Statistics

Dr. Htike Myat Phyu

Statistics

- *Statistics* – A scientific discipline which is concerned with systematic collection, compilation, processing, analysis, interpretation and presentation of data.

  (Daniel W, 2005)
- *Biostatistics* – The application of statistics in biological sciences and medicine
- *Medical Statistics* – applications of statistics to medicine and the health sciences

Biostatistics

- The science of quantifying relationships observed in public health and medical statistics can be assumed in two senses.

1) Descriptive Statistics - Summarization of data describing in statistical idea
2) Inferential Statistics - Making of inferences about a population based on information contained in a sample
   - Estimation - Point estimation, Interval estimation
   - Hypothesis testing

Descriptive Statistics or Data Summarization

Learning Outcomes

1. Describe data (mean, median, standard deviation, IQR)
2. Present the results in a proper table or figures

Data summarization

- Percentages, ratios, proportions and rates
- Frequency measures
- Measures of central tendency
- Measures of dispersion
- Measures of location
- Measures of skewness
- Measures of peakedness

Categorical vs. Numerical

- Categorical – frequency, proportion, ratio, rate, %
- Numerical - measures of central tendency (mean, median, mode) & dispersion (SD, IQR)

Summarizing Categorical data

- *Ratio* – x/y [e.g. Male-female ratio = 1:5]
- *Proportions* - x/(x+y) [e.g. Proportion of male = 5/30]
- *Rate* [e.g. Crude Birth rate]
- *Frequency* [e.g. no. of males = 5]
- *Percentage* - x/(x+y) * 100 [e.g. % of males = (5/30) * 100 = 16.67 %]

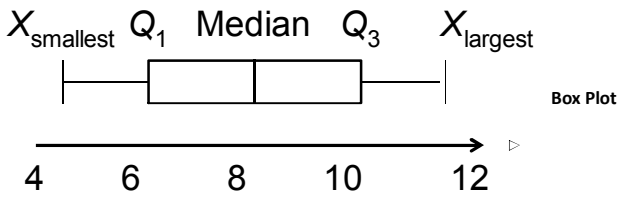
Summarizing numerical data

A. Measures of central tendency
   - Mean - Calculated by adding up all the values and dividing this sum by the number of values in the set
   - Median - middle most value of arranged data
   - Mode – the value that occurs most frequently in the data set

B. Measures of dispersion
   - Standard deviation (SD) & variance
   - Range (Min – Max) & Inter-quartile range (IQR)

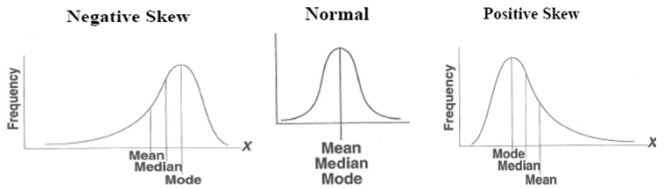
Measures of location

- Percentiles
- Quartiles
- Quintiles
- Deciles



Measures of skewness

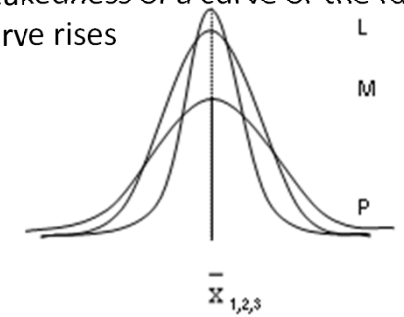
- Skew refers to the symmetry of a distribution



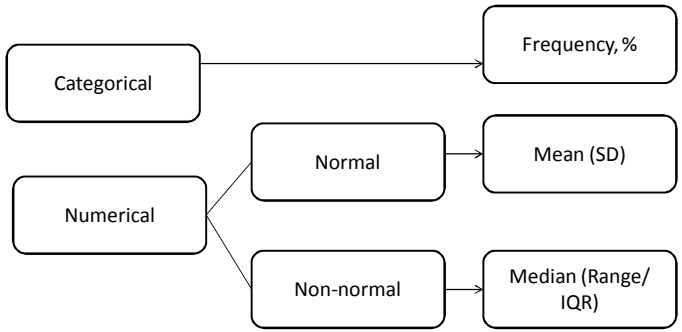
Measures of peakedness

Kurtosis

- describe the peakedness of a curve or the rate at which the curve rises
  - Laptokurtosis
  - Mesokurtosis
  - Platykurtosis



How to describe data



DATA PRESENTATION BY TABLES

Example

3.1 Sociodemographic characteristics of study sample

From the 102 selected residents, 100 (98% response rate) had agreed to participate. Details of the sociodemographic characteristics of study respondents are shown in Table 3.1.

Table 3.1: Sociodemographic characteristics of study sample (100 respondents)

Variable	Mean (SD)	Median (IQR)	Freq. (%)
Age (year)	40.6 (5.36)	-	-
Family income (RM)	-	2500 (2000)[a]	
Gender			
Male			48 (48.0)
Female			52 (52.0)
Ethnicity			
Malay			76 (76.0)
Chinese			13 (13.0)
Indian			10 (10.0)
Others			1 (1.0)
Education level			
No schooling			34 (34.0)
Primary school			32 (32.0)
Secondary/higher			34 (34.0)

SD = Standard Deviation; IQR = Interquartile range; Freq. = frequency
[a] The distribution is skewed to the right

Example

What is dummy table?

Table 2 Prevalence of the MS by age group and age standardized preval

	IDF criteria				
Age groups	Men	Women	P	Total	Me
<45	11 (52%)	9 (60%)	0.65	20 (56%)	9 (43
45-54	31 (62%)	47 (89%)	0.002	78 (76%)	24 (48
55-64	31 (54%)	37 (84%)	0.002	68 (67%)	22 (3
>=65	18 (62%)	37 (95%)	0.001	55 (81%)	12 (41
Total[a]	91 (58%)	130 (86.1%)	<0.001	221 (71.7%)	67 (42.
Total[b]	55.7%	72.1%		64.5%	43.1
P[c]	0.78	0.02		0.03	0.8

a crude prevalence.
b age standardized prevalence.
c p-value for the difference by age group within a subgroup.

- ✓ Title
- ✓ Nice presentation
- ✓ Clear Row & column title
- ✓ Numeric formatting, justification
- ✓ Bold, Italics
- ✓ Gridlines
- ✓ Footnotes – abbreviation or Sourcing

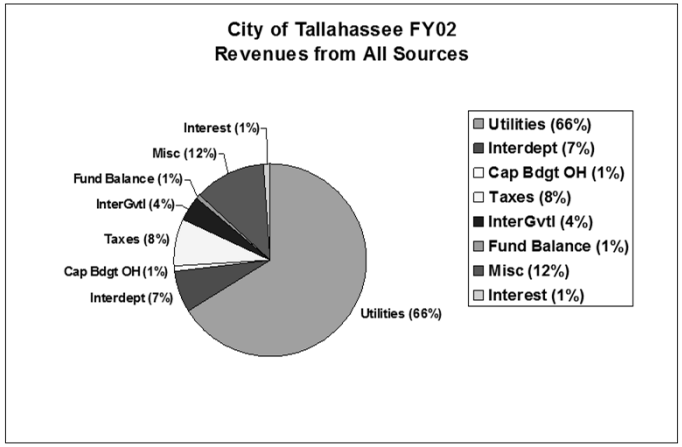
DATA PRESENTATION BY USING FIGURES/GRAPHS/CHARTS

How to display data

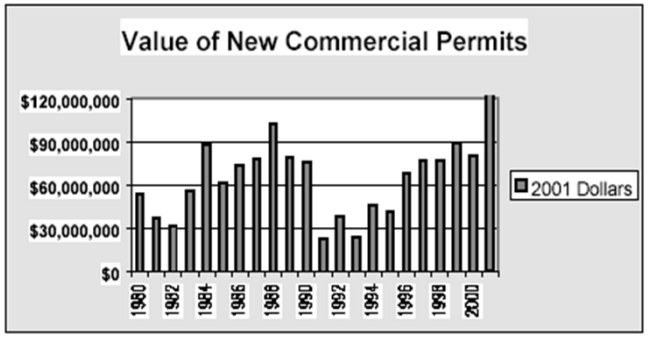
- Categorical
  - Pie
  - Bar chart - Simple or Multiple or composite (stacked), Horizontal or vertical
- Numerical
  - Dot plot or Scatterplot
  - Histogram
  - Line graph
  - Box & whisker plot

CATEGORICAL DATA PRESENTATION

Pie chart

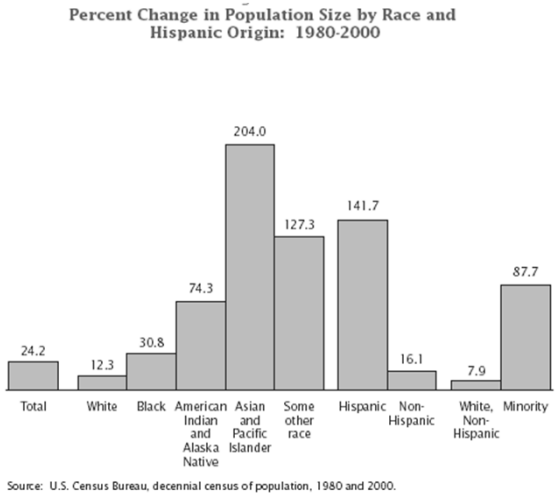


Bar Chart

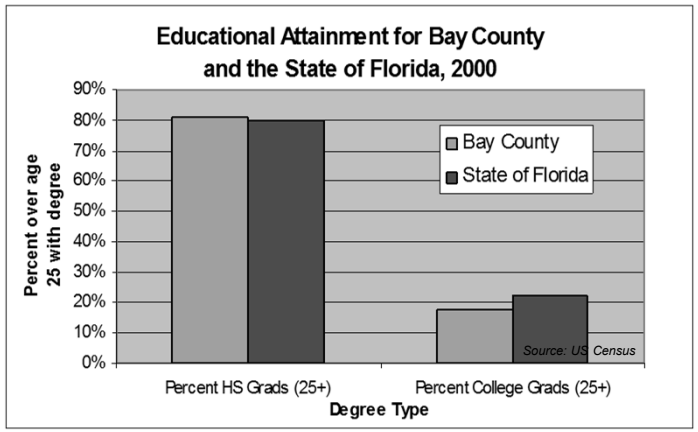


Taken from the *Tallahassee Statistical Digest, 2001*

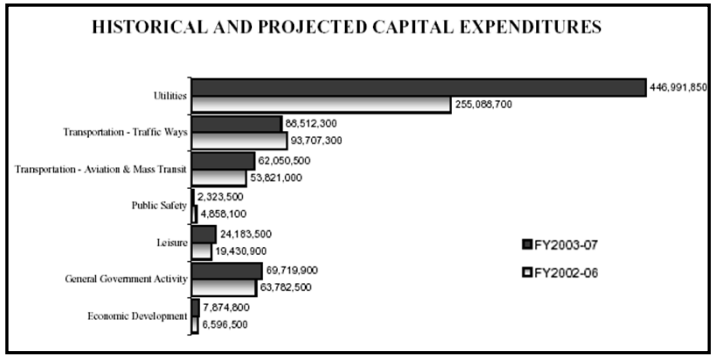
Bar Chart



Clustered Bar Chart

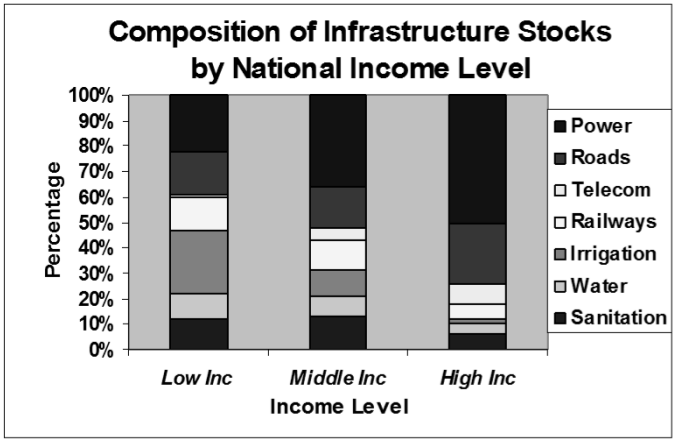


Clustered Bar Chart

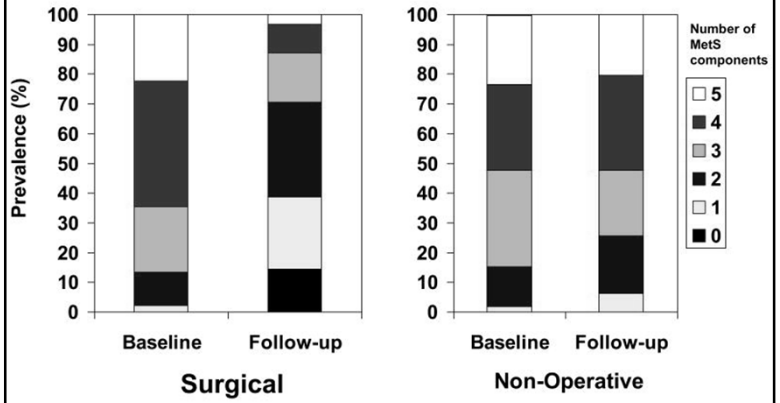


Source: Tallahassee 2003 CIP

Stacked Column Chart

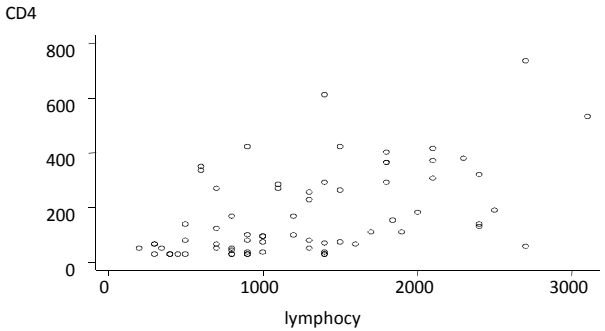


Stacked Column Chart (For Comparing)



NUMERICAL DATA PRESENTATION

Scattered or Dot Plot



Relationship between CD4 counts and lymphocytes

Histogram

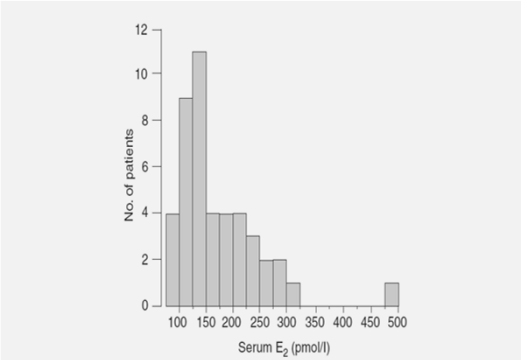
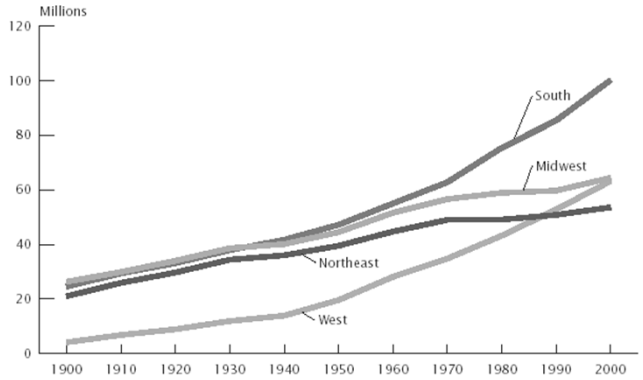


Figure 4.2 An example of positive skew. Serum E2 levels in 45 patients in a study of HRT for the prevention of osteoporosis. Reproduced with permission of the *British Journal of General Practice* (1997, Vol. **47**, pages 161–165)

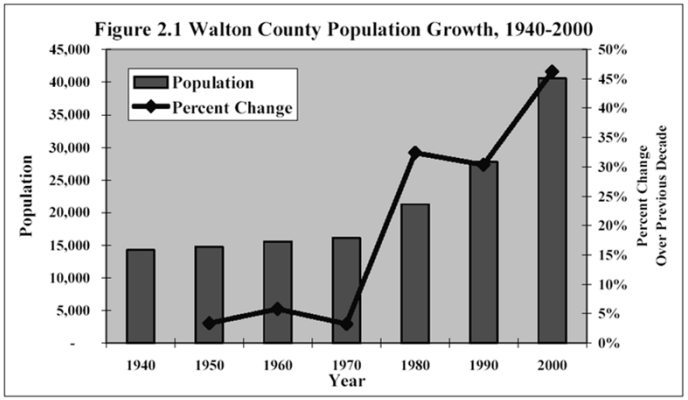
Line Chart

Total Population by Region: 1900 to 2000



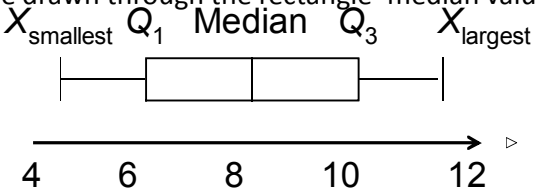
Source: U.S. Census Bureau, decennial census of population, 1900 to 2000.

Line-Column Chart Example



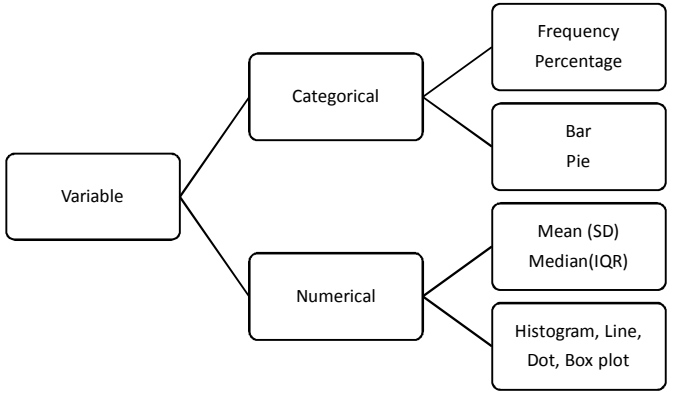
Box plot

- Often called a box-and-whisker plot
- Graphical display of data using 5-Number Summary
- A vertical or horizontal rectangle
- The ends of the rectangle = the upper and lower quartiles
- A line drawn through the rectangle=median value



Take Home Message

Descriptive Statistics



Thank you