LECTURE NOTES

DATA PRESENTATION

BY

MR. LEONARD THUO

Overview

- Variable types
- Descriptive statistics
- Methods for presenting data

Descriptive Statistics

Descriptive statistics are used to describe the data set itself without reference to the population from which it is derived. The use of statistics to describe, summarize, and explain or make sense of a given set of data Examples:

- graphing, calculating, averages, looking for extreme scores.
- Exploratory/Initial data analysis

Variable types

Numerical summaries (mean, median, frequency tables, percentiles)

Graphical summaries (boxplots, empirical CDF, histograms)

Variable types

- Categorical data (qualitative) Nominal data (sex male, female; blood group 0, A, B, AB) Ordinal data (cancer stage I, II, III, IV)
- Numerical data (quantitative) Discrete data (number of children 0, 1, 2, 3, 4, 5+) Continuous data (blood pressure; height in cm)
- Other types of data Ranks, percentages, rates and ratios, scores, visual analog scale, censored data, time to event data

Note: It is important to know the data type since representation and analysis are dependent on this type.

Continuous

Continuous variables: conveys both order and scale.
Quantitative data measured on a continuous scale. Examples

- Weight
- Age
- Time

Usually recorded using some method of rounding
Many methods for summarizing

- mean, median, mode
- variation, skewness
- histograms, distribution functions

Categorical variables

Three types: nominal, ordinal, ranked

Nominal. unordered categories (e.g. gender, race, blood type)

Ordinal. ordered categories (e.g. faculty rank, cancer stages, socio-economic status)

Ranked. rank in a list (conveys order but not scale)

Usually summarized via frequency tabulations

Numerical summaries for categorical data

Frequency tabulations

Gender	Freq.	Percent	Cum.
M	149	65.93	65.93
F	77	34.07	100
Total	226	100	

rank	Freq.	Percent	Cum.
1	64	28.32	28.32
2	105	46.46	74.78
3	57	25.22	100
Total	226	100	

Cross tabulation of gender and rank

	Rank			
gender	1	2	3	Total
M	35	63	51	149
	15.49	27.88	22.57	65.93
F	29	42	6	77
	12.83	18.58	2.65	34.07
Total	64	105	57	226
	28.32	46.46	25.22	100

Frequencies

- Absolute frequency: Number of observation k bearing the same value or fall within a given class from the number n of total observations

$$f_{abs} = k$$

- Relative frequency: Estimate of the probability of a single event for discrete data:

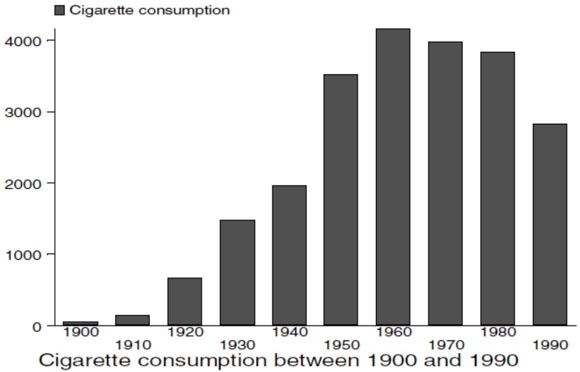
$$f_{rel} = \frac{k}{n}$$

$$0 \leq frel \leq 1$$

- Relative frequency in percent:

$$f_{rel\%} = f_{rel} * 100\%$$

Bar chart


Cigarette consumption
4000
3000
2000
1000
0
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
Cigarette consumption between 1900 and 1990

Summaries for Continuous Data

Measures of central tendency

- mean, median, mode

Percentile summaries; cumulative distribution function
Measures of dispersion, variation, and shape

- interquartile range, standard deviation, variance, skewness

Measures of central tendency

- Mean (sample average)
- Median (middle value of sorted list)
- Mode (most frequent value)

Some necessary notation

List of numbers: $x_1, x_2, ..., x_n$

Sum of these: $\sum_{i=1}^{n} x_i$

Sample mean: $\sum_{i=1}^{n} x_i / n$

Properties:

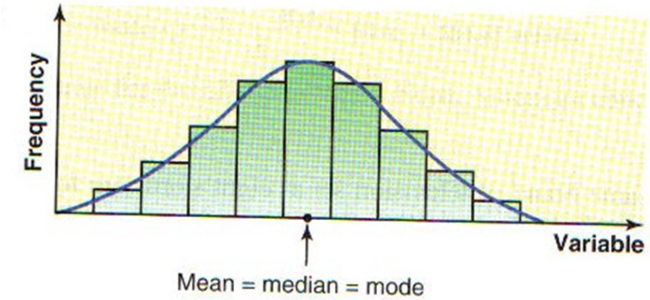
- Easy to calculate
- Nice statistical properties
- Sensitive to extreme observations

Sample median: middle value of ranked list (50th percentile)
Properties:

- Not sensitive to extreme observations
- Tedious to calculate for large data sets
- Not as amenable to statistical manipulation

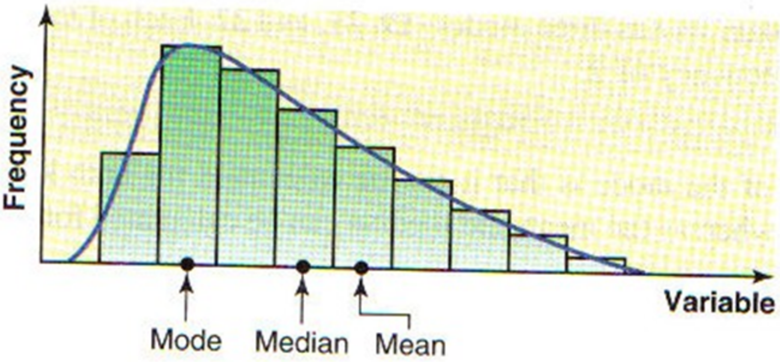
Relationship among mean , median and mode (1)

- For a symmetric histogram and frequency curve with one peak , the values of the mean, median, and mode are identical, and they lie at the center of the distribution



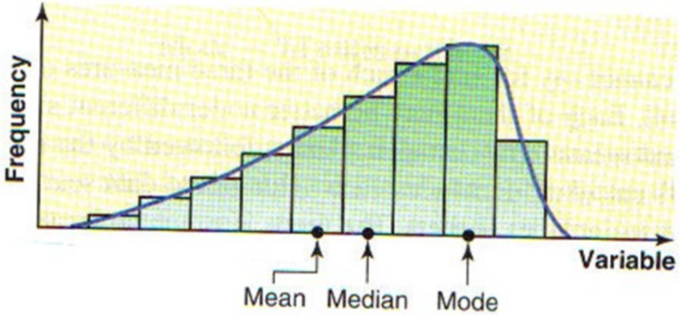
Relationship among mean , median and mode (2)

- For a histogram and a frequency curve skewed to the right, the value of the mean is the largest, that of the mode is the smallest, and the value of the median lies between these two.
- Notice that the mode always occurs at the peak point.
- The value of the mean is the largest in this case because it is sensitive to outliers that occur in the right tail.
- These outliers pull the mean to the right.


Frequency
Variable
Mode
Median
Mean

Relationship among mean , median and mode (3)

- If a histogram and a distribution curve are skewed to the left , the value of the mean is the smallest and that of the mode is the largest, with the value of the median lying between these two.
- In this case, the outliers in the left tail pull the mean to the left


Frequency
Variable
Mean
Median
Mode

Measures of variability

Rank, rank list The sample $x_1, x_2, ..., x_n$ sorted by the size of the values is $x_{(1)}, x_{(2)}, ..., x_{(n)}$ and called rank list, where the indices (1), ...(n) are the ranks $R(x_i)$ of the values.

Range Span width (Range): $r = x_{max} - x_{min} = x_{(n)} - x_{(1)}$

Percentiles The $p\%$ percentile (Q_p) means that $p\%$ of the values are smaller than or equal to the $p\%$ percentile.

$$Q_p = \begin{cases} x_{(k)} & : n * p \text{ is not an integer}(k = int(n * p) + 1 \\ \frac{1}{2}(x_{(k)} + x_{(k+1)}) & : n * p \text{ is an integer}(k = n * p \end{cases}$$

Measures of variability

- Quartiles
  - $1stquartile = Q1 = Q25$
  - $2ndquartile = Q2 = Q50 = median$
  - $3rdquartile = Q3 = Q75$
- Interquartile range: IQR = Q3 - Q1 = Q75 - Q25
- Outlier detection
  - Mild outlier
    $x_i \geq Q75 + 1.5 * IQR$ or $x_i \leq Q25 - 1.5 * IQR$
  - Extreme outlier.
    $x_i \geq Q75 + 3.0 * IQR$ or $x_i \leq Q25 - 3.0 * IQR$
  - This approach could be misleading for small number of observations.
  - There are also other methods for outlier detection and for determination of quartiles. E.g.:
    $Q_p = (1 - j) * x_{(k+1)} + j * x_{(k+2)} : k = \text{int}((n - 1) * p); j = (n - 1) * p - k$

Measures of dispersion: spread

Variance:

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

or

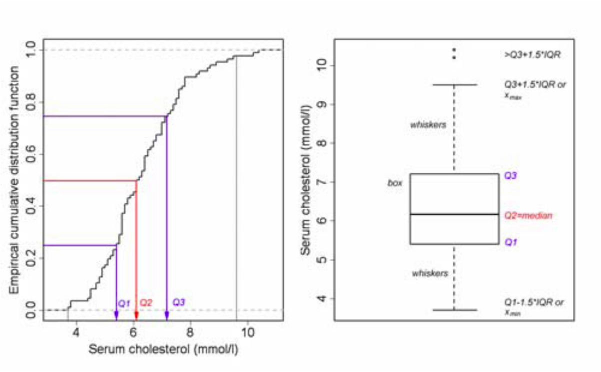
$$\frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum X_i)^2}{n}}{n-1}$$

Standard deviation

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Properties:

Box Whisker plots



Measures of variability

Coefficient of variance

$$CV = \frac{s}{x}$$

provides a measure if the variability is high or not ($CV < 10\%$ means low and $CV > 25\%$ means high variability).

Standard error of mean

$$SE(\bar{X}) = \frac{s}{\sqrt{n}}$$

describes not the data, but the accuracy of the estimation.

Measures of shape

Skewness

$$g_1 = \frac{m_3}{m_2^3} == \frac{\frac{1}{n}\sum_1^n (X_i - \bar{X})^3}{\sqrt{(\frac{1}{n}\sum_1^n (X_i - \bar{X})^2)^3}}$$

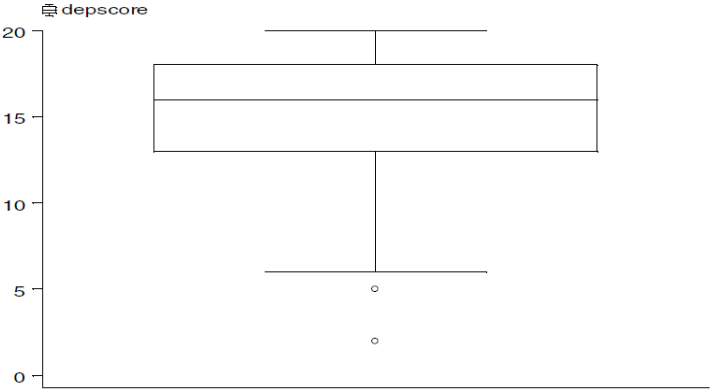
g1 $=$ 0 means the distribution is symmetrical, $g_1 > 0$ right skewed, and $g_1 < 0$ left skewed and mi is the i-th central moment.

Measures of shape

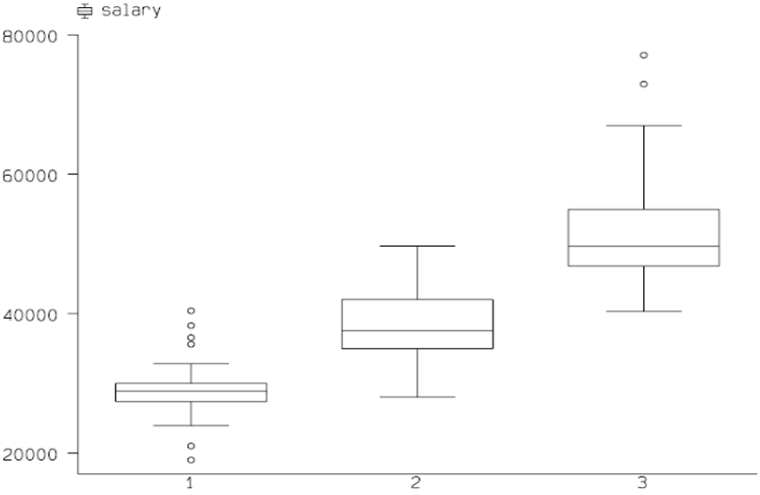
Kurtosis

$$g_2 = \frac{m_4}{m_2^2} - 3 == \frac{\frac{1}{n}\sum_1^n (X_i - \bar{X})^4}{(\frac{1}{n}\sum_1^n (X_i - \bar{X})^2)^2}$$

For normal distribution g2 $= 0$. If $g2 > 0 (g2 < 0)$ within the center of the distribution lies more(less) values than for the normal distribution.



Box plot of Koopmans depression scores


salary
80000
60000
40000
20000
1
2
3

Presentation of continuous data

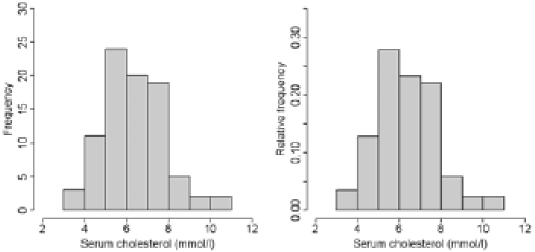
- A simple graphical way of depicting a complete set of observations is by means of the histogram in which the number (or frequency) of observation is plotted for different values or groups of values.
- Example Serum cholesterol levels (mmol/l) of a sample of 86 stroke patients (Markus et. al. 1995)

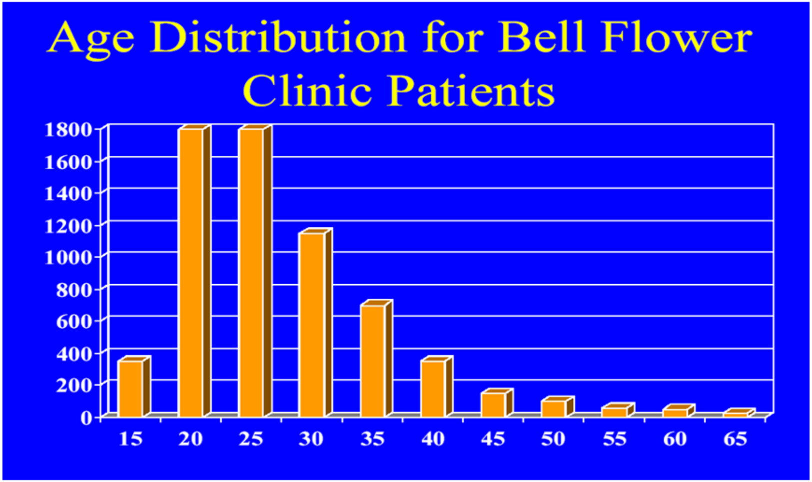
3.7	3.8	3.8	4.4	4.5	4.5	4.5	4.7	4.7	4.8	4.8	4.9	4.9
4.9	5.0	5.1	5.1	5.2	5.3	5.3	5.4	5.4	5.5	5.5	5.5	5.6
5.6	5.6	5.6	5.6	5.6	5.6	5.7	5.7	5.7	5.8	5.8	5.9	6.0
6.1	6.1	6.1	6.1	6.2	6.3	6.3	6.4	6.4	6.4	6.4	6.4	6.5
6.5	6.6	6.7	6.7	6.8	6.8	7.0	7.0	7.0	7.0	7.1	7.1	7.2
7.3	7.4	7.4	7.5	7.5	7.6	7.6	7.6	7.7	7.8	7.8	7.8	8.2
8.3	8.6	8.7	8.9	9.3	9.5	10.2	10.4					

Histogram

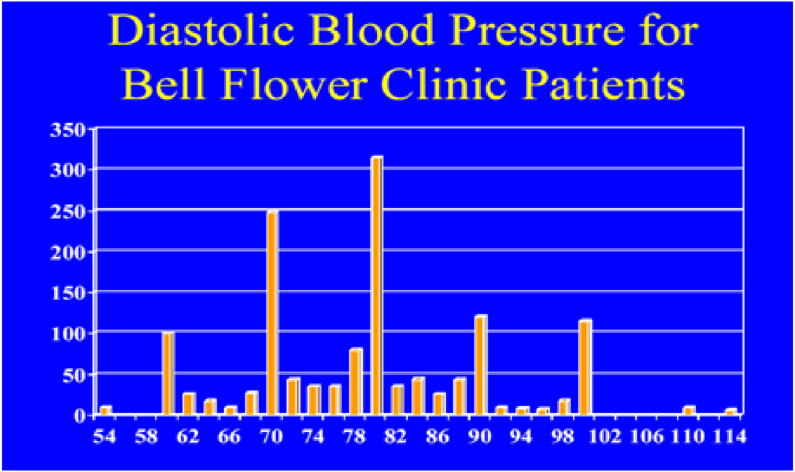
- Give relative frequency of different values
- Conveys shape of distribution
- Typically useful only for continuous data
- Smoothness important for presentation

Histograms of cholesterol levels from stroke patients

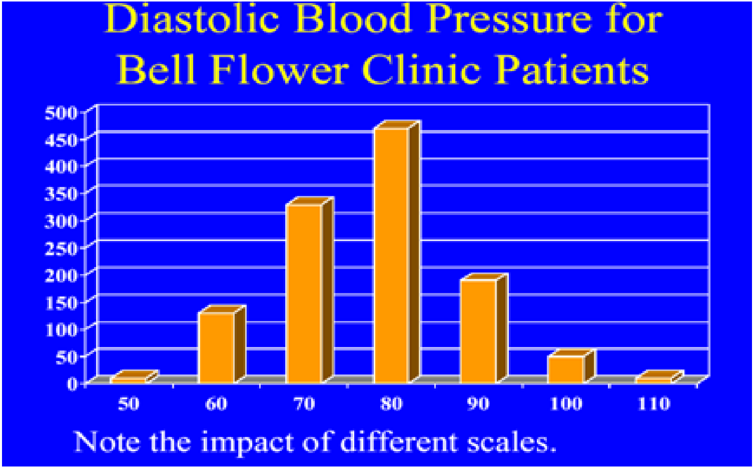



Age Distribution for Bell Flower Clinic Patients
1800
1600
1400
1200
1000
800
600
400
200
0
15
20
25
30
35
40
45
50
55
60
65

Frequency Histogram Example

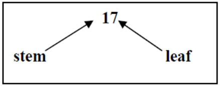


Frequency Histogram Example



Stem and Leaf Plot

- A stem and leaf plot is a method used to organize statistical data.
- The greatest common place value of the data is used to form the stem.
- The next greatest common place value is used to form the leaves.



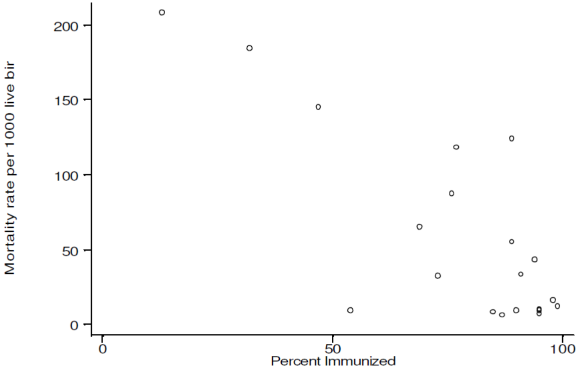
Example

- Make a stem and leaf plot of the algebra test scores given below. 56, 65, 98, 82, 64, 71, 78, 77, 86, 95, 91, 59, 69, 70, 80, 92, 76, 82, 85, 91, 92, 99, 73
- Step1: Sort the data 56, 59, 64, 65, 69, 70, 71, 73, 76, 77, 78, 80, 82, 82, 85, 86, 91, 91, 92, 92, 95, 98, 99
- Since the data range from 56 to 99, the stems range from 5 to 9.
- To plot the data,make a vertical list of the stems.
- Each number is assigned to the graph by pairing the units digit, or leaf, with the correct stem.
- The score 56 is plotted by placing the units digit, 6, to the right of stem 5.

Stem	Leaf
5	6 9
6	4 5 9
7	0 1 3 6 7 8
8	0 2 2 5 6
9	1 1 2 2 5 8 9

Scatter plot

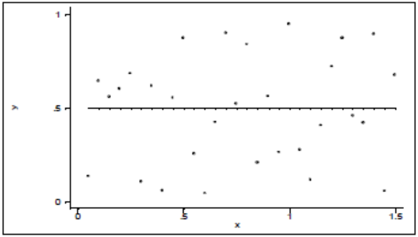
Example: DPT Immunization and Infant Mortality
Consider the following two-way scatter plot of the under-5 mortality rate on the y axis and the DPT levels (percent of the population immunized) on the x axis (under five mortality rate data set).



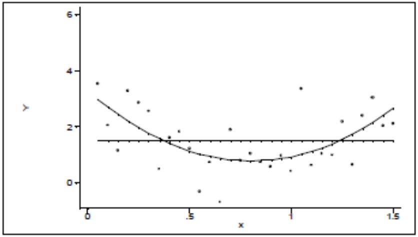
By simple inspection of the graph it is clear that as the proportion of infants immunized against DPT *increases*, the infant mortality rate *decreases*.

Scatter plots

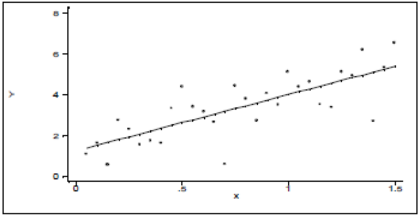
Some examples of relationships between two measures



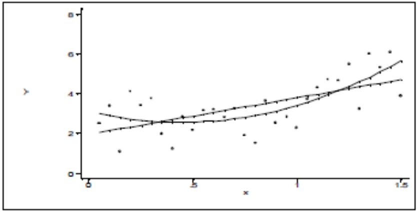
1. No relatioship between X and Y



2. Non-linear relationship between X and Y



3. Linear positive relationship between X and Y



4.Non-linear (but positive) relationship between X and Y

Grouped Data

Example Obtain the mean, variance and median of the following dataset

daily commuting time	Number of employees
0 to $<$10	4
10 to $<$20	9
20 to $<$30	6
30 to $<$40	4
40 to $<$50	2

Summaries for grouped data

- Mean

$$\bar{X} = \frac{\sum mf}{n}$$

- Variance

$$s^2 = \frac{\sum m^2 f - \frac{(\sum mf)^2}{n}}{n-1}$$

where m is the midpoint and f is the frequency

Solution

Obtain midpoint m

daily commuting time	f	m	mf	m^2f
0 to $<$10	4	5	20	100
10 to $<$20	9	15	135	2025
20 to $<$30	6	25	150	3750
30 to $<$40	4	35	140	4900
40 to $<$50	2	45	90	4050
	n=25		$\sum mf$=535	$\sum m^2f$=14825

Summaries for grouped data

- Median

$$l + \left(\frac{\frac{n}{2} - cf}{f}\right) * h$$

where

- l is the lower limit of median class
- cf cumulative frequency of class prior to the median class
- f frequency of median class
- h class size

Solution

daily commuting time	f	cf
0 to <10	4	4
10 to <20	9	13
20 to <30	6	19
30 to <40	4	23
40 to <50	2	25

- l=10,n=25,cf=4,h=10,f=9

$$\text{Median} = 10 + \frac{12.5 - 4 * 10}{9} = 19.4$$