Stat 201: Statistics I
Week 4

Metropolitan State University

Week 4
Examining and Summarizing Data

Section 4.1
Summarizing and Plotting Data Distributions

Frequency distributions

A **frequency** is the number of times a particular value occurs in a set of data, i.e. the count.

A **frequency distribution** (or **frequency table**) summarizes a set of data by listing the frequencies of data in categories or classes (groups).

- For categorical data, the categories are simply the possible values of the data.
- For quantitative data, the classes are usually ranges of possible values.

Frequency distribution for categorical data

Example

Favorite kind of taco $=$ {Chicken, Fish, Fish, Veggie, Chicken, Beef }

Kind of taco	Frequency
Beef	1
Chicken	2
Pork	0
Fish	2
Veggie	1

Frequency distribution for quantitative data

Example

Tacos eaten $= \{3, 0, 17, 6, 4, 3, 5\}$

Number of tacos eaten	Frequency
0 - 4	4
5 - 9	2
10 - 14	0
15 -20	1

Relative frequency

Relative frequency is the proportion (fraction) of the whole data set that resides in each category or class. When expressed as a percent it is called **percentage frequency**.

To calculate: For each class,

$$\text{Relative frequency} = \frac{\text{class frequency}}{\text{total count}}$$

$$\text{Percentage frequency} = \frac{\text{class frequency}}{\text{total count}} \times 100$$

Relative frequency example

Example

Tacos eaten	Frequency	Relative	Percentage
0 - 4	4	0.5714	57.14 %
5 - 9	2	0.2857	28.57 %
10 - 14	0	0	0 %
15 -20	1	0.1428	14.28 %
Total	7	1	100 %

Cumulative frequency

Cumulative frequency is the frequency for a class and *all previous classes*.

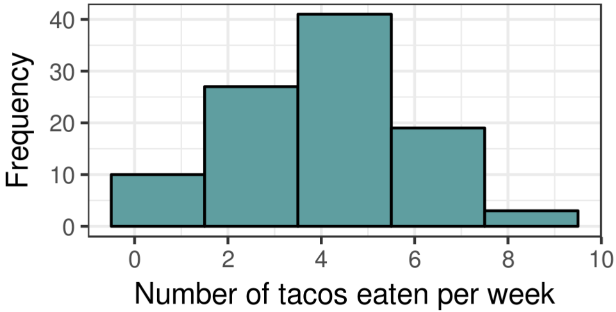
Example

Tacos eaten	Frequency	Cumulative
0 - 4	4	4
5 - 9	2	6
10 - 14	0	6
15 -20	1	7

Histograms

A **histogram** is a graphical representation of a frequency distribution of quantitative data. This allows the distribution of the data to be more easily visualized.

Num tacos	Freq
< 2	10
2 -3	27
4 - 5	41
6 - 7	19
≥ 8	3



Properties of histograms

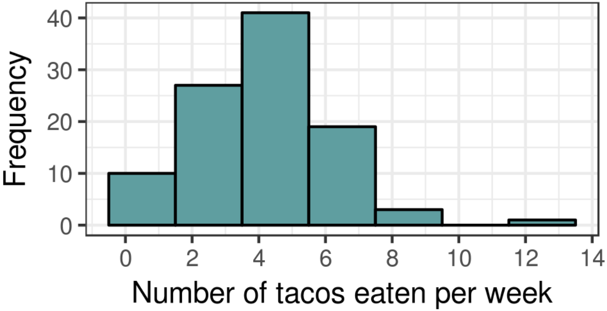
- A graph of bars of equal width drawn adjacent to each other.
- The horizontal scale (x-axis) represents values of the quantitative data. Each bar represents a class, or range of values, from a frequency table.
- The vertical scale (y-axis) represents frequency (counts), or proportions (relative frequency) or percentages (percentage frequency).
- The number of bars is largely an aesthetic choice. There should be enough bars to adequately show the shape of the distribution, but too many can make a "busy" graph that's hard to read. Most software will automatically choose the number of bars.

Outliers

An **outlier** is a data point that is distant from other data or that deviates from an established pattern.

- Outliers can result from chance, an unusual subject, or error.

Num tacos	Freq
< 2	10
2 -3	27
4 - 5	41
6 - 7	19
8 - 9	3
10 - 11	0
≥ 12	1



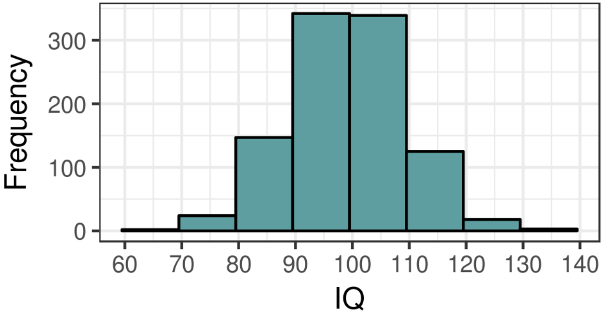
Normal distributions

A **normal distribution** can be identified from a frequency table that has the following characteristics:

- The frequencies start low, increase to a high point and then decrease to low frequencies at the end
- The frequencies are approximately symmetric around the high point.

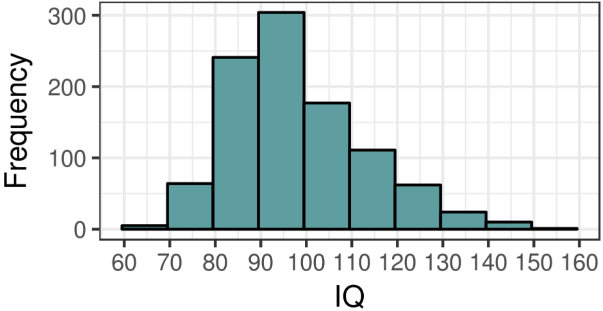
Normal distributions, example

IQ	Freq
< 70	2
70 - 80	24
80 - 90	147
90 - 100	342
100 - 110	339
110 - 120	125
120 - 130	18
130 - 140	3



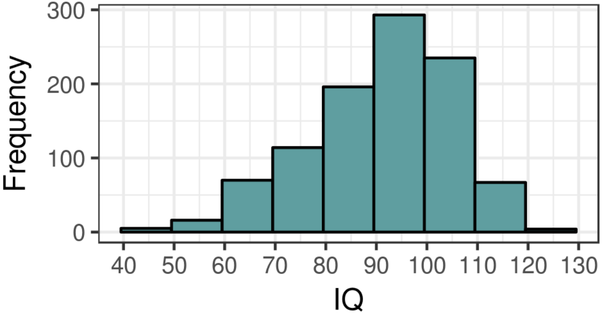
Skewed distributions, right skew example

IQ	Freq
60 - 70	5
70 - 80	64
80 - 90	241
90 - 100	304
100 - 110	177
110 - 120	111
120 - 130	62
130 - 140	24
140 - 150	10
150 - 160	1



Skewed distributions, left skew example

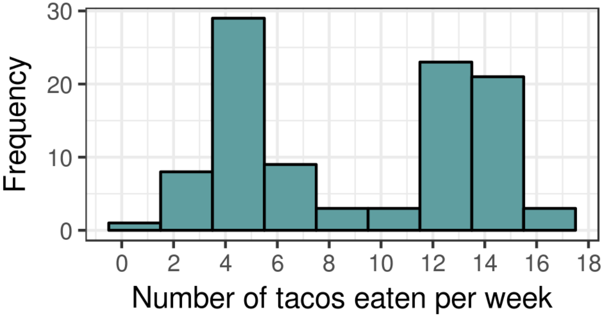
IQ	Freq
40 - 50	5
50 - 60	16
60 - 70	70
70 - 80	114
80 - 90	196
90 - 100	293
100 - 110	235
110 - 120	67
120 - 130	4



Bimodal distributions

Distributions with two peaks are known as **bimodal**. They might indicate that the data come from two different populations.

Num tacos	Freq
0 - 1	1
2 - 3	8
4 - 5	29
6 - 7	9
8 - 9	3
10 -11	3
12 - 13	23
14 - 15	21
16 - 17	3



Section 4.2
Summary Statistics

Measures of center

In order to understand a data set, values are calculated which summarize the distribution of the data or describe various properties of the data. These are, unsurprisingly, called **descriptive** or **summary** statistics.

Perhaps the most important of these, **measures of center** are a way of representing the value of the middle of the data.

There are four measures of center discussed in this section:

- mean
- median
- mode
- midrange

Mean

The **mean** (the arithmetic mean) is the measure of center calculated by adding the values of the data set and dividing by the size of the data set. Also known as the average.

- Only makes sense with quantitative data
- Sensitive to outliers (extreme or unusual values) and skewed data distributions.

To calculate

Let X be sample of size n of quantitative data with values $x_1, \ldots, x_n$. Then, the mean designated $\bar{x}$ (pronounced **x bar**), is

$$\bar{x} = \frac{\sum x_i}{n}$$

- $\sum$ means add all the x_i's, where i is between 1 and n

Mean, example

Example

Suppose we find a sample of 8 students and ask for their ages.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$
- The sample size is $n = 8$
- The sum of the data is

$$\sum x_i = 22 + 32 + 46 + 50 + 33 + 38 + 20 + 24 = 265$$

- The mean is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{265}{8} = 33.125 \text{ years}$$

Median

The **median** is the value that is greater than or equal to at least 50% of the data and less than or equal to at least 50% of the data.

- Can be used with quantitative and ordinal data (usually)
- Not sensitive to extreme values (**resistant** measure of center)

To calculate

Arrange the data in order, from lowest to highest.

- If n is odd, the middle value is the median.
- If n is even, the median is the mean of the two middle values.

Median, example

Example

Returning to the ages of 8 students.

- The sample is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$
- Arranged in order, the sample looks like

$$20 \quad 22 \quad 24 \quad 32 \quad 33 \quad 38 \quad 46 \quad 50$$

- Since n is even, find the the mean of the two middle values.

$$20 \quad 22 \quad 24 \quad \underbrace{32 \quad 33}_{(32+33)/2=32.5} \quad 38 \quad 46 \quad 50$$

- The median is $\tilde{x} = 32.5$ years.

Mean vs. Median

Suppose in our age data set, we replaced the 50 with a 85.

- Mean goes from 33.125 to 37.5
- Median remains unchanged at 32.5

This is why median is called a **resistant** statistic.

- Median is used when we don't want a few extreme values to distort a more reasonable middle, such as house prices or incomes.

Mean vs. Median, cont.

Suppose instead of calculating a grade point average (mean), we calculated a grade point median. Consider a student who got A's in 3 classes and D's in 2.

- The median grade point is 4, an A.
- The GPA for such a student would be 2.8.

The median does not consider all values of a data set. The mean does.

- Mean is used when all values are important or when we expect to have roughly symmetric data.

Mode

The **mode** is the data value with highest frequency.

- Can be used with any kind of data.
- A data set might have more than one mode, or there might not be any mode.

Mode, example

Examples

- The age data, $\{22, 32, 46, 50, 33, 38, 20, 24\}$, has no mode.
- From a TACO survey, favorite kind of taco had these responses:

  {Beef, Beef, Fish, Shrimp, Beef, Pork, Chicken, Beef, Chicken, Beef}

  The mode is “Beef” with a frequency of five.
- Suppose a sample from a class got the following grades on a quiz:

  {A, C, B, A, A, B, C, B}

  The modes are A and B, with frequencies of three each.

Midrange

The **midrange** is the value half way between the minimum and maximum values. Calculate by finding the mean of the minimum and maximum.

- Only makes sense with quantitative data.
- *Very* sensitive the extreme values.
- Easy to calculate, but rarely used.

Example

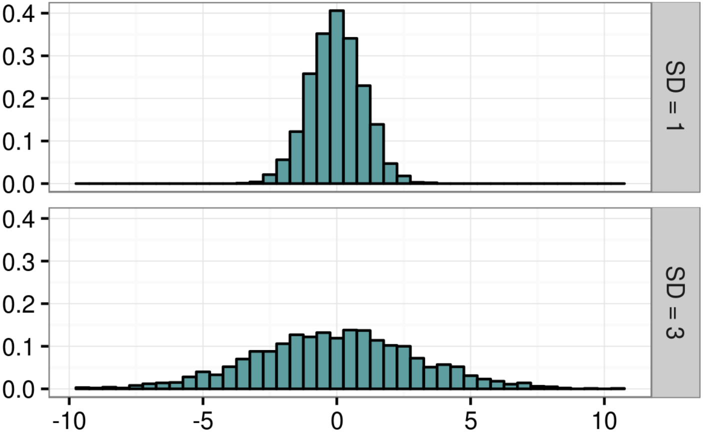
The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$.

- The minimum age is 20 and the maximum age is 50.
- The midrange is

$$\frac{\min(X) + \max(X)}{2} = \frac{20 + 50}{2} = 35$$

Variation

Center is not the only important way to describe a distribution.



Measures of variation

Another important class of descriptive statistics are **measures of variation** which describe how much the data is spread out.

There are three measures of variation discussed in this section:

- Range
- Variance
- Standard deviation

Range

The **range** is the difference between the maximum and minimum values.

- Like the midrange, very sensitive to extreme values.

Example

The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$.

- The minimum age is 20 and the maximum age is 50.
- The range is

$$\max(X) - \min(X) = 50 - 20 = 30 \text{ years}$$

Variance and standard deviation

The **variance** is the mean of the squared difference of the data from the mean. The **standard deviation** is the square root of the variance.

- More simply, the standard deviation is the average distance of the data from the data mean (the center).
- Always non-negative. A zero standard deviation means all the data are the same value.
- Sensitive to extreme values.
- The units of standard deviation are the same as the data. Variance units are the data units squared.

Variance and standard deviation, calculation

To calculate

Let X be sample of size n of quantitative data with values $x_1, \ldots, x_n$ and sample mean $\bar{x}$. Then,

$$\mathrm{Var}\,(X) = s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \qquad \text{and} \qquad \mathrm{SD}\,(X) = s = \sqrt{s^2}$$

- Note: Never calculate this by hand. Use technology.

Variance and standard deviation, example

Example

The age data is $X = \{22, 32, 46, 50, 33, 38, 20, 24\}$. The sample size is $n = 8$ and the sample mean is $\bar{x} = 33.125$

- The variance is

$$\begin{aligned} s^2 &= \frac{\sum (x_i - \bar{x})^2}{n-1} \\ &= \frac{(22 - 33.125)^2 + \cdots + (24 - 33.125)^2}{7} \\ &= 122.125 \text{ years}^2 \end{aligned}$$

- The standard deviation is

$$s = \sqrt{s^2} = \sqrt{122.125} = 11.05 \text{ years}$$

Notation

Recall, values that describe the properties of populations are called **parameters** and values that describe samples are called **statistics**. Notationally, in math formulas or when abbreviating, Greek letters are used to refer to parameters and Latin letters are used to refer to statistics.

Property	Parameter		Statistic
Mean	μ	(mu)	$\bar{x}$
Variance	σ^2	(sigma-squared)	s^2
Standard deviation	σ	(sigma)	s

Section 4.3
Summarizing Data with Graphs

Types of graphs

There are many types of graphs. Deciding which to use depends on the type of data involved and the message to be delivered.

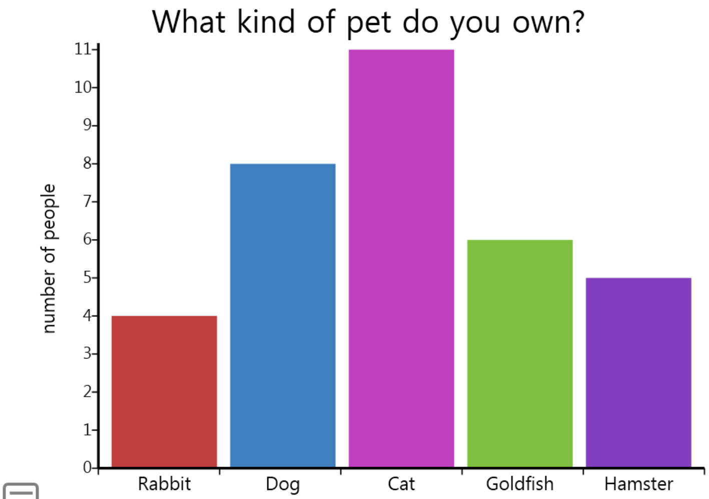
- Continuous data:
    - Histograms
    - Boxplots (to be covered later)
- Categorical data:
    - Bar graphs
    - Pareto charts
    - Pie charts
- Paired data:
    - Scatterplots
    - Time series

Types of graphs: bar graph

A **bar graph** displays frequencies of categorical data.

- The horizontal scale (x-axis) represents values of the categorical data.
- The vertical scale (y-axis) represents frequencies (or proportions or percentages).
- Often, but not always, bars are drawn with a gap between values.

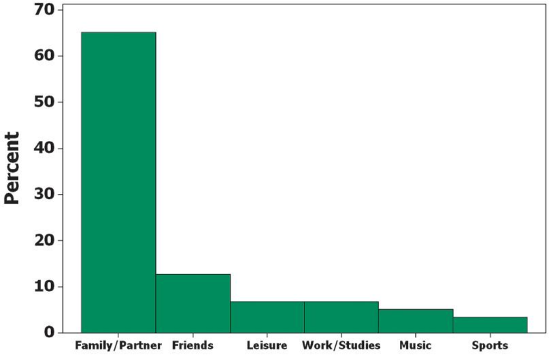
Bar graph, example



Types of graphs: Pareto charts

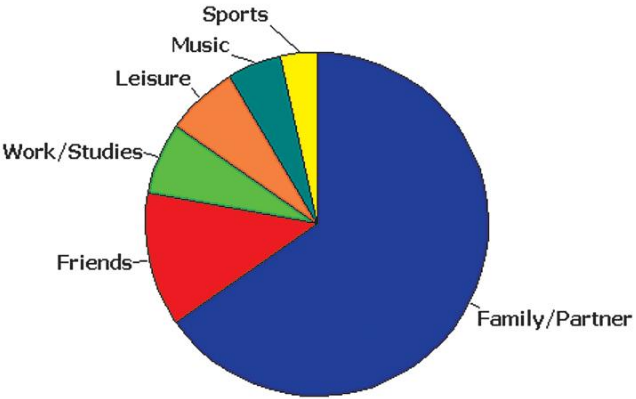
A **Pareto chart** is very similar to a bar graph, except the bars are arranged from most frequent to least, left to right.

- Can be confusing if used with ordinal data.



Types of graphs: pie charts

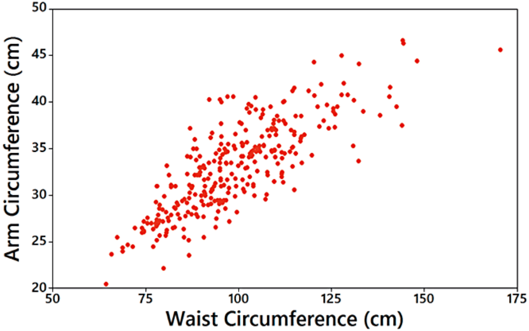
A **pie chart** displays relative frequencies of categorical data as “slices” of a whole circle. The “slices” must be labelled or distinguished by color.



Types of graphs: scatterplots

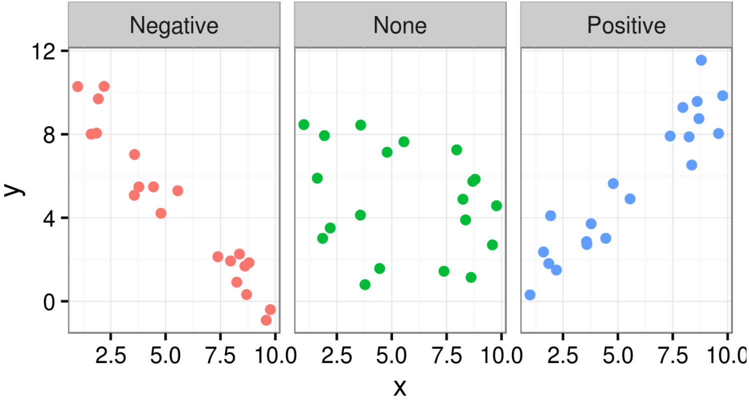
A **scatterplot** displays the relationship between paired quantitative variables.

- The x-axis represents one variable and the y-axis the other.
- A dot (or other symbol) for each data pair is placed at the appropriate x and y values.



Scatterplots and correlation

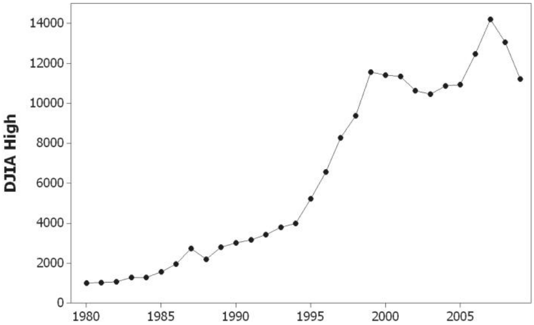
A question that can be answered with a scatterplot is whether there is an association or correlation between variables.



Types of graphs: time series

A graph of paired quantitative data where one variable represents time is called a **time series**. It is much like a scatterplot, except...

- The x-axis always represents the time variable.
- Often a line is drawn between the points.



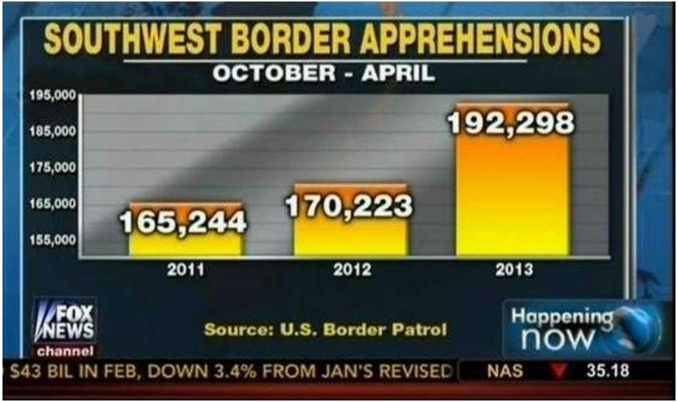
Graphs that deceive

There are two types of bad graphs:

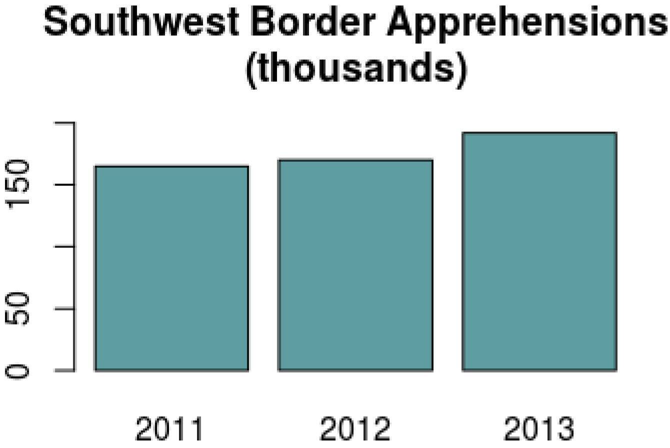
- Sometimes a graph is factually incorrect, whether because of errors in the data or a mistake in creating the graph. This is often difficult to detect without access to the original data.
- Sometimes graphs are technically correct, but designed to give a false impression of the data. Part of being a critical consumer of statistics is learning to recognize these misleading graphs.

Misleading graphs: non-zero axis

A **non-zero axis** is when one of the axis has a scale which does not include zero. This can make the relative sizes of the graph items to be distorted, especially in histograms or bar graphs.



Non-zero axis, fixed



Misleading graphs: pictographs

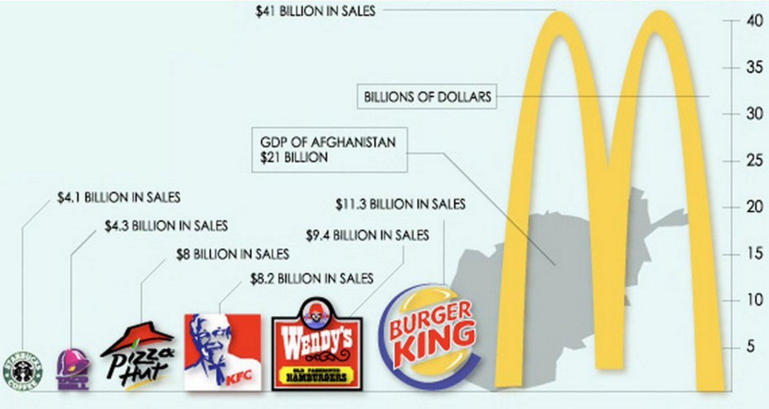
A **pictograph** uses pictures or 3D objects to represent size, rather than simple bars or points. This can also distort relative sizes.

Example

Suppose we wanted to graph the difference in sales between two oil companies, one of which is has twice the sales as the other. If we created a pictograph, we would draw the height of the larger sales twice as tall as the other.

- If we used a pictures, such as a company logos, the larger would have 4 times the area.
- If we used a 3D object, such as an oil barrel, the larger would have 8 times the volume.

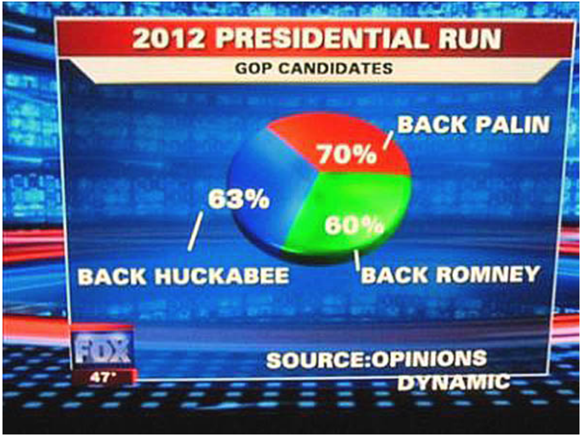
Pictograph, example



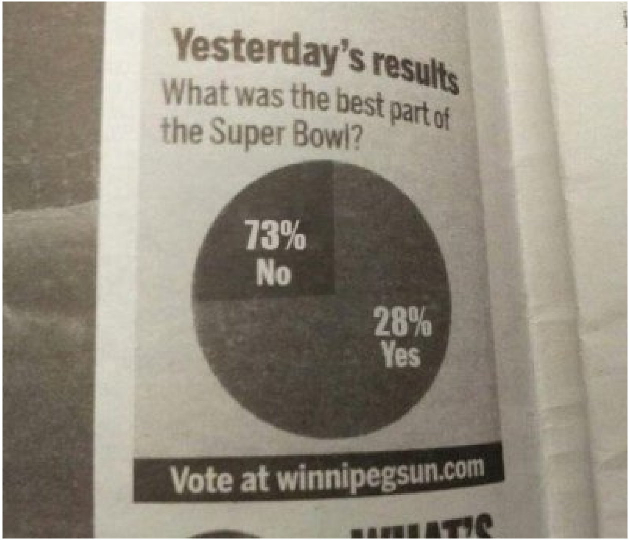
Note that KFC has twice the sales of Starbucks and McDonald's is about 4 times Burger King, but both differences appear much greater.

Misleading graphs: pie chart abuse

Since pie charts represent portions of a whole, the slices should always add up to 100%.



No. Just no.



Yesterday's results

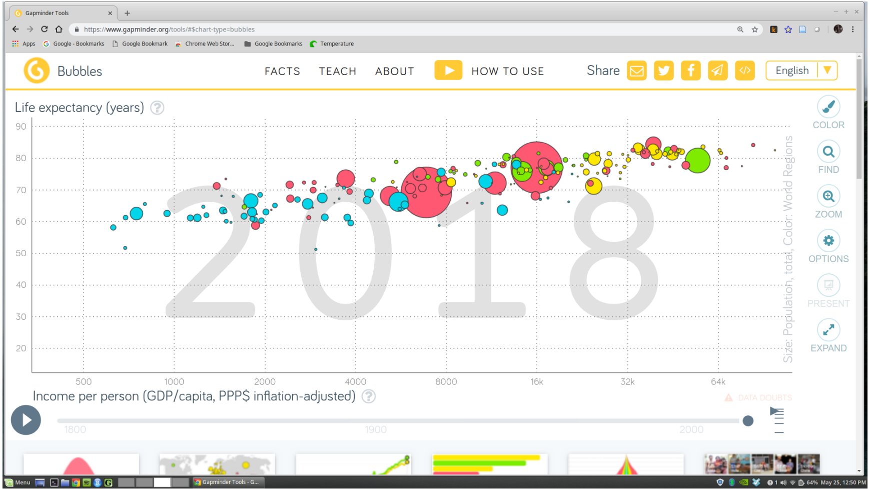
What was the best part of the Super Bowl?

Vote at winnipegsun.com

WHAT'S

Gapminder

`https://www.gapminder.org/tools/`



Gapminder Tools
https://www.gapminder.org/tools/#$chart-type=bubbles
Apps
Google - Bookmarks
Google Bookmark
Chrome Web Stor...
Google Bookmarks
Temperature
Bubbles
FACTS
TEACH
ABOUT
HOW TO USE
Share
English
Life expectancy (years)
90
80
70
60
50
40
30
20
2018
500
1000
2000
4000
8000
16k
32k
64k
Income per person (GDP/capita, PPP$ inflation-adjusted)
Size: Population, total, Color: World Regions
COLOR
FIND
ZOOM
OPTIONS
PRESENT
EXPAND
DATA DOUBTS
1800
1900
2000
Menu
Gapminder Tools - G...
64% May 25, 12:50 PM