Unit 1
Summarizing Data

"It is difficult to understand why statisticians commonly limit their enquiries to averages, and do not revel in more comprehensive views. Their souls seem as dull as the charm of variety as that of the native of one of our flat English counties, whose retrospect of Switzerland was that, if its mountains could be thrown into its lakes, two nuisances would be got rid of at once"

- Sir Francis Galton (England, 1822-1911)

This unit introduces ***variables*** and the variety of ***types of data*** possible (nominal, ordinal, interval, and ratio).

It also introduces ***numerical*** ways of summarizing data. Numerical summaries of data include those that describe central tendency (eg – mode, mean, median), those that describe dispersion (eg – range and standard deviation), and those that describe the shape of the distribution (eg – 25th and 75th percentiles).

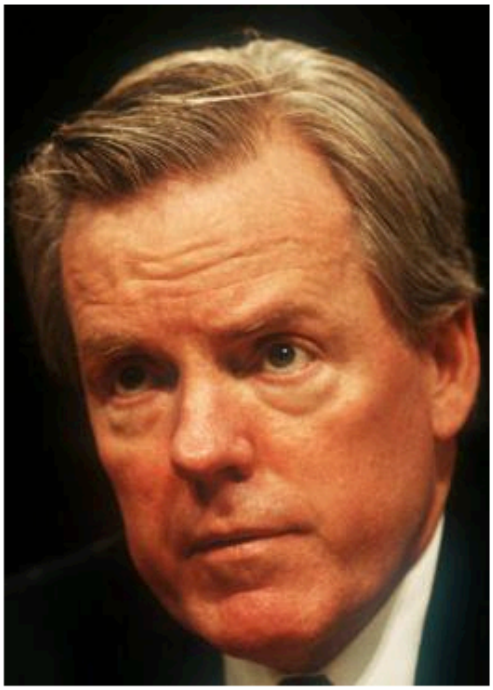
Graphical summaries (***data visualizations***) are introduced in the next unit, Unit 2.

Cheers!

Barksdale on Opinion versus Data

If we have data, let's look at data. If all we have are opinions, let's go with mine.

James Love Barksdale (1943 -)

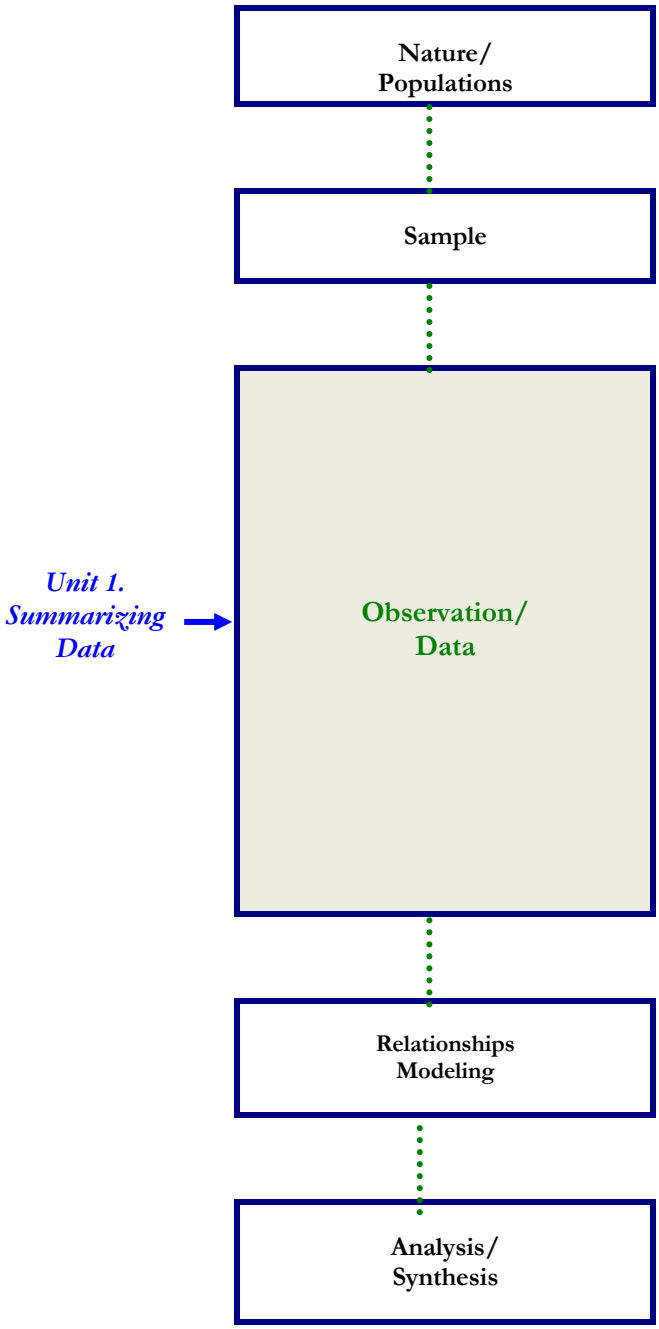


Jim Barksdaler

Source: With permission, download from CAUSEweb.org

Table of Contents

1. Unit Roadmap



Observation and data are not the same. Think about it. What you record as data is ***only some of*** what you actually observe!! Thus, data are the result of selection. ***Data*** are the values you obtain by measurement. A ***variable*** is something whose value can vary.

The ***purpose*** of summarizing data is to communicate the relevant aspects of the data. **Tip – Ask yourself, "what is relevant?"**

Goal - Aim for summaries that are simple, clear, and thorough.

Tip – Avoid misleading summaries.

2. Learning Objectives

When you have finished this unit, you should be able to:

- Explain the distinction between variable and data value.
- Explain the distinction between qualitative and quantitative data.
- Identify the type of variable represented by a variable and its data values.
- Understand and know how to compute: percentile, five number summary, and interquartile range, IQR.
- Understand and know how to compute summary measures of central tendency: mode, mean, median.
- Understand and know how to compute other summary measures of dispersion: range, interquartile range, standard deviation, sample variance, standard error.
- Understand somewhat the distinction between standard deviation and standard error ***Note –We will discuss this again in Unit 3 (Probability Basics) and in Unit 5 (Populations and Samples).***
- Understand the importance of the type of data and the shape of the data distribution when choosing which data summary to obtain.

HOMEWORK due Friday September 23, 2022
Question #1 of 5

Dear all. This question checks that you have read the syllabus! The solutions are in the syllabus.

a) Are the exams "in-class"/proctored or are they take-home?

b) How are the exam grades weighted in the final course grade determination?

c) How are the homeworks graded?

d) Is attendance in Zoom classes required?

e) Your course score is ***not determined by the columns in Blackboard.*** How is the course score determined?

f) How are the final course letter grades determined?

g) Is it possible to obtain the exam questions early?

h) Are late homework and exam submissions allowed (yes or no)?

i) What is the policy on late homework and late exam submissions?

3. Variables and Types of Data

Data can be of different types, and it matters…

Variables versus Data

A **variable** is something whose value can vary. It is a characteristic that is being measured. Examples of variables are:

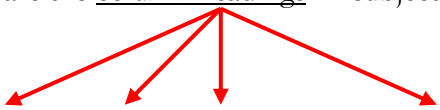
- AGE
- SEX
- BLOOD TYPE

A **data value** is the value of a variable ("realization" - a number or text response) that you obtain upon measurement. Examples of data values are:

- 54 years
- female
- A

Consider the following little data set that is stored in a spreadsheet:

Variables are the column headings – "subject", "age", "sex", "bloodtype"



subject	age	sex	bloodtype
1	54	female	A
2	32	male	B
3	24	female	AB

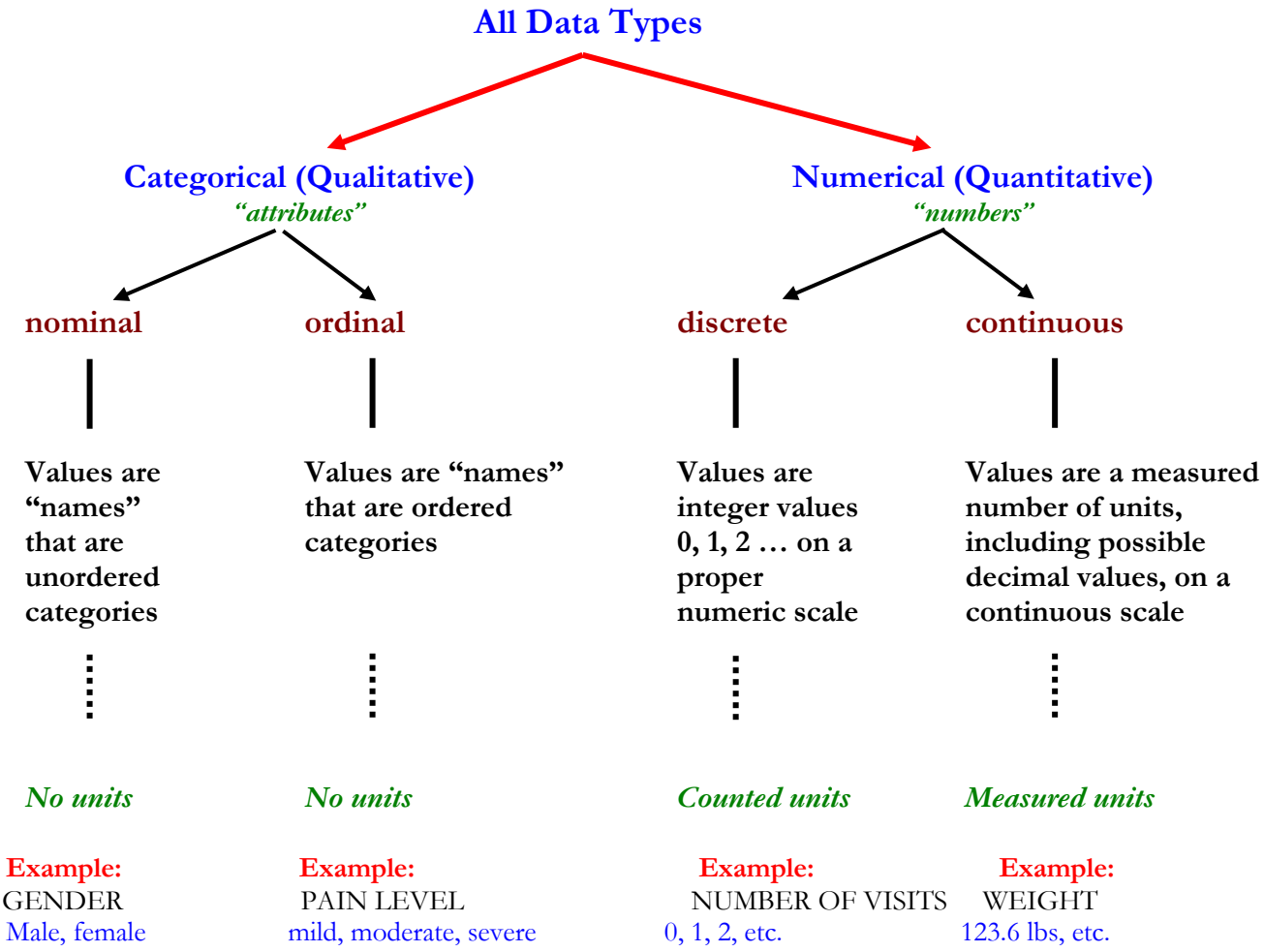
Data values are the table cell entries – "54", "female", "A", etc.

This data table (spreadsheet) has three observations (rows), four variables, and 12 data values.

The different data types are distinct because the scales of measurement are distinct.

There are a variety of schemes for organizing distinct data types. Nevertheless, they all capture the point that differences in scale of measurement are what distinguish distinct data types.

- Whitlock MC and Schluter D (*"The Analysis of Biological Data, Second Edition"*) classify data types as follows:



The distinction between categorical versus numerical is straightforward:

Categorical: Attributes that do ***NOT*** have magnitude on a numerical scale
Numerical: Attributes or scores that ***DO*** have magnitude on a numerical scale

Example - To describe a flower as pretty is a categorical (qualitative) assessment while to record a child's age as 11 years is a numerical (quantitative) measurement.

Consider this …
We can reasonably refer to the child's 22 year old cousin as being twice as old as the 11 year old child whereas we cannot reasonably describe an orchid as being twice as pretty as a dandelion.

We encounter similar stumbling blocks in statistical work. Depending on the type of the variable, its scale of measurement type, some statistical methods are meaningful while others are not.

- CATEGORICAL ▶Nominal Scale: **Values are names which cannot be ordered.**

  Example: Cause of Death
  - Cancer
  - Heart Attack
  - Accident
  - Other

  Example: Gender
  - Male
  - Female

  Example: Race/Ethnicity
  - Black
  - White
  - Latino
  - Other

  Other Examples: Eye Color, Type of Car, University Attended, Occupation

- CATEGORICAL ▶ Ordinal Scale: **Values are attributes (names) that are naturally ordered.**

Example: Size of Container
- Small
- Medium
- Large

Example: Pain Level
- None
- Mild
- Moderate
- Severe

For analysis in the computer, both nominal and ordinal data might be stored using numbers rather than text.

Example of nominal: Race/Ethnicity
- 1 = Black
- 2 = White
- 3 = Latino
- 4 = Other

Nominal - The numbers "1", "2", etc. have NO meaning
They are labels ONLY

Example of ordinal: Pain Level
- 1 = None
- 2 = Mild
- 3 =Moderate
- 4 = Severe

Ordinal – The numbers have LIMITED meaning
4 > 3 > 2 > 1 says ONLY that "severe" is worse than "moderate" and so on. ***You cannot do math on these!***

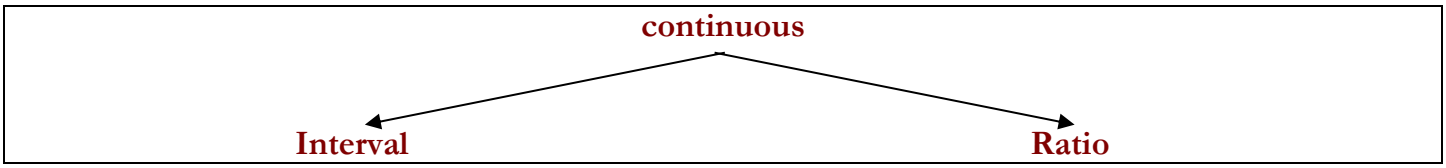
NUMERICAL ▶ Discrete Scale: **Values are counts of the number of times some event occurred and are thus whole numbers: 0, 1, 2, 3, etc.** …

Examples:

Number of children a woman has had
Number of clinic visits made in one year

The numbers are meaningful. We can actually compute with these numbers.

NUMERICAL ▶ Continuous: A further classification of data types is possible for numerical data that are continuous



NUMERICAL ▶ Continuous → Interval **("no true zero")**: Continuous interval data are generally measured on a continuum and differences between any two numbers on the scale are of known size but ***there is no true zero***.

Example: Temperature in °F on 4 successive days

Day:	A	B	C	D
Temp °F:	50	55	60	65

"5 degrees difference" makes sense. For these data, not only is day A with 50° cooler than day D with 65°, but it is 15° cooler. Also, day A is cooler than day B by the same amount that day C is cooler than day D (i.e., 5°).

"0 degrees cannot be interpreted as absence of temperature". In fact, we think of 0 degrees as quite cold! Or, we might think of it as the temperature at which molecules are no longer in motion. Either way, it's not the same as "0 apples" or "0 Santa Claus". Thinking about mathematics, for data that are continuous and interval (such as temperature and time), the value "0" is arbitrary and doesn't reflect absence of the attribute.

NUMERICAL ► Continuous → Ratio **("meaningful zero")**: Continuous ratio data are also measured on a meaningful continuum. The distinction is that ratio data have a meaningful zero point.

Example: Weight in pounds of 6 individuals
136, 124, 148, 118, 125, 142

Note on meaningfulness of "ratio"-
Someone who weighs 142 pounds is two times as heavy as someone else who weighs 71 pounds. This is true even if weight had been measured in kilograms.

- In the sections that follow, we will see that the possibilities for meaningful description (tables, charts, means, variances, etc) are lesser or greater depending on the scale of measurement.
- The chart on the next page gives a sense of this idea.
- For example, we'll see that we can compute relative frequencies for a nominal random variable (eg. Hair color: e.g. "7% of the population has red hair") but we cannot make statements about cumulative relative frequency for a nominal random variable (eg. it would not make sense to say "35% of the population has hair color less than or equal to blonde")

Chart showing data summarization methods, by data type:

	All Data Types			
	Categorical *"qualitative"*		Numerical *"quantitative"*	
Type	Nominal	Ordinal	Discrete	Continuous
Descriptive Methods **Coming soon! Unit 2, Data Visualization**	Bar chart Pie chart - -	Bar chart Pie chart - -	Bar chart Pie chart Dot diagram Scatter plot (2 variables) Stem-Leaf Histogram Box Plot Quantile-Quantile Plot	- - Dot diagram Scatter plot (2 vars) Stem-Leaf Histogram Box Plot Quantile-Quantile Plot
Numerical Summaries **This unit! Unit 1, Summarizing Data**	Frequency Relative Frequency	Frequency Relative Frequency Cumulative Frequency	Frequency Relative Frequency Cumulative Frequency means, variances, percentiles	- - - means, variances, percentiles

Note – This table is an illustration only. It is not intended to be complete.

HOMEWORK Due Friday September 23, 2022
Question #2 of 5

For each of the following variables indicate whether it is quantitative or qualitative and specify the measurement scale that is employed when taking measurements on each:

a) Class standing of members of this class relative to each other.

b) Admitting diagnosis of patients admitted to a mental health clinic.

c) Weights of babies born in a hospital during a year.

d) Gender of babies born in a hospital during a year.

e) Range of motion of elbow joint of students enrolled in a university health sciences curriculum.

f) Under-arm temperature of day-old infants born in a hospital.

4. The Summation Notation

Why this?

The summation notation (and the product notation, by the way) is handy and lots of folks use it! So it's good to know. Quite simply, it is nothing more than a secretarial convenience. We use it to avoid having to write out long expressions.

To get ourselves going,

- Here are five (5) values of age, all in years: 15, 31, 75, 52, and 84
- Now we "tag" or "index" them as follows: X_1=15, X_2=31, X_3=75, X_4=52, X_5=84

Here is how summation notation works:

Instead of writing the sum $x_1 + x_2 + x_3 + x_4 + x_5$,

We write $\sum_{i=1}^{5} x_i$

And here is how product notation works:

Instead of writing out the product of five terms $x_1 * x_2 * x_3 * x_4 * x_5$,

We write $\prod_{i=1}^{5} x_i$

This is actually an example of the product notation

The summation notation

$\sum$ The Greek symbol sigma says "add up some items"

$\sum_{\text{STARTING HERE}}$ Below the sigma symbol is the starting point

$\overset{\text{END}}{\sum}$ Up top is the ending point

Example – Consider the 5 values of age at the top of this page. Using summation notation, what is the sum of the 2nd, 3rd, and 4th values?

x_1=15 x_2=31 x_3=75 x_4=52 x_5=84 →

$$\sum_{i=2}^{4} x_i = x_2 + x_3 + x_4 = 31 + 75 + 52 = 158$$

Additional Resources to Help you Learn Summation Notation

__1. Video. Youtube Tutorial (With apologies - the sound quality is not so great and there is an ad)
PatrickJMT. Summation Notation (Youtube: 10:15)

__2. Columbia University Tutorial
http://www.columbia.edu/itc/sipa/math/summation.html

__3. Khan Academy – Some Exercises to Practice What You Have Learned
https://www.khanacademy.org/math/algebra2/sequences-and-series/copy-of-sigma-notation/e/evaluating-basic-sigma-notation

HOMEWORK Due Friday September 23, 2022
Question #3 of 5

Let $x_1=3$, $x_2=1$, $x_3=4$, and $x_4=6$

3a. Express the following sum in sigma notation and evaluate numerically.

$$(x_1 + x_2 + x_3 + x_4)^2$$

3b. Express the following sum in sigma notation and evaluate numerically.

$$x_1^2 + x_2^2 + x_3^2 + x_4^2$$

3c. Evaluate the following numerically.

$$\Sigma\ (X_i - 1)^2 \text{ for } i=1\ldots4.$$

3d. Evaluate the following numerically.

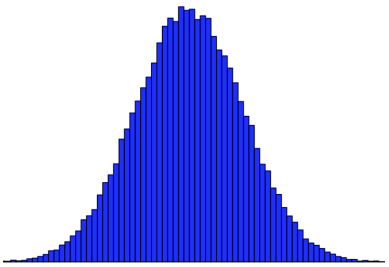
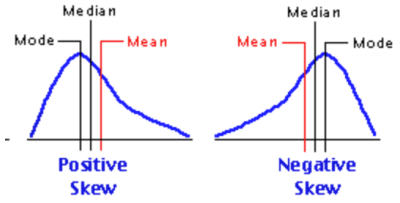
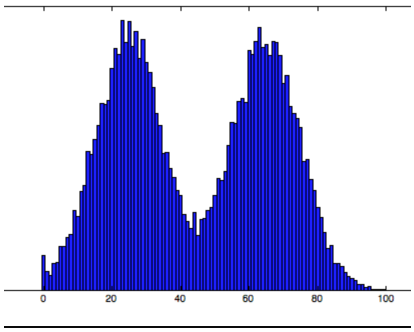
$$\Sigma\ 3X_i \text{ for } i=1\ldots4.$$

5. Numerical Summaries for Quantitative Data - Central Tendency

Among the important tools of description are those that address

- What is typical (location or central tendancy)
- What is the scatter (dispersion)

Recall - "Good" choices for summarizing location and dispersion are not always the same and depend on the pattern of scatter.

		
Symmetric	**Skewed**	**Bimodal**

Mode. The mode is the most frequently occurring value. It is not influenced by extreme values. Often, it is not a good summary of the majority of the data.

Mean. The mean is the arithmetic average of the values. It is sensitive to extreme values.

$$\text{Mean} = \frac{\text{sum of values}}{\text{sample size}} = \frac{\Sigma\ (\text{values})}{n}$$

Median. The median is the middle value when the sample size is odd. For samples of even sample size, it is the average of the two middle values. It is not influenced by extreme values.

We consider each one in a bit more detail …

5a. Mode

Mode. The mode is the most frequently occurring value. It is not influenced by extreme values. Often, it is not a good summary of the majority of the data.

Example

- Data are: 1, 2, 3, 4, 4, 4, 4, 5, 5, 6
- Mode is 4

Example

- Data are: 1, 2, 2, 2, 3, 4, 5, 5, 5, 6, 6, 8
- There are two modes – value 2 and value 5
- This distribution is said to be "bimodal"

Modal Class

- For grouped data, it may be possible to speak of a modal class
- The modal class is the class with the largest frequency

Example – Data set of n=80 values of age (years)

Interval/Class of Values (age, years")	Frequency, f (# times)
31-40	1
41-50	2
51-60	5
61-70	15
71-80	25
81-90	20
91-100	12

- The modal class is the interval of values 71-80 years of age, because values in this range occurred the most often (25 times) in our data set.

5b. Mean

Mean. The mean is the arithmetic average of the values. It is sensitive to extreme values.

$$\text{Mean} = \frac{\text{sum of values}}{\text{sample size}} = \frac{\Sigma \text{ (values)}}{n}$$

Examples: Calculation of a "mean" or "average" is familiar; e.g. -

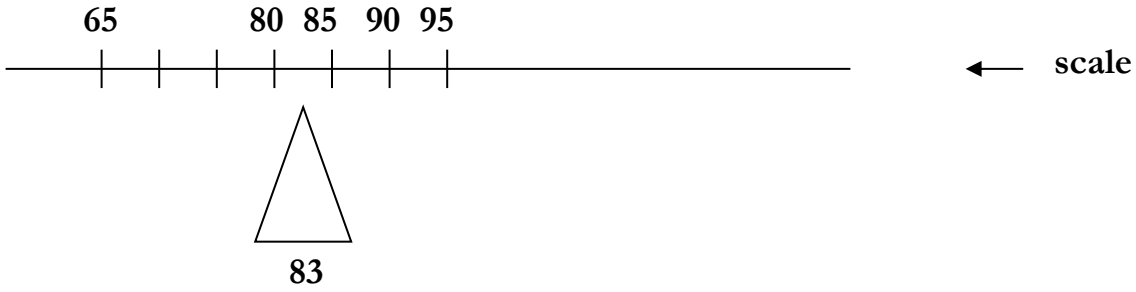
grade point average
mean annual rainfall
average weight of a catch of fish
average family size for a region

A closer look using summation notation introduced on page 15

- Suppose data are: 90, 80, 95, 85, 65
- $\text{sample mean} = \frac{90+80+95+85+65}{5} = \frac{415}{5} = 83$
- sample size, n = 5
- $x_1 = 90, x_2 = 80, x_3 = 95, x_4 = 85, x_5 = 65$
- $\overline{X}$ = sample mean
- $\overline{X} = \frac{\sum_{i=1}^{5} x_i}{n} = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5} = \frac{90+80+95+85+65}{5} = 83$

5c. The Mean as a "Balancing Point" and Introduction to Skewness

The mean can be thought of as a "balancing point", "center of gravity"



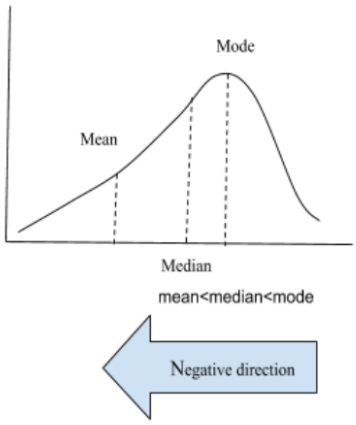
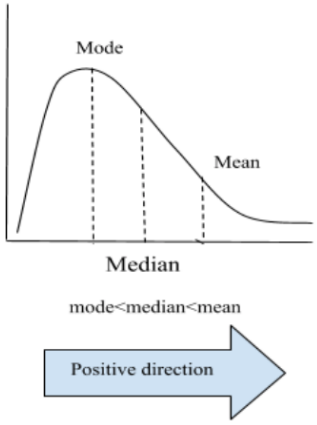
- By "balance", it is meant that the sum of the departures from the mean to the left balance out the sum of the departures from the mean to the right.

  Sum of departures from the mean to the LEFT: (**83**-65) + (**83**-80) = 21
  Sum of departures from the mean to the RIGHT: (85-**83**) + (90-**83**) + (95-**83**) = 21

- In this example, sample mean $\overline{X} = 83$
- **TIP!!** Often, the value of the sample mean is not one that is actually observed

Skewness

When the data are skewed, the mean is "dragged" in the direction of the skewness

Negative Skewness (Left tail) **mean is dragged left**	**Positive Skewness (Right tail)** **mean is dragged right**
	

https://alevelmaths.co.uk/statistics/skewness/

5d. The Mean of Grouped Data

- Sometimes, data values occur multiple times and it is more convenient to group the data than to list the multiple occurrence of "like" values individually.
- The calculation of the sample mean in the setting of grouped data is an extension of the formula for the mean that you have already learned.
- Each unique data value is multiplied by the frequency with which it occurs in the sample.
- **Example**

Value of variable X =	Frequency in sample is =
$X_1 = 96$	$f_1 = 20$
$X_2 = 84$	$f_2 = 20$
$X_3 = 65$	$f_3 = 20$
$X_4 = 73$	$f_4 = 10$
$X_5 = 94$	$f_5 = 30$

$$\text{Grouped mean} = \frac{\sum(\text{data value})(\text{frequency of data value})}{\sum(\text{frequencies})}$$

$$= \frac{\sum_{i=1}^{n}(f_i)(X_i)}{\sum(f_i)}$$

$$= \frac{(20)(96)+(20)(84)+(20)(65)+(10)(73)+(30)(94)}{(20)+(20)+(20)+(10)+(30)}$$

$$= 84.5$$

Tip – The use of the weighted mean is often used to estimate the mean in a sample of data that have been summarized in a frequency table. The values used are the interval midpoints. The weights used are the interval frequencies.

5e. The Median

Median. The median is the middle value when the sample size is odd. For samples of even sample size, it is the average of the two middle values. It is not influenced by extreme values. Recall:

If the sample size n is ODD	$\text{median} = \frac{n+1}{2}\text{th largest value}$
If the sample size n is EVEN	$\text{median} = \text{average of}\left(\left[\frac{n}{2}\right]th, \left[\frac{n+2}{2}\right]th\right)\text{values}$

Example

- Data, from smallest to largest, are: 1, 1, 2, 3, 7, 8, 11, 12, 14, 19, 20
- The sample size, n=11
- Median is the $\frac{n+1}{2}$th largest $= \frac{12}{2} = 6$th largest value
- Thus, reading from left (smallest) to right (largest), the median value is = 8

1, 1, 2, 3, 7, 8, 11, 12, 14, 19, 20

↑

- Five values are smaller than 8; five values are larger.

Example

- Data, from smallest to largest, are: 2, 5, 5, 6, 7, **10, 15**, 21, 22, 23, 23, 25
- The sample size, n=12
- Median =

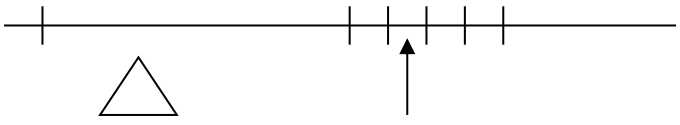
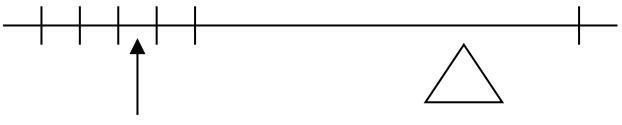
$$\text{average}\left[\frac{n}{2}\text{th largest},\ \frac{n+2}{2}\text{th largest}\right] = \text{average}\left[\text{of 6th and 7th largest values}\right]$$

- Thus, median value is = the average of [10, 15] = 12.5

Skewed Data – When the data are skewed the median is a better description of the majority than the mean

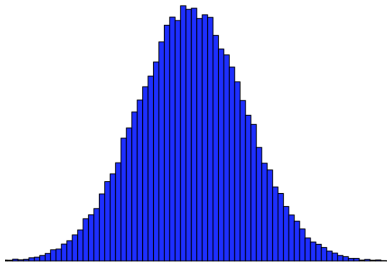
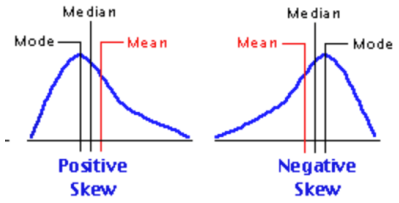
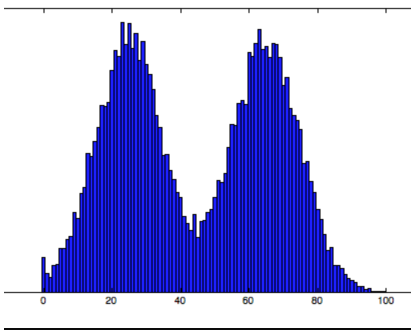
Example

- Data are: 14, 89, 93, 95, 96
- Skewness is reflected in the outlying low value of 14
- The sample mean is 77.4
- The median is 93

Negative Skewness (Left tail)	Positive Skewness (Right tail)
MEAN < Median	**MEAN > Median**
 **Mean is dragged left** **MEDIAN**	 **MEDIAN** **Mean is dragged right**

6. Numerical Summaries for Continuous Data - Dispersion

There are choices for describing ***dispersion***, too. As before, a "good" choice will depend on the shape of the distribution.

		
Symmetric	**Skewed**	**Bimodal**

6a. Variance

Two quick reminders: (1) a parameter is a numerical fact about a population (eg – the average age of every citizen in the United States population); (2) a statistic is a number calculated from a sample (eg – the average age of a random sample of 50 citizens).

Population Mean, μ. One example of a parameter is the population mean. It is written as μ and, for a finite sized population, it is the average of all the values for a variable, taken over all the members of the population.

Population Variance, σ^2. The population variance is also a parameter. It is written as σ^2 and is a summary measure of the squares of individual departures from the mean *in a population*. If we're lucky and we're dealing with a population that is finite in size (yes, it's theoretically possible to have a population of infinite size … more on this later) and of size N, there exists a formula for population variance. This formula makes use of the mean of the population which is represented as μ.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$$

How to interpret the *population* variance: It is the average of the individual squared deviations from the mean. Think of it as answering the question "Typically, how scattered are the individual data points?"

Sample variance, s^2 A sample variance is a statistic; thus, it is a number calculated from the data in a sample. The sample variance is written as S^2 and is a summary measure of the squares of individual departures from the *sample mean* in a *sample*. For a simple random sample of size n (recall – we use the notation "N" when we speak of the size of a finite population and we use the notation "n" when we speak of the size of a sample)

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{(n-1)}$$

Notice that the formula for the sample variance is very similar to the formula for a finite population variance… (1) N is replaced by (n-1) and the (2) μ is replaced by $\bar{X}$. The idea here is

- We are replacing the population **N** by the "sample size minus 1" **(n-1)**
- We are replacing the population mean μ with the sample mean $\bar{X}$

Why (n-1) and not simply (n):

This has to do with the long run average of S^2 being equal to its target $\sigma^2 = \dfrac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}$

6b. Standard Deviation

Standard Deviation, s. The population standard deviation (σ) and a sample standard deviation (S or SD) are the square roots of σ^2 and S^2. As such, they are additional choices for summarizing variability. The advantage of the square root operation is that the resulting summary has the same scale as the original values.

$$\textbf{Sample Standard Deviation (S or SD)} = \sqrt{\frac{\sum(X-\bar{X})^2}{n-1}}$$

Disparity between individual and average …	$(X-\bar{X})$
Disparity between individual and average …	$(X-\bar{X})^2$
The average of these …	$\dfrac{\sum(X-\bar{X})^2}{n}$
The sample variance S^2 is an "almost" average	$S^2 = \dfrac{\sum(X-\bar{X})^2}{n-1}$
The related measure S (or SD) returns measure of dispersion to original scale of observation …	$\text{S or SD} = \sqrt{\dfrac{\sum(X-\bar{X})^2}{n-1}}$

Example of Sample Variance (S^2) and Standard Deviation (S) Calculation –

Consider the following sample of survival times (X) of n=11 patients after heart transplant surgery. Interest is to calculate the sample variance and standard deviation.

- Patients are identified numerically, from 1 to 11.
- The survival time for the "ith" patient is represented as X_i for i= 1, …, 11.

Patient Identifier, "i"	Survival (days), X_i	Mean for sample, $\overline{X}$	Deviation , $(X_i - \overline{X})$	Squared deviation $(X_i - \overline{X})^2$
1	135	161	-26	676
2	43	161	-118	13924
3	379	161	218	47524
4	32	161	-129	16641
5	47	161	-114	12996
6	228	161	67	4489
7	562	161	401	160801
8	49	161	-112	12544
9	59	161	-102	10404
10	147	161	-14	196
11	90	161	-71	5041
TOTAL	**1771**		**0**	**285236**

Dear Class,
Are you new to this sort of table and not quite sure how to navigate? No worries! Consider the first row:
Key – Patient #1 is the person for whom i=1. Patient #1 survived 135 days. So we write X_1 = 135. And so on….

- $\sum_{i=1}^{11} X_i = 1771 \text{ days}$
- **Sample mean is** $\overline{X} = \frac{1771}{11} = 161$ **days**
- **Sample variance is** $S^2 = \frac{\sum_{i=1}^{11} (X_i - \overline{X})^2}{n-1} = \frac{285236}{10} = 28523.6$ **days²**
- **Sample standard deviation is** $s = \sqrt{s^2} = \sqrt{28523.6} = 168.89$ **days**

6c. Median Absolute Deviation About the Median (MADM)

Median Absolute Deviation about the Median (MADM) - Another measure of variability is helpful when we wish to describe scatter among data that is skewed.

Recall that the median is a good measure of location for skewed data because it is not sensitive to extreme values.

Distances are measured about the median, not the mean.

We compute deviations rather than squared differences.

Thus

Median Absolute Deviation about the Median (MADM)

$$\text{MADM} = \text{median of } [\ |X_i - \text{median of } \{X_1,\ldots,X_n\}\ |\ ]$$

Example.

Original data: { 0.7, 1.6, 2.2, 3.2, 9.8 }

Median = 2.2

X_i	$\lvert X_i - \text{median} \rvert$
0.7	1.5
1.6	0.6
2.2	0.0
3.2	1.0
9.8	7.6

MADM = median { 0.0, 0.6, 1.0, 1.5, 7.6 } = 1.0

6d. Standard Deviation (S or SD) versus Standard Error (SE)

Tip **– The standard deviation (s or sd) and the standard error (se or sem) are often confused.**

> The standard deviation (SD or S) addresses questions about variability of individuals in nature (imagine a collection of individuals), whereas
>
> The standard error (SE) addresses questions about the variability of a summary statistic among many replications of your study (imagine a collection of values of a sample statistic such as the sample mean that is obtained by repeating your whole study over and over again)

The distinction has to do with the idea of **sampling distributions** which are introduced on page 34 (stay tuned!) and which are re-introduced several times throughout this course. Consider the following illustration of the idea.

Example

Suppose you conduct a study that involves obtaining a simple random sample of size n=11. Suppose further that, from this one sample, you calculate the sample mean (*note – you might have calculated other sample statistics, too, such as the median or sample variance*). Now imagine replicating the entire study 5000 times. You would then have 5,000 sample means, each based on a sample of size n =11.

If instead of replicating your study 5000 times, the study were replicated infinitely many times, the resulting collection of infinitely many sample means has a name: the **sampling distribution of** $\bar{X}_{n=11}$. Notice the subscript "n=11". This is a reminder to us that the particular study design that we have replicated infinitely many times calls for drawing a sample of size n=11 each time.

So what? Why do we care?
We care because, often, we're interested in knowing if the results of our one study conduct are similar to what would be obtained if someone else were to repeat it!!

Distinction between Standard Deviation (s or sd) and Standard Error of the Mean (se or sem).

We're often interested in the (theoretical) behavior of the **sample mean** $\bar{X}_n$ from one replication of our study to the next.

So, whereas, the typical variability among individual values can be described using the standard deviation (SD).

The typical variability of the sample mean from one replication of the study is described using the standard error (SE) of the mean:

$$SE(\bar{X}) = \frac{SD}{\sqrt{n}}$$

Note* – *A limitation of the SE is that it is a function of both the natural variation (SD in the numerator) and the study design (n in the denominator). More on this later!

Example, continued

Previously, we summarized the results of one study that enrolled n=11 patients after heart transplant surgery. For that one study, we obtained an average survival time of $\overline{X} = 161$ days.

What happens if we repeat the study? What will our next $\overline{X}$ be? Will it be close? How different will it be? We care about this question because it pertains to the generalizability of our study findings.

> The behavior of $\overline{X}$ from one replication of the study to the next replication of the study is referred to as the sampling distribution of $\overline{X}$.
>
> (We could just as well have asked about the behavior of the median from one replication to the next (sampling distribution of the median) or the behavior of the SD from one replication to the next (sampling distribution of SD).)

Thus, interest is in a measure of the "noise" that accompanies $\overline{X} = 161$ days. The measure we use is the standard error measure. This is denoted SE. For this example, in the heart transplant study

$$SE(\overline{X}) = \frac{SD}{\sqrt{n}} = \frac{168.89}{\sqrt{11}} = 50.9$$

We interpret this to mean that a similarly conducted study might produce an average survival time that is near 161 days, give or take 50.9 days.

6e. A Feel for Sampling Distributions

Source: Thinkstock

Sample

$\bar{X}_n$ = sample average

Sample

$\bar{X}_n$ = sample average

.....

sample

$\bar{X}_n$ = sample average

Top Picture

Here is a **population** of **individuals in nature.**

Each individual in this population has their own value of some variable X.

To make this concrete, suppose X = 2021 income ($)

Suppose we want to know the average income in this population

μ = average income ($) over entire population

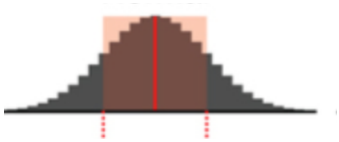
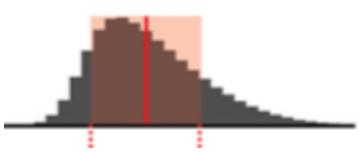
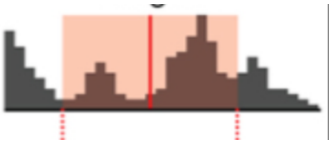
Bottom Picture

Here we imagine 3 separate samples drawn from the population, each with sample size = n.

We have 3 separate averages

$\bar{X}_n$ = average 2021 income in the sample, each based on a sample size of n

Source/Population Distribution. The source/population distribution is the pattern of scatter of the **individual** incomes x among the entirety of individuals in the population in nature. It could be anything! Here are four (4) possibilities. In each picture: x-axis (horizontal) = income and y-axis (vertical) = how often

			
The incomes x are distributed symmetrically about some central value	Or, maybe the distribution is mostly symmetric but there are some with really large incomes (tail to the right)	Or, maybe for every distinct income, the proportion with that income is the same	Or, maybe the distribution of income is just some weird pattern

Source: https://mat117.wisconsin.edu/wp-content/uploads/2014/12/section7-1.png

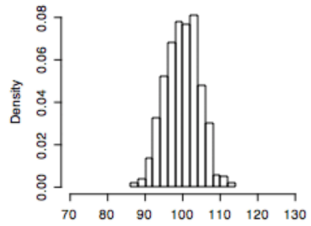
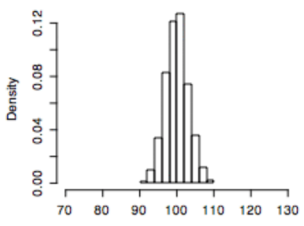
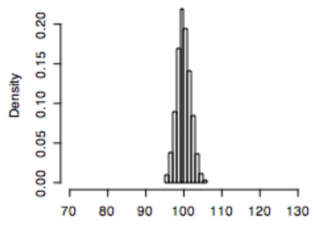
Distribution of all possible sample means.

Hack: When we talk about **all possible individuals in nature** we refer to this as the **population distribution of X** But, when we talk about **all possible sample means** of $\bar{X}_n$, we refer to this as the **sampling distribution** of $\bar{X}_n$

Sampling Distribution of $\bar{X}_n$ ***= Collecting together all possible*** $\bar{X}_n$ ***…***

This the result of repeating the sampling game over and over and over for forever…

Sampling Distribution. The sampling distribution refers to the distribution of some calculated statistic, taken over all possible samples drawn from the source population in nature. Here we are talking about the sampling distribuiton of the sample mean $\bar{X}_n$.

		
Collection of all possible $\bar{X}_n$ when sample size for each sampling is **n=5**	Collection of all possible $\bar{X}_n$ when sample size for each sampling is **n=10**	Collection of all possible $\bar{X}_n$ when sample size for each sampling is **n=30**

Source: https://mse.redwoods.edu/darnold/math15/UsingRInStatistics/CentralLimit3.php

There are lots of sampling distributions, actually.

Sampling distribuiton of the mean. So far, we have imagined calculating the sample mean for a sample of data. From there, we imagined all possible sample means that would be produced by doing the sampling of size=n over and over and over for forever, each time calculating a new sample mean.. When we collected them all, the result was the sampling distribution of the mean.

Sampling distribution of the variance. But we might have instead calculated the sample variance for a sample of data. From there, we can just as easily imagine all possible sample variances that would be produced by doing the sampling of size=2 over and over for forever, each time calculating a new sample variance.

And so on and so on… "sampling distribution of the median", "sampling distribution of the estimated slope", you get the idea…

Another perspective on standard deviation versus standard error is the following:

Standard Deviation	***Standard Error***
• Describes variation in values of ***individuals***.	• Describes variation in values of a ***statistic*** from one conduct of study to the next. • Often, it is the variation in the ***sample mean*** that interests us.
• In the population of *individuals*: σ	• In the population of all possible ***sample means*** ("sampling distribution of mean"): $\sigma/\sqrt{n}$
• **Our "guess" is S**	• **Our "guess" of the SE of the sample mean is** $S/\sqrt{n}$

6f. The Coefficient of Variation

The **coefficient of variation** is the ratio of the standard deviation to the mean of a distribution.

- It is a measure of the spread of the distribution relative to the mean of the distribution
- In the population, coefficient of variation is denoted ξ and is defined

$$\xi = \frac{\sigma}{\mu}$$

- The coefficient of variation ξ can be estimated from a sample. Using the hat notation to indicate "guess". It is also denoted CV

$$\text{cv} = \hat{\xi} = \frac{S}{\overline{X}}$$

Example – "Cholesterol is more variable than systolic blood pressure"

	S	$\overline{X}$	$\text{cv} = \hat{\xi} = s/\overline{x}$
Systolic Blood Pressure	**15 mm**	**130 mm**	**.115**
Cholesterol	**40 mg/dl**	**200 mg/dl**	**.200**

Example – "Diastolic is relatively more variable than systolic blood pressure"

	S	$\overline{X}$	$\text{cv} = \hat{\xi} = s/\overline{x}$
Systolic Blood Pressure	**15 mm**	**130 mm**	**.115**
Diastolic Blood Pressure	**8 mm**	**60 mm**	**.133**

6g. The Range

The range is the difference between the largest and smallest values in a data set.

- It is a quick measure of scatter but not a very good one.
- Calculation utilizes only two of the available observations.
- As n increases, the range can only increase. Thus, the range is sensitive to sample size.
- The range is an unstable measure of scatter compared to alternative summaries of scatter (e.g. S or MADM)
- HOWEVER – when the sample size is very small, it may be a better measure of scatter than the standard deviation S.

Example –

- Data values are 5, 9, 12, 16, 23, 34, 37, 42
- range = 42-5=37

HOMEWORK DUE Friday September 23, 2022
Question #4 of 5

The following are behavioral ratings as measured by the Zang Anxiety Scale (ZAS) for 26 persons with a diagnosis of panic disorder:

53	51	46	45	40	35
59	51	45	60	35	
45	38	53	43	31	
36	40	41	41	38	
69	41	46	38	36	

4a. ***By any means you like.*** Compute the mean, median, mode, range, variance, and standard deviation, and the 25th and 75th percentiles.

Tip!!!!!! See page 50 before you start!!!!

4b. The following are behavioral ratings as measured by the Zang Anxiety Scale (ZAS) for 21 healthy controls:

26	26	25	25	25
28	26	26	25	
34	30	31	28	
26	34	25	25	
25	28	25	25	

By any means you like. Compute the mean, median, mode, range, variance, and standard deviation, and the 25th and 75th percentiles.

7. Some Other Important Numerical Summaries

In this section I consider both categorical and numerical data

Example - Consider a study of 25 consecutive patients entering the general medical/surgical intensive care unit at a large urban hospital.

- For each patient the following data are collected:

Variable Label (**Variable**)	Code
• Age, years (AGE)	
• Type of Admission (**TYPE_ADM**):	1= Emergency 0= Elective
• ICU Type (**ICU_TYPE**):	1= Medical 2= Surgical 3= Cardiac 4= Other
• Systolic Blood Pressure, mm Hg (**SBP**)	
• Number of Days Spent in ICU (**ICU_LOS)**	
• Vital Status at Hospital Discharge (**VIT_STAT)**:	1= Dead 0= Alive

The actual data are provided on the following page.

ID	Age	Type_Adm	ICU_Type	SBP	ICU_LOS	Vit_Stat
1	15	1	1	100	4	0
2	31	1	2	120	1	0
3	75	0	1	140	13	1
4	52	0	1	110	1	0
5	84	0	4	80	6	0
6	19	1	1	130	2	0
7	79	0	1	90	7	0
8	74	1	4	60	1	1
9	78	0	1	90	28	0
10	76	1	1	130	7	0
11	29	1	2	90	13	0
12	39	0	2	130	1	0
13	53	1	3	250	11	0
14	76	1	3	80	3	1
15	56	1	3	105	5	1
16	85	1	1	145	4	0
17	65	1	1	70	10	0
18	53	0	2	130	2	0
19	75	0	3	80	34	1
20	77	0	1	130	20	0
21	52	0	2	210	3	0
22	19	0	1	80	1	1
23	34	0	3	90	3	0
24	56	0	1	185	3	1
25	71	0	2	140	1	1

Categorical (Qualitative) data:

- Type of Admission (Type_Adm)
- ICU Type (ICU_Type)
- Vital Status at Hospital Dicharge (Vit_Stat)

Numerical (Quantitative) data:

- Age, years (Age)
- Number of days spent in ICU (ICU_LOS)
- Systolic blood pressure (SBP)

7a. Frequencies, Relative Frequencies and More

For nominal variables, we can compute frequencies and relative frequencies.

A tally of the possible outcomes, together with "how often" and "proportionately often" is called a **frequency and relative frequency distribution.**

- Appropriate for - nominal, ordinal, count data types.
- For the variable ICU_Type, the frequency distribution is the following:

Frequency & Relative Frequency Table

ICU_Type	Frequency ("how often")	Relative Frequency ("proportionately often")
Medical	12	0.48
Surgical	6	0.24
Cardiac	5	0.20
Other	2	0.08
TOTAL	25	1.00

- This summary will be useful in constructing two graphical displays, the bar chart and the pie chart.

For ordinal variables, we can compute frequencies and relative frequencies + cumulative frequencies and cumulative relative frequencies.

The Glasgow Coma Score (GCS) measures severity of a coma on an ordinal scale, with ***lower*** values corresponding ***to greater severity*** of coma. Suppose we have this information for 35 patients. The following table tallies the number of patients with each GCS score (frequency) together with the proportion of the sample of patients with each score (relative frequencies). But it also tallies what are called cumulative tallies and allows us to answer such questions as "how many patients have a GCS score of 5 or less?" (**cumulative frequency)** and "what proportion of the sample have GCS scores of 5 or less?" (**cumulative relative frequency**)

Frequency, Cumulative Frequency, Relative Frequency, Cumulative Relative Frequency Table

GCS Score	Frequency ("how often")	Relative Frequency ("proportionately often")	Cumulative Frequency	Cumulative Relative Frequency
3	**10**	**.285**	**10**	**.285**
4	**5**	**.143**	**15** (=10+5)	**.429**
5	**6**	**.171**	**21**	**.600**
6	**2**	**.057**	**23**	**.719**
7	**12**	**.343**	**35**	**1.000**
TOTAL	**35**			

HOMEWORK DUE Friday September 23, 2022
Question #5 of 5

The following table shows the age distribution of cases of a certain disease reported during a year in a particular state.

Age	Number of Cases
5-14	5
15-24	10
25-34	20
35-44	22
45-54	13
55-64	5
TOTAL	75

5a. ***By any means you like.*** Construct a frequency table with columns for class endpoints, class midpoint, frequency, relative frequency, cumulative frequency, and cumulative relative frequency.

5b. ***By any means you like.*** Estimate the values of the mean, median, variance, and standard deviation.

Tip -
Use the midpoints of each age interval as your values and use number of cases as their frequencies. For example, the value 10 has an estimated frequency of 5, the value 20 has an estimated frequency of 10, and so on.

7b. Percentiles (and Quantiles)

Percentiles are one way to summarize the range and shape of values in a distribution. Percentile values communicate various "cut-points". For example:

Suppose that 50% of a cohort survived at least 4 years.

This also means that 50% survived at most 4 years.

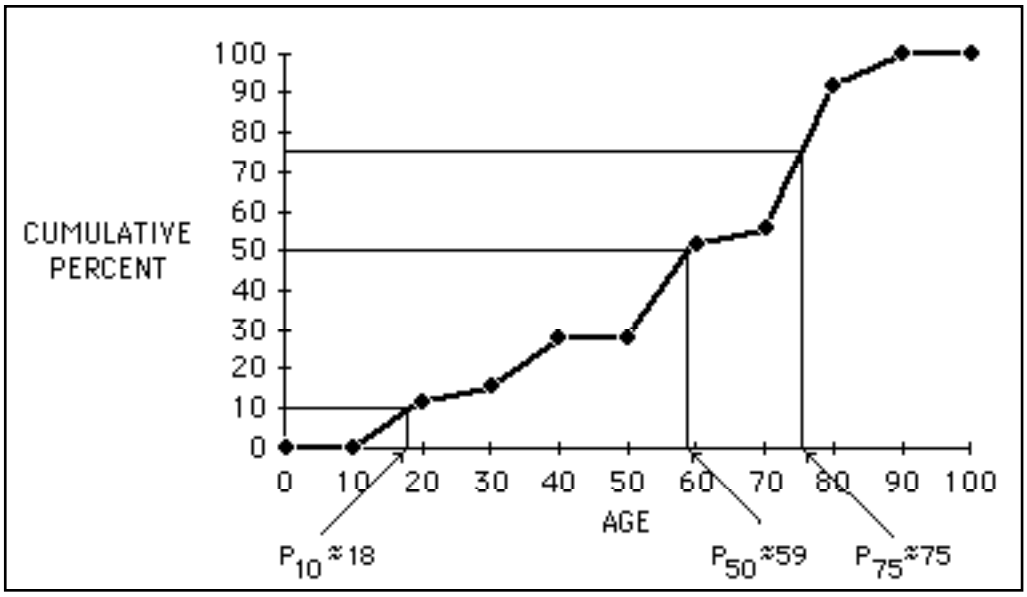
We say 4 years is the median.

The median is also called the 50^{th} percentile, or the 0.50 quantile. We write $P_{50} = 4$ years.

Similarly we could speak of other percentiles:

P_{25}: 25% of the sample values are less than or equal to this value. This is the 0.25 quantile
P_{75}: 75% of the sample values are less than or equal to this value. This is the 0.75 quantile
P_0: The minimum.
P_{100}: The maximum.

It is possible to estimate the values of percentiles from a cumulative frequency polygon (no worries – we'll come to this in Unit 2, Data Visualization).



Example – Consider $P_{10} = 18$. It is interpreted as follows: ***"10% of the sample is age ≤ 18"*** **or** ***"The 10th percentile of age in this sample is 18 years".***

How to Determine the Values of Q1, Q2, Q3 – the 25th, 50th, and 75th Percentiles in a Data Set

Often, it is the quartiles we're after. An easy solution for these is the following. Obtain the median of the entire sample. Then obtain the medians of each of the lower and upper halves of the distribution.

Step 1 - Preliminary:

Arrange the observations in your sample in order, from smallest to largest, with the smallest observation at the left.

Step 2 – Obtain median of entire sample:

Solve first for the value of Q2 = 50th percentile ("median):

	Sample Size is ODD	Sample Size is EVEN
Q2 = 50th Percentile ("median")	$Q2 = \left[\frac{n+1}{2}\right]^{th}$ ordered observation	$Q2 = \text{average}\left(\left[\frac{n}{2}\right], \left[\frac{n}{2}\right]+1\right)^{st}$ ordered observation

Step 3 – Q1 is the median of the lower half of the sample:

To obtain the value of Q1 = 25th percentile, solve for the median of the lower 50% of the sample.

Step 4 – Q3 is the median of the upper half of the sample:

To obtain the value of Q3 = 75th percentile, solve for the median of the upper 50% of the sample:

Example

Consider the following sample of n=7 data values

1.47	2.06	2.36	3.43	3.74	3.78	3.94

Solution for Q2

$$Q2 = 50^{th}\text{Percentile} = \left[\frac{7+1}{2}\right]^{th} = \left[4^{th}\text{ordered observation}\right] = 3.43$$

Solution for Q1

The lower 50% of the sample is thus, the following

1.47	2.06	2.36	3.43

$$Q1 = 25^{th}\text{Percentile} = \text{average}\left[\frac{4}{2}, \frac{4}{2}+1\right]^{st} = \text{average}\left[2\text{nd},3\text{rd observation}\right] = \text{average}(2.06, 2.36) = 2.21$$

Solution for Q3

The upper 50% of the sample is the following

3.43	3.74	3.78	3.94

$$Q3 = 75^{th}\text{Percentile} = \text{average}\left[\frac{4}{2}, \frac{4}{2}+1\right]^{st} = \text{average}\left[2\text{nd}, 3\text{rd observation}\right] = \text{average}(3.74, 3.78) = 3.76$$

How to determine the values of other Percentiles in a Data Set

Important Note – Unfortunately, there exist multiple formulae for doing this calculation. Thus, there is no single correct method

Consider the following sample of n=40 data values

0	1	1	3	17	32	35	44	48	86
87	103	112	121	123	130	131	149	164	167
173	173	198	208	210	222	227	234	245	250
253	256	266	277	284	289	290	313	477	491

Step 1:
Order the data from smallest to largest

Step 2:

Compute $L = n\left[\frac{p}{100}\right]$ where

n = size of sample (eg; n=40 here)
p = desired percentile (eg p=25th)

L is NOT a whole number

Step 3:

Change L to next whole number.

Pth percentile = Lth ordered value in the data set.

L is a whole number

Step 3:

Pth percentile = average of the Lth and (L+1)st ordered value in the data set.

7c. Five Number Summary

A "five number summary" of a set of data is, simply, a particular set of five percentiles:

P_0: The minimum value.
P_{25}: 25% of the sample values are less than or equal to this value.
P_{50}: The median. 50% of the sample values are less than or equal to this value.
P_{75}: 75% of the sample values are less than or equal to this value.
P_{100}: The maximum.

Why bother? **This choice of five percentiles is actually a good summary, since:**

The minimum and maximum identify the extremes of the distribution, and

The 1st and 3rd quartiles identify the middle "half" of the data, and

Altogether, the five percentiles are the values that define the quartiles of the distribution, and

Within each interval defined by quartile values, there are an equal number of observations.

Example, continued –

We're just about done since on page 46, the solution for P_{25} , P_{50}, and P_{75} was shown. Here is the data again.

1.47	2.06	2.36	3.43	3.74	3.78	3.94

Thus,

P_0 = the minimum value = **1.47**
P_{25} = 1st quartile = 25th percentile = **2.21**
P_{50} = 2nd quartile = 50th percentile (median) = **3.43**
P_{75} = 3rd quartile = 75th percentile = **3.76**
P_{100} = the maximum value = **3.94**

7d. Interquartile Range (IQR)

The interquartile range is simply the difference between the 1st and 3rd quartiles:

$$\text{IQR} = \text{Interquartile Range} = \left[P_{75} - P_{25}\right]$$

The IQR is a useful summary also:

It is an alternative summary of dispersion (sometimes used instead of standard deviation)

The range represented by the IQR tells you the spread of the middle 50% of the sample values

Example, continued –

Here is the data again.

1.47	2.06	2.36	3.43	3.74	3.78	3.94

P_0 = the minimum value = **1.47**
P_{25} = 1st quartile = 25th percentile = **2.21**
P_{50} = 2nd quartile = 50th percentile (median) = **3.43**
P_{75} = 3rd quartile = 75th percentile = **3.76**
P_{100} = the maximum value = **3.94**

IQR = Interquartile Range = [P75 – P25] = **[3.76 – 2.21] = 1.55**

Activity – Introduction to "Art of Stat"

Recall again the behavioral ratings data of question #3b on page 39:

26	**26**	**25**	**25**	**25**
28	**26**	**26**	**25**	
34	**30**	**31**	**28**	
26	**34**	**25**	**25**	
25	**28**	**25**	**25**	

In question #4b, you were asked to compute by hand the values of the following sample statistics: mean, median, mode, range, variance, and standard deviation, and the 25th and 75th percentiles.

In this exercise (activity, really), I invite you to play with this same data in a wonderful online application.

Step 1 – Launch "artofstat.com" and click at right on **Online Web Apps**

Step 2 - From the main welcome window, middle, click on
EXPLORE QUANTITATIVE DATA

Step 3 - From this menu, at top left, under ENTER DATA, choose YOUR OWN

Step 4 - Now enter your data. One way is to do this by hand. Alternatively, if you're lazy (like me), you could also highlight to select the data in these course notes and then do an ***EDIT>COPY*** (control-C) followed by an ***EDIT>PASTE.***
Tip - Not sure, but you might just have to do a little tweaking to be sure the data values themselves are separated by just one space

AND NOW - just play around with this app and enjoy.