

Understanding and serving users

Class 12 - Data analysis

Statistics, probability, distributions
and inferential test choice

Basic types of statistical treatment

- **Descriptive** statistics which summarize the characteristics of a *sample* of data
- **Inferential** statistics which attempt to say something about a *population* on the basis of a *sample* of data - infer to *all* on the basis of *some*

Two kinds of descriptive statistic:

- Measures of central tendency
 - mean
 - median
 - mode
- Measures of dispersion (variation)
 - range
 - interquartile range
 - variance/standard deviation

Symbol check

Σ

- Sigma: Means the 'sum of'

$\sum_{i=1}^n x_i$

- Sigma (1 to n) x of i: means add all values of i from 1 to n in a data set
- X_i = the i^{th} data point

Mean (arithmetic mean, or average)

Sum of all observations divided by the number of observations

In notation:

$$\frac{\sum_{i=1}^n x_i}{n}$$

Mean uses every item of data but is sensitive to extreme 'outliers'

Suppose we ask 10 people to perform a word processing task and measure the time in minutes which it takes them to complete it. The raw data are as follows:

2 4 3 2 5 2 3 2 4 2

To find the mean, we sum the observations and divide by the number of observations:

$$\frac{\sum_{i=1}^n x_i}{n} = \frac{2 + 4 + 3 + 2 + 5 + 2 + 3 + 2 + 4 + 2}{10} = 29 / 10 = 2.9$$

Mean (cont.)

- Now suppose the person who took 5 minutes had *real* problems and took 60 minutes.
- The mean now becomes:

$$\frac{2 + 4 + 3 + 2 + 60 + 2 + 3 + 2 + 4 + 2}{10} = 84 / 10 = 8.4$$

Mean is sensitive to *every* data point in the set₇

Median (the 50th percentile point)

To determine the median, put the observations in size order:

2 2 2 2 2 3 3 4 4 5
 ↑

Half-way point

Hence, median = 2.5

Same is true even if last observation is 60 rather than 5
Thus even if mean shifts from 2.9 to 8.4 (as seen),
median remains 2.5.

When to use the Median?

Median is appropriate when

- Data are ordinal
- Data contain extreme observations which will distort the mean as a measure of central tendency
e.g., testers' salaries:

\$45k, \$50k, \$55k, \$57k, \$60k, \$65k, \$150k

Median = \$57k

Mean = \$69k

Mode -the most frequent observation

2 2 2 2 2 3 3 4 4 5

<u>N°</u>	<u>Frequency</u>
2	5
3	2
4	2
5	1

Mode = 2

The mode is the only permissible measure of central tendency for nominal data

Measures of dispersion

- Range - the difference between least and largest scores:

<u>Sample</u>	<u>Scores</u>	<u>Mean</u>
I	6,4,7,5	5.5
II	7,10,1,4	5.5

Range for : Sample I: $7 - 4 = 3$
 Sample II: $10 - 1 = 9$

Interquartile deviation

- Range is sensitive to outliers:

2 2 2 2 2 3 3 4 4 5 6 60

$$\text{Range} = 60 - 2 = 58$$

25th percentile (Q1)



75th percentile (Q3)



2 2 2 2 2 3 3 4 4 5 6 60

- Interquartile range = $Q3 - Q1 = 4.5 - 2 = 2.5$

When to use Interquartile deviation

- Useful where extremes render range un-descriptive
- Less sensitive to outliers, tells us more about the *internal spread* of data

How do we handle outliers?

- Usually worth doing two analyses:
 - One with the outlier
 - One without the outlier
- Outlier might be caused by test design, some odd user behavior *or* might be a true reflection of some usability issue - so you need to investigate this further.....

Variance and standard deviation

- A deviation is a measure of how far from the mean is a score in our data
 - Sample: 6,4,7,5 mean =5.5
 - Each score can be expressed in terms of distance from 5.5
 - 6,4,7,5, => 0.5, -1.5, 1.5, -0.5 (these are distances from mean)
 - Since these are measures of distance, some are positive (greater than mean) and some are negative (less than the mean)
 - TIP: Sum of these distances ALWAYS = 0

Symbol check

$$\bar{x}$$

- Called 'x bar'; refers to the 'mean'

$$(x - \bar{x})$$

- Called 'x minus x-bar'; implies subtracting the mean from a data point x. also known as a deviation from the mean

Variance

The average squared deviation from the mean

$$\text{var} = \frac{\sum (x - \bar{x})^2}{n}$$

Standard Deviation

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

Computational formula

$$s = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$$

Two ways to get SD

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

- Sum the sq. deviations from the mean
- Divide by No. of observations
- Take the square root of the result

$$sd = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}$$

- Sum the squared raw scores
- Divide by N
- Subtract the squared mean
- Take the square root of the result

Task time:

x	x^2
2	4
2	4
2	4
2	4
2	4
3	9
3	9
4	16
4	16
5	25
$\Sigma x = 29$	$\Sigma x^2 = 95$

$$\begin{aligned} s &= \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2} \\ &= \sqrt{\frac{95}{10} - 2.9^2} \\ &= \sqrt{9.5 - 8.41} \\ &= \sqrt{1.09} \\ &= 1.044 \end{aligned}$$

Exercise 2

- Calculate the s.d. using the 'other' method

▪ 2 2 2 2 2 3 3 4 4 5

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

Exercise 2

Step:	1	2	3	4	5
X (scores)	Subtract each X from mean	Square the result	Sum the squares	Divide by N	Take square root
2	-0.9	0.81	10.9	10.9/10 = 1.09	$\sqrt{1.09}$ = 1.04
2	-0.9	0.81			
2	-0.9	0.81			
2	-0.9	0.81			
2	-0.9	0.81			
3	0.1	0.01			
3	0.1	0.01			
4	1.1	1.21			
4	1.1	1.21			
5	2.1	4.41			

Both SD formulae yield the same answer....1.04

If we include the outlier:

x	x^2
2	4
2	4
2	4
2	4
2	4
3	9
3	9
4	16
4	16
60	3600
$\Sigma x = 84$	$\Sigma x^2 = 3670$

$$s = \sqrt{\frac{\Sigma x^2}{n} - \bar{x}^2}$$

$$= \sqrt{\frac{3760}{10} - 8.4^2}$$

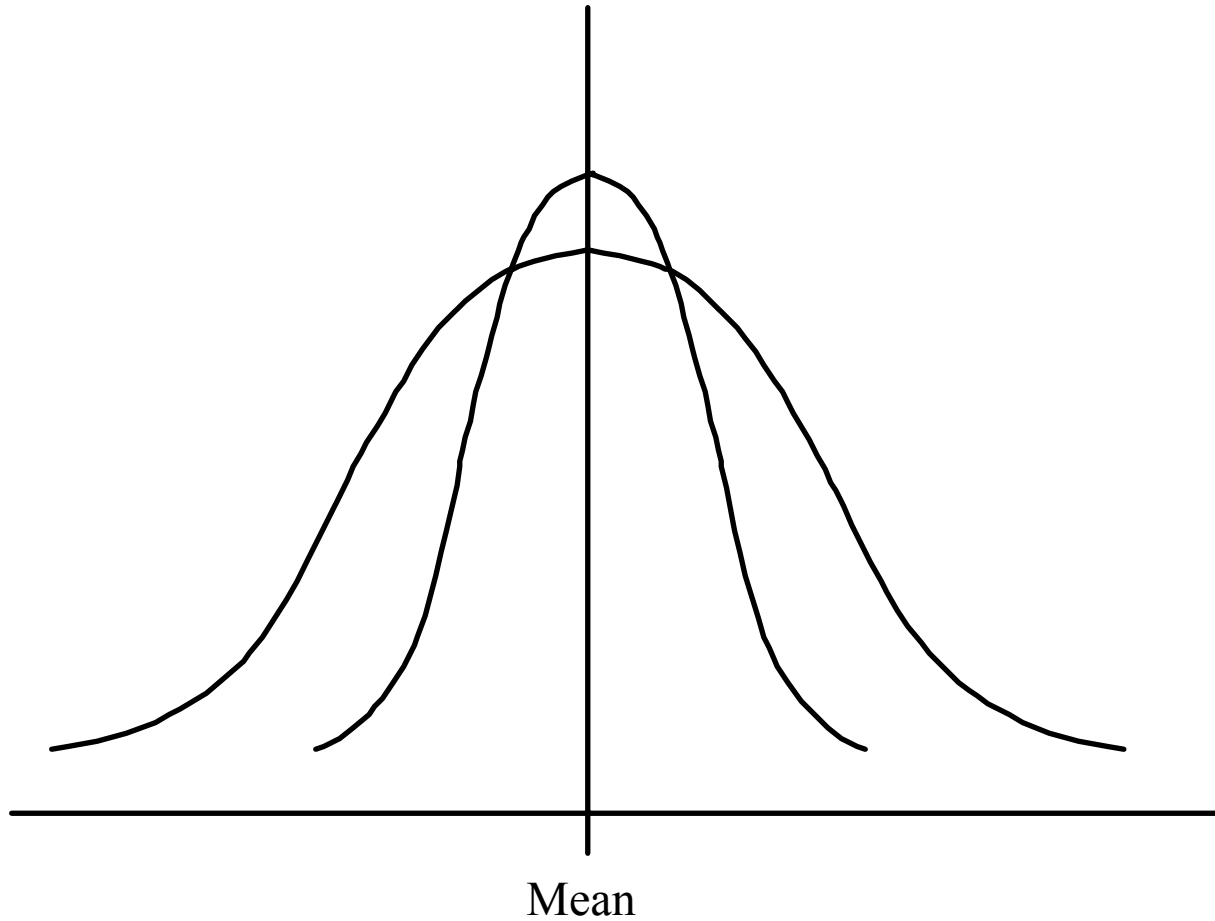
$$= \sqrt{367 - 70.56}$$

$$= \sqrt{296.44}$$

$$= 17.22$$

Note increase in SD





Two sets of data can have the same mean but different standard deviations.

The bigger the SD, the more s-p-r-e-a-d out are the data.

On the use of N or N-1

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

$$sd = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

- When your observations are the complete set of people that could be measured
- When you are observing only a sample of potential users, the use of N-1 increases size of sd slightly

Summary

Measures of Central Tendency

- Mode • Most frequent observation. Use with nominal data
- Median • 'Middle' of data. Use with ordinal data or when data contain outliers
- Mean • 'Average'. Use with interval and ratio data if no outliers

Measures of Dispersion

- Range • Dependent on two extreme values
- Interquartile Range • More useful than range. Often used with median
- Variance / Standard Deviation • Same conditions as mean. With mean, provides excellent summary of data

Deviation units: Z scores

Any data point can be expressed in terms of its
Distance from the mean in SD units:

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

A positive z score implies a value above the mean
A negative z score implies a value below the mean₂₆

Interpreting Z scores

- Mean = 70, SD = 6
 - Then a score of 82 is 2 sd [$(82-70)/6$] above the mean, or 82 = Z score of 2
 - Similarly, a score of 64 = a Z score of -1
- By using Z scores, we can standardize a set of scores to a scale that is more intuitive
 - Many IQ tests and aptitude tests do this, setting a mean of 100 and an SD of 10 etc.

Comparing data with Z scores

You score 49 in class A but 58 in class B
How can you compare your performance in both?

Class A:

Mean =45

SD=4

Class B:

Mean =55

SD = 6

49 is a Z=1.0

58 is a Z=0.5

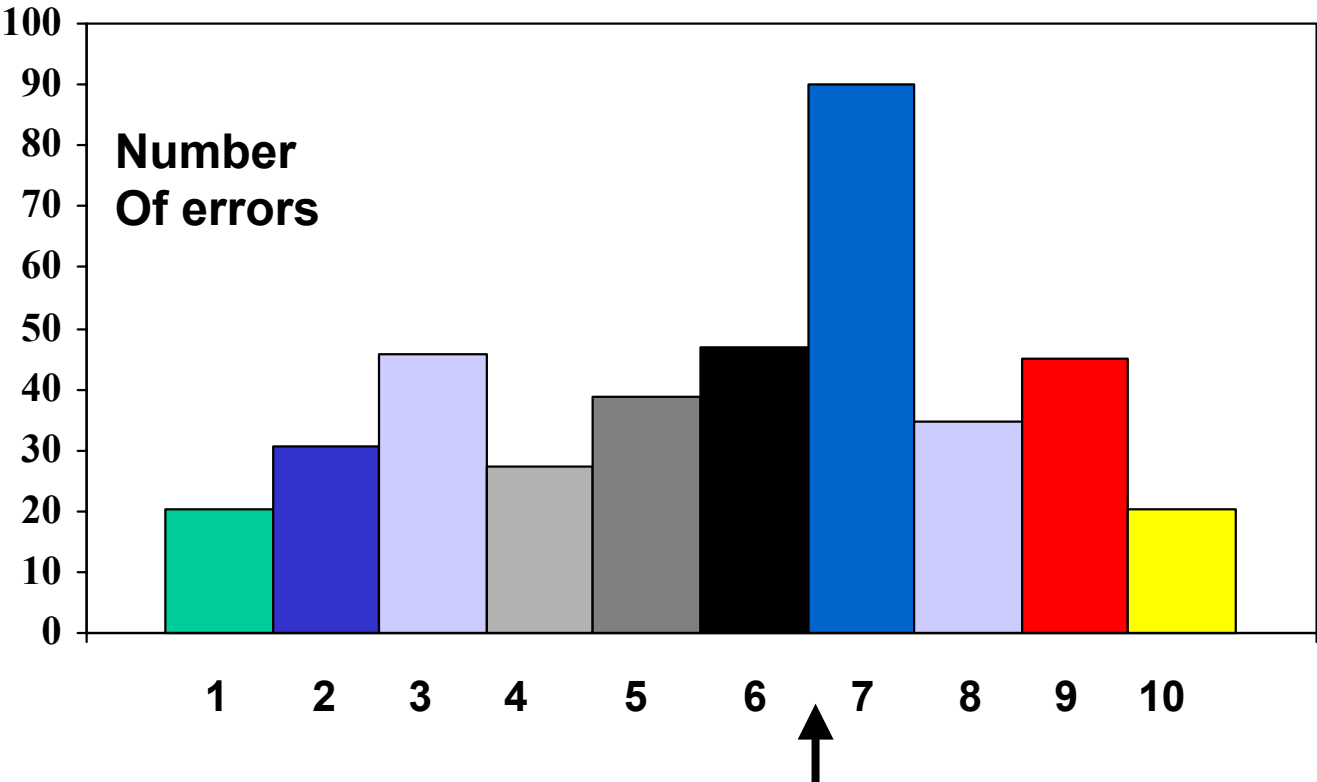
With normal distributions

*Mean,
SD and
Z tables*

In combination provide powerful means of estimating what your data indicates

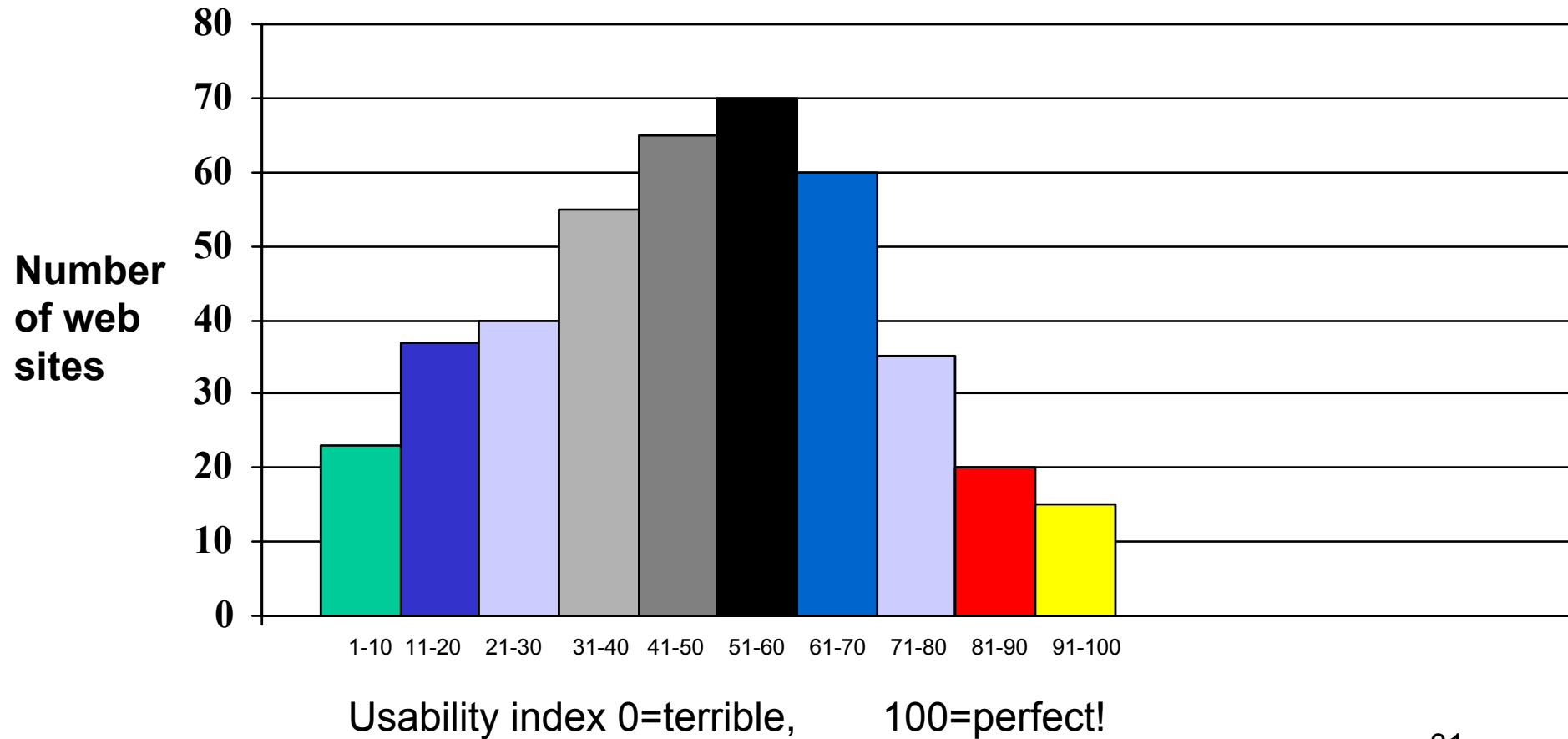
Graphing data - the histogram

The frequency of occurrence for measure of interest, e.g., errors, time, scores on a test etc. →

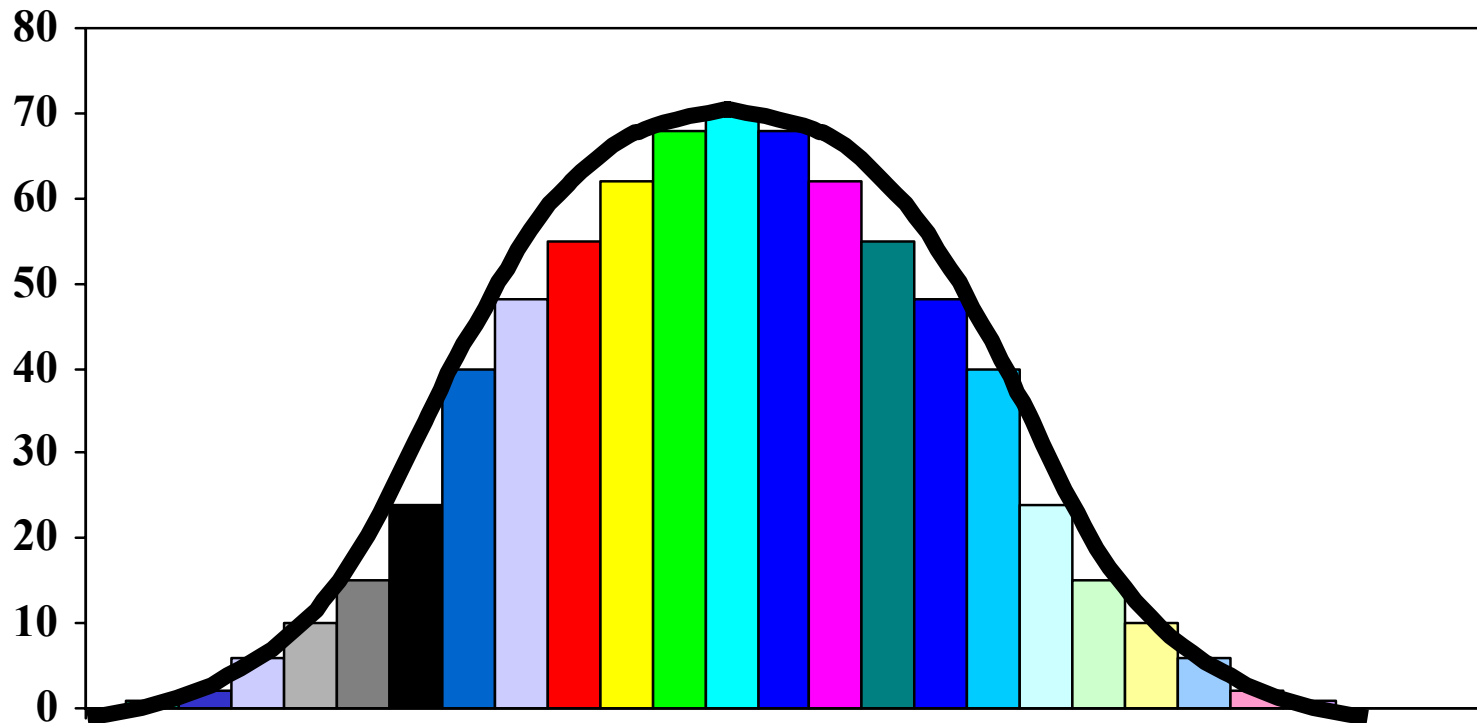


↑
The categories of data we are studying, e.g., task or interface, or user group etc.

Example distribution of data - usability scores of 420 web sites



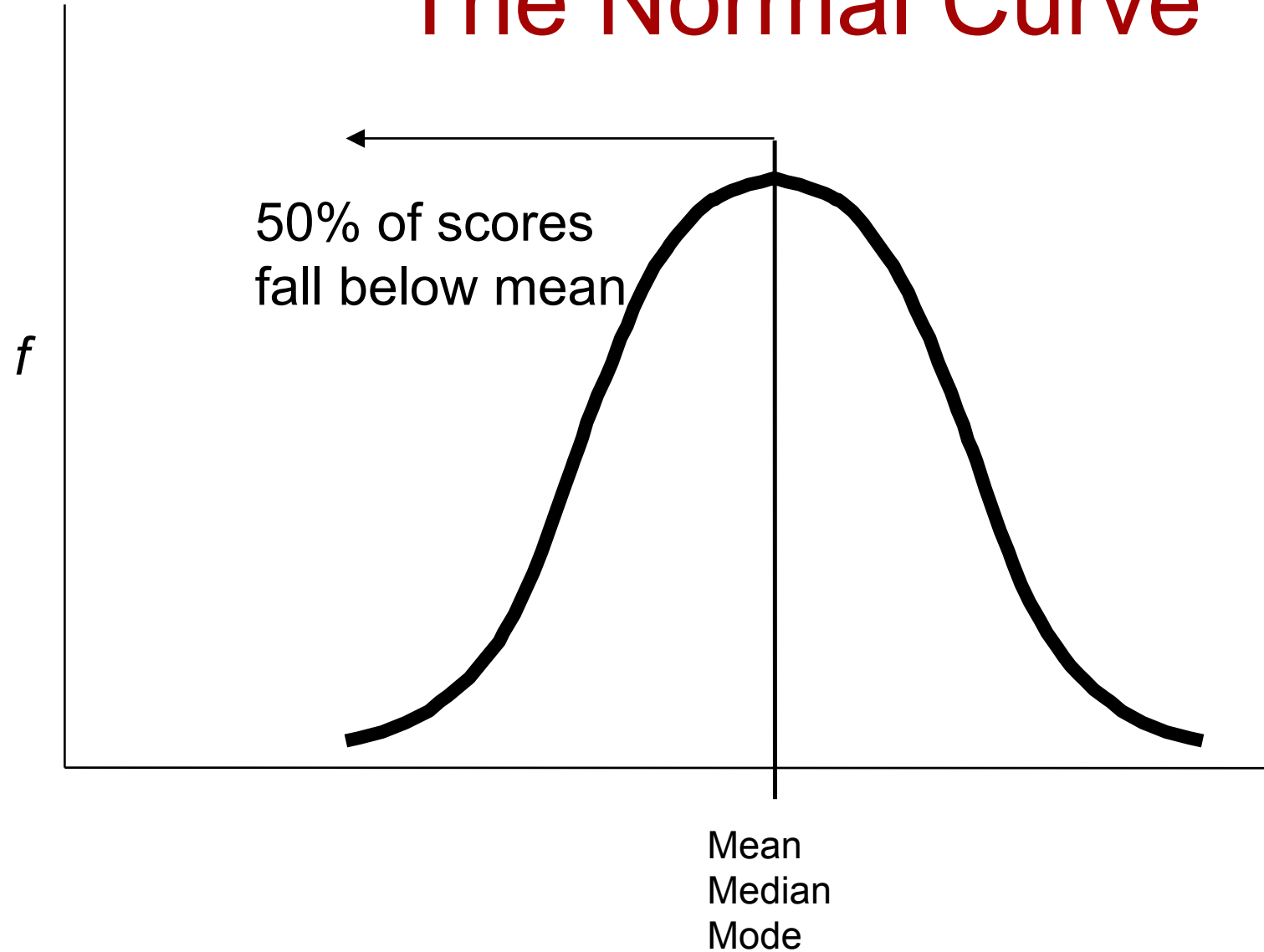
Very large data sets tend to have distinct shape:



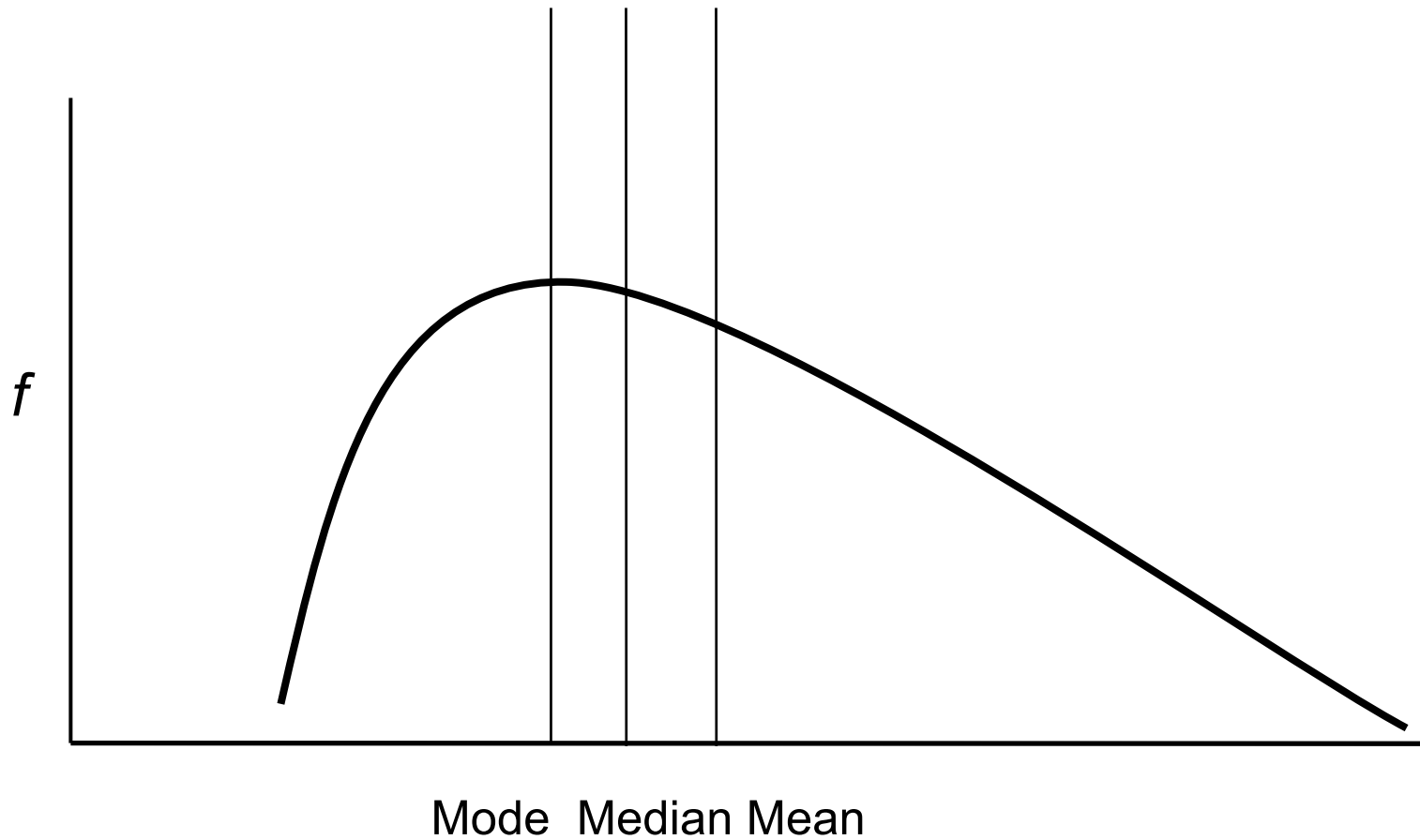
Normal distribution

- Bell shaped, symmetrical, measures of central tendency converge
 - mean, median, mode are equal in normal distribution
 - Mean lies at the peak of the curve
- Many events in nature follow this curve
 - IQ test scores, height, tosses of a fair coin, user performance in tests,

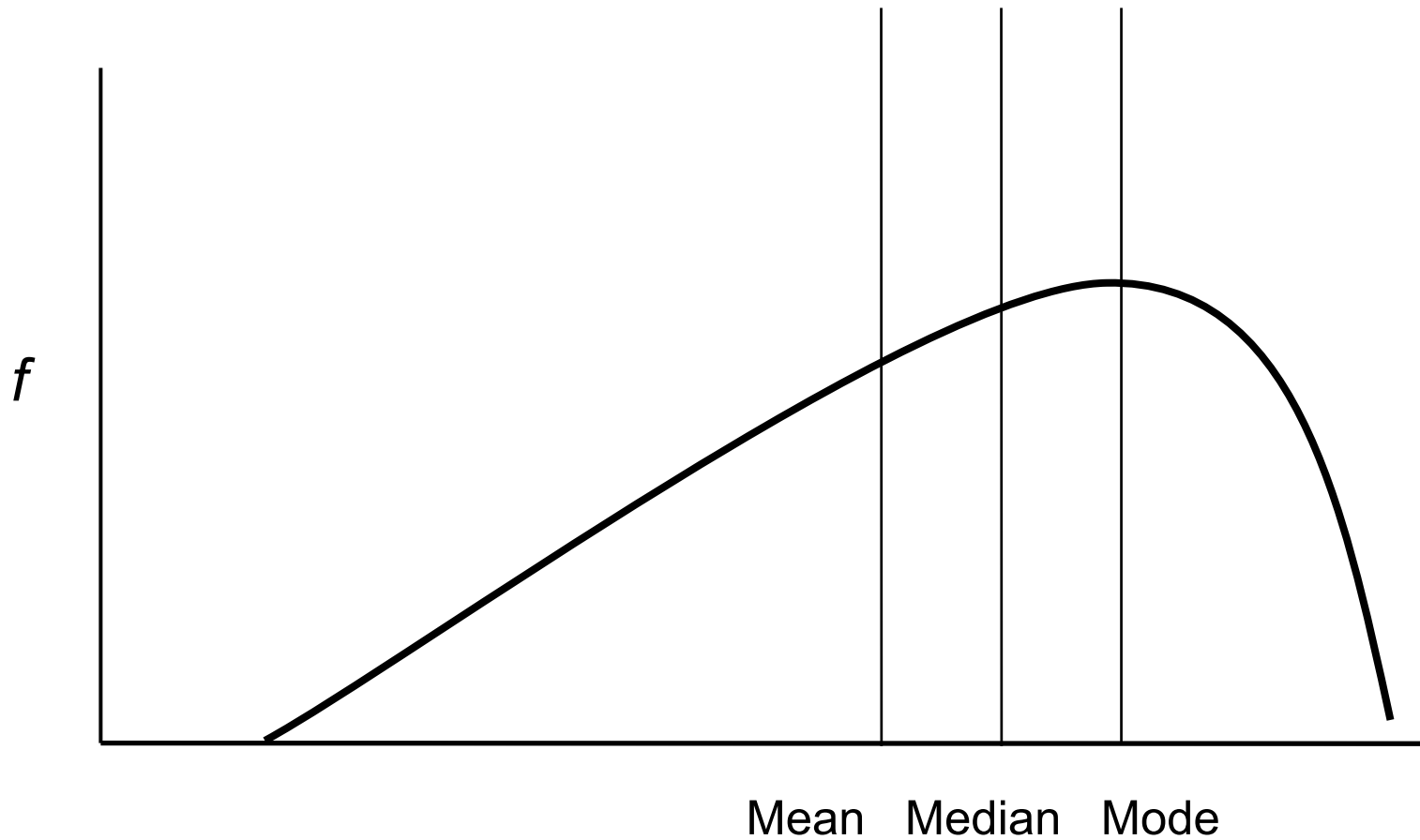
The Normal Curve



Positively skewed distribution



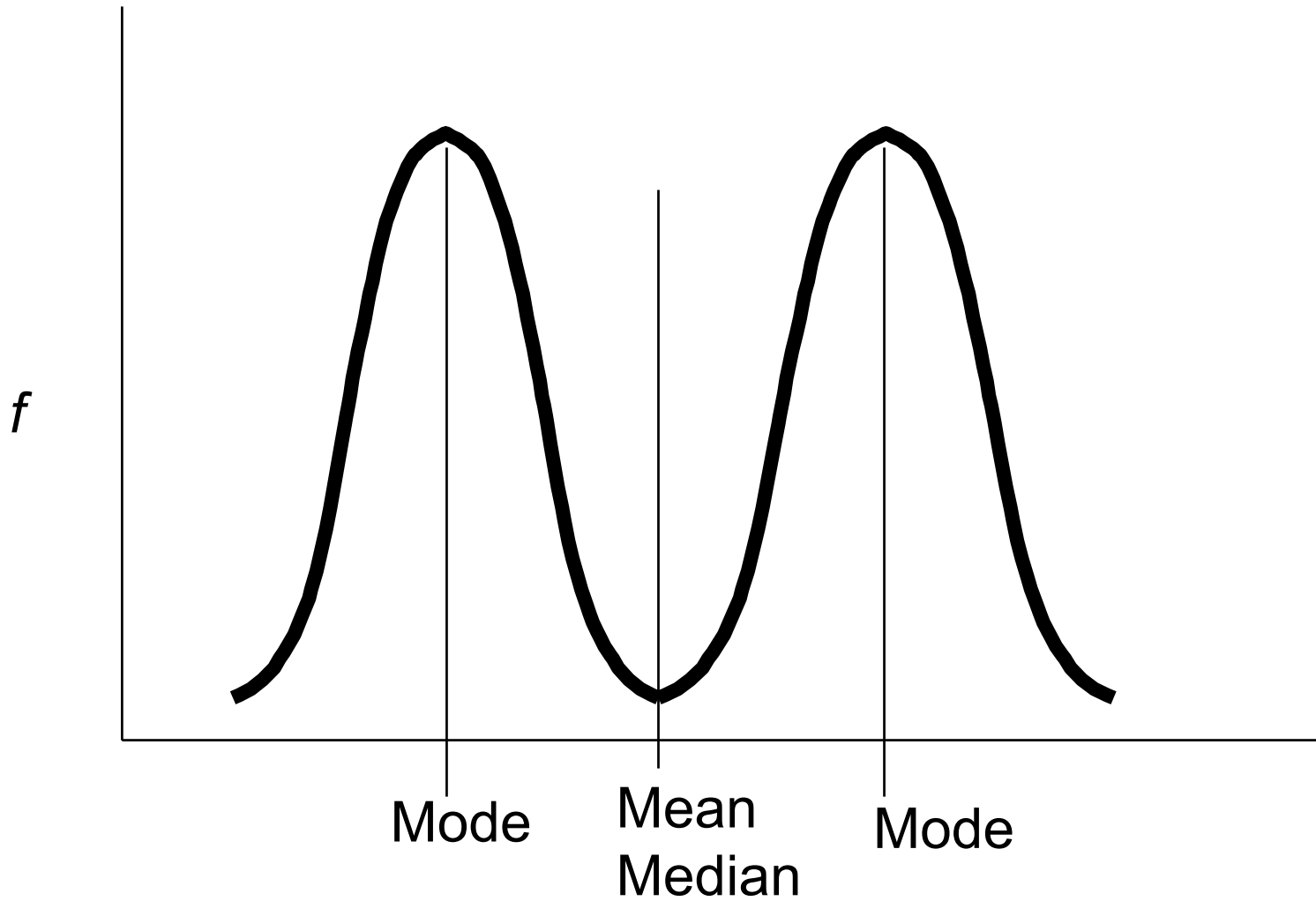
Negatively skewed distribution



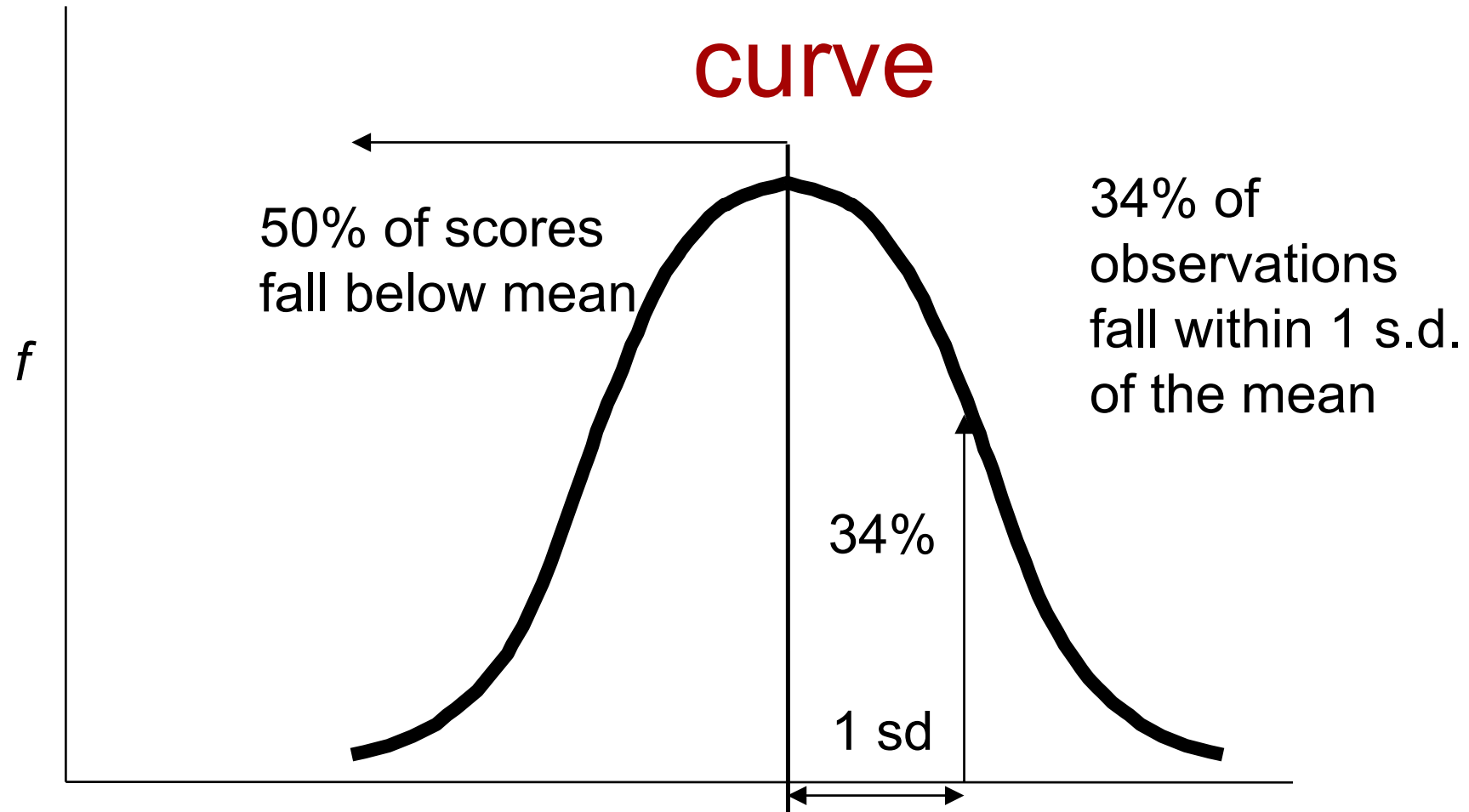
Other distributions

- Bimodal
 - Data shows 2 peaks separated by trough
- Multimodal
 - More than 2 peaks

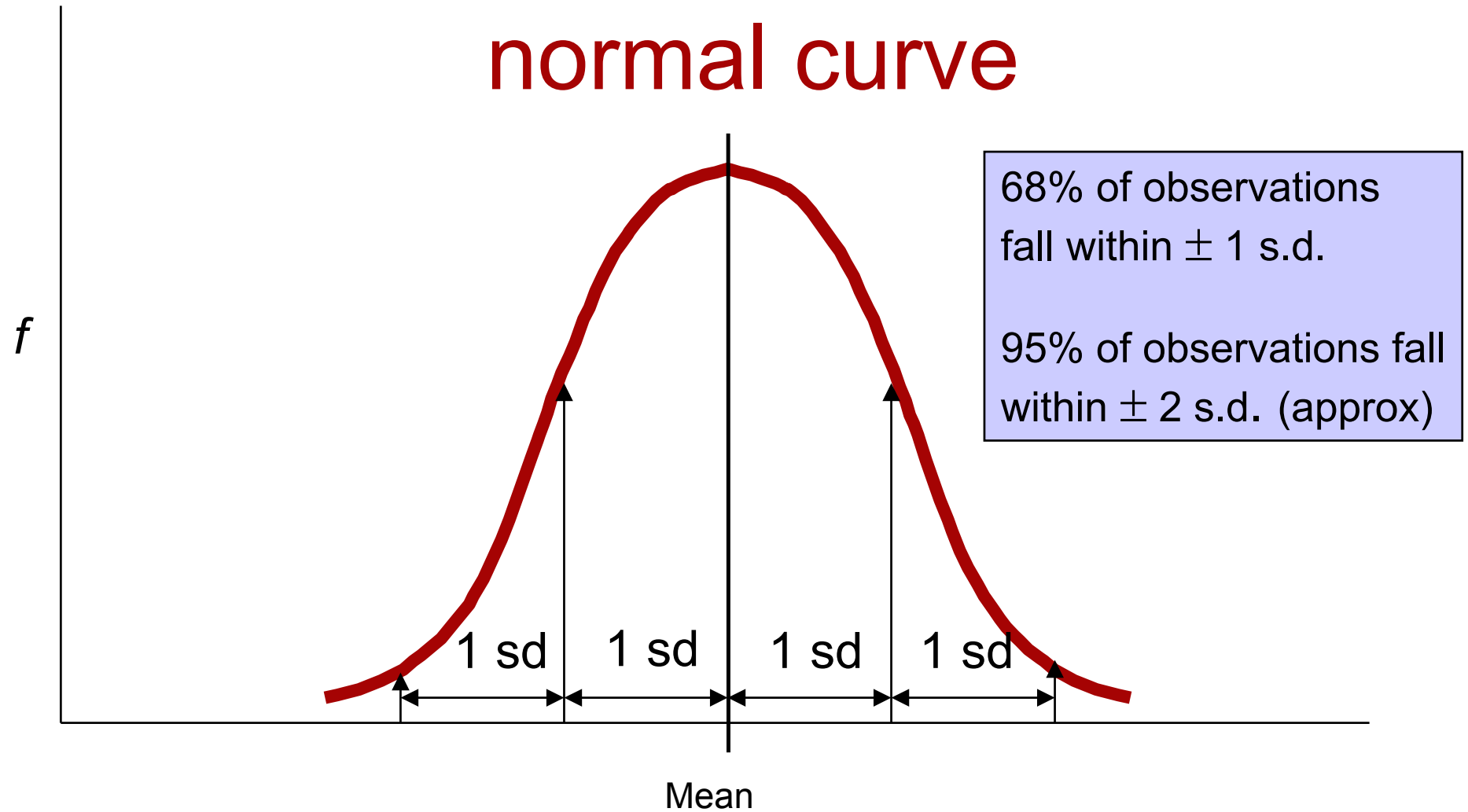
Bimodal



Proportions under the normal curve



Standard deviations and the normal curve



Z scores and tables

Knowing a Z score allows you to determine where under the normal distribution it occurs

Z score between:

0 and 1 = 34% of observations

1 and -1 = 68% of observations etc.

Or 16% of scores are >1 Z score above mean

Check out Z tables in any basic stats book

Remember:

- A Z score reflects position in a normal distribution
- The Normal Distribution has been plotted out such that we know what proportion of the distribution occurs above or below any point

Importance of distribution

- Given the mean, the standard deviation, and some reasonable expectation of normal distribution, we can establish the confidence level of our findings
- With a distribution, we can go beyond descriptive statistics to inferential statistics (tests of significance)

So - for your tests:

- Always summarize the data by graphing it - look for general pattern of distribution
- Then, determine the mean, median, mode and standard deviation
- From these we know a LOT about what we have observed

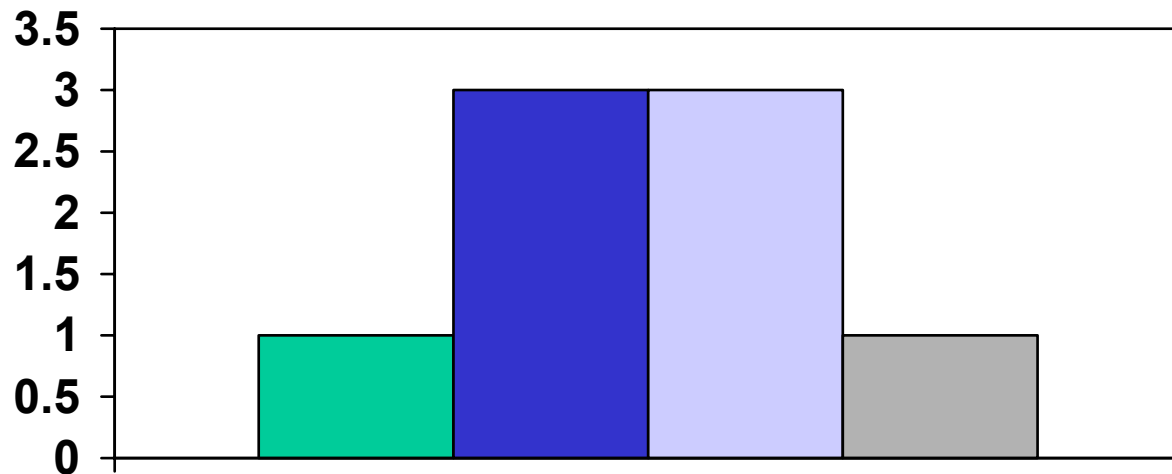
Probability

- Inferential statistics rely on the laws of probability to determine the 'significance' of the data we observe.
- Statistical significance is NOT the same as practical significance
- In statistics, we generally consider 'significant' those differences that occur less than 1:20 by chance alone

Calculating probability

- Probability refers to the likelihood of any given event occurring out of all possible events e.g.:
 - Tossing a coin - outcome is either head or tail
 - Therefore probability of head is $1/2$
 - Probability of two heads on two tosses is $1/4$ since the other possible outcomes are two tails, and two possible sequences of head and tail.
- The probability of any event is expressed as a value between 0 (no chance) and 1 (certain)

Sampling distribution for 3 coin tosses



■ 0 heads	1
■ 1 head	3
■ 2 heads	3
■ 3 heads	1

Probability and normal curves

- Q? When is the probability of getting 10 heads in 10 coin tosses the same as getting 6 heads and 4 tails?
 - HHHHHHHHHH
 - HHTHTHHTHT
- Answer: when you specify the precise order of the 6 H/4T sequence:
 - $(1/2)^{10} = 1/1024$ (specific order)
 - But to get 6 heads, in any order it is: $2^{10}/1024$ (or about 1:5)

What use is probability to us?

- It tells us how likely is any event to occur by chance
- This enables us to determine if the behavior of our users in a test is just chance or is being affected by our interfaces

Determining probability

- Your statistical test result is plotted against the distribution of all scores on such a test.
- It can be looked up in stats tables or is calculated for you in EXCEL or SPSS etc
- This tells you its probability of occurrence
- The distributions have been determined by statisticians.

What is a significance level?

- In research, we estimate the probability level of finding what we found by chance alone.
- Convention dictates that this level is 1:20 or a probability of .05, usually expressed as : $p < .05$.
- However, this level is negotiable

What levels might we chose?

- In research there are two types of errors we can make when considering probability:
 - Claiming a significant difference when there is none (type 1 error)
 - Failing to claim a difference where there is one (type 2 error)
- The $p < .05$ convention is based on addressing type 1 concerns

Using other levels

- Type 1 and 2 errors are interwoven, if we lessen the probability of one occurring, we increase the chance of the other.
- If we think that we really want to find any differences that exist, we might accept a probability level of .10 or higher

Thinking about p levels

- The $p < .x$ level means we believe our results could occur by chance alone (not because of our manipulation) at least $x/100$ times
 - $P < .10 \Rightarrow$ our results should occur by chance 1 in 10 times
 - $P < .20 \Rightarrow$ our results should occur by chance 2 in 10 times
- WHICH also means, our results are NOT chance 9 or 8 out of 10 times!

Putting probability to work

- Understanding the probability of gaining the data you have can guide your decisions
- Determine how precise you need to be IN ADVANCE, not after you see the result
- It is like making a bet....you cannot play the odds after the event!

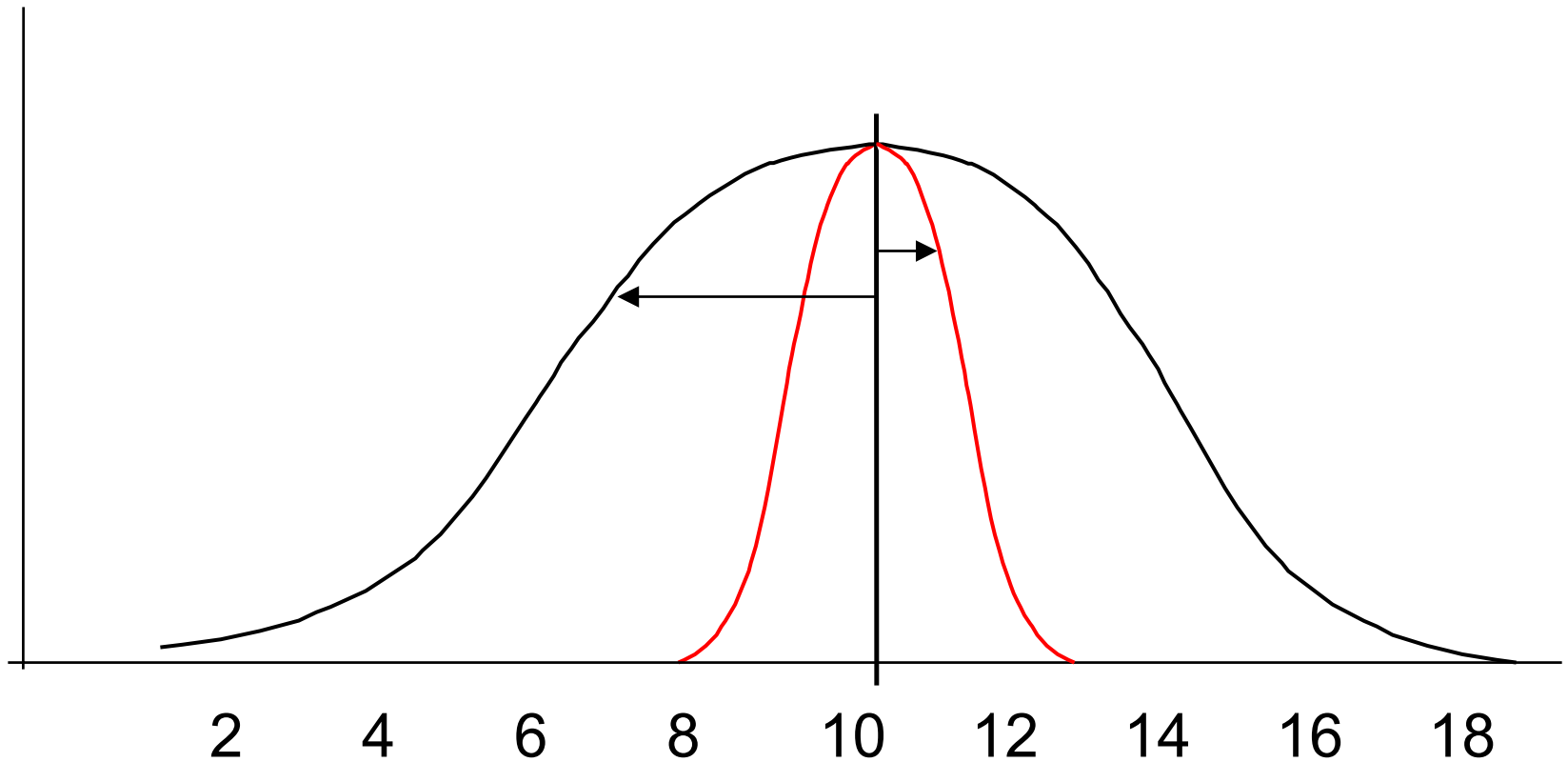
Sampling error and the mean

- Usually, our data forms only a small part of all the possible data we could collect
 - All possible users do not participate in a usability test
- The mean we observe therefore is unlikely to be the exact mean for the whole population
 - The scores of our users in a test are not going to be an exact index of how all users would perform

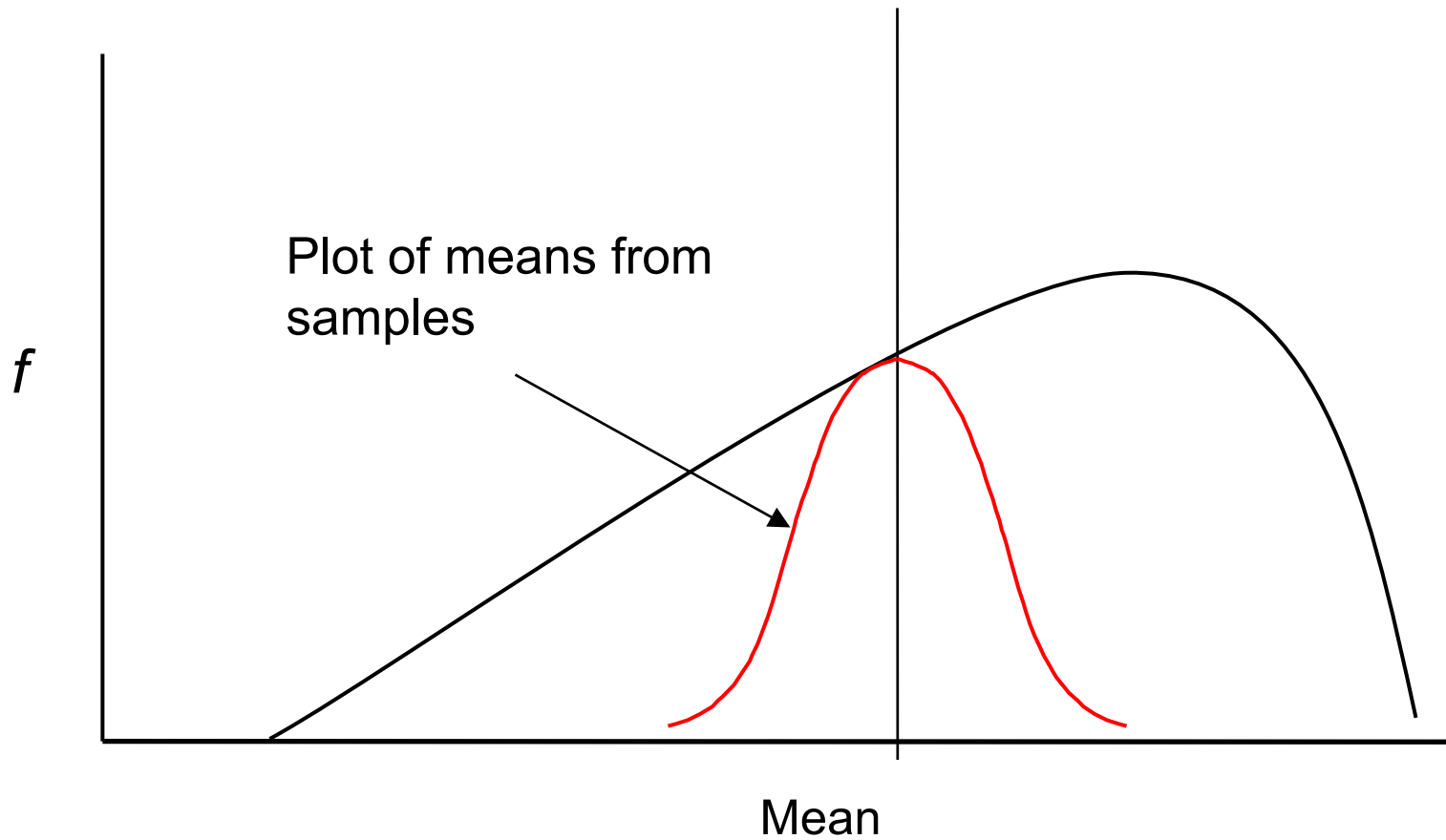
How can we relate our sample to everyone else?

- Central limit theorem
 - If we repeatedly sample and calculate means from a population, our list of means will itself be normally distributed
 - Holds true even for samples taken from a skewed population distribution
- This implies that our observed mean follows the same rules as all data under the normal curve

The distribution of the means forms a smaller normal distribution about the true mean:



True for skewed distributions too



How means behave..

- A mean of any sample belongs to a normal distribution of possible means of samples
- Any normal distribution behaves lawfully
- If we calculate the SD of all these means, we can determine what proportion (%) of means fall within specific distances of the 'true' or population mean

But...

- We only have a sample, not the population...
- We use an estimate of this SD of means known as the Standard Error of the Mean

$$SE = \frac{SD}{\sqrt{N}}$$

Implications

- Given a sample of data, we can estimate how confident we are in it being a true reflection of the 'world' or...
- If we test 10 users on an interface, we can estimate how much variability about our mean score we will find within the intended full population of users

Example

- We test 20 users on a new interface:
 - Mean error score: 10, sd: 4
 - What can we infer about the broader user population?
- According to the central limit theorem, our observed mean (10 errors) is itself 95% likely to be within 2 s.d. of the ‘true’ (but unknown to us) mean of the population

The Standard Error of the Means

$$SE = \frac{s.d.(sample)}{\sqrt{N}}$$
$$= \frac{4}{\sqrt{20}} = \frac{4}{4.47} = 0.89$$

If standard error of mean = 0.89

- Then observed (sample) mean is *within a normal distribution* about the 'true' or population mean:
 - So we can be
 - 68% confident that the true mean = 10 ± 0.89
 - 95% confident our population mean = 10 ± 1.78
 - 99% confident it is within 10 ± 2.67
- This offers a strong method of interpreting of our data

Issues to note

- If s.d. is large and/or sample size is small, the estimated deviation of the population means will appear large.
 - e.g., in last example, if $n=9$, $SE\ mean=1.33$
 - So confidence interval becomes 10 ± 2.66 (i.e., we are now 95% confident that the true mean is somewhere between 7.44 and 12.66.
 - Hence confidence improves as sample increases and variability lessens
 - Or in other words: the more users you study, the more sure you can be.....!

Exercise 3:

- If the mean = 10 and the s.d.=4, what is the 68% confidence interval when we have:
 - 16 users?
 - 9 users?
- If the s.d. = 12, and mean is still 10, what is the 95% confidence interval for those N?

Recap

- Summarizing data effectively informs us of central tendencies
- We can estimate how our data deviates from the population we are trying to estimate
- We can establish confidence intervals to enable us to make reliable 'bets' on the effects of our designs on users