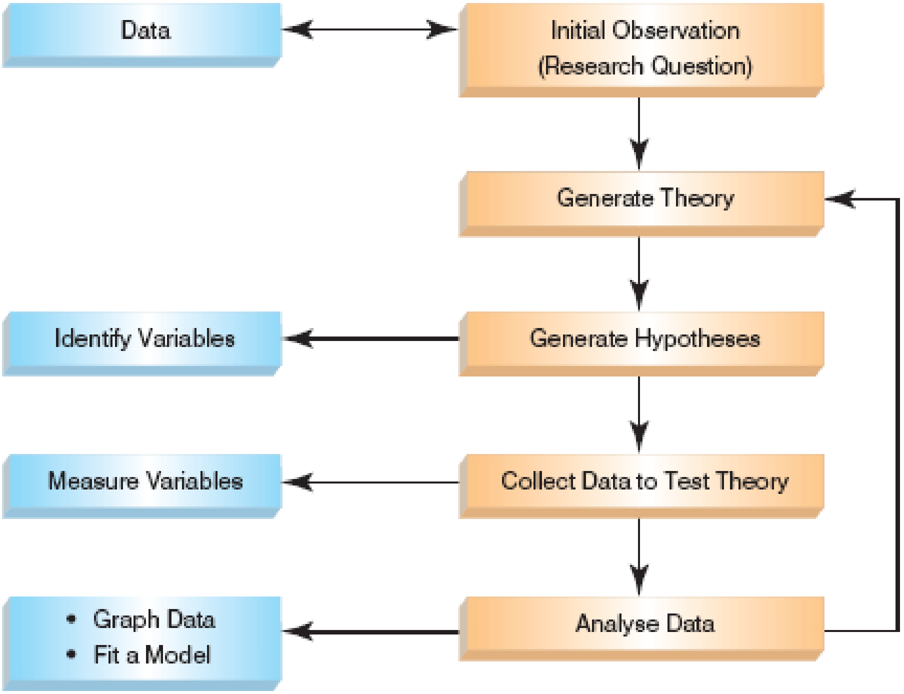
MATH9102 Fundamentals of Data Analysis*

DESCRIBING YOUR DATASET

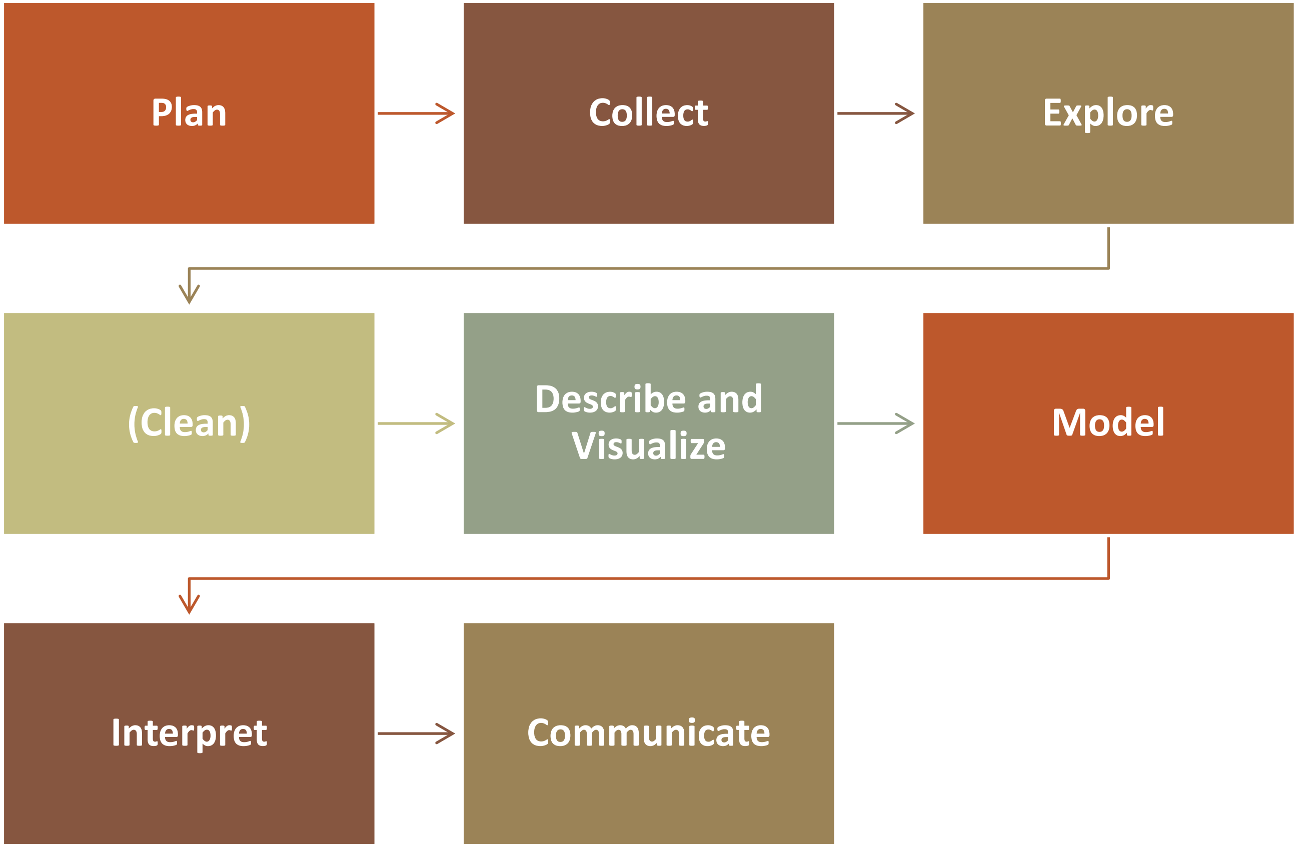
* Formerly Probability and Statistical Inference


Data
Initial Observation (Research Question)
Generate Theory
Identify Variables
Generate Hypotheses
Measure Variables
Collect Data to Test Theory
• Graph Data
• Fit a Model
Analyse Data

The Process

"Now, keep in mind that these numbers are only as accurate as the fictitious data, ludicrous assumptions and wishful thinking they're based upon!"

Data Analysis Pipeline



Plan

Collect

Explore

(Clean)

Describe and Visualize

Model

Interpret

Communicate

Getting Started With Statistical Study Conceptual Framework

Before you start

- You need to decide your research question and the nature of the study
- Then decide what individuals or objects are of interest
    - Conceptual Framework
- You need a detailed understanding of the data needed to conduct the analysis
- Create a careful advance plan of data collection/manipulation
- Choose the correct analytic approach is needed to answer the question under investigation in a scientific way

Getting Started With Statistical Study Population and Sample

Start with understanding the **population**

- All possible cases that meet certain criteria.
- The total collection of cases that you are interested in studying
- Usually constantly changing, difficult to collect

Getting Started With Statistical Study Population and Sample

Work with a **sample** (set of samples)

- Subset or portion of the population
- Substitute for the population
- Representativeness and size are key

Getting started with Statistical Study Experiments

A statistical **experiment** or observation is any process through which measurements are obtained.

- We conduct the experiment and generate data about the person, place, thing, situation or idea which is the subject of our theory.

Getting started with Statistical Study Experiments

An experiment is a procedure we conduct working with an appropriate sample sample with the intention of conducting an analysis of that sample to make some decisions about our question(s) of interest for our population of interest.

Getting started with Statistical Study Experiments

Experiments can be:

- Observational studies (observing and recording data about concepts of interest)
  - Example: A researcher records the number of hours university students spend in the library each week and compares it with their GPA.
  - Key point: No intervention, just observing naturally occurring behaviour.
- Surveys (using questions to elicit data about concepts of interest)
  - Example: Participants complete a wellbeing questionnaire (like survey.dat) where they report on stress levels, optimism, and smoking habits.
  - Key point: Data is elicited directly through structured questions.

Getting started with Statistical Study Experiments

Experiments can be:

- Scientific experiments (under controlled conditions may involve manipulation of some concepts)
  - Example: Researchers randomly assign participants to two groups.
    - Group A listens to relaxing music for 20 minutes.
    - Group B sits in silence for 20 minutes.
  - After the session, both groups complete a validated stress scale questionnaire (empirical measurement of perceived stress), and their heart rate variability is recorded using a monitor (physiological empirical measurement).
  - The goal is to test whether music reduces stress compared to silence.
  - Key point: Researcher manipulates a condition and controls the environment.
- Meta-analytical studies (combining findings of other statistical experiments)
  - Example: A systematic review collects results from 50 published studies examining whether smoking is linked to life satisfaction, then applies statistical methods to combine the findings into one overall effect size.
  - Key point: Integrates and re-analyses data/results from multiple independent studies.

Getting started with Statistical Study Variables

For an experiment we identify concepts/criteria of interest.

- These need to be represented these in our sample
- Data must be collected to represent these concepts

A variable/feature

- Represents a concept of interest in an experiment
  - Describes or quantifies an aspect of a person, place, thing, situation, or idea.
- Takes on a statistical type appropriate for that concept of interest?
  - Why do we care about the type?
    - Different statistical procedures/techniques suit different types of variables

Getting started with Statistical Study Variables

A **variable/feature**

- The value of a variable can "vary" from one entity to another.
- A variable is **random** if the value it takes on in an experiment or observation is determined by chance.
  - Your job is to ensure your design is sufficiently robust that this is what happens when we create our sample.

Selecting variables for your study

Looking for indicators that represent your concepts of interest.

You need to ensure the indicators chosen demonstrate:

- Reliability
  - Degree to which an indicator is a consistent measuring device
  - E.g.
    - Is asking a student how well they did in school a reliable indicator of their ability to learn?
- Validity
  - Extent to which an indicator measures what it is intended to measure
  - E.g.
    - Is a student’s IQ value a valid indicator of their educational achievement?

Variables

Some we **measure indirectly**

- Sometimes it is not possible to measure something directly, so we work with approximations
  - E.g. psychological tests are approximate measures

- Sometimes there will be a difference between values we use to represent a thing we are measuring and the actual value of the thing
- We must always be aware of **Measurement error**

Variables

Some we **control** for

- Recognising that some measurements are influenced by other factors which we may also have measured
  - Confounding variables

Variables

May also be:

- Things we can manipulate to derive a measure for other concepts
  - E.g. Dimension Reduction
- Things we can compute from measures of other concepts
  - E.g. Using a set of variables as a scale or index

Variable Types

Determine the types of measurement that each variable can hold

- Qualitative
- Quantitative

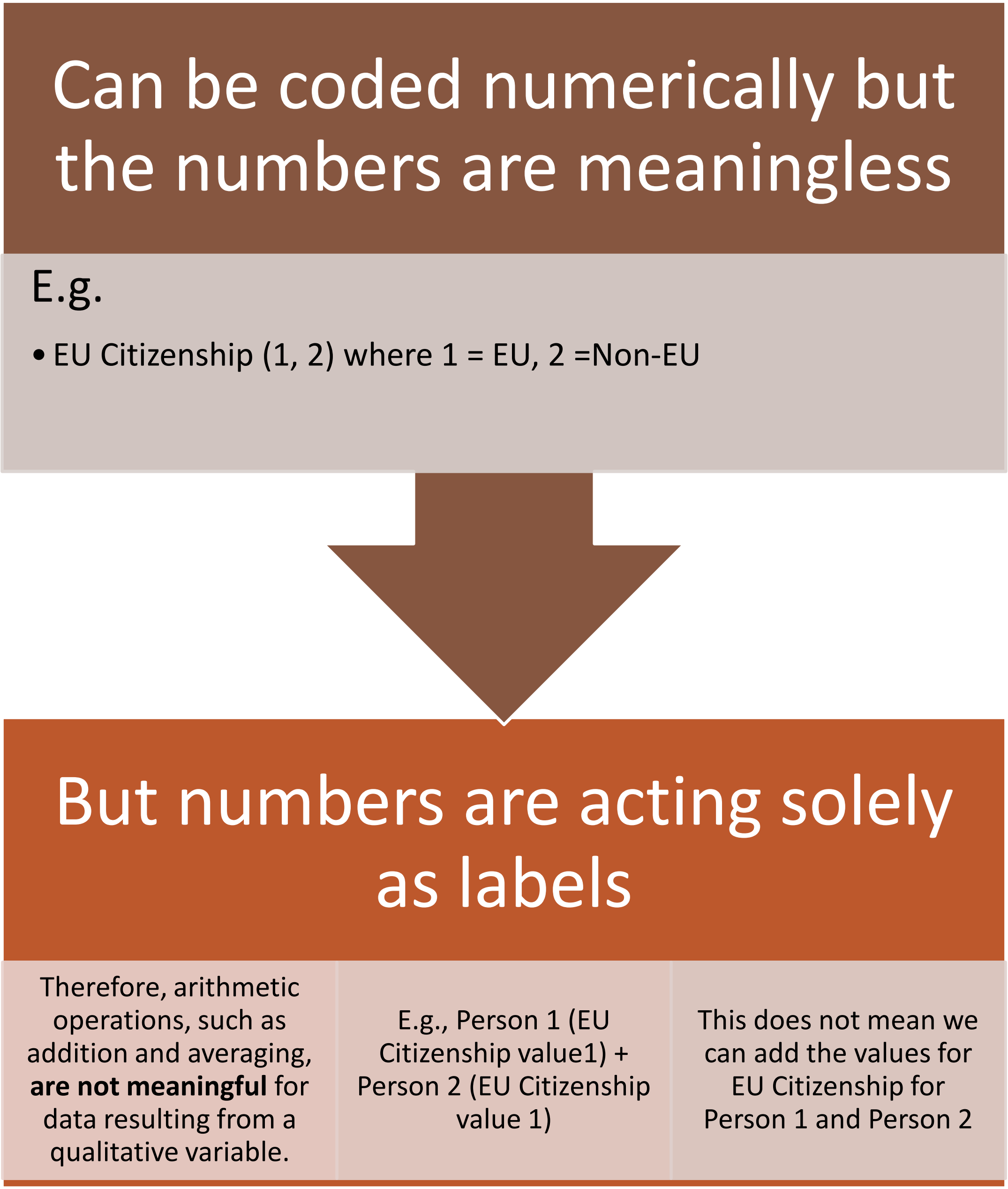
Qualitative (Attribute, or Categorical)

Variable categorizes or describes an element of a population.

Take on values that are names or labels

- E.g.
  - Hair Colour (Black, Light-Brown, Dark-Brown, Gray, Red, Blonde)
  - Creativity (Very Creative, Somewhat Creative, Not Creative)
  - Grade (A, B, C, D, E, F)
  - Grade Point (1, 2, 3, 4)

Qualitative, or Attribute, or Categorical, Variable



Can be coded numerically but the numbers are meaningless

E.g.

- EU Citizenship (1, 2) where 1 = EU, 2 =Non-EU

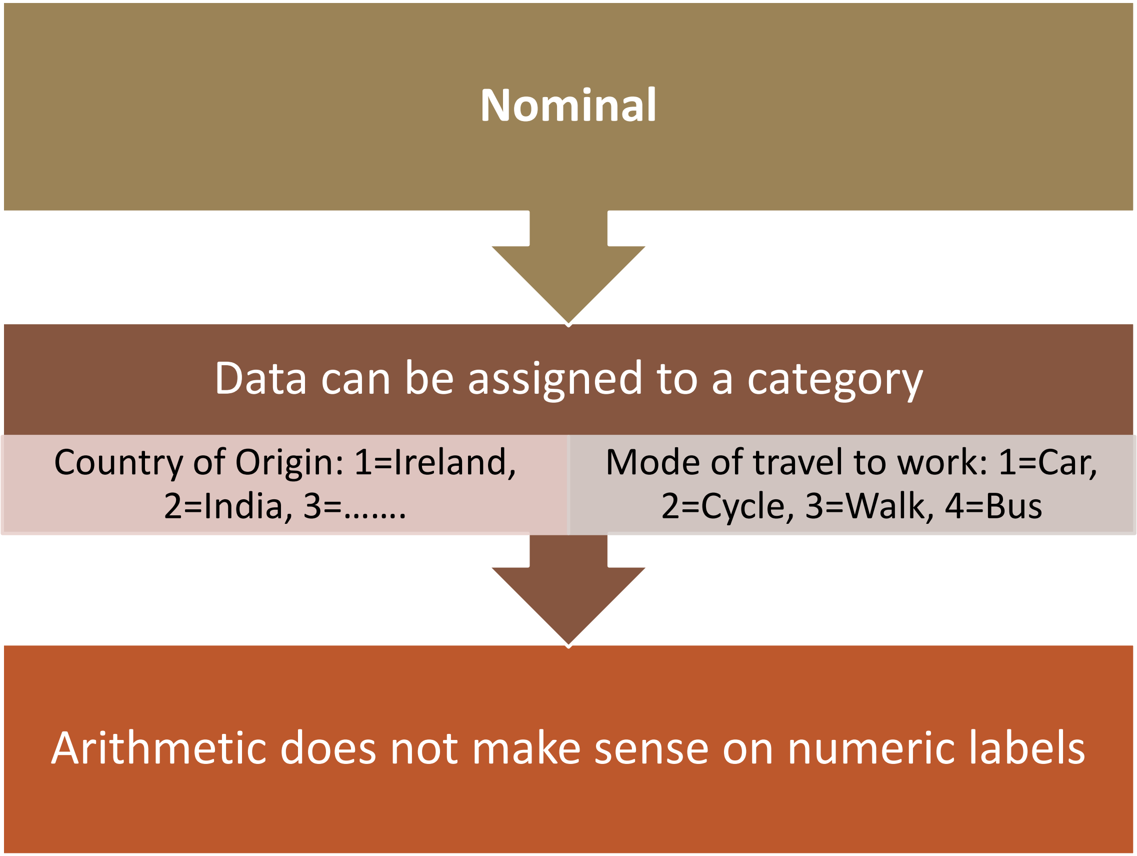
But numbers are acting solely as labels

Therefore, arithmetic operations, such as addition and averaging, **are not meaningful** for data resulting from a qualitative variable.

E.g., Person 1 (EU Citizenship value1) + Person 2 (EU Citizenship value 1)

This does not mean we can add the values for EU Citizenship for Person 1 and Person 2

Qualitative variable - Levels of measurement



Nominal

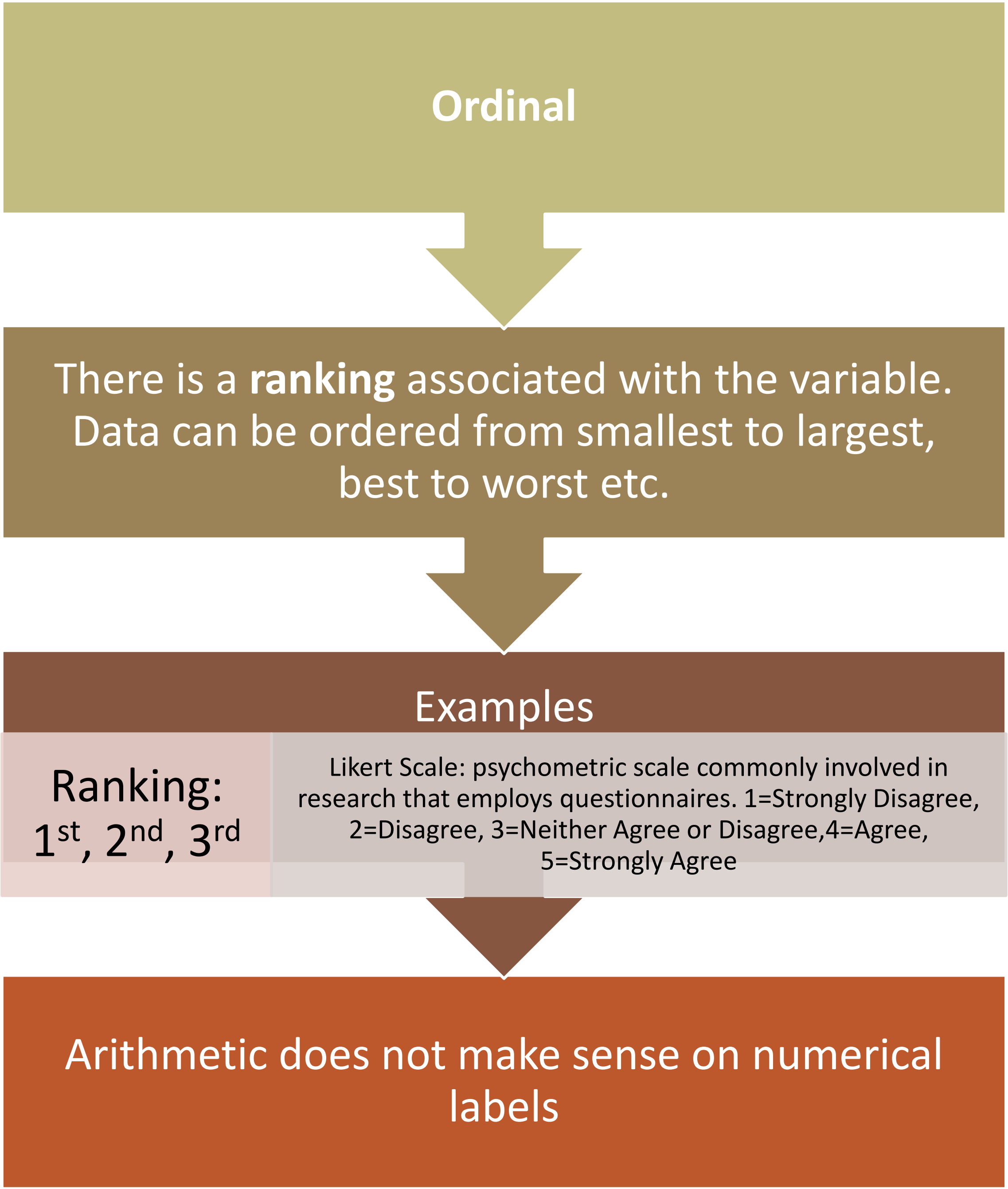
Data can be assigned to a category

Country of Origin: 1=Ireland, 2=India, 3=…….

Mode of travel to work: 1=Car, 2=Cycle, 3=Walk, 4=Bus

Arithmetic does not make sense on numeric labels

Qualitative variable - Levels of measurement



Ordinal

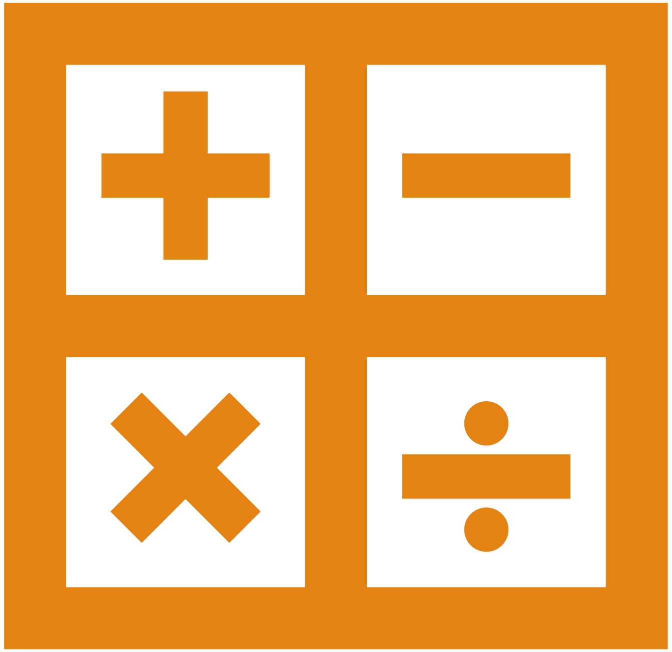
There is a **ranking** associated with the variable. Data can be ordered from smallest to largest, best to worst etc.

Examples

Ranking: 1st, 2nd, 3rd

Likert Scale: psychometric scale commonly involved in research that employs questionnaires. 1=Strongly Disagree, 2=Disagree, 3=Neither Agree or Disagree,4=Agree, 5=Strongly Agree

Arithmetic does not make sense on numerical labels



Quantitative (Numerical Variable)

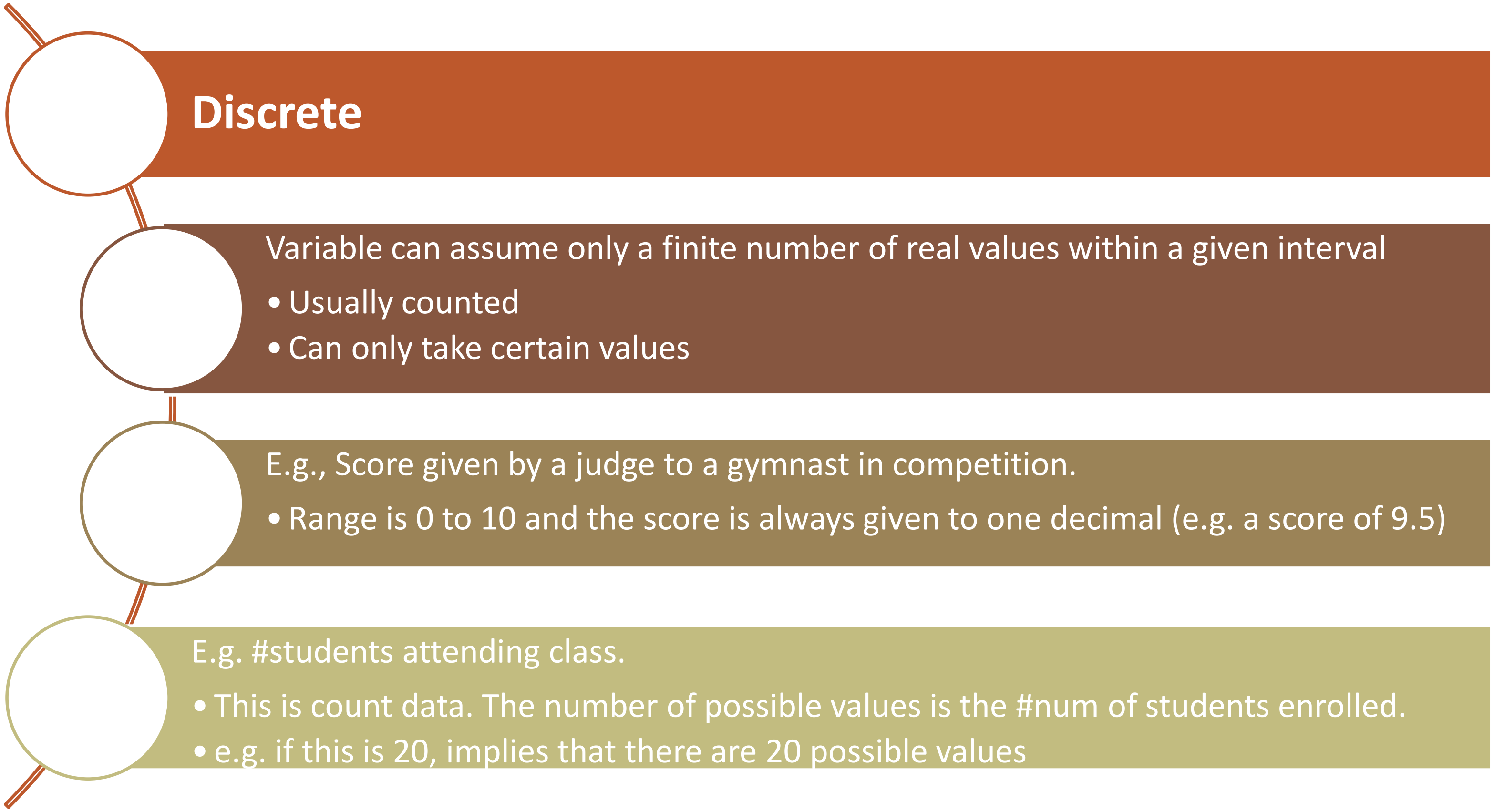
A variable that **quantifies** an element/characteristic of a population.

Observations or measurements take on numerical values

Note:

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

Quantitative, or Numerical, Variable (Broad Types)



Discrete

Variable can assume only a finite number of real values within a given interval

- Usually counted
- Can only take certain values

E.g., Score given by a judge to a gymnast in competition.

- Range is 0 to 10 and the score is always given to one decimal (e.g. a score of 9.5)

E.g. #students attending class.

- This is count data. The number of possible values is the #num of students enrolled.
- e.g. if this is 20, implies that there are 20 possible values

Quantitative, or Numerical, Variable (Broad Types)

Continuous

Variable that can be any value in a given range

Usually measured

E.g. , Heights of students attending class

- Can’t be any values, can’t be negative, can’t be higher than three metres. BUT on the scale between 0 and 3, the number of possible values is theoretically infinite;

E.g., Time spent concentrating in class

- Can’t be any values, can’t be negative, can’t be higher than 180mins. BUT on the scale between 0 and 180, the number of possible values is
- Theoretically infinite.

Quantitative Variable – Levels of Measurement

Interval

Data is ordered but there is meaning between the values of order

- The distance between two points on the scale is standardised and equal
- Allows comparison between the data values
- Can be added and subtracted.

E.g., Temperature (Farenheit), temperature (Celcius).

- The difference between 10° and 20° is quantitatively the same as the difference between 20° and 30°

E.g., IQ Test. No one can have an IQ of zero (according to studies so it is not ratio).

- Therefore, the difference between an IQ of 110 and 115 is quantitatively the same as the difference between an IQ of 115 and 120.

Quantitative Variable -Levels of measurement

Ratio

Has all the properties of interval data

PLUS

- Has an **absolute 0**
- It makes sense to say for example one value is twice as large as another
- Can be added and subtracted, multiplied and divided.

E.g., The difference between 1 and 2 is the same as the difference between 3 and 4

- 4 is also twice as much as 2.

A person’s age – Is it Ratio?

Technically YES.

- Does have a zero and
- (if measured with the required accuracy) is continuous

Practically, IT DEPENDS

- Need to consider the measurement
- Culture?
- Age categories?
- Age as a discrete count (e.g. number of years)?

Variables and Data

Statistical analysis revolves around the “collection, organization, and interpretation of data according to well-defined procedures.”

- Kachigan, S. K. (1991). Multivariate statistical analysis: A conceptual introduction (2nd ed.),New York: Radius.

Data: information about different variables.

Data point: individual piece of information related to one variable.

Dataset: collection and organisation of information relative to specific variables

Data distribution: a listing of the values or responses associated with a particular variable in a data set.

Frequency distribution: a table or graph that indicates how many times a value appears in a dataset of values.

Getting Started With Statistical Study

We must be able to describe our sample

- Be able to describe it simply
  - Using the appropriate summary statistics for the variables of interest
- Be able to represent it visually
  - Using the appropriate type of graph/plot for variables of interest

Then we will be able to analyse it appropriately

- Using the appropriate statistical tests to draw inference

Important Step - Describing your data

Must describe before analyzing

Need to include sufficient detail that your consumer can

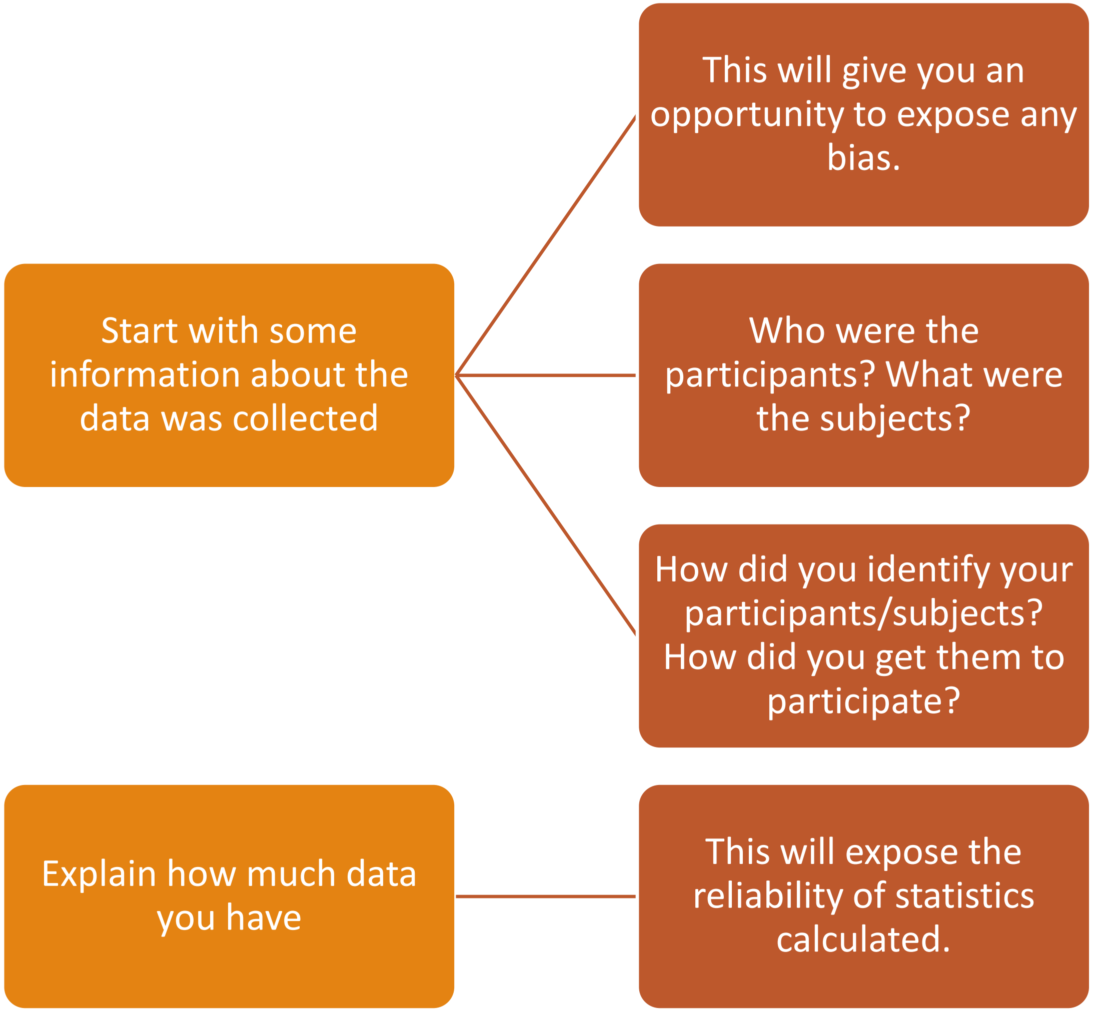
- See what you are basing your analysis/conclusions on
- Understand anything that may constrain those analysis/conclusions

You are trying to present information about a large body of data so that your consumer/reader can understand it without having to view every individual case you have collected

Reporting on your dataset

- ❑Put yourself in the readers shoes
- ❑Are you providing enough information to make the experiment/analysis reproducible?
- ❑Have you made clear any assumptions, sources of potential bias, mechanisms used to overcome flaws in the sample?
- ❑Have your provide references to any sources used?

Reporting on your dataset



Start with some information about the data was collected

This will give you an opportunity to expose any bias.

Who were the participants? What were the subjects?

How did you identify your participants/subjects? How did you get them to participate?

Explain how much data you have

This will expose the reliability of statistics calculated.

Reporting on your dataset

- ❑Start with some information about the participants/subjects
  - ❑Who were the participants? What were the subjects?
  - ❑How did you identify your participants/subjects?
  - ❑Example:

  The participants in this study included 134 cisgender men between 18 and 25 years old attending electronic engineering undergraduate degree programmes in a university in Dublin. All participants were fluent in English, and first-generation college students.

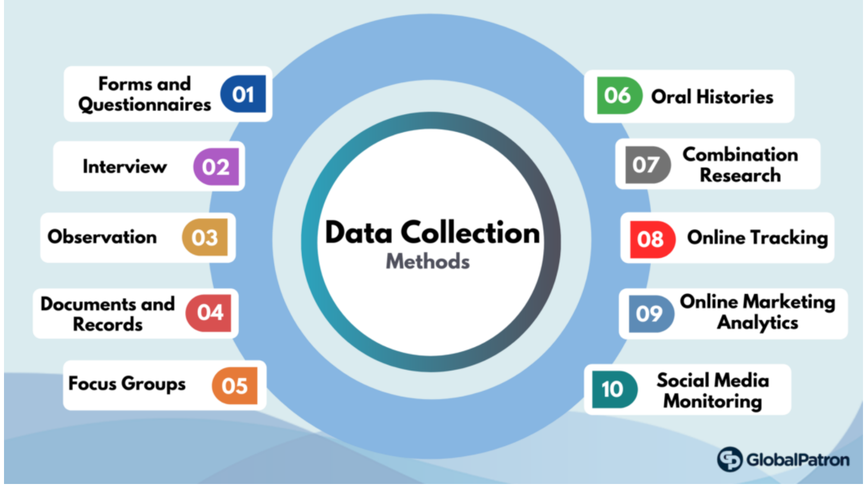
Reporting on your dataset

Data Collection
Methods
Forms and Questionnaires 01
Interview 02
Observation 03
Documents and Records 04
Focus Groups 05
06 Oral Histories
07 Combination Research
08 Online Tracking
09 Online Marketing Analytics
10 Social Media Monitoring
GlobalPatron

- ❑ Explain your sampling procedure. How did you decide what to include/exclude?
  - ❑ Did you invite participants? Were incentives provided? Did you need permission to conduct the study?
  - ❑ Is it a secondary data source? If so what do you know about how it was collected.
  - ❑ Example:

Ethics approval was obtained from the University Ethics committee before any participants were recruited. Current first-generation college students were invited to participate. The study was advertised through general emails sent to university-wide mailing lists, social media posts, and flyers across campus. Participants were self-selected and compensated with a voucher for their participation in the hour-long study.

Reporting on your dataset



❑Explain any procedures used in the data collection.

❑The design of the instruments/mechanisms used

❑Measures of reliability and validity

All participants were informed that the survey concerned students' general knowledge and would take a maximum of thirty minutes. After arriving at the laboratory individually, they were assured confidentiality, and they provided informed consent.

Participants were randomly assigned to a control or experimental condition. To begin, all participants were given the AAI and a demographics questionnaire to complete, followed by an unrelated filler task. In the control condition, participants completed a short general knowledge test immediately after the filler task. In the experimental condition, participants were asked to visualize themselves taking the test for 3 minutes before they actually did. For more details on the exact instructions and tasks given, see supplementary materials.

Reporting on your dataset

- ❑ State the variables and their statistical type.
  - ❑ You only need to describe the variables for the concepts of interest
- ❑ For each variable
  - ❑ State how much of it you have
    - ❑ You will be missing some data, and you need to acknowledge that (we will learn how to address this later)
  - ❑ State how you derived the measure (and if necessary, provide measures of validity and reliability)
  - ❑ You need to provide some information on the potential values that variables can take.

Examplar Case Study

PSYCHOLOGICAL ADJUSTMENT AND WELLBEING OF RESPONDENTS IN MELBOURNE AUSTRALIA

Example

Dataset: survey.dat

This is a real data set condensed for use by Julie Pallant to accompany her book Surviving Statistics.

This dataset contains data collected from respondents who completed a survey as part of a study to explore the factors that impact on respondents' psychological adjustment and wellbeing.

This dataset contains data collected from 439 respondents who completed a survey as part of a study to explore the factors that impact on respondents' psychological adjustment and wellbeing.

Example Description

Data was collected using a survey distributed to members of the general public in Melbourne, Australia and surrounding districts.

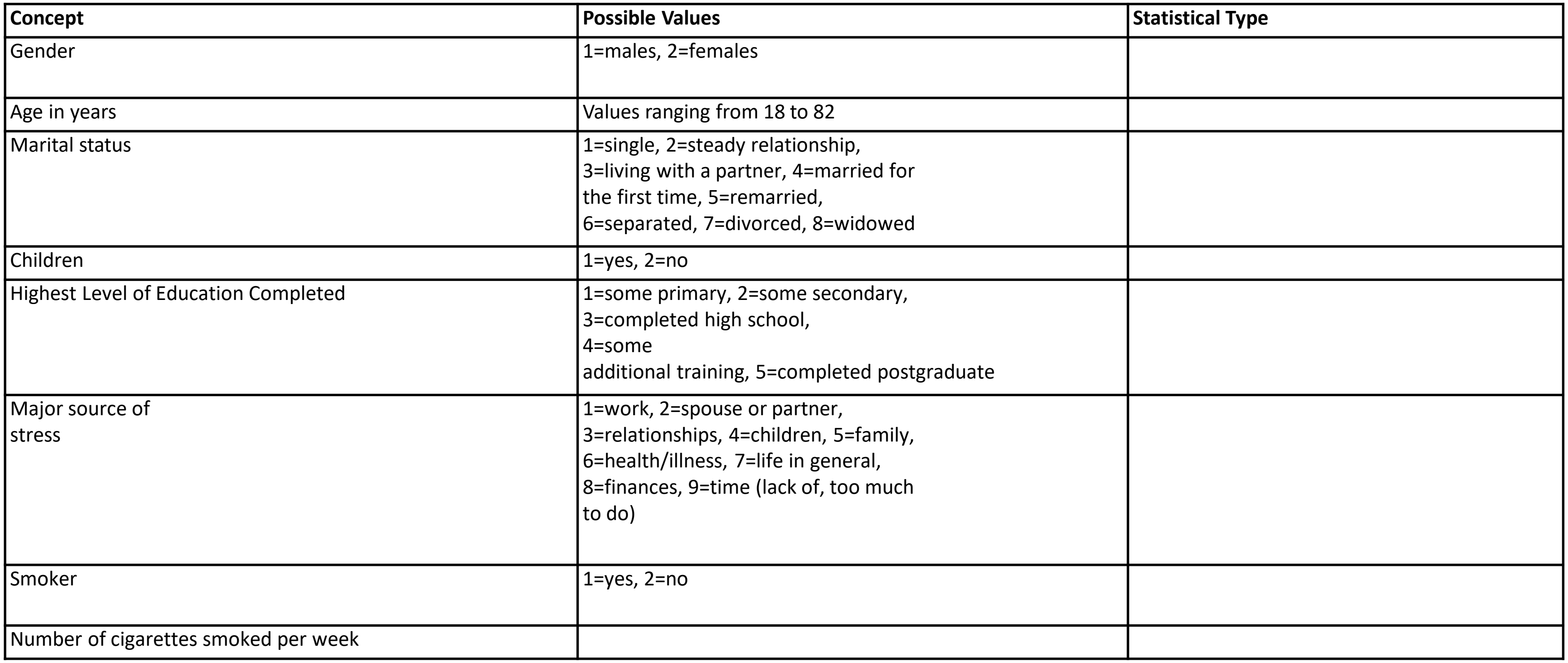
Participation in the study was voluntary and respondents were asked to complete a survey distributed via email.

The survey contained a variety of validated scales measuring constructs that the extensive literature on stress and coping suggest influence people's experience of stress. The scales measured self-esteem, optimism, perceptions of control, perceived stress, positive and negative affect, and life satisfaction. A scale was also included that measured people's tendency to present themselves in a favourable or socially desirable manner.

A range of demographic information was also collected from each respondent.

The final sample size was 439 with 42 per cent identifying as male and 58 per cent identifying as female. Respondent ages ranged from 18 to 82.

Survey.dat - Variables



Concept	Possible Values	Statistical Type
Gender	1=males, 2=females	
Age in years	Values ranging from 18 to 82	
Marital status	1=single, 2=steady relationship, 3=living with a partner, 4=married for the first time, 5=remarried, 6=separated, 7=divorced, 8=widowed	
Children	1=yes, 2=no	
Highest Level of Education Completed	1=some primary, 2=some secondary, 3=completed high school, 4=some additional training, 5=completed postgraduate	
Major source of stress	1=work, 2=spouse or partner, 3=relationships, 4=children, 5=family, 6=health/illness, 7=life in general, 8=finances, 9=time (lack of, too much to do)	
Smoker	1=yes, 2=no	
Number of cigarettes smoked per week		

Concept	Possible Values	Statistical Type
Gender	1=males, 2=females	Nominal
Age in years	Values ranging from 18 to 82	Ratio
Marital status	1=single, 2=steady relationship, 3=living with a partner, 4=married for the first time, 5=remarried, 6=separated, 7=divorced, 8=widowed	Nominal
Children	1=yes, 2=no	Nominal
Highest Level of Education Completed	1=some primary, 2=some secondary, 3=completed high school, 4=some additional training, 5=completed postgraduate	Ordinal
Major source of stress	1=work, 2=spouse or partner, 3=relationships, 4=children, 5=family, 6=health/illness, 7=life in general, 8=finances, 9=time (lack of, too much to do)	Nominal
Smoker	1=yes, 2=no	Nominal
Number of cigarettes smoked per week scale		Ratio

Concept	Possible Values	Statistical Type
Major Source of Stress	1= work; 2 = spouse or partner; 3 = relationships; 4 = children; 5 = family; 6 = health / illness; 7 = life in general	
Marital Status	1 = single; 2 = steady relationship; 3 = living with a partner; 4 = married for the first time; 5 = remarried; 6 = separated; 7 = divorced; 8 = widowed	
Age Group Categories	1=18-29yrs, 2=30-44yrs, 3=45+yrs	

Concept	Possible Values	Statistical Type
Major Source of Stress	1= work; 2 = spouse or partner; 3 = relationships; 4 = children; 5 = family; 6 = health / illness; 7 = life in general	Nominal
Marital Status	1 = single; 2 = steady relationship; 3 = living with a partner; 4 = married for the first time; 5 = remarried; 6 = separated; 7 = divorced; 8 = widowed	Nominal
Age Group Categories	1=18-29yrs, 2=30-44yrs, 3=45+yrs	Nominal

Concept	**Possible Values**	Staitstical Type
Optimism measured using the 6 item Life Orientation Test instrument developed by Scheier, M.F. & Carver, C.S. (1985). Optimism, coping and health: An assessment and implications of generalized outcome expectancies. Health Psychology, 4, 219–47.	Each Item: 1=strongly disagree, 5=strongly agree	
Mastery measured using the 7 item Master test instrument developed by Pearlin, L. & Schooler, C. (1978). The structure of coping. Journal of Health and Social Behavior, 19, 2–21.	Each Item: 1=strongly disagree, 4=strongly agree	
Positive and Negative Affect measured using the 10 item PANAS test instrument developed by Watson, D., Clark, L.A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54, 1063–70.	Each Item: 1=very slightly, 5=extremely	
Life Satisfaction measured using the 5 item Satisfaction with Life instrument developed by Diener, E., Emmons, R.A., Larson, R.J. & Griffin, S. (1985). The Satisfaction with Life scale. Journal of Personality Assessment, 49, 71–6.	Each Item: 1 =strongly disagree, 7=strongly agree scale	
Perceived Stress measured using the 10 item Perceived Stress test developed by Cohen, S., Kamarck, T. & Mermelstein, R. (1983). A global measure of perceived stress. Journal of Health and Social Behavior, 24, 385–96.	Each Item: 1=never, 5=very often	
Self-esteem measured using the 10 item Self-esteem test instrument developed by Rosenberg, M. (1965). Society and the adolescent self-image. Princeton, NJ: Princeton University Press.	Each Item: 1=strongly disagree, 4=strongly agree	
Social Desirability measured using the 10 item Social Desirability test instrument developed by Crowne, D.P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 24, 349–54.	Each Item: 1=true, 2=false	
Perceived Control of Internal States (PCOISS) measured by the 18 item PCOISS test instrument developed by Pallant, J. (2000). Development and validation of a scale to measure perceived control of internal states. Journal of Personality Assessment, 75, 2, 308–37.	Each Item: 1=strongly disagree, 5=strongly agree	

Concept	Possible Values	Statistical Type
Optimism measured using the 6 item Life Orientation Test instrument developed by Scheier, M.F. & Carver, C.S. (1985). Optimism, coping and health: An assessment and implications of generalized outcome expectancies. Health Psychology, 4, 219–47.	1=strongly disagree, 5=strongly agree	Ordinal
Mastery measured using the 7 item Master test instrument developed by Pearlin, L. & Schooler, C. (1978). The structure of coping. Journal of Health and Social Behavior, 19, 2–21.	1=strongly disagree, 4=strongly agree	Ordinal
Positive and Negative Affect measured using the 10 item PANAS test instrument developed by Watson, D., Clark, L.A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54, 1063–70.	1=very slightly, 5=extremely	Ordinal
Life Satisfaction measured using the 5 item Satisfaction with Life instrument developed by Diener, E., Emmons, R.A., Larson, R.J. & Griffin, S. (1985). The Satisfaction with Life scale. Journal of Personality Assessment, 49, 71–6.	1 =strongly disagree, 7=strongly agree scale	Ordinal
Perceived Stress measured using the 10 item Perceived Stress test developed by Cohen, S., Kamarck, T. & Mermelstein, R. (1983). A global measure of perceived stress. Journal of Health and Social Behavior, 24, 385–96.	1=never, 5=very often	Ordinal
Self-esteem measured using the 10 item Self-esteem test instrument developed by Rosenberg, M. (1965). Society and the adolescent self-image. Princeton, NJ: Princeton University Press.	1=strongly disagree, 4=strongly agree	Ordinal
Social Desirability measured using the 10 item Social Desirability test instrument developed by Crowne, D.P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 24, 349–54.	1=true, 2=false	Nominal
Perceived Control of Internal States (PCOISS) measured by the 18 item PCOISS test instrument developed by Pallant, J. (2000). Development and validation of a scale to measure perceived control of internal states. Journal of Personality Assessment, 75, 2, 308–37.	1=strongly disagree, 5=strongly agree	Ordinal

How would we work with all these Ordinal variables?

We have lots of different items for all the psychometric scales

Conducting individual analyses would be very time-consuming and may not be the focus of the study.

We need a way to establish a single measure for each of the major concepts.

This is where **dimension reduction** comes in (we will work on that later in the module).

In the dataset survey.dat measures have already been calculated for each one of these (as scale variables):

“tposaff"   "tnegaff"   "tlifesat"   "tpstress"   "tslfest"   "tmarlow"

"tpcoiss" “toptim” “tmast”

All of these are **Ratio** variables

Concept	**Statistical Type**
toptim created using dimension reduction of variables representing answers to the 6 item Life Orientation Test instrument developed by Scheier, M.F. & Carver, C.S. (1985). Optimism, coping and health: An assessment and implications of generalized outcome expectancies. Health Psychology, 4, 219–47.	Ratio
tmast created using dimension reduction of variables representing answers to the 7 item Master test instrument developed by Pearlin, L. & Schooler, C. (1978). The structure of coping. Journal of Health and Social Behavior, 19, 2–21.	Ratio
tposaff and **tnegaff** created using dimension reduction of variables representing answers to the 10 item PANAS test instrument developed by Watson, D., Clark, L.A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54, 1063–70.	Ratio
tlifesat created using dimension reduction of variables representing answers to the 5 item Satisfaction with Life instrument developed by Diener, E., Emmons, R.A., Larson, R.J. & Griffin, S. (1985). The Satisfaction with Life scale. Journal of Personality Assessment, 49, 71–6.	Ratio
tpstress created using dimension reduction of variables representing answers to the 10 item Perceived Stress test developed by Cohen, S., Kamarck, T. & Mermelstein, R. (1983). A global measure of perceived stress. Journal of Health and Social Behavior, 24, 385–96.	Ratio
tslfest created using dimension reduction of variables representing answers to the 10 item Self-esteem test instrument developed by Rosenberg, M. (1965). Society and the adolescent self-image. Princeton, NJ: Princeton University Press.	Ratio
tmarlow created using dimension reduction of variables representing answers to the 10 item Social Desirability test instrument developed by Crowne, D.P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 24, 349–54.	Ratio
topcoiss created using dimension reduction of variables representing answers to the 8 item PCOISS test instrument developed by Pallant, J. (2000). Development and validation of a scale to measure perceived control of internal states. Journal of Personality Assessment, 75, 2, 308–37.	Ratio

Two Research Questions

Does optimism predict life satisfaction, controlling for stress and self-esteem?

Do age, gender, optimism, and life satisfaction predict whether someone is a smoker (yes/no)?

What variables do we need from our survey dataset?

Concept	**Possible Values**	**Statistical Type**
Gender	1=males, 2=females	Nominal
Age in years	Values ranging from 18 to 82	Ratio
Marital status	1=single, 2=steady relationship, 3=living with a partner, 4=married for the first time, 5=remarried, 6=separated, 7=divorced, 8=widowed	Nominal
Children	1=yes, 2=no	Nominal
Highest Level of Education Completed	1=some primary, 2=some secondary, 3=completed high school, 4=some additional training, 5=completed postgraduate	Ordinal
Major source of stress	1=work, 2=spouse or partner, 3=relationships, 4=children, 5=family, 6=health/illness, 7=life in general, 8=finances, 9=time (lack of, too much to do)	Nominal
Smoker	1=yes, 2=no	Nominal
Number of cigarettes smoked per week scale		RAtio

Concept	Possible Values	Statistical Type
Major Source of Stress	1= work; 2 = spouse or partner; 3 = relationships; 4 = children; 5 = family; 6 = health / illness; 7 = life in general	Nominal
Marital Status	1 = single; 2 = steady relationship; 3 = living with a partner; 4 = married for the first time; 5 = remarried; 6 = separated; 7 = divorced; 8 = widowed	Nominal
Age Group Categories	1=18-29yrs, 2=30-44yrs, 3=45+yrs	Nominal

Concept	Statistical Type
toptim created using dimension reduction of variables representing answers to the 6 item Life Orientation Test instrument developed by Scheier, M.F. & Carver, C.S. (1985). Optimism, coping and health: An assessment and implications of generalized outcome expectancies. Health Psychology, 4, 219–47.	Ratio
tmast created using dimension reduction of variables representing answers to the 7 item Master test instrument developed by Pearlin, L. & Schooler, C. (1978). The structure of coping. Journal of Health and Social Behavior, 19, 2–21.	Ratio
tposaff and tnegaff created using dimension reduction of variables representing answers to the 10 item PANAS test instrument developed by Watson, D., Clark, L.A. & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. Journal of Personality and Social Psychology, 54, 1063–70.	Ratio
tlifesat created using dimension reduction of variables representing answers to the 5 item Satisfaction with Life instrument developed by Diener, E., Emmons, R.A., Larson, R.J. & Griffin, S. (1985). The Satisfaction with Life scale. Journal of Personality Assessment, 49, 71–6.	Ratio
tpstress created using dimension reduction of variables representing answers to the 10 item Perceived Stress test developed by Cohen, S., Kamarck, T. & Mermelstein, R. (1983). A global measure of perceived stress. Journal of Health and Social Behavior, 24, 385–96.	Ratio
tslfest created using dimension reduction of variables representing answers to the 10 item Self-esteem test instrument developed by Rosenberg, M. (1965). Society and the adolescent self-image. Princeton, NJ: Princeton University Press.	Ratio
tmarlow created using dimension reduction of variables representing answers to the 10 item Social Desirability test instrument developed by Crowne, D.P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. Journal of Consulting Psychology, 24, 349–54.	Ratio
Tpcoiss created using dimension reduction of variables representing answers to the 8 item PCOISS test instrument developed by Pallant, J. (2000). Development and validation of a scale to measure perceived control of internal states. Journal of Personality Assessment, 75, 2, 308–37.	Ratio

Does optimism predict life satisfaction, controlling for stress and self-esteem?

TOPTIM,
TLIFESAT,
TPSTRESS,
TSLFEST

Do age, gender, optimism, and life satisfaction predict whether someone is a smoker (yes/no)?

AGE,
GENDER,
TOPTIM,
TLIFESAT,
SMOKE

What do we need to describe for each variable?

Quantitative:

- Centre: mean, median, mode.
- Spread: range, SD, IQR.
- Shape: distribution - normal/skewed, outliers, multimodality.

Qualitative:

- Distribution across categories for qualitative data.

Population and Samples (Again)

Parameter- Population characteristic

Fixed value about a population typically unknown

Suppose we want to know the MEAN length of all the fish in Lough Mask . . .

Parameter- A Population characteristic

Fixed value about a population typically unknown

Suppose we want to know the MEAN length of all the fish in Lough Mask . . .

At any given point in time, how many values are there for the mean length of fish in the lake?

Parameter - population characteristic

Fixed value about a population typically unknown

Suppose we want to know the MEAN length of all the fish in Lough Mask . . .

Is this a value that is known?

Parameter - Population characteristic

Fixed value about a population typically unknown

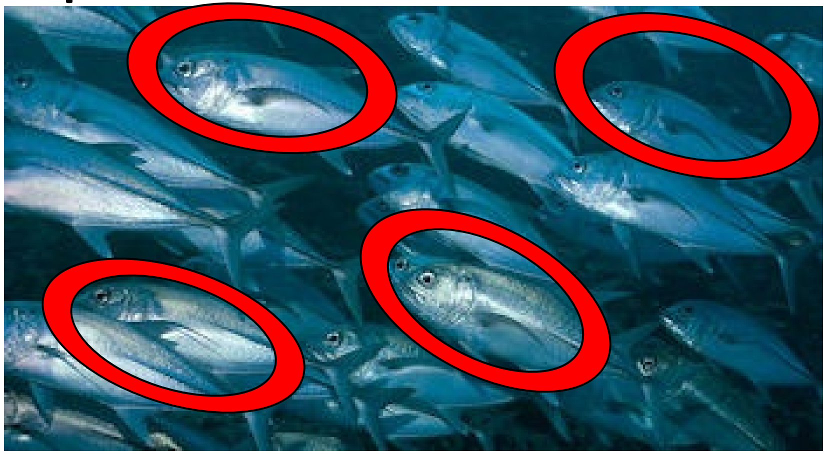
Suppose we want to know the MEAN length of all the fish in Lough Mask . . .

Can we find it out?

Statistic

Suppose we want to know the MEAN length of all the fish in Lough Mask.

What can we do to estimate this unknown population characteristic?



We can calculate a **statistic** – a value calculated from a **sample**

Statistic

Suppose we want to know the MEAN length of all the fish in Lough Mask.

How can we provide a more reliable estimate of the **population mean**?

Statistic

Suppose we want to know the MEAN length of all the fish in Lough Mask.

How can we provide a more reliable estimate of the population mean?

- Using multiple samples
  - Calculate the mean for each
  - Calculate the mean of the means

Suppose that everyone in the class caught a sample of 6 fish from the lake.
Would each of our samples contain the same fish?

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?

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?

And the samples would also have different ranges!

Going beyond the data

We ideally want to collect data from all members of the population

- But can't

We usually collect a number of samples

- Each sample could have a different mean
  - **Sampling variation**
  - Variation in the observed values of the sampling statistic
- We can plot the sample means into a frequency distribution
  - **Sample distribution**

The Sampling Distribution

Any statistic, being calculated from a sample, will vary between samples and consequently will have a distribution

Consider the mean

- If it is calculated repeatedly for different samples drawn from a population, these values of the mean will vary and will be distributed in some way
- The distribution of a statistic is called a **sampling distribution**
- This has a **standard error** (corresponding to the standard deviation of a value distribution)
- And has an **expected value** (corresponding to the mean of a value distribution)

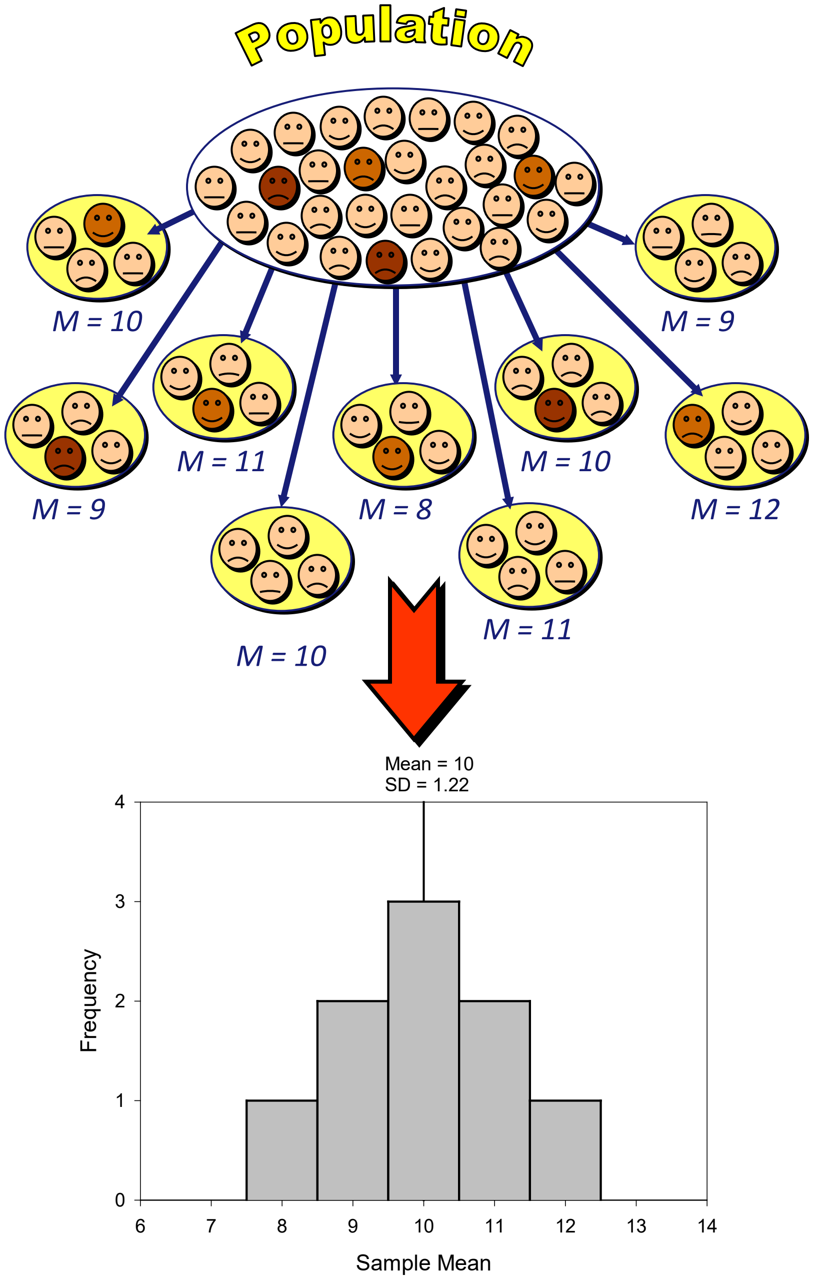
Going beyond the data

So what?

- If we have enough samples, we can estimate 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)**

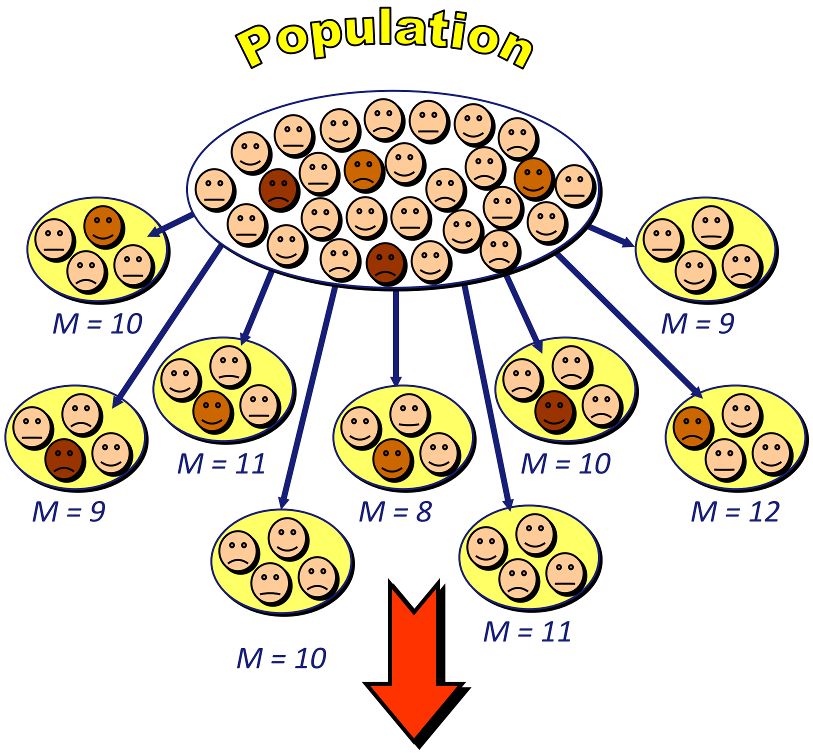


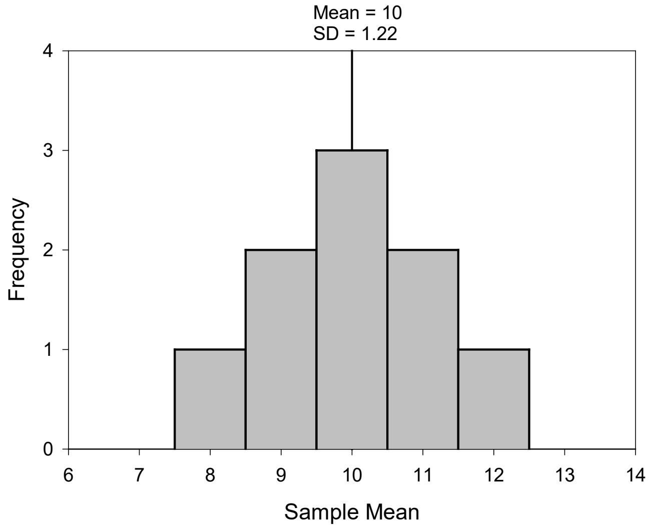
Taking a set of samples from the population

We can calculate the standard error of the mean:

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$

s= sample standard deviation





Why use square root of N rather than N?

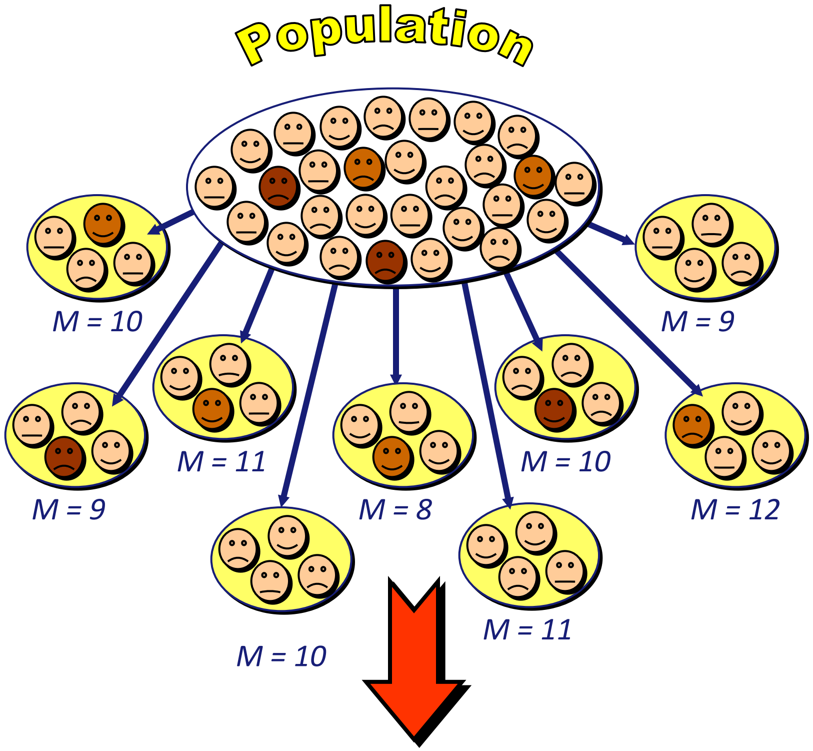
Sample size influences variability and the distribution of sample means.

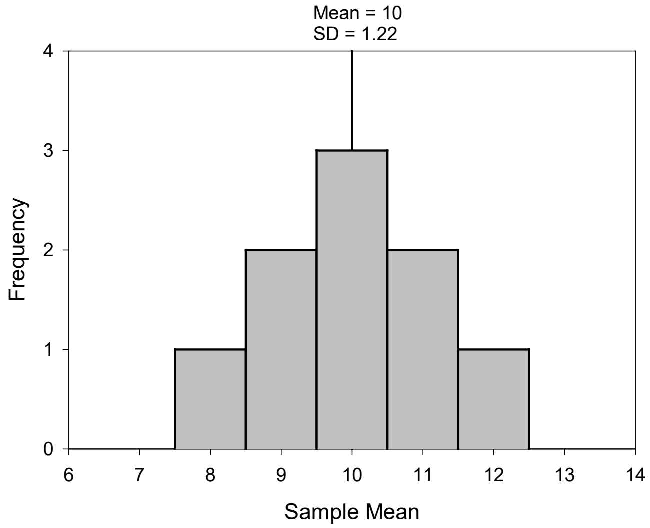
If you used n directly, the SE of the Mean (SEM) would decrease too quickly as the sample size increases.

In reality, the relationship between sample size and the reduction in variability is not linear, but rather proportional to the square root of the sample size.

s= sample standard deviation

$$\boxed{\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}}$$





The reduction follows the square root law, which reflects the way randomness averages out as you take larger samples.
Using square root of n instead adjusts the SEM to reflect the fact that each additional data point provides diminishing returns in terms of reducing the variability of the sample mean.

s= sample standard deviation

$$\sigma_{\bar{X}} = \frac{s}{\sqrt{N}}$$

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**

- Concerned with drawing finite samples of size *n* from a population with a known mean μ, and a known standard deviation, σ.
- Two alternative forms
  1. If we collect samples of size *n* with a "large enough *n*," calculate each sample's mean, and create a histogram of those means, then the resulting histogram will tend to have an approximate normal bell shape.
  2. If we again collect samples of size *n* that are "large enough," calculate the sum of each sample and create a histogram, then the resulting histogram will again tend to have a normal bell-shape.

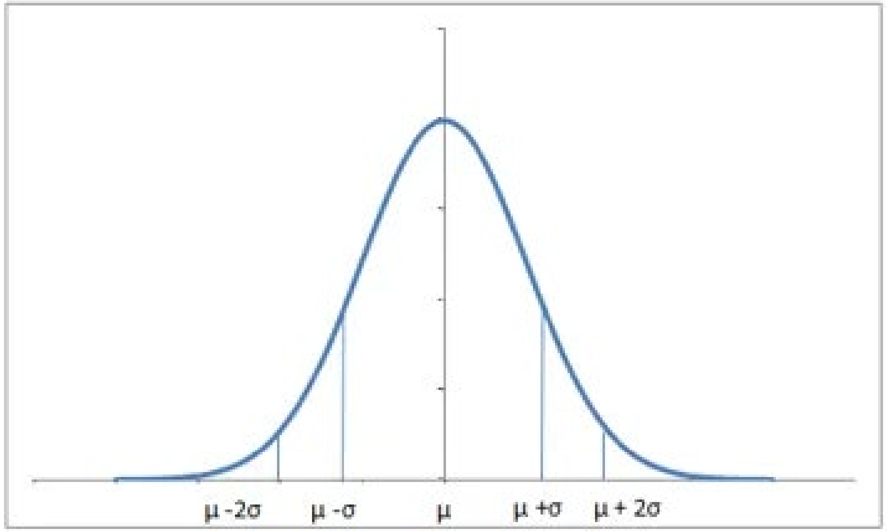
Going beyond the data

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 μ, and a standard deviation σ of* :

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

Normal Distribution



The normal distribution is perfectly symmetrical about the mean.
The probabilities move similarly in both directions around the mean.
The total area under the curve is 1, since summing up all the possible probabilities would give 1.

So what does this mean?

If our distribution follows the normal distribution

- We can use the standard deviation of the sampling distribution as the approximation of the sample error
  - For other shapes of distribution, we have other ways of approximating the population mean and standard error.

Probability Frequency Distribution

Shows how often an event will happen

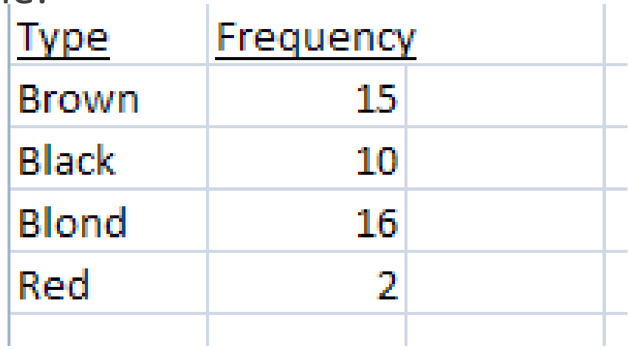
Sample question:

In a sample of 43 students:

- 15 had brown hair.
- 10 had black hair.
- 16 had blond hair.
- 2 had red hair.
- find the **probability** a person has neither red nor blond hair.

Probability Frequency Distribution

Frequency distribution table:



Type	Frequency
Brown	15
Black	10
Blond	16
Red	2

Brown = 15/43 (15 out of 43 students have brown hair).

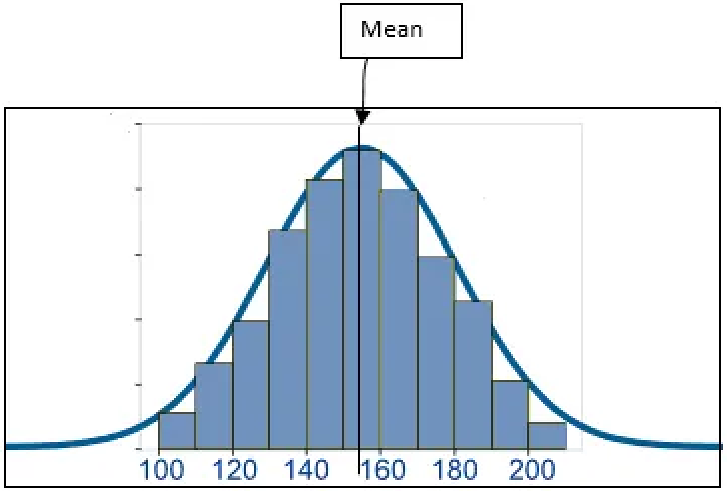
Black = 10/43 (10 out of 43 students have black hair).

Add these together to get the total number of students who have either brown or black hair

- 15/43 + 10/43 = 25/43
  - (25 out of 43 students have either brown or black hair).

Frequency Distribution to Probability

Frequency distribution can be displayed as a histogram- gives us an idea about how frequently a given data point occurs



From this, probability can be calculated and using **probability density function** (equation/function)

- Using this you can plot the probability distribution

Sampling

When you have huge amount of data, it is difficult to make sense of it (or even collect it).

To tackle this problem, what we do is take a small chunk of data and look at it.

But we won’t be satisfied with just a single chunk.

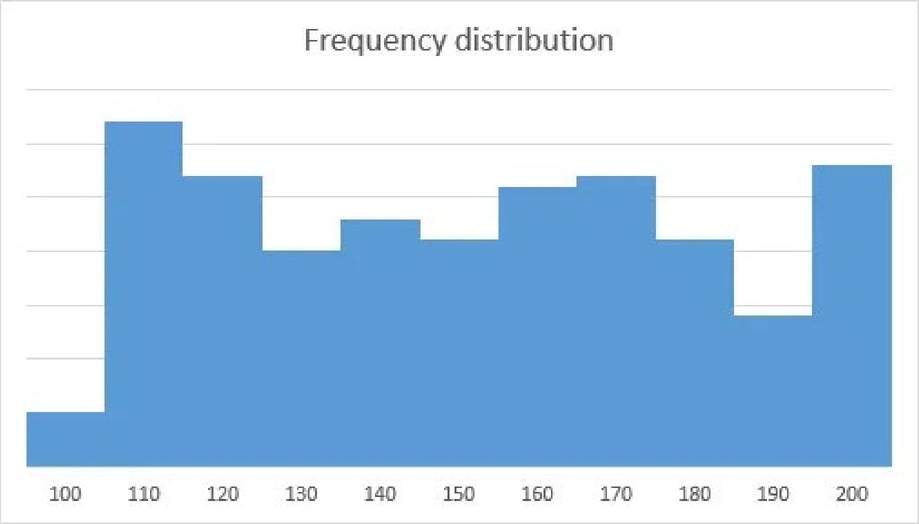
We’d try to look at multiple chunks to be sure of results.

Sampling

Let’s say we have the cholesterol levels of all the people in a country

We can look at the mean, median and mode of the data.

Suppose we plot the data, and it looks like this.



We calculate the mean is 153.2

Sampling

But it is difficult to process this large amount of data.

We can take the data of some 50 people and calculate their mean.

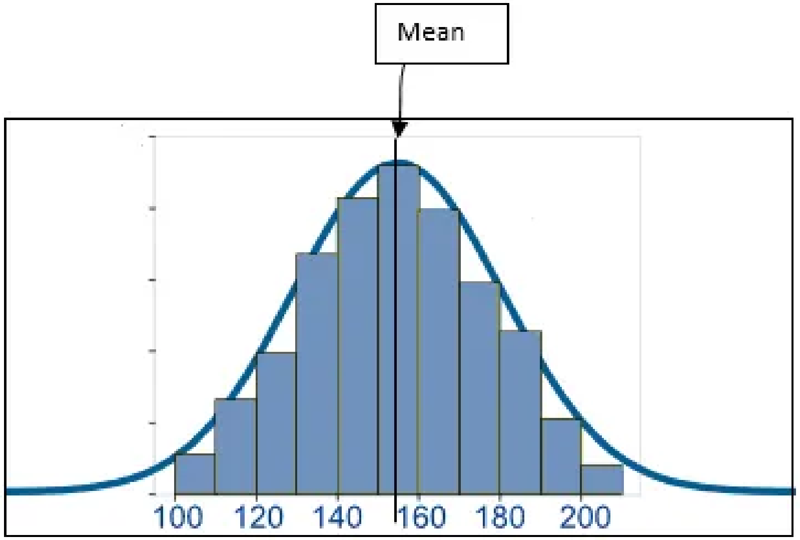
- We then take a new sample of 50 people and calculate the mean.
- We then take another new sample of 50 people and calculate the mean

- We keep doing that to collect a quantity of samples….

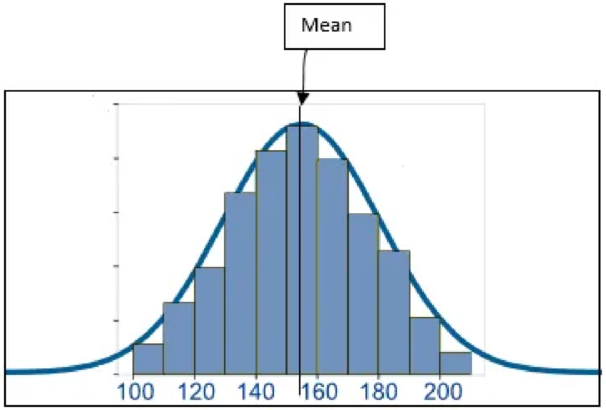
Sampling

We can then plot the means (or any other statistic) of these samples.

We get a symmetrical frequency distribution.



Sampling



When we take means of cholesterol levels of 50 people, again and again, we observe the mean values are around 150-160.

- Only a few mean values are more than 170 and less than 140.
- There are very, very few over 190 or less than 110.

Sampling – Frequency to probability

We can easily convert these frequencies to see probabilities.

- If we divide the frequency of a bin (range like 110 to 120) by the total number of data points, we get the probabilities of each bin.
- This converts the frequency distribution to a probability distribution of the same shape.

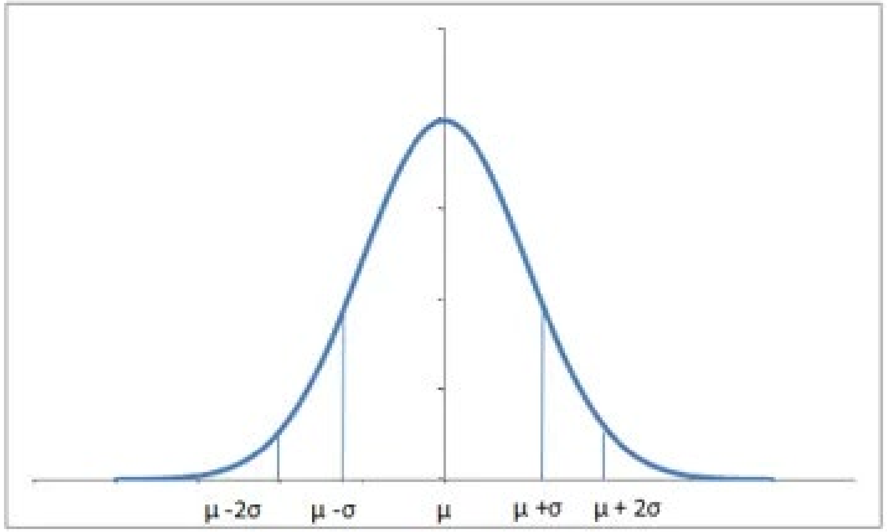
Sampling – Frequency to probability

The probability distribution becomes more and more symmetrical when the sample size that we use to create those means is very large.

As the sample size approaches infinity, the probability distribution becomes a perfectly symmetrical where the center of the curve is the mean of the population **– the central limit theorem.**

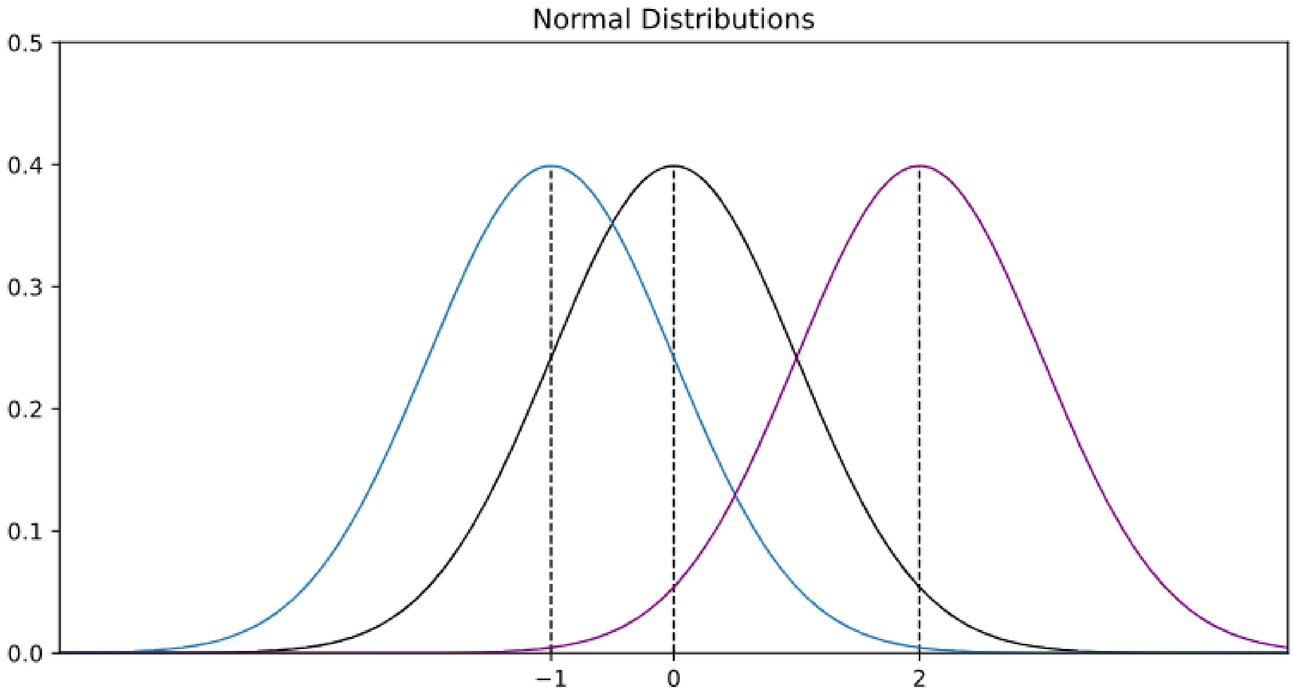
The curve is known as **normal distribution.**

Normal Distribution



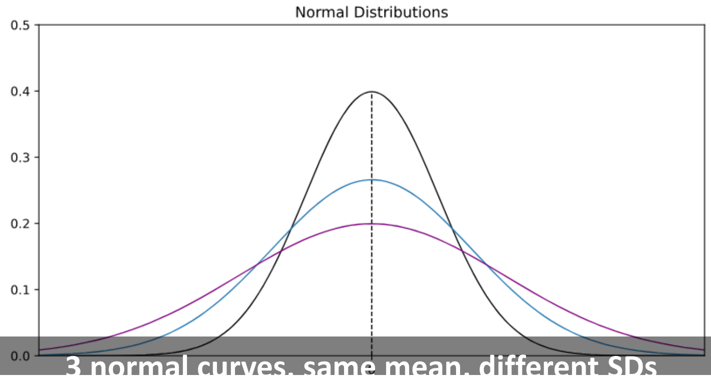
The normal distribution is perfectly symmetrical about the mean.
The probabilities move similarly in both directions around the mean.
The total area under the curve is 1, since summing up all the possible probabilities would give 1.

Normal Distribution



3 normal curves, same SD different mean

Normal Distribution



3 normal curves, same mean, different SDs

The distribution might vary a bit depending upon how spread the data is.

If the data has a very large standard deviation, the curve would be spread out and flatter since more of the values would be away from the mean.

Probability Distribution

If a lot of values are away from the mean, the probability for data being around the population mean also drops.

Probability Distribution

Similarly, if the standard deviation is low,

- which means most of the values are near around the mean

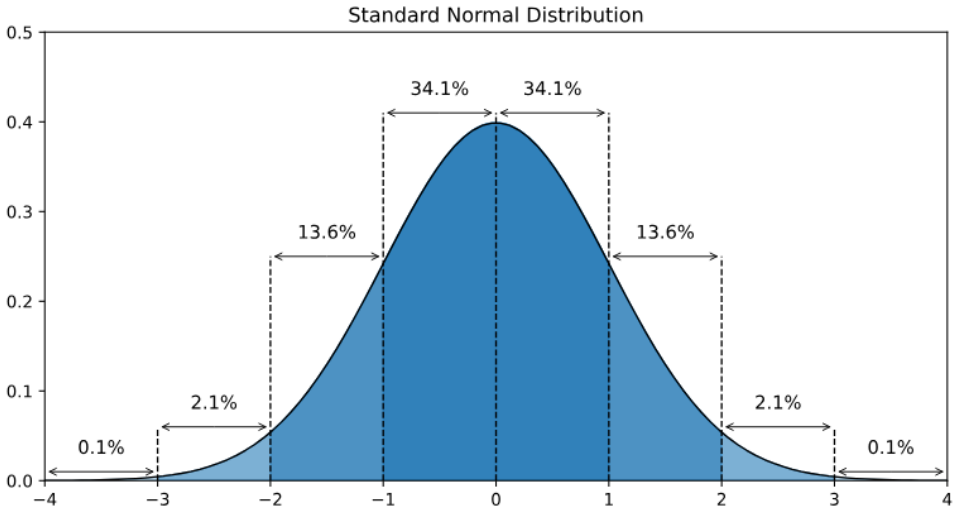
Using the laws of probability, there is high probability of the sample mean being around the population mean and the distribution is a lot thinner.

The Standard Normal

A normal distribution with a mean of 0 and a standard deviation of 1 is called a **standard normal distribution**.

Any normal distribution in any scale can be converted to the standard normal by **transforming the scale to be in units of standard deviation**.

- We change the units of measurement



The Standard Normal

Benefits

- Allows us to quantitatively compare concepts which were measured using different scales
- Allows us to use pre-calculated statistical tables to draw conclusions about statistics calculated from a sample that conforms to the normal distribution

Standard normal distribution

So what?

- If we have data shaped like the normal distribution
  - 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

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

Z-scores

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

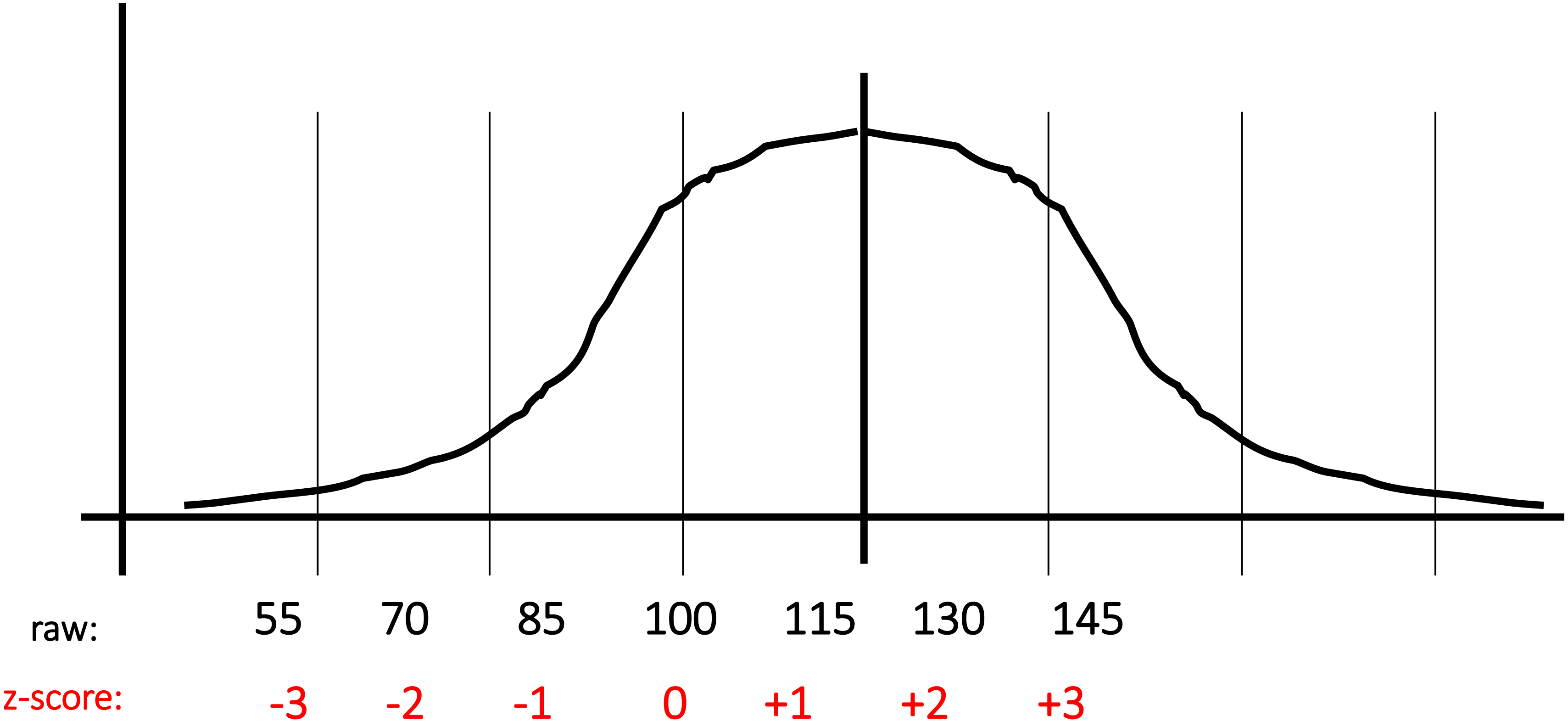
z-scores

Transform an original IQ scores into scores with a mean of 0 and an SD of 1.

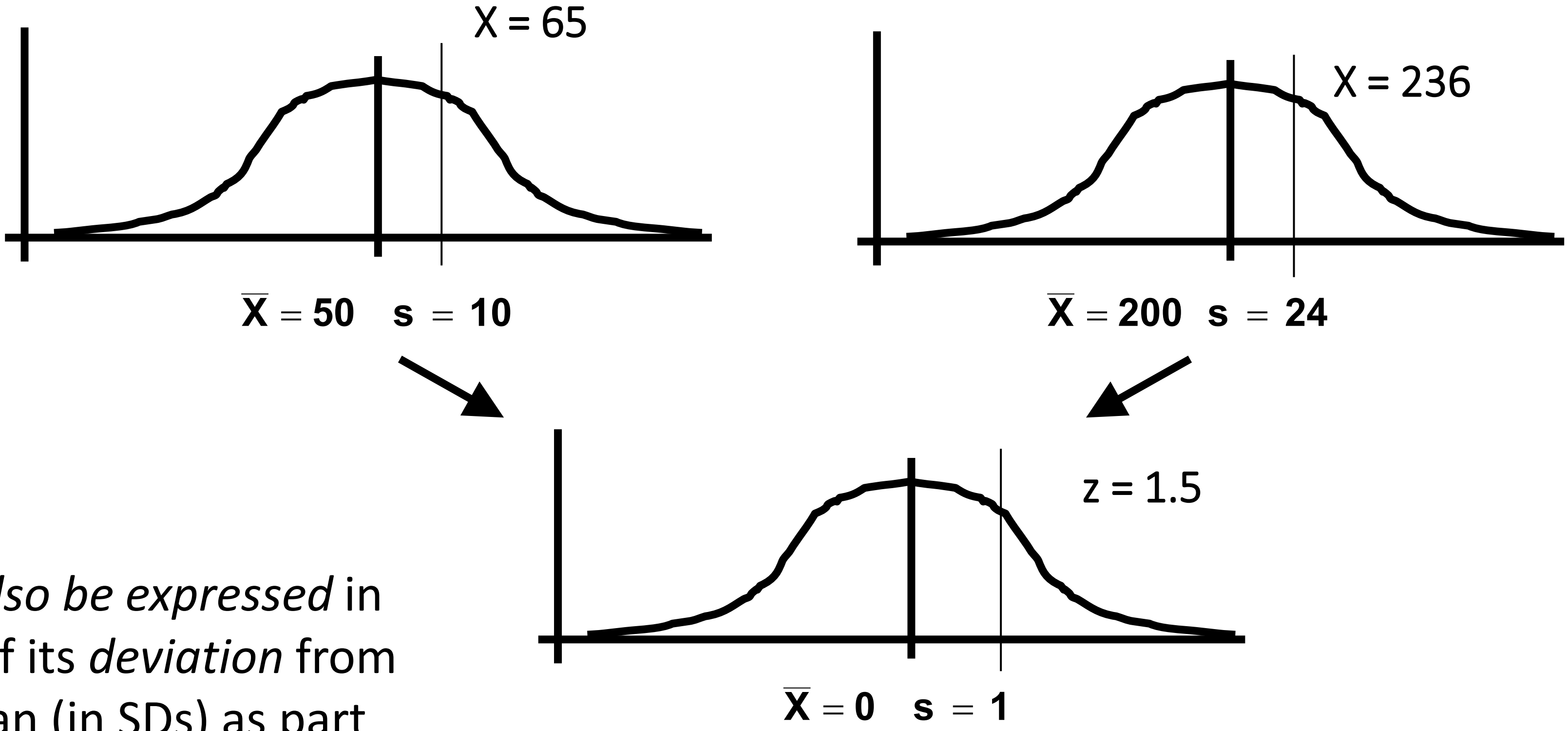
Raw IQ scores (mean = 100, SD = 15)

z for 100 = (100-100) / 15 = 0,   z for 115 = (115-100) / 15 = 1,

z for 70 = (70-100) / -2, etc.



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


X = 65
$\bar{X} = 50 \quad s = 10$
X = 236
$\bar{X} = 200 \quad s = 24$
z = 1.5
$\bar{X} = 0 \quad s = 1$

It can also be expressed in terms of its *deviation* from the mean (in SDs) as part of a *z-score distribution*

The Standard Normal Distribution

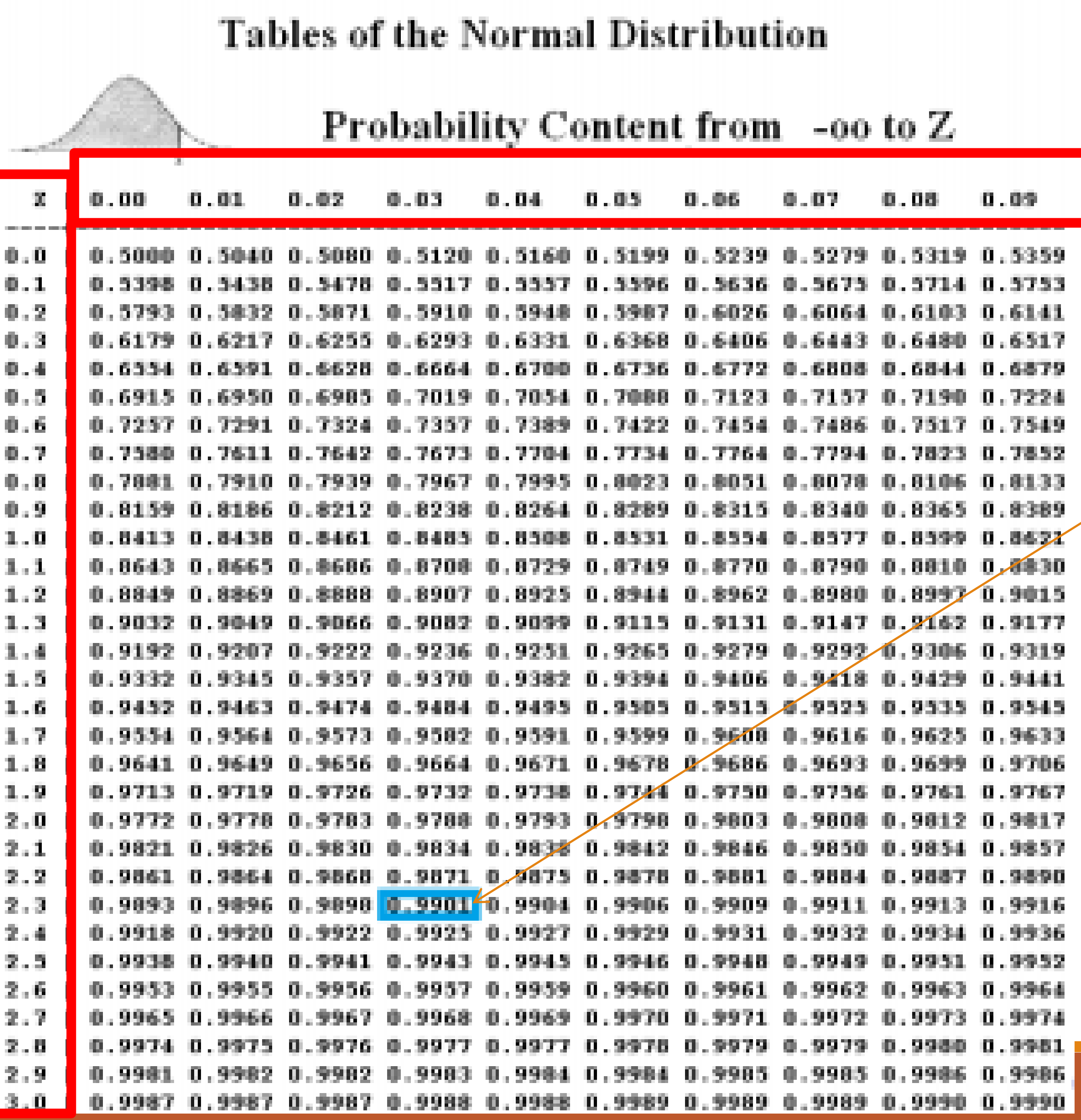
The distribution of a normal variable with mean equal to zero and standard deviation equal to 1 looks identical to that of the normal but uses a different measurement scale.

The Standard Normal Distribution

So what?

- It is the fact that we can now have access to a table showing, for each point in $[-\infty, +\infty]$, the probability that we have a realization of a variable to the left and to the right of that point.

Stem and Leaf



Tables of the Normal Distribution

Probability Content from -oo to Z

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	0.9925	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	0.9959	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
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

This is a stem and leaf table

The probability that a realization is lower than point 2.33 = 0.99

Then the probability that the realization 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:

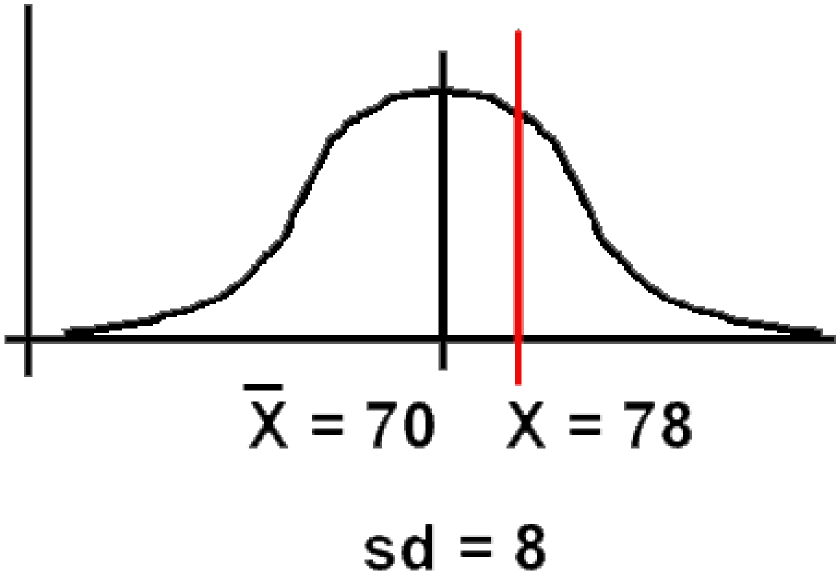
- Test A: Fred scores 78. Mean score = 70, SD = 8.
- Test B: Fred scores 78. Mean score = 66, SD = 6.

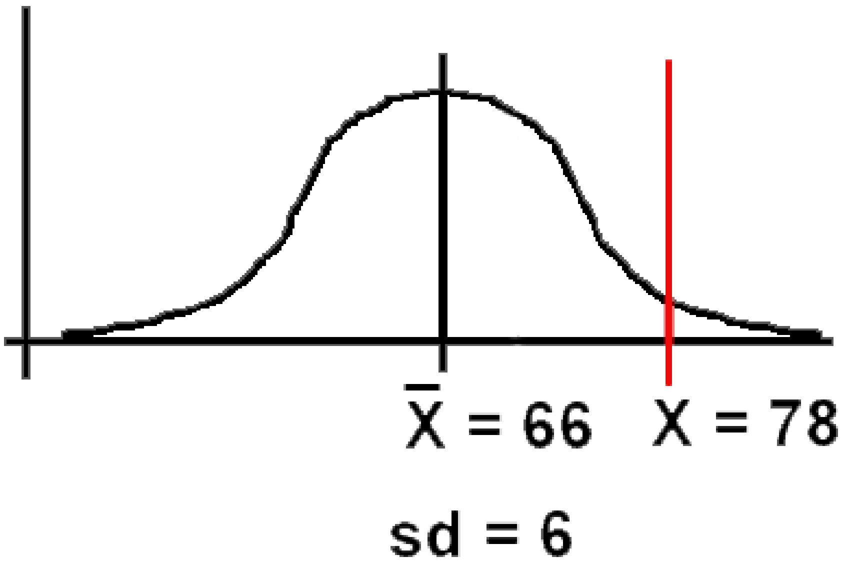
- Did Fred do better or worse in comparison to the rest of the class on the second test?

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.

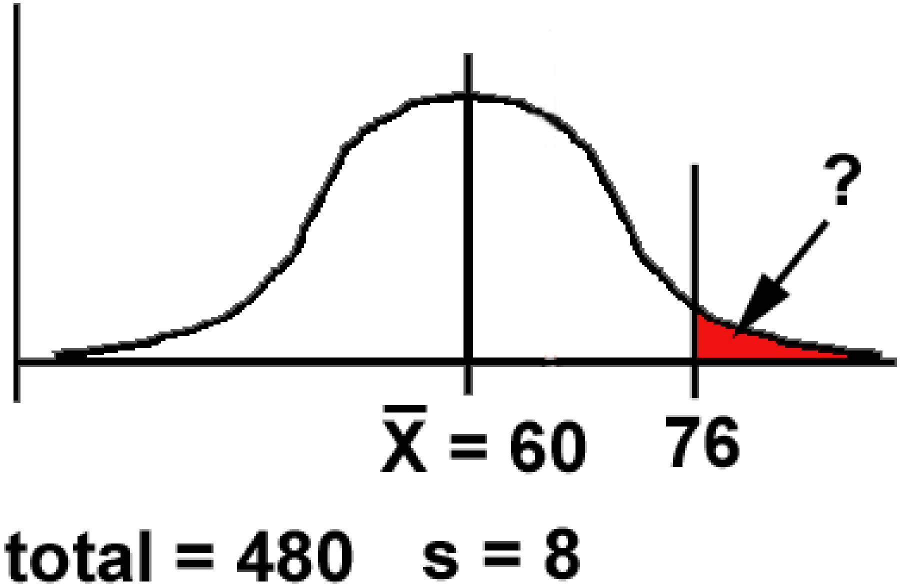


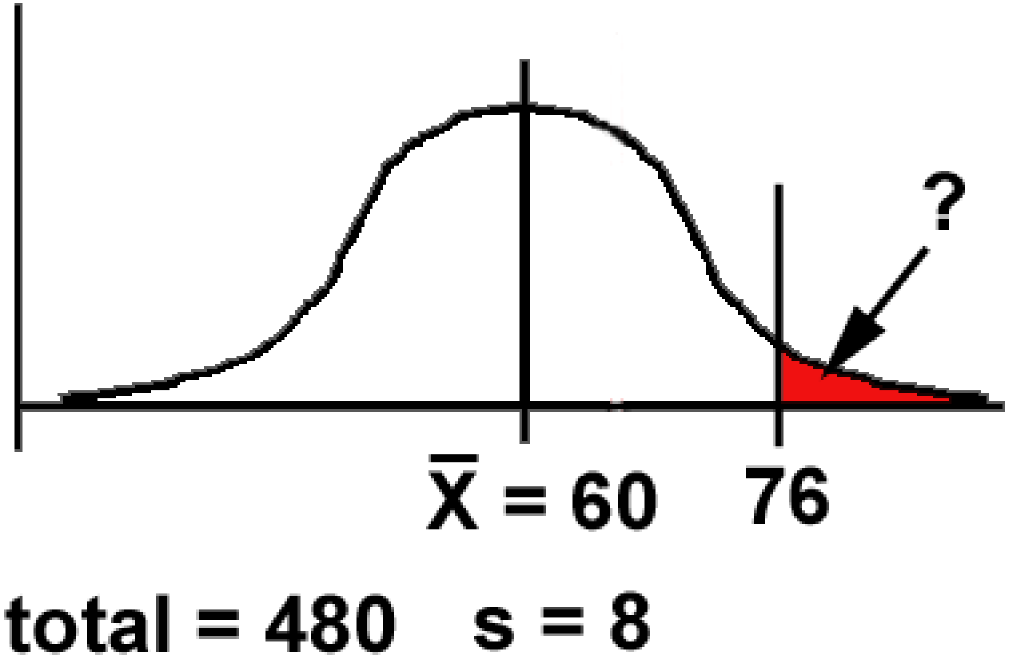


z-scores enable us to determine the relationship between one score and the rest of the scores, using just one table for *all* normal distributions.

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:





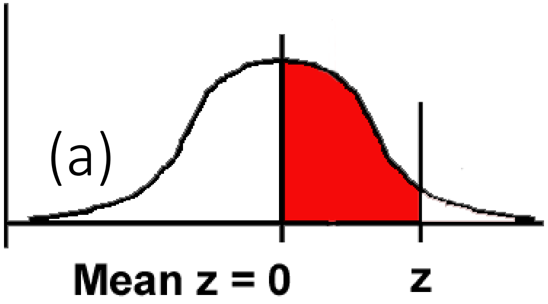
Work out the z-score for 76:

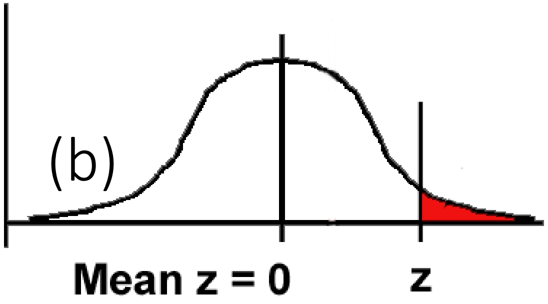
$$z = (X - \bar{X}) / s \quad = \quad (76 - 60) / 8 \quad = \quad 16 / 8 \quad = 2.00$$

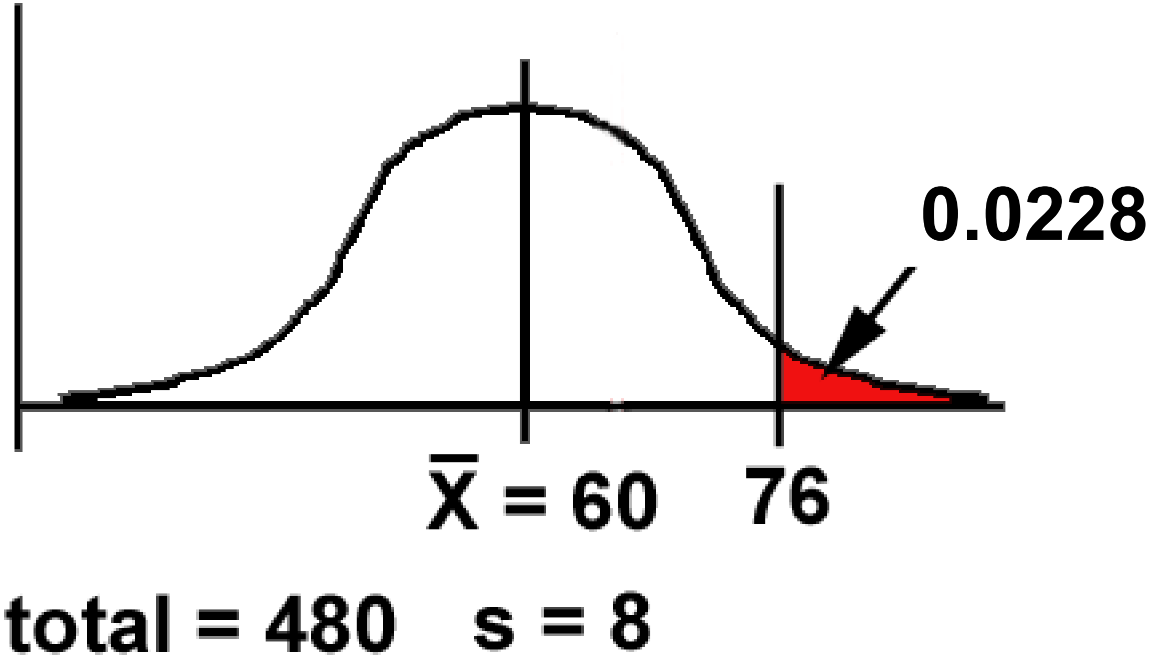
We need to know the size of the area beyond z (remember - the area under the Normal curve corresponds directly to the proportion of scores).

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

z	(a) Area between mean and z	(b) Area beyond z
0.00	0.0000	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







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

As a percentage = **2.28%**

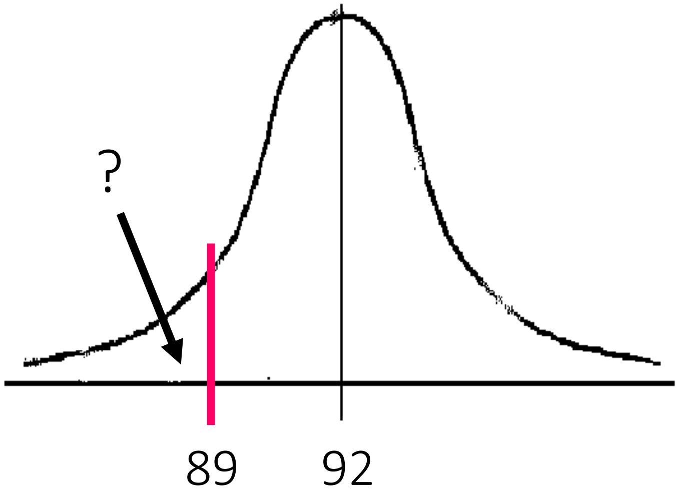
As a number= 0.0228 * 480 = **10.94** scores.

Conducting a word comprehension test to investigate the impact a head injury has on a person’s word comprehension.

Have created a dataset of scores from people with NO head injury
No. correct: mean = 92, SD = 6 out of 100

Have conducted the test with a person WITH a head injury:
No. correct: 89 out of 100.
Is this person's comprehension significantly impaired?

Step 1: Graph the problem:



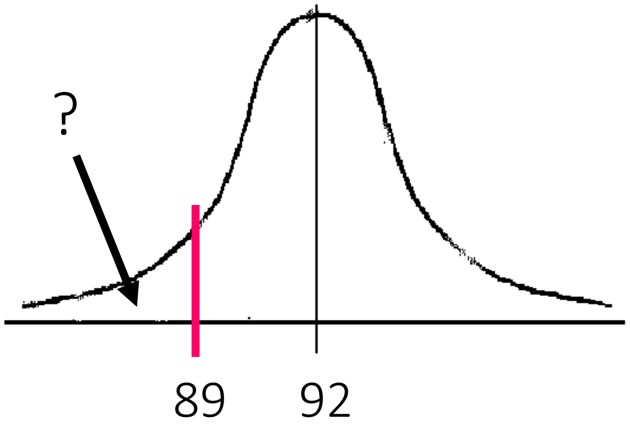
Step 2: Convert 89 into a z-score:

$$z = (89 - 92) / 6 = -3 / 6 = -0.5$$

Step 3: use the table to find the "area beyond z" for our z of - 0.5:

Area beyond z = 0.3085

Conclusion: .31 (31%) of people without a head injury are likely to have a comprehension score this low or lower.



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

Normal Distribution

A *density curve* describes the overall pattern of a distribution.

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.

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)

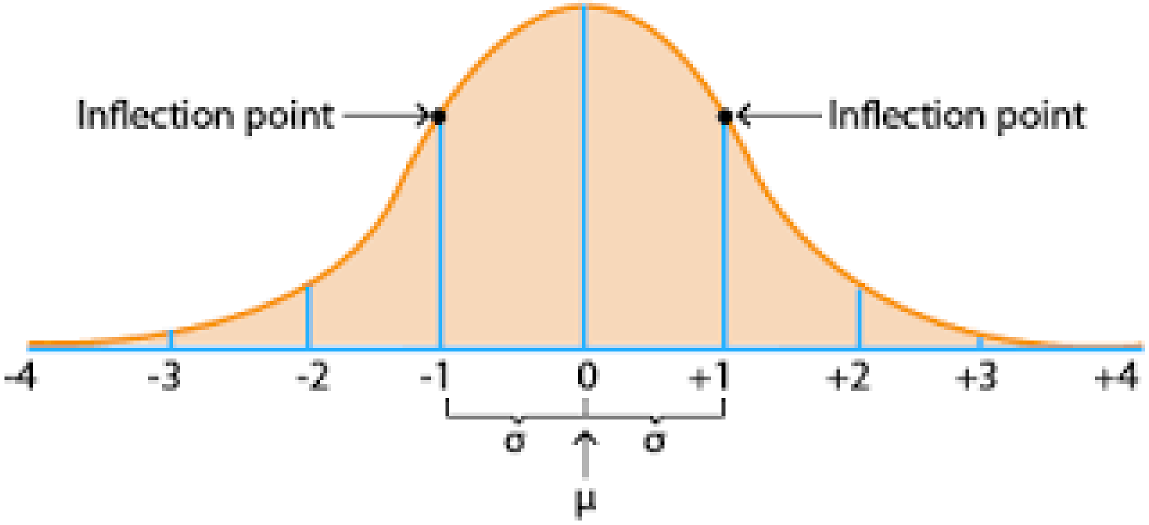
The spread of the curve is determined by the standard deviation

- 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



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

The 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

We can now translate this to z-scores.

Properties of z-scores

Using our empirical rule we know that:

- 95% of z-scores lie between −1.96 and 1.96.
  - 1.96 cuts off the top 2.5% of the distribution.
  - −1.96 cuts off the bottom 2.5% of the distribution.
- 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

- (a) comparing scores on different raw-score scales;
- (b) showing how a given score stands in relation to the overall set of scores.
- (c) using probability tables to calculate likelihood of particular scores.

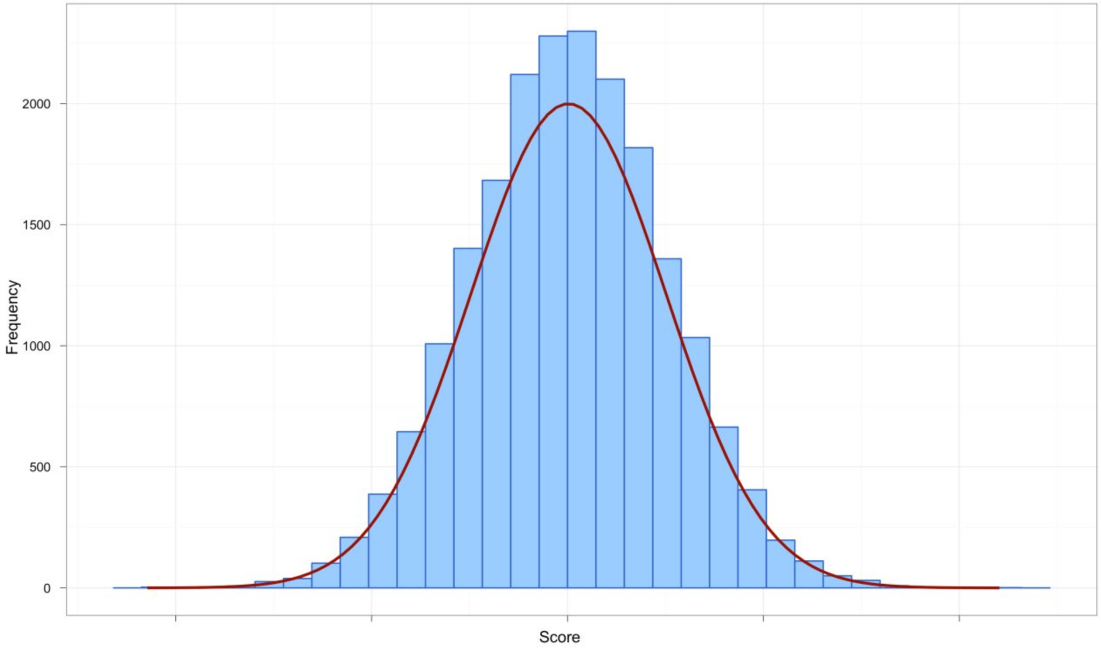
Normal distribution in summary

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.

The Normal Distribution



The curve shows the idealized shape.

It is important that our data is close to this shape if we wish to use Parametric tests.

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!

Purpose: Describe and Summarize Data

Key ideas

Provide the maximum description with the minimum statistics

Generally, you need to provide**relevant statistics** and **visualization** for key concepts.

Key ideas - Qualitative

Summarize the distribution of categories: How the data is distributed across categories.

Provide insight into trends or patterns

Compare different categories

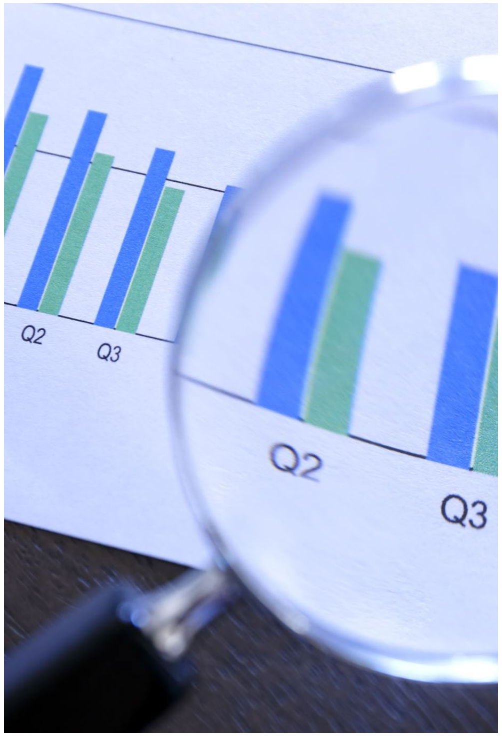
Use:

- **Frequencies**: How often each category or value appears.
- **Percentages/Proportions**: Relative frequency of each category to the total.
- **Mode**: Most frequently occurring category.

Key ideas - Quantitative

Give a sense of the "average" value or **typical case**.

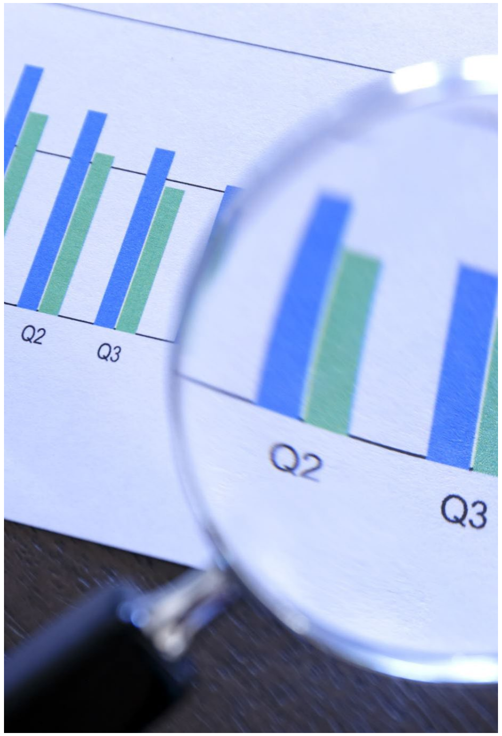
- **Mean**: Arithmetic average of the data.
- **Median**: Middle value, especially useful when data is skewed (more of a particular type of data).
- **Mode**: Most frequently occurring value (if any).



Key ideas - Quantitative

Summarize the **Variability (Spread)**: Spread or variability of the data.

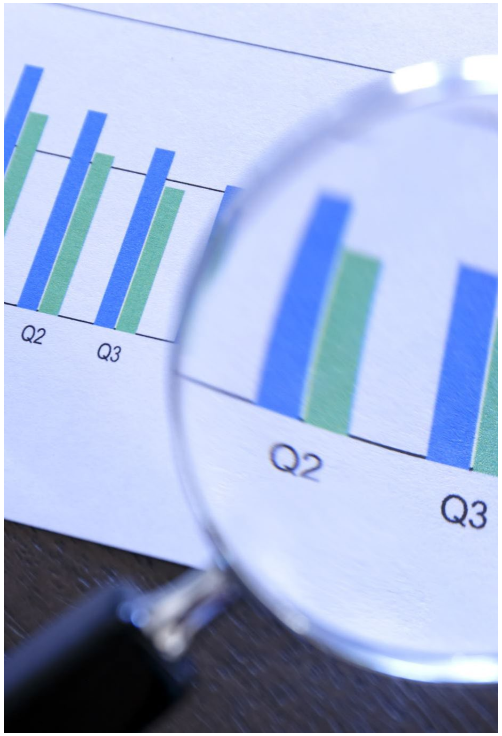
- **Variance**: How much the data varies from the mean.
- **Standard Deviation**: A measure of how spread out the values are around the mean.
- **Interquartile Range** (IQR): Spread of the middle 50% of the data, useful for skewed distributions.
- **Range**: The difference between the largest and smallest values.

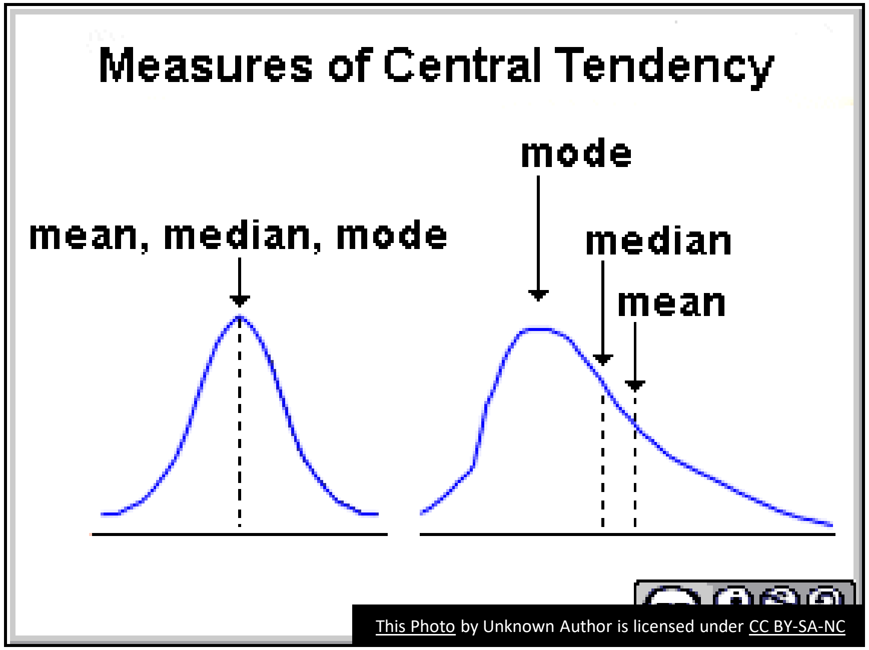


Key ideas - Quantitative

Shape of the Data Distribution: Describe the shape of the data distribution (e.g., symmetric, skewed, bell-shaped).

- **Bell Shaped:** Indicates that the data is equally distributed either side of the mean (aligns with the normal distribution).
- **Skewness**: Indicates if the data is more spread out on one side of the mean.
- **Kurtosis**: Describes whether the distribution has heavy tails or is more peaked than a normal distribution.
- Identify **Outliers**: Values that are significantly higher or lower than the rest of the data.





Central Tendency

Measure of Central Tendency

A descriptive statistic for numerical data – a single variable.

A single number to serve as a representative value around which all the numbers in the set tend to cluster.

Measure of Central Tendency

Mode:

- The value that occurs most frequently for a variable in a set of data.
- Undefined for sequences in which no observation is repeated.

Median:

- The value in the middle; half of the values for a variable are larger than the median and half of the values are smaller than the median
- The middle value of a sequence of all the values in a distribution arranged from lowest to highest.
- In case of an even number of observations the average of the two middle most values is the median.

Mean:

- The arithmetic average of a group of values; the sum of the values divided by the number of values.

$$\overline{X} = \frac{\sum X}{N}$$

Calculating the Mean

Calculate the mean of the following data:

1 5 4 3 2

Sum the scores (ΣX):

$1 + 5 + 4 + 3 + 2 = 15$

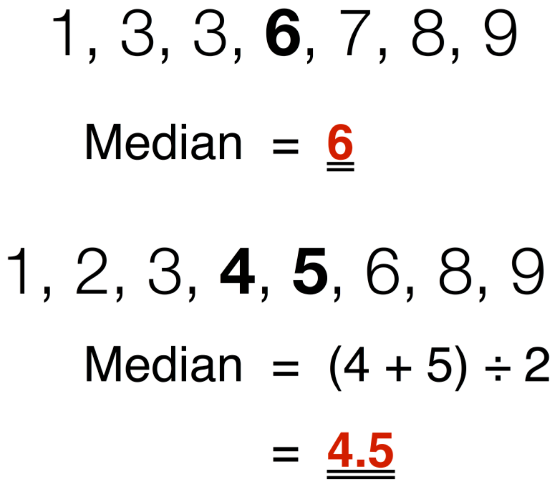
Divide the sum ($\Sigma X = 15$) by the number of scores ($N = 5$):

$15 / 5 = 3$

Mean = $\overline{\mathbf{X}} = 3$

The mean is sensitive to extreme values

The Median



1, 3, 3, **6**, 7, 8, 9

$$\text{Median} = \underline{\underline{6}}$$

1, 2, 3, **4**, **5**, 6, 8, 9

$$\text{Median} = (4 + 5) \div 2$$

$$= \underline{\underline{4.5}}$$

This Photo by Unknown Author is licensed under CC BY-SA

Just another name for the 50th percentile

- It is the score in the middle; half of the scores are larger than the median and half of the scores are smaller than the median
- The middle score of a sequence of all the scores in a distribution arranged from lowest to highest.

Sort the data from highest to lowest

Find the score in the middle

- middle = (N + 1) / 2
- If N, the number of scores, is even the median is the average of the middle two scores
- $\widetilde{X}$

Median Example

What is the median of the following scores:
10 8 14 15 7 3 3 8 12 10 9

Sort the scores:
15 14 12 10 10 9 8 8 7 3 3

Determine the middle score:
middle = (N + 1) / 2 = (11 + 1) / 2 = 6

Middle score (6th score)

Median = 9

Median Example

What is the median of the following scores:
24 18 19 42 16 12

Sort the scores:
42 24 19 18 16 12

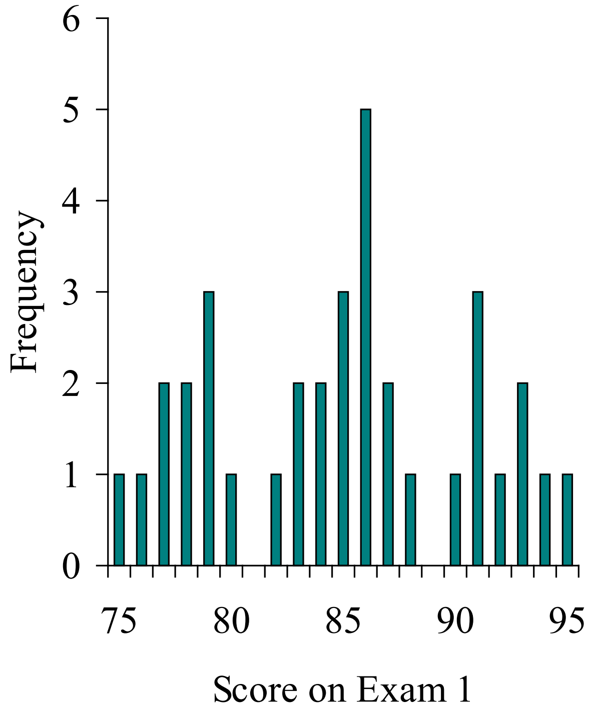
Determine the middle score:
middle = (N + 1) / 2 = (6 + 1) / 2 = 3.5

Median = average of 3rd and 4th scores:
(19 + 18) / 2 = 18.5

The Mode

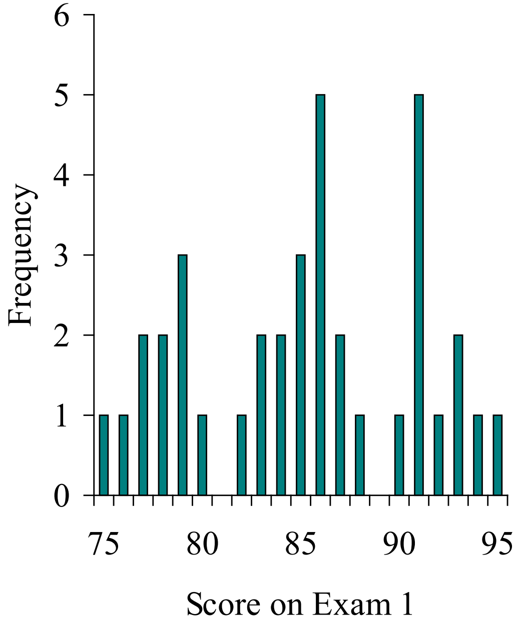
The mode is the score that occurs most frequently in a set of data

The value with the greatest frequency on the distribution.



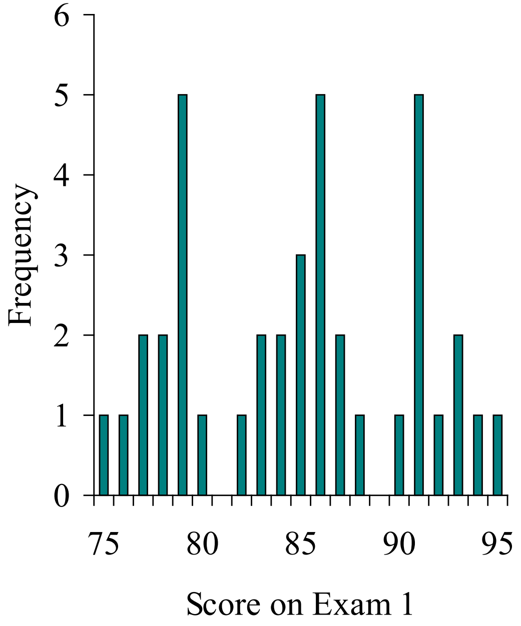
Bimodal Distributions

When a distribution has two "modes," it is called *bimodal*



Multimodal Distributions

If a distribution has more than 2 “modes,” it is called *multimodal*

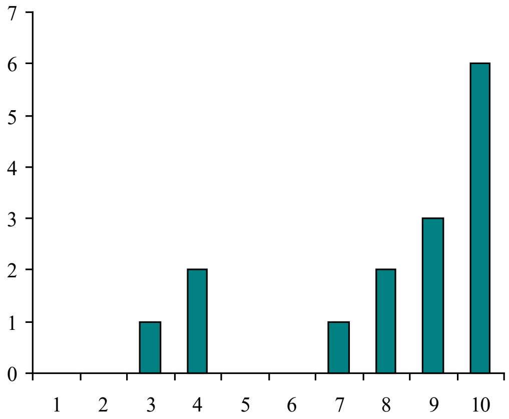


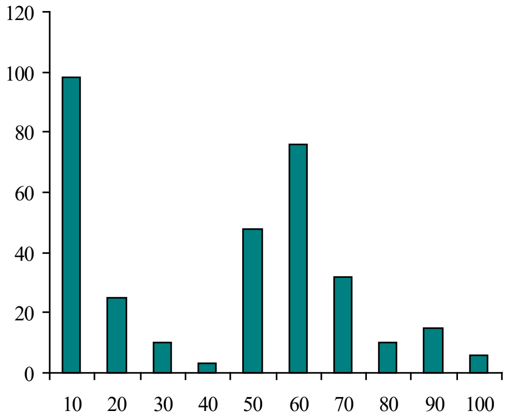
When To Use the Mode

The mode is not a very useful measure of central tendency for numerical/quantitative variables

It is insensitive to large changes in the data set

- That is, two data sets that are very different from each other can have the same mode





When To Use the Mode

The mode is **primarily used with nominal variables** that use numerical values as labels

Example:

- 3, 7, 5, 13, 20, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29
- In order these numbers are:
- 3, 5, 7, 12, 13, 14, 20, **23, 23, 23, 23**, 29, 39, 40, 56
- In this case the mode is 23.

Mean or Median

The median is less sensitive to outliers (extreme scores) than the mean

Therefore, it is a better measure than the mean for data that is strongly influenced by outliers

- E.g. family income.

Mean or Median

Example

- Sample: 20, 30, 40, and 990
- Mean is (20+30+40+990)/4 =270.
- Median (990, 40, 30, 20)
  - (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.

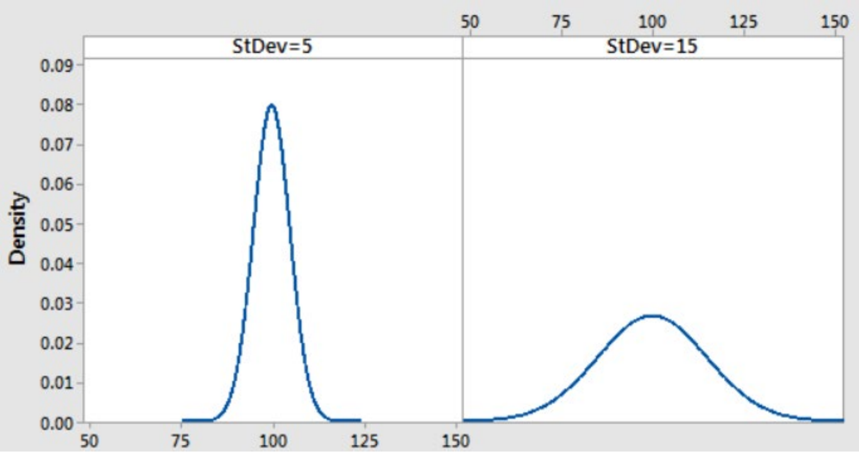
Measures of Dispersion/ Variability

Variability/Dispersion

How spread out are values of a variable ?

Reported with a measure of central tendency

Why is it important to know about variability?



When the population variability is small

- All of the values for a variable are clustered close together
- Any individual value or sample will necessarily provide a good representation of the entire set.

When population variability is large and values are widely spread

- It is easy for one or two extreme values to give a distorted picture of the general population.

Why is the study of variability important?

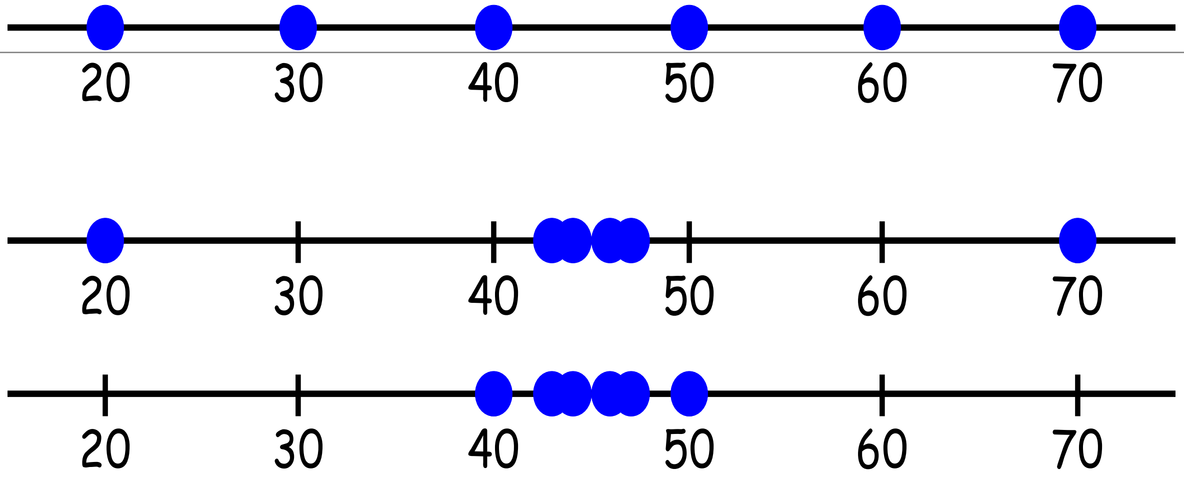
Variability (or dispersion) measures the amount of scatter in a dataset.

There is variability in virtually everything

Allows us to distinguish between usual and unusual values

Reporting only a measure of centrality doesn't provide a complete picture of the distribution.

Disappointed Dad finds single Wotsit in his packet of crisps
Source: https://metro.co.uk/2019/09/10/disappointed-dad-finds-single-wotsit-packet-crisps-10715789/



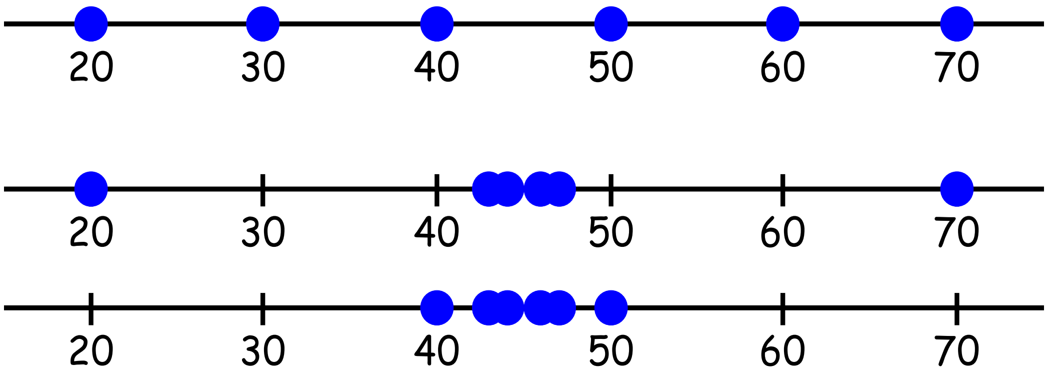
Notice that these three data sets all have the same mean and median (at 45), but they have very different amounts of variability.

Measures of Variability/Dispersion

The simplest numeric measure of variability is range.

- Its a crude measure of variability.

Range = largest observation – smallest observation



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

Measures of Variability/Dispersion

A common measure of the variability in a data set uses the deviations from the mean (*x* – *x*).

If we have a sample of 6 fish that we caught from the lake . . .
Suppose they were the following lengths:

3”, 4”, 5”, 6”, 8”, 10”

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

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

(3-6)+(4-6)+(5-6)+(8-6)+(10-6)=-3-2-1+2+4=0

Is this helpful?

Measures of Variability/Dispersion - Variance

A sample of 6 fish that we caught from the lake . . .
They were the 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?

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

Can we find an average deviation?

(3-6)+(4-6)+(5-6)+(8-6)+(10-6)=-3-2-1+2+4=0

Measures of Variability/Dispersion - Variance

Uses the square of deviations from the mean ($x - x$).

A sample of 6 fish that we caught from the lake . . .
They were the following lengths:

3”, 4”, 5”, 6”, 8”, 10”

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

What can we do to the deviations so that we could find an average?

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

Find the variance of the length of fish.

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

First square the deviations

What is the sum of the deviations squared?

Divide this by 5 to get the variance.

$s^2 = 6.8$

Measures of Variability/Dispersion

Variance: concerned with deviations from the mean (X-μ)

- First subtract the mean from each of the values gives use a *deviate* or a *deviation value* - how far a given value is from the typical, or average, value
- Then *square* the result
    - If we just added up the differences from the mean the negatives would cancel the positives
    - If we used absolute values we wouldn't get an accurate measure of spread
    - Squaring is the best option
    - *Variance* is defined as the average of the deviations from the mean squared:

$$s^2 or\sigma^2 = \frac{\sum(X-\mu)^2}{N\ -1}$$

N-1 here is the degrees of freedom = the number of independent pieces of information on which the estimate is based

Measures of Variability/Dispersion

Degrees of Freedom

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

Population variance is denoted by s^2.

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

Degree of freedom

Measures of Variability – Standard Deviation

The square root of variance is called **standard deviation**.

A typical deviation from the mean is the **standard deviation.**

This is the more common measure since it makes it will be in the units of measurement rather than squared units of measurement – so is more intuitive

Measures of Variability – Standard Deviation

Our fish example: s^2 = 6.8 inches^2 so standard deviation = 2.608 inches

The fish in our sample deviate from the mean of 6 by an average of 2.608 inches.

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 df for the estimate, you have to subtract 1 from the number of items.

Degrees of freedom Why n-1?

Fundamentally it gives you a better estimate

We are working with a sample to make a statement about the population

It is unlikely that the standard deviation of a sample will give us the exact population standard deviation (or even be very close to it).

The standard deviation is affected by the number of cases used to calculate it.

Degrees of freedom Why n-1?

Imagine a population that has substantial variability in it e.g. the distribution of 23,000

students' ages at a large university.

There will be some unusually young students in the population and there will be some unusually old students

in the population.

If you selected a sample of students, it would be hard to ensure you collect all the variability that actually exists

in the population.

Most of your sample cases would likely come from the portion of the population that has most of the cases to begin with (say 20 to 25)

Your sample probably wouldn't reflect all of the variability that really exists in the population.

So your sample standard deviation would likely be slightly less than the true standard deviation of the population.

Degrees of freedom Why n-1?

We need to make some adjustment

Statisticians deal with this situation by using a small correction factor.

When calculating the standard deviation of a sample (or the variance of a sample, for that matter), they change the n in the denominator to n – 1.

This slight reduction in the denominator results in a larger standard deviation—one that better reflects the true standard deviation of the population.

When you use n – 1 in the denominator of the formula for the standard deviation of a sample, you not only slightly increase the final answer (or value of the standard deviation), you do so in a way that is sensitive to sample size.

The smaller the size of the sample, the more of an impact the adjustment will make.

Measures of Dispersion

Standard Deviation: the most useful and the most popular measure of dispersion.

- Concept was first introduced by Karl Pearson in 1893.
- Standard deviation = the **square root** of the **Variance**
- Use Greek symbol sigma σ
- The larger the value the more spread out around the mean the data is, smaller means less spread.
- 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.
- Allows us to see how spread out *on average* individual cases are from the mean

Measures of Variability – Interquartile Range

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

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

The IQR is resistant to extreme values

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

Find the interquartile range for this set of data.

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

52 values
Median is the 26th value = 26

First put the data in order & find the median.

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

13th value =24

Find the lower quartile (Q_1) by finding the median of the lower half.

39th value =30

Find the upper quartile (Q_3) by finding the median of the upper half.

iqr = IQ3 –IQ1

=30 – 24 = 6

Measures of Dispersion

Measures of Dispersion are descriptive statistics that describe how similar a set of values are to each other (or the range of values)

- The more similar the values are to each other, the lower the measure of dispersion will be
- The less similar the values are to each other, the higher the measure of dispersion will be
- In general, the more spread out a distribution is, the larger the measure of dispersion will be

Needs to be considered in relation to the measure of central tendency

Summary

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).

What do I need to describe for interval/ratio data?

Centre

- Discuss where the middle of the data falls
- Measures of central tendency
  - mean, median and mode

Spread/Dispersion

- Discuss how spread out the data is
- Refers to the variability in the data
  - Range, standard deviation, IQR

Shape/Pattern

- Refers to the overall shape of the distribution
- Symmetrical, uniform, skewed, or bimodal
- We will talk about shape next

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.

Visualization

Sources used in creation of this lecture:
Discovering Statistics Using R Field, Miles and Field;
Understanding Basic Statistics, Brase and Brase;
Statistics and Data Analysis, Peck, Olsen and Devore

Plots and Graphs

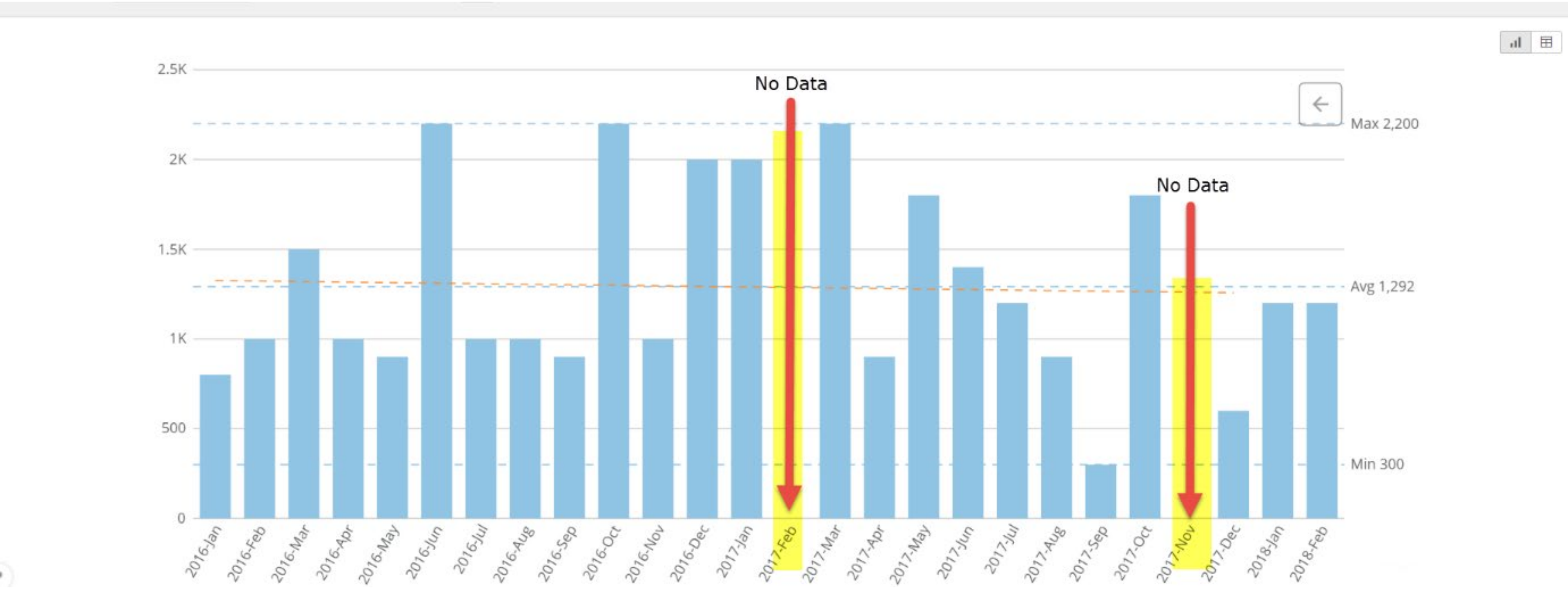
The Art of Presenting Data

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, 2nd edition,2001

Examples



Missing bars for years from a timeline

Graphical Presentation – Choose the correct type for each variable

Interval or Ratio numerical data

- Histogram
- Stem and Leaf diagrams
- Box-plot
  - Depending on dispersion

KEY
SLIDE

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.

Creating 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

Step 1: Group the data by tens digits.

Step 2: Order the data from least to greatest.

75 79

83 84 86 86 88

91 94 99

Creating Stem-and-Leaf Plots

Step 3: List the tens digits of the data in order from least to greatest. Write these in the "stems" column.

75 79

83 84 86 86 88

91 94 99

Step 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.

Step 5: Title the graph and add a key.

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 (40 + … + 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 Plot

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

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

Constructed differently for discrete versus continuous data

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

Queen honey bees mate shortly after they become adults.

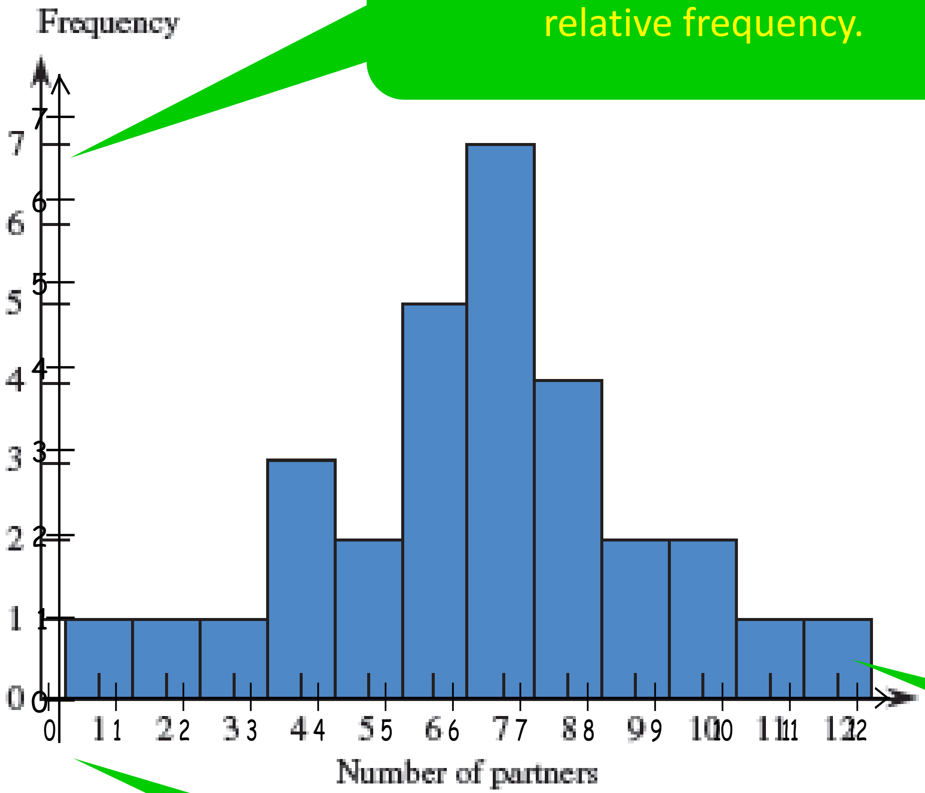
During a mating flight, the queen usually takes several partners.

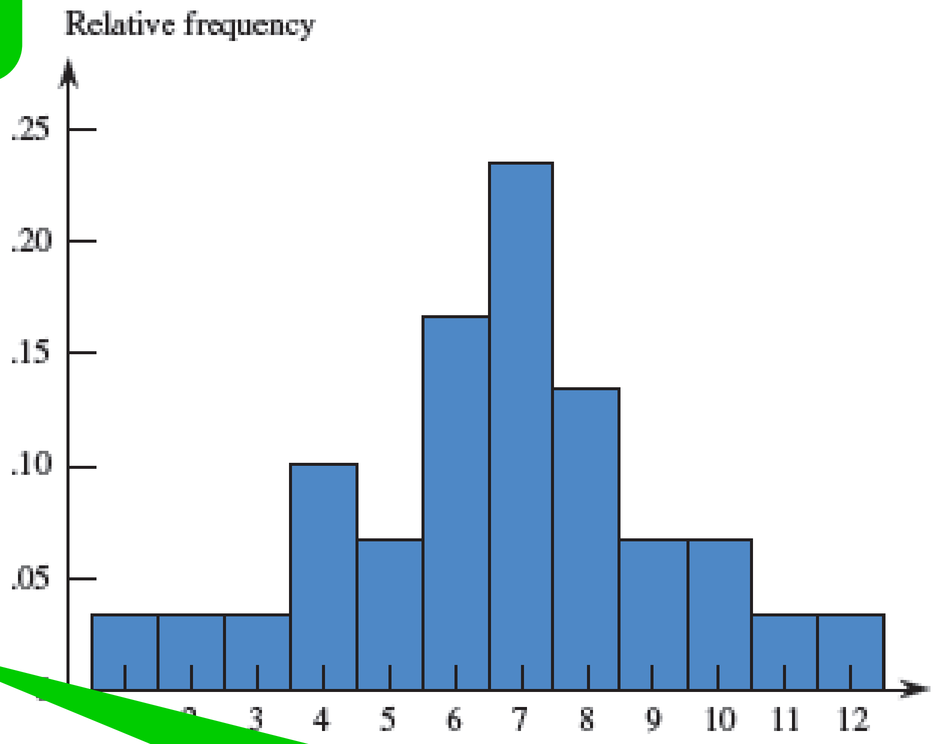
A study on honey bees provided the following data on the number of partners for 30 queen bees.

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

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

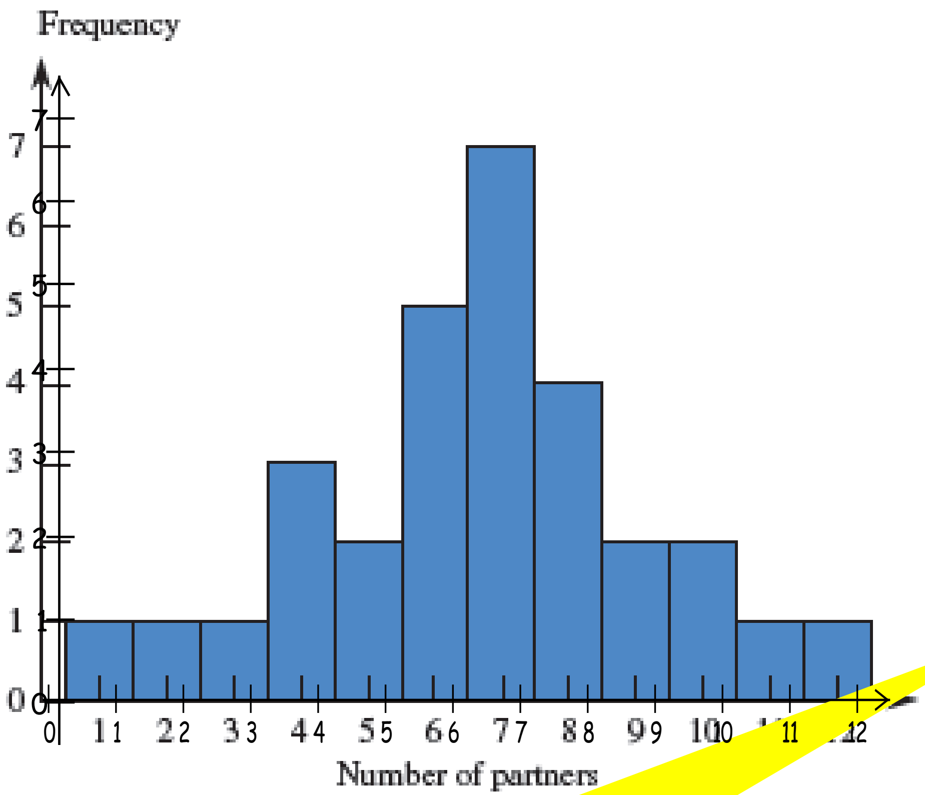
Next draw a vertical axis, scaled with frequency or relative frequency.

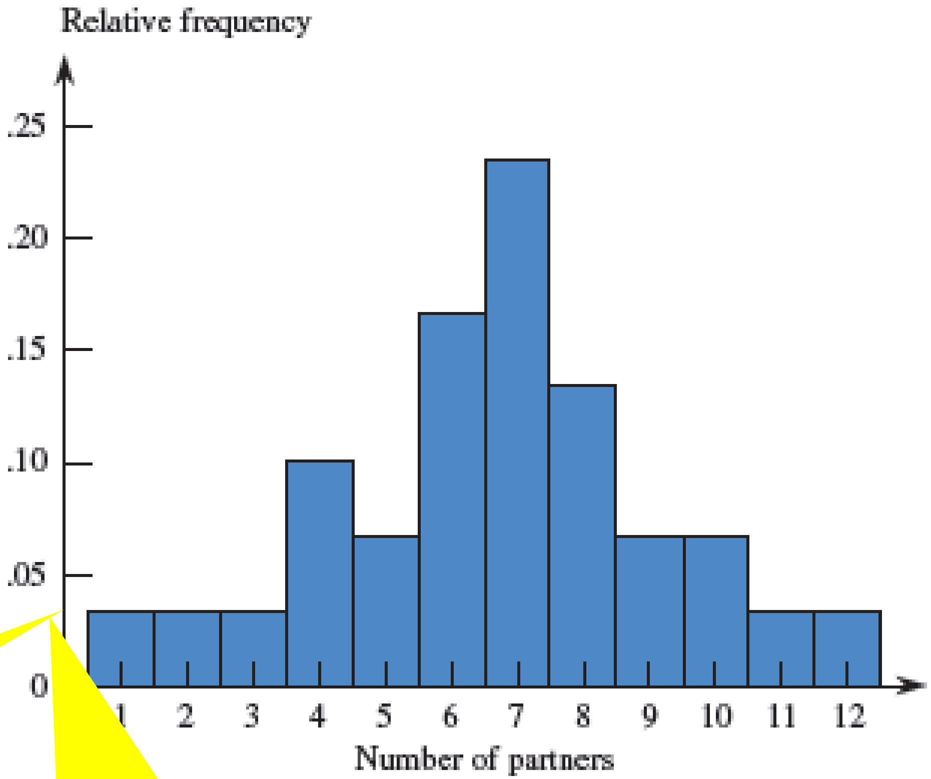




Draw a rectangle above each value with a height corresponding to the frequency.

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





Suppose we use relative frequency instead of frequency on the vertical axis.

We can see how often something happens divided by all outcomes

This is the type of histogram that most people are familiar with.

Histograms

When to Use

- Univariate numerical data (one variable)

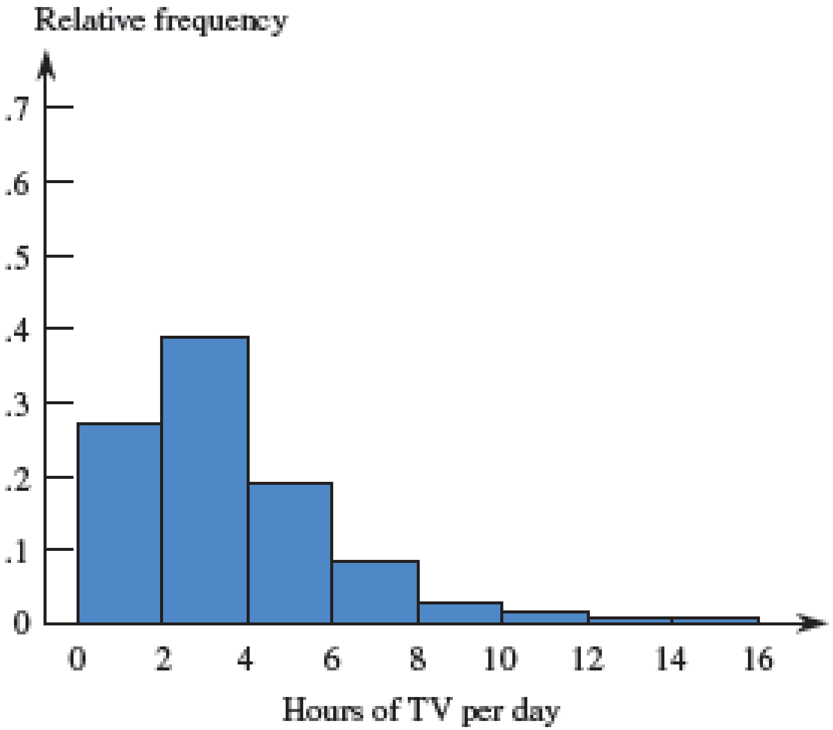
Continuous data

- 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

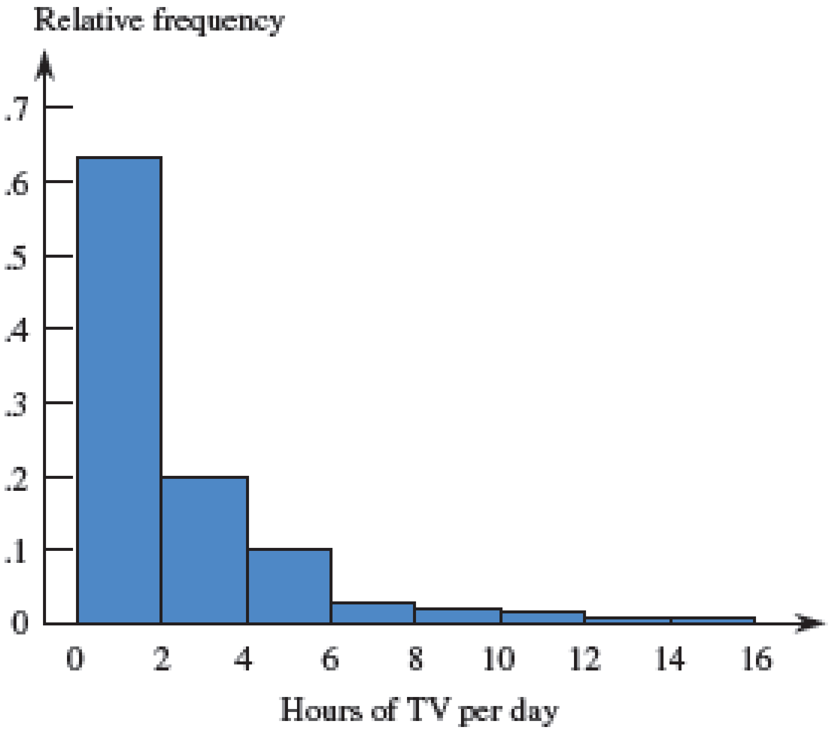
To describe

- Comment on the centre spread, and shape of the distribution and if there are any unusual features

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.



Children Age 1



Children Age 3

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)

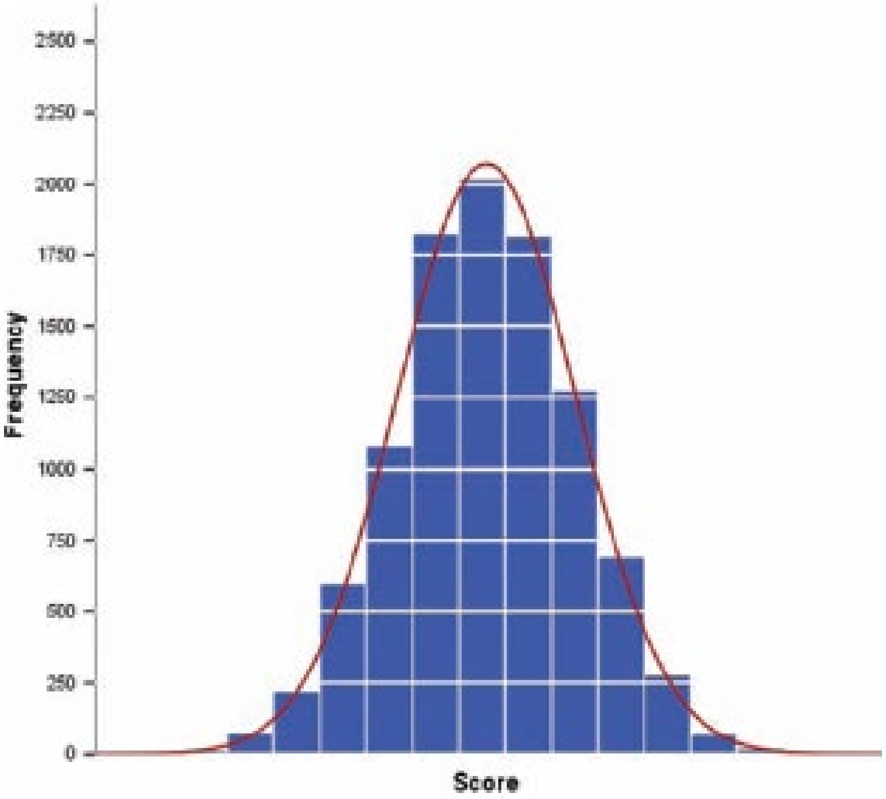
Skewed

- Lack of symmetry
- Data pulled towards one end of the graph

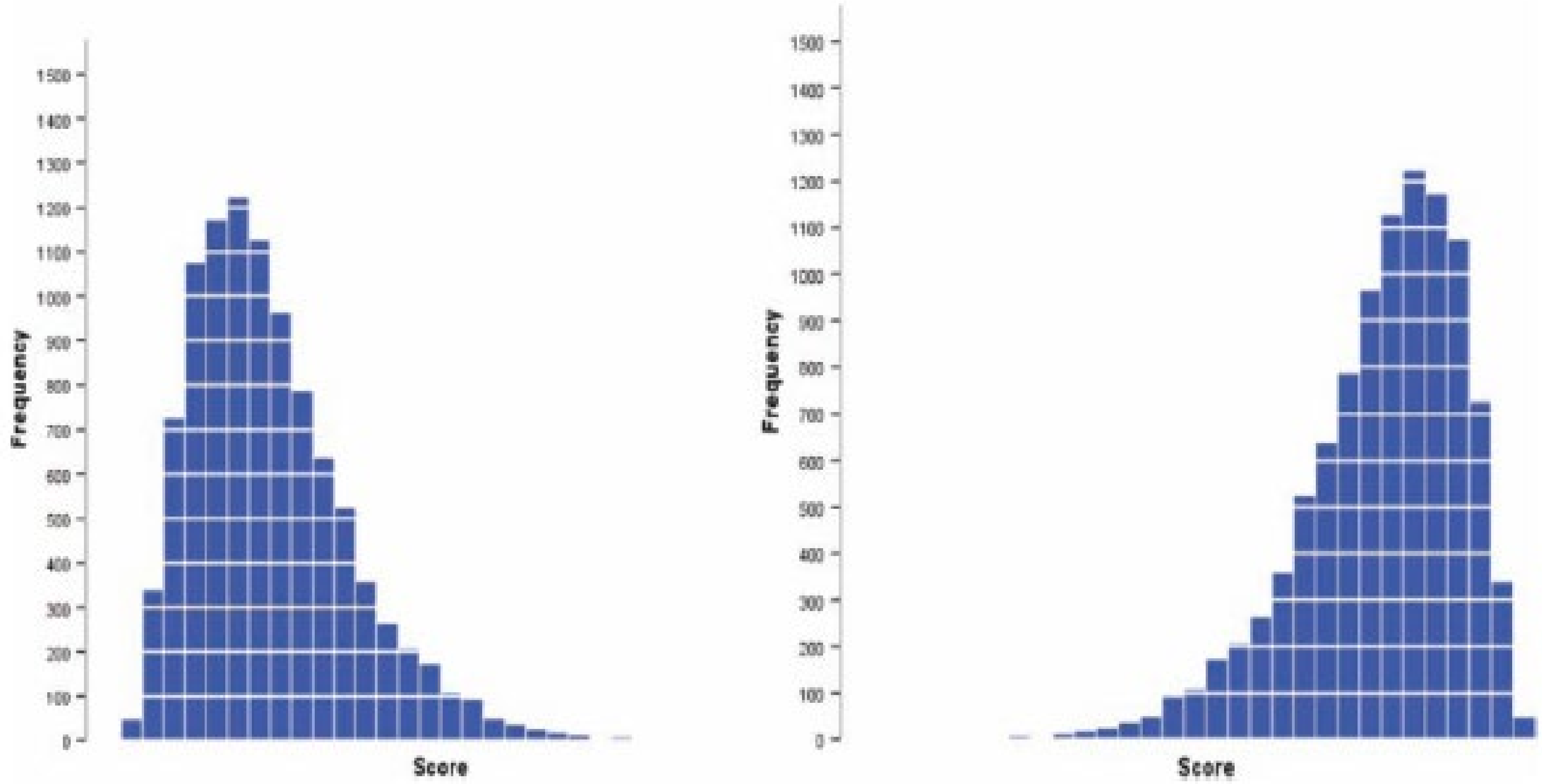
Kurtosis

- Pointyness

The Normal Distribution



Skew



A positively (left figure) and negatively (right figure) skewed distribution

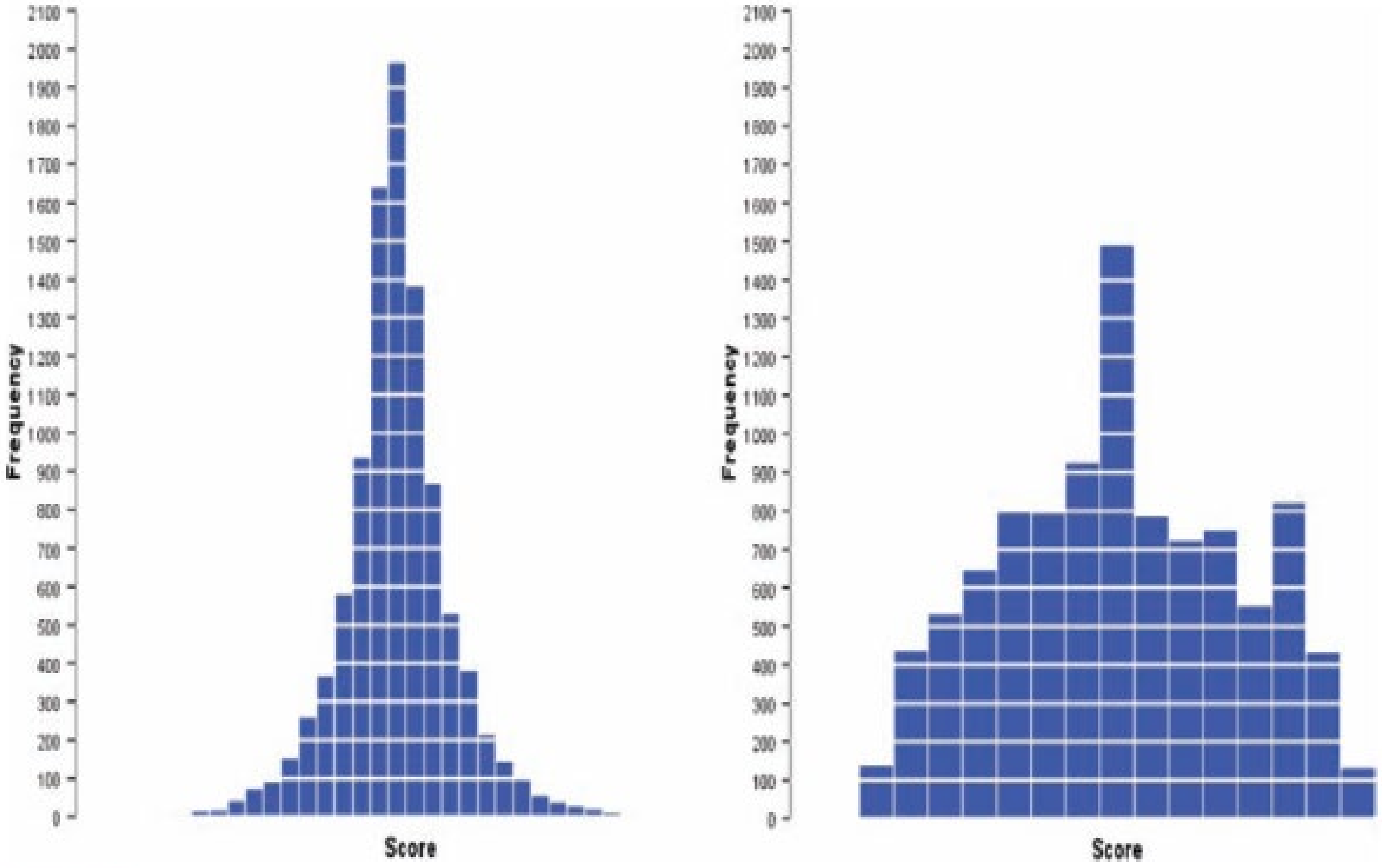
Properties of Frequency Distributions

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).

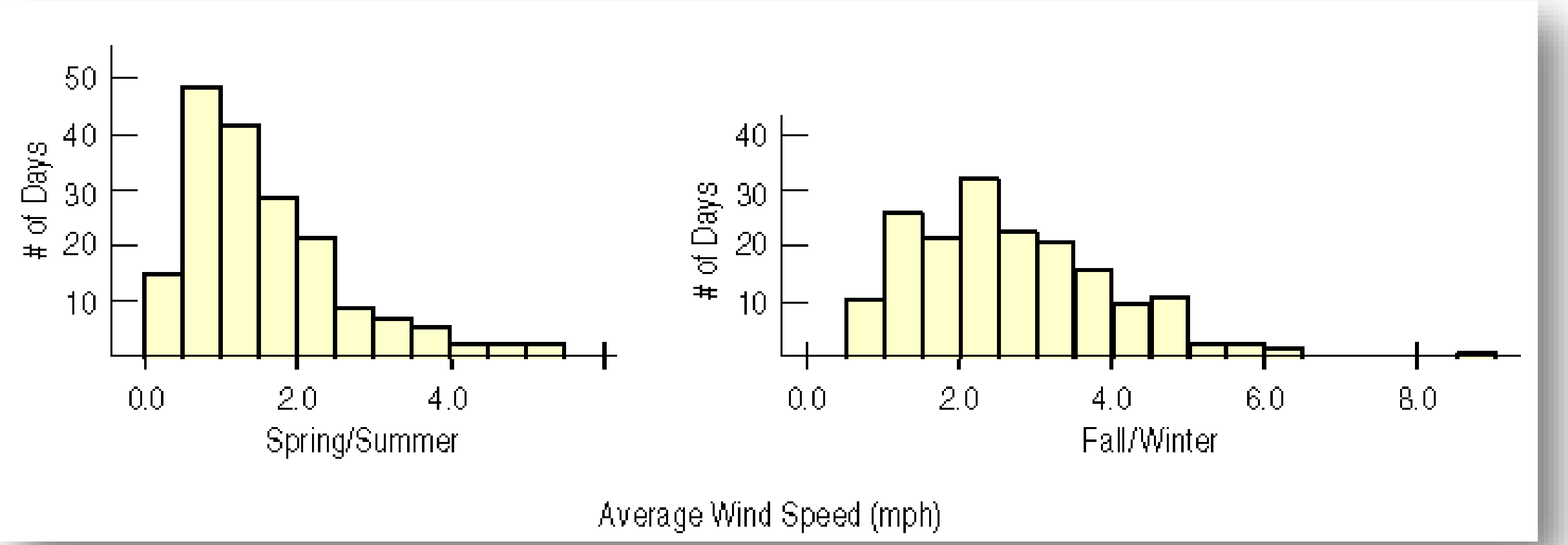


Distributions with positive kurtosis (leptokurtic, left figure) and negative kurtosis (platykurtic, right figure)

Comparing Groups

It is almost always more interesting to compare groups.

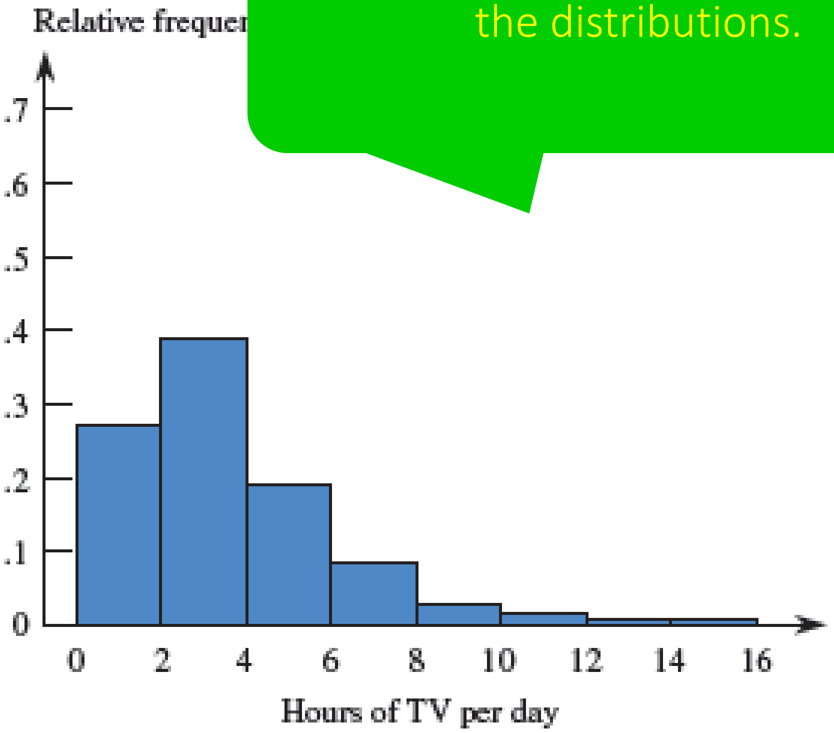
With histograms, note the shapes, centers, and spreads of the two distributions.



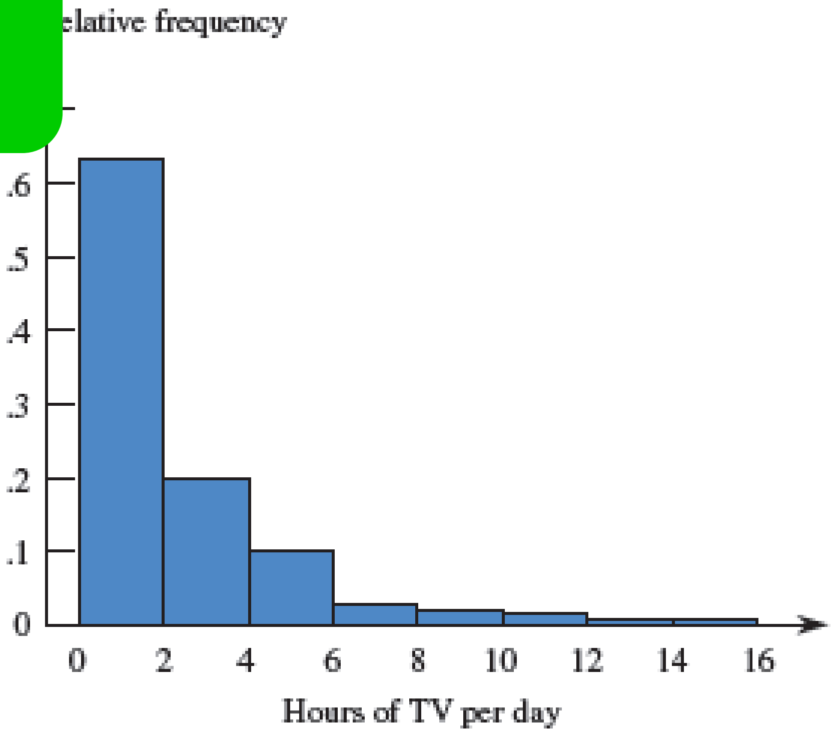
What does this graphical display tell you?

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.

Write a few sentences comparing the distributions.

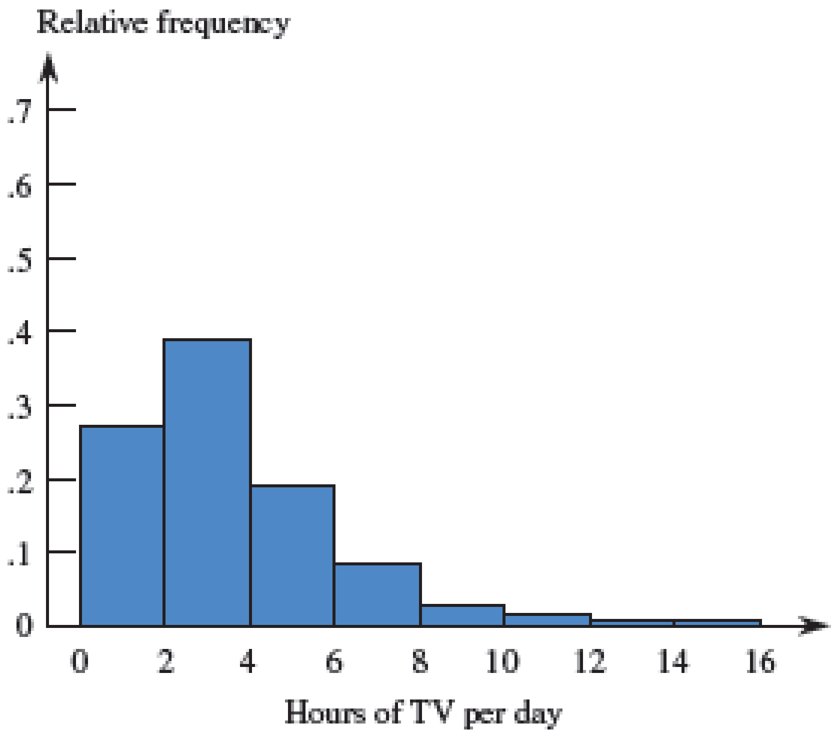


Children Age 1

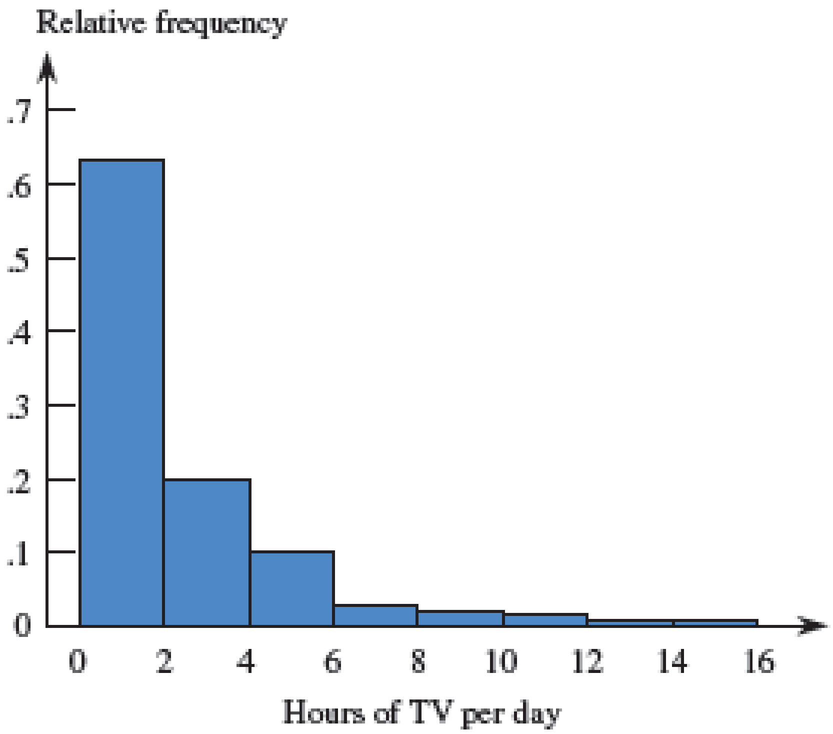


Children Age 3

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 distributions had similar ranges.

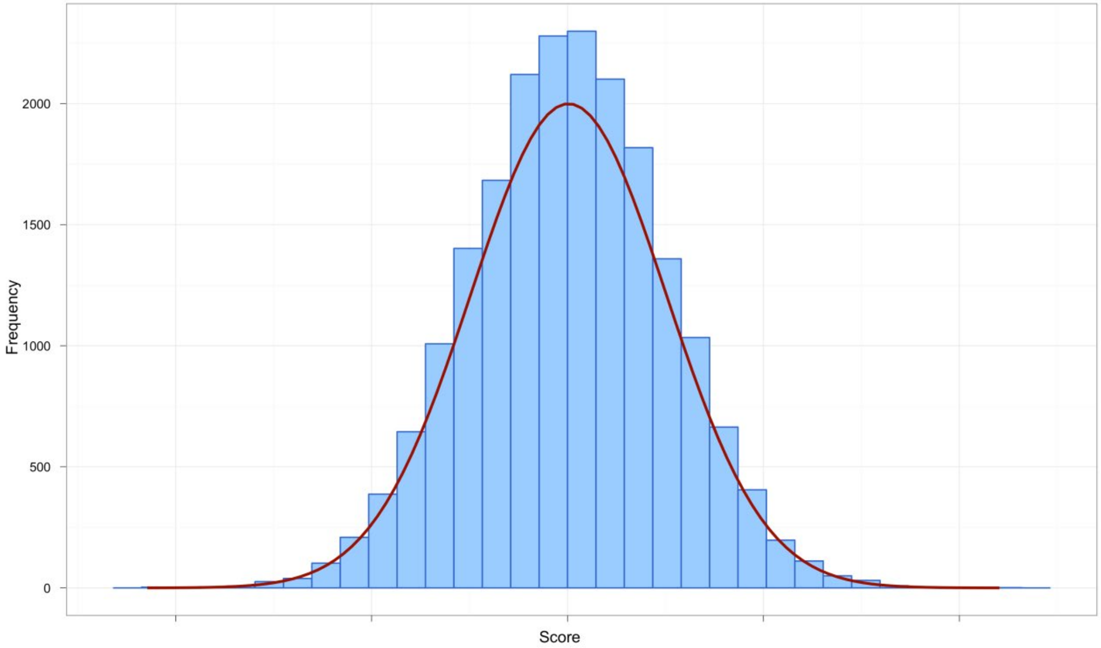


Children Age 1



Children Age 3

The Normal Distribution



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

Density Curve/Density Plot

A *density curve* describes the overall pattern of a distribution.

A density curve is a graphical representation of a numerical distribution where the outcomes are continuous.

- A smoothed version of a histogram
- Aims to (smoothly) show how data is spread out

The **Kernel Density Estimation (KDE) function** chosen provides a way to estimate the probability distribution of your data

- Gives you a smooth curve that represents where data points are likely to be found.

The area under a density curve represents **probability**.

The area under a density curve = 1.

Density Curve/Plot

Example:

Suppose we have a crowded room full of people standing at different heights.

We can create a **density plot** is like a smooth map that shows where most people are standing and where fewer people are.

Data Points: Individual heights (people in the room).

Kernel: The shape of each person's light (e.g., bell-shaped).

KDE Function: The process of adding up all the individual lights.

Density Plot: The final brightness map showing height distribution.

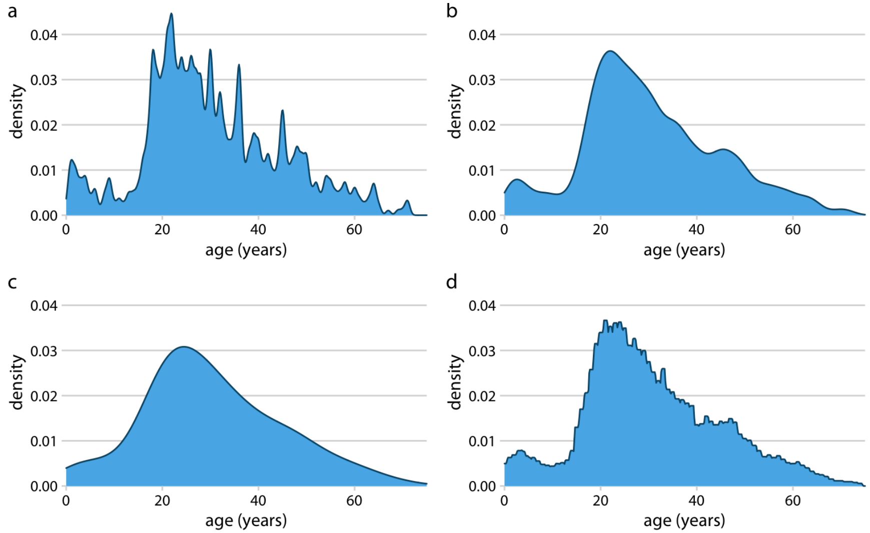
1. **Collect Data:**
   - Imagine you have the heights of 100 people.
2. **Choose a Kernel**:
   - Let's say you choose a Gaussian (bell-shaped) kernel.
3. **Apply KDE**:
   - For each person's height, place a small bell-shaped curve centred at their height.
4. **Combine Curves**:
   - Add up all these small curves to create one smooth hill.
5. **Create Density Plot**:
   - The resulting hill shows where most people's heights are concentrated and where there are fewer.

Normal/Gaussian Density Curve

Formula used to generate the shape of the curve is the *normal density function*

A Gaussian kernel will tend to produce density estimates that look Gaussian-like (normal curve), with smooth features and tails.

By contrast, a rectangular kernel can generate the appearance of steps in the density curve



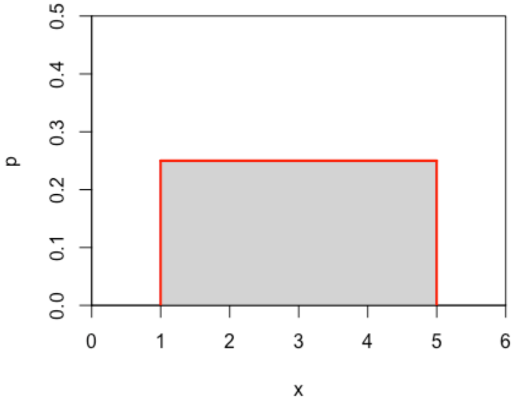
Kernel density estimates depend on the chosen kernel and bandwidth. Here, the same age distribution of Titanic passengers is shown for four different combinations of these parameters: (a) Gaussian kernel, bandwidth = 0.5; (b) Gaussian kernel, bandwidth = 2; (c) Gaussian kernel, bandwidth = 5; (d) Rectangular kernel, bandwidth = 2.

https://clauswilke.com/dataviz/histograms-density-plots.html

Density Curve

Curves are "uniform" when the probabilities for all outcomes are the same.

In a uniform distribution each outcome has the same frequency. Because of this, the height at each point on the x-axis is identical and the shape of a uniform density curve becomes a rectangle.



Density Curve

In a uniform distribution the probability of $x=a$ is zero and probability that $x < a$ is equal to the probability that $x \leq a$.

Probability is equal to the area.

For $P(x < a)$, one shades to the left of the point and thus creates a rectangle of area.

For a single point $x = a$, there would be nothing to shade left causing the rectangle of area to have no width and thus no probability.

Therefore, the probability that $x = a$ is zero, and the probability that $x < a$ must be the same as the probability that $x \leq a$.

Density Curve

Example: Given a uniform density curve with a length of 50.

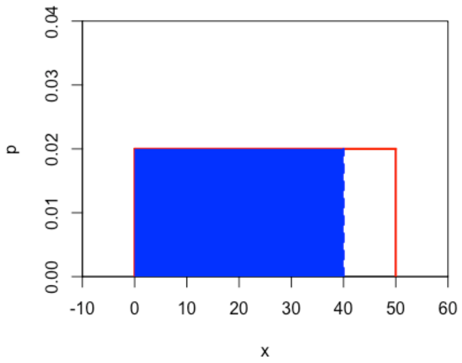
We can calculate the height of density curve. The area must equal 1, and uniform density curves are rectangular.

Thus, we can set up the equation base *x* height = 1 → (50)(h) = 1 → height = 0.02

Calculate probability P(x < 40):

Step 1) Shade in the area.

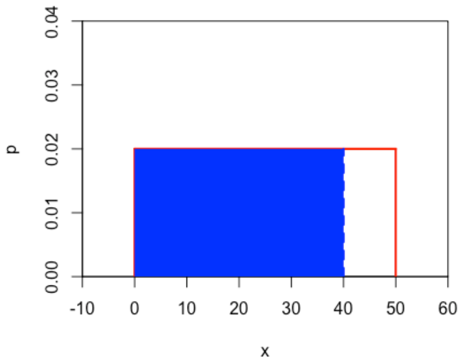
Step 2) Find the area of the shaded rectangle using the height you calculated. → area = (40)(0.02) = 0.8



Density Curve

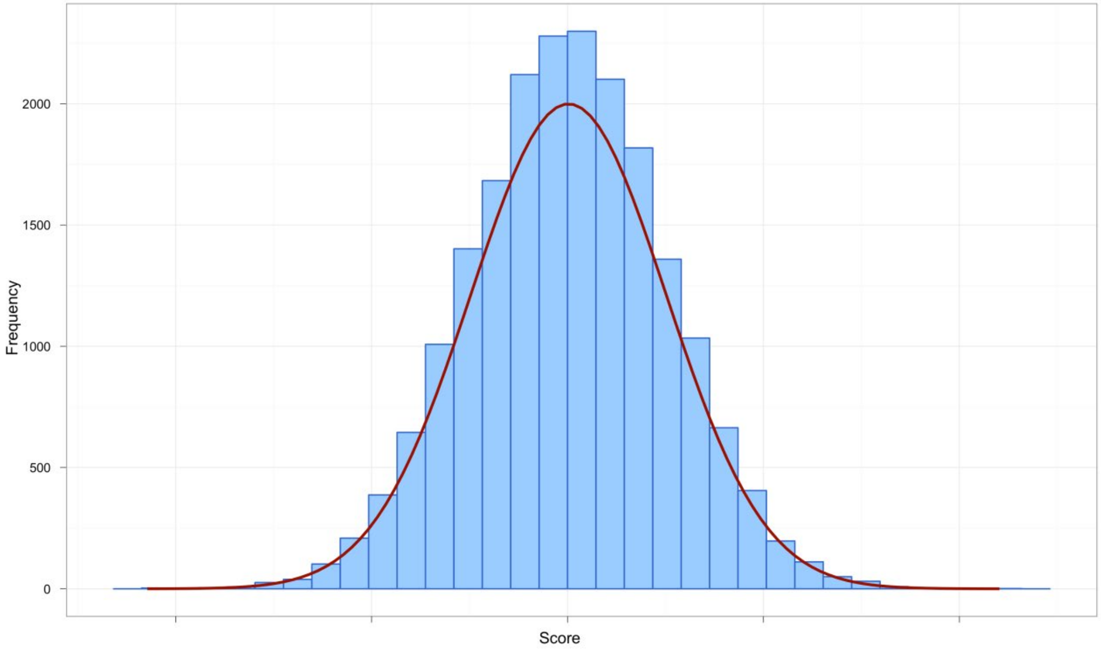
Calculate P(x = 40)

If you attempt to follow the steps from the previous slide you will see that there is no area that can be shaded in since this rectangle of area would have no width. Thus, the probability is equal to zero.

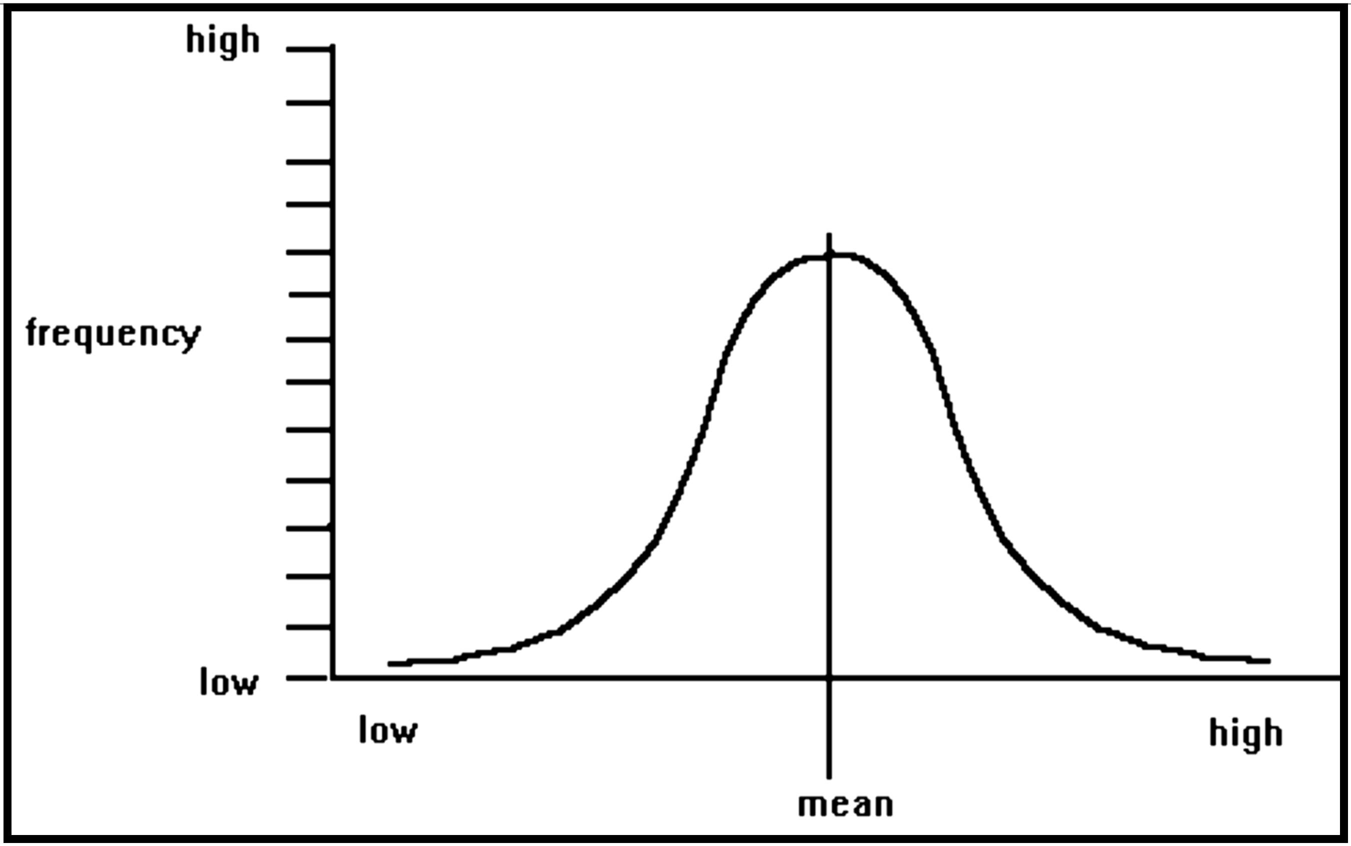


Normal Distribution

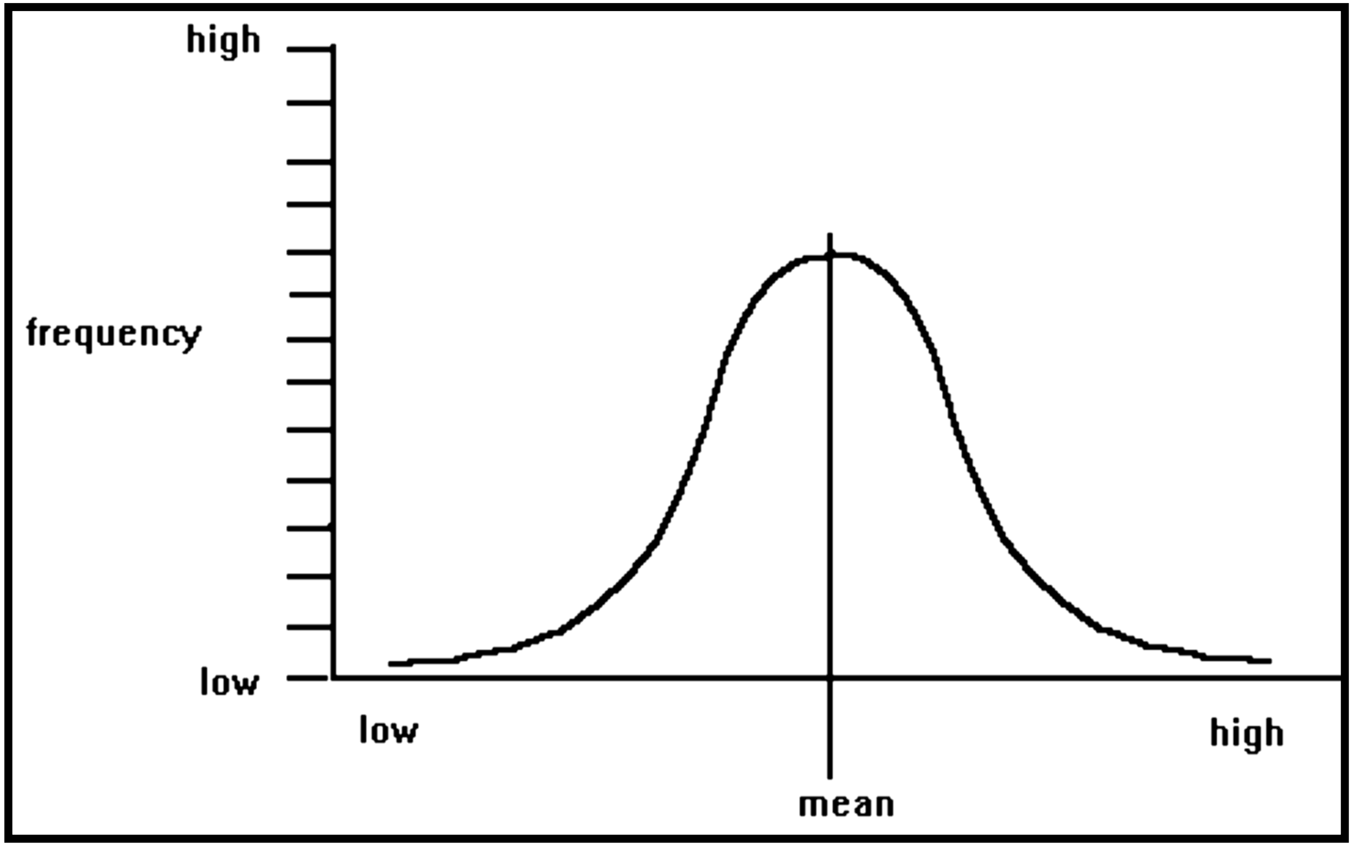
A distribution is **normal** if its density curve is symmetric, single-peaked and bell-shaped.



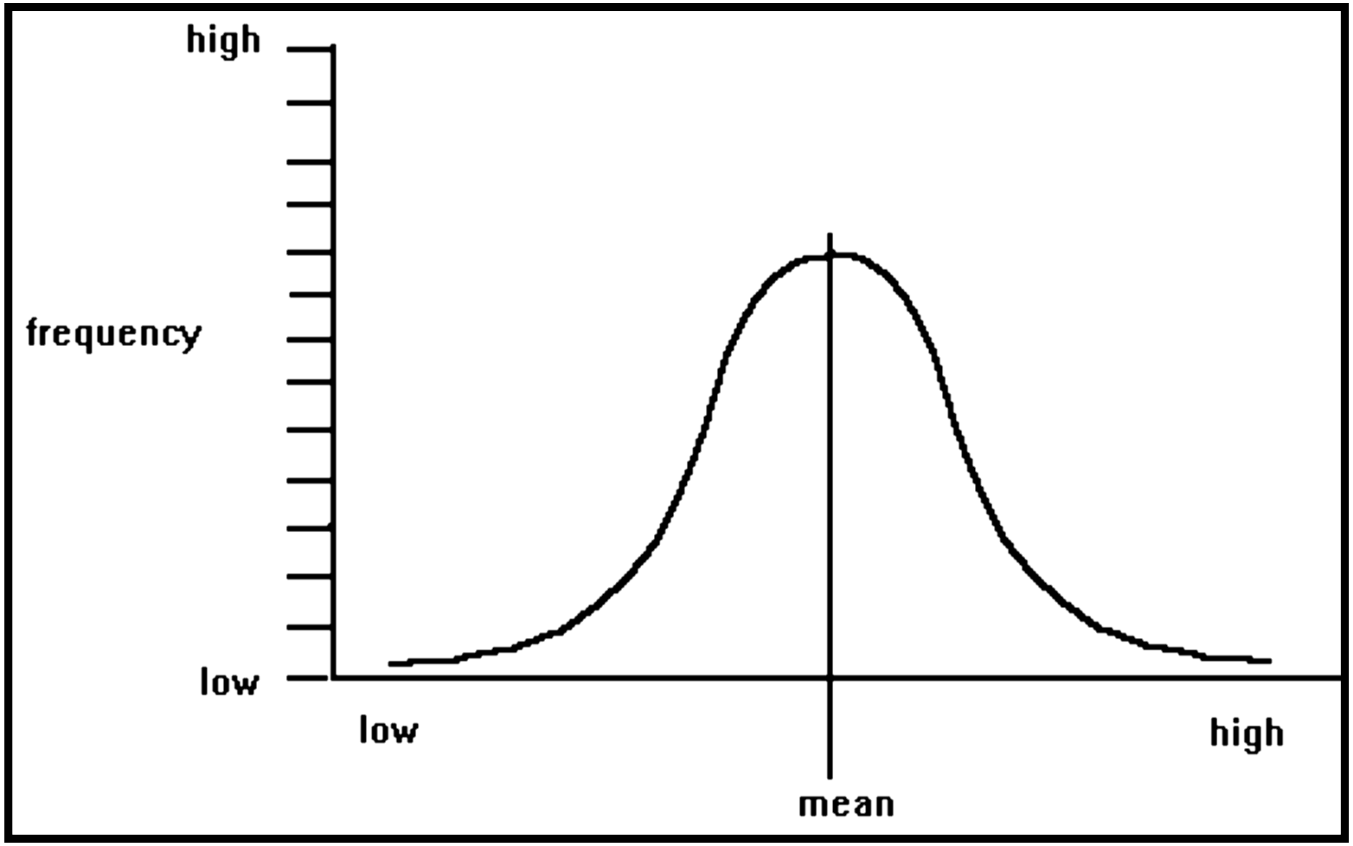
Properties of the Normal Distribution:



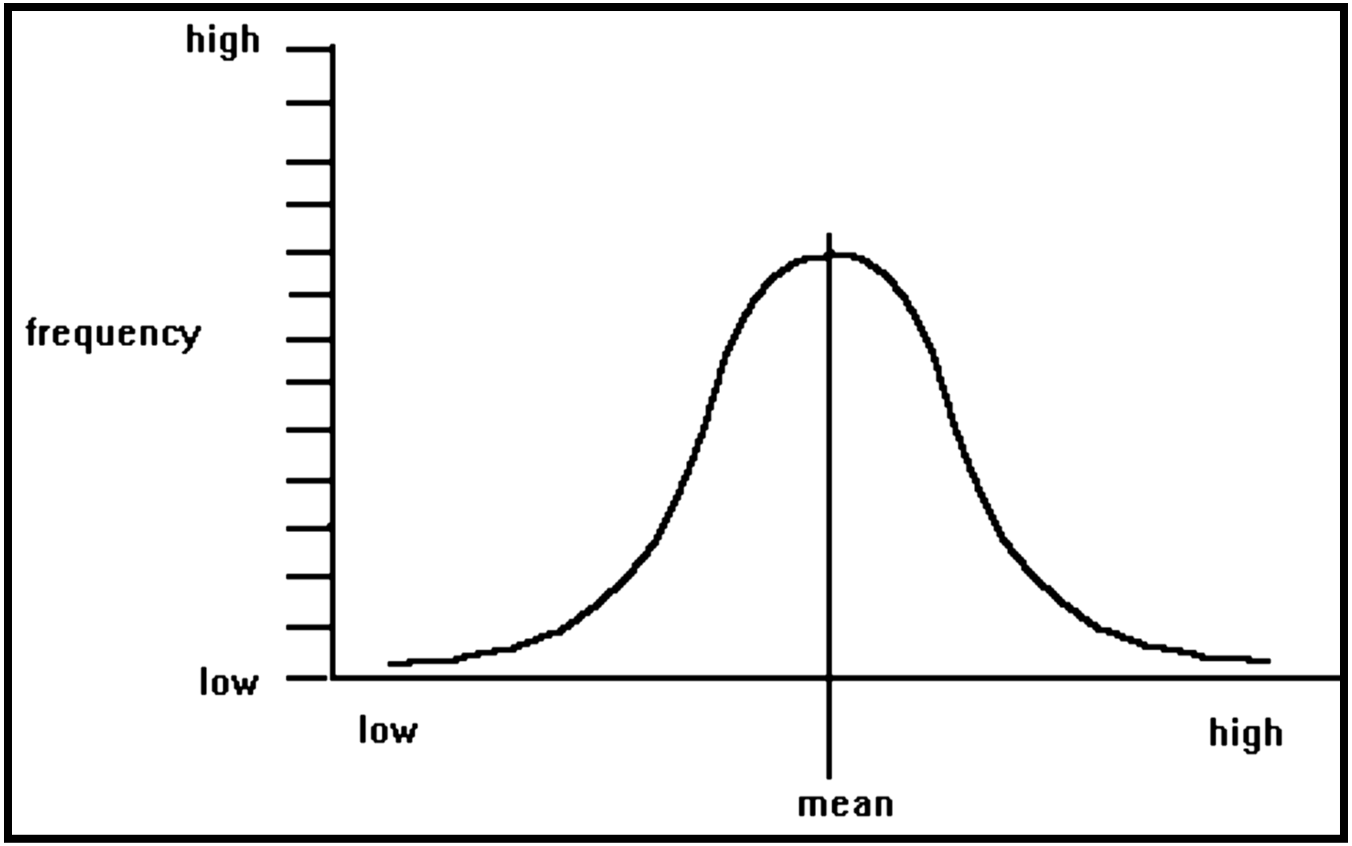
1. It is bell-shaped and asymptotic at the extremes.



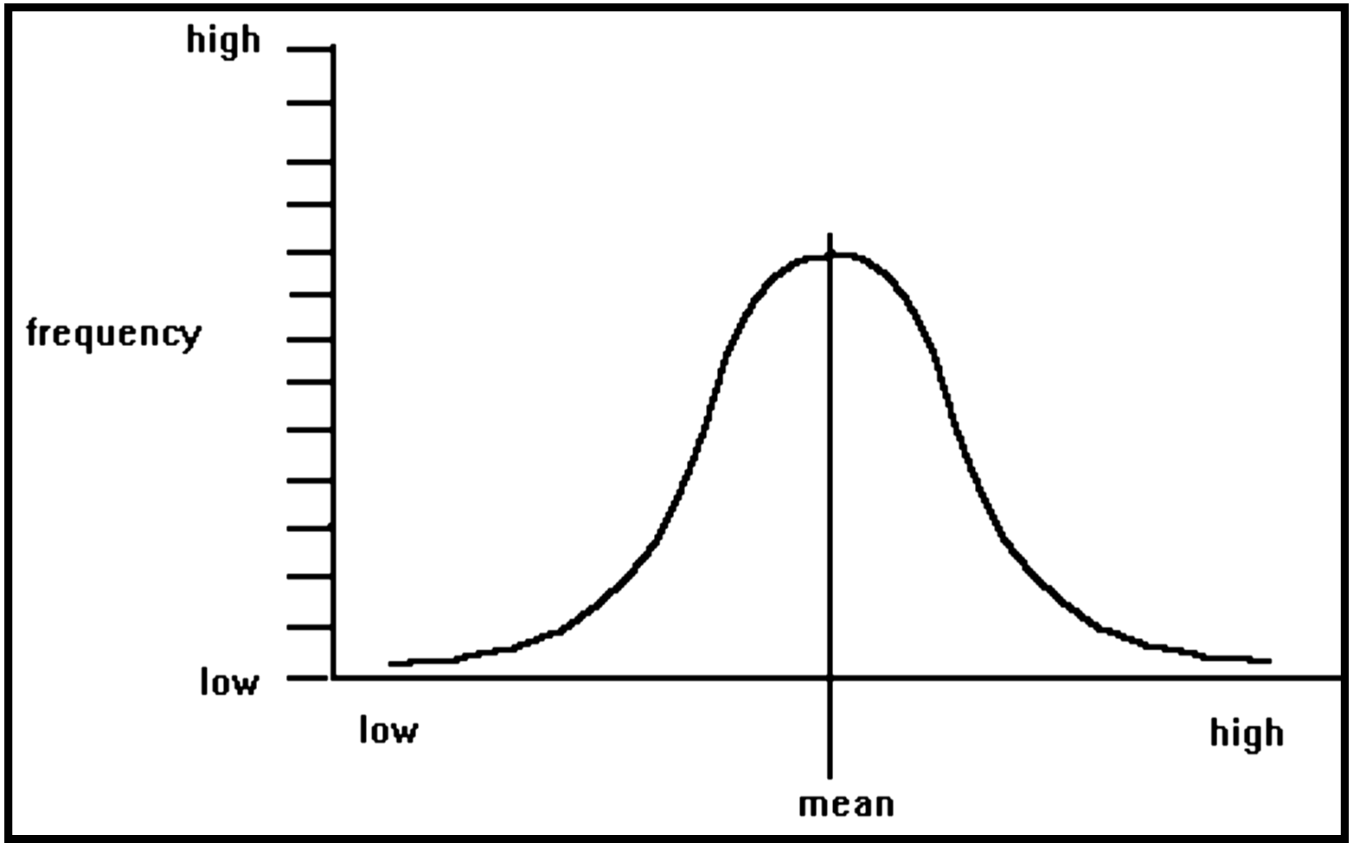
2. It's symmetrical around the mean.



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

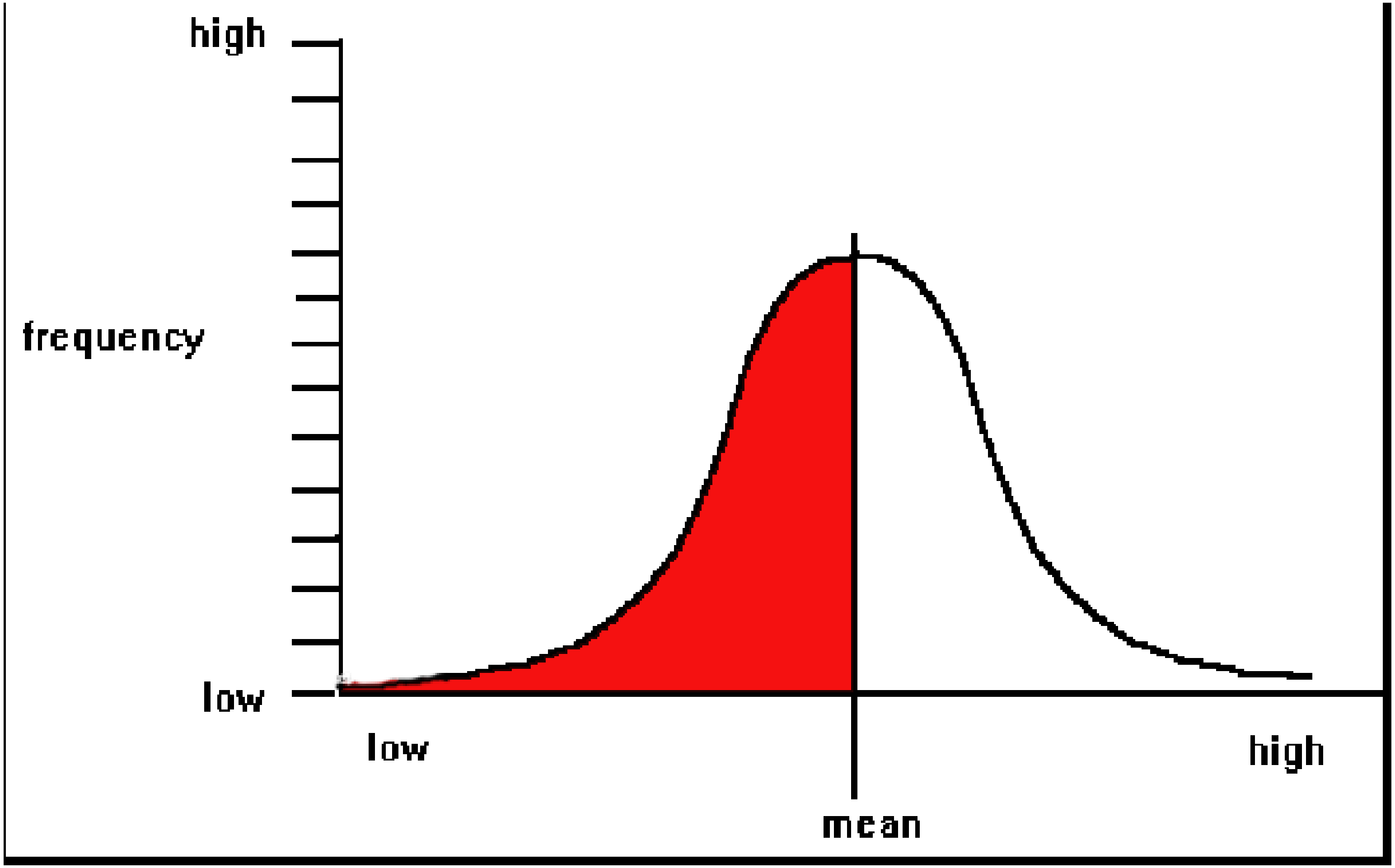


4. It can be specified completely, once mean and SD are known.

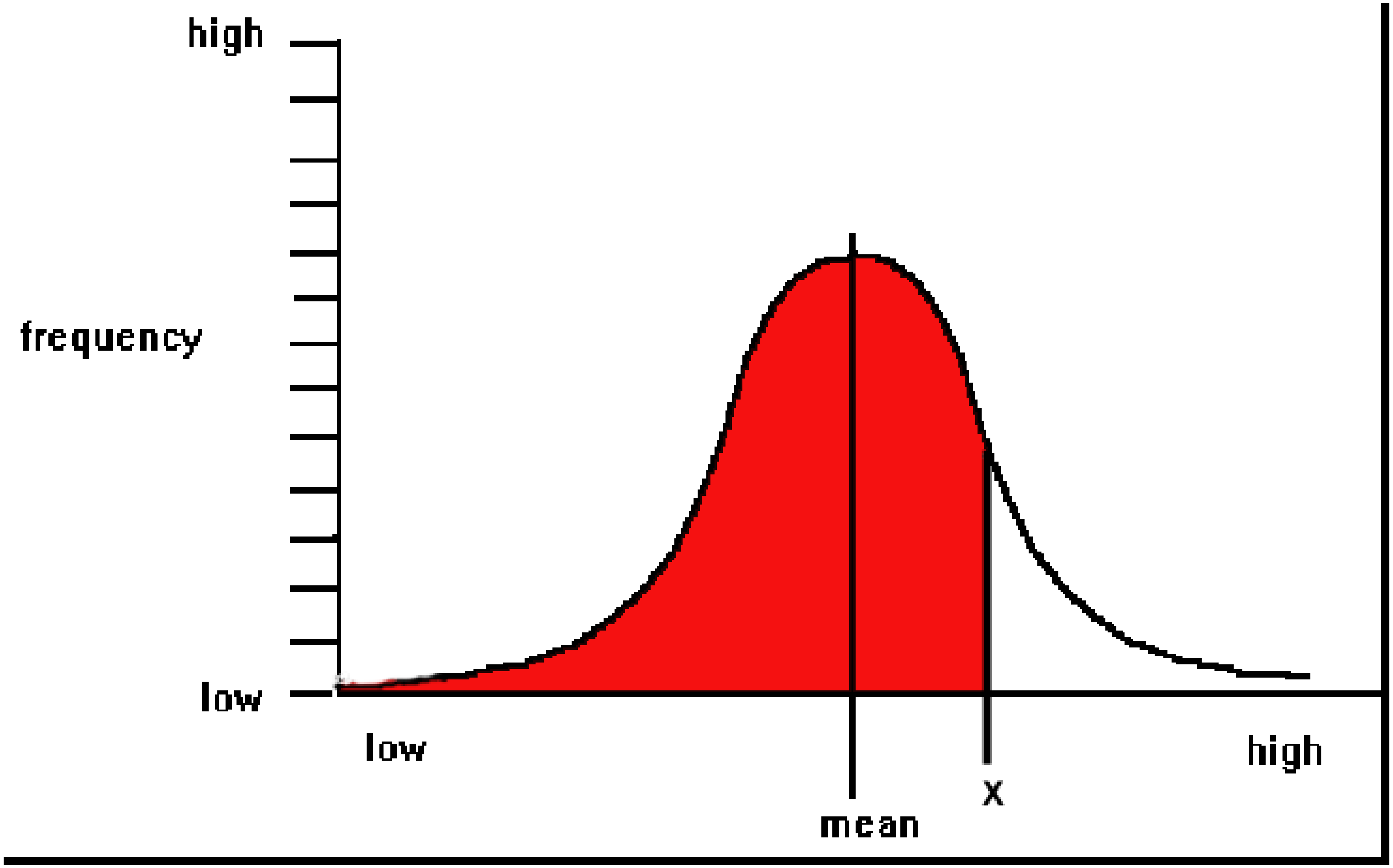


5. The area under the curve is directly proportional to the relative frequency of observations.

Thus we can calculate the probability of observations occurring in a population

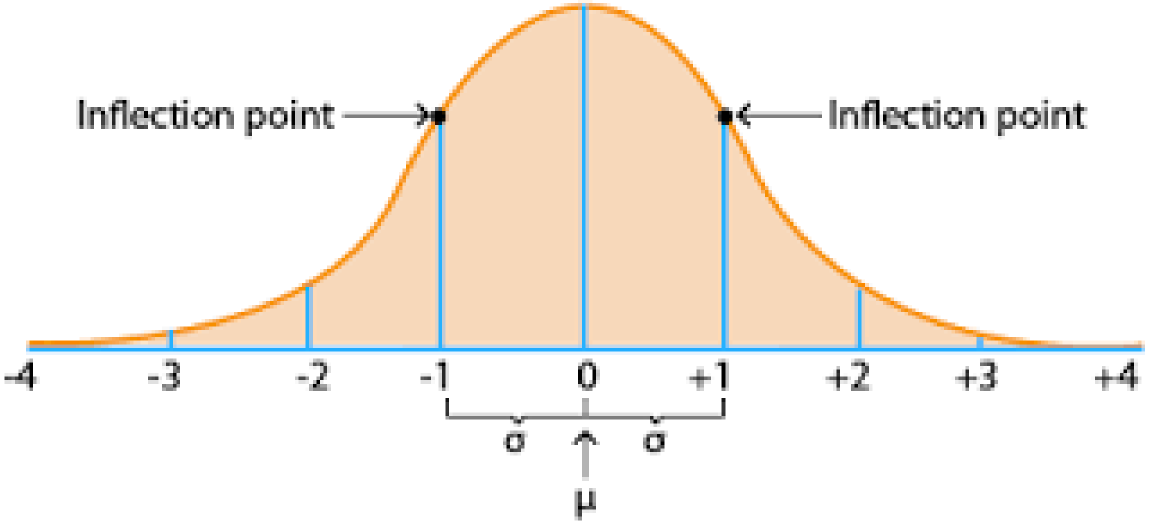


e.g. here, 50% of scores fall below the mean, as does 50% of the area under the curve.



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

Normal Distribution



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.

The 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

Normal Distribution in summary

Many psychological/biological properties are normally distributed.

This is very important for statistical inference (extrapolating from samples to populations)

Normal distribution is central to choosing the correct statistical test

Parametric Tests

- Assume a normal distribution

Non-parametric Tests

- Non normal distribution

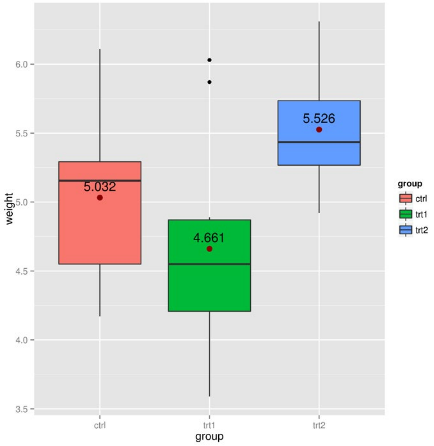
Always start by looking at the data!

Boxplot

A boxplot is a graphical display of the five-number summary.

Boxplots are useful when comparing groups.

Boxplots are particularly good at pointing out outliers.



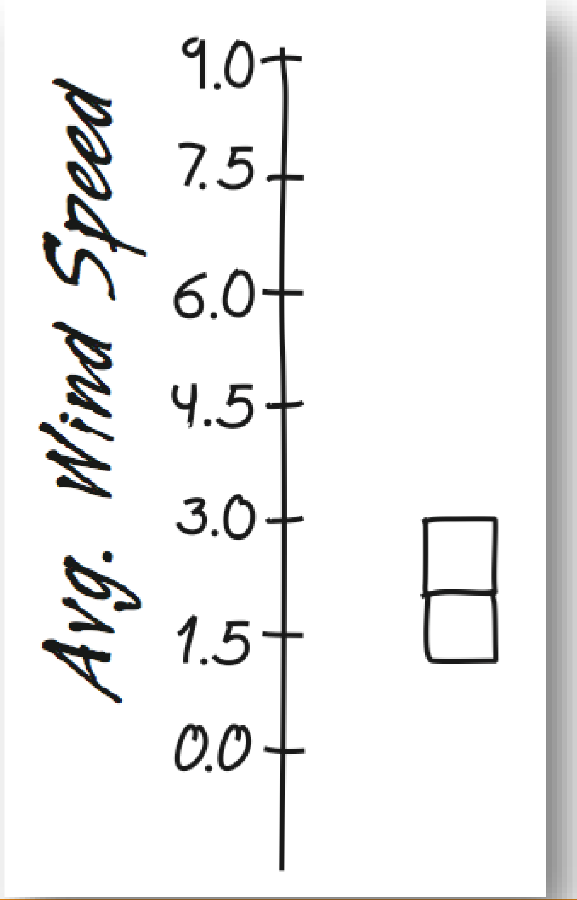
This Photo by Unknown Author is licensed under CC BY-SA

Constructing Boxplots

Draw a single vertical axis spanning the range of the data.

Draw short horizontal lines at the lower and upper quartiles and at the median.

Then connect them with vertical lines to form a box.



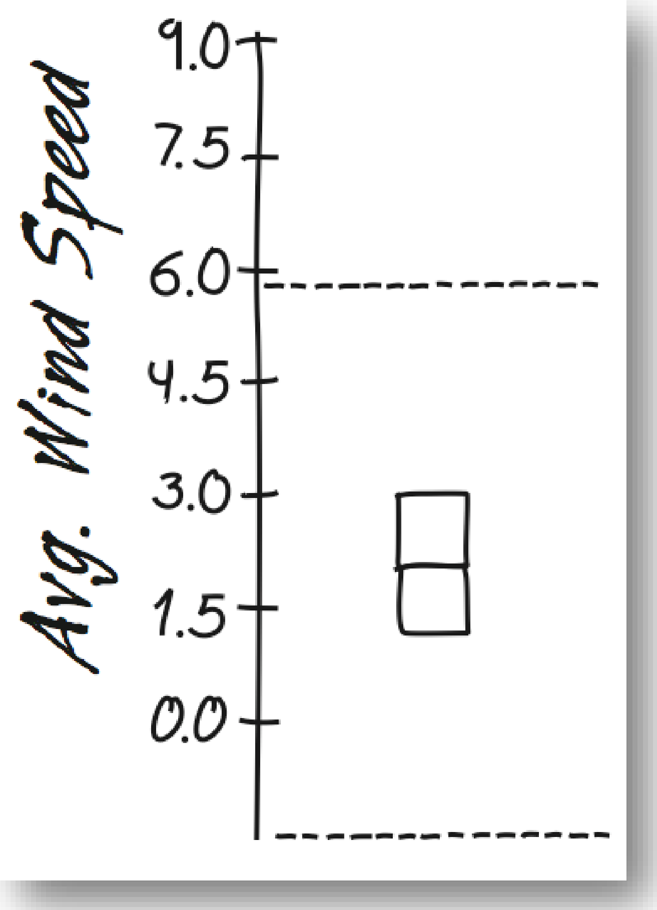
Constructing Boxplots

Draw "fences" around the main part of the data.

The upper fence is 1.5*(IQR) above the upper quartile.

The lower fence is 1.5*(IQR) below the lower quartile.

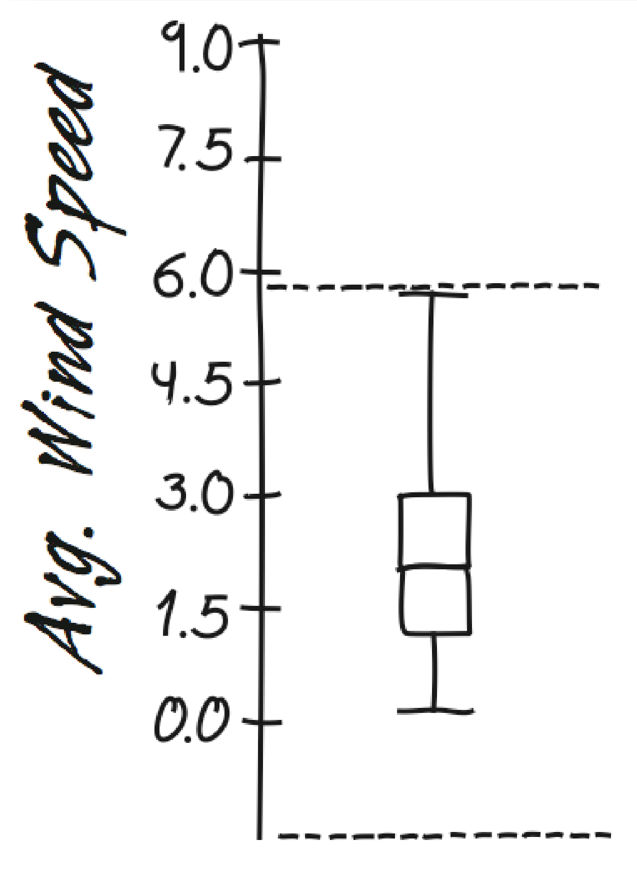
Note: the fences only help with constructing the boxplot and should not appear in the final display.



Constructing Boxplots

Use the fences to grow "whiskers."

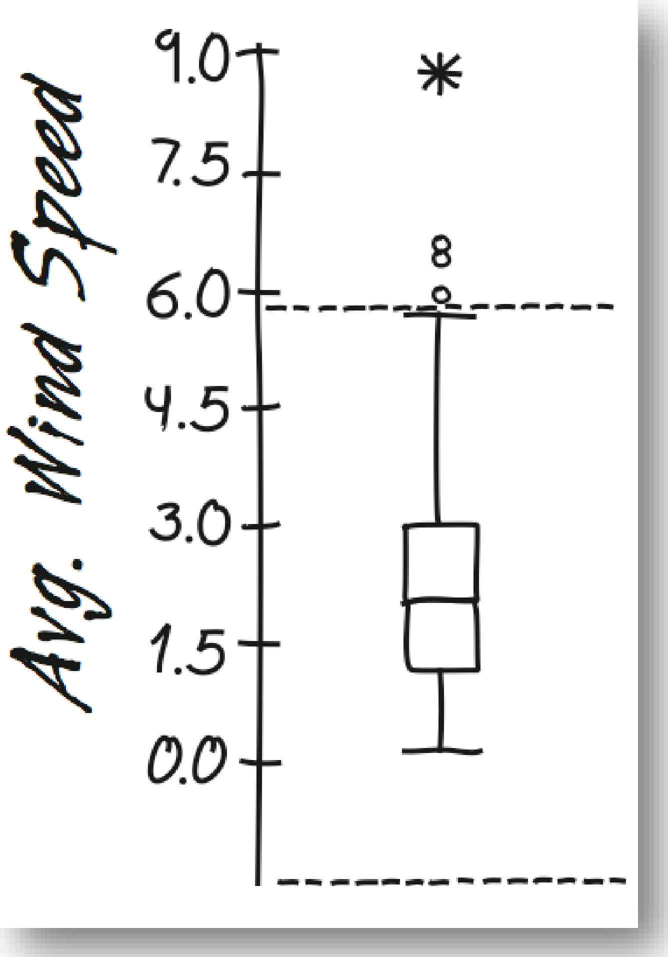
- Draw lines from the ends of the box up and down to the *most extreme data values found within the fences*.
- If a data value falls outside one of the fences, we do *not* connect it with a whisker.



Constructing Boxplots

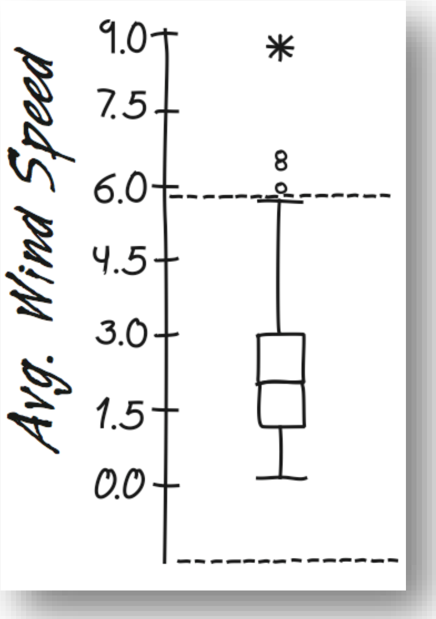
Add the **outliers** by displaying any data values beyond the fences with special symbols.

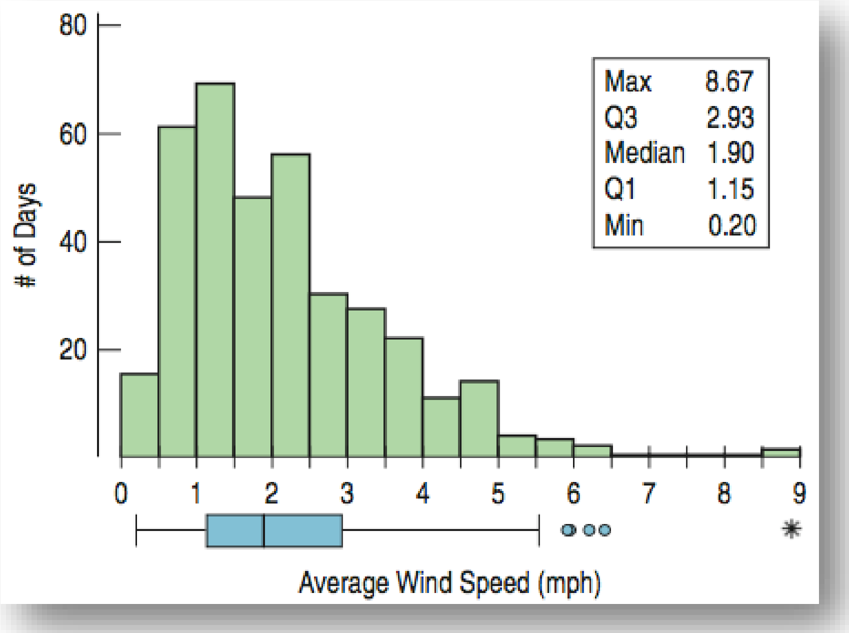
We often use a different symbol for “far outliers” that are farther than 3 IQRs from the quartiles.



Boxplots v Histograms

Compare this histogram and boxplot for daily wind speeds:





How does each display represent the distribution?

What Do Boxplots Tell Us?

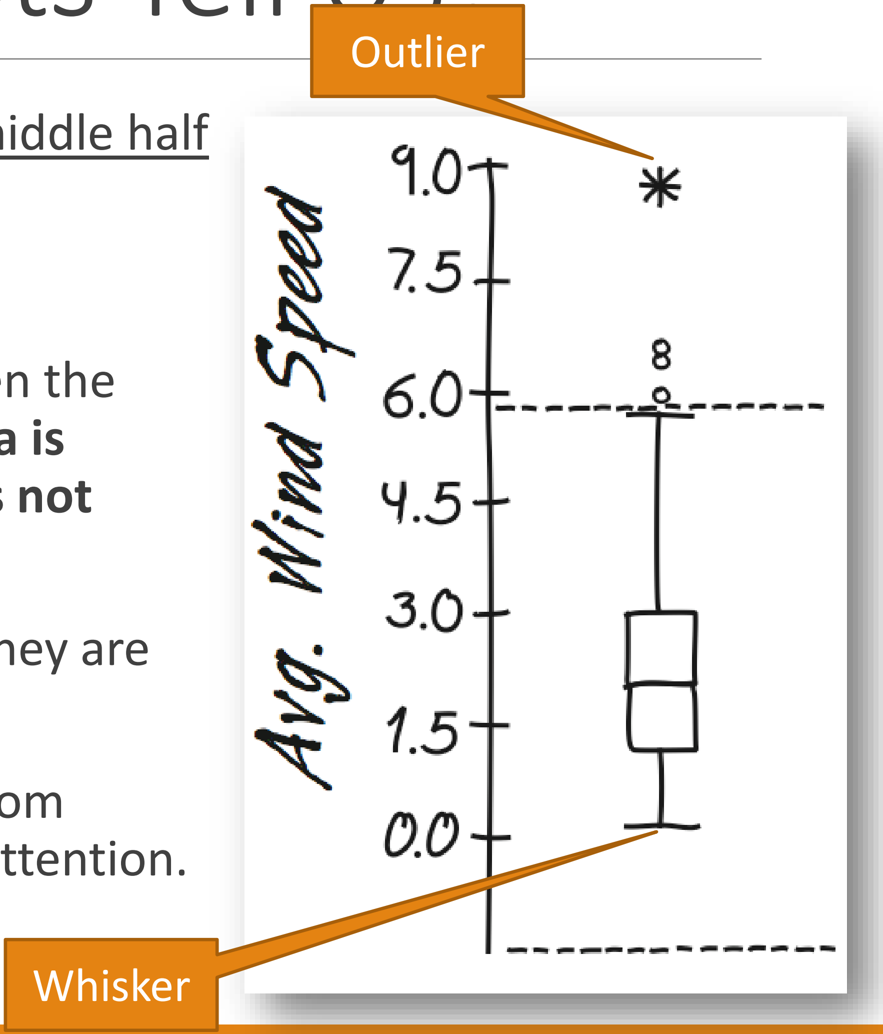
The **center of the boxplot** shows us the middle half of the data between the quartiles.

The **height of the box** is equal to the **IQR**.

If the **median is roughly centered** between the quartiles, then the **middle half of the data is roughly symmetric**. Thus, if the **median is not centered**, the **distribution is skewed**.

The **whiskers** also show the **skewness** if they are not the same length.

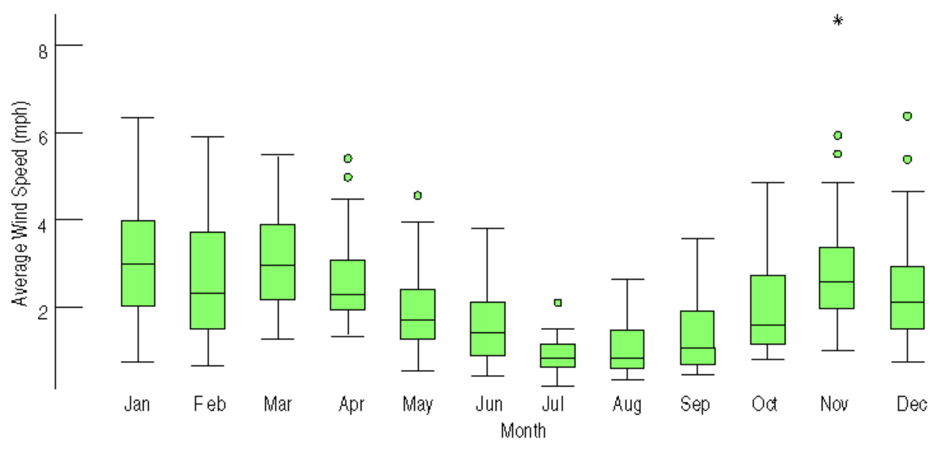
Outliers are out of the way to keep you from judging skewness, but give them special attention.



Comparing Groups

Boxplots offer an ideal balance of information and simplicity, hiding the details while displaying the overall summary information.

We often plot them side by side for groups or categories we wish to compare.



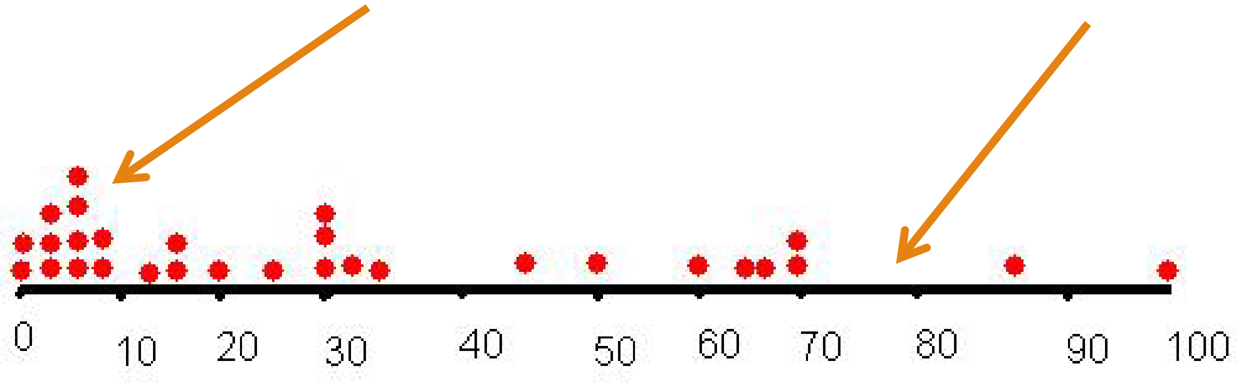
What do these boxplots tell you?

Dot Plot

A data display in which each data item is shown as a dot above a number line

In a dot plot a **cluster** shows where a group of data points fall.

A **gap** is an interval where there are no data items.



Which Descriptive Statistic to use?

Depends on **measurement type** and **data dispersion**

Interval or Ratio (Scale)

- Normally distributed
  - **Mean and Standard Deviation**
- Skewed/Kurtotic
  - **Median and Interquartile Range**

KEY
SLIDE

Which Visualisation to use?

Depends on **measurement type** and **data dispersion**

Interval or Ratio (Scale)

- Normally distributed
  - **Histogram**
- Skewed/Kurtotic
  - **Histogram/BoxPlot**

KEY SLIDE

Probability and Statistical Inference

DESCRIBING AND VISUALISING QUALITATIVE DATA

Sources used in creation of this lecture:
Discovering Statistics Using R Field, Miles and Field;
Understanding Basic Statistics, Brase and Brase;
Statistics and Data Analysis, Peck, Olsen and Devore

What do I need to describe for categorical data?

Set of all possible values

Frequency

- The number of occurrences of each possible value
- Maybe a graph or a table
- The most commonly occurring value

Unusual Occurrences

- Gaps
- Clusters

Graphical Presentation – Choose the correct type for each variable

Nominal or Ordinal data

- Bar charts
- Pie charts
- Frequency Tables

KEY
SLIDE

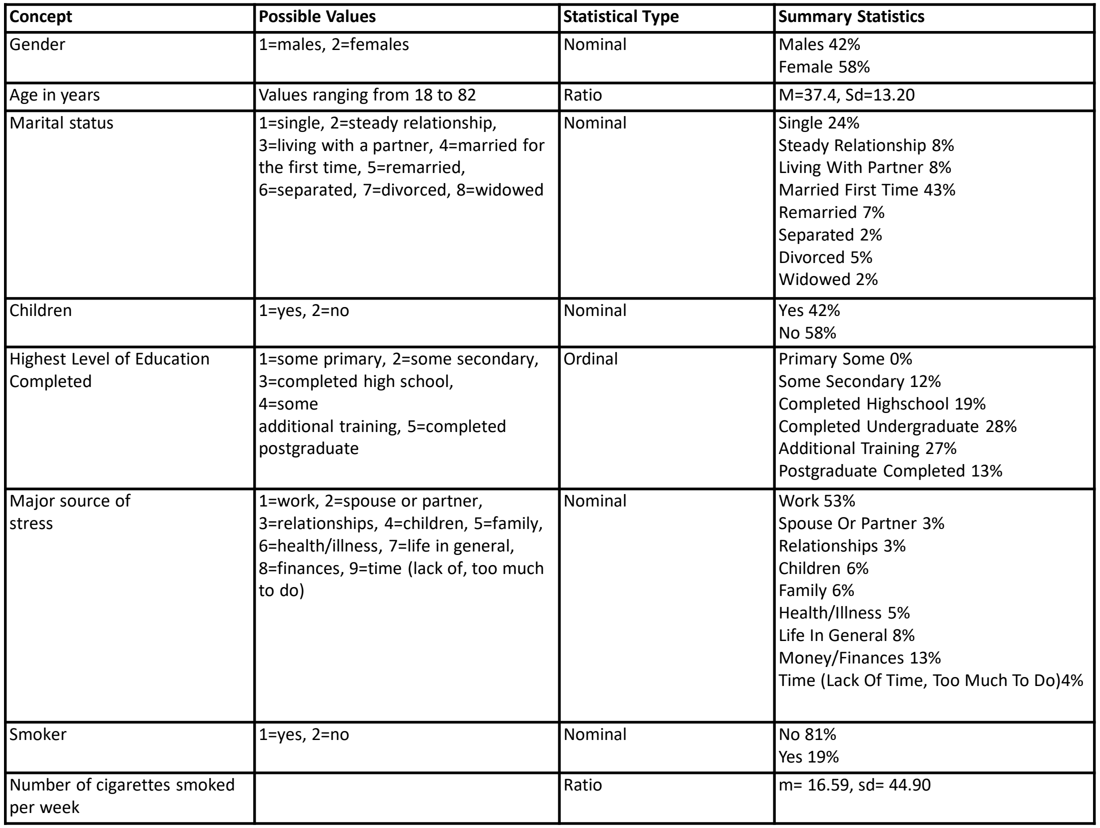
Which Descriptive Statistic to use?

Ordinal or nominal

- **Mode and/or simple frequencies**

KEY SLIDE

Survey.dat



Concept	Possible Values	Statistical Type	Summary Statistics
Gender	1=males, 2=females	Nominal	Males 42% Female 58%
Age in years	Values ranging from 18 to 82	Ratio	M=37.4, Sd=13.20
Marital status	1=single, 2=steady relationship, 3=living with a partner, 4=married for the first time, 5=remarried, 6=separated, 7=divorced, 8=widowed	Nominal	Single 24% Steady Relationship 8% Living With Partner 8% Married First Time 43% Remarried 7% Separated 2% Divorced 5% Widowed 2%
Children	1=yes, 2=no	Nominal	Yes 42% No 58%
Highest Level of Education Completed	1=some primary, 2=some secondary, 3=completed high school, 4=some additional training, 5=completed postgraduate	Ordinal	Primary Some 0% Some Secondary 12% Completed Highschool 19% Completed Undergraduate 28% Additional Training 27% Postgraduate Completed 13%
Major source of stress	1=work, 2=spouse or partner, 3=relationships, 4=children, 5=family, 6=health/illness, 7=life in general, 8=finances, 9=time (lack of, too much to do)	Nominal	Work 53% Spouse Or Partner 3% Relationships 3% Children 6% Family 6% Health/Illness 5% Life In General 8% Money/Finances 13% Time (Lack Of Time, Too Much To Do)4%
Smoker	1=yes, 2=no	Nominal	No 81% Yes 19%
Number of cigarettes smoked per week		Ratio	m= 16.59, sd= 44.90