

Lecture 21: Displaying Data

Dewayne E Perry

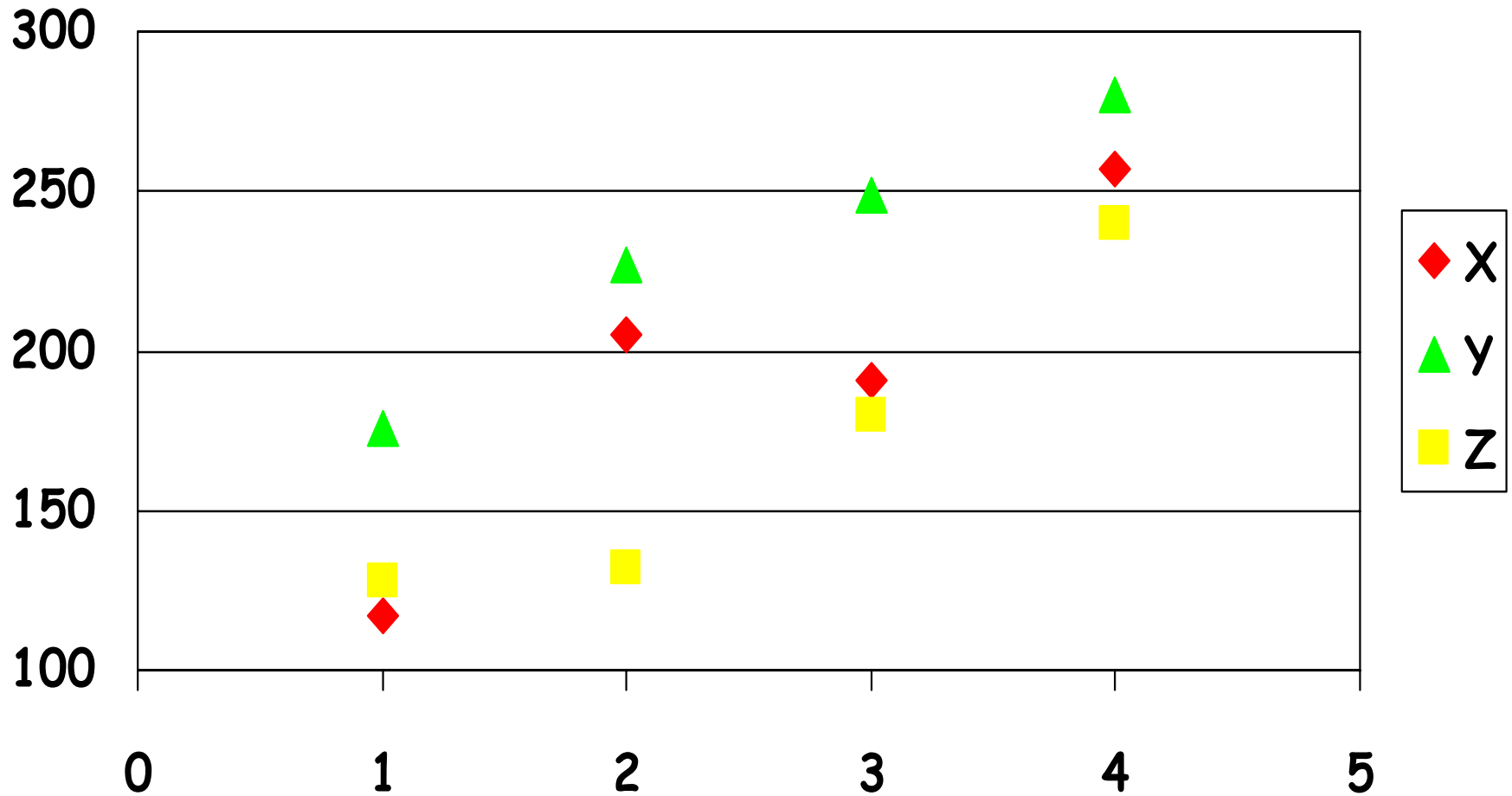
ENS 623

perry@ece.utexas.edu

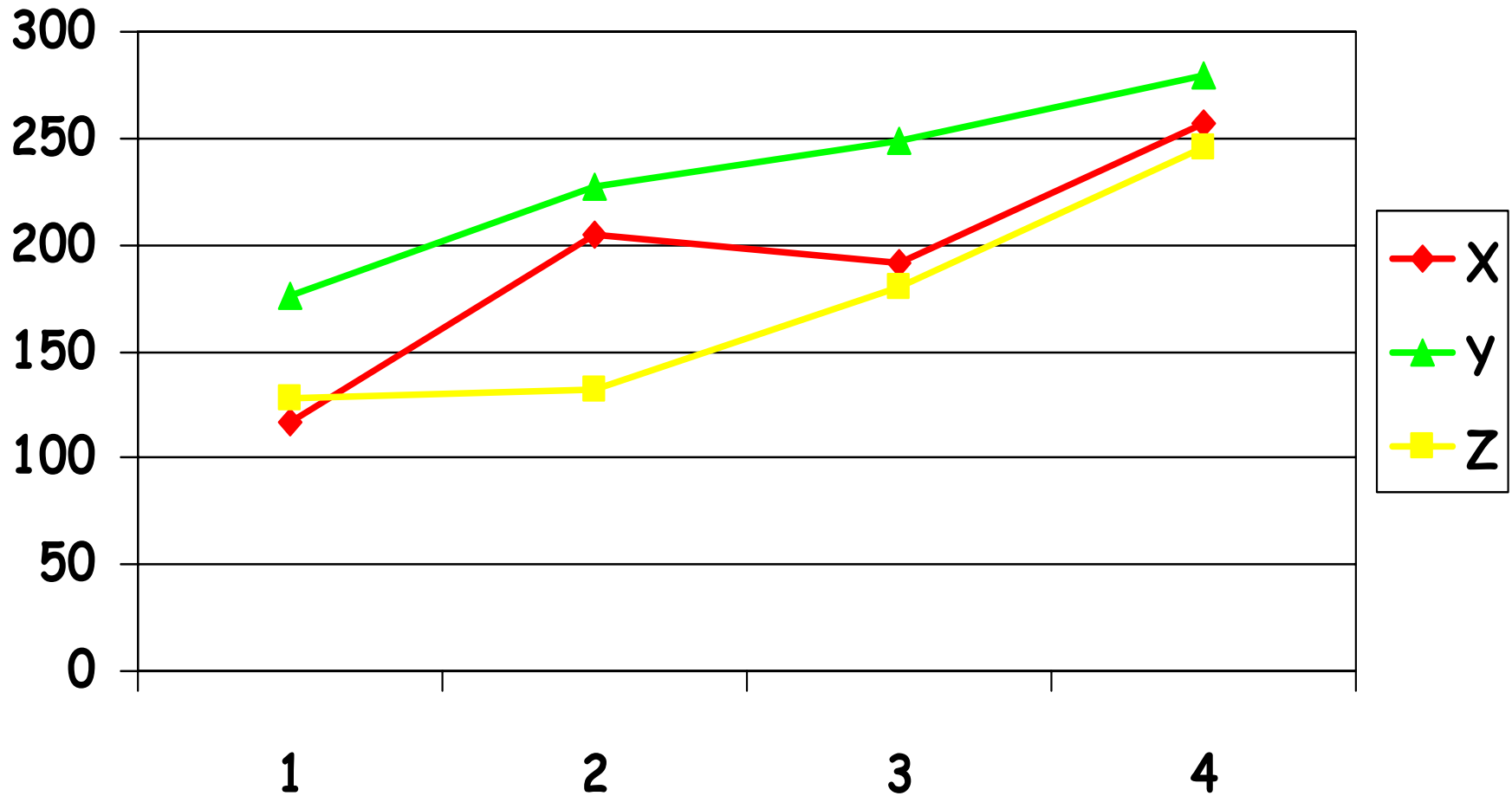
Fundamental Work

- **Description of a group of sampling units**
 - ↪ Sampling units: the things being studied
 - ↪ For each variable
 - Some type of number assigned to each unit
 - ↪ Describing data = summarizing the numbers
- **Central tendencies: mean, median and mode**
- **Distributions**
 - ↪ X axis is the independent variable
 - ↪ Y axis is the dependent variable
- **Various forms:**
 - ↪ **Plots**
 - Scatter, Connected points, Curves
 - ↪ **Bar charts**
 - Standard, Pictorial
 - ↪ **Stem and leaf displays**
 - ↪ **Box plots**
 - ↪ **Creative** (Tufte's Napoleonic Russian Campaign)

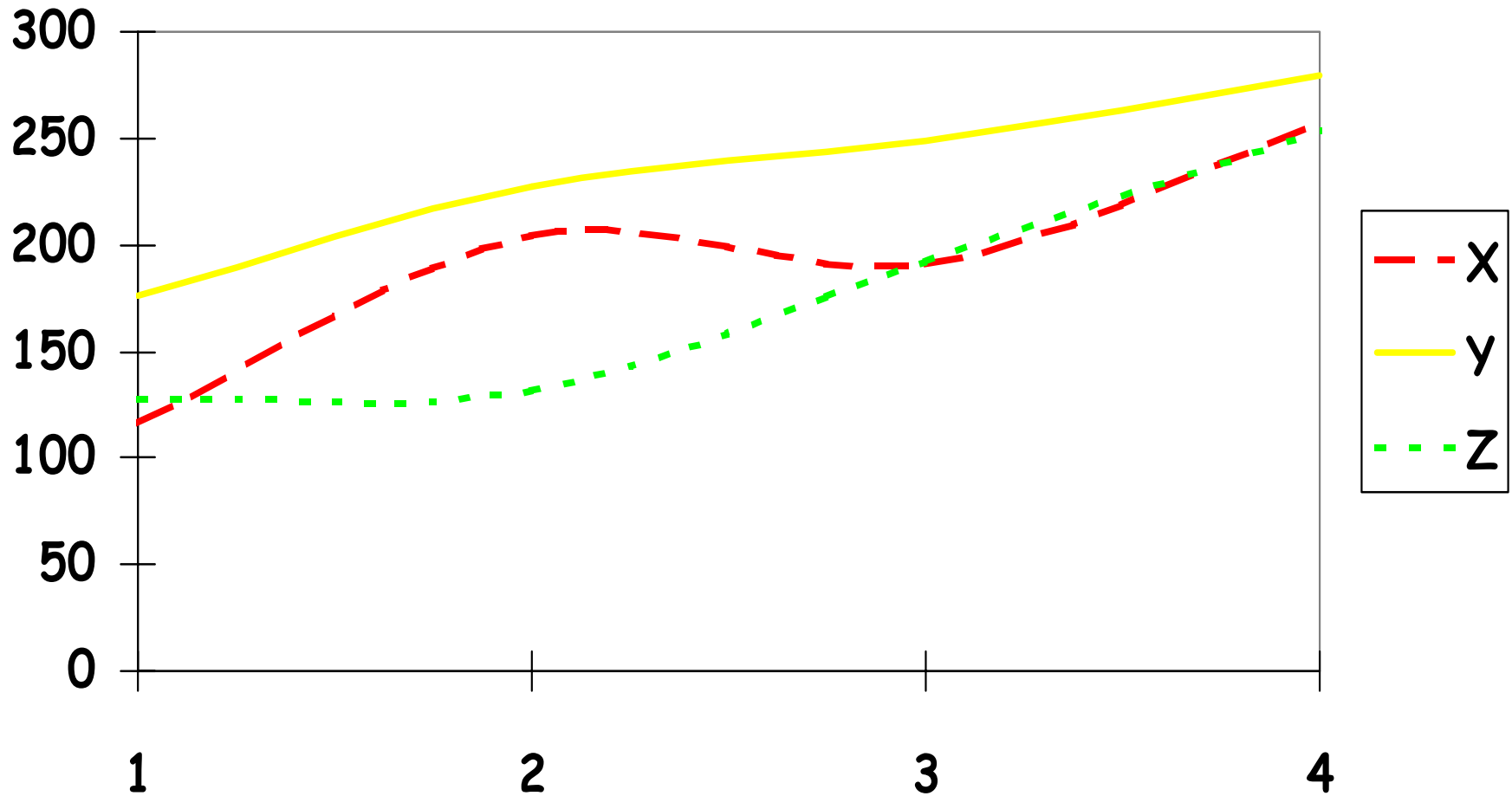
Plot: Scatter



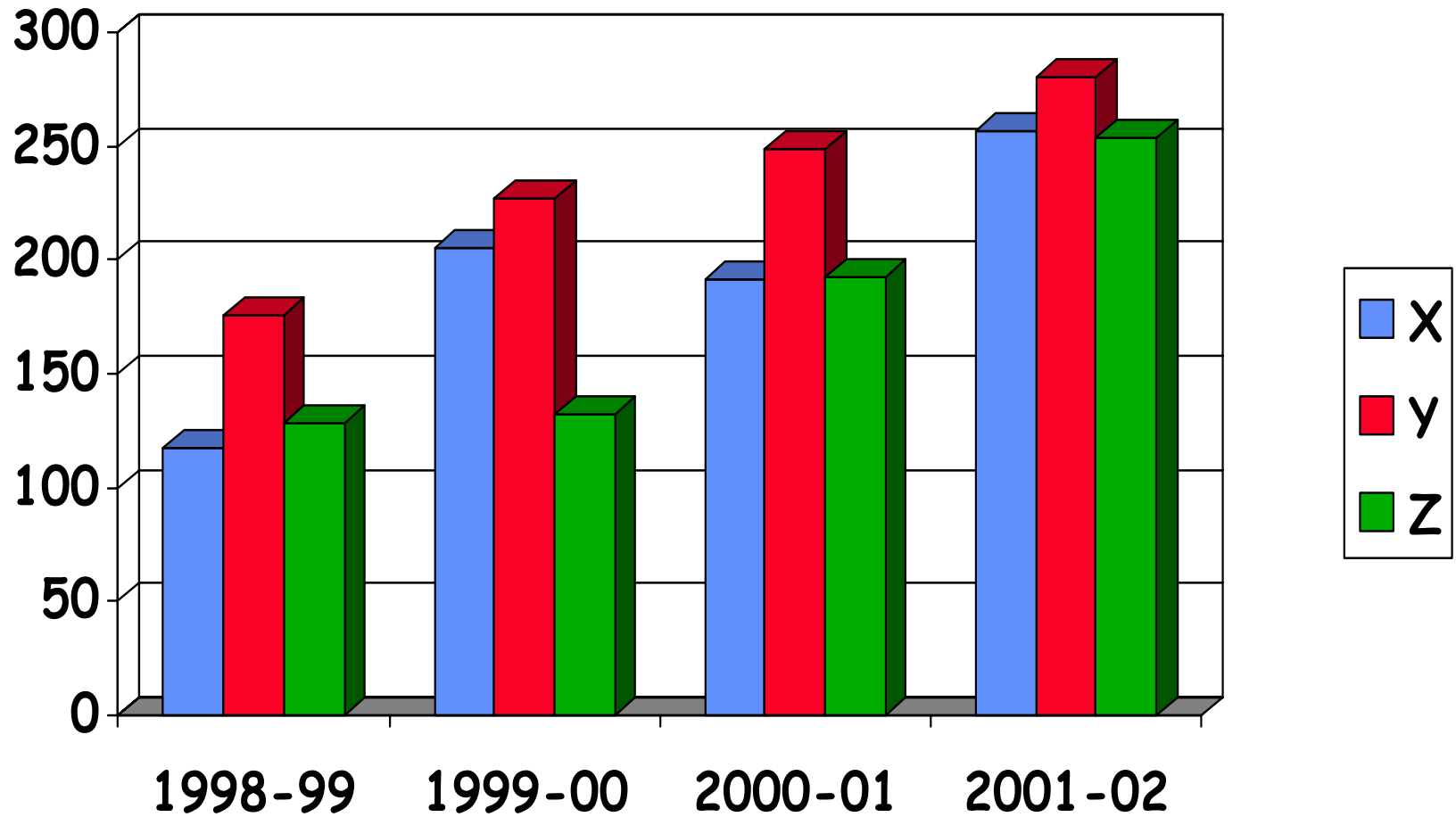
Plot: Connected Points



Plot: Curves



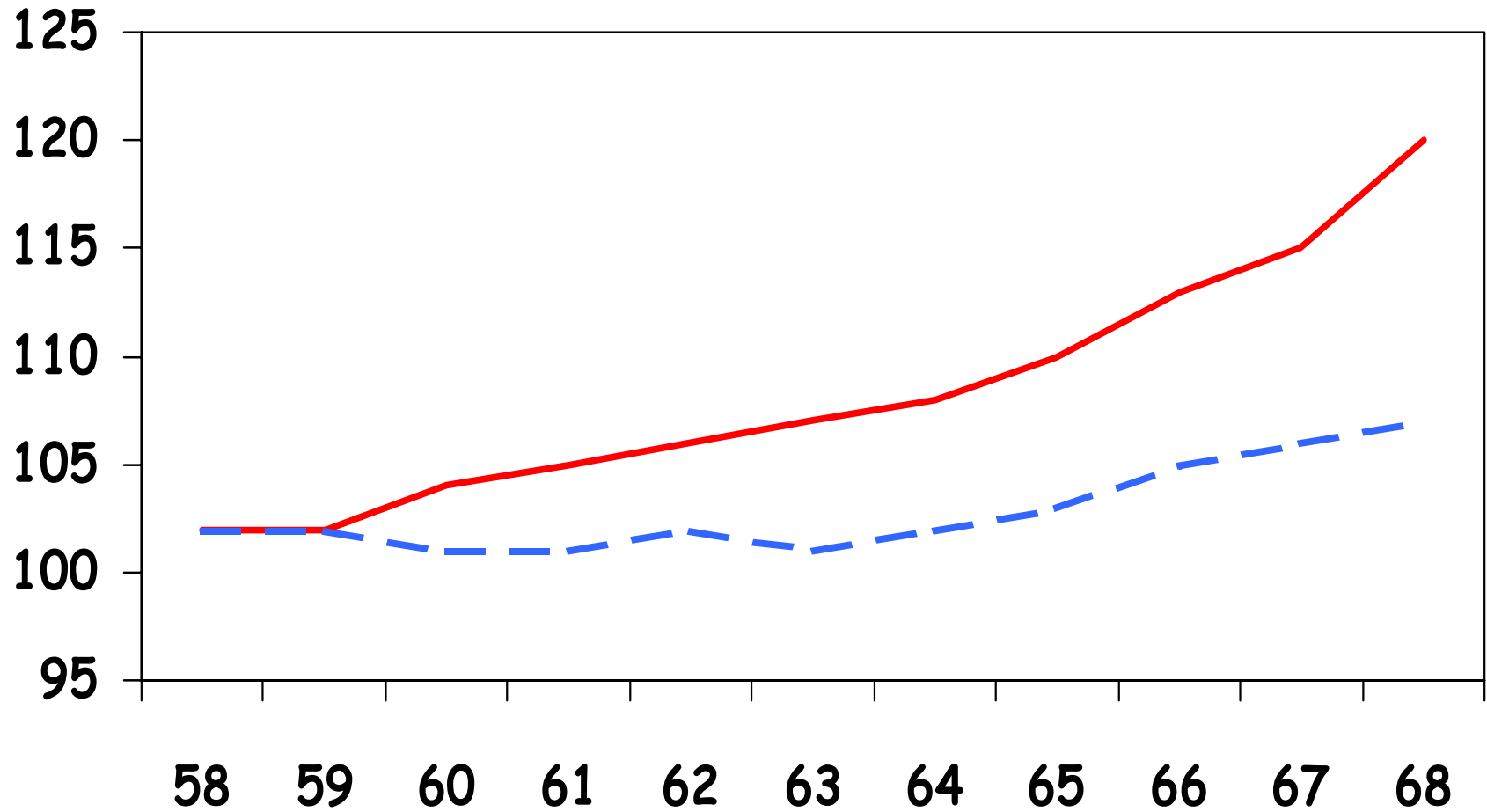
Sample Bar Chart



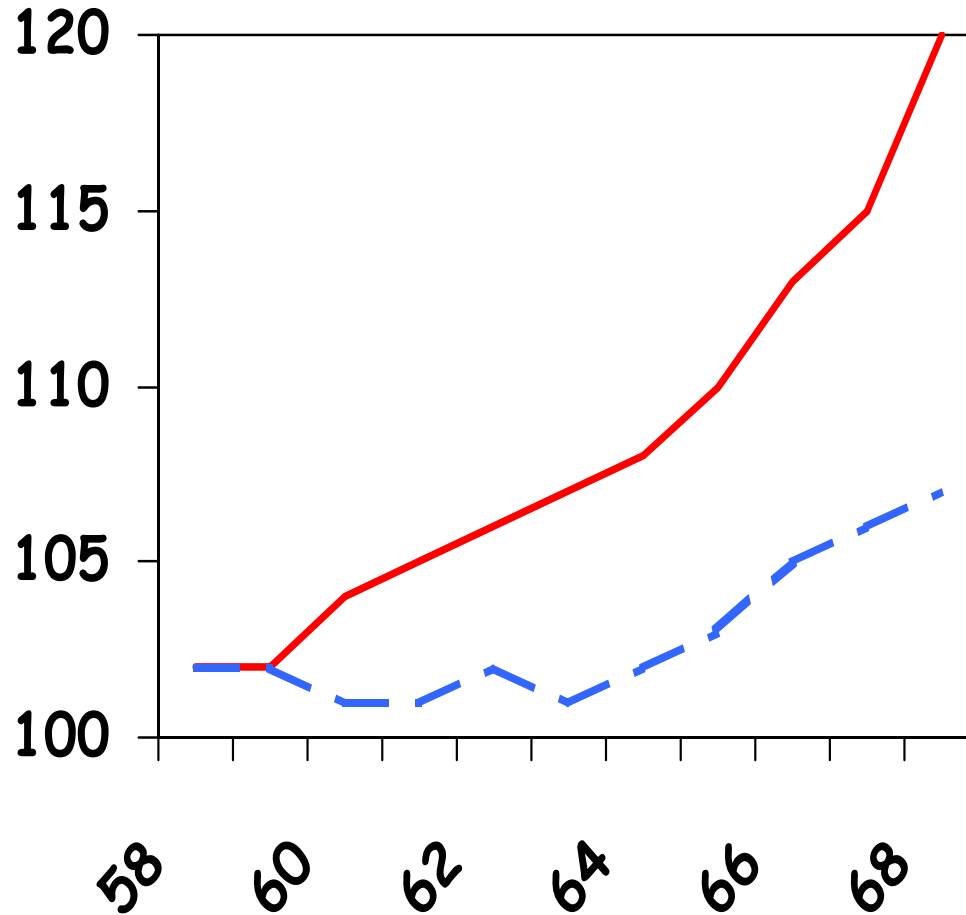
Misuse of Graphs

- Over emphasize vertical scale
- Correct vertical scale, over emphasized volume or horizontal scale
- Misleading pictures

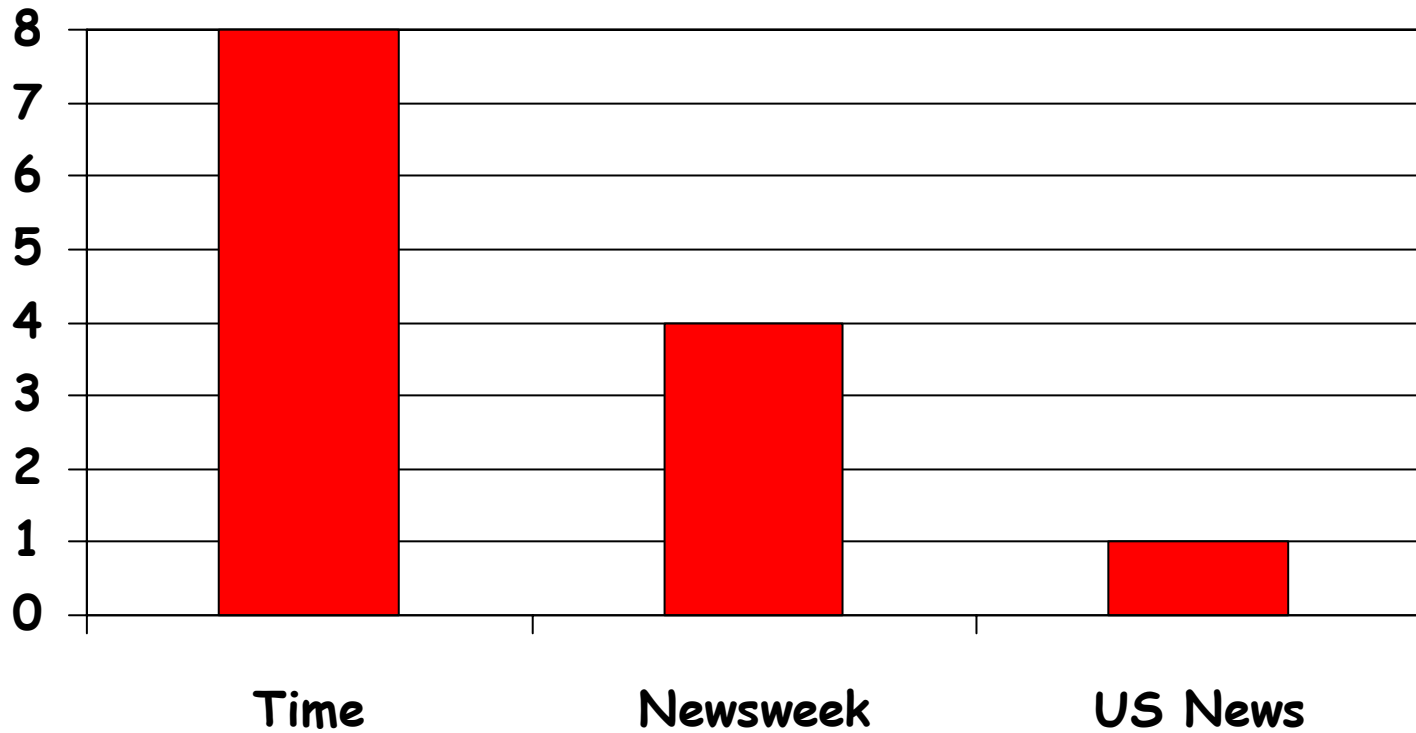
More Neutral



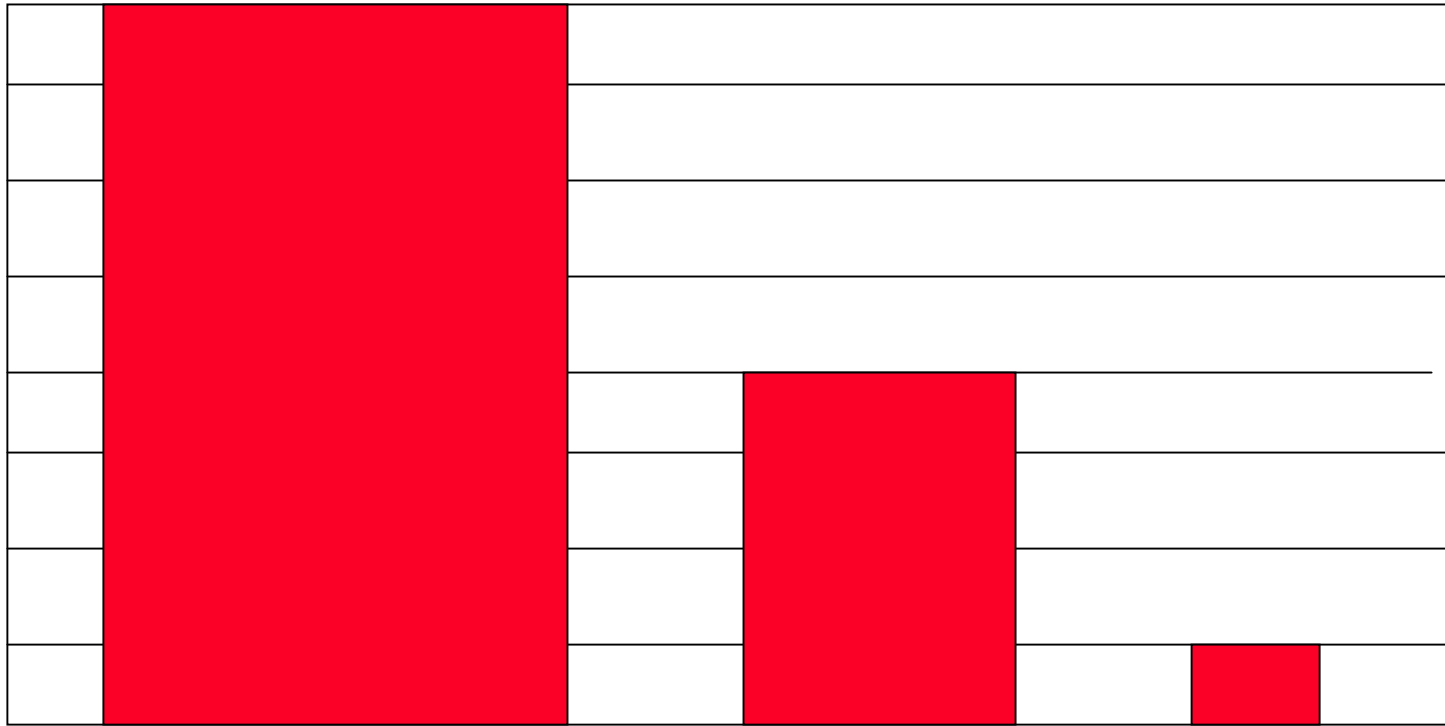
Vertical Emphasis



Normal Comparison



Volume Overemphasis



Stem and Leaf

→ Wainer and Thissen 81:

↳ "the most important device for the analysis of small batches of numbers to appear since the t -test" (pg 199)

→ Purposes:

↳ Storing data for current and later use

↳ Plotting data as a distribution from high to low scores

↳ Making it easier to discern patterns

→ Standard approach

↳ Leading digit

↳ Second digit (possibly arranged in order)

Stem and leaf

9	1100
8	8513536104
7	39974539
6	759890
5	6655
4	5

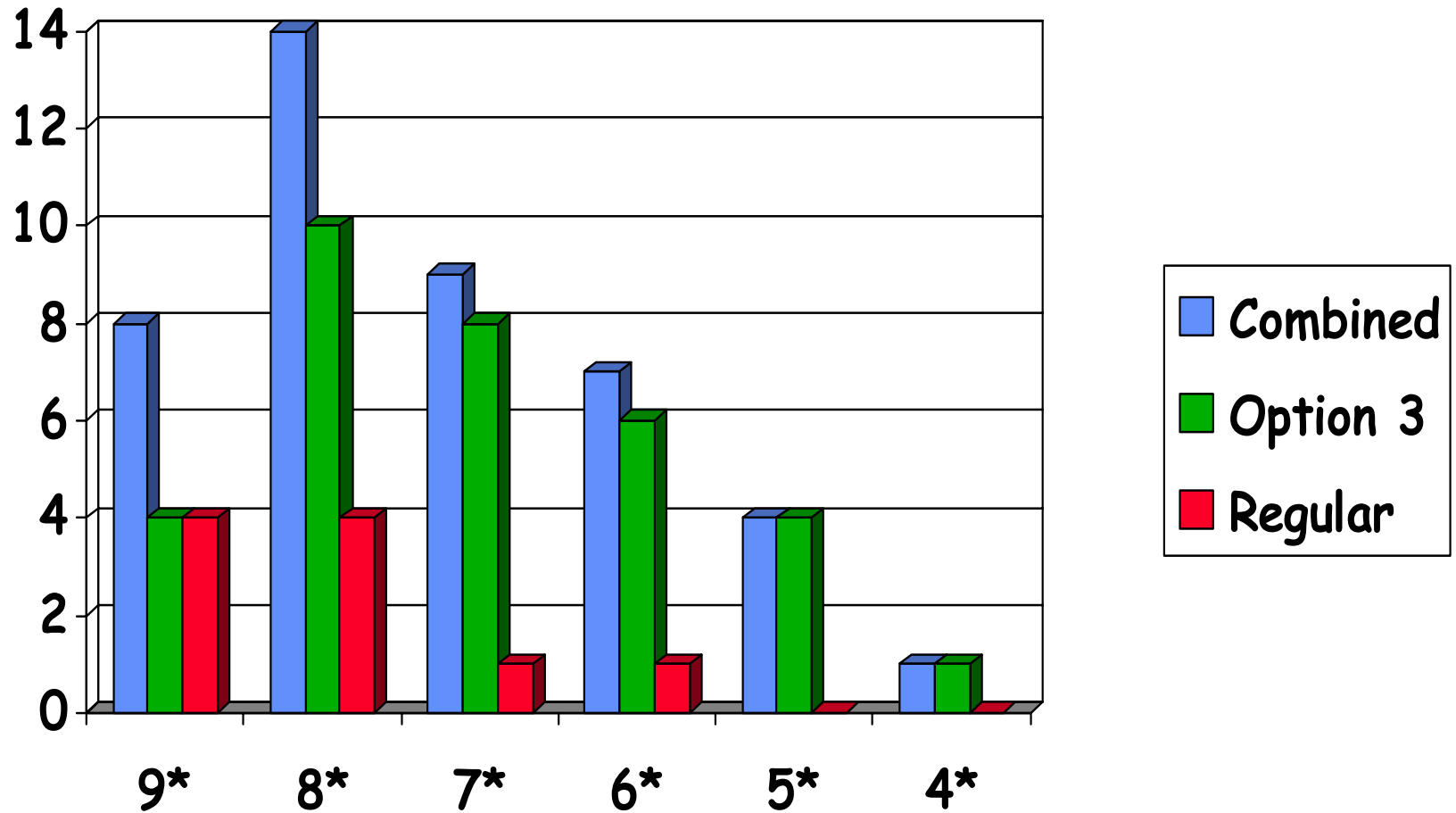
Stem and Leaf

9	00111122
8	01123334556788
7	334578999
6	0678899
5	5566
4	5

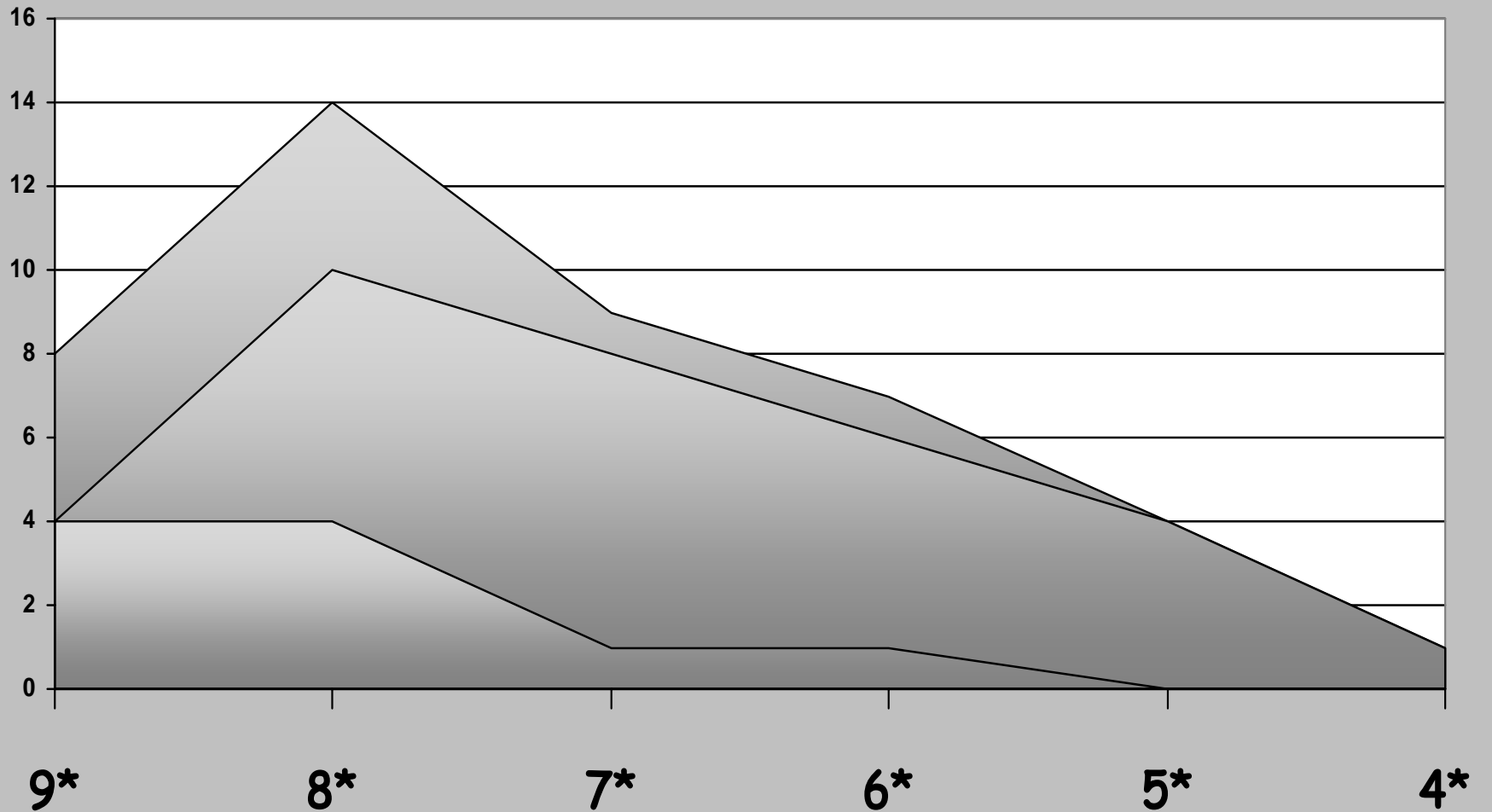
Distribution Comparison

2211	9	0011
8732	8	0113345568
8	7	33457999
8	6	067899
5		5566
4		5

Score Comparison: Bar Chart



Score Comparison: Shaded Plot



Summary Data

→ Mode

↩	R	91, 92
↩	O3	79
↩	C	91,79

→ Median

↩	R	87.5
↩	O3	79
↩	C	81

→ Mean

↩	R	85.2
↩	O3	74.94
↩	C	77.33

Summary Data

→ Maximum	92
→ 75th percentile	87
→ Median	81
→ 25th percentile	69
→ Minimum	45

Box Plots

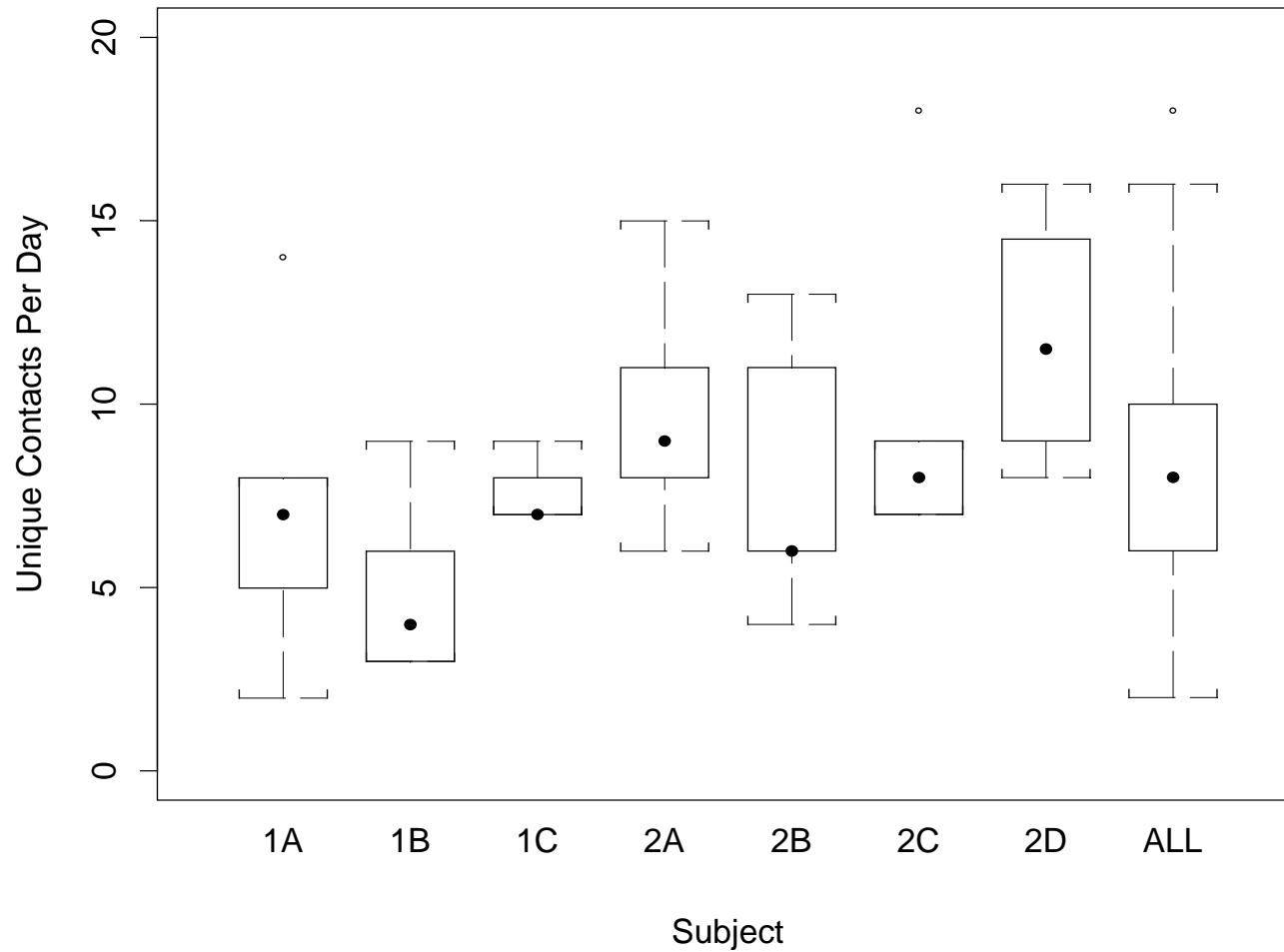
→ Useful when

- ↪ Large amount of data
 - Ie, stem and leaf too large to display easily
- ↪ Several distributions to compare
 - Eg, in the time studies

→ Tukey 77

- ↪ Called it a “box and whisker” plot
- ↪ Top and bottom of the box are 75^{th} and 25^{th} percentile respectively
- ↪ Line or dot in dividing the box represents *median*
- ↪ Whiskers:
 - 90^{th} and 10^{th} with outliers as dots
 - 100^{th} and 0^{th}

Box Plots



Central Tendency

→ Mode

- ↳ Score of greatest frequency
- ↳ May be *unimodal*, *bimodal* or *multimodal*

→ Median

- ↳ Midmost score in the series
- ↳ If N is even, median is half the distance between the two midmost scores

→ Mean

- ↳ Arithmetic average: $\frac{\sum X}{N}$
- ↳ Trimmed average: drop x% of each end's scores

Measures of Spread

→ Given the central tendency, we want to know how far the scores deviate

↳ Spread, dispersion, or variability

→ **Range**

↳ Crude range: highest score - lowest score

➤ $92 - 45 = 47$

↳ Extended range: extend upper/lower, say .5

➤ $92.5 - 44.5 = 48$

➤ Used when range is small and

➤ measurement not very accurate

↳ Trimmed range: reduce effect of extreme scores

➤ Often use the quartile ranges

➤ Or the 90th and 10th as the trim lines

→ **Deviations**

↳ Average:

➤ Average distance from mean of all scores:

$$\bar{D} = \frac{\sum |X - \bar{X}|}{N} = \frac{\sum |D|}{N}$$

Measures of Spread

↪ For regular students, $\bar{D} = 59.6/10 = 5.96$

→ Variance

↪ Mean of the squared deviations of the scores from their mean

$$\sigma^2 = \frac{\sum (X - \bar{X})^2}{N} = \frac{533.6}{10} = 53.36$$

→ Standard deviation: $\sigma = \sqrt{\sigma^2} = \sqrt{53.36} = 7.3$

→ Unbiased estimator of the population value of σ^2

$$S^2 = \frac{\sum (X - \bar{X})^2}{N - 1} = \frac{533.6}{9} = 59.32$$

→ S then is the estimator of σ

$$S = \sqrt{S^2} = \sqrt{59.32} = 7.7$$

The Normal Distribution

- Special bell shaped curve, defined by
 - ↳ Mean
 - ↳ Standard deviation
- Useful for a wide variety of statistical procedures
 - ↳ Can specify what proportion of the areas is to be found in any region of the curve
 - ↳ Many attributes are distributed in a normal or nearly normal manner
- About 2/3rds of the scores fall between -1σ and $+1\sigma$
- About 95% of the scores fall between -2σ and $+2\sigma$
- Over 99% of the scores fall between -3σ and $+3\sigma$

Bell Shaped Curve

