Unit 2
Data Visualization

"Always graph the data."
(Source: - van Belle, G. Statistical Rules of Thumb, Second Edition. John Wiley & Sons, 2008)

In Unit 1, we learned how to summarize data in tables. Unit 2 is an introduction to the principles and methods of producing good ***visual*** summaries of data. In both units 1 and 2, the goal is to capture the "take home" message of the individual observations. The trick is to do it well and not mislead!

Here's a nice quote (and stated much better than I could do) to get us started:

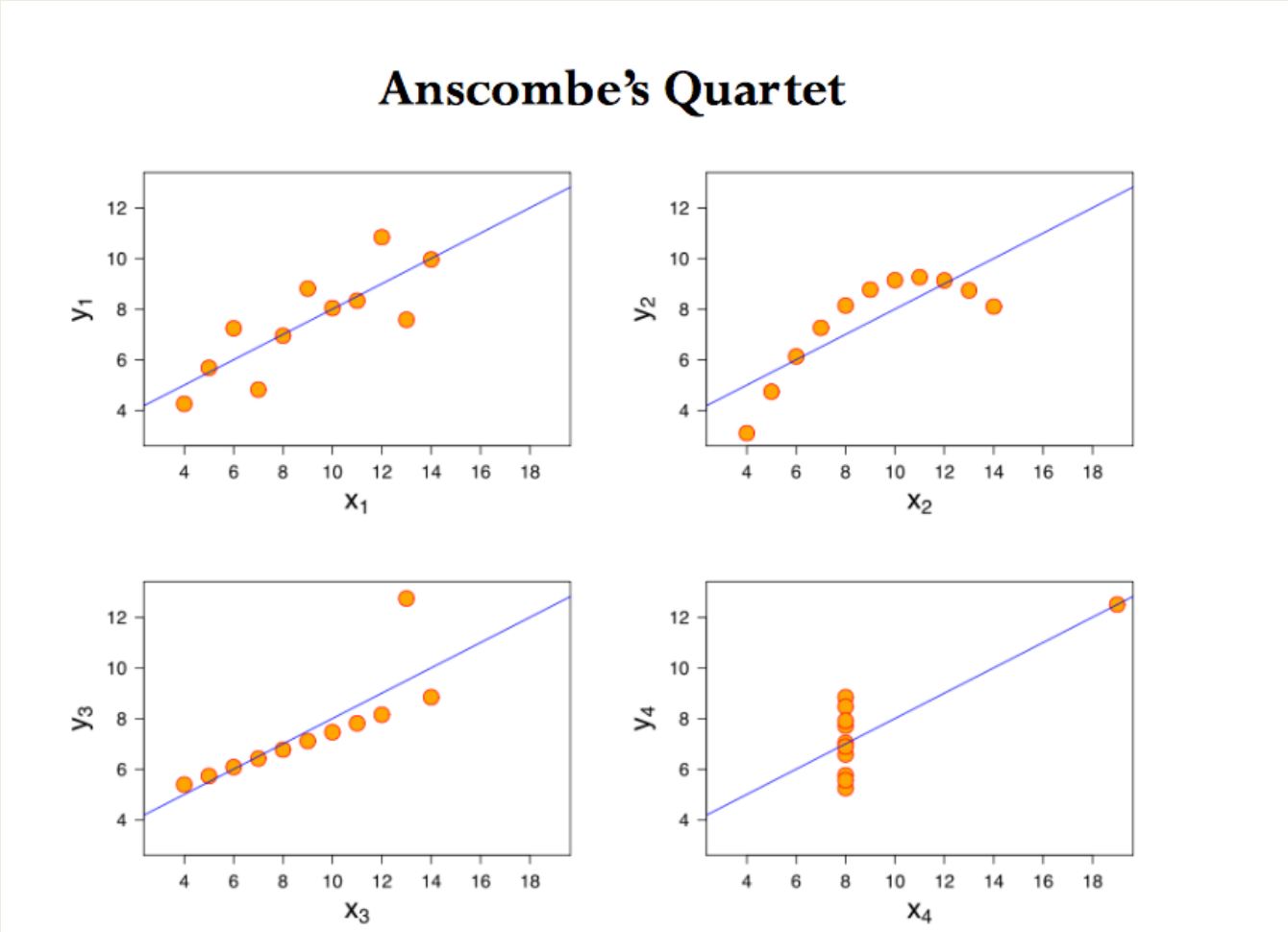
"Excellence in statistical graphics consists of complex ideas communicated with clarity, precision, and efficiency. Graphical displays should

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else
- avoid distorting what the data have to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure
- serve a reasonably clear purpose: description, exploration, tabulation, or decoration
- be closely integrated with the statistical and verbal descriptions of a data set."

source: E.R Tufte. *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, Connecticut, 1983. Page 121.

Want to follow along?
Right click to download from the course website:
#1. ICU data in excel format: https://people.umass.edu/biep540w/datasets/icu_540.xlsx
#2. ICU data as an R data set: https://people.umass.edu/biep540w/datasets/icu_540.Rdata

Thank you, Francis Anscombe!

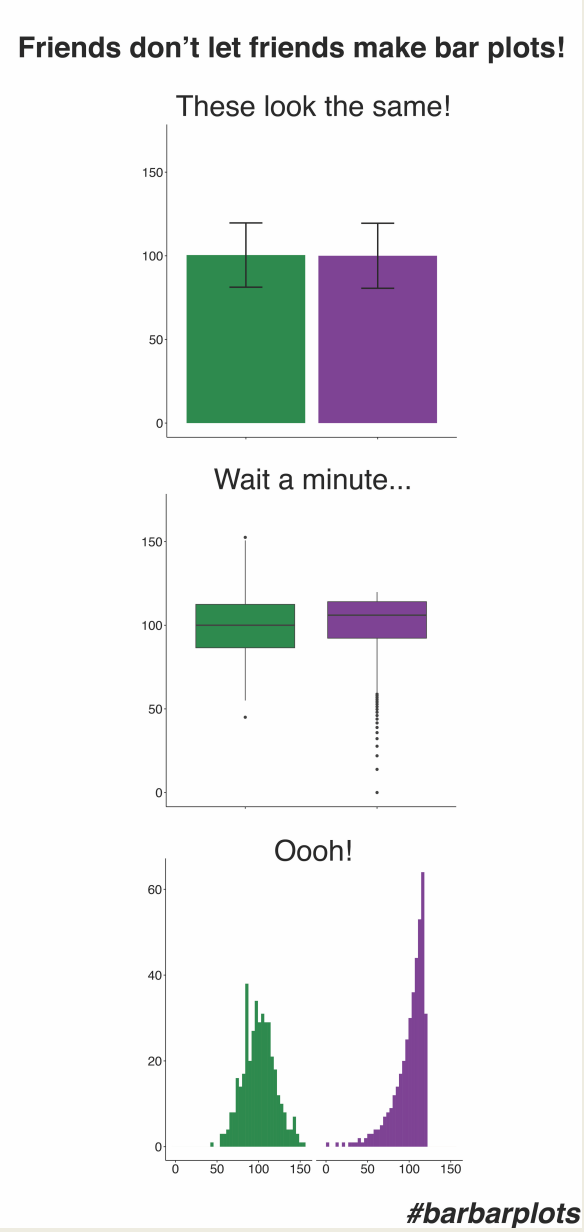


Always plot the data!

The four data sets are clearly very different. *And yet, all four data sets have the same: 1) sample mean and sample variance of X; 2) sample mean and sample variance of Y; 3) correlation between X and Y; and 4) linear regression of Y on X. To read more, visit:*

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Beware the Bar ± Error Plot (also known as "detonator plot")

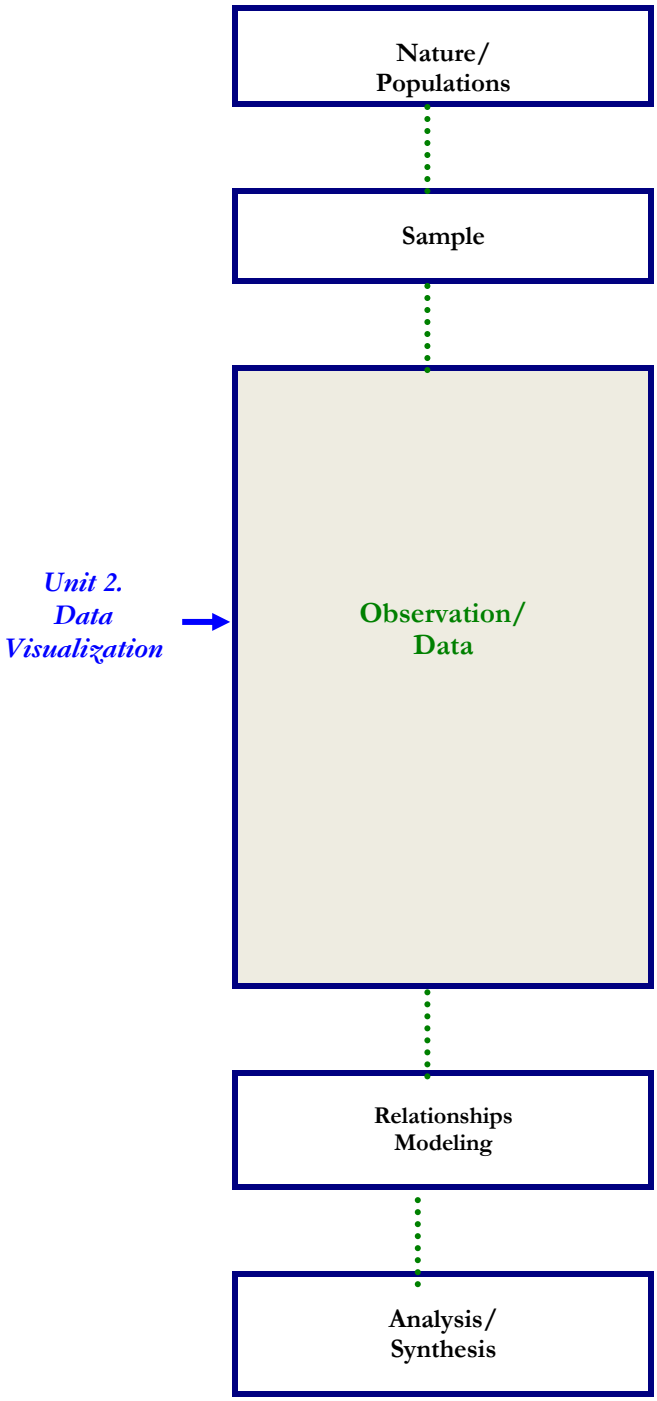


Do not construct bar plots of continuous data!

Bar plots are just fine for some purposes (e.g. nominal data) but are not meaningful for others (e.g. means ± error)

Table of Contents

1. Unit Roadmap



With luck, "Anscombe's Quartet" and the side-by-side "detonator plot" convinced you of the importance of data visualization. "Always graph the data", indeed! This unit is an introduction to producing visual summaries of data and emphasizes the importance of "good graphs".

A **"good graph"** is one that emphasizes the raw data, is not misleading (oh, yes, misleading graphs are quite possible!) and that reveals: 1) patterns and relationships; 2) potentially incorrect measurements, and; 3) unexpected patterns and relationships

Tips – Always provide a title and label your axes.

2. Learning Objectives

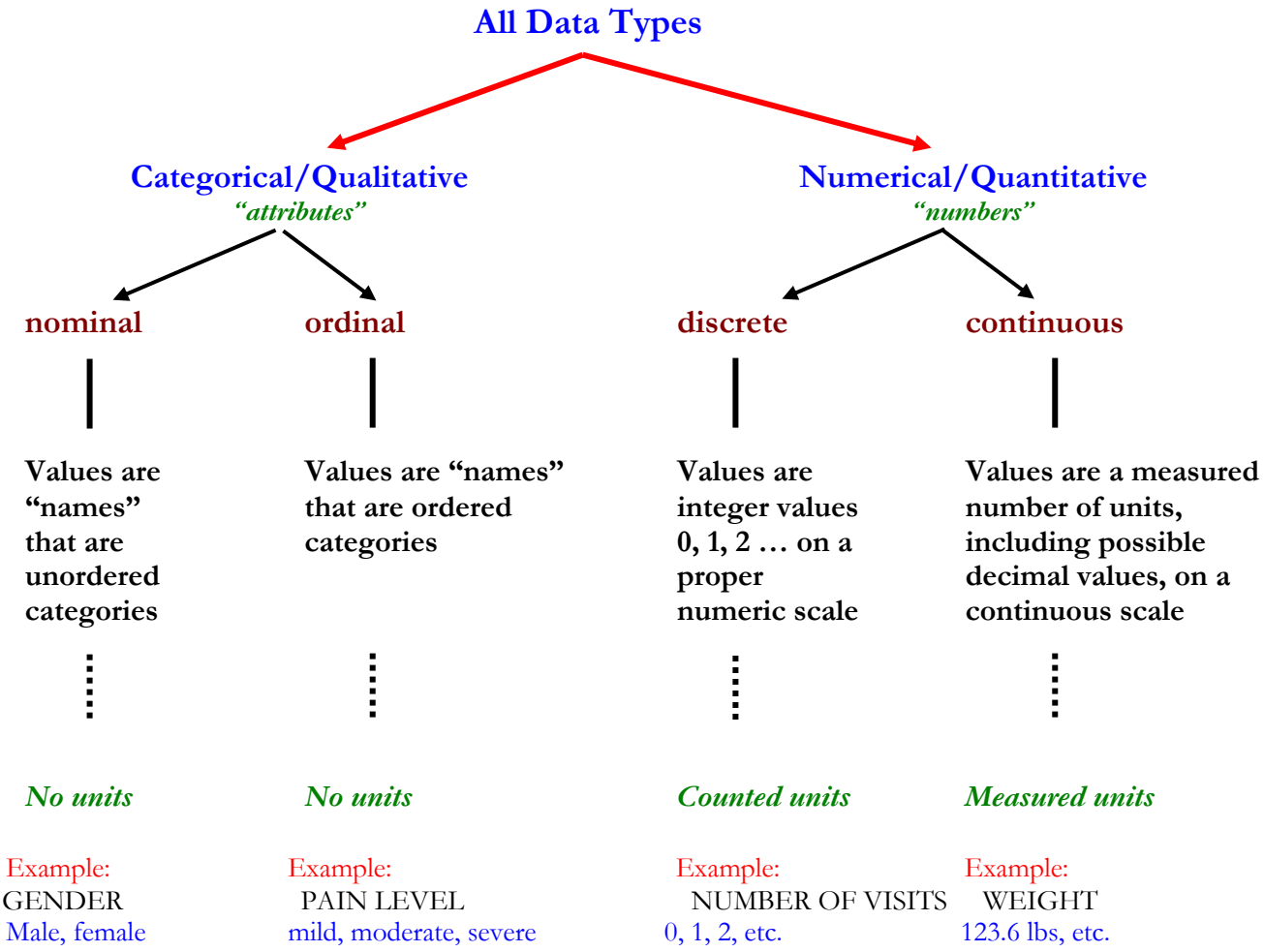
When you have finished this unit, you should be able to:

- State the facts shown and interpret a variety of basic graphs;
- Critically assess a published graph and know when it is misleading;
- For nominal/ordinal data: Understand and know "when and how" to construct bar charts;
- 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;
- Avoid pie charts;
- Avoid bar plots ± SE (also called "detonator" plots); and
- Understand the importance of the type of data when choosing which graph to produce.

3. Review: Variables and Types of Data

This review is a summary of selected lecture notes posted previously, "Variables and Types of Data" that can be found in the course notes for Unit 1 (Summarizing Data).

Review: It's important to appreciate that data may be of different data types because the appropriateness of any graph depends on the data type.



Source:
Adapted from: Daniel, Wayne W (*"Biostatistics – A Foundation for Analysis in the Health Sciences"*)

Another thing that is important to remember - 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

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

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.

- CATEGORICAL/QUALITATIVE ▶Nominal Scale: **Values are names which cannot be ordered.**
  Example: Cause of Death
  - Cancer
  - Heart Attack
  - Accident
  - Other

- CATEGORICAL/QUALITATIVE ▶Ordinal Scale: **Values are attributes (names) that are naturally ordered.**
  Example: Pain Level
  - None
  - Mild
  - Moderate
  - Severe

- NUMERICAL/QUANTITATIVE ▶Discrete Scale: **Values are counts of the number of times some event occurred**
  Example: Number of children a woman has had

- NUMERICAL/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

- NUMERICAL/QUANTITATIVE ▶ Continuous → Ratio **("meaningful zero")**: Continuous ratio data are also measured on a meaningful continuum with a meaningful zero point.
  Example: Weight in pounds

Finally, we learned that depending on the variable type, our options for how to summarize are different.

	All Data Types			
	Categorical *"qualitative"*		Numerical *"quantitative"*	
Type	Nominal	Ordinal	Discrete	Continuous
Graphical Summaries **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 was 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.

4. Single Sample: Categorical/Qualitative (*"attribute"*) Data

Example – We will use the same example that was used in notes, 1. Summarizing Data. Recall. This is a study of 25 consecutive patients entering the general medical/surgical intensive care unit at a large urban hospital.

- For each patient the following measurements (data) were obtained:

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

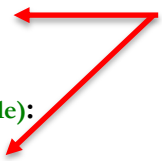
The actual data are provided on the following page.

Want to follow along?
Right click to download from the course website:
#1. ICU data in excel format: https://people.umass.edu/biep540w/datasets/icu_540.xlsx
#2. ICU data as an R data set: https://people.umass.edu/biep540w/datasets/icu_540.Rdata

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 **(scores do *NOT* represent magnitudes on a numerical scale):**

- **type_adm**: Type of Admission
- **icu_type**: ICU Type
- **vit_stat**: Vital Status at Hospital Discharge



Reminder to R learners:
R is case sensitive!

Quantitative data **(values *DO* represent magnitudes on a numerical scale):**

- **age**: Age, years
- **icu_los** Number of days spent in ICU
- **sbp**: Systolic blood

4a. Review: Frequency Table, Relative Frequency Table

Recall. We summarize nominal data by counting the number of occurrences of each outcome ("frequency") together with the percentages of each outcome ("relative frequency") in the sample. The possible outcomes, together with "how often" and "proportionately often" is called a **frequency and relative frequency distribution**.

4b. Bar Chart

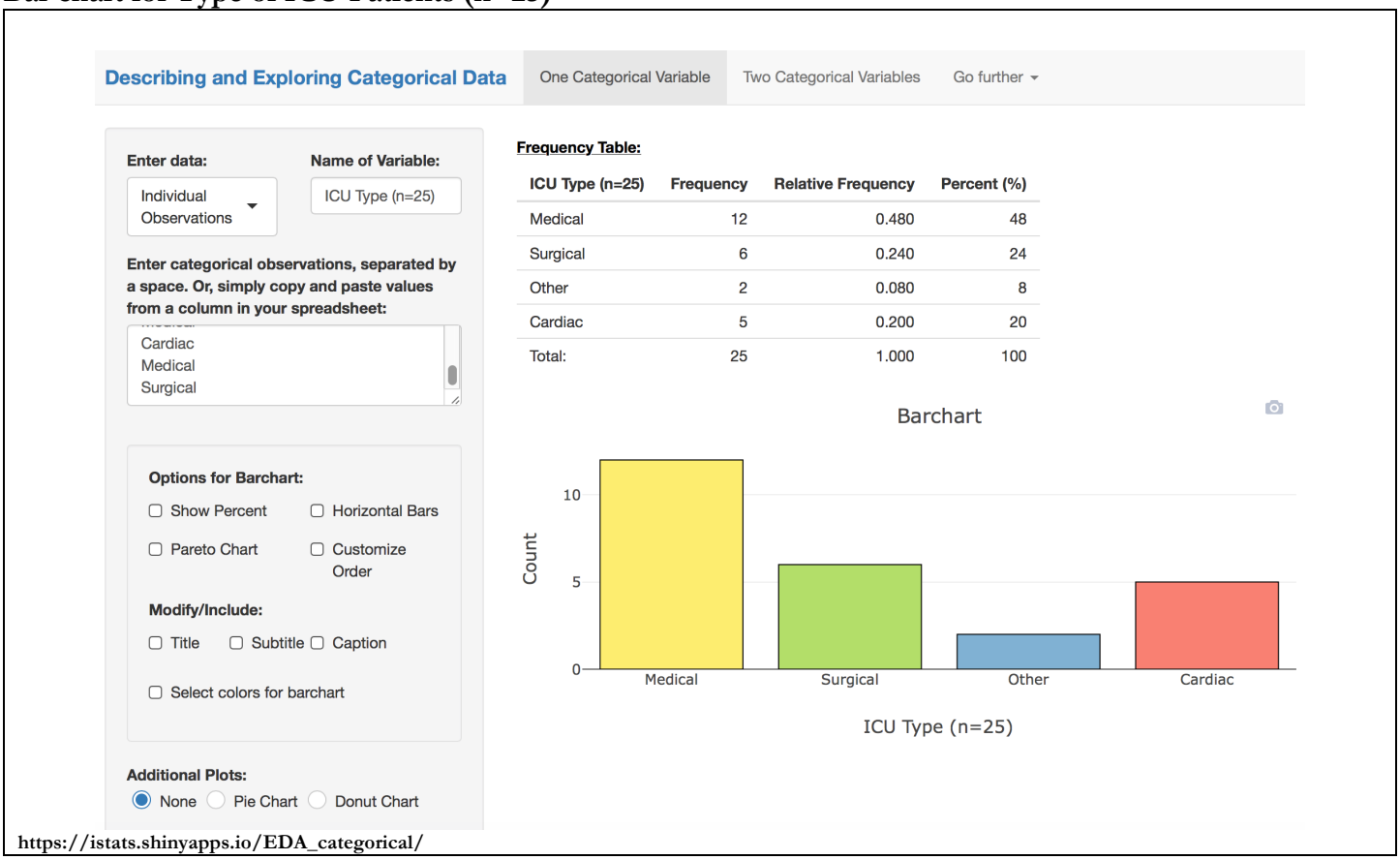
Definition: On the horizontal axis (also called the x-axis) plot the names of the nominal outcome.
On the vertical axis (also called the y-axis) plot the frequency and/or the relative frequency

Illustration Using Online Application – Bar Chart

Right click to download ICU data in excel format: https://people.umass.edu/biep540w/datasets/icu_540.xlsx

Step 1: Launch http://www.artofstat.com
Step 2: At right click ***Online Web Apps*** > ***EXPLORE CATEGORICAL DATA***
Step 3: At top, choose tab **One Categorical Variable**
Step 4: At left: At drop down box, **enter data**: Choose ***Individual Observations***
Step 5: Paste your data from excel (sheet 1)
Step 6: Play with the various options that you can change.

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



https://istats.shinyapps.io/EDA_categorical/

Tip (nifty!) – If you hover over one of the bars, artofstat.com will reward you with its explanation; eg -

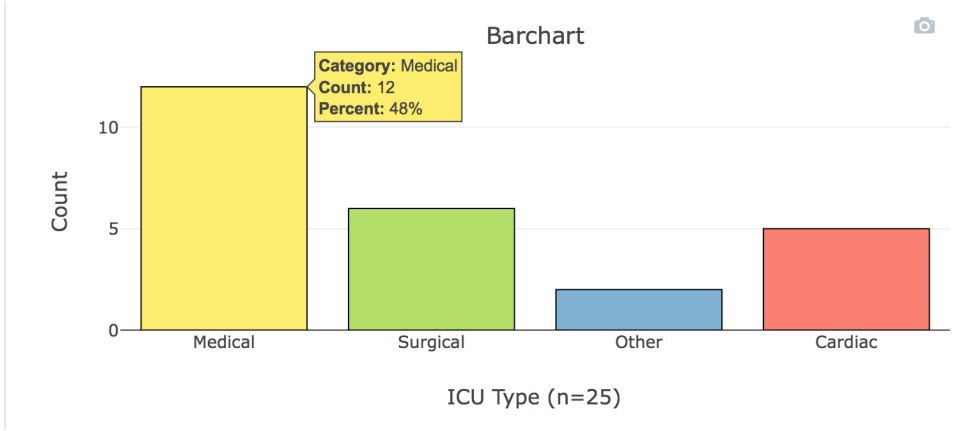
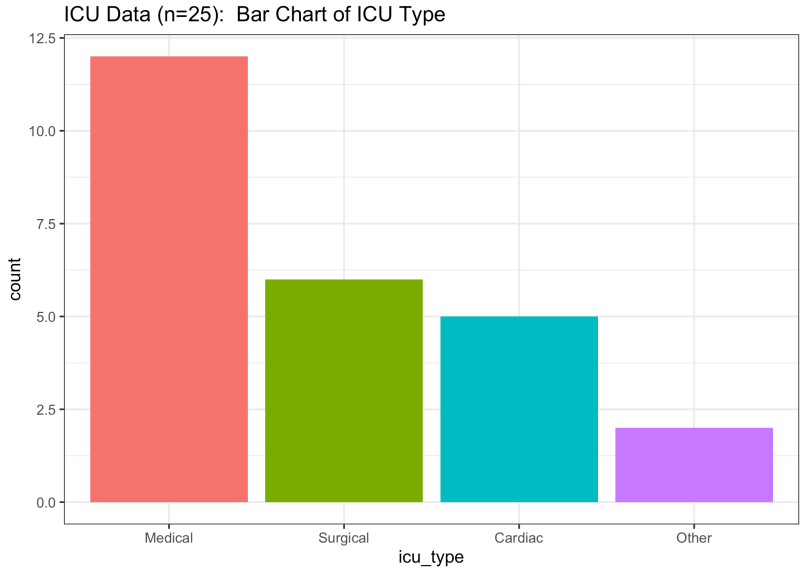


Illustration Using R – Bar Chart

Right click to download ICU data as an R data set: https://people.umass.edu/biep540w/datasets/icu_540.Rdata

```
# ggplot(data=DATAFRAME, aes(x=NOMINALVARIABLE,fill=NOMINALVARIABLE)) + geom_bar()
library(ggplot2)
ggplot(data=icudata,aes(x=icu_type,fill=as_factor(icu_type))) +
     geom_bar(show.legend = FALSE) +
     ggtitle("ICU Data (n=25):  Bar Chart of ICU Type") +
     theme(legend.position = "none") +
     theme_bw()
```

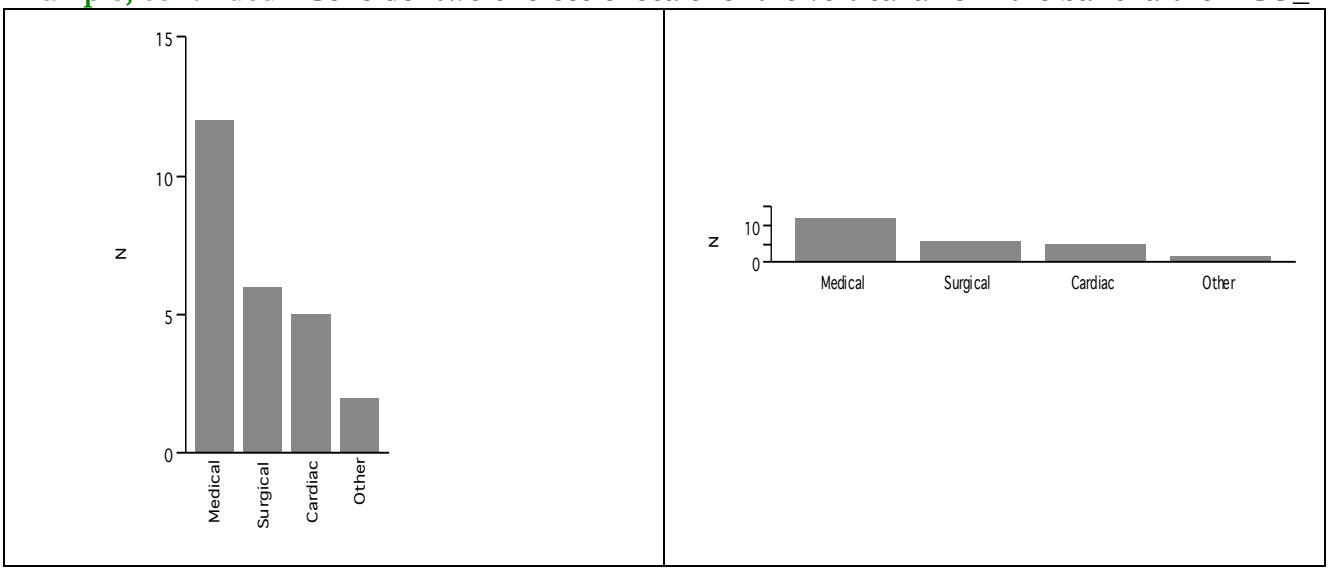


Rules for Construction of Bar Charts

- Label both axes clearly
- Leave space between bars
- Leave space between the left-most bar and the vertical axis
- 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. **Want to be misleading?** 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:

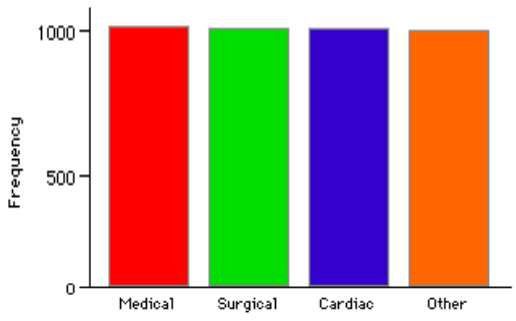
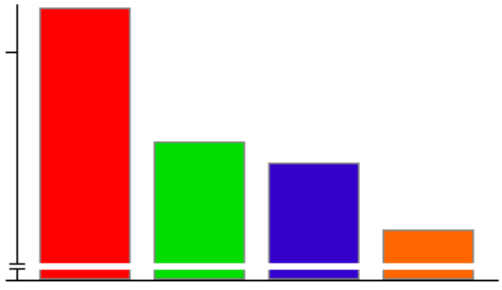


It's difficult to know how to construct a bar chart when the frequencies are very high.

- 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 -

Vertical Axis Starts at Zero	Vertical Axis has "Break"
	

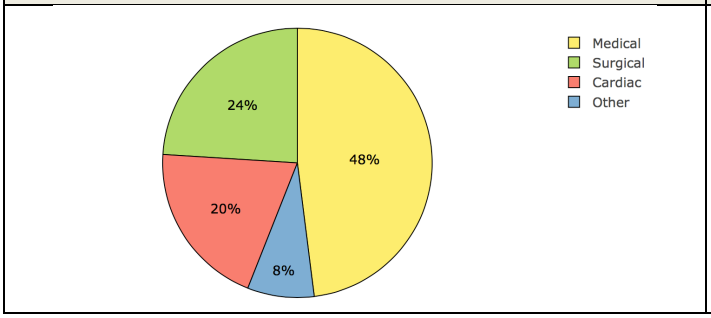
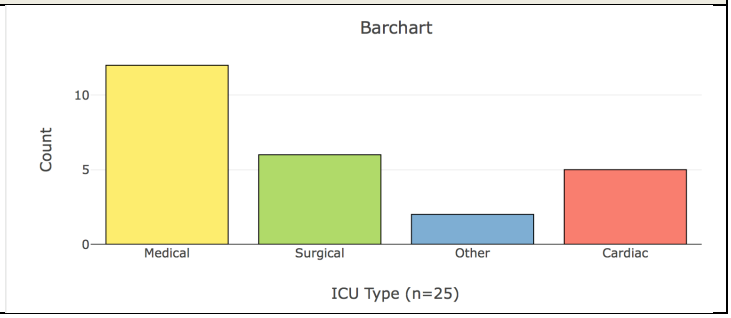
4c. Pie Chart

Pie charts are not recommended!!!!, but here is their explanation for completeness sake and so you appreciate why they are disliked. Briefly, whereas in a bar chart we plot bars rising up vertically (bar heights are either frequencies or relative frequencies), in a pie chart, we instead plot "shares" of a pie. How big are the "shares" or "wedges" of the pie? Recalling that a circle has 360 degrees (100% total), 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) x (100%) x 360 degrees

Example, continued –

A pie chart is one of the options in the Art of Stat online calculator we used to obtain a bar chart on page 12. Here, I show the bar chart and the pie chart side-by-side:

Pie Chart	Bar Chart
	

https://istats.shinyapps.io/EDA_categorical/

This is why pie charts are disliked - 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.

HOMEWORK DUE Friday September 30, 2022
Question #1 of 4

Download data (if needed) before you begin.
Art of Stat Users: nothing to download (right? You can enter this into excel, right?)
R Users: Right click to download https://people.umass.edu/biep540w/datasets/q1data.Rdata
Note – This data set is also located in Blackboard Learn and on the public course website

The World Health Organization (WHO) records the annual number of confirmed cases of human Avian Influenza A/(H5N1). Following are a subset of their data:

Year:	2003	2004	2005	2006	2007	2008
# cases:	4	32	43	79	59	26

a) By any means you like, construct a **frequency/relative frequency table** summarization of these data.

b) By any means you like, construct a **bar graph** summarization of these data. Include a title and label your axes.

c) State the **facts** of your **bar graph**.

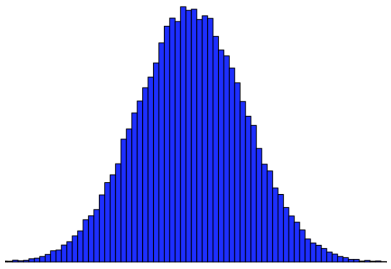
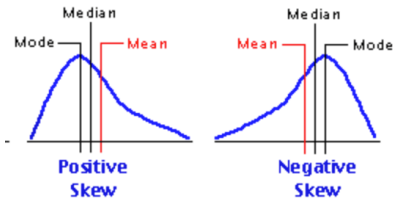
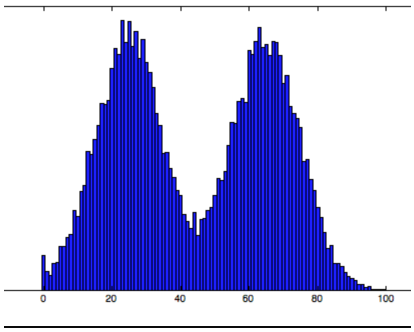
d) In 1-2 sentences, **interpret** the table and bar graph you have produced.

5. Single Sample: Numerical/Quantitative Data

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

- 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.

		
Symmetric	**Skewed**	**Bimodal**

The 3 patterns are quite different. The leftmost pattern is symmetric and bell shaped. The middle pattern shows two kinds of skewness, right (tail to the right) and left (tail to the left). The rightmost pattern is an illustration of bimodal data, here comprised of two symmetric bell shaped patterns that are separated in their central location.

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

5a. Histogram

Tip – Do you have numerical data that is measured on a continuum? If so, do NOT produce a frequency distribution!!!!!
WHY: 1) the result is a catalogue, not a summarization; and 2) the list will be very long (you'll be really annoyed, I promise).

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 numerical continuous random variable. More formally, it is a graphical summary of the frequency distribution.**

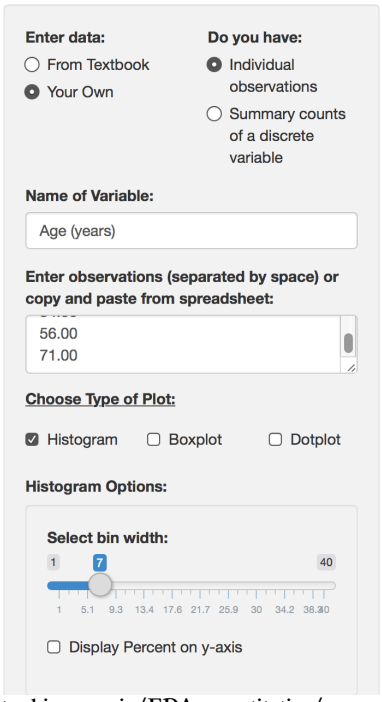
- **It is analogous to the bar graph summary for the distribution of a discrete random variable**

Illustration Using Online Application – Histogram

Right click to download ICU data in excel format: https://people.umass.edu/biep540w/datasets/icu_540.xlsx

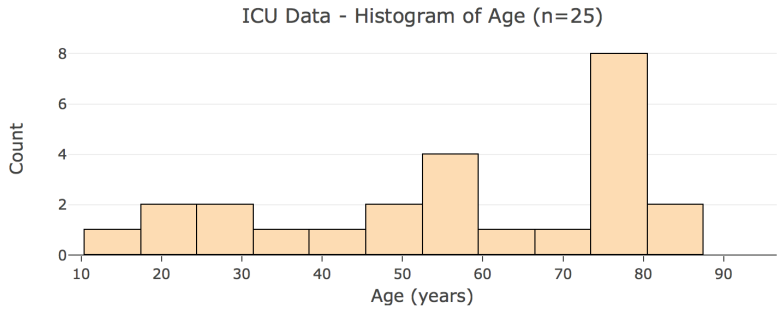
Step 1: Launch http://www.artofstat.com
Step 2: At right click ***Online Web Apps*** > ***EXPLORE QUANTITATIVE DATA***
Step 3: At top, choose tab **Single Group**
Step 4 At drop down box, **enter data**: Choose ***Your Own*** and ***Individual Observations***
Step 5: (If needed) Paste your data from excel
Step 6: Scroll down to select TYPE OF PLOT: histogram

Histogram of Age of ICU Patients (n=25)



Descriptive Statistics:

Sample Size	Mean	Std. Dev.	Minimum	Q1	Median	Q3	Maximum	IQR
25	56.920	22.179	15.000	39.000	56.000	76.000	85.000	37.000



https://istats.shinyapps.io/EDA_quantitative/

Similarly… – If you hover over one of the bars, artofstat.com will reward you with its explanation. Here we see that the ICU data includes a count (frequency) 4 observations of age that are between 53 and 59 years.

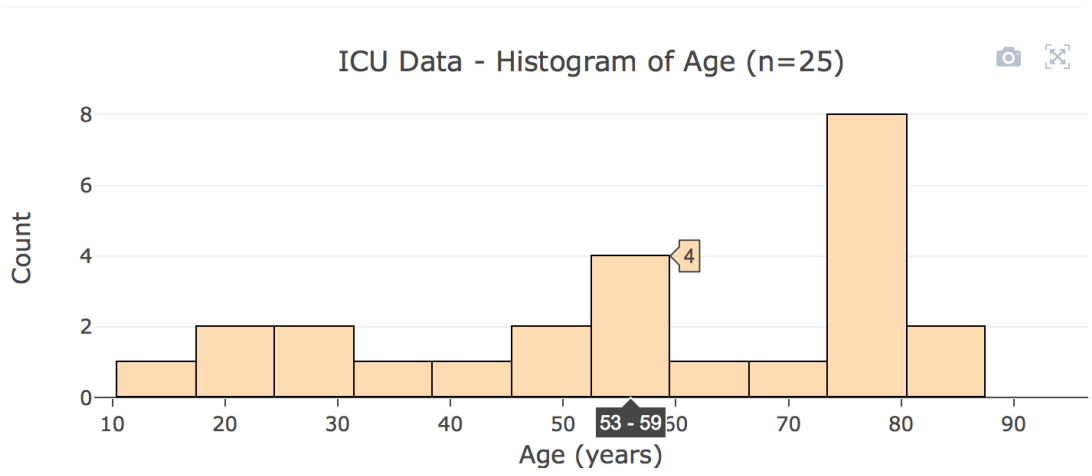
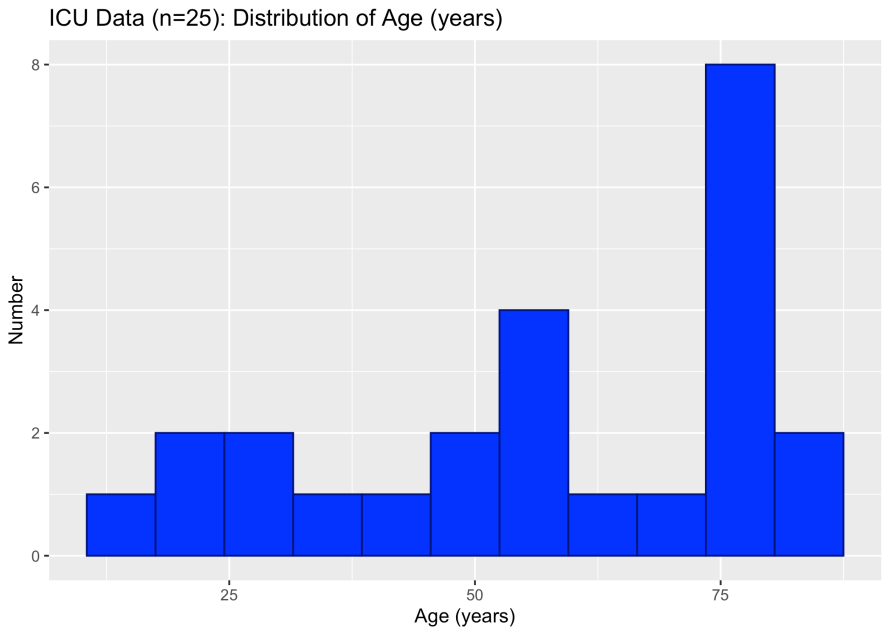


Illustration Using R – Single Histogram

Right click to download ICU data as an R data set: https://people.umass.edu/biep540w/datasets/icu_540.Rdata

```
# ggplot(data=DATAFRAME, aes(x=NOMINALVARIABLE,fill=NOMINALVARIABLE)) + geom_histogram()
ggplot(data=icudata, aes(x=age)) +
  geom_histogram(binwidth=7, color="navy", fill="blue") +
  ggtitle("ICU Data (n=25): Distribution of Age (years)") +
  labs(y="Number", x="Age (years)")
```



Rules for How to Construct a Histogram

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

- The choice is arbitrary. A formula is provided in the text (Sturge's Rule) but the authors note that this is not a rigid rule. Instead, aim for a number of intervals that will allow you to see patterns in the data, in particular ….
- K too small over-summarizes. K too large under-summarizes.
- Sometimes, the choices of intervals are straightforward; they just make sense – eg 10 year intervals, 7 day intervals, 30 day intervals.
- Artofstat.com, R, and Stata all let you try different choices of number of intervals, k.
- Example – For the age data, we might 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 ***un***equal widths, take care to plot "area proportional to relative frequency". This is explained on the next page.

IMPORTAT POINT: In a histogram, "Area is always proportional to relative frequency"

To understand what this is saying, 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 is plotted out as a frequency of 2 over 10 units of age done twice: 2 over the interval 10-19 and then the other 2 over the interval 20-29. →
- Thus, putting them together we have that, for the interval 10-29, a height of 2 would be plotted.

Histograms with bins of varying width are possible but is beyond the scope of this course.

Illustration Using R – Side by Side Histogram

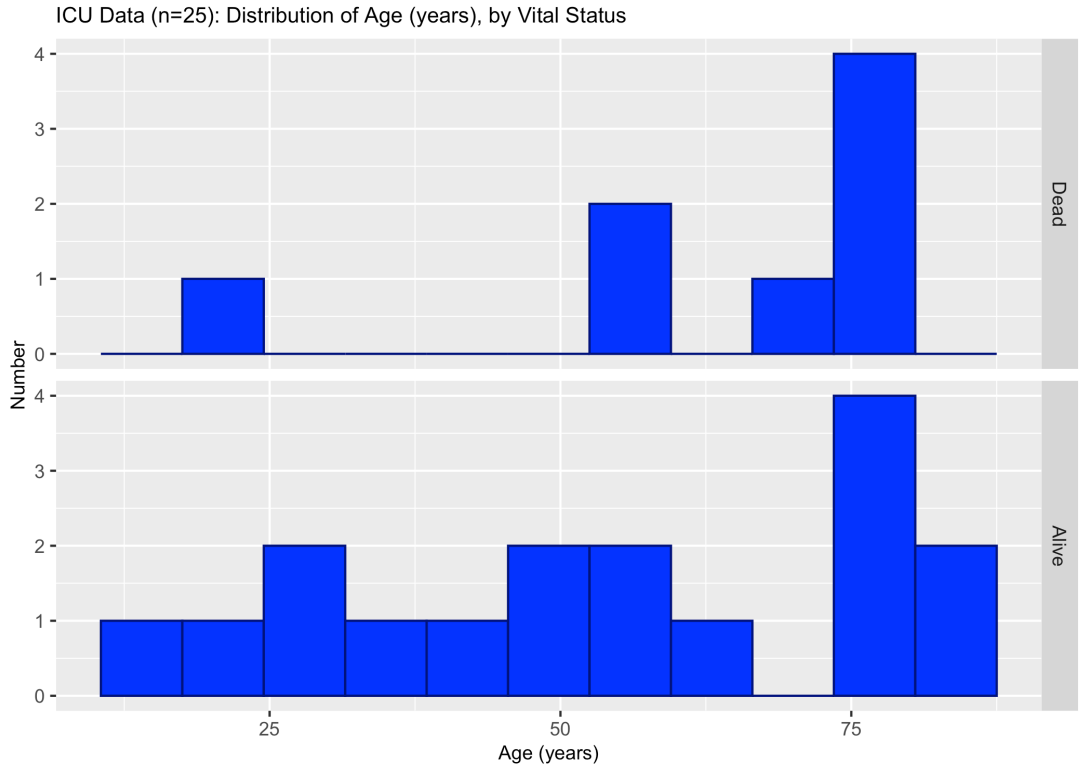
Right click to download ICU data as an R data set: https://people.umass.edu/biep540w/datasets/icu_540.Rdata

```
library(ggplot2)

# ggplot(data=DATAFRAME, aes(x=NOMINALVARIABLE,fill=NOMINALVARIABLE)) +
         geom_histogram() + facet(GROUPVARIABLE ~ .)

# Preliminary - label the levels of the grouping variable
icudata$vit_stat <- factor(icudata$vit_stat,
                           levels=c(1,0),
                           labels=c("Dead","Alive"))

ggplot(data=icudata, aes(x=age)) +
     geom_histogram(binwidth=7, color="navy", fill="blue") +
     facet_grid(vit_stat ~ .) +
     ggtitle("ICU Data (n=25): Distribution of Age (years), by Vital Status") +
     labs(y="Number", x="Age (years)") +
     scale_color_grey()+scale_fill_grey()
```



HOMEWORK DUE Friday September 30, 2022
Question #2 of 4

Download data (if needed) before you begin.
Art of Stat Users: Right click to download https://people.umass.edu/biep540w/datasets/cholesterol_540.xlsx
R Users: Right click to download TWO R data sets:
https://people.umass.edu/biep540w/datasets/cholesterol_smokers.Rdata
https://people.umass.edu/biep540w/datasets/cholesterol_nonsmokers.Rdata
Note – These data sets are also located in Blackboard Learn and on the public course website
A study examining the health risks of smoking measured the cholesterol (mg/dL) levels of people in two independent groups: 1) SMOKERS: those who had smoked for at least 25 years; and 2) NON-SMOKERS: persons of similar ages who had never smoked. The following are the data.

Smokers			
225	211	209	284
258	216	196	288
250	200	209	280
225	256	243	200
213	246	225	237
232	267	232	216
216	243	200	155
216	271	230	309
183	280	217	305
287	217	246	351
200	280	209	

NON-Smokers		
250	213	300
249	213	310
175	174	328
160	188	321
213	257	292
200	271	227
238	163	263
192	242	249
242	267	243
217	267	218
217	183	228

a) By any means you like, produce a **histogram** of cholesterol for **SMOKERS**
b) By any means you like, produce a **histogram** of cholesterol for **NON-SMOKERS**
c) (Be adventurous – there is not an example in the lecture notes. I want you to explore a bit on your own) Play with artofstat.com to produce side-by-side histograms (Hint: at top click on "several groups")
d) In 1-2 sentences**, state the facts** of these histograms.
e) In 1-2 sentences, provide an **interpretation** of the comparison of these histograms.

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

Tips -

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.

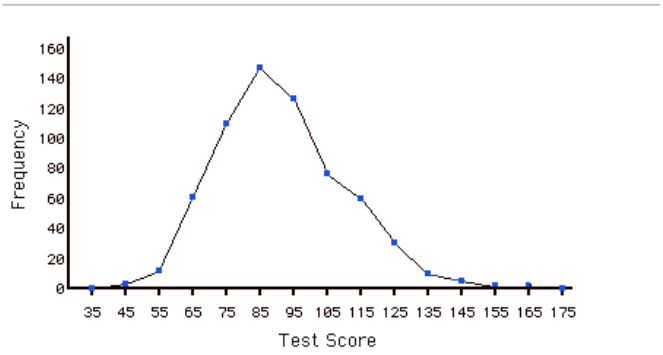


Figure 1: Frequency polygon for the psychology test scores.

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

5c. Cumulative Frequency Polygon

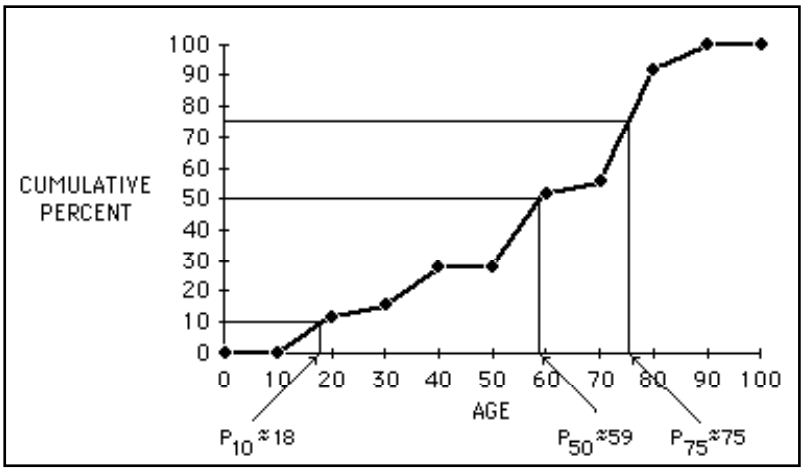
As you might think, a cumulative frequency polygon depicts accumulating data.

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)	Cumulative through Interval 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. Review: Percentiles (Quantiles)

Recall. 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.

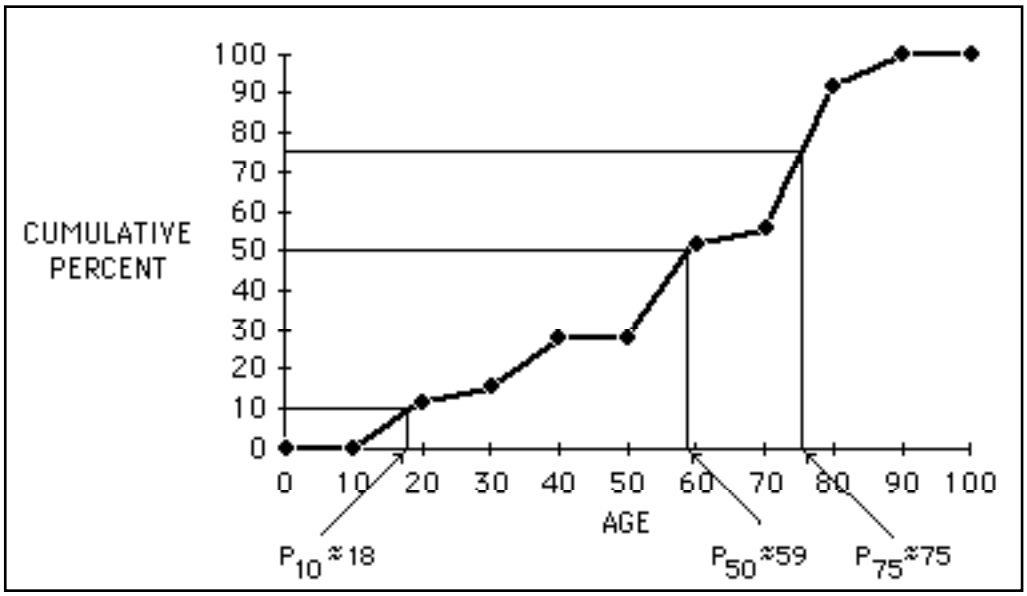
We say 4 years is the median.

The median is also called the 50th percentile, or the .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.
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 translated 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 (other than letting the computer do it!)
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. Review: Five Number Summary

Recall. 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 27, 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. Review: Interquartile Range (IQR)

Recall. 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**

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

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

- We might ask: *"Can I assume that the distribution of my data is normal (Gaussian)?"*
  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**

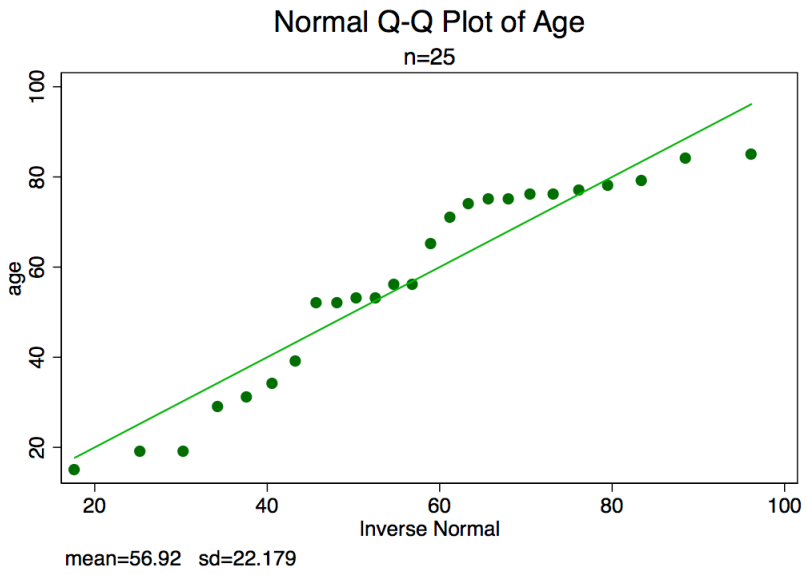
Desired Quantile	**10th**	**20th**	**…**	**99th**
X = Value in Sample of Data				
Y = Value in Reference Distribution				

- **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!**

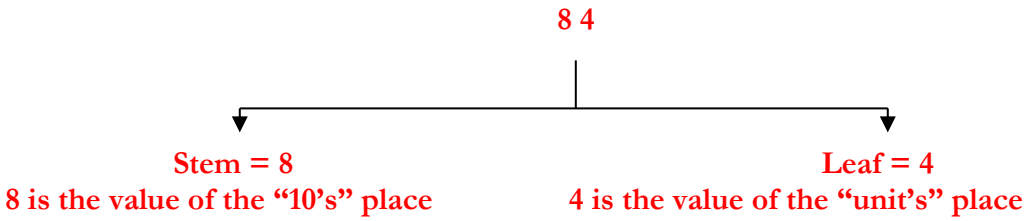
Example – Normal Quantile-Quantile Plot of the ICU data on age that is on page 11



5h. Stem and Leaf Diagram

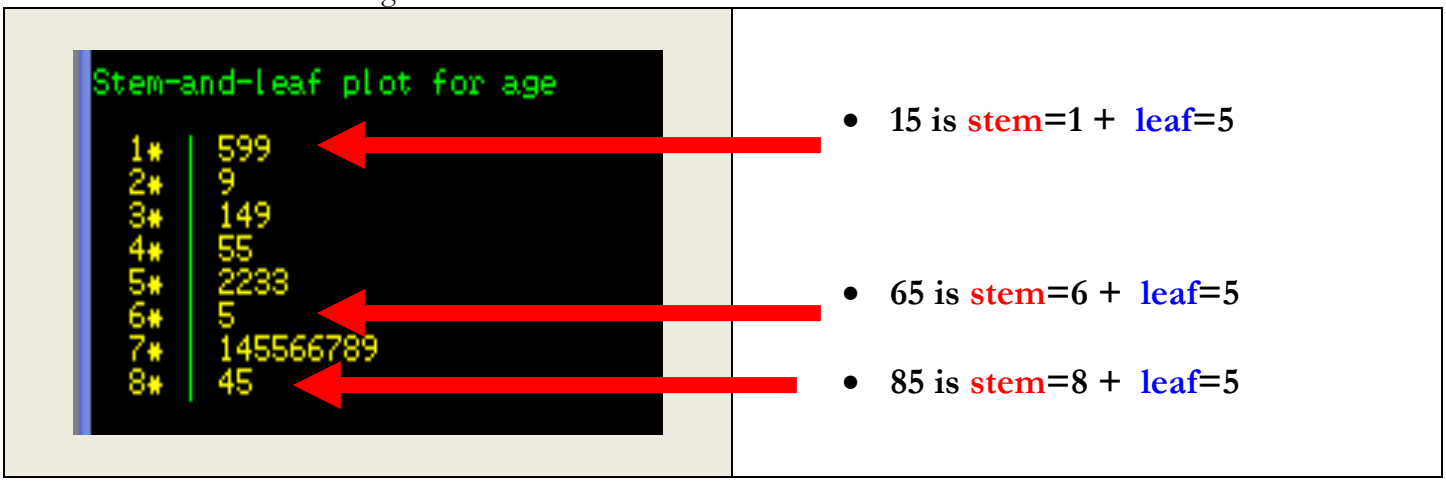
A stem and leaf diagram is a "quick and dirty" histogram. It is also a quick and easy way to sort data. Each actual data point is de-constructed into a stem and a leaf. There are a variety of ways to do this, depending on the sample size and range of values.

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



Example - Stem and Leaf Plot of the ICU data on 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.



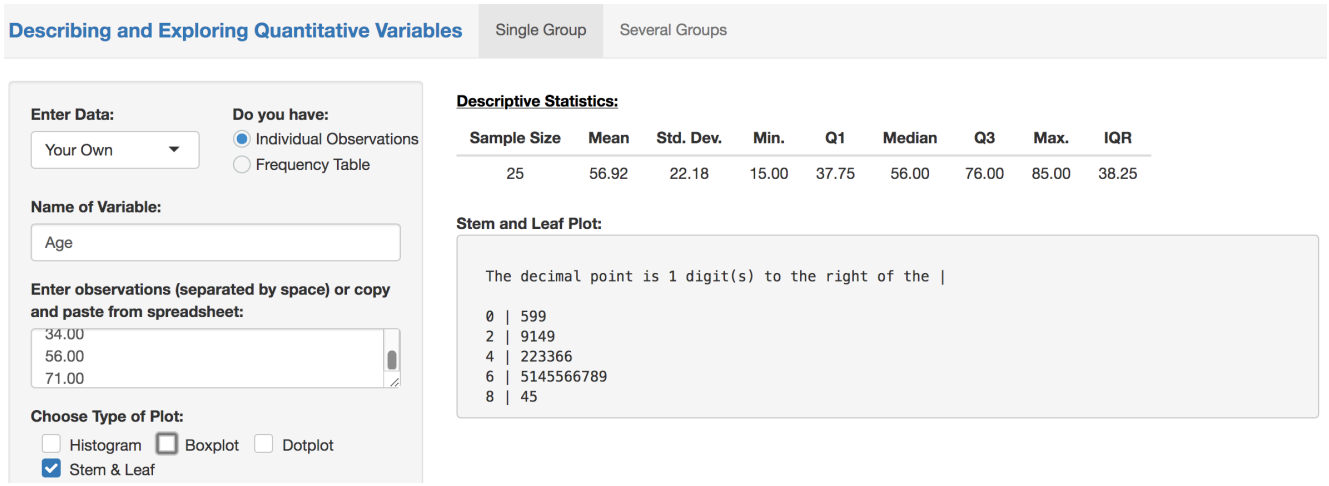
Key - Among other things, we see that the minimum age in this sample is 15 years (Note the unit=5 in the row for stem=1) and the maximum age is 85 years (Unit=5 in the row for stem=8). We can also see that most of the subjects in this sample are in their seventies (Actual ages are 71, 74, 75, 75, 76,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

Illustration Using Online Application – Stem & Leaf

Right click to download ICU data in excel format: https://people.umass.edu/biep540w/datasets/icu_540.xlsx

Step 1: Launch http://www.artofstat.com
Step 2: At right click ***Online Web Apps*** > ***EXPLORE QUANTITATIVE DATA***
Step 3: At top, choose tab **Single Group**
Step 4 At drop down box, **enter data**: Choose ***Your Own*** and ***Individual Observations***
Step 4: (If needed) Paste your data from excel
Step 6: Scroll down to choose TYPE OF PLOT: Stem & Leaf

Stem and Leaf Diagram of Distribution of Age of ICU Patients (n=25)



https://istats.shinyapps.io/EDA_quantitative/

Illustration Using R – Single Variable Stem and Leaf

Dear Reader – This is not a ggplot and no file is produced. I did a screen capture and pasted it to here

Right click to download ICU data as an R data set: https://people.umass.edu/biep540w/datasets/icu_540.Rdata

```
# stem(DATAFRAME$VARIABLENAME)
stem(icudata$age)
```

```
The decimal point is 1 digit(s) to the right of the |

0 | 599
2 | 9149
4 | 223366
6 | 5145566789
8 | 45
```

Illustration Using R – Side by Side Variable Stem and Leaf

Right click to download ICU data as an R data set: https://people.umass.edu/biep540w/datasets/icu_540.Rdata

```
# NOTE:  The following code assumes that you have run the following code that converts vit_stat to a factor variable
icudata$vit_stat <- factor(icudata$vit_stat,
                           levels=c(1,0),
                           labels=c("Dead","Alive"))

dead <- subset(icudata,vit_stat=="Dead",select=c(age))          # extract distribution of age for vit_stat="Dead"
alive <- subset(icudata,vit_stat=="Alive",select=c(age))        # extract distribution of age for vit_stat="Alive"

library(aplpack)                                                # CHECK:  Have you first installed the package aplpack?
stem.leaf.backback(dead$age,alive$age)                          # stem.leaf.backback( ) for back to back stem and leaf
```

```
_____________________________
  1 | 2: represents 12, leaf unit: 1
    dead$age     alive$age
_____________________________
   1        9| 1 |59       2
             | 2 |9        3
             | 3 |149      6
             | 4 |
   3       66| 5 |2233    (4)
             | 6 |5        7
  (5)   65541| 7 |6789     6
             | 8 |45       2
             | 9 |
_____________________________
n:           8     17
_____________________________
```

Personally, I don't find this picture helful

```
stem(dead$age)
stem(alive$age)
```

Dead

```
The decimal point is 1 digit(s) to the right of the

  0 | 9
  2 |
  4 | 66
  6 | 14556
```

Alive

```
The decimal point is 1 digit(s) to the right of the

  0 | 59
  2 | 9149
  4 | 2233
  6 | 56789
  8 | 45
```

Now, don't you think this is a better visual?
Sometimes, simpler really is better.

HOMEWORK DUE Friday September 30, 2022
Question #3 of 4

Download data (if needed) before you begin.
Art of Stat Users: Right click to download https://people.umass.edu/biep540w/datasets/cholesterol_540.xlsx
R Users: Right click to download TWO R data sets:
https://people.umass.edu/biep540w/datasets/cholesterol_smokers.Rdata
https://people.umass.edu/biep540w/datasets/cholesterol_nonsmokers.Rdata
Note – These data sets are also located in Blackboard Learn and on the public course website

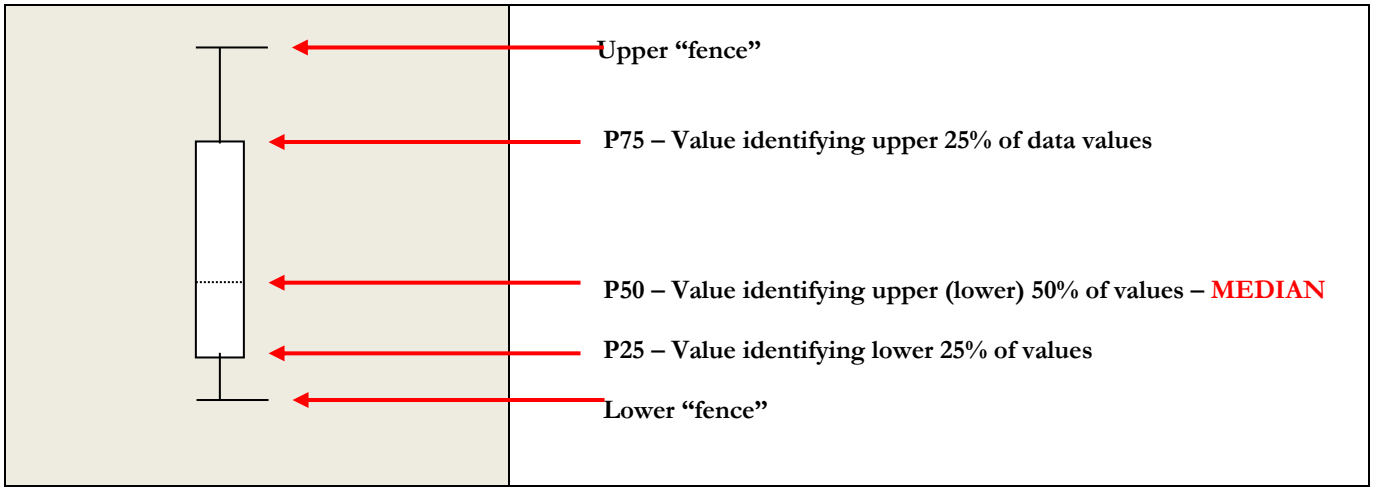
Consider again the cholesterol (mg/dL) data in smokers and non-smokers introduced in question #2.

a) By any means you like, produce a **side-by-side stem and leaf** diagram.
b) By any means you like, complete the following table:

	Smokers	NON-Smokers
Number in group, n =		
P_{25} = Q1 = Lower Quartile =		
P_{50} = Q2 = Median Quartile =		
P_{75} = Q3 = Upper Quartile =		
Interquartile Range (IQR) =		
1.5* IQR=		
Value of Lower Fence =		
Value of Upper Fence =		
Outliers (if any) below lower fence (LIST) =		
Outliers (if any) above upper fence (LIST) =		

5i. Box and Whisker Plot

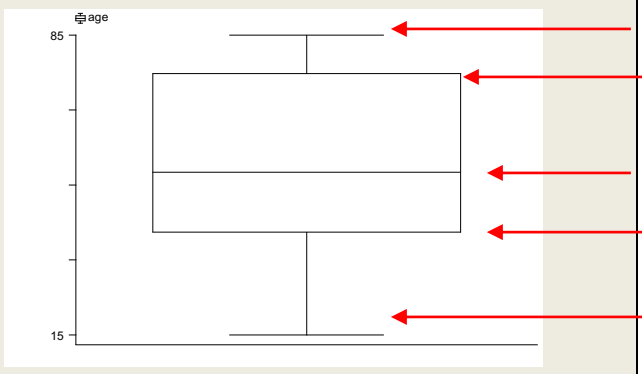
- The box and whisker plot (also called box plot) is a wonderful schematic summary of the distribution of values in a data set.
- It shows you a number of features, including: extremes, 25th and 75th percentile, median and sometimes the mean.
- Side-by-side box and whisker plots are a wonderful way to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

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



85 years = largest ≤ P_{75} + 1.5*(P_{75} – P_{25})

P_{75} = 76 years

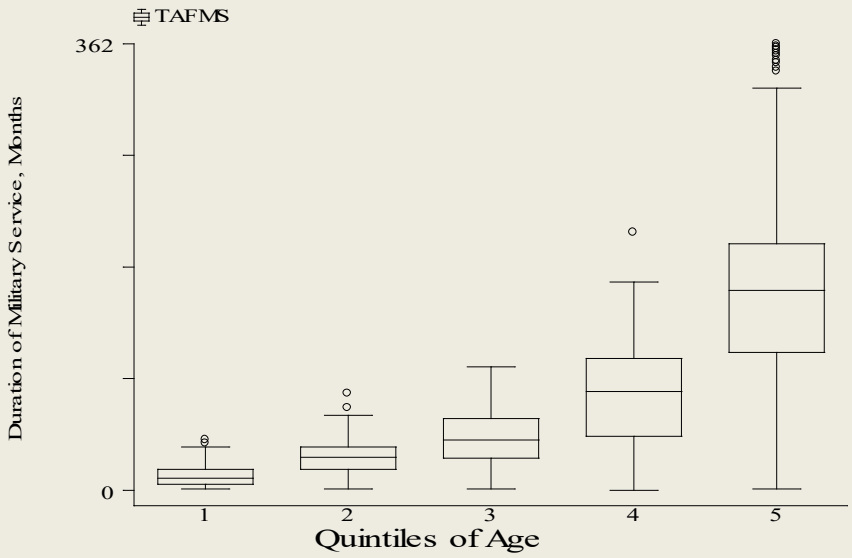
Median, P_{50} = 53 years

P_{25} = 39 years

15 years = smallest ≥ P_{25} - 1.5*(P_{75} – P_{25})

Note: min=15, P_{25}=39, P_{50}=53, P_{75}=76, max=85, and 1.5*(P_{75}-P_{25})=55.5

Example of Interpretation – The following is an example of side-by-side box and whisker plots. They are plots of values of duration of military service (Y-axis) across 5 groups defined by age (X-axis).



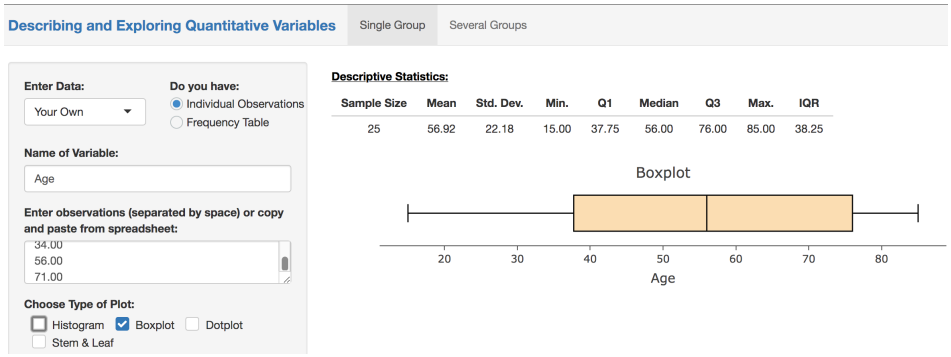
- Duration of military service increases with age.
- With age, the variability in duration of military service is greater.
- The individual circles represent extreme values.

Illustration Using Online Application – Box & Whisker

Right click to download ICU data in excel format: https://people.umass.edu/biep540w/datasets/icu_540.xlsx

Step 1: Launch http://www.artofstat.com
Step 2: At right click ***Online Web Apps*** > ***EXPLORE QUANTITATIVE DATA***
Step 3: At top, choose tab **Single Group**
Step 4 At drop down box, **enter data**: Choose ***Your Own*** and ***Individual Observations***
Step 5: (If needed) Paste your data from excel
Step 6: Scroll down to choose TYPE OF PLOT: Boxplot

Box & Whisker Plot of Distribution of Age of ICU Patients (n=25)



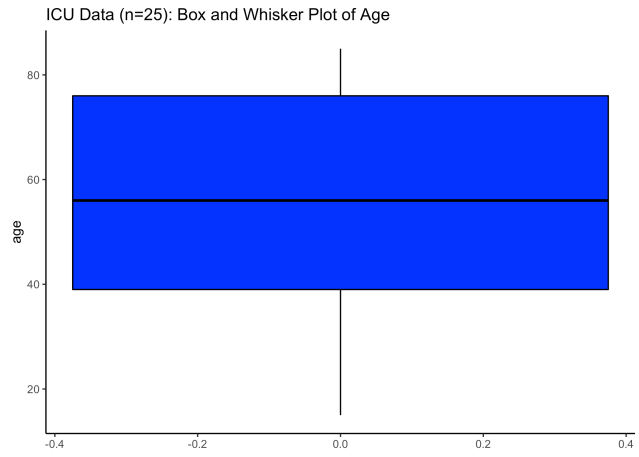
https://istats.shinyapps.io/EDA_quantitative/

Illustration Using R – Single Variable Box & Whisker Plot

Right click to download ICU data as an R data set: https://people.umass.edu/biep540w/datasets/icu_540.Rdata

```
library(ggplot2)

# ggplot(data=DATAFRAME, aes(y=CONTINUOUSVARIABLE)) + geom_boxplot()
ggplot(data=icudata, aes(y=age)) +
  geom_boxplot(fill='blue', color="black") +
  ggtitle("ICU Data (n=25): Box and Whisker Plot of Age") +
     theme_classic()
```



HOMEWORK DUE Friday September 30, 2022
Question #4 of 4

Download data (if needed) before you begin.
Art of Stat Users: Right click to download https://people.umass.edu/biep540w/datasets/cholesterol_540.xlsx
R Users: Right click to download TWO R data sets:
https://people.umass.edu/biep540w/datasets/cholesterol_smokers.Rdata
https://people.umass.edu/biep540w/datasets/cholesterol_nonsmokers.Rdata
Note – These data sets are also located in Blackboard Learn and on the public course website

Consider again the cholesterol (mg/dL) data in smokers and non-smokers introduced in question #2.

Smokers			
225	211	209	284
258	216	196	288
250	200	209	280
225	256	243	200
213	246	225	237
232	267	232	216
216	243	200	155
216	271	230	309
183	280	217	305
287	217	246	351
200	280	209	

NON-Smokers		
250	213	300
249	213	310
175	174	328
160	188	321
213	257	292
200	271	227
238	163	263
192	242	249
242	267	243
217	267	218
217	183	228

a) By any means you like, produce a **side-by-side boxplot**
Hint to Art of Stat Users - You will need to click on tab (at top right) to enter multiple groups

Appendix
Summary of R Code

```
#1.  Single Variable, Nominal – Bar Chart

# Basic command (user edits yellow)
library(ggplot2)
ggplot(data=DATAFRAME) +                  # dataset to use
       aes(x=NOMINALVARIABLE) +           # nominal variable to plot
       geom_bar(option, option) +         # geom_bar( ) to obtain bar plot
       additions as you like

# Example – with added aesthetics
library(ggplot2)
ggplot(data=icudata) +
    aes(x=icu_type) +
    geom_bar(fill="steelblue") +
    ggtitle("ICU  Data (n=25)")
```

```
#2.  Single Variable, Continuous – Histogram

# Basic command (user edits yellow)
library(ggplot2)
ggplot(data=DATAFRAME) +                  # dataset to use
        aes(x=CONTINUOUSVARIABLE) +       # continuous variable to plot
        geom_histogram(option, option) +  # geom_histogram( ) to obtain histogram
        additions as you like

# Example – with added aesthetics
library(ggplot2)
ggplot(data=icudata) +
      aes(x=age) +
      geom_histogram(binwidth=10, color="navy", fill="blue") +
      ggtitle("ICU Data (n=25): Distribution of Age (years)") +
      xlab("Age(years)") +
      ylab("Number") +
      ggtitle("ICU Data (n=25)")
```

```
#3.  Single Continuous Variable (age) & One Grouping Variable (vit_stat), – Side by Side Histogram

# Basic command (user edits yellow)
library(ggplot2)
ggplot(data=DATAFRAME) +                   # dataset to use
       aes(x=CONTINUOUSVARIABLE) +         # continuous variable to plot
       geom_histogram(option, option) +    # geom_histogram( ) to obtain histogram
       facet_grid(GROUPVARIABLE ~ .) +     # facet_grid(groupingvariable ~ .) will produce side-by-side by grouping var.
       additions as you like


# Example – with added aesthetics
library(ggplot2)

icudata$vit_stat <- factor(icudata$vit_stat,            # factor( ) to make the graph readable
                    levels=c(1,0),                      # the values of 1 and 0
                    labels=c("Dead","Alive"))           # will now be labelled “Dead” and “Alive”
ggplot(data=icudata) +
    aes(x=age) +
    geom_histogram(binwidth=7, color="navy", fill="blue") +
    facet_grid(vit_stat ~ .) +
    xlab("Age(years)") +
    ylab("Number") +
    ggtitle("Distribution of Age, by Vital Status")
```

```
#4  Single Variable, Continuous – Stem & Leaf

# Basic command (user edits yellow)
stem(DATAFRAME$VARIABLENAME)

# Example –
stem(icudata$age)
```

```
#5.  Single Continuous Variable & One Grouping Variable – "Back to back" Side by Side Stem and Leaf

library(aplpack)                              # Be sure to have installed the package {aplpack}
# Basic command (user edits yellow)
stem.leaf.backback(DATAFRAME$VARIABLE1, DATAFRAME$VARIABLE2)


# Example –
# Preliminary – Extract the two distributions (group 1 = "dead", group 2 = "alive")
library(aplpack)
dead <- subset(icudata,vit_stat=="Dead",select=c(age))
alive <- subset(icudata,vit_stat=="Alive",select=c(age))
stem.leaf.backback(dead$age,alive$age)
```

```
#6.  Single Continuous Variable – Box & Whisker Plot

# Basic command (user edits yellow)
library(ggplot2)
ggplot(data=DATAFRAME) +                          # dataset name
      aes(y=CONTINUOUSVARIABLE) +                 # TIP:  Use y= to get boxplot on the vertical, x= for horizontal
      geom_boxplot(option, option) +              # geom_boxplot( ) to get boxplot
      additions as you like

# Example – with added aesthetics
library(ggplot2)
ggplot(data=icudata) +
    aes(y=age) +
    geom_boxplot(color="black", fill='blue') +
    ggtitle("ICU Data (n=25): Box and Whisker Plot of Age")
```

```
#7.  Single Continuous Variable & One Grouping Variable – Side by Side Box & Whisker Plot

# Basic command (user edits yellow)
library(ggplot2)
ggplot(data=DATAFRAME)  +
       aes(x=GROUPVARIABLE, y=CONTINUOUSVARIABLE, fill=GROUPVARIABLE) +
       geom_boxplot(option, option) +
       additions as you like


# Example – with added aesthetics
library(ggplot2)
ggplot(data=icudata) +
     aes(x=vit_stat, y=age, fill=vit_stat)) +
     geom_boxplot() +
     ggtitle("ICU Data (n=25): Box and Whisker Plot of Age, by Vital Status")
```