Statistics and research

Usaneya Perngparn
Chitlada Areesantichai

Drug Dependence Research Center
(WHOCC for Research and Training in Drug Dependence)
College of Public Health Sciences
Chulolongkorn University, Bangkok, THAILAND

Statistics used in research

1. Sampling technique
2. Sample size
3. Reliability check
4. Data analysis

Data analysis

1. Building file from questionnaire
2. Reliability check
3. Summarisation of major characteristics of population
4. Summarisation of major characteristics of sample
5. Causal finding
6. Classify groups/cross sectional
7. Prediction

Data analysis

Reliability check

1. Editing - possible check
   - cross check
2. Cronbach Alpha

Data analysis

Sumarised data

Nominal and Ordinal Scale

- number

- percentage

Interval and Ratio Scale

- Mean, Median, Standard diviation
- Graph
- Boxplot
- etc

Nominal and Ordinal Scale

1. Univariate Variable

-frequency table and percentage

Contraceptive use	N	%
Yes	300	30
No	700	70
Total	1000	100

Incident Rate: IR

$$IR = \frac{\text{Number of new case during the period of study}}{\text{Total population}}$$

2. Bivariate Variable

- Ratio : R

$$R = \frac{\text{Male}}{\text{Female}}$$

(Sex ratio)

- Risk Ratio : RR

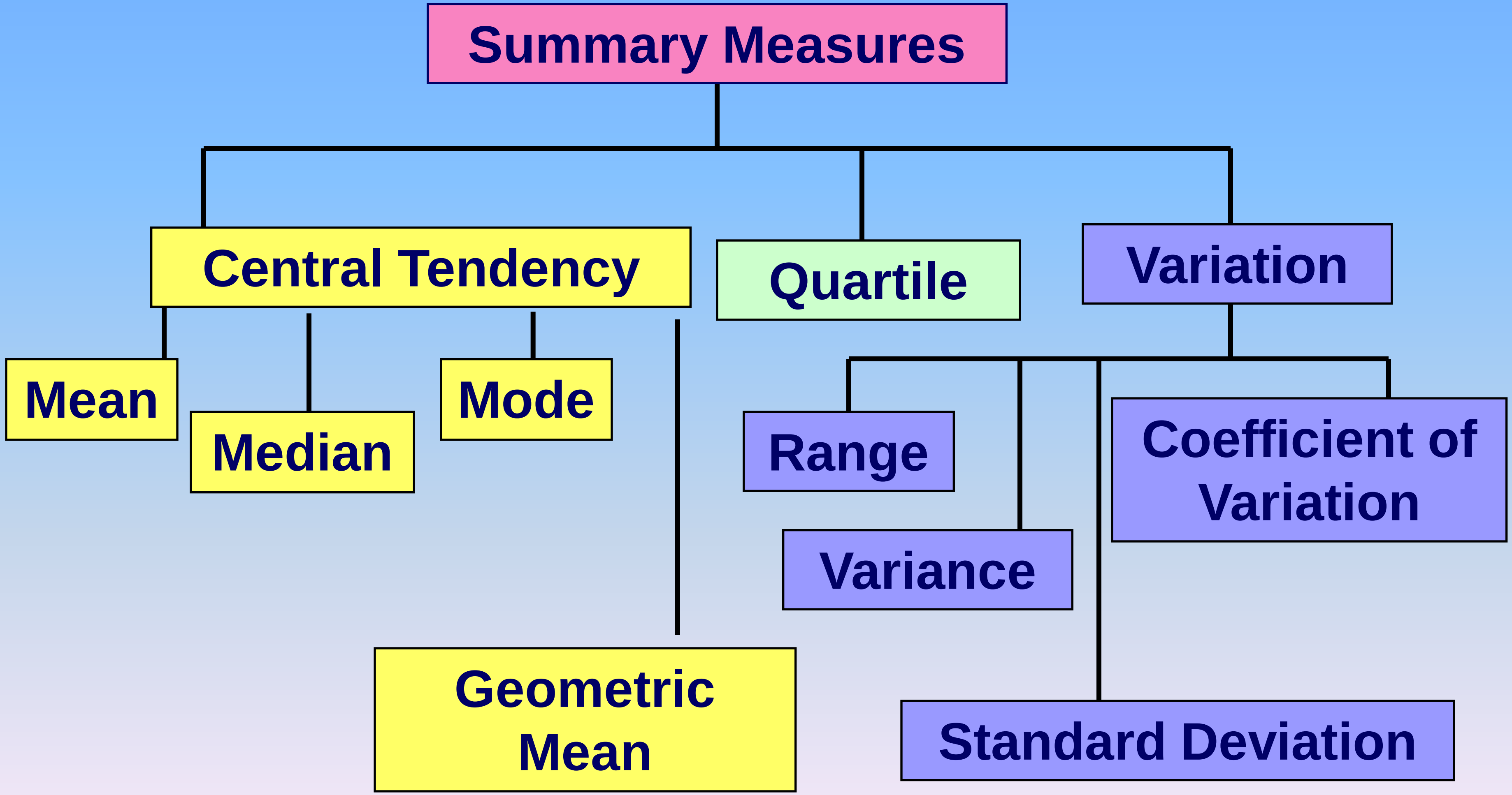
$$RR = \frac{p_{\text{exposed}}}{p_{\text{non-exposed}}}$$

RR is the risk of an event (or of developing a disease) relative to exposure. Relative risk is a ratio of the probability of the event occurring in the exposed group versus a non-exposed group.

Descriptive Statistics

- Measures of central tendency
  - Mean, median, mode, geometric mean, midrange
- Measure of variation
  - Range, Interquartile range, variance and standard deviation, coefficient of variation
- Shape
  - Symmetric, skewed, using box-and-whisker plots

Summary Measures



Summary Measures

Central Tendency

Mean

Median

Mode

Geometric Mean

Quartile

Variation

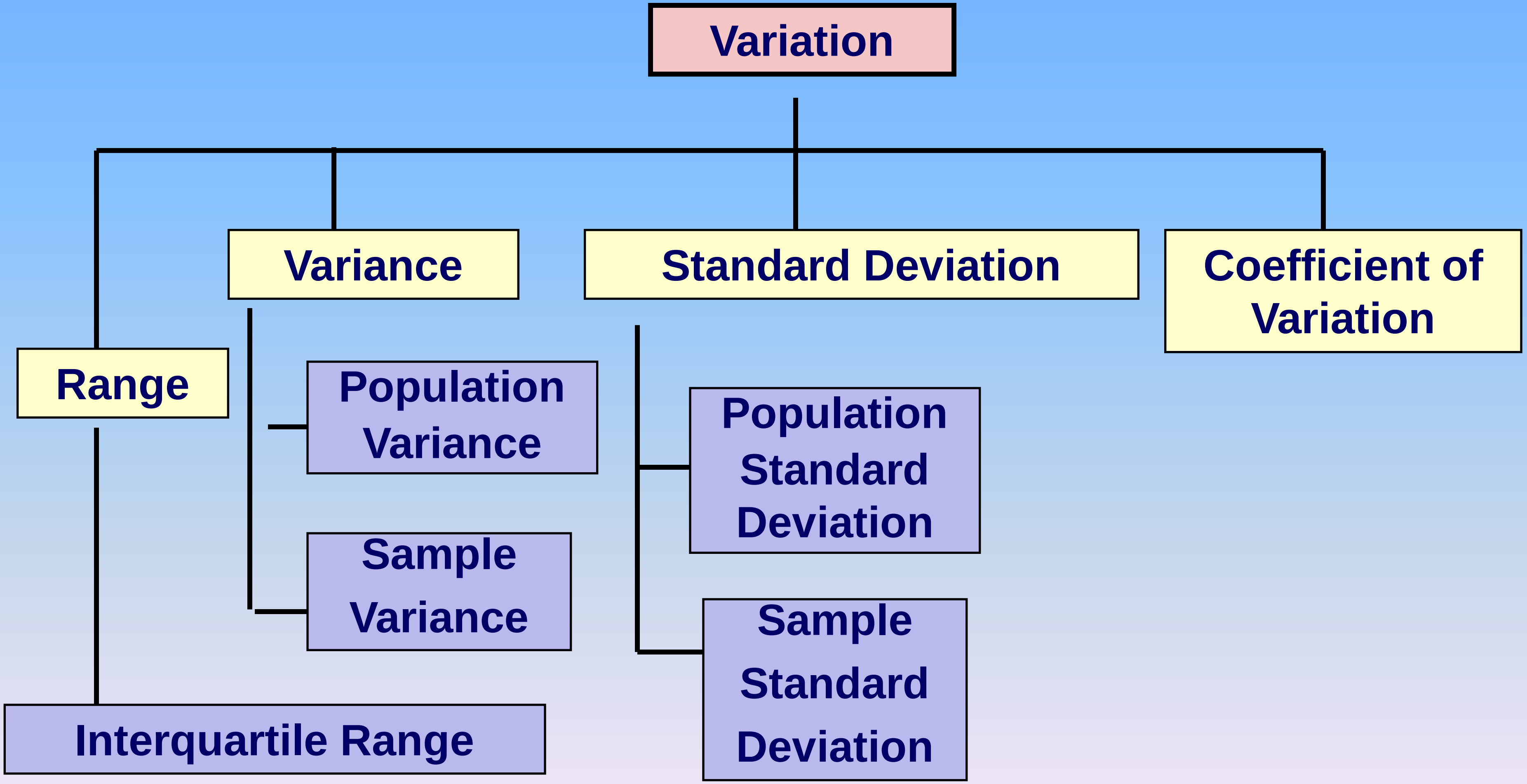
Range

Variance

Standard Deviation

Coefficient of Variation

Measures of Variation


Variation

Range

Interquartile Range

Variance

Population Variance

Sample Variance

Standard Deviation

Population Standard Deviation

Sample Standard Deviation

Coefficient of Variation

Coefficient of Variation

- Measures relative variation
- Always in percentage (%)
- Shows variation relative to mean
- Is used to compare two or more sets of data measured in different units
-

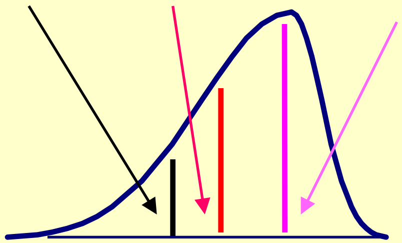
$$CV = \left(\frac{S}{\bar{X}}\right)100\%$$

Shape of a Distribution

- Describes how data is distributed
- Measures of shape
  - □ Symmetric or skewed

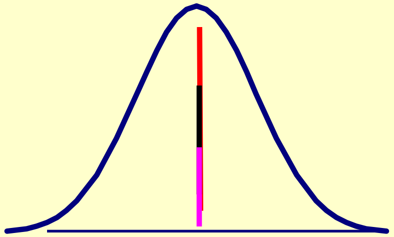
Left-Skewed

Mean < Median < Mode



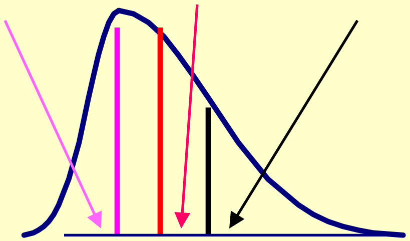
Symmetric

Mean = Median =Mode



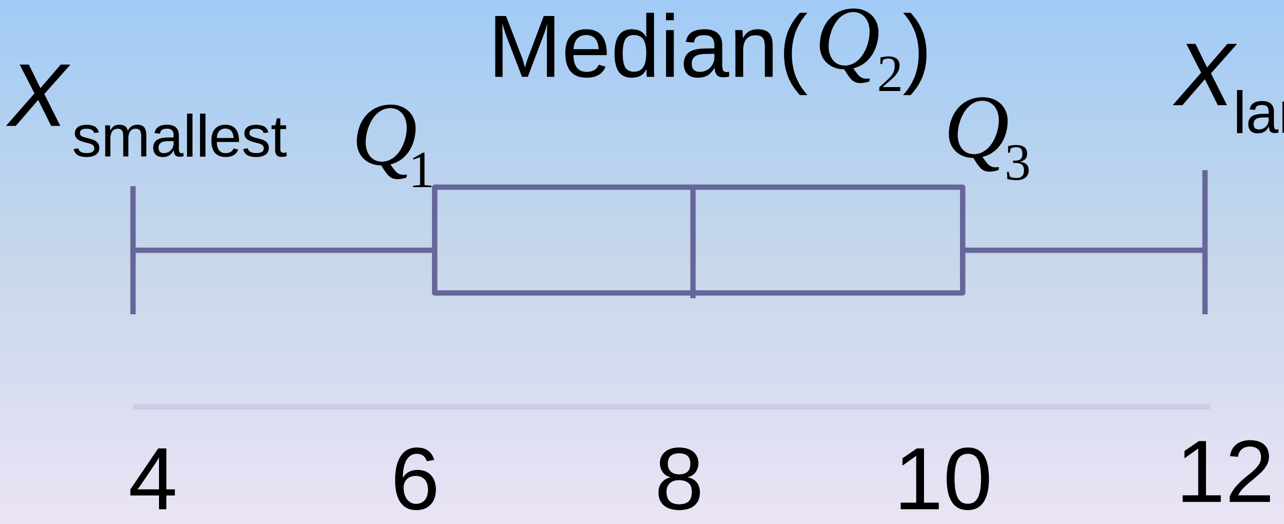
Right-Skewed

Mode < Median < Mean



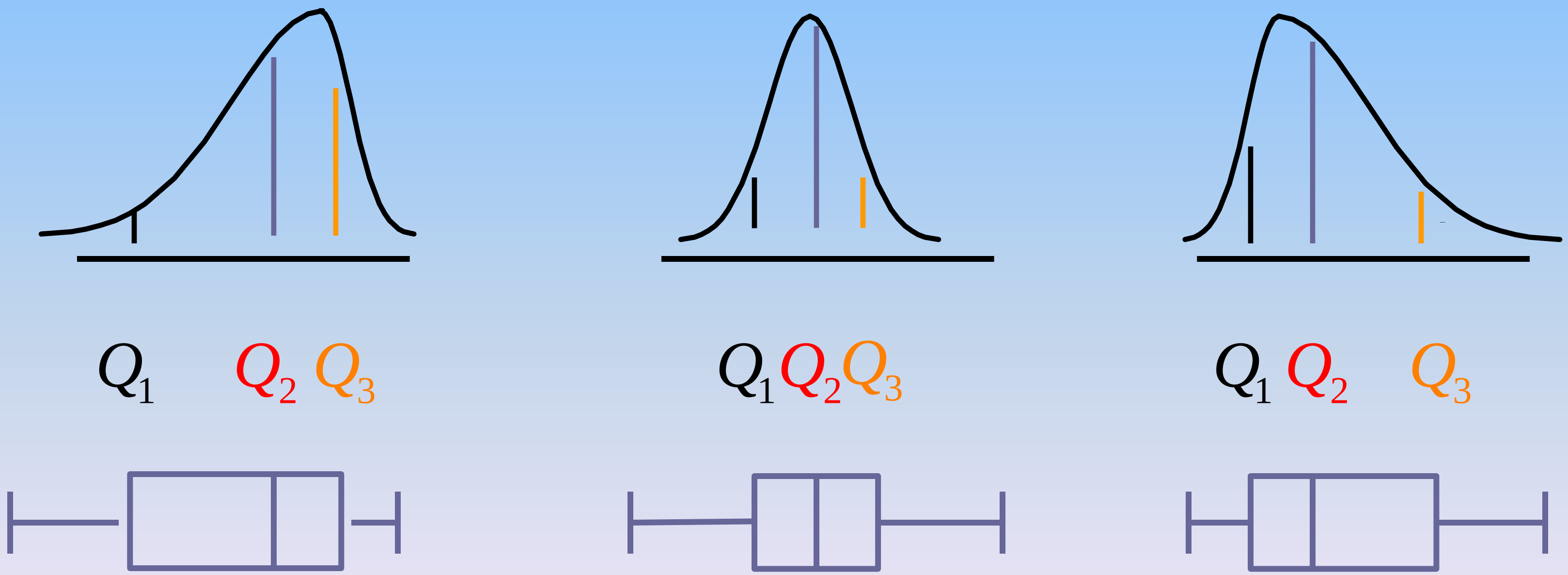
Exploratory Data Analysis

- Box-and-whisker plot
  - □ Graphical display of data using 5-number summary



Distribution Shape and Box-and-Whisker Plot

Left-Skewed **Symmetric** **Right-Skewed**



Data analysis

Testing Hypothesis:

All hypothesis tests are conducted the same way. The researcher states a hypothesis to be tested, formulates an analysis plan, analyzes sample data according to the plan, and accepts or rejects the null hypothesis, based on results of the analysis.

Statistical Testing Hypothesis

Every hypothesis test requires the analyst to state

-a null hypothesis (H_o) and an alternative hypothesis (H_1)

-Significance level = 0.01, 0.05, or 0.10; but any value between 0 and 1 can be used.

-Test method: The test method involves a test statistic and a sampling distribution. Computed from sample data, the test statistic might be a mean score, proportion, difference between means, difference between proportions, z-score, t-score, chi-square, etc.

Statistical Testing Hypothesis

- Quantitative data

$$\mu, \sigma^2, \mu_1 - \mu_2, \frac{\sigma_1^2}{\sigma_2^2}$$

- Qualitative data: p , p_1-p_2
- Correlation test

Statistical Testing Hypothesis

1. Null Hypothesis: H_o

2. Alternative Hypothesis: H_1 or H_a

H_1 or H_a must be in the difference direction

H_o : average income =10,000 US$/month $(\mu = 10{,}000)$

H_1 : average income ≠ 10,000 US$/month $(\mu \neq 10{,}000)$

$(\mu < 10{,}000)$ $(\mu \geq 10{,}000)$

ERROR

1. Type I error is the error of rejecting the null hypothesis when it is true -- of saying an effect or event is statistically significant when it is not. The projected probability of committing type I error is called the level of significance. For example, for a test comparing two samples, a 5% level of significance (a = .05) means that when the null hypothesis is true (i.e. the two samples are part of the same population), you believe that your test will conclude "there's a significant difference between the samples" 5% of the time.

$$\alpha = \mathbf{P\ (\ reject\ H_0\ /\ H_0\ \ true\ )}$$

$$\alpha = \mathbf{Level\ of\ Significance}$$

ERROR

2. Type II error is the error of accepting the null hypothesis (H_0) when alternative hypothesis (H_1) is true

H_0 not true

$\beta = P(\text{accept } H_0 / H_0 \text{ not true})$

Reducing α made β increased

Reducing β made α increased

Adjusting the sample size can reduce both α and β

Proportional test

1. Binomial Test : small n

2. Z test : large n

3. Chi-Square Test for 1-way classification

    3.1 Pearson Chi-Square test

    3.2 Likelihood ratio test

    3.3 Fisher’s Exact test

Correlation analysis

>2 Variables

- Multiple Regression
- Discriminant Analysis
- Logistic Regression
- ANOVA (Analysis of Variance)
- ANOVA (Analysis of Covariance)
- MANOVA (Multivariate Analysis of Variance)
- Log-Linear Model

H_o: Var(gr 1.) = Var(gr. 2)= . . . . . = var(gr.k)

H_1: At least 1 pairs of var(gr.i) # Var(gr. J)

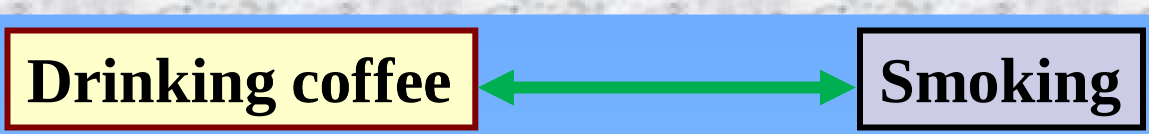
Reject H_o(Accept H_1) if Sig. < alpha

Multiple Comparison

1. Variance of each group is equal:
   LSD, Bonferronni, Tukey, Duncan

2. Variance of each group is unequal:
   Dunnett, Tamhane etc.

Two-Way table or Contingency Table

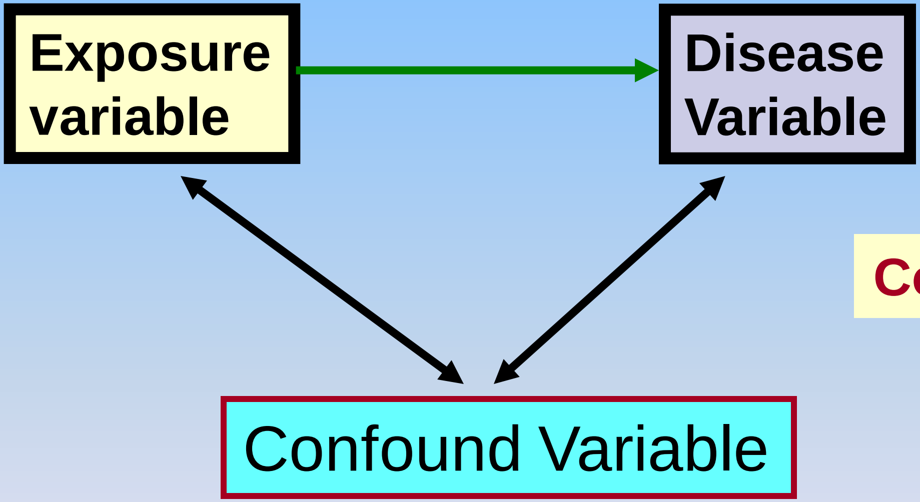

Drinking coffee

Smoking

Drink coffee (# per day)	Smoking (# per day)			
	0	1-2	3+	Total
0	4	9	15	28
1-5	6	10	13	29
6 +	5	9	16	30
Total	15	28	44	87

- Pearson Chi-Square test
- Likelihood ratio Chi-Square
- Fisher Exact test
- Etc.

Analysis of 2-group correlation for 2X2 table ĉ control other factors


Exposure variable

Disease Variable

Confound Variable

Cochran’s Mantel-Haenszel

Multivariate Analysis

- Discriminant Analysis ,
- Regression and Correlation Analysis
- Cluster Analysis
- Factor Analysis
- Principal Analysis
- Multidimensional Scaling
- Path Analysis
- etc.

Statistics for Research
By using
SPSS for Windows

Chitlada Areesantichai

Usaneya Perngparn

WHO CC for Research and Training in Drug Dependence
College of Public Health Sciences

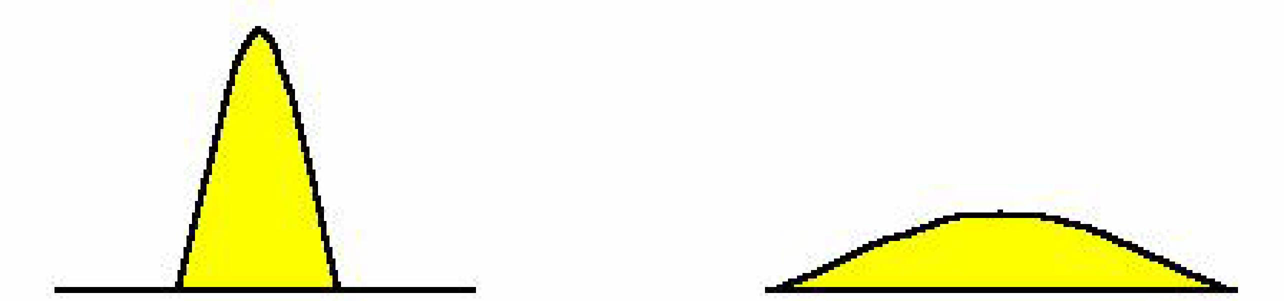
Chulalongkorn University

Bangkok THAILAND

- Mean is the average and is computed as the sum of all the observed outcomes from the sample divided by the total number of events.
- Median is the 'middle value' in your list.
- Mode: the *mode* of a set of data is the number with the highest frequency.
- Range of a set of data is the difference between the highest and lowest values in the set.

Level of Data	Stat used for Central Tendency measurement
Interval / Ratio	Mode, Median, Mean
Ordinal	Mode, Median
Nominal	Mode

Variance: In statistics, a variance is also called the mean squared error. The variance is one of several measures that statisticians use to characterize the dispersion among the measures in a given population. To calculate the variance, it is necessary to first calculate the mean or average of the scores. The next step is to measure the amount that each individual score deviates or is different from the mean. Finally, you square that deviation by multiplying the number by itself. Numerically the variance equals the average of the squared deviations from the mean. *(http://www.answers.com/topic/variance-1)*



Small variance

Large variance

- Standard Deviation or Std dev or SD or S**:** A measure of the dispersion of a set of data from its mean. The more spread apart the data is, the higher the deviation. Standard deviation can also be calculated as the square root of the variance.

  http://www.answers.com/topic/standard-deviation

- Standard Error or Std err or SE of a statistic is the *standard* deviation of the sampling distribution of that statistic.

Sample random sampling

Small sample size causes higher error

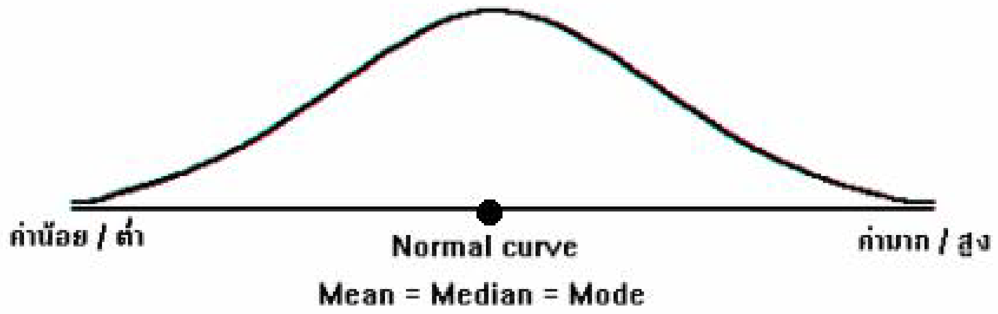
Large sample size causes low error

Skewness is a measure of the degree of asymmetry of a distribution. If the left tail (tail at small end of the distribution) is more pronounced than the right tail (tail at the large end of the distribution), the function is said to have negative skewness. If the reverse is true, it has positive skewness. If the two are equal, it has zero skewness.

Normal distribution

Symmetric

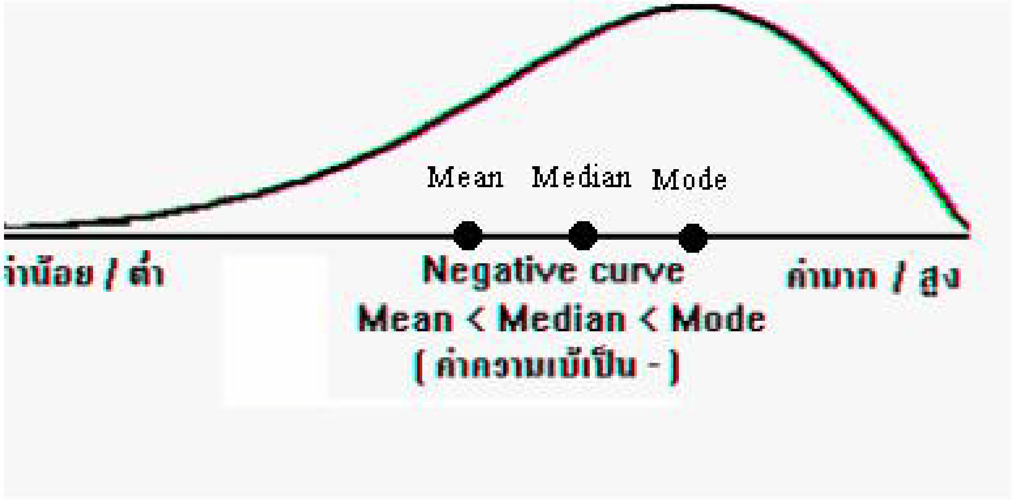
Mean = Median = Mode



ภาพ 3.1ก เส้นโค้งปกติ

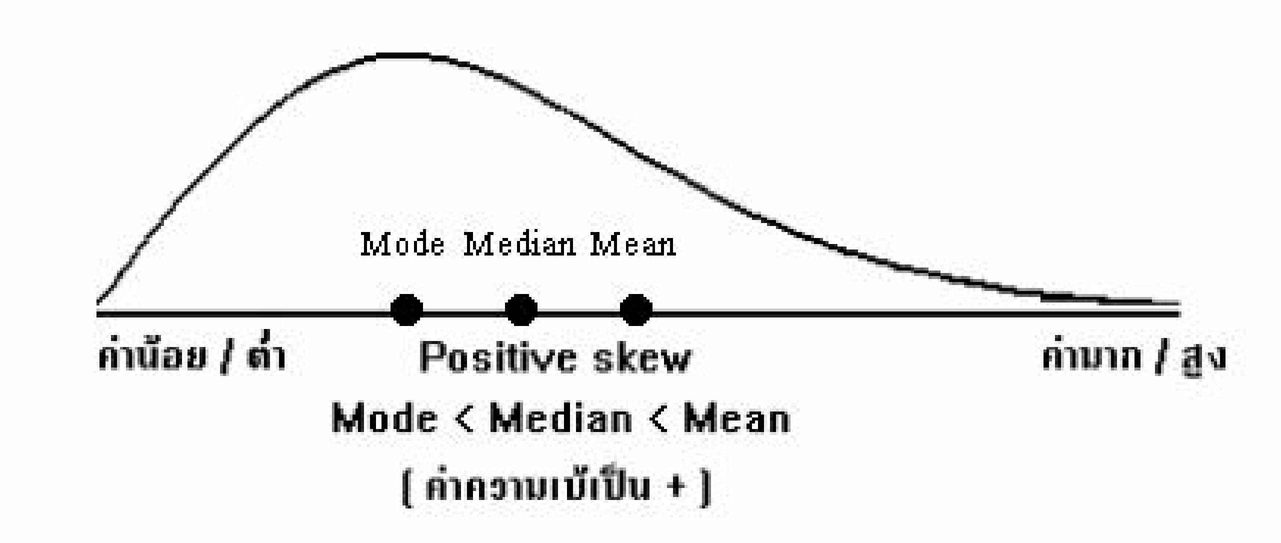
Skewed to the left or negative skew

Mean < Median < Mode



ภาพ 3.1ข เส้นโค้งเบ้ซ้ายหรือเบ้ลบ

Skewed to the right or positive skew
Mode < Median < Mean



Note:

Skewness = - Skewed to the left

Skewness = 0 No skew

Skewness = + Skewed to the right

- Normal distribution
- Reliability test
- Relation test
- Comparison test

t-test

1.Independent t-test

2. Paired t-test

Normal distribution test

SPSS: Analyze → Descriptive Statistics → Descriptive

Analyze → Descriptive Statistics → Explore

Reliability(Alpha Coefficient or Cronbach)

Use for Rating Scale questions.

Cronbach's alpha can be written as a function of the number of test items and the average inter-correlation among the items

1. SPSS: Analyze → Scale → Reliability Analysis
2. Choose item in Data, Put into item
3. Choose Alpha Model
4. If the statistic detail is needed

   Click Statistic button
5. If the individual case is needed

   Choose List item labels

Output Alpha Coefficient or Cronbach

```
****** Method 1 (space saver) will be used for this analysis ******

RELIABILITY ANALYSIS - SCALE (ALPHA)


Reliability Coefficients

N of Cases =     50.0        N of Items = 10

Alpha =   .7358
```

Pearson Correlation

1. SPSS: Analyze →Correlate →Bivariate →Correlation
2. 2 Variables or more, put variable math, stat
3. At Correlation Coefficients, Choose Pearson
4. Test of significance, Choose Two-tailed
5. At Flag significant correlation, Set a symbol to identify the significant correlation of those two variables.
6. Click option to identify other values
7. Click ok

Chi-Square

Test the distribution ratio

H_0 : Population ratio is equal

H_1 : Population ratio is not equal

1. SPSS: Analyze → Nonparametric test
2. Choose variable in Test variable list, identify value
3. Click Option to identify other related values, Click OK

Chi-Square test

1. Observed number
2. Expected number
3. Residual is the different between observation and expectation number
4. Chi-square test (χ^2 test) is any hypothesis test in which the sampling distribution of the test statistic is a chi-square distribution when the null hypothesis is true, or any in which this is *asymptotically* true, meaning that the sampling distribution can be made to approximate a chi-square distribution as closely as desired by making the sample size large enough.

Chi-Square Ratio

1. SPSS: Analyze → Nonparametric test
2. Move variables in Test variable list, specify value
3. Define Expected Values, click value
4. Set expected ratio
5. Click option, press OK

Mann-Whitney U-test

For 2 independent groups

1. SPSS: Analyze → Nonparametric test
2. Move variables in Test variable list, put variable score_M
3. Move target variables into Grouping variable
4. Click Define Group to verify variable value, e.g. sex : 1=male 2=female
5. In Test Type, choose Mann-Whitney Utest
6. Define others
7. Click OK

Independent t-test

Independent t-test is used to test for a difference between two independent groups (like males and females) on the means of a continuous variable.

SPSS: Analyze → Compare Means → Independent Sample t-test

Paired t-test

Paired t test provides an hypothesis test of the difference between population means for a pair of random samples whose differences are approximately normally distributed. The most common design is that one nominal variable represents different individuals, while the other is "before“ and "after" some treatment.

SPSS: Analyze → Compare Means → Paired Sample T Test

ANOVA (Analysis of Variance)

At least 2 independent variables and one dependent variable

SPSS: Analyze →Compare Means → One-Way ANOVA

1. Move to variable in Dependent list and move another variable to factor

2. Click Post Hoc to set sig level. If it is significant, then analyze Post Hoc Comparison to find out which couple are significant. If the samples are not equal, choose Scheffe

3. Click continue

4. Click contrast or option

5. Click ok