USGS
science for a changing world

Basic statistical concepts and methods for earth scientists

By Ricardo A. Olea

Open-File Report 2008–1017

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
KENNETH L. SALAZAR, Secretary

U.S. Geological Survey
Marcia K. McNutt, Director

U.S. Geological Survey, Reston, Virginia 2008

For product and ordering information:
World Wide Web: http://www.usgs.gov/pubprod
Telephone: 1-888-ASK-USGS (1-888-275-8747)

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment: World Wide Web: http://www.usgs.gov
Telephone: 1-888-ASK-USGS (1-888-275-8747)

Suggested citation:
Olea, R.A., 2008, Basic statistical concepts for earth scientists: U.S. Geological Survey, Open-File Report 2008-1017, 191 p.

Revised version, February 2010

Basic statistical concepts and methods for earth scientists

Ricardo A. Olea

Open-File Report 2008–1017

U.S. Department of the Interior
U.S. Geological Survey

CONTENTS

1.	Introduction	3
2.	Data Collection Issues	11
3.	Univariate Statistics	35
4.	Bivariate Statistics	78
5.	Nondetect Statistics	100
6.	Statistical Testing	111
7.	Multivariate Statistics	153

These notes have been prepared for teaching a one-day course intended to refresh and upgrade the statistical background of the participants.

USGS

1. INTRODUCTION

STATISTICS

Statistics is the science of collecting, analyzing, interpreting, modeling, and displaying masses of numerical data primarily for the characterization and understanding of incompletely known systems.

Over the years, these objectives have led to a fair amount of analytical work to achieve, substantiate, and guide descriptions and inferences.

WHY STATISTICS?

- Given any district, time and economics ordinarily preclude acquiring perfect knowledge of a single attribute of interest, let alone a collection of them, resulting in uncertainty and sometimes into bulky datasets.
- Note that uncertainty is not an intrinsic property of geological systems; it is the result of incomplete knowledge by the observer.
- To a large extent, earth sciences aim to inferring past processes and predicting future events based on relationships among attributes, preferably quantifying uncertainty.

Statistics is an important component in the emerging fields of data mining and geoinformatics.

USGS

WHEN NOT TO USE STATISTICS?

- There are no measurements for the attribute(s) of interest.
- There are very few observations, say 3 or 4.
- The attribute is perfectly known and there is no interest in having associated summary information, preparing any generalizations, or making any type of quantitative comparisons to other attributes.

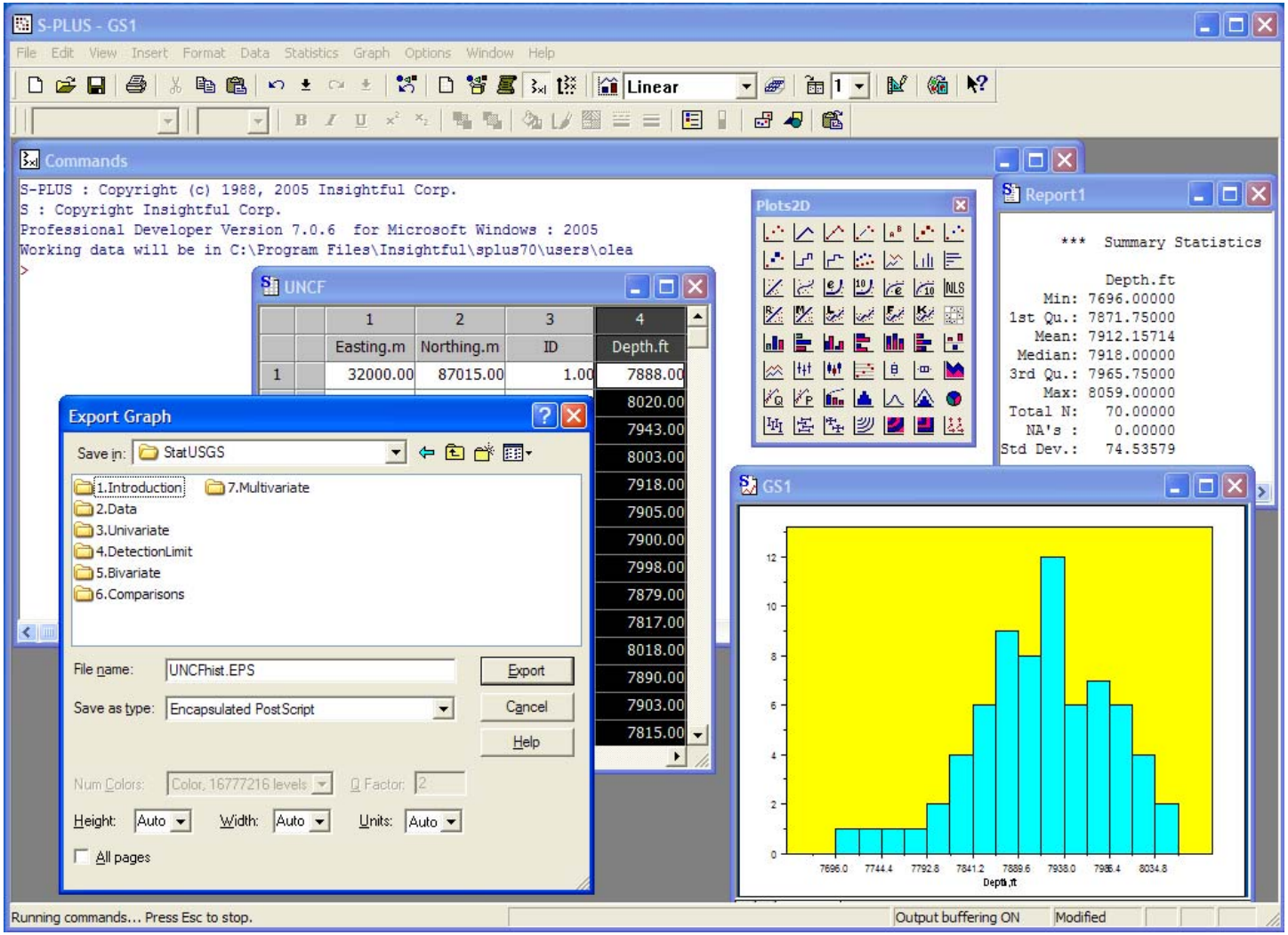
CALCULATIONS

As shown in this course, most of the statistical calculations for the case of one attribute are fairly trivial, not requiring more than a calculator or reading a table of values.

Multivariate methods can be computationally intensive, suited for computer assistance. Computer programs used to be cumbersome to utilize, some requiring the mastering of special computer languages.

Modern computer programs are easy to employ as they are driven by friendly graphical user interfaces. Calculations and graphical display are performed through direct manipulation of graphical icons.

EXAMPLE OF MODERN PROGRAM



COURSE OBJECTIVES

This short course is intended to refresh basic concepts and present various tools available for the display and optimal extraction of information from data.

At the end of the course, the participants:

- should have increased their ability to read the statistical literature, particularly those publications listed in the bibliography, thus ending better qualified to independently apply statistics;
- may have learned some more theoretically sound and convincing ways to prepare results;
- might be more aware both of uncertainties commonly associated with geological modeling and of the multiple ways that statistics offers for quantifying such uncertainties.

USGS

"That's the gist of what I want to say. Now get me some statistics to base it on."

2. DATA COLLECTION ISSUES

ACCURACY AND PRECISION

ACCURACY

Accuracy is a property of measured and calculated quantities that indicates the quality of being close to the actual, true value.

- For example, the value 3.2 is a more accurate representation of the constant π (3.1415926536 …) than 3.2564.
- Accuracy is related to, but not synonymous with precision.

PRECISION

In mathematics, precision is the number of significant digits in a numerical value, which may arise from a measurement or as the result of a calculation.

The number 3.14 is less precise representation of the constant π than the number 3.55745.

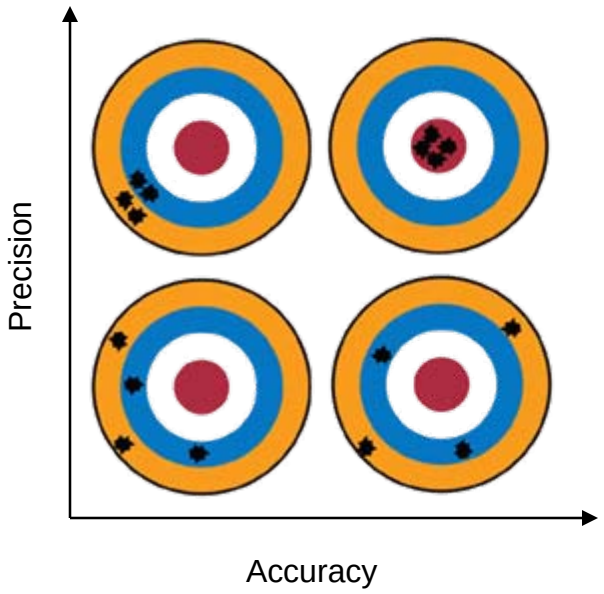
For experimentally derived values, precision is related to instrumental resolution, thus to the ability of replicating measurements.

ACCURACY AND PRECISION

Mathematically, a calculation or a measurement can be:

- Accurate and precise. For example 3.141592653 is both an accurate and precise value of π.
- Precise but not accurate, like $\pi = 3.5893627002$.
- Imprecise and inaccurate, such as $\pi = 4$.
- Accurate but imprecise, such as $\pi = 3.1$.

Experimental context



While precision is obvious to assess, accuracy is not. To a large extent, what is behind statistics is an effort to evaluate accuracy.

USGS

SIGNIFICANT DIGITS

SIGNIFICANT DIGITS

Significant digits are the numerals in a number that are supported by the precision of a measurement or calculation.

- The precision of a measurement is related to the discriminating capability of the instrument.
- The precision of a calculation is related to the numbers involved in the arithmetic processing and is decided by a few simple rules.

SIGNIFICANT DIGITS IN A NUMBER

- Digits from 1-9 are always significant.
- Zeros between two other significant digits are always significant.
- Zeros used solely for spacing the decimal point (placeholders) are not significant. For example, the value 0.00021 has two significant digits.
- Trailing zeros, both to the right of the decimal place and of another significant digit, are significant. For example, 2.30 has three significant digits.
- For real numbers, zeros between the rightmost significant digit and the decimal place are not significant. Hence, 4000**.** is automatically understood to have one significant digit. If, for example, indeed there are two, use the notation $4.0 \cdot 10^3$.
- Whole numbers have unlimited number of significant digits. So, 4000 has infinite significant digits.

USGS

SIGNIFICANT DIGITS IN CALCULATIONS

Reporting results often requires some manual rounding, especially if using calculators or computers.

In all four basic operations of addition, substraction, multiplication, and division, the number with the least significant digits in decimal places determines the significant decimal digits to report in the answer. Examples:

- 2.1 + 2.08 = 4.2
- (8.1·2.08)/4 = 4.2

Easy way corollary:

If all the operands have the same number of significant decimal places, the significant decimal places in the result are the same as those of the operands. Example: 0.38·27.90 - 4.28/10.25 = 10.18

USGS

DETECTION LIMIT

LIMITS IN CHEMICAL ANALYSES

In analytical chemistry, the detection limit is the lowest quantity of a substance at which it can be decided whether an analyte is present or not, a problem that arises from the fact that instruments produce readings even in the complete absence of the analyte.

The limit actually measured, x_L, is:

$$x_L = x_{bl} + k \cdot s_{bl}$$

where x_{bl} is the mean of blank measurements, s_{bl} their standard deviation, and *k* a reliability factor commonly taken equal to 3. If *S* is the sensitivity of the calibration curve, then the detection limit, *LOD*, is:

$$LOD = k \cdot s_{bl} \cdot S$$

USGS

EXAMPLE OF COPPER DETECTION (1)

In this case the interest is in the detection limit for Cu in Fe. From the measurements in Fe with actually any Cu:

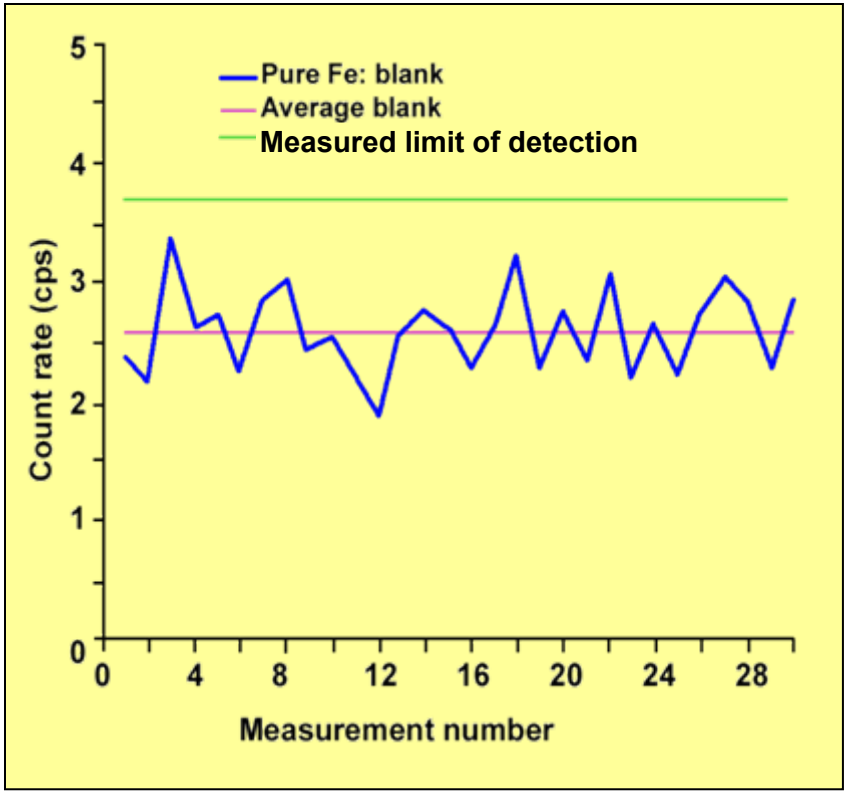
$$x_{bl} = 2.594 \ \text{cps}$$

$$s_{bl} = 0.366 \ \text{cps}$$

So, taking $k = 3$:

$$x_L = 2.594 + 3 \cdot 0.366$$

$$= 3.692$$



EXAMPLE OF COPPER DETECTION (2)

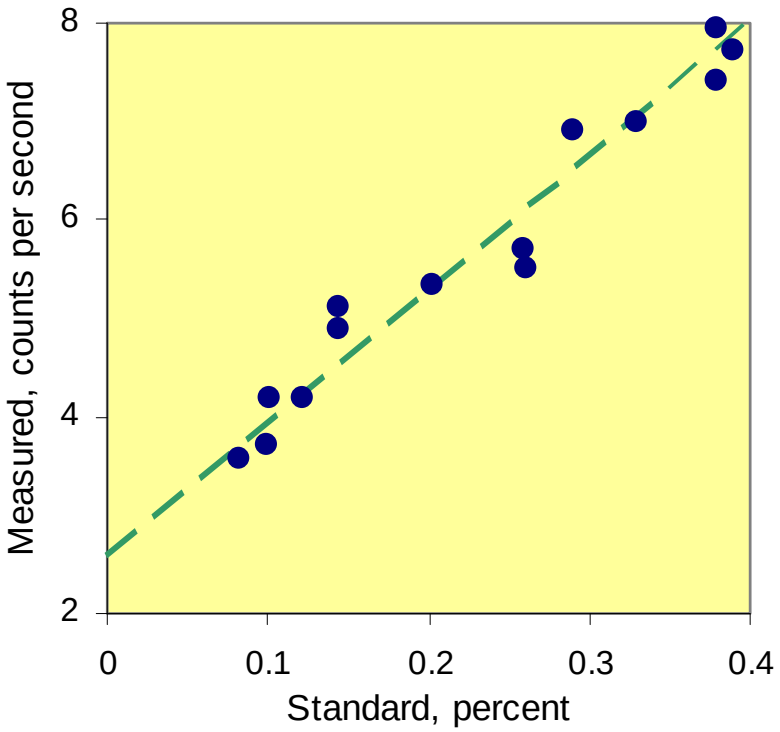
The sensitivity of the calibration curve, *S*, is the inverse of the rate of change of instrumental reading with changes in the concentration of the analyte above the detection limit, *LOD*. In our example,

$$S = 0.074\ \%/\text{cps}.$$

Thus

$$LOD = 3 \cdot 0.366 \cdot 0.074$$

$$= 0.08\ \%\text{Cu}$$



INSTRUMENTAL DETECTION LIMIT

The Limit of Detection presumes a matrix clean of other interfering substances. In such a case, the Limit of Detection is a special case of Instrumental Detection Limit, IDL:

$$LOD = IDL\,.$$

METHOD DETECTION LIMIT

In the presence of complex matrices that can complicate the calculation of the background properties, it is common to use the Method Detection Limit, *MDL*, instead of simply the Instrumental Detection Limit, *IDL*.

In the detection of copper, iron is a clean matrix and stainless steel would be a complex matrix with several alloys.

The Method Detection Limit can be anywhere from 2 to 5 times higher than the Instrumental Detection Limit

$$MDL = (2 \text{ to } 5) \cdot IDL$$

USGS

LIMIT OF QUANTITATION

A common misconception is that the *LOD* is the minimum concentration that can be measured. Instead, LOD is the concentration at which one can be reasonably certain that the concentration is not zero.

Just because one can tell something from noise does not mean that one can necessarily know how much of the analyte there actually is. Quantitation is generally agreed to start at 5 times the Method Detection Limit. This higher concentration is called Practical Quantitation Limit, *PQL*. Thus

$$PQL = 5 \cdot MDL$$

USGS

SUMMARY

In summary:

$$LOD : IDL : MDL : PQL = 1 : 1 : (2 \text{ to } 5) : (10 \text{ to } 25)$$

LEGACY DATA

LEGACY DATA

Legacy data are information in the development of which an organization has invested significant resources in its preparation and that has retained its importance, but that has been created or stored using software and/or hardware that is perceived outmoded or obsolete by current standards.

- Working with legacy data is ordinarily a difficult and frustrating task.
- Customarily, legacy data are too valuable to be ignored.
- The nature of problems is almost boundless, yet it is possible to group them in some general categories.

USGS

COMMON PROBLEMS WITH LEGACY DATA

- Data quality challenges
- Database design problems
- Data architecture problems
- Processing difficulties

TYPICAL DATA QUALITY PROBLEMS

- Different technologies in acquiring the data.
- Obsolete methods and conventions were used to prepare the data.
- The purpose of a column in a tabulation is determined by the value of one or several other columns.
- Inconsistent data formatting.
- Frequently missing data.
- Multiple codes have the same meaning.
- Varying default values.
- Multiple source for the same type of information.

COMMON DATABASE DESIGN PROBLEMS

- Inadequate documentation.
- Ineffective or no naming conventions.
- Text and numbers appear in the same column.
- Original design goals are at odds with current project needs.

COMMON ARCHITECTURE PROBLEMS

- Different hardware platforms.
- Different storage devices.
- Redundant data sources.
- Inaccessible data in obsolete media.
- Poor or lacking security.

LEGACY DATA EVALUATION

Issues to consider include:

- Are the data needed to achieve an established goal?
- What will be lost if this information is eliminated?
- Are the data consistent?
- Are the data accurate and up-to-date?
- How much data are missing?
- What is the format of the data?
- Can the new system support the same format?
- Is the information redundant?
- Is this information stored in multiple ways or multiple times?

We are all generating legacy data. Be visionary!

USGS

3. UNIVARIATE STATISTICS

EVERYTHING AND A PIECE

In statistics, population is the collection of all possible outcomes or individuals comprising the complete system of our interest; for example all people in the United States.

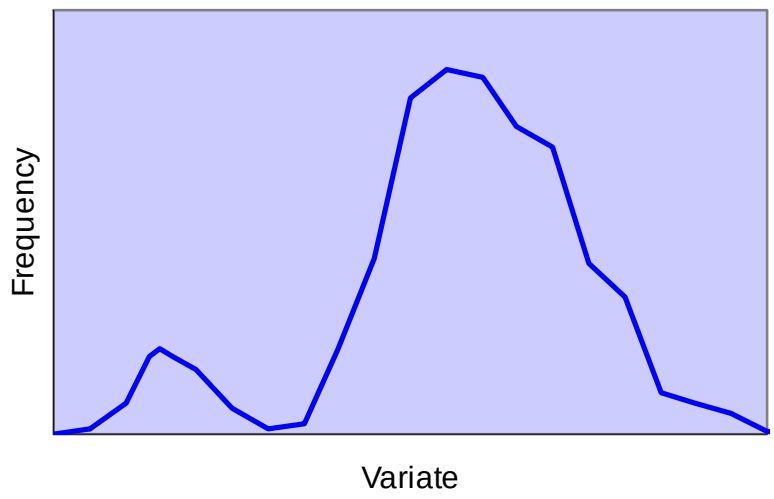
Populations may be hard or impossible to analyze exhaustively. In statistics, a limited collection of measurements is called sample; for example, a Gallup Poll.

Unfortunately, the term “sample” is employed with different meanings in geology and statistics.

Geology	Statistics
collection	sample
sample	observation

The statistical usage of the term sample is observed in what follows.

RANDOM VARIABLE



- A random variable or variate is a quantity that may take any of the values within a given set with specified relative frequencies.
- The concept is heavily utilized in statistics to characterize a population or convey the unknown value that an attribute may take.

DESCRIPTIVE ANALYSIS

- A sample of an attribute ordinarily comprises several measurements, which are best understood when organized in some way, which is an important aspect of statistics.
- The number of measurements in a sample is the sample size.
- There are multiple options to make the data more intelligible, some more convenient than others, depending on factors such as the sample size and the ultimate objectives of the study.

SAMPLE VISUALIZATION

FREQUENCY TABLE

Given some numerical information, if the interval of variation of the data is divided into class intervals— customarily of the same lengths—and all observations are assigned to their corresponding classes, the result is a count of relative frequency of the classes.

UNCF EXAMPLE FREQUENCY TABLE

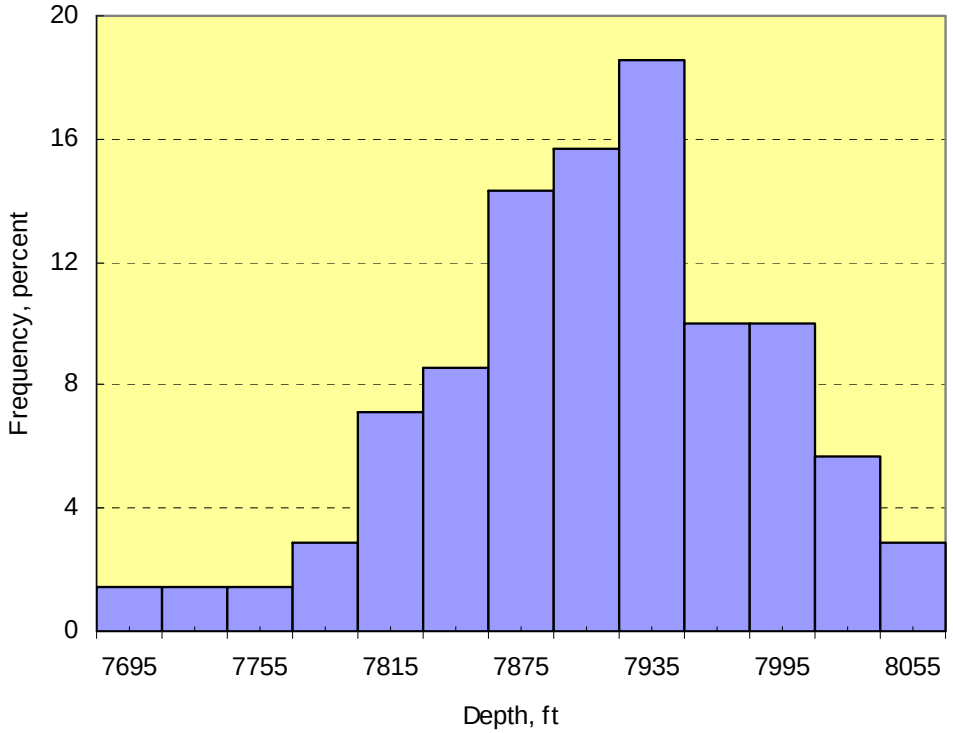
Class		Count	Frequency, %
1	7,680-7,710	1	1.43
2	7,710-7,740	1	1.43
3	7,740-7,770	1	1.43
4	7,770-7,800	2	2.86
5	7,800-7,830	5	7.14
6	7,830-7,860	6	8.57
7	7,860-7,890	10	14.29
8	7,890-7,920	11	15.71
9	7,920-7,950	13	18.57
10	7,950-7,980	7	10.00
11	7,980-8,010	7	10.00
12	8,010-8,040	4	5.71
13	8,040-8,070	2	1.43
Total		70	100.00

This example from a major oil company relates to depth in feet to an unconformity (UNCF) in an undisclosed area.

It will be used as a common reference to graphically illustrate other definitions in this chapter.

USGS

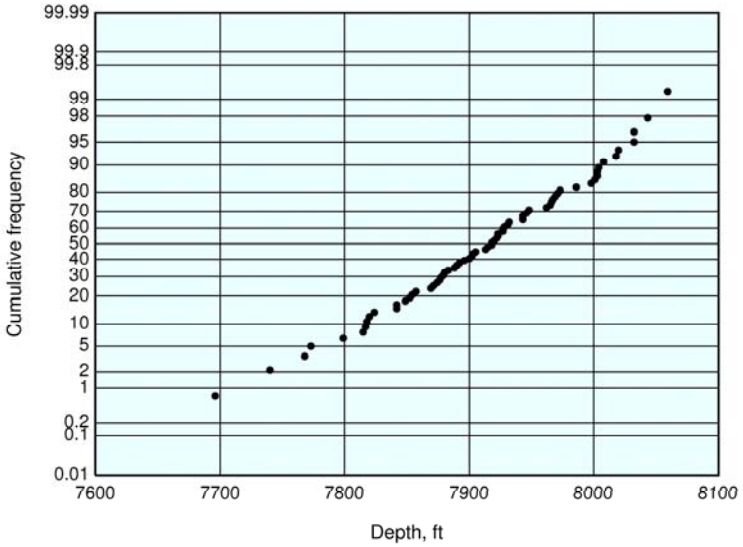
HISTOGRAM



A histogram is a graphical representation of a frequency table.

CUMULATIVE FREQUENCY

Summaries based on frequency tables depend on the selection of the class interval and origin.



Given a sample of size n, this drawback is eliminated by displaying each observation z_i versus the proportion of the sample that is not larger than z_i.

Each proportion is a multiple of $100/n$. The vertical axis is divided in n intervals and the data are displayed at the center of the corresponding interval.

Customarily, the vertical axis is scaled so that data from a normal distribution (page 69) display as a straight line.

SUMMARY STATISTICS

SUMMARY STATISTICS

Summary statistics are another alternative to histograms and cumulative distributions.

A statistic is a synoptic value calculated from a sample of observations, which is usually but not necessarily an estimator of some population parameter.

Generally, summary statistics are subdivided into three categories:

- Measures of location or centrality
- Measures of dispersion
- Measures of shape

USGS

MEASURES OF LOCATION

Measures of location give an idea about the central tendency of the data. They are:

- mean
- median
- mode

MEANS

The arithmetic mean or simply the mean, $\hat{m}$, of a sample of size n is the additive average of all the observations, z_i:

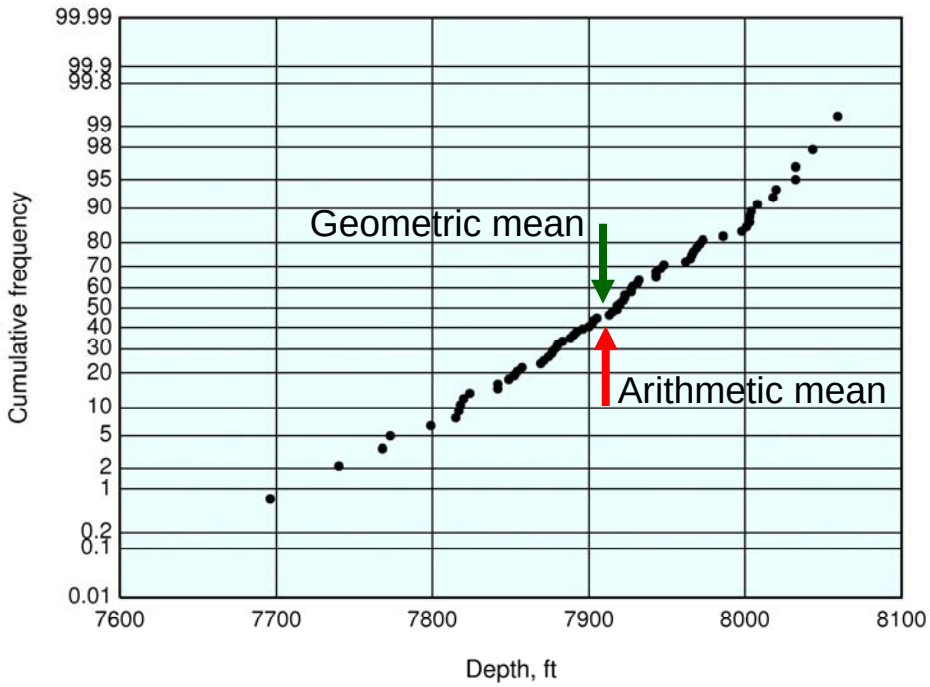
$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} z_i .$$

The less frequently used geometric mean, $\hat{m}_g$, is the nth root of the product:

$$\hat{m}_g = \left(\prod_{i=1}^{n} z_i\right)^{1/n} = \sqrt[n]{z_1 \cdot z_2 \cdot \ldots \cdot z_n} .$$

Always:

$$\hat{m} \geq \hat{m}_g .$$

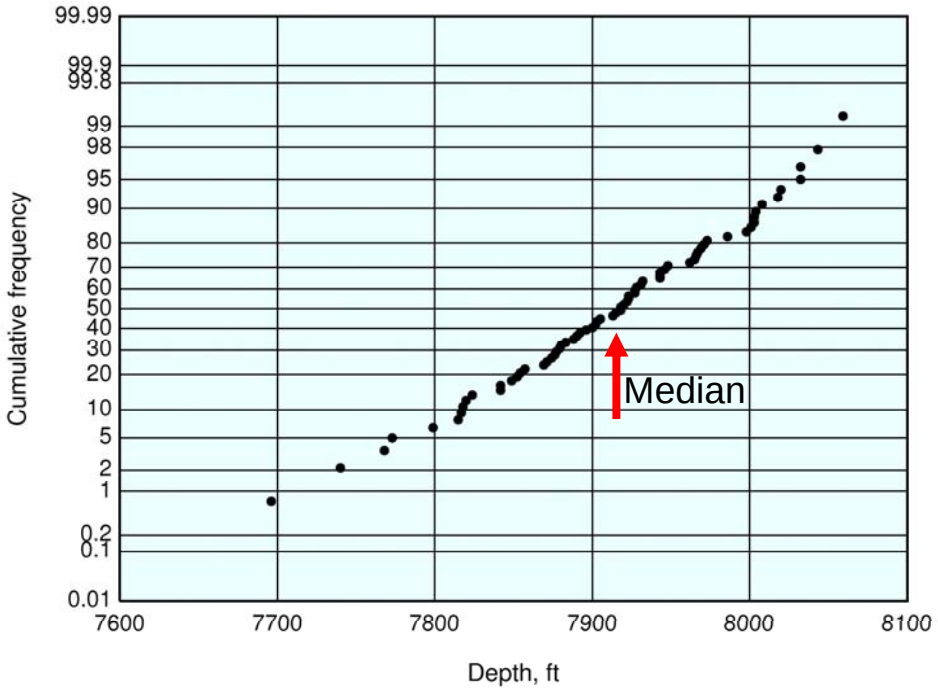


The arithmetic mean of the UNCF sample is 7912.2 ft and its geometric mean is 7911.8 ft.

THE MEDIAN

The median, Q_2, of a sample is the value that evenly splits the number of observations z_i into a lower half of smallest observations and the upper half of largest measurements.

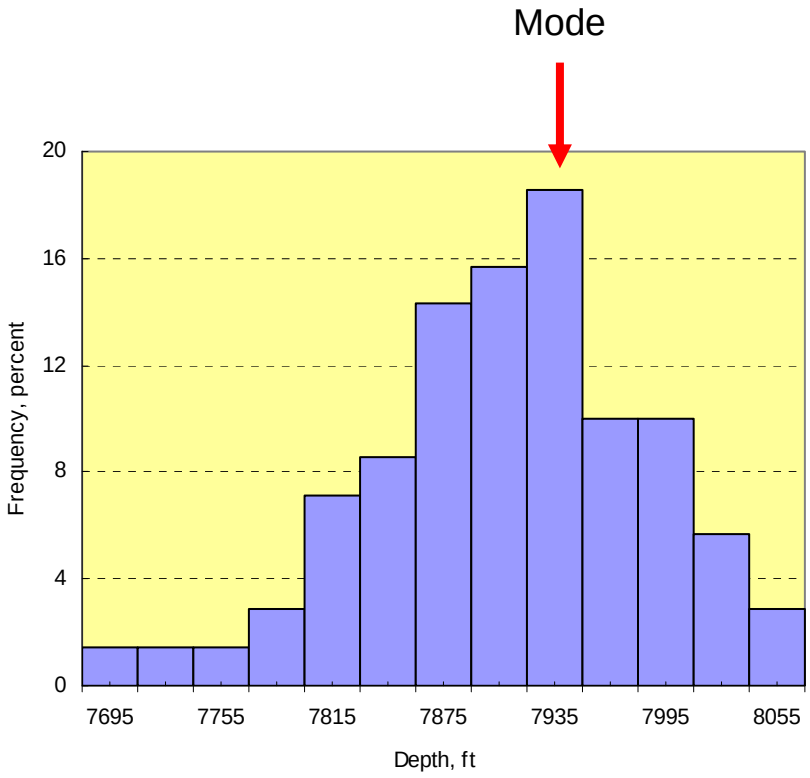
If z_i is sorted by increasing values, then



$$Q_2 = \begin{cases} z_{(n+1)/2}, \text{ if } n \text{ is odd}, \\ 0.5 \cdot \left(z_{n/2} + z_{(n/2)+1} \right), \text{ if } n \text{ is even}. \end{cases}$$

The median of the UNCF sample is 7918 ft.

THE MODE



The mode of a sample is the most probable or frequent value, or equivalently, the center point of the class containing the most observations.

For the UNCF sample, the center of the class with the most observations is 7935 ft.

ROBUSTNESS

Robustness denotes the ability of statistical methods to work well not only under ideal conditions, but in the presence of data problems, mild to moderate departures from assumptions, or both.

For example, in the presence of large errors, the median is a more robust statistic than the mean.

MEASURES OF SPREAD

Measures of spread provide an idea of the dispersion of the data. The most common measures are:

- variance
- standard deviation
- extreme values
- quantiles
- interquartile range

USGS

VARIANCE

The variance, $\hat{\sigma}^2$, is the average squared dispersion around the mean:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(z_i - m)^2 = \frac{1}{n}\left[\sum_{i=1}^{n} z_i^{\,2} - n \cdot m^2\right],$$

expressions that are commonly restricted to estimate variances of finite populations.

When dealing with samples, the denominator is often changed to *n* - 1.

Because this is a quadratic measure, it is less robust than most other measures of spread.

The variance of the UNCF sample is 5,474 sq ft.

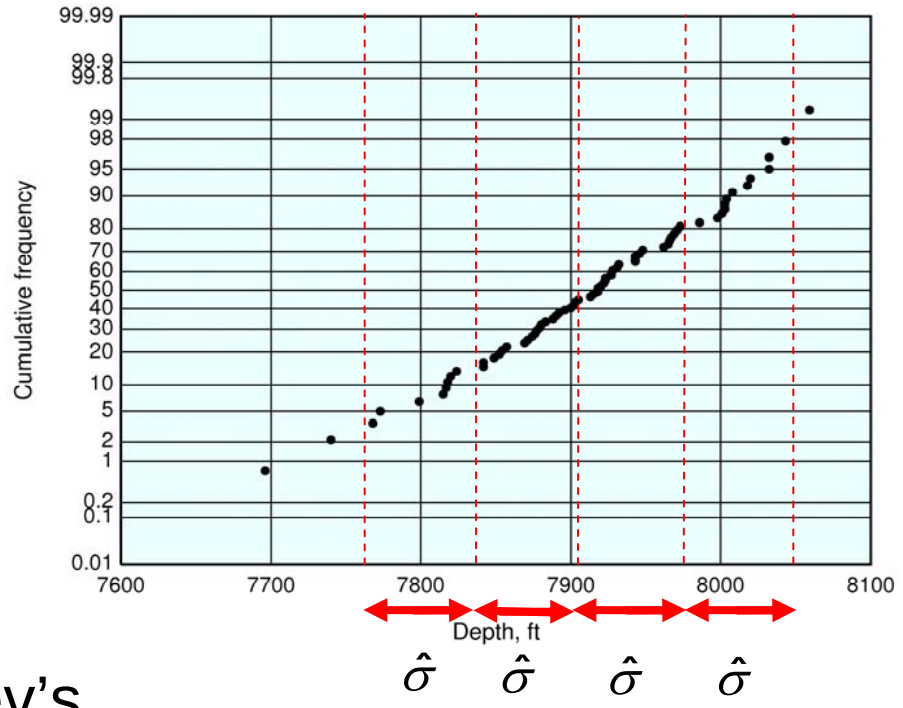
USGS

STANDARD DEVIATION

The standard deviation is the positive square root of the variance.

It has the advantage of being in the same units as the attribute.

The standard deviation of the UNCF sample is 74.5 ft.



According to Chebyshev's theorem, for any sample and $t > 1$, the proportion of data that deviates from the mean $\hat{m}$ at least $t \cdot \hat{\sigma}$ is at most t^{-2}:

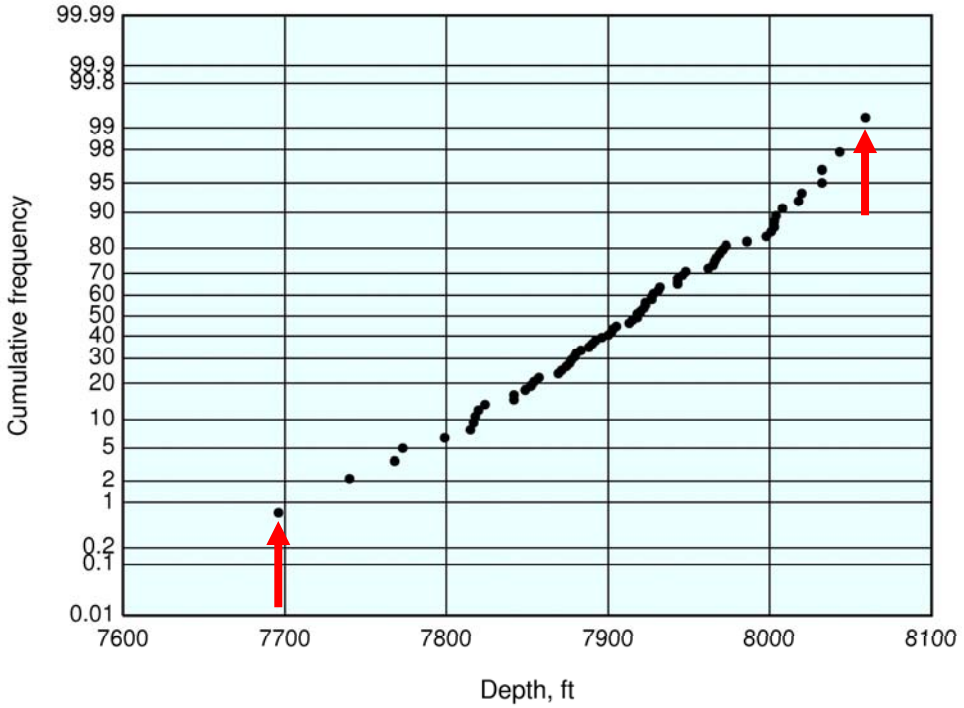
$$\text{Prop}\left(\left|X - \hat{m}\right| \geq t \cdot \hat{\sigma}\right) \leq \frac{1}{t^2}$$

EXTREME VALUES

The extreme values are the minimum and the maximum.

For the UNCF sample, the minimum value is 7,696 ft and the maximum value is 8,059 ft.

This measure is not particularly robust, especially for small samples.



QUANTILES

The idea of the median splitting the ranked sample into two equal-size halves can be generalized to any number of partitions with equal number of observations. The partition boundaries are called quantiles or fractiles. The names for the boundaries for the most common quantiles are:

- Median, for 2 partitions
- Quartiles, for 4 partitions
- Deciles, for 10 partitions
- Percentiles, for 100 partitions

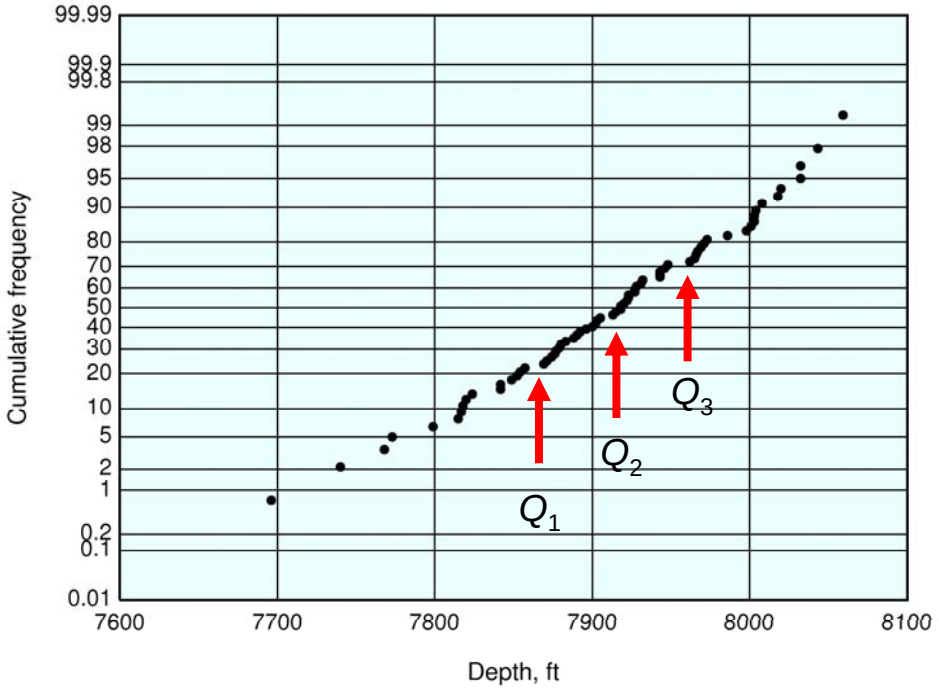
There is always one less boundary than the number of partitions.

UNCF QUARTILES

Q_1 = 7871.75 ft

Q_2 = 7918 ft

Q_3 = 7965.75 ft



Q_2 coincides with the median.

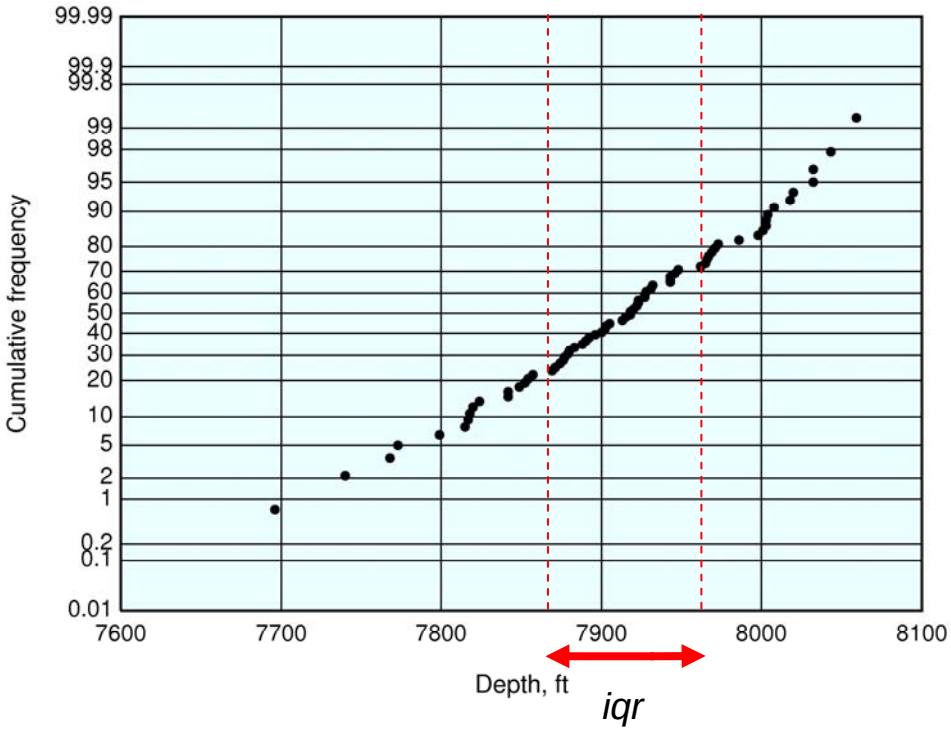
INTERQUARTILE RANGE

The interquartile range, *iqr*, is the difference between the upper and the lower quartiles

$$iqr = Q_3 - Q_1,$$

thus measuring the data central spread.

For the UNCF sample, the interquartile range is 94 ft.

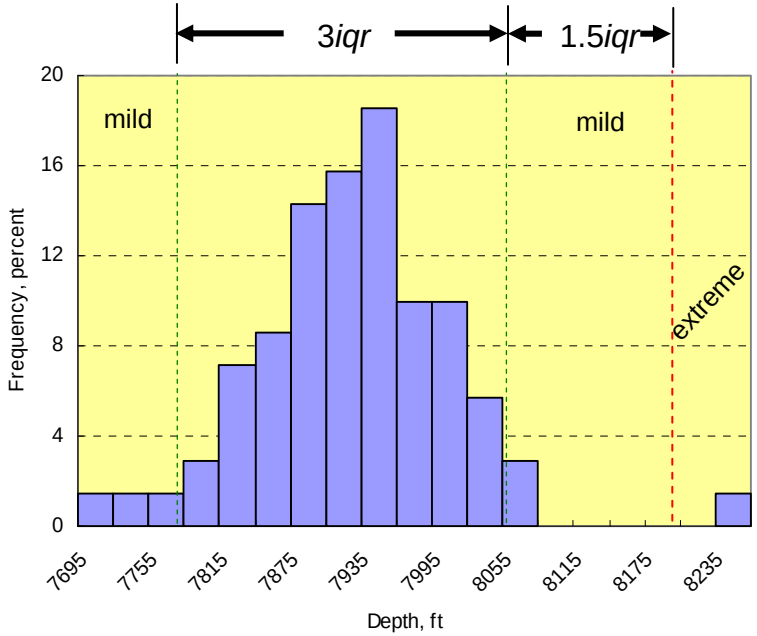


The interquantile range is more robust than the variance but insensitive to values in the lower and upper tails.

OUTLIER

Outliers are values so markedly different from the rest of the sample that they rise the suspicion that they may be from a different population or that they may be in error, doubts that frequently are hard to clarify. In any sample, outliers are always few, if any.

A practical rule of thumb is to regard any value deviating more than 1.5 times the interquartile range, *iqr*, from the median as a mild outlier and a value departing more than 3 times *iqr* as an extreme outlier.



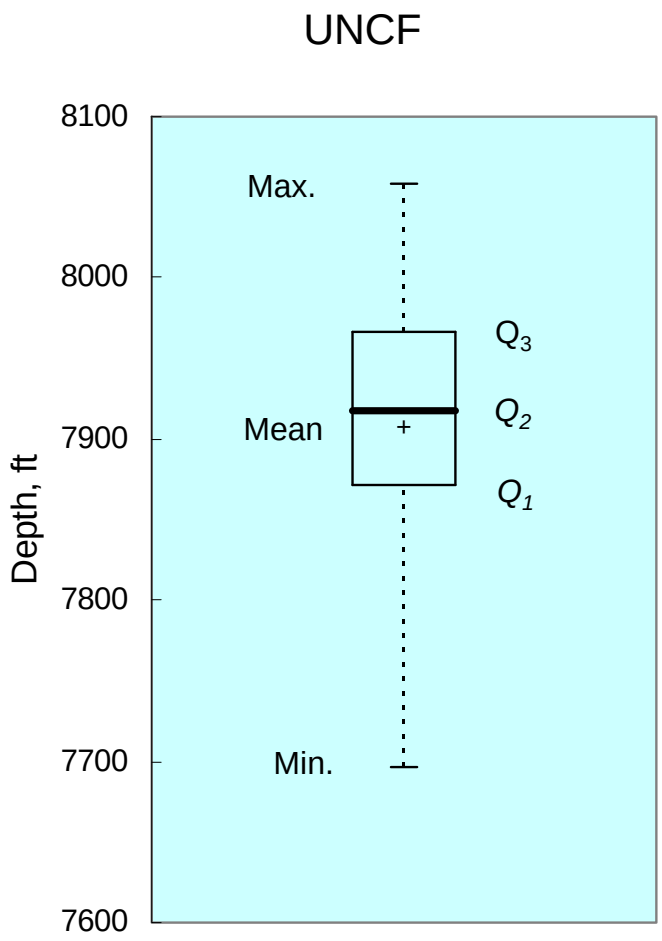
For the UNCF sample, all mild outliers seem to be legitimate values, while the extreme outlier of 8,240 ft is an error.

BOX-AND-WHISKER PLOT

The box-and whisker plot is a simple graphical way to summarize several of the statistics:

- Minimum
- Quartiles
- Maximum
- Mean

Variations of this presentation abound. Extremes may exclude outliers, in which case the outliers are individually posted as open circles. Extremes sometimes are replaced by the 5 and 95 percentiles.



MEASURES OF SHAPE

The most commonly used measures of shape in the distribution of values are:

- Coefficient of skewness
- Quartile skew coefficient
- Coefficient of kurtosis

USGS

COEFFICIENT OF SKEWNESS

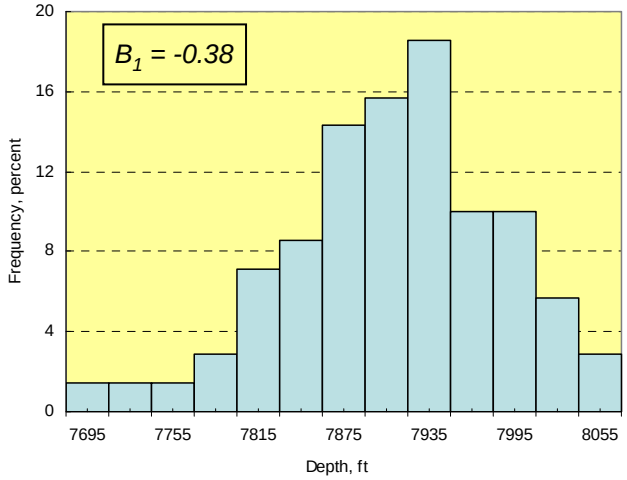
The coefficient of skewness is a measure of asymmetry of the histogram. It is given by:

$$B_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - m)^3}{\sigma^3}$$

If : $B_1 < 0$, the left tail is longer;

$B_1 = 0$, the distribution is symmetric;

$B_1 > 0$, the right tail is longer.

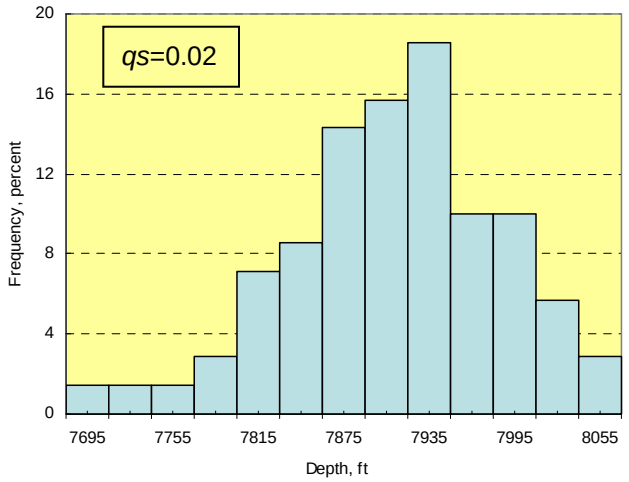


The UNCF coefficient of skewness is -0.38.

QUARTILE SKEW COEFFICIENT

The quartile skew coefficient serves the same purpose as the coefficient of skewness, but it is more robust, yet only sensitive to the central part of the distribution. Its definition is:

$$qs = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{iqr}$$



If : $qs < 0$, the left tail is longer;

$qs = 0$, the distribution is symmetric;

$qs > 0$, the right tail is longer.

The UNCF quartile skew coefficient is 0.02.

COEFFICIENT OF KURTOSIS

This statistic measures the concentration of values around the mean. Its definition is:

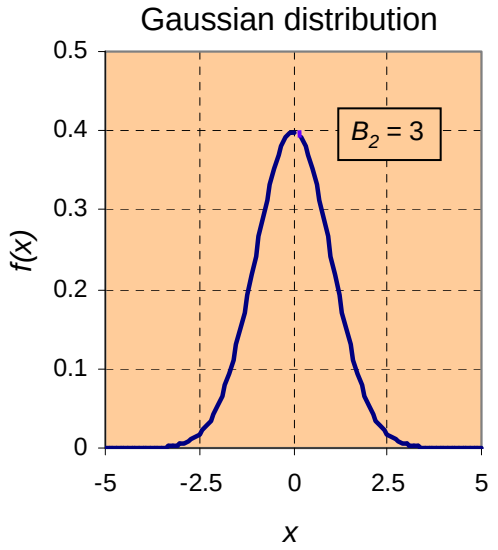
$$B_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(z_i - m\right)^4}{\sigma^4}$$

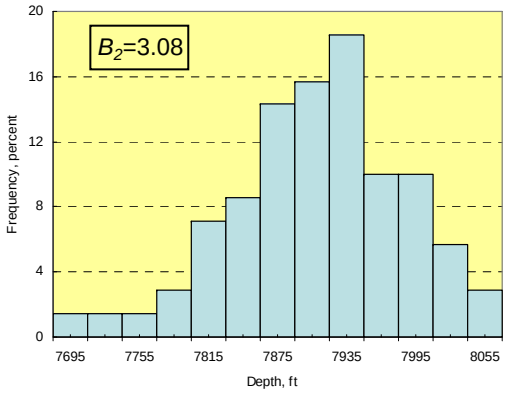
If : $B_2 < 3$, the distribution is more peaked than Gaussian;

$B_2 = 3$, it is as peaked as Gaussian;

$B_2 > 3$, it is less peaked than Gaussian.

The UNCF coefficient of kurtosis is 3.08.





MODELS

PROBABILITY

Probability is a measure of the likelihood that an event, *A*, may occur. It is commonly denoted by Pr[*A*].

- The probability of an impossible event is zero, Pr[*A*] = 0. It is the lowest possible probability.
- The maximum probability is Pr[*A*] = 1, which denotes certainty.
- When two events *A* and *B* cannot take place simultaneously, Pr[*A* or *B*] = Pr[*A*] + Pr[*B*].
- Frequentists claim that Pr[*A*] = N_A/N, where *N* is total number of outcomes and N_A the number of outcomes of *A.* The outcomes can be counted theoretically or experimentally.
- For others, a probability is a degree of belief in *A*, even if no random process is involved nor a count is possible.

USGS

BAYES'S THEOREM

This is one of the most widely used probability relationships. If event *B* already occurred, the conditional probability of event *A,* Pr[*A*|*B*], is:

$$\Pr[A \mid B] = \frac{\Pr[B \mid A]}{\Pr[B]} \Pr[A]$$

Example
Suppose we have two boxes. A blue ball is drawn (event *B)*. What is the probability it came from box #1 (event *A*)?

Box	Number of balls		
	Blue	Red	Total
#1	20	5	25
#2	12	18	30
	32	23	55

- In the absence of additional information, Pr[*A*] = ½ = 0.5.
- Pr[*B*|*A*] = 20/25 = 0.8.
- Pr[*B*] = 32/55 = 0.59. Hence
- $$\Pr[A \mid B] = \frac{0.8}{0.59} 0.5 = 0.69$$

USGS

PROBABILITY FUNCTIONS

Analytical functions approximating experimental fluctuations are the alternative to numerical descriptors and measures. They provide approximations to general conditions. The drawback is loss of fine details in favor of simpler models.

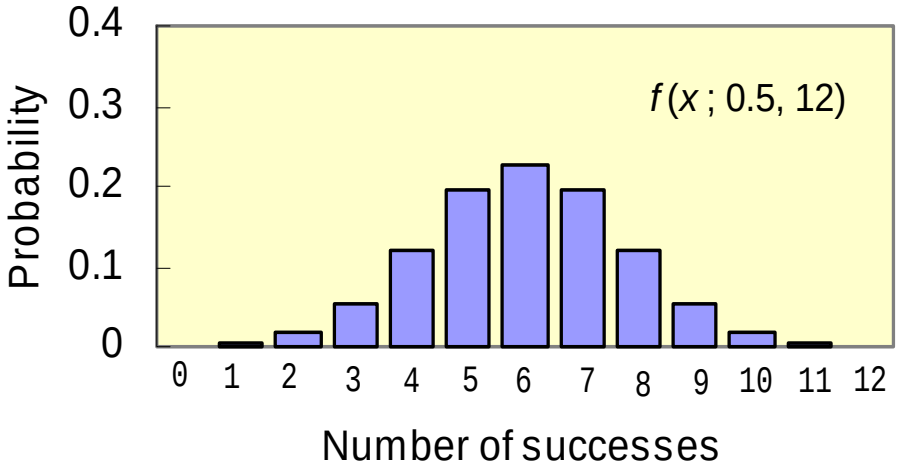
Analytical expressions approximating histograms are called probability density functions; those modeling cumulative frequencies are the cumulative distribution functions.

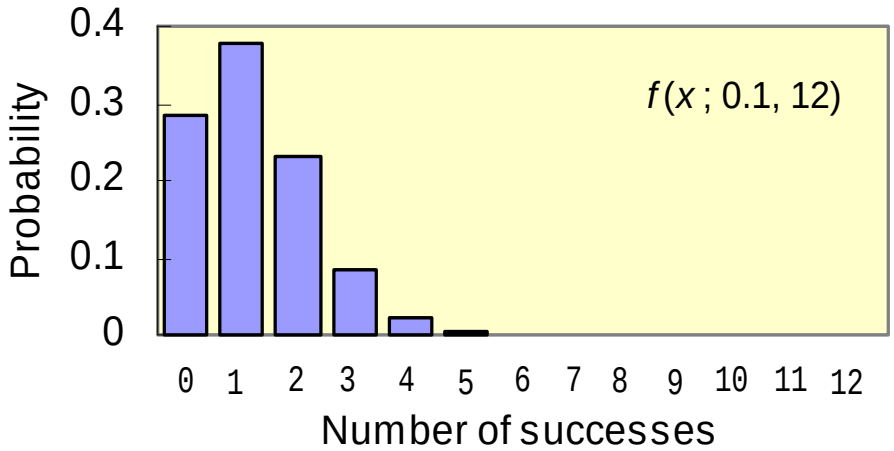
Variations in the parameters in an expression allow the generation of a family of functions, sometimes with radically different shapes.

Main subdivision is into discrete and continuous functions, of which the binomial and normal distribution, respectively, are two typical and common examples.

BINOMIAL DISTRIBUTION

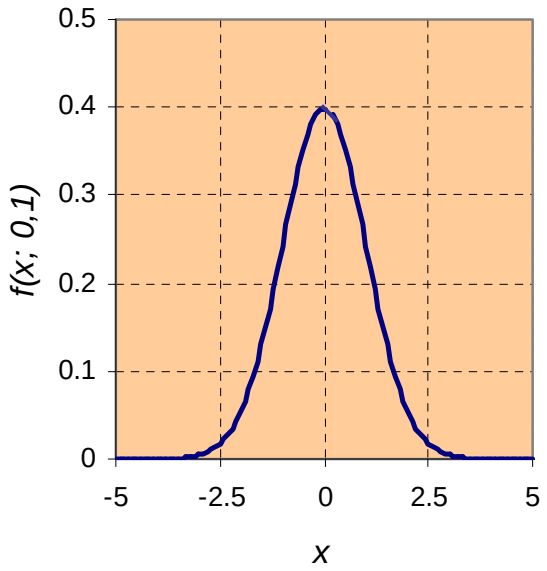
This is the discrete probability density function, $f(x;p,n)$, of the number of successes in a series of independent (Bernoulli) trials, such as head or tails in coin flipping. If the probability of success at every trial is p, the probability of x successes in n independent trials is





$$f(x;p,n)=\frac{n!}{x!(n-x)!}p^{x}(1-p)^{n-x},\ x=0,1,2,\ldots,n$$

NORMAL DISTRIBUTION



The most versatile of all continuous models is the normal distribution, also known as Gaussian distribution. Its parameters are μ and σ, which coincide with the mean and the standard deviation.

$$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},-\infty<x<\infty$$

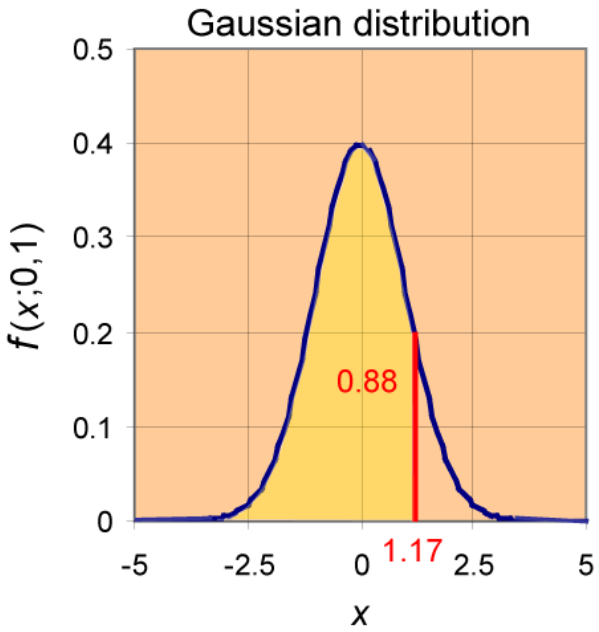
If $X = \log(Y)$ is normally distributed, Y is said to follow a lognormal distribution. Lognormal distributions are positively defined and positively skewed.

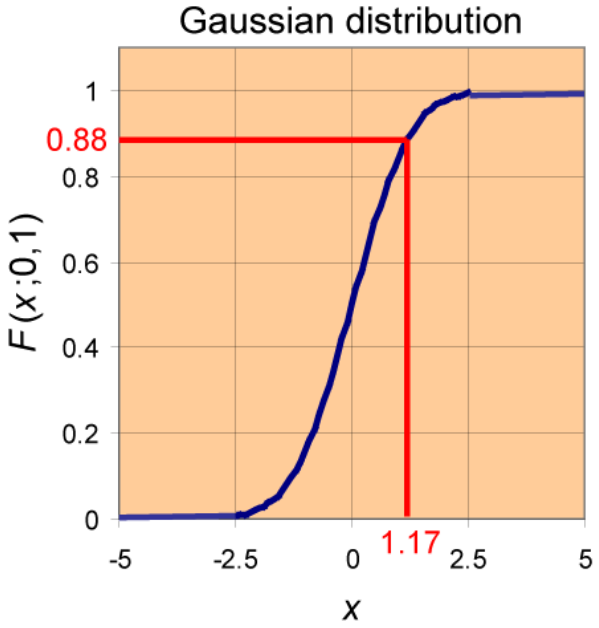
PROBABILITY FROM MODELS

$$\text{Prob}[X \le x_1] = \int_{-\infty}^{x_1} f(x)dx$$

$$\text{Prob}[X \le x_1] = F(x_1)$$

Examples:





$$\text{Prob}[X \le 1.17] = \int_{-\infty}^{1.17} \text{Normal}(x;0,1)dx = 0.88$$

$$\text{Prob}[X \le 1.17] = F(1.17) = 0.88$$

EXPECTED VALUE

Let X be a random variable having a probability distribution $f(x)$, and let $u(x)$ be a function of x. The expected value of $u(x)$ is denoted by the operator $\mathrm{E}[u(x)]$ and it is the probability weighted average value of $u(x)$.

Continuous case, such as temperature:

$$\mathrm{E}[u(x)] = \int_{-\infty}^{\infty} u(x) f(x) dx$$

Discrete case, like coin flipping:

$$\mathrm{E}[u(x)] = \sum_{x} u(x) f(x)$$

EXAMPLE

For the trivial case

$$u(x) = x$$

in the discrete case, if all values are equally probable, the expected value turns into

$$\mathrm{E}[x] = \frac{1}{n}\sum_{x} x$$

which is exactly the definition of the mean.

USGS

MOMENT

Moment is the name given to the expected value when the function of the random variable, if it exists, takes the form $(x-a)^k$, where k is an integer larger than zero, called the order.

If a is the mean, then the moment is a central moment.

The central moment of order 2 is the variance. For an equally probable discrete case,

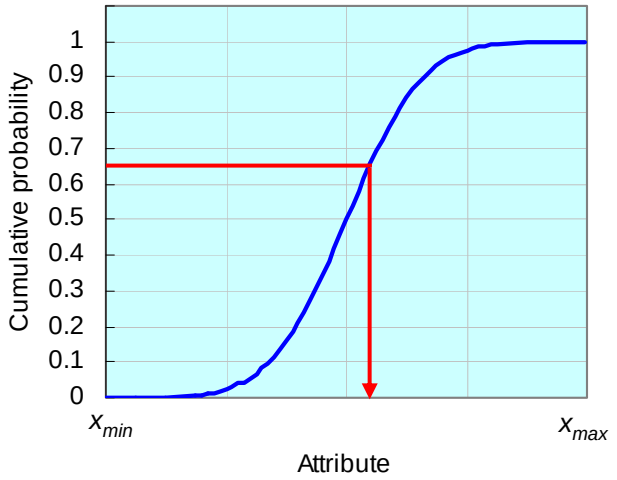
$$M_2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2$$

SIMULATION

SIMULATION

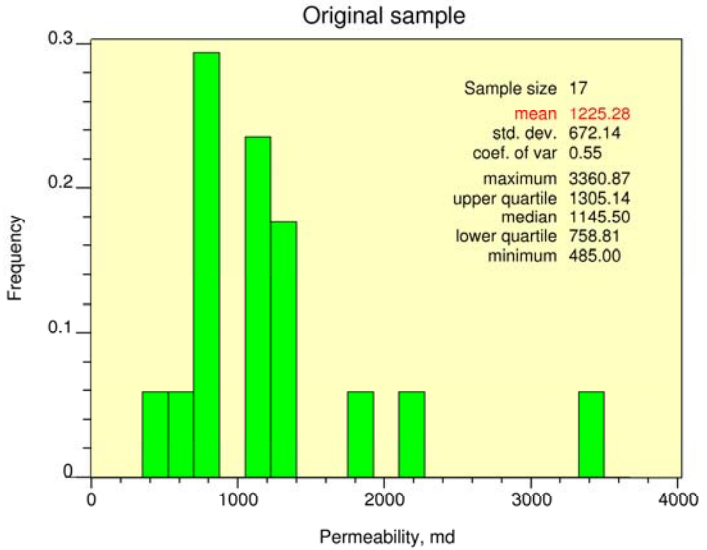
Reality is often too complex to be able to analytically derive probabilities as in flipping coins. Modern approaches use computers to model a system and observe it repeatedly, with each repetition being the equivalent of flipping a coin.

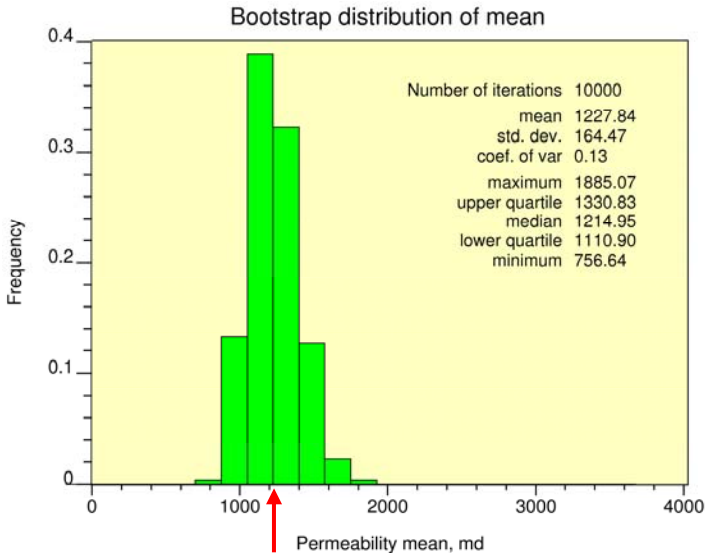
- Relevant attributes take the form of random variables instead of being deterministic variables.
- Generators of random sequences of numbers between 0-1 play an important part in simulation.
- The other important component is the Monte Carlo method, which allows drawing values from the distributions.
- The final result is the histogram(s) for the output variable(s).

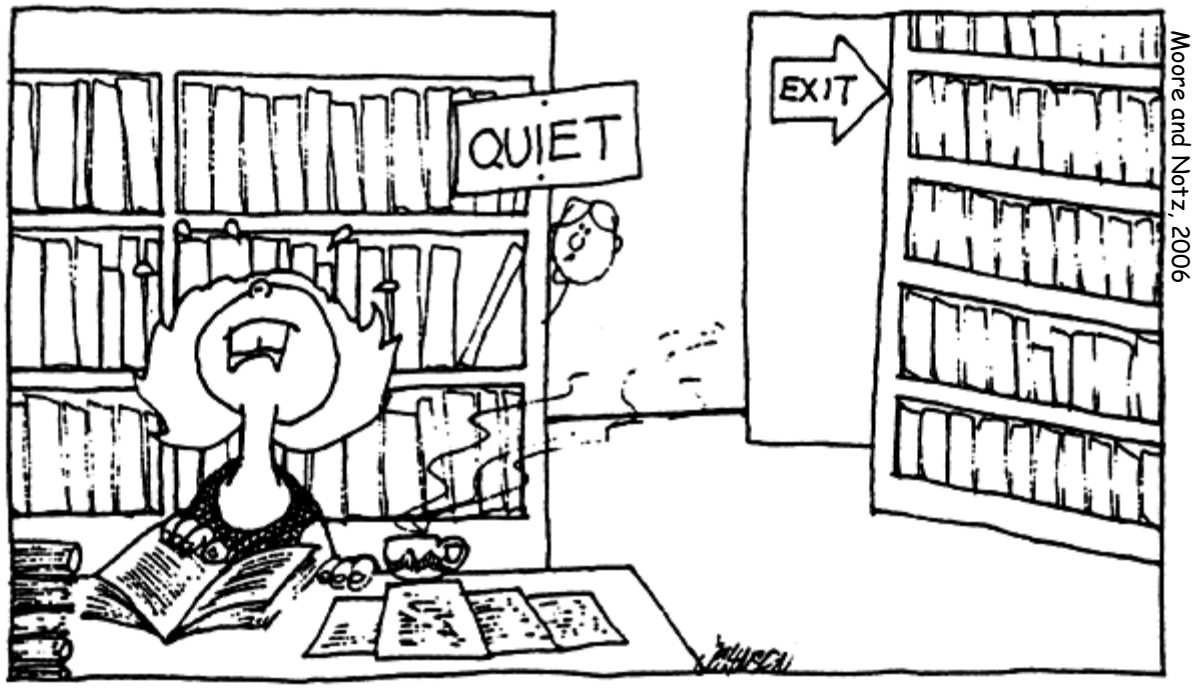


BOOTSTRAP

- This is a resampling form of the Monte Carlo method for the numerical modeling of distributions.
- It works directly with the sample instead of its distribution.
- In its simplest form, the general steps are:
  1. Randomly pick as many measurements as the sample size. Some values may be taken more than once, while others not at all.
  2. Calculate and save as many statistics as interested.
  3. Go back to step 1 and repeat the process at least 1,000 times.







Moore and Notz, 2006

"I've had it! Simulated wood, simulated leather, simulated coffee, and now simulated probabilities!"

4. BIVARIATE STATISTICS

TOOLS

Frequently there is interest in comparing two or more measurements made for the same object or site. Among the most common alternatives, we have:

- Scatterplot
- Correlation coefficient
- Regression
- Quantile-Quantile plot
- Probability-Probability plot

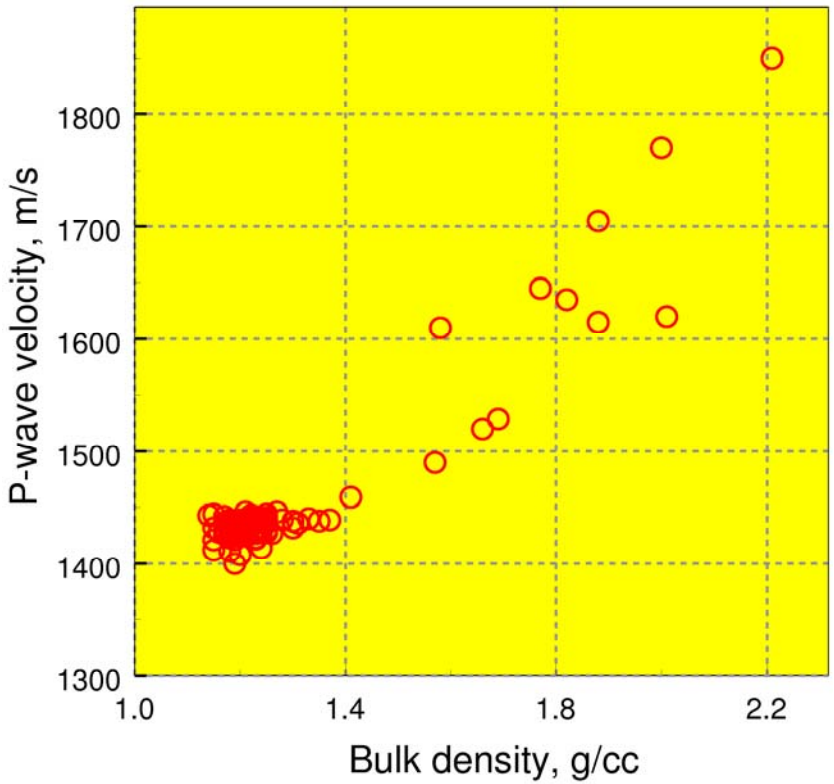
Some of these concepts can be generalized to more than two variables, and all multivariate techniques in Chapter 7 are valid for two variables.

SCATTERPLOTS

A bivariate scatterplot is a Cartesian posting in which the abscissa and the ordinate are any two variables consistently measured for a series of objects.

Scatterplots are prepared for exploring or revealing form, direction, and strength of association between two attributes.

Mecklenburg Bay seafloor, Germany



COVARIANCE

Given two random variables *X* and *Y* with means μ_x and μ_Y, their covariance is the expected value:

$$\mathrm{Cov}_{X,Y} = \mathrm{E}\left[\left(X - \mu_X\right)\left(Y - \mu_Y\right)\right]$$

The covariance estimator when using a sample of point measurements is:

$$\hat{\mathrm{Cov}}_{X,Y} = \frac{1}{n-1}\sum_{i=1}^{n} x_i \cdot y_i - \frac{1}{n \cdot (n-1)}\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i$$

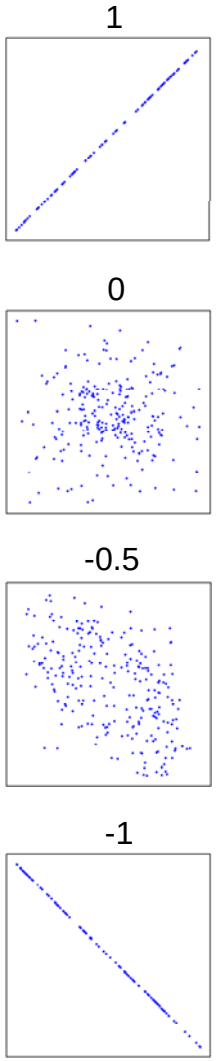
The covariance is a measure of joint variation.

CORRELATION COEFFICIENT

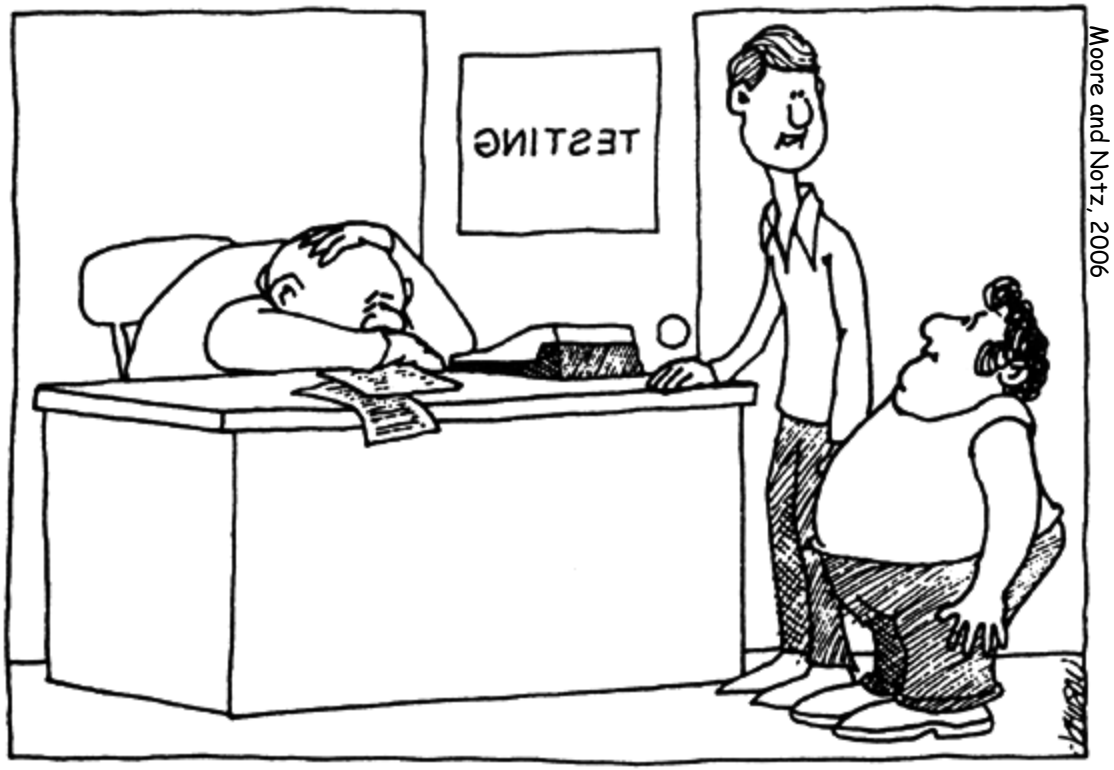
This coefficient is the number most commonly used to summarize bivariate comparisons. If σ_X and σ_Y are the standard deviations for two variables, their correlation coefficient, ρ, is given by:

$$\rho = \frac{\text{Cov}_{X,Y}}{\sigma_X \cdot \sigma_Y}$$

- It only makes sense to employ ρ for assessing linear associations.
- ρ varies continuously from -1 to 1:
  1, perfect direct linear correlation
  0, no linear correlation
  -1, perfectly inverse correlation



SENSITIVITY TO OUTLIERS



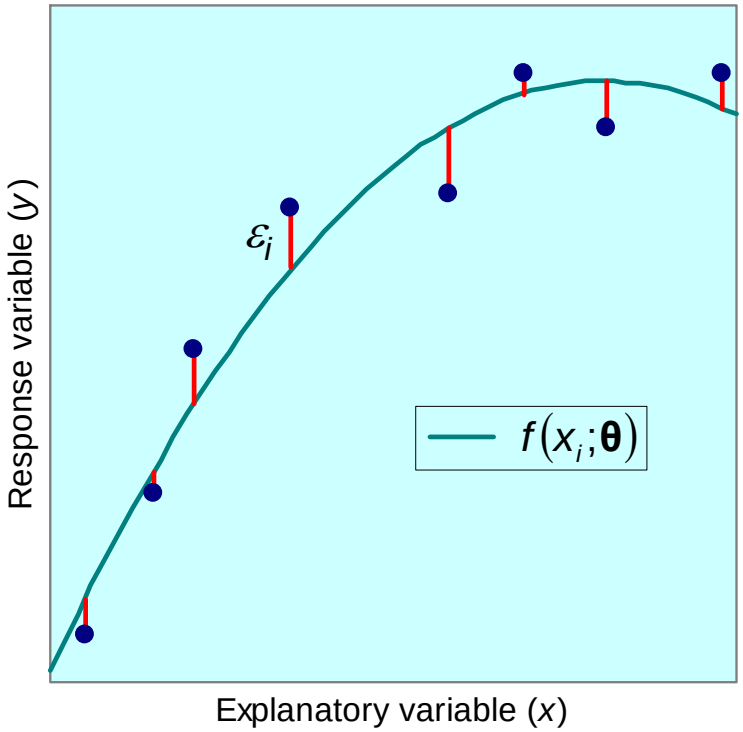
"He says we've ruined his positive correlation between height and weight."

REGRESSION

Regression is a method for establishing analytical dependency of one or more variables on another mainly to determine the degree of their dependency, to estimate values not included in the sampling, or to summarize the sampling.

- The variable in the abscissa is called the regressor, independent, or explanatory variable.
- The variable in the ordinate is the regressed, dependent, or response variable.
- In many studies, which variable goes into which axis is an arbitrary decision, but the result is different. Causality or the physics of the process may help in solving the indetermination.

REGRESSION MODEL



- The model is:

$$y_i = f(x_i;\boldsymbol{\theta}) + \varepsilon_i$$

- $f(x_i;\boldsymbol{\theta})$ is any continuous function of x that is judiciously selected by the user. $\boldsymbol{\theta}$ are unknown parameters.
- Term ε is a random variable accounting for the error.
- Having selected $f(x_i;\boldsymbol{\theta})$, parameters $\boldsymbol{\theta}$ are calculated by minimizing total error, for which there are several methods.

Avoid using the model outside the sample range of the explanatory variable.

USGS

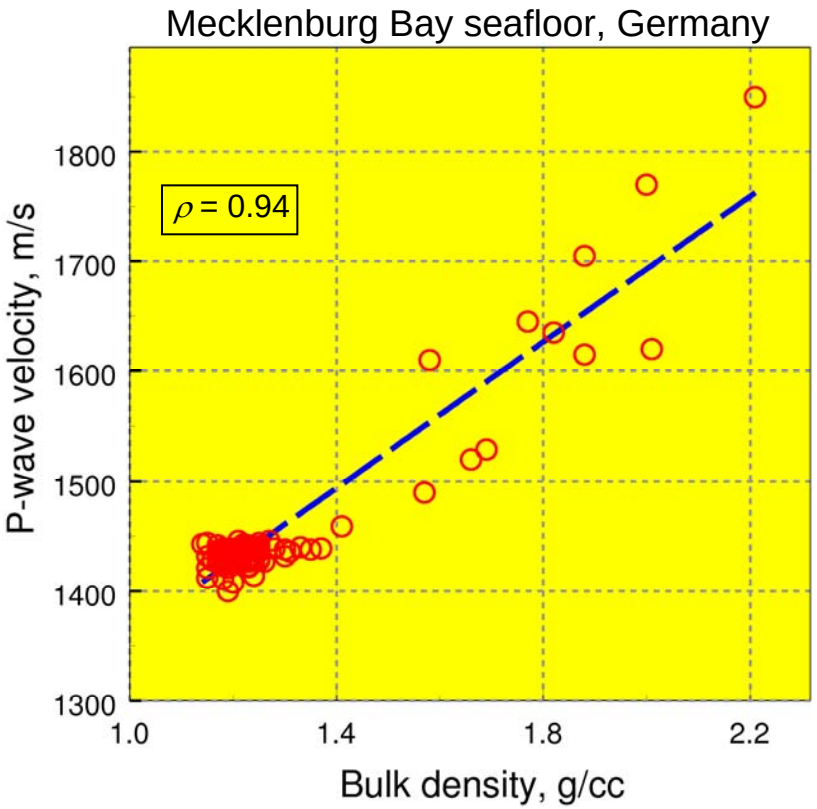
LINEAR REGRESSION

The simplest case is linear regression with parameters obtained minimizing the mean square error, MS_E:

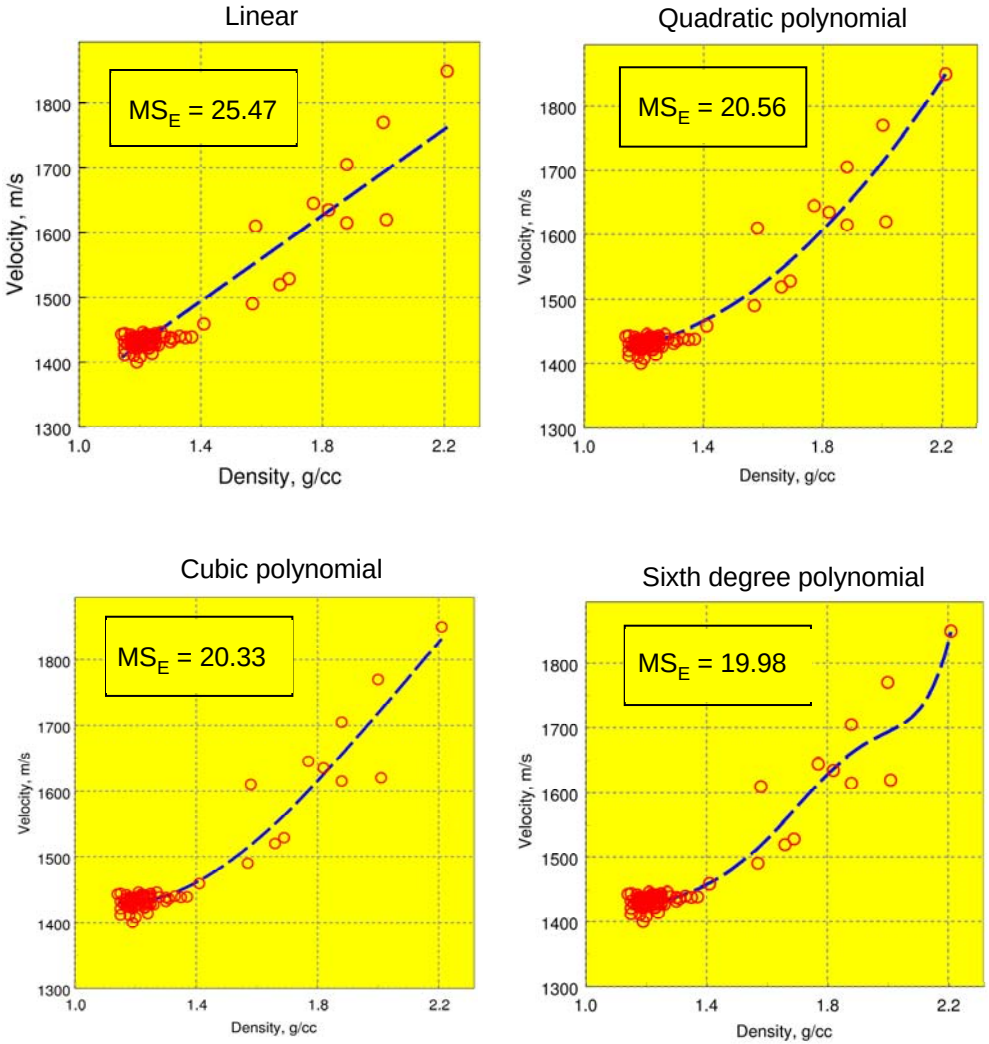
$$MS_E = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$$

In this special situation, ρ^2 accounts for the proportion of variation accounted by the model.

In the example, $\rho = 0.94$. Hence, in this case, the linear regression explains 88% $\left(100 \cdot \rho^2\right)$ of the variation.

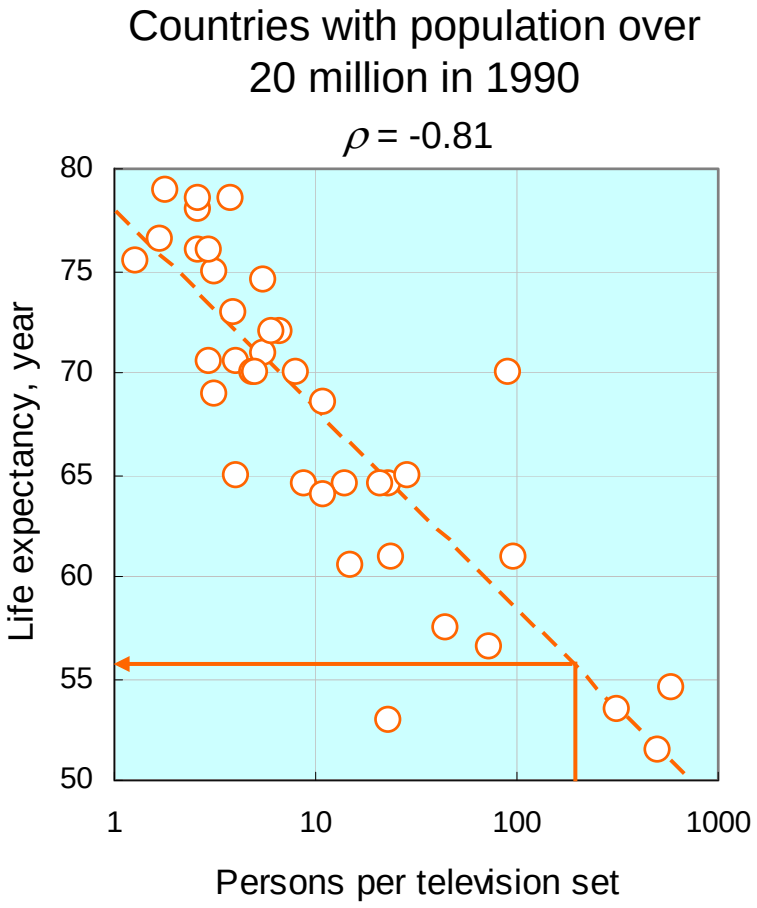


NONLINEAR REGRESSION



- In theory, the higher the polynomial degree, the better the fit.
- In practice, the higher the polynomial, the less robust the solution.
- Overfitting may capture noise and not systematic variation.

IMPLICATIONS



High to good correlation:

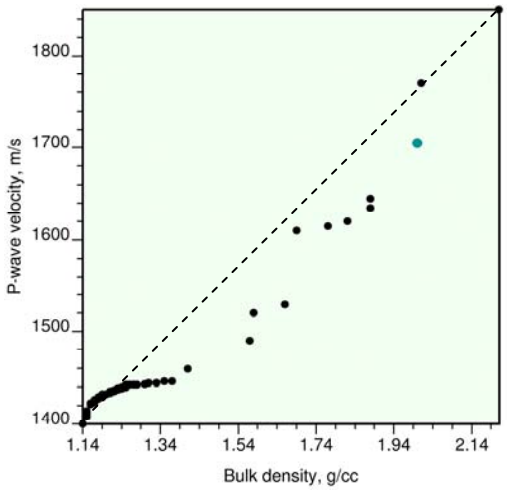
- allows prediction of one variable when only the other is known, but the inference may be inaccurate, particularly if the correlation coefficient is low;
- means the variables are related, but the association may be caused by a common link to a third lurking variable, making the relationship meaningless;
- does not necessarily imply cause-and-effect.

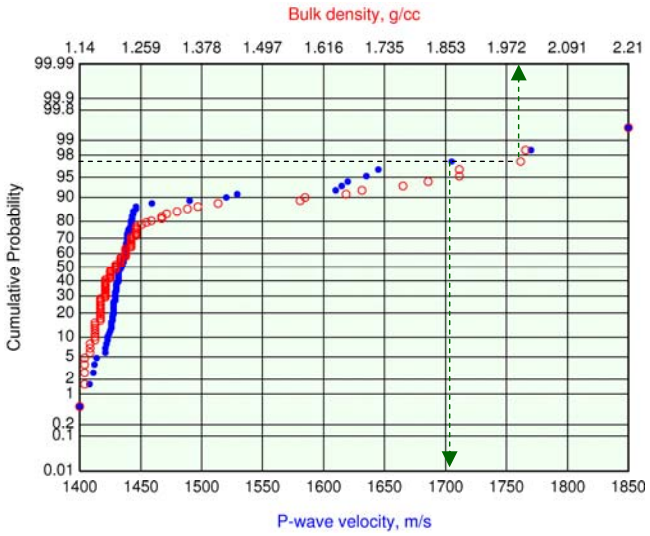
After Moore and Notz, 2006

"In a new effort to increase third-world life expectancy, aid organizations today began delivery of 100,000 television sets."

QUANTILE-QUANTILE PLOT

- A quantile-quantile or Q-Q plot is a scatterplot based on ranked data.
- The paring is independent from the object or site where the observations were taken. The first pair of coordinates has the minimum value for each attribute, the second pair is made of the second smallest readings, and so on until finishing with the two maximum values. Interpolations are necessary for different size samples.
- Identical distributions generate a straight line along the main diagonal.
- Q-Q plots are sensitive to a shift and scaling of the distributions.

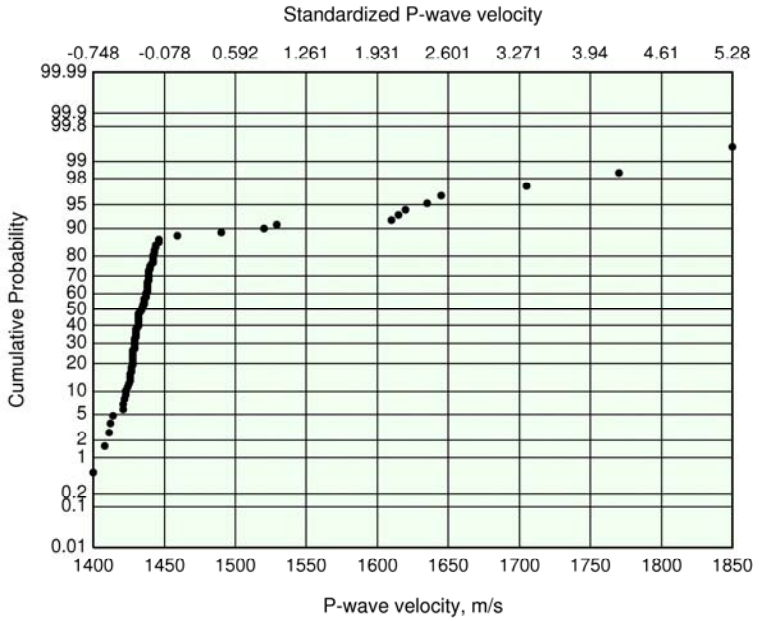




STANDARDIZED VARIATE

If X is a random variable with mean μ and standard deviation σ, the standardized variate, Z, is the transformation:

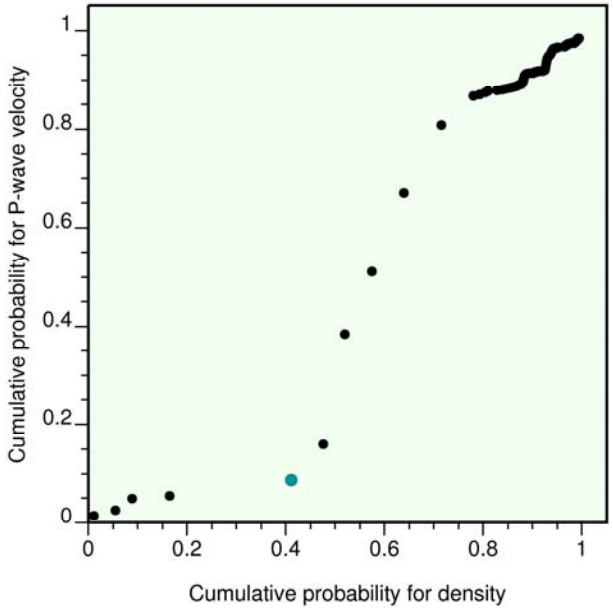
$$Z = \frac{X - \mu}{\sigma}$$

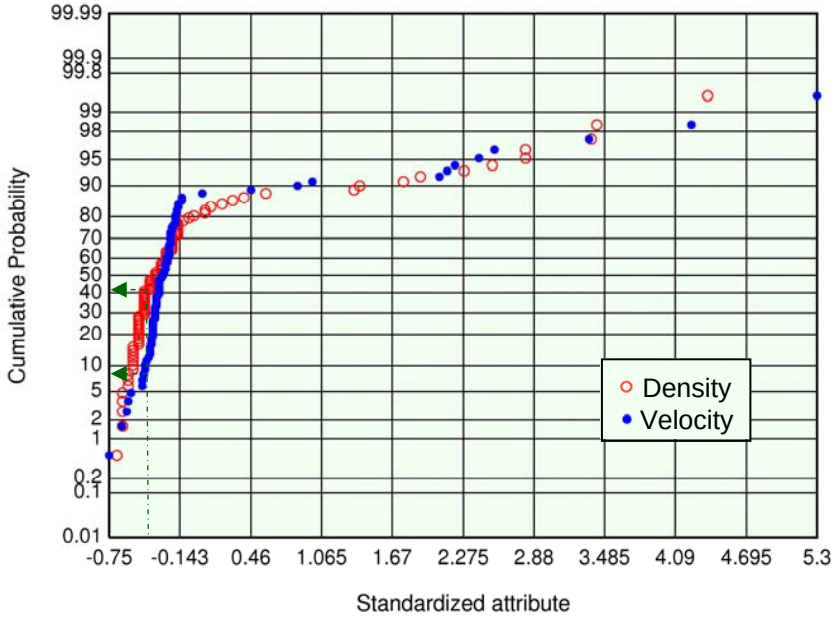


A standardized variate always has a mean of zero and variance of one.

A standardized Gaussian distribution is called a standard normal distribution. It is often denoted by $N(0,1)$.

PROBABILITY-PROBABILITY (P-P) PLOT





A P-P plot is another scatterplot prepared by extracting information from the cumulative distributions of two variates.

- If the variates are in different units, preliminary standardization is necessary.
- For given thresholds, the axes show the cumulative probabilities for the two distributions being compared.

SPECIAL TOPICS

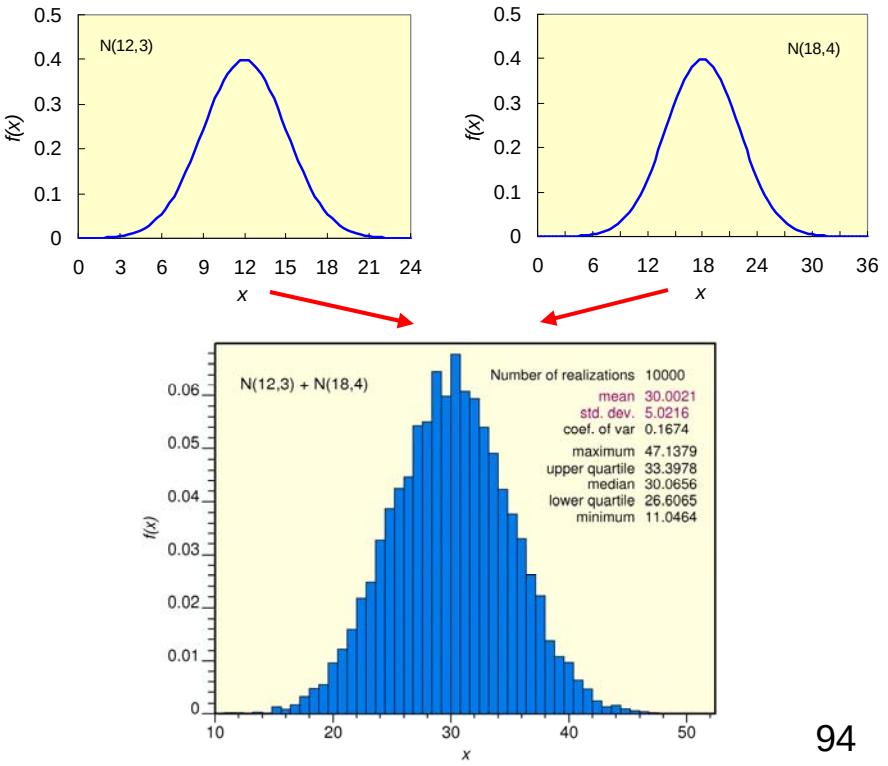
PROPAGATION OF ERRORS

Error propagation is a rather archaic term referring to the influence that uncertainty in variables have on a function of them.

- Under certain simple conditions, it is possible to analytically express the impact of variable uncertainty on the function uncertainty.
- Today the situation is resolved through numerical simulation.

Analytically:
N(30,5) = N(12,3) + N(18,4)

Monte Carlo simulation:



BIAS

Bias in sampling denotes preference in taking the measurements.

Bias in an estimator implies that the calculations are preferentially loaded in one direction relative to the true value, either systematically too high or too low.

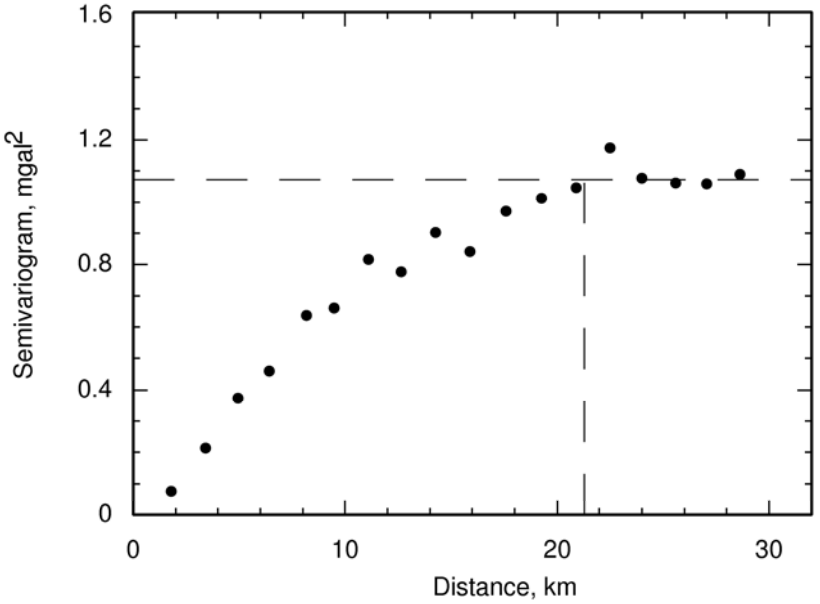
- When undetected or not compensated, bias induces erroneous results.
- The opposite of the quality of being biased is to be unbiased.

SEMIVARIOGRAM

Let $\mathbf{s}_i$ be the geographical location of a sampling site.

The semivariogram is a squared average comparison of the degree of dissimilarity with geographical distance, **h**, for pairs of measurements of *Z* at sites **s** and **s** + **h**:

$$\gamma(\mathbf{h}) = \frac{1}{2}\mathrm{E}\left[\{Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})\}^2\right]$$



If N_h is the number of pairs a distance **h** apart, the most popular unbiased estimate is:

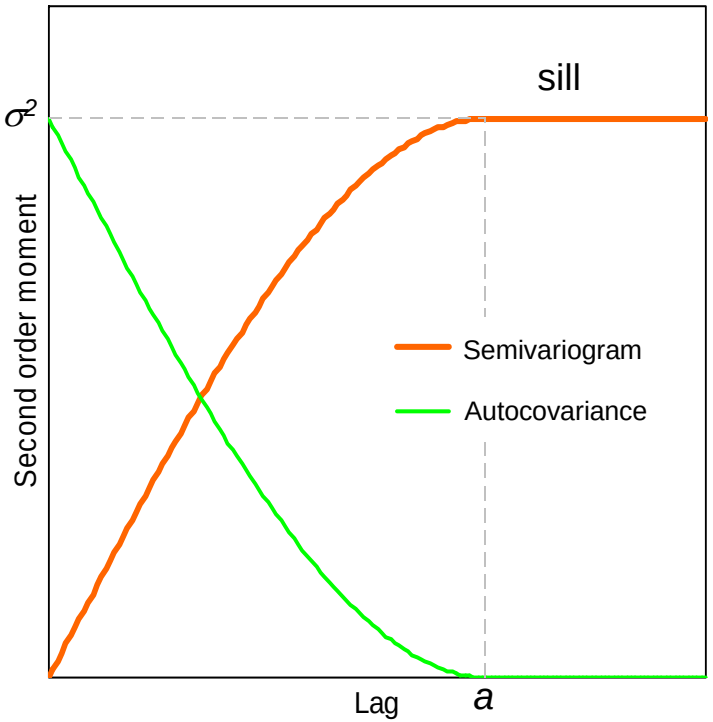
$$\hat{\gamma}(\mathbf{h}) = -\frac{1}{2N_h}\sum_{i=1}^{N_h}\left[Z(\mathbf{s}_i+\mathbf{h}) - Z(\mathbf{s}_i)\right]^2$$

LEARNING FROM THE SEMIVARIOGRAM

- The autocovariance is a function that results from grouping covariance values by distance class.
- The semivariogram and the autocovariance are related through:

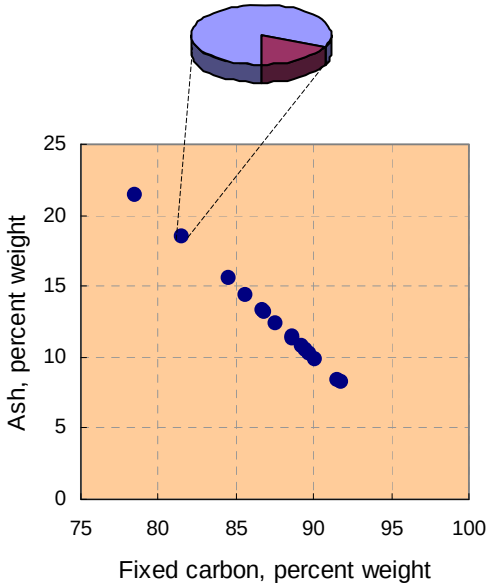
$$\gamma(\mathbf{h}) = \sigma^2 - \mathrm{Cov}(\mathbf{h}).$$

- The sill is equal to the population variance, σ^2.
- The range, *a*, gives a quantitative indication to the notion of radius of influence of a measurement.



COMPOSITIONS

In some studies, an object can be decomposed in constituent parts. By definition, these parts add to the whole object. Compositional data are measurements of these components.



Compositional data only contain relative information.

Compositional data are common in geochemical analysis, taking units such as parts per million and percent.

Only positive measurements are possible, between zero and the value denoting the whole object, say, 100 if the parts are percentages of the whole.

The fact that compositions are constraint data brings correlation to the components.

CAUTION

In general, straight application of statistics to data adding to a constant produces suboptimal or inconsistent results. Most successful ad hoc developments still employ classical statistics, but after taking special precautions.

The most common approach is to transform the data. One possibility is the isometric log-ratio transformation, Y, which for two components, X_1 and X_2, is:

$$Y = \frac{1}{\sqrt{2}} \log \frac{X_1}{X_2} \quad \text{or} \quad Y = \frac{1}{\sqrt{2}} \log \frac{X_2}{X_1}, \quad \text{with } X_1 + X_2 = c$$

where c is a known closure constant. This transformation:

- brings the range of variation to $(-\infty, \infty)$;
- eliminates the correlation due to closure;
- properly handles distances in attribute space.

Results require backtransformation.

5. NONDETECT STATISTICS

STATISTICS OF THE UNDETECTED

- Undetected values have peculiar characteristics that put them apart from those actually measured.
- Values below detection limit require different treatment than precise measurements.
- There have been several attempts to analyze samples with values below detection limit.

STATISTICAL METHODS

The main approaches for the calculation of summary statistics are:

- Replacement of the values below detection by an arbitrary number.
- Maximum likelihood method
- Kaplan-Meier method
- Bootstrap

REPLACEMENT

- This is the prevailing practice.
- This is not really a method. It has poor to no justification.
- It can lead to completely wrong results.
- The most common practice is to replace the values below detection by half that value.

MAXIMUM LIKELIHOOD METHOD
THE GENERAL CASE

This is a method for the inference of parameters, $\boldsymbol{\theta}$, of a parent probability density function $f(x \mid \boldsymbol{\theta})$ based on multiple outcomes, x_i.

The function $L(\boldsymbol{\theta}) = f(x_1,\ldots,x_n \mid \theta_1,\ldots,\theta_m)$ is called the likelihood function.

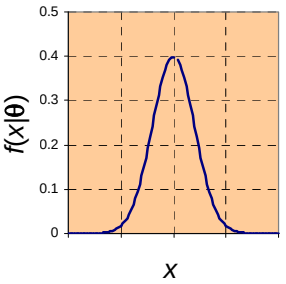
Given outcomes x_i, the optimal parameters $\boldsymbol{\theta}$ are those that maximize $L(\boldsymbol{\theta})$, which come from solving $\partial L(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = 0$.

For example, for the normal distribution,

$$f(x \mid \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\theta_2} e^{-\frac{(x-\theta_1)^2}{2\theta_2^2}} \qquad L(\boldsymbol{\theta}) = \left(\frac{1}{2\pi\theta_2^2}\right)^{\frac{n}{2}} e^{-\frac{\sum_{i=1}^{n}(x_i-\theta_1)^2}{2\theta_2^2}}$$

the solution is:

$$\theta_1 = \frac{1}{n}\sum_{i=1}^{n} x_i = \hat{\mu} \qquad \theta_2^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2 = \hat{\sigma}^2$$

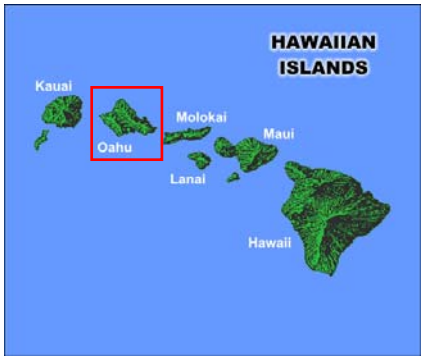


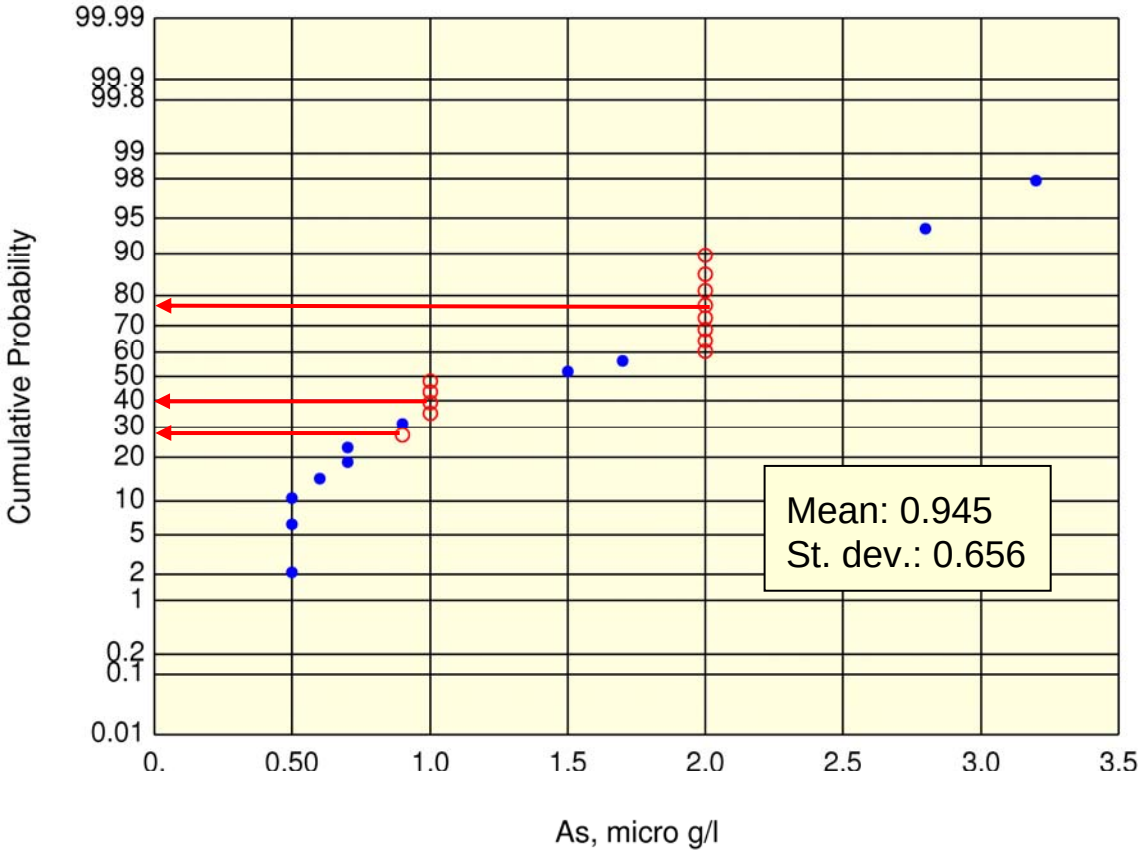
USGS

MAXIMUM LIKELIHOOD OF THE UNDETECTED

- The method is parametric: it requires assuming the distribution that the data would have followed in case all values would have been accurately measured.
- Essential for the success of the method is the assumption of the correct distribution.
- Most commonly selected models are the normal and the lognormal distributions.
- It is commonly used to estimate mean, variance, and quartiles.
- The method tends to outperform the others for samples of size larger than 50, or when 50 to 80% of the values are below detection limit.

ARSENIC CONCENTRATION, MANOA STREAM, OAHU, HAWAII



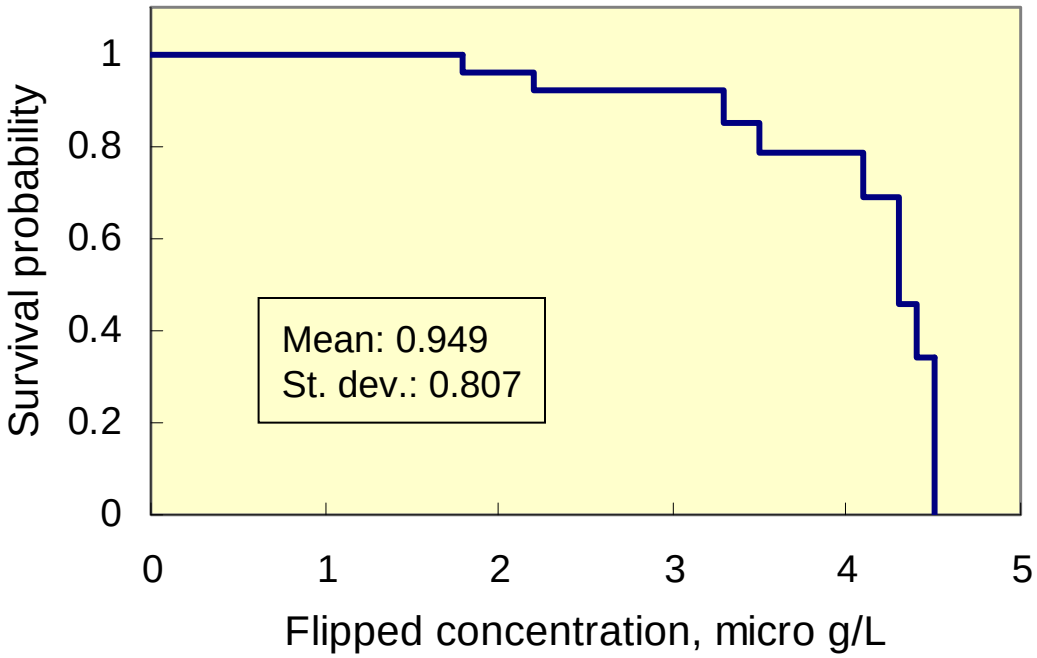


KAPLAN-MEIER METHOD

- The Kaplan-Meier method works in terms of a survival function, a name derived from clinical applications in which the interest is on survival time after starting medical treatment.
- The Kaplan-Meier method was developed for right censored data—exact measurements are missing above certain values. Left censored data must be flipped using a convenient but arbitrary threshold larger than the maximum.
- Calculated parameters must be backtransformed.
- The Kaplan-Meier method tends to work better when there are more exact measurements than data below detection limit.

USGS

OAHU DATA SURVIVAL FUNCTION

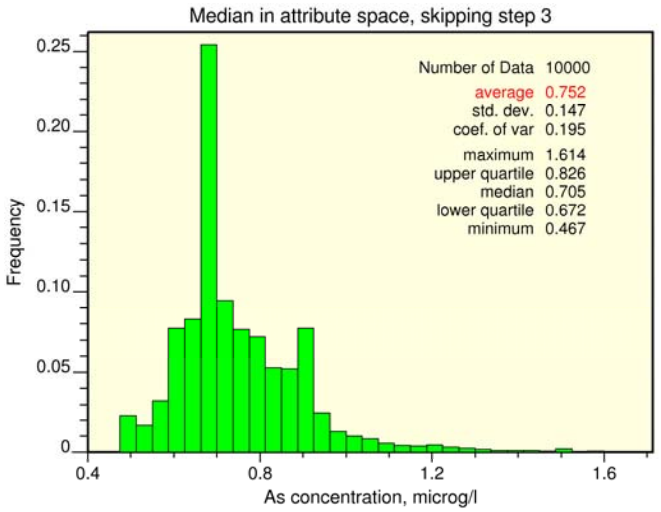


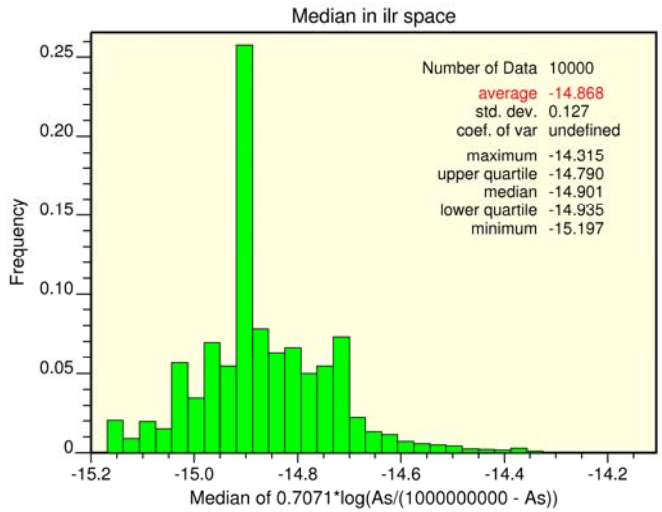
Flipping constant: 5 micro g/L

BOOTSTRAPPING FOR NONDETECTS

Remember that the bootstrap is a resampling form of the Monte Carlo method. The adjusted steps for handling nondetects are:

1. Randomly pick as many values as the sample size.
2. If a picked value is a nondetect, replace it with a number randomly selected between zero and the detection limit.
3. Make a log-ratio transformation.
4. Calculate all statistics of interest.
5. Save the statistics.
6. Go back to step 1 and repeat the process at least 1,000 times.





COMPARISON OF THE FOUR METHODS EXAMPLE FROM OAHU, HAWAII

	Mean	St. dev.	Q_1	Median	Q_3
	µg/L				
Substituting with zero	0.567	0.895	0.0	0.0	0.7
Subs. half detection lim.	1.002	0.699	0.5	0.95	1.0
Subs. with detection lim.	1.438	0.761	0.75	1.25	2.0
Lognormal Max. Likelihd.	0.945	0.656	0.509	0.777	1.185
Kaplan-Meier, nonparm.	0.949	0.807	0.5	0.7	0.9
Bootstrap average	1.002	0.741	0.501	0.752	1.370
Log-ratio bootstrap aver.	0.938	0.729	0.492	0.738	1.330

USGS

6. STATISTICAL TESTING

STATISTICAL TESTING

The classical way to make statistical comparisons is to prepare a statement about a fact for which it is possible to calculate its probability of occurrence. This statement is the null hypothesis and its counterpart is the alternative hypothesis.

The null hypothesis is traditionally written as H_0 and the alternative hypothesis as H_1.

A statistical test measures the experimental strength of evidence against the null hypothesis.

Curiously, depending on the risks at stake, the null hypothesis is often the reverse of what the experimenter actually believes for tactical reasons that we will examine.

USGS

EXAMPLES OF HYPOTHESES

Let μ_1 and μ_2 be the means of two samples. If one wants to investigate the likelihood that their means are the same, then the null hypothesis is:

$$H_0 : \mu_1 = \mu_2$$

and the alternative hypothesis is:

$$H_1 : \mu_1 \neq \mu_2$$

but it could also be:

$$H_1 : \mu_1 > \mu_2$$

The first example of H_1 is said to be two-sided or two-tailed because includes both $\mu_1 > \mu_2$ and $\mu_1 < \mu_2$. The second is said to be one-sided or one-tailed.

The number of sides has implications on how to formulate the test.

POSSIBLE OUTCOMES

	H_0 is correct	H_0 is incorrect
H_0 is accepted	Correct decision Probability: $1-\alpha$	Type II error Probability: β
H_0 is rejected	Type I error Probabilty: α	Correct decision Probability: $1-\beta$

- The probability α of committing a Type I error is called the level of significance.
- α is set before performing the test.
- In a two-sided test, α is split between the two options.
- Often, H_0 and α are designed with the intention of rejecting H_0, thus risking a Type I error and avoiding the unbound Type II error. The more likely this is, the more power the test has. Power is $1-\beta$.

USGS

IMPORTANCE OF CHOOSING H_0

Selection of what is null and what is alternative hypothesis has consequences in the decision making. Customarily, tests operate on the left column of the contingency table and the harder to analyze right column remains unchecked. Consider environmental remedial action:

		H_o: Site is clean	
		True	False
Test action	Accept	Correct	
	Reject	Wrong	

A. Wrong rejection means the site is declared contaminated when it is actually clean, which should lead to unnecessary cleaning.

		H_o: Site is contaminated	
		True	False
Test action	Accept	Correct	
	Reject	Wrong	

B. Now, the wrong decision declares a contaminated site clean. No action prolongs a health hazard.

In both cases, $\Pr[\text{Type I error}] \leq \alpha$.

PARTITION

The level of significance is employed to partition the range of possible values of the statistic into two classes.

- One interval, usually the longest one, contains those values that, although not necessarily satisfying the null hypothesis exactly, are quite possibly the result of random variation. If the statistic falls in this interval—the green interval in our cartoon—the null hypothesis is accepted.

- The red interval comprises those values that, although possible, are highly unlikely to occur. In this situation, the null hypothesis is rejected. The departure from the null hypothesis most likely is real, significant.
- When the test is two-sided, there are two rejection zones.

STATISTIC

A key step in the feasibility of being able to run a test is the ability of finding an analytical expression for a statistic such that:

- It is sensitive to all parameters involved in the null hypothesis.
- It has an associated probability distribution.

Given some data, the *p*-value is the probability that the statistic takes values beyond the value calculated using the data while H_0 is still true. Hence:

- If the *p*-value is larger than the level of significance, H_0 is accepted.
- The lower the *p*-value, the stronger is the evidence provided by the data against the null hypothesis.

The *p*-value allows to convert the statistic to probability units.

USGS

SAMPLING DISTRIBUTION

The sampling distribution of a statistic is the distribution of values taken by the statistic for all possible random samples of the same size from the same population. Examples of such sampling distributions are:

- Standard normal and the t-distributions for the comparison of two means.
- The F-distribution for the comparison of two variances.
- The χ^2 distribution for the comparison of two distributions.

USGS

PROBABILITY REMINDER AND P-VALUES

$$\text{Prob}[X \le x_1] = \int_{-\infty}^{x_1} f(x)dx$$

$$\text{Prob}[X > x_1] = \int_{x_1}^{\infty} f(x)dx$$

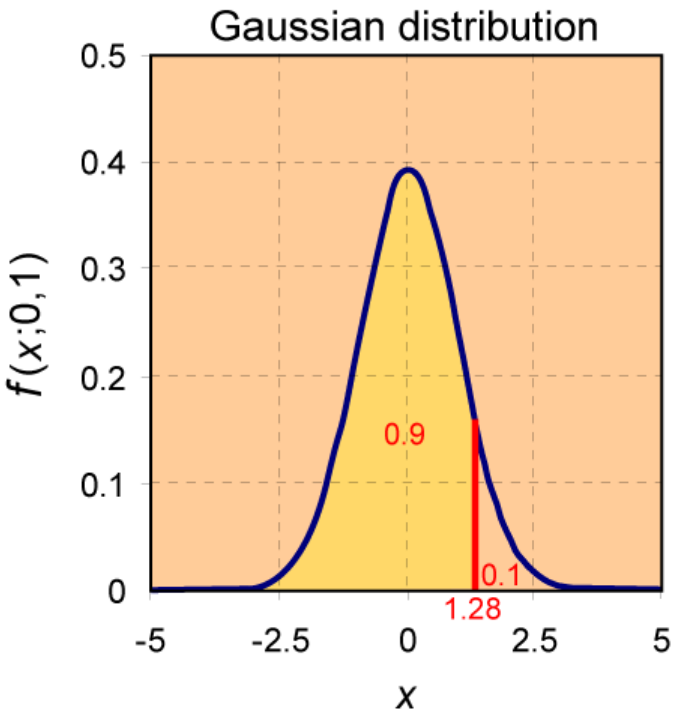
Example:

$$\text{Prob}[X \le 1.28] = \int_{-\infty}^{1.28} \text{Normal}(x;0,1)dx$$

$$= 0.90$$

$$\text{Prob}[X > 1.28] = \int_{1.28}^{\infty} \text{Normal}(x;0,1)dx$$

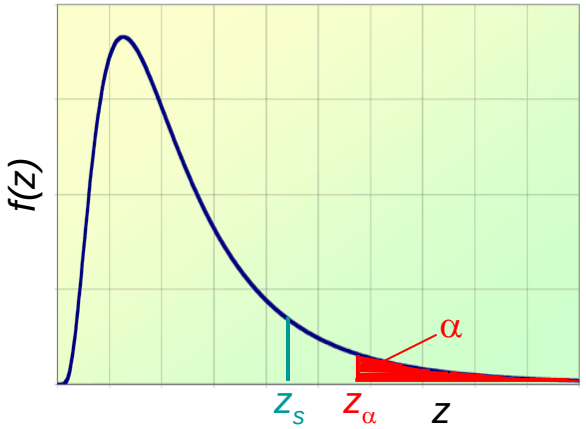
$$= 0.10$$



In this example, assuming a one-sided test, when the statistic is 1.28, the *p*-value is 0.1 (10%).

TESTING PROCEDURE

1. Select the null hypothesis H_0 and the alternative hypothesis H_1.
2. Choose the appropriate statistic.
3. Set the level of significance α.
4. Evaluate the statistic for the case of interest, z_s.



5. Use the distribution for the statistic in combination with the level of significance to define the acceptance and rejection intervals. Find out either the corresponding:
   - *p*-value of the statistic in the probability space, or
   - level of significance in the statistic space, z_α.
6. Accept the null hypothesis if $z_s < z_\alpha$. Otherwise, reject H_0 because its chances to be true are less than α.

PARAMETER TESTING

TESTING TWO MEANS

INDEPENDENCE

The random outcomes *A* and *B* are statistically independent if knowing the result of one does not change the probability for the unknown outcome of the other.

$$\text{Prob}[A \mid B] = \text{Prob}[A]$$

For example:

- Head and tail in successive flips of a fair coin are independent events.
- Being dealt a king from a deck of cards and having two kings on the table are dependent events.

USGS

TESTING PARENT POPULATIONS

Given a sample, the interest is on whether the values are a partial realization of a population with known mean and variance.

A one-sided test would be:

$$H_0 : m_n \leq \mu_p$$

$$H_1 : m_n > \mu_p$$

where

m_n is the sample mean,

μ_p is the population mean,

σ_p^2 is the population variance.

CENTRAL LIMIT THEOREM

Let $X_1, X_2, \ldots, X_n$ be *n* random variables that:

- share the same probability distribution *D*;
- are independent;
- the mean of the probability distribution is μ;
- the standard deviation of the distribution σ exists and is finite; and
- let $m_n = \left(X_1 + X_2 + \ldots + X_n\right)/n$.

Then, for any *D*:

$$\lim_{n \to \infty} F\left(m_n\right) = \text{Normal}\left(\mu, \sigma^2/n\right)$$

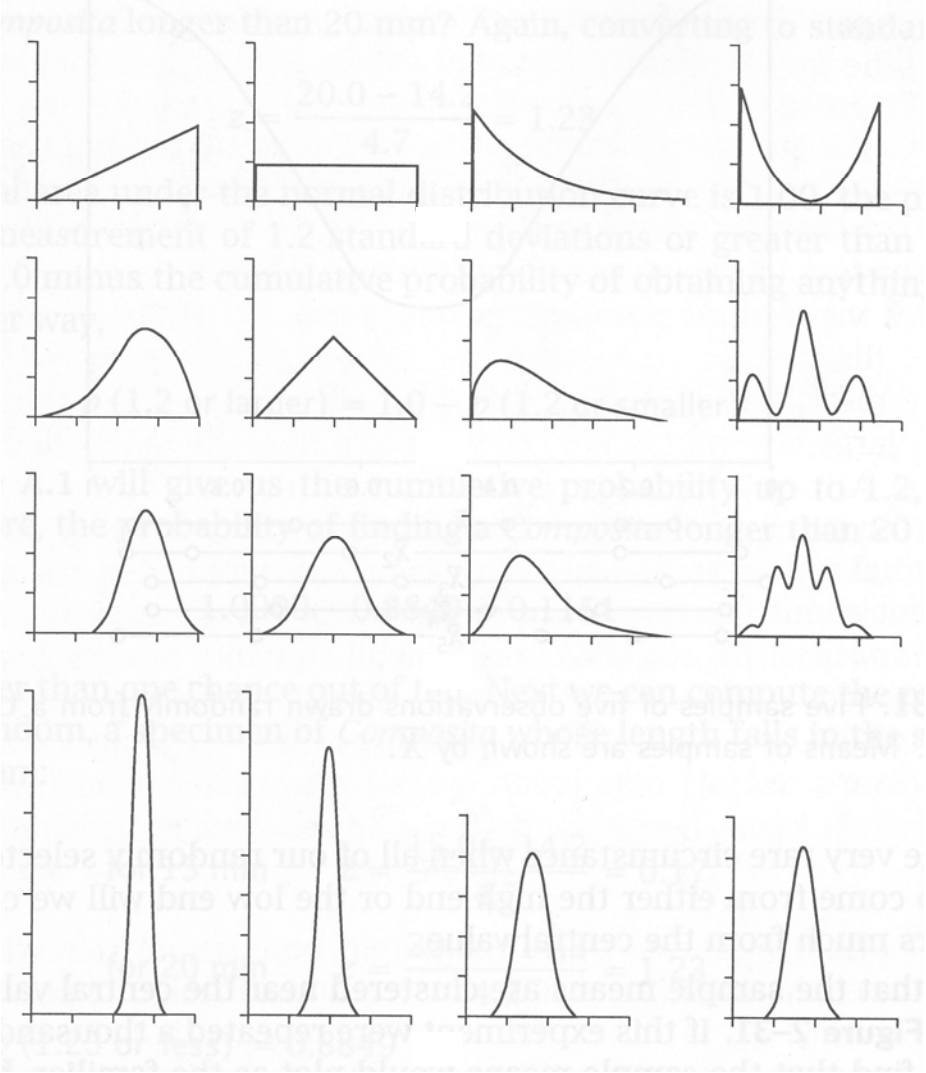
CENTRAL LIMIT THEOREM IN ACTION

Examples

$n = 2$

$n = 4$

$n = 25$

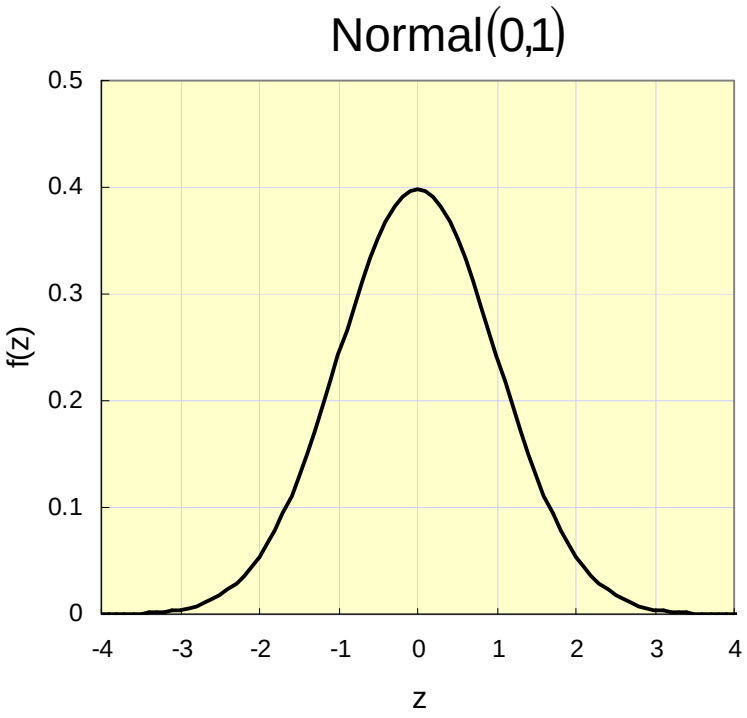


STANDARD NORMAL DISTRIBUTION

As mentioned before, standardization is the transformation obtained by subtracting the mean from each observation and dividing the result by the standard deviation.

In the Central Limit case:

$$z = \frac{m_n - \mu_p}{\sigma_p \sqrt{1/n}}$$



Thus, in this case, the appropriate statistic is the sample mean standardized by employing the population parameters. By the Central Limit theorem, this statistic follows a standard normal distribution, *N*(0,1).

EXAMPLE

	Size	Mean	Stand. dev.
Population		16.5	6.0
Sample	25	18.0	

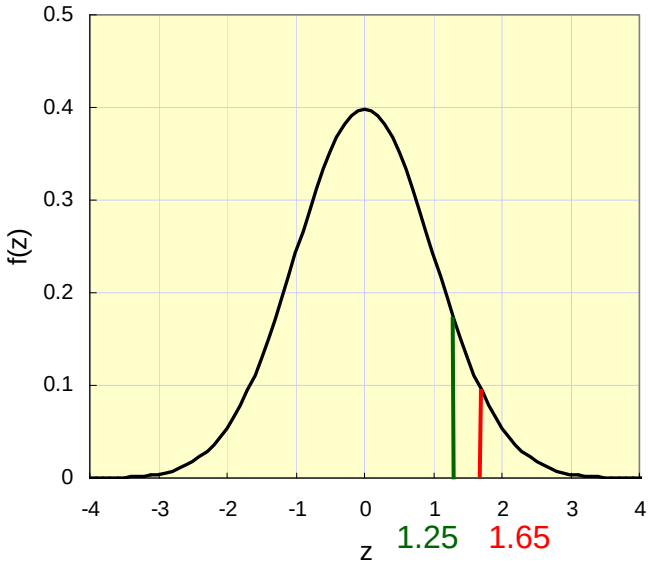
1. $H_0 : m_n \leq \mu_p$

   $H_1 : m_n > \mu_p$

2. $\alpha = 0.05\ (5\%)$

3. The statistic is z-score.

4. $z_s = \dfrac{18 - 16.5}{6\sqrt{1/25}} = 1.25$



5. For a Normal(0,1) distribution, the cumulative probability is 0.05 above $z_\alpha = 1.65$.

6. $z_s < z_\alpha$, therefore, H_0 is accepted.

INFLUENCE OF SAMPLE SIZE

The table shows sensitivity of the results to sample size when the experimental mean remains fixed at 18.0 in the previous example.

Sample size	Statistic z_s	*P*-value, percent	$H_o(\alpha=5\%)$
10	0.79	21.48	Accepted
25	1.25	10.16	Accepted
100	2.50	0.62	Rejected
250	3.95	0.01	Rejected

The larger the sample size, the more likely a rejection.

- Specification of sample size adds context to a test.
- Specificity of a test is poor for small sample sizes.
- For large sample sizes, findings can be statistically significant without being important.

USGS

DEGREES OF FREEDOM

In the calculation of a statistic, the number of degrees of freedom is the number of values free to vary.

The number of degrees of freedom for any estimate is the number of observations minus all relationships previously established among the data. The number of degrees of freedom is at least equal to the number of other parameters necessary to compute for reaching to the calculation of the parameter of interest.

For example, for a sample of size *n*, the number of degrees of freedom for estimating the variance is *n* – 1 because of the need first to estimate the mean, after which one observation can be estimated from the others. For example:

$$x_1 = n \cdot \underbrace{\frac{x_1 + x_2 + \ldots + x_n}{n}}_{\text{mean}} - \left(x_2 + x_3 + \ldots + x_n\right)$$

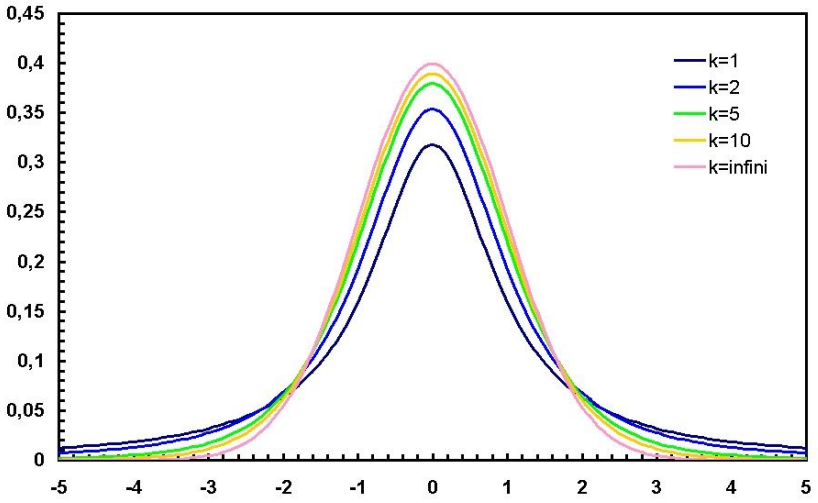
STUDENT'S T-DISTRIBUTION

We have seen that the mean of any independent and identically distributed random variables is normal provided that:

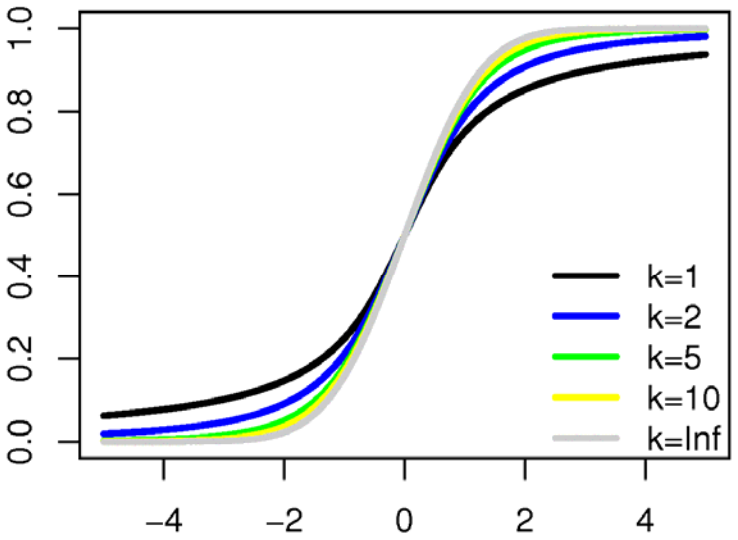
- One knows the means and variance of the population.
- The sample size is large. The rule of thumb is that a size above 30 is large.

The Student's t distribution operates analogously to the standard normal distribution, $N(0,1)$, and should be used instead when any of the requirements above is not met.

EXAMPLES OF T-DISTRIBUTIONS



Probability density function



Cumulative distribution function

k is the degrees of freedom.

TESTING TWO VARIANCES

F-DISTRIBUTION

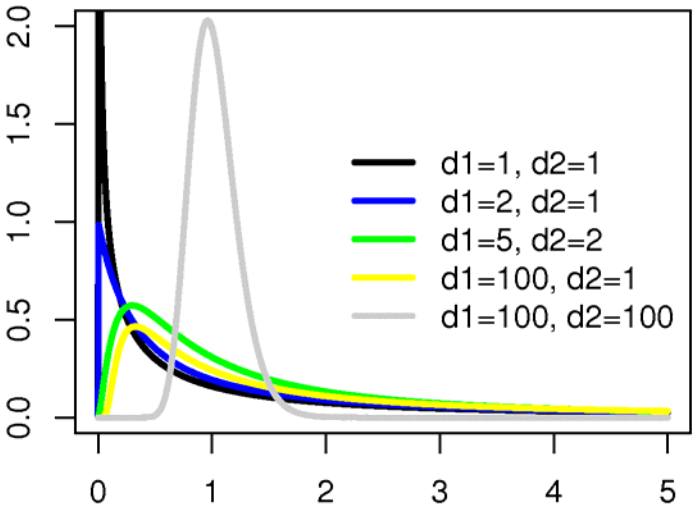
The F-distribution is another distribution particularly developed for statistical testing.

Its parameters are two degrees of freedom that vary independently from each other.

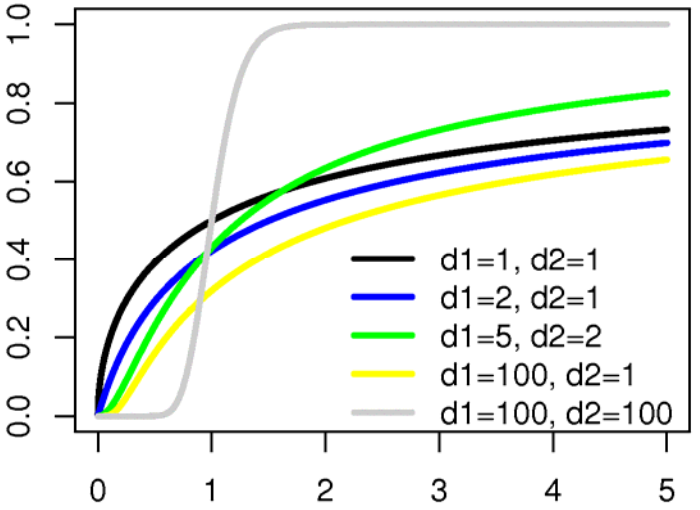
It is employed to test variances and it operates completely analogously to the other distributions: it provides a reference value based on which the null hypothesis is accepted or rejected according to a level of significance.

The main limitation is that the samples must come from normal distributions.

EXAMPLES OF F-DISTRIBUTIONS



Probability density function



Cumulative distribution function

d_1 *and* d_2 are the degrees of freedom.

EXAMPLE OF VARIANCE TESTING

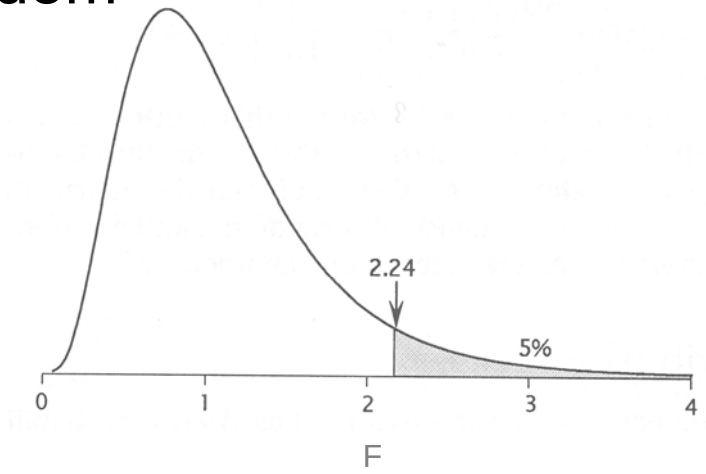
	Size	Variance
Sample 1	11	22.4
Sample 2	26	10.0

$$H_0 : \sigma_1 = \sigma_2$$
$$H_1 : \sigma_1 > \sigma_2$$

In this case, the degrees of freedom are the sample sizes minus 1.

The test statistic is the ratio of largest to smallest variance

$$F = \frac{22.4}{10}$$
$$= 2.24$$



Above $F_{10,25} = 2.24$, the *p*-value is 5%. Thus H_0 is rejected if $\alpha > 5.0\%$, when z_α will be less than 2.24. Otherwise, it is accepted.

Note that using the *p*-value allows expressing the result independently of α, z_α, and z_s.

TESTING MORE THAN TWO MEANS

ANALYSIS OF VARIANCE

- Analysis of variance is an application of the F-test for the purpose of testing the equality of several means.
- Strategies and types of variances abound, but it all revolves around calculating some variances and then proving or disproving their equality.
- The interest is on deciding whether the discrepancies in measurements are just noise or part of a significant variation.
- The simplest form is the one-way ANOVA, in which the same attribute is measured repeatedly for more than two specimens:

$$H_0 : \mu_1 = \mu_2 \cdots = \mu_k, k \geq 3$$

$$H_1 : \text{at least one mean is different.}$$

- In this case, the variances going into the statistic are the variance for the mean among localities and the variance within specimens.

ONE-WAY ANOVA TABLE

Source of variation	Sum of squares	Degrees of freedom	Variance	F test statistic
Among localities	SS_A	m - 1	MS_A	MS_A / MS_E
Within specimens ("Error")	SS_E	$N - m$	MS_E	
Total Variation	SS_T	N - 1		

m : number of localities

n : number of specimens

x_{ij}: specimen i for locality j

$$N = n \cdot m$$

$$\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

$$\overline{\overline{X}} = \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{n} x_{ij}$$

$$SS_A = \sum_{j=1}^{m}\left(\overline{X}_j - \overline{\overline{X}}\right)^2$$

$$SS_E = \sum_{j=1}^{m}\left(\sum_{i=1}^{n} x_{ij} - \overline{X}_j\right)^2$$

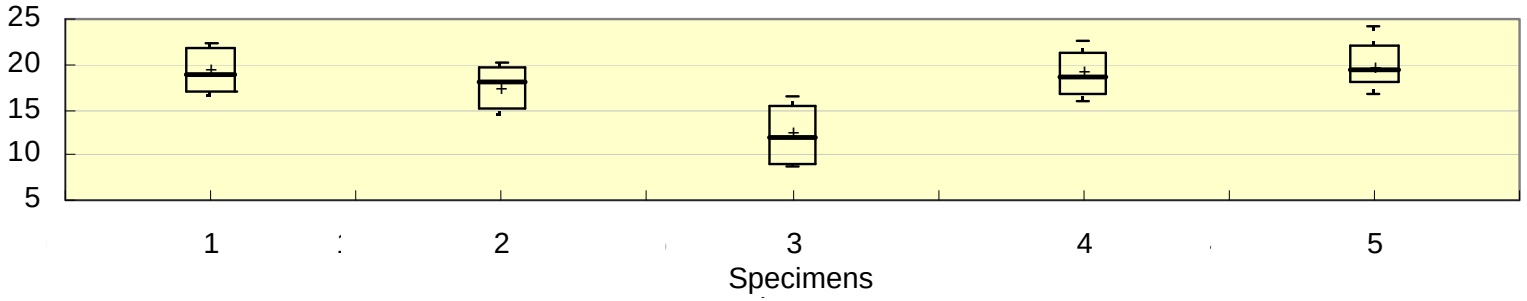
$$SS_T = \sum_{j=1}^{m}\sum_{i=1}^{n}\left(x_{ij} - \overline{\overline{X}}\right)^2$$

$$MS_A = \frac{SS_A}{m-1}$$

$$MS_E = \frac{SS_E}{N-m}$$

EXAMPLE

Shell width, mm					
Specimen	Locality 1	Locality 2	Locality 3	Locality 4	Locality 5
1	16.5	14.3	8.7	15.9	17.6
2	17.3	16.1	9.5	17.6	18.4
3	18.7	17.6	11.4	18.4	19.1
4	19.2	18.7	12.5	19.0	19.9
5	21.3	19.3	14.3	20.3	20.2
6	22.4	20.2	16.5	22.5	24.3



Source of variation	Sum of squares	Degrees of freedom	Variances	F test statistic
Among localities	237.42	4	59.35	10.14
Within specimens ("Error")	146.37	25	5.85	
Total Variation	383.79	29		

When $F_{4,25} = 10.14$, $p = 0.00005$, so for $\alpha > 0.00005$, at least one mean is significantly different from the others.

DISTRIBUTIONAL TESTING

THE χ^2 DISTRIBUTION

Let $X_1, X_2, \ldots, X_k$ be *k* random variables that:

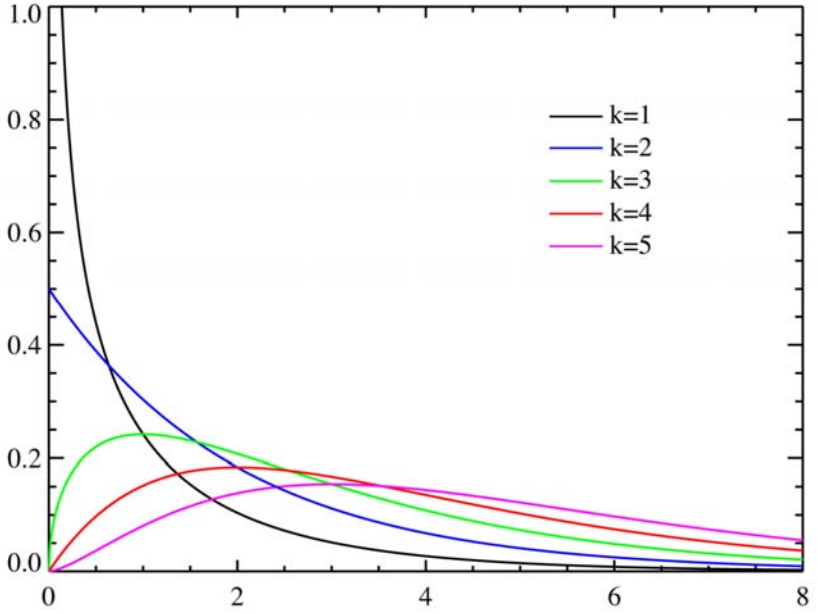
- are independent;
- are all normally distributed;
- the means μ_i are not necessarily equal;
- each distribution has a standard deviation σ_i,

then, the sum of the standardized squares

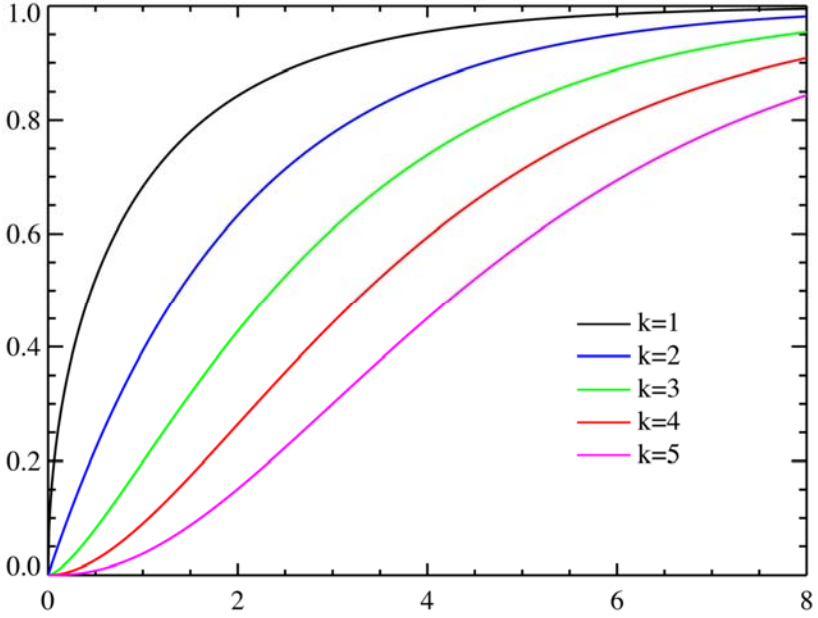
$$\sum_{i=1}^{k} z_i^2 = \sum_{i=1}^{k} \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2$$

follows a χ_k^2 distribution with *k* degrees of freedom. Like the t-distribution, $\lim_{k \to \infty} \chi_k^2 = N(0,1)$.

EXAMPLES OF χ^2 DISTRIBUTIONS



Probability density function



Cumulative distribution function

k is the degrees of freedom.

χ^2 TEST OF FITNESS

χ_k^2 can be used to test whether the relative frequencies of an observed event follow a specified frequency distribution. The test statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

O_i is an observed frequency for a given class,

E_i is the frequency corresponding to the target distribution.

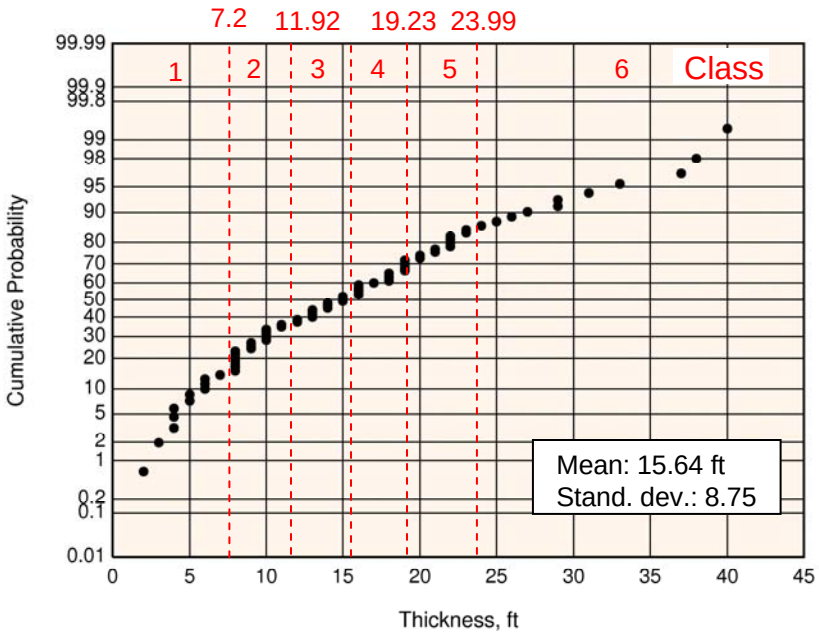
As usual, if the statistic lands beyond the value of χ_k^2 corresponding to the degrees of freedom *k* and level of significance α, the null hypothesis is rejected.

Weaknesses of the test are:

- result depends on the number of classes,
- no class can be empty,
- there must be at least 5 observations per class for at least 80% of the classes.

WEST LYONS FIELD, KANSAS (1)

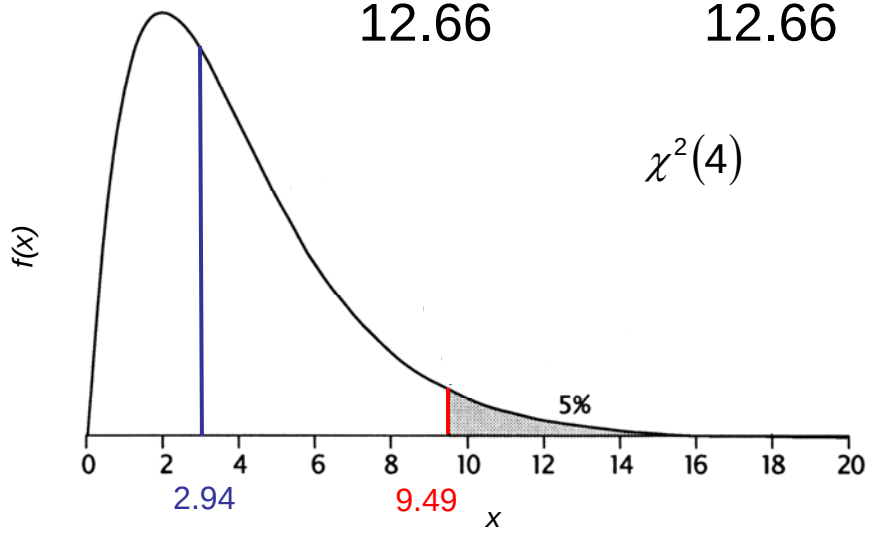
1. The testing will be for normality, 6 classes, and a level of significance of 5%.
2. In the normal distribution, class width should be 100/6=16.66% probability. Boundaries for those equiprobable classes are 7.20, 11.92, 15.64, 19.23, and 23.99. Each class contains E_i = 76/6=12.66 measurements exactly. Values for O_i must be counted.



WEST LYONS FIELD, KANSAS (2)

4. In this case, the statistic is:

$$\chi_s^2 = \frac{(11-12.66)^2}{12.66} + \frac{(17-12.66)^2}{12.66} + \frac{(12-12.66)^2}{12.66} + \frac{(10-12.66)^2}{12.66} + \frac{(15-12.66)^2}{12.66} + \frac{(11-12.66)^2}{12.66} = 2.94$$



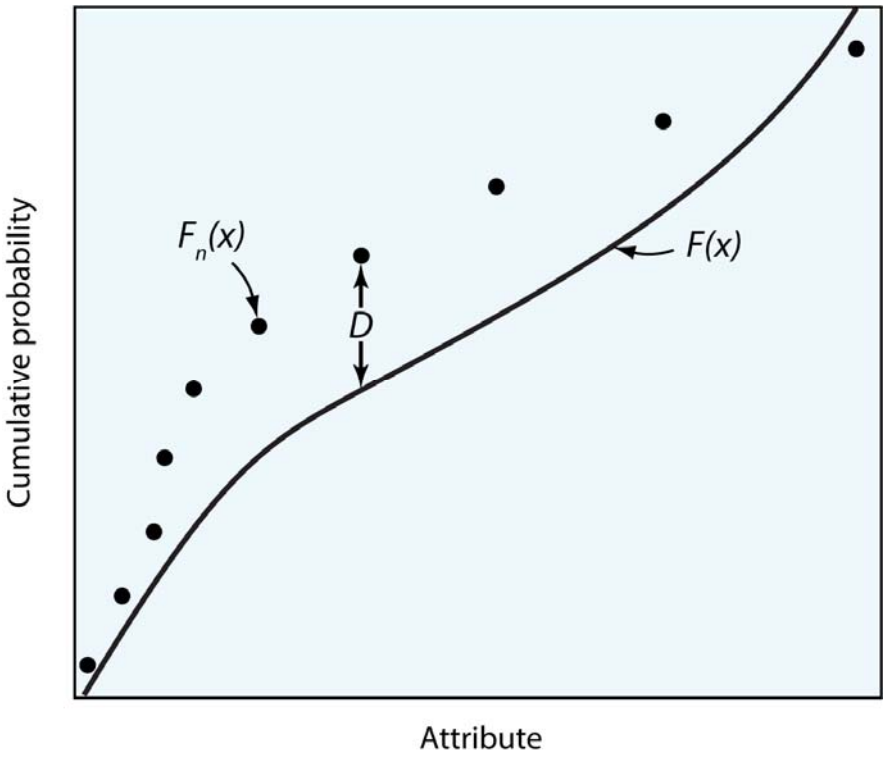
5. We have calculated two parameters (mean and standard deviation). So the there are 6 - 2 = 4 degrees of freedom.

6. For a level of significance of 5%, $\chi_\alpha^2(4, 0.05) = 9.49$. So, because $\chi_s^2 < \chi_\alpha^2$, there is no evidence to suggest that the thickness values are not normally distributed.

KOLMOGOROV-SMIRNOV TEST

This is a much simpler test than the chi-square test because:

- The result does not depend on the number of classes.
- It is nonparametric.

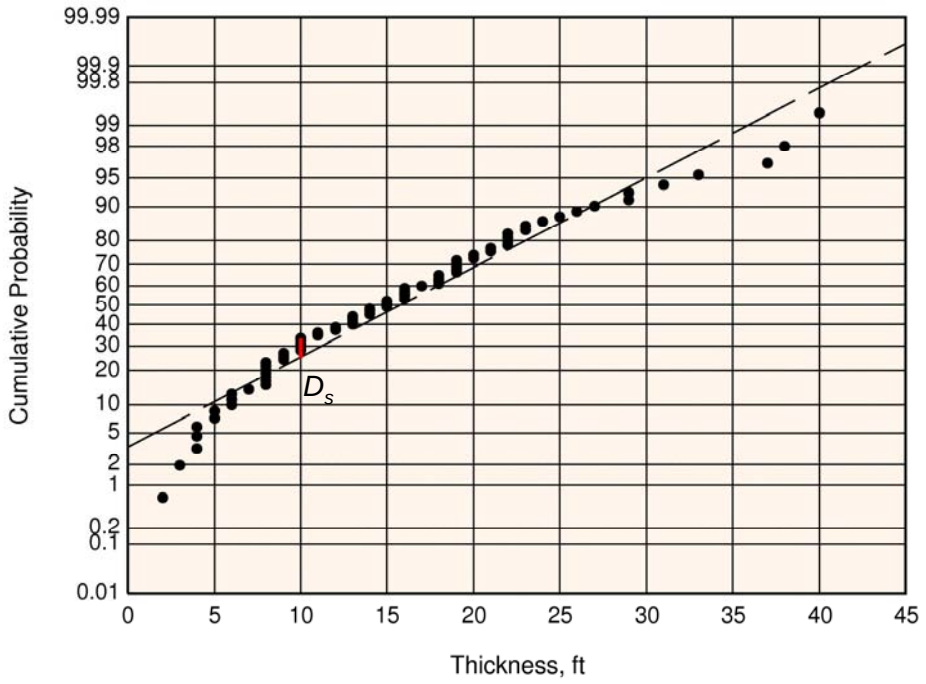


The statistic is the maximum discrepancy, D, between the two cumulative distributions $F_n(x)$ and $F(x)$ under comparison.

$$D = \max\left(F_n(x) - F(x)\right)$$

WEST LYONS FIELD, KANSAS

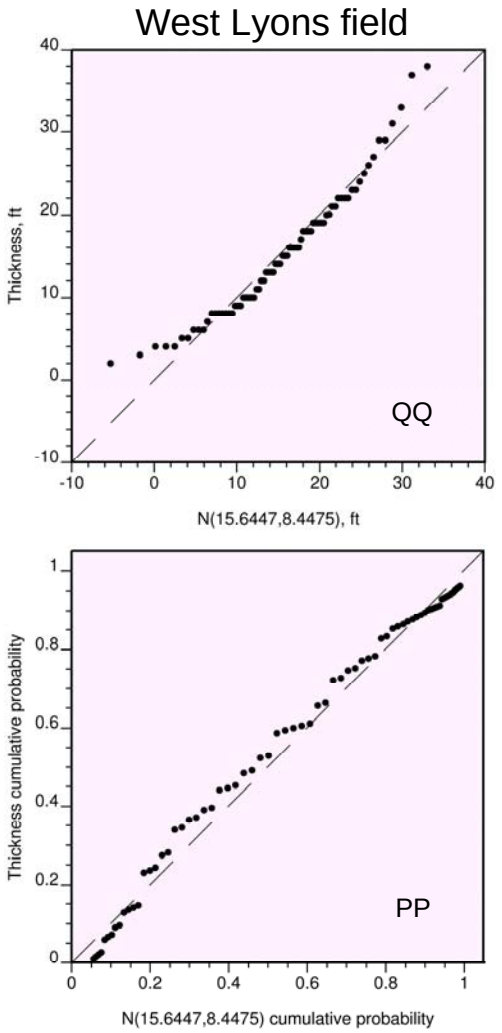
1. We want to test that the data are N(15.64,8.75) for a level of significance of 5%.
2. D_s = 0.090 (9%).
3. Critical value for $D_\alpha(76,0.05)$ = 0.102.
4. There is no evidence to postulate that the thickness values are not normally distributed because $D_s < D_\alpha$. The test, however, presumes the sample comes from independent and identically distributed observations.



Q-Q AND P-P PLOTS

Scatterplots of the quantiles and the cumulative probabilities of the two distributions are the ultimate tests in terms of simplicity.

- If the distributions are the same, the points align along the main diagonal.
- There is no statistic or level of significance for evaluating the results.
- P-P plots are insensitive to shifting and scaling, and the vertical scale is in the same units as in Kolmogorov-Smirnov test.
- The Q-Q plot is good here at calling the user's attention about the normal distribution being able to take negative values.



FINAL REMARKS

SPATIOTEMPORAL DEPENDENCE

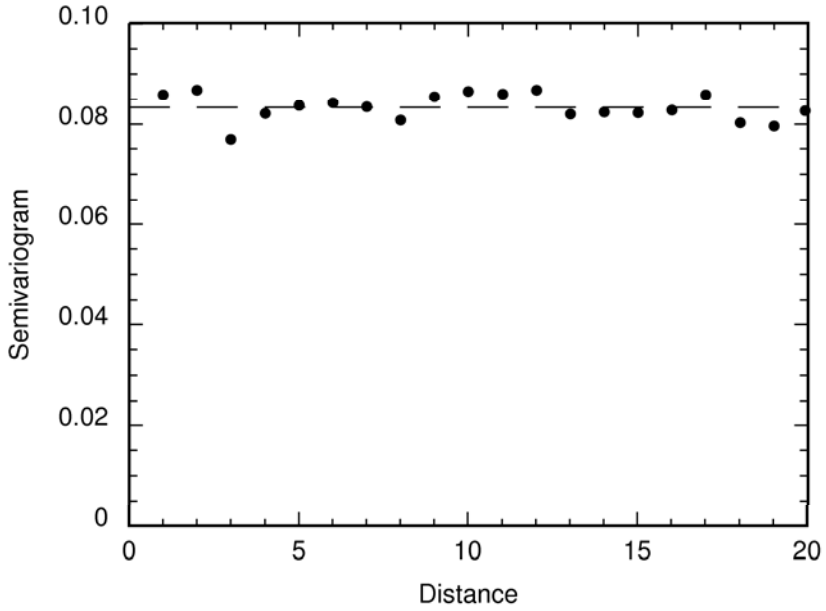
Many natural phenomena exhibit fluctuations that show continuity in space and time. Continuity denotes the common experience that in proximity of any observation, another measurement will be approximately the same.

- Given a spatiotemporally continuous phenomenon and the location of an observation, close proximity does tell something about the outcome of a second observation; they are not necessarily independent.
- The degree and extent of this dependence can be estimated through the semivariogram.
- Modeling spatiotemporally dependent data, often in the form of maps, is the realm of geostatistics.

USGS

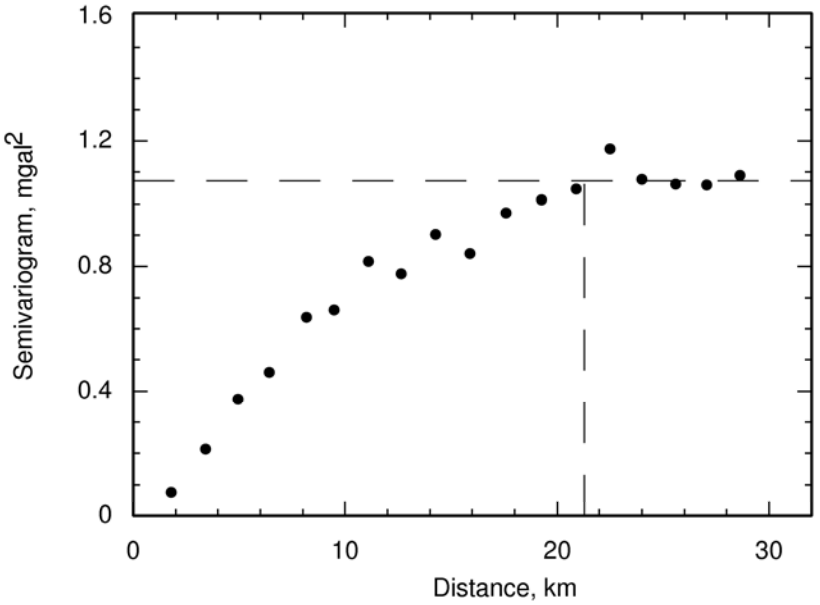
SEMIVARIOGRAM

1000 random numbers



Spatial independence

Gravimetric anomaly, Elk Co., KS



Spatial dependence

7. MULTIVARIATE STATISTICS

METHODS

Multivariate statistics deals with the analysis and display of objects with two or more attributes consistently measured for several specimens. The main motivations are better understanding of the data and interest in making simplifications.

The main families of methods available are:

- cluster analysis
- discriminant analysis
- principal component analysis
- factor analysis

While missing values are not an insurmountable problem, they are a situation to avoid. Often one missing value in a record requires dropping the entire record.

All methods are complex enough to require a computer for performing the calculations.

MATRICES

A matrix is a rectangular array of numbers, such as **A**. When $n = m$, **A** is a square matrix of order n.

Matrices are a convenient notation heavily used in dealing with large systems of linear equations, which notably reduce in size to just **A X** = **B**.

Transposing all rows and columns in **A** is denoted as $\mathbf{A}^T$.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} b_1 & b_2 & \cdots & b_m \end{bmatrix}^T$$

$$\mathbf{X} = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix}^T$$

Main diagonal of a square matrix is the sequence of element a_{11}, a_{22}, …, a_{nn} from upper left to lower right.

CLUSTER ANALYSIS

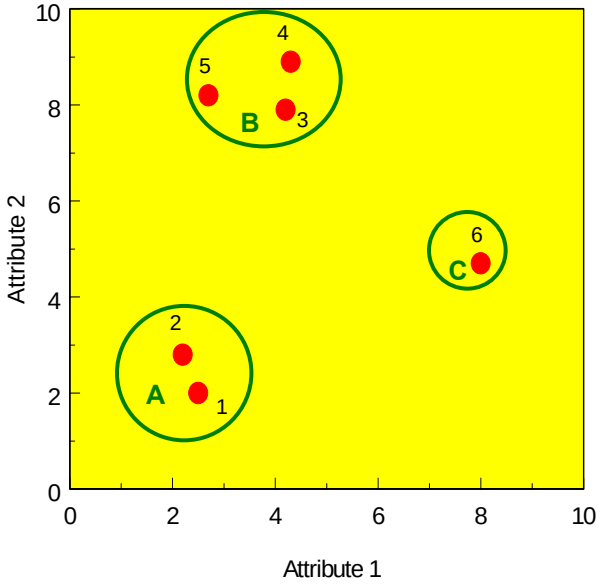
AIM

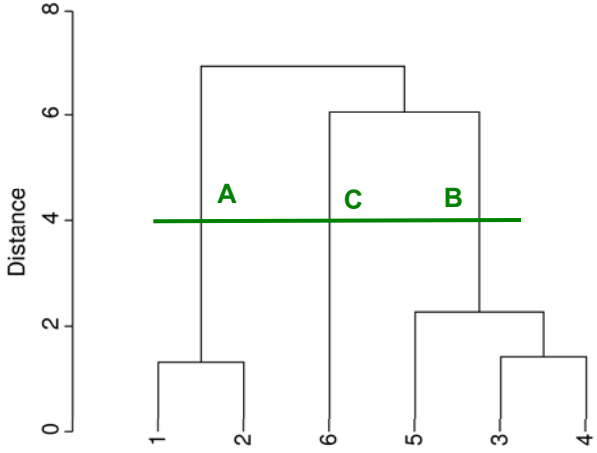
The general idea is to group objects in attribute space into clusters as internally homogeneous as possible and as different form the other clusters as possible.

Different types of distances, proximity criteria, and approaches for preparing the clusters have resulted in several methods.

Some methods render results as dendrograms, which allow displaying the data in two dimensions.

Large distance increments provide the best criteria for deciding on the natural number of clusters.





DISSIMILARITIES

If $\boldsymbol{\Sigma}$ is the covariance matrix for a multivariate sample of *m* attributes, the following distances are the most common measurements for dissimilarity between vectors ***p*** and ***q***:

- Euclidean, $\sqrt{\sum_{i=1}^{m}(p_i - q_i)^2}$
- Manhattan, $\sum_{i=1}^{m}|p_i - q_i|$
- Mahalanobis, $\sqrt{\sum_{i=1}^{m}(\mathbf{p}-\mathbf{q})'\boldsymbol{\Sigma}^{-1}(\mathbf{p}-\mathbf{q})}$.

Distances can be in original data space or standardized.

Mahalanobis distances account for distance relative to direction and global variability through the covariance matrix.

Euclidean distances can be regarded as a special case of Mahalanobis distance for a covariance matrix with ones along the main diagonal and zeros anywhere else.

USGS

PROXIMITY AND METHODS

Proximity between clusters is commonly decided based on the average inter-cluster distance. The most common methods are:

- Agglomerative hierarchical
- Divisive hierarchical
- K-means

Choice of dissimilarity measure may have greater influence in the results than the selection of the method.

DATA SET TO BE CLUSTERED

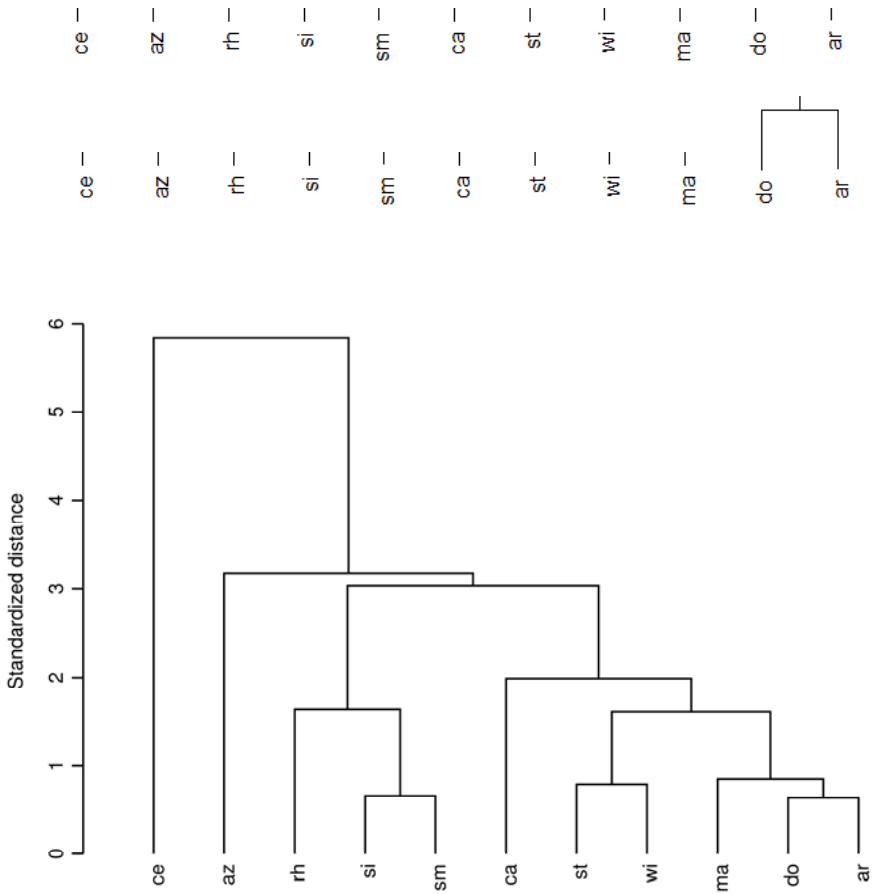
Physical properties of carbonates

Mineral	Spec. gravity g/cc	Refractive index Smallest	Largest	Hardness
Aragonite (ar)	2.94	1.530	1.685	3.7
Azurite (az)	3.77	1.730	1.838	3.7
Calcite (ca)	2.72	1.486	1.658	3.0
Cerusite (ce)	6.57	1.803	2.076	3.0
Dolomite (do)	2.86	1.500	1.679	3.7
Magnesite (mg)	2.98	1.508	1.700	4.0
Rhodochrosite (rh)	3.70	1.597	1.816	3.7
Smithsonite (sm)	4.43	1.625	1.850	4.2
Siderite (si)	3.96	1.635	1.875	4.3
Strontianite (st)	3.72	1.518	1.667	3.5
Witherite (wi)	4.30	1.529	1.677	3.3

Often all attributes are standardized to avoid the dominance of results by those with the largest numerical values.

AGGLOMERATIVE CLUSTERING

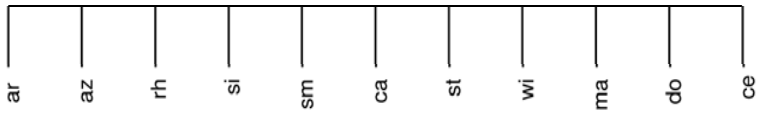
- Initially there are as many clusters as records.
- At any stage, the two closest clusters are merged together, reducing the number of clusters by one.
- The procedure ends with the last two clusters merging into a single cluster.



Euclidean distance is the most common choice (Ward's method).

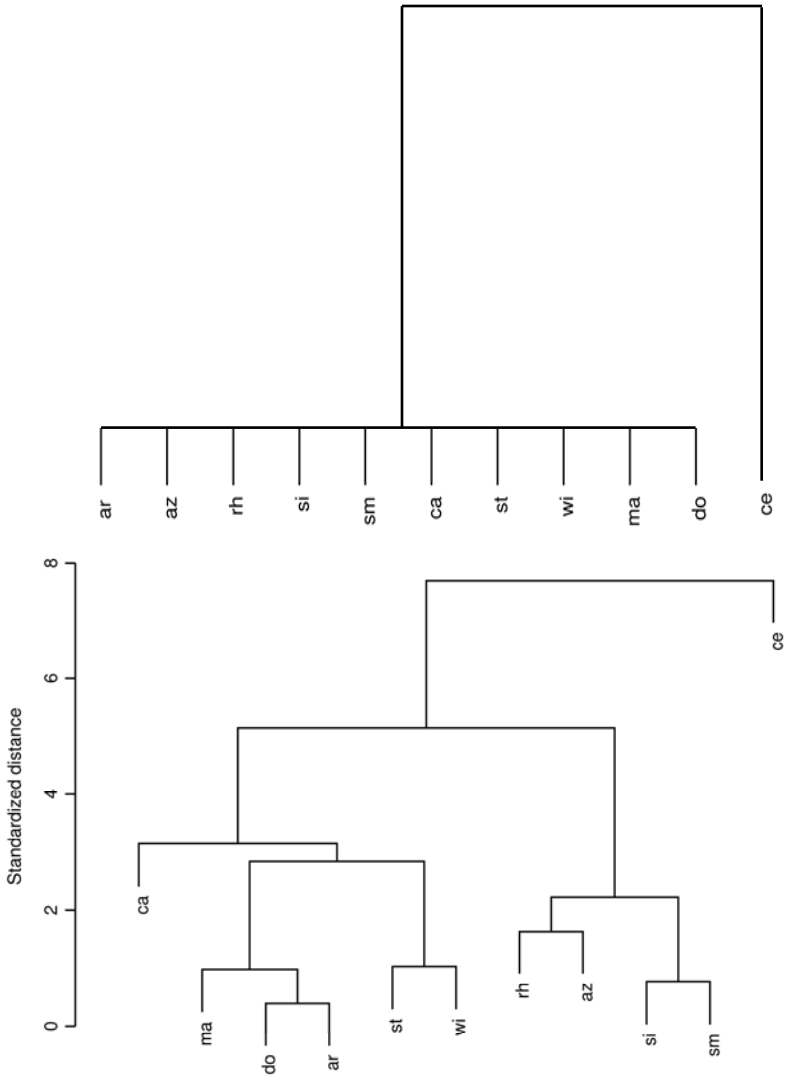
DIVISIVE CLUSTERING

1. Initially all records are in one cluster.
2. At any stage, all distances inside each cluster are calculated.
3. The cluster with the largest specimen-to-specimen distance is broken apart, increasing the number of clusters by one.



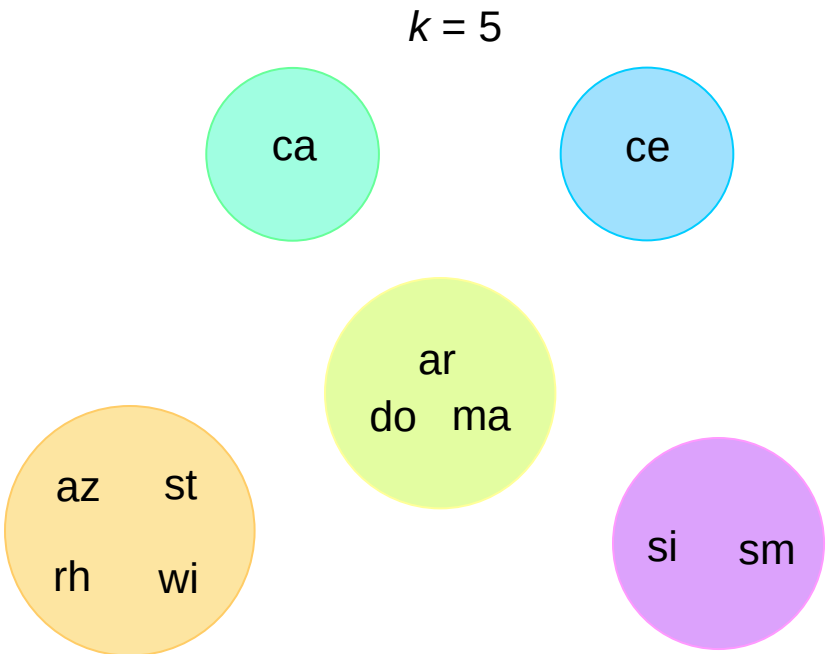
DIVISIVE CLUSTERING

4. These specimens with the largest distance become the seed of the two new clusters. All other specimens in the breaking apart cluster are assigned to the closest seed.
5. The procedure ends with the last two true clusters breaking apart into individual specimens.



K-MEANS CLUSTERING

The final number, *k*, of clusters is here decided at the outset. The algorithm is:



1. Select the location of *k* centroids at random.
2. All objects are assigned to the closest centroid.
3. Recalculate the location of the *k* centroids.
4. Repeat steps 2 and 3 until reaching convergence.

The method is fast, but the solution may be sensitive to the selection of the starting centroids.

CLUSTERING REMARKS

- The method is primarily a classification tool devoid of a statistical background.
- The more complex the dataset, the more likely that different methods will generate different results.
- The k-means methods is an extreme case, as even different runs for the same number of clusters may produce different results.
- Often solutions are suboptimal, failing to perfectly honor the intended objective of the method.
- In the absence of clear cut selection criteria, convenience in the eye of the user remains as the ultimate consideration on choosing the number of final clusters and clustering method.
- If totally lost, go for the Ward's method followed by the k-means method starting from the cluster generated by Ward's method.

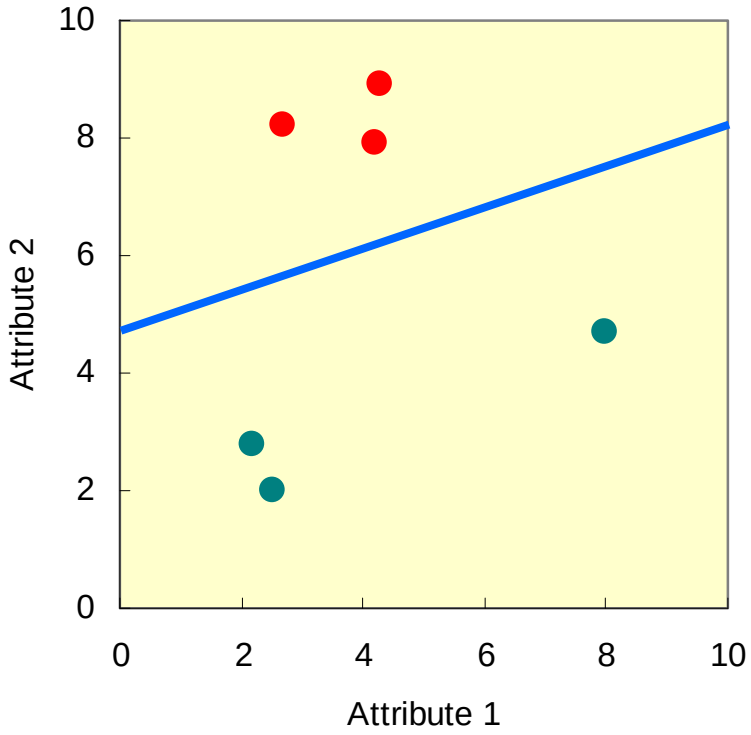
USGS

DISCRIMINANT ANALYSIS

BASIC IDEA

Discriminant analysis is a mixture of classification and prediction method under different availability of data than in cluster analysis.

- The classes are known for all objects in a training set.
- The training set is a data set intended for a second stage classification of objects without class assignments.
- The problem is solved by minimizing misclassification, which starts by finding class geometric boundaries for the training set in the data space.



USGS

ASSUMPTIONS

Discriminant analysis is a true statistical method based on multivariate distribution modeling. Important assumptions for making the analysis possible are:

- The data are a sample from a multivariate normal distribution. Thus, all attributes are normally distributed within each class.
- Any specimen has the same probability of belonging to any of the classes.
- None of the attributes is a linear combination of the others.
- The means for attributes across groups are not correlated with the variances.

Although in practice all assumptions are never simultaneously satisfied, the formulations are robust enough to tolerate departures.

VARIANTS

There are different approaches to discriminant analysis, but the main difference is in the type of surfaces employed to establishing the class boundaries.

- Linear methods, in which the surfaces are hyperplanes in an *m* dimensional space, where *m* is the number of attributes considered in the analysis. Linear discriminant analysis results from assuming all classes have the same covariance matrix.
- Quadratic methods, in which the boundaries are polynomials of order up to 2, with each class having a different covariance matrix.

TESTS

The following tests are integral parts of the procedure:

- Homogeneity of covariances
- Equality of means
- Normality

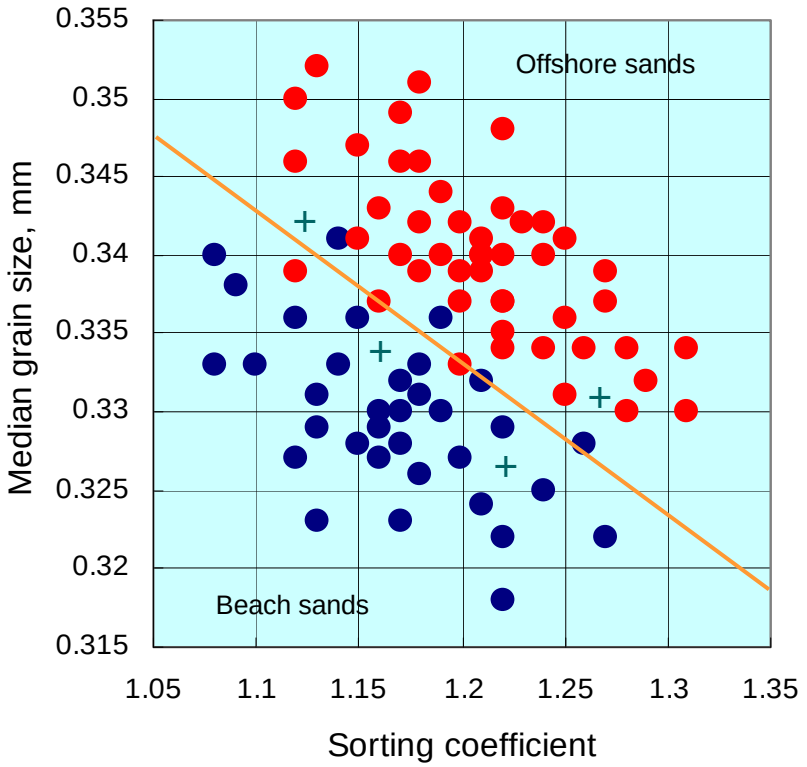
TEXAS GULF COAST SAND EXAMPLE

Test for homogeneity of covariances

$p = 0.85$

Test for equality of means

$p = 0.00$



Classification table

	Beach	Offshore
Beach	31	3
Offshore	3	44

Kolmogorov-Smirnov test for normality

	Statistic	Probability
Beach	0.056	0.96390
Offshore	0.034	0.99999

Assignments (+)

Sorting	Med. size	Pr[Beach]	Pr[Offsh.]
1.22	0.327	0.97	0.03
1.17	0.334	0.85	0.15
1.12	0.342	0.40	0.60
1.27	0.331	0.12	0.88

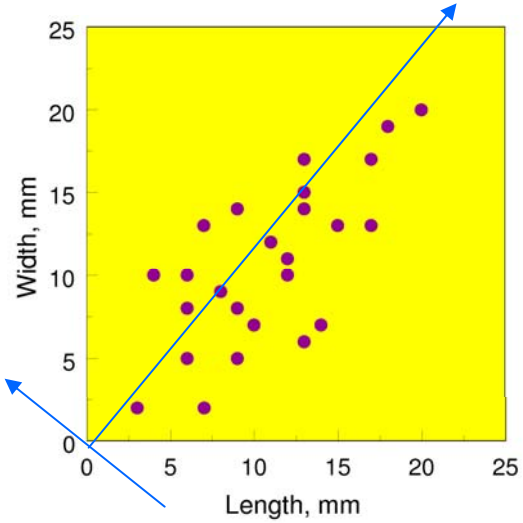
USGS

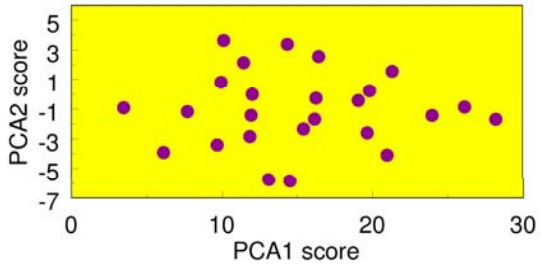
PRINCIPAL COMPONENT ANALYSIS

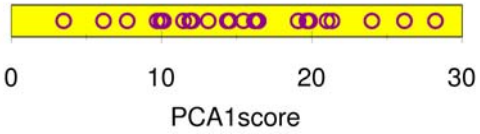
BASIC IDEA

The main objective of principal component analysis is to reduce the dimensionality of the sample.

- The new variables are linear combinations of the old ones.
- The new axes are orthogonal and oriented in directions of maximum variation.
- The new variables account for the same total variance, but in decreasing proportions.
- If explanation of less than 100% of the total variation is acceptable, one can drop the less relevant new variables.







USGS

MATRIX TERMINOLOGY

The determinant of a square matrix, $|\mathbf{A}|$, of order m is an expression having additive and subtractive products of m coefficients. The exact rules are complicated, but for order 2, for example, $|\mathbf{A}| = a_{11} \cdot a_{22} - a_{12} \cdot a_{21}$.

Given a symmetric square matrix, its eigenvalues are the solution, $\mathbf{\Lambda}$, to the equation that results from subtracting the unknown value λ from the main diagonal of its determinant.

$$\begin{vmatrix} a_{11}-\lambda & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22}-\lambda & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mm}-\lambda \end{vmatrix} = 0$$

$$\mathbf{\Lambda} = \begin{bmatrix} \lambda_1 & \lambda_2 & \cdots & \lambda_m \end{bmatrix}^T$$

An eigenvector $\mathbf{X}_i$ is the solution to equation system obtained by subtracting an eigenvalue from all diagonal terms of **A** when **B** = **0**.

$$\begin{bmatrix} a_{11}-\lambda_i & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22}-\lambda_i & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mm}-\lambda_i \end{bmatrix} \begin{bmatrix} x_{i_1} \\ x_{i_2} \\ \vdots \\ x_{i_m} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

USGS

METHOD

Solution to the problem comes from finding the orthogonal combination of attributes with maximum variance.

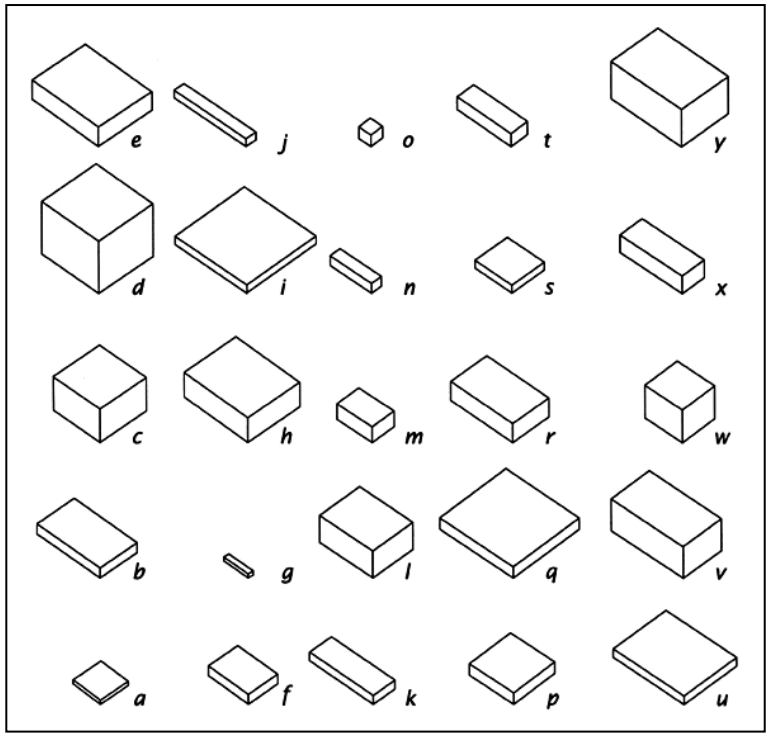
Given *n* records containing measurements for *m* attributes, one can calculate an *m* by *m* symmetric covariance matrix. The actual multivariate distribution is irrelevant.

The directions of maximum variance are provided by the eigenvectors in the form of directional cosines. The axes in the new system of reference remain orthogonal.

The length of each axis is taken as twice the eigenvalue.

Borrowing from the properties of matrices, the new axes are oriented along the directions of maximum variance and ranked by decreasing length.

BOX EXAMPLE



Variables

$x_1 = \text{long edge}$

$x_2 = \text{middle edge}$

$x_3 = \text{short edge}$

$x_4 = \text{longest diagonal}$

$$x_5 = \frac{\text{radius of smallest circumscribe sphere}}{\text{radius of largest inscribe sphere}}$$

$$x_6 = \frac{\text{long edge} + \text{intermediate edge}}{\text{short edge}}$$

$$x_7 = \frac{\text{surface area}}{\text{volume}}$$

USGS

PRINCIPAL COMPONENTS EXAMPLE

Coefficient matrix

$$\mathbf{A} = \begin{bmatrix} 5.400 & & & & & & \\ 3.260 & 5.846 & & & & & \\ 0.779 & 1.465 & 2.774 & & & & \\ 6.391 & 6.083 & 2.204 & 9.107 & & & \\ 2.155 & 1.312 & -3.839 & 1.611 & 10.714 & & \\ 3.035 & 2.877 & -5.167 & 2.783 & 14.774 & 20.776 & \\ -1.996 & -2.370 & -1.740 & -3.283 & 2.252 & 2.622 & 2.594 \end{bmatrix}$$

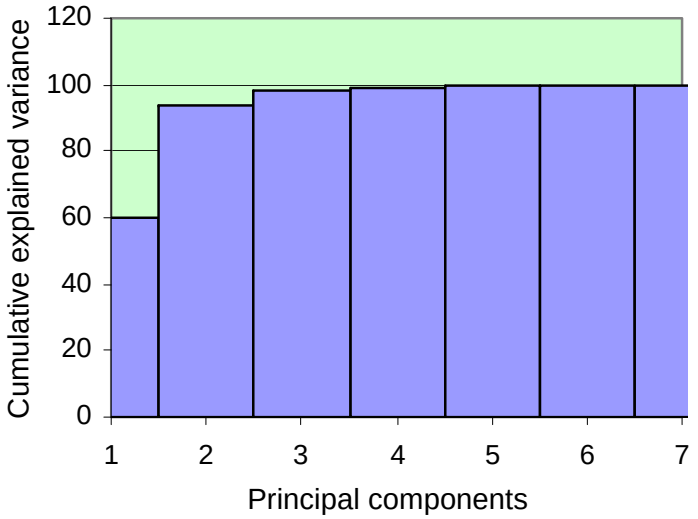
Eigenvalues

$$\mathbf{\Lambda} = \begin{bmatrix} 34.491 & 18.999 & 2.539 & 0.806 & 0.341 & 0.033 & 0.003 \end{bmatrix}^T$$

Eigenvectors

$$\mathbf{X} = \begin{bmatrix} 0.164 & -0.422 & -0.645 & & 0.225 & 0.415 & 0.385 \\ 0.142 & -0.447 & 0.713 & & 0.395 & & 0.329 \\ -0.173 & -0.257 & 0.130 & -0.629 & -0.607 & 0.280 & 0.211 \\ 0.170 & -0.650 & -0.146 & -0.212 & & -0.403 & -0.565 \\ 0.546 & 0.135 & -0.105 & -0.164 & -0.161 & -0.596 & 0.514 \\ 0.768 & 0.133 & 0.149 & & -0.207 & 0.465 & -0.327 \\ & 0.313 & & -0.719 & 0.596 & 0.107 & \end{bmatrix}$$

Orientation of second new axis



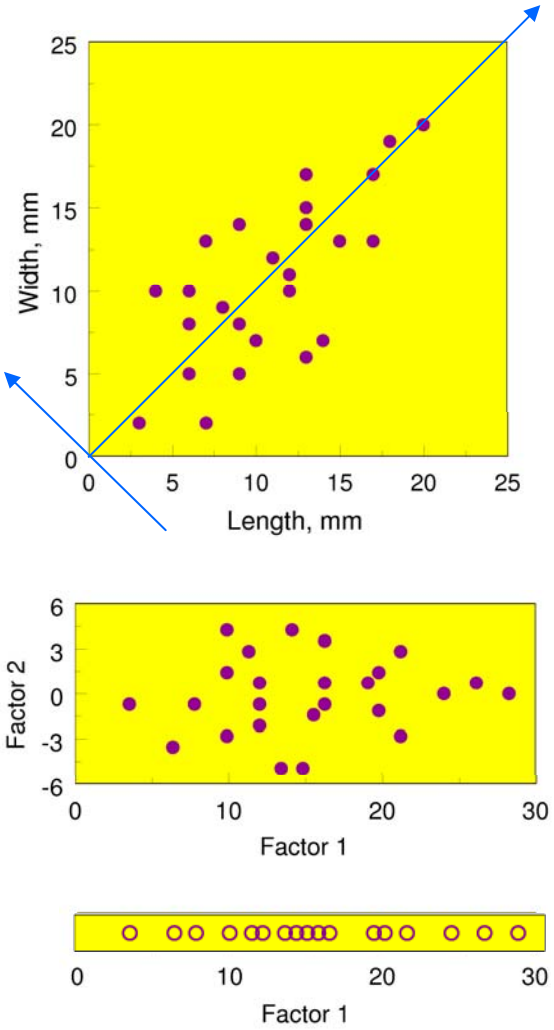
FACTOR ANALYSIS

INTENDED PURPOSE

Factor analysis was originally formulated in psychology to assess intangible attributes, like intelligence, in terms of abilities amenable of testing, such as reading comprehension.

Principal component analysis can be used to run a factor analysis, which sometimes contributes to confusion of the two methods.

Principal component factor analysis employs a correlation coefficient matrix obtained after standardizing the data.



GENERAL ASSUMPTIONS

Factor analysis, unlike principal component analysis, follows the canons of statistical modeling, thus

- opening possibilities in terms of testing, yet
- creating restrictions in utilization: data must comply with some assumption for the method to be applicable.

Analogously to regression, the model is:

$$x_i = \sum_{k=1}^{p} a_{ik} \cdot f_k + \varepsilon_i, i = 1,2,\ldots,m$$

where x_i is the ith observed attribute.

a_{ik} and f_k are the loadings and factors to come from the analysis.

ε_i is a random error.

It is also assumed that all variables are multinormally distributed.

VARIANTS

The two main approaches to factor analysis are:

- principal components
- maximum likelihood

In addition to the general assumptions, maximum likelihood factor analysis assumes that:

- Factors follow a standard normal distribution, N(0, 1).
- All of the factors are independent.
- The correlation coefficient matrix prepared with the data is a good approximation to the population correlation coefficient matrix .

USGS

BOX EXAMPLE REVISITED

Variables

$x_1 = \text{long edge}$

$x_2 = \text{middle edge}$

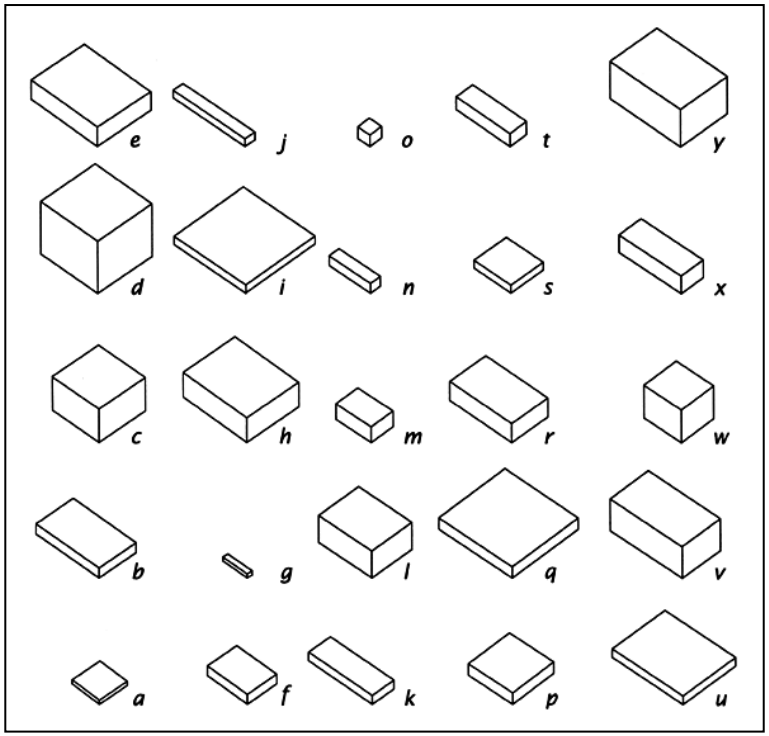
$x_3 = \text{short edge}$

$x_4 = \text{longest diagonal}$

$$x_5 = \frac{\text{radius of smallest circumscribe sphere}}{\text{radius of largest inscribe sphere}}$$

$$x_6 = \frac{\text{long edge} + \text{intermediate edge}}{\text{short edge}}$$

$$x_7 = \frac{\text{surface area}}{\text{volume}}$$



PRINCIPAL COMPONENT FACTOR ANALYSIS FOR THE BOX EXAMPLE

Coefficient matrix

$$\mathbf{A} = \begin{bmatrix} 1.000 & & & & & & \\ 0.580 & 1.000 & & & & & \\ 0.201 & 0.364 & 1.000 & & & & \\ 0.911 & 0.834 & 0.439 & 1.000 & & & \\ 0.283 & 0.166 & -0.704 & 0.163 & 1.000 & & \\ 0.287 & 0.261 & -0.681 & 0.202 & 0.990 & 1.000 & \\ -0.533 & -0.609 & -0.649 & -0.676 & 0.427 & 0.357 & 1.000 \end{bmatrix}$$

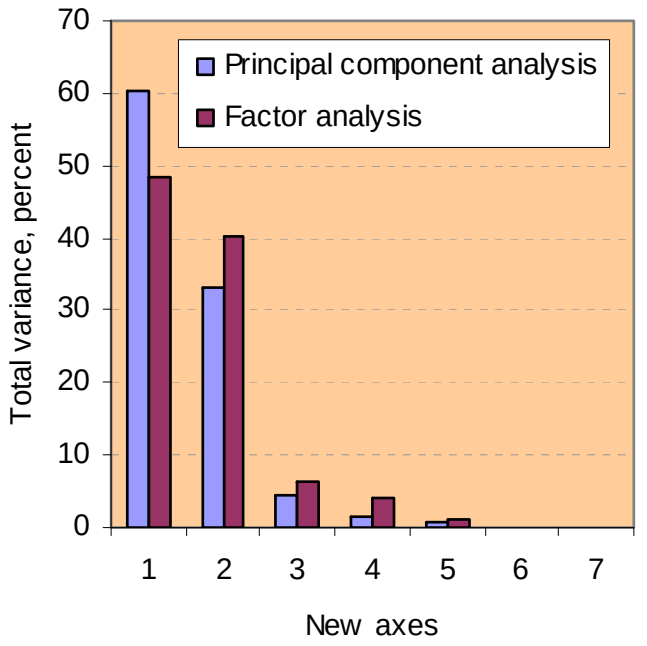
Eigenvalues

$$\mathbf{\Lambda} = \begin{bmatrix} 3.395 & 2.805 & 0.437 & 0.278 & 0.081 & 0.003 & 0.000 \end{bmatrix}^T$$

Eigenvectors (Factors)

$$\mathbf{X} = \begin{bmatrix} -0.405 & 0.293 & -0.667 & & -0.227 & -0.410 & 0.278 \\ -0.432 & 0.222 & 0.698 & & -0.437 & -0.144 & 0.254 \\ -0.385 & -0.356 & 0.148 & 0.628 & 0.512 & -0.188 & 0.108 \\ -0.494 & 0.232 & -0.119 & 0.210 & -0.105 & 0.588 & -0.536 \\ 0.128 & 0.575 & & 0.111 & 0.389 & 0.423 & 0.556 \\ & 0.580 & 0.174 & & 0.355 & -0.500 & -0.497 \\ 0.481 & 0.130 & & 0.735 & -0.455 & & \end{bmatrix}$$

Orientation of second new axis



ACKNOWLEDGMENTS

I am indebted for the valuable suggestions received as part of the review process of these notes from Emil Attanasi (U.S. Geological Survey), Geoffrey Bohling (Kansas Geological Survey), and Vera Pawlowsky-Glahn (Universitat de Girona, Spain). I am also grateful to early participants of my lectures who contributed with their remarks to make corrections and reformulate some topics. Eric Morrissey (U. S. Geological Survey) added tags to facilitate use of the notes by visually impaired readers.

BIBLIOGRAPHY

Aitchison, J., 2003, The statistical analysis of compositional data: Blackburn Press, Caldwell, NJ, reprint, 416 p.

Ambler. S., 2003, Agile Database Techniques: Effective Strategies for the Agile Software Developer: John Wiley & Sons, New York, 480 p.

Buccianti, A., G. Mateu-Figueras and V. Pawlowsky-Glahn, editors, 2006, Compositional data analysis from theory to practice: The Geological Society, London, Special Publication No. 264, 212 p.

Chernick, M. R., 2008, Bootstrap Methods: A Guide for Practitioners and Researchers: Wiley Interscience, Hoboken, NJ, second edition, 369 p.

Davis, J. C., 2002, Statistics and Data Analysis in Geology: John Wiley and Sons, New York, third edition, 638 p.

Donnelly, R. A., Jr., 2007, The Complete Idiot's Guide to Statistics: Alpha, New York, second edition, 395 p.

Good, P. I. and J. W. Hardin, 2006, Common Errors in Statistics (and How to Avoid Them): Wiley Interscience, Hoboken, NJ, 254 p.

Hastie, T., R. Tibshirani, and J. Friedman, 2001, The Elements of Statistical Learning—Data Mining, Inference, and Prediction: Springer, New York, 533 p.

Helsel, D. R., 2004, Nondetects and Data Analysis: Wiley Interscience, Hoboken, NJ, 268 p.

Holmgren, E. B., 1995, The P-P Plot as a Method for Comparing Treatment Effects: *Journal of the American Statistical Association*, vol. 90, no. 429, p. 360-365.

Jackson, S. L., 2005, Statistics Plain and Simple: Thomson Wadsworth, Belmont, CA, 269 p.

Moore, D. S. and W. I. Notz, 2006, Statistics—Concepts and Controversies: Freeman, New York, sixth edition, 561 p.

Salkind, N. J., 2007, Statistics for People Who (Think They) Hate Statistics: Sage Publications, Thousand Oaks, CA, 402 p.

Thomsen, V., D. Schatzlein, and D. Mercuro, 2003, Limits of detection in spectroscopy: www.spectroscopyonline.com, vol. 18, no. 12, p. 112-114,

Venables, W. N. and B. D. Ripley, 2003, Modern Applied Statistics with S: Springer, New York, 495 p.

Walpole, R. E., R. H. Myers, S. L. Myers, and K. Ye, 2006, Probability and Statistics for Engineers and Scientists: Prentice Hall, eighth edition, 848 p.

INDEX

accuracy 13, 15
agglomerative hierarchical method 159, 161
alternative hypothesis 112–113, 115, 120
analysis
  cluster 156–165, 167
  descriptive 38
  discriminant analysis 167-171
  factor 179–183
  principal component 173–177, 179–181, 183
  variance 138-140
arithmetic mean 47
assignment 167, 171
autocovariance 97
Bayes's theorem 66
bias 95
binomial distribution 68
bootstrap method 76, 109–110
box-and-whisker plot 59
causality 88
central limit theorem 125–127
Chebyshev's theorem 53
chi-square
  distribution 142–143
  test 144–147
classification 167, 169
cluster analysis 156–165, 167
coefficient
  kurtosis 63
  quartile skew 62
  skewness 61, 69
coin flipping 68, 71, 75, 123
compositional data 98–99
conditional probability 66
correlation 98
  coefficient 82, 86, 88, 179
covariance 81–82, 97, 158, 169–171
  matrix 158, 169, 175
cumulative distribution 43
decile 56
degrees of freedom 130, 132, 134–135, 142–144, 146
dendrogram 157
dependent variable 84
detection limit 21–27, 101–110
  instrumental 24
  method 25
  quantitation 26
determinant 174
discriminant analysis 167–171
dissimilarity 96, 159
distance 157–158
  Euclidean 158, 161
  Mahalanobis 158
  Manhattan 158
distribution 67, 125
  binomial 68
  chi-square 142–143
  continuous 67
  discrete 67
  F 134–135
  Gaussian 43, 63, 69–70, 104–105, 125–128, 131, 134, 142, 145, 148, 168, 170, 180
  lognormal 69, 105
  normal 43, 63, 69–70, 104–105, 125–128, 131,134, 142, 145, 148, 168, 170, 180
  sampling 119
  student's t 131–132
  testing 142–149
divisive hierarchical method 159, 162–163
eigenvalue 174–175, 177, 183
eigenvector 174–175, 177, 183
error 85
  mean square 86–87
  propagation 94
  Type I 114
  Type II 114
Euclidean distance 158, 161
expected value 71–73, 81
explanatory variable 84–85
extreme value 54, 59
factor 180–181
  analysis 179–183
F-distribution 134–136, 140
fractile 56
frequency table 40–41

Gaussian distribution 43, 63, 69–70, 104–105, 125–128, 131, 134, 142, 145, 148, 168, 170, 180
geometric mean 47, 110
geostatistics 151
histogram 42, 67, 75
hypothesis 120
  alternative 112–113, 115, 120
  null 112–117, 120, 128–129, 134, 144
independence 123, 148, 151, 152, 181
independent variable 84, 142
instrumental detection limit 24
interquartile range 57–58
k-means method 159, 164–165
Kaplan-Meier method 107, 110
Kolmogorov-Smirnov test 147–149, 171
kurtosis 63
legacy data 29–34
level of significance 114, 116–117, 120, 134, 144, 146, 149
limit of quantitation 26
linear
  correlation 82
  method 169
  regression 86–87
lognormal distribution 69, 105
log-ratio transformation 99, 109
lurking variable 88
Mahalanobis distance 158
main diagonal 155, 174
Manhattan distance 158
matrix 155, 174–175, 177, 183
  correlation coefficient 179, 181
  covariance 158, 169, 175
  order 155, 174
  square 155
maximum 54
  likelihood method 104–105, 110, 181
mean 47, 50, 59, 69, 72, 81, 91, 105, 124–125, 131, 142, 146, 168, 170–171
  arithmetic mean 47
  geometric mean 47, 110
  testing 124–128, 138–140
measure
  centrality 46–49
  dispersion 51–58
  shape 60–63
median 48, 50, 56
method
  agglomerative hierarchical 159, 161
  bootstrap 76, 109–110
  detection limit 25
  divisive hierarchical 159, 162–163
  k-means 159, 164–165
  linear 169
  Kaplan-Meier 107, 110
  maximum likelihood 104–105, 110, 181
  Monte Carlo 75–76
  quadratic 169
  replacement 103, 110
  Ward's 161, 165
minimum 54
mode 49
moment 73
Monte Carlo method 75–76, 94, 109
noise 87, 152
nonlinear regression 87
normal distribution 43, 63, 69–70, 104–105, 125–128, 131, 134, 142, 145, 148, 168, 170, 180
null hypothesis 112–117, 120, 128–129, 134, 144
outlier 58
order 155, 174
P-P plot 92, 149
p-value 117, 119–120, 136
percentile 56
plot
  P-P 92, 149
  Q-Q 90, 149
population 36, 124
power 114
precision 14–15, 101

principal component analysis 173–177, 179–181, 183
probability 65, 70, 117–118
  conditional 66
  density function 67
probability-probability plot 92, 149
proximity 159
Q-Q plot 90, 149
quadratic method 169
quantile 56
quantile-quantile plot 90, 149
quartile 56–57, 59, 105
radius of influence 97
random variable 37, 75
range 97
regressed variable 84
regression 84–88, 180
regressor 84
replacement method 103, 110
response variable 84
robustness 50, 87, 168
sample 36, 95
  size 38, 105, 129–131
sampling distribution 118
scatterplot 80, 149
semivariogram 96–97, 151–152
significant digit 17–19
sill 97
simulation 75
skewness 61, 69
standard deviation 53, 69, 91, 125, 142, 146
standardization 91–92, 127, 142, 158, 160, 179
statistic 45, 76, 109, 116–120, 144, 146, 147, 149
statistics 4–8, 37, 127, 136
student's t distribution 131–132
test
  chi-square 144–146
  distribution 142–149
  Kolmogorov-Smirnov 147–148, 171
  one-sided 113, 116, 123
  two-sided 113, 116, 140
  mean 124–128, 138–140
  variance 134–136
theorem
  Bayes's 66
  central limit 125–127
  Chebyshev's theorem 53
unbiasedness 95–96
uncertainty 5, 94
variable
  dependent 84
  explanatory 84–85
  independent 84
  lurking 88
  random 37, 75
  regressed 84
  response 84
variance 52–53, 73, 97, 105, 124, 134, 136, 138, 168, 173, 175
variate 37, 75
  testing 134–136
Ward's method 161, 165