Introduction to Informatics

Lecture 17: Inductive Model Building

Centrality and Dispersion

It is proven that the celebration of birthdays is healthy. Statistics show that those people who celebrate the most birthdays become the oldest.

The Japanese eat very little ***fat*** and suffer fewer heart attacks than the British or the Americans.

On the other hand, the French eat a lot of ***fat*** and also suffer fewer heart attacks than the British or the Americans.

The Japanese drink very little ***red wine*** and suffer fewer heart attacks than the British or the Americans.

The Italians drink excessive amounts of ***red wine*** and also suffer fewer heart attacks than the British or the Americans.

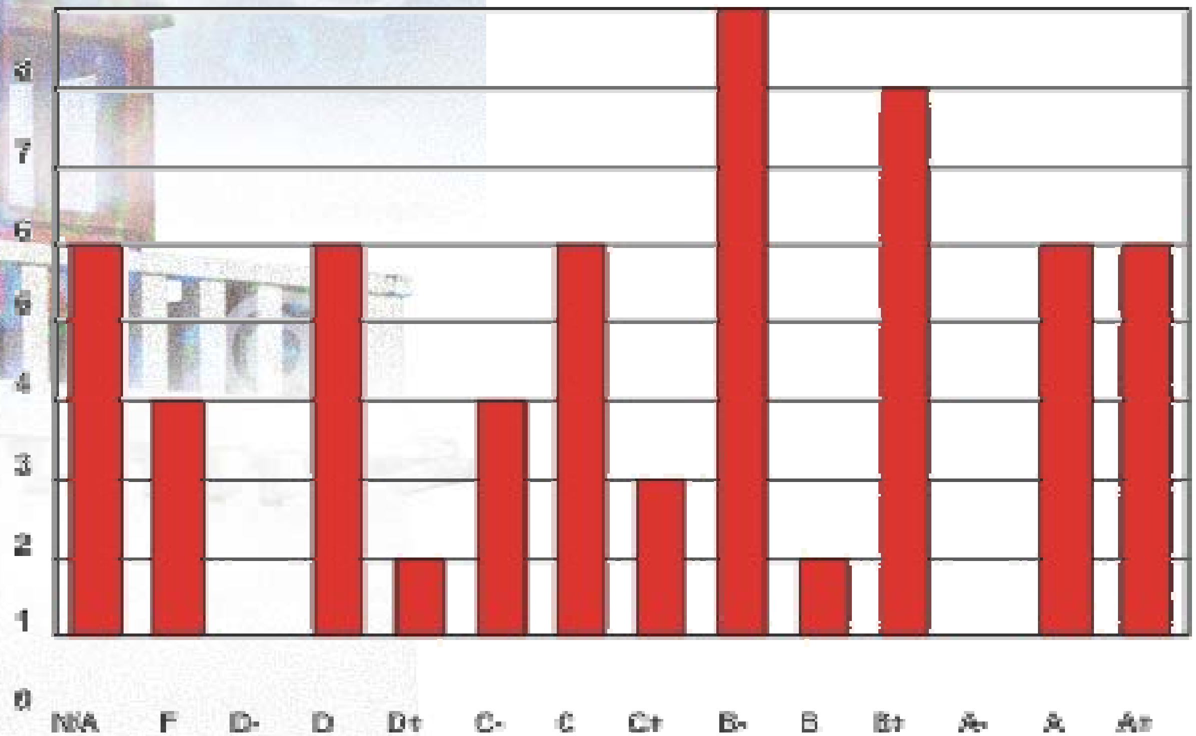
Conclusion: Eat and drink whatever you like. It's speaking English that kills you

Luis M.Rocha and Santiago Schnell

Readings until now

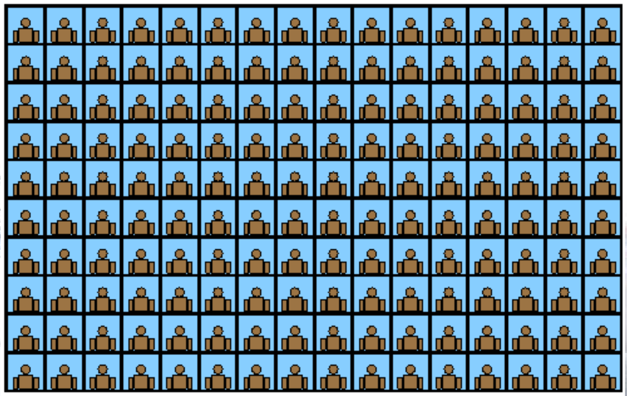
- Lecture notes
  - Posted online
    - http://informatics.indiana.edu/rocha/i101
      - *The Nature of Information*
      - *Technology*
      - *Modeling the World*
  - *@ infoport*
    - *http://infoport.blogspot.com*
  - From course package
    - Von Baeyer, H.C. [2004]. *Information: The New Language of Science*. Harvard University Press.
      - Chapters 1, 4 (pages 1-12)
    - From Andy Clark's book "*Natural-Born Cyborgs*"
      - Chapters 2 and 6 (pages 19 - 67)
    - From Irv Englander's book "*The Architecture of Computer Hardware and Systems Software*"
      - Chapter 3: Data Formats (pp. 70-86)
    - Klir, J.G., U. St. Clair, and B.Yuan [1997]. Fuzzy Set Theory: foundations and Applications. Prentice Hall
      - Chapter 2: Classical Logic (pp. 87-97)
      - Chapter 3: Classical Set Theory (pp. 98-103)
    - Norman, G.R. and D.L. Streinrt [2000]. *Biostatistics: The Bare Essentials*.
      - Chapters 1-3 (pages 105-129)

Midterm Exam

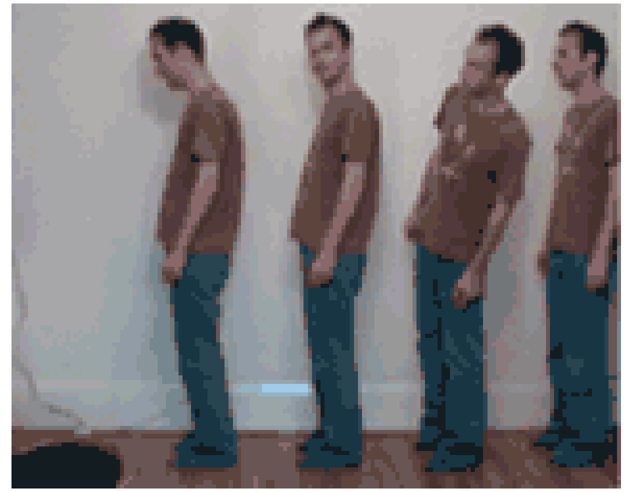


- Final Exam
    - Thursday, May 3rd, 7:15-9:15 p.m.

- James Waugh, Anna Wong, Andrew Kim, John Oglesby, and Jake Marsh
    - Excellent!



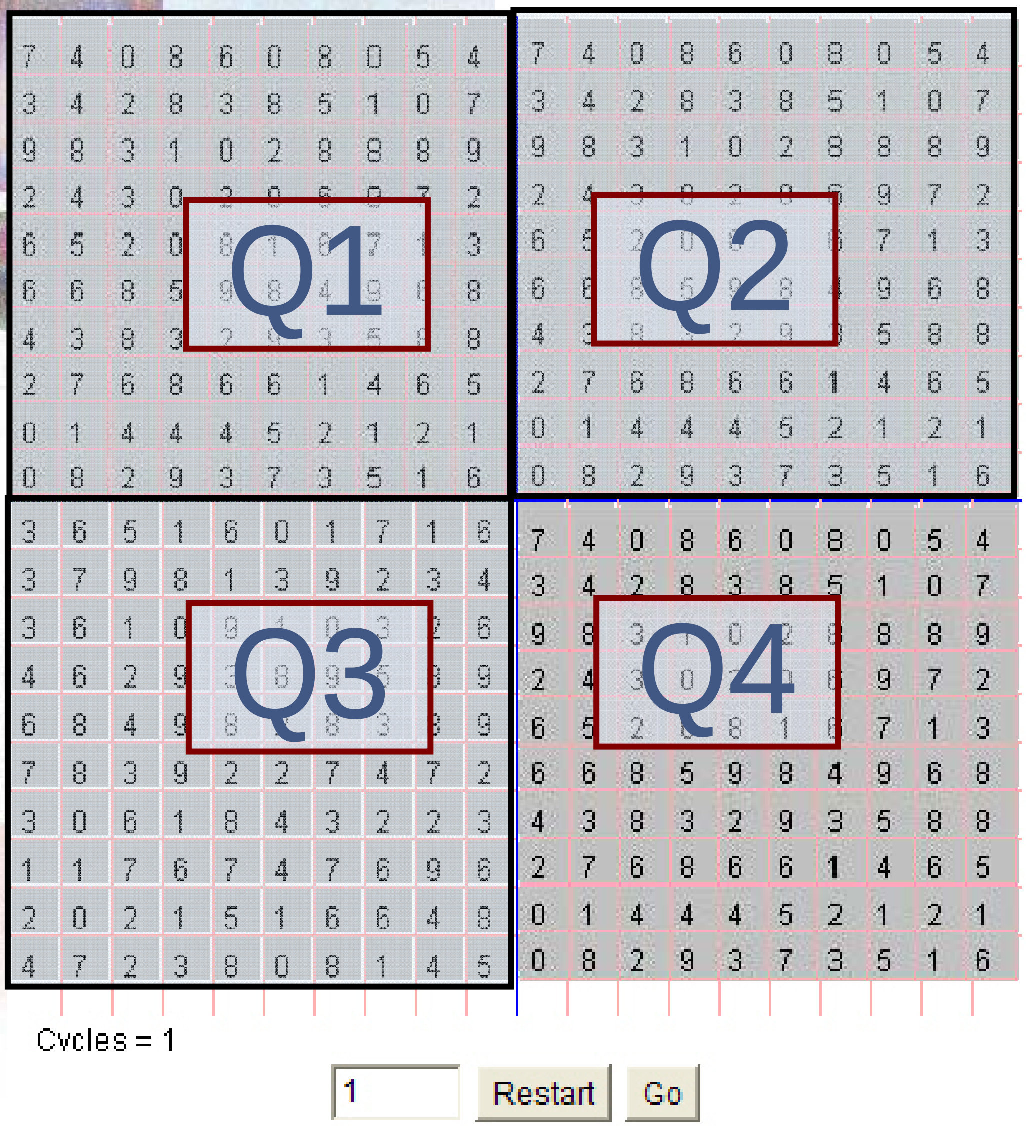
Assignment Situation



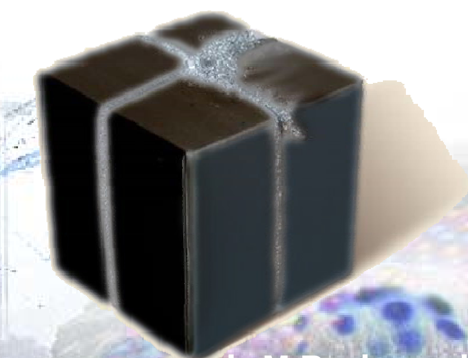
- Labs
  - Past
    - Lab 1: Blogs
      - Closed (Friday, January 19): Grades Posted
    - Lab 2: Basic HTML
      - Closed (Wednesday, January 31): Grades Posted
    - Lab 3: Advanced HTML: Cascading Style Sheets
      - Closed (Friday, February 2): Grades Posted
    - Lab 4: More HTML and CSS
      - Closed (Friday, February 9): Grades Posted
    - Lab 5: Introduction to Operating Systems: Unix
      - Closed (Friday, February 16): Grades Posted
    - Lab 6: More Unix and FTP
      - Closed (Friday, February 23): Grades Posted
    - Lab 7: Logic Gates
      - Closed: due Friday, March 9
  - Next: Lab 8
    - Intro to Statistical Analysis using Excel
      - March 22 & 23, Due Friday, March 30

- Assignments
  - Individual
    - First installment
      - Closed: February 9: Grades Posted
    - Second Installment
      - Past: March 2, Being Grades Posted
    - Third installment
      - Presented on March 8^{th}, Due on March 30^{th}
  - Group
    - First Installment
      - Past: March $9^{th,}$ Being graded
    - Second Installment
      - March 29; Due Friday, April 6

Individual Assignment – Part III

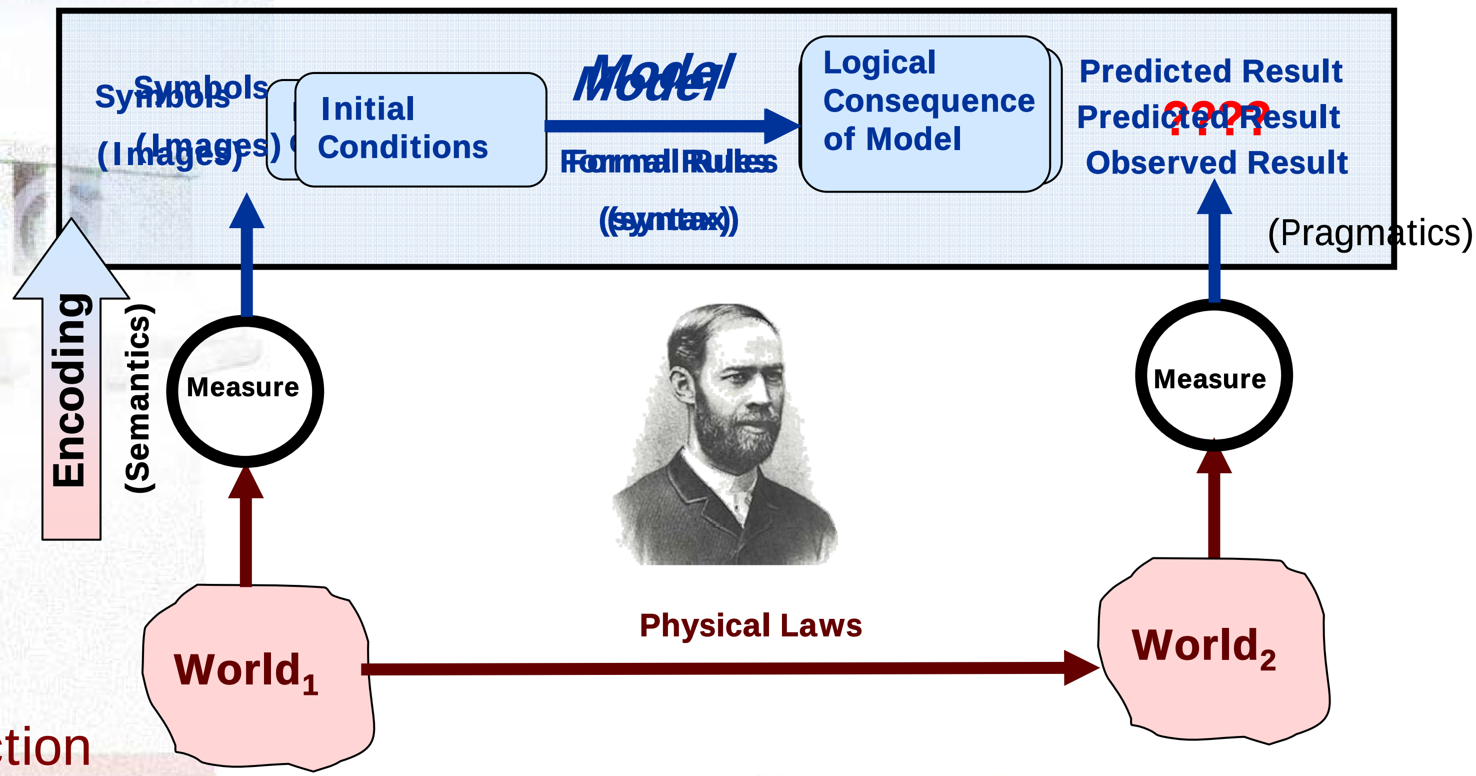


- Step by step analysis of "dying" squares
  - 3rd Installment
    - Presented: March 8th
    - Due: March 30th
  - 4th Installment
    - Presented: April 5th
    - Due: April 20th
- Use descriptive statistics
  - To uncover rules inductively
    - E.g. the behavior of evens and odds, individual numbers, or ranges of cycles, etc.



The Modeling Relation

Hertz' Modeling Paradigm



- Induction
  - Requires attention to data collection and ***description***
- Rules from Inference
  - From Data analysis
  - Produce Conclusions

Deduction vs. Induction

- Deductive Inference ← **Logic**
  - If the premises are true, we have absolute ***certainty*** of the conclusion
- Inductive Inference ← **Uncertainty**
  - Conclusion supported by ***good evidence*** (significant number of examples/observations) but not full certainty -- ***likelihood***

Measuring Central Tendency

The number that is meant to convey the idea of a ***typical*** or representative value for the data array or distribution

Ideas about Central Tendency

- Average life expectancy (in seconds) of an enemy soldier in a Chuck Norris film : 4
- Average Salary of Pro Wrestlers: $47,500 /yr.
  - If Pro Wrestling didn't exist: $4.25/hr.
- Average miles per gallon you can expect if a car maker's ad says " 30 mpg, city": 23
- The 50-50-90 rule:
  - Anytime you have a 50-50 chance of getting something right, there's a 90% probability you'll get it wrong.
- Did you hear about the statistician who put her head in the oven and her feet in the refrigerator?
  - She said, "On average, I feel just fine.“

Mean (Ungrouped Data)

Population Mean

$$\mu = \frac{\sum X}{N}$$

Sum of values of all observations

Number of elements in the population

Sample Mean

$$\bar{X} = \frac{\sum X}{n}$$

Sum of values of all observations

Number of elements in the sample

Mean Example

Percentile Increase in SAT Verbal Scores	**Student**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
	Increase	**9**	**7**	**7**	**6**	**4**	**4**	**2**

$$\bar{X} = \frac{\sum X}{n} = \frac{9+7+7+6+4+4+2}{7}$$

$$= \frac{39}{7}$$

$$= 5.6 \text{ points per student}$$

Observations about the Mean

- Advantages:
  - Familiar and intuitively clear to most people
  - Every data set has one and only one mean
  - Useful for performing statistical procedures
- Disadvantages:
  - May be affected by extreme values
  - Tedious to compute
  - Difficult to compute for data set with open-ended classes

Weighted Mean Example

Sorted Data: 30 values (Yards Produced by Carpet Looms)

Frequency Distribution

Class		Frequency
15.2	x	1
15.3		1
15.5		1
15.6	...	2
15.7		2
15.8	x	4
15.9		5
16.0		2
16.1		1
16.2	...	2
16.3		3
16.4		3
16.8		3

30

15.2

\+

63.2

\+

480.3

$$\bar{X}_f = \frac{\sum (f \times X)}{\sum f}$$

Weighted Mean:
16.01

The Weighted Mean

Takes into account the *importance* of each value to the overall total

$$\overline{X}_f = \frac{\sum(f \times X)}{\sum f}$$

where:

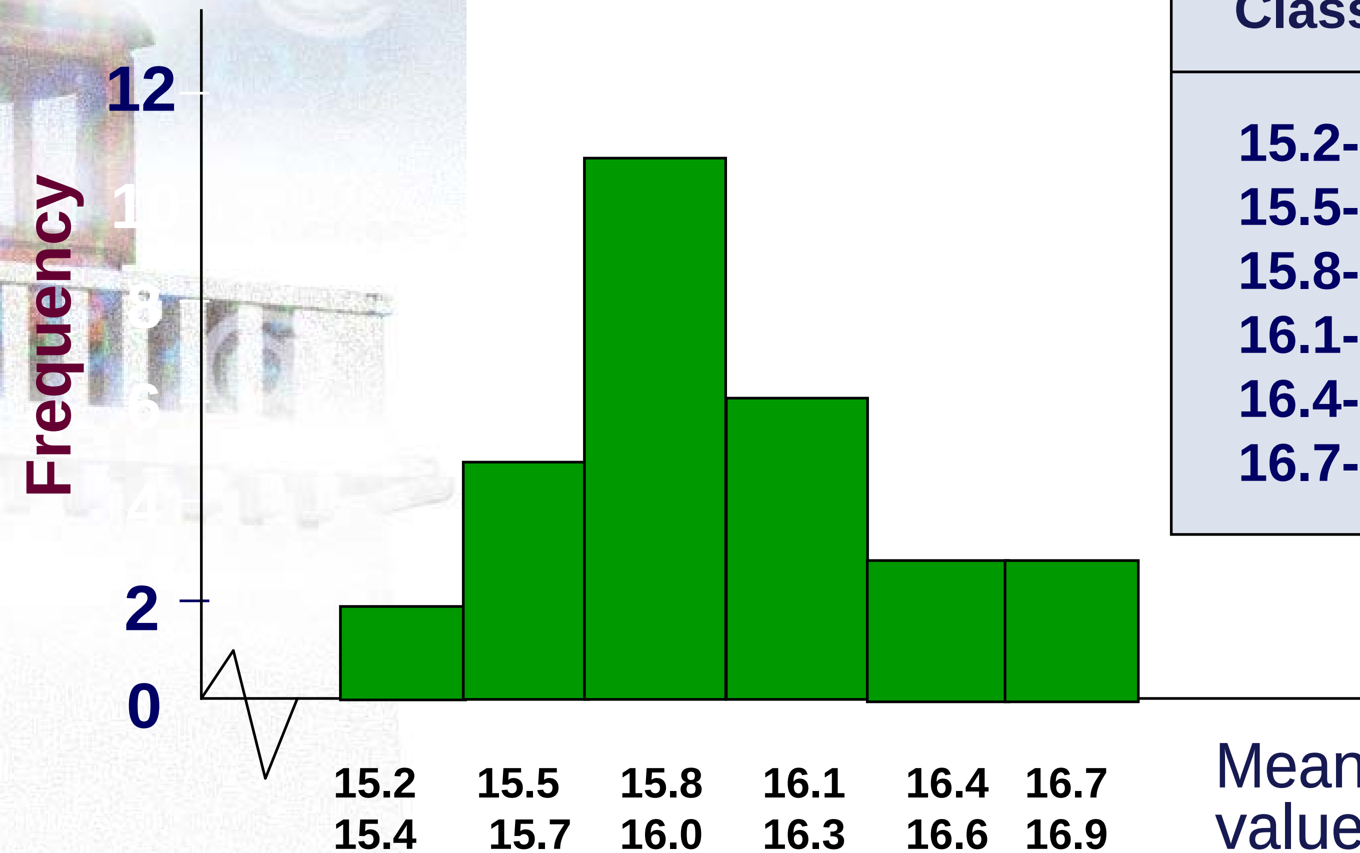
$\overline{X}_f$ = symbol for the weighted mean

f = frequency of each value/class

$\sum(f \times X)$ = sum of the frequency of each element times that element

$\sum w$ = sum of all the frequencies = N

Weighted Mean in Grouped Frequency Distributions



Class	Frequency
15.2-15.4	2
15.5-15.7	5
15.8-16.0	11
16.1-16.3	6
16.4-16.6	3
16.7-16.9	3

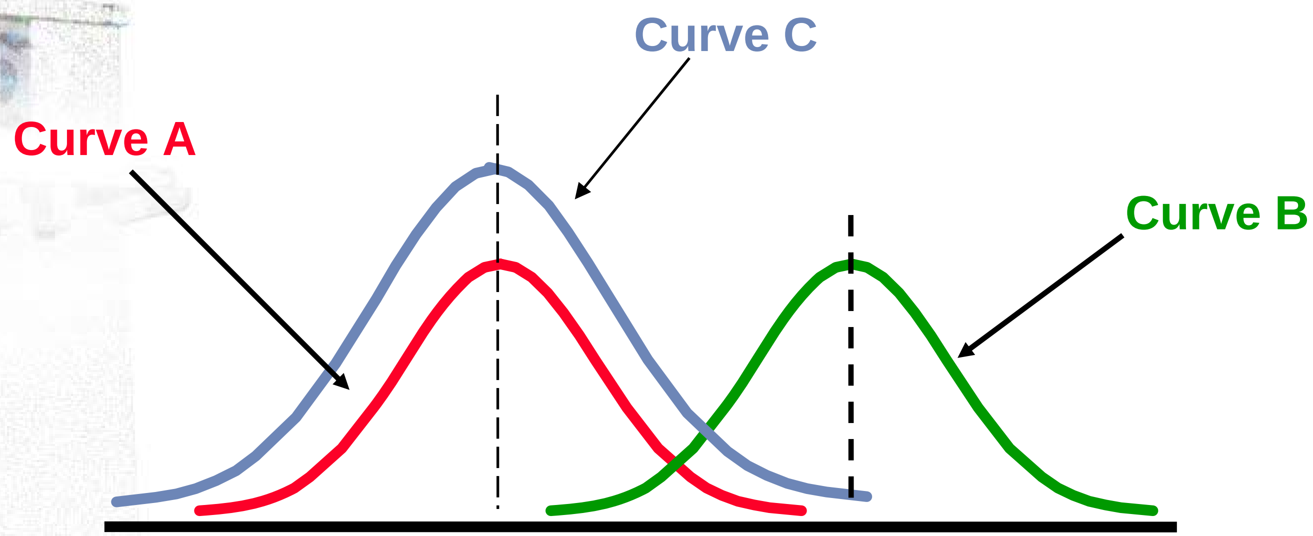
Mean is the mid-value of classes

$$\overline{X}_f = \frac{\sum (f \times X)}{\sum f}$$

((15.3 x 2) + (15.6 x 5) + (15.9 x 11) + (16.2 x 6) + (16.5 x 3) + (16.8 x 3))/30 = 16.02

Luis M.Rocha and Santiago Schnell

Comparison of Mean for 3 Frequency Distributions



- The mean is the balancing point of the frequency distribution
- Histograms balance at the mean

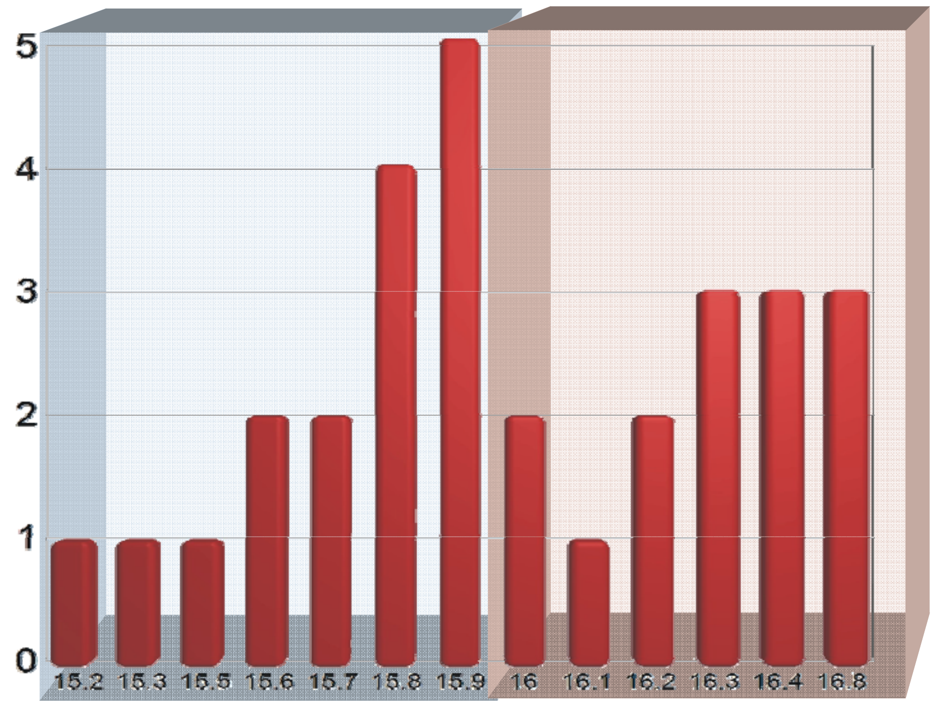
Mean

Mean: 16.01

Frequency Distribution

Carpet Loom Example

Class	Frequency	
15.2	1	16
15.3	1	
15.5	1	
15.6	2	
15.7	2	
15.8	4	
15.9	5	
16.0	2	14
16.1	1	
16.2	2	
16.3	3	
16.4	3	
16.8	3	
480.3	30	



- The mean is the balancing point of the frequency distribution
- Histograms balance at the mean

Median

- The number that is exactly in the middle of the data array
  - Middlemost or most central item in the set of ordered numbers
    - If *odd n*, middle value of sequence
    - If *even n*, average of 2 middle values

$$\tilde{x}$$

Number of items in the array

$$\text{Median} = \left(\frac{n+1}{2}\right)th \text{ item in the data array}$$

Median: Odd Sample Size

Times for track-team members	Item in data array	1	2	3	4	5	6	7
	Time (in minutes)	4.2	4.3	4.7	4.8	5.0	5.1	9.0

Median

$$\text{Positioning Point} = \frac{n+1}{2} = \frac{7+1}{2} = 4.0$$

$$\text{Median} = 4.8$$

Median of Even Sample Size

Patients treated in E.R.									
	Item in data array	1	2	3	4	5	6	7	8
	Number of patients	86	52	49	43	35	31	30	11

Median = 39

$$\text{Positioning Point} = \frac{n+1}{2} = \frac{8+1}{2} = 4.5$$

$$\text{Median} = \frac{43+35}{2} = 39$$

Mode

- Value that occurs most frequently

Ungrouped Data

Delivery trips per day in one 20-day period	Trips Arrayed in Ascending Order									
	0	0	1	1	2	2	4	4	5	5
	6	6	7	7	8	12	***15***	***15***	***15***	19

Mode (15)

Mode: More Examples

- **No Mode**

Raw Data: 10.3 4.9 8.9 11.7 6.3 7.7

- **One Mode**

Raw Data: 6.3 **4.9** 8.9 6.3 **4.9** **4.9**

- **More Than 1 Mode**

Raw Data: 21 **28** **28** 41 **43** **43**

Comparing Measures of Central Tendency

Frequency Distribution

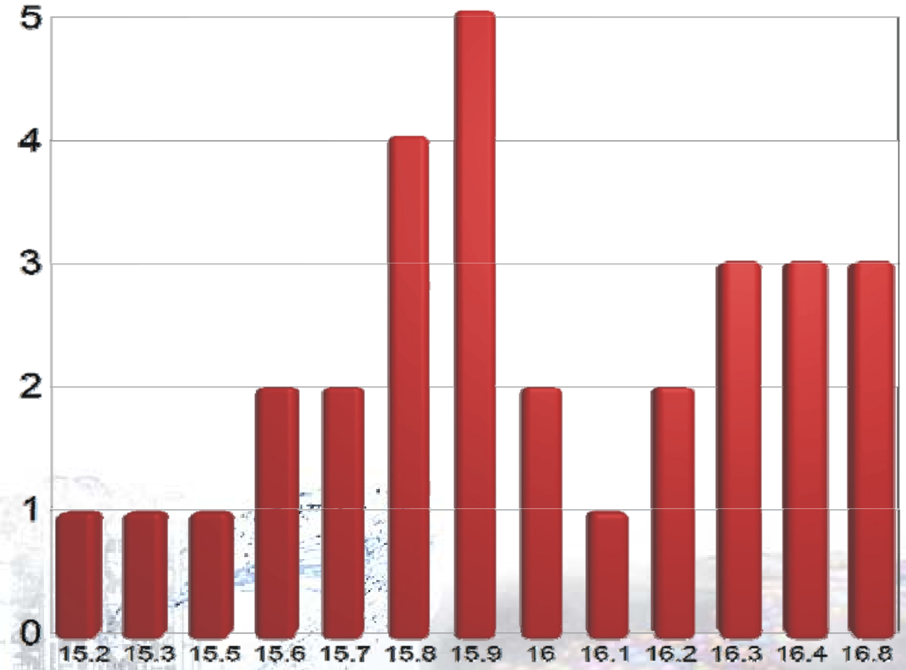
Carpet Loom Example

Class	Frequency
15.2	1
15.3	1
15.5	1
15.6	2
15.7	2
15.8	4
15.9	5
16.0	2
16.1	1
16.2	2
16.3	3
16.4	3
16.8	3

Mean: 16.01

Median: 15.9

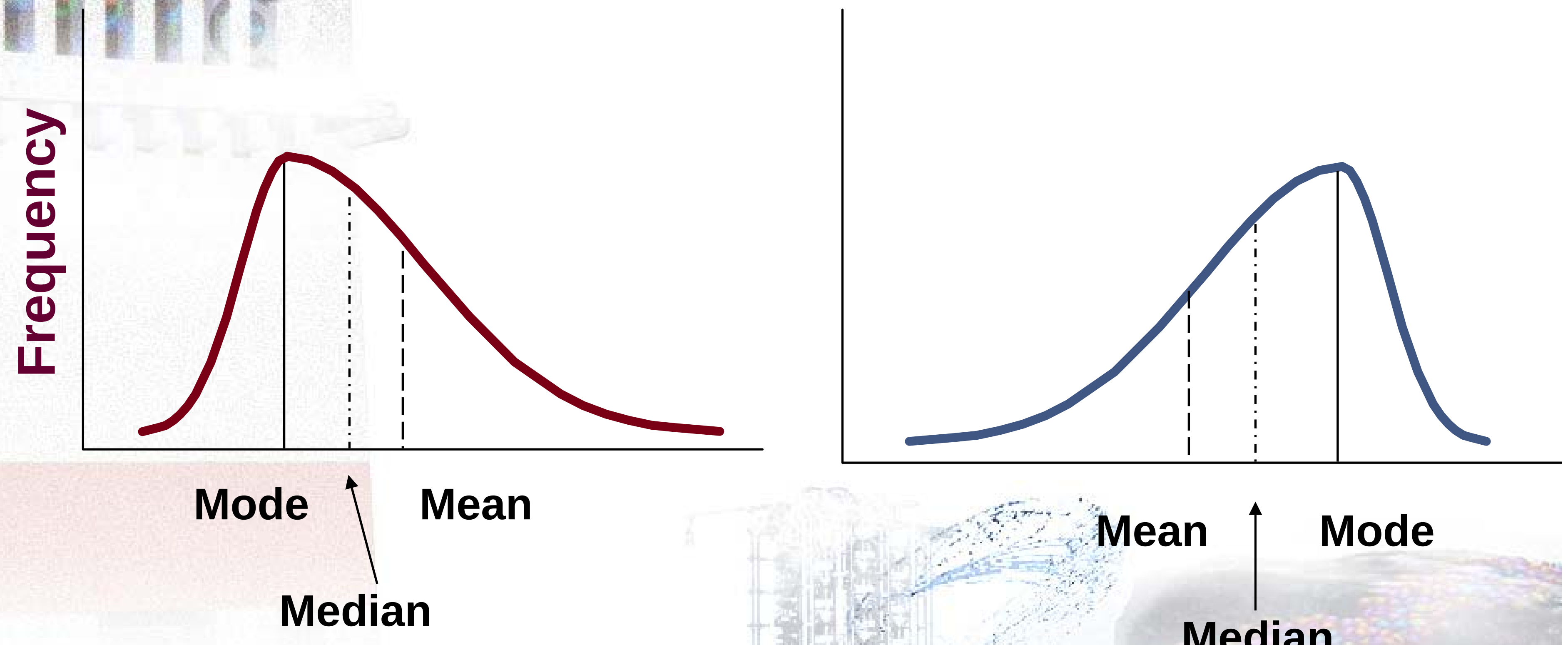
Mode: 15.9



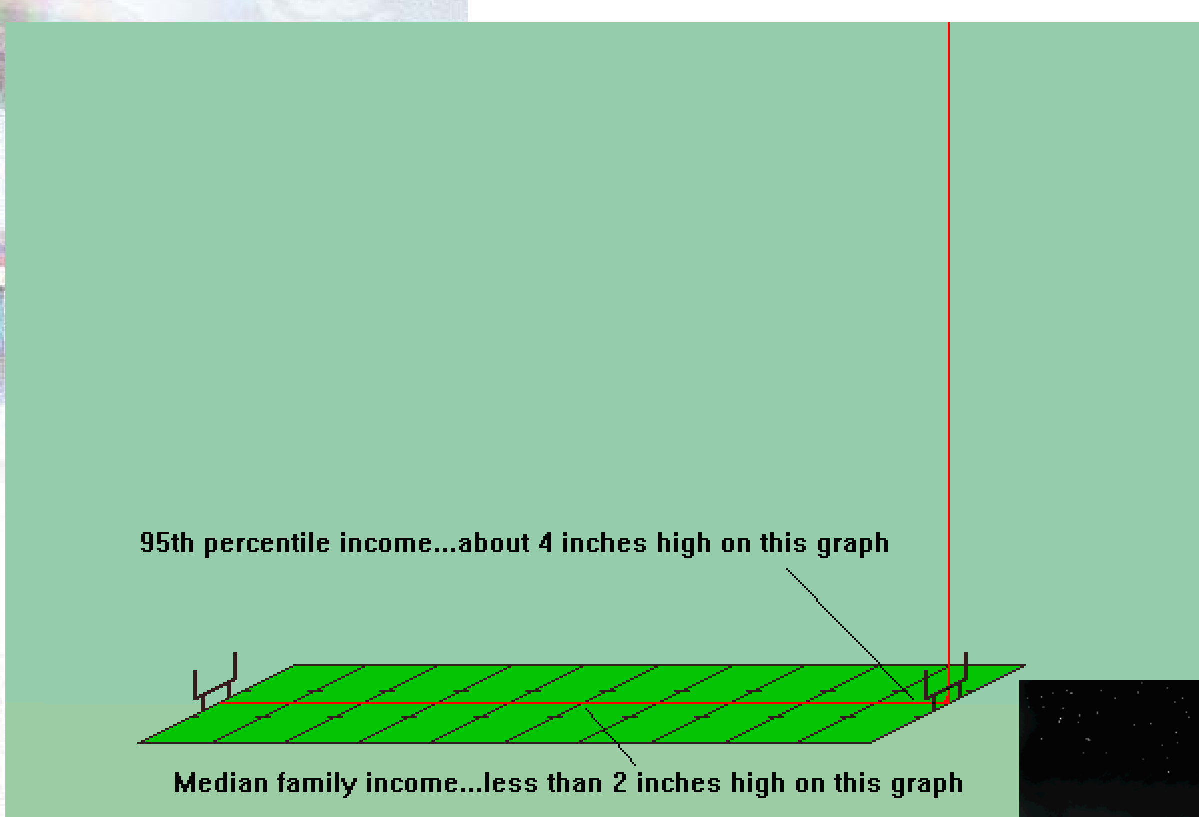
Notes on central Tendency

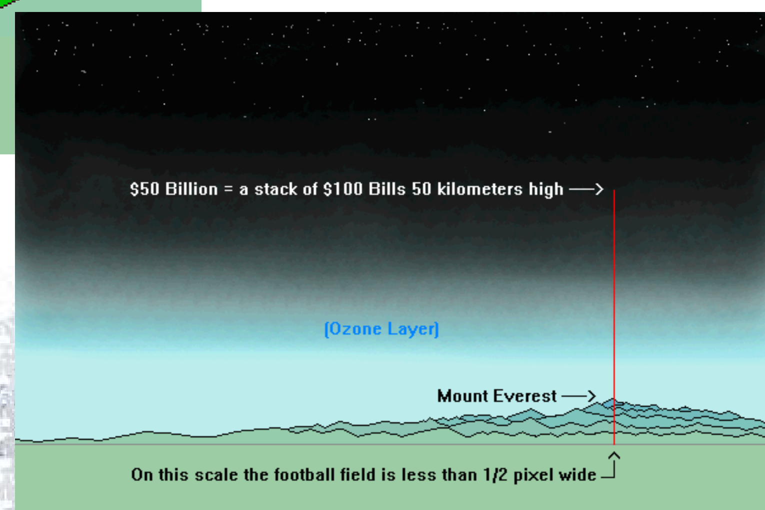
- ***Mean*** is the measure that varies the least from one sample to the next in a population
  - In most populations we encounter
- Simple formula for algebraic manipulations
- Uses all the information contained in the data
  - Mode depends on a single value
  - Median depends only in the middle position
- But in skewed frequency distributions we may want to downplay certain values
  - Salary (K$): 10, 12, 13, 13.5, 14, 14, 14.5, 15, 16, 16, 60
  - Mean: 18K; Median: 14K

Comparing the Mean, Median, and Mode



US income distribution

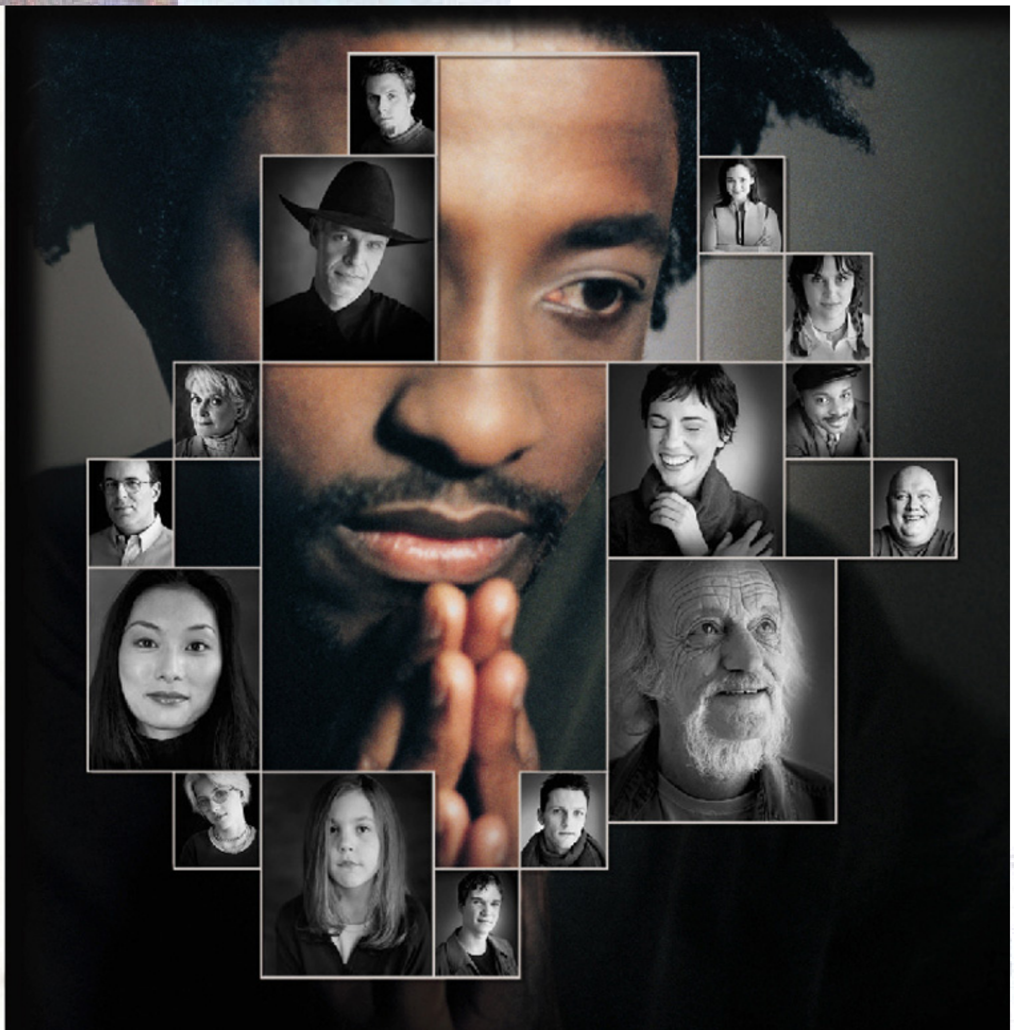




Central Tendency Measures

Measure	Equation	Description
Mean	$\Sigma x \;/\; n$	Balance Point
Median	$\frac{(n+1)}{2}$ *th* item in array	Middle value in ordered array
Mode	none	Most frequent

Measuring Dispersion



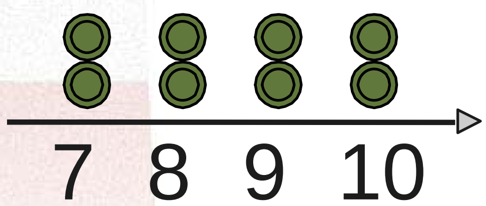
The number that conveys an idea of how much ***spread*** or ***variability*** exists among the data values

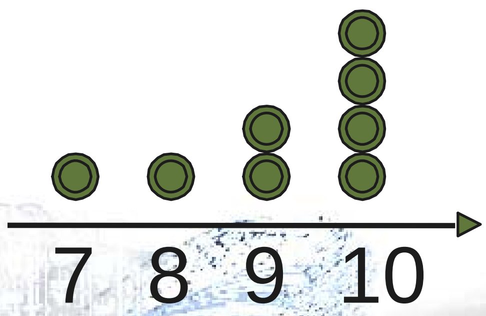
Range

- Difference between largest & smallest observations

$$\text{Range} = X_{largest} - X_{smallest}$$

- Ignores how data are distributed





Variance & Standard Deviation

- Most commonly used measures
- Consider how data are distributed
- Show variation about mean ($\bar{X}$ or μ)
  - Deviation from mean

	Population	Sample
Variance	σ^2	s^2
Standard Deviation	σ	s

Population Variance Formula

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

$$= \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}$$

Where:

N **- Population size**

X ***- item or observation***

μ ***- Population mean***

Σ ***- sum of the values***

Population Standard Deviation Formula

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

$$= \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

$$= \sqrt{\frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_N - \mu)^2}{N}}$$

Sample Variance Formula

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

$$= \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n - 1}$$

n - 1 in denominator!
(Use ***N*** if **Population** Variance)

Where:
***n* - sample size**
x - item or observation
***$\bar{x}$* - sample mean**
$\sum$ - sum of the values

Computation of Variance and Standard Deviation: Ungrouped Data

Given a sample consisting of 12 annual Blue Cross-Blue Shield payments to Cumberland Hospital, compute the variance and standard deviation.

Payments ($000):

863	903	957
1,041	1,138	1,204
1,354	1,624	1,698
1,745	1,802	1,883

Solution

$$s^2 = \frac{\sum (x - \overline{x})^2}{n - 1}$$

Step 1: Compute the Sample Mean

$$\overline{x} = \frac{\sum x}{n}$$

$$= 1{,}351$$

Observations (x) (1)
863
903
957
1,041
1,138
1,204
1,354
1,624
1,698
1,745
1,802
1,883
16,212

Solution

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

Step 2: Compute the sum of $(x - \bar{x})^2$

Observation (x) (1)	Mean $(\bar{x})$ (2)	$x - \bar{x}$ (1)-(2)	$(x - \bar{x})^2$ $[(1)-(2)]^2$
863	1,351	- 488	238,144
903	1,351	- 448	200,704
957	1,351	- 394	155,236
1,041	1,351	- 310	96,100
1,138	1,351	- 213	45,369
1,204	1,351	- 147	21,609
1,354	1,351	3	9
1,624	1,351	273	74,529
1,698	1,351	347	120,409
1,745	1,351	394	155,236
1,802	1,351	451	203,401
1,883	1,351	532	283,024
			1,593,770

Solution

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

$$= \frac{1,593,770}{(12 - 1)}$$

$$= 144,888$$

$$s = \sqrt{144,888}$$

= 380.64 or $380,640

Observation (x) (1)	Mean ($\bar{x}$) (2)	$x - \bar{x}$ (1)-(2)	$(x - \bar{x})^2$ $[(1)-(2)]^2$
863	1,351	- 488	238,144
903	1,351	- 448	200,704
957	1,351	- 394	155,236
1,041	1,351	- 310	96,100
1,138	1,351	- 213	45,369
1,204	1,351	- 147	21,609
1,354	1,351	3	9
1,624	1,351	273	74,529
1,698	1,351	347	120,409
1,745	1,351	394	155,236
1,802	1,351	451	203,401
1,883	1,351	532	283,024
			1,593,770

Example: Sample Dispersion

Frequency Distribution

Class	Frequency	
15.2	1	$(15.2 - 16.01)^2 = 0.6561$
15.3	1	$(15.3 - 16.01)^2 = 0.5041$
15.5	1	$(15.5 - 16.01)^2 = 0.2601$
15.6	2	$2 \times (15.6 - 16.01)^2 = 0.3362$
15.7	2	$2 \times (15.7 - 16.01)^2 = 0.1922$
15.8	4	$4 \times (15.8 - 16.01)^2 = 0.1764$
15.9	5	
16.0	2	• • •
16.1	1	
16.2	2	
16.3	3	
16.4	3	
16.8	3	
24	1	

Carpet Loom Example

Mean: 16.01

8.8

Range = 16.8-15.2 = 1.6

Standard Deviation = $\sqrt{0.15636} = 0.39542$

1.4651

Variance 4.847/(30+1) = 0.15636

2.1465

Uses of Standard Deviation

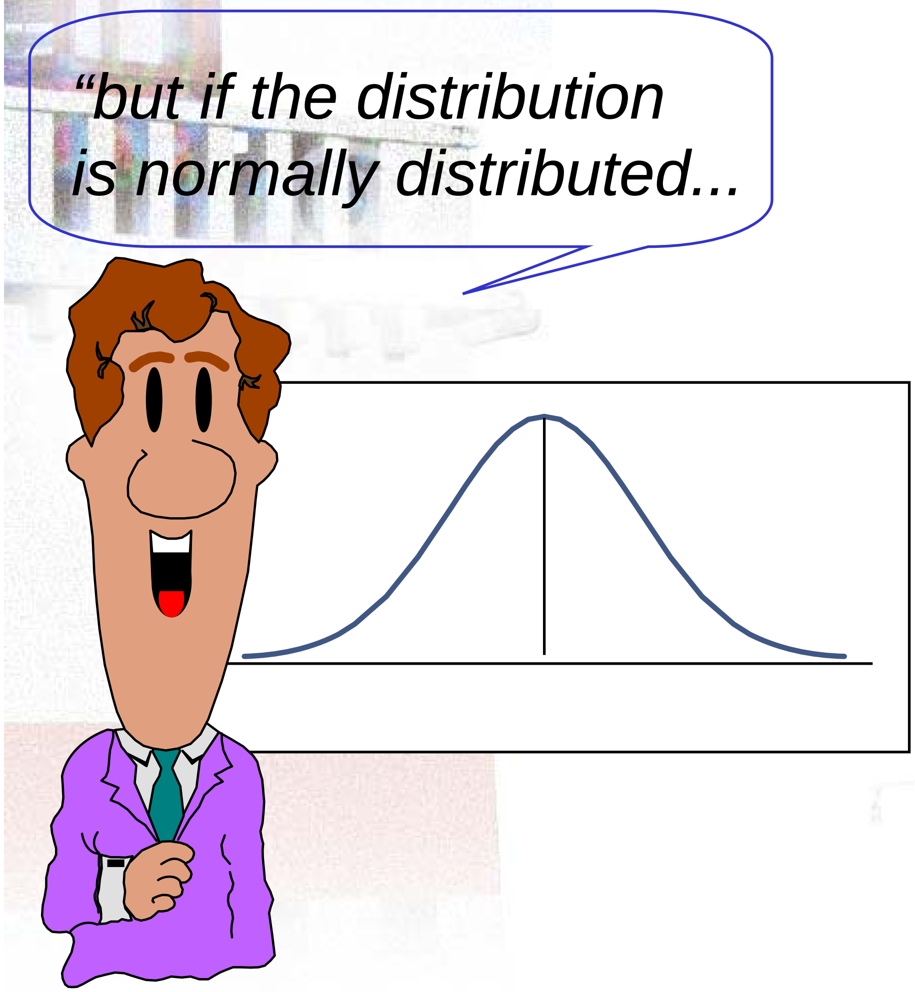
- Aside from measure of dispersion…
  - Determines where values of frequency distribution are in relation to mean ("standard scores")
- Measures percentage of items within specific ranges
  - Chebyshev's Theorem
  - Normal distribution

Chebyshev's Theorem

"Regardless of original distribution...

- At least 75% of all observations fall within two standard deviations of the mean
- At least 89% fall within three standard deviations
- At least $1-1/k^2$ fall within k standard deviations

Empirical Rule (Normal Distribution)



- 68% Within one Standard Deviation
- 95% Within two Standard Deviations
- 99% Within three Standard Deviations

Example: Sample Dispersion

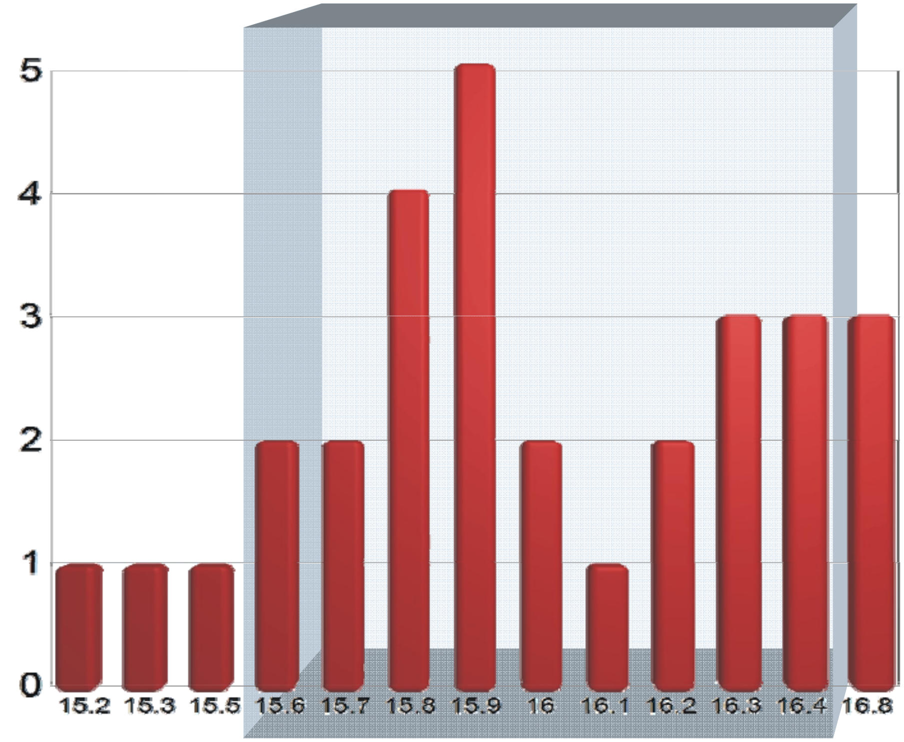
Frequency Distribution

Carpet Loom Example

Class	Frequency
15.2	1
15.3	1
15.5	1
15.6	2
15.7	2
15.8	4
15.9	5
16.0	2
16.1	1
16.2	2
16.3	3
16.4	3
16.8	3

1 sd (15.6 – 16.4)

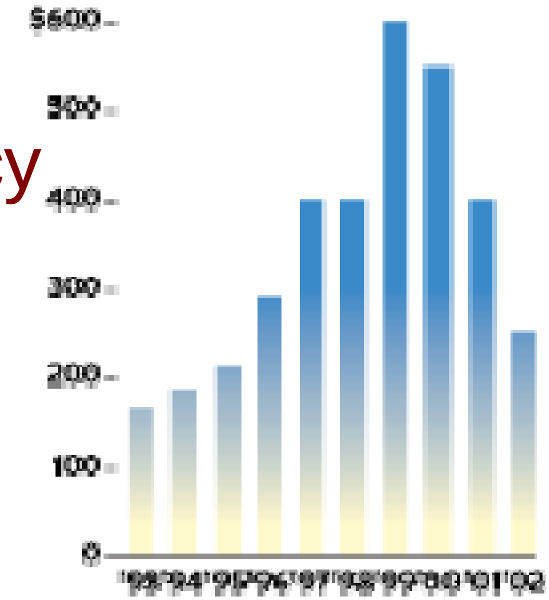
80 %



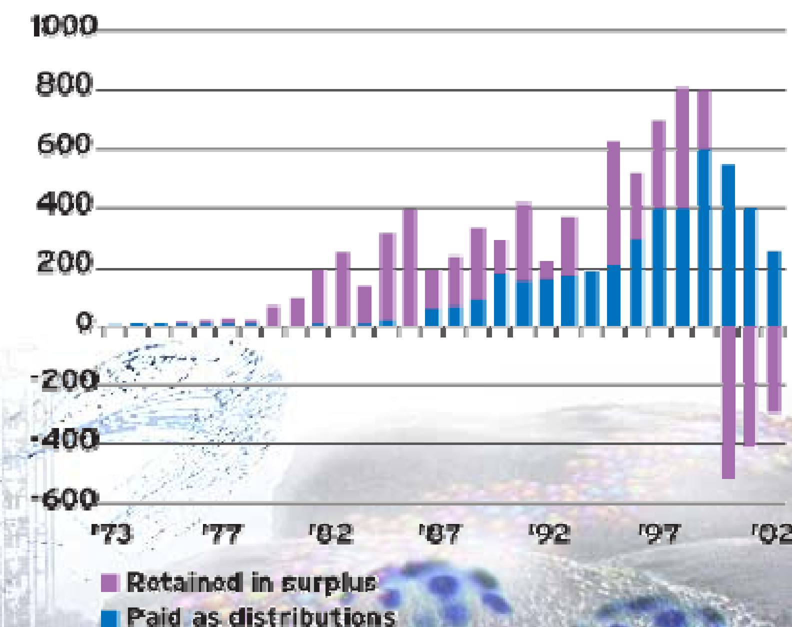
Mean: 16.01

Standard Deviation = 0.39542

Summarizing Data



- Measures of Central Tendency
  - Mean
  - Median
  - Mode
- Measures of Dispersion
  - Range
  - Standard Deviation
  - Variance
- Measures of Shape
  - Skewness
  - Kurtosis



Frequency Analysis and Cryptography

- Cryptography
  - Derived from the Greek word *Kryptos*: hidden
- See Simon Singh's The Code Book CD-ROM
  - The Vigenère Code

Next Class!

- Topics
  - More Inductive Reasoning Modeling
    - Linear Regression
- Readings for Next week
  - @ *infoport*
  - From course package
    - Norman, G.R. and D.L. Streinrt [2000]. *Biostatistics: The Bare Essentials*.
      - Chapters 1-3 (pages 105-130)
      - OPTIONAL: Chapter 4 (pages 131-136)
      - Chapter 13 (pages 147-155)
- Lab 8
  - Data analysis with Excel (linear regression)