

Data Analysis and Visualisation

Lecture 2

Descriptive Statistics

Descriptive Statistics

Frequency distributions - graphical view of data

Drawback:

Not enough or right kind of information

“What is the average GRE score of this class?”

“How much money does the average football player make?”

- To do this we need to look at
 - Measures of central tendency – Mean, Median, Mode
 - Measures of variability- Range, Variance(SD^2), Standard Deviation(SD)

Measures of Central tendency

Measure of central tendency- **Mean**

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

where X is the set of scores and N is the number of scores

Mean is denoted by $\bar{X}$

Example



Microsoft Excel
17-2003 Worksheet

Maths Score

14

19

13

9

13

$$\bar{X} = \frac{\sum X}{N} = \frac{68}{5} = 13.6$$

Reading Score

25

37

26

20

19

$$\bar{Y} = \frac{\sum Y}{N} = \frac{127}{5} = 25.4$$

Measures of central tendency

The ***Median*** – Middle value(divides the frequency distribution in half)

Example 1: odd set of scores

113, 95, 94, 133, 112

- Sorting : 94, 95, 112, 113, 133
- Middle value: 112

Example 2: even set of scores

113, 95, 94, 133, 112, 102

- Sorting: 94, 95, 102, 112, 113, 133
- Mean of the middle two values:
 $(102 + 112)/2 = 107$

Measures of Central Tendency

Mode - most frequently occurring score in a set

Example 1:

Given a set of scores

110, 105, 100, 100, 100, 100, 99, 98

- 100 appears 4 times

• Example 2:

Given a set of scores

110, 105, 105, 103, 100, 100, 99, 98

- 105 and 100 both appear twice-Bimodal
- Multimodal distributions are also possible

Exercise

- Determine the mean, median and mode(s) of the midterm examination dataset (80 students)

Midterm Examination Scores									
44	35	20	40	38	52	29	36	38	38
38	38	41	35	42	50	31	43	30	37
32	47	43	41	47	32	38	29	41	26
41	51	48	49	37	26	34	48	23	29
38	47	41	33	39	48	38	20	48	34
29	44	29	33	35	50	41	38	35	35
32	26	24	38	38	56	56	48	41	29
26	26	38	37	44	24	44	47	32	41

Selecting a measure of central tendency

Symmetrical unimodal distribution

Mean = = Median == Mode

Mean is the 'balance point' of a distribution

Skewed Dataset → Find Median and not Mean

- A skewed dataset - unbalanced Dataset

- Consider a small University Football team,
5 persons - four professors, and a football coach.

Annual income

Professors' Salary	Coach's Salary
Professor 1 \$60,000	\$1,000,000
Professor 2 \$55,000	
Professor 3 \$55,000	
Professor 4 \$50,000	

Average Salary = \$ 244,000

Is it Mean or Median or Mode?

The Normal curve

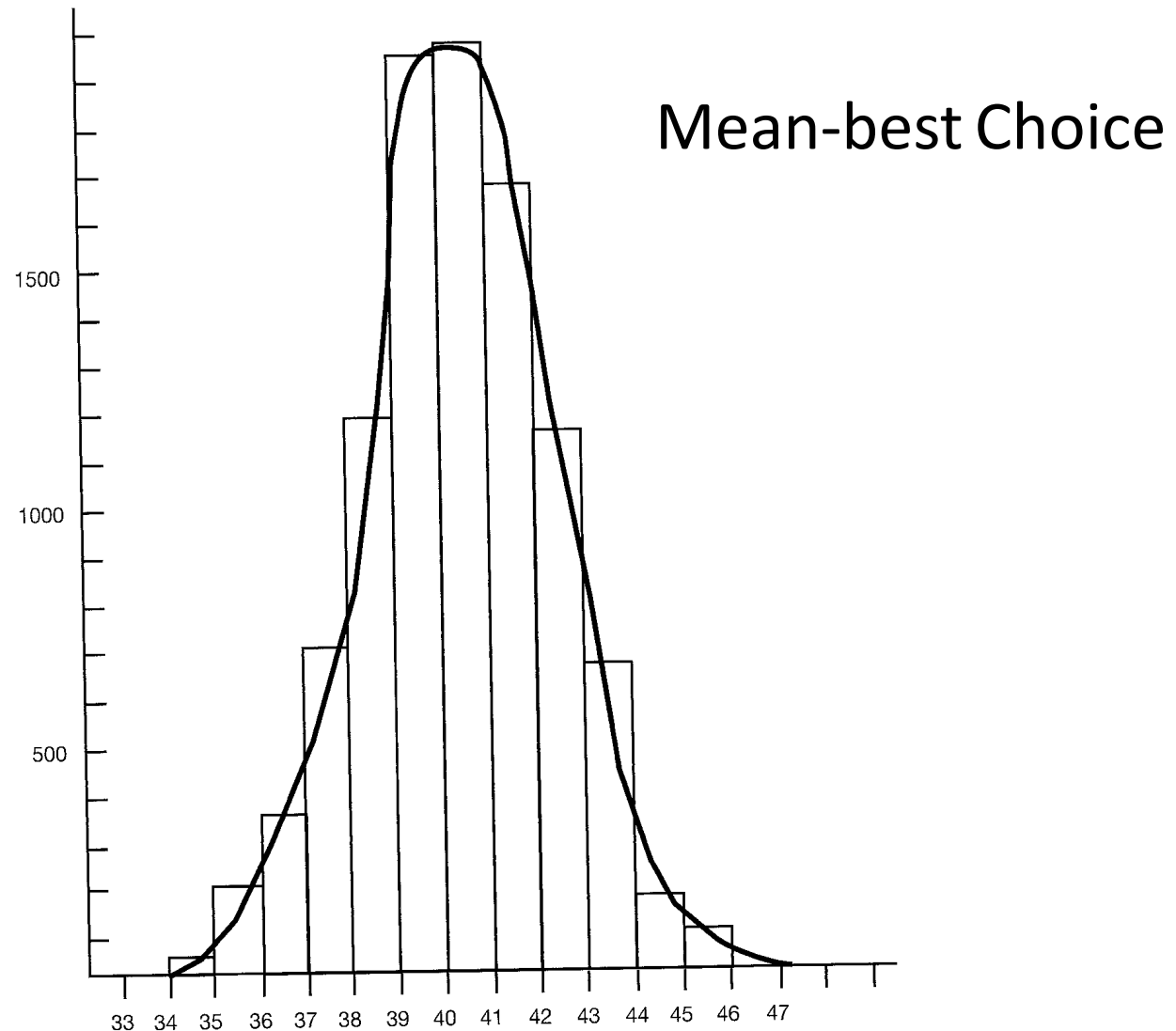


FIGURE 4-3 Chest sizes of 5,738 soldiers.

Skewed distributions

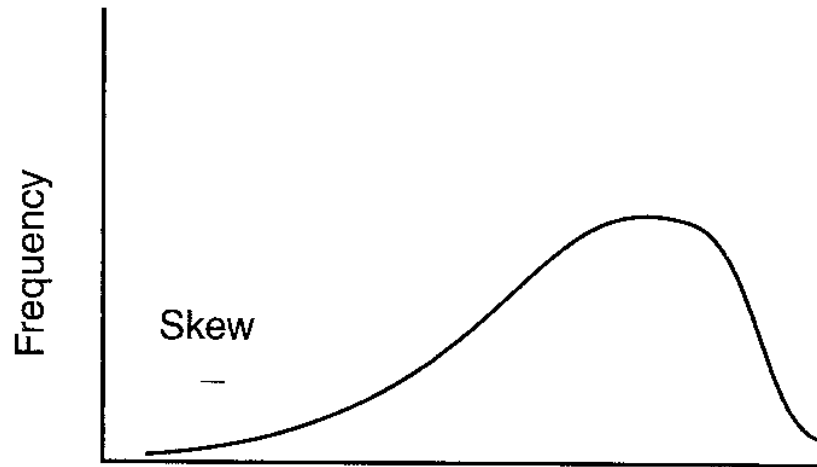


FIGURE 4-4 A negatively skewed distribution.

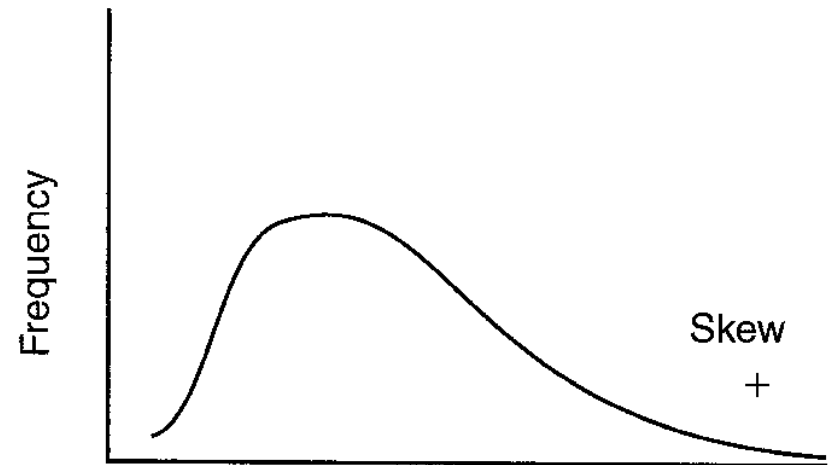


FIGURE 4-5 A positively skewed distribution.

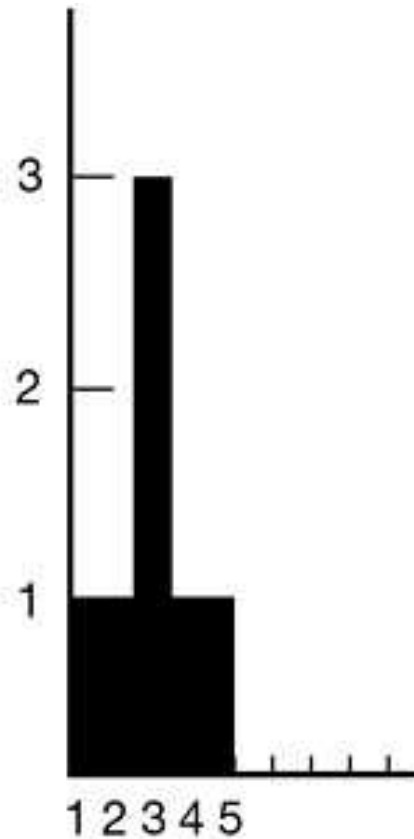
Variation with Skewness

Scores (X)

1
2
3
3
3
4
5

21

$M_X = 3$
 $Md = 3$
Mode = 3

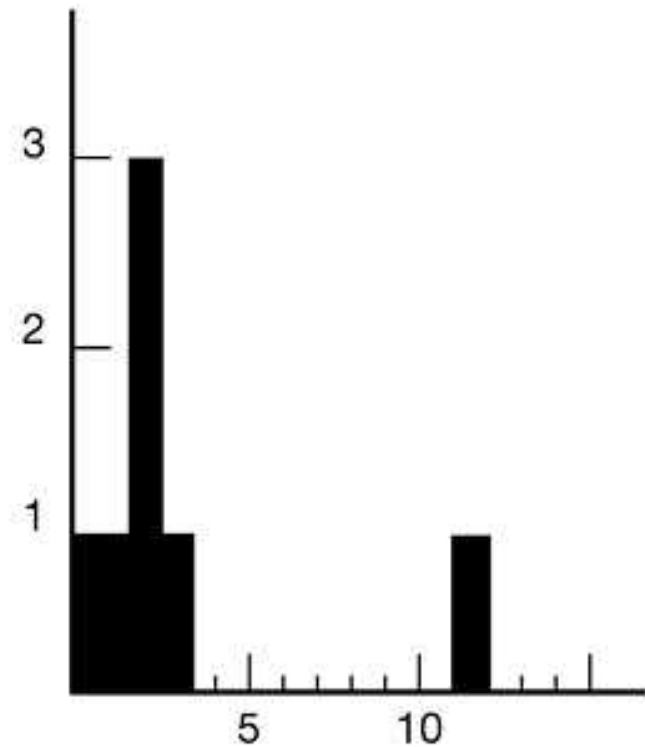


Scores (Y)

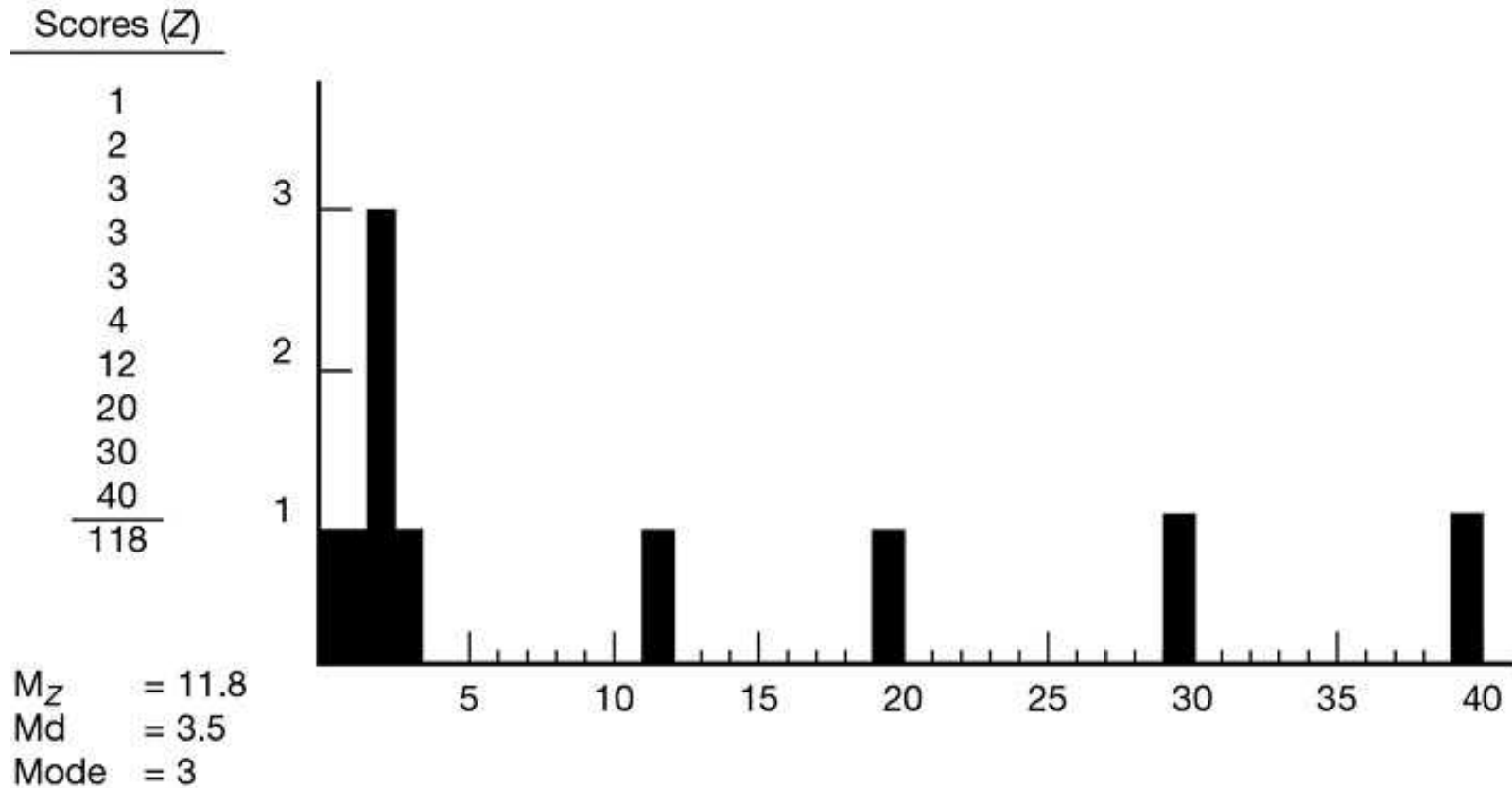
1
2
3
3
3
4

12
28

$M_Y = 4$
 $Md = 3$
Mode = 3



Variation with Skewness



Measures of variability

- Mean, Median & Mode → single score for entire set
- Consider the final exam scores earned by students in each of two statistics classes:

Classroom X Exam Score

160

130

100

70

40

$$\bar{X} = \frac{\sum X}{N} = \frac{500}{5} = 100$$

Classroom Y Exam Score

102

101

100

99

98

$$\bar{Y} = \frac{\sum Y}{N} = \frac{500}{5} = 100$$

Difference in Variability!!!

variability = Differences in set of scores.

Measures of Variability

- Range
- Variance (SD^2)
- Standard Deviation (SD)

Range

$$\text{Range} = H - L$$

Where H- highest score & L – Lowest score

Classroom X Exam Score

160

130

100

70

40

$$\begin{aligned}\text{Range} &= 160 - 40 \\ &= 120\end{aligned}$$

Classroom Y Exam Score

102

101

100

99

98

$$\begin{aligned}\text{Range} &= 102 - 98 \\ &= 4\end{aligned}$$

- Skewness not considered

Variance

$$SD^2 = \frac{\sum (X - \bar{X})^2}{N - 1}$$

- $(X - \bar{X})$ is the *deviation score*

Variance

To calculate the variance first calculate the deviations χ then square and sum them, then divide by N-1

Scores (X)	$(X - \bar{X})$	χ	$(X - \bar{X})^2$
9	$9 - 5$	4	16
7	$7 - 5$	2	4
5	$5 - 5$	0	0
3	$3 - 5$	-2	4
1	$1 - 5$	-4	16
$\Sigma X = 25$	$\Sigma(X - \bar{X}) = 0$	$\Sigma \chi = 0$	$\Sigma(X - \bar{X})^2 = 40$

$$\begin{aligned} SD_X^2 &= \frac{\Sigma(X - \bar{X})^2}{N - 1} \\ &= \frac{40}{5 - 1} = 10 \end{aligned}$$

Variance

- An alternative formula exists that eliminates rounding errors and is easier to compute

$$SD^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1}$$

Scores (X)	Scores Squared (X) ²
9	81
7	49
5	25
3	9
1	1
$\sum X = 25$; $(\sum X)^2 = 625$	$\sum X^2 = 165$

$$SD^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N - 1} = \frac{165 - \frac{625}{5}}{4} = \frac{165 - 125}{4} = \frac{40}{4} = 10$$

Standard Deviation (S.D.)

Scores (X)	
9	Mean $\bar{X} = 5$
7	Variance $SD^2 = 10$
5	Standard Deviation $SD = \sqrt{10}$
3	$= 3.16$
1	

Exercises

A school psychologist is interested in knowing whether students in the gifted and talented program study more hours per day, on average, than students who are in regular education. What measure of central tendency would be best for the school psychologist to use? Explain.

Exercises

A school psychologist is interested in knowing whether students in the gifted and talented program study more hours per day, on average, than students who are in regular education. What measure of central tendency would be best for the school psychologist to use? Explain.

Solution:

Mean is the measure that best reflects the central tendency of both groups.

A researcher gathers data and finds that her distribution of scores is skewed. She decides that she does not want her measure of central tendency to be unduly affected by the most extreme scores. Which measure will she use? Explain.

A researcher gathers data and finds that her distribution of scores is skewed. She decides that she does not want her measure of central tendency to be unduly affected by the most extreme scores. Which measure will she use? Explain.

Solution:

The median because it is a reasonably stable measure of central tendency that is not so responsive to extreme scores as the mean.

A researcher gathers IQ data for a large group of undergraduate students. The mean of his distribution of scores is 110, and the median is 130. What can he assume about the shape of his distribution?

A researcher gathers IQ data for a large group of undergraduate students. The mean of his distribution of scores is 110, and the median is 130. What can he assume about the shape of his distribution?

Solution:

Negatively skewed; because the mean is lower than the median, low scores must be pulling it down.

Exercise

- Find the mean $\bar{X}$, range R, variance SD^2 and standard deviations for each of the following groups of scores. Carry out all calculations to three decimal places and round them correct to two places

$$SD^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N-1} \quad \text{or} \quad SD^2 = \frac{\sum (X - \bar{X})^2}{N-1}$$

Classroom X Exam Score

160

130

100

70

40

Classroom Y Exam Score

102

101

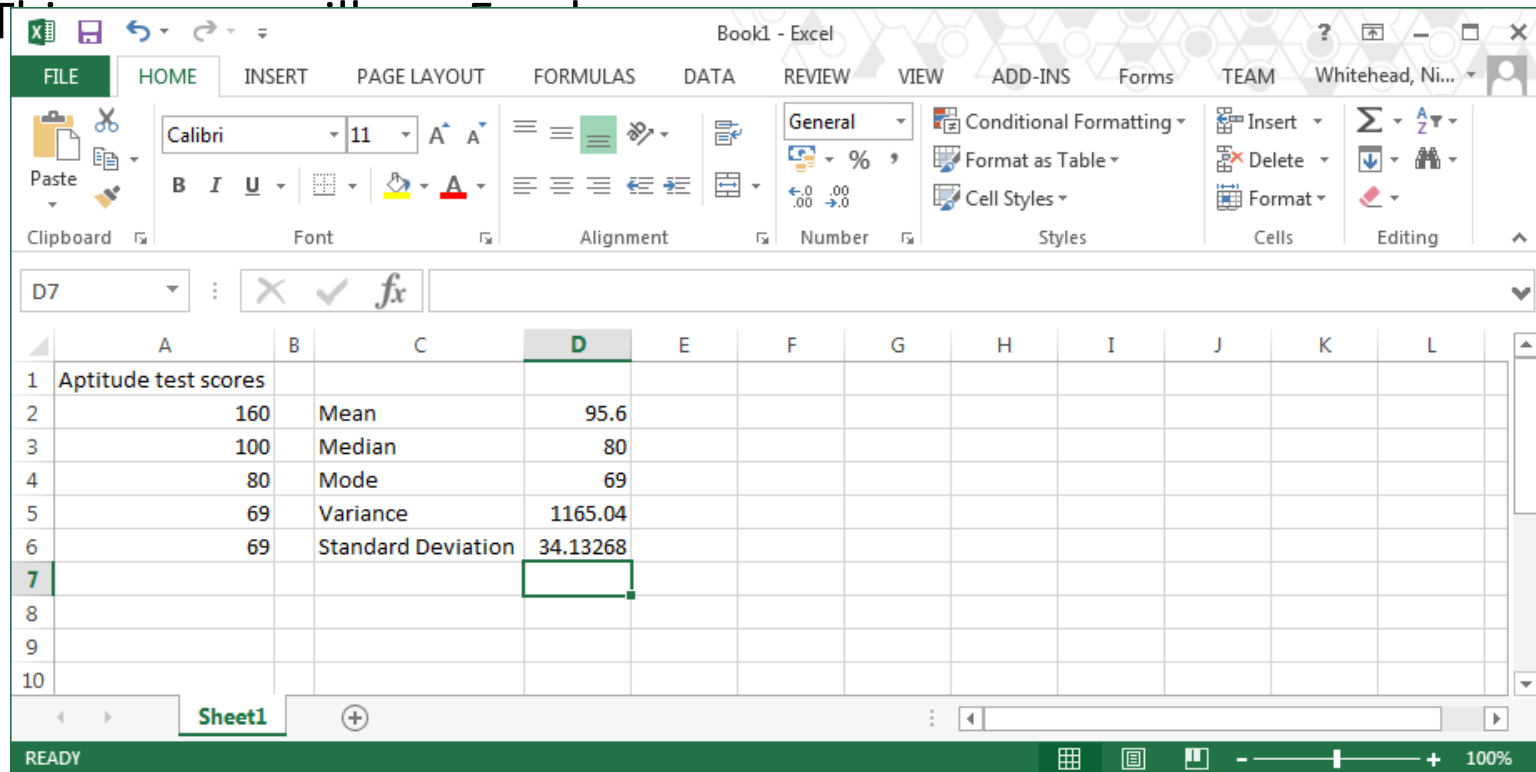
100

99

98

Using Excel

- Calculating the means, variances and standard deviations can be done by hand but is better done with a spreadsheet or a statistical package
- T



Using Excel

- The five main Excel functions are:
 - Mean: AVERAGE(range)
 - Median: MEDIAN(range)
 - Mode: MODE(range)
 - If there is no mode then it will return the value #N/A
 - Variance: VAR.S(range)
 - This will ignore any non-numerical cells in the range
 - Standard Deviation: STDEV.S(range)

	A	B	C	D	E
1	Aptitude test scores				
2	160	Mean		=AVERAGE(A2:A6)	
3	100	Median		=MEDIAN(A2:A6)	
4	80	Mode		=MODE(A2:A6)	
5	69	Variance		=VAR.S(A2:A6)	
6	69	Standard Deviation		=STDEV.S(A2:A6)	

Exercise

- Download the file Midterms.xlsx and calculate the mean, median, mode, variance and standard deviation of the data therein

Summary

- All of today's material looked at descriptive statistics of univariate data
 - Measures of central tendency
 - Selecting a measure of central tendency
 - Measures of variability
 - Using Excel for statistics
- Next session we'll look at the Normal curve and how this can be used to tell us more about a dataset

Coming soon

- **Next session:** The Normal curve
- **Homework:** Read chapter 5 of *Statistics for the Terrified*.