

## Measures of Central Tendency: Ungrouped Data

- Measures of central tendency yield information about “particular places or locations in a group of numbers.”
- Common Measures of Location
  - Mode
  - Median
  - Mean
  - Percentiles
  - Quartiles

Stat 1300, UCLA, Ivo Dinov

1

## Mode

- The most frequently occurring value in a data set
- Applicable to all levels of data measurement (nominal, ordinal, interval, and ratio)
- Bimodal -- Data sets that have two modes
- Multimodal -- Data sets that contain more than two modes

Stat 1300, UCLA, Ivo Dinov

2

## Mode -- Example

- The mode is 44.
- There are more 44s than any other value.

35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

Stat 1300, UCLA, Ivo Dinov

3

## Median

- Middle value in an ordered array of numbers.
- Applicable for ordinal, interval, and ratio data
- Not applicable for nominal data
- Unaffected by extremely large and extremely small values.

Stat 1300, UCLA, Ivo Dinov

4

## Median: Computational Procedure

- First Procedure
  - Arrange the observations in an ordered array.
  - If there is an odd number of terms, the median is the middle term of the ordered array.
  - If there is an even number of terms, the median is the average of the middle two terms.
- Second Procedure
  - The median's position in an ordered array is given by  $(n+1)/2$ .

Stat 1300, UCLA, Ivo Dinov

5

## Median: Example with an Odd Number of Terms

Ordered Array

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21 22

- There are 17 terms in the ordered array.
- Position of median =  $(n+1)/2 = (17+1)/2 = 9$
- The median is the 9th term, 15.
- If the 22 is replaced by 100, the median is 15.
- If the 3 is replaced by -103, the median is 15.

Stat 1300, UCLA, Ivo Dinov

6

## Median: Example with an Even Number of Terms

Ordered Array

3 4 5 7 8 9 11 14 15 16 16 17 19 19 20 21

- There are 16 terms in the ordered array.
- Position of median =  $(n+1)/2 = (16+1)/2 = 8.5$
- The median is between the 8th and 9th terms, 14.5.
- If the 21 is replaced by 100, the median is 14.5.
- If the 3 is replaced by -88, the median is 14.5.

Stat 1300, UCLA, Ivo Dinov

7

## Arithmetic Mean

- Commonly called 'the mean'
- is the average of a group of numbers
- Applicable for interval and ratio data
- Not applicable for nominal or ordinal data
- Affected by each value in the data set, including extreme values
- Computed by summing all values in the data set and dividing the sum by the number of values in the data set

Stat 1300, UCLA, Ivo Dinov

8

## Population Mean

$$\begin{aligned}\mu &= \frac{\sum X}{N} = \frac{X_1 + X_2 + X_3 + \dots + X_N}{N} \\ &= \frac{24 + 13 + 19 + 26 + 11}{5} \\ &= \frac{93}{5} \\ &= 18.6\end{aligned}$$

Stat 1300, UCLA, Ivo Dinov

9

## Sample Mean

$$\begin{aligned}\bar{X} &= \frac{\sum X}{n} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n} \\ &= \frac{57 + 86 + 42 + 38 + 90 + 66}{6} \\ &= \frac{379}{6} \\ &= 63.167\end{aligned}$$

Stat 1300, UCLA, Ivo Dinov

10

## Percentiles

- Measures of central tendency that divide a group of data into 100 parts
- At least  $n\%$  of the data lie below the  $n^{\text{th}}$  percentile, and at most  $(100 - n)\%$  of the data lie above the  $n^{\text{th}}$  percentile
- Example: 90th percentile indicates that at least 90% of the data lie below it, and at most 10% of the data lie above it
- The median and the 50th percentile are the same.
- Applicable for ordinal, interval, and ratio data
- Not applicable for nominal data

Stat 1300, UCLA, Ivo Dinov

11

## Percentiles: Computational Procedure

- Organize the data into an ascending ordered array.
- Calculate the percentile location:

$$i = \frac{P}{100}(n)$$

- Determine the percentile's location and its value.
- If  $i$  is a whole number, the percentile is the average of the values at the  $i$  and  $(i+1)$  positions.
- If  $i$  is not a whole number, the percentile is at the  $(i+1)$  position in the ordered array.

Stat 1300, UCLA, Ivo Dinov

12

### Percentiles: Example

- Raw Data: 14, 12, 19, 23, 5, 13, 28, 17
- Ordered Array: 5, 12, 13, 14, 17, 19, 23, 28
- Location of 30th percentile:  $i = \frac{30}{100}(8) = 2.4$
- The location index,  $i$ , is not a whole number;  $i+1 = 2.4+1=3.4$ ; the whole number portion is 3; the 30th percentile is at the 3rd location of the array; the 30th percentile is 13.

Stat 130D, UCLA, Ivo Dinov

13

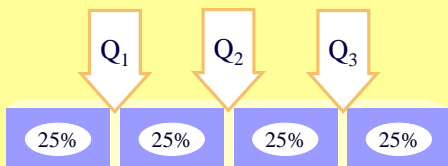
### Quartiles

- Measures of central tendency that divide a group of data into four subgroups
- $Q_1$ : 25% of the data set is below the first quartile
- $Q_2$ : 50% of the data set is below the second quartile
- $Q_3$ : 75% of the data set is below the third quartile
- $Q_1$  is equal to the 25th percentile
- $Q_2$  is located at 50th percentile and equals the median
- $Q_3$  is equal to the 75th percentile
- Quartile values are not necessarily members of the data set

Stat 130D, UCLA, Ivo Dinov

14

### Quartiles



Stat 130D, UCLA, Ivo Dinov

15

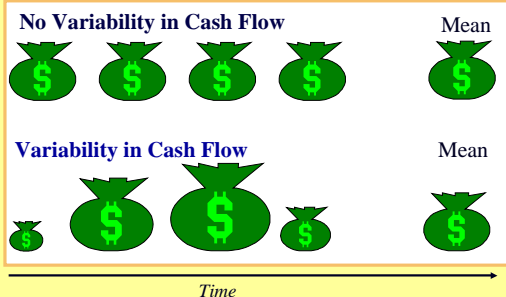
### Quartiles: Example

- Ordered array: 106, 109, 114, 116, 121, 122, 125, 129
- $Q_1$ :  $i = \frac{25}{100}(8) = 2$       $Q_1 = \frac{109+114}{2} = 111.5$
- $Q_2$ :  $i = \frac{50}{100}(8) = 4$       $Q_2 = \frac{116+121}{2} = 118.5$
- $Q_3$ :  $i = \frac{75}{100}(8) = 6$       $Q_3 = \frac{122+125}{2} = 123.5$

Stat 130D, UCLA, Ivo Dinov

16

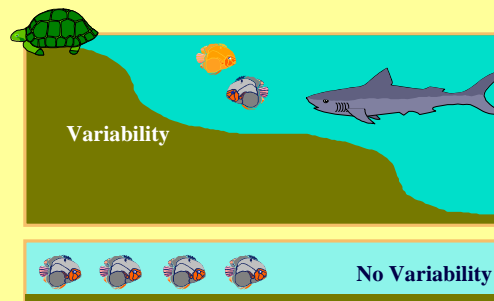
### Variability



Stat 130D, UCLA, Ivo Dinov

17

### Variability



Stat 130D, UCLA, Ivo Dinov

18

## Measures of Variability: Ungrouped Data

- Measures of variability describe the spread or the dispersion of a set of data.
- Common Measures of Variability
  - Range
  - Interquartile Range
  - Mean Absolute Deviation
  - Variance
  - Standard Deviation
  - Z scores
  - Coefficient of Variation

Stat 1300, UCLA, Ivo Dinov

19

## Range

- The difference between the largest and the smallest values in a set of data
- Simple to compute
- Ignores all data points except two extremes
- Example:  
Range  
Largest - Smallest  
 $48 - 35 = 13$

35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

Stat 1300, UCLA, Ivo Dinov

20

## Interquartile Range

- Range of values between the first and third quartiles
- Range of the “middle half”
- Less influenced by extremes

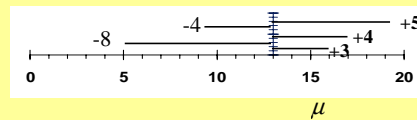
$$\text{Interquartile Range} = Q_3 - Q_1$$

Stat 1300, UCLA, Ivo Dinov

21

## Deviation from the Mean

- Data set: 5, 9, 16, 17, 18
- Mean:  $\mu = \frac{\sum X}{N} = \frac{65}{5} = 13$
- Deviations from the mean: -8, -4, 3, 4, 5



Stat 1300, UCLA, Ivo Dinov

22

## Mean Absolute Deviation

- Average of the absolute deviations from the mean

$X$	$X - \mu$	$ X - \mu $
5	-8	+8
9	-4	+4
16	+3	+3
17	+4	+4
18	+5	+5
	0	24

$$\begin{aligned}
 M.A.D. &= \frac{\sum |X - \mu|}{N} \\
 &= \frac{24}{5} \\
 &= 4.8
 \end{aligned}$$

Stat 1300, UCLA, Ivo Dinov

23

## Population Variance

- Average of the squared deviations from the arithmetic mean

$X$	$X - \mu$	$(X - \mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\begin{aligned}
 \sigma^2 &= \frac{\sum (X - \mu)^2}{N} \\
 &= \frac{130}{5} \\
 &= 26.0
 \end{aligned}$$

Stat 1300, UCLA, Ivo Dinov

24

## Population Standard Deviation

- Square root of the variance

$X$	$X - \mu$	$(X - \mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

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

$$= \frac{130}{5}$$

$$= 26.0$$

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

$$= \sqrt{26.0}$$

$$= 5.1$$

Stat 1300, UCLA, Ivo Dinov

25

## Sample Variance

- Average of the squared deviations from the arithmetic mean

$X$	$X - \bar{X}$	$(X - \bar{X})^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

$$S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

$$= \frac{663,866}{3}$$

$$= 221,288.67$$

Stat 1300, UCLA, Ivo Dinov

26

## Sample Standard Deviation

- Square root of the sample variance

$X$	$X - \bar{X}$	$(X - \bar{X})^2$
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

$$S^2 = \frac{\sum (X - \bar{X})^2}{n - 1}$$

$$= \frac{663,866}{3}$$

$$= 221,288.67$$

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

$$= \sqrt{221,288.67}$$

$$= 470.41$$

Stat 1300, UCLA, Ivo Dinov

27

## Uses of Standard Deviation

- Indicator of financial risk
- Quality Control
  - construction of quality control charts
  - process capability studies
- Comparing populations
  - household incomes in two cities
  - employee absenteeism at two companies

Stat 1300, UCLA, Ivo Dinov

28

## Standard Deviation as an Indicator of Financial Risk

Financial Security	Annualized Rate of Return	
	$\mu$	$\sigma$
A	15%	3%
B	15%	7%

Stat 1300, UCLA, Ivo Dinov

29

## Empirical Rule

- Data are normally distributed (or approximately normal)

Distance from the Mean	Percentage of Values Falling Within Distance
$\mu \pm 1\sigma$	68
$\mu \pm 2\sigma$	95
$\mu \pm 3\sigma$	99.7

Stat 1300, UCLA, Ivo Dinov

30

### Chebyshev's Theorem

- Applies to all distributions

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}$$

for  $k > 1$

Stat 130B, UCLA, Ivo Dinov

31

### Chebyshev's Theorem

- Applies to all distributions

Number of Standard Deviations	Distance from the Mean	Minimum Proportion of Values Falling Within Distance
$K = 2$	$\mu \pm 2\sigma$	$1 - 1/2^2 = 0.75$
$K = 3$	$\mu \pm 3\sigma$	$1 - 1/3^2 = 0.89$
$K = 4$	$\mu \pm 4\sigma$	$1 - 1/4^2 = 0.94$

Stat 130B, UCLA, Ivo Dinov

32

### Coefficient of Variation

- Ratio of the standard deviation to the mean, expressed as a percentage
- Measurement of relative dispersion

$$C.V. = \frac{\sigma}{\mu}(100)$$

Stat 130B, UCLA, Ivo Dinov

33

### Coefficient of Variation

$$\begin{aligned} \mu_1 &= 29 \\ \sigma_1 &= 4.6 \\ C.V._1 &= \frac{\sigma_1}{\mu_1}(100) \\ &= \frac{4.6}{29}(100) \\ &= 15.86 \end{aligned}$$

$$\begin{aligned} \mu_2 &= 84 \\ \sigma_2 &= 10 \\ C.V._2 &= \frac{\sigma_2}{\mu_2}(100) \\ &= \frac{10}{84}(100) \\ &= 11.90 \end{aligned}$$

Stat 130B, UCLA, Ivo Dinov

34

### Measures of Central Tendency and Variability: Grouped Data

- Measures of Central Tendency
  - Mean
  - Median
  - Mode
- Measures of Variability
  - Variance
  - Standard Deviation
  - Mean Absolute Deviation

Stat 130B, UCLA, Ivo Dinov

35

### Mean of Grouped Data

- Weighted average of class midpoints
- Class frequencies are the weights

$$\begin{aligned} \mu &= \frac{\sum fM}{\sum f} \\ &= \frac{\sum fM}{N} \\ &= \frac{f_1 M_1 + f_2 M_2 + f_3 M_3 + \dots + f_i M_i}{f_1 + f_2 + f_3 + \dots + f_i} \end{aligned}$$

Stat 130B, UCLA, Ivo Dinov

36

## Calculation of Grouped Mean

Class Interval	Frequency	Class Midpoint	fM
20-under 30	6	25	150
30-under 40	18	35	630
40-under 50	11	45	495
50-under 60	11	55	605
60-under 70	3	65	195
70-under 80	1	75	75
	50		2150

$$\mu = \frac{\sum fM}{\sum f} = \frac{2150}{50} = 43.0$$

Stat 1300, UCLA, Ivo Dinov

37

## Median of Grouped Data

$$\text{Median} = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

Where:

L = the lower limit of the median class

cf<sub>p</sub> = cumulative frequency of class preceding the median class

f<sub>med</sub> = frequency of the median class

W = width of the median class

N = total of frequencies

Stat 1300, UCLA, Ivo Dinov

38

## Median of Grouped Data -- Example

Class Interval	Frequency	Cumulative Frequency
20-under 30	6	6
30-under 40	18	24
40-under 50	11	35
50-under 60	11	46
60-under 70	3	49
70-under 80	1	50

N = 50

$$\begin{aligned} Md &= L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W) \\ &= 40 + \frac{25 - 24}{11}(10) \\ &= 40.909 \end{aligned}$$

Stat 1300, UCLA, Ivo Dinov

39

## Mode of Grouped Data

- Midpoint of the modal class
- Modal class has the greatest frequency

Class Interval	Frequency
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1

$$\text{Mode} = \frac{30 + 40}{2} = 35$$

Stat 1300, UCLA, Ivo Dinov

40

## Variance and Standard Deviation of Grouped Data

Population

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

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

Sample

$$S^2 = \frac{\sum f(M - \bar{X})^2}{n - 1}$$

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

Stat 1300, UCLA, Ivo Dinov

41

## Population Variance and Standard Deviation of Grouped Data

Class Interval	f	M	fM	M-μ	(M-μ) <sup>2</sup>	f(M-μ) <sup>2</sup>
20-under 30	6	25	150	-18	324	1944
30-under 40	18	35	630	-8	64	1152
40-under 50	11	45	495	2	4	44
50-under 60	11	55	605	12	144	1584
60-under 70	3	65	195	22	484	1452
70-under 80	1	75	75	32	1024	1024
	50		2150			7200

$$\sigma^2 = \frac{\sum f(M - \mu)^2}{N} = \frac{7200}{50} = 144$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{144} = 12$$

Stat 1300, UCLA, Ivo Dinov

42

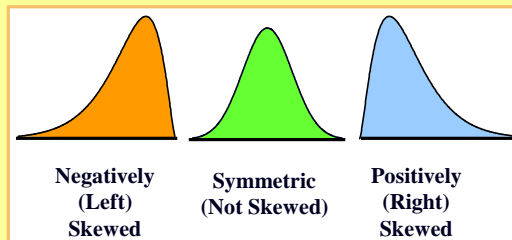
## Measures of Shape

- Skewness
  - Absence of symmetry
  - Extreme values in one side of a distribution
- Kurtosis
  - Peakedness of a distribution
  - Leptokurtic: high and thin
  - Mesokurtic: normal shape
  - Platykurtic: flat and spread out
- Box and Whisker Plots
  - Graphic display of a distribution
  - Reveals skewness

Stat 130D, UCLA, Ivo Dinov

43

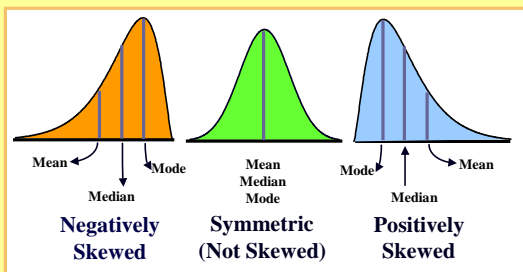
## Skewness



Stat 130D, UCLA, Ivo Dinov

44

## Skewness



Stat 130D, UCLA, Ivo Dinov

45

## Coefficient of Skewness

- Summary measure for skewness
- If  $S < 0$ , the distribution is negatively skewed (skewed to the left).
- If  $S = 0$ , the distribution is symmetric (not skewed).
- If  $S > 0$ , the distribution is positively skewed (skewed to the right).

Stat 130D, UCLA, Ivo Dinov

46

## Skewness & Kurtosis

- What do we mean by symmetry and positive and negative skewness? Kurtosis?

Properties?!?

$$\text{Skewness} = \frac{\sum_{k=1}^N (y_k - \bar{y})^3}{(N-1)SD^3}; \quad \text{Kurtosis} = \frac{\sum_{k=1}^N (y_k - \bar{y})^4}{(N-1)SD^4}$$

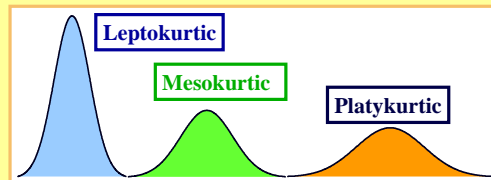
- Skewness is linearly invariant  
 $Sk(aX+b) = Sk(X)$
- Skewness is a measure of unsymmetry
- Kurtosis is a measure of flatness
- Both are used to quantify departures from StdNormal
- Skewness(StdNorm)=0; Kurtosis(StdNorm)=3

Stat 130D, UCLA, Ivo Dinov

47

## Kurtosis

- Peakedness of a distribution
  - Leptokurtic: high and thin
  - Mesokurtic: normal in shape
  - Platykurtic: flat and spread out



Stat 130D, UCLA, Ivo Dinov

48

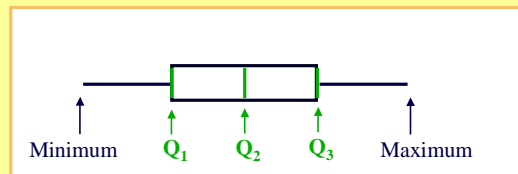
## Box and Whisker Plot

- Five specific values are used:
  - Median,  $Q_2$
  - First quartile,  $Q_1$
  - Third quartile,  $Q_3$
  - Minimum value in the data set
  - Maximum value in the data set
- Inner Fences
  - $IQR = Q_3 - Q_1$
  - Lower inner fence =  $Q_1 - 1.5 IQR$
  - Upper inner fence =  $Q_3 + 1.5 IQR$
- Outer Fences
  - Lower outer fence =  $Q_1 - 3.0 IQR$
  - Upper outer fence =  $Q_3 + 3.0 IQR$

Stat 1300, UCLA, Ivo Dinov

49

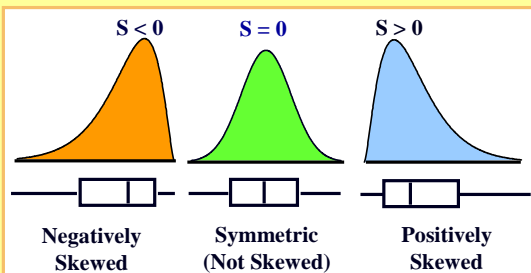
## Box and Whisker Plot



Stat 1300, UCLA, Ivo Dinov

50

## Skewness: Box and Whisker Plots, and Coefficient of Skewness



Stat 1300, UCLA, Ivo Dinov

51

## Pearson Product-Moment Correlation Coefficient

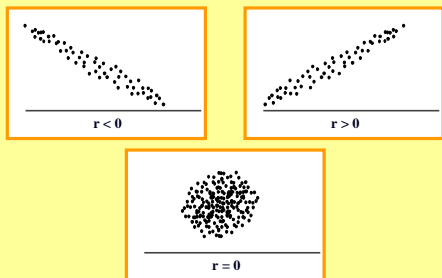
$$r = \frac{SSXY}{\sqrt{(SSX)(SSY)}} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right] \left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}$$

**$-1 \leq r \leq 1$**

Stat 1300, UCLA, Ivo Dinov

52

## Three Degrees of Correlation



Stat 1300, UCLA, Ivo Dinov

53

## Computation of $r$ for the Economics Example (Part 1)

Day	Interest X	Futures Index Y	$X^2$	$Y^2$	$XY$
1	7.43	221	55,205	48,841	1,642.03
2	7.48	222	55,950	49,284	1,660.56
3	8.00	226	64,000	51,076	1,808.00
4	7.75	225	60,063	50,625	1,743.75
5	7.60	224	57,760	50,176	1,702.40
6	7.63	223	58,217	49,729	1,701.49
7	7.68	223	58,982	49,729	1,712.64
8	7.67	226	58,829	51,076	1,733.42
9	7.59	226	57,608	51,076	1,715.34
10	8.07	235	65,125	55,225	1,896.45
11	8.03	233	64,481	54,289	1,870.99
12	8.00	241	64,000	58,081	1,928.00
<b>Summations</b>	<b>92.93</b>	<b>2,725</b>	<b>720,220</b>	<b>619,207</b>	<b>21,115.07</b>

Stat 1300, UCLA, Ivo Dinov

54

## Computation of r for the Economics Example (Part 2)

$$r = \frac{\sum XY - \frac{(\sum X)(\sum Y)}{n}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{n}\right] \left[\sum Y^2 - \frac{(\sum Y)^2}{n}\right]}}$$

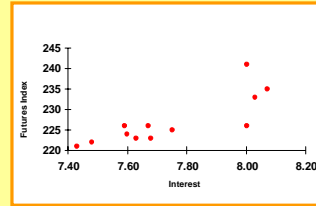
$$= \frac{(21,115,07) - \frac{(92,93)(2725)}{12}}{\sqrt{\left[720,22 - \frac{(92,93)^2}{12}\right] \left[619,207 - \frac{(2725)^2}{12}\right]}}$$

$$= .815$$

Stat 1300, UCLA, Ivo Dinov

55

## Scatter Plot and Correlation Matrix for the Economics Example



	Interest	Futures Index
Interest	1	0.815254
Futures Index	0.815254	1

Stat 1300, UCLA, Ivo Dinov

56

```
// Palindrome Testing Program (1 of 5)
// test for palindrome property – Cstrings vs Strings!
#include <iostream>
#include <string>
#include <cctype>
using namespace std;

void swap(char& lhs, char& rhs);
//swaps char args corresponding to parameters lhs and rhs

string reverse(const string& str);
//returns a copy of arg corresponding to parameter
//str with characters in reverse order.

string removePunct(const string& src,
                  const string& punct);
//returns copy of string src with characters
//in string punct removed

string makeLower(const string& s);
//returns a copy of parameter s that has all upper case
//characters forced to lower case, other characters unchanged.
//Uses <string>, which provides tolower

bool isPal(const string& this_String);
//uses makeLower, removePunct.
//if this_String is a palindrome,
// return true;
//else
// return false;
```

57

```
// Palindrome Testing Program (2 of 5)
int main()
{
    string str;
    cout << "Enter a candidate for palindrome test "
         << "\nfollowed by pressing return.\n";
    getline(cin, str);
    if (isPal(str))
        cout << "\"" << str << "\" is a palindrome ";
    else
        cout << "\"" << str << "\" is not a palindrome ";
    cout << endl;
    return 0;
}

void swap(char& lhs, char& rhs)
{
    char tmp = lhs;
    lhs = rhs;
    rhs = tmp;
}
```

58

```
// Palindrome Testing Program (3 of 5)

string reverse(const string& str)
{
    int start = 0;
    int end = str.length();
    string tmp(str);
    while (start < end)
    {
        end--;
        swap(tmp[start], tmp[end]);
        start++;
    }
    return tmp;
}

//Returns arg that has all upper case characters forced to lower case,
//other characters unchanged. makeLower uses <string>, which provides
//tolower
string makeLower(const string& s) //uses <cctype>
{
    string temp(s); //This creates a working copy of s
    for (int i = 0; i < s.length(); i++)
        temp[i] = tolower(s[i]);
    return temp;
}
```

59

```
// Palindrome Testing Program (4 of 5)

//returns a copy of src with characters in punct removed
string removePunct(const string& src, const string& punct)
{
    string no_punct;
    int src_len = src.length();
    int punct_len = punct.length();
    for(int i = 0; i < src_len; i++)
    {
        string aChar = src.substr(i,1);
        int location = punct.find(aChar, 0);
        //find location of successive characters
        //of src in punct
        if (location < 0 || location >= punct_len)
            no_punct = no_punct + aChar; //aChar not in punct -- keep it
    }
    return no_punct;
}
```

60

```
// Palindrome Testing Program (5 of 5)
```

```
//uses functions makeLower, removePunct.
```

```
//Returned value:
```

```
//if this_String is a palindrome,
```

```
// return true;
```

```
//else
```

```
// return false;
```

```
bool isPal(const string& this_String)
```

```
{
```

```
    string punctuation(",:;?!\" "); //includes a blank
```

```
    string str(this_String);
```

```
    str = makeLower(str);
```

```
    string lowerStr = removePunct(str, punctuation);
```

```
    return lowerStr == reverse(lowerStr);
```

```
}
```

61