Chapter 3: Describing, Exploring, and Comparing Data

Section 3.1: Measures of Center

will define these more precisely later

Def **Measure of Center** a value at the center or middle of a data set.

The three most widely-used measures of center are the mean, median, and mode.

The **(arithmetic) mean** of a data set is computed by adding all of the values of the variable in the data set and dividing by the number of observations.

The **population** arithmetic mean, μ, is computed using ALL of the individuals in a population. The population mean is a parameter.	The **sample** arithmetic mean, $\bar{x}$, is computed by using some of the individuals in a population. The sample mean is a statistic.
$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum x}{N}$	$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$

pop. mean ⇓ μ para

sample mean $\bar{x} \to$ stat

$\sum$ = "sum"

Ex: Of the 42 students enrolled in an Introductory Statistics course, the data below are the first 10 exam scores. Treat the 10 students as a sample of the population, which means you use $\bar{x}$ to find the mean.

Student	Score
Michelle	82
Ryanne	77
Bilal	90
Pam	71
Jennifer	62
Dave	68
Joel	74
Sam	84
Justine	94
Juan	88

$n=10$

$$\bar{x} = \frac{\sum x}{n} = \frac{82+77+90+\cdots+88}{n} \qquad n=10$$

$$= (82+77+\cdots+88)/10 = 79$$

calc

2nd enter trick

The **median** of a data set is the value that lies in the middle of the data when arranged in ascending order. We use M to represent the median.

ODD number of data 1st: Arrange the data in ascending (↑) order 2nd: The median will be the middle number *Ex:* 11, 14, 16, 19, 28 already ascending M $M=16$	**EVEN** number of data 1st: Arrange the data in ascending order 2nd: The median will be the mean of the middle numbers *Ex:* 14, 18, 20, 26, 31, 39 ← max min $M = \frac{20+26}{2} = \boxed{23}$

The **midrange** of a data set is the value midway between the minimum and maximum values.

$$Midrange = \frac{min\ value + max\ value}{2} \qquad \text{Midrange} = \frac{14+39}{2} = \boxed{26.5}$$

Ex: Use the data from the Introductory Statistics example from above to find the median and midrange.

Use calculator! • Enter data: (STAT → edit ↳ L1)

Sort: • Sort: (STAT → SORTA(L1))

sorted: min 62 68 71 74 77 82 84 88 90 94 max

Median: $M = \frac{77+82}{2} = \boxed{79.5}$

$n=10$ even "middle two" $= 10/2 = 5$, "+1" $= 6$

midrange: $\frac{62+94}{2} = \boxed{78}$

Round-Off Rule: Carry one more decimal place than is present in the original set of values.

↳ "STAT RULE OF ROUNDING"

Ex: Data whole #s, rounding rule is to one decimal place.

The **mode** of a variable is the most frequent observation of the variable that occurs in the data set.
*If no observation occurs more than once, we say that the data have no mode (or mode is N/A)
*If the data set has more than one observation that repeat the same number of time, then it is considered ______________.

Ex: Find the mode for each example below.

a) The following data represent the number of O-ring failures on the shuttle *Columbia* for its 17 flights prior to its fatal flight:	b) The data of the test scores from above:	c) Hair color of ten people in line: (qualitative)
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 2, 3 $\boxed{\text{mode} = 0}$	82, 77, 90, 71, 62, 68, 74, 84, 94, 88 no mode	Brown, Blonde, Red, Brown, Brown, Blonde, Brown, Blonde, Blonde, Red Brown : 4 Blonde : 4 $\boxed{\text{Mode: Brown \& Blonde}}$ "bimodal"

Mean from a Frequency Distribution

Formula: $$\boxed{\bar{x} = \frac{\Sigma(f \cdot x)}{n}}$$

M CB
LCL UCL

Ex: The following table gives the weights of a sample of 100 babies born at a local hospital.

Weight (in lbs)	*Freq* (f)	*class midpt* (x)	$f \cdot x$
3-4.9	5	$\frac{3+4.9}{2} = 3.95$	$(f)(x)$ $5 \cdot 3.95 = 19.75$
5-6.9	32	5.95	190.4
7-8.9	40	7.95	318
9-10.9	18	9.95	179.1
11-12.9	5	11.95	59.75
	$n = \Sigma f =$ 100		$\Sigma(f \cdot x) =$ 767

{this represents alot of data!}

Find the sample mean. $\bar{x}$

$$\bar{x} = \frac{767}{100} = \boxed{7.67 \text{ lbs}}$$

! important! Units!

Resistance Statistics

A numerical summary of data is said to be resistant if extreme values (very large or small) relative to the data do not affect its value substantially.

Ex: The following are wait times (in minutes) at a dentist office: 1, 1, 2, 2, 3, 5. (a) Find the mean ($\bar{x}$) and median (M).

$$\bar{x} = \frac{\Sigma x}{n} = \frac{14}{6} = \boxed{2.3} \qquad M = \frac{2+2}{2} = \boxed{2}$$

b) Note the value of 102 minutes added to this data. Find the mean and median. Which measure is resistant to the added value?

1, 1, 2, 2, 3, 5, 102

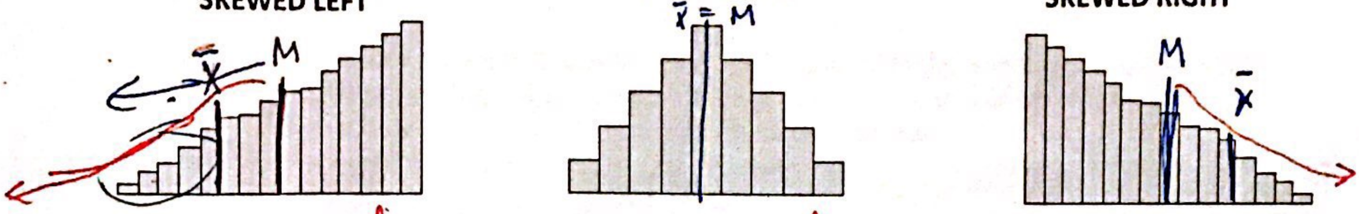
$$\bar{x} = \frac{116}{7} = \boxed{16.6} \qquad \boxed{M = 2}$$

{so mean changed from 2.3 to 16.6} {Median is resistant!}

(Explanation of skewness : tail moves the mean)

When data are skewed, there are extreme values in the tail, which tend to pull the mean in the direction of the tail.

SKEWED LEFT | SYMMETRIC | SKEWED RIGHT



mean < median | mean = median | median > mean

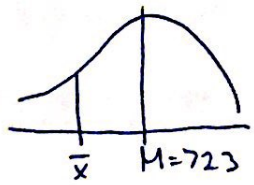
General rule: If the data are symmetric use the MEAN as the best measure of center.

If the data are skewed use the MEDIAN as the best measure of center.

Ex: FICO scores range in value from 300 to 850, with a higher score indicating a more creditworthy individual. The distribution of FICO scores is skewed left with a median score of 723.

(a) Do you think the mean FICO score is greater than, less than, or equal to 723? Justify your response.

mean < 723 $\boxed{\text{mean is less than 723}}$



(b) What proportion of individuals have a FICO score above 723?

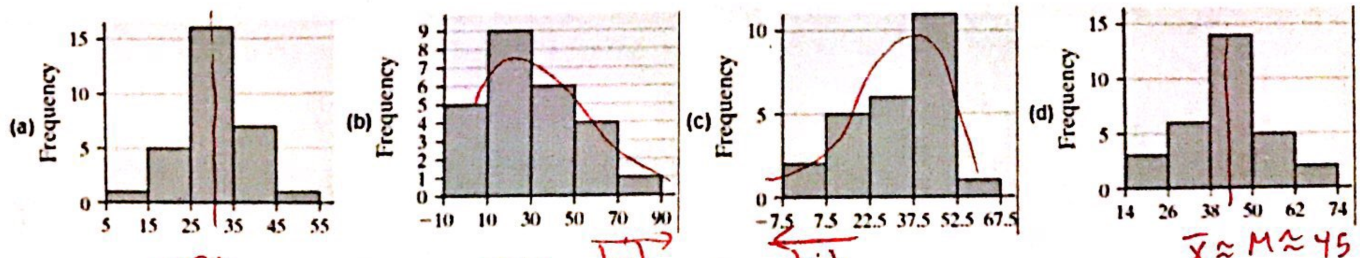
"fraction" since 723 is median = middle

$\boxed{\text{proportion above 723 is 1/2 or 0.5}}$

Ex: Match the histograms shown to the appropriate summary statistics by writing the appropriate number under each histogram.

	Mean	Median
1	42	42
2	31	36
3	31	26
4	31	32

→ $\bar{x} = M$ → symmetric
→ $\bar{x} < M$ → skewed LEFT
→ $\bar{x} > M$ → skewed RIGHT
→ $\bar{x} \approx M$ → symmetric



$\bar{x} = M \approx 30$ (4) | (3) | (2) | $\bar{x} \approx M \approx 45$ (1)

"change/spread"

3.2 Measures of Variation

Importance of Variation

Ex: Advil and Motrin IB produce the same headache relief medication with the active ingredient ibuprofen. Each pill should contain 200 mg of ibuprofen. A health agency obtains a sample of ten tablets from both manufacturers and measures how much ibuprofen each pill actually contains.

	Number of milligrams measured									
Advil	199.25	198.50	200.10	200.75	201.00	198.00	200.10	199.00	201.10	202.20
Motrin IB	205.00	195.80	195.20	203.20	205.80	194.40	204.60	194.60	207.20	194.20

$\bar{x}$ ± 2 from 200

± 7 from $\bar{x}$

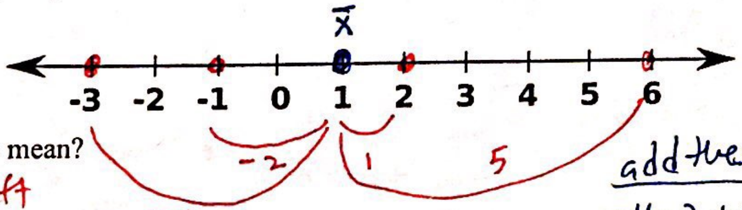
Each sample has a mean value of 200 mg. However, based on the given sample values, which company would you prefer to buy from?

Advil Because all the values are much closer to 200 mg ($\pm \approx 2$ mg from $\bar{x}$) whereas MotrinIB has a much larger variation .... ($\pm \approx 7$ mg from $\bar{x}$)

Ex: The following are temperatures (in degrees) on four consecutive days in Mongolia in January: $-3, -1, 2, 6$

(a) Find the mean. $\bar{x} = \frac{\Sigma x}{n} = \frac{4}{4} = 1$

$\boxed{\bar{x} = 1}$



(b) How far away is each number from the mean?

use +/− to represent right/left side of $\bar{x}$

add these up

$-4 - 2 + 1 + 5 = \underline{\underline{0}}$

not helpful!

Measures of variation

Def The **range** of a data set is the difference between the maximum and minimum data values.

$$\boxed{\textit{range} = \textit{maximum value} - \textit{minimum value}}$$

Standard Deviation of a Sample

s – sample standard dev. σ – population st. dev.

Def The **standard deviation** (denoted by s) of a set of sample values is a measure of variation of values about the mean. It is a type of average deviation of values from the mean.

Formula: $$\boxed{s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{(n-1)}}}$$ "*Shortcut:*" $$\boxed{s = \sqrt{\frac{n(\Sigma x^2) - (\Sigma x)^2}{n(n-1)}}}$$

Note: *Each form can be tricky, but the alternative form tends to be easier.*

Standard Deviation of a Population

Def The **standard deviation** (denoted by σ) of a complete set of values is a measure of variation of values about the mean. It is a type of average deviation of values from the mean.

Formula: $$\boxed{\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}}$$

/4

Note

Note: *It's rare to compute a population standard deviation. Therefore, when using technology, be sure to use the sample standard deviation unless otherwise noted.*

→ **Variance** s^2 or σ^2

Def The variance (denoted by s^2 or σ^2) of a set of values is a measure of variation equal to the square of the standard deviation.

Ex: Find the range, standard deviation, and variance for the following sample of the number of chips in nine randomly sampled fun-sized bags of Doritos.

data → 25 31 28 19 24 26 29 32 20

• $\bar{x} = \frac{\Sigma x}{n} = \frac{234}{9} = 26$

• range

$32 - 19 = 13$

$\boxed{\text{range} = 13}$

x	$x - \bar{x}$	$(x - \bar{x})^2$
19	$19-26 = -7$	$(-7)^2 = 49$
20	$20-26 = -6$	$(-6)^2 = 36$
24	$24-26 = -2$	$(-2)^2 = 4$
25	$25-26 = -1$	$(-1)^2 = 1$
26	$26-26 = 0$	$(0)^2 = 0$
28	$28-26 = 2$	$(2)^2 = 4$
29	$29-26 = 3$	$(3)^2 = 9$
31	$31-26 = 5$	$(5)^2 = 25$
32	$32-26 = 6$	$(6)^2 = 36$
$\Sigma x = 234$		$\Sigma(x - \bar{x})^2 = 164$

• sample standard dev.

$$s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \quad n = 9, \ n-1 = 8$$

$$s = \sqrt{\frac{164}{8}} = \sqrt{20.5}$$

$$\boxed{s \approx 4.5}$$

"stat rule rounding"

Use the alternative form to find the standard deviation.

$$s = \sqrt{\frac{n(\Sigma x^2) - (\Sigma x)^2}{n(n-1)}}$$

$$s = \sqrt{\frac{9(6248) - (234)^2}{9(9-1)}}$$

$$s = \sqrt{20.5}$$

$$\boxed{s \approx 4.5}$$

SAME! :)

x	x^2
19	361
20	400
24	576
25	625
26	676
28	784
29	841
31	961
32	1024
$\Sigma x = 234$	$\Sigma x^2 = 6248$

Question: If you bought a bag of chips everyday, would you prefer to have a small or large standard deviation between bags?

Small! so bag of chips can count on the size of chips being consistent!

Standard Deviation from a Frequency Distribution

Formula:

$$\boxed{s = \sqrt{\frac{n\left[\Sigma(f \cdot x^2)\right] - \left[\Sigma(f \cdot x)\right]^2}{n(n-1)}}}$$

Ex: Recall: The following table gives the weights of a sample of 100 babies born at a local hospital.

Weight (in lbs)	*Freq* (f)	*class midpt* (x)	$f \cdot x$	$f \cdot x^2$
3-4.9	5	3.95	19.75	$5 \cdot (3.95)^2 = 78.0125$
5-6.9	32	5.95	190.4	$32 \cdot (5.95)^2 = 1132.88$
7-8.9	40	7.95	318	$40 \cdot (7.95)^2 = 2528.1$
9-10.9	18	9.95	179.1	$18 \cdot (9.95)^2 = 1782.045$
11-12.9	5	11.95	59.75	$5 \cdot (11.95)^2 = 714.0125$
	$n = \Sigma f = 100$		$\Sigma(f \cdot x) = 767$	$\Sigma(f \cdot x^2) = 6235.05$

Find the sample standard deviation and variance. $\rightarrow$ $s^2 = \left(\sqrt{3.5571}\right)^2 = 3.5571 \rightarrow$

st. dev $\quad s = \sqrt{\dfrac{100[6235.05] - [767]^2}{100(99)}} = \sqrt{3.5571} \approx 1.8860 \quad \boxed{s = 1.89}$

↑ stats rule rounding

Variance: $\boxed{s^2 = 3.56}$

Range Rule of Thumb

Data are significantly LOW if the value is $\mu - 2\sigma$ or *lower*.	***Data are not significant*** if the value is between $\mu - 2\sigma$ and $\mu + 2\sigma$.	***Data are significantly HIGH*** if the value is $\mu + 2\sigma$ or *higher*.

Ex: The data below are free download wifi speeds (in Mbps) from ten of the busiest international airports.

$n = 10$
$n - 1 = 9$

AIRPORT CODE	WIFI SPEED (x)	(x^2)	AIRPORT CODE	WIFI SPEED (x)	(x^2)
DEN	78.2	6115.24	YYC	41.8	1747.24
YVR	55.1	3036.01	BOS	32.2	1036.84
PHL	48.4	2342.56	DFW	32.0	1024
SFO	45.3	2052.09	MEX	27.7	767.29
SEA	43.7	1909.69	DTW	22.9	524.41

$\Sigma x = 427.3$

$\Sigma x^2 = 20555.37$

https://www.speedtest.net/insights/blog/fastest-airports-north-america-2017/

If an international airport began to provide new 60.8 Mbps free wifi, and claimed that their speed is "miles above" many others, would you agree with their claim? → check if significantly hight if "$60.8 > \bar{x} + 2s$"

- $\bar{x} = \dfrac{\Sigma x}{n} = \dfrac{427.3}{10} = 42.73$
- $s = \sqrt{\dfrac{\Sigma(x-\bar{x})^2}{n-1}} = \sqrt{\dfrac{n(\Sigma x^2) - (\Sigma x)^2}{n(n-1)}} = \sqrt{\dfrac{10(20555.37) - (427.3)^2}{10 \cdot 9}} = 15.98 = s$
- $\bar{x} + 2s = 42.73 + 2(15.98) = 74.69.$ $\boxed{\text{Conclusion}}$ "60.8 is not significantly high!"

/6

{Remember: statistical significance is within 95%}

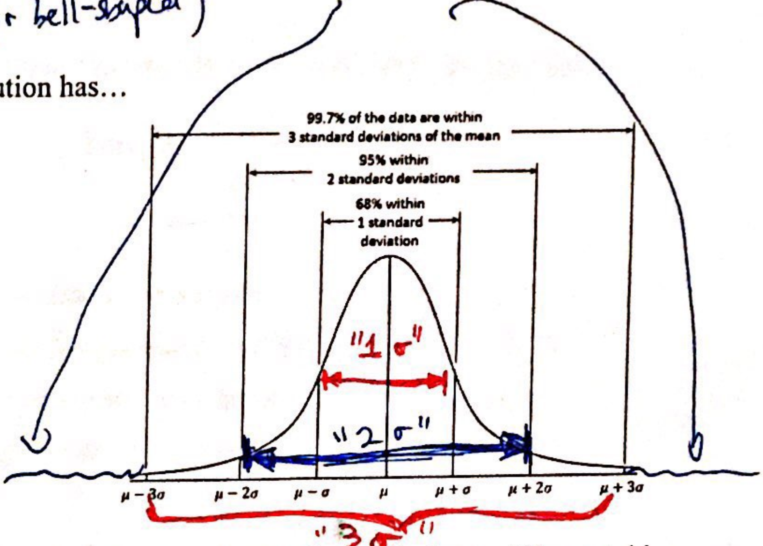
"68-95-99.7 Rule"

Empirical Rule

Empirical Rule says that a symmetric (or normal) (or bell-shaped) distribution has…

- Approximately 68 % of the data will lie within 1 standard deviation of the mean.
- Approximately 95 % of the data will lie within 2 standard deviation of the mean.
- Approximately 99.7 % of the data will lie within 3 standard deviation of the mean.

don't forget thin tails "go forever"

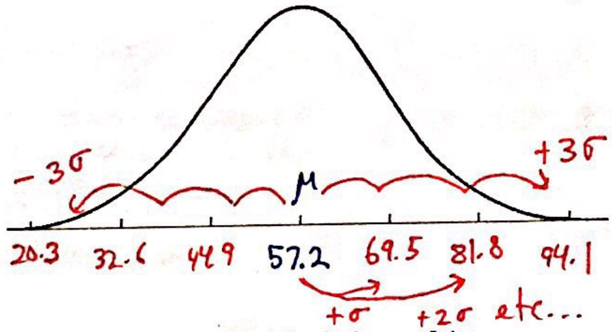


Ex: The following data represent the ages of all of the forty-two female patients of a family doctor. We are told that the data has a bell-shaped distribution and that the population mean, μ, is 57.2 years and the population standard deviation, σ, is 12.3 years.

41	48	43	38	35	37	44
62	75	77	58	82	39	85
67	69	69	70	65	72	74
60	60	60	61	62	63	64
54	54	55	56	56	56	57
45	47	47	48	48	50	52

$\mu = 57.2$

$\sigma = 12.3$



(a) Determine the percentage of all patients whose age is within 3 standard deviations of the mean.

Approximately 99.7 % by Empirical Rule

(b) Between what two values will this percentage lie?

Between 20.3 years and 94.1 years.

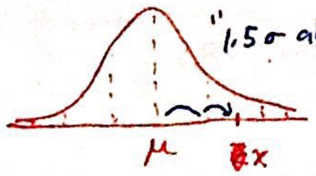
units!!

Finding summary statistics for a data set

Use Graphing Calculator (TI-84 Plus)

Instructions:	(a)	STAT ⇒ 1: Edit... ⇒ Enter data into a list
	(b)	STAT ⇒ CALC ⇒ 1: 1-Var Stats ⇒ List: (Choose list) ⇒ Calculate

Scanned with CamScanner



3.3 Measures of Relative Standing and Boxplots

z Scores

Def A **z score** is the number of standard deviations that a given value x is above or below the mean.

Formula: Sample : $z = \dfrac{x - \bar{x}}{s}$ Population: $z = \dfrac{x - \mu}{\sigma}$

Round-Off Rule: *Round z scores to two decimal places.*

The z score is a standardized value that describes a data value's relative standing.

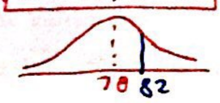
1. A negative z score corresponds to a data value below the mean.
2. Unusual data values are more than two standard deviations from the mean.
   - Ordinary values: $-2 \le z \text{ score} \le 2$
   - Unusual data values: $z < -2$ or $z > 2$
3. The z score allows us to compare data values drawn from different samples or populations.

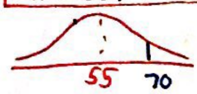
purpose of a z-score!

<u>Ex</u>: Two college roommates are taking different physics courses at a university. They agree to a wager regarding their midterm scores, whereby the loser must do the dishes for a month. After scoring an 82, Jacob insists that Michael lost since he earned a 70 on his exam. However, Michael argues that he performed better relative to the rest of his class than did Jacob. Use the given class results to determine who won the bet? → use z-score to compare

Jacob's class: $\bar{x} = 78,\ s = 6$ Michael's class: $\bar{x} = 55,\ s = 12$

more variation in Michael's class





$z = \dfrac{x - \bar{x}}{s} = \dfrac{82 - 78}{6} \cong 0.67$ $z = \dfrac{x - \bar{x}}{s} = \dfrac{70 - 55}{12} = 1.25$

Michael scored better! b/c Michael's z-score was z = 1.25 which is larger than Jacob's z-score of z = 0.67

Percentiles

Def Percentiles (denoted P_k) are measures of location in relation to all the other data values.

Notation

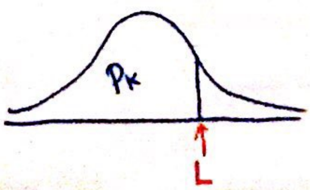
Symbol	Represents
L	locator that gives the rank of a value
P_k	k^{th} percentile

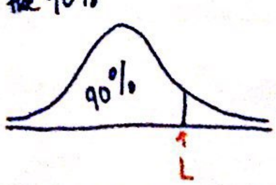
- k percentage (b/w 0% & 100%)
- $P_k = \dfrac{k}{100}$ (decimal)

Finding a Percentile Associated to a Given Score

Formula: $\text{percentile of score } (P_k) = \dfrac{\#\text{ of scores less than given score}}{\text{total number of scores}}$

Ex: the 90%





Also: $\boxed{k = \frac{L}{n} * 100}$ ⟷ $P_k = \frac{L}{n}$

↕

Finding the Score Associated to a Given Percentile

Formula: $$\boxed{L = \frac{k}{100} \cdot n}$$

Note: *If L is a decimal, then always round up to find the score with that specific rank.* i.e. the one in that ordered spot

If L is a whole number, then average the k^{th} score and the next higher score. so 128 is really 128,000

Ex: The following data set represents the selling price (in thousands) of 38 randomly selected homes. ← n

min, 23, M, 10th, 29th, 33rd spot, L, max

128	135	138	145	149	152	155	158	159	163	163	165	167
168	170	170	172	173	176	177	180	180	185	188	191	193
199	205	210	212	215	229	233	250	325	450	500	525	

(a) Find the percentile corresponding to a selling price of $188,000.

$$\% \text{ score} = \frac{\#\text{ less than } 188}{\text{total}} = \frac{23}{38} \approx 60.5\% = \boxed{61\%} \text{ or } \boxed{61^{st} \text{ percentile}}$$

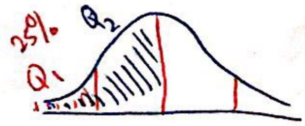
(b) Find the home price corresponding to the 85^{th} percentile. (k)

$$L = \frac{k}{100} \cdot n = \frac{85}{100} \cdot (38) = 32.3 = 33^{rd} \text{ spot} \longrightarrow$$ $233,000 is the price of home at the 85^{th} percentile.

Quartiles

Def **Quartiles** (denoted Q_1, Q_2, and Q_3) are measures of location which divide a data set into four groups with about 25% of the values in each group.

Note: $\boxed{Q_1 = P_{25}}$, $\boxed{Q_2 = \textit{the median}}$, and $\boxed{Q_3 = P_{75}}$



$Q_4 = P_{100}$ = everything!

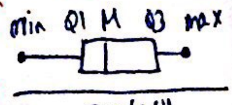
Boxplot

Def A **boxplot** is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile, the median, and the third quartile.

Note: *A five-number summary refers to the five values used to draw the boxplot.*

BOX PLOT / 5 # Summary

· min
· Q1
· M
· Q3
· max



Ex: Find the five-number summary for the previous data regarding home prices.

(a) Find the minimum and maximum values in the data set.

$\boxed{\min = \$128{,}000}$, $\boxed{\max = \$525{,}000}$

(b) Find Q_2 (the median).

$$Q_2 = M = \frac{176 + 177}{2} = 176.5 \rightarrow \boxed{Q2 = \$176{,}500}$$

(c) Find $Q_1 = P_{25}$.

$$Q1 = P_{25} = \frac{25}{100} \cdot (38) = 9.5 \rightarrow 10^{th} \text{ spot}$$

$$Q1 = \boxed{\$163{,}000}$$

(d) Find $Q_3 = P_{75}$.

$$Q3 = P_{75} = \frac{75}{100}(38) = 28.5 \rightarrow 29^{th} \text{ spot}$$

$$Q3 = \boxed{\$210{,}000}$$

(e) List the five-number summary.

$\boxed{\$128{,}000,\ \$163{,}000,\ \$176{,}500,\ \$210{,}000,\ \$525{,}000}$

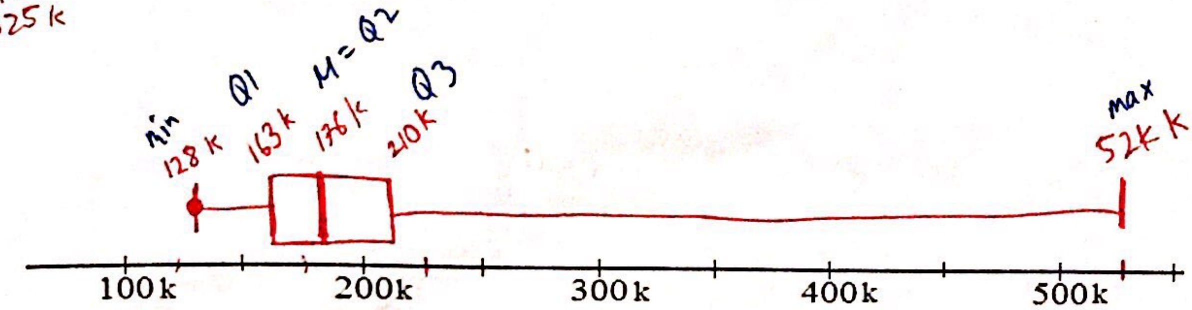
Ex: Construct a boxplot for the previous data set regarding home prices.

Graphing a boxplot:

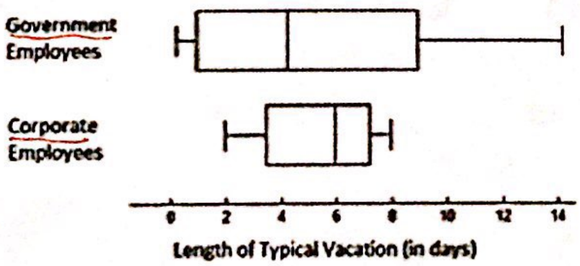
(a) Place the five number summary for a data set on a number line, and draw a straight line connecting them.
(b) Draw vertical lines at each of the five number summary values.
(c) Draw a rectangle connecting Q_1 to Q_3.

5# Summary
- 128 k
- 163 k
- 176 k
- 210 k
- 525 k

important: draw an accurate scale (horizontal axis) first, then draw box-plot.



Ex: Below, boxplots are shown for the length of a typical vacation for California residents who work for the government in some capacity and for those who work for a private company.



Based on these graphs, would you prefer a government job or corporate job based solely on the length of their vacations? There is no one right answer, but please consider **center, spread**, and any other relevant statistics or values from the boxplots shown to support your data.

Pro Government

* larger spread so can sometimes have short vacay (0 days) but sometimes have long vacay (14 days)!
* median = 4 days, so 50% of employees have ~~long~~ vacay length 4 to 14 days

Pro-Coorporate

* less spread so pretty consistent vacay length of about 6 days (better to plan!)
* at least 2 days off guarenteed where govt can have 0 days off
* median = 6 days so median higher
  50% have vacay 2 to 6 days
  50% have vacay 6 to 8 days

/10