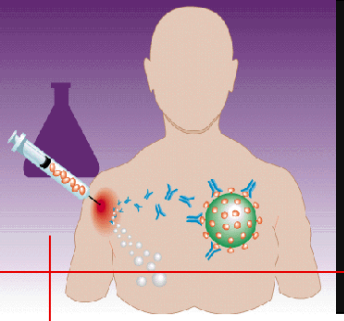
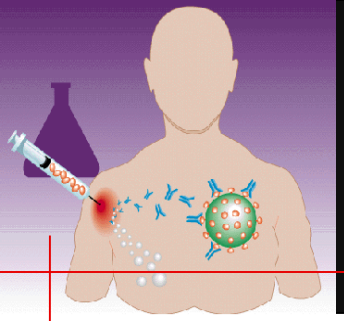


FK6163 EXEC 2018

Exploratory Data Analysis

Assc. Prof. Dr. Azmi Mohd Tamil
Dept of Community Health
Universiti Kebangsaan Malaysia



Introduction

Method of Exploring Data differs According to Types of Variables

Data Types

Quantitative
(Takes numerical values)

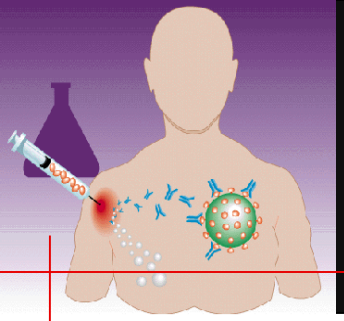
- discrete
  (whole numbers)
  e.g. Number of children

- continuous
  (takes decimal places)
  e.g. Height, Weight

Qualitative
(Takes coded numerical values)

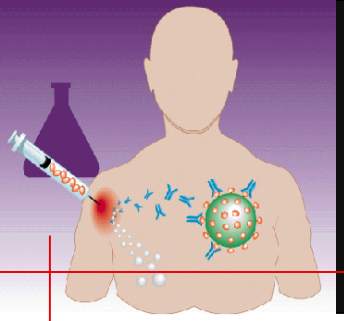
- ordinal
  (ranking order exists)
  e.g. Pain severity

- nominal
  (no ranking order)
  e.g. Race, Gender



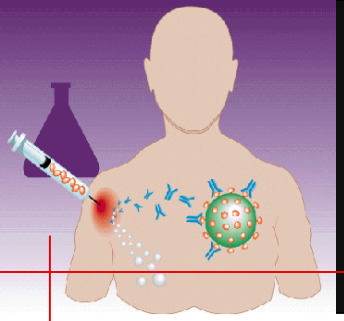
Explore

- It is the first step in the analytic process
- to explore the characteristics of the data
- to screen for errors and correct them
- to look for distribution patterns - normal distribution or not
- May require transformation before further analysis using parametric methods
- Or may need analysis using non-parametric techniques



Data Screening

- By running frequencies, we may detect inappropriate responses
- How many in the audience have 15 children and currently pregnant with the 16th?



Data Screening

occupati	age	totalmem	parity	abortion	stilbrth
HOUSEWI	44	17	15	0	0
HOUSEWI	39	12	11	0	0
HOUSEWI	36	11	10		0
HOUSEWI	25	7	10		
HOUSEWI	44	6	10		0
HOUSEWI	34	10	9	0	0
TEACHER	37	10	9	0	0
HOUSEWI	46	10	9	1	0
HOUSEWI	43	10	9	0	0
HOUSEWI	37	9	9	0	1
HOUSEWI	38	10	8	0	0
HOUSEWI	35	9	8	2	0
HOUSEWI	42	7	8	0	0
HOUSEWI	37	7	8	0	2
HOUSEWI	37	10	8	0	0
HOUSEWI	33	10	8	0	0
HOUSEWI	41	9	8	0	0

- See whether the data make sense or not.
- E.g. Parity 10 but age only 25.

Table II. Height of subjects.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.30	20	26.3	26.3	26.3
	1.40	14	18.4	18.4	44.7
	1.50	28	36.8	36.8	81.6
	1.60	10	13.2	13.2	94.7
	1.70	3	3.9	3.9	98.7
	3.70	1	1.3	1.3	100.0
	Total	76	100.0	100.0	

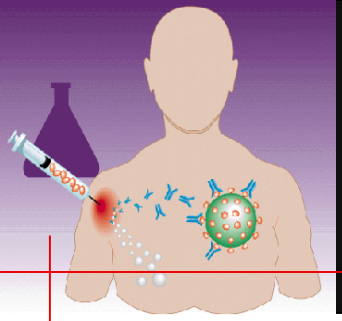
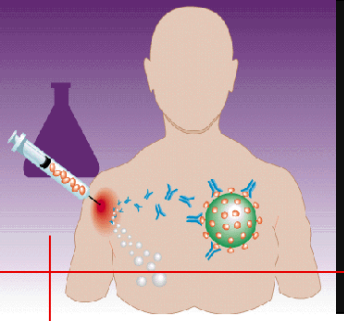


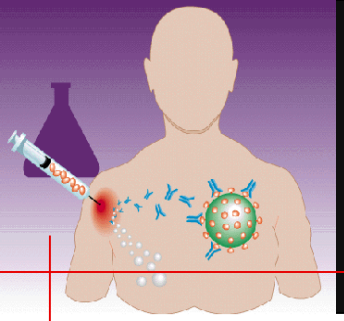
Table 1. Using Strings/Text for Categorical variables.

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	female	38	50.0	50.0	50.0
	male	13	17.1	17.1	67.1
	Male	25	32.9	32.9	100.0
	Total	76	100.0	100.0	



Data Screening

- By looking at measures of central tendency and range, we can also detect abnormal values for quantitative data



Interpreting the Box Plot

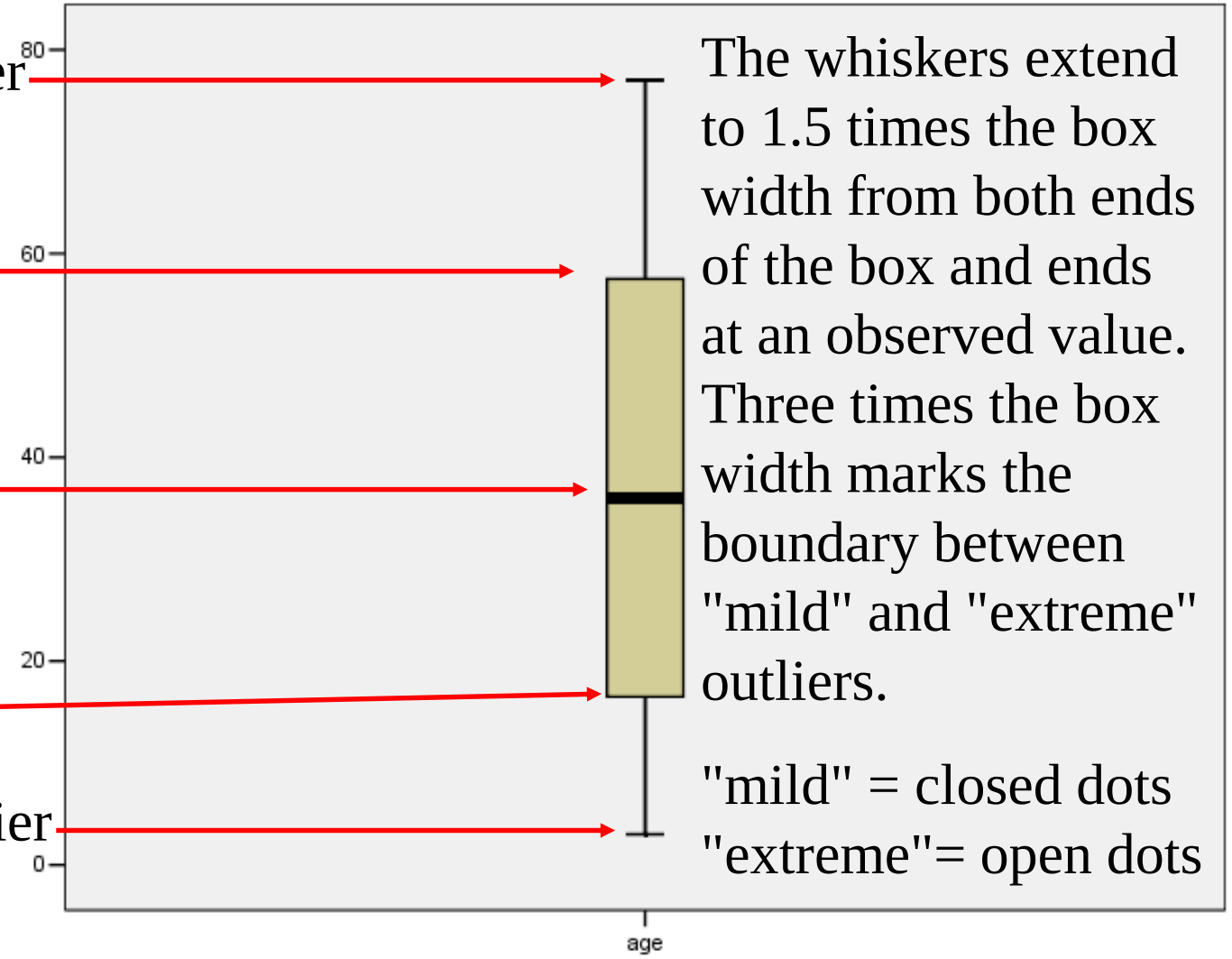
Largest non-outlier

Upper quartile

Median

Lower quartile

Smallest non-outlier

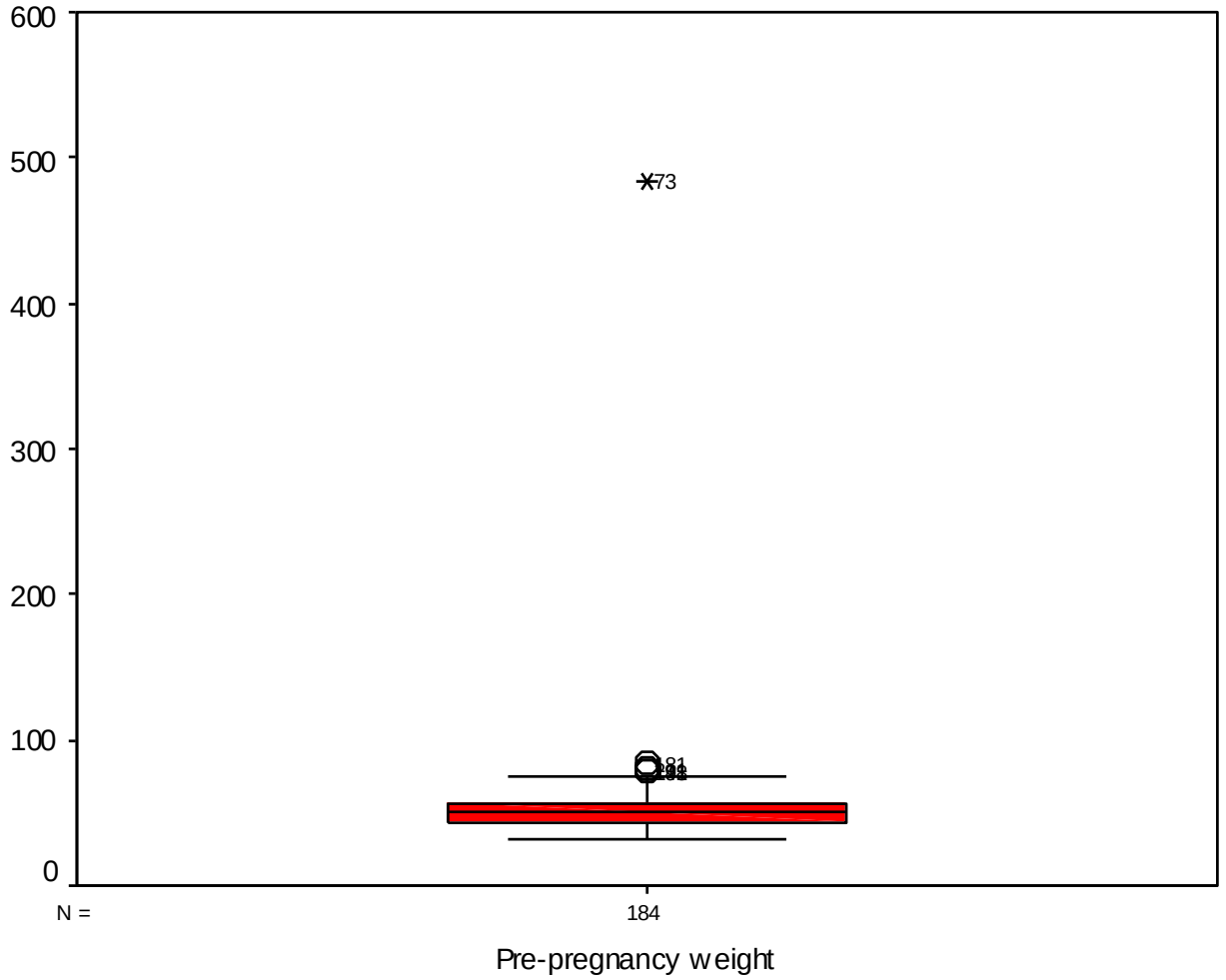


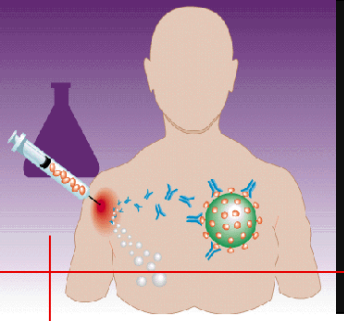
The whiskers extend to 1.5 times the box width from both ends of the box and ends at an observed value. Three times the box width marks the boundary between "mild" and "extreme" outliers.

"mild" = closed dots
"extreme"= open dots

Data Screening

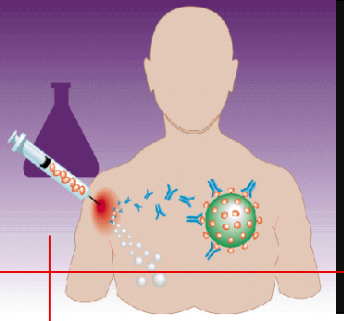
- We can also make use of graphical tools such as the box plot to detect wrong data entry





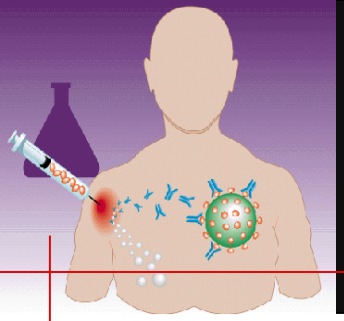
Data Cleaning

- Identify the extreme/wrong values
- Check with original data source – i.e. questionnaire
- If incorrect, do the necessary correction.
- Correction must be done before transformation, recoding and analysis.



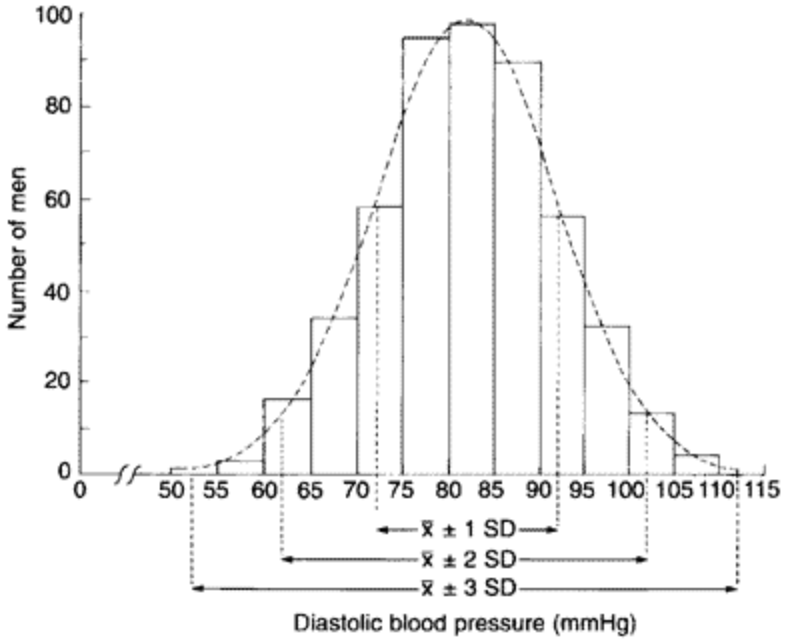
Parameters of Data Distribution

- Mean – central value of data
- Standard deviation – measure of how the data scatter around the mean
- Symmetry (skewness) – the degree of the data pile up on one side of the mean
- Kurtosis – how far data scatter from the mean



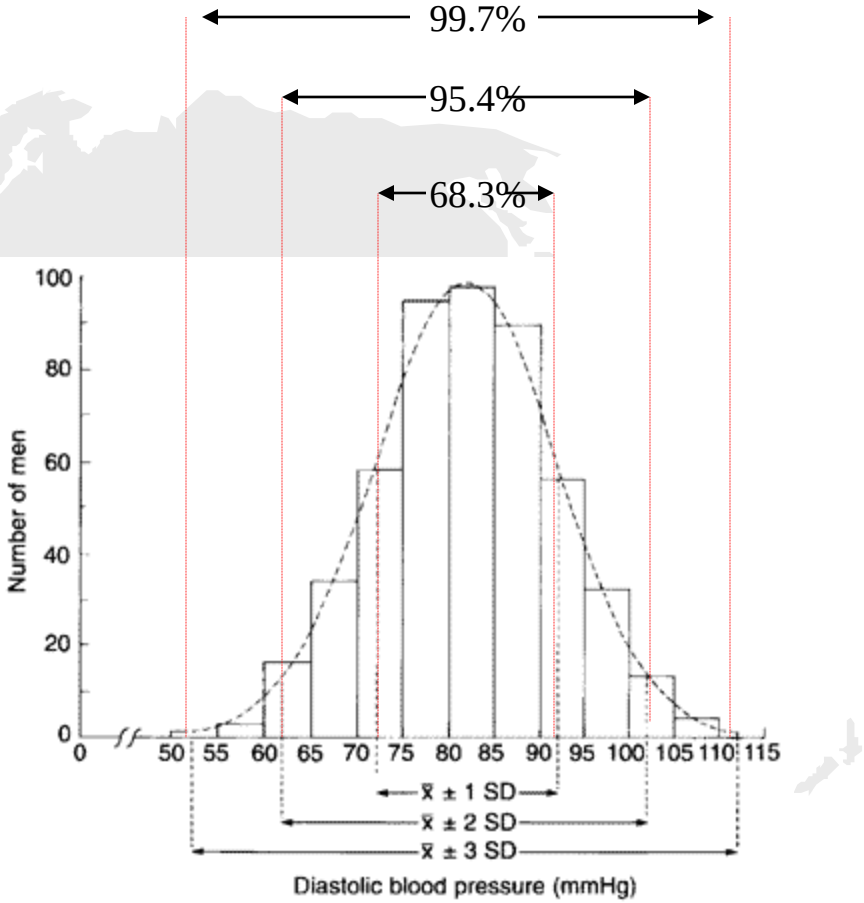
Normal distribution

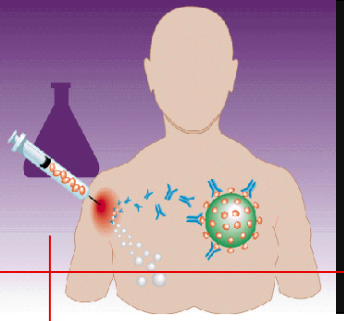
- The Normal distribution is represented by a family of curves defined uniquely by two parameters, which are the mean and the standard deviation of the population.
- The curves are always symmetrically bell shaped, but the extent to which the bell is compressed or flattened out depends on the standard deviation of the population.
- However, the mere fact that a curve is bell shaped does not mean that it represents a Normal distribution, because other distributions may have a similar sort of shape.



Normal distribution

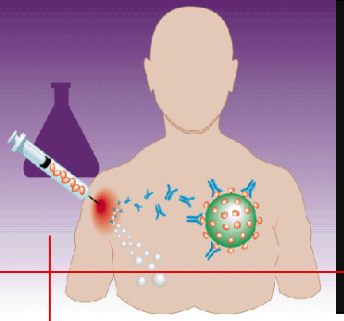
- If the observations follow a Normal distribution, a range covered by one standard deviation above the mean and one standard deviation below it includes about 68.3% of the observations;
- a range of two standard deviations above and two below (± 2sd) about 95.4% of the observations; and
- of three standard deviations above and three below (± 3sd) about 99.7% of the observations





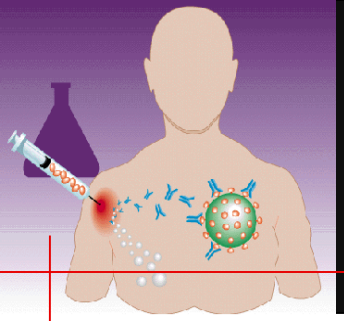
Normality

- Why bother with normality??
- Because it dictates the type of analysis that you can run on the data



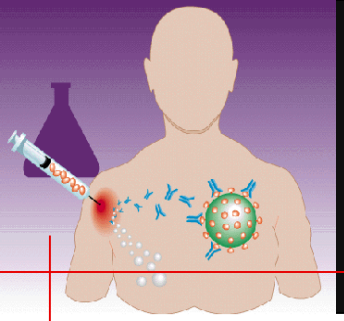
Normality-Why?
Parametric

Qualitative Dichotomus	Quantitative	Normally distributed data	Student's t Test
Qualitative Polinomial	Quantitative	Normally distributed data	ANOVA
Quantitative	Quantitative	Repeated measurement of the same individual & item (e.g. Hb level before & after treatment). Normally distributed data	Paired t Test
Quantitative - continous	Quantitative - continous	Normally distributed data	Pearson Correlation & Linear Regresssion



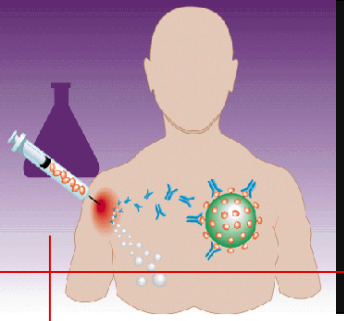
Normality-Why?
Non-parametric

Qualitative Dichotomus	Quantitative	Data not normally distributed	Wilcoxon Rank Sum Test or U Mann-Whitney Test
Qualitative Polinomial	Quantitative	Data not normally distributed	Kruskal-Wallis One Way ANOVA Test
Quantitative	Quantitative	Repeated measurement of the same individual & item	Wilcoxon Rank Sign Test
Quantitative - continous/ordinal	Quantitative - continous	Data not normally distributed	Spearman/Kendall Rank Correlation



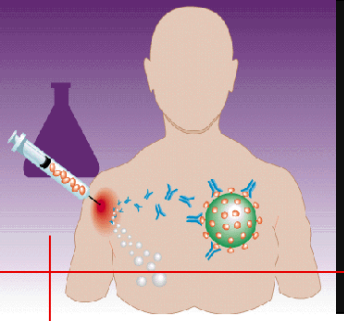
Normality-How?

- Explored graphically
  - Histogram
  - Stem & Leaf
  - Box plot
  - Normal probability plot
  - Detrended normal plot
- Explored statistically
  - Kolmogorov-Smirnov statistic, with Lilliefors significance level and the Shapiro-Wilks statistic
  - Skew ness (0)
  - Kurtosis (0)
    - + leptokurtic
    - 0 mesokurtik
    - - platykurtic



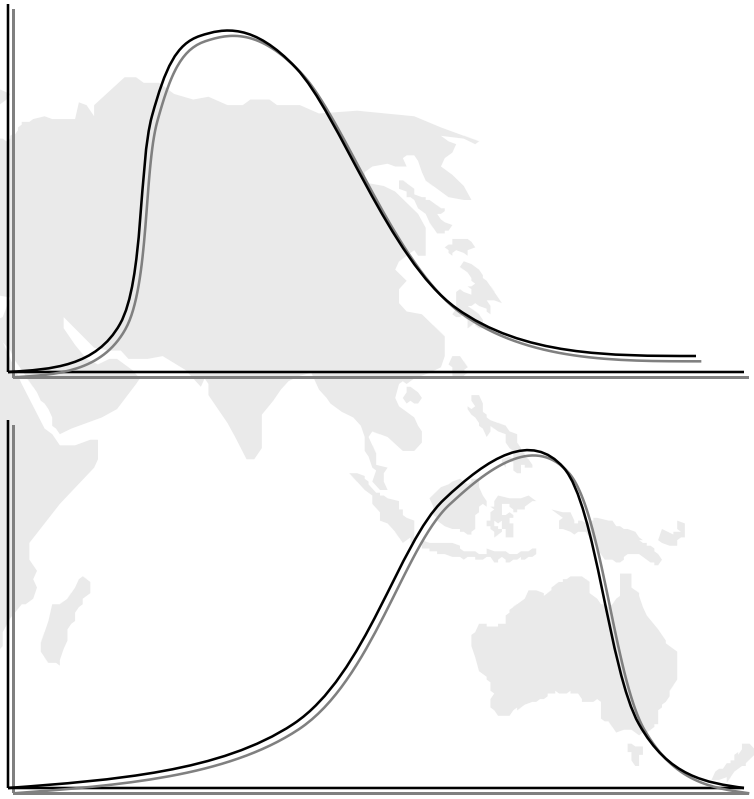
Kolmogorov- Smirnov

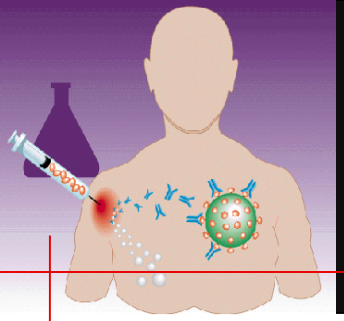
- In the 1930's, Andrei Nikolaevich Kolmogorov (1903-1987) and N.V. Smirnov (his student) came out with the approach for comparison of distributions that did not make use of parameters.
- This is known as the Kolmogorov-Smirnov test.



Skew ness

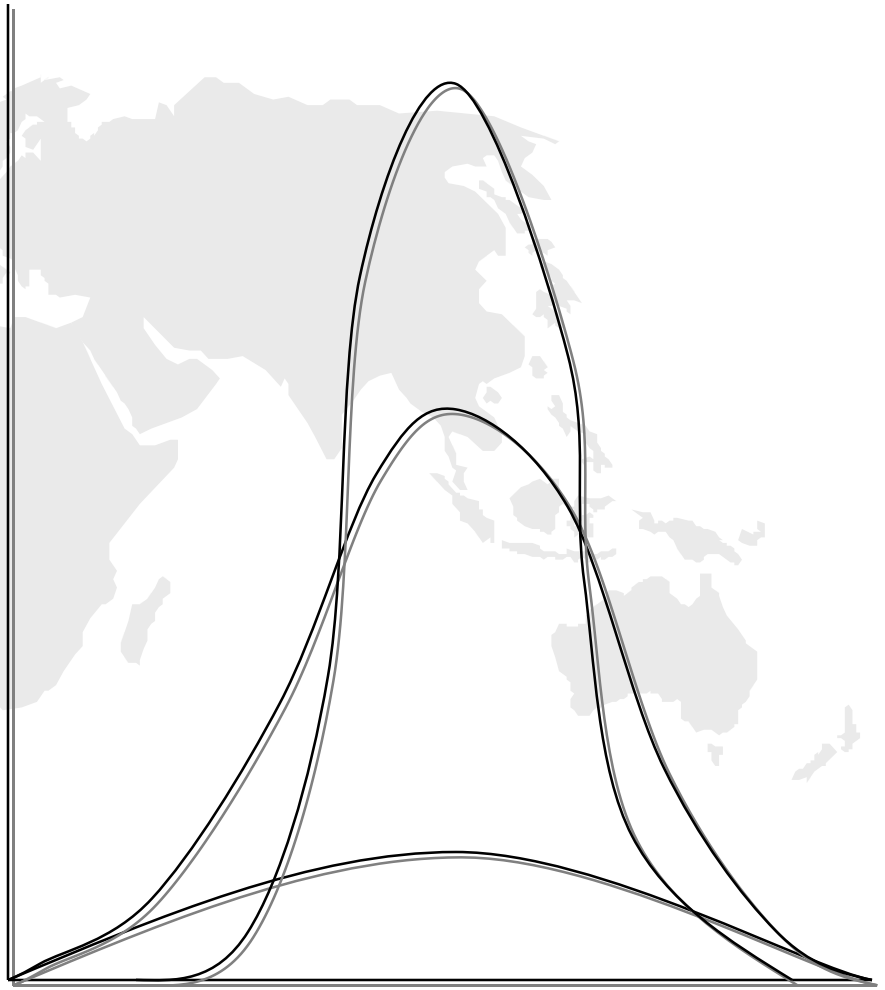
- Skewed to the right indicates the presence of large extreme values
- Skewed to the left indicates the presence of small extreme values

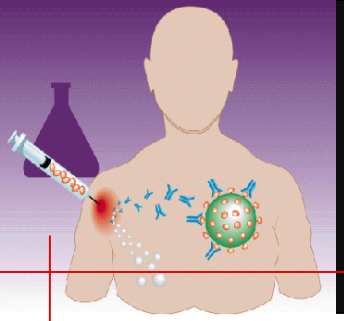




Kurtosis

- For symmetrical distribution only.
- Describes the shape of the curve
- Mesokurtic - average shaped
- Leptokurtic - narrow & slim
- Platikurtic - flat & wide

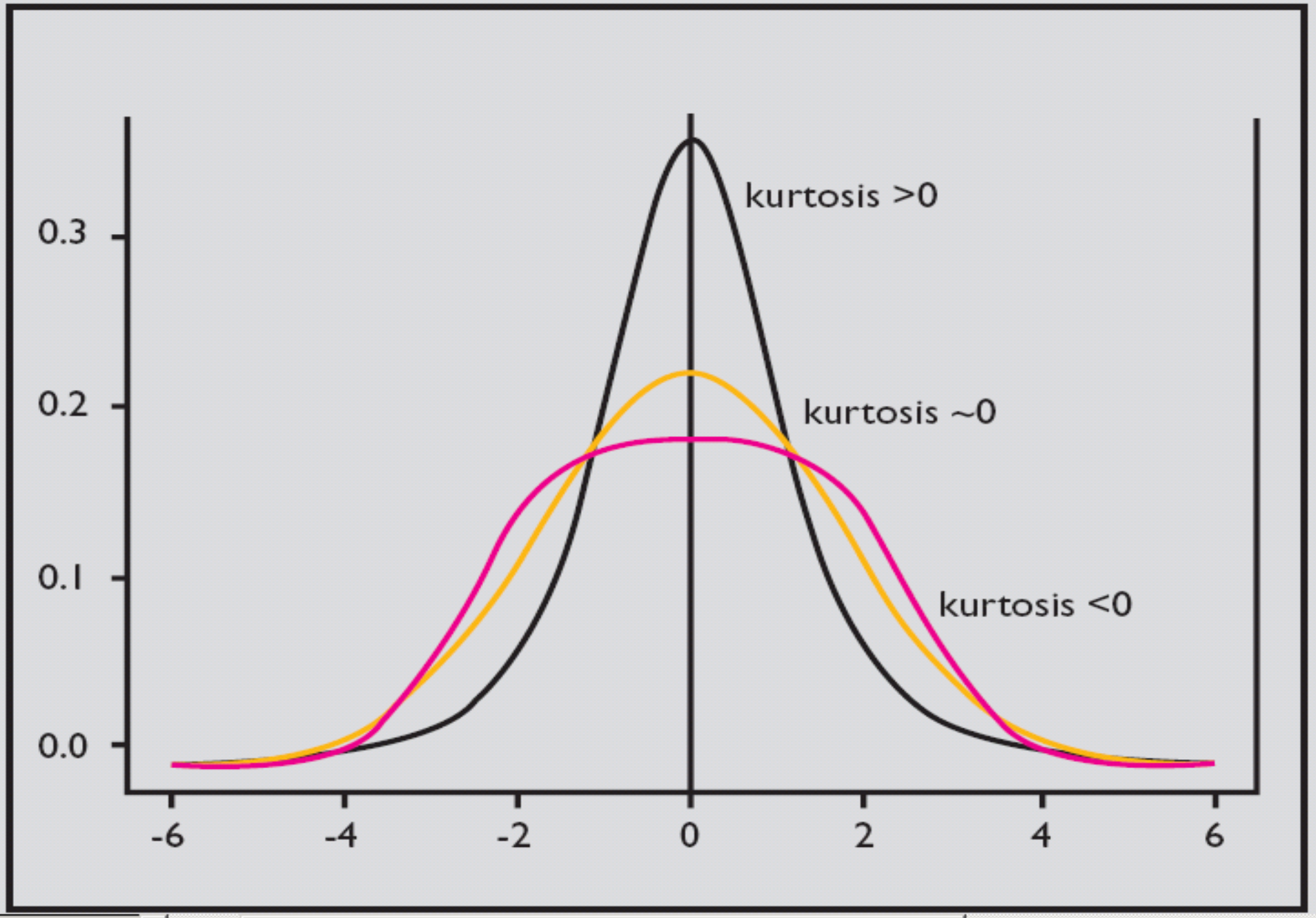


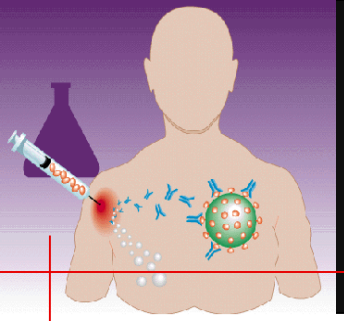


Skew ness & Kurtosis

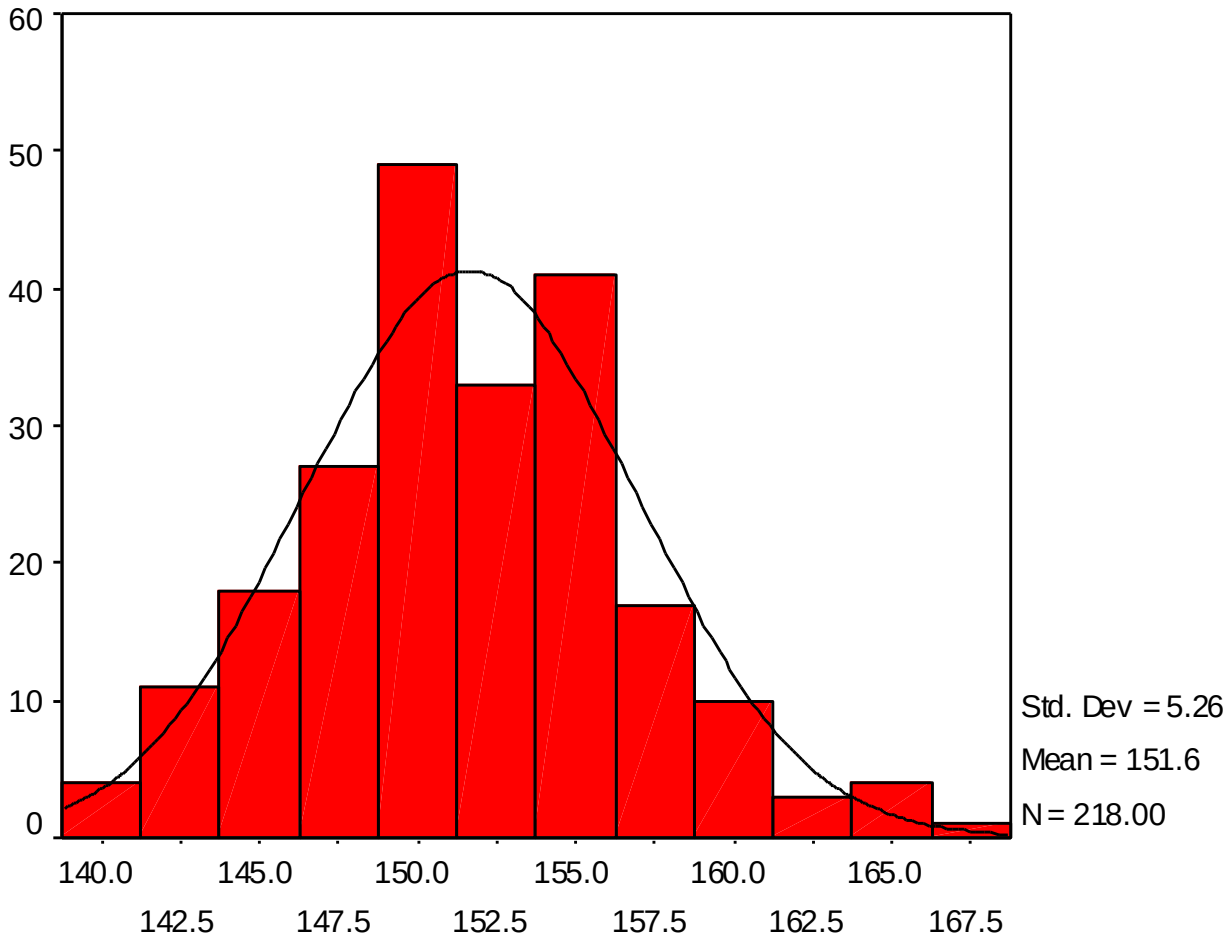
- Skew ness ranges from -3 to 3.
- Acceptable range for normality is skew ness lying between -1 to 1.
- Normality should not be based on skew ness alone; the kurtosis measures the “peak ness” of the bell-curve (see Fig. 4).
- Likewise, acceptable range for normality is kurtosis lying between -1 to 1.

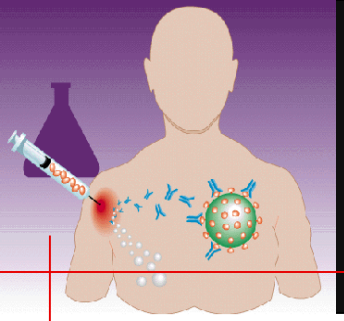
Fig. 4





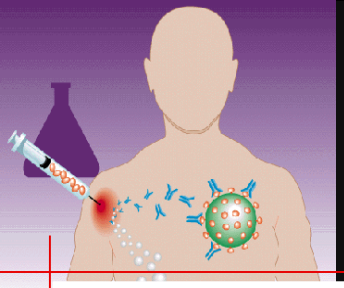
Normality - Examples
Graphically



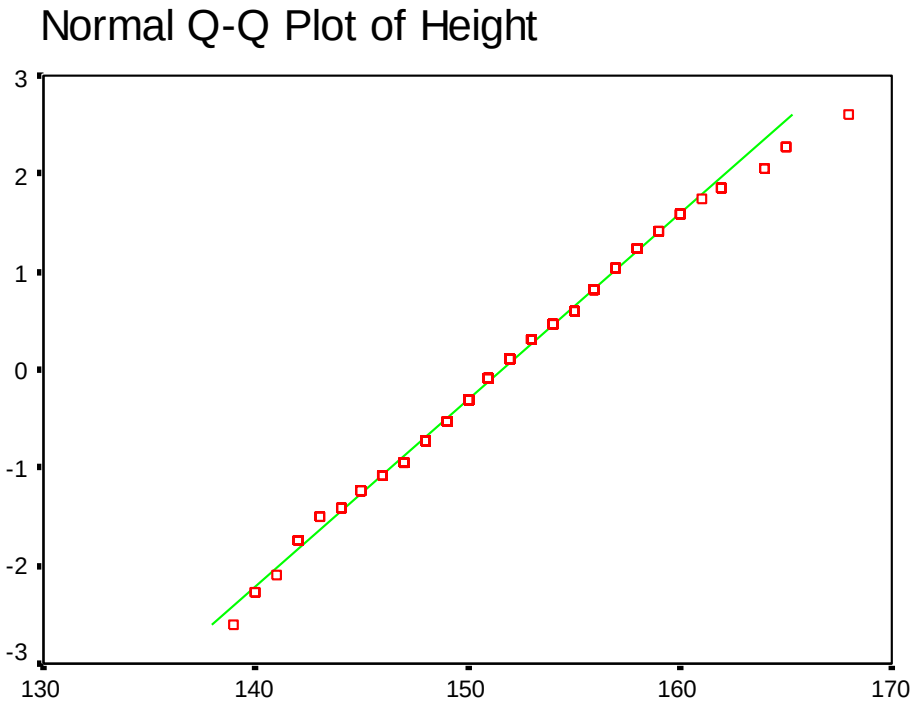


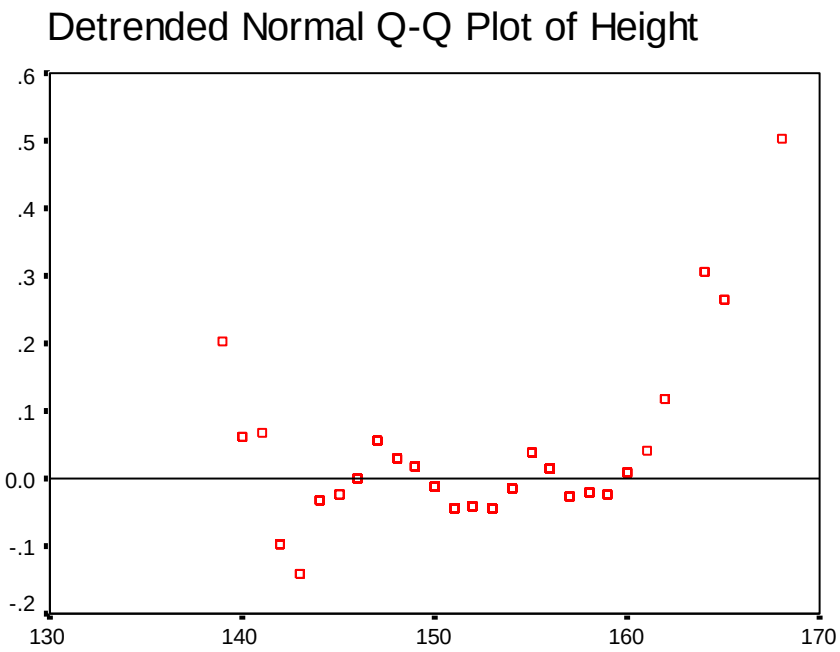
Q&Q Plot

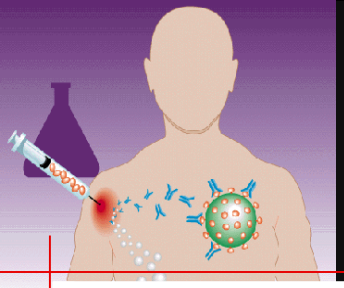
- This plot compares the quintiles of a data distribution with the quintiles of a standardised theoretical distribution from a specified family of distributions (in this case, the normal distribution).
- If the distributional shapes differ, then the points will plot along a curve instead of a line.
- Take note that the interest here is the central portion of the line, severe deviations means non-normality. Deviations at the “ends” of the curve signifies the existence of outliers.



Normality - Examples
Graphically







Normality - Examples Statistically

Descriptives

			Statistic	Std. Error
Height	Mean		151.65	.356
	95% Confidence Interval for Mean	Lower Bound	150.94	
		Upper Bound	152.35	
	5% Trimmed Mean		151.59	
	Median		151.50	
	Variance		27.649	
	Std. Deviation		5.258	
	Minimum		139	
	Maximum		168	
	Range		29	
	Interquartile Range		8.00	
	Skewness		.148	.165
	Kurtosis		.061	.328

Normal distribution
Mean=median=mode

Skewness & kurtosis within ± 1

$p > 0.05$, so normal distribution

Shapiro-Wilks; only if sample size less than 100.

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
Height	.060	218	.052

a. Lilliefors Significance Correction

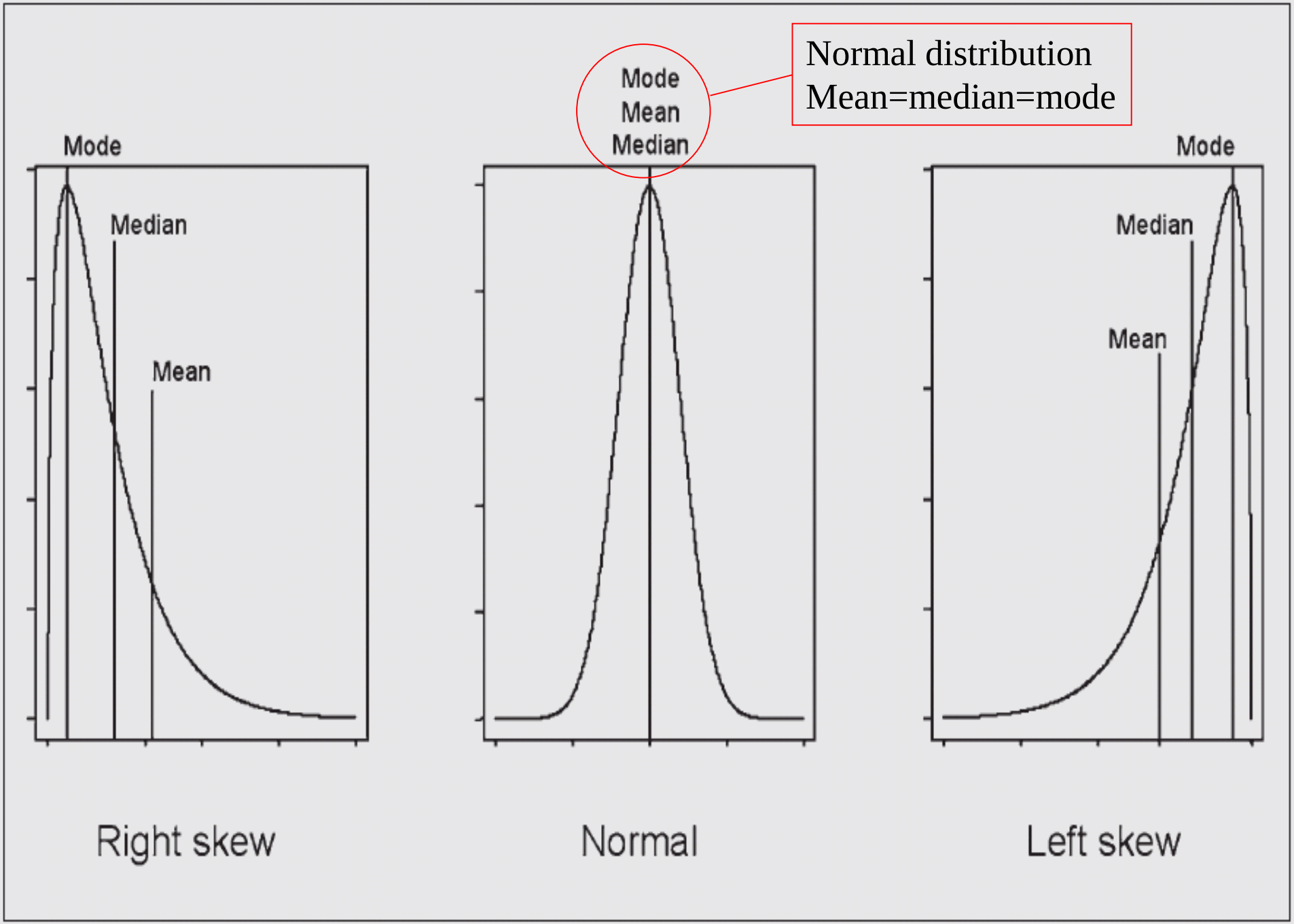
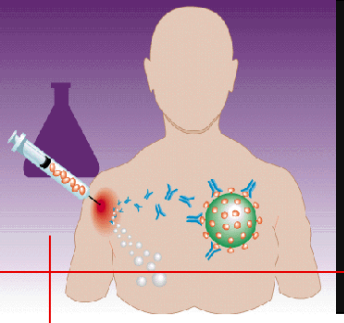


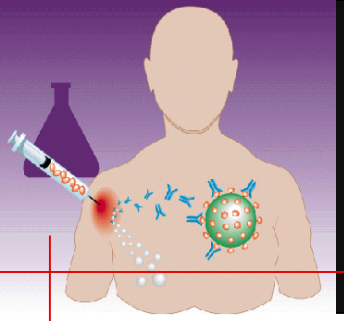
Fig. 2 Distributions of Quantitative Data.



K-S Test

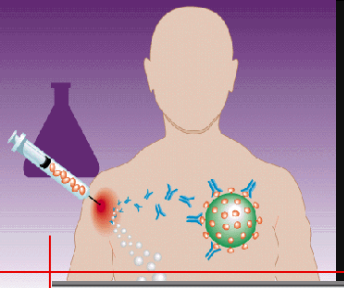
Table III. Normality tests.

	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Right Skew	0.187	76	0.000	0.884	76	0.000
Normal	0.079	76	0.200	0.981	76	0.325
Left skew	0.117	76	0.012	0.927	76	0.000



K-S Test

- very sensitive to the sample sizes of the data.
- For small samples (n<20, say), the likelihood of getting p<0.05 is low
- for large samples (n>100), a slight deviation from normality will result in being reported as abnormal distribution



Guide to deciding on normality

Table IV. Flowchart for normality checking.

1. Small samples* (n<30): always assume not normal.

2. Moderate samples (30-100).

   If formal test is significant, accept non-normality otherwise double-check using graphs, skewness and kurtosis to confirm normality.

3. Large samples (n>100).

   If formal test is not significant, accept normality otherwise Double-check using graphs, skewness and kurtosis to confirm non-normality.

* Reminder: not ethical to do small sized studies[12].