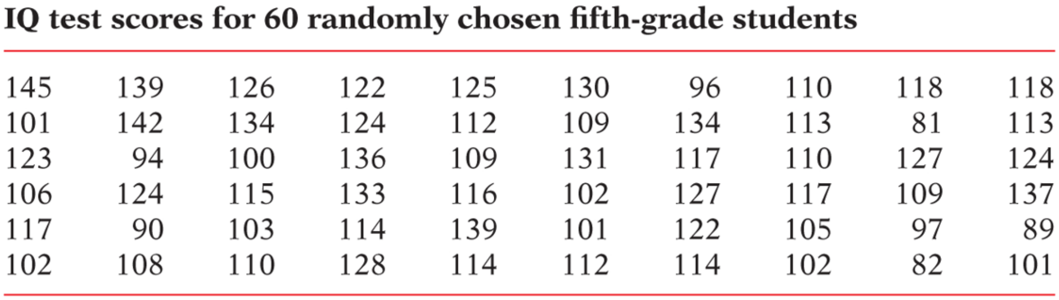
Lecture 5: Examining Numerical Data

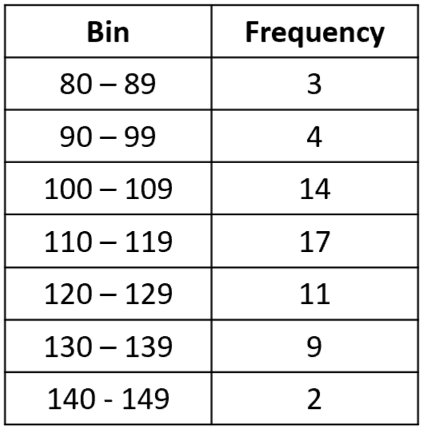
Histograms



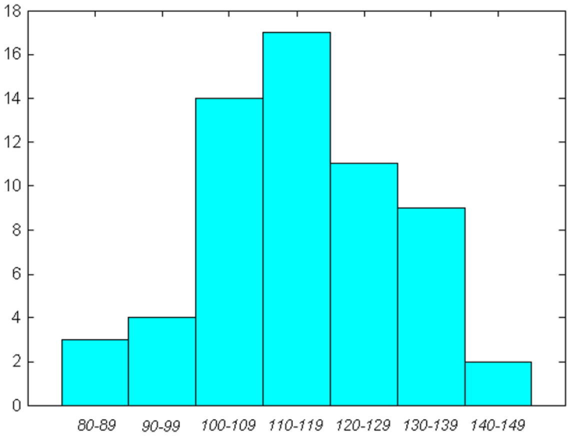
IQ test scores for 60 randomly chosen fifth-grade students

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

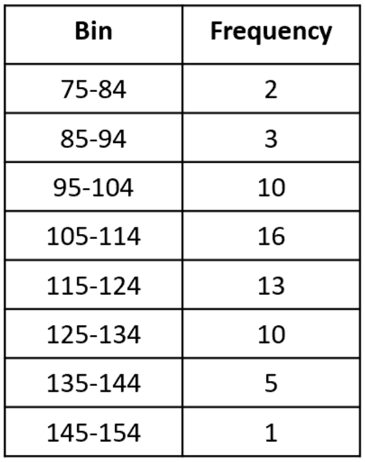
We can use a **histogram** to visually inspect the *distribution* of these IQ scores.



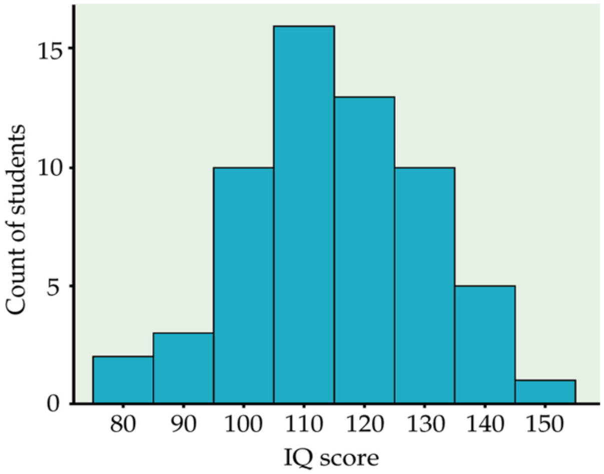
Bin	Frequency
80 – 89	3
90 – 99	4
100 – 109	14
110 – 119	17
120 – 129	11
130 – 139	9
140 - 149	2



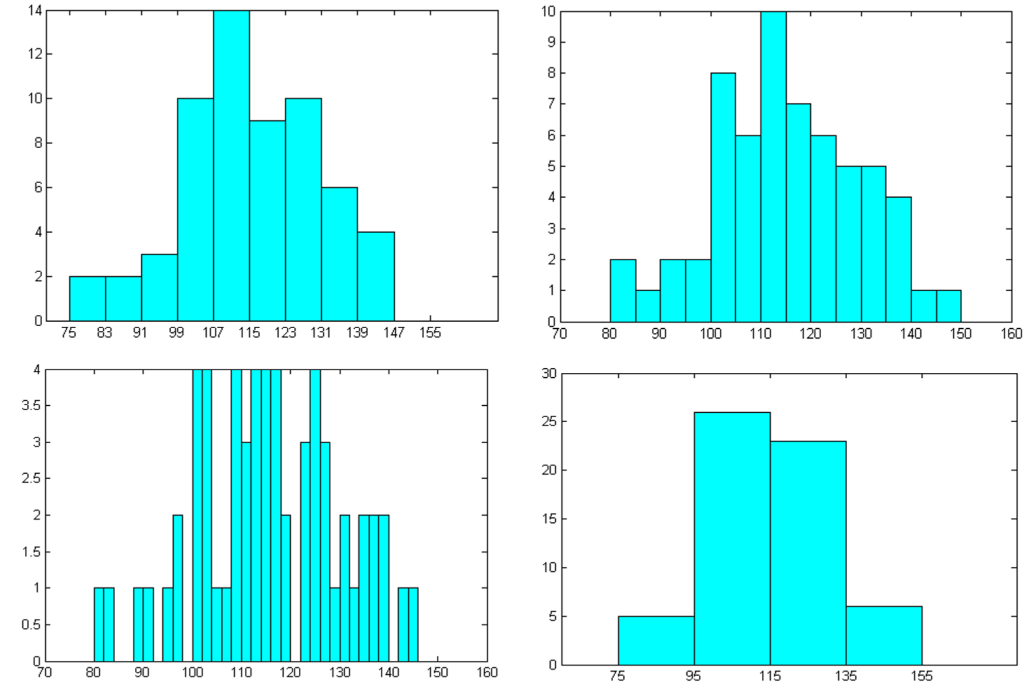
Slightly different choice of bins:



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

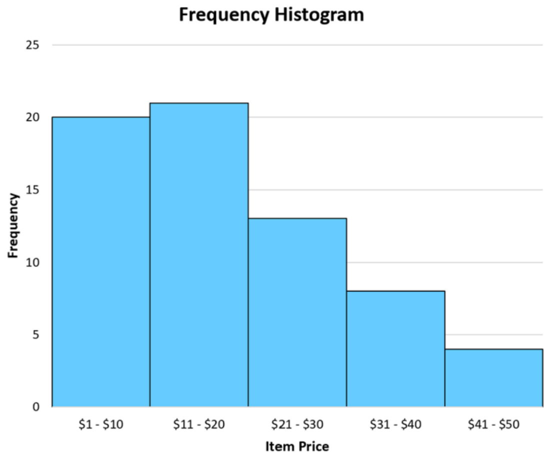


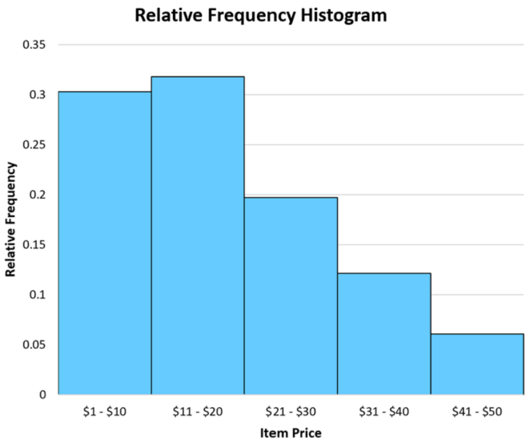
Other choices of bins:



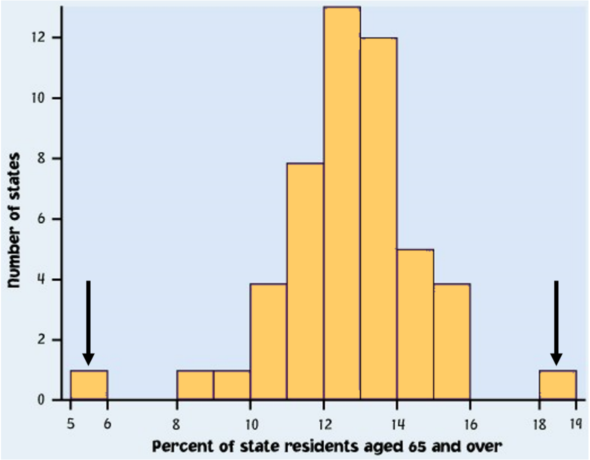
Using relative frequencies instead:

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





Using a histogram to identify **outliers**:



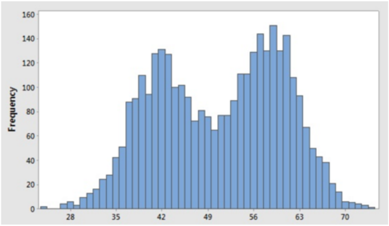
Describing Distributions

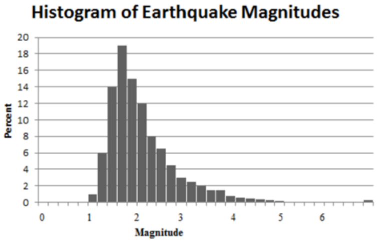
- Shape

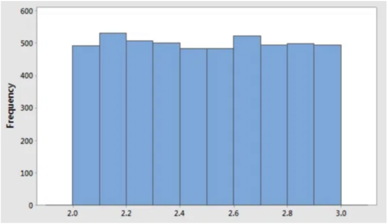
- Center

- Spread

- Outliers







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

- Notation

- Measures of center

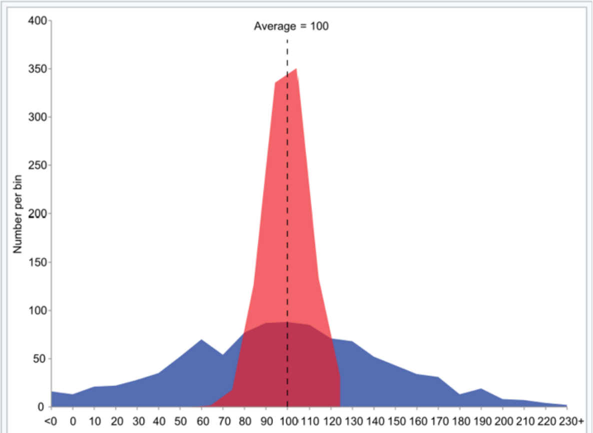
  How does adding a 64-year old to the group change mean and median?

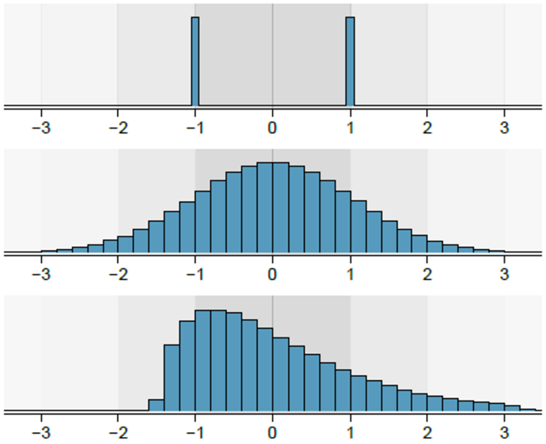
  Effect of outliers on mean and median:

- Median as a percentile

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: open sheet `unc2017.xlsx`