3.2 Measures of Spread

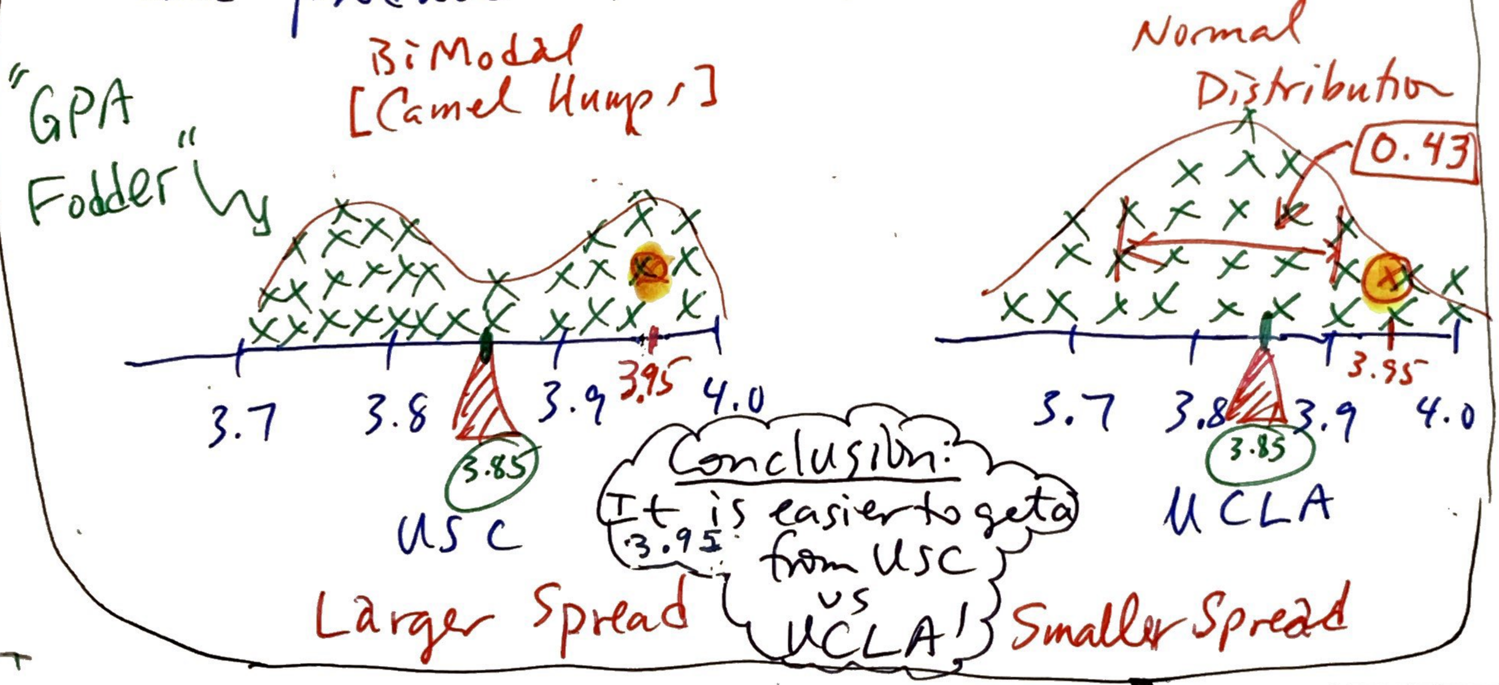
Scenario: You Are on the Yale Med. School acceptance panel. you have selected all of the next years med students except one slot is left.

There are two outstanding matched students to compete for that slot. Both have a 3.95 GPA's, Both schools have mean 3.85 GPA one from USC and one from UCLA.

Q: What more can you look at to break the tie?

A: Having had statistics you know that you need to examine the GPA distribution from the two schools.

The assistant collects dozens of GPAs from current and recent UCLA & USC applicants and produces these distributions



Ex Ave. monthly temperatures

	Jan	F	M	A	M	Ju	Jl	A	S	O	N	Dec
San Fran	51	54	55	56	58	60	60	61	63	62	58	52
Kansas City	30	35	44	57	66	75	79	78	70	59	45	35

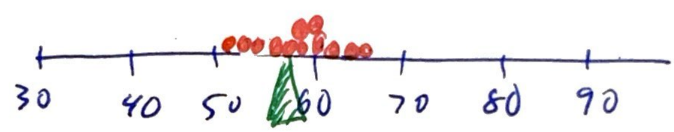
- mean value :

$$SF = \frac{51+54+55+\cdots+58+52}{12} = \boxed{57.5}$$

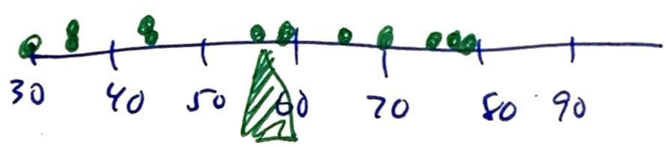
$$KC = \frac{30+35+\cdots+45+35}{12} = \boxed{56.1}$$

- Dot Plots

SF



KC



Range

Range = max value of a data set
− min value of a data set

S.F : Range = 63 − 51 = $\boxed{12\,°F}$ spread

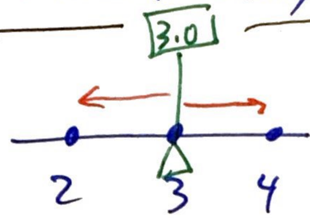
K.C. Range = 79 − 30 = $\boxed{49\,°F}$ spread

Unless there is an outlier the range is a quick measure of the spread of a group of data.

* Variance

For this measure of spread we look at how far each datapoint is from the mean. So as to not have data points to the left of the mean cancel out data point equally to the right of the mean, we square the sum the differen

EX



$\bar{x} = \frac{2+3+4}{3} = 3.0$

$$\begin{aligned}\sum \Delta x &= (\bar{x}-2) + (\bar{x}-3) + (\bar{x}-4)\\ &= (3-2) + (3-3) + (3-4)\\ &= 1 + 0 + -1\\ &= \boxed{0} \text{ No spread!! ???}\end{aligned}$$

Now lets square the Δx:

$$\begin{aligned}\sum \Delta x^2 &= (3-2)^2 + (3-3)^2 + (3-4)^2\\ &= 1^2 + 0^2 + 1^2\\ &= \boxed{2} \text{ ok, better, } \neq 0\end{aligned}$$

Now lets average these

$$\sum \left(\frac{\Delta x^2}{N}\right) = \frac{1^2 + 0^2 + 1^2}{3} = \boxed{\frac{2}{3}} \text{ average spread of data}$$

- We call this Variance.

$$\boxed{\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}}$$

For populations (census! every data is polled)

Ex Find the variance of the following:

2, 2, 3, 4, 5, 7

(i) $\mu = \dfrac{2+2+3+4+5+7}{6} = \underline{3.83\bar{3}}$

TI-30Xa : 2 $\boxed{\Sigma+}$ 2 $\boxed{\Sigma+}$ 3 $\boxed{\Sigma+}$ 4 $\boxed{\Sigma+}$ 5 $\boxed{\Sigma+}$ 7 $\boxed{\Sigma+}$

μ : $\boxed{2^{nd}}$ $\boxed{\bar{x}}$ 3.833

(ii) To caclulate the variance we use a table

x	$x-\mu$	$(x-\mu)^2$
2	$2-3.83=-1.83$	3.35
2	$2-3.83=-1.83$	3.35
3	$3-3.83=-0.83$	0.69
4	$4-3.83=+0.17$	0.29
5	$5-3.83=1.17$	1.37
7	$7-3.83=3.17$	10.05

sum

$\sum \Delta x^2$ $\boxed{19.10}$

capital sigma "sum"

$\sum \dfrac{\Delta x^2}{N} = \dfrac{19.10}{6}$

variance

$\boxed{\sigma^2 = 3.18}$

baby "sigma"

$\sigma = \sqrt{3.18} = 1.78 \approx 1.8$

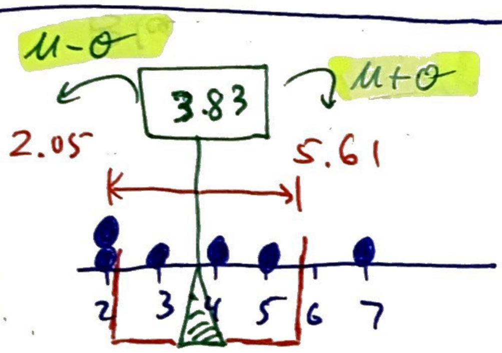
Standard Deviation $= \sqrt{\sigma^2}$

• On The TI-30Xa "σ_{xn}"

$\boxed{2^{nd}}$ $\boxed{\sigma_{xn}}$ $\boxed{x^2}$ $\rightarrow 1.77^2$

$\boxed{\sigma^2 = 3.14}$

?



$$\begin{array}{r} 3.83 \\ -1.78 \\ \hline 2.05 \end{array} \qquad \begin{array}{r} 3.83 \\ +1.78 \\ \hline 5.61 \end{array}$$

Notes: Descriptive stats on populations

μ = mean

σ^2 = variance

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Descriptive statistics on samples from population

$\bar{x}$ = mean

s^2 = variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

Same proceedupe, same results

* Small data group sizes:

- For small data sets statisticians prefere to divide by N-1 vs. N.

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N-1} \qquad s^2 = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

- on the TI-30Xa you can choose which eqn

$\boxed{2^{nd}}$ $\boxed{\sigma_{xn}}$ or $\boxed{2^{nd}}$ $\boxed{\sigma_{xn-1}}$

- To convert from N-1 versin to N version {Stadisk only uses σ_{xn-1}}

$$\boxed{\sigma_{xn}^2 = \sigma_{xn-1}^2 \cdot \left(\frac{N-1}{N}\right)}$$

((Apply this to Variance NOT S.D: ~~to get the~~ Std. Deviation))

Ex Find the Variance and Std. Dev for the S.F and K.C. Temperature data

6

SF

- calculator :

51 Σ+
54 Σ+
55 Σ+
⋮
58 Σ+
52 Σ+

2nd $\boxed{\bar{x}}$ = 57.5

2nd σ_{xn} = 3.752

variance $\boxed{x^2}$ = $\boxed{14.08}$

÷N

- Stat Disk : $\boxed{S^2 = 15.36}$ ÷N-1

$$\boxed{\sigma^2 = (S^2)(N-1)/N}$$

Convert sample variance to census variance

⇒ Data Editor

Col 1
51
54
⋮
58
52

Data → Explore Data

$\bar{x}$ =

S^2 =

range =

median =

KC

- calculator :

30 $\boxed{\Sigma+}$
35 $\boxed{\Sigma+}$
⋮
45 $\boxed{\Sigma+}$
35 $\boxed{\Sigma+}$

2nd $\boxed{\bar{x}}$ = 56.08

2nd $\boxed{\sigma_{xn}}$ = 17.08

$\boxed{\text{Variance} = 291.91}$

NOTE : We skip the part in the text where it is shown how to read a Histogram and estimate σ.

⊛ Estimating unimodal and Symmetric Data distributions

- **Emperical Rule**

When a data set's histogram (dotplot) shows that a quantity has a unimodal and symmetric shape {A Bell Curve} we can estimate the number of data points that lie in certain boundaries.

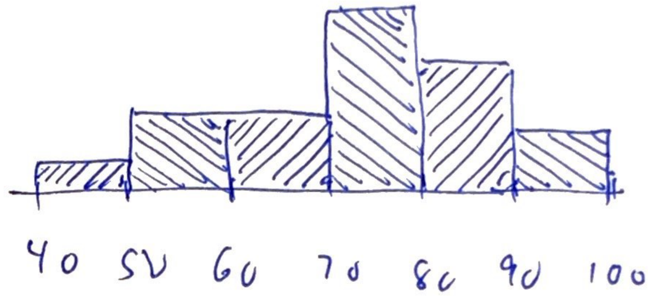
Emperical Rule of Thumb.

- 68% of the data will lie between $\bar{x} - s$ to $\bar{x} + s$
- 95% of the data will lie between $\bar{x} - 2s$ to $\bar{x} + 2s$
- 99.5% of the data lies between $\bar{x} - 3s$ to $\bar{x} + 3s$

Def: **Usual Data** falls between $\bar{x} \pm 2s$

Unusual Data falls outside of $\bar{x} \pm 2s$

Ex The histogram of 200 History Student's final Exam is shown below



$\bar{x} = 75$

$s = 10$

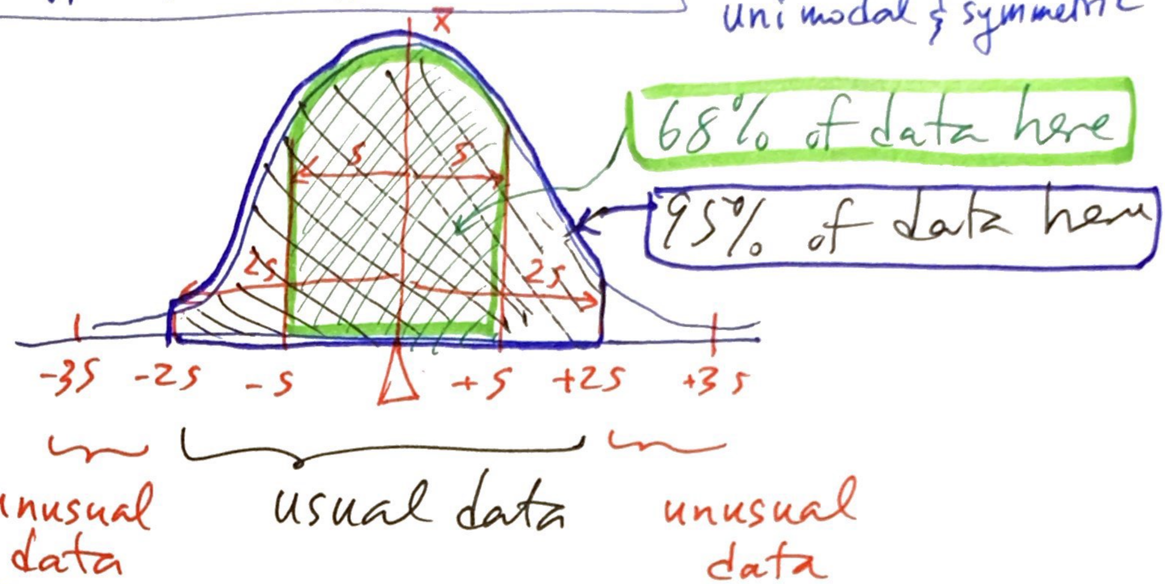
Since the shape is roughly unimodal & sym we can apply the $\boxed{\text{emperical rule}}$ of

- 68% of Students's scores fell between
  $\bar{x} - s$ to $\bar{x} + s$
  $75 - 10$ to $75 + 10$
  ie $\boxed{65 \text{ to } 85 \text{ points}}$

- 95% of Students had scores between
  $\bar{x} - 2s$ to $\bar{x} + 2s$
  $75 - 2(10)$ to $75 + 2(10)$
  ie $75 - 20$ to $75 + 20$
  ie. $\boxed{55 \text{ to } 95}$

- Scores below 55 and above 95 are considered unusual

Summary of Rule of Thumbs for "Bell Curves"
unimodal & symmetric



Coefficient of variation

The coefficient of variation is a scaling of the standard deviation by the mean

I.E. How many std. deviations fit within the mean

$$\boxed{CV \equiv \frac{\sigma}{\mu}} \quad \boxed{CV \equiv \frac{s}{\bar{x}}}$$

- CV is dimensionless
- CV can compare two parameters within one pop.

Ex Rainfall vs. Temperature in San Diego

Rainfall in SD

$\mu = 10.5$ in/yr

$\sigma = 1.2$ in/yr

$\boxed{CV = \frac{1.2}{10.5} = 0.11}$

Temperature in S.D.

$\mu = 51.5$°F

$\sigma = 3.75$°F

$\boxed{CV = \frac{3.75}{51.5} = 0.19}$

Analysis: Rainfall variation ~~matches temp~~ variation

CV perhaps is less effective when comparing 2 pops.

EX Temps in S.F. vs. Temps in K.C.

Temps in S.F.	Temps in K.C.
$\mu = 57.5°F$	$\mu = 56.08$
$\sigma = 3.75°F$	$\sigma = 17.08$
$C.V. = \frac{3.75}{57.5}$	$C.V. = \frac{17.08}{56.08}$
$\boxed{CV = 0.065}$	$\boxed{CV = 0.305}$
narrow dist	wide dist

narrow dist vs wide dist

Summary:

Distribution	Prefered measure of Center	Prefered measure of Spread
• Sym & Unimodal	mean	std. Deviation
• Skewed but Unimodal	median	Inter Quartile Range*
• Scatterd about	mode	Range

* Covered in Section 3.3