such as the average age, the proportion of subjects with each gender, and the proportion of subjects with related comorbidities.

In research involving comparisons between groups, a major emphasis is often placed on the significance level for the hypothesis that the groups being compared differ to a greater degree than would be expected by chance. This significance level is often represented as a p-value, or sometimes as the standard score of a test statistic. In contrast, an effect size is a descriptive statistic that conveys the estimated magnitude and direction of the difference between groups, without regard to whether the difference is statistically significant. Reporting significance levels without effect sizes is often criticized, since for large sample sizes even small effects of little practical importance can be highly statistically significant.

Examples of descriptive statistics

Most statistics can be used either as a descriptive statistic, or in an inductive analysis. For example, we can report the average reading test score for the students in each classroom in a school, to give a descriptive sense of the typical scores and their variation. If we perform a formal hypothesis test on the scores, we are doing inductive rather than descriptive analysis.

Some statistical summaries are especially common in descriptive analyses. Some examples follow.

- Measures of central tendency , Measures of dispersion , Measures of association , Cross-tab, contingency table , Histogram , Quantile, Q-Q plot , Scatterplot , Box plot .

INFERENTIAL STATISTICS – that body of methods used to draw conclusions about characteristics of a population based on information available from a sample taken scientifically from that population, e.g., given that UTL has 70,000 subscribers and 30,000 potential subscribers. If UTL wanted to introduce new communication methods, it my sample only 10% of its population. This leaves a chance of making errors.

However, statistical methods have ways of determining the reliability of statistical inference

Reliability: where a repeated measurement gives a similar if not exact value as before. This depends on: - the tool of measurement, -the competence of the person doing the measurement and -the consistency of the data.

The sample should be drawn using a probabilistic method (representative and luck of bias). Non probabilistic methods is not applicable in inferential statistics.

Population: is the total set of elements or characters under obsevation/study. It may consist of measurements, companies, a set of accounts, etc.

Sample: is a sub set of a population and a descriptive measure of the sample is known as a **statistic.**

ESTIMATION

Samples estimate parameters. Whereas a parameter for a specific population is a constant, a statistic is a variable.

The process of estimating, forecasting or making decisions about the population from sample information is called statistical inference and is the primary purpose of statistics. Populations are often large making it impractical to inquire from every member of a population.

Decision makers use statistics to estimate parameters. because of the uncertainty sorrounding the estimation techniques, each statistic must be acconpanied by a measure of reliability of inference.

Average

In mathematics, an **average**, **central tendency**[1] of a data set is 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. These include means, the median and the mode. Other statistical measures such as the standard deviation and the range are called measures of spread and describe how spread out the data is.

An average is a single value that is meant to typify a list of values. If all the numbers in the list are the same, then this number should be used. If the numbers are not all the same, an easy way to get a representative value from a list is to randomly pick any number from the list. However, the word 'average' is usually reserved for more sophisticated methods that are generally found to be more useful. In the latter case, the average is calculated by combining the values from the set in a specific way and computing a single number as being the average of the set.

The most common method is the arithmetic mean but there are many other types of averages, such as median (which is used most often when the distribution of the

values is skewed with some small numbers of very high values, as seen with house prices or incomes).[2]

Calculation

Arithmetic mean

Main article: Arithmetic mean

If n numbers are given, each number denoted by a_i, where $i = 1, \ldots, n$, the arithmetic mean is the [sum] of the a_i's divided by n or

The arithmetic mean, often simply called the mean, of two numbers, such as 2 and 8, is obtained by finding a value A such that 2 + 8 = A + A. One may find that $A = (2 + 8)/2 = 5$. Switching the order of 2 and 8 to read 8 and 2 does not change the resulting value obtained for A. The mean 5 is not less than the minimum 2 nor greater than the maximum 8. If we increase the number of terms in the list for which we want an average, we get, for example, that the arithmetic mean of 2, 8, and 11 is found by solving for the value of A in the equation $2 + 8 + 11 = A + A + A$. One finds that $A = (2 + 8 + 11)/3 = 7$.

Changing the order of the three members of the list does not change the result: A (8 + 11 + 2)/3 = 7 and that 7 is between 2 and 11. This summation method is easily generalized for lists with any number of elements. However, the mean of a list of integers is not necessarily an integer. "The average family has 1.7 children" is a jarring way of making a statement that is more appropriately expressed by "the average number of children in the collection of families examined is 1.7".

Geometric mean

The geometric mean of n numbers is obtained by multiplying them all together and then taking the nth root. In algebraic terms, the geometric mean of $a_1, a_2, \ldots, a_n$ is defined as

Geometric mean can be thought of as the antilog of the arithmetic mean of the logs of the numbers.

Example: Geometric mean of 2 and 8 is

Harmonic mean

Harmonic mean for a set of numbers $a_1, a_2, \ldots, a_n$ is defined as the reciprocal of the arithmetic mean of the reciprocals of a_i's:

One example where it is useful is calculating the average speed. For example, if the speed for going from point A to B was 60 km/h, and the speed for returning from B to A was 40 km/h, then the average speed is given by

Inequality concerning AM, GM, and HM

A well known inequality concerning arithmetic, geometric, and harmonic means for any set of positive numbers is

It is easy to remember noting that the alphabetical order of the letters A, G, and H is preserved in the inequality. See Inequality of arithmetic and geometric means.

Mode and median

The most frequently occurring number in a list is called the mode. The mode of the list (1, 2, 2, 3, 3, 3, 4) is 3. The mode is not necessarily well defined, the list (1, 2, 2, 3, 3, 5) has the two modes 2 and 3. The mode can be subsumed under the general

method of defining averages by understanding it as taking the list and setting each member of the list equal to the most common value in the list if there is a most common value. This list is then equated to the resulting list with all values replaced by the same value. Since they are already all the same, this does not require any change. The mode is more meaningful and potentially useful if there are many numbers in the list, and the frequency of the numbers progresses smoothly (e.g., if out of a group of 1000 people, 30 people weigh 61 kg, 32 weigh 62 kg, 29 weigh 63 kg, and all the other possible weights occur less frequently, then 62 kg is the mode).

The mode has the advantage that it can be used with non-numerical data (e.g., red cars are most frequent), whilst other averages cannot.

The median is the middle number of the group when they are ranked in order. (If there are an even number of numbers, the mean of the middle two is taken.)

Thus to find the median, order the list according to its elements' magnitude and then repeatedly remove the pair consisting of the highest and lowest values until either one or two values are left. If exactly one value is left, it is the median; if two values, the median is the arithmetic mean of these two. This method takes the list 1, 7, 3, 13 and orders it to read 1, 3, 7, 13. Then the 1 and 13 are removed to obtain the list 3, 7. Since there are two elements in this remaining list, the median is their arithmetic mean, (3 + 7)/2 = 5.

Definitions

Mean

Mode The most frequent value in the data set

Median The middle value that separates the higher half from the lower half of the data set

Truncated Mean The arithmetic mean of data values after a certain number or proportion of the highest and lowest data values have been discarded

Interquartile mean A special case of the truncated mean, using the interquartile range

Winsorized mean Similar to the truncated mean, but, rather than deleting the extreme values, they are set equal to the largest and smallest values that remain

Geometric mean A rotation invariant extension of the median for points in R^n

Solutions to variational problems

Several measures of central tendency can be characterized as solving a variational problem, in the sense of the calculus of variations, namely minimizing variation from the center. That is, given a measure of statistical dispersion, one asks for a measure of central tendency that minimizes variation: such that variation from the center is minimal among all choices of center. In a quip, "dispersion precedes location". In the sense of L^p spaces, the correspondence is:

L^p	Dispersion	central tendency

L^1	average absolute deviation	median
L^2	standard deviation	mean
L^∞	maximum deviation	midrange

Thus standard deviation about the mean is lower than standard deviation about any other point, and the maximum deviation about the midrange is lower than the maximum deviation about any other point. The uniqueness of this characterization of mean follows from convex optimization. Indeed, for a given (fixed) data set x, the function represents the dispersion about a constant value c relative to the L^2 norm. Because the function f_2 is a strictly convex coercive function, the minimizer exists and is unique.

Note that the median in this sense is not in general unique, and in fact any point between the two central points of a discrete distribution minimizes average absolute deviation. The dispersion in the L^1 norm, given by

is not *strictly* convex, whereas strict convexity is needed to ensure uniqueness of the minimizer. In spite of this, the minimizer is unique for the L^∞ norm.

Miscellaneous types

Other more sophisticated averages are: trimean, trimedian, and normalized mean.

One can create one's own average metric using generalized f-mean:

where f is any invertible function. The harmonic mean is an example of this using $f(x) = 1/x$, and the geometric mean is another, using $f(x) = \log x$. Another example, expmean (exponential mean) is a mean using the function $f(x) = e^x$, and it is inherently biased towards the higher values. However, this method for generating means is not general enough to capture all averages. A more general method for defining an average, y, takes any function of a list $g(x_1, x_2, \ldots, x_n)$, which is symmetric under permutation of the members of the list, and equates it to the same function with the value of the average replacing each member of the list: $g(x_1, x_2, \ldots, x_n) = g(y, y, \ldots, y)$. This most general definition still captures the important property of all averages that the average of a list of identical elements is that element itself. The function $g(x_1, x_2, \ldots, x_n) = x_1 + x_2 + \ldots + x_n$ provides the arithmetic mean. The function $g(x_1, x_2, \ldots, x_n) = x_1 \cdot x_2 \cdot \ldots \cdot x_n$ provides the geometric mean. The function $g(x_1, x_2, \ldots, x_n) = x_1^{-1} + x_2^{-1} + \ldots + x_n^{-1}$ provides the harmonic mean. (See John Bibby (1974) "Axiomatisations of the average and a further generalisation of monotonic sequences," Glasgow Mathematical Journal, vol. 15, pp. 63–65.)

In data streams

The concept of an average can be applied to a stream of data as well as a bounded set, the goal being to find a value about which recent data is in some way clustered. The stream may be distributed in time, as in samples taken by some data acquisition system from which we want to remove noise, or in space, as in pixels in an image from which we want to extract some property. An easy-to-understand and widely used application of average to a stream is the simple moving average in which we compute the arithmetic mean of the most recent N data items in the stream. To advance one position in the stream, we add 1/N times the new data item and subtract 1/N times the data item N places back in the stream.

Averages of functions

The concept of average can be extended to functions.[3] In calculus, the average value of an integrable function f on an interval $[a,b]$ is defined by

Etymology

An early meaning (c. 1500) of the word *average* is "damage sustained at sea". The root is found in Arabic as *awar*, in Italian as *avaria* and in French as *avarie*. Hence an *average adjuster* is a person who assesses an insurable loss.

Marine damage is either *particular average*, which is borne only by the owner of the damaged property, or general average, where the owner can claim a proportional contribution from all the parties to the marine venture. The type of calculations used in adjusting general average gave rise to the use of "average" to mean "arithmetic mean".

However, according to the Oxford English Dictionary, the earliest usage in English (1489 or earlier) appears to be an old legal term for a tenant's day labour obligation to a sheriff, probably anglicised from "avera" found in the English Domesday Book (1085). This pre-existing term thus lay to hand when an equivalent for *avarie* was wanted.

Statistical dispersion

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 interquartile range.

Dispersion is contrasted with location or central tendency, and together they are the most used properties of distributions.

Measures of statistical dispersion

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 metres or seconds, the measure of dispersion has the same units. Such measures of dispersion include:

ESTIMATES OF SCALE

- Standard deviation
- Interquartile range
- Range
- Mean difference
- Median absolute deviation
- Average absolute deviation (or simply called average deviation)

These are frequently used (together with scale factors) as estimators of scale parameters, in which capacity they are called **estimates of scale.**

All the above measures of statistical dispersion have the useful property that they are **location-invariant**, as well as linear in scale. So if a random variable X has a

dispersion of S_X then a linear transformation $Y = aX + b$ for real a and b should have dispersion $S_Y = |a| S_X$.

Other measures of dispersion are **dimensionless (scale-free)**. In other words, they have no units even if the variable itself has units. These include:

- Coefficient of variation
- Quartile coefficient of dispersion
- Relative mean difference, equal to twice the Gini coefficient

There are other measures of dispersion:

- Variance (the square of the standard deviation) — location-invariant but not linear in scale.
- Variance-to-mean ratio — mostly used for count data when the term coefficient of dispersion is used and when this ratio is dimensionless, as count data are themselves dimensionless: otherwise this is not scale-free.

Some measures of dispersion have specialized purposes, among them the Allan variance and the Hadamard variance.

For categorical variables, it is less common to measure dispersion by a single number. See qualitative variation. One measure which does so is the discrete entropy.

Sources of statistical dispersion

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.

Association (statistics)

In statistics, an **association** is any relationship between two measured quantities that renders them statistically dependent.[1] The term "association" refers broadly to any such relationship, whereas the narrower term "correlation" refers to a linear relationship between two quantities.

There are many statistical measures of association that can be used to infer the presence or absence of an association in a sample of data. Examples of such measures include the product moment correlation coefficient, used mainly for quantitative measurements, and the odds ratio, used for dichotomous measurements. Other measures of association are the tetrachoric correlation coefficient and Goodman and Kruskal's lambda

In quantitative research, the term "association" is often used to emphasize that a relationship being discussed is not necessarily causal (see correlation does not imply causation).

Cross tabulation

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.

The following is a fictitious example of a 3 × 2 contingency table. The variable "Wikipedia usage" has three categories: heavy user, light user, and non user. These categories are all inclusive so the columns sum to 100%. The other variable "underpants" has two categories: boxers, and briefs. These categories are not all inclusive so the rows need not sum to 100%. Each cell gives the percentage of subjects who share that combination of traits.

	boxers	briefs
heavy Wiki user	70%	5%
light Wiki user	25%	35%
non Wiki user	5%	60%

Cross tabs are frequently used because:

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

Statistics related to cross tabulations

The following list is not comprehensive.

- **Chi-square** - This tests the statistical significance of the cross tabulations. Chi-squared should not be calculated for percentages. The cross tabs must be converted back to absolute counts (numbers) before calculating chi-squared. Chi-squared is also problematic when any cell has a joint frequency of less than five. For an in-depth discussion of this issue see Fienberg, S.E. (1980). "The Analysis of Cross-classified Categorical Data." 2nd Edition. M.I.T. Press, Cambridge, MA.
- **Contingency coefficient** - This tests the strength of association of the cross tabulations. It is a variant of the **phi coefficient** that adjusts for statistical significance. Values range from 0 (no association) to 1 (the theoretical maximum possible association).

- **Cramer's V** - This tests the strength of association of the cross tabulations. It is a variant of the **phi coefficient** that adjusts for the number of rows and columns. Values range from 0 (no association) to 1 (the theoretical maximum possible association).
- **Lambda coefficient** — This tests the strength of association of the cross tabulations when the variables are measured at the nominal level. Values range from 0 (no association) to 1 (the theoretical maximum possible association). **Asymmetric lambda** measures the percentage improvement in predicting the dependent variable. **Symmetric lambda** measures the percentage improvement when prediction is done in both directions.
- **phi coefficient** - If both variables instead are nominal and dichotomous, phi coefficient is a measure of the degree of association between two binary variables. This measure is similar to the correlation coefficient in its interpretation. Two binary variables are considered positively associated if most of the data falls along the diagonal cells. In contrast, two binary variables are considered negatively associated if most of the data falls off the diagonal.
- **Kendall tau**:
    - **Tau b** - This tests the strength of association of the cross tabulations when both variables are measured at the ordinal level. It makes adjustments for ties and is most suitable for square tables. 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.
    - **Tau c** - This tests the strength of association of the cross tabulations when both variables are measured at the ordinal level. It makes adjustments for ties and is most suitable for rectangular tables. 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.
- **Gamma** - This tests the strength of association of the cross tabulations 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.
- **Uncertainty coefficient**, **entropy coefficient** or **Theil's U**

Histogram

In statistics, a **histogram** is a graphical display of tabulated frequencies, shown as bars. It shows what proportion of cases fall into each of several categories: it is a form of data binning. The categories are usually specified as non-overlapping intervals of some variable. The categories (bars) must be adjacent. The intervals (or bands, or bins) are generally of the same size.[1]

Histograms are used to plot density of data, and often for density estimation: estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the x-axis are all 1, then a histogram is identical to a relative frequency plot.

An alternative to the histogram is kernel density estimation, which uses a kernel to smooth samples. This will construct a smooth probability density function, which will in general more accurately reflect the underlying variable.

The histogram is one of the seven basic tools of quality control, which also include the Pareto chart, check sheet, control chart, cause-and-effect diagram, flowchart, and scatter diagram.

Etymology

The word *histogram* derived from the Greek *histos* 'anything set upright' (as the masts of a ship, the bar of a loom, or the vertical bars of a histogram); and *gramma* 'drawing, record, writing'. The term was introduced by Karl Pearson in 1895.[2]
Examples

As an example we consider data collected by the U.S. Census Bureau on time to travel to work (2000 census, [1], Table 2). The census found that there were **124 million people** who work outside of their homes. This rounding is a common phenomenon when collecting data from people.

This histogram shows the number of cases per unit interval so that the height of each bar is equal to the proportion of total people in the survey who fall into that category. The area under the curve represents the total number of cases (124 million). This type of histogram shows absolute numbers.

In other words a histogram represents a frequency distribution by means of rectangles whose widths represent class intervals and whose areas are proportional to the corresponding frequencies. They only place the bars together to make it easier to compare data.

Activities and demonstrations

The SOCR resource pages contain a number of hands-on interactive activities demonstrating the concept of a histogram, histogram construction and manipulation using Java applets and charts.

Mathematical definition

An ordinary and a cumulative histogram of the same data. The data shown is a random sample of 10,000 points from a normal distribution with a mean of 0 and a standard deviation of 1.

In a more general mathematical sense, a histogram is a mapping m_i that counts the number of observations that fall into various disjoint categories (known as *bins*), whereas the graph of a histogram is merely one way to represent a histogram. Thus,

if we let n be the total number of observations and k be the total number of bins, the histogram m_i meets the following conditions:

Cumulative histogram

A cumulative histogram is a mapping that counts the cumulative number of observations in all of the bins up to the specified bin. That is, the cumulative histogram M_i of a histogram m_i is defined as:

Number of bins and width

There is no "best" number of bins, and different bin sizes can reveal different features of the data. Some theoreticians have attempted to determine an optimal number of bins, but these methods generally make strong assumptions about the shape of the distribution. You should always experiment with bin widths before choosing one (or more) that illustrate the salient features in your data.

The number of bins k can be calculated directly, or from a suggested bin width h:

The braces indicate the ceiling function.

Sturges' formula[3]

,

which implicitly bases the bin sizes on the range of the data, and can perform poorly if $n < 30$.

Scott's choice[4]

,

where σ is the sample standard deviation.

Freedman-Diaconis' choice[5]

,

which is based on the interquartile range.

Continuous data

The idea of a histogram can be generalized to continuous data. Let (see Lebesgue space), then the cumulative histogram operator H can be defined by:

$H(f)(y)$ = with only finitely many intervals of monotony this can be rewritten as

$h(f)(y)$ is undefined if y is the value of a stationary point.

Density estimation

- Kernel density estimation, a smoother but more complex method of density estimation

Quantile

Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. Dividing ordered data into q essentially equal-sized data subsets is the motivation for q-quantiles; the quantiles are the data values marking the boundaries between consecutive subsets. Put another way, the kth q-quantile for a random variable is the value x such that the probability that the random variable will be less than x is at most k / q and the probability that the

random variable will be more than x is at most $(q - k) / q$. There are $q - 1$ quantiles, with k an integer satisfying $0 < k < q$.

Median of the order statistics

Alternatively, one may use estimates of the *median* of the order statistics, which one can compute based on estimates of the median of the order statistics of a uniform distribution and the quantile function of the distribution; this was suggested by (Filliben 1975).[3]

This can be easily generated for any distribution for which the quantile function can be computed, but conversely the resulting estimates of location and scale are no longer precisely the least squares estimates, though these only differ significantly for n small.

Statistical inference

Statistical inference or **statistical induction** comprises the use of statistics and random sampling to make inferences concerning some unknown aspect of a population. It is distinguished from descriptive statistics.

Two schools of statistical inference are frequency probability and Bayesian inference.

Definition

Statistical inference is inference about a population from a random sample drawn from it or, more generally, about a random process from its observed behavior during a finite period of time. It includes:

1. Point estimation
2. Interval estimation
3. Hypothesis testing (or statistical significance testing)
4. Prediction – see predictive inference

There are several distinct schools of thought about the justification of statistical inference. All are based on some idea of what real world phenomena can be reasonably modeled as probability.

1. Frequency probability
2. Bayesian probability
3. Fiducial probability

The topics below are usually included in the area of **statistical inference**.

1. Statistical assumptions
2. Statistical decision theory
3. Estimation theory
4. Statistical hypothesis testing
5. Revising opinions in statistics
6. Design of experiments, the analysis of variance, and regression

7. Survey sampling
8. Summarizing statistical data

Exploratory data analysis

From Wikipedia, the free encyclopedia

Jump to: navigation, search

Exploratory data analysis (EDA) is an approach to analyzing data for the purpose of formulating hypotheses worth testing, complementing the tools of conventional statistics for testing hypotheses[1]. It was so named by John Tukey to contrast with Confirmatory Data Analysis, the term used for the set of ideas about hypothesis testing, p-values, confidence intervals etc. which formed the key tools in the arsenal of practicing statisticians at the time.

EDA development

Tukey held that too much emphasis in statistics was placed on statistical hypothesis testing (confirmatory data analysis); more emphasis needed to be placed on using data to suggest hypotheses to test. In particular, he held that confusing the two types of analyses and employing them on the same set of data can lead to systematic bias owing to the issues inherent in testing hypotheses suggested by the data.

The objectives of EDA are to:

- Suggest hypotheses about the causes of observed phenomena
- Assess assumptions on which statistical inference will be based
- Support the selection of appropriate statistical tools and techniques
- Provide a basis for further data collection through surveys or experiments

Many **EDA** techniques have been adopted into data mining and are being taught to young students as a way to introduce them to statistical thinking.[2]

Variance

In probability theory and statistics, the **variance** of a random variable, probability distribution, or sample is a measure of statistical dispersion, averaging the squares of the deviations of its possible values from its 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.

Definition

If a random variable X has expected value (mean) $\mu = E(X)$, then the variance $\mathrm{Var}(X)$ of X is given by:

This definition encompasses random variables that are discrete, continuous, or neither. Of all the points about which squared deviations could have been calculated, the mean produces the minimum value for the averaged sum of squared deviations.

This definition is expanded as follows:

The variance of random variable X is typically designated as $\operatorname{Var}(X)$, , or simply σ^2 (pronounced "sigma squared"). If a distribution does not have an expected value, as is the case for the Cauchy distribution, it does not have a variance either. Many other distributions for which the expected value does exist do not have a finite variance because the relevant integral diverges. An example is a Pareto distribution whose Pareto index k satisfies $1 < k \leq 2$.

Continuous case

If the random variable X is continuous with probability density function $p(x)$,

where

and where the integrals are definite integrals taken for x ranging over the range of X.

Discrete case

If the random variable X is discrete with probability mass function

$x_1 \mapsto p_1, \ldots, x_n \mapsto p_n$, then

(When such a discrete weighted variance is specified by weights whose sum is not 1, then one divides by the sum of the weights.) That is, it is the expected value of the square of the deviation of X from its own mean. In plain language, it can be expressed as "The average of the square of the distance of each data point from the mean". It is thus the *mean squared deviation*.

Examples

Exponential distribution

The exponential distribution with parameter λ is a continuous distribution whose support is the semi-infinite interval $[0,\infty)$. Its probability density function is given by:

and it has expected value $\mu = \lambda^{-1}$. Therefore the variance is equal to:

So for an exponentially distributed random variable $\sigma^2 = \mu^2$.

Fair die

A six-sided fair die can be modelled with a discrete random variable with outcomes 1 through 6, each with equal probability $1/6$. The expected value is $(1+2+3+4+5+6)/6 = 3.5$. Therefore the variance can be computed to be:

Properties

Variance is non-negative because the squares are positive or zero. The variance of a constant random variable is zero, and the variance of a variable in a data set is 0 if and only if all entries have the same value.

Variance is invariant with respect to changes in a location parameter. That is, if a constant is added to all values of the variable, the variance is unchanged. If all values are scaled by a constant, the variance is scaled by the square of that constant. These two properties can be expressed in the following formula:

The variance of a finite sum of **uncorrelated** random variables is equal to the sum of their variances. This stems from the identity:

and that for uncorrelated variables covariance is zero.

In general, for the sum of N variables: , we have:

1. Suppose that the observations can be partitioned into equal-sized **subgroups** according to some second variable. Then the variance of the total group is equal to the mean of the variances of the subgroups plus the variance of the means of the subgroups. This property is known as variance decomposition or the law of total variance and plays an important role in the analysis of variance. For example, suppose that a group consists of a subgroup of men and an equally large subgroup of women. Suppose that the men have a mean body length of 180 and that the variance of their lengths is 100. Suppose that the women have a mean length of 160 and that the variance of their lengths is 50. Then the mean of the variances is (100 + 50) / 2 = 75; the variance of the means is the variance of 180, 160 which is 100. Then, for the total group of men and women combined, the variance of the body lengths will be 75 + 100 = 175. Note that this uses N for the denominator instead of N − 1.
   In a more general case, if the subgroups have unequal sizes, then they must be weighted proportionally to their size in the computations of the means and variances. The formula is also valid with more than two groups, and even if the grouping variable is continuous.
   This formula implies that the variance of the total group cannot be smaller than the mean of the variances of the subgroups. Note, however, that the total variance is not necessarily larger than the variances of the subgroups. In the above example, when the subgroups are analyzed separately, the variance is influenced only by the man-man differences and the woman-woman differences. If the two groups are combined, however, then the men-women differences enter into the variance also.
2. Many computational formulas for the variance are based on this equality: **The variance is equal to the mean of the squares minus the square of the mean.** For example, if we consider the numbers 1, 2, 3, 4 then the mean of the squares is (1 × 1 + 2 × 2 + 3 × 3 + 4 × 4) / 4 = 7.5. The mean is 2.5, so the square of the mean is 6.25. Therefore the variance is 7.5 − 6.25 = 1.25, which is indeed the same result obtained earlier with the definition formulas. Many pocket calculators use an algorithm that is based on this formula and that allows them to compute the variance while the data are entered, without storing all values in memory. The algorithm is to adjust only three variables when a new data value is entered: The number of data entered so far (n), the sum of the values so far (S), and the sum of the squared values so far (SS). For example, if the data are 1, 2, 3, 4, then after entering the first value, the algorithm would have n = 1, S = 1 and SS = 1. After entering the second value (2), it would have n = 2, S = 3 and SS = 5. When all data are entered, it would have n = 4, S = 10 and SS = 30. Next, the mean is computed as $M = S / n$, and finally the variance is computed as $SS / n - M \times M$. In this example the outcome would be 30 / 4 − 2.5 × 2.5 = 7.5 − 6.25 = 1.25. If the unbiased sample estimate is to be computed, the outcome will be multiplied by $n / (n - 1)$, which yields 1.667 in this example.

Properties, formal

Sum of uncorrelated variables (Bienaymé formula)

One reason for the use of the variance in preference to other measures of dispersion is that the variance of the sum (or the difference) of uncorrelated random variables is the sum of their variances:

This statement is called the Bienaymé formula.[1] and was discovered in 1853. It is often made with the stronger condition that the variables are independent, but uncorrelatedness suffices. So if the variables have the same variance σ^2, then, since division by n is a linear transformation, this formula immediately implies that the variance of their mean is

That is, the variance of the mean decreases with n. This fact is used in the definition of the standard error of the sample mean, which is used in the central limit theorem.

Sum of correlated variables

In general, if the variables are correlated, then the variance of their sum is the sum of their covariances:

(Note: This by definition includes the variance of each variable, since $\operatorname{Cov}(X,X)=\operatorname{Var}(X)$.)

Here Cov is the covariance, which is zero for independent random variables (if it exists). The formula states that the variance of a sum is equal to the sum of all elements in the covariance matrix of the components. This formula is used in the theory of Cronbach's alpha in classical test theory.

So if the variables have equal variance σ^2 and the average correlation of distinct variables is ρ, then the variance of their mean is

This implies that the variance of the mean increases with the average of the correlations. Moreover, if the variables have unit variance, for example if they are standardized, then this simplifies to

This formula is used in the Spearman-Brown prediction formula of classical test theory. This converges to ρ if n goes to infinity, provided that the average correlation remains constant or converges too. So for the variance of the mean of standardized variables with equal correlations or converging average correlation we have

Therefore, the variance of the mean of a large number of standardized variables is approximately equal to their average correlation. This makes clear that the sample mean of correlated variables does generally not converge to the population mean, even though the Law of large numbers states that the sample mean will converge for independent variables.

Weighted sum of variables

Properties 6 and 8, along with this property from the covariance page: $\operatorname{Cov}(aX, bY) = ab \operatorname{Cov}(X, Y)$ jointly imply that

This implies that in a weighted sum of variables, the variable with the largest weight will have a disproportionally large weight in the variance of the total. For example, if X and Y are uncorrelated and the weight of X is two times the weight of Y, then the weight of the variance of X will be four times the weight of the variance of Y.

Decomposition

The general formula for variance decomposition or the law of total variance is: If X and Y are two random variables and the variance of X exists, then

Here, $\mathrm{E}(X|Y)$ is the conditional expectation of X given Y, and $\mathrm{Var}(X|Y)$ is the conditional variance of X given Y. (A more intuitive explanation is that given a particular value of Y, then X follows a distribution with mean $\mathrm{E}(X|Y)$ and variance $\mathrm{Var}(X|Y)$. The above formula tells how to find $\mathrm{Var}(X)$ based on the distributions of these two quantities when Y is allowed to vary.) This formula is often applied in analysis of variance, where the corresponding formula is

$$SS_{\text{Total}} = SS_{\text{Between}} + SS_{\text{Within}}.$$

It is also used in linear regression analysis, where the corresponding formula is

$$SS_{\text{Total}} = SS_{\text{Regression}} + SS_{\text{Residual}}.$$

This can also be derived from the additivity of variances (property 8), since the total (observed) score is the sum of the predicted score and the error score, where the latter two are uncorrelated.

Computational formula

The **computational formula for the variance** follows in a straightforward manner from the linearity of expected values and the above definition:

This is often used to calculate the variance in practice, although it suffers from catastrophic cancellation if the two components of the equation are similar in magnitude.

Characteristic property

The second moment of a random variable attains the minimum value when taken around the first moment (i.e., mean) of the random variable, i.e. . Conversely, if a continuous function satisfies for all random variables X, then it is necessarily of the form , where $a > 0$. This also holds in the multidimensional case.[2]

Calculation from the CDF

The population variance for a non-negative random variable can be expressed in terms of the cumulative distribution function F using

where $H(u) = 1 - F(u)$ is the right tail function. This expression can be used to calculate the variance in situations where the CDF, but not the density, can be conveniently expressed.

Approximating the variance of a function

The delta method uses second-order Taylor expansions to approximate the variance of a function of one or more random variables. For example, the approximate variance of a function of one variable is given by

provided that f is twice differentiable and that the mean and variance of X are finite.[*citation needed*]

Population variance and sample variance

In general, the population variance of a *finite* population of size N is given by

or if the population is an abstract population with probability distribution Pr:

where is the population mean. This is merely a special case of the general definition of variance introduced above, but restricted to finite populations.

In many practical situations, the true variance of a population is not known *a priori* and must be computed somehow. When dealing with infinite populations, this is generally impossible.

A common task is to estimate the variance of a population from a sample. We take a sample with replacement of n values from the population, and estimate the variance on the basis of this sample. There are several good estimators. Two of them are well known:

and

Both are referred to as ***sample variance***.

The two estimators only differ slightly as we see, and for larger values of the sample size n the difference is negligible. While the first one may be seen as the variance of the sample considered as a population, the second one is the unbiased estimator of the population variance, meaning that its expected value $E[s^2]$ is equal to the true variance of the sampled random variable; the use of the term $n - 1$ is called Bessel's correction. The sample variance with $n - 1$ is a U-statistic for the function $f(x_1,x_2) = (x_1 - x_2)^2 / 2$ meaning that it is obtained by averaging a 2-sample statistic over 2-element subsets of the population.

While,

Distribution of the sample variance

Being a function of random variables, the sample variance is itself a random variable, and it is natural to study its distribution. In the case that y_i are independent observations from a normal distribution, Cochran's theorem shows that s^2 follows a scaled chi-square distribution:

As a direct consequence, it follows that

If the y_i are independent and identically distributed, but not necessarily normally distributed, then s^2 is unbiased for σ^2. If the conditions of the law of large numbers hold, s^2 is a consistent estimator of σ^2.

Generalizations

Unbiased estimate for expected error in the mean of A for a sample of M data points with sample bias coefficient ρ. The log-log slope -½ line for ρ=0 is the unbiased standard error.

If X is a vector-valued random variable, with values in , and thought of as a column vector, then the natural generalization of variance is , where and is the transpose of X, and so is a row vector. This variance is a positive semi-definite square matrix, commonly referred to as the covariance matrix.

If X is a complex-valued random variable, with values in , then its variance is , where X^* is the complex conjugate of X. This variance is also a positive semi-definite square matrix.

If one's (real) random variables are defined on an n-dimensional continuum $\mathbf{x}$, the cross-covariance of variables A[$\mathbf{x}$] and B[$\mathbf{x}$] as a function of n-dimensional vector displacement (or lag) $\mathbf{\Delta x}$ may be defined as $\sigma_{AB}[\mathbf{\Delta x}] \equiv \langle (A[\mathbf{x}+\mathbf{\Delta x}]-\mu_A)(B[\mathbf{x}]-\mu_B) \rangle_{\mathbf{x}}$. Here

the population (as distinct from sample) average over **x** is denoted by angle brackets $\langle\ \rangle_{\mathbf{x}}$ or the Greek letter μ.

This quantity, called a second-moment correlation measure because it's a generalization of the second-moment statistic *variance*, is sometimes put into dimensionless form by normalizing with the population standard deviations of A and B (e.g. $\sigma_A \equiv \text{Sqrt}[\sigma_{AA}[0]]$). This results in a correlation coefficient $\rho_{AB}[\mathbf{\Delta x}] \equiv \sigma_{AB}[\mathbf{\Delta x}]/(\sigma_A \sigma_B)$ that takes on values between plus and minus one. When A is the same as B, the foregoing expressions yield values for autocovariance, a quantity also known in scattering theory as the pair-correlation (or Patterson) function.

If one defines *sample bias coefficient* ρ as an average of the autocorrelation-coefficient $\rho_{AA}[\mathbf{\Delta x}]$ over all point pairs in a set of M sample points[3], an unbiased estimate for *expected error in the mean* of A is the square root of: sample variance (taken as a population) times $(1+(M-1)\rho)/((M-1)(1-\rho))$. When ρ is much greater than 1/(M-1), this reduces to the square root of: sample variance (taken as a population) times $\rho/(1-\rho)$. When $|\rho|$ is much less than 1/(M-1) this yields the more familiar expression for standard error, namely the square root of: sample variance (taken as a population) over (M-1).

Moment of inertia

The variance of a probability distribution is analogous to the moment of inertia in classical mechanics of a corresponding mass distribution along a line, with respect to rotation about its center of mass. It is because of this analogy that such things as the variance are called *moments* of probability distributions. The covariance matrix is related to the moment of inertia tensor for multivariate distributions. The moment of inertia of a cloud of *n* points with a covariance matrix of Σ is given by

This difference between moment of inertia in physics and in statistics is clear for points that are gathered along a line. Suppose many points are close to the *x* and distributed along it. The covariance matrix might look like

That is, there is the most variance in the *x* direction. However, physicists would consider this to have a low moment *about* the *x* axis so the moment-of-inertia tensor is

Skewness

Example of experimental data with non-zero skewness (gravitropic response of wheat coleoptiles, 1,790)

In probability theory and statistics, **skewness** is a measure of the asymmetry of the probability distribution of a real-valued random variable.

Introduction

Consider the distribution in the figure. The bars on the right side of the distribution taper differently than the bars on the left side. These tapering sides are called *tails*, and they provide a visual means for determining which of the two kinds of skewness a distribution has:

1. **negative skew**: The left tail is longer; the mass of the distribution is concentrated on the right of the figure. It has relatively few low values. The distribution is said to be *left-skewed*. Example (observations): 1,1000,1001,1002,1003

2. **positive skew**: The right tail is longer; the *mass* of the distribution is concentrated on the left of the figure. It has relatively few high values. The distribution is said to be *right-skewed*. Example (observations): 1,2,3,4,100.

In a skewed (unbalanced, lopsided) distribution, the mean is farther out in the long tail than is the median. If there is no skewness or the distribution is symmetric like the bell-shaped normal curve then the mean = median = mode.

Many textbooks teach a rule of thumb stating that the mean is right of the median under right skew, and left of the median under left skew. This rule fails with surprising frequency. It can fail in multimodal distributions, or in distributions where one tail is long but the other is heavy. Most commonly, though, the rule fails in discrete distributions where the areas to the left and right of the median are not equal. Such distributions not only contradict the textbook relationship between mean, median, and skew, they also contradict the textbook interpretation of the median.[1]

Definition

Skewness, the third standardized moment, is written as γ_1 and defined as

where μ_3 is the third moment about the mean and σ is the standard deviation. Equivalently, skewness can be defined as the ratio of the third cumulant κ_3 and the third power of the square root of the second cumulant κ_2:

This is analogous to the definition of kurtosis, which is expressed as the fourth cumulant divided by the fourth power of the square root of the second cumulant.

The skewness of a random variable X is sometimes denoted Skew[X].

Sample skewness

For a sample of n values the *sample skewness* is

where x_i is the i^{th} value, is the sample mean, m_3 is the sample third central moment, and m_2 is the sample variance.

Given samples from a population, the equation for the sample skewness g_1 above is a biased estimator of the population skewness. The usual estimator of skewness is

where k_3 is the unique symmetric unbiased estimator of the third cumulant and k_2 is the symmetric unbiased estimator of the second cumulant. Unfortunately G_1 is, nevertheless, generally biased. Its expected value can even have the opposite sign from the true skewness; compare unbiased estimation of standard deviation.

Properties

If Y is the sum of n independent random variables, all with the same distribution as X, then it can be shown that Skew[Y] = Skew[X] / $\sqrt{n}$.

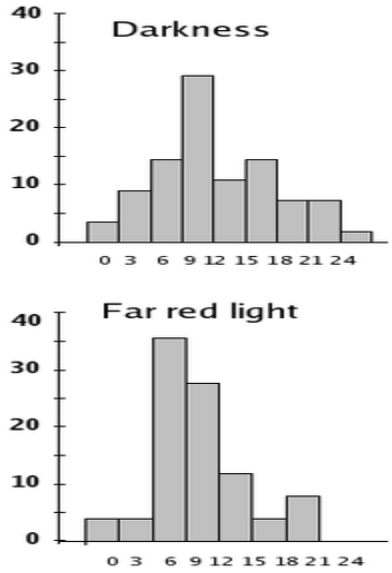
Kurtosis

From Wikipedia, the free encyclopedia

Jump to: navigation, search

In probability theory and statistics, **kurtosis** (from the Greek word κυρτός, *kyrtos* or *kurtos*, meaning bulging) is a measure of the "peakedness" of the probability

distribution of a real-valued random variable. Higher kurtosis means more of the variance is due to infrequent extreme deviations, as opposed to frequent modestly-sized deviations.



The far red light has no effect on the average speed of the gravitropic reaction in wheat coleoptiles, but it changes kurtosis from platykurtic to leptokurtic (−0.194 → 0.055)

Definition

The fourth standardized moment is defined as

$$\frac{\mu_4}{\sigma^4},$$

where μ_4 is the fourth moment about the mean and σ is the standard deviation. This is sometimes used as the definition of kurtosis in older works, but is not the definition used here.

Kurtosis is more commonly defined as the fourth cumulant divided by the square of the second cumulant, which is equal to the fourth moment around the mean divided by the square of the variance of the probability distribution minus 3,

$$\gamma_2 = \frac{\kappa_4}{\kappa_2^2} = \frac{\mu_4}{\sigma^4} - 3,$$

which is also known as **excess kurtosis**. The "minus 3" at the end of this formula is often explained as a correction to make the kurtosis of the normal distribution equal to zero. Another reason can be seen by looking at the formula for the kurtosis of the sum of random variables. Because of the use of the cumulant, if Y is the sum of n independent random variables, all with the same distribution as X, then $\mathrm{Kurt}[Y] = \mathrm{Kurt}[X] / n$, while the formula would be more complicated if kurtosis were defined as μ_4 / σ^4.

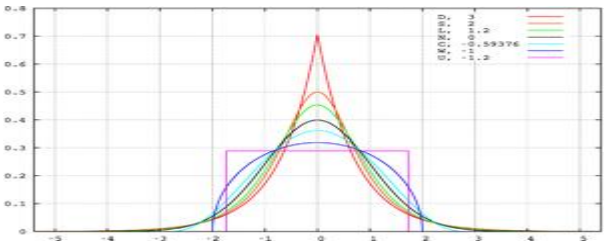
More generally, if $X_1, \ldots, X_n$ are independent random variables all *having the same variance*, then

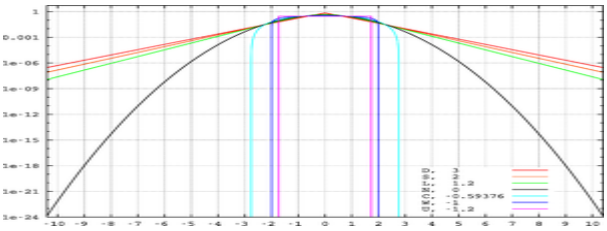
$$\operatorname{Kurt}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Kurt}(X_i),$$

whereas this identity would not hold if the definition did not include the subtraction of 3.

The fourth standardized moment must be at least 1, so the excess kurtosis must be −2 or more (the lower bound is realized by the Bernoulli distribution with p = ½, or "coin toss"); there is no upper limit and it may be infinite.

Kurtosis of well-known distributions





In this example we compare several well-known distributions from different parametric families. All densities considered here are unimodal and symmetric. Each has a mean and skewness of zero. Parameters were chosen to result in a variance of unity in each case. The images on the right show curves for the following seven densities, on a linear scale and logarithmic scale:

- D: Laplace distribution, a.k.a. double exponential distribution, red curve (two straight lines in the log-scale plot), excess kurtosis = 3
- S: hyperbolic secant distribution, orange curve, excess kurtosis = 2
- L: logistic distribution, green curve, excess kurtosis = 1.2
- N: normal distribution, black curve (inverted parabola in the log-scale plot), excess kurtosis = 0
- C: raised cosine distribution, cyan curve, excess kurtosis = −0.593762...
- W: Wigner semicircle distribution, blue curve, excess kurtosis = −1
- U: uniform distribution, magenta curve (shown for clarity as a rectangle in both images), excess kurtosis = −1.2.

Note that in these cases the platykurtic densities have bounded support, whereas the densities with positive or zero excess kurtosis are supported on the whole real line.

There exist platykurtic densities with infinite support,

- e.g., exponential power distributions with sufficiently large shape parameter b

and there exist leptokurtic densities with finite support.

- e.g., a distribution that is uniform between −3 and −0.3, between −0.3 and 0.3, and between 0.3 and 3, with the same density in the (−3, −0.3) and (0.3, 3) intervals, but with 20 times more density in the (−0.3, 0.3) interval

Sample kurtosis

For a sample of *n* values the **sample kurtosis** is

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3$$

where m_4 is the fourth sample moment about the mean, m_2 is the second sample moment about the mean (that is, the sample variance), x_i is the i^{th} value, and $\overline{x}$ is the sample mean.

The formula

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

,

$$E = \frac{1}{nD^2}\sum_{i=1}^{n}(x_i - \bar{x})^4 - 3$$

is also used, where n—the sample size, *D*—the pre-computed variance, x_i—the value of the x'th measurement and $\bar{x}$—the pre-computed arithmetic mean.

Mean absolute error

From Wikipedia, the free encyclopedia

Jump to: navigation, search

In statistics, the **mean absolute error** is a quantity used to measure how close forecasts or predictions are to the eventual outcomes. The mean absolute error (MAE) is given by

As the name suggests, the mean absolute error is an average of the absolute errors $e_i = f_i - y_i$, where f_i is the prediction and y_i the true value. Note that alternative formulations may include relative frequencies as weight factors.

The mean absolute error is a common measure of forecast error in time series analysis, where the terms "mean absolute deviation" is sometimes used in confusion with the more standard definition of mean absolute deviation. The same confusion exists more generally.

Interquartile range

From Wikipedia, the free encyclopedia

Jump to: navigation, search

In descriptive statistics, the **interquartile range** (**IQR**), also called the **midspread** or **middle fifty**, is a measure of statistical dispersion, being equal to the difference between the third and first quartiles.

Unlike the (total) range, the interquartile range is a robust statistic, having a breakdown point of 25%, and is thus often preferred to the total range.

The IQR is used to build box plots, simple graphical representations of a probability distribution.

For a symmetric distribution (so the median equals the midhinge, the average of the first and third quartiles), half the IQR equals the median absolute deviation (MAD).

The median is the corresponding measure of central tendency.

From this table, the width of the interquartile range is 115 − 105 = 10.

Data set in a plain-text box plot

```
                    +-----+-+
 o        *     |-------|     | |---|
                    +-----+-+


+---+---+---+---+---+---+---+---+---+---+---+---+---+   number line
0   1   2   3   4   5   6   7   8   9  10  11  12
```

For this data set:

- lower (first) quartile ($Q1$, $x_{.25}$) = 7
- median (second quartile) (Med, $x_{.5}$) = 8.5
- upper (third) quartile ($Q3$, $x_{.75}$) = 9
- interquartile range, IQR = $Q3 - Q1$ = 2

Interquartile range of distributions

The interquartile range of a continuous distribution can be calculated by integrating the probability density function (which yields the cumulative distribution function—any other means of calculating the CDF will also work). The lower quartile, Q1, is a number such that integral of the PDF from -∞ to Q1 equals 0.25, while the upper quartile, Q3, is such a number that the integral from Q3 to ∞ equals 0.25; in terms of the CDF, the quartiles can be defined as follows:

$Q1 = \text{CDF}^{-1}(0.25)$

$Q3 = \text{CDF}^{-1}(0.75)$

The interquartile range and median of some common distributions are shown below

Distribution	Median	IQR
Normal	μ	$2\,\Phi^{-1}(0.75) \approx 1.349$
Laplace	μ	$2b\ln(2)$
Cauchy	μ	

Range (statistics)

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.

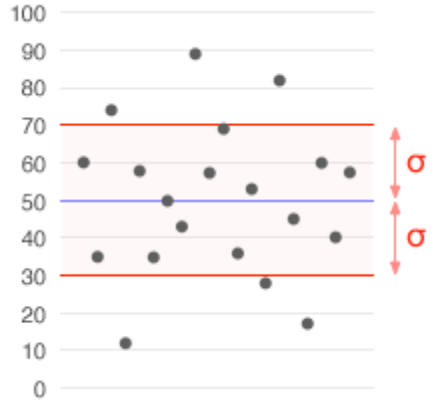
For a population, the range is greater than or equal to twice the standard deviation, which equality only for the coin toss (Bernoulli distribution with $p = ½$).

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.

Standard deviation

A plot of a normal distribution (or bell curve). Each colored band has a width of one standard deviation.



A data set with a mean of 50 (shown in blue) and a standard deviation (σ) of 20.

In probability theory and statistics, **standard deviation** is a measure of the variability or dispersion of a statistical population, a data set, or a probability distribution. A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values.

For example, the average height for adult men in the United States is about 70 inches (178 cm), with a standard deviation of around 3 in (8 cm). This means that most men (about 68 percent, assuming a normal distribution) have a height within 3 in (8 cm) of the mean (67–73 in (170–185 cm)), whereas almost all men (about 95%) have a height within 6 in (15 cm) of the mean (64–76 in (163–193 cm)). If the standard

deviation were zero, then all men would be exactly 70 in (178 cm) high. If the standard deviation were 20 in (51 cm), then men would have much more variable heights, with a typical range of about 50 to 90 in (127 to 229 cm).

In addition to expressing the variability of a population, standard deviation is commonly used to measure confidence in statistical conclusions. 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. The reported margin of error is typically 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—normal random error or variation in the measurements is in this way distinguished from causal variation. Standard deviation is also important in finance, where the standard deviation on the rate of return on an investment is a measure of the volatility of the investment.

The term *standard deviation* was first used[1] in writing by Karl Pearson[2] in 1894, following his use of it in lectures. This was as a replacement for earlier alternative names for the same idea: for example Gauss used "mean error".[3] 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 quantity called the sample standard deviation, explained below.

Basic example

Consider a population consisting of the following values:

There are eight data points in total, with a mean (or average) value of 5:

To calculate the population standard deviation, we compute the difference of each data point from the mean, and square the result:

Next we average these values and take the square root, which gives the standard deviation:

Therefore, the above has a population standard deviation of 2.

Note that we are assuming that we are dealing with a complete population. If our 8 values are obtained by random sampling from some parent population, we might prefer to compute the **sample standard deviation** using a denominator of 7 instead of 8. See below for an explanation.

Definition

Probability distribution or random variable

Let X be a random variable with mean value μ:

Here the operator E denotes the average or expected value of X. Then the **standard deviation** of X is the quantity

That is, the standard deviation σ (sigma) is the square root of the average value of $(X - \mu)^2$.

In the case where X takes random values from a finite data set , with each value having the same probability, the standard deviation is

or, using summation notation,

The standard deviation of a (univariate) probability distribution is the same as that of a random variable having that distribution. Not all random variables have a standard deviation, since these expected values need not exist. For example, the standard deviation of a random variable which follows a Cauchy distribution is undefined because its expected value is undefined.

[edit] Continuous random variable

Continuous distributions usually give a formula for calculating the standard deviation as a function of the parameters of the distribution. In general, the standard deviation of a continuous real-valued random variable X with probability density function $p(x)$ is

where

and where the integrals are definite integrals taken for x ranging over the range of X.

Discrete random variable or data set

The standard deviation of a discrete random variable is the root-mean-square (RMS) deviation of its values from the mean.

If the random variable X takes on N values (which are real numbers) with equal probability, then its standard deviation σ can be calculated as follows:

1. Find the mean, , of the values.
2. For each value x_i calculate its deviation () from the mean.
3. Calculate the squares of these deviations.
4. Find the mean of the squared deviations. This quantity is the variance σ^2.
5. Take the square root of the variance.

This calculation is described by the following formula:

where is the arithmetic mean of the values x_i, defined as:

If not all values have equal probability, but the probability of value x_i equals p_i, the standard deviation can be computed by:

and

where

and N' is the number of non-zero weight elements.

The standard deviation of a data set is the same as that of a discrete random variable that can assume precisely the values from the data set, where the point mass for each value is proportional to its multiplicity in the data set.

Example

Suppose we wished to find the standard deviation of the data set consisting of the values 3, 7, 7, and 19.

Step 1: find the arithmetic mean (average) of 3, 7, 7, and 19,

Step 2: find the deviation of each number from the mean,

Step 3: square each of the deviations, which amplifies large deviations and makes negative values positive,

Step 4: find the mean of those squared deviations,

Step 5: take the non-negative square root of the quotient (converting squared units back to regular units),

So, the standard deviation of the set is 6. This example also shows that, in general, the standard deviation is different from the mean absolute deviation (which is 5 in this example).

Note that if the above data set represented only a sample from a greater population, a modified standard deviation would be calculated (explained below) to estimate the population standard deviation, which would give 6.93 for this example.

Simplification of formula

The calculation of the sum of squared deviations can be simplified as follows:

Applying this to the original formula for standard deviation gives:

This can be memorized as taking the square root of (the average of the squares less the square of the average).

Estimating population standard deviation

In the real world, finding the standard deviation of an entire population is unrealistic except in certain cases, (such as standardized testing), where every member of a population is sampled. In most cases, the standard deviation σ is estimated by examining a random sample taken from the population. Some estimators are given below:

With standard deviation of the sample

An estimator for σ sometimes used is the **standard deviation of the sample**, denoted by "s_n" and defined as follows:

This estimator has a uniformly smaller mean squared error than the "sample standard deviation" (see below), and is the maximum-likelihood estimate when the population is normally distributed. But this estimator, when applied to a small or moderately-sized sample, tends to be too low: it is a biased estimator.

With sample standard deviation

The most common estimator for σ used is an adjusted version, the **sample standard deviation,** denoted by "s" and defined as follows:

where is the sample and is the mean of the sample. This correction (the use of $N - 1$ instead of N) is known as Bessel's correction. The reason for this correction is that s^2 is an unbiased estimator for the variance σ^2 of the underlying population, if that variance exists and the sample values are drawn independently with replacement. However, s is *not* an unbiased estimator for the standard deviation σ; it tends to underestimate the population standard deviation.

Note that the term "standard deviation of the sample" is used for the *uncorrected* estimator (using N) whilst the term "sample standard deviation" is used for the *corrected* estimator (using $N - 1$). The denominator $N - 1$ is the number of degrees of freedom in the vector of residuals, .

With interquartile range

The statistic

(1.35 is an approximation) where IQR is the interquartile range of the sample, is a consistent estimate of σ. The interquartile range IQR is the difference of the 3rd quartile of the data and the 1st quartile of the data. The asymptotic relative efficiency (ARE) of this estimator with respect to the one from sample standard deviation is 0.37. Hence, for normal data, it is better to use the one from sample standard deviation; when data is with thicker tails, this estimator can be more efficient.[4][*not in citation given*][*dubious – discuss*]

Other estimators

Further information: Unbiased estimation of standard deviation

Although an unbiased estimator for σ is known when the random variable is normally distributed, the formula is complicated and amounts to a minor correction: see Unbiased estimation of standard deviation for more details. Moreover, unbiasedness, (in this sense of the word), is not always desirable: see bias of an estimator.

if we take all weights equal to 1.

Mean difference

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.

Calculation

For a population of size n, with a sequence of values y_i, i = 1 to n:

$$MD = \frac{1}{n(n-1)} \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} |y_i - y_j|$$

For a discrete probability function $f(y)$, where y_i, i = 1 to n, are the values with nonzero probabilities:

$$MD = \Sigma_{i=1}^{n} \Sigma_{j=1}^{n} f(y_i) f(y_j) |y_i - y_j|$$

For a probability density function $f(x)$:

$$MD = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x)\, f(y)\, |x - y|\, dx\, dy$$

For a cumulative distribution function F(x) with inverse x(F):

$$MD = \int_{0}^{1} \int_{0}^{1} |x(F_1) - x(F_2)|\, dF_1\, dF_2$$

The inverse $x(F)$ may not exist because the cumulative distribution function has jump discontinuities or intervals of constant values. However, the previous formula can still apply by generalizing the definition of $x(F)$:

$x(F_1) = \inf \{y : F(y) \geq F_1\}$.

Relative mean difference

When the probability distribution has a finite and nonzero arithmetic mean, the relative mean difference, sometimes denoted by ∇ or RMD, is defined by

$$RMD = \frac{MD}{\text{arithmetic mean}}.$$

The relative mean difference quantifies the mean difference in comparison to the size of the mean and is a dimensionless quantity. The relative mean difference is equal to twice the Gini coefficient which is defined in terms of the Lorenz curve. This relationship gives complementary perspectives to both the relative mean difference and the Gini coefficient, including alternative ways of calculating their values.

Compared to standard deviation

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.

Sample estimators

For a random sample S from a random variable $\boldsymbol{X}$, consisting of n values y_i, the statistic

$$MD(S) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |y_i - y_j|}{n(n-1)}$$

is a consistent and unbiased estimator of MD($\boldsymbol{X}$).

The statistic:

$$RMD(S) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |y_i - y_j|}{(n-1)\sum_{i=1}^{n} y_i}$$

is a consistent estimator of RMD($\boldsymbol{X}$), but is not, in general, unbiased.

Confidence intervals for RMD($\boldsymbol{X}$) can be calculated using bootstrap sampling techniques.

There does not exist, in general, an unbiased estimator for RMD($\boldsymbol{X}$), in part because of the difficulty of finding an unbiased estimation for multiplying by the inverse of the mean. For example, even where the sample is known to be taken from a random variable $\boldsymbol{X}(p)$ for an unknown p, and $\boldsymbol{X}(p)$ - 1 has the Bernoulli distribution, so that $\Pr(\boldsymbol{X}(p) = 1) = 1 - p$ and $\Pr(\boldsymbol{X}(p) = 2) = p$, then

$$\text{RMD}(\boldsymbol{X}(p)) = 2p(1 - p)/(1 + p).$$

But the expected value of any estimator $R(\boldsymbol{S})$ of RMD($\boldsymbol{X}(p)$) will be of the form:

$$E(R(S)) = \sum_{i=0}^{n} p^i (1-p)^{n-i} r_i$$

where the r_i are constants. So E(R(**S**)) can never equal RMD(***X***(*p*)) for all *p* between 0 and 1.

References

 Moses, Lincoln E. (1986) *Think and Explain with Statistics*, Addison-Wesley, ISBN 978-0-201-15619-5 . pp. 1–3

 ^ Hays, William Lee, (1973) *Statistics for the Social Sciences*, Holt, Rinehart and Winston, p.xii, ISBN 978-0-03-077945-9

 ^ Moore, David (1992). "Teaching Statistics as a Respectable Subject". In F. Gordon and S. Gordon. *Statistics for the Twenty-First Century*. Washington, DC: The Mathematical Association of America. pp. 14–25. ISBN 978-0-88385-078-7.

 ^ Chance, Beth L.; Rossman, Allan J. (2005). "Preface". *Investigating Statistical Concepts, Applications, and Methods*. Duxbury Press. ISBN 978-0-495-05064-3.

 ^ Anderson, D.R.; Sweeney, D.J.; Williams, T.A.. (1994) *Introduction to Statistics: Concepts and Applications*, pp. 5–9. West Group. ISBN 978-0-314-03309-3

 ^ Singh, Simon (2000). *The code book : the science of secrecy from ancient Egypt to quantum cryptography* (1st Anchor Books ed.). New York: Anchor Books. ISBN 0-385-49532-3.[*page needed*]

 ^ Al-Kadi, Ibrahim A. (1992) "The origins of cryptology: The Arab contributions", *Cryptologia*, 16(2) 97–126. doi:10.1080/0161-119291866801

 ^ Willcox, Walter (1938) *The Founder of Statistics*. Review of the International Statistical Institute 5(4):321–328.

 ^ Leo Breiman (2001). "Statistical Modelling: the two cultures", *Statistical Science* **16** (3), pp.199-231.

 ^ Lindley, D. (2000) "The Philosophy of Statistics", *Journal of the Royal Statistical Society, Series D (The Statistician)*, 49 (3), 293-337 JSTOR 2681060 doi:10.1111/1467-9884.00238

 ^ Huff, Darrell (1954) *How to Lie With Statistics*, WW Norton & Company, Inc. New York, NY. ISBN 0-393-31072-8