Basic Statistics I

Hui Bian

Office for Faculty Excellence

Basic statistics

- My contact information:
  - Hui Bian, Statistics & Research Consultant
  - Office for Faculty Excellence, 1001 Joyner library, room 1006
  - Email: bianh@ecu.edu
  - Website: http://core.ecu.edu/ofe/StatisticsResearch/

Basic statistics

- 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.

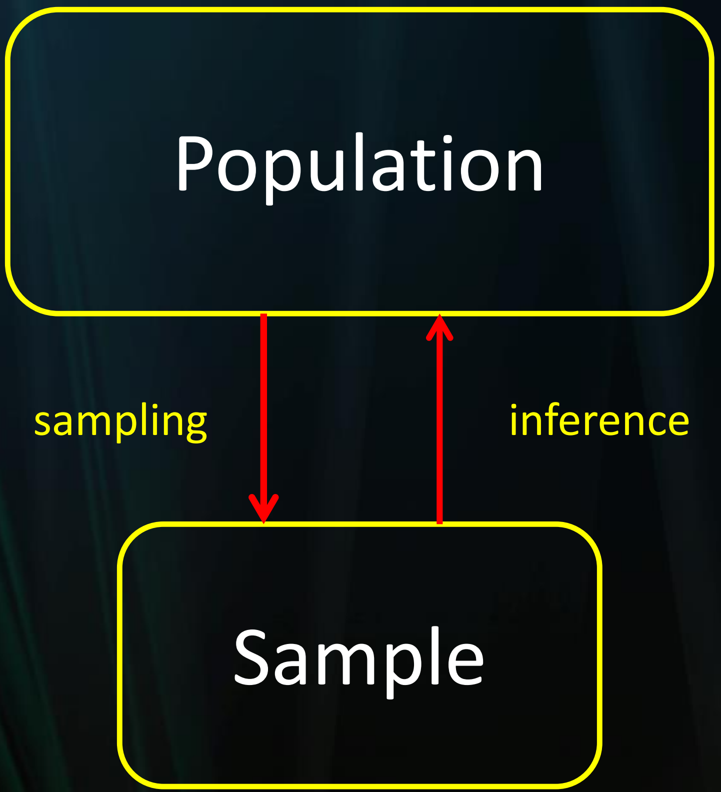
Population and sample

- 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 a certain university or the population of freshmen attending colleges similar to a certain university.

Population and sample



Types of measurement

- Discrete: Quantitative data are called discrete if the sample space contains a finite or countably 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 degree 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 variables.
  - Quantitative variables: are numeric variables.

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.”

Descriptive statistics

- Use descriptive statistics to describe, summarize, and organize set of measurements.
- Use descriptive statistics to communicate with other researchers and the public.
- Descriptive statistics: Central tendency and Dispersion

Descriptive statistics

- 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

Descriptive statistic

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

Descriptive statistics

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

Descriptive statistics

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

Descriptive statistics

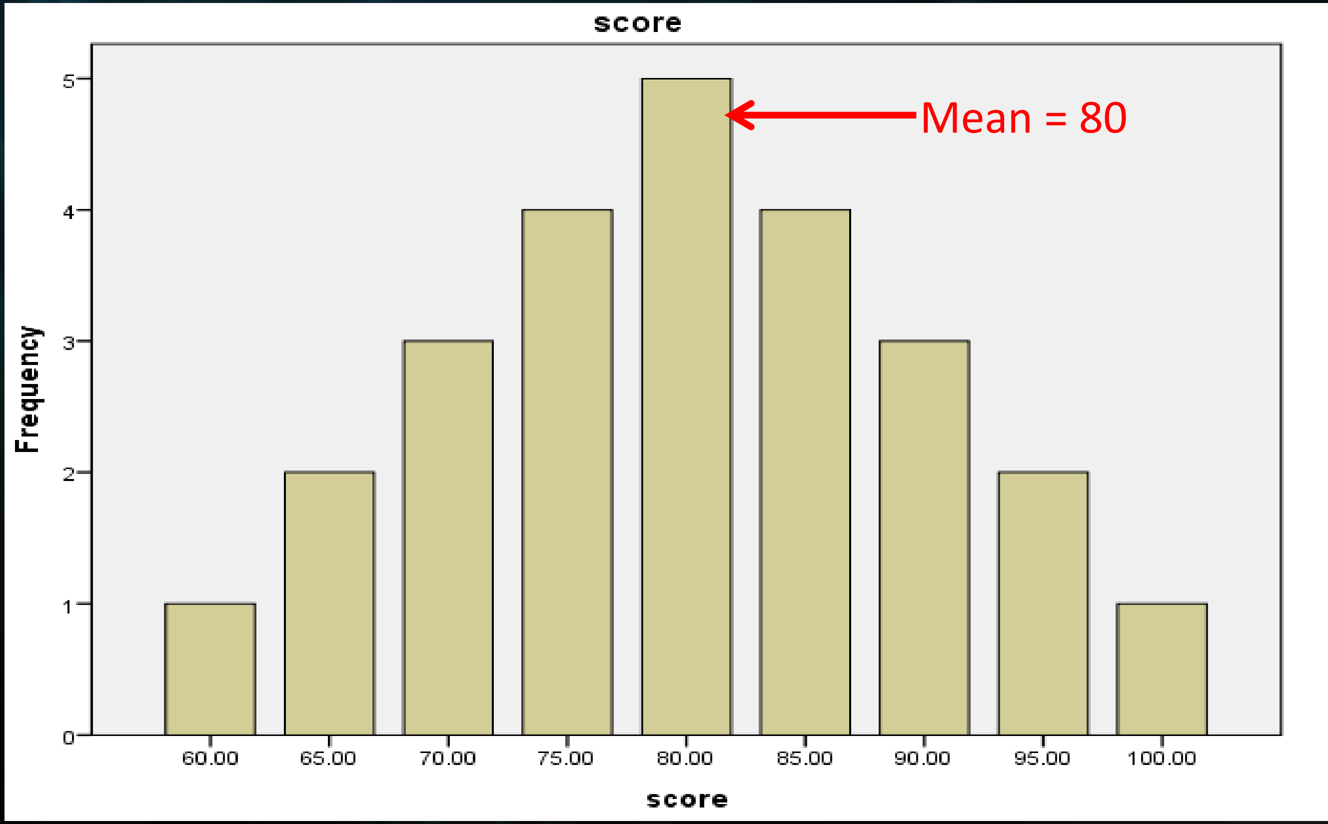
Score (X)	Frequency (f)	fX
60	1	60
65	2	130
70	3	210
75	4	300
80	5	400
85	4	340
90	3	270
95	2	190
100	1	100
Sum	25	2000

Descriptive statistics

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

Descriptive statistics

- Distribution of Example 1



Descriptive statistics

- 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.

Descriptive statistics

- 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. median = 6

Descriptive statistics

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

Descriptive statistics

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

Descriptive statistics

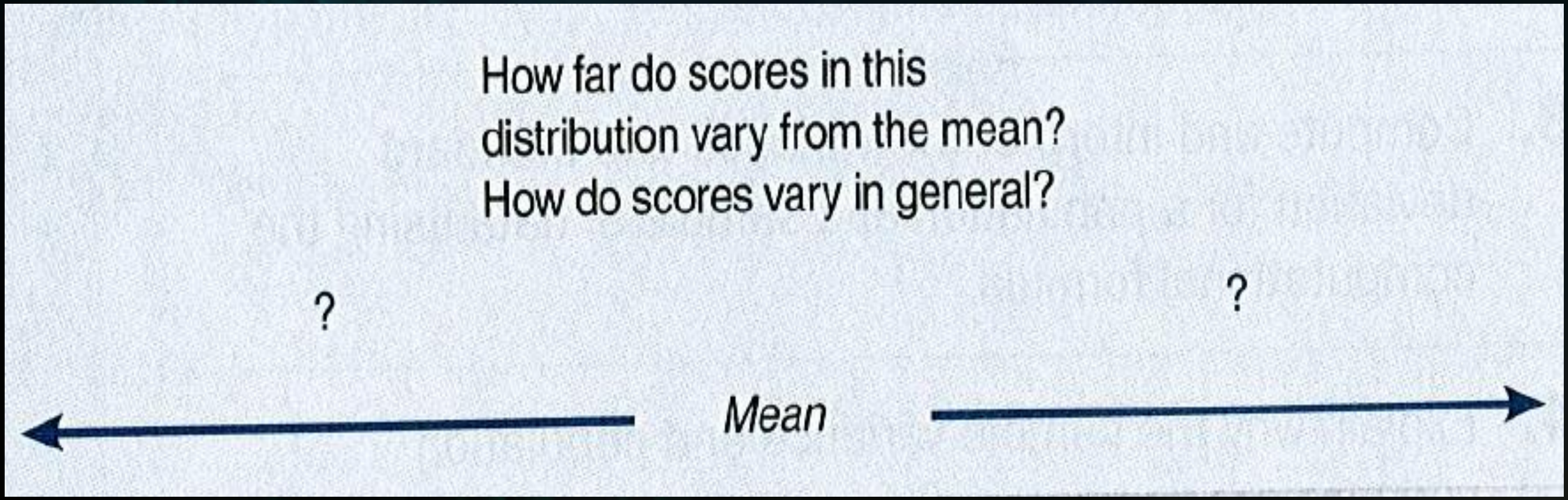
- Median = ?
- Position 18.5
- Which number is at position 18.5?
- Median = $45k

Descriptive statistics

- Mode
  - The value in a data set that occurs most often or most frequently.
  - Example: 2,3,3,3,4,4,4,4,7,7,8,8,8. Mode = 4

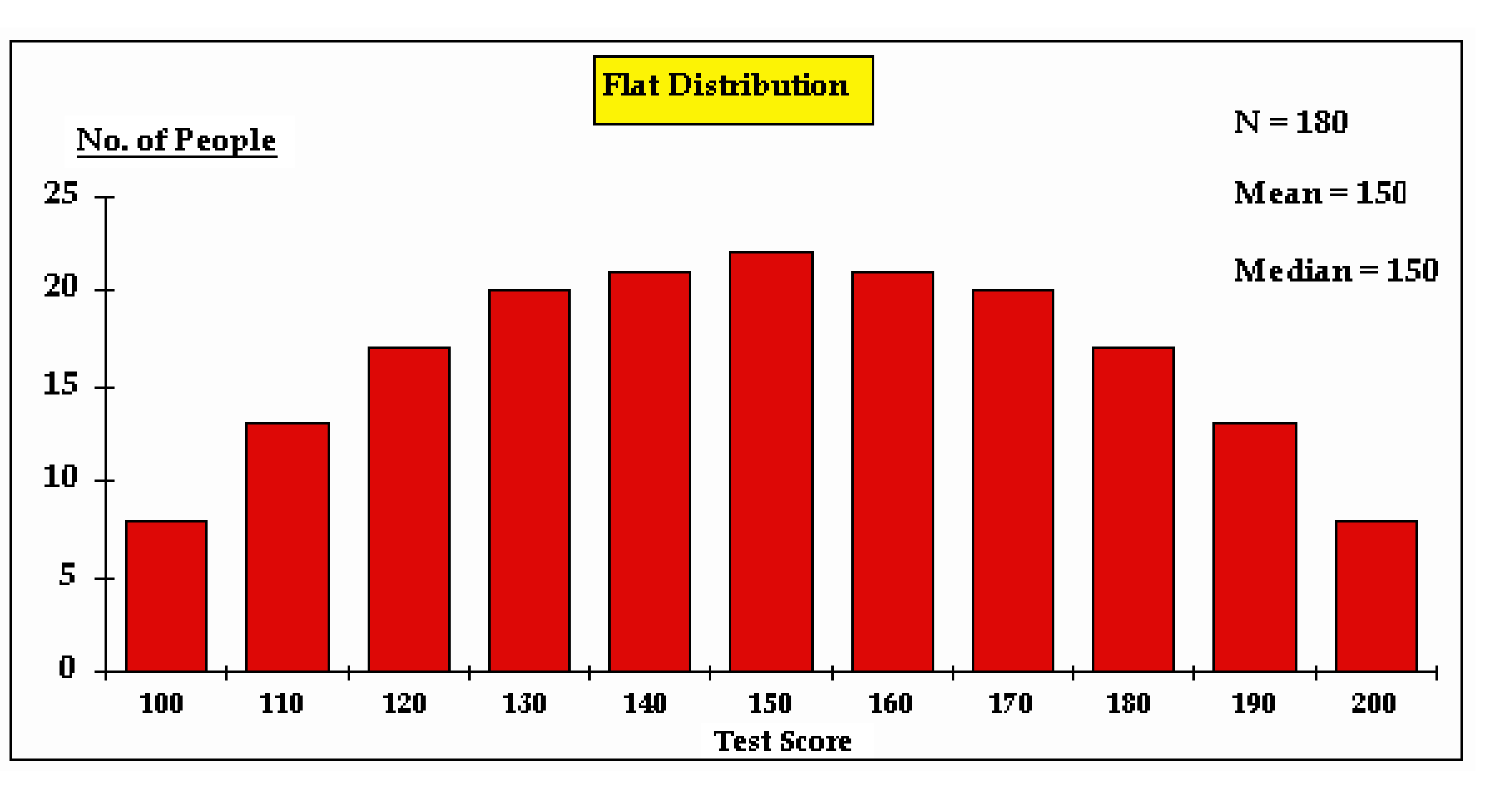
Descriptive statistics

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



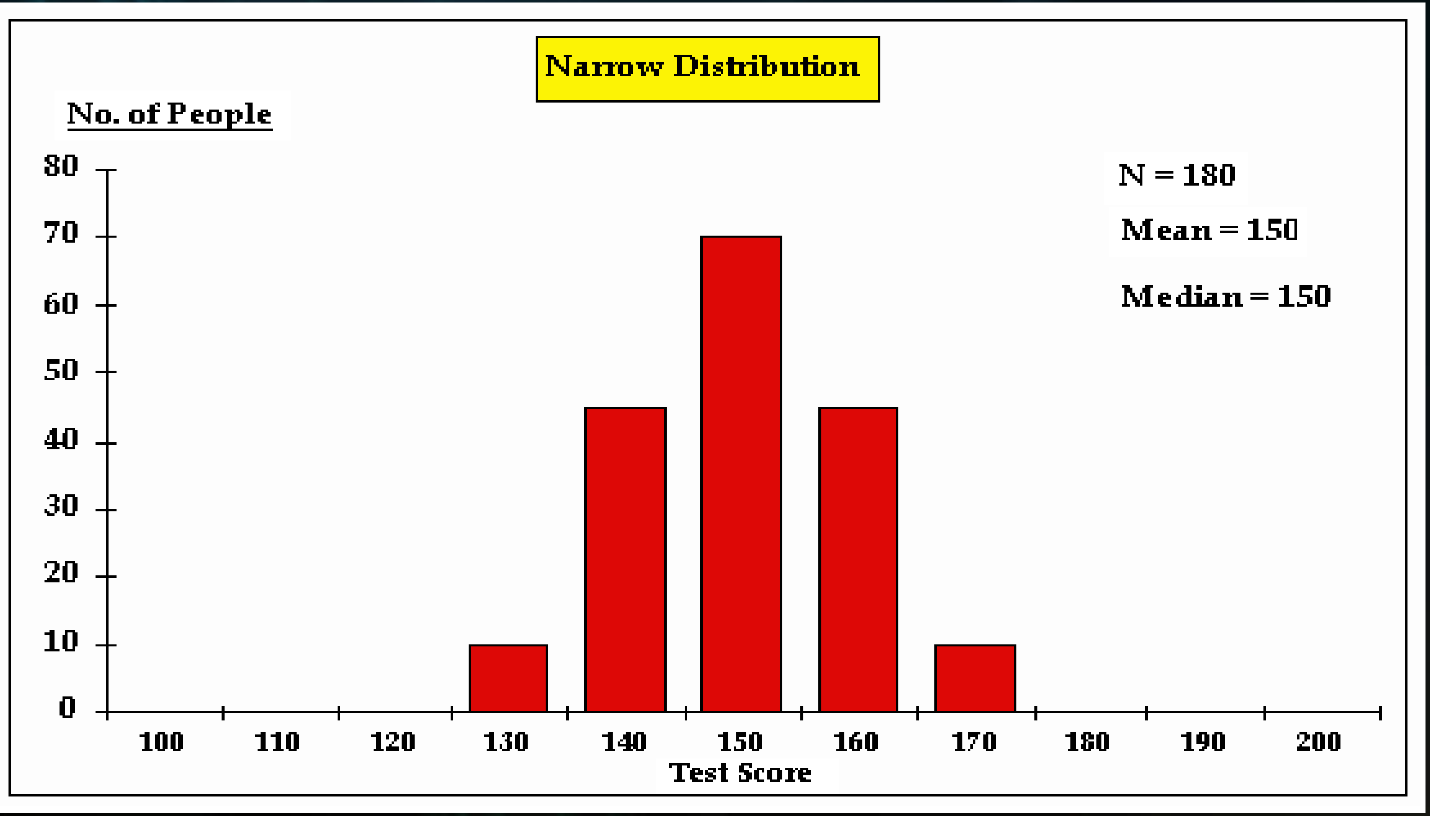
Descriptive statistics

- Compare different distributions



Descriptive statistics

- Compare different distributions



Descriptive statistics

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

Descriptive statistics

- Dispersion
  - Range
  - Variance
  - Standard deviation

Descriptive statistics

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

Descriptive statistics

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

Descriptive statistics

- 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.

Descriptive statistics

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

Descriptive statistics

- Degree of freedom
  - Example 1. we have five scores: 1, 2, 3, and two unknown scores: x and y. The mean of five values is equal to 3. So $x + y = 9$.
  - Example 2. we have five scores: 1, 2, and three unknown scores: x, y, and z. The mean of five values is equal to 3. $x + y + z = 12$.

Descriptive statistics

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

Descriptive statistics

- Example 3: calculate standard deviation

Scores (x)	Frequency(f)	$x - \bar{x}$ (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

Descriptive statistics

- 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.

Descriptive statistics

- 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.

Descriptive statistics

- 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)

Descriptive statistics

- Probability

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

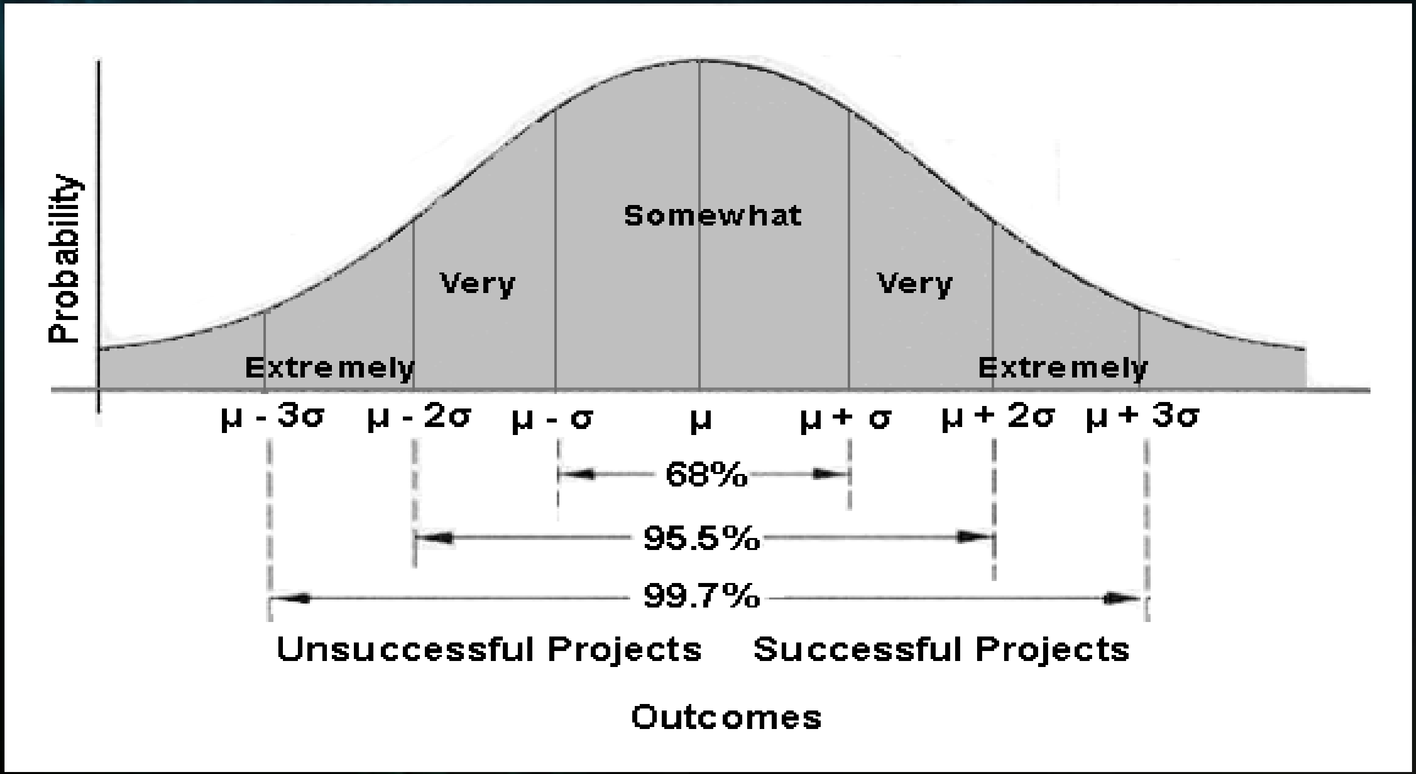
1. What is the probability of Fail? 4/10 =.4
2. What is the probability of Pass? 6/10 = .6
3. What is the probability of Fail among males? 3/5 = .6
4. What is the probability of Pass among females? 4/5 = .8

Descriptive statistics

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

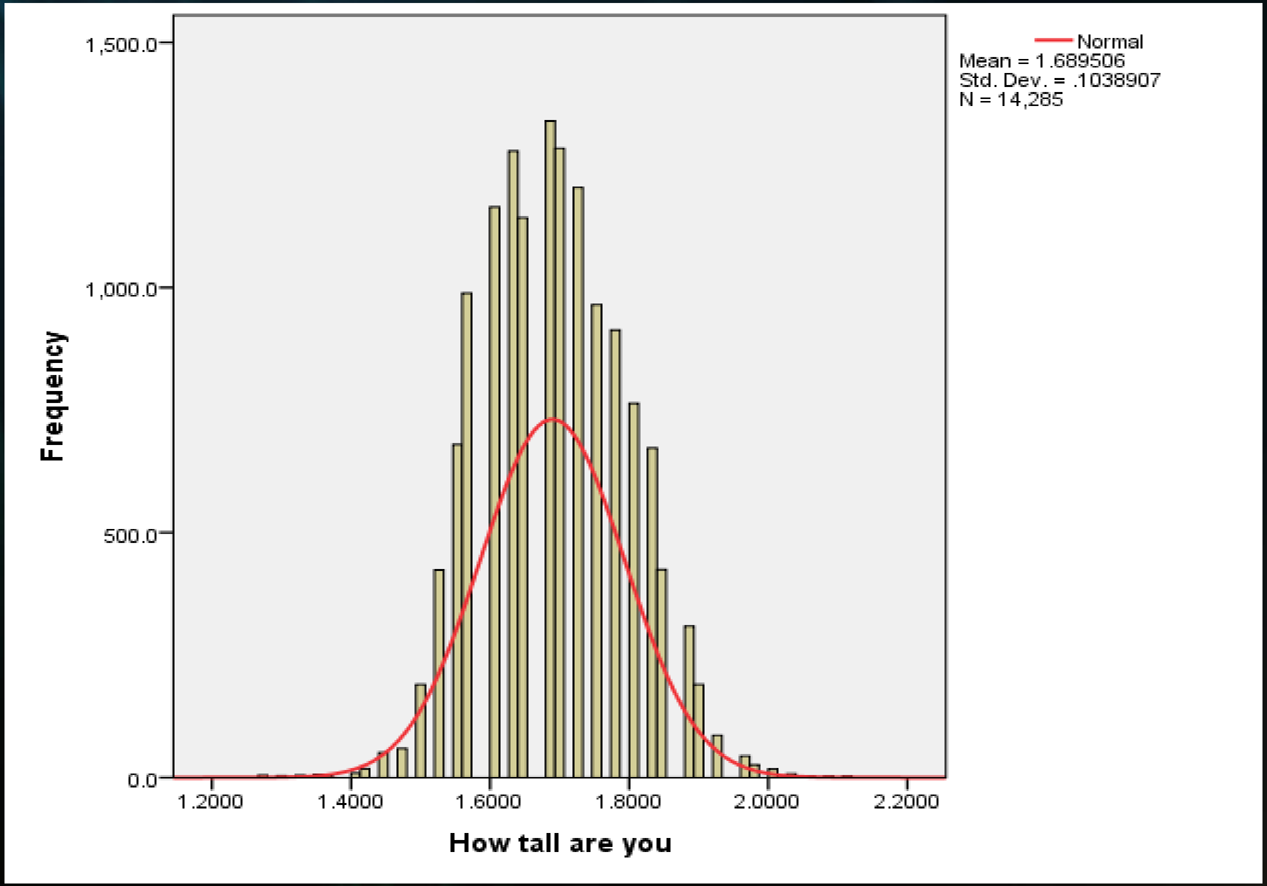
Descriptive statistics

- Normal curve



Descriptive statistics

- Normal curve



Descriptive statistics

- Characteristics of normal distribution
  - 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.

Descriptive statistics

- Characteristics of normal distribution
  - 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.

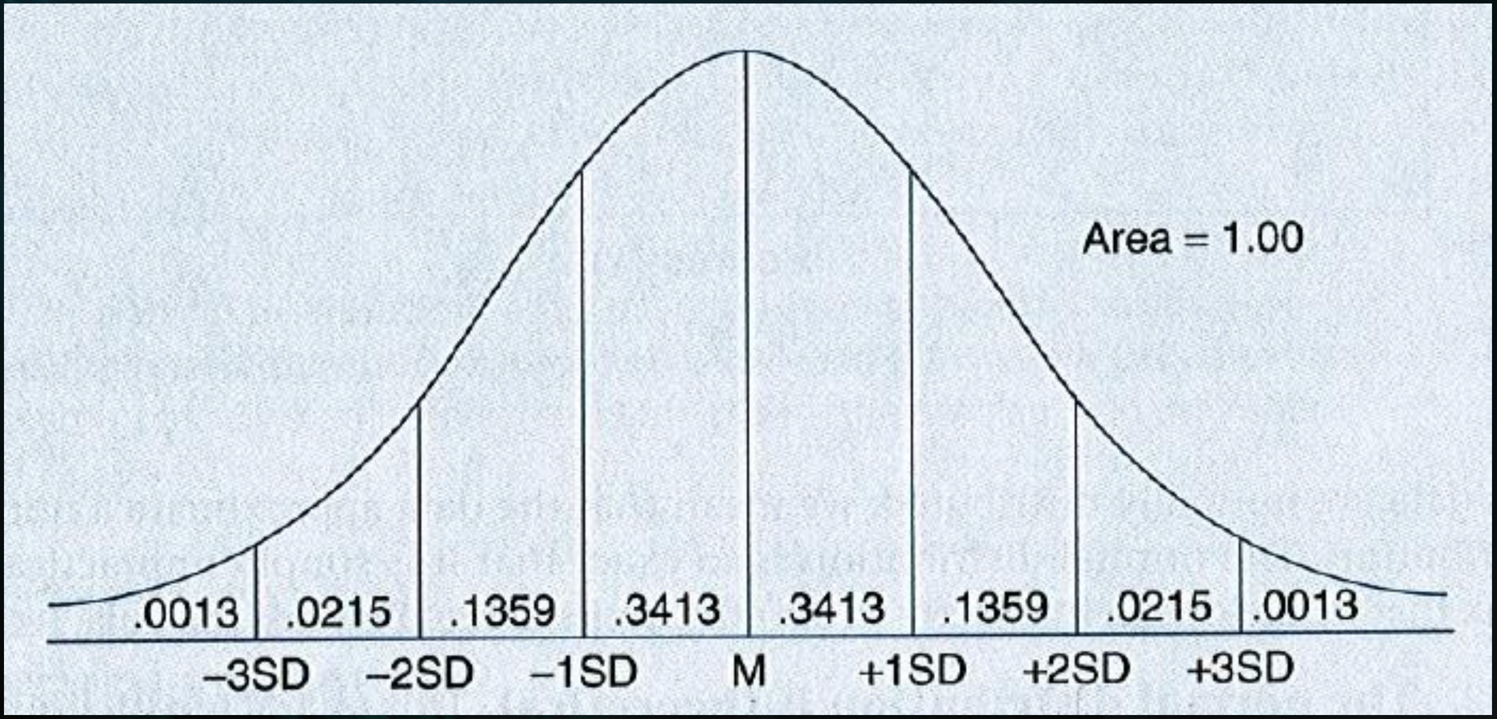
Descriptive statistics

- Characteristics of normal distribution
  - 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.

Descriptive statistics

- 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.

Descriptive statistics



Proportions of area under the normal curve

Descriptive statistics

- 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.

Descriptive statistics

- 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.

Descriptive statistics

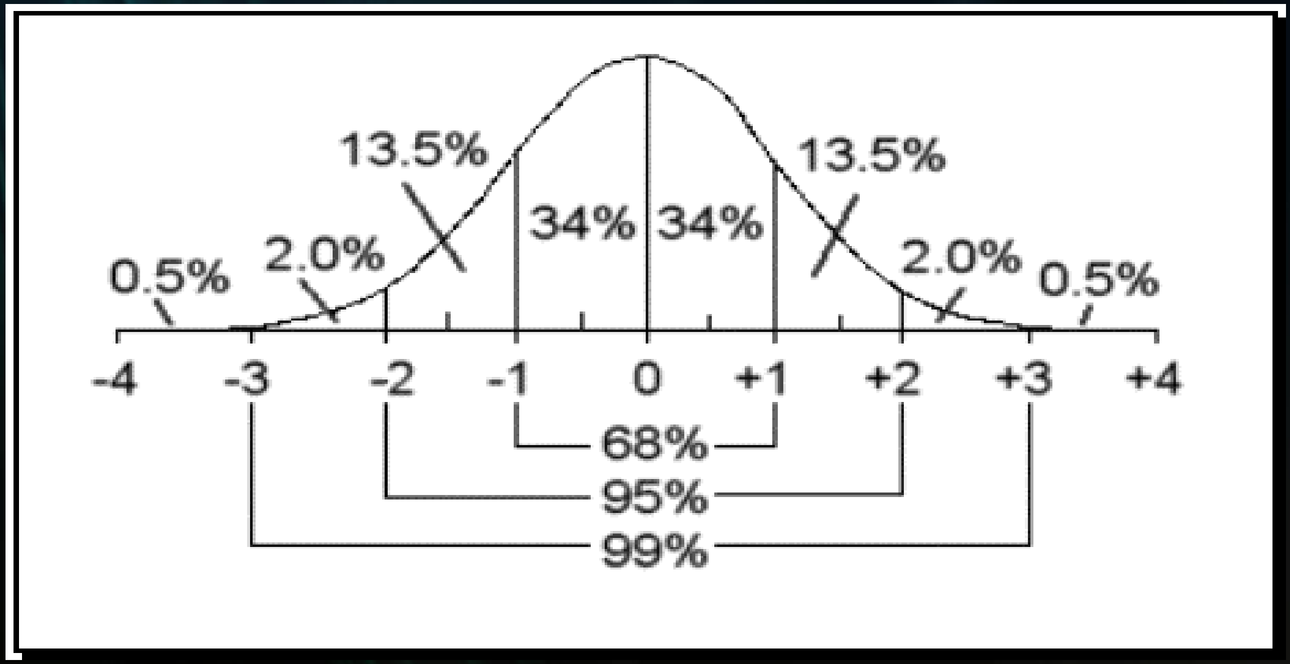
- In another word, we have 95% chance of selecting a score that is within 2 standard deviation of mean.
- Less than 5% scores are far from the mean (NOT normal scores).

Descriptive statistics

- 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

Descriptive statistics

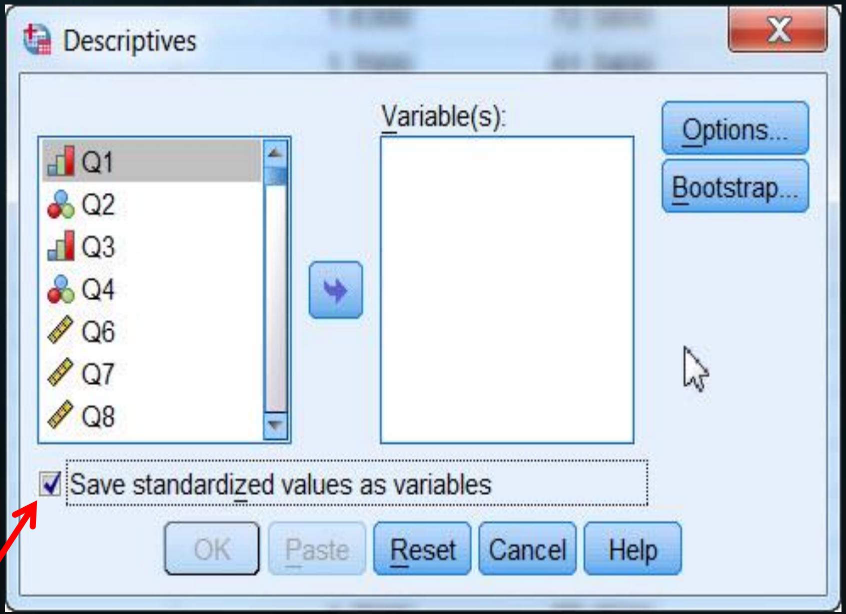
- Standard normal distribution or Z distribution



Descriptive statistics

- z transformation

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



X means individual value, M is mean and SD is standard deviation.

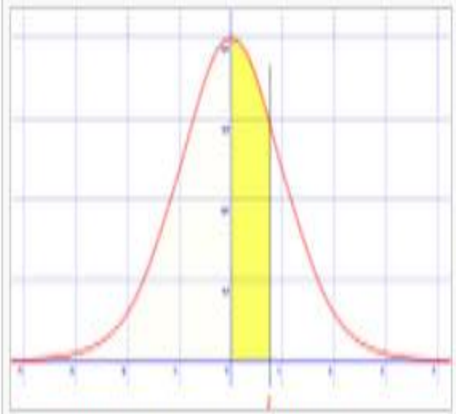
In SPSS, go to Analyze > Descriptive Statistics > Descriptives to get Z scores

Descriptive statistics

- Normal table/z table

Cumulative from mean (0 to Z) [edit source | edit beta]

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	1.33
0.0	0.00000	0.00399	0.00798	0.01197	0.01595	0.01994	0.02392	0.02790	0.03188	0.03586	
0.1	0.03983	0.04380	0.04776	0.05172	0.05567	0.05962	0.06356	0.06749	0.07142	0.07535	
0.2	0.07926	0.08317	0.08706	0.09095	0.09483	0.09871	0.10257	0.10642	0.11026	0.11409	
0.3	0.11791	0.12172	0.12552	0.12930	0.13307	0.13683	0.14058	0.14431	0.14803	0.15173	
0.4	0.15542	0.15910	0.16276	0.16640	0.17003	0.17364	0.17724	0.18082	0.18439	0.18793	
0.5	0.19146	0.19497	0.19847	0.20194	0.20540	0.20884	0.21226	0.21566	0.21904	0.22240	
0.6	0.22575	0.22907	0.23237	0.23565	0.23891	0.24215	0.24537	0.24857	0.25175	0.25490	
0.7	0.25804	0.26115	0.26424	0.26730	0.27035	0.27337	0.27637	0.27935	0.28230	0.28524	
0.8	0.28814	0.29103	0.29389	0.29673	0.29955	0.30234	0.30511	0.30785	0.31057	0.31327	
0.9	0.31594	0.31859	0.32121	0.32381	0.32639	0.32894	0.33147	0.33398	0.33646	0.33891	
1.0	0.34134	0.34375	0.34614	0.34849	0.35083	0.35314	0.35543	0.35769	0.35993	0.36214	
1.1	0.36433	0.36650	0.36864	0.37076	0.37286	0.37493	0.37698	0.37900	0.38100	0.38298	
1.2	0.38493	0.38686	0.38877	0.39065	0.39251	0.39435	0.39617	0.39796	0.39973	0.40147	
1.3	0.40320	0.40490	0.40658	0.40824	0.40988	0.41149	0.41308	0.41466	0.41621	0.41774	
1.4	0.41924	0.42073	0.42220	0.42364	0.42507	0.42647	0.42785	0.42922	0.43056	0.43189	
1.5	0.43319	0.43448	0.43574	0.43699	0.43822	0.43943	0.44062	0.44179	0.44295	0.44408	
1.6	0.44520	0.44630	0.44738	0.44845	0.44950	0.45053	0.45154	0.45254	0.45352	0.45449	
1.7	0.45543	0.45637	0.45728	0.45818	0.45907	0.45994	0.46080	0.46164	0.46246	0.46327	



This table gives a probability that a statistic is between 0 (mean) and Z.

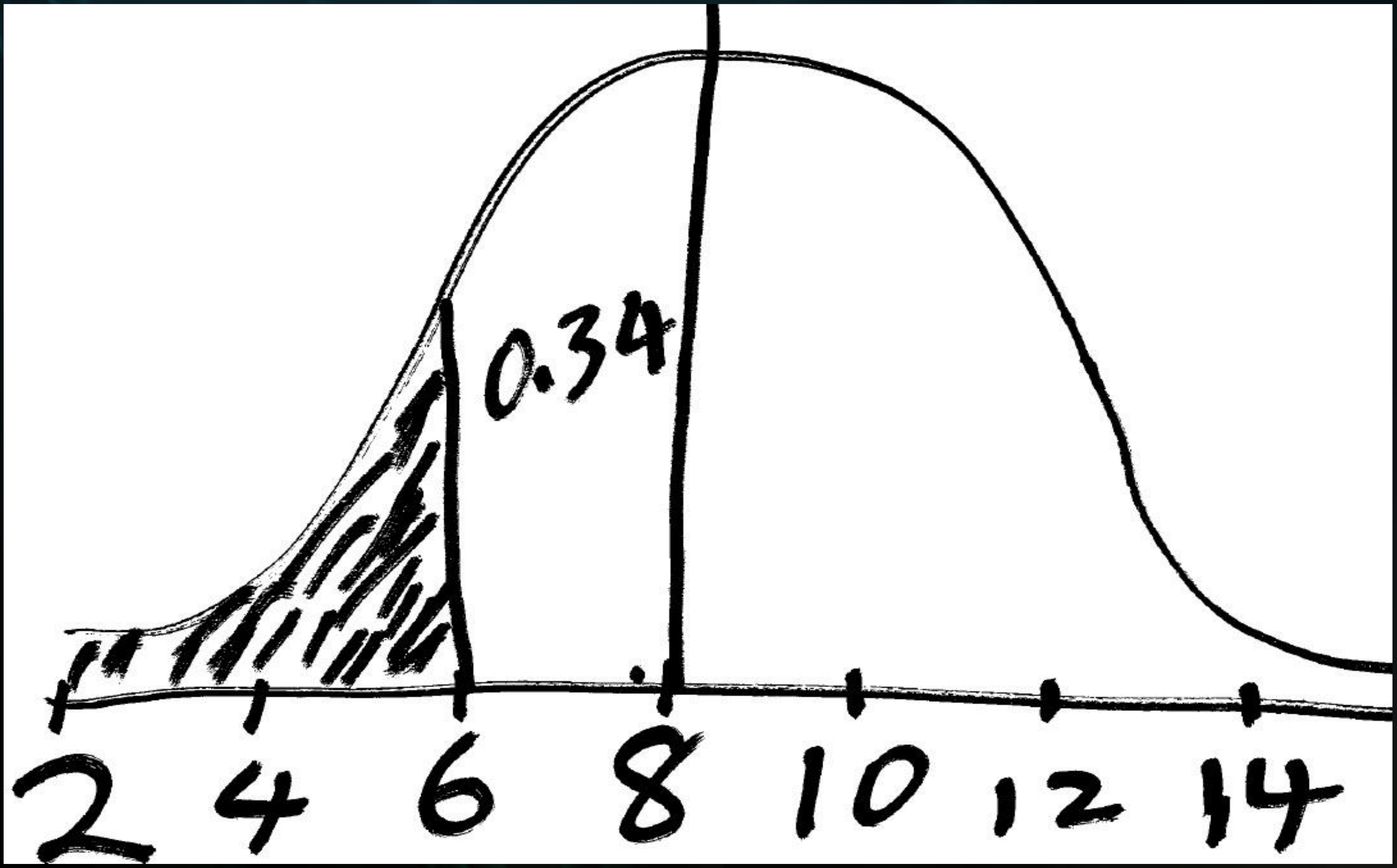
Descriptive statistics

- How to use z table?
  - Example: a sample of scores are approximately distributed normally with mean 8 and standard deviation 2. What is the probability of score lower than 6?

Descriptive statistics

- How to use z table?
  - Transform a raw score 6 into a z score
  - $z = (6-8)/2=-1$
  - Check the normal table p (probability) $= 0.5-0.34=0.16$
  - The probability of obtaining score less than 6 is 16%

Descriptive statistics



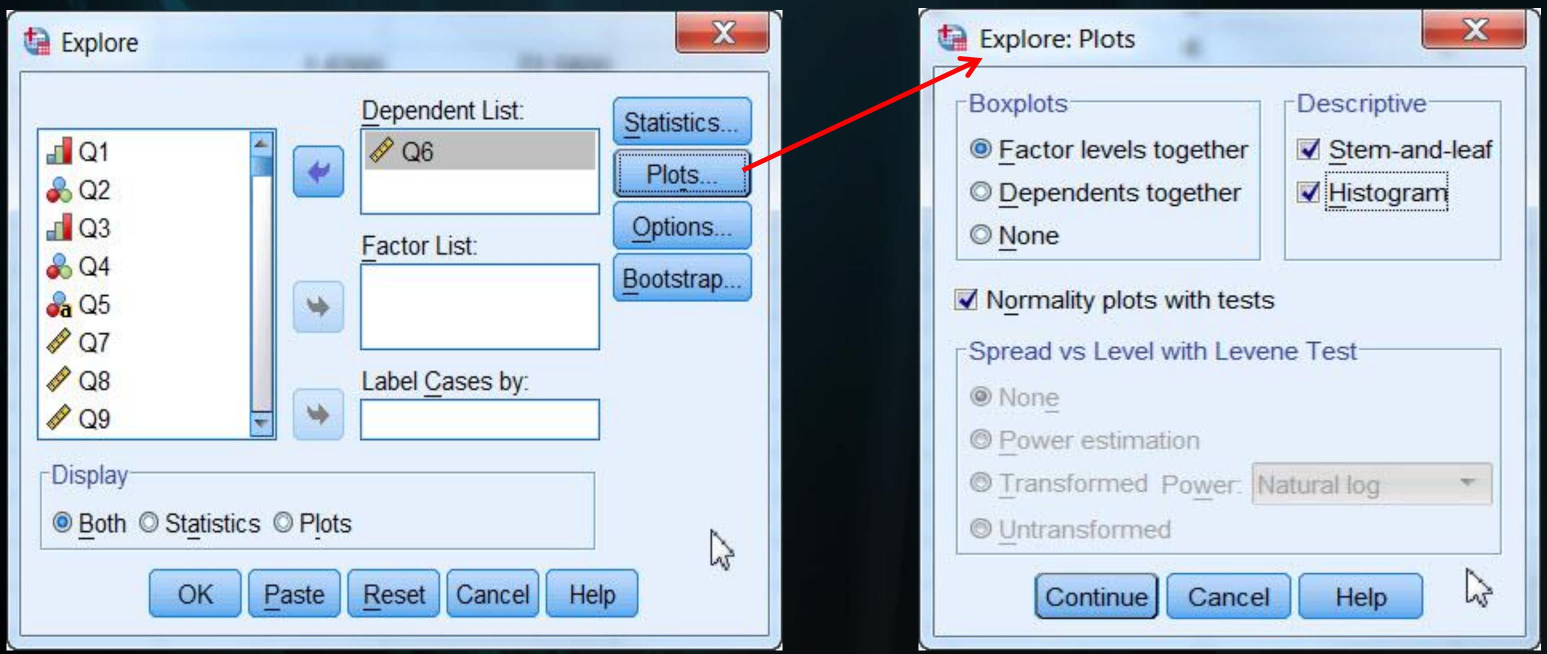
Descriptive statistics

- Descriptive statistics in SPSS
  - Frequencies
  - Descriptives
  - Explore

Descriptive statistics

- Exercise: use 2015 YRBSS data
  - Use Explore function to get descriptive statistics for Q6 (height)
  - Analyze > Descriptive Statistics > Explore

Descriptive statistics



Descriptive statistics

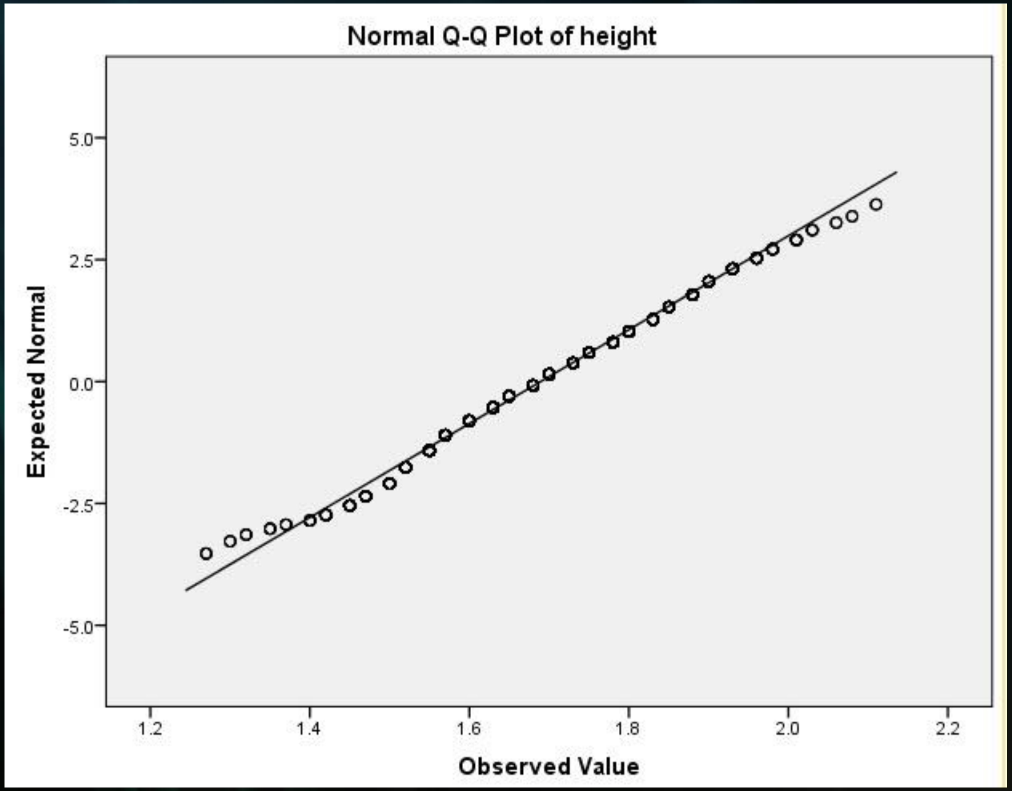
- SPSS output

Descriptives

			Statistic	Std. Error
Q6 height	Mean		1.689506	.0008692
	95% Confidence Interval for Mean	Lower Bound	1.687802	
		Upper Bound	1.691210	
	5% Trimmed Mean		1.688306	
	Median		1.680000	
	Variance		.011	
	Std. Deviation		.1038907	
	Minimum		1.2700	
	Maximum		2.1100	
	Range		.8400	
	Interquartile Range		.1500	
	Skewness		.150	.020
	Kurtosis		-.099	.041

Descriptive statistics

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

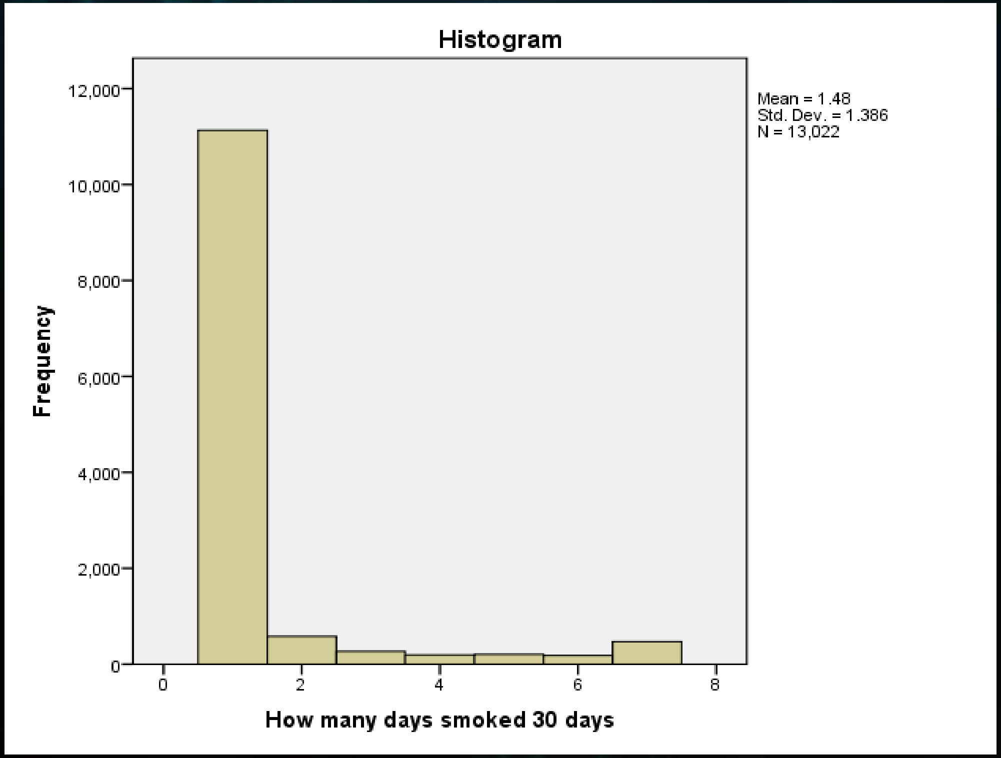


Graphs

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

Histogram

- Example: Q33 (how many days smoked during the last 30days)



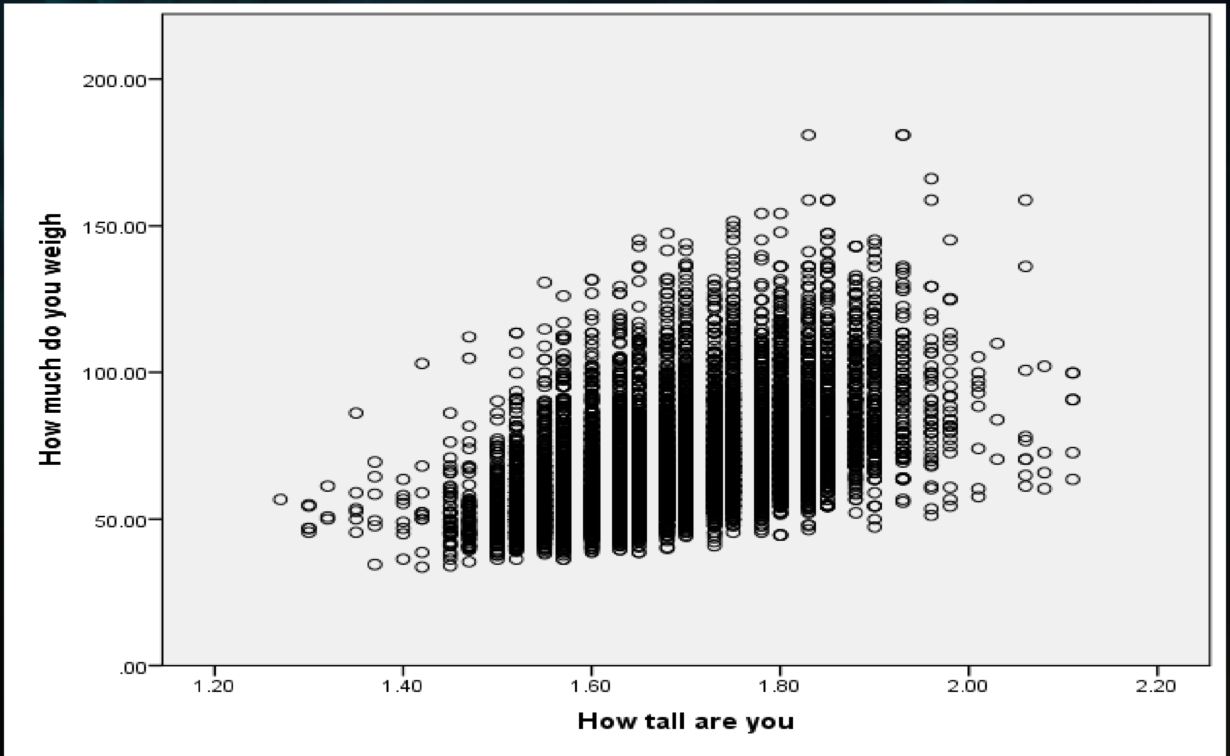
We use histogram to know the distribution of Q33.
Y axis represents frequency and X axis represents the responses.

Scatter Plot

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

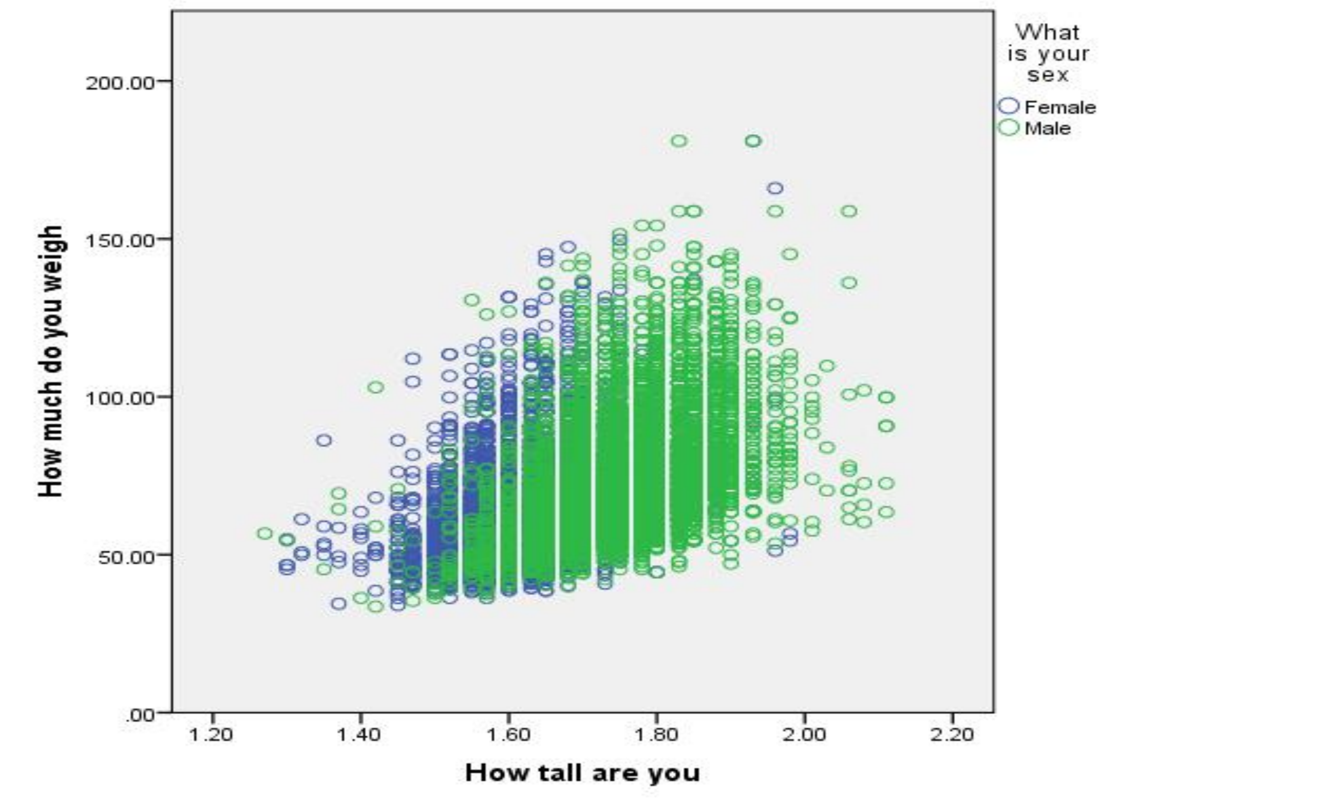
Scatter Plot

- 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

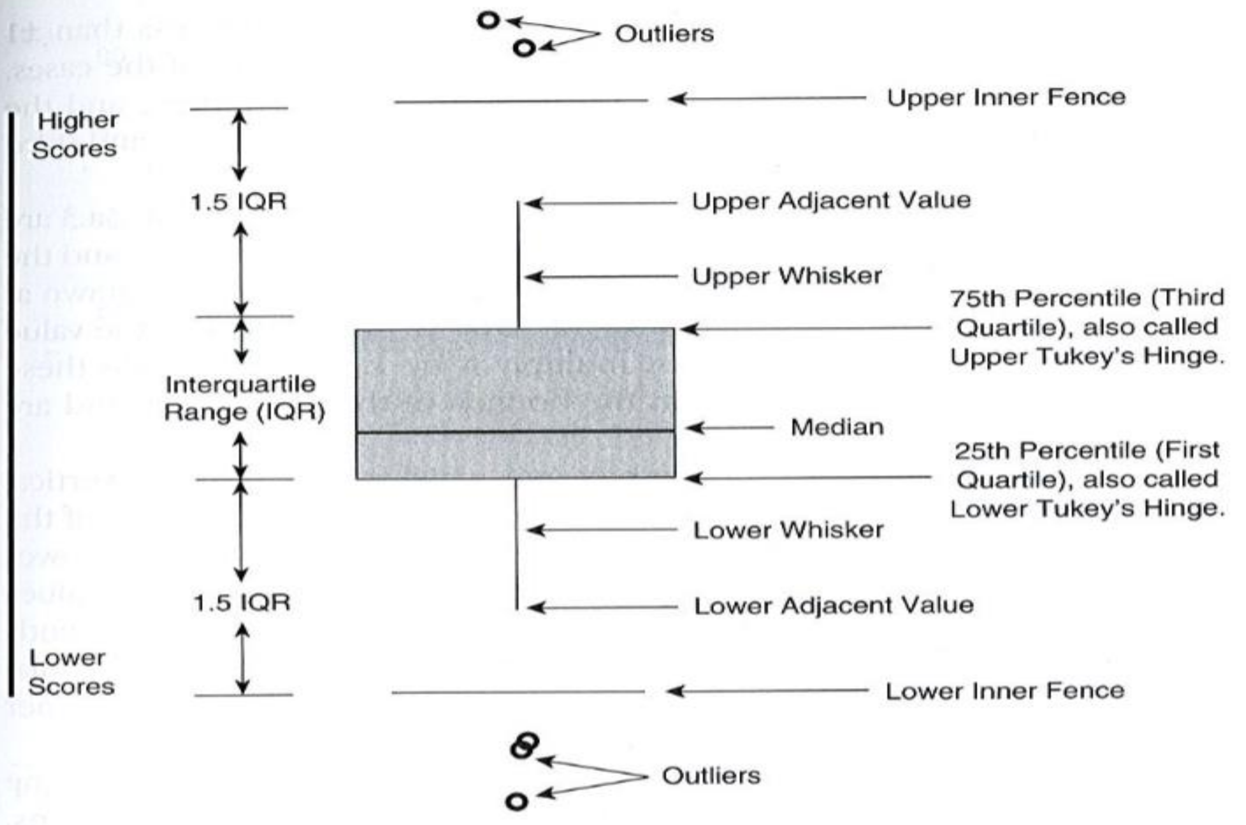
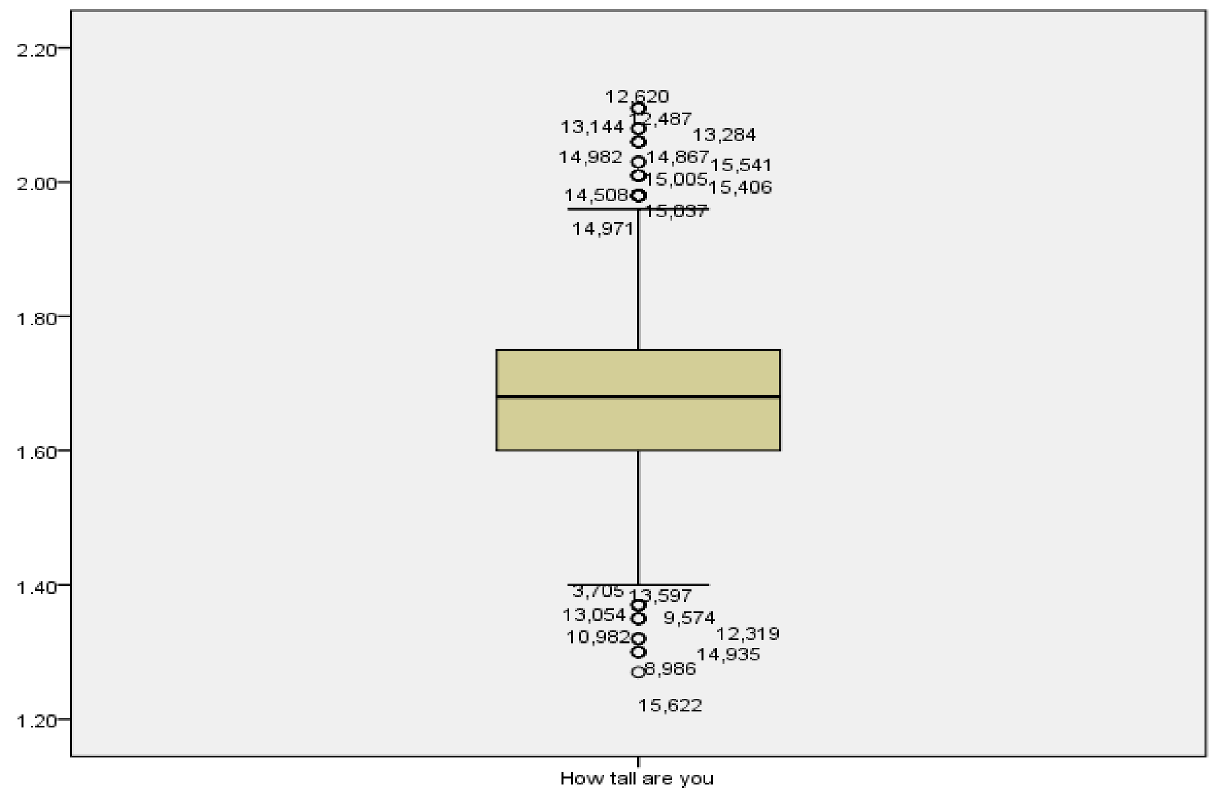


Figure 3a.3 The General Form of a Box and Whiskers Plot Based on Cohen's (1996) Description

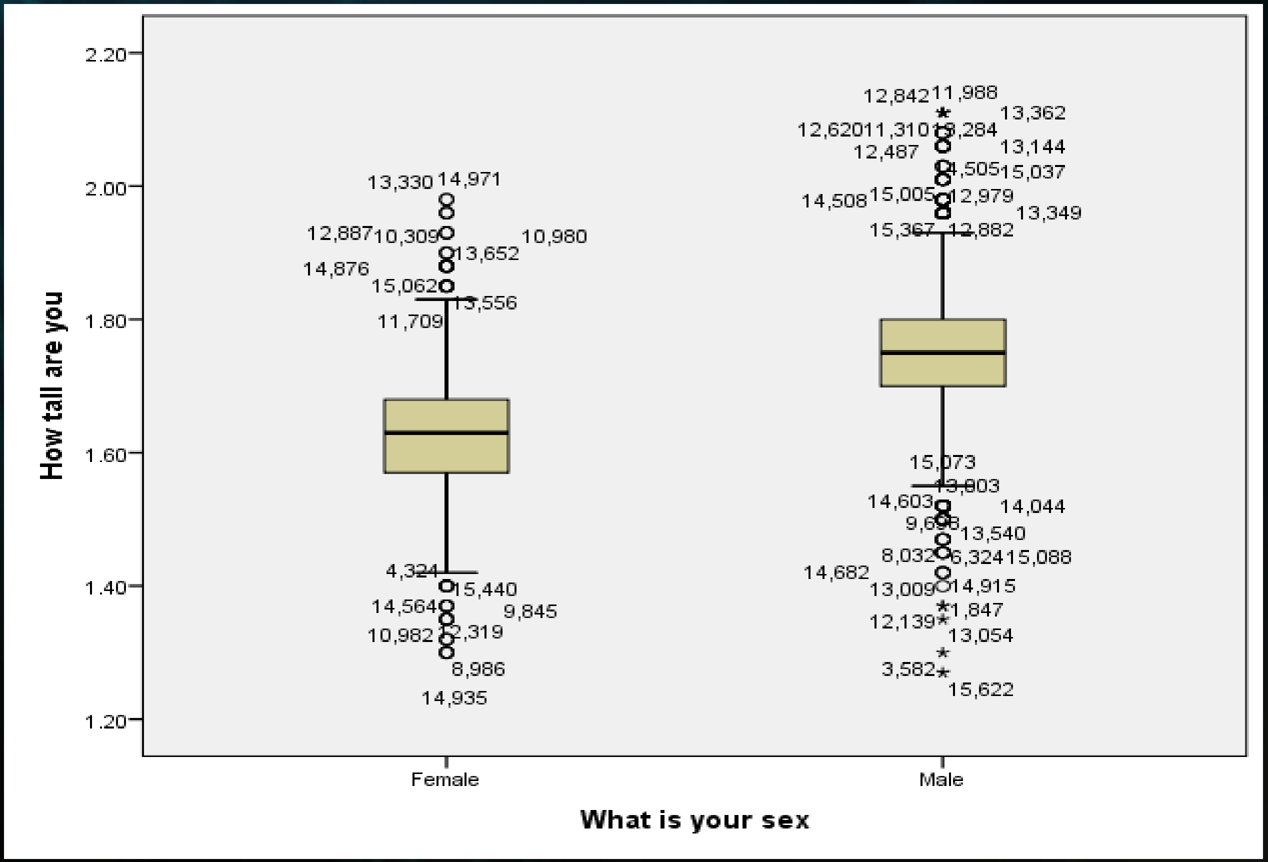
Box Plot

- Box plot of Q6 without Q2



Box Plot

- Box plot of Q6 by Q2

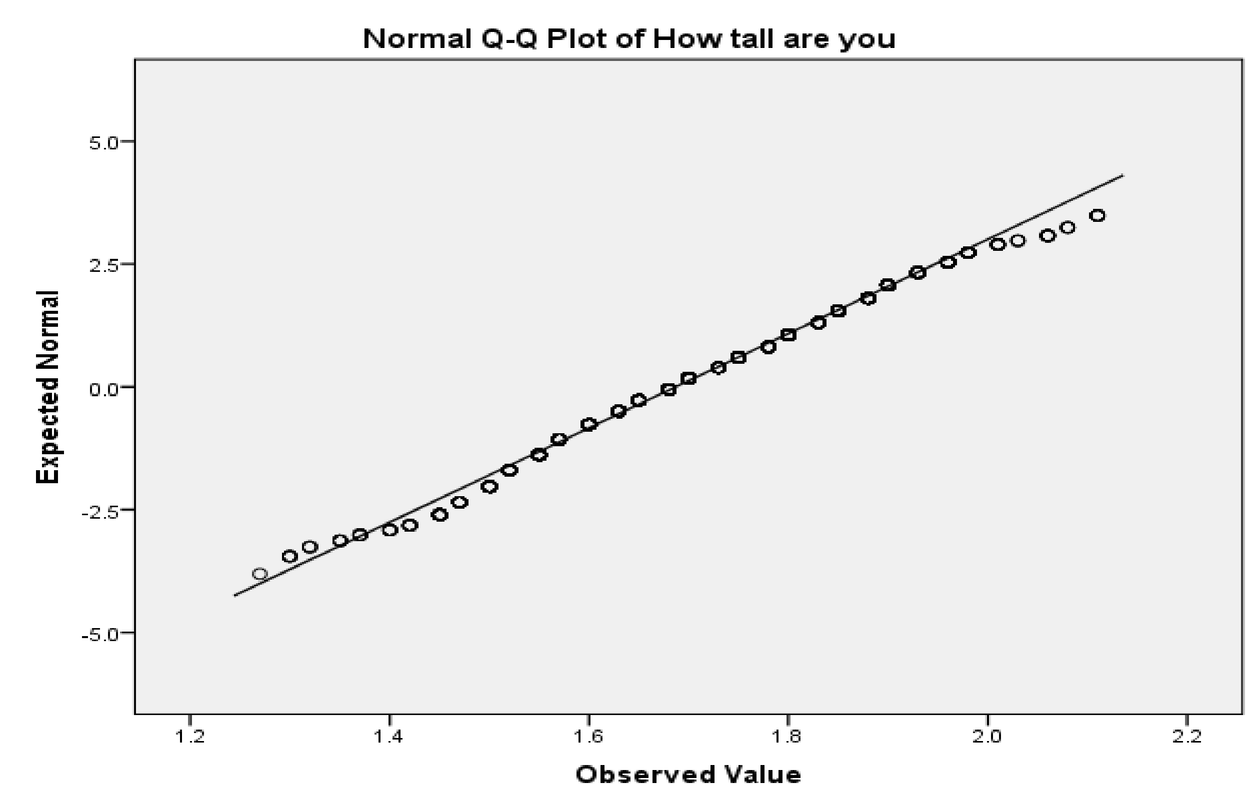


Normal Q-Q plot

- 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

- 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.