

# Describing Data with Charts

Gbenga Ogunfowokan  
Lead, Nigerian Regional Faculty  
The Global Health Network  
19<sup>th</sup> May 2017

# Preparation for Data description

- Data analysis must be done before the summary can be described using charts.
- Before data analysis is done, ensure
- 1) Data processing: Collecting, Collating, Entry. The process must be carefully done to avoid error.
- 2) Data Editing- to detect and correct all errors.
- 3) Data Analysis using statistical software

# Data Description

- **To describe / interpret the data, consider the following.**
- **Shape** of the Distribution
  - Symmetry
  - Modality: most frequently occurring value
  - Unimodal or bimodal or uniform
  - Skewness
  - Centrality – mid range of values
- **Spread** – range of values
- Extreme values - outliers

# Describing Data with graphs

## **Basic Principles for Constructing All Plots**

- Data should stand out clearly from background.
- Keep the graph as simple as possible.
- The information should be clearly labeled and include:
  - title
  - axes, bars, pie segments, etc. - include units that are needed to interpret data
  - axis labels

# Tips

- scale including starting points. The vertical axis will typically begin at 0.
- Sources of data should be identified, as appropriate.
- Do not clutter the graphs with unnecessary information and graphical components that are really not necessary.
- Do not put too much information or data on one graph.
- Sometimes, you have to try several approaches before selecting an appropriate graph.

# Some practical advice for constructing graphs includes the following.

- Every bit of ink on a graphic requires a reason. And nearly always that reason should be that the ink presents new information. In most cases, non-data ink clutters up the data. Avoid content-free decoration, including chart junk.
- Type should be clear, precise, and modest. Usually - type in upper and lower case.
- The grid should usually be muted or completely suppressed so that its presence is only implicit - lest it compete with the data. Dark lines are chart junk. They carry no information, clutter up the graphic and generate graphic activity unrelated to data information.

# Tips

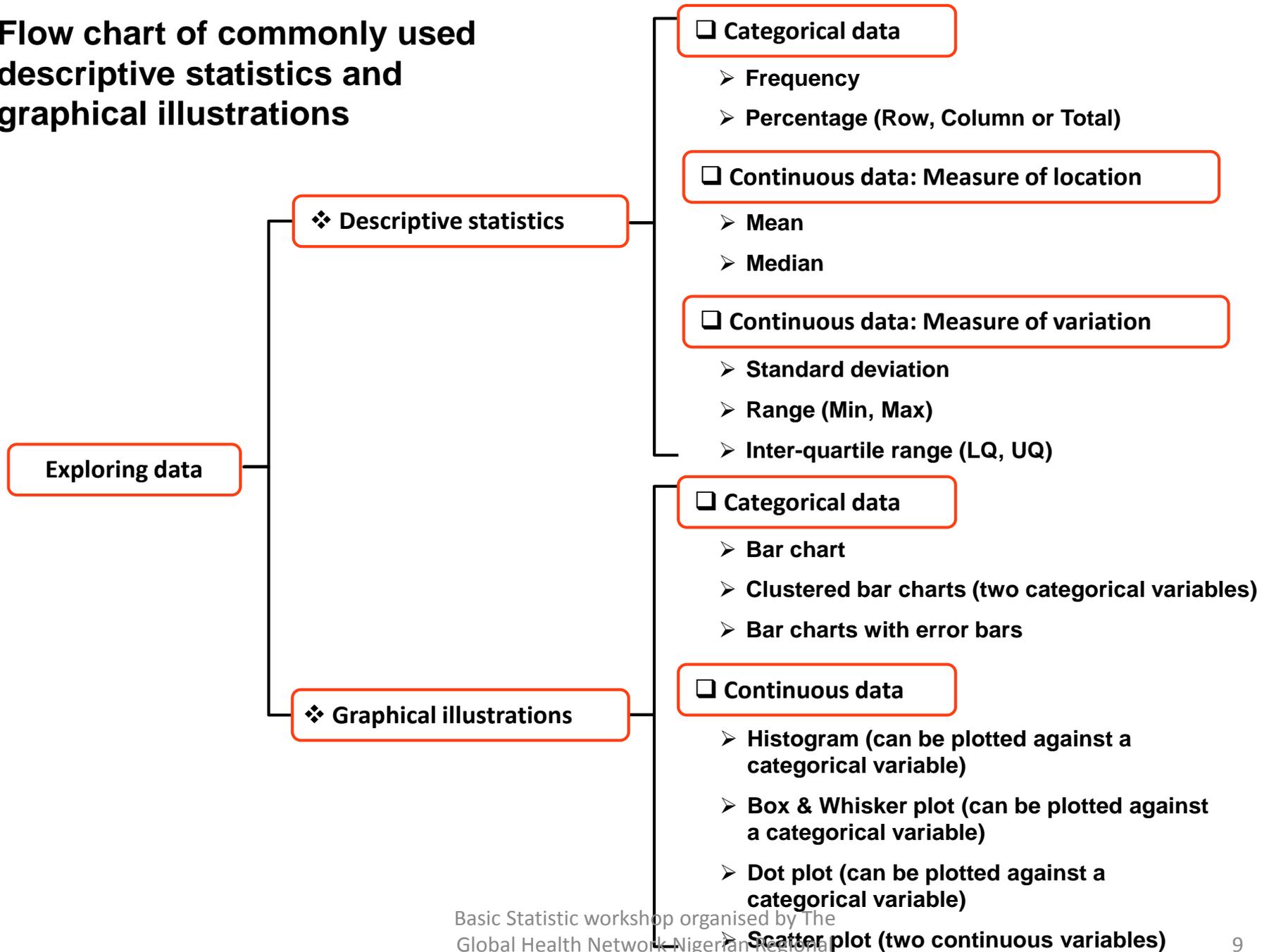
- The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the numerical quantities represented.
- Clear, detailed, and thorough labeling should be used to defeat graphical distortion and ambiguity. Label important events in the data.
- **Show data variation, not design variation.**
- Graphical elegance is often found in simplicity of design and complexity of data

# Tips

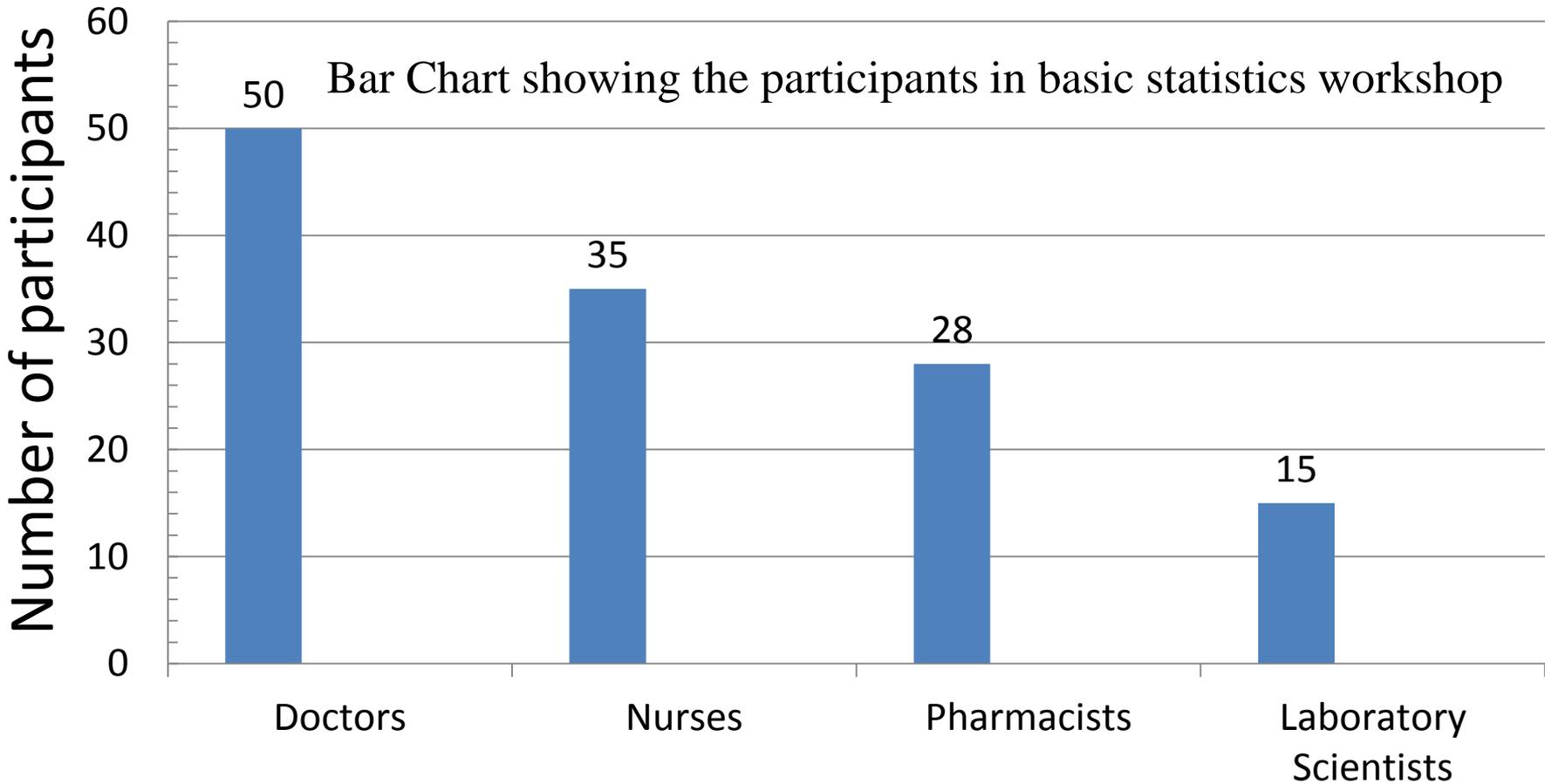
## **In interpreting graphs, consider:**

- Horizontal and vertical scales; what is the relationship - are the distances between, for example, 10 and 20, the same on each axis?
- The center point - of particular importance in comparing two histograms. Look at the starting point of the vertical scale - does it start at 0? How could this affect the interpretation of the graph

# Flow chart of commonly used descriptive statistics and graphical illustrations



# Presenting categorical data using Excel



Category of participants

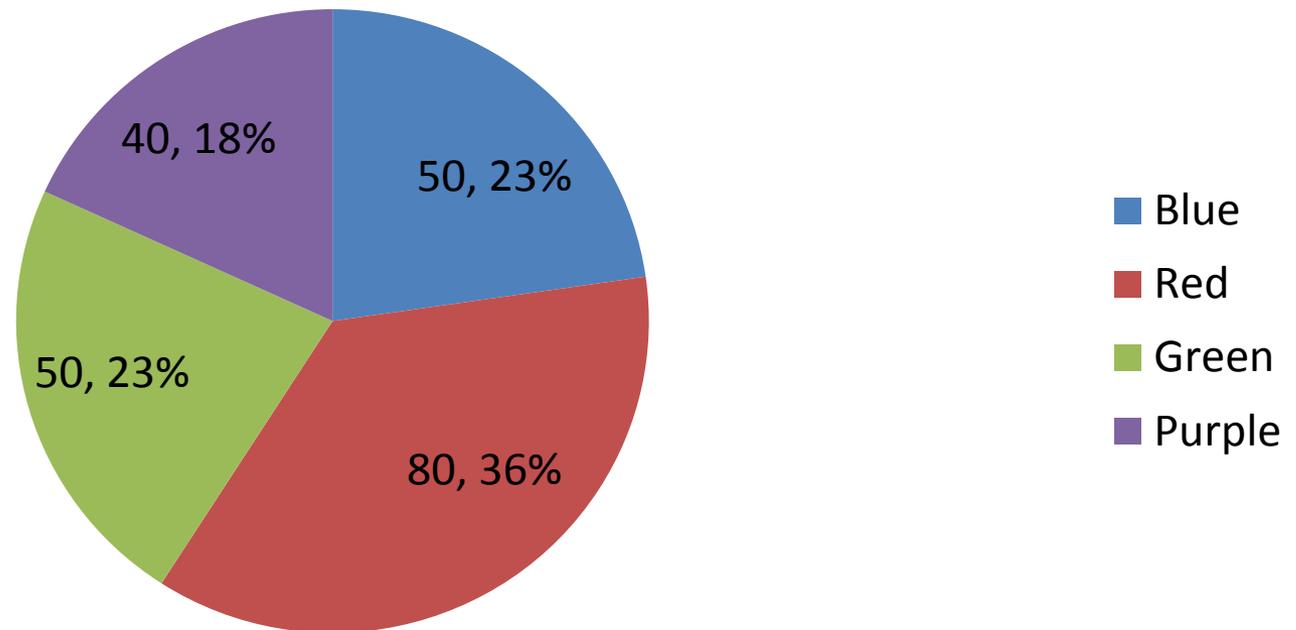
Basic Statistics workshop organised by The  
Global Health Network Nigerian Regional  
Faculty

# Bar Charts

- Bar charts also show percentages in various categories and allow comparison between categories.
- The vertical scale is frequencies, relative frequencies, or percentages.
- The horizontal scale shows categories.
- Consider the following in constructing bar charts.
- all boxes should have the same width
- leave gaps between the boxes (because there is no connection between them)
- boxes can be in any order.
- Bar charts can be used to represent two categorical variables simultaneously
- As presented above, the bar chart is also called a **Pareto** chart because the vertical bars are plotted in descending order by frequency.

# Categorical variables

## Colour preference of participants in basic statistic workshop



# Pie Charts

- A circle is divided proportionately and shows what percentage of the whole falls into each category.
- The size of each slice of the pie varies according to the percentage in each category.
- These charts are simple to understand. They convey information regarding the relative size of groups more readily than does a table.

# Graphs for Measured/Continuous Quantitative Data

- Stem and Leaf
- Histograms
- Percentage Polygons
- XY Scatter Charts (2 variables)
- Line Graphs (e.g. time series)
- Box plots

# Stem and Leaf

## Stem and Leaf Plots

A stem-and-leaf plot puts data into groups (called stems) so that the values within each group (the leaves) branch out to the right on each row. The advantage of a stem and leaf plot is that it utilizes the data as a part of the graph.

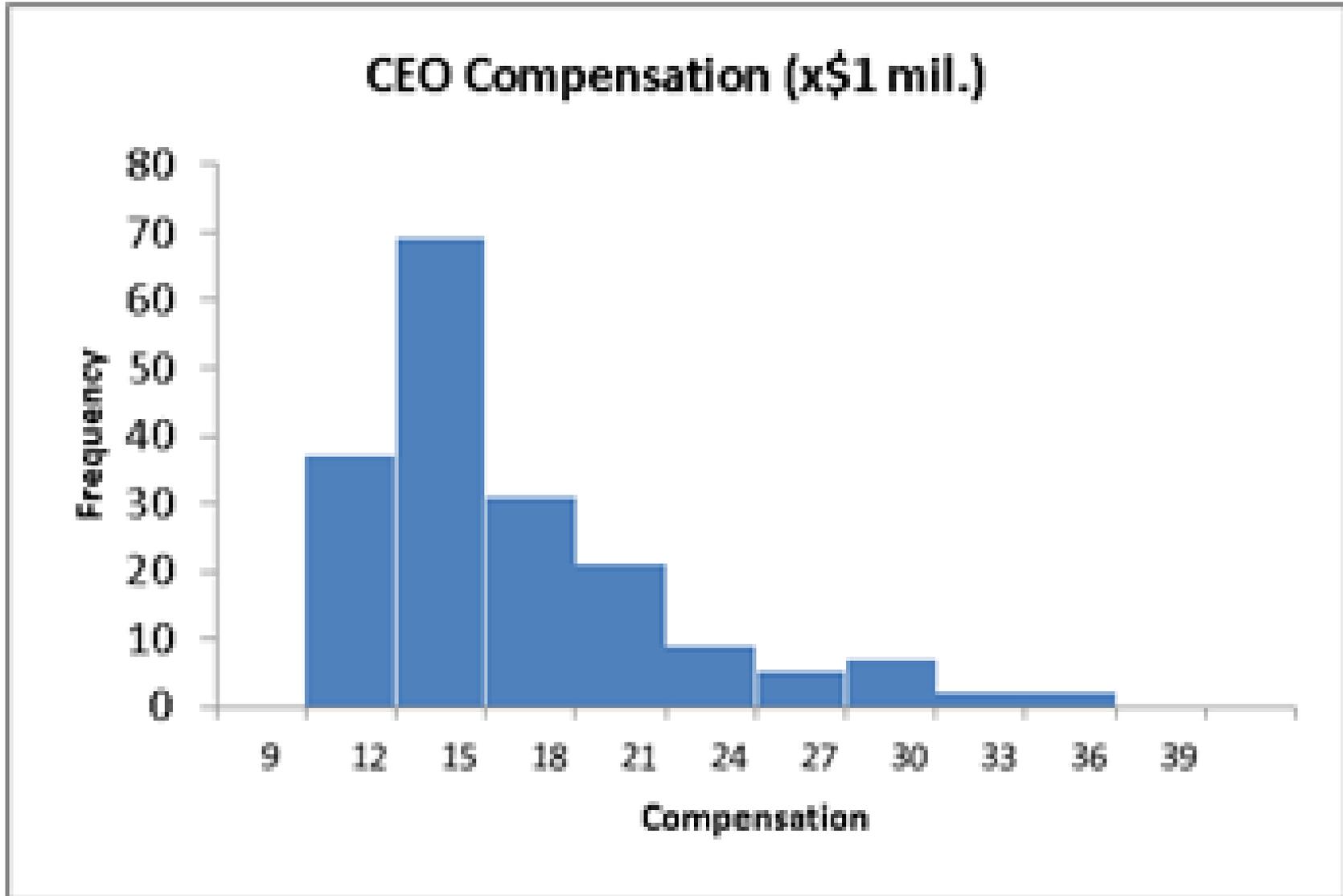
| CEO Compensation (x\$1mil.) |                               |
|-----------------------------|-------------------------------|
| Stem                        | unit 1                        |
| 10                          | 9999                          |
| 11                          | 0011112223333444556667899999  |
| 12                          | 00000111122223334444455678999 |
| 13                          | 000000222334444456777888999   |
| 14                          | 011223355577888999            |
| 15                          | 114666778                     |
| 16                          | 0111223345568                 |
| 17                          | 023567999                     |
| 18                          | 1145678999                    |
| 19                          | 0245                          |
| 20                          | 25577                         |
| 21                          | 00279                         |
| 22                          | 59                            |
| 23                          | 0124                          |
| 24                          | 2                             |
| 25                          | 1223                          |
| 26                          |                               |
| 27                          |                               |
| 28                          | 4                             |
| 29                          | 334569                        |
| 30                          |                               |
| 31                          | 47                            |

Stem →

← leaf

Note the first line. The first stem is 10. It is followed by four leaves, each 9. This means that the original data has four values of 10.9.

# Histograms



# Histogram

- Histograms show the frequency distributions of continuous variables. They look similar to Bar Charts, but they are drawn without gaps between the bars because the x-axis is used to represent the class intervals (on a continuum).  
**However, many of the current software packages do easily not make this distinction (e.g. Excel).**
- The data is divided into non-overlapping intervals (usually use from 5 to 15).
- Intervals generally have the same length
- The number of values in each interval is counted (the class frequency).
- Sometimes relative frequencies or percentages are used. (Divide the cell total by the grand total.)

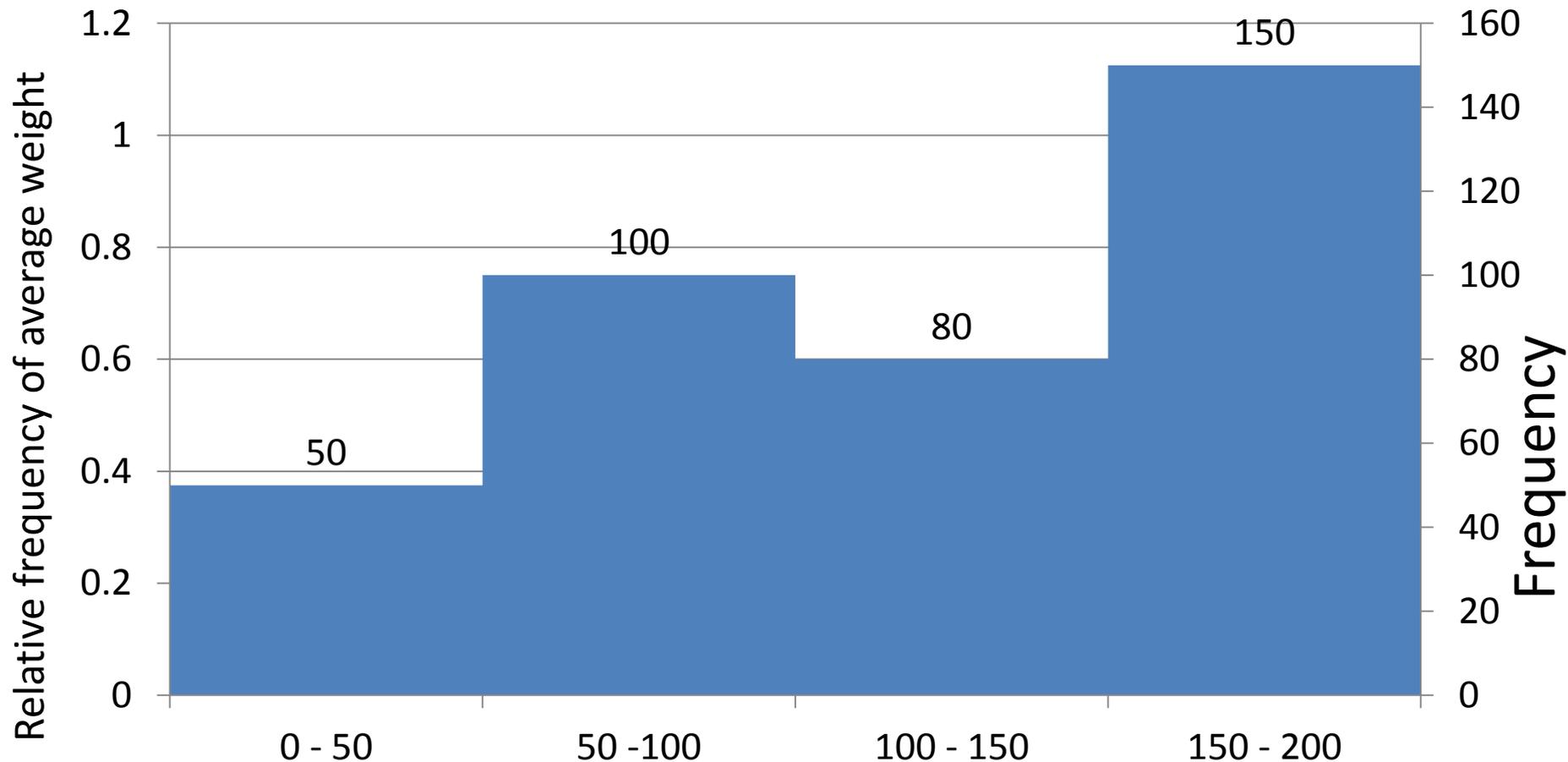
# Histograms

- Rectangles are drawn over each interval. (The area of rectangle = relative frequency of the interval. If intervals are not all of the same length then heights have to be scaled so that each area is proportional to the frequency for that interval. )
- Shifts in data concentration may show up when different class boundaries are chosen. As the size of the data set increases, the impact of alterations in the selection of class boundaries is greatly reduced
- When comparing two or more groups with different sample sizes, you must use either a relative frequency or a percentage distribution

# Histograms

- **Note:** XL does not give mid points; it uses bins – which represent a range of values.
- The upper boundary of a bin is explicitly given – no value in the bin exceeds the upper boundary.
- All the values in the bin are greater than the lower boundary.

# Histogram of Height range of participants and their average weight



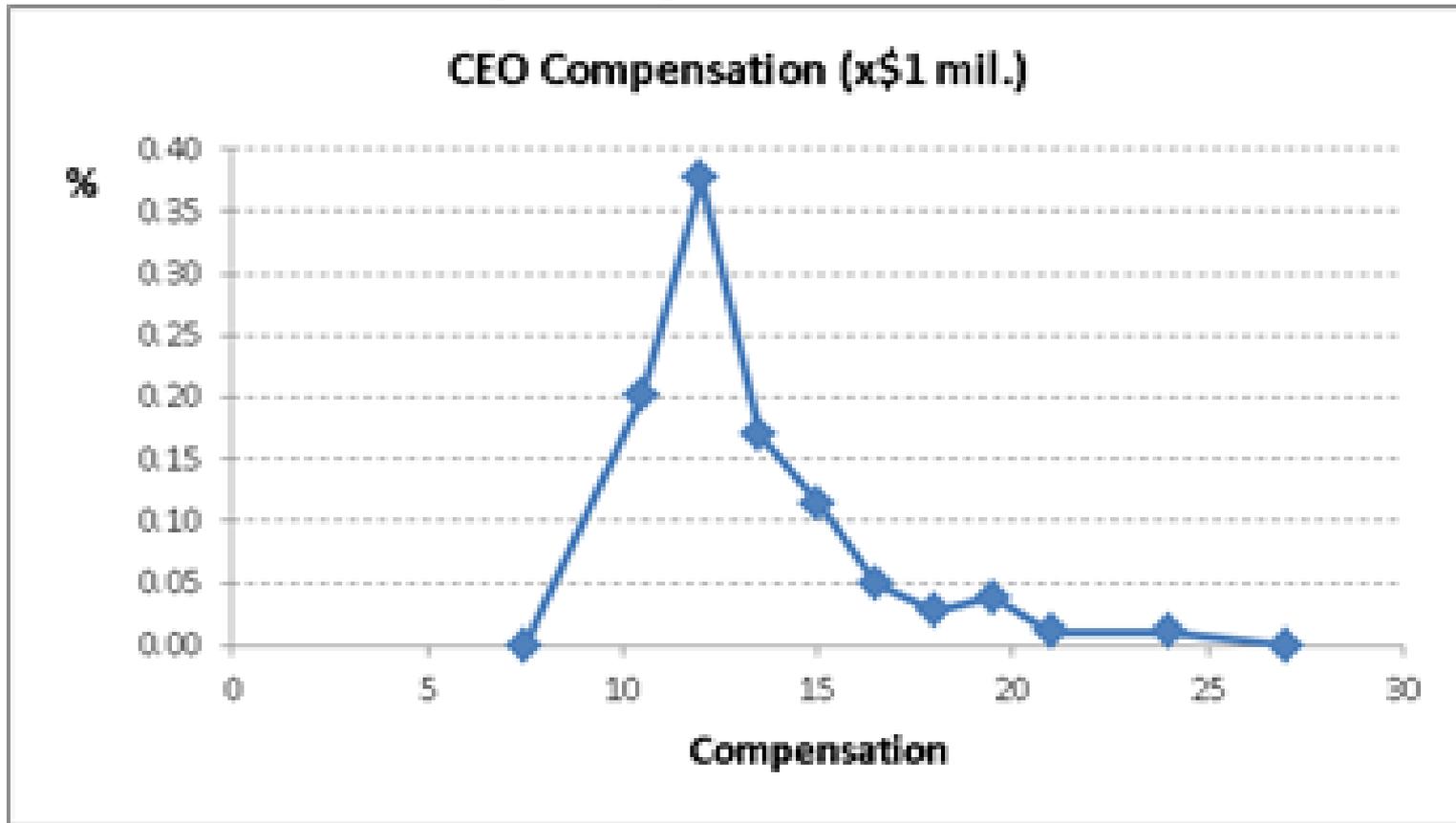
Height range of participants in cm

Basic Statistics workshop organised by The  
Global Health Network Nigerian Regional

Faculty

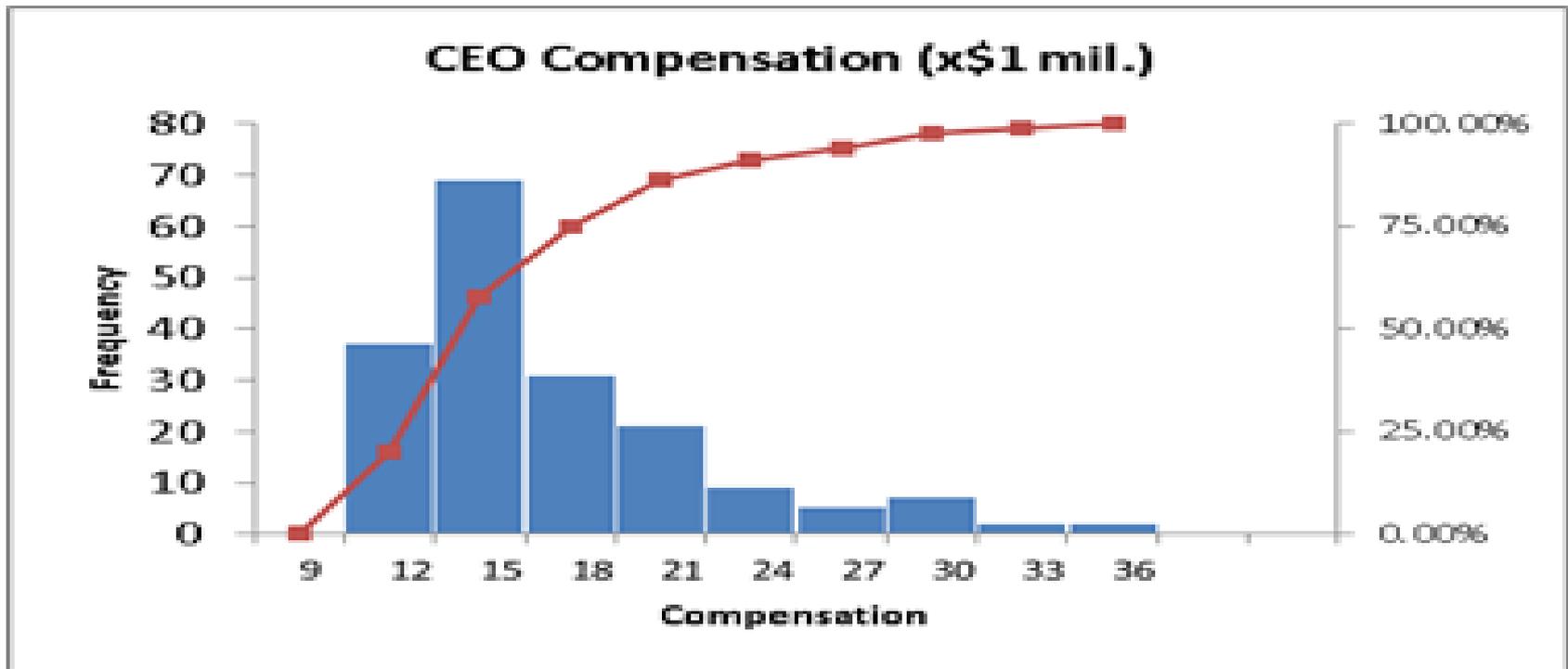
# Frequency Polygon for CEO Compensation

Use midpoints to represent the data.



# Ogive for CEO Compensation

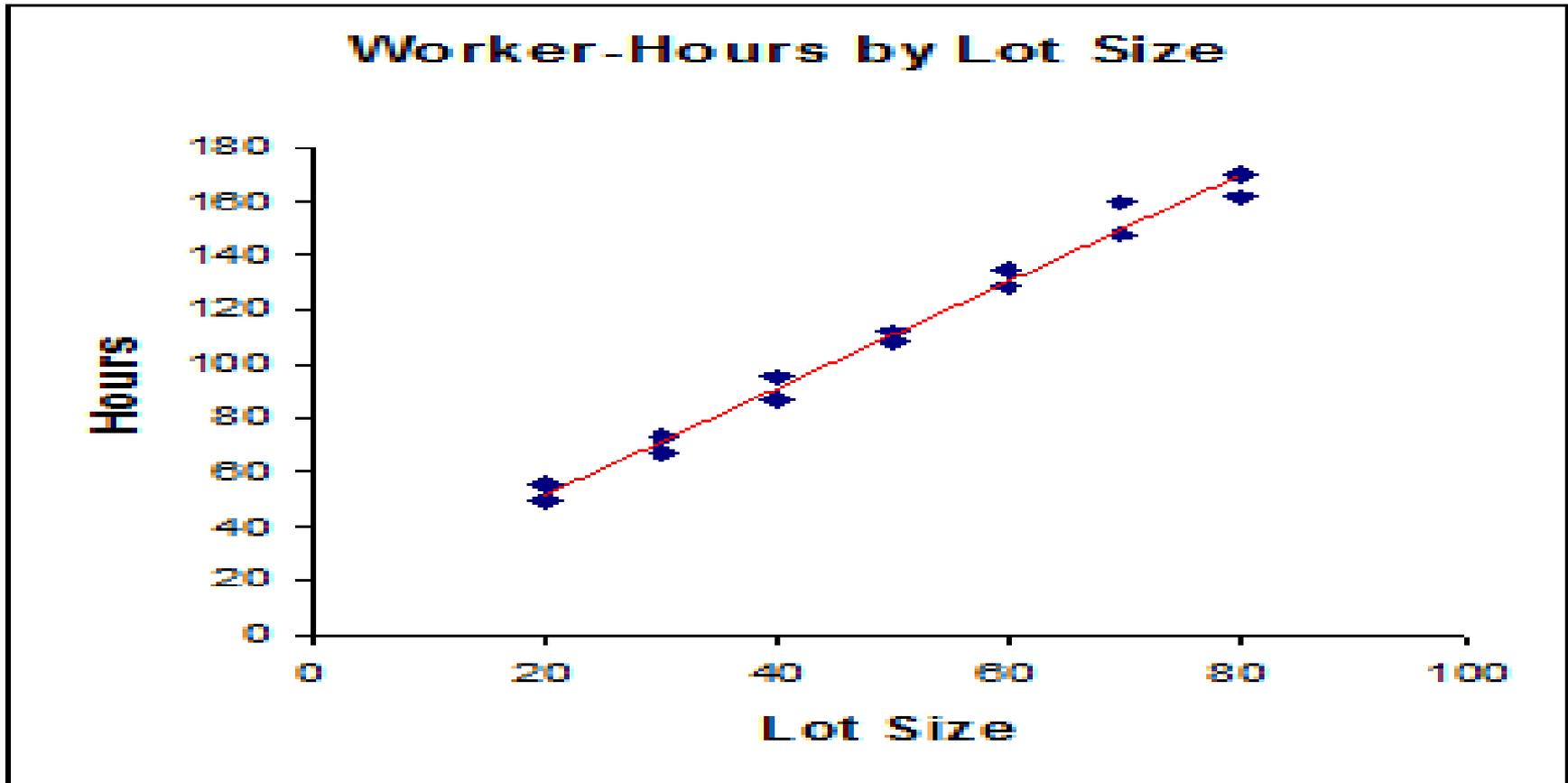
Cumulative percentages are plotted along the Y axis.



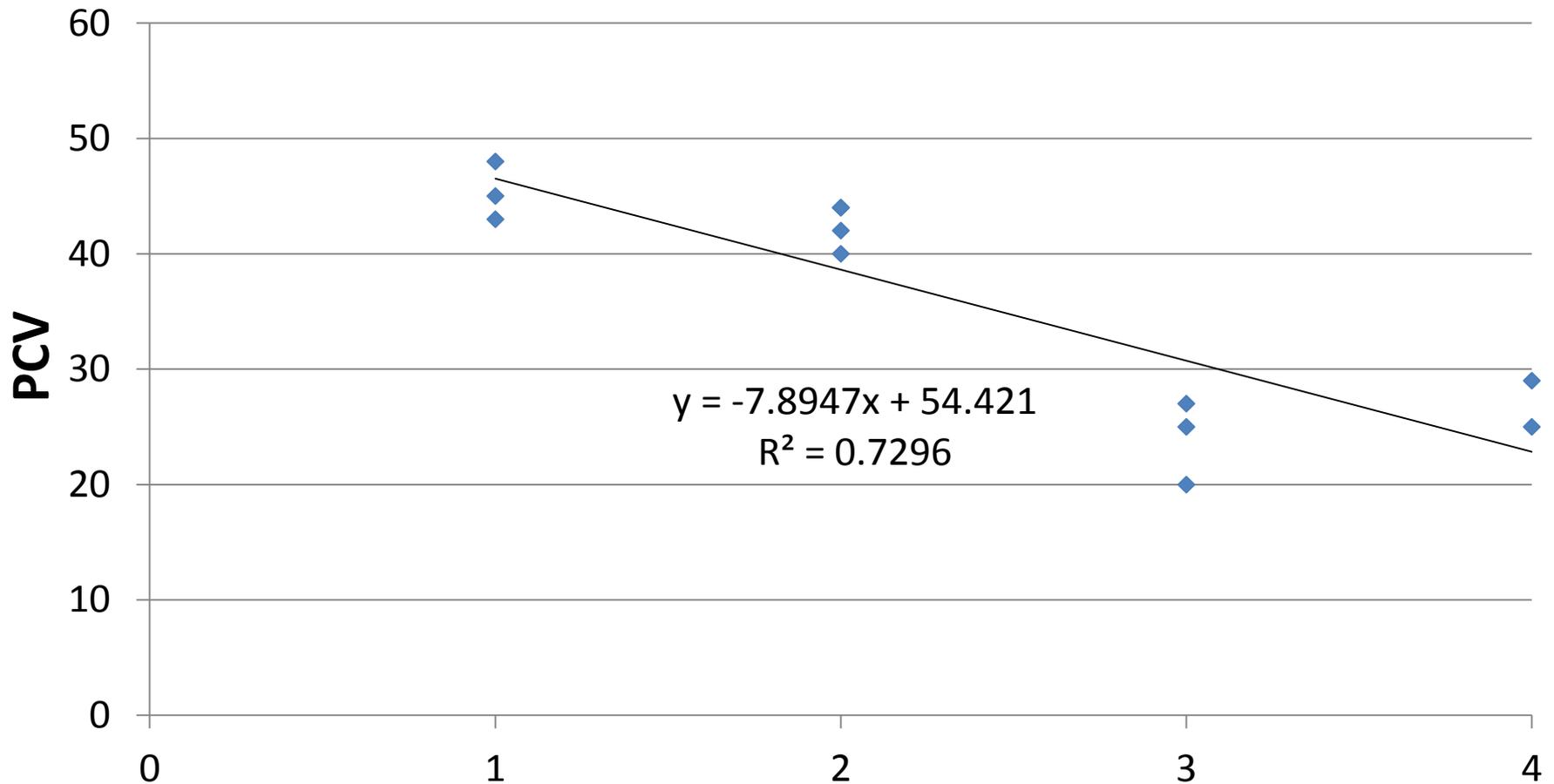
# XY Scatter Chart

- This type of chart should be used with two variables when both of the variables are quantitative and continuous.
- Plot pairs of values using the rectangular coordinate system to examine the relationship between two values.

# XY Scatter Chart



# Scattered plot of the correlation between malaria parasite count and PCV of U5 children



# Scatter plot

- **Measure of Relationships between Two Quantitative Variables**
- 
- **Correlation**
- 
- Correlation ( $r$ ) is used in describing the strength of the relationship between two (or more) variables.
- $r$  can vary from a low of  $-1$  (perfect negative correlation) to  $+1$  (perfect positive relationship). A value of  $0$  means there is no correlation
- Correlation coefficients reflect whether the relationship between variables is:
- positive (i.e. as one variable increases, the other variable increases)  
or

# Scatter Plot

- negative (i.e. as one variable increases, the other variable decreases).
- It also may indicate that there is no relationship.
- There are many different types of correlation coefficients and selection of the appropriate one depends on the variables. We will consider Pearson Product-moment Correlation Coefficient which assumes continuous quantitative data.

# Scatter Plot

- Borg and Gall, *Educational Research* from Longman Publishing, provide the following information for **interpreting correlation coefficients**.
- Correlations coefficients ranging from 0.20 to 0.35 show a slight relationship between the variables; they are of little value in practical prediction situations.

# Scatter Plot

- With correlations around 0.50, crude group prediction may be achieved. In describing the relationship between two variables, correlations that are this low do not suggest a good relationship.
- Correlations coefficients ranging from 0.65 to 0.85 make possible group predictions that are accurate enough for most purposes. Near the top of this correlation range, individual predictions can be made that are more accurate than would occur if no such selection procedure were used.
- Correlations coefficients over 0.85 indicate a close relationship between the two variables.

# Scatter Plot

- It is important to understand that **even a high correlation coefficient does not establish a cause and effect relationship.** There may be other factors that relate to both of the variables.
- **Line of Best Fit and Other Considerations**
- It is always good to look at an **XY scatter plot** to see what you think about the relationship between the variables.

# Scatter Plot

- In comparing two variables, you can take the square root of the correlation to get the correlation coefficient; this measure gives the percent of variation in the dependent variable that is 'explained' by the independent variable.

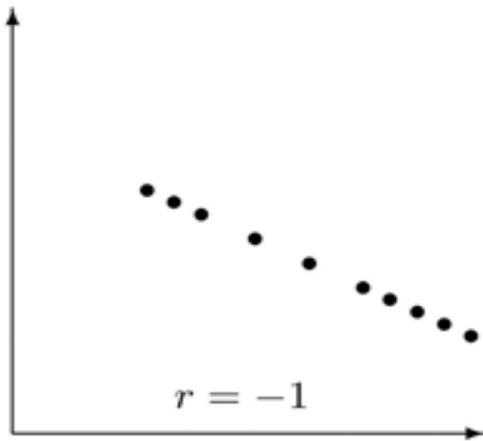
# Scatter Plot

- Excel will not only give you a correlation coefficient, but it will also give you the equation for the **Least Square line** which can be useful in describing the relationship between the two variables and in making predictions of the dependent variable from the independent variable. Note the **slope** of the line; it tells how much the y value changes for each unit change in X.
- .

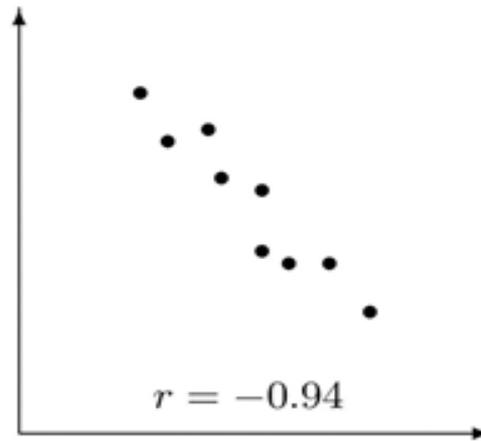
# Scatter Plot

- Note that in making predictions of  $y$  based on  $x$ , stay close to the data set in your selection of  $x$ ; the function may not look the same outside of the given data range

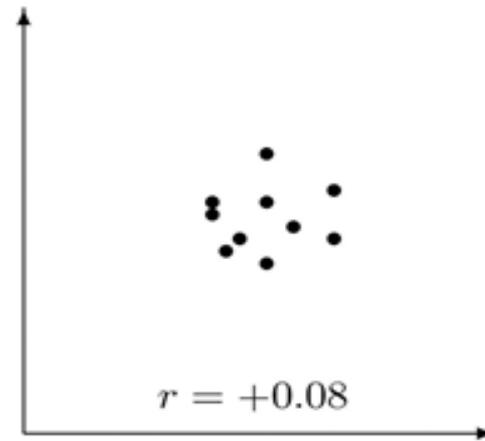
# Scatter Plot



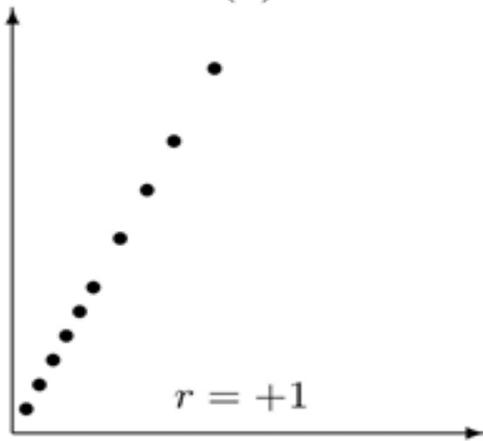
(a)



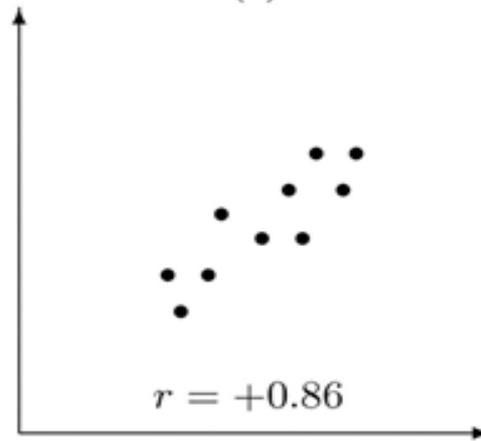
(b)



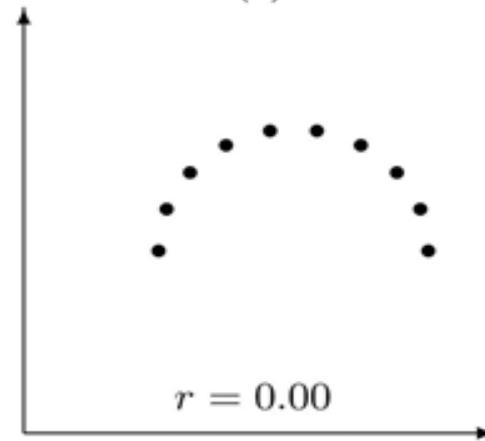
(c)



(d)



(e)



(f)

# Scatter Plot

- In Excel, use the function wizard to find the **correlation coefficient**:
- 
- `CORREL(A2:A16,B2:B16)` g *insert/highlight the correct range*

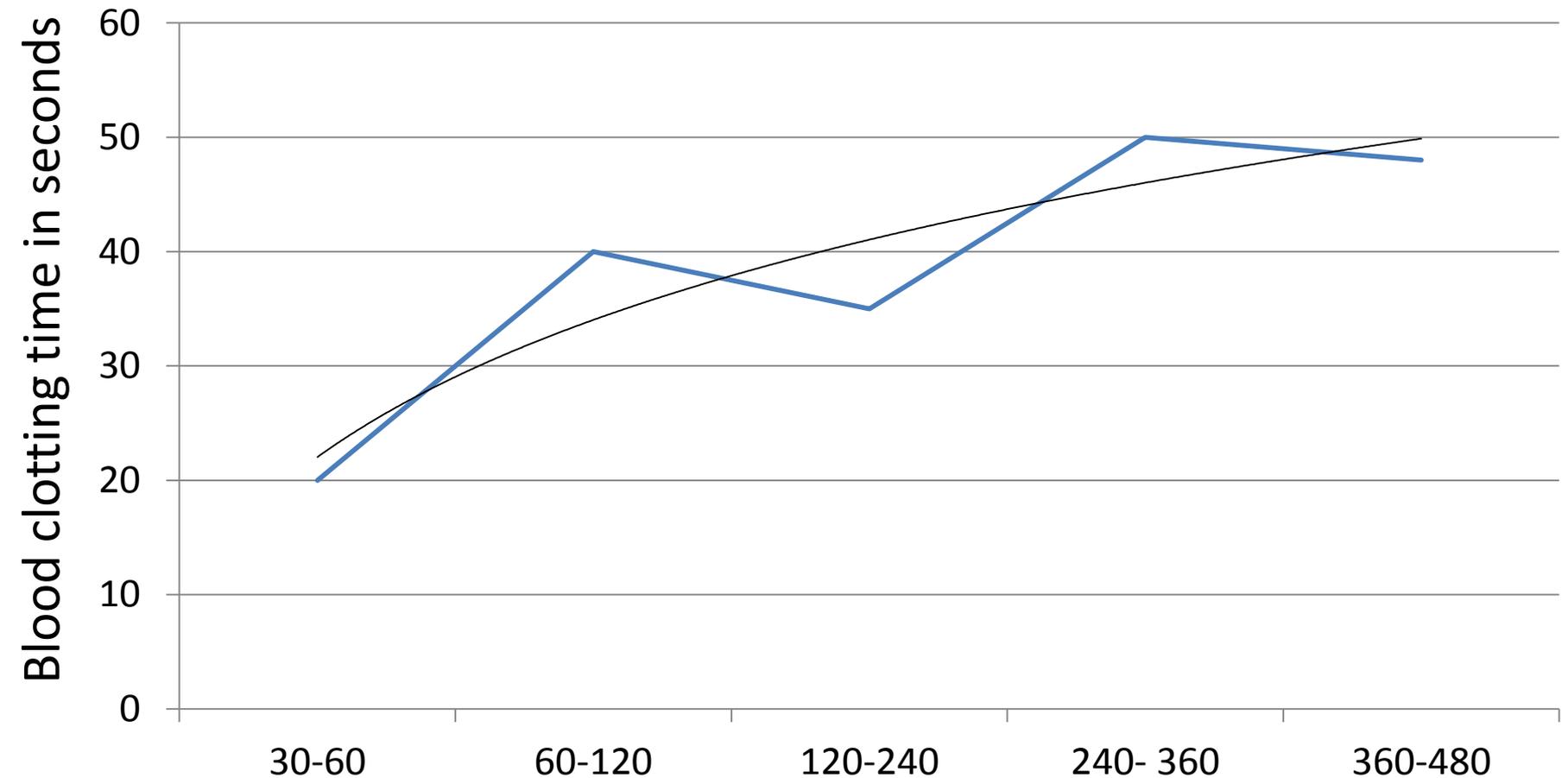
# For Discussion

- Would you expect the correlation between engine size and gas mileage to be positive or negative? Why?
- The following plot shows the relationship between a test for employment (Score 1) and the results of a test given after training (Score 2). Interpret - Consider factors such as slope, coefficient of determination, and correlation.

# A Line Chart

- is similar to the scatter chart; however, it can be used when the values of the independent variable (shown on the horizontal axis) are ranked values (i.e. they do not have to be continuous variables). It is also used for time series plots.

# Blood clotting time of patients bitten by viper by bite-to-hospital time



1. In the following examples, what features of the 'Good Presentation' make it better than the 'Bad Presentation'?

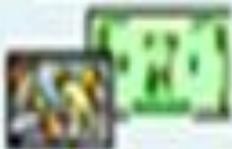


**Bad Presentation**

**Minimum Wage**



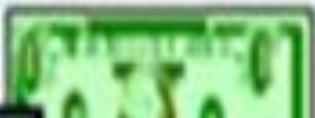
1960: \$1.00



1970: \$1.60



1980: \$3.10



1990: \$3.80



**Good Presentation**

**Minimum Wage**



Basic Statistic workshop organised by The  
Global Health Network Nigerian Regional

Faculty

# Answer

- **Answer to #1**
- **Useless Information** – Don't show decimals if they are not needed
- **Poor Alignment** – Make sure alignment makes sense
  - Don't center numbers, always right justify – try to align decimal points
  - Consider the appropriate placement of row titles
  - Headers are left justified – align with the numbers?
- **Difficult to Read** – Use commas used when the number exceeds a thousand

## 2. In the following examples, what features of the 'Good Presentation' make it better than the 'Bad Presentation'?



### Bad Presentation

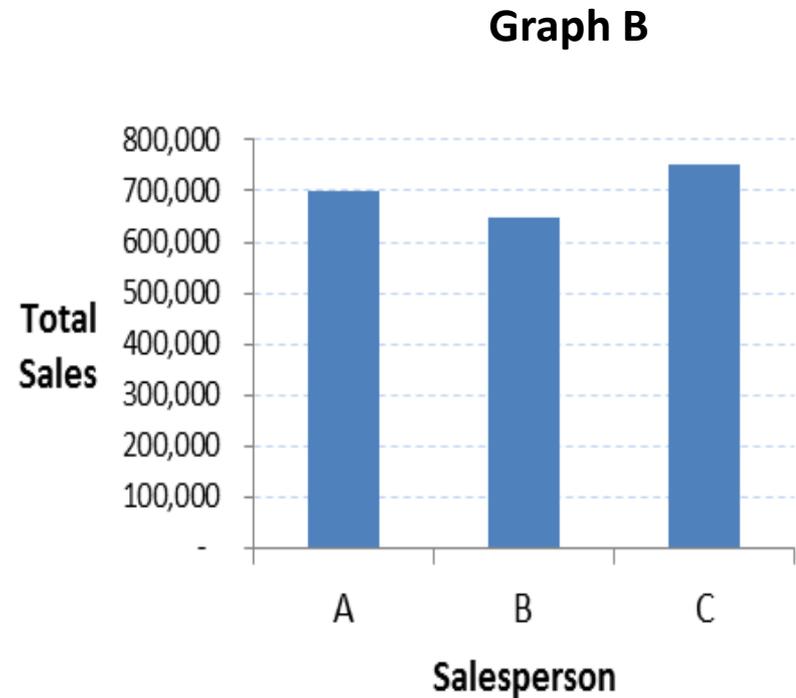
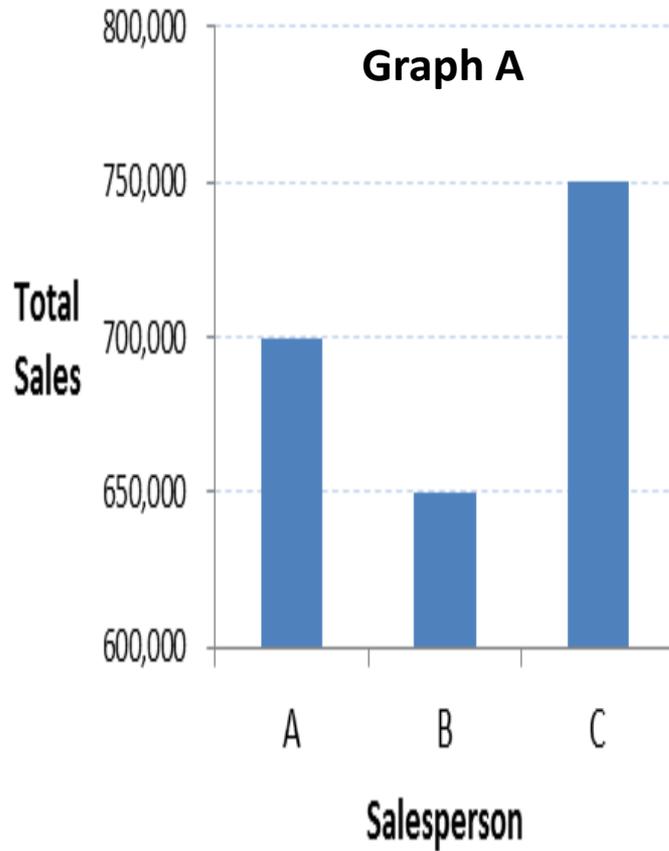


### Good Presentation

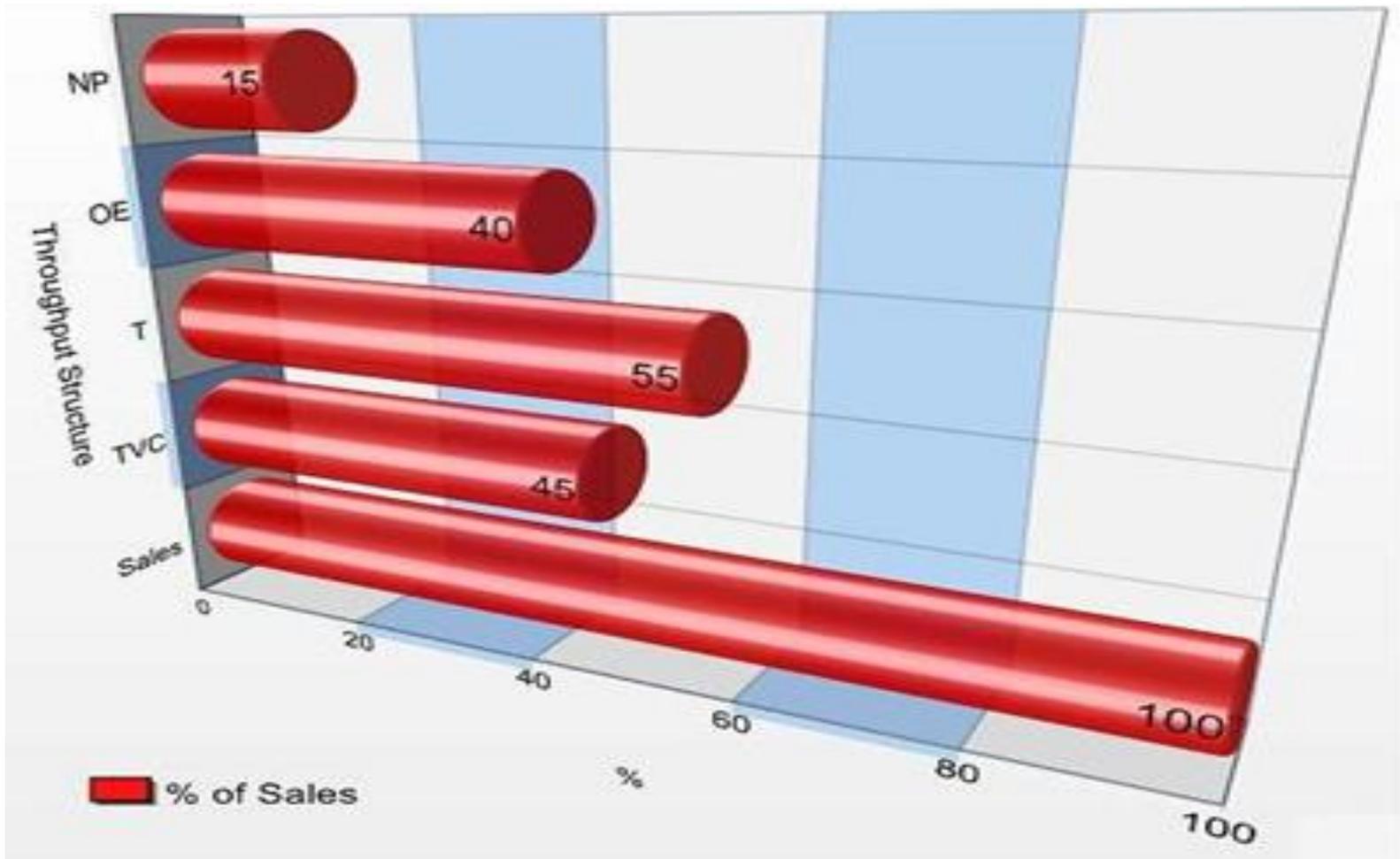


### 3. Compare Graph A and Graph B.

What message is conveyed by Graph A; how does that compare with the message conveyed in Graph B? Which is the most accurate in terms of presenting the info?



# How can you improve the graph?





**CORBIS/Brian Leng (05065)**

basic statistic workshop organised by the  
Global Health Network Nigerian Regional  
Faculty