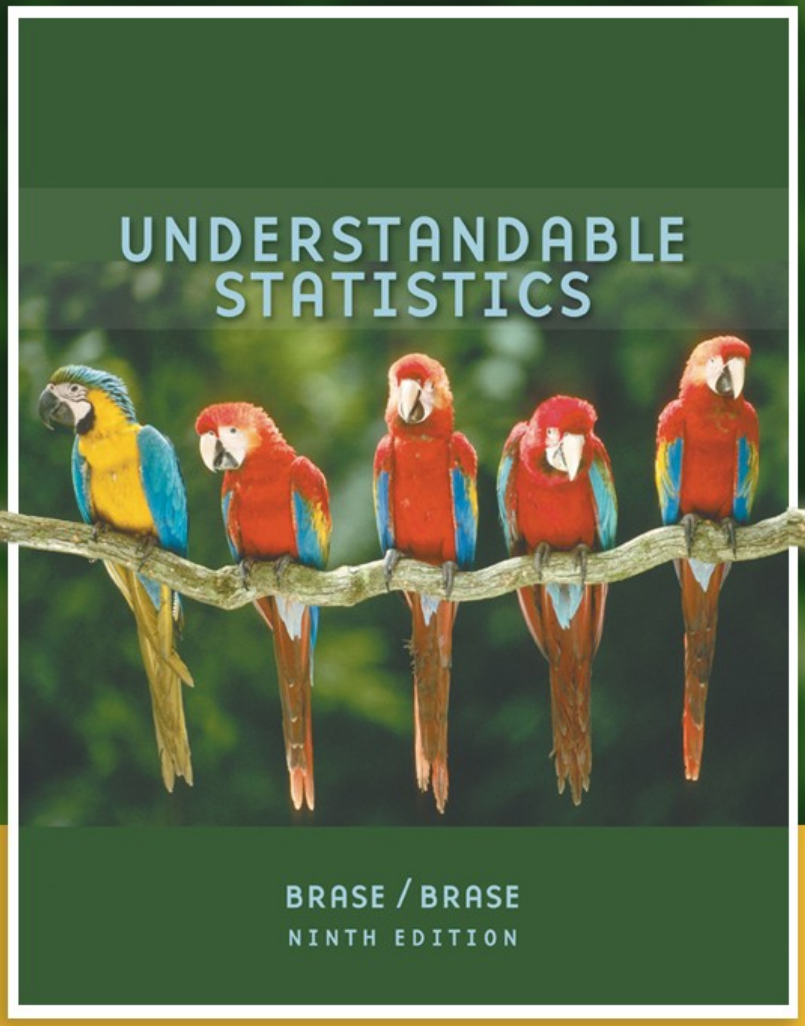


Chapter 2

Organizing Data

Understandable Statistics
Ninth Edition

By Brase and Brase
Prepared by Yixun Shi
Bloomsburg University of Pennsylvania

Frequency and Relative Frequency

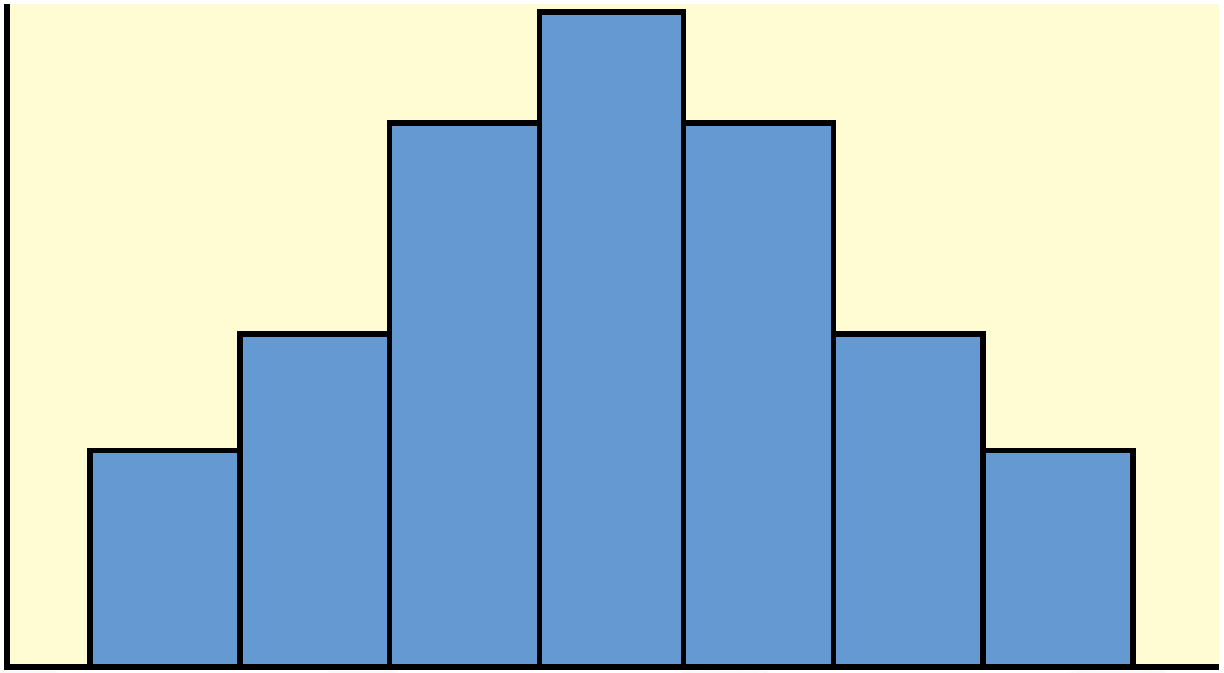
- Class
- Frequency
- Class width
- Lower class limit, upper class limit, and midpoint
- Tally data

Frequency and Relative Frequency

- Class boundaries for integer data
- Frequency table
- Relative frequency
- Relative frequency table
- Frequency histogram
- Relative frequency histogram

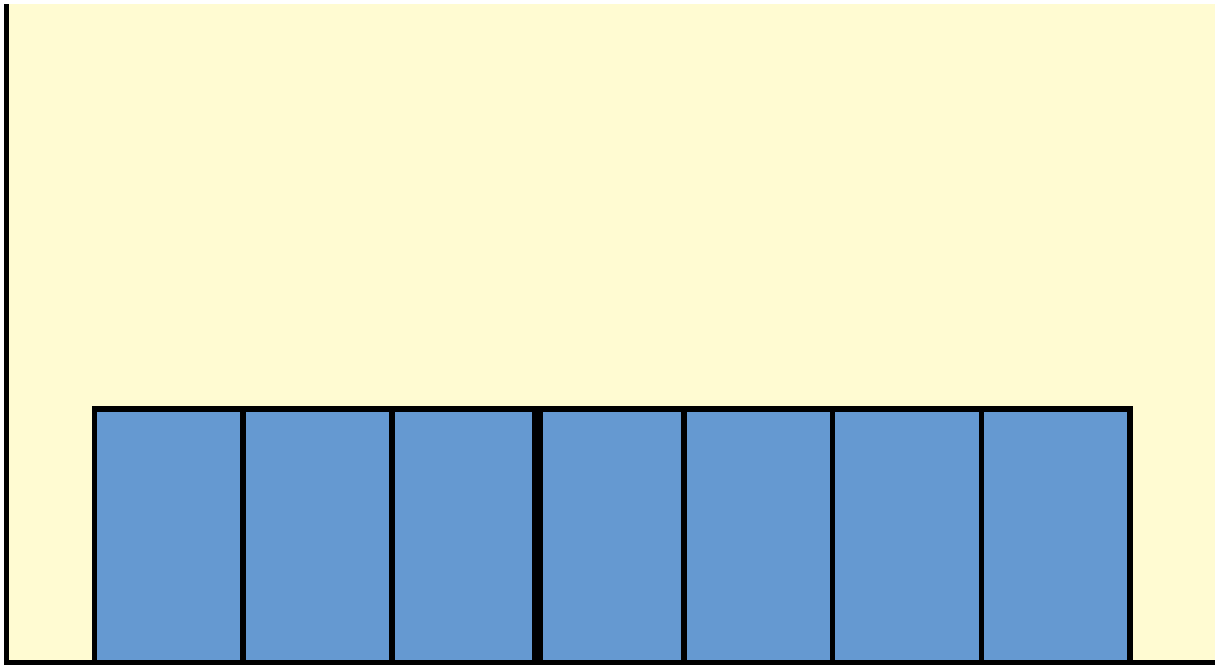
Distribution Shapes

- Symmetric – The distribution's shape is generally the same if folded down the middle.



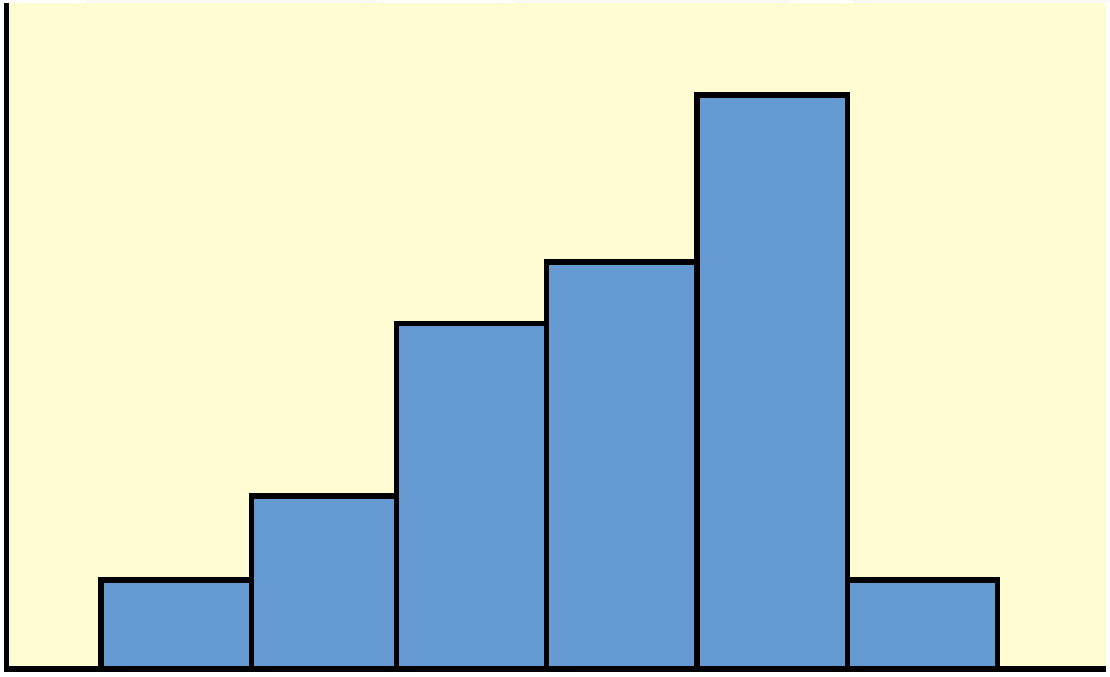
Distribution Shapes

- Uniform or rectangular (also symmetric)



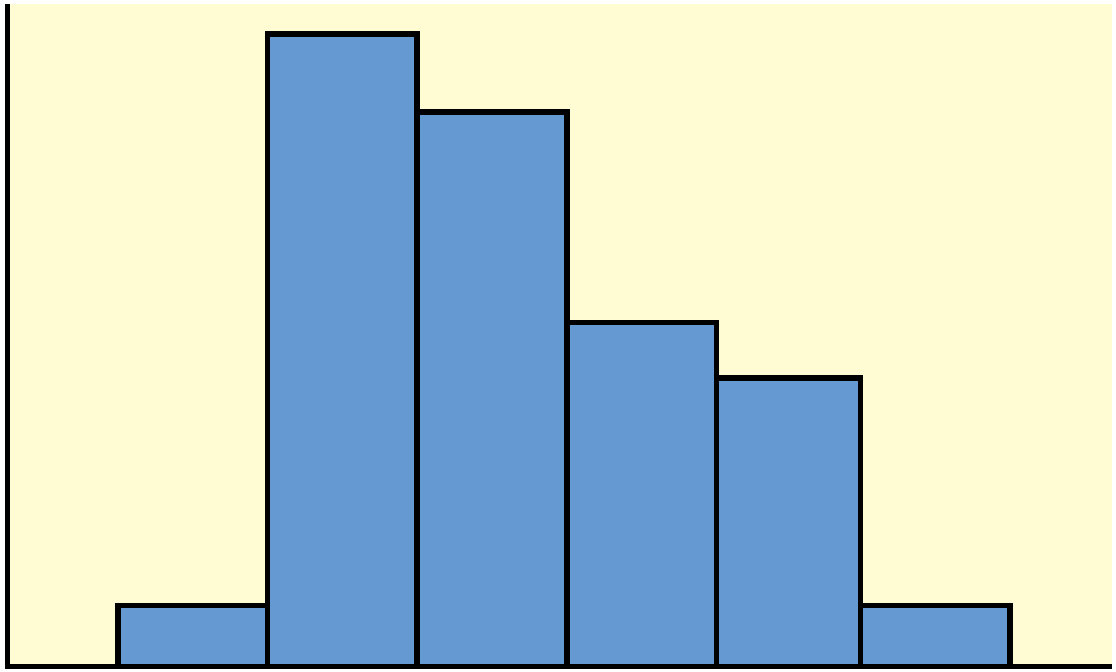
Distribution Shapes

- Skewed Left



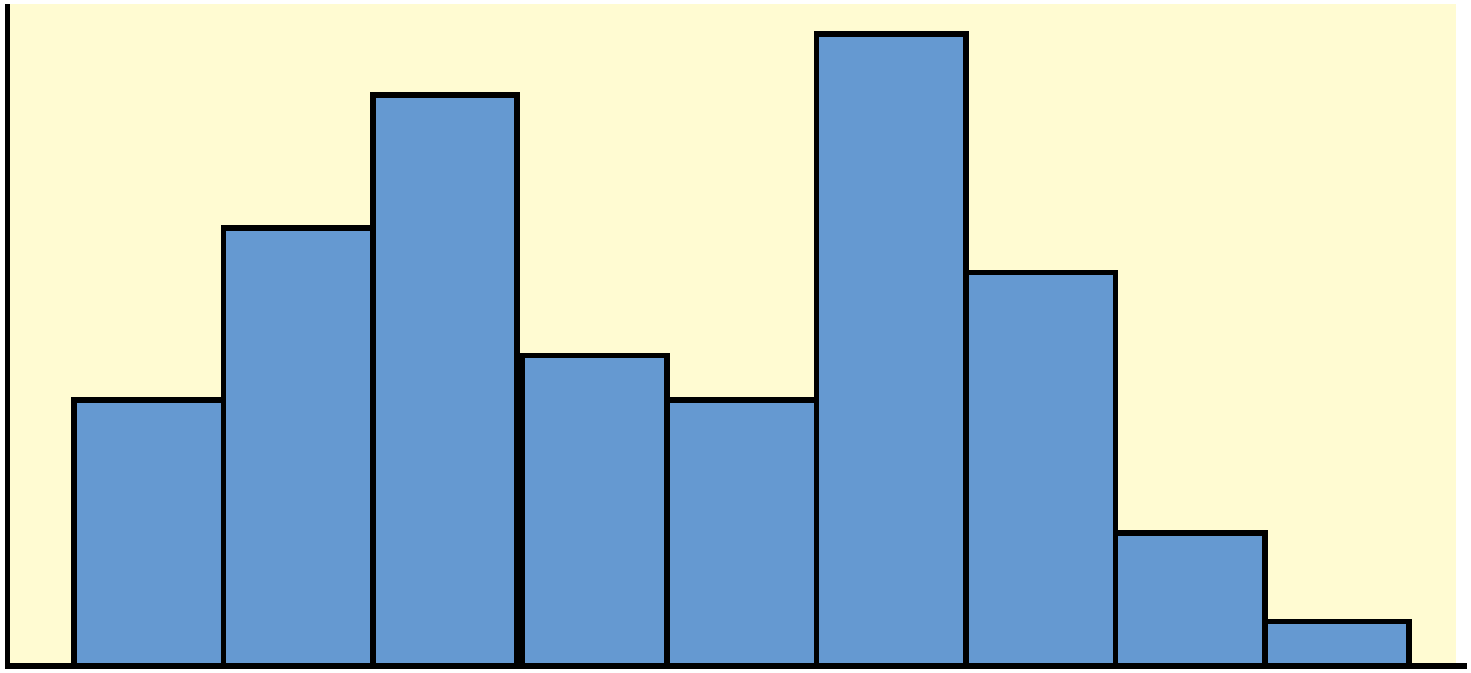
Distribution Shapes

- Skewed Right



Distribution Shapes

- Bimodal



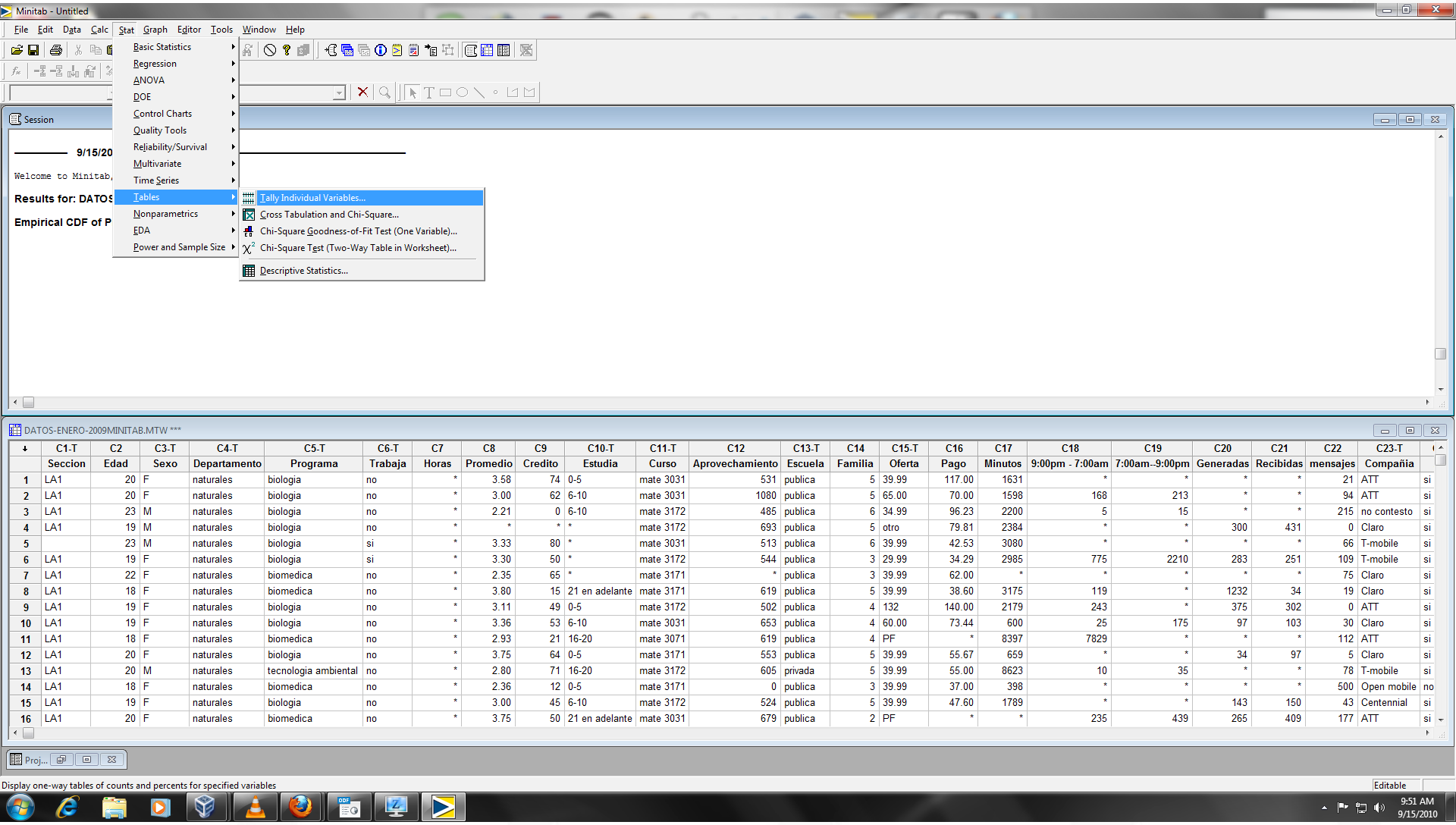
Critical Thinking

- A bimodal distribution shape might indicate that the data are from two different population.
- Outliers – data values that are very different from other values in the data set.
- Outliers may indicate data recording errors.

Histogram

Visit http://math.uprag.edu/clase-histograma.html.

- Remember numbere of clases square root of n.
- Histogram never has more than 20 clases and less than 5 clases.



	C1-T	C2	C3-T	C4-T	C5-T	C6-T	C7	C8	C9	C10-T	C11-T	C12	C13-T	C14	C15-T	C16	C17	C18	C19	C20	C21	C22	C23-T
	Seccion	Edad	Sexo	Departamento	Programa	Trabaja	Horas	Promedio	Credito	Estudia	Curso	Aprovechamiento	Escuela	Familia	Oferta	Pago	Minutos	9:00pm - 7:00am	7:00am--9:00pm	Generadas	Recibidas	mensajes	Compañia
1	LA1	20	F	naturales	biologia	no	*	3.58	74	0-5	mate 3031	531	publica	5	39.99	117.00	1631	*	*	*	*	21	ATT
2	LA1	20	F	naturales	biologia	no	*	3.00	62	6-10	mate 3031	1080	publica	5	65.00	70.00	1598	168	213	*	*	94	ATT
3	LA1	23	M	naturales	biologia	no	*	2.21	0	6-10	mate 3172	485	publica	6	34.99	96.23	2200	5	15	*	*	215	no contesto
4	LA1	19	M	naturales	biologia	no	*	*	*	*	mate 3172	693	publica	5	otro	79.81	2384	*	*	300	431	0	Claro
5		23	M	naturales	biologia	si	*	3.33	80	*	mate 3031	513	publica	6	39.99	42.53	3080	*	*	*	*	66	T-mobile
6	LA1	19	F	naturales	biologia	si	*	3.30	50	*	mate 3172	544	publica	3	29.99	34.29	2985	775	2210	283	251	109	T-mobile
7	LA1	22	F	naturales	biomedica	no	*	2.35	65	*	mate 3171	*	publica	3	39.99	62.00	*	*	*	*	*	75	Claro
8	LA1	18	F	naturales	biomedica	no	*	3.80	15	21 en adelante	mate 3171	619	publica	5	39.99	38.60	3175	119	*	1232	34	19	Claro
9	LA1	19	F	naturales	biologia	no	*	3.11	49	0-5	mate 3172	502	publica	4	132	140.00	2179	243	*	375	302	0	ATT
10	LA1	19	F	naturales	biologia	no	*	3.36	53	6-10	mate 3031	653	publica	4	60.00	73.44	600	25	175	97	103	30	Claro
11	LA1	18	F	naturales	biomedica	no	*	2.93	21	16-20	mate 3071	619	publica	4	PF	*	8397	7829	*	*	*	112	ATT
12	LA1	20	F	naturales	biologia	no	*	3.75	64	0-5	mate 3171	553	publica	5	39.99	55.67	659	*	*	34	97	5	Claro
13	LA1	20	M	naturales	tecnologia ambiental	no	*	2.80	71	16-20	mate 3172	605	privada	5	39.99	55.00	8623	10	35	*	*	78	T-mobile
14	LA1	18	F	naturales	biomedica	no	*	2.36	12	0-5	mate 3171	0	publica	3	39.99	37.00	398	*	*	*	*	500	Open mobile
15	LA1	19	F	naturales	biologia	no	*	3.00	45	6-10	mate 3172	524	publica	5	39.99	47.60	1789	*	*	143	150	43	Centennial
16	LA1	20	F	naturales	biomedica	no	*	3.75	50	21 en adelante	mate 3031	679	publica	2	PF	*	*	235	439	265	409	177	ATT

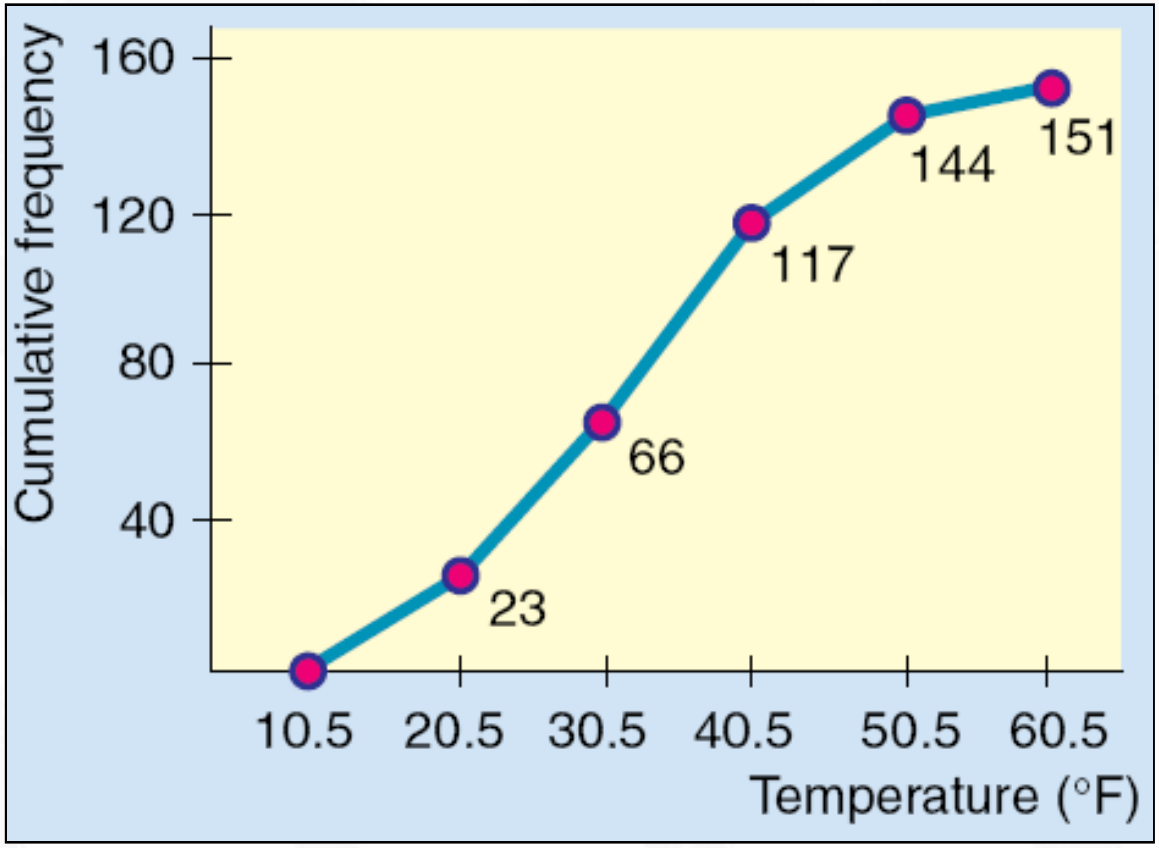
Cumulative Frequency Tables

- Cumulative frequencies for a class are the sums of all the frequencies up to and including that class.
- Example

Class Boundaries			
Lower	Upper	Frequency	Cumulative Frequency
10.5	20.5	23	23
20.5	30.5	43	66 (sum 23 + 43)
30.5	40.5	51	117 (sum 66 + 51)
40.5	50.5	27	144 (sum 117 + 27)
50.5	60.5	7	151 (sum 144 + 7)

Ogives

- Graph that displays cumulative frequencies



Exploratory Data Analysis

- EDA is the process of learning about a data set by creating graphs.
- EDA specifically looks for patterns and trends in the data.
- EDA also identifies extreme values.

Graphical Displays

- Represent the data
- Induce the viewer to think about the substance of the graphic
- Avoid distorting the message of the data

Bar Graphs

- Used for qualitative or quantitative data
- Can be vertical or horizontal
- Bars are uniformly spaced and have equal widths.
- Length/height of bars indicate counts or percentages of the variable.
- Including titles and units and labeling axes are good practices.

Pareto Charts

- A bar chart with two specific features:
  - Heights of bars represent frequencies.
  - Bars are vertical and are ordered from tallest to shortest.

Circle Graphs/Pie Charts

- Used for qualitative data
- Wedges of the circle represent proportions of the data that share a common characteristic.
- Including a title and legend is a good practice.

Time-Series

- Time-Series Data – Measurements of the same variable for the same individual over regular intervals of time.
- Time-Series Graphs

Critical Thinking – which type of graph to use?

- Bar graphs are useful for quantitative or qualitative data.
- Pareto charts identify the frequency in decreasing order.
- Circle graphs display how a *total* is dispersed into several categories.
- Time-series graphs display how data change over time

Stem and Leaf Plots

- Displays the distribution of the data while maintaining the actual data values.
- Each data value is split into a stem and a leaf.

Stem and Leaf Plot Construction

PROCEDURE

How to make a stem-and-leaf display

1. Divide the digits of each data value into two parts. The leftmost part is called the *stem* while the rightmost part is called the *leaf*.
2. Align all the stems in a vertical column from smallest to largest. Draw a vertical line to the right of all the stems.
3. Place all the leaves with the same stem on the same row as the stem, and arrange the leaves in increasing order.
4. Use a label to indicate the magnitude of the numbers in the display. We include the decimal position in the label rather than with the stems or leaves.

Critical Thinking

- By looking at the stem-and-leaf display “sideways”, we can see the distribution shape of the data.
- Large gaps between stems containing leaves, especially at the top or bottom, suggest the existence of outliers.
- Watch the outliers – are they data errors or simply unusual data values?