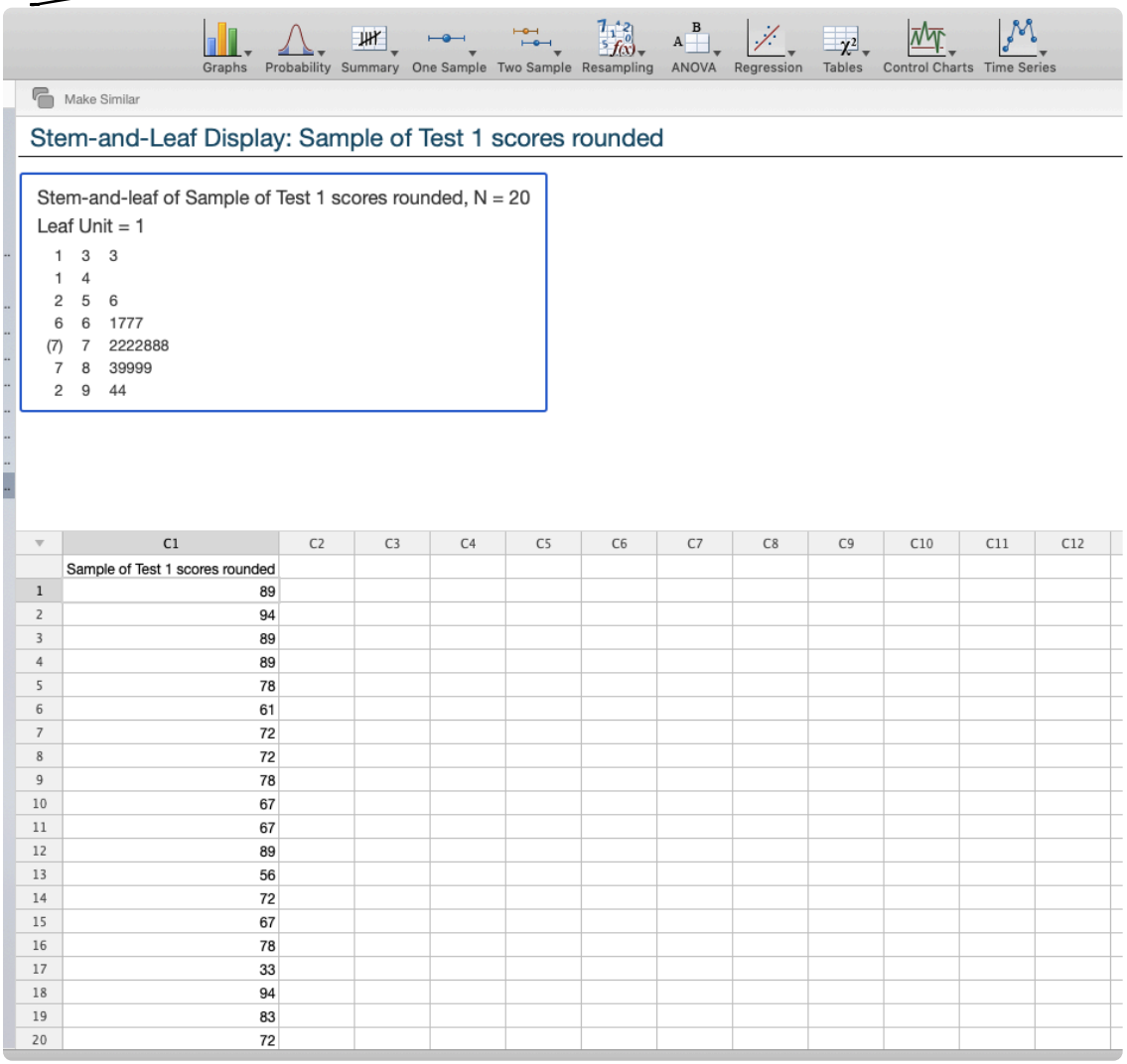
Lecture 17.

Recall that given a set of data points $\{x_1, \ldots, x_n\}$ where each x_i has at least two digits, we can construct a stem and leaf plot to visualize the data.

Ex: (from last time)



	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
	Sample of Test 1 scores rounded											
1	89											
2	94											
3	89											
4	89											
5	78											
6	61											
7	72											
8	72											
9	78											
10	67											
11	67											
12	89											
13	56											
14	72											
15	67											
16	78											
17	33											
18	94											
19	83											
20	72											

Some useful information we can just read off of a stem and leaf diagram is the quartiles and percentiles:

- 1st quartile is the number, q_1, such that 25% of the data is (less than or equal to) than q_1.
- 2nd quartile, q_2, is the number such that 50% of the data is less than or equal to q_2. (q_2 is also called MEDIAN!).
- 3rd quartile, q_3, is the number such that 75% of the data is less or equal to q_3.

More generally, the nth percentile is the number such that n% of the data lies below.

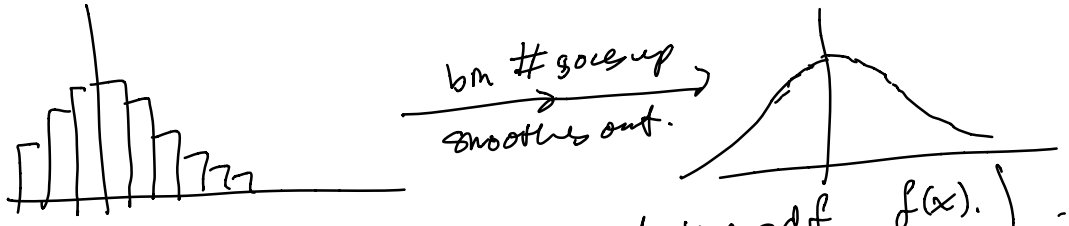
Defn The interquartile range is

$$IQR = q_3 - q_1.$$

It is usually a better measure of spread than range.

Frequency Distributions and Histograms

- Frequency distributions give a compact way of visualizing data.
- divide data into bins / class intervals / cells of equal width.
- how many bins? → too many, loose shape
  → too few, lose detail.



(- as # of bins $\to \infty$, get the pdf $f(x)$.)

General rule: for n data points, take $\sqrt{n}$ bins.

	Industrial Building Permits Issued in Hamilton by Year		
	YEAR	PERMITS ISSUED	
1	1998	86	
2	1999	90	
3	2000	73	* min
4	2001	170	
5	2002	128	
6	2003	140	
7	2004	112	
8	2005	122	
9	2006	188	
10	2007	158	
11	2008	142	
12	2009	172	
13	2010	157	
14	2011	213	* max
15	2012	146	
16	2013	178	
17	2014	183	
18	2015	183	
19	2016	172	
20	2017	193	

- 20 data points.
- $\sqrt{20} \approx 4.47 \rightarrow$ 5 bins.

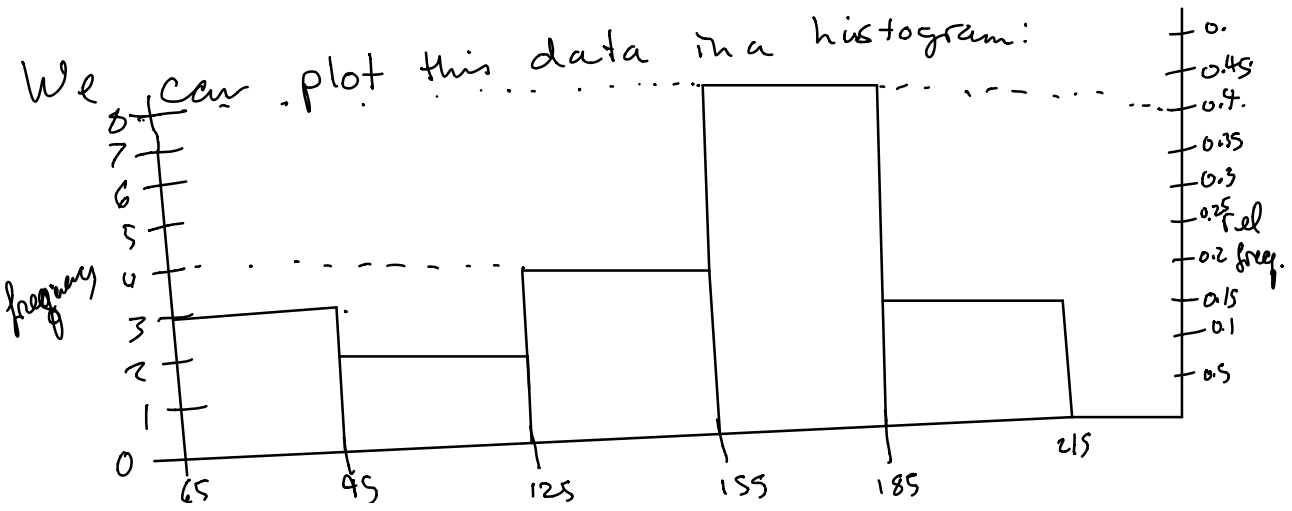
Range of data $= 213 - 73$
$= 140.$

$140/5 = 28 \sim 30$
bin width.

A frequency distribution for this data is:

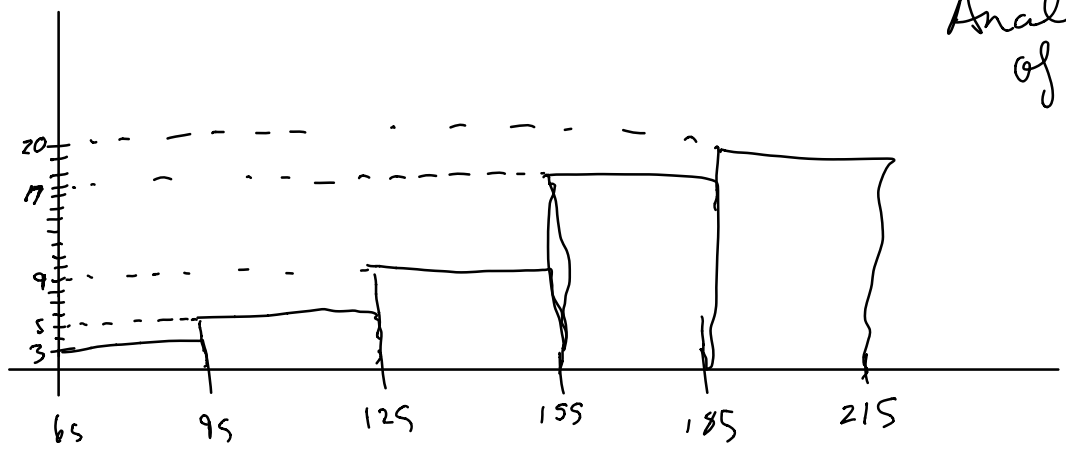
Bins	$65 \le x < 95$	$95 \le x < 125$	$125 \le x < 155$	$155 \le x < 185$	$185 \le x \le 215$
Frequency f	3	2	4	8	3
Relative Frequency $= f/20$	$3/20 = 0.15$	$\frac{2}{20} = 0.1$	$\frac{4}{20} = 0.2$	$\frac{8}{20} = 0.4$	$\frac{3}{20} = 0.15$
Cummulative Frequency	3	5	9	17	20

We can plot this data in a histogram:



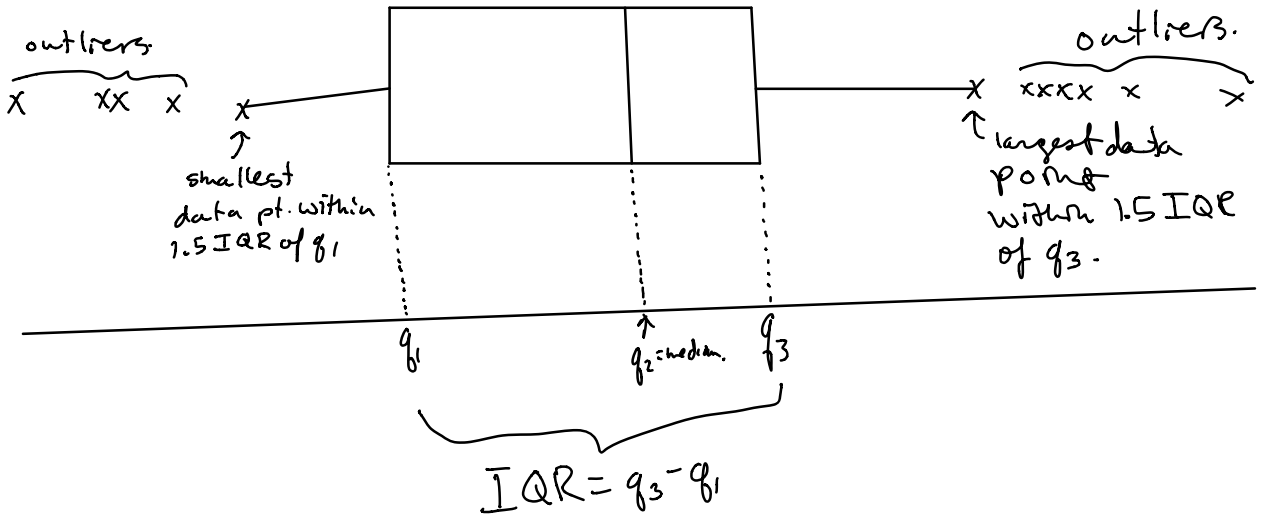
Can also plot cumulative frequency:

Analogue of cdf.



Box plots (Box-and-whisker plots).

- Combines many pieces of data into a single diagram:



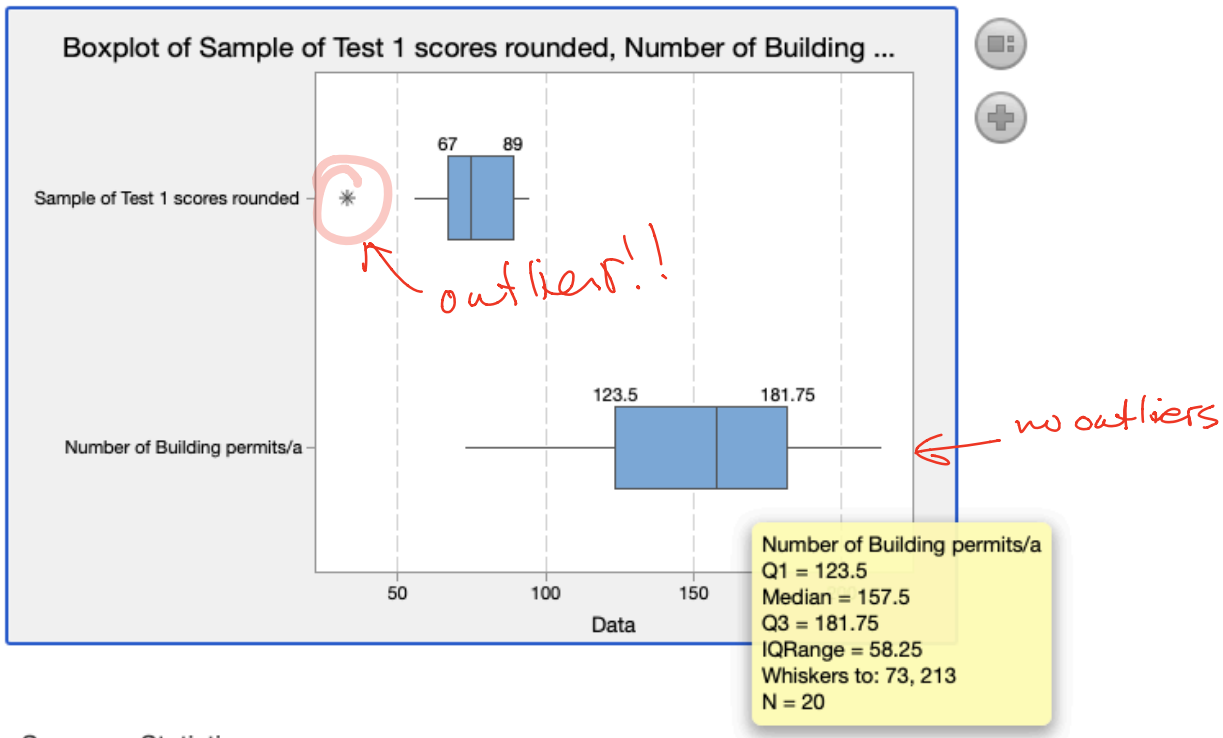
An outlier is anything greater than 1.5 IQR away from the box ends.

An extreme outlier is anything 3 IQR from box ends.

Can compare data sets with side by side box plots.

Ex:

Boxplot of Sample of Test 1 scores rounded, Number of Building permits/a



Summary Statistics

Variable	N	Minimum	Q1	Median	Q3	Maximum	95% Median CI
Sample of Test 1 scores rounded	20	33.000	67.000	75.000	89.000	94.000	(68.176, 87.589)
Number of Building permits/a	20	73.000	123.500	157.500	181.750	213.000	(130.823, 176.589)