



Let's learn about

1.Histograms

2.Measure of Central Tendency

3.Measure of Dispersion

4.Percentiles and Quantiles

5. 5 Number Summary





# Histograms (Continuous values)

Let's see how to plot histograms.

**ages** = {10, 20, 25, 30, 35}

- first we have to sort the data
- define the bins (no of groups)
- find bins size (size of bins)

min = 10      max = 35      bins = 10

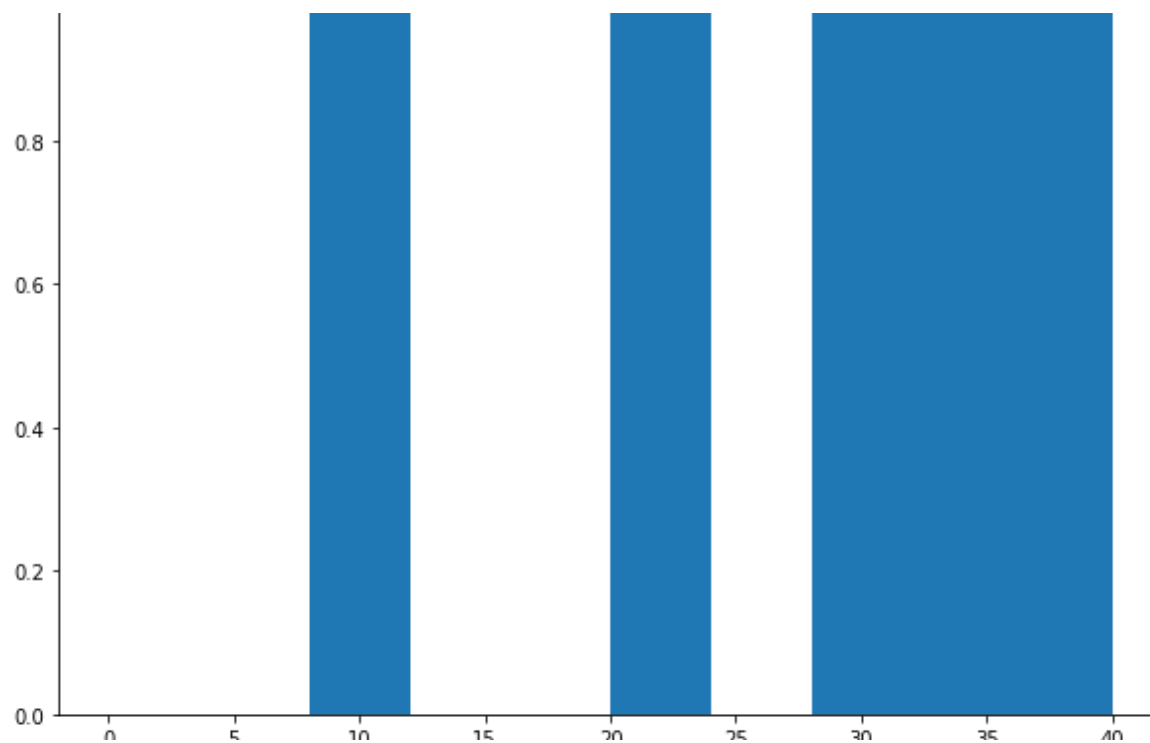


if you want to start the histogram from 0.

$$\text{bins size} = \frac{\text{max}}{\text{bins}}$$

from the above example :

$$\text{bins size} = 40/10 = 4$$

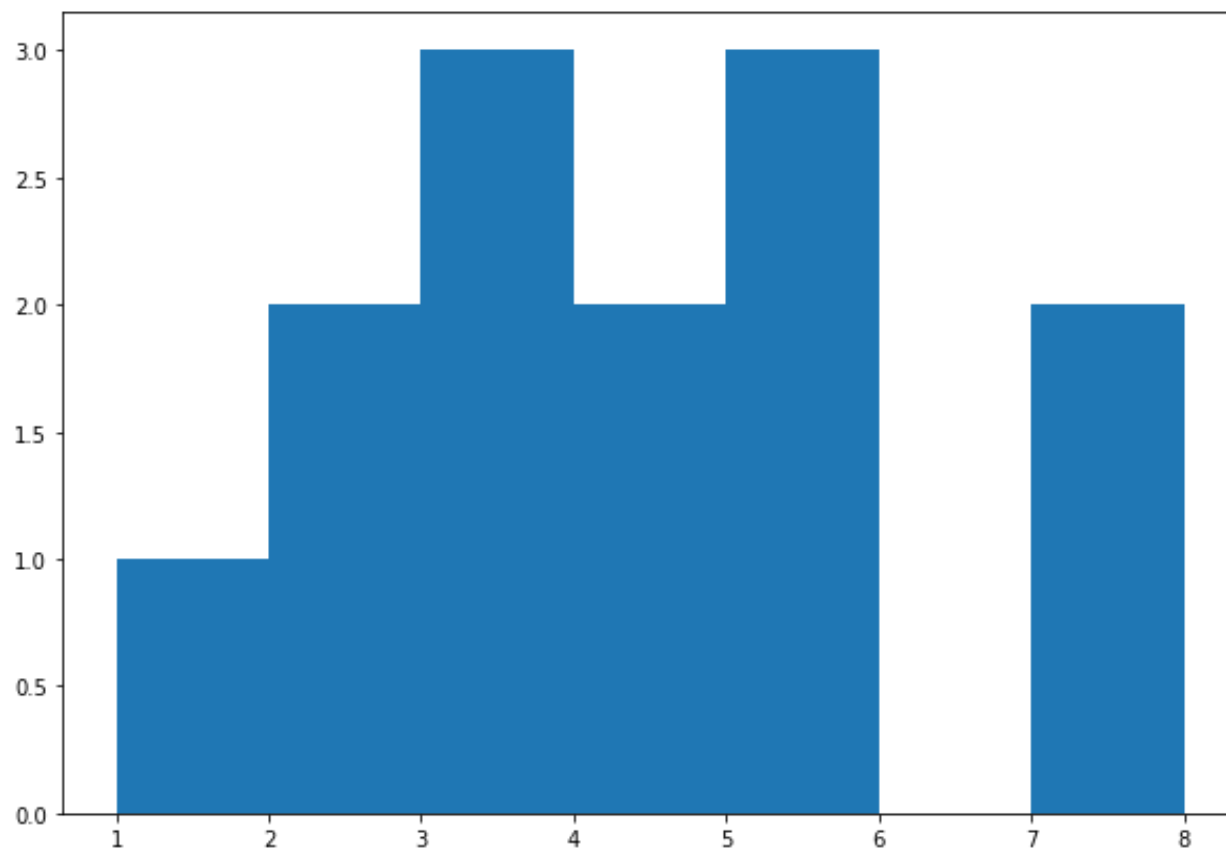


by smotthing the histogram with  
contineous values we will get pdf

# Histograms (Discrete values)

Let's see how to plot histogram with discrete values

bank accounts = [2,3,5,1,4,5,3,7,8,3,2,4,5]

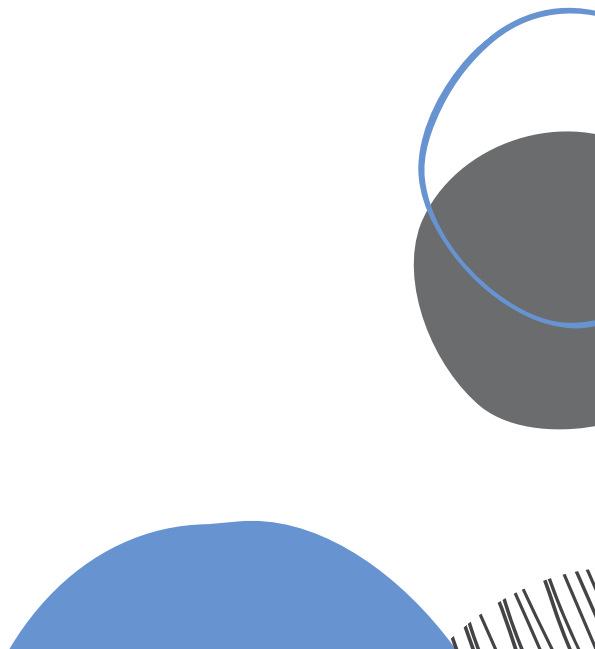




# Measure of central Tendency

A measure of central tendency is a single value that attempts to describe a set of data identifying the central position.

There are three ways to calculate central tendency

- Mean
  - Median
  - Mode
- 



# Mean

sum of all values divided by the total number of values.

$$\text{mean/average} = \frac{\text{sum of the terms}}{\text{number of terms}}$$

$$x = \{1, 2, 3, 4, 5\}$$

$$\text{mean} = (1+2+3+4+5) / 5$$

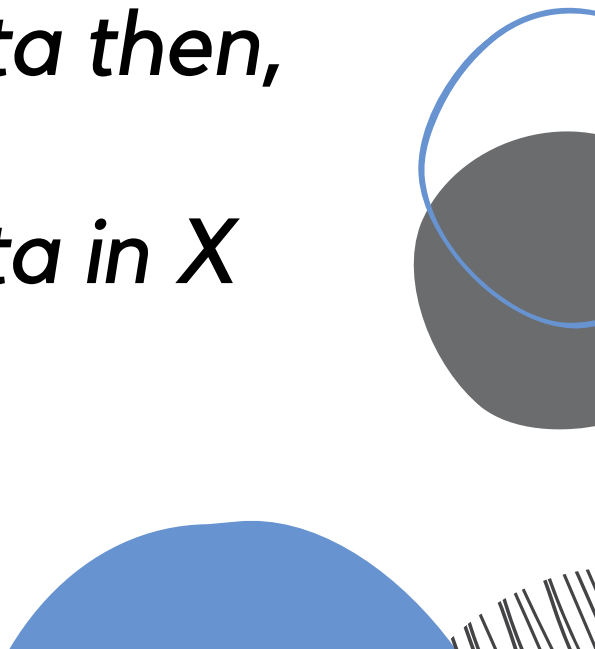
$$\text{mean} = 15/5 = 3$$




# Population Mean

Population can be calculated by the sum of values in the given data or population divided by the total number of values in the given data or population

*The population mean is denoted by the symbol  $\mu$ .  
Let  $X$  be the variable that holds data then,  
**Population Mean ( $\mu$ ) =  $\Sigma X / N$**   
Where,  $\Sigma X$  is the summation of data in  $X$   
And,  $N$  is the count of data in  $X$ .*





**example:**

population age = {24, 23, 2, 1, 28, 27}

$$\text{population mean } (\mu) = \frac{24+23+2+1+28+27}{6}$$

**population mean ( $\mu$ ) = 17.5**







# Sample Mean

The sample mean for a data set is defined as the sum of all the terms divided by the total number of terms. It is denoted by the symbol  $\bar{x}$ .

$$\text{sample mean } \bar{x} = \Sigma xi / n$$

*where,*

*$\Sigma xi$  is the sum of terms in the sample,  
 $n$  is the number of terms in the sample.*





**example:**

Sample age = {24,2,1,27}

$$\text{sample mean } (\bar{x}) = \frac{24+2+1+27}{4}$$

$$\text{sample mean } (\bar{x}) = 17.5$$

**observations:**

population(N) > sample (n)

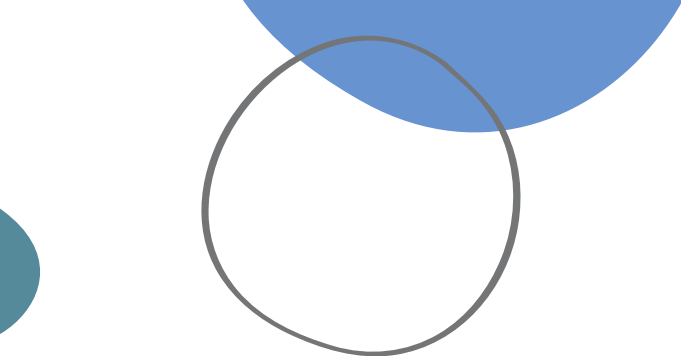
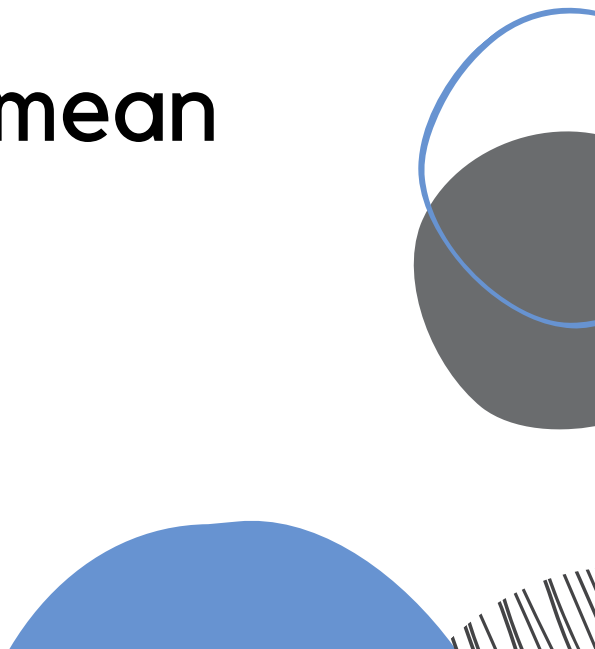
$$\mu \geq \bar{x} \quad \text{or} \quad \bar{x} \geq \mu$$

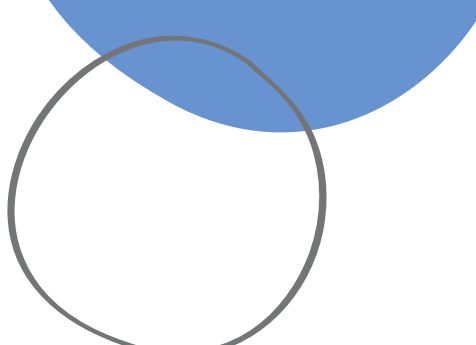

## Practical usage of mean in Ds:

for example we have a dataset

| Age | Salary | family size |
|-----|--------|-------------|
| 21  | 20k    | 4           |
| nan | 25k    | 6           |
| 28  | 28k    | 2           |
| 28  | 30k    | nan         |

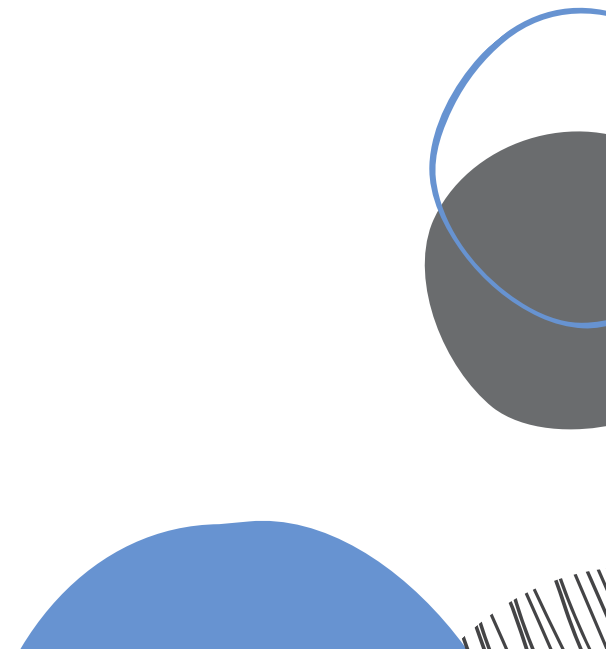
- nan is nothing but a null value not a number

- 
- if we have huge number of data points and very less number of missing values then we can drop the null values
  - if we don't have many data points and a very less number of missing values then we can replace them with mean.
  - let's replace the average/mean value in place of nan
- 



mean of Age = 26 and family  
size=4 lets replace them

| Age       | Salary | family size |
|-----------|--------|-------------|
| 21        | 20k    | 4           |
| <b>26</b> | 25k    | 6           |
| 28        | 28k    | 2           |
| 28        | 30k    | <b>4</b>    |





# Median

what is the usage of median when we already have mean.

when we see any **outliers** in our data we can use **Median**

**question raised?**

what are outliers??

**an outlier is a data point that differs significantly from other observations.**





example:

age = {25, 21, 26, 28, 30, 36}

normally people have these ages

age = {25, 21, 26, 28, 30, 36, **200**}

**200 is an outlier**

**Does any one lived for 200 years?**

if you know anyone who lived for  
200 years comment below





## Median

The middle number, found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers)

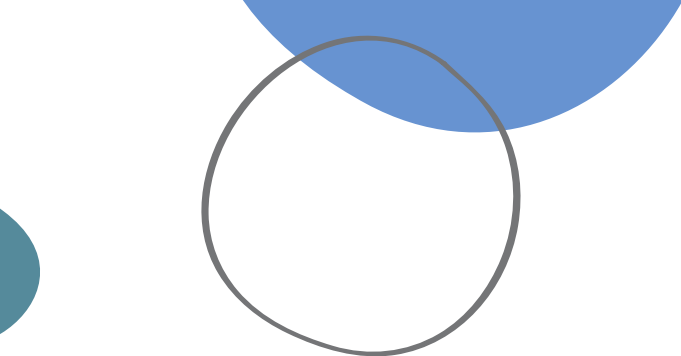
example:

age = {25, 21, 26, 28, 30, 36}

**$\bar{x} = 27.6$**








age = {25, 21, 26, 28, 30, 36, **200**}

***$\bar{x}$  when we have outlier 52.2***

we can observe that there is a huge difference in mean when we have outlier

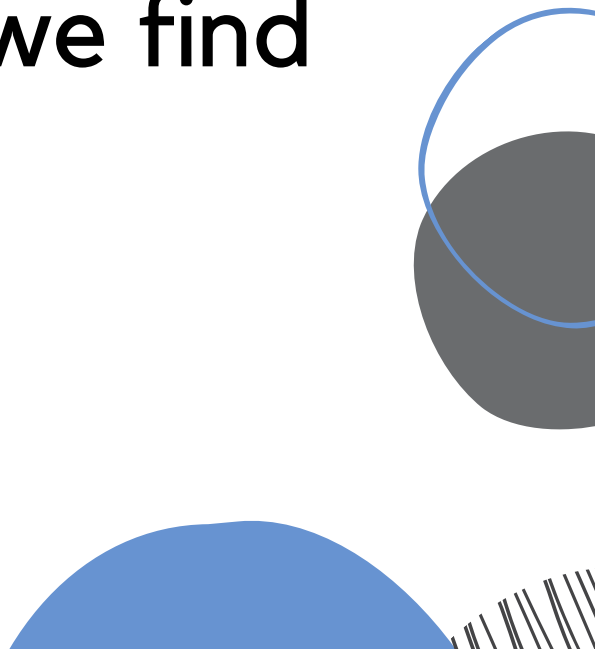
age = {25, 21, 26, 28, 30, 36, **200**}

for same data when we use median  
median is **28**. we can observe that there is no much difference between mean and median even when we have an outlier.





## **Steps to find median.**

- sort the numbers.
  - find the central element.
  - if the number of elements are even then we find the average of central elements.
  - if the elements are odd we find the central element.
- 



example:

age = {25, 21, 26, 28, 30, 36}

here we have even number of elements

**median = (26+28) / 2**

age = {25, 21, 26, 28, 30, 36, **200**}

here we have odd number of elements

**median = 28**

**if no outliers use mean.**

**if outliers present then median.**





# Mode

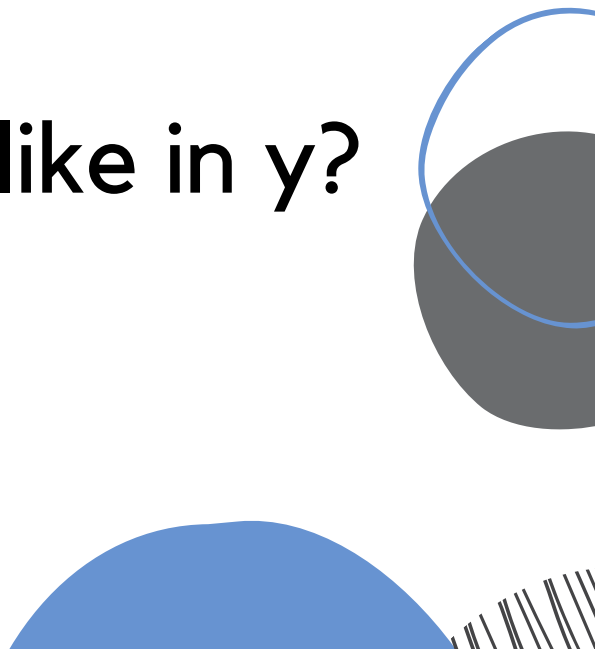
mode is the most frequent occurring element.

$$x = \{1, 1, 2, 2, 3, 3, 3, 3, 4, 4, 5, 6\}$$

here the mode is 3 because it occurred four times.

$$y = \{1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 5, 6\}$$

what if we have equal number like in y?  
here the mode is 3 and 4.





# Practical usage of Mode

**mode is generally used for categorical variables.**

**flowers**

Rose

Jasmine

nan



**This nan is replaced with mode known as Rose.**

Rose

Lily

Rose





# Measure of d

Measures of dispersion measure the scatter of the data, that is how far the values in the distribution are

**1. Variance ( $\sigma^2$ )**

**2. Standard deviation**






## Variance( $\sigma^2$ )

Variance is used to find how data has been spread out.

## Population Variance( $\sigma^2$ )

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

Here N is the population size and the  $x_i$  are data points.  $\mu$  is the population mean.



# Sample Variance( $S^2$ )

Sample Variance ( $S^2$ )

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

$s^2$  = variance  
 $x_i$  = term in data set  
 $\bar{x}$  = Sample mean  
 $\sum$  = Sum  
 $n$  = Sample size

Here  $n$  is the sample size and the  $x_i$  are data points.  $\bar{x}$  is the sample mean.





## Example to find Population Variance

Find the population variance of the age of children in a family of five children aged 16, 11, 9, 8, and 1:

Step 1: Find the mean,  $\mu$ :

$$\mu = 9.$$

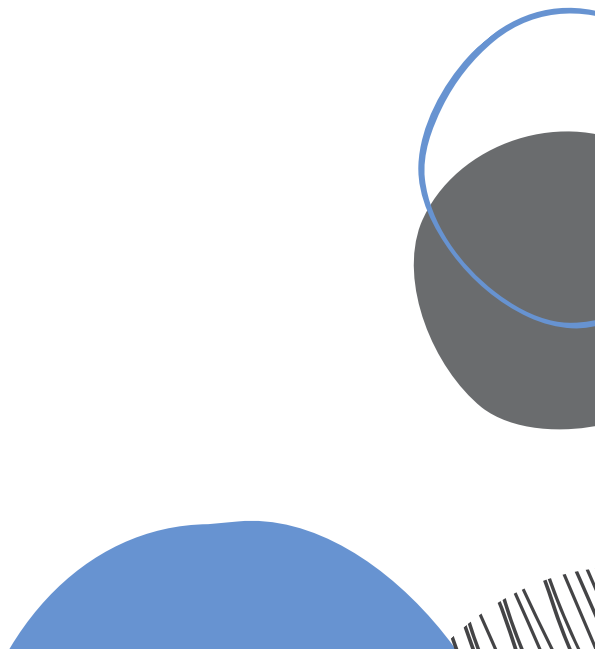
Step 2: Subtract each data point from the mean, then square the result:

$$(16-9)^2 = 49$$

$$(11-9)^2 = 4$$

$$(9-9)^2 = 0$$

$$(8-9)^2 = 1$$

$$(1-9)^2 = 64.$$





Step 3: Add up all of the squared differences from Step 2:

$$(16-9)^2 + (11-9)^2 + (9-9)^2 + (8-9)^2 + (1-9)^2 = 118.$$

Step 4: Divide Step 3 by the number of items.  $118/5$  gives a population variance of 23.6.

in the same way we are going to find sample variance.

but here there is a small difference in denominator for **population** we are using **N** and for **sample variance** we are using **(n-1)**.let's know the difference.






## **difference between $N$ and $(n-1)$**

Bessel's Correction: Why Use  $N-1$  For Variance

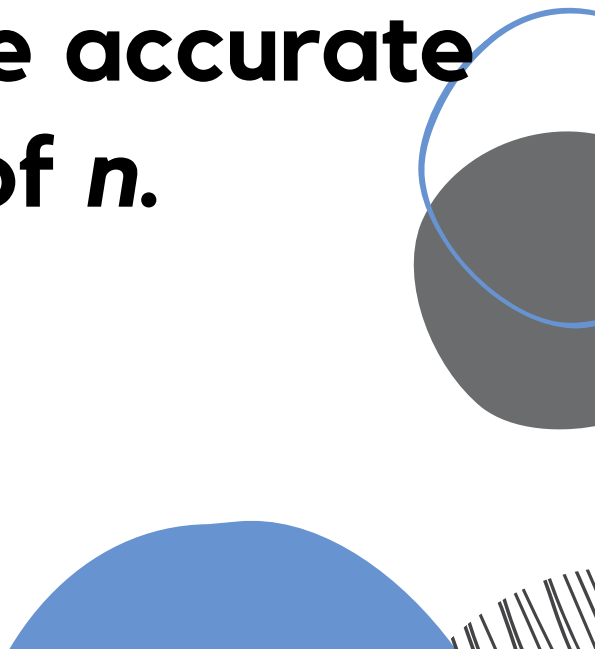
what is Bessel's correction?

- **Bessels' correction refers to the " $n-1$ " found in several formulas, including the sample variance and sample standard deviation formulas. This correction is made to correct for the fact that these sample statistics tend to underestimate the actual parameters found in the population.**
  - **Bessel's is also found in calculations for the Student's T Test.**
- 



**So why do we subtract 1 when using these formulas?**

**The simple answer:** the calculations for both the sample standard deviation and the sample variance both contain a little bias (that's the statistics way of saying "error"). Bessel's correction (i.e. subtracting 1 from your sample size) corrects this bias. In other words, **you'll usually get a more accurate answer if you use  $n-1$  instead of  $n$ .**





# Standard deviation

we will find standard deviation by square root variance.

$$\sigma = \sqrt{v}$$

$$x = \{1, 2, 3, 4, 5\}$$

$$\mu = 3$$

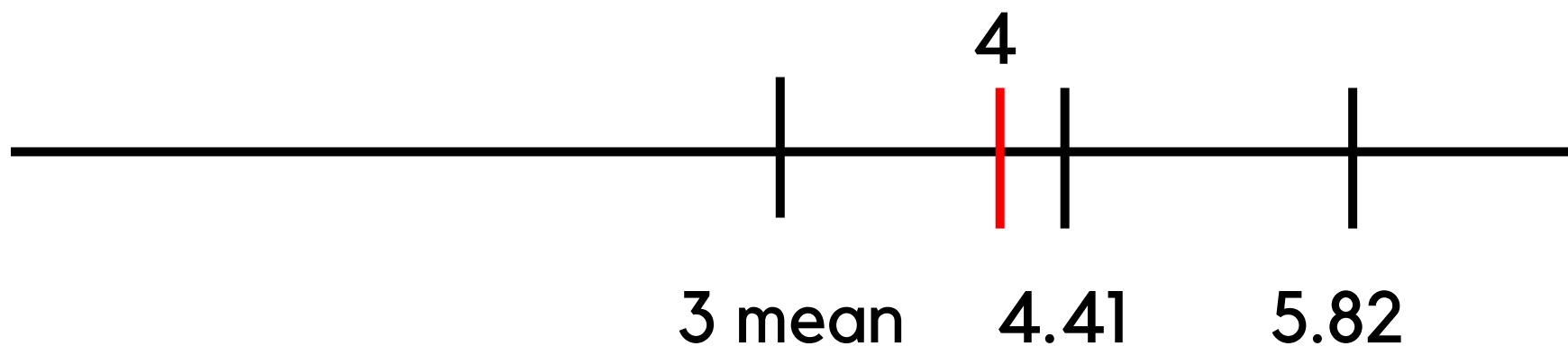
$$\sigma^2 = 2$$

$$\sigma = \sqrt{2} = 1.41$$


# Standard deviation

with the help of standard deviation we find the element in which standard deviation it is in.

let's find where 4 is in with three standard deviations.



4 is in first standard deviations.



# Percentiles

lets know about percentages first

$$x = \{1, 2, 3, 4, 5, 6, 7, 8\}$$

**percentage of even  
numbers =**

**no of even numbers**



**Total no of numbers**

$$\text{percentage of even numbers} = 4 / 8 = 0.5 = 50\%$$




# Percentiles

let's learn about percentiles

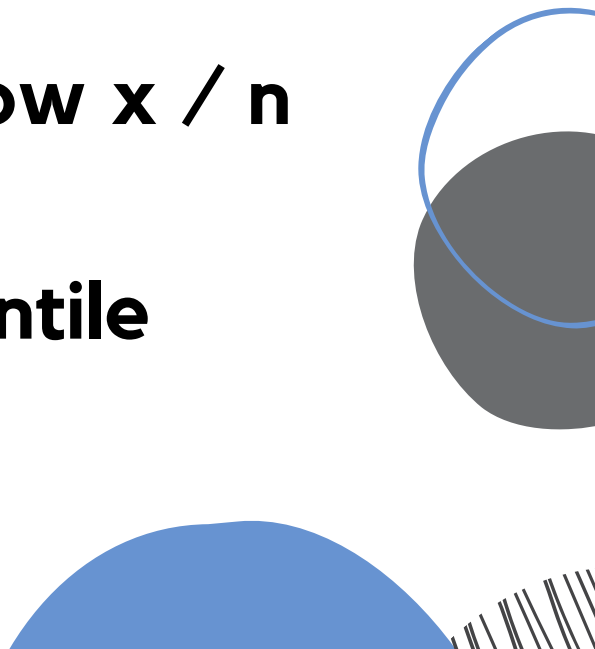
A percentile is a value below which a certain percentage of observation lie

99 percentile = It means the person has got better marks than 99% of the entire students.

dataset = {2,2,3,4,5,5,6,7,8,9}

percentile rank of x = **No of values below x / n**

percentile rank of 8 = 9 / 10 = **90 percentile**





**ex: what is the value that can exists at 25th percentile**

dataset = {2,2,3,4,5,5,6,7,8,9}

value = (percentile / 100) \* n+1

value =  $(25/100) * 11 = 2.75$  which is a index value which is approximately 3


dataset = {2,2,3,4,5,5,6,7,8,9}  
          | | | |  
index= {0,1,2,3}

4 is the value which is present at the 25th index.



## Questions for you

dataset = {2,2,3,4,5,5,5,6,6,6,7,8,9,9}

1. find the 95th percentile value and comment below??
  2. find the 60th percentile value and comment below??
- 




## 5 Number Summary

1. Minimum
2. First Quantile.
3. Median
4. Third Quantile.
5. Maximum

$$x = \{1, 2, 2, 2, 3, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 27\}$$

5 number summary is mainly used find the lower fence and upper fence to exclude outliers.



## 5 Number Summary

$$\text{lower fence} = Q1 - 1.5(IQR)$$

$$\text{Upper fence} = Q3 + 1.5(IQR)$$

first lets find  $Q1 = 25 \text{ percentile} / 100 * (n+1)$

$$Q1 = 25 / 100 * (20) = 5^{\text{th}} = 3$$

lets find  $Q3 = 75 \text{ percentile} / 100 * (n+1)$

$$Q3 = 75 / 100 * (20) = 15^{\text{th}} = 7$$

# 5 Number Summary

$$\text{IQR} = Q3 - Q1 = 7 - 3 = 4$$

$$\text{lower fence} = 3 - 1.5(4) = -3$$

$$\text{Upper fence} = 7 + 1.5(4) = 13$$

the values which are not present between -3 and 13 are outliers.

with the help of box plot we can identify outliers



outlier 27 from  
above data





Thank You

for your motivation sir and your appreciation.

