The Online R-FETPV 1st Module : Basic Epidemiology and Surveillance Data Analysis

5 April -28 May 2021

USAID FROM THE AMERICAN PEOPLE

Department of Livestock Development

Food and Agriculture Organization of the United Nations

R-FETPV Regional Field Epidemiology Training Program for Veterinarian

Surveillance data analysis and bias: principles of data handling individually and as part of a field and laboratory network

Suwicha Kasemsuwan

FETPV

USAID FROM THE AMERICAN PEOPLE

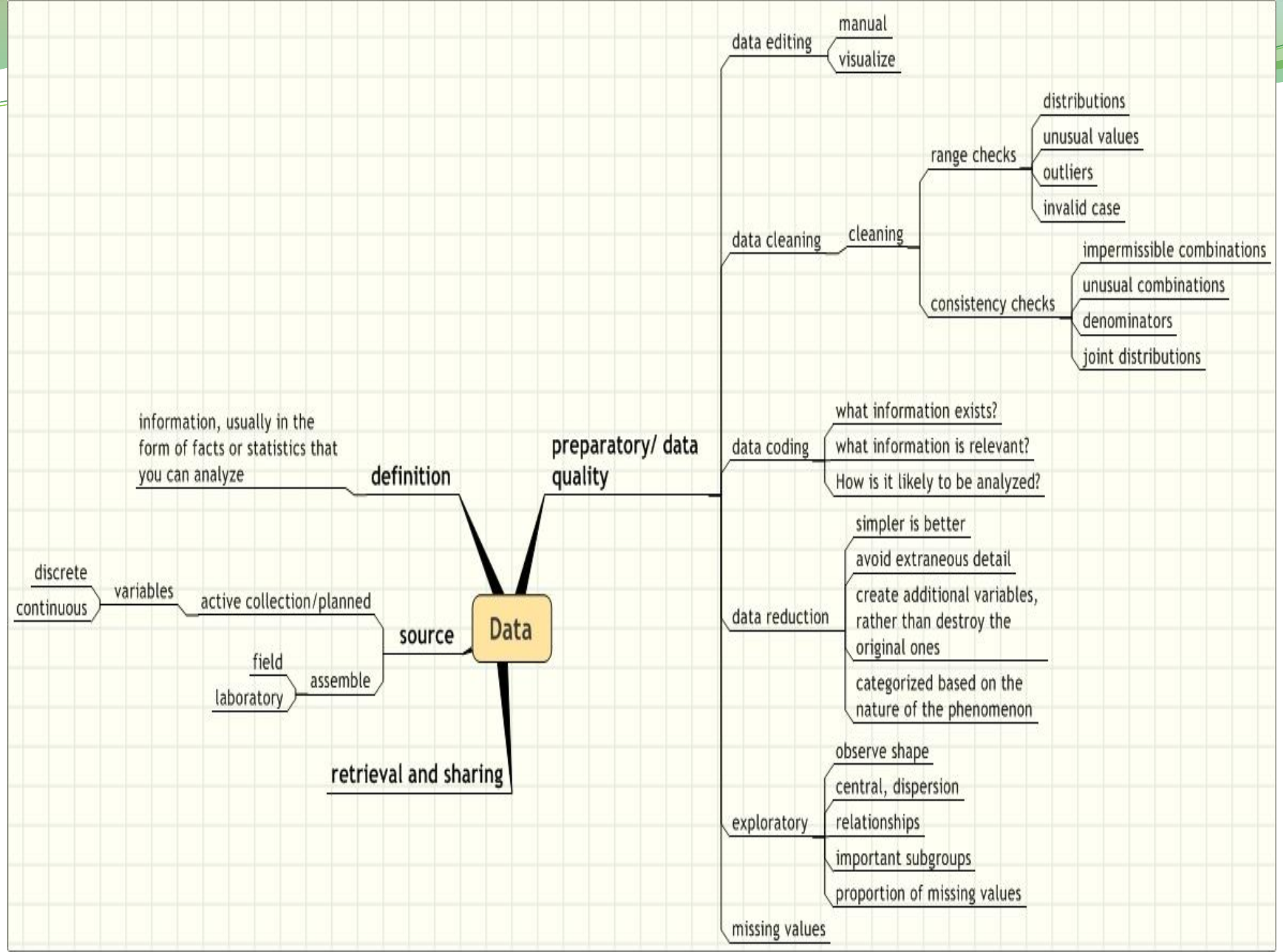
Department of Livestock Development

Food and Agriculture Organization of the United Nations

R-FETPV Regional Field Epidemiology Training Program for Veterinarians

References

- Victor J. Schoenbach, Waynbe D. Rosamond. 2000. Understanding the Fundamentals of Epidemiology and Evolving text. Chapel, North Carolina
- THEILERIOSIS in Eastern, Central and Southern Africa, proceedings of a workshop on East Coast fever immunization held in Lilongwe, Malawi 20-22 September 1988 *Organized by* The International Laboratory for Research on Animal Diseases, The Food and Agriculture Organization of the United Nations, The Organization of African Unity *With support from* The Government of Malawi *Edited by* T.T. Dolan *Published by* THE INTERNATIONAL LABORATORY FOR RESEARCH ON ANIMAL DISEASESBOX 30709 · NAIROBI · KENYA


Data
definition
information, usually in the form of facts or statistics that you can analyze
source
active collection/planned
variables
discrete
continuous
assemble
field
laboratory
retrieval and sharing
preparatory/ data quality
data editing
manual
visualize
data cleaning
cleaning
range checks
distributions
unusual values
outliers
invalid case
consistency checks
impermissible combinations
unusual combinations
denominators
joint distributions
data coding
what information exists?
what information is relevant?
How is it likely to be analyzed?
data reduction
simpler is better
avoid extraneous detail
create additional variables, rather than destroy the original ones
categorized based on the nature of the phenomenon
exploratory
observe shape
central, dispersion
relationships
important subgroups
proportion of missing values
missing values

Aim for participants to learn

- Data quality
- Data assembly from field and laboratory sources for basic descriptive statistics
- Data storage
- Data retrieval
- Data sharing

data

- Facts, statistics or information either historical or derived by calculation or experimentation. (Webster's Universal Dictionary)
- Is information, usually in the form of facts or statistics that you can analyse. (Collins Cobuild Essential English Dictionary)

Data

- Do not speak for themselves.
- Begin with data exploratory, descriptive analyses
- Feel for the data
- Address specific questions from the study aims.
- Data reflect
  - Relationship among host, agent, environment
  - A competent investigator will often be convinced of a claim from simply "eyeballing" data or plotting a simple graph.

Data collection

- Reliable process: paper records or computerized
- Survey or non-survey purposes (e.g. during disease control intervention, inspections for movement control or during disease eradication schemes)
- Consistency and quality of data collection
- Format that facilitates analysis

Terrestrial Animal Health Code - 2007

Quality of data

- Distribution of those involved in generating and transferring data from the field to a centralized location
- The ability of the data processing system to detect missing, inconsistent or inaccurate data, and to address these problems
- Maintenance of disaggregated data rather than the compilation of summary data
- Minimization of transcription errors during data processing and communication

Terrestrial Animal Health Code - 2007

Preparatory work

- Data editing
- Data cleaning
- Exploratory of the data
- Missing values
- Data coding
- Data reduction

Data editing

- Before and after computerized
- Manual or visual editing
  - Get a sense for how well the forms were filled out.
- Keyed
  - A data entry program: avoid illegal values from entering the dataset.

Data cleaning

- Range checks
  - Detect and correct invalid values
  - Note and investigate unusual values
  - Not outliers
  - Check reasonableness of distributions
- Consistency check
  - Detect and correct impermissible combinations
  - Note and investigate unusual combination
  - Check consistency of denominators
  - Check reasonableness of joint distributions

Data coding

- Translating information into values suitable for computer entry and statistical analysis.
- Coding decisions:
  - What information exists?
  - What information is relevant?
  - How is it likely to be analyzed?

- Absence = 0, present = 1
- No = 0, yes = 1
- Male = 1, female = 2
- Skipped responses
  - Question not applicable for this respondent
  - Respondent declined to answer
  - Respondent not know
  - Respondent skipped without reason

- easier when done at once
- One may ignore coded or judged as meaningless
- Not code → not analyse
- More detail → more recode → more chance for error

Data reduction

- Reduce the number of variables for analysis by combining single variables into compound variables that better quantify the construct.
- Height and weight → overweight, underweight
- Collapsing possible values to a smaller number: breed → small, medium, large

- Create additional variables, never overwrite the raw data
- Verify accuracy of derived variables and recodes by examination cross tabulations between the original and derived variables.

Exploring the data

- Feel the data
- Cross tabulation, scatter plots
- Observe shape: symmetry?, discontinuities
- Summary statistics
  - Location: mean, median, percentage above a cut-point
  - Disperson: standard deviation, quantiles
- Look for relationships in data
- Look within important subgroups
- Note proportion of missing values

Missing values

- Confusing, tiresome
- Denominators differ
- When involve multiple variables: may exclude an entire observation (listwise deletion) if missing is in pattern, beware of bias
- Imputation for missing values
  - Replace missing value by the mean or median for that variable (not reduce bias)
  - By making use of the values of variables for which data are present and which are related to the variable being imputed: predictive (i.e. BP and age in human)

Sample size

- Too large
  - Comparing groups and using unnecessarily large sample sizes can lead to "statistical significance", which is not the same as "biological significance".
- Too small
  - Choosing sample sizes that are too small is a common error and frequently leads to a total waste of study resources. When it occurs, it can lead to claims of no significant difference between groups because the experiment was designed with inadequate data.

Bias and confounding

- Bias
  - One treatment group may be advantaged simply because animals have not been randomized to ensure comparable groups at the start of the experiment.
  - Data from certain animals may be omitted in the belief that these animals produced "outliers".
- Confounding
  - Detected differences among treatment groups may be confounded by other factors such as breed, genotype or age.
  - Randomization helps to eliminate confounding, which frequently occurs when animals in different treatment groups are not managed similarly.

What can analysis offer?

- Experimenter must be protected from prejudices.
- Statistical methods provide a scientific standard for planning studies, exploring data patterns and reporting scientific claims.

Data retrieve and sharing

- Data stored electronically in databases also offer a convenient medium for the exchange of information.
- Key field

- A wide range of software packages exists for data handling.
  - Database packages
  - Spreadsheet packages
- The data can be exported from the database to other statistical software packages for analysis.
- Graphics packages can be used to express trends in the data and produce visual images of the relationships among parameters.
- The fast rate of processing data stored electronically means that analyses can routinely take place during the course of the study and final analyses can be completed within days of the end of the study.

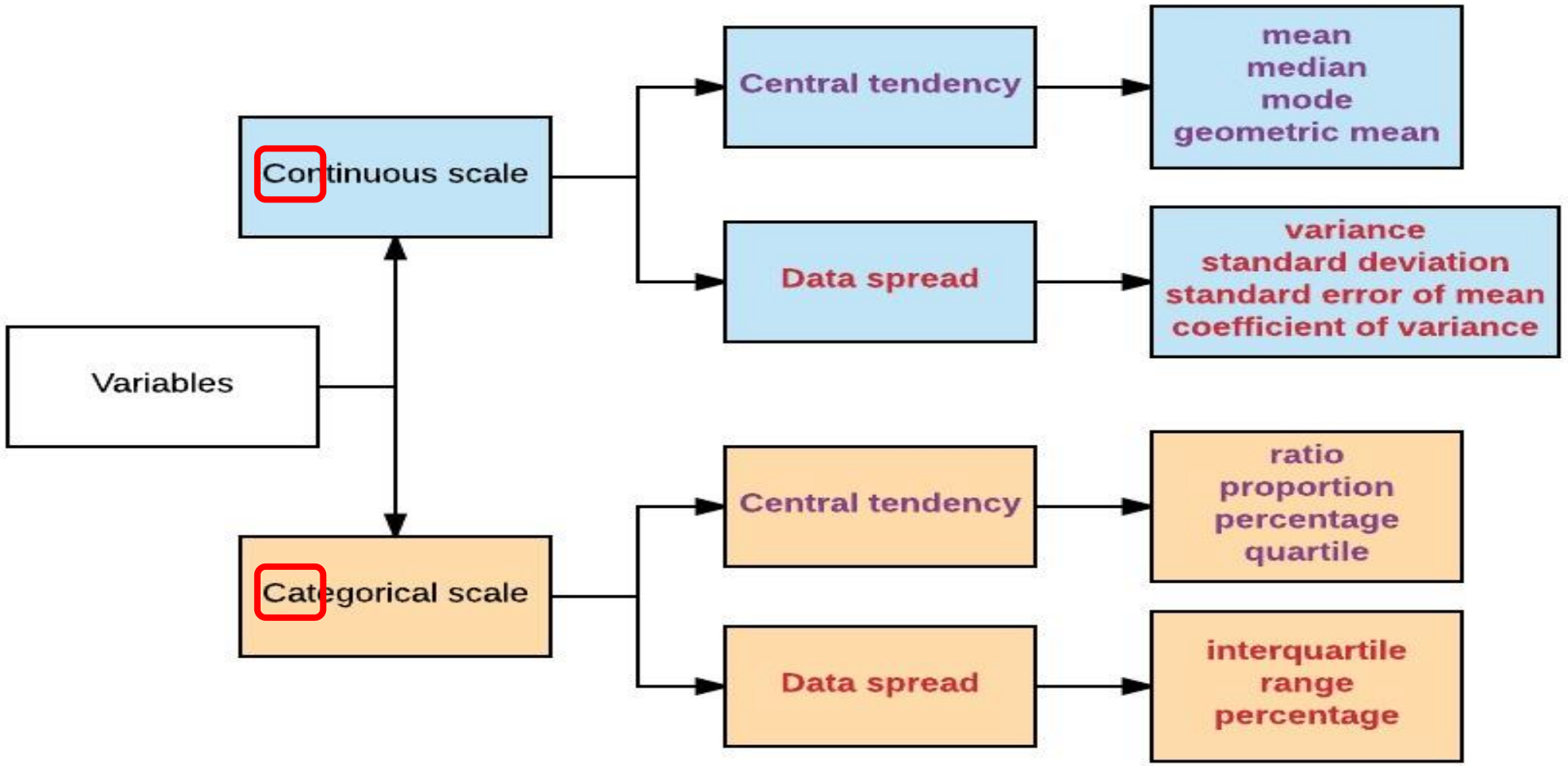
Descriptive statistics

Types of variables

Quantitative	
Continuous	**Discrete**
Blood pressure, height, weight, age	Number of children Number of FMD cases per week

Categorical	
Ordinal (Ordered categories)	**Nominal (Unordered categories)**
Grade of breast cancer Better, same, worse	Sex (male/female) Alive or dead

Variables


Variables
Continuous scale
Categorical scale
Central tendency
Data spread
Central tendency
Data spread
mean
median
mode
geometric mean
variance
standard deviation
standard error of mean
coefficient of variance
ratio
proportion
percentage
quartile
interquartile
range
percentage

Measurement of the middle

- Mean

$$\overline{X} = \frac{\sum X}{n}$$

- Median = 50 percentile

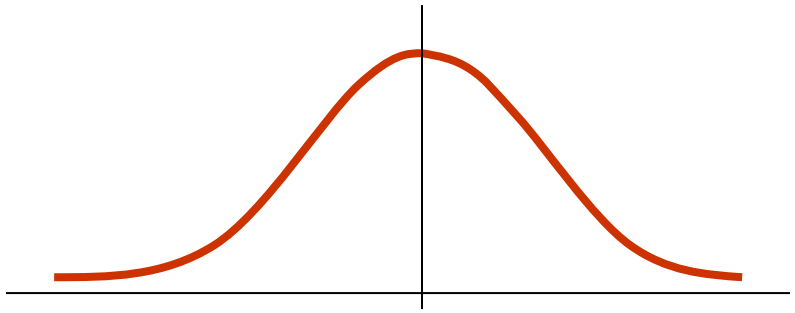
- Mode

- Geometric mean

$$GM = \sqrt[n]{(X_1)(X_2)(X_3)\ldots(X_n)}$$

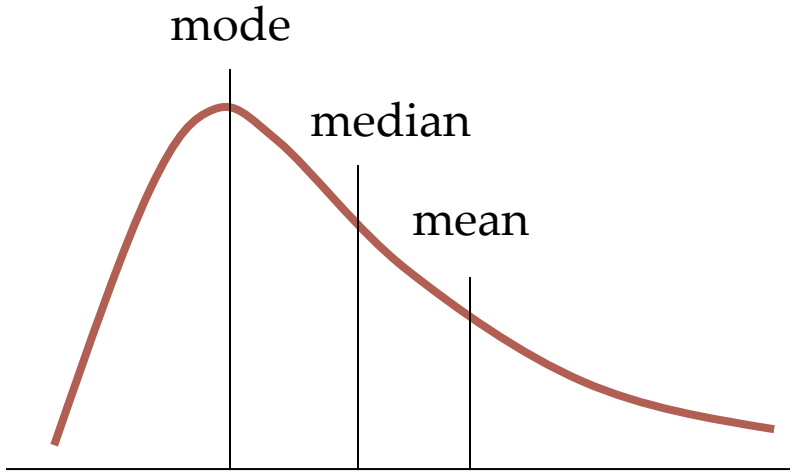
$$\log GM = \sum \frac{\log X}{n}$$

Mean = median = mode

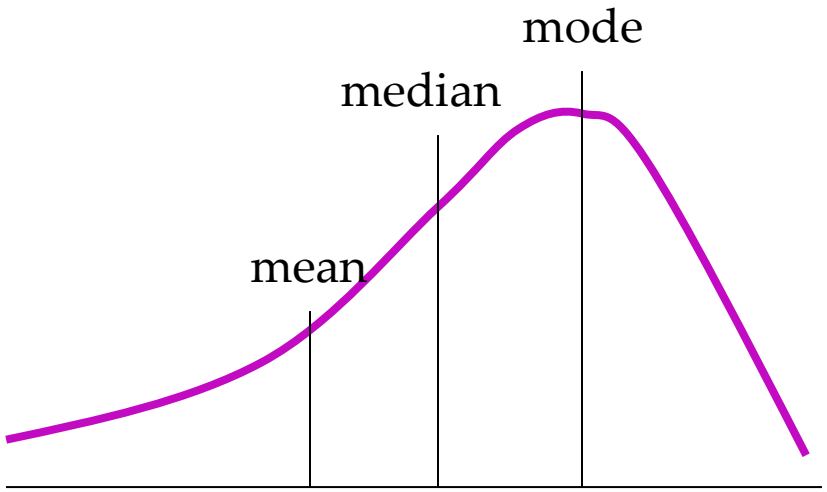


Symmetric

Symmetric



Positively skewed



Negatively skewed

Measures of spread: variance, s, se, CV

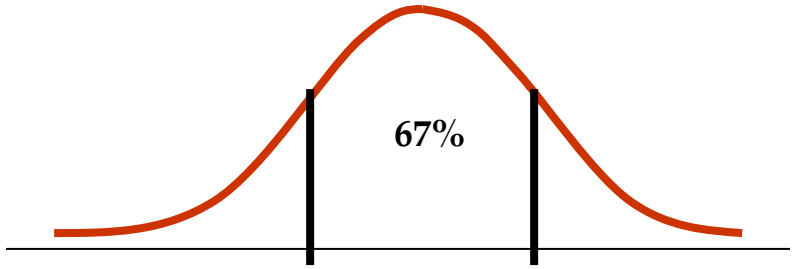
- Range **max. – min.**
- Standard deviation (ของ mean) $s = \sqrt{\frac{\sum(X - \overline{X})^2}{n-1}}$
- Standard error (ของ mean) SEM = s / sqrt n
- Coefficient of variation percentiles $CV = \frac{SD}{\overline{X}}(100\%)$
  - relative spread in data
- Percentile
- Inter-quartile range Different between 25th and 75th percentile

Variance / standard deviation / standard error of a proportion

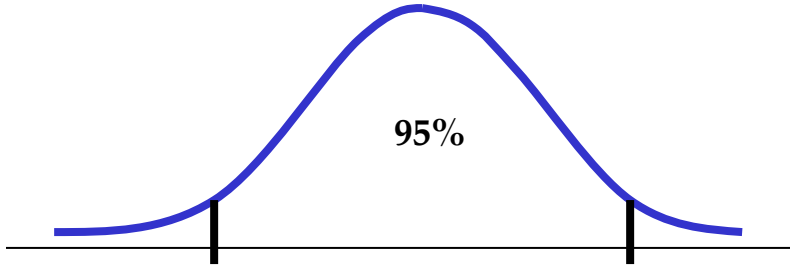
- **Variance of a proportion**
  - $\sigma^2 = PQ$ $\quad s^2 = pq$ $\quad q = 1-p$
- **Standard deviation of a proportion**
  - $\sigma = \sqrt{\sigma^2} = \sqrt{PQ}$ $\quad s = \sqrt{s^2} = \sqrt{pq}$
- **Standard error of a proportion**
  - $\sigma_p = \sqrt{\frac{PQ}{N}}$

Standard normal (z) distribution

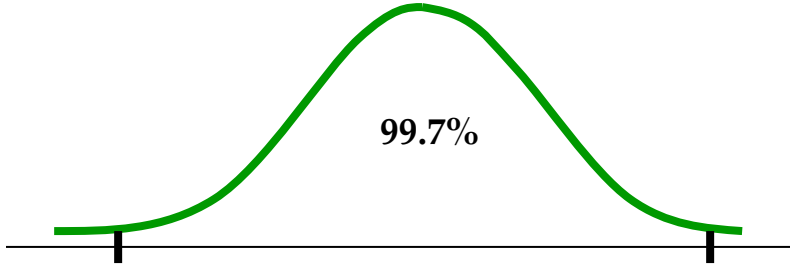
1 s: 68.2%

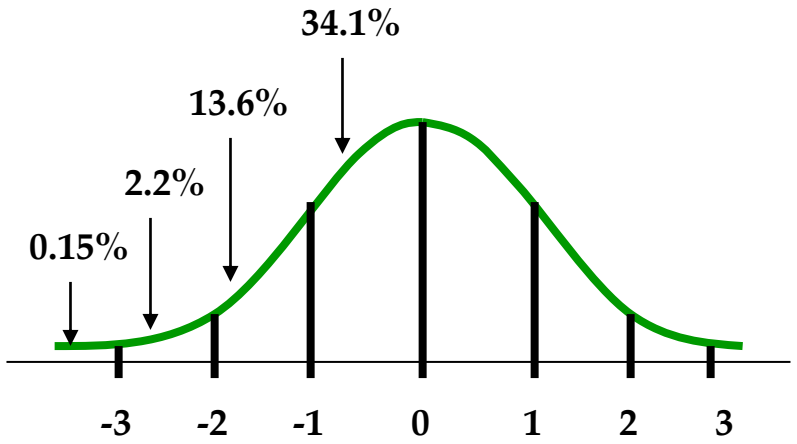


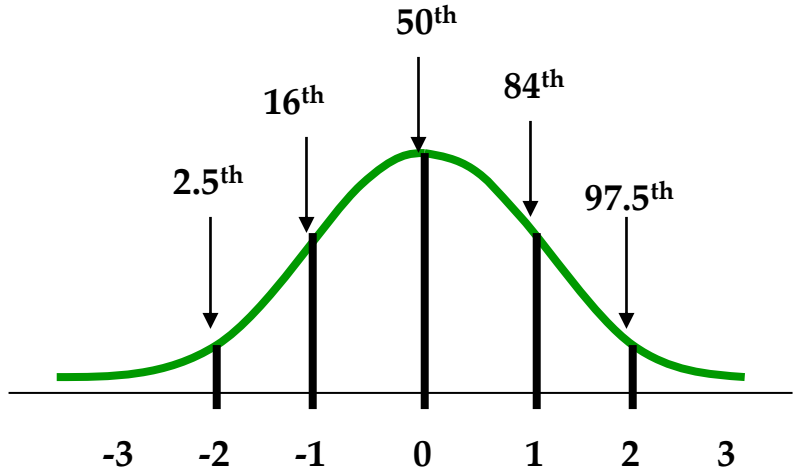
2 s: 95.4%



3 s: 99.6%







$$SEM = \frac{s}{\sqrt{n}}$$

Sample size matters

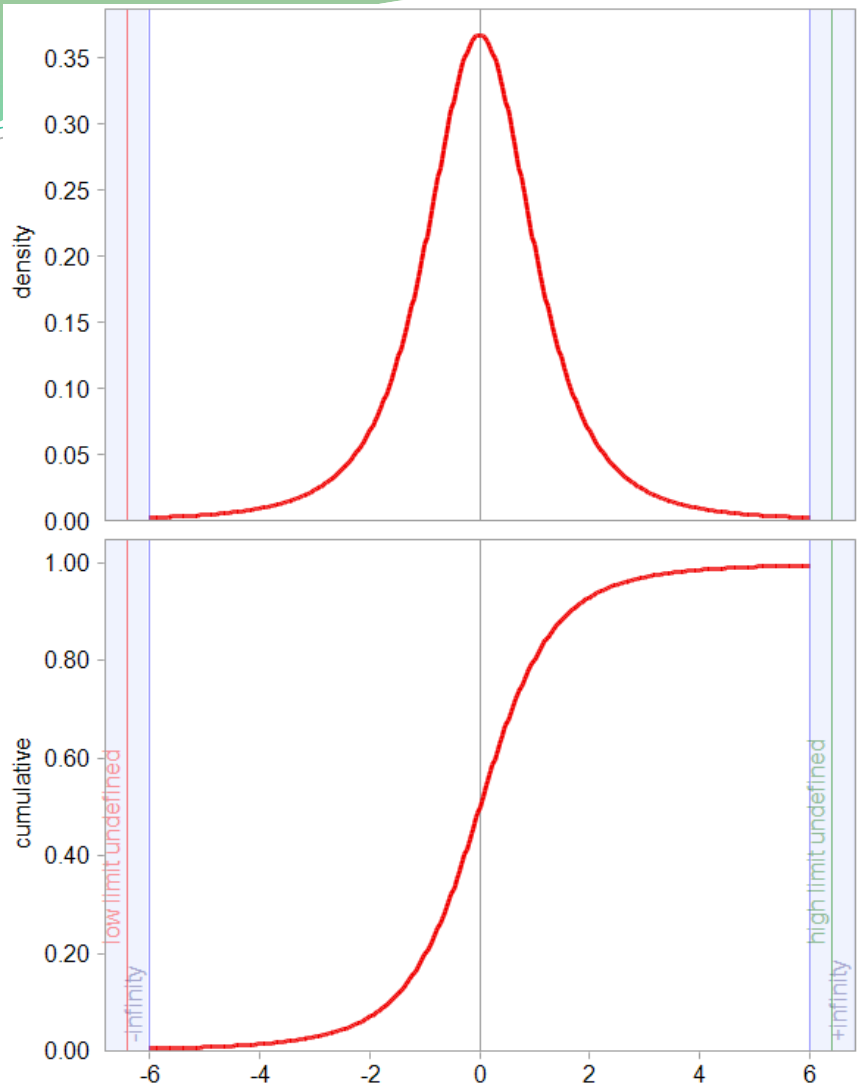
- When sample size increased, sample mean and standard deviation would be more accuracy in representing population mean and standard deviation.
- SEM does not indicate variability of original population (standard deviation indicates variability)
- SEM indicates certainty that sample mean assess true population mean

outliers

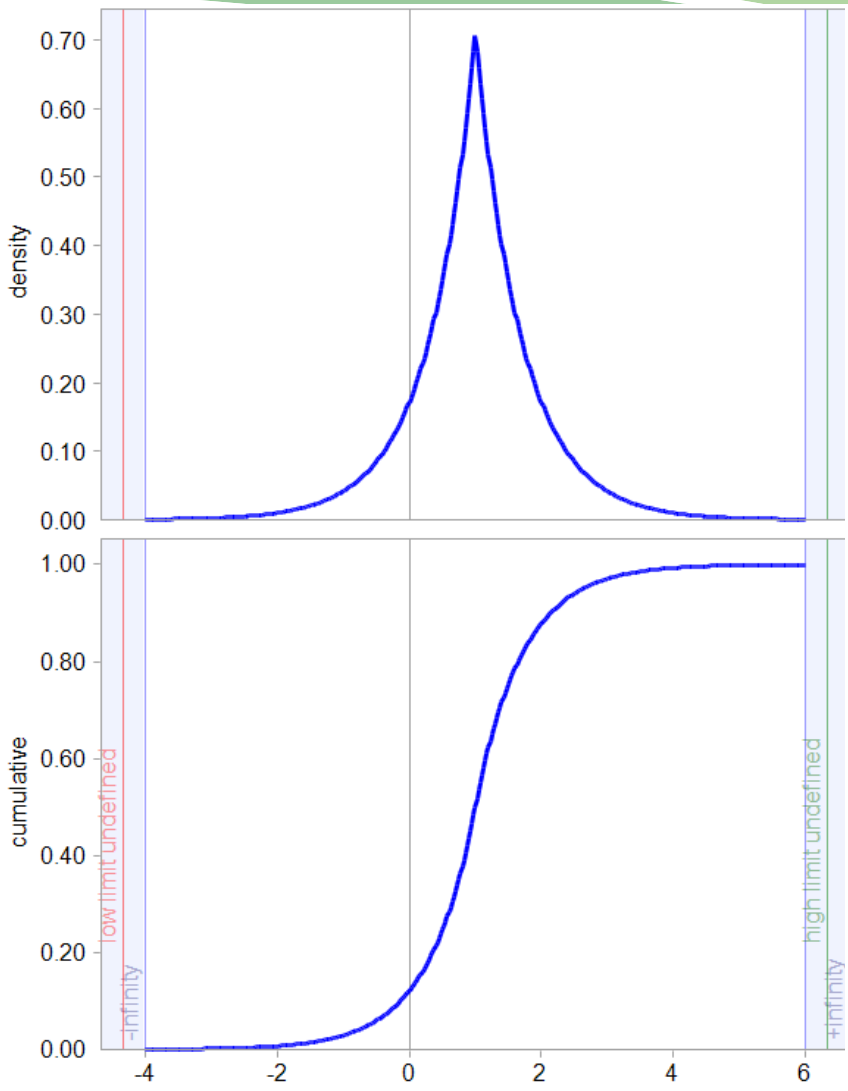
- Outlier is any values that are less than 1.5 IQR of the first quartile (Q1) or more than 1.5 IQR of the 3rd quartile (Q3)
  - High = Q3 + 1.5 IQR
  - Low = Q1 – 1.5 IQR

skewness

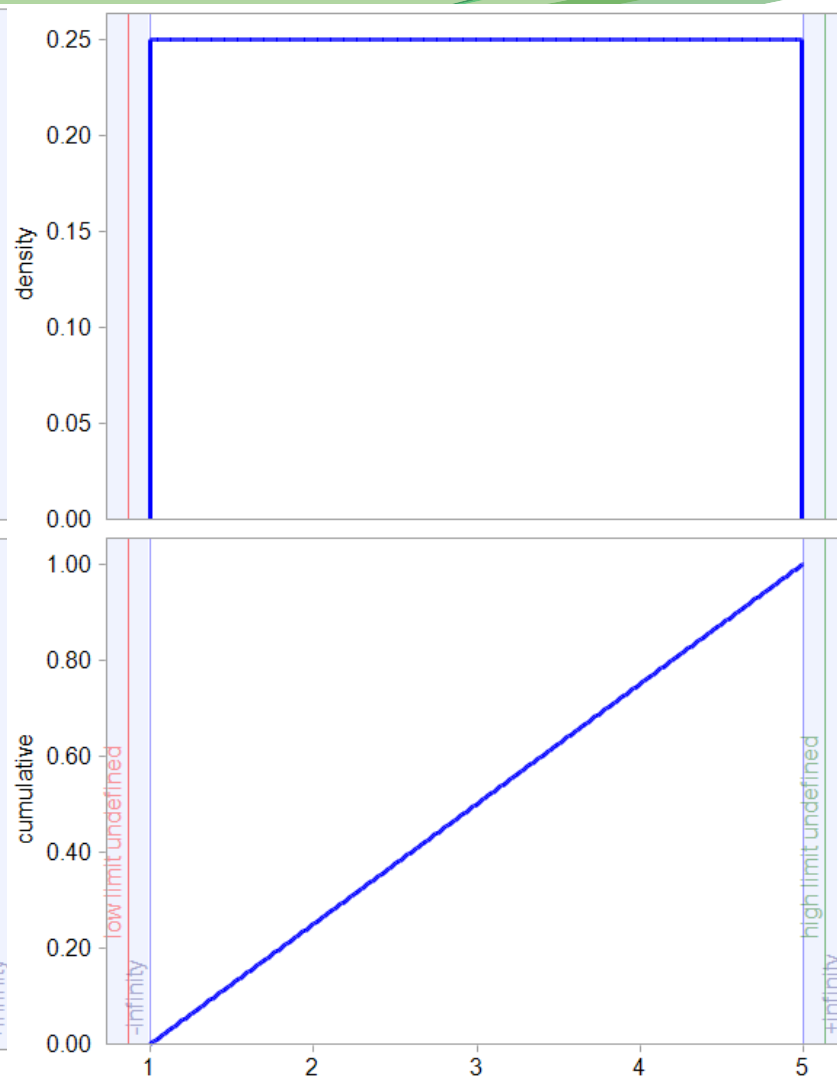
- **Skewness measures symmetry of distribution**
- **standard normal distribution is perfectly symmetry, and the skew is 0.**
- **Other distributions that have zero skew**
  - **= T distribution, uniform distribution, Laplace distribution**
- Let-skewed distribution (negative-skewed)
  - Tail is longer on the left, mean is pulled to the left.
- Right-skewed distribution (positive-skewed)
  - Tail is longer on the right, mean is pulled to the right.



T distribution

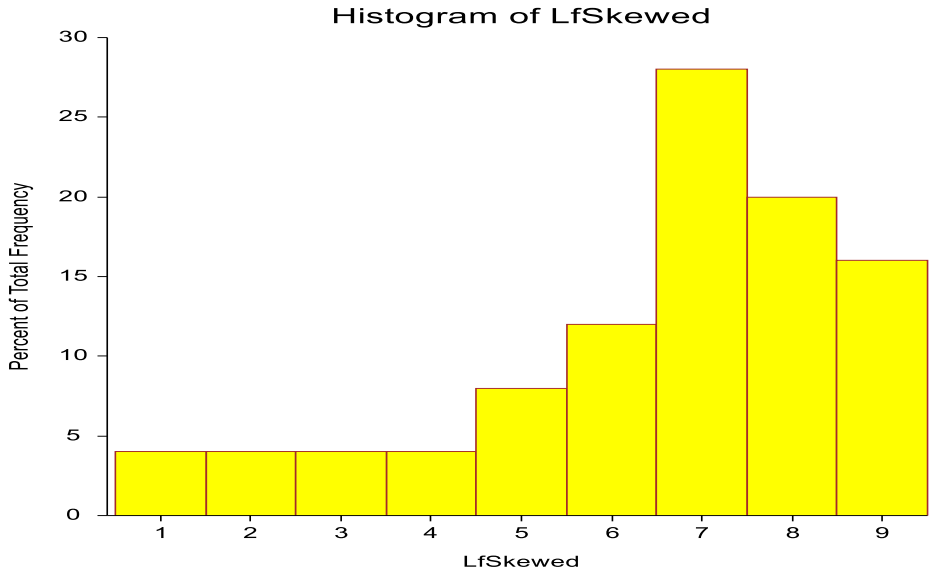


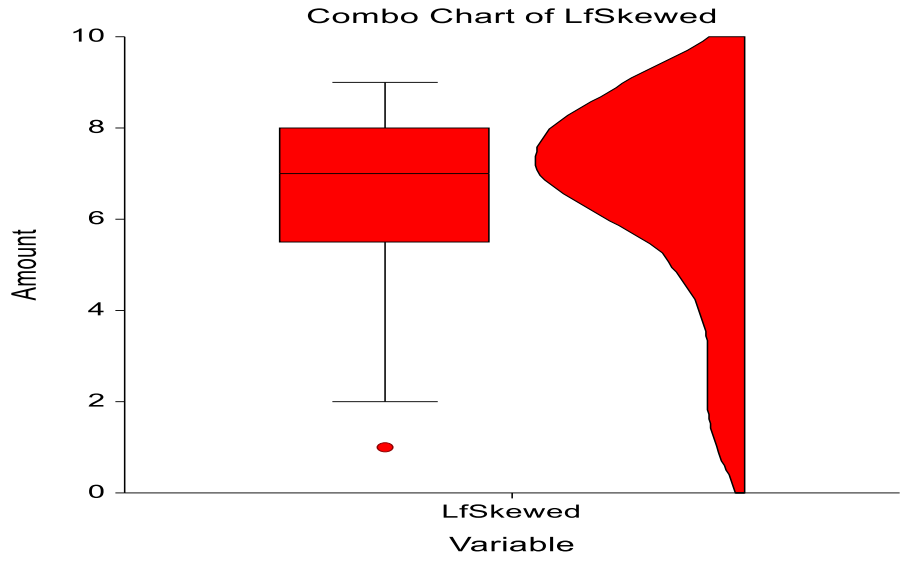
Laplace distribution



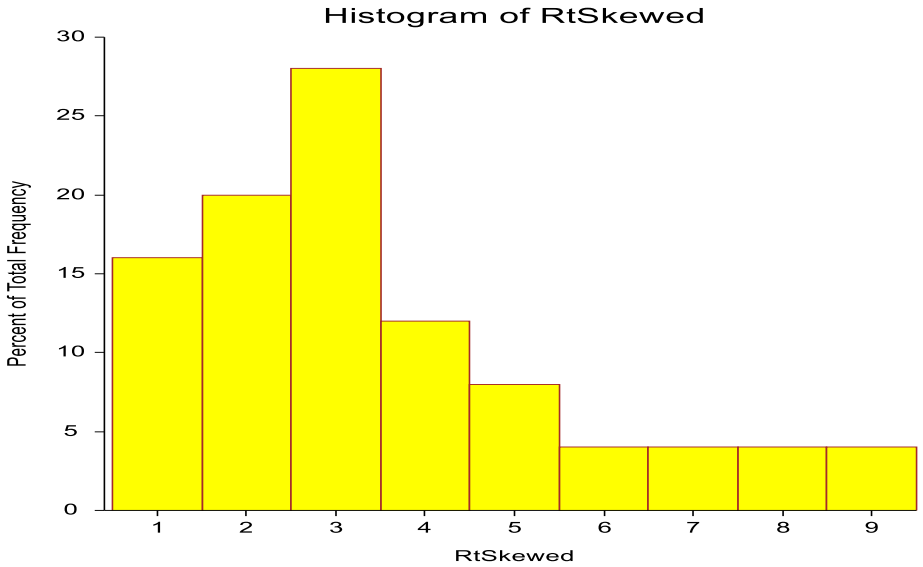
uniform distribution

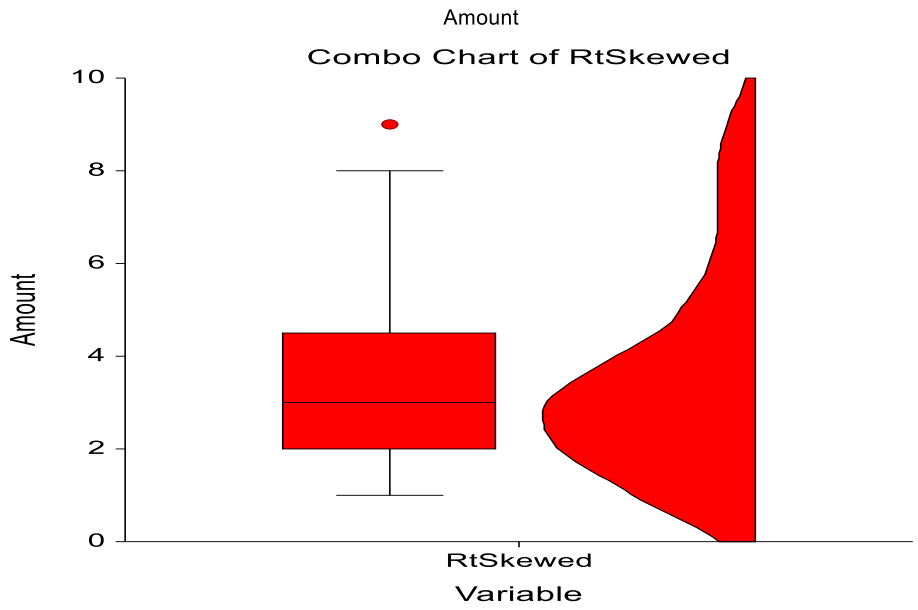
Left-skewed distribution





Right-skewed distribution





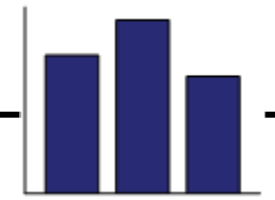
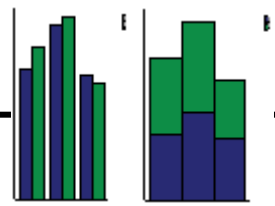
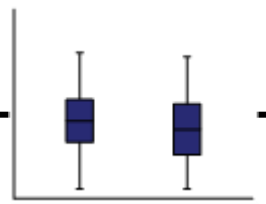
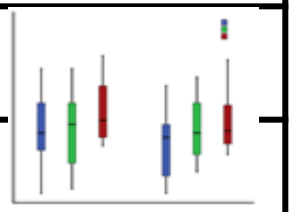
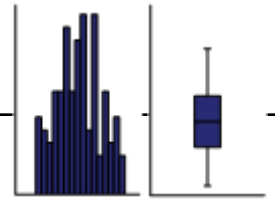
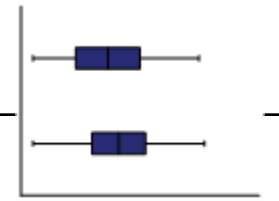
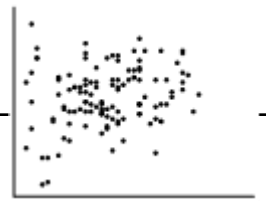
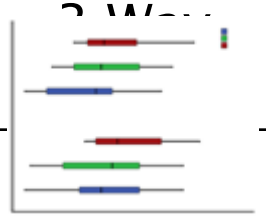
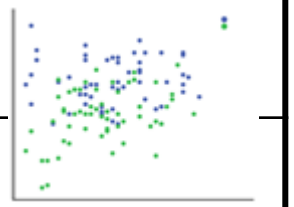
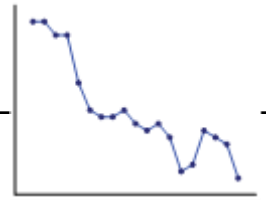
Right skewed = positive skewed

- Log
- Square root: sqrt(x)
- Cube root ที่ a: x^(1/a)
  - ใช้สำหรับค่าติดลบได้

Left skewed = negative skewed

- Square
- 10^(x)
- หรือ ใช้ตัวเลขมากกว่าค่าที่มากที่สุด (เป็นค่าคงที่) ลบทุกค่าออกค่าคงที่นั้น จะทำให้ข้อมูลกลับจาก left skewed ไปอีกด้าน แล้วใช้ log

Using appropriate statistics and graphs…

				Z=Cat.	Z=Cat.
X=Cat.					
X=Cont.					
X=Time	N/A	N/A		N/A