STAT20060 - Statistics and Probability
Handout 2 - Descriptive Statistics

Damien McParland

Statistical Process

1. Hypothesis
2. Study Design
3. Collect Data
4. Analyze Data
    - Descriptive Statistics
    - Inference
5. Present Results

Population and Samples

- **Population:** Entire set of objects of interest.
- **Sample:** Subset of the set of objects of interest.
- e.g. Interested in the average height of UCD students.
  - Population: Entire UCD student body.
  - Sample: Students in one lecture theatre.
- Time, expense and efficiency means that we analyze samples to make inferences about a population.
- This is **statistical inference**.

Study Types

Designed Experiment

Researcher exerts control over the experimental units. e.g. Tensile strength of beams randomly assigned to different treatments.

Observational Study

Researcher observes the experimental units and records the variables of interest. e.g. RPM of an engine at the failure time of a component.

- In a designed experiment we create differences in the explanatory variable and then examine the results. In an observational study we observe differences in the explanatory variable and then notice whether these are related to differences in the response variable.

- Experimental studies are complex to implement, but they are more informative. There can be ethical issues with experimental studies.

Data Types

Quantitative Data

Measurements that are recorded on a natural numerical scale.

- Continuous Data are measurements which can fall anywhere on the real line. (Or an interval of the real line)
  e.g. Weight, temperature, volume ...

- Discrete Data are measurements which can only take one of a finite set of values.
  e.g. Number of server crashes in a week, number of coin flips until a head is observed, ...

Data Types

Qualitative Data (Categorical)

Measurements that cannot be recorded on a natural numerical scale.

- Nominal Data is qualitative data with no meaningful ordering.
  e.g. Eye colour, (Blue, Green, Brown...).
- Ordinal Data is qualitative data which has an inherent order.
  e.g. Your grade at the end of the semester (B, B+, A-, etc.).
- Qualitative data is very common in survey data.
- Note that there is no concept of distance with ordinal data.

Data Types

- What data types do the following have?
  - The amount of active ingredient in a pharmaceutical pill.
  - The price of a share in Bank of Ireland.
  - The number of students in UCD.
  - The degree what a randomly selected member of the class is studying.
  - The rating that you give a song on iTunes.

Warning!

- The distinction between categorical or numerical data can be difficult to determine in some cases.
- Consider the iTunes rating question on the previous slide.
- The ratings on iTunes take the values 1,2,3,4,5. *It could be argued that these are a categorical value in disguise!*
- This issue can arise in a lot of survey applications. *For example someone might code*
  - 1 – Strongly Disagree
  - 2 – Disagree
  - 3 – No opinion
  - 4 – Agree
  - 5 – Strongly Agree
- This is called a Likert scale.

Warning 2!

- The distinction between discrete and continuous numerical data can be ambiguous too.
  1. Suppose that the weight of an item is recorded to the nearest kilogram.
  2. Suppose the area of a piece of land is recorded to the nearest cm^2.

- It could be argued that
  1. Although weight is continuous it is recorded in a discrete manner.
  2. Although the land area is recorded in a discrete manner the number of possible values are vast, so it's almost continuous.

Descriptive Statistics

Numerical Summaries

- These report some numbers that provide information about the data
- There are many numerical summaries but the two main types are:
  - Measures of location: Where is the data "centered"?
  - Measures of spread: How spread out are the data?

Graphical Summaries

- These provide a pictorial summary of the data.
  "A picture is worth a thousand words."
  - Barchart: Shows the distribution of categorical values.
  - Histogram: Shows the distribution of numerical values.
  - Scatter plot: Shows the relationship between variables

Examples: Graphical Summaries

- It's worth keeping an eye out for good and bad examples of plots.
- Good places to start on the web are:
  - Hans Rosling's talks (`http://www.open.ac.uk/openlearn/whats-on/the-joy-stats`)
  - Andrew Gelman's blog (`http://www.stat.columbia.edu/~cook/movabletype/mlm/`)
  - Edward Tufte's webpage (`http://www.edwardtufte.com`)
  - Media sources like the Economist and the NY Times.

Barcharts

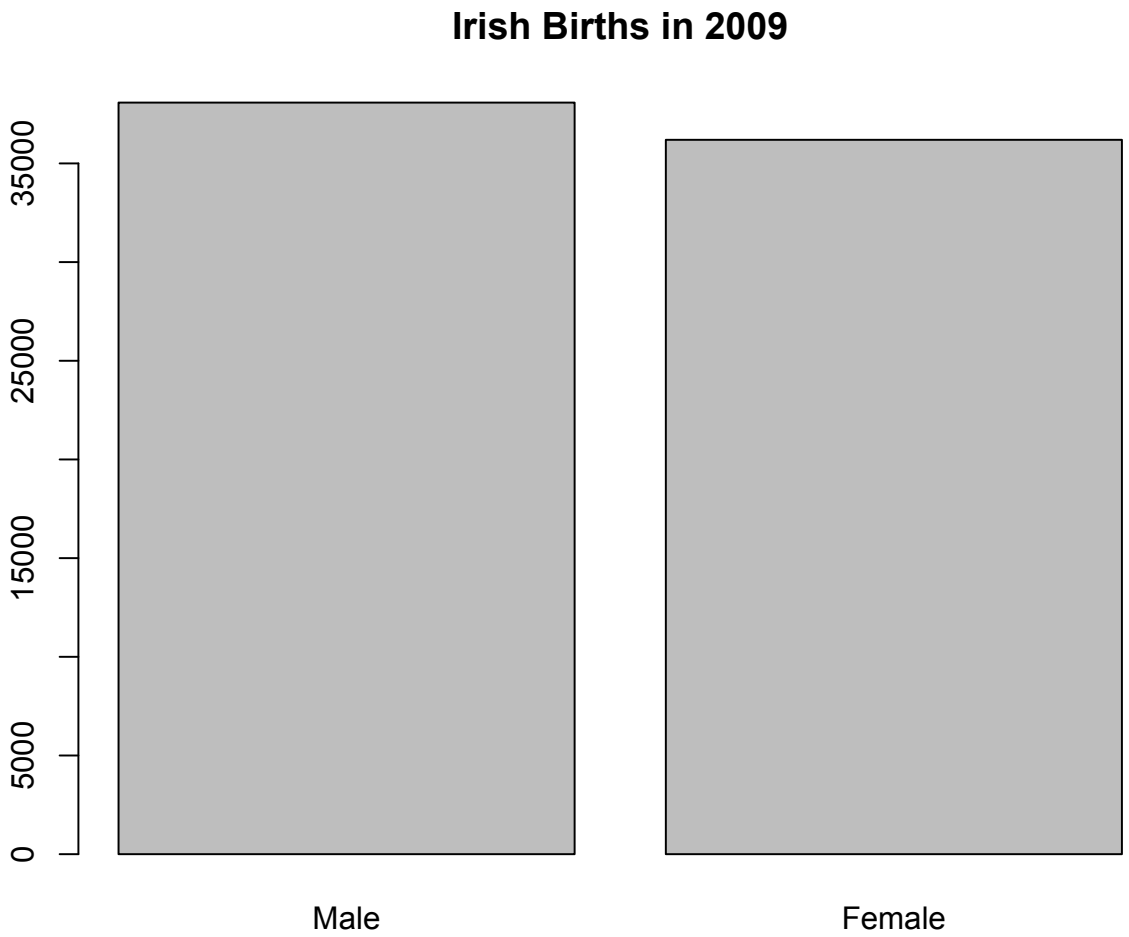
- These provide a graphical summary of **categorical data**.
- They are very easy to produce:
  1. Count the number of observations in each category (frequency).
  2. Draw a plot where each category is an equal width rectangle and the height is proportional to frequency.
- The height of each bar could also be the relative frequency,

$$RF = \frac{Frequency}{N}$$

  where N is the total number of observations.
- If your data is ordinal make sure that the rectangles are in a sensible order. *Warning: Software may reorder the categories into alphabetical order*

Example: Births

- There were 74278 children born in Ireland in 2009.
- How did it break down into Male/Female?



Example: Births 2

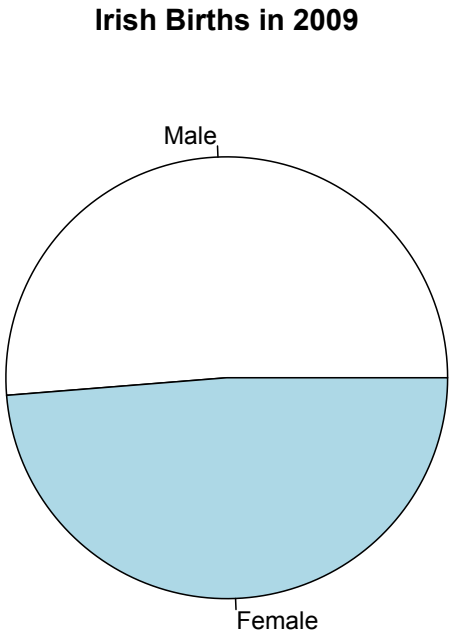
- If we look at the age of the mother, coded into categorical bands.

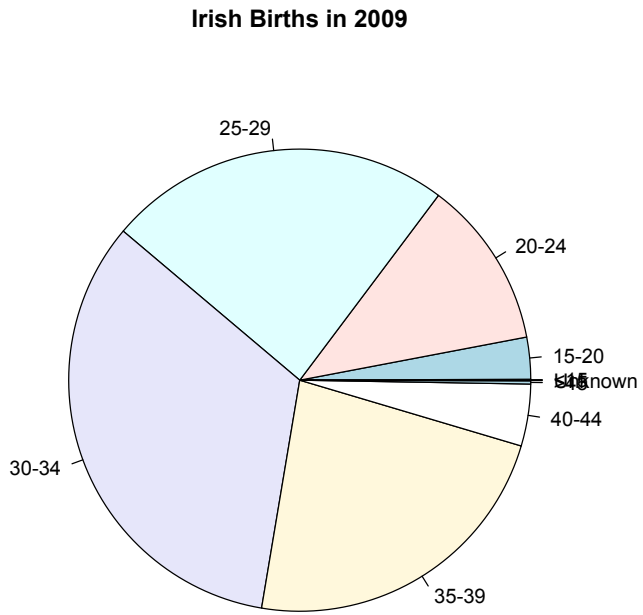
Example: Births 3

- This plot is technically equivalent to the previous one, but it's not very good!

Piechart

- A piechart is an alternative plot. The area of each slice (equivalently angle) is proportional to frequency.





- Lots of people argue that they're a poor plot choice.

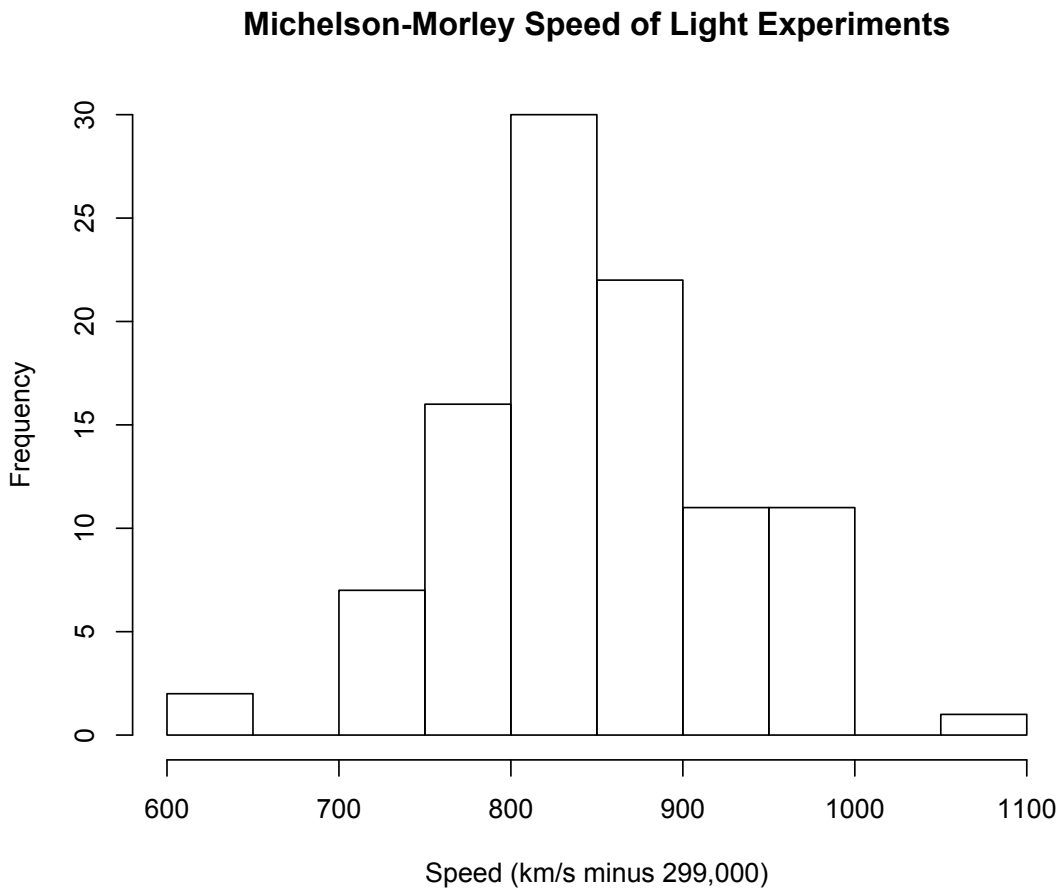
Histogram

- It could be argued that the barchart for the mother's age is almost a histogram.

- Histograms are used to represent numerical data.
  1. Divide the range of the data into bins. *The bins are usually, but not necessarily equal width.*
  2. Count how many observations fall into each bin.
  3. Plot a rectangle for each bin, where the base length is proportional to the bin width and the **area** of the rectangle is proportional to frequency.

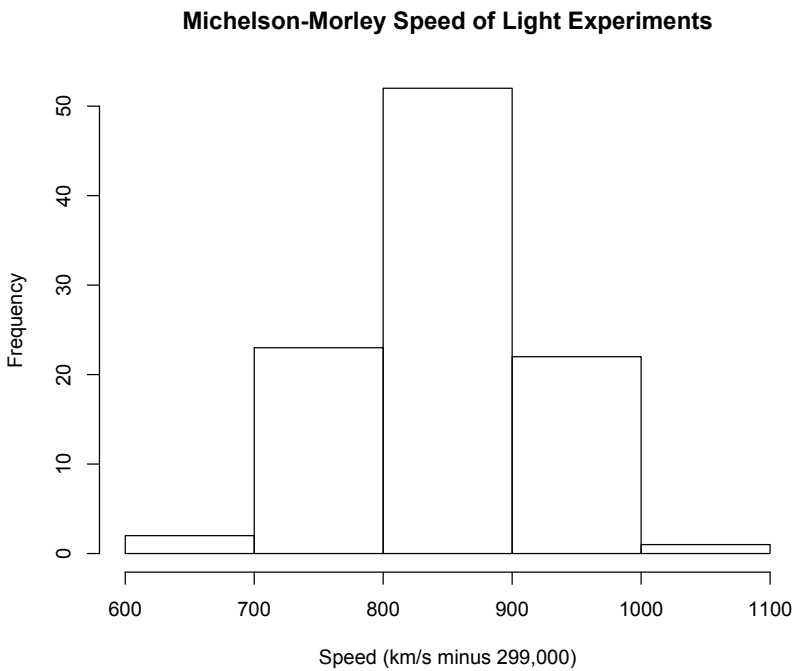
Example: Speed of Light

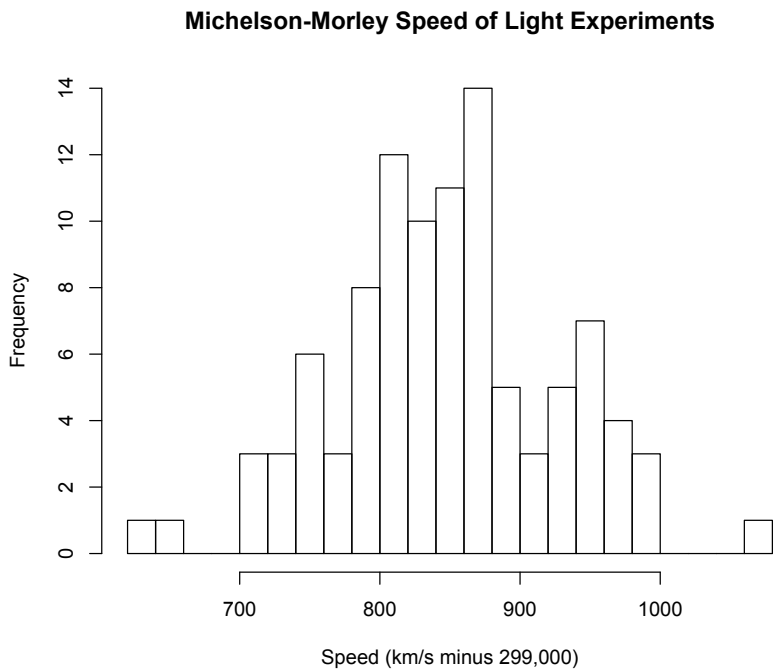
- In 1887 Michelson and Morley conducted five experiments to determine the speed of light. In each experiment they made twenty measurements of the speed of light.
- A histogram of the one hundred measurements is as follows:



Example: Other Histograms

- The number of bins in a histogram can change its appearance.

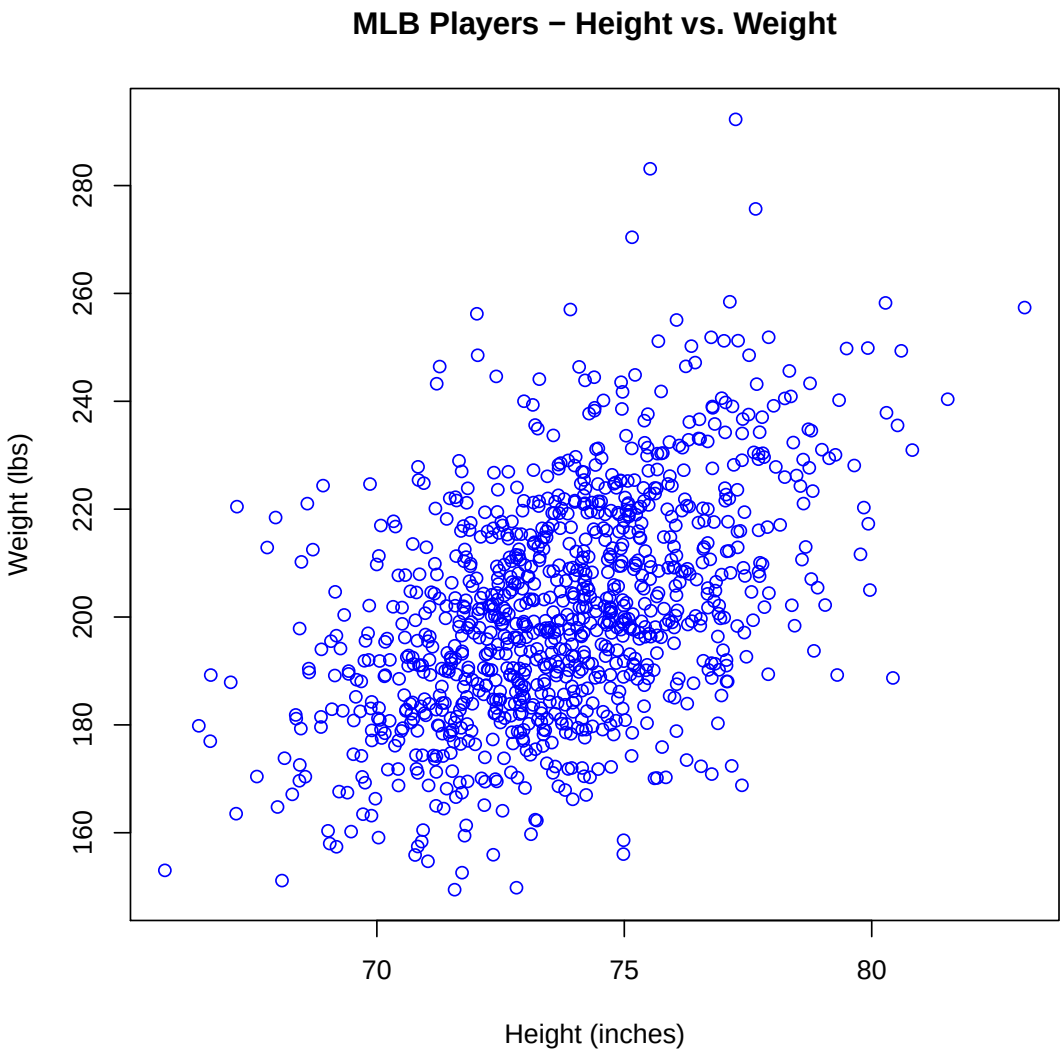




Scatterplots

- If we have measurements on two variables, plotting one against the other can be useful for finding patterns.
- Scatterplots can be a useful graphical representation of the relationship between variables.
- These plots will be very useful for regression analysis. (Later in the course)

Scatterplot



Numerical Summaries

- We have two main types of numerical summaries:
  - Measures of location
  - Measures of spread
- Some summaries (eg. quantiles) have a dual purpose.
- Measures of location (or central tendency) give an idea of where the data are "centered".
- Measures of spread give an idea of the range of "most" of the data.

Location

- The mean *or average* is a measure of central tendency.
- Suppose we have a sample of size n from a population of size N.
- The population mean is

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

- The sample mean is

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

- The sample mean is sensitive to outliers.

Median

- The median is the middle value.
- The population median has the same number of values greater or less than it.
- The sample median has the same number of values greater or less than it.
- In the case of an odd number of values, the median is simply the middle value.
- In the case of an even number of values, there isn't a unique median (why?) so the convention is to average the middle two values.
- The sample median is insensitive to outliers

Mode

- The mode is the most common value.
- For numerical values, the mode may not be well defined.
- It is possible to have more than one mode.
- The word “mode” is also used to refer to the highest point in the histogram.

Example: CPI

- The consumer price index was recorded for ten consecutive years.

  `3.295837 3.295837 3.401197 3.610918 3.871201`
  `3.850148 3.784190 3.761200 3.713572 3.688880`

- How can we summarize the data?

Example: Old Faithful Geyser

- The time (in minutes) between eruptions of the Old Faithful geyser in Yellowstone national park were recorded for 272 eruptions.

79 54 74 62 85 55 88 85 51 . . . 60 75 81 46 90 46 74

- How can we represent the distribution of the time between eruptions?
- What characteristics does the distribution have?

Example: Speed of Light

- Michelson and Morley conducted five experiments in 1887 to study the speed of light.

- Their experiments yielded 100 measurements of the speed of light.
- How can we represent their measurements?

Measures of Spread

- Measures of spread give a numerical summary of the "typical range" of the data.

- A number of alternatives exist including:
  1. Range: This records the full spread of the data.
  2. Interquartile Range: This records the spread of the middle of the data.
  3. Standard Deviation: This measures how values differ from the mean value.

Range

- The range records the difference between the minimum and the maximum value.
- This is a good statistic because it records the full range of the data.
- However, it is highly sensitive to the extreme values in the data.

Data	Min	Max	Range
CPI	3.295837	3.850148	0.554311
Michelson-Morley	620	1070	450
Old Faithful	43	96	53

- Note that the range has the same "units" as the data.

Interquartile Range

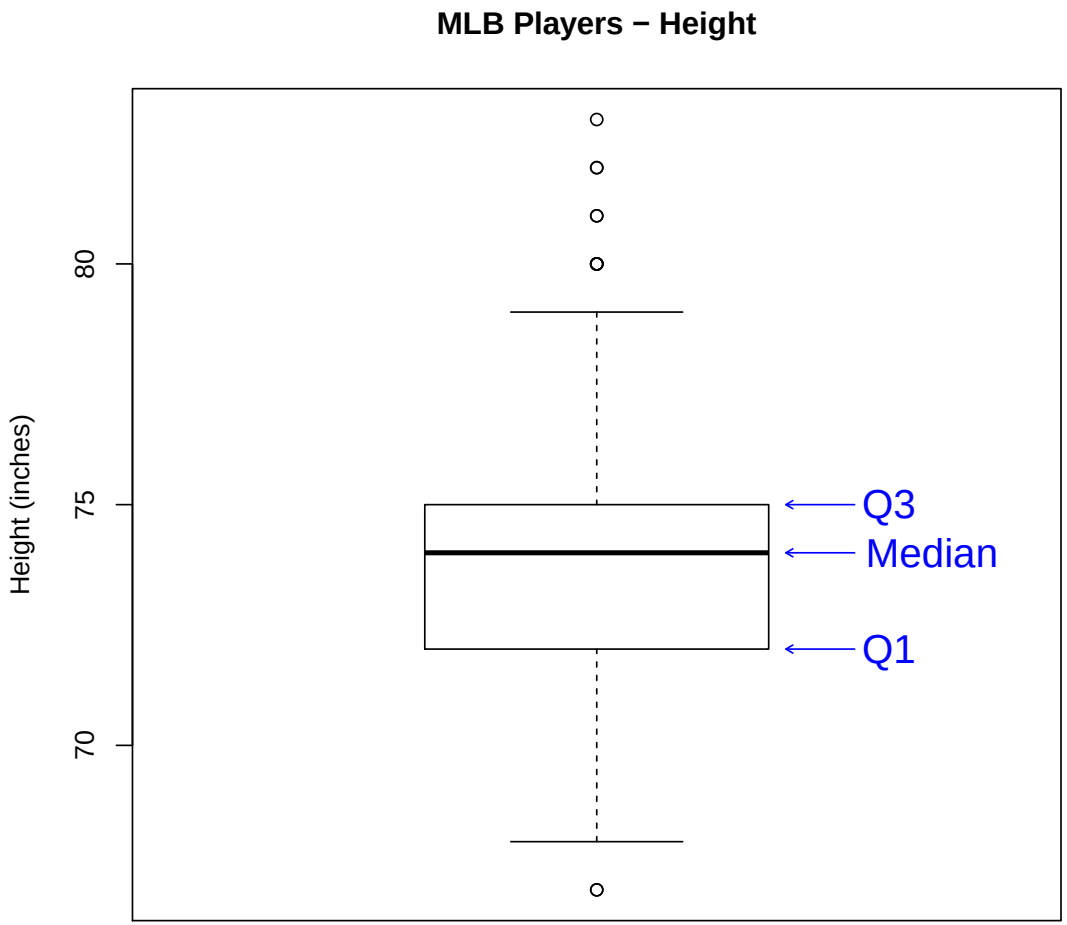
- The interquartile range (IQR) records the difference between the 75% percentile and the 25% percentile.
- The 25% percentile (also known as the 1st Quartile) has 25% of the values less than it.
- The 75% percentile (also known as the 3rd Quartile) has 75% of the values less than it.
- It records the range of the middle 50% of the data.
- It is very robust to extreme values in the data.

Data	Q1	Q3	IQR
Failure	46.25	144.50	98.25
CPI	3.453627	3.871201	0.3248153
Michelson-Morley	807.5	892.5	85
Old Faithful	58	82	24

- Note that the IQR has the same "units" as the data.

Box Plots

- Box plots provide a useful graphical summary of the data.
- They show where there data are located and also indicate the spread of the data.



Standard Deviation

- The standard deviation records the root mean squared deviation of values from the mean.
- Sounds nasty!
- The population standard deviation is

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}.$$

- The sample standard deviation is

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

- The reason for the mysterious -1 will be clear later!

Standard Deviation (2)

- For our data we get:

Data	SD
Failure	54.1
CPI	0.219
Michelson-Morley	79.0
Old Faithful	13.5

- Note that the SD has the same "units" as the data.

- The *variance* is the squared standard deviation.
  It does not have the same units!

Using Summaries in Conjunction

- In many situations we find that "most" of the data lie in the range

$$\text{Median} \pm 1.5\text{IQR}$$

and/or

$$\overline{x} \pm 2s$$

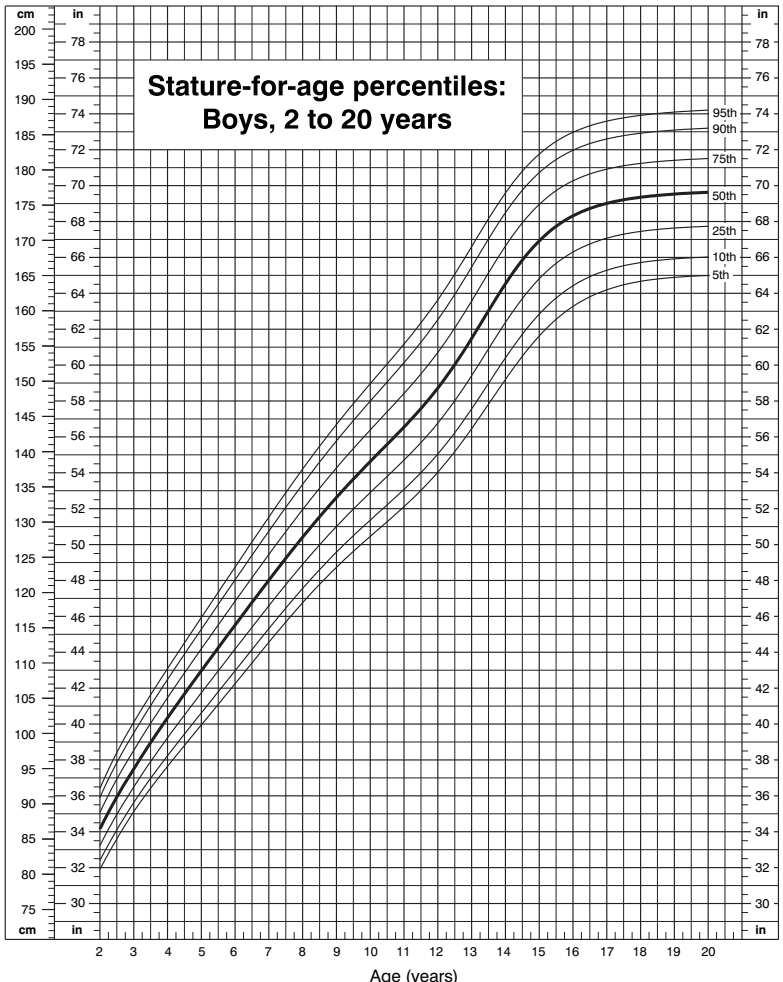
- Later, we will see a theorem that explains why 75% of the data lie in the second range.

Percentiles

- Percentiles are used to give a value where a particular percentage of a sample (or population) fall below this value.
- The 1% percentile will have 1% of values below it and 99% of values above it.
- The 90% percentile will have 90% of values below it and 10% of values above it.

Growth Charts

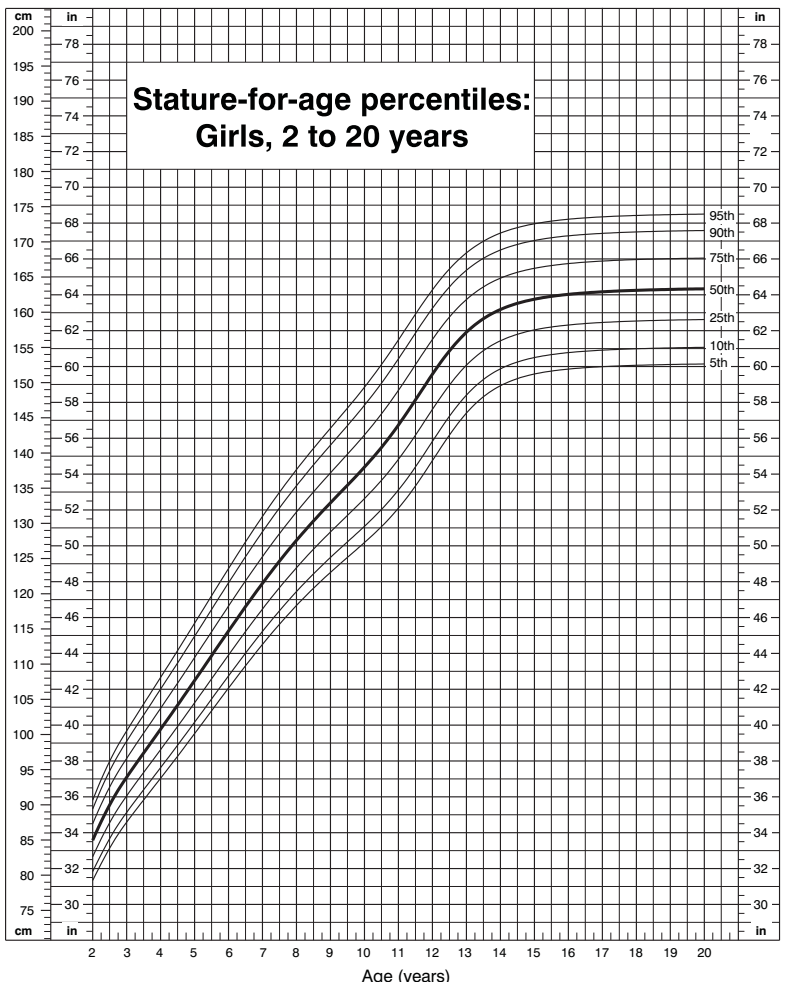
CDC Growth Charts: United States



Published May 30, 2000.
SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).

SAFER · HEALTHIER · PEOPLE™

CDC Growth Charts: United States



Published May 30, 2000.
SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).

SAFER · HEALTHIER · PEOPLE™