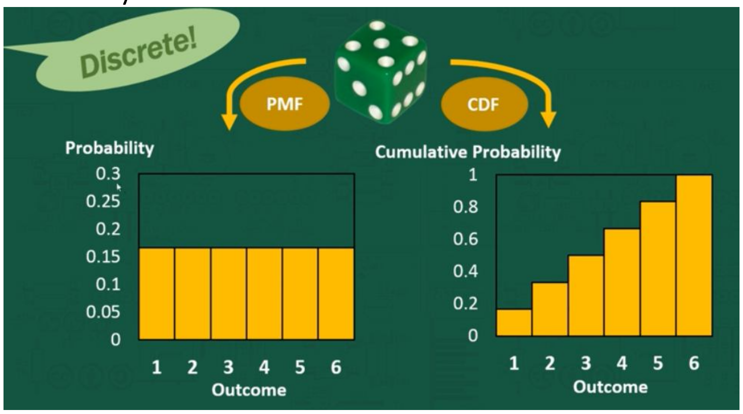
ECDF : Empirical Cumulative Distribution Function (ECDF)
https://www.youtube.com/watch?v=3xAIWiTJCvE

https://www.youtube.com/watch?v=YXLVjCKVP7U

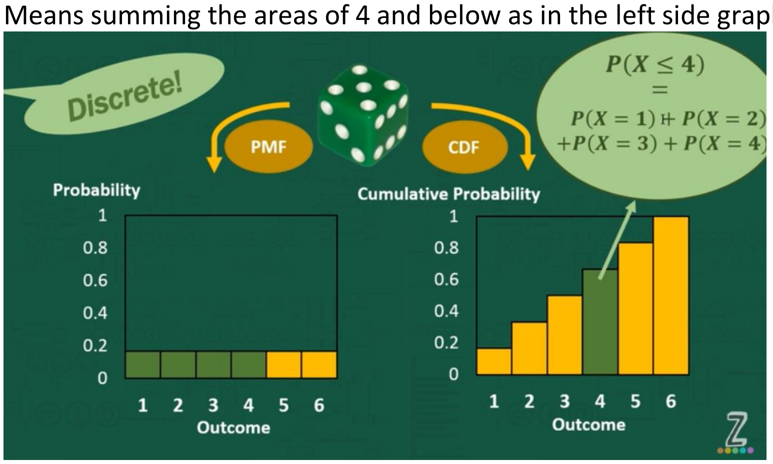
http://www.zstatistics.com/videos

Probability Mass Function Vs Cumulative Distribution Function

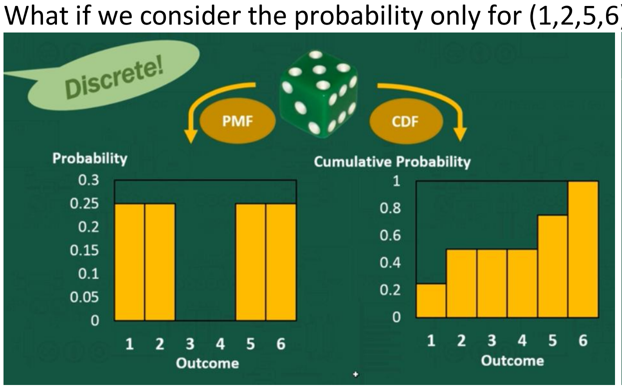


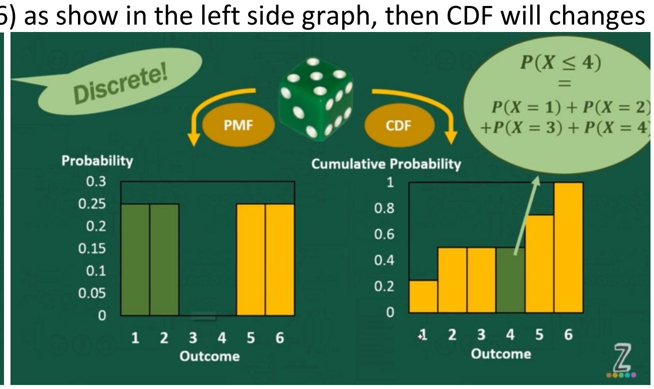
By change the scale on the left side to match the right side

The height of 4 = the probability at 4 + the probability at 3 + the probability at 2 + at 1
Means summing the areas of 4 and below as in the left side graph

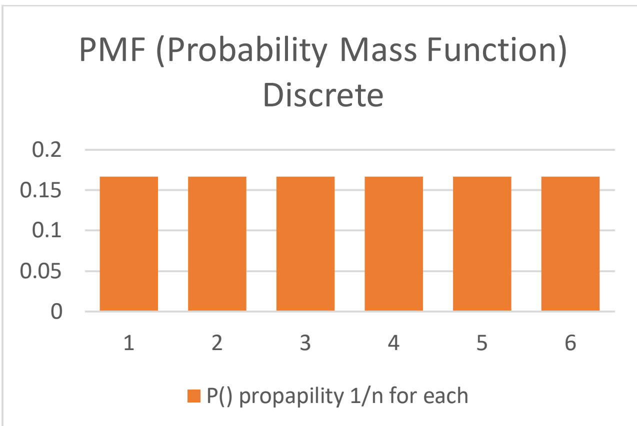


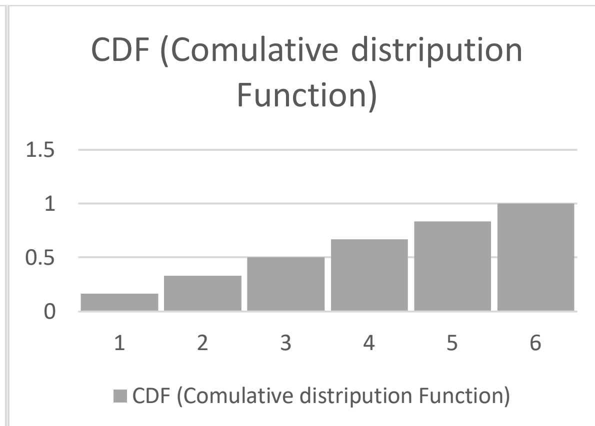
So in CDF final bar need to be 1
What if we consider the probability only for (1,2,5,6) as show in the left side graph, then CDF will changes



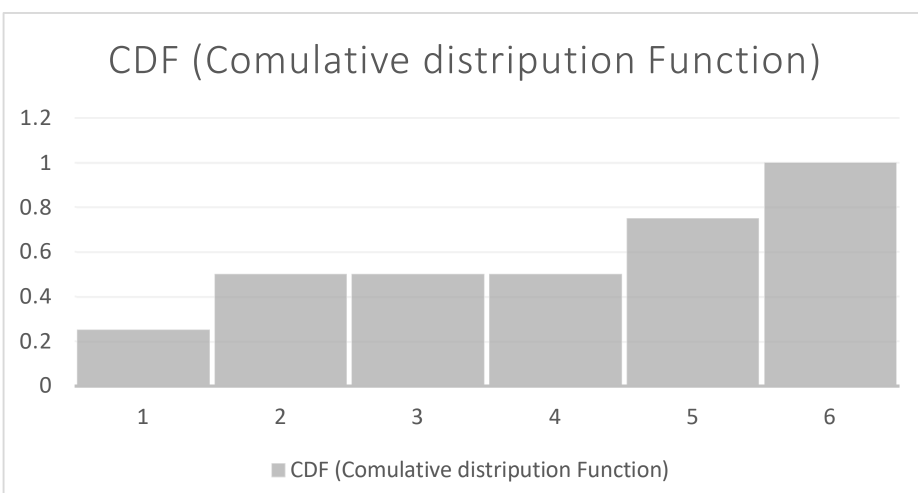


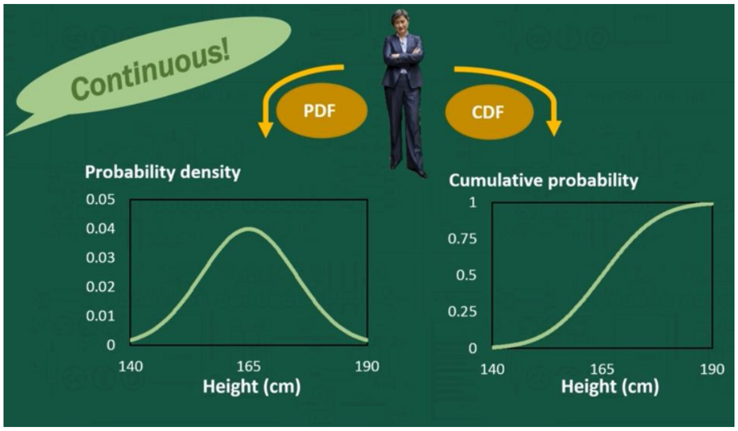
	PMF (Probability Mass function)	CDF (Comulative Distripution Function)	
dice	P() propapility 1/n for each	CDF (Comulative distripution Function)	CDF
1	0.166666667	0.166666667	p(1)
2	0.166666667	0.333333333	p(1)+p(2)
3	0.166666667	0.5	P(1)+P(2)+P(3)
4	0.166666667	0.666666667	P(1)+P(2)+P(3)+P(4)
5	0.166666667	0.833333333	P(1)+P(2)+P(3)+P(4)+P(5)
6	0.166666667	1	P(1)+P(2)+P(3)+P(4)+P(5)+P(6)

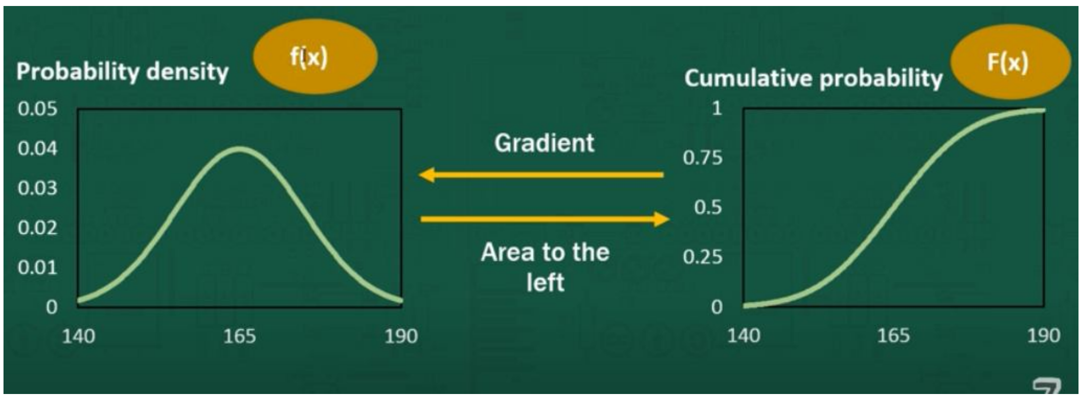




	PMF (Probability mass function)	CDF (Comulative Distripution Function)	
dice	P() propapility 1/4 for each	CDF (Comulative distripution Function)	CDF
1	0.25	0.25	p(1)
2	0.25	0.5	p(1)+p(2)
3	0	0.5	
4	0	0.5	
5	0.25	0.75	P(1)+P(2)+P(5)
6	0.25	1	P(1)+P(2)+P(5)+P(6)



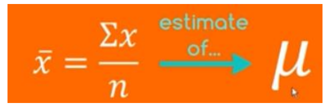

Continuous!
PDF
CDF
Probability density
0.05
0.04
0.03
0.02
0.01
0
140
165
190
Height (cm)
Cumulative probability
1
0.75
0.5
0.25
0
140
165
190
Height (cm)


Probability density
f(x)
0.05
0.04
0.03
0.02
0.01
0
140
165
190
Gradient
Area to the left
Cumulative probability
F(x)
1
0.75
0.5
0.25
0
140
165
190

The MEAN μ:

https://www.youtube.com/watch?v=bfQLNyiDPsk&list=PLTNMv857s9WVStKLco6ZBOsfSGXzJ1L0f&index=1

$$\bar{x} : \mu = \frac{\sum X}{n}$$



The total of the samples over the number of samples of the data

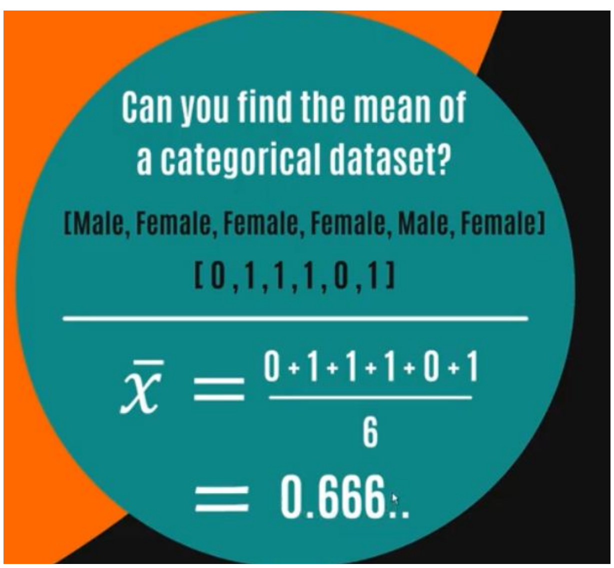
Example:
10, 28, 28, 33, 54

$$\bar{x} = \mu = \frac{10+28+28+33+54}{5} = \frac{153}{5} = 30.6$$

Weighted Mean:
Use the unique values in a table counting the frequency, so we weighted each of the number with the frequency that occurs

X	F(X)
10	1
28	2
33	1
54	1

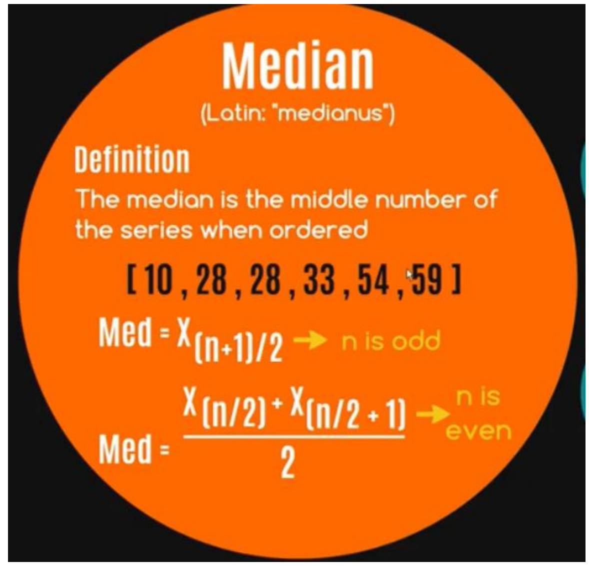
$$\bar{x} : \mu = \frac{\sum XF(X)}{\sum F(X)} = \frac{10(1) + 28(2) + 33(1) + 54(1)}{1 + 2 + 1 + 1}$$



The MEDIAN:

https://www.youtube.com/watch?v=rvBqEEGtJY4&list=PLTNMv857s9WVStKLco6ZBOsfSGXzJ1L0f&index=3

The middle number of the series when ordered, the center of the data, so if odd it will be the middle number , if even it will be the average of the 2 middle numbers.

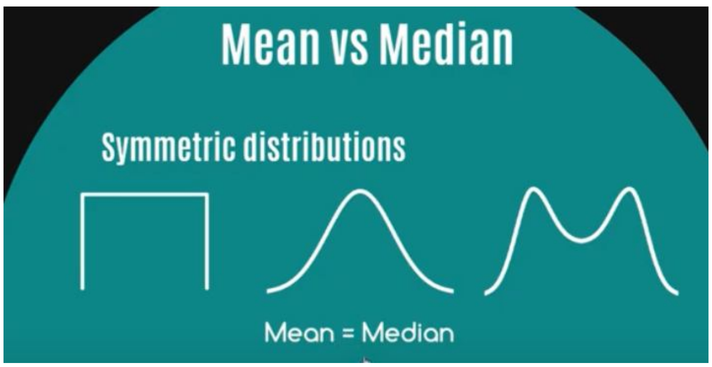


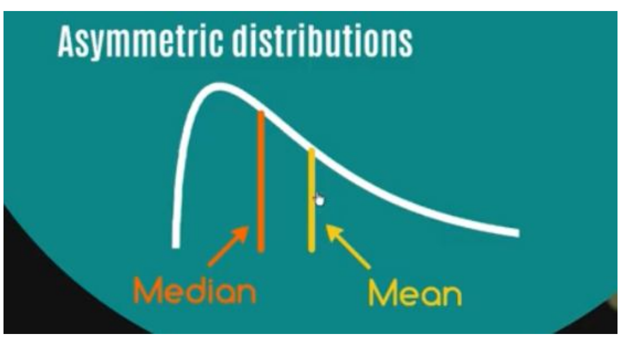
Example:
10, 28, 28, 33, 54

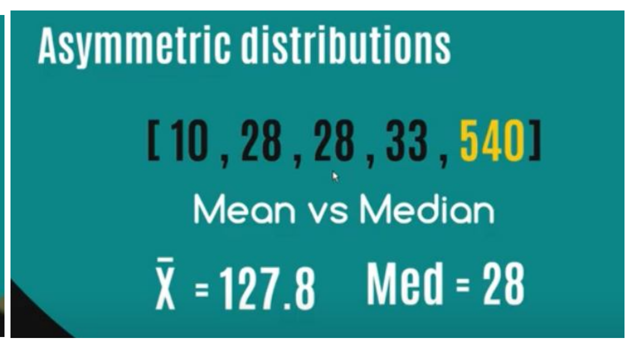
M or MED = 28 , mean as calculated before for the same is 30.6

Example:
10, 28, 28, 33, 54, 59

MED = 28+33/2=





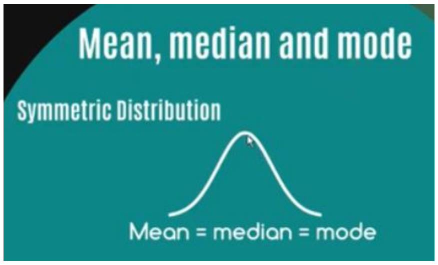


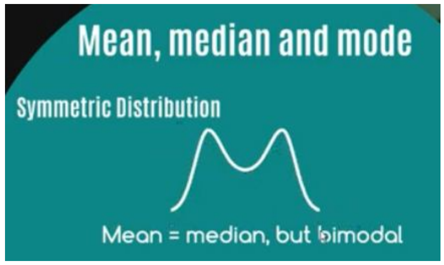
The MODE:

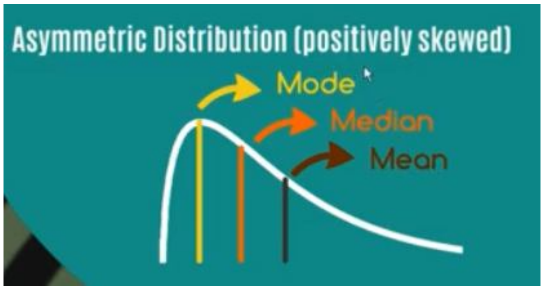
The observation of the highest frequency, more suitable for large data set

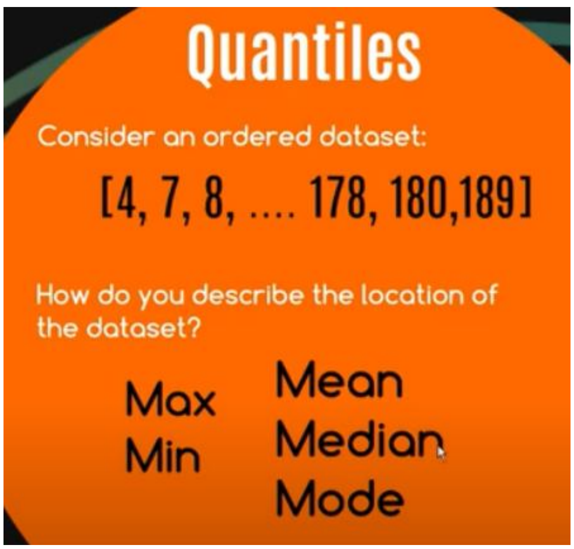
Example:
10, 28, 28, 33, 54

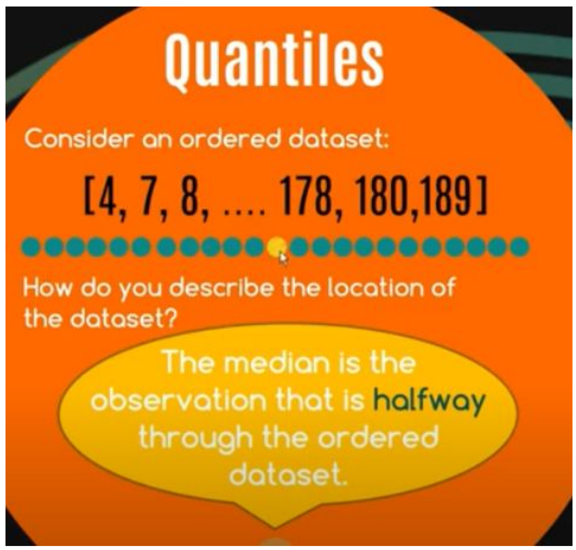
Mode = 28 (highest frequency)

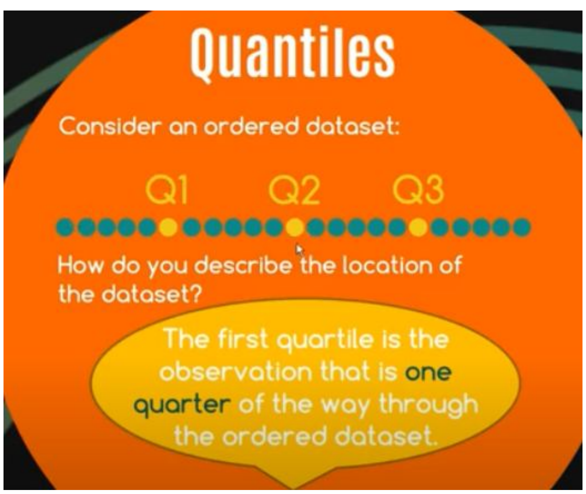

Mean, median and mode
Symmetric Distribution
Mean = median = mode

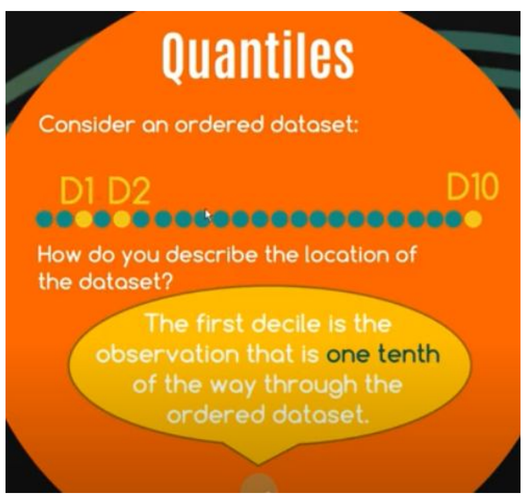

Mean, median and mode
Symmetric Distribution
Mean = median, but bimodal

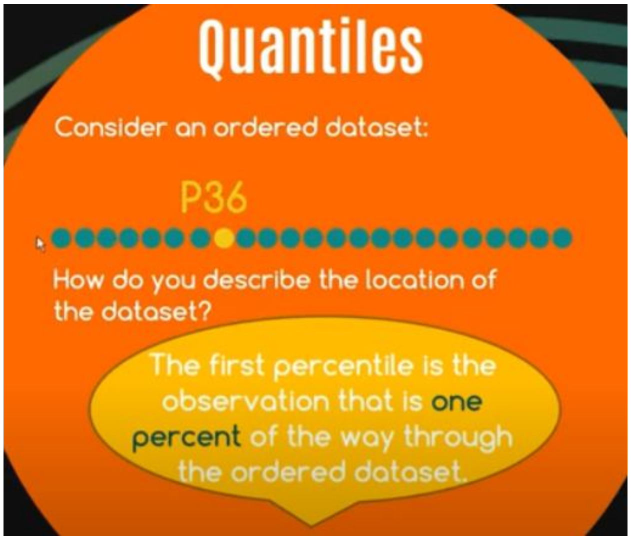

Asymmetric Distribution (positively skewed)
Mode
Median
Mean


Quantiles
Consider an ordered dataset:
[4, 7, 8, ... 178, 180,189]
How do you describe the location of the dataset?
Max
Min
Mean
Median
Mode


Quantiles
Consider an ordered dataset:
[4, 7, 8, ... 178, 180,189]
How do you describe the location of the dataset?
The median is the observation that is halfway through the ordered dataset.

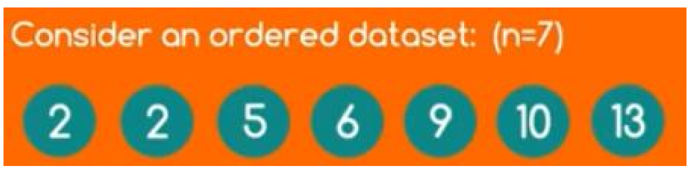

Quantiles
Consider an ordered dataset:
Q1 Q2 Q3
How do you describe the location of the dataset?
The first quartile is the observation that is one quarter of the way through the ordered dataset.


Quantiles
Consider an ordered dataset:
D1 D2 D10
How do you describe the location of the dataset?
The first decile is the observation that is one tenth of the way through the ordered dataset.



Quartile: split the dataset into quarter (4)
Decile: split the dataset in to tens (10)
Percentile: split the dataset in to hundreds (100)

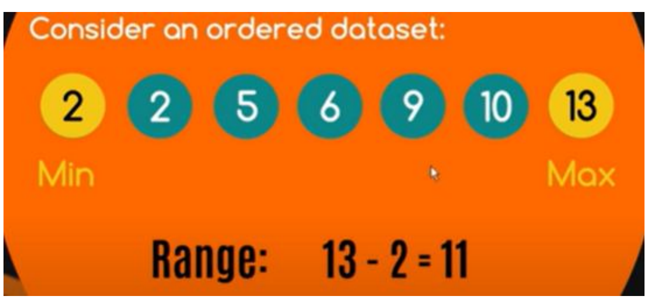
Example:
Calculate the quantile for the below small dataset (quartile, decile, percentile)

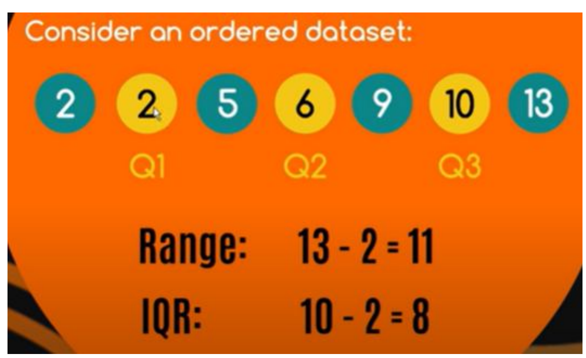


Find the five number summary (Min, Q1, Q2, Q3, Max) for the series above.

Solution:
Min = 2
Q1 = the median of the left side from 6 = 2
Med = 6
Q3 = the median of the right side from 6 = 10
Max= 13

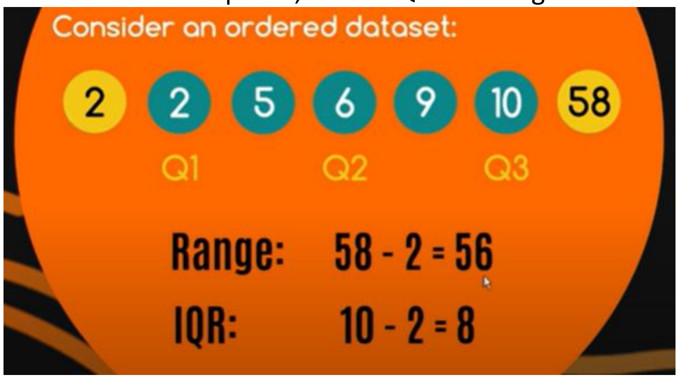
Interquartile Range (IQR) | Box and whisker plot



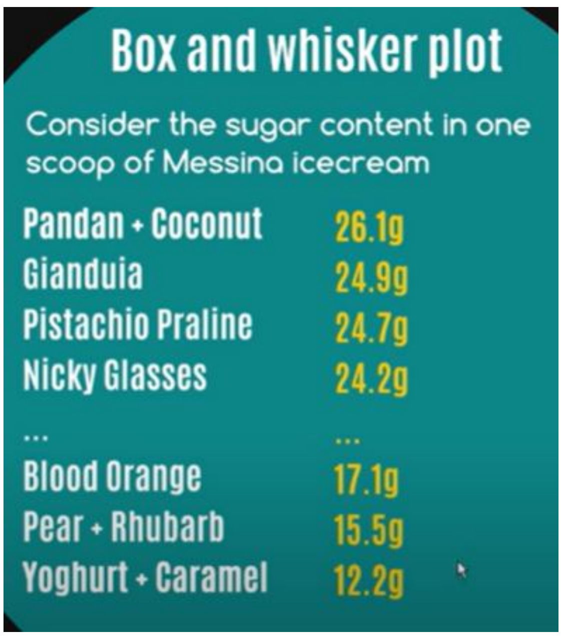


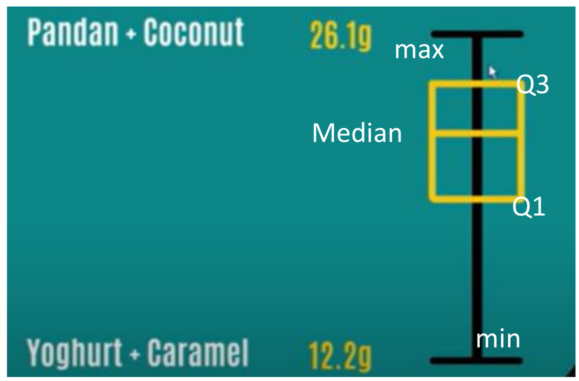
Range = max - min
IRQ = Q3 – Q1

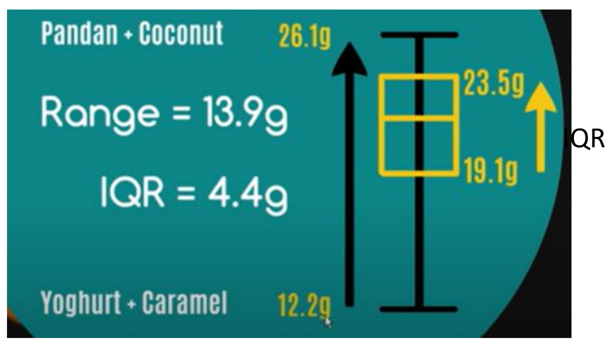
Suppose max value is to far from the range of the numbers then the Range value will be to high as this data set has too much spread, then IRQ is showing the most common value



Example:







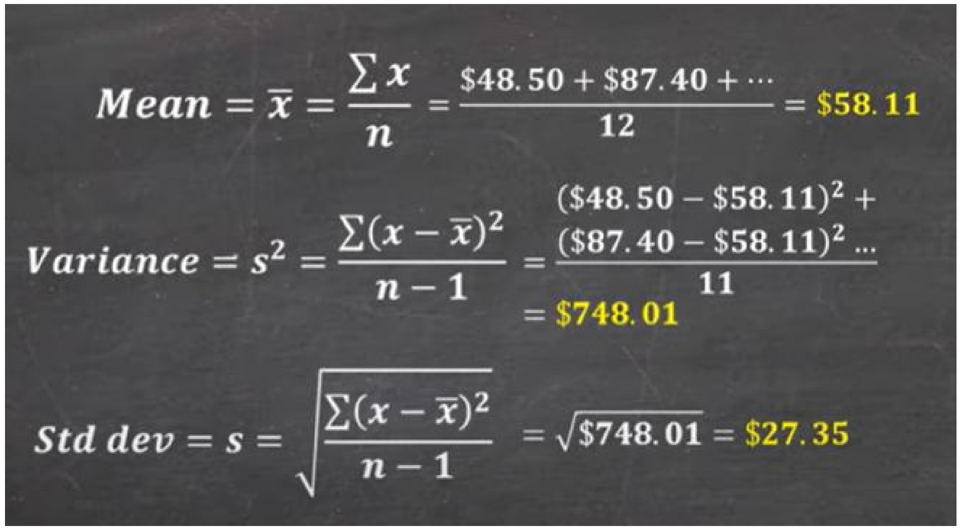
Variance and Standard Deviation

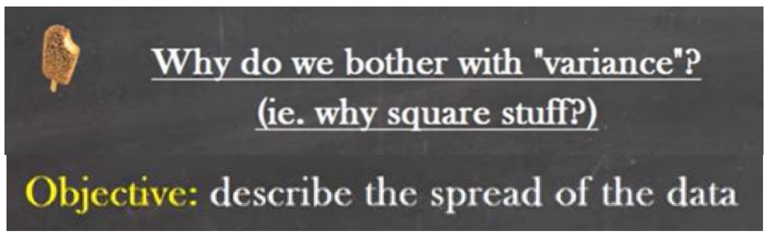
$$Mean = \overline{x} = \frac{\sum x}{n}$$

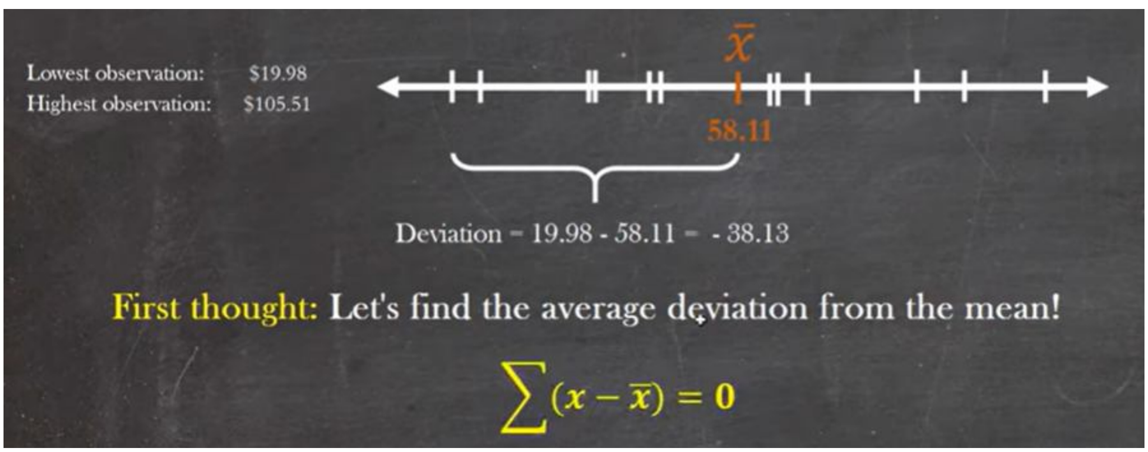
$$Variance = s^2 = \frac{\sum (x - \overline{x})^2}{n - 1}$$

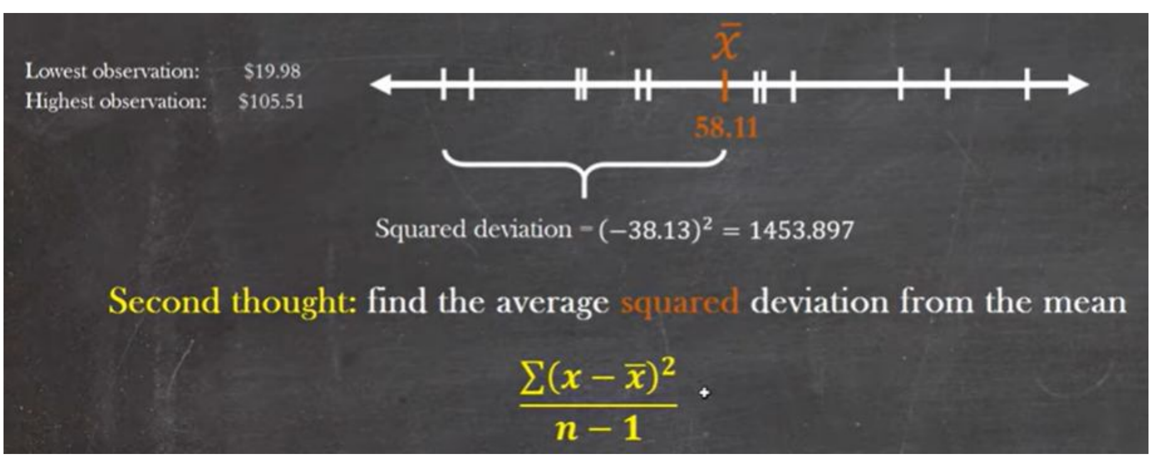
$$Std\ dev = s = \sqrt{\frac{\sum (x - \overline{x})^2}{n - 1}}$$

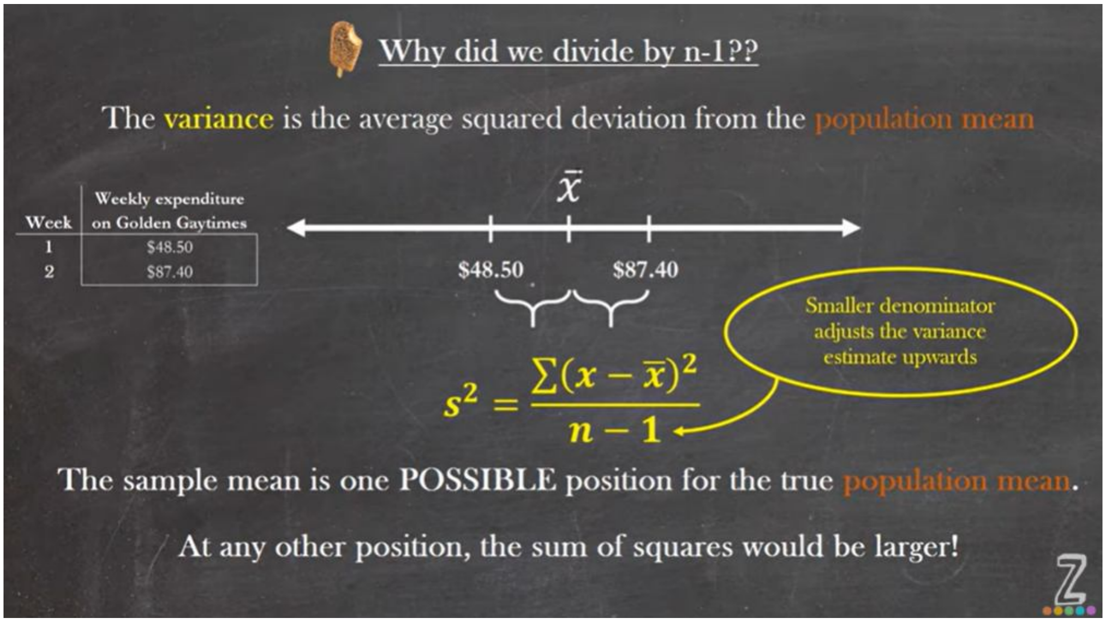

Week
Weekly expenditure on Golden Gaytimes
1 $48.50
2 $87.40
3 $19.98
4 $59.74
5 $40.87
6 $105.51
7 $40.80
8 $23.10
9 $98.10
10 $60.54
11 $64.81
12 $48.01


$Mean = \bar{x} = \frac{\sum x}{n} = \frac{\$48.50 + \$87.40 + \cdots}{12} = \$58.11$
$Variance = s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{(\$48.50 - \$58.11)^2 + (\$87.40 - \$58.11)^2 \ldots}{11} = \$748.01$
$Std\ dev = s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} = \sqrt{\$748.01} = \$27.35$


Why do we bother with "variance"?
(ie. why square stuff?)
Objective: describe the spread of the data


Lowest observation: $19.98
Highest observation: $105.51
$\bar{x}$
58.11
Deviation = 19.98 - 58.11 = - 38.13
First thought: Let's find the average deviation from the mean!
$\sum(x-\bar{x}) = 0$


Lowest observation: $19.98
Highest observation: $105.51
$\bar{x}$
58.11
Squared deviation = $(-38.13)^2 = 1453.897$
Second thought: find the average squared deviation from the mean
$\frac{\sum(x-\bar{x})^2}{n-1}$

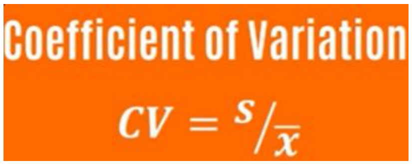


Example:

Week	Expenditure	x-mean	(x-mean)^2
3	19.98	-38.13333333	1454.15111
8	23.1	-35.01333333	1225.93351
7	40.8	-17.31333333	299.751511
5	40.87	-17.24333333	297.332544
12	48.01	-10.10333333	102.077344
1	48.5	-9.613333333	92.4161778
4	59.74	1.626666667	2.64604444
10	60.54	2.426666667	5.88871111
11	64.81	6.696666667	44.8453444
2	87.4	29.28666667	857.708844
9	98.1	39.98666667	1598.93351
6	105.51	47.39666667	2246.44401
		total	8228.12867

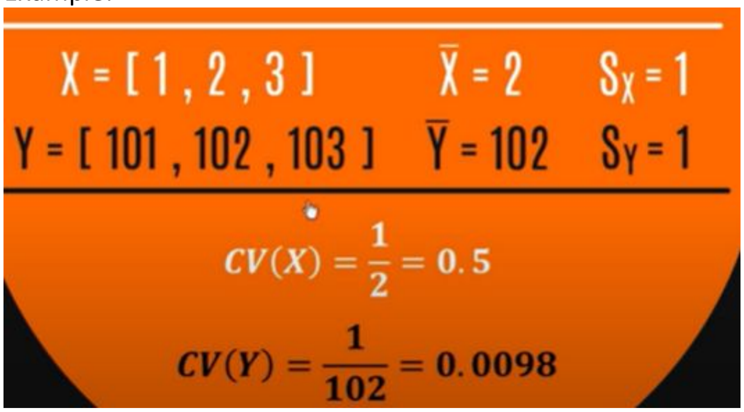
Min	19.98	Minimum
Max	105.51	Maximum
Median	54.12	average of the 2 numbers in the middle if dataset is even
Mean	58.11333333	Average
Range	85.53	Maximum - Minimum
Varience	748.011697	Variance = total(x-mean)^2/n-1
std Deviation	27.3498025	std deviation = SQRT(Variance)

Coefficient Of Variation

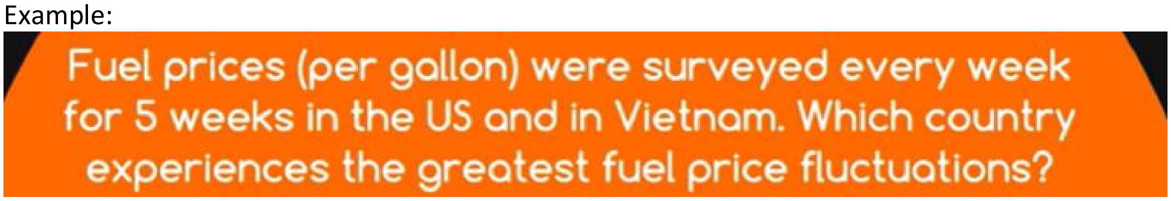


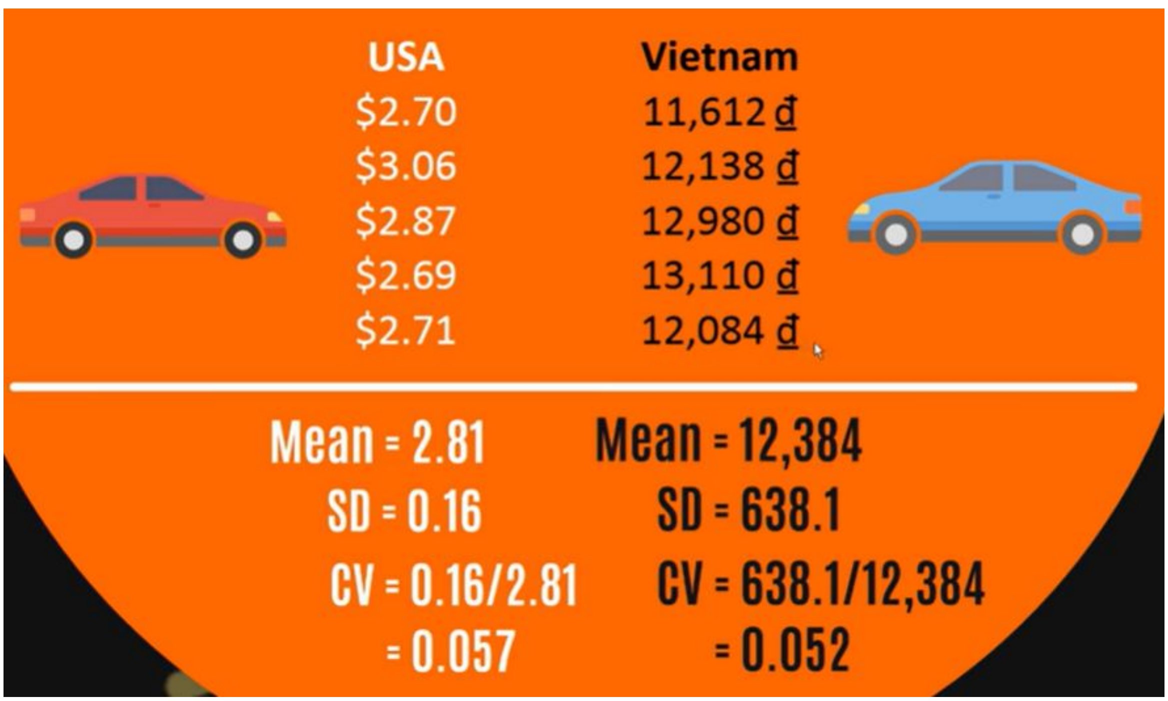
Coefficient of Variation = Standard deviation / Main

Example:



Example:



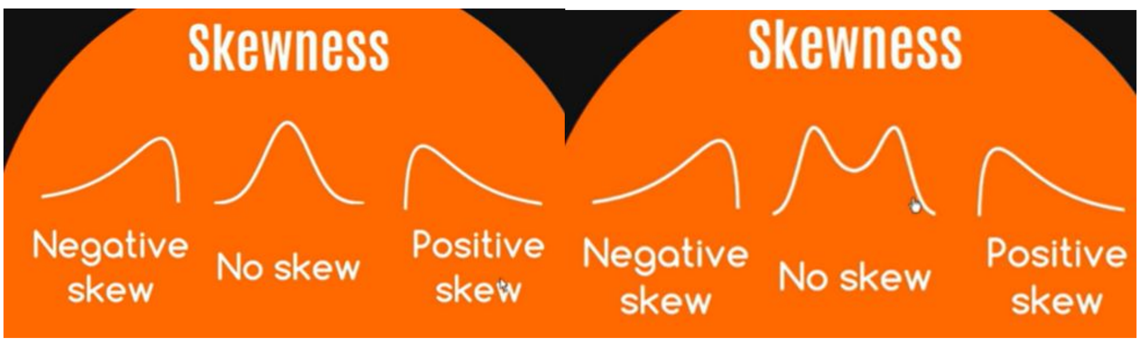


USA	deviation	(deviation)^2
2.7	-0.11	0.0112
3.06	0.25	0.0645
2.87	0.06	0.0041
2.69	-0.12	0.0135
2.71	-0.10	0.0092
	variation	0.0256
	std deviation	0.1601
	CV	0.0571

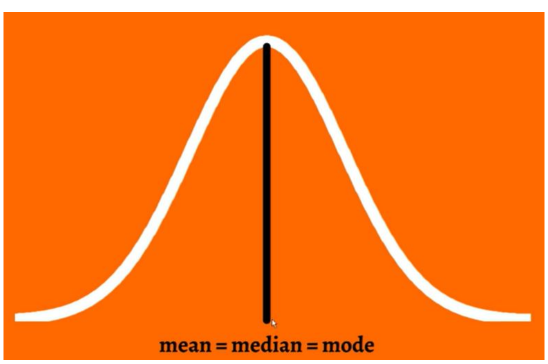
Vietnam	deviation	(deviation)^2
11,612	-773	597,219.840
12,138	-247	60,910.240
12,980	595	354,263.040
13,110	725	525,915.040
12,084	-301	90,480.640
	variation	407,197.200
	std deviation	638.120
Coefficient of Variation	CV = (std deviation)/mean	0.052

mean	2.806
Min	2.69
Max	3.06
range	0.37

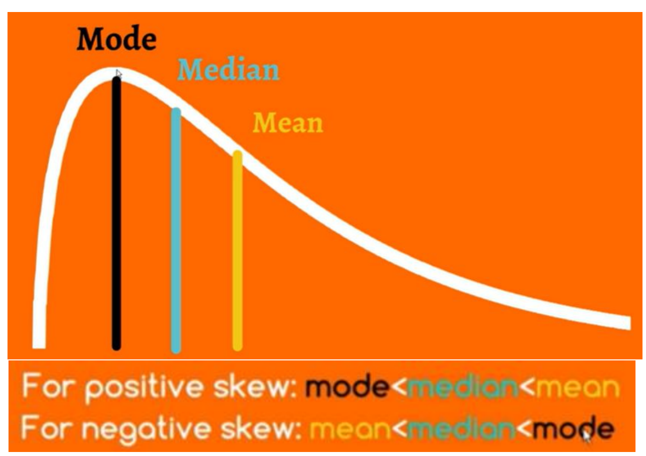
mean	12,385
Min	11612
Max	13110
range	1498



No Skew



left (negative) skew: more observations in the right side of the mode



Example :
Consider the grades from two students attending a same school:
A : {17, 16, 16, 16, 15, 16, 18, 17, 16, 16}
B : {16, 15, 14, 17, 14, 17, 14, 14, 14, 14}
What can we say about the two students?

We could start by counting:

Grade	14	15	16	17	18
A	0	1	6	2	1
B	6	1	1	2	0

We could also report the relative frequencies:

Grade	14	15	16	17	18
A	0.0	0.1	0.6	0.2	0.1
B	0.6	0.1	0.1	0.2	0.0

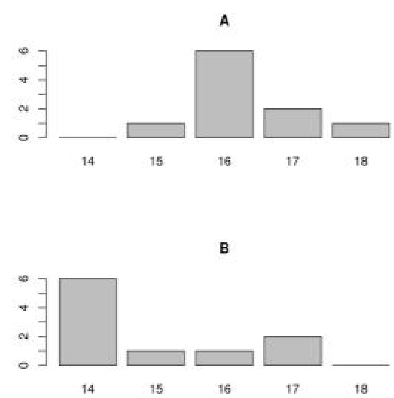


Figure: Distribution of grades of students A and B

Definition
Given a qualitative variable X with levels $\{x_i\}$, we define the *frequency distribution* of X as the following table:

level of X	absolute frequency	relative frequency
x_1	n_1	$f_1 = n_1/n$
	...	
x_i	n_i	$f_i = n_i/n$
	...	
x_k	n_k	$f_k = n_k/n$

Specifically,
$\{n_i\}$ $(\sum_i n_i = n)$ are the *absolute frequencies*
$\{f_i\}$ $(\sum_i f_i = 1)$ are the *relative frequencies*

In R: cfr. `table`, `prop.table`, `barplot`

Example

Consider again the grades from student A, this time *sorting* his grades:

A : $\{15, 16, 16, 16, 16, 16, 16, 17, 17, 18\}$

We previously handled them as discrete, though we can think of the grades as real numbers, and observe:

$1/10$ of the grades are ≤ 15, $7/10$ are ≤ 16, $9/10$ are ≤ 17, and $10/10$ are ≤ 18.
The computation is more straightforward if we start from the previously computed relative frequencies f_i:

i	x_i	n_i	f_i	$\sum_{j=1}^{i} f_i$
1	15	1	0.1	0.1
2	16	6	0.6	0.7
3	17	2	0.2	0.9
4	18	1	0.1	1.0

We just described the empirical cumulative distribution function of student's A grades

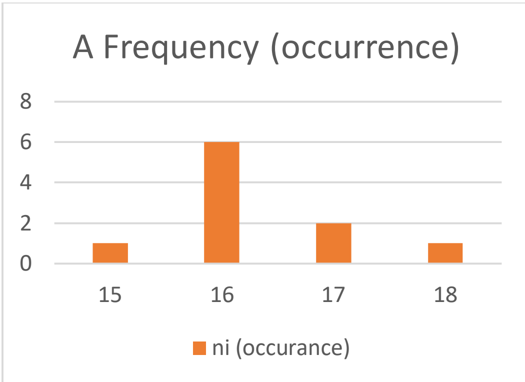
ni : the number of occurrence ,
fi : ni/n the number of occurrence / the total number
the sum of fi the total > f2 = f2 + f1 , f3 = f3+f2+f1

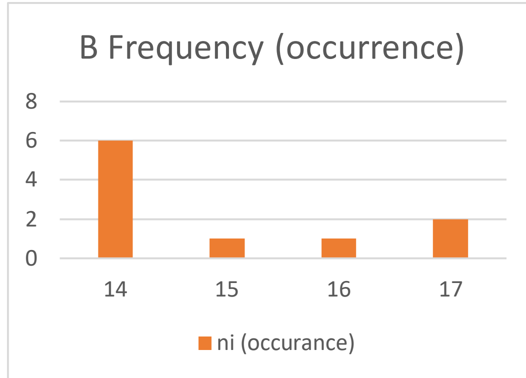
A

i	grades	ni (occurrence)	fi (frequency) ni/total n	total frequency (fi+fi-1+fi-2+…+f1)
1	15	1	0.1	0.1
2	16	6	0.6	0.7
3	17	2	0.2	0.9
4	18	1	0.1	1

B

i	grades	ni (occurrence)	fi (frequency) ni/total n	total frequency (fi+fi-1+fi-2+…+f1)
1	14	6	0.6	0.6
2	15	1	0.1	0.7
3	16	1	0.1	0.8
4	17	2	0.2	1





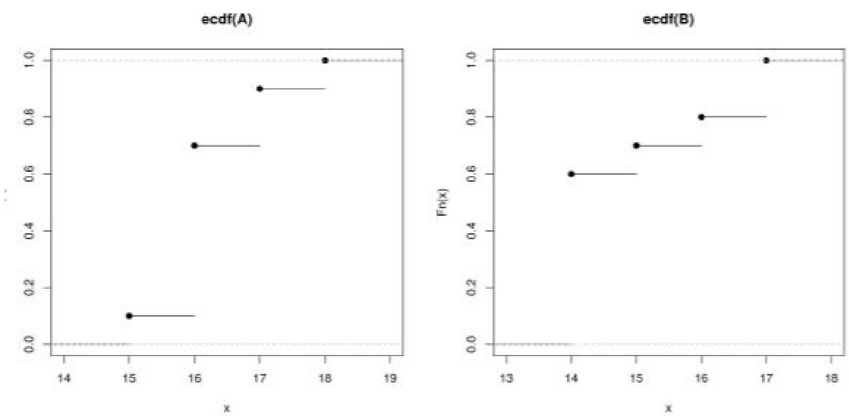


Figure: ECDF of the student's grades

ECDF some properties:

- $0 \leq F_n(t) \leq 1, \quad t \in \mathcal{R}$
- non decreasing
- right continuous
- $F_n(-\infty) = 0,\ F_n(+\infty) = 1$

Probability:
Probability of an event is a *chance* that this event will happen.

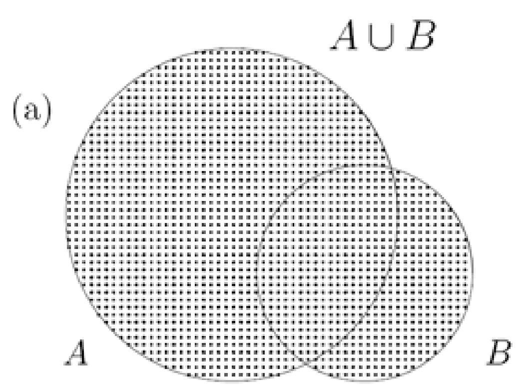
Example.

If there are 5 communication channels in service, and a channel is selected at random when a telephone call is placed, then each channel has a probability $1/5 = 0.2$ of being selected

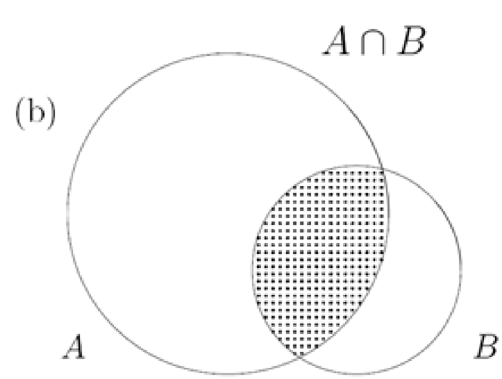
Sample space
A collection of all elementary results, or **outcomes** of an experiment

Event
Any set of outcomes is an **event**. Thus, events are subsets of the sample space

A **union** of events *A, B, C, . . .*
is an event consisting of *all* the outcomes in all these events. It occurs if *any* of *A. B. C.* . . . occurs, and therefore, corresponds to the word
"OR": *A* or *B* or *C* or…



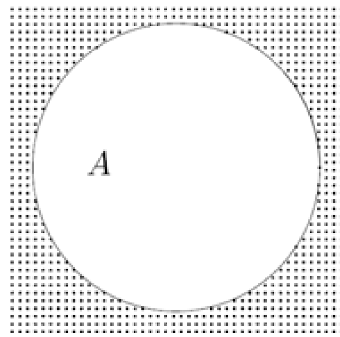
An **intersection** of events *A, B, C, . . .*
is an event consisting of outcomes that are *common* in all these events. It occurs if *each A, B, C,* . . . occurs, and therefore, corresponds to the word
"AND": *A* and *B* and *C* and ..



A **complement** of an event *A*
is an event that occurs every time when *A* does not occur. It consists of outcomes excluded from *A*, and therefore, corresponds to the word
"NOT": not *A*

$\overline{A}$

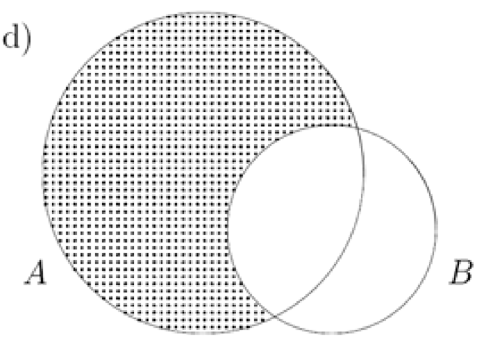
(c)



A **difference** of events *A* and *B*:
Consists of all outcomes included in *A* but excluded from *B*. It occurs when *A* occurs and *B* does not, and corresponds to
"BUT NOT": A but not *B*

$A\backslash B$

(d)



$$\begin{aligned} A \cup B &= \text{union} \\ A \cap B &= \text{intersection} \\ \overline{A} \text{ or } A^c &= \text{complement} \\ A\backslash B &= \text{difference} \end{aligned}$$

Events *A* and *B* are **disjoint** if their intersection is empty, A ∩ B = Φ

Events *A, B, C . . .* are **exhaustive** if their union equals the whole sample space, A U B U C U.. = Ω

Example:
Any event *A* and its complement $\bar{A}$ represent a classical example of disjoint and exhaustive events.

it is often easier to compute probability of an intersection than probability of a union. Taking complements converts unions into intersections

$$\boxed{\overline{E_1 \cup ... \cup E_n} = \overline{E_1} \cap ... \cap \overline{E}_n, \overline{E_1 \cap ... \cap E_n} = \overline{E}_1 \cup ... \cup \overline{E}_n}$$