Descriptive Analysis

HERTANTO WAHYU SUBAGIO

Univariate Analysis

- Univariate analysis involves the examination across cases of one variable at a time.
- There are three major characteristics of a single variable :

✓ **distribution**
✓ **central tendency**
✓ **dispersion** / variability

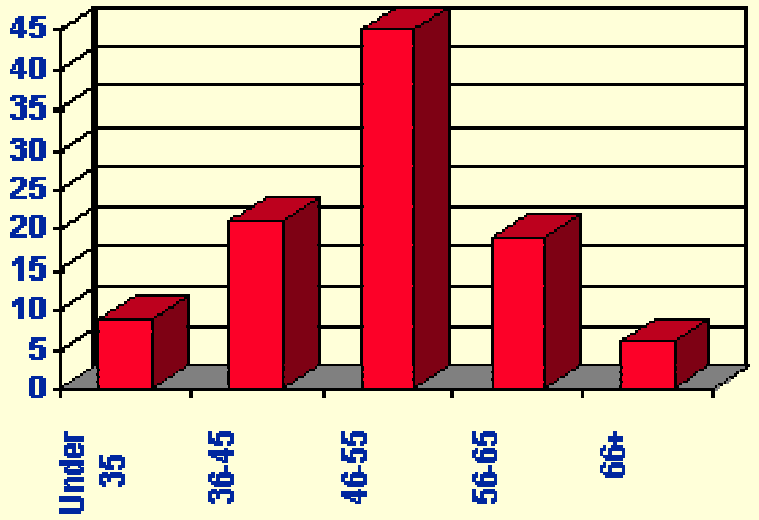
Distribution

Distribution.

is a summary of the frequency of individual values or ranges of values for a variable.

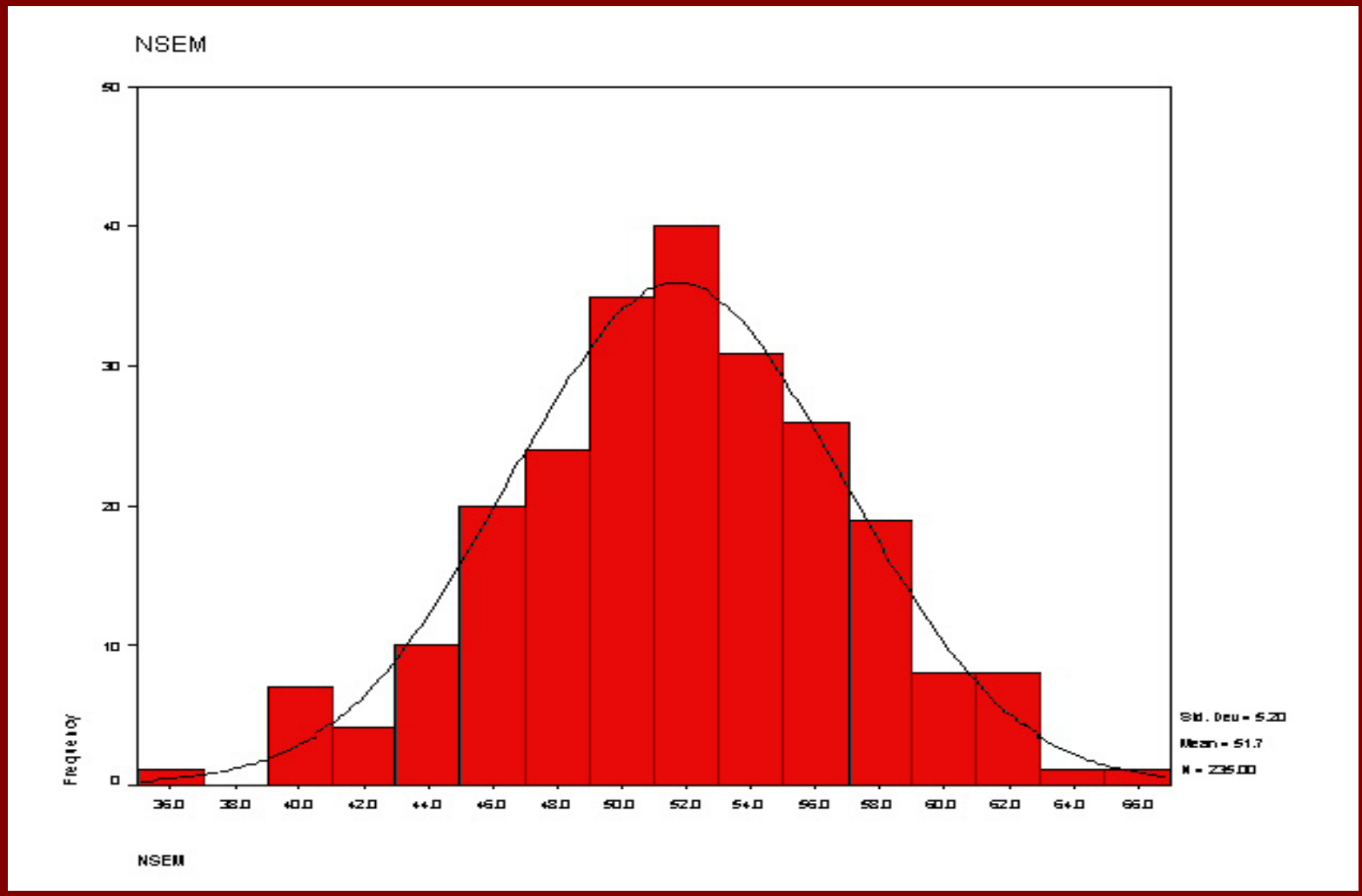
Category	Percent
Under 35	9%
36-45	21
46-55	45
56-65	19
66+	6

Frequency distribution table.



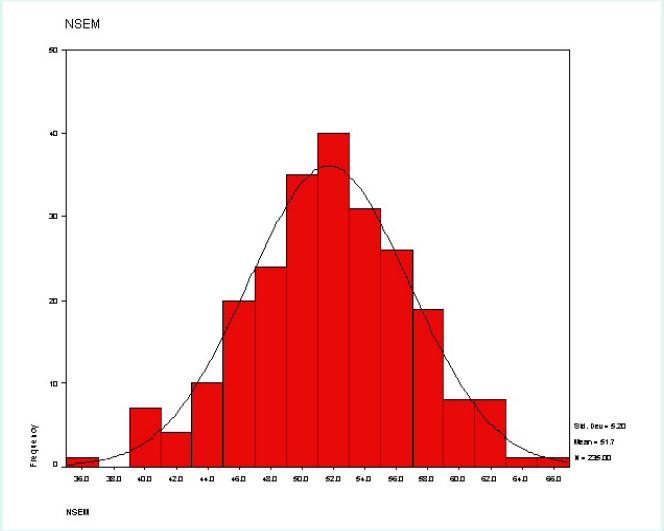
Frequency distribution bar chart.

Histogram of normal distributed data



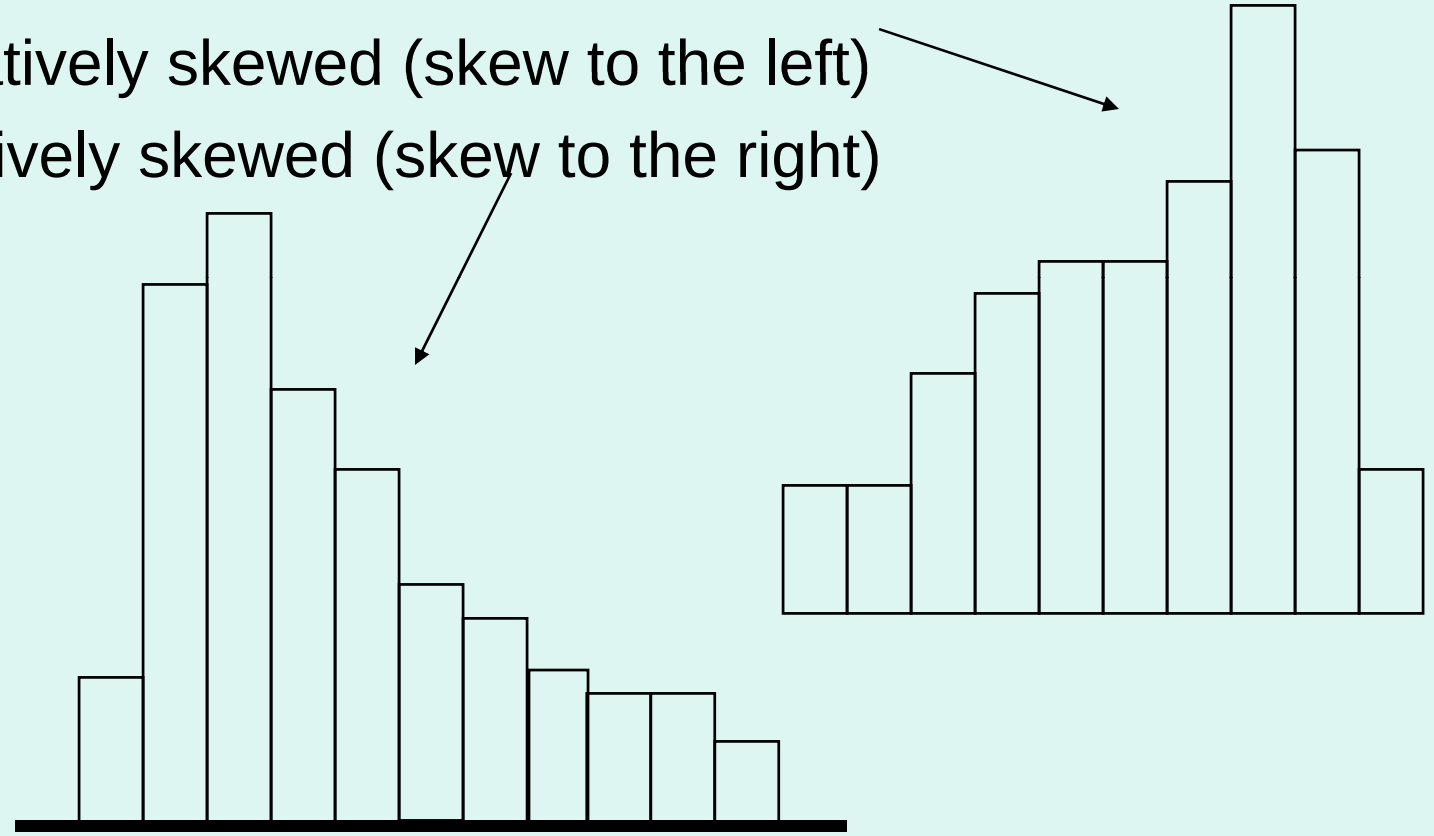
Normal or bell-shaped

- ✓ **approximately 69% of the scores in the sample fall within one standard deviation of the mean**
- ✓ **approximately 95% of the scores in the sample fall within two standard deviations of the mean**
- ✓ **approximately 99% of the scores in the sample fall within three standard deviations of the mean**



Skewness of a distribution

- Negatively skewed (skew to the left)
- Positively skewed (skew to the right)



Central tendency

Central Tendency

- is an estimate of the "center" of a distribution of values.
- There are three major types of estimates of central tendency :

✓ mean

✓ median

✓ mode

Mean

Mean or average is probably the most commonly used method of describing central tendency.

Add up all the values and divide by the number of values.

For example, consider the test score values:

15, 20, 21, 20, 36, 15, 25, 15

The sum of these 8 values is 167,
so the mean is 167/8 = 20.875.

Median

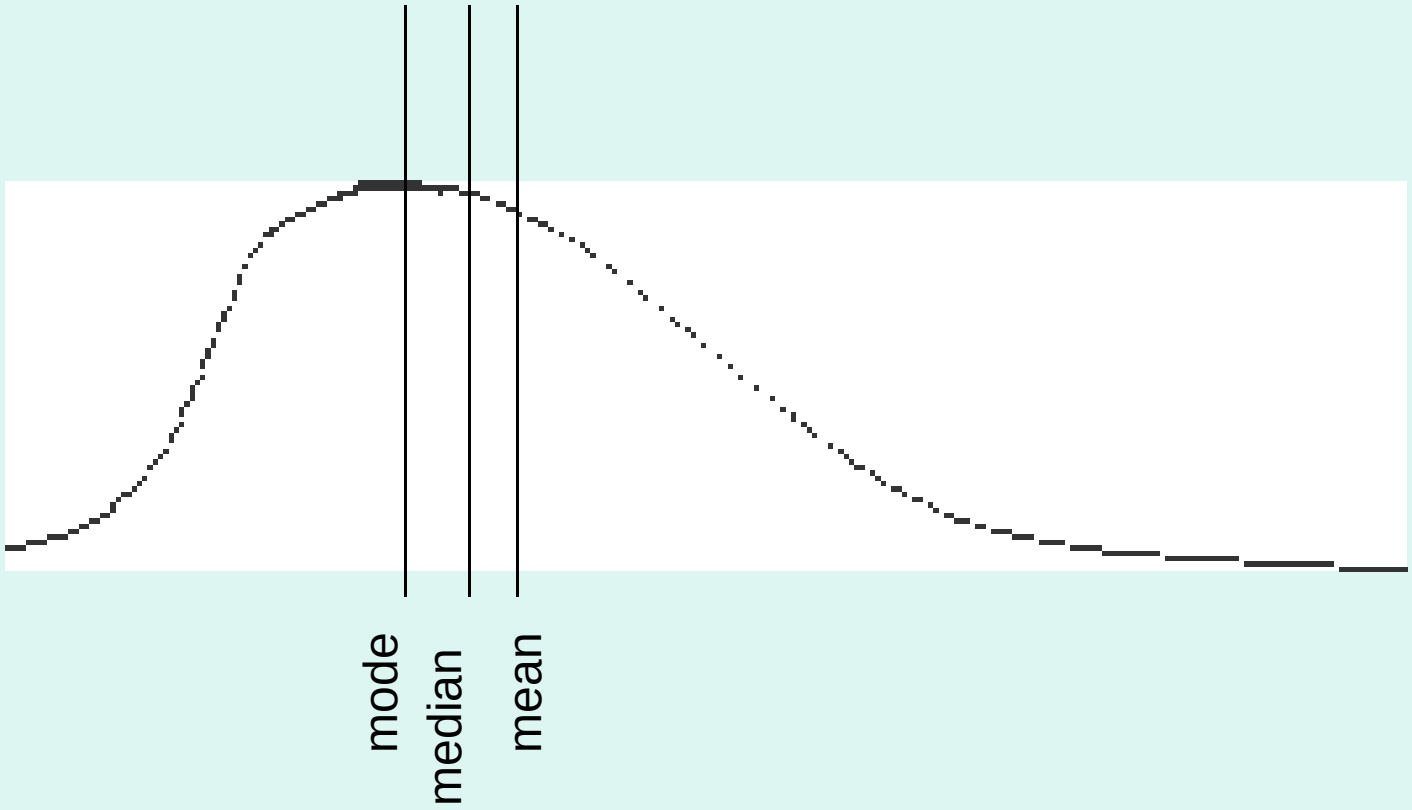
- Median is the score found at the exact middle of the set of values.
- If we order the 8 scores shown above, we would get:
  15, 15, 15, 20, 20, 21, 25, 36
- There are 8 scores and score #4 and #5 represent the halfway point. Since both of these scores are 20, the median is 20.
- If the two middle scores had different values, you would have to interpolate to determine the median.

Mode

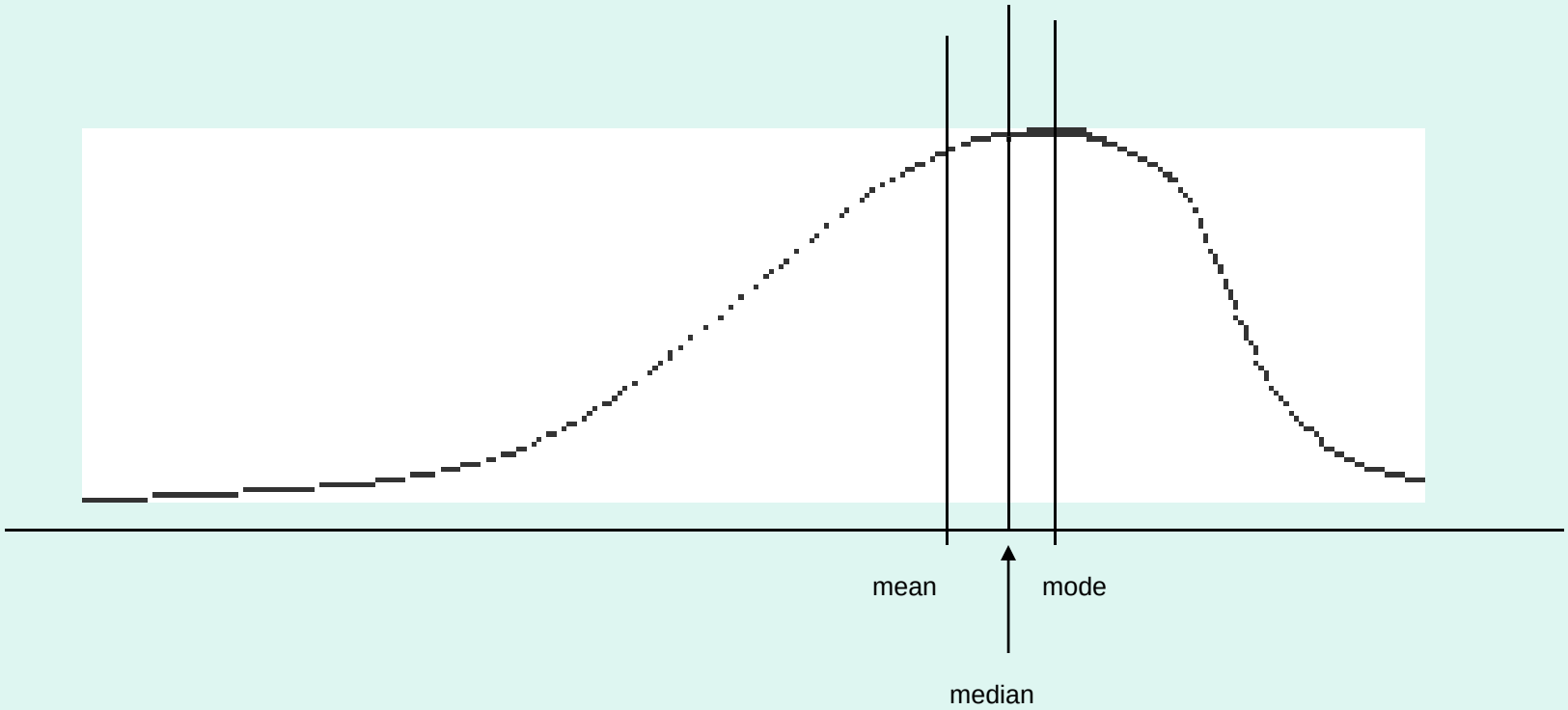
- is the most frequently occurring value in the set of scores.
- In our example, the value 15 occurs three times and is the mode.
- In some distributions there is more than one modal value. For instance, in a bimodal distribution there are two values that occur most frequently.

15, 15, 15, 20, 20, 21, 25, 36

Positively skewed distribution



Negatively skewed distribution



Variability

Variability

- Range
- Variance
  - Measure of dispersion

Population Variance $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$ parameter

Sample Variance $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ Statistic (μ is unknown)

  - Symbols will be defined in class

- Standard Deviation : square root of the variance
- Coefficient of Variation (CV) : SD / mean

Range

- the highest value minus the lowest value.

15, 20, 21, 20, 36, 15, 25, 15

- The high value is 36 and the low is 15, so the range is

36 - 15 = 21.

Standard Deviation

$$\sqrt{\frac{\Sigma(X - \overline{X})^2}{(n - 1)}}$$

where:

X = each score

$\overline{X}$ = the mean or average

n = the number of values

Σ means we sum across the values

Computing SD

15,20,21,20,36,15,25,15

to compute the standard deviation, we first find the <u>distance between each value and the mean</u> (20.875).

So, the differences from the mean are:

15 - 20.875 = -5.875
20 - 20.875 = -0.875
21 - 20.875 = +0.125
20 - 20.875 = -0.875
36 - 20.875 = 15.125
15 - 20.875 = -5.875
25 - 20.875 = +4.125
15 - 20.875 = -5.875

Notice that values that are below the mean have negative discrepancies and values above it have positive ones.

Next, we square each discrepancy:

-5.875 * -5.875 = 34.515625
-0.875 * -0.875 = 0.765625
+0.125 * +0.125 = 0.015625
-0.875 * -0.875 = 0.765625
15.125 * 15.125 = 228.765625
-5.875 * -5.875 = 34.515625
+4.125 * +4.125 = 17.015625
-5.875 * -5.875 = 34.515625

Now, we take these "squares" and sum them to get the Sum of Squares (SS) value. Here, the sum is 350.875.

Next, we divide this sum by the number of scores minus 1. Here, the result is 350.875 / 7 = 50.125.

This value is known as the **variance.**

To get the **standard deviation**, we take the square root of the variance. This would be SQRT(50.125) = 7.079901129253

Coefficient of variation

- The coefficient of variation of a distribution is the ratio of standard deviation to the mean

- Useful for comparing spread (variability) of distribution

$$\text{Sample coefficient of variation: } cv = \frac{s}{\bar{x}}$$

$$\text{Population coefficient of variation: } \mathrm{CV} = \frac{\sigma}{\mu}$$

To Obtain Frequencies and Statistics

From the menus choose:

Analyze
Descriptive Statistics
Frequencies...

Select one or more categorical or quantitative variables.

Optionally, you can:

Click Statistics for descriptive statistics for quantitative variables.
Click Charts for bar charts, pie charts, and histograms.
Click Format for the order in which results are displayed.

Descriptive Statistics
SPSS PC. 10.1

Statistics

NSEM

N	Valid	235
	Missing	2
Mean		51.6915
Median		51.8333
Mode		50.83[a]
Std. Deviation		5.1982
Variance		27.0209
Skewness		-.152
Std. Error of Skewness		.159
Kurtosis		.012
Std. Error of Kurtosis		.316
Range		30.33
Percentiles	10	45.1000
	90	58.6667

a. Multiple modes exist. The smallest value is shown

Measure of shape

- **Coefficient of Skewness**

  A measure of symmetry.

  a symmetric distribution has a coefficient of skewness=0

- **Coefficient of Kurtosis**

  A measure of the peakednes of a distribution

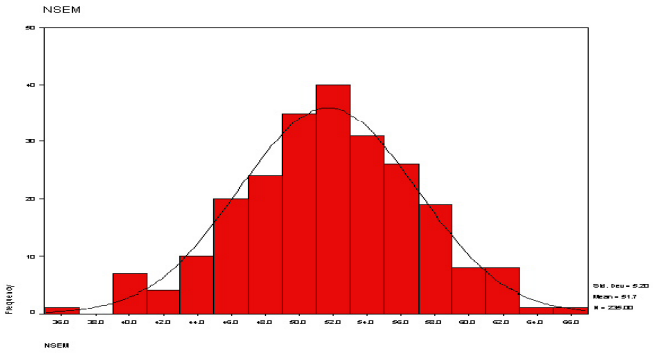
  the normal distribution=0.

Normal distribution

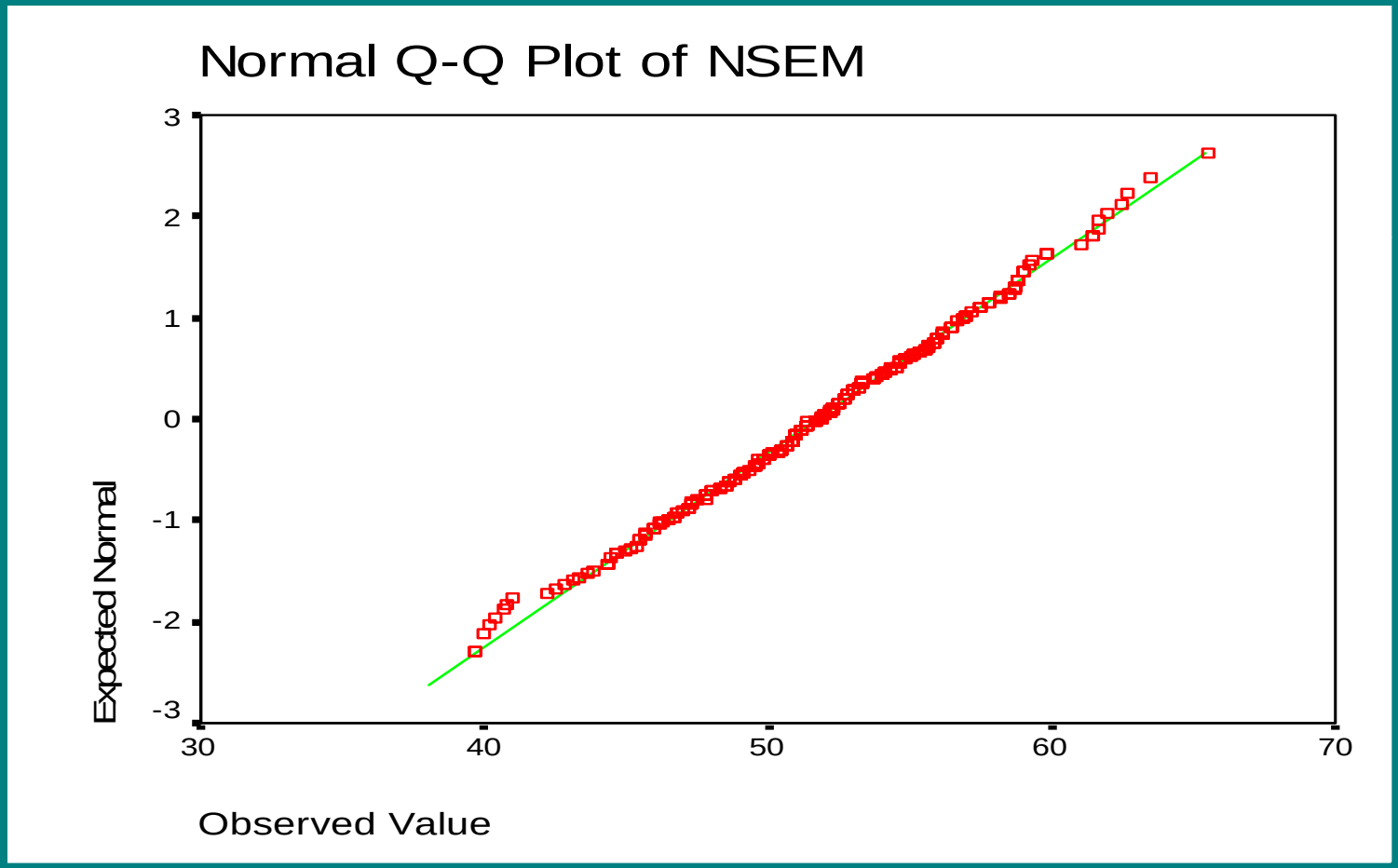
- **index of kurtosis and index of skewness**

  (between –2 and + 2 : normal)
- **normal Q-Q plot and detrended normal Q-Q plot**
- **Kolmogorov-Smirnov test / Shapiro Wilks :**

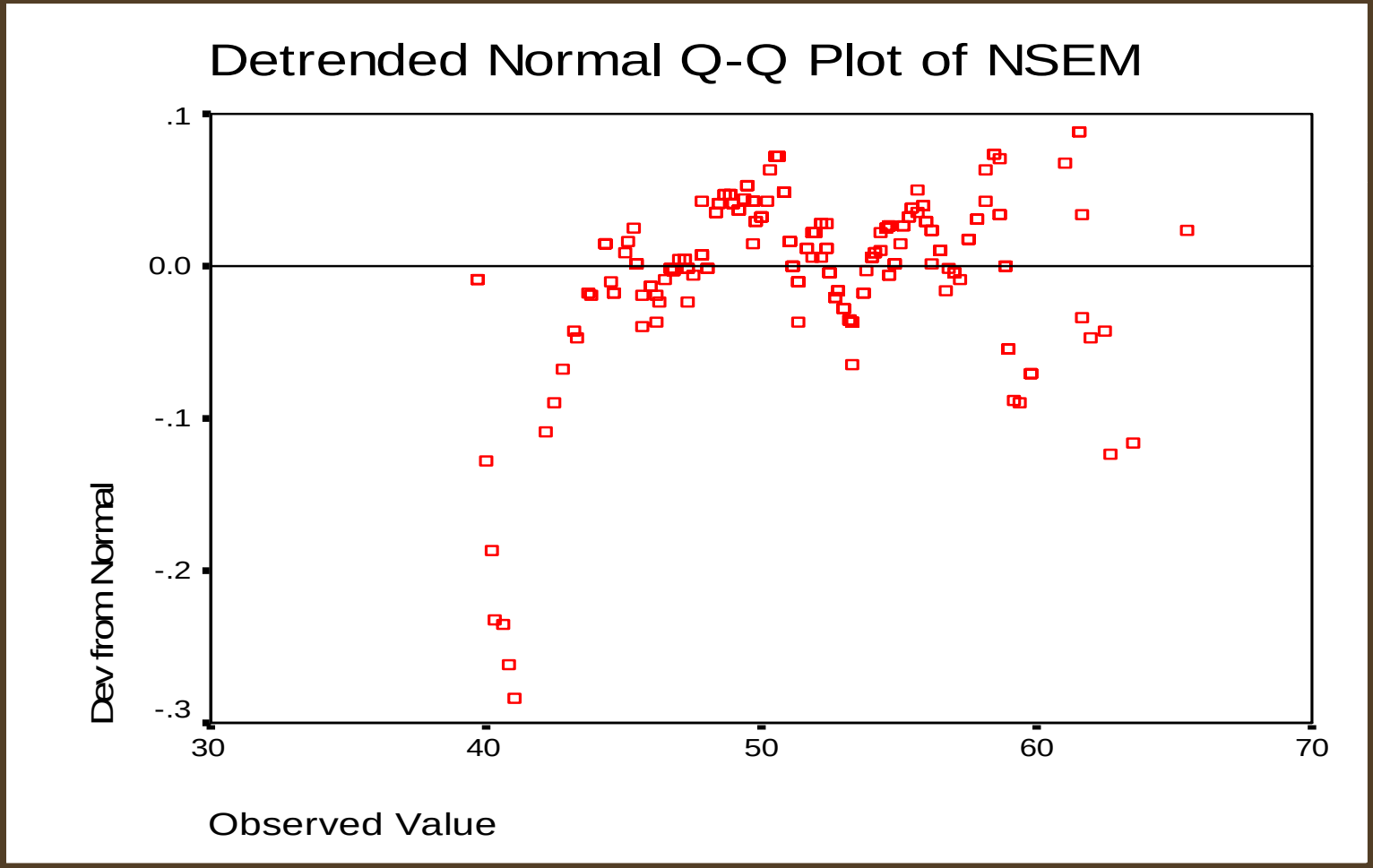
  p > 0.05 : normal distributed



Normal Q-Q plot of normal distributed data



Detrended normal Q-Q plot of normal distributed data



Normality Test
Kolmogorov-Smirnov

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
NSEM	.035	235	.200*

*. This is a lower bound of the true significanc

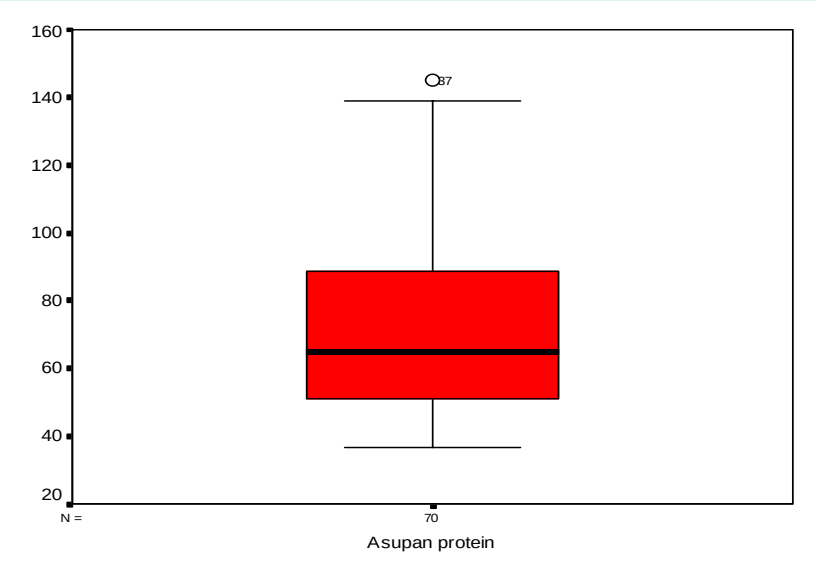
a. Lilliefors Significance Correction

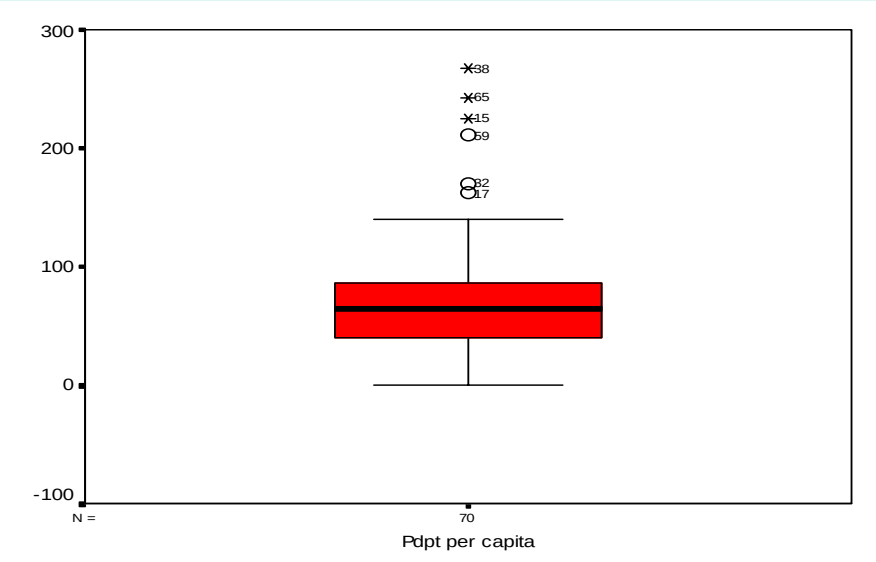
Testing normality of data

From the menus choose:

Analyze
descriptive statistic
explore
plot
normality plot

Boxplot


160
140
120
100
80
60
40
20
87
N =
70
Asupan protein


300
200
100
0
-100
38
65
15
59
82
17
N =
70
Pdpt per capita