Introduction to Statistics I

Instructor: Jodin Morey moreyj@lemoyne.edu

Previous Lecture

- Defined: Mean, Median
- Dist Shape: Symmetric/Skewed Data
- Mean vs. Median

Topic 9: Measures of Spread

Now that we have measures of center for distributions, let's explore measures of **spread** (how is the data spread out?).

Example: SCOTUS Tenure by Party of Nominating President. Below is a sampJle of 23 recent ustices.

Justice	Tenure	Justice	Tenure
Ruth Bader Ginsburg (D)	27	Arthur Goldberg (D)	3
David Souter (R)	19	Byron White (D)	31
Anthony Kennedy (R)	30	Potter Stewart (R)	23
Antonin Scalia (R)	29	Charles Evans Whittaker (R)	5
William Rehnquist (R)	32	William Brennan (R)	34
Sandra Day O'Connor (R)	24	John Marshall Harlan (R)	17
John Paul Stevens (R)	34	Earl Warren (R)	16
Lewis Powell (R)	15	Sherman Minton (D)	7
Harry Blackmun (R)	24	Tom C. Clark (D)	18
Warren Burger (R)	17	Fred Vinson (D)	8
Thurgood Marshall (D)	24	Harold Hitz Burton (D)	13
Abe Fortas (D)	4		

Dem Nom'd:

27	24	4	3	31	7	18	8	13

GOP Nom'd:

19	30	29	32	24	34	15	24	17	23	5	34	17	16

Recall Center

Dem Sorted:

3	4	7	8	13	18	24	27	31

Median/Mean ??

Median is 13, Mean is 15.

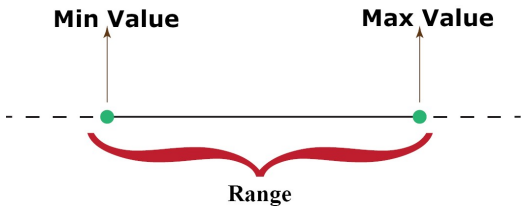
GOP Sorted:

5	15	16	17	17	19	23	24	24	29	30	32	34	34

Median is 23.5, Mean is 22.8.

Measures of Spread: Range, Interquartile Range (IQR), Standard Deviation (SD, *s*).

Range: Max – Min.



Dems Sorted:

3	4	7	8	13	18	24	27	31

Min is 3, Max is 31, Range is 31 – 3 = 28.

GOP Sorted:

5	15	16	17	17	19	23	24	24	29	30	32	34	34

Min is 5, Max is 34, Range is 34 – 5 = 29.

We sometimes *say*, "Dem nominated justices' tenures range from 3 to 31".

But in stats, *the range* is always a single number (max minus min), the length (not location) of the data.

So the range of tenure of Dem. nomt'd justices is 28.

Interquartile Range (IQR)

IQR is how much space the middle data takes up, **not** its location.

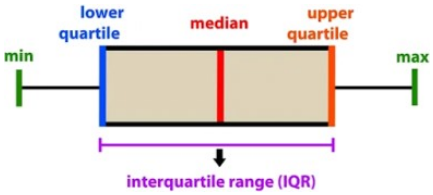
IQR is range of (approx.) the middle 50% of data.

IQR excludes the extremes (so, not sensitive to outliers).

How to Calculate?

Upper Quartile (UQ): Value such that ≈ ¼ of data is above it, and ¾ is below.

Lower Quartile (LQ): Value such that ≈ ¾ of data is above it, and ¼ is below.



So, $IQR = UQ - LQ$.

Back to Example (how to calculate UQ and LQ):

Recall Dems:

3	4	7	8	13	18	24	27	31

Median is 13, so upper half is

18	24	27	31

.

UQ **is median of upper half**: $UQ = \frac{24+27}{2} = 25.5$.

Lower half is

3	4	7	8

.

LQ **is median of lower half**: $LQ = \frac{4+7}{2} = 5.5$.

So, $IQR = 25.5 - 5.5 = 20$ yrs.

Range of the middle 50% of data (*IQR*) is 20 years.

Recall GOP:

5	15	16	17	17	19	23	24	24	29	30	32	34	34

Median is 23.5, so upper half is

24	24	29	30	32	34	34

.

UQ is median of upper half: $UQ = 30$.

Lower half is

5	15	16	17	17	19	23

.

LQ is median of lower half: $LQ = 17$.

So $IQR = 13$.

Recall *IQR* is range of the middle 50% of the data.

Dems: range of middle 50% is 20 years.

GOP: range of middle 50% is 13 years.

So Dem nominees are more varied in how long they stay on the court.

Activity: 9-1

IQR is a measure of spread related to the median.

There's also a measure of spread related to the population mean (μ): it's called **Standard Deviation** (**SD**, s).

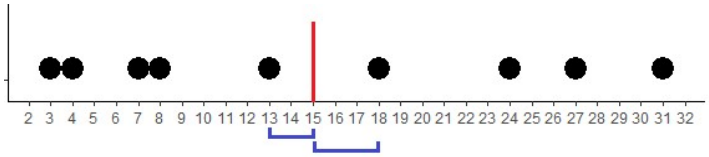
Standard Deviation (SD, s)

Roughly, SD is "average distance (or deviation) of data from the mean."

SD is ideally related to the pop. mean μ. But we usually don't know μ!

All we have is a sample, so we'll have to settle on using the sample mean ($\bar{x}$).

Dem nomt'd tenures: We'll begin w/average distance each data pt is from sample mean $\bar{x}$ (marked in red).



Tenures of justices nomt'd by Dem. presidents

If pts are:

- close and clustered around mean ⇒ Low SD,
- spread out, and far from mean ⇒ High SD.

Dem nomt'd tenures

Data	Data − Mean ($\bar{x}$ = 15)
3	3 − 15 = −12
4	4 − 15 = −11
7	7 − 15 = −8
8	8 − 15 = −7
13	13 − 15 = −2
18	18 − 15 = 3
24	24 − 15 = 9
27	27 − 15 = 12
31	31 − 15 = 16

Average distance from mean?

$-12 + (-11) + (-8) + (-7) + (-2) + 3 + 9 + 12 + 16 = -40 + 40 = 0.$ (!?!)

Sum of deviations from a mean is **always** zero.

So, we want to remove the negative signs to consider **distances** instead.

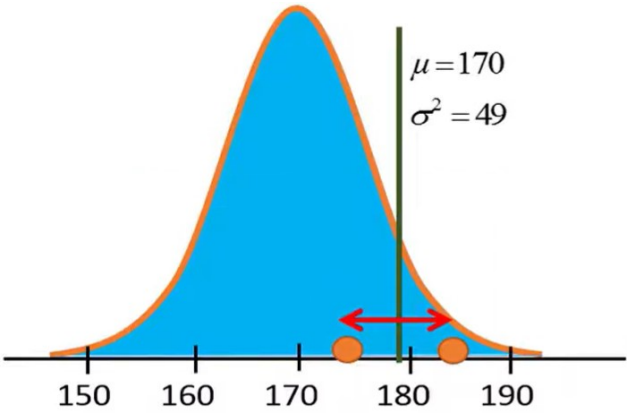
Squared Distances from Sample Mean

Data	Data – Mean ($\bar{x} = 15$)	(Data – Mean)^2
3	$3 - 15 = -12$	$(-12)^2 = 144$
4	$4 - 15 = -11$	$(-11)^2 = 121$
7	$7 - 15 = -8$	$(-8)^2 = 64$
8	$8 - 15 = -7$	$(-7)^2 = 49$
13	$13 - 15 = -2$	$(-2)^2 = 4$
18	$18 - 15 = 3$	$3^2 = 9$
24	$24 - 15 = 9$	$9^2 = 81$
27	$27 - 15 = 12$	$12^2 = 144$
31	$31 - 15 = 16$	$16^2 = 256$

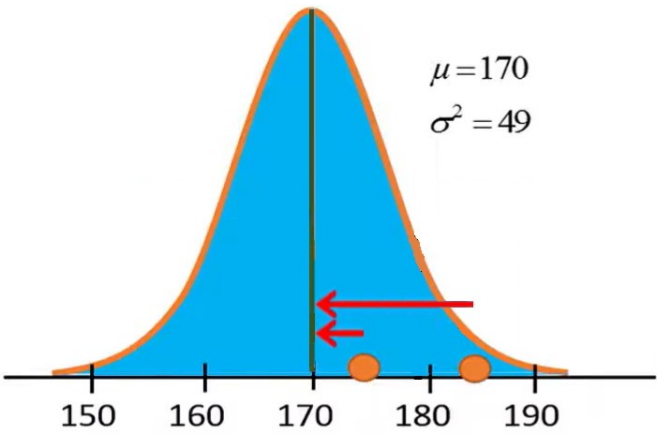
$144 + 121 + 64 + 49 + 4 + 9 + 81 + 144 + 256 = 872$ (makes more sense)

Average squared distance to $\bar{x}$: $\frac{872}{n} = \frac{872}{9} \approx 96.9.$

How does this compare to using population mean μ?



Distance to sample mean $\bar{x}$: underestimate



Distance to population mean μ: actual

So our calculations were underestimates. How should we adjust them to be more accurate?

Adjustment: Approximate average of *squared distances* from μ is: $\frac{872}{n-1} = \frac{872}{8} = 109$.

See vid on Canvas for more justification.

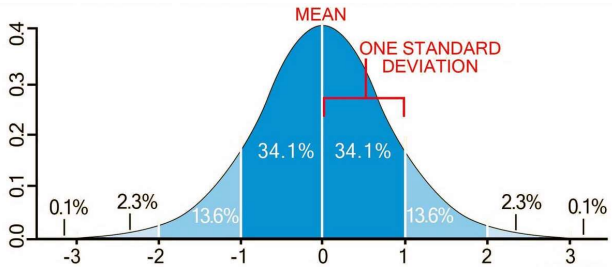
So, approximate average of *distances* from μ should be
square root of average squared distance: $\sqrt{109} \approx 10.4$.

This number is called the **Standard Deviation** (**SD** or s).

SD is approx average distance of each pt from the pop. mean μ: $s = \sqrt{\frac{\text{sum the (data pt} - \bar{x})^2}{n-1}}$.

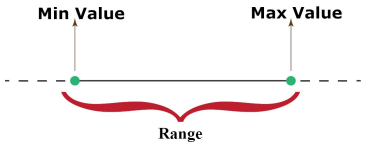
Steps to SE:

- Calculate sample mean $\bar{x}$.
- Subtract $\bar{x}$ from every data pt (deviations).
- Square the deviations.
- Sum the squared deviations, divide by $n - 1$.
- Take square root (this retrieves the unit: *yrs* instead of yrs^2 for this example).

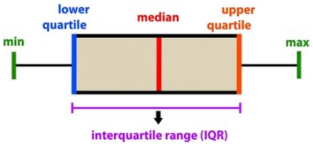


Three main measures of spread:

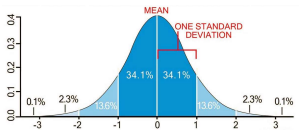
- **Range**: Max – Min. Quick to calculate, very influenced by outliers.



- **IQR**: Range of middle 50%, resistant to outliers.



- **SD**: Average (approx) distance of pts from pop. mean μ. Influenced by outliers.

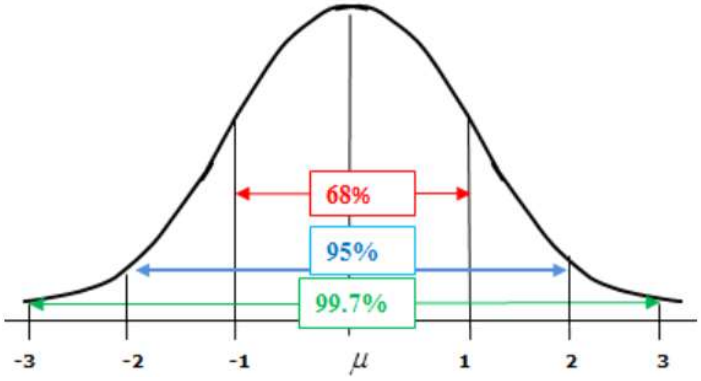


SD is most commonly used.

Activity: 9-2

For mound shaped, symmetric distributions, SD helps predict behavior of data. How? ...

The Empirical Rule



# of SDs above/below μ. (1,2,3)

Empirical Rule:

- 68% of data pts are within 1 SD of μ.
- 95% are within 2 SDs of μ.
- Nearly all are within 3 SDs of μ.

Applies only to data which is approx. mound shaped!!

Example: An IQ test has mean of 100 and SD of 15 pts.
If we take a random sample, we'd expect to find 68% of subjects w/IQs between what values?

$100 - 15 = 85$

$100 + 15 = 115$

We'd expect to find 68% of subjects w/IQs between 85 and 115.

What range would we expect to find 95% of students?

Empirical rule allowed us to identify "unusual-ness" of a data pt by how far it is from μ.

- Pts within 1 SD of mean are *common*.
- Pts between 1 & 2 SDs of mean are *relatively common*.
- Pts between 2 & 3 SDs of mean are *rare*.
- Pts > 3 SDs from mean are *very rare*.

---

z - Score

Calculates how many SDs a pt is (approx) from μ. (how far is data pt from avg value?)

$z = \frac{x-\bar{x}}{s}$, where x is data pt, $\bar{x}$ is sample mean, and s is SD.

Also applies only to data which is approx. mound shaped!!

Example (z-score): A college accepts two different math placement exams.

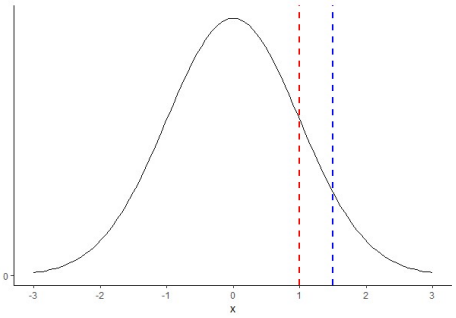
- Exam A scores students from 0 – 20, has $\bar{x}$ of 10 and SD of 4.
- Exam B scores students from 0 – 100, has $\bar{x}$ of 70 with SD of 10.

Arthur scores 16 on Exam A. Bertha scores 80 on Exam B.

Who did better?

Arthur z-score: $z = \frac{16-10}{4} = 1.5$.

Bertha z-score: $z = \frac{80-70}{10} = 1.0$.



Bertha (red), Arthur (blue)

Example (neg z-score): A college accepts two different math placement exams.

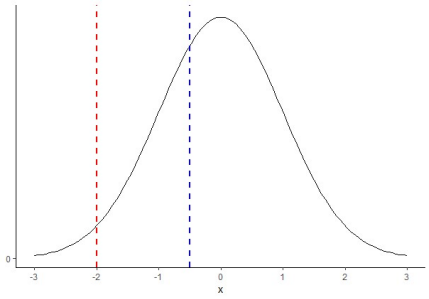
- Exam A scores students from 0 – 20, has mean of 10 and SD of 4.
- Exam B scores students from 0 – 100, has mean of 70 and SD of 10.

Dolly scores 8 on Exam A. Cristobal scores 50 on Exam B.

Who did worse?

Dolly z-score: $z = \frac{8-10}{4} = -0.5$.

Cristobal z-score: $z = \frac{50-70}{10} = -2.0$.



Cristobal (red), Dolly (blue)

Activity: **9-4**

What did we learn?

- Measures of Spread: Range, IQR, SD
- Empirical Rule: 1SD: 68%, 2SD: 95%, 3SD: Almost All
- z-score: $\frac{x-\bar{x}}{s}$