Data Management

CT051-3-M

ASIA PACIFIC UNIVERSITY OF TECHNOLOGY & INNOVATION

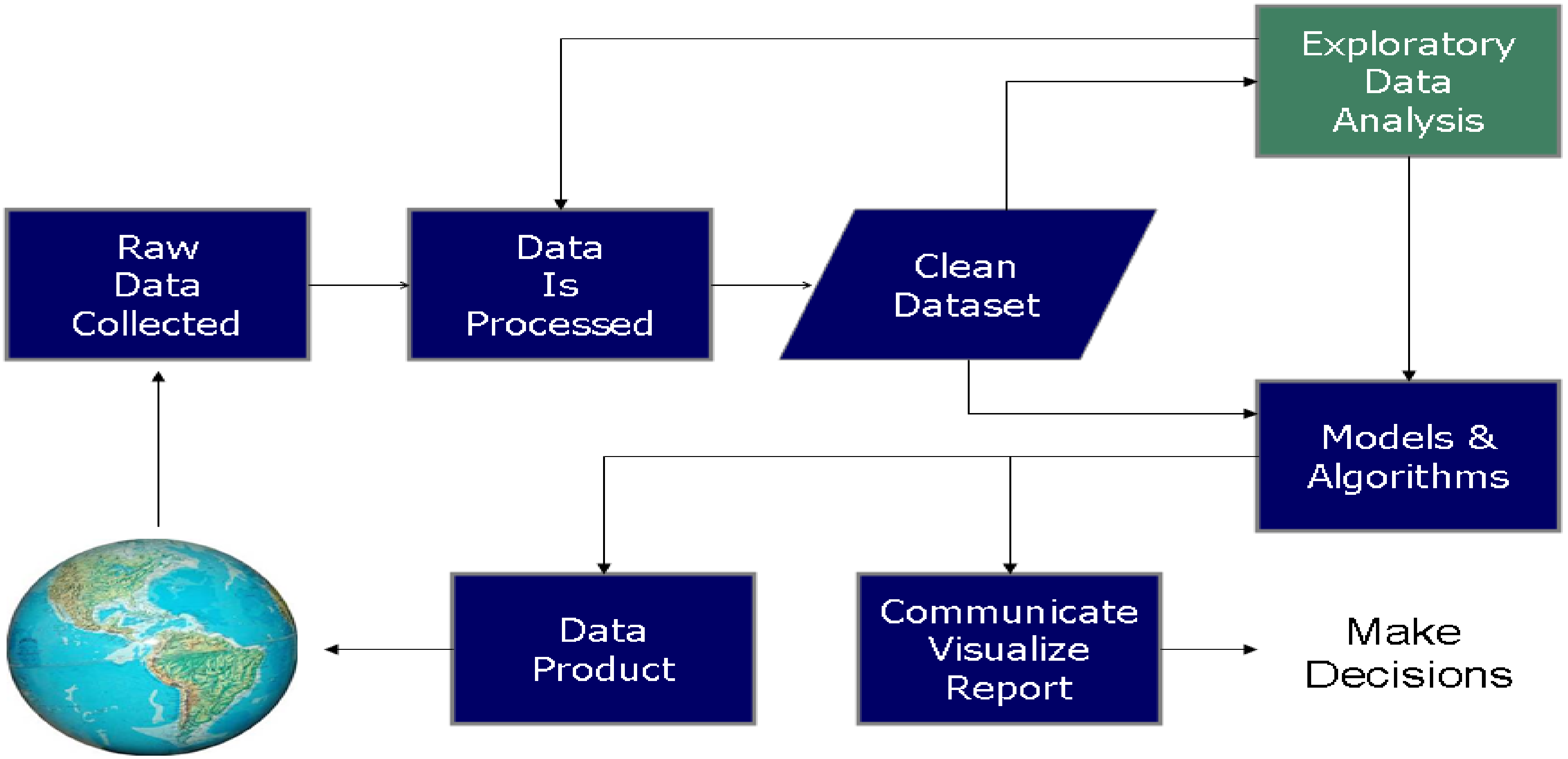
Topic 5 – Exploratory Data Analysis

Exploratory Data Analysis

- John Tukey (1970s)
- data
  - two components:
    - smooth + rough
    - patterned behaviour + random variation
- **resistant** measures/displays
  - little influenced by changes in a small proportion of the total number of cases
  - resistant to the effects of outliers
  - emphasizes *smooth* over *rough* components
- concepts apply to statistics and to graphical methods

Exploratory Data Analysis

Data Science Process



15,000	2
20,000	5
30,000	10
35,000	1
40,000	20
50,000	10

6 category

4 bins

15k -20 k = rank 1(7)

21k – 30k = rank 2 (10)

31-40 = rank 3 (21)

41-50 = rank 4 (30)

4 bins

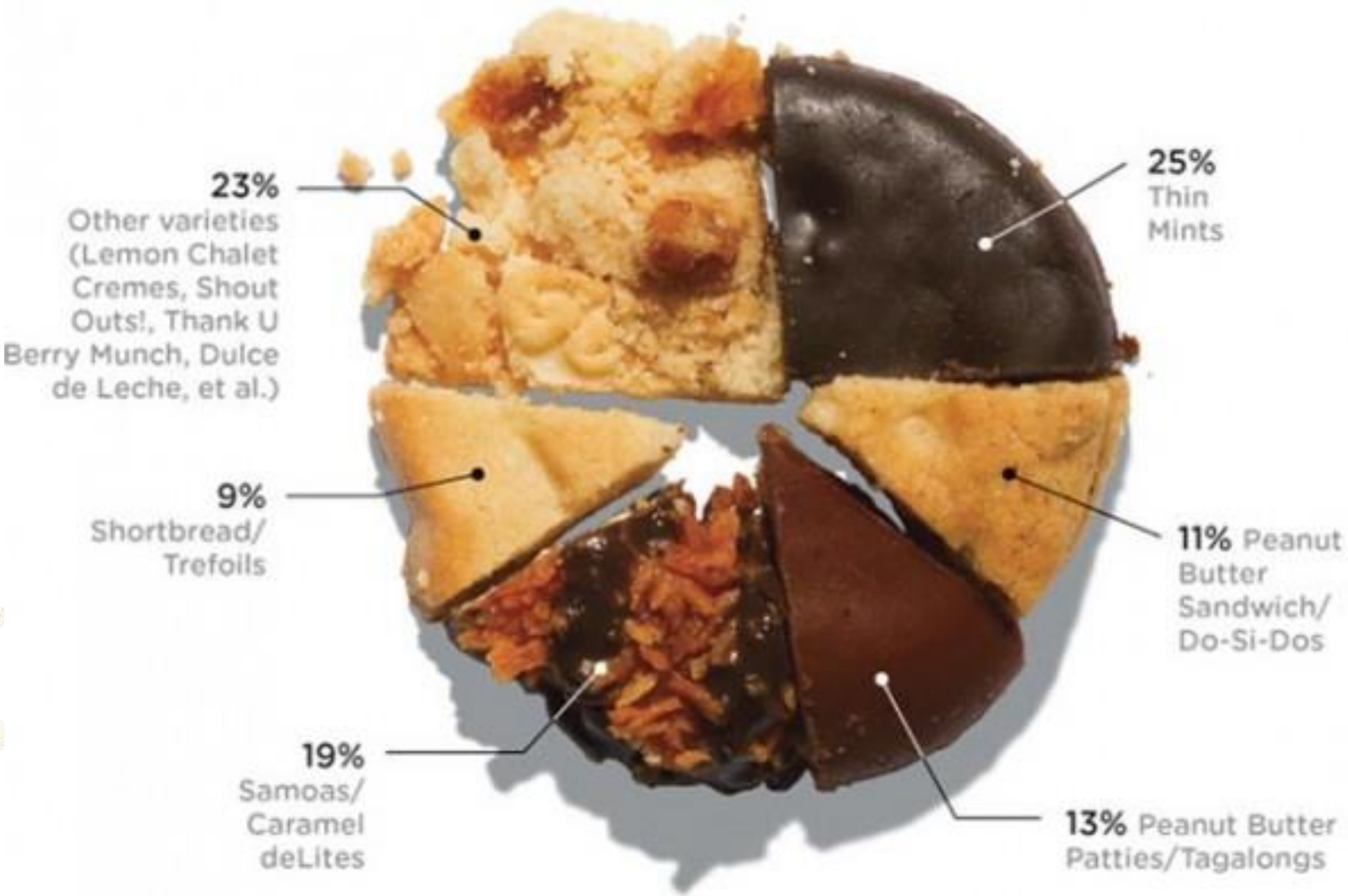
EDA and Visualization

- Exploratory Data Analysis (EDA) and Visualization are very important steps in any analysis task.

- get to know your data!
  - distributions (symmetric, normal, skewed)
  - data quality problems
  - outliers
  - correlations and inter-relationships
  - subsets of interest
  - suggest functional relationships

- Sometimes EDA or viz might be the goal!

Data Visualization – cake bakery



Exploring Data

- **Exploring data**
  - ❖ **Descriptive statistics**
    - ❑ **Categorical data**
      - ➢ **Frequency**
      - ➢ **Percentage (Row, Column or Total)**
    - ❑ **Continuous data: Measure of location**
      - ➢ **Mean**
      - ➢ **Median**
    - ❑ **Continuous data: Measure of variation**
      - ➢ **Standard deviation**
      - ➢ **Range (Min, Max)**
      - ➢ **Inter-quartile range (LQ, UQ)**
  - ❖ **Graphical illustrations**
    - ❑ **Categorical data**
      - ➢ **Bar chart**
      - ➢ **Clustered bar charts (two categorical variables)**
      - ➢ **Bar charts with error bars**
    - ❑ **Continuous data**
      - ➢ **Histogram (can be plotted against a categorical variable)**
      - ➢ **Box & Whisker plot (can be plotted against a categorical variable)**
      - ➢ **Dot plot (can be plotted against a categorical variable)**
      - ➢ **Scatter plot (two continuous variables)**

	Continuous	Discrete		
	Quantitative data	Qualitative / Categorical / Attribute data		
Measurement	**Units (example)**	**Ordinal (example)**	**Nominal (example)**	**Binary (example)**
Time of day	Hours, minutes, seconds	1, 2, 3, etc.	N/A	a.m./p.m.
Date	Month, date, year	Jan., Feb., Mar., etc.	N/A	Before / After
Cycle time	Hours, minutes, seconds, month, date, year	10, 20, 30, etc.	N/A	Before / After
Speed	Miles per hour/centimeters per second	10, 20, 30, etc.	N/A	Fast / Slow
Brightness	Lumens	Light, medium, dark	N/A	On / Off
Temperature	Degrees C or F	10, 20, 30, etc.	N/A	Hot / Cold
<Count data>	Number of things	10, 20, 30, etc.	N/A	Large / Small
Test scores	Percent, number correct	F, D, C, B, A	N/A	Pass / Fail
Defects	N/A	Number of cracks	N/A	Good / Bad
Defects	N/A	N/A	Oversized, missing	Good / Bad
Color	N/A	N/A	Red, blue, green	N/A
Location	N/A	N/A	East, West, South	Domestic / International
Groups	N/A	N/A	HR, Legal, IT	Exempt / Non-exempt
Anything	Percent	10, 20, 30, etc.	N/A	Above / Below

Exercise

1. Calculate the Mean, Median, Variance and Standard deviation.

Five students enrolled for Data Management module. The students score can range from 0 to 100.

The following are the students score:

83, 94, 30, 63, 66

Exercise

2. For the given two vector, identify the Mean, Median, Mode and Standard Deviation

A = {2, 2, 4, 4, 2, 5, 6}

B = {2, 2, 4, 400, 4, 2, 5, 6}

InterQuartile Range (IQR)

- When a dataset has outliers or extreme values, we summarize a typical value using the median as opposed to the mean.
- Variability is often summarized by a statistic called the interquartile range.

$$\text{Interquartile Range} = Q_3 - Q_1$$

Interquartile Range with an Odd Sample Size

Median = 72

Lower half | Upper half

63	64	64	70	72	76	77	81	81

Lower quarter

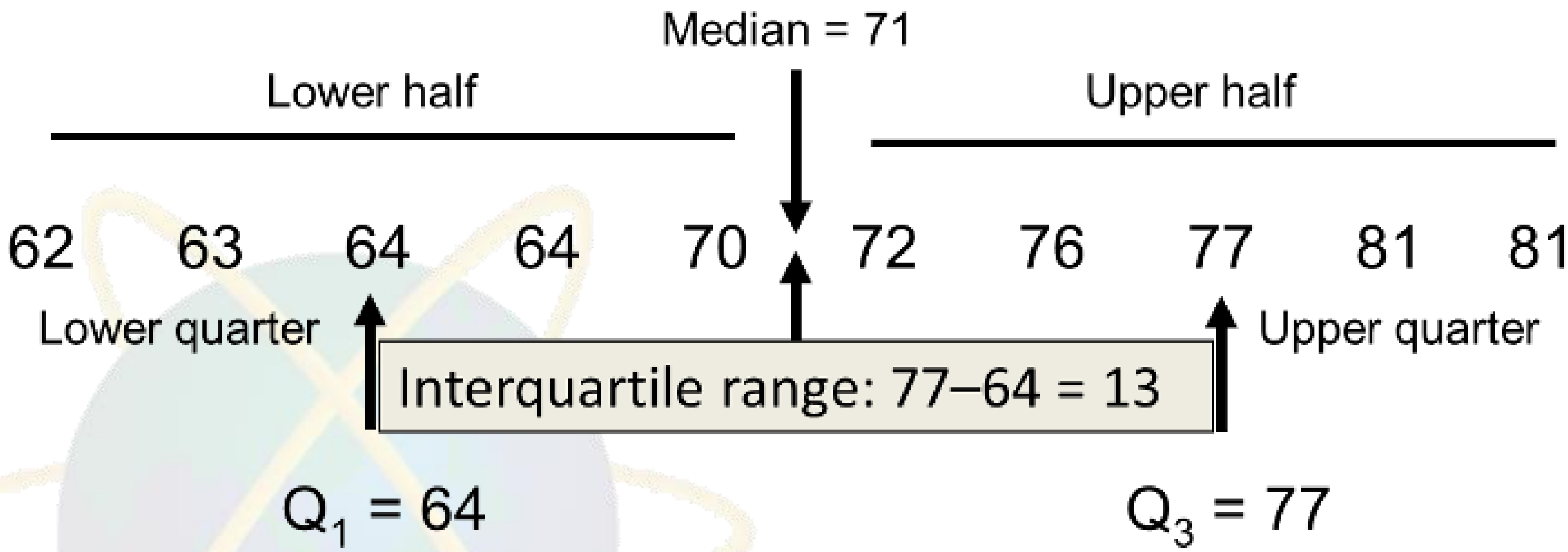
Interquartile range: 79–64 = 15

Upper quarter

$Q_1 = (64+64)/2=64$

$Q_3 = (77+81)/2=79$

Interquartile Range with Even Sample Size



Exercise

3. Apply the Interquartile Ranges (IQR) to determine the outliers that exist in the above distribution.

- Height measures of the graduate students are reported as follows:

130	132	138	136	131	153	131	133	129	133	110	132	129	134	135
132	135	134	133	132	130	131	134	135	135	134	136	133	133	130

Categorical Data

- Frequency Distribution
- Frequency Table

Categorical Data

- Organize qualitative data into a frequency table.
- Present a frequency table as a bar chart or a pie chart.
- Organize quantitative data into a frequency distribution.
- Present a frequency distribution for quantitative data using histograms, frequency polygons, and cumulative frequency polygons.

Frequency Distribution

Selling Prices ($ thousands)	Frequency
15 up to 18	8
18 up to 21	23
21 up to 24	17
24 up to 27	18
27 up to 30	8
30 up to 33	4
33 up to 36	2
Total	80

A **Frequency distribution** is a grouping of data into mutually exclusive categories showing the number of observations in each class. The table shows a frequency distribution for a set of quantitative data.

Frequency Table

FREQUENCY TABLE A grouping of qualitative data into mutually exclusive classes showing the number of observations in each class.

TABLE 2–1 Frequency Table for Vehicles Sold at Whitner Autoplex Last Month

Car Type	Number of Cars
Domestic	50
Foreign	30

Descriptive Statistics

Continuous Data: Measure of Location

- Mean
- Median
- Mode

Central value

- Give information concerning the average or typical score of a number of scores
  - mean
  - median

Central value: The Mean

- The Mean is a measure of *central value*
  - What most people mean by “average”
  - Sum of a set of numbers divided by the number of numbers in the set

$$\frac{1+2+3+4+5+6+7+8+9+10}{10}=\frac{55}{10}=5.5$$

Central value: The Mean

Arithmetic average:

Sample

$$\overline{X} = \frac{\Sigma x}{n}$$

Population

$$\mu = \frac{\Sigma x}{N}$$

$$X = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$$

$$\sum X / n = 5.5$$

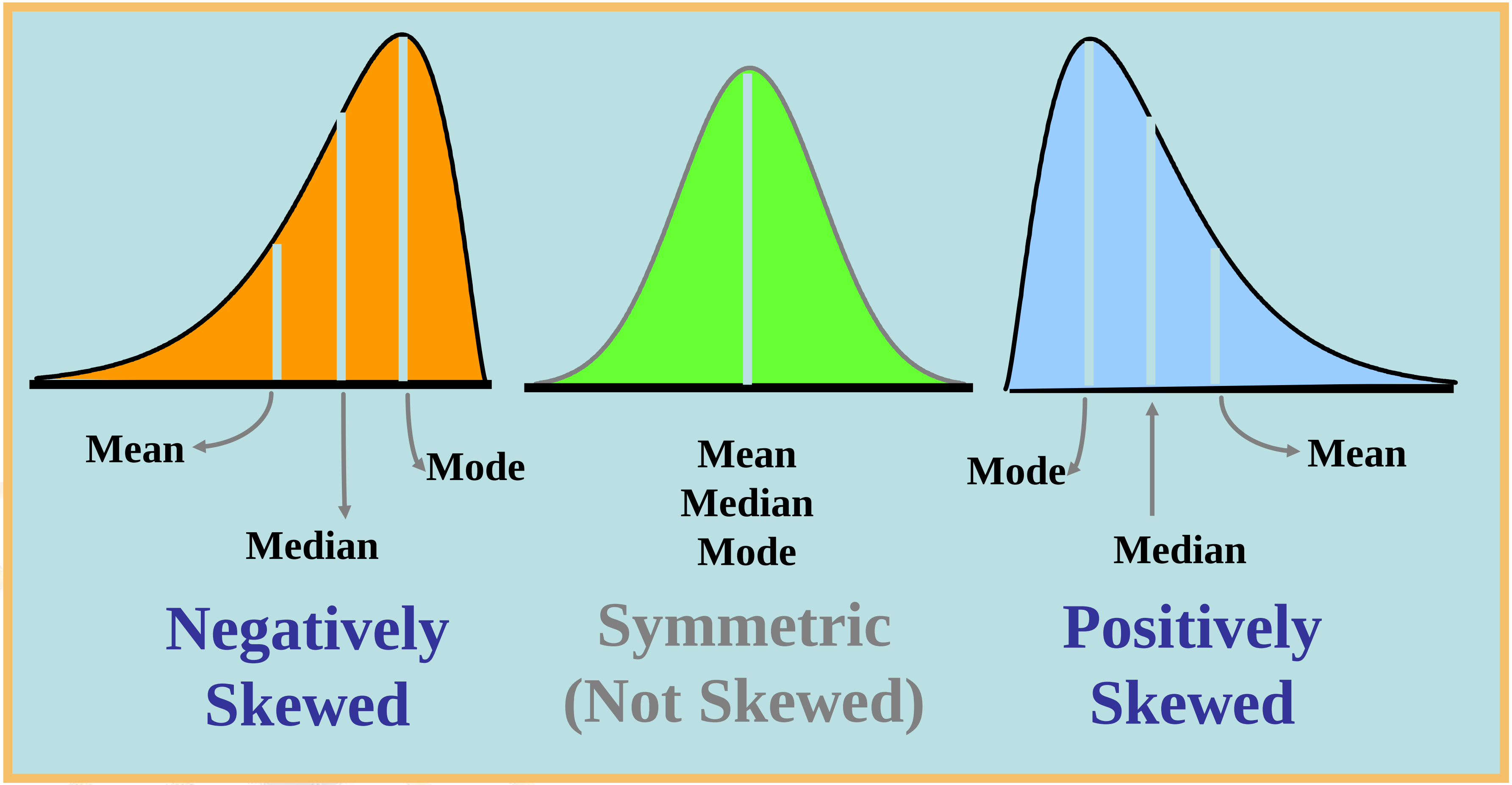
Central value: The Median

- Middlemost or most central item in the set of ordered numbers; it separates the distribution into two equal halves
- If *odd n*, middle value of sequence
  - if $X = [1,2,4,6,9,10,12,14,17]$
  - then 9 is the median
- If *even n*, average of 2 middle values
  - if $X = [1,2,4,6,9,10,11,12,14,17]$
  - then 9.5 is the median; i.e., (9+10)/2
- Median is not affected by extreme values

When to Use What

- Mean is a great measure. But, there are time when its usage is inappropriate or impossible.
  - Nominal data: Mode
  - The distribution is bimodal: Mode
  - You have ordinal data: Median or mode
  - Are a few extreme scores: Median

Mean, Median, Mode



Descriptive Statistics

Continuous Data: Measure of Variance

- Standard deviation
- Range
- Interquartile

Range

The spread, or the distance, between the lowest and highest values of a variable.

To get the range for a variable, you subtract its lowest value from its highest value.

Class A--IQs of 13 Students

102	115
128	109
131	89
98	106
140	119
93	97
110	

Class A Range = 140 - 89 = 51

Class B--IQs of 13 Students

127	162
131	103
96	111
80	109
93	87
120	105
109	

Class B Range = 162 - 80 = 82

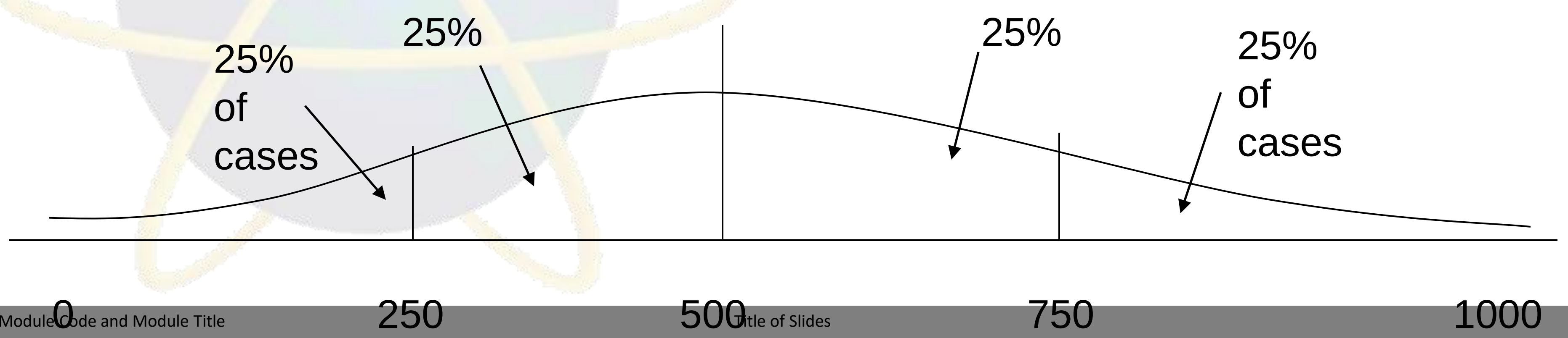
Interquartile Range

A quartile is the value that marks one of the divisions that breaks a series of values into four equal parts.

The median is a quartile and divides the cases in half.

25^{th} percentile is a quartile that divides the first ¼ of cases from the latter ¾.
75^{th} percentile is a quartile that divides the first ¾ of cases from the latter ¼.

The interquartile range is the distance or range between the 25^{th} percentile and the 75^{th} percentile. Below, what is the interquartile range?



Standard Deviation

To convert variance into something of meaning, let's create standard deviation.

The square root of the variance reveals the average deviation of the observations from the mean.

$$\text{s.d.} = \sqrt{\frac{\Sigma(Y_i - \text{Y-bar})^2}{n - 1}}$$

Standard Deviation

For Class A, the standard deviation is:

$$\sqrt{235.45} = 15.34$$

The average of persons' deviation from the mean IQ of 110.54 is 15.34 IQ points.

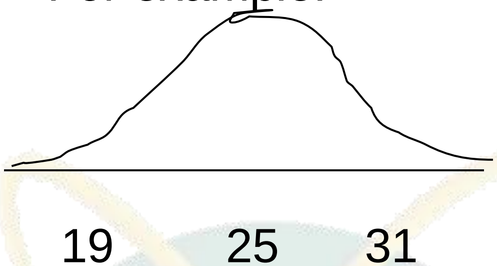
Review:

1. Deviation
2. Deviation squared
3. Sum of squares
4. Variance
5. Standard deviation

Standard Deviation

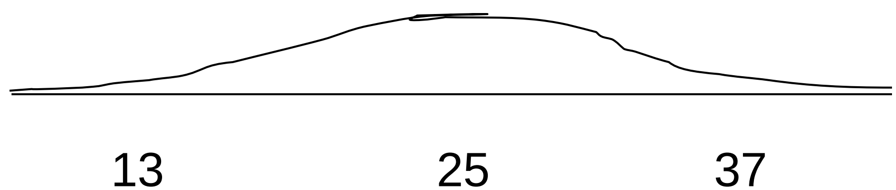
1. Larger s.d. = greater amounts of variation around the mean.

   For example:



$\bar{Y} = 25$

s.d. = 3



$\bar{Y} = 25$

s.d. = 6

2. s.d. = 0 only when all values are the same (only when you have a constant and not a “variable”)
3. If you were to “rescale” a variable, the s.d. would change by the same magnitude—if we changed units above so the mean equaled 250, the s.d. on the left would be 30, and on the right, 60
4. Like the mean, the s.d. will be inflated by an outlier case value.

Standard Deviation

- Note about computational formulas:
  - Your book provides a useful short-cut formula for computing the variance and standard deviation.
  - This is intended to make hand calculations as quick as possible.
  - They obscure the conceptual understanding of our statistics.
  - SPSS and the computer are “computational formulas” now.

Practical Application for Understanding Variance and Standard Deviation

Even though we live in a world where we pay real dollars for goods and services (not percentages of income), most American employers issue raises based on percent of salary.

Why do supervisors think the most fair raise is a percentage raise?

Answer: 1) Because higher paid persons win the most money.

2) The easiest thing to do is raise everyone's salary by a fixed percent.

If your budget went up by 5%, salaries can go up by 5%.

The problem is that the flat percent raise gives unequal increased rewards. . .

Practical Application for Understanding Variance and Standard Deviation

Acme Toilet Cleaning Services

Salary Pool: $200,000

Incomes:

President: $100K; Manager: 50K; Secretary: 40K; and Toilet Cleaner: 10K

Mean: $50K

Range: $90K

Variance: $1,050,000,000

Standard Deviation: $32.4K

These can be considered "measures of inequality"

Now, let's apply a 5% raise.

Practical Application for Understanding Variance and Standard Deviation

After a 5% raise, the pool of money increases by $10K to $210,000

Incomes:

President: $105K; Manager: 52.5K; Secretary: 42K; and Toilet Cleaner: 10.5K

Mean: $52.5K – went up by 5%

Range: $94.5K – went up by 5%

Variance: $1,157,625,000

Standard Deviation: $34K –went up by 5%

Measures of Inequality

The flat percentage raise increased inequality. The top earner got 50% of the new money. The bottom earner got 5% of the new money. Measures of inequality went up by 5%.

Last year's statistics:

Acme Toilet Cleaning Services annual payroll of $200K

Incomes:

$100K, 50K, 40K, and 10K

Mean: $50K

Range: $90K; Variance: $1,050,000,000; Standard Deviation: $32.4K

Practical Application for Understanding Variance and Standard Deviation

The flat percentage raise increased inequality. The top earner got 50% of the new money. The bottom earner got 5% of the new money. Inequality increased by 5%.

Since we pay for goods and services in real dollars, not in percentages, there are substantially more new things the top earners can purchase compared with the bottom earner for the rest of their employment years.

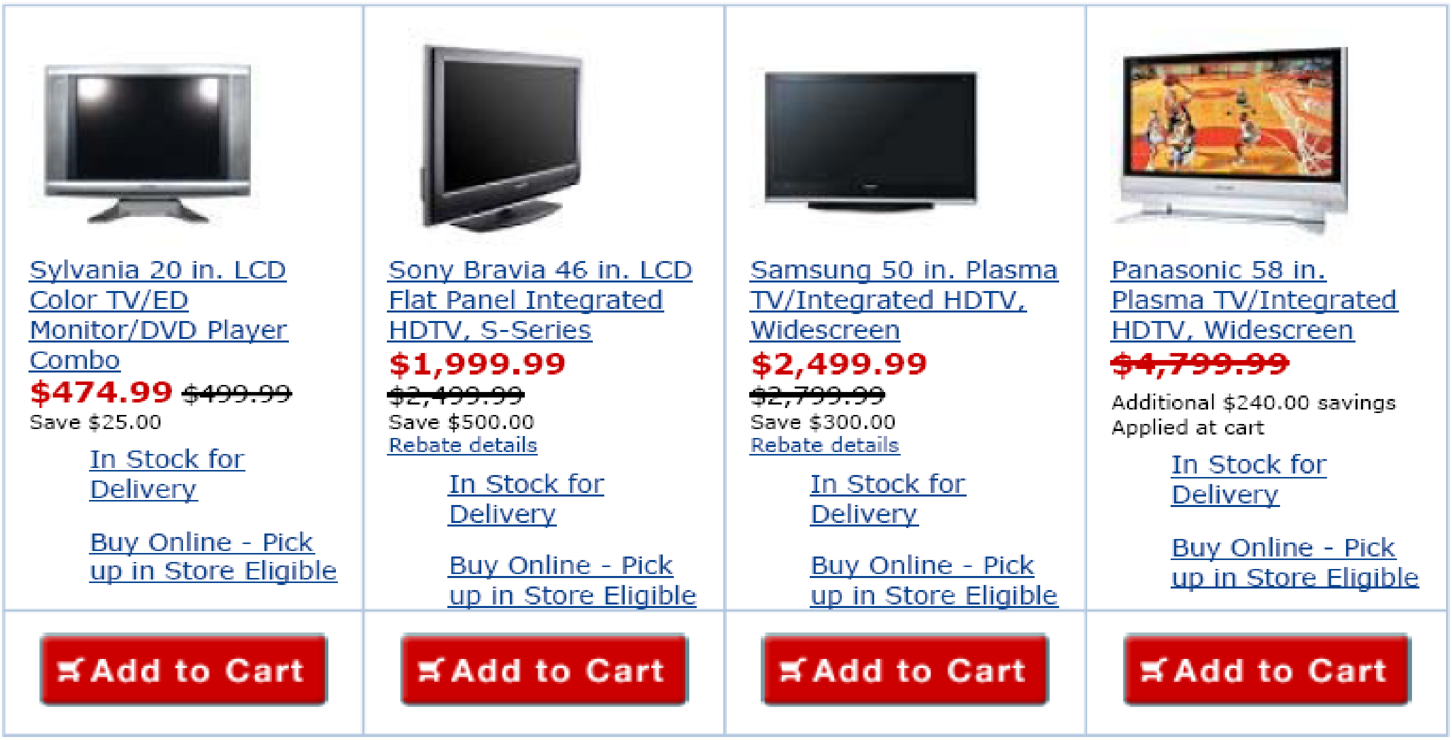
Acme Toilet Cleaning Services is giving the earners $5,000, $2,500, $2,000, and $500 more respectively ***each and every year forever***.

What does this mean in terms of compounding raises?

Acme is essentially saying: "Each year we'll buy you a new TV, in addition to everything else you buy, here's what you'll get:"

Practical Application for Understanding Variance and Standard Deviation

Toilet Cleaner	Secretary	Manager	President



The gap between the rich and poor expands.

This is why some progressive organizations give a percentage raise with a flat increase for lowest wage earners. For example, 5% or $1,000, whichever is greater.

Visualization

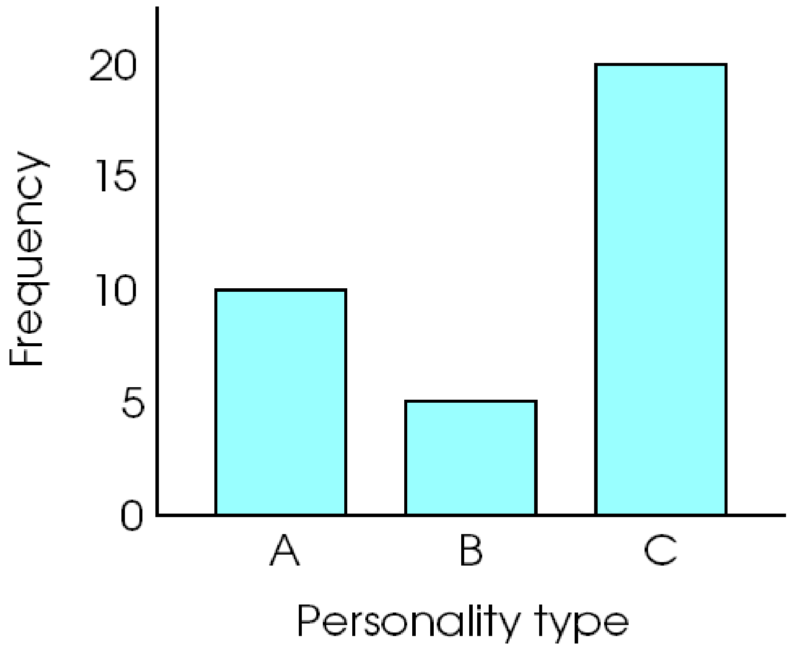
Visualization

Categorical Data

- Bar Chart

Bar Graphs

- For categorical data
- Like a histogram, but with gaps between bars
- Useful for showing two samples side-by-side



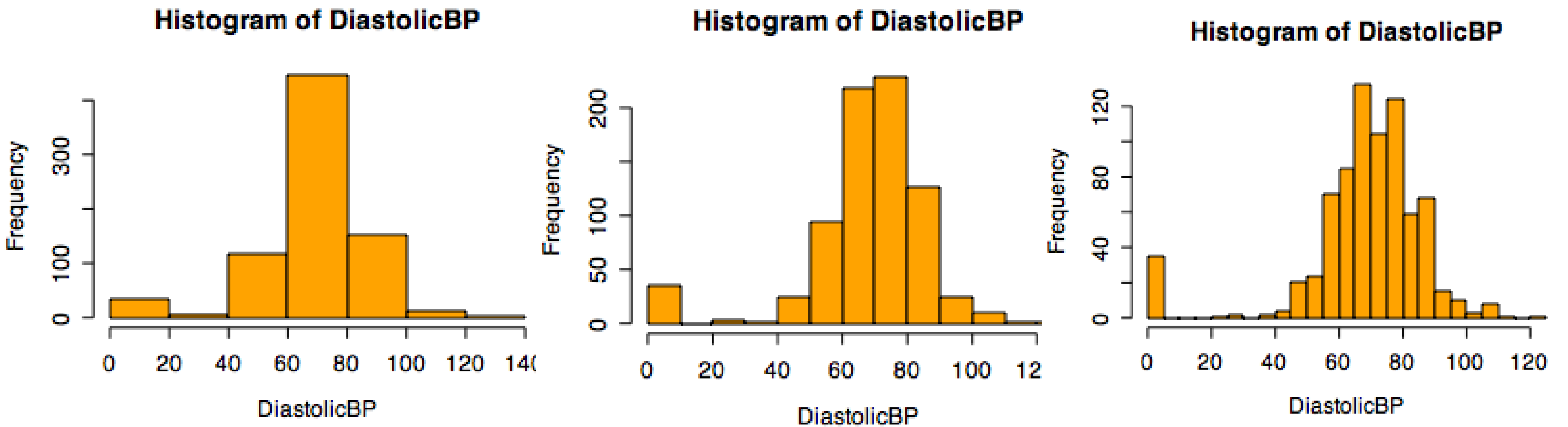
Visualization

Continuous Data

- Histogram
- Box & Whisker Plot

Single Variable Visualization

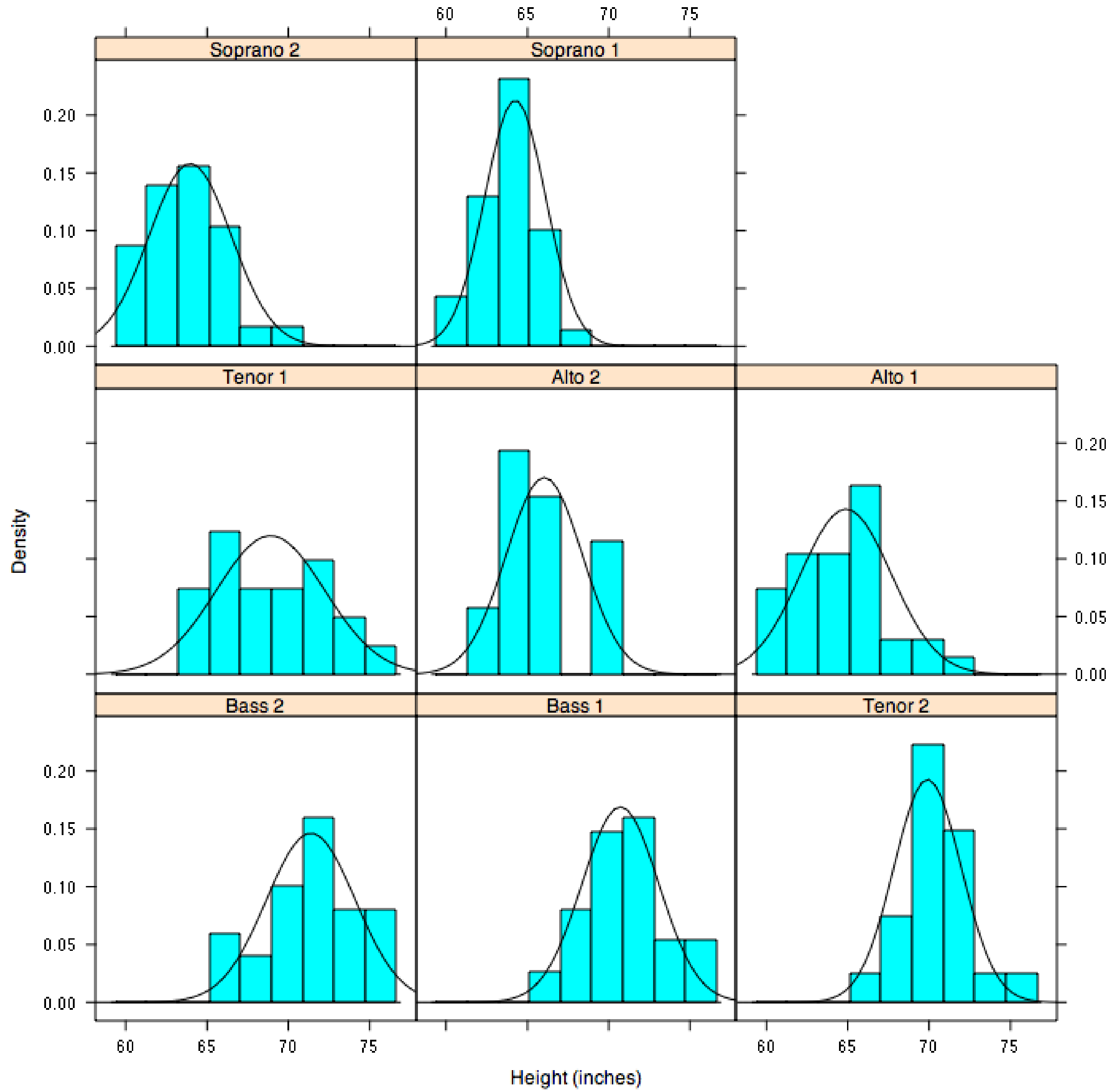
- Histogram:
  - Shows center, variability, skewness, modality,
  - outliers, or strange patterns.
  - Bin width and position matter
  - Beware of real zeros
  - ***No gaps between bars***



Issues with Histograms

- For small data sets, histograms can be misleading.
  - Small changes in the data, bins, or anchor can deceive
- For large data sets, histograms can be quite effective at illustrating general properties of the distribution.
- Histograms effectively only work with 1 variable at a time
  - But ‘small multiples’ can be effective

But be careful with axes and scales!



Box-Plots

A way to graphically portray almost all the descriptive statistics at once is the box-plot.

A box-plot shows: Upper and lower quartiles

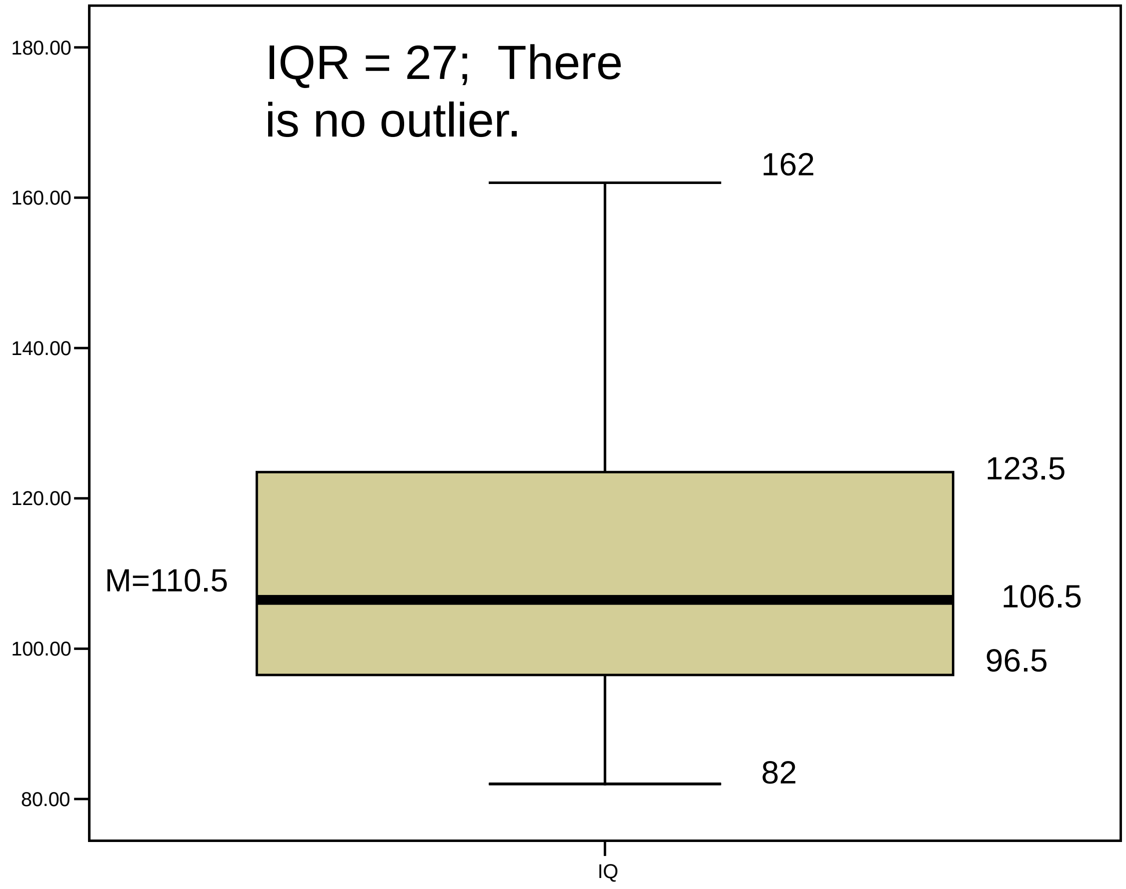
Mean

Median

Range

Outliers (1.5 IQR)

Box-Plots



IQR = 27; There is no outlier.
162
123.5
M=110.5
106.5
96.5
82
180.00
160.00
140.00
120.00
100.00
80.00
IQ

IQV—Index of Qualitative Variation

- For nominal variables
- Statistic for determining the dispersion of cases across categories of a variable.
- Ranges from 0 (no dispersion or variety) to 1 (maximum dispersion or variety)
- 1 refers to even numbers of cases in all categories, NOT that cases are distributed like population proportions
- IQV is affected by the number of categories

IQV—Index of Qualitative Variation

To calculate:

$$IQV = \frac{K(100^2 - \Sigma\, cat.\%^2)}{100^2(K-1)}$$

K=# of categories

Cat.% = percentage in each category

IQV—Index of Qualitative Variation

Problem: Is SJSU more diverse than UC Berkeley?

Solution: Calculate IQV for each campus to determine which is higher.

SJSU:

Percent	Category
00.6	Native American
06.1	Black
39.3	Asian/PI
19.5	Latino
34.5	White

UC Berkeley:

Percent	Category
00.6	Native American
03.9	Black
47.0	Asian/PI
13.0	Latino
35.5	White

What can we say before calculating? Which campus is more evenly distributed?

$$IQV = \frac{K\,(100^2 - \Sigma \text{ cat.}\%^2)}{100^2(K-1)}$$

IQV—Index of Qualitative Variation

Problem: Is SJSU more diverse than UC Berkeley? YES

Solution: Calculate IQV for each campus to determine which is higher.

SJSU:

Percent	Category	$\%^2$
00.6	Native American	0.36
06.1	Black	37.21
39.3	Asian/PI	1544.49
19.5	Latino	380.25
34.5	White	1190.25

$K = 5$ $\quad \Sigma \text{ cat.}\%^2 = 3152.56$

$100^2 = 10000$

$$IQV = \frac{K\,(100^2 - \Sigma \text{ cat.}\%^2)}{100^2(K - 1)}$$

$5(10000 - 3152.56) = 34237.2$

$10000(5 - 1) = 40000$ SJSU IQV = .856

UC Berkeley:

Percent	Category	$\%^2$
00.6	Native American	0.36
03.9	Black	15.21
47.0	Asian/PI	2209.00
13.0	Latino	169.00
35.5	White	1260.25

$k = 5$ $\quad \Sigma \text{ cat.}\%^2 = 3653.82$

$5(10000 - 3653.82) = 31730.9$

$10000(5 - 1) = 40000$ UCB IQV =.793

Question & Answer Session

Q & A