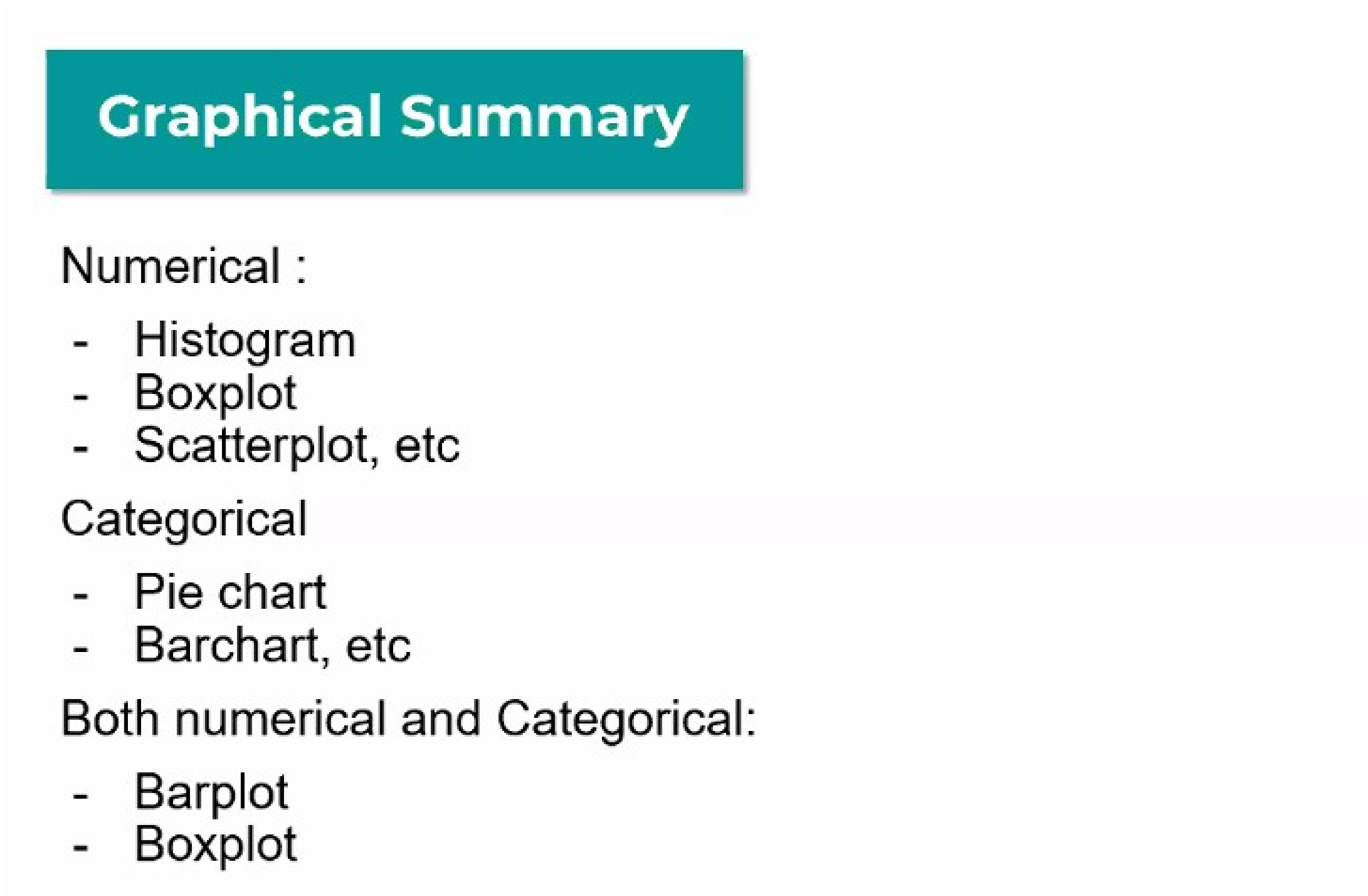


Graphical Summary

Numerical :

- Histogram
- Boxplot
- Scatterplot, etc

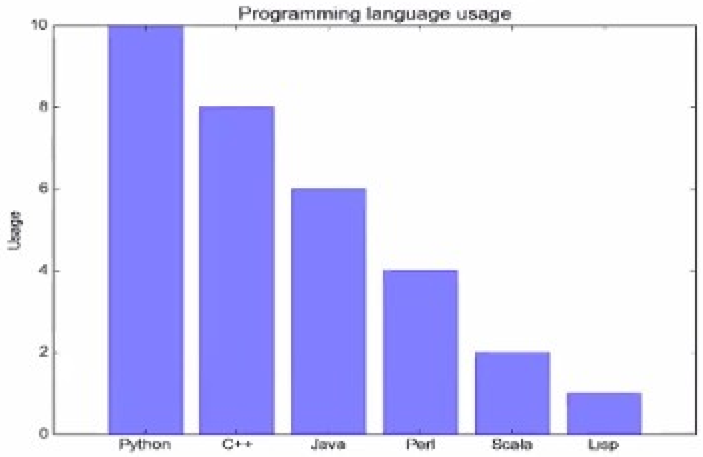
Categorical

- Pie chart
- Barchart, etc

Both numerical and Categorical:

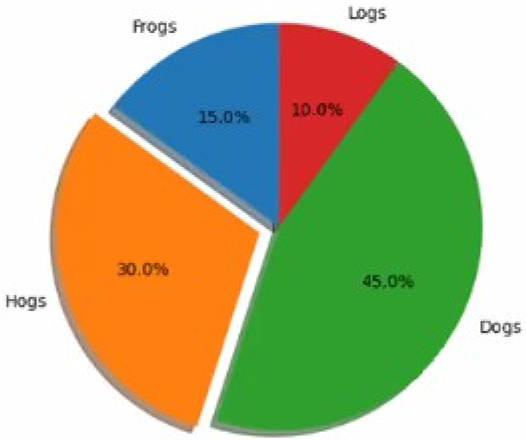
- Barplot
- Boxplot

Bar chart



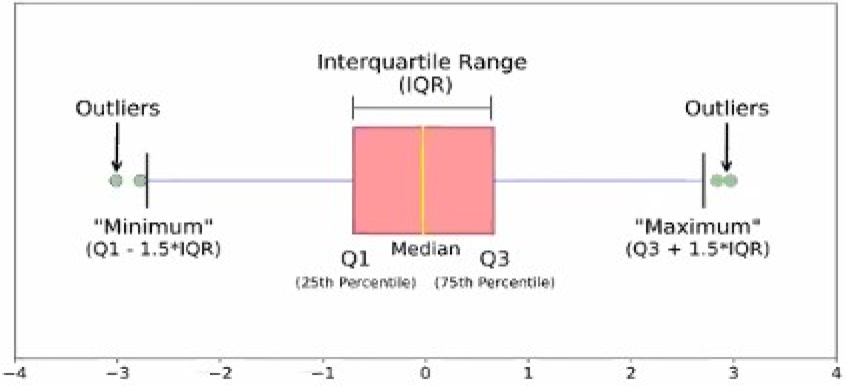
- Represents **categorical data** with rectangular bars. Each bar has a height corresponds to the value it represents. It's useful when we want to **compare** a given numeric value on different **categories**.
- Each category can be consecutive and overlapping
- Can be used to see composition or comparison

Pie chart



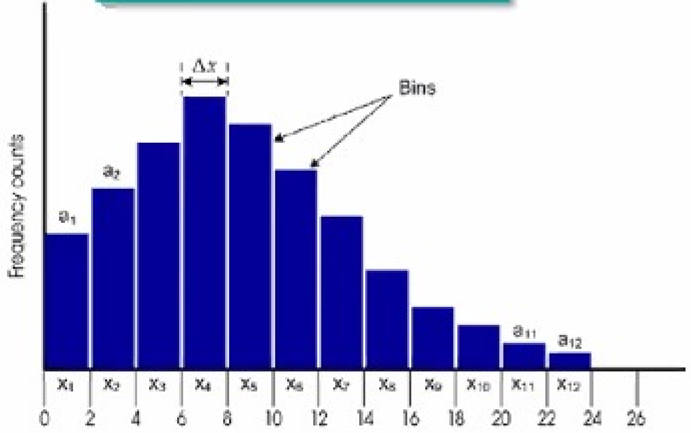
- A circular plot, divided into slices to show numerical proportion of the categorical data. They are widely used in the business world.
- Each category are consecutive and non-overlapping
- Main purpose is composition
- Not recommended if there are too many categories

Boxplot



- Box plot, also called the box-and-whisker plot: a way to show the **distribution of values based on the five-number summary**: minimum, first quartile, median, third quartile, and maximum.
- Can be used to detect anomaly data/outliers

Histogram

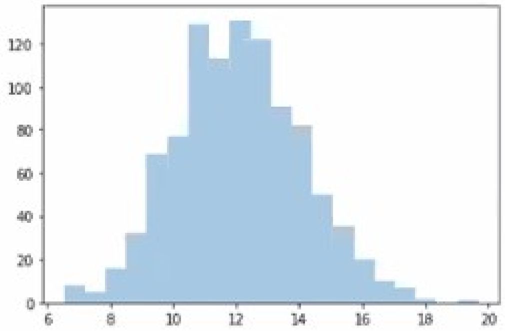


- **Histogram** is an accurate representation of the **distribution of numeric data**.
- A histogram is a graph that uses bars to portray the frequencies or the relative frequencies of the possible outcomes for a quantitative variable.

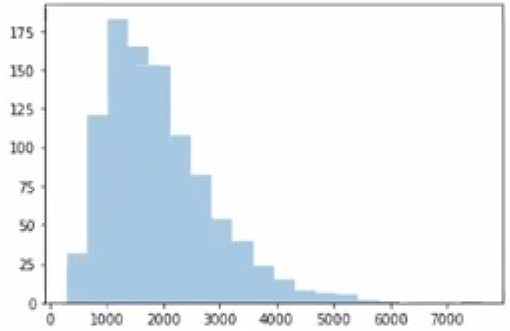
Histogram

Using histogram we can see how data spread.

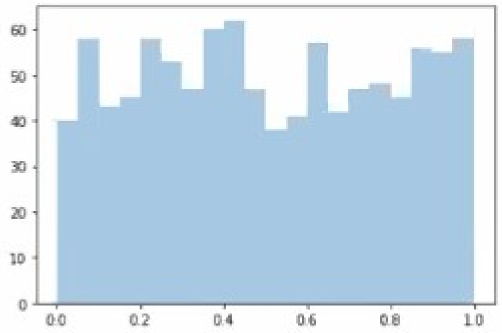
Symmetric or Normally Distributed



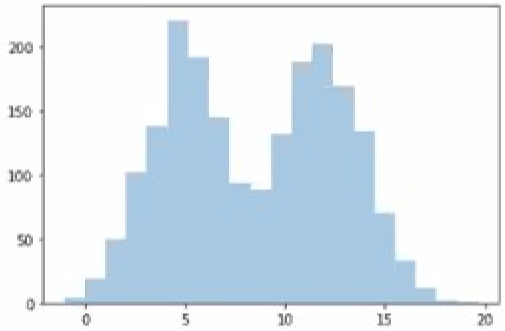
Right Skewed



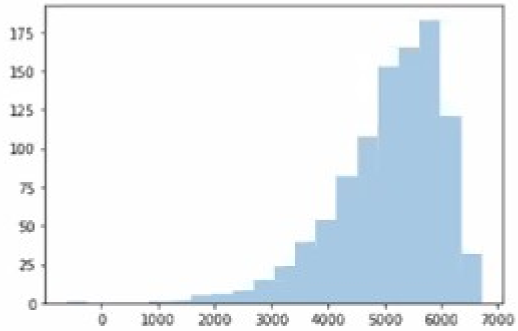
Uniform



Bimodal



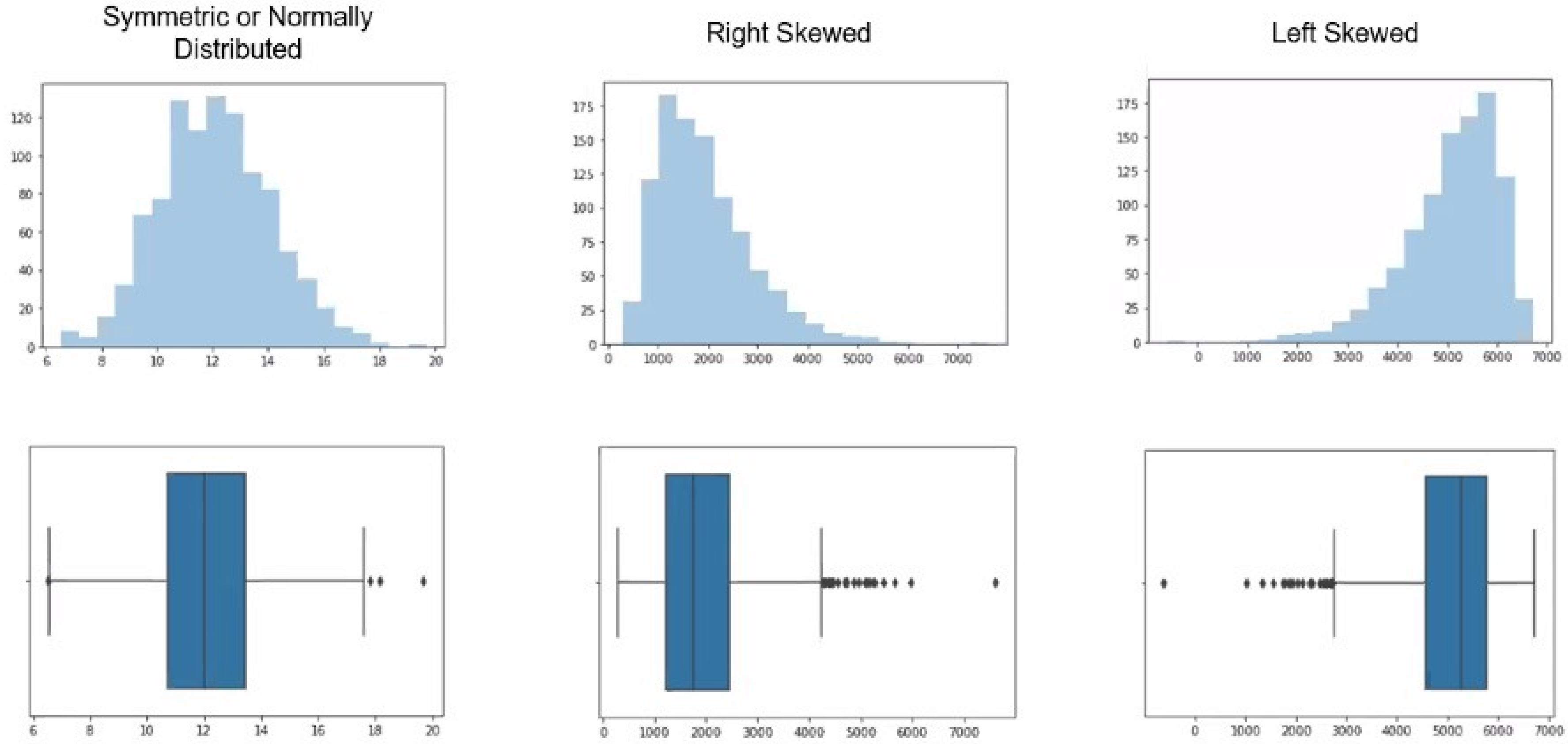
Left Skewed



Purwadhika
Startup and Coding School

Boxplot

Using boxplot we can detect outliers

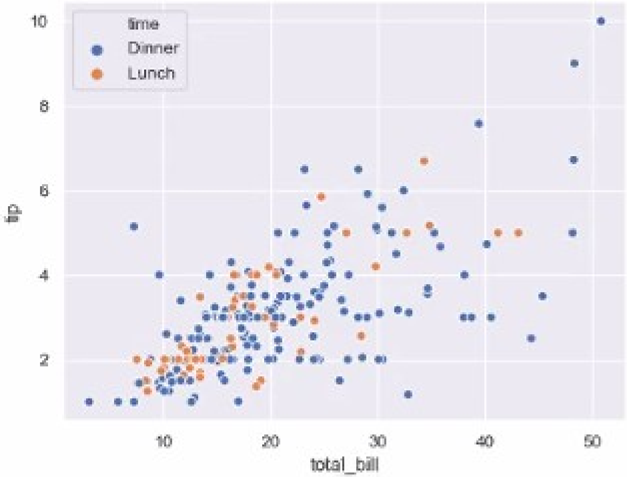


Symmetric or Normally Distributed

Right Skewed

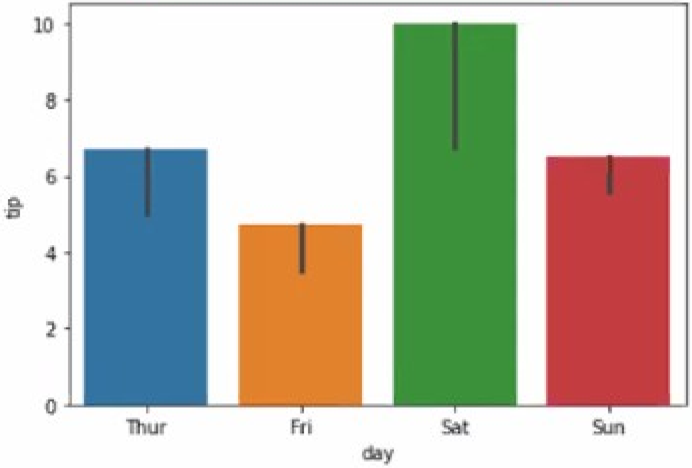
Left Skewed

Scatterplot



- This type of plot shows **all individual data points**. Here, they aren't connected with lines.
- Each data point has the value of the x-axis value and the value from the y-axis values.
- This type of plot can be used to display **trends or correlations**.
- In data science, it shows relationship between **two numerical variables**.

Barplot

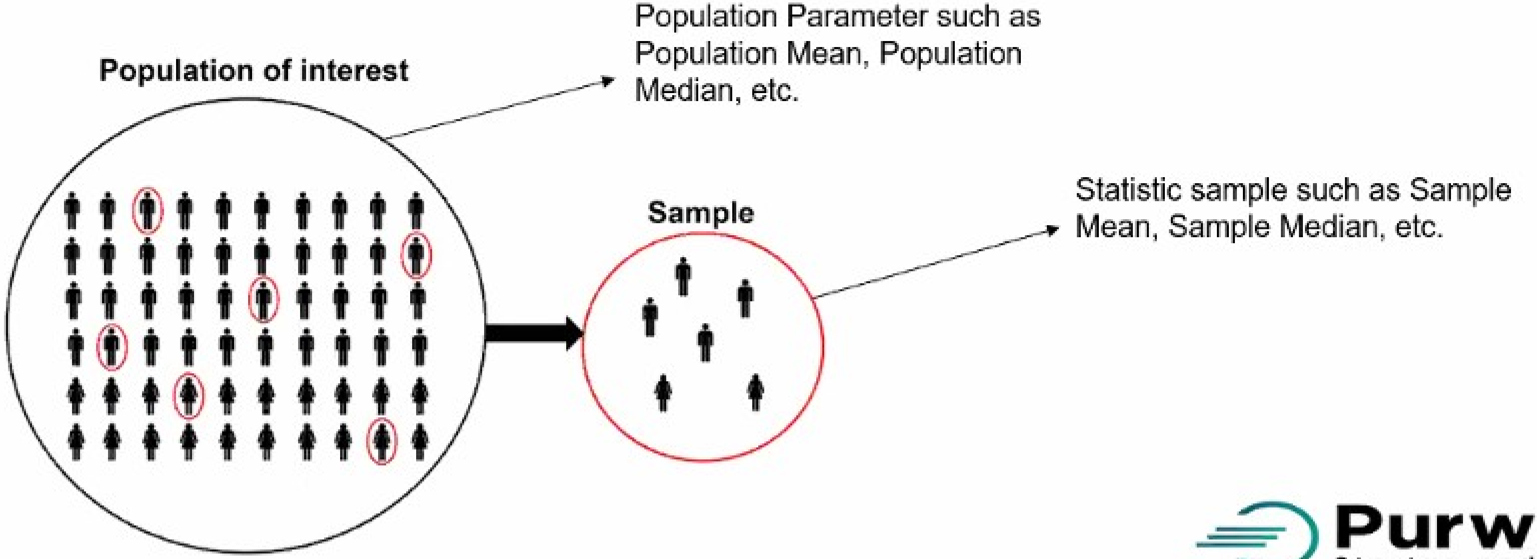


- **Barplot** is a general plot that allows you to aggregate some values in the categorical data based on some function (mean, sum, min, max, std, etc)
- In data science, it shows composition and relationship between **a numerical variables** and **a categorical variables**.

Purwadhika
Startup and Coding School

Statistics and Parameter

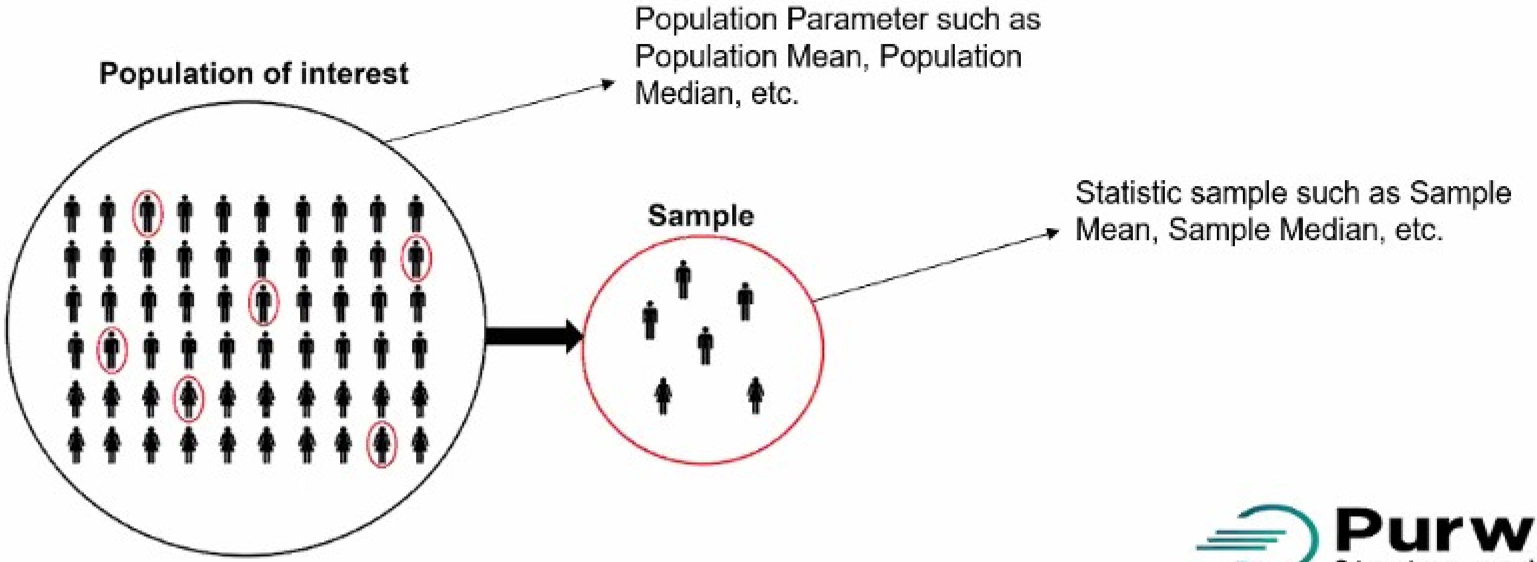
- **A parameter** is a numerical summary of the population. **A statistic** is a numerical summary of a sample taken from the population.
- Population parameter are unknown and sample statistic used to make inference about it



Population of interest
Population Parameter such as Population Mean, Population Median, etc.
Sample
Statistic sample such as Sample Mean, Sample Median, etc.

Statistics and Parameter

- **A parameter** is a numerical summary of the population. **A statistic** is a numerical summary of a sample taken from the population.
- Population parameter are unknown and sample statistic used to make inference about it



Population of interest

Population Parameter such as Population Mean, Population Median, etc.

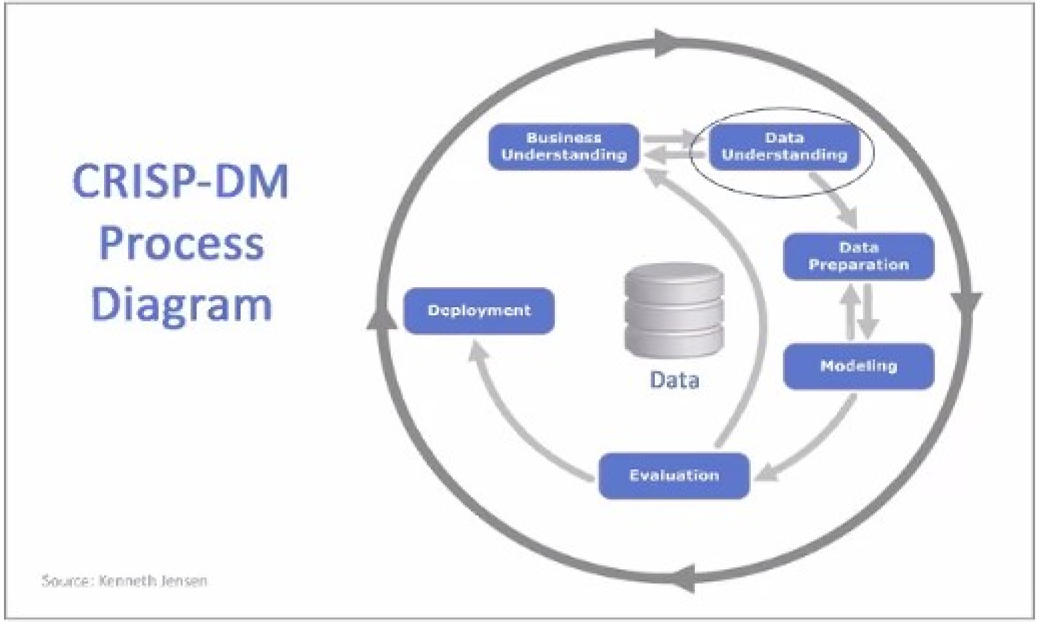
Sample

Statistic sample such as Sample Mean, Sample Median, etc.

Probability Statistic

Outline

- What is probability?
- Probability distribution
  - Discrete
  - Continuous
- Sampling Distribution
- Central Limit Theorem



CRISP-DM Process Diagram

Source: Kenneth Jensen

Probability

Probability study about randomness.

Many aspect in Statistics involve randomness (sampling from observation or assign subjects to treatments, assumption in hypothesis testing)

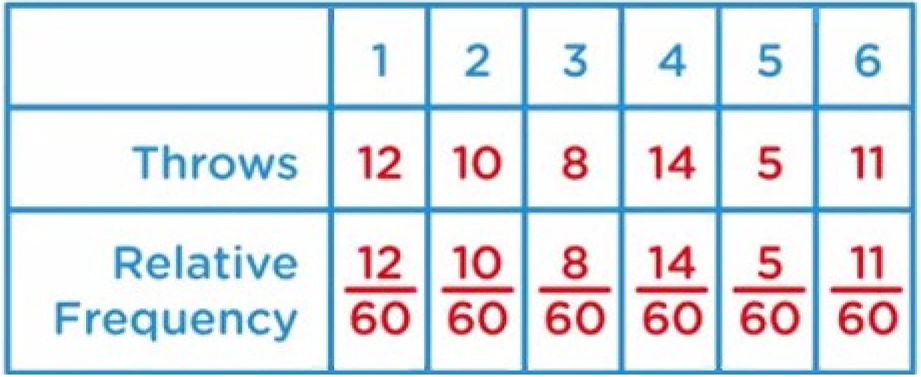
- When we roll a dice we know every possible outcome
  - 1, 2, 3, 4, 5, or 6
- but we don’t know exactly what will occur

The proportion of any outcome (1, 2, 3, 4, 5, or 6) if we roll dice as much as possible. That is an example of probability.

P(any event) ranged from 0 to 1

Rolling Dice

- If the dice is balance the probability of each number occur should be ⅙.
- There is a way to estimate the probability by experimenting.
- We roll dice 60 times and record what is the record. We expect each number show up 10 times. But what we got is



	1	2	3	4	5	6
Throws	12	10	8	14	5	11
Relative Frequency	$\frac{12}{60}$	$\frac{10}{60}$	$\frac{8}{60}$	$\frac{14}{60}$	$\frac{5}{60}$	$\frac{11}{60}$

By looking at the result:

1. does it mean that the probability of each number occur is not equal to ⅙ ?
2. is it surprising we got one of the dice number 5 times ?

How about compute mean from any population?

Let's say we want to take sample (30) from a Purwadhika student population and compute the mean of the age

- Person 1 take sample and the result is 25.4
- Person 2 take sample again, are you sure the result is 25.4 again ?
- How about the next person ?

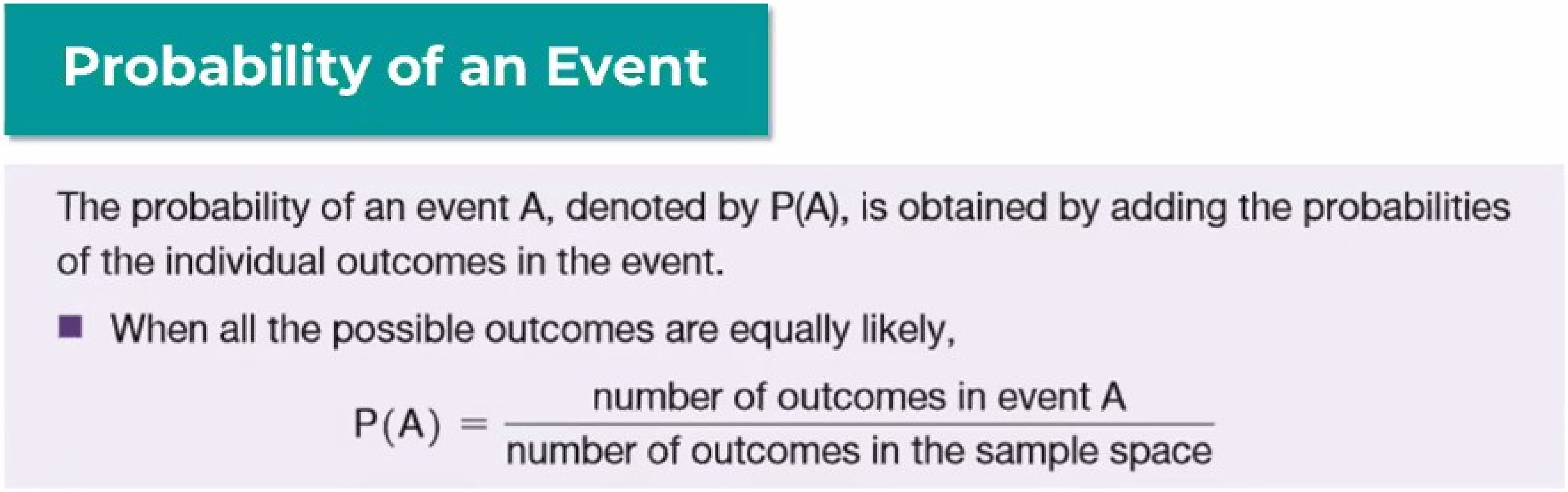
There are so many possibilities. You can say that the possibility is infinite if the population is very large.

Some Terminology in Probability

- When rolling a dice the possible outcome are 1,2,3,4,5, or 6.
- When rolling two dice the possible outcome are (1,1),(1,2),(1,3),....,(5,6), (6,6).
- Every possible outcome is called **sample space**.

Subset of sample space is called **Event**.

- odd number (1,3,5)
- prime number (2,3,5)
- total number of dice = 4 ((2,2), (1,3), (3,1))



Probability of an Event

The probability of an event A, denoted by P(A), is obtained by adding the probabilities of the individual outcomes in the event.

- When all the possible outcomes are equally likely,

$$P(A) = \frac{\text{number of outcomes in event A}}{\text{number of outcomes in the sample space}}$$

For example:
event A = odd number occur (1, 3, 5)

P(A) = 3/6

Probability Distribution

Probability distribution is a way to summarize every possible outcome and their probability.

There are two types of probability distribution:

- **Discrete Value, Probability Mass Distribution**

  Ex. Dice number

- **Continuous Value, Probability Density Distribution**

  Ex. mean from samples.

Some Terminology in Probability Distribution

From the definition we know that: Probability distribution is a way to summarize every possible outcome and their probability.

Every possible outcome is quantified into numerical value that is called **random variable**.

Ex: when we toss two coins. Here is the sample space :

(Face, Tail), (Tail, Face), (Face, Face), (Tail, Tail)

Let’s say a random variable X = #number of face occur. it will be 0 1 and 2

So, X is random variable with every possible values are 0, 1, or 2

Probability Mass Distribution Illustration

two-tosses	head-head	head-tail	tail-head	tail-tail
probability	0.25	0.25	0.25	0.25

X = #numer of head occur

0, 1, 2 (discrete value)

This is the **Probability Mass Distribution**

X	2	1	0
probability	0.25	0.5	0.25

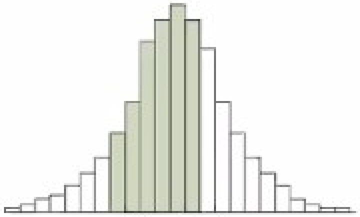
Rules:

1. All add up to 1
2. Must be disjoint (mutually exclusive)
3. Each should be between 0 and 1

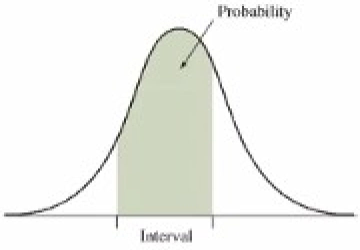
Probability Density Distribution Illustration

Probability distribution is also a way to summarize every possible outcome and their probability but when the **random variable** is **continuous**.

For example, probability to measure someone height exactly at 175 cm (no more or less than few picometer) is 0 or at least close to 0.



Many intervals



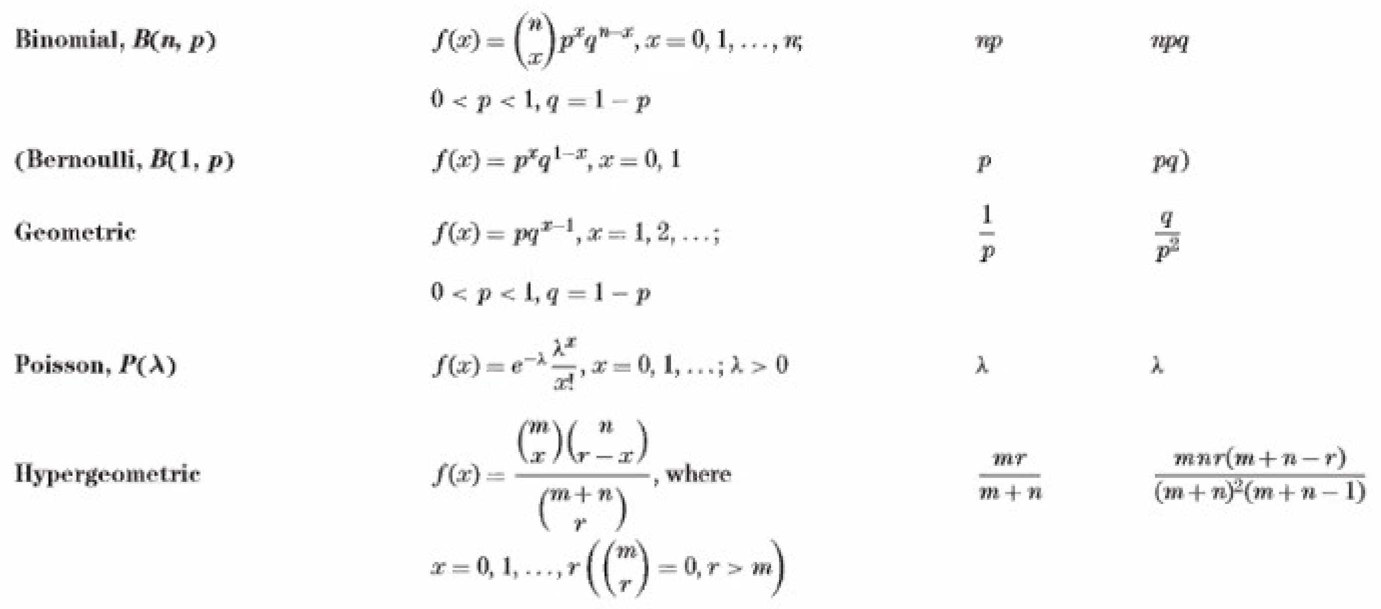
Smooth curve approximation

So, are measured based on the value range, not just one value. In the left picture we divide height into so many interval and compute the probability in each interval (similar like histogram right?)

In the next picture, it is an illustration of the smooth version of the histogram (infinite interval) which is drawn using a curve.

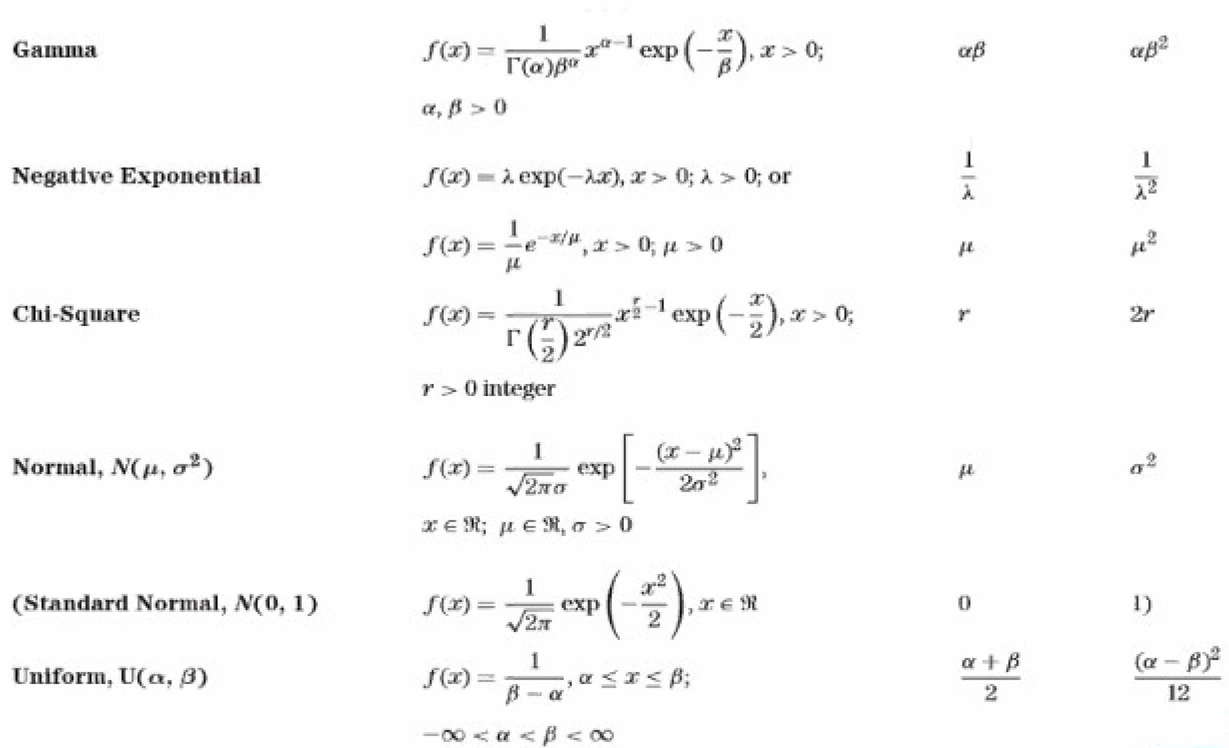
- each interval 175 - 180, 180 - 185 has a value of probability between 0 and 1
- an interval contain all possible value has a value of probability equal to 1.

Probability Mass Function (PMF) Example



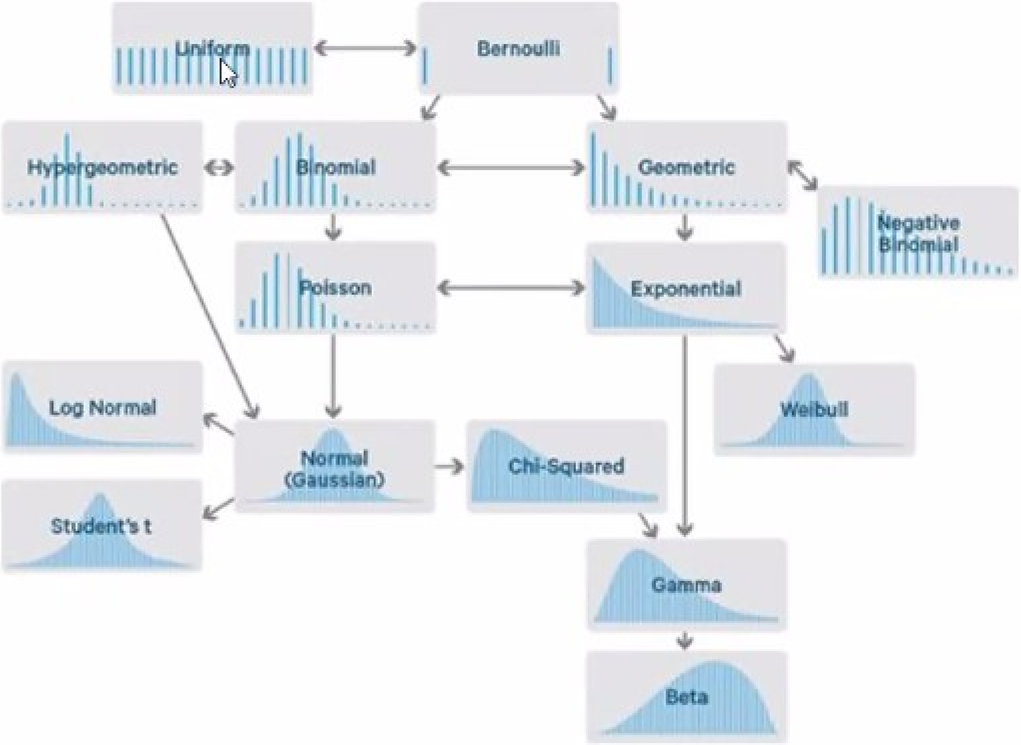
Binomial, $B(n, p)$	$f(x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, \ldots, n;$	np	npq
	$0 < p < 1, q = 1 - p$		
(Bernoulli, $B(1, p)$	$f(x) = p^x q^{1-x}, x = 0, 1$	p	$pq)$
Geometric	$f(x) = pq^{x-1}, x = 1, 2, \ldots;$	$\frac{1}{p}$	$\frac{q}{p^2}$
	$0 < p < 1, q = 1 - p$		
Poisson, $P(\lambda)$	$f(x) = e^{-\lambda} \frac{\lambda^x}{x!}, x = 0, 1, \ldots; \lambda > 0$	λ	λ
Hypergeometric	$f(x) = \frac{\binom{m}{x}\binom{n}{r-x}}{\binom{m+n}{r}}$, where	$\frac{mr}{m+n}$	$\frac{mnr(m+n-r)}{(m+n)^2(m+n-1)}$
	$x = 0, 1, \ldots, r \left(\binom{m}{r} = 0, r > m\right)$		

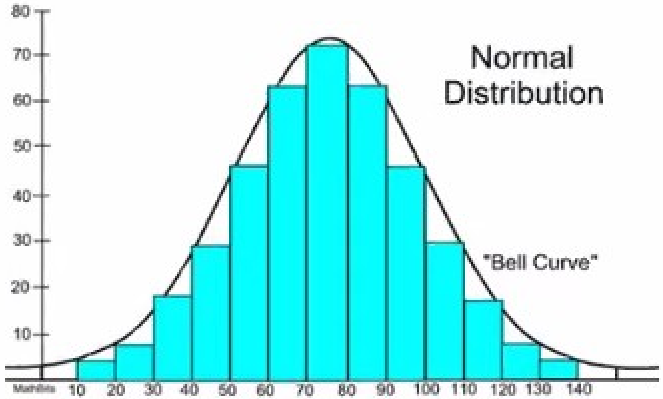
Probability Density Function (PDF) Example



Gamma	$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} \exp\left(-\frac{x}{\beta}\right), x > 0;$	$\alpha\beta$	$\alpha\beta^2$
	$\alpha, \beta > 0$		
Negative Exponential	$f(x) = \lambda \exp(-\lambda x), x > 0; \lambda > 0;$ or	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
	$f(x) = \frac{1}{\mu} e^{-x/\mu}, x > 0; \mu > 0$	μ	μ^2
Chi-Square	$f(x) = \frac{1}{\Gamma\left(\frac{r}{2}\right) 2^{r/2}} x^{\frac{r}{2}-1} \exp\left(-\frac{x}{2}\right), x > 0;$	r	$2r$
	$r > 0$ integer		
Normal, $N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right],$	μ	σ^2
	$x \in \Re; \ \mu \in \Re, \sigma > 0$		
(Standard Normal, $N(0, 1)$	$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), x \in \Re$	0	$1)$
Uniform, U(α, β)	$f(x) = \frac{1}{\beta - \alpha}, \alpha \le x \le \beta;$	$\frac{\alpha + \beta}{2}$	$\frac{(\alpha - \beta)^2}{12}$
	$-\infty < \alpha < \beta < \infty$		

Statistical Distributions

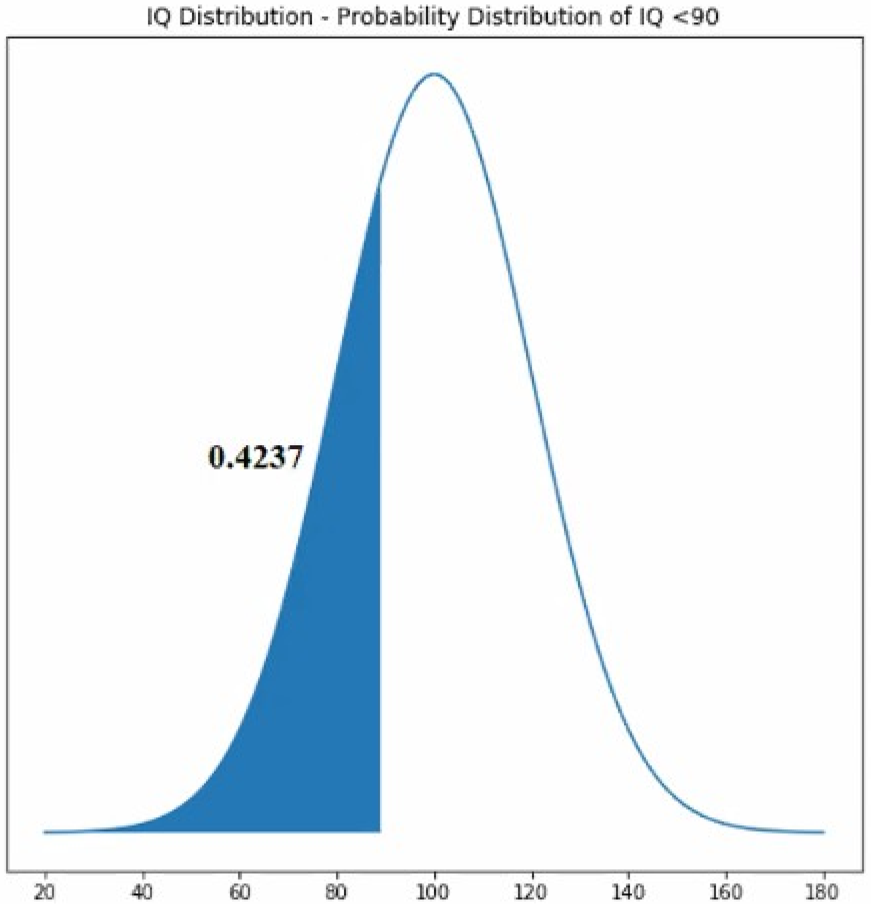




Purwadhik
Startup and Coding Scho

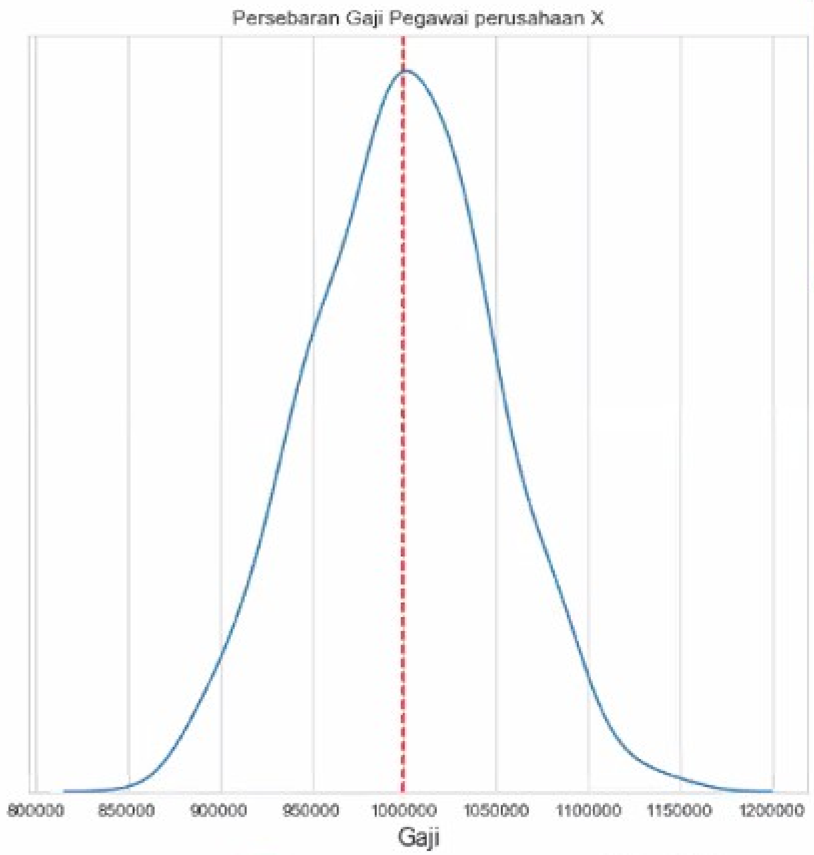
PDF

- In picture beside, we could for example want to know the probability to acquire a sample with IQ less than 90.
- We can measure the probability by measure the total area of the colored area or *area under the curve*.
- You could check this article to know more about the measurement.
- From the data, if we take a random sample, the probability to get a sample with IQ less than 90 is 42.37%.
- $P(X < 90) = 0.4237$
- X = random variable IQ



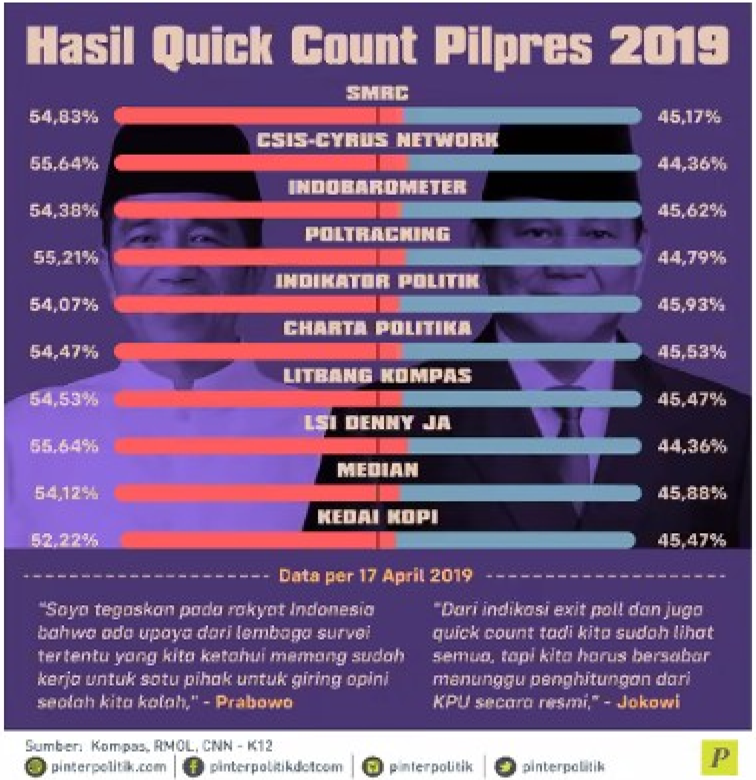
Probability Distribution

- *Probability Distribution* is useful for inferring the most probable event to be happen, likelihood of an event to occur, and the likelihood interval for event to occur. For example, company X employ 1000 employees and I want to take a sample of the salary for each employee. If I create the distribution plot of the company X employees salary, it could be visualized in the image beside.
- Picture beside giving us the *Probability Distribution* of the salary likelihood in the company X. We know if we randomly take a salary sample from company X, we probably would get a higher chance to acquire value that close to the red line or the mean. In other hand, the probability for us to get the random sample with salary value less than 900000 or more than 1100000 is small.



Sampling Distribution

- Estimation from quick count is done based on random sample.
- Estimation result vary from sample to sample.
- We can use histogram to infer the shape distribution of estimation result.
- The shape of the distribution estimation result is called sampling distribution.



Data per 17 April 2019

"Saya tegaskan pada rakyat Indonesia bahwa ada upaya dari lembaga survei tertentu yang kita ketahui memang sudah kerja untuk satu pihak untuk giring opini seolah kita kalah," - **Prabowo**

"Dari indikasi exit poll dan juga quick count tadi kita sudah lihat semua, tapi kita harus bersabar menunggu penghitungan dari KPU secara resmi," - **Jokowi**

Sumber: Kompas, RMOL, CNN - K12

Central Limit Theorem

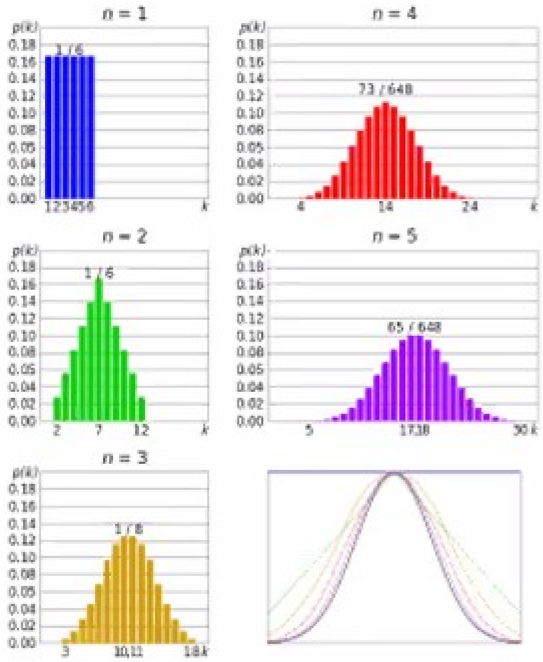
In the illustration before, we that sample taken 10 times and we have 10 estimation result.

For each sample, if **sample size large** enough (n > 30). **Distribution of the estimation** result from each sample should follow **normal** distribution.

This law also applied:

- to any statistics like mean, sum, median, etc (in this case proportion).
- when the population distribution is not normal.

Here's what the Central Limit Theorem is saying, graphically. The picture below shows one of the simplest types of test: rolling a die. The **more times you roll the die**, the more likely the shape of the distribution of the means tends to look like a **normal distribution graph**.



Why is Central Limit Theorem is so important ?

- Hypothesis testing like t-test, z-test, Anova F-test each of them need normality assumption to infer any population parameter using statistics
- With central limit theorem we know that any statistics if the **sample size large enough** the sampling distribution will follow normal distribution
- So, if we do any test statistics (t-test, z-test, F-test) and the **sample size** is **large** enough the result will be valid.