Topic: Exploratory Data Analysis (EDA)

Measures of Variability - Part A

School of Mathematics and Applied Statistics

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Measuring Variability: Motivating Example

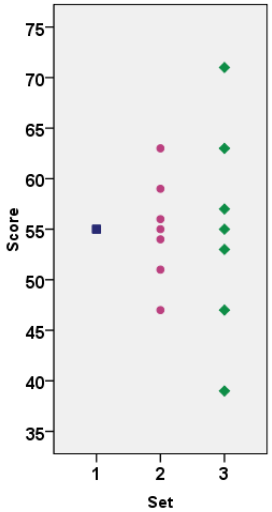
Consider the following data sets: $n = 7$

		Total
Data set 1:	55, 55, 55, 55, 55, 55, 55	385
Data set 2:	47, 51, 54, 55, 56, 59, 63	385
Data set 3:	39, 47, 53, 55, 57, 63, 71	385

For each data set

- Median = 55 = Q_2
- Mean = $385/7 = 55 = \bar{x}$

But the spread of the scores vary.



How do we Measure Variability?

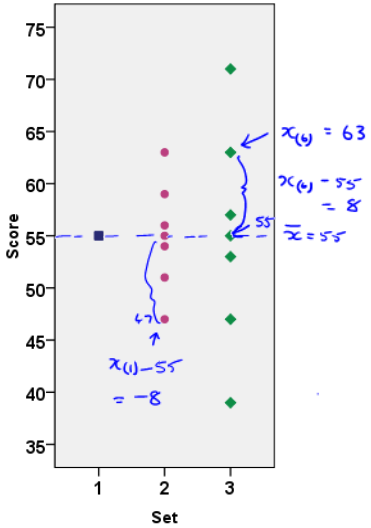
Variability (spread) can be measured by:

- **Variance σ^2 or s^2**
  - uses all data values but is inflated by outliers
- **Standard deviation σ or s**
  - uses all data values but is inflated by outliers
- **Range** $=$ maximum $-$ minimum = $x_{(n)} - x_{(1)}$
  - unreliable measure, depends on extreme values
- **Interquartile range**: $IQR = Q_3 - Q_1$
  - spans middle 50% of data,
  - unaffected by outliers, ignores variation in tails

Variance and Standard Deviation

- The mean is used as a reference point. $\bar{x} = 55$
- Consider the deviation from the ith point to the mean: $x_i - \overline{x}$
- Deviations may be positive or negative
- **Variance** is based on the squared deviations.
- We can also describe the difference in spread using a notion of average distance from the mean.
- This measure of variability is called the standard deviation.

- sum of deviations = 0.



Variance

Variance is based on the squared distances of individual data points from the mean.

- **Population Variance**

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- **Sample Variance**

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

- measurement is in squared units
- is never negative, and
- is only zero when all data values are identical
- s^2 is an unbiased estimator of σ^2.

Standard Deviation

Standard deviation σ or s

- is the square root of the variance
- is measured in same units of measurement as data

- **Population standard deviation**

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

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

Sample standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$s = \sqrt{s^2}$$

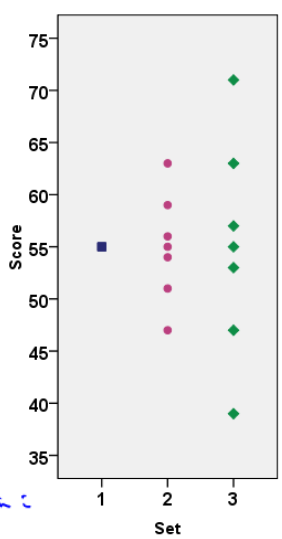
- **On calculator,** enter data then use: STAT mode

pop sd. | σ_n *or* $x\sigma_n$ (σ_x) or σ_{n-1} *or* $x\sigma_{n-1}$ (s_x) *or* s_{n-1} *or* s key | sample sd.

Example

- Data set 1: 55, 55, 55, 55, 55, 55, 55
- Data set 2: 47, 51, 54, 55, 56, 59, 63
- Data set 3: 39, 47, 53, 55, 57, 63, 71

	Median	Mean	SD
Set 1	55	55	
Set 2	55	55	
Set 3	55	55	



Discuss: What is the SD for Set 1? Why?

$s_1 = 0$ (5)

$x_i - \bar{x} = 0$

$x_i = 55$ for each i

$\bar{x} = 55$

Activity: Calculate Variance and SD

Exercise: What is the sample variance and standard deviation for Set 2 and Set 3? Calculate s^2 and s:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \qquad s = \sqrt{s^2}$$

Data Set 2 $\overline{x} = 55$

47, 51, 54, 55, 56, 59, 63

x_i	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
47	−8	64
51	−4	16
54	−1	1
55	0	0
56	1	1
59	4	16
63	8	64
	0	162 $= \sum_{i=1}^{7}(x_i - \overline{x})^2$

$$s_2^2 = \frac{1}{7-1} \times 162$$

$$= 27$$

$$s = \sqrt{27}$$

$$= 5.196\ldots$$

Data Set 3

39, 47, 53, 55, 57, 63, 71

Activity: Calculate Variance and SD cont.

Set 3.

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
39	-16	256
47	-8	64
53	-2	4
55	0	0
57	2	4
63	8	64
71	16	256
		648

$= \sum_{i=1}^{7} (x_i - \bar{x})^2$

$s_3^2 = \frac{1}{7-1} \times 648$

$= 108$

$s_3 = \sqrt{108}$

$= 10.392.$

	Set 1	Set 2	Set 3
s	0	5.196	10.392

In R: Calculate mean, variance and sd

```
> Set2 <- c(47, 51, 54, 55, 56, 59, 63)
> mean(Set2)
[1] 55
> var(Set2)
[1] 27
> sd(Set2)
[1] 5.196152
```

```
> Set3 <- c(39, 47, 53, 55, 57, 63, 71)
> mean(Set3)
[1] 55
> var(Set3)
[1] 108
> sd(Set3)
[1] 10.3923
```