Introduction to Statistics I

Textbook: Elementary Statistics (4th Edition, by Navidi and Monk), and Workshop Statistics (4th Edition by Rossman and Chance).

Previous Lecture

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

§3.2: Measures of Spread

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

Example: Supreme Court (SCOTUS) Tenure by Party of Nominating President. Below is a sample of 23 recent retired justices.

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

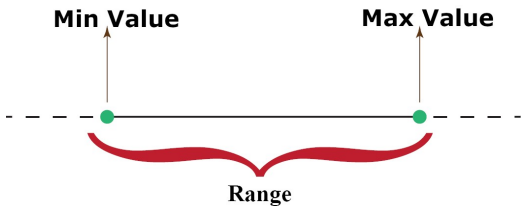
GOP Nom'd:

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

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

Simplest measure of spread is range:

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 tenures range from 3 to 31".

But in stats, *the range* is always a *single number* (max minus min). It's 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 length the *middle* data takes up (**not** its location).

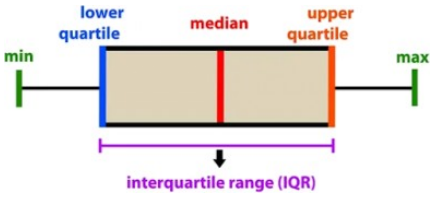
So, IQR is the range of the middle 50% of data (approx.).

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

How to Calculate?

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

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



So, IQR is $UQ - LQ$.

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

Recall Dems:

27	24	4	3	31	7	18	8	13

First, how do we calculate median (the middle of the data)?

Dems Sorted:

3	4	7	8	13	18	24	27	31

Median is 13, so upper half is $\boxed{18 \mid 24 \mid 27 \mid 31}$.

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

Lower half is $\boxed{3 \mid 4 \mid 7 \mid 8}$.

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

Thus, IQR is $UQ - LQ = 25.5 - 5.5 = 20$ yrs.

Thus, the range of the middle 50% of data (IQR) is 20 years.

Recall GOP (sorted): $\boxed{5 \mid 15 \mid 16 \mid 17 \mid 17 \mid 19 \mid 23 \mid 24 \mid 24 \mid 29 \mid 30 \mid 32 \mid 34 \mid 34}$ Median?

Median is 23.5, so upper half is $\boxed{24 \mid 24 \mid 29 \mid 30 \mid 32 \mid 34 \mid 34}$.

UQ is median (middle) of upper half: So, UQ is 30.

Lower half is $\boxed{5 \mid 15 \mid 16 \mid 17 \mid 17 \mid 19 \mid 23}$.

LQ is median of lower half: So, LQ is 17.

Thus, IQR is 13.

Dems: range of middle 50% is 20 years.

GOP: range of middle 50% is 13 years.

So Dem nominees are *more varied* (spread out) in how long they stay on the court.

Activity: 3.2a

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).

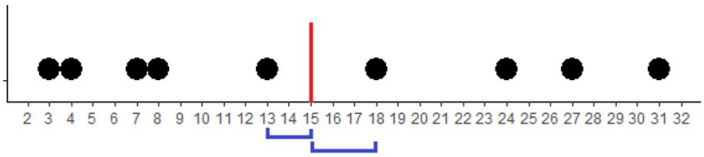
Variance/Standard Deviation (SD, s)

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

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

If we only have a sample, we'll have to approximate using the sample mean ($\bar{x}$).

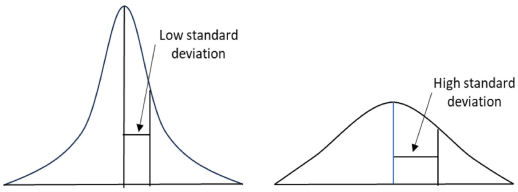
Dem. nomt'd: We'll begin by calculating the average distance each data pt is from the sample mean $\bar{x} = 15$ (marked in red below).



Tenures of justices nomt'd by Dem. presidents

If pts are:

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



Calculating deviations from data to mean

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?

$$\frac{-12+(-11)+(-8)+(-7)+(-2)+3+9+12+16}{9} = \frac{-40+40}{9} = 0. \qquad (!?!)$$

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

So, we want positive values in order to consider **distances** instead.

Squared Deviations 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	$31 - 15 = 16$	$16^2 = 256$

$\frac{\text{Sum of squared dist's}}{n} = \frac{144+121+64+49+4+9+81+144+256}{9} = \frac{872}{9}$ (makes more sense)

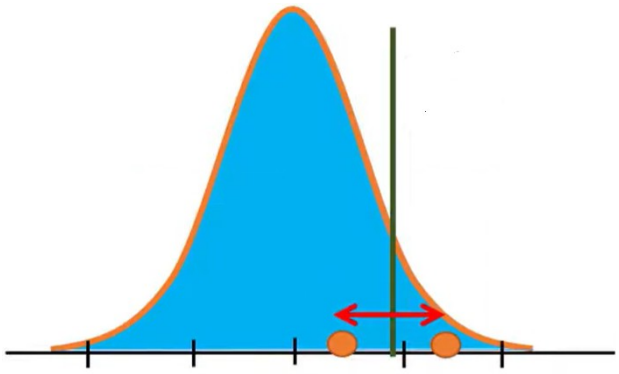
Average *squared* distance to $\bar{x}$: ≈ 96.9.

If we had been looking at *all* the **population** data, and it's distance to the **population** mean μ, we'd call this the **Population Variance**, and it's square root the **Population SD**.

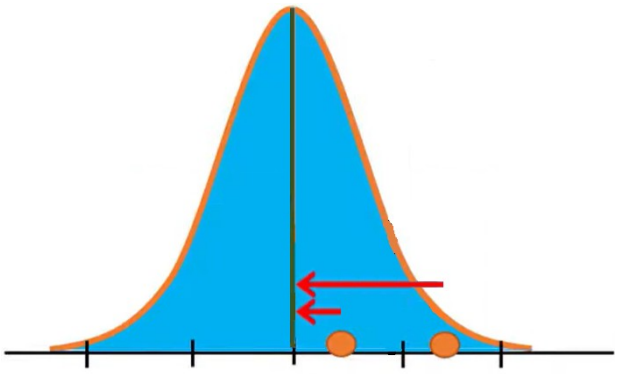
But since this is the squared distance of **sample** data to the **sample** mean $\bar{x}$, it is called the **Sample Variance**, and it's square root the **Sample SD**.

But recall our goal is to estimate the **Population Variance/SD**.

So, how does the Sample Variance differ from the Population Variance?



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



Distance to population mean μ: actual

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

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

This is our best estimate of the **Population Variance**: $\frac{\text{sum the (data pt} - \bar{x})^2}{n-1}$.

See vid on Canvas for more justification of "$n-1$".

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

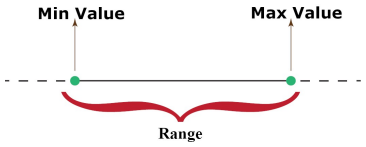
This is our best estimate of the **Population Standard Deviation** (**SD** or s): $s = \sqrt{\frac{\text{sum the (data pt} - \bar{x})^2}{n-1}}$.

Steps to Population SD:

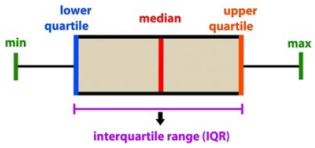
- Calculate sample mean $\bar{x}$.
- Subtract $\bar{x}$ from every data pt (deviations).
- Square the deviations.
- Sum the squared deviations, divide by $n-1$.
- Take square root.

Three main measures of spread:

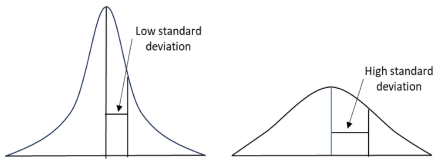
- **Range**: Max – Min. Quick to calculate, very influenced by outliers.



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



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

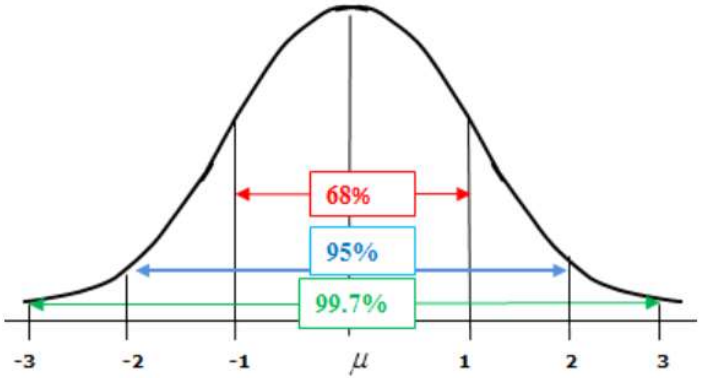


SD is most commonly used.

(Activity 3.2b: On Canvas, insufficient class time to do here. Good practice, though.)

For mound shaped distr's, SD helps predict behavior of data. How? ...

The Empirical Rule



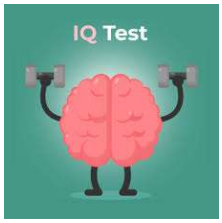
# of SDs above/below μ. (1,2,3)

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?

Activity: 3.2c

What did we learn?

- Measures of Spread: Range, IQR, SD
- Empirical Rule - 1SD: 68%, 2SD: 95%, 3SD: Almost All