Topic 4: Examining Numerical Data

Histograms

Histogram: a graphical method for analyzing the distribution of 1 numerical variable

IQ test scores for 60 randomly chosen fifth-grade students

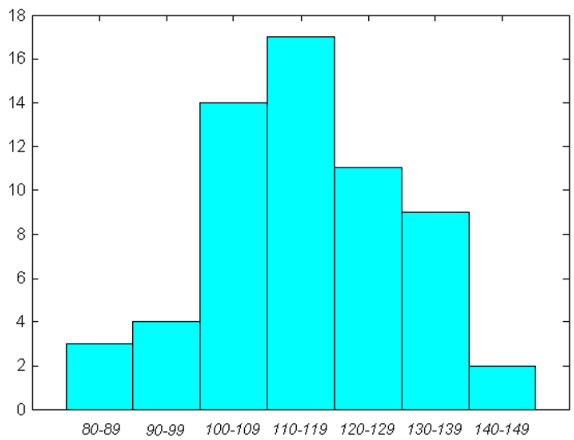
145	139	126	122	125	130	96	110	118	118
101	142	134	124	112	109	134	113	81	113
123	94	100	136	109	131	117	110	127	124
106	124	115	133	116	102	127	117	109	137
117	90	103	114	139	101	122	105	97	89
102	108	110	128	114	112	114	102	82	101

We can use a **histogram** to visually inspect the *distribution* of these IQ scores.

"Bins" = "classes" / "ranges"

Bin	Frequency
80 – 89	3
90 – 99	4
100 – 109	14
110 – 119	17
120 – 129	11
130 – 139	9
140 - 149	2

Frequency table: a sorting of data values into bins ("classes") of even width such that each value goes in exactly 1 bin

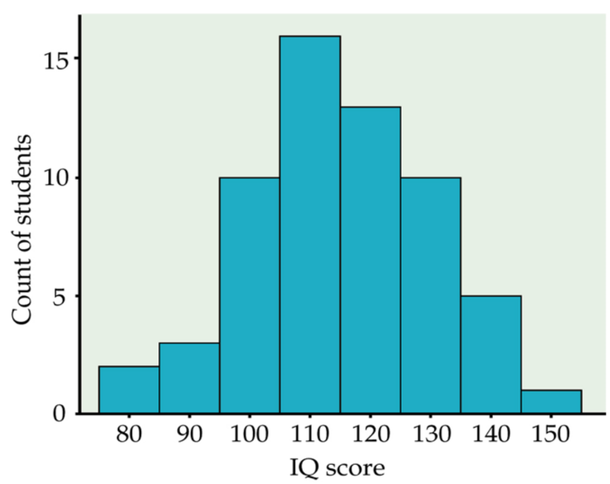


This is a histogram!

Bin etiquette: equal width/range, no gaps between their specified ranges, no overlap

Slightly different choice of bins:

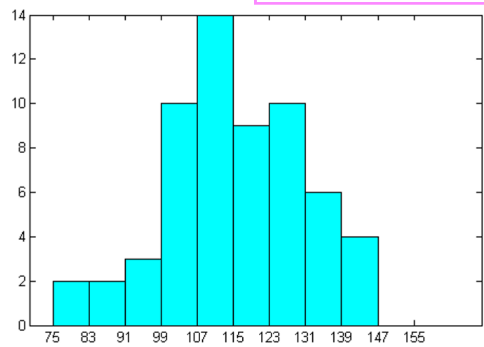
Bin	Frequency
75-84	2
85-94	3
95-104	10
105-114	16
115-124	13
125-134	10
135-144	5
145-154	1

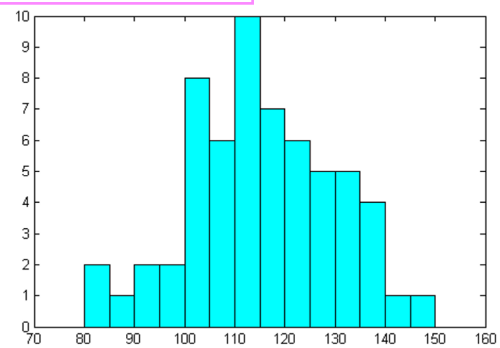


A single dataset has many valid bin construction options, as long as you obey the rules above!

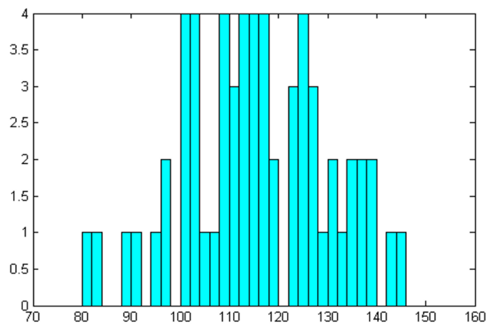
Other choices of bins:

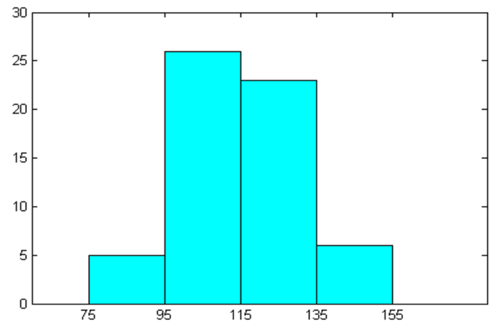
These top two both look like good bin choices with regards to width.





These bins are too narrow - the many gaps between bins and big jumps between bin heights obscure the distribution.





These bins are too wide - there's so few bins that we can't see the shape formed by the data very well.

Using relative frequencies instead:

Rel. freq. = freq. / n

n = total number of data values

Item Price	Frequency	Relative Frequency
$1 - $10	20	0.303
$11 - $20	21	0.318
$21 - $30	13	0.197
$31 - $40	8	0.121
$41 - $50	4	0.061

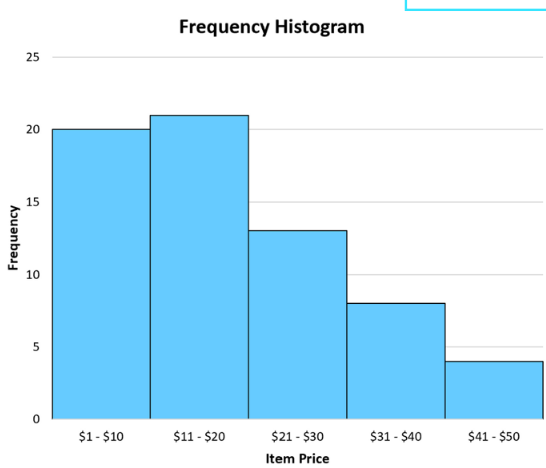
The relative frequencies will always all add up to 1.

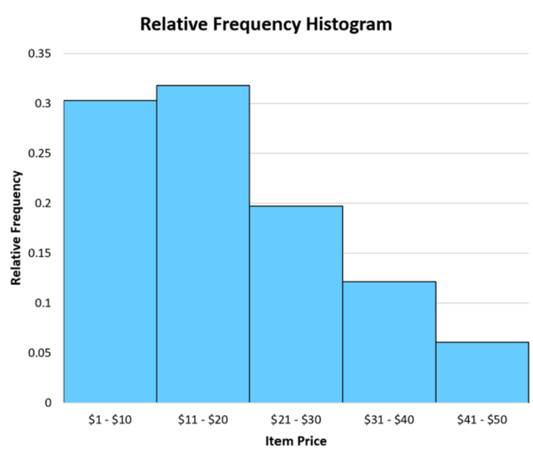
You can always find n by adding up all frequencies!

Relative frequency of "$41-$50":

freq / n = 4/66 = 0.061 (approx.)

n = 20+21+13+8+4 = 66





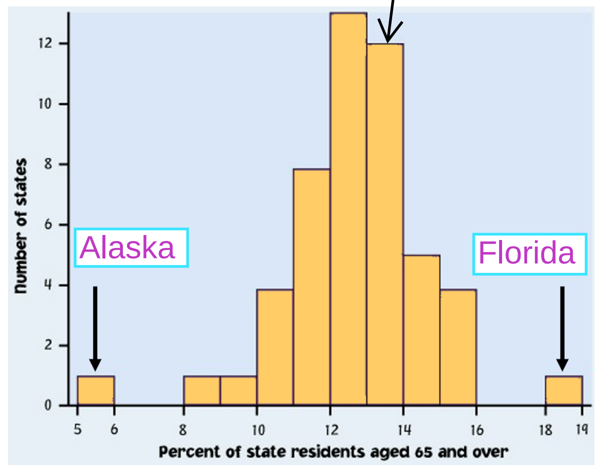
Histograms have the SAME SHAPE regardless of frequency vs relative frequency.

Using relative frequency allows you to COMPARE histograms of multiple datasets with DIFFERENT amounts of data (n values)

Using a histogram to identify **outliers**:

Outlier: a data value that is far away from the rest of the data

"12 states have 13%-14% of residents aged 65 and over"



Look for bars of height 1 with empty space between them and the rest of the (otherwise-well-structured) histogram

Describing Distributions

- Shape

  Peaks: unimodal (1), bimodal (2), multimodal (3+), or uniform(0)?

  Layout: symmetric or skew?

  WATCH OUT!
  "Right-skew" means the TAIL is on the RIGHT, and the peak is on the left.
  "Left-skew" means the TAIL is on the LEFT, and the peak is on the right.

- Center

  Mean - average
  Median - middle data value when sorted smallest to largest
  Mode - most frequent (can be local)

  Symmetric: mean ≈ median
  Right-skew: mean > median
  Left-skew: mean < median

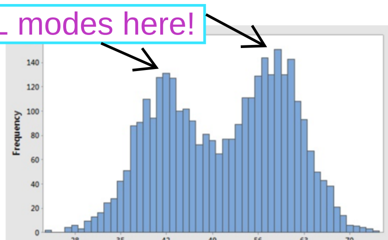
- Spread

  Variance and/or standard deviation
  Inter-Quartile Range (IQR)

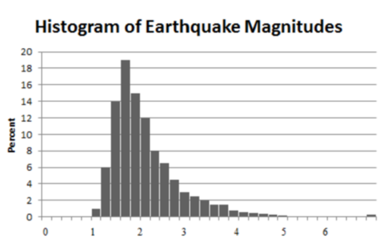
- Outliers

  Are there outliers?

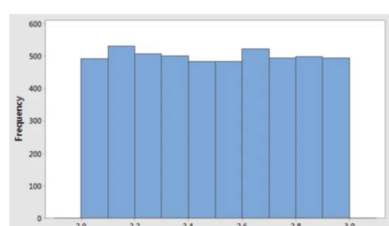
2 LOCAL modes here!



Bimodal
Symmetric (roughly)
Mean ≈ median, around 49-50
1 minor outlier to left



Unimodal
Skew: right-skew
Mean > median
1 outlier far to right



Uniform
Symmetric
Mean ≈ median, around 2.5
No outliers

Example: A child's birthday party has 9 attendees of the following ages: 7, 1, 3, 4, 4, 6, 3, 5, 3

- Notation

Observations: label as $x_1, x_2, x_3, \ldots x_n$
n = total # of data points

Here: n=9, and $x_1 = 7$, $x_2 = 1$, etc.

Refer to a specific data point as "x_i" where "i" is an integer in the set {1, 2, ... n}

- Measures of center

Mean: average of all data points

Summation notation!
"The sum from i=1 to i=n of all values x_i"

Notation: $\bar{x}$

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\bar{x} = \frac{1}{9}(7+1+3+4+4+6+3+5+3)$$

$$= \boxed{4 \text{ years old}}$$

Median: "middle data value", according to sorted data smallest to largest

1, 3, 3, 3, **4,** 4, 5, 6, 7

Here: since there are 9 values, the median is equal to the 5th one when ordered smallest to largest

m = 4 years old

Mode: "most frequent data value"

Here: the value 3 occurs the most in the data (three times)

How does adding a 64-year old to the group change mean and median?

$$\bar{x} = \frac{1}{10}(7+1+3+4+4+6+3+5+3+64)$$

$$= 10 \text{ years old}$$

1, 3, 3, 3, **4, 4,** 5, 6, 7, 64

For n even, have TWO middle values... Just average them

$$\text{Median} = \frac{4+4}{2} = 4 \text{ years old}$$

Effect of outliers on mean and median:

Mean is **sensitive** to outliers; outliers have a large effect on the mean.

Median is **robust** to outliers; outliers do NOT have a large effect on the median.

- Median as a percentile

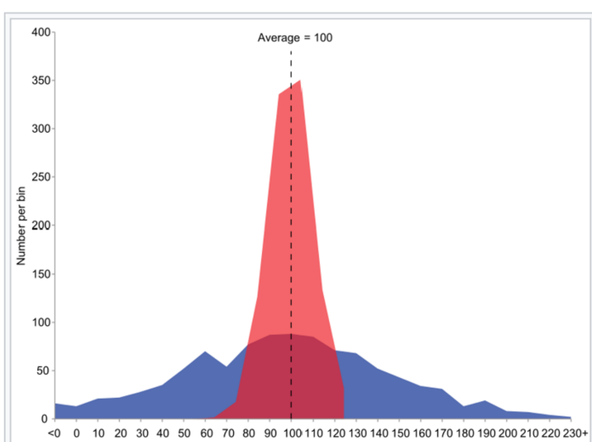
Other concept for median: median is **the 50th percentile**

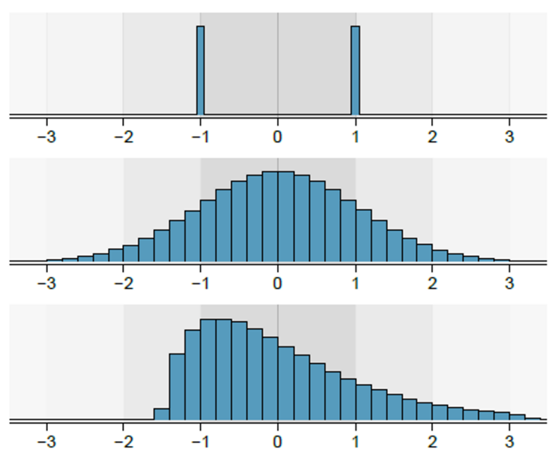
x^{th} percentile: the value greater than or equal to (≥) x% of the data values

Median is **the value ≥ 50%** of the data values (which also means it is less than 50% of the data values)

Same example: Birthday party attendees aged 7, 1, 3, 4, 4, 6, 3, 5, 3

- Measures of spread: variance and standard deviation





Same example: Birthday party attendees aged 7, 1, 3, 4, 4, 6, 3, 5, 3

- Another measure of spread: IQR

- IQR criterion for outliers:

- 5-number summary and box plot:

Data analysis in Excel: some easy commands to try on sheet `unc2017.xlsx` (posted on Canvas)!