Describing and Presenting Data

Probability and Statistical Inference

Bojan Božić

TU Dublin

Winter 2020

Sources

Sources used in creation of this lecture:

- Peck, Roxy, Chris Olsen, and Jay L. Devore. Introduction to statistics and data analysis. Cengage Learning, 2015.
- Field, Andy. Discovering statistics using IBM SPSS statistics. sage, 2013.
- Brase, Charles Henry, and Corrinne Pellillo Brase. Understanding basic statistics. Cengage Learning, 2013.

Main Literature

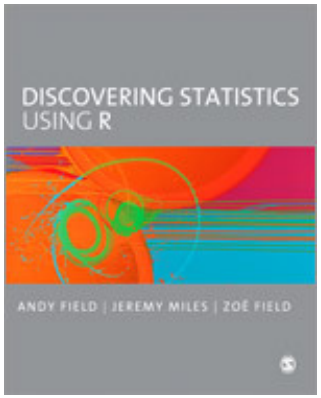


Figure: Discovering Statistics Using R - Field, Miles & Field: `https://studysites.uk.sagepub.com/dsur/main.htm`

Student Copies

There is a 30% discount if you use the DTS code UK19STTUD after following this link: `https://uk.sagepub.com/en-gb/eur/discovering-statistics-using-r/book236067`.

Basics of Statistics

- Science of collection, presentation, analysis, and reasonable interpretation of data.
- Statistics provides a rigorous scientific method for gaining insight into data.

Experiments and Variables

- For our purposes a statistical experiment or observation is any process through which measurements are obtained.
- Common to use the letter x to represent the quantitative results of an experiment or observation.
  - X is a variable, x is the value of the variable.

Random Variables

- A variable is considered a random variable if the value it takes on in a given experiment or observation is determined by chance.
  - A variable whose realisation is determined by chance.
- A discrete random variable.
  - May only take on a finite number of values or countable number of values.
  - Result of a count.
- A continuous random variable.
  - May take on any number of values in a line interval.
  - Measured on a continuous scale.

Variables

- Not only something we measure.
- Others we measure indirectly.
  - There will sometimes be a difference between the numbers we use to represent a thing we are measuring and the actual value of the thing (if we were measuring it directly).
- Measurement error.
  - E.g. psychological tests are approximate measures.
- Can be:
  - Things we can manipulate,
  - Compute or
  - Control.

Study Design

- A careful advance plan of data collection and the analytic approach is needed to answer the question under investigation in a scientific way.
- The basic elements of a study design:
  - Selecting an appropriate sample size for a specified level of power and level of significance.
  - Select appropriate measures.
  - Selecting methods of sampling, data collection, and analysis appropriate to the study's objectives.

Guidelines for Presenting Descriptive Statistics

- Ensure that you are using the most appropriate way of summarizing and presenting your data.
- Be as efficient as possible when presenting your findings.
  - All charts and tables should, as far as possible, be self-explanatory.
  - Use appropriate visualisation for the variables of interest.
- Be consistent in the way you present your findings.
- Ensure that your data are not presented in a way that may be misleading and/or confusing.

What do I need to describe for numerical data?

- Centre:
  - Discuss where the middle of the data falls.
  - Measures of central tendency: mean, median and mode.
- Spread:
  - Discuss how spread out the data is.
  - Refers to the variability in the data: Range, standard deviation, IQR.
- Shape:
  - Refers to the overall shape of the distribution.
  - Symmetrical, uniform, skewed, or bimodal.

What do I need to describe for numerical data?

- Unusual Occurrences:
  - Outliers (value that lies away from the rest of the data).
  - Gaps.
  - Clusters.
- Context:
  - You must write your answer in reference to the context in the problem, using **correct statistical vocabulary** and using complete sentences.

Methods of Centre Measurement

- Mean:
  - Summing up all the observation and dividing by number of observations.
- 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.
- Mode:
  - The value that is observed most frequently.
  - 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.
- Example:
  - Mean of 20, 30, 40, and 990 is $(20+30+40+990)/4 = 270$.
  - Median of these four observations is $(30+40)/2 = 35$.
  - 3 observations out of 4 lie between 20-40.
  - Mean of 270 really fails to give a realistic picture of the major part of the data. It is influenced by extreme value 990.
  - Median is more reflective of the data.

Population Characteristic

- Suppose we want to know the **mean** length of all fish in Lough Mask ...

Population Characteristic

- Suppose we want to know the **mean** length of all fish in Lough Mask ...
- Is this a known value?

Population Characteristic

- Suppose we want to know the **mean** length of all fish in Lough Mask ...
- Is this a known value?
- Can we find it out?

Population Characteristic

- Suppose we want to know the **mean** length of all fish in Lough Mask ...
- Is this a known value?
- Can we find it out?
- At any given point in time, how many values are there for the mean length of fish in the lake?

Population Characteristic

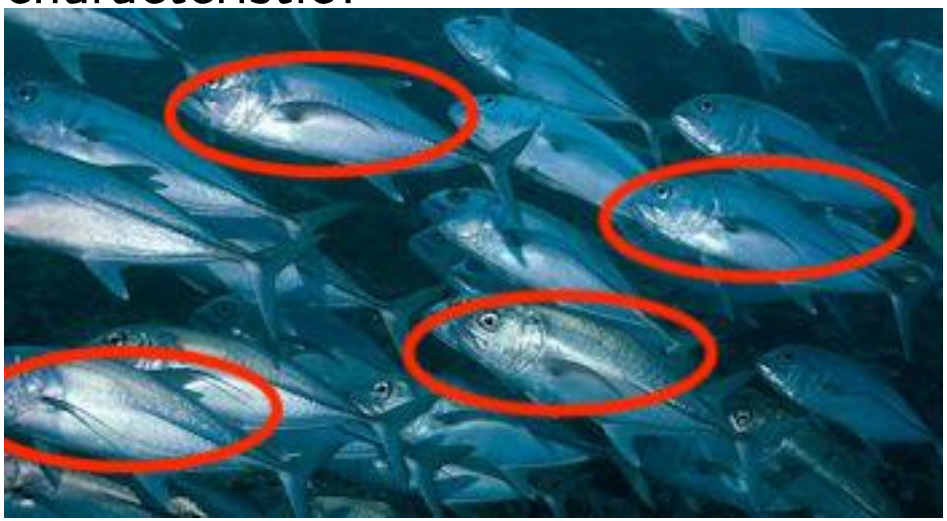
- Suppose we want to know the **mean** length of all fish in Lough Mask ...
- Is this a known value?
- Can we find it out?
- At any given point in time, how many values are there for the mean length of fish in the lake?
- **Fixed value about a population is typically unknown.**

Statistic

- What can we do to estimate this unknown population characteristic?

Statistic

- What can we do to estimate this unknown population characteristic?



- **Calculate value from a sample.**

Review

- In a symmetrical distribution, the mean and median are equal.
    - In a symmetrical distribution, you should report the mean.
- In a skewed distribution, the mean is pulled in the direction of the skewness.
    - In a skewed distribution, you should report the median.

Trimmed Mean

- Purpose is to remove outliers from a data set.
- To calculate a trimmed mean:
  - Multiply the percent to trim by n (number in the sample).
  - Truncate that many observations from BOTH ends of the distribution (when listed in order).
  - Calculate the mean with the shortened data set.
- Not often used for large datasets.
  - Example Olympic Diving/Gymnastics scoring.
    - Used to eliminate extreme scores/bias from judges.

Example

Find the mean of the following set of data:

12, 14, 19, 20, 22, 24, 25, 26, 26, 50

Example

Find the mean of the following set of data:

12, 14, 19, 20, 22, 24, 25, 26, 26, 50

Arithmetic Mean

$$\overline{x} = \frac{12+14+19+20+22+24+25+26+26+50}{10} = 23.8$$

Example

Find the mean of the following set of data:

12, 14, 19, 20, 22, 24, 25, 26, 26, 50

Arithmetic Mean

$$\overline{x} = \frac{12+14+19+20+22+24+25+26+26+50}{10} = 23.8$$

Find a 10% trimmed mean:

Example

Find the mean of the following set of data:

12, 14, 19, 20, 22, 24, 25, 26, 26, 50

Arithmetic Mean

$$\overline{x} = \frac{12+14+19+20+22+24+25+26+26+50}{10} = 23.8$$

Find a 10% trimmed mean:

~~12~~, 14, 19, 20, 22, 24, 25, 26, 26, ~~50~~

Example

Find the mean of the following set of data:

12, 14, 19, 20, 22, 24, 25, 26, 26, 50

Arithmetic Mean

$$\overline{x} = \frac{12+14+19+20+22+24+25+26+26+50}{10} = 23.8$$

Find a 10% trimmed mean:

~~12~~, 14, 19, 20, 22, 24, 25, 26, 26, ~~50~~

Trimmed Mean

$$\overline{x}_T = \frac{14+19+20+22+24+25+26+26}{8} = 22$$

Why is the study of variability important?

- Variability (or dispersion) measures the amount of scatter in a dataset.
- There is variability in virtually everything.
- Allows us to distinguish between usual and unusual values.
- Reporting only a measure of centre doesn't provide a complete picture of the distribution.

Does this can of cola contain exactly 330ml?

Types of Variation

Systematic Variation

Differences in performance created by a specific experimental manipulation.

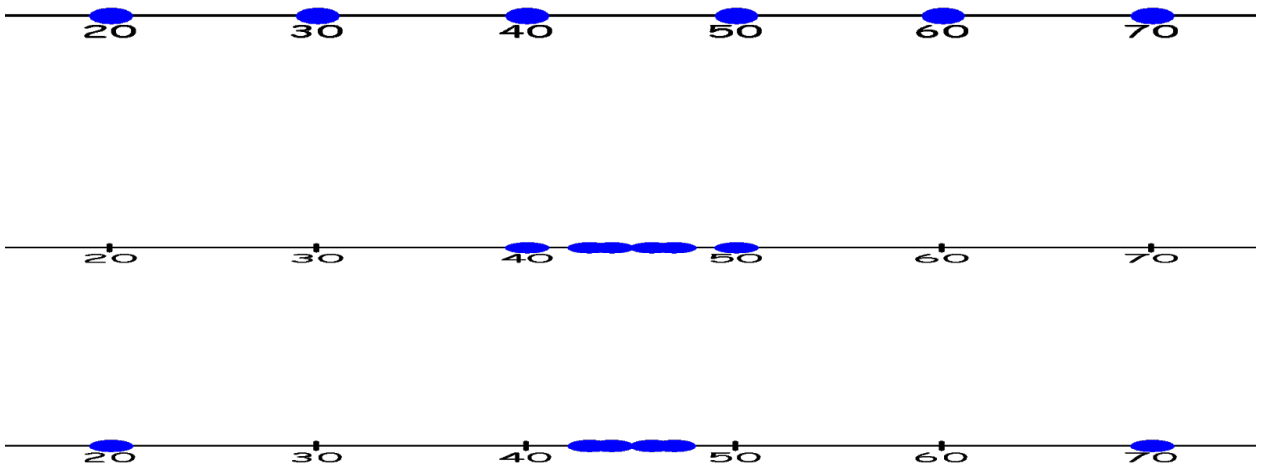
Unsystematic Variation

Differences in performance created by unknown factors. E.g. Age, gender, IQ, time of day, measurement error, etc.

Randomization

Minimizes unsystematic variation.

Variability



Notice

These three datasets all have the **same mean and median (at 45)**, but they have very **different amounts of variability**.

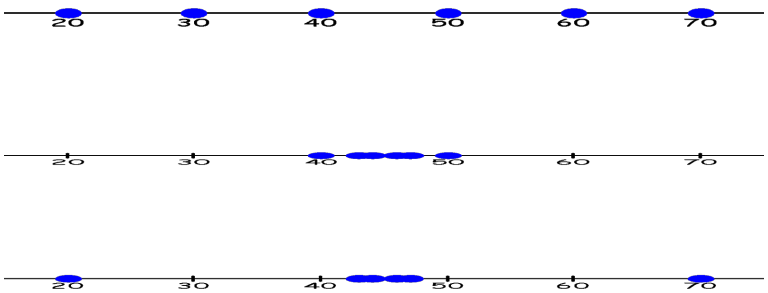
Methods of Variability/Dispersion Measurement

Commonly used methods

Range, variance, standard deviation, interquartile range, coefficient of variation etc.

Measures of Variability

- The simplest numeric measure of variability is range.
- It's a crude measure of variability though.
- Range = largest observation – smallest observation



The first and last data sets have a range of 50 (70-20) but the middle data set has a much smaller range of 10.

Measure of Variability

Another measure of variability in a dataset uses the deviations from the mean $(x - \overline{x})$.

Examples

A sample of 6 fish caught from the lake with following lengths ...

$$3'', 4'', 5'', 6'', 8'', 10''.$$

The mean length was 6 inches. We can calculate the deviations from the mean. What was the sum of these deviations?

- Can we find an average deviation?
- How can we find an average by using deviations?
- **Variance** is estimated average of the deviations squared:

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

Measures of Variability

Find the variance of the length of the 6 fish.

x	$(x - \bar{x})$	$(x - \bar{x})^2$
3	-3	9
4	-2	4
5	-1	1
6	0	0
8	2	4
10	4	16
Sum	0	34

- Finding the average of deviations would always equal 0 (in a symmetrical distribution).
- First square the deviations.
- What is the sum of deviations squared?
- Divide this by 5.
- $s^2 = 6.8$
- Standard Deviation.

Measures of Variability

- The square root of variance is called **standard deviation**.
- A typical deviation from the mean is the **standard deviation**.
- Our fish example: $s^2 = 6.8 inches^2$ so $s = 2.608 inches$
- The fish in our sample deviate from the mean of 6 by an average of 2.608 inches.

Example

Examples

Suppose that everyone in the class caught a sample of 6 fish from the lake. Would each of our samples contain the same fish? Would our mean lengths be the same? No, we would also have different ranges!

When calculating sample variance, we use degrees of freedom $(n-1)$ in the denominator instead of n because this tends to produce better estimates.

Degrees of Freedom

- Degrees of freedom of an estimate is **the number of independent pieces of information that went into calculating the estimate**.
  - It's not quite the same as the number of items in the sample.
- In order to get the degree of freedom for the estimate, you have to subtract 1 from the number of items.
- Why subtract 1?
  - Another way to look at degrees of freedom is that they are the number of values that are free to vary in a dataset.
  - Or the number of values that need to be known in order to know all the values needed to achieve a particular value.

Degrees of Freedom

- What does "free to vary" mean?
- An example using the mean:
  - Pick a set of numbers that have a mean of 10.
  - Some sets of numbers you might pick: 9, 10, 11 or 8, 10, 12 or 5, 10, 15.
  - Once you have chosen the first two numbers in the set, the third is fixed.
  - In other words, you can't choose the third item in the set.
    - The only numbers that are free to vary are the first two.
    - You can pick $9 + 10$ or $5 + 15$, but once you've made that decision you must choose a particular number that will give you the mean you are looking for.
    - So degrees of freedom for a set of three numbesr is TWO (n-1).

Degrees of Freedom

- Two Samples.
- If you have two samples and want to find a parameter, like the mean, you have two "N"'s to consider $(sample1(N_1) and Sample2(N_2))$.
- Degrees of freedom in that case are: $(N_1 + N_2) - 2$.

Measures of Variability

- Interquartile range (IQR) is the range of the middle half of the data.
- Lower quartile (Q1) is the median of the lower half of the data.
- Upper quartile (Q3) is the median of the upper half of the data.
- $iqr = Q3 - Q1$

Measures of Variability

- Interquartile range (IQR) is the range of the middle half of the data.
- Lower quartile (Q1) is the median of the lower half of the data.
- Upper quartile (Q3) is the median of the upper half of the data.
- $iqr = Q3 - Q1$

Question

What advantage does the interquartile range have over the standard deviation?

Measures of Variability

- Interquartile range (IQR) is the range of the middle half of the data.
- Lower quartile (Q1) is the median of the lower half of the data.
- Upper quartile (Q3) is the median of the upper half of the data.
- $iqr = Q3 - Q1$

Question

What advantage does the interquartile range have over the standard deviation?

The IQR is resistant to extreme values.

Example

Examples

The chronicle of Higher Education (2009-2010 issue) published the accompanying data on the percentage of the population with a bachelor's or higher degree in 2007 for each of the 50 states and the District of Columbia.

Example

Examples

The chronicle of Higher Education (2009-2010 issue) published the accompanying data on the percentage of the population with a bachelor's or higher degree in 2007 for each of the 50 states and the District of Columbia.

21 27 26 19 30 35 35 26 47 26 27 30 24 29 22 24 29 20 20 27 35
38 25 31 19 24 27 27 23 34 25 32 26 24 22 28 26 30 23 25 22 25
29 33 34 30 17 25 23 34 26

Example

Examples

The chronicle of Higher Education (2009-2010 issue) published the accompanying data on the percentage of the population with a bachelor's or higher degree in 2007 for each of the 50 states and the District of Columbia.

21 27 26 19 30 35 35 26 47 26 27 30 24 29 22 24 29 20 20 27 35
38 25 31 19 24 27 27 23 34 25 32 26 24 22 28 26 30 23 25 22 25
29 33 34 30 17 25 23 34 26

Question

Find the interquartile range for this set of data.

Example

21 27 26 19 30 35 35 26 47 26 27 30 24 29 22 24 29 20 20 27 35
38 25 31 19 24 27 27 23 34 25 32 26 24 22 28 26 30 23 25 22 25
29 33 34 30 17 25 23 34 26

Question

- First put the data in order and find the median.

Example

17 19 19 20 20 21 22 22 22 23 23 23 24 24 24 24 25 25 25 25 25
26 26 26 26 26 26 27 27 27 27 27 28 29 29 29 30 30 30 30 31 32
33 34 34 34 35 35 35 38 47

Question

- First put the data in order and find the median.

Example

17 19 19 20 20 21 22 22 22 23 23 23 24 24 24 24 25 25 25 25 25
26 26 26 26 26 26 27 27 27 27 27 28 29 29 29 30 30 30 30 31 32
33 34 34 34 35 35 35 38 47

Question

- First put the data in order and find the median.

Example

17 19 19 20 20 21 22 22 22 23 23 23 24 24 24 24 25 25 25 25 25
26 26 26 26 26 26 27 27 27 27 27 28 29 29 29 30 30 30 30 31 32
33 34 34 34 35 35 35 38 47

Question

- First put the data in order and find the median.
- Find the lower quartile (Q_1) by finding the median of the lower half.

Example

17 19 19 20 20 21 22 22 22 23 23 23 24 24 24 24 25 25 25 25 25
26 26 26 26 26 26 27 27 27 27 27 28 29 29 29 30 30 30 30 31 32
33 34 34 34 35 35 35 38 47

Question

- First put the data in order and find the median.
- Find the lower quartile (Q_1) by finding the median of the lower half.

Example

17 19 19 20 20 21 22 22 22 23 23 23 24 24 24 24 25 25 25 25 25
26 26 26 26 26 26 27 27 27 27 27 28 29 29 29 30 30 30 30 31 32
33 34 34 34 35 35 35 38 47

Question

- First put the data in order and find the median.
- Find the lower quartile (Q_1) by finding the median of the lower half.
- Find the upper quartile (Q_3) by finding the median of the upper half.

Example

17 19 19 20 20 21 22 22 22 23 23 23 24 24 24 24 25 25 25 25 25
26 26 26 26 26 26 27 27 27 27 27 28 29 29 29 30 30 30 30 31 32
33 34 34 34 35 35 35 38 47

Question

- First put the data in order and find the median.
- Find the lower quartile (Q_1) by finding the median of the lower half.
- Find the upper quartile (Q_3) by finding the median of the upper half.

Which Descriptive Statistic to use?

- Depends on **measurement type** and **data dispersion**.
- Interval or Ration (Scale).
  - Normally distributed.
    - Mean and Standard Deviation.
  - Skewed.
    - Median and Interquartile Range.
- Ordinal or nominal.
  - Mode and/or simple frequencies.

The Research Process (Step 4)

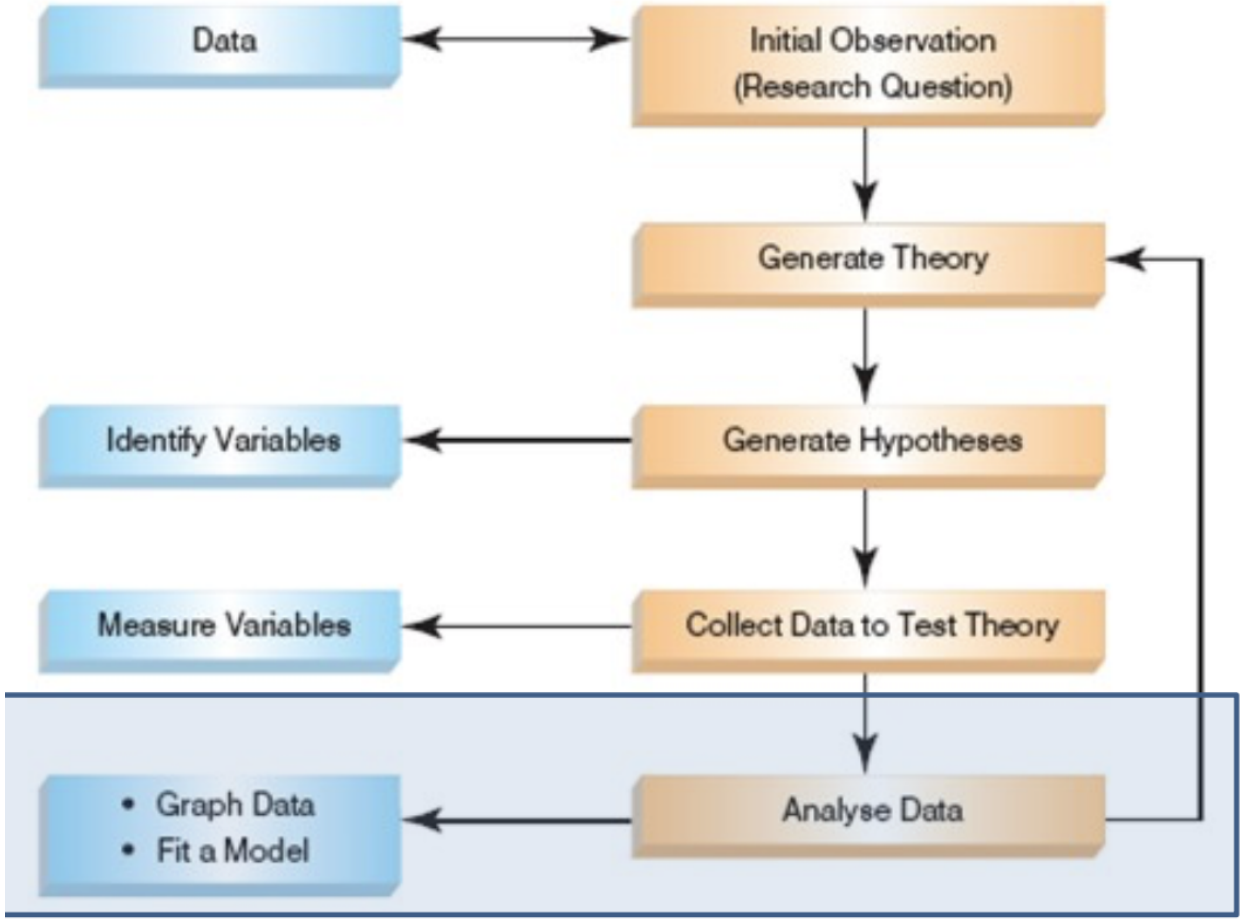


Figure: The research process - step 4

Analysing Data

- First Step: Graph the Data.
- Frequency Distributions (aka Histograms).
  - A graph plotting values of observations on the horizontal axis, with a bar showing how many times each value occurred in the data set.
- Ideal: The Normal Distribution.
  - Bell shaped.
  - Symmetrical around the centre.

The Normal Distribution

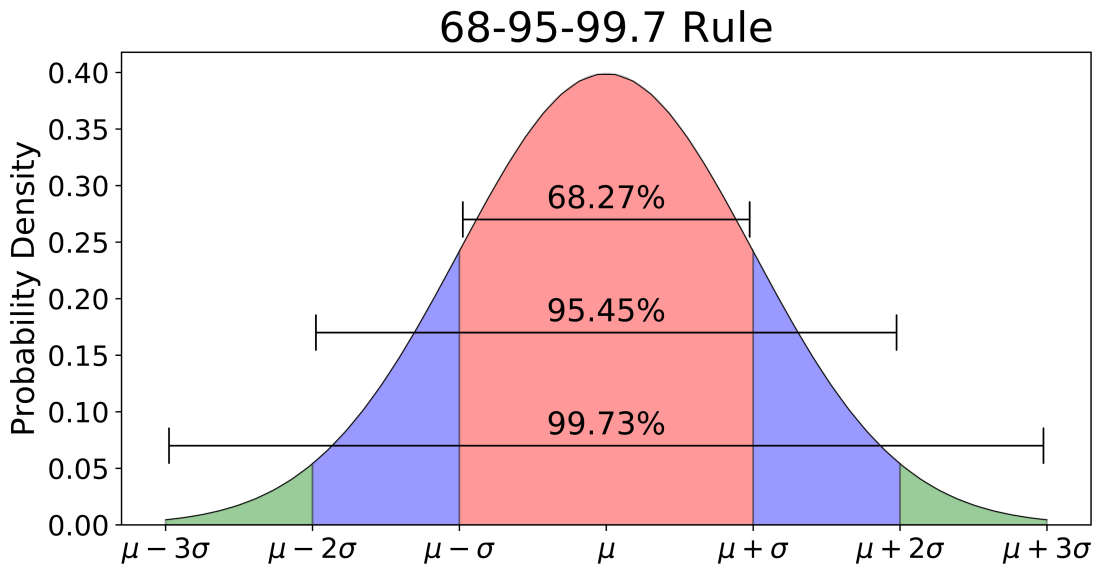


Figure: The normal distribution (with 68-95-99.7 rule.

Skew

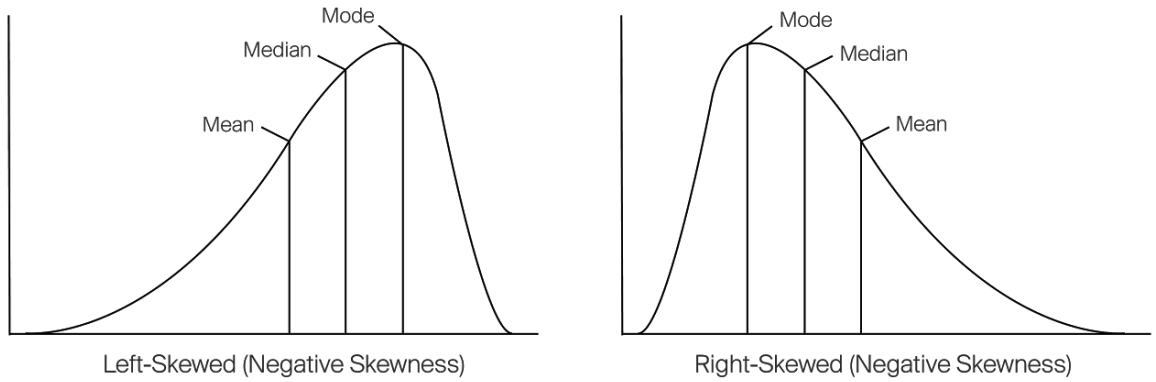


Figure: A positively and negatively skewed distribution.

Properties of Frequency Distributions

- Skew
  - The symmetry of the distribution.
  - Positive skew (scores bunched at low values with the tail pointing to high values).
  - Negative skew (scores bunched at high values with the tail pointing to low values).
- Kurtosis
  - The 'heaviness' of the tails.
  - Leptokurtic = heavy tails (more scores in the tails).
  - Platykurtic = light tails (more scores in the middle).

Kurtosis

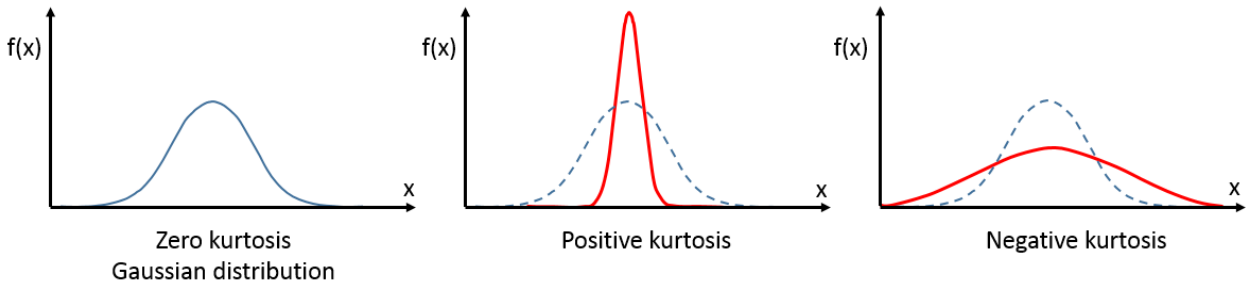


Figure: Examples for kurtosis.

Going beyond the data ...

Frequency Distribution

- Not only useful for descriptive purposes.
- Can be used to calculate likelihood of particular values occurring – **probability**.
- For any distribution we could calculate the probability of achieving any of the possible values:
  - Tedious, time consuming.
  - Statisticians have created a range of idealized distributions **probability distributions** and from these we can calculate the likelihood of achieving particular values if our data distribution matches.

Probability

- Chance behaviour is unpredictable in the short term, but has a regular and predictable pattern in the long term.
- The probability of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions.
- Sample Space:
  - The set of all possible outcomes of a random phenomenon.
- Event:
  - Any set of outcomes of interest.
- Probability of an event:
  - The relative frequency of this set of outcomes over an infinite number of trials.
- $P(A)$ is the probability of event A.

Probability Distributions

- X represents the random variable X.
- $P(X)$ represents the probability of X.
- $P(X = x)$ refers to the probability that the discrete random variable X is equal to a particular value, denoted by x.
  - As an example, $P(X = 1)$ refers to the probability that the random variable X is equal to 1.
- Cumulative probability is the probability that a value falls within a particular range or interval:
  - $P(X <= x)$

Probability Distributions

- The probability distribution for a random variable X gives the possible values for X, and the probabilities associated with each possible value (i.e., the likelihood that the values will occur).
  - Has a probability assigned to each distinct value of the variable.
- A **cumulative probability** refers to the probability that the value of a random variable falls within a specified range.
- The methods used to specify discrete probability distributions are similar to (but slightly different from) those used to specify continuous probability distributions.

Discrete Random Variable

- Has a probability assigned to each distinct value of the variable.
- The sum of all assigned probabilities must be 1.
- Probability distribution can be considered a relative-frequency distribution and therefore has a mean and standard deviation.
- Mean is often called the **expected value**:
  - Represents a cluster point for the entire distribution.
  - Need not be an actual value of a point of the sample space.
- Standard deviation is represented as a measure of **risk**.
  - The larger the standard deviation, the more likely it is that a random variable x is different from the expected value.

Discrete Probability Distribution

- Shows us the complete space on which the distribution is based.
- The corresponding probability of each event in the sample space.

Probability Distributions

- Suppose you flip a coin two times.
- This simple statistical experiment can have four possible outcomes:
  - HH (two heads), HT (heads and tails), TH (tails and heads), and TT (tails and tails).
- Let the variable X represent the number of Heads that result from this experiment.
  - X can take on the values 0, 1, or 2.
- In this example, X is a random variable because its value is determined by the outcome of a statistical experiment.

Probability Distribution

A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.

Number of heads (X)	Probability
0	0.25
1	0.5
2	0.25

Table: Probability of X = the number of Heads that result from this experiment.

Probability Distributions

- A **cumulative probability** refers to the probability that the value of a random variable falls within a specified range.
- This can be represented by a table or an equation which refers to the probability that the random variable X **is less than or equal** to x.
- If we flip a coin two times, what is the probability that the coin flips would result in one or fewer heads?
- The answer would be a cumulative probability.
  - It would be the probability that the coin flip experiment results in zero heads plus the probability that the experiment results in one head.
  - $P(X < 1) = P(X = 0) + P(X = 1) = 0.25 + 0.50 = 0.75$

Probability Distribution

Number of heads	$P(X = x)$	$P(X \leq x)$
0	0.25	0.25
1	0.5	0.75
2	0.25	1

Table: Probability distribution for probability and cumulative probability.

Probability Distribution

- The simplest probability distribution occurs when all of the values of a random variable occur with equal probability.
- This probability distribution is called the **uniform distribution**.
- Suppose the random variable X can assume k different values.
- Suppose also that the $P(X = x_k)$ is constant. Then, $P(X = x_k) = \frac{1}{k}$.
- Suppose a die is tossed.
  - What is the probability that the die will land on 5?
  - There are 6 possible outcomes represented by: $S = 1, 2, 3, 4, 5, 6$.
  - Each possible outcome is a random variable (X), and each outcome is equally likely to occur.
  - Thus, we have a uniform distribution. Therefore, $P(X = 5) = \frac{1}{6}$.

Probability Distribution

- Suppose we undertake a dice tossing experiment.
- This time, we ask what is the probability that the die will land on a number that is smaller than 5?
- There are still 6 possible outcomes represented by: $S = 1, 2, 3, 4, 5, 6$.
- This problem involves a cumulative probability.
- The probability that the die will land on a number smaller than 5 is equal to: $P(X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{2}{3}$.

Continuous Probability Distribution

- If a random variable is a **continuous variable** (variable can take on any value between two specified values), its probability distribution is called a continuous probability distribution.
- A continuous probability distribution cannot be expressed in tabular form.
  - An equation or formula (**probability density function**) is used.
  - Hypothesis testing relies extensively on the idea that, having such a function, one can compute the probability of all the corresponding events i.e. probability of a X taking a value less than or equal to a particular value (a).

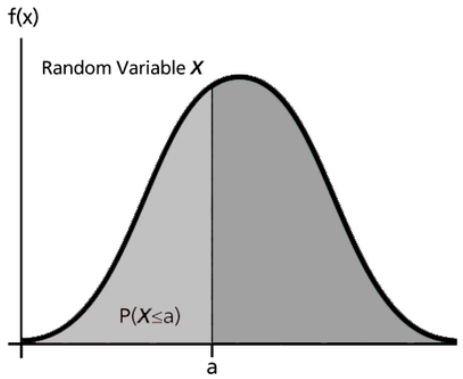
Continuous Probability Distribution

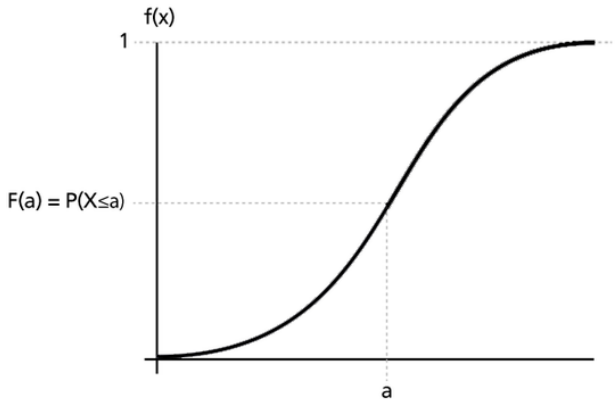
The density function has the following properties:

- Since the continuous random variable is defined over a continuous range of values (called the domain of the variable), the graph of the density function will also be continuous over that range.
- The area bounded by the curve of the density function and the x-axis is equal to 1, when computed over the domain of the variable.
- The probability that a random variable assumes a value between a and b is equal to the area under the density function graph bounded by a and b.

Continuous Probability Function

- Consider the probability density function shown in the graph.
- Suppose we wanted to know $P(X \leq a)$.
- This is equal to the area under the curve bounded by a and $-\infty$ (as indicated by the lighter area).
- The lighter area in the graph represents the probability that the random variable X is less than or equal to a, while the darker area equals probability of X being greater than a.
- This is a **cumulative probability**.





Calculating Probability from a Frequency Distribution

- Statisticians have described several common frequency distributions.
- For each they have created mathematical formulae (**probability density functions**) that specify idealized versions of these distributions.
- We can draw the function by plotting the value of a variable x against the probability of it occurring y which gives us the probability distribution.
- The area under the curve of this distribution tells us something about the probability of a value occurring .
- We can use the area under the curve between two values to tell us how likely it is that a score falls between these two values.

So what does this mean for us?

Our frequency distribution gives us the opportunity to calculate likelihood of particular values occurring – **probability** using relevant probability calculations.

- Tedious, time consuming.
- Statisticians have created a range of idealized **probability distributions** and from these we can calculate the likelihood of achieving particular values if our data distribution matches.

Standard Normal

- Statisticians have calculated the probability of scores occurring in a distribution with a mean of 0 and a standard deviation of 1.
- So what?
  - If we have data shaped like the normal distribution then the mean can be mapped to 0 and the standard deviation to 1.
  - We can then use the tables of probability created by these statisticians to work out the probability of particular scores occurring within that distribution.
- How do we map our scores to fit the standard normal?

Z Scores and Raw Scores

- If we want to compare samples with normal distributions then mean of each may be located anywhere on the x axis and the scores more/less spread out as determined by the standard deviation.
- This causes difficulties when calculating the area under the curve and hence the probability that a measurement will fall into the interval of interest.
- We could have sets of tables that calculate the area under the curve for each combination of μ and σ but this would be quite an onerous task to compile or use.

Z Scores and Raw Scores

- We need a way to standardise the distributions so we can use one table for all normal distributions.
- We can use the standard deviation as the measurement scale.
  - We consider how many standard deviations a measure is from the mean.
  - This allows comparison between a value in one normal distribution with a value in another.

Going beyond the data: Z-scores

- Standardising a score with respect to the other scores in the group.
- Expresses a score in terms of how many standard deviations it is away from the mean.
- The distribution of z-scores has a mean of 0 and $SD = 1$.

z-scores

z-score

States the position of a raw score in relation to the mean of the distribution, using the standard deviation as the unit of measurement.

$$z = \frac{rawscore - mean}{standarddeviation} \quad (2)$$

This is a z-test. The z-score is therefore a test statistic.

For a population:

$$z = \frac{X - \mu}{\sigma} \quad (3)$$

For a sample:

$$z = \frac{X - \overline{X}}{s} \quad (4)$$

1. Find the *difference* between a score and the mean of the set of scores.
2. Divide this difference by the SD (in order to assess how big it really is).

Example: Transforming IQ scores

Z-scores transform our original IQ scores into scores with a mean of 0 and an SD of 1. Raw IQ scores: mean = 100, SD = 15. z for 100 = (100-100) / 15 = 0, z for 115 = (115-100) / 15 = 1, z for 70 = (70-100) / -2, etc.

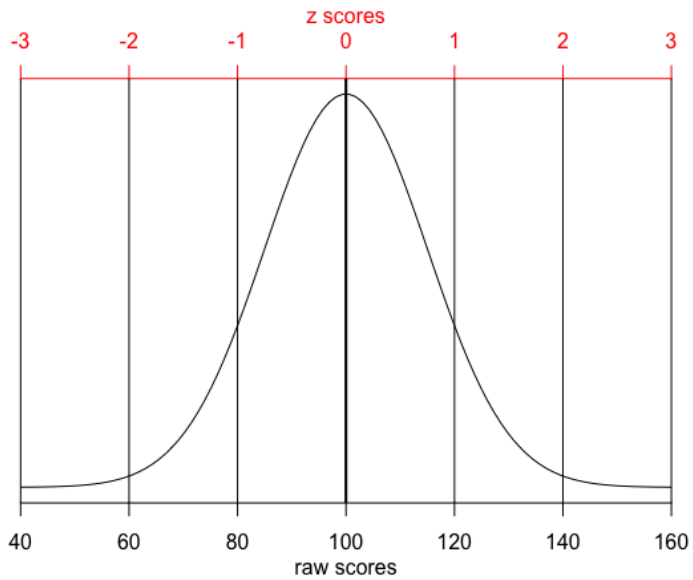
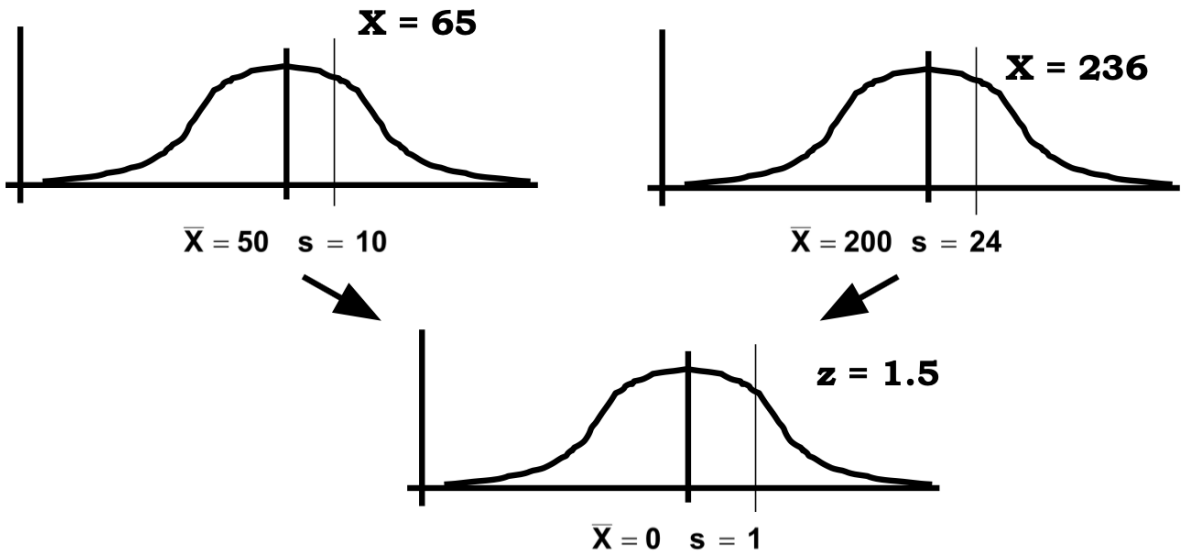


Figure: IQ Scores with raw scores and z scores.

Raw score distributions vs Z Score distributions

A score, X, is expressed in the original units of measurement:



Z-score Distribution

X is expressed in terms of its deviation from the mean (in SDs). So plotting the z scores we are plotting the test statistic.

The Standard Normal Distribution

- The distribution of a normal variable with mean equal to zero and standard deviation equal to 1.
  - Looks identical to that of the normal but uses a different measurement scale.
- So what?
  - It is the fact that we can now have a table showing, for each point in $[-\infty, +\infty]$, the probability that we have a realisation of a variable to the left and to the right of that point.

Normal Distribution Table

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9958	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Figure: The probability that a realisation is lower than point 2.33 = 0.99. Then the probability that the realisation is above 2.33 (1-0.99) = 0.01.

Why use z-scores?

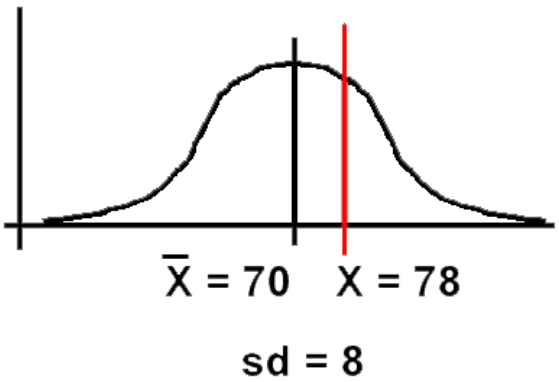
- z-scores make it easier to compare scores from distributions using different scales.
- e.g. two tests:
  - Test A: Fred scores 78. Mean score = 70, SD = 8.
  - Test B: Fred scores 78. Mean score = 66, SD = 6.
  - Did Fred do better or worse in comparison to the rest of the class on the second test?

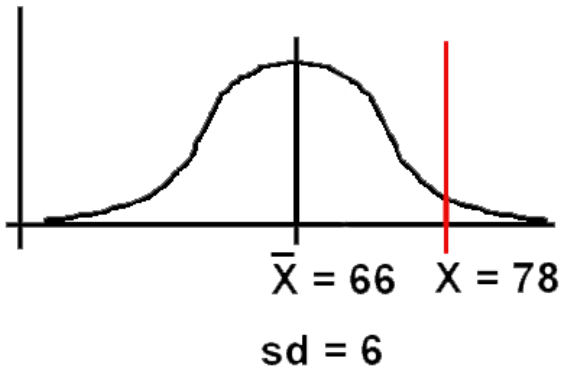
Solution

Test A: as a z-score, $z = (78 - 70)/8 = 1.00$

Test B: as a z-score , $z = (78 - 66)/6 = 2.00$

Conclusion: Fred comparatively did much better on Test B.





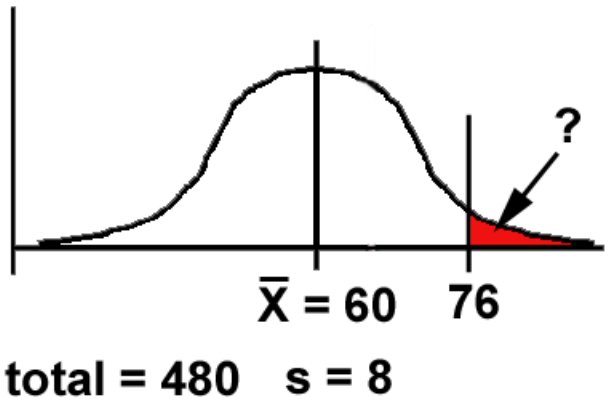
Problem

z-scores enable us to determine the relationship between one score and the rest of the scores, using just one table for all normal distributions.

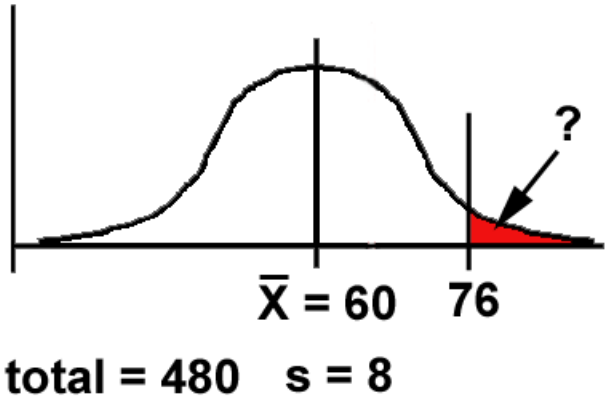
Examples

If we have 480 scores, normally distributed with a mean of 60 and an SD of 8, how many would be 76 or above?

Graph the problem:



Problem



Work out the z-score for 76:

$$z = \frac{(X - \overline{X})}{s} = \frac{(76 - 60)}{8} = \frac{16}{8} = 2.00$$

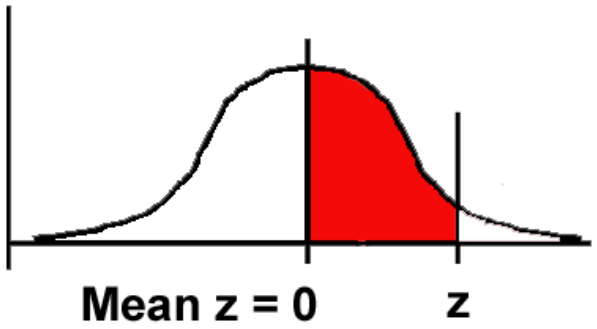
We need to know the size of the area beyond z (remember - the area under the Normal curve corresponds directly to the proportion of scores).

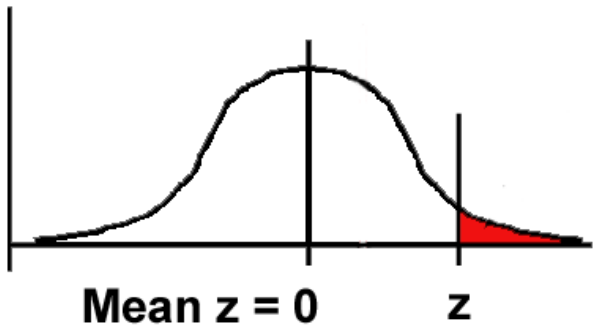
Problem

Many statistics books have z-score tables, giving us this information:

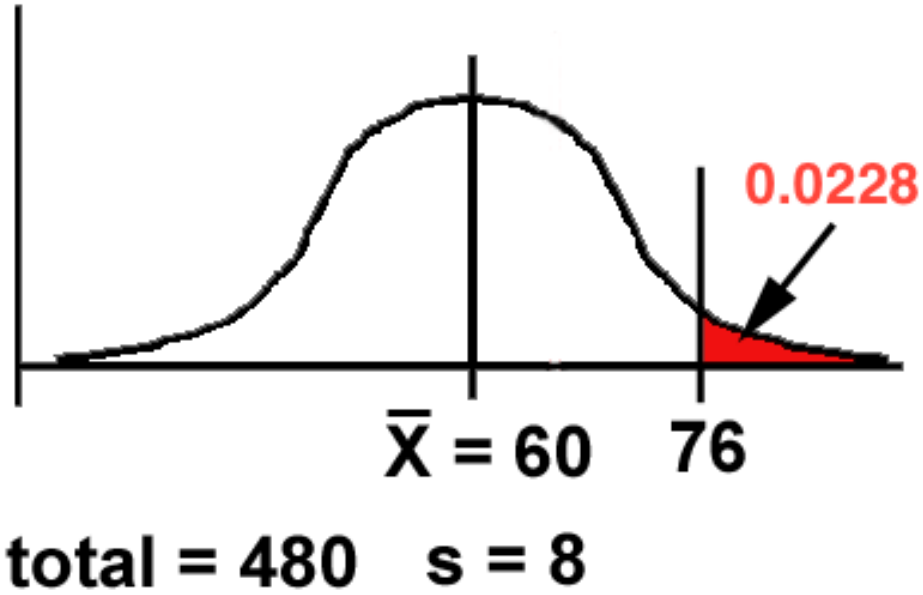
z	(a) Area between mean and z	(b) Area beyond z
0.00	0.000	0.5000
0.01	0.0040	0.4960
0.02	0.0080	0.4920
...	...	...
1.00	0.3413	0.1587
...	...	...
2.00	0.4772	0.0228
...	...	...
3.00	0.4987	0.0013

(a)





Problem



So: as a proportion of 1, 0.0228 of scores are likely to be 76 or more.

As a percentage: 2.28%

As a number: 0.0228 * 480 = 10.94 scores.

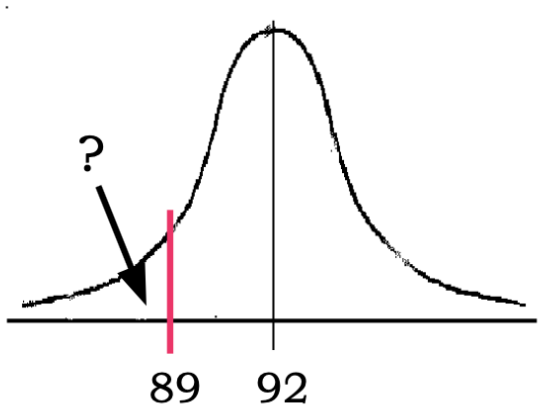
Problem

Word comprehension test scores:

A person has a score of 89 out of 100. Mean = 92, SD = 6. Is this person's comprehension significantly impaired?

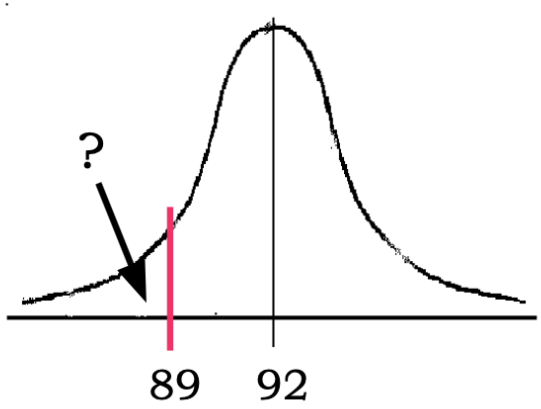
1. Graph the problem.
2. Convert 89 into a z-score.

$$z = \frac{(89-92)}{6} = \frac{-3}{6} = -0.5$$



Problem

3. Use the table to find the "area beyond z" for our z of - 0.5.



Area beyond z = 0.3085.

z-score value:	Area between the mean and z:	Area beyond z:
0.44	0.17	0.33
0.45	0.1736	0.3264
0.46	0.1772	0.3228
0.47	0.1808	0.3192
0.48	0.1844	0.3156
0.49	0.1879	0.3121
→ 0.5	0.1915	→ 0.3085
0.51	0.195	0.305
0.52	0.1985	0.3015
0.53	0.2019	0.2981
0.54	0.2054	0.2946
0.55	0.2088	0.2912
0.56	0.2123	0.2877
0.57	0.2157	0.2843
0.58	0.219	0.281
0.59	0.2224	0.2776
0.6	0.2257	0.2743
0.61	0.2291	0.2709

Conclusion

.31 (31%) of people are likely to have a comprehension score this low or lower.

The Normal Distribution

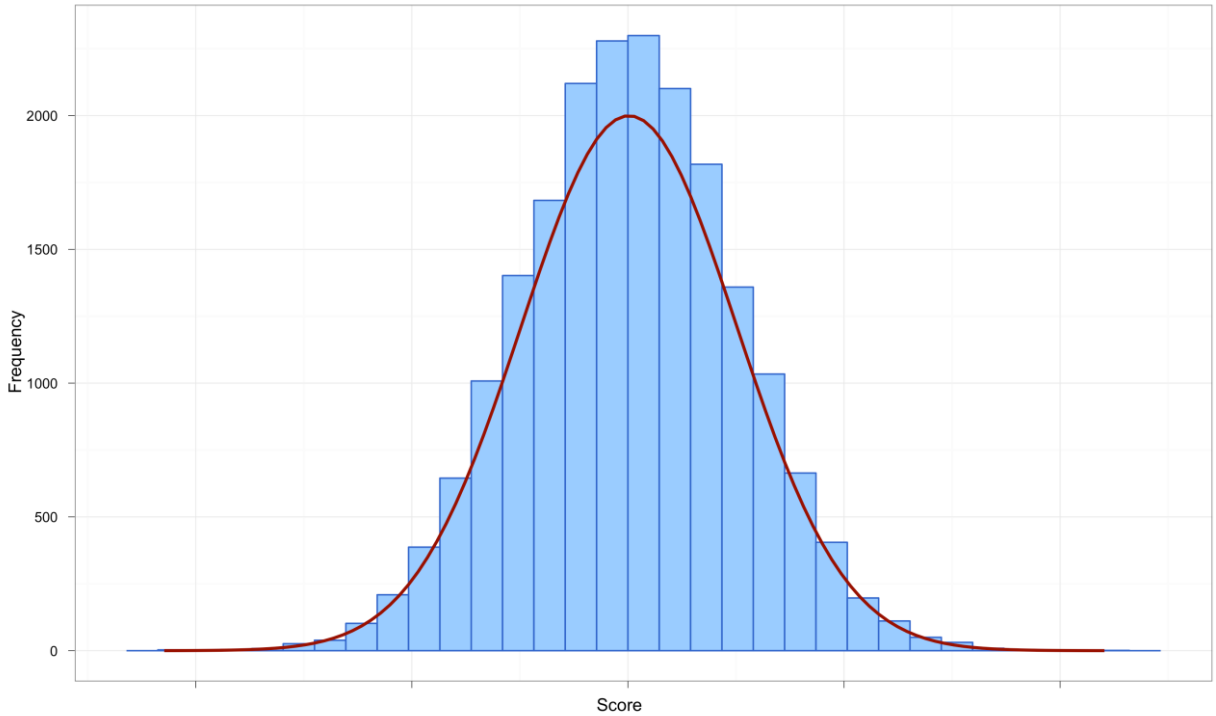


Figure: The curve shows the idealized shape. **It is important that our data is close to this shape if we wish to use Parametric tests.**

The Normal Distribution

- Normal Curve or Bell-shaped Curve.
  - Key players: Abraham DeMoivre (1667-1754) and Carl Frederick Gauss (1777-1855).
  - Sometimes normal distribution is referred to as a Gaussian distribution.
- Smooth, symmetrical curve about the mean which is the highest point of the curve.
- Approaches the horizontal axis but never touches it (asymptotic).
- The spread of the curve is determined by the standard deviation.
  - Larger this value the more spread out the curve is, smaller the more peaked it is.
- The inflection points where it starts to transition are determined by the mean +/- one standard deviation.
- The area under the curve is 1.

Normal Distribution

- A *density curve* describes the overall pattern of a distribution.
- The Formula used to generate the shape of the curve is the *normal density function*.
- A distribution is **normal** if its density curve is symmetric, single-peaked and bell-shaped.
  - Mean, median, and mode are same for a normal distribution.

Properties of the Normal Distribution

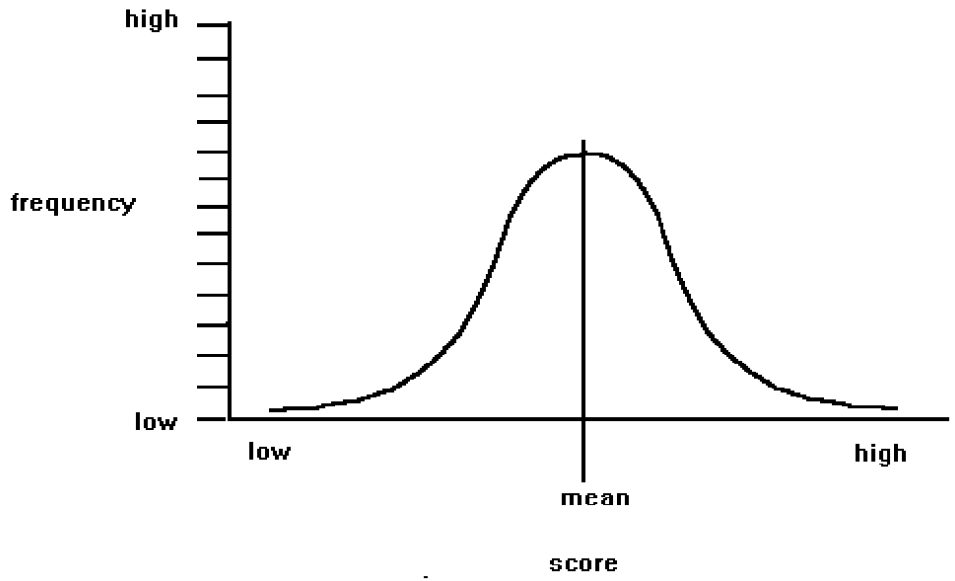


Figure: 1. It is bell-shaped and asymptotic at the extremes.

Properties of the Normal Distribution

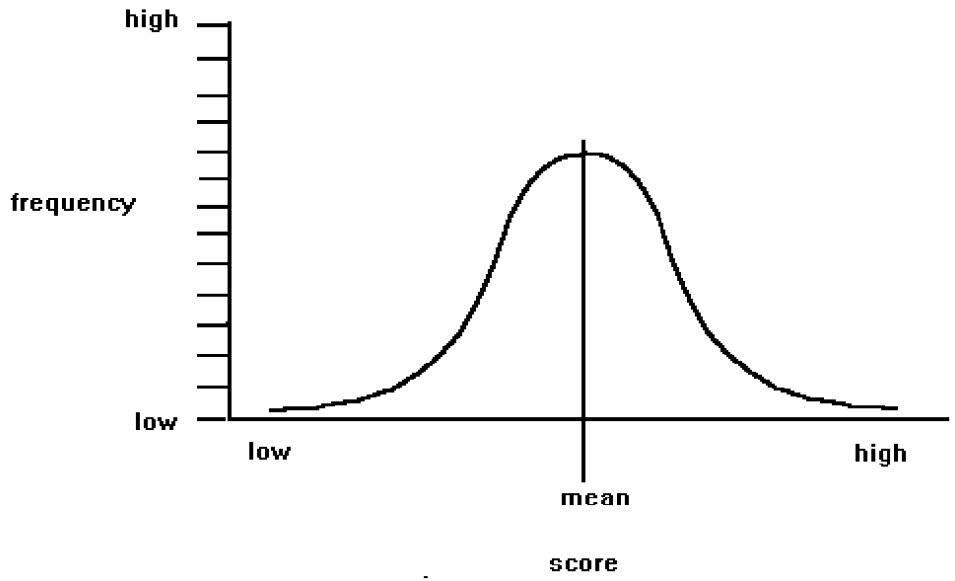


Figure: 2. It's symmetrical around the mean.

Properties of the Normal Distribution

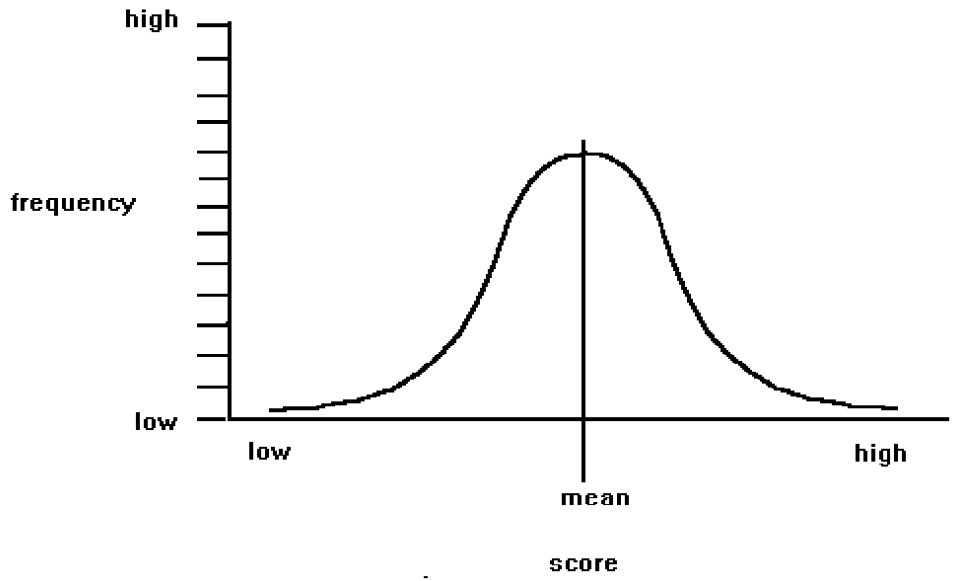


Figure: 3. The mean, median and mode all have same value.

Properties of the Normal Distribution

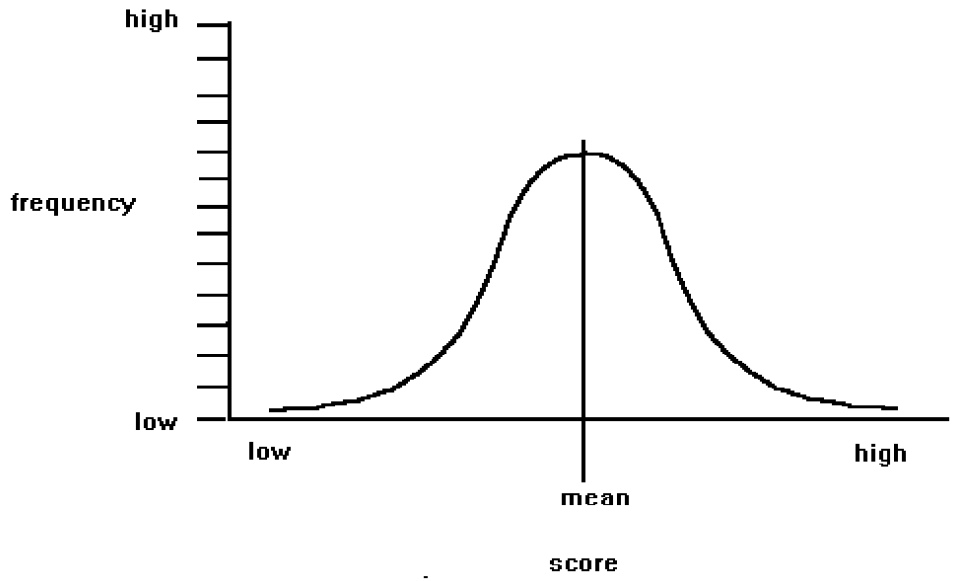


Figure: 4. It can be specified completely, once mean and SD are known.

Properties of the Normal Distribution

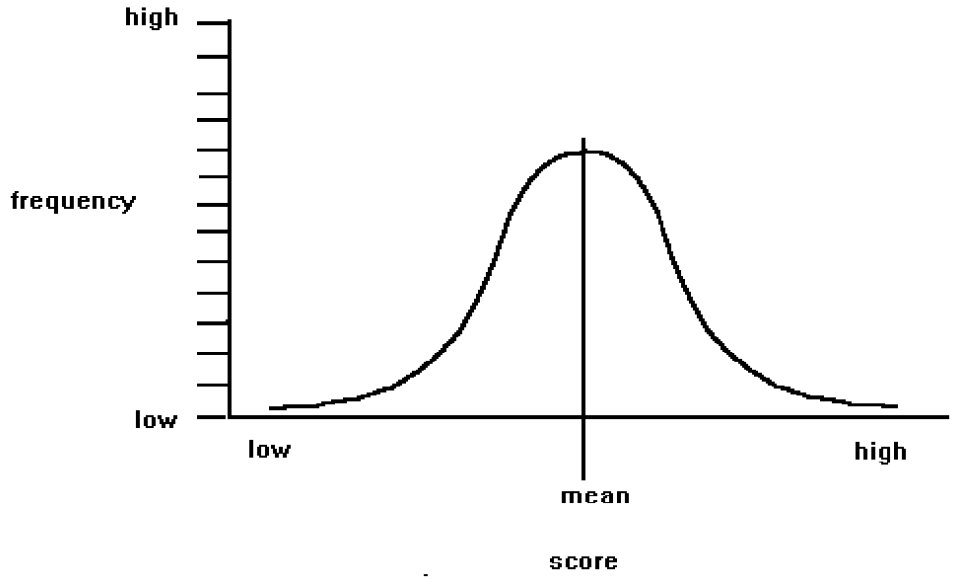


Figure: 5. The area under the curve is directly proportional to the relative frequency of observations. Thus we can calculate the probability of observations occurring in a population.

Properties of the Normal Distribution

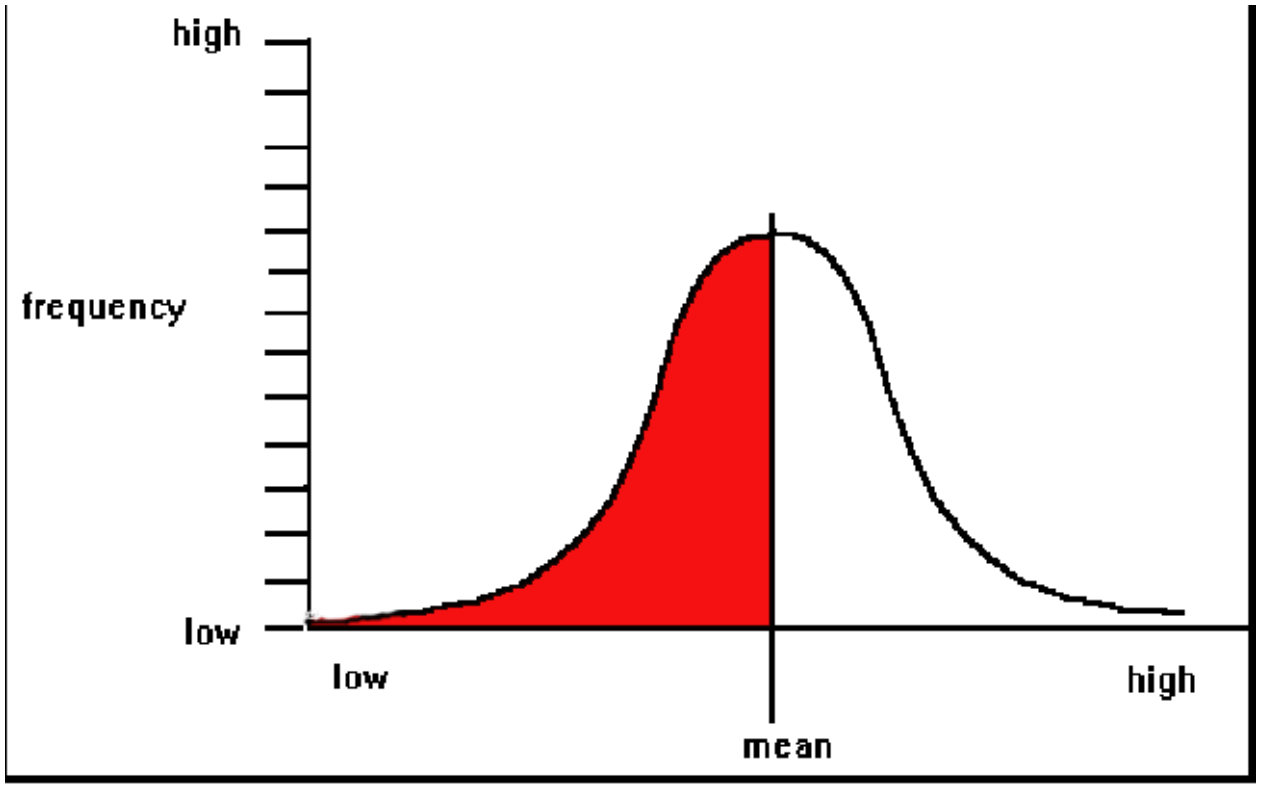


Figure: e.g. here, 50% of scores fall below the mean, as does 50% of the area under the curve.

Properties of the Normal Distribution

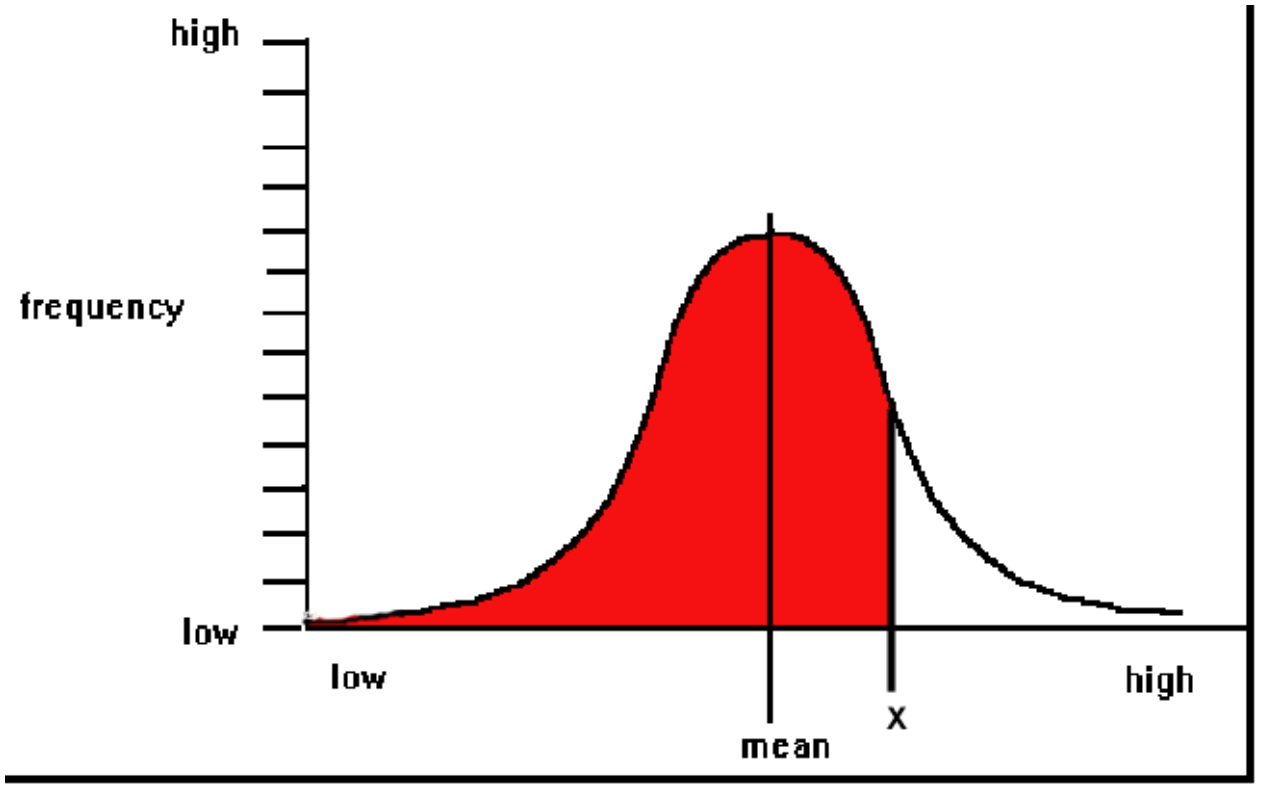


Figure: e.g. here, 85% of scores fall below score X, corresponding to 85% of the area under the curve.

Normal Distribution

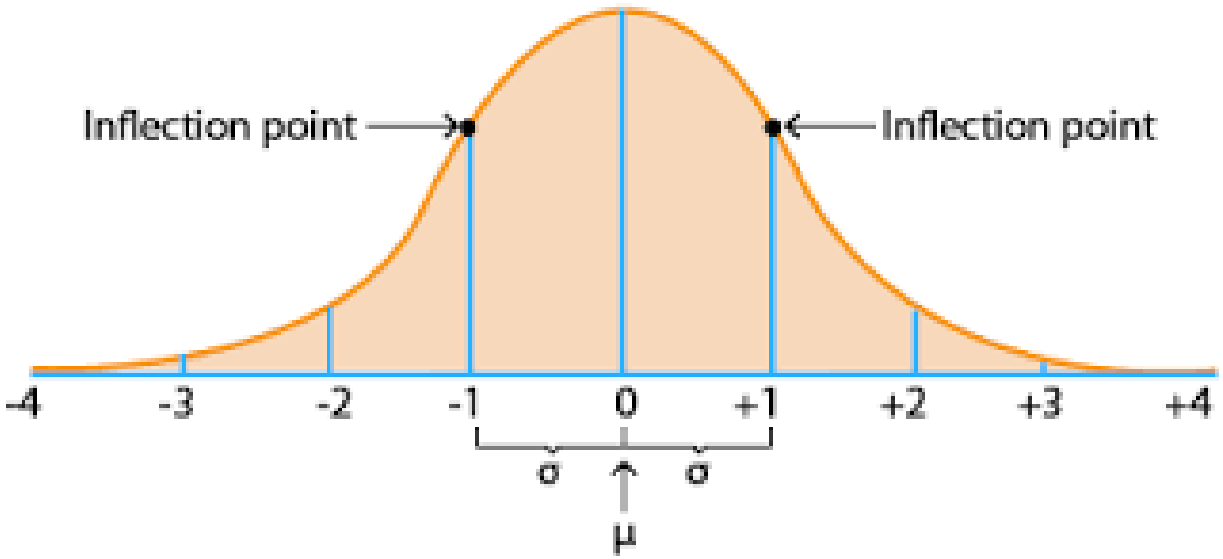
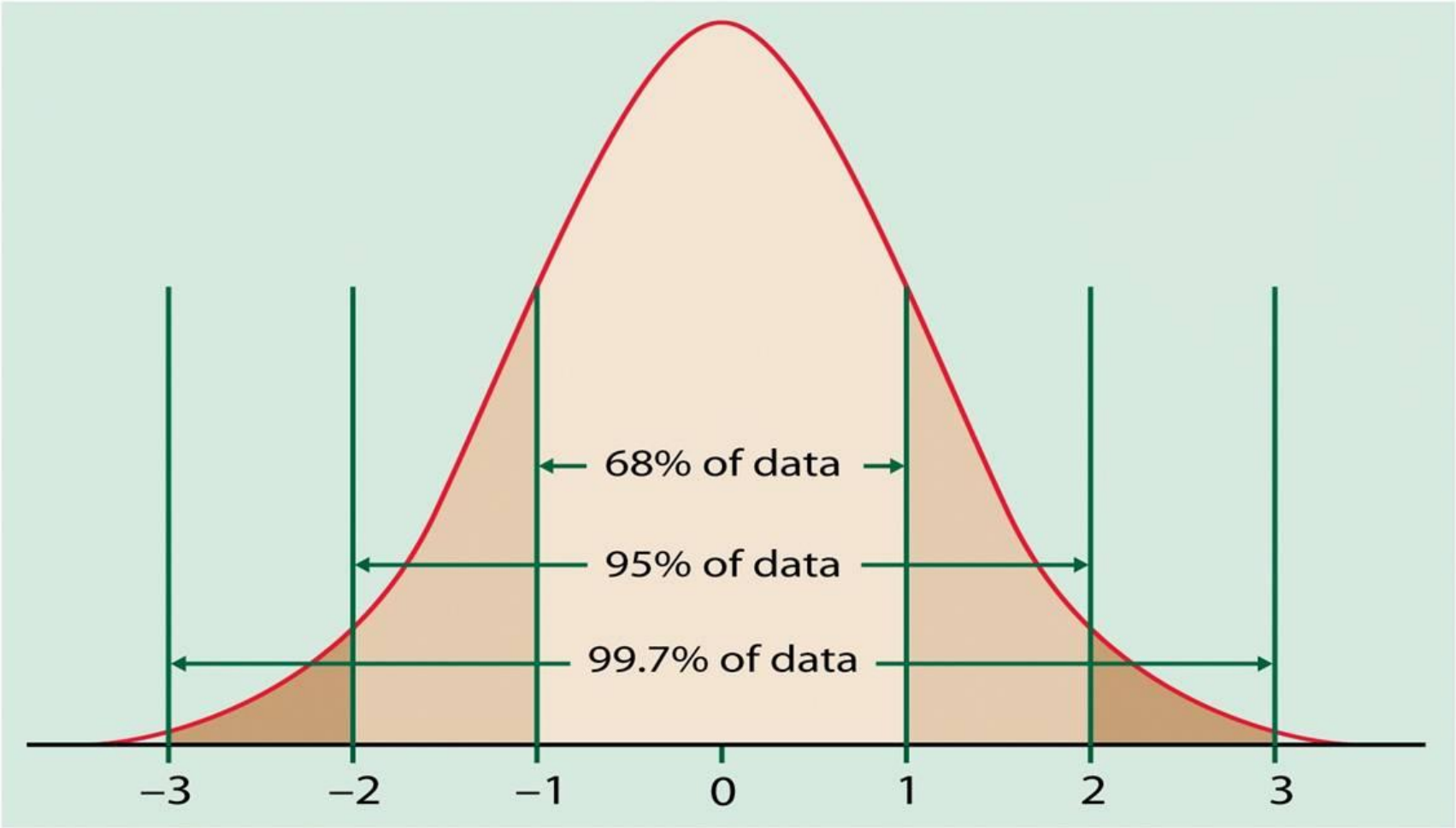


Figure: If we know μ and σ, we derive a lot of additional information about the data with a normal distribution.

Normal Distribution

- The Empirical Rule - The 68-95-99.7 Rule .
- In the normal distribution with mean μ and standard deviation σ:
  - 68% of the observations fall within σ of the mean μ.
  - 95% of the observations fall within 2σ of the mean μ.
  - 99.7% of the observations fall within 3σ of the mean μ.

Normal Distribution



Normal Distribution

If a variable is normally distributed, then:

- within one standard deviation of the mean there will be approximately 68% of the data.
- within two standard deviations of the mean there will be approximately 95% of the data.
- within three standard deviations of the mean there will be approximately 99.7% of the data.

Properties of z-scores

- 1.96 cuts off the top 2.5% of the distribution.
- -1.96 cuts off the bottom 2.5% of the distribution.
- As such, 95% of z-scores lie between -1.96 and 1.96.
- 99% of z-scores lie between -2.58 and 2.58,
- 99.9% of them lie between -3.29 and 3.29.

Normal Distribution in Summary

- Many psychological/biological properties are normally distributed.
- This is very important for statistical inference (extrapolating from samples to populations).
- z-scores provide a way of
  1. comparing scores on different raw-score scales;
  2. showing how a given score stands in relation to the overall set of scores;
  3. using probability tables to calculate likelihood of particular scores.

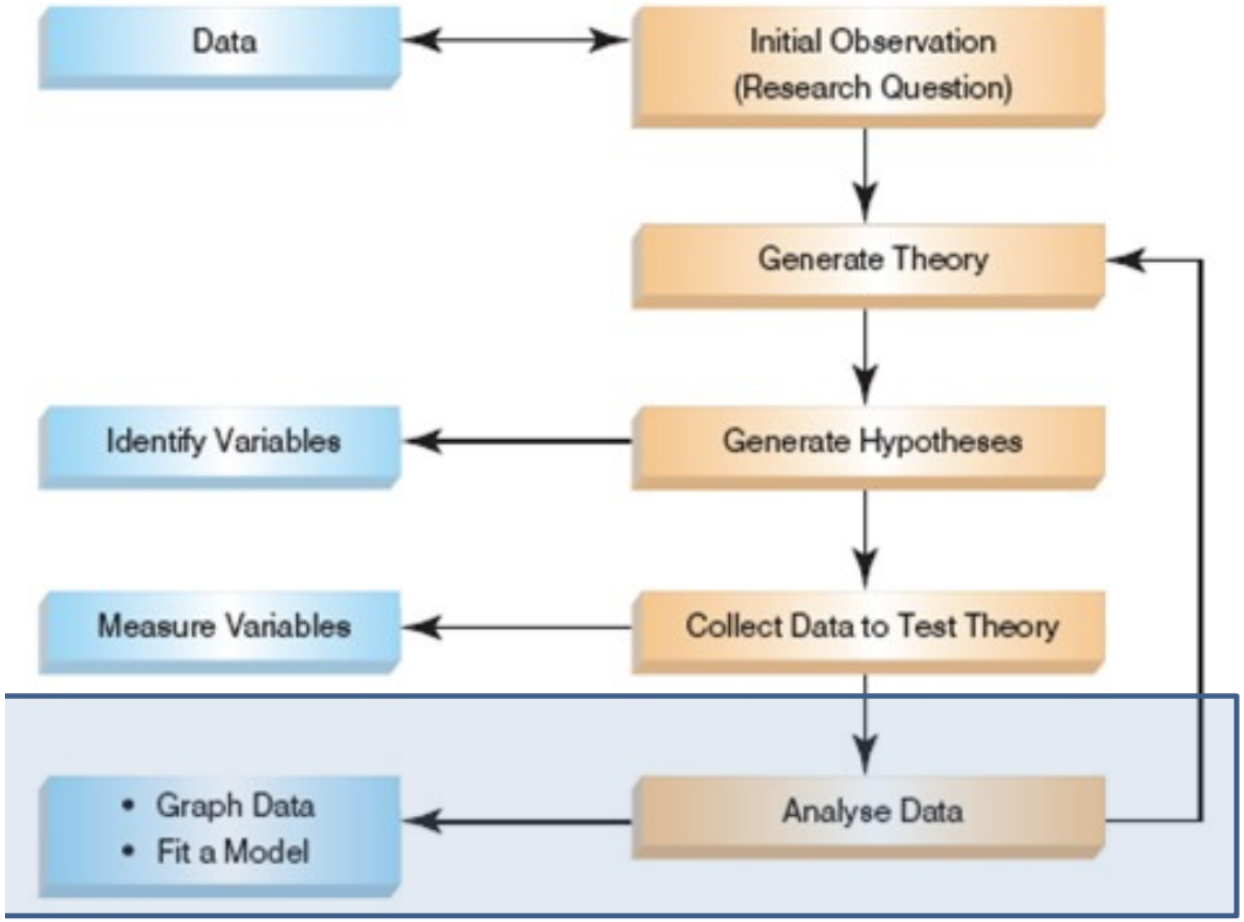
Normal Distribution in Summary

- The logic of z-scores underlies many statistical tests:
  1. Scores are normally distributed around their mean.
  2. Sample means are normally distributed around the population mean.
  3. Differences between sample means are normally distributed around zero ("no difference").
- We can exploit these phenomena in devising tests to help us decide whether or not an observed difference between sample means is due to chance.

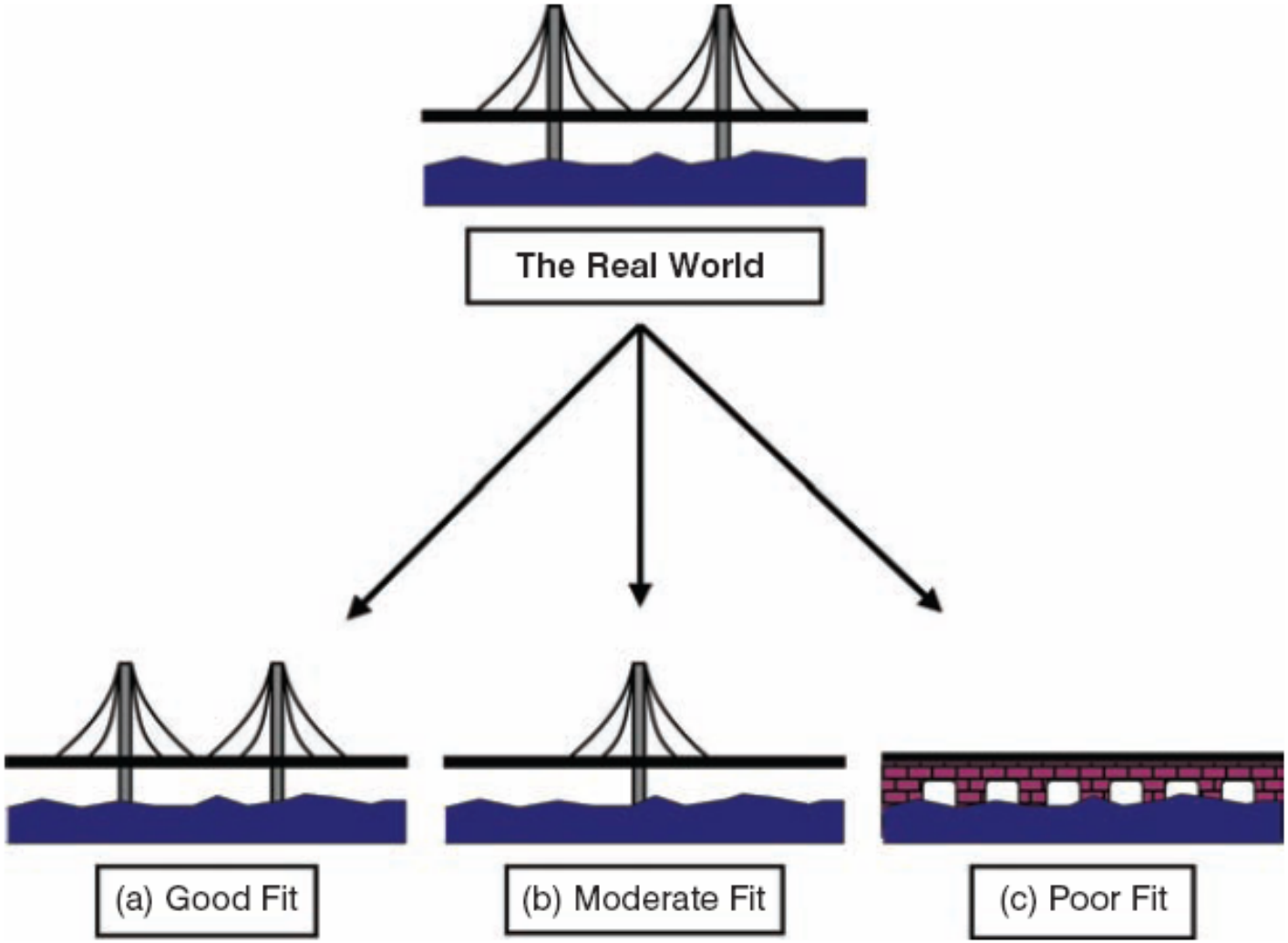
Distribution is central to choosing the correct test

- Parametric Tests
  - Normal distribution
- Non-parametric Tests
  - Non-normal distribution
- Always start by looking at the data!

The Research Process



Why do we build statistical models?



Populations and Samples

- Population
  - The collection of units (be they people, plankton, plants, cities, suicidal authors, etc.) to which we want to generalize a set of findings or a statistical model.
- Sample
  - A smaller (but hopefully representative) collection of units from a population used to determine truths about that population.

The only equation you will ever need ...

$$outcome_i = (model) + error_i$$

A Simple Statistical Model

- In statistics we fit models to our data (i.e. we use a statistical model to represent what is happening in the real world).
- The mean is a hypothetical value (i.e. it doesn't have to be a value that actually exists in the data set).
- As such, the mean is simple statistical model.

The Mean

- The mean is the sum of all scores divided by the number of scores.
- The mean is also the value from which the (squared) scores deviate least (it has the least error).

Measuring the 'Fit' of the Model

- The mean is a *model* of what happens in the real world: the *typical* score.
- It is not a perfect representation of the data.
- How can we assess how well the mean represents reality?

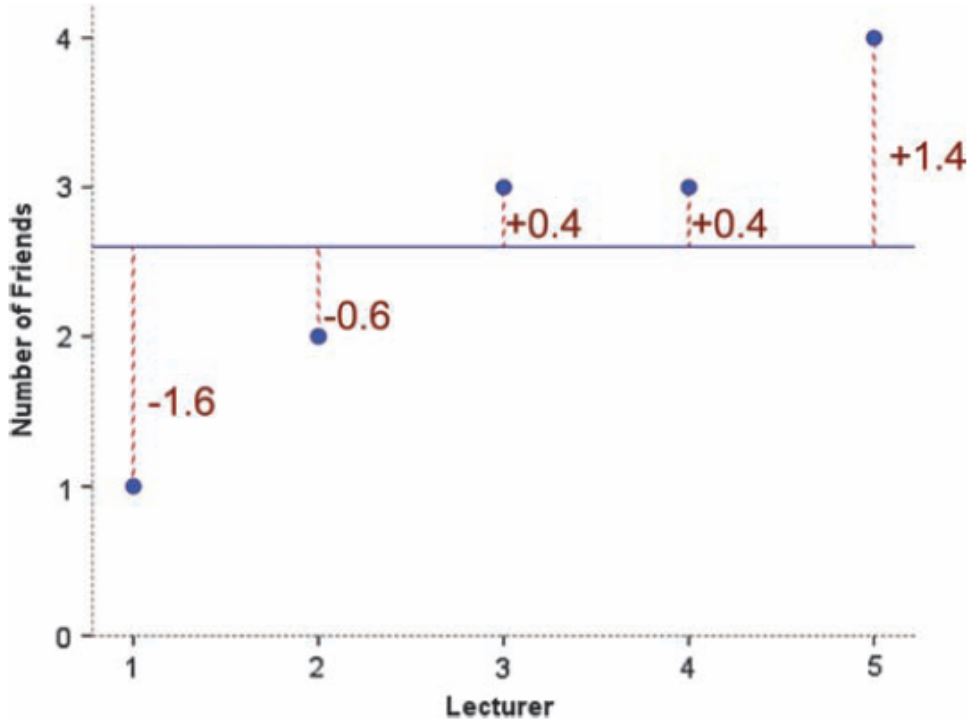
Calculating 'Error'

- A deviation is the difference between the mean and an actual data point.
- Deviations can be calculated by taking each score and subtracting the mean from it:

$$deviation = x_i - \overline{x}$$

Difference

FIGURE 2.4
Graph showing the difference between the observed number of friends that each statistics lecturer had, and the mean number of friends



Use the Total Error?

We could just take the error between the mean and the data and add them.

Score	Mean	Deviation
1	2.6	-1.6
2	2.6	-0.6
3	2.6	0.4
3	2.6	0.4
4	2.6	1.4
		Total = 0

Is this a good solution?

The formula for that would be: $\sum(X - \overline{X}) = 0$

Sum of Squared Errors

- We could add the deviations to find out the total error.
- Deviations cancel out because some are positive and others negative.
- Therefore, we square each deviation.
- If we add these squared deviations we get the **sum of squared errors (SS)**.

Sum of Squared Errors

Score	Mean	Deviation	Squared Deviation
1	2.6	-1.6	2.56
2	2.6	-0.6	0.36
3	2.6	0.4	0.16
3	2.6	0.4	0.16
4	2.6	1.4	1.96
			Total = 5.2

Seems like a much better solution.

The formula for that would be: $SS = \sum(X - \overline{X})^2 = 5.2$

Variance

- The sum of squares is a good measure of overall variability, but is dependent on the number of scores.
- We calculate the average variability by dividing by the number of scores (n).
- This value is called the variance (s^2).

Examples

$$variance(s^2) = \frac{SS}{N-1} = \frac{\sum(x_i - \bar{x})^2}{N-1} = \frac{5.2}{4} = 1.3$$

Standard Deviation

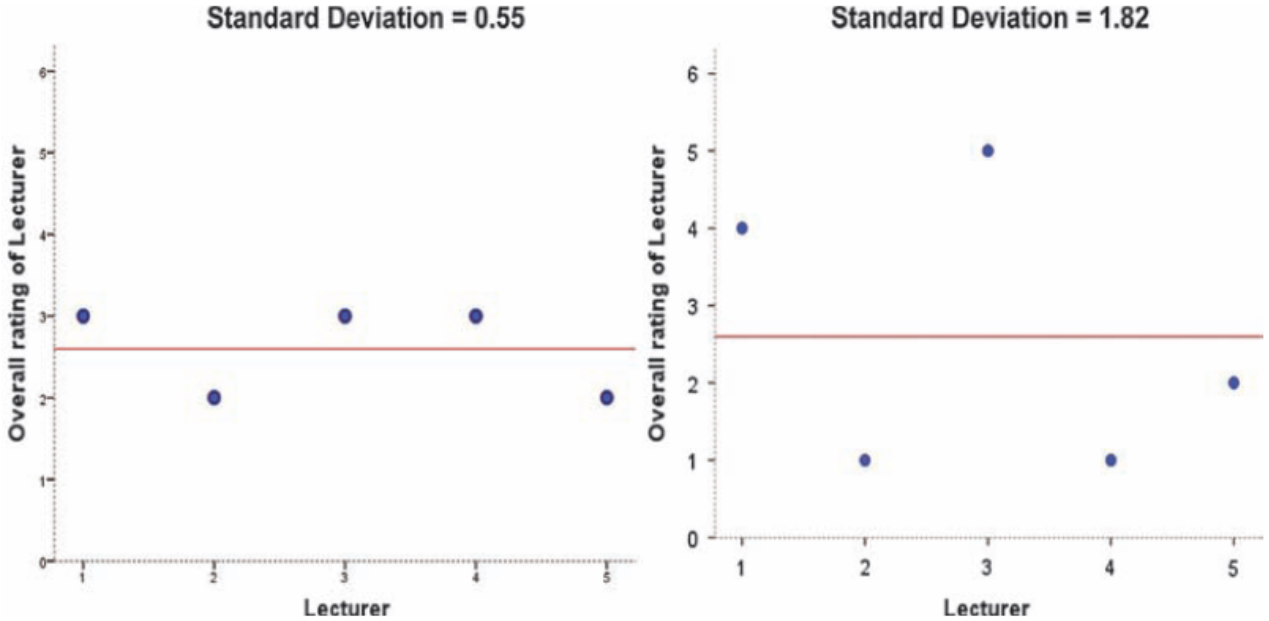
- The variance has one problem: it is measured in units squared.
- This isn't a very meaningful metric so we take the square root value (measured in units).
- This is the standard deviation (s).

Examples

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\frac{5.2}{5}} = 1.02$$

Same Mean, Different SD

FIGURE 2.5
Graphs illustrating data that have the same mean but different standard deviations



The SD and the Shape of a Distribution

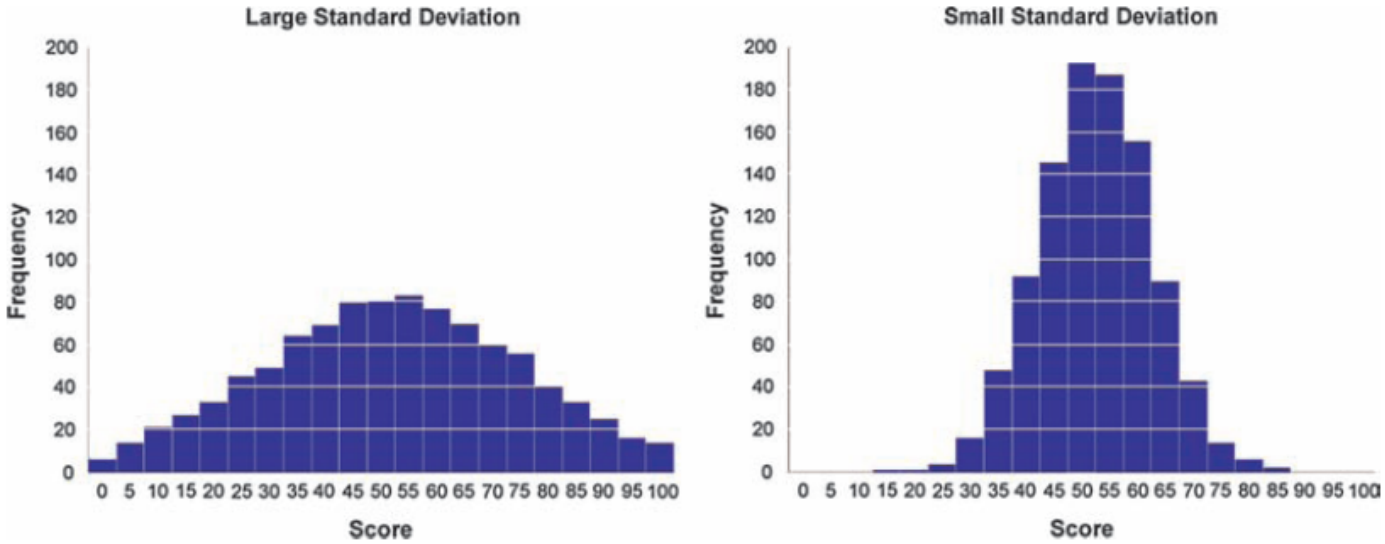


FIGURE 2.6 Two distributions with the same mean, but large and small standard deviations

So what is the mean a model of?

- We have used it to model a summary of a set of data.
- The standard deviation in this case represents how good a 'fit' that model is to the set of data.
- So we are assessing the fit of the model by comparing the data we have to the model we've 'fittted' to the data.
- This is a fundamental idea within the linear statistical model

Important Things to Remember

The sum of squares, variance, and standard deviation represent the same thing:

- The 'fit' of the mean to the data.
- The variability in the data.
- How well the mean represents the observed data.
- Error.

Samples vs. Populations

- Sample
  - Mean and SD describe only the sample from which they were calculated.
- Population
  - Mean and SD are intended to describe the entire population (very rare in most studies).
- Sample to Population:
  - Mean and SD are obtained from a sample, but are used to estimate the mean and SD of the population (very common).

Going beyond the data

- We now know how to fit a simple model to our data.
- But usually we want to move beyond our data to the wider world the data represents and say something about the world.
  - Based on our sample.
- So we need to look at whether the model is a good fit for the population from which it came.

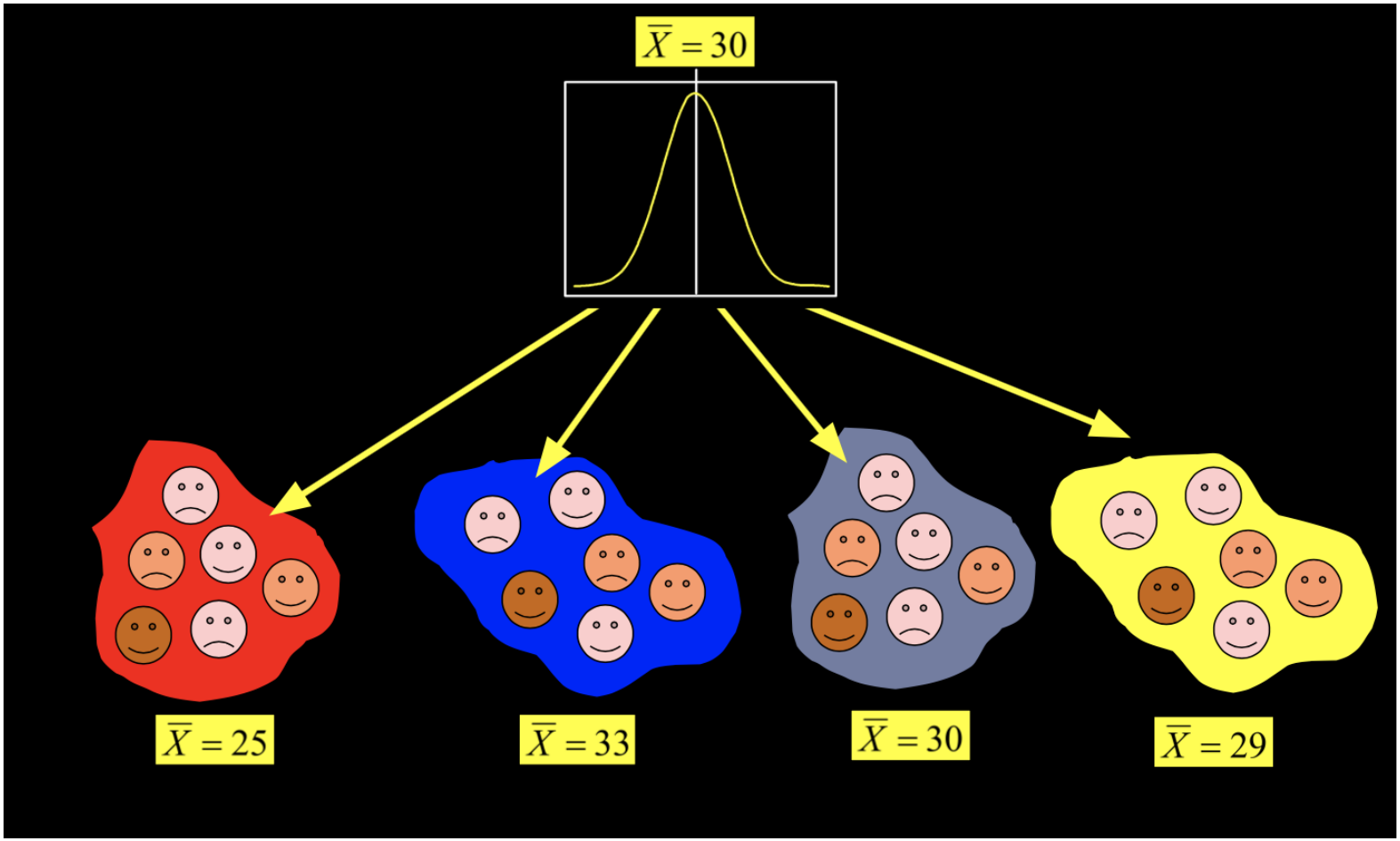
Going beyond the data

- We ideally want to collect data from all members of the population.
- We usually collect a number of samples.
  - Each sample could have a different mean - **sampling variation**.
- We can plot the sample means into a frequency distribution.
  - **Sample distribution**.

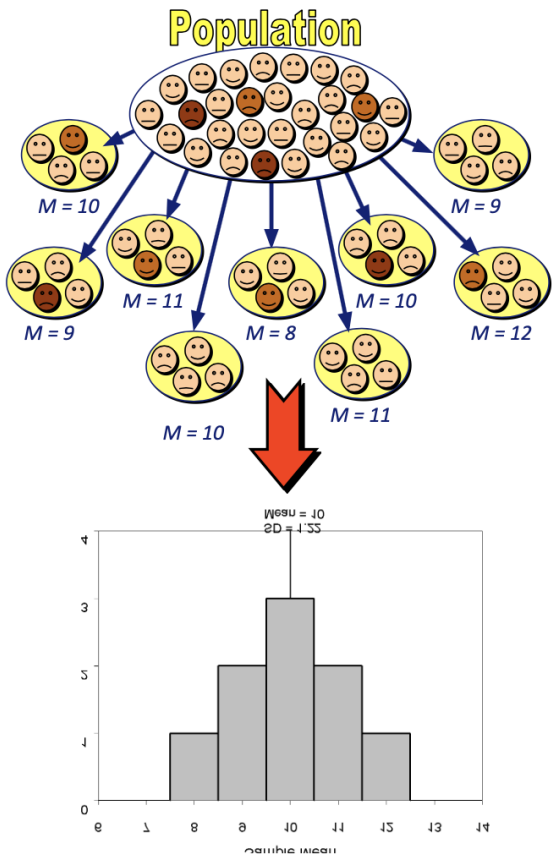
Going beyond the data

- So what?
  - If we have enough samples we can calculate the population mean.
  - But how well does it fit?
- Need to calculate the standard deviation of the sample means.
  - Standard error of the mean (SE).

Example


$\overline{X} = 30$
$\overline{X} = 25$
$\overline{X} = 33$
$\overline{X} = 30$
$\overline{X} = 29$

Example



$$\sigma_{\overline{X}} = \frac{S}{\sqrt{N}}$$

Going beyond the data

- In reality we can't collect enough samples.
- Instead we rely on an approximation of the sample mean and sample error.
- Based on the **Central Limit Theorem**.
  - As samples get large, the sampling distribution has a normal distribution with a sample mean equal to the population mean and a standard deviation of

$$\sigma_{\overline{X}} = \frac{S}{\sqrt{N}}$$

So what does this mean?

We can use the standard deviation of the sampling distribution as the approximation of the sample error.

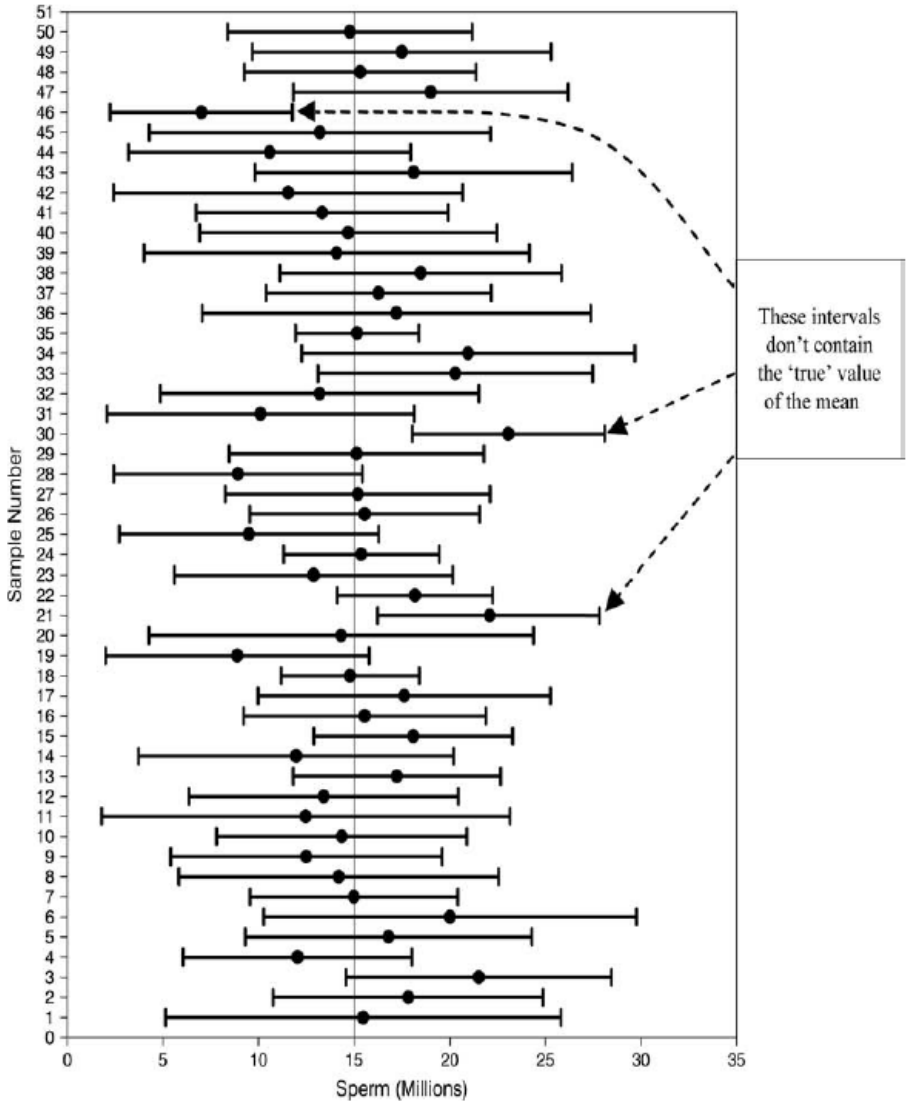
- If our distribution follows the normal distribution.
- For other shapes of distribution we have other ways of approximating the population mean and standard error.

Confidence Intervals

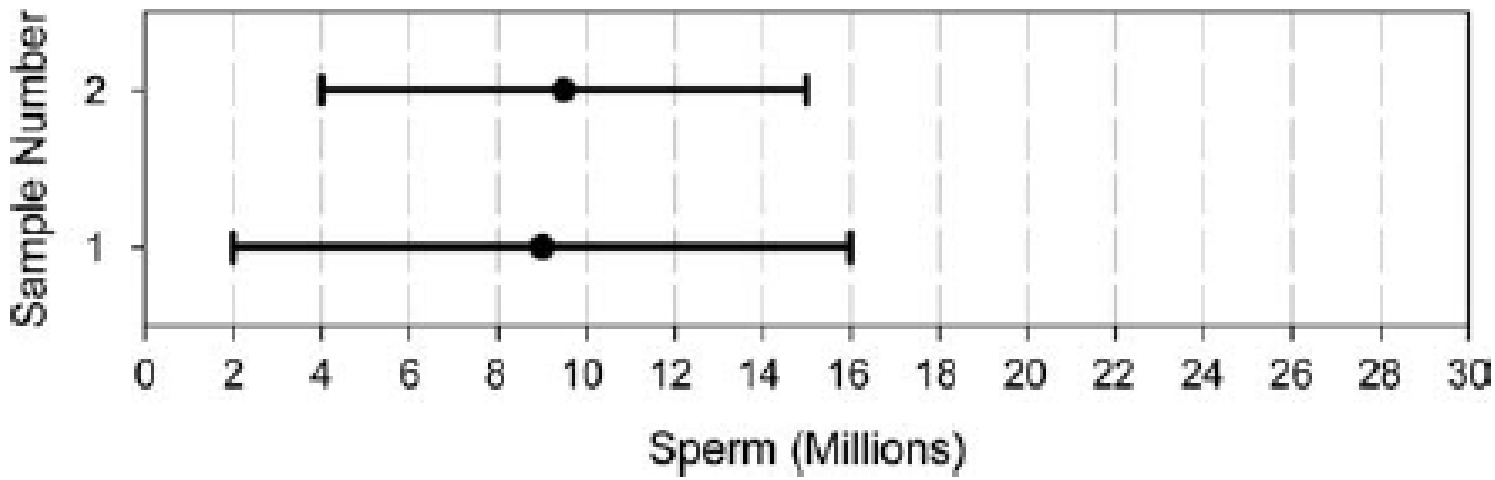
- CI represents a range of values between which we think a population value will fall.
- Suppose we are looking at our fish in Lough Mask.
- True mean: 15 thousand fish.
- Sample mean: 17 thousand fish.
- Interval estimate:
  - 12 to 22 thousand(contains true value).
  - 16 to 18 thousand (misses true value).
  - CIs constructed such that 95% contain the true value.

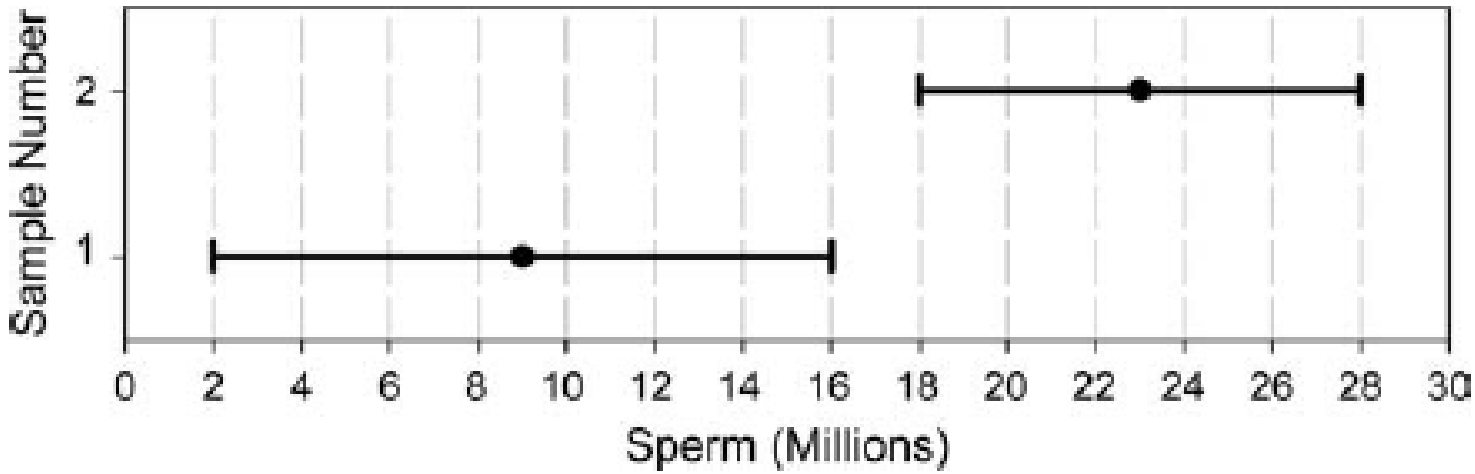
Confidence Intervals

FIGURE 2.8
The confidence intervals of the sperm counts of Japanese quail (horizontal axis) for 50 different samples (vertical axis)



Confidence Intervals





How to construct a CI?

- Typically look at 95% CI but can also look at 99%.
- What does this mean?
  - If we say CI is 95% then if we collected 100 samples, calculated the mean.
  - Then a CI of 95% means we are confident that 95 of these would contain the true mean.
- How to calculate?
  - Need to know the limits within which 95% of the means fall.
  - Go back to the normal distribution: 95% of scores fall between +-1.96.
  - Once we know the mean and standard deviation we can calculate any score and therefore the CI.

CI

- Lower boundary: $\overline{X} = 1 - SE$
- Higher boundary: $\overline{X} = 1 + SE$
- $\overline{X}$ is population mean.
- SE is standard error of the mean.
- And we can use our approximations for this also.

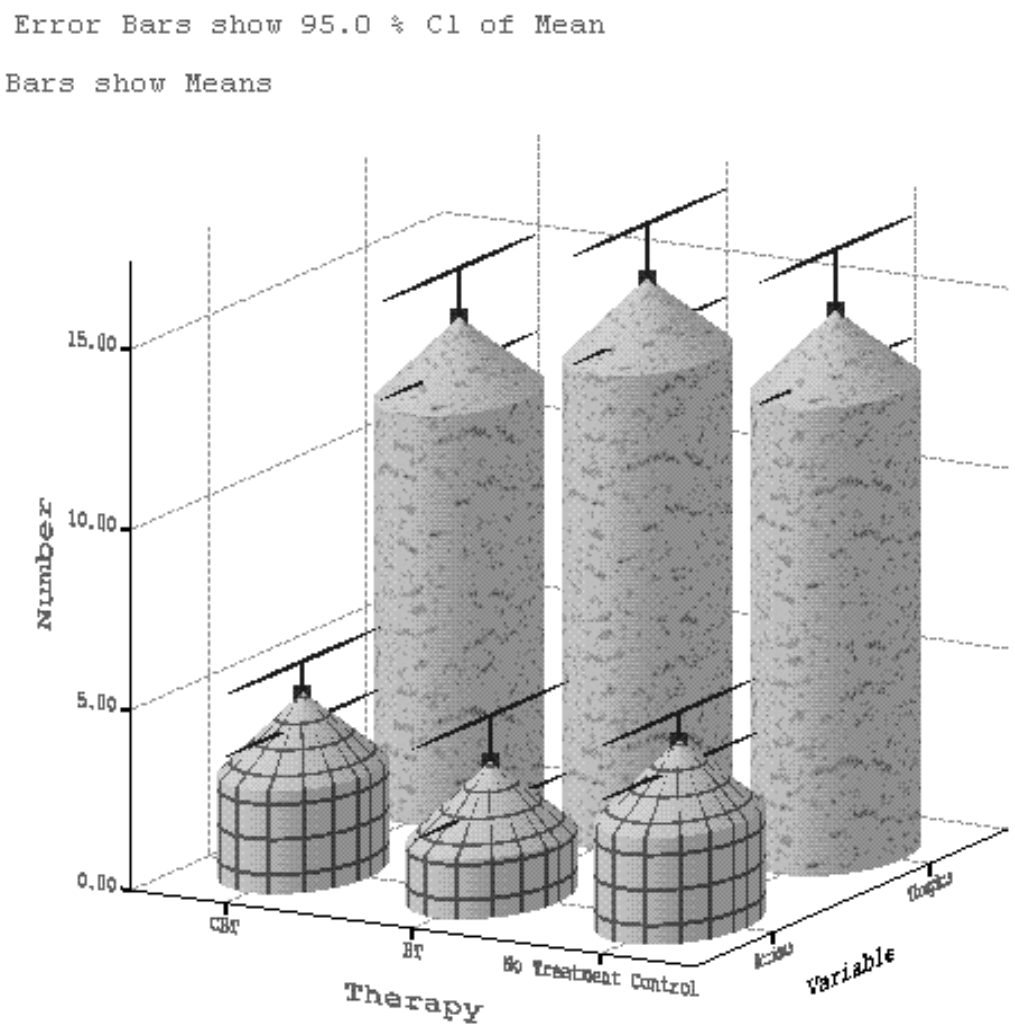
The Art of Presenting Data

Graphs should (Tufte, 2001):

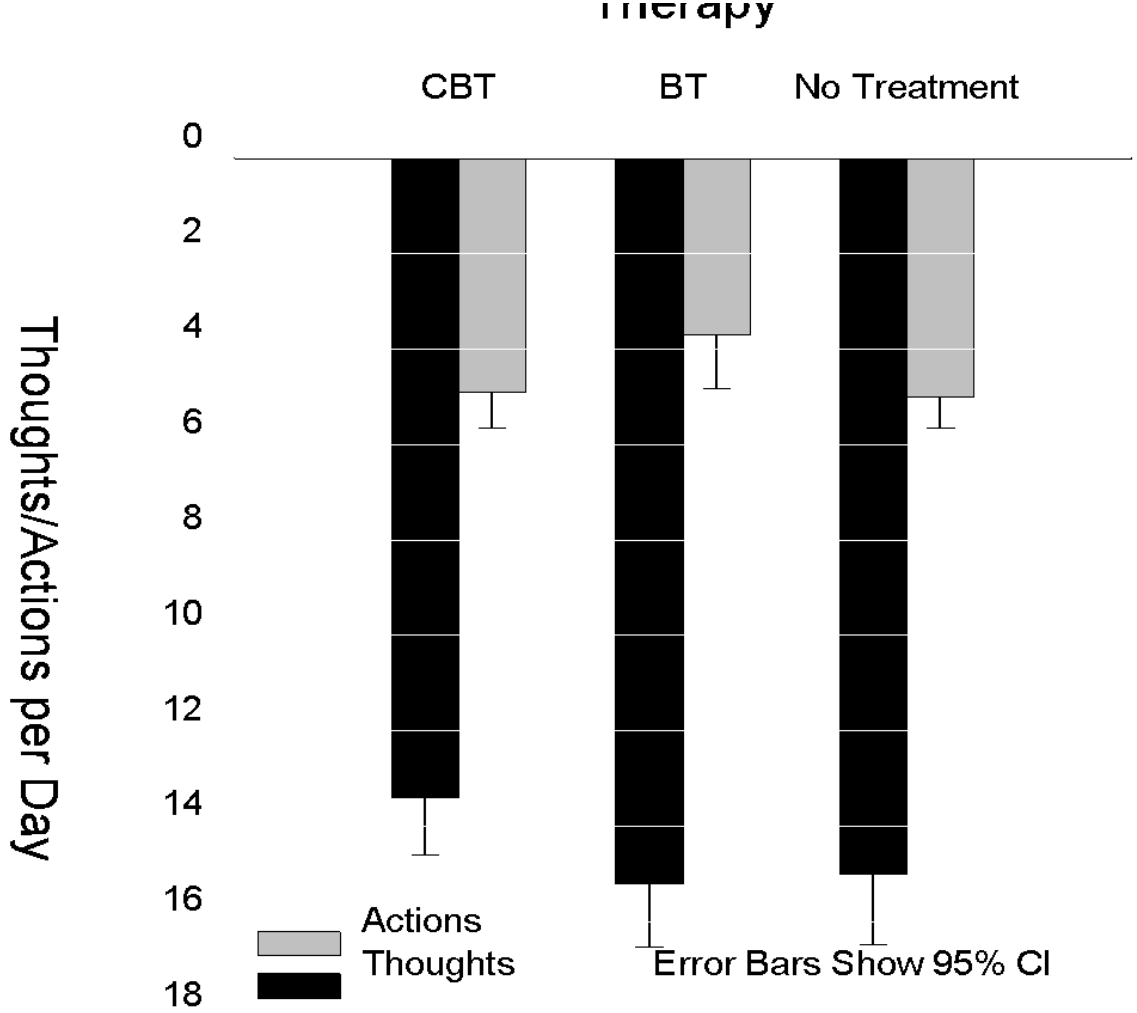
- Show the data.
- Induce the reader to think about the data being presented (rather than some other aspect of the graph).
- Avoid distorting the data.
- Present many numbers with minimum ink.
- Make large data sets (assuming you have one) coherent.
- Encourage the reader to compare different pieces of data.
- Reveal data.

Tufte(2001) Edward Tufte, The Visual Display of Quantitative Information, Graphics Press, 2^{nd} edition, 2001.

Why is this graph bad?



Why is this graph better?



Be careful with your scales ...

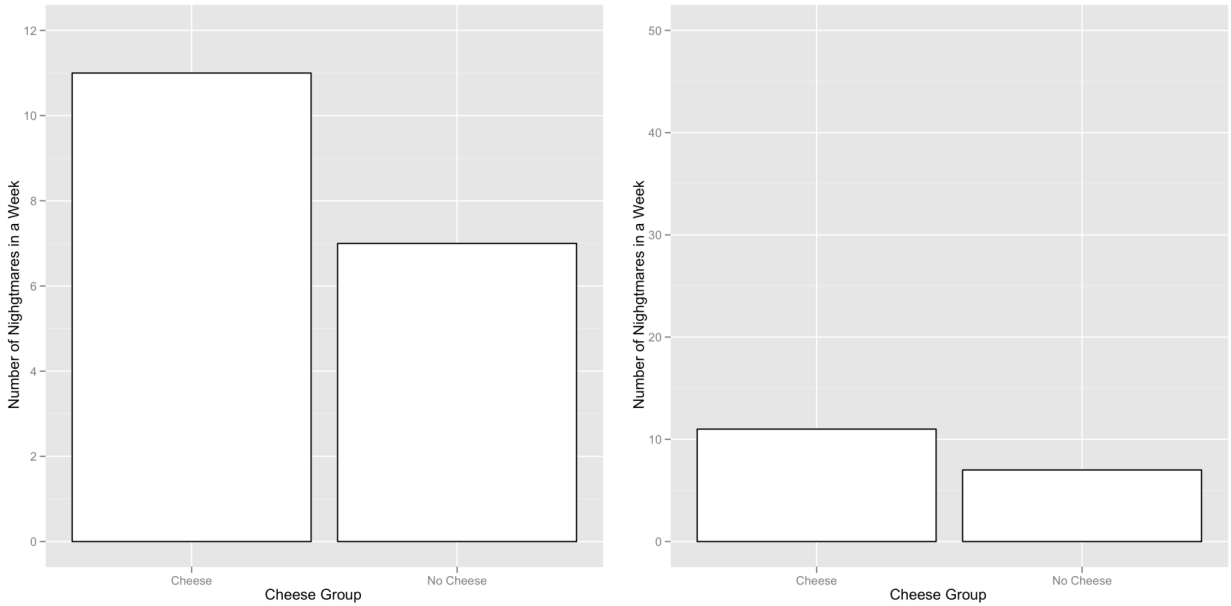


Figure: Two graphs about cheese.

Graphical Presentation

- Nominal or Ordinal:
  - Bar charts, pie charts or frequency tables.
- Interval or Ratio numerical:
  - Histogram, stem and leaf diagrams or box plot (depending on dispersion).

Stem and Leaf Plot

- Shows data arranged by place value.
- You can use a stem-and-leaf plot when you want to display data in an organized way that allows you to see each value.
- Use for small to moderate sized data sets. Doesn't work well for large data sets.
- Accompany with a comment on the centre, spread, and shape of the distribution and if there are any unusual features.

Create Stem and Leaf Plots

Use the data in the table to make a stem and leaf plot.

Test Scores				
75	86	83	91	94
88	84	99	79	86

1. Group the data by ten digits.
2. Order the data from least to greatest.

75 79
83 84 86 86 88
91 94 99

Creating Stem and Leaf Plots

3. List the tens digits of the data in order from least to greatest. Write these in the "stems" column.
4. For each tens digit, record the ones digits of each data value in order from least to greatest. Write these in the "leaves" column.
5. Title the graph and add a key.

75 79
83 84 86 86 88
91 94 99

Test Scores

Stems	Leaves
7	5 9
8	3 4 6 6 8
9	1 4 9

Key 7 | 5 means 75.

Reading Stem and Leaf Plots

Find the least value, greatest value, mean, median, mode, and range of the data.

The least stem and least leaf give the least value, 40.

The greatest stem and greatest leaf give the greatest value, 94.

Use the data values to find the mean $\frac{(40+\ldots+94)}{23} = 64$.

Stems	Leaves
4	0 0 1 5 7
5	1 1 2 4
6	3 3 3 5 9 9
7	0 4 4
8	3 6 7
9	1 4

Key: 4 | 0 means 40.

Reading Stem and Leaf Plots

The median is the middle value in the table, 63.

To find the mode, look for the number that occurs most often in a row of leaves. Then identify its stem. The mode is 63.

The range is the difference between the greatest and the least value: 94-40=54.

Stems	Leaves
4	0 0 1 5 7
5	1 1 2 4
6	3 3 3 5 9 9
7	0 4 4
8	3 6 7
9	1 4

Key: 4 | 0 means 40.

Histograms

- When to Use:
  - Univariate (single variable) numerical data.
- Discrete data:
  - May only take on a finite number of values or countable number of values.
  - Draw a horizontal scale and mark it with the possible values for the variable.
  - Draw a vertical scale and mark it with frequency or relative frequency.
  - Above each possible value, draw a rectangle centred at that value with a height corresponding to its frequency or relative frequency.
- To describe:
  - Comment on the centre, spread, and shape of the distribution and if there are any unusual features.

Example

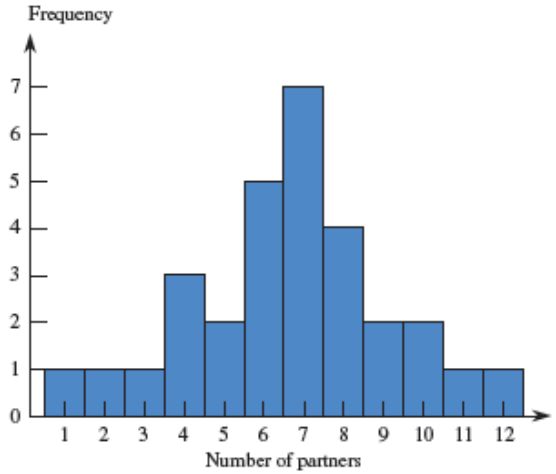
Queen honey bees mate shortly after they become adults. During a mating flight, the queen usually takes several partners, collecting sperm that she will store and use throughout the rest of her life. A study on honey bees provided the following data on the number of partners for 30 queen bees.

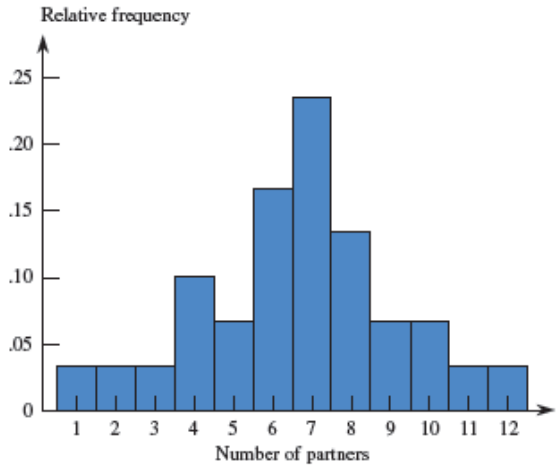
12 2 4 6 6 7 8 7 8 11
8 3 5 6 7 10 1 9 7 6
9 7 5 4 7 4 6 7 8 10

Task

Create a histogram for the number of partners of the queen bees.

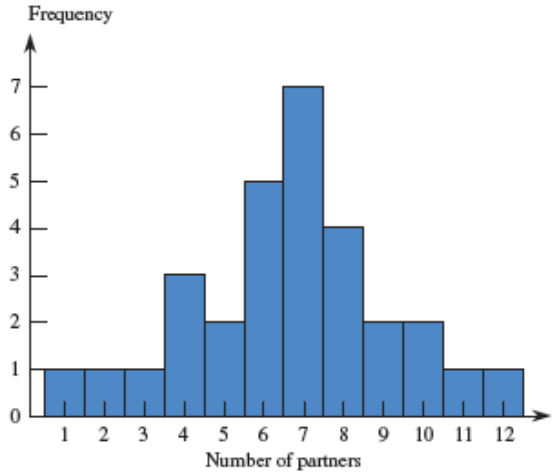
Example

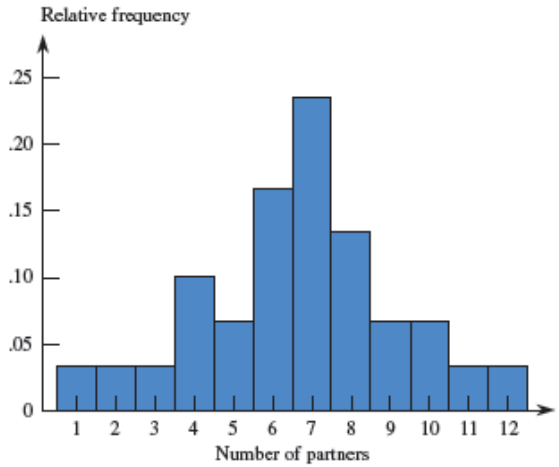




1. First draw a horizontal axis, scaled with the possible values of the variable of interest.

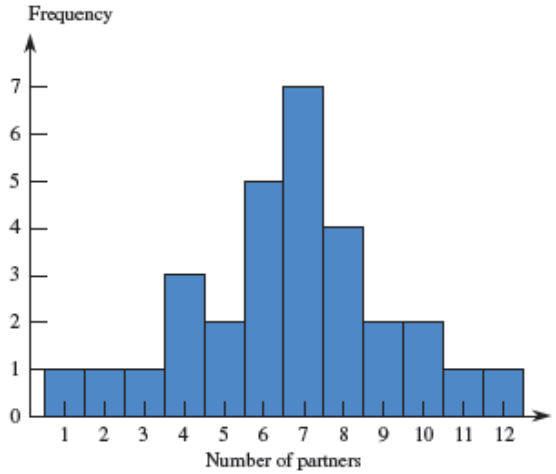
Example

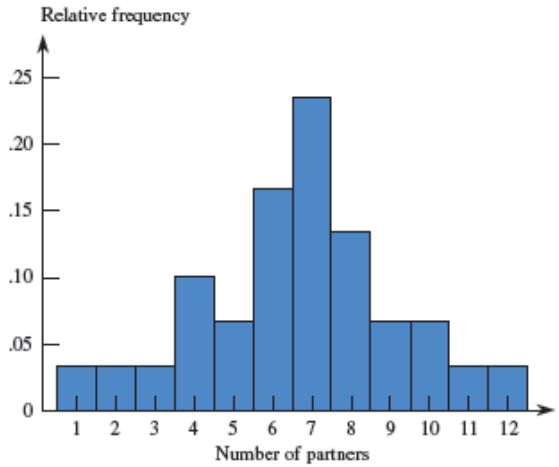




1. First draw a horizontal axis, scaled with the possible values of the variable of interest.
2. Next draw a vertical axis, scaled with frequency or relative frequency.

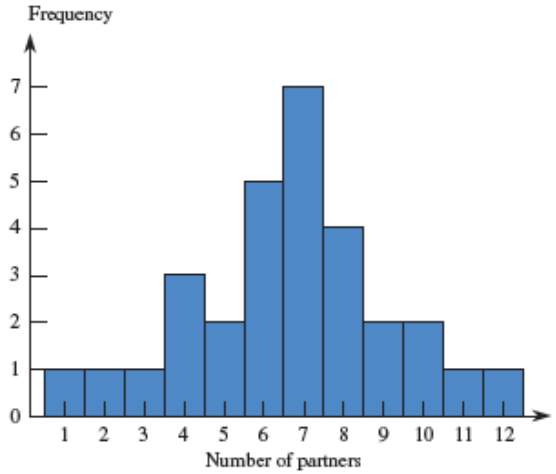
Example

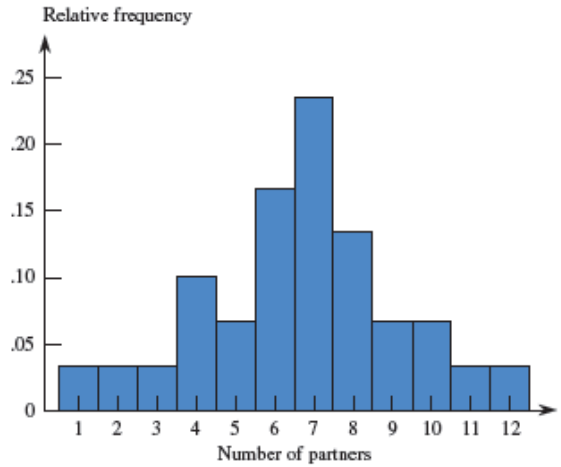




1. First draw a horizontal axis, scaled with the possible values of the variable of interest.
2. Next draw a vertical axis, scaled with frequency or relative frequency.
3. Draw a rectangle above each value with a height corresponding to the frequency.

Example





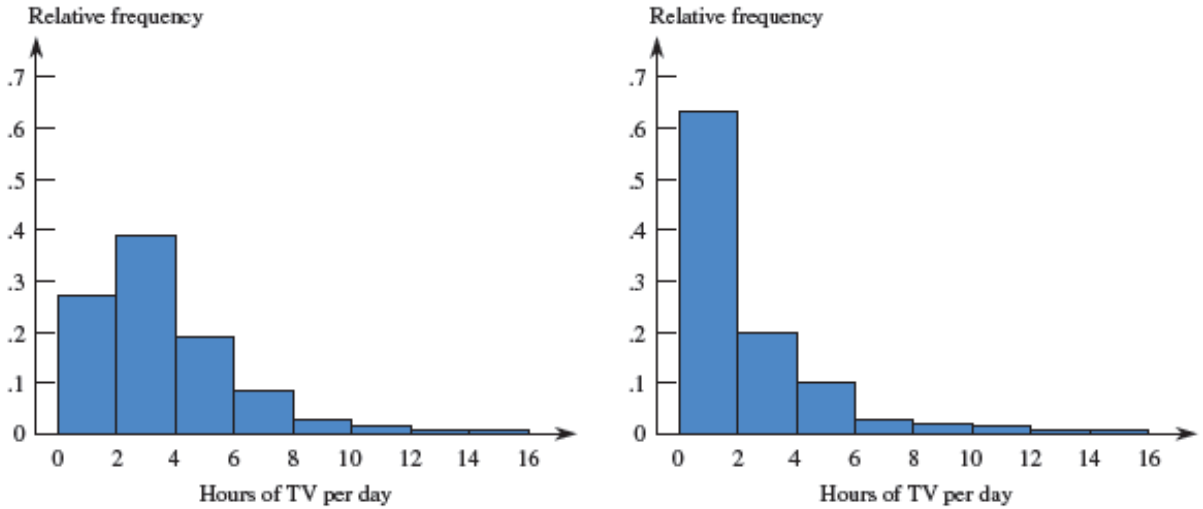
1. First draw a horizontal axis, scaled with the possible values of the variable of interest.
2. Next draw a vertical axis, scaled with frequency or relative frequency.
3. Draw a rectangle above each value with a height corresponding to the frequency.
4. Suppose we use relative frequency instead of frequency on the y axis.

Histograms

- When to Use:
  - Univariate numerical data (one variable).
- Continuous data:
  - Mark the boundaries of the class intervals on the horizontal axis.
  - Draw a vertical scale and mark it with frequency or relative frequency.
  - Draw a rectangle directly above each class interval with a height corresponding to its frequency or relative frequency.
- To describe:
  - Comment on the centre spread, and shape of the distribution and if there are any unusual features.

Example

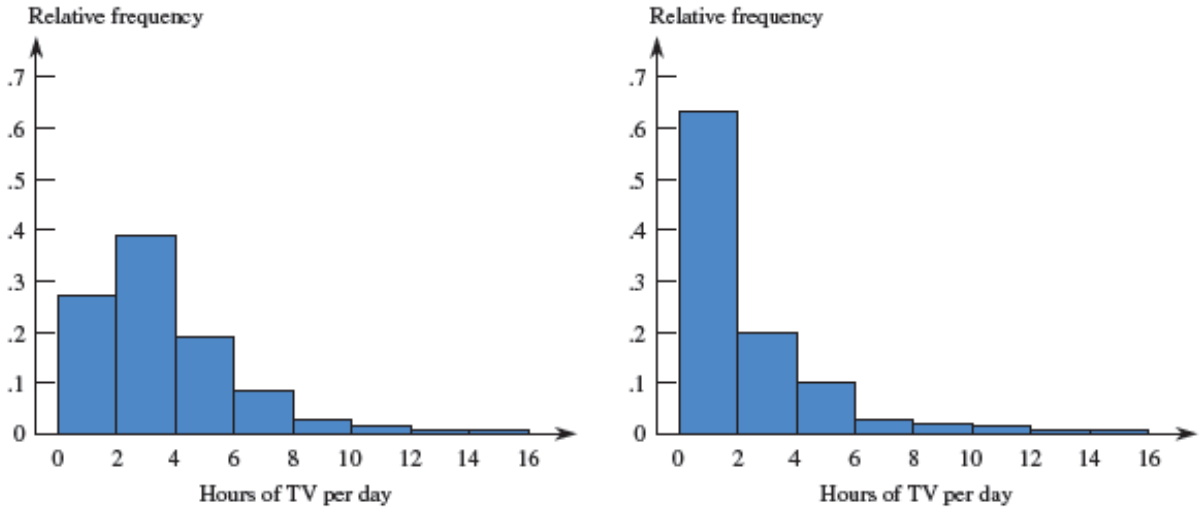
A study examined the length of hours spent watching TV per day for a sample of children age 1 and for a sample of children age 3. Below are comparative histograms.



1. Notice the common scale on the horizontal axis.

Example

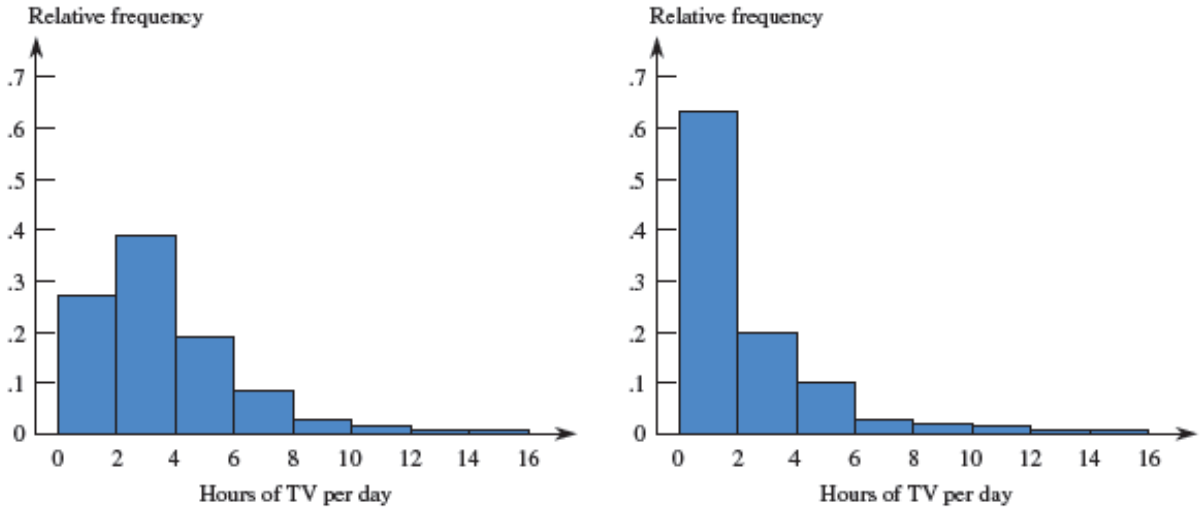
A study examined the length of hours spent watching TV per day for a sample of children age 1 and for a sample of children age 3. Below are comparative histograms.



1. Notice the common scale on the horizontal axis.
2. Write a few sentences comparing the distributions.

Example

A study examined the length of hours spent watching TV per day for a sample of children age 1 and for a sample of children age 3. Below are comparative histograms.



1. Notice the common scale on the horizontal axis.
2. Write a few sentences comparing the distributions.
3. The median number of hours spent watching TV per day was greater for the 1-year-olds than for the 3-year-olds. The distribution for the 3-year-olds was more strongly skewed right than the distribution for the 1-year-olds, but the two

Frequency Distribution

- Graphs are useful in assisting us in assessing the distribution in a set of data for a particular variable.
- Frequency distribution shows the relative frequencies of values for variables of interest in a dataset.
  - Where values have been binned into groups (e.g. 10 to 20, 21 to 30 etc).
  - The height of each bar is proportional to the relative frequency in the data set of the group it represents.
- Normal distribution:
  - Bell-shaped, scores equally distributed around a central value (mean).
- Skewed:
  - Lack of symmetry.
  - Data pulled towards one end of the graph.
- Kurtosis:
  - Pointyness.

15mins Break

Check Your Progress

- Do the following self assessment:
  `https://www.med.soton.ac.uk/stats_eLearning/level-quiz-basic-summary-statistics.html`
- If you have any questions use the chat.
- Feel free to discuss your progress or how you felt answering the questions.
- Time: 30mins

Assessment 1

- In Brightspace select the module and click **Assessment** → **Assignments**.
- Click **Assignment 1**.
- Read the description carefully and look at the attachments to find the problems and data.
- The solution should be one R or Rmd file with comments and markup addressing all questions.

Plots in R

ggplot2

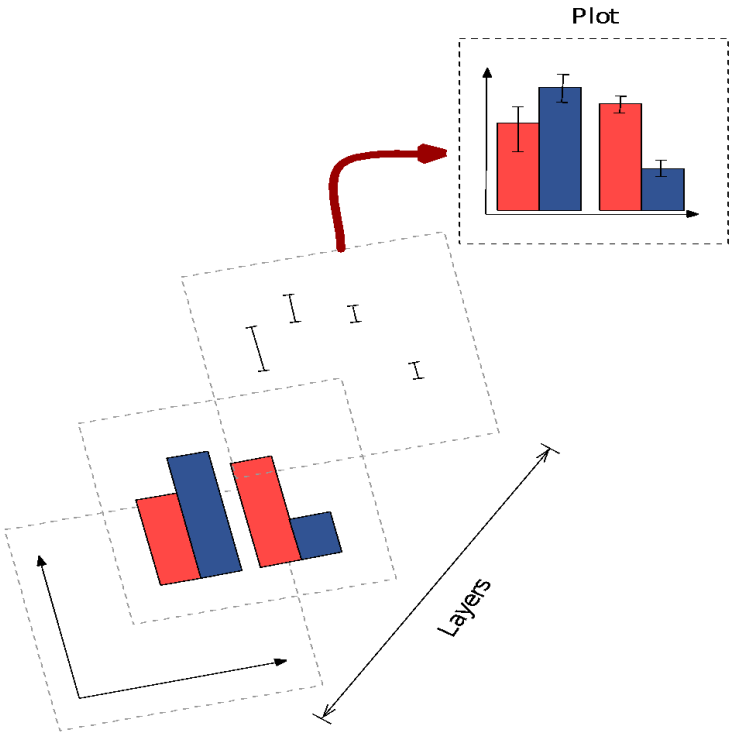


Figure: In ggplot2 a plot is made up of layers.

Ggplot Components

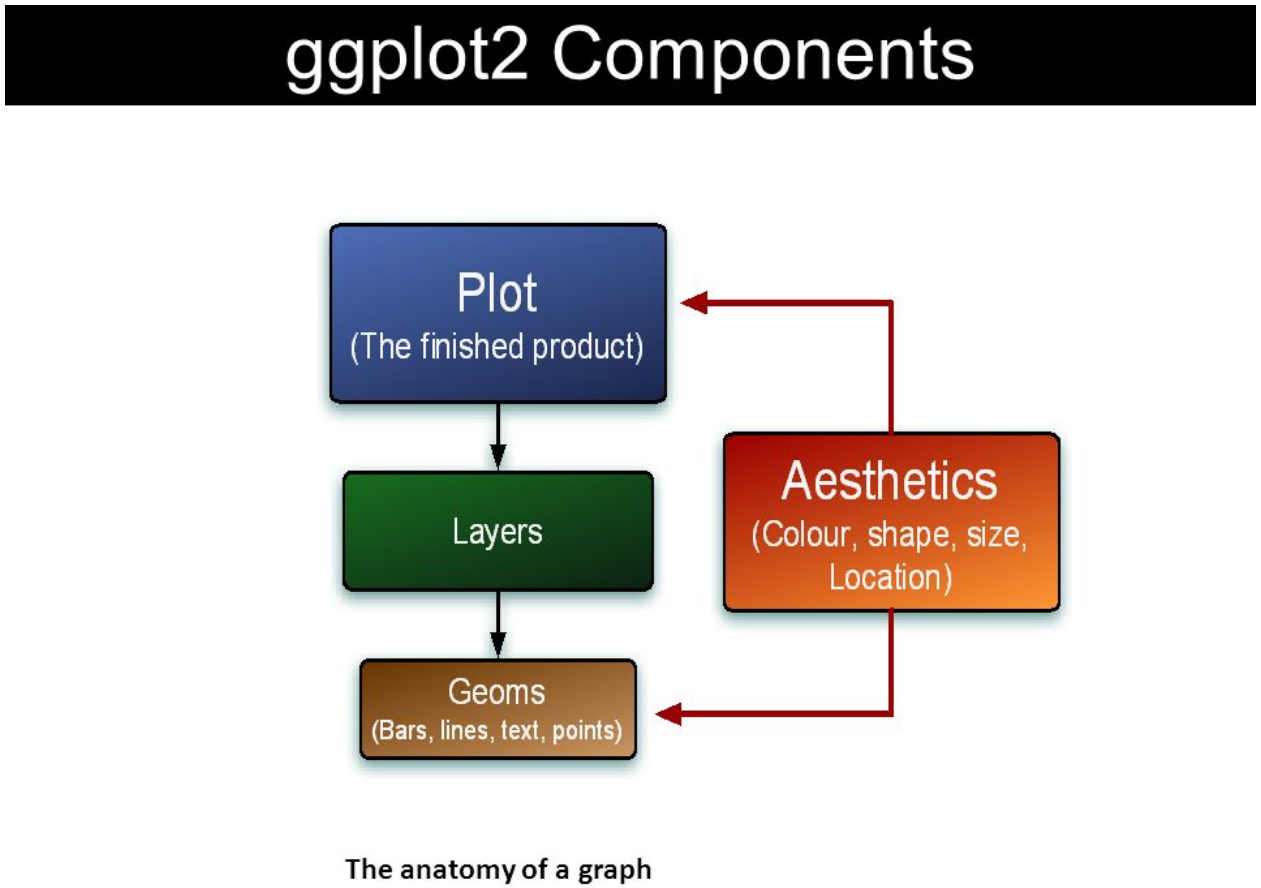


Figure: The anatomy of a graph.

Geometric Objects (geoms)

- geom_bar(): creates a layer with bars.
- geom_point(): creates a layer with data points.
- geom_line(): creates a layer with lines.
- geom_histogram(): creates a layer with a histogram.
- geom_boxplot(): creates a layer with a box-whisker diagram.
- geom_text(): creates a layer with text on it.
- geom_density(): creates a layer with a density plot on it.

Geometric Object Functions

- Each takes different parameters.
- See handout for pages from Andy Field re properties associated with common geom + some common aesthetics.

Using ggplot

Create an object that specifies the plot:

1. Pass in the data and set whatever aesthetics you want to apply to all layers (if any).
2. Adjust the layers.
3. Display/save the graph.

Lab Exercise

- Sample code: PSIWeek2.R
- Datasets:
  - facebookNarcissism.dat: Contains data from a study that looked at ratings of Facebook profile pictures (on coolness, fashion, atrractiveness and glamour) and predicting from this how high the person who posted the picture rates on narcissim.
  - DownloadFestival.dat: Data collected from the Download Music Festival. Study measured the hygiene of attendees on the three days of the festival. Not all participants were measured each day so there is some missing data. Gender of the concert-goer is recorded.
  - ChickFlick.dat: Based on a study of 20 men and 20 women. Half the sample watched Bridget Jones Diary, half watched Memento. Measured their interest in the movie. Also stored is their gender.

Lab Exercise

1. We are going to split up in 2 groups per dataset.
2. Every group will analyse the assigned dataset and produce some graphs to find something interesting about the dataset.
3. They will elect a speaker who will present their findings in a short 3-5 min presentation by discussing the created graphs.

Your Mission

Get the datasets loaded, the sample code running, and see if you can come up with some new graphs which show your own findings from the data!