2.2: Frequency Distributions and Their Graphs

Frequency distributions: A *frequency distribution* is a way of summarizing quantitative data by grouping it into *classes*. For each class, we write the frequency, the number of data points that fall into that class. Usually the class is a range of data values, rather than just a single data value.

Example 1: This is a frequency distribution of family income data.
Source: U.S. Census Bureau, 2013-2017 American Community Survey 5-Year Estimates.
https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_17_5YR_DP03&src=pt

Families	78,298,703	+/-221,073	78,298,703
Less than $10,000	3,293,360	+/-12,758	4.2%
$10,000 to $14,999	2,193,054	+/-10,839	2.8%
$15,000 to $24,999	5,545,106	+/-21,850	7.1%
$25,000 to $34,999	6,402,558	+/-20,269	8.2%
$35,000 to $49,999	9,576,021	+/-22,542	12.2%
$50,000 to $74,999	14,362,139	+/-43,806	18.3%
$75,000 to $99,999	11,095,863	+/-51,635	14.2%
$100,000 to $149,999	13,551,675	+/-86,860	17.3%
$150,000 to $199,999	5,872,950	+/-40,109	7.5%
$200,000 or more	6,405,977	+/-43,463	8.2%
Median family income (dollars)	70,850	+/-215	(X)
Mean family income (dollars)	95,031	+/-228	(X)

The *relative frequency* for a class is the proportion of the data that fall into that class. Relative frequencies are often written as percentages.

$$\text{Relative frequency} = \frac{\text{Frequency}}{\text{Sum of all frequencies}}$$

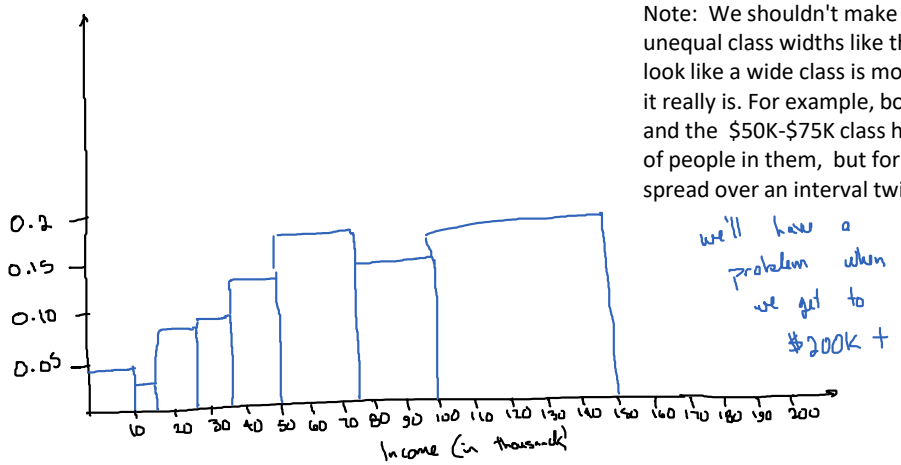
Note: The last class in Example 1, incomes of $200,000 or more, is called an *open-ended class*.

Histograms:

A histogram is a specific type of bar graph used to represent quantitative data. The values for the variable are on the horizontal axis, and the frequencies or relative frequencies are on the vertical axis. All values within the range of the variable are represented on the *x*-axis.

In a histogram (bar chart for grouped quantitative data), the bars must touch.

Example 2: Create a histogram for family income, based on the data in Example 1.



Note: We shouldn't make a histogram with unequal class widths like this, because it makes it look like a wide class is more common than what it really is. For example, both the $100K-$150K class and the $50K-$75K class have about the same number of people in them, but for $100K-$150K, they are spread over an interval twice as wide.

Creating a frequency distribution:

To create a frequency distribution, you need to decide the number of classes and the class width.

- Usually (but not always!) the classes will have equal width.
- The classes cannot overlap.
- Every data point (observation) must fall into one of the classes.
- There cannot be gaps between the classes.

(classes are often called "bins")

The *lower class limit of a class* is the smallest value that can appear in the class.
The *upper class limit* of a class is the largest value that can appear in the class.
The *class width* is the difference between consecutive lower class limits.

If we know how many classes we want, we can divide the range of data by the desired number of classes.

$$\text{Class width} \approx \frac{\text{Largest data value} - \text{Smallest data value}}{\text{Number of classes}}$$

The *midpoint* of a class is the value halfway between the lower limit of a class and the lower limit of the next class.

$$\text{Class midpoint} = \frac{\text{Lower limit of class} + \text{Lower limit of next class}}{2}$$

A *frequency polygon*, or a *relative frequency polygon*, is created by drawing line segments between the midpoints of the classes.

Example 1: Create a frequency distribution, relative frequency distribution, frequency histogram, and relative frequency histogram for the current Astros' batting average.

2019 | All-Time By Year | All-Time Totals | Regular Season | All Time | Active | All Players | Qualifiers

MLB | AL | NL | Houston Astros | All Positions | Select Split

Timeframe: YTD | Yesterday | Last 7 | Last 30 | Pre All-Star | Post All-Star

Next Stats

RK	Player	Team	Pos	G	AB	R	H	2B	3B	HR	RBI	BB	SO	SB	CS	AVG	OBP	SLG	OPS
1	Gurriel, Y	HOU	1B	123	487	73	150	36	2	26	93	30	55	4	3	.308	.348	.550	.898
2	Brantley, M	HOU	LF	123	483	81	163	39	2	19	80	44	54	3	2	.337	.397	.545	.942
3	Bregman, A	HOU	3B	127	458	101	131	28	2	32	90	94	71	4	1	.286	.410	.566	.975
4	Reddick, J	HOU	RF	118	430	47	114	16	3	10	41	28	53	4	2	.265	.305	.386	.691
5	Springer, G	HOU	RF	98	394	75	116	18	3	27	69	53	91	5	2	.294	.383	.561	.944
6	Altuve, J	HOU	2B	97	390	75	118	22	2	24	60	36	64	5	3	.303	.363	.554	.917
7	Chirinos, R	HOU	C	98	317	49	75	18	0	15	47	43	112	0	2	.237	.343	.435	.778
8	Correa, C	HOU	SS	72	270	39	75	16	1	19	56	34	70	1	0	.278	.358	.556	.914
9	Marisnick, J	HOU	CF	99	246	42	60	16	2	9	31	14	80	10	3	.244	.301	.435	.736
10	Alvarez, Y	HOU	1B	60	219	40	72	19	0	21	62	35	61	0	0	.329	.420	.703	1.123
11	White, T	HOU	1B	71	218	16	49	14	0	3	21	32	74	0	0	.225	.320	.330	.650
12	Kemp, T	HOU	CF	66	163	23	37	6	2	7	17	16	29	4	3	.227	.308	.417	.725
13	Diaz, A	HOU	2B	46	137	25	37	5	1	7	29	11	18	0	0	.270	.323	.474	.797
14	Stassi, M	HOU	C	31	90	4	15	1	0	1	3	7	34	0	0	.167	.235	.211	.446
15	Straw, M	HOU	CF	43	86	19	20	3	1	0	4	16	22	7	1	.233	.353	.291	.644
16	Fisher, D	HOU	RF	17	53	9	12	2	1	1	5	7	14	4	1	.226	.317	.358	.675
16	Mayfield, J	HOU	2B	19	53	6	7	5	0	1	3	1	13	0	0	.132	.148	.283	.431
18	Maldonado, M	HOU	C	11	37	9	6	1	0	3	4	6	13	0	0	.162	.279	.432	.712
19	Stubbs, G	HOU	C	9	25	2	4	1	0	0	2	2	5	1	0	.160	.222	.200	.422
20	Toro, A	HOU	3B	5	19	3	4	0	0	0	1	2	3	0	1	.211	.286	.211	.496

Min Batting Average 0.132

Max Batting Avg 0.337

Total width: 0.337 − 0.132 = 0.205

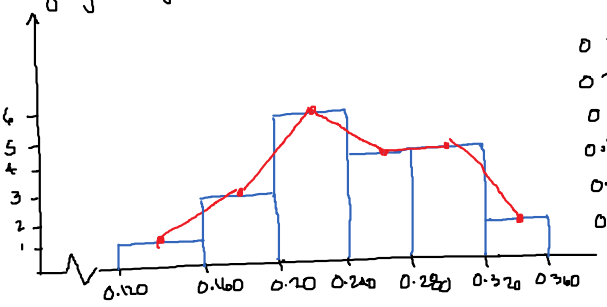
Let's do 11 classes

$\frac{0.200}{10} \Rightarrow 0.02$ class width

Average: 0.120, 0.140, 0.160, 0.180, 0.200, 0.220, 0.240, 0.260, 0.280, 0.300

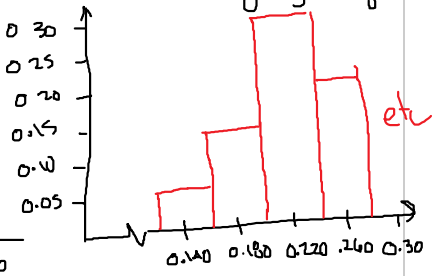
Midpoint	Class	Frequency	Relative Frequency
0.140	0.120 − 0.159	1	1/20 = 0.05
0.180	0.160 − 0.199	3	3/20 = 0.15
0.220	0.200 − 0.239	6	6/20 = 0.30 ← modal class
0.260	0.240 − 0.279	4	4/20 = 0.20
0.300	0.280 − 0.319	4	4/20 = 0.20
0.320	0.320 − 0.359	2	2/20 = 0.10
		n = 20	Total = 1.0 ⇒ 100%

Frequency Histogram



Could use midpoints to label

Relative Frequency Histogram



Histograms for discrete data:

When constructing a histogram for discrete (countable) data, the classes can either be a single value or a range of values.

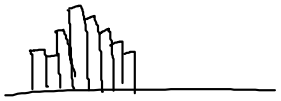
Example 3: Create a histogram representing the number of siblings in the families of everyone in our class.

Batting Average: continuous data. Can (theoretically) take on any value between 0 and 1.

Number of Siblings: Discrete data (countable)

Number of Siblings	Frequency
1	
2	
3	
4	
5	
6	

Shapes of histograms:



Symmetric

Skewed Left

Skewed Right

Unimodal

Bimodal

Uniform

The *mode* is the most frequent data value.