

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”

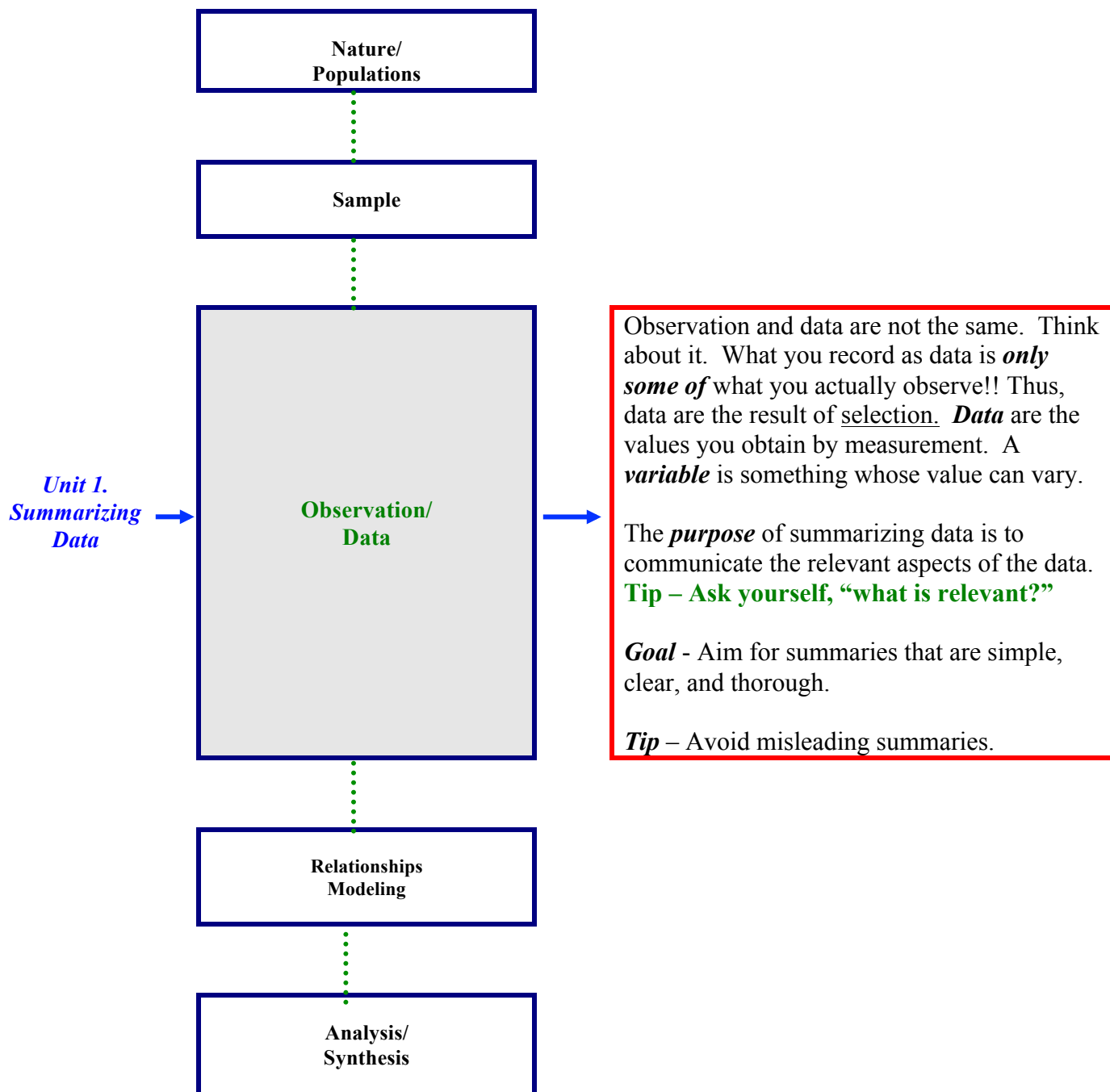
- Sir Francis Galton (England, 1822-1911)

This unit introduces **variables** and the variety of **types of data** possible (nominal, ordinal, interval, and ratio). **Graphical** and **numerical** ways of summarizing data are described. Graphs are encouraged and it is emphasized that these should be as simple and as clear as possible. 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).

Table of Contents

Topics		
	1. Unit Roadmap	3
	2. Learning Objectives	4
	3. Variables and Types of Data	5
	4. Summaries for Qualitative Data	12
	a. Frequency Table, Relative Frequency Table	14
	b. The Bar Chart	14
	c. The Pie Chart	17
	5. Graphical Summaries for Quantitative Data	18
	a. The Histogram	19
	b. The Frequency Polygon	23
	c. The Cumulative Frequency Polygon	24
	d. Percentiles (Quantiles)	25
	e. Five Number Summary	28
	f. Interquartile Range, IQR	29
	g. Quantile Quantile Plot	30
	h. Stem and Leaf Diagram	31
	i. Box and Whisker Plot	32
	6. The Summation Notation.....	34
	7. Numerical Summaries for Quantitative Data -Central Tendency	35
	a. The mode	37
	b. The mean	38
	c. The mean as a “balancing” point and skewness	39
	d. The mean of grouped data	40
	e. The median	41
	8. Numerical Summaries for Quantitative Data - Dispersion.....	43
	a. Variance	44
	b. Standard Deviation	45
	c. Median Absolute Deviation from Median	47
	d. Standard Deviation v Standard Error	48
	e. A Feel for Sampling Distributions	51
	f. The Coefficient of Variation	53
	g. The Range	54

1. Unit Roadmap



Nature ——— Population/
Sample ——— Observation/
Data ——— Relationships/
Modeling ——— Analysis/
Synthesis

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 “when and how” to construct bar charts. *Tip! -Pie charts are not recommended.*
- Understand and know how to compute: percentile, five number summary, and interquartile range, IQR.
- Understand and know “when and how” to construct the following graphical summaries for quantitative data: histograms, frequency and cumulative frequency polygons, quantile-quantile plot, stem and leaf diagrams, and box and whisker plots.
- 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 2 (Introduction to Probability) and in Unit 3 (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.

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 “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.

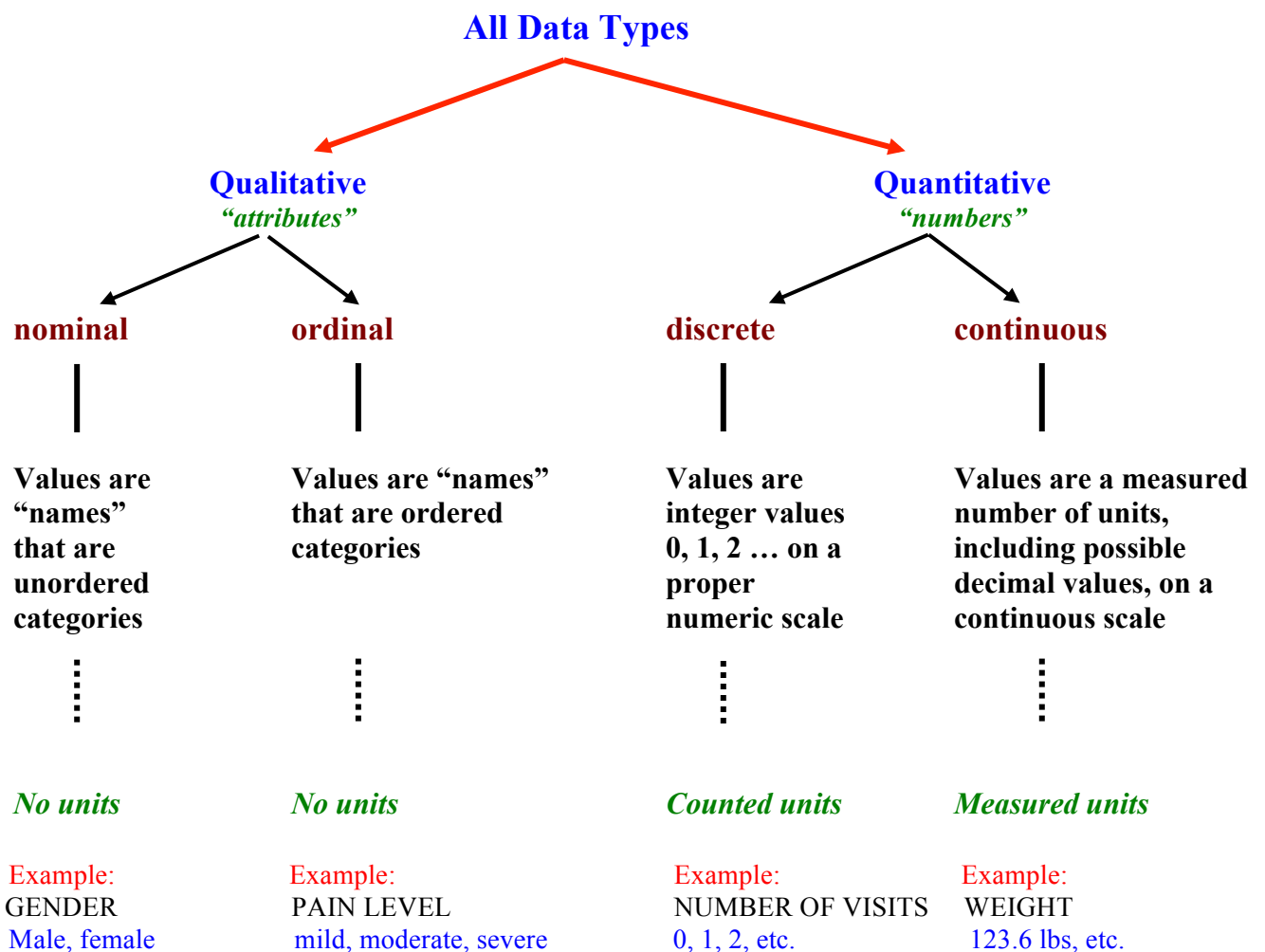
This data table (spreadsheet) has three observations (rows), four variables, and 12 data values.

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

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.

- Daniel, Wayne W (*“Biostatistics – A Foundation for Analysis in the Health Sciences”*) classifies data types as follows:



The distinction between qualitative versus quantitative is straightforward:

Qualitative: Attributes, characteristics that cannot be reported as numbers

Quantitative: Numbers

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

Consider this ...

We can reasonably refer to the child's 22 year old cousin as being twice as old as the 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.

- **QUALITATIVE ► 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

Nature ——— Population/
Sample ——— Observation/
Data ——— Relationships/
Modeling ——— Analysis/
Synthesis

- **QUALITATIVE ► 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 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 is all we know
apart from their utility as labels.

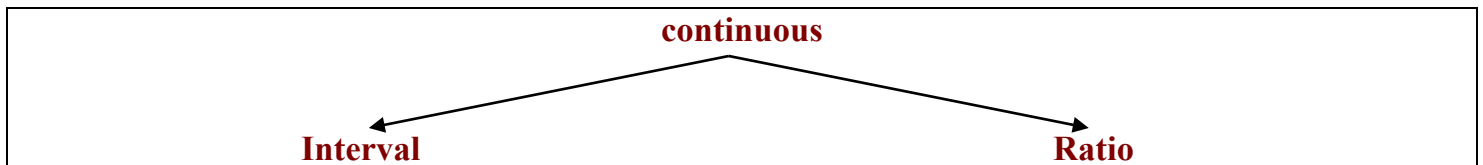
QUANTITATIVE ► Discrete Scale: Values are **counts** of the number of times some event occurred and are thus whole numbers: 0, 1, 2, 3, etc. ...

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.

QUANTITATIVE ► Continuous: A further classification of data types is possible for quantitative data that are continuous



QUANTITATIVE ► 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.

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

QUANTITATIVE ► **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				
Type	Qualitative “attribute”		Quantitative “number”	
	Nominal	Ordinal	Discrete	Continuous
Descriptive Methods	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	Frequency Relative Frequency Frequency	Frequency Relative Frequency Cumulative Frequency 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.

Nature ——— Population/ Sample ——— Observation/ Data ——— Relationships/ Modeling ——— Analysis/ Synthesis

4. Summaries for Qualitative (“*attribute*”) 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:

<u>Variable Label (Variable)</u>	<u>Code</u>
• 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

Qualitative data:

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

Quantitative data:

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

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

4a. Frequency Table, Relative Frequency Table

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:

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

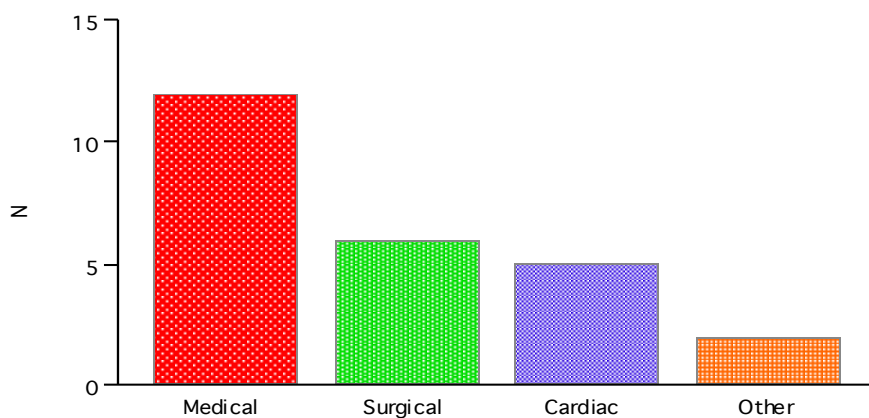
- ◆ This summary will be useful in constructing two graphical displays, the bar chart and the pie chart.

4b. Bar Chart

On the horizontal are the possible outcomes. **Important** – In a bar chart, along the horizontal axis, the possible outcomes are separated by spaces. (This makes sense – the spaces remind us that outcomes “inbetween” are not possible) On the vertical is plotted either

- ◆ “how often” - frequency
- ◆ “how proportionately often” - relative frequency

Example - Bar chart for Type of ICU Patients (n=25)



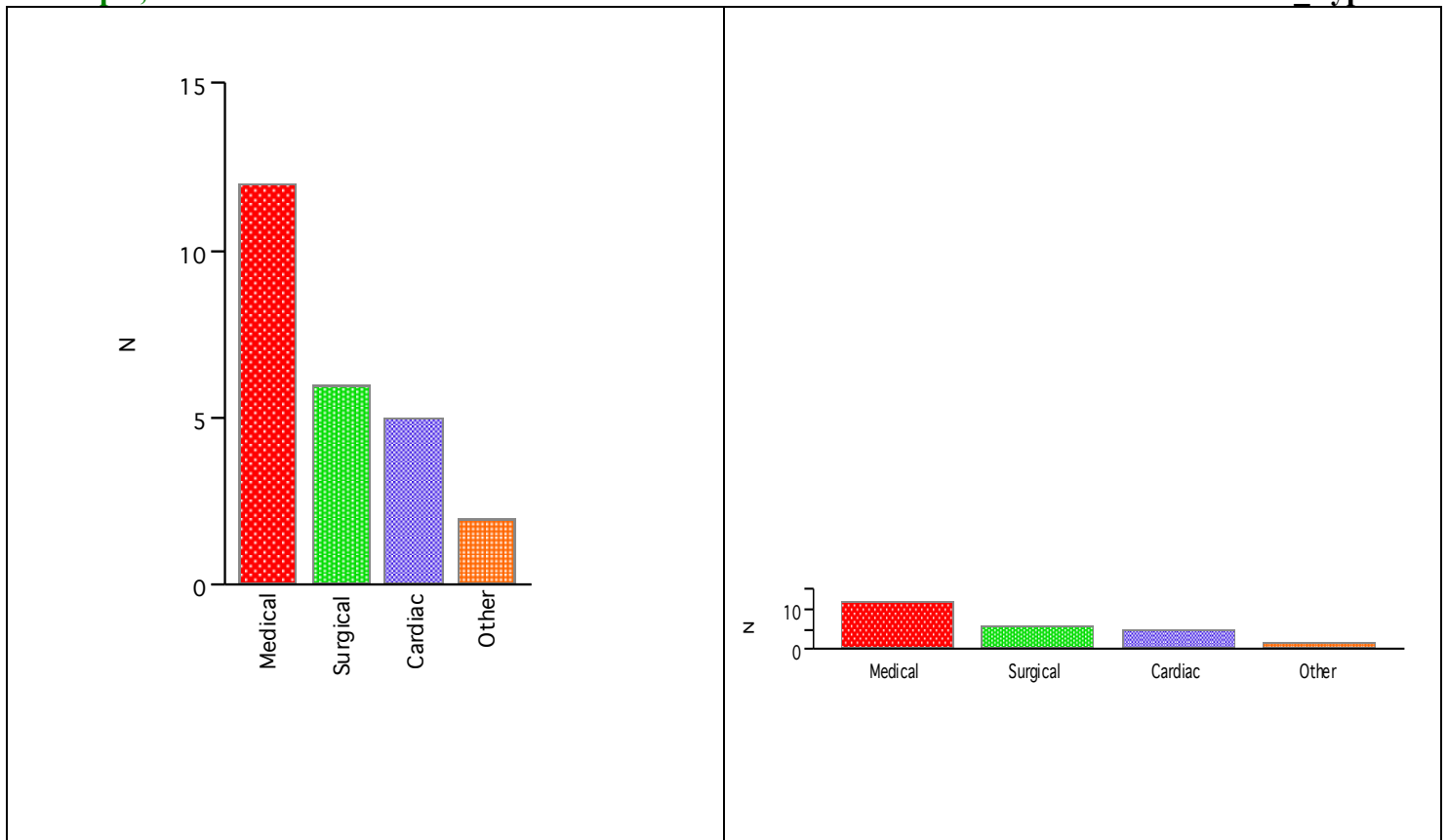
Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

Guidelines for Construction of Bar Charts

- Label both axes clearly
- Leave space between bars
- Leave space between the left-most bar and the vertical axis
- When possible, begin the vertical axis at 0
- All bars should be the same width

There's no reason why the bar chart can't be plotted horizontally instead of vertically. And, not surprisingly, if you change the choice of scale, you can communicate to the eye a very different message.

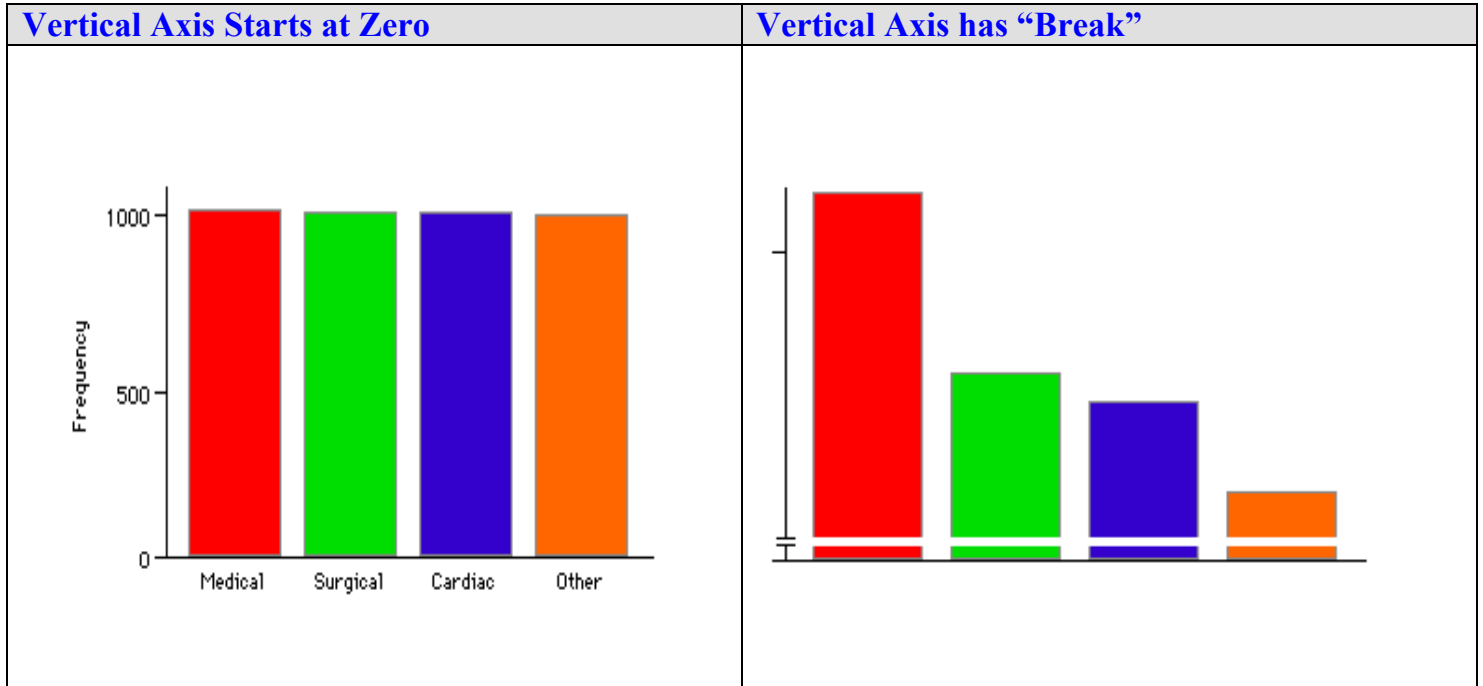
Example, continued - Consider two choices of scale for the vertical axis in the bar chart for ICU_Type:



It's difficult to know how to construct a bar chart when the frequencies are very high.

- Suppose the frequencies were 1012 medical, 1006 surgical, 1005 cardiac and 1002 other type.
- Is it better to have a vertical axis start at zero or to “break” the axis?

Example, continued -



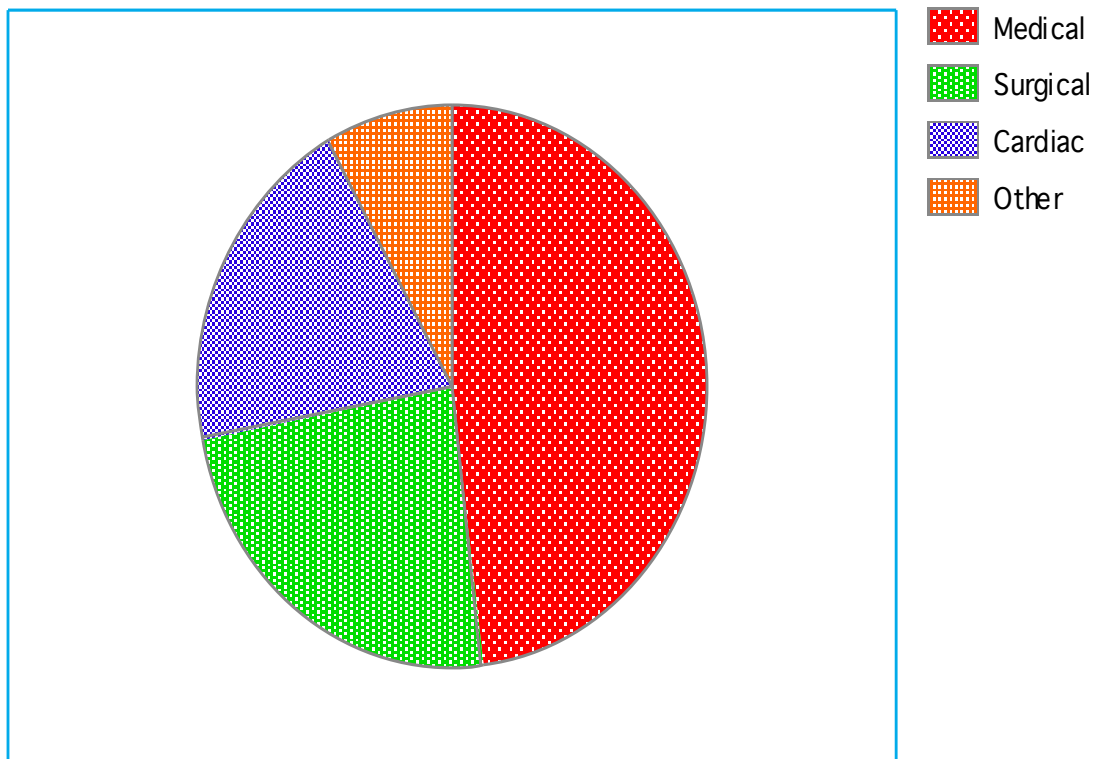
4c. Pie Chart

Instead of bars rising up from the horizontal (as in a bar chart), we could plot instead the shares of a pie.

Recalling that a circle has 360 degrees, that 50% means 180 degrees, 25% means 90 degrees, etc, we can identify “wedges” according to relative frequency

Relative Frequency	Size of Wedge, in degrees
0.50	50% of 360 = 180 degrees
0.25	25% of 360 = 90 degrees
p	$(p) \times (100\%) \times 360$ degrees

Example, continued –



Tip - Avoid pie charts! It is harder for the eye to compare pie slices than to compare bars. Also, the eye has to move around the circle, thus violating the rule of simplicity.

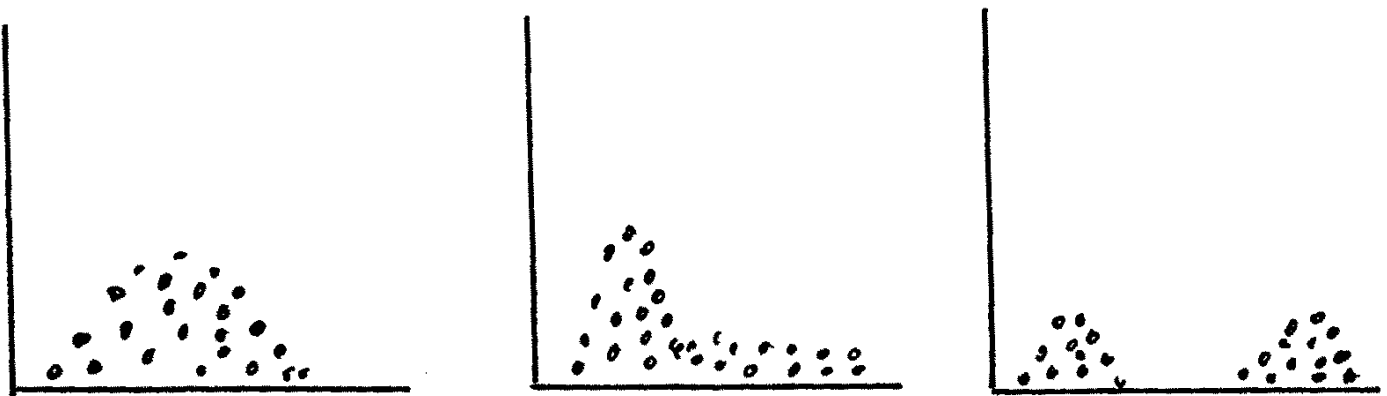
5. Graphical Summaries for Quantitative Data

Summaries for quantitative data address many things. Among the most important are:

- What is typical (location)
- What is the scatter (dispersion)

Another important aspect of quantitative data, however, is the *shape* of the distribution of values.

The following are 3 scenarios for the patterns of values (eg – values of age) in a simple random sample (eg – a simple random sample of cholesterol values where the sample size is $n=25$)



The 3 patterns are quite different. The leftmost pattern is symmetric and bell shaped. The middle pattern has a tail to the right. The rightmost pattern is comprised of two symmetric bell shaped patterns that are separated in their location.

“Good” choices for summarizing location and dispersion are not always the same and depend on the pattern of scatter.

5a. Histogram

For a continuous variable (e.g. – age), the frequency distribution of the individual ages is not so interesting.

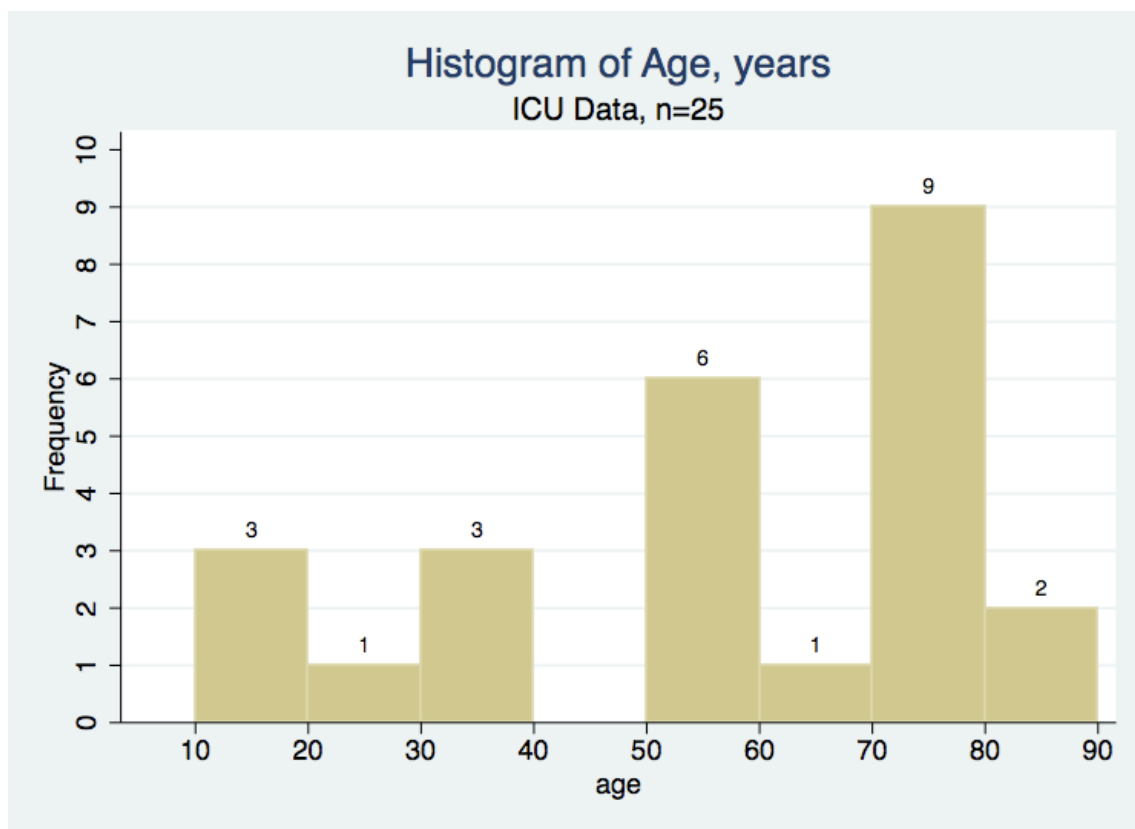
Age	Frequency
15	1
19	2
29	1
31	1
34	1
39	1
52	2
53	2
56	2
65	1
71	1
74	1
75	2
76	2
77	1
78	1
79	1
84	1
85	1

We “see more” in frequencies of age values in “groupings”. Here, 10 year groupings make sense.

Age Interval	Frequency
10-19	3
20-29	1
30-39	3
40-49	0
50-59	6
60-69	1
70-79	9
80-89	2
TOTAL	25

A plot of this “grouped” frequency table gives us a better feel for the pattern of ages with respect to both location and scatter. This plot is called a [histogram](#)

- A **histogram** is a graphical summary of the pattern of values of a **continuous random variable**. More formally, it is a graphical summary of the frequency distribution
- It is **analogous** to the bar graph summary for the distribution of a discrete random variable



Stata command that produced this histogram:

```
.histogram age, width(10) start(10) frequency addlabels xlabel(10 (10) 90) ylabel(0 (1) 10) title("Histogram of Age, years") subtitle("ICU Data, n=25")
```

Stata command for a very basic histogram:

```
.histogram age
```

How to Construct a Histogram

Step 1: Choose the number of groupings (“class intervals”). Call this k .

- ♦ The choice is arbitrary.
- ♦ K too small over-summarizes. K too large under-summarizes.
- ♦ Sometimes, the choices of intervals are straightforward – eg 10 year intervals, 7 day intervals, 30 day intervals.
- ♦ Some text books provide formulae for k . Use these if you like, provided the resulting k makes sense.
- ♦ Example – For the age data, we use $k=8$ so that intervals are sensible 10 year spans.

Step 2: Rules for interval beginning and end values (“boundaries”)

- ♦ Boundaries should be such that each observation has exactly one “home”.
- ♦ Equal widths are not necessary. WARNING – If you choose to plot intervals that are of Unequal widths, take care to plot “area proportional to relative frequency”. This is explained in Step 3.

Step 3: In a histogram, always plot “Area Proportional to Relative Frequency”

- ♦ Example – Suppose the first two age intervals are combined:

Age Interval	Frequency
10-29	4
30-39	3
40-49	0
50-59	6
60-69	1
70-79	9
80-89	2
TOTAL	25

- ◆ For the intervals 30-39, 40-49, 50-59, 60-69, 70-79, 80-89: the widths are all the same and span 10 units of age. Heights plotted are 3, 0, 6, 1, 9, and 2.
- ◆ The new, combined, **10-29 spans 20 units of age**. A frequency of 4 over 20 units of age corresponds to a frequency of 2 over 10 units of age.
- ◆ For the interval 10-29, a height of 2 would be plotted.

Histograms with bins of varying width are possible in Stata but require some additional coding that is beyond the scope of this course.

5b. Frequency Polygon

The frequency polygon is an alternative to the histogram. It's not used often. Its companion, the cumulative frequency polygon (next page), is more commonly used.

- ◆ Both the histogram and frequency polygon are graphical summaries of the frequency distribution of a continuous random variable
- ◆ Whereas in a histogram ...
 - ◆ X-axis shows intervals of values
 - ◆ Y-axis shows bars of frequencies
- ◆ In a frequency Polygon:
 - ◆ X-axis shows midpoints of intervals of values
 - ◆ Y-axis shows dot instead of bars

Some guidelines:

- i. The graph title should be a complete description of the graph
- ii. Clearly label both the horizontal and vertical axes
- iii. Break axes when necessary
- iv. Use equal class widths
- v. Be neat and accurate

Example -

The following frequency polygon is similar in interpretation to a histogram.

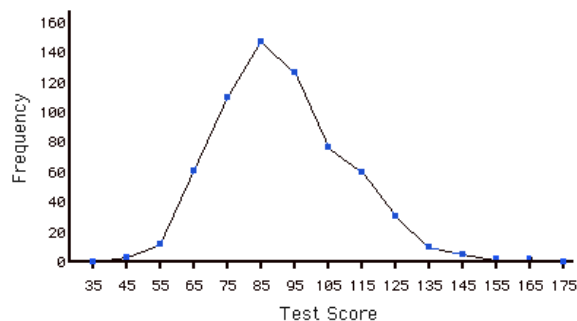


Figure 1: Frequency polygon for the psychology test scores.

Source: <http://cnx.org/content/m11214/latest/>

5c. Cumulative Frequency Polygon

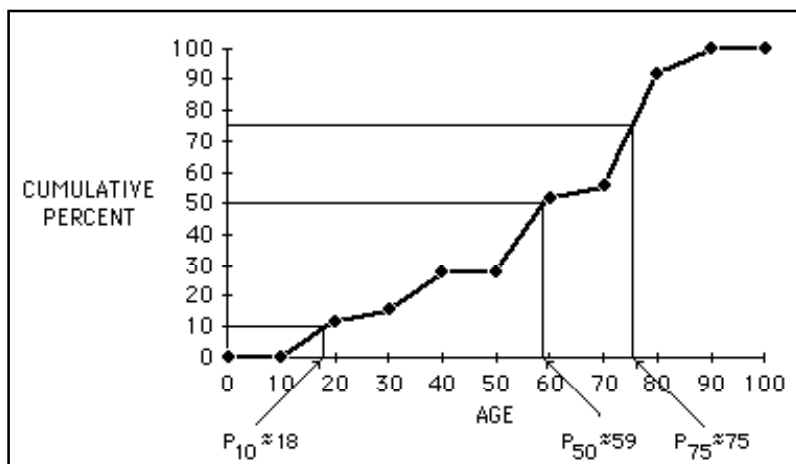
As you might think, a **cumulative** frequency polygon depicts **accumulating** data.

This is actually useful in that it lets us see **percentiles** (eg, median)

Its plot utilizes cumulative frequency information (right hand columns below in blue)

Age Interval	Frequency (count)	Relative Frequency (%)	Cumulative through Interval	
			Frequency (count)	Relative Frequency (%)
10-19	3	12	3	12
20-29	1	4	4	16
30-39	3	12	7	28
40-49	0	0	7	28
50-59	6	24	13	52
60-69	1	4	14	56
70-79	9	36	23	92
80-89	2	8	25	100
TOTAL	25	100		

Notice that it is the **ENDPOINT** of the interval that is plotted on the horizontal. This makes sense inasmuch as we are keeping track of the **ACCUMULATION** of frequencies to the **end of the interval**.



5d. Percentiles (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 50th percentile, or the 50th quantile. We write $P_{50} = 4$ years.

Similarly we could speak of other percentiles:

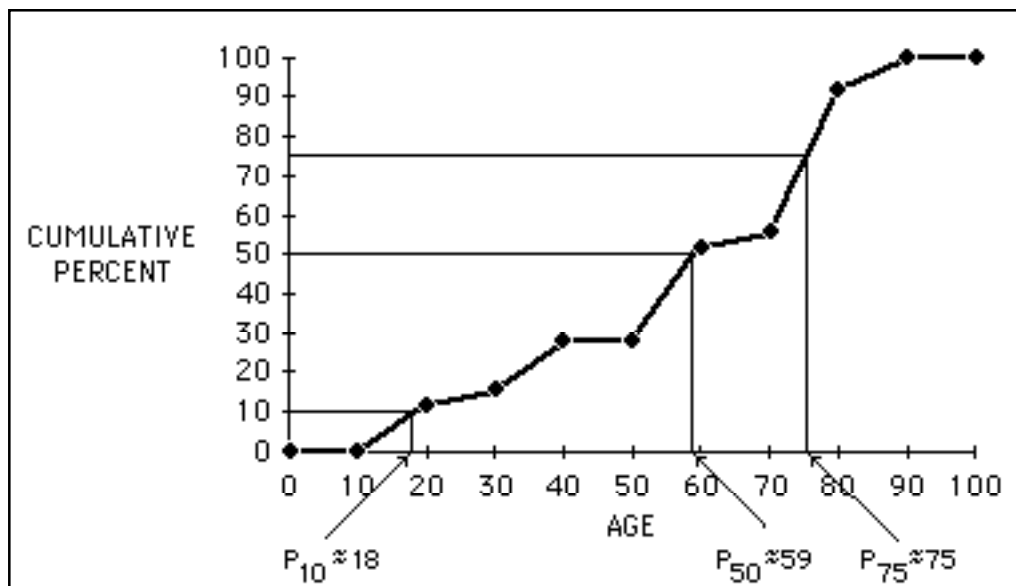
P_{25} : 25% of the sample values are less than or equal to this value.

P_{75} : 75% of the sample values are less than or equal to this value.

P_0 : The minimum.

P_{100} : The maximum.

It is possible to estimate the values of percentiles from a cumulative frequency polygon.



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”.

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

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

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

Solution for Q1

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

1.47	2.06	2.36	3.43
------	------	------	------

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

Solution for Q3

The upper 50% of the sample is the following

3.43	3.74	3.78	3.94
------	------	------	------

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

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.

5e. 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 26, 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**

5f. Interquartile Range (IQR)

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

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

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

$$\text{IQR} = \text{Interquartile Range} = [P_{75} - P_{25}] = [3.76 - 3.21] = 1.55$$

5g. Quantile-Quantile (QQ) and Percentile-Percentile (PP) Plots

- We might ask: “Is the distribution of my data normal?”
QQ and PP plots are useful when we want to compare the percentiles of our data with the percentiles of some reference distribution (eg- reference is normal)

- QQ Plot:** **X= quantile in sample** **Y=quantile in reference**

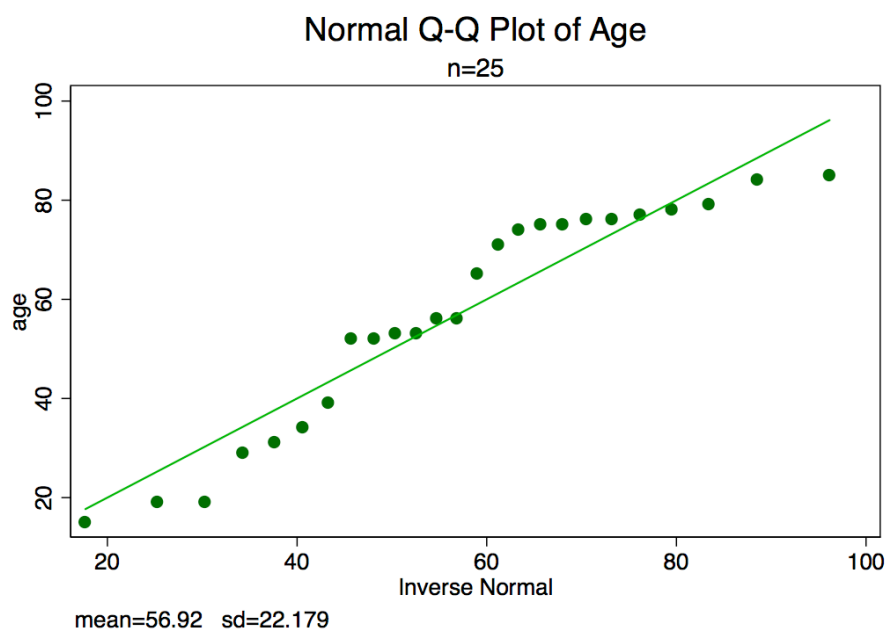
Desired Quantile	10th	20th	...	99 th
X = Value in Sample of Data				
Y = Value in Reference Distribution				

- PP Plot:** **X= percentile rank in sample** **Y= percentile rank in reference**

Data value	##	##	...	##
X = Rank (percentile) in Sample of Data				
Y = Rank (percentile) in Reference Distribution				

- What to look for:** A straight line suggests that the two distributions are the same

Example – Normal Quantile-Quantile Plot of Age for data on page 13



Stata command that produced this Q-Q Plot:

```
. qnorm age, title("Normal Q-Q Plot of Age") subtitle("n=25") caption("mean=56.92 sd=22.179")
```

Stata command to produce a basic Q-Q Plot:

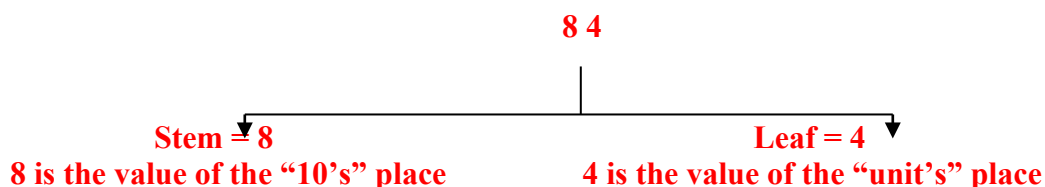
```
. qnorm age
```

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

5h. Stem and Leaf Diagram

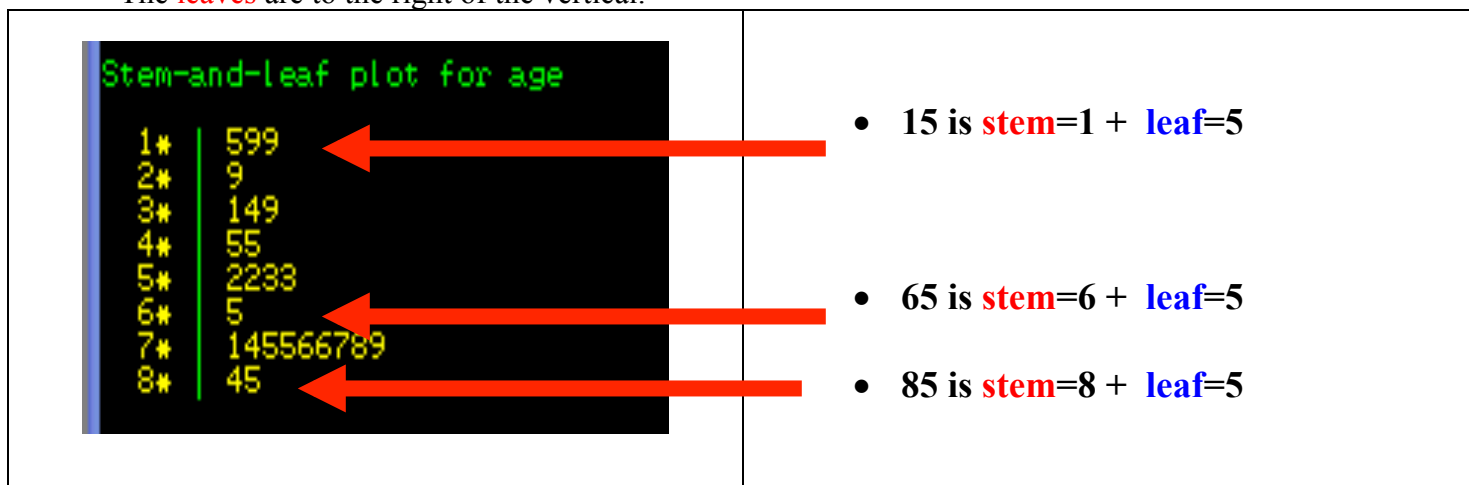
A stem and leaf diagram is a “quick and dirty” histogram. It is also a quick and easy way to sort data. Each actual data point is de-constructed into a stem and a leaf. There are a variety of ways to do this, depending on the sample size and range of values.

- For example, the data value 84 might be de-constructed as follows



Example - Stem and Leaf Plot of Age of 25 ICU Patients:

- The **stems** are to the left of the vertical. In this example, each value of stem represents a multiple of 10.
- The **leaves** are to the right of the vertical.



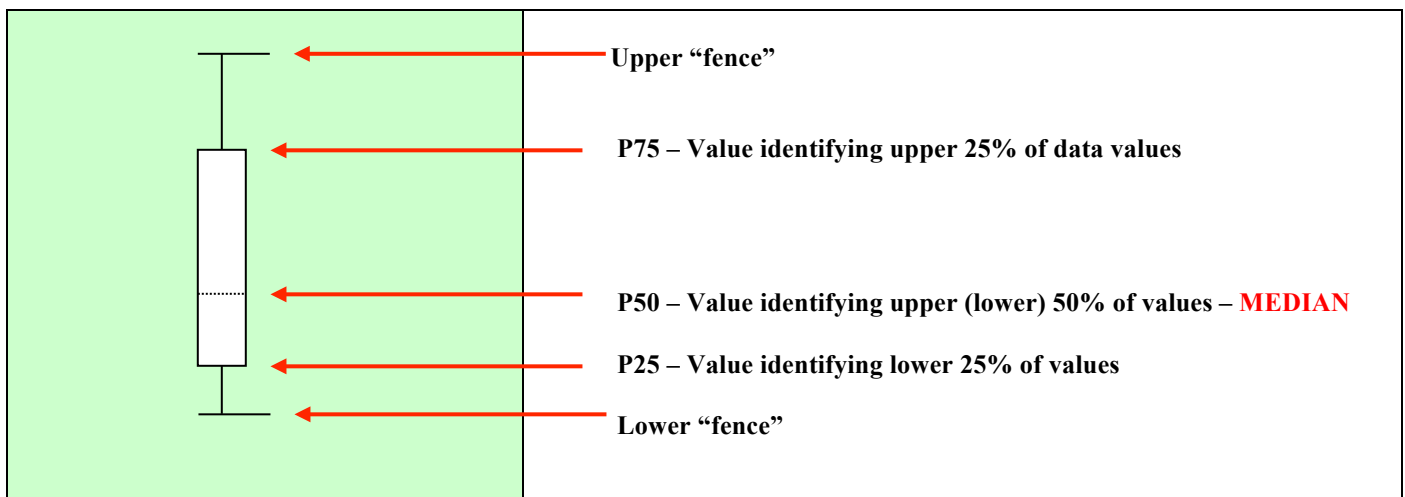
Key - Among other things, we see that the minimum age in this sample is 15 years (Note the unit=5 in the row for stem=1) and the maximum age is 85 years (Unit=5 in the row for stem=8). We can also see that most of the subjects in this sample are in their seventies (Actual ages are 71, 74, 75, 75, 76, 76, 77, 78, 79).

5i. Box and Whisker Plot

- The box and whisker plot (also called box plot) is a wonderful schematic summary of the distribution of values in a data set.
- It shows you a number of features, including: extremes, 25th and 75th percentile, median and sometimes the mean.
- Side-by-side box and whisker plots are a wonderful way to compare multiple distributions.

- **Definition**

- The central box spans the interquartile range and has P_{25} and P_{75} for its limits. It spans the middle half of the data.
- The line within the box identifies the median, P_{50} . Sometimes, an asterisk within the box is shown. It is the mean. The lines coming out of the box are called the “whiskers”. The ends of these “whiskers” are called “fences”.
- Upper “fence” = The largest value that is still less than or equal to $P_{75} + 1.5*(P_{75} - P_{25}) = P_{75} + 1.5*(IQR)$.
- Lower “fence” = The smallest value that is still greater than or equal to $P_{25} - 1.5*(P_{75} - P_{25}) = P_{25} - 1.5*(IQR)$.

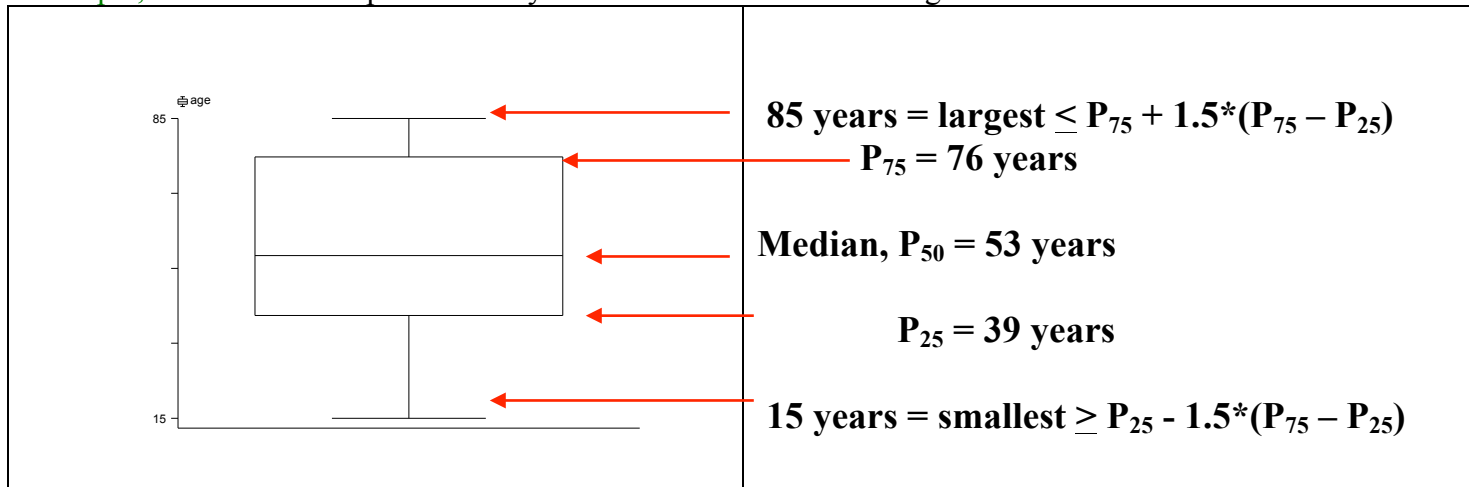


Note on the multiplier 1.5

This is a convenient multiplier if we are interested in comparing the distribution of our sample values to a normal (Gaussian) distribution in the following way. If the data are from a normal distribution, then 95% of the data values will fall within the range defined by the lower and upper fences.

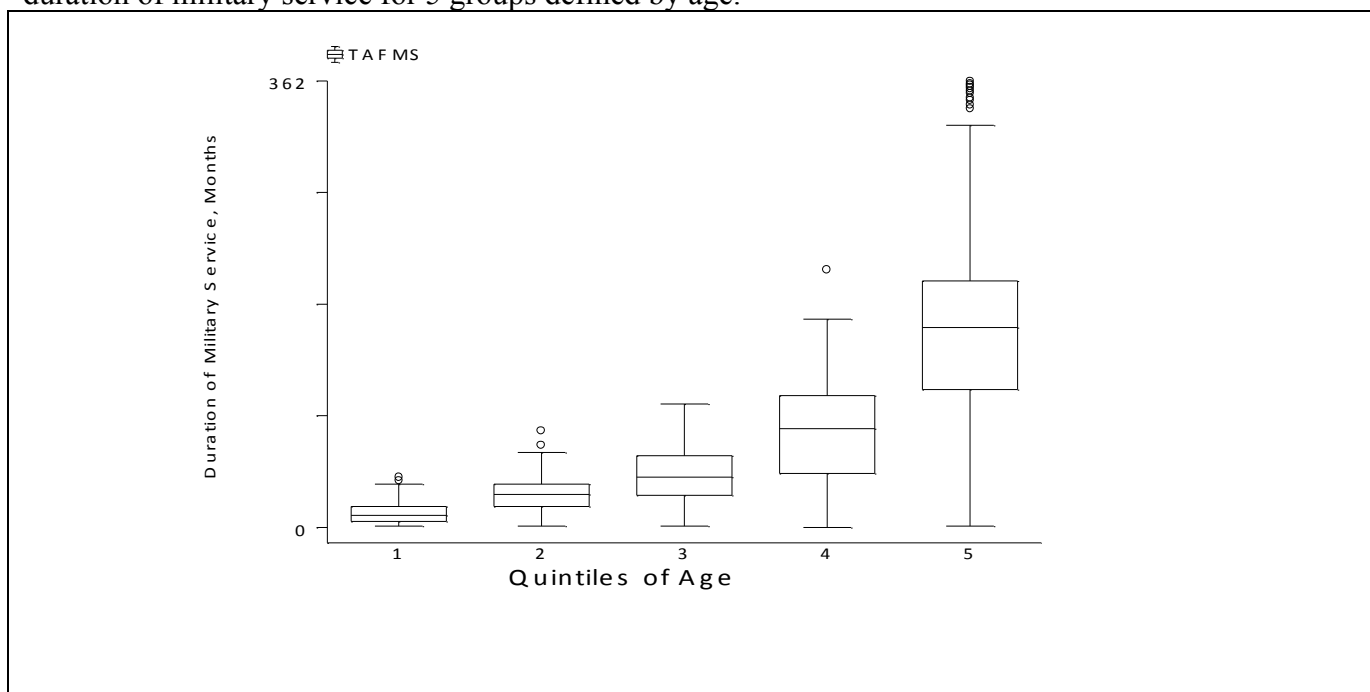
Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

Example, continued - Box plot summary of the distribution of the 25 ages.



Note: min=15, P₂₅=39, P₅₀=53, P₇₅=76, max=85, and $1.5*(P_{75}-P_{25})=55.5$

Example – The following is an example of side-by-side box and whisker plots. They are plots of values of duration of military service for 5 groups defined by age.



- Duration of military service increases with age.
- With age, the variability in duration of military service is greater.
- The individual circles represent extreme values.

6. The Summation Notation

The summation notation is nothing more than a secretarial convenience. We use it to avoid having to write out long expressions.

For example,

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

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

Another example –

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

Σ

The Greek symbol sigma says “add up some items”

Σ

STARTING HERE

Below the sigma symbol is the starting point

Σ

END

Up top is the ending point

Example – The first 5 values of age on page 13. Using summation notation, what is the sum of the 2nd, 3rd, and 4th values?

$x_1=15 \quad x_2=31 \quad x_3=75 \quad x_4=52 \quad x_5=84 \quad \rightarrow$

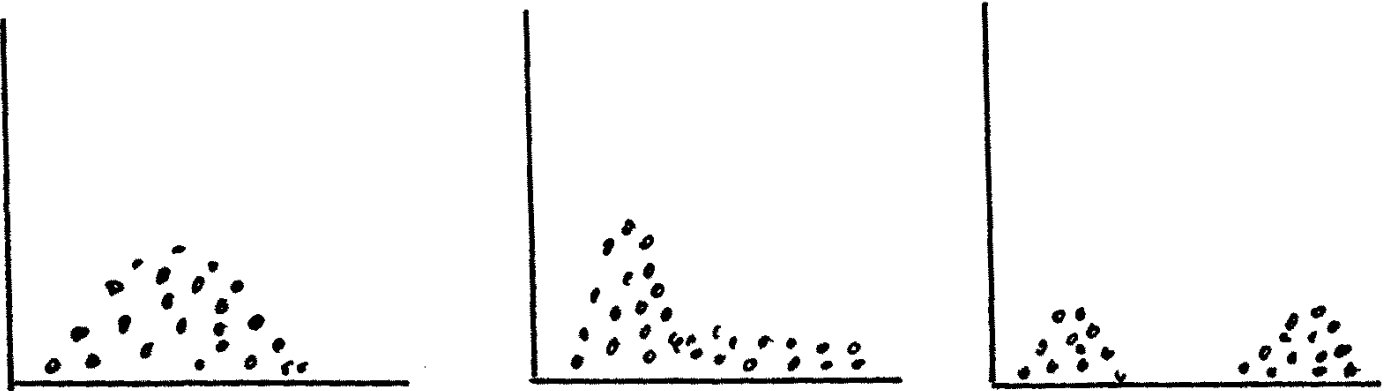
$$\sum_{i=2}^4 x_i = x_2 + x_3 + x_4 = 31 + 75 + 52 = 158$$

7. Numerical Summaries for Quantitative Data - Central Tendency

Previously we noted that among the most important tools of description are that address

- What is typical (location)
- 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**.



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.

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

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 ...



7a. 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 “bi-modal”

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.

7b. Mean

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

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

Calculation of a “mean” or “average” is familiar; e.g. -

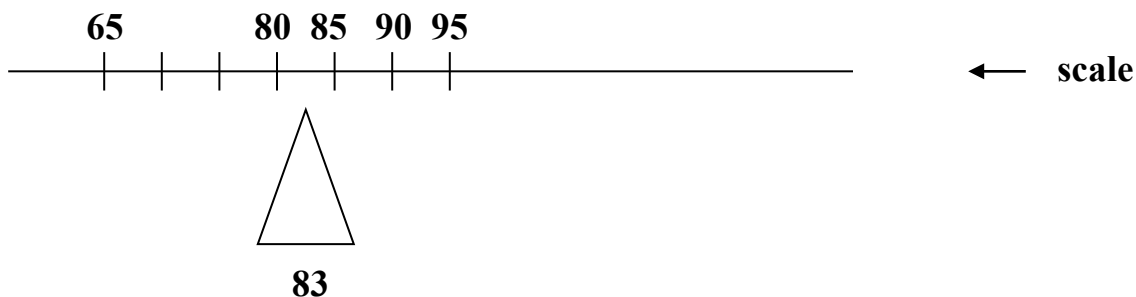
grade point average
mean annual rainfall
average weight of a catch of fish
average family size for a region

A closer look using summation notation introduced on page 34)

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

7c. 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.

$$\begin{aligned} \text{LEFT: } (83-65) + (83-80) &= 21 \\ \text{RIGHT: } (85-83) + (90-83) + (95-83) &= 21 \end{aligned}$$

- In this example, sample mean $\bar{X} = 83$
- 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)	Positive Skewness (Right tail)
<p>Mean is dragged left</p>	<p>Mean is dragged right</p>

7d. 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.

7e. 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	median = $\frac{n+1}{2}$ th largest value
If the sample size n is EVEN	median = average of $\left(\left[\frac{n}{2} \right] \text{th}, \left[\frac{n+2}{2} \right] \text{th} \right)$ 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 =

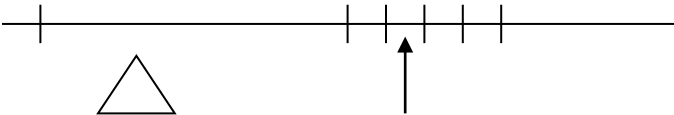
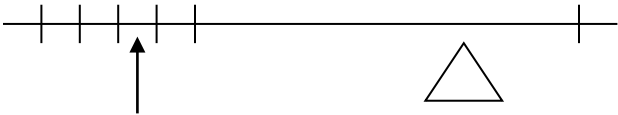
$$\text{average} \left[\frac{n}{2} \text{th largest}, \frac{n+2}{2} \text{th largest} \right] = \text{average} [\text{of 6th and 7th largest values}]$$

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

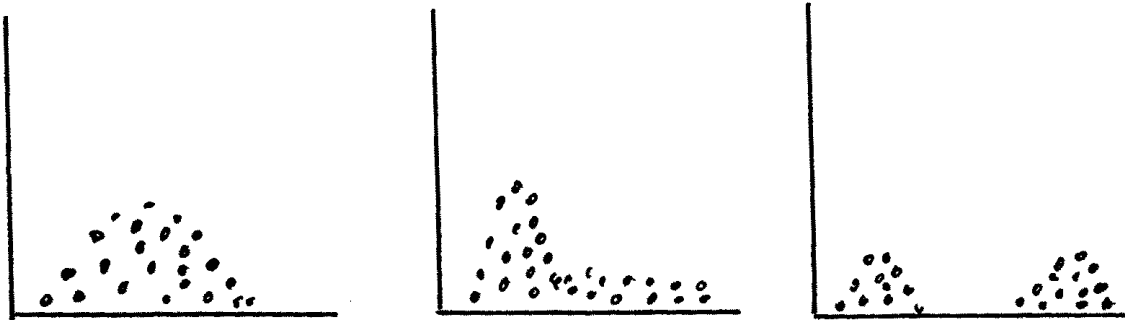
Skewed Data – When the data are skewed the median is a better description of the majority than the mean

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
 <p>Mean is dragged left</p> <p>MEDIAN</p>	 <p>MEDIAN</p> <p>Mean is dragged right</p>

8. 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.

8a. 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 data 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 (2) μ is replaced by \bar{X} and (3) the divisor N is replaced by the divisor (n-1). This suggests that, provided the sample is a representative one, the sample variance might be a good guess (estimate) of the population variance.

8b. 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.

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

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

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 $i = 1, \dots, 11$.

Patient Identifier, “i”	Survival (days), X_i	Mean for sample, \bar{X}	Deviation , $(X_i - \bar{X})$	Squared deviation $(X_i - \bar{X})^2$
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

◆ $\sum_{i=1}^{11} X_i = 1771 \text{ days}$

◆ Sample mean is $\bar{X} = \frac{1771}{11} = 161 \text{ days}$

◆ Sample variance is $S^2 = \frac{\sum_{i=1}^{11} (X_i - \bar{X})^2}{n-1} = \frac{285236}{10} = 28523.6 \text{ days}^2$

◆ Sample standard deviation is $s = \sqrt{S^2} = \sqrt{28523.6} = 168.89 \text{ days}$

8c. 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)

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

Example.

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

Median = 2.2

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

$$\text{MADM} = \text{median } \{ 0.0, 0.6, 1.0, 1.5, 7.6 \} = 1.0$$

8d. 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 48 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!!

Nature — Population/ Sample — Observation/ Data — Relationships/ Modeling — Analysis/ Synthesis

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 $\bar{X} = 161$ days.

What happens if we repeat the study? What will our next \bar{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 \bar{X} from one replication of the study to the next replication of the study is referred to as the sampling distribution of \bar{X} .

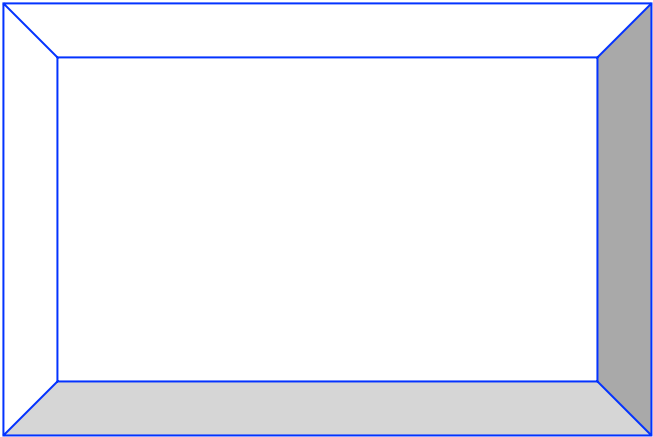
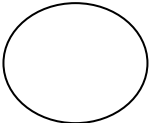
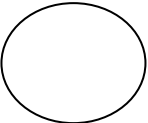
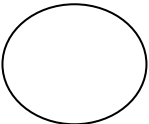
(We could just as well have asked about the behavior of the median from one replication to the next (sampling distribution of the median) or the behavior of the SD from one replication to the next (sampling distribution of SD).)

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

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

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.

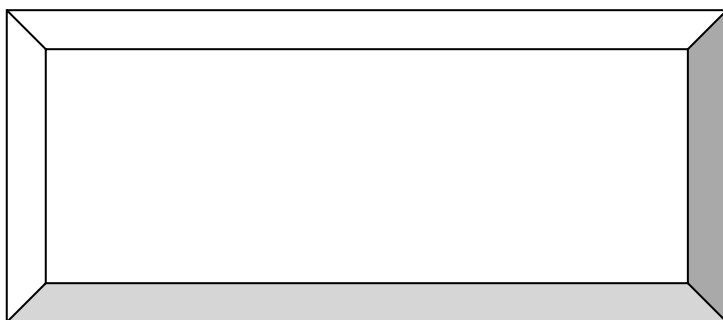
8e. A Feel for Sampling Distributions

				<p>This is a schematic of a population in nature. It is comprised of the universe of all possible individual values.</p> <p>Population mean = μ</p> <p>Population variance = σ^2</p>
sample	sample	sample	<p>This is a schematic representing the theoretical replication of study where each study replication (red arrow) involves a new sample of size n and the calculation of a new sample mean \bar{X}_n</p>
				
\bar{X}_n	\bar{X}_n		\bar{X}_n	



Collecting together all the \bar{X}_n ...

Nature ——— Population/ Sample ——— Observation/ Data ——— Relationships/ Modeling ——— Analysis/ Synthesis



This is a schematic of the resulting **sampling distribution of \bar{X}_n** . It is the collection of all possible sample means.

Sampling distribution has mean = $\mu_{\bar{X}_n}$

Sampling distribution has variance = $\sigma_{\bar{X}_n}^2$

Standard Deviation

- Describes variation in values of *individuals*.
- In the population of *individuals*: σ
- Our “guess” is S

Standard Error

- 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 all possible *sample means* (“sampling distribution of mean”): σ/\sqrt{n}
- Our “guess” of the SE of the sample mean is S/\sqrt{n}

8f. 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

$$cv = \hat{\xi} = \frac{S}{\bar{X}}$$

Example – “Cholesterol is more variable than systolic blood pressure”

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

8g. 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
- $\text{range} = 42 - 5 = 37$