# Exploratory Data Analysis: Describing Sample Data

Johannes Wissel

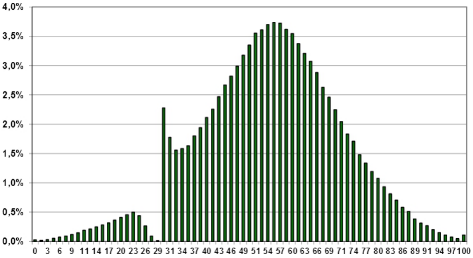
School of Operations Research & Information Engineering
Cornell University

Spring 2024

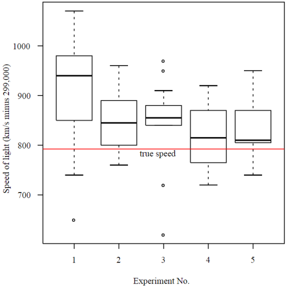
Reading: Devore Chapter 1

# Visual Summaries (aka. Plots)

Use them to *summarize & understand* data, and to *detect outliers*.



Histogram of scores from the Matura exam (2013, Poland).



Boxplot of speed-of-light measurements; Michelson experiment (1882).

# Stem & Leaf Plots

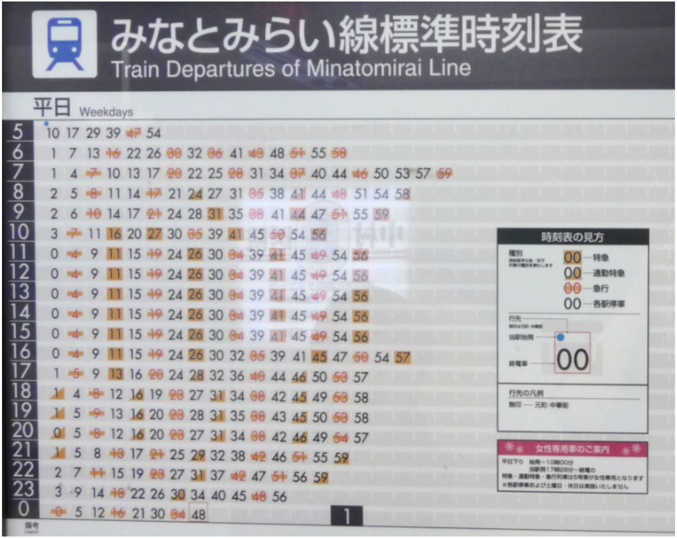
**Stems** from leading digits, and **leaves** from trailing digit.

**Example:** The income of 10 individuals (in thousands of dollars) is 49, 53, 57, 59, 61, 68, 68, 69, 72, 74. The stem & leaf plot for this data is:

| | | | | |
|---|---|---|---|---|
| 4 | 9 | | | |
| 5 | 3 | 7 | 9 | |
| 6 | 1 | 8 | 8 | 9 |
| 7 | 2 | 4 | | |

- *Advantage:* Gives a sense of the distribution of the data values at a glance.
- *Disadvantage:* Only useful for small datasets.

# Another stem & leaf example



みなとみらい線標準時刻表
Train Departures of Minatomirai Line

平日 Weekdays

| | |
|---|---|
| 5 | 10 17 29 39 47 54 |
| 6 | 1 7 13 16 22 26 30 32 36 41 43 48 51 55 58 |
| 7 | 1 4 7 10 13 17 20 22 25 28 31 34 37 40 44 46 50 53 57 59 |
| 8 | 2 5 8 11 14 17 21 24 27 31 35 38 41 44 48 51 54 58 |
| 9 | 2 6 10 14 17 21 24 28 31 35 38 41 44 47 51 55 59 |
| 10 | 3 7 11 16 20 27 30 35 39 41 45 50 54 56 |
| 11 | 0 4 9 11 15 19 24 26 30 34 39 41 45 49 54 56 |
| 12 | 0 4 9 11 15 19 24 26 30 34 39 41 45 49 54 56 |
| 13 | 0 4 9 11 15 19 24 26 30 34 39 41 45 49 54 56 |
| 14 | 0 4 9 11 15 19 24 26 30 34 39 41 45 49 54 56 |
| 15 | 0 4 9 11 15 19 24 26 30 34 39 41 45 49 54 56 |
| 16 | 0 4 9 11 15 19 24 26 30 32 35 39 41 45 47 50 54 57 |
| 17 | 1 5 9 13 16 20 24 28 32 36 40 44 46 50 53 57 |
| 18 | 1 4 8 12 16 19 23 27 31 34 38 42 45 49 53 58 |
| 19 | 1 5 9 13 16 20 23 28 31 35 38 43 45 50 53 58 |
| 20 | 0 5 8 12 16 20 23 27 31 34 38 42 46 49 54 57 |
| 21 | 1 5 8 13 17 21 25 29 32 38 42 46 51 55 59 |
| 22 | 2 7 11 15 19 23 27 31 37 42 47 51 56 59 |
| 23 | 3 9 14 18 22 26 30 34 40 45 48 56 |
| 0 | 0 5 12 16 21 30 34 48 |

1

時刻表の見方

種別
時刻数字の色・形で列車の種別を表わします

| | |
|---|---|
| 00 | 特急 |
| 00 | 通勤特急 |
| 00 | 急行 |
| 00 | 各駅停車 |

行先
無印は元町・中華街

当駅始発

終電車

00

行先の凡例

無印 ― 元町・中華街

女性専用車のご案内

平日下り 始発～10時00分
当駅発17時28分～終電の
特急・通勤特急・急行列車は5号車が女性専用となります
※各駅停車および土曜日・休日は実施いたしません

# Scatter Plots

Useful for *understanding relationships* between two variables.

To make a scatter plot for two sets of values $\{x_1, \dots, x_n\}$ and $\{y_1, \dots, y_n\}$, plot the points $(x_1, y_1), \dots, (x_n, y_n)$.

**Example:** For financial data like

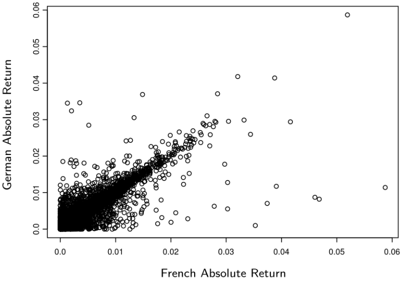
$$P_t = \text{price at } t^{\text{th}} \text{ measurement time},$$

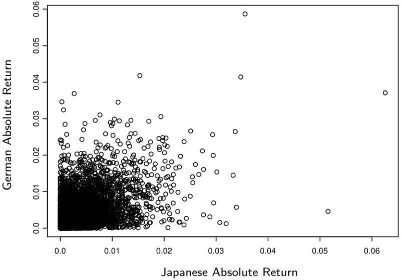
the *log return* is

$$R_t = \log \frac{P_t}{P_{t-1}} = \log P_t - \log P_{t-1},$$

which measures the change in price.

# Scatter Plots Example: Exchange rate log-returns (relative to U.S. dollar)





## What does the data say?

A. There's no relationship between the currencies.

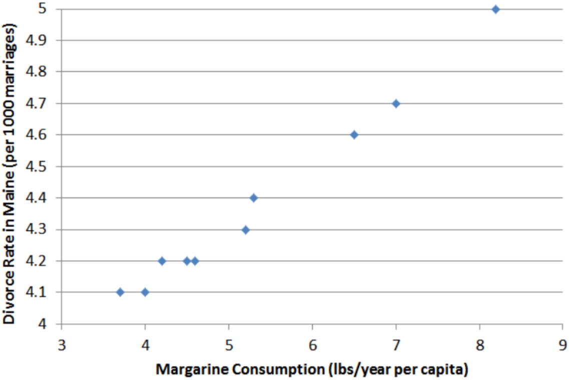
B. The French & German currencies tend to move together, because they share a major border in Europe.

C. The French & German currencies tend to move together more than the German & Japanese currencies.

D. None of the above.

# Scatter Plots

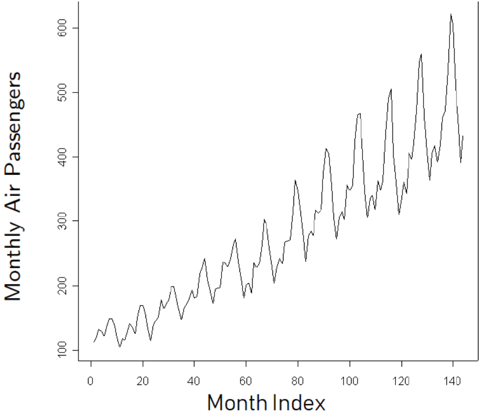
**Warning:** Correlation does not imply causation! See http://www.tylervigen.com/spurious-correlations.



Other examples: Nobel Prize vs. Chocolate, Murder vs. Internet Explorer. https://www.businessinsider.com/chocolate-consumption-vs-nobel-prizes-2014-4

# Time Series Plots

Given data $x_1, \ldots, x_T$ for times $1, \ldots, T$ , its *time series plot* is the plot of the points $(1, x_1), \ldots, (T, x_T)$.

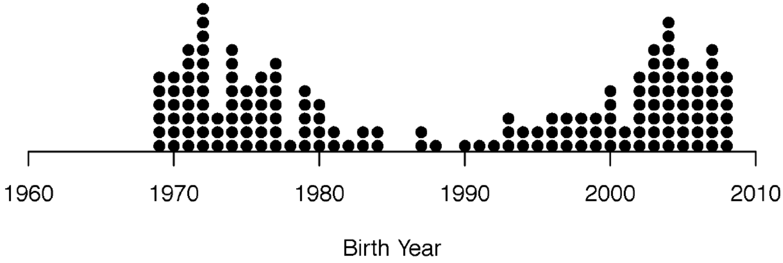


- Does it make sense to extrapolate?
- Is there a trend? Seasonality?

# Dotplots

Draw a number line and place a dot for each data point, stacking if there're repeats.

**Example:** Birth years of 160 individuals.



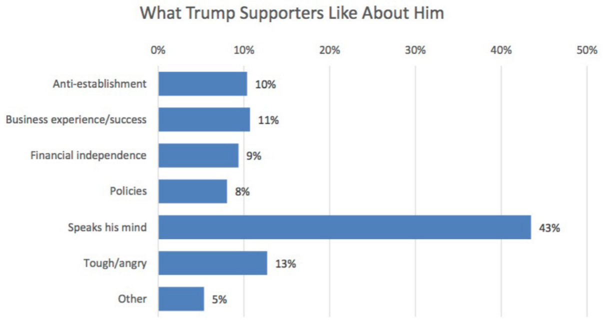
Sometimes used for *small data sets*.

# Bar Charts

For *qualitative* data.

**Examples:**

- Clicker responses;
- Numbers of people who use R, Python, Haskell, Scala, Java;
- What Trump supporters like about him.



What Trump Supporters Like About Him

n=303. Source: Working America

# Histograms

For *quantitative* data.

**Procedure:** For $n$ real-valued data points:

1. Find the smallest value $\underline{x}$ and largest value $\overline{x}$ that the data takes.
2. Partition the interval $[\underline{x}, \overline{x}]$ into disjoint intervals

$$I_1 = [\underline{x}, a_2),\ I_2 = [a_2, a_3), \ldots,\ I_{k-1} = [a_{k-1}, a_k),\ I_k = [a_k, \overline{x}].$$

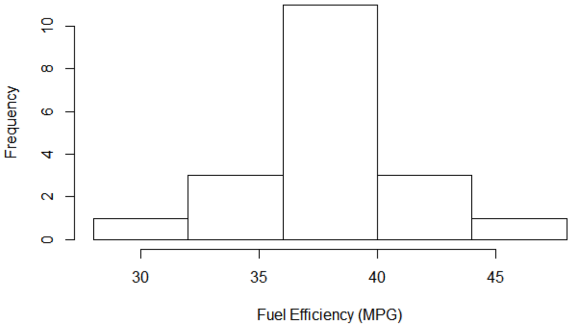
3. For each interval $I_j$, find the number of data points $n_j$ that fall in $I_j$.
4. For each interval $I_j$, compute the *relative frequency* $f_j = n_j/n$.
5. For each interval $I_j$, draw a rectangle centered at $I_j$'s midpoint whose *area is proportional to $f_j$*.

# Histograms

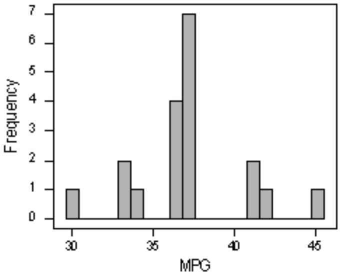
**Example:** Fuel efficiency (miles per gallon) of 19 small cars:

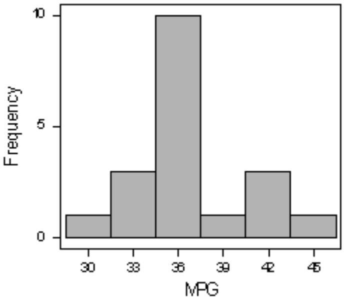
30.0 32.9 33.2 33.6 36.3 36.5 36.6 36.7 36.8 36.9
37.1 37.2 37.3 37.4 37.5 41.0 41.2 42.1 44.9

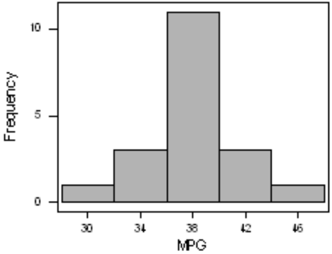
*Bin choice*: [28, 32), [32, 36), [36, 40), [40, 44), [44, 48]

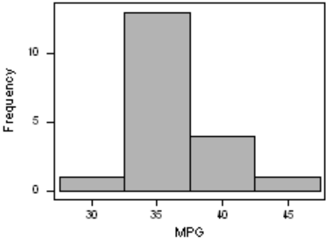


# Histograms: Bin Selection









# Histograms

Note the following properties of relative frequencies $f_j$:

- $0 \leq f_j \leq 1$;
- $f_1 + f_2 + \cdots + f_{k-1} + f_k = 1$.

**Questions:** Suppose the width of each bin is 1.

1. If the height of each rectangle is relative frequency, what's the total area of the histogram's rectangles?

2. If the height of each rectangle is frequency, what's the total area of the histogram's rectangles?

# Measures of Location

Measures that indicate *where the data tends to be situated*.

The measures of location that we'll cover are the:

1. Sample Mean;
2. Sample Median;
3. Sample Mode.

**Notation:** Consider a sample

$$x_1, \ldots, x_n$$

of size $n$ from some population.

# Sample Mean

The **sample mean** $\bar{x}$ is the average of the observations:

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

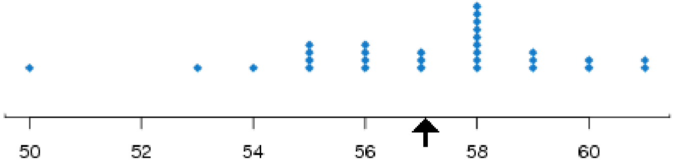
**Example:** Class of 2015 mean starting salary. http://www.career.cornell.edu/resources/surveys/upload/2015_Navy_Camel.pdf

| College | Salary $000 |
|---|---|
| Agriculture and Life Sciences | 58 |
| Architecture, Art, and Planning | 51 |
| Arts and Sciences | 57 |
| Engineering | 75 |
| Hotel | 54 |
| Human Ecology | 45 |
| Industrial and Labor Relations | 58 |

$$\bar{x} = \frac{58 + 51 + 57 + 75 + 54 + 45 + 58}{7} = 56.9.$$

# Sample Mean

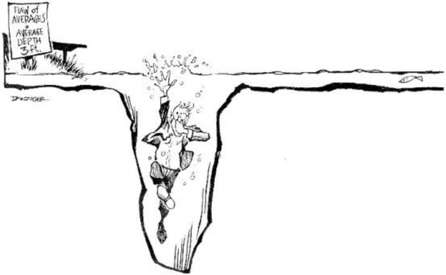
**Physical Interpretation:** Center of mass (based on the dotplot of the data):



**Question:** Is the average we computed over the 7 Cornell colleges equal to the average salary of the Cornell Class of 2015 graduate?

# Sample Mean: Caveats

1. It doesn't say much about *how the data is distributed*.
   - from http://flawofaverages.com/: "The average depth of the pool is 3 feet."



2. It's *sensitive to extreme values*.
   - If 1000 is added to the starting salary data, then $\bar{x}$ increases from about 57 to about 175.

# Sample Median

A measure of location that's less sensitive to extremes than the sample mean. (i.e., it's more *robust*)

List the sample values $x_1, \dots, x_n$ in increasing order. The **sample median** $\tilde{x}$ is

$$\tilde{x} = \begin{cases} \text{middle value in the list} & \text{if } n \text{ is odd,} \\ \text{average of the two middle values} & \text{if } n \text{ is even.} \end{cases}$$

- About 1/2 of the observations are above the sample median, and about 1/2 are below it.

**Example:** For the salary data on slide 15,

$$\tilde{x} = 57.$$

If we add 1000 to this data, then $\tilde{x} =$ .

# Sample Mode

The **sample mode** is the most frequently occurring value(s) in a sample.

**Example:** The mode of the salary sample from slide 15 is 58.

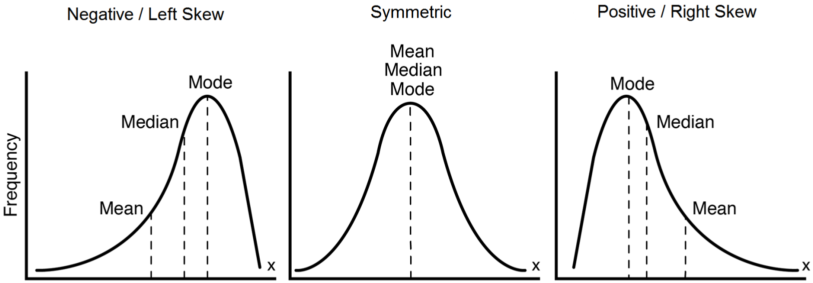
- The sample mode *might not be unique*; for example, for the sample

  $$3 \quad 3 \quad 7 \quad 14 \quad 14 \quad 23 \quad 27$$

  both 3 and 14 are sample modes.

# Mean vs. Median vs. Mode

The mean, median, and mode differ when the data is **skewed**, i.e. when one “tail” of its histogram is heavier than the other:



# Measures of Variability

Measures that indicate *how spread out the data is*.

The measures of variability that we'll cover are the:

1. Sample Range;
2. Sample Variance & Standard Deviation;
3. Interquartile Range.

# Sample Range

The **sample range** is

$$\text{(largest observation)} - \text{(smallest observation)}.$$

**Example:** The range of the salary sample

58 51 57 75 54 45 58

from slide 15 is $75 - 45 = 30$.

- Crude measure of spread (only uses extremes).
- Often increases without bound as the sample size $n$ grows – not that useful as a general measure of variability.

# Sample Variance & Standard Deviation

**Idea:** Measure the data's variability by quantifying how it's spread out around the mean $\bar{x}$.

The **sample variance** $\sigma^2$ is

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

and the **sample standard deviation** is

$$\sigma = \sqrt{\sigma^2}$$

**Example:** The sample variance of the dataset

$$3 \quad 5 \quad 7 \quad 9$$

is $\frac{1}{4-1}[(3-6)^2 + (5-6)^2 + (7-6)^2 + (9-6)^2] = \frac{20}{3}$.

# Scaling properties

The dataset

$$3 \quad 5 \quad 7 \quad 9$$

has sample variance $\frac{20}{3}$ and sample standard deviation $\sqrt{20/3}$.

The dataset

$$30 \quad 50 \quad 70 \quad 90$$

has sample variance and sample standard deviation

If you scale every value in a data set by a constant $c$ (think of changing units from, e.g., inches to centimeters), then the sample mean is scaled by , the sample variance is scaled by and the sample standard deviation is scaled by

# Order Statistics

Given the data

$$x_1, x_2, \ldots, x_n,$$

$x_{(i)}$ refers to the $i^{\text{th}}$ largest value of the data. Note that

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}.$$

The $x_{(i)}$'s are called the **order statistics** of the data.

# Quantiles: Rough Definition

Roughly, a $q$-quantile of the data $x_1, \ldots, x_n$ is the smallest number greater than or equal to $100q$ percent of the values in the data.

A $p$-percentile is the smallest number that's greater than or equal to $p$ percent of the values in the data (i.e. $p$-percentile $=$ $(p/100)$-quantile).

**Example:** For the salary data

58 51 57 75 54 45 58

The 0.5 quantile = the 50th percentile is

Suppose this salary data gave the median (0.5 quantile) starting salary within each college, and not the mean. Question: Would $57,000 be the median starting salary of Cornell graduates overall?

# Quantiles: Standard Definition

For the $i^{\text{th}}$ smallest value $x_{(i)}$ of the data, let

$$q_i = \frac{i - 0.5}{n}.$$

The ***q*-quantile** is defined as follows:

- If there's an $i$ where $q = q_i$, then the $q$-quantile is $x_{(i)}$.
- If $q$ is between $q_i$ and $q_{i+1}$, the $q$-quantile is

  $$\frac{x_{(i)} + x_{(i+1)}}{2}.$$

- If $q \leq q_1$, the $q$-quantile is $x_{(1)}$.

# Quantiles: Example

**Salary data:**

| i | $x_{(i)}$ | $\frac{i-0.5}{n}$ |
|---|---|---|
| 1 | 45 | 0.071 |
| 2 | 51 | 0.214 |
| 3 | 54 | 0.357 |
| 4 | 57 | 0.5 |
| 5 | 58 | 0.643 |
| 6 | 58 | 0.786 |
| 7 | 75 | 0.929 |

0.5-quantile $=$ 57; 0.25-quantile $=(51+54)/2=52.5$.

# Interquartile Range (IQR)

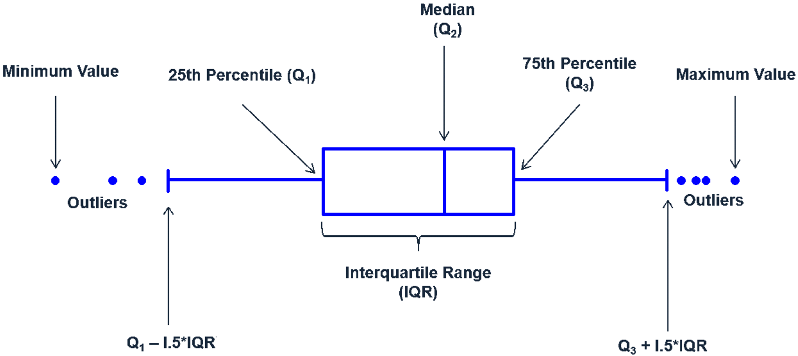
The third measure of variability (others were range, sample variance & standard deviation) we'll cover.

The **interquartile range (IQR)** is

$$\text{IQR} = (0.75\text{-quantile}) - (0.25\text{-quantile}).$$

- Much less sensitive (i.e. more robust) to extreme values than the range and sample variance & standard deviation.

# Boxplots


Median
($Q_2$)
Minimum Value
25th Percentile ($Q_1$)
75th Percentile
($Q_3$)
Maximum Value
Outliers
Outliers
Interquartile Range
(IQR)
$Q_1$ – I.5*IQR
$Q_3$ + I.5*IQR

# Summary

- We covered some basic data *visualization* techniques.
  - When you get some data, plot it every way you can think of.
- We went over three *measures of location* and three *measures of variability*.
- Watch out for how sensitive your measure is to extreme values.
- Don't rely on numerical summaries without visual ones!