

**INTRODUCTION TO
CORPUS-BASED
STATISTICS-ORIENTED
(CBSO)
TECHNIQUES
(PART II: BASIC CONCEPTS)**

Keh-Yih Su
Tung-Hui Chiang
Jing-Shin Chang

Department of Electrical Engineering
National Tsing-Hua University
Hsinchu, TAIWAN 30043, R.O.C.
email: kysu@bdc.com.tw

CONTENTS

INTRODUCTION TO PROBABILITY

- Basic terminology
- Random variables
- Distribution and density function
- Mean and variance

INTRODUCTION TO STATISTICS

- Basic terminology
- Estimation
- Test of hypotheses

INTRODUCTION TO STOCHASTIC PROCESS

- Markov chains

INTRODUCTION TO INFORMATION THEORY

- Entropy
- Mutual information
- Perplexity

INTRODUCTION TO PROBABILITY

Basic Terminology

Experiment:

- The process of observing a phenomenon that has variation in its outcomes.
- Example:
Observing the outcomes of tossing a fair coin twice.

Sample Space:

- The totality of the possible outcomes of a random experiment.
- Example:
The sample space $S=\{HH, HT, TH, TT\}$, where H: head; T: tail.

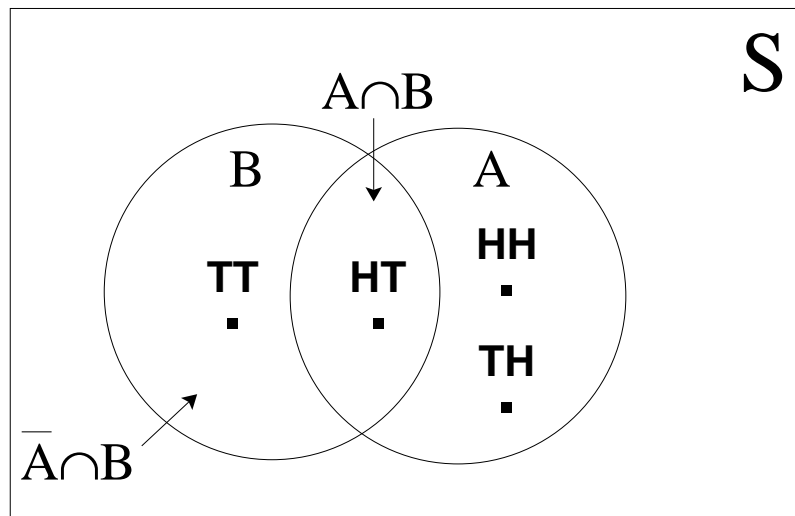
Event:

- An event is a subset of the sample space.
- Example:

A : at least one head in the two tosses.

B : tail at the second toss.

$A=\{HT, TH, HH\}$, $\bar{A}=\{TT\}$, $B=\{HT, TT\}$



$$P(A) = \frac{3}{4}, P(\bar{A}) = \frac{1}{4}, P(B) = \frac{1}{2}$$

$$P(A \cap B) = \frac{1}{4}, P(\bar{A} \cap B) = \frac{1}{4}$$

$$P(A | B) = \frac{1/4}{1/2} = \frac{1}{2}$$

$$P(\bar{A} | B) = \frac{1/4}{1/2} = \frac{1}{2}$$

$$P(B | A) = \frac{1/4}{1/3} = \frac{3}{4}$$

$$P(B | \bar{A}) = \frac{1/4}{1/4} = 1$$

Probability of an Event

- The probability of an event expresses the long-run frequency for the event occurring in many repeated experiments.
- Example:

$$P(A) = \frac{3}{4}, P(\bar{A}) = \frac{1}{4}, P(B) = \frac{1}{2}.$$

Probability of a Joint Event

- The probability of the joint event A and B is $P(A, B)$ (or $P(A \cap B)$)
- Example:

$$A \cap B = \{HT\}, P(A \cap B) = \frac{1}{4}$$

$$\bar{A} \cap B = \{TT\}, P(\bar{A} \cap B) = \frac{1}{4}$$

Conditional Probability

- The conditional Probability of the event A given that B event has occurred:

$$P(A | B) = \frac{P(A, B)}{P(B)}$$

- Example:

$$P(A | B) = \frac{1/4}{1/2} = \frac{1}{2}$$

$$P(\bar{A} | B) = \frac{1/4}{1/2} = \frac{1}{2}$$

$$P(B | A) = \frac{1/4}{1/3} = \frac{3}{4}$$

$$P(B | \bar{A}) = \frac{1/4}{1/4} = 1$$

- A and B are independent if and only if

$$P(AB) = P(A) \times P(B) \\ \Rightarrow P(A|B) = P(A)$$

Multiplication Theorem of Probability

Theorem:

$$\begin{aligned}P(A, B) &= P(A | B) \times P(B) \\ &= P(B | A) \times P(A)\end{aligned}$$

Example:

$$\begin{aligned}P(A, B) &= \frac{1}{4} \\ P(B) \times P(A | B) &= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4} \\ P(A) \times P(B | A) &= \frac{3}{4} \times \frac{1}{3} = \frac{1}{4}\end{aligned}$$

Generalization:

$$\begin{aligned}P(A_1, A_2, \dots, A_k) &= P(A_1 | A_2, \dots, A_k) \\ &\quad \times P(A_2 | A_3, \dots, A_k) \\ &\quad \dots \\ &\quad \times P(A_{k-1} | A_k) \times P(A_k)\end{aligned}$$

Bayes' Rule

$$\begin{aligned} P(A | B) &= \frac{P(A, B)}{P(B)} = \frac{P(A, B)}{P(A, B) + P(\bar{A}, B)} \\ &= \frac{P(A) \times P(B | A)}{P(A) \times P(B | A) + P(\bar{A}) \times P(B | \bar{A})} \end{aligned}$$

□ Example:

$$\begin{aligned} P(\bar{A}) &= 1 - P(A) = \frac{1}{4} \\ P(B | \bar{A}) &= \frac{P(B, \bar{A})}{P(\bar{A})} = \frac{1/4}{1/4} = 1 \\ P(A | B) &= \frac{\frac{3}{4} \times \frac{1}{3}}{\left(\frac{3}{4} \times \frac{1}{3}\right) + \left(\frac{1}{4} \times 1\right)} = \frac{1}{2} \end{aligned}$$

□ Generalization:

$$P(A_k | B) = \frac{P(B | A_k) \times P(A_k)}{\sum_{A_i} P(B | A_i) \times P(A_i)}$$

where A_1, A_2, \dots, A_n are partitions of the sample space; i.e.,

$$\begin{aligned} A_1 \cup A_2 \cup \dots \cup A_n &= S, \\ A_i \cap A_j &= \phi \quad \forall i \neq j. \end{aligned}$$

Discrete Random Variable

- A random variable X on a sample space S is a function $X : S \rightarrow \mathbb{R}$ that assigns a real number $X(s)$ to each sample point $s \in S$.
- Example:
 X : number of head in the two tosses,
 $[X = 0] = \{TT\}$
 $[X = 1] = \{TH, HT\}$
 $[X = 2] = \{TT\}$.

Continuous Random Variable

- A random variable X on a probability space (S, \mathcal{F}, P) is a function $X : S \rightarrow \mathbb{R}$ that assigns a real number $X(s)$ to each sample point $s \in S$, such that for every real number x , the set $\{s | X(s) \leq x\}$ is an event, where \mathcal{F} denotes the class of measurable subsets of S .

Distribution Function

- The distribution function F_X of a random variable X is defined to be a function

$$F_X(x) = P(X \leq x), \quad -\infty < x < \infty.$$

- Example: the continuous uniform distribution:

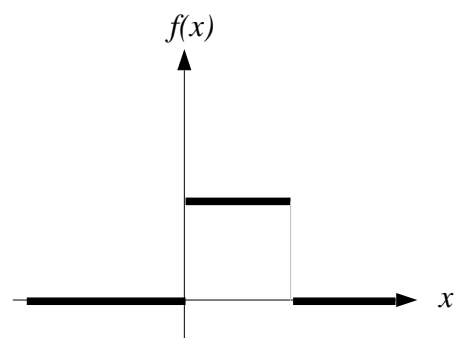
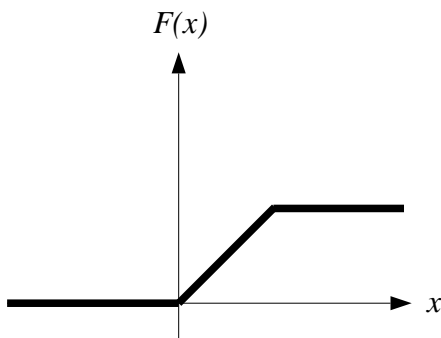
$$F_X(x) = \begin{cases} 0, & x < 0, \\ x, & 0 \leq x < 1, \\ 1, & x \geq 1. \end{cases}$$

Probability Density Function

- For a continuous random variable, X , $f(x) = \frac{dF_X(x)}{dx}$ is called the probability density function (pdf) of X

- Example: the continuous uniform distribution:

$$f(x) = \begin{cases} 0, & x < 0, \\ 1, & 0 \leq x < 1, \\ 0, & x \geq 1. \end{cases}$$



Mean

- The measure of central tendency or expected value of a random variable.
- A weighted average of the possible values of the random variable.

$$\mu_X = E[X] = \begin{cases} \sum_{x_i} x_i P(x_i), & (\text{discrete}) \\ \int_{-\infty}^{\infty} x f(x) dx, & (\text{continuous}). \end{cases}$$

Variance

- The measure of dispersion for a random variable.
- A weighted average which indicates how much individual values differ from the center of the distribution.

$$\text{Var}(X) = E[(X - \mu_X)^2] = \begin{cases} \sum_{x_i} (x_i - \mu_X)^2 P(x_i), & (\text{discrete}) \\ \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x) dx, & (\text{continuous}). \end{cases}$$

Special Discrete Distributions

□ Bernoulli Distribution:

- A random variable X has the Bernoulli distribution if (for some $p, (0 \leq p \leq 1)$)

$$P(X = x) = \begin{cases} p^x (1 - p)^{1-x}, & x = 0, 1 \\ 0, & \textit{otherwise.} \end{cases}$$

- A Bernoulli random variable X can be interpreted as the number of successes in one trial of an experiment where the probability of success is p .
- $E[X] = p; \quad Var[X] = p(1 - p).$

□ Binomial Distribution:

- A random variable X has the binomial distribution if (for some integer n , and some $p, (0 \leq p \leq 1)$)

$$P(X = x) = \begin{cases} \binom{n}{x} p^x (1 - p)^{n-x}, & x = 0, 1, \dots, n \\ 0, & \textit{otherwise.} \end{cases}$$

- A binomial random variable can be considered as a sum of n Bernoulli random variables, that is as the number of successes in n Bernoulli trials.
- When sampling from finite populations, the binomial distribution arises only when the sampling is done with replacement.
- $E[X] = np; \quad Var[X] = np(1 - p).$

□ Multinomial Distribution:

- Consider an experiment consisting of n independent and identical trials, in which each trial can result in any one of r possible outcomes.
- Let random variables X_i denote the number of trials resulting in outcome i ($i = 1, \dots, r$). The joint distribution of X_1, \dots, X_r has the *multinomial distribution*:

$$P(x_1, \dots, x_r) = \begin{cases} \frac{n!}{(x_1!)(x_2!) \cdots (x_r!)} p_1^{x_1} p_2^{x_2} \cdots p_r^{x_r}, & x_i = 1, 2, \dots, \text{ and } \sum_{i=1}^r x_i = n, \\ 0, & \text{otherwise.} \end{cases}$$

□ Poisson Distribution:

- A random variable X has the Poisson distribution if (for some $\mu > 0$, called a *parameter* of the distribution)

$$P(X = x) = \begin{cases} e^{-\mu} \frac{\mu^x}{x!}, & x = 0, 1, \dots \\ 0, & \text{otherwise.} \end{cases}$$

- The parameter μ can be interpreted as the “average” number of occurrences of the event)
- $E[X] = Var[X] = \mu$.

Special Continuous Distribution

□ Normal Distribution (Gaussian Distribution):

- A random variable X has the normal distribution $N(\mu, \sigma^2)$ if (for some $\sigma^2 > 0$ and $-\infty < \mu < \infty$)

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \quad -\infty < x < \infty.$$

- $N(0, 1)$ distribution is called the *standard normal* distribution.
- $E[X] = \mu; \quad Var[X] = \sigma^2.$

□ Chi-Square Distribution:

- A random variable X has the chi-square distribution with ν degrees of freedom (for some $\nu \in \mathbf{N}$)

$$f_X(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\frac{\nu}{2})} x^{(\nu/2)-1} e^{-x/2}, & x > 0 \\ 0, & \textit{otherwise.} \end{cases}$$

- Let X_1, X_2, \dots, X_ν be ν i.i.d. random variables with p.d.f. $N(0, 1)$, the random variable

$$\chi^2 = X_1^2 + X_2^2 + \dots + X_\nu^2$$

has a chi-square distribution with ν degrees of freedom.

- $E[X] = \nu; \quad Var[X] = 2\nu.$

□ t-distribution

- A random variable X has the t-distribution with n degrees of freedom (for some integer $n > 0$)

$$f_X(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi} \Gamma\left(\frac{n}{2}\right)} \frac{1}{\left(1 + \frac{x^2}{n}\right)^{(n+1)/2}}, \quad -\infty < x < \infty.$$

- Properties:
 - t-distribution curve is bell-shaped and centered at 0.
 - As the degree of freedom n increases, the spread of the corresponding distribution curve decreases.
 - Each t-distribution curve is more spread out than the standard normal curve.
 - As the degree of freedom $n \rightarrow \infty$, the sequence of t-distribution curve approaches the standard normal curve.

Joint Distribution and Density

□ Joint Distribution:

- The joint distribution $F(x, y)$ of two random variables X and Y is the probability of the event $\{X \leq x, Y \leq y\}$, i.e.,

$$F(x, y) = P(X \leq x, Y \leq y).$$

□ Joint Density:

- The joint density $f(x, y)$ of two random variables X and Y is

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}.$$

Marginal Distribution and Density

□ Marginal Distribution:

- The marginal distribution $F(x)$ of X is

$$F_X(x) = F(x, \infty).$$

- The marginal distribution $F_Y(y)$ of Y is

$$F_Y(y) = F(\infty, y).$$

□ Marginal Density:

- The marginal density $f_X(x)$ of X is

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy.$$

- The marginal density $f_Y(y)$ of Y is

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

Conditional Distribution and Density

□ Conditional Distribution:

- The conditional distribution $F(x|\mathcal{U})$ of the random variable X assuming \mathcal{U} is

$$F(x|\mathcal{U}) = \frac{P(X \leq x, \mathcal{U})}{P(\mathcal{U})}.$$

□ Conditional Density:

- The conditional density $f(x|\mathcal{U})$ of the random variable X assuming \mathcal{U} is

$$f(x|\mathcal{U}) = \frac{dF(x|\mathcal{U})}{dx}.$$

Covariance and Correlation Coefficient

Covariance:

- The covariance C_{xy} of two random variables X and Y is defined as follows:

$$C_{xy} = E[(X - \mu_X)(Y - \mu_Y)].$$

Correlation Coefficient:

- The correlation coefficient ρ_{xy} of two random variables X and Y is defined as follows:

$$\rho_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}.$$

Independence vs. Uncorrelatedness

Independence:

- The random variables X and Y are called *independent* if

$$f_{XY}(x, y) = f_X(x) f_Y(y) \\ \left(f_{X|Y}(x|y) = f_X(x) \right).$$

Uncorrelatedness:

- The random variables X and Y are called *uncorrelated* if their covariance is zero, i.e.,

$$C_{xy} = 0, \rho_{xy} = 0, E[XY] = E[X] E[Y].$$

- If two random variables are *independent*, then they are *uncorrelated* (but vice versa).

Orthogonality:

- The random variables X and Y are said to be *orthogonal* if

$$E[XY] = 0.$$

Random Vector

- A *random vector* is a vector $\mathbf{X} : [X_1, \dots, X_n]$ whose components X_i are random variables.
- The probability that \mathbf{X} is in a region D of the n -dimensional space is

$$P(\mathbf{X} \in D) = \int_D f(x_1, \dots, x_n) dx_1 \cdots dx_n.$$

- The random variables X_1, \dots, X_n are (mutually) independent if the events $\{X_1 \leq x_1\}, \dots, \{X_n \leq x_n\}$ are independent, i.e.,

$$\begin{aligned} F(x_1, \dots, x_n) &= F(x_1) \cdots F(x_n) \\ f(x_1, \dots, x_n) &= f(x_1) \cdots f(x_n). \end{aligned}$$

Let X_1, X_2, \dots, X_n be *normally distributed* random variables having means $\mu_1, \mu_2, \dots, \mu_n$ and variances $\sigma_1, \sigma_2, \dots, \sigma_n$, respectively. Let a_1, a_2, \dots, a_n be constants. Then the random variable Y which is the linear combination of the X 's, i.e., $Y = a_1X_1 + a_2X_2 + \dots + a_nX_n$ is also normally distributed with mean μ_Y and variance σ_Y^2 , where

$$\begin{aligned} \mu_Y &= a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n \\ \sigma_Y^2 &= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2. \end{aligned}$$

INTRODUCTION TO STATISTICS

Basic Terminology

□ Point Estimate:

- A *point estimate* is a single number that is used as an estimate of a population parameter or population characteristic.
- Example:
 - Head appears 15 time in 25 independent tosses, then the estimated probability for appearing head is $\hat{p} = \frac{15}{25} = 0.6$.

□ Interval Estimate:

- An *interval estimate* is an interval that provides an upper and lower bound for a specific population parameter whose value is unknown.
- This interval estimate has an associated degree of confidence of containing the population parameters. Such interval estimates are also called **confidence intervals**.
 - e.g., $\mu = \bar{x} \pm 0.03$ with 95% confidence interval

$$P(|\mu - \bar{x}| \leq 0.03) \geq 0.95.$$

□ Estimators:

- An *estimator* is a random variable calculated from sample data that provides either point estimates or interval estimates for some population parameter.
- **Unbiasedness:**
 - An estimator $\hat{\theta}$ is *unbiased* if its mean is equal to the population parameter being estimated θ , i.e., $E[\hat{\theta}] = \theta$.
- **Efficiency:**
 - An estimator $\hat{\theta}$ of θ is said to be more *efficient* than any other unbiased estimator $\hat{\theta}$ if $Var(\hat{\theta}) \leq Var(\hat{\theta})$.
 - An estimator is a **minimum variance unbiased estimator** if the variance of its sampling distribution is the smallest of all other unbiased estimators.
- **Consistency:**
 - An estimator is said to be a *consistent estimator* if it approaches the parameter to be estimated *in a probability sense* as the sample size n gets large, i.e.,

$$\lim_{n \rightarrow \infty} P\left(\left|\hat{\theta}(n) - \theta\right| \geq \epsilon\right) = 0,$$

where ϵ is a small positive number.

Maximum Likelihood Estimation

- To choose a set of parameters θ in a way that maximizes the likelihood function $L(\theta)$:

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

where x_1, x_2, \dots, x_n is a set of random samples from the distribution of a random variable X with density f and associated parameter θ .

- The ML estimation $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$ is the set of estimated values that satisfies the equations

$$\frac{\partial L(\theta)}{\partial \theta_i} = 0, \quad i = 1, \dots, k.$$

Properties:

- Maximum-likelihood estimates are (1) consistent, and (2) asymptotically efficient.
- Let $\hat{\theta}_{ML}$ be a MLE of θ , then $g(\hat{\theta}_{ML})$ is a MLE of $g(\theta)$, i.e.,

$$\left[\widehat{g(\theta)} \right]_{ML} = g(\hat{\theta}_{ML}),$$

where $g(\cdot)$ is a monotonic function.

□ Examples:

- The MLE for the “success” probability p of Bernoulli distribution is

$$\hat{p}_{ML} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

where x_i ($= 0$ or 1) is the outcome of the i -th Bernoulli trial.

— \hat{p}_{ML} can be interpreted as the relative frequency of success over the n trials.

- The MLEs for the mean and variance of the normal density are:

$$\hat{\mu}_{ML} = \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$
$$\hat{\sigma}_{ML}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Bayesian Estimation

- To choose the parameters which maximizes the likelihood function $L(\pi(\theta), \theta)$:

$$L(\pi(\theta), \theta) = \pi(\theta) f(x_1, \dots, x_n | \theta),$$

where $\pi(\theta)$ is the prior probability density of θ before sampling.

- The Bayesian estimation $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$ is the set of estimated values that satisfies the equations

$$\frac{\partial L(\pi(\theta), \theta)}{\partial \theta_i} = 0, \quad i = 1, \dots, k.$$

- The Bayesian estimation $\hat{\theta}$ of parameter θ is the expected value of the parameter taken with respect to the *posterior distribution* of θ given the outcome of the random sample x , i.e.,

$$\hat{\theta} = E[\theta | x].$$

- Example:

- To estimate the mean μ of a normal density $N(\mu, \sigma^2)$ with the known value of σ . Let $\pi(\theta) \sim N(\mu_0, \alpha)$, then the Bayesian estimate $\hat{\mu}_{Baye}$ of μ is

$$\hat{\mu}_{Baye} = \frac{\sigma^2 \mu_0 + \alpha^2 n \bar{x}}{\sigma^2 + \alpha^2 n},$$

- n is the number of samples,
- \bar{x} is the sample mean.

Least Squares Estimation

- To estimate θ by the point $\tilde{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$ which makes the corresponding expected values $g_1(\theta), g_2(\theta), \dots, g_n(\theta)$ as close as possible to the observed samples x_1, x_2, \dots, x_n .
- The LS estimation $\tilde{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k)$ is the set of estimated values that satisfies the equations

$$x_i = g_i(\theta) + \epsilon_i, \quad 1 \leq i \leq n,$$
$$\frac{\partial}{\partial \theta_j} \sum_{i=1}^n [x_i - g_i(\theta)]^2 = 0, \quad j = 1, \dots, k.$$

Hypothesis Testing

Goal:

To make a binary decision on a hypothesis based on the given observations.

Hypotheses:

- **Null Hypothesis (H_0):**
The hypothesis that we are interested in rejecting or refuting.
- **Alternative Hypothesis (H_1):**
The contradictory hypothesis of H_0 .

Decision Regions:

The observation space is partitioned into acceptance region $R(H_0)$ and rejection region $R(H_1)$; if the observed features fall within the acceptance region, hypothesis H_0 is confirmed, otherwise, H_0 is rejected.

Types of errors:

- **Type I error:**
 H_0 is true but the observation suggests H_1
- **Type II error:**
 H_0 is false but the observation suggests H_0 .

□ **Level of Significance:**

The *level of significance*, denoted by α , is the maximum probability of making a *Type I error*.

□ **One-Tailed Test vs. Two-Tailed Test:**

For a test statistic T computed on the sample data:

- a *upper one-tailed test* has the decision rule:
— Reject H_0 if $T > T_U$; otherwise accept H_0 .
- a *lower one-tailed test* has the decision rule:
— Reject H_0 if $T < T_L$; otherwise accept H_0 .
- a *two-tailed test* has the decision rule:
— Reject H_0 if $T > T_U$ or $T < T_L$;
Accept H_0 , otherwise.
- The values of T_U and T_L are critical values that are selected so that the test will have the desired level of significance α .

□ **p-Value:**

The *p-value* associated with a test statistic is the *smallest* level of significance that would have allowed the null hypothesis to be rejected.

□ **Procedures:**

1. State the null hypothesis, H_0 .
2. State the alternative hypothesis , H_1 .
3. Decide on the level of significance, α .
4. Choose an appropriate testing procedure and determine the acceptance region.
5. Compute the test statistic from the sample data.
6. Make the decision: reject H_0 if the p-value is less than the level of significance α ; otherwise accept H_0 .

□ **Example:**

To test $H_0: p = 0.6$ against $H_1: p \neq 0.6$.

For a two-tailed test of the level of significance $\alpha = 0.05$, the critical values of the normal distribution are $T_U = 1.96$; $T_L = -1.96$.

Suppose the computed test statistic $T = 2.06$ which corresponds to the p-value of 0.0394.

We will accept H_0 of the level of significance $\alpha = 0.05$.

However, we will reject H_0 if the level of significance $\alpha = 0.01$.

Likelihood Ratio Test

□ The **likelihood ratio** λ :

$$\lambda = \frac{f_0(X)}{f_1(X)},$$

where

- H_0 : the pdf of the data is $f_0(X)$,
- H_1 : the pdf of the data is $f_1(X)$.

Accept H_0 if $\lambda > \lambda_T$ (λ_T is a preset threshold), otherwise accept H_A .

□ Example:

For automatic compound noun extraction [Su 94]:

- H_0 : the feature vector \vec{x} for the input pattern is generated by a compound model M_c .
- H_A : the feature vector for the input pattern is generated by a *non-compound* model M_{nc} .

the likelihood ratio λ is

$$\lambda = \frac{P(M_c | \vec{x})}{P(M_{nc} | \vec{x})}$$

INTRODUCTION TO STOCHASTIC PROCESS

- A *stochastic process* $\{X(t), t \in T\}$ is a collection of random variables; i.e, for each $t \in T$, $X(t)$ is a random variable.
- Interpretations:
 - A stochastic process $X(t) = X(t, \zeta)$ is a single time function (a sample of the given process) if ζ is fixed.
 - $X(t, \zeta)$ becomes a random variable equal to the *state* of the given process at time t , if t is fixed.
 - If t and ζ are fixed, then $X(t, \zeta)$ is a number.
- The set T is called the *index* set of the process.
 - $\{X(t)\}$ is a *discrete-time* process, if T is a countable set; e.g., $\{X_n, n = 0, 1, \dots\}$.
 - $\{X(t)\}$ is a *continuous-time* process, when T is an interval of the real line; e.g., $\{X(t), t \geq 0\}$.
- Example:
 - $\{X(t)\}$ might be equal to the total number of customers that have entered a supermarket by time t .

Markov Chains

□ A discrete-time discrete-state stochastic process $\{X_n, n = 0, 1, \dots\}$, having the property that given the present state, the past states have no influence on the future, is called a discrete-time *Markov chain*

□ The Markov property:

$$\begin{aligned} P(X_n = j \mid X_{n-1} = i, X_{n-2} = i_{n-2}, \dots, X_0 = i_0) \\ = P(X_n = j \mid X_{n-1} = i), \end{aligned}$$

- $P(X_n = j \mid X_{n-1} = i)$ are called the *transition probabilities* of the chain.

□ Example:

The formula for tagging part-of-speech is approximated as:

$$\max \prod_{i=1}^n P(w_i \mid t_i) \cdot P(t_i \mid t_{i-1}),$$

where t_i corresponds to the part-of-speech attached to the i -th word w_i .

- The probability $P(t_i \mid t_{i-1})$ in the above formula is the transition probability of the assumed Markov model.

INTRODUCTION TO INFORMATION THEORY

Entropy

Let each possible outcome x_k of a stationary source X occur with a probability of $P(x_k)$.

□ **Self-information** $I(x_k)$:

$$I(x_k) = -\log P(x_k)$$

- $I(x_k)$ is the amount of information associated with the known occurrence of output x_k .

□ **Entropy** $H(X)$:

$$H(X) = -\sum_i [P(x_i) \cdot \log P(x_i)]$$

- $H(X)$ is the average information (or uncertainty) of the source X .

Mutual Information

□ **Mutual Information** $I(x; y)$:

$$I(x; y) \equiv \log_2 \left[\frac{P(x, y)}{P(x) \cdot P(y)} \right]$$

- $I(x; y)$ is the information that the reception of y supplies about x .

□ **Example:** Use $I(w_x; w_y)$ as a measure for the preference of "strong economy" and "powerful economy" [K. Church 89]:

- $I(w_x; w_y) \gg 0$, w_x and w_y are highly associated.
- $I(w_x; w_y) \approx 0$, w_x and w_y are independent.
- $I(w_x; w_y) \ll 0$, w_x and w_y are in complementary distribution.

Perplexity

- The perplexity is a measure of the constraint imposed by the grammar, or the level of uncertainty given the grammar.
- Let $P(w | s)$ be the probability that w will be next word when the current state is s .

- The entropy, $H_s(w)$, associated with state s is

$$H_s(w) = -\sum_w P(w | s) \log_2 P(w | s).$$

- The entropy $H(w)$ of the task is the average value of $H_s(w)$, i.e.,

$$H(w) = \sum_s \pi(s) H_s(w),$$

where $\pi(s)$ is the probability of being in state s during the production of a sentence.

- The **perplexity** $S(w)$ of the task [Bahl 83]

$$S(w) = 2^{H(w)}.$$