Basic Statistics

By Hui Bian

Office for Faculty Excellence

Introduction

- Statistics: "a bunch of mathematics used to summarize, analyze, and interpret a group of numbers or observations."

  *It is a tool.

  *Cannot replace your research design, your research questions, and theory or model you want to use.

Population and sample

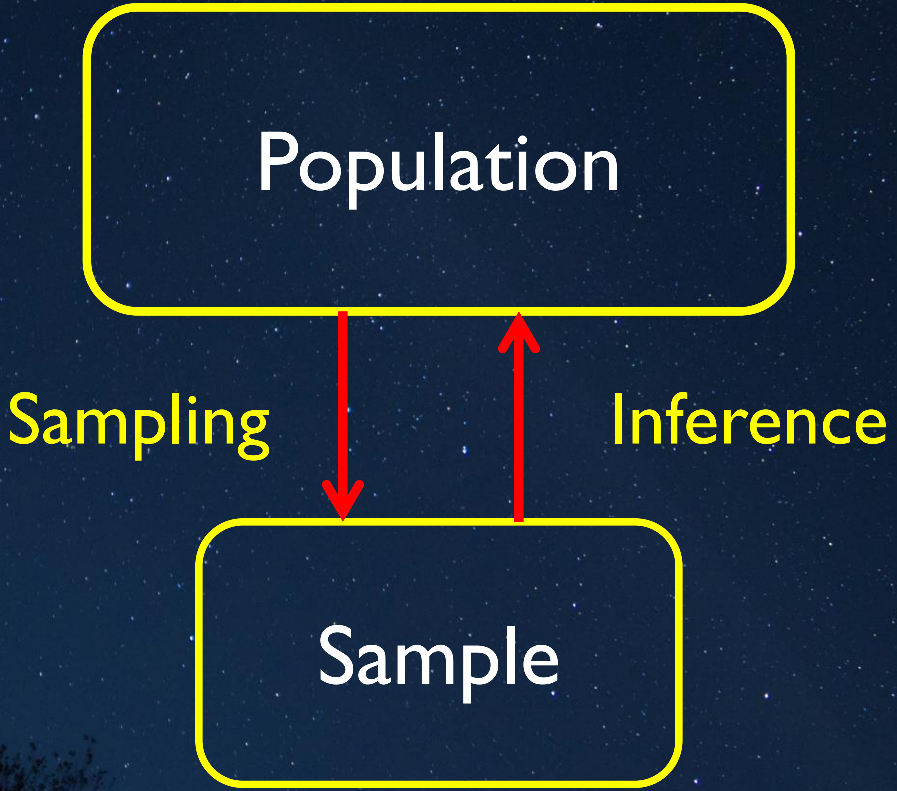
- Population: any group of interest or any group that researchers want to learn more about.
    - Population parameters (unknown to us): characteristics of population

- Sample: a group of individuals or data are drawn from population of interest.
    - Sample statistics: characteristics of sample

Population and sample

- We are much more interested in the population from which the sample was drawn.
  - Example: 30 GPAs as a representative sample drawn from the population of GPAs of the freshmen currently in attendance at ECU.

Population and sample



Population

Sampling

Inference

Sample

Sampling

- Sampling is the procedure of selecting a sample from a population.
- There are different strategies we use to select our sample
    - Convenience sampling, Quota sampling, snowball sampling, stratified sampling, simple random sampling, etc.

Inference

- Statistical inference uses sample data to draw conclusions about the entire population.

Types of measurement

- Discrete: Quantitative data are called discrete if the sample space contains a finite or countable infinite number of values.
  - How many days did you smoke during the last 7 days

Types of measurement

- Continuous: Quantitative data are called continuous if the sample space contains an interval or continuous span of real numbers.
  - Weight, height, temperature
  - Height: 1.72 meters, 1.7233330 meters

Types of measurement

- Nominal
  - Categorical variables. Numbers that are simply used as identifiers or names represent a nominal scale of measurement such as female vs. male.

Types of measurement

- Ordinal
  - An ordinal scale of measurement represents an ordered series of relationships or rank order.
    - Likert-type scales (such as "On a scale of 1 to 10, with one being no pain and ten being high pain, how much pain are you in today?") represent ordinal data.

Types of measurement

- Interval: A scale that represents quantity and has equal units but for which zero represents simply an additional point of measurement.
  - The Fahrenheit scale is a clear example of the interval scale of measurement. Thus, 60 degrees Fahrenheit or -10 degrees Fahrenheit represent interval data.

Types of measurement

- Ratio: The ratio scale of measurement is similar to the interval scale in that it also represents quantity and has equality of units. However, this scale also has an absolute zero (no numbers exist below zero).
  - For example, height and weight.

Types of measurement

- Qualitative vs. Quantitative variables
  - Qualitative variables: values are texts (e.g., Female, male), we also call them string/character variables.
  - Quantitative variables: are numeric variables.

Types of measurement

- Response/outcome variables: measure outcomes of a study. They are also called Dependent variables.
- Explanatory variables: explain or influence changes in a response variable. They are also called Independent variables.
  - Gender, sex, or other demographics., treatments, etc.

Basic statistics

- Two types of statistics
  - Descriptive statistics
  - Inferential statistics

Basic statistics

- Descriptive statistics: "are procedures used to summarize, organize, and make sense of a set of scores or observations."

Basic statistics

- Inferential statistics:

  "are procedures used that allow researchers to infer or generalize observations made with samples to the larger population from which they were selected."

Basic statistics

- Descriptive statistics for scale variables:
  - Central tendency
  - Dispersion

Central Tendency

- Measures of Central tendency: we use statistical measures to locate a single score that is most representative of all scores in a distribution.
  - Mean
  - Median
  - Mode

Mean

- The notations used to represent population parameters and sample statistics are different.
    - For example
        - Population size : N
        - Sample size : n

Mean

- Mean
  - $\bar{X}$ (or M) for sample mean and μ for population mean
  - $\bar{X}$ (x bar) = $\frac{\sum x}{n}$
  - ∑x means sum of all individual scores of x_1-x_n
  - n means number of scores

Mean

- Example 1: we want to know how 25 students performed in math tests.
- Data are in the next slide.

Mean

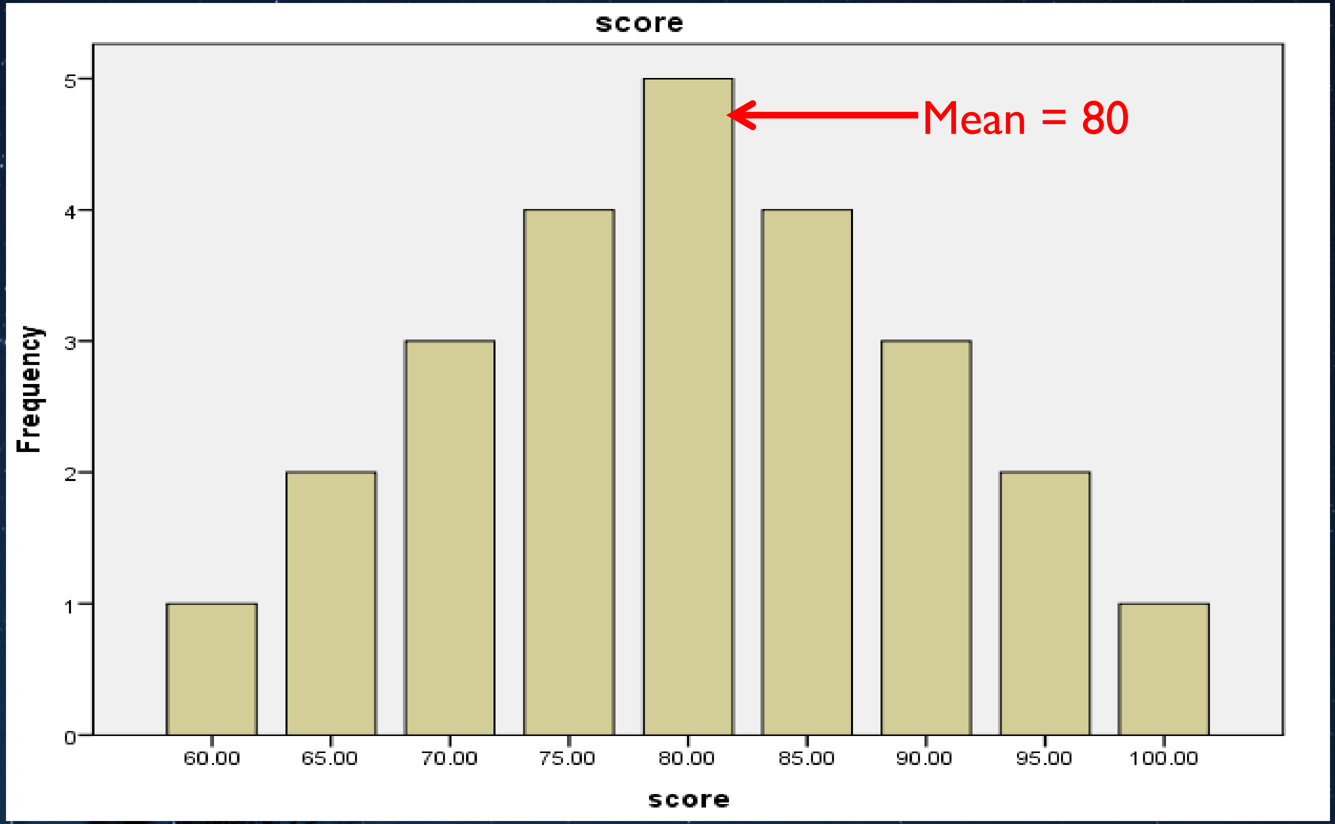
Score (X)	Frequency (f)	fX
60	1	60
65	2	130
70	3	210
75	4	300
80	5	400
85	4	340
90	3	270
95	2	190
100	1	100
Sum	25	2000

Mean

- How to calculate mean for those 25 scores?
- $\bar{X} = \frac{\sum fx}{n} = \frac{2000}{25} = 80.00$

Mean

- Distribution of Example 1



Median

- Median
  - Data: 2, 3, 4, 5, 7, 10, 80. Mean of those scores is 15.86.
  - 80 is an outlier.
  - Mean fails to reflect most of the data. We use median instead of mean to remove the influence of an outlier.
  - Median is the middle value in a distribution of data listed in a numeric order.

Median

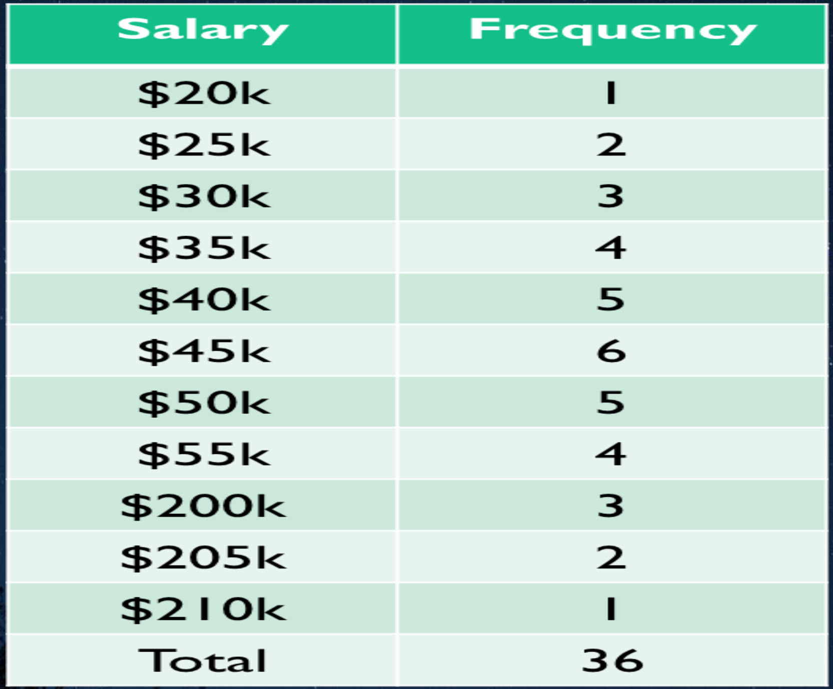
- Median
  - Position of median = $\frac{n+1}{2}$
  - For odd –numbered sample size: 3,6,5,3,8,6,7. First place each score in numeric order: 3,3,5,6,6,7,8. Position 4: (7+1)/2 is the median. Median = 6

Median

- Median
  - For even-numbered sample size: 3,6,5,3,8,6. First place each score in a numeric order: 3,3,5,6,6,8. Position = (6+1)/2= 3.5. Median = $\frac{5+6}{2}$ = 5.5
  - Example 2: we want to know average salary of 36 cases.

Median

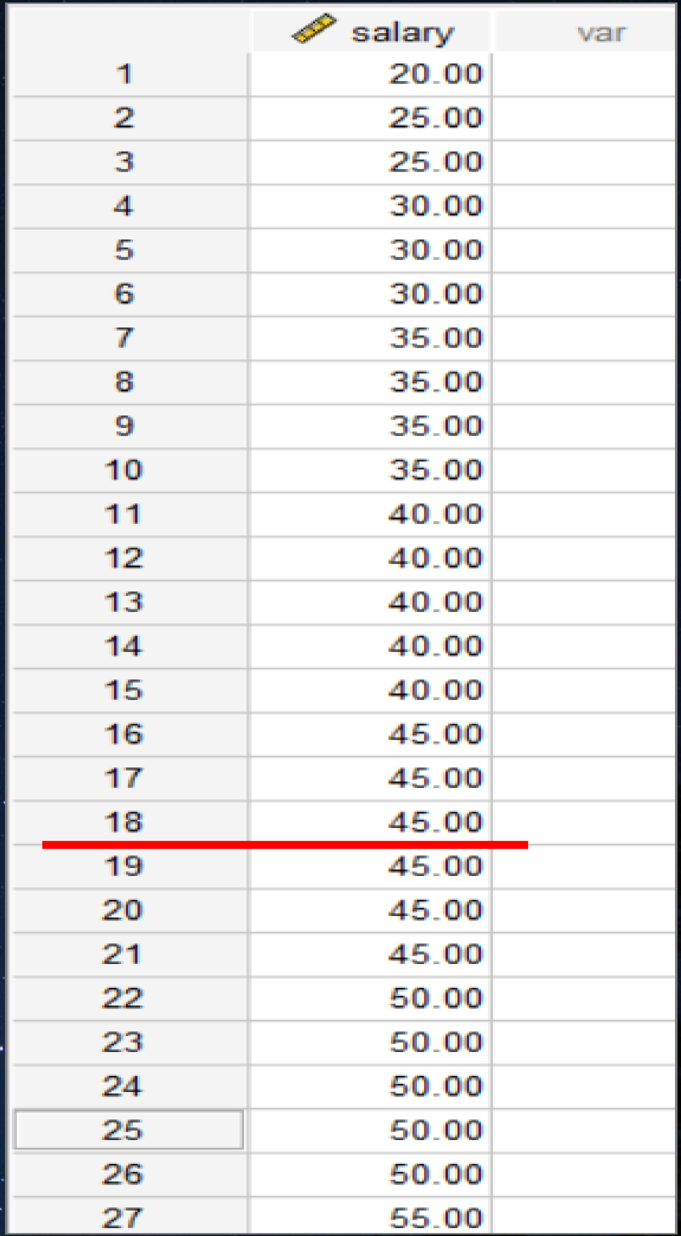
- Median
  - Example 2: we want to know average salary of 36 cases.



Salary	Frequency
$20k	1
$25k	2
$30k	3
$35k	4
$40k	5
$45k	6
$50k	5
$55k	4
$200k	3
$205k	2
$210k	1
Total	36

Median

- Median = ?
- Position = (36 +1)/2 = 18.5
- Which number is at position 18.5?
- Median = $45k



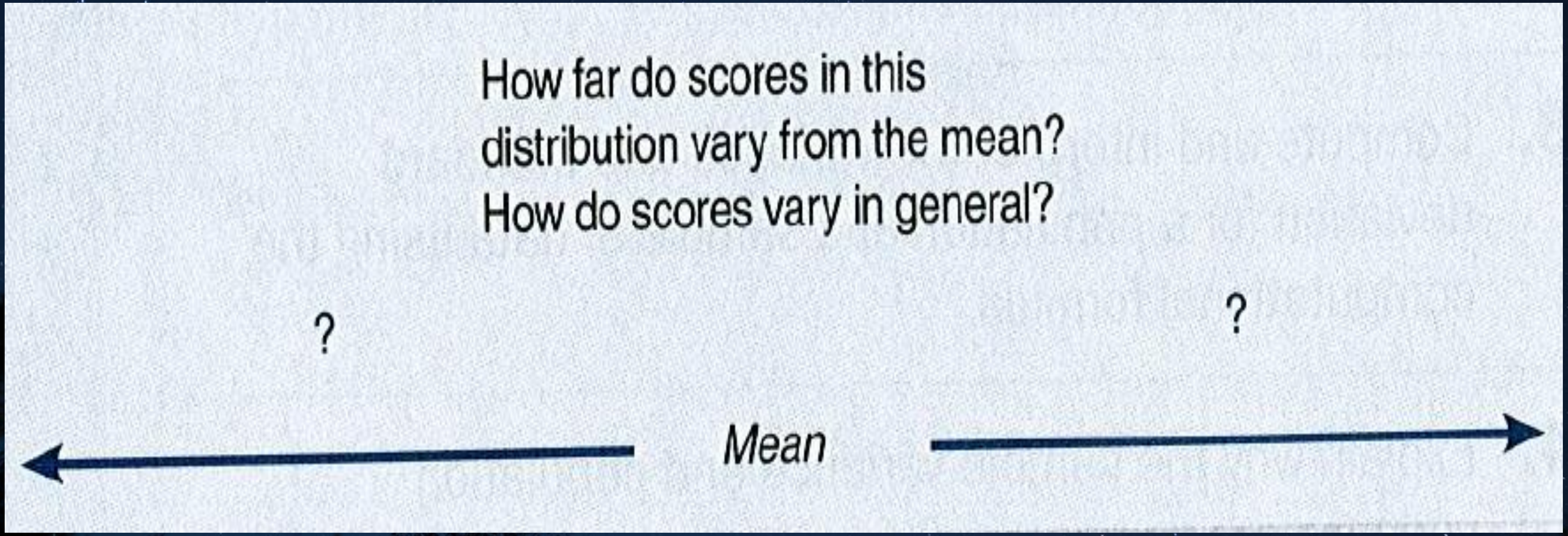
	salary	var
1	20.00	
2	25.00	
3	25.00	
4	30.00	
5	30.00	
6	30.00	
7	35.00	
8	35.00	
9	35.00	
10	35.00	
11	40.00	
12	40.00	
13	40.00	
14	40.00	
15	40.00	
16	45.00	
17	45.00	
18	45.00	
19	45.00	
20	45.00	
21	45.00	
22	50.00	
23	50.00	
24	50.00	
25	50.00	
26	50.00	
27	55.00	

Percentile Values

- Rank all observations from the lowest to the highest
  - 25th percentile value: 25% values lie below that value
  - 50th percentile value: this is median. 50% values lie below that value
  - 75th percentile value: 75% values lie below that value

Dispersion

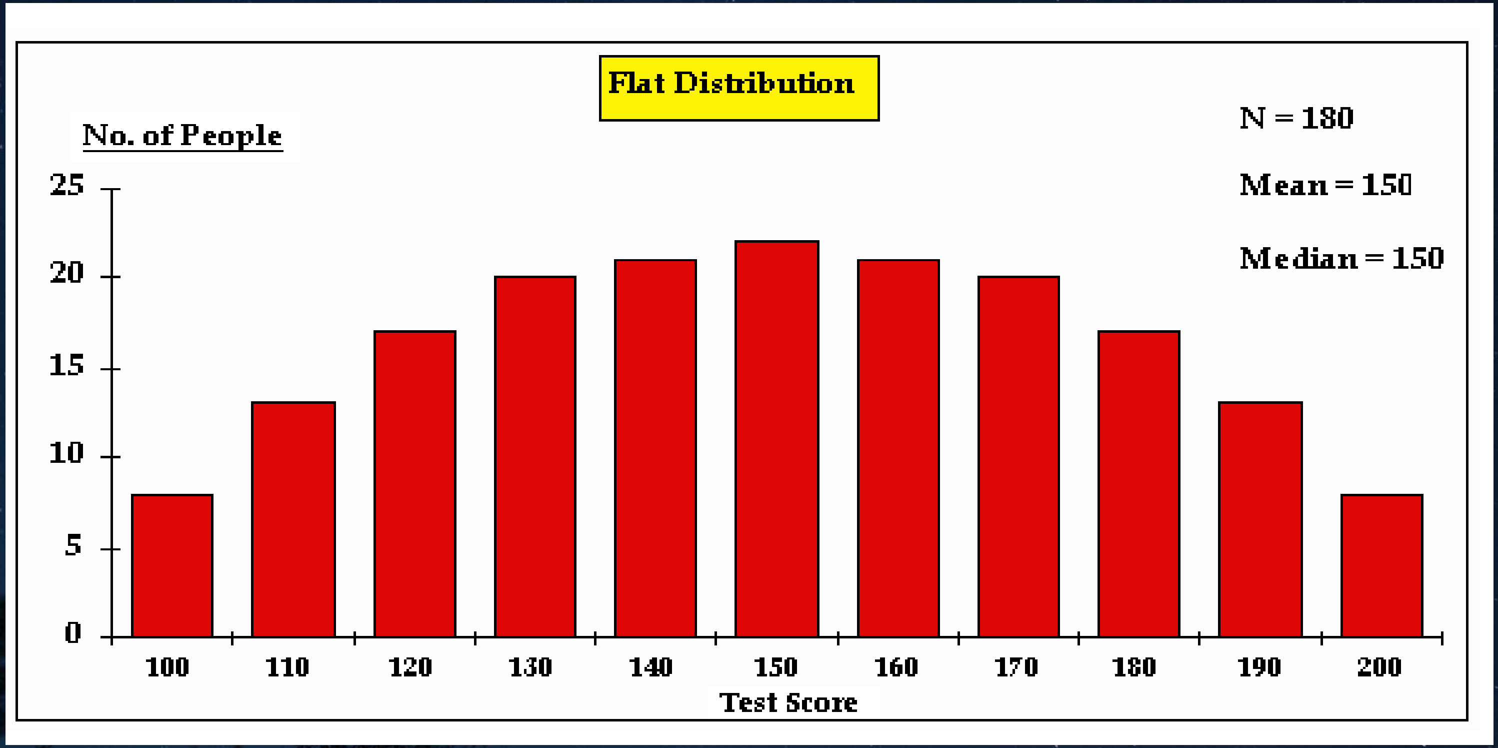
- Dispersion (Variability): a measure of the spread of scores in a distribution.



How far do scores in this distribution vary from the mean? How do scores vary in general?

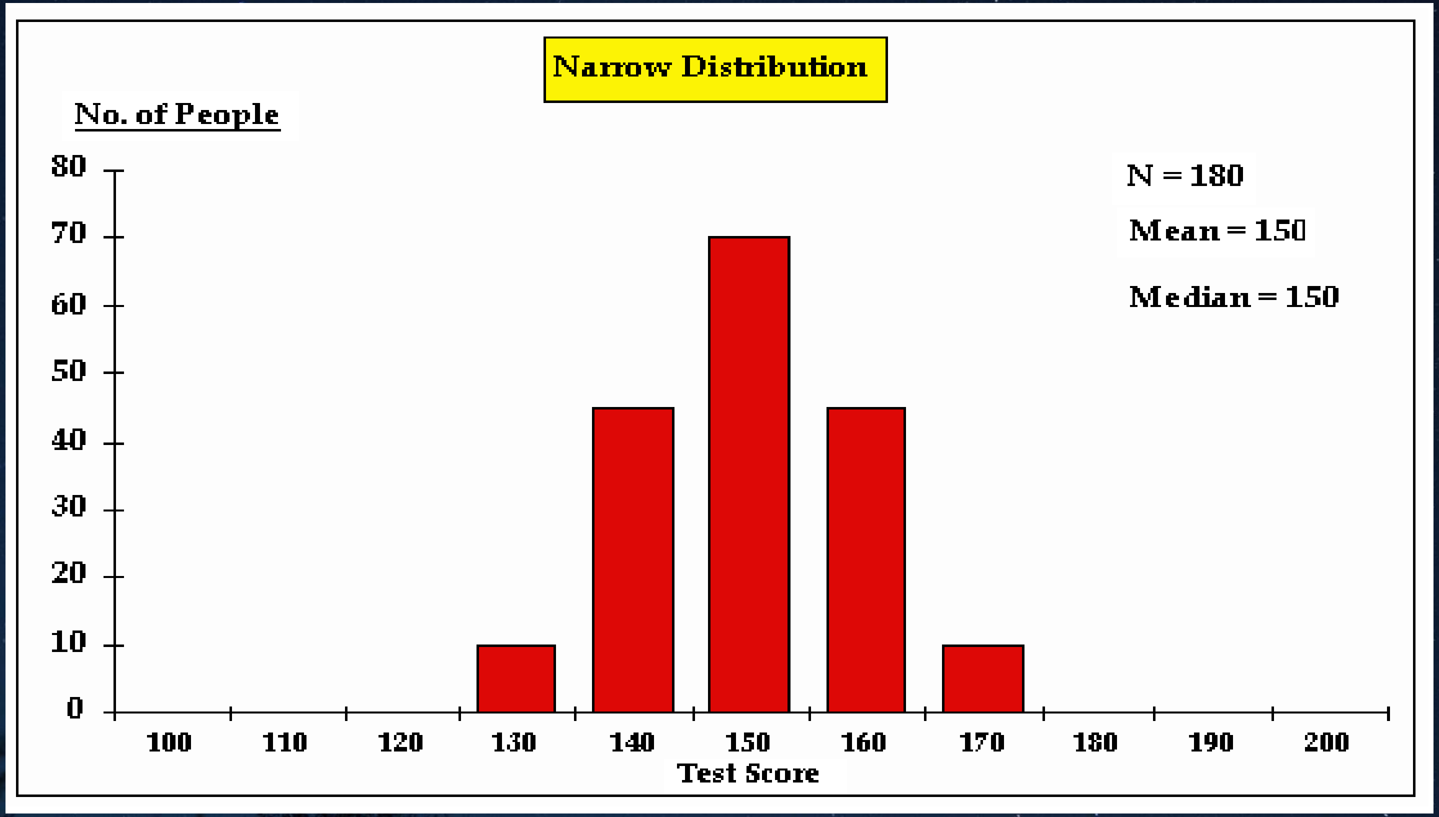
Dispersion

- Compare different distributions



Dispersion

- Compare different distributions



Dispersion

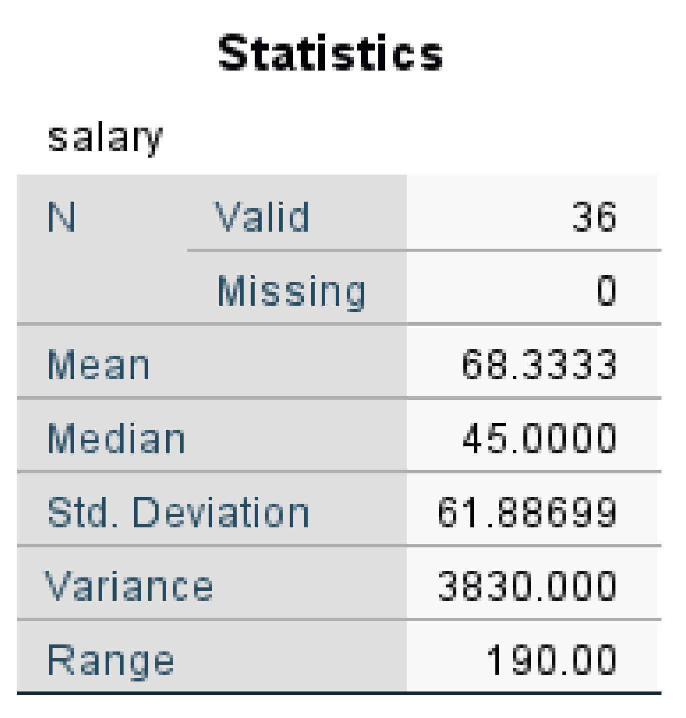
- Two sets of data have the same sample size, mean, and median.
- But they are different in terms of variability.

Dispersion

- Dispersion
  - Range
  - Variance
  - Standard deviation

Dispersion

- In our previous example about salary:



Statistics

salary

N	Valid	36
	Missing	0
Mean		68.3333
Median		45.0000
Std. Deviation		61.88699
Variance		3830.000
Range		190.00

Dispersion

- Range
  - It is the difference between the largest value and smallest value.
  - It is informative for data without outliers.

Variance

- It measures the average squared distance that scores deviate from their mean.
- Sample variance: s^2 (population variance σ^2 sigma)

Variance

- How to calculate variance?
  - $s^2 = \frac{\sum(x - \bar{x})^2}{n-1}$ or $\frac{ss}{n-1}$: ss means sum of squares.
  - n-1 means: degree of freedom: the number of scores in a sample that are free to vary.

Variance

- Example: five scores: 5, 10, 7, 8, 15
  - Mean = (5 + 10 + 7 + 8 + 15)/5 = 9
  - Let's calculate variance
    - *SS* = $(5-9)^2 + (10-9)^2 + (7-9)^2 + (8-9)^2 + (15-9)^2 = 58$
    - Sample variance = 58/(5-1) = 14.5

Standard Deviation

- Standard deviation (s, σ)
  - It is the square root of variance.
  - It is average distance that scores deviate from their mean.
  - $s = \sqrt{\frac{ss}{n-1}}$

Standard Deviation

- Example 3: calculate standard deviation

Scores (x)	Frequency(f)	$x - \overline{x}$ (d)	d^2	fd^2(ss)
100	6	100-115.5=-15.5	240.25	6*240.25
110	12	110-115.5= -5.5	30.25	12*30.25
120	16	120-115.5=4.5	20.25	16*20.25
130	6	130-115.5=14.5	210.25	6*210.25
Sum	40			3390.0

Standard Deviation

- s = $\sqrt{\frac{3390}{40-1}} = 9.32$
- $\bar{X}$ = 115.5
- Summary:
  - When individual scores are close to mean, the standard deviation (SD) is smaller.

Standard Deviation

- Summary
  - When individual scores are spread out far from the mean, the standard deviation is larger.
  - SD is always positive
  - It is typically reported with mean.

Descriptive statistics

- Choosing proper measure of central tendency depends on:
  - the type of distribution
  - the scale of measurement

Descriptive statistics

- Mean describes data that are normally distributed and measures on an interval or ratio scale.
- Median is used when the data are not normally distributed.

Normal distribution

- Normal distribution
  - Probability: the frequency of times an outcome is likely to occur divided by the total number of possible outcomes.
    - It varies between 0 and 1.
    - Example (next slide)

Probability

- Probability

	Fail	Pass	Total
Male	3	2	5
Female	1	4	5
Total	4	6	10

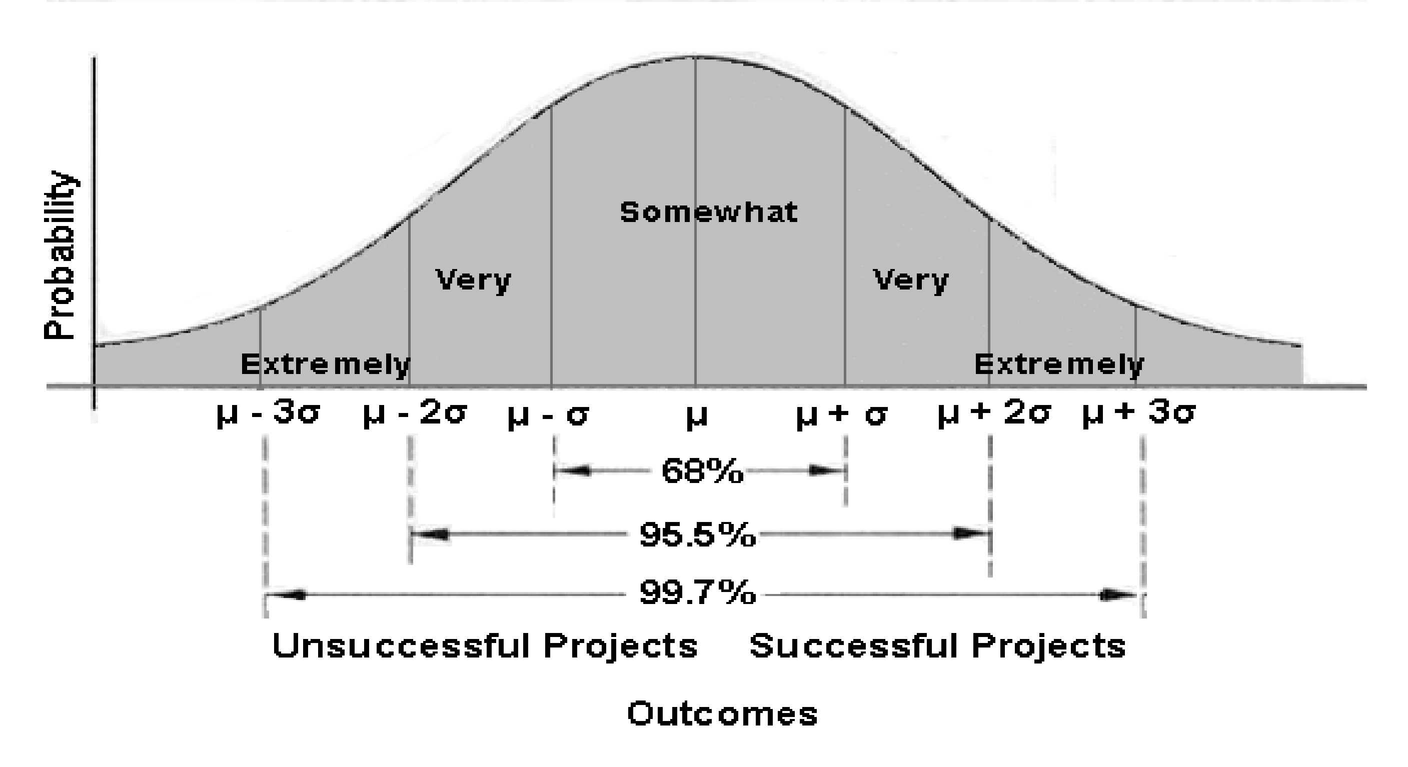
1. What is the probability of Fail? 4/10 =.4
2. What is the probability of Pass? 6/10 = .6
3. What is the probability of Fail among males? 3/5 = .6
4. What is the probability of Pass among females? 4/5 = .8

Normal Distribution

- Normal distribution/Normal curve
  - Data are symmetrically distributed around mean, median, and mode.
  - Also called the symmetrical, Gaussian, or bell-shaped distribution.

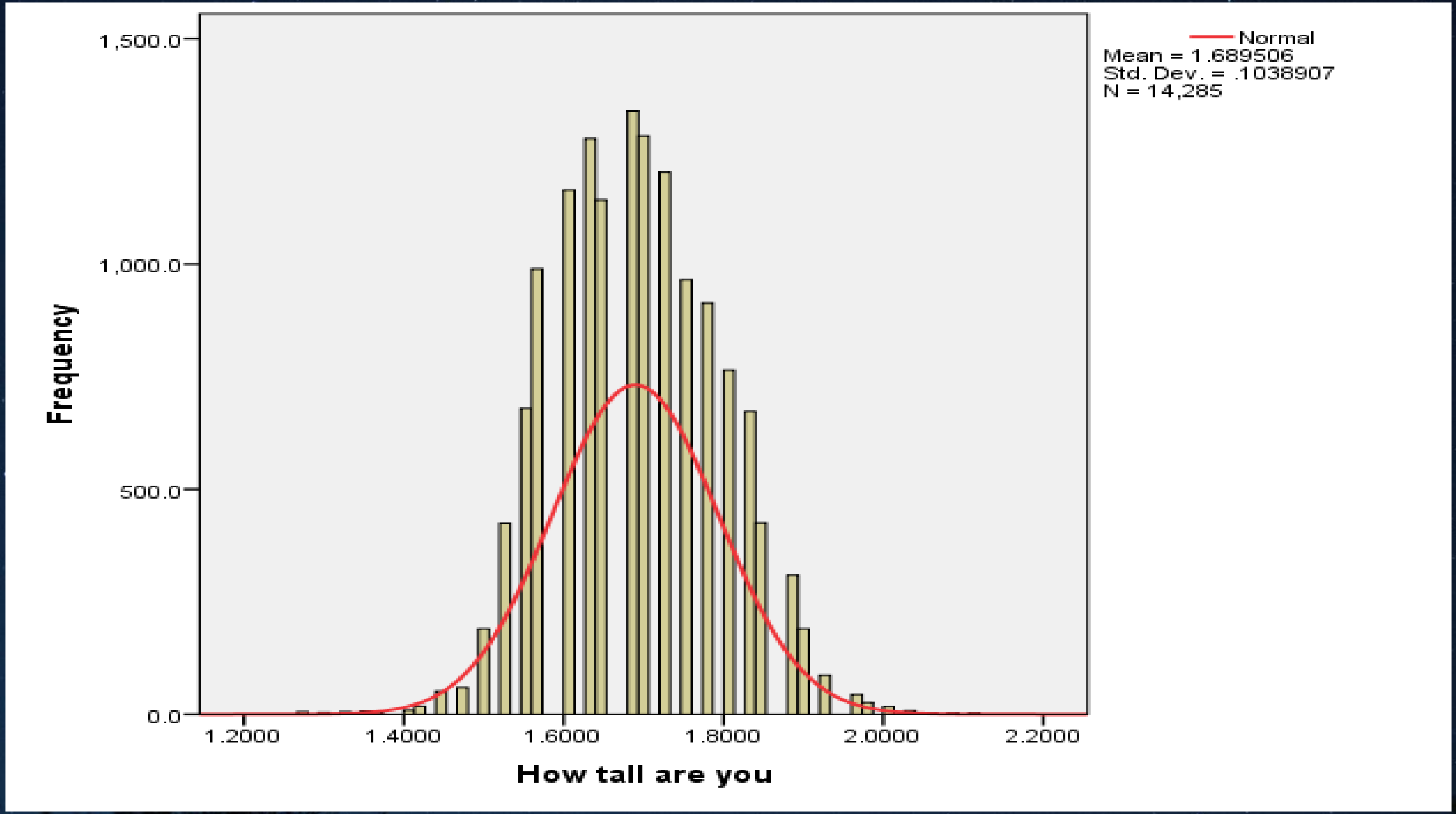
Normal Distribution

- Normal curve



Normal Distribution

- Normal curve



Normal Distribution

- Characteristics of normal distribution curve
  - The normal distribution is mathematically defined.
  - The normal distribution is theoretical.
  - The mean, median, and mode are all the same value at the center of the distribution.

Normal Distribution

- Characteristics of normal distribution curve
  - The normal distribution is symmetrical.
  - The form of a normal distribution is determined by its mean and standard deviation.
  - Standard deviation can be any positive value.

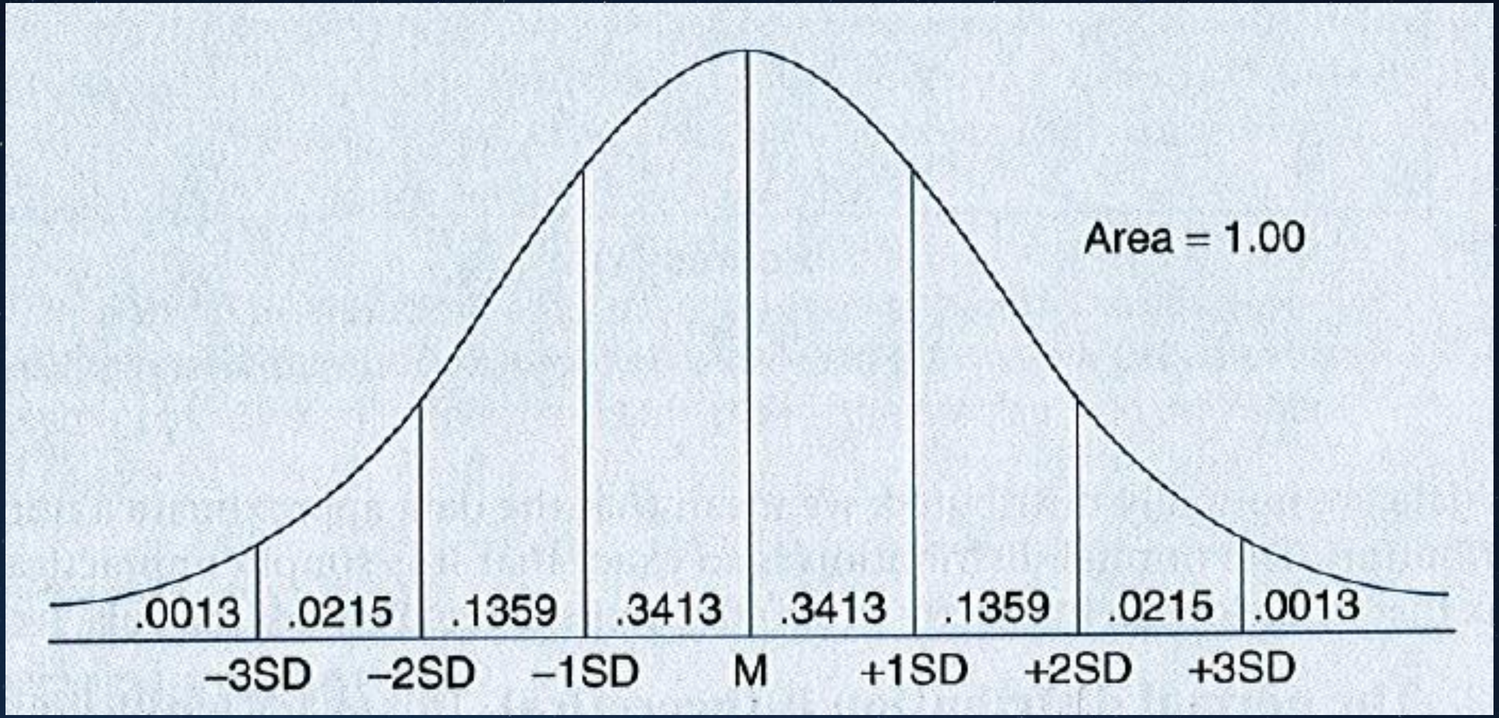
Normal Distribution

- Characteristics of normal distribution curve
  - The total area under the curve is equal to 1.
  - The tails of normal distribution are always approaching to x axis, but never touch it.

Normal Distribution

- Normal distribution/Normal curve
  - We use normal distribution to locate probabilities for scores.
  - The area under the curve can be used to determine the probabilities at different points.

Normal Distribution



Proportions of area under the normal curve

Normal Distribution

- Normal distribution: the standard deviation indicates precisely how the scores are distributed. Empirical rule:
    - About 68% of all scores lie within one standard deviation of the mean. In another word, roughly two thirds of the scores lie between one standard deviation on either side of the mean.

Normal Distribution

- Normal distribution
  - About 95% of all scores lie within two standard deviation of the mean (Normal scores: close to the mean).
  - About 99.7% of all scores lie within three standard deviation of the mean.

Normal Distribution

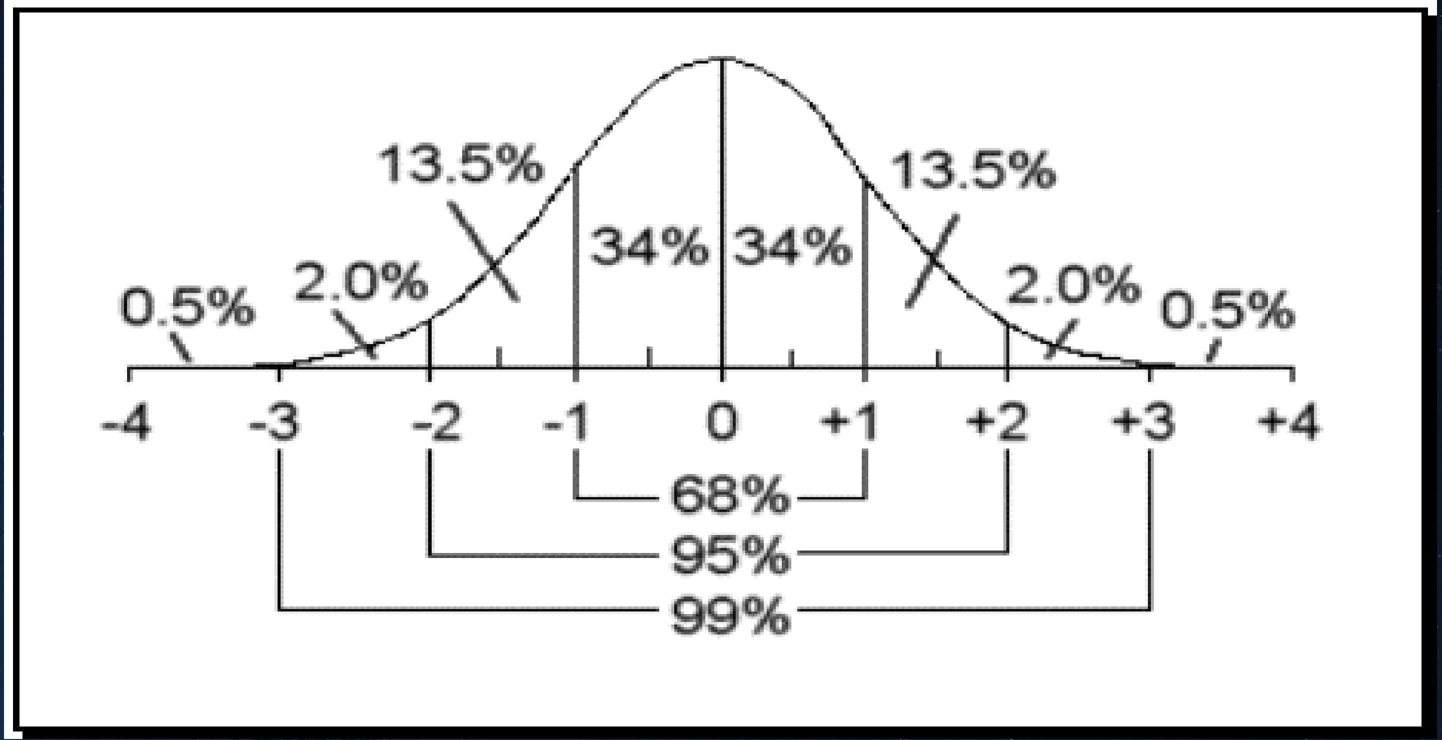
- In another word, we have 95% chance of selecting a score that is within 2 standard deviation of mean.
- Less than 5% scores are far from the mean (NOT normal scores).

Normal Distribution

- Standard normal distribution or Z distribution
  - A normal distribution with mean = 0, and standard deviation = 1.
  - A Z score is a value on the x-axis of a standard normal distribution

Normal Distribution

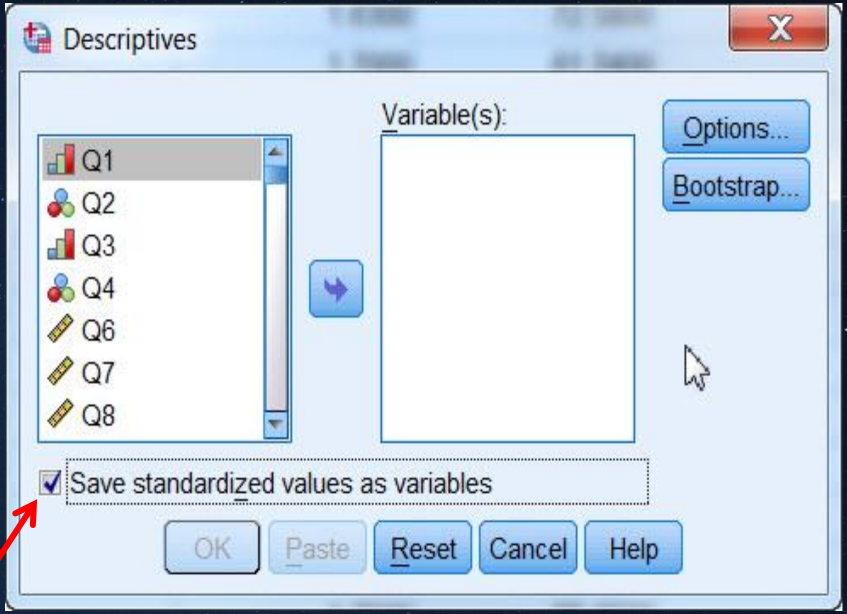
- Standard normal distribution or Z distribution



Normal Distribution

- z transformation

$$z = \frac{X - M}{SD}$$



☑ Save standardized values as variables

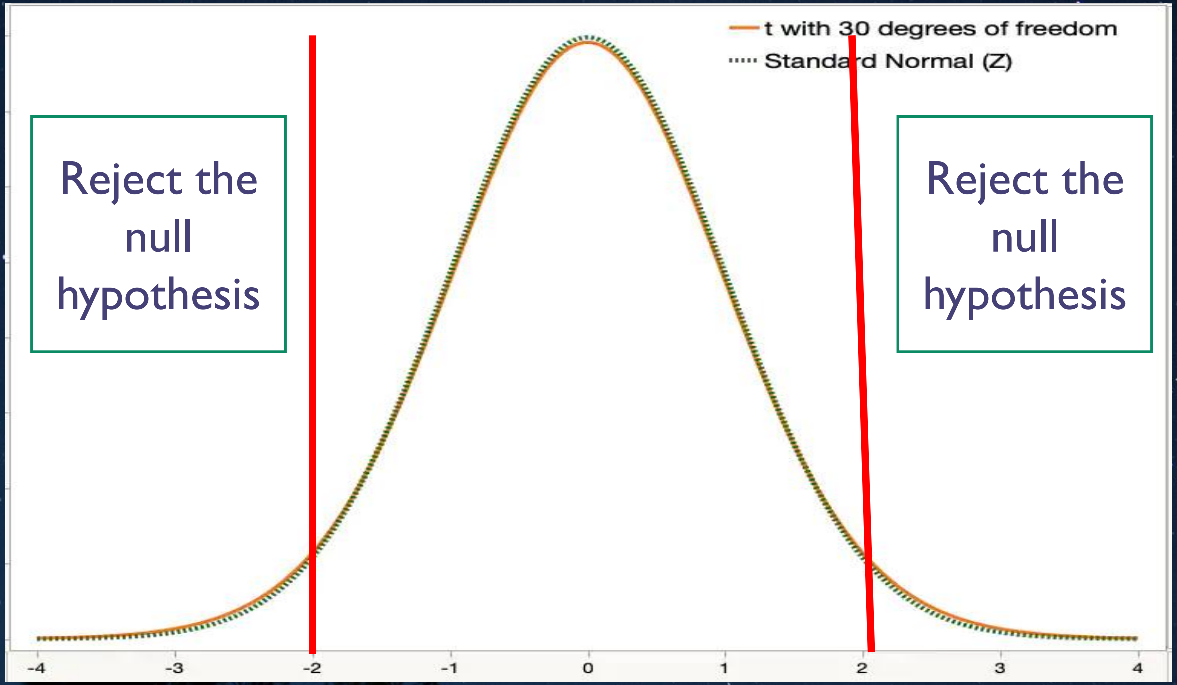
X means individual value, M is mean and SD is standard deviation.
In SPSS, go to Analyze > Descriptive Statistics > Descriptives to get Z scores

Normal Distribution

- In inferential statistics for example in independent samples t test.
- Null hypothesis (H_0): two means are equal
- Alternative hypothesis (H_A): two means are not equal

Normal Distribution

- The t distribution curve shows the distribution of t statistic for two-tailed test when H_0 is true.



Confidence Interval

Mean = 1.89 for example, it is a point estimate. But we want to know the probable accuracy of the estimate. How close that estimate is likely to fall to the true parameter value.

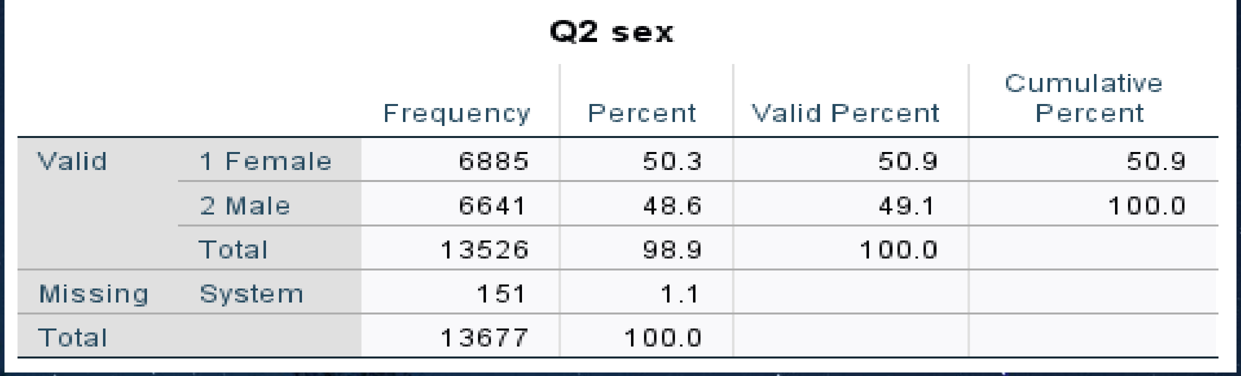
Confidence Interval

A confidence interval (CI) is a range of numbers which contains the parameter. The probability of the accuracy is called confidence coefficient: 95% or 99% . We use standard error to calculate CI.

95% CI for a mean: 95% times the interval contains population mean.

Descriptive Statistics

- For nominal and ordinal variables, we use Frequency and Percentage to describe the data.
- For example: description of Q2 (sex).



Q2 sex

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 Female	6885	50.3	50.9	50.9
	2 Male	6641	48.6	49.1	100.0
	Total	13526	98.9	100.0	
Missing	System	151	1.1		
Total		13677	100.0		

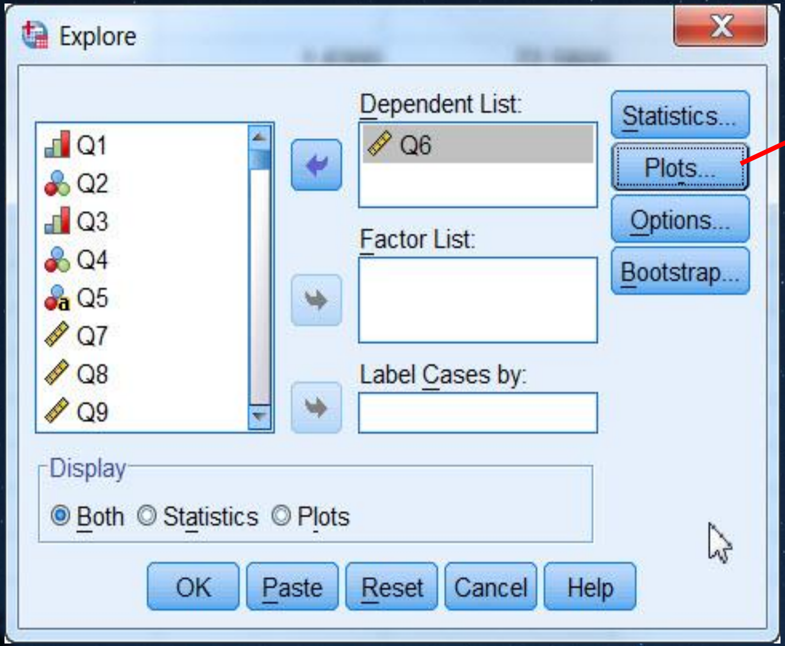
Descriptive statistics in SPSS

- Descriptive statistics in SPSS
  - Frequencies
  - Descriptives
  - Explore

Descriptive statistics in SPSS

- Exercise: use 2019 YRBSS data
  - Use Explore function to get descriptive statistics for Q6 (height)
  - Analyze > Descriptive Statistics > Explore

Descriptive statistics in SPSS



Explore

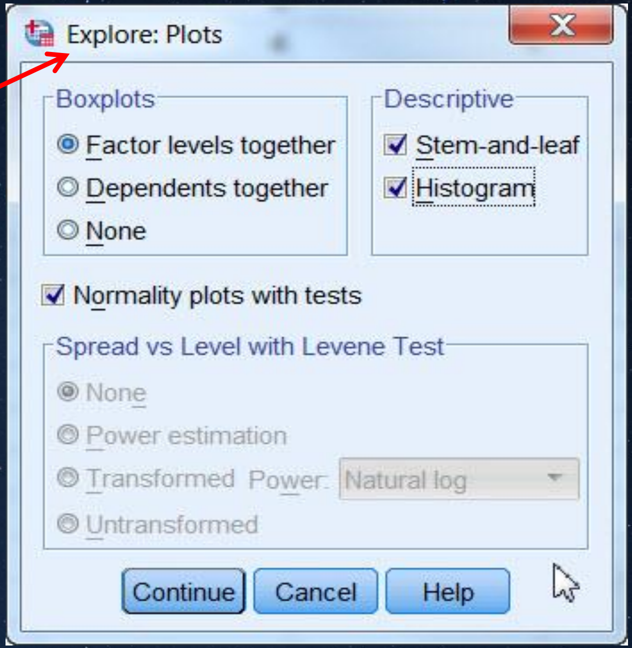
Q1
Q2
Q3
Q4
Q5
Q7
Q8
Q9

Dependent List: Q6

Factor List:

Label Cases by:

Display
● Both ○ Statistics ○ Plots



Explore: Plots

Boxplots
● Factor levels together
○ Dependents together
○ None

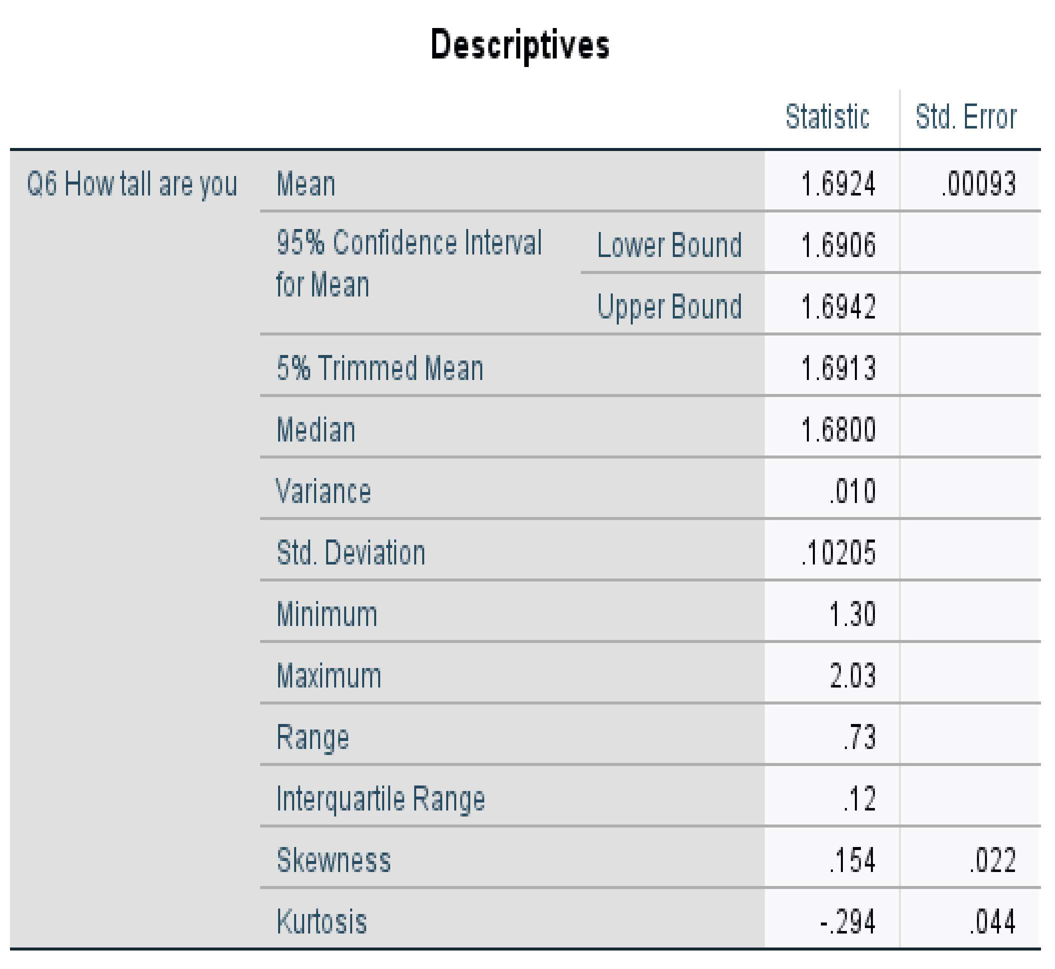
Descriptive
☑ Stem-and-leaf
☑ Histogram

☑ Normality plots with tests

Spread vs Level with Levene Test
● None
○ Power estimation
○ Transformed Power: Natural log
○ Untransformed

Descriptive statistics in SPSS

- SPSS output

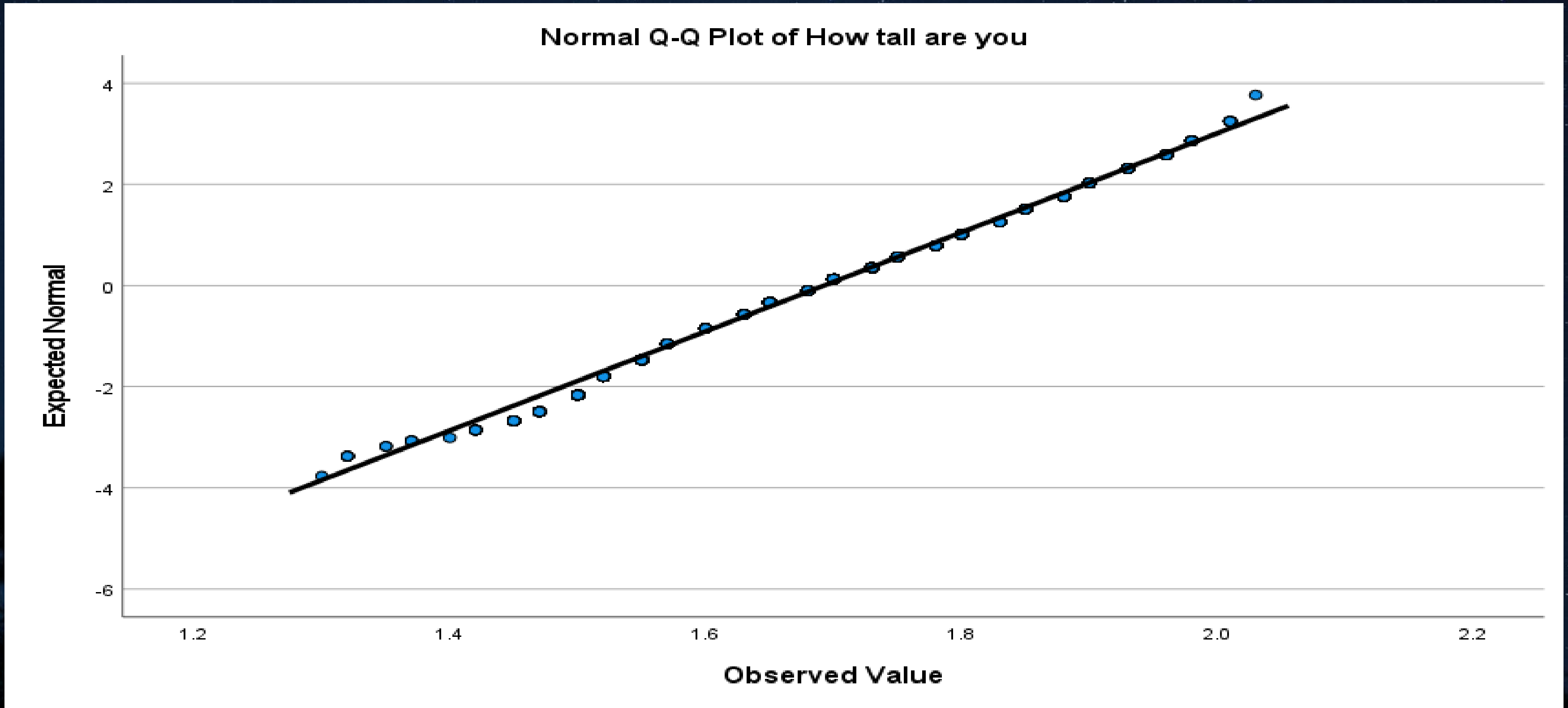


Descriptives

			Statistic	Std. Error
Q6 How tall are you	Mean		1.6924	.00093
	95% Confidence Interval for Mean	Lower Bound	1.6906	
		Upper Bound	1.6942	
	5% Trimmed Mean		1.6913	
	Median		1.6800	
	Variance		.010	
	Std. Deviation		.10205	
	Minimum		1.30	
	Maximum		2.03	
	Range		.73	
	Interquartile Range		.12	
	Skewness		.154	.022
	Kurtosis		-.294	.044

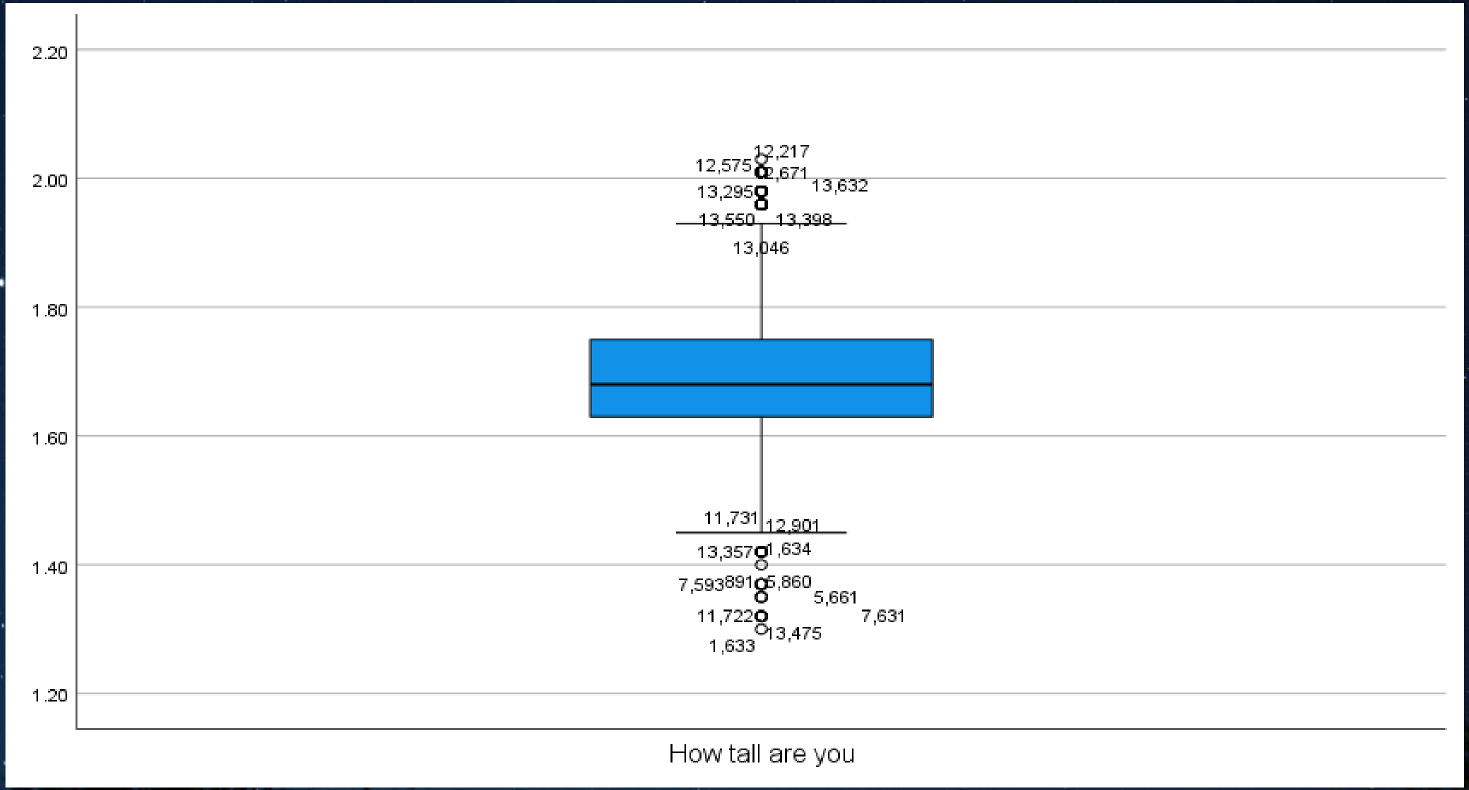
Descriptive statistics in SPSS

- SPSS output: Normal Quantile-Quantile (Q-Q) plot



Descriptive statistics in SPSS

- SPSS output: boxplot

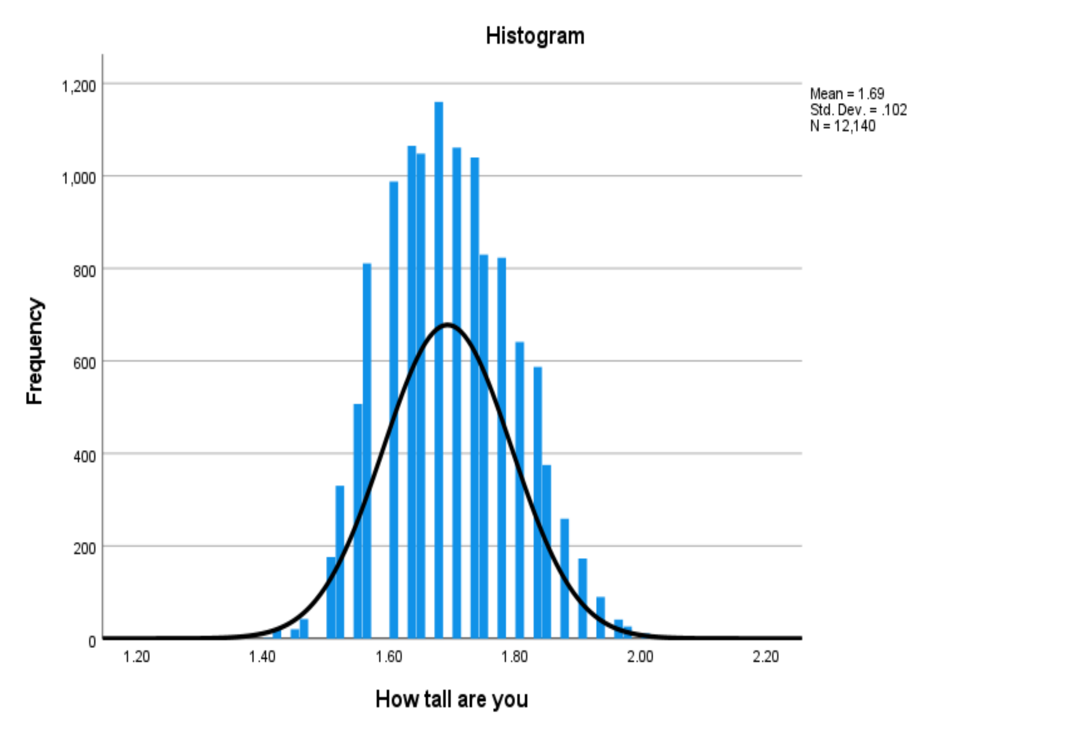


Graphs

- Summarize quantitative data graphically
  - It depends on the type of data
- Histogram: we use Histogram to summarize scale data.

Histogram

- Example: Q6 (height)



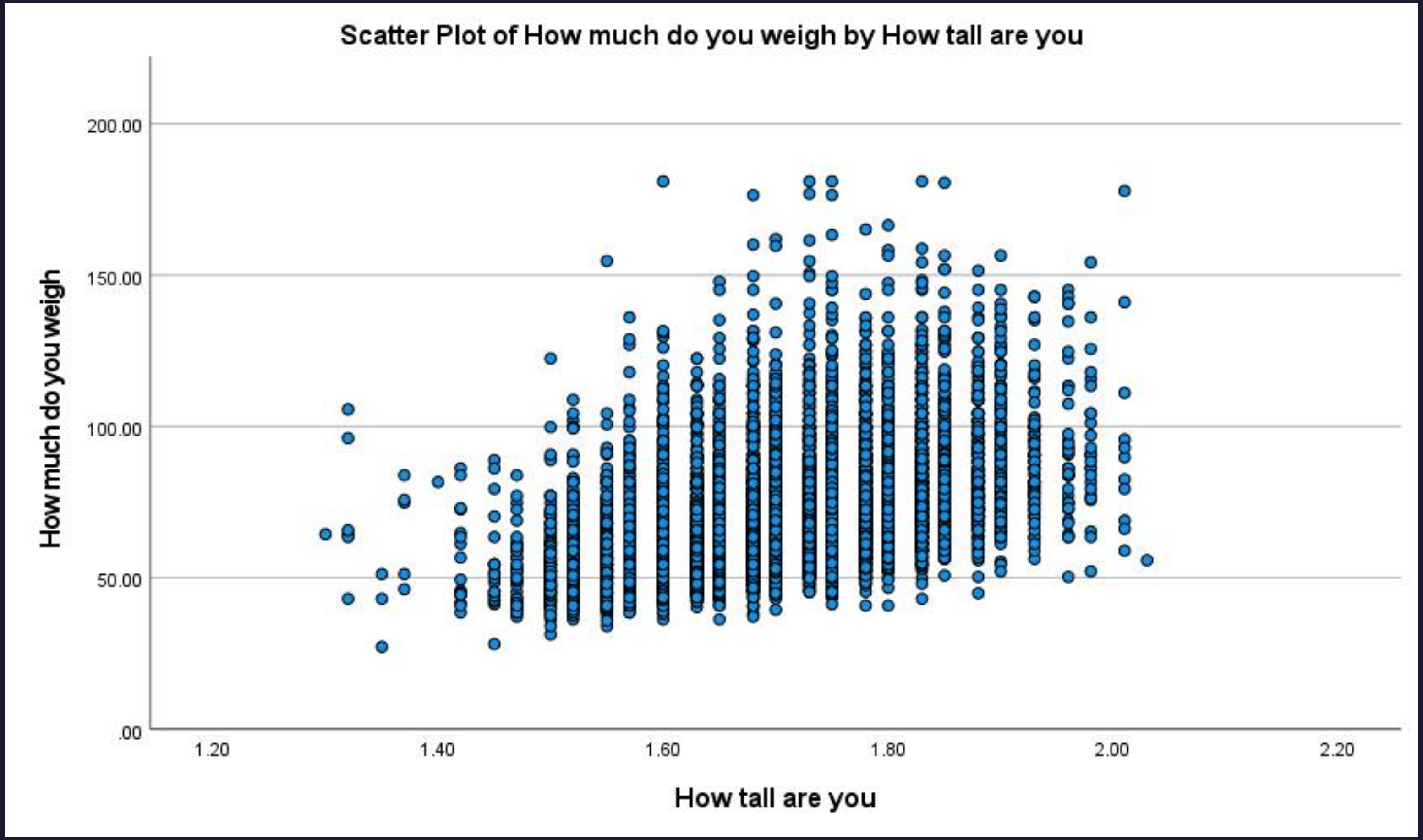
We use histogram to know the distribution of Q6.
Y axis represents frequency and X axis represents the responses.

Scatter Plot

- We use scatter plot to check linear relationship between two scale variables
- Example: Q6 (height) and Q7 (weight) by Q2 (gender)

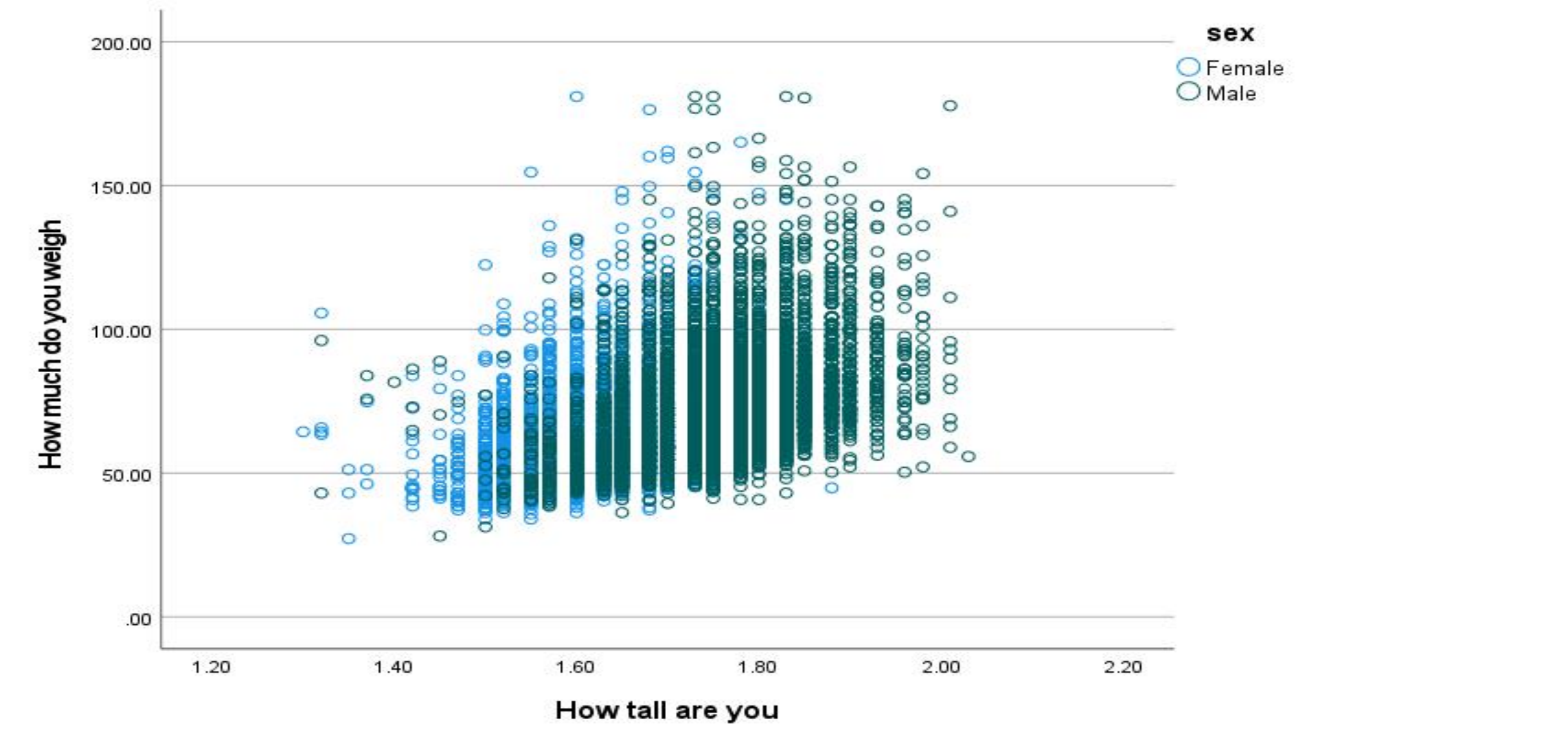
Scatter Plot

Scatter Plot: without grouping variable (Q2)



Scatter Plot

- Scatter plot by gender



Box Plot

- We can use either Explore function or Graphs to get box plot
- Example: box plot for Q6 (height) by Q2 (gender)

Box Plot

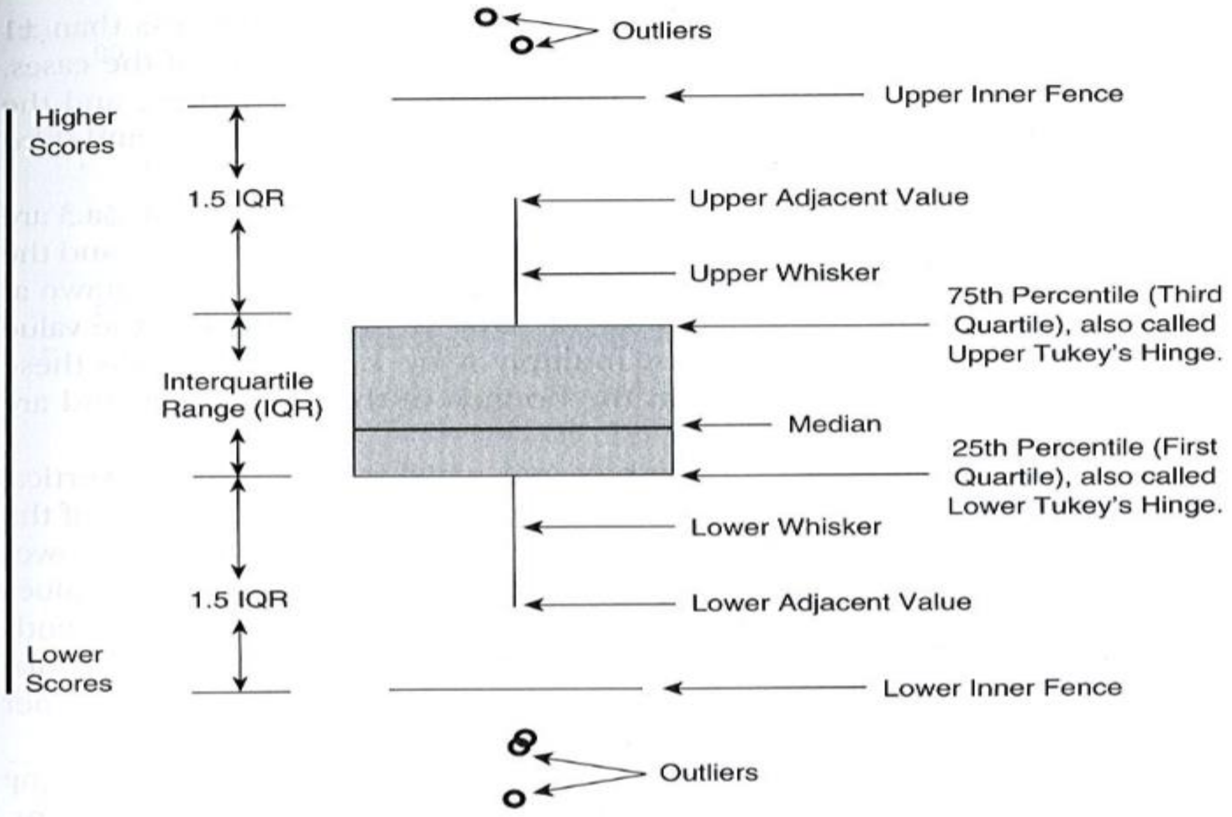
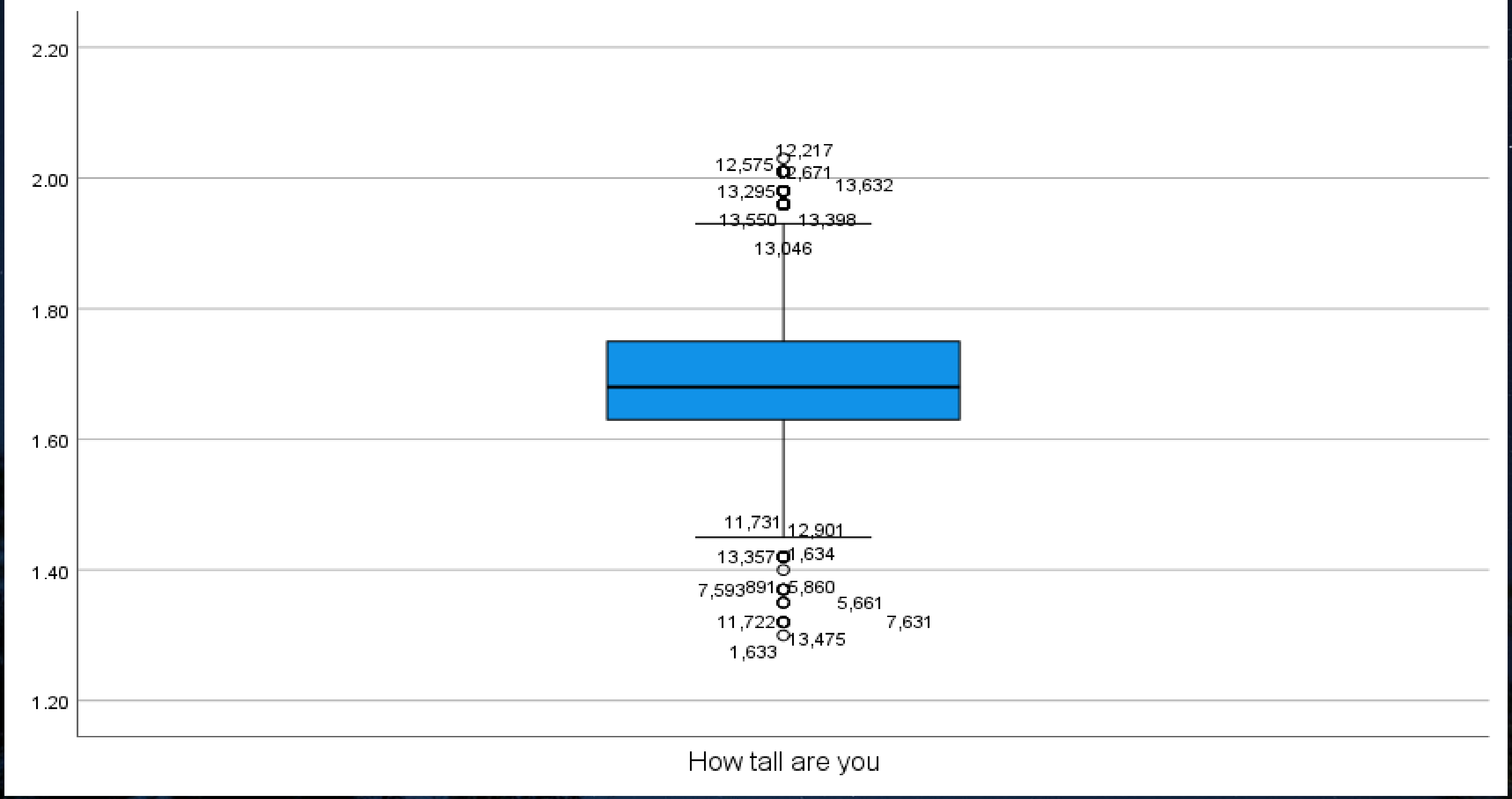


Figure 3a.3 The General Form of a Box and Whiskers Plot Based on Cohen's (1996) Description

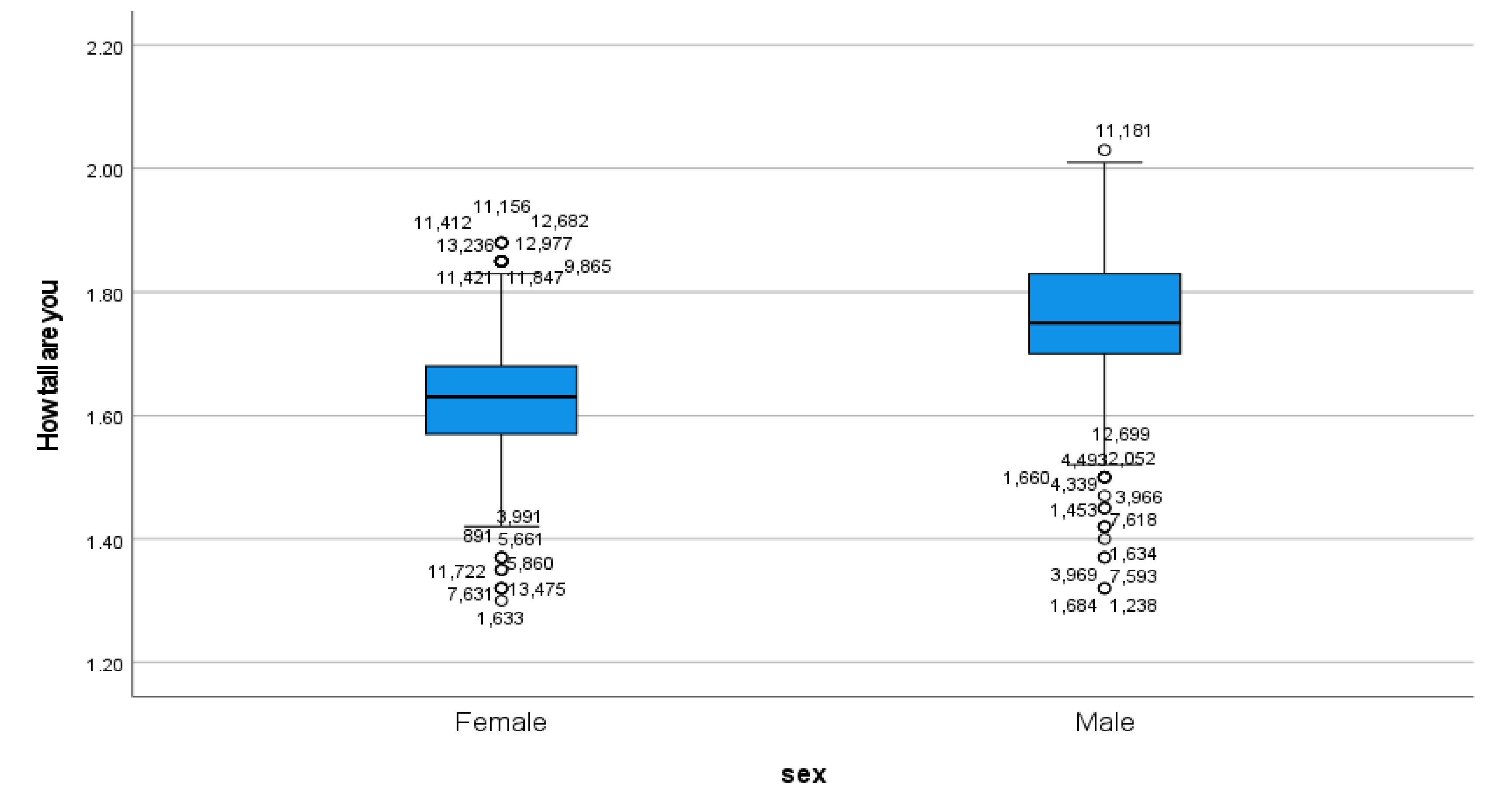
Box Plot

- Box plot of Q6 without Q2



Box Plot

- Box plot of Q6 by Q2

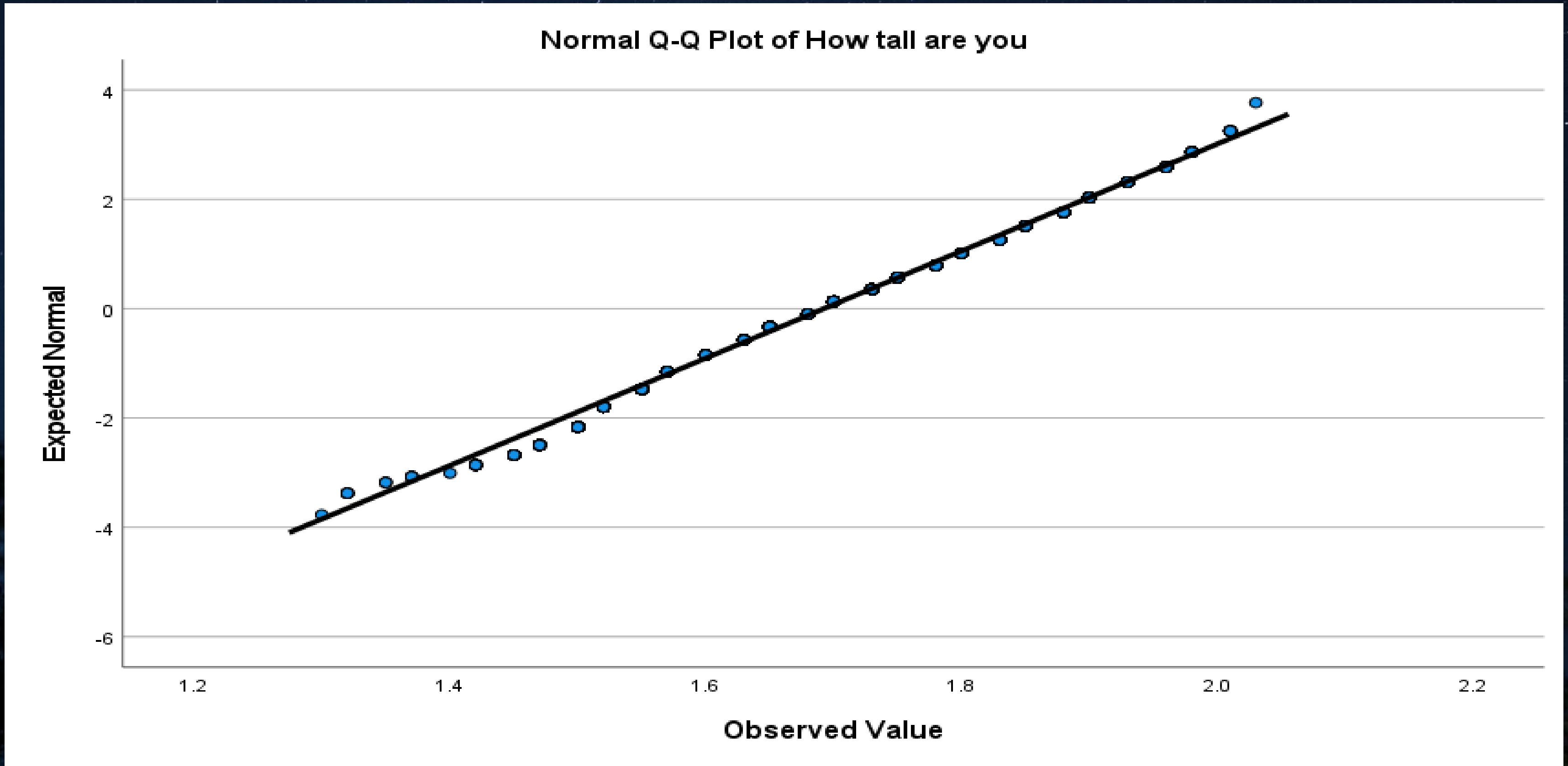


Normal Q-Q plot

- Normal Q-Q plot or quantile-quantile plot
- We use Normal Q-Q plot to check normality assumption: we assume that Q6 is normally distributed.
- If the data indeed follow the normal distribution, then the points on the Q-Q plot will fall approximately on a straight line.

Normal Q-Q plot

- Example: normal Q-Q plot for Q6 (height)



Basic Statistics

- Compare means
  - T test, ANOVA
- Compare medians
  - Non-parametric tests

Basic statistics

- References
  - Agresti, A. & Finlay, B. (1997). Statistical methods for the social sciences. Upper Saddle River, NJ. Prentice Hall, Inc.
  - Neutens, J. J., & Rubinson, L. (1997). *Research techniques for the health sciences*. Needham Heights, MA. Allyn & Bacon.

Basic statistics

- References
  - Privitera, G. J. (2012). *Statistics for the behavioral sciences*. Thousand Oaks, CA. SAGE Publications, Inc.

Thank You