ADVANCED PHARMACEUTICAL BIOSTATISTICS

Dr. Mohammed S. Al-Lami

PhD Pharmaceutics

18 Sep. 2023

Aims of Class

- Provide the basic skills necessary to analyse the data gathered in your research project.

- Help you in the experimental design of your project.

- Become familiar with the statistics and graph plotting software packages available.

Week	Topic
1	Basic Definitions and Concepts
2-3	Descriptive statistics
4-5	Intro to inferential statistics, 2-sample tests
6-7	Analysis of variance (ANOVA)
8-9	Multi-factor ANOVA
10-11	Curve fitting
12-13	Non-parametric statistics

Statistics and Graph Plotting Software

Package	Type	Availability
Excel	General purpose statistics package.	On setting
Minitab	General purpose statistics package.	On Download
SPSS	General purpose statistics package.	On Download
Origin Pro	Scientific graph plotting with statistics. Good curve fitting.	???
GraphPad Prism	Scientific graph plotting with statistics. Good curve fitting. Easy to use statistics.	On Download

References

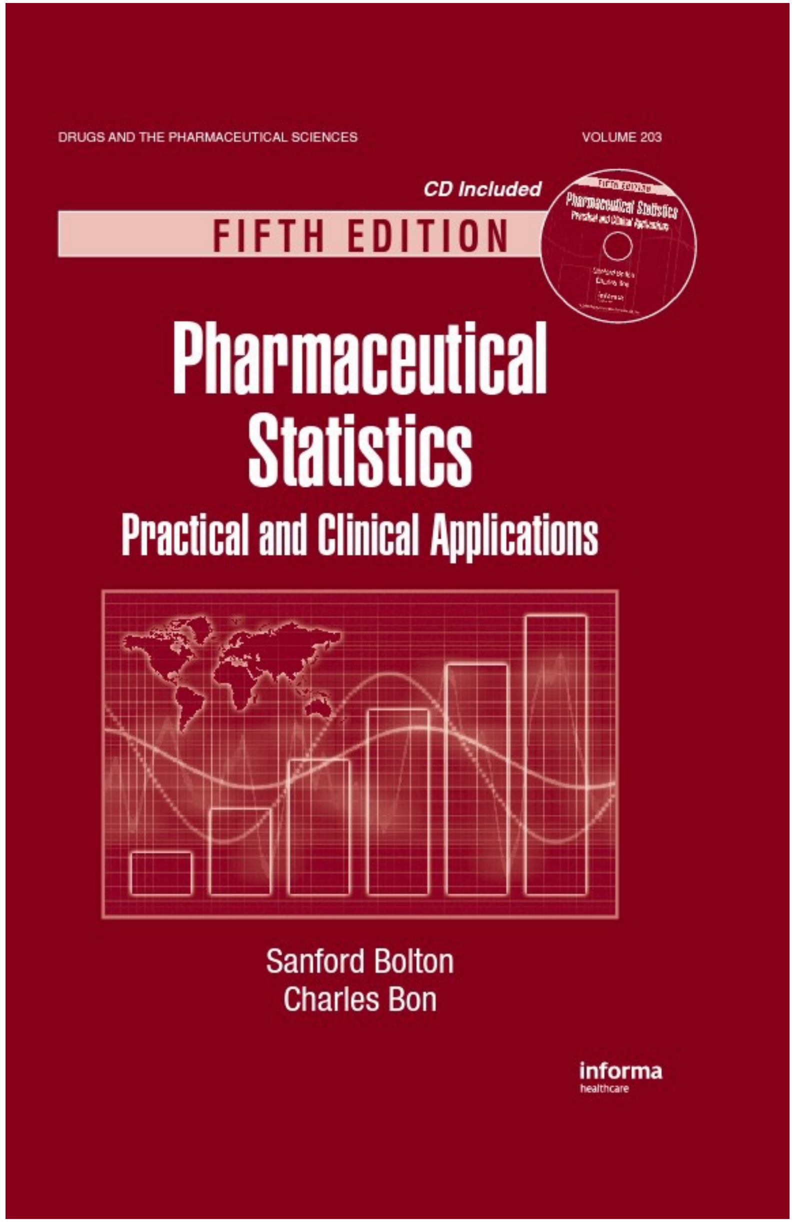
Pharmaceutical Statistics

Practical and Clinical Applications

5th Ed.

Sanford Bolton
Consultant
Tucson, Arizona, USA

Charles Bon
Biostudy Solutions, LLC
Wilmington, North Carolina, USA



- **Between subject variability**
  - No two living creatures are exactly the same; e.g. height, weight, blood pressure
- **Within subject variability**
  - Repeated measurement gives different values; e.g. blood pressure, heart rate
- **Measurement error**
  - Instrument readings vary due to instrument noise

Biological Variability: Why you need Statistics

- It is difficult to draw conclusions from single measurements from biological subjects
- Because …
  - The subject chosen may not be typical of the population as a whole.
  - The measurement may be unusually high or low for that subject on the day of the measurement.
  - There may be an unknown degree of measurement error
- It is therefore essential to make **repeated measurements** on more than one subject.
- Statistics summarise the results of multiple measurements and allow conclusions to be drawn.

Statistics

- **Descriptive Statistics**
  - Statistical measurements that summarise a data set
  - **Measures of 'central tendency'**
    Values that are representative of the population
    - Mean, median, mode
  - **Measures of variability**
    - Range, percentile, standard deviation, standard error

- **Inferential Statistics**
  - Statistical Measurements that allow conclusions to be drawn
    - T Test, Analysis of Variance (ANOVA)

Populations, Variables & Data

- **Population**
  - A group of people or things with a measurable characteristic in common
- **Variables**
  - a measurable factor, characteristic, or attribute of an individual or a system
- **Data**
  - the raw facts (numbers or words) that come from the measurement of a variable.

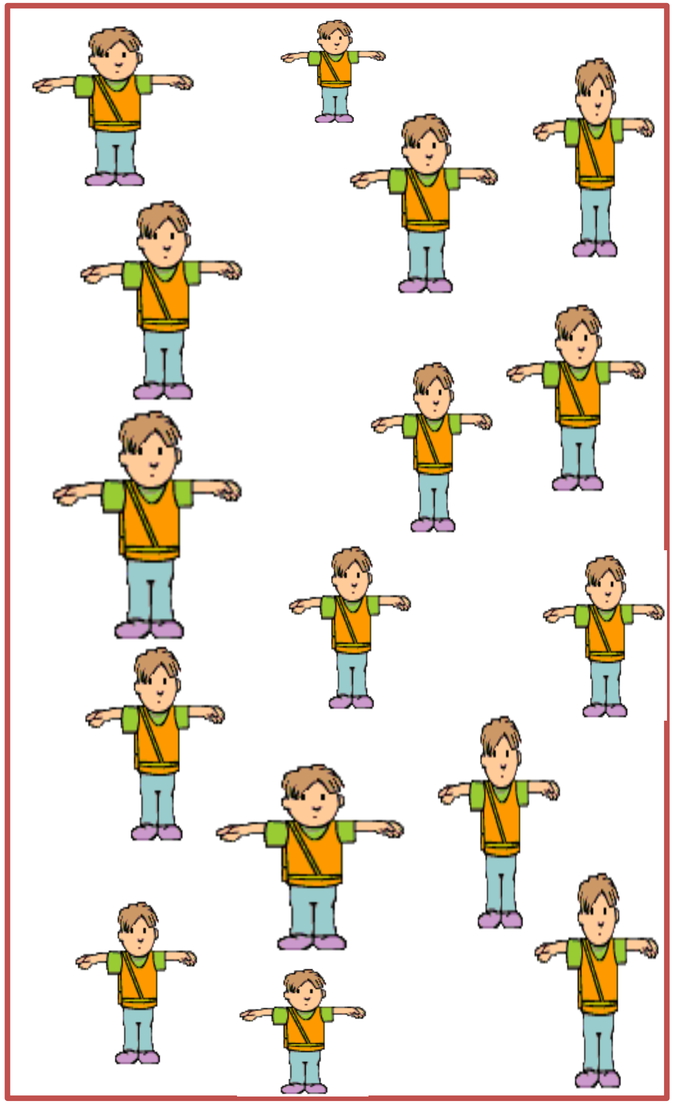
Nominal	**Non-quantitative** classification into 2 or more exclusive categories. e.g. Male/female, Smoker/Non-smoker
Ordinal	**Non-quantitative** classification into rank order, 1,2,3,4,… e.g. 0= 0=Strongly disagree, 1=Disagree, 2=Agree, 3=Strongly Agree
Interval	**Quantitative** scale of equal unit intervals without an absolute zero point Fahrenheit & Centigrade temperature scales
Ratio	**Quantitative** scale of equal unit intervals with an absolute zero point, e.g. Kelvin (absolute) temperature scale, height

The Experimental Process

Formulate Question
↓
Design Experiment to Answer It
↓
Perform Experiment
↓
Tabulate Results
↓
Analyse Results (Statistics, T-tests)
↓
Present Results (Bar charts, Scatter Graph)
↓
Think???
↓
Write Discussion/Conclusions

Example: The height of male students

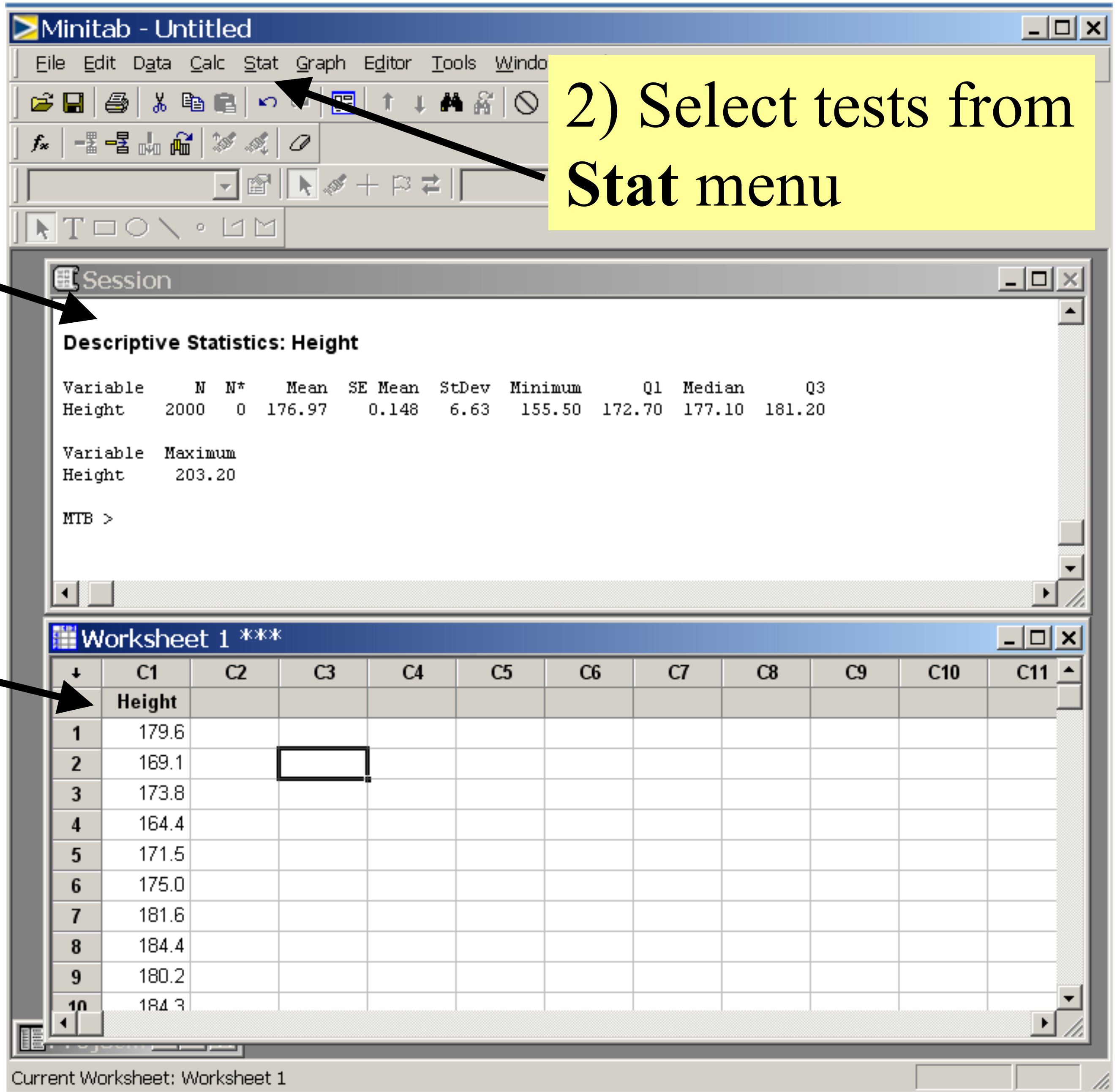
- **Question**
  - What is the average height of young male adult?
- **Population studied**
  - First year students (age 19-20 years)
  - 2000 male
- **Variable**
  - Height (cm)



Using Minitab

3) Results displayed in **Session** window

2) Select tests from **Stat** menu



1) Enter data into **Worksheet** columns

Raw Data

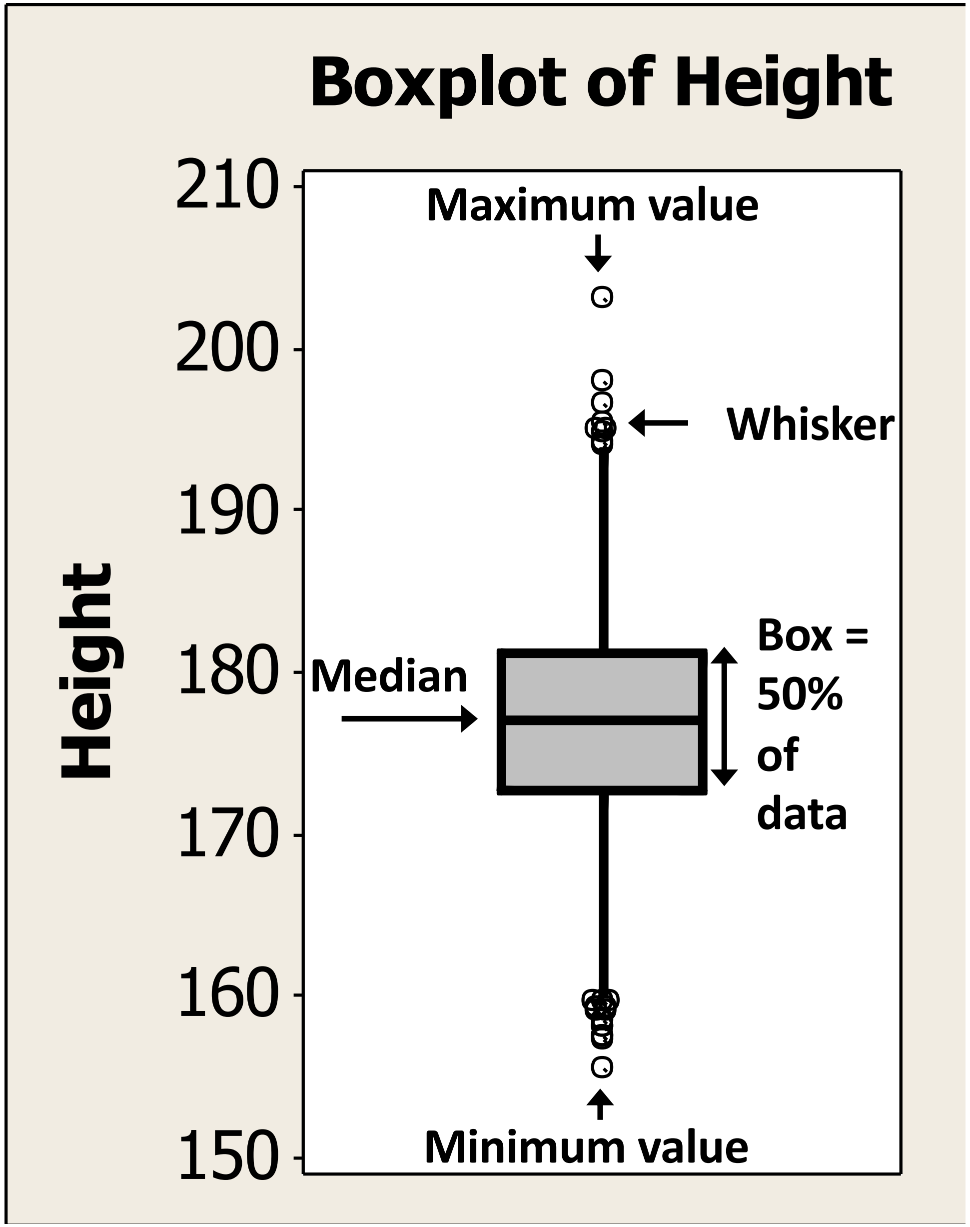
- Height (cm) of 200 male first year students

179.6	172.3	176	177.6	174.1	184.5	175.8	179.9
169.1	180.6	176.9	180	172.8	181.2	172.7	184.4
173.8	183	179.5	174.9	189.7	177	183.1	176.6
164.4	179.3	178.8	188.4	174.9	172.9	164.4	181
171.5	183	182.3	168.7	175.4	173	179.8	175.5
175	177	174.8	171.5	165.4	180.1	175.9	177.5
181.6	173.2	178.7	179.5	178.2	176.1	182	184
184.4	180.3	183.6	180.6	179.6	178.3	175	178.4
180.2	176.1	172	179.8	176.8	183.9	174.4	168.2
184.3	180.3	174.6	164.9	169.3	178.3	169.6	165.1
170.8	172.4	172	183.3	173.2	172.4	189.6	180.9
175.6	183	177.3	170	181.5	181.6	181.5	175.3
182.3	164.4	174.9	178	174.9	170.8	185.9	172.4
176.6	181.8	173.3	167.8	177	180.1	181.8	177.7
181.4	162.4	195.1	180.7	179.6	193.6	171.8	175.7
176.9	180.7	173.8	175.6	182.4	175.7	172.7	196.7
177	182.4	181.4	180.3	183.3	178.7	180.4	178.4
177.1	174.1	178.1	178.6	181.7	171.5	174.9	168.4
184.9	185.5	175.7	172.8	175.5	181.2	157.3	180.3
183.5	171.4	185	170.1	182.5	172.7	182.2	184.5
178.8	182.1	190.4	189.1	182.3	177.1	165.6	173.4
188.1	164.3	188.3	181.4	186.9	184	176.3	180.8
179.7	173.3	173.6	180.5	173.8	174.2	166.8	193.9
183.8	183.5	178.8	174.7	185.5	176.1	185.9	177

Visual Display of Data: Box Plot

- Summarises the range of values within the data set
- Median (line)
- Range of 50% of data (box)
- Highest and lowest data values (indicated by either whiskers or outlier points)
- Outliers: Individual values which are more than box lengths away from the edges of the box.

NOTE! Looking at a box plot of your data is a good way to see whether there are any unusual values/errors in it.

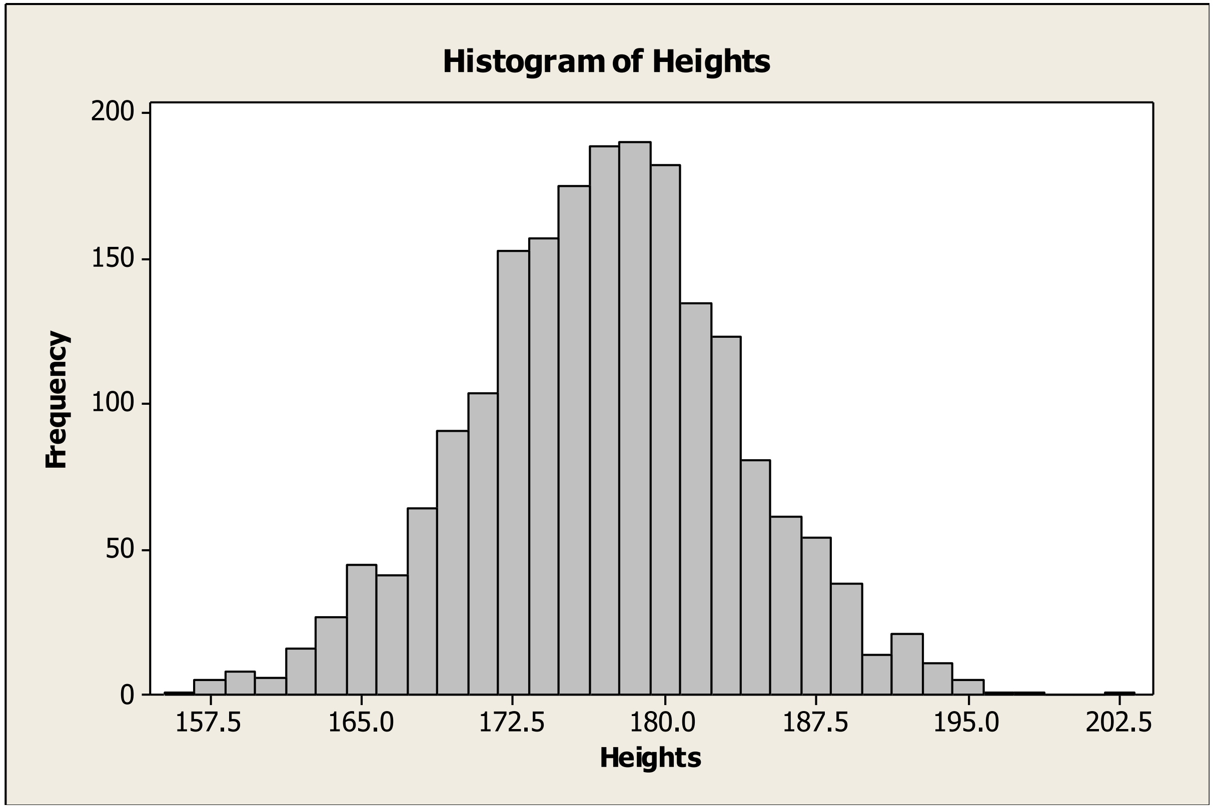


Minitab: Graph / Box Plot / Simple

Visual Display of Data: Frequency Distribution

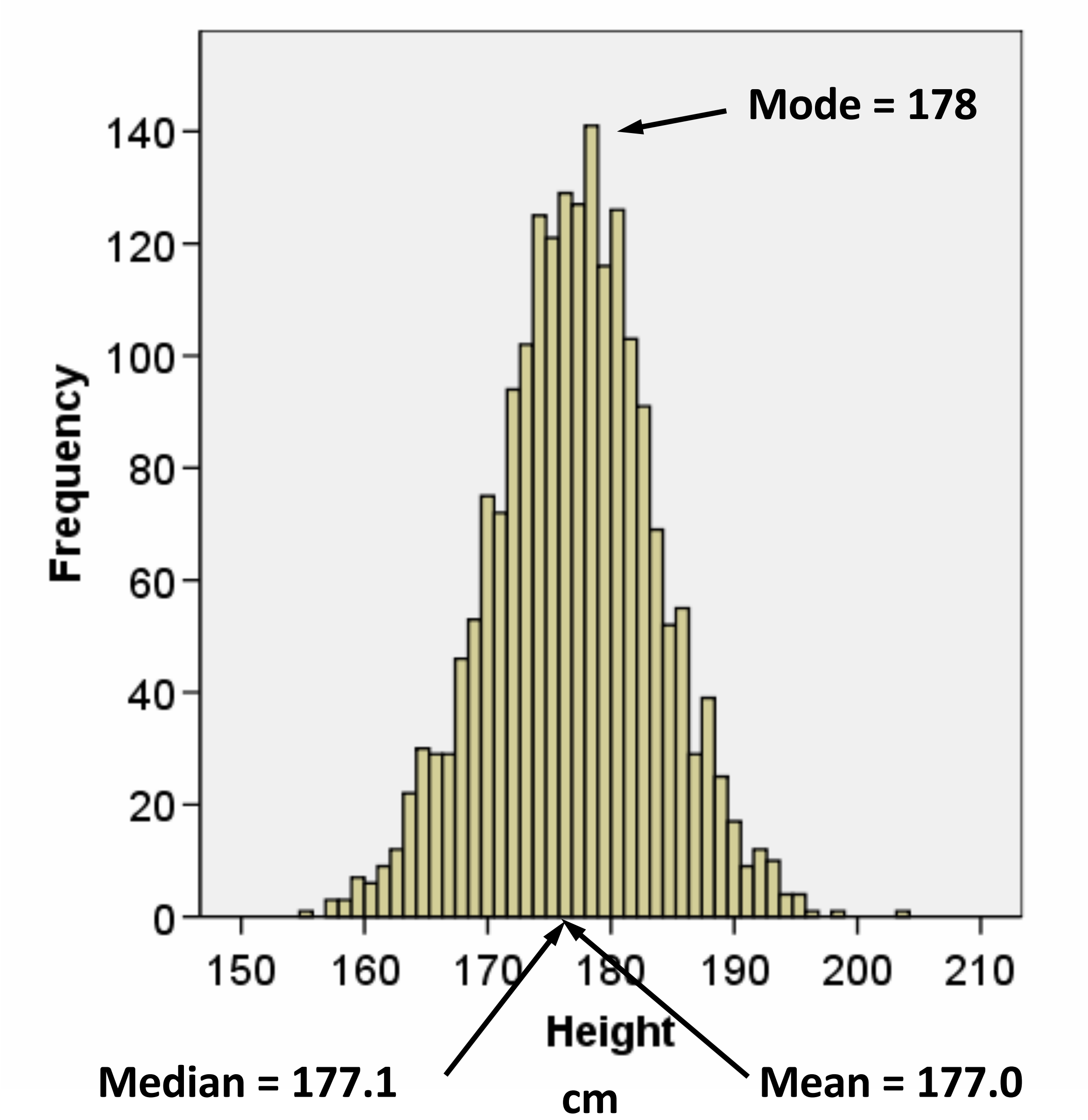
- A plot of the frequency with which data points have specific values or fall within a specific range of values.
- The range of data values contained within a data set is sub-divided into a series of equal sized bins and the number of data points falling into each bin is counted.

Frequency histogram showing distribution of male student height data.



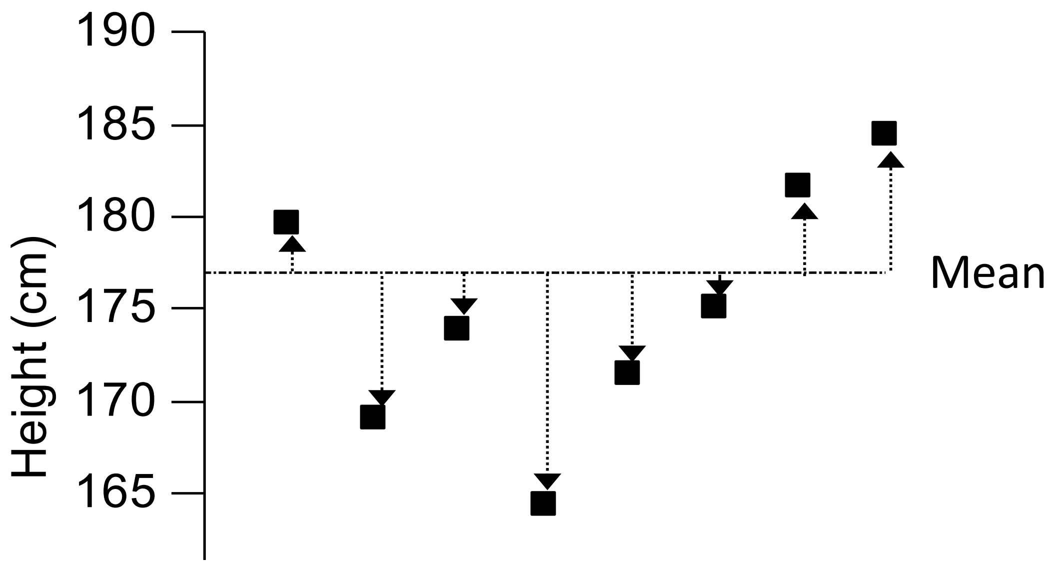
Minitab: Graph / Histogram / Simple

- **Mean**
  - Arithmetic average
- **Median**
  - Middle value when data is ranked into ascending order
- **Mode**
  - Most common value or histogram bin value



- Variance
  - Average of the squares of the differences between each data point and the mean

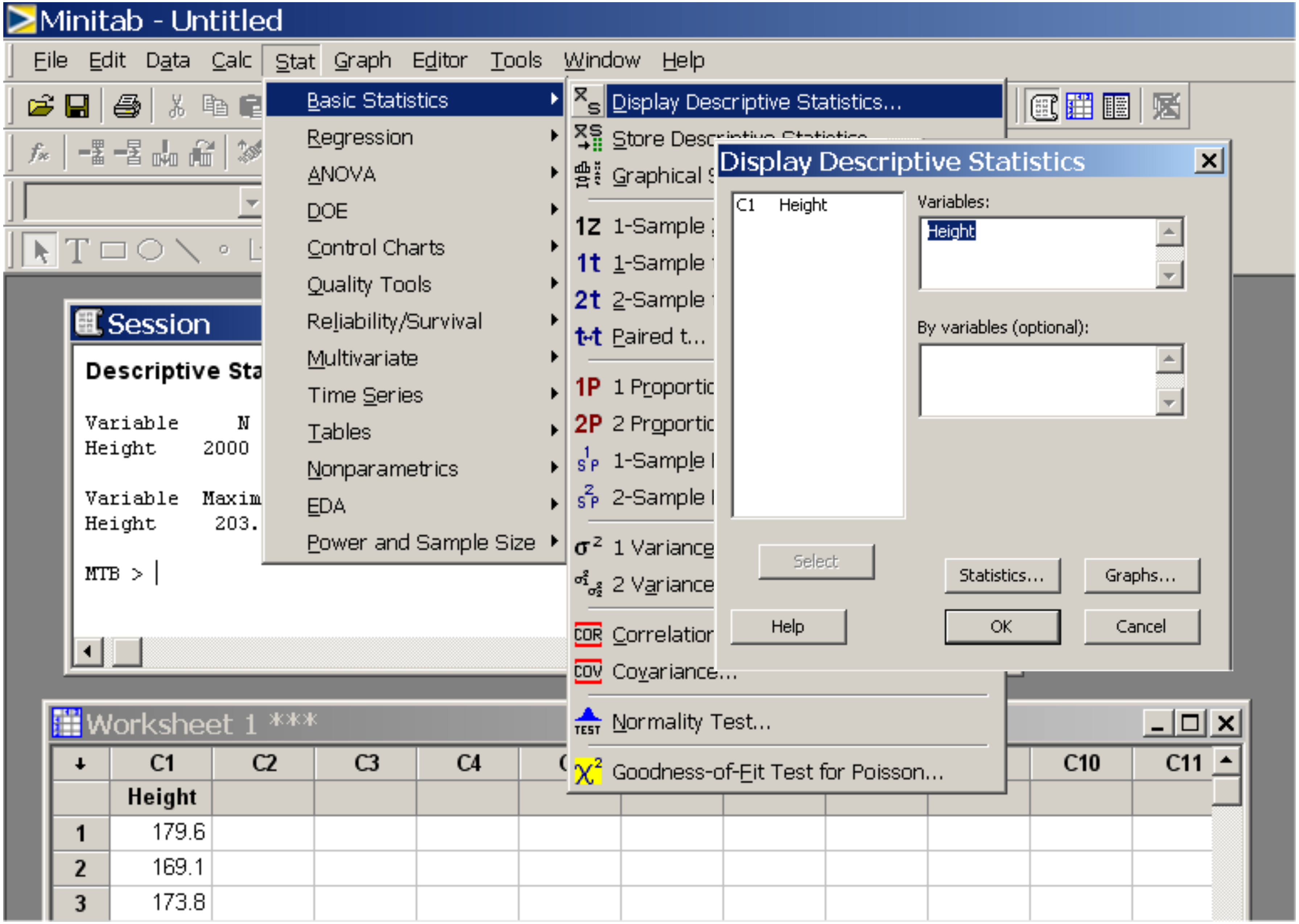
$$SD^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n\text{-}1}$$



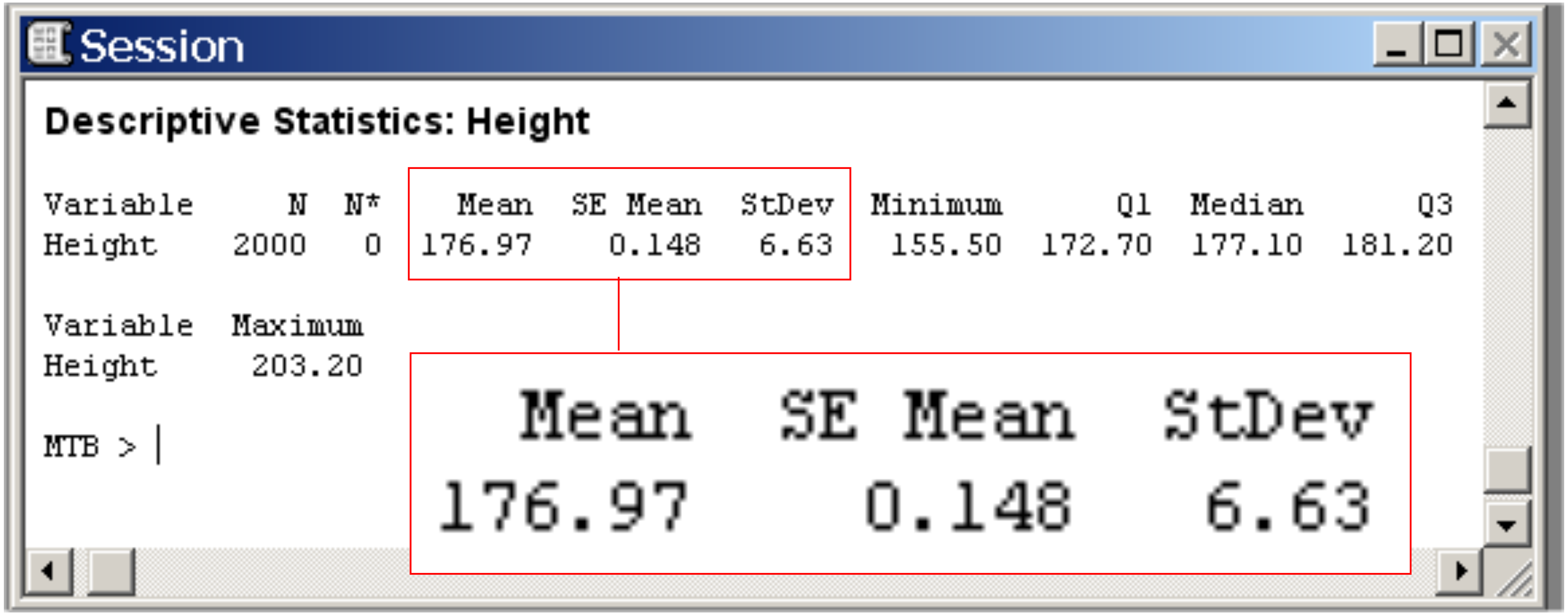
- Standard deviation
  - Square root of variance

$$\text{SD} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n\text{-}1}}$$

Minitab – Displaying Descriptive Statistics



Descriptive Statistics of Male Height Data



Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Height	2000	0	176.97	0.148	6.63	155.50	172.70	177.10	181.20

Variable	Maximum
Height	203.20

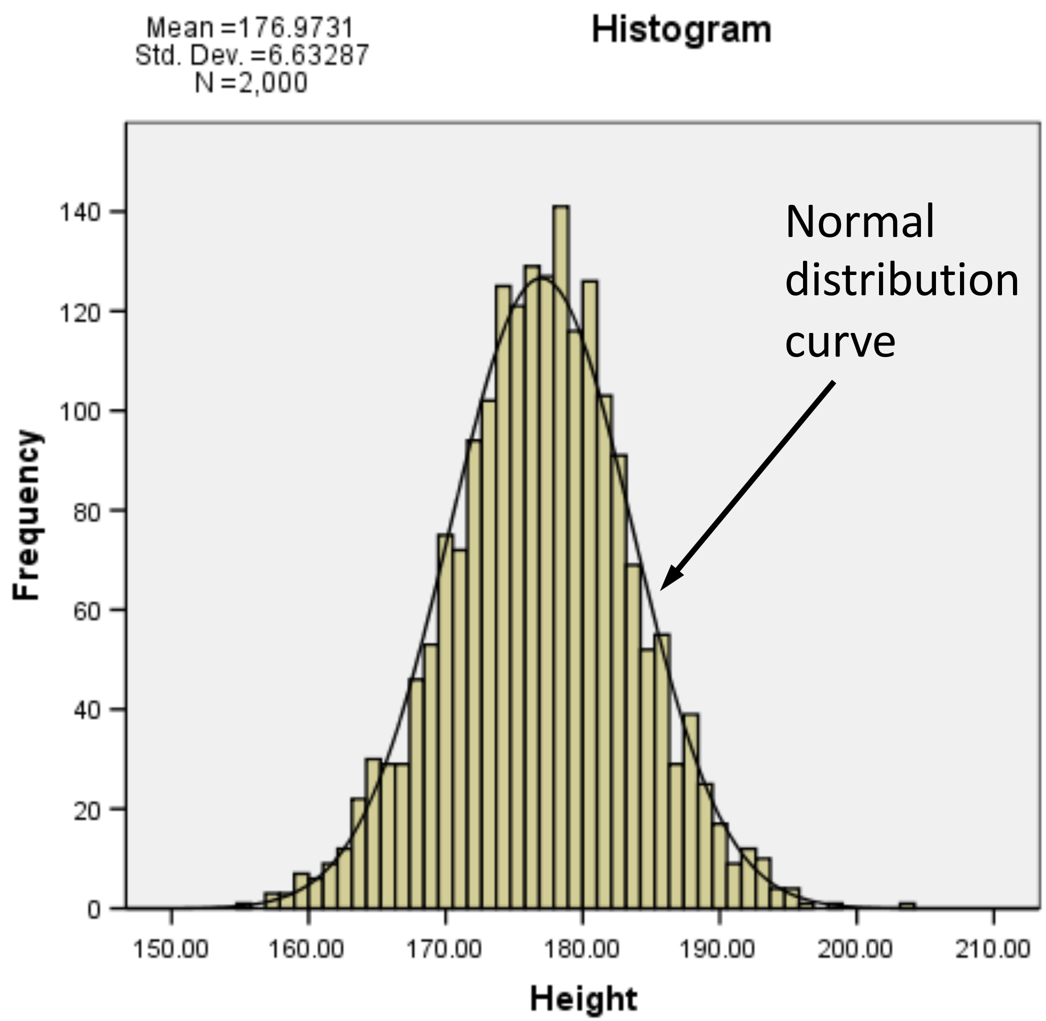
MTB >

Mean	SE Mean	StDev
176.97	0.148	6.63

Minitab: Stat / Basic Statistics / Display Descriptive Statistics / C1 → Variables

The Normal Distribution

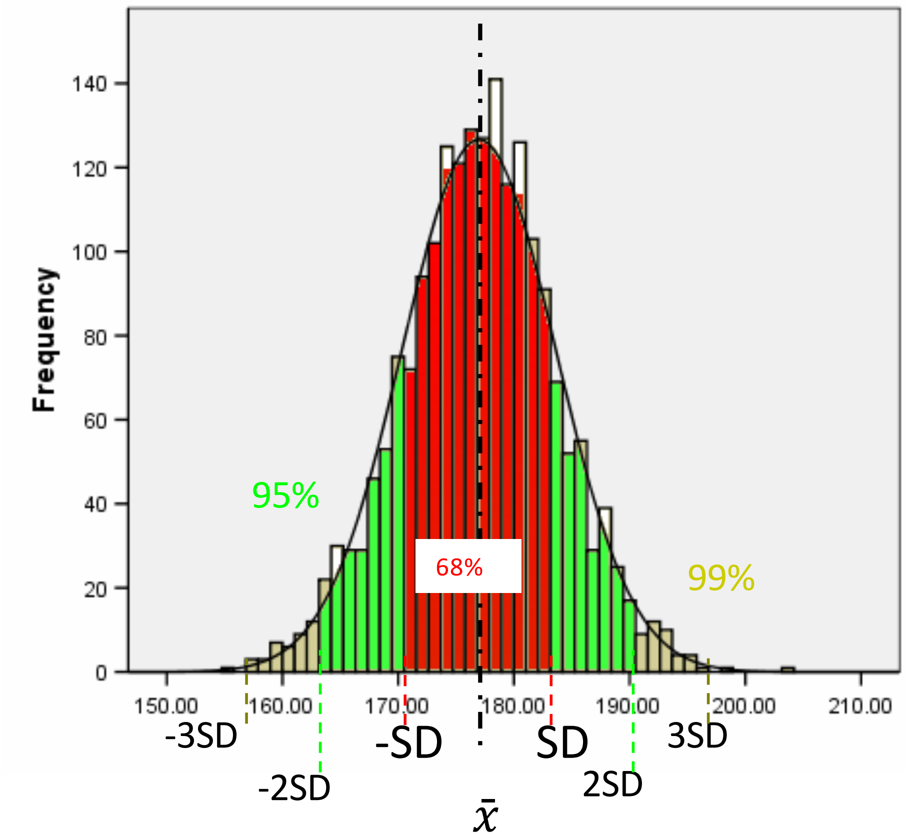
- The frequency distribution of the height data has a **<u>bell-shape</u>**, symmetrically distributed about the population mean.
- The percentage of data values above and below the mean can often be represented by mathematical function known as the **normal distribution**
- Mean, median and mode are equal in a normal distribution



The Normal Distribution

- When a population is 'normally' distributed the percentage of data within a specific range can be predicted from the mean ($\bar{x}$) and standard deviation (σ)

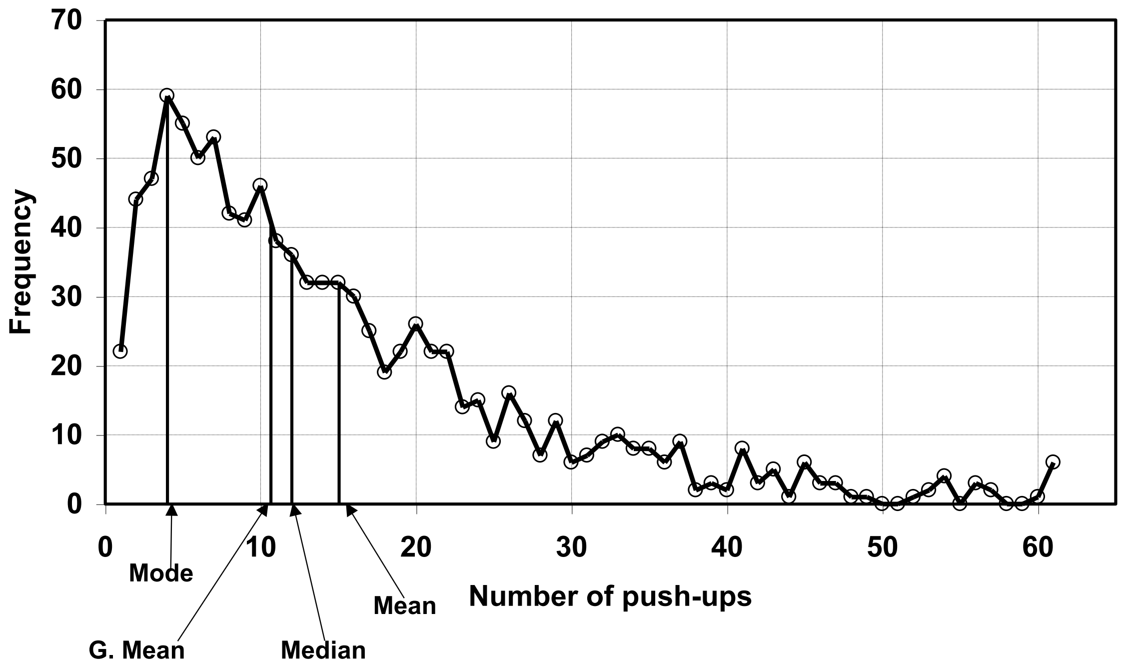
Range	% of data
$\bar{x}$ - SD to $\bar{x}$ + SD	68%
$\bar{x}$-2SD to $\bar{x}$ + 2SD	95%
$\bar{x}$-3SD to $\bar{x}$ + 3SD	99%



Height Range	% of data
170.4 to 183.6	68%
163.8 to 190.2	95%
157.2 to 196.8	99%

Not all data is normally distributed!!

- Data distributions can sometimes be **<u>skewed</u>** with excess numbers of high or low values.
- In a skewed distribution, the mean, median and mode may be significantly different.
- The percentage of data within specific ranges cannot be predicted from the mean and standard deviation.



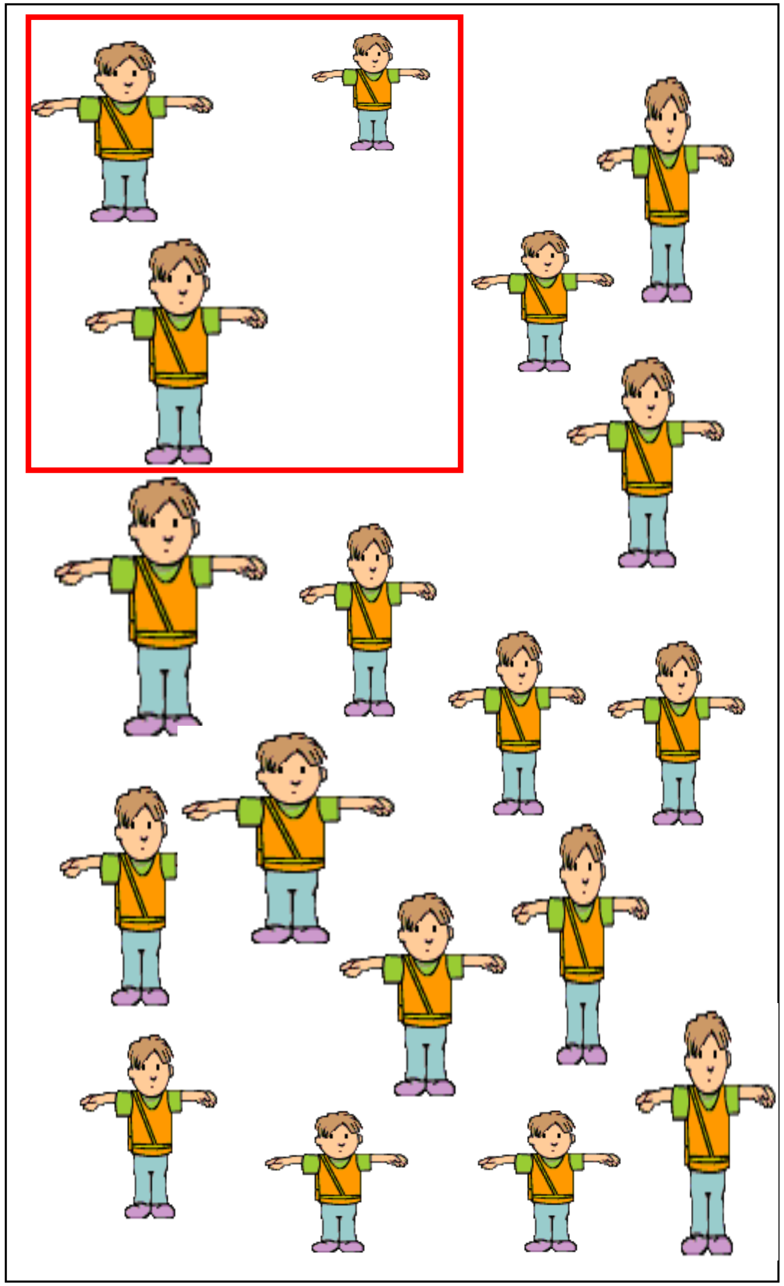
Estimating a population mean from a sample

- It is not practical to determine the mean and standard deviation of a population by measuring every member.
- An **estimate** has to be made from a small, randomly selected, **sample** of the members of a population
  - Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

  - Sample standard deviation

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

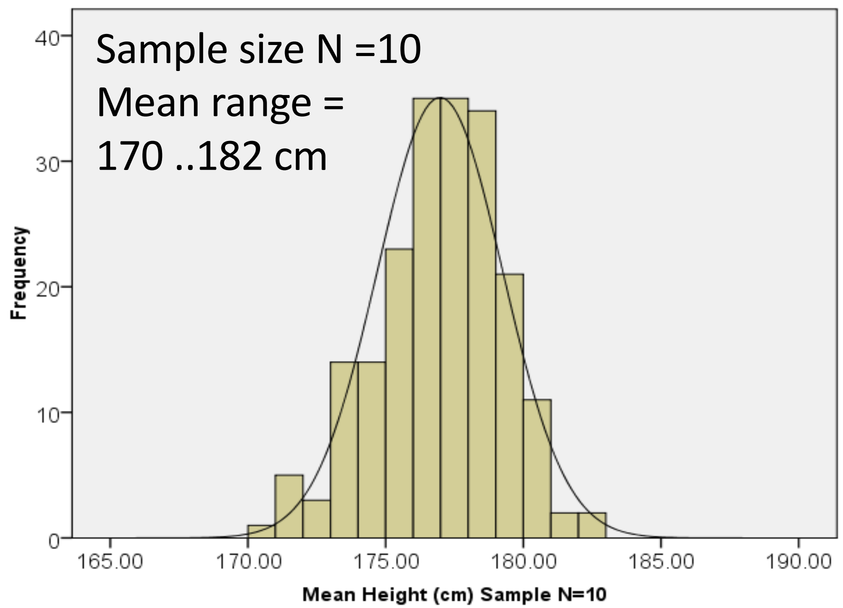


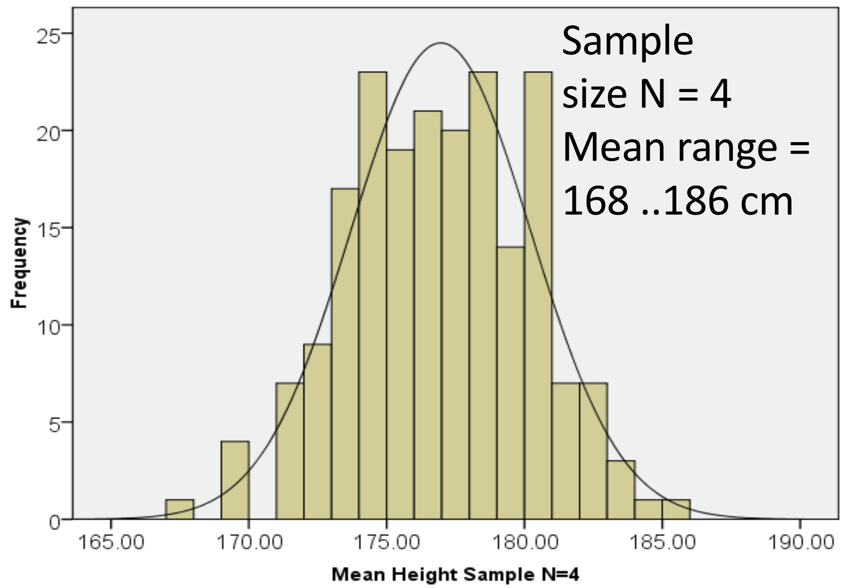
The Standard Error of the Mean

- The larger the sample size, the less variation there is in the estimated mean.
- If we repeatedly took many samples from the population, the standard deviation of the means of these samples – **<u>the standard error</u>** - would tell us how variable the estimate of the mean was.
- The standard error can be estimated from a single sample by

$$SE = \frac{SD}{\sqrt{N}}$$

Means from 200 sets of samples





Estimated means vary from sample to sample

- Different samples give different estimates for the population mean.

- Before we can say anything meaningful about the population mean (the average male height), we need to determine how variable the estimate of the population mean obtained from a sample is.

Mean height values from 3 different samples of 4 students from first year male student population.

	Heights (cm)		
n = 4	**Sample 1**	**Sample 2**	**Sample 3**
	179	181	181
	177	176	181
	172	167	178
	179	176	176
Mean	**176.7**	**175.1**	**178.0**

Confidence Intervals

- The distribution of sample means of a normal distribution is also a normal distribution, so the percentage of sample means within a specific range of the population mean can be calculated.

- In 68% of the samples, the sample mean is within one standard error of the population mean. This is known as the 68% **confidence interval (μ)** . (since we can be 68% confident that the mean of our sample is within one SE of the population mean)

$$\mu = \bar{x} \pm SE$$

- The 95% confidence interval is

$$\mu = \bar{x} \pm 1.96 \times SE$$

- The 99% confidence interval is

$$\mu = \bar{x} \pm 2.58 \times SE$$

Confidence Intervals: Male Heights

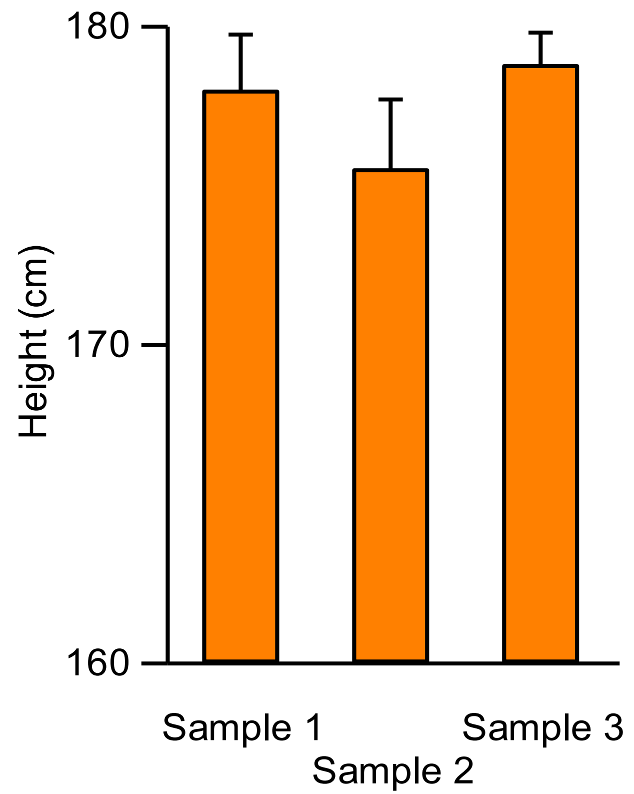
	N	Mean (cm)	SD (cm)	SE (cm)	95% C.I. (cm)
Sample 1	4	176.7	3.3	1.6	173.5-180
Sample 2	4	175	5.8	2.9	169.3-180.7
Sample 3	4	179	2.4	1.2	176.6-181.4
Population		177	6.6		177

Note. Although sample means vary (176.7, 175, 179), the confidence intervals of the 3 samples contains the population mean (177). This will be the case for 95% (1 in 20) of the samples of size 5 taken from the population.

We can say with 95% confidence (i.e. that we will be right 19 out of 20 times) that the mean of the population lies within the confidence limits.

Visual Display of Error Bars

- Most scientists put error bars on graphs and bar charts – *but not always the same sort of error bar*.
- Standard deviation
  - Shows the spread of the data in the sample or population. (Not commonly used)
- Standard error of the mean
  - Shows the accuracy of the sample mean. The range of values around the sample mean where the population mean can be predicted to lie with 68% confidence **(recommended)**
- 95% confidence interval
  - The range of values around the sample mean where the population mean can be predicted to lie with 95% confidence



Bar chart of means of samples 1-3 with error bars showing standard errors

Basic Definitions and Concepts

- Specifically, such terms:
- discrete and continuous variables,
- frequency distribution, population, sample,
- mean, median, standard deviation,
- Variance (SD^2), coefficient of variation (CV),
- range,
- accuracy, and precision

Practical Exercise #1
Descriptive Statistics & Confidence Intervals

The clinical laboratories of 5 hospitals (A to E), tested a portion of the same standard sample of pooled human blood serum containing 42.0 g/l of albumin.

Each laboratory did 6 determinations (on the same day) of the albumin concentration, with the following results (in g/l)

A	B	C	D	E
42.5	39.8	43.5	35.0	42.2
41.6	43.6	42.8	43.0	41.6
42.1	42.1	43.8	37.1	42
41.9	40.1	43.1	40.5	41.8
41.1	43.9	42.7	36.8	42.6
42.2	41.9	43.3	42.2	39.0

- Exercise
  - Enter data into excel sheet
  - Produce box plots
  - Produce descriptive statistics
  - Calculate confidence intervals for mean albumin conc.
  - Which labs. produce accurate determinations.
  - Which labs. have best/worst precision.

For descriptivre analysis:
Data/Data analysis/ Descriptive statistics

Column1		Column2		Column3		Column4		Column5	
Mean	41.9	Mean	41.9	Mean	43.2	Mean	39.1	Mean	41.5333333
Standard Error	0.201659779	Standard Error	0.697137	Standard Error	0.17126977	Standard Error	1.32765457	Standard Error	0.52577984
Median	42	Median	42	Median	43.2	Median	38.8	Median	41.9
Mode	#N/A	Mode	#N/A	Mode	#N/A	Mode	#N/A	Mode	#N/A
Standard Deviation	0.493963561	Standard Deviation	1.70762994	Standard Deviation	0.41952354	Standard Deviation	3.25207626	Standard Deviation	1.28789234
Sample Variance	0.244	Sample Variance	2.916	Sample Variance	0.176	Sample Variance	10.576	Sample Variance	1.65866667
Kurtosis	0.298978769	Kurtosis	-1.7798918	Kurtosis	-1.1428202	Kurtosis	-2.15647	Kurtosis	4.62270345
Skewness	-0.716851178	Skewness	-0.1308579	Skewness	0.21940474	Skewness	0.02412633	Skewness	-2.0542549
Range	1.4	Range	4.1	Range	1.1	Range	8	Range	3.6
Minimum	41.1	Minimum	39.8	Minimum	42.7	Minimum	35	Minimum	39
Maximum	42.5	Maximum	43.9	Maximum	43.8	Maximum	43	Maximum	42.6
Sum	251.4	Sum	251.4	Sum	259.2	Sum	234.6	Sum	249.2
Count	6	Count	6	Count	6	Count	6	Count	6

For descriptivre analysis:

Box plot:

Insert/ Histogram/ box or whisker