Chapter 3   Video and Popper 3

3.1 Measures of Center

These are the numbers that describe what is normal, usual, and in the middle or the center. These terms are very loose and need firming up mathematically, of course.

Mode $\dot{x}$ dot

Median $\tilde{x}$ tilde

Mean $\bar{x}$ bar

Mode

One measure of central tendency is the **Mode**. Symbolized: x under a dot.

This is the number that occurs **most frequently** in a data set.

A data set doesn't always have a mode – if each data point is a different number the set is mode-free. The mode is always a number in the data set, if there is one.
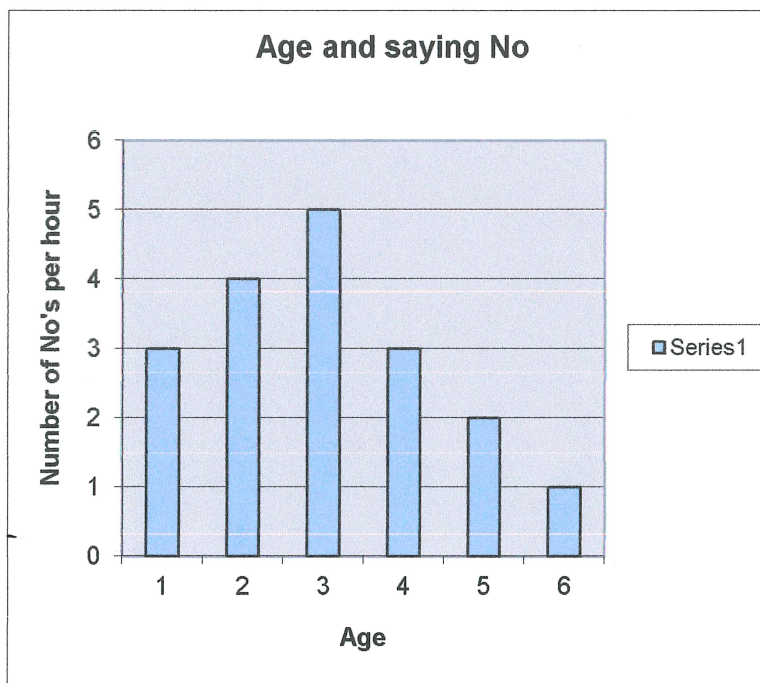
Some data sets have a mode; some are bi-modal or multimodal.

Let's look at a histogram of young children saying no. The age is on the horizontal axis and the frequency per hour is on the vertical axis. Note it is easy to see the mode on a histogram.

Example: Mode

Which of the following bars shows the mode in this histogram?

Note frequency on the vertical axis

excel

Age and saying No

(Vertical axis: Number of No's per hour, values 0 to 6)
(Horizontal axis: Age, values 1 to 6)
(Legend: Series1)

Median

Another measure of central tendency is the **Median**:

The median is the value that is at the numerical middle of the data if there are an odd number of data points and they are arranged in order by size. It is the mean of the 2 middle data points if the number of data points is even and arranged in order by size. So sometimes it's a number in the set; sometimes not.

The formula for finding the location of the median for n data points is
$$0.5(n + 1).$$
The process is to order the data and then find the measurement at that location.

✶ Symbolized by x under a tilde.

Example: Median and Mean

Find the median location for

Data set A.   n = 19 data points

Half of 19 +1 is 10. The 10th measurement whatever that number is

Data set B.   n = 52 data points

Half of 53 is 26.5, the average of measurement 27 and measurement 26. Might not even be in the data set!

Chapter 3 Popper Question 1

Is the median the same number as it's location in a data set that is organized from smallest to largest?

A.   Unlikely

B.   Absolutely 100% of the time

Homework Problem 1

In golf the holes are rated for a recommended number of strokes needed to sink the golf ball into the hole. A score of *par* means the golfer used the recommended number, a *birdie* is one fewer than recommended, a *bogey* is one more than the recommended number, an *eagle* is 2 fewer strokes.

At a recent televised tournament, 7 golfers had the following scores, ranked alphabetically by last name: par, birdie, par, par, birdie, bogey, and eagle.

Where is the median score located?   What is the median score? Make a dot diagram of this and note that it's ok to have negative numbers on the horizontal axis of a dot diagram.

Mean

The most popular measure of "centeredness" is the **Mean** (sometimes called the average).
The mean of n numbers is the sum of the numbers divided by n. If you are working with a data set of measurements, the mean is denoted: $\bar{x}$.

There are some very cogent reasons for its popularity:

It can always be calculated and it's easy to calculate.

It is unique: there is only ONE mean for a data set.

It uses EVERY data point; nothing is eliminated.

It doesn't depend on chance or luck.

There is an equally important reason to take the mean with a grain of salt:

It is heavily affected by outliers!   *curve wreckers*

Homework Problem 2

Here is a list of home prices:

$77,500 $78,200 $110,500 $137,000 $1,800,300

What is the MEAN?   What the the MEDIAN?
Which is a better measure of center, usual, normal?
Why are they different dollar amounts?

Chapter 3 Popper Question 2

Which of the following is a negative about the mean?

A.   It is unique.
B.   It is easy to calculate.
C.   It uses every data point.
D.   It can be pulled up or down by outliers.
E.   It doesn't depend or chance of luck.

Let's look at this problem together:

An elevator in PGH is designed to carry a maximum load of 3,200 pounds. If it is loaded with 18 people with a mean weight of 166 pounds, is it in any danger of being overloaded?   18(166) is 2988 Hint

Relationships among Mean, Median, and Mode, an example

The data shown in the table are the median prices of existing homes in the USA from 1981 through 1986. If the average prices of existing homes were calculated for each of these years, how do you think these values would compare to the median prices shown?

Would the average price be higher, lower, or the same?

Year	Median
1981	66,460
1982	67,800
1983	70,300
1984	72,400
1985	75,500
1986	80,300

Median is not affected by outliers the way average (or mean) is

Example Mean Median Mode:

Here are 3 data sets in one table. The graphs for them follow.

Let's look at the titles for a minute. The first column is the horizontal values, the inputs. The next three columns are FREQUENCY data for the x values in three different data sets.

H.   ✓   ✓   ✓

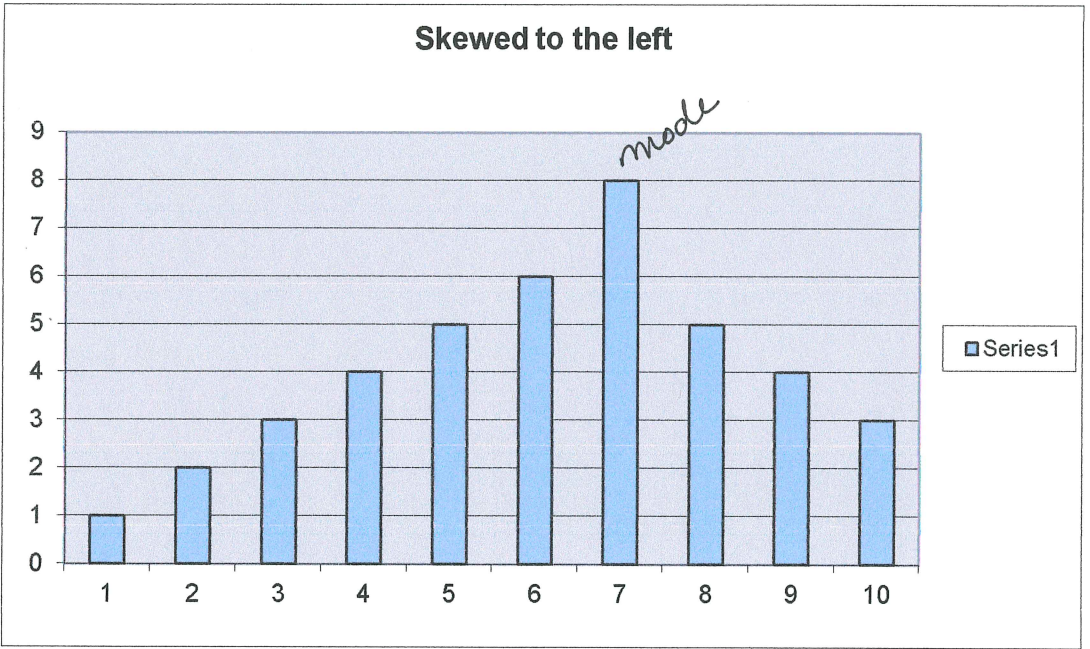| x axis	STTR	STTL	Symm
1	1	1	1
2	2	2	2
3	4	3	3
4	5	4	4
5	4	5	5
6	3	6	5
7	2	8	4
8	2	5	3
9	1	4	2
10	1	3	1

Skewed to the right

Skewed to the left

Symmetric

Calculate mean, median, and mode for these 3 charts. Mark on the x-axis where each goes. How many data points in each set?
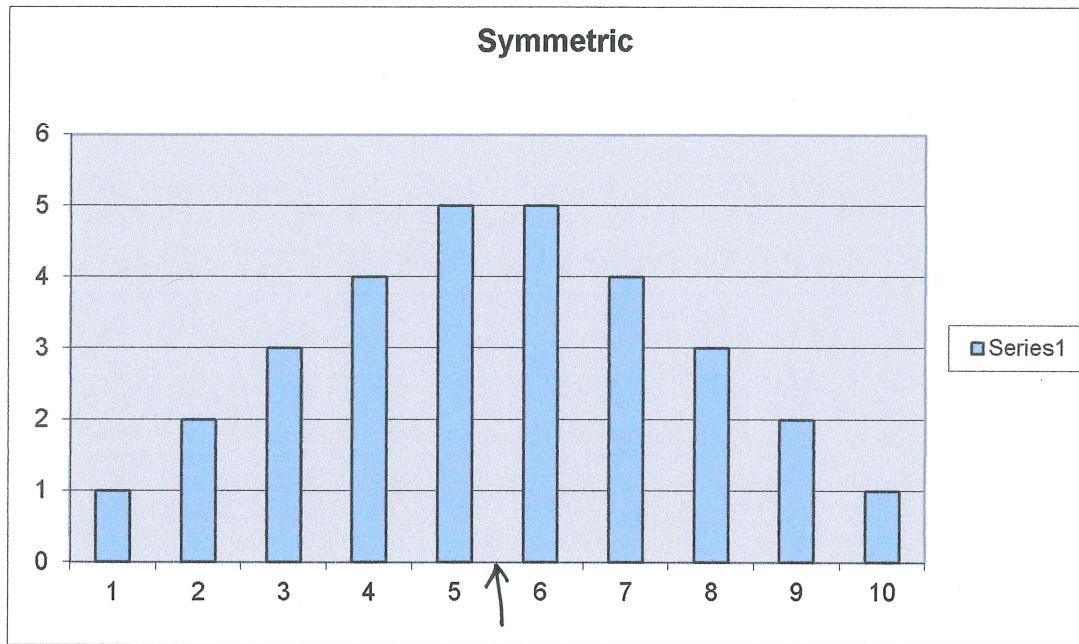
Skewed to the right
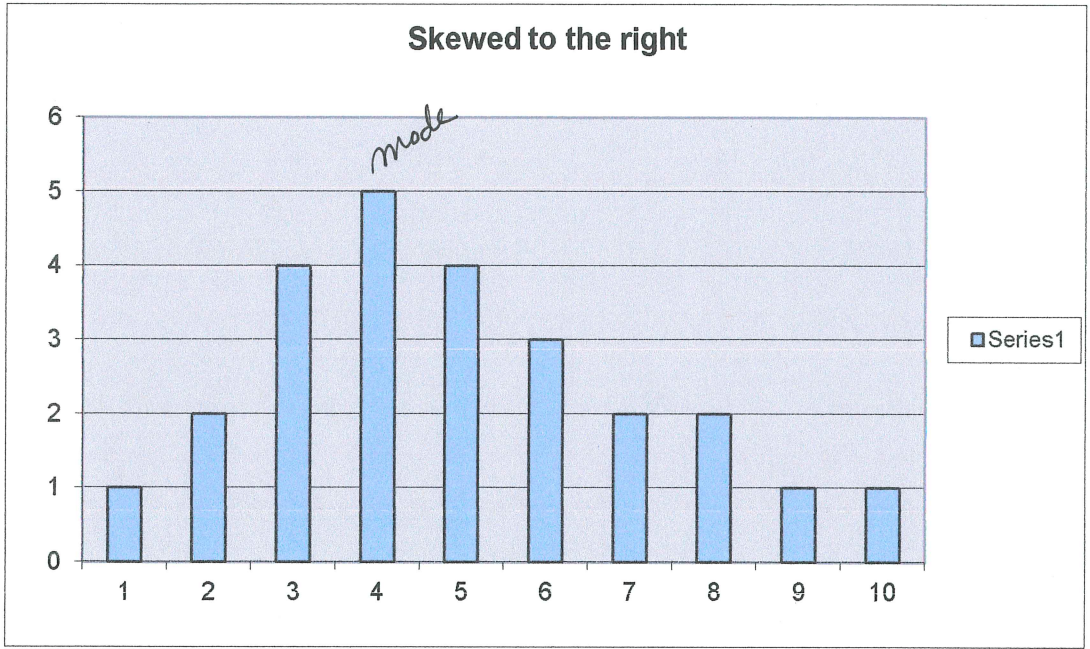
N = 25 median is 13th measurement

Skewed to the left

N = 41 median is 21st measurement

$1 + 2 + 3 = 6$
$+ 4 + 5 + 6$
$\boxed{21}$

the sixth 6 is the median

mean will be pulled leftward a bit

7

Symmetric

N = 30 Median is avg 14th and 15th

OYO: Summarize your results with a mnemonic device.

Chapter 3 Popper Question 3

Which measurement is more sensitive to outliers?

A.   Mean

B.   Median

3.2 Measures of Spread or Variability

Range   Max - Min

***Graphing Calculator, page 60

Variance:

Mean deviation   p. 58

$$\frac{\sum |x - \bar{x}|}{n}$$

The mean deviation is calculated by doing the following:

Calculate the mean.

Subtract the mean from each data point. Take the absolute value of each difference.

Add up the positive differences.

Divide by n, the number of data points.

Standard deviation   p. 60

Variance: $$\frac{\sum (x - \bar{x})^2}{n - 1}$$

The standard deviation for a set of data is the square root of the variance.

graphing calculator p. 61

The sample variance is calculated by doing the following:

First calculate the sample mean,

then subtract the mean from each measurement individually and

square the answer.

Add up all the squares and divide by $n - 1$.

The **Standard Deviation** is the square root of the Variance.

It is preferred to the Mean Deviation because it is more conservative in tagging outliers. It is a larger number (see the denominators n vs. n – 1)…and thus tags fewer measurements.
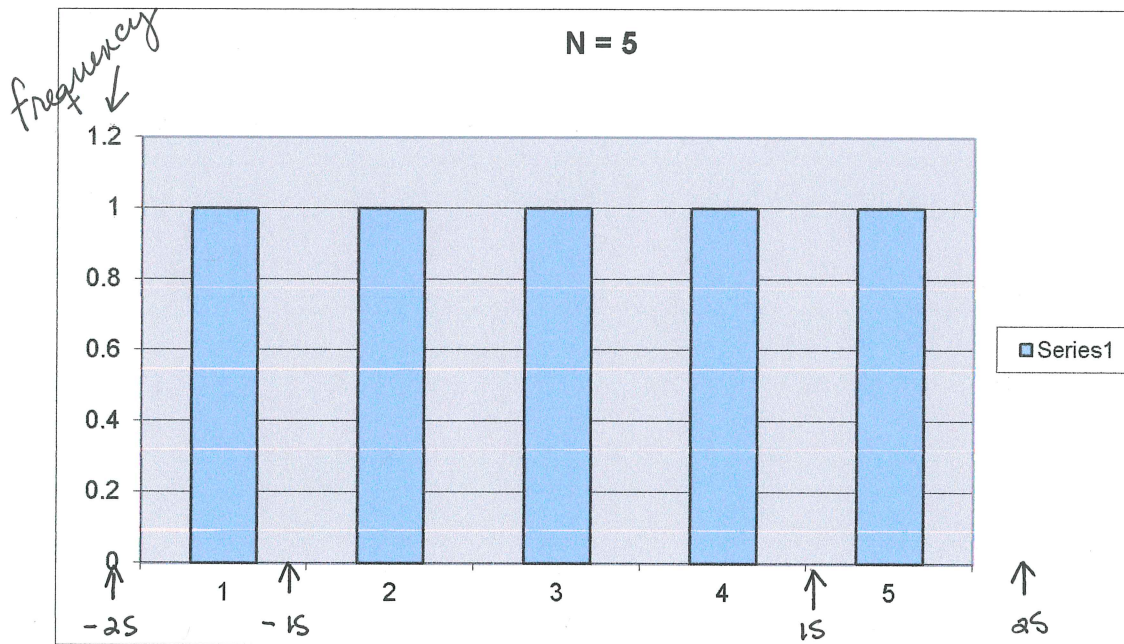
Homework Problem 3

Given the following data points find the mean deviation and the standard deviation along with the measures of central tendency. What is the range?
Display the data…why did you choose what you did for the display?

5, 6, 9, 0, 1, 6, 11, 5

Example: Measures of Variablitiy

What is the data set for this histogram?

Excel

frequency ↓

N = 5

1.2

1

0.8

0.6

0.4

0.2

0

1   2   3   4   5

□ Series1

↑   ↑   ↑   ↑
-2S   -1S   1S   2S
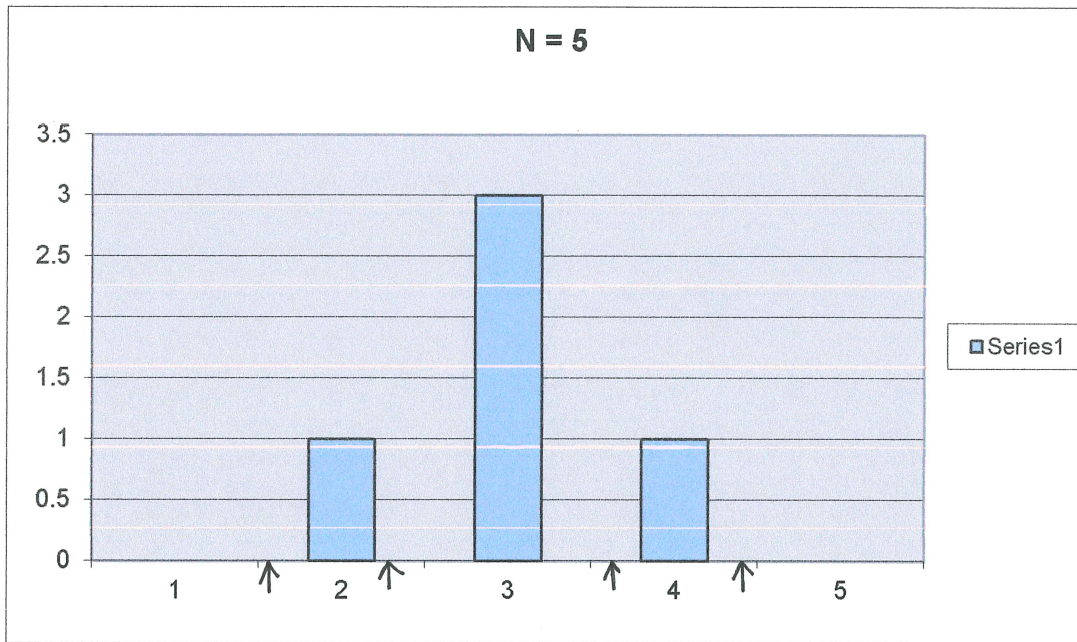
{1, 2, 3, 4, 5}

What is the range? 5 − 1 = 4

What is the mean and standard deviation? How can you get this quickly?

use technology !

Mean is 3 and sd is about 1.58. Mark off two sd on the histogram.

3 + 1.58 = 4.58   plus another sd = 6.16

3 − 1.58 = 1.42   minus another sd = -0.16

all data points are w/in 2S of the mean
both high and low

11

N = 5

Excel

Here's another one:

Range: 4 – 2 = 2

Mean is 3

S = about .71

Mark of two s on the histogram

3.71 and 4.42 along with 2.29 and 1.58

Which of these two is MORE VARIABLE?

The one w/. the bigger SD.!
ie higher variance   1.58 vs. .71

Homework Problem 4

The data in the following table are for the inner diameters of some tubes manufactured by a machine. This table is called a "distribution" because it gives the values and their frequency. Find the mean diameter and the variance for the tubes.

D, inches	frequency
2.0	2
2.2	4
2.3	6
2.8	3
3.0	5

Let's break here and meet up in Video B for Chapter 3

Percentile Rank

Decile   10ths

Quartile   25%

Percentile   singles

A fractile ranking means that a given number of measurements lie below the given measurement and a given number above.

Suppose your child comes home to tell you that she's in the 90th percentile of her class on a particular test. This means that 90% of the children have lower scores or the same score as she does and 10% have higher scores. You do need to be a little careful with these measurements of relative ranking, though. It could be that 91% of the children failed the test and 9% passed. In this scenario, of course, being in the 90% percentile isn't much to brag about. You need absolute measures AND relative measures to evaluate a situation about fractiles.

Deciles divide the measurements into 10ths and quartiles divide the measurements into quarters (breaks at the 25% marks). The median is both a decile and a quartile ranking.

Let's look at quartiles:

Q1 is the median of all measurements less than the median of the data set.

Q3 is the median of all measurements greater than the median of the data set.

25% 25% 25% 25%

MIN   Q1   median   Q3   MAX
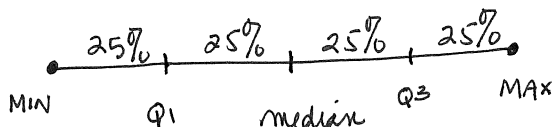
on an interval of real numbers. Not so even w/. real data often

And deciles:

D1 is the measurement such that 90% of the measurements are BIGGER than it.

Example: Measure of Position

The following numbers are weekly lumber production (in million board feet) for a company in Oregon. Find the first quartile and the 10th percentile for the data.

390   406   447   410   370   338   410   320   359   392   315   480

Well the first thing to do is put the numbers in order!

315   320   338   359   370   390   392   406   410   410   447   480
  ↑   ↑   ↑
 316   Q1   391

12 measurements 10percent of 12 is 1.2 so we'll find out what is 1.2 along the way between 315 and 320. 2 tenths of 5, the difference is 1. So 316 is the first decile measurement.

The median is half of the 12 + 1th measurement: the average of measurements 6 and 7 which are NOT 6 and 7 in reality but are 390 and 392…their average is 391 so that is the median. These do not need to be measurements in the data set, folks.

Q1
Now the first quartile is the median of the first 6 measurements which is the average of the third and fourth measurements: half of 338 + 359: 348.5

Not in the book, but handy to know!

Percentage change in a measurement:

The percent change in a measurement is often of interest to managers, doctors, and teachers. It is used as a measure of efficacy.

The calculation is

$$\frac{\text{final - initial}}{\text{initial}}$$

Suppose you have a student who was reading poorly – 15 words a minute. You train the student using your favorite method and test him again to find him reading 27 words a minute.

initial 15
final 27

The percent change is

$$\frac{f - i}{i}$$

$$\frac{27 - 15}{15} \qquad \text{which is 80\%.}$$

You would then report an 80% improvement in speed.

Chapter 3 Homework Problem 5

You've been looking at a sweater in the store but it costs $135 and that's too much. BUT one day you go and check and it's been marked <u>down</u> to $65…what is the percent change?

$- \% \Delta$

Homework Problem 6

A student has been working with a tutor on his math skills. His weekly quiz average was a 65% when he started with the help program.

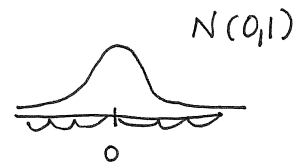His quizzes are 30 points each. During the program his weekly grades are

20, 23, 21, 28, 27, 29

What is the percent change in his average? Would you say that the tutoring helped?
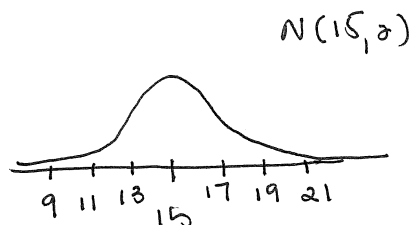
The Empirical Rule   page 71

$N(0,1)$

Given a normal distribution (continuous, symmetric, mound-shaped)
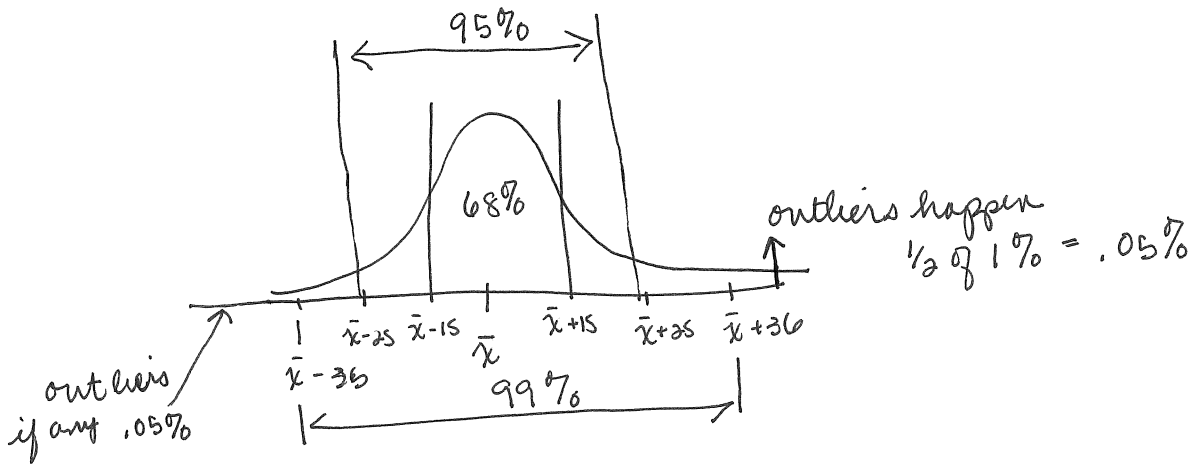
0

68% of the data will lie inside 1 standard deviation from the mean $\mu = 0$
95% of the data will lie inside 2 standard deviations from the mean $\sigma = 1$
99% of the data will lie inside 3 standard deviations from the mean

Let's sketch a normal distribution and then head to the next page for the Empirical Rule:

$N(15, 2)$

9 11 13 15 17 19 21

A Normal Distribution with the Empirical Rule on it
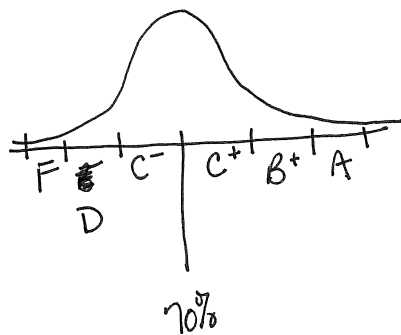

Now for something very IMPORTANT

Z-score – a number that tells you how far a measurement is from the mean.

    Usual, unexceptional data points will be $\pm1-\pm1.5\,s$
        Think C's on the positive end

    Unusual will be $\pm1.2-\pm2.5$

    Rare and outliers will be ±2.5 and up or down

    Think of a grading scheme and standard deviations here: let's put in standard deviations and letter grades:

Here is one of my classes, a listing of the grades on the final…raw data and real
This is a stem-and-leaf diagram.

10	123
09	45779
08	327758
07	459
06	78
05	354

101   102   103   ③
94   95   97 97 99   5 more students
⋮

53   55   54   my F's

How many students were in my class? N = 22

What is the mean and the standard deviation?

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Mean   82.3   S   15.7   excel

Check the empirical rule: +/- 1s   66.6 to 88

+/- 2s   50.9 to 93.7

+/- 3s   45.2 to 99.4   outliers up top. extra credit at work

Discuss the grades at the extremes… 103, 102, 101, and 53, 55, and 54

outliers   not outliers

Which grade is at the 80% percentile? 80% of 22 is the 17.6^{th} measurement…count them out: 97 is 17th and 97 is 18th so .6 along will make it 97.
most students wouldn't want to make 97, the median, a C.

How far is the 85 from the mean in terms of the standard deviation?

Mean = 82.3   SD = 15.7   calculated in Excel

Know by heart formula:

(measurement – mean) divided by standard deviation

(85 – 82.3)/ 15.7 = .17 from the mean   not even 1s out. Quite normal. If you're being tough: it's a C!

Z score example:

If you have 2 students applying for entrance to a G&T program and you have room for only one, which one will you pick based on the following test information?

Gina got a 78 on a test with an average of 72 and a standard deviation of 5.

Mike got an 87 on a test with an average of 85 and standard deviation 1.5.

Gina's zscore: 78-72 divide by 5 = 1.2

Mike's zscore: 87-85 divide by 1.5 = 1.33

Who is the stronger student and how do you know?   *Don't go by gut feelings! have a "defendable" reason*

Summarizing to here

Be able to discuss

the measures of central tendency

- mean
- median
- mode
-

the measures of variability

- range
- variance
- standard deviation
-

and give

- the z score for a measurement.

and

Be able to verify the Empirical Rule by making a dot or bar chart of the data and marking off where each of the standard deviations from the mean are with respect to the data points . ($\pm s$, $\pm 2s$, $\pm 3s$).

Homework Problem 7

The mean salary of the employees at a high school in Missouri is $28, 500 with a standard deviation of $2,100.

Discuss the Empirical Rule and who might fit where on a bar chart of employee salaries.

The state announces a flat raise of $500 per employee for the next year. Find the mean and standard deviation of the new salaries.

Who will benefit the most in a percentage change analysis?

Z score and Range

A rough estimate of the range is the mean $+/-$ 2 standard deviations from the mean. Why is this true?

95%

Could you use 3 sd? What would the difference be? *longer range*

So you can ESTIMATE the standard deviation by taking the range and dividing by 4...let's do this. It's rough, but sometimes you just have to take what you can get!

If the range is 24, the estimate of the sd is 6.

Chapter 3 Popper Question 4

If the range is 16 what is a rough estimate of one SD?

A.   1ish

B.   2ish

C.   3ish

D.   4ish

E.   16exactly

Chapter 3 Popper Question 5

If the mean is 4 and the s is 1.2 what is a rough estimate of the range?

A.   2.8 to 5.2

B.   1.6 to 6.4

C.   0.4 to 7.6

So here we'll end video B and start on the last Chapter 3 video: C.

are sometimes called "box plots".   They use the

Five Number Summary
in a visual way:

Minimum value in the data set
Lower Quartile value $Q1$
Median M
Upper Quartile value $Q3$
Maximum value
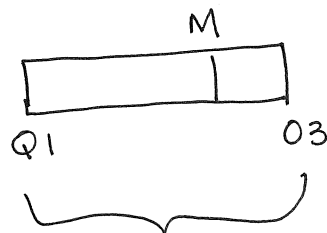
***Graphing Calculator instructions, page 79

Definitions:

Lower Quartile: Q1:   the median of the values below the median
Upper Quartile: Q3:   the median of the values above the median

It is possible to replace the minimum and maximum with prescribed values and have "outliers" marked.

Sketch: horizontal showing the quartile box only, not the full box plot

real data not always
evenly spaced

the middle 50% is here
kind of like $\bar{x} \pm 1s$

Outside the box

IQR: Interquartile Range: is the difference between the upper quartile and the lower quartile. It is where the most "normal" measurements are and we use the IQR to define where is normal and where the outliers are.

$$IQR = Q3 - Q1$$

1.5 times IQR defines the boundaries for normal: $1.5(IQR)$

Illustration box plot with "whiskers" and outliers! All 5 numbers showing!
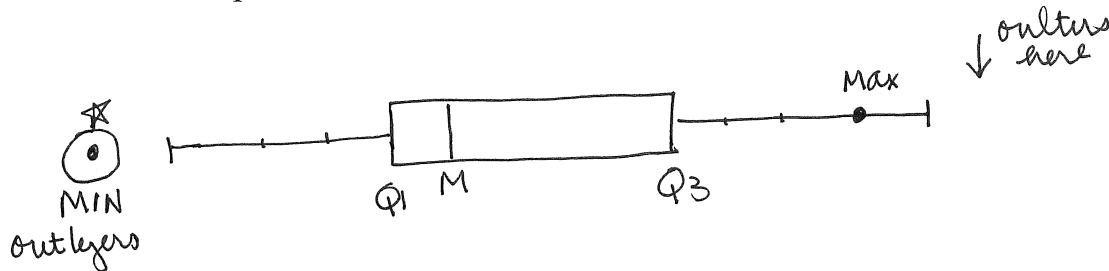
Box plots are often used to compare data sets! It's so easy to see how categories compare with them.

Constructing a box plot with specified "fences" and "outliers"
as opposed to the Five Number Summary only

> Put the data set in numerical order.
> Mark the Five Number Summary right on the list.
> Construct the box with Q1, the median, and Q3
> Find the length of the fences (upper and lower, Qx $\pm$ 1.5(IQR))
> Identify any data points that lie outside the fences and mark them *

Example Box and Whiskers plot:

Here is one of my classes, a listing of the grades on the final…raw data and real
This is a stem-and-leaf diagram. Same one from video B.

```
10 | 123
09 | 45779
08 | 327758
07 | 459
06 | 78
05 | 354
```

$N = 22$

median $\dfrac{n+1}{2} = \dfrac{23}{2} = \dfrac{11+12}{2} = 11.5$

$\dfrac{87+87}{2} = 87$

$Q1$: middle of 11 on bottom
#6 (74)

$Q3$ middle of 11 on top
#6 97

What is the Five Number Summary? The IQR?

MAX 103

Q3 97

M 87

Q1 74

MIN 53

What is the estimated SD? And the estimated z-score for 67?
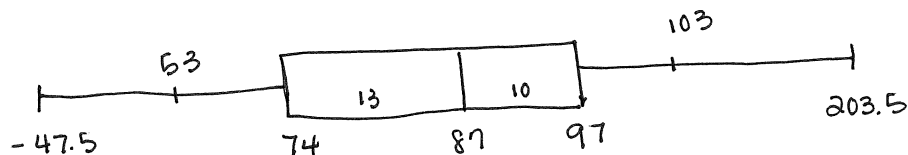
Mean is 82.3   real SD is 15.7

Calculate the whisker length and what numbers they stretch to:
Whiskers are sometimes called "fences"

$Q3 - Q1$ is $97 - 74 = 67$ $1.5(67) = 100.5$

max + 100.5   min − 100.5
103 + 100.5   203.5 upper fence $53 - 100.5 = -47.5$

Sketch the box and whisker plot! Were there any outliers?
How do you know they're outliers? Use the next page for this

no outliers

BW continued

oops! on other page
moving on

Chapter 3 Popper Question 6

The 5 number summary and the IQR are what you need to construct a Box and Whiskers plot.

A.   True
B.   False

Chapter 3 Homework Problem 8

Make a box and whiskers plot with the following data set:

{0, 1, 1, 3, 6, 8, 8, 9, 10, 13}.

Show the 5 Number Summary, the calculations for the whiskers (the "fences"). Circle any outliers, show the 5 numbers on the plot.

Ok now wrapping up Chapter Three

The popper has 6 questions.
There are 7 homework problems in this script PLUS two problems to do from the book:

Chapter 3 Homework from the book:

Page 83 number 9, 13

Join me in the Chapter 4 video next up!