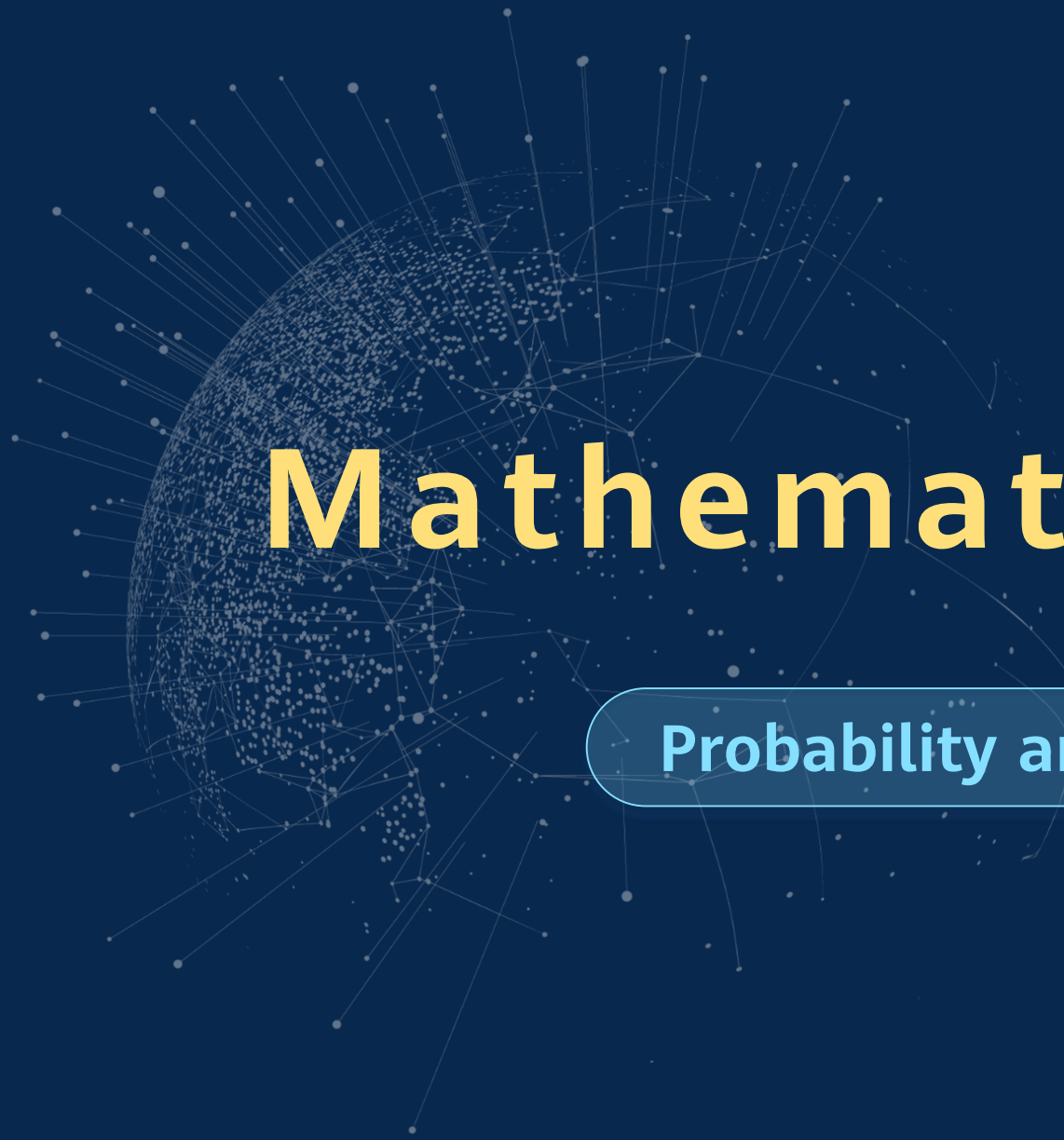


Mathematics Basics

Probability and Statistics

HUAWEI

Foreword

- This document introduces the mathematics basics used in AI, including linear algebra, probability and statistics as well as optimization problem.

Contents

- Mathematics and AI
  - Linear Algebra
  - Probability and Statistics
    - Basic Concepts of Probability and Statistics
    - Random Variable and Probability Distribution
    - Law of Large Numbers and Law of Central Limit
    - Parameter Estimation and Hypothesis Test
  - Optimization Problems

Statistical Basics

- Population: The set of data (numeric or otherwise) corresponding to the entire collection of units about which information is sought.
- Sample: A subset of the population data that are actually collected in the course of a study.
- Example:
  - Population--Blood pressure readings of ALL people in China.
  - Sample--Blood pressure readings of 1000 randomly selected people in China.
- In most studies, it is difficult to obtain information about the whole population. That is why we rely on samples to make estimates and inferences related to the whole population.

Parameters vs Statistics

- A parameter is a number that describes a population.
- A statistic is a number that describes a sample.
- Parameters are usually denoted using Greek letters $\{\mu, \sigma\}$ while statistics are usually denoted using Roman letters $\{x, s\}$.
- A parameter is a fixed number (usually unknown). A statistic is a variable whose value varies from sample to sample.

Descriptive Statistics

- Many methods are available for summarising data in numeric form.
- Numeric
  - Measure of location

    Mean, Median, Mode
  - Measure of spread

    Standard deviation, MAD (median absolute deviation), IQR
  - Others:

    Min, Max, Quartile, Five number summaries (used later in boxplot)

Measure of Location

- Consider n samples of data drawn from population. $\{x_1, x_2, \dots, x_n\}$
- Sample mean: the sum of all the observations divided by the number of observations. It is written in symbols as:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Sample median:
  - The $\frac{(n+1)^{th}}{2}$ largest observation if n is odd.
  - The average of the $(\frac{n}{2})^{th}$ and $(\frac{n}{2}+1)^{th}$ largest observation if n is even.
- Sample mode: the most frequently occurring value among all the observations in a sample.

Median or Mean?

- Both the median and the mean are measures of location, but which is preferable?
- For symmetric data, the mean is usually less variable from sample to sample than the median.
- For skewed data, the median is a better measure of location.
- The median does not reacted as much as the mean by outliers. This property of the median is known as ‘robustness’.
- The mean is easier to compute than the median and is much easier to handle theoretically.

Measure of Spread-Standard deviation

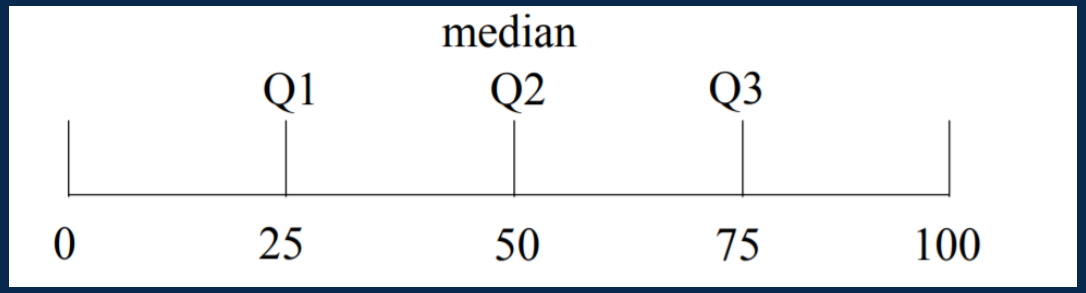
- The **standard deviation (SD)** or **variance** measures how spread out the data are around the mean.
- Steps to calculate sample variance
  1. Find the mean of the data.
  2. Make a list of **deviations** from the mean.
  3. Calculate the average of the squares of deviations.

$$var = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}, \qquad SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

- The sample SD is defined as: $SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

Measure of Spread – IQR

- The three quartiles, Q1, Q2, and Q3, approximately divide an ordered data set into four equal parts.



- The **Inter-quartile range (IQR)** is defined as the upper quartile (Q3 ;75th percentile) minus the lower quartile (Q1; 25th percentile). It is the width of the interval that contains the middle 50% of the data

$$IQR = Q3 - Q1$$

SD vs IQR

- Similar to median vs mean
- Sample standard deviations and the $IQR\ (=\ Q3 -\ Q1)$ are both measures of spread. The IQR is robust, like the median, but it is harder to handle theoretically than the standard deviation.

Probability Basics: Experiment, Sample Space, Probability

- A **random experiment** must satisfy the following properties:
  - can be repeated under the same conditions with more than one possible outcome.
  - The outcome of each trial cannot be predicted with certainty.
- **Sample space S**: The set of all the possible outcomes of a random experiment.
- **Sample point**: each possible outcome of a random experiment.
- **Random event A**: a set of possible outcomes of the experiment, thus A is the subset of S. $P(A) = \sum_{i \in A} p_i$
- Example:
  - Random experiment E_1: Observe the possible outcomes of numbers when a dice is rolled.
  - Sample space: $S = \{1, 2, 3, 4, 5, 6\}$
  - Sample point: $e_i = 1, 2, 3, 4, 5, 6$
  - Random event A_1: "The outcome is 5" can be denoted as $A_1 = \{x | x = 5\}$.

Conditional Probability and Bayes' Theorem

- Conditional probability of Y given X has occurred, assuming $P(X) \neq 0$

$$P(Y|X) = \frac{P(YX)}{P(X)}$$

- **Bayes' Theorem**:

$$P(Y|X) = \frac{P(YX)}{P(X)} ; P(X|Y) = \frac{P(XY)}{P(Y)}$$

$$\because P(YX) = P(XY)$$

$$\therefore P(Y|X)P(X) = P(X|Y)P(Y)$$

$$\therefore P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$$

$P(X|Y)$ indicates $Posterior$; $P(Y|X)$ indicates $Likelihood$; $P(X)$ indicates $Prior$; $P(Y)$ indicates $Normalizer$.

$$P(hypothesis|data) = \frac{P(data|hypothesis)P(hypothesis)}{P(data)}$$

Conditional Probability and Bayes' Theorem

- If X is a probability space $\{X_1, X_2, \ldots, X_n\}$ consisting of mutually independent events, then $P(Y)$ can be calculated by the **total probability theorem**

$$P(Y) = P(Y|X_1)P(X_1) + P(Y|X_2)P(X_2) + \cdots + P(Y|X_n)P(X_n)$$

- **Bayes' Theorem**:

$$P(X_i|Y) = \frac{P(Y|X_i)P(X_i)}{\sum_{i=1}^{n} P(Y|X_i)P(X_i)}$$

- Application of Bayes' Theorem: statistical machine translation, Bayesian network, and others.

Independence

- Two events A and B are independent if

$$P(YX) = P(Y)P(X)$$

- If X and Y are independent then Independence

$$P(Y \mid X) = \frac{P(YX)}{P(X)} = \frac{P(\mathrm{Y})P(X)}{P(X)} = P(Y)$$

Expectation and Variance

- **Mathematical expectation:** the sum of: [(each of the possible outcomes) × (the probability of the outcome occurring)]. The expectation is one of the most basic mathematical characteristics of a probability distribution. It represents the average value of a random variable.
  - For a discrete random variable, $E(X) = \sum_{k=1}^{\infty} x_k p_k, k = 1, 2, \ldots$.
  - For a continuous random variable, $E(X) = \int_{-\infty}^{\infty} xf(x)dx$.
- **Variance:** measures the dispersion of random variables or a set of numbers. This means that it measures the gap between random variables and their expectations.

$$D(X) = Var(X) = E\{[X - E(X)]^2\}$$

In addition, $\sqrt{D(X)}$ is often represented by $\sigma(X)$, and is called the standard deviation or mean square deviation. $X^* = \frac{X - E(X)}{\sigma(X)}$ is called the standardized variable of X.

Covariance, Correlation Coefficient, and Covariance Matrix

- **Covariance**: somewhat measures the linear relationship between two variables.

$$Cov(X,Y) = E\big[\big(X - E(X)\big)\big(Y - E(Y)\big)\big]$$

- **Correlation coefficient**: sometimes called linear correlation coefficient, which measures the linear relationship between two variables.

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X)}\sqrt{D(Y)}}$$

- The **covariance matrix** of a random variable (X_1, X_2):

$$C = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

where $c_{ij} = Cov\big(X_i, X_j\big) = E\big\{[X_i - E(X_i)]\big[X_j - E\big(X_j\big)\big]\big\}, i, j = 1, 2, \cdots, n$. The elements along the diagonal of the covariance matrix are the variance of X_1, X_2, and the rest are the covariance of X_1, X_2.

Contents

- Mathematics and AI
- Linear Algebra
- Probability and Statistics
  - Basic Concepts of Probability and Statistics
  - Random Variable and Probability Distribution
  - Law of Large Numbers and Law of Central Limit
  - Parameter Estimation and Hypothesis Test
- Optimization Problems

Random Variables

- **Random variable**: a function mapping the subset of a sample space to real numbers. Random variables give numbers to outcomes of random events. The outcomes of some random experiments may not be numbers. We usually use uppercase letters to denote random variables, and lowercase letters to denote their values.
- Example 1: Random experiment to observe the sum of possible outcomes when two dice are rolled. The sample space is $S = \{e\} = \{(i,j)|i,j = 1,2,3,4,5,6\}$, where i and j respectively represent the outcomes of the first dice and the second dice. X is known as a random variable when it is used to represent the sum of the two outcomes:

$$X = X(e) = X(i,j) = i + j, \qquad i,j = 1,2,\dots,6$$

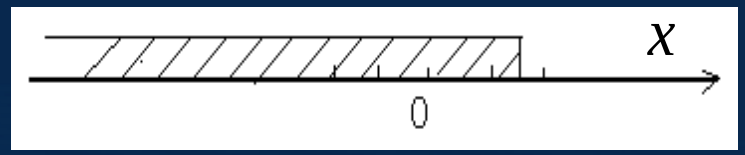
- Random variables are divided into two types based on the possible values:
  - Discrete random variable: has a finite or countable infinite number of values, for example, the number of births in a certain place in a certain year.
  - Continuous random variable: takes an uncountable and infinite number of possible values, for example, the amount of milk a cow produces on a given day (It may be a value in a given interval of numbers.)

Discrete Distributions

- For any random variable X with a discrete distribution, there is a sample space S with finite number of possible values $x = \{x_1, x_2, \dots, x_n\}$ and associated probabilities $\{p_1, p_2, \dots\}$.

X	x_1	x_2	...	x_n	...
p_k	p_1	p_2	...	p_n	...

- The point probabilities for each random real number x of a discrete random variable X is $f(x) = P(X = x)$, and the **cumulative distribution function(CDF)** $F(x) = P(X \leq x), -\infty < x < \infty$
- Significance of distribution function $F(x)$:
  - $F(x)$ at x indicates the probability that X falls in the interval $(-\infty, x]$, that is, the probability that the random variable X is less than or equal to x.



- Properties:
  - There is a countable number of possible values;
  - $\sum_i p_i = 1$

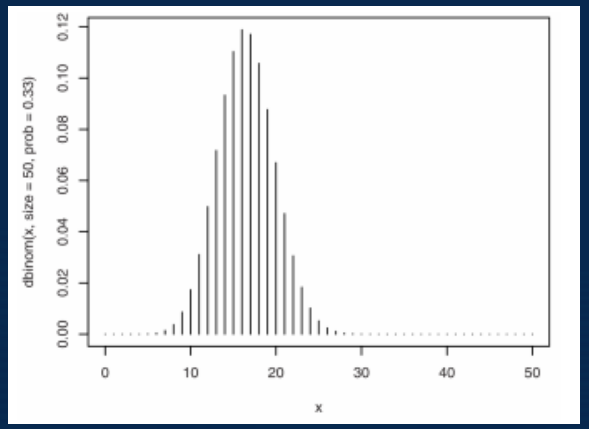
Discrete Distributions — Binomial Distribution

- The binomial distribution consists of **n** repeated Bernoulli trials.
- If X represents the number of times that event A occurs in a Bernoulli experiment with n trials, then the probability of event A occurring for k times in n trials is described as

$$P(X = k) = C_n^k p^k (1-p)^{n-k}, k = 0, 1, 2, \cdots n,$$

where $X \sim B(n, p)$, $E(X) = np$, $Var(x) = np(1-p)$

In this case, C_n^k is known as binomial coefficients. The parameter p is the probability of a successful outcome in an individual trial (called a Bernoulli Trial).



Discrete Distributions — Poisson Distribution

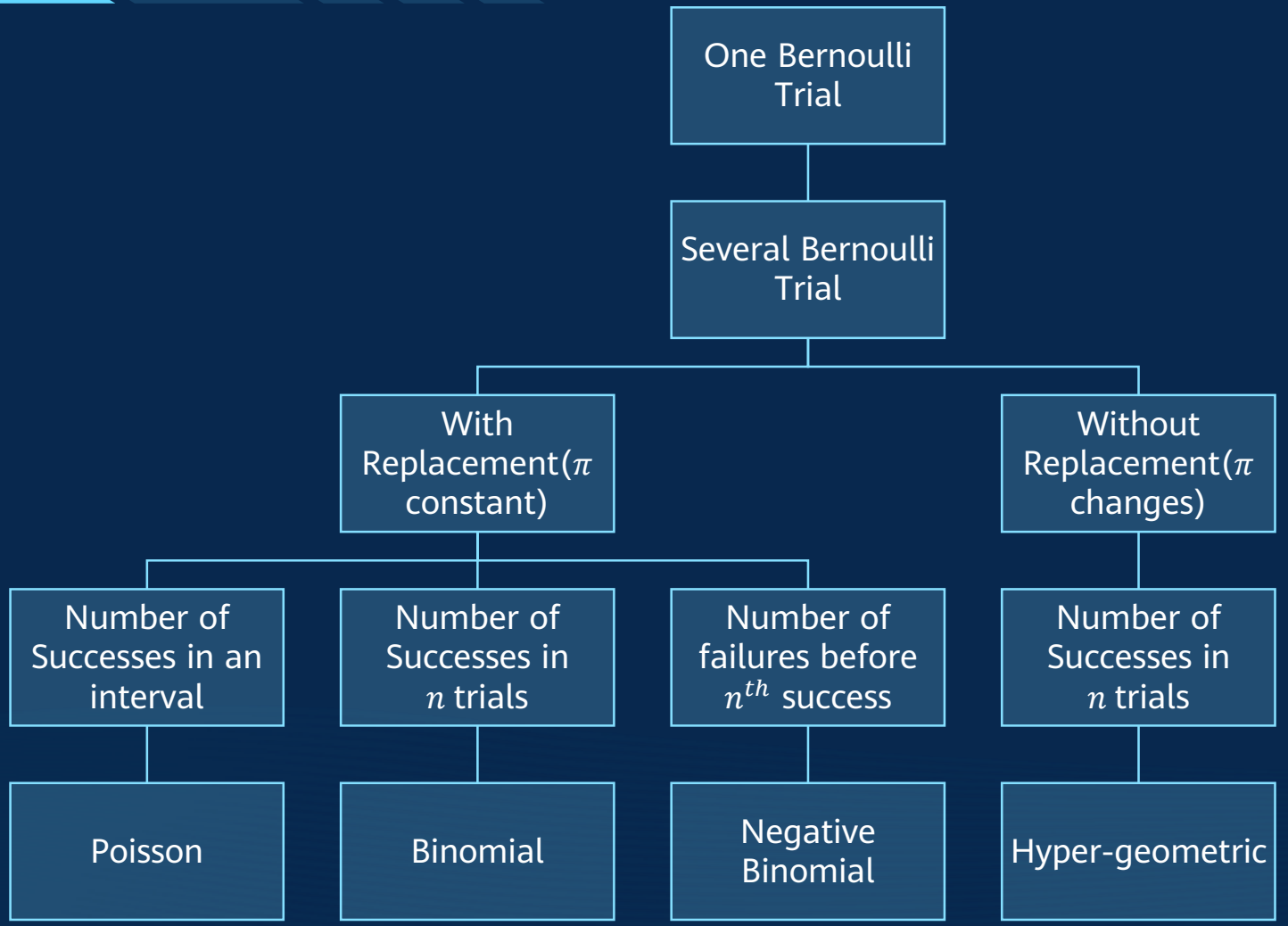
- **Poisson distribution**: If all the possible values of a random variable are $0, 1, 2, ...$, the probability of the variable taking each value can be written as

$$P\{X = k\} = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, ...,$$

  then X is called a random variable that follows the Poisson distribution with parameter λ, and is described by $X \sim P(\lambda)$, where $E(X) = \lambda$ and $D(X) = \lambda$. The parameter λ represents the average occurrence rate of a random event during an interval or in a unit area.

- The Poisson distribution is a limiting case of binomial distribution when n becomes very large and p gets very small.
- Poisson distribution describes the number of times that a random event will occur within an interval, for example, the number of service requests received by a service center, the number of passengers waiting at a bus stop, the number of faults that occur in a machine, the number of times that a natural disaster happens, and the number of variations in the DNA sequence.
- In image processing, Poisson noise that follows the Poisson distribution usually occurs due to the uncertainty caused by the imaging equipment. We often add Poisson noise to images for data augmentation.

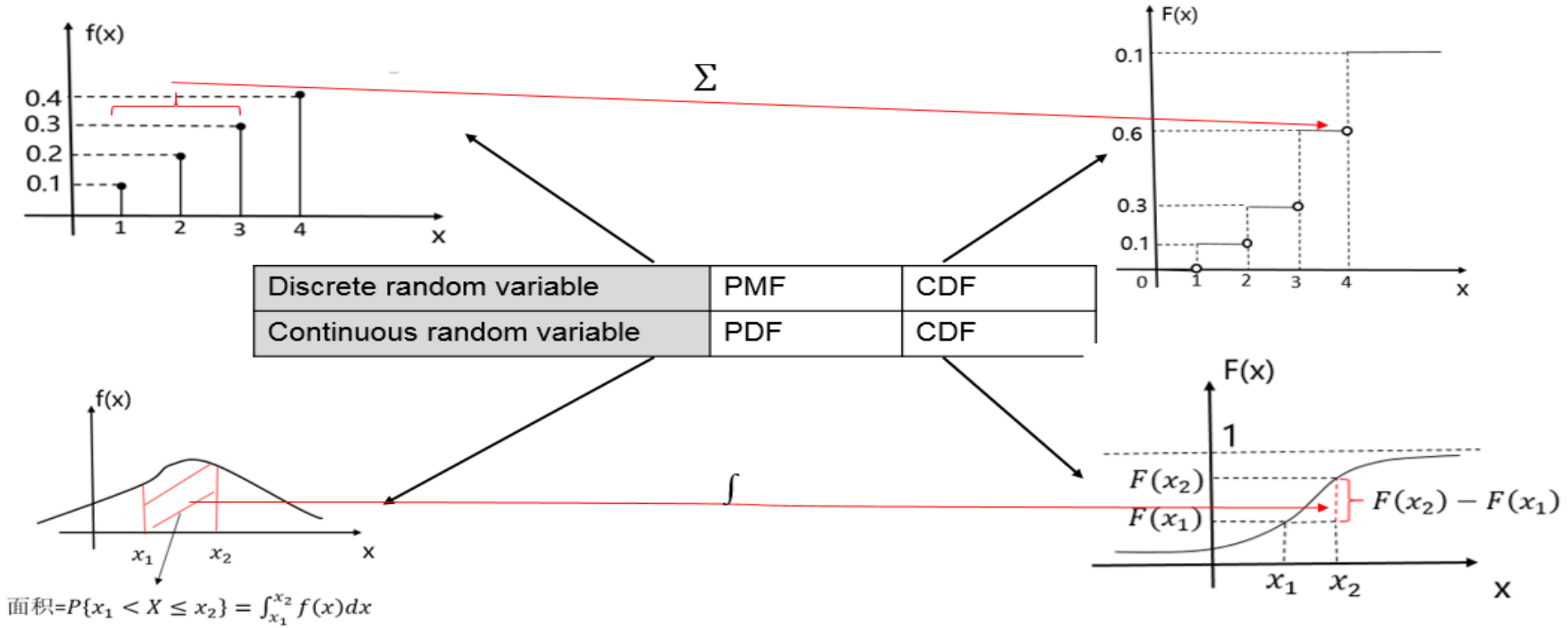
Summary of Special Discrete Distributions



Continues Distributions

- For any random variable X with a continues distribution, there is an infinite number of possible values; These values may be within a fixed interval such as the height of male in cm may be within the range of $[50, 300]$.
- The point probabilities for each value of x is $P(X = x) = 0$ and the cumulative distribution function $F(x) = \int_{-\infty}^{x} f(x)dx$
- Properties:
  - There are infinite number of possible values;
  - $f(x)$ is called the **probability density function (PDF)** and its integration from $-\infty$ to $+\infty$ with respect to x is 1.

PMF, PDF and CDF



PMF: **probability mass function,** i.e. probability distribution of a discrete random variable
PDF: probability density function
CDF: cumulative distribution function

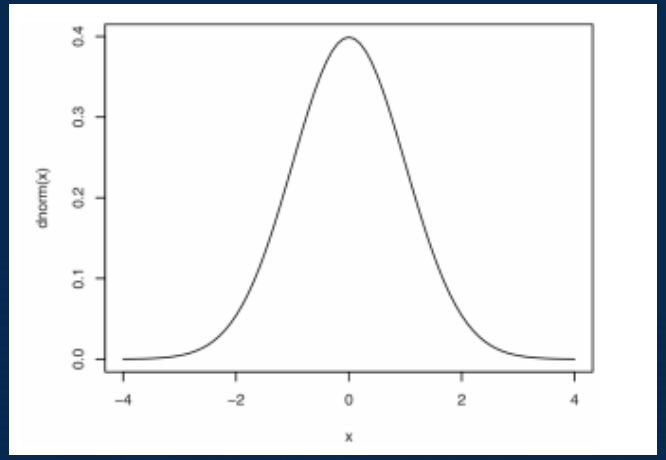
Special Distribution — Normal (Gaussian) Distribution

- If the probability density function of a continuous random variable X is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty,$$

where parameter μ and $\sigma(\sigma > 0)$ are the mean and standard deviation, followed by $X \sim N(\mu, \sigma^2)$.

When $\mu = 0, \sigma = 1$, the random variable X is said to follow the standard normal distribution, followed by $X \sim N(0, 1)$.



Statistical Notation

- We use capital letters to denote random variables, such as Y or X, and lowercase letters to represent specific realized values of random variables such as y or x.
  - The probability density function of X is denoted f;
  - the cumulative distribution function is F.
  - We use the notation $\mathrm{X} \sim f(x)$ to mean that X is distributed with density $f(x)$. Frequently, the dependence of $f(x)$ on one or more parameters also will be denoted with a conditioning bar, as in $f(x|\alpha, \beta)$.
  - The density functions for X and Y are f_X and f_Y, respectively.

Probability Distribution Review

- Random variables (both continuous and discrete) are associated with distributions.
- Learn the parameters of a distribution from data.

Example: Poisson Noise and Gaussian Noise in Images

Original

Gaussian noise that follows $\mu = 0,\ \sigma = 10$ is added.

Poisson noise that follows $\lambda = 15$ is added.

Random Vectors

- In practice, we usually need to use multiple vectors to describe problems. We group together multiple random variables to form a vector, known as a multivariate random variable or random vector.
- **Definition**: If $X_1(\omega), X_2(\omega), \dots, X_n(\omega)$ are n random variables defined in the same sample space $\Omega = \{\omega\}$, then we have

$$X(\omega) = (X_1(\omega), X_2(\omega), \dots, X_n(\omega)),$$

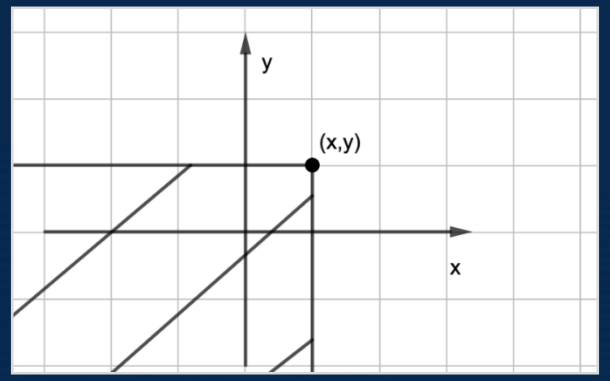
  which is called **an n-dimensional random variable or random vector**.
- For example, we usually estimate a person's age based on multiple features (random variables), such as his/her face shape, skin texture, senile plaques, skin elasticity, and hairline. These features are combined and mapped to a real number, that is, the age.

Joint Cumulative Distribution Function

- A random variable has a corresponding CDF. Likewise, a random vector has a corresponding joint CDF.
- Definition: For any n real numbers $x_1, x_2, \ldots, x_n$, the probability of **n events $\{X_1 \le x_1\}$, $\{X_2 \le x_2\}$,...,$\{X_n \le x_n\}$ occurring simultaneously** is

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \ldots, X_n \le x_n),$$

  which is called the joint CDF of an n-dimensional random variable.
- The joint CDF of a two-dimensional random variable $F(x, y) = P(X \le x, Y \le y)$ represents the probability that a random point (X, Y) falls in the semi-infinite rectangle lying to the left and below the point (x, y).



Joint Probability Density

- A one-dimensional random variable has a corresponding probability density function. Likewise, a random vector has a corresponding joint probability density function.
- Definition: If a binary non-negative function $p(x, y)$ satisfies the CDF of a two-dimensional random variable (X, Y), and can be denoted as

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} p(u, v) du dv,$$

then (X, Y) is called a two-dimensional continuous random variable, and the function $p(u, v)$ is the joint probability density of (X, Y).

Contents

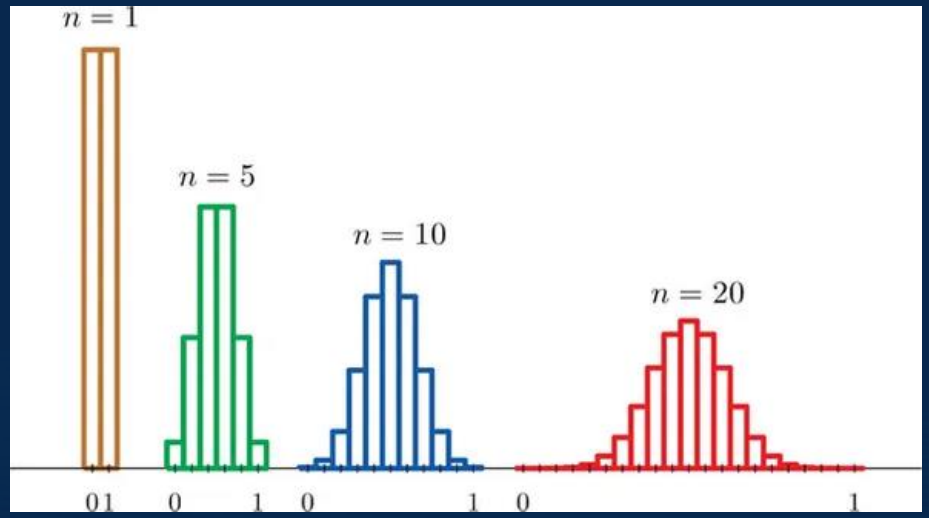
Law of Large Numbers

- **Law of large numbers(LLN)：** As the number of identically distributed, randomly generated variables increases with repeating experiment independently a large number of times, and average of the result should be converge to the expected value.
- Example: Coin flipping with two results “head” and “tail”.

$$\frac{\sum_{I=1}^{n} 1_{\{x_i="head"\}}}{n} \rightarrow \int P(x = "head")\, 1_{\{x="head"\}} dx = P(x = "head")$$

Central Limit Theorem

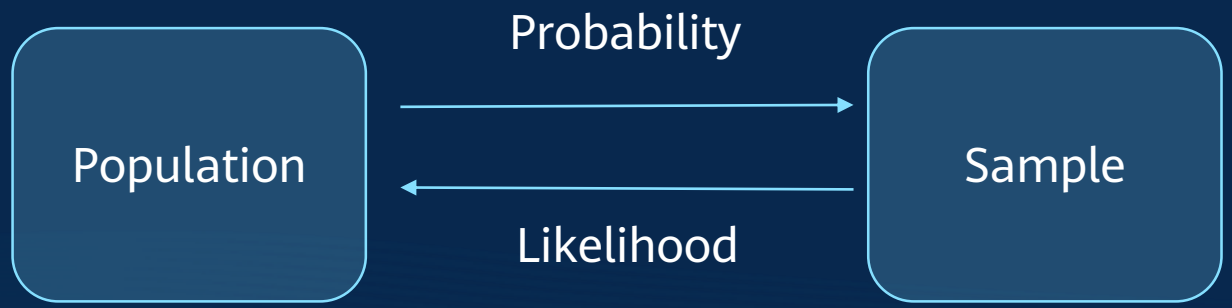
- The distribution of the sum of n $i.i.d.$ random variables becomes increasingly Gaussian as n grows. As a sample size increases, the sample mean and standard deviation will be closer to the population mean μ and standard deviation σ .
- Example: n uniform $[0,1]$ random variables.



Contents

Parameter Estimation-Maximum Likelihood

- Maximum likelihood estimation (MLE) is a method of estimating the parameters of a probability distribution by maximizing a likelihood function, so that under the assumed statistical model the observed data is most probable. The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.
- What is likelihood?



MLE-Discrete Probability Distribution

- Let X be a discrete random variable with probability mass function p depending on parameter(s) θ.
- Likelihood function:

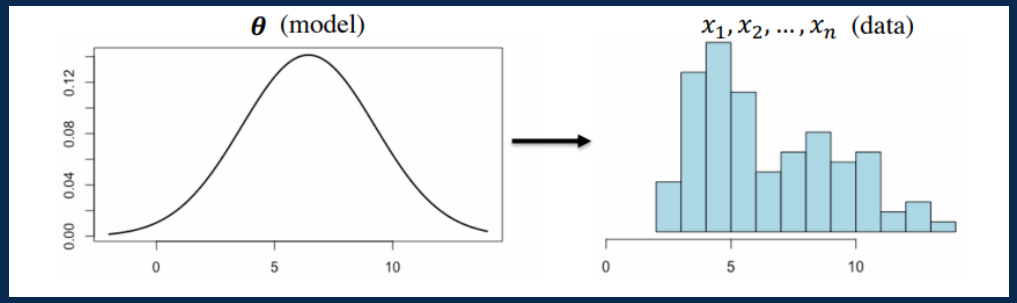
$$\mathcal{L}(\theta|x) = \prod_{i=1}^{n} p(x_i|\theta)$$

- Since good predictions are better, a natural approach to parameter estimation is to choose the set of parameter values that yields the best predictions—that is, the parameter that maximizes the likelihood of the observed data. This value is called the maximum likelihood estimate (MLE), defined formally as:

$$\tilde{\theta}_{MLE} \stackrel{\text{def}}{=} \arg\max_{\theta} \mathcal{L}(\theta|x_i)$$

MLE-Continuous Probability Distribution

- Let X be a random variable following an absolutely continuous probability distribution with density function f depending on parameter(s) θ. $f(x_1, x_2, \dots, x_n|\theta)$ Probability of observing $x_1, x_2, \dots, x_n$ given parameter(s) θ.



- Likelihood function(if every predictor is i.i.d)：

$$\mathcal{L}(\theta|x_i) = f(x_1, x_2, \dots, x_n|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdot \cdots \cdot f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

- Because maximize log-likelihood is often easier so we commonly maximize the following:

$$\mathcal{L}(\theta|x_i) = \prod_{i=1}^{n} f(x_i|\theta) \rightarrow \ln \mathcal{L}(\theta|x_i) = \sum_{i=1}^{n} \ln f(x_i|\theta)$$

Expectation–Maximization(EM) Algorithm

- “A general technique for finding maximum likelihood estimators in latent variable models is the expectation-maximization (EM) algorithm.”

  —Pattern Recognition and Machine Learning, 2006.

- Observed data $X = \{x_1, x_2, \dots, x_n\}$; latent(Incomplete) data $Y = \{y_1, y_2, \dots, y_n\}$;
- Complete data Z=$\{X, Y\}$, Complete data likelihood:

$$\mathcal{L}(\theta|Z) = p(Z|\theta) = p(X, Y|\theta) = p(Y|X, \theta)p(X|\theta)$$

  - $\mathcal{L}(\theta|Z)$: if we are given θ, a function of random variable Y;
  - $p(Y|X, \theta)$: the function of latent variable Y and parameter θ, the result is in terms of random variable Y.
  - $p(X|\theta)$: the function of parameter θ, which is computable.

Expectation–Maximization(EM) Algorithm

- Complete data likelihood: $\mathcal{L}(\theta|Z) = p(X, Y|\theta)$
- Let $\theta^{(i-1)}$ be the parameter vector obtained at $(i-1)^{th}$ time step.
- **E-Step**. Estimate the missing variables in the dataset.
  - Define (conditional expectation of log likelihood of complete data)

$$Q(\theta, \theta^{(i-1)}) = E[log\mathcal{L}(\theta|Z)X, \theta^{(i-1)}]$$

- **M-Step**. Maximize the parameters of the model in the presence of the data.

$$\mathrm{Q}^{(i)} = arg\max_{\theta} Q\left(\theta, \theta^{(i-1)}\right)$$

- The EM algorithm is widely-used in machine learning to solve unsupervised learning problems, such as density estimation and clustering.

Hypothesis Test

- Hypothesis testing is a formal procedure for investigating the ideas using statistics, which used to test specific predictions, called hypothesis.
- There are 5 main steps in hypothesis testing:
  - State the research hypothesis as a null (H_0) and alternate (H_a) hypothesis.
    - The alternate hypothesis: the initial hypothesis that predicts a relationship between variables.
    - The null hypothesis: the prediction of no relationship between the variables.
  - Collect data in a way designed to test the hypothesis.
  - Perform an appropriate statistical test.
  - Decide whether the null hypothesis is supported or rejected.
  - Present the findings in the results and discussion section.

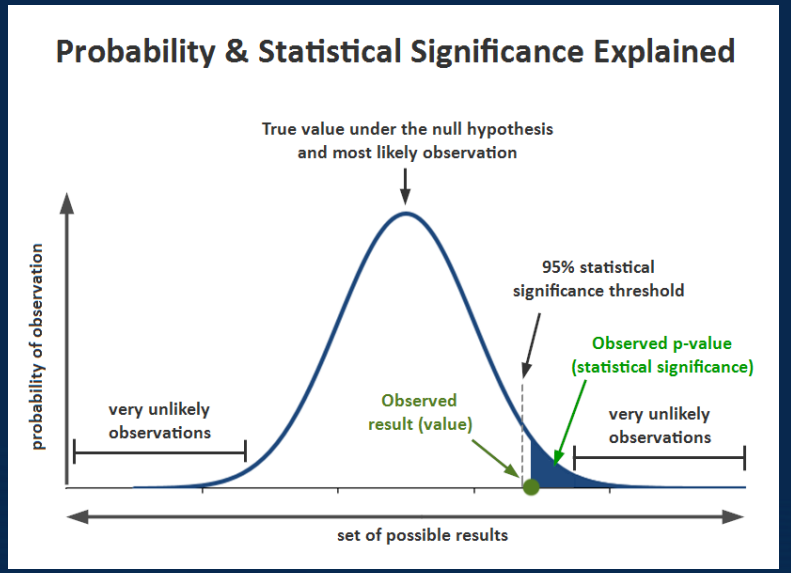
Test Statistics

- A test statistic describes how closely the observed data match the distribution expected under the null hypothesis of that statistical test. The test statistic is used to calculate the p-value of the results, helping to decide whether to reject null hypothesis.

Test statistic	Null H_0 and alternative H_a hypotheses	Statistical tests that use it
z-value	$\boldsymbol{H_0}$: The means of two groups are equal $\boldsymbol{H_a}$: The means of two groups are not equal	• Z-test

$P-value$

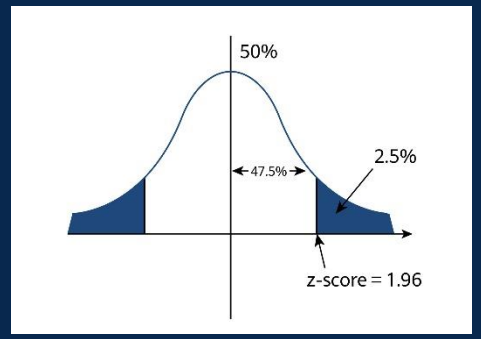
- The **significance level** is denoted by the Greek letter $\boldsymbol{\alpha}$, usually 1%, 5%, 10%.
- $\boldsymbol{P-value}$ stands for conditional probability value, which is the probability of getting a test statistic at least as extreme as the one given null hypothesis was true.
  - $p-value < \alpha \Longrightarrow reject\ H_0;$
  - $p-value \geq \alpha \Longrightarrow Do\ not\ reject\ H_0.$



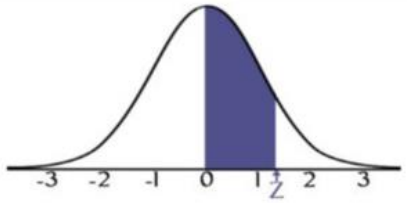
http://blog.analytics-toolkit.com/2017/statistical-significance-ab-testing-complete-guide/

Example-$Z\ Test$(1)

- Blood glucose levels for obese patients have a mean $\bar{x} = 100$ with $\sigma = 15$. A researcher thinks that a diet high in raw cornstarch will have a positive or negative effect on blood glucose levels. A sample size $n = 30$ patients who have tried the raw cornstarch diet have a mean glucose level of 140. Test the hypothesis that the raw cornstarch had an effect.
  - Step 1: State the hypothesis, H0: $\mu = 100$, H1: $\mu \neq 100$
  - Step 3: State alpha level $\alpha = 0.05$. As a two-tailed test, split α into two. $\frac{0.05}{2} = 0.025$
  - Step 4: Find Table z-score associated with α.
  - Step 5: Find the test statistic using this formula: $Z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ $\mathrm{Z} = \frac{140-100}{15/\sqrt{30}} = 14.6$
  - Step 6: since $14.6 > 1.96$, reject the null hypothesis.



Example-*Z Test* (2)



STANDARD NORMAL TABLE (Z)

Entries in the table give the area under the curve between the mean and *z* standard deviations above the mean. For example, for *z* = 1.25 the area under the curve between the mean (0) and *z* is 0.3944.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0190	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2969	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3513	0.3554	0.3577	0.3529	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817

Confidence Intervals

- The **confidence interval(CI)** is the range of values that expect the estimate to fall between a certain percentage of the time if re-sample the population in the same way.
- The significance level is denoted by the Greek letter α, usually 1%, 5%, 10%.
- The **confidence level** is the percentage of times that expect to reproduce an estimate between the upper and lower bounds of the confidence interval.

$$Confidence\ level = 1 - \alpha$$

Calculating Confidence Interval

- The point estimate that are constructing the confidence interval for
- The critical values for the test statistic
  - Choose alpha value α.
  - Decide if need a one-tailed interval or a two-tailed interval.
  - Look up the critical value that corresponds with the alpha value α.
- The standard deviation of the sample $SD = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$
- The sample size: If data follows a normal distribution in a large sample size ($n > 30$), then use the z-distribution to find the critical values with population data. Otherwise use t-distribution in a small dataset with sample data.

CI for mean of normally-distributed data(1)

- The confidence interval for data which follows a standard normal distribution is:

$$CI = \bar{X} \pm Z^* \frac{\sigma}{\sqrt{n}}$$

- Where:
  - CI=the confidence interval
  - $\bar{X}$= the population mean
  - Z^*= the critical value of the *z*-distribution
  - σ = the population standard deviation
  - $\sqrt{n}$ = the square root of the population size

CI for mean of normally-distributed data(2)

- The confidence interval for the t-distribution follows the same formula, but replaces the $Z^*$with the t^*
- In real life, we replace the population values with sample values unless we do a complete census:

$$CI = \hat{X} \pm Z^* \frac{s}{\sqrt{n}}$$

- Where:
  - $\hat{X}$=the sample mean
  - s = the sample standard deviation

CI for proportions

- The confidence interval for a proportion follows the same pattern as the confidence interval for means, but place of the standard deviation you use the sample proportion times one minus the proportion:

$$CI = \hat{p} \pm Z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- Where:
  - $\hat{p}$ = the proportion in your sample (e.g. the proportion of respondents who said they watched any television at all)
  - Z^*= the critical value of the z-distribution
  - n = the sample size

Example: Calculating a confidence interval

- In one survey, there are more than 30 observations and the data follow an approximately normal distribution (bell curve), so we can use the *z*-distribution in population dataset for test statistics.
- For a two-tailed 95% confidence interval, $\alpha = 0.025$, the corresponding critical value z-statistic=1.96.
- To calculate the upper and lower bounds of the 95% confidence interval:

$$CI = \bar{X} \pm Z^* \frac{\sigma}{\sqrt{n}} = mean \pm 1.96 \frac{standard\ deviation}{\sqrt{sample\ size}}$$

Confidence level	90%	95%	99%
α for one-tailed CI	0.1	0.05	0.01
α for two-tailed CI	0.05	0.025	0.005
z-statistic	1.64	1.96	2.57

Look-up Table for *z*-distribution

Thank you.

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home, and organization for a fully connected, intelligent world.

HUAWEI