

ITCS 121 Statistics

Lecture 1

Describing Data With Graphs

Why study statistics?

- ▶ Basic math knowledge is useful for getting a good job, managing your finances, understanding your taxes etc.
- ▶ Statistics specifically is important for lots of jobs you might want to do as an ICT graduate. E.g.
 - ▶ Some programming jobs need statistics, e.g. for testing software.
 - ▶ Data scientists need to know lots of statistics.
 - ▶ Traditional analyst jobs at banks etc.
- ▶ Statistics is one of the most important tools we use to understand the world.
- ▶ Many arguments in politics, science, economics etc. are statistical.

What is statistics for?

- ▶ **Question:** Is the New Year period unusually dangerous for traffic accidents in Thailand?

Media thinks yes.



HOME NEWS THAILAND WORLD NEWS TOURISM HEALTH LEARNING
FEATURED VIDEO
ABOUT CHIANGRAI ABOUT US ADVERTISE WITH US

Road Safety Center Opened for the 7 Dangerous Days of the New Year Holiday

By **Editor** on December 28, 2018 • Comments Off on Road Safety Center Opened for the 7 Dangerous Days of the New Year Holiday

The interior minister said Thailand ranks ninth among 175 nations in terms of road fatalities and most accidents were involved motorcycles.

Thailand

Jamie Fullerton

@jamiefullerton1

Thu 3 Jan 2019 09:14 GMT

NEWS / THAILAND

Almost 500 people killed on Thai roads during New Year holidays

Drink-driving main cause of accidents as 4,000 injured between the 'seven dangerous days', December 27 and January 2.

3 Jan 2019



Day 1: 42 dead, 432 injured

28 Dec 2018 at 14:31 43 comments
WRITER: ONLINE REPORTERS



463 killed in collisions on Thailand's roads in new year's week

Death toll from period police call 'Seven Dangerous Days' is close to record



What are the numbers?

- ▶ World Health Organization (WHO) has figures for traffic deaths in Thailand for 2013.

[filter table](#) | [reset table](#)

Last updated: 2018-05-16

Download filtered data as: [CSV table](#) | [XML \(simple\)](#) | [JSON \(simple\)](#)

Download complete data set as: [CSV table](#) | [Excel](#) | [CSV list](#) | [more...](#)

	Estimated number of road traffic deaths ⁱ	Estimated road traffic death rate (per 100 000 population) ⁱ
Country	2013	2013
Suriname	103 ⁱ	19.1
Sweden	272 ⁱ	2.8
Switzerland	269 ⁱ	3.3
Syrian Arab Republic		20.0
Tajikistan	1 543 [1 387 - 1 699] ⁱ	18.8
Thailand	24 237 ⁱ	36.2
The former Yugoslav republic of Macedonia	198 ⁱ	9.4
Timor-Leste	188 [158 - 219] ⁱ	16.6
Togo	2 123 [1 719 - 2 526] ⁱ	31.1
Tonga	8 ⁱ	7.6
Trinidad and Tobago	189 ⁱ	14.1

Comparing the numbers.

- ▶ According to the WHO, in 2013 there were 24,237 deaths.
- ▶ That's $\frac{24,237}{365} = 66.4$ deaths per day.
- ▶ According to Thai government there were 463 traffic deaths over New Year Week 2019.
- ▶ That's $\frac{463}{7} = 66.14$.
- ▶ So deaths per day was slightly *less* than average in New Year period...

Does this answer the question?

- ▶ From the numbers we have seen, it looks like there is no reason to believe the New Year week is more dangerous than any other week.
- ▶ But...

Does this answer the question?

- ▶ Is it ok to compare numbers from 2013 with numbers from 2018/2019?
- ▶ How do the WHO calculate their number? Is it different from the way the Thai government calculate their number?
 - ▶ Is it reasonable to compare numbers from these different sources?
- ▶ What is an 'average day' in Thailand for road deaths?
 - ▶ Maybe the New Year week is more dangerous than most weeks, but there's another period which is also very dangerous which takes the overall average up.

Does this answer the question?

- ▶ To really answer this question we need to do more work, which is what we need statistics for.
- ▶ We can see that the story the media tells needs more justification before we should believe it.

Variables and Data

- ▶ A variable is a characteristic that can be different for different objects, or change over time for one object.
- ▶ Variables can be numbers, but they don't have to be.
- ▶ Examples:
 - ▶ Hair colour
 - ▶ Weight
 - ▶ CPU speed
 - ▶ Nationality
- ▶ When we're doing math we often use abstract symbols like x or y to represent variables.

Populations and Measurements

- ▶ In statistics we start with a collection of objects we are interested in.
- ▶ The objects in this collection are called **experimental units**.
- ▶ We find the value of a variable for a given experimental unit by taking a **measurement**.
- ▶ The collection of all these measurements is called the **population**.
- ▶ E.g.
 - ▶ Experimental units are the individual Thai people.
 - ▶ If the variable is height, we take a measurement by finding how tall someone is.
 - ▶ Population = the heights of all Thai people.

Example: computer prices

- ▶ Variable = price (in THB).
- ▶ Experimental unit = computer.
- ▶ Typical measurements:
 - ▶ 10,000 THB
 - ▶ 20,000 THB
 - ▶ 31,499 THB

Samples

- ▶ The point of statistics is to learn about the world by looking at population data.
- ▶ Usually the population of interest is too big for us to deal with all of it.
- ▶ So, we try to understand populations by looking at **samples**.
- ▶ **Samples** are subsets taken from a population in some way.
- ▶ This will be very important, and we will come back to it in later classes.

Types of variables

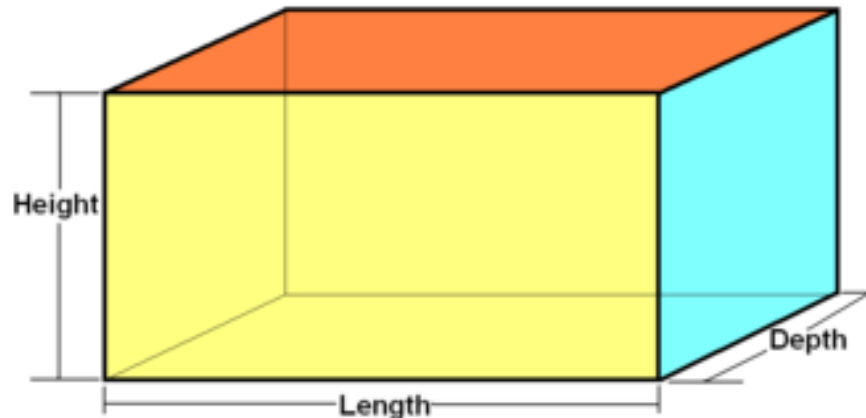
- ▶ Variables can be numbers (**Quantitative**).
- ▶ Variables can also be properties (**Qualitative**).
 - ▶ E.g. colour, flavour, nationality.
- ▶ Quantitative variables can be either **discrete** or **continuous**.
- ▶ **Discrete** variables can only take whole number values.
 - ▶ E.g. 1,3,7,-5,287...
- ▶ **Continuous** variables can take any real number value (i.e. any number we can represent with a possibly infinite decimal).

How many variables?

- ▶ Experimental units can be associated with more than one variable.
 - ▶ E.g. you can measure a person's height and weight.
- ▶ If we measure one variable for each experimental unit then we have **univariate** data.
- ▶ If we measure two variables we have **bivariate** data.
- ▶ If we measure three or more variables we have **multivariate** data.

Example: Shapes

- ▶ A line has a length. This is univariate data.
- ▶ A rectangle has a length and a height. This is bivariate data.
- ▶ A cuboid has length, height and depth. This is multivariate data.



Class activity 1

1. Qualitative or Quantitative?
 - a) Amount of time it takes to solve a simple puzzle.
 - b) Number of students in a first-grade classroom.
 - c) Country where a person lives.

2. Discrete or Continuous ?
 - a) Number of people in a poll of 1000 who support the government.
 - b) Time to complete an exam.
 - c) Number of brothers and sisters a person has.

Class activity 2

An educational researcher wants to evaluate the effectiveness of a new method for teaching. Achievement at the end of a period of teaching is measured by a student's score on a final exam.

- a) What is the variable to be measured?
What type of variable is it?
- b) What is the experimental unit?
- c) Identify the population of interest to the experimenter.

Visualizing data

- ▶ The point of collecting data is to help humans understand things.
- ▶ Often we want to present data in a way that makes it easy for humans to understand it.
- ▶ We will look at some important ways to do this.

Graphing qualitative variables

- ▶ There are several graphs we can draw for qualitative variables.
- ▶ These are useful in different situations.
- ▶ Here we will look at **bar** and **pie charts**.
- ▶ These are both based on the idea of looking at the frequency of occurrences.

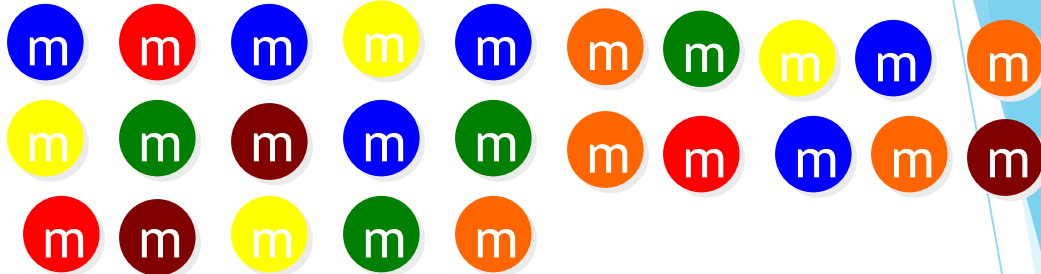
Frequency and relative frequency

- ▶ Suppose we have a qualitative variable that can take a finite number of values.
 - ▶ E.g. red, blue, green, orange, brown, yellow.
- ▶ We take all our measurements to get the population.
- ▶ We count the number of times each variable value occurs.
- ▶ This gives us the **frequency** for each value.
- ▶ To calculate the **relative frequency** we take the frequency and divide by the total number of experimental units.







Example: M&Ms

▶ A bag of M&Ms contains 25 candies:

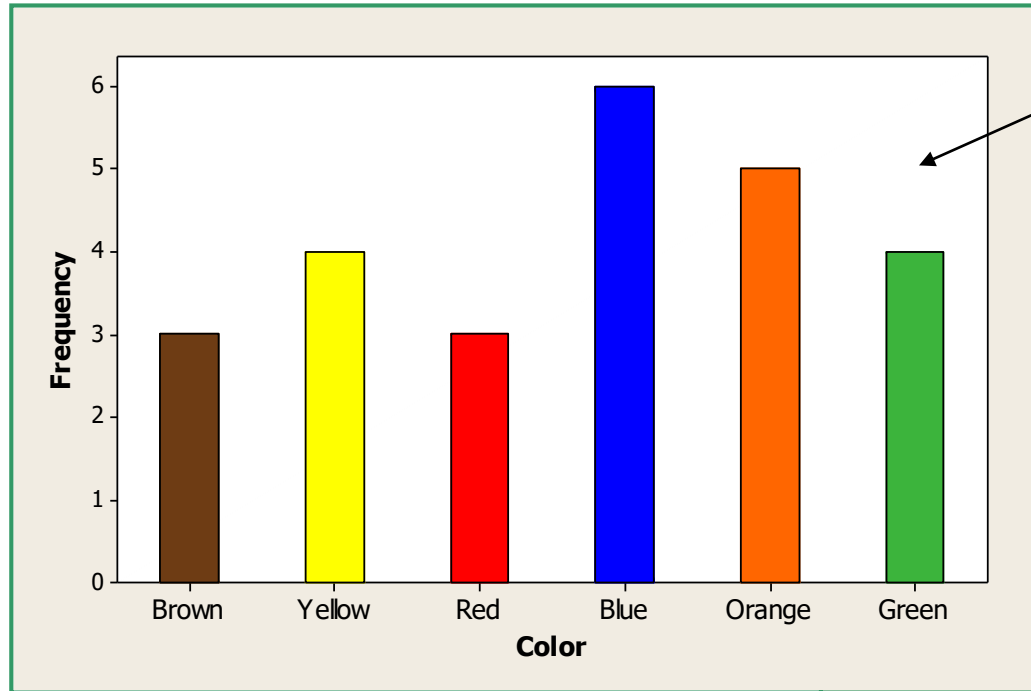
▶ Raw Data:



▶ Statistical Table:

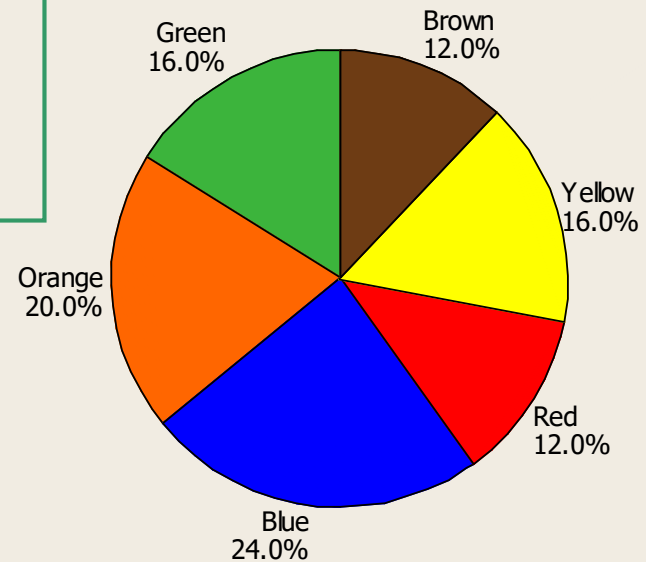
Color	Tally	Frequency	Relative Frequency	Percent
Red		3	$3/25 = .12$	12%
Blue		6	$6/25 = .24$	24%
Green		4	$4/25 = .16$	16%
Orange		5	$5/25 = .2$	20%
Brown		3	$3/25 = .12$	12%
Yellow		4	$4/25 = .16$	16%

Representing the data



Bar Chart

Pie Chart

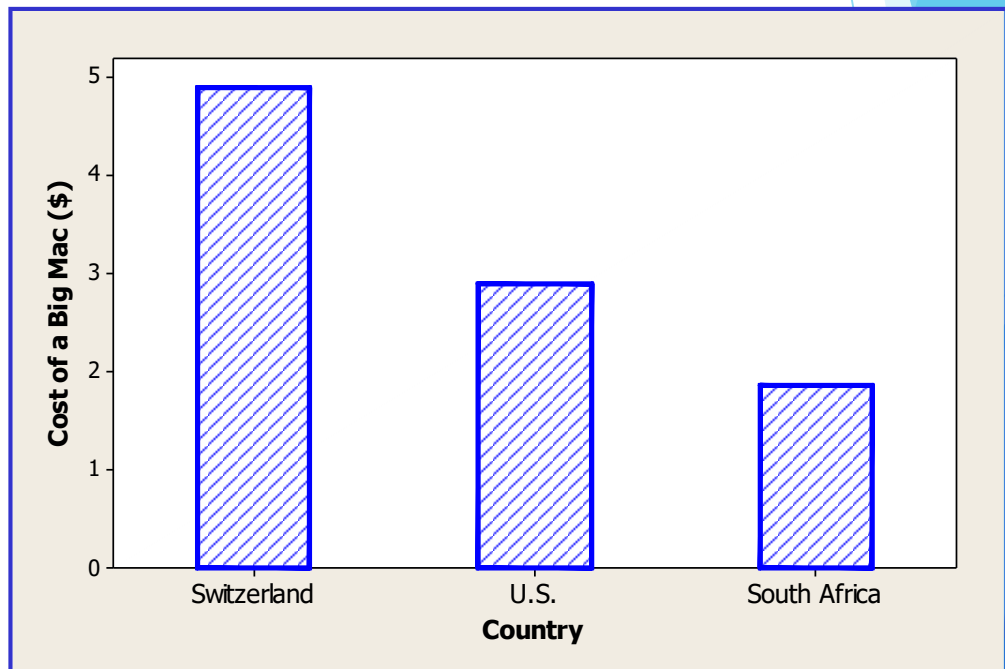


Graphing quantitative variables

- ▶ There are many important techniques for graphing quantitative variables.
- ▶ In this class we will discuss:
 - ▶ Bar charts
 - ▶ Time series
 - ▶ Dot plots
 - ▶ Stem and leaf plots
 - ▶ Histograms

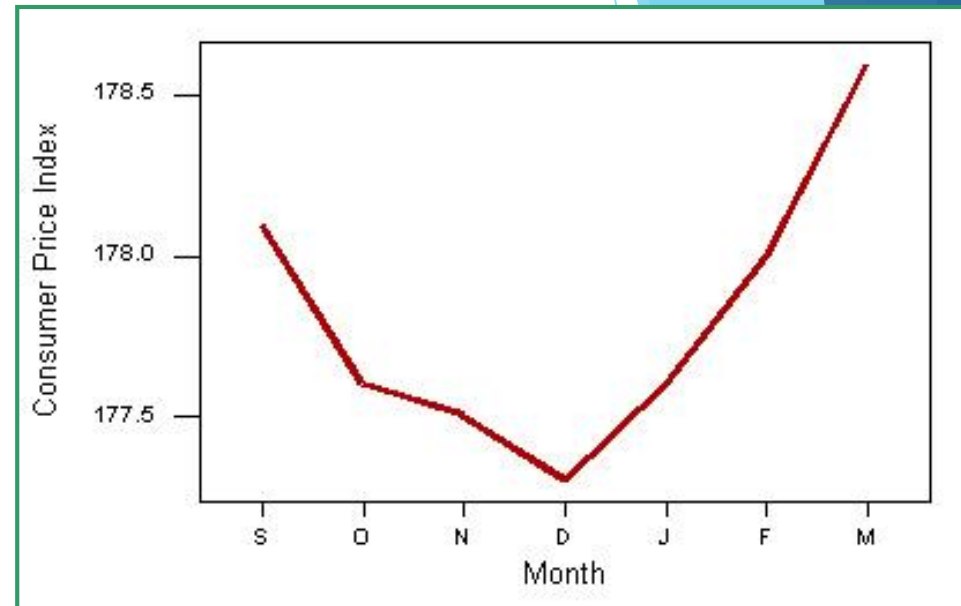
Bar charts for quantitative variables

- ▶ If the population is very small, we can represent it using a bar chart.
- ▶ E.g. We look at the price of a Big Mac burger in three different countries.
- ▶ Data:
 - ▶ Switzerland - \$4.90
 - ▶ USA - \$2.90
 - ▶ South Africa - \$1.86



Time series

- ▶ Suppose we take a measurement for a quantitative variable at different times.
- ▶ We usually call this a **time series**.
- ▶ We could represent the values using a bar chart.
- ▶ Usually we use a line.

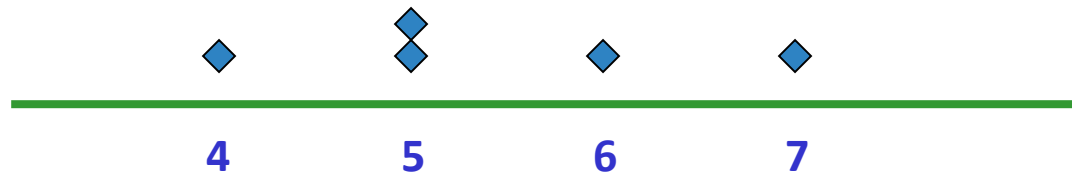


CPI: All Urban Consumers-Seasonally Adjusted

September	October	November	December	January	February	March
178.10	177.60	177.50	177.30	177.60	178.00	178.60

Dot plots

- ▶ Suppose we have a discrete quantitative variable, and a fairly small population.
- ▶ The possible values form the horizontal axis.
- ▶ For each value that occurs in the data we draw a dot in the appropriate place.
- ▶ If there's already a dot in that place we draw the new dot above it. This gives us a **dot plot**.
- ▶ E.g. For the data 4, 5, 5, 7, 6.



Class activity 4

A discrete variable can take on only the values 0, 1, or 2. A set of 20 measurements on this variable is shown here:

1	2	1	0	2	2	1	1	0	0
2	2	1	1	0	0	1	2	1	1

Draw a dot plot to describe the data.

Stem and leaf plots

- ▶ Stem and leaf plots let us present quantitative data in a way that gives information about the frequency of occurrences of values.
- ▶ It works best when the population isn't too big.
- ▶ The easiest way to understand it is to see an example.

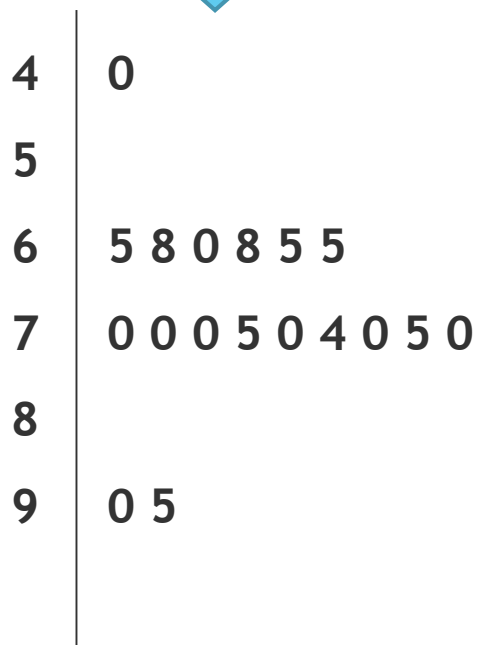
Example: Stem and leaf plots

The population:

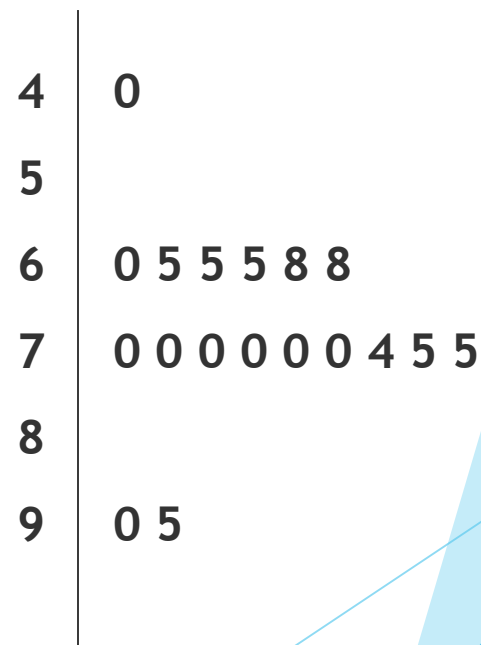
90 70 70 70 75 70 65 68 60
74 70 95 75 70 68 65 40 65



represent data



reorder



Class activity 5

Consider this data:

4.5	3.2	3.5	3.9	3.5	3.9	4.3	4.8	3.6	3.3	4.3	4.2
3.9	3.7	4.3	4.4	3.4	4.2	4.4	4.0	3.6	3.5	3.9	4.0

- a) Construct a stem and leaf plot by using the leading digit as the stem.
- b) Construct a stem and leaf plot by using each leading digit twice. I.e. have one version of the leading digit for second digit 0-4, and another for second digit 5-9. Does this technique improve the presentation of the data? Explain.

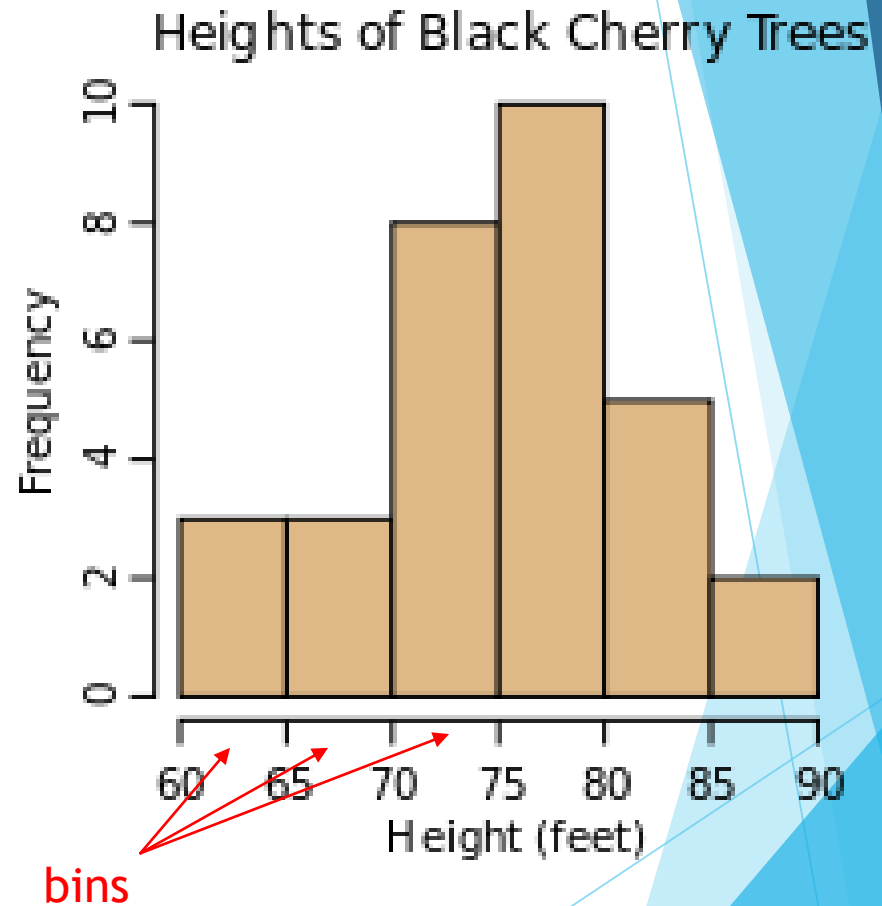
Histograms

- ▶ A histogram counts the number of times the value of a variable falls within certain ranges known as **bins**.
- ▶ When this is represented as a chart the x-axis is divided into intervals representing the bins.
- ▶ The y-axis usually measures either the **frequency**, the **relative frequency**, or the **frequency density**.
- ▶ For frequency histograms the *height* of each bar equals the number of data points that fall inside that bin (the frequency).
- ▶ In relative frequency histograms the height of the bar represents the number of data points divided by the total number.
- ▶ For frequency density the idea is that the *area* of each bar of the histogram is equal to the frequency.
- ▶ Frequency density histograms are good when extreme values occur only rarely.
- ▶ We can collect extreme values together without the height of the bar being misleading.

Cherry Tree Height Histogram

(frequency)

Height (feet)	Frequency
60-65	3
65-70	3
70-75	8
75-80	10
80-85	5
85-90	2

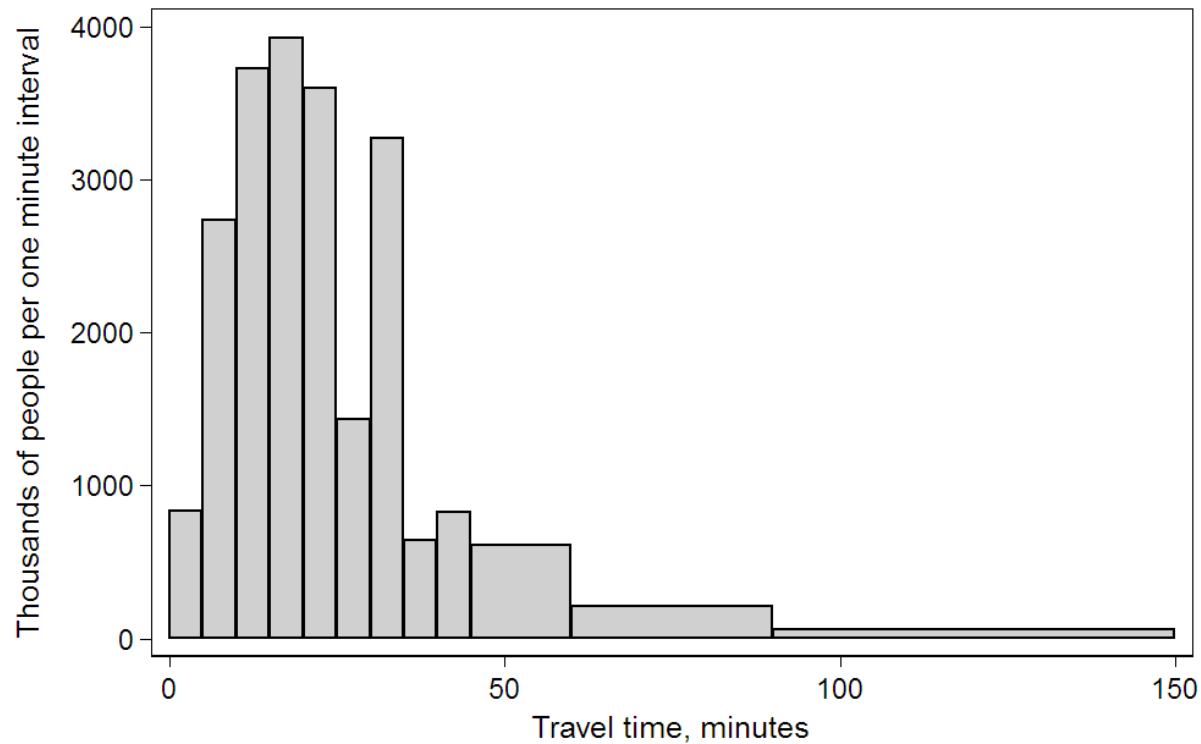


Graphic and data taken from 'Histogram' Wikipedia page.

Travel Time Histogram

(frequency density)

Interval	Width	Quantity	Quantity/ width
0	5	4180	836
5	5	13687	2737
10	5	18618	3723
15	5	19634	3926
20	5	17981	3596
25	5	7190	1438
30	5	16369	3273
35	5	3212	642
40	5	4122	824
45	15	9200	613
60	30	6461	215
90	60	3435	57



Graphic and data taken from 'Histogram' Wikipedia page.

Preparing to draw a histogram

- ▶ Divide the range of the data into **5-12 subintervals** of equal length.
- ▶ Calculate the **approximate width** of the subintervals (bins) as $\text{Range}/\text{number of subintervals}$.
- ▶ Round the approximate width up to a convenient value.
- ▶ Choose a sensible starting point. I.e. so all the data is contained nicely in the range covered by the bins
- ▶ Use the principle of **left inclusion**. I.e. each bin contains the left endpoint, but not the right.
- ▶ Create a **statistical table** including the subintervals, their frequencies and relative frequencies.

Example: Faculty

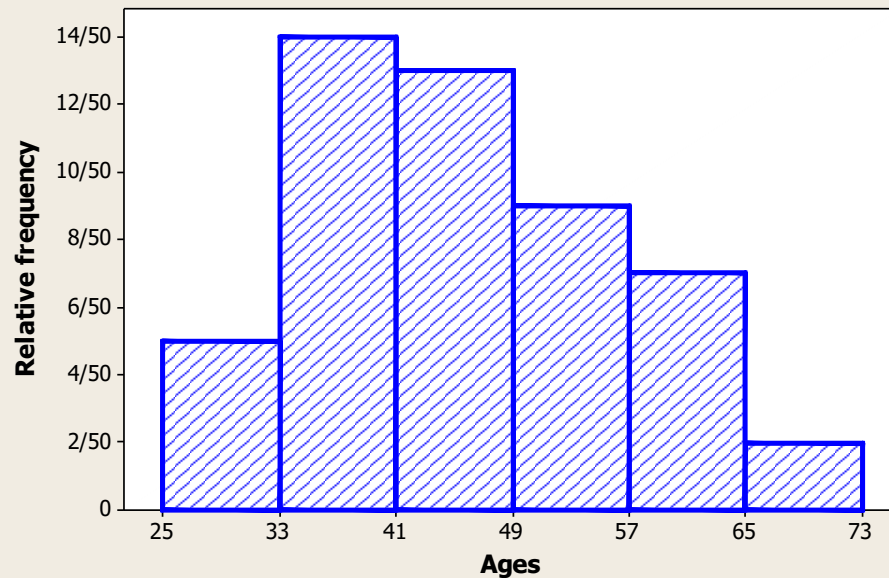
- ▶ The ages of 50 faculty members at a university.

34 48 **70** 63 52 52 35 50 37 43 53 43 52 44
42 31 36 48 43 **26** 58 62 49 34 48 53 39 45
34 59 34 66 40 59 36 41 35 36 62 34 38 28
43 50 30 43 32 44 58 53

- We choose to use **6** intervals.
- Minimum class width = $(70 - 26)/6 = 7.33$
- Convenient class width = **8**
- Use **6** classes of length **8**, starting at **25**.

Example: Faculty

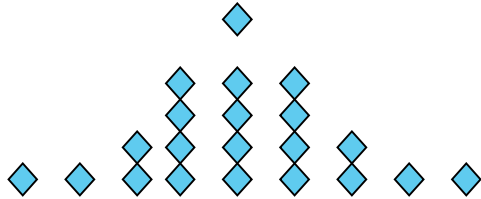
Age	Frequency	Relative Frequency	Percent
25 to < 33	5	$5/50 = .10$	10%
33 to < 41	14	$14/50 = .28$	28%
41 to < 49	13	$13/50 = .26$	26%
49 to < 57	9	$9/50 = .18$	18%
57 to < 65	7	$7/50 = .14$	14%
65 to < 73	2	$2/50 = .04$	4%



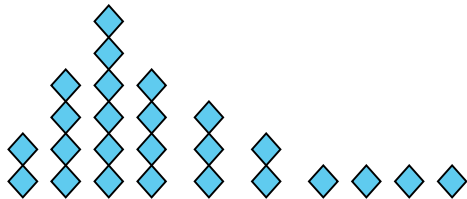
Describing distributions

- ▶ When we draw a dotplot or a histogram, we can talk about the shape of the graph.
- ▶ For example:
 - ▶ How many peaks does it have?
 - ▶ How flat or peaked is it?
 - ▶ Is it symmetric, or lopsided in some way?
- ▶ In statistics there are technical terms to describe these kinds of features precisely, but we can think about them intuitively.

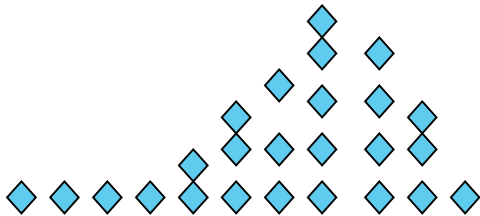
Important graph shapes



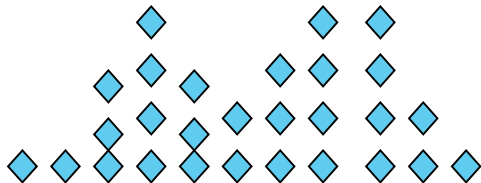
Mound shaped and symmetric



Skewed right: a few unusually large measurements



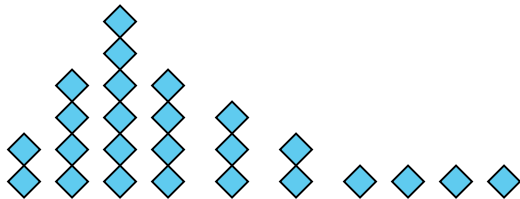
Skewed left: a few unusually small measurements



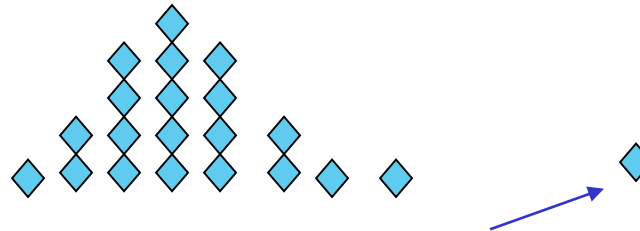
Bimodal: two peaks

Outliers

- ▶ Are there any strange or unusual measurements that stand out in the data set?



No Outliers



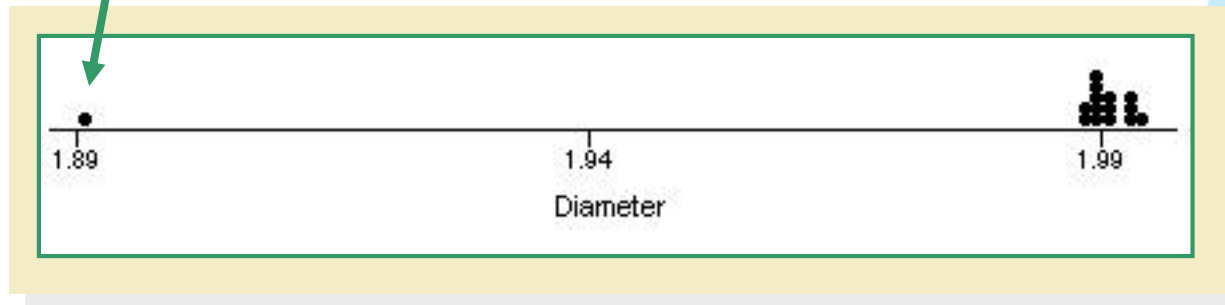
Outlier

- ▶ If there are outliers in the data, you have to think about whether you should include or exclude them.
- ▶ Sometimes outliers represent problems with the data collection, so should be excluded.
- ▶ Other times, outliers represent real events, and excluding them could be misleading.
- ▶ No easy answer.

Example: Outlier

- ▶ A quality control process measures the diameter of a gear being made by a machine (cm). The technician records 15 diameters, but she makes a typing mistake on the second entry.

1.991 1.891 1.991 1.988 1.993 1.989 1.990 1.988
1.988 1.993 1.991 1.989 1.989 1.993 1.990 1.994



Class activity 6

A group of 50 biomedical students recorded their pulse rates by counting the number of beats for 30 seconds and multiplying by 2.

80	70	88	70	84	66	84	82	66	42
52	72	90	70	96	84	96	86	62	78
60	82	88	54	66	66	80	88	56	104
84	84	60	84	88	58	72	84	68	74
84	72	62	90	72	84	72	110	100	58

- a) Why are all of the measurements even numbers? *a frequency histogram*
- b) Construct a relative frequency histogram for the data.
- c) Briefly describe the shape of the distribution.

Videos

- ▶ Populations, parameters, samples and statistics:
<https://www.youtube.com/watch?v=MYjgfoNAKkk>
- ▶ Stem and leaf plots:
<https://www.youtube.com/watch?v=OaJXJduRiIE>
- ▶ Frequency histograms:
<https://www.youtube.com/watch?v=gSEYtAjuZ-Y>
- ▶ Frequency density histograms:
<https://www.youtube.com/watch?v=wtECBdpSyDQ>