Yay!! Calculations!

Unit 3 – Numerical Summaries

Your Calculating Professor Colton

Unit 3 Outline

Measures of Center

- Mean
- Median
- Best Measure of Center
- Mode
- Weighted Mean
- Examples with Calc

Numerical Summaries Intro

- We got data.
- We learned how to display it.
- But numbers provide great info too.
- How can we **describe** our data??

```
 [1] 0.35 0.30 1.25 0.30 0.30 1.51 0.57 0.73 1.01 0.54 0.42 2.01 1.71 2.01 0.51
[16] 0.31 0.35 0.70 0.90 0.90 2.48 0.74 1.12 0.51 0.86 2.01 0.50 0.41 0.38 0.90
[31] 1.02 1.13 2.01 1.00 1.50 0.73 1.34 1.01 1.20 1.36 1.04 1.27 0.30 0.23 1.01
[46] 0.40 1.54 0.40 1.52 0.93 1.02 1.03 0.30 0.41 0.31 0.76 0.55 0.58 0.61 0.42
[61] 0.91 1.00 0.70 0.71 0.31 0.70 0.57 0.42 0.56 0.32 0.70 0.60 0.40 0.32 0.30
[76] 0.50 0.43 0.35 1.74 1.02 1.24 0.31 1.07 1.50 0.33 0.42 0.31 2.50 1.23 1.25
[91] 0.31 0.34 1.14 1.52 0.60 1.06 0.90 1.14 0.41 1.09
```

Carat	Frequency	Relative Frequency
[0.2,0.487]	70	0.350
(0.487,0.775]	43	0.215
(0.775,1.06]	31	0.155
(1.06,1.35]	28	0.140
(1.35,1.64]	15	0.075
(1.64,1.92]	4	0.020
(1.92,2.21]	7	0.035
(2.21,2.5]	2	0.010

- If we get ONE number, what do we choose?
- Edge numbers (min, max), middle (how define middle?), spread numbers (range, others?).
- Each choice has its advantages and disadvantages

Mean

Data: 1, 5, 2, 9, 0, ~~3~~, 90

$$\bar{x} = \frac{\sum x_i}{n} = \frac{1 + 5 + 2 + 9 + 0 + \cancel{3}\ 90}{6} \approx 3.33$$

≈ 17.83

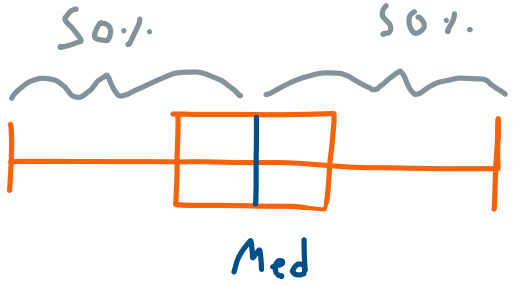
Quantitative Data!

Mean (Average)

- Most common and most informative measure of center.
- Simple, arithmetic average of the data.
  - Sum all numbers ($\sum x_i$) and divide by the sample size (n)!
- Notation:
  - Sample mean = $\bar{x}$ (pronounced "x-bar"), this is both the writing notation and your calculator notation
  - Population mean = μ (the Greek letter mu, keyboard trick: option + m = μ)
  - THERE IS A BIG DIFFERENCE BETWEEN THESE TWO!
    - $\bar{x}$ is a statistic because it describes a sample! While μ is a parameter because it is referring to a population

- **Mean is NOT a resistant measure.**
  - This means this summary statistic can be *heavily* influenced by extreme observations.
  - Gets pulled in the direction of any existing extreme data values.

Median



Median

- Another measure of center. Less informative than the mean.
- It is the middle of the data.
  - It's literally just the middle data value when data is ordered smallest to largest
  - Mini caveat when there is an even number of observations
    - In which case the median is the average (midpoint) of the two middle values.
- Notation:
  - NOT really regular symbol for median, although it is technically the second Quartile Q_2
  - On your calculator, it is Med.

- **Median is considered a resistant measure.**
  - This descriptive statistic is NOT affected by extreme data.

Case 1 – Odd *n*

9 Obs: 10, 5, 6, 1, 3, 9, 8, 4, 4

Sorted: ~~1, 3, 4, 4~~, 5, ~~6, 8, 9, 10~~

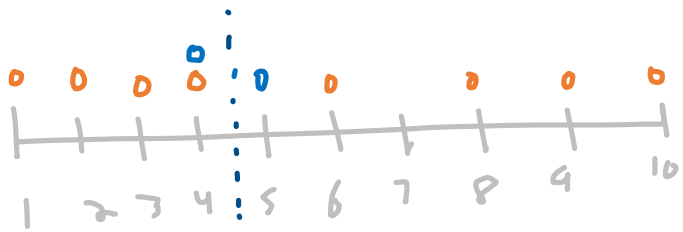
-> Med (Q_2) = 5

Case 2 – Even *n*

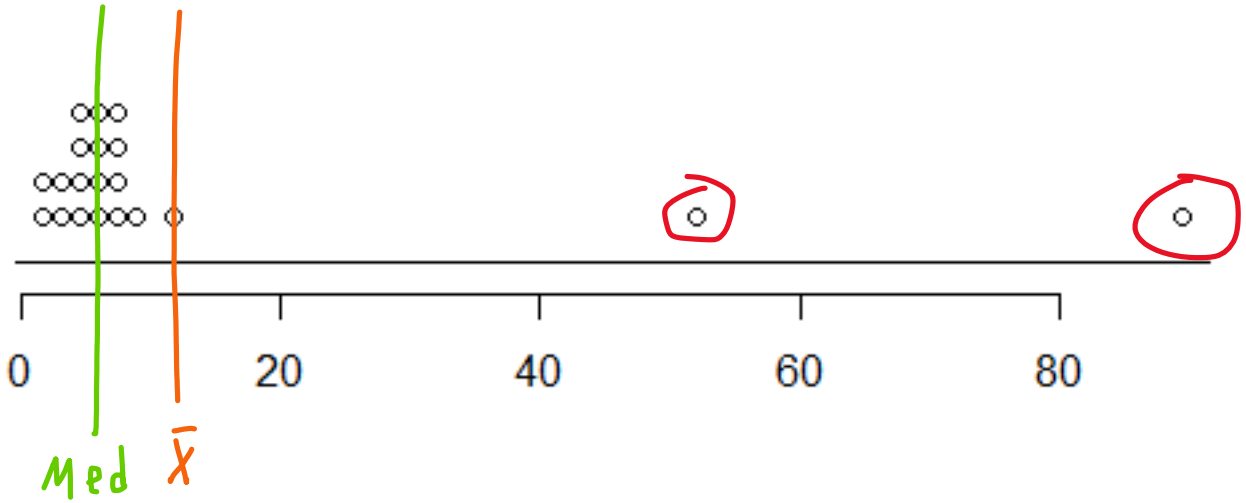
10 Obs: 10, 5, 6, 1, 3, 9, 8, 4, 5, 2

Sorted: ~~1, 2, 3, 4~~, 4, 5, ~~6, 8, 9, 10~~

-> Med (Q_2) = $\frac{4+5}{2} = 4.5$



Best Measure of Center - Visual



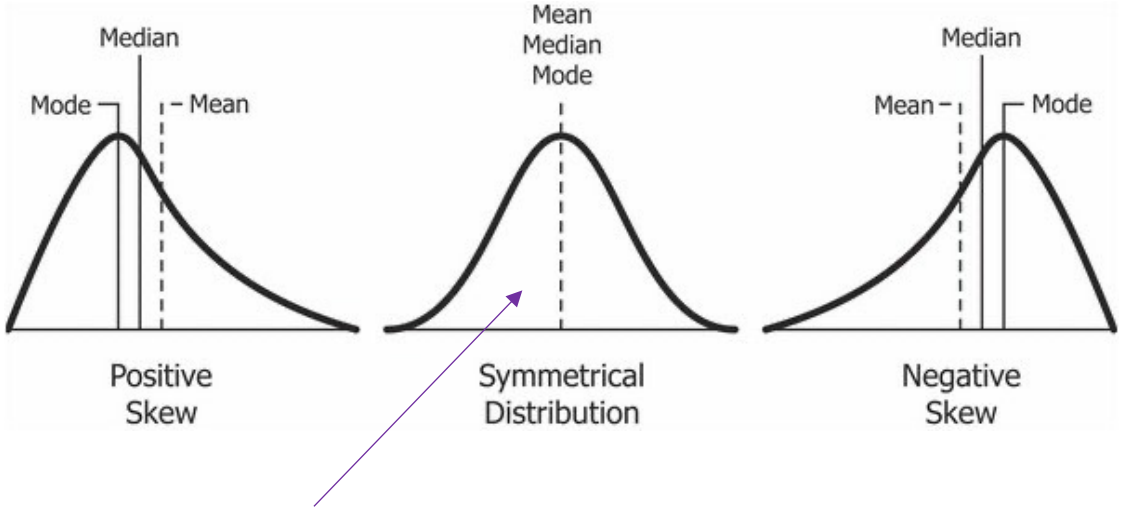
GOAL: Choose the best measure of center!

- Mean
  - It was pulled upwards from the two outliers
  - So there are only three observations above it and 16 observations below
  - Doesn't really capture the middle of the bulk of the data
- Median
  - It is actually in the middle of the majority of the data because unaffected by the outliers
  - Gives us the best indication of where the data is centered and is the most appropriate measure

Best Measure of Center

Comparison

- BOTH measure center, but differently.
  - Mean actually uses the VALUES of each data point → MOST INFORMATIVE
  - Median (for the most part) ONLY uses the POSITION of each data point.
- This is why mean is *NOT resistant* while median *resistant*.



Mean and median are equal here, so use the most informative one!

https://www.quora.com/What-does-SKEWED-DISTRIBUTION-mean

When to Use

- **Mean** is best when…
  - Data is symmetric (NOT skewed).
  - AND NO outliers to pull the mean away from the center.
- **Mean gets pulled towards the tail of skewed data.**
  - So it no longer accurately represents the majority of the data
  - It is "falsely inflated" from the few observations in the tail
- When this is the case, use the **Median.**
  - The median is *unaffected by the skewness* and therefore would now be the MOST APPROPRIATE measure of center.
  - In other words, it better represents the center point where the data is spread around

EXCEL DEMO

IMPORTANT LCQ: Best Measure of Center

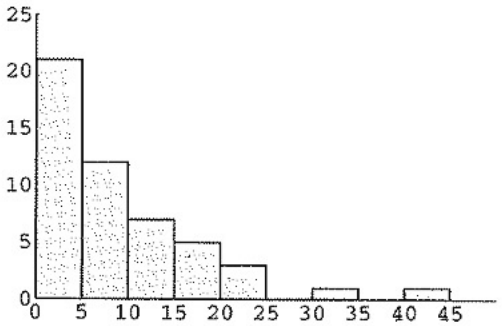
Which **measure of center** is needed?????

a) Stem-and-leaf plot for the hotel rate data

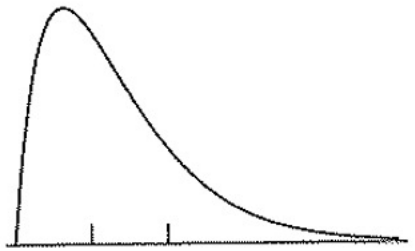
5	0 4
5	5 8
6	0 1 1 3 4
6	5 8 9 9
7	0 1 2 2 2 3 3 3 3 4
7	5 5 5 5 5 6 7 7 7 7 8 8 8 8 9 9 9 9
8	0 1 1 1 3 4
8	5 5 6 9 9
9	3 3 4
9	0 1

Stem = tens
Leaf = ones

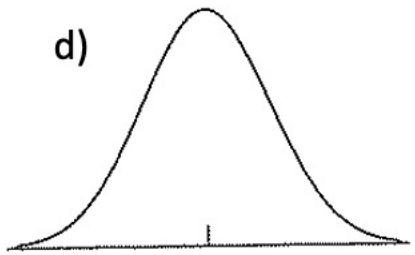
b)



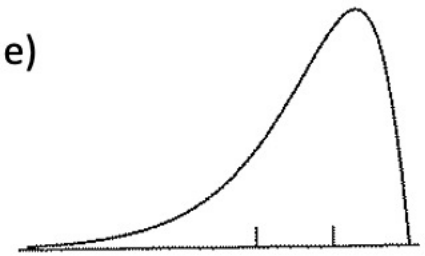
c)



d)



e)



IMPORTANT LCQ: Best Measure of Center

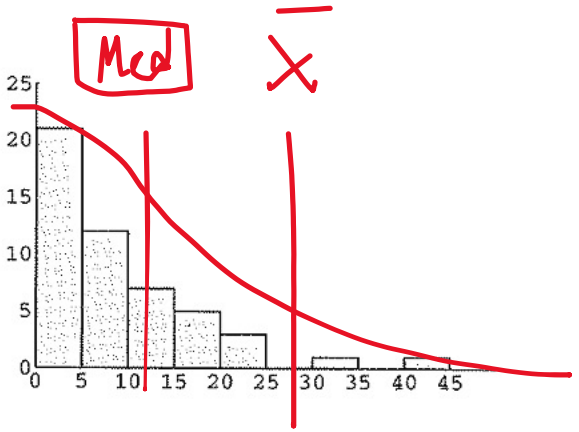
Which **measure of center** is needed?????

a) Stem-and-leaf plot for the hotel rate data

5	0 4
5	5 8
6	0 1 1 3 4
6	5 8 9 9
7	0 1 2 2 2 3 3 3 3 4
7	5 5 5 5 5 6 7 7 7 7 8 8 8 8 9 9 9 9
8	0 1 1 1 3 4
8	5 5 6 9 9
9	3 3 4
9	0 1

Stem = tens
Leaf = ones

b)

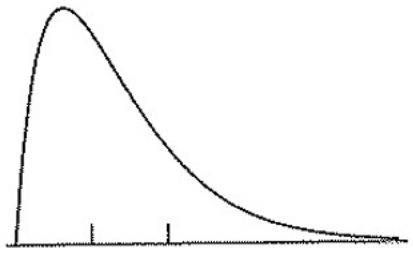


Decisions

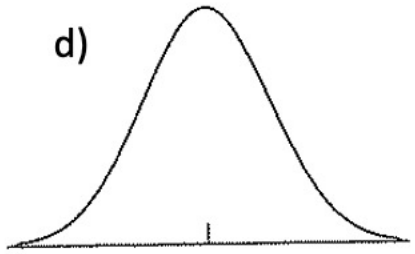
- *Start by thinking we should use the mean because it is the most informative*
- *Then check if it is still appropriate to use the mean with the two below (can think of these as "tests" or "conditions" that the distribution needs to pass)*

1. *Shape: Symmetric or skewed?*
2. *No outliers or outliers?*

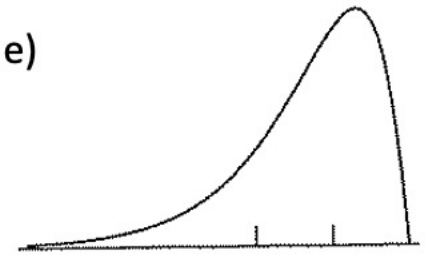
c)



d)



e)



a) Mean
b) Median
c) Median
d) Mean
e) Median

LCQ: Measure of Center and Shape

Use the values of the sample mean and the sample median to determine whether the distribution is symmetric, skewed left, or skewed right.

a) Sample mean = 67 Sample Median = 49

b) Sample mean = 23.5 Sample Median = 23.5

c) Sample mean = -17 Sample Median = -6

LCQ: Measure of Center and Shape

Use the values of the sample mean and the sample median to determine whether the distribution is symmetric, skewed left, or skewed right.

a) Sample mean = 67 Sample Median = 49

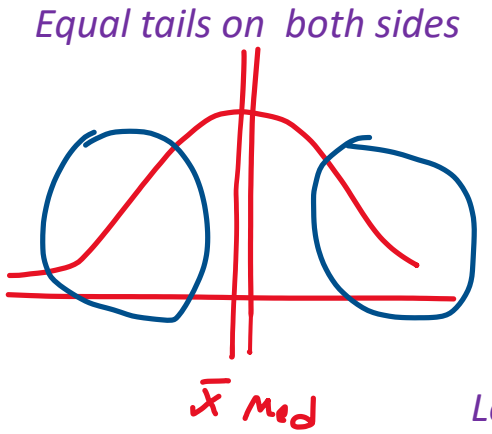
Sample mean > Sample Median, ***Skewed right***

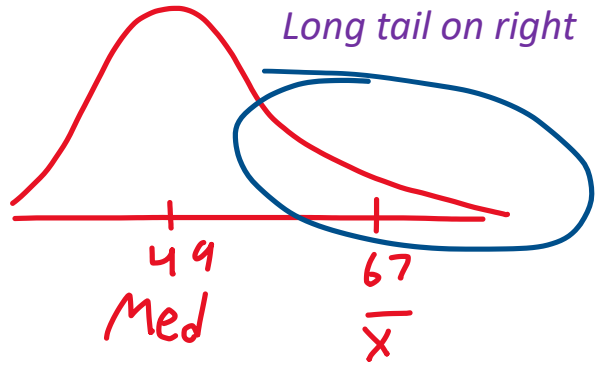
b) Sample mean = 23.5 Sample Median = 23.5

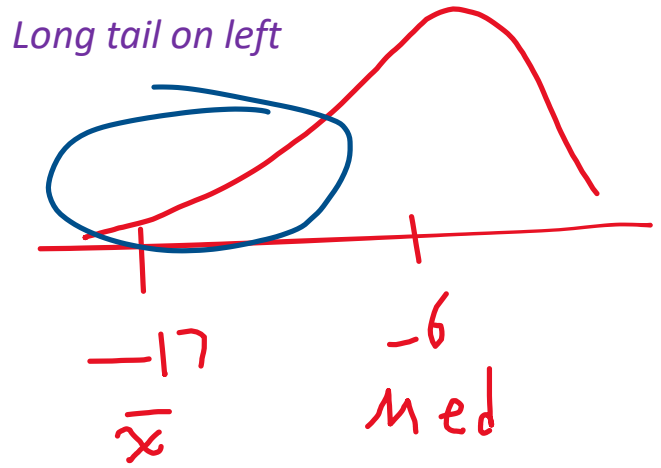
Sample Mean ≈ Sample Median, ***Symmetric***

c) Sample mean = -17 Sample Median = -6

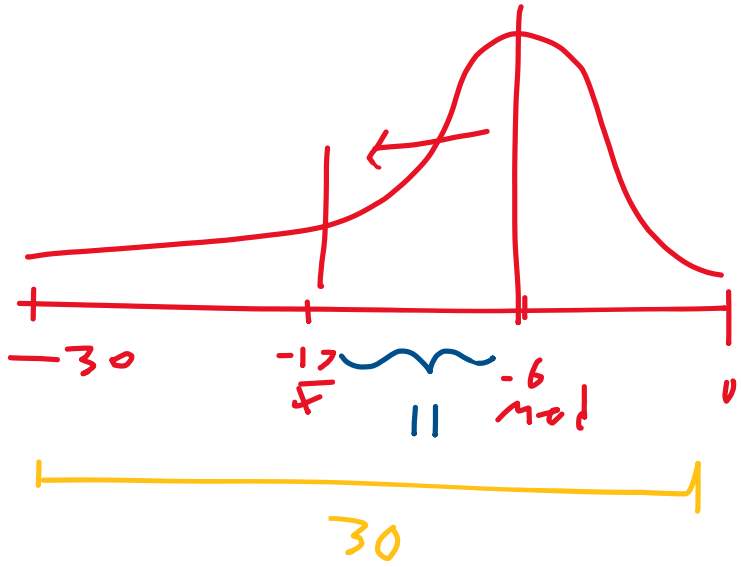
Sample Mean < Sample Median, ***Skewed left***

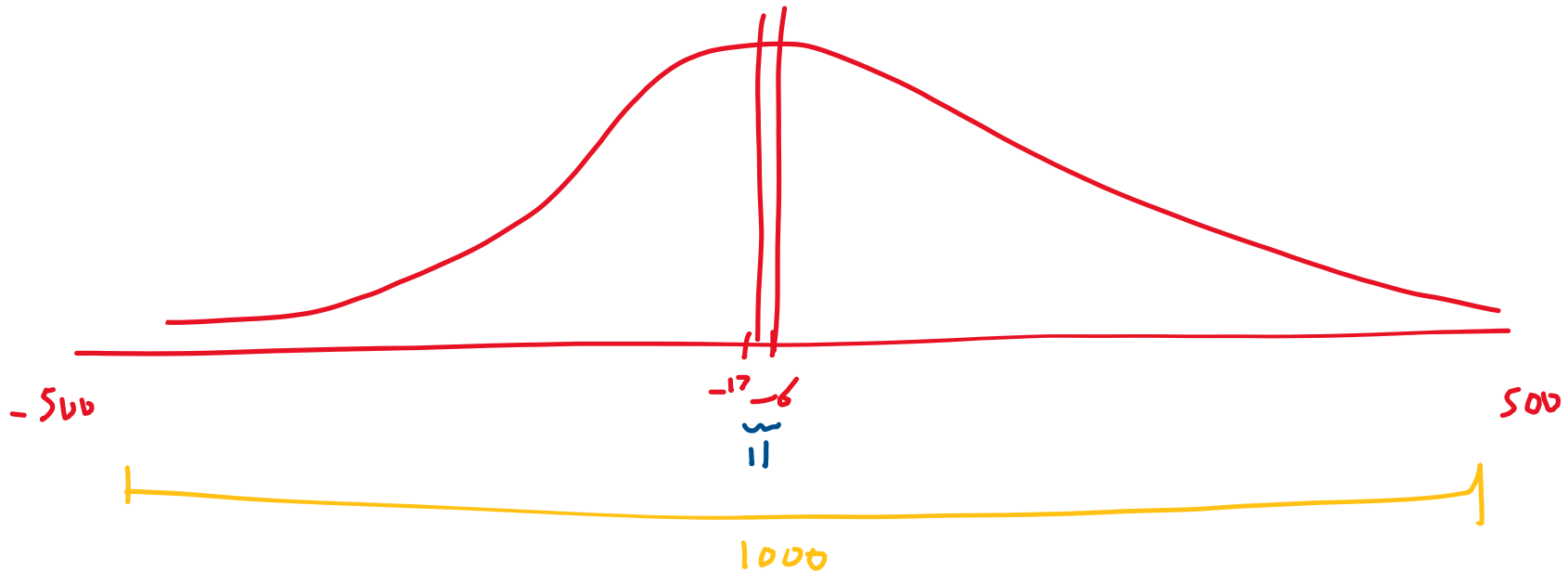






- *Just comparing the mean vs the median to determine the shape of a distribution could be misleading*
- *We also have to take into account the scale or the range of the entire distribution to get a clear idea of the shape*

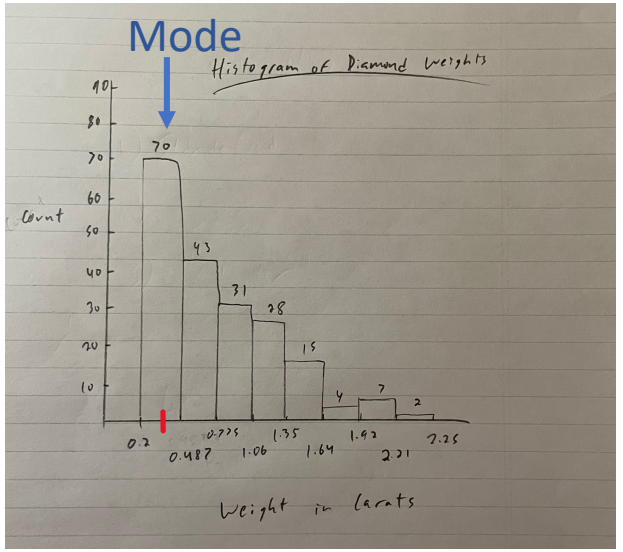




Mode

Bin	Midpoint	Frequency	Relative Frequency
0-10	5	3	0.15
10-20	15	2	0.1
20-30	25	4	0.25
30-40	35	4	0.25
40-50	45	7	0.35
Total:		20	1

Mode



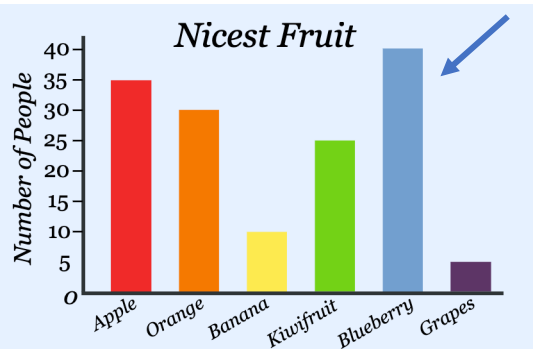
Mode

- Most common data value.
  - If all data occur the same number of times, there is no mode.
  - Can have more than one mode.

- Frequency Tables and Histograms
  - The actual value for the mode is identified by the midpoint of the class (bin) that contains the most data.
  - It is sometimes called the *modal class.*

- The Mode is the ONLY measure of center that works with qualitative data as well.

Mode



https://www.mathsisfun.com/data/bar-graphs.html

Weighted Mean

Weighted Mean (Mean of Grouped Data)

- Used for Frequency Tables, when we don't have the raw data.
- Uses midpoints to estimate each class and frequencies to provide "weight" for each class.

Simple Example:

Data (10 obs):
0, 1, 2, 3, 3, 4, 4, 5, 6, 6

Bin	Midpoint	Frequency
0-2	(0+2)/2 = 1	3
2-4	3	2
4-6	5	5

Regular $\bar{x} = \frac{\sum x_i}{n} = \frac{1}{10}(0 + 0 + 1 + 2 + 3 + 4 + 4 + 5 + 6 + 6) = \mathbf{3.1}$

Weighted $\bar{x} = \frac{\sum m_1 + \sum m_2 + \sum m_3}{n} = \frac{1}{10}(1 + 1 + 1 + 3 + 3 + 5 + 5 + 5 + 5 + 5) = \mathbf{3.4}$

Calculator Fun Session 2!!!!

LCQ / Example 1: 1-Var Stats

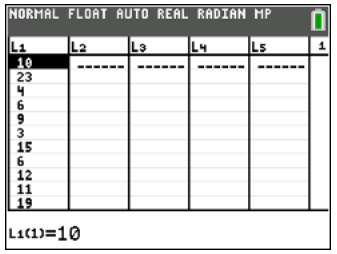
GOAL: Find some Descriptive (Summary) Stats!

1. Enter data in L_1 (15 obs):
   - 10, 23, 4, 6, 9, 3, 15, 6, 12, 11, 19, 10, 6, 8, 15
2. 1-Var Stats
   a) List is L_1.
   b) Leave FreqList blank.
   c) Calculate!

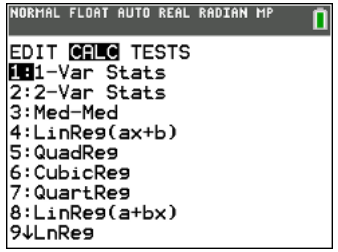
Interpret results:
- Let's say this data is from a sample of the number of eggs hens laid in the month.
- **Find** the mean. **Show** your work. **Interpret** in context.

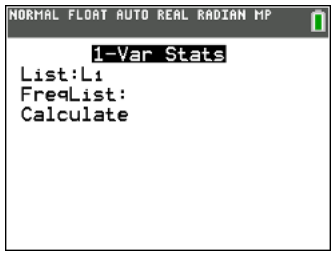
LCQ / Example 1: 1-Var Stats

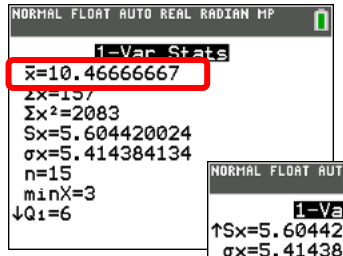
GOAL: Find some Descriptive (Summary) Stats!

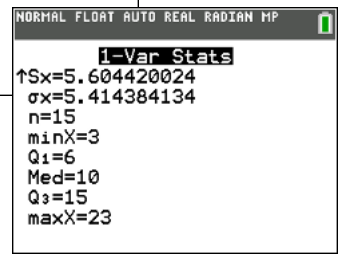


1. Enter data in L_1 (15 obs):
   - 10, 23, 4, 6, 9, 3, 15, 6, 12, 11, 19, 10, 6, 8, 15
2. 1-Var Stats
   a) List is L_1.
   b) Leave FreqList blank.
   c) Calculate!









Interpret results:

- Let's say this data is from a sample of the number of eggs hens laid in the month.
- **Find** the mean. **Show** your work. **Interpret** in context.

$\bar{x}$ = 10.467, used calc to do 1-Var Stat on L_1 → *this is a words explanation of what you did*

or

$\bar{x}$ = 10.467, 1-Var Stats(List = L_1) → *this is kinda like writing the calc function and inputs*

- The average number of eggs that hens laid in the month was approximately 10.47

Could also calculate and show work via Excel, such as x-bar = AVERAGE(A1:A15)

- *But need to get comfortable using calc and calc is just easier and gives more info ☺*

LCQ / Example 2: Weighted Mean

GOAL: Find the Weighted Mean!

Bin	*Midpoint*	Frequency
0-10		3
10-20		2
20-30		5
30-40		4
40-50		6

1. Calculate Midpoints of each bin.
2. Enter data.
   a) Midpoints (our estimates) go in L_1.
   b) Frequencies (our weights) go in L_2.
3. 1-Var Stats
   a) List is L_1 (Midpoints).
   b) FreqList is L_2 (Frequencies).
   c) Calculate!

Interpret results:

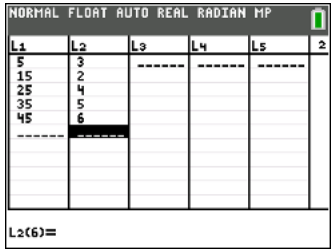
- Let's say this data is about the number of homework assignments students had last semester in all of their classes.
- **Find** the weighted mean. **Show** you work. **Interpret** in context.

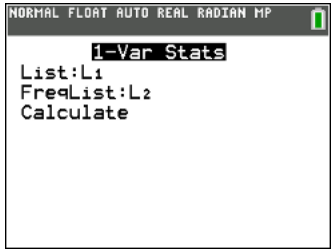
LCQ / Example 2: Weighted Mean

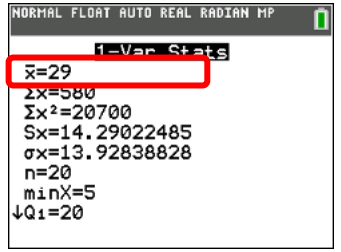
Bin	*Midpoint*	Frequency
0-10	*(0+10)/2 = 5*	3
10-20	*(10+20)/2 = 15*	2
20-30	*25*	5
30-40	*35*	4
40-50	*45*	6

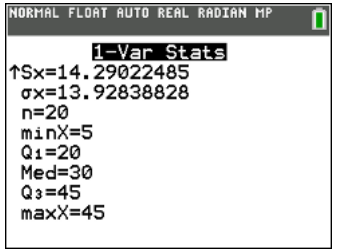
GOAL: Find the Weighted Mean!

1. Calculate Midpoints of each bin.
2. Enter data.
   a) Midpoints (our estimates) go in L_1.
   b) Frequencies (our weights) go in L_2.
3. 1-Var Stats
   a) List is L_1 (Midpoints).
   b) FreqList is L_2 (Frequencies).
   c) Calculate!









Interpret results:

- Let's say this data is about the number of homework assignments students had last semester in all of their classes.
- **Find** the weighted mean. **Show** you work. **Interpret** in context.

Showing midpoints.

Weighted $\bar{x} = 29.5$, 1-Var Stats(List = L_1, FreqList = L_2) *(could also say used graphing calc to run a 1-Var Stats on Midpoints and Frequencies)*

- The weighted average of number of homework assignments students had last semester was 29.5

LCQ / Example 1: Mode and 5-Number Summary

GOAL: Find the Mode and use the 5-Number Summary to sketch a boxplot!

1. Enter data in L_1 (12 obs):
   - 90, 86, 60, 73, 77, 94, 83, 90, 90, 87, 80, 71
2. Sort data.
   - Then can look at new list to find the Mode easily:
3. 1-Var Stats
4. Report 5-Number Summary
5. Sketch a boxplot!

LCQ / Example 1: Mode and 5-Number Summary

GOAL: Find the Mode and use the 5-Number Summary to sketch a boxplot!

1. Enter data in L_1 (12 obs):
   - 90, 86, 60, 73, 77, 94, 83, 90, 90, 87, 80, 71
2. Sort data.
   - Then can look at new list to find the Mode easily: *Mode = 90*
3. 1-Var Stats
4. Report 5-Number Summary
   - *Min = 60*
   - *Q1 = 75*
   - *Med = 84.5*
   - *Q3 = 90*
   - *Max = 94*
5. Sketch a boxplot!