CSE 487/587
Data Intensive Computing

Lecture 9: Basics of Statistics

Vipin Chaudhary

vipin@buffalo.edu

716.645.4740
305 Davis Hall

Overview of Today's Lecture

- Statistical Methods for Big Data

Statistics?

- Methods for organizing, summarizing, and interpreting information
- Many tool boxes exist
  - How do you know which tool to use?
    1. What do you want to know?
    2. What type of data do you have?
  - Two main branches:
    - *Descriptive statistics*
    - *Inferential statistics*

Descriptive and Inferential statistics

Descriptive Statistics:

Tools for summarizing, organizing, simplifying data

Tables & Graphs

Measures of Central Tendency

Measures of Variability

Examples:

Average rainfall in Manchester last year

Number of car thefts in last year

Your test results

Percentage of males in our class

Inferential Statistics:

Data from *sample* used to draw inferences about *population*

Generalizing beyond actual observations

Generalize from a *sample* to a *population*

Statistical terms

- Population
  - complete set of individuals, objects or measurements
- Sample
  - a sub-set of a population
- Variable
  - a characteristic which may take on different values
- Data
  - numbers or measurements collected
- A parameter is a characteristic of a population
  - e.g., the *average* height of all Britons.
- A statistic is a characteristic of a sample
  - e.g., the *average* height of a sample of Britons.

Measurement scales

- Measurements can be qualitative or quantitative and are measured using four different scales

1. Nominal or categorical scale

- uses numbers, names or symbols to classify objects
- e.g. types of properties
  - Houses, condos, bungalows, co-ops
  - Where people live in - states

Measurement scales

2. Ordinal Scale

- ranking scale
- objects are placed in order
- divisions or gaps between objects may no be equal

Example: Patient Pain Scale from 1-10

- Pain difference between 3 and 4 might be very different than from 7 to 8.

Measurement scales

3. Interval Scale

- equality of length between objects
- no true zero
- Difference between two values is meaningful

Example: Temperature scales

Fahrenheit: Fahrenheit established 0°F as the stabilised temperature when equal amounts of ice, water, and salt are mixed. He then defined 96°F as human body temperature.

Celsius: 0 and 100 are arbitrarily placed at the melting and boiling points of water.

To go between scales is complicated: $$T(°C) = \frac{5}{9} \times \left[T(°F) - 32\right]$$

Interval Scale. You are also allowed to quantify the difference between two interval scale values but there is no natural zero. For example, temperature scales are interval data with 25C warmer than 20C and a 5C difference has some physical meaning. Note that 0C is arbitrary, so that it does not make sense to say that 20C is twice as hot as 10C.

Measurement scales

4. Ratio Scale

- an interval scale with a true zero
- ratio of any two scale points are independent of the units of measurement

Example: Length (metric/imperial)

- inches/centimetres = 2.54
- miles/kilometres = 1.609344

Ratio Scale. You are also allowed to take ratios among ratio scaled variables. It is now meaningful to say that 10 m is twice as long as 5 m. This ratio hold true regardless of which scale the object is being measured in (e.g. meters or yards).

Discrete and Continuous data

- Data consisting of numerical (quantitative) variables can be further divided into two groups: (1) *discrete* and (2) *continuous*.
  1. If the set of all possible values, when pictured on the number line, consists only of isolated points.
  2. If the set of all values, when pictured on the number line, consists of intervals.
- The most common type of discrete variable we will encounter is a *counting variable*.

Accuracy and precision

- Accuracy is the degree of conformity of a measured or calculated quantity to its actual (true) value
- Accuracy is closely related to precision, also called reproducibility or repeatability, the degree to which further measurements or calculations will show the same or similar results.

High accuracy but low precision

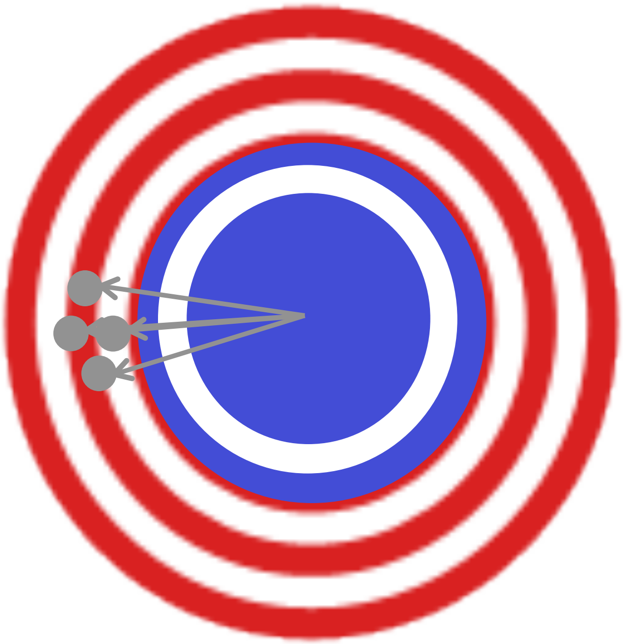
High precision but low accuracy

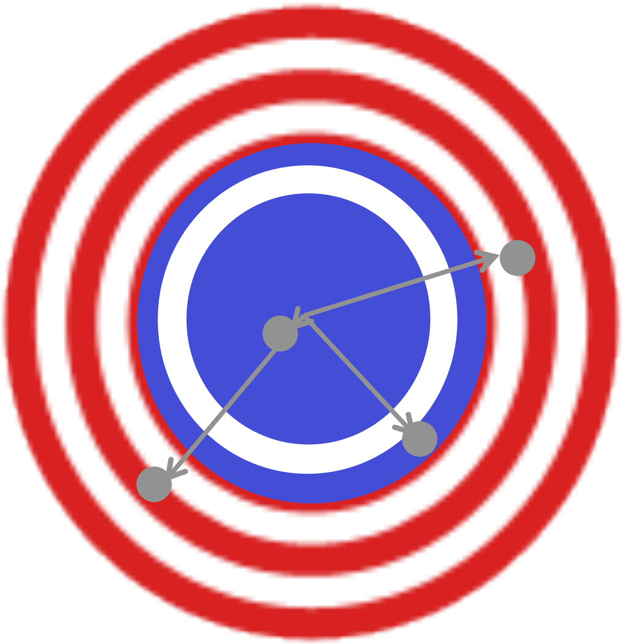
Accuracy and precision: The target analogy

High accuracy and high precision

Two types of error

- Systematic error
  - Poor accuracy
  - Definite causes
  - Reproducible
- Random error
  - Poor precision
  - Non-specific causes
  - Not reproducible





Systematic error

- Diagnosis
  - Errors have consistent signs
  - Errors have consistent magnitude
- Treatment
  - Calibration
  - Correcting procedural flaws
  - Checking with a different procedure

Random error

- Diagnosis
  - Errors have random sign
  - Small errors more likely than large errors
- Treatment
  - Take more measurements
  - Improve technique
  - Higher instrumental precision

Statistical graphs of data

- A picture is worth a thousand words!

- Graphs for numerical data:
  Histograms
  Frequency polygons
  Pie

- Graphs for categorical data
  Bar graphs
  Pie

Describing data

	Moment	Non-mean based measure
Center	Mean	Mode, median
Spread (Dispersion)	Variance (standard deviation)	Range, Interquartile range
Skew	Skewness	--
Peaked	Kurtosis	--

Central value

- Give information concerning the average or typical score of a number of scores
  - mean
  - median
  - mode

Central value: The Mean

- The Mean is a measure of *central value*
  - What most people mean by "average"
  - Sum of a set of numbers divided by the number of numbers in the set

$$\frac{1+2+3+4+5+6+7+8+9+10}{10} = \frac{55}{10} = 5.5$$

Central value: The Mean

Arithmetic average:

Sample

$$\overline{X} = \frac{\Sigma \mathrm{x}}{\mathrm{n}}$$

Population

$$\mu = \frac{\Sigma \mathrm{x}}{\mathrm{N}}$$

$$X = [1,2,3,4,5,6,7,8,9,10]$$

$$\sum X / n = 5.5$$

Central value: The Median

- Middlemost or most central item in the set of ordered numbers; it separates the distribution into two equal halves
- If *odd n*, middle value of sequence
  - if X = [1,2,4,6,9,10,12,14,17]
  - then 9 is the median
- If *even n*, average of 2 middle values
  - if X = [1,2,4,6,9,10,11,12,14,17]
  - then 9.5 is the median; i.e., (9+10)/2

- Median is not affected by extreme values

Central value: The Mode

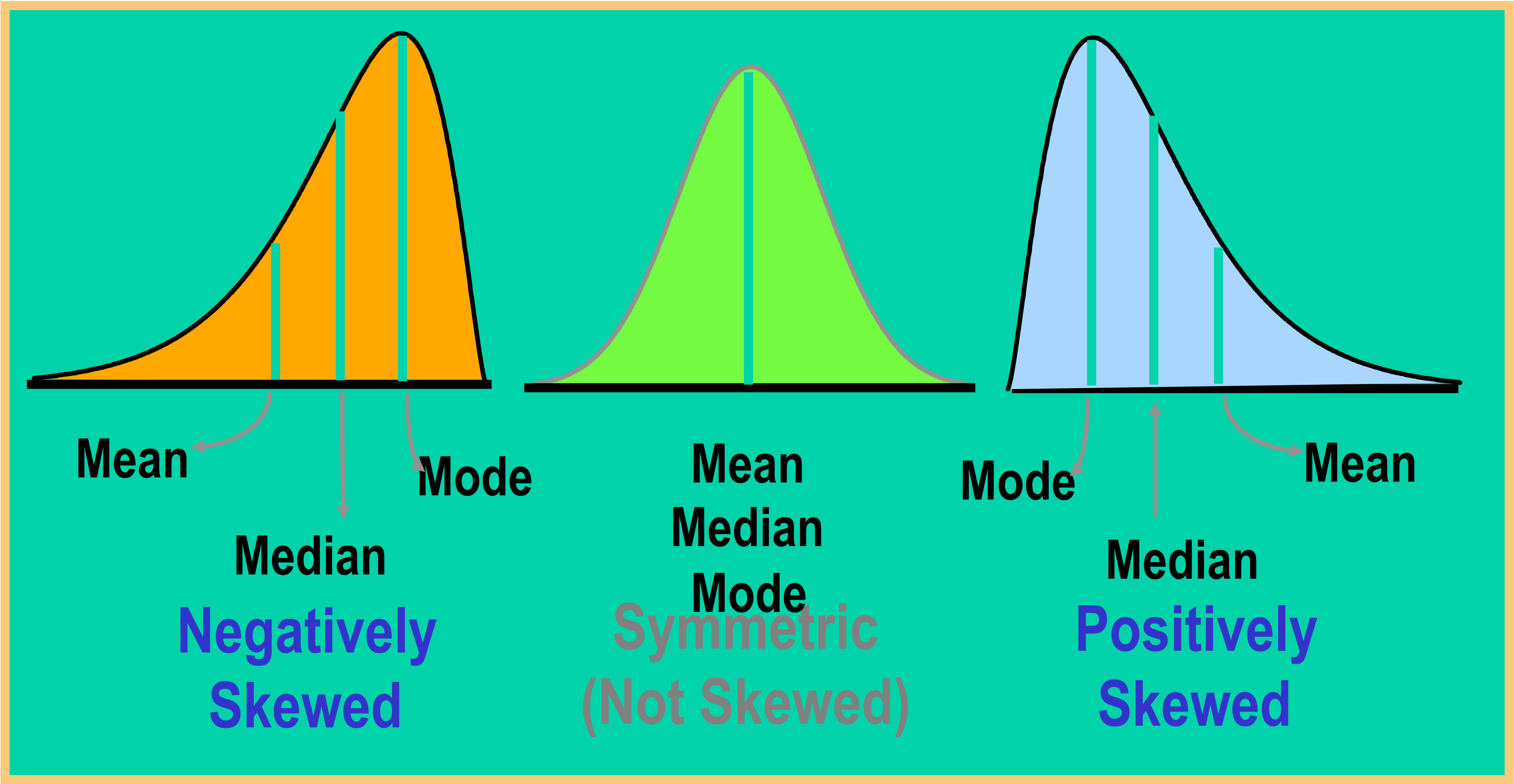
- The mode is the most frequently occurring number in a distribution
  - if X = [1,2,4,7,7,7,8,10,12,14,17]
  - then 7 is the mode
- Easy to see in a simple frequency distribution
- Possible to have no modes or more than one mode
  - *bimodal* and *multimodal*
- Don't have to be exactly equal frequency
  - *major mode, minor mode*

- Mode is not affected by extreme values

When to Use What

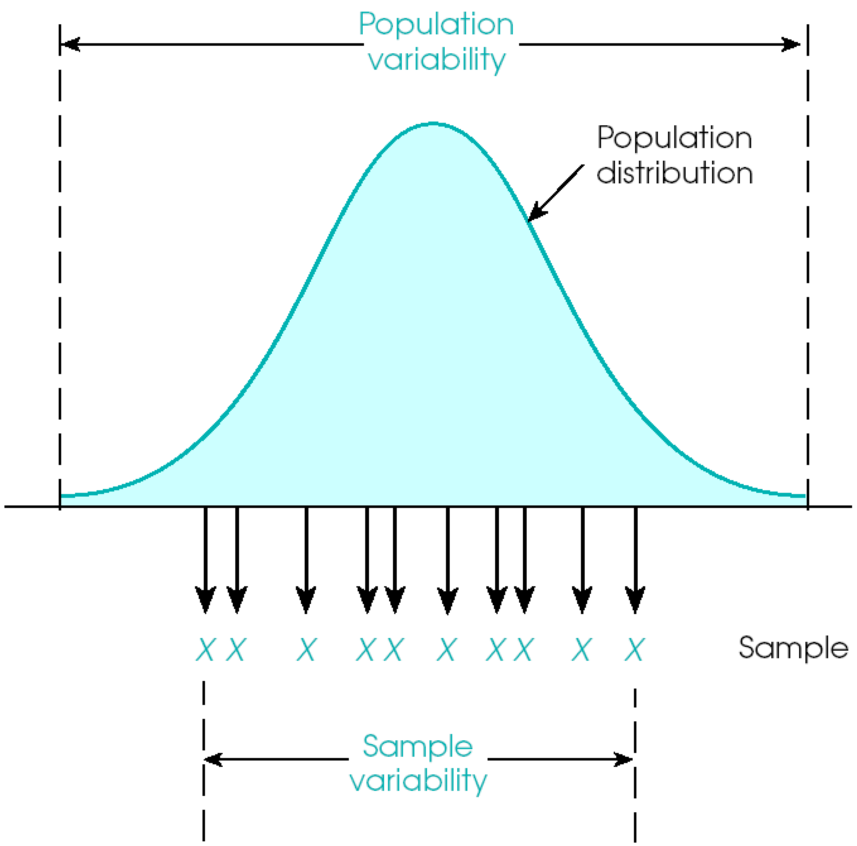
- Mean is a great measure. But, there are time when its usage is inappropriate or impossible.
  - Nominal data: Mode
  - The distribution is bimodal: Mode
  - You have ordinal data: Median or mode
  - Are a few extreme scores: Median

Mean, Median, Mode



Dispersion (Spread)

- Dispersion
  - How tightly clustered or how variable the values are in a data set.
- Example
  - Data set 1: [0,25,50,75,100]
  - Data set 2: [48,49,50,51,52]
  - Both have a mean of 50, but data set 1 clearly has greater *Variability* than data set 2.



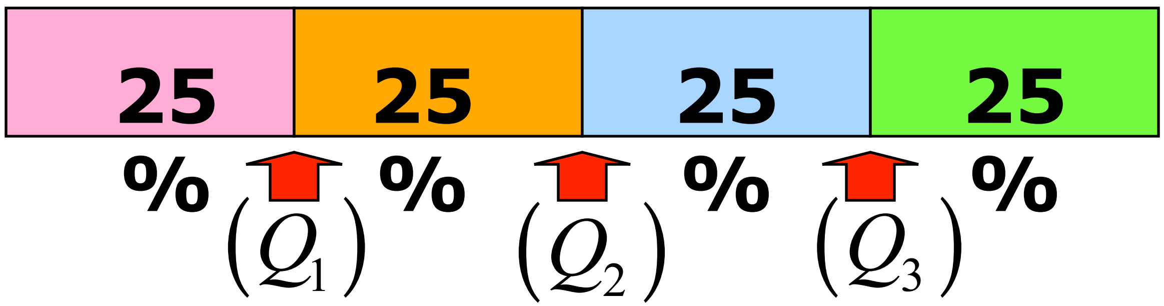
Dispersion: The Range

- The *Range* is one measure of dispersion
  - The range is the difference between the maximum and minimum values in a set
- Example
  - Data set 1: [1,25,50,75,100]; R: 100-1 +1 = 100
  - Data set 2: [48,49,50,51,52]; R: 52-48 + 1= 5
  - *The range ignores how data are distributed and only takes the extreme scores into account*

- $RANGE = (X_{largest} - X_{smallest}) + 1$

Quartiles

- Split Ordered Data into 4 Quarters



- Q_1 = first quartile
- Q_2 = second quartile= Median
- Q_3 = third quartile

Dispersion: Interquartile Range

- Difference between third & first quartiles
  - Interquartile Range = $Q_3 - Q_1$
- Spread in middle 50%
- Not affected by extreme values

Variance and standard deviation

Variance – average of squared deviates of each observation from mean of observations in a group of data

Population variance $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \mu\right)^2$$

Standard deviation is square root of variance

Sample variance $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

It is easy to show that $\sum_{i=1}^{n}\left(x_i - \bar{x}\right) = 0$ and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n\right]$$

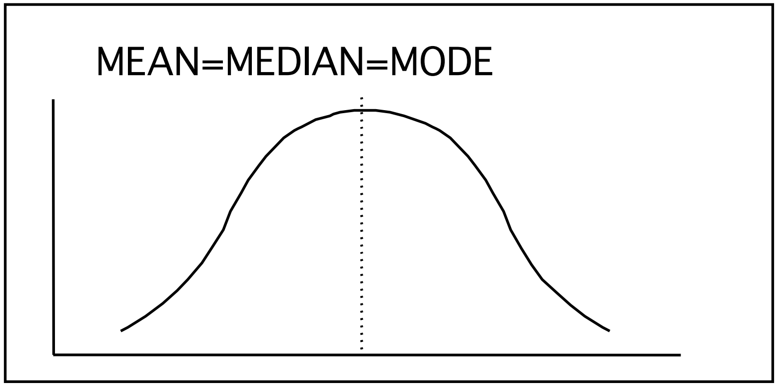
Dispersion: Standard Deviation

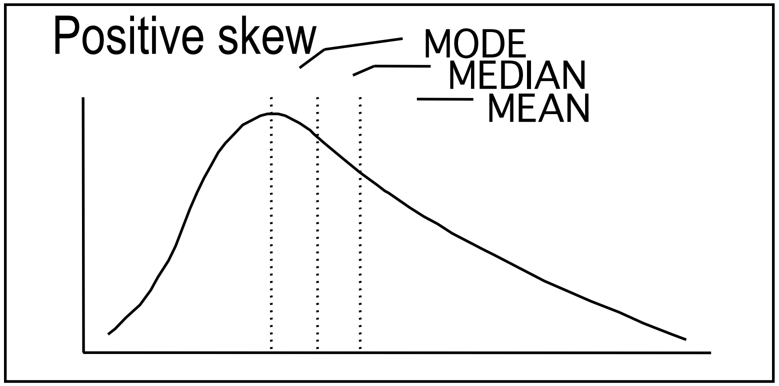
- let X = [3, 4, 5 ,6, 7]
- $\overline{X} = 5$
- $(X - \overline{X})$ = [-2, -1, 0, 1, 2]
  - ⇧ subtract $\overline{x}$ from each number in X
- $(X - \overline{X})^2$ = [4, 1, 0, 1, 4]
  - ⇧ squared deviations from the mean
- $\Sigma (X - \overline{X})^2 = 10$
  - ⇧ sum of squared deviations from the mean (SS)
- $\Sigma (X - \overline{X})^2 / n-1 = 10/5 = 2.5$
  - ⇧ average squared deviation from the mean
- $\sqrt{\Sigma (X - \overline{X})^2 / n-1} = \sqrt{2.5} = 1.58$
  - ⇧ square root of averaged squared deviation

$$s = \sqrt{\frac{\sum (X - \overline{X})^2}{n-1}}$$

Symmetry

Skew - asymmetry





$$\gamma_1 = \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3} = \frac{\mathrm{E}\left[(X-\mu)^3\right]}{\left(\mathrm{E}\left[(X-\mu)^2\right]\right)^{3/2}} = \frac{\kappa_3}{\kappa_2^{3/2}},$$

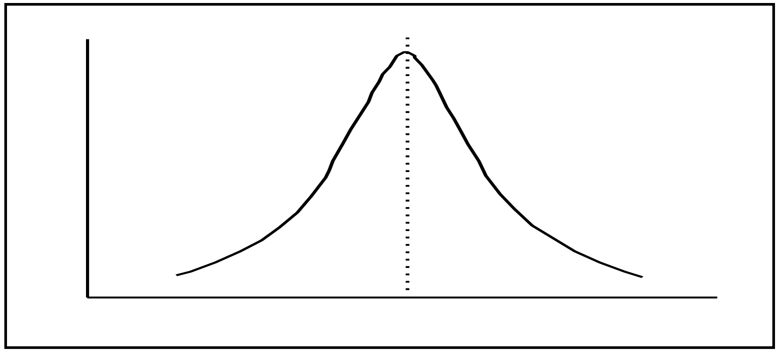
- IQ, SAT
  - "No skew"
  - "Zero skew"
  - Symmetrical

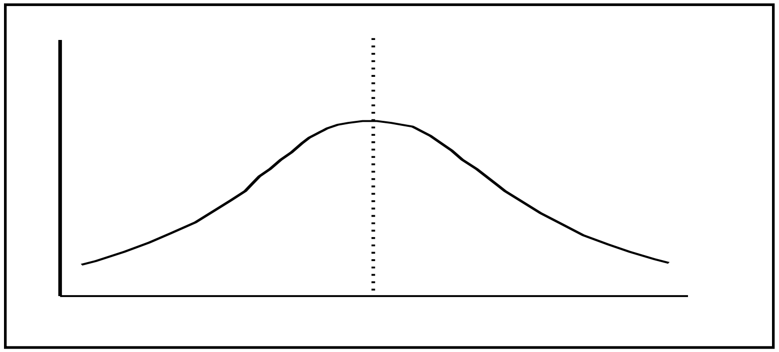
- GPA of CSE 487/587 students
  - "Negative skew"
  - " Left skew"

- Income
- Contribution to candidates
- Populations of countries
  - "Positive skew"
  - "Right skew"

Symmetry

Kurtosis - peakedness or flatness



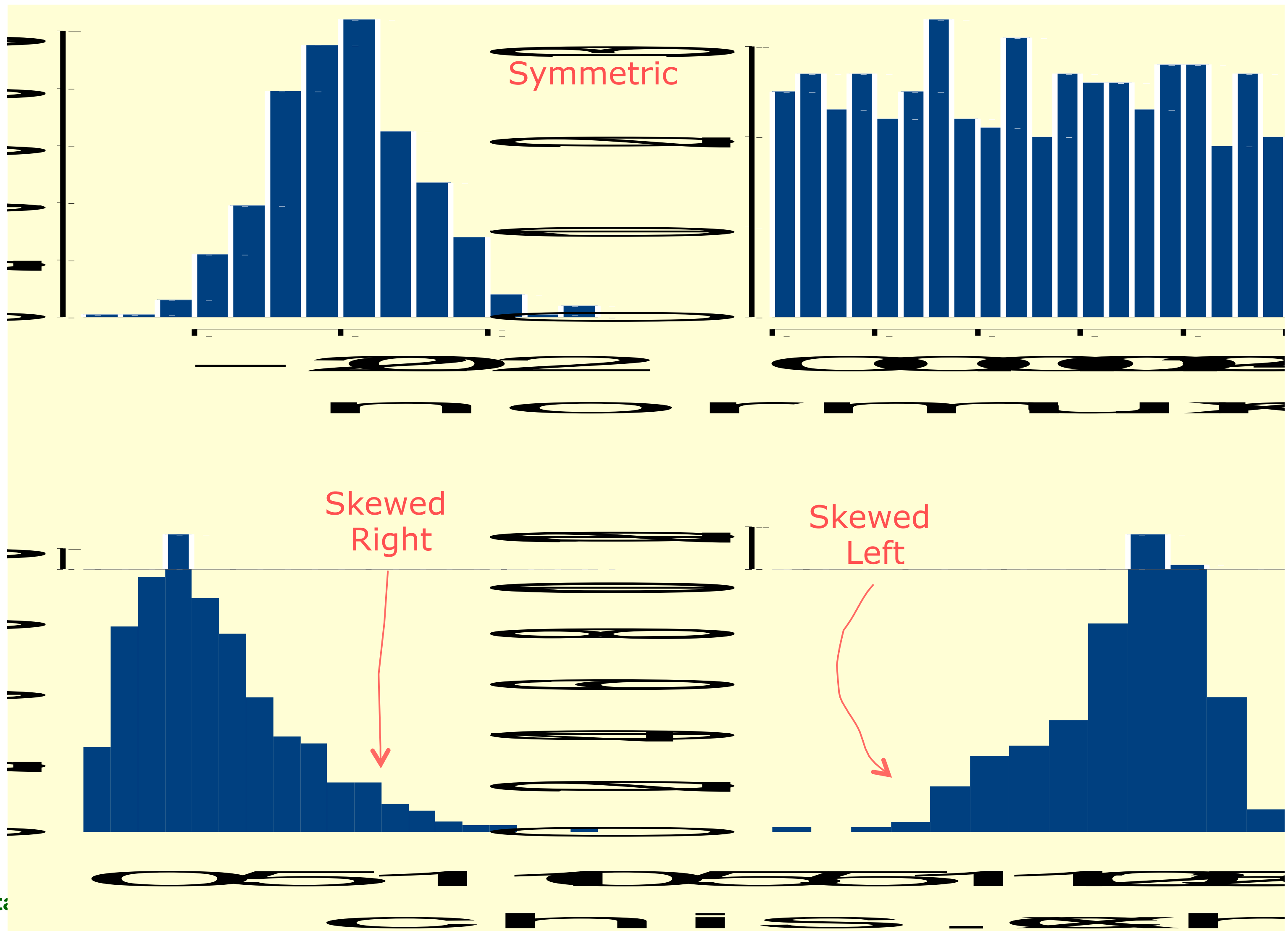


$$\beta_2 = \frac{\mathrm{E}[(X-\mu)^4]}{(\mathrm{E}[(X-\mu)^2])^2} = \frac{\mu_4}{\sigma^4}$$

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

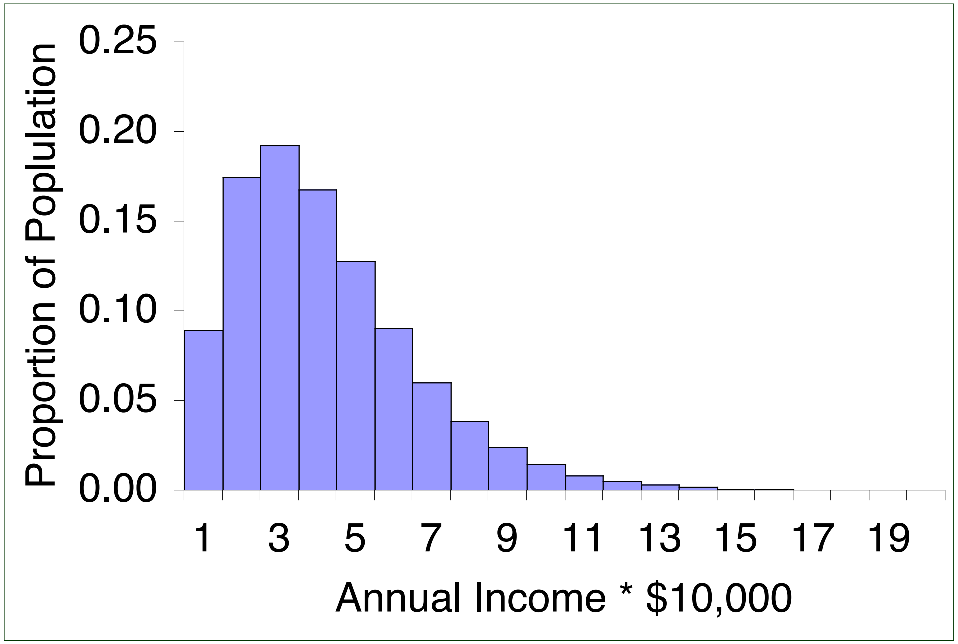
which is also known as **excess kurtosis**. The "minus 3" at the end of this formula is often explained as a correction to make the kurtosis of the normal distribution equal to zero.

Symmetrical vs. Skewed



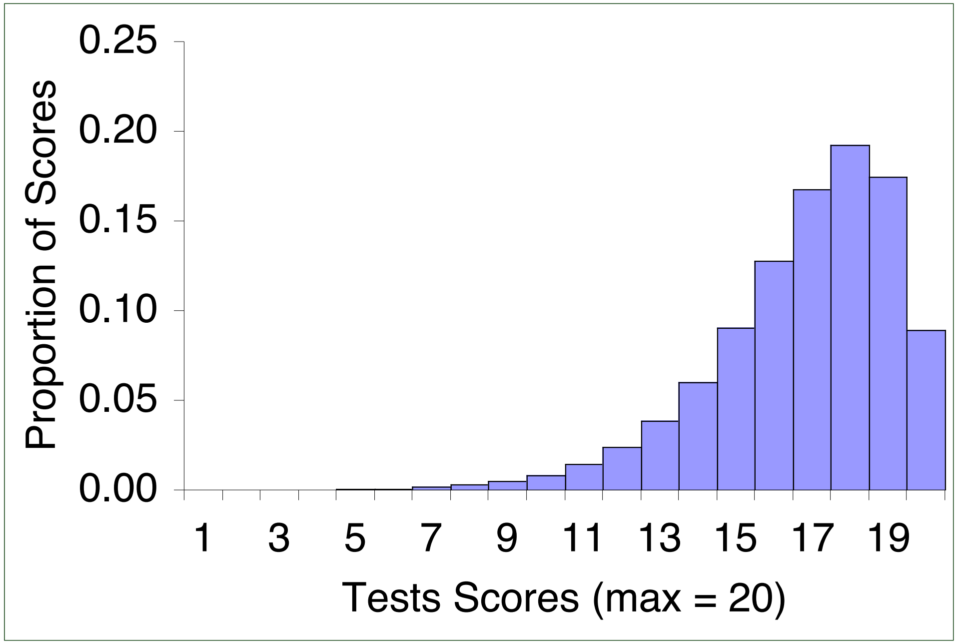
Skewed Frequency Distributions

- Positively skewed
  - AKA Skewed right
  - Tail trails to the right
  - ****** The skew describes the skinny end ******



Skewed Frequency Distributions

- Negatively skewed
  - Skewed left
  - Tail trails to the left



Symmetry: Skew

- The third `moment' of the distribution
- Skewness is a measure of the asymmetry of the probability distribution.
- Roughly speaking, a distribution has positive skew (right-skewed) if the right (higher value) tail is longer and negative skew (left-skewed) if the left (lower value) tail is longer (confusing the two is a common error).

$$g_1 = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}}$$

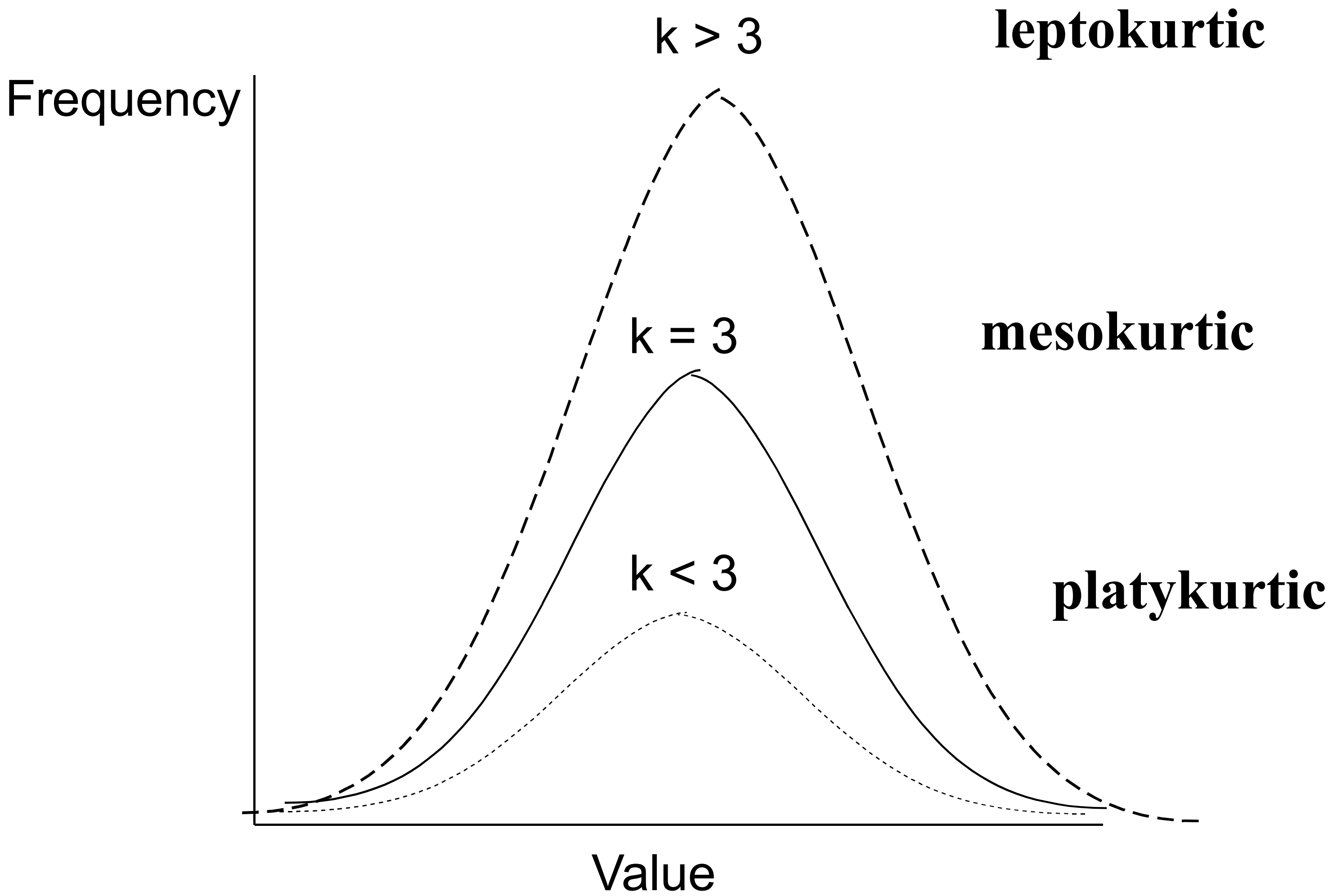
Symmetry: Kurtosis

- The fourth `moment' of the distribution

- A high kurtosis distribution has a sharper "peak" and fatter "tails", while a low kurtosis distribution has a more rounded peak with wider "shoulders".

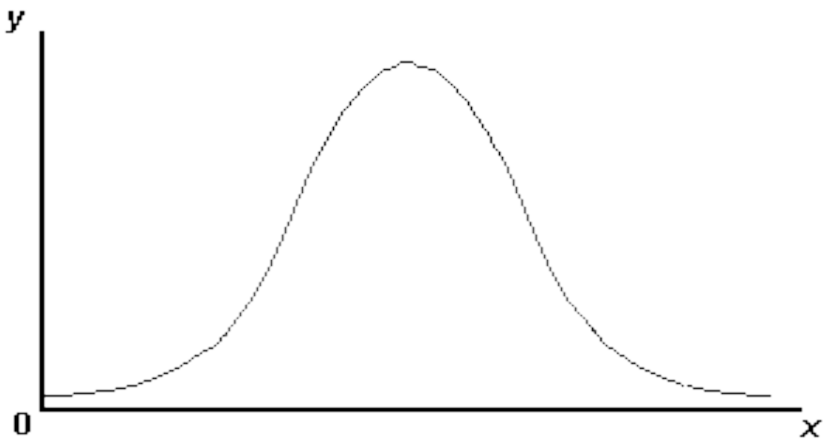
$$g_2 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

Kurtosis



k > 3
leptokurtic
Frequency
k = 3
mesokurtic
k < 3
platykurtic
Value

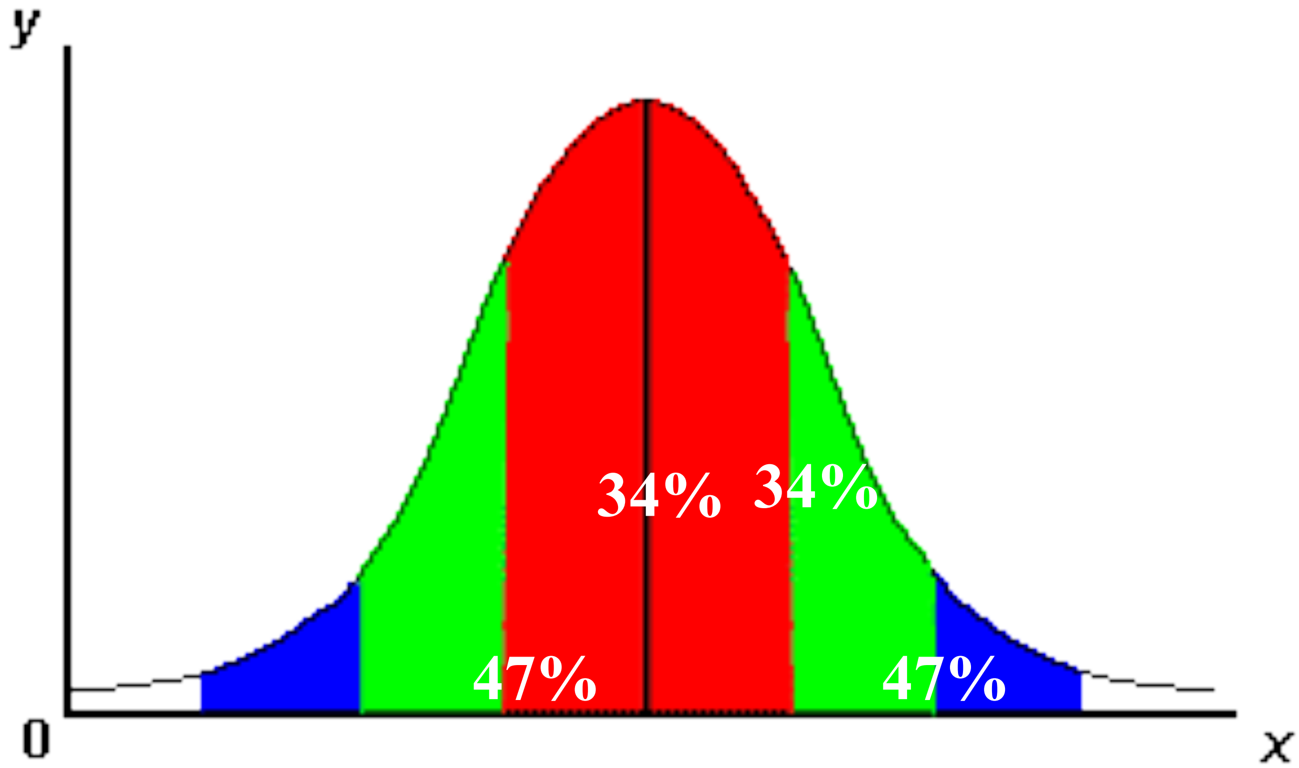
A few words about the normal curve



- Skewness = 0
- Kurtosis = 3

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$$

Standard Deviation in normal curve



Accuracy

- Accuracy:
  - the closeness of the measurements to the “actual” or “real” value of the physical quantity.

- Statistically this is estimated using the standard error of the mean

Standard error of the mean

- The mean of a sample is an estimate of the true (population) mean.

$$\boxed{\bar{x} \approx \mu}$$

- The extent to which this estimate differs from the true mean is given by the

standard error of the mean

$$\boxed{SE(\bar{x}) = \frac{s}{\sqrt{N}}}$$

s = standard deviation of the sample mean and
describes the extent to which any single measurement is liable to differ from the mean

- The standard error depends on the standard deviation and the number of measurements

$$\boxed{\frac{1}{\sqrt{N}}}$$

Often it is not possible to reduce the standard deviation significantly (which is limited instrument precision) so repeated measurements (high N) may improve the resolution.

Precision

- *Precision:* is used to indicate the closeness with which the measurements agree with one another
  - Statistically the precision is estimated by the standard deviation of the mean
- The assessment of the possible error in any measured quantity is of fundamental importance in science
  - Precision is related to random errors that can be dealt with using statistics
  - Accuracy is related to systematic errors and are difficult to deal with using statistics

Weighted Average Error

- A set of measurements of the same quantity, each given with a known error

$$\begin{array}{l} x_1 \pm s_1 \\ x_2 \pm s_2 \\ x_3 \pm s_3 \\ x_4 \pm s_4 \\ \ldots\ldots\ldots \end{array}$$

- The mean value is calculated by "weighting" each of the measurements (x-values) according to its error.

$$\bar{x}_{tot} = \frac{\sum x_i / s_i^2}{\sum 1 / s_i^2}$$

with a standard deviation given by

$$s_{tot} = \sqrt{\frac{1}{\sum 1 / s_i^2}}$$

IQV—Index of Qualitative Variation

- For nominal variables
- Statistic for determining the dispersion of cases across categories of a variable.
- Ranges from 0 (no dispersion or variety) to 1 (maximum dispersion or variety)
- 1 refers to even numbers of cases in all categories, NOT that cases are distributed like population proportions
- IQV is affected by the number of categories

IQV—Index of Qualitative Variation

To calculate:

$$IQV = \frac{K(100^2 - \Sigma \text{ cat.}\%^2)}{100^2(K - 1)}$$

K=# of categories

Cat.% = percentage in each category

IQV—Index of Qualitative Variation

Problem: Is SJSU more diverse than UC Berkeley?

Solution: Calculate IQV for each campus to determine which is higher.

<u>SJSU:</u>

Percent	Category
00.6	Native American
06.1	Black
39.3	Asian/PI
19.5	Latino
34.5	White

<u>UC Berkeley:</u>

Percent	Category
00.6	Native American
03.9	Black
47.0	Asian/PI
13.0	Latino
35.5	White

What can we say before calculating? Which campus is more evenly distributed?

IQV—Index of Qualitative Variation

Problem: Is SJSU more diverse than UC Berkeley? **YES**

Solution: Calculate IQV for each campus to determine which is higher.

SJSU:

Percent	Category	%^{2}
00.6	Native American	0.36
06.1	Black	37.21
39.3	Asian/PI	1544.49
19.5	Latino	380.25
34.5	White	1190.25

UC Berkeley:

Percent	Category	%^{2}
00.6	Native American	0.36
03.9	Black	15.21
47.0	Asian/PI	2209.00
13.0	Latino	169.00
35.5	White	1260.25

K = 5 Σ cat.%^{2} = 3152.56

k = 5 Σ cat.%^{2} = 3653.82 $100^2 = 10000$

$$IQV = \frac{K\,(100^2 - \Sigma \text{ cat.}\%^2)}{100^2(K-1)}$$

5(10000 – 3152.56) = 34237.2

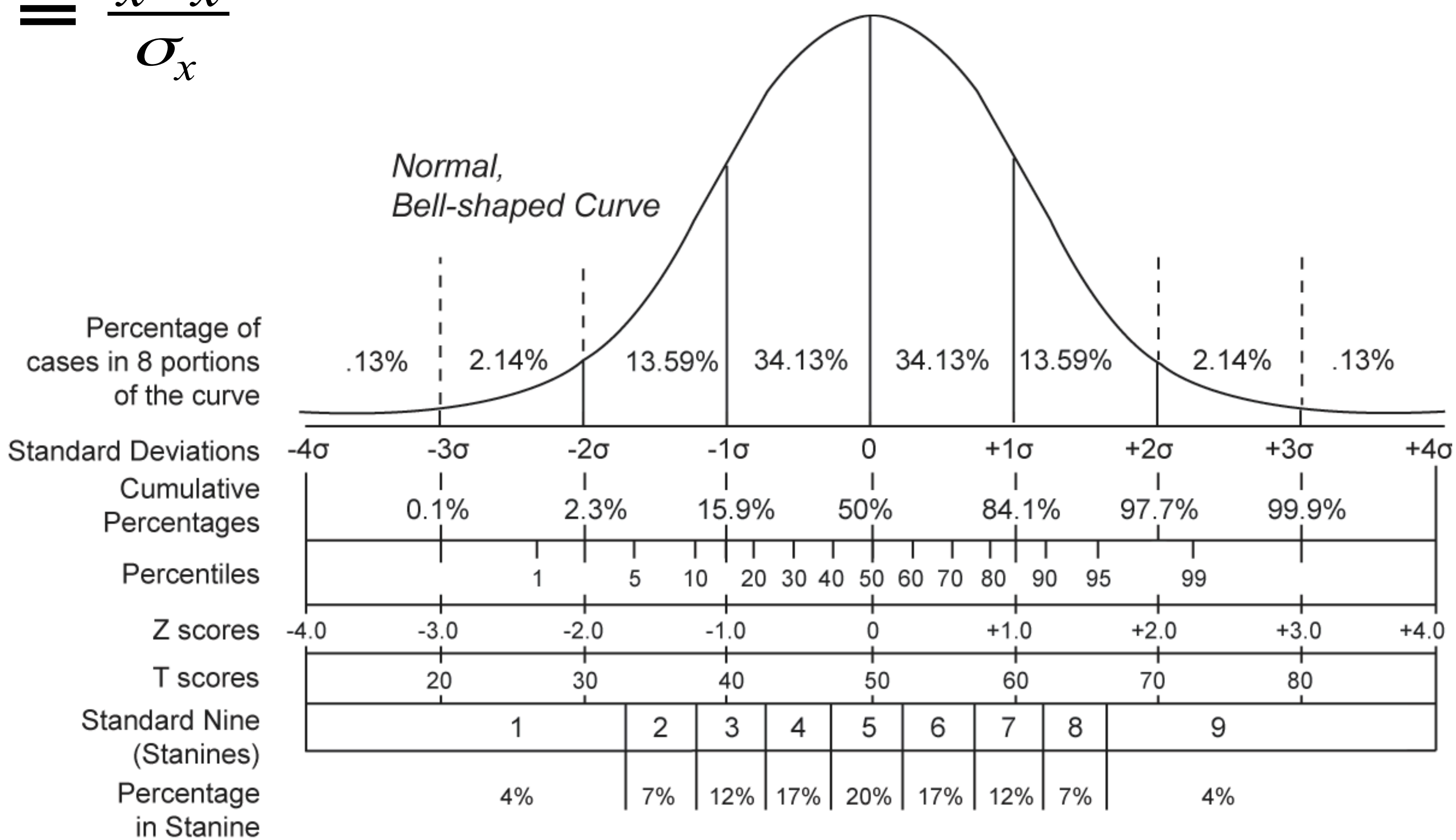
10000(5 – 1) = 40000 **SJSU IQV = .856**

5(10000 – 3653.82) = 31730.9

10000(5 – 1) = 40000 **UCB IQV =.793**

The z-score: "standardized score"

$$z = \frac{x - \bar{x}}{\sigma_x}$$



Comparing various grading methods in a normal distribution

Correlation

How can we quantify the strength and direction of a *linear* relationship between *X* and *Y* variables?

Pearson's Coefficient

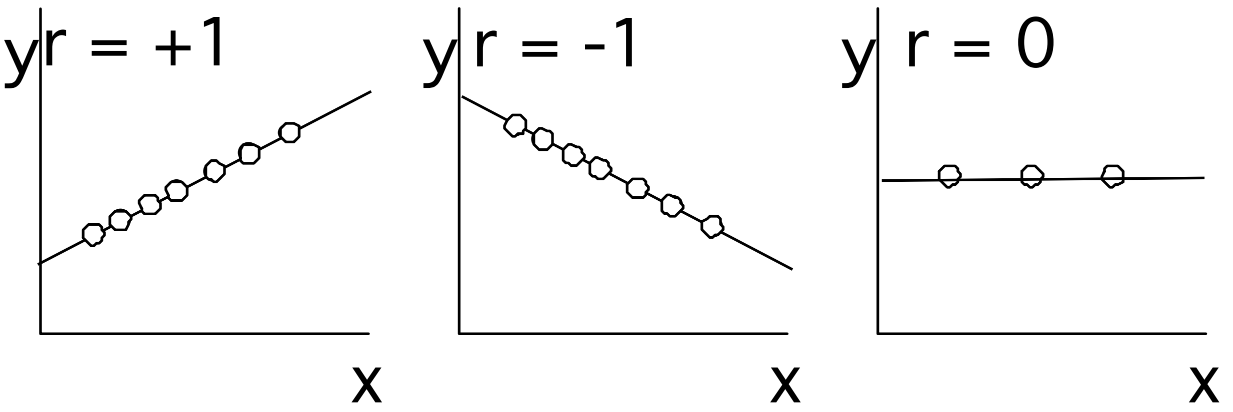
$$r = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{N}\right] \times \left[\sum y^2 - \frac{(\sum y)^2}{N}\right]}}$$

- Σ y = sum of all y-values
- Σ x = sum of all x-values
- Σ x^2 = sum of all x^2 values
- Σ y^2 = sum of all y^2 values
- Σ xy = sum of the x times y values

Like other numerical measures, the population correlation coefficient is   (the Greek letter ``rho'', ρ) and the sample correlation coefficient is denoted by *r*.

Correlation

- Values of r



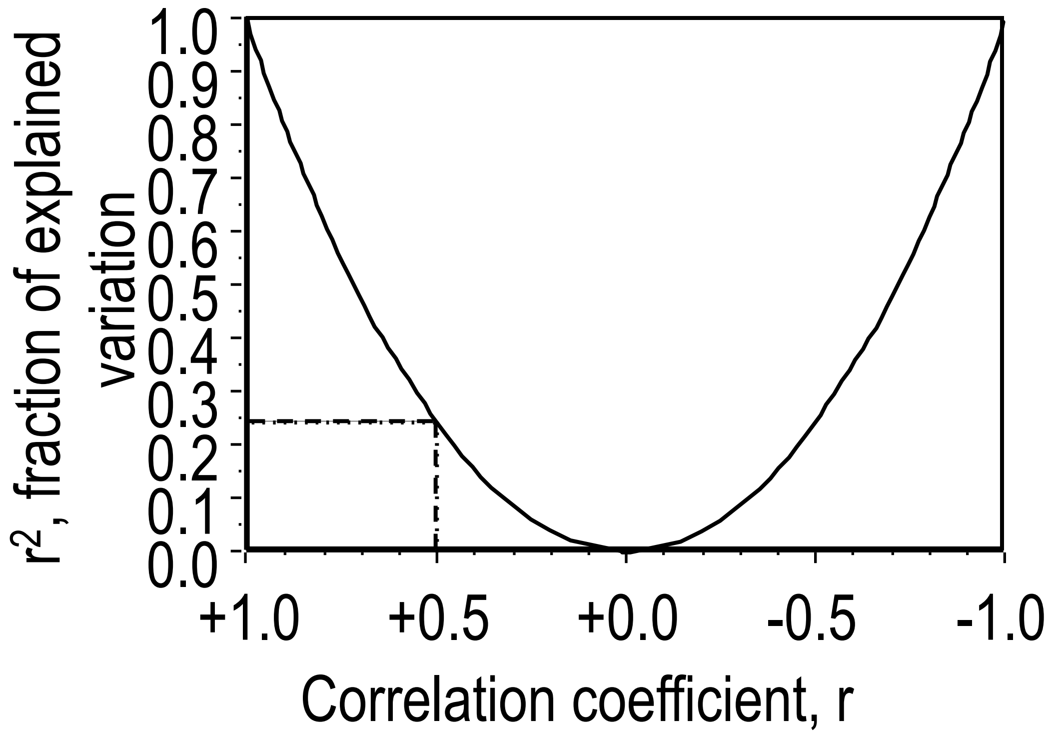
Perfect positive correlation

Perfect negative correlation

No correlation

Correlation

- r^2 is the amount of variation in x and y that is explained by the linear relationship. It is often called the `goodness of fit'
- E.g. if an r = 0.97 is obtained then r^2 = 0.95
  - so 100x0.95=95% of the total variation in x and y is explained by the linear relationship,
  - but the remaining 5% variation is due to “other” causes.



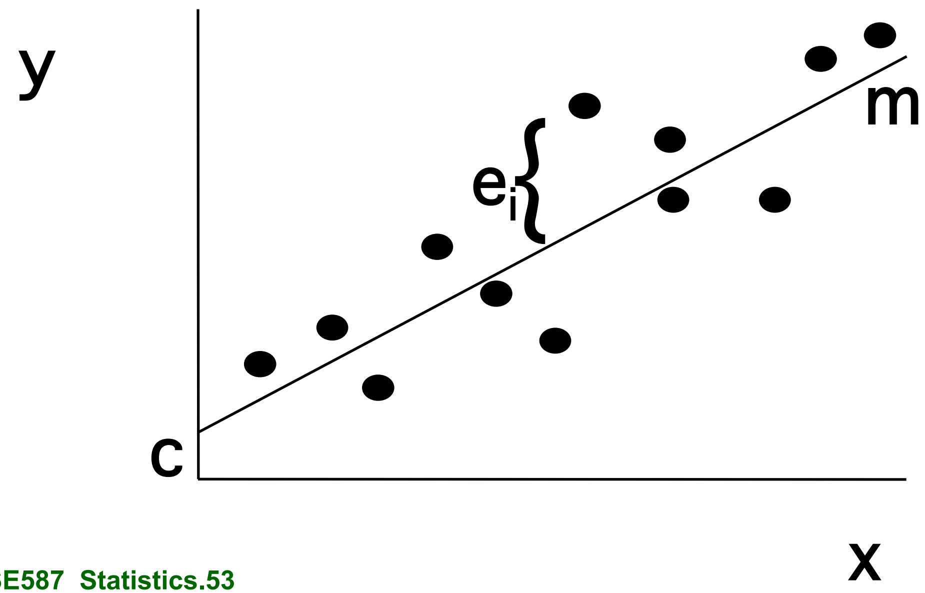
Regression analysis

How can we fit an equation to a set of numerical data x, y such that it yields the best fit for all the data?

Linear Regression: An approximate fit yields a straight line that passes through the set of points in the *best possible manner* without being required to pass exactly through any of the points.

Linear Regression

$y=mx+c$



- Where e_i is the deviation of the data point from the fit line, c is the intercept, m is the gradient.
- Assumes that the error is present only in y.

How do we define a good fit?

- If the sum of all deviations is a minimum? Σe_i
- If the sum of all the absolute deviations is a minimum? $\Sigma |e_i|$
- If the maximum deviation is a minimum? e_{max}
- If the sum of all the squares of the deviations is a minimum? Σe_i^2

Classical linear regression

- The best way is to minimise the sum of the squares of the deviation. Formally this involves some Mathematics:
- At each value of x_i:

$$y_i = mx_i + c$$

- Therefore the deviations from the curve are:

$$e_i = (Y_i - y_i)$$

- The sum of the squares:

$$S(c,m) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - c - mx_i)^2$$

Classical linear regression

- How do you find the minimum of a function?

- Use calculus
- Differentiate and set to zero

$$\frac{\partial S(c,m)}{\partial c} = \sum_{i=1}^{N} 2(Y_i - c - mx_i)(-1) = 0$$

$$\frac{\partial S(c,m)}{\partial m} = \sum_{i=1}^{N} 2(Y_i - c - mx_i)(-x_i) = 0$$

- Two simultaneous equations

$$cN + m\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} Y_i$$

$$c\sum_{i=1}^{N} x_i + m\sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} x_i Y_i$$

Classical linear regression

- Solving the two equations yields:

$$c = \frac{\sum_{i=1}^{N} Y_i \left(\sum_{i=1}^{N} x_i\right)^2 - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i Y_i}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}$$

$$m = \frac{N \sum_{i=1}^{N} x_i Y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} Y_i}{N \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}$$

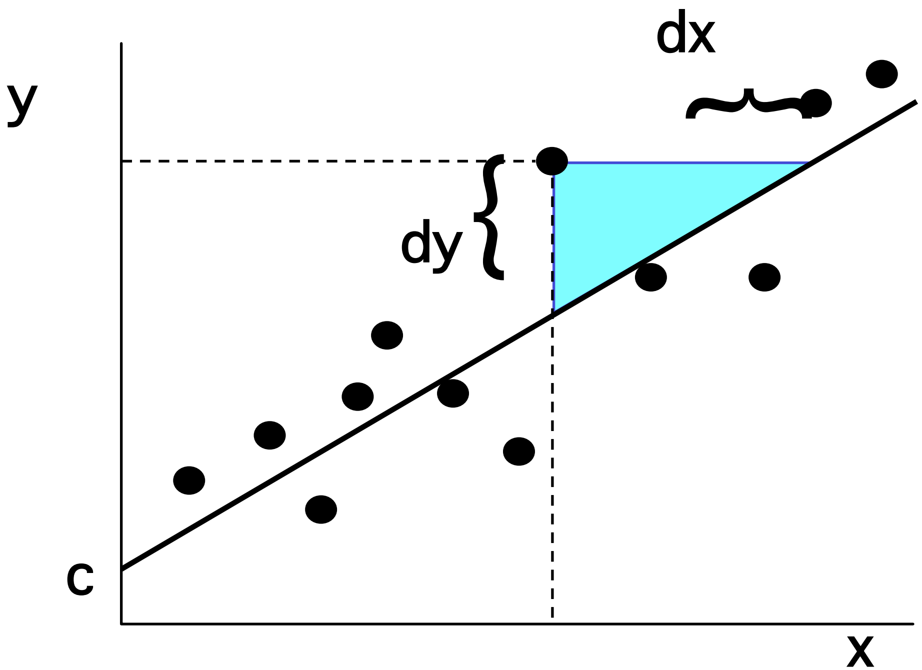
Classical linear regression

- Classical linear regression only considered errors in the Y values of the data.

- How can we consider errors in both x and y values?

- Use Reduced major axis regression

Reduced major axis regression



$$m = \frac{\sigma_y}{\sigma_x} = \sqrt{\frac{\sum y^2 - \frac{\left(\sum y\right)^2}{N}}{\sum x^2 - \frac{\left(\sum x\right)^2}{N}}}$$

$$c = \bar{y} - m\bar{x}$$

- Method to quantify a linear relationship where both variables are dependent and have errors
- Instead of minimising $e^2=(Y-y)^2$ we minimize $e^2=dy^2+dx^2$.

Error propagation

- Every measurement of a variable has an error.
- Often the error quoted is one standard deviation of the mean (mean ± standard deviation)
- The standard deviation of the sample mean is usually our best estimate of the population standard deviation

Error propagation

- Error propagation is a way of combining two or more **random** errors together to get a third. The equations assume that the errors are Gaussian in nature.
- It can be used when you need to measure more than one quantity to get at your final result.
- How then do we combine variables which have errors?

Error propagation - quoted

Relationship	Error propagation
$z = x + y$	$(\sigma_z)^2 = (\sigma_x)^2 + (\sigma_y)^2$
$z = x - y$	$(\sigma_z)^2 = (\sigma_x)^2 + (\sigma_y)^2$
$z = xy$	$\left(\frac{\sigma_z}{z}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2$
$z = \frac{x}{y}$	$\left(\frac{\sigma_z}{z}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2$
$z = kx$	$\sigma_z = k\sigma_x$ (k=constant)
$z = x^n$	$\frac{\sigma_z}{z} = n\frac{\sigma_x}{x}$
$z = \log_e x$	$\sigma_z = \frac{\sigma_x}{x}$
$z = e^x$	$\frac{\sigma_z}{z} = \sigma_x$

Example of propagation of error

- Suppose we measure the thickness of a rock bed using a tape measure.
- The tape measure is shorter then the bed thickness so we have to do it in two steps x and y.
- We repeat the measurements 100 times and obtain the following mean and standard deviation values for x and y:

$$\boxed{\begin{gathered} x=12.1\pm0.3 \text{ cm} \\ y=4.2\pm0.2 \text{ cm} \end{gathered}}$$

- The thickness of the bed should be simply:

$$\boxed{x+y=16.3 \text{ cm}}$$

- But what about the error on the total thickness?

Example of propagation of error

- It is given by propagating the individual errors as follows:

$$Relationship: z = x + y$$

$$Error\ propagation: \sigma_z^2 = \sigma_x^2 + \sigma_y^2$$

$$\sigma_z^2 = 0.3^2 + 0.2^2 = 0.13$$

$$\sigma_z^2 = \sqrt{0.13} = 0.36$$

- So the final answer for the total thickness of the bed is:

$$\boxed{16.3 \pm 0.4 \text{ cm}}$$

- Error propagation formulae are non-intuitive and understanding how they are derived requires some mathematical knowledge

More complex examples

- What if we have several functions of several variables?
- E.g. calculating density using Archimedes Principle:

$$Density = \frac{wt.\,in\,air\,(A)}{wt.\,in\,air(A)\text{ - }wt\,in\,water(W)}$$

- This equation contains two functions and two variables

- Error propagation is best done in parts, so first work out value and error in denominator:

$$x = A - W$$

- Then the value and error of:

$$Density = \frac{A}{x}$$