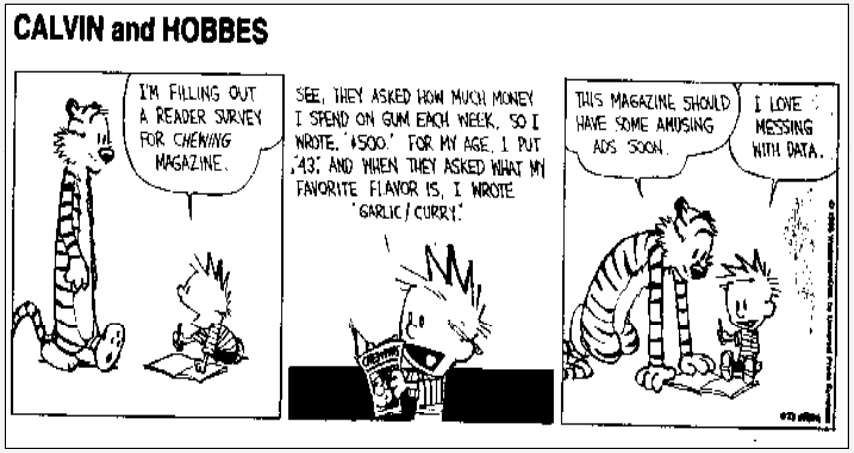
Introduction to Descriptive and Inferential Statistics

Jaranit Kaewkungwal, Ph.D.
Faculty of Tropical Medicine
Mahidol University

Data & Variables



Types of Data

QUALITATIVE

Data expressed by type

Data that has been described

QUANTITATIVE

Data classified by numeric value

Data that has been measured or counted

QUALITATIVE and QUANTITATIVE data are not mutually exclusive

Types of Data: Qualitative (Categorical) Data

NOMINAL DATA

- values that the data may have do not have specific order
- values act as labels with no real meaning
- Binomial: two possible values (categories, states)
- Multinomial: more than two possible values (categories, states)

e.g. Health status healthy =1 sick=2
e.g. Treatment new regimen = 1 standard regimen = 2
e.g. hair colour brown =1 blond =2 black =100

ORDINAL DATA

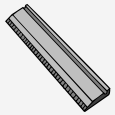
- values with some kind of ordering
- data that has been measured or counted

e.g. social class: upper=1 middle = 2 working = 3
e.g. glioblastoma tumor grade: 1 2 3 4 5
e.g. position in a race: 1 st 2 nd 3 rd

Types of Data: Quantitative Data

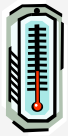
DISCRETE
- distinct or separate parts, with no finite detail

e.g children in family



CONTINUOUS
- between any two values, there would be a third

e.g between meters there are centimetres



INTERVAL
- equal intervals between values and an arbitrary zero on the scale

e.g temperature gradient

RATIO
- equal intervals between values and an absolute zero

e.g body mass index

Adapted from: Dr. Craig Jackson, University of Central England

UCE Birmingham

Examples of Data Coding

Site Number ___ Subject Number ___ Chk ___ Initials ___

Demographics

1. Gender

1 ☐ Male at birth 2 ☐ Female at birth

3 ☐ Surgical / medical correction to male

4 ☐ Surgical / medical correction to female

2. Date of Birth

dd mm yy

Nominal/Cat. Var

3. Race / Ethnicity

1 ☐ Thai 2 ☐ Hill-Tribe 3 ☐ Unknown 4 ☐ Other, specify: _____

15. ในช่วง 6 เดือนที่ผ่านมา ท่านมีเพศสัมพันธ์กับคู่นอนที่ท่านอยู่ด้วยบ่อยแค่ ไหน

(In the last six months, how often have you had sexual intercourse with your live-in partner? Would you say:)

1 ☐ น้อยกว่าเดือนละครั้ง (Less than once a month)

2 ☐ ทุกเดือนแต่ไม่ทุกสัปดาห์ (Every month, but not every week)

3 ☐ ทุกสัปดาห์แต่ไม่ทุกวัน (Every week, but not every day)

4 ☐ เกือบทุกวัน (About every day)

~~88~~ ☐ ไม่ทราบ (Don't Know)

~~99~~ ☐ ไม่ตอบ (Non-response)

Exclude from Analysis?

Ordinal/Cat. Var

Examples of Data Coding

Site Number ___ Subject Number ___ Chk ___ Initials ___ Date of Exam ___ dd mm yy

Physical Exam

1. Height ___ cm
2. Weight ___ kg
3. Temperature ___ °C Pulse ___ /min
5. Respiration ___ /min
6. Blood Pressure ___ / ___ mmHg (systolic / diastolic)

Cont.

Instructions: *Physical exam directed by history, system review and symptoms only.*

1 2 99

Body System	Normal	Abnormal	Not Done	Comment on all abnormalities
7. HEENT	☐	☐	☐	
8. Lymph Nodes	☐	☐	☐	
9. Cardiovascular	☐	☐	☐	

see pp. 4-7 Labs pg 6

Page 50

Genentech, Inc. Genentech, Inc. Genentech, Inc. Genentech, Inc. Genentech, Inc.

Study #: V0633g
Drug: MN Recombinant Glycoprotein 120/HIV-1 Vaccine
Study: Phase I/II Recovering IVDU in Thailand

Subject Identification
Subject Number
Subject Initials
Specimen Date: ___ ___ ___ (Mo Day Yr)

LABORATORY DATA (MONTH 12.5)

		Results	ND
CHEMISTRY	Creatinine		☐
	ALT/SGPT		☐
HEMATOLOGY	WBC		☐
	RBC		☐
	Hemoglobin		☐
	Hematocrit		☐
	Platelet Count		☐
	Segmented Neutrophils		☐
	Banded Neutrophils		☐
	Lymphocytes		☐
	Monocytes		☐
	Eosinphils		☐
	Basophils		☐
	Other, Specify: ___		☐

		N/A	Pos	Neg	ND
HEPATITIS	HBs Ag	☐	☐	☐	☐
	Anti HBc	☐	☐	☐	☐
	HCV Ab	☐	☐	☐	☐

		Results	ND
URINALYSIS	Urine Albumin		☐
	Urine Nitrite		☐
	Urine Esterase		☐
	Urine Glucose		☐

		N/A	Pos	Neg	ND
	Urine Pregnancy	☐	☐	☐	☐
	Urine Opiates	☐	☐	☐	☐

Cont. 99

If "Not Done", please check ND boxes to indicate this.
N/A means "Not Applicable" (for example, Pregnancy - males).

Example of Descriptive Statistics

Pergamon
Soc. Sci. Med. Vol. 47, No. 3, pp. 663–673, 1998
PII: S0277-9536(98)00143-9

0277-9536/98 $19.00 + 0.00

FACTORS RELATED TO PHYSICAL ACTIVITY: A STUDY OF ADOLESCENTS

RUNAR VILHJALMSSON* and THOROLFUR THORLINDSSON†
*Department of Nursing, University of Iceland, Eiriksgotu 34, 101 Reykjavik, Iceland and †Faculty of Social Sciences, University of Iceland, Sturlugotu, 101 Reykjavik, Iceland

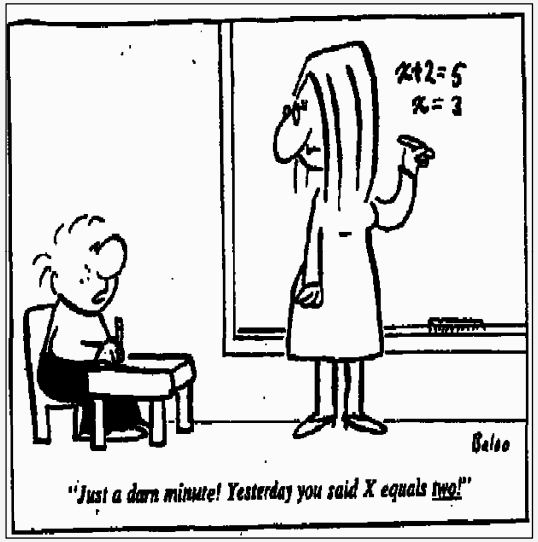
Table 1. Descriptive statistics (N = 1131)

Category and variable	Mean	Range	Standard deviation
Background			
Sex (1 = female)	0.488	0–1	0.499
Social class	1.948	1–3	0.812
Residence (1 = Reykjavik area)	0.444	0–1	0.494
Attitudes/beliefs			
Importance of sport	0.863	0–2	0.744
Importance of health improvement	1.805	0–2	0.445
Internal health locus of control	3.437	0–4	0.825
Other's physical activity			
Father	0.732	0–2	0.822
Mother	0.740	0–2	0.830
Older brother	0.603	0–2	0.767
Older sister	0.462	0–2	0.688
Best friend	1.448	0–2	0.755
Main teacher	1.352	0–2	0.677

Constant vs. Variable

Variables are the specific properties that have the ability to take different values.

Constants are the specific properties that cannot vary or won't be made to vary.



Terminology - Variables

INDEPENDENT

(syn: treatment, experimental, predictor, input, exposure, explanatory variable) is a stimulus or activity that is ***identified or manipulated*** to predict the dependent variable; they are considered as the causal factors, or that you may manipulate.

e.g. new drug, working hours, exposure, worker attitudes, policies

DEPENDENT

(syn: Effect, criterion, criterion measure, outcome, output variable) is a response that the researcher wanted to predict; they are considered as the ***outcomes of the treatments or the responses*** to changes in the independent variables.

e.g. Symptomotology, productivity, accident rates, attitudes, health status, performance on neuropsychological test

Adapted from: Dr. Craig Jackson, University of Central England

UCE Birmingham

Terminology - Variables

EXTRANEOUS

Extraneous variable is a variable that has a potential to ***distorts the relationship between dependent and independent*** **variables.**

• **Controlled extraneous variables** are recognized before the study is initiated and are *controlled in the design and selection criteria*.

• **Uncontrolled extraneous variables** are recognized before the study is initiated or, sometimes, even if recognized cannot be controlled in the design and selection phase. Usually an attempt is made to *assess and adjust them through sophisticated statistical* tools.

e.g., Working hours, temperatures, extraneous exposure, diet, class, income, Ambient noise and temperature in testing room

Adapted from: Dr. Craig Jackson, University of Central England

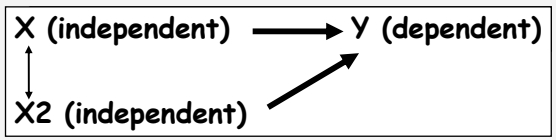
UCE Birmingham

Study Variables

Independent Variables & **Dependent variables**

X ⟷ Y X (independent) ⟶ Y (dependent)

Extraneous variable



Study Variables

Confounding Variable

- When the *effects of two or more variables cannot be separated.*

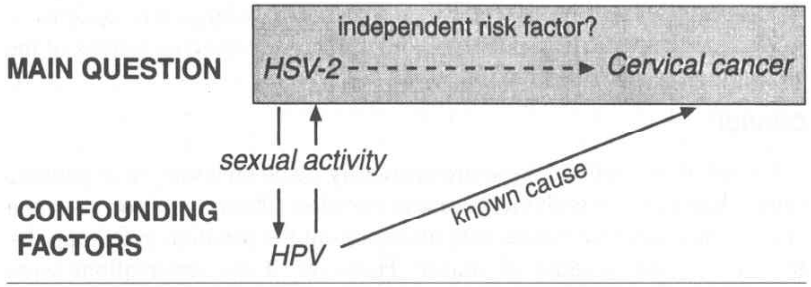


Figure 1.1. Confounding bias: Is herpesvirus 2 (HSV-2) a possible cause of cervical cancer? Only if its association with cervical cancer is independent of human papillomavirus (HPV) infection, known to be a cause of cervical cancer. Both viruses are related to increased sexual activity.

Study Variables

Confounding Variable

- When the *effects of two or more variables cannot be separated.*

		STD rate	
Condom Use	Yes	55/95	(61%)
	No	45/105	(43%)

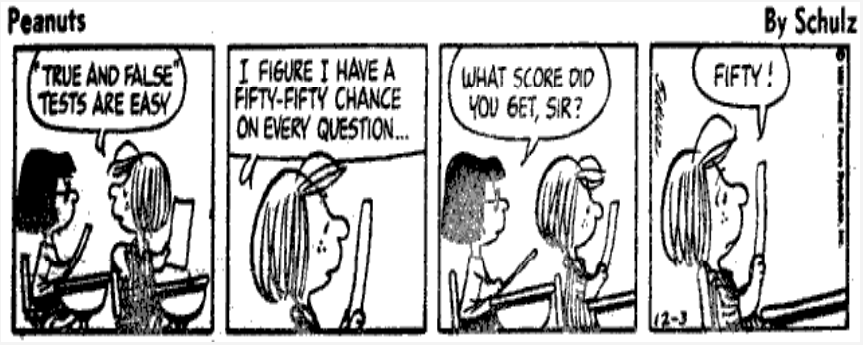
"Condom Use increases the risk of STD"

BUT ...

		STD rate	
# Partners < 5			
Condom Use	Yes	5/15	(33%)
	No	30/82	(37%)
# Partners ≥ 5			
Condom Use	Yes	50/80	(62%)
	No	15/23	(65%)

Explanation: Individuals with more partners are more likely to use condoms. But individuals with more partners are also more likely to get STD.

Bias & Chance



Measuring Outcomes:

Observed vs. Truth

Possible Explanations of Outcome Measured

Bias | **Chance** | **Truth**

Observed = Truth + Error

Systematic error + Random error

(Bias) (Chance)

Bias vs. Chance

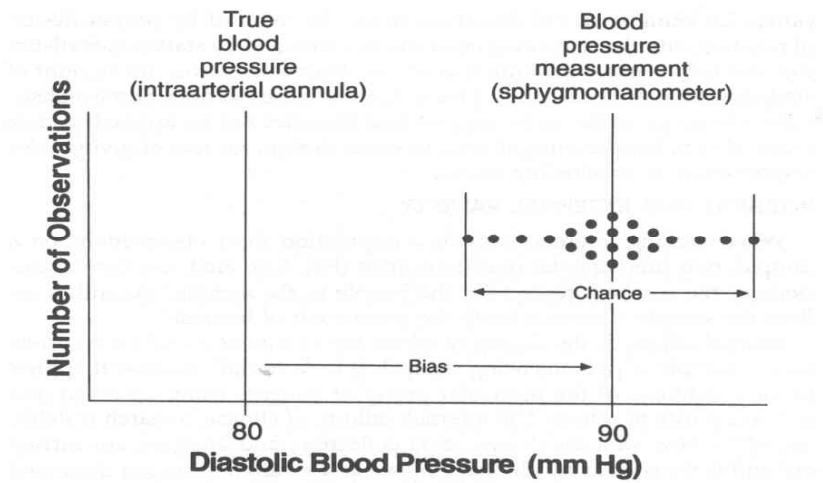


Figure 1.2. Relationship between bias and chance: Blood pressure measurements by intraarterial cannula and sphygmomanometer.

Bias:

- A process at any stage of inference tending to produce results that depart systematically from the true values.

Chance:

- The divergence of an observation on a sample from the true population value in either direction.
- The divergence due to chance alone is called *random variation*

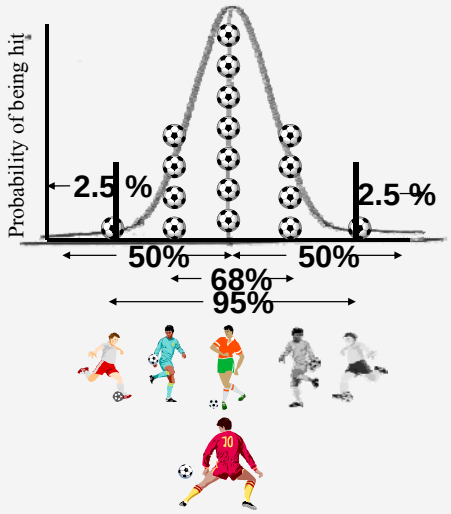
Bias and chance- are not mutually exclusive.

"A well designed, carefully executed study usually gives results that are obvious without a formal analysis and if there are substantial flaws in design or execution a formal analysis will not help."

Johnson AF. Beneath the technological fix. *J Chron Dis* 1985 (38), 957-961

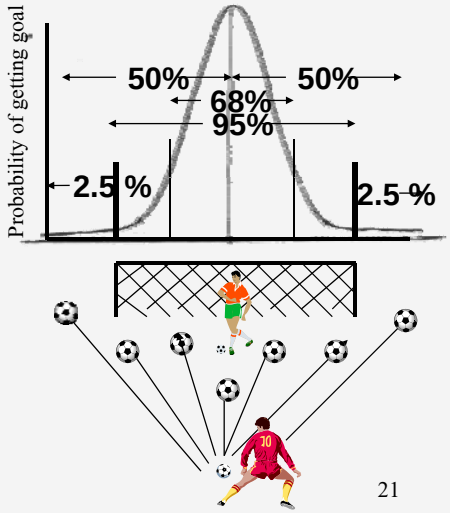
Chance

"Free kick"



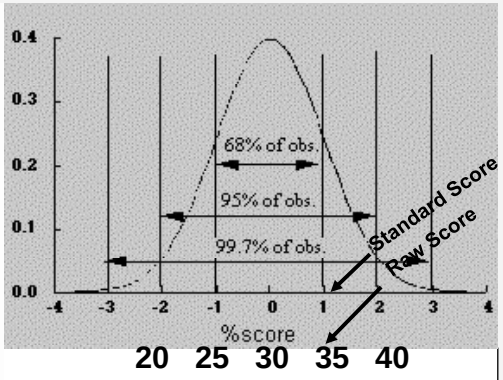
Chance

"Free kick"



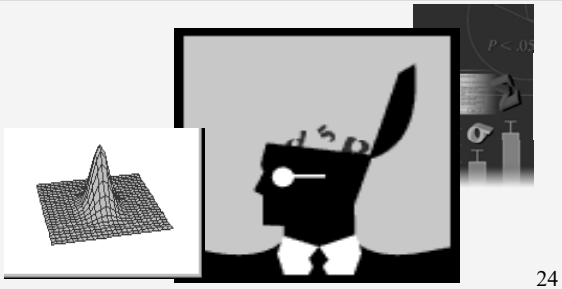
Normal Distribution

Normal Distribution in Descriptive Statistics



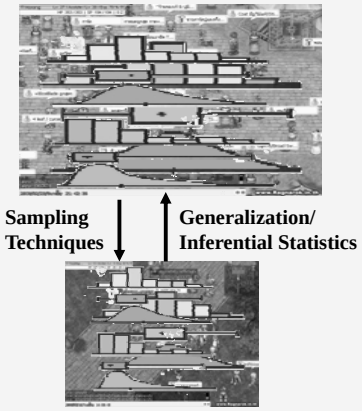
$\bar{X} = 30$;
SD = 5

Types of Statistical Methods



Types of Statistics

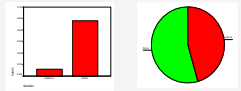
- By Level of Generalization
  - Descriptive Statistics
  - Inferential Statistics
    - Parameter Estimation
    - Hypothesis Testing
      - Comparison
      - Association
      - Multivariable data analysis
- By Level of Underlying Distribution
  - Parametric Statistics
  - Non-parametric Statistics



Descriptive Statistics

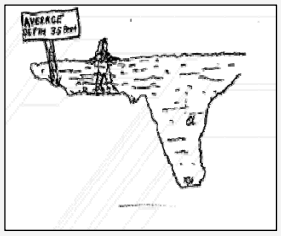
Descriptive Statistics

- **Measure of Location (Categorical Vars)**
  - **Frequency (f)**



- **Measure of Location (Continuous Vars)**
  - **Mean**
    Average

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \textit{or} \quad \mu = \frac{\sum_{i=1}^{n} x_i}{N}$$



  - **Median**
    Mid-point

$$X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9$$

  - **Mode**
    The Most Frequent

$$X_1 X_2 X_3 X_4 X_5 X_6 X_7 X_8 X_9$$

(1 2 2 2 2 3 3 4 5)

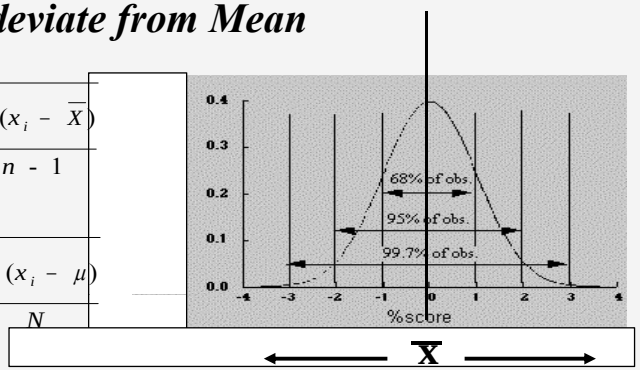
Descriptive Statistics

- **Measure of Spread**
  - **Range**
    Max - Min
  - **Standard Deviation / Variance**
    X_is deviate from Mean

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

or

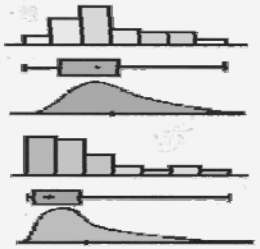
$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)}{N}}$$



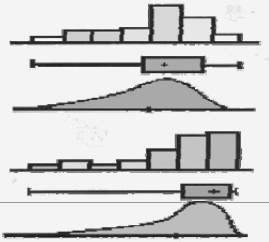
Descriptive Statistics

- **Measure of Shape**
  - **Normal Distribution**

  - **Skewed Distribution**
    - **Positively skewed**



Negatively skewed



Inferential Statistics

Inferential Statistics

- **Purpose of Inferential Statistics**
  - **Generalisabiliy of research results from Sample Statistics to Population Parameters**
- **Types of Inferential Statistical Methods**
  - **Parameter Estimation - to estimate the range of values that is likely to include the true value in population**

$\overline{X} \rightarrow \mu$

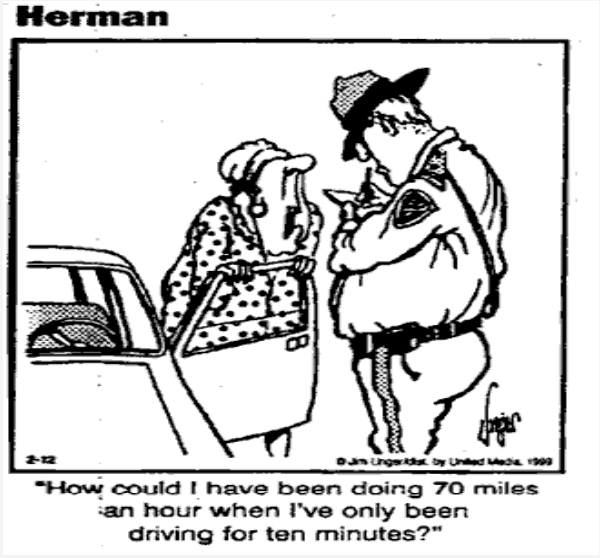
proportion $\rightarrow \Pi$

  - **Hypothesis Testing - to ask whether an effect (difference) is present or not among different groups**

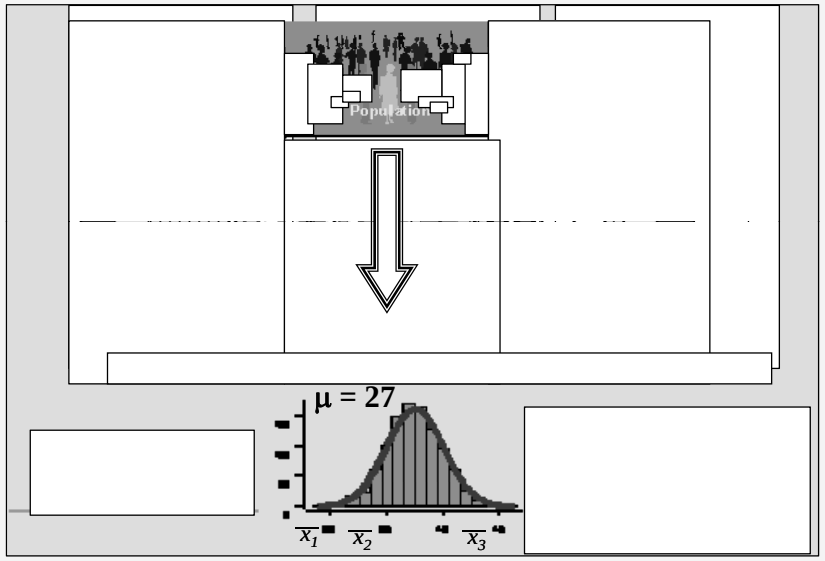
Ho: $\overline{X}_1 = \overline{X}_2 \rightarrow$ Ho: $\mu_1 = \mu_2$

Ho: $r_{xy} = 0 \rightarrow$ Ho: $\rho_{xy} = 0$

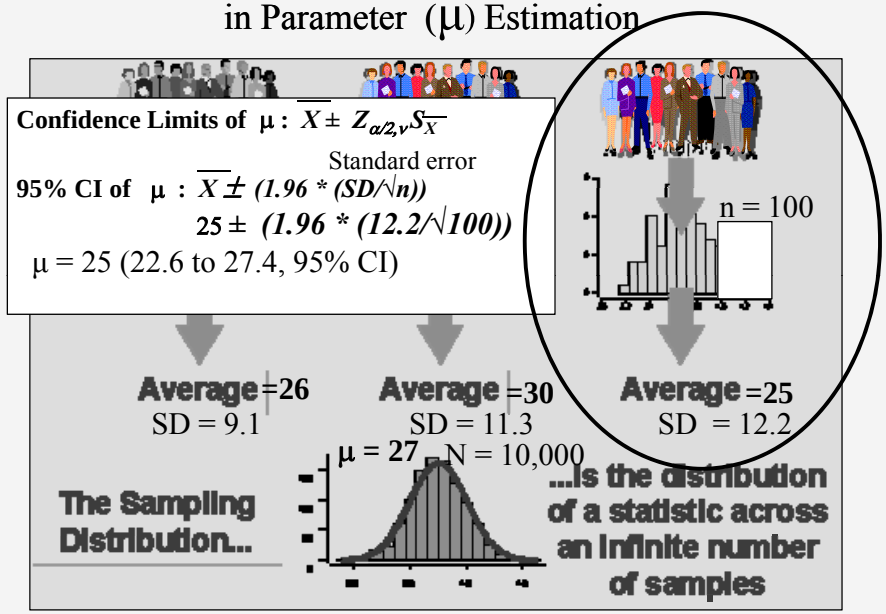
Inferential Statistics



Sampling Distribution



Sampling Distribution (Normal Distribution) in Parameter (μ) Estimation



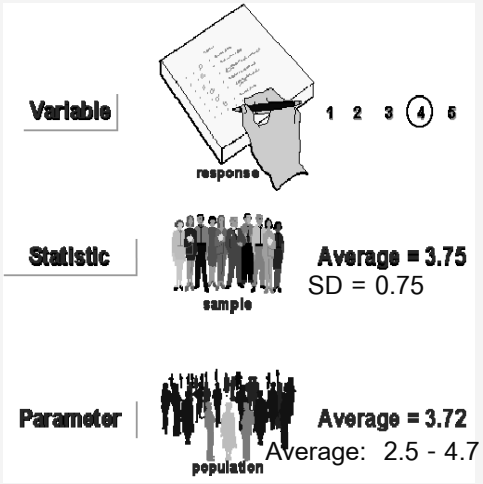
Parameter Estimation

Point Estimates:

Single values (Mean, Variance, Correlation, treatment effect, relative risk, etc.) representing characteristics in the whole population

Interval Estimates:

Ranges of values, usually centered around point estimates, indicating bounds within which we expect the true values for the whole population to lie (stability of the estimate)



Example Confidence Limits of μ: $\overline{X} \pm Z_{\alpha/2,\upsilon} S_{\overline{X}}$

Point estimates and confidence intervals are used to characterise the statistical precision of any rate (incidence, prevalence), comparisons of rates (e.g., relative risk), and other statistics.

• US adults have used unconventional therapy = 34% (31% - 37%, 95%CI)

• Sensitivity of clinical examination of splenomegaly = 27% (19 - 36%, 95%CI)

•Relative risk of lung cancer of smoker vs. non-smoker = 5.6 (2.1 - 8.9, 95%CI)

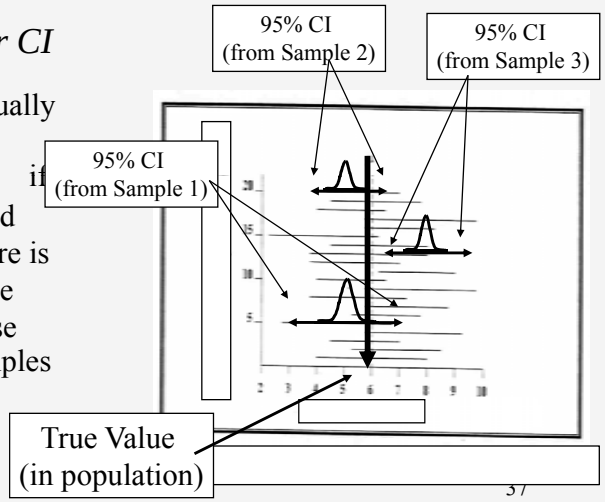
•Relative risk of HIV infected of male vs. female = 2.1 (0.5 - 6.9, 95%CI)

Parameter Estimation

Consideration in confidence level of the estimate

Select a cut-point for CI

Confidence Interval, usually set at 95% CI, can be interpreted such that - if the study is unbiased and repeated 100 times, there is 95% chance that the true value is included in these intervals of the 100 samples



Hypothesis Testing

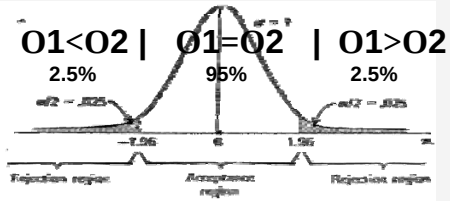
Hypothesis Testing

- ***Hypothesis & Tail of the test***
  - ***One-sided vs. Two-sided Test***

Two-sided test:

Ex Ho: Outcome 1 = Outcome 2

Ha: Outcome 1 ≠ Outcome 2



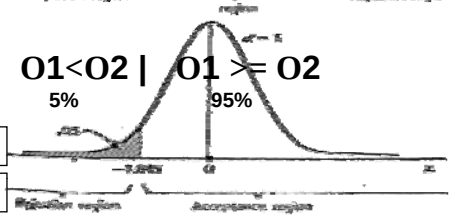
One-sided test:

Ex **Ho: Outcome 1 ≤ Outcome 2**

Ha: Outcome 1 > Outcome 2

Ho: Outcome 1 ≥ Outcome 2

Ha: Outcome 1 < Outcome 2



Hypothesis Testing

- Basic steps in hypothesis testing

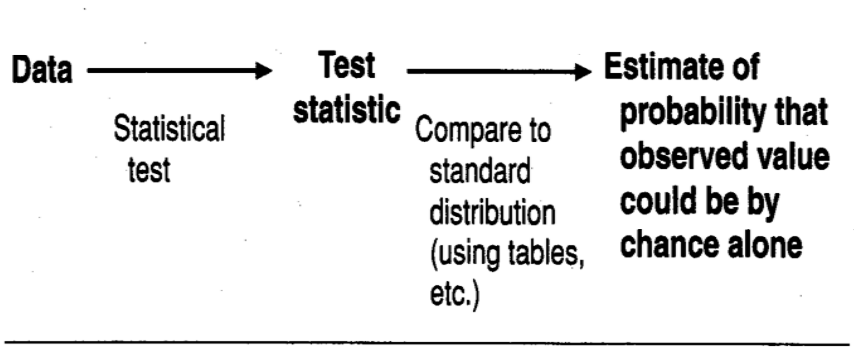
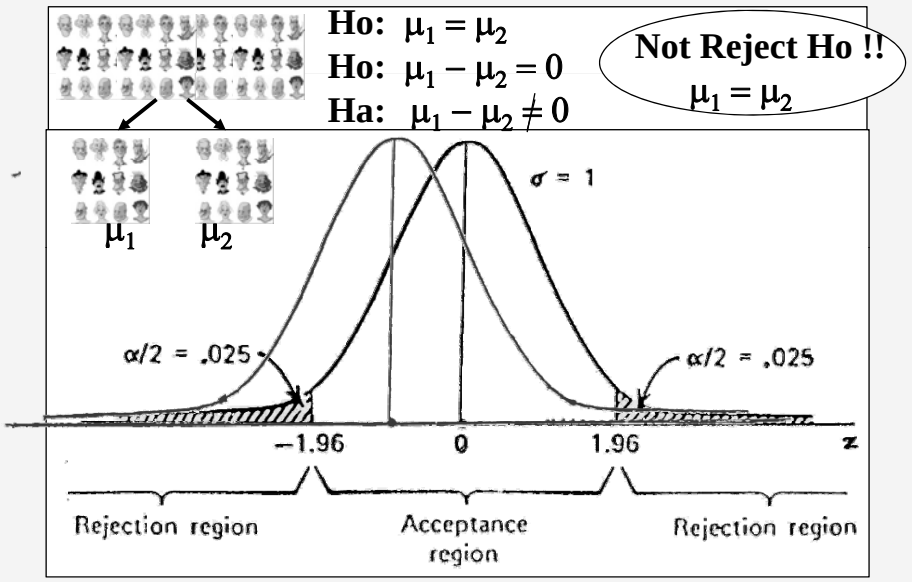
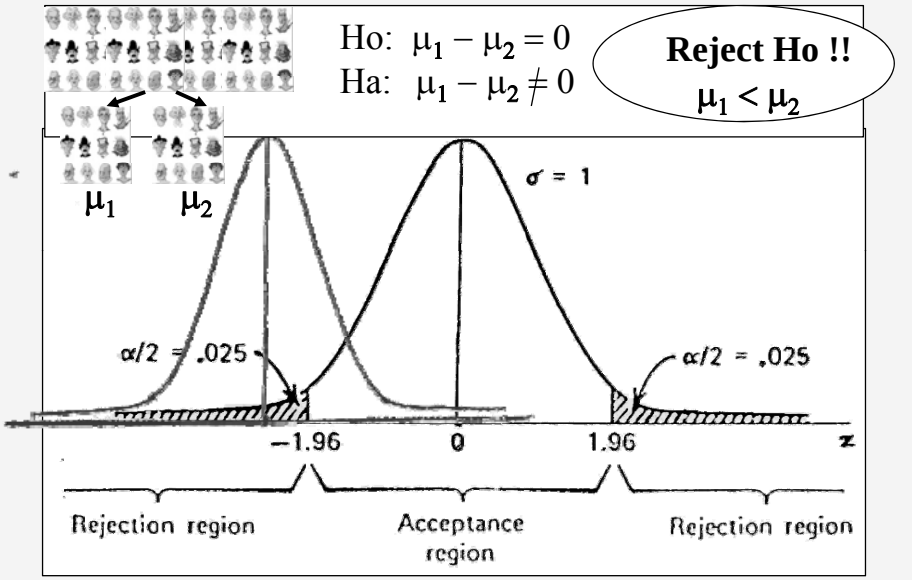


Figure 9.2. Statistical testing.

Hypothesis Testing



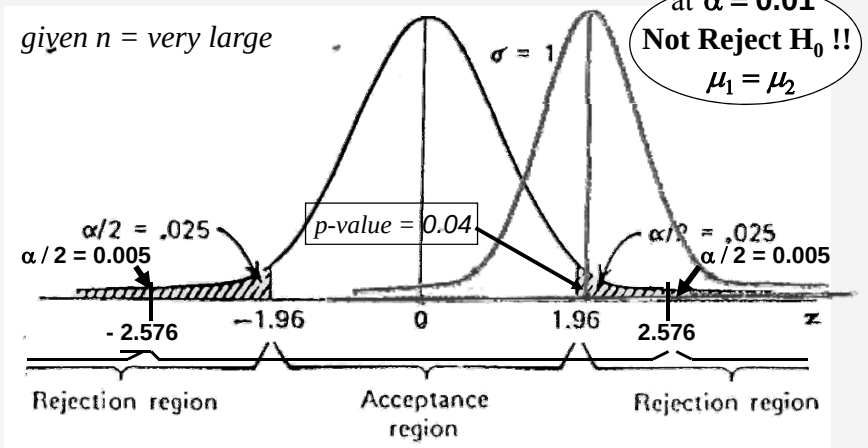
Hypothesis Testing



H_0: $\mu_1 - \mu_2 = 0$
H_a: $\mu_1 - \mu_2 \neq 0$

at α = 0.05
Reject H_0 !!
$\mu_1 > \mu_2$

at α = 0.01
Not Reject H_0 !!
$\mu_1 = \mu_2$



- ***Type I & Type II errors***

H_0: $G_1 = G_2$

Decision	**Reality/Truth: H_0 True** *($G_1=G_2$)*	**Reality/Truth: H_0 False** *($G_1<>G_2$)*
Accept H_0 *(G1=G2)*	**Correct** **Confidence : 1 - α** *0.99, 0.95* **A**	**B** **Type II Error** β *0.10, 0.20*
Reject H_0 *(G1<>G2)*	**C** **Type I Error** α *0.01, 0.05*	**D** **Correct** **Power : 1 - β** *0.90, 0.80*

The End of Inferential Statistics