Descriptive Statistics -
plus a discussion about SEM

Kathleen Torkko
September 9, 2019

STATISTICS

↙ ↘

DESCRIPTIVE STATISTICS

Graphs & tables showing frequency distributions

Numeric summary measures

- Location (central tendency)
- Dispersion (variability)
- Shape (data distribution)

INFERENTIAL STATISTICS

Estimation

- Sampling distributions
- Standard errors
- Confidence intervals

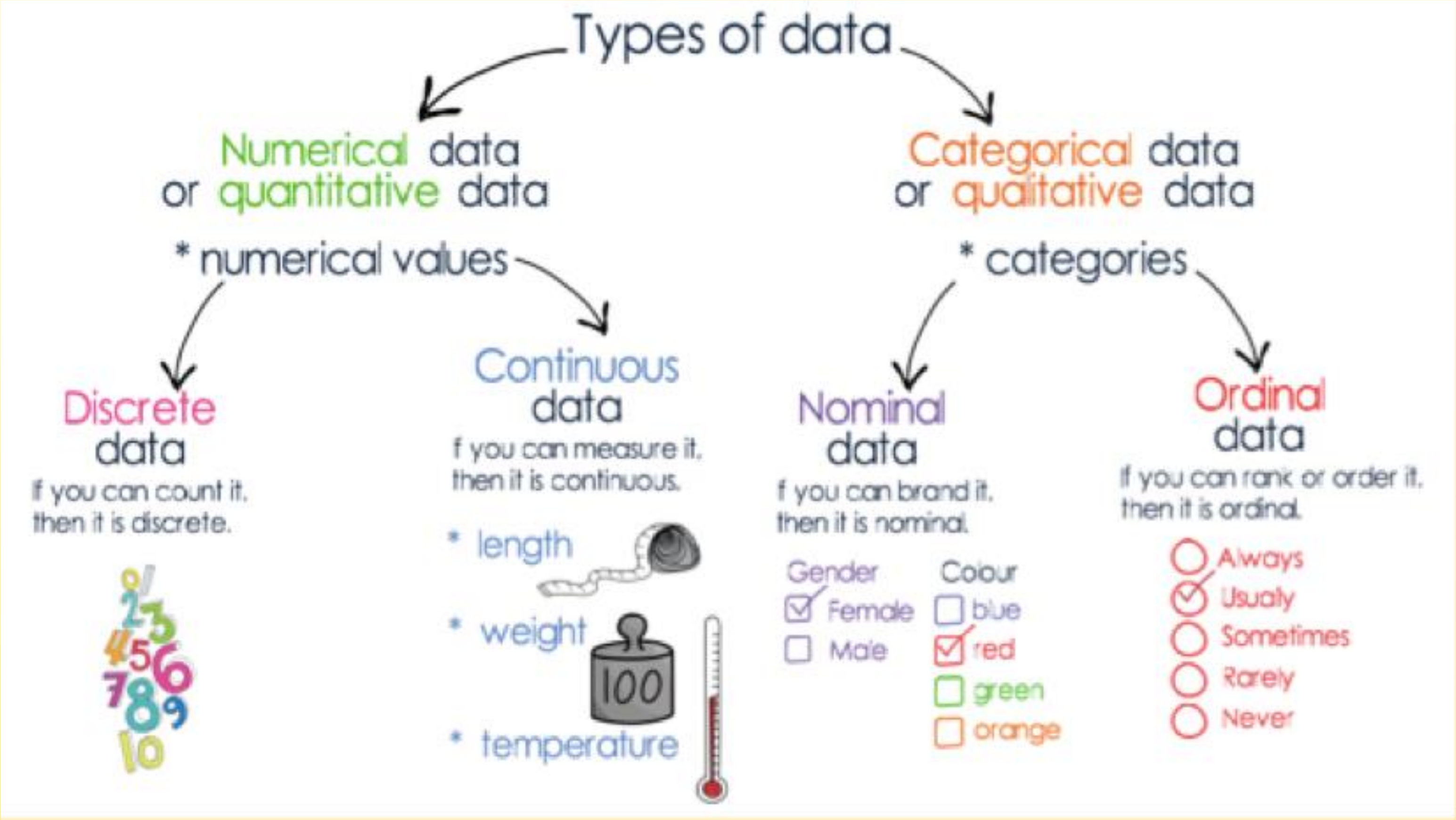
Hypothesis testing

- Parametric tests
- Nonparametric tests

Objectives

- Learn about different types of data
- Learn about measures of central tendency (AKA location)
- Learn about measures of variability (dispersion)
- Understand differences between SD, SEM, CI

A comment about TA office hours…



Types of data

Numerical data or quantitative data

* numerical values

Discrete data

If you can count it, then it is discrete.

Continuous data

f you can measure it, then it is continuous.

* length

* weight

* temperature

Categorical data or qualitative data

* categories

Nominal data

f you can brand it, then it is nominal.

Gender

☑ Female

□ Male

Colour

□ blue

☑ red

□ green

□ orange

Ordinal data

If you can rank or order it, then it is ordinal.

○ Always

✓ Usualy

○ Sometimes

○ Rarely

○ Never

legac.com.au/blogs/further-mathematics-exam-revision/further-mathematics-unit-3-data-analysis-types-of-data

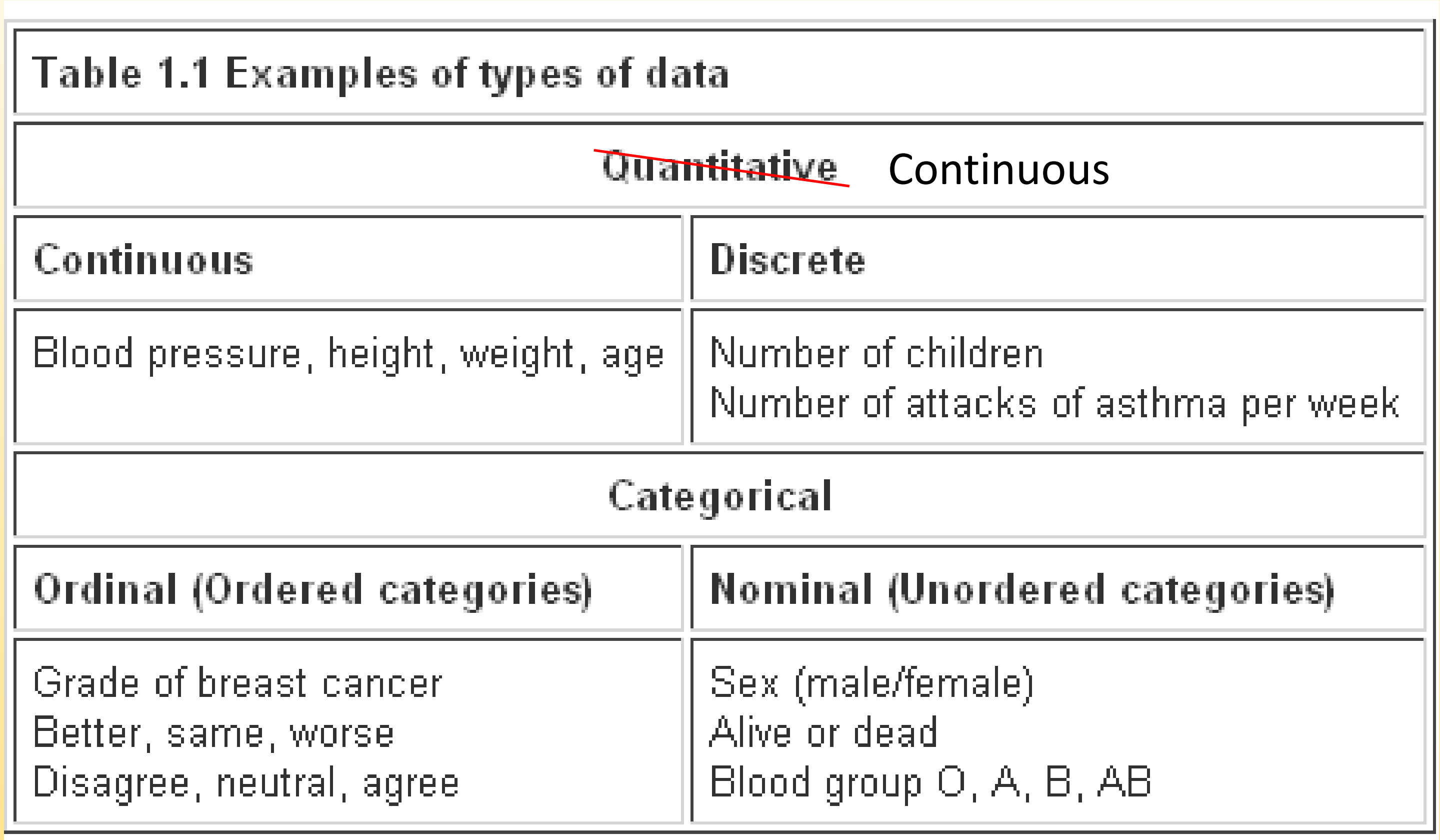


Table 1.1 Examples of types of data

~~Quantitative~~ Continuous	
Continuous	**Discrete**
Blood pressure, height, weight, age	Number of children Number of attacks of asthma per week
Categorical	
Ordinal (Ordered categories)	**Nominal (Unordered categories)**
Grade of breast cancer Better, same, worse Disagree, neutral, agree	Sex (male/female) Alive or dead Blood group O, A, B, AB

www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/1-data-display-and-summary

Types of Descriptive Statistics Depends on Types of Data

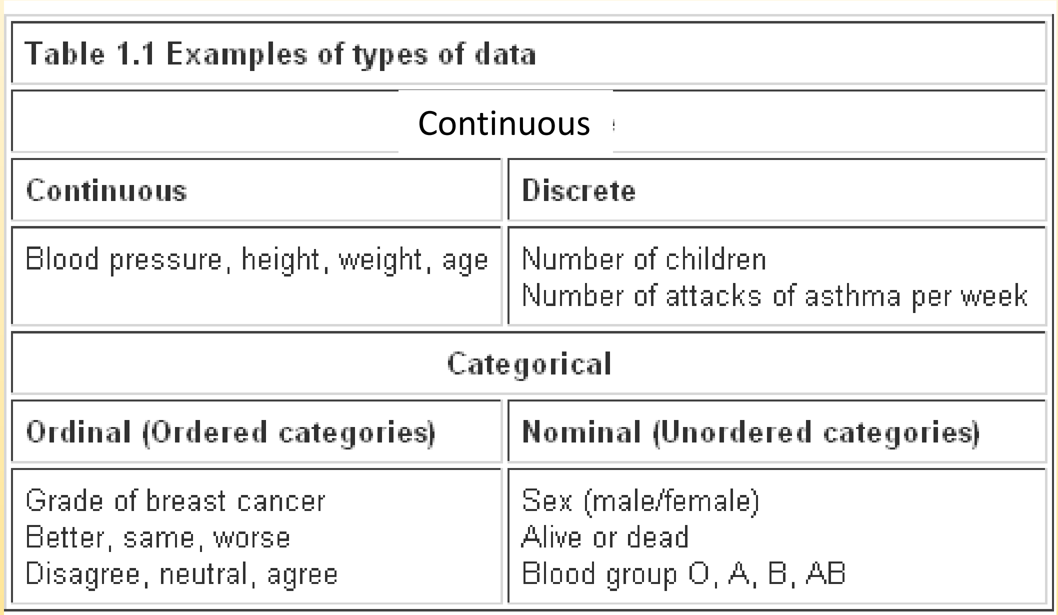


Table 1.1 Examples of types of data	
Continuous	
Continuous	**Discrete**
Blood pressure, height, weight, age	Number of children Number of attacks of asthma per week
Categorical	
Ordinal (Ordered categories)	**Nominal (Unordered categories)**
Grade of breast cancer Better, same, worse Disagree, neutral, agree	Sex (male/female) Alive or dead Blood group O, A, B, AB

Mean, median, standard deviation, range

Percentages, counts, median (ordinal)

Types of Graphic Presentation Depends on Types of Data

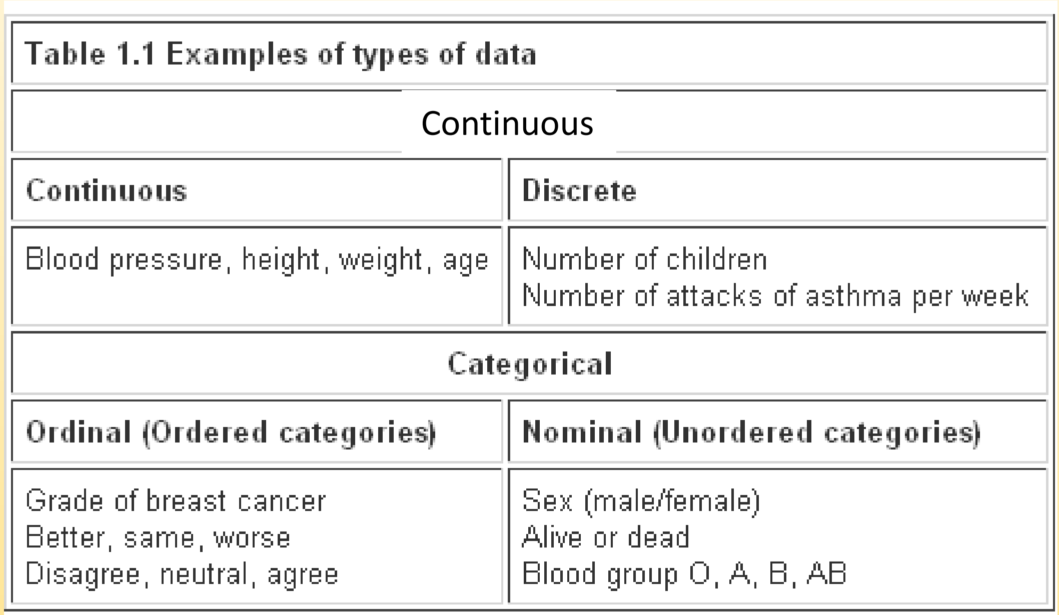


Table 1.1 Examples of types of data

Continuous	
Continuous	**Discrete**
Blood pressure, height, weight, age	Number of children Number of attacks of asthma per week
Categorical	
Ordinal (Ordered categories)	**Nominal (Unordered categories)**
Grade of breast cancer Better, same, worse Disagree, neutral, agree	Sex (male/female) Alive or dead Blood group O, A, B, AB

Histogram, scatter plot, line graph, box plots

Table, bar graph, pie chart

Types of Analyses Depends on Types of Data

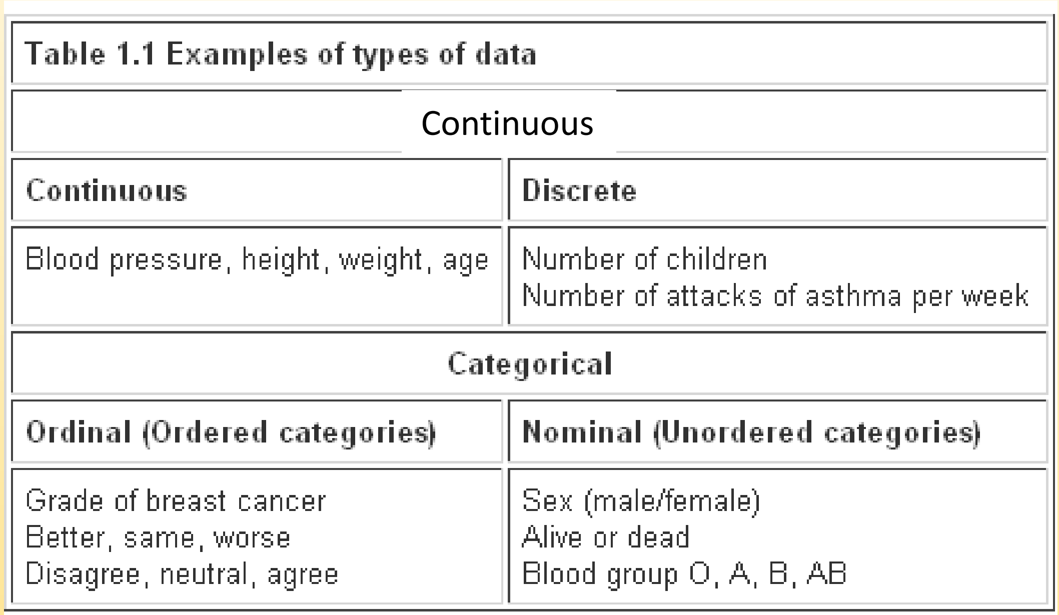


Table 1.1 Examples of types of data

Continuous	
Continuous	**Discrete**
Blood pressure, height, weight, age	Number of children Number of attacks of asthma per week
Categorical	
Ordinal (Ordered categories)	**Nominal (Unordered categories)**
Grade of breast cancer Better, same, worse Disagree, neutral, agree	Sex (male/female) Alive or dead Blood group O, A, B, AB

T-tests, ANOVA, Correlation, Linear Regression, Wilcoxon Rank Sum, Kruskal Wallis

Chi-square, Fishers exact tests

Weights (in Pounds) of Elementary Students
Which class is the heaviest?

Class A		Class B	
102	115	127	162
128	109	131	103
131	89	96	111
98	106	80	109
140	119	93	87
93	97	120	105
110		109	

	Class A	Class B
Mean	110.5	110.2

A data summary helps us better understand the data set and more easily see differences between data sets.

Descriptive Statistics

A fundamental task in many statistical analyses is to characterize the sample data set.

Descriptive statistics summarize data

Measures of central tendency:

mean, median, mode

Measures of variability:

range, quartiles, standard deviation, variance

Graphical techniques

Measures of Central Tendency

The mean, median and mode are all valid measures of central tendency,

Some measures of central tendency are more appropriate to use than others

MEAN - the mean (AKA average) is the sum of the data points divided by the number of data points.

MEDIAN - the median is the value of the point which has half the data smaller than that point and half the data larger than that point.

MODE - the mode is the value of the random sample that occurs with the greatest frequency. It is not necessarily unique.

Can be referred to as “point estimates” or a “measure of effect”

Measures of Central Tendency: Mean

Population Mean vs. Sample Mean

Population mean:

Computed using all values in the population of interest

Usually impractical or impossible to actually compute it

It is usually estimated via the *sample mean*

Symbolized by Greek letter μ (mu)

Called a parameter

Sample mean:

Computed using a sample drawn from a population

Symbolized by Roman letter with bar over it, e.g.,

$\overline{x}$ = the mean of a sample of values

Called a statistic or estimation of the population mean

Measures of Central Tendency: Mean

$$\bar{x} = \frac{(x_1 + x_2 + \cdots + x_n)}{n}$$

$$\bar{x} = \frac{\sum x}{n}$$

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

This includes every value in your data set as part of the calculation.

One main disadvantage: it is particularly susceptible to the influence of outliers (extreme values is better).

Staff	1	2	3	4	5	6	7	8	9	10
Salary	$35K	$38K	$36K	$34K	$35K	$35K	$32K	$37K	$190K	$195K

Mean of this group is $67K, but most earn <$39K
The mean is being skewed by the two large salaries.

The Median

The ***median*** is a point that divides the data such that 50% of the values fall below it and 50% fall above it.

Same thing as ***the 50th percentile***

Less affected by outliers and skewed data.

To find the median:

Sort the values in ascending order

For an odd number of values, the median = the middle value

For an even number of values, the median = the mean of the two middle values

Staff	7	4	1	5	6	3	8	2	9	10
Salary	$32K	$34K	$35K	$35K	$35K	$36K	$37K	$38K	$190K	$195K

Mean of this group is $67K, the median is $35.5K

Measures of Central Tendency: Median

If *odd n*, middle value of sequence

Median = 56

1.	14	
2.	35	
3.	45	
4.	54	
5.	55	5 values below in value
6.	$\boxed{56}$	
7.	57	5 values above
8.	65	
9.	87	
10.	89	
11.	92	

If *even n*, average of 2 middle values

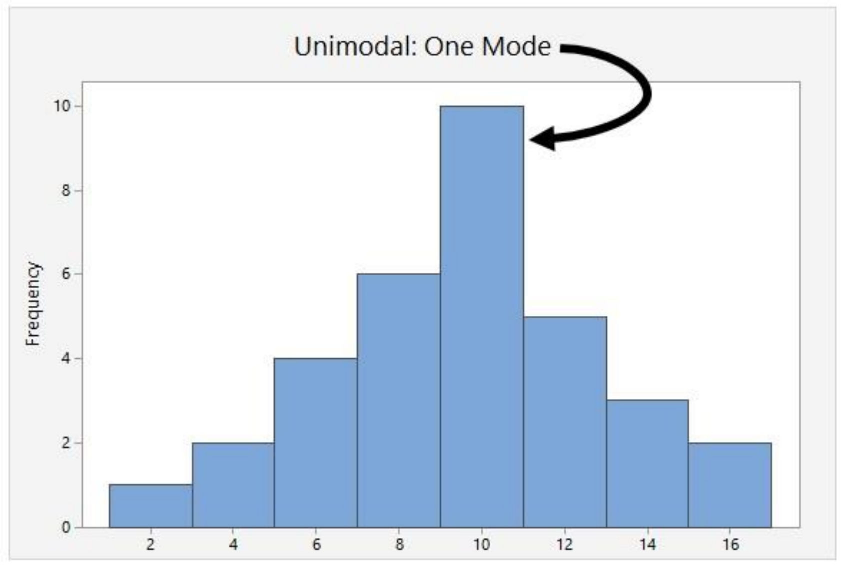
Median = 55.5

1.	14	
2.	35	
3.	45	
4.	54	
5.	55	5 values below ← 55.5
6.	56	5 values above
7.	57	
8.	65	
9.	87	
10.	89	

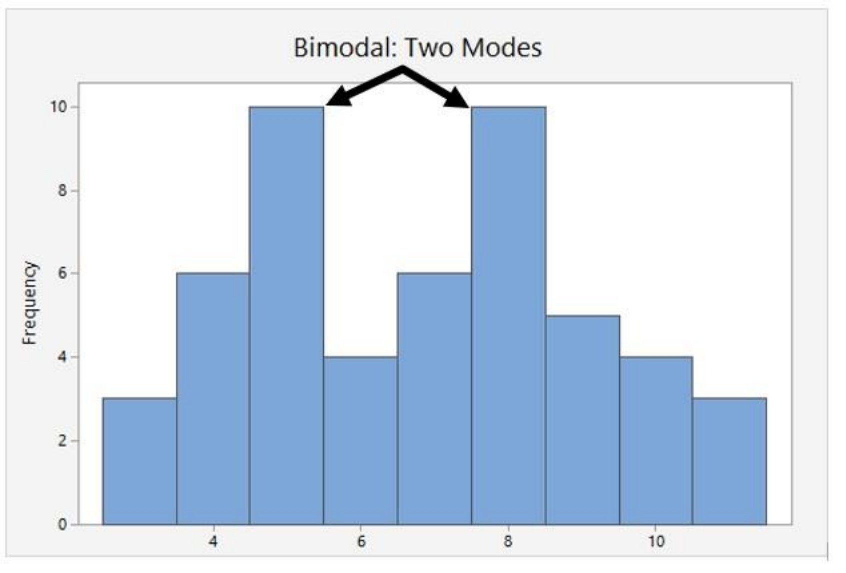
Measures of Central Tendency: Mode

The mode is the most common variable in the dataset [the highest peak(s)] in a frequency distribution chart

Unimodal:
A distribution with one prominent peak



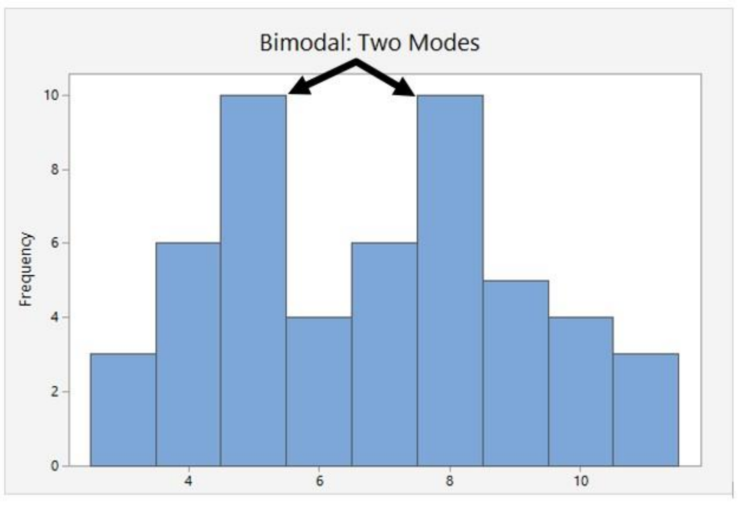
Bimodal:
A distribution with two prominent peaks of the same magnitude

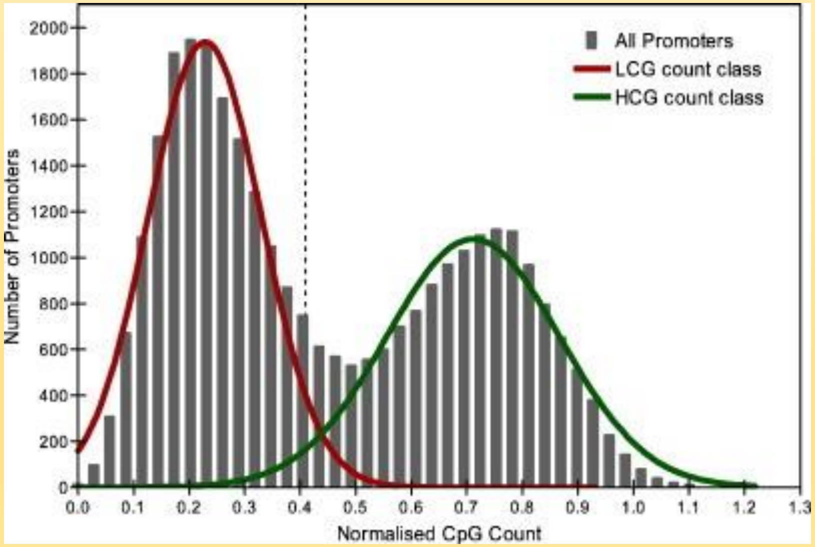


Two modes in a distribution (bimodal) vs. a bimodal distribution

A **bimodal distribution** is a chart of frequency that has two different peaks from two different distributions

It is not related to the other common usage of "mode," which refers to the most frequent number found in a distribution.





DNA sequence and structural properties as predictors of human and mouse promoters.
Akan P, Deloukas P - Gene (2007)

TRY DOING BEFORE CLASS

Using BIOS 6606 Student Years in CO 2017-19.xlxs

Create frequency distributions of the data

1. All data
2. By born in Colorado or not

Which measure of location to use?

For ***continuous*** variables: mean, median, or mode

The mean is preferred if the distribution is reasonably symmetrical

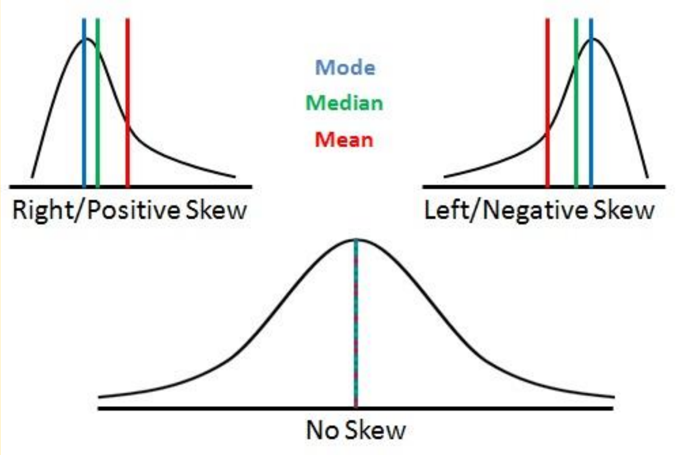
For skewed distributions (*e.g*., salaries) or data with outliers, the median is preferred

For ***categorical*** (nominal or ordinal) variables (*e.g*., stage of cancer), the mode (nominal or ordinal) or median (for ordinal only) is appropriate

Type of Variable	**Best measure of central tendency**
Nominal	Mode
Ordinal	Median
Continuous (not skewed, no outlier)	Mean
Continuous (skewed, outliers)	Median

Effect of skew in data on mean median and mode

When the data are perfectly symmetrical, the mean, median and mode are identical.



As data become skewed the mean loses its ability to provide the best central location for the data because the skewed data are dragging it away from the typical value.

Median is better than the mean (or mode) when the data are skewed

The median is not as strongly influenced by the skewed values.

Why is Skew Important?

Skewness is a measure of symmetry, or more precisely, the lack of symmetry.

A distribution is symmetric if it looks the same to the left and right of the center point

Skewness is not necessarily abnormal

Some data have a naturally skewed distribution

Problem is, some statistical tests assume a symmetrical normal distribution

TRY DOING BEFORE CLASS

Do a frequency distribution of BIOS 6606 Student Number of Siblings 2017-19.xlsx for all years combined.

Skewness & Kurtosis

Skewness has to do with the symmetry (or lack thereof) of a distribution

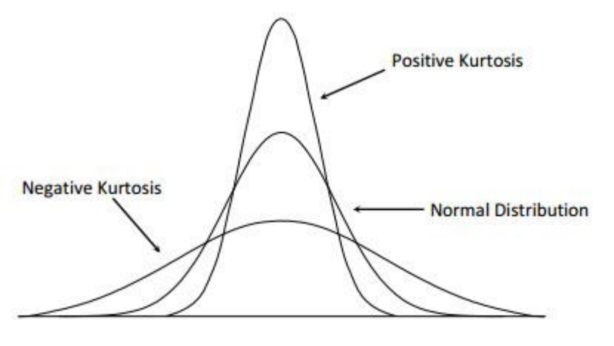
As a general rule of thumb:

- If **skewness** is less than -1 or greater than 1, the distribution is highly skewed.
- If **skewness** is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
- If **skewness** is between -0.5 and 0.5, the distribution is approximately symmetric.

Kurtosis is another shape parameter—it describes the shape of a distribution relative to the Normal distribution (more about the Normal distribution soon!)

Flatter than Normal distribution= **negative** kurtosis

More peaked than Normal distribution= **positive** kurtosis



Data Always Come with Noise

Variability is good, because we need variability in predictors to explain variability in outcomes.

e.g., to find genetic associations with diseases, we need genetic variation;
to find associations of markers with cancer, we need the expression levels to vary between cancer and non-cancer tissues.

Variability can also be annoying

Nearly all biological measurements, when repeated, exhibit variation
Need to look for and control likely sources of systematic error (bias)

The standard deviation and variance are descriptive measures of variability

(Standard errors of the mean or confidence intervals are inferential measures of uncertainty)

Variability: Dispersion or Spread of Data

Measures of variability:

variance
standard deviation
range
quartiles
interquartile range (IQR)

The standard deviation is one of the most commonly used measure of variability.

The standard deviation is heavily influenced by outliers and therefore may not be the most appropriate measure of variability when data are not symmetrical.

Variance formula

$$s^2 = \frac{\sum\left(x - \bar{X}\right)^2}{n-1}$$

Standard deviation formula

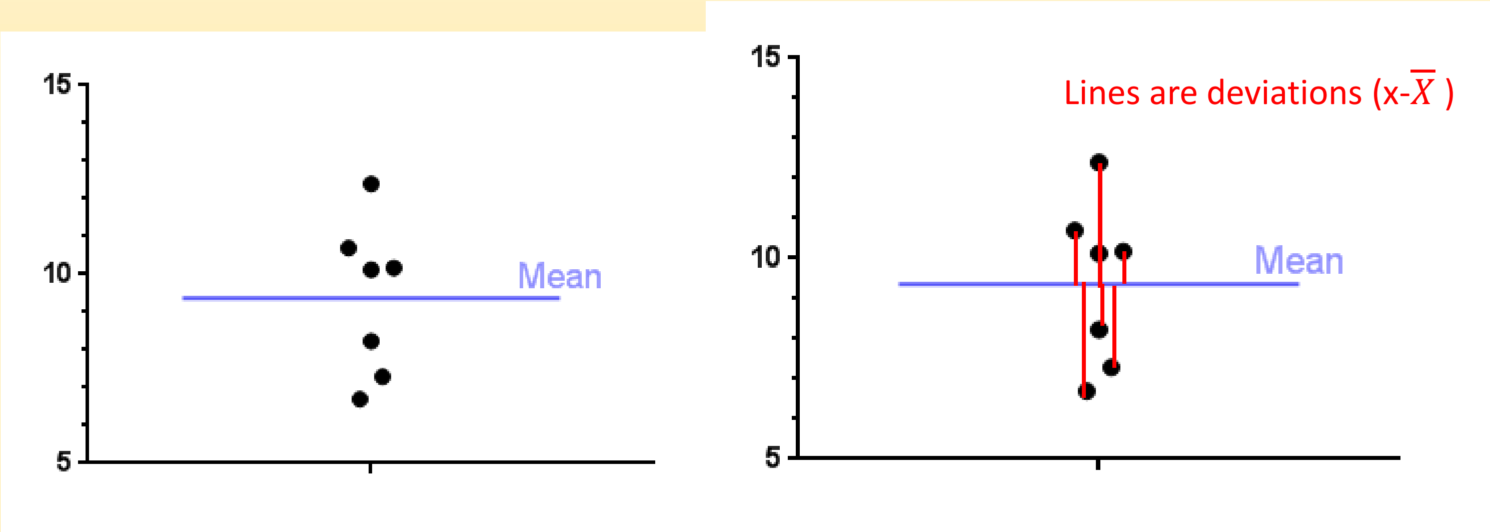
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Taking the square root of the variance reverts the units to match those of the sample data mean

Variance

A measure of the spread of the values of a variable.

Based on *deviations* of individual observations about the central value (the mean)



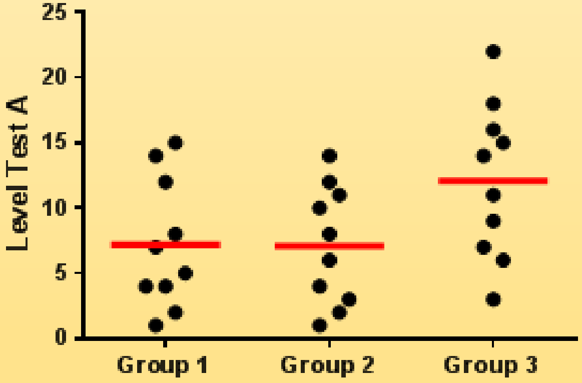
$$s^2 = \frac{\sum\left(x - \bar{X}\right)^2}{n-1}$$

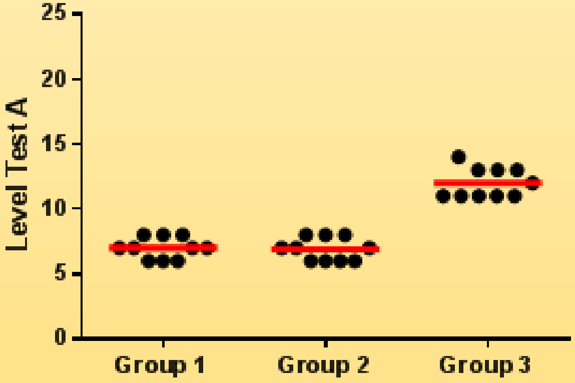
Variance

The further the individual values are from the mean, the larger the variance

$$s^2 = \frac{\sum\left(x-\bar{X}\right)^2}{n-1}$$

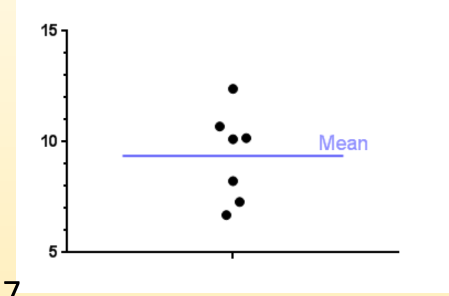
The closer the values are to the mean, the smaller the variance





Variance Calculation by Hand

12.40
10.17
10.12
7.29
8.23
6.70
10.70



Mean=9.37

$(x-\bar{X})$	$(x-\bar{X})^2$
12.40 - 9.37 = 3.03	9.18
10.17 - 9.37 = 0.80	0.64
10.12 - 9.37 = 0.75	0.56
7.29 - 9.37 = -2.08	4.33
8.23 - 9.37 = -1.14	1.30
6.70 - 9.37 = -2.67	7.13
10.70 - 9.37 = 1.33	1.77
$\Sigma(x-\bar{X})^2$	24.91

$$s^2 = \frac{\sum(x-\bar{X})^2}{n-1}$$

= 24.91/(7-1)

=24.91/6

= 4.15

Population vs. sample variances

$$\sigma^2 = \frac{\sum(x - \mu)^2}{N}$$

$$s^2 = \frac{\sum(x - \bar{X})^2}{n-1}$$

Why “n-1” for the sample variance?

It makes the variance calculation a little bigger to account for the fact we are using a sample to estimate the population parameter.

“n-1” increases the uncertainty slightly.

Standard Deviation

The ***standard deviation*** (SD) is the square root of the variance

It conveys the same information as the variance, but in the same units as the original values

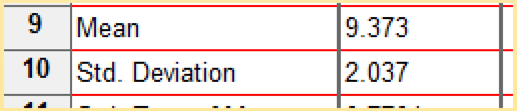
Therefore, it is easier to interpret

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

$$s = \sqrt{\frac{\sum(x-\bar{X})^2}{n-1}}$$

$s^2 = 4.15$

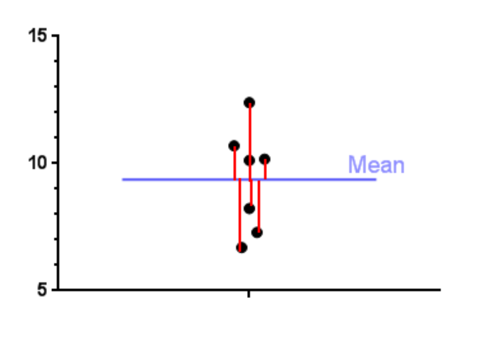
$s = 2.037$



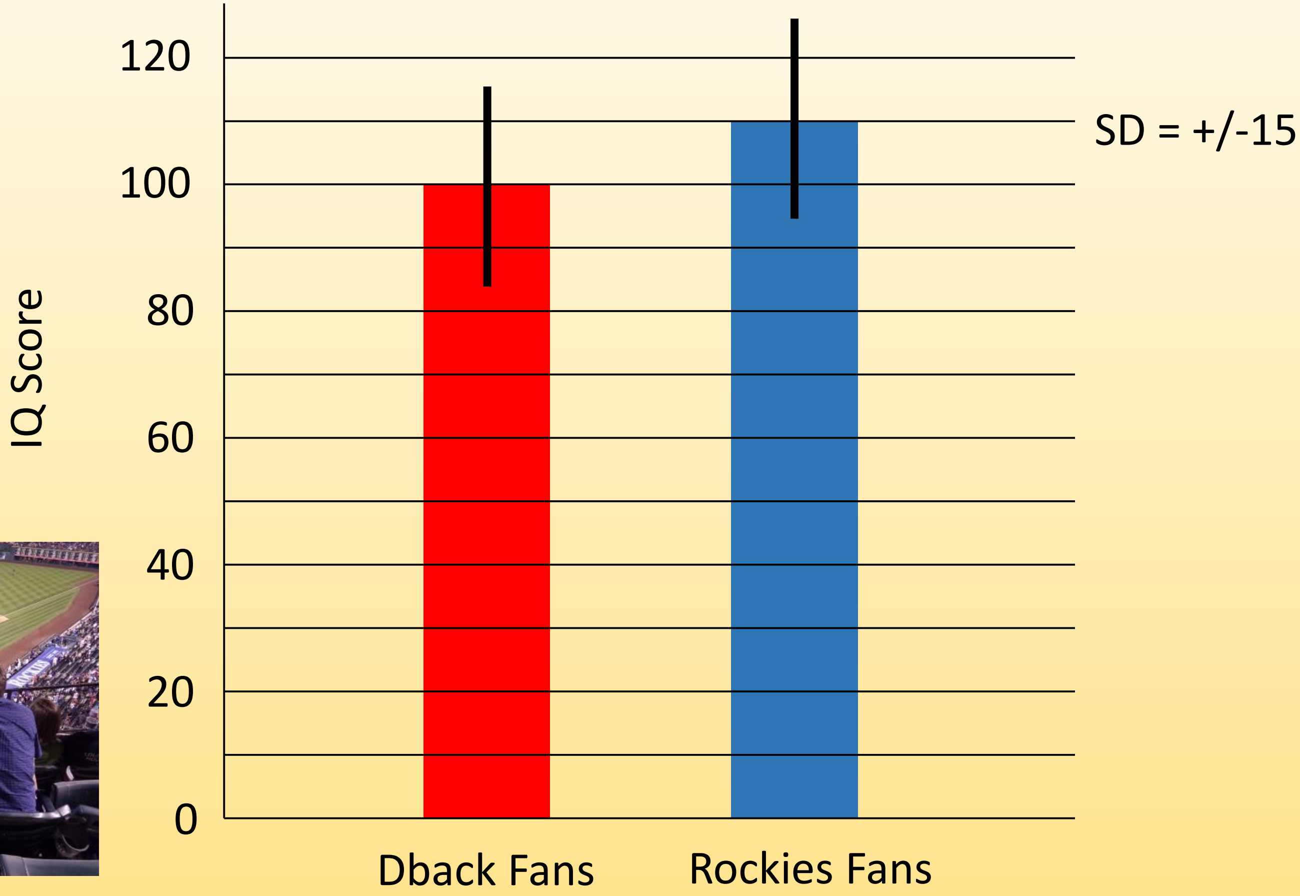
9	Mean	9.373
10	Std. Deviation	2.037

Deviation: An individual value minus the mean

Sum of squared deviations: Deviations are first squared and then added together, also known as sum of squares or SS



What Does Standard Deviation Tell You?



SD estimates the variability around the mean

Based on *z distribution*, 95% lie within +/- 1.96 SDs

95% of Dback fans have an IQ between ~70 and 130

95% of Rockies fans have an IQ between ~80 and 140

Totally made up data!

The Range

The spread, or the distance, between the lowest and highest values of a variable.

It can be reported by giving the minimum and maximum values themselves, or just the difference between them

min = 18, max = 89 (**18 to 89**)

The range = 89-18 = **71**

Range

Class A--Weights of 13 Students

102
128
131
98
140
93
110
115
109
89
106
119
97

Class A Range = 140 - 89 = 51

Class A Range = 89 to 140

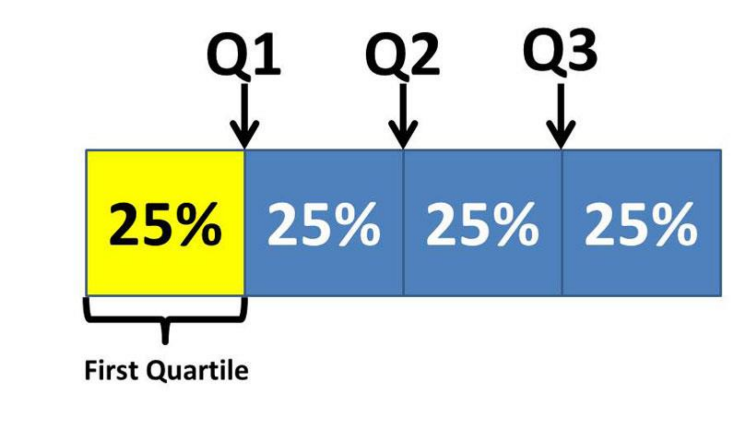
Class B--Weight of 13 Students

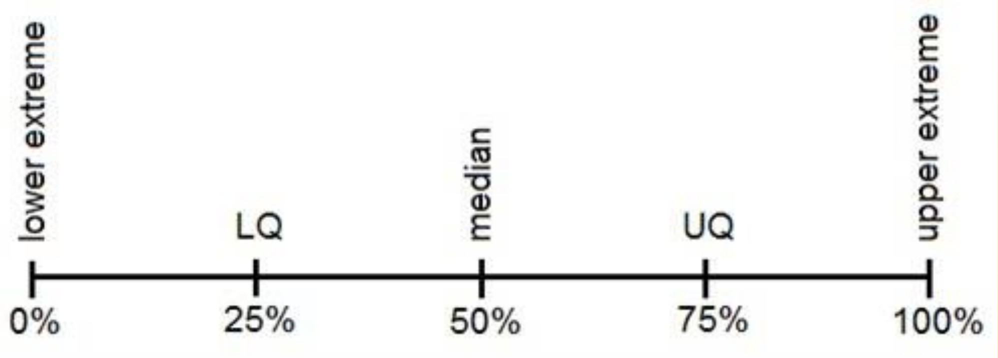
127
131
96
80
93
120
109
162
103
111
109
87
105

Class B Range = 162 - 80 = 82

Class B Range = 80 to 162

Quartiles

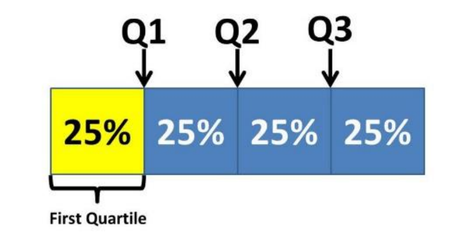




25th percentile is a quartile that divides the first ¼ of cases from the latter ¾.
50th percentile (AKA median) divides the first ½ of cases from the latter ½.
75th percentile is a quartile that divides the first ¾ of cases from the latter ¼.

quantile One of several equal parts of an ordered data set.
3 equal parts = tertiles
4 equal parts = quartiles
5 equal parts = quintiles

Interquartile Range (IQR)



Difference between third & first quartiles

Interquartile Range = $Q_3 - Q_1$

The interquartile range is the distance or range between the 25th percentile and the 75th percentile.

Describes the spread in middle 50%

The “box” in box plots

Not affected by extreme values

The IQR can be reported by giving the 1st and 3rd quartiles themselves, or the difference between them

Example, the 1st and 3rd quartiles are 32 and 60

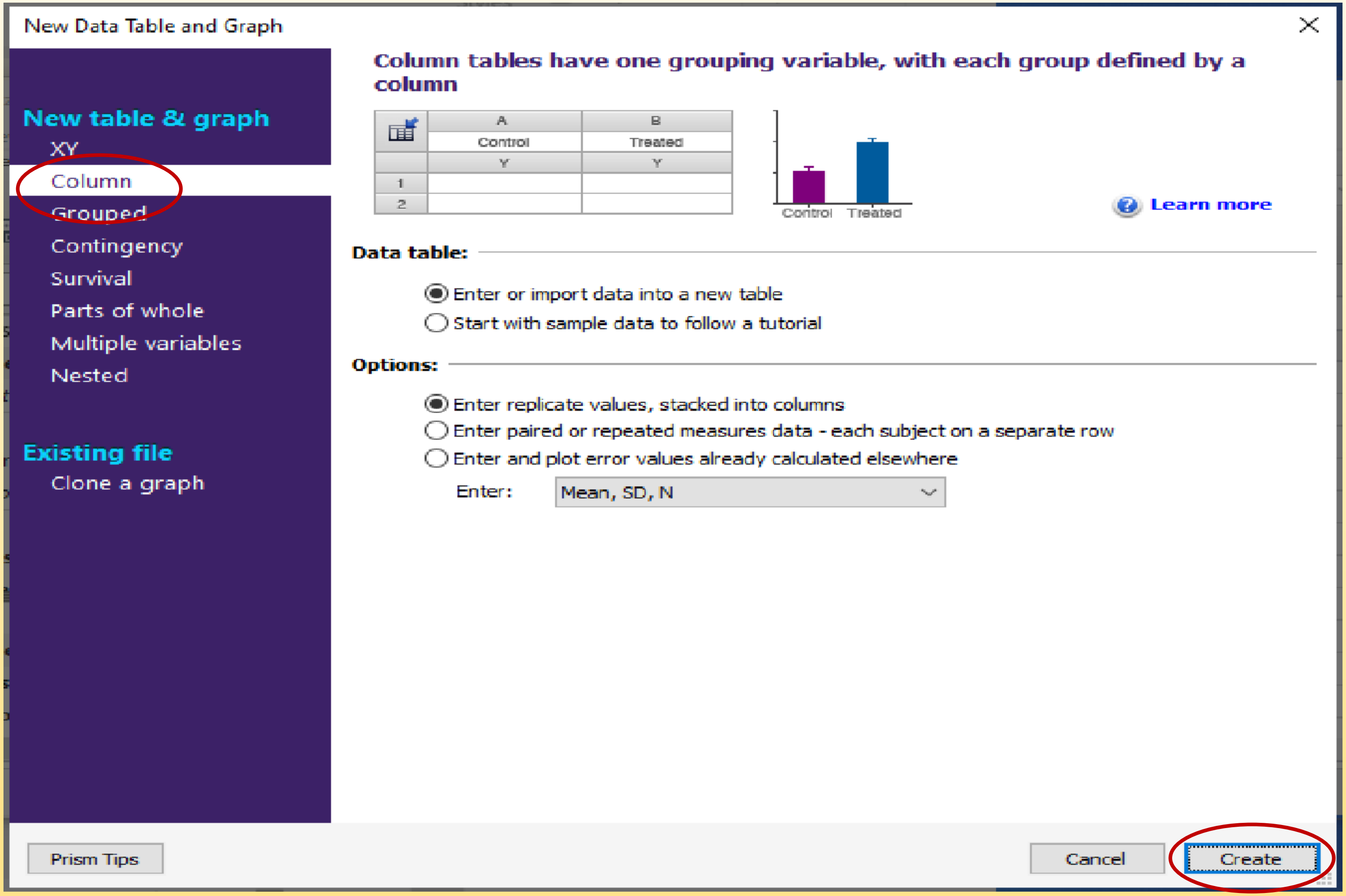
Therefore, *IQR* = 60 - 32 = **28**

Using Prism to Calculate Measures of Central Tendency and Variability

Mean
Median
Mode

Variance
Standard deviation
Range
Quartiles
Interquartile range (IQR)

I will used data from last two classes as a demonstration



Column tables have one grouping variable, with each group defined by a column

	A	B
	Control	Treated
	Y	Y
1		
2		

Data table:

● Enter or import data into a new table

○ Start with sample data to follow a tutorial

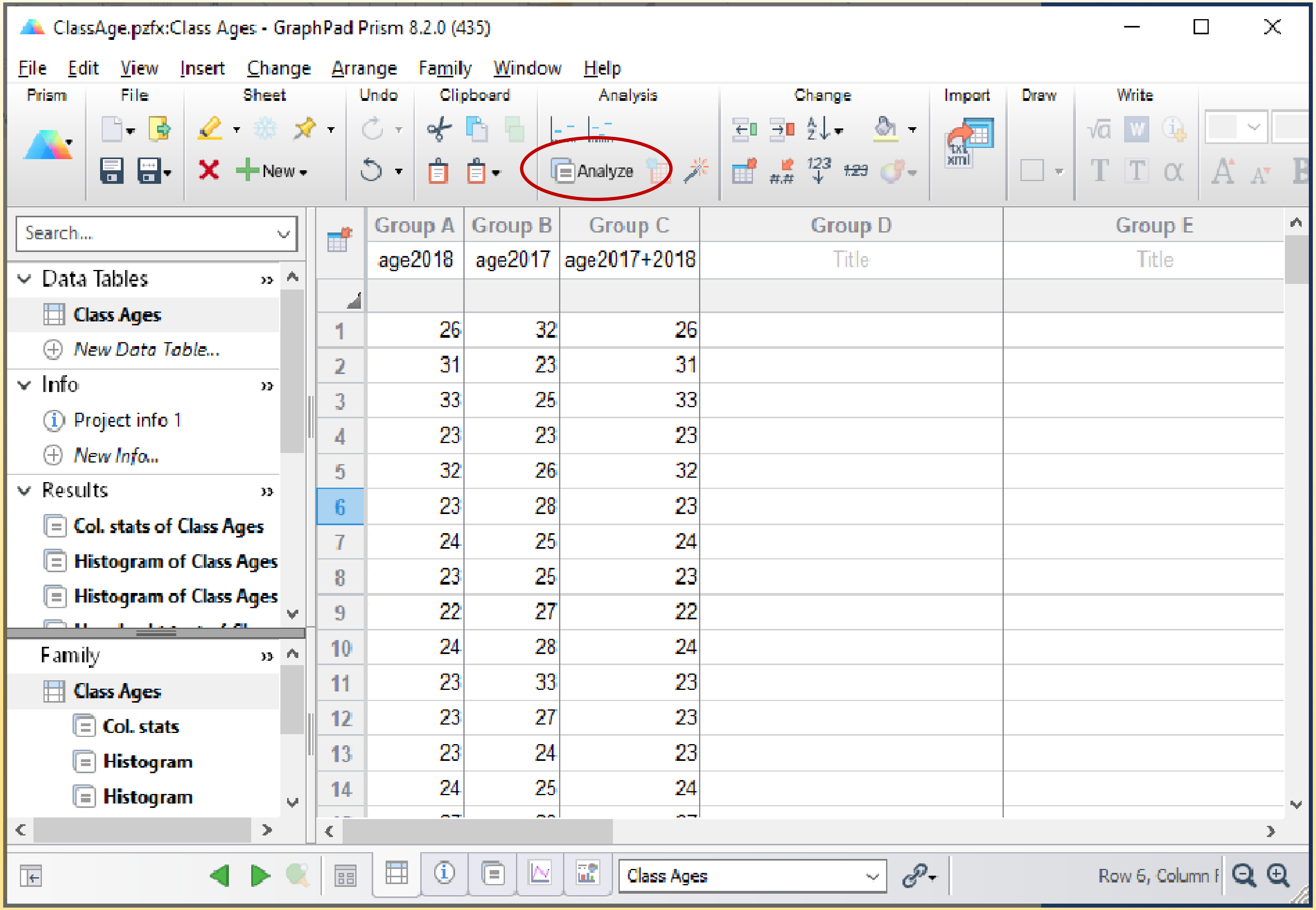
Options:

● Enter replicate values, stacked into columns

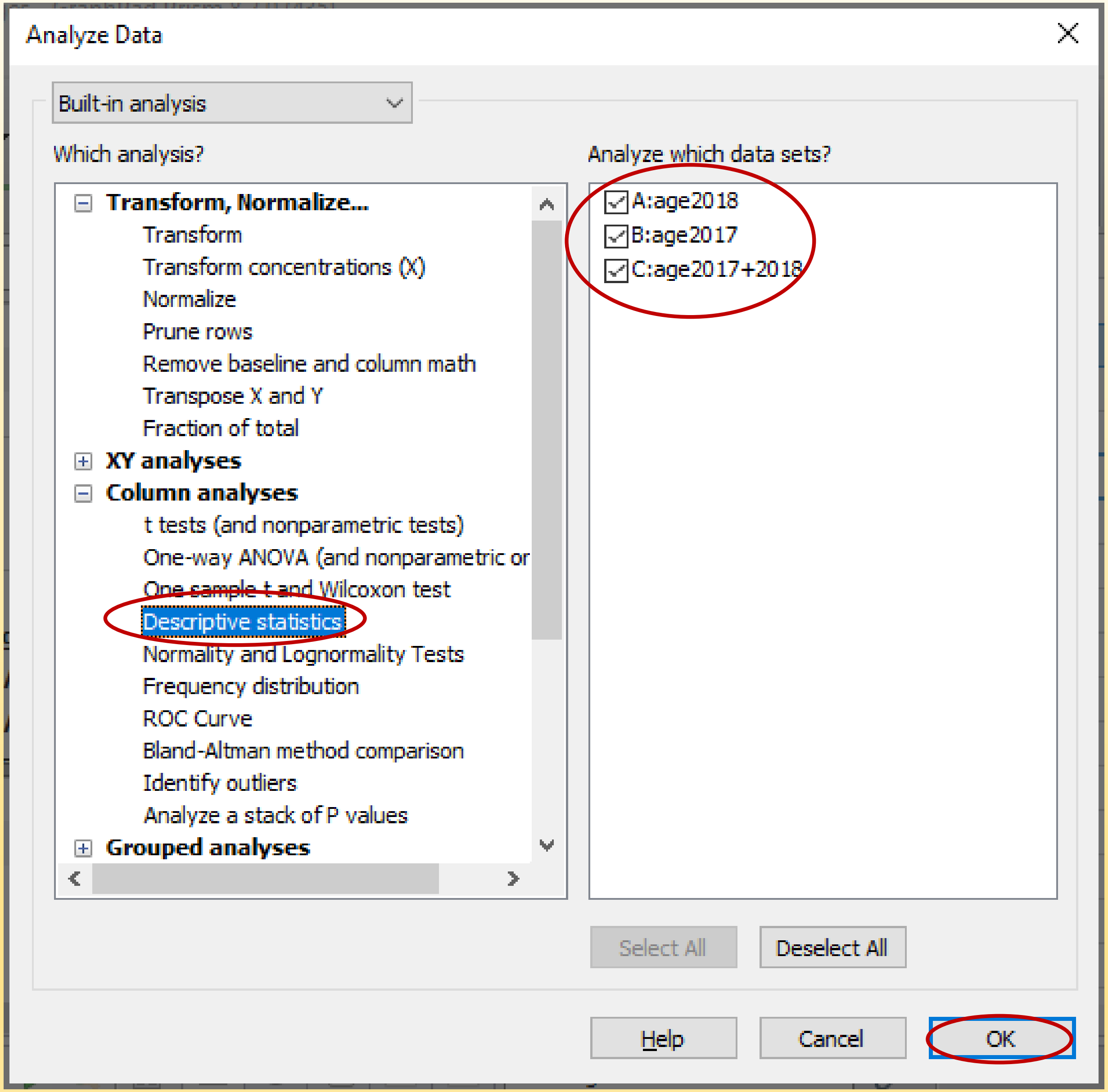
○ Enter paired or repeated measures data - each subject on a separate row

○ Enter and plot error values already calculated elsewhere

Enter: Mean, SD, N



	Group A	Group B	Group C	Group D	Group E
	age2018	age2017	age2017+2018	Title	Title
1	26	32	26		
2	31	23	31		
3	33	25	33		
4	23	23	23		
5	32	26	32		
6	23	28	23		
7	24	25	24		
8	23	25	23		
9	22	27	22		
10	24	28	24		
11	23	33	23		
12	23	27	23		
13	23	24	23		
14	24	25	24		



Analyze Data

Built-in analysis

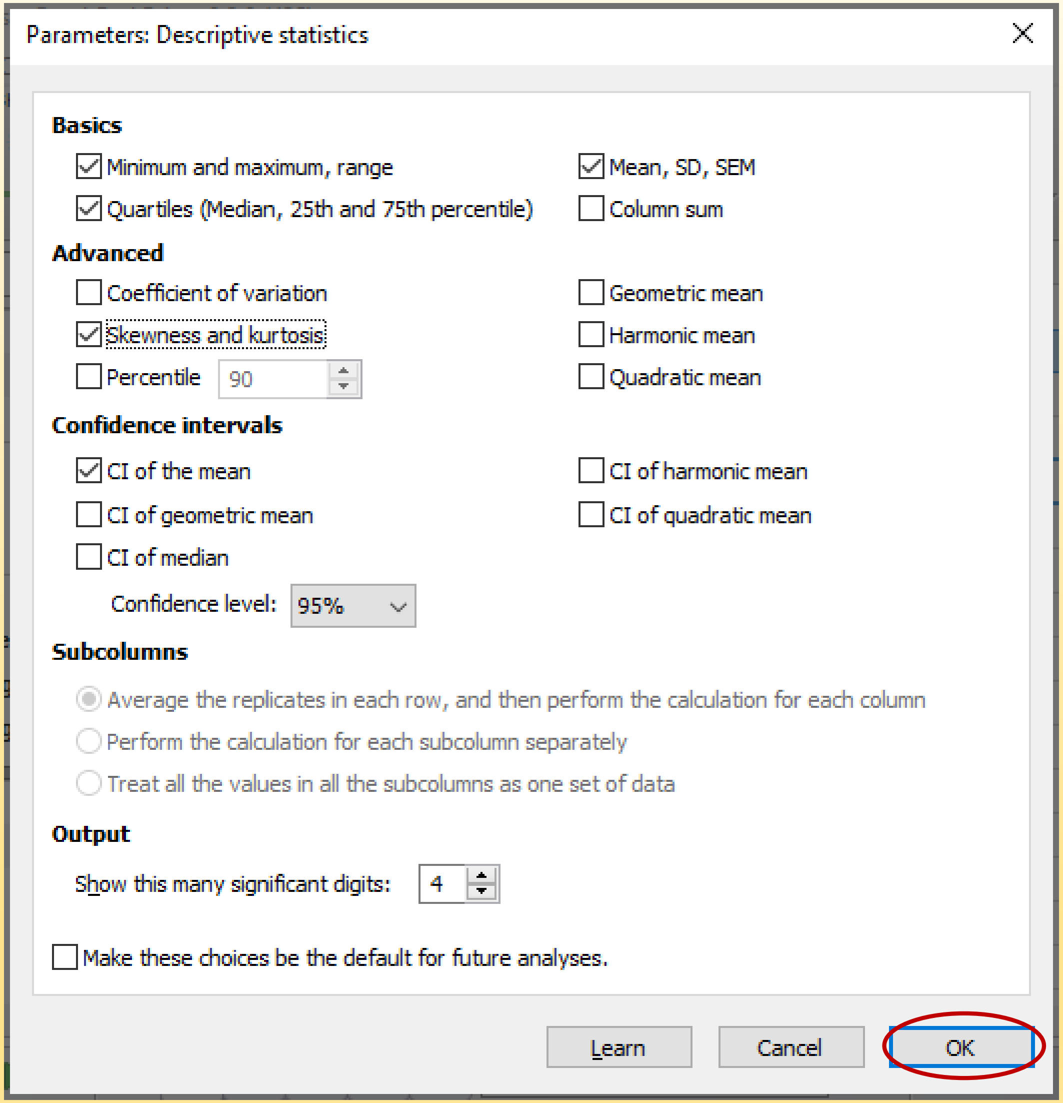
Which analysis?

- ⊟ **Transform, Normalize...**
  - Transform
  - Transform concentrations (X)
  - Normalize
  - Prune rows
  - Remove baseline and column math
  - Transpose X and Y
  - Fraction of total
- ⊞ **XY analyses**
- ⊟ **Column analyses**
  - t tests (and nonparametric tests)
  - One-way ANOVA (and nonparametric or
  - One sample t and Wilcoxon test
  - Descriptive statistics
  - Normality and Lognormality Tests
  - Frequency distribution
  - ROC Curve
  - Bland-Altman method comparison
  - Identify outliers
  - Analyze a stack of P values
- ⊞ **Grouped analyses**

Analyze which data sets?

☑ A:age2018
☑ B:age2017
☑ C:age2017+2018

Some boxes will already be ticked



Parameters: Descriptive statistics

Basics

☑ Minimum and maximum, range
☑ Quartiles (Median, 25th and 75th percentile)
☑ Mean, SD, SEM
□ Column sum

Advanced

□ Coefficient of variation
☑ Skewness and kurtosis
□ Percentile 90
□ Geometric mean
□ Harmonic mean
□ Quadratic mean

Confidence intervals

☑ CI of the mean
□ CI of geometric mean
□ CI of median
□ CI of harmonic mean
□ CI of quadratic mean

Confidence level: 95%

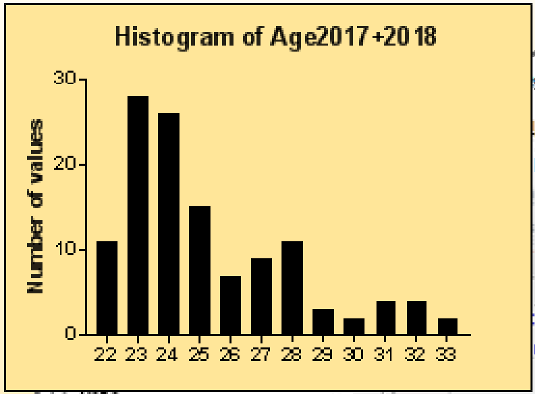
Subcolumns

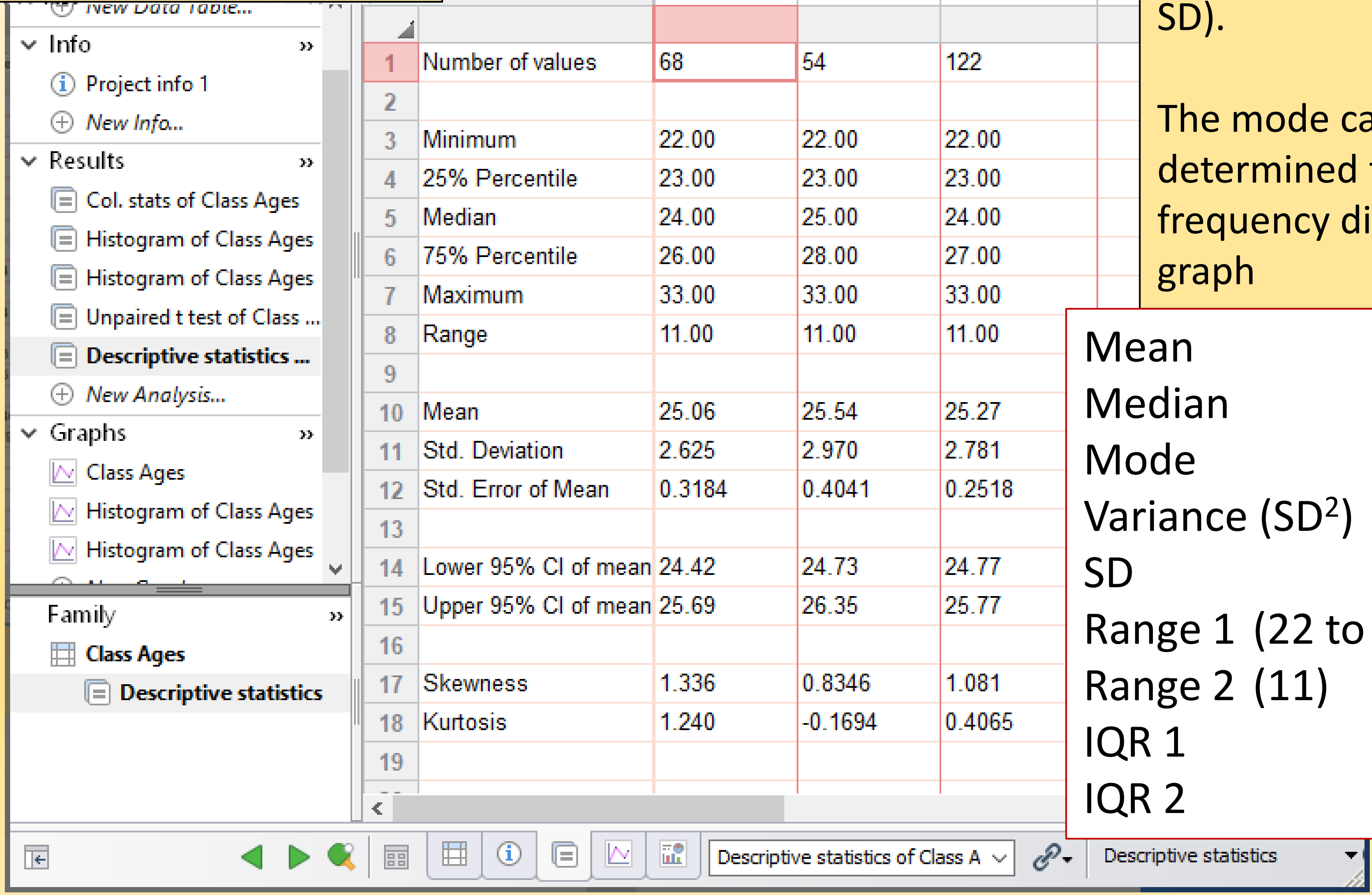
● Average the replicates in each row, and then perform the calculation for each column
○ Perform the calculation for each subcolumn separately
○ Treat all the values in all the subcolumns as one set of data

Output

Show this many significant digits: 4

□ Make these choices be the default for future analyses.





	escriptive statistics	A age2018	B age2017	C age2017+2018
1	Number of values	68	54	122
2				
3	Minimum	22.00	22.00	22.00
4	25% Percentile	23.00	23.00	23.00
5	Median	24.00	25.00	24.00
6	75% Percentile	26.00	28.00	27.00
7	Maximum	33.00	33.00	33.00
8	Range	11.00	11.00	11.00
9				
10	Mean	25.06	25.54	25.27
11	Std. Deviation	2.625	2.970	2.781
12	Std. Error of Mean	0.3184	0.4041	0.2518
13				
14	Lower 95% CI of mean	24.42	24.73	24.77
15	Upper 95% CI of mean	25.69	26.35	25.77
16				
17	Skewness	1.336	0.8346	1.081
18	Kurtosis	1.240	-0.1694	0.4065
19				

Almost everything is here. You will need to calculate the IQR and the variance (from the SD).

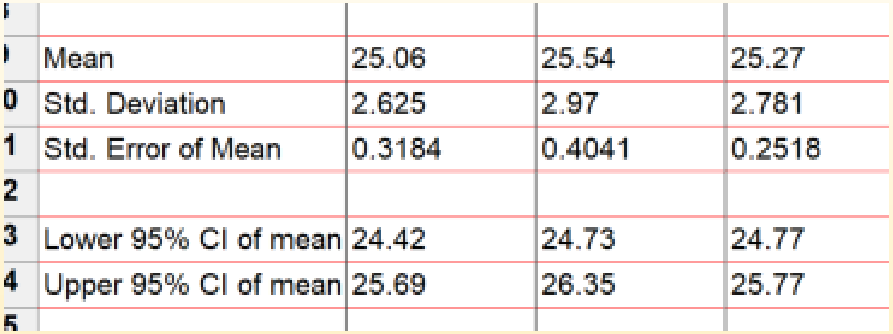
The mode can be determined from the frequency distribution graph

Mean
Median
Mode
Variance (SD^2)
SD
Range 1 (22 to 33)
Range 2 (11)
IQR 1
IQR 2

TRY DOING BEFORE CLASS

Using BIOS 6606 Student Ages 2017to2019.xlsx calculate measures of central tendency and variability for each year

Plot a frequency distribution for each year.



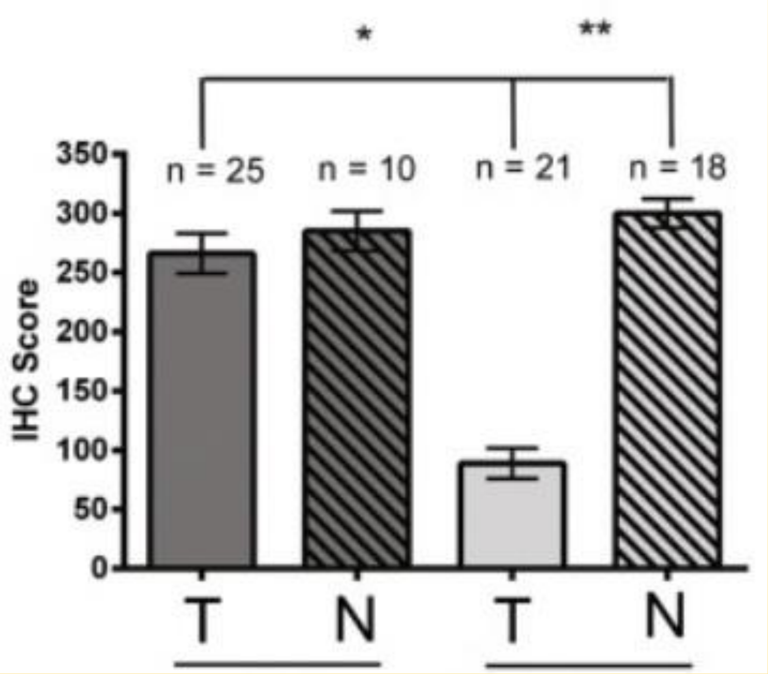
	Mean	25.06	25.54	25.27
0	Std. Deviation	2.625	2.97	2.781
1	Std. Error of Mean	0.3184	0.4041	0.2518
2				
3	Lower 95% CI of mean	24.42	24.73	24.77
4	Upper 95% CI of mean	25.69	26.35	25.77
5				

We know the SD is a measure of data variability.
How does it compare to the SEM and CI?

Crossing over to inferential statistics….

Plus some more information about the normal distribution and the central limit theorem

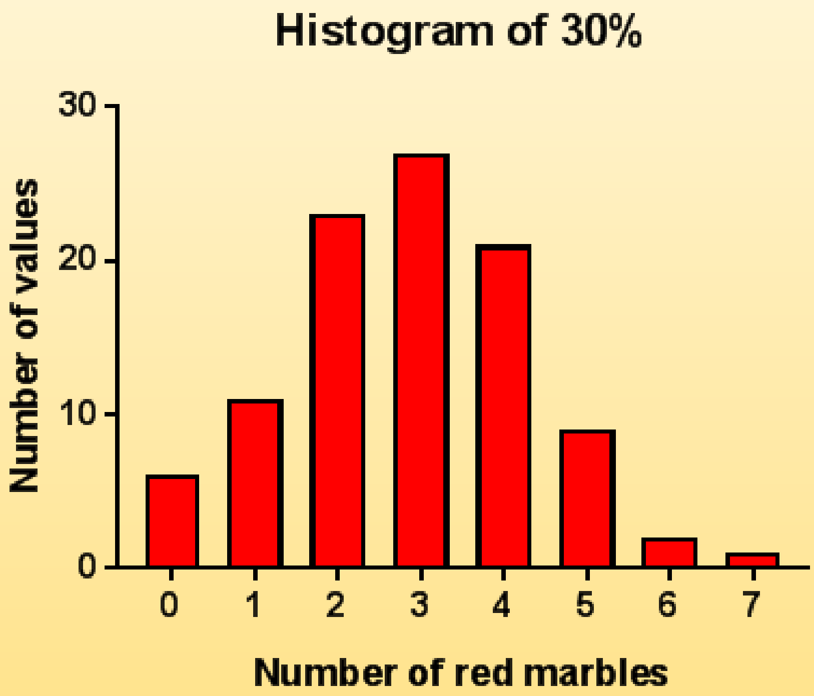
Basic scientists frequently use bar graphs with the Standard Error of the Mean (SEM)



To better understand what the SEM is, first a brief return to sampling distributions

P.S. I don't think this bar chart is not a good example of the best way to present your results

Remember this sampling distribution?
An empirical experiment resulting in this sampling distribution

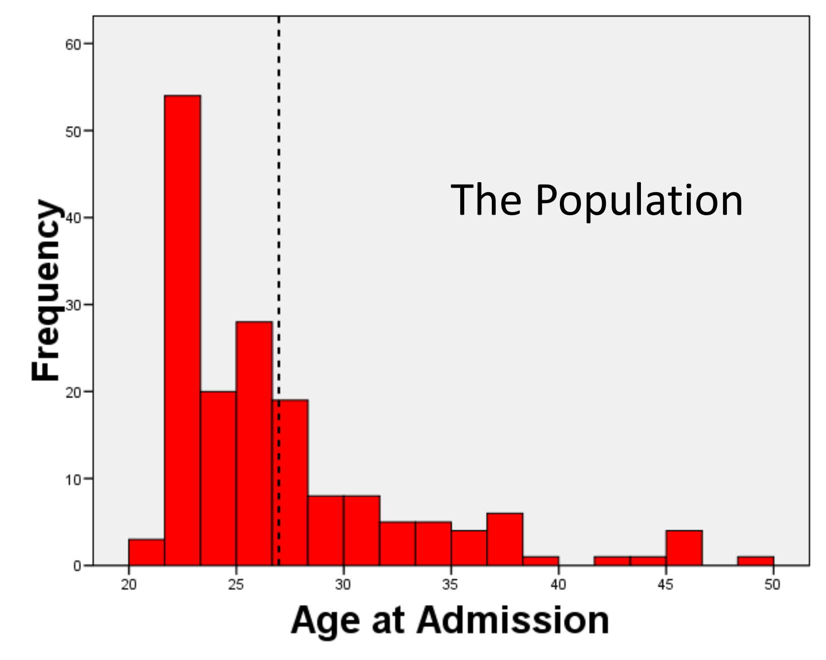


The sampling distribution: the beginning of the SEM

Example: Suppose that our population consists of all students admitted to a medical school in Northern Ontario in its first few years of operation

The variable of interest is **age** upon admission

Age at admission to medical schools, Ontario



Population Parameters

n = 168

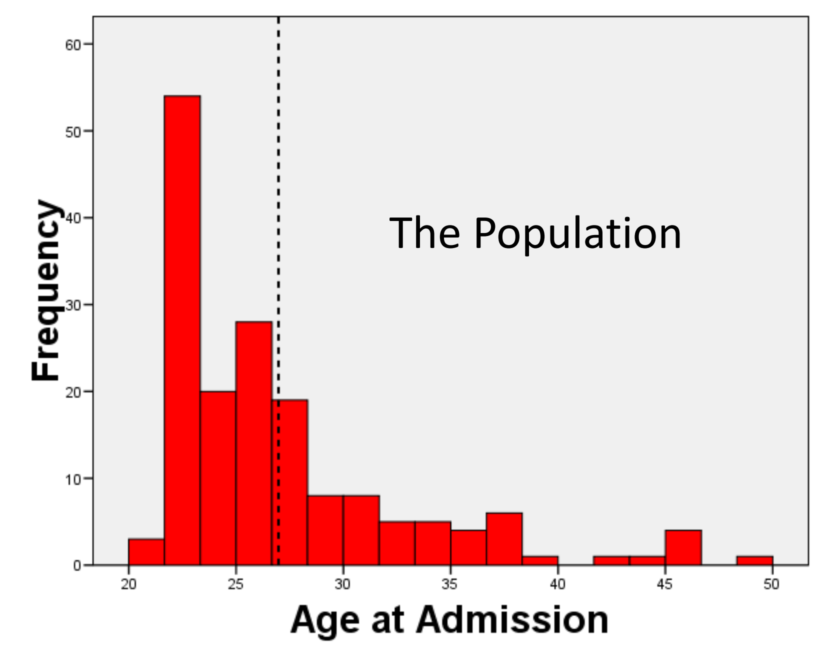
Pop. Mean (μ) = 26.98

Pop. SD (σ) = 5.62

The population parameters are usually not known—we only know them here because it is a contrived example.

https://sites.google.com/a/lakeheadu.ca/bweaver/Home/statistics/introductory-biostatistics

What happens if we take many random samples of 16 from the population to create a frequency distribution of these samples (aka sampling distribution)



Draw a Random Sample

Draw a random sample of *n* = 16 from the population

Calculated the sample mean and sample SD

Sample mean ($\bar{x}$)= 25.06

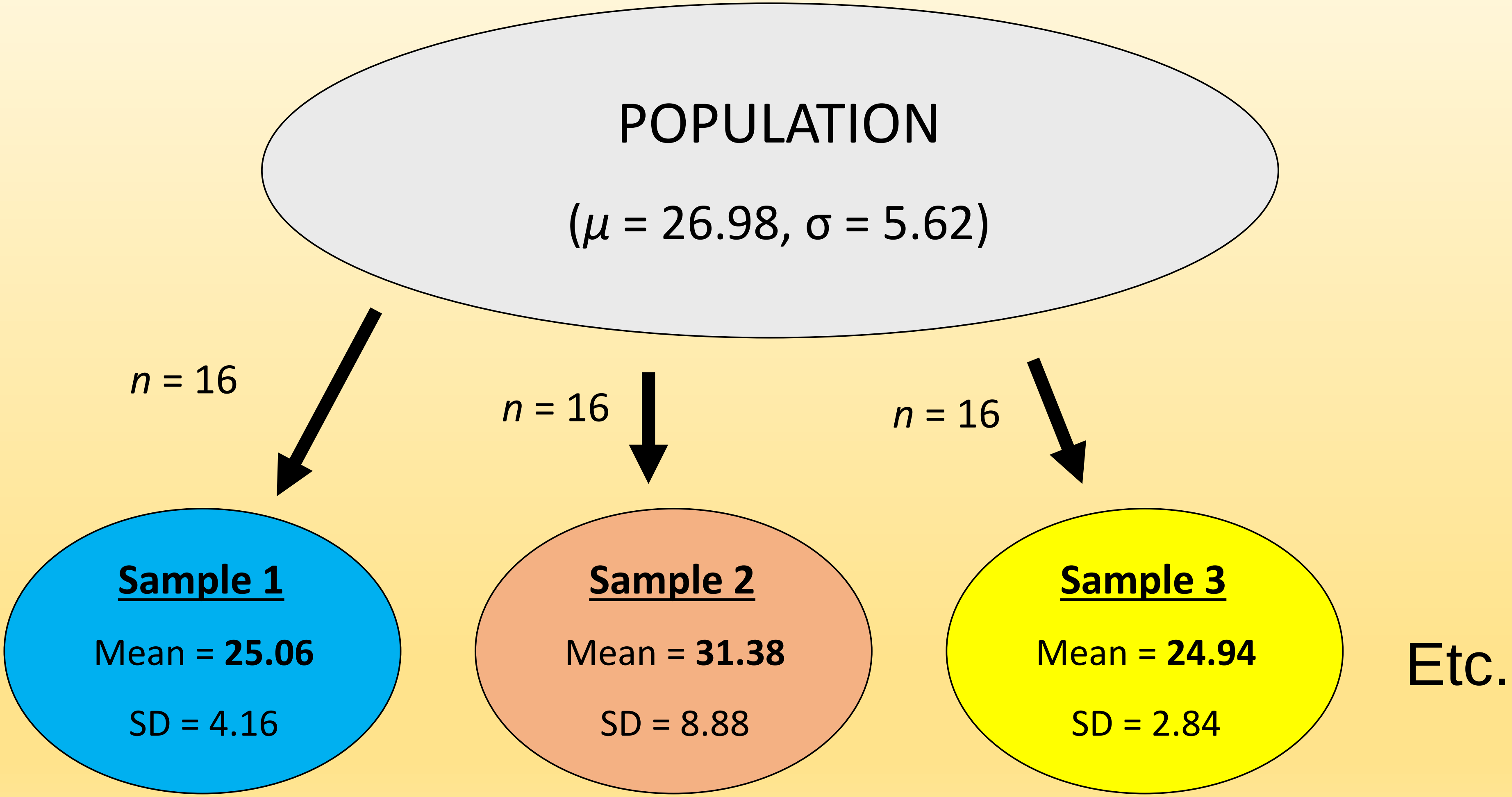
Sample SD (s) = 4.16

}

The population mean (μ=26.98)
and SD (σ=5.62)

Notice that these estimates of the parameters are in error
(aka sampling error).

More Random Samples of n = 16



POPULATION

(μ = 26.98, σ = 5.62)

n = 16

n = 16

n = 16

Sample 1

Mean = **25.06**

SD = 4.16

Sample 2

Mean = **31.38**

SD = 8.88

Sample 3

Mean = **24.94**

SD = 2.84

Etc.

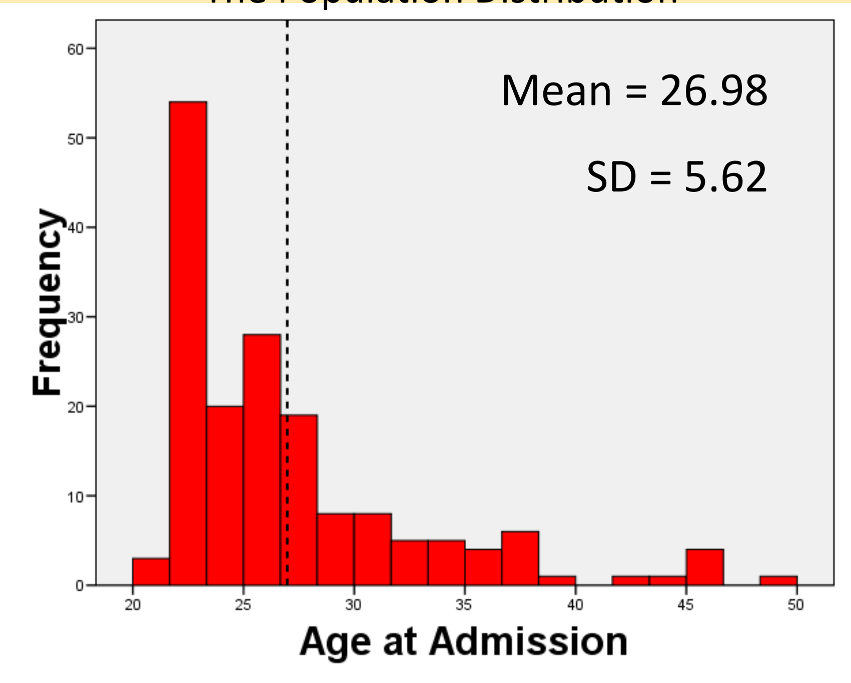
10,000 Random Samples of *n* = 16
An *in silico* experiment

Use a computer to generate **10,000 random samples** of *n* = 16

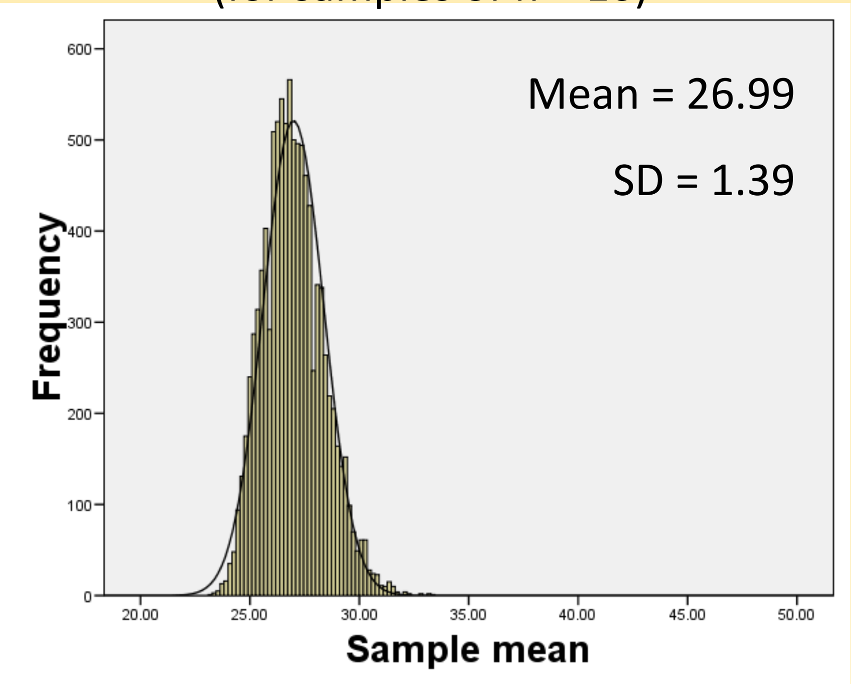
For each sample, compute the mean

Plot a frequency histogram of the 10,000 sample means

The Population Distribution



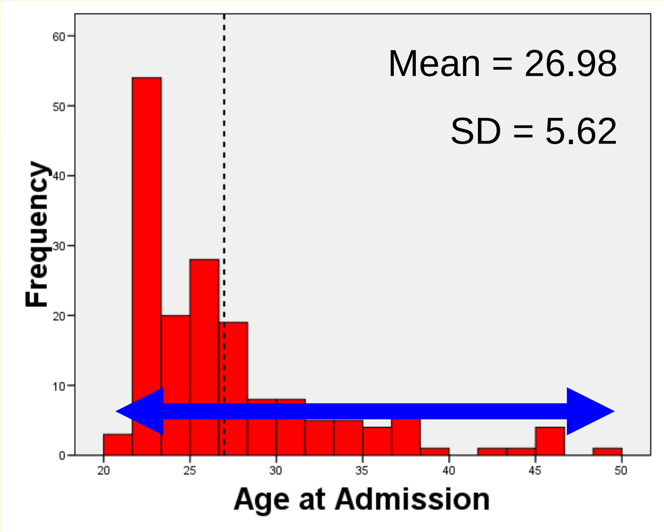
Distribution of 10,000 **Sample Means**
(for samples of n = 16)



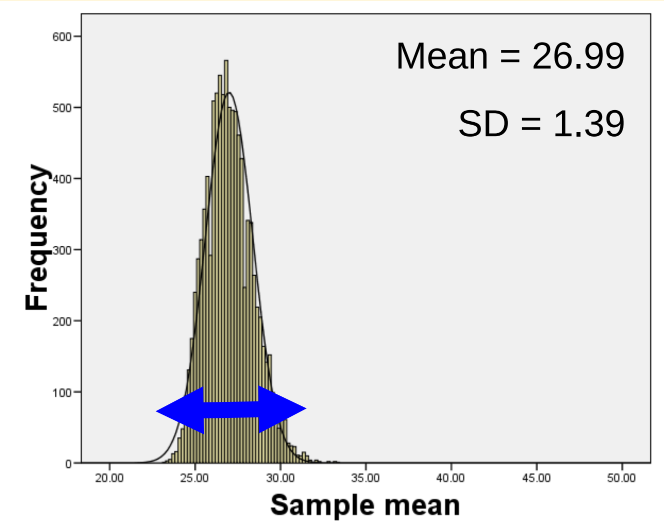
Points to Notice

The distribution of sample means is much less variable than the population distribution

The Population Distribution



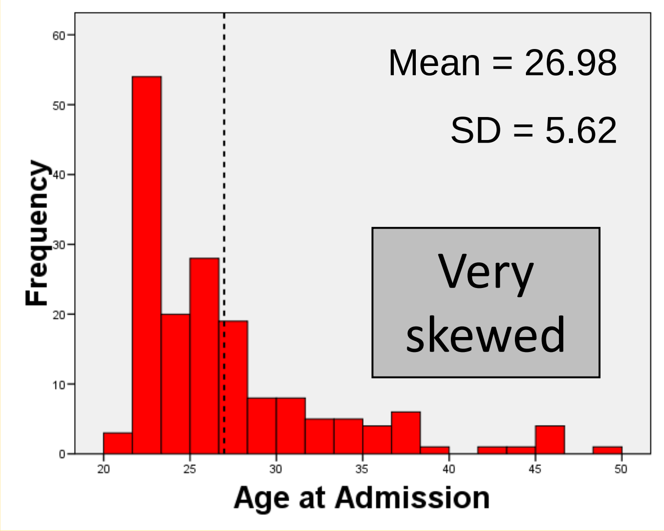
Distribution of **Sample Means** (for samples of n = 16)



Points to Notice

Despite the positive skew in the population, the distribution of 10,000 sample means is symmetrical (approximately normal)

The Population Distribution

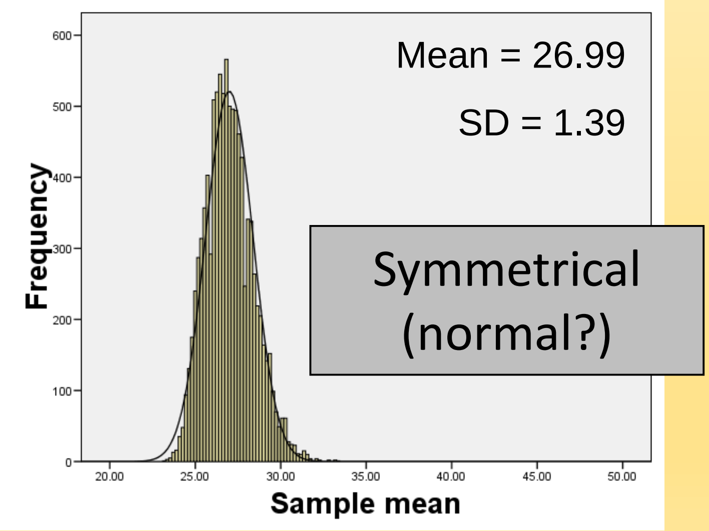


Mean = 26.98

SD = 5.62

Very skewed

Distribution of **Sample Means** (for samples of n = 16)



Mean = 26.99

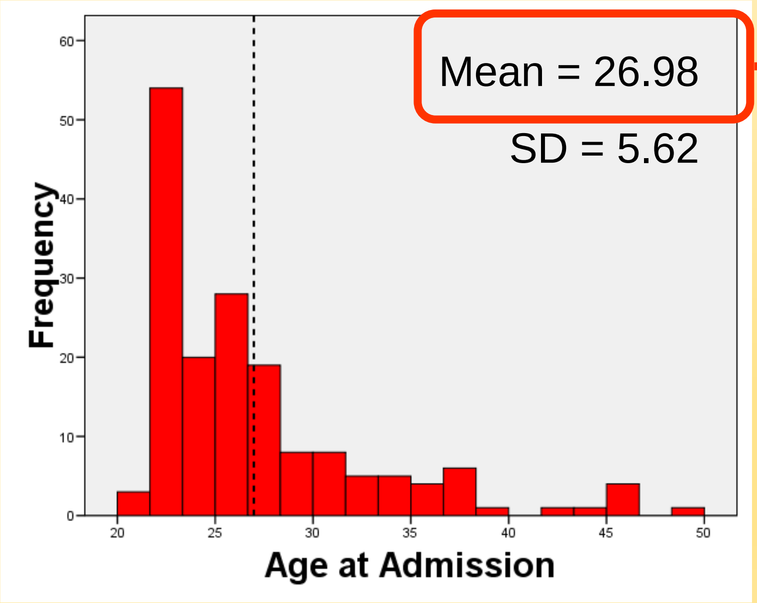
SD = 1.39

Symmetrical (normal?)

Points to Notice

The mean of the 10,000 sample means (26.99) is nearly identical to the population mean (26.98)

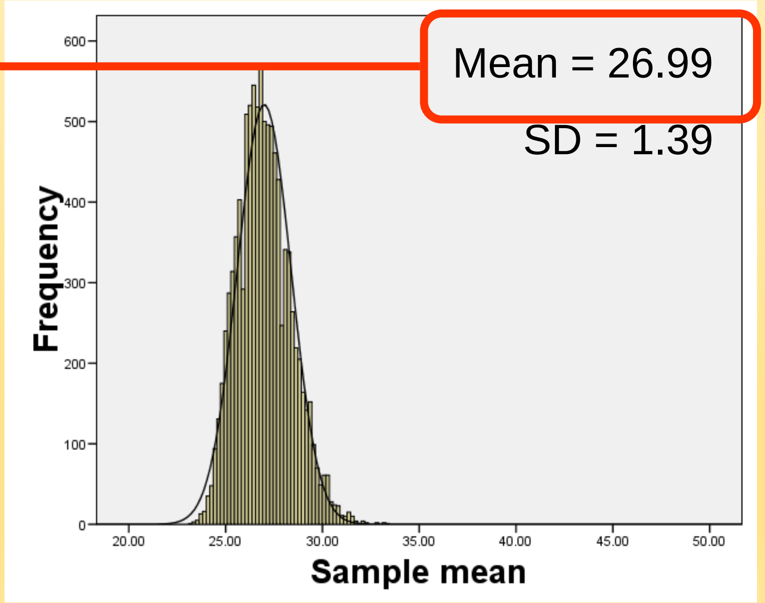
The Population Distribution



Mean = 26.98

SD = 5.62

Distribution of **Sample Means** (for samples of n = 16)



Mean = 26.99

SD = 1.39

Summary of Points to Notice

1. The distribution of sample means is much less variable than the population distribution

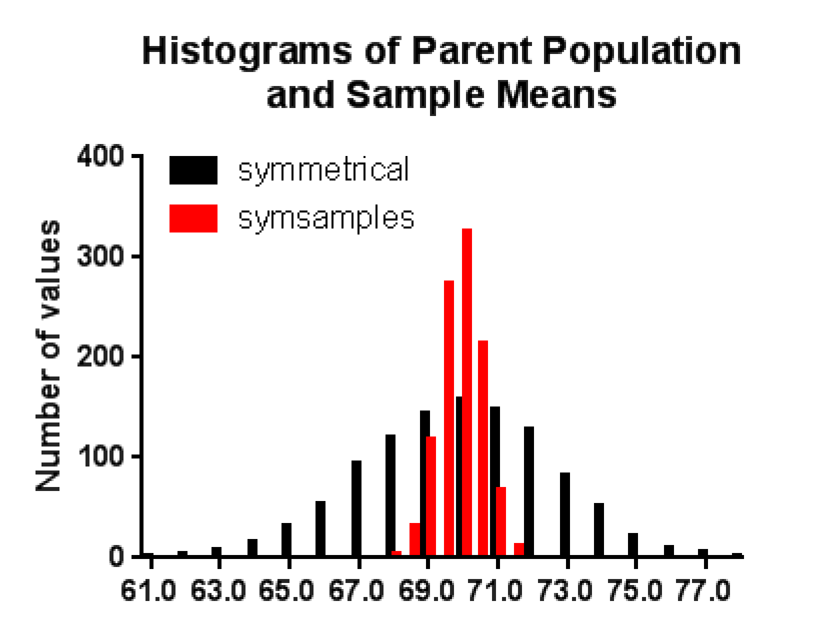
2. Despite the positive skew in the population, the distribution of 10,000 sample means is symmetrical (approximately normal)

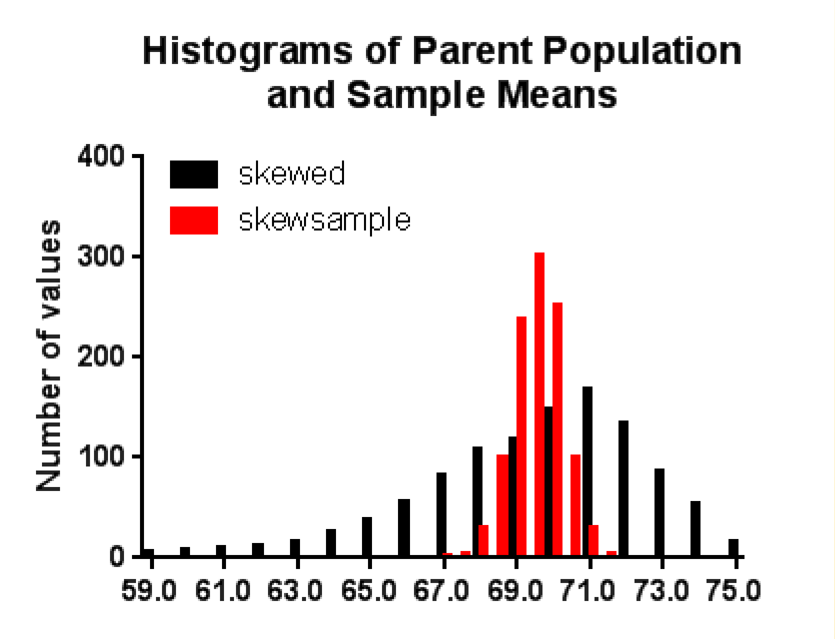
    A demonstration of the central limit theorem

3. The mean of the 10,000 sample means (26.99) is nearly identical to the population mean (26.98)

4. As the size of the sample increases, the SD of the sample means decreases

Results from first year: Using their random samples





	n	Mean	SD
Symmetrical	1075	69.85	2.70
Sample means	1044	69.88	0.60

	n	Mean	SD
Skewed	1075	69.47	3.05
Sample means	1058	69.50	0.66

A great example of the CLT

As quick discussion about the Central Limit Theorem

In probability theory, the central limit theorem (CLT) states that,

> the mean of a sufficiently large number of independent random variables will be usually be approximately normally distributed, regardless of the underlying distribution.

The CLT simplifies problems in statistics by allowing us to work with a sample distribution that is approximately normal.

The Central Limit Theorem

Because of the CLT, even though the shape of the distribution where the data comes from is unknown, the sampling distribution can be treated as normal and variability of the data sample can be determined for statistical tests.

In order for the conclusions of the theorem to hold, the sample size needs to large enough. Rule of thumb:

$$n = 30$$

The Standard Deviation of the Mean of the Sampling Distribution
AKA Standard Error of Mean (SEM)

The SEM is a special kind of standard deviation

It is the standard deviation of a ***sampling distribution of the sample means***

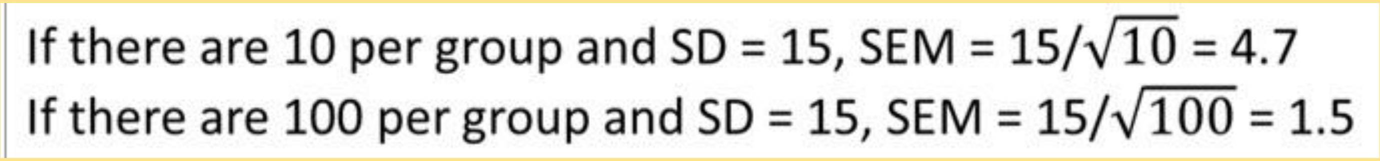
Does not give much information about the variability of your sample

For sample of size **n** with standard deviation (s):

$$SEM = \frac{s}{\sqrt{n}}$$

Expressed as sample mean ± the SEM

As n increases, SEM *estimate* decreases, *i.e.*, estimate of population mean improves



If there are 10 per group and SD = 15, SEM = $15/\sqrt{10}$ = 4.7

If there are 100 per group and SD = 15, SEM = $15/\sqrt{100}$ = 1.5

What Does Standard Error of the Mean Tell You?

The SEM error bars are expected to contain the true population mean ~68% of the time.

Example: You run 100 different experiments of 50 samples each

the true population mean will be within the SEM error bars for 68 of the experiments

the remaining 32 SEM bars will not accurately estimate the truth

Why do people use them then? They’re the smallest!

The Sample SEM vs. the Sample SD

In biomedical journals, SEM and SD are often used interchangeably to express the variability

But they measure different things

SEM quantifies uncertainty in the estimate of the population mean from your sample

is an estimate of how far the sample mean is likely to be from the unknown *population* mean

SD describes variability of the sample data from the sample mean

the degree to which individual values within the sample differ from the calculated *sample* mean

The Sample SEM vs. the Sample SD

The SEM is always smaller than the SD

The SEM gets smaller as your samples get larger

Huge sample = more precision even if the data are scattered

The SD is relatively constant even as you take larger samples

Which do I Use: SD or SEM?

Use the SD to show how widely scattered the measurements are

Better, show a graph of all data points especially if your sample size is small.

Use the SEM to give a sense of how well you have determined the mean (indicates the uncertainty around the estimate of the mean measurement)

If you want to show differences ($p \leq 0.05$) between groups,

Do not use SD error bars

SEM or CI error bars can be useful but are open to interpretation (next lecture – hint CI are better)

Another Uncertainty Measure: Confidence Intervals (CI) of a Mean

A confidence interval gives an ***estimated range of values that is likely to include the unknown population mean***

The estimated range is calculated from a given sample.

The confidence interval (CI) of a point estimate (*i.e.*, mean) from a sample describes the precision of this estimate

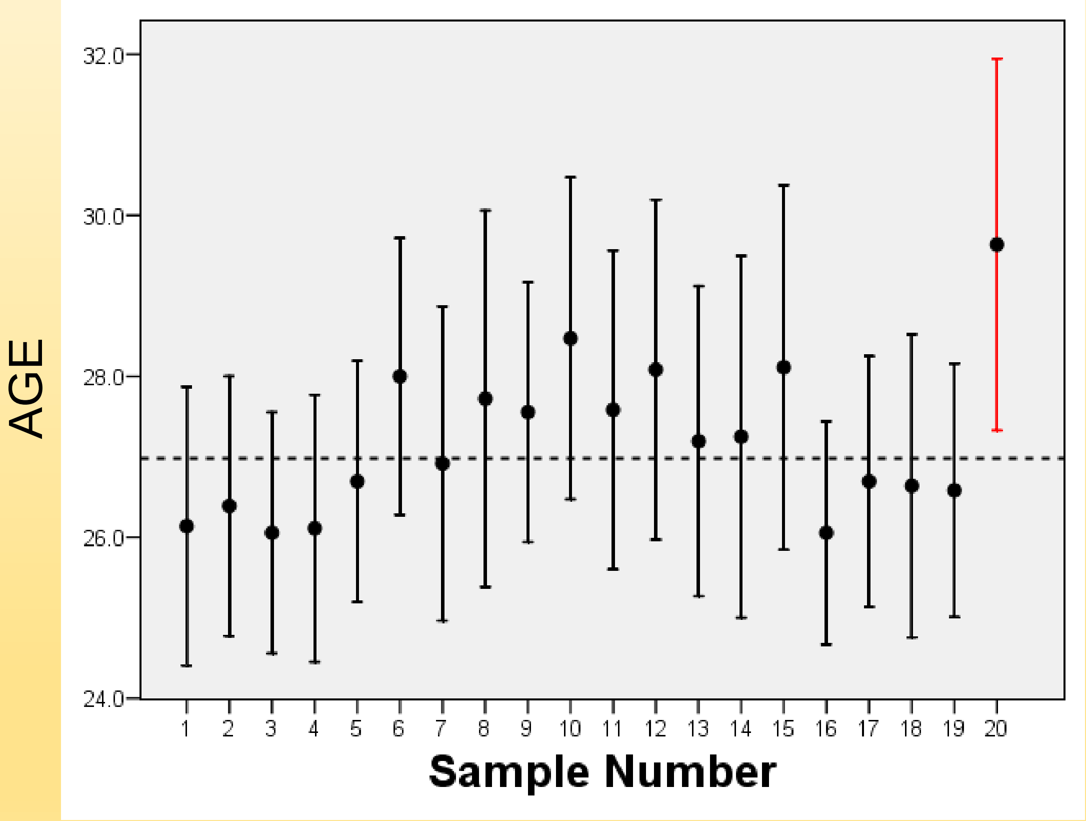
The CI represents a range of values on either side of the estimate. The narrower the CI, the more precise the point estimate

A 95% Confidence Interval is expected to contain the population mean 95% of the time (*i.e.*, of 95% CIs from 100 samples, 95 will contain the population mean)

$$\overline{X} \pm t_{95\%,n-1} SEM$$

95% CI for the Mean from Random Samples

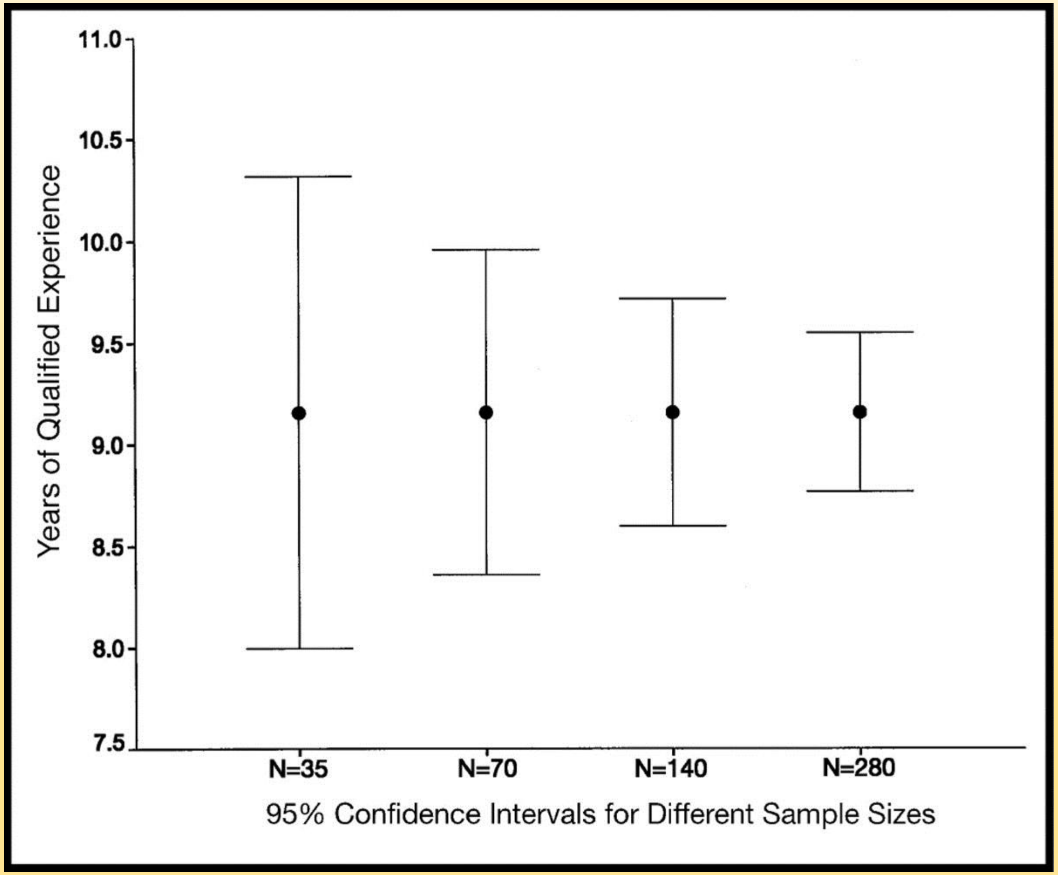
95% CIs for 20 random samples from the population of medical students



Notice that 19 out of 20, or 95% of the 95% confidence intervals, contain the population mean.

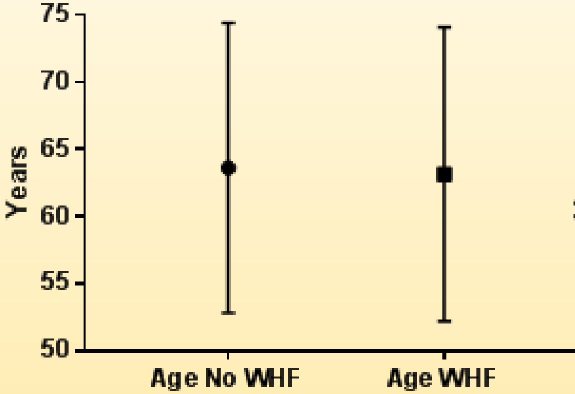
← Population Mean (26.98)

Confidence intervals and mean for progressively larger samples drawn from a single population

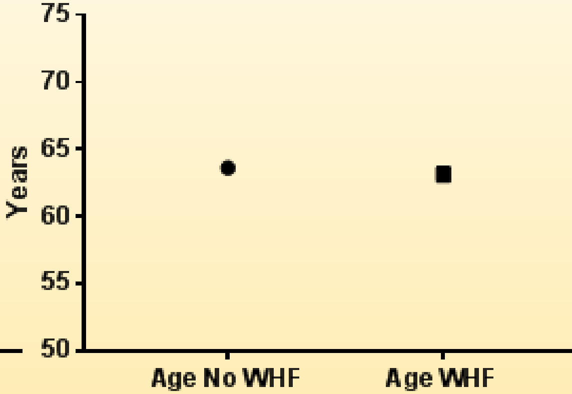


Julius Sim, and Norma Reid PHYS THER 1999;79:186-195

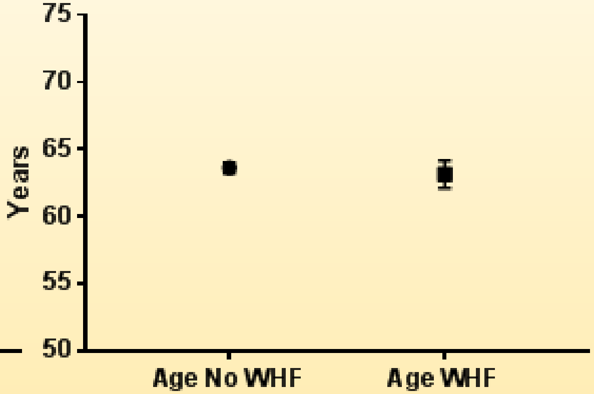
Descriptive


Age (Mean with SD)

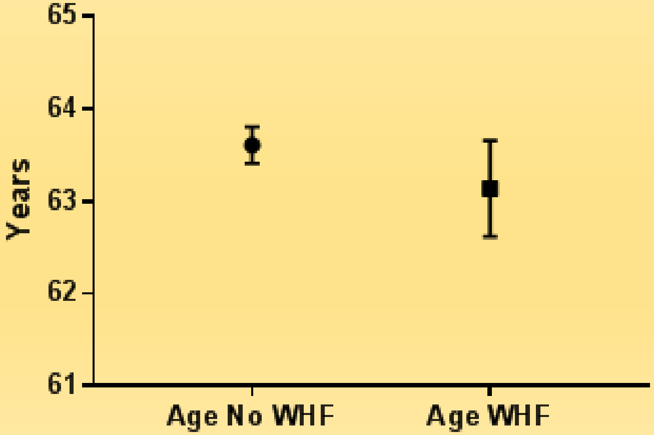
Inferential

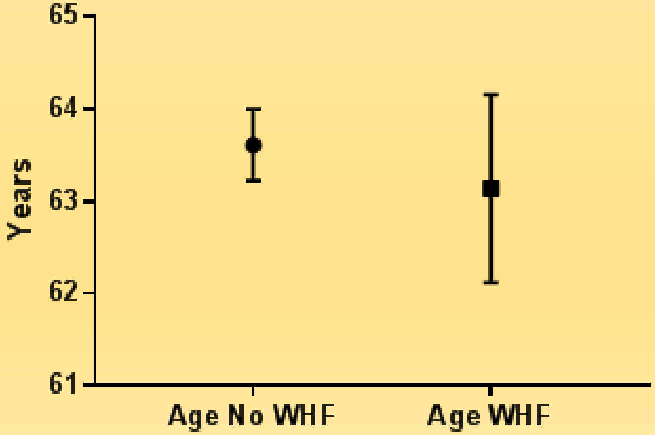

Age (Mean with SEM)

Inferential


Age (Mean with 95% CI)

Expanding the y-axis to see SEM and CI bars


Age (Mean with SEM)


Age (Mean with 95% CI)