Unit 3
Shapes of Distribution & Graphs

Dr Mahmoud Alhussami

Shapes of Distribution

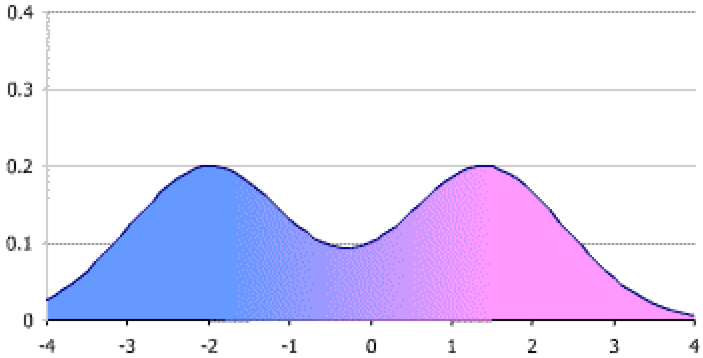
- A third important property of data – after location and dispersion - is its shape
- Distributions of quantitative variables can be described in terms of a number of features, many of which are related to the distributions' physical appearance or shape when presented graphically.
  - modality
  - Symmetry and skewness
  - Degree of skewness
  - Kurtosis

Modality

- The modality of a distribution concerns with how many peaks or high points there are.
- A distribution with a single peak, one value a high frequency is a unimodal distribution.

Modality

- A distribution with two or more peaks called multimodal distribution.



Symmetry and Skewness

- A distribution is symmetric if the distribution could be split down the middle to form two haves that are mirror images of one another.
- In asymmetric distributions, the peaks are off center, with a bull of scores clustering at one end, and a tail trailing off at the other end. Such distributions are often describes as skewed.
  - When the longer tail trails off to the right this is a positively skewed distribution. E.g. annual income.
  - When the longer tail trails off to the left this is called negatively skewed distribution. E.g. age at death.

Symmetry and Skewness

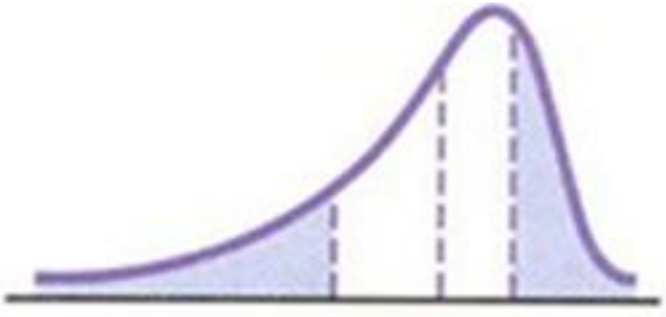
- Shape can be described by degree of asymmetry (i.e., skewness).

mean > median	positive or right-skewness
mean = median	symmetric or zero-skewness
mean < median	negative or left-skewness

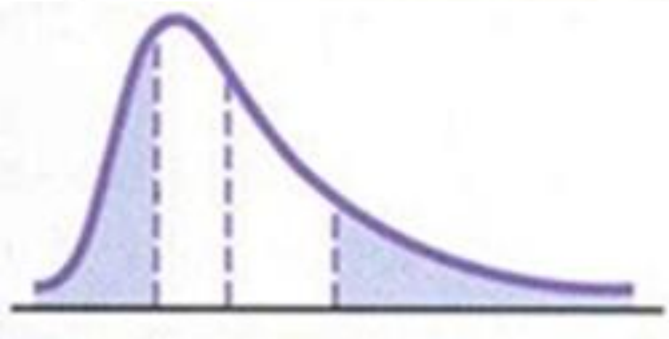
- Positive skewness can arise when the mean is increased by some unusually high values.
- Negative skewness can arise when the mean is decreased by some unusually low values.

Skewness

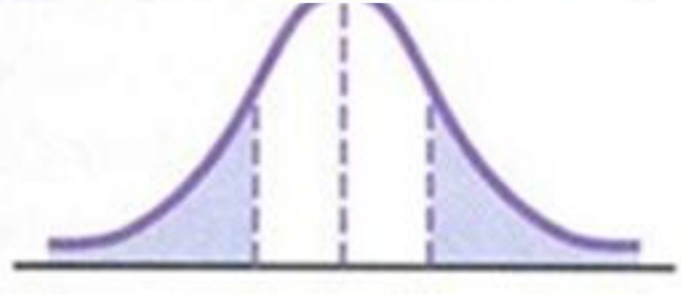
- Left skewed:



- Right skewed:

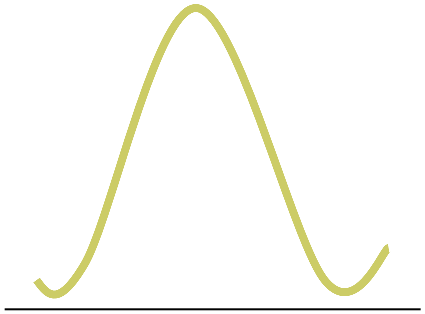


- Symmetric:



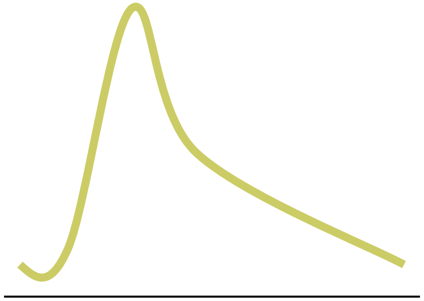
Shapes of the Distribution

- Three common shapes of frequency distributions:



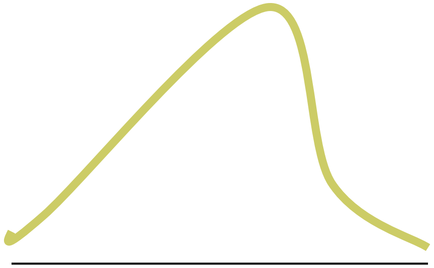
A

Symmetrical and bell shaped



B

Positively skewed or skewed to the right

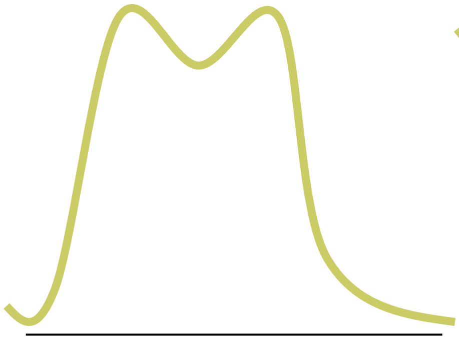


C

Negatively skewed or skewed to the left

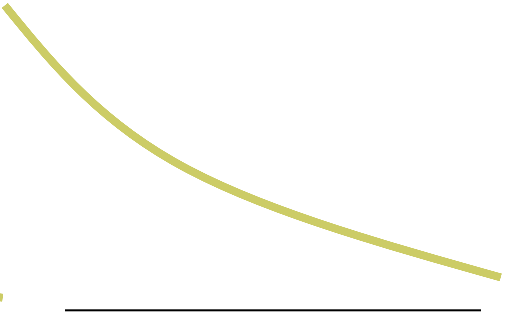
Shapes of the Distribution

- Three less common shapes of frequency distributions:



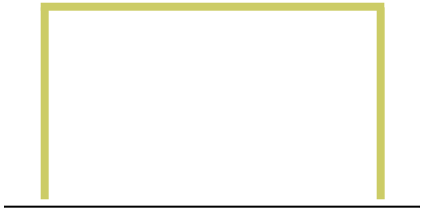
A

Bimodal



B

Reverse J-shaped



C

Uniform

Example: # hours to complete a task

This guy took a VERY long time!

Data (for n=12 employees):

2 3 8 ¦ 8 9 10 ¦ 10 12 15 ¦ 18 22 63

$\overline{X}$= 180/12 = 15 hours

Median = 10 hours

The (extremely slow) employee who took 63 hours to complete the task skewed the entire distributon to the right.

s^2 = 2868 / 11 = 260.79

s = 16.25 hours

CV = 107.7%

Degree of Skewness

- A skewness index can readily be calculated most statistical computer program in conjunction with frequency distributions
- The index has a value of 0 for perfectly symmetric distribution.
- A positive value if there is a positive skew, and negative value if there is a negative skew.
- A skewness index that is more than twice the value of its standard error can be interpreted as a departure from symmetry.

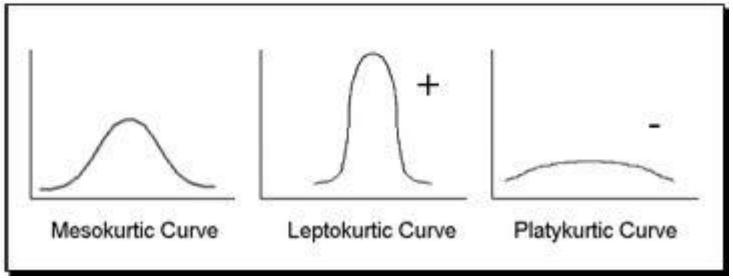
Measures of Skewness or Symmetry

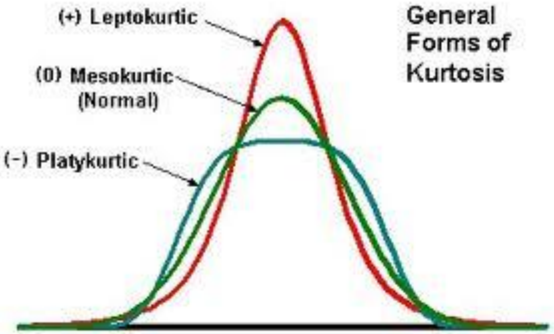
- Pearson's skewness coefficient
  - It is nonalgebraic and easily calculated. Also it is useful for quick estimates of symmetry .
  - It is defined as:

    skewness = mean-median/SD
- Fisher's measure of skewness.
  - It is based on deviations from the mean to the third power.

Kurtosis

- The distribution's kurtosis is concerns how pointed or flat its peak.
- Two types:
  - Leptokurtic distribution (mean thin).
  - Platykurtic distribution (means flat).





Kurtosis

- There is a statistical index of kurtosis that can be computed when computer programs are instructed to produce a frequency distribution
- For kurtosis index, a value of zero indicates a shape that is neither flat nor pointed.
- Positive values on the kurtosis statistics indicate greater peakedness, and negative values indicate greater flatness.

Fishers' measure of Kurtosis

- Fisher's measure is based on deviation from the mean to the fourth power.
- A z-score is calculated by dividing the measure of kurtosis by the standard error for kurtosis.

Graphical Methods

- Frequency Distribution
- Histogram
- Frequency Polygon
- Cumulative Frequency Graph
- Pie Chart.

Presenting Data

- **Table**
  - Condenses data into a form that can make them easier to understand;
  - Shows many details in summary fashion;

BUT

  - Since table shows only numbers, it may not be readily understood without comparing it to other values.

Principles of Table Construction

- Don’t try to do too much in a table
- Use white space effectively to make table layout pleasing to the eye.
- Make sure tables & test refer to each other.
- Use some aspect of the table to order & group rows & columns.

Principles of Table Construction

- If appropriate, frame table with summary statistics in rows & columns to provide a standard of comparison.
- Round numbers in table to one or two decimal places to make them easily understood.
- When creating tables for publication in a manuscript, double-space them unless contraindicated by journal.

Frequency Distributions

- A useful way to present data when you have a large data set is the formation of a frequency table or frequency distribution.

- Frequency – the number of observations that fall within a certain range of the data.

Frequency Table

Age	Number of Deaths
<1	564
1-4	86
5-14	127
15-24	490
25-34	66
35-44	806
45-54	1,425
55-64	3,511
65-74	6,932
75-84	10,101
85+	9825
Total	34,524

Frequency Table

Data Intervals	Frequency	Cumulative Frequency	Relative Frequency (%)	Cumulative Relative Frequency (%)
10-19	5	5		
20-29	18	23		
30-39	10	33		
40-49	13	46		
50-59	4	50		
60-69	4	54		
70-79	2	56		
Total				

Number of Intervals

- There is no clear-cut rule on the number of intervals or classes that should be used.
- Too many intervals – the data may not be summarized enough for a clear visualization of how they are distributed.
- Too few intervals – the data may be over-summarized and some of the details of the distribution may be lost.

Presenting Data

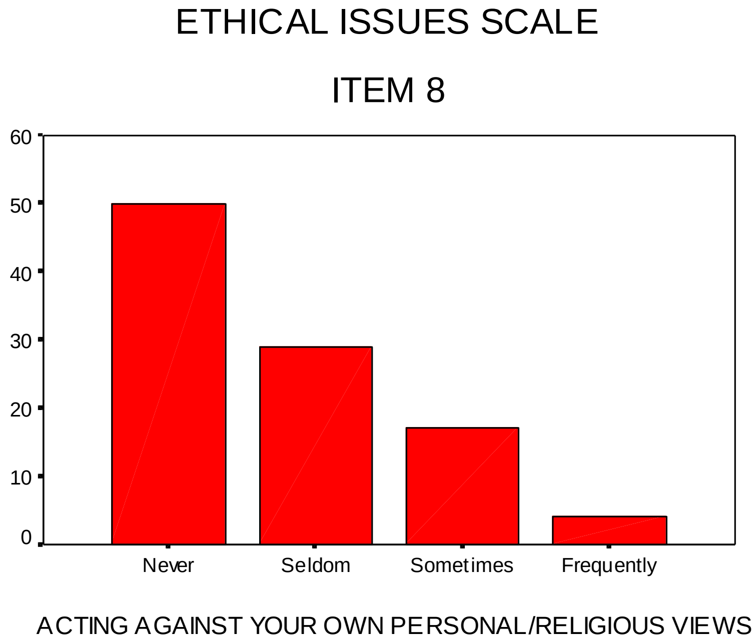
Chart

- Visual representation of a frequency distribution that helps to gain insight about what the data mean.
- Built with lines, area & text: bar charts

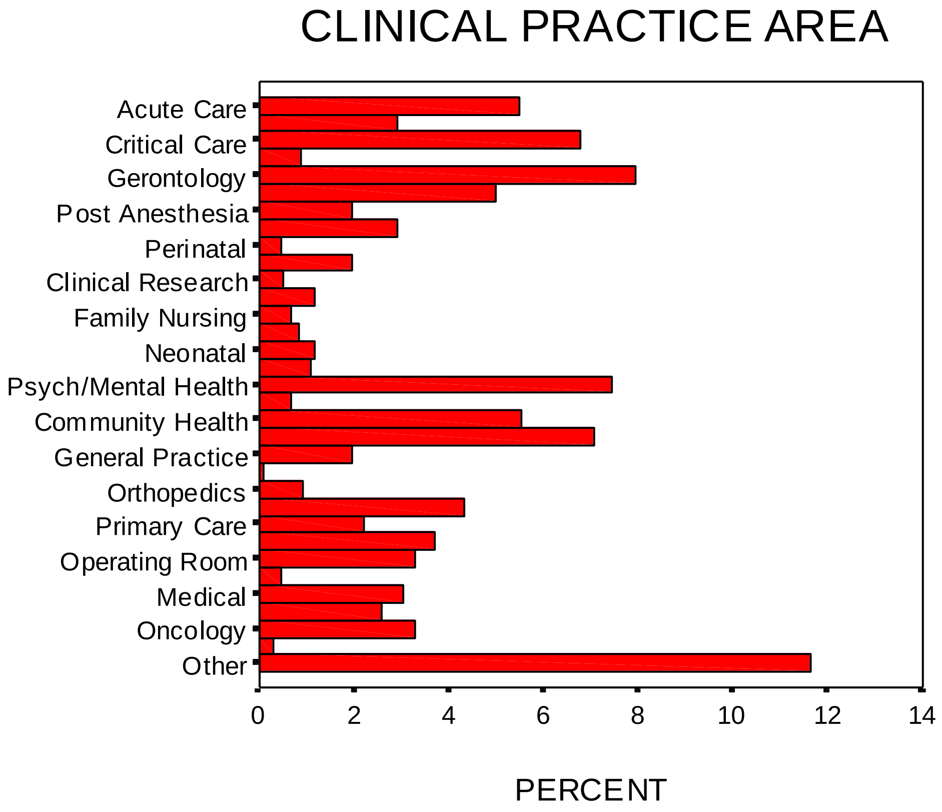
  Ex: bar chart, pie chart

Bar Chart

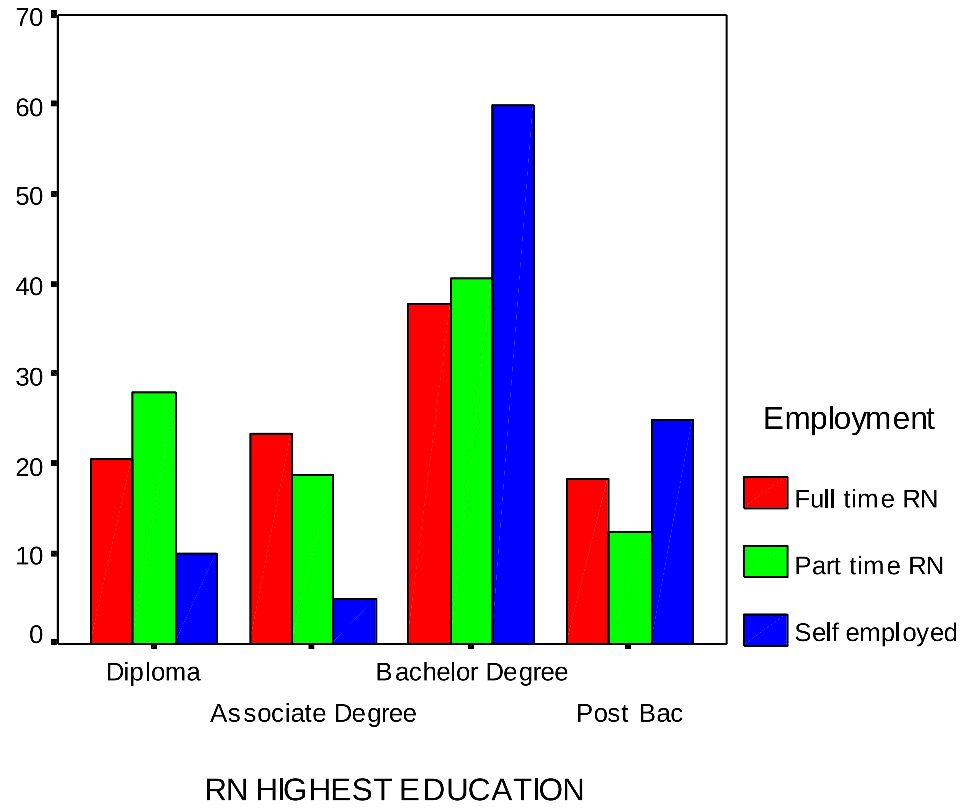
- Simplest form of chart
- Used to display nominal or ordinal data



Horizontal Bar Chart

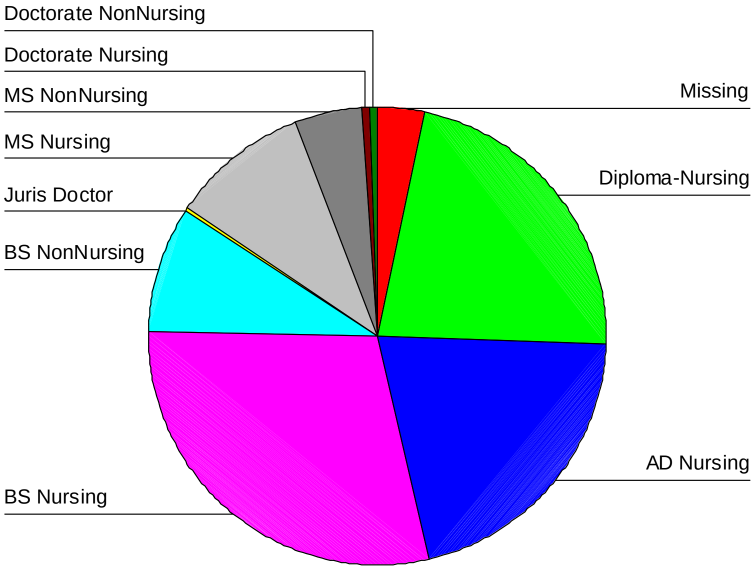


Cluster Bar Chart



Pie Chart

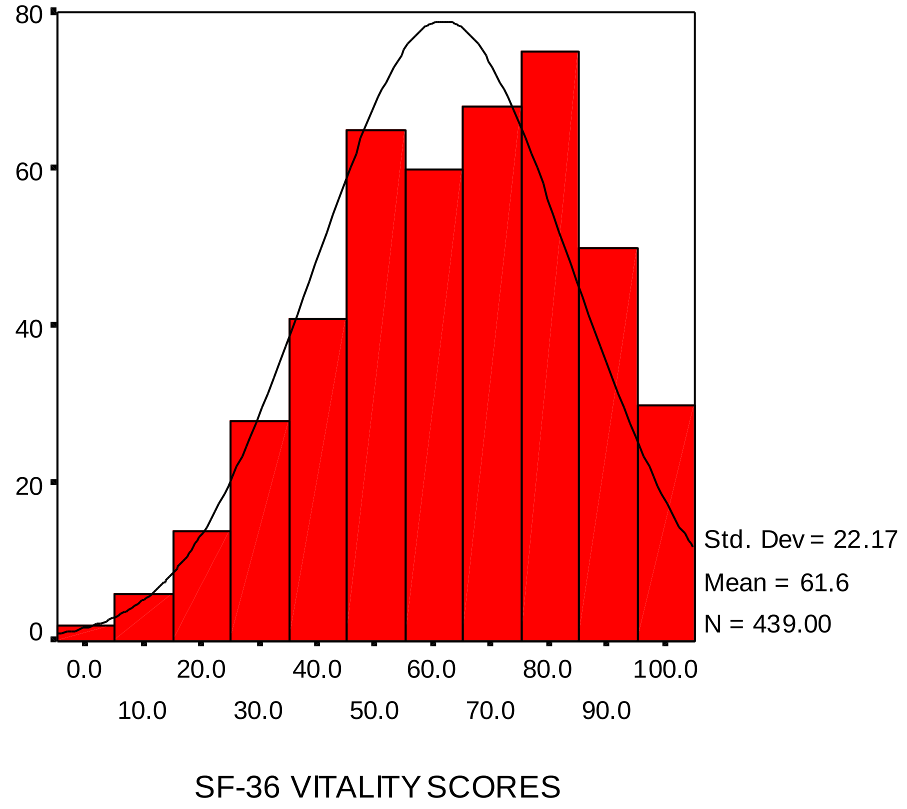
- Alternative to bar chart
- Circle partitioned into percentage distributions of qualitative variables with total area of 100%



Histogram

- Appropriate for interval, ratio and sometimes ordinal data
- Similar to bar charts but bars are placed side by side
- Often used to represent both frequencies and percentages
- Most histograms have from 5 to 20 bars

Histogram

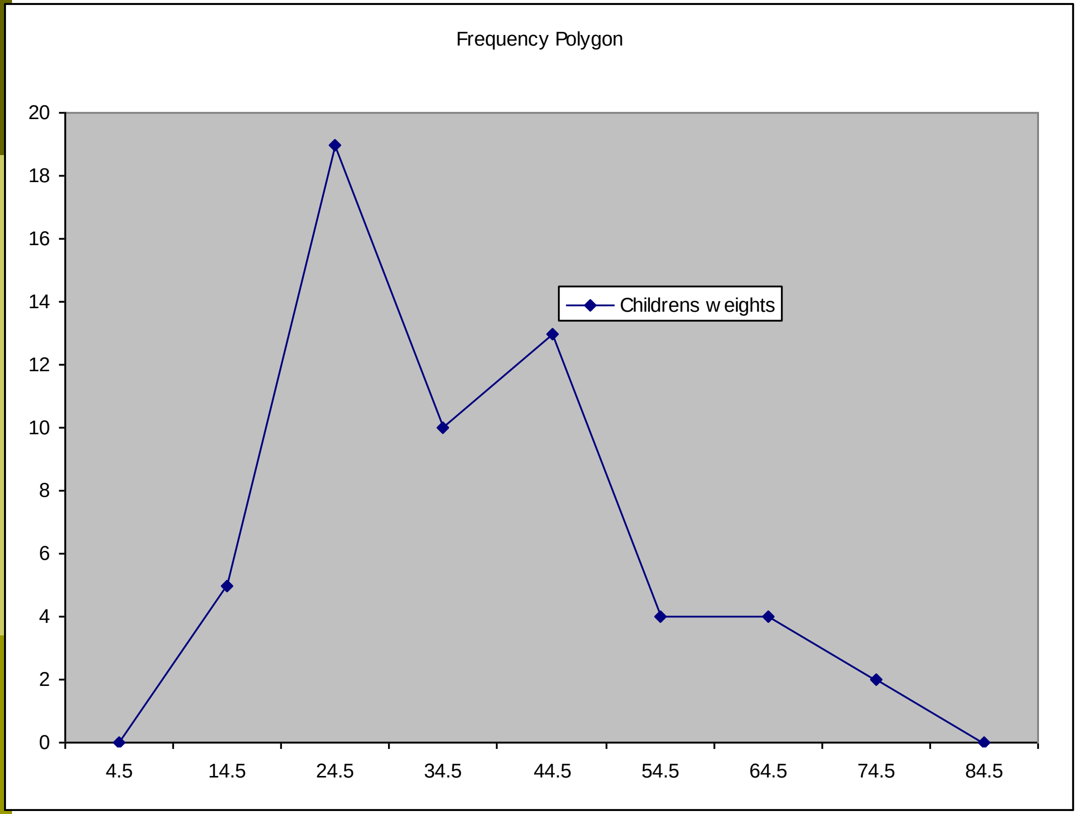

80
60
40
20
0
Std. Dev = 22.17
Mean = 61.6
N = 439.00
0.0
20.0
40.0
60.0
80.0
100.0
10.0
30.0
50.0
70.0
90.0
SF-36 VITALITY SCORES

Frequency Polygon

- The ***frequency polygon*** is a line graph. It is made by connecting the top center points of each of the bars.

- *The ends of the line must be anchored on the x-axis.* This requires an additional class interval with a value of 0 (zero) at each end of the table of class intervals.

Frequency Polygon



- First place a dot at the midpoint of the upper base of each rectangular bar.
- The points are connected with straight lines.
- At the ends, the points are connected to the midpoints of the previous and succeeding intervals (these intervals have zero frequency).

Hallmarks of a Good Chart

- Simple & easy to read
- Placed correctly within text
- Use color only when it has a purpose, not solely for decoration
- Make sure others can understand chart; try it out on somebody first
- Remember: A poor chart is worse than no chart at all.

Outliers

- Are values that are extreme relative to the bulk of scores in the distribution.
- They appear to be inconsistent with the rest of the data.
- Advantages:
  - They may indicate characteristics of the population that would not be known in the normal course of analysis.
- Disadvantages:
  - They do not represent the population
  - Run counter to the objectives of the analysis
  - Can distort statistical tests.

Sources of Outliers

- An error in the recording of the data.
- A failure of data collection, such as not following sample criteria (e.g. inadvertently admitting a disoriented patient into a study), a subject not following instructions on a questionnaire, or equipment failure.
- An actual extreme value from an unusual subjects.

Methods to Identify Outliers

- Traditional way of labeling outliers, any value more than 3SD from the mean.

Handling Outliers

- Analyze the data two ways:
  - With the outliers in the distribution
  - With outliers removed.
- If the results are similar, as they are likely to be if the sample size is large, then the outliers may be ignored.
- If the results are not similar, then a statistical analysis that is resistant to outliers can be used (e.g. median and IQR).
- If you want to use a mean with outliers, then the trimmed mean is an option. If calculated with a certain percentage of the extreme values removed from both ends of the distribution (e.g. n=100, then 5% trimmed mean is the mean of the middle 90% of the observation).

Missing Data

- Any systematic event external to the respondent (such as data entry errors or data collection problems) or action on the part of the respondent (such as refusal to answer) that leads to missing data.
- It means that analyses are based on fewer study participants than were in the full study sample. This, in turn, means less statistical power, which can undermine statistical conclusion validity-the degree to which the statistical results are accurate.
- Missing data can also affect internal validity-the degree to which inferences about the causal effect of the dependent variable on the dependent variable are warranted, and also affect the external validity-generalizability.

Strategies to avoid Missing Data

- Persistent follow-up
- Flexibility in scheduling appointments
- Paying incentives.
- Using well-proven methods to track people who have moved.
- Performing a thorough review of completed data forms prior to excusing participants.

Techniques for Handling Missing Data

- Deletion techniques. Involve excluding subjects with missing data from statistical calculation.
- Imputation techniques. Involve calculating an estimate of each missing value and replacing, or imputing, each value by its respective estimate.
- Note: techniques for handling missing data often vary in the degree to which they affect the amount of dispersion around true scores, and the degree of bias in the final results. Therefore, the selection of a data handling technique should be carefully considered.

SHOKRAN