Math 1070-2: Spring 2008

Lecture 1

Davar Khoshnevisan

Department of Mathematics
University of Utah
http://www.math.utah.edu/~davar

January 9, 2009

The Course

- Syllabus: http://www.math.utah.edu/~davar
- Weekly assignments (click . . . bottom of the page)
- . . . and please:
  1. Turn off 
  2.  OK but ssssh

What is Statistics? [Sta•tis•tic: a piece of data]

- Is there a “the truth”?
  - What is the population of the U.S. today?
  - What is the average age in France today?
  - Are the “laws of physics” laws of nature?
- Two common methods for finding “the truth”:
  - A priori belief:
    - Personal/philosophical ideals
    - Scientific hunches
    - Informed opinions
  - Some sort of “inference” made from a “sample”:
    - Data gathering
    - Data analysis
    - Data presentation

Applications

- Some obvious ones:
  - Predicting elections
  - Scientific/engineering research
  - Learning about public opinion
  - Advertising . . .
- Some not-so-obvious ones:
  - National security
  - Public planning
  - Quality control
  - Public health . . .

Data [da•tum: A piece of information]

- Good data:
  - Representational
  - Non-judgemental
  - No external influences . . .
- Bad data:
  - Judgmental
  - Poor quality
  - Small size . . .

A first example

- “Do people like freshly-baked cookies”?
- Stand on 9th and 9th tomorrow from 9:00 to 11:00 a.m., and ask the first 50 people whether they do

This method has a h u g e number of faults

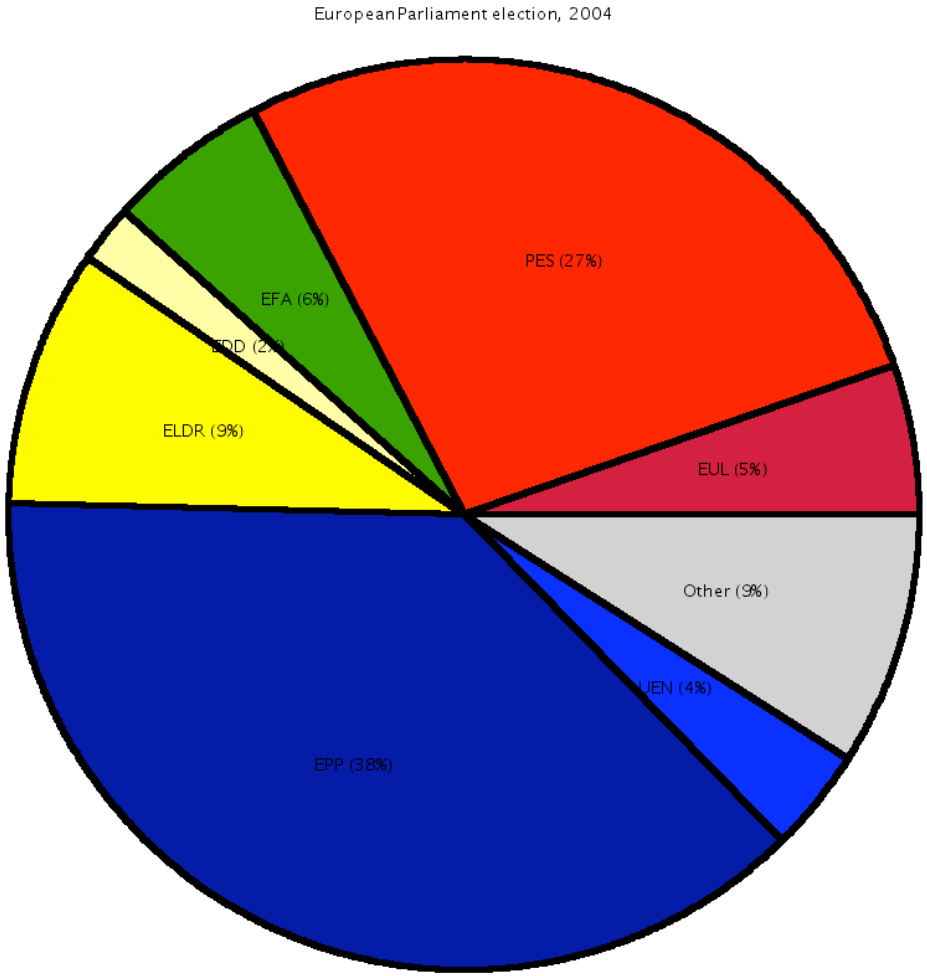
Data recap

- Gathering
- Analysis
- Representation, as well as presentation (today; why?)

Some real election data ☹

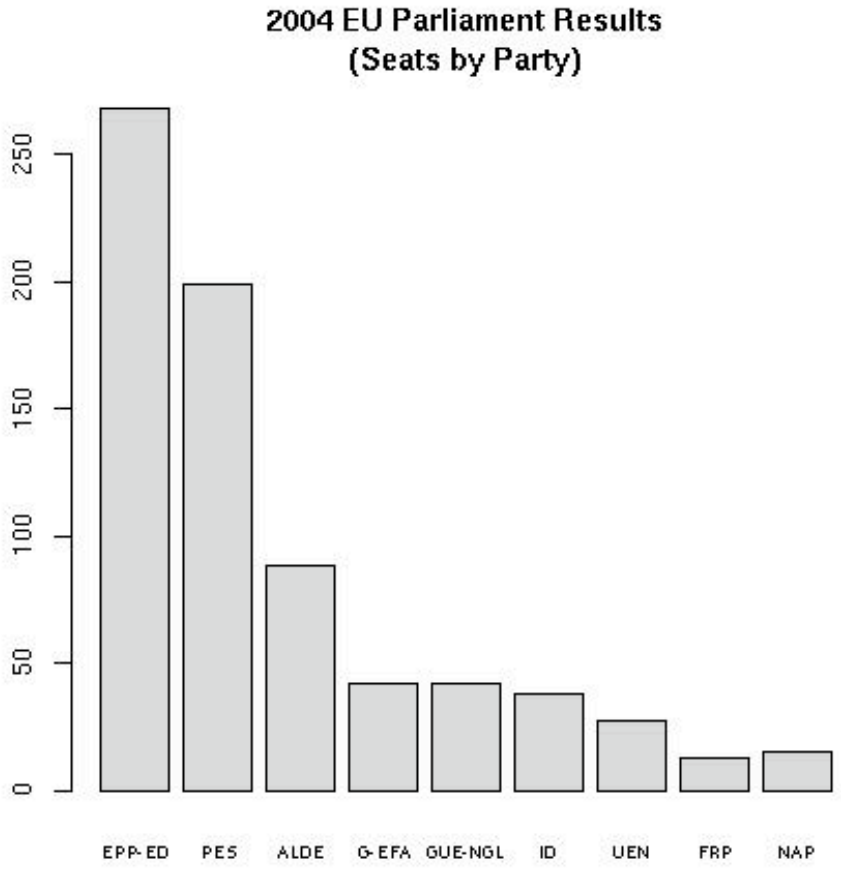
```
county technology columns under over Bush Gore Browne Nader Harris Hagelin Buchanan McReynolds Philli
Moorehead Chote McCarthy Total
Alachua Optical 1 217 105 34124 47365 658 3226 6 42 263 4 20 21 0 0 85729
Baker Optical 1 79 46 5610 2392 17 53 0 3 73 0 3 3 0 0 8154
Bay Optical 1 541 141 38637 18850 171 828 5 18 248 3 18 27 0 0 58805
Bradford Optical 2 41 695 5414 3075 28 84 0 2 65 0 2 3 0 0 8673
Brevard Optical 1 277 136 115185 97318 643 4470 11 39 570 11 72 76 0 0 218395
Broward Votomatic 1 4946 7826 177902 387703 1217 7104 54 135 795 37 74 122 0 0 575143
Calhoun Optical 1 78 0 2873 2155 10 39 0 1 90 1 2 3 0 0 5174
Charlotte Optical 2 170 2985 35426 29645 127 1462 6 15 182 3 18 12 0 0 66896
Citrus Optical 1 154 54 29767 25525 194 1379 5 16 270 0 18 28 2 0 57204
Clay Optical 1 223 157 41736 14632 204 562 1 14 186 3 6 9 0 0 57353
Collier Votomatic 1 2070 1134 60450 29921 185 1400 7 34 122 4 10 29 0 0 92162
Columbia Optical 1 76 615 10964 7047 127 258 1 7 89 2 8 5 0 0 18508
DeSoto Datavote 2 66 568 4256 3320 23 157 0 0 36 3 8 2 3 3 7811
Dixie Datavote 1 22 306 2697 1826 32 75 0 2 29 0 3 2 0 0 4666
Duval Votomatic 2 5090 21855 152098 107864 952 2757 37 162 652 15 58 41 0 0 264636
Escambia Optical 1 679 3680 73017 40943 296 1727 6 24 502 3 110 20 0 0 116648
Flagler Optical 1 60 7 12613 13897 60 435 1 4 83 3 3 12 0 0 27111
Franklin Optical 2 70 350 2454 2046 17 85 1 3 33 0 3 2 0 0 4644
Gadsden Optical 2 121 1946 4767 9735 24 139 3 4 38 4 7 6 0 0 14727
Gilchrist Datavote 1 47 241 3300 1910 52 97 0 1 29 0 2 4 0 0 5395
Glades Datavote 1 68 281 1841 1442 12 56 0 3 9 1 0 1 0 0 3365
Gulf Optical 2 47 362 3550 2397 21 86 2 4 71 2 2 9 0 0 6144
Hamilton Optical 2 31 373 2146 1722 12 37 4 1 23 8 7 4 0 0 3964
Hardee Datavote 1 84 323 3765 2339 17 75 0 2 30 0 2 3 0 0 6233
Hendry Optical 2 39 760 4747 3240 11 104 3 1 22 2 7 2 0 0 8139
Hernando Optical 1 83 148 30646 32644 116 1501 8 26 242 4 10 22 0 0 65219
Highlands Votomatic 1 466 520 20206 14167 64 545 6 16 127 3 7 8 0 0 35149
Hillsborough Votomatic 1 5431 3640 180760 169557 1138 7490 35 217 847 29 68 154 0 0 360295
Holmes Optical 1 97 40 5011 2177 18 94 1 7 76 3 6 2 0 0 7395
IndianRiver Votomatic 1 1044 790 28635 19768 122 950 4 13 105 2 13 10 0 0 49622
Jackson Optical 2 94 998 9138 6868 40 138 0 2 102 1 4 7 0 0 16300
Jefferson Datavote 1 30 540 2478 3041 14 76 2 1 29 1 0 0 0 1 5643
```

Graphical representation (Pie charts)



http://commons.wikimedia.org

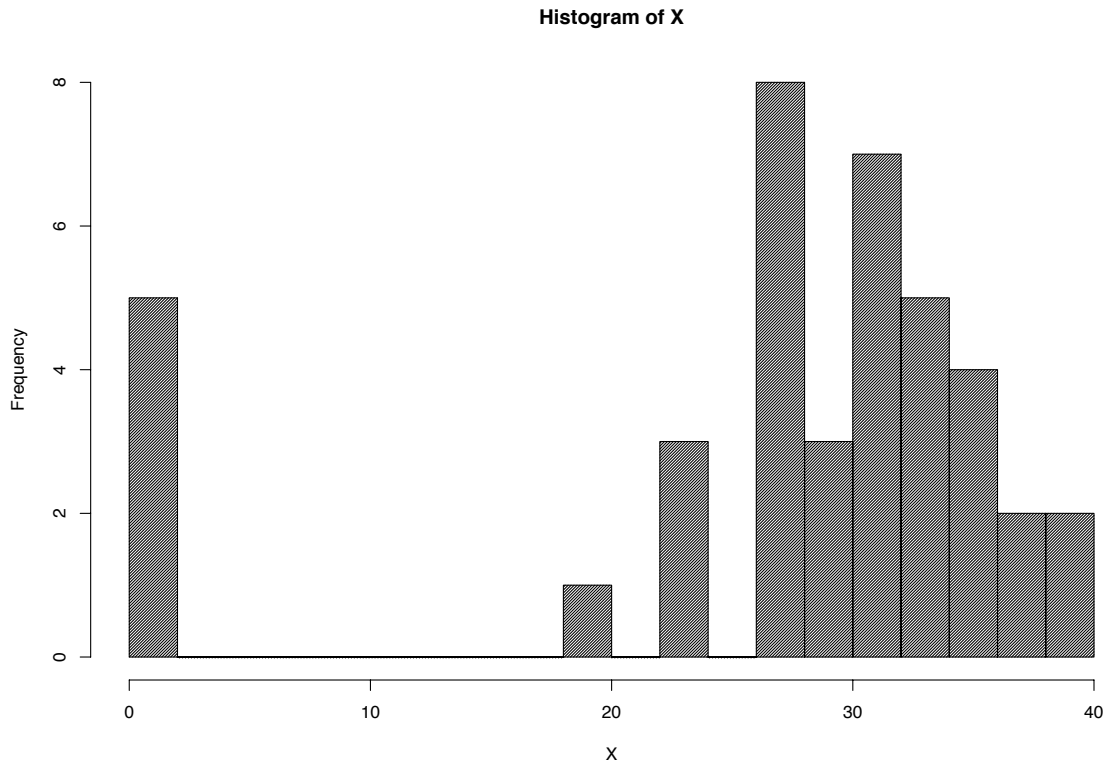
Graphical representation (Bar graphs)



http://freshmeat.net

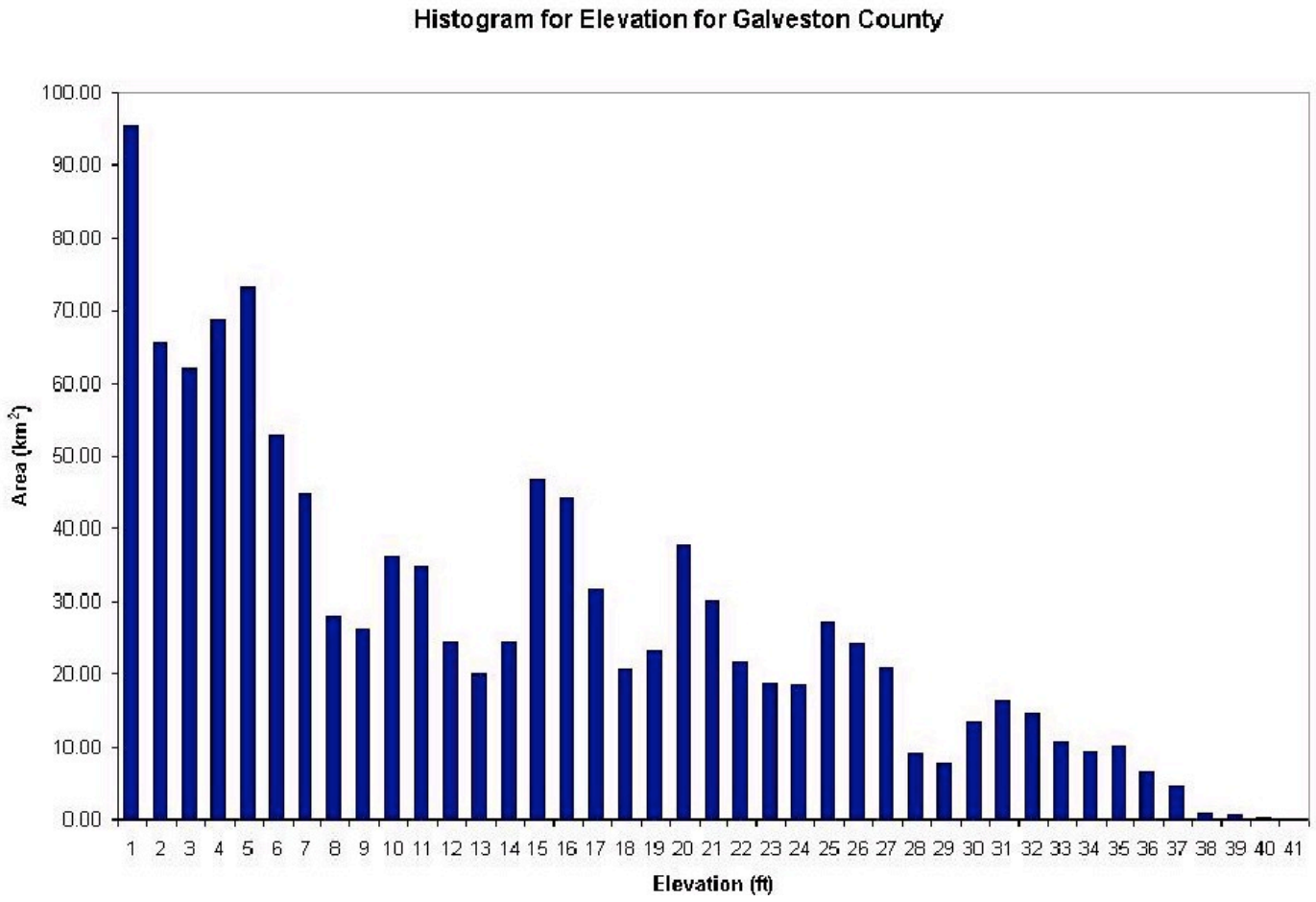
Graphical representation (Histograms)

- ☹ (old 1070 final exam) 30.8 33.1 26.3 00 30.8 34.8 36.5 27.4 35.4 40 35.4 37.7 33.1 00 24 26.3 32.6 27.4 30.8 33.7 32 22.8 30.3 28.5 34.8 26.3 28.6 27.4 33.1 22.8 00 00 30.3 00 29.7 27.4 30.8 27.4 19.4 40



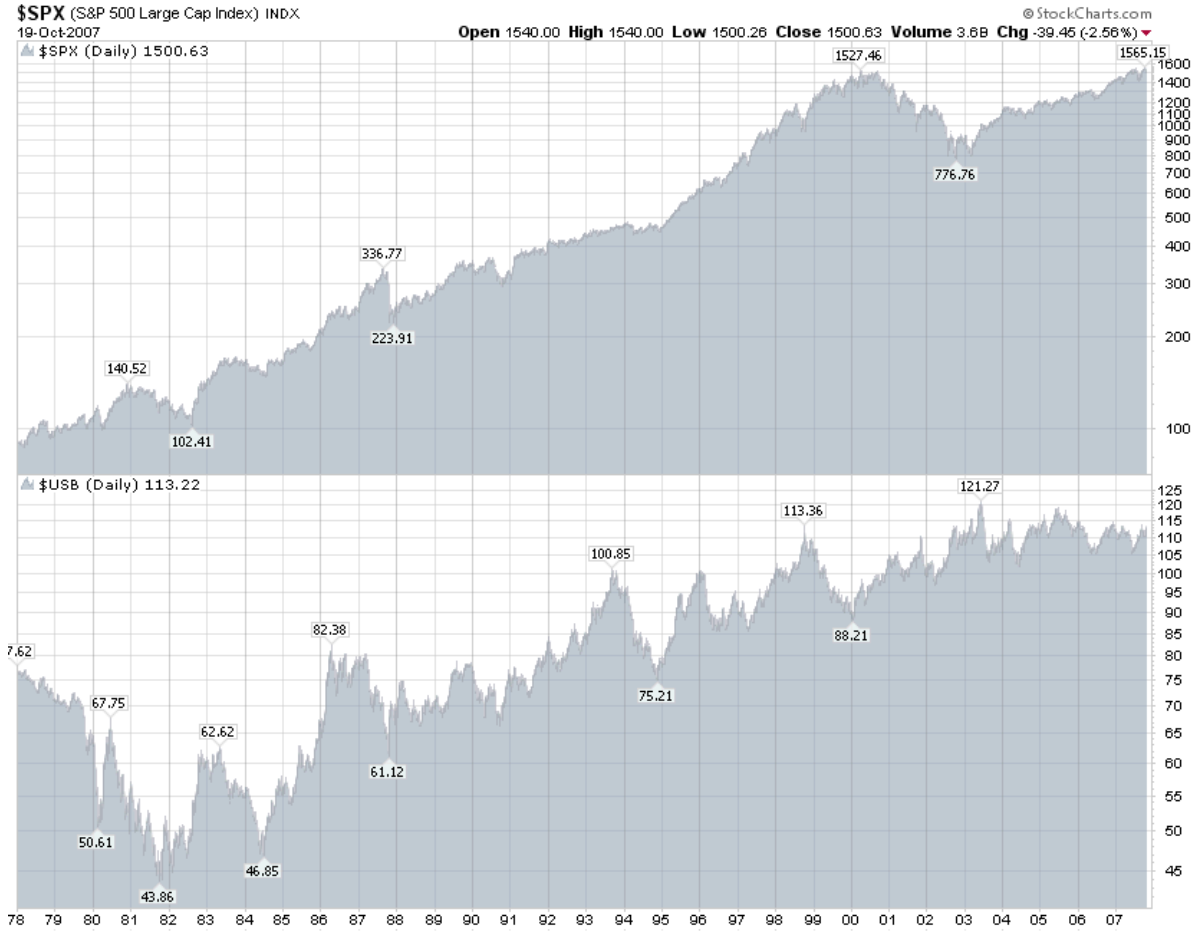
- ☺

Is this a histogram? ☹



http://www.crwr.utexas.edu

Graphical representation (time-plots) ☺



http://stockcharts.com

Sample vs. population

- Population [the generally unknown]
  - Wages of all US teens (in USD)
  - Ages of all smokers (in years)
- Sample [knowable, but is it useful?]
  - Wages of all US teens in our study
  - Ages of all smokers in our hospital
- Statistics: sample $\stackrel{?}{\longrightarrow}$ population

Numerical summaries (for the sample)

- Popular measures of centrality:
  - Mean (or average)
  - Median
- Popular measures of spread:
  - Standard deviation (or SD)
  - Interquartile range (or IQR)
- Popular 5-no. summary:
  min | 25th percentile | median | 75th percentile | max
- Range = max − min

Numerical summaries (for the sample)

- Popular measures of centrality:
  - Mean (or average)
  - Median
- Popular measures of spread:
  - Standard deviation (or SD)
  - Interquartile range (or IQR)
- Popular 5-no. summary:
  min | 25th percentile | median | 75th percentile | max
- Range = max − min N.B.: The range is a single number!

Measures of centrality (mean or average)

- Data: $x_1, \ldots, x_n$
- Avg: $\bar{x} = (x_1 + \cdots + x_n)/n$
- 1070 old final: 30.8 33.1 26.3 00 30.8 34.8 36.5 27.4 35.4 40 35.4 37.7 33.1 00 24 26.3 32.6 27.4 30.8 33.7 32 22.8 30.3 28.5 34.8 26.3 28.6 27.4 33.1 22.8 00 00 30.3 00 29.7 27.4 30.8 27.4 19.4 40
- Avg $\approx$ 26.6925 (out of 40)
- Censored Avg $\approx$ 30.50571 (removed the zero "outliers")

Measures of centrality (median)

- Order your data (from small to large, say)
- Median is the middle number if data size = odd
- Ex: 7 1 4 4 6

Measures of centrality (median)

- Order your data (from small to large, say)
- Median is the middle number if data size = odd
- Ex: 7 1 4 4 6$\longrightarrow$ Sort: 1 4 4 6 7

Measures of centrality (median)

- Order your data (from small to large, say)
- Median is the middle number if data size = odd
- Ex: 7 1 4 4 6$\longrightarrow$ Sort: 1 4 4 6 7$\longrightarrow$ Median = 4
- Median is the average of the two middle no.s if data size = even
- Ex: 6 7 1 4 4 6

Measures of centrality (median)

- Order your data (from small to large, say)
- Median is the middle number if data size = odd
- Ex: 7 1 4 4 6$\longrightarrow$ Sort: 1 4 4 6 7$\longrightarrow$ Median = 4
- Median is the average of the two middle no.s if data size = even
- Ex: 6 7 1 4 4 6$\longrightarrow$ Sort: 1 4 4 6 6 7

Measures of centrality (median)

- Order your data (from small to large, say)
- Median is the middle number if data size = odd
- Ex: 7 1 4 4 6$\longrightarrow$ Sort: 1 4 4 6 7$\longrightarrow$ Median = 4
- Median is the average of the two middle no.s if data size = even
- Ex: 6 7 1 4 4 6$\longrightarrow$ Sort: 1 4 4 6 6 7$\longrightarrow$ Median = 4.5

Measures of centrality (outliers and robustness)

- Data: 7 1 4 4 6 3

Measures of centrality (outliers and robustness)

- Data: 7 1 4 4 6 3 (sorted: 1 3 4 4 6 7)

Measures of centrality (outliers and robustness)

- Data: 7 1 4 4 6 3 (sorted: 1 3 4 4 6 7)

$$\text{Median} = 4, \qquad \text{Mean} = \frac{7+1+4+4+6+3}{6} \approx 4.17$$

- Corrupted data: 7 1 4 4 6 34

Measures of centrality (outliers and robustness)

- Data: 7 1 4 4 6 3 (sorted: 1 3 4 4 6 7)

$$\text{Median} = 4, \qquad \text{Mean} = \frac{7+1+4+4+6+3}{6} \approx 4.17$$

- Corrupted data: 7 1 4 4 6 34 (sorted: 1 4 4 6 7 34)

Measures of centrality (outliers and robustness)

- Data: 7 1 4 4 6 3 (sorted: 1 3 4 4 6 7)

$$\text{Median} = 4, \qquad \text{Mean} = \frac{7+1+4+4+6+3}{6} \approx 4.17$$

- Corrupted data: 7 1 4 4 6 34 (sorted: 1 4 4 6 7 34)

$$\text{Median} = 5, \qquad \text{Mean} = \frac{7+1+4+4+6+34}{6} \approx 9.33$$

- Badly corrupted data: 7 1 4 4 6 54

Measures of centrality (outliers and robustness)

- Data: 7 1 4 4 6 3 (sorted: 1 3 4 4 6 7)

$$\text{Median} = 4, \qquad \text{Mean} = \frac{7+1+4+4+6+3}{6} \approx 4.17$$

- Corrupted data: 7 1 4 4 6 34 (sorted: 1 4 4 6 7 34)

$$\text{Median} = 5, \qquad \text{Mean} = \frac{7+1+4+4+6+34}{6} \approx 9.33$$

- Badly corrupted data: 7 1 4 4 6 54 (sorted: 1 4 4 6 7 54)

Measures of centrality (outliers and robustness)

- Data: 7 1 4 4 6 3 (sorted: 1 3 4 4 6 7)

$$\text{Median} = 4, \qquad \text{Mean} = \frac{7+1+4+4+6+3}{6} \approx 4.17$$

- Corrupted data: 7 1 4 4 6 34 (sorted: 1 4 4 6 7 34)

$$\text{Median} = 5, \qquad \text{Mean} = \frac{7+1+4+4+6+34}{6} \approx 9.33$$

- Badly corrupted data: 7 1 4 4 6 54 (sorted: 1 4 4 6 7 54)

$$\text{Median} = 5, \qquad \text{Mean} = \frac{7+1+4+4+6+54}{6} \approx 12.67$$

Measures of spread (SD)

- For data = $x_1, \ldots, x_n$,

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Measures of spread (SD)

- For data = $x_1, \ldots, x_n$,

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\text{sum of squared deviations}}{n-1}}$$

Measures of spread (SD)

- For data = $x_1, \ldots, x_n$,

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\text{sum of squared deviations}}{n-1}}$$

- The "$n-1$" is here for technical reasons. If it were n, then SD would be the distance between the average and the typical number in the data
- Might look bad, but easy to get (esp. via a calculator)

Measures of spread (IQR)

- $Q_1 =$ 25th percentile (bigger than 25% of the data)
- Ex. Data: 7 1 4 3 6 5

Measures of spread (IQR)

- $Q_1 =$ 25th percentile (bigger than 25% of the data)
- Ex. Data: 7 1 4 3 6 5 Sort: 1 3 4 5 6 7

Measures of spread (IQR)

- $Q_1 =$ 25th percentile (bigger than 25% of the data)
- Ex. Data: 7 1 4 3 6 5 Sort: 1 3 4 5 6 7
  Data size = 6

Measures of spread (IQR)

- $Q_1 =$ 25th percentile (bigger than 25% of the data)
- Ex. Data: 7 1 4 3 6 5 Sort: 1 3 4 5 6 7
  Data size = 6 therefore 25% of the data is
  $6 \times 0.25 \approx 1.5$

Measures of spread (IQR)

- $Q_1 =$ 25th percentile (bigger than 25% of the data)
- Ex. Data: 7 1 4 3 6 5 Sort: 1 3 4 5 6 7
  Data size = 6 therefore 25% of the data is
  $6 \times 0.25 \approx 1.5$
  $Q_1 = 4$

Measures of spread (IQR)

- $Q_1 =$ 25th percentile (bigger than 25% of the data)
- Ex. Data: 7 1 4 3 6 5 Sort: 1 3 4 5 6 7
  Data size = 6 therefore 25% of the data is
  $6 \times 0.25 \approx 1.5$
  $Q_1 = 4$ (why not 3?)

Measures of spread (IQR)

- $Q_1 =$ 25th percentile (bigger than 25% of the data)
- Ex. Data: 7 1 4 3 6 5 Sort: 1 3 4 5 6 7
  Data size = 6 therefore 25% of the data is
  $6 \times 0.25 \approx 1.5$
  $Q_1 = 4$ (why not 3?)
- $Q_2 =$ median = 50th percentile (in Ex. $Q_3 = 4.5$)
- $Q_3 =$ 75th percentile (in Ex. $Q_3 = 5$)
- IQR = $Q_3 - Q_1$ (in Ex. $Q_3 = 1$; compare with SD = 1.5)
- IQR more robust against outliers than SD

A rule of thumb

- If x falls away from Q_1 or Q_3 by more than $1.5\times$ IQR, then x might be an outlier
- Is it, really? ☹
- Ex. Earlier we had $Q_1 = 4$, $Q_3 = 5$, IQR = 1.
  - $1.5 \times \text{IQR} = 1.5$
  - If $x < 4 - 1.5 = 2.5$ or $x > 5 + 1.5 = 6.5$ then we declare it an

A rule of thumb

- If x falls away from Q_1 or Q_3 by more than $1.5\times$ IQR, then x might be an outlier
- Is it, really? ☹
- Ex. Earlier we had $Q_1 = 4$, $Q_3 = 5$, IQR = 1.
  - $1.5 \times \text{IQR} = 1.5$
  - If $x < 4 - 1.5 = 2.5$ or $x > 5 + 1.5 = 6.5$ then we declare it an

Outlier

A rule of thumb

- If x falls away from Q_1 or Q_3 by more than $1.5\times$ IQR, then x might be an outlier
- Is it, really? ☹
- Ex. Earlier we had $Q_1 = 4$, $Q_3 = 5$, IQR = 1.
  - $1.5 \times \text{IQR} = 1.5$
  - If $x < 4 - 1.5 = 2.5$ or $x > 5 + 1.5 = 6.5$ then we declare it an

Outlier