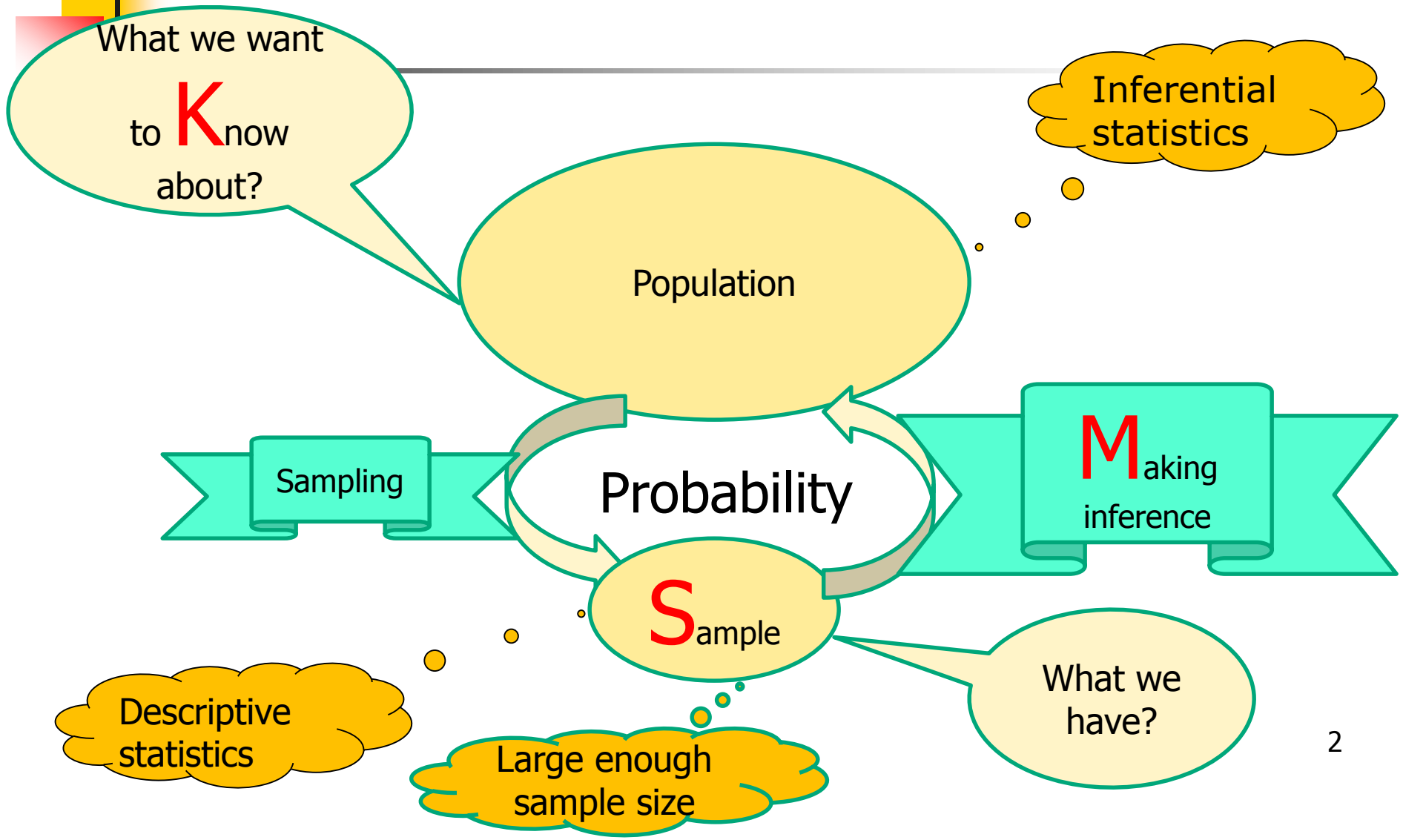
# Descriptive statistics

Dr. Kyaw Swa Mya
AP/Head
Department of Biostatistics
University of Public Health

# RESEARCH OVERVIEW



# Specific learning objectives

- Ability to choose the appropriate summary statistic measures depending on different types of variables
- Ability to choose the appropriate graphical displays depending on different types of variables
- To understand the role of exploratory data analysis

# Descriptive statistics

- To describe the variables scientifically
- Descriptive measures of the variables of the sample

Descriptive measures

- May be frequency (%) for qualitative variables
- May be summary statistic (mean, median, SD, variance, interquartile range) of quantitative variables
- May use diagram/graph

# Age of 189 people

**TABLE 2.2.1 Ordered Array of Ages of Subjects from Table 1.4.1**

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 30 | 34 | 35 | 37 | 37 | 38 | 38 | 38 | 38 | 39 | 39 | 40 | 40 | 42 | 42 |
| 43 | 43 | 43 | 43 | 43 | 43 | 44 | 44 | 44 | 44 | 44 | 44 | 44 | 45 | 45 |
| 45 | 46 | 46 | 46 | 46 | 46 | 46 | 47 | 47 | 47 | 47 | 47 | 47 | 48 | 48 |
| 48 | 48 | 48 | 48 | 48 | 49 | 49 | 49 | 49 | 49 | 49 | 49 | 50 | 50 | 50 |
| 50 | 50 | 50 | 50 | 50 | 51 | 51 | 51 | 51 | 52 | 52 | 52 | 52 | 52 | 52 |
| 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 | 53 |
| 53 | 53 | 54 | 54 | 54 | 54 | 54 | 54 | 54 | 54 | 54 | 54 | 54 | 55 | 55 |
| 55 | 56 | 56 | 56 | 56 | 56 | 56 | 57 | 57 | 57 | 57 | 57 | 57 | 57 | 58 |
| 58 | 59 | 59 | 59 | 59 | 59 | 59 | 60 | 60 | 60 | 60 | 61 | 61 | 61 | 61 |
| 61 | 61 | 61 | 61 | 61 | 61 | 61 | 62 | 62 | 62 | 62 | 62 | 62 | 62 | 63 |
| 63 | 64 | 64 | 64 | 64 | 64 | 64 | 65 | 65 | 66 | 66 | 66 | 66 | 66 | 66 |
| 67 | 68 | 68 | 68 | 69 | 69 | 69 | 70 | 71 | 71 | 71 | 71 | 71 | 71 | 71 |
| 72 | 73 | 75 | 76 | 77 | 78 | 78 | 78 | 82 | | | | | | |

# Sturges Rule

- $k = 1 + 3.322 (\log_{10} n)$, where k stands for the number of class intervals and n is the number of values in the data set under consideration
- $w = R/k$, where w is width of class interval and R is the range of data set
- $k = 1 + 3.322 (2.4393) \approx 9$
- $w = 82\text{-}30/9 = 5.778$

**TABLE 2.3.1 Frequency Distribution of Ages of 189 Subjects Shown in Tables 1.4.1 and 2.2.1**

| Class Interval | Frequency |
|---|---|
| 30–39 | 11 |
| 40–49 | 46 |
| 50–59 | 70 |
| 60–69 | 45 |
| 70–79 | 16 |
| 80–89 | 1 |
| Total | 189 |

# Histogram

**TABLE 2.3.3 The Data of Table 2.3.1 Showing True Class Limits**

| True Class Limits | Frequency |
|---|---|
| 29.5–39.5 | 11 |
| 39.5–49.5 | 46 |
| 49.5–59.5 | 70 |
| 59.5–69.5 | 45 |
| 69.5–79.5 | 16 |
| 79.5–89.5 | 1 |
| Total | 189 |

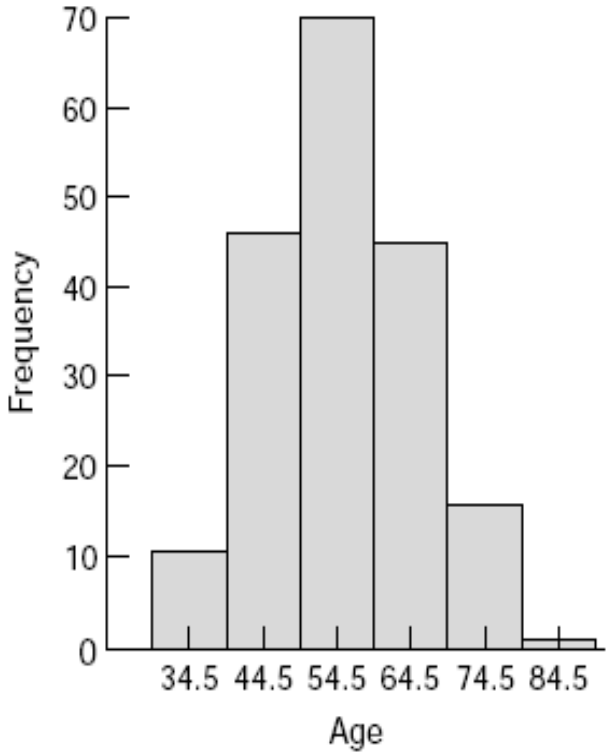



**FIGURE 2.3.2** Histogram of ages of 189 subjects from Table 2.3.1.

# Frequency polygon

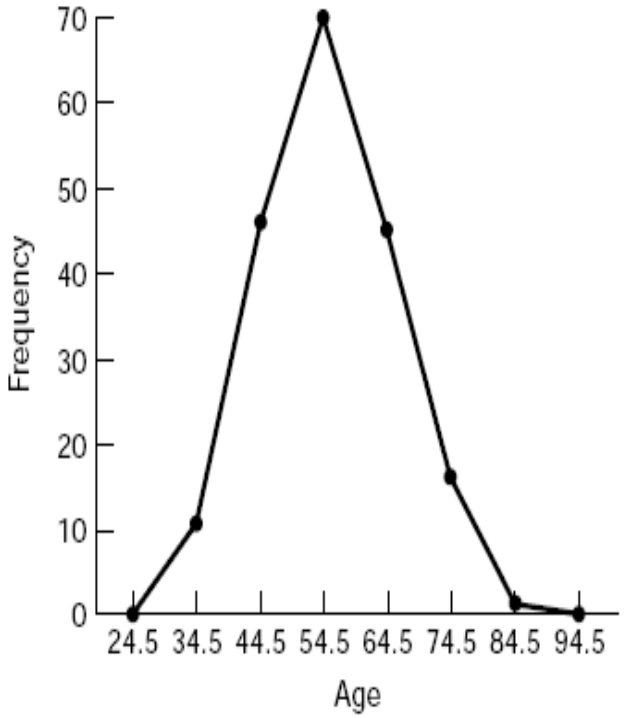



**FIGURE 2.3.4** Frequency polygon for the ages of 189 subjects shown in Table 2.2.1.

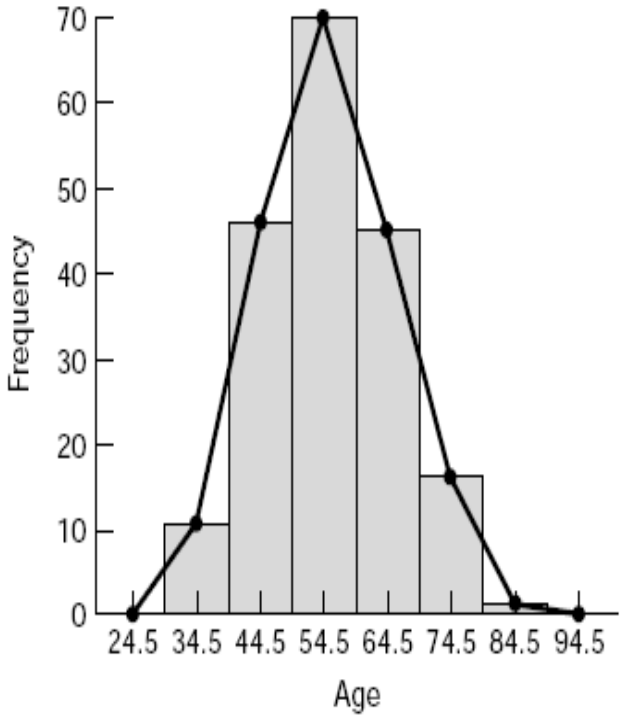



**FIGURE 2.3.5** Histogram and frequency polygon for the ages of 189 subjects shown in Table 2.2.1.

# Statistic vs. Parameter

- A descriptive measure computed from the data of a sample is called a statistic.

- A descriptive measure computed from the data of a population is called a parameter.

# Measures of central tendency

- Mean – average (Unique, simple, influenced by outliers)
- Median - value which divides the set into two equal parts (Unique, simple, not influenced by outliers)
- Mode - value which occurs most frequently

# Measures of dispersion

- Range - the difference between the largest and smallest value in a set of observations
- Variance - dispersion relative to the scatter of the values about their mean but expressed in square unit
- Standard deviation - dispersion relative to the scatter of the values about their mean but expressed in original unit

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

- Coefficient of variation – use to compare the variation of two or more dataset

  CV = SD/mean × 100

- Interquartile range - the difference between the third and first quartiles: that is,
- IQR = Q3 - Q1

$$Q_1 = \frac{n+1}{4} \text{ th ordered observation}$$

$$Q_2 = \frac{2(n+1)}{4} = \frac{n+1}{2} \text{ th ordered observation}$$

$$Q_3 = \frac{3(n+1)}{4} \text{ th ordered observation}$$

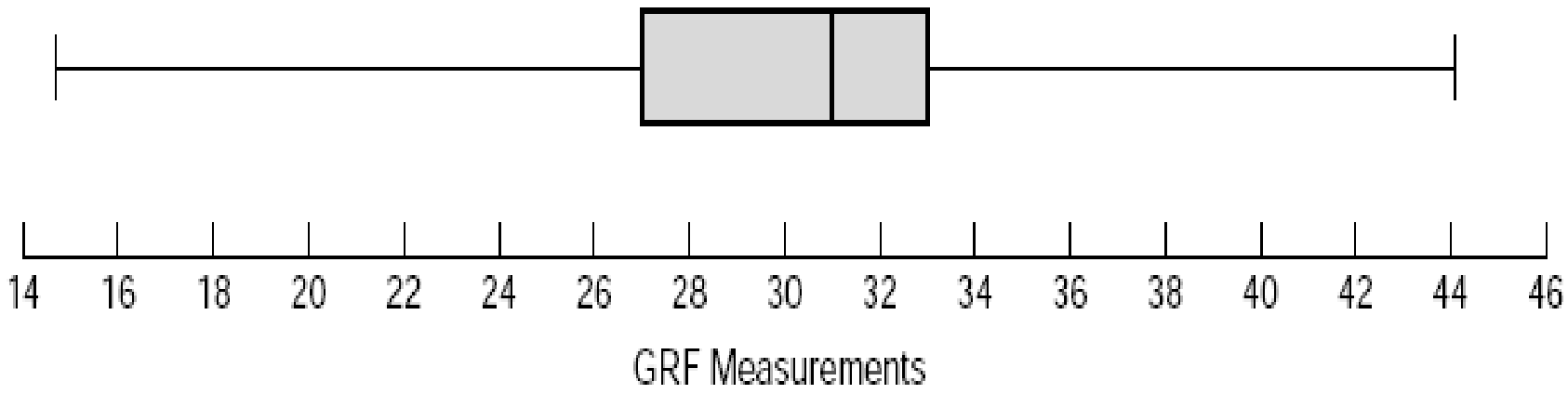



**FIGURE 2.5.5** Box-and-whisker plot for Example 2.5.5.

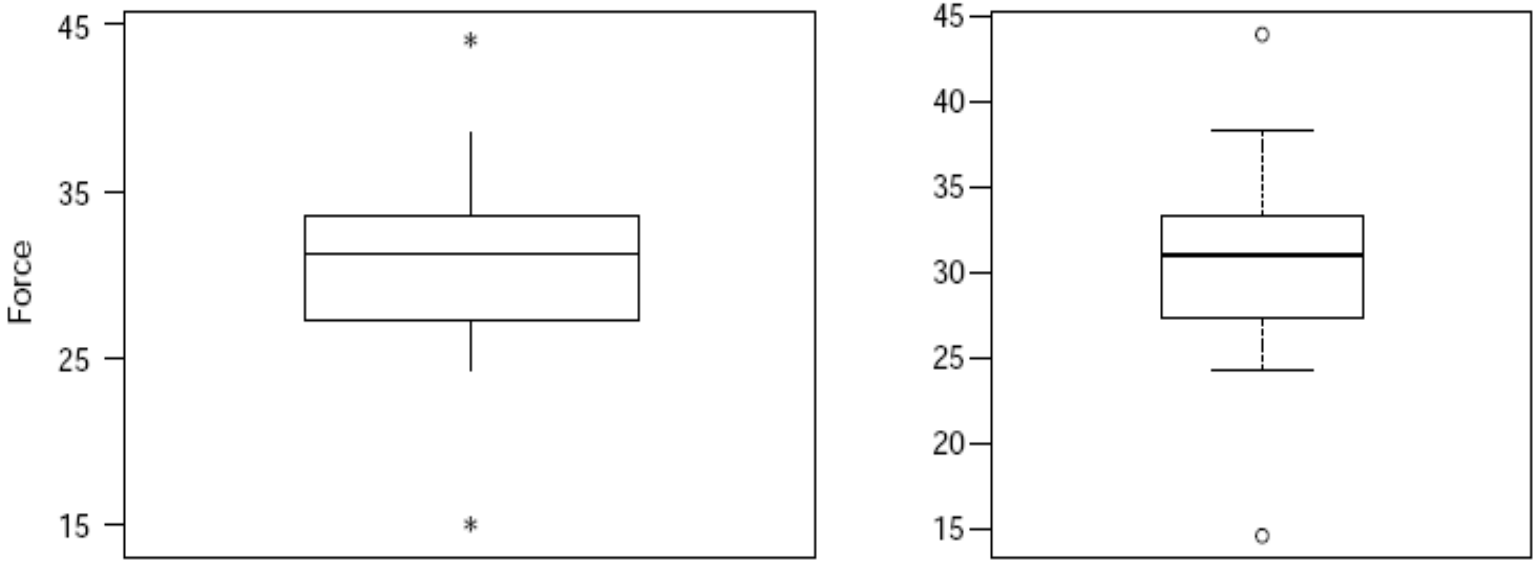
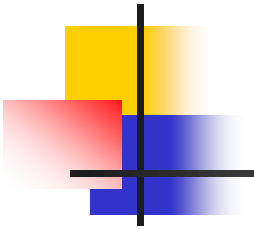



**FIGURE 2.5.6** Box-and-whisker plot constructed by MINITAB (left) and by R (right) from the data of Table 2.5.1.



- Skew - If the graph (histogram or frequency polygon) of a distribution is asymmetric, the distribution is said to be skewed.

# Skew distribution

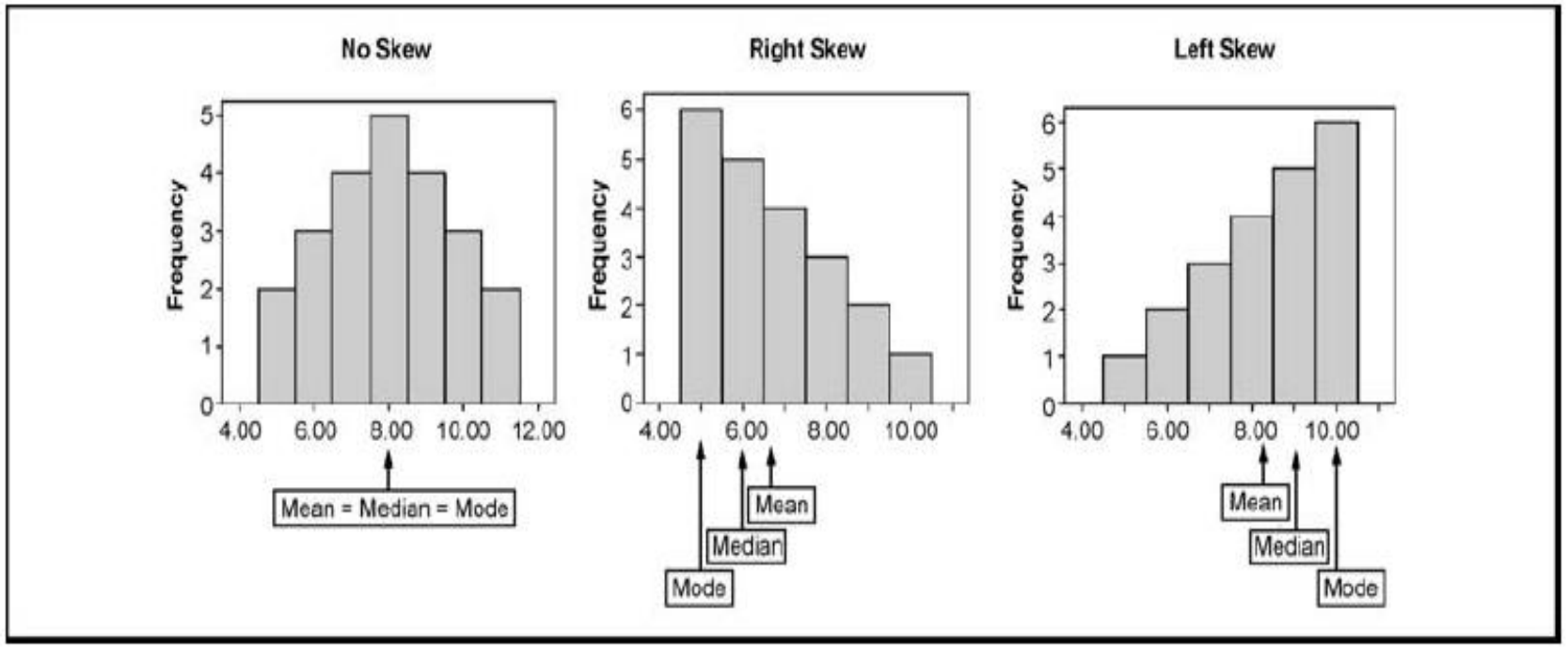
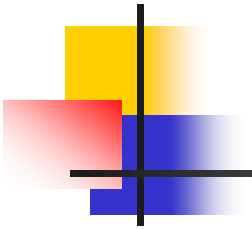



**FIGURE 2.4.1** Three histograms illustrating skewness.



- Kurtosis – a measure of the degree to which a distribution is "peaked" or flat in comparison to a normal distribution whose graph is characterized by a bell-shaped appearance

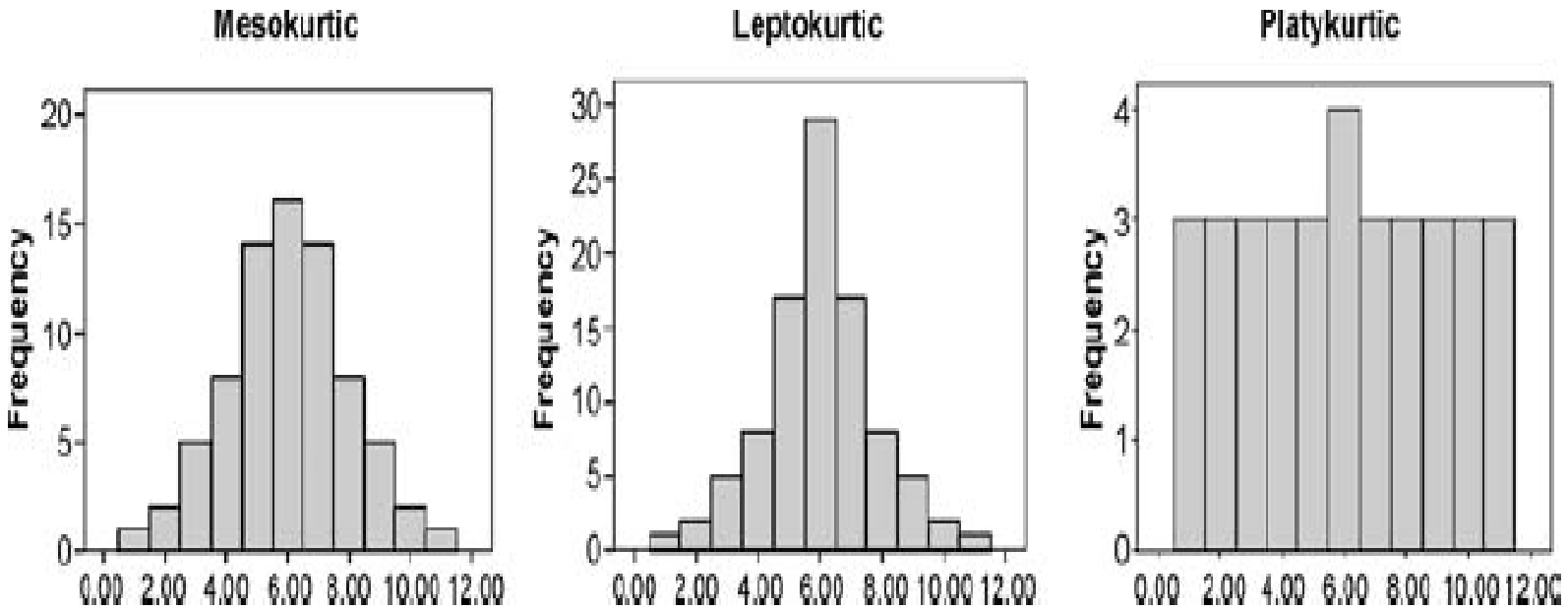



**FIGURE 2.5.4** Three histograms representing kurtosis.

# Exploratory Data Analysis

- Box-and-whisker plots and stem-and-leaf displays are examples of what are known as exploratory data analysis techniques.
- These techniques allow the investigator to examine data in ways that reveal trends and relationships, identify unique features of data sets, and facilitate their description and summarization.

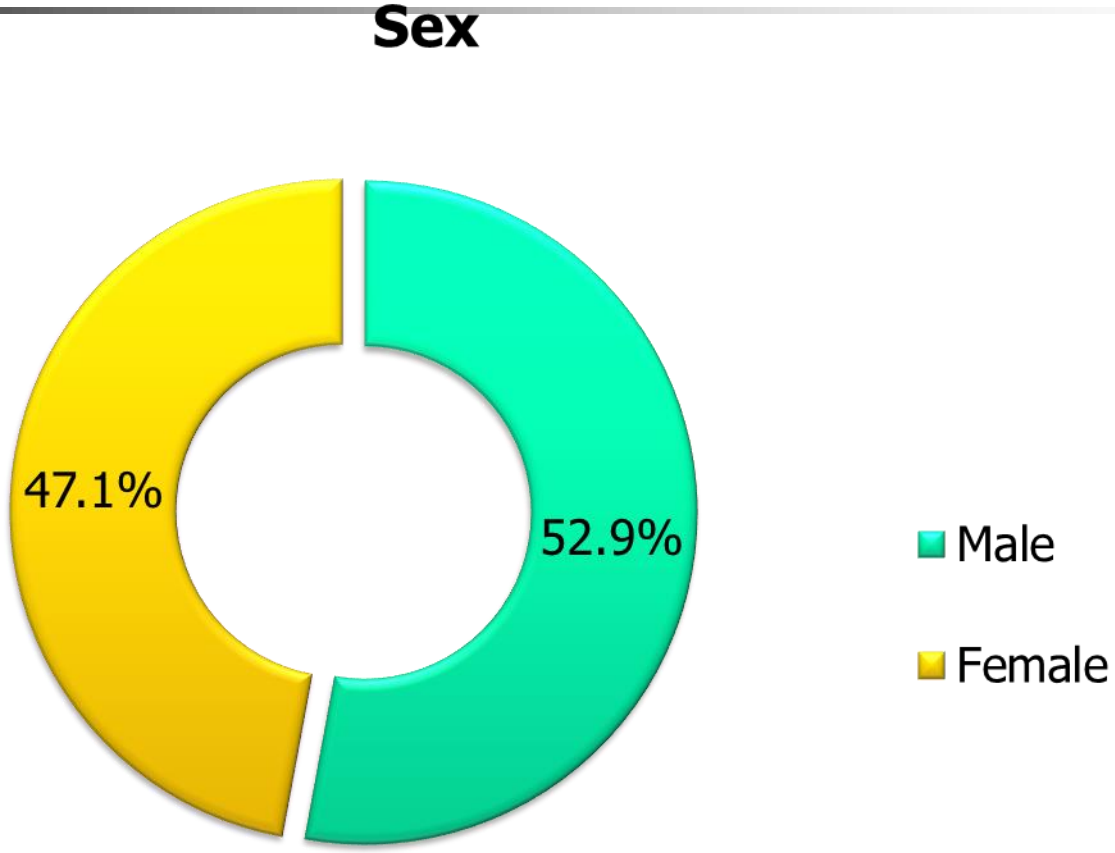
# For qualitative variables

- Among 189, how about sex distribution and how about occupation
- Male 100 vs. Female 89
- Dependent 100, Manual workers 50 and Government staff 39

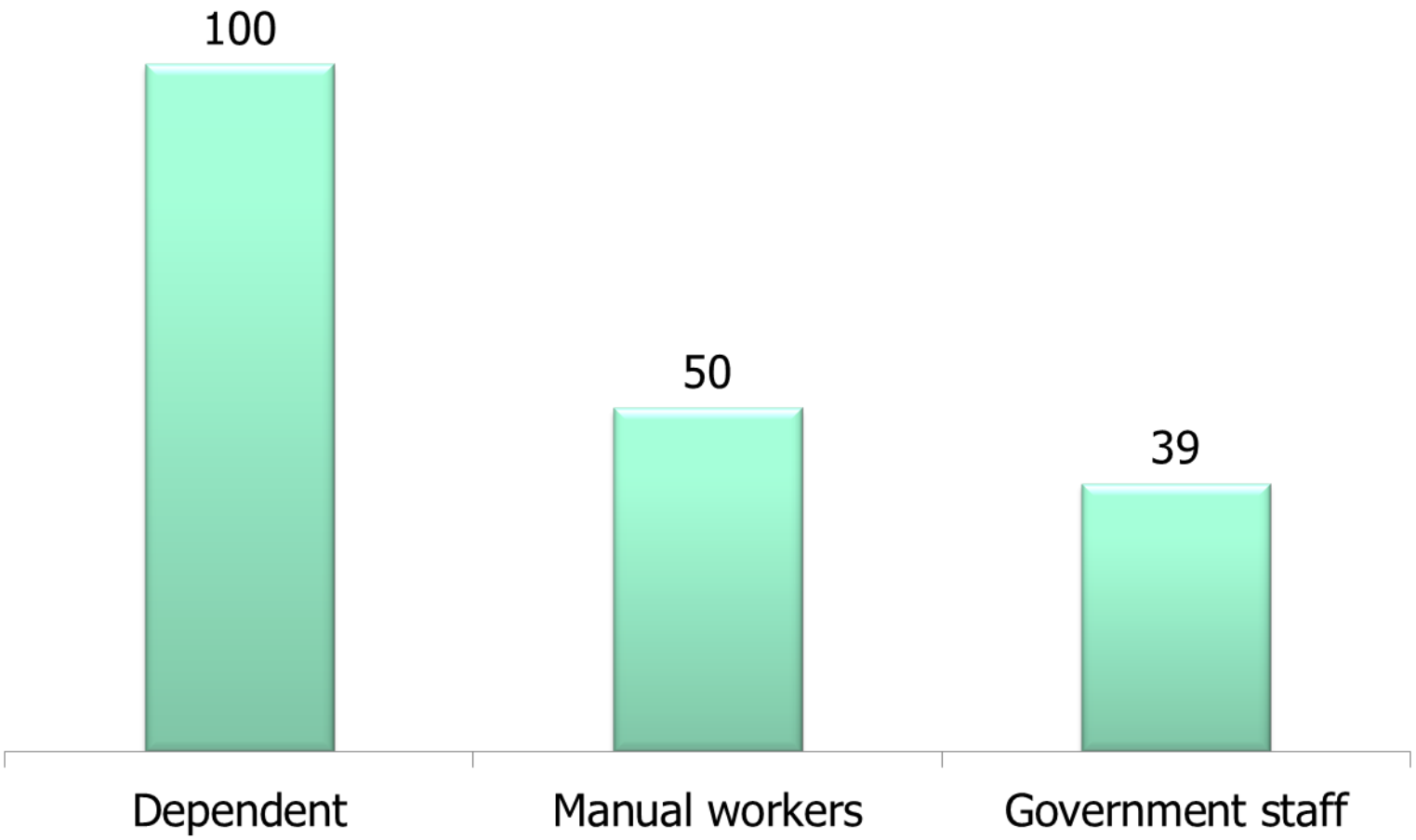
# Frequency distribution table

| Variable | Frequency | Percentage |
|---|---|---|
| Sex | | |
| Male | 100 | 52.9 |
| Female | 89 | 47.1 |
| Occupation | | |
| Dependent | 100 | 52.9 |
| Manual workers | 50 | 26.5 |
| Government staff | 39 | 20.6 |

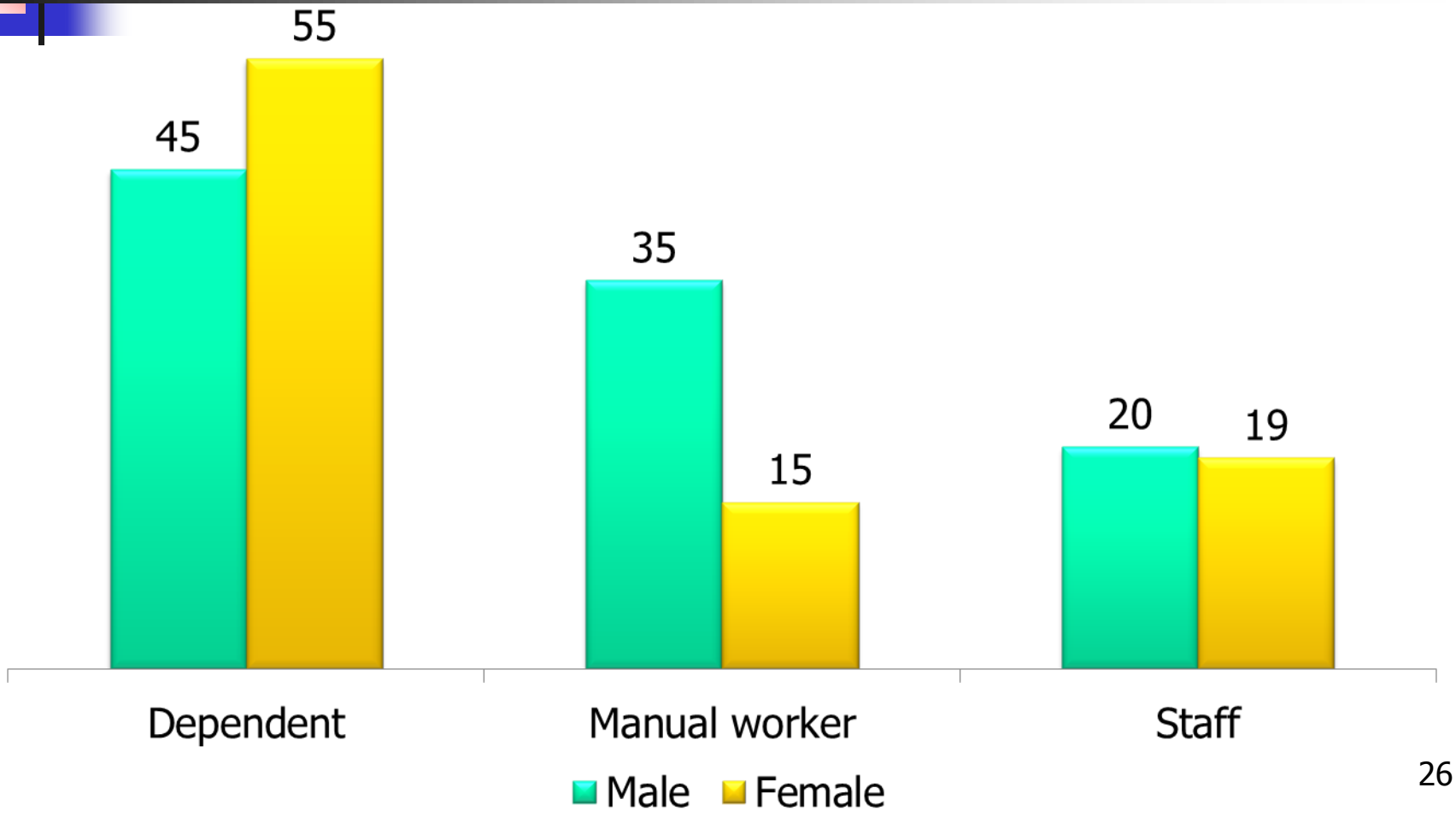
# Figure . Frequency distribution of sex



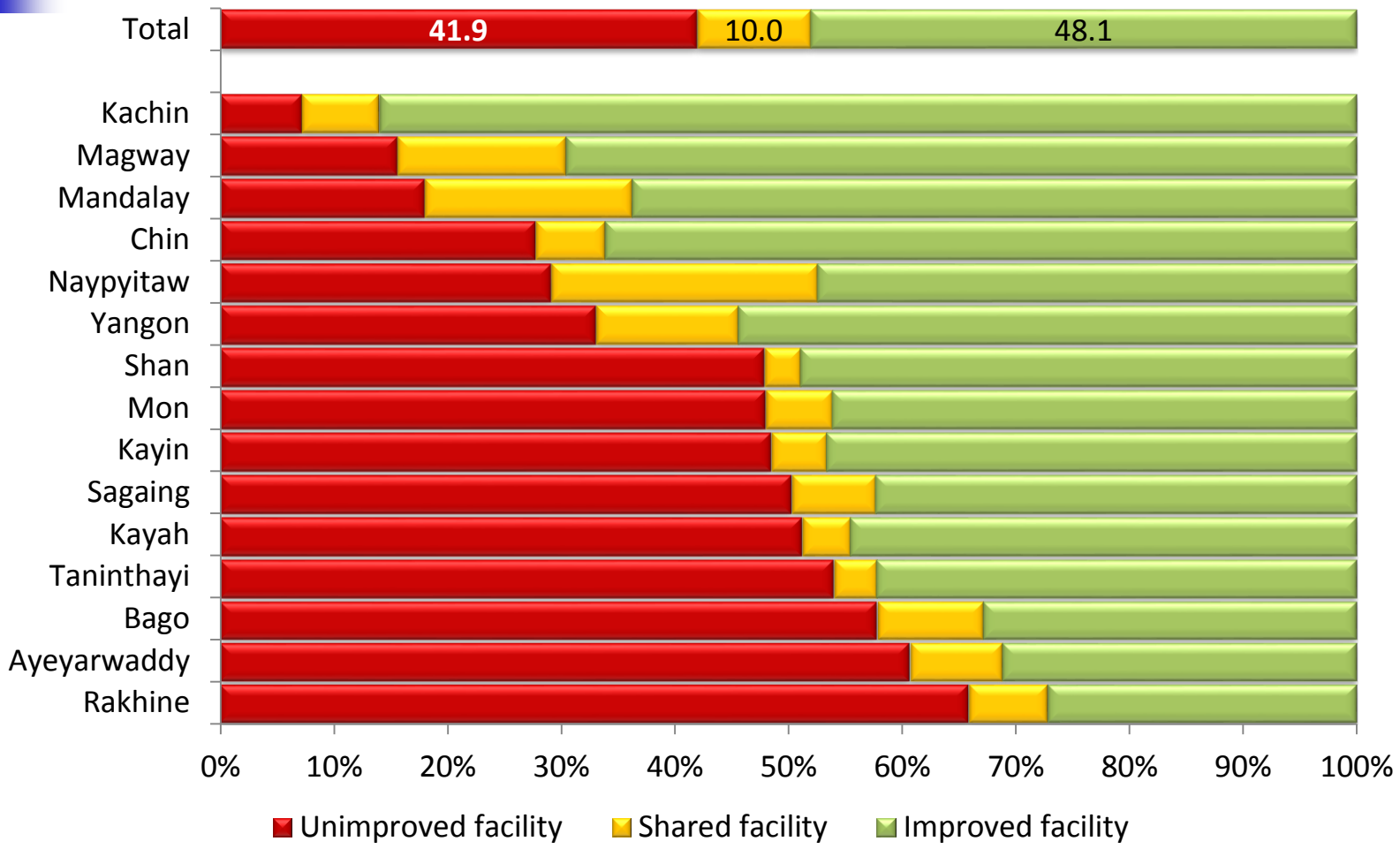
# Figure . Frequency distribution of occupation (Bar Chart)



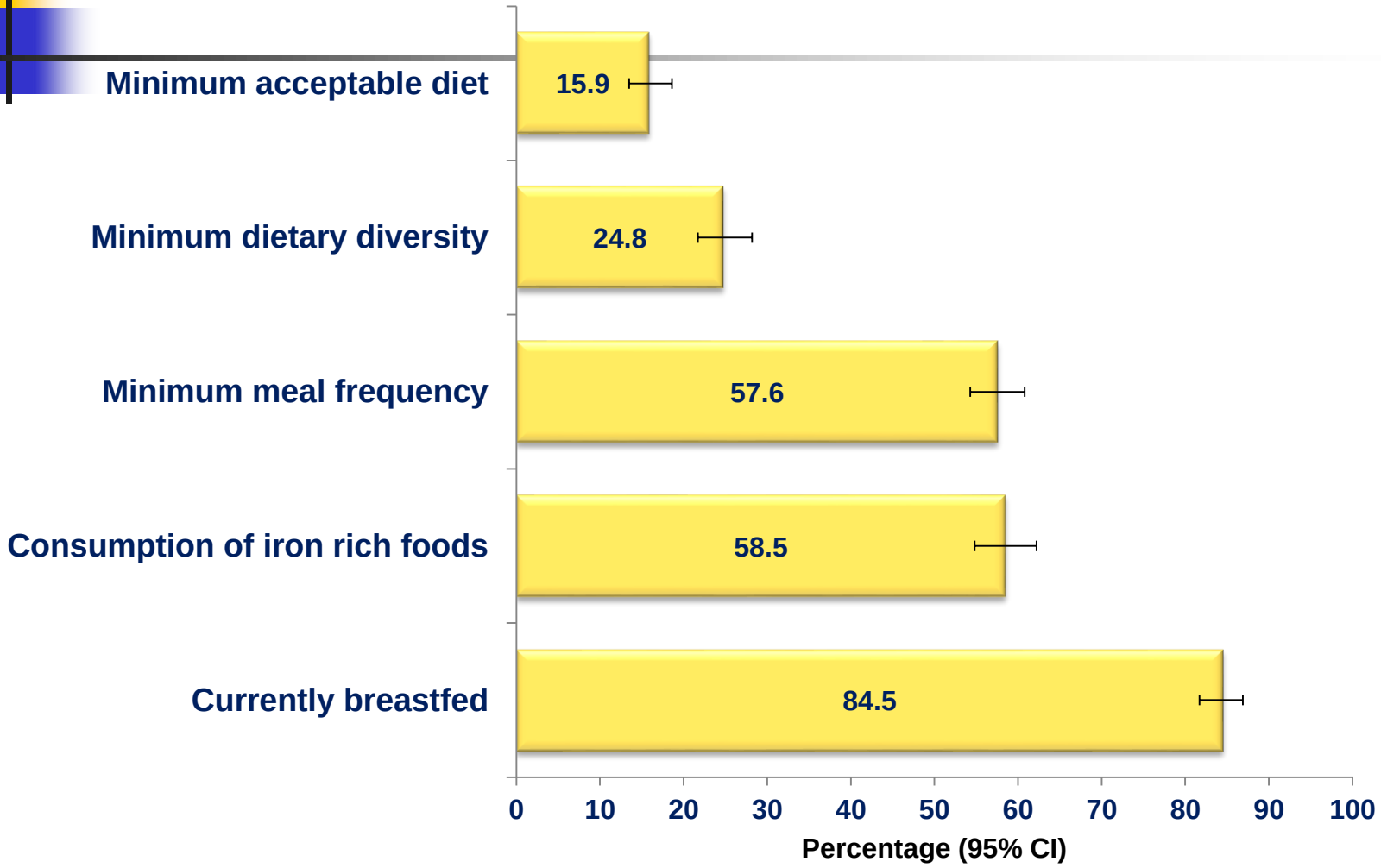
# Figure . Frequency distribution of occupation by sex (Cluster Bar)



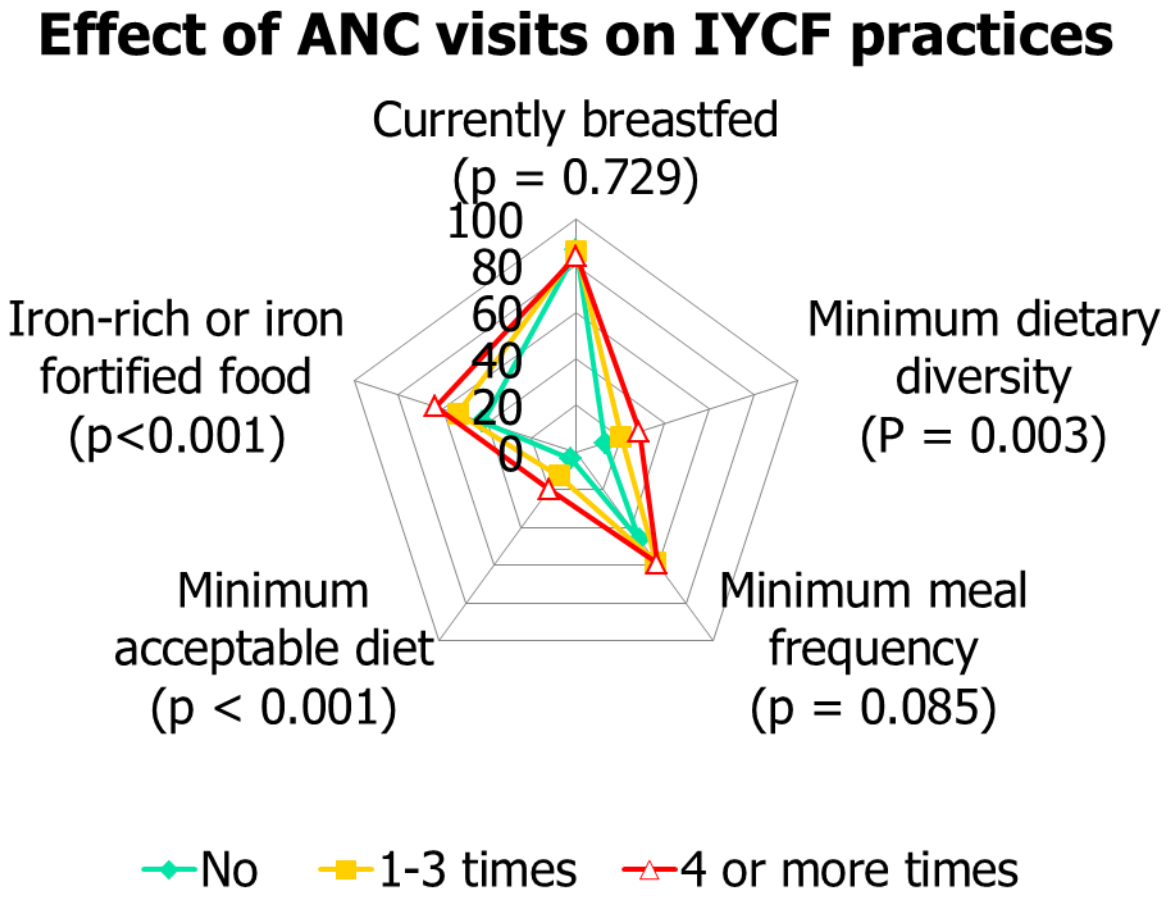
# National and regional distribution of sanitation facilities, Myanmar DHS 2015-16



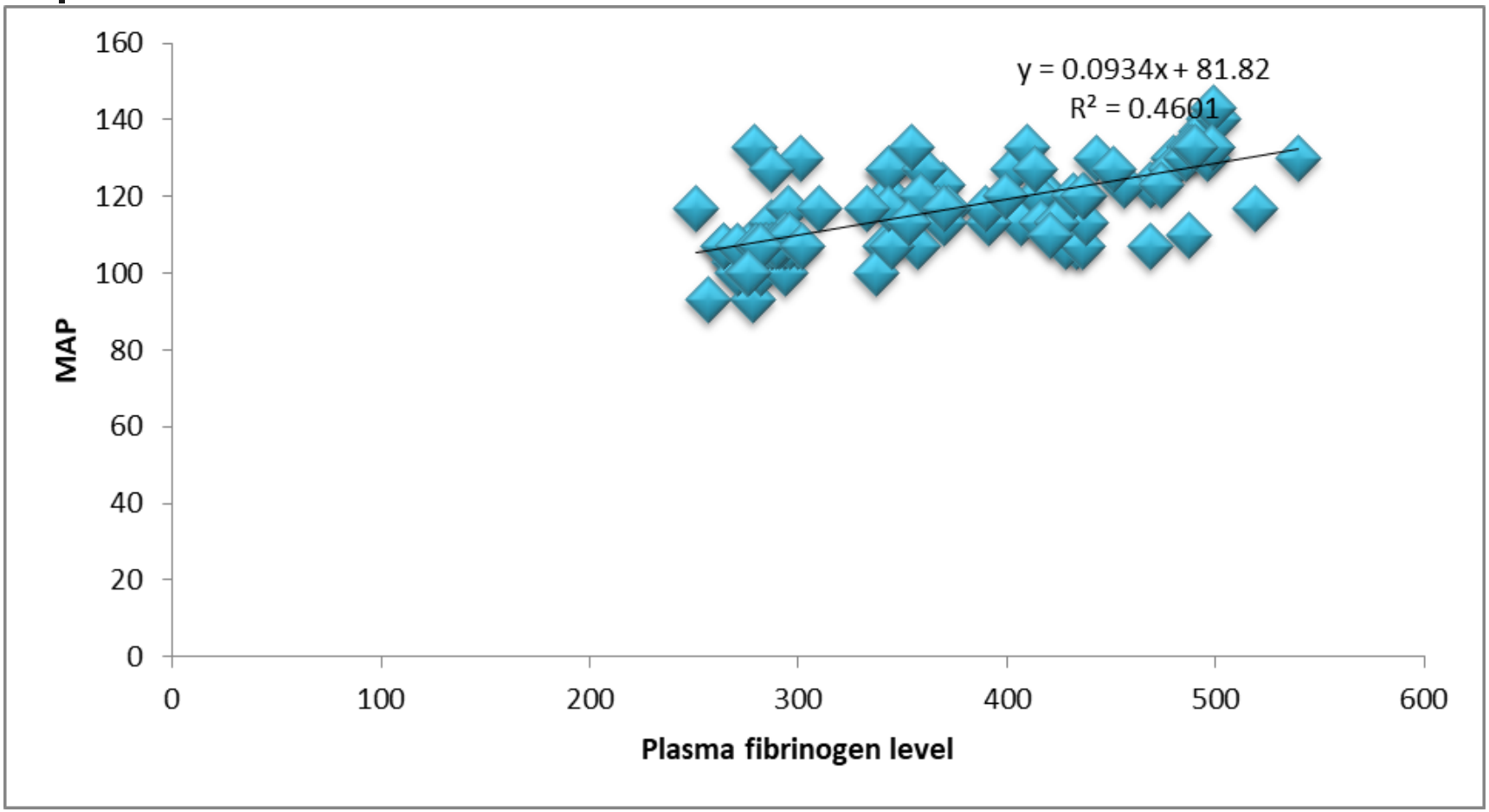
# Prevalence of IYCF practices among children age 6-23 months

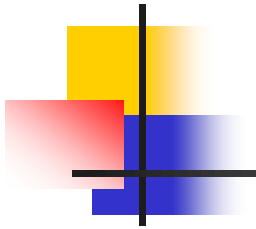


# Radar


Effect of ANC visits on IYCF practices
Currently breastfed
(p = 0.729)
100
80
60
40
20
0
Minimum dietary
diversity
(P = 0.003)
Minimum meal
frequency
(p = 0.085)
Minimum
acceptable diet
(p < 0.001)
Iron-rich or iron
fortified food
(p<0.001)
No
1-3 times
4 or more times

# Scattered diagram





# Thank you!

# Mean Vs. Median

**TABLE 1. Baseline Characteristics of Participants According to Soft Drink Consumption (n=8997)**

| | No. of Soft Drinks Consumed Per Day | | | |
|---|---|---|---|---|
| Characteristic | <1 (n=5840) | 1 (n=1918) | ≥2 (n=1239) | *P** |
| Age, y | 56±10 | 53±10 | 51±9 | ... |
| Men, % | 42.8 | 50.2 | 53.4 | ... |
| Systolic BP, mm Hg | 127±19 | 125±17 | 126±18 | <0.0001 |
| Diastolic BP, mm Hg | 76±10 | 77±10 | 78±11 | <0.0001 |
| BP ≥130/85 mm Hg or on treatment, % | 48.9 | 46.7 | 48.4 | <0.0001 |
| Hypertension, % | 22.5 | 18.7 | 21.6 | 0.0014 |
| Treatment for hypertension, % | 18.9 | 16.1 | 17.6 | 0.0011 |
| BMI, kg/m$^2$ | 26.8±4.8 | 27.8±5.1 | 28.5±5.4 | <0.0001 |
| BMI ≥30 kg/m$^2$, % | 20.9 | 27.1 | 32.1 | <0.0001 |
| Weight, kg | 75.5±16.1 | 79.4±16.9 | 82.1±18.1 | <0.0001 |
| Waist circumference, in | 36.0±5.6 | 36.9±5.7 | 37.8±6.1 | <0.0001 |
| Increased waist circumference, %† | 33.9 | 37.2 | 41.1 | <0.0001 |
| Men | 36.3 | 40.9 | 48.1 | <0.0001¶ |
| Women | 32.0 | 33.4 | 33.2 | <0.0001¶ |
| Total cholesterol, mg/dL | 206±37 | 204±37 | 202±38 | 0.72 |
| Low-density liporotein cholesterol, mg/dL | 129±34 | 128±33 | 127±34 | 0.30 |
| Triglycerides, mg/dL | 127±83 | 141±119 | 148±118 | <0.0001 |
| High triglycerides, %‡ | 28.3 | 32.7 | 35.9 | <0.0001 |
| HDL-C, mg/dL | 52±16 | 50±15 | 47±14 | <0.0001 |

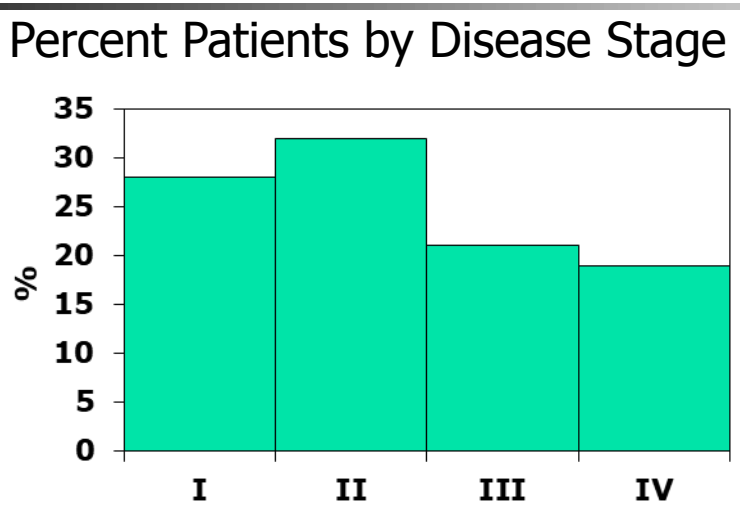
We want to study whether individuals over 45 years are at greater risk of diabetes than those younger than 45. What kind of variable is age?

1. Dichotomous
2. Ordinal
3. Categorical
4. Continuous

# We are interested in assessing disparities in infant morbidity by race/ethnicity. What kind of variable is race/ethnicity?

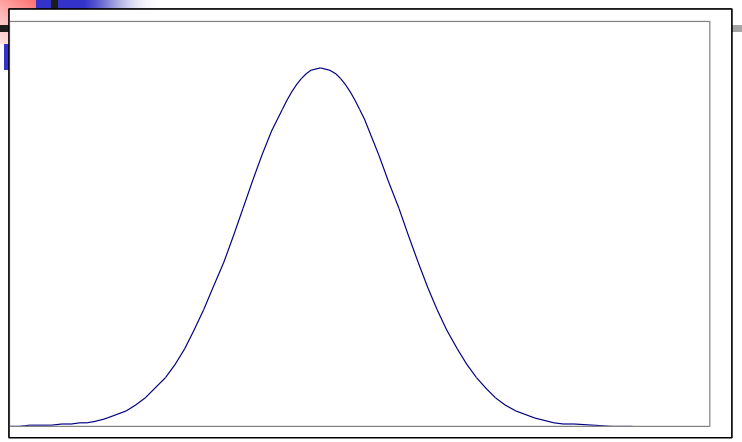
1. Dichotomous
2. Ordinal
3. ✓ Categorical
4. Continuous

# What type of display is shown below?



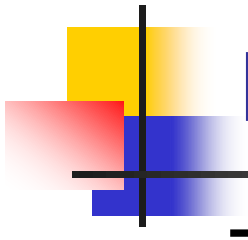

1. Frequency bar chart
2. Relative frequency bar chart
3. Frequency histogram
4. Relative frequency histogram

# The distribution of SBP in men, 20-29 years is shown below. What is the best summary of a typical value



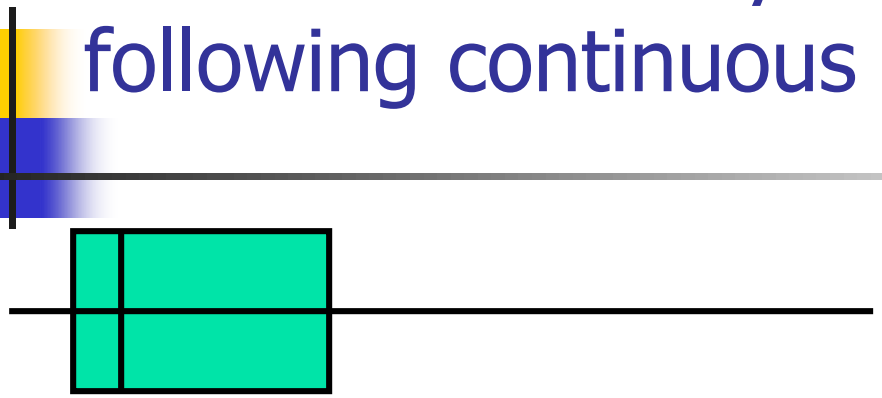
1. Mean ✓
2. Median
3. Interquartile range
4. Standard Deviation

# When data are skewed, the mean is higher than the median.



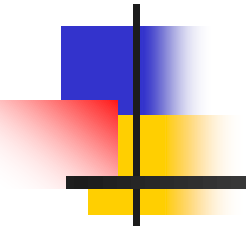
1. True
2. False ✓

# The best summary of variability for the following continuous variable is



1. Mean
2. Median
3. ✓ Interquartile range
4. Standard Deviation

# Probability distribution and sampling distribution



# What is the probability of selecting a male with optimal blood pressure?

| | Blood Pressure Category | | | | |
|---|---|---|---|---|---|
| | Optimal | Normal | Pre-Htn | Htn | Total |
| Male | 20 | 15 | 15 | 30 | 80 |
| Female | 5 | 15 | 25 | 25 | 70 |
| Total | 25 | 30 | 40 | 55 | 150 |

1. 20/25
2. 20/80
3. ✓ 20/150

# What is the probability of selecting a patient with Pre-Htn or Htn?

| | Blood Pressure Category | | | | |
|---|---|---|---|---|---|
| | Optimal | Normal | Pre-Htn | Htn | Total |
| Male | 20 | 15 | 15 | 30 | 80 |
| Female | 5 | 15 | 25 | 25 | 70 |
| Total | 25 | 30 | 40 | 55 | 150 |

1. 95/150
2. 45/80
3. 55/150

# What proportion of men have prevalent CVD?

| | CVD | Free of CVD |
|---|---|---|
| Men | 35 | 265 |
| Women | 45 | 355 |

1. 35/80
2. 35/265
3. ✓ 35/300

# What proportion of patients with CVD are men ?

| | CVD | Free of CVD |
|---|---|---|
| Men | 35 | 265 |
| Women | 45 | 355 |

1. 35/700
2. 35/80
3. 80/300

# Are Family History and Current Status Independent?

**Example.** Consider the following table which cross classifies subjects by their family history of CVD and current (prevalent) CVD status.

| | **Current CVD** | |
|---|---|---|
| **Family History** | No | Yes |
| No | 215 | 25 |
| Yes | 90 | 15 |

P(Current CVD| Family Hx) = 15/105 = 0.143

P(Current CVD| No Family Hx) = 25/240 = 0.104

# Are symptoms independent of disease?

| | Disease | No Disease | Total |
|---|---|---|---|
| Symptoms | 25 | 225 | 250 |
| No Symptoms | 50 | 450 | 500 |

1. No
2. Yes

# Probability Models – Binomial Distribution

- Two possible outcomes: success and failure
- Replications of process are independent
- P(success) is constant for each replication

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

- Mean=np, variance=np(1-p)

# Probability Models – Poisson Distribution

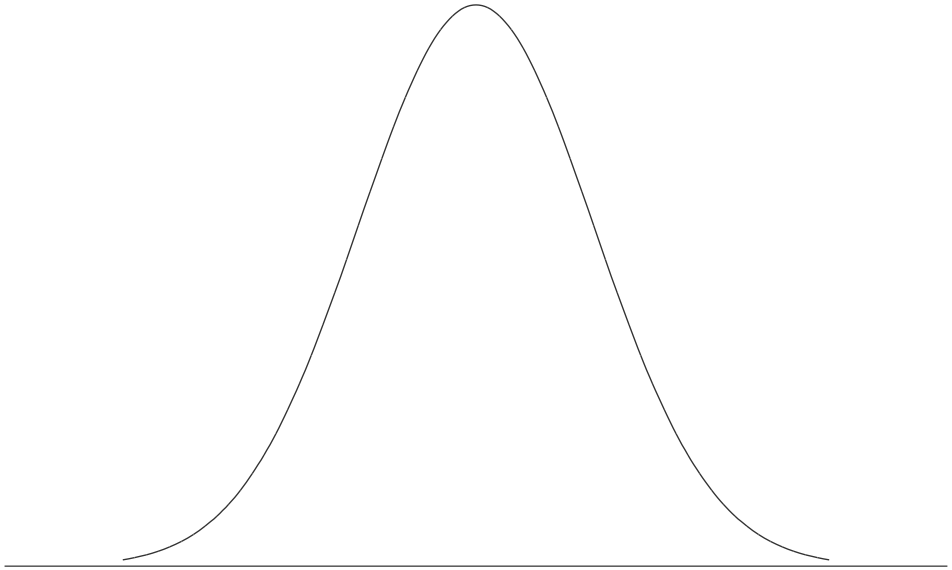
- Two possible outcomes: success and failure
- Replications of process are independent
- Often used to model counts (often used to model rare events)

$$P(x) = (e^{-\mu}\ \mu^{x}) / x!$$

- Mean=$\mu$, variance=$\mu$

# Probability Models – Normal Distribution

- Model for continuous outcome
- Mean=median=mode



# Normal Distribution

Properties of Normal Distribution

I) The normal distribution is symmetric about the mean (i.e., $P(X > \mu) = P(X < \mu) = 0.5$).

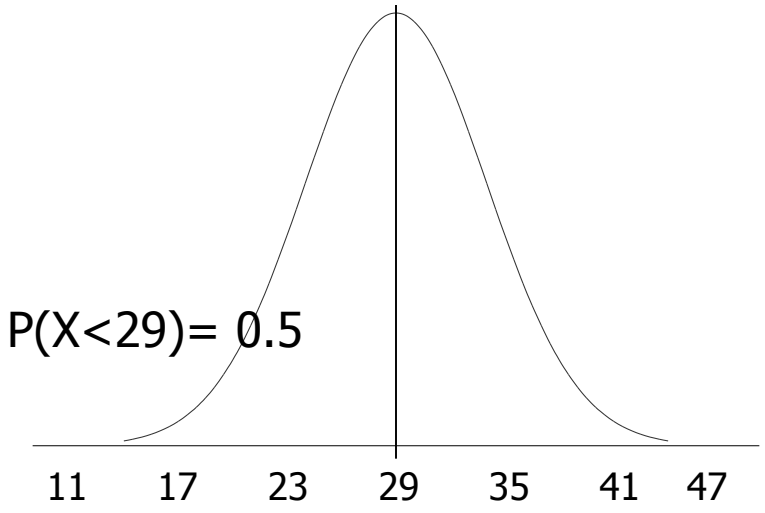
ii) The mean and variance ($\mu$ and $\sigma^2$) completely characterize the normal distribution.

iii) The mean = the median = the mode

iv) Approximately 68% of obs between mean $\pm$ 1 sd 95% between mean $\pm$ 2 sd, and >99% between mean $\pm$ 3 sd

# Normal Distribution

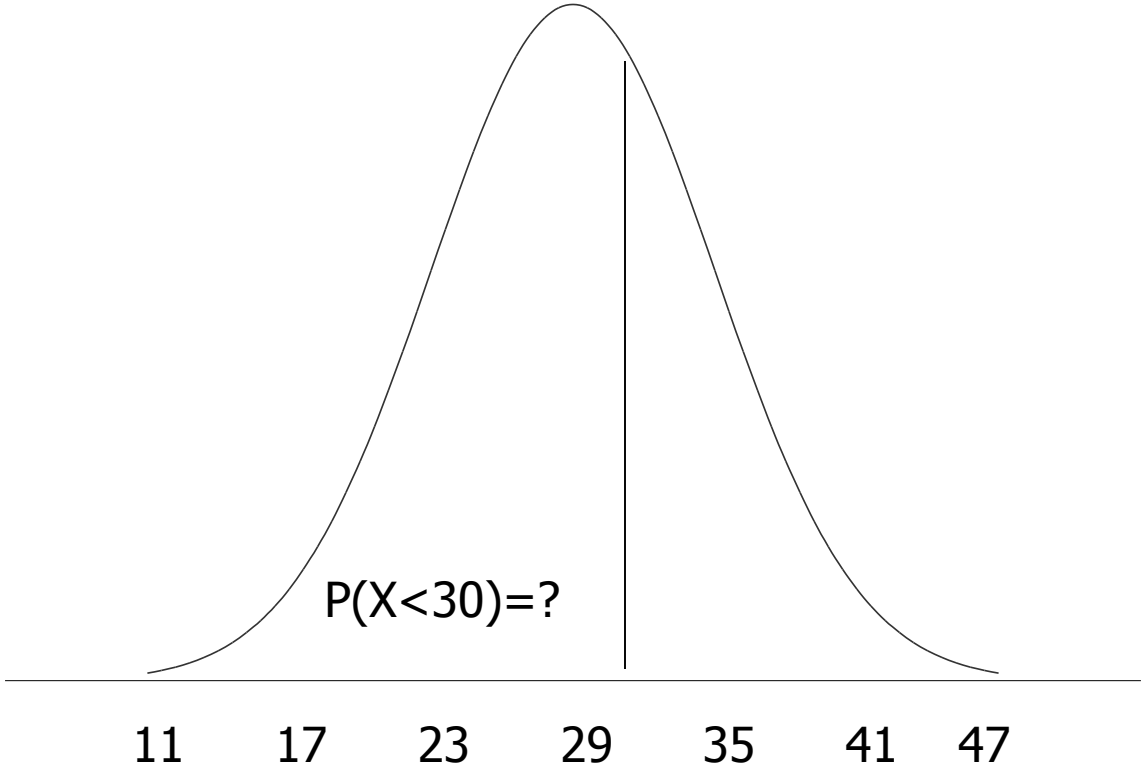
Body mass index (BMI) for men age 60 is normally distributed with a mean of 29 and standard deviation of 6.




What is the probability that a male has BMI < 29?

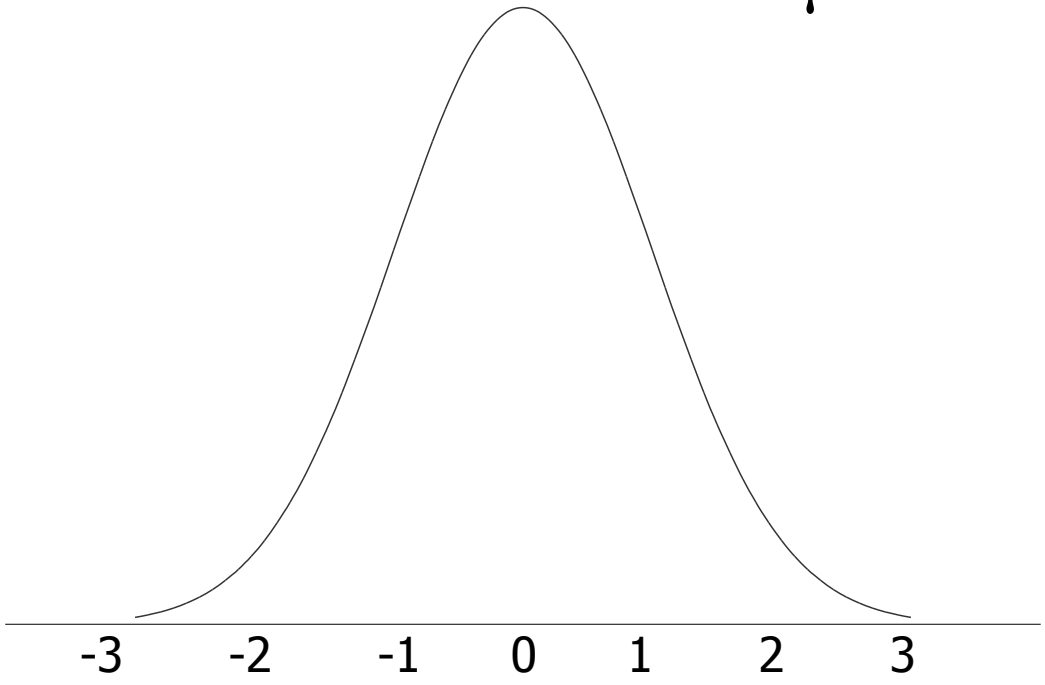
# Normal Distribution

What is the probability that a male has BMI less than 30?



# Standard Normal Distribution Z

Normal distribution with $\mu=0$ and $\sigma=1$



# Normal Distribution

$$Z = \frac{x - \mu}{\sigma} = \frac{30 - 29}{6} = 0.17$$

P(X<30)= P(Z<0.17) = 0.5675

From a table of standard normal probabilities or statistical computing package.

# Comparing Systolic Blood Pressure (SBP)

Comparing systolic blood pressure (SBP)

- Suppose for Males Age 50, SBP is approximately normally distributed with a mean of 108 and a standard deviation of 14
- Suppose for Females Age 50, SBP is approximately normally distributed with a mean of 100 and a standard deviation of 8

If a Male Age 50 has a SBP = 140 and a Female Age 50 has a SBP = 120, who has the "relatively" higher SBP ?

# Normal Distribution

$Z_M = (140 - 108) / 14 = 2.29$

$Z_F = (120 - 100) / 8 = 2.50$

Which is more extreme?

# Percentiles of the Normal Distribution

- The $k^{th}$ *percentile* is defined as the score that holds k percent of the scores below it.
- Eg., $90^{th}$ percentile is the score that holds 90% of the scores below it.
- Q1 = $25^{th}$ percentile, median = $50^{th}$ percentile, Q3 = $75^{th}$ percentile

# Percentiles

For the normal distribution, the following is used to compute percentiles:

$$X = \mu + Z\,\sigma$$

where

$\mu$ = mean of the random variable X,

$\sigma$ = standard deviation, and

Z = value from the standard normal distribution for the desired percentile (e.g., 95$^{th}$, Z=1.645).

95$^{th}$ percentile of BMI for Men: 29+1.645(6) = 38.9

# Central Limit Theorem

- (Non-normal) population with $\mu$, $\sigma$
- Take samples of size n – as long as n is sufficiently large (usually n $\geq$ 30 suffices)
- The distribution of the sample mean is approximately normal, therefore can use Z to compute probabilities

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$ (Standard error)

# Sampling Error

- We want to estimate population parameter, so we collect enough sample size randomly.
- Although we randomly selected enough sample size, the sample statistic may not identical with population parameter.
- This is due to sampling error.
- Sampling error is not a bias but it is random error.

# Sampling Distribution

- The distribution of all possible values that can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population, is called the sampling distribution of that statistic.

# Construction

1. From a finite population of size N, randomly draw all possible samples of size n.
2. Compute the statistic of interest for each sample.
3. Draw a distribution curve for all computed sample statistic.

# Characteristics

1. The distribution of sample means will be normal.

2. The mean of the distribution of sample mean will be equal to the mean of the population from which the samples were drawn.

3. The variance of the distribution of sample mean will be equal to the variance of the population divided by the sample size.

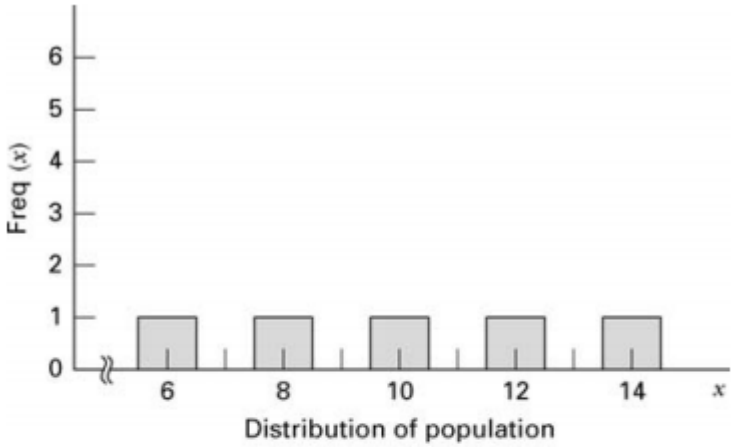
# Population (N=5) = 6, 8, 10, 12, 14
# Mean = 10, variance = 8

**TABLE 5.3.1 All Possible Samples of Size $n = 2$ from a Population of Size $N = 5$. Samples Above or Below the Principal Diagonal Result When Sampling Is Without Replacement. Sample Means Are in Parentheses**

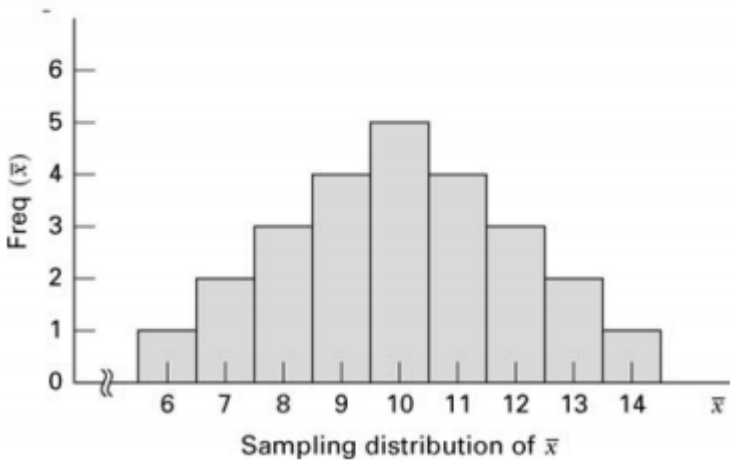
| | | Second Draw | | | | |
|---|---|---|---|---|---|---|
| | | **6** | **8** | **10** | **12** | **14** |
| | **6** | 6, 6 | 6, 8 | 6, 10 | 6, 12 | 6, 14 |
| | | (6) | (7) | (8) | (9) | (10) |
| | **8** | 8, 6 | 8, 8 | 8, 10 | 8, 12 | 8, 14 |
| | | (7) | (8) | (9) | (10) | (11) |
| **First Draw** | **10** | 10, 6 | 10, 8 | 10, 10 | 10, 12 | 10, 14 |
| | | (8) | (9) | (10) | (11) | (12) |
| | **12** | 12, 6 | 12, 8 | 12, 10 | 12, 12 | 12, 14 |
| | | (9) | (10) | (11) | (12) | (13) |
| | **14** | 14, 6 | 14, 8 | 14, 10 | 14, 12 | 14, 14 |
| | | (10) | (11) | (12) | (13) | (14) |

**TABLE 5.3.2 Sampling Distribution of $\bar{x}$ Computed from Samples in Table 5.3.1**

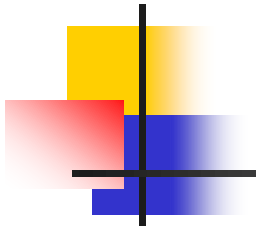
| $\bar{x}$ | Frequency | Relative Frequency |
|---|---|---|
| 6 | 1 | 1/25 |
| 7 | 2 | 2/25 |
| 8 | 3 | 3/25 |
| 9 | 4 | 4/25 |
| 10 | 5 | 5/25 |
| 11 | 4 | 4/25 |
| 12 | 3 | 3/25 |
| 13 | 2 | 2/25 |
| 14 | 1 | 1/25 |
| Total | 25 | 25/25 |



Distribution of population



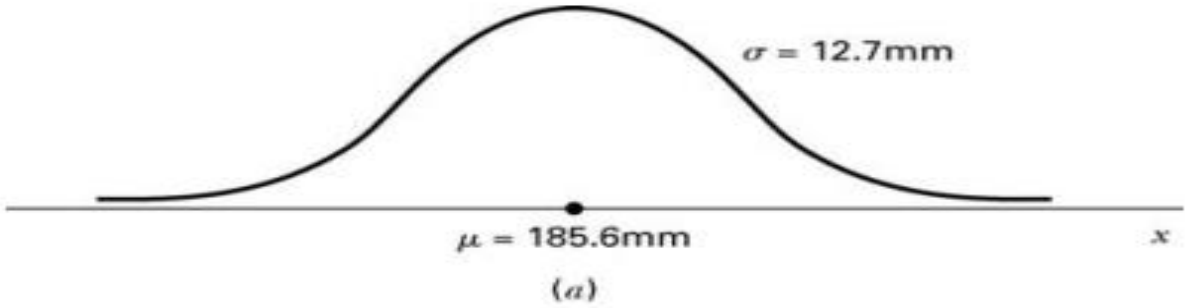
Sampling distribution of $\bar{x}$



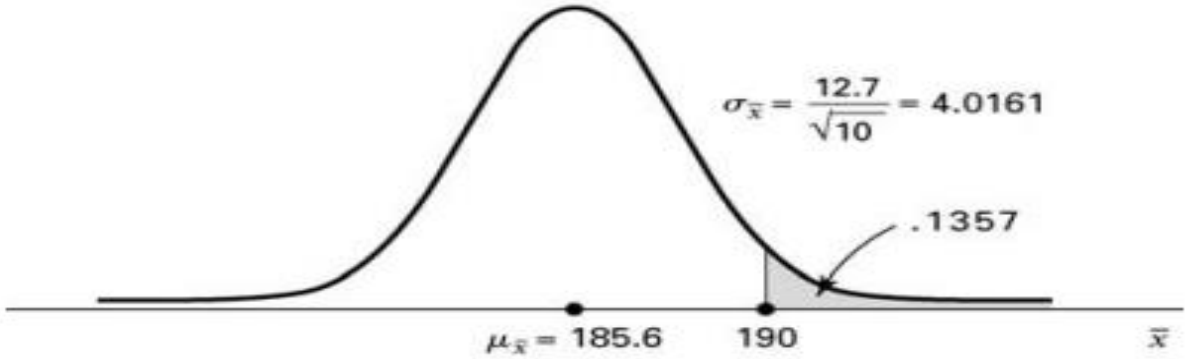
- Mean of sampling distribution = 10
  (Equal with population mean)
- Variance of sampling distribution = 4
  (not equal with population variance)
- Variance of sample mean = Variance (pop)/n
  = 8/2
  = 4

# Standard Error of Mean (SEM)

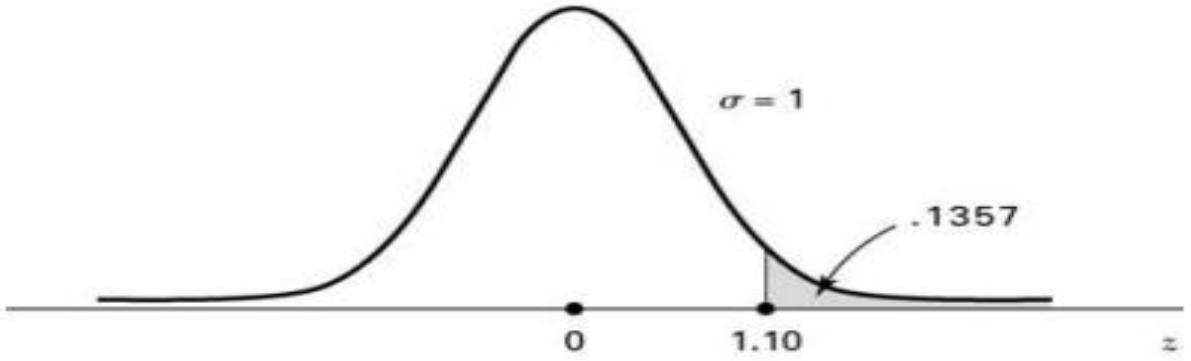
- $\sigma_{(\text{sample mean})} = \sigma / \sqrt{n}$
- Standard deviation of sample means
- Measure the variation of sample means
- We have to use standard error (SE) whenever we want to make inference about population based on sample finding to compensate the sampling error


σ = 12.7mm
μ = 185.6mm
x

(a)


σx̄ = 12.7/√10 = 4.0161
.1357
μx̄ = 185.6
190
x̄

(b)


σ = 1
.1357
0
1.10
z

(c)

# Inferential statistic

- Estimation
- Hypothesis testing

# Estimation

- Point estimation
- Interval estimation

Estimator ± Reliability coefficient × SE

# Hypothesis testing

- Hypothesis – Simple statement about population
- Research hypothesis and Statistical hypothesis

Statistical hypothesis

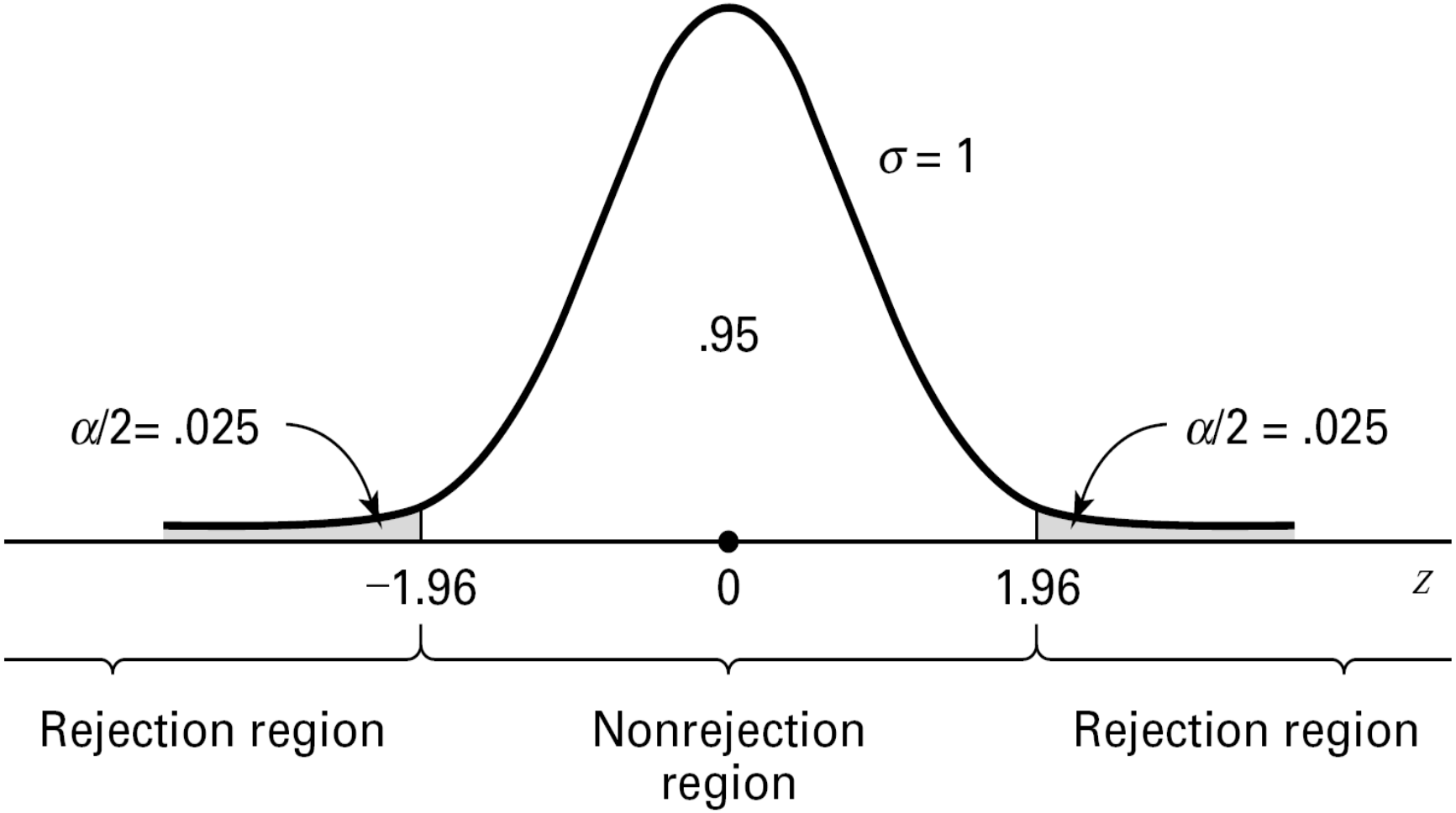
Null – No difference/no association (=, ≥, ≤)

Alternative – Difference/association (≠, >, <)

# Level of significance (α)

- The level of significance a is a probability and, in fact, is the probability of rejecting a true null hypothesis.

# Statistical decision



# Types of error

| Possible Action | Condition of Null Hypothesis: True | Condition of Null Hypothesis: False |
|---|---|---|
| Fail to reject $H_0$ | Correct action | Type II error |
| Reject $H_0$ | Type I error | Correct action |

# P value

- The p value is the smallest value of alpha (α) for which we can reject a null hypothesis.