Review: Types of Summary Statistics

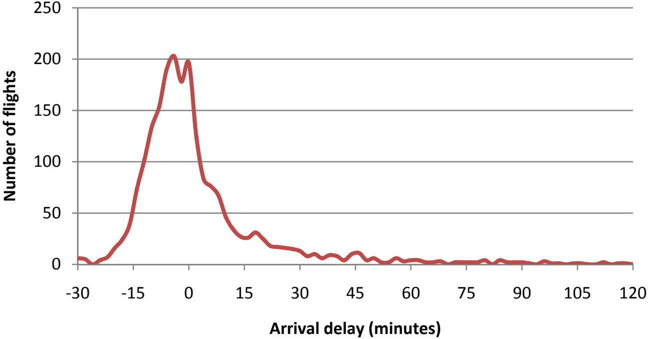
We're often interested in describing the following characteristics of the distribution of a data series:

- **Central tendency** - where is the middle of the distribution? ✓
- **Dispersion** - how spread out is the data? ✓
- **Skewness (asymmetry)** - how symmetric (or assymetric) is the distribution?
- **Peakedness** - how fat are the tails, how tall is the peak?

Measuring Symmetry (or Asymmetry)

- Typically use skewness to measure symmetry
- Right-skewed: distribution has a long right tail and data are concentrated to the left
- Left-skewed: distribution has a long left tail and data are concentrated to the right
- One way to test for right- or left-skewed is to compare median to mean:
  - Symmetric: $\bar{x} = median(x)$
  - Right-skewed: $\bar{x} > median(x)$
  - Left-skewed: $\bar{x} < median(x)$

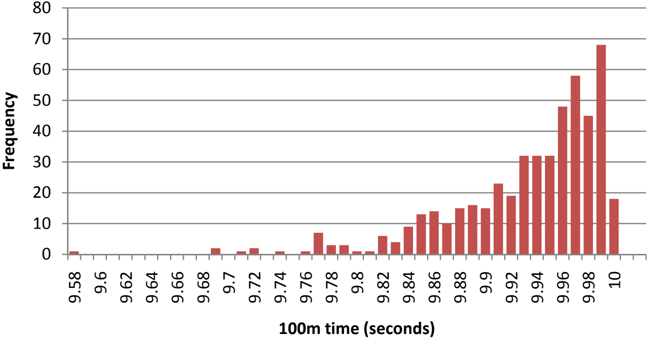
A Right-Skewed Distribution



Distribution of arrival delays for Southwest flights into SMF, January 2010

Mean = 3.4 min , Median = -2 min , Skewness = 5.0

A Left-Skewed Distribution



Distribution of the 500 fastest 100m times as of December 2010

Mean = 9.93 sec , Median = 9.95 sec, Skewness = -1.6

Quantifying Skewness

- The basic idea is to compare the mean with the median
- How we actually do it:

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3$$

- Interpretation of statistic: 0 if symmetric, greater than 0 if right-skewed, less than zero if left skewed
- Excel: use SKEW() function

Measuring “Peakedness”

- Peakedness is a question of how fat the tails of a distribution are
- Formally, we use kurtosis:

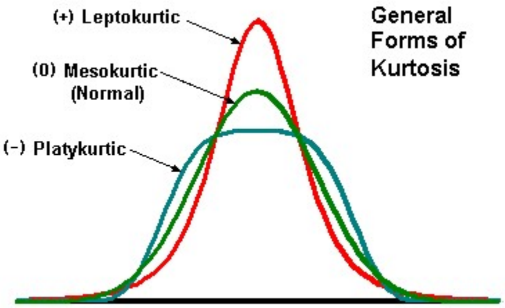
$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4-\frac{3(n-1)^2}{(n-2)(n-3)}$$

- Excel: use KURT() function

Interpreting Kurtosis

- Kurtosis has no units (because $x_i - \bar{x}$ is divided by s)
- If kurtosis is equal to 0, the distribution has the shape of the normal distribution
- If kurtosis is greater than 0, the distribution is peaked relative to the normal distribution and has fat tails
- If kurtosis is less than 0, the distribution is less peaked relative to the normal distribution and has skinny tails

Interpreting Kurtosis



Excel Demonstration

To practice generating and interpreting summary statistics, we'll use some flight delay data from SMF:

- Data are for all Southwest flights departing SMF in January and July of 2010
- These are panel data (multiple observations for each flight)
- Data are available on Smartsite (southwest-flights-2010.xlsx)

Excel Demonstration

Before we switch over to Excel, a couple of quick notes:

- Make certain that you have installed the data analysis toolpack for Excel (while not necessary for the summary statistics, it will be necessary later in the course)
- I'll show you how to add it when we switch over to Excel
- You can calculate summary statistics three ways:
  - Enter the formula as a function
  - Use the predefined function (AVERAGE, SKEW, etc.)
  - Use the descriptive statistics function under data analysis

Now to Excel ...

Graphical Representations of Univariate Data

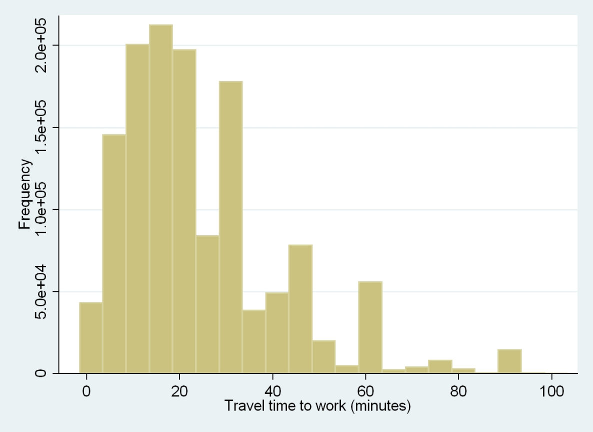
“I see you brought the pie charts.”

Graphical Representations of Univariate Data

With univariate data, we have a few different options for graphing the data. The most common are:

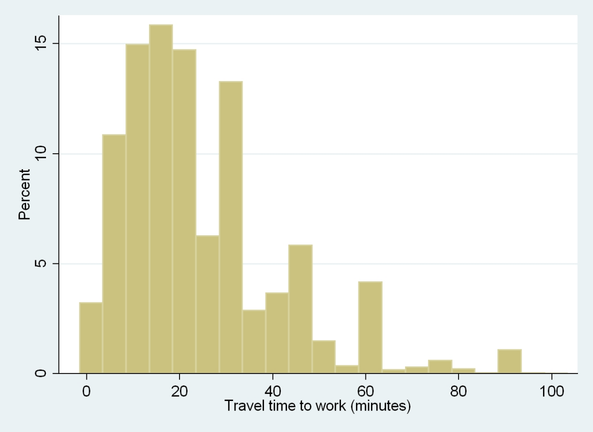
- Histograms - graphs showing the frequency of occurrence of different values
- Pie charts, bar charts, column charts - various ways to present observations that are measured in different categories
- Line charts - plots of the variable value against the observation number

A Histogram Example Using Absolute Frequencies



Data are from the 2008 American Community Survey downloaded from usa.ipums.org

A Histogram Example Using Relative Frequencies



Data are from the 2008 American Community Survey downloaded from usa.ipums.org

Histograms

There are a few choices to make when constructing a histogram.

- Whether to use *absolute frequency* or *relative frequency* for the vertical axis
  - Absolute frequency - just the number of times a particular value is observed in the data
  - Relative frequency - the number of times a value is observed as a percentage of all observations
  - Either choice will lead to the same shape for the histogram
- How large to make the bin sizes
  - If the data take on many different values, you'll want to group data into *bins*
  - In general, the more observations you have, the more bins you use

Constructing a Histogram in Excel

- Choose ‘Data Analysis’ and then select ‘Histogram’
- For input range, select the values you want to plot a histogram of
- Leave‘bin range’ blank to get automatic bins, or specify your own bin range
- Select a cell with space below and to the right of it as the ‘output range’
- Click on ‘chart output’ and optionally ‘cumulative percentage’

Now back to our flight data Excel...

Pie and Bar/Column Charts

Histograms are good for representing numerical univariate data. For categorical univariate data, we typically use pie charts or bar/column charts.

- Pie charts are perhaps the easiest way for people to visualize percentages
- Bar/column charts have the advantage of being able to show both relative and absolute frequencies
- Bar/column charts will become more useful as we start adding more variables

Creating Pie Charts in Excel

- The first step is to get frequencies for the different categories
- You can do this using the FREQUENCY() function in Excel (remember that this is an array function)
- Once you have a column of category names and a column of frequencies, highlight the values then select 'Inset', then 'Pie Chart' and choose your preferred options
- It's the same method for bar/column charts, just specificy the appropriate chart type

Back to Excel and American Community Survey data on travel to work (travel-to-work.csv)...

Line Charts

When the observations in a univariate dataset have a natural order, it often makes sense to use a line chart

- A line chart plots successive values of the data against the successive index values
- This offers an easy way to visualize whether values are getting larger or smaller
- Line charts are most common with tme series data

Constructing a Line Chart in Excel

To practice constructing a line chart, we'll use time series data on employment in California.

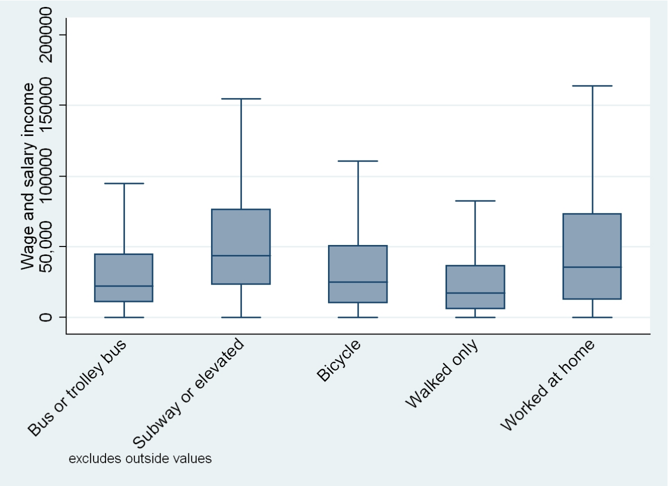
- The data are available on Smartsite (ca-urate-2000-2010.csv)
- They are monthly time series data from January of 2000 to November of 2010
- The data were downloaded from the Bureau of Labor Statistics (www.bls.gov)

To Excel...

Constructing a Line Chart in Excel

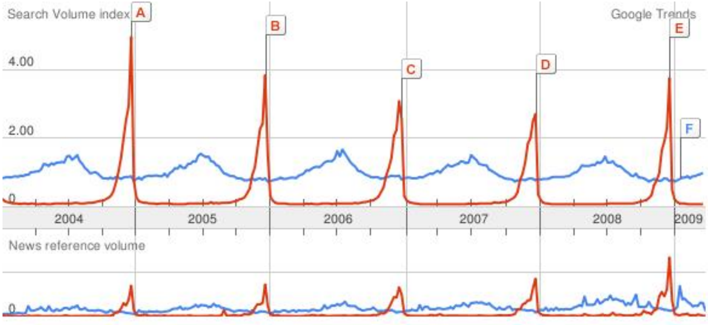
- Begin by selecting the data values that you want to graph
- Select ‘Insert’ and then ‘Line’ and then whichever type of line chart you prefer
- To get the x-axis values you want, right click on the chart and choose ‘Select data...’
- Click on the ‘Edit’ box under ‘Horizontal (category) Axis Label’ and select the cells containing your labels
- If you have graphed multiple data series on the same graph, be certain to include a legend

Summary Statistics as a Graph: The Box Plot



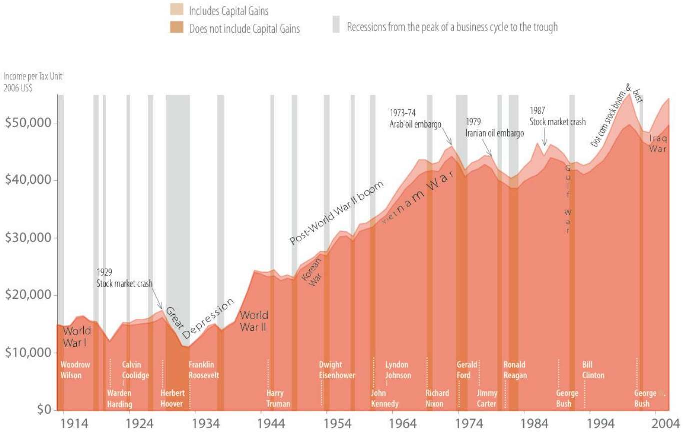
Box plot of income by form of transportation used, 2008 American Community Survey

Some Other Examples of Visual Representations of Data



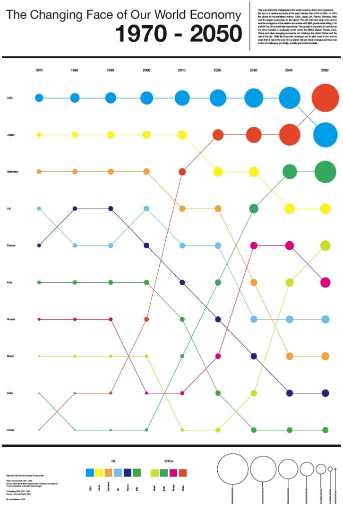
Google Trends data for the phrase "ice cream" (blue line) and the word "Santa" (red line).

Some Other Examples of Visual Representations of Data



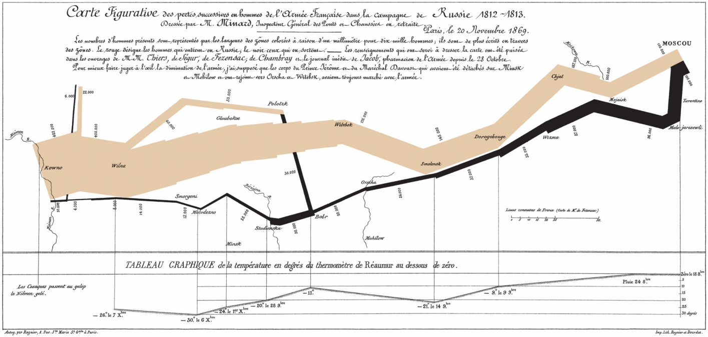
From visualizingeconomics.com

Some Other Examples of Visual Representations of Data



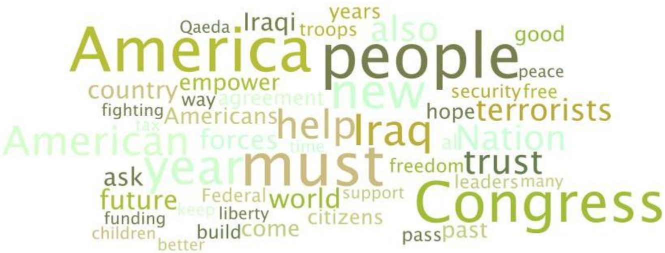
From joeswainson.blogspot.com

Some Other Examples of Visual Representations of Data


Carte Figurative des pertes successives en hommes de l'Armée Française dans la campagne de Russie 1812–1813.
Dressée par M. Minard, Inspecteur Général des Ponts et Chaussées en retraite.
Paris, le 20 Novembre 1869.
Les nombres d'hommes présents sont représentés par les largeurs des zônes colorées à raison d'un millimètre pour dix mille hommes; ils sont de plus écrits en travers des zônes. Le rouge désigne les hommes qui entrent en Russie, le noir ceux qui en sortent. —— Les renseignements qui ont servi à dresser la carte ont été puisés dans les ouvrages de M.M. Thiers, de Ségur, de Fezensac, de Chambray et le journal inédit de Jacob, pharmacien de l'Armée depuis le 28 Octobre.
Pour mieux faire juger à l'œil la diminution de l'armée, j'ai supposé que les corps du Prince Jérôme et du Maréchal Davoust qui avaient été détachés sur Minsk et Mohilow et ont rejoint vers Orscha et Witebsk, avaient toujours marché avec l'armée.
MOSCOU
Niémen R.
Kowno
Wilna
Gloubokoe
Polotzk
Witebsk
Smolensk
Dorogobouge
Chjat
Mojaisk
Tarantino
Malo-jarosewli
Wizma
Orscha
Bobr
Studienska
Molodezno
Smorgoni
Minsk
Mohilow
Berezina R.
Moskowa R.
422.000
400.000
6.000
22.000
60.000
33.000
175.000
145.000
127.100
100.000
100.000
96.000
87.000
55.000
37.000
24.000
30.000
50.000
28.000
12.000
14.000
8.000
4.000
10.000
Lieues communes de France (Carte de M. de Fezensac)
0 5 10 15 20 25 30
TABLEAU GRAPHIQUE de la température en degrés du thermomètre de Réaumur au dessous de zéro.
Les Cosaques passent au galop le Niémen gelé.
— 26.° le 7 X.bre
— 30.° le 6 X.bre
— 24. le 1.er X.bre
— 20.° le 28 9.bre
— 11.°
— 21.° le 14 9.bre
— 9.° le 9 9.bre
Pluie 24 8.bre
Zéro le 18 8.bre
5
10
15
20
25
30 degrés
Autog. par Regnier, 8. Pas. S.te Marie S.t G.ain à Paris.
Imp. Lith. Regnier et Dourdet.

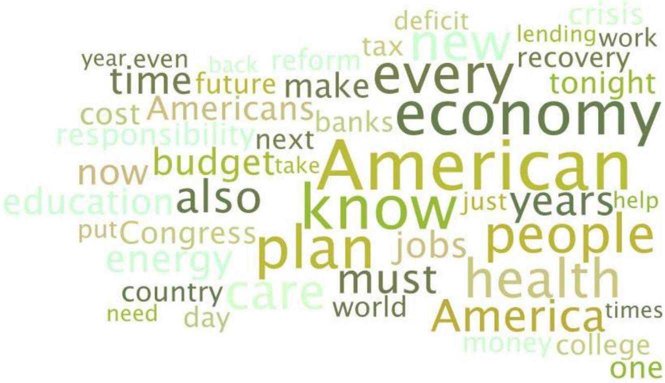
Map of Napoleon's Russian campaign of 1812, Charles Joseph Minard (1861)

Some Other Examples of Visual Representations of Data



Wordle generated from Bush's 2002 State of the Union address (after 9/11).

Some Other Examples of Visual Representations of Data



Wordle generated from Obama's 2009 State of the Union address (after start of recession).