Statistical Graphics:

What is Exploratory Data Analysis?

Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Describe exploratory data analysis

Exploratory Data Analysis

Approach to analyzing data sets to summarize main characteristics

+

Elements include:

- Data visualization
- Residual analysis
- Data transformations/ re-expression
- Resistance procedures

→

Exploratory Data Analysis

Data Visualization

- Data visualization facilitates advanced data analysis
- Checks distributional and other assumptions
- Observes time-based processing
- Spots outliers
- Examines relationships
- Discriminates clusters
- Compares mean differences

Data Distributions

- **The type of data distribution affects**
  - How it should be analyzed
  - How it should be visualized
- **Key step is pre-conditioning data**

The Normal Distribution

Normal (Gaussian) Distribution

- Popular
- Fully characterizes with two parameters
- Probability is determined knowing distance from mean
- Many measures and tests are designed for this

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Mean and Standard Deviation

- For sample population X = {x_1,…, x_n} the mean is defined as:

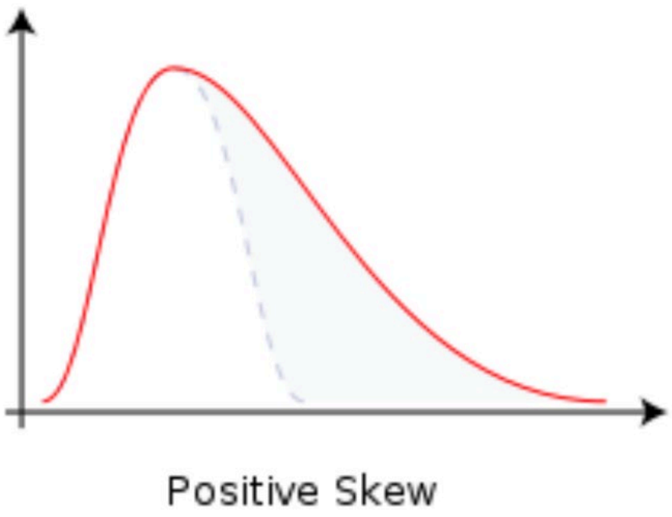
$$\mu = \frac{1}{N}\sum_{i=0}^{N} x_i$$

- The standard deviation is defined as:

$$\sqrt{\frac{1}{N}\sum_{i=0}^{N}(x_i - \mu)^2}$$

Skewness

Measure of the asymmetry of the probability distribution



Skewed Data

For a sample of N values, the sample skewness is:

$$\gamma = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)^{3/2}}$$

Statistical Graphics:

Designing Pie Charts

Ira A. Fulton Schools of Engineering
Arizona State University

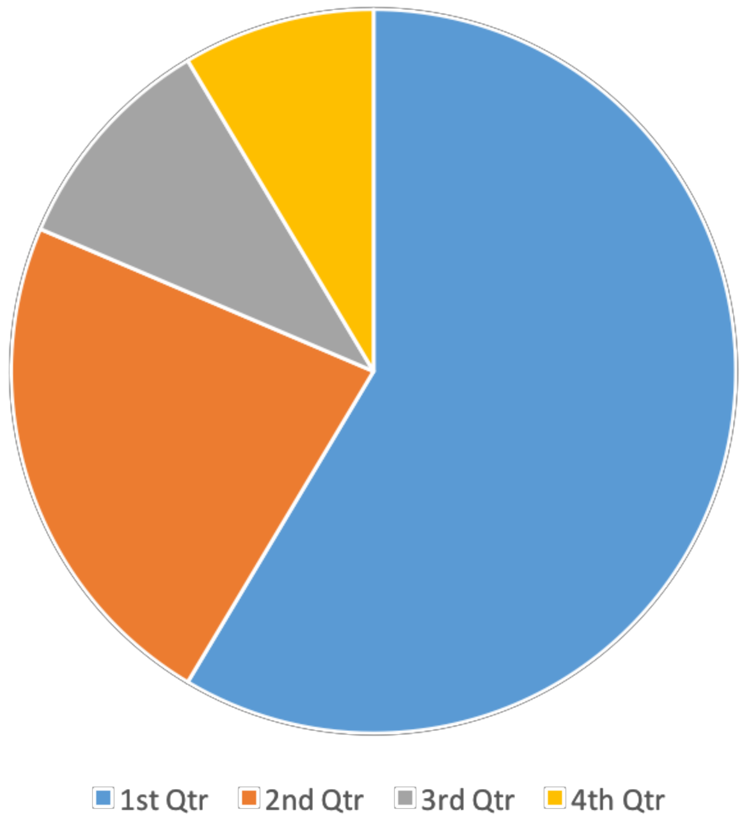
When to Use a Pie Chart

Categorical data

- Each slice can represent a different category

How many categories do you have?

- Good rule of thumb is ~7 categories maximum

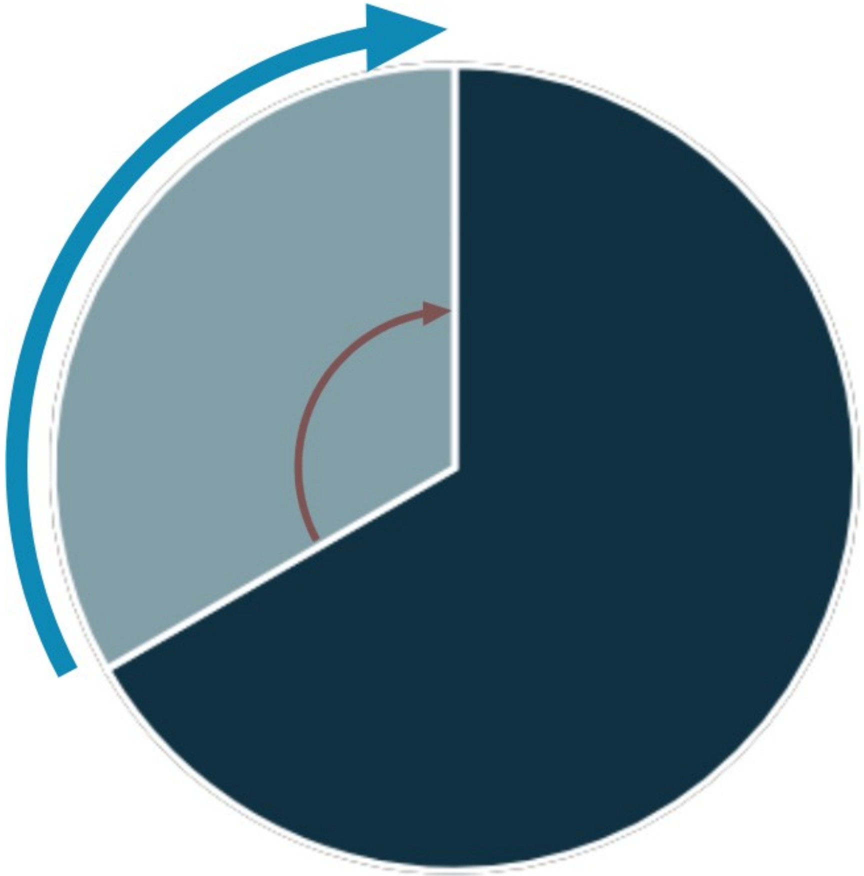


Pie Charts

Use:

- Angle
- Area
- Arc Length

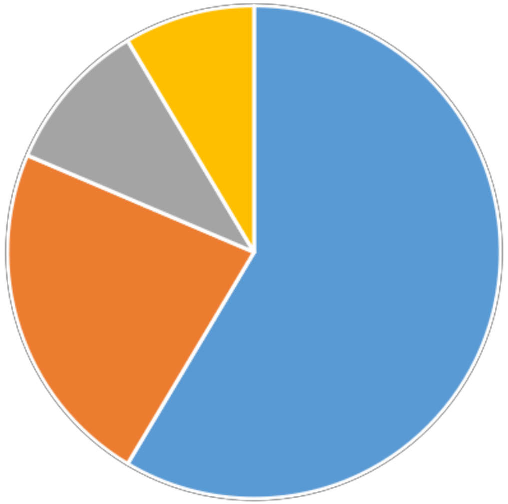
to encode values

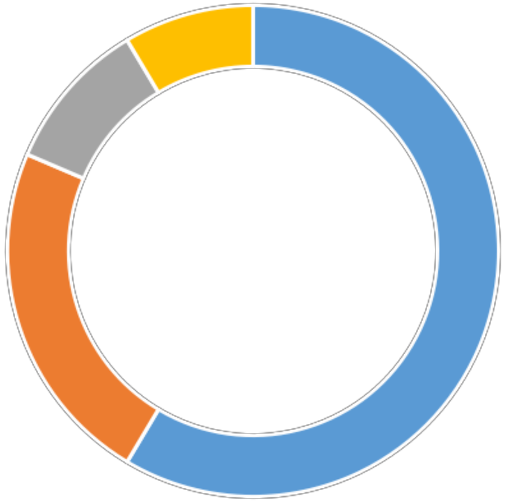


Interpreting Pie Charts

- Pie charts lead to an **overestimation** of small values and **underestimation** of large ones.
- Perceptual research shows that error is high when estimating values.
- Pie charts should be presented with values as pie slice labels.

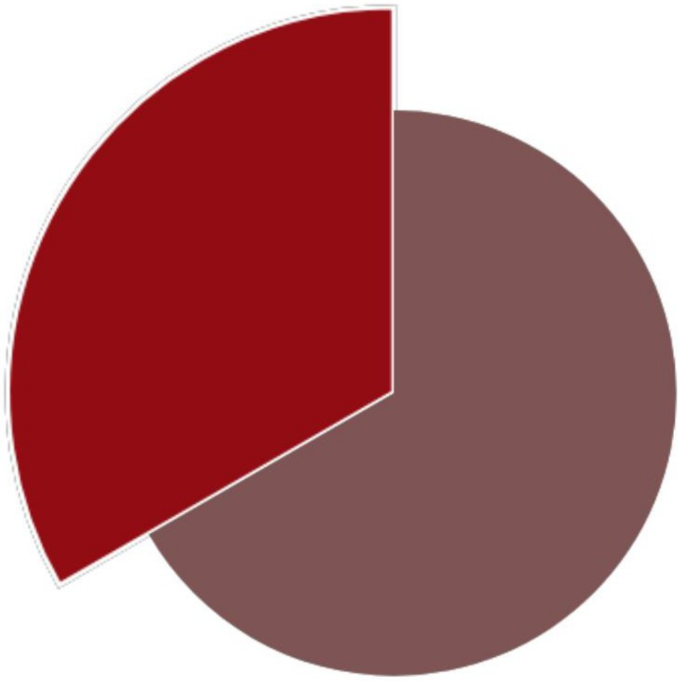
Visual Variables



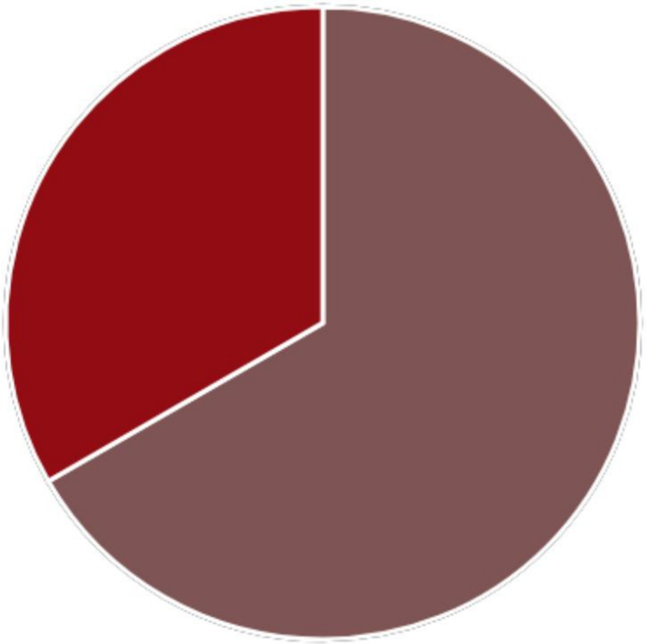


- Angle is not a key visual clue
- Arc Length and Area are important
- Doughnut charts are just as effective

Interpreting Pie Chart Variants



Larger Slice



Exploded Pie

Bar and Line Charts

Ross Maciejewski, Ph.D

Associate Professor

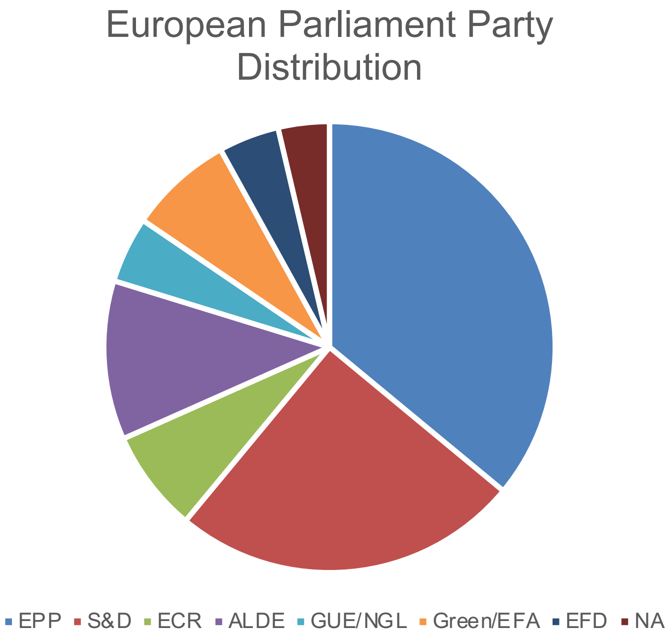
Arizona State University

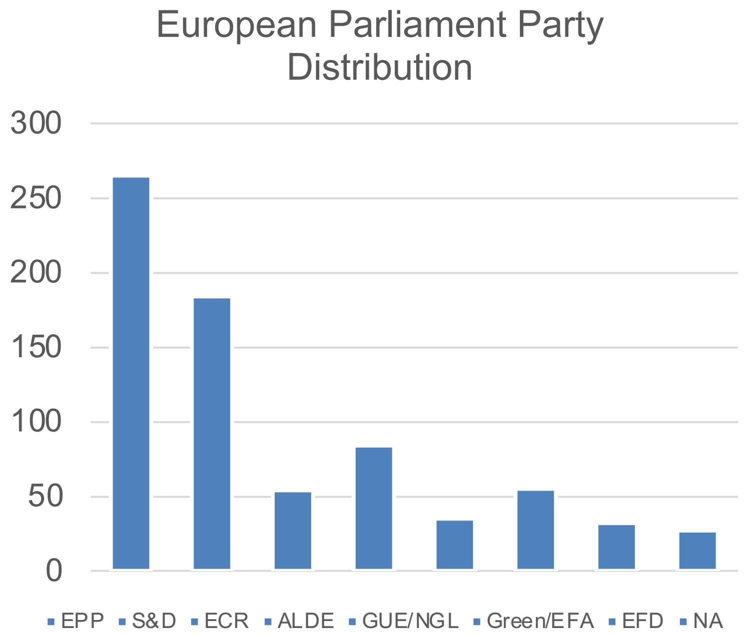
Ira A. Fulton Schools of Engineering
Arizona State University

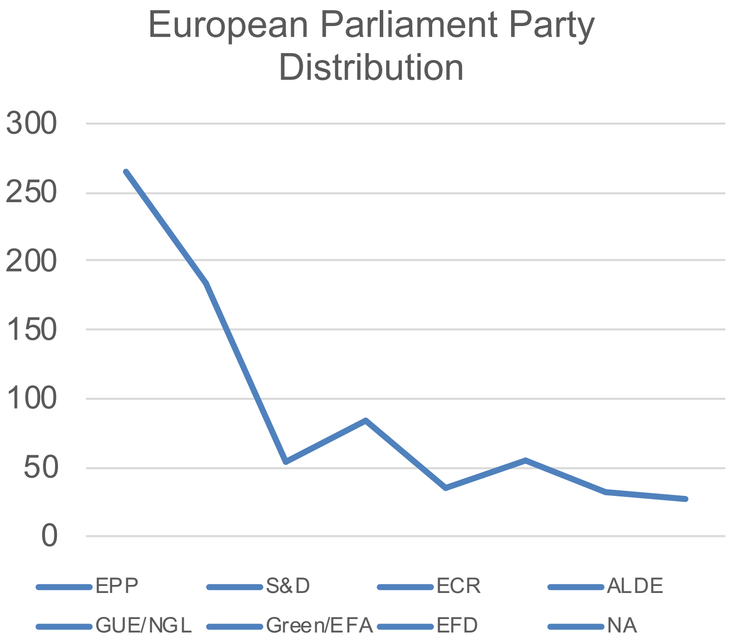
Which Type of Graph Should I Use?

- Pie Charts
  - Comparing parts of a whole
- Bar Charts
  - Comparing between groups or over time
- Line Chart
  - Changes over time

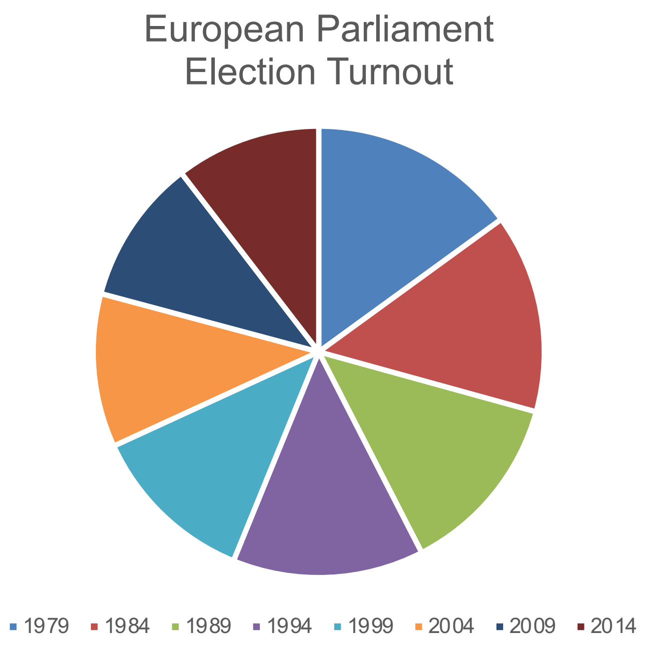
Non-Time Series Data

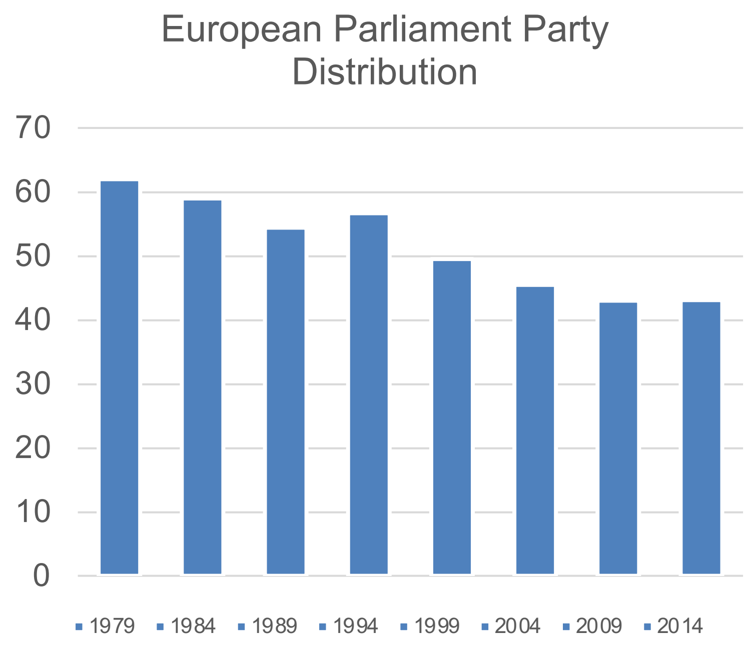


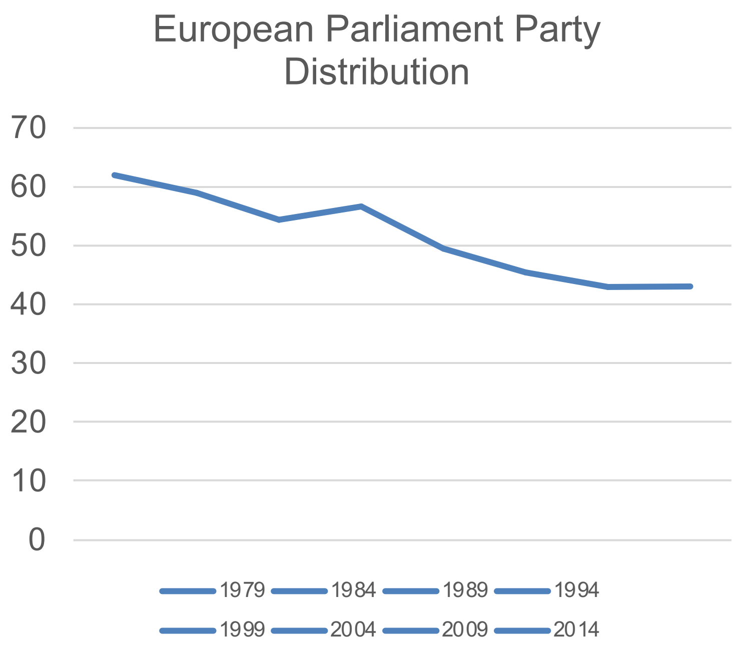




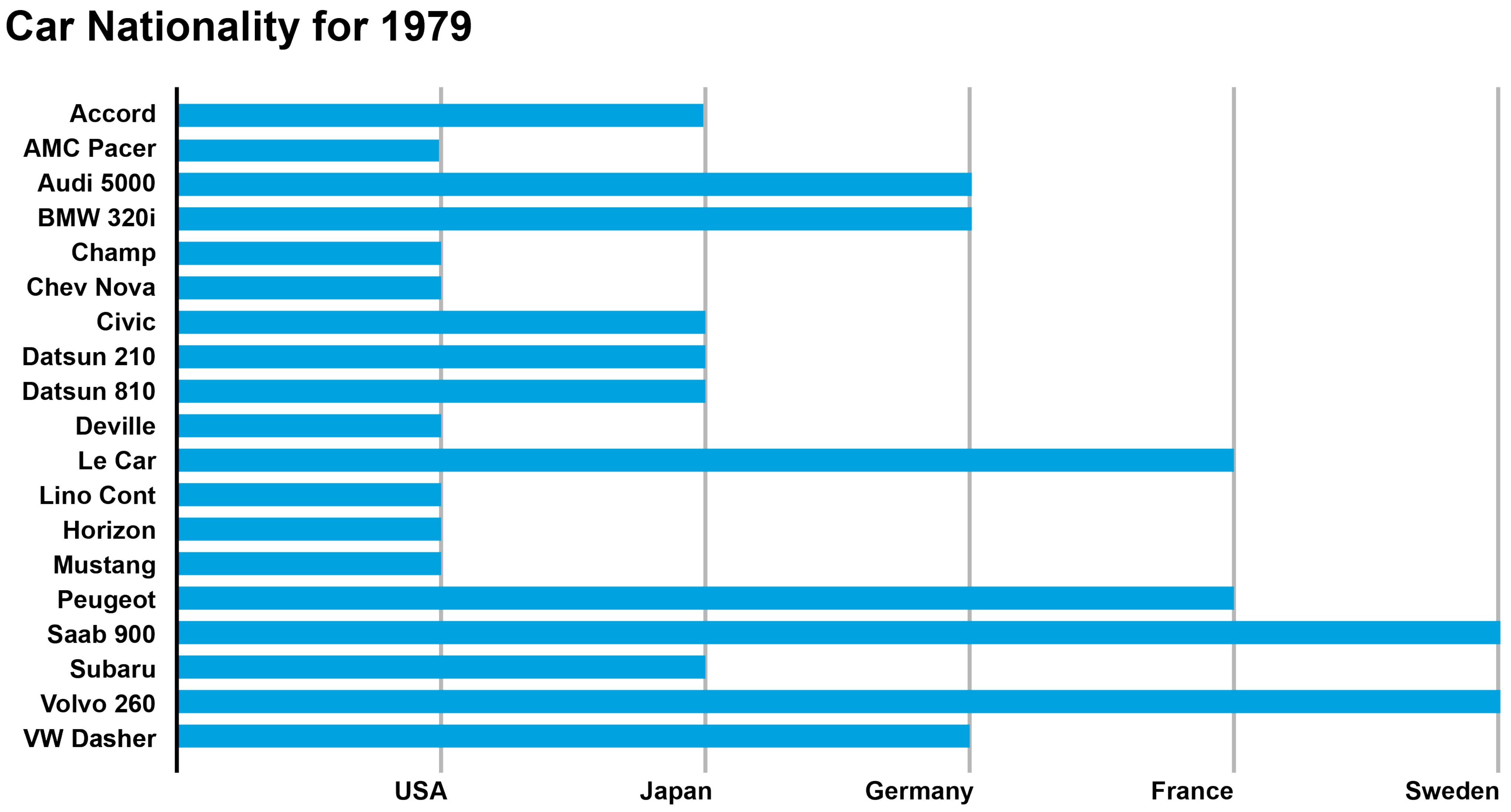
Time Series Data





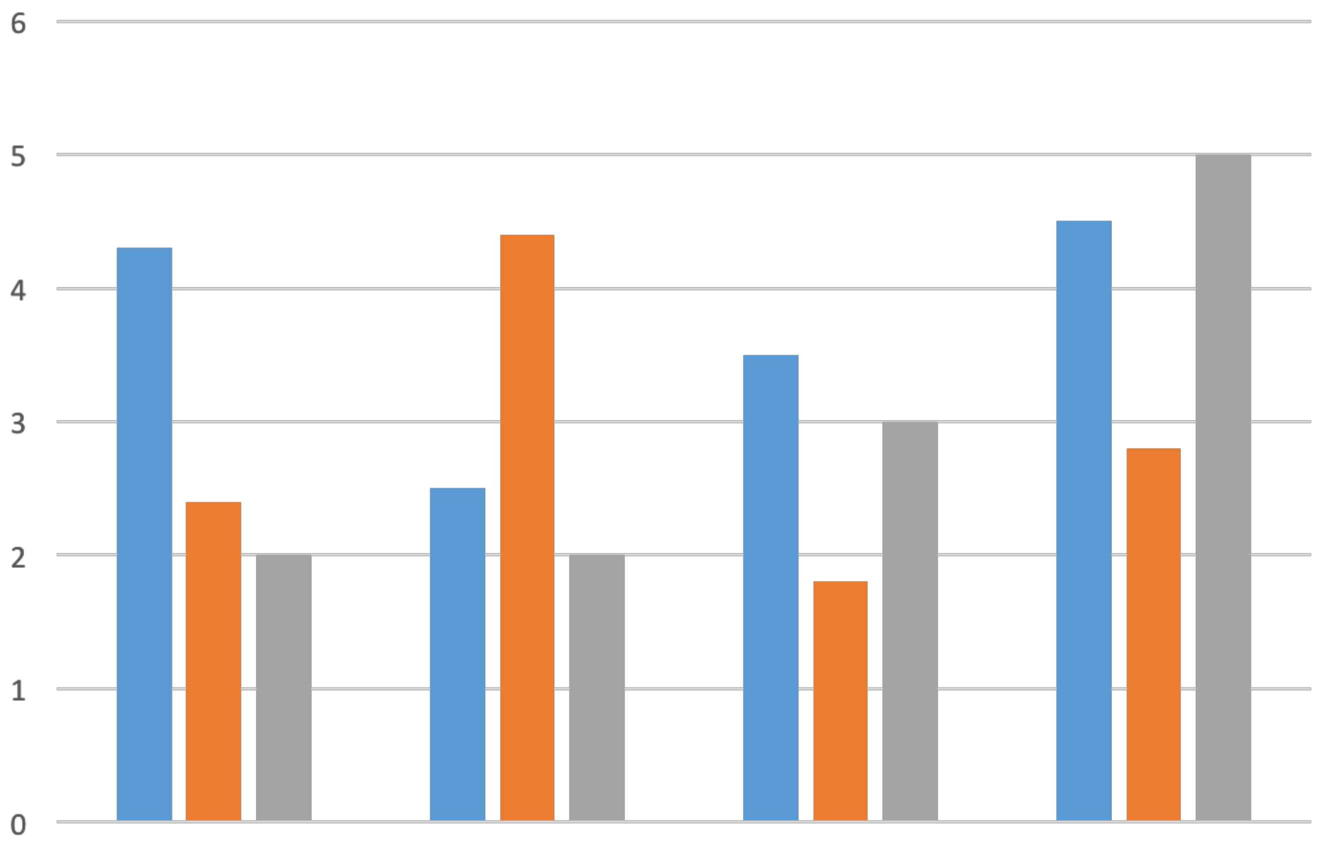


When to Not Use Bar Charts



Graph Aspect Ratios

- Perception of trends and patterns is heavily influenced by the aspect ratio.
- Aspect ratios affects:
  - Densities
  - relative distances orientations

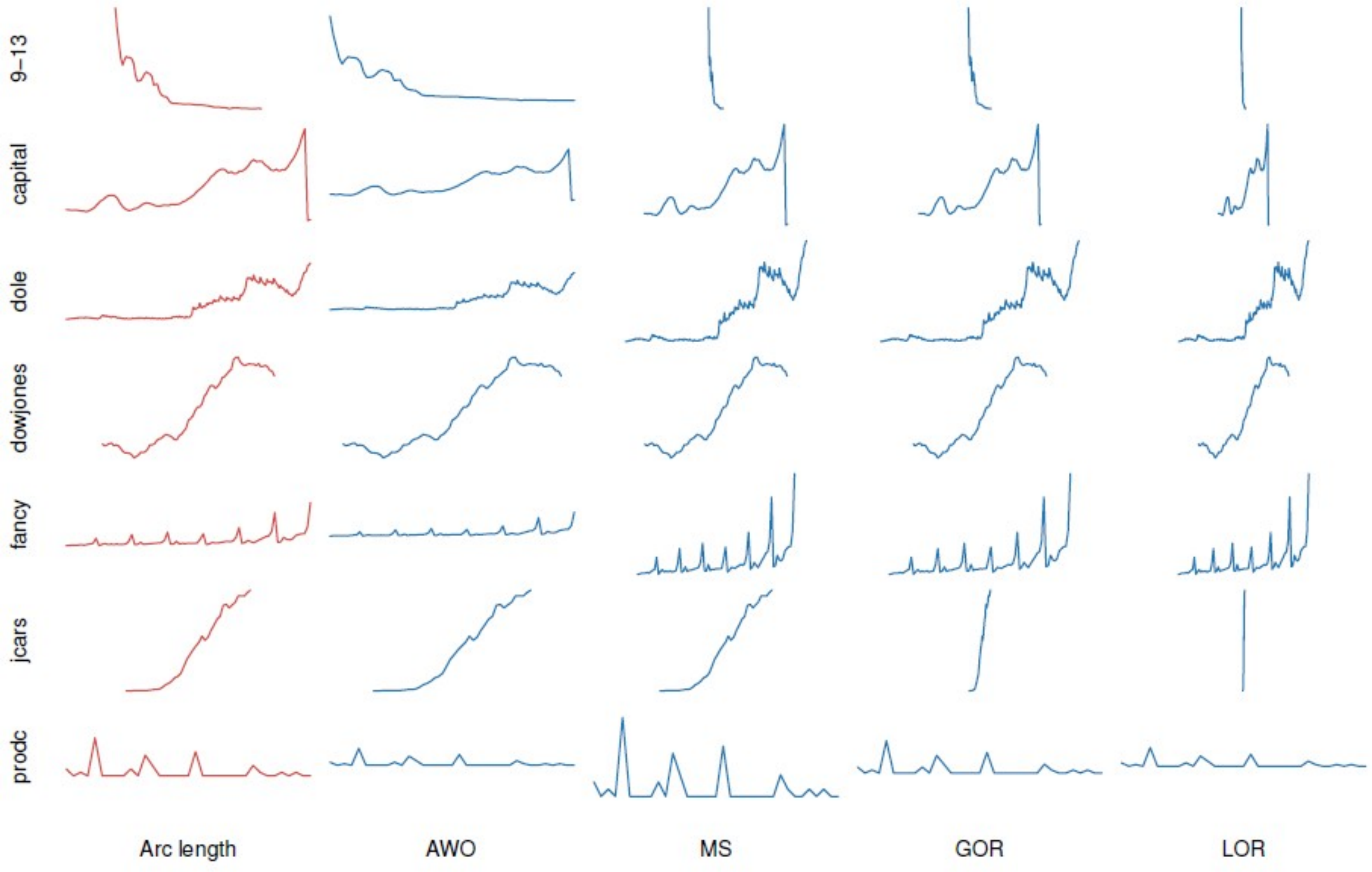


Aspect ratio: a = width/height

Arc Length-Based Aspect Ratio

$$max_a \frac{\sum_i \sum_j \left|\sin \theta_{ij}\right| l_i(a) l_j(a)}{\sum_i \sum_j l_i(a) l_j(a)}$$

Arc Length-Based Aspect Ratio Selection



Aspect Ratios

Ross Maciejewski, Ph.D

Associate Professor

Arizona State University

Ira A. Fulton Schools of Engineering
Arizona State University

Statistical Graphics:

Non Data Components of Graphs

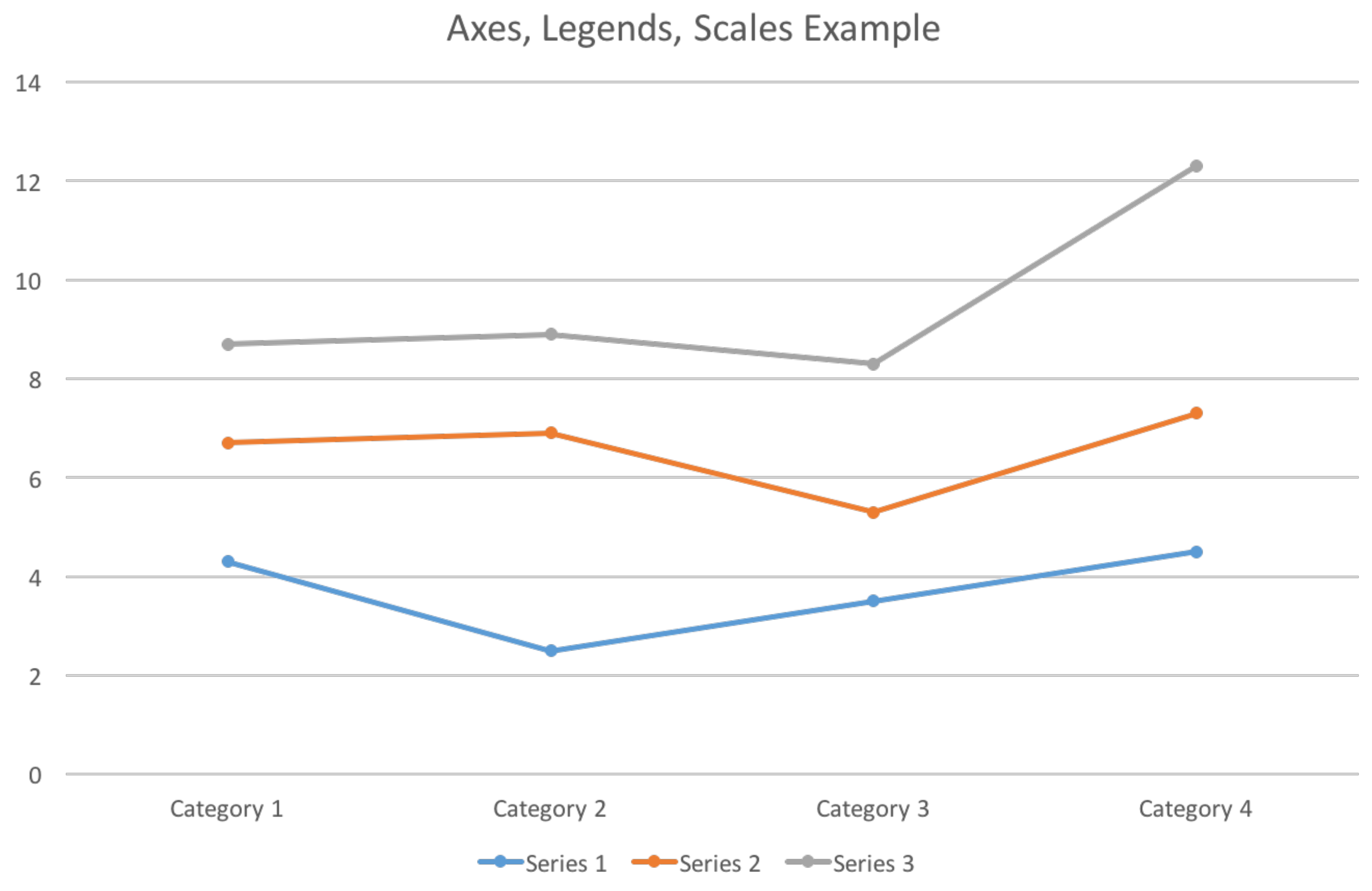
Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Define design principles for bar charts and line charts

Axes, Legends and Scales



Heckbert's Labeling Algorithm

Data range

105 - 543

Data range

2.03-2.17

Nice Numbers

```
const ntick ← 5;                              desired number of tick marks

loose_label: label the data range from min to max loosely.
  (tight method is similar)

procedure loose_label(min, max: real);
nfrac: int;
d: real;                                      tick mark spacing
graphmin, graphmax: real;                     graph range min and max
range, x: real;
begin
  range ← nicenum(max − min, false);
  d ← nicenum(range/(ntick − 1), true);
  graphmin ← floor(min/d)*d;
  graphmax ← ceiling(max/d)*d;
  nfrac ← max( − floor(log10(d)), 0);         number of fractional digits to show
  for x ← graphmin to graphmax + .5*d step d do
    put tick mark at x, with a numerical label showing nfrac fraction digits
    endloop;
  endproc loose_label;

nicenum: find a "nice" number approximately equal to x.
Round the number if round = true, take ceiling if round = false.

function nicenum(x: real; round: boolean): real;
exp: int;                                     exponent of x
f: real;                                      fractional part of x
nf: real;                                     nice, rounded fraction
begin
  exp ← floor(log10(x));
  f ← x/expt(10., exp);                       between 1 and 10
```

```
if round then
  if f < 1.5 then nf ← 1.;
  else if f < 3. then nf ← 2.;
  else if f < 7. then nf ← 5.;
  else nf ← 10.;
else
  if f ≤ 1. then nf ← 1.;
  else if f ≤ 2. then nf ← 2.;
  else if f ≤ 5. then nf ← 5.;
  else nf ← 10.;
return nf*expt(10., exp);
endfunc nicenum;
```

P. Heckbert. Nice numbers for graph labels. In A. Glassner, editor, Graphics Gems, pages 61–63 657–659. Academic Press, Boston, 1990.

Heckbert’s Labeling Algorithm

Problem

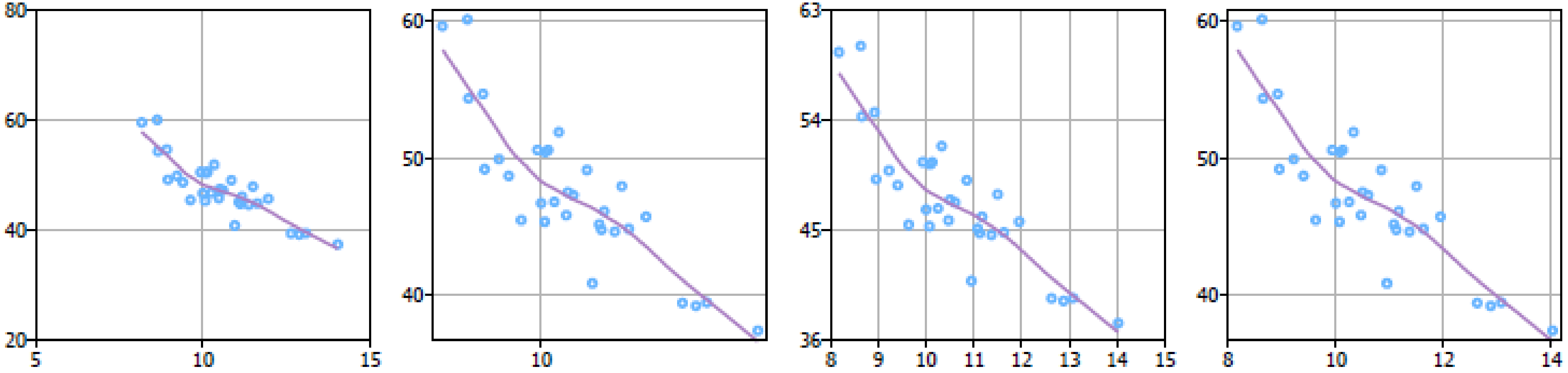
For small numbers, the range of labels can be much larger than the data range.

Solution

Drop labels which overlap or fall outside the data range

This leads to unevenly spaced labels or axes with only one label

Extension of Wilkinson’s Algorithm



(a) Heckbert (b) R’s pretty (c) Wilkinson (d) Extended

Extension of Wilkinson's Algorithm

$$\text{Coverage} = 1 - \frac{1}{2}\frac{(d_{max}-l_{max})^2+(d_{min}-l_{min})^2}{[.1(d_{max}-d_{min})]^2}$$

$$\text{Legibility} = \frac{format+font_{size}+orientation+overlap}{4}$$

Statistical Graphics:

Creating Histograms

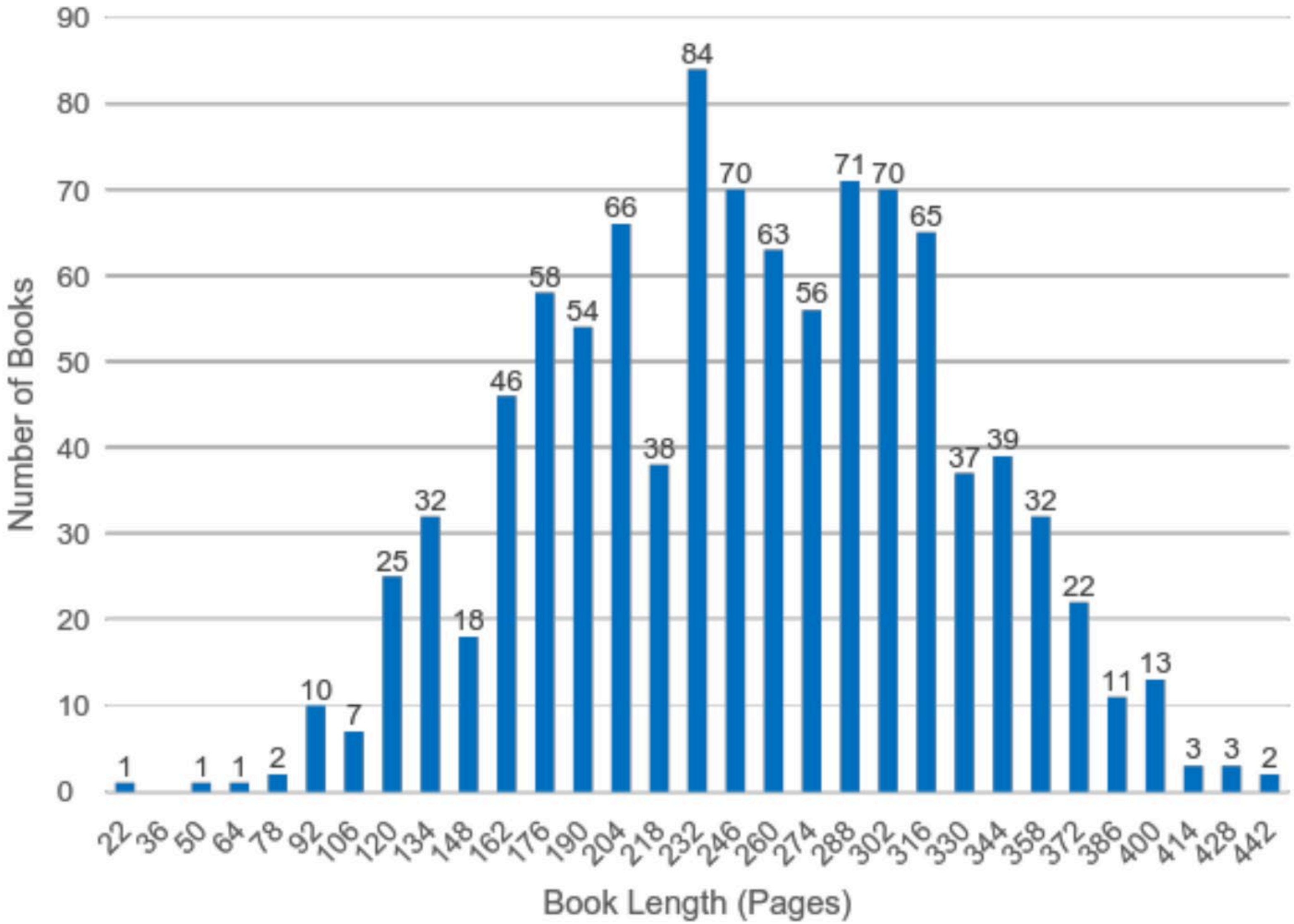
Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Define design principles for Histograms and the impact of parameter choices on the visualization.

Histograms



Number of Books
90
80
70
60
50
40
30
20
10
0
1
1
1
2
10
7
25
32
18
46
58
54
66
38
84
70
63
56
71
70
65
37
39
32
22
11
13
3
3
2
22 36 50 64 78 92 106 120 134 148 162 176 190 204 218 232 246 260 274 288 302 316 330 344 358 372 386 400 414 428 442
Book Length (Pages)

Histogram Binning

Number of bins (k) can be user-specified or chosen from a suggested bin width (h) such that:

$$k = \left\lceil \frac{\max x - \min x}{h} \right\rceil$$

Histogram Binning

Common choices for k include the square-root choice where:

$$k = \sqrt{N}$$

Histogram Binning

Sturge's formula

$$k = \lceil \log N + 1 \rceil$$

Scott's choice

$$h = \frac{3.5\sigma}{N^{\frac{1}{3}}}$$

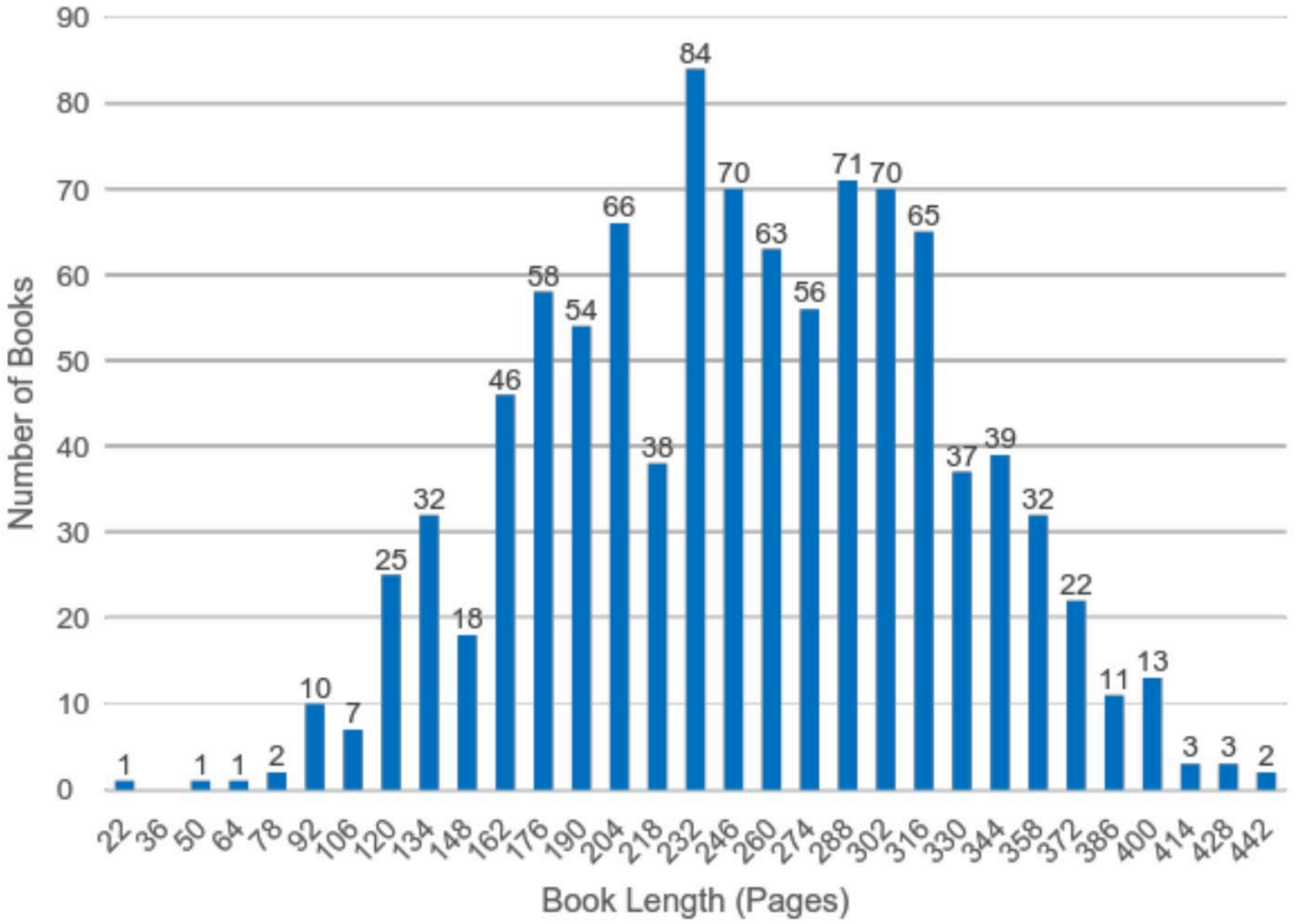
Freedman-Diaconis rule

$$h = 2IQR(x)N^{-\frac{1}{3}}$$

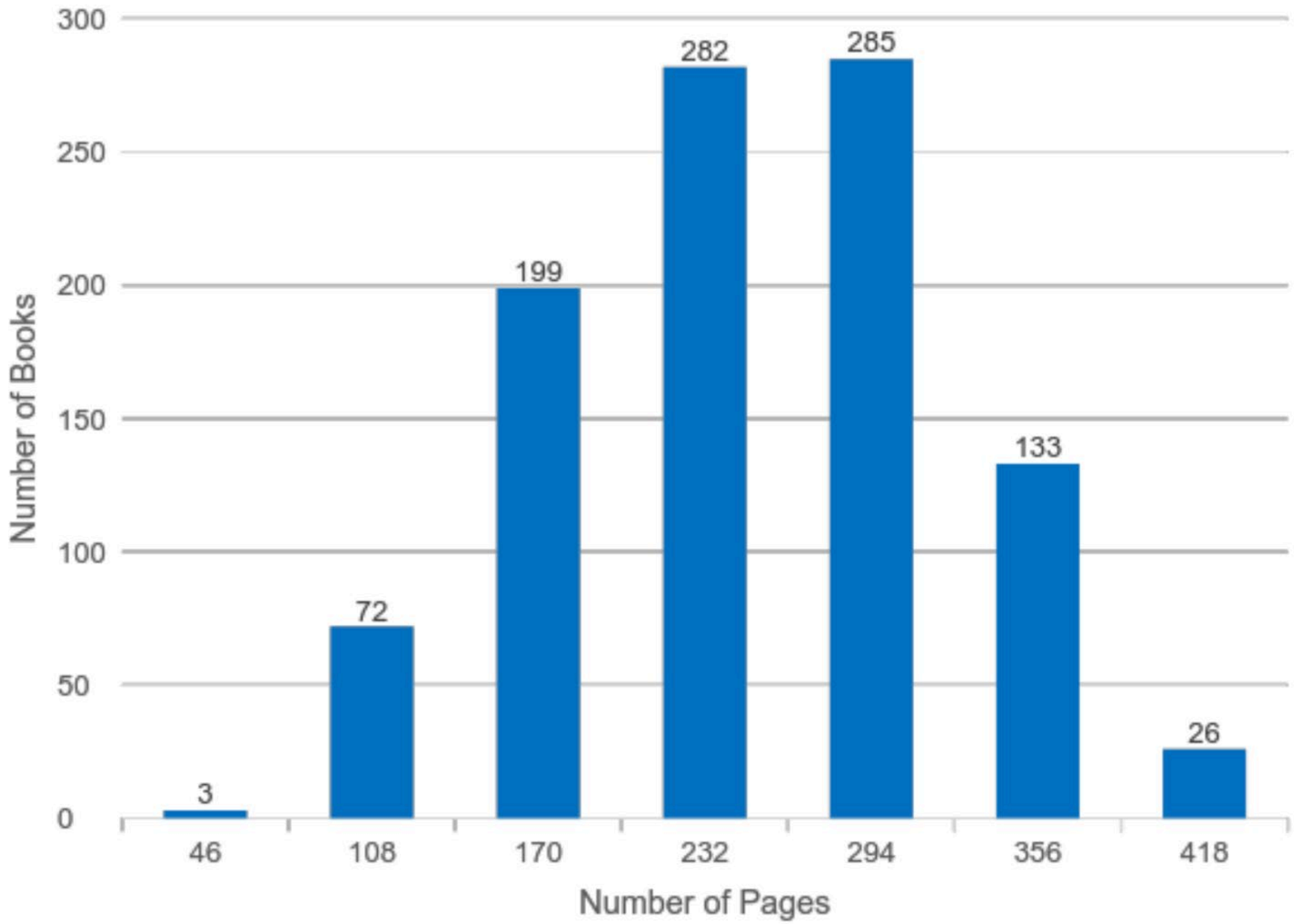
Histogram Example

- Plot a histogram of 1000 book lengths.
- Use all four common choices for k or h.
- All x-axis labels indicate the center of the histogram bin.

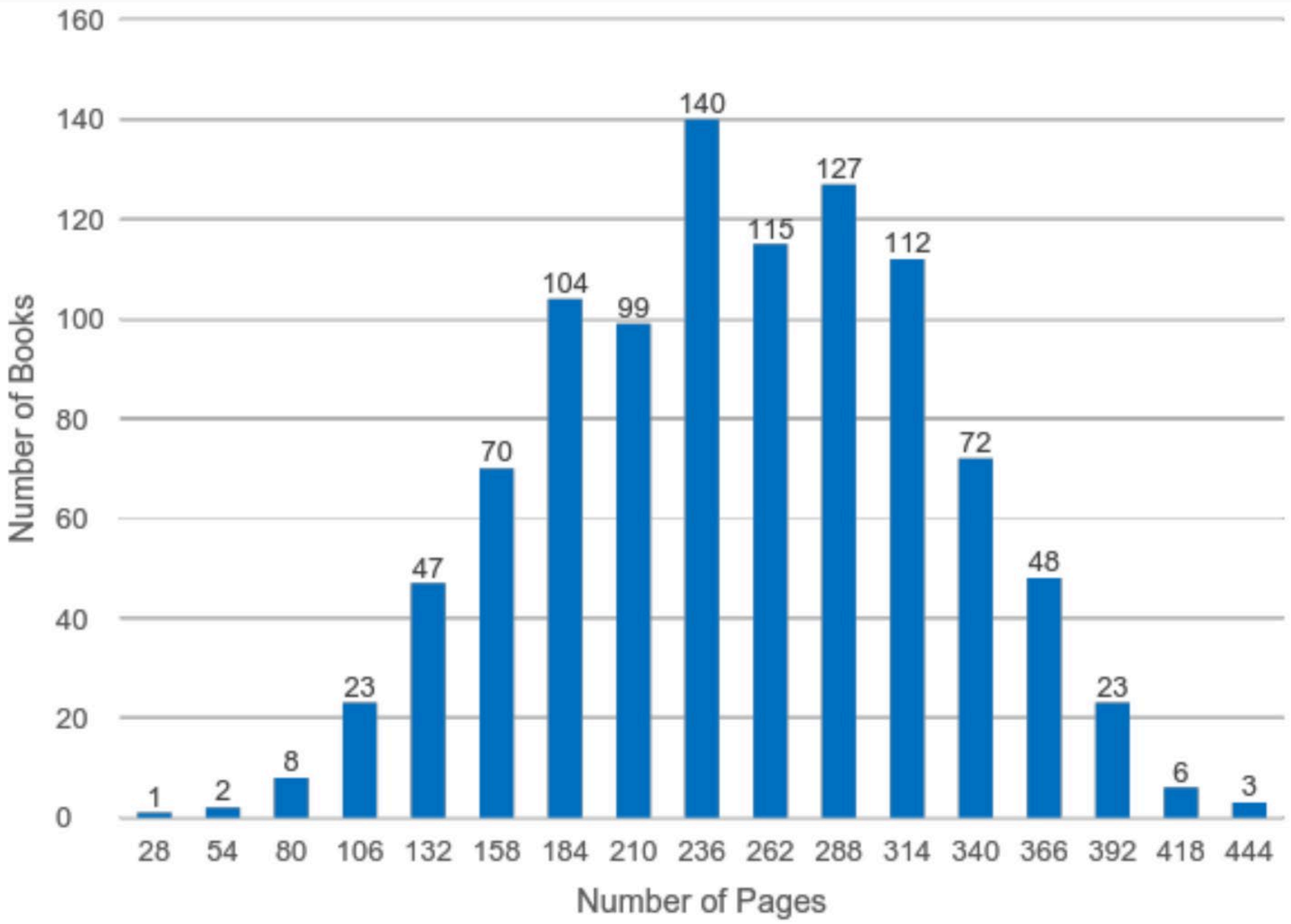
Square-Root Choice



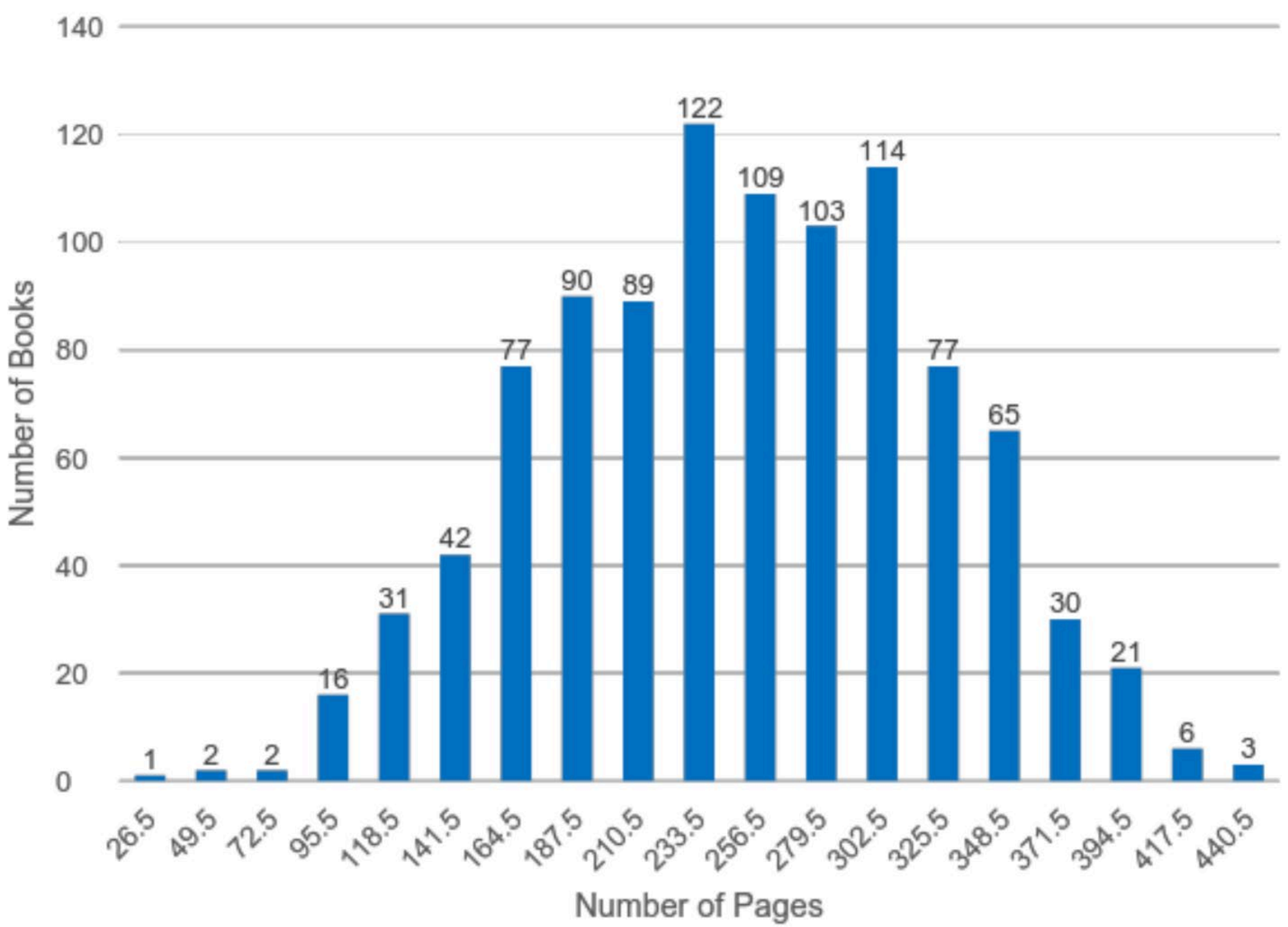
Sturge’s Formula



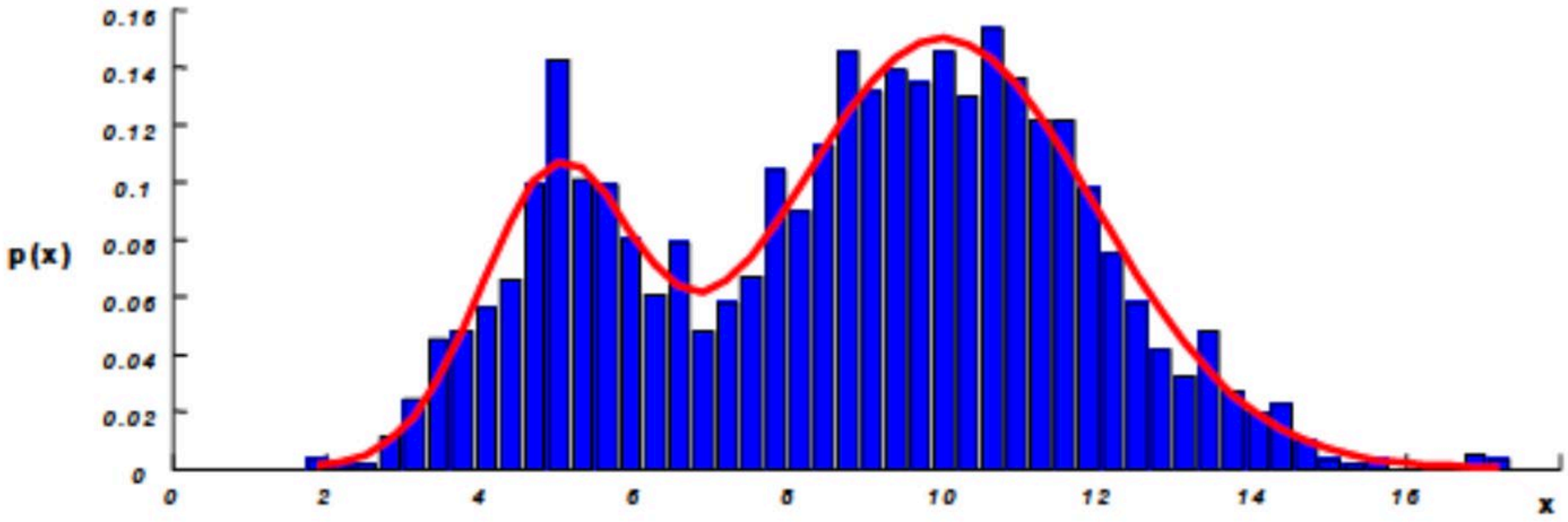
Scott's Choice



Freedman-Diaconis Rule



Histograms



Statistical Graphics:
Understanding Quantiles

Ira A. Fulton Schools of Engineering
Arizona State University

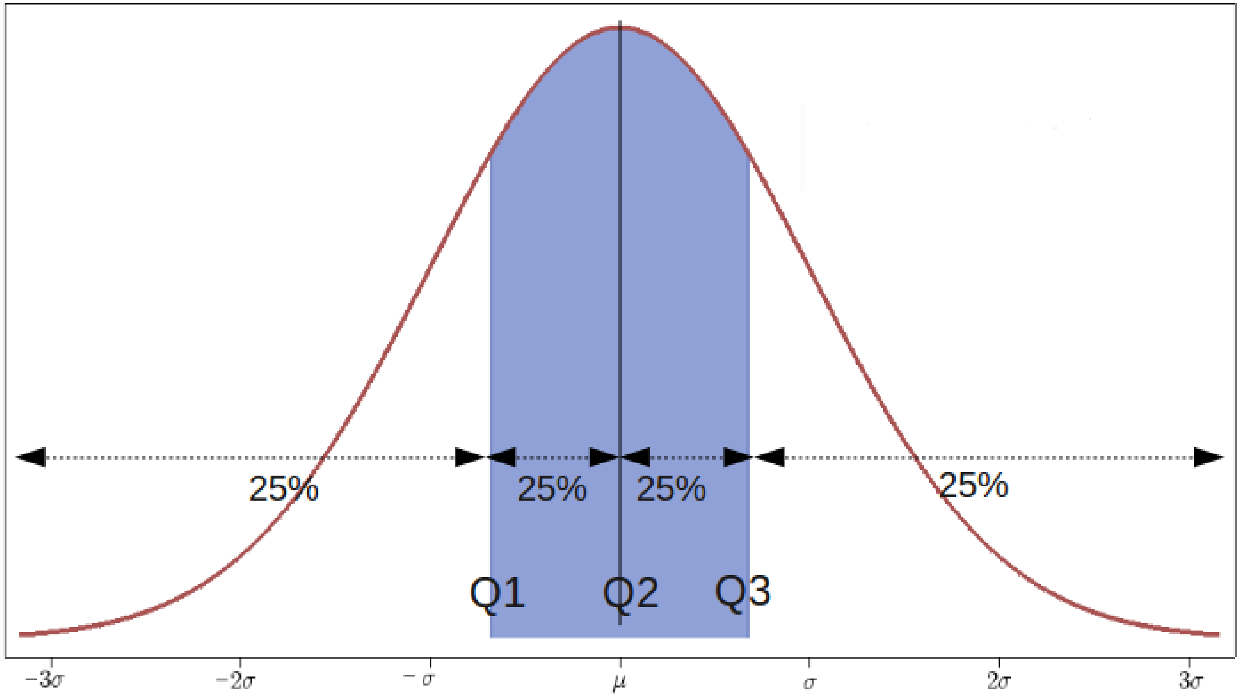
Objective

Objective

Understand how to create and use Box-plots and Q-Q Plots

Quantiles

- **Quantiles** – points taken at **regular intervals** from the **cumulative distribution** function of a random variable



Calculating Quantiles

Distribution of at-bats

587	547	471	470	596	587	599	525	619	463	543	554	591	554	580	517	579	569	514	589

Calculate Quantiles

463	470	471	514	517	525	543	547	554	554	569	579	580	587	587	589	591	596	599	619

Quantiles

Useful measures because they are less susceptible to long-tailed distributions and outliers.

May be more descriptive statistics than means and other moment-related statistics.

Quantiles of a random value are preserved under increasing transformations.

Can be used where only ordinal data are available.

Statistical Graphics:
Box and Whisker Plots

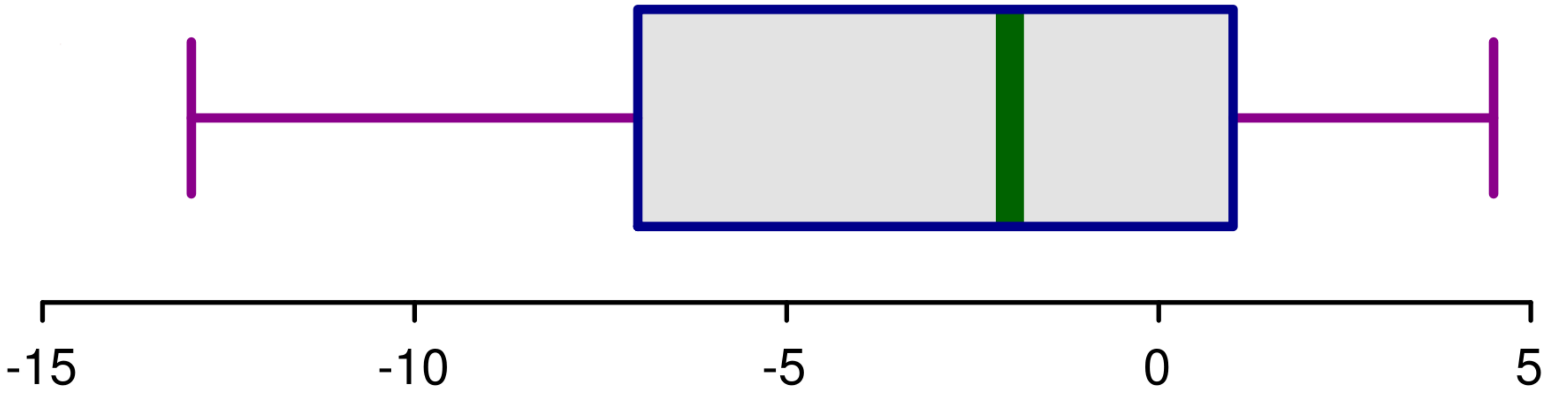
Ira A. Fulton Schools of Engineering
Arizona State University

Objective

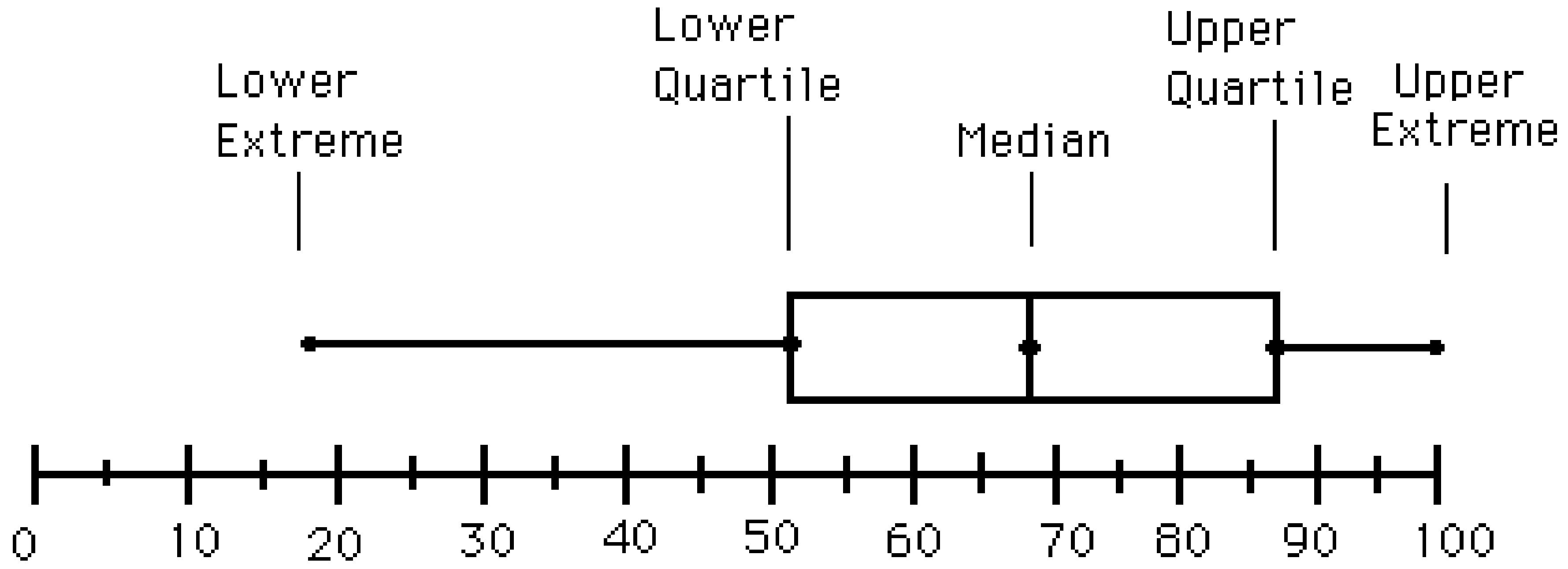
Objective

Understand how to create and use Box-plots and Q-Q Plots

Box and Whisker Plot



Summaries


Lower Extreme
Lower Quartile
Median
Upper Quartile
Upper Extreme
0
10
20
30
40
50
60
70
80
90
100

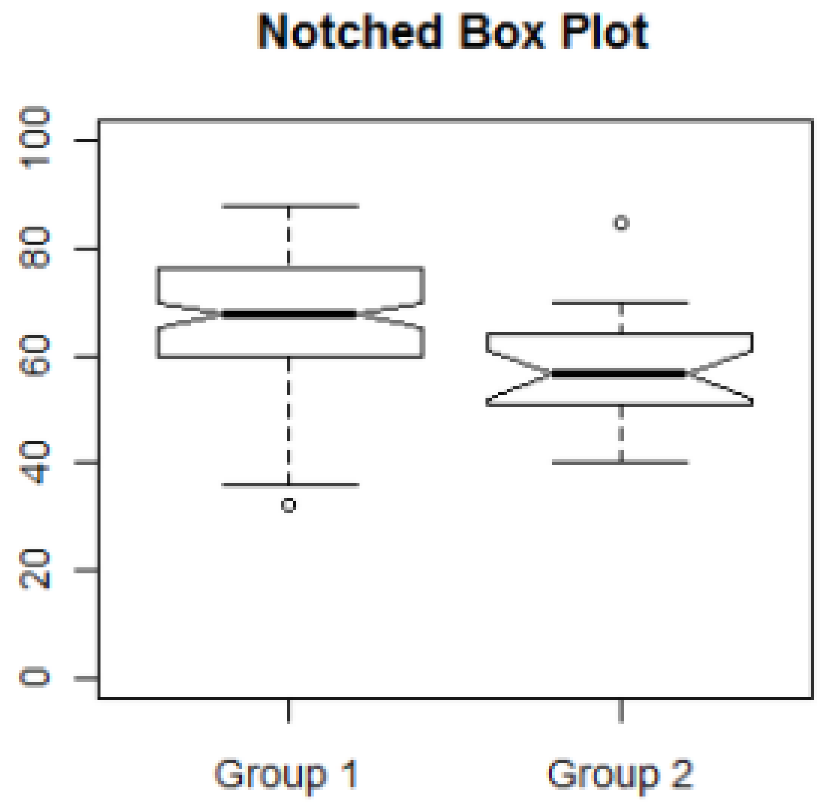
Alternate forms of Box and Whisker Plots

Width of the box

- mapped to the size of the group
- can make width proportional to the square root of the group size

Notched box plot

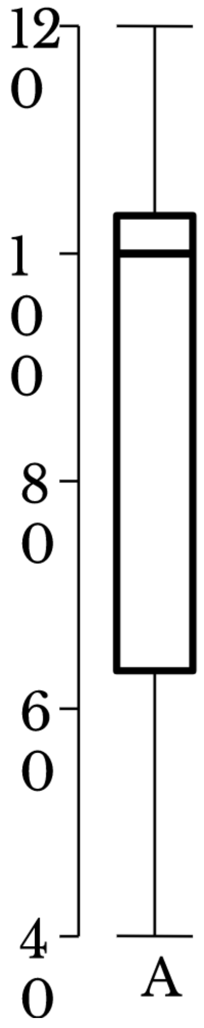
- Width of notches is proportional to IQR



Box and Whisker Plots: Baseball Example

RBI	Batting Ave
117	0.336
113	0.324
47	0.321
95	0.315
103	0.312
118	0.312
66	0.307
85	0.307
103	0.304
41	0.3

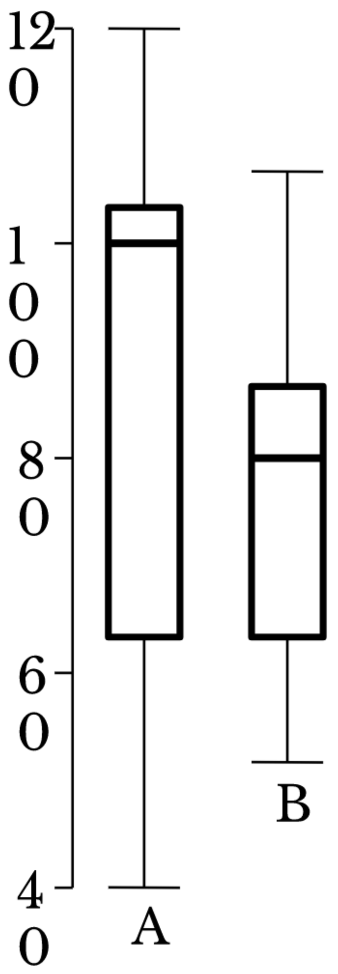
A: 1-10



Box and Whisker Plots: Baseball Example

RBI	Batting Ave
76	0.3
52	0.298
101	0.298
85	0.296
66	0.293
82	0.292
69	0.29
86	0.29
59	0.288
105	0.287

B: 11-20



Statistical Graphics:

Q-Q Plots

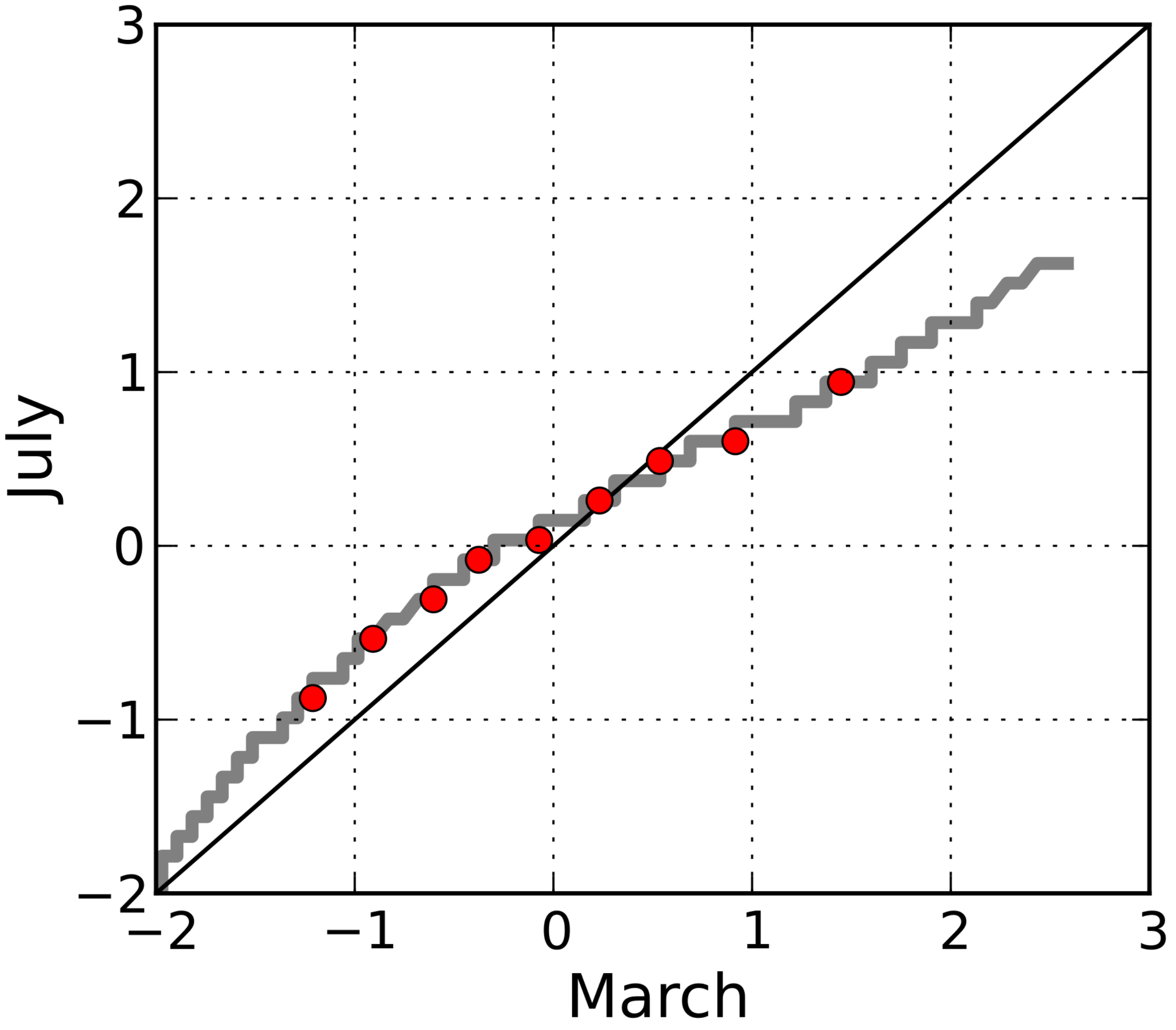
Ira A. Fulton Schools of Engineering
Arizona State University

Objective

Objective

Understand how to create and use Box-plots and Q-Q Plots

Q-Q Plots



QQ Plot

More powerful comparing two distributions than histograms

Sample sizes do not need to be equal

Alternative is a probability plot

Creating a Q-Q Plot

RBI	Batting Ave
117	0.336
113	0.324
47	0.321
95	0.315
103	0.312
118	0.312
66	0.307
85	0.307
103	0.304
41	0.3

A: 1-10