Faculty: BioScienze e Tecnologie Agro-Alimentari e Ambientali
MASTER DEGREE IN FOOD SCIENCE AND TECHNOLOGY
I YEAR

Course:
EXPERIMENTAL DESIGN AND CHEMOMETRICS IN FOOD
(5 credits – 38 hours)

Teacher: Marcello Mascini

(mmascini@unite.it)

The Teacher is available to answer questions at the end of the lesson, or on request by mail

The course is split in 4 units

UNIT 1: statistical regression

Data, information, models, data types, analytical representation of data

Calibration and regression, Introduction to Statistics

Average & Variance

The Normal distribution, theory of measurement errors, the central limit theorem and the theorem of Gauss

Maximum likelihood, method of least squares, Generalization of the method of least squares

Polynomial regression, non-linear regression, the χ2 method, Validation of the model

UNIT 3: Data Matrices and sensor arrays

Correlation

Multiple linear regression

Principal component analysis (PCA)

Principal component regression (PCR) and Partial least squares regression - (PLS)

UNIT 2: Design of Experiments

Basic design of experiments and analysis of the resulting data

Analysis of variance, blocking and nuisance variables

Factorial designs

Fractional factorial designs

Overview of other types of experimental designs (Plackett–Burman designs, D-optimal designs, Supersaturated designs, Asymmetrical designs)

Response surface methods and designs

Applications of designed experiments from various fields of food science

UNIT 4: Elements of Pattern recognition

Cluster analysis

Normalization

The space representation (PCA) Examples of PCA

Discriminant analysis (DA) PLS-DA

Examples of PLS-DA

UNIT 1: statistical regression

Data, information, models, data types, analytical representation of data

Calibration and regression, Introduction to Statistics

Average & Variance

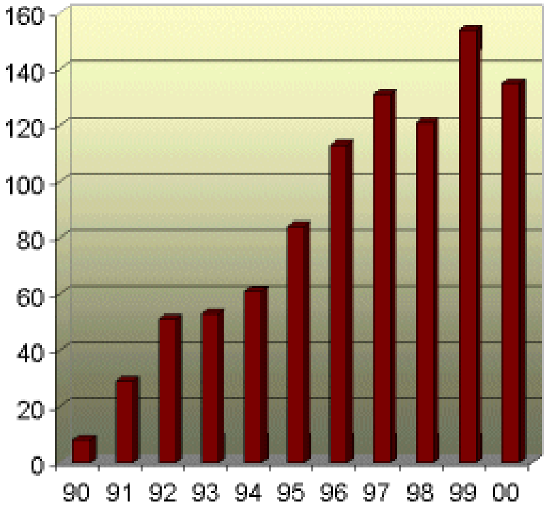
The Normal distribution, theory of measurement errors, the central limit theorem and the theorem of Gauss

Maximum likelihood, method of least squares, Generalization of the method of least squares

Polynomial regression, non-linear regression, the $\chi 2$ method, Validation of the model

CHEMOMETRICS

- The science of extracting information from chemical systems by data-driven means.
- It is a highly interfacial discipline, using methods frequently employed in core data-analytic disciplines such as multivariate statistics, applied mathematics, and computer science, in order to address problems in chemistry, biochemistry, medicine, biology and chemical engineering.
- The goal is using data from multidimensional signals for examples spectrometers or chromatograms



Number of Published Chemometrics Papers

Dedicated journals

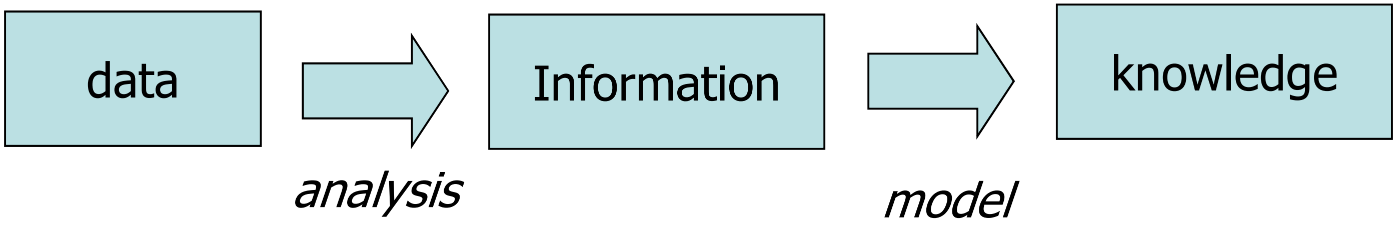
- Chemometrics and Intelligent Laboratory systems
- Journal of Chemometrics

Articles are published also in:

- Analytical Chemistry
- Analytica Chimica Acta
- Trends in Analytical Chemistry
- J. computer aided molecular design
- ………………

DATA

- Data are are individual pieces of information.
- For Example human being data:
  - Height, weight, chemical blood analysis DNA composition, hair color…
- Data can be qualitative or quantitative
- Data must be analysed to have information and to increase knowledge
  - Example: a chemical blood analysis has to be supported by a human being model



data

analysis

Information

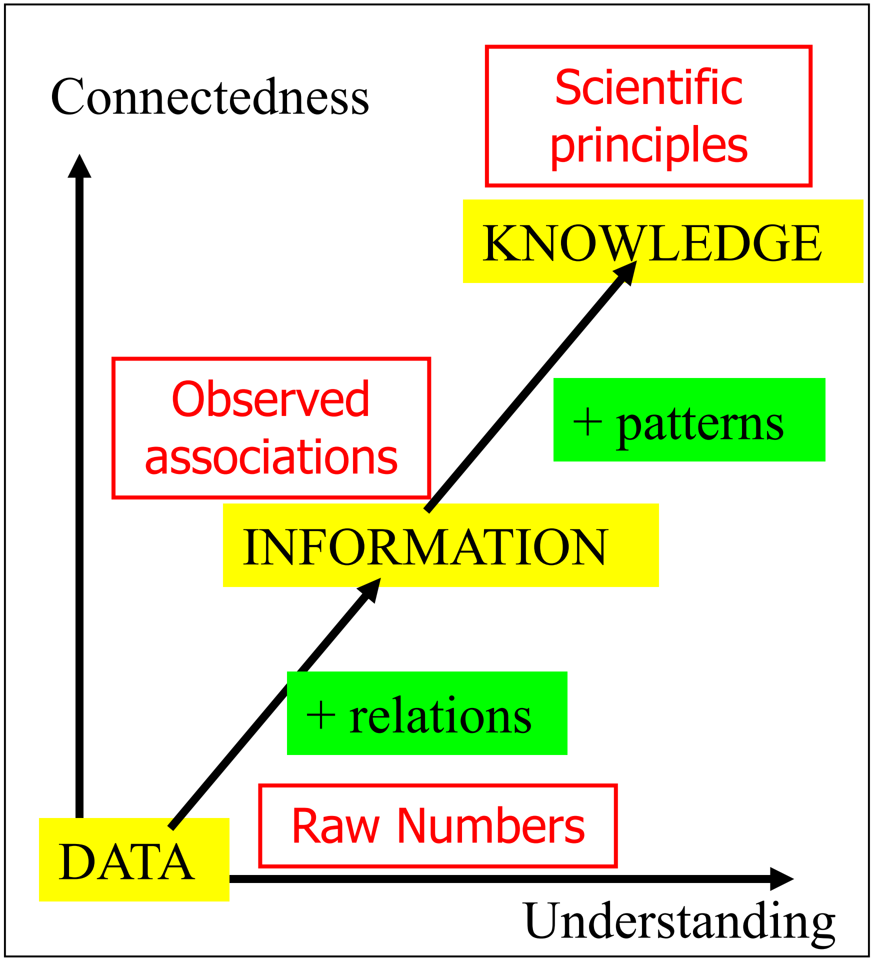
model

knowledge

Data ⇨ Information ⇨ Knowledge

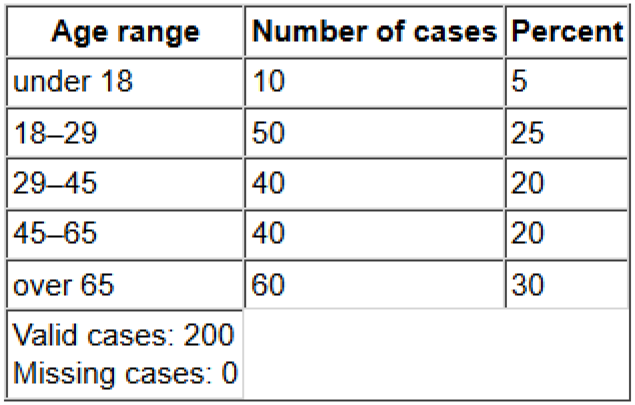
The aim of data-mining can be illustrated graphically as follows:

- Data
  - unrelated *facts*
- Information
  - facts plus *relations*
- Knowledge
  - information plus *patterns*



Univariate analysis

- Describing the distribution of a single variable, including its central tendency (including the mean, median, and mode) and dispersion (including the range and quantiles of the data-set, and measures of spread such as the variance and standard deviation). Characteristics of a variable's distribution may also be depicted in graphical or tabular format, including histograms and stem-and-leaf display.



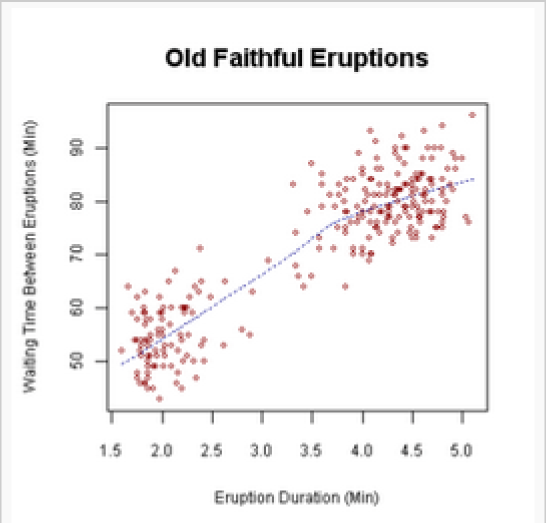
Age range	Number of cases	Percent
under 18	10	5
18–29	50	25
29–45	40	20
45–65	40	20
over 65	60	30
Valid cases: 200 Missing cases: 0		

- Any measurement can be judged by the following meta-measurement criteria values: level of measurement (which includes magnitude), dimensions (units), and uncertainty:
  - Electrical resistance is 100KΩ
  - The apple weight is è 80g
  - The K^+ concentration in water is 1.02 mg/l

Bivariate analysis

- It involves the analysis of two variables (often denoted as X, Y), for the purpose of determining the empirical relationship between them. In order to see if the variables are related to one another, it is common to measure how those two variables simultaneously change together (covariance).

- The major differentiating point between univariate and bivariate analysis, in addition to the latter's looking at more than one variable, is that the purpose of a bivariate analysis goes beyond simply descriptive: it is the analysis of the relationship between the two variables. Bivariate analysis is a simple (two variable) special case of multivariate analysis (where multiple relations between multiple variables are examined simultaneously)



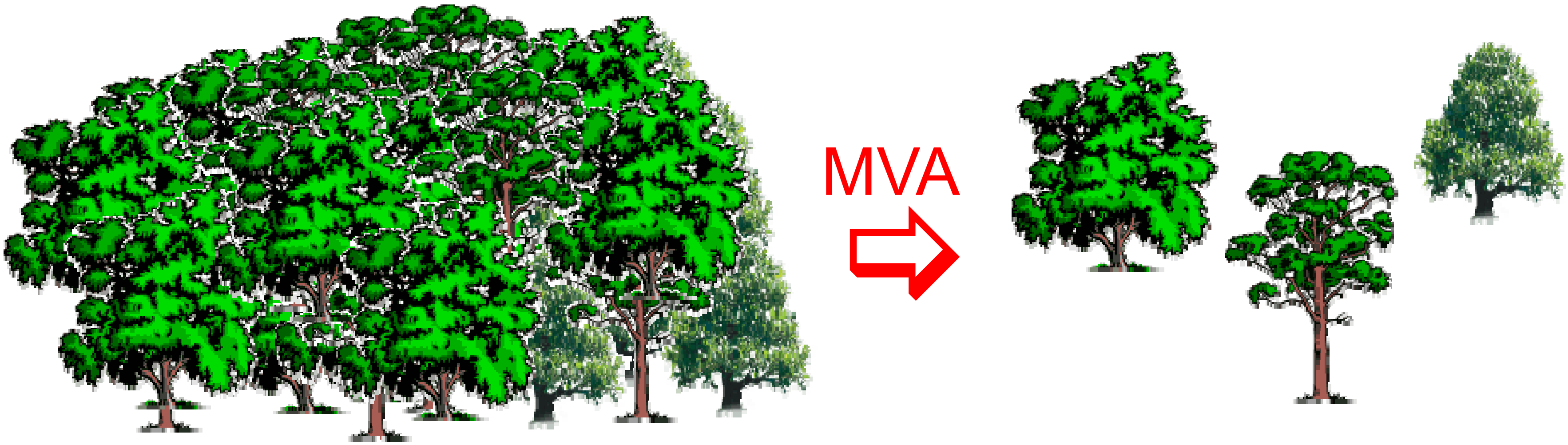
Why Multivariate?

- Typically more than one measurement is taken on a given experimental unit
- Need to consider all the measurements together so that one can understand how they are related
- Need to consider all the measurements together so that one can extract essential structure

What is MVA?

Multivariate analysis (MVA) is defined as the simultaneous analysis of more than five variables. Some people use the term “megavariate” analysis to denote cases where there are more than a hundred variables.

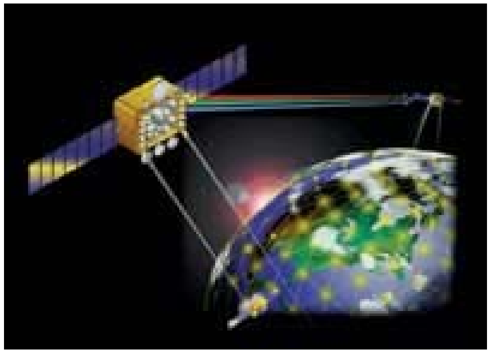
MVA uses ALL available data to capture the most information possible. The basic principle is to boil down hundreds of variables down to a *mere handful.*



Process Integration Challenge: Make sense of masses of data

Many organisations today are faced with the same challenge: TOO MUCH DATA. These include:

- –Business - *customer transactions*
- –Communications - *website use*
- –Government - *intelligence*
- –Science - *astronomical data*
- –Pharmaceuticals - *molecular configurations*
- –Industry - *process data*



It is the last item that is of interest to us as chemical engineers…

Graphical representation of MVA

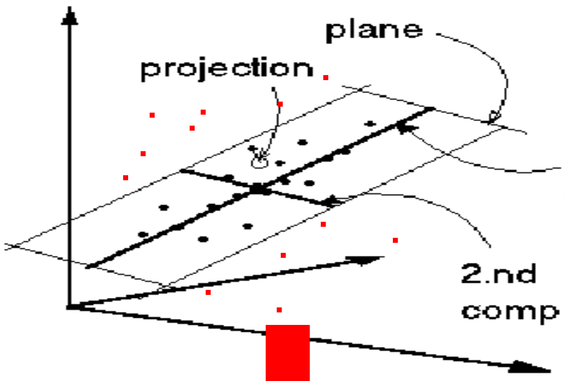
Tmt	X1	X4	X5	Rep	Y avec	Y sans
1	-1	-1	-1	1	2.51	2.74
1	-1	-1	-1	2	2.36	3.22
1	-1	-1	-1	3	2.45	2.56
2	-1	0	1	1	2.63	3.23
2	-1	0	1	2	2.55	2.47
2	-1	0	1	3	2.65	2.31
3						2.67
3						2.45
3						2.98
4						3.22
4						2.57
4	0	-1	1	3	2.97	2.63
5	0	0	0	1	2.89	3.16
5	0	0	0	2	2.56	3.32
5	0	0	0	3	2.52	3.26
6	0	1	-1	1	2.44	3.1
6	0	1	-1	2	2.22	2.97
6	0	1	-1	3	2.27	2.92

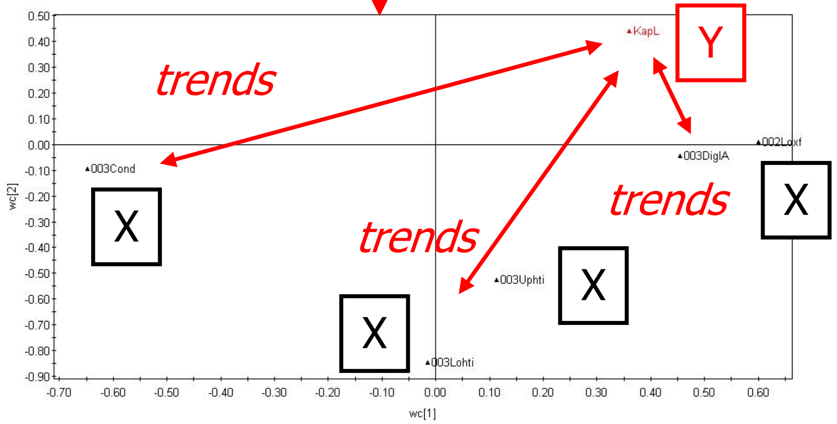
Raw Data:

impossible to interpret

Statistical Model

(internal to software)

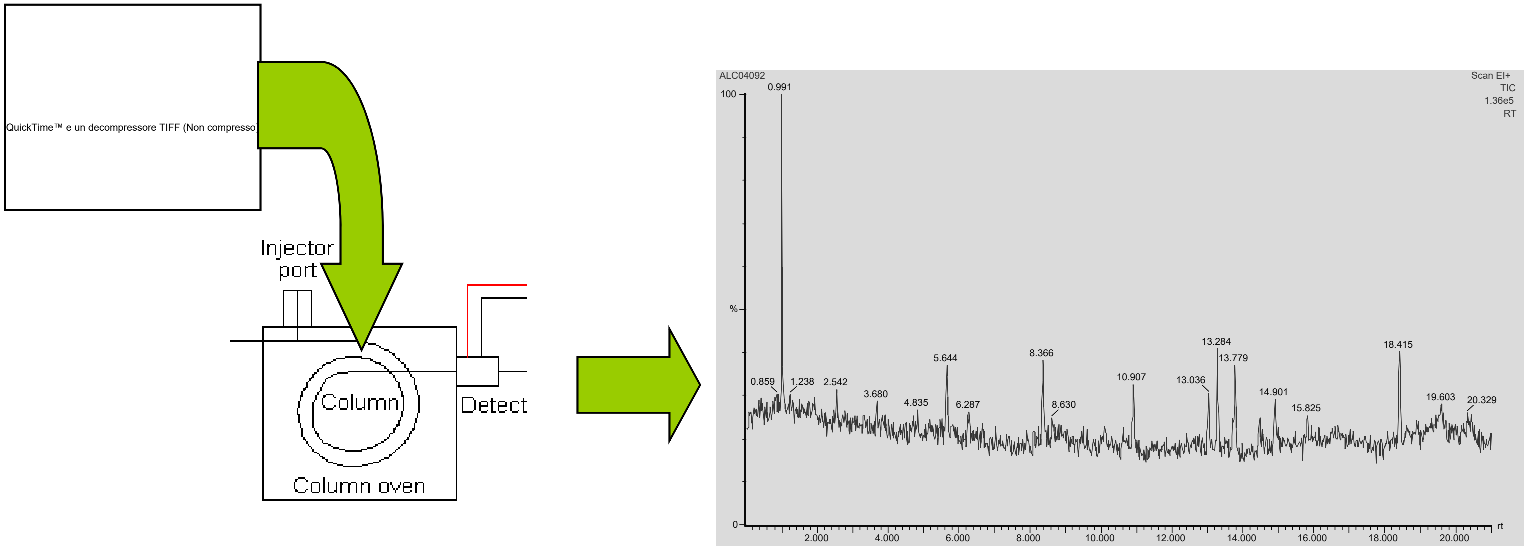




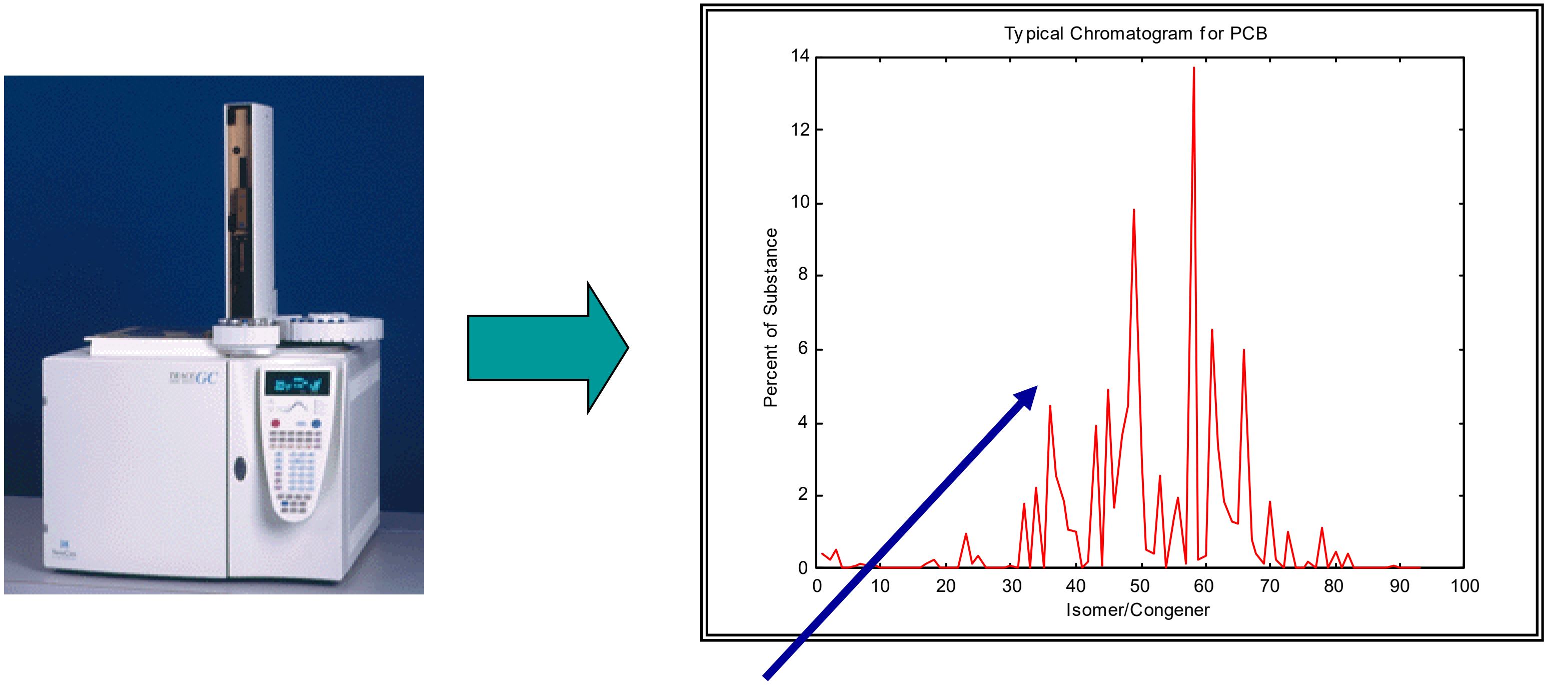
2-D Visual Outputs

Multidimensional Instruments

- **Gas chromatography**



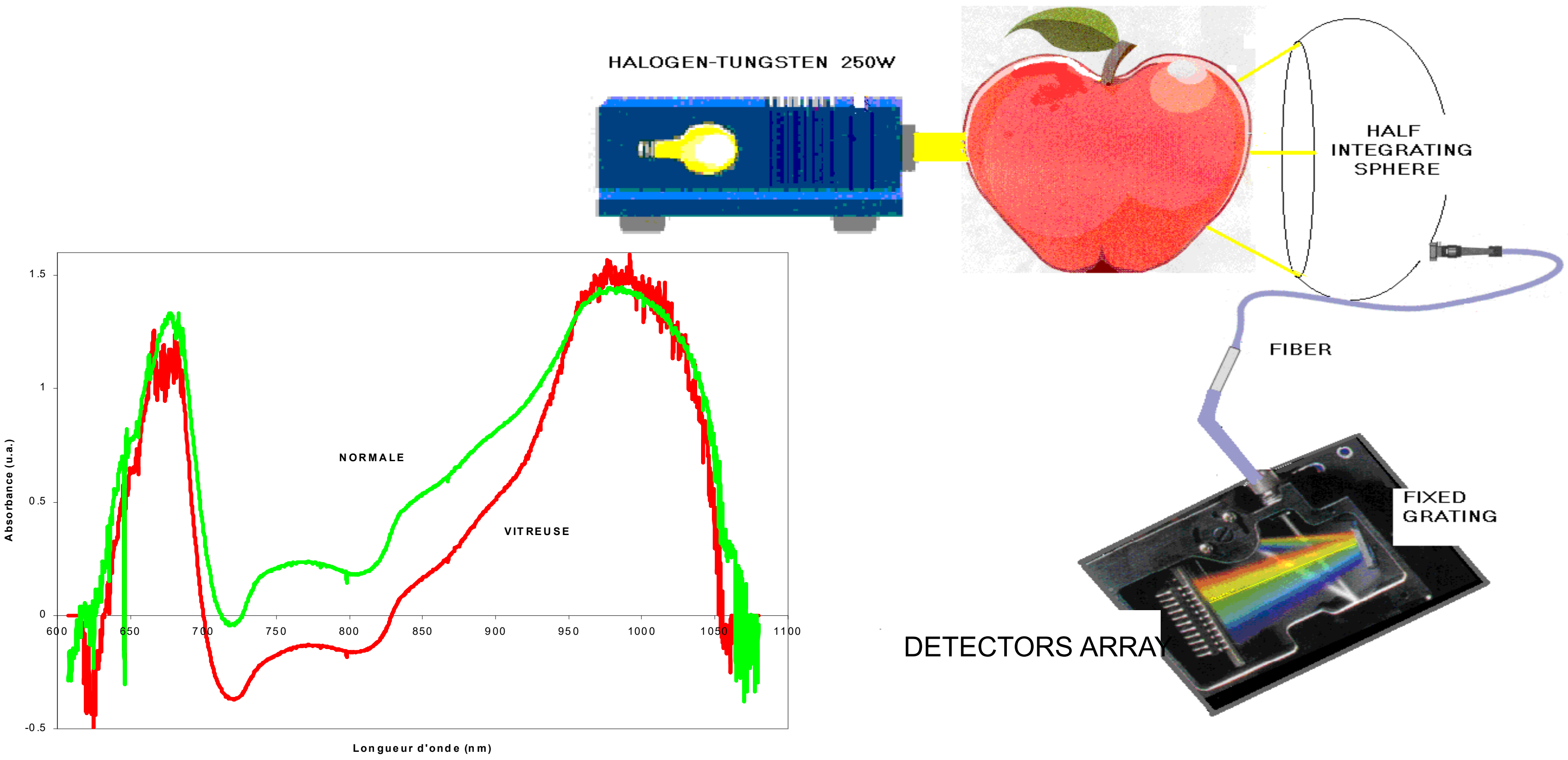
In Chromatography



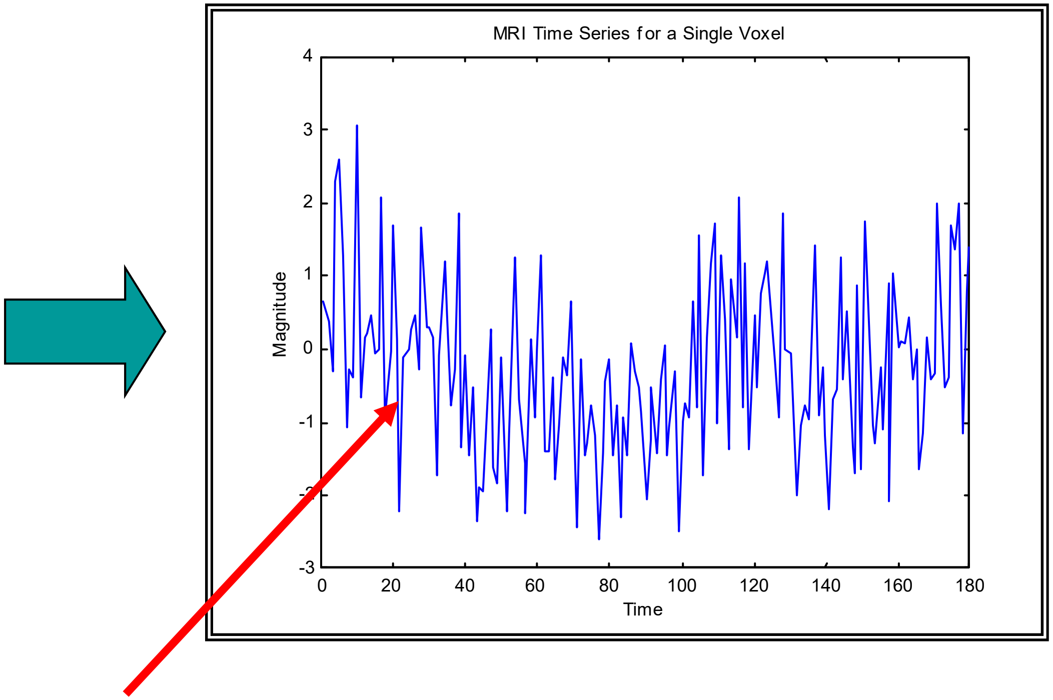
one observation

Multidimensional Instruments

- **Spectroscopy**
  - Vis/NIR of an apple



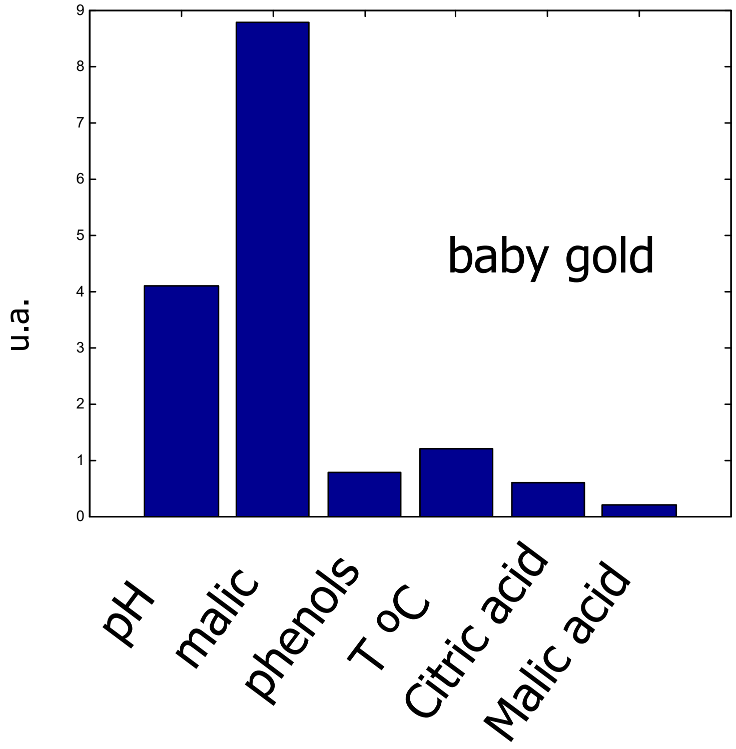
In Neuroimaging

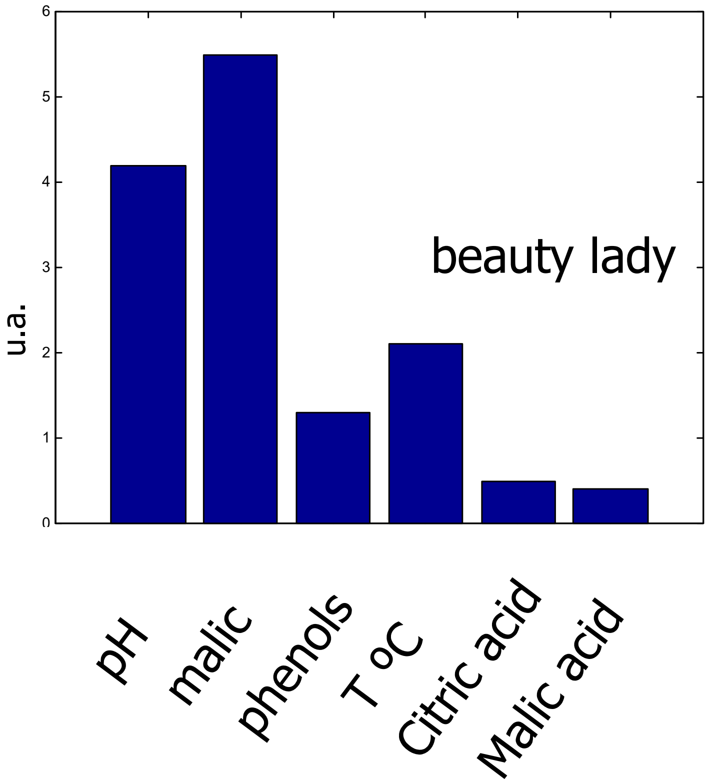


one observation

Multidimensionali Instruments

- **Sensors Array**

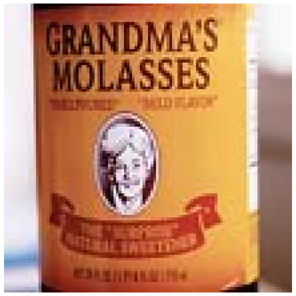




Illustrative Data Set: Food Consumption in European Countries

To illustrate these concepts, we take an easy-to-understand example involving food.

Data on food preferences in 16 different European countries are considered, involving the consumption patterns for 18 different food groups.



Look at the table on the following page. Can you tell anything from the raw numbers? Of course not. No one could.

Data Table: Food Consumption in European Countries

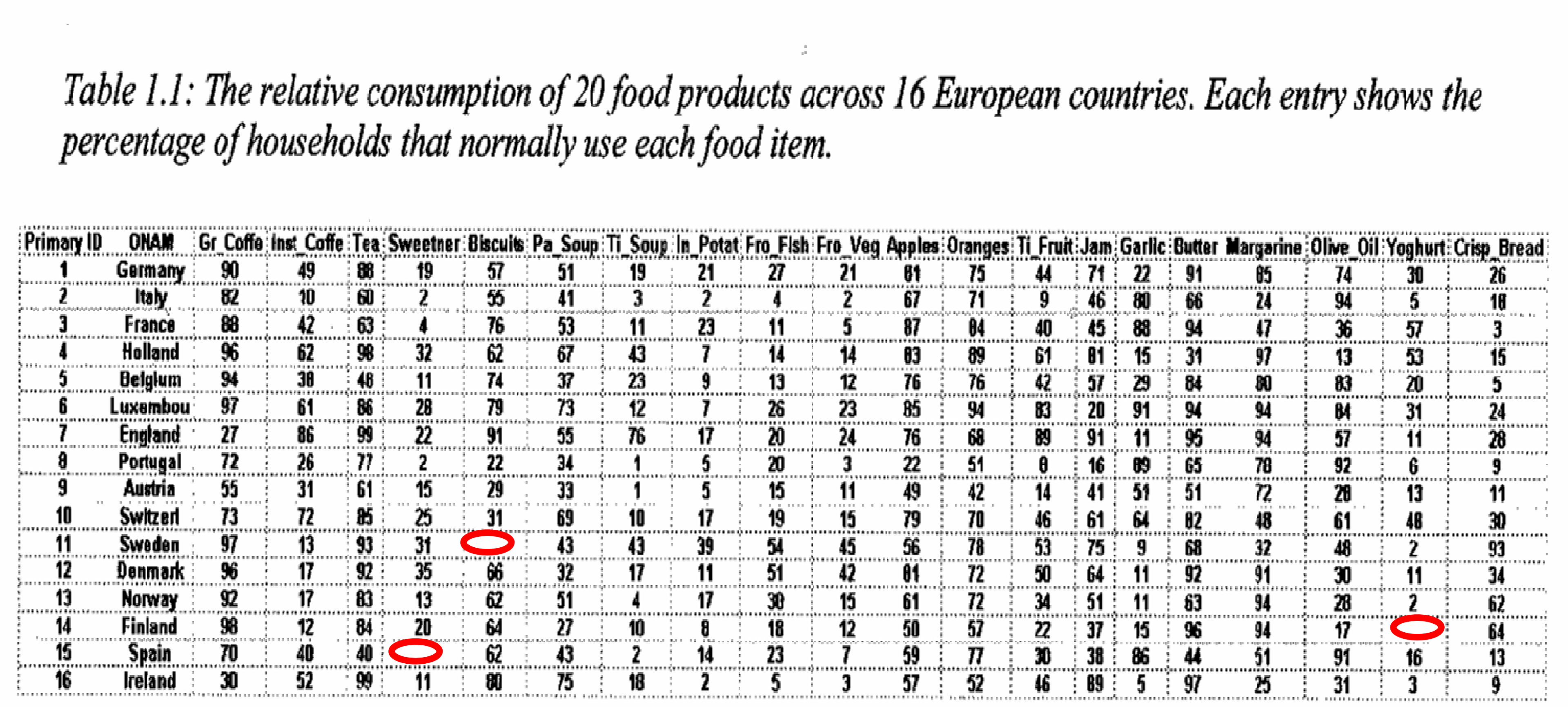


Table 1.1: The relative consumption of 20 food products across 16 European countries. Each entry shows the percentage of households that normally use each food item.

Primary ID	ONAM	Gr_Coffe	Inst_Coffe	Tea	Sweetner	Biscuits	Pa_Soup	Ti_Soup	In_Potat	Fro_Fish	Fro_Veg	Apples	Oranges	Ti_Fruit	Jam	Garlic	Butter	Margarine	Olive_Oil	Yoghurt	Crisp_Bread
1	Germany	90	49	88	19	57	51	19	21	27	21	81	75	44	71	22	91	85	74	30	26
2	Italy	82	10	60	2	55	41	3	2	4	2	67	71	9	46	80	66	24	94	5	18
3	France	88	42	63	4	76	53	11	23	11	5	87	84	40	45	88	94	47	36	57	3
4	Holland	96	62	98	32	62	67	43	7	14	14	83	89	61	81	15	31	97	13	53	15
5	Belgium	94	38	48	11	74	37	23	9	13	12	76	76	42	57	29	84	80	83	20	5
6	Luxembou	97	61	86	28	79	73	12	7	26	23	85	94	83	20	91	94	94	84	31	24
7	England	27	86	99	22	91	55	76	17	20	24	76	68	89	91	11	95	94	57	11	28
8	Portugal	72	26	77	2	22	34	1	5	20	3	22	51	8	16	89	65	78	92	6	9
9	Austria	55	31	61	15	29	33	1	5	15	11	49	42	14	41	51	51	72	28	13	11
10	Switzerl	73	72	85	25	31	69	10	17	19	15	79	70	46	61	64	82	48	61	48	30
11	Sweden	97	13	93	31		43	43	39	54	45	56	78	53	75	9	68	32	48	2	93
12	Denmark	96	17	92	35	66	32	17	11	51	42	81	72	50	64	11	92	91	30	11	34
13	Norway	92	17	83	13	62	51	4	17	30	15	61	72	34	51	11	63	94	28	2	62
14	Finland	98	12	84	20	64	27	10	8	18	12	50	57	22	37	15	96	94	17		64
15	Spain	70	40	40		62	43	2	14	23	7	59	77	30	38	86	44	51	91	16	13
16	Ireland	30	52	99	11	80	75	18	2	5	3	57	52	46	89	5	97	25	31	3	9

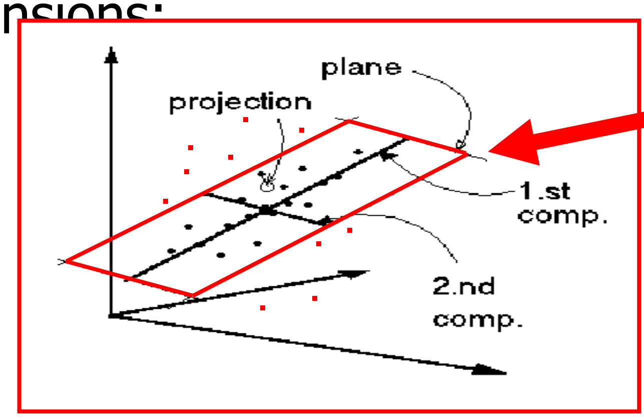
Note that MVA can handle up to 10-20% missing data

Courtesy of Umetrics corp.

Score Plot

The MVA software generates two main types of plots to represent the data: *Score* plots and *Loadings* plots.

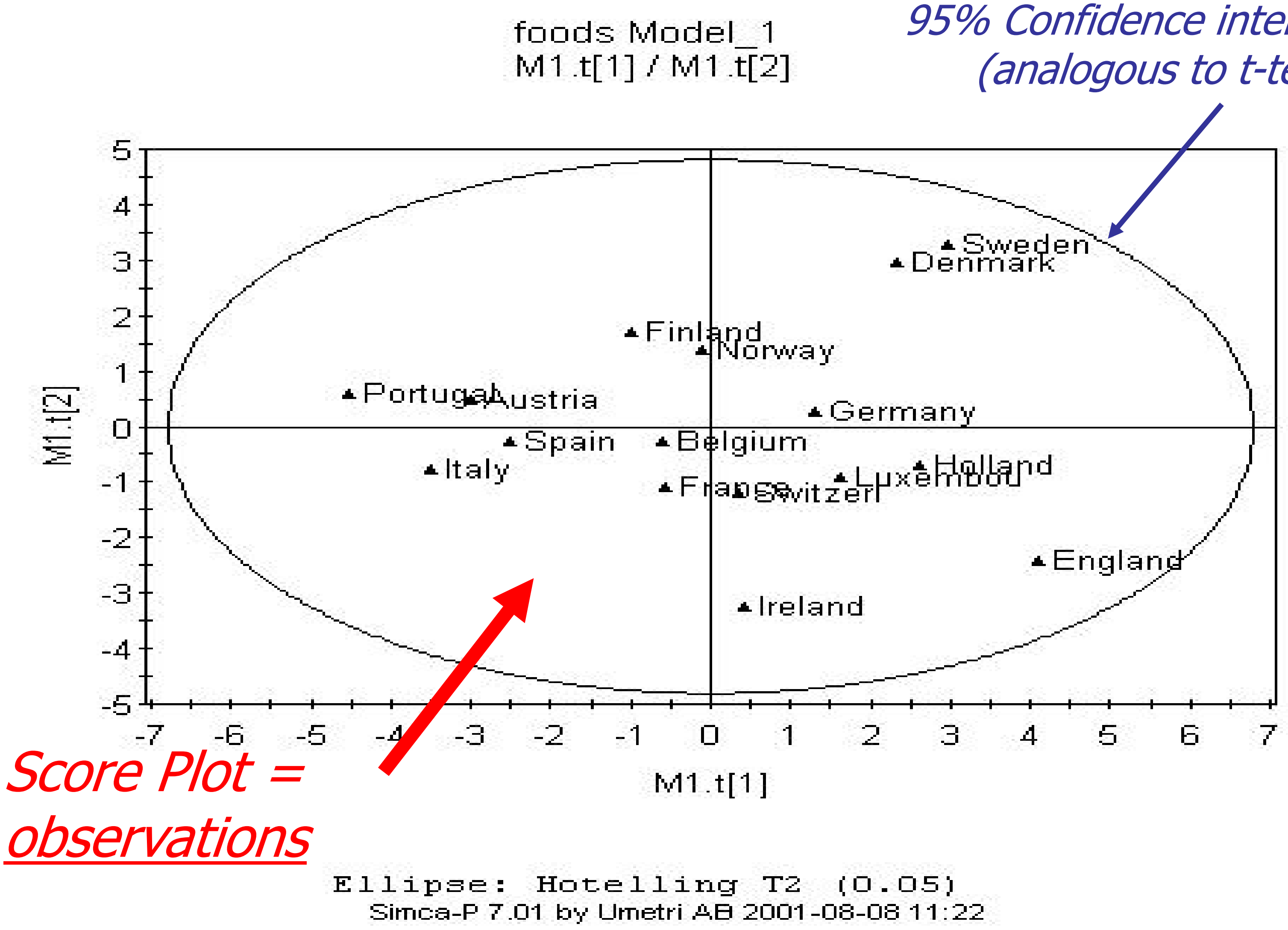
The first of these, the Score plot, shows all the original data points (observations) in a new set of coordinates or *components*. Each score is the value of that data point on one of the *new* component dimensions:



The *Score Plot* is the projection of the original data points onto a plane defined by two new *components*.

A score plot shows how the observations are arranged in the new component space. The score plot for the food data is shown on the next page. Note how similar countries cluster together…

Score Plot for Food Example



95% Confidence interval
(analogous to t-test)

Score Plot =
observations

Loadings Plot

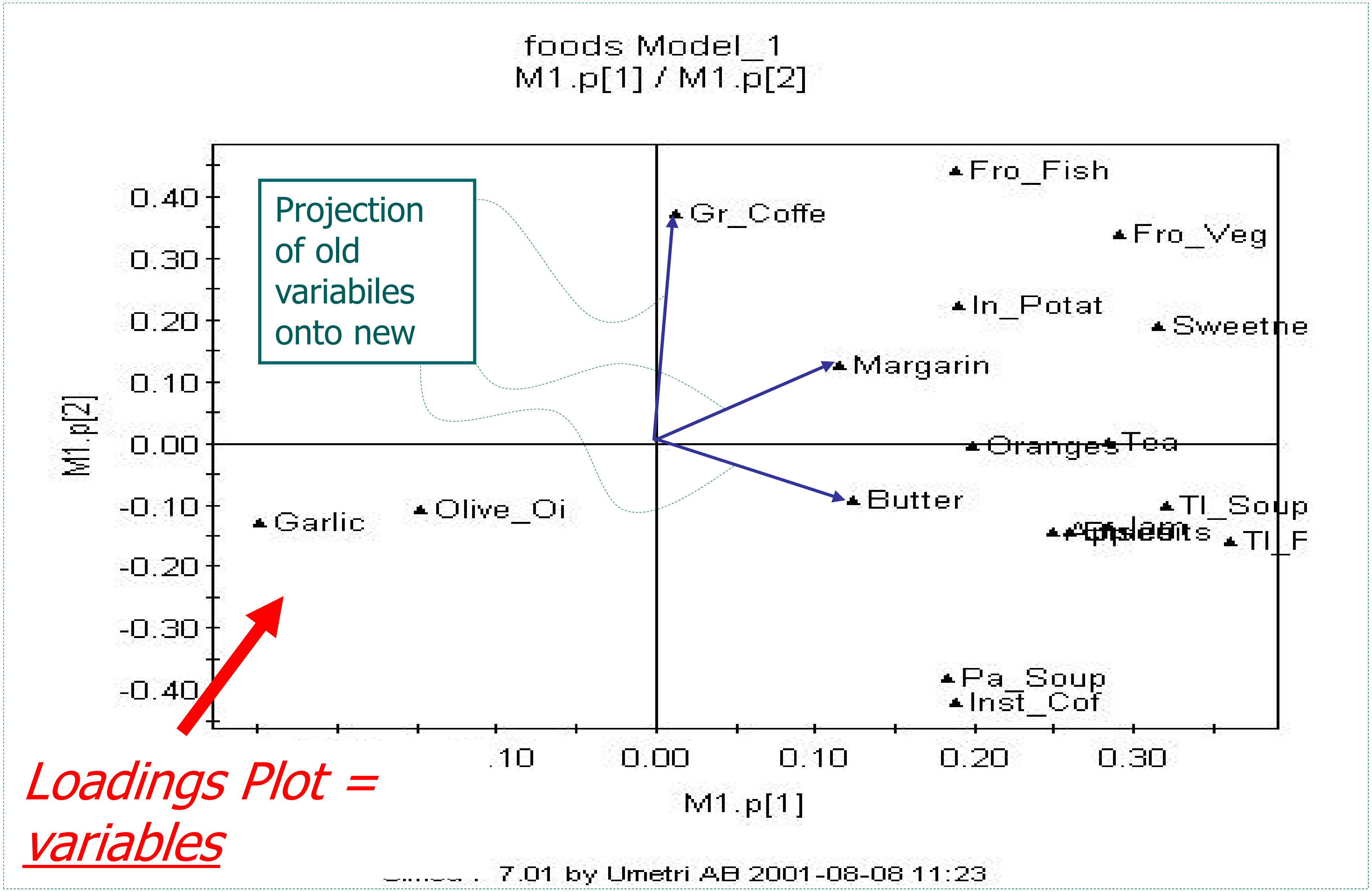
The second type of data plot generated by the MVA software is the *Loadings* plots. This is the equivalent to the score plot, only from the point of view of the original *variables*.

Each component has a set of *loadings* or weights, which express the projection of each original variable onto each new component.

Loadings show how strongly each variable is associated with each new component. The loadings plot for the food example is shown on the next page. The further from the origin, the more significant the correlation.

Note that the *quadrants* are the same on each type of plot. Sweden and Denmark are in the top-right corner; so are frozen fish and vegetables. Using both plots, variables and observations can be correlated with one another.

Use of loadings (illustration)



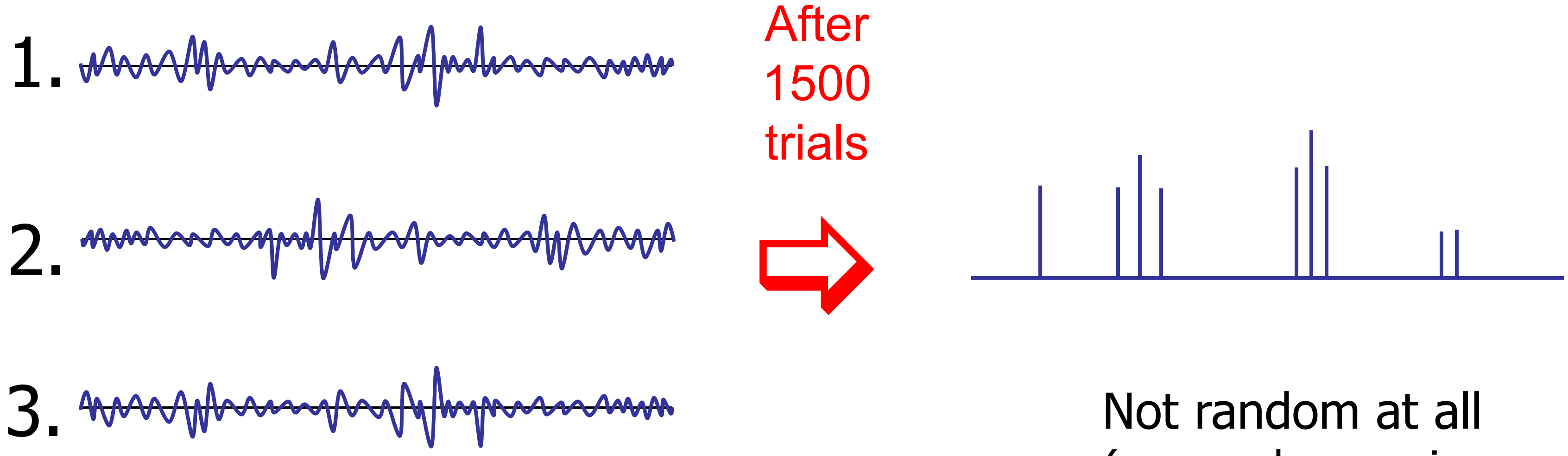
Projection of old variabiles onto new

Loadings Plot = variables

To MVA, Data Overload is Good!

One great advantage of MVA is that the more data are available, the less noise matters (assuming that the noise is normally distributed). This is one of the reasons MVA is used to mine huge amounts of data.

This is analogous to NMR measurements in a laboratory. The more trials there are, the clearer the spectrum becomes:



1.

2.

3.

Looks random

After 1500 trials

Not random at all
(+ve and –ve noise cancels out)

Multivariate Analysis: Benefits

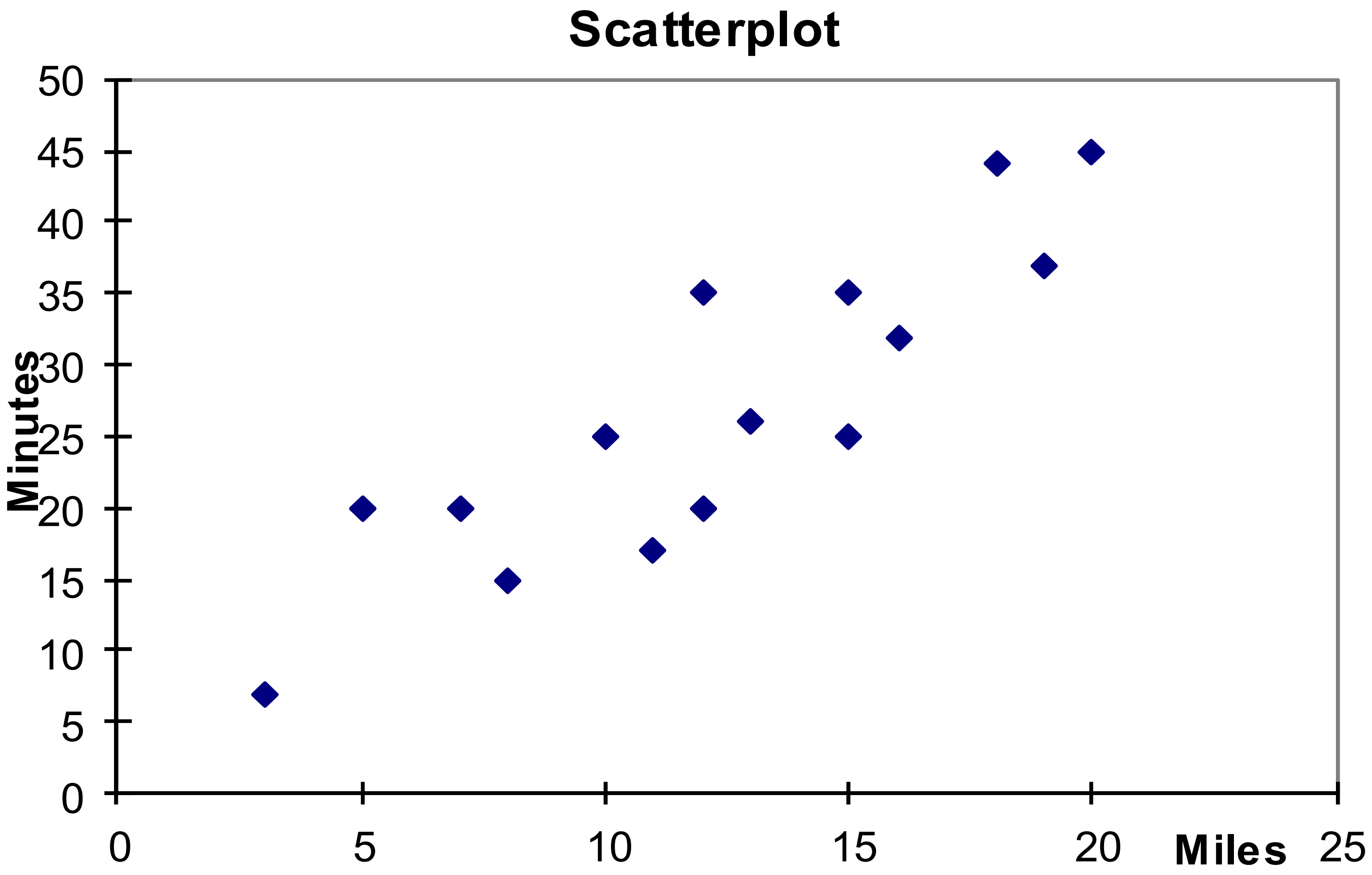
What is the point of doing MVA?

The first potential benefit is to explore the inter-relationships between different process variables. It is well known that simply creating a model can provide insight in the process itself ("Learn by modelling").

Once a representative model has been created, the engineer can perform "what if?" exercises without affecting the real process. This is a low-cost way to investigate options.

Some important parameters, like final product quality, cannot be measured in real time. They can, however, be inferred from other variables that are measured on-line. When incorporated in the process control system, this inferential controller or "soft sensor" can greatly improve process performance.

Statistics


Scatterplot
Minutes
50
45
40
35
30
25
20
15
10
5
0
0
5
10
15
20
Miles
25

What is Statistics?

Statistics: The science of collecting, describing, and interpreting data.

Two areas of statistics:

Descriptive Statistics: collection, presentation, and description of sample data.

Inferential Statistics: making decisions and drawing conclusions about populations.

Example: A recent study examined the math and verbal SAT scores of high school seniors across the country. Which of the following statements are descriptive in nature and which are inferential.

- The mean math SAT score was 492.
- The mean verbal SAT score was 475.
- Students in the Northeast scored higher in math but lower in verbal.
- 80% of all students taking the exam were headed for college.
- 32% of the students scored above 610 on the verbal SAT.
- The math SAT scores are higher than they were 10 years ago.

Introduction to Basic Terms

Population: A collection, or set, of individuals or objects or events whose properties are to be analyzed.

Two kinds of populations: *finite* or *infinite*.

Sample: A subset of the population.

Variable: A characteristic about each individual element of a population or sample.

Data (singular): The value of the variable associated with one element of a population or sample. This value may be a number, a word, or a symbol.

Data (plural): The set of values collected for the variable from each of the elements belonging to the sample.

Experiment: A planned activity whose results yield a set of data.

Parameter: A numerical value summarizing all the data of an entire population.

Statistic: A numerical value summarizing the sample data.

Example: A college dean is interested in learning about the average age of faculty. Identify the basic terms in this situation.

The *population* is the age of all faculty members at the college.

A *sample* is any subset of that population. For example, we might select 10 faculty members and determine their age.

The *variable* is the "age" of each faculty member.

One *data* would be the age of a specific faculty member.

The *data* would be the set of values in the sample.

The *experiment* would be the method used to select the ages forming the sample and determining the actual age of each faculty member in the sample.

The *parameter* of interest is the "average" age of all faculty at the college.

The *statistic* is the "average" age for all faculty in the sample.

Two kinds of variables:

Qualitative, or Attribute, or Categorical, Variable: A variable that categorizes or describes an element of a population.

Note: Arithmetic operations, such as addition and averaging, are *not* meaningful for data resulting from a qualitative variable.

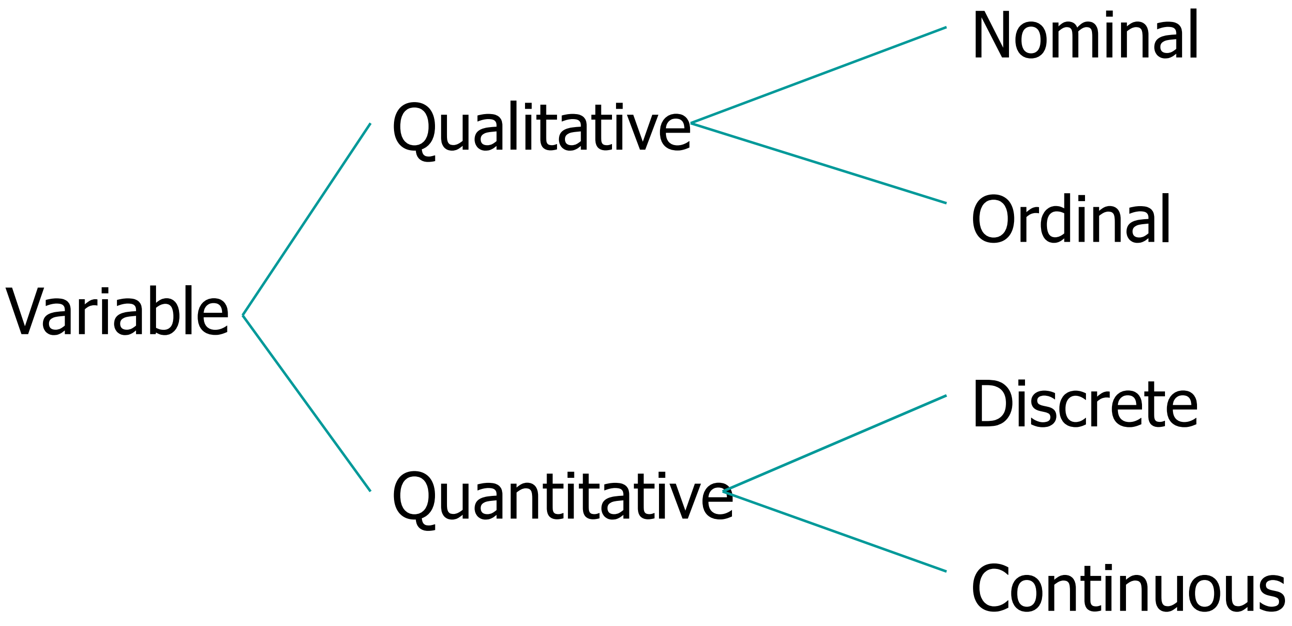
Quantitative, or Numerical, Variable: A variable that quantifies an element of a population.

Note: Arithmetic operations such as addition and averaging, are meaningful for data resulting from a quantitative variable.

Example: Identify each of the following examples as attribute (qualitative) or numerical (quantitative) variables.

1. The residence hall for each student in a statistics class. (Attribute)
2. The amount of gasoline pumped by the next 10 customers at the local Unimart. (Numerical)
3. The amount of radon in the basement of each of 25 homes in a new development. (Numerical)
4. The color of the baseball cap worn by each of 20 students. (Attribute)
5. The length of time to complete a mathematics homework assignment. (Numerical)
6. The state in which each truck is registered when stopped and inspected at a weigh station. (Attribute)

Qualitative and quantitative variables may be further subdivided:



Variable

Qualitative

Nominal

Ordinal

Quantitative

Discrete

Continuous

Nominal Variable: A qualitative variable that categorizes (or describes, or names) an element of a population.

Ordinal Variable: A qualitative variable that incorporates an ordered position, or ranking.

Discrete Variable: A quantitative variable that can assume a countable number of values. Intuitively, a discrete variable can assume values corresponding to isolated points along a line interval. That is, there is a gap between any two values.

Continuous Variable: A quantitative variable that can assume an uncountable number of values. Intuitively, a continuous variable can assume any value along a line interval, including every possible value between any two values.

Note:

1. In many cases, a discrete and continuous variable may be distinguished by determining whether the variables are related to a count or a measurement.
2. Discrete variables are usually associated with counting. If the variable cannot be further subdivided, it is a clue that you are probably dealing with a discrete variable.
3. Continuous variables are usually associated with measurements. The values of discrete variables are only limited by your ability to measure them.

Measure and Variability

- No matter what the response variable: there will always be **variability** in the data.
- One of the primary objectives of statistics: measuring and characterizing variability.
- Controlling (or reducing) variability in a manufacturing process: statistical process control.

Example: A supplier fills cans of soda marked 12 ounces. How much soda does each can really contain?

- It is very *unlikely* any one can contains exactly 12 ounces of soda.
- There is variability in any process.
- Some cans contain a little more than 12 ounces, and some cans contain a little less.
- On the average, there are 12 ounces in each can.
- The supplier hopes there is little variability in the process, that most cans contain *close* to 12 ounces of soda.

Data Collection

- First problem a statistician faces: how to obtain the data.
- It is important to obtain *good*, or *representative*, data.
- Inferences are made based on statistics obtained from the data.
- Inferences can only be as good as the data.

Biased Sampling Method: A sampling method that produces data which systematically differs from the sampled population. An **unbiased sampling method** is one that is not biased.

Sampling methods that often result in biased samples:

1. **Convenience sample**: sample selected from elements of a population that are easily accessible.
2. **Volunteer sample**: sample collected from those elements of the population which chose to contribute the needed information on their own initiative.

Process of data collection:

1. Define the objectives of the survey or experiment.

   Example: Estimate the average life of an electronic component.

2. Define the variable and population of interest.

   Example: Length of time for anesthesia to wear off after surgery.

3. Defining the data-collection and data-measuring schemes. This includes sampling procedures, sample size, and the data-measuring device (questionnaire, scale, ruler, etc.).

4. Determine the appropriate descriptive or inferential data-analysis techniques.

Methods used to collect data:

Experiment: The investigator controls or modifies the environment and observes the effect on the variable under study.

Survey: Data are obtained by sampling some of the population of interest. The investigator does not modify the environment.

Census: A 100% survey. Every element of the population is listed. Seldom used: difficult and time-consuming to compile, and expensive.

Sampling Frame: A list of the elements belonging to the population from which the sample will be drawn.

Note: It is important that the sampling frame be representative of the population.

Sample Design: The process of selecting sample elements from the sampling frame.

Note: There are many different types of sample designs. Usually they all fit into two categories: judgment samples and probability samples.

Judgment Samples: Samples that are selected on the basis of being “typical.”

Items are selected that are representative of the population. The validity of the results from a judgment sample reflects the soundness of the collector’s judgment.

Probability Samples: Samples in which the elements to be selected are drawn on the basis of probability. Each element in a population has a certain probability of being selected as part of the sample.

Random Samples: A sample selected in such a way that every element in the population has a equal probability of being chosen. Equivalently, all samples of size *n* have an equal chance of being selected. Random samples are obtained either by sampling with replacement from a finite population or by sampling without replacement from an infinite population.

Note:

1. Inherent in the concept of randomness: the next result (or occurrence) is not predictable.
2. Proper procedure for selecting a random sample: use a random number generator or a table of random numbers.

Example: An employer is interested in the time it takes each employee to commute to work each morning. A random sample of 35 employees will be selected and their commuting time will be recorded.

There are 2712 employees.

Each employee is numbered: 0001, 0002, 0003, etc. up to 2712.

Using four-digit random numbers, a sample is identified: 1315, 0987, 1125, etc.

Systematic Sample: A sample in which every *k*th item of the sampling frame is selected, starting from the first element which is randomly selected from the first *k* elements.

Note: The systematic technique is easy to execute. However, it has some inherent dangers when the sampling frame is repetitive or cyclical in nature. In these situations the results may not approximate a simple random sample.

Stratified Random Sample: A sample obtained by stratifying the sampling frame and then selecting a fixed number of items from each of the strata by means of a simple random sampling technique.

Proportional Sample (or Quota Sample): A sample obtained by stratifying the sampling frame and then selecting a number of items in proportion to the size of the strata (or by quota) from each strata by means of a simple random sampling technique.

Cluster Sample: A sample obtained by stratifying the sampling frame and then selecting some or all of the items from some of, but not all, the strata.

Numerical Presentation

A fundamental concept in summary statistics is that of a *central value* for a set of observations and the extent to which the central value characterizes the whole set of data. Measures of central value such as the mean or median must be coupled with measures of data dispersion (e.g., average distance from the mean) to indicate how well the central value characterizes the data as a whole.

To understand how well a central value characterizes a set of observations, let us consider the following two sets of data:

A: 30, 50, 70
B: 40, 50, 60

The mean of both two data sets is 50. But, the distance of the observations from the mean in data set A is larger than in the data set B. Thus, the mean of data set B is a better representation of the data set than is the case for set A.

Methods of Center Measurement

Center measurement is a summary measure of the overall level of a dataset

Commonly used methods are mean, median, mode, geometric mean etc.

Mean: Summing up all the observation and dividing by number of observations. Mean of 20, 30, 40 is (20+30+40)/3 = 30.

Notation : Let $x_1, x_{2,} ... x_n$ are n observations of a variable x. Then the mean of this variable,

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Methods of Center Measurement

Median: The middle value in an ordered sequence of observations. That is, to find the median we need to order the data set and then find the middle value. In case of an even number of observations the average of the two middle most values is the median. For example, to find the median of {9, 3, 6, 7, 5}, we first sort the data giving {3, 5, 6, 7, 9}, then choose the middle value 6. If the number of observations is even, e.g., {9, 3, 6, 7, 5, 2}, then the median is the average of the two middle values from the sorted sequence, in this case, (5 + 6) / 2 = 5.5.

Mode: The value that is observed most frequently. The mode is undefined for sequences in which no observation is repeated.

Mean or Median

The median is less sensitive to outliers (extreme scores) than the mean and thus a better measure than the mean for highly skewed distributions, e.g. family income. For example mean of 20, 30, 40, and 990 is (20+30+40+990)/4 =270. The median of these four observations is (30+40)/2 =35. Here 3 observations out of 4 lie between 20-40. So, the mean 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.

Methods of Variability Measurement

Variability (or dispersion) measures the amount of scatter in a dataset.

Commonly used methods: *range*, *variance*, *standard deviation*, *interquartile range*, *coefficient of variation etc*.

Range: The difference between the largest and the smallest observations. The range of 10, 5, 2, 100 is (100-2)=98. It’s a crude measure of variability.

Methods of Variability Measurement

Variance: The variance of a set of observations is the average of the squares of the deviations of the observations from their mean. In symbols, the variance of the n observations x_1, x_2,…x_n is

$$S^2 = \frac{(x_1 - \overline{x})^2 + ... + (x_n - \overline{x})^2}{n-1}$$

Variance of 5, 7, 3? Mean is (5+7+3)/3 = 5 and the variance is

$$\frac{(5-5)^2 + (3-5)^2 + (7-5)^2}{3-1} = 4$$

Standard Deviation: Square root of the variance. The standard deviation of the above example is 2.

Methods of Variability Measurement

Quartiles: Data can be divided into four regions that cover the total range of observed values. Cut points for these regions are known as quartiles.

In notations, quartiles of a data is the ((n+1)/4)qth observation of the data, where q is the desired quartile and n is the number of observations of data.

The first quartile (Q1) is the first 25% of the data. The second quartile (Q2) is between the 25th and 50th percentage points in the data. The upper bound of Q2 is the median. The third quartile (Q3) is the 25% of the data lying between the median and the 75% cut point in the data.

Q1 is the median of the first half of the ordered observations and Q3 is the median of the second half of the ordered observations.

Methods of Variability Measurement

In the following example Q1= ((15+1)/4)1 =4th observation of the data. The 4th observation is 11. So Q1 is of this data is 11.

An example with 15 numbers

```
        3 6 7 11 13 22 30 40 44 50 52 61 68 80 94
              Q1        Q2        Q3
```

The first quartile is Q1=11. The second quartile is Q2=40 (This is also the Median.) The third quartile is Q3=61.

Inter-quartile Range: Difference between Q3 and Q1. Inter-quartile range of the previous example is 61- 40=21. The middle half of the ordered data lie between 40 and 61.

Deciles and Percentiles

Deciles: If data is ordered and divided into 10 parts, then cut points are called Deciles

Percentiles: If data is ordered and divided into 100 parts, then cut points are called Percentiles. 25th percentile is the Q1, 50th percentile is the Median (Q2) and the 75th percentile of the data is Q3.

In notations, percentiles of a data is the ((n+1)/100)p th observation of the data, where p is the desired percentile and n is the number of observations of data.

Coefficient of Variation: The standard deviation of data divided by it's mean. It is usually expressed in percent.

Coefficient of Variation = $\frac{\sigma}{\bar{x}} \times 100$

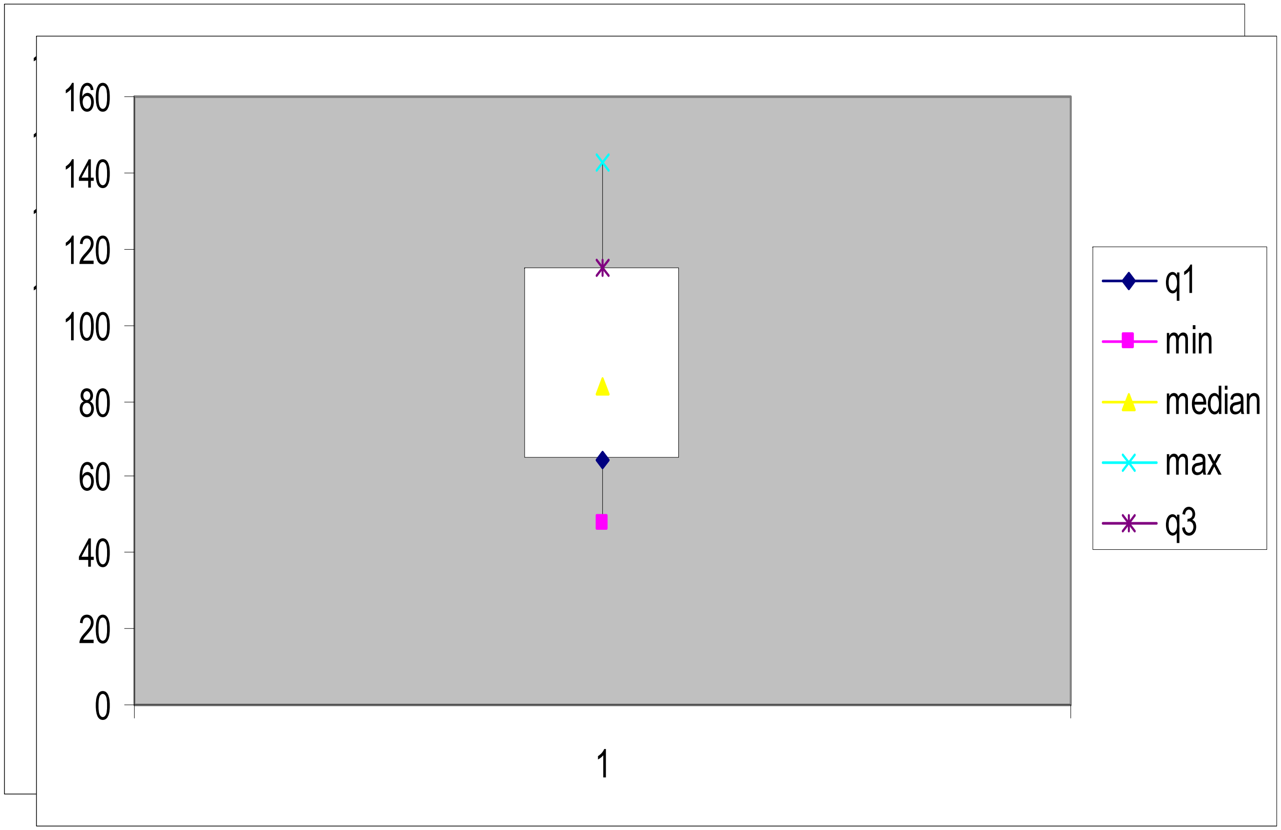
Five Number Summary

Five Number Summary: The five number summary of a distribution consists of the smallest (Minimum) observation, the first quartile (Q1),
The median(Q2), the third quartile, and the largest (Maximum) observation written in order from smallest to largest.

Box Plot: A box plot is a graph of the five number summary. The central box spans the quartiles. A line within the box marks the median. Lines extending above and below the box mark the smallest and the largest observations (i.e., the range). Outlying samples may be additionally plotted outside the range.

Boxplot

Distribution of Age in Month



Choosing a Summary

The five number summary is usually better than the mean and standard deviation for describing a skewed distribution or a distribution with extreme outliers. The mean and standard deviation are reasonable for symmetric distributions that are free of outliers.

In real life we can’t always expect symmetry of the data. It’s a common practice to include number of observations (n), mean, median, standard deviation, and range as common for data summarization purpose. We can include other summary statistics like Q1, Q3, Coefficient of variation if it is considered to be important for describing data.

Shape of Data

- Shape of data is measured by
  - Skewness
  - Kurtosis

Skewness

- Measures asymmetry of data
  - Positive or right skewed: Longer right tail
  - Negative or left skewed: Longer left tail

Let $x_1, x_2, ... x_n$ be n observations. Then,

$$\text{Skewness} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^{3/2}}$$

Kurtosis

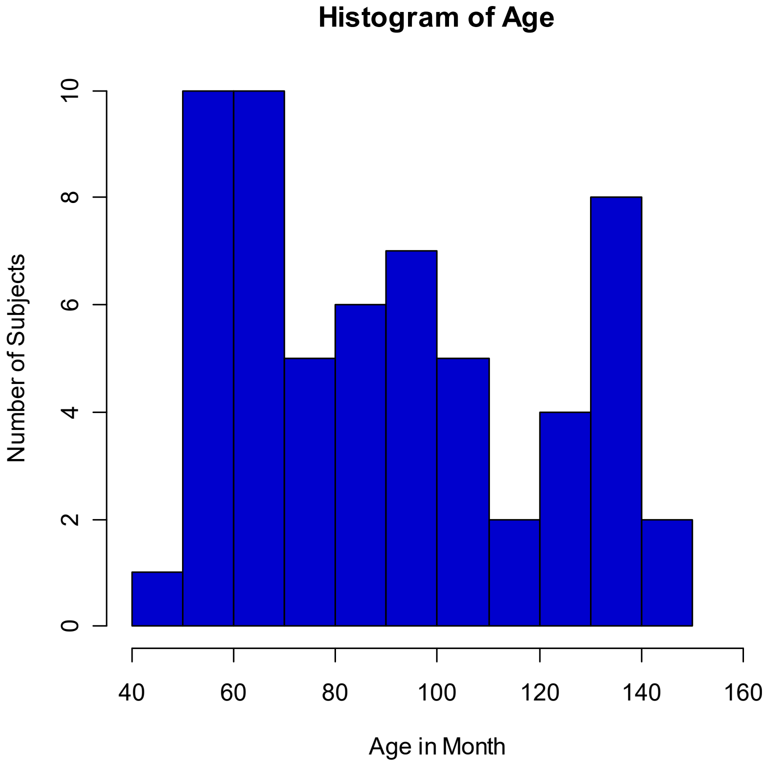
- Measures peakedness of the distribution of data. The kurtosis of normal distribution is 0.

$$\text{Let } x_1, x_2, \ldots x_n \text{ be } n \text{ observations. Then,}$$

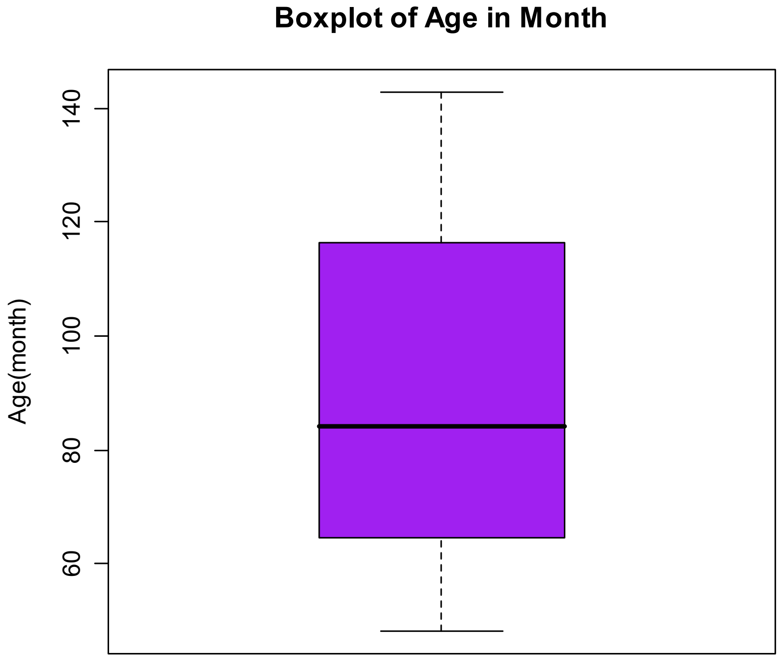
$$\text{Kurtosis} = \frac{n\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3$$

Summary of the Variable 'Age' in the given data set

Mean	90.41666667
Standard Error	3.902649518
Median	84
Mode	84
Standard Deviation	30.22979318
Sample Variance	913.8403955
Kurtosis	-1.183899591
Skewness	0.389872725
Range	95
Minimum	48
Maximum	143
Sum	5425
Count	60



Summary of the Variable ‘Age’ in the given data set



Microsoft Excel

A Spreadsheet Application. It features calculation, graphing tools, pivot tables and a macro programming language called VBA (Visual Basic for Applications).

There are many versions of MS-Excel. Excel XP, Excel 2003, Excel 2007 are capable of performing a number of statistical analyses.

Starting MS Excel: Double click on the Microsoft Excel icon on the desktop or Click on Start --> Programs --> Microsoft Excel.

Worksheet: Consists of a multiple grid of cells with numbered rows down the page and alphabetically-tilted columns across the page. Each cell is referenced by its coordinates. For example, A3 is used to refer to the cell in column A and row 3. B10:B20 is used to refer to the range of cells in column B and rows 10 through 20.

Microsoft Excel

Opening a document: File → Open (From a existing workbook). Change the directory area or drive to look for file in other locations.

Creating a new workbook: File→New→Blank Document

Saving a File: File→Save

Selecting more than one cell: Click on a cell e.g. A1), then hold the Shift key and click on another (e.g. D4) to select cells between and A1 and D4 or Click on a cell and drag the mouse across the desired range.

Creating Formulas: 1. Click the cell that you want to enter the formula, 2. Type = (an equal sign), 3. Click the Function Button, 4. Select the formula you want and step through the on-screen instructions.

Microsoft Excel

Entering Date and Time: Dates are stored as MM/DD/YYYY. No need to enter in that format. For example, Excel will recognize jan 9 or jan-9 as 1/9/2007 and jan 9, 1999 as 1/9/1999. To enter today's date, press Ctrl and ; together. Use a or p to indicate am or pm. For example, 8:30 p is interpreted as 8:30 pm. To enter current time, press Ctrl and : together.

Copy and Paste all cells in a Sheet: Ctrl+A for selecting, Ctrl +C for copying and Ctrl+V for Pasting.

Sorting: Data → Sort→ Sort By …

Descriptive Statistics and other Statistical methods: Tools→Data Analysis→ Statistical method. If Data Analysis is not available then click on Tools→ Add-Ins and then select Analysis ToolPack and Analysis toolPack-Vba

Microsoft Excel

Statistical and Mathematical Function: Start with ‘=‘ sign and then select function from function wizard f_x.

Inserting a Chart: Click on Chart Wizard (or Insert→Chart), select chart, give, Input data range, Update the Chart options, and Select output range/ Worksheet.

Importing Data in Excel: File →open →FileType →Click on File→ Choose Option (Delimited/Fixed Width) →Choose Options (Tab/ Semicolon/ Comma/ Space/ Other) → Finish.

Limitations: Excel uses algorithms that are vulnerable to rounding and truncation errors and may produce inaccurate results in extreme cases.

Statistical Inference

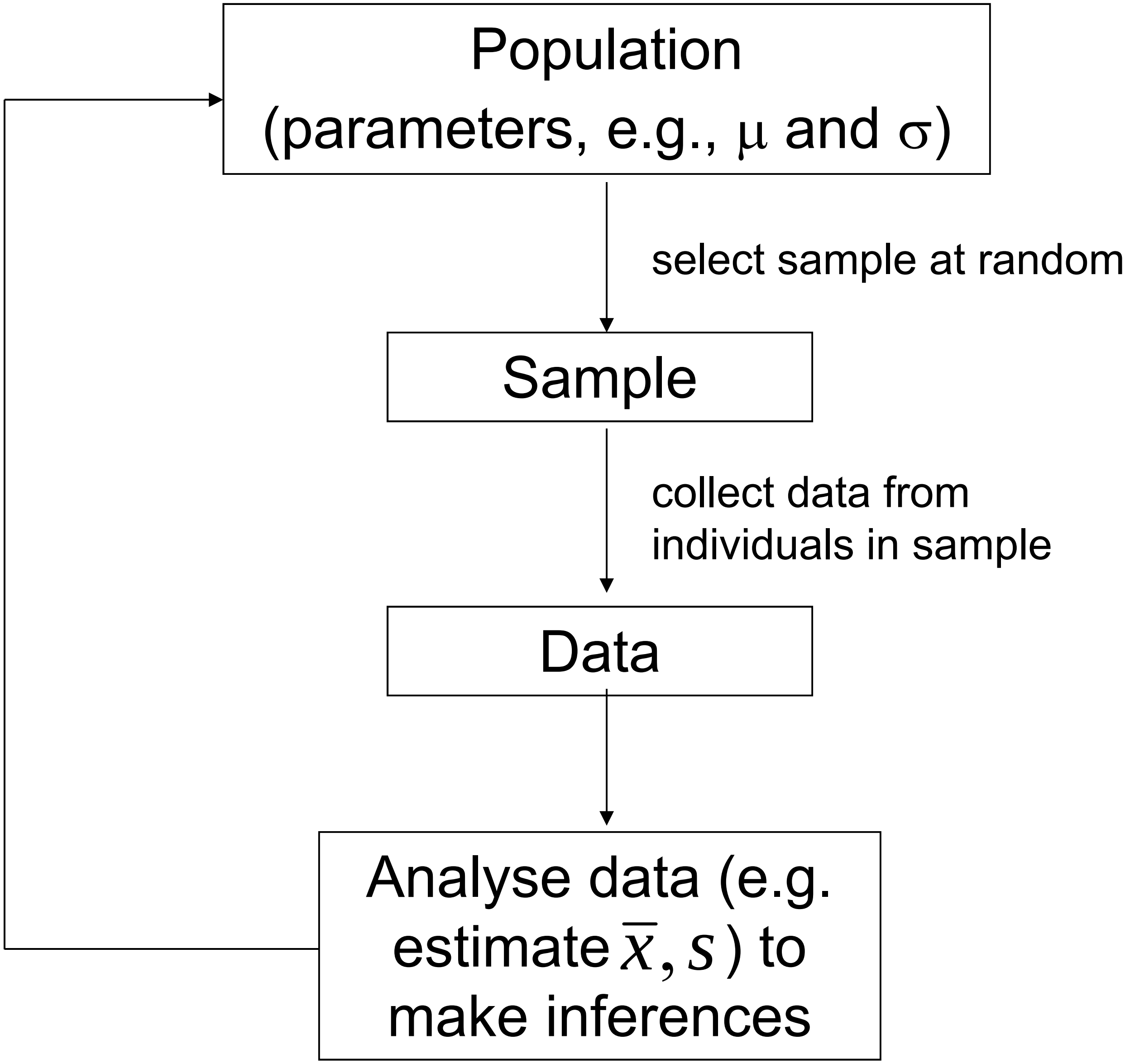
- **Statistical Inference** – the process of drawing conclusions about a population based on information in a sample
- Unlikely to see this published…

  > "In our study of a new antihypertensive drug we found an effective 10% reduction in blood pressure for those on the new therapy. However, the effects seen are only specific to the subjects in our study. We cannot say this drug will work for hypertensive people in general".

Describing a population

- Characteristics of a population, e.g. the population mean μ and the population standard deviation σ are never known exactly
- Sample characteristics, e.g. $\overline{x}$ and s are **estimates** of population characteristics μ and σ
- A sample characteristic, e.g. $\overline{x}$,is called a **statistic** and a population characteristic, e.g. μ is called a **parameter**

Statistical Inference



Population
(parameters, e.g., μ and σ)

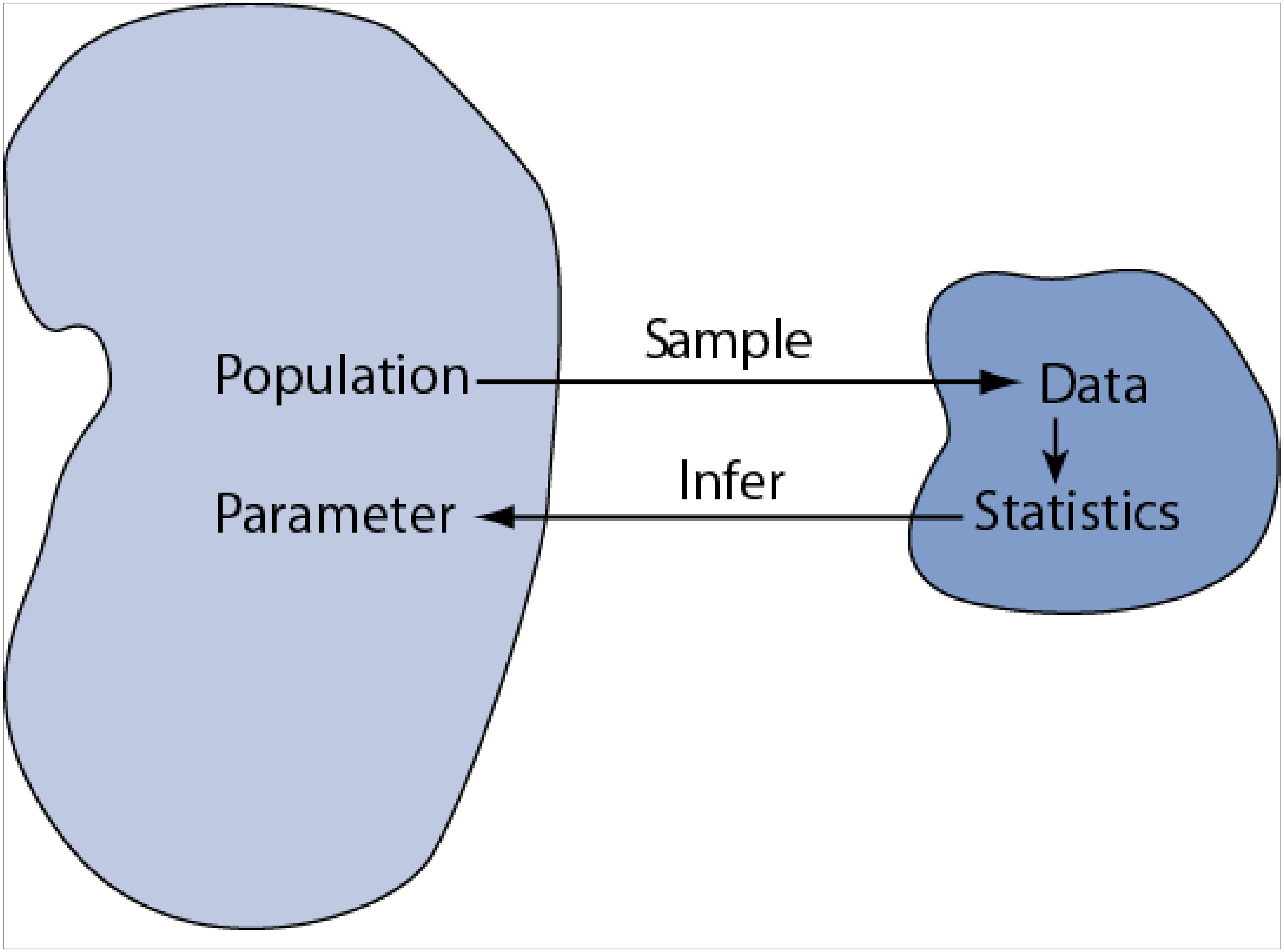
select sample at random

Sample

collect data from
individuals in sample

Data

Analyse data (e.g.
estimate $\overline{x}, s$) to
make inferences


Sample
Population
Data
Infer
Parameter
Statistics

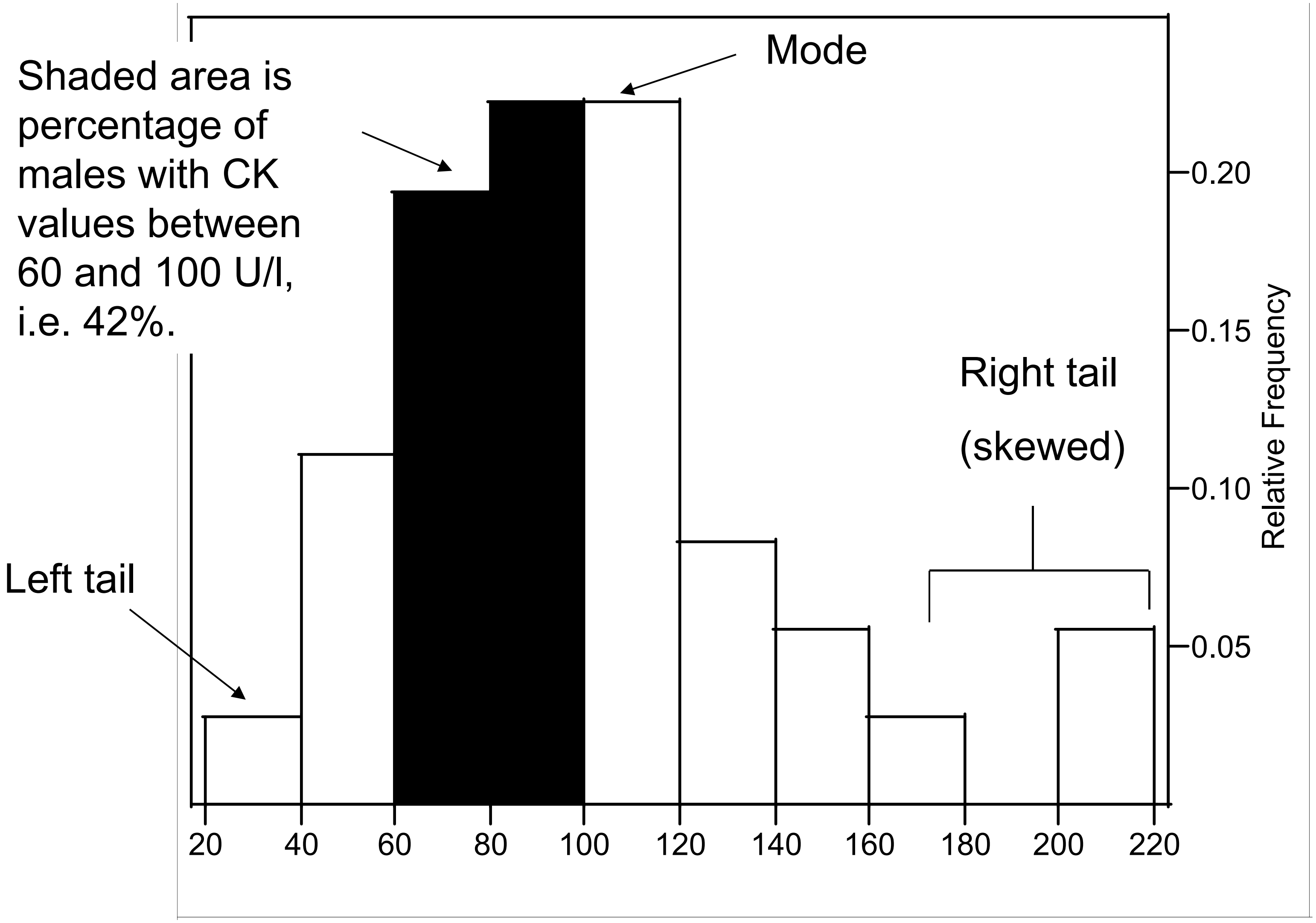
Distributions

- As sample size increases, histogram class widths can be narrowed such that the histogram eventually becomes a smooth curve
- The population histogram of a random variable is referred to as the **distribution** of the random variable, i.e. it shows how the population is distributed across the number line

Density curve

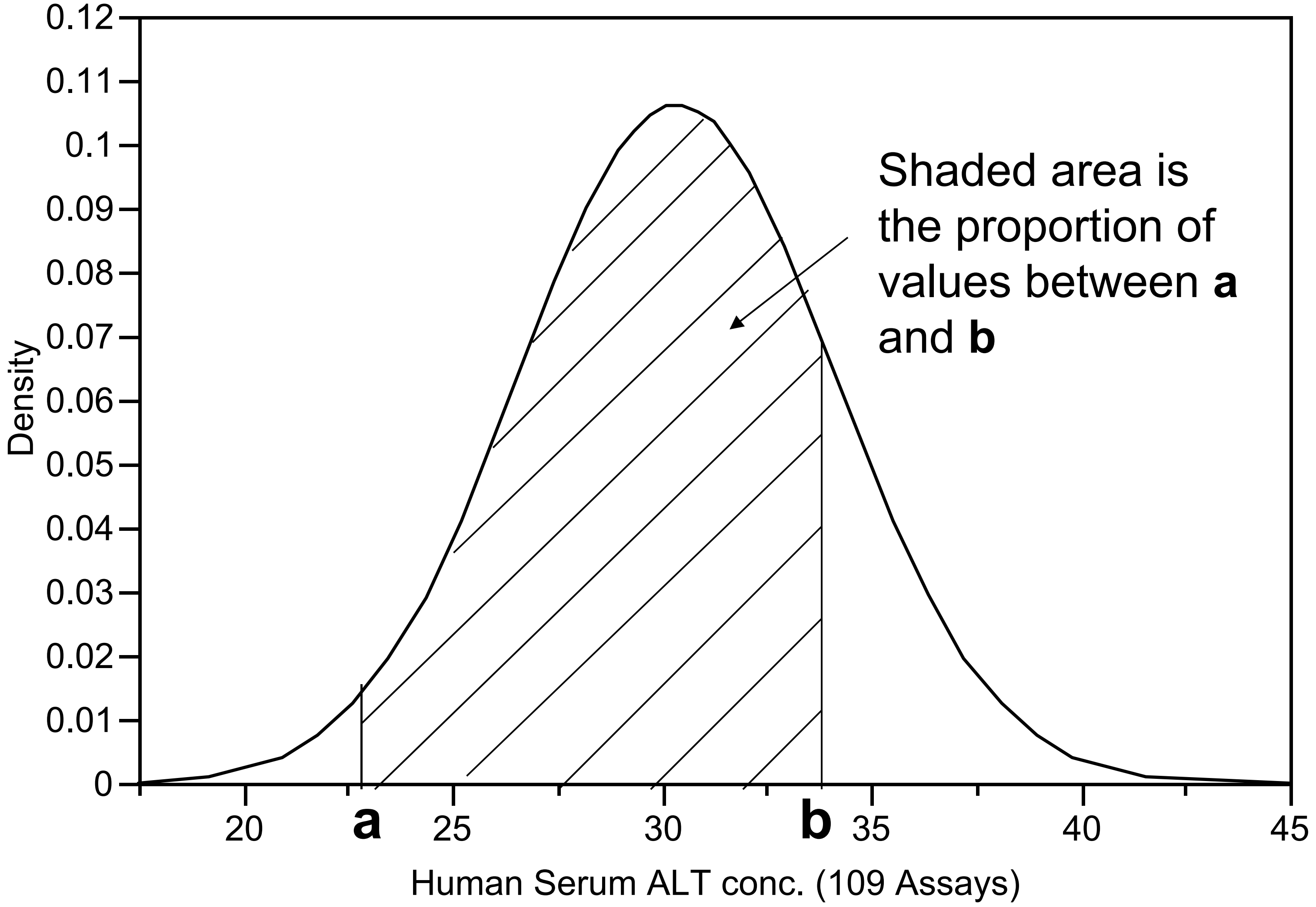
- A smooth curve representing a relative frequency distribution is called a **density** curve
- The area under the density curve between any two points **a** and **b** is the proportion of values between **a** and **b**.

Sample Relative Frequency Distribution



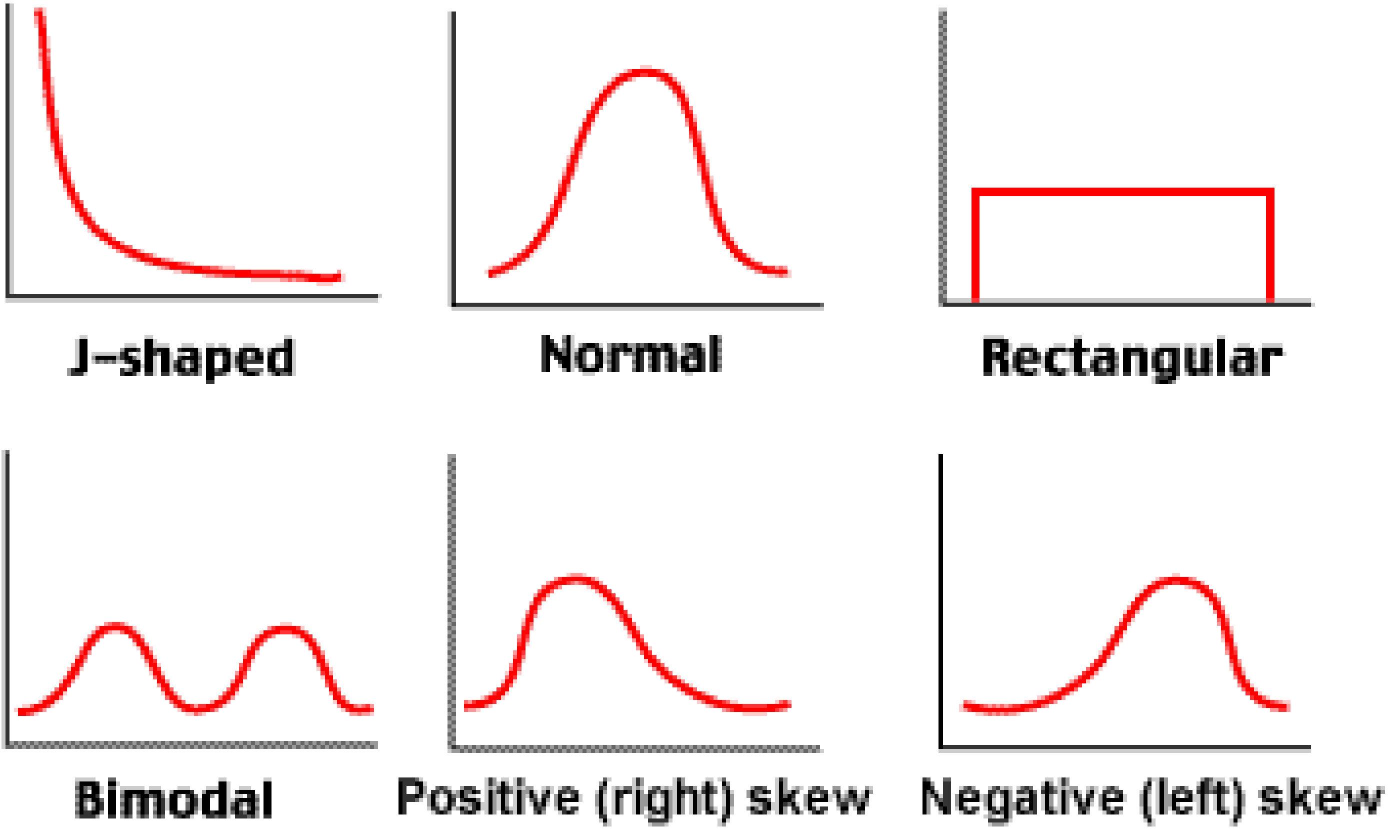
Shaded area is percentage of males with CK values between 60 and 100 U/l, i.e. 42%.

Population Relative Frequency Distribution (Density)



Shaded area is the proportion of values between **a** and **b**

Distribution Shapes



J-shaped

Normal

Rectangular

Bimodal

Positive (right) skew

Negative (left) skew

The Normal Distribution

- The **Normal distribution** is considered to be the most important distribution in statistics

- It occurs in “nature” from processes consisting of a very large number of elements acting in an **additive** manner

- However, it would be very difficult to use this argument to assume normality of your data
  - Later, we will see exactly why the Normal is so important in statistics

Normal Distribution

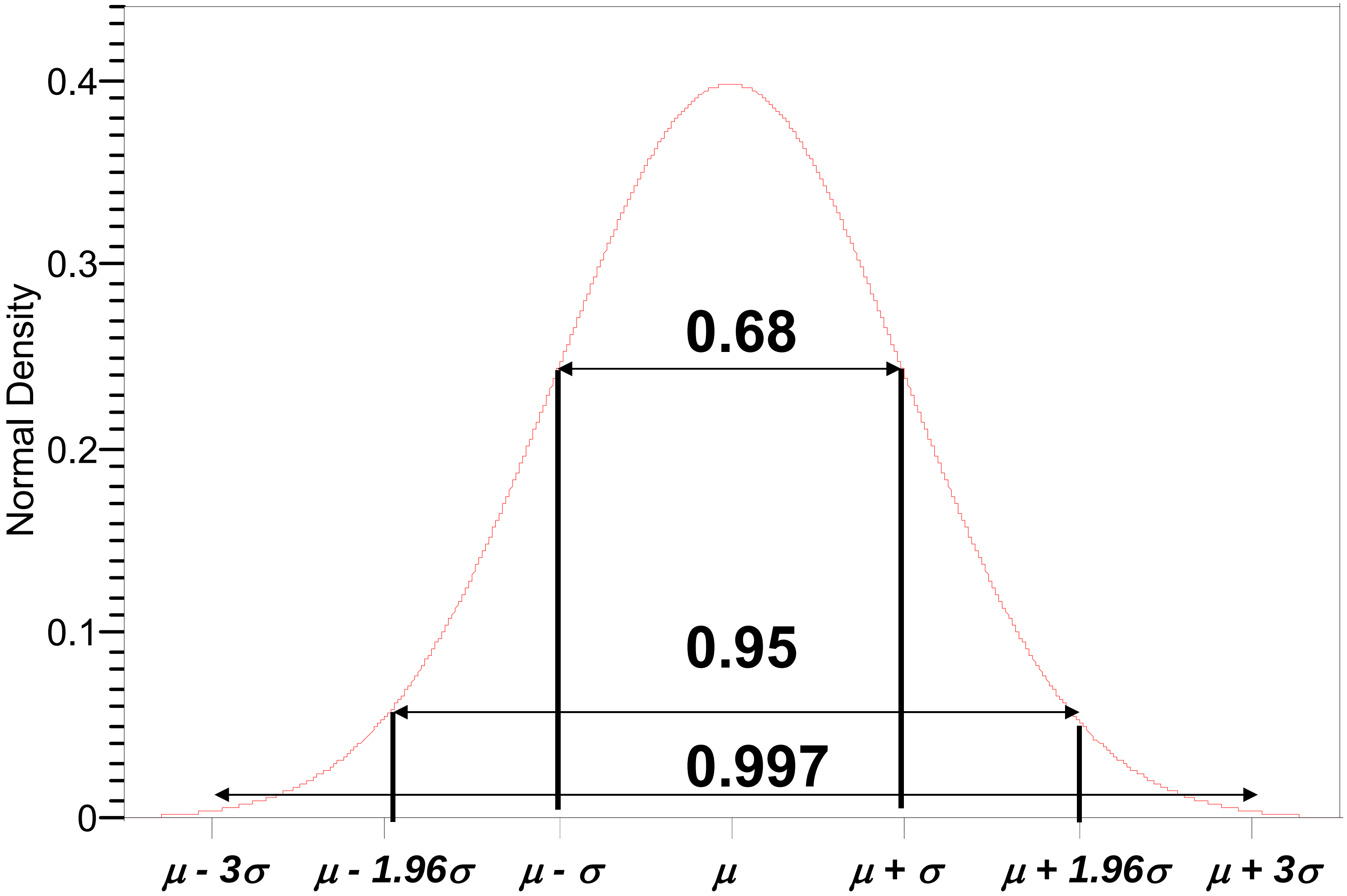
- Closely related is the **log-normal** distribution, based on factors acting **multiplicatively**. This distribution is right-skewed.
  - Note: The logarithm of the data is thus normal.

- The log-transformation of data is very common, mostly to eliminate skew in data

Properties of the Normal Distribution

- The Normal distribution has a symmetric bell-shaped density curve
- Characterised by two parameters, i.e. the mean μ, and standard deviation σ
  - 68% of data lie within 1σ of the mean μ
  - 95% of data lie within 2σ of the mean μ
  - 99.7% of data lie within 3σ of the mean μ

Normal curve



Probability Density Functions…

- Unlike a discrete random variable which we studied in Chapter 7, a ***continuous random variable*** is one that can assume an **uncountable** number of values.
- ➔ We cannot list the possible values because there is an infinite number of them.
- ➔ Because there is an infinite number of values, the probability of each individual value is virtually 0.

Point Probabilities are Zero

➔ Because there is an infinite number of values, the probability of each individual value is virtually 0.

Thus, we can determine the probability of a ***range of values*** only.

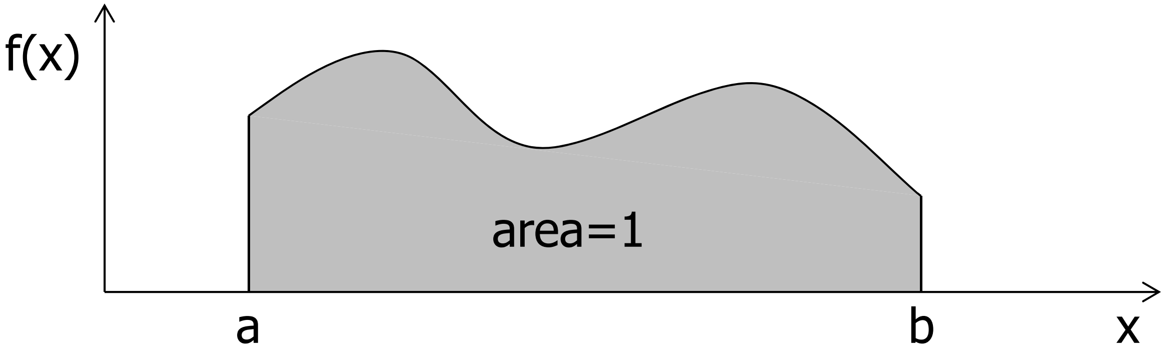
- E.g. with a **discrete** random variable like tossing a die, it is meaningful to talk about $P(X=5)$, say.
- In a **continuous** setting (e.g. with time as a random variable), the probability the random variable of interest, say task length, takes **exactly** 5 minutes is infinitesimally small, hence $P(X=5) = 0$.
  - ***It is meaningful to talk about $P(X \leq 5)$.***

Probability Density Function…

- A function f(x) is called a ***probability density function*** (over the range **a ≤ x ≤ b** if it meets the following requirements:

1) f(x) ≥ 0 for all **x** between **a** and **b**, and

2) The total area under the curve between **a** and **b** is 1.0

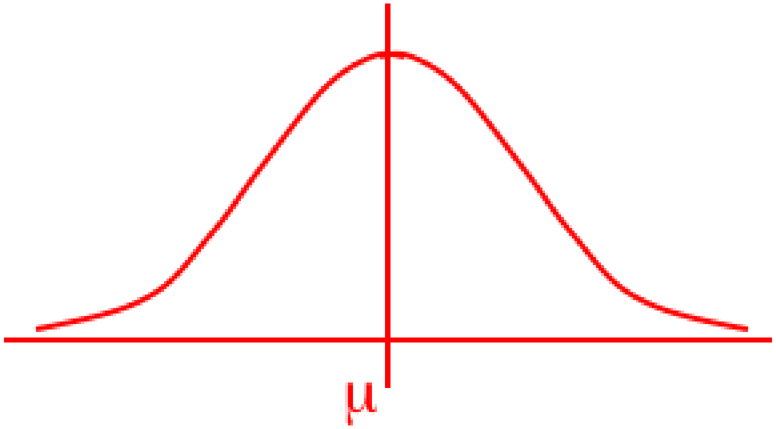


The Normal Distribution…

- The ***normal distribution*** is the most important of all probability distributions. The probability density function of a ***normal random variable*** is given by:

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

- It looks like this:
- Bell shaped,
- Symmetrical around the mean …



μ

The Normal Distribution…

The normal distribution is fully defined by two parameters:
its standard deviation and mean

- **Important things to note:**

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

The normal distribution is bell shaped and
symmetrical about the mean

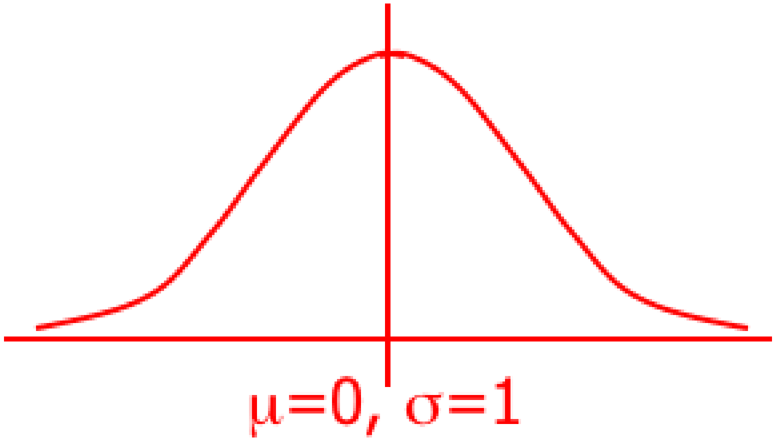
Unlike the range of the uniform distribution (a ≤ x ≤ b)
Normal distributions ***range from minus infinity to plus infinity***

Standard Normal Distribution…

- A normal distribution whose mean is zero and standard deviation is one is called the ***standard normal distribution***.

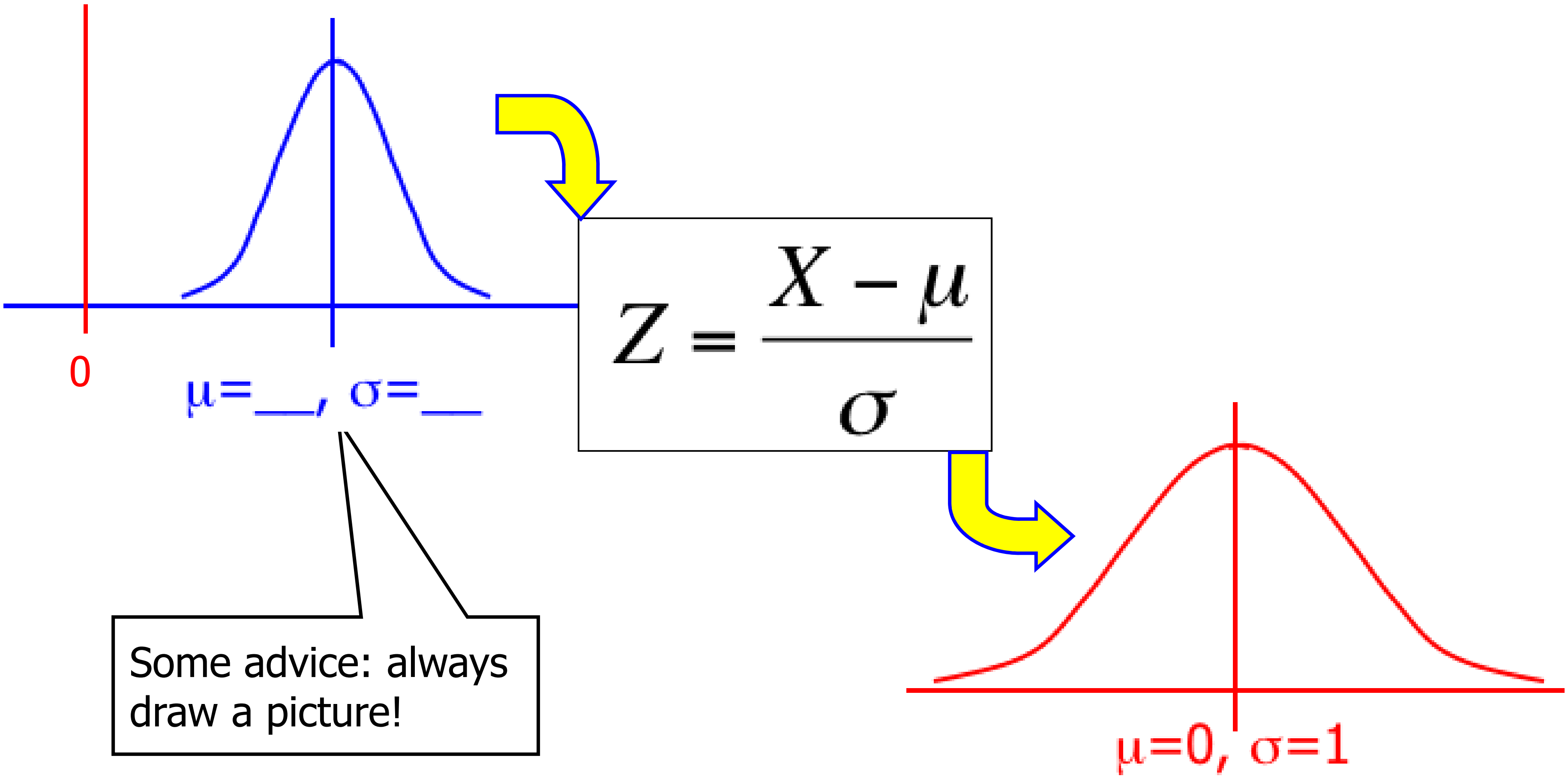
$$f(x) = \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-0}{1}\right)^2} \qquad -\infty < x < \infty$$

- As we shall see shortly, any normal distribution can be ***converted*** to a standard normal distribution with simple algebra. This makes calculations much easier.



Calculating Normal Probabilities…

- We can use the following function to convert any normal random variable to a **standard** normal random variable…

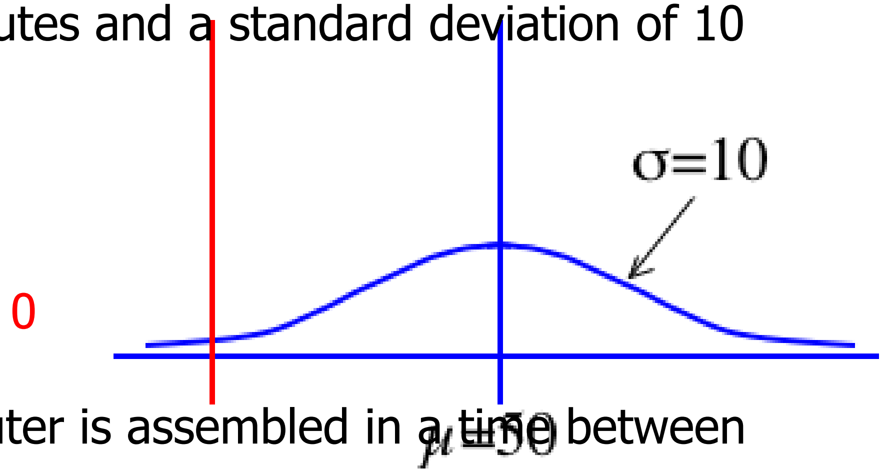


$$Z = \frac{X - \mu}{\sigma}$$

Some advice: always draw a picture!

Calculating Normal Probabilities…

- Example: The time required to build a computer is ***normally distributed*** with a mean of 50 minutes and a standard deviation of 10 minutes:

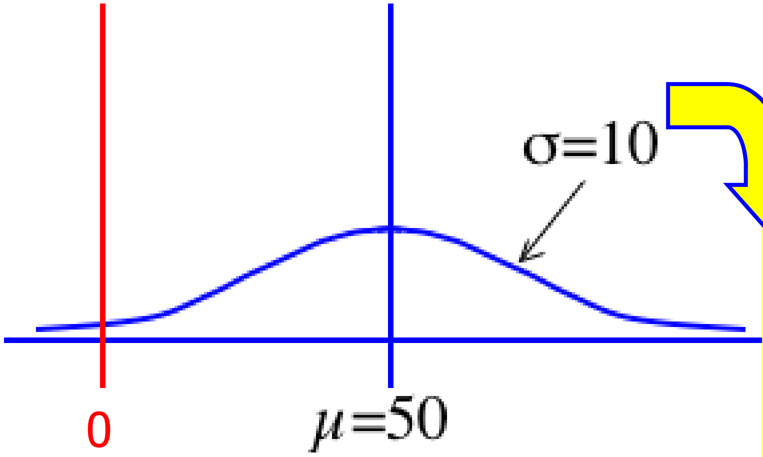


- What is the probability that a computer is assembled in a time between 45 and 60 minutes?

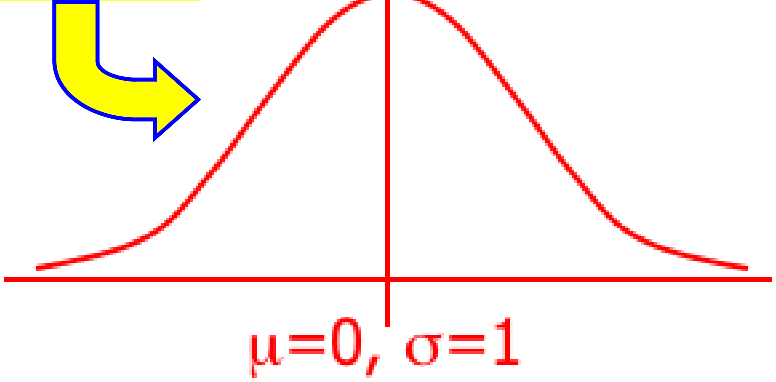
- Algebraically speaking, what is **P(45 < X < 60)** ?

Calculating Normal Probabilities…

…mean of 50 minutes and a standard deviation of 10 minutes…



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



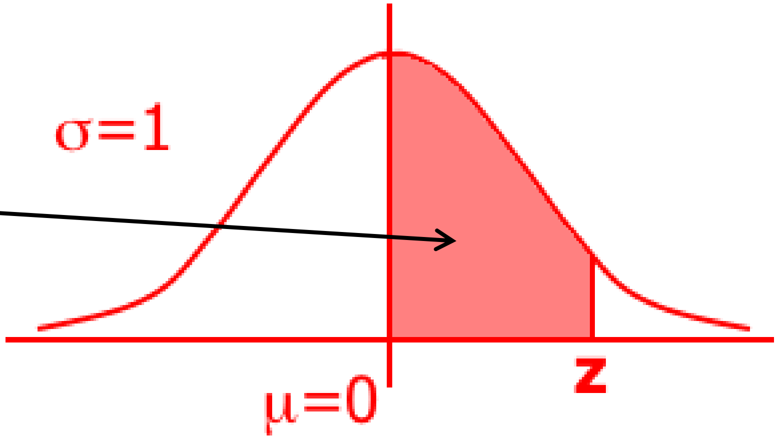
$$P(45 < X < 60) =$$

$$P\left(\frac{45-50}{10} < \frac{X-\mu}{\sigma} < \frac{60-50}{10}\right) =$$

$$P(-.5 < Z < 1)$$

Calculating Normal Probabilities…

- We can use Table 3 in
- Appendix B to look-up
- probabilities $P(0 < Z < z)$



- We can break up $P(-.5 < Z < 1)$ into:
- $P(-.5 < Z < 0) + P(0 < Z < 1)$

- The distribution is ***symmetric*** around zero, so we have:
- $P(-.5 < Z < 0) = P(0 < Z < .5)$
- Hence: $P(-.5 < Z < 1) = P(0 < Z < .5) + P(0 < Z < 1)$

Calculating Normal Probabilities…

- How to use Table …

This table gives probabilities **P(0 < Z < z)**

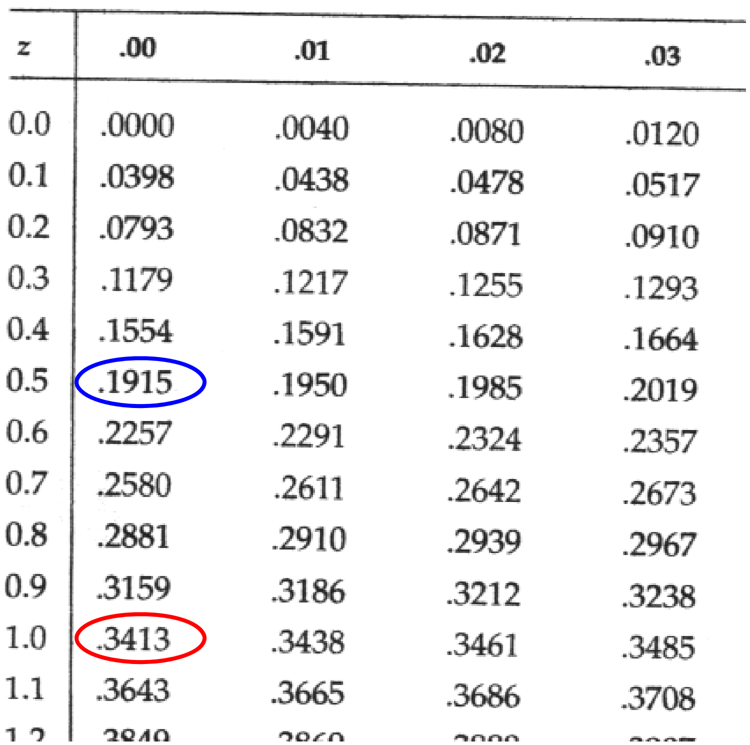
First column = integer + first decimal

Top row = second decimal place

P(0 < Z < 0.5)

P(0 < Z < 1)

P(–.5 < Z < 1) = .1915 + .3414 = .5328

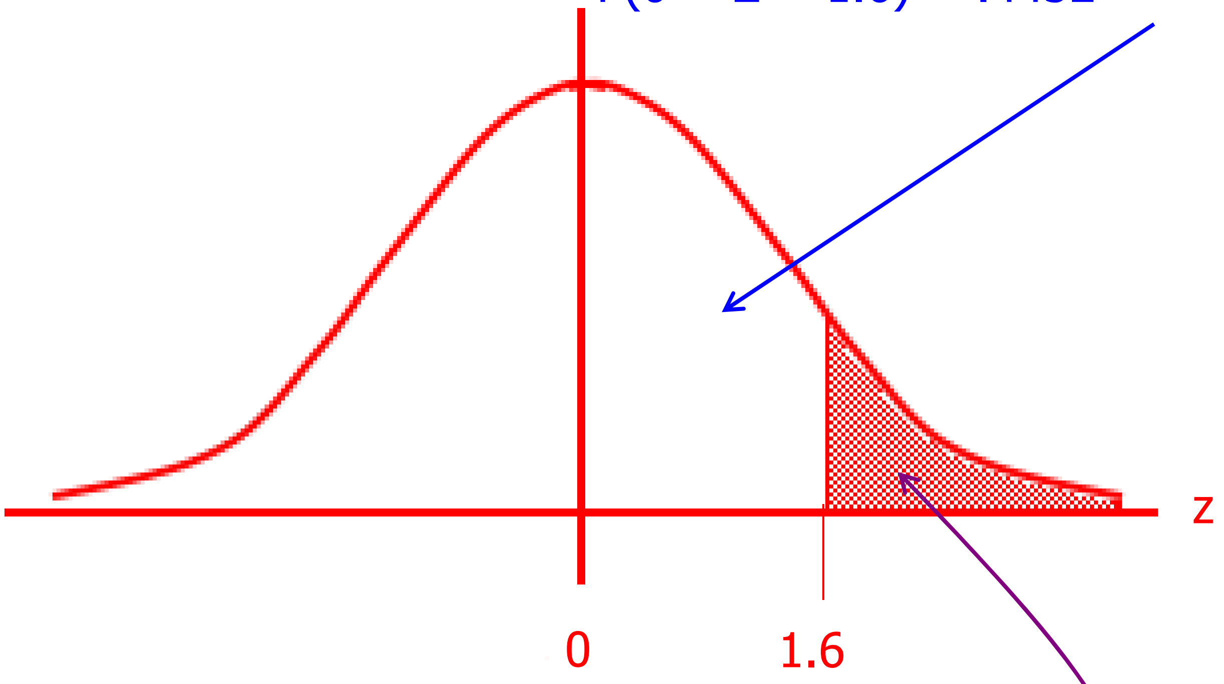


z	.00	.01	.02	.03
0.0	.0000	.0040	.0080	.0120
0.1	.0398	.0438	.0478	.0517
0.2	.0793	.0832	.0871	.0910
0.3	.1179	.1217	.1255	.1293
0.4	.1554	.1591	.1628	.1664
0.5	.1915	.1950	.1985	.2019
0.6	.2257	.2291	.2324	.2357
0.7	.2580	.2611	.2642	.2673
0.8	.2881	.2910	.2939	.2967
0.9	.3159	.3186	.3212	.3238
1.0	.3413	.3438	.3461	.3485
1.1	.3643	.3665	.3686	.3708
1.2	.3849	.3869	.3888	.3907

Using the Normal Table

- What is **P(Z > 1.6)** ?

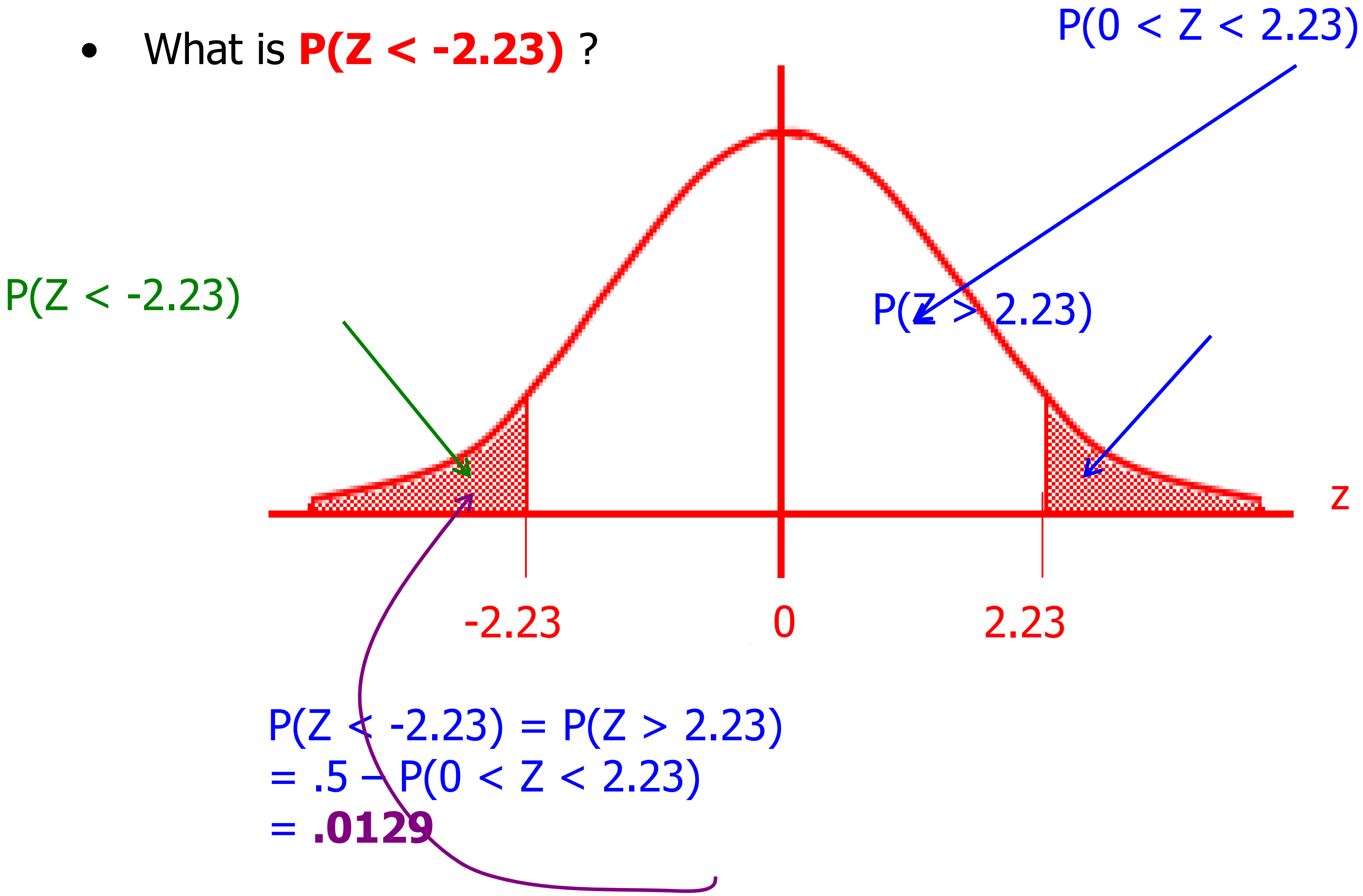
P(0 < Z < 1.6) = .4452



P(Z > 1.6) = .5 – P(0 < Z < 1.6)
= .5 – .4452
= **.0548**

Using the Normal Table (Table 3)…

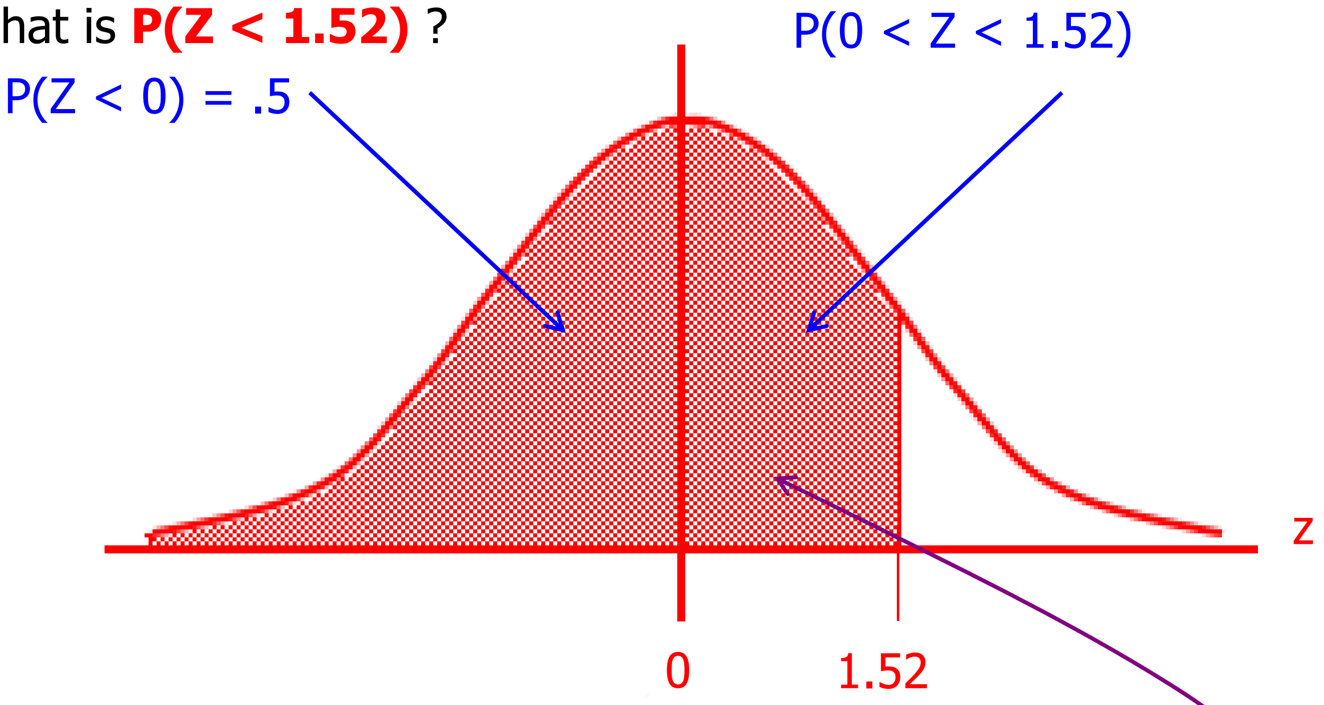
- What is **P(Z < -2.23)** ?



P(Z < -2.23) = P(Z > 2.23)
= .5 – P(0 < Z < 2.23)
= **.0129**

Using the Normal Table (Table 3)…

- What is **P(Z < 1.52)** ?

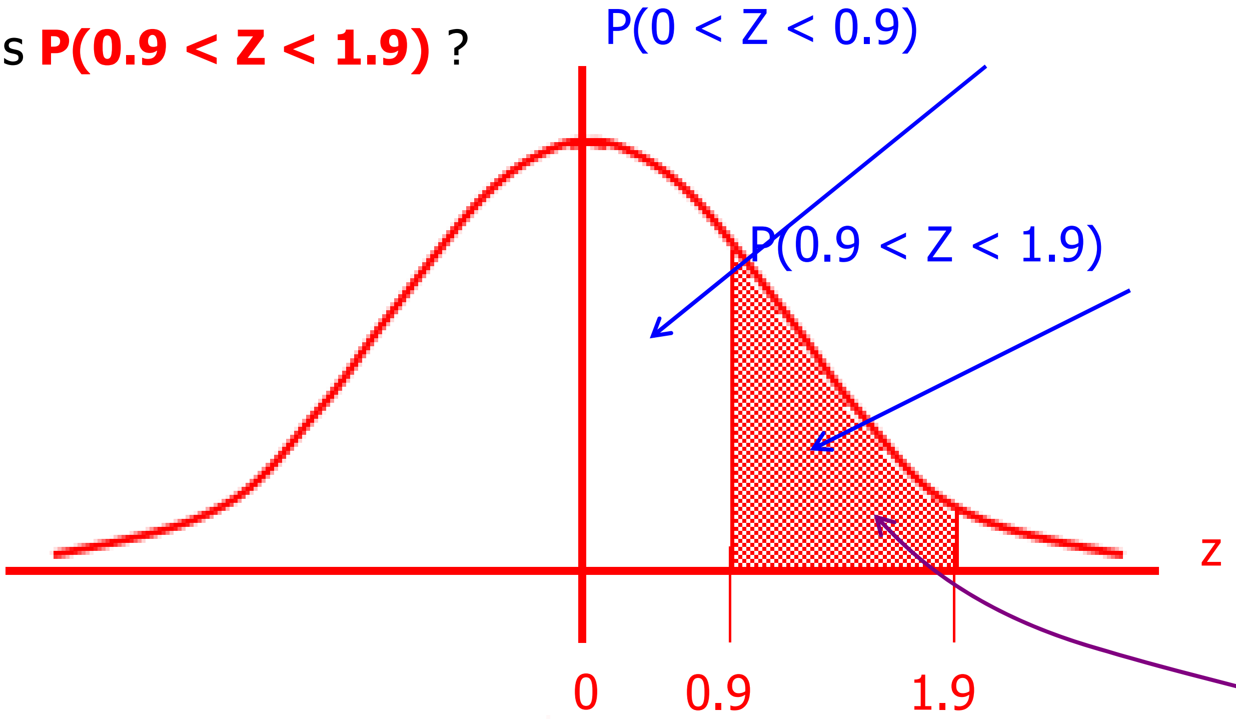

P(Z < 0) = .5

P(0 < Z < 1.52)

P(Z < 1.52) = .5 + P(0 < Z < 1.52)
= .5 + .4357
= **.9357**

Using the Normal Table (Table 3)…

- What is **P(0.9 < Z < 1.9)** ?



$P(0.9 < Z < 1.9) = P(0 < Z < 1.9) – P(0 < Z < 0.9)$
$=.4713 – .3159$
$=$ **.1554**

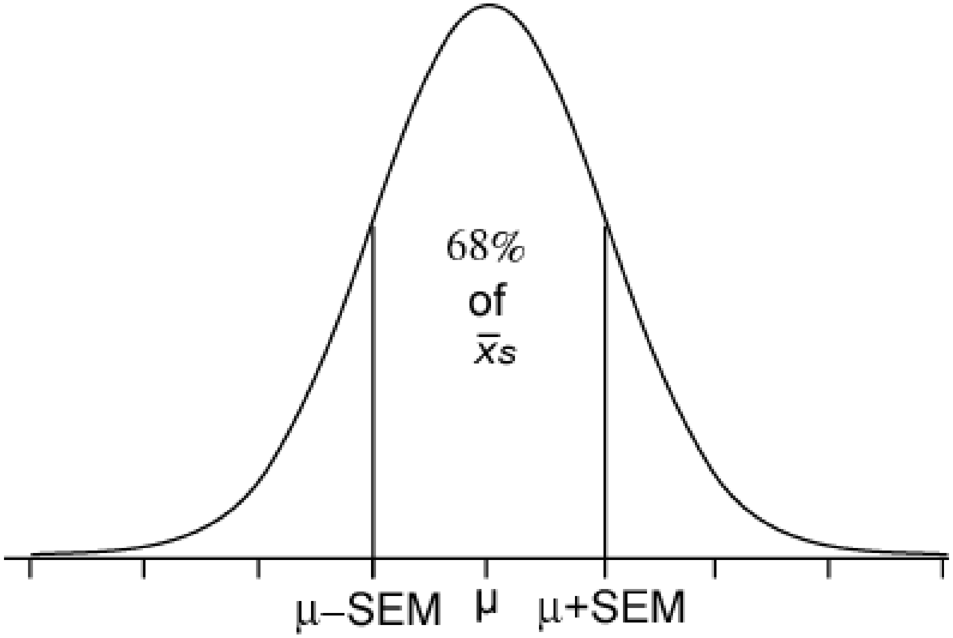
Sampling Distributions of a Mean

The **sampling distributions of a mean (SDM)** describes the behavior of a sampling mean

SE=standard error

$$\bar{x} \sim N\left(\mu, SE_{\bar{x}}\right)$$

$$\text{where } SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$



Standard Normal distribution

- If X is a Normally distributed random variable with mean = μ and standard deviation = σ, then X can be converted to a Standard Normal random variable Z using:

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

Standard Normal distribution (contd.)

- Z has mean = 0 and standard deviation = 1
- Using this transformation, we can calculate areas under **any** normal distribution

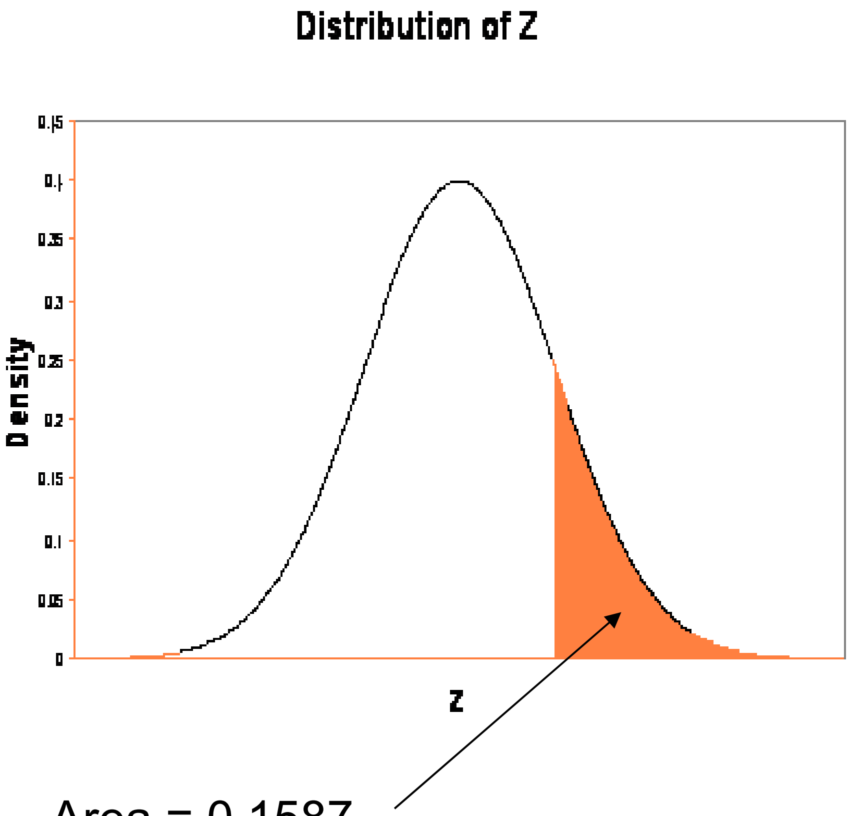
Example

- Assume the distribution of blood pressure is Normally distributed with μ = 80 mm and σ = 10 mm
- What percentage of people have blood pressure greater than 90?
- Z score transformation:

  Z=(90 - 80) /10 = 1

Example (contd.)

- The percentage greater than 90 is equivalent to the area under the Standard Normal curve greater then Z = 1.
- From tables of the Standard Normal distribution, the area to the right of Z=1 is 0.1587 (or 15.87%)



Area = 0.1587

Central Limit Theorem (CLT)

- Suppose you take any random sample from a population with mean μ and variance σ^2
- Then, for large sample sizes, the CLT states that the distribution of sample means is the Normal Distribution, with mean μ and variance σ^2/n (i.e. standard deviation is $\sigma/\sqrt{n}$)
- If the original data is Normal then the sample means are Normal, irrespective of sample size

What is it really saying?

(1) It gives a relationship between the sample mean and population mean

- This gives us a framework to extrapolate our sample results to the population *(statistical inference);*

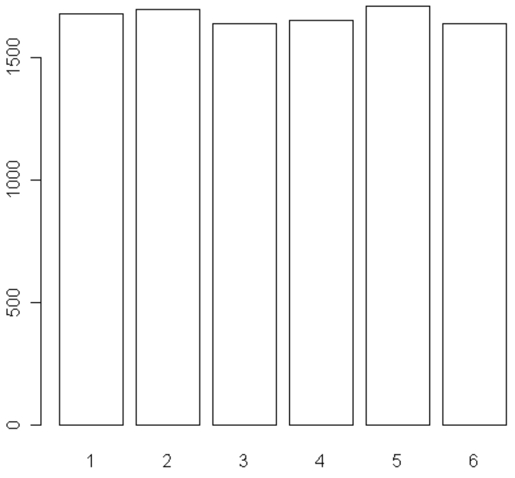
(2) It doesn’t matter what the distribution of the original data is, the sample mean will always be Normally distributed when n is large.

- This why the Normal is so central to statistics

Example: Toss 1, 2 or 10 dice (10,000 times)

Toss 1 dice

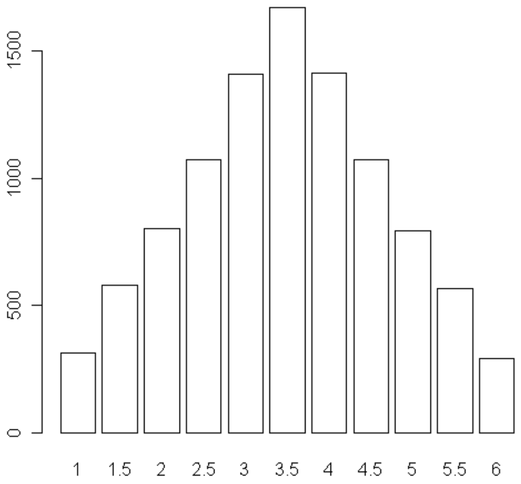
Histogram of data



Distribution of data is far from Normal

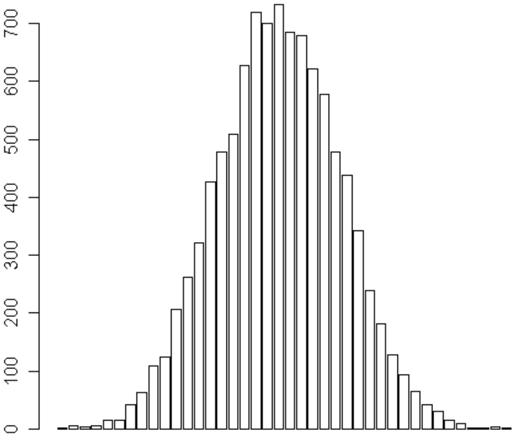
Toss 2 dice

Histogram of averages



Toss 10 dice

Histogram of averages



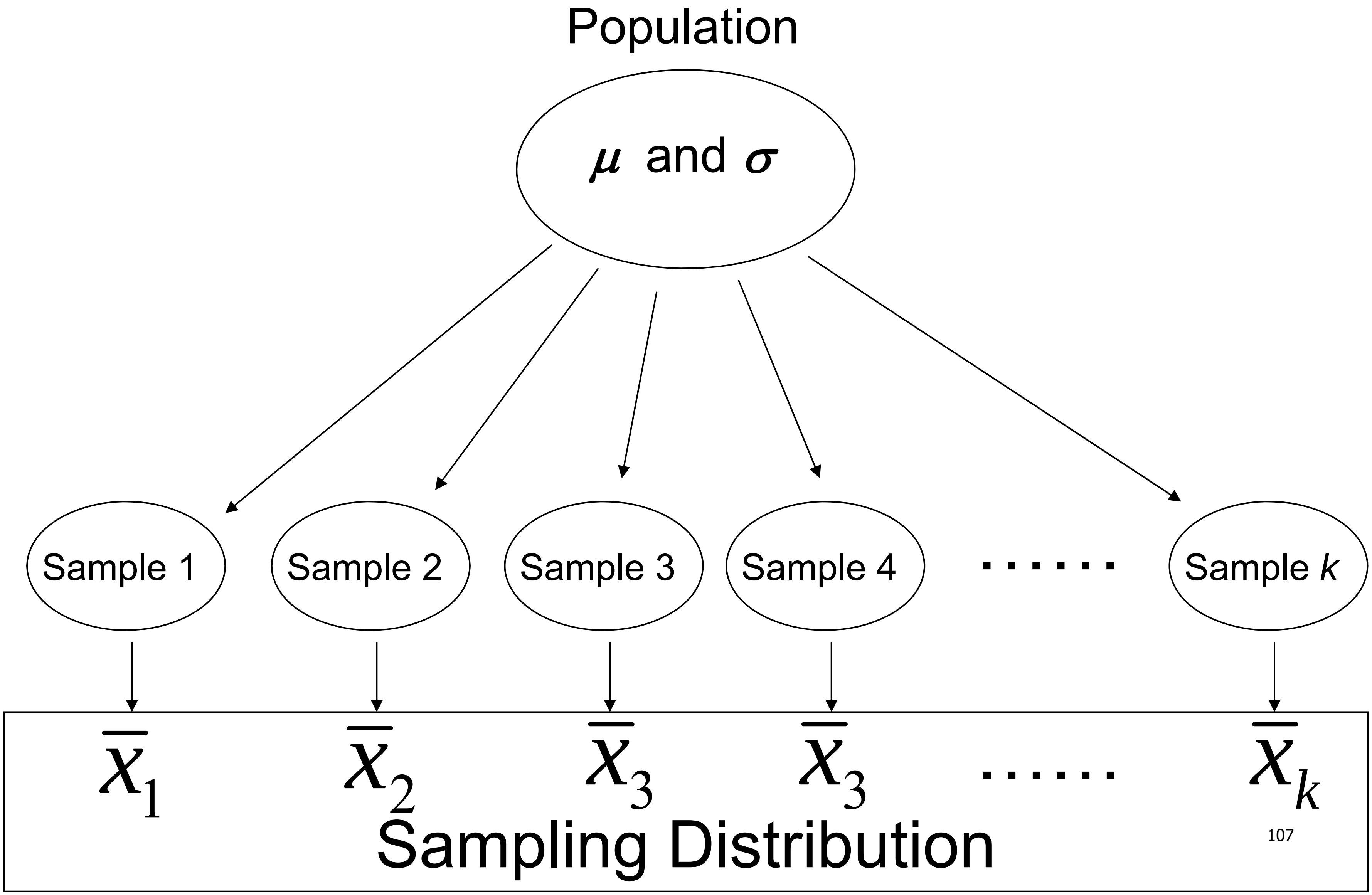
Distribution of averages approach Normal as sample size (no. of dice) increases

CLT cont’d

(3) It describes the distribution of the sample mean

- The values of $\overline{x}$ obtained from repeatedly taking samples of size *n* describe a separate population
- The distribution of any statistic is often called the **sampling distribution**

Sampling distribution of $\overline{X}$



Population

μ and σ

Sample 1 Sample 2 Sample 3 Sample 4 ······ Sample *k*

$\overline{x}_1$ $\overline{x}_2$ $\overline{x}_3$ $\overline{x}_3$ ······ $\overline{x}_k$

Sampling Distribution

CLT continued

(4) The mean of the sampling distribution of $\overline{X}$ is equal to the population mean, i.e.

(5) Standard deviation of the sampling distribution of is the population standard deviation ÷ square root of sample size, i.e.

$$\mu_{\overline{X}} = \mu$$

$$\overline{X}$$

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Estimates

- Since ***s*** is an estimate of σ, an estimate of $\frac{\sigma}{\sqrt{n}}$ is $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

- This is known as the **standard error of the mean**

- Be careful not to confuse the standard deviation and the standard error !
  - Standard deviation describes the variability of the data
  - Standard error is the measure of the precision of $\overline{X}$ as a measure of μ

Confidence Interval

- A **confidence interval** for a population characteristic is an interval of plausible values for the characteristic. It is constructed so that, with a chosen degree of confidence (the **confidence level**), the value of the characteristic will be captured inside the interval
- E.g. we claim with 95% confidence that the population mean lies between 15.6 and 17.2

Methods for Statistical Inference

Confidence Intervals

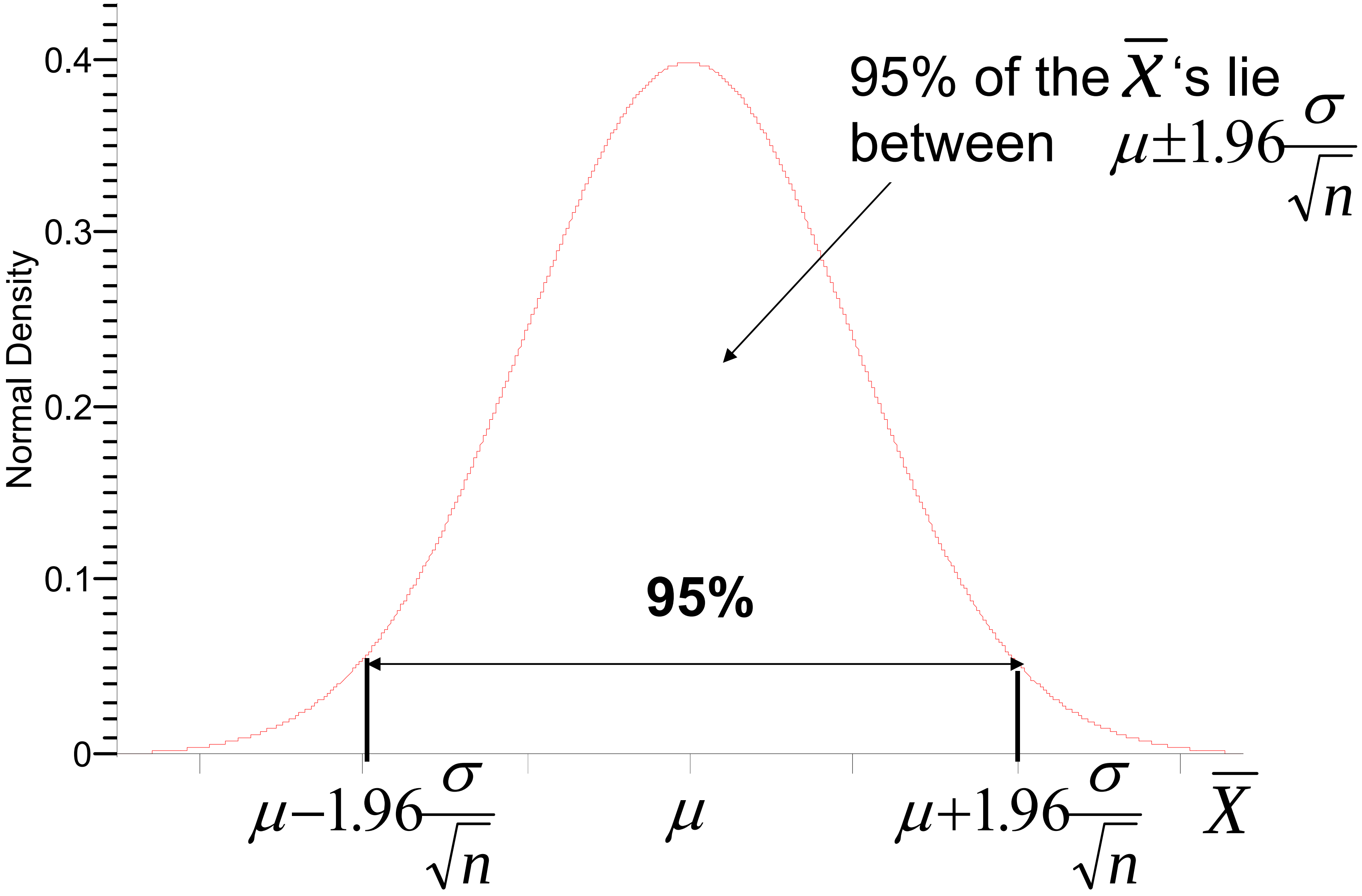
Hypothesis Tests

Confidence Interval for μ when σ is known

- A 95% confidence interval for μ if σ is known is given by:

$$\bar{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$$

Sampling distribution of $\overline{X}$



95% of the $\overline{X}$'s lie between $\mu \pm 1.96\frac{\sigma}{\sqrt{n}}$

Rationale for Confidence Interval

- From the sampling distribution of $\overline{X}$ conclude that μ and $\overline{X}$ are within 1.96
- standard errors ($\frac{\sigma}{\sqrt{n}}$) of each other 95% of the time
- Otherwise stated, 95% of the intervals contain μ
- So, the interval $\overline{x} \pm 1.96 \times \frac{\sigma}{\sqrt{n}}$ can be taken as an interval that typically would include μ

Example

- A random sample of 80 tablets had an average potency of 15mg. Assume σ is known to be 4mg.
- $\overline{X}$ =15, σ =4, n=80
- A 95% confidence interval for μ is

$$15 \pm 1.96 \times \frac{4}{\sqrt{80}}$$

= (14.12 , 15.88)

Confidence Interval for μ when σ is unknown

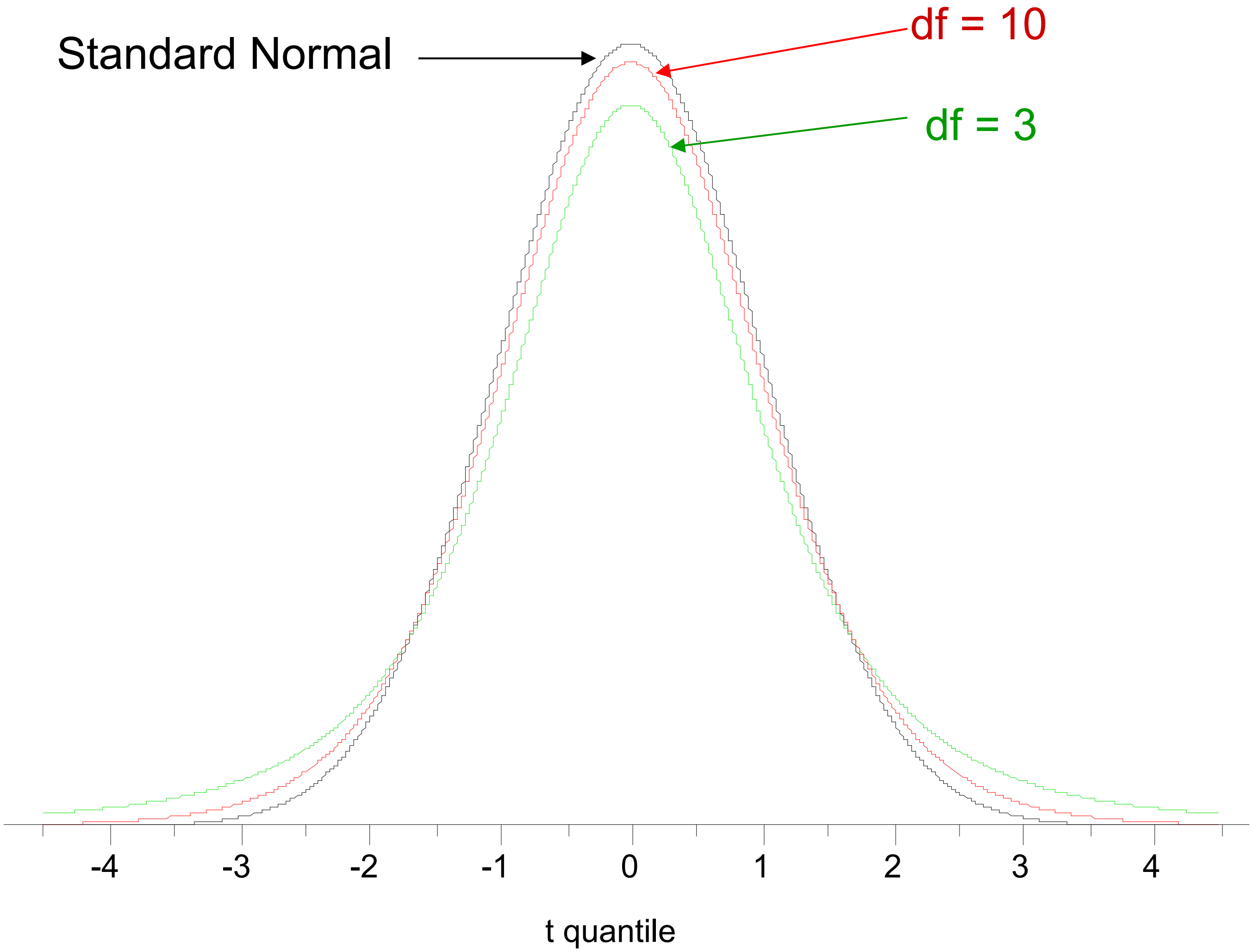
- Nearly always σ is unknown and is estimated using sample standard deviation $\boldsymbol{s}$
- The value 1.96 in the confidence interval is replaced by a new quantity, i.e., $t_{0.025}$
- The 95% confidence interval when σ is unknown is:

$$\overline{x} \pm t_{0.025} \times \frac{s}{\sqrt{n}}$$

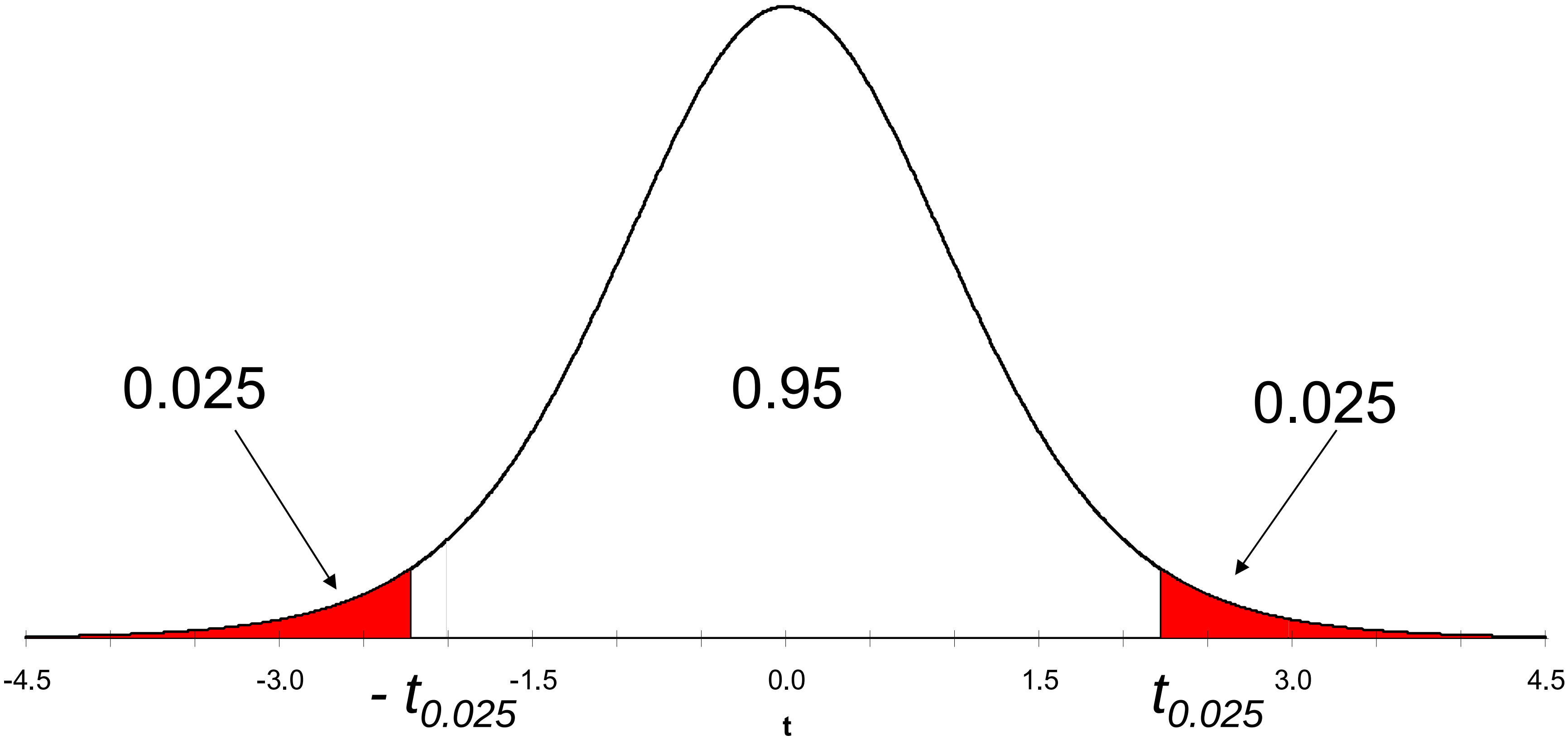
Student's *t* Distribution

- Closely related to the standard normal distribution *Z*
  - Symmetric and bell-shaped
  - Has mean = 0 but has a larger standard deviation
- Exact shape depends on a parameter called **degrees of freedom** (df) which is related to sample size
  - In this context df = *n-1*

Student's *t* distribution for 3, 10 df and standard Normal distribution



Definition of $t_{0.025}$ values



Example

- 26 measurements of the potency of a single batch of tablets in mg per tablet are as follows

498.38	489.31	505.50	495.24	490.17	483.2
488.47	497.71	503.41	482.25	488.14	
492.22	483.96	473.93	463.40	493.65	
499.48	496.05	494.54	508.58	488.42	
463.68	492.46	489.45	491.57	489.33	

Example (contd.)

- $\bar{x} = 490.096, \text{and } s = 10.783$ mg per tablet
- $t_{0.025}$ with df = 25 is 2.06

$$\bar{x} \pm t_{0.025} \times \frac{s}{\sqrt{n}} = 490.096 \pm 2.06 \times \frac{10.783}{\sqrt{26}}$$

$$= 490.096 \pm 4.356$$

- So, the batch potency lies between 485.74 and 494.45 mg per tablet

General Form of Confidence Interval

Estimate ±(critical value from distribution).(standard error)

Hypothesis testing

- Used to investigate the validity of a claim about the value of a population characteristic
- For example, the mean potency of a batch of tablets is 500mg per tablet, i.e.,

  μ_0 = 500mg

Procedure

- Specify Null and Alternative hypotheses
- Specify test statistic
- Define what constitutes an exceptional outcome
- Calculate test statistic and determine whether or not to reject the Null Hypothesis

Step 1

- Specify the hypothesis to be tested and the alternative that will be decided upon if this is rejected
  - The hypothesis to be tested is referred to as the **Null Hypothesis** (labelled H_0)
  - The alternative hypothesis is labelled H_1
- For the earlier example this gives:

$$H_0 : \mu = 500\text{mg}$$
$$H_a : \mu \neq 500\text{mg}$$

Step 1 (continued)

- The Null Hypothesis is assumed to be true unless the data clearly demonstrate otherwise

Step 2

- Specify a test statistic which will be used to measure departure from

$$H_0 : \mu = \mu_0$$

where μ_0 is the value specified under the Null Hypothesis, e.g. $\mu_0 = 500$ in the earlier example.

- For hypothesis tests on sample means the test statistic is:

$$t = \frac{\overline{x} - \mu_0}{s / \sqrt{n}}$$

Step 2 (contd.)

- The test statistic $$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

  is a ‘signal to noise ratio’, i.e. it measures how far is from $\bar{x}$ in terms of standard error units μ_0
- The *t* distribution with df = *n*-1 describes the distribution of the test statistics **if** the Null Hypothesis is true
- In the earlier example, the test statistic *t* has a *t* distribution with df = 25

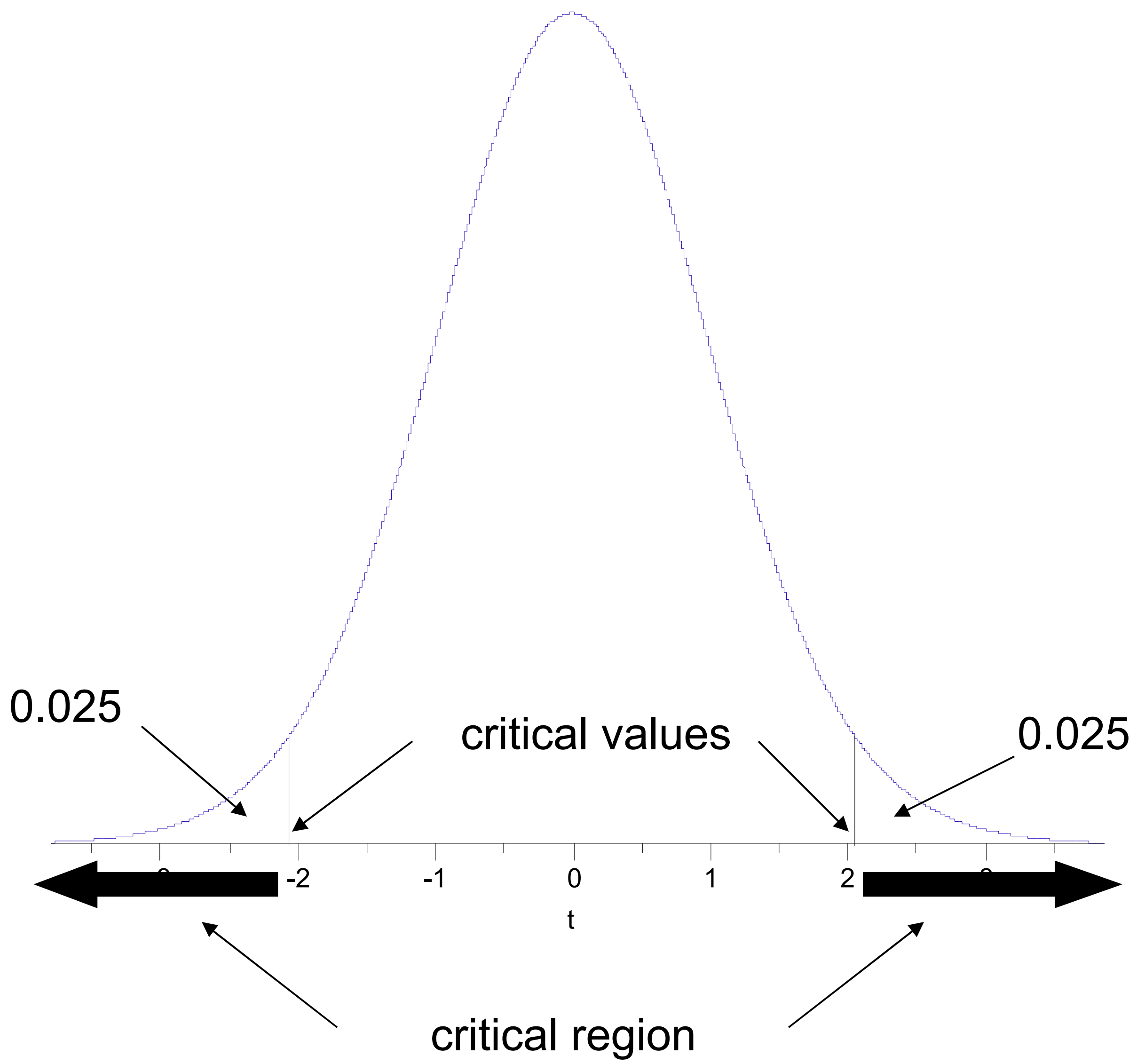
Step 3

- Define what will be an exceptional outcome
  - a value of the test statistic is exceptional if it has only a small chance of occurring when the null hypothesis is true
- The probability chosen to define an exceptional outcome is called the **significance level** of the test and is labelled α
  - Conventionally, α is chosen to be = 0.05

Step 3 (contd.)

- $\alpha = 0.05$ gives cut-off values on the sampling distribution of *t* called **critical values**
  - values of the test statistic *t* lying beyond the critical values lead to rejection of the null hypothesis
- For the earlier example the critical value for a *t* distribution with df = 25 is 2.06

t distribution with df=25 showing critical region



0.025

critical values

0.025

critical region

Step 4

- Calculate the test statistic and see if it lies in the critical region
- For the example

$$t = \frac{490.096 - 500}{10.783 \Big/ \sqrt{26}}$$

$$= -4.683$$

- *t =* -4.683 is < -2.06 so the hypothesis that the batch potency is 500 mg/tablet is rejected

P value

The **P value** associated with a hypothesis test is the probability of getting sample values **as extreme or more extreme** than those actually observed, assuming null hypothesis to be true

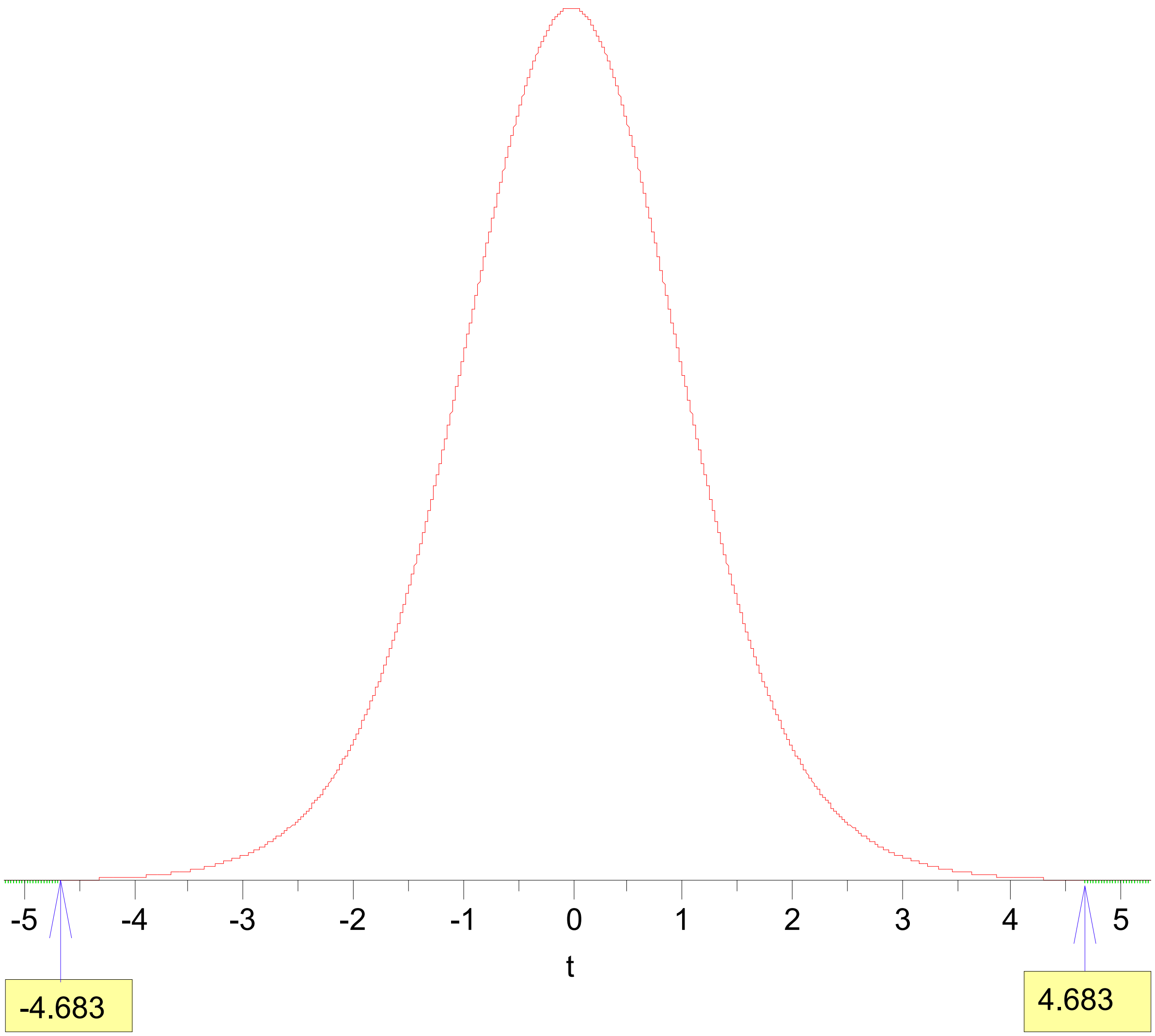
Example (contd)

- P value = probability of observing a more extreme value of *t*
- The observed *t* value was -4.683, so the P value is the probability of getting a value more extreme than ± 4.683
- This P value is calculated as the area under the *t* distribution below -4.683 plus the area above 4.683, i.e., 0.00008474 !

Example (contd)

- Less than 1 in 10,000 chance of observing a value of t more extreme than -4.683 if the Null Hypothesis is true
- Evidence in favour of the alternative hypothesis is very strong

P value (contd.)



Two-tail and One-tail tests

- The test described in the previous example is a **two-tail** test
  - The null hypothesis is rejected if either an unusually large or unusually small value of the test statistic is obtained, i.e. the rejection region is divided between the two tails

One-tail tests

- Reject the null hypothesis only if the observed value of the test statistic is
  - Too large
  - Too small
- In both cases the critical region is entirely in one tail so the tests are **one-tail** tests

Statistical versus Practical Significance

- When we reject a null hypothesis it is usual to say the result is **statistically significant** at the chosen level of significance
- But should also always consider the **practical significance** of the **magnitude** of the difference between the estimate (of the population characteristic) and what the null hypothesis states that to be

Hypothesis Testing

- Is also called *significance testing*
- Tests a claim about a parameter using evidence (data in a sample
- The technique is introduced by considering a one-sample z test
- The procedure is broken into four steps
- *Each* element of the procedure must be understood

Hypothesis Testing Steps

A. Null and alternative hypotheses

B. Test statistic

C. P-value and interpretation

D. Significance level (optional)

Null and Alternative Hypotheses

- Convert the research question to null and alternative hypotheses
- The **null hypothesis (H_0)** is a claim of “no difference in the population”
- The **alternative hypothesis (H_a)** claims “H_0 is false”
- Collect data and seek evidence against H_0 as a way of bolstering H_a (deduction)

Illustrative Example: “Body Weight”

- **The problem:** In the 1970s, 20–29 year old men in the U.S. had a mean μ body weight of 170 pounds. Standard deviation σ was 40 pounds**.** We test whether mean body weight in the population now differs.
- **Null hypothesis** $H_{0:}$ μ = 170 (“no difference”)
- The **alternative hypothesis** can be either $H_{a:}$ μ > 170 (**one-sided test**) or
  $H_{a:}$ μ ≠ 170 (**two-sided test**)

§9.2 Test Statistic

This is an example of a one-sample test of a mean when σ is known. Use this statistic to test the problem:

$$\mathrm{z}_{\mathrm{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}}$$

$$\text{where } \mu_0 \equiv \text{population mean assuming } H_0 \text{ is true}$$

$$\text{and } SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Illustrative Example: *z* statistic

- For the illustrative example, μ_0 = 170
- We know σ = 40
- Take an SRS of *n* = 64. Therefore

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{64}} = 5$$

- If we found a sample mean of 173, then

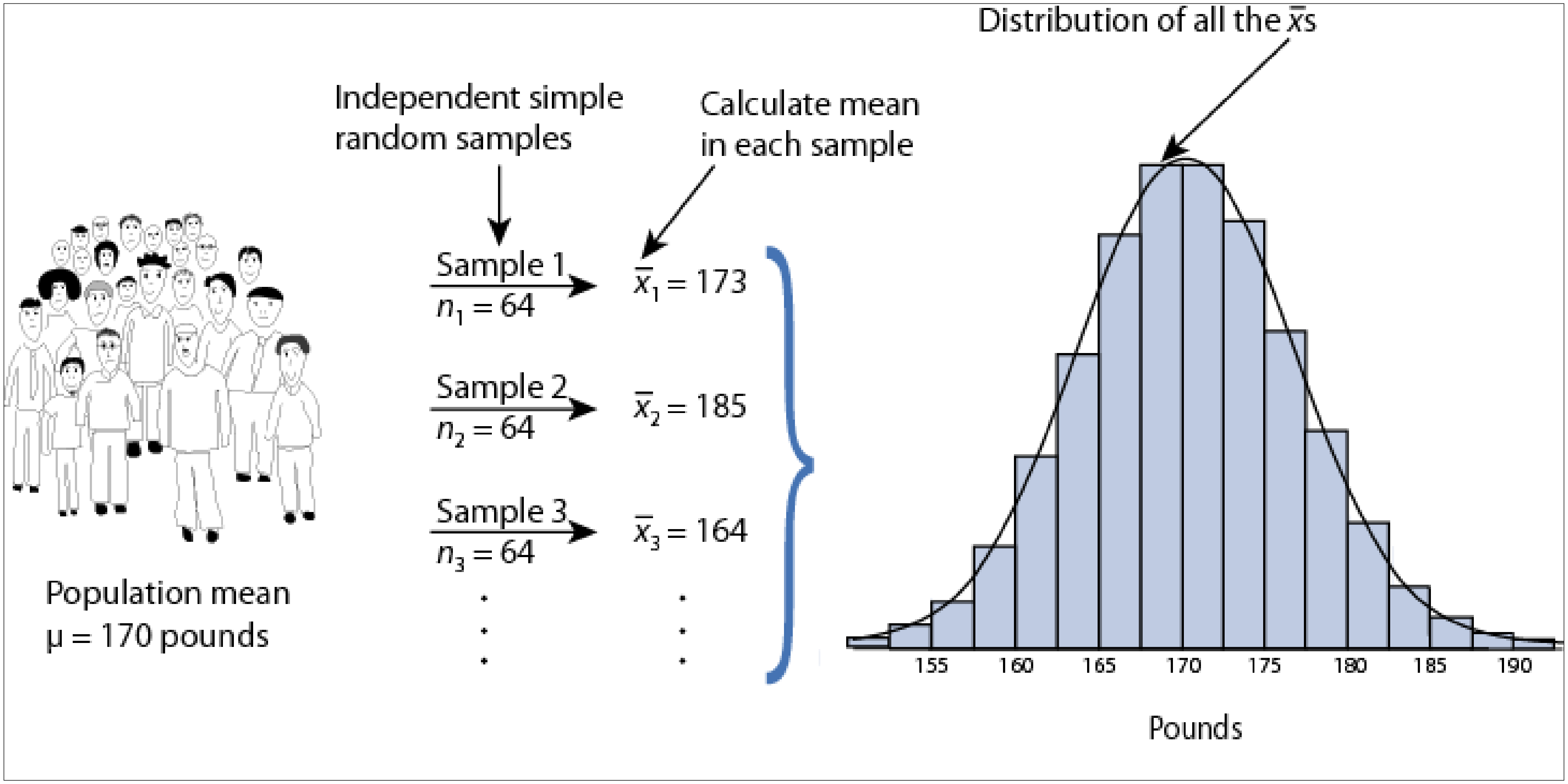
$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{173 - 170}{5} = 0.60$$

Illustrative Example: *z* statistic

If we found a sample mean of 185, then

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{185 - 170}{5} = 3.00$$

Reasoning Behin μz_{stat}



Sampling distribution of xbar under H_0: μ = 170 for n = 64 ⇒ $\bar{x} \sim N(170,5)$

P-value

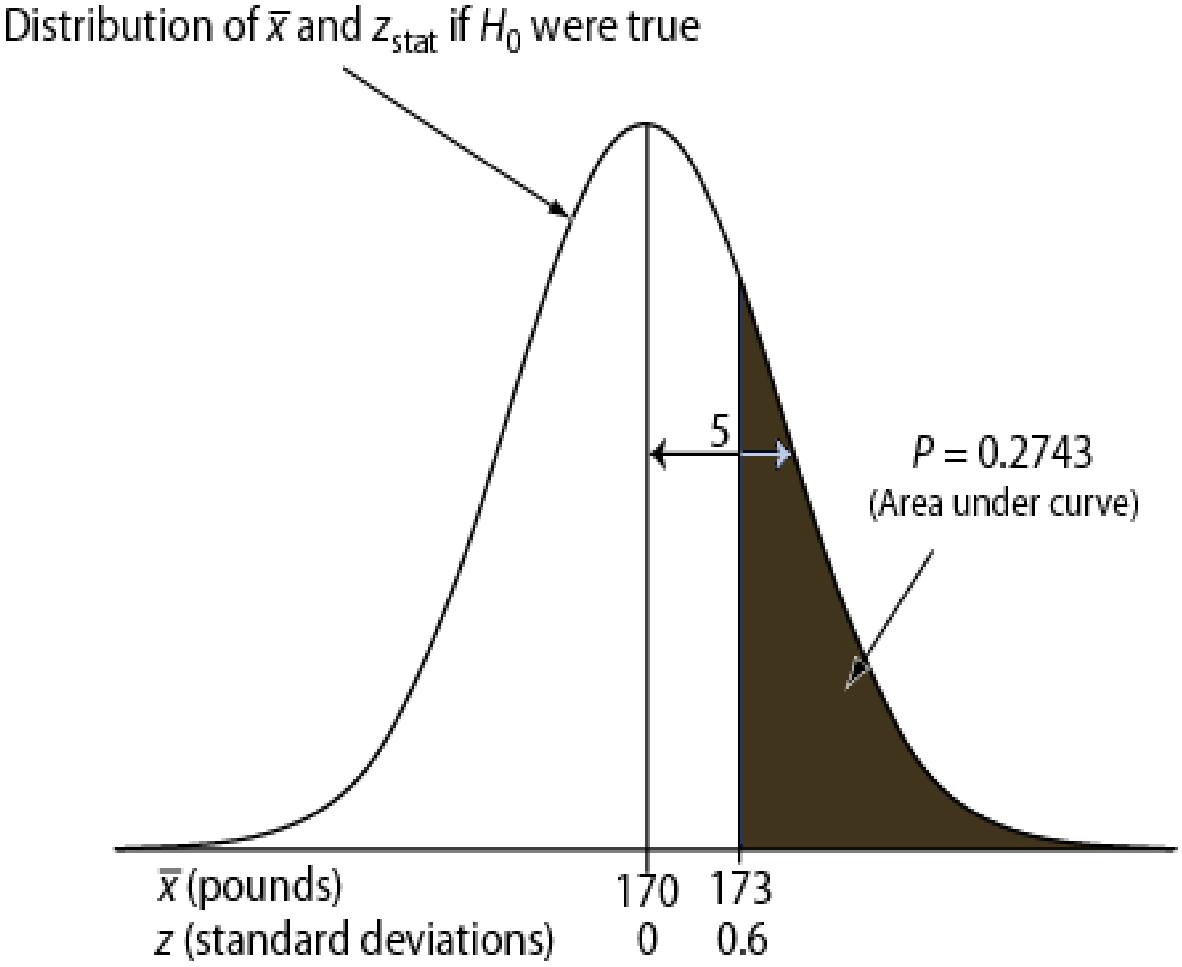
- The *P*-value answer the question: What is the probability of the observed test statistic or one more extreme **when *H_0 is true***?
- This corresponds to the AUC in the tail of the Standard Normal distribution beyond the $z_{\text{stat.}}$
- Convert *z* statistics to *P*-value :

  For H_a: $\mu > \mu_0 \Rightarrow P = \Pr(Z > z_{\text{stat}})$ = right-tail beyond z_{stat}

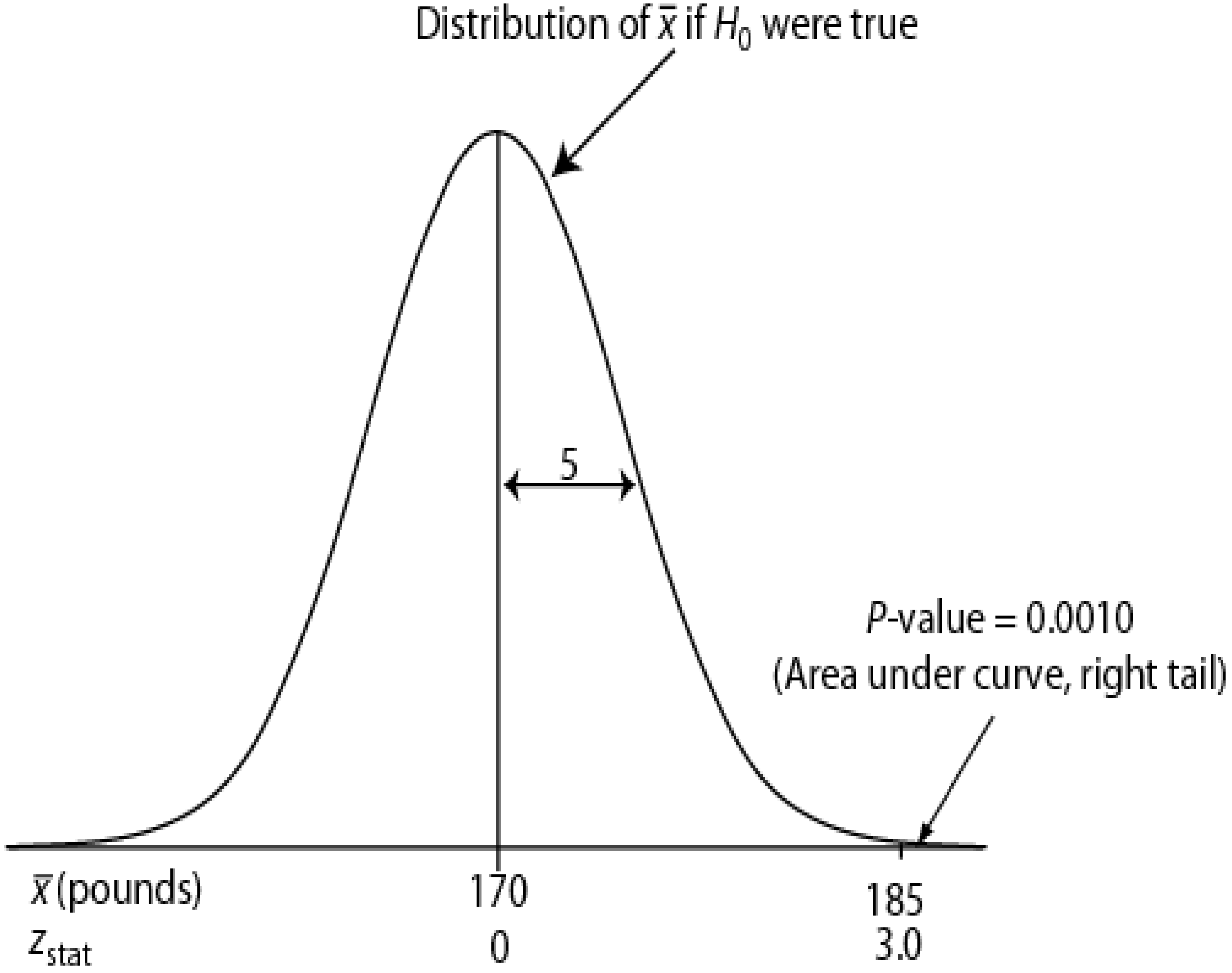
  For H_a: $\mu < \mu_0 \Rightarrow P = \Pr(Z < z_{\text{stat}})$ = left tail beyond z_{stat}

  For H_a: $\mu \neq \mu_0 \Rightarrow P$ = 2 × one-tailed *P*-value
- Use Table B or software to find these probabilities (next two slides).

One-sided *P*-value for z_{stat} of 0.6

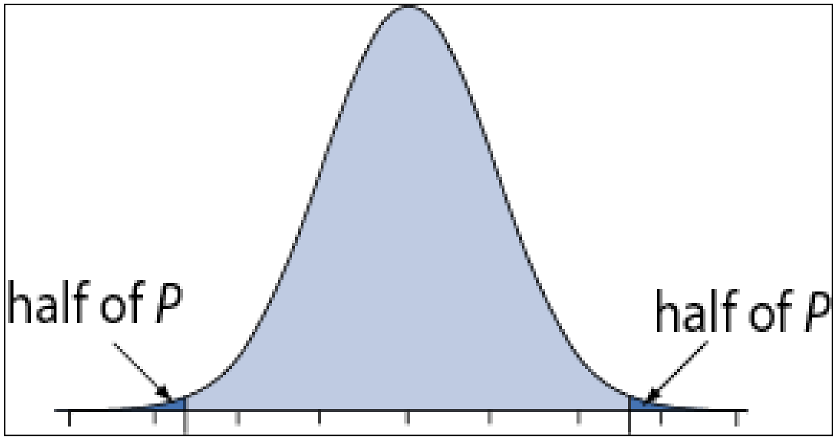


One-sided *P*-value for z_{stat} of 3.0



Two-Sided *P*-Value

- One-sided H_a ⇒ AUC in tail beyond z_{stat}
- Two-sided H_a ⇒ consider potential deviations in both directions ⇒ double the one-sided *P*-value



Examples: If one-sided *P* = 0.0010, then two-sided *P* = 2 × 0.0010 = 0.0020. If one-sided *P* = 0.2743, then two-sided *P* = 2 × 0.2743 = 0.5486.

Interpretation

- *P*-value answer the question: What is the probability of the observed test statistic … **when *H_0 is true*?**
- Thus, smaller and smaller *P*-values provide stronger and stronger evidence against H_0
- Small *P*-value $\Rightarrow$ strong evidence

Interpretation

Conventions*

$P > 0.10 \Rightarrow$ non-significant evidence against H_0

$0.05 < P \leq 0.10 \Rightarrow$ marginally significant evidence

$0.01 < P \leq 0.05 \Rightarrow$ significant evidence against H_0

$P \leq 0.01 \Rightarrow$ highly significant evidence against H_0

Examples

$P = .27 \Rightarrow$ non-significant evidence against H_0

$P = .01 \Rightarrow$ highly significant evidence against H_0

*** It is *unwise* to draw firm borders for "significance"**

α-Level (Used in some situations)

- Let α ≡ probability of erroneously rejecting H_0
- Set α threshold (e.g., let α = .10, .05, *or whatever*)
- Reject H_0 when $P \leq \alpha$
- Retain H_0 when $P > \alpha$
- Example: Set α = .10. Find $P = 0.27 \Rightarrow$ retain H_0
- Example: Set α = .01. Find $P = .001 \Rightarrow$ reject H_0

(Summary) One-Sample *z* Test

A. Hypothesis statements
   H_0: μ = μ_0 vs.
   H_a: μ ≠ μ_0 (two-sided) or
   H_a: μ < μ_0 (left-sided) or
   H_a: μ > μ_0 (right-sided)

B. Test statistic

C. P-value: convert z_{stat} to P value

D. Significance statement (usually not necessary)

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} \text{ where } SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Two-Sample Inferences

- So far, we have dealt with inferences about μ for a **single** population using a **single** sample.
- Many studies are undertaken with the objective of comparing the characteristics of two populations. In such cases we need two samples, one for each population
- The two samples will be **independent** or dependent (**paired**) according to how they are selected

Example

- Animal studies to compare toxicities of two drugs

2 independent samples:

Select sample of rats for drug 1 and another sample of rats for drug 2

2 paired samples:

Select a number of pairs of litter mates and use one of each pair for drug 1 and drug 2

Two Sample t-test

- Consider inferences on 2 independent samples
- We are interested in testing whether a difference exists in the population means, μ_1 and μ_2

$$\text{Formulate hypotheses}$$

$$\mathrm{H_0} : \mu_2 - \mu_1 = 0$$

$$\mathrm{H_a} : \mu_2 - \mu_1 \neq 0$$

Two Sample t-Test

- It is natural to consider the statistic $\overline{x}_2 - \overline{x}_1$ and its sampling distribution
- The distribution is centred at μ_2-μ_1, with standard error

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- If the two populations are normal, the sampling distribution is normal
- For large sample sizes (n_1 and $n_2 > 30$), the sampling distribution is approximately normal even if the two populations are not normal (CLT)

Two Sample t-Test

- The two-sample t-statistic is defined as

$$t = \frac{(\overline{x}_2 - \overline{x}_1) - (\mu_2 - \mu_1)}{s_p \sqrt{\frac{1}{n_1} + \frac{1,}{n_2}}} \quad \text{where} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- The two sample standard deviations are combined to give a pooled estimate of the population standard deviation σ

Two-sample Inference

- The t statistic has n_1+n_2-2 degrees of freedom
- Calculate critical value & p value as per usual
- The 95% confidence interval for $\mu_2-\mu_1$ is

$$(\bar{x}_2 - \bar{x}_1) \pm t_{0.025} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example

Population	n	mean	s
Drug 1	20	35.9	11.9
Drug 2	38	36.6	12.3

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{(19)(141.61) + (37)(151.29)}{56}$$

$$= 148.01$$

Example (contd)

$$t = \frac{(\overline{x}_2 - \overline{x}_1) - 0}{s_p^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$= \text{-}0.21$$

- Two-tailed test with 56 df and α=0.05 therefore we reject the null hypthesis if t>2 or t<-2
- Fail to reject - there is insufficient evidence of a difference in mean between the two drug populations
- Confidence interval is -7.42 to 6.02

Paired t-test

- Methods for independent samples are **not** appropriate for paired data.
- Two related observations (i.e. two observations per subject) and you want to see if the means on these two normally distributed interval variables differ from one another.
- Calculation of the t-statistic, 95% confidence intervals for the mean difference and P-values are estimated as presented previously for one-sample testing.

Example

- 14 cardiac patients were placed on a special diet to lose weight. Their weights (kg) were recorded before starting the diet and after one month on the diet
- Question: Do the data provide evidence that the diet is effective?

Patient	**Before**	**After**	**Difference**
1	62	59	3
2	62	60	2
3	65	63	2
4	88	78	10
5	76	75	1
6	57	58	-1
7	60	60	0
8	59	52	7
9	54	52	2
10	68	65	3
11	65	66	-1
12	63	59	4
13	60	58	2
14	56	55	1

Example

$$H_0 : \mu_d = 0$$

$$H_a : \mu_d > 0$$

$$\bar{x}_d = 2.5 \quad s_d = 2.98 \quad n = 14$$

$$t = \frac{\bar{x}_d - 0}{s_d \big/ \sqrt{n}} = \frac{2.5}{2.98 \big/ \sqrt{14}} = \boxed{3.14}$$

Example (contd)

- Critical Region (1 tailed) t > 1.771
- Reject H_0 in favour of H_a
- P value is the area to the right of 3.14

= 1-0.9961=0.0039

- 95% Confidence Interval for $\mu_d = \mu_1 - \mu_2$

2.5 ± 2.17 (2.98/√14)

= 2.5 ±1.72

=0.78 to 4.22

Example (cont)

- Suppose these data were (incorrectly) analysed as if the two samples were independent…

➔ t=0.80

Example (contd)

- We calculate t=0.80
- This is an upper tailed test with 26 df and α=0.05 (5% level of significance) therefore we reject H_0 if t>1.706
- Fail to reject - there is not sufficient evidence of a difference in mean between ‘before’ and ‘after’ weights

Wrong Conclusions

- By ignoring the paired structure of the data, we incorrectly conclude that there was no evidence of diet effectiveness.
- When pairing is ignored, the variability is inflated by the subject-to-subject variation.
- The paired analysis eliminates this source of variability from the calculations, whereas the unpaired analysis includes it.
- Take home message: NB to use the right test for your data. If data is paired, use a test that accounts for this.

Analysis of Variance (ANOVA)

- Many investigations involved a comparison of **more than two** population means
- Need to be able to extend our two sample methods to situations involving more than two samples
- i.e. equivalent of the paired samples t-test, but allows for two or more levels of the categorical variable
- Tests whether the mean of the dependent variable differs by the categorical variable
- Such methods are known collectively as the **analysis of variance**

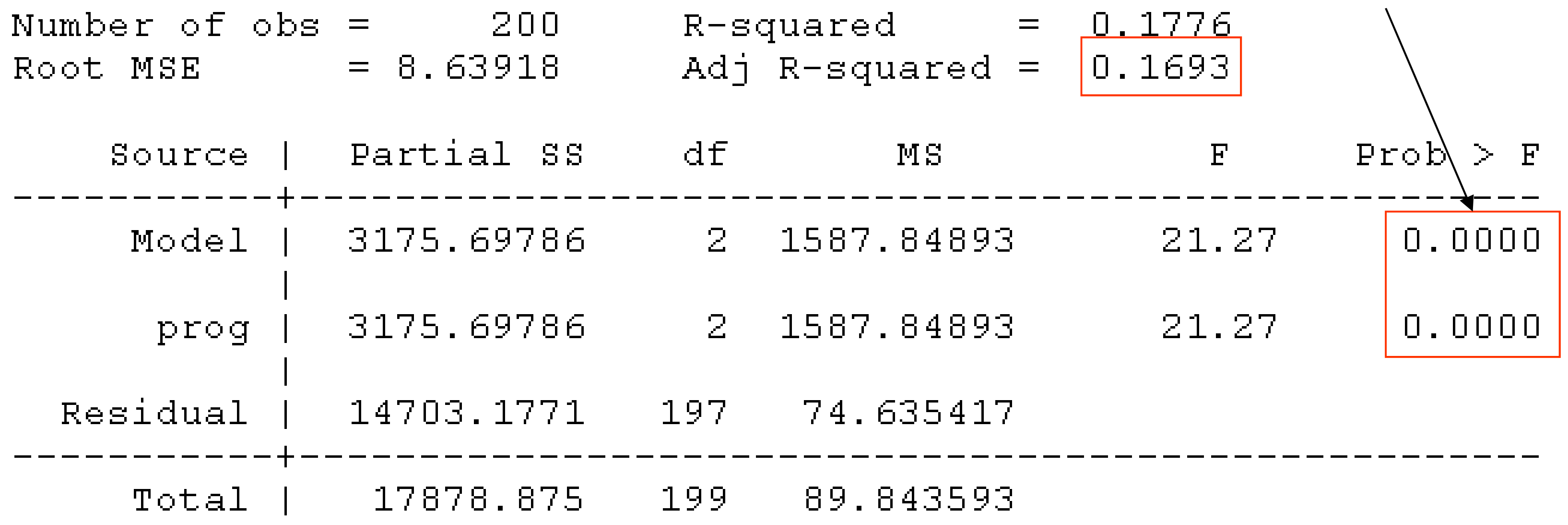
Completely Randomised Design/one-way ANOVA

- Equivalent to independent samples design for two populations
- A completely randomised design is frequently referred to as a **one-way ANOVA**
- Used when you have a categorical independent variable (with two or more categories) and a **normally distributed** interval dependent variable (e.g. $10,000,$15,000,$20,000) and you wish to test for differences in the means of the dependent variable broken down by the levels of the independent variable
- e.g. compare three methods for measuring tablet hardness. 15 tablets are randomly assigned to three groups of 5 and each group is measured by one of these methods

ANOVA example

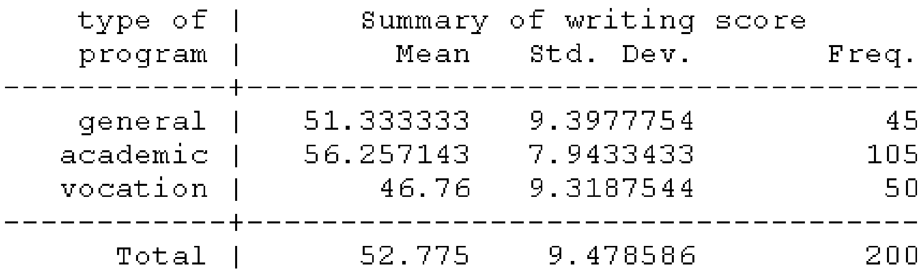
Mean of the dependent variable differs significantly among the levels of program type. However, we do not know if the difference is between only two of the levels or all three of the levels.

```
anova write prog
```



```
Number of obs =      200     R-squared     =  0.1776
Root MSE      = 8.63918     Adj R-squared =  0.1693

      Source |  Partial SS    df       MS           F     Prob > F
-------------+----------------------------------------------------
       Model |  3175.69786     2  1587.84893      21.27     0.0000
             |
        prog |  3175.69786     2  1587.84893      21.27     0.0000
             |
    Residual |  14703.1771   197   74.635417
-------------+----------------------------------------------------
       Total |   17878.875   199   89.843593
```

```
tabulate prog, summarize(write)
```



```
    type of |      Summary of writing score
    program |        Mean   Std. Dev.       Freq.
------------+------------------------------------
    general |   51.333333   9.3977754          45
   academic |   56.257143   7.9433433         105
   vocation |       46.76   9.3187544          50
------------+------------------------------------
      Total |      52.775    9.478586         200
```

See that the students in the academic program have the highest mean writing score, while students in the vocational program have the lowest.

Example

Compare three methods for measuring tablet hardness.
15 tablets are randomly assigned to three groups of 5

Method A	Method B	Method C
102	99	103
101	100	100
101	99	99
100	101	104
102	98	102

Hypothesis Tests: One-way ANOVA

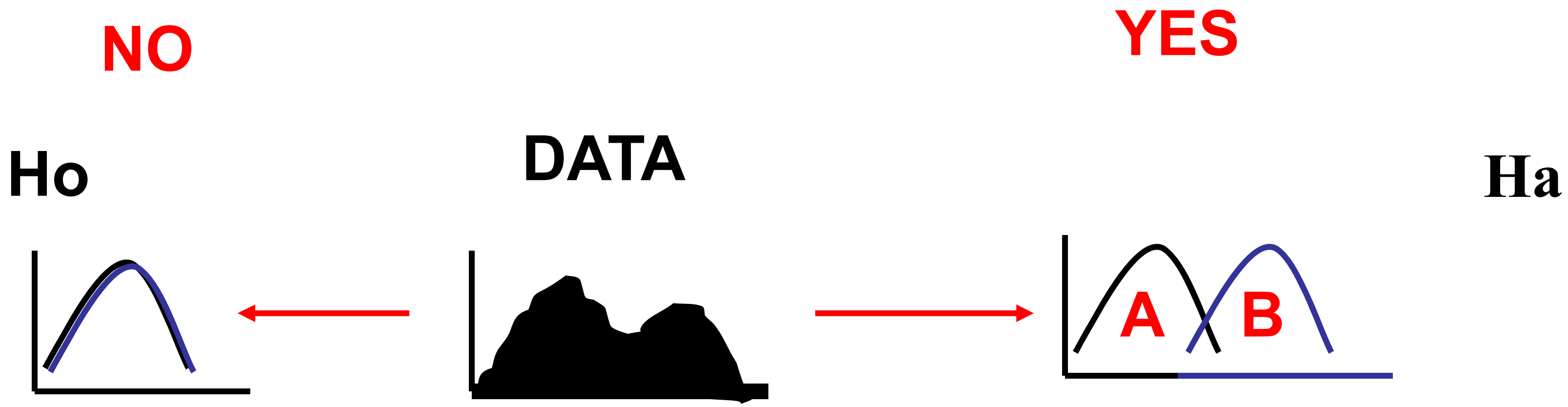
- K populations

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

$$H_A : at\, least\, one\, \mu\, is\, different$$

Do the samples come from different populations?

- Two-sample (t-test)



NO

YES

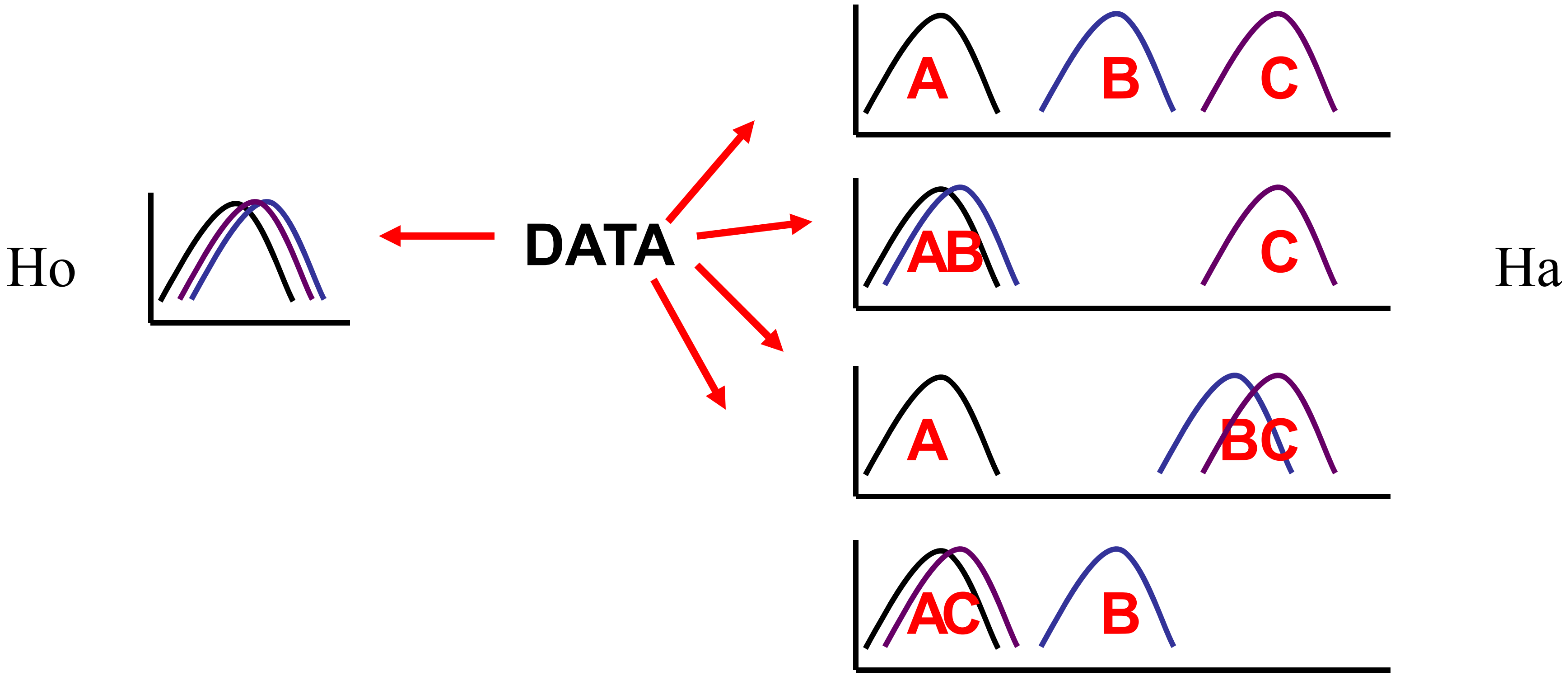
Ho

DATA

Ha

Do the samples come from different populations?

- One-way ANOVA (F-test)



Ho

DATA

Ha

F-test

- The ANOVA extension of the t-test is called the **F-test**
- Basis: We can decompose the total variation in the study into sums of squares
- Tabulate in an **ANOVA table**

Decomposition of total variability (sum of squares)

Assign subscripts to the data

- i is for treatment (or method in this case)
- j are the observations made within treatment

e.g.

- y_{11}= first observation for Method A i.e. 102
- $y_{1.}$ = average for Method A

Using algebra

Total Sum of Squares (SST)=Treatment Sum of Squares (SSX) + Error Sum of Squares (SSE)

$$\sum (y_{ij} - \overline{y})^2 = \sum (\overline{y}_{i.} - \overline{y})^2 + \sum (y_{ij} - \overline{y}_{i.})^2$$

ANOVA table

	df	SS	MS	F	P-value
Treatment (between groups)	*df (X)*	*SSX*	$\frac{SSX}{df\,(X)}$	$\frac{MSX}{MSE}$	*Look up !*
Error (within groups)	*df (E)*	*SSE*	$\frac{SSE}{df\,(E)}$		
Total	*df (T)*	*SST*			

Example (Contd)

- Are any of the methods different?
- P-value=0.0735
- At the 5% level of significance, there is no evidence that the 3 methods differ

Two-Way ANOVA

- Often, we wish to study <u>2 (or more) independent variables (factors)</u> in a single experiment
- An ANOVA of observations each of which can be classified in two ways is called a *two-way ANOVA*

Randomised Block Design

- This is an extension of the paired samples situation to more than two populations
- A block consists of homogenous items and is equivalent to a pair in the paired samples design
- The randomised block design is generally more powerful than the completely randomised design (/one way anova) because the variation between blocks is removed from the test statistic

Decomposition of sums of squares

$$\sum(y_{ij}-\bar{y})^2=\sum(\bar{y}_{i.}-\bar{y})^2+\sum(\bar{y}_{.j}-\bar{y})^2+\sum(\bar{y}_{ij}-\bar{y}_{i.}-\bar{y}_{j.}+\bar{y})^2$$

Total SS = Between Blocks SS + Between Treatments SS + Error SS

- Similar to the one-way ANOVA, we can decompose the overall variability in the data (total SS) into components describing variation relating to the factors (block, treatment) & the error (what’s left over)
- We compare Block SS and Treatment SS with the Error SS (a signal-to-noise ratio) to form F-statistics, from which we get a p-value

Example

- An experiment was conducted to compare the mean bioavailabilty (as measured by AUC) of three drug products from laboratory rats.
- Eight litters (each consisting of three rats) were used for the experiment. Each litter constitutes a block and the rats within each litter are randomly allocated to the three drug products

Example (cont’d)

Litter	Product A	Product B	Product C
1	89	83	94
2	93	75	78
3	87	75	89
4	80	76	85
5	80	77	84
6	87	73	84
7	82	80	75
8	68	77	75

Example (cont'd): ANOVA table

Source	df	SS	MS	F-ratio	P-value
Product	**2**	**200.333**	**100.167**	**3.4569**	**0.0602**
Litter	**7**	**391.833**	**55.9762**	**1.9318**	**0.1394**
Error	**14**	**405.667**	**28.9762**		
Total	**23**	**997.833**			

Interactions

- The previous tests for block and treatment are called tests for *main effects*

- ***Interaction effects*** happen when the effects of one factor are different depending on the level (category) of the other factor

Example

- 24 patients in total randomised to either Placebo or Prozac
- Happiness score recorded
- Also, patients gender may be of interest & recorded
- There are <u>two factors</u> in the experiment: treatment & gender
  - <u>Two-way ANOVA</u>

Example

- Tests for Main effects:
  - Treatment: are patients happier on placebo or prozac?
  - Gender: do males and females differ in score?
- Tests for Interaction:
  - Treatment x Gender: Males may be happier on prozac than placebo, but females not be happier on prozac than placebo. Also vice versa. Is there any evidence for these scenarios?
  - Include interaction in the model, along with the two factors treatment & gender

More jargon: factors, levels & cells

Happiness score

Factor 2 Treatment →

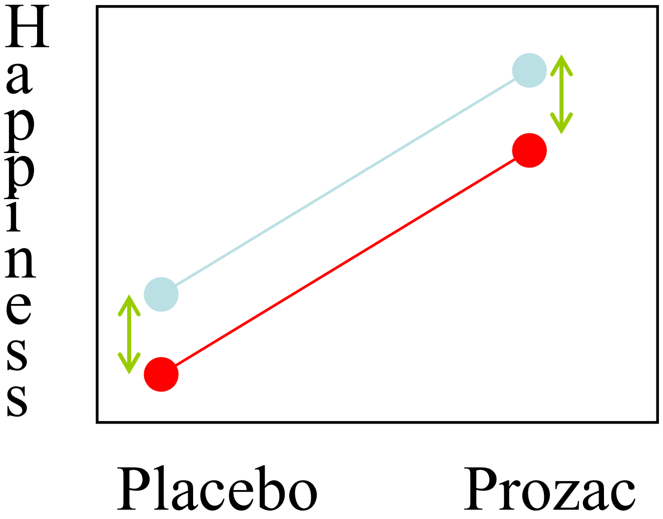
Levels

Cells

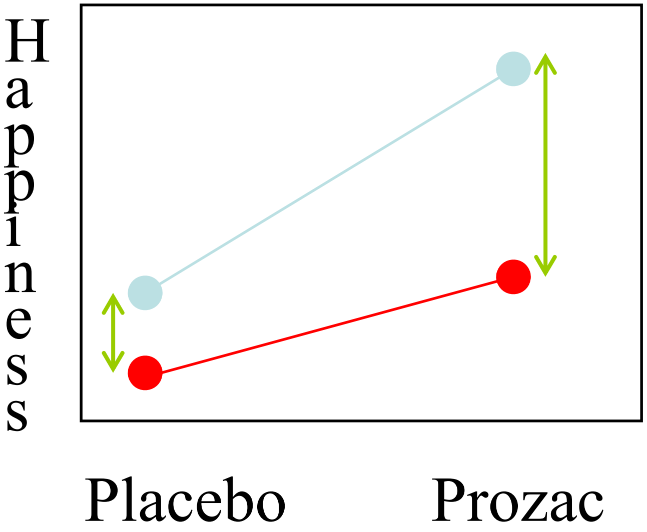
Factor 1 Gender ↓

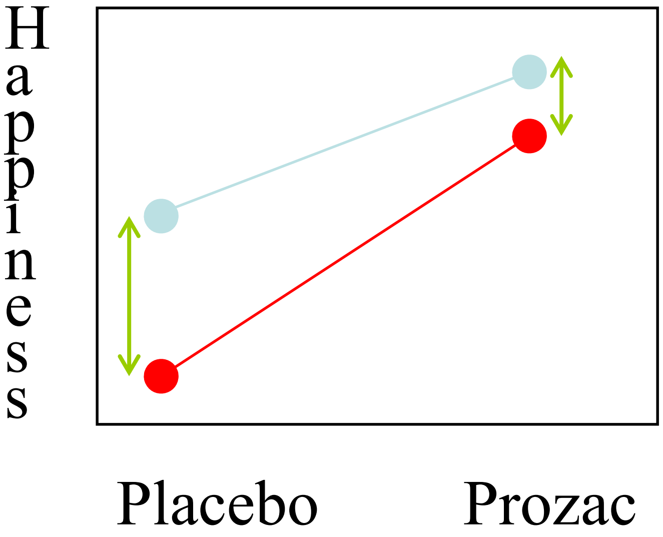
	Placebo	Prozac
Male	3 4 2 3 4 3	7 7 6 5 6 6
Female	4 5 4 6 6 4.5	5 5 5 4 6 6

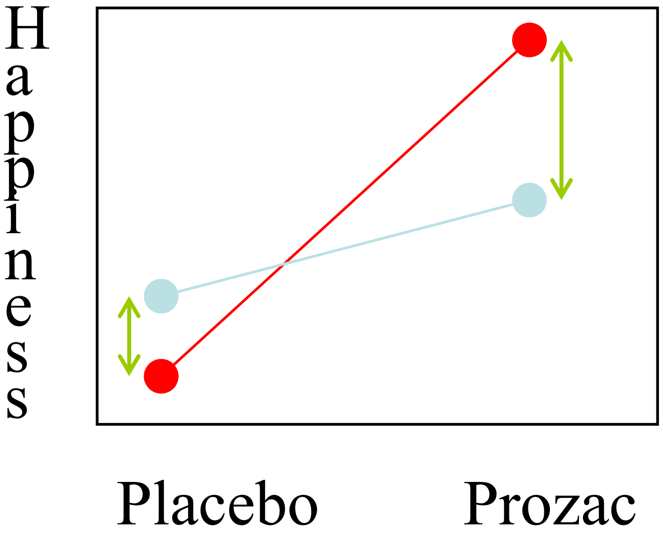
What do interactions looks like?



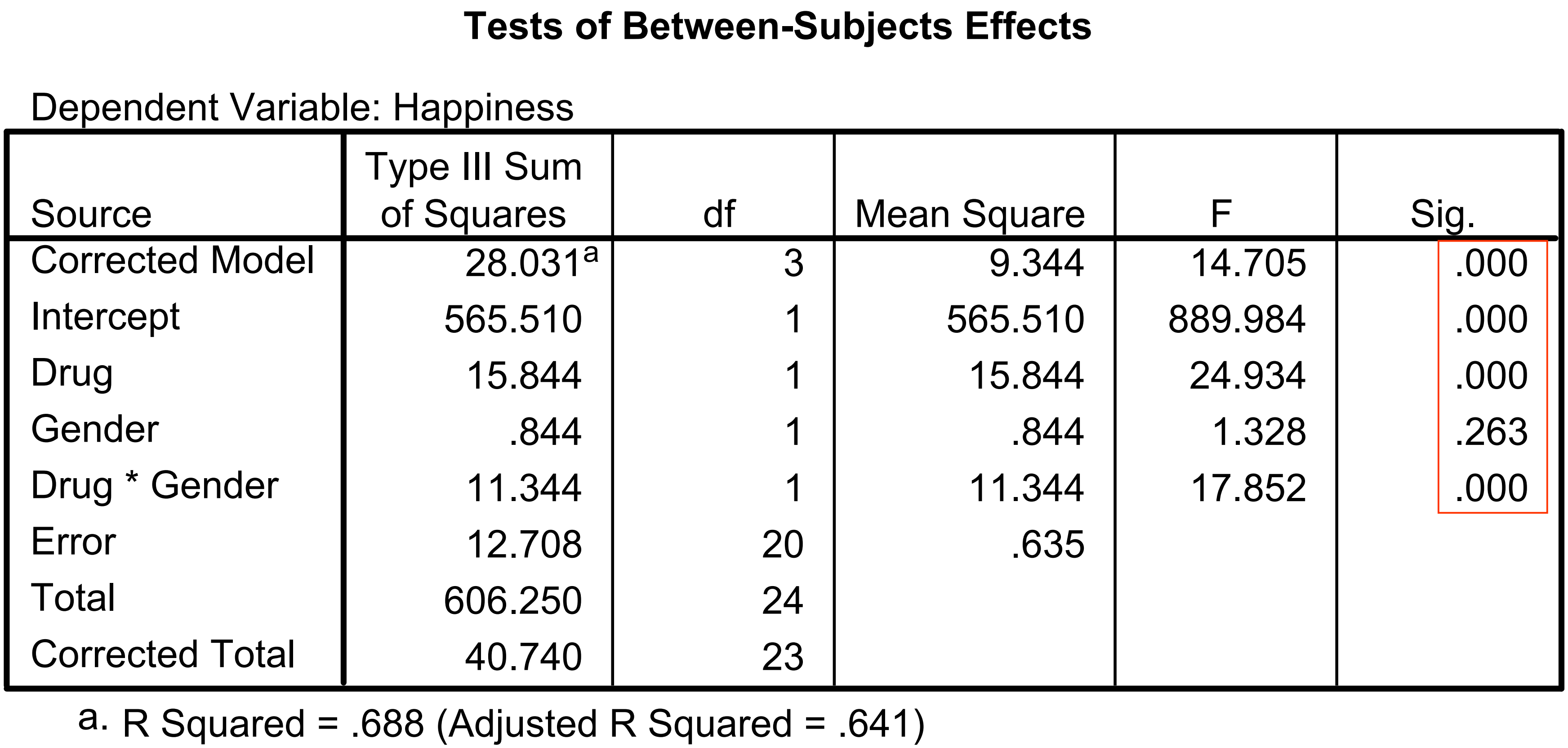
NO INTERACTION!







Results



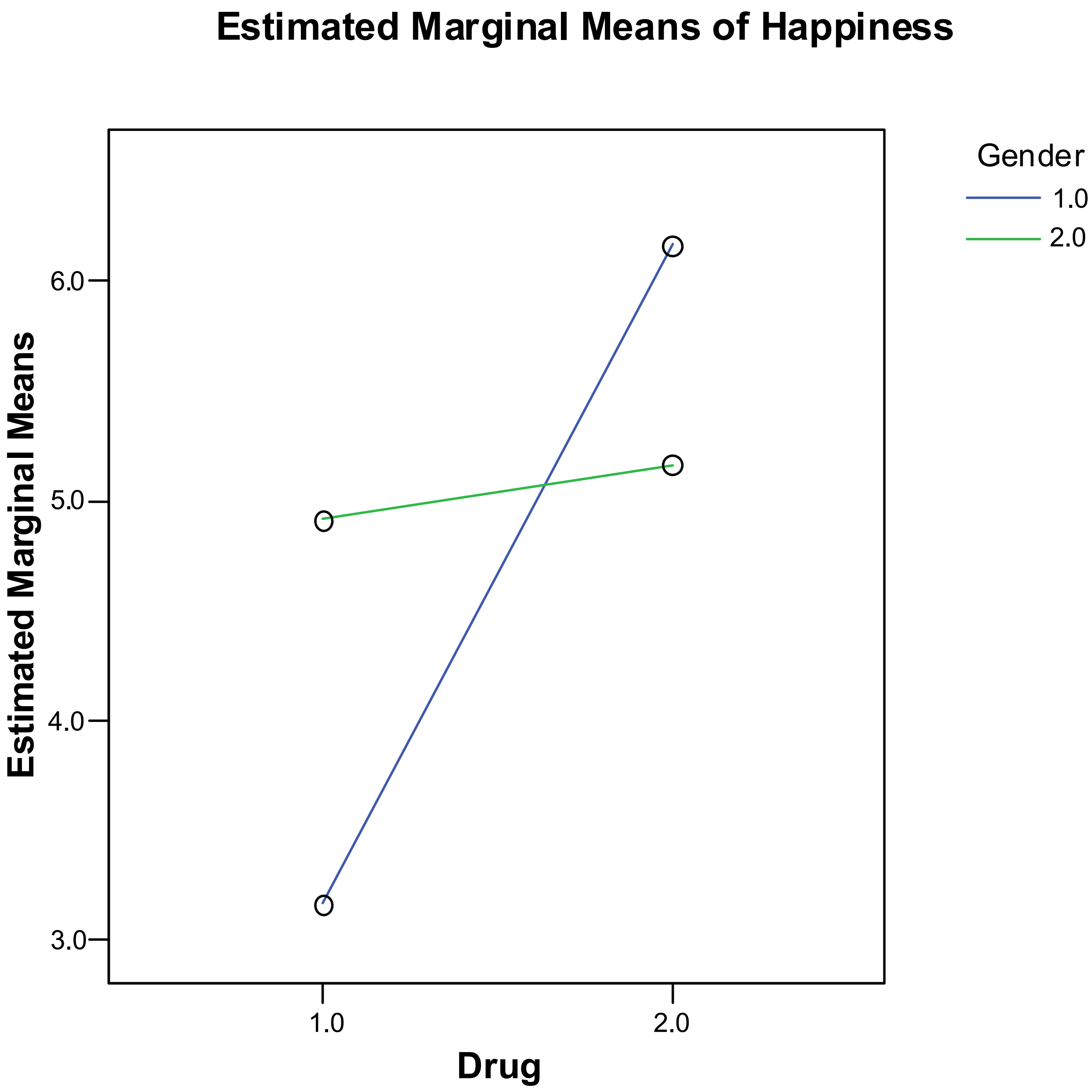
Tests of Between-Subjects Effects

Dependent Variable: Happiness

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	28.031[a]	3	9.344	14.705	.000
Intercept	565.510	1	565.510	889.984	.000
Drug	15.844	1	15.844	24.934	.000
Gender	.844	1	.844	1.328	.263
Drug * Gender	11.344	1	11.344	17.852	.000
Error	12.708	20	.635		
Total	606.250	24			
Corrected Total	40.740	23			

[a] R Squared = .688 (Adjusted R Squared = .641)

Interaction? Plot the means



Example: Conclusions

- Significant evidence that drug treatment affects happiness in depressed patients (p<0.001)
  - Prozac is effective, placebo is not
- No significant evidence that gender affects happiness (p=0.263)
- Significant evidence of an interaction between gender and treatment (p<0.001)
  - Prozac is effective in men but not in women!!*

Introduction to Linear Regression and Correlation Analysis

Goals

After this, you should be able to:

- Calculate and interpret the simple correlation between two variables
- Determine whether the correlation is significant
- Calculate and interpret the simple linear regression equation for a set of data
- Understand the assumptions behind regression analysis
- Determine whether a regression model is significant

Goals

(continued)

After this, you should be able to:

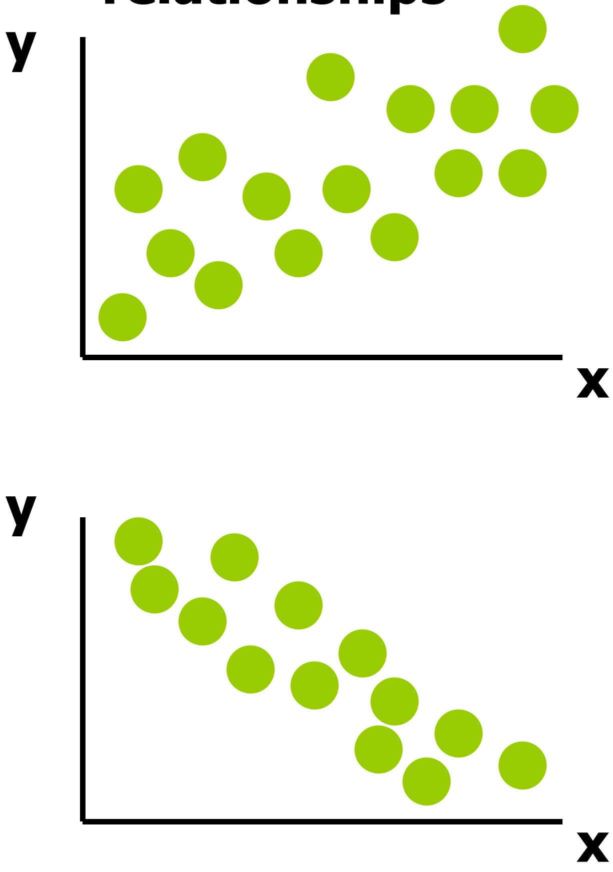
- Calculate and interpret confidence intervals for the regression coefficients
- Recognize regression analysis applications for purposes of prediction and description
- Recognize some potential problems if regression analysis is used incorrectly
- Recognize nonlinear relationships between two variables

Scatter Plots and Correlation

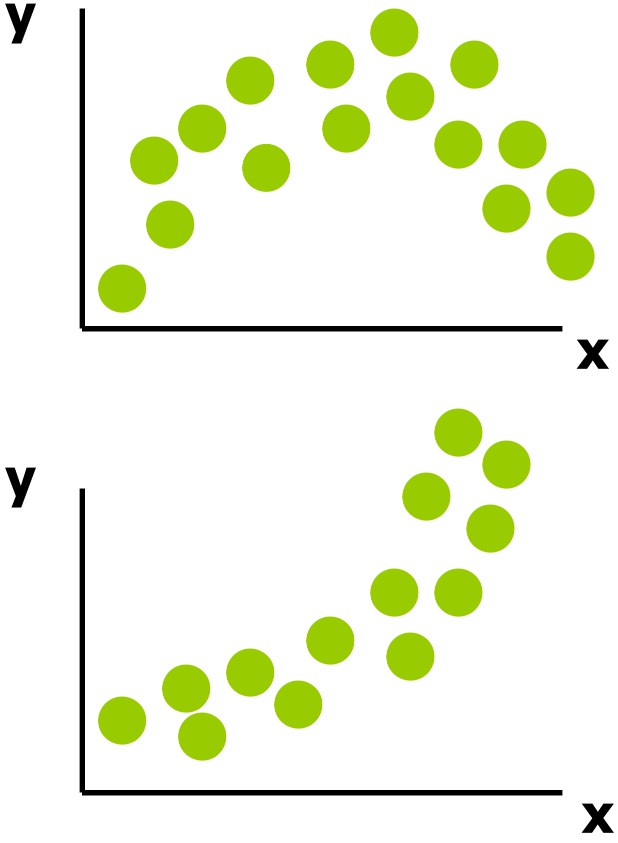
- A scatter plot (or scatter diagram) is used to show the relationship between two variables
- Correlation analysis is used to measure strength of the association (linear relationship) between two variables
  - Only concerned with strength of the relationship
  - No causal effect is implied

Scatter Plot Examples

Linear relationships



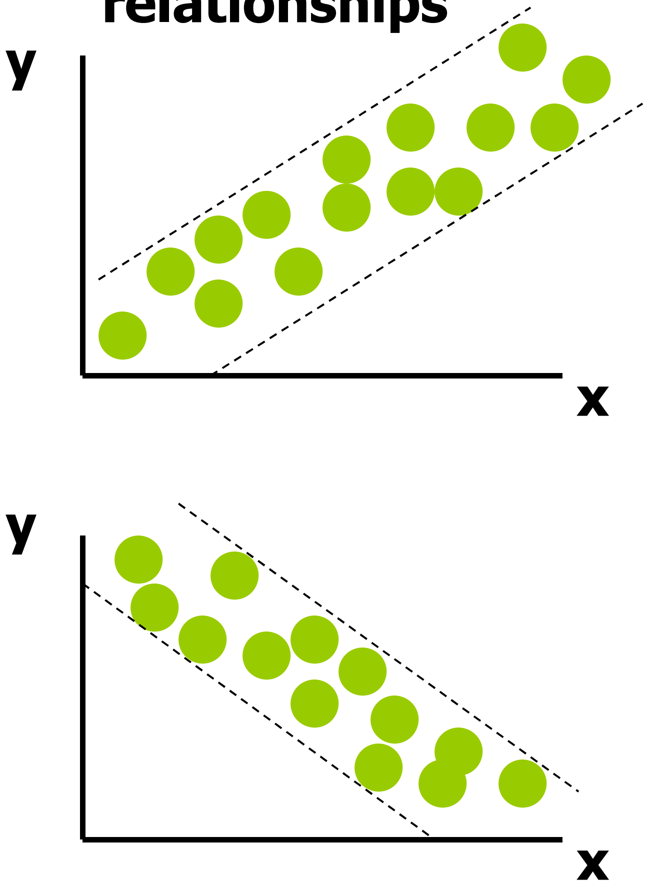
Curvilinear relationships



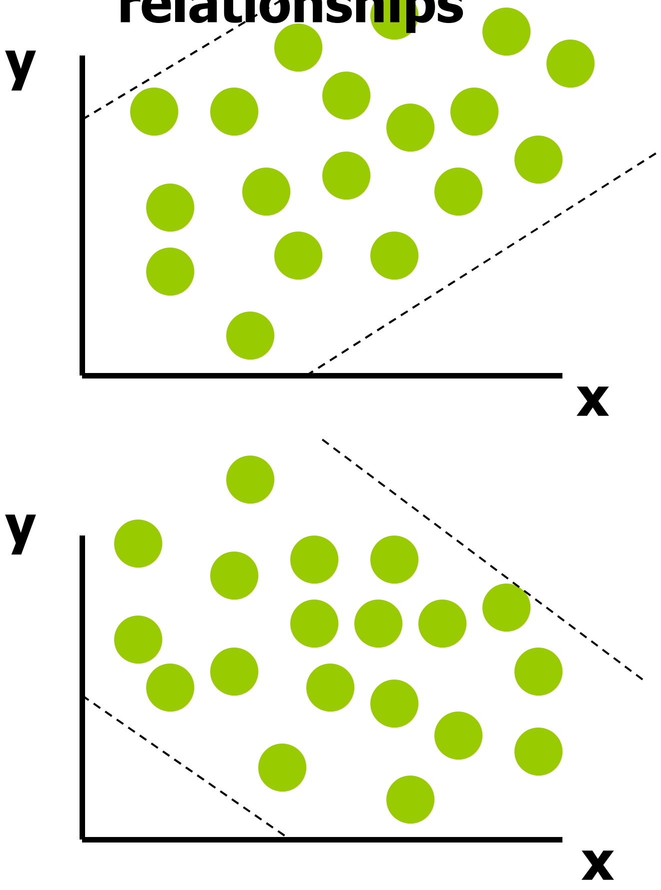
Scatter Plot Examples

(continued)

Strong relationships



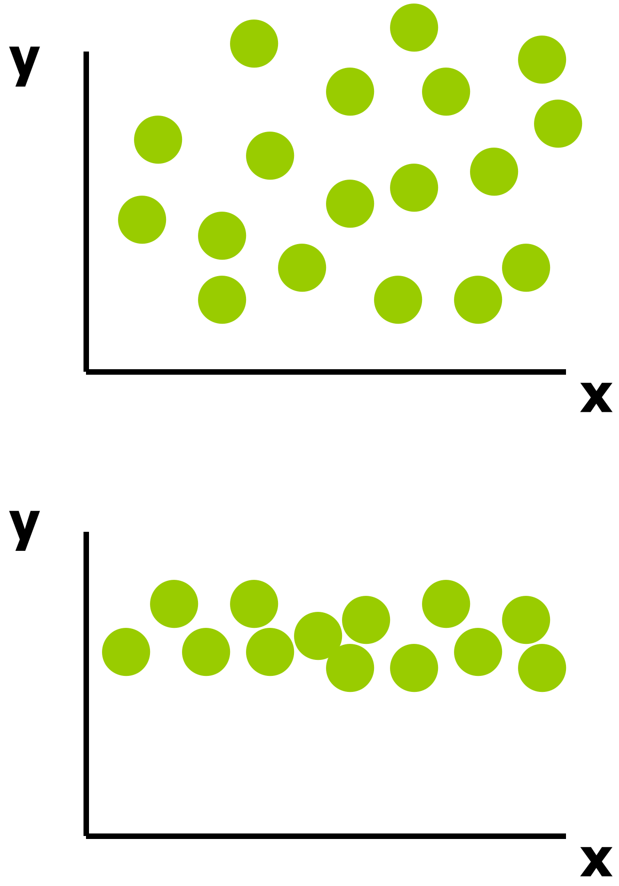
Weak relationships



Scatter Plot Examples

(continued)

No relationship



Correlation Coefficient

(continued)

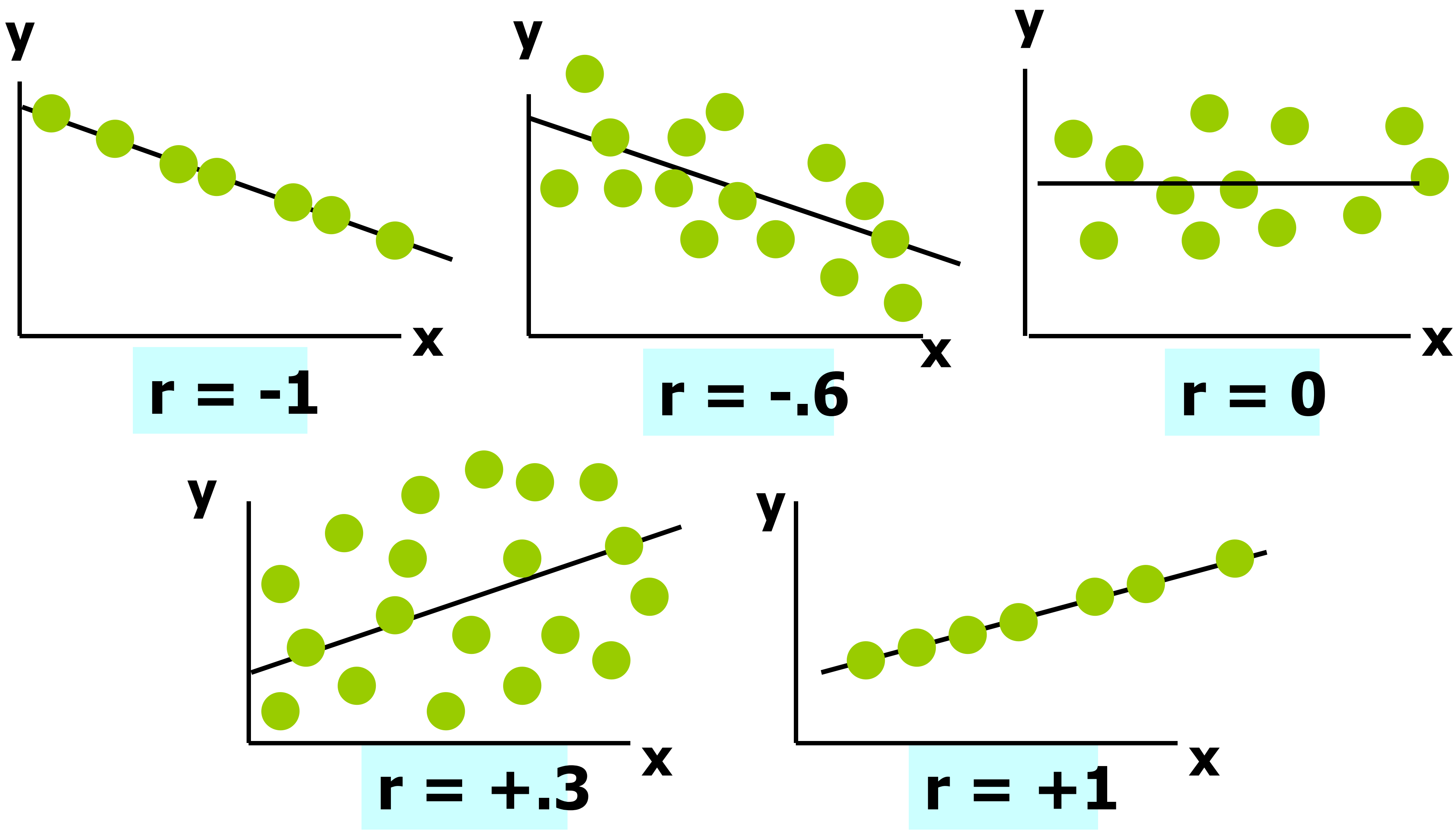
- The population correlation coefficient ρ (rho) measures the strength of the association between the variables

- The sample correlation coefficient r is an estimate of ρ and is used to measure the strength of the linear relationship in the sample observations

Features of ρ and r

- Unit free
- Range between -1 and 1
- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

Examples of Approximate r Values



Calculating the Correlation Coefficient

Sample correlation coefficient:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{[\sum (x - \bar{x})^2][\sum (y - \bar{y})^2]}}$$

or the algebraic equivalent:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

where:

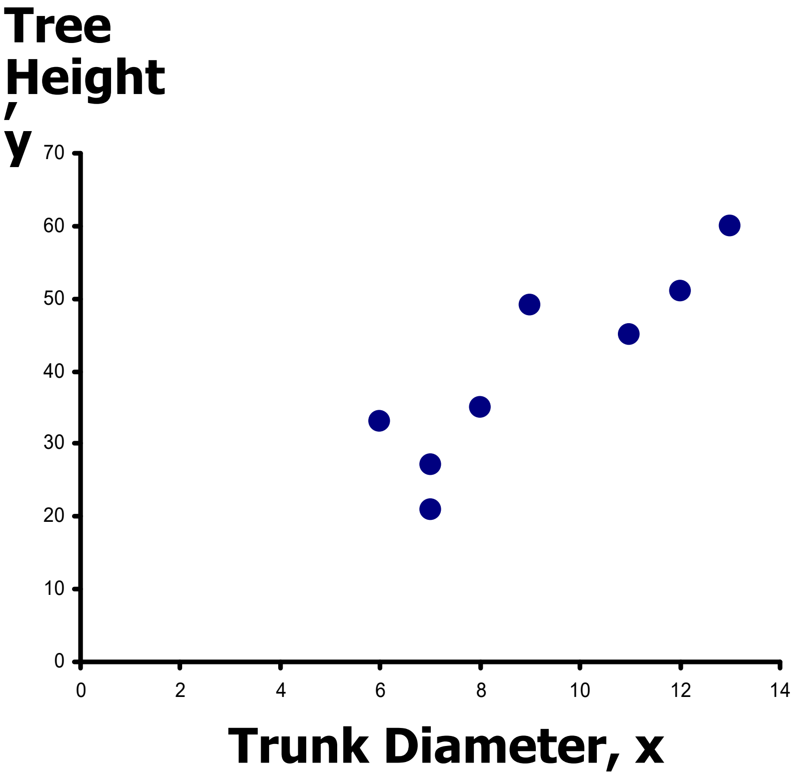
r = Sample correlation coefficient
n = Sample size
x = Value of the independent variable
y = Value of the dependent variable

Calculation Example

Tree Height	**Trunk Diameter**			
y	**x**	**xy**	**y^2**	**x^2**
35	8	280	1225	64
49	9	441	2401	81
27	7	189	729	49
33	6	198	1089	36
60	13	780	3600	169
21	7	147	441	49
45	11	495	2025	121
51	12	612	2601	144
Σ=321	**Σ=73**	**Σ=3142**	**Σ=14111**	**Σ=713**

Calculation Example

(continued)



$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]}}$$

$$= \frac{8(3142) - (73)(321)}{\sqrt{[8(713) - (73)^2][8(14111) - (321)^2]}}$$

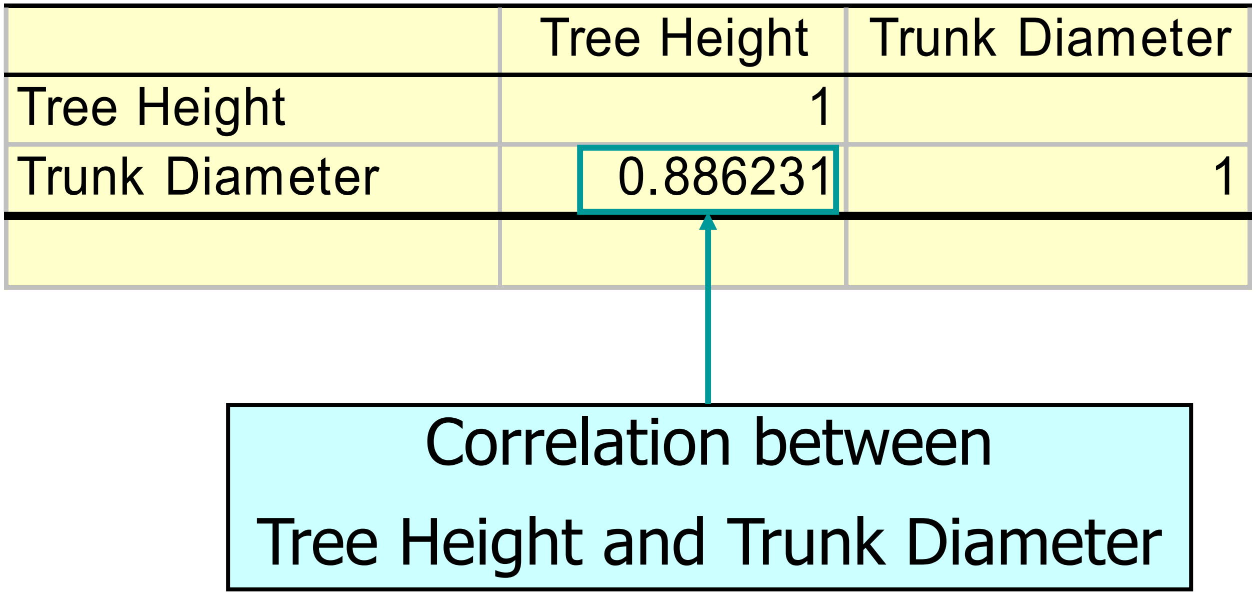
$$= 0.886$$

r = 0.886 → relatively strong positive linear association between x and y

Excel Output

Excel Correlation Output

Tools / data analysis / correlation…



	Tree Height	Trunk Diameter
Tree Height	1	
Trunk Diameter	0.886231	1

Correlation between

Tree Height and Trunk Diameter

Significance Test for Correlation

- Hypotheses

$$H_0: \rho = 0 \quad \text{(no correlation)}$$

$$H_A: \rho \neq 0 \quad \text{(correlation exists)}$$

- Test statistic
  - $$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

(with n – 2 degrees of freedom)

Example: Produce Stores

Is there evidence of a linear relationship between tree height and trunk diameter at the .05 level of significance?

H_0: ρ = 0   (No correlation)

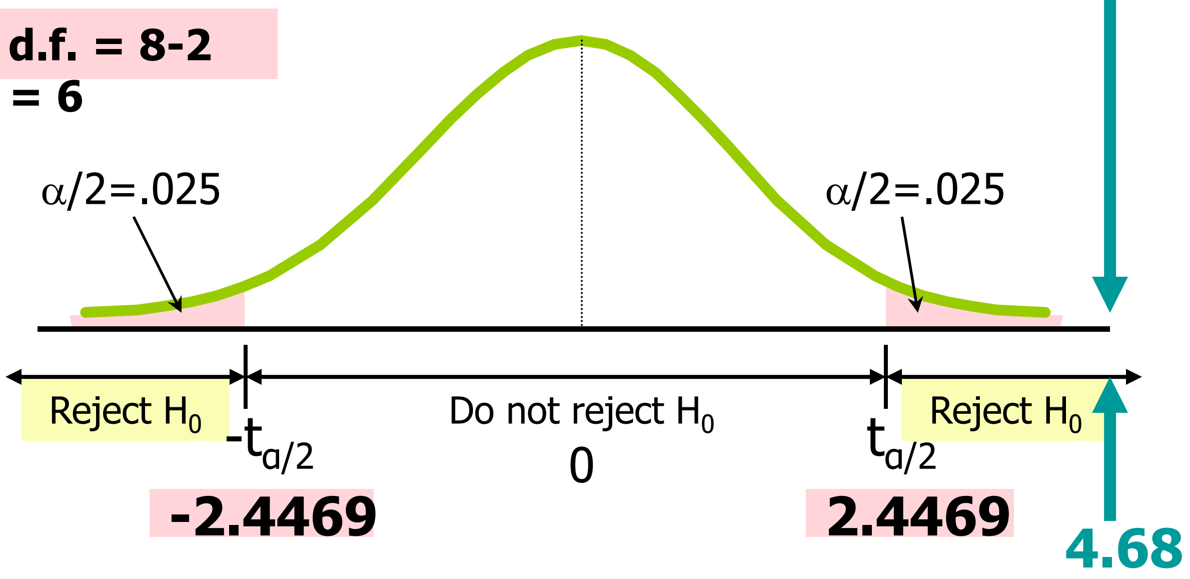
H_1: ρ ≠ 0   (correlation exists)

α =.05 ,   df = 8 - 2 = 6

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = 4.68$$

Example: Test Solution

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{.886}{\sqrt{\frac{1-.886^2}{8-2}}} = \boxed{4.68}$$



Decision:
Reject H_0

Conclusion:
There **is evidence** of a linear relationship at the 5% level of significance

Introduction to Regression Analysis

- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

Dependent variable: the variable we wish to explain

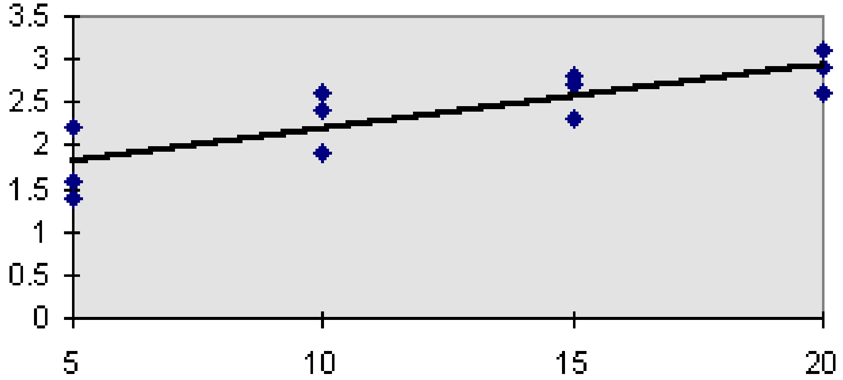
Independent variable: the variable used to explain the dependent variable

Simple Linear Regression Model

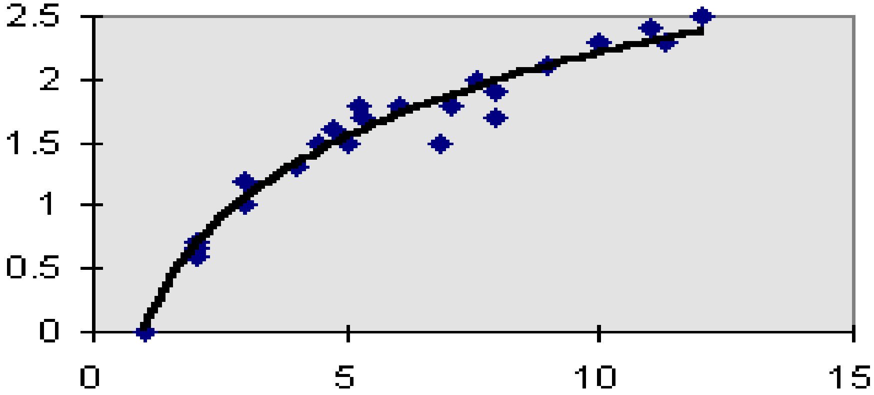
- Only **one** independent variable, x
- Relationship between x and y is described by a linear function
- Changes in y are assumed to be caused by changes in x

Types of Regression Models

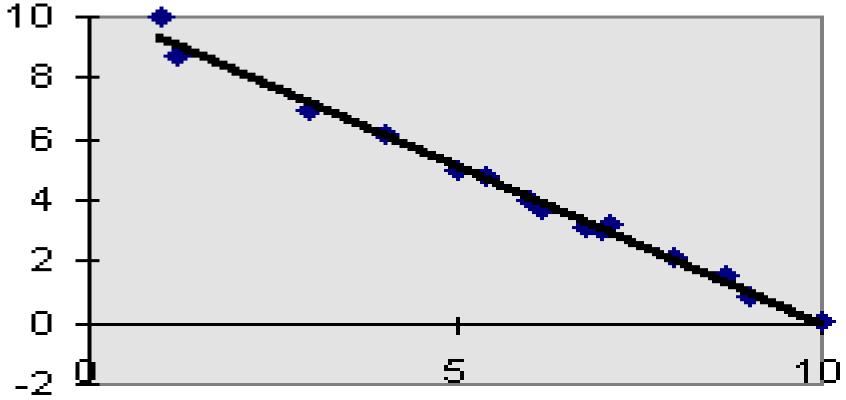
Positive Linear Relationship



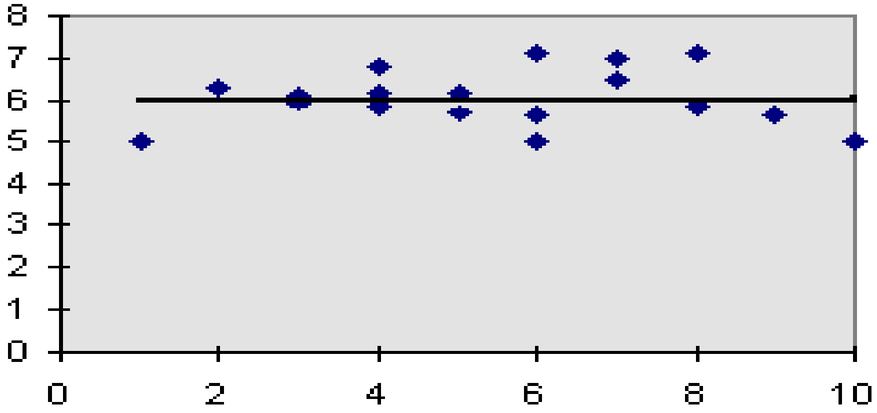
Relationship NOT Linear



Negative Linear Relationship



No Relationship



Population Linear Regression

The population regression model:

Dependent Variable

Population y intercept

Population Slope Coefficient

Independent Variable

Random Error term, or residual

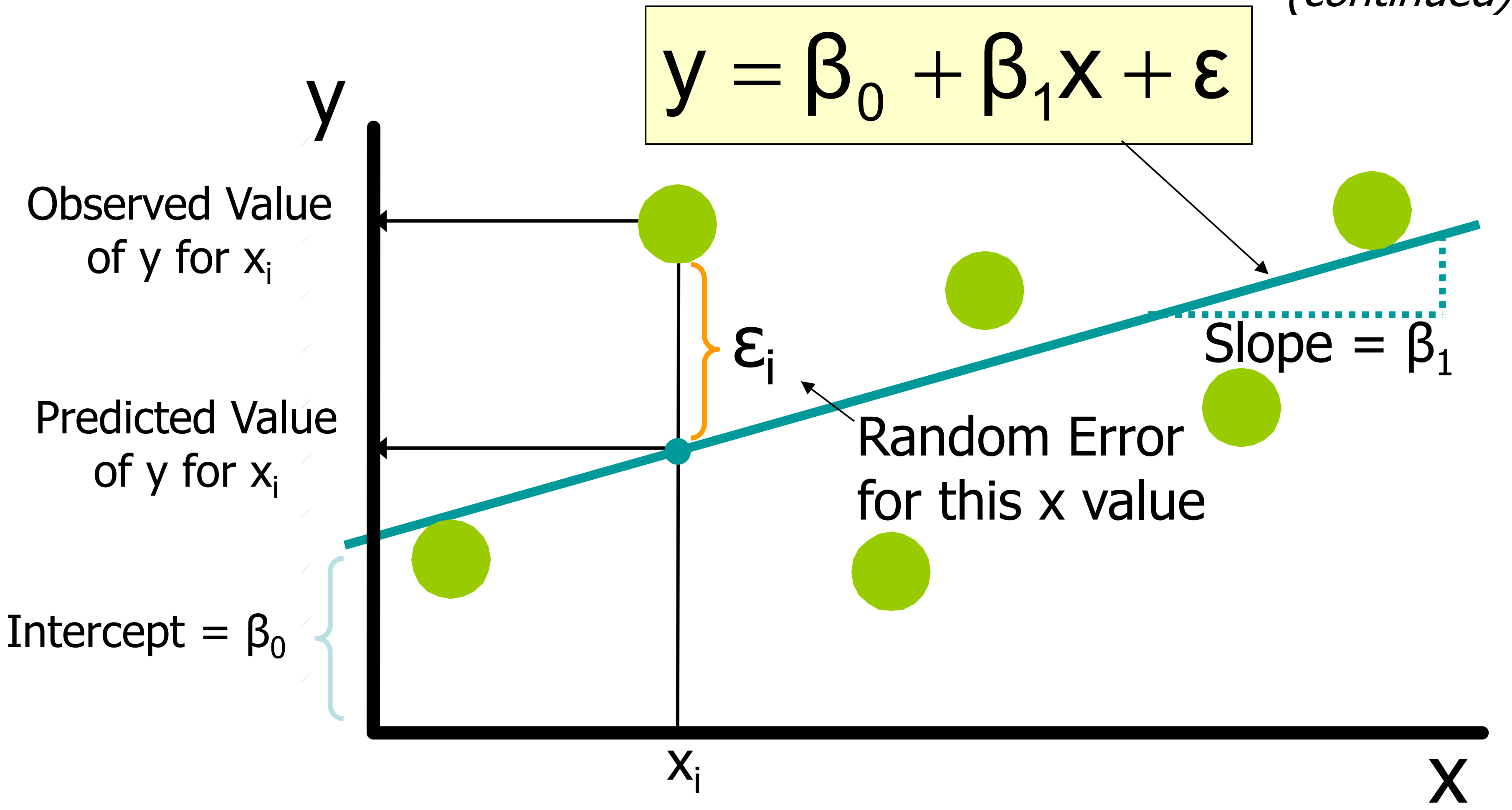
$$\boxed{y = \underbrace{\beta_0 + \beta_1 x}_{\text{Linear component}} + \underbrace{\varepsilon}_{\text{Random Error component}}}$$

Linear Regression Assumptions

- Error values (ε) are statistically independent
- Error values are normally distributed for any given value of x
- The probability distribution of the errors is normal
- The probability distribution of the errors has constant variance
- The underlying relationship between the x variable and the y variable is linear

Population Linear Regression

(continued)



$$y = \beta_0 + \beta_1 x + \varepsilon$$

y

Observed Value of y for x_i

ε_i

Predicted Value of y for x_i

Slope = β_1

Random Error for this x value

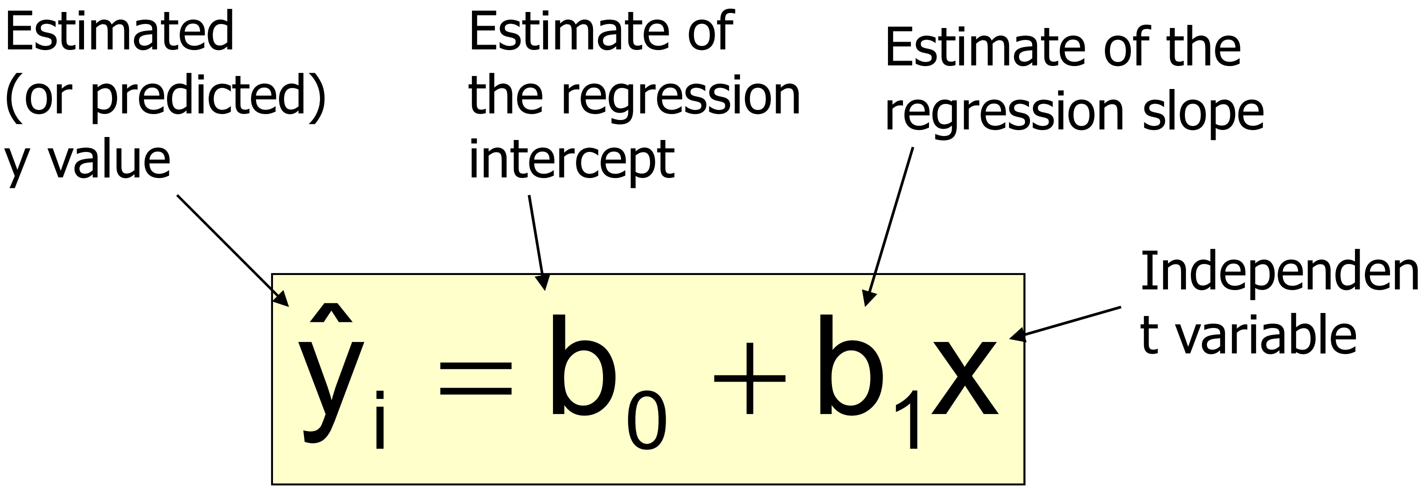
Intercept = β_0

x_i

x

Estimated Regression Model

The sample regression line provides an estimate of the population regression line



Estimated
(or predicted)
y value

Estimate of
the regression
intercept

Estimate of the
regression slope

Independen
t variable

$$\hat{y}_i = b_0 + b_1 x$$

The individual random error terms e_i have a mean of zero

Least Squares Criterion

- b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared residuals

$$\begin{aligned}\sum e^2 &= \sum (y - \hat{y})^2 \\ &= \sum (y - (b_0 + b_1 x))^2\end{aligned}$$

The Least Squares Equation

- The formulas for b_1 and b_0 are:

$$b_1 = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

algebraic
equi

$$b_1 = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sum x^2 - \frac{(\sum x)^2}{n}}$$

and

$$b_0 = \bar{y} - b_1 \bar{x}$$

Interpretation of the Slope and the Intercept

- b_0 is the estimated average value of y when the value of x is zero

- b_1 is the estimated change in the average value of y as a result of a one-unit change in x

Finding the Least Squares Equation

- The coefficients b_0 and b_1 will usually be found using computer software, such as Excel or Minitab

- Other regression measures will also be computed as part of computer-based regression analysis

Simple Linear Regression Example

- A real estate agent wishes to examine the relationship between the selling price of a home and its size (measured in square feet)

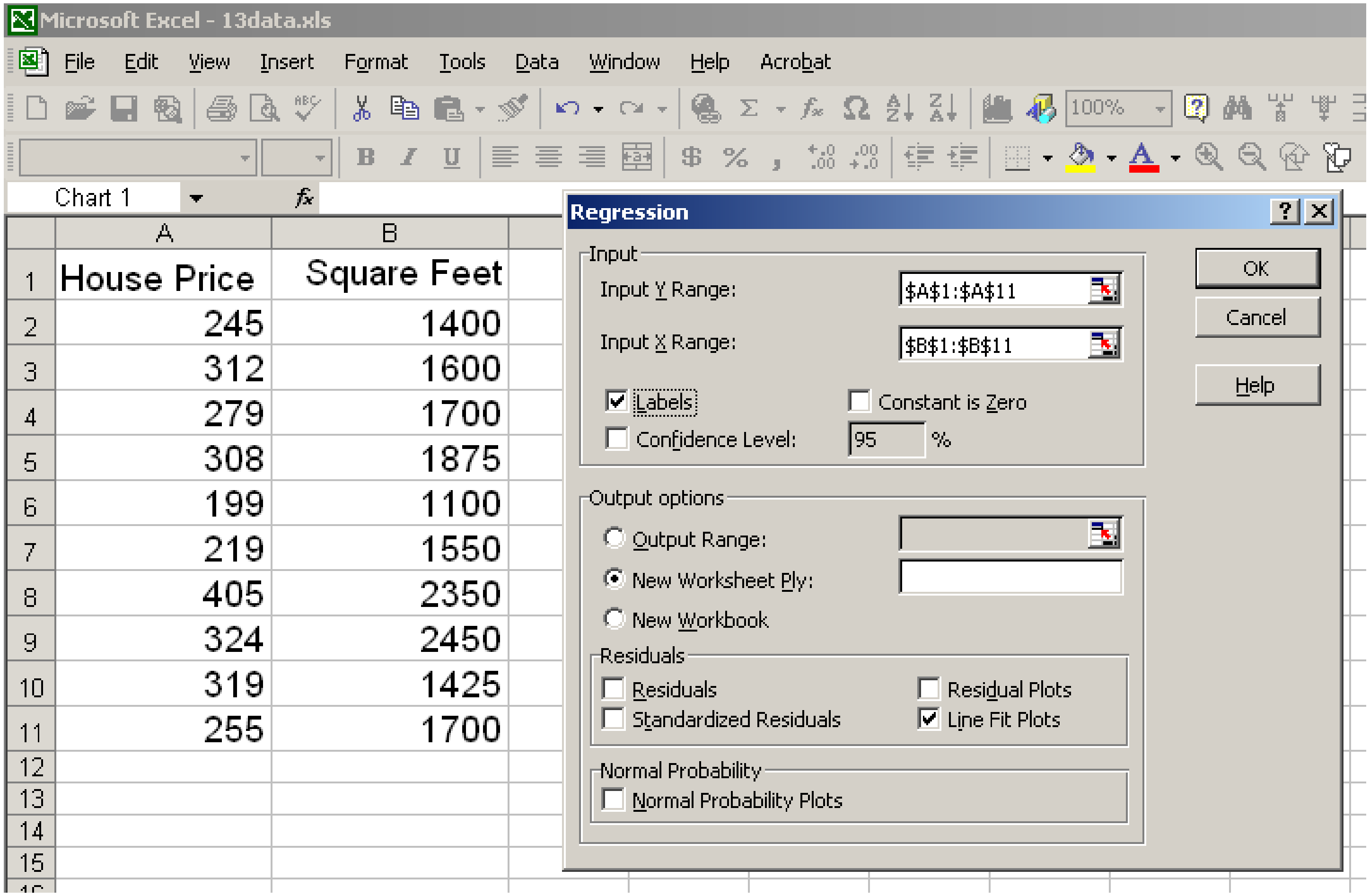
- A random sample of 10 houses is selected
  - Dependent variable (y) = house price in $1000s
  - Independent variable (x) = square feet

Sample Data for House Price Model

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Regression Using Excel

- Tools / Data Analysis / Regression


Microsoft Excel - 13data.xls
File Edit View Insert Format Tools Data Window Help Acrobat
100%
Chart 1
fx

	A	B
1	House Price	Square Feet
2	245	1400
3	312	1600
4	279	1700
5	308	1875
6	199	1100
7	219	1550
8	405	2350
9	324	2450
10	319	1425
11	255	1700
12		
13		
14		
15		

Regression

Input

Input Y Range: $A$1:$A$11

Input X Range: $B$1:$B$11

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☐ Residuals ☐ Residual Plots

☐ Standardized Residuals ☑ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

Excel Output

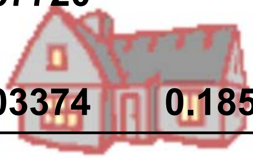
Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

The regression equation is:

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

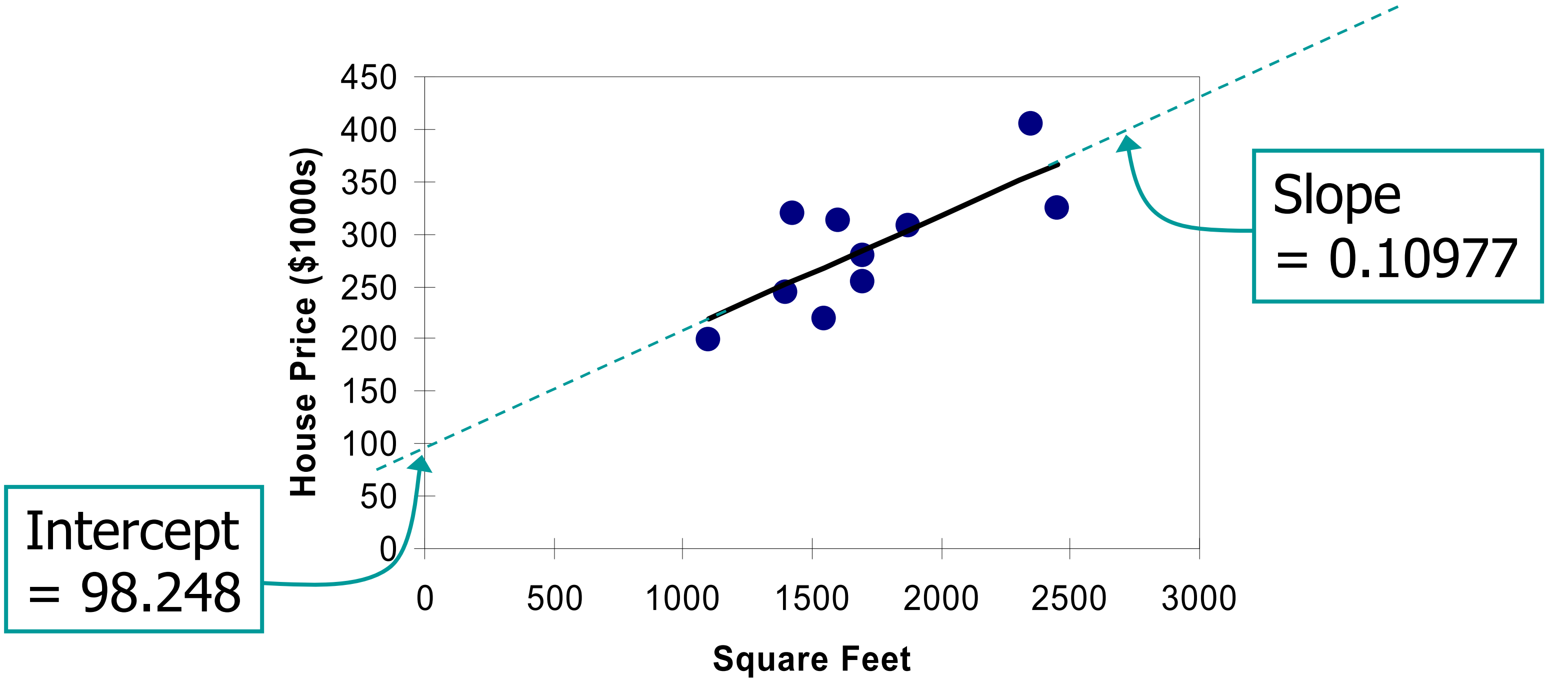
ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580



Graphical Presentation

- House price model: scatter plot and regression line



Intercept = 98.248

Slope = 0.10977

$$\widehat{\text{house price}} = 98.24833 + 0.10977\ (\text{square feet})$$

Interpretation of the Intercept, b_0

$$\widehat{\text{house price}} = \boxed{98.24833} + 0.10977\ (\text{square feet})$$

- b_0 is the estimated average value of Y when the value of X is zero (if x = 0 is in the range of observed x values)
  - Here, no houses had 0 square feet, so b_0 = 98.24833 just indicates that, for houses within the range of sizes observed, $98,248.33 is the portion of the house price not explained by square feet

Interpretation of the Slope Coefficient, b_1

$$\widehat{\text{house price}} = 98.24833 + \boxed{0.10977}(\text{square feet})$$

- b_1 measures the estimated change in the average value of Y as a result of a one-unit change in X
  - Here, b_1 = .10977 tells us that the average value of a house increases by .10977($1000) = $109.77, on average, for each additional one square foot of size

Least Squares Regression Properties

- The sum of the residuals from the least squares regression line is 0 ($\sum (y - \hat{y}) = 0$)
- The sum of the squared residuals is a minimum (minimized $\sum (y - \hat{y})^2$)
- The simple regression line always passes through the mean of the y variable and the mean of the x variable
- The least squares coefficients are unbiased estimates of β_0 and β_1

Explained and Unexplained Variation

- Total variation is made up of two parts:

$$\text{SST} = \text{SSE} + \text{SSR}$$

Total sum of Squares

Sum of Squares Error

Sum of Squares Regression

$$\text{SST} = \sum (y - \bar{y})^2$$

$$\text{SSE} = \sum (y - \hat{y})^2$$

$$\text{SSR} = \sum (\hat{y} - \bar{y})^2$$

where:

$\bar{y}$ = Average value of the dependent variable

y = Observed values of the dependent variable

$\hat{y}$ = Estimated value of y for the given x value

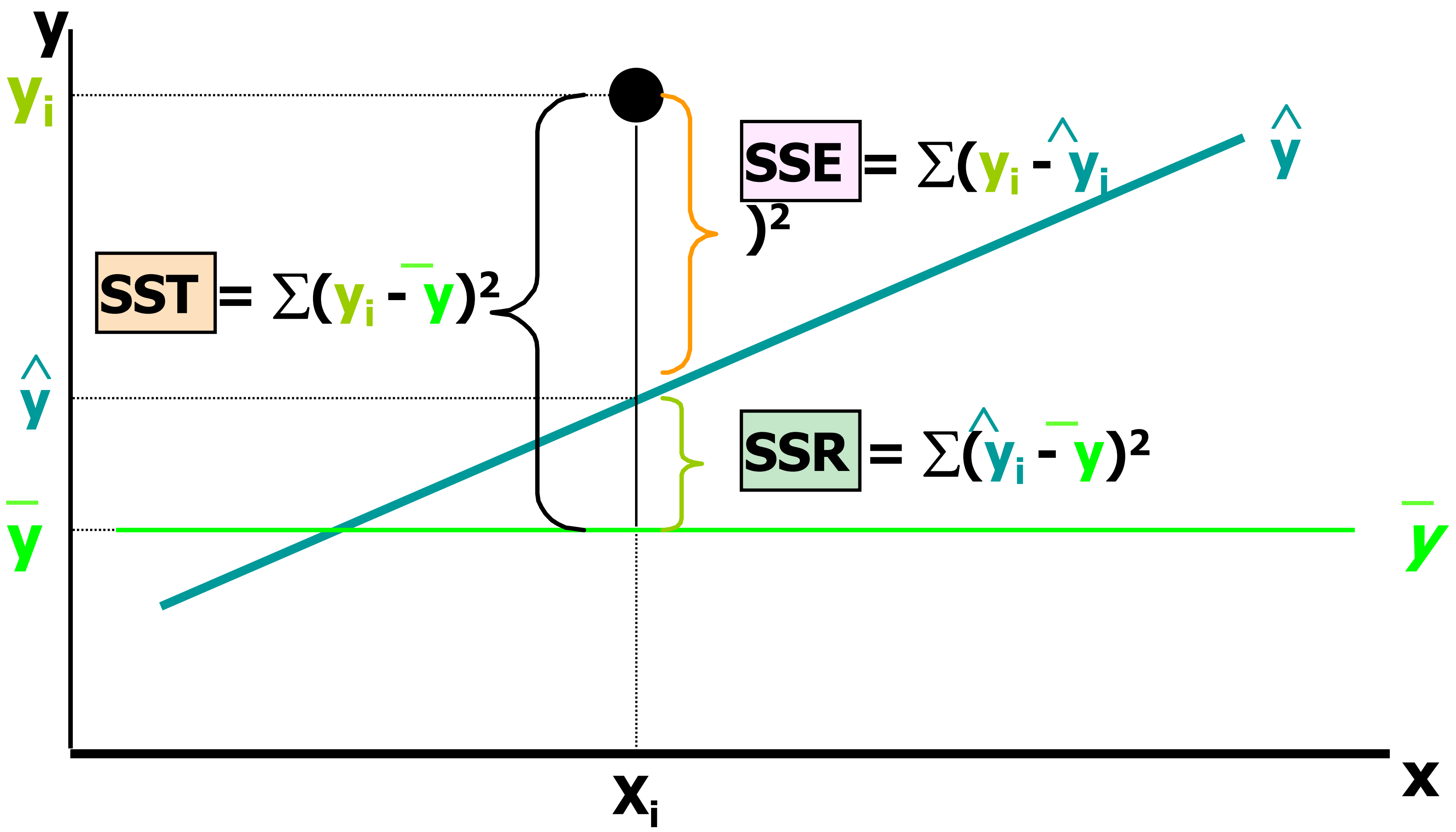
Explained and Unexplained Variation

(continued)

- SST = total sum of squares
  - Measures the variation of the y_i values around their mean y
- SSE = error sum of squares
  - Variation attributable to factors other than the relationship between x and y
- SSR = regression sum of squares
  - Explained variation attributable to the relationship between x and y

Explained and Unexplained Variation

(continued)



SST $= \Sigma(y_i - \bar{y})^2$

SSE $= \Sigma(y_i - \hat{y}_i)^2$

SSR $= \Sigma(\hat{y}_i - \bar{y})^2$

Coefficient of Determination, R^2

- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable
- The coefficient of determination is also called R-squared and is denoted as R^2

$$\boxed{R^2 = \frac{SSR}{SST}} \quad \text{where} \quad \boxed{0 \le R^2 \le 1}$$

Coefficient of Determination, R^2

(continued)

Coefficient of determination

$$\boxed{R^2 = \frac{SSR}{SST} = \frac{\text{sum of squares explained by regression}}{\text{total sum of squares}}}$$

Note: In the single independent variable case, the coefficient of determination is

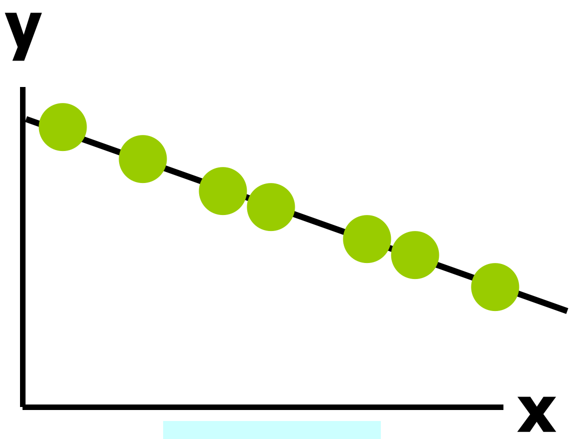
$$\boxed{R^2 = r^2}$$

where:

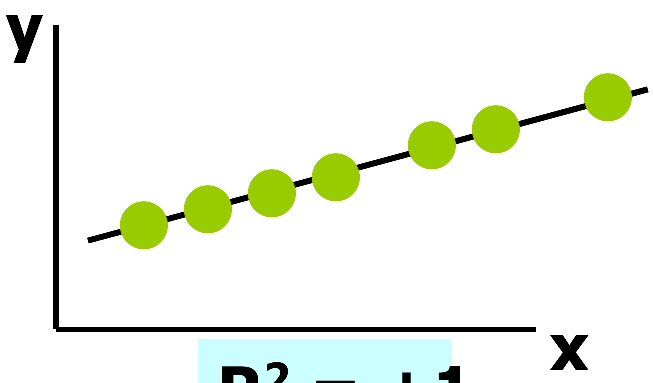
R^2 = Coefficient of determination

r = Simple correlation coefficient

Examples of Approximate R^2 Values



$R^2 = 1$



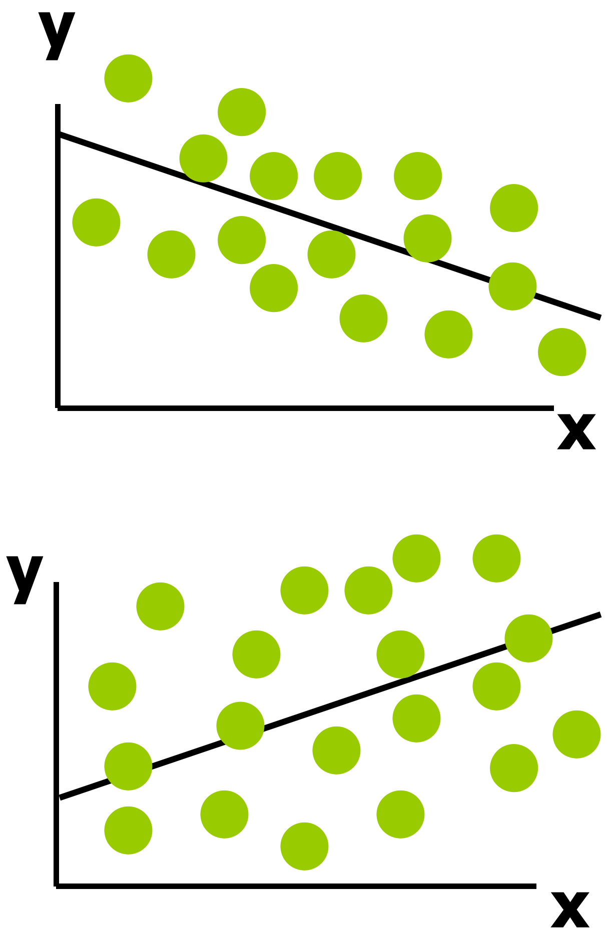
$R^2 = +1$

$R^2 = 1$

Perfect linear relationship between x and y:

100% of the variation in y is explained by variation in x

Examples of Approximate R^2 Values

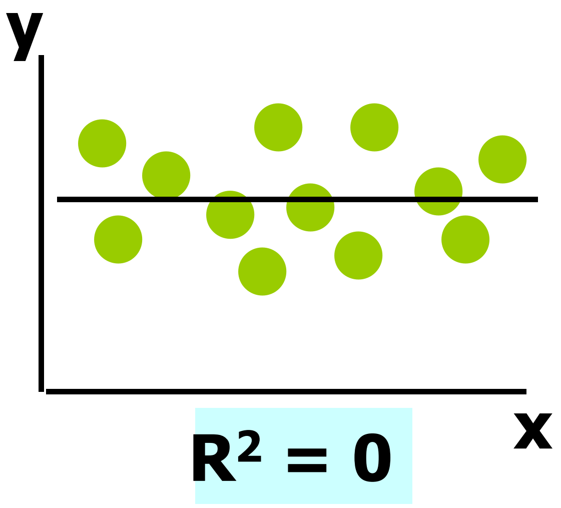


$0 < R^2 < 1$

Weaker linear relationship between x and y:

Some but not all of the variation in y is explained by variation in x

Examples of Approximate R^2 Values



$R^2 = 0$

No linear relationship between x and y:

The value of Y does not depend on x. (None of the variation in y is explained by variation in x)

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$R^2 = \frac{SSR}{SST} = \frac{18934.9348}{32600.5000} = 0.58082$$

58.08% of the variation in house prices is explained by variation in square feet

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$\boxed{s_{\varepsilon} = \sqrt{\frac{SSE}{n-k-1}}}$$

Where

SSE = Sum of squares error

n = Sample size

k = number of independent variables in the model

The Standard Deviation of the Regression Slope

- The standard error of the regression slope coefficient (b_1) is estimated by

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

where:

s_{b_1} = Estimate of the standard error of the least squares slope

$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ = Sample standard error of the estimate

Excel Output

Regression Statistics	
Multiple R	0.76211
R Square	0.58082
Adjusted R Square	0.52842
Standard Error	41.33032
Observations	10

$$s_{\varepsilon} = 41.33032$$

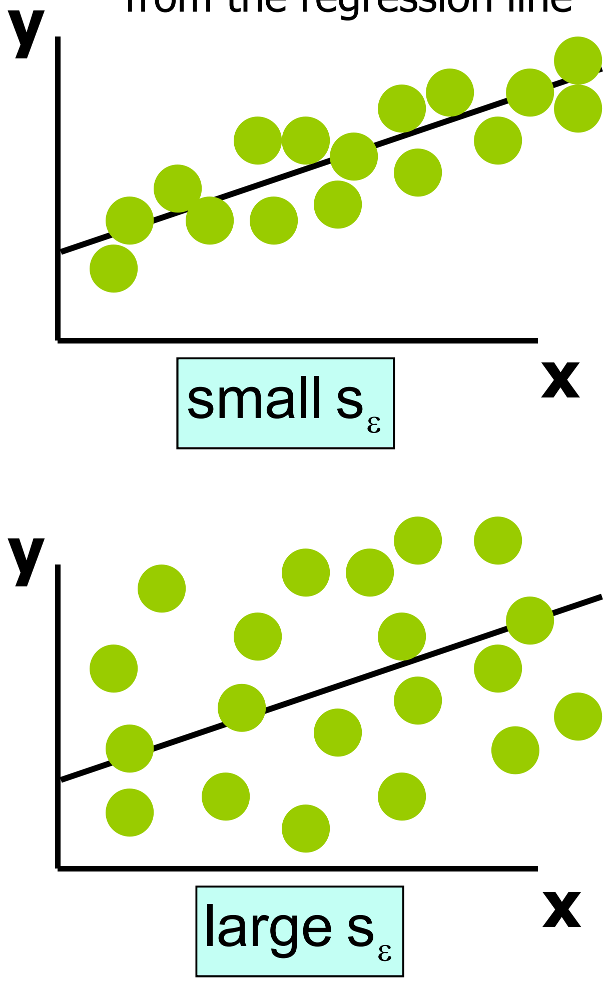
$$s_{b_1} = 0.03297$$

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	18934.9348	18934.9348	11.0848	0.01039
Residual	8	13665.5652	1708.1957		
Total	9	32600.5000			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	98.24833	58.03348	1.69296	0.12892	-35.57720	232.07386
Square Feet	0.10977	0.03297	3.32938	0.01039	0.03374	0.18580

Comparing Standard Errors

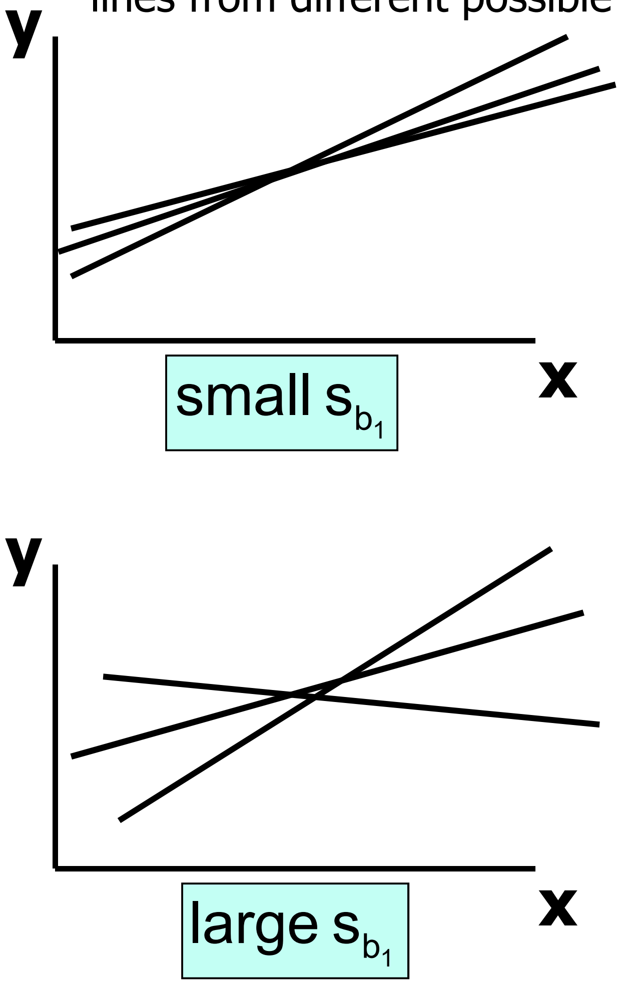
Variation of observed y values from the regression line



small s_ε

large s_ε

Variation in the slope of regression lines from different possible samples



small s_{b_1}

large s_{b_1}

Inference about the Slope: t Test

- t test for a population slope
  - Is there a linear relationship between x and y?
- Null and alternative hypotheses
  - H_0: $\beta_1 = 0$ (no linear relationship)
  - H_1: $\beta_1 \neq 0$ (linear relationship does exist)
- Test statistic
  -
  -

$$\boxed{t = \frac{b_1 - \beta_1}{s_{b_1}}}$$

$$d.f. = n - 2$$

where:

b_1 = Sample regression slope coefficient

β_1 = Hypothesized slope

s_{b1} = Estimator of the standard error of the slope

Inference about the Slope: t Test

(continued)

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\text{house price} = 98.25 + 0.1098\,(\text{sq.ft.})$$

The slope of this model is 0.1098

Does square footage of the house affect its sales price?

Inferences about the Slope: t Test Example

Test Statistic: **t = 3.329**

H_0: β_1 = 0

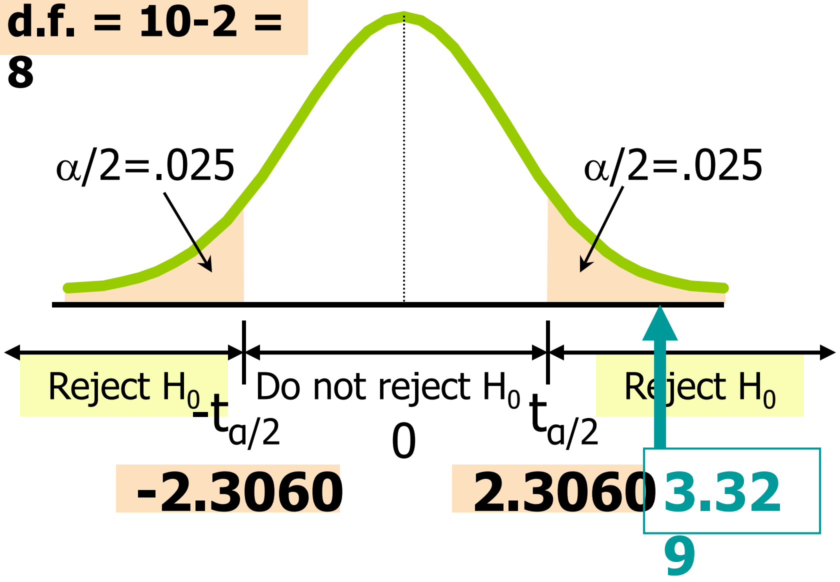
H_A: $\beta_1 \neq 0$

From Excel output:

b_1 s_{b_1} t

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**

d.f. = 10-2 = 8



Decision:

Reject H_0

Conclusion:

There is sufficient evidence that square footage affects house price

Regression Analysis for Description

Confidence Interval Estimate of the Slope:

$$b_1 \pm t_{\alpha/2} s_{b_1}$$

d.f. = n - 2

Excel Printout for House Prices:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

At 95% level of confidence, the confidence interval for the slope is (0.0337, 0.1858)

Regression Analysis for Description

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**98.24833**	**58.03348**	**1.69296**	**0.12892**	**-35.57720**	**232.07386**
Square Feet	**0.10977**	**0.03297**	**3.32938**	**0.01039**	**0.03374**	**0.18580**

Since the units of the house price variable is $1000s, we are 95% confident that the average impact on sales price is between $33.70 and $185.80 per square foot of house size

This 95% confidence interval does not include 0.

Conclusion: There is a significant relationship between house price and square feet at the .05 level of significance

Confidence Interval for the Average y, Given x

Confidence interval estimate for the **mean of y** given a particular x_p

Size of interval varies according to distance away from mean, $\bar{x}$

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{\overbrace{(x_p - \bar{x})^2}}{\sum (x - \bar{x})^2}}$$

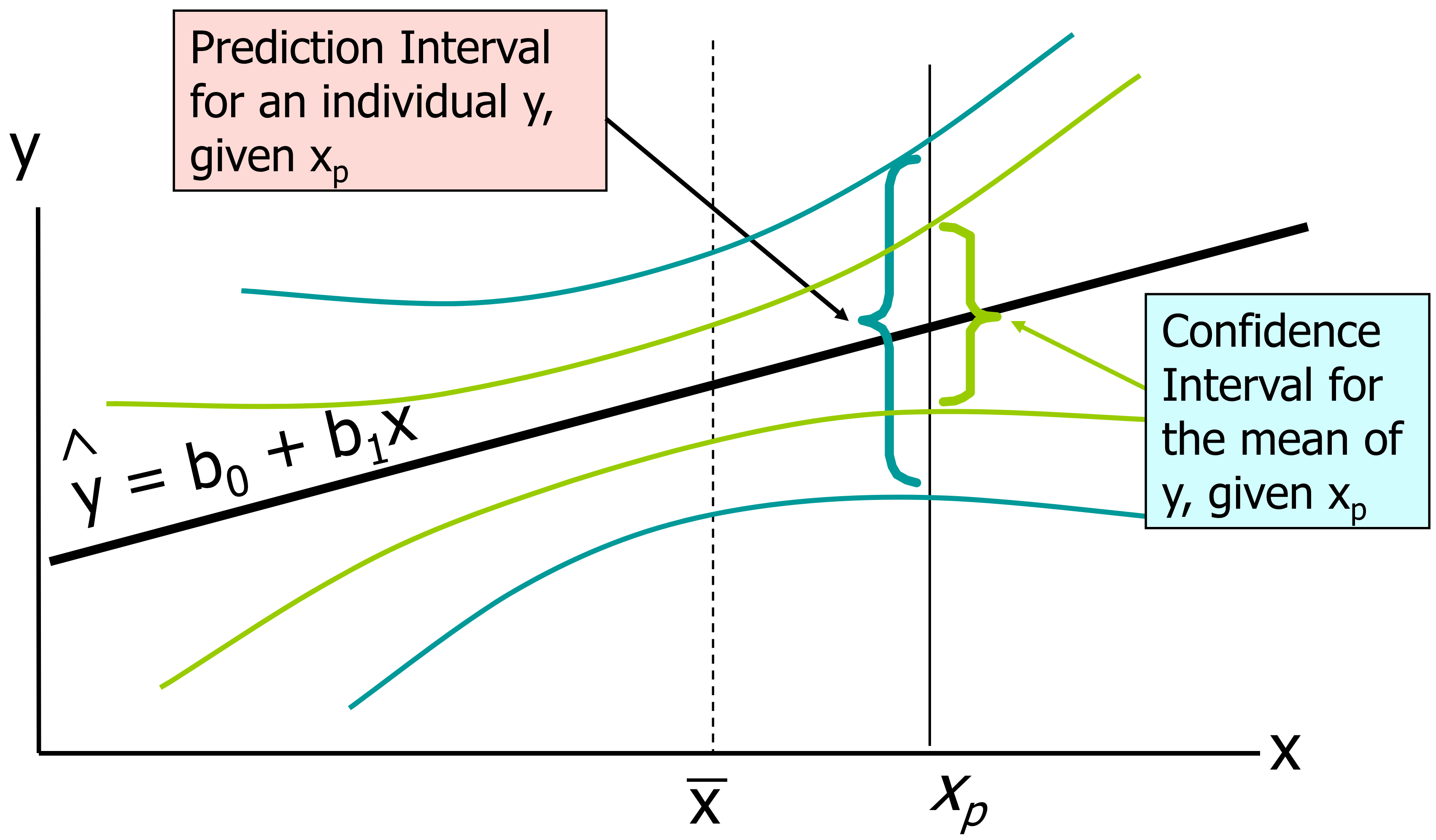
Confidence Interval for an Individual y, Given x

Confidence interval estimate for an **Individual value of y** given a particular x_p

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

This extra term adds to the interval width to reflect the added uncertainty for an individual case

Interval Estimates for Different Values of x



Prediction Interval for an individual y, given x_p

Confidence Interval for the mean of y, given x_p

Example: House Prices

House Price in $1000s (y)	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Estimated Regression Equation:

$$\text{house price} = 98.25 + 0.1098\,(\text{sq.ft.})$$

Predict the price for a house with 2000 square feet

Example: House Prices

(continued)

Predict the price for a house with 2000 square feet:

$$\widehat{\text{house price}} = 98.25 + 0.1098\,(\text{sq.ft.})$$

$$= 98.25 + 0.1098(2000)$$

$$= 317.85$$

The predicted price for a house with 2000 square feet is 317.85($1,000s) = $317,850

Estimation of Mean Values: Example

Confidence Interval Estimate for $E(y)|x_p$

Find the 95% confidence interval for the average price of 2,000 square-foot houses

Predicted Price $\hat{Y}_i$ = 317.85 ($1,000s)

$$\hat{y} \pm t_{\alpha/2} s_{\varepsilon} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}} = 317.85 \pm 37.12$$

The confidence interval endpoints are 280.66 -- 354.90, or from $280,660 -- $354,900

Estimation of Individual Values: Example

Prediction Interval Estimate for $y|x_p$

Find the 95% confidence interval for an individual house with 2,000 square feet

Predicted Price $\hat{Y}_i$ = 317.85 ($1,000s)

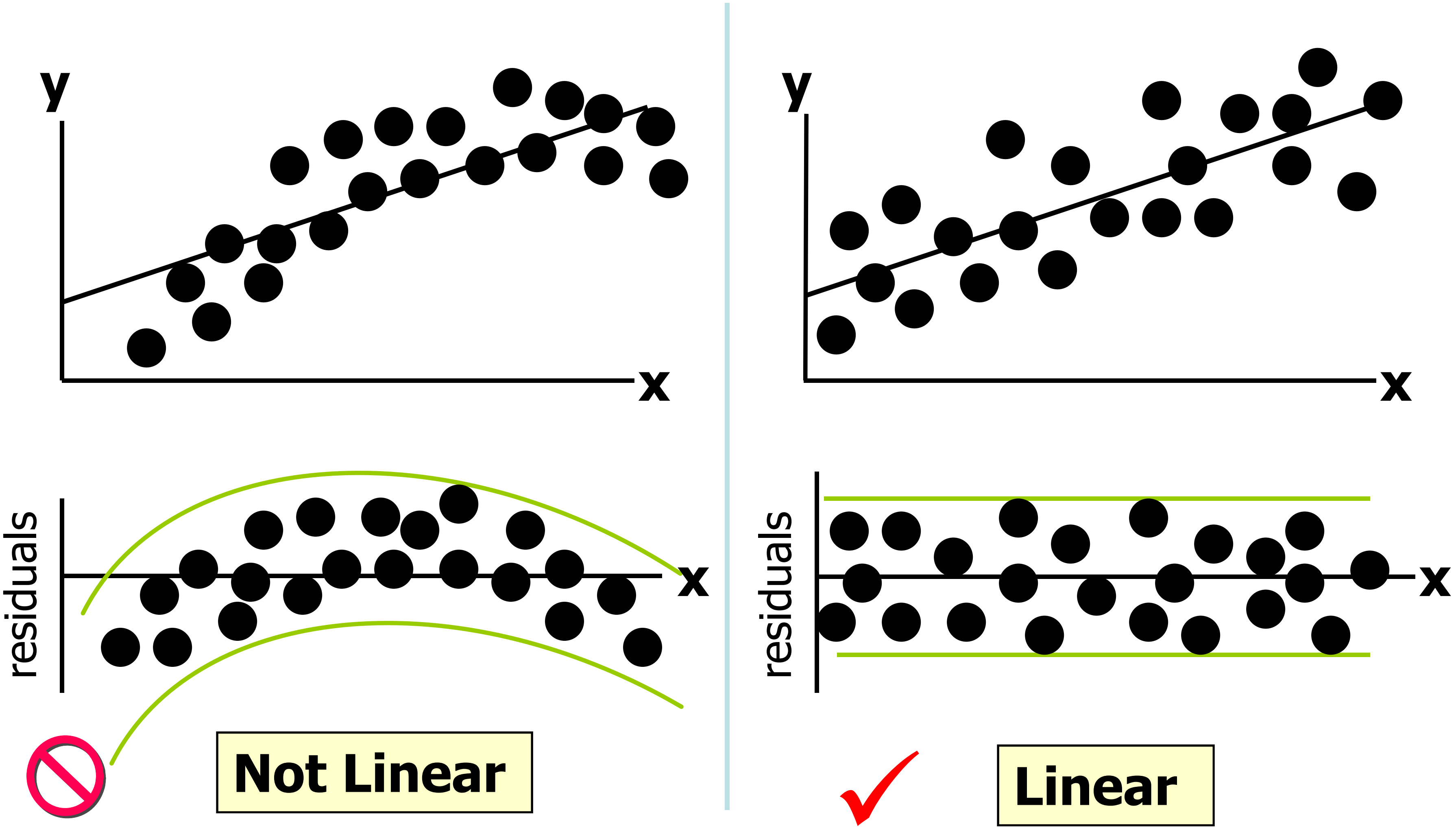
$$\hat{y} \pm t_{\alpha/2} s_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}} = 317.85 \pm 102.28$$

The prediction interval endpoints are 215.50 -- 420.07, or from $215,500 -- $420,070

Residual Analysis

- Purposes
  - Examine for linearity assumption
  - Examine for constant variance for all levels of x
  - Evaluate normal distribution assumption
- Graphical Analysis of Residuals
  - Can plot residuals vs. x
  - Can create histogram of residuals to check for normality

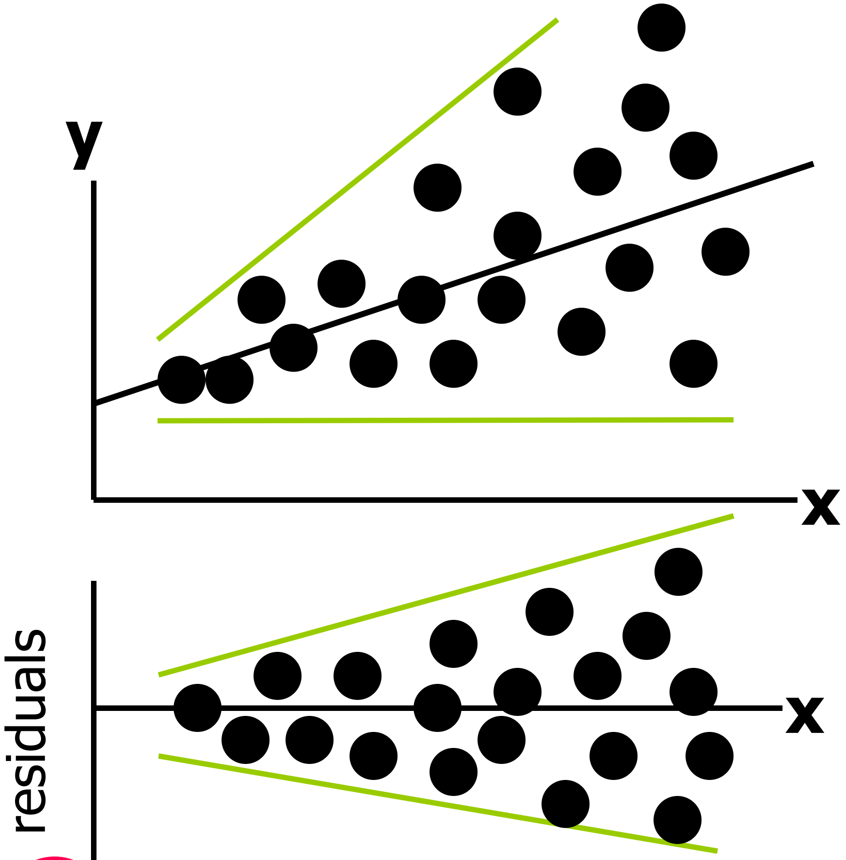
Residual Analysis for Linearity



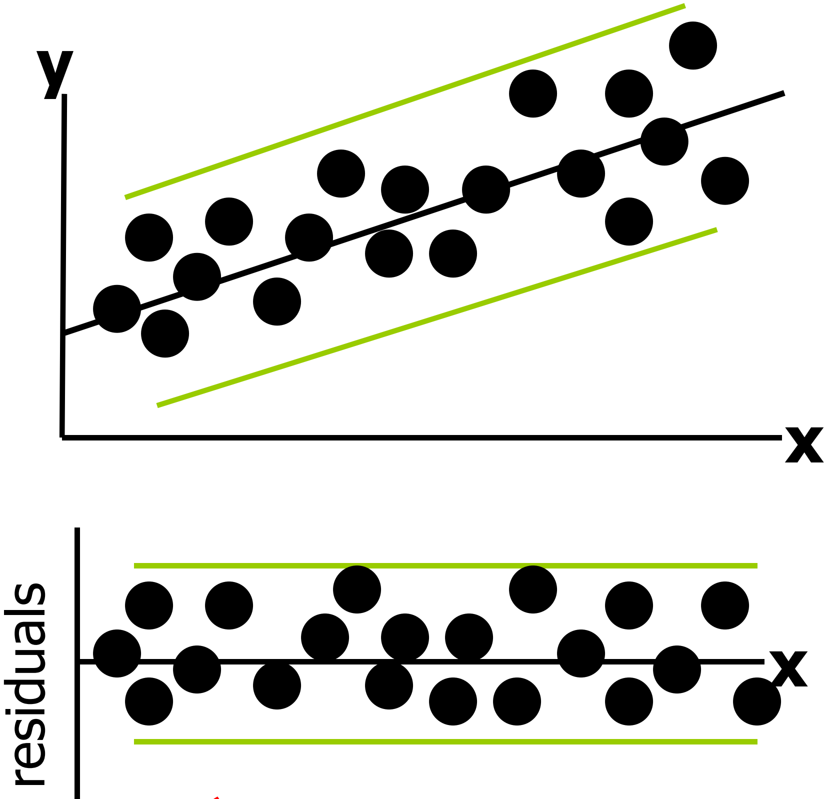
Not Linear

✓ Linear

Residual Analysis for Constant Variance

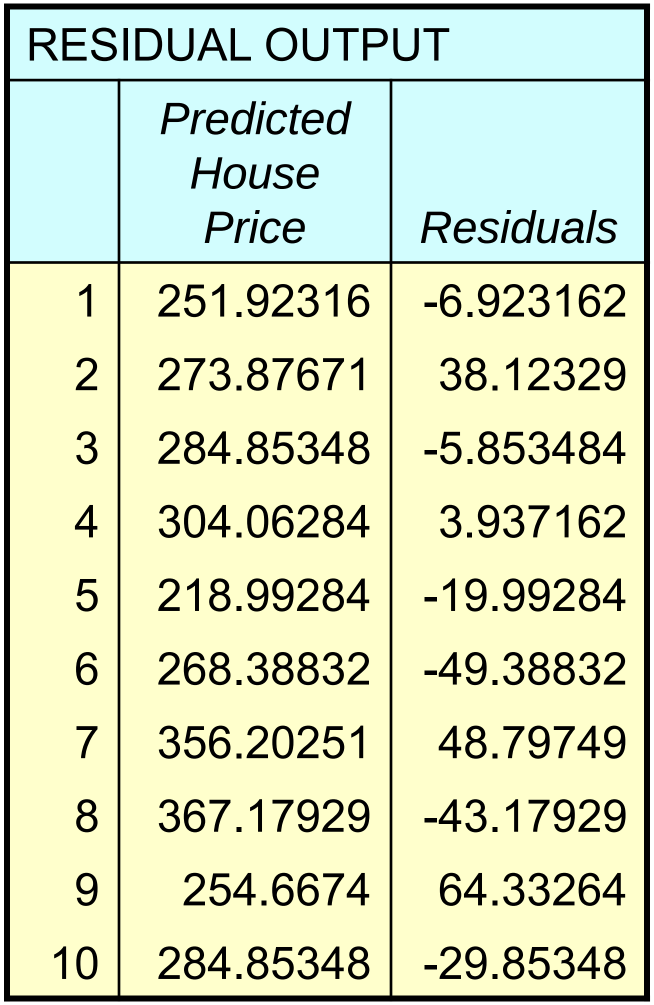


Non-constant variance



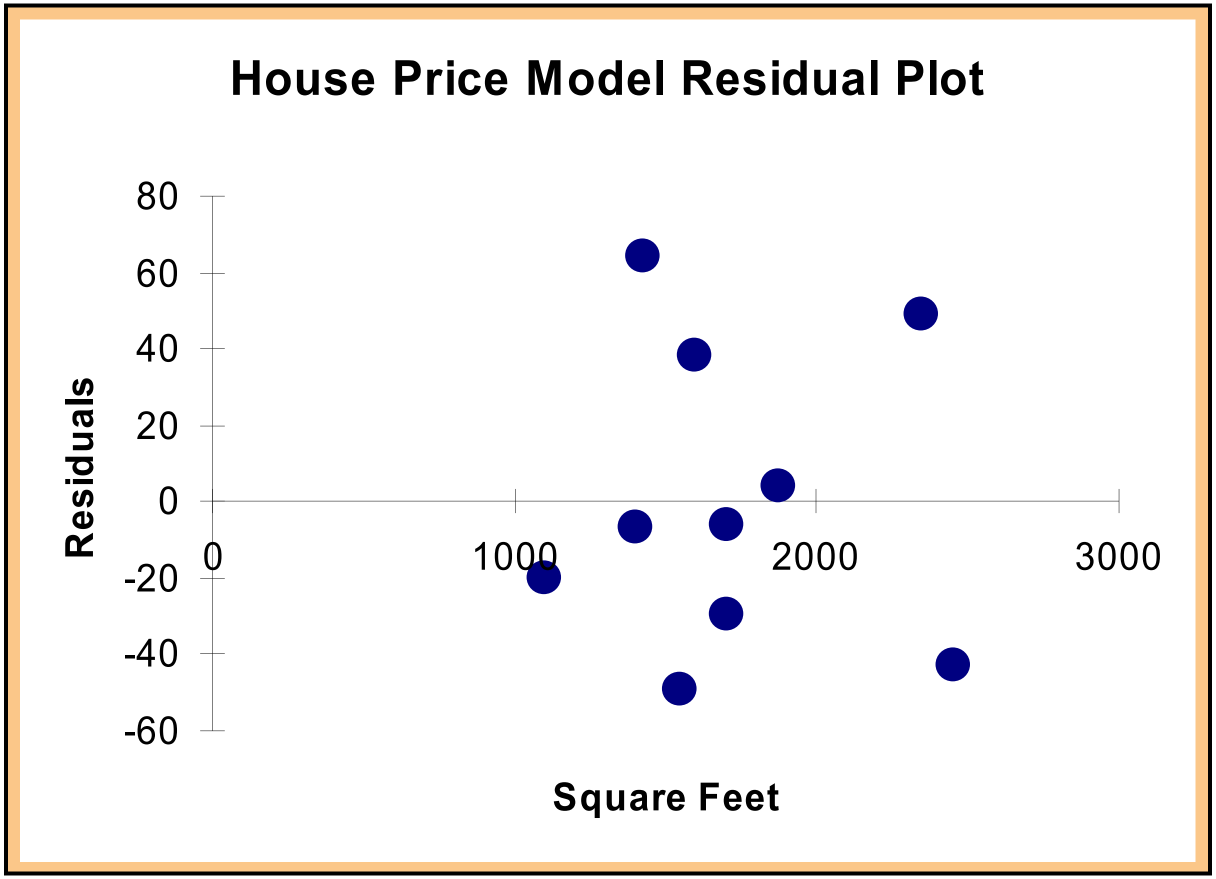
✓ Constant variance

Excel Output



RESIDUAL OUTPUT

	Predicted House Price	*Residuals*
1	251.92316	-6.923162
2	273.87671	38.12329
3	284.85348	-5.853484
4	304.06284	3.937162
5	218.99284	-19.99284
6	268.38832	-49.38832
7	356.20251	48.79749
8	367.17929	-43.17929
9	254.6674	64.33264
10	284.85348	-29.85348



Summary

- Introduced correlation analysis
- Discussed correlation to measure the strength of a linear association
- Introduced simple linear regression analysis
- Calculated the coefficients for the simple linear regression equation
- measures of variation (R^2 and s_ε)
- Addressed assumptions of regression and correlation

Summary

(continued)

- Described inference about the slope
- Addressed estimation of mean values and prediction of individual values
- Discussed residual analysis

Example

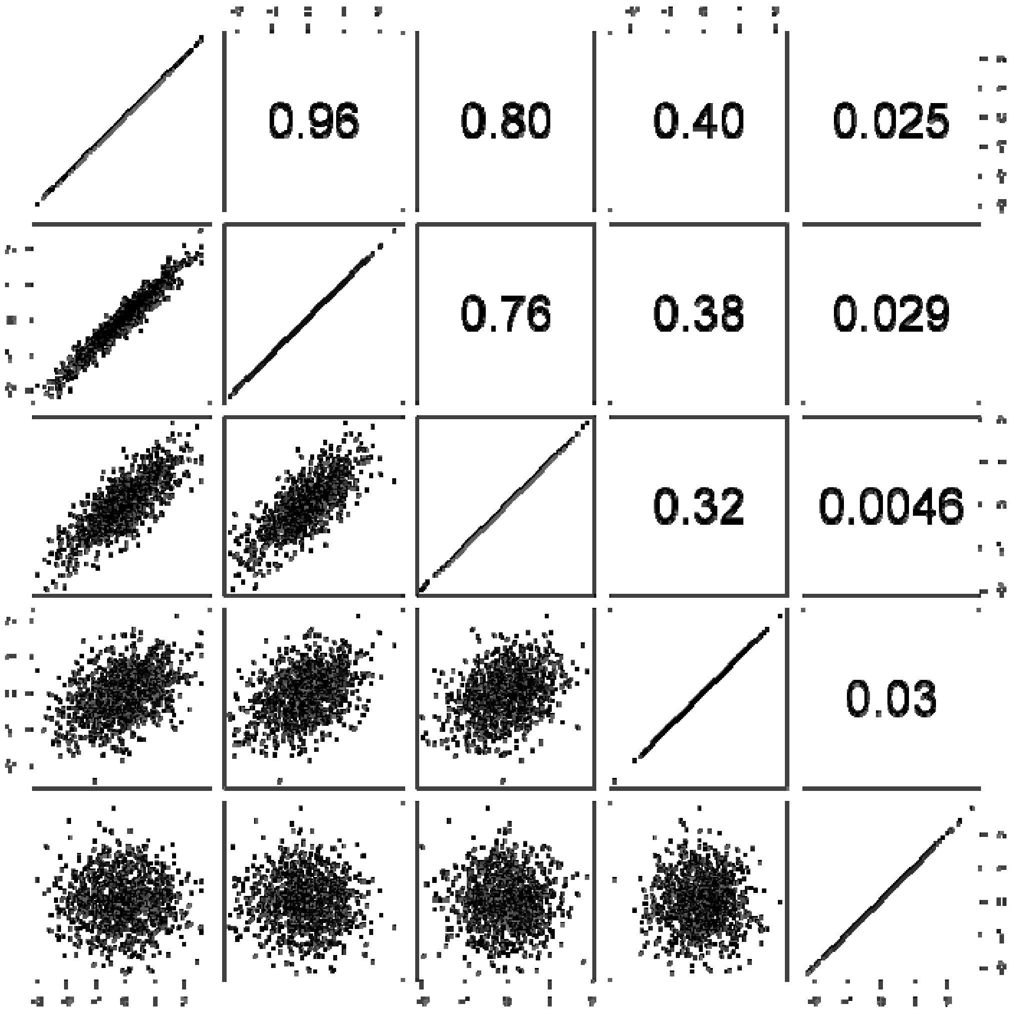
- Daytime SBP (systolic blood pressure) and age collected for 447 hypertensive males.

SBP	Age
115	**34**
130	**40**
128	**28**
123	**21**
126	**39**
…	…

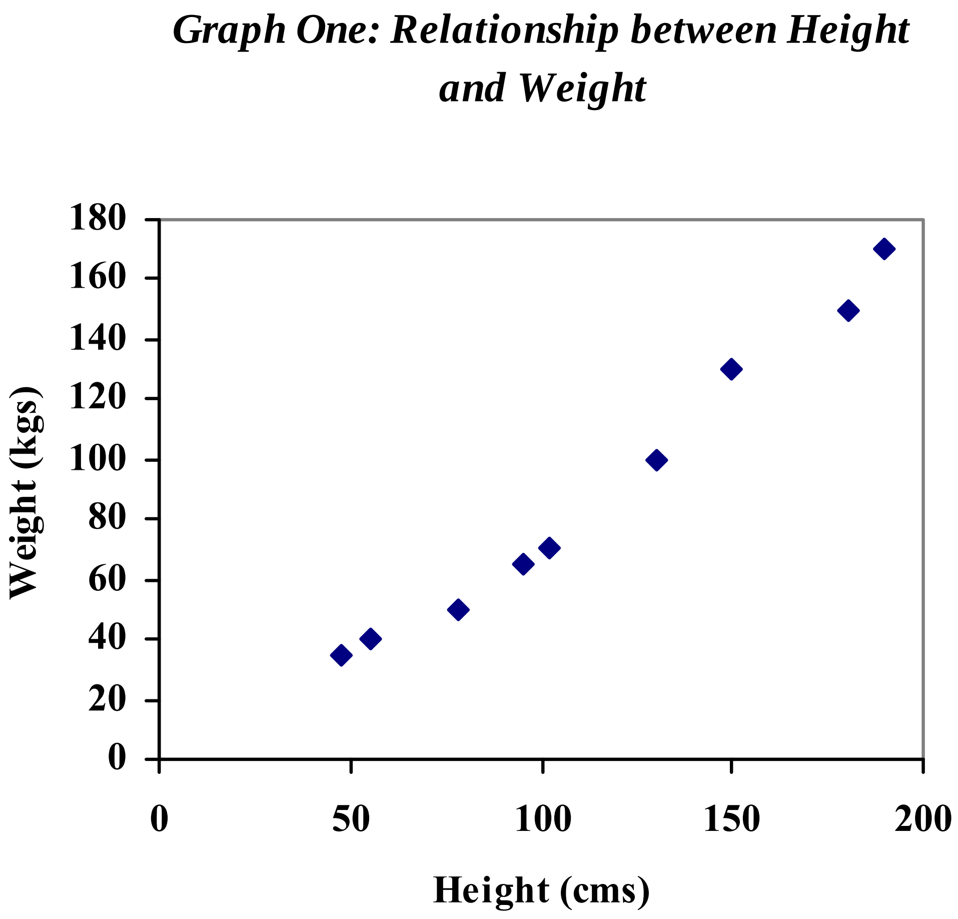
Example (contd)

- Is there a linear relationship between SBP and Age?
- r=0.145 ➔ weak positive relationship

Correlation examples



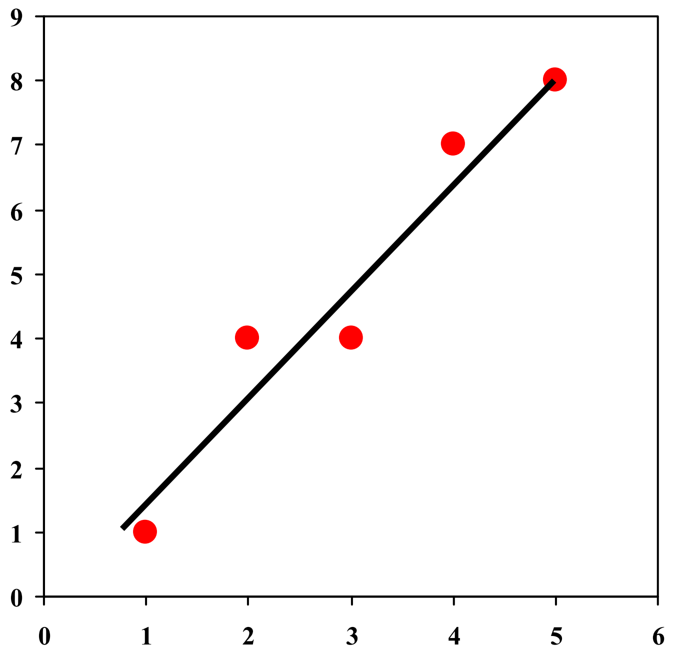
Example 2: Height vs. Weight



- Strong positive correlation between height and weight
- Can see how the relationship works, but cannot predict one from the other
- If 120cm tall, then how heavy?

Regression

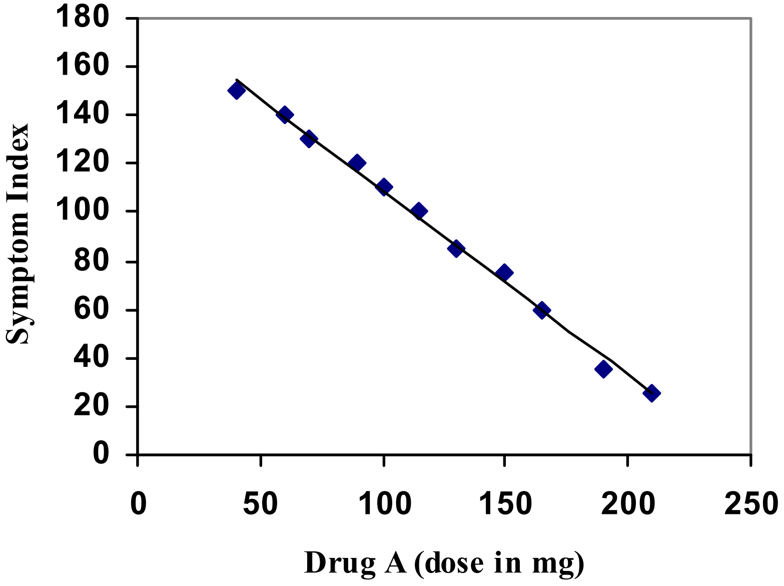
Problem: to draw a straight line through the points that best explains the variance



Line can then be used to predict Y from X

Example: Symptom Index *vs* Drug A

Graph Three: Relationship between Symptom Index and Drug A (with best-fit line)



- "Best fit line"
- allows us to describe relationship between variables more accurately.
- We can now predict specific values of one variable from knowledge of the other
- All points are close to the line

Simple Linear Regression

- Assume the population regression line:

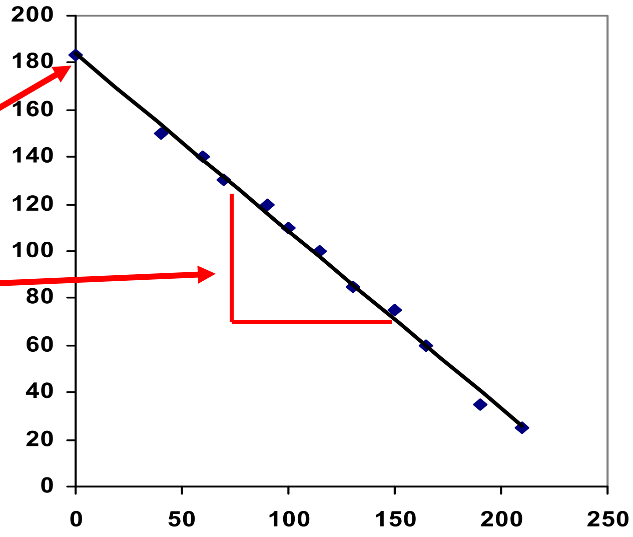
$$y = \alpha + \beta x$$

Where: α = y *intercept (constant)*

β = *slope of line*

y = *dependent variable*

x = *independent variable*



- $y_i = \alpha + \beta x_i + \varepsilon_i$

Regression

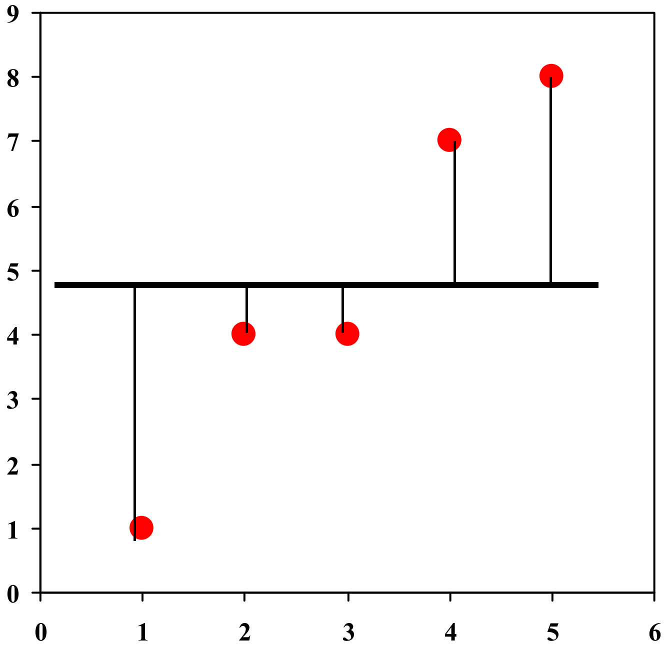
- Establish equation for the **best-fit line**:

$$y = a+bx$$

- Best-fit line same as **regression** line
- b is the **regression coefficient** for **x**
- x is the **predictor** or **regressor** variable for y

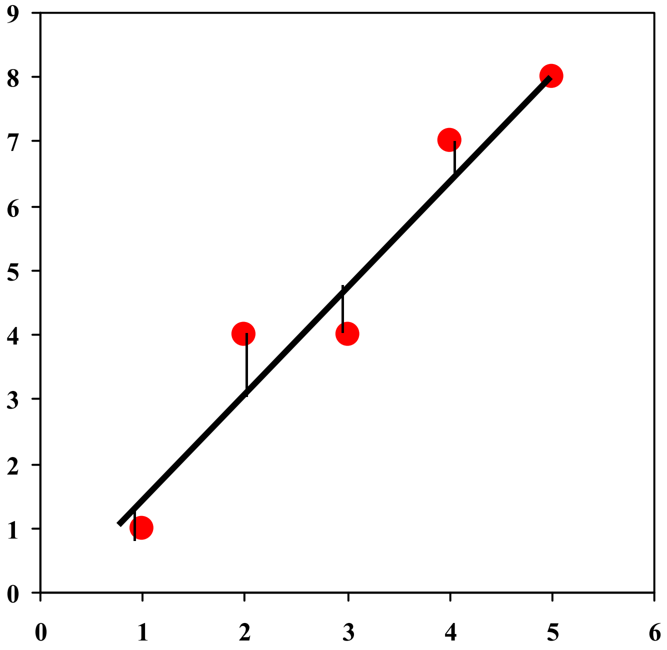
Fit a line to the data:

- Not great:



Fit a line to the data:

- Better:



Least Squares

- Minimise the (squared) distance between the points and the line
- a and b are the estimates of α and β which minimise

$$\sum \left\{ y_i - (\alpha + \beta x_i) \right\}^2$$

Least Squares Estimates

- Using calculus (partial derivatives), we get

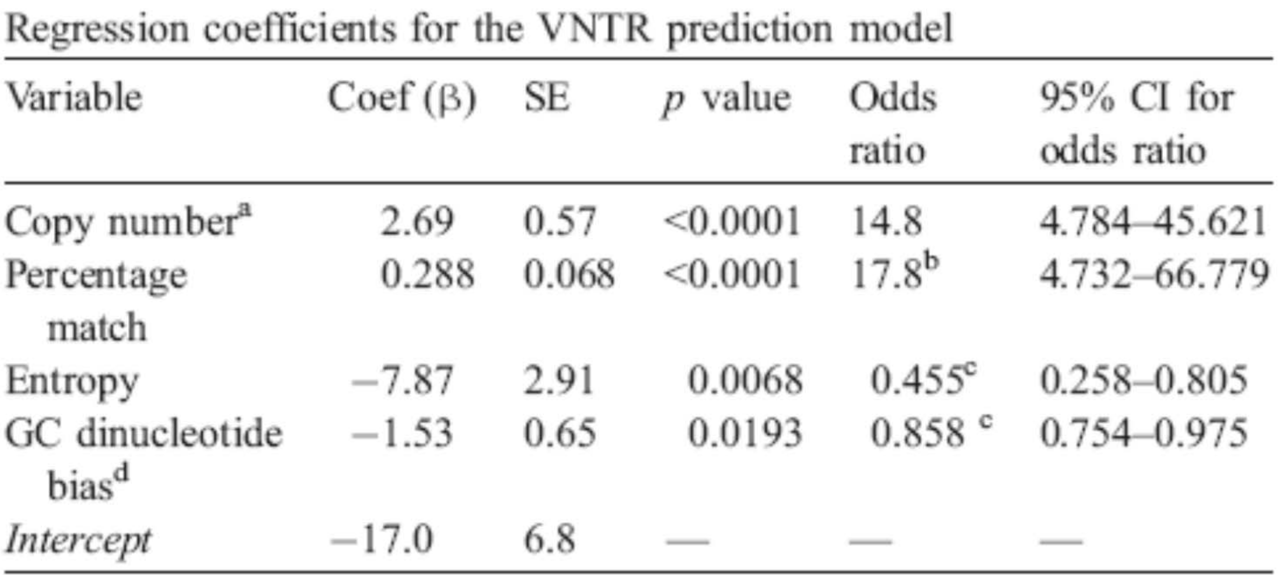
$$b = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

- Note b is related to the correlation coefficient r (same numerator)- if x and y are positively correlated then the slope is positive

Example from the literature

- Predicting tandem repeat variability

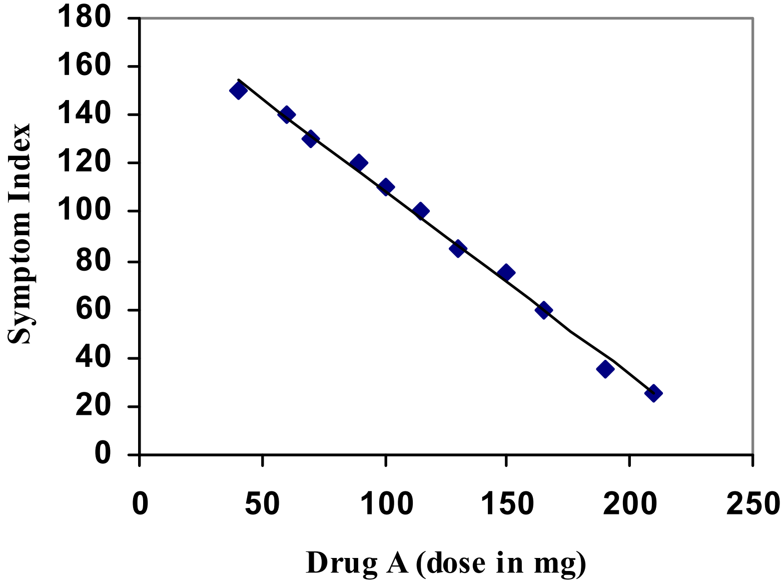


Regression coefficients for the VNTR prediction model

Variable	Coef (β)	SE	*p* value	Odds ratio	95% CI for odds ratio
Copy number[a]	2.69	0.57	<0.0001	14.8	4.784–45.621
Percentage match	0.288	0.068	<0.0001	17.8[b]	4.732–66.779
Entropy	−7.87	2.91	0.0068	0.455[c]	0.258–0.805
GC dinucleotide bias[d]	−1.53	0.65	0.0193	0.858 [c]	0.754–0.975
Intercept	−17.0	6.8	—	—	—

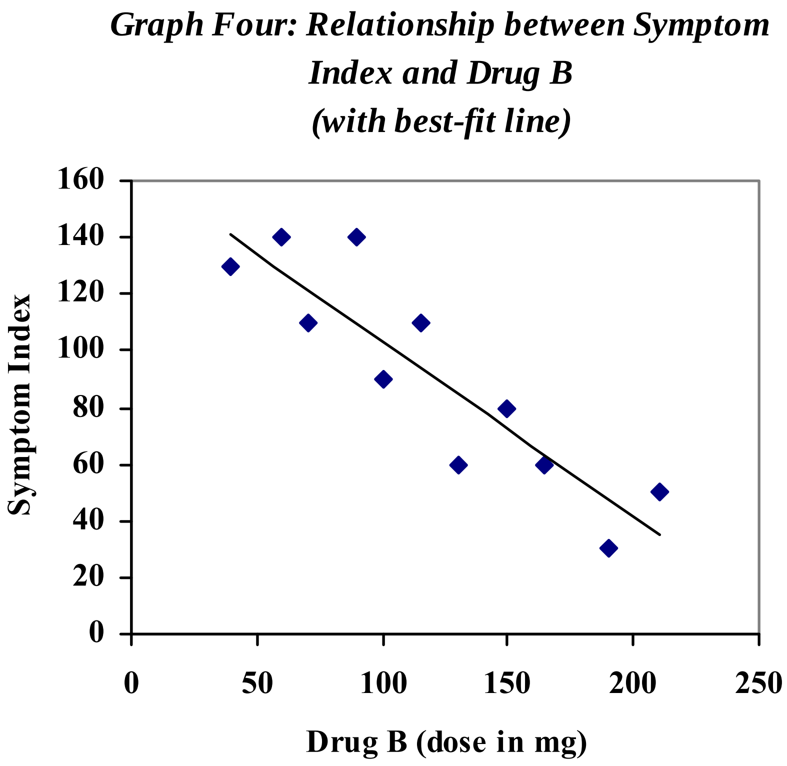
Example: symptom Index versus Drug A dose

Graph Three: Relationship between Symptom Index and Drug A (with best-fit line)



- "Best fit line"
- Allows us to describe relationship between variables more accurately.
- We can now predict specific values of one variable from knowledge of the other
- All points are close to the line

Example: Symptom Index versus Drug B dose



- We can still predict specific values of one variable from knowledge of the other
- Will predictions be as accurate?
- Why not?
- Large “residual” variation (random error)
  - = Difference between **observed** data and that **predicted** by the equation

Regression Hypothesis Tests

- Hypotheses about the intercept

  H_0: $\alpha = 0$ H_A: $\alpha \neq 0$

- But most research focuses on the slope

  H_0: $\beta = 0$ H_A: $\beta \neq 0$

  This addresses the general question “Is X predictive of Y?”

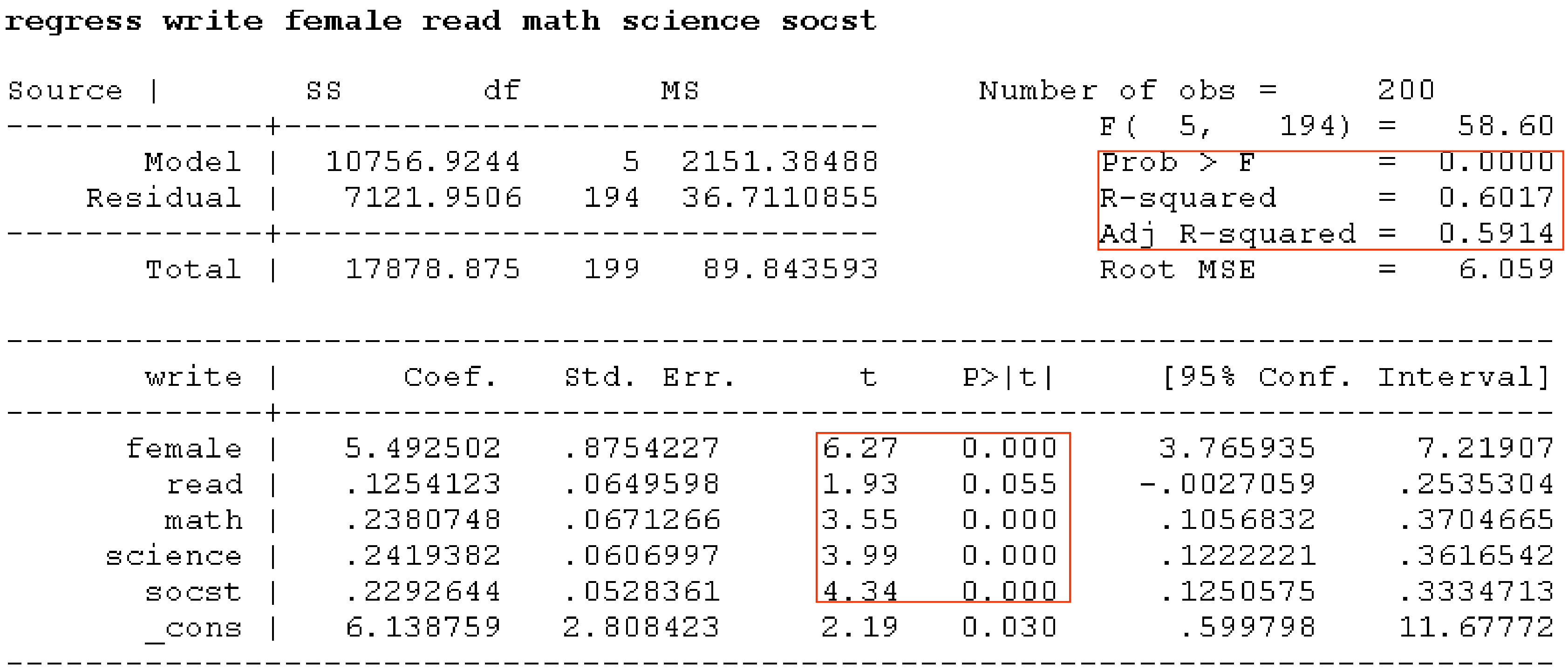
Regression

- Estimates of a slope (b) have a sampling distribution, like any other statistic
- If certain assumptions are met (NB normality, homogeneity of variance) the sampling distribution approximates the t-distribution
- Thus, we can assess the probability that a given value of b would be observed, if $\beta = 0$

→ hypothesis tests & confidence intervals

Regression

- R^2, the **coefficient of determination**, is the percentage of variation explained by the "regression".
- $R^2 > 0.6$ is deemed reasonably good.
- Note, the model must also be significant, e.g.



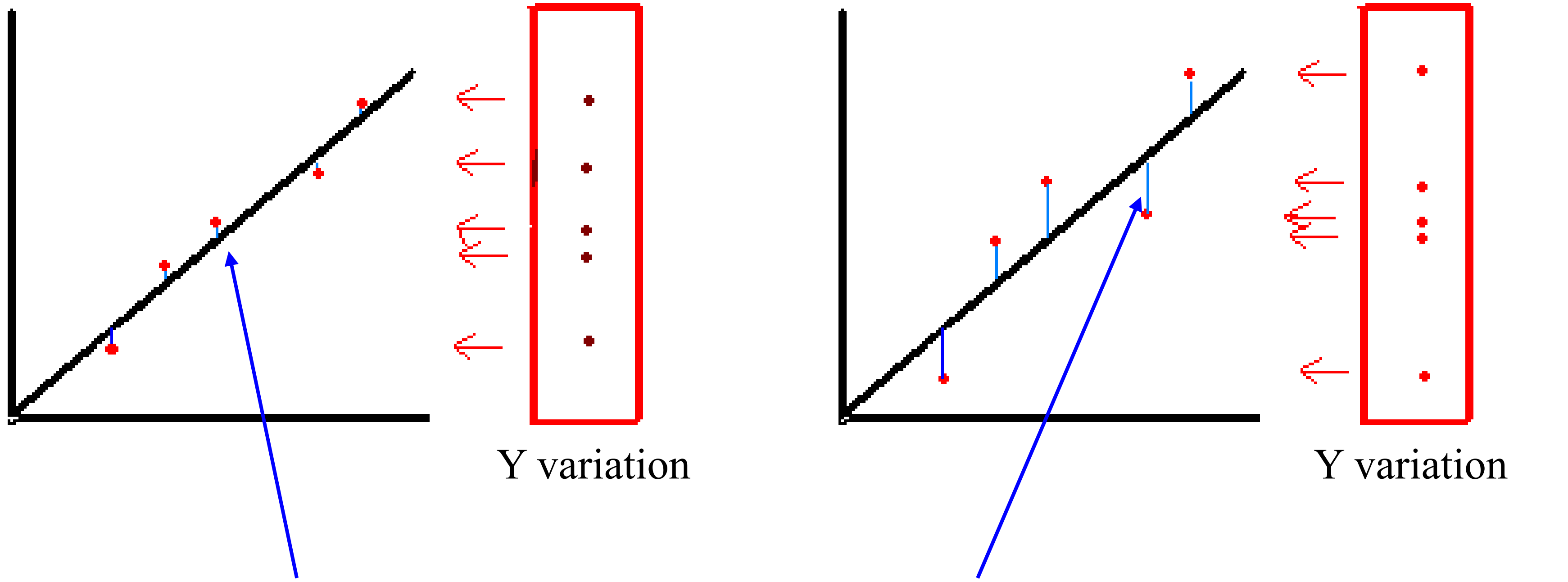
```
regress write female read math science socst

Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  5,   194) =   58.60
       Model |  10756.9244     5  2151.38488           Prob > F      =  0.0000
    Residual |   7121.9506   194  36.7110855           R-squared     =  0.6017
-------------+------------------------------           Adj R-squared =  0.5914
       Total |   17878.875   199   89.843593           Root MSE      =   6.059

------------------------------------------------------------------------------
       write |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   5.492502   .8754227     6.27   0.000     3.765935     7.21907
        read |   .1254123   .0649598     1.93   0.055    -.0027059    .2535304
        math |   .2380748   .0671266     3.55   0.000     .1056832    .3704665
     science |   .2419382   .0606997     3.99   0.000     .1222221    .3616542
       socst |   .2292644   .0528361     4.34   0.000     .1250575    .3334713
       _cons |   6.138759   2.808423     2.19   0.030      .599798    11.67772
------------------------------------------------------------------------------
```

Back to SBP and Age example

- a=123 and b=0.159 approximately
- What does b mean?
- Is age predictive of BP? i.e. is there evidence that b ≠ 0?
- How good is the fit of the regression line?

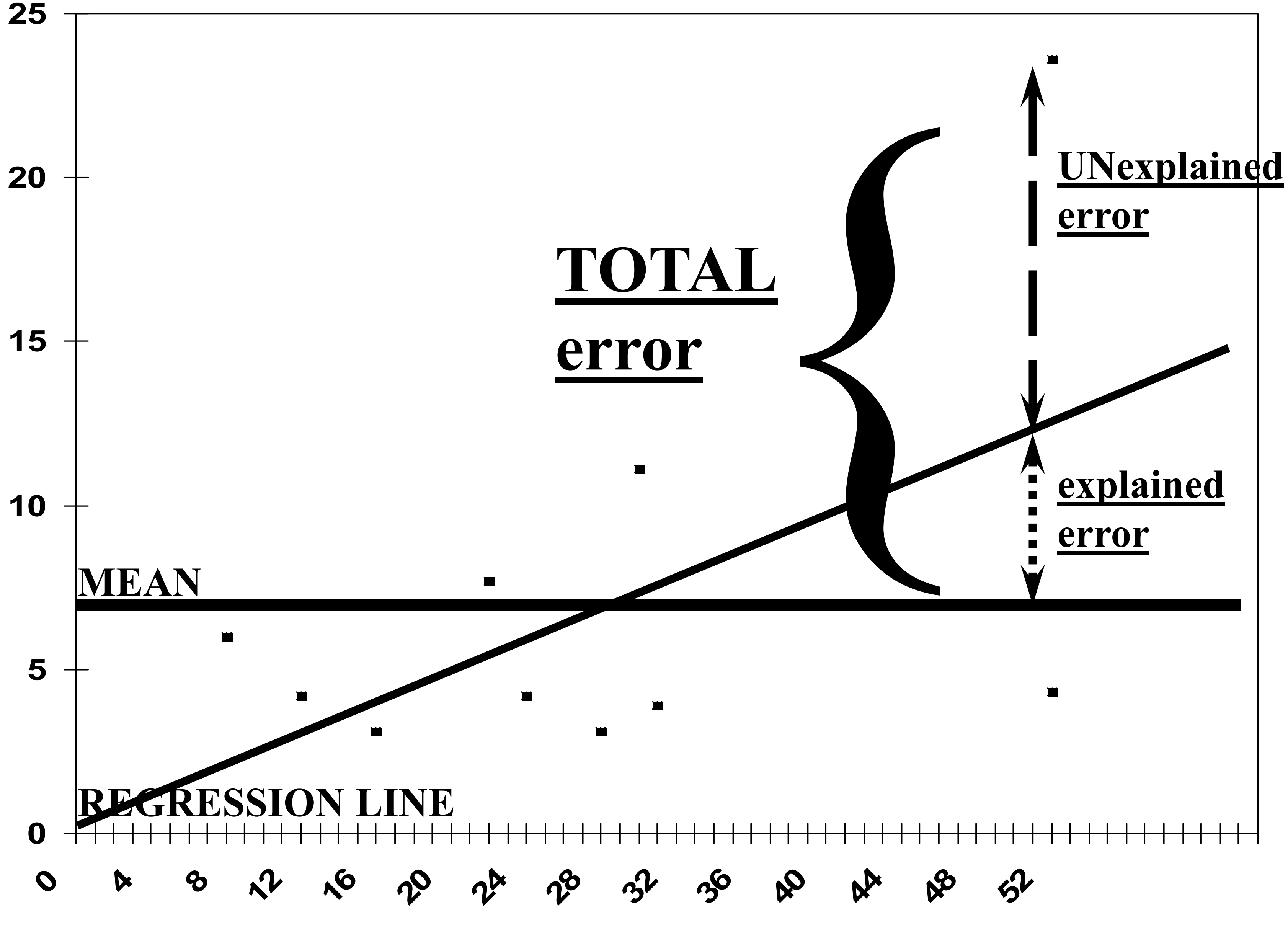
Regression R^2 Interpretation

- R^2= proportion of variation explained by (or predictive ability of) the regression



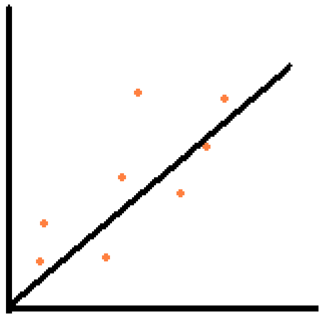
Variation in y is almost fully explained by x: $R^2 \approx 1$

Still some variation in y left over (not explained by x): $R^2 < 1$



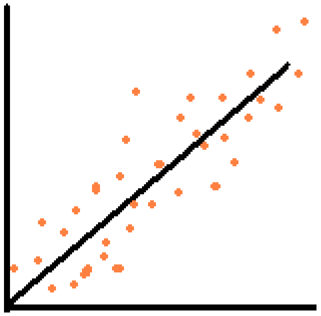
25
20
15
10
5
0
0
4
8
12
16
20
24
28
32
36
40
44
48
52
TOTAL
error
UNexplained
error
explained
error
MEAN
REGRESSION LINE

Regression – four possibilities



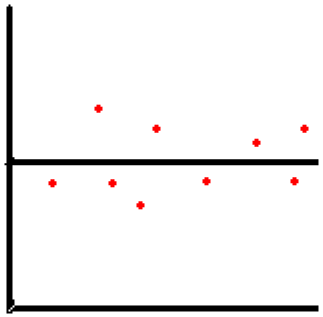
$b \neq 0$
P-value non-significant

Relationship but not much evidence



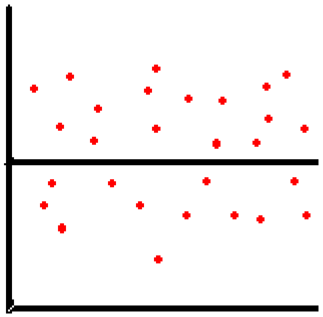
$b \neq 0$
P-value significant

Plenty of evidence for a relationship



$b \approx 0$
P-value non-significant

No relationship & not much evidence

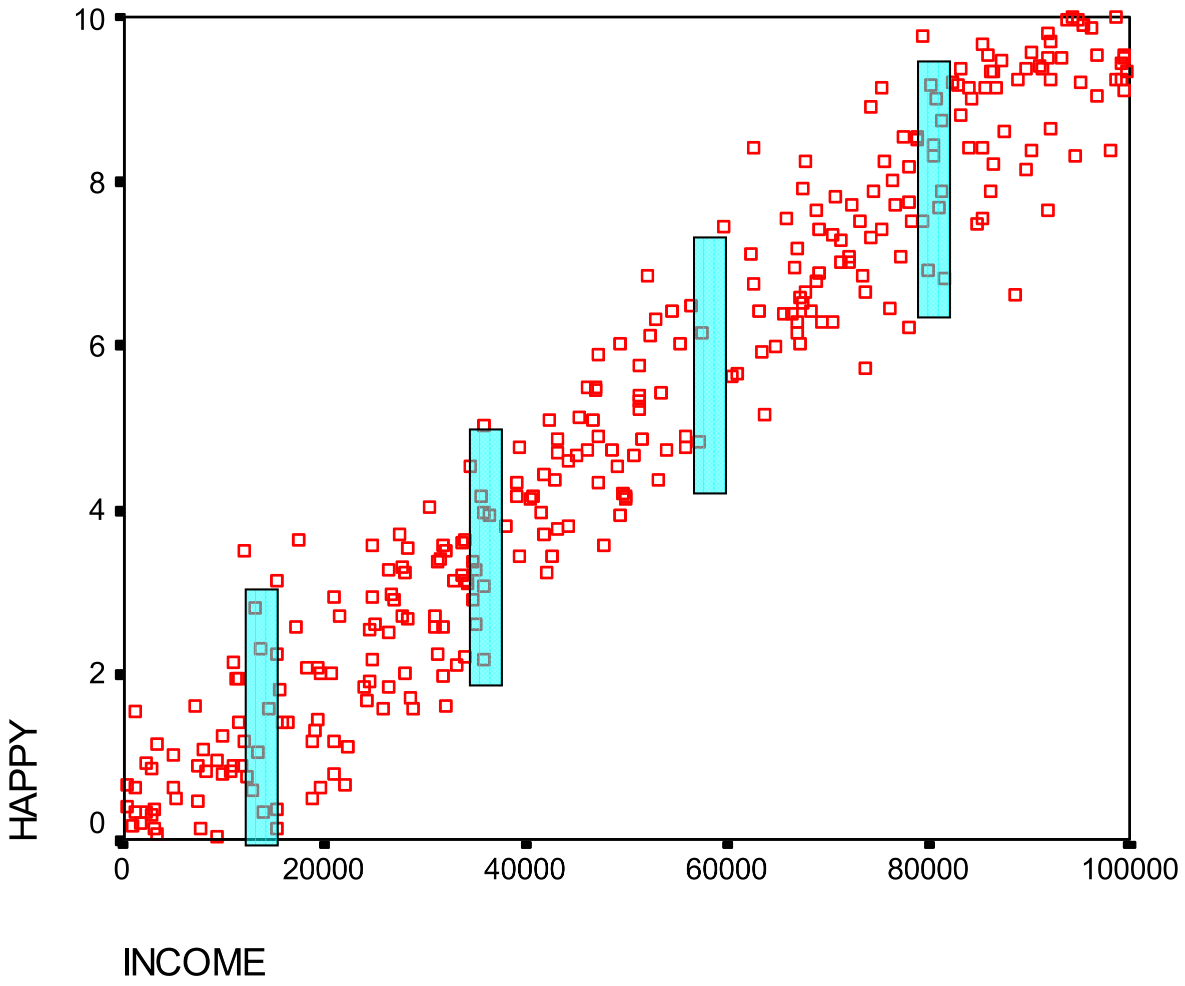


$b \approx 0$
P-value significant

Plenty of evidence for no relationship

Regression Assumption:

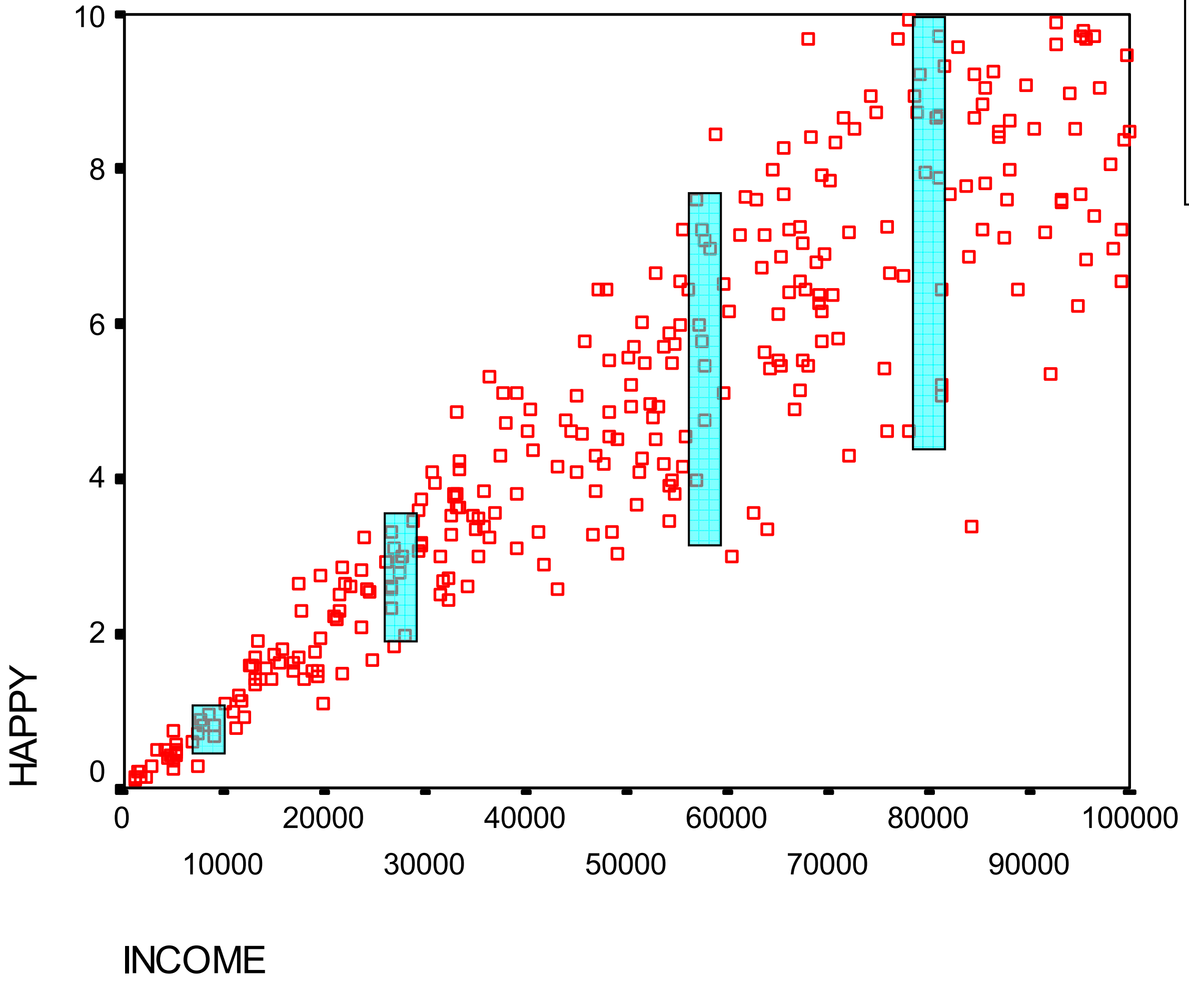
Homoscedasticity (Equal Error Variance)



Examine error at different values of X. Is it roughly equal?

Here, things look pretty good.

Heteroscedasticity: **Un**equal Error Variance



At higher values of X, error variance increases a lot.

A transformation of data (e.g. log) can remove heterskedasticity

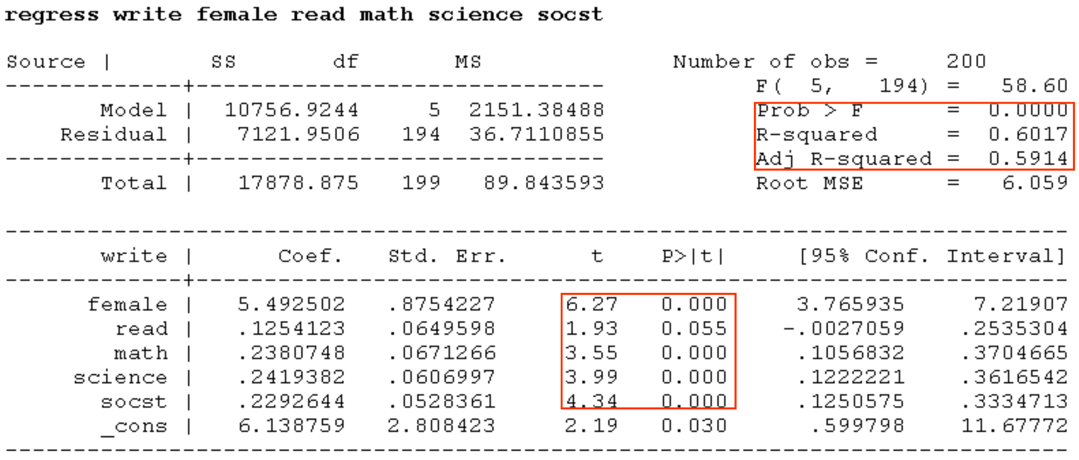
Multiple Regression

- Extension of simple linear regression to more than one (continuous/ordinal) independent variables
- We use least squares in exactly the same way to obtain estimates of the regression coefficients
- e.g. with 2 independent variables x and z, we fit the regression

  $$y=a+bx+cz…$$

  where a,b and c are the regression coefficients. This represents a plane in 3d space

Previous example



```
regress write female read math science socst

Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  5,   194) =   58.60
       Model |  10756.9244     5  2151.38488           Prob > F      =  0.0000
    Residual |   7121.9506   194  36.7110855           R-squared     =  0.6017
-------------+------------------------------           Adj R-squared =  0.5914
       Total |   17878.875   199   89.843593           Root MSE      =   6.059

------------------------------------------------------------------------------
       write |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   5.492502   .8754227     6.27   0.000     3.765935     7.21907
        read |   .1254123   .0649598     1.93   0.055    -.0027059    .2535304
        math |   .2380748   .0671266     3.55   0.000     .1056832    .3704665
     science |   .2419382   .0606997     3.99   0.000     .1222221    .3616542
       socst |   .2292644   .0528361     4.34   0.000     .1250575    .3334713
       _cons |   6.138759   2.808423     2.19   0.030      .599798    11.67772
------------------------------------------------------------------------------
```

Notes on multiple regression

- Make sure variables are normal. If not, transform them. If still not, can split into 2 groups (categories (0/1)) for e.g. high vs. low responders
- Can combine with “stepwise selection”: instead of using every variable and forcing them into a final model, can drop out variables automatically, e.g. petri dish temperature, that are not predictive

Example

- Study to evaluate the effect of the duration of anesthesia and degree of trauma on percentage depression of lymphocyte transformation
- 35 patients
- Trauma factor classified as 0, 1, 3 and 4, depending upon severity

Duration	Trauma	Depression	Duration	Trauma	Depression
4	3	36.7	3	3	29.9
6	3	51.3	4	3	76.1
1.5	2	40.8	3	3	11.5
4	2	58.3	3	3	19.8
2.5	2	42.2	7	4	64.9
3	2	34.6	6	4	47.8
3	2	77.8	2	2	35
2.5	2	17.2	4	2	1.7
3	3	-38.4	2	2	51.5
3	3	1	1	1	20.2
2	3	53.7	1	1	-9.3
8	3	14.3	2	1	13.9
5	4	65	1	1	-19
2	2	5.6	3	1	-2.3
2.5	2	4.5	4	3	41.6
2	2	1.6	8	4	18.4
1.5	2	6.2	2	2	9.9
1	1	12.2			

Example (con't)

- Fitted regression line is

$$y=-2.55+10.375x+1.105z$$

or

Depression= -2.55+10.375*Trauma+1.105*Duration

- Both slopes are non-significant (p-value=0.1739 for trauma, 0.7622 for duration)
- R^2=16.6% of the variation in lymphocyte depression is explained by the regression
- Conclusion: Trauma score and duration of anesthesia are poor explanations for lymphocyte depression

Collinearity

- If two (independent) variables are closely related its difficult to estimate their regression coefficients because they tend to get confused
- This difficulty is called *collinearity*
- Solution is to exclude one of the highly correlated variables

Example

- Correlation between trauma and duration= 0.762 (quite strong)
- Drop trauma from regression analysis

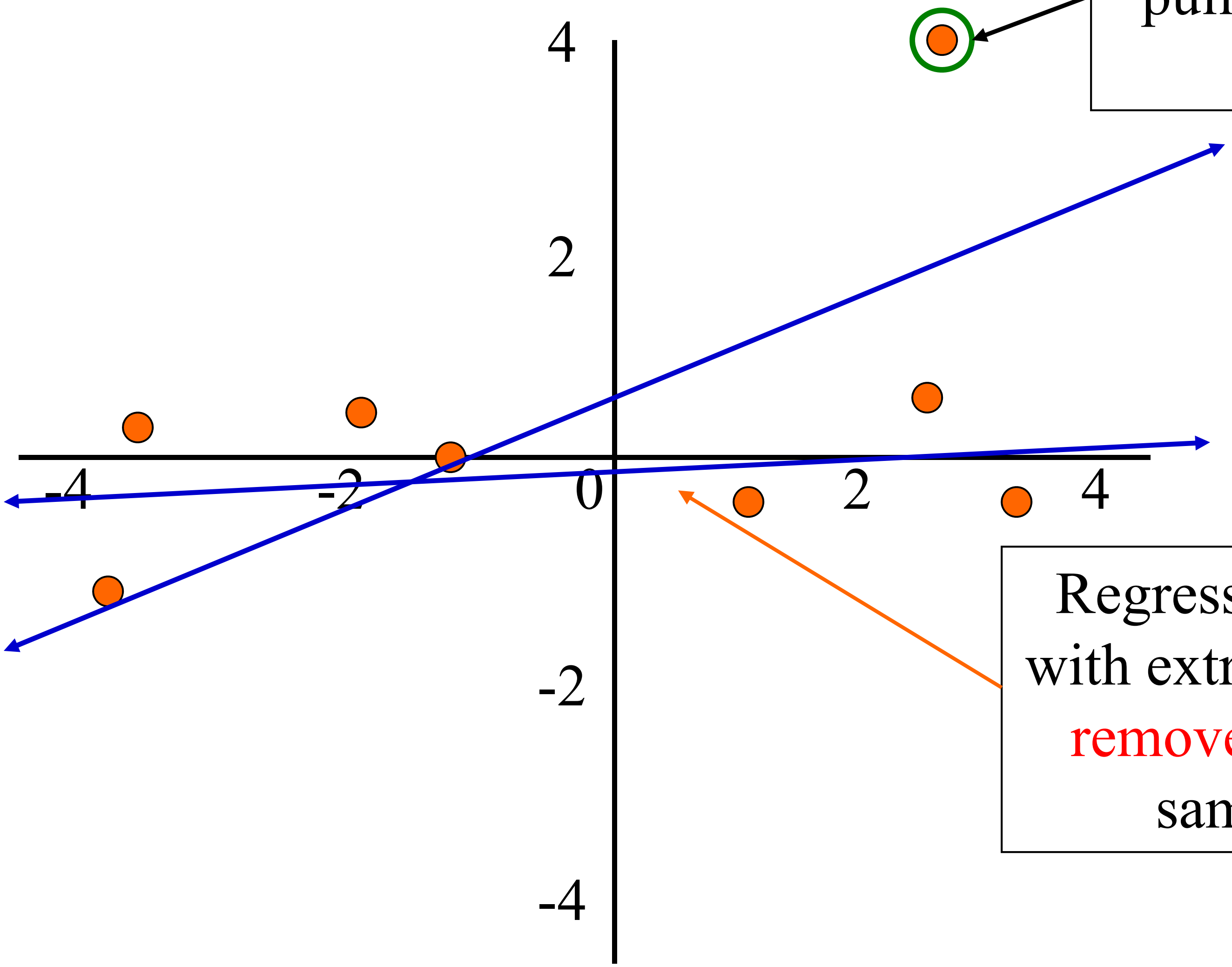
Depression=9.73+4.94*Duration

- P-value for duration is 0.0457, statistically significant!
- However, the R^2 is still small (11.6%)
- Conclusion: Although there is evidence for a non-zero slope or linear relationship with duration, there is still considerable variation not explained by the regression.

Outliers in Regression

- Outliers: cases with extreme values that differ greatly from the rest of your sample
- Even a few outliers can dramatically change estimates of the slope (b)
- Outliers can result from:
  - Errors in coding or data entry (→rectify)
  - Highly unusual cases (→exclude?)
  - Or, sometimes they reflect important "real" variation (→include?)

Outliers: Example

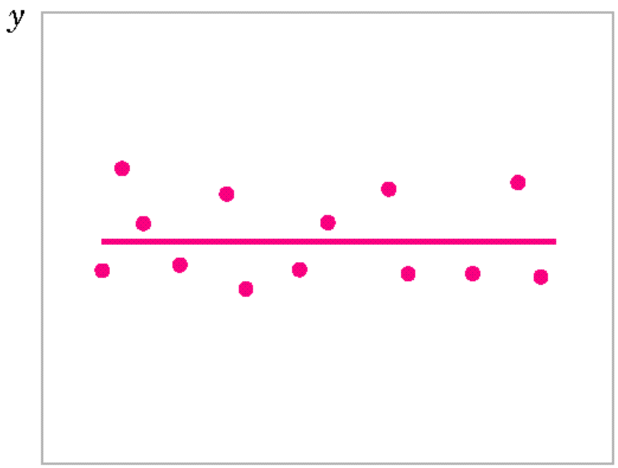


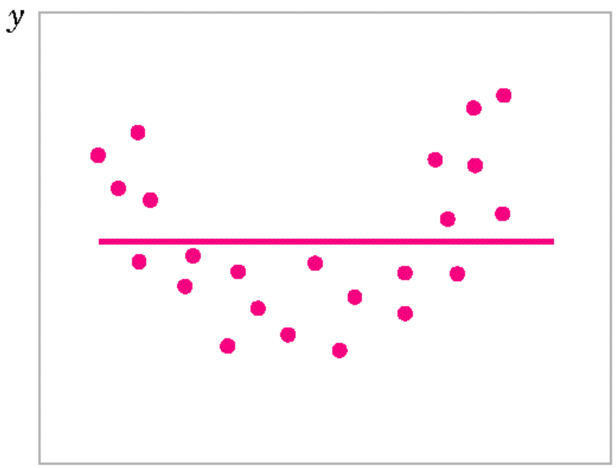
Extreme case that pulls regression line up

Regression line with extreme case removed from sample

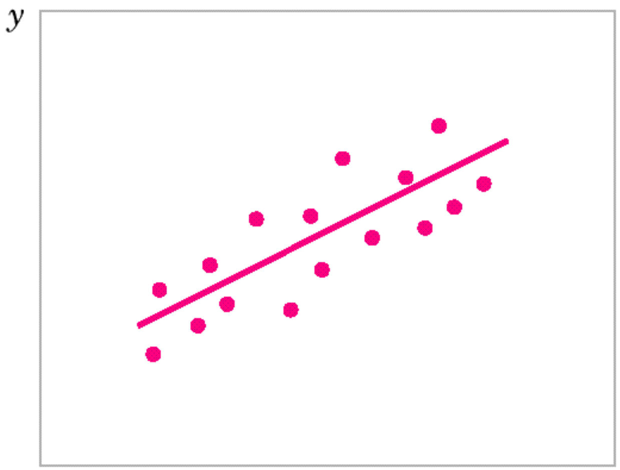
What about non-linear relationships?

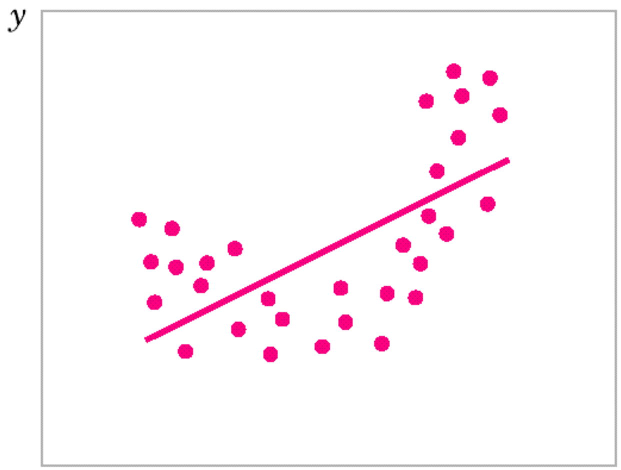
Fail to reject β=0





Reject β=0



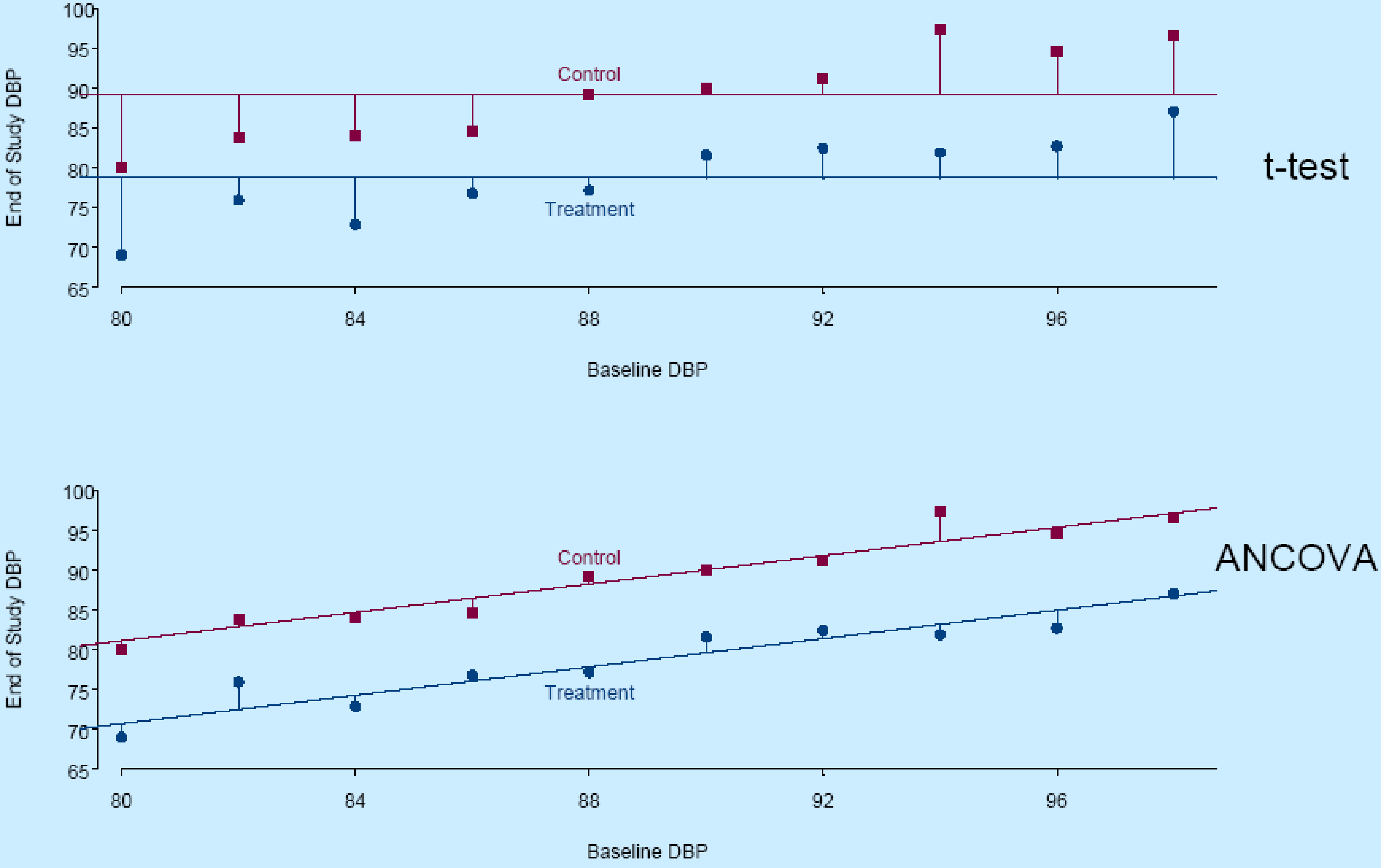


Non-linear models

- Linear Regression fits a straight line to your data
- Non-linear models may, in some cases, be more appropriate
- Non-linear models are usually used in situations where non-linear relationships have some *biological explanation*
- e.g. non-linear models are commonly used in pharmacokinetics studies and compartmental analysis
- Computationally intensive - estimates may not "converge" and results may be difficult to interpret

Analysis of Covariance-ANCOVA

- Modelling both categorical and continuous independent variables (*covariates*)
- Justification: Consideration of a covariate may improve the precision (reduce variability) of the comparisons of categories made in ANOVA
- Often used as a correction for baseline imbalances
- ANOVA + regression combined



Within-arm variability smaller with ANCOVA than with t-test

Example

- Antihypertensive Drug Clinical Trial: 84 patients randomly allocated to either Aliskiren or Losartan. Blood pressure measured at baseline and after several weeks treatment. Age and gender was also recorded for each patient. Is there a significant difference in the two treatments for BP reduction?

Age	Treatment	Ba_daySBP	Red_daySBP	gender
66	SPP75	131.1	0.6	Female
62	SPP75	160.4	-0.7	Male
48	Los100	147.3	17.9	Male
32	Los100	144.8	-2.4	Male
61	Los100	150.7	21.6	Female
68	SPP75	152.4	6	Male
60	Los100	143.6	15.3	Male
33	SPP75	143.2	5.1	Male
69	Los100	166.6	24.6	Male
53	SPP75	147.6		Female
63	SPP75	163.7	-0.2	Male
64	Los100	145.7	15.5	Male
58	SPP75	168.3	0.5	Male
52	SPP75	156.8	0.6	Male
54	SPP75	154.9	7.3	Male
43	SPP75	170.5	-2.7	Female
46	SPP75	155.5	18.3	Female
46	Los100	173.1	26.7	Male
66	Los100	151.2	-1.7	Male
29	SPP75	139.8	-1.5	Male
50	SPP75	162.6	13	Male
49	SPP75	178.8	17.2	Female
40	SPP75	146.8	0	Male
52	Los100	157.1	-0.2	Female
68	Los100	152	8.9	Male
35	SPP75	145.4	2.8	Male
49	Los100	153.7	14.9	Male
47	SPP75	139.2	10	Male
45	Los100	156	17	Male
68	Los100	149.9	6.8	Female
48	Los100	147	16	Female
69	SPP75	145.3	-1.2	Male
64	Los100	142.5	20.6	Female
40	SPP75	168.6	2.6	Male
61	SPP75	165.6	8.1	Male
47	Los100	156.3	2.3	Female
60	Los100	147.7	11.5	Female
35	SPP75	157.4	12.5	Male
61	SPP75	143.7	-1.8	Male
62	Los100	148.6	6.6	Male
54	Los100	164.6	39.6	Female
45	Los100	145.3	-6.1	Female
57	SPP75	143.9	7	Female
48	SPP75	144.3	2.4	Male
59	Los100	147.8	1.1	Female
47	SPP75	150.4	-2.3	Male
54	Los100	143.9	-0.2	Male
45	SPP75	145.1	6.6	Male
61	SPP75	158	3	Male
69	Los100	154.8		Male
21	SPP75	142.1		Male
69	SPP75	171.3	-2	Male
66	Los100	140.3	2.8	Male
42	SPP75	146	5.7	Male
47	SPP75	159.3	14.6	Female
60	SPP75	157.8	6.8	Male

Age	Treatment	Ba_daySBP	Red_daySBP	gender
45	SPP75	162.9	10.7	Male
62	SPP75	173.4	55.9	Female
57	SPP75	141.1	0.2	Female
54	Los100	147.6	11.1	Male
54	Los100	140.5	-16.5	Male
63	Los100	156	19.3	Female
35	SPP75	150.9	26.8	Male
52	SPP75	143.8	-9.7	Female
66	Los100	150.9	0.6	Male
55	SPP75	155.8	-3	Female
61	Los100	162.1	21	Male
35	Los100	149.1	21.3	Male
52	Los100	177	-2.7	Male
60	SPP75	157.8	6.8	Male
37	SPP75	143.4	-3.9	Male
54	Los100	163.9	26.6	Male
62	Los100	147.4		Female
55	Los100	141.9	8.7	Female
57	Los100	163.1	3.6	Male
40	SPP75	153.3	0.9	Male
56	SPP75	165.9	11.9	Male
53	Los100	144.1	-1.9	Female
52	Los100	144.1	18.6	Male
46	SPP75	147.7	-5.6	Male
58	Los100	153	-10.3	Male
41	SPP75	149.9	4.3	Male
42	Los100	165.8	35.5	Male
57	SPP75	154.8	16.1	Female
56	Los100	146.6	18.1	Female

Analysis without covariates

- Since treatment has only 2 levels (Losartan & Aliskiren), the ANOVA is equivalent to the two-sample t-test
- Treatment difference (Losartan-Aliskiren)=5.06 with P-value=0.0554
- Borderline non-significant at the 5% level of significance

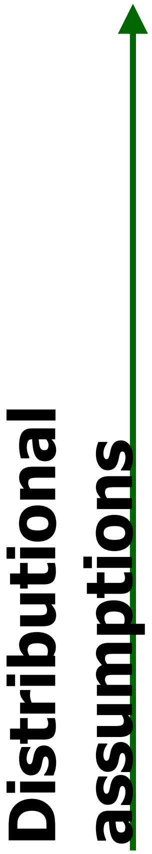
ANCOVA analysis

- We have as factors (categorical independent variables)
  - Treatment
  - Gender
- As covariates (continuous independent variables)
  - Age
  - Baseline BP

Results: ANOVA table

Source	df	SS	MS	F	P-value
Baseline BP	1	1572.90	1572.90	14.36	0.0005
Treatment	1	651.73	651.73	5.58	0.0208
Age	1	71.57	71.57	0.61	0.4363
Gender	1	279.02	279.02	2.39	0.1264

Classification of statistical methods based on distributional assumptions



Distributional assumptions

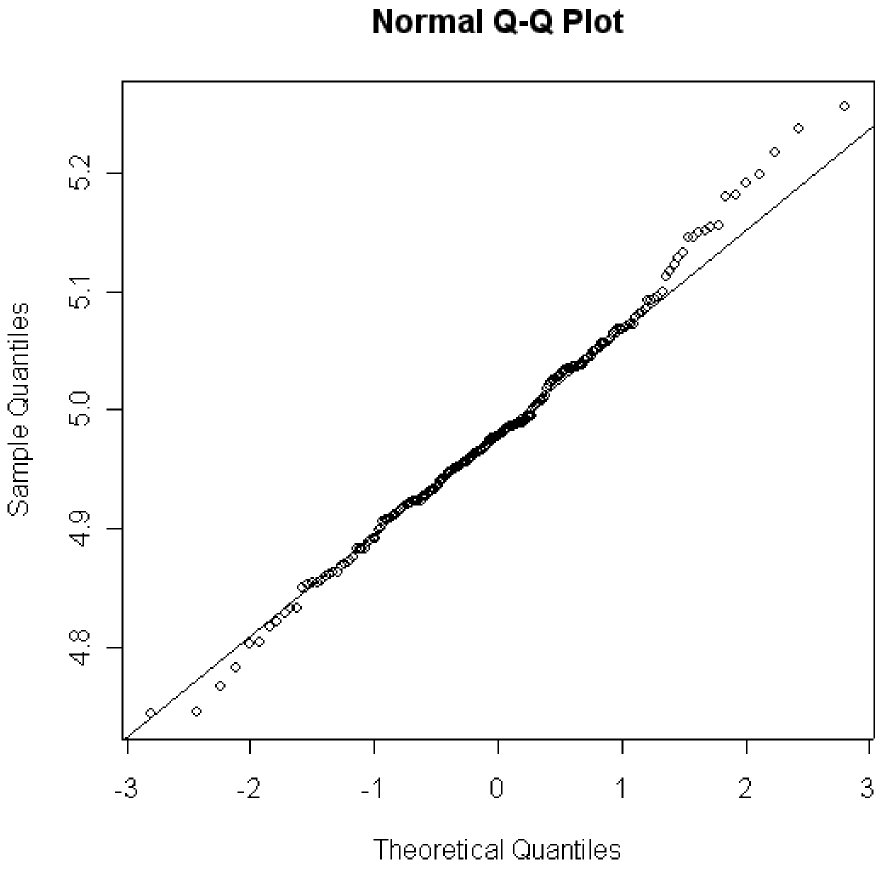
- Likelihood Based Methods
  - Assume a distribution of data explicitly at the outset
  - Can fit very complex models e.g. model correlation structures with the data, allow for unequal variances, etc.

- T-tests, ANOVA, regression, etc. rely on normality of data or that the data be of a sufficiently large size (Central Limit Theorem)

- Non-parametric methods
  - Relaxes assumptions about the shape of the distribution
  - Methods are based on ranking of the data points

So how normal is your data?

- Difficult to see visually if a histogram looks normal
- Use normal probability (quantile-quantile) plots
- Points must lie along a line
- Also useful for detecting outliers



Significant Deviations from Normality: Alternatives

- Transform your data to make it more “normal” and use standard parametric tests
  - e.g. log transform (eliminates skew)
  - Difficulty - may wish to reverse-transform results back (e.g. exponentiate parameter estimates)
- Use non-parametric methods
  - Make less assumptions on distribution shape
  - These may have less power than parametric methods

Non-parametric Equivalents

Parametric	Non-parametric
Paired T-test	Wilcoxon Signed Rank Test
Two-sample T-test	Wilcoxon Rank-Sum
ANOVA	Kruskal Wallis Analysis
Pearson’s Correlation	Spearman’s Rank Correlation
ANCOVA	ANCOVA on ranked data

Other Multivariate Methods

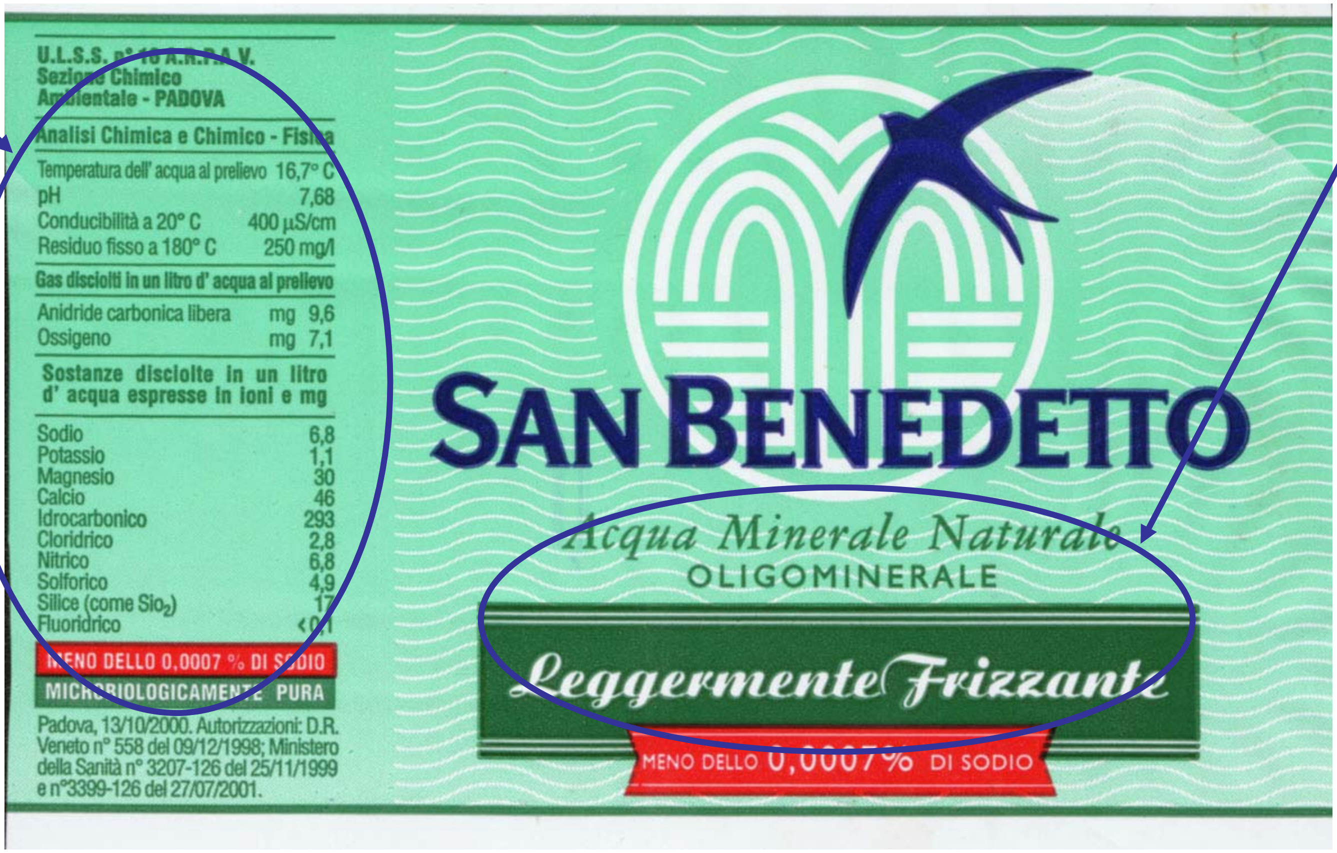
- Multivariate Analysis of Variance (MANOVA)
  - ANOVA with more than one dependent variable or response
  - Also MANCOVA
  - Low power is a problem
  - Not so widely used

- The objectives behind multivariate analyses can be quite different (to those presented), namely
  - Discriminant Analysis
  - Classification
  - Clustering
  - Pattern Recognition (principal components analysis)

Multivariate analysis: mineral water

Pattern

Pattern recogni



U.L.S.S. n° 16 A.R.P.A.V.
Sezione Chimico
Ambientale - PADOVA

Analisi Chimica e Chimico - Fisica

Temperatura dell' acqua al prelievo	16,7° C
pH	7,68
Conducibilità a 20° C	400 µS/cm
Residuo fisso a 180° C	250 mg/l

Gas disciolti in un litro d' acqua al prelievo

Anidride carbonica libera	mg 9,6
Ossigeno	mg 7,1

Sostanze disciolte in un litro d' acqua espresse in ioni e mg

Sodio	6,8
Potassio	1,1
Magnesio	30
Calcio	46
Idrocarbonico	293
Cloridrico	2,8
Nitrico	6,8
Solforico	4,9
Silice (come Sio_2)	17
Fluoridrico	<0,1

MENO DELLO 0,0007 % DI SODIO

MICROBIOLOGICAMENTE PURA

Padova, 13/10/2000. Autorizzazioni: D.R. Veneto n° 558 del 09/12/1998; Ministero della Sanità n° 3207-126 del 25/11/1999 e n°3399-126 del 27/07/2001.

Correlation coefficients in data

- The square root of coefficient R is C.
- Example: we measure 5 variables of a population of peaches :
  - Total acidity, anthocyanin, brix, carotene e chlorophyll
  - We want to know the correlation between them
- The correlation can be shown as a matrix 5*5

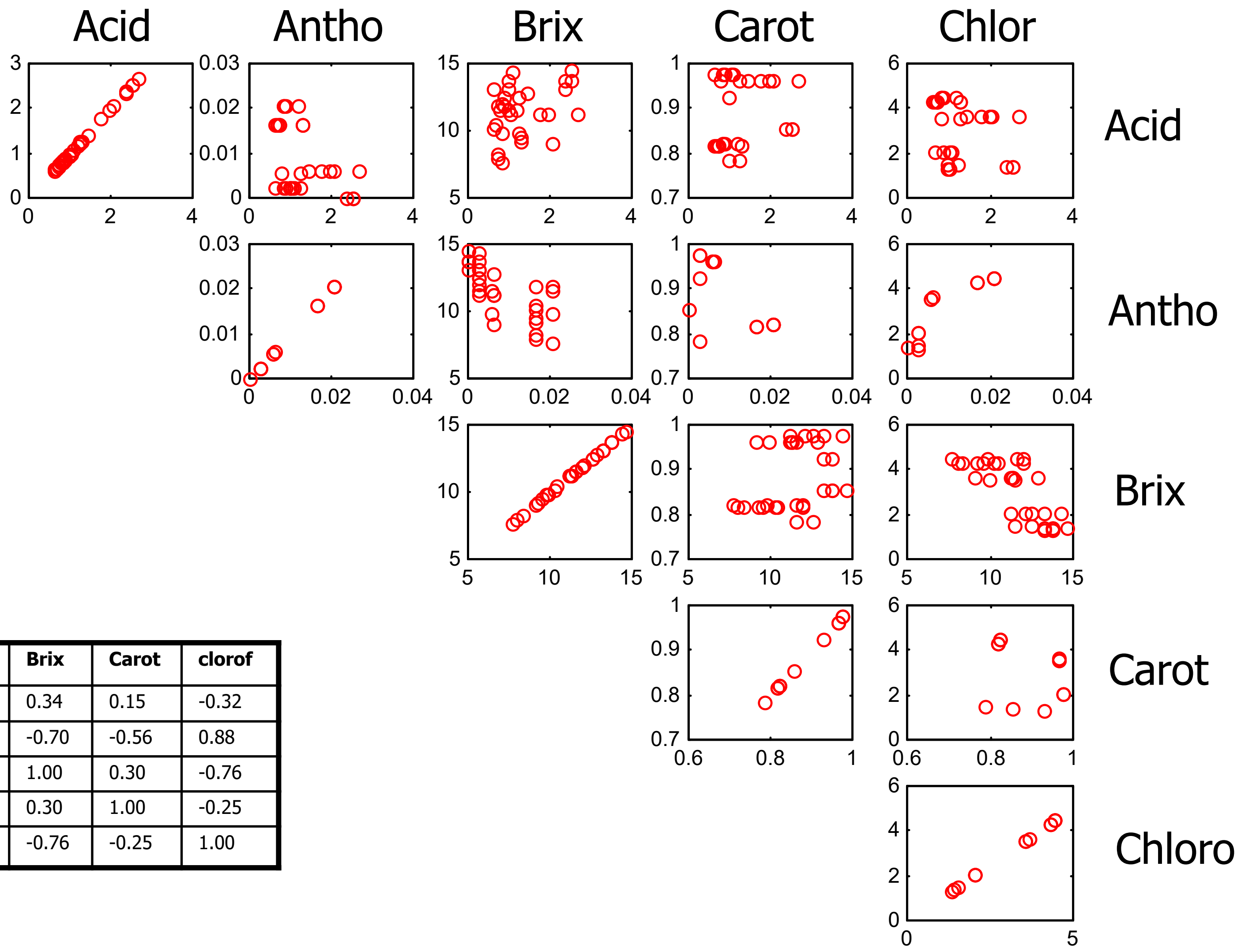
	Acid	**Anth**	**Brix**	**Carot**	**chlo**
acid	1.00	-0.48	0.34	0.15	-0.32
anth	-0.48	1.00	-0.70	-0.56	0.88
brix	0.34	-0.70	1.00	0.30	-0.76
carot	0.15	-0.56	0.30	1.00	-0.25
chlor	-0.32	0.88	-0.76	-0.25	1.00

correlation matrix

	Acid	**Antho**	**Brix**	**Carot**	**chlo**
acid	1.00	-0.48	0.34	0.15	-0.32
antoc	-0.48	1.00	-0.70	-0.56	0.88
brix	0.34	-0.70	1.00	0.30	-0.76
carot	0.15	-0.56	0.30	1.00	-0.25
clorof	-0.32	0.88	-0.76	-0.25	1.00

- The matrix is symmetric
  - Y to X has the same correlation of X to Y
- I the values are between -1 (anticorrelation) and 1 (correlation)
- Usually to variables are not 100% correlated (c=1) or not at all correlated c=0) there are always a partial correlation between variables

Correlation graph



	Acid	Antoc	Brix	Carot	clorof
acid	1.00	-0.48	0.34	0.15	-0.32
antoc	-0.48	1.00	-0.70	-0.56	0.88
brix	0.34	-0.70	1.00	0.30	-0.76
carot	0.15	-0.56	0.30	1.00	-0.25
clorof	-0.32	0.88	-0.76	-0.25	1.00

Multivariate probability

- What is the probability that a peach has at the same time a concentration of carotene of 0.40±0.02 and of chlorophyll of 4.31±0.23?
- To answer this question we have to know the joint probability.
  - Thera re two possibilities :
    - Y and X are independents $\Rightarrow$ P(X,Y)=P(X)+P(Y)
    - Y and X are dependents to each other $\Rightarrow$ P(X,Y)
  - In the first case we have the product of the PDF monovariate functions (PDF)
  - In the second case we have to introduce the bivariate PDF

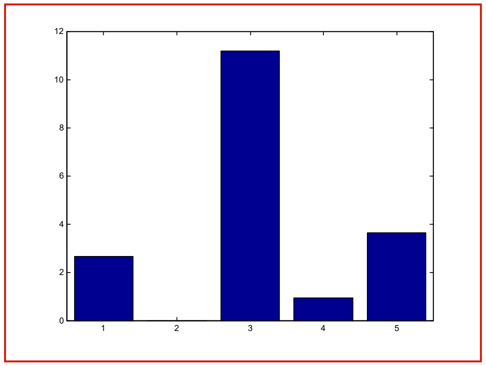
Multivariate data

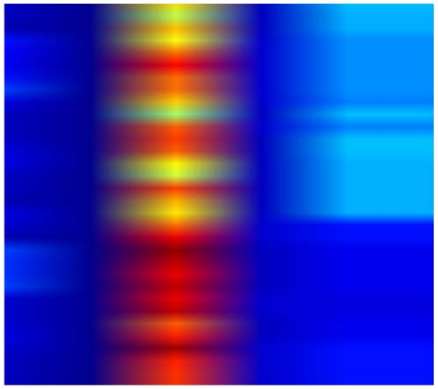
- The multivariate value is a vector
- The multivariate data are shown as a matrix
- Example :
  - A population of 20 peaches having measured 5 variables:
  - Total acidity, anthocyanin, brix, carotene e chlorophyll

variabili

campioni

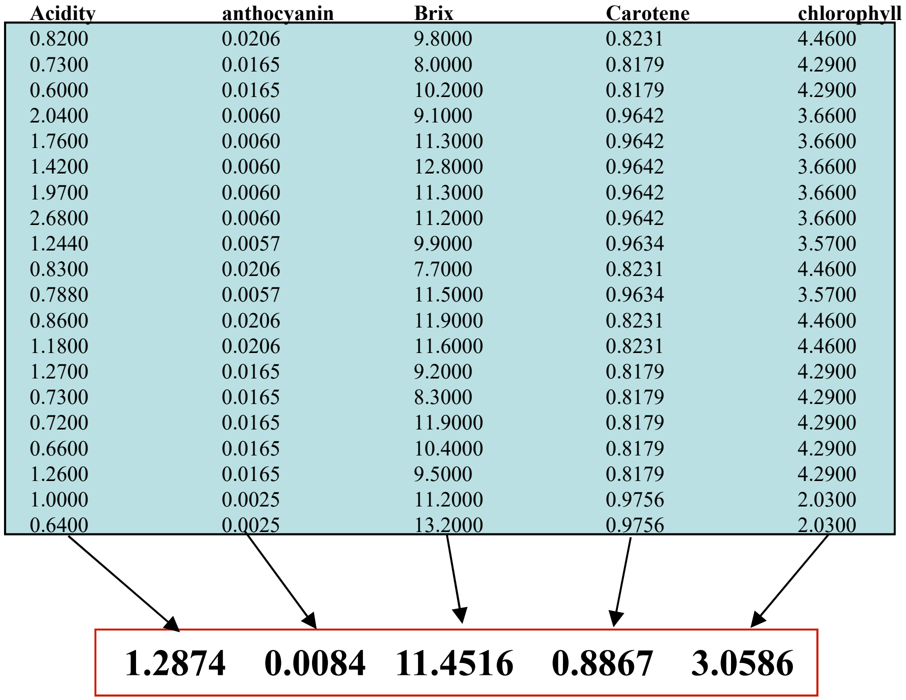
Acidity	anthocyanin	Brix	Carotene	chlorophyll
0.8200	0.0206	9.8000	0.8231	4.4600
0.7300	0.0165	8.0000	0.8179	4.2900
0.6000	0.0165	10.2000	0.8179	4.2900
2.0400	0.0060	9.1000	0.9642	3.6600
1.7600	0.0060	11.3000	0.9642	3.6600
1.4200	0.0060	12.8000	0.9642	3.6600
1.9700	0.0060	11.3000	0.9642	3.6600
2.6800	0.0060	11.2000	0.9642	3.6600
1.2440	0.0057	9.9000	0.9634	3.5700
0.8300	0.0206	7.7000	0.8231	4.4600
0.7880	0.0057	11.5000	0.9634	3.5700
0.8600	0.0206	11.9000	0.8231	4.4600
1.1800	0.0206	11.6000	0.8231	4.4600
1.2700	0.0165	9.2000	0.8179	4.2900
0.7300	0.0165	8.3000	0.8179	4.2900
0.7200	0.0165	11.9000	0.8179	4.2900
0.6600	0.0165	10.4000	0.8179	4.2900
1.2600	0.0165	9.5000	0.8179	4.2900
1.0000	0.0025	11.2000	0.9756	2.0300
0.6400	0.0025	13.2000	0.9756	2.0300





Multivariate average

- The average multivariate data is a vector made by the average of the single variable (the column).



Acidity	anthocyanin	Brix	Carotene	chlorophyll
0.8200	0.0206	9.8000	0.8231	4.4600
0.7300	0.0165	8.0000	0.8179	4.2900
0.6000	0.0165	10.2000	0.8179	4.2900
2.0400	0.0060	9.1000	0.9642	3.6600
1.7600	0.0060	11.3000	0.9642	3.6600
1.4200	0.0060	12.8000	0.9642	3.6600
1.9700	0.0060	11.3000	0.9642	3.6600
2.6800	0.0060	11.2000	0.9642	3.6600
1.2440	0.0057	9.9000	0.9634	3.5700
0.8300	0.0206	7.7000	0.8231	4.4600
0.7880	0.0057	11.5000	0.9634	3.5700
0.8600	0.0206	11.9000	0.8231	4.4600
1.1800	0.0206	11.6000	0.8231	4.4600
1.2700	0.0165	9.2000	0.8179	4.2900
0.7300	0.0165	8.3000	0.8179	4.2900
0.7200	0.0165	11.9000	0.8179	4.2900
0.6600	0.0165	10.4000	0.8179	4.2900
1.2600	0.0165	9.5000	0.8179	4.2900
1.0000	0.0025	11.2000	0.9756	2.0300
0.6400	0.0025	13.2000	0.9756	2.0300

1.2874 0.0084 11.4516 0.8867 3.0586

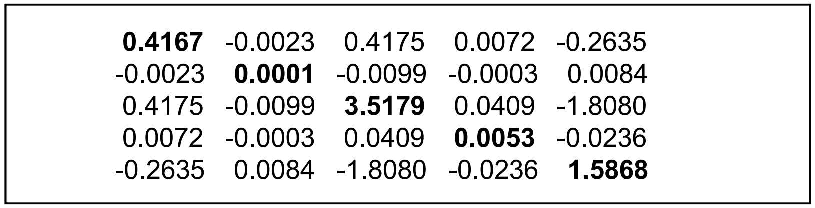
variance in Multivariate data

- In multivariate data the variance is a matrix called covariance matrix
- The covariance matrix is linked to correlation and correlation matrix
- The covariance matrix is defined as:

$$cov(X) = \Sigma = E\left[(x-m)^T \cdot (x-m)\right]$$

- The covariance matrix is symmetric and quadratic. The dimensions are equal to the variables measured
- each element on the principal diagonal of the covariance matrix is just the variance of each of the elements in the vector
- The other elements of the covariance matrix are proportional to the correlation coefficients (ρ)

$$\Sigma_{ii} = \sigma_i^2 \quad ; \quad \Sigma_{ik} = \rho_{ik}\sigma_i\sigma_k$$



0.4167	-0.0023	0.4175	0.0072	-0.2635
-0.0023	**0.0001**	-0.0099	-0.0003	0.0084
0.4175	-0.0099	**3.5179**	0.0409	-1.8080
0.0072	-0.0003	0.0409	**0.0053**	-0.0236
-0.2635	0.0084	-1.8080	-0.0236	**1.5868**

Covariance and Correlation

- The covariance matrix can be written as:

$$\Sigma = \Gamma \cdot R \cdot \Gamma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \\ \dots & & \dots & \\ 0 & & & \sigma_n \end{bmatrix} \cdot \begin{bmatrix} 1 & \rho_{21} & \dots & \rho_{n1} \\ \rho_{12} & 1 & & \\ \dots & & \dots & \\ \rho_{1n} & & & 1 \end{bmatrix} \cdot \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & \\ \dots & & \dots & \\ 0 & & & \sigma_n \end{bmatrix}$$

  - Where R is the correlation matrix

Correlation and independence

- Two variables are independent if :

$$E\left[X \cdot Y\right] = E\left[X\right] \cdot E\left[Y\right]$$

  - This condition is also called linear independence

- Two variables are independent if:

$$P\left[X \cdot Y\right] = P\left[X\right] \cdot P\left[Y\right]$$

PDF multivariate 1

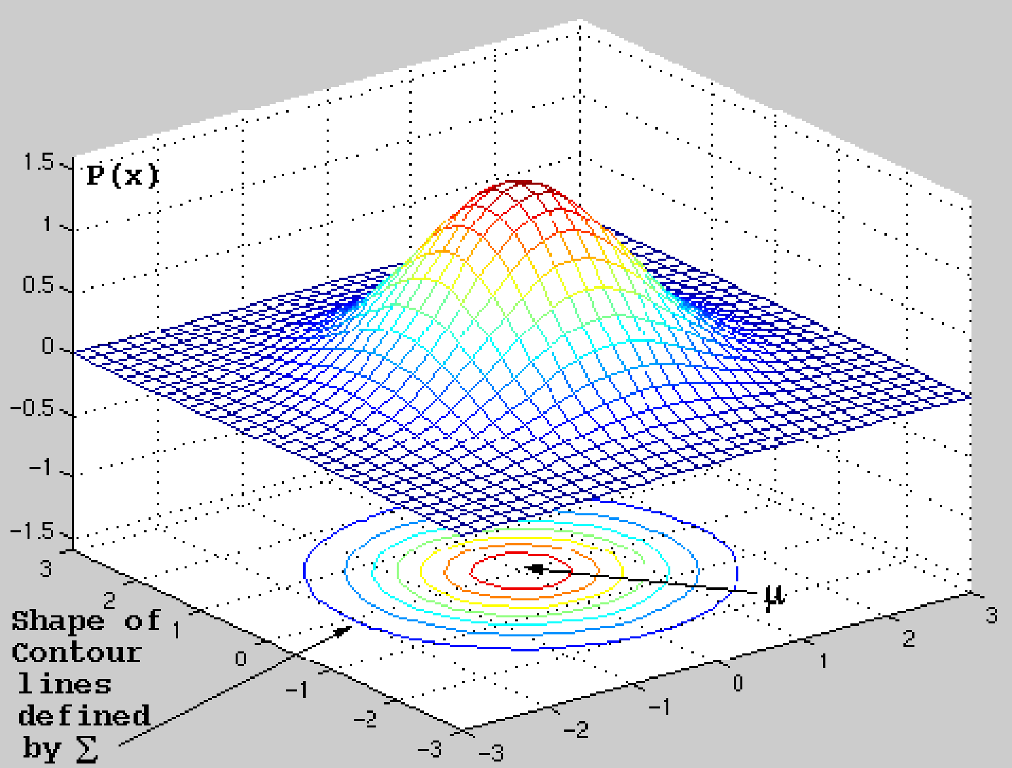
$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1d} \\ \vdots & & \vdots \\ \sigma_{d1} & \cdots & \sigma_{dd} \end{bmatrix}$$

$$f_{\vec{X}}(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n |\mathrm{C}|}} \exp\left(-\frac{1}{2} (\vec{x} - \vec{\mu})^{\mathrm{T}} \mathrm{C}^{-1} (\vec{x} - \vec{\mu}) \right)$$

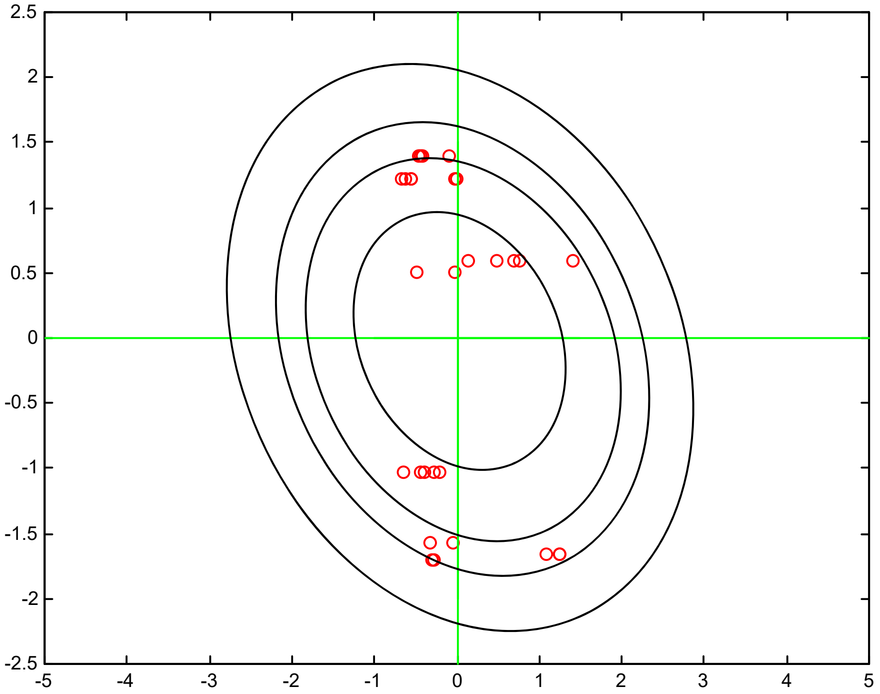
where n is the dimensionality of the space under consideration.

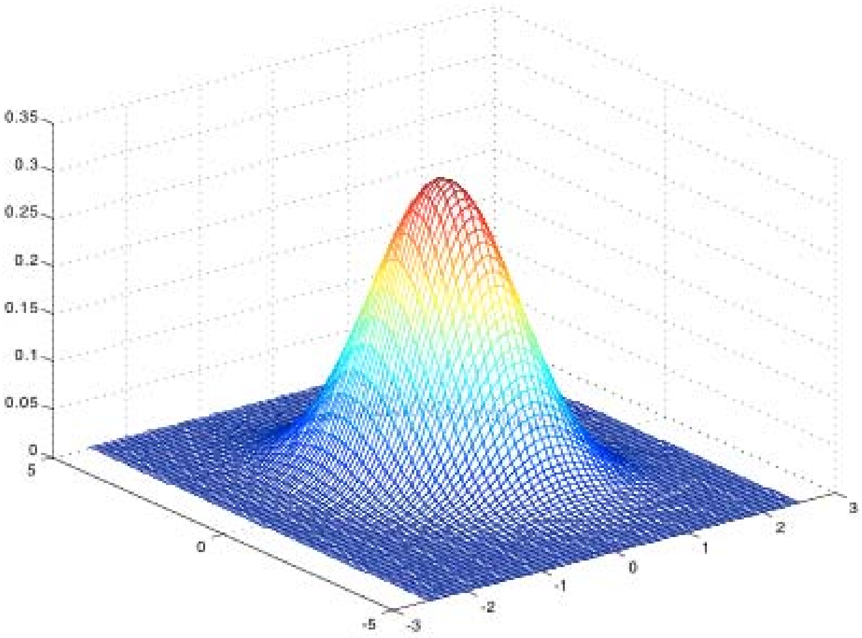
$$p(x) = \frac{1}{\sqrt{2\Pi}\sqrt{\Sigma}} \exp\left[\frac{-1}{2} (x - \mu)^t \Sigma^{-1} (x - \mu) \right]$$

The probability points are quadratic forms. **Ellipse** for the PDF bivariate.



Shape of Contour lines defined by Σ


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-5
-4
-3
-2
-1
0
1
2
3
4
5


0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
5
0
-5
-3
-2
-1
0
1
2
3

Instruments review

Instrument Converts information stored in the physical or chemical characteristics of the analyte into useful information

Require a source of energy to stimulate measurable response from analyte

Data domains

- Methods of encoding information electrically
- Nonelectrical domains
- Electrical domains
  - Analog, Time, Digital

Detector

Device that indicates a change in one variable in its environment (eg., pressure, temp, particles)

Can be mechanical, electrical, or chemical

Sensor

Analytical device capable of monitoring specific chemical species continuously and reversibly

Transducer

Devices that convert information in nonelectrical domains to electrical domains and the converse

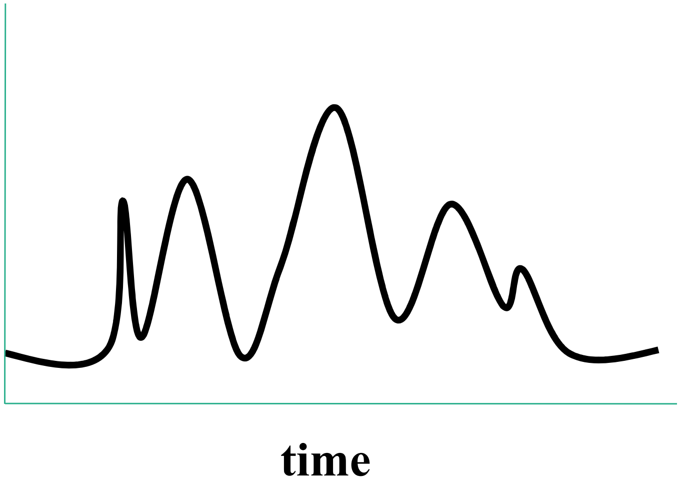
Method Validation

- Specificity
- Linearity
- Accuracy
- Precision
- Range
- Limits of Detection and Quantitation

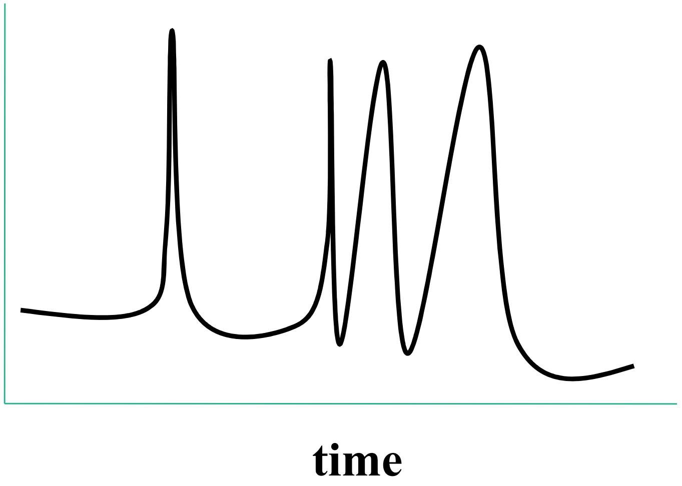
Method Validation - Specificity

- How well an analytical method distinguishes the analyte from everything else in the sample.

- Baseline separation



vs.

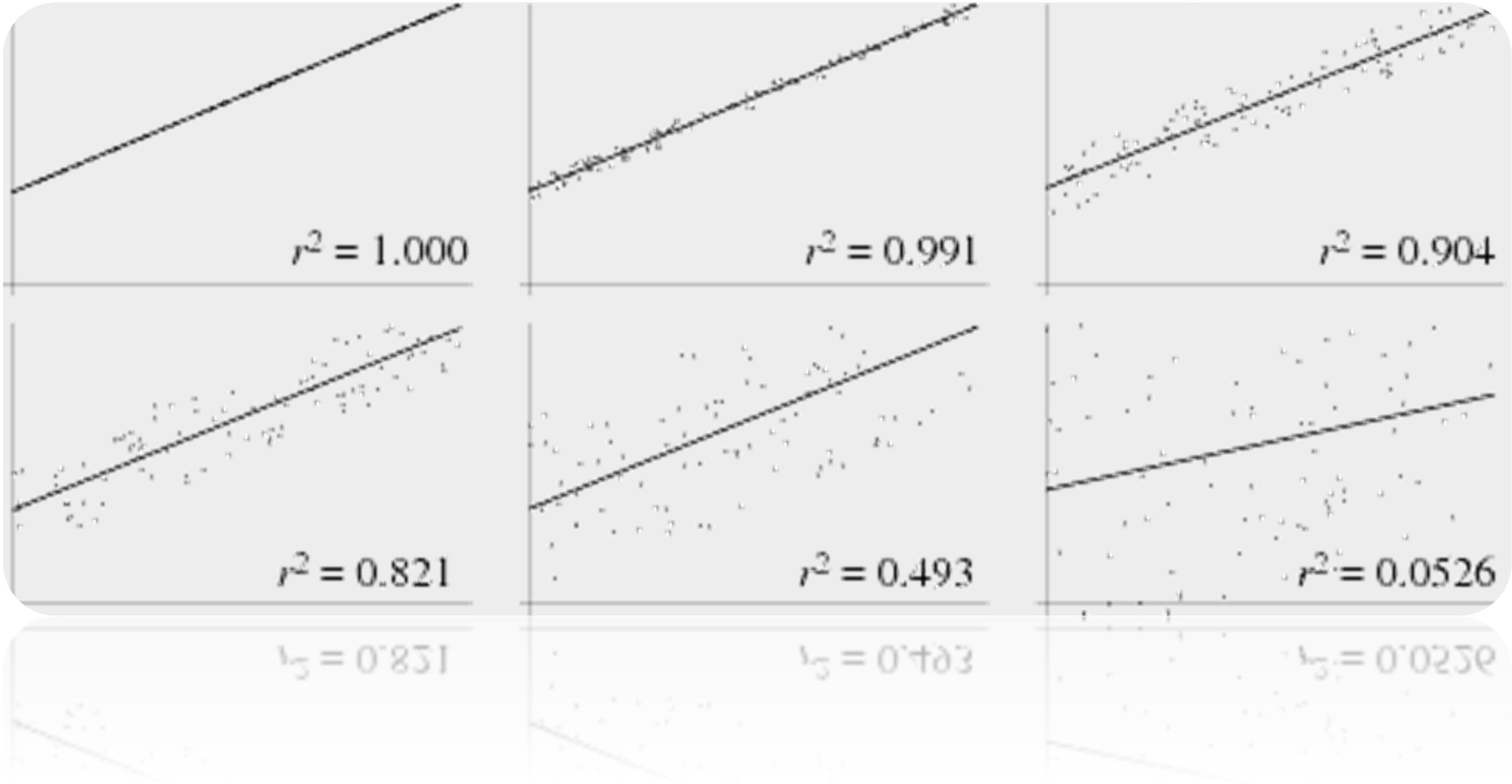


Method Validation- Linearity

- How well a calibration curve follows a straight line.
- R^2 (Square of the correlation coefficient)

$$R^2 = \frac{[\sum(x_i - \overline{x})(y_i - \overline{y})]^2}{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}$$

Method Validation- Linearity

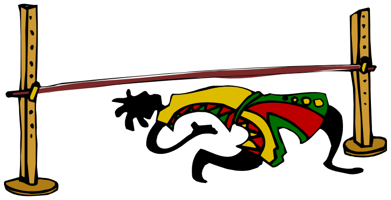


Method Validation- LOD and LOQ

Sensitivity

- Limit of detection (LOD) – "the lowest content that can be measured with reasonable statistical certainty."

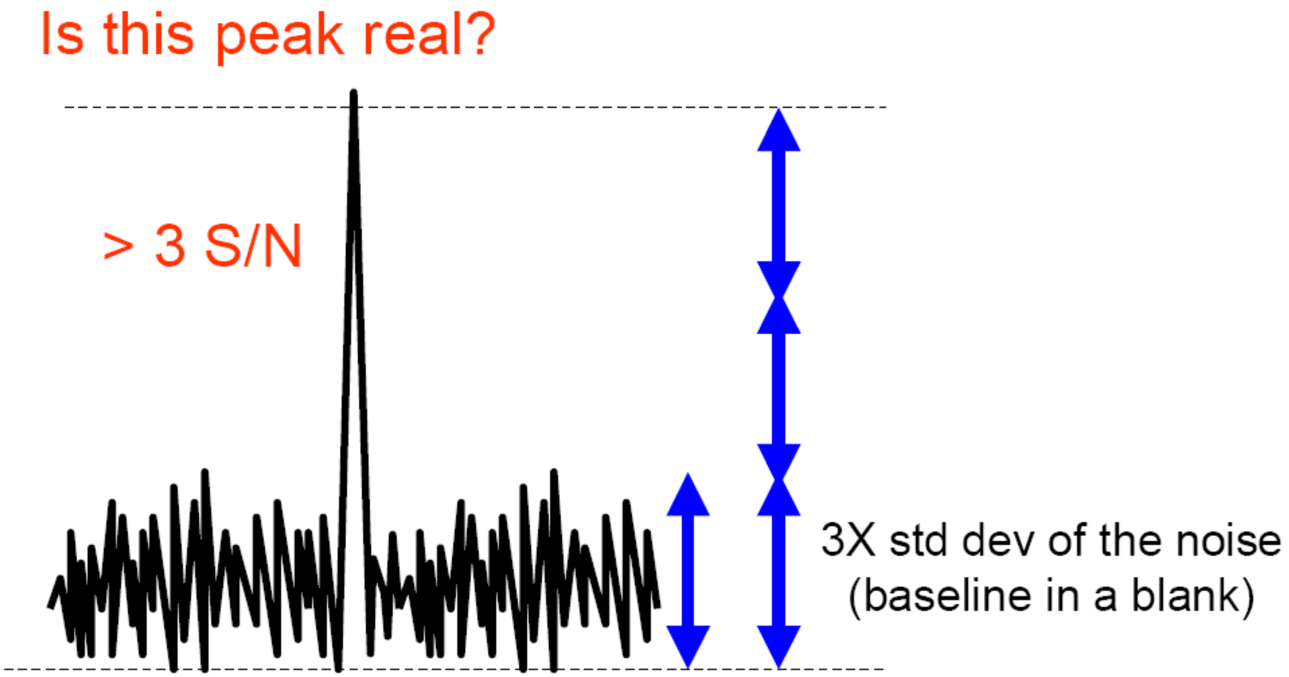
- Limit of quantitative measurement (LOQ) – "the lowest concentration of an analyte that can be determined with acceptable precision (repeatability) and accuracy under the stated conditions of the test."



- How low can you go?

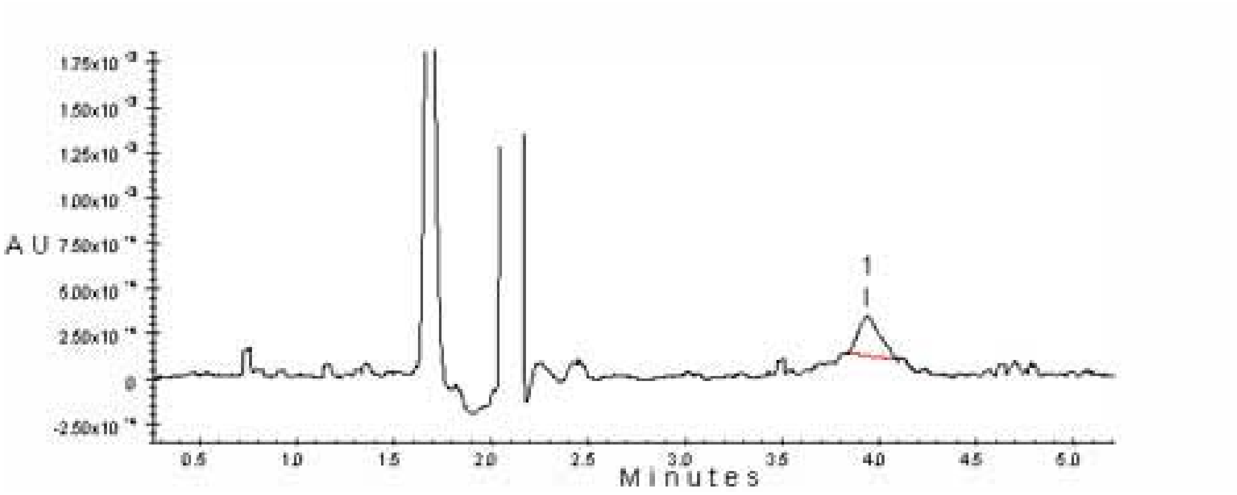
Limit of Detection (LOD)

- Typically 3 times the signal-to-noise (based on standard deviation of the noise)

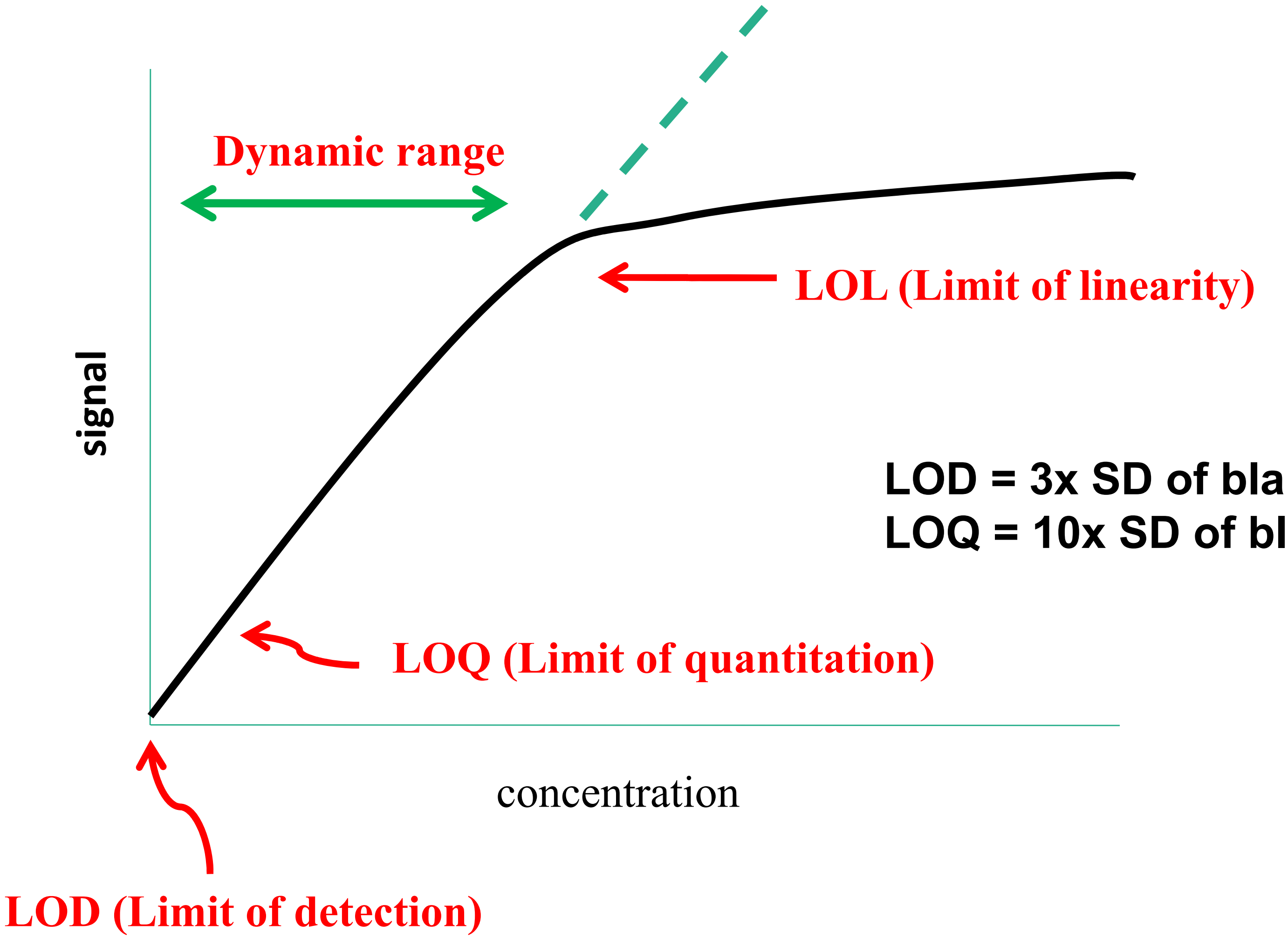


Limit of Linear Response (LOL)

- Point of saturation for an instrument detector so that higher amounts of analyte do not produce a linear response in signal.



Useful Range of an Analytical Method



Dynamic range

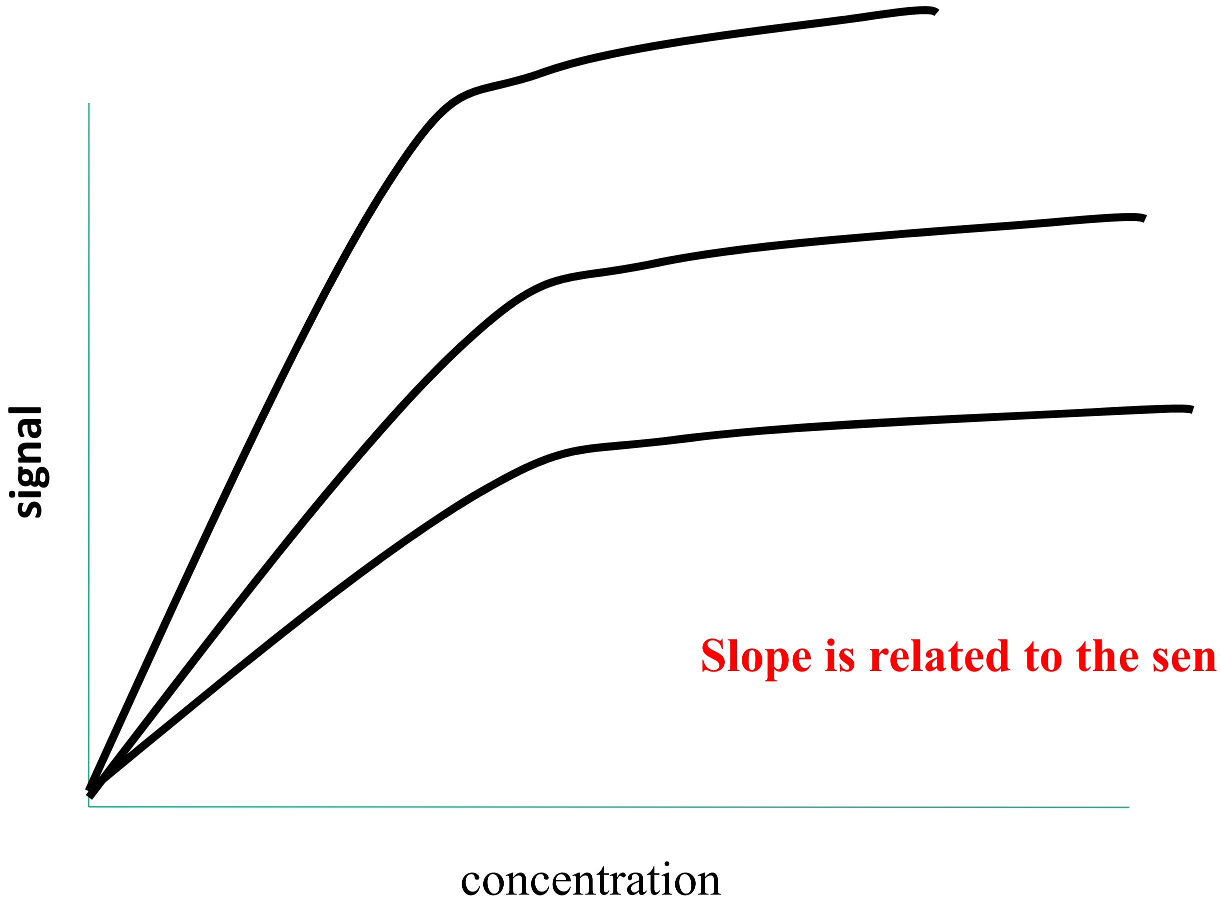
LOL (Limit of linearity)

LOQ (Limit of quantitation)

LOD (Limit of detection)

LOD = 3x SD of blank
LOQ = 10x SD of blank

Method Validation- Linearity



Slope is related to the sensitivity

Method Validation- Accuracy and Precision

- Accuracy – nearness to the truth
- Compare results from more than one analytical technique
- Analyze a blank spiked with known amounts of analyte.

Precision - reproducibility

Method Validation- LOD and LOQ

- Detection limit (lower limit of detection – smallest quantity of analyte that is “statistically” different from the blank.

- HOW TO:
- Measure signal from n replicate samples (n > 7)
- Compute the standard deviation of the measurments
- Signal detection limit: $y_{dl} = y_{blank} + 3s$
- $y_{sample} - y_{blank} = m \cdot$ sample concentration

- Detection limit: 3s/m
- Lower limit of quantitation (LOQ) : 10s/m

Example: sample concentrations: 5.0, 5.0, 5.2, 4.2, 4.6, 6.0, 4.9 nA
Blanks: 1.4, 2.2, 1.7, 0.9, 0.4, 1.5, 0.7 nA
The slope of the calibration curve for high conc. m= 0.229 nA/μM
What is the signal detection limit and the minimum detectable concentration?
What is the lower limit of quantitation?

Standard Addition

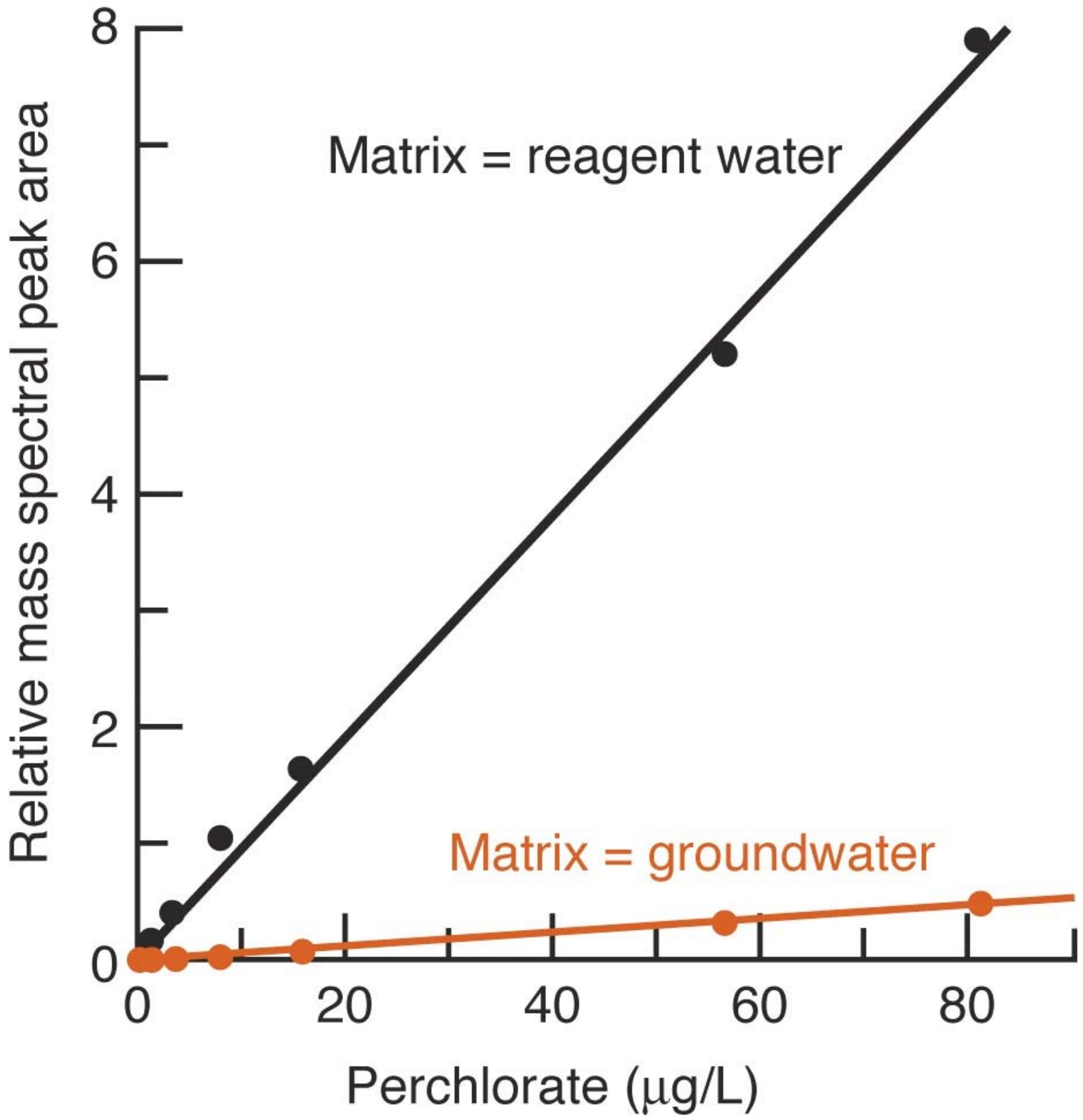
- **Standard addition** is a method to determine the amount of analyte in an unknown.
  - In standard addition, known quantities of analyte are added to an unknown.
  - We determine the analyte concentration from the increase in signal.
- Standard addition is often used when the sample is unknown or complex and when species other than the analyte affect the signal.
  - The **matrix** is everything in the sample other than the analyte and its affect on the response is called the **matrix effect**

The Matrix Effect

- The matrix effect problem occurs when the unknown sample contains many impurities.

- If impurities present in the unknown interact with the analyte to change the instrumental response or themselves produce an instrumental response, then a calibration curve based on pure analyte samples will give an incorrect determination

Calibration Curve for Perchlorate with Different Matrices



Perchlorate (ClO_4^-) in drinking water affects production of thyroid hormone. ClO_4^- is usually detected by mass spectrometry (Ch. 22), but the response of the analyte is affected by other species, so you can see the response of calibration standards is very different from real samples

Calculation of Standard Addition

- The formula for a standard addition is:

$$\frac{[X]_i}{[S]_f + [X]_f} = \frac{I_x}{I_{S+X}}$$

[X] is the concentration of analyte in the initial (i) and final (f) solutions, [S] is the concentration of standard in the final solution, and I is the response of the detector to each solution.

- But,

$$[X]_f = [X]_i\left(\frac{V_0}{V_f}\right) \quad \text{and} \quad [S]_f = [S]_i\left(\frac{V_s}{V_f}\right)$$

If we express the diluted concentration of analyte in terms of the original concentration, we can solve the problem because we know everything else.

Standard Addition Example

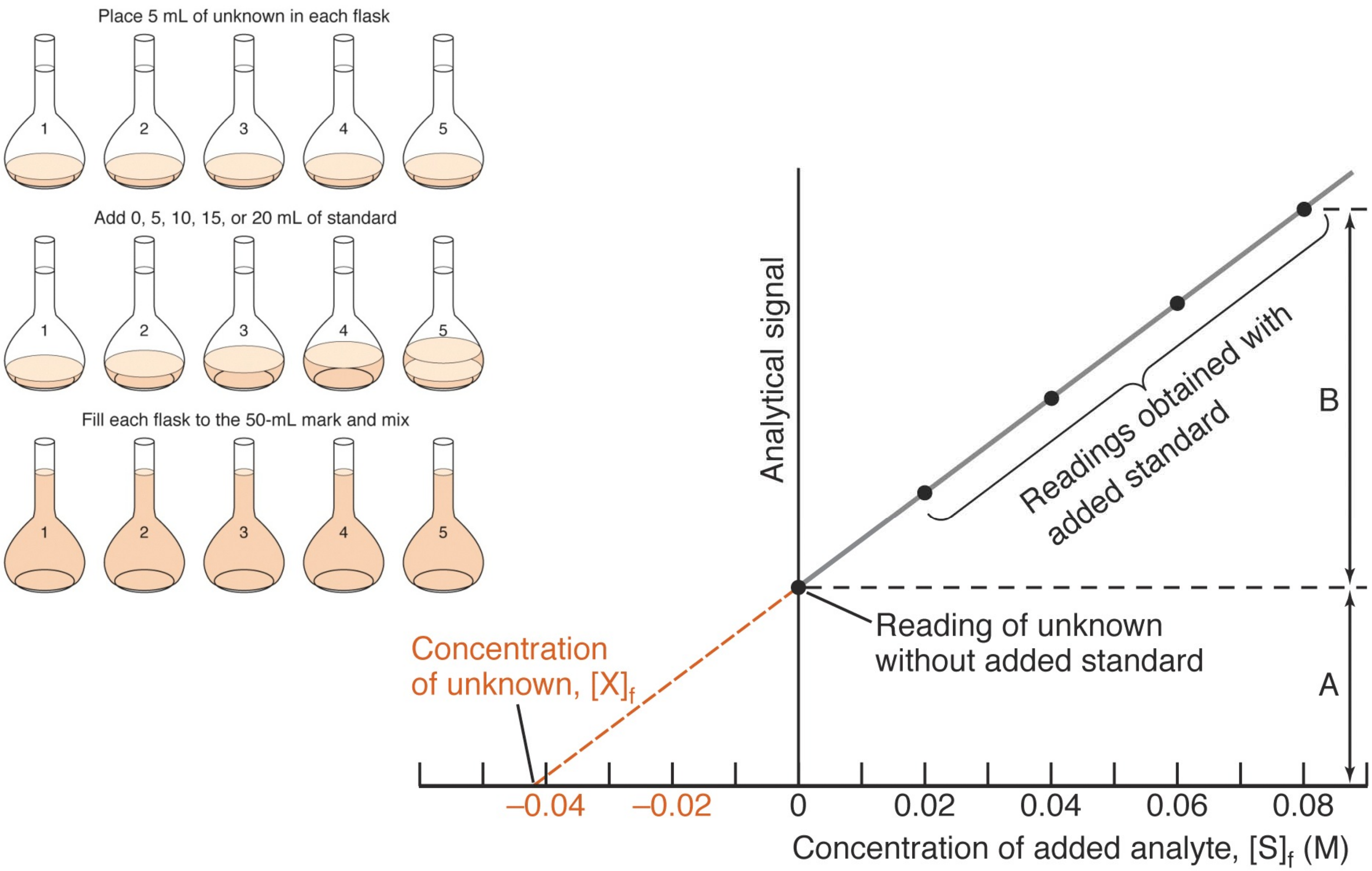
- Serum containing Na$^+$ gave a signal of 4.27 mv in an atomic emission analysis. 5.00 mL of 2.08 M NaCl were added to 95.0 mL of serum. The spiked serum gave a signal of 7.98 mV. How much Na$^+$ was in the original sample?

$$[\mathrm{X}]_f = [\mathrm{X}]_i \left(\frac{95.0 \text{ mL}}{100.0 \text{ mL}} \right) = 0.950[\mathrm{X}]_i$$

$$[\mathrm{S}]_f = [\mathrm{S}]_i \left(\frac{V_s}{V_f} \right) = (2.08 \text{ M}) \frac{5.00 \text{ mL}}{100.0 \text{ mL}} = 0.104\text{M}$$

$$\frac{[\mathrm{Na}^+]_i}{0.104 \text{ M} + 0.950[\mathrm{Na}^+]_f} = \frac{4.27 \text{ mV}}{7.98 \text{ mV}} \qquad [\mathrm{Na}^+]_i = 0.113 \text{ M}$$

Standard Additions Graphically


Place 5 mL of unknown in each flask
1
2
3
4
5
Add 0, 5, 10, 15, or 20 mL of standard
1
2
3
4
5
Fill each flask to the 50-mL mark and mix
1
2
3
4
5
Analytical signal
Readings obtained with added standard
B
A
Reading of unknown without added standard
Concentration of unknown, $[X]_f$
–0.04
–0.02
0
0.02
0.04
0.06
0.08
Concentration of added analyte, $[S]_f$ (M)

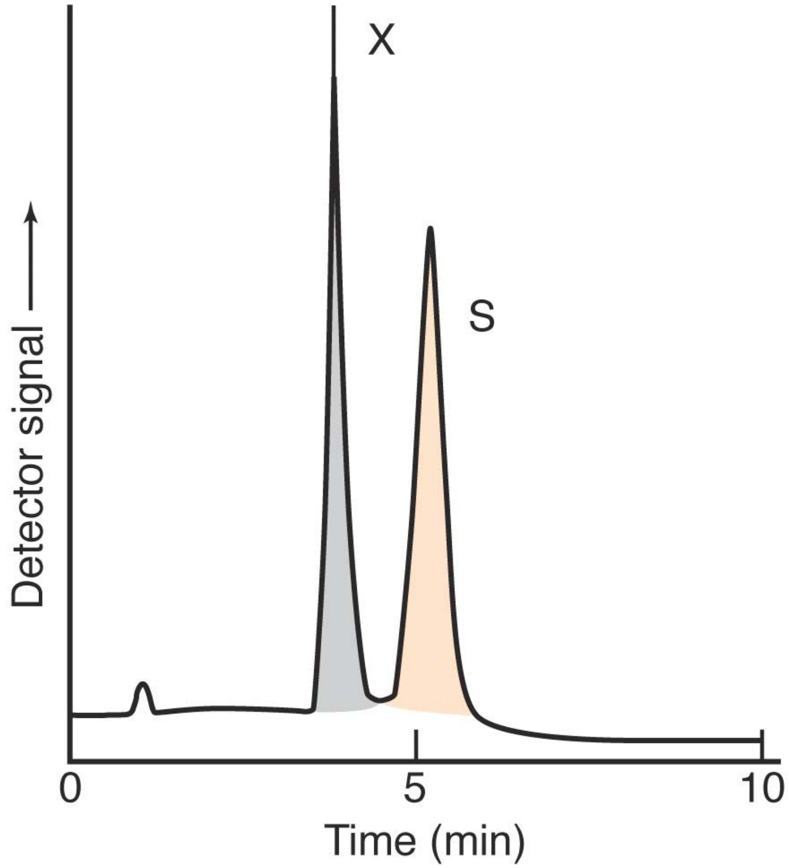
Internal Standards

- An **internal standard** is a known amount of a compound, different from the analyte, added to the unknown sample.

- Internal standards are used when the detector response varies slightly from run to run because of hard to control parameters.
  - *e.g.* Flow rate in a chromatograph

- But even if absolute response varies, as long as the *relative* response of analyte and standard is the same, we can find the analyte concentration

Response Factors



For an internal standard, we prepare a mixture with a known amount of analyte and standard. The detector usually has a different response for each species, so we determine a **response factor** for the analyte:

$$\frac{\mathrm{A_X}}{[\mathrm{X}]} = F\left(\frac{\mathrm{A_S}}{[\mathrm{S}]}\right)$$

[X] and [S] are the concentrations of analyte and standard after they have been mixed together.

$$\frac{\text{Area of analyte signal}}{\text{Concentration of analyte}} = F\left(\frac{\text{area of standard signal}}{\text{Concentration of standard}}\right)$$

Internal Standard Example

- In an experiment, a solution containing 0.0837 M Na^+ and 0.0666 M K^+ gave chromatographic peaks of 423 and 347 (arbitrary units) respectively. To analyze the unknown, 10.0 mL of 0.146 M K^+ were added to 10.0 mL of unknown, and diluted to 25.0 mL with a volumetric flask. The peaks measured 553 and 582 units respectively. What is $[Na^+]$ in the unknown?

- First find the response factor, *F*

$$\frac{A_{Na}}{[Na^+]} = F\left(\frac{A_K}{[K^+]}\right)$$

$$F = \left(\frac{A_{Na}}{[Na^+]}\right) / \left(\frac{A_K}{[K^+]}\right) = \frac{423}{0.0837} / \frac{347}{0.0666} = 0.970$$

Internal Standard Example (Cont.)

- Now, what is the concentration of K⁺ in the mixture of unknown and standard?

$$\left[\mathrm{K}^+\right] = (0.146\mathrm{M})\left(\frac{10\ \mathrm{mL}}{25.0\ \mathrm{mL}}\right) = 0.05484\ \mathrm{M}$$

- Now, you know the response factor, *F*, and you know how much standard, K⁺ is in the mixture, so we can find the concentration of Na⁺ in the mixture.

$$\frac{\mathrm{A_{Na}}}{\left[\mathrm{Na}^+\right]} = F\left(\frac{\mathrm{A_K}}{\left[\mathrm{K}^+\right]}\right) \qquad \frac{553}{\left[\mathrm{Na}^+\right]} = (0.970)\left(\frac{582}{0.0584\ \mathrm{M}}\right) \qquad \left[\mathrm{Na}^+\right] = 0.0572\ \mathrm{M}$$

- Na⁺ unknown was diluted in the mixture by K⁺, so the Na⁺ concentration in the unknown was:

$$\left[\mathrm{Na}^+\right] = (0.0572\ \mathrm{M})\left(\frac{25\ \mathrm{mL}}{10.0\ \mathrm{mL}}\right) = 0.143\ \mathrm{M}$$