Big Data Fundamentals and Applications

Statistical Analysis (I)

Descriptive Statistics – Indicators

Asst. Prof. Chan, Chun-Hsiang

Master program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Intelligent Computing and Big Data, Chung Yuan Christian University, Taoyuan, Taiwan
Undergraduate program in Applied Artificial Intelligence, Chung Yuan Christian University, Taoyuan, Taiwan

Outlines

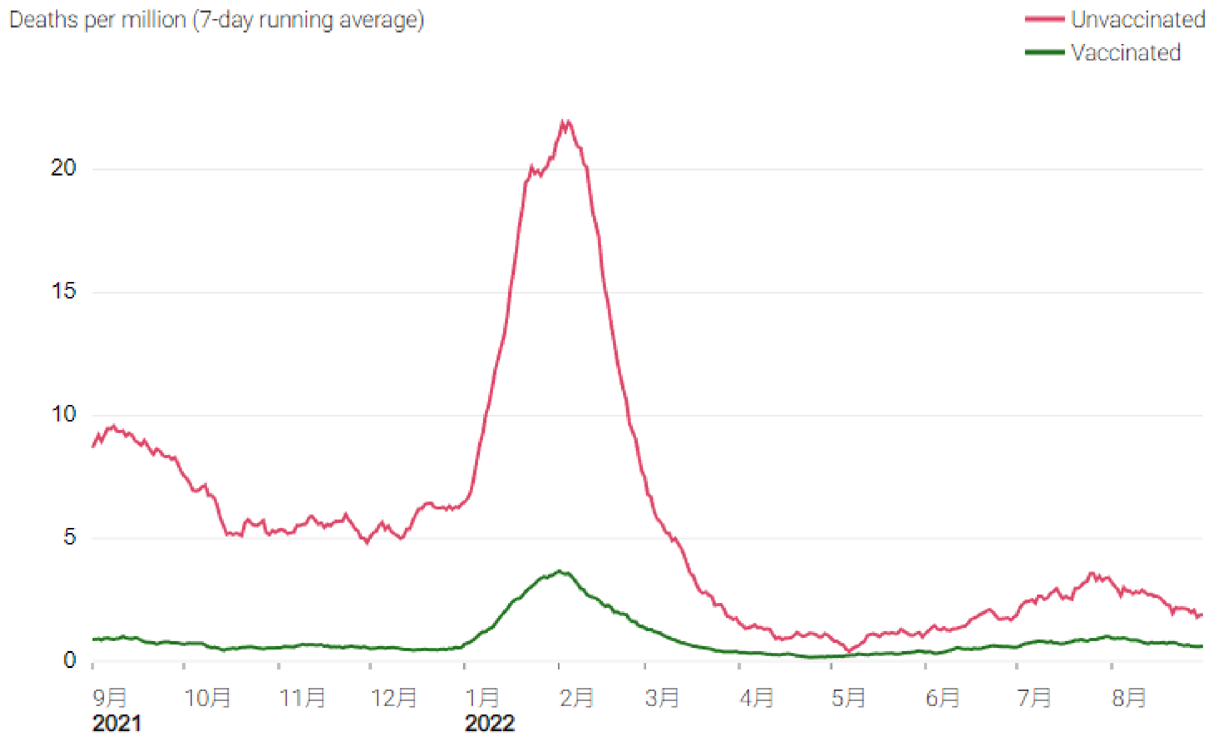
1. Introduction to Statistics
2. Descriptive Statistics
3. Central Tendency
4. Dispersion
5. Heterogeneity
6. Shape
7. Question Time

Introduction to Statistics

- **Statistics** plays an vital role in data science.
- In some cases, we may directly conduct data exploration approach (e.g., data visualization) to understand the distribution of your dataset, and even differentiate the characteristics between different features.
- However, we always face a dilemma that we cannot directly determine whether one feature is significantly different from another. Therefore, inferential statistics quantitatively present the difference between one distribution to another through a hypothesis testing.
- Due to time limitation, we will focus on descriptive statistics in the first two weeks, then inferential statistics.

Descriptive Statistics

- Descriptive statistics are used to describe the characteristics of data from a distribution perspective, including center tendency, dispersion, shape, heterogeneity, and graphs.



Source: https://covid19.ca.gov/state-dashboard/

Central Tendency – Indicators

Indicators	Meanings
Mean	The expectation/average in a set of data
	Arithmetic mean (AM)
	Geometric mean (GM)
	Harmonic mean (HM)
Mid-range	The arithmetic mean of the maximum and minimum values of the data set
Median	The center value in a set of data
Mode	The most often value in a set of data
Sum	The total value of the data

Source: https://en.wikipedia.org/wiki/Mean

Central Tendency – Q1

Question 1

Give one practical example for each statistic (i.e., mean, median, mode, and sum) and calculate their value by self-defined function.

Source: https://en.wikipedia.org/wiki/Mean

Central Tendency – Mean

Arithmetic mean (AM)	
The arithmetic mean (or simply mean) of a list of numbers, is the sum of all of the numbers divided by the number of numbers.	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$
Geometric mean (GM)	
The geometric mean is an average that is useful for sets of positive numbers, that are interpreted according to their product (as is the case with rates of growth) and not their sum (as is the case with the arithmetic mean)	$\bar{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = (x_1 x_2 \cdots x_n)^{\frac{1}{n}}$
Harmonic mean (HM)	
The harmonic mean is an average which is useful for sets of numbers which are defined in relation to some unit, as in the case of speed (i.e., distance per unit of time)	$\bar{x} = n\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1}$

Source: https://en.wikipedia.org/wiki/Mean

Central Tendency

Question 2

Design a script to calculate and test the regularity (sorting by its value) of average values based on three mean definitions, including arithmetic, geometric, and harmonic mean. You may obtain three testing datasets from the internet or generating from random variables. Please notice that the testing data should be representative; otherwise, it will be meaningless.

Source: https://en.wikipedia.org/wiki/Mean

Central Tendency – Mid-range & Median

- **Mid-range** represents the center value of the dataset based on minimum and maximum value.

$$mid - range = \frac{\min(x_i) + \max(x_i)}{2}, \forall i > 0$$

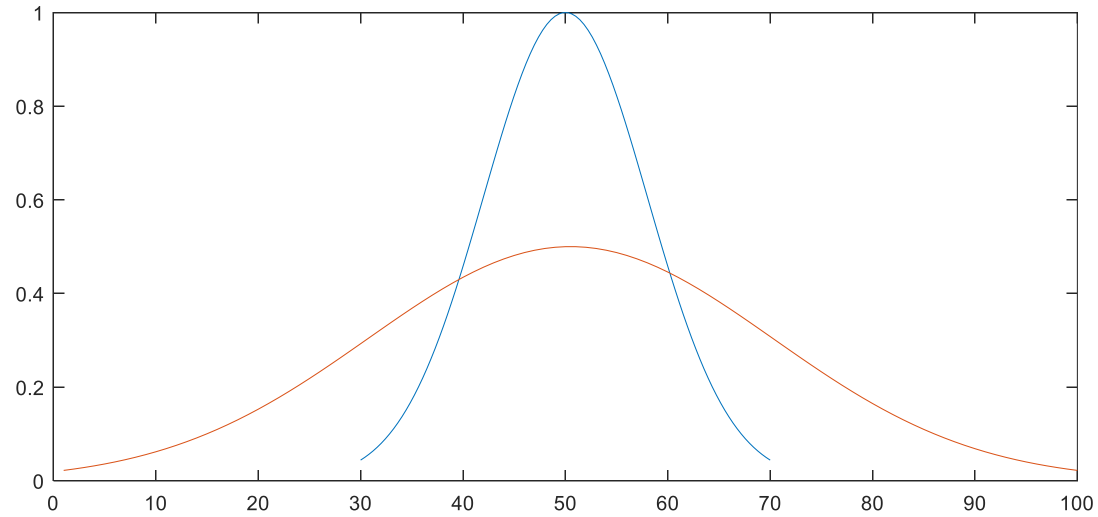
- Unlike mid-range, **median** is also a common statistic to describe the center location of the dataset based on values.
- 1,2,3,4,5,6,7 ➔ median = 4
- 1,2,3,4,5,6 ➔ median = ?

Central Tendency – Mode & Sum

- **Mode** is usually used to present the concept of consensus. For instance, we have a meeting to decide the catering company for our international conference; therefore, we need to vote for your favorite company. The catering company with the highest number of votes will be selected for our conference. The physical meaning of the highest number of votes is the same as the definition of mode.
- Sometimes, we want to know the overall performance between features or datasets; therefore, we may obtain the **sum**mation of all values together for comparison.

Dispersion

- In most cases, center tendency cannot represent the distribution or characteristics of dataset because of its variation. The figure provided below demonstrates that two distributions have the same mean but their variations are quite different. Therefore, if you only observe these datasets without variation, then you will obtain a biased explanation.



Dispersion - Indicators

Indicator	Equation $X = \{x_1, x_2, \ldots, x_n\}$
Standard deviation	$\sigma = \sqrt{\frac{(x_i - \bar{x})^2}{n}}$
Interquartile range (IQR)	$IQR = Q3 - Q1$
Maximum and minimum	$\max(X),\ \min(X)$
Range	$range = \max(X) - \min(X)$
Average absolute deviation (AAD) Mean absolute deviation (MAD)	$AAD = \frac{1}{n}\sum_{i=1}^{n} \lvert x_i - \bar{x} \rvert$
Median absolute deviation (MAD)	$MAD = median(\lvert x_i - median(X) \rvert)$

Dispersion – Dimensionless

- All descriptive statistics are affected by the sample sizes or unit.
- To overcome this dilemma, we can adopt **dimensionless quantity** concept to measure the dispersion characteristics of the dataset.

Coefficient of Variance (CV)	Quartile Coefficient of Dispersion	Variance	Variance-to-mean Ratio (VMR) [1]
$CV = \frac{s}{\bar{x}}$	$\frac{Q_3 - Q_1}{Q_3 + Q_1}$	$var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$	$D = \frac{s^2}{\bar{x}}$

[1] index of dispersion, dispersion index, coefficient of dispersion, relative variance, or variance-to-mean ratio (VMR)

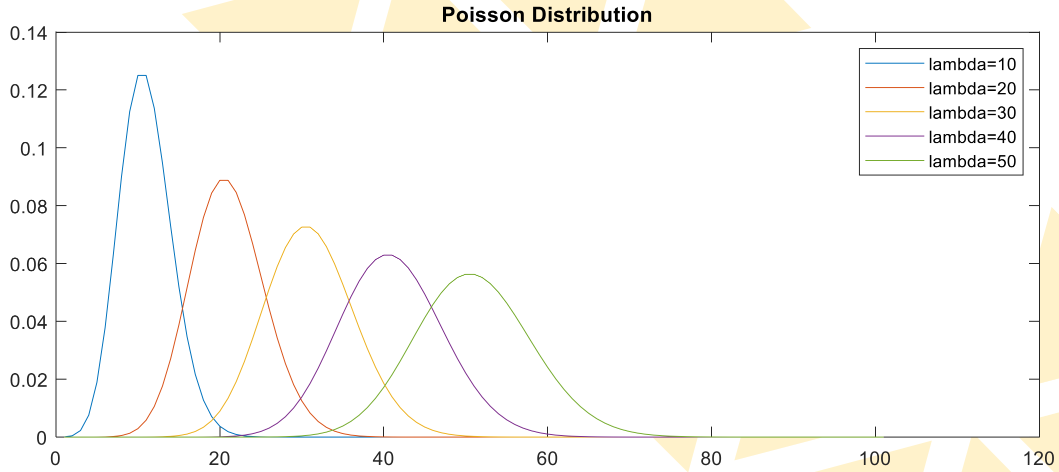
Dispersion – Dimensionless

- **Variance-to-mean Ratio (VMR)**

$$D = \frac{s^2}{\bar{x}}$$

Constant random variable	VMR = 0	not dispersed
Binomial distribution	0 < VMR < 1	under-dispersed
Poisson distribution	VMR = 1	
Negative binomial distribution	VMR > 1	over-dispersed

Poisson Distribution

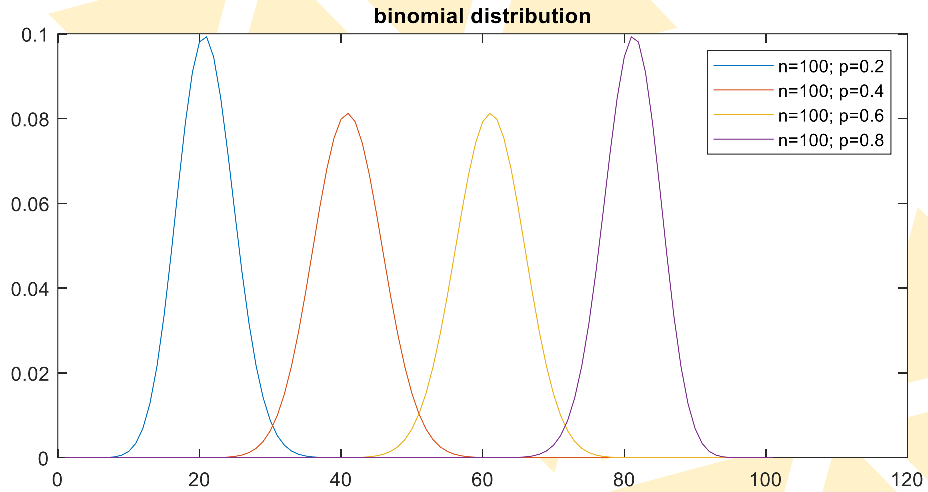


- **From Wiki:**

The **Poisson distribution** is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant mean rate and independently of the time since the last event.

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Binomial Distribution



- **From Wiki:**

  The **binomial distribution** with Indicators n and p is the discrete probability distribution of the number of successes in a sequence of n independent experiments, each asking a yes–no question, and each with its own Boolean-valued outcome: success (with probability p) or failure (with probability $q = 1 - p$).

$$\Pr(X = x) = \binom{n}{k} p^k (1 - p)^{n-k},$$

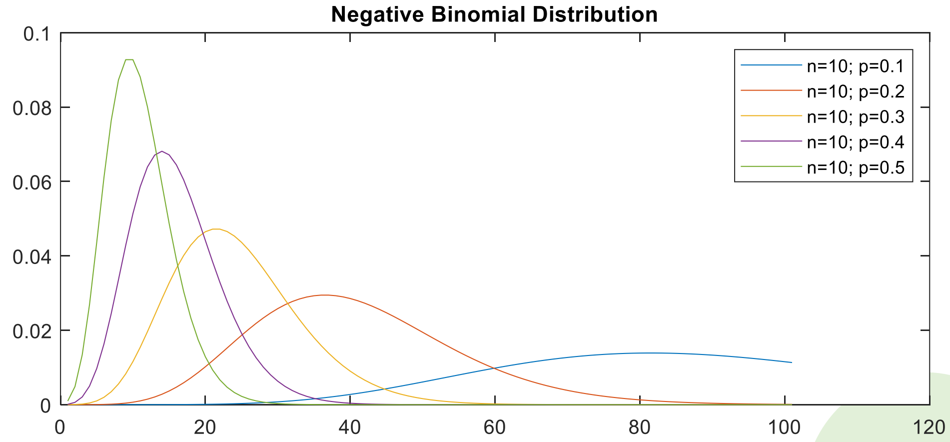
$$where \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Negative Binomial Distribution

- **From Wiki:**

  The **negative binomial distribution** is a discrete probability distribution that models the number of failures (denoted k) in a sequence of independent and identically distributed Bernoulli trials before a specified (non-random) number of successes (denoted r) occurs.

$$\Pr(X = k) = \binom{k + r + 1}{r - 1} p^r (1 - p)^k$$



Dispersion – Variance

Question 3

The variance of random variable X is the expected value of the squared deviation from the mean of X. $\mu = E[X]$:

$$Var(X) = Cov(X, X) = E[(X - \mu)^2]$$

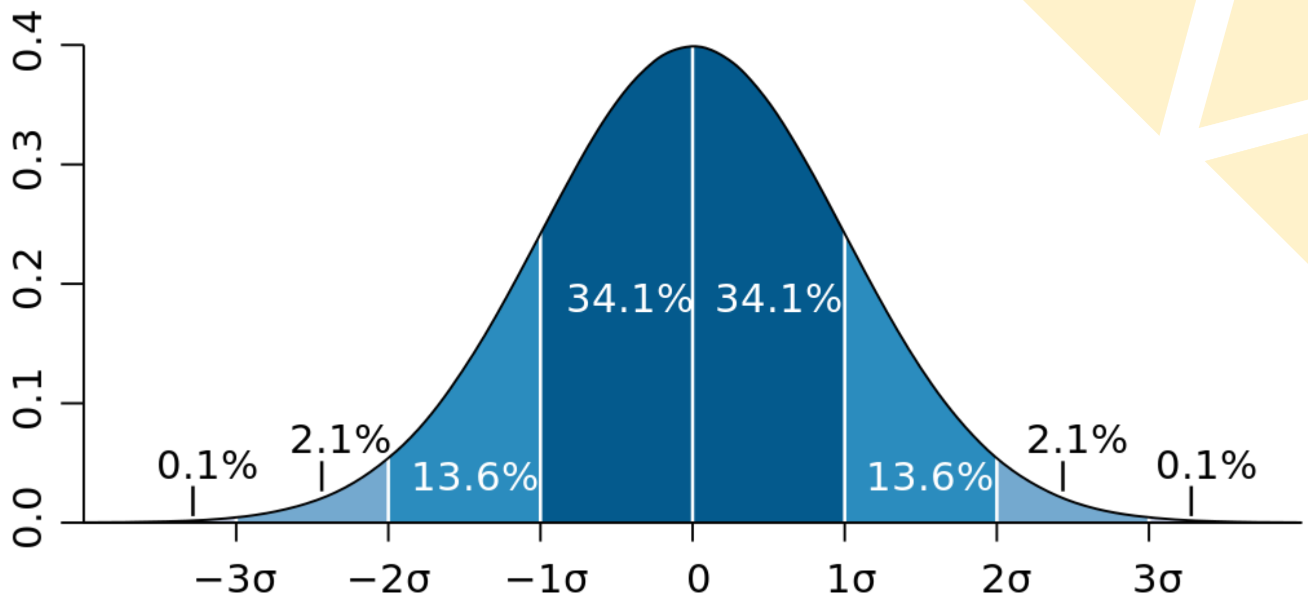
Please expand the variance to the simplest form.

Percentile in Normal Distribution

- For a very large population following a normal distribution, it might be plotted as right-hand-side figure.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- We can use standard deviation to present the percentile.



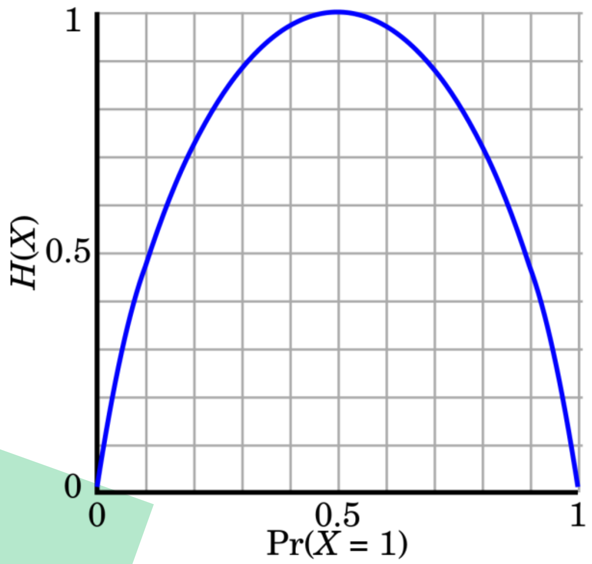
Heterogeneity

- **Heterogeneity** is one of the crucial features to describe the internal differences. For example, there are 100 people in the party A, where 50% are doctors, 20% are sales, 10% are engineers, 10% are consultants, and 10% are secretaries. In the party B, all participants are doctors. How do you quantitatively describe the job distribution differences between party A and party B?
- Here, we will introduce three common indictors: (information) entropy, Gini coefficient, and Herfindahl-Hirschman Index

Entropy

- Entropy (information entropy or Shannon entropy) is a mathematical form to demonstrate the heterogeneity between samples.

$$H(X) := -\sum_{x \in X} p(x) log_b p(x) = \mathbb{E}[-\log p(X)], where\ b = 2, e, or\ 10.$$



Question 5

What do you observe the relationship between probability and entropy from the left-hand-side figure?

Source: https://en.wikipedia.org/wiki/Entropy_(information_theory)

Gini Coefficient

- **From Wiki:**

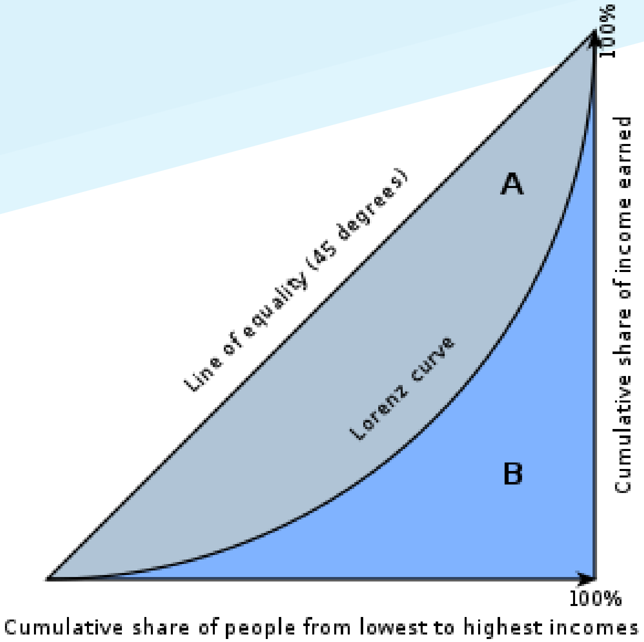
  The **Gini coefficient** is an index for the degree of inequality in the distribution of income/wealth, used to estimate how far a country's wealth or income distribution deviates from an equal distribution.

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{2\sum_{i=1}^{n}\sum_{j=1}^{n} x_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{2n\sum_{j=1}^{n} x_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{2n^2\bar{x}},$$

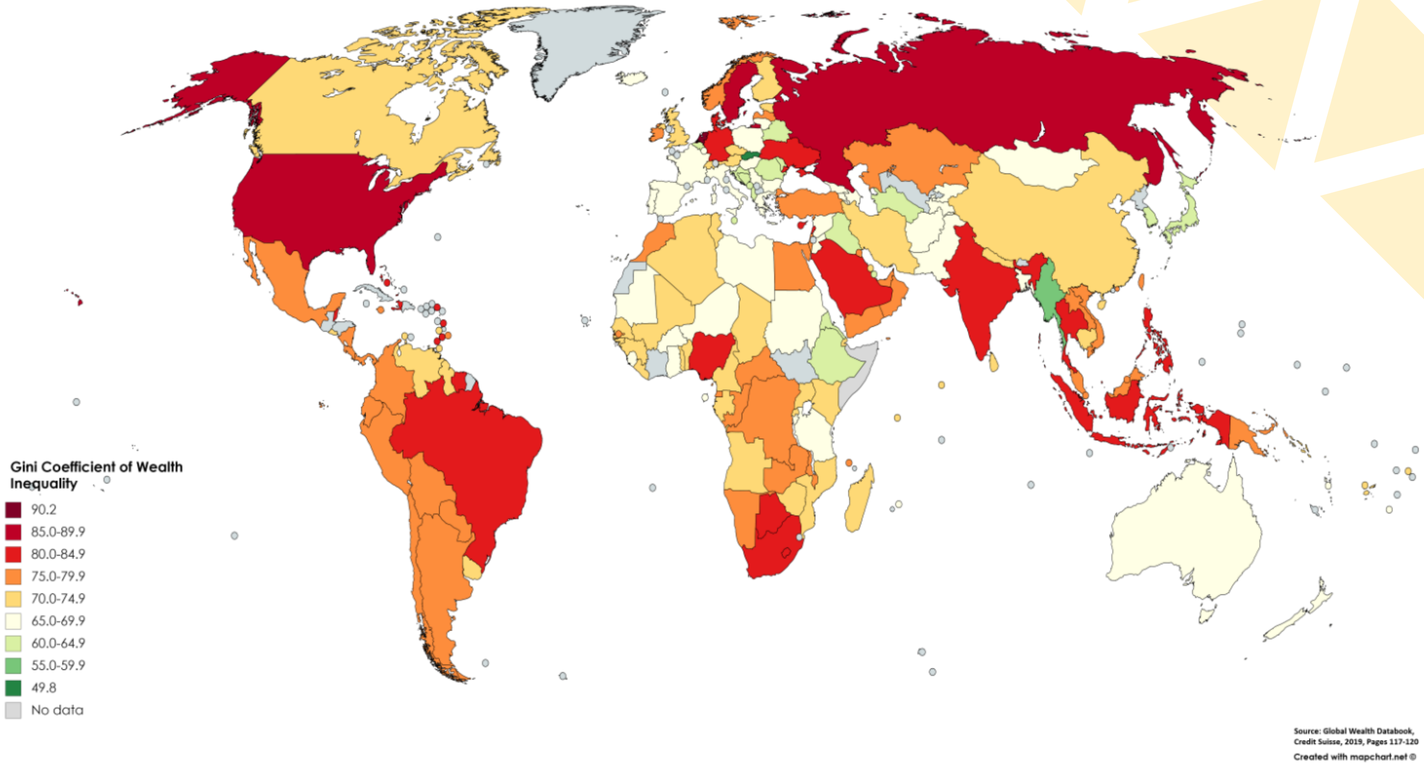
$$G = \frac{1}{2\mu}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(x)p(y)|x-y|dxdy$$

Source: https://en.wikipedia.org/wiki/Gini_coefficient

Gini Coefficient



Graphical representation of the Gini coefficient: The graph shows that the Gini coefficient is equal to the area marked A divided by the sum of the areas marked A and B, that is, Gini = A/(A + B). It is also equal to 2A and to 1 − 2B due to the fact that A + B = 0.5 (since the axes scale from 0 to 1).



Source: https://en.wikipedia.org/wiki/Gini_coefficient

Gini Coefficient

	Country	Subregion	Region	Gini[5][6] %	Gini[5][6] Year
1	Afghanistan	Southern Asia	Asia		
	World				
2	Slovakia	Eastern Europe	Europe	23.2	2019
3	Belarus	Eastern Europe	Europe	24.4	2020
4	Slovenia	Southern Europe	Europe	24.4	2019
5	Armenia	Western Asia	Asia	25.2	2020
6	Czech Republic	Eastern Europe	Europe	25.3	2019
7	Ukraine	Eastern Europe	Europe	25.6	2020
8	Moldova	Eastern Europe	Europe	26.0	2019
9	United Arab Emirates	Western Asia	Asia	26.0	2018
10	Iceland	Northern Europe	Europe	26.1	2017
11	Belgium	Western Europe	Europe	27.2	2019
12	Algeria	Northern Africa	Africa	27.6	2011
13	Denmark	Northern Europe	Europe	27.7	2019
14	Finland	Northern Europe	Europe	27.7	2019
15	Norway	Northern Europe	Europe	27.7	2019
16	Kazakhstan	Central Asia	Asia	27.8	2018
17	East Timor	South-eastern Asia	Asia	28.7	2014
18	Croatia	Southern Europe	Europe	28.9	2019
19	Kosovo	Eastern Europe	Europe[a]	29.0	2017
47	Portugal	Southern Europe	Europe	32.8	2019
48	Tunisia	Northern Africa	Africa	32.8	2015
49	Japan	Eastern Asia	Asia	32.9	2013
50	Bosnia and Herzegovina	Southern Europe	Europe	33.0	2011
51	North Macedonia	Southern Europe	Europe	33.0	2018
52	Greece	Southern Europe	Europe	33.1	2019
53	Switzerland	Western Europe	Europe	33.1	2018
54	Canada	Northern America	Americas	33.3	2017
55	Taiwan	Eastern Asia	Asia	33.6	2014
56	Azerbaijan	Western Asia	Asia	33.7	2008
57	Jordan	Western Asia	Asia	33.7	2010
58	Tajikistan	Central Asia	Asia	34.0	2015
59	Luxembourg	Western Europe	Europe	34.2	2019
60	Sudan	Northern Africa	Africa	34.2	2014
61	Australia	Australia, New Zealand	Oceania	34.3	2018
62	Spain	Southern Europe	Europe	34.3	2019
63	Georgia	Western Asia	Asia	34.5	2020
64	Latvia	Northern Europe	Europe	34.5	2019
144	Singapore	South-eastern Asia	Asia	45.9	2017
145	Nicaragua	Central America	Americas	46.2	2014
146	Cameroon	Middle Africa	Africa	46.6	2014
147	Burkina Faso	Western Africa	Africa	47.3	2018
148	Ecuador	South America	Americas	47.3	2020
149	Honduras	Central America	Americas	48.2	2019
150	Guatemala	Central America	Americas	48.3	2014
151	Brazil	South America	Americas	48.9	2020
152	Congo	Middle Africa	Africa	48.9	2011
153	Costa Rica	Central America	Americas	49.3	2020
154	Belize	Central America	Americas	49.8	2014
155	Panama	Central America	Americas	49.8	2019
156	Zimbabwe	Eastern Africa	Africa	50.3	2019
157	Saint Lucia	Caribbean	Americas	51.2	2016
158	Angola	Middle Africa	Africa	51.3	2018
159	Botswana	Southern Africa	Africa	53.3	2015
160	Hong Kong	Eastern Asia	Asia	53.9	2016
161	Mozambique	Eastern Africa	Africa	54.0	2014
162	Colombia	South America	Americas	54.2	2020
163	Eswatini	Southern Africa	Africa	54.6	2016
164	Central African Republic	Middle Africa	Africa	56.2	2008
165	Zambia	Eastern Africa	Africa	57.1	2015
166	Suriname	South America	Americas	57.9	1999
167	Namibia	Southern Africa	Africa	59.1	2015
168	South Africa	Southern Africa	Africa	63.0	2014

Source: https://en.wikipedia.org/wiki/List_of_countries_by_income_equality

Gini Coefficient

- **Question 4**

How to define the equality level between wealth or income within a country via Gini coefficient?

- Below 0.2

- 0.2-0.29

- 0.3-0.39

- 0.4-0.59

- Higher than 0.6

Source: https://en.wikipedia.org/wiki/List_of_countries_by_income_equality

Herfindahl-Hirschman Index (HHI)

- **From Wiki:**

  Herfindahl-Hirschman Index (HHI) is a measure of the size of firms in relation to the industry they are in and is an indicator of the amount of competition among them.

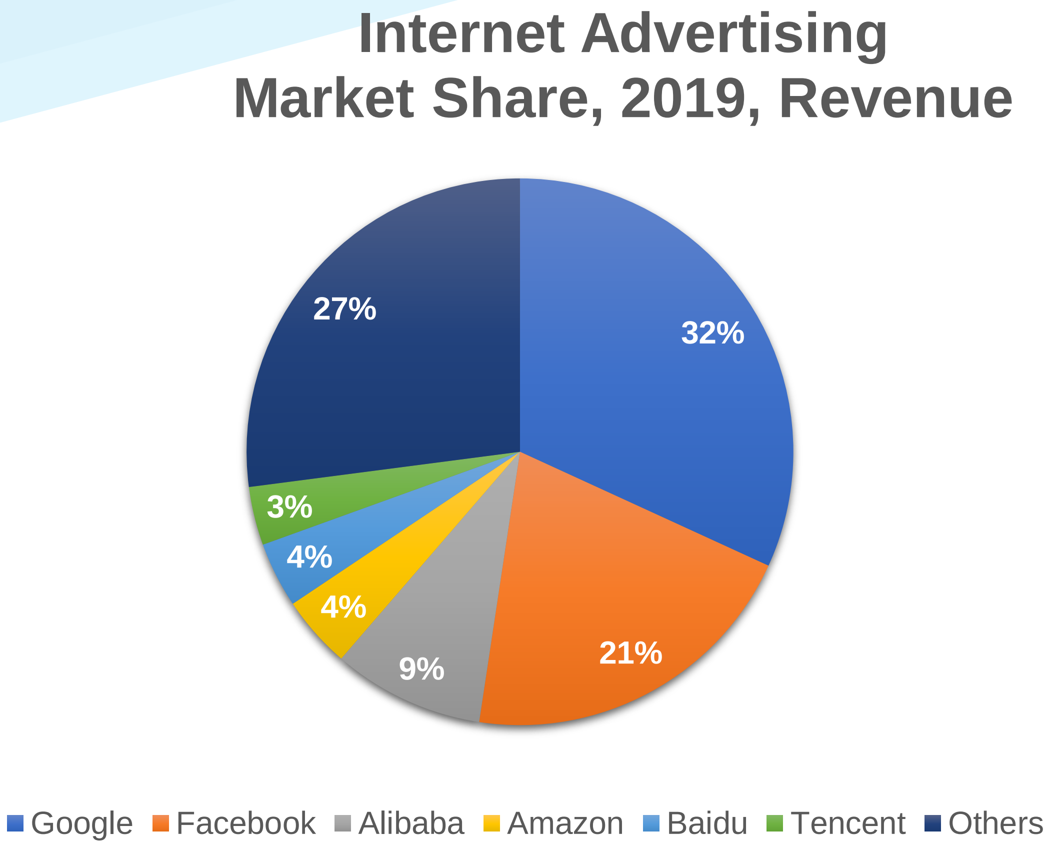
$$HHI = \sum_{i=1}^{N} \left(\frac{x_i}{\sum_{i=1}^{N} x_i} \right)^2 = \sum_{i=1}^{N} S_i^2 ,$$

where N is the number of company, x_i is the market scale of the $i-th$ company, and S_i is the market share of the $i-th$ company.

Herfindahl-Hirschman Index (HHI)

Level	Nature of Competition	Range of Herfindahl
1	Perfect competition	Usually below 0.2
2	Monopolistic competition	Usually below 0.2
3	Oligopoly	0.2 – 0.6
4	Monopoly	0.6 and above

Herfindahl-Hirschman Index (HHI)



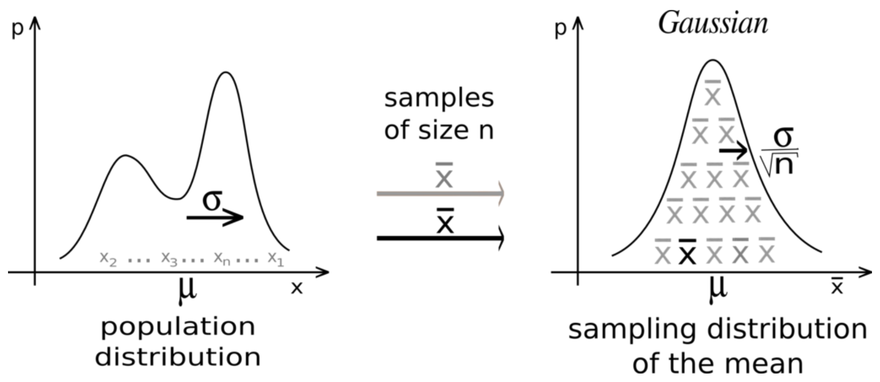
Question 6

Design a function to calculate the HHI of internet advertising market share in 2019 by revenue.

Source: https://www.t4.ai/industry/internet-advertising-market-share

Shape

- For each type of distribution, they have their own variables to describe the shape of distribution, such as lambda for Poisson distribution, mean and standard deviation for normal distribution.
- In many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution even if the original variables themselves are not normally distributed – central limit theorem (CLT).



Source: https://en.wikipedia.org/wiki/Central_limit_theorem

Shape – Indicators

- If the function is a probability distribution, then the first moment is the **expected value**, the second central moment is the **variance**, the third standardized moment is the **skewness**, and the fourth standardized moment is the **kurtosis**.

	Expected Value	Variance	Skewness	Kurtosis
Discrete	$\mu = \sum_{i=1}^{\infty} P(X = x_i)$	$\sigma^2 = \sum_{i=1}^{\infty} P(x_i)(x_i - \mu)^2$	$\gamma = \frac{M_3}{\sigma^3}$	$\kappa = \frac{M_4}{\sigma^4}$
Continuous	$\mu = \int_{-\infty}^{\infty} x f(x) dx$	$\sigma^2 = \int_{-\infty}^{\infty} (x_i - \mu)^2 f(x) dx$		

$$Kth\ central\ moment\ for\ discrete \Rightarrow M_k = \sum_{i=1}^{\infty} P(x_i)(x_i - \mu)^k$$

$$Kth\ central\ moment\ for\ continuous \Rightarrow M_k = \int_{-\infty}^{\infty} (x_i - \mu)^k f(x) dx$$

Shape – Q7

	Expected Value	Variance	Skewness	Kurtosis
Discrete	$\mu = \sum_{i=1}^{\infty} P(X = x_i)$	$\sigma^2 = \sum_{i=1}^{\infty} P(x_i)(x_i - \mu)^2$	$\gamma = \frac{M_3}{\sigma^3}$	$\kappa = \frac{M_4}{\sigma^4}$
Continuous	$\mu = \int_{-\infty}^{\infty} xf(x)dx$	$\sigma^2 = \int_{-\infty}^{\infty} (x_i - \mu)^2 f(x)dx$		

$$Kth\ central\ moment\ for\ discrete \Rightarrow M_k = \sum_{i=1}^{\infty} P(x_i)(x_i - \mu)^k$$

$$Kth\ central\ moment\ for\ continuous \Rightarrow M_k = \int_{-\infty}^{\infty} (x_i - \mu)^k f(x)dx$$

Question 7

Describe the characteristics of the following distributions.
(1) Skewness = 0; (2) Skewness < 0; (3) Skewness > 0;
(4) Kurtosis = 0; (5) Kurtosis < 0; (6) Kurtosis > 0.

Question Time

If you have any questions, please do not hesitate to ask me.

The End

Thank you for your attention))