ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

1

ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

APPLIED STATISTICS AND EXPERIMENTAL DESIGN

Elements of Statistics

2

TITLE AND CONTENT SLIDE

Elements of Statistics

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

3

III. Statistics – parameter estimation

- 3.1. Introduction to Statistics
- 3.2. Parameter estimation
- 3.3. Hypothesis testing

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

4

3.1. Introduction to Statistics

- Definitions
  - Probability – building an abstract models and its conclutions are deductions based on system of axioms
  - Statistics – applications of theory to real world problems and its conclutions are inferences based on observations.
    - Statistics: analysis + design
    - Analysis – mathematical statistics – involving repeated trials and events the probability of which close to 0 or 1.
    - Design –applied statistics deals with data collection and construction of experiments that can be adequately described by probabilistic models
    - Scope of study: mathematical statistics

5

3.1. Introduction to Statistics

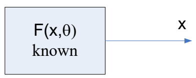
- Probabilistic concepts and reality
  - Probability of event A:
    - P(A) is estimated by $P(A) \approx N_A/N$
    - This empirical formula is used for relative frequency interpretation of all probabilistic concepts
  - Example:
    - the mean η of a r.v can be estimated by
    - $\eta\hat{} = (1/n)\sum x_i$, where x_i are observed value of a r.v X.
    - Distribution function $F_X(x)$ can be estimated by
    - $F_X\hat{}(x) = n_x/n$, where n_x is number of $x_i \leq x$.
    - The relationship are empirical point estimates of the parameter η and $F_X(x)$ and a major objective of statistics is to give them an exact interpretation

SOICT VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

6

3.1. Introduction to Statistics

- Problems of statistics:
  - First class of problems:
    - Predict future observations when probabilistic model is known.
    - We proceed from models to observations.
    - Example:
      - Distribution functtion $F_X(x)$ of X is known, and we wish to predict average $\bar{X}$ of its n future samples
      - Probability P of an event A is known, and we wish to predict number of occurences of A in n future trials

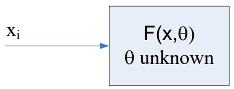


Predict x

SOICT VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

7

3.1. Introduction to Statistics



Estimate θ

- Second class of problems:
  - One or more parameters θ_i of the model are unknown
  - Estimate values of parameters (parameter estimation)
  - Decide whether θ_i is a set of known constants θ_{0i} (hypothesis testing).
  - We proceed from observations to models

SOICT VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

8

3.1. Introduction to Statistics

- Example:
  - A coin is tossed 1000 times and heads show 475 times.
    + Estimate value of the probability of heads or
    + Decide whether the coin is fair.
  - The values x_i of r.v X are observed.
    + Estimate mean η of X or
    + Decide whether to accept the hypothesis that η=5.3.

9

3.1. Introduction to Statistics

- Prediction problems
  - Given r.v X with known $F_X(x)$
    - Predict its value at future trials
  - Point prediction of X:
    - Determine constant c: (X-c) min.
    - If criterion of selecting *c* is to minimize the MS error $E\{(X-c)^2\}$, then $c=E\{X\}$ MSE -mean square error
      - $MSE = E\{(X-c)^2\} = E\{X^2- 2Xc + c^2\} = \psi(c)$ min.
      - $d\psi(c)/dc = 0$: $E\{2c – 2X\} = 0 \Rightarrow E\{c-X\} = 0 \Rightarrow$ $c – E\{X\} = 0$: $c = E\{X\}$.

10

3.1. Introduction to Statistics

- Interval prediction of X:
  - Determine constants c_1 and c_2 :
    $$P\{c_1<X<c_2\} = \gamma = 1 - \delta \quad (1)$$
    - Where γ is a given constant called the confidence coefficient. γ100%
  - Predict: $x_i \in (c_1, c_2)$,
    - Correct prediction in 100γ% of the cases.
  - Interval prediction: find c_1, c_2: $(c_2 – c_1)$ min and (1)

11

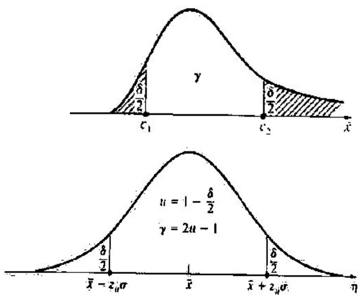
3.1. Introduction to Statistics

- Selection of γ :
  - If $\gamma \approx 1$, $X \in (c_1, c_2)$ is reliable, but $(c_2 – c_1)$ is large.
  - If $\gamma \downarrow \Rightarrow (c_2 – c_1) \downarrow$, the estimation is less reliable.
  - For optimum prediction: assign value to γ and determine c_1, c_2: $(c_2 – c_1)$ min and (1).
  - If $f_X(x)$ has single maximum $(c_2 – c_1)$ is minimum if $f_X(c_1) = f_X(c_2)$

12

3.1. Introduction to statistics

- Suboptimal solution: if we determine c_1 and c_2:
  $P\{X < c_1\} = \delta/2$ and $P\{X > c_2\} = \delta/2$
- $\Rightarrow c_1 = x_{\delta/2}$ and $c_2 = x_{1-\delta/2}$
- This solution is optimum if the $f_X(x)$ is symmetrical about its mean η
- If X is normal:
  $X_u = \eta + z_u\sigma$



3.2. Parameter estimation

- Parameter estimation problem
  - X – r.v with p.d.f $F_X(x, \theta)$ of known form which is depends on parameter θ.
  - θ - scalar or vector
  - Estimate parameter θ.
  - Repeat experiment *n* time and x_i is observed value of x.
  - Based on observed value, find point estimate and interval estimate of θ.

3.2. Parameter estimation

- Point estimate
  - Point estimate: $\hat{\theta} = g(X)$, X=[x_1, ..., x_n] – observation vector;
  - R.v $\hat{\theta} = g(X)$ - point estimator of θ;
  - Any function of vector X=[x_1, ..., x_n] – statistic;
  - Point estimator is statistic.
  - $\hat{\theta}$ - unbias estimator of parameter θ if $E\{\hat{\theta}\} = \theta$
  - Otherwise – biased estimator with bias $b = E\{\hat{\theta}\} - \theta$
  - If *g(X)* is properly selected, the estimation error ↓ when *n*↑.
  - If estimation error $\hat{\theta} - \theta$ $\hat{\theta} - \theta \to 0$ when $n \to \infty$ then $\hat{\theta}$ is called consistent estimator

3.2. Parameter estimation

- Example: sample mean $\bar{X}$ of X
  - $\bar{X}$ is unbiased estimator of η_X
  - Its variance $\sigma^2/n \to 0$ when $n \to \infty$
  - $\bar{X} \to \eta_X$ in MS sense, also in probability.
  - $\bar{X}$ is consistent estimator of η_X
- The best estimator $\hat{\theta} = g(X)$ MS error min

$$e = E\{[g(X) - \theta]^2\} = \int_R [g(X) - \theta]^2 f(X, \theta) dX$$

  - g(X) is usualy selected empirically.

3.2. Parameter estimation

- Empirically determination of g(X)
  - Suppose θ is the mean θ =E{q(X)} of some function q(X) of X.
  - Sample mean of q(X) is consistent estimator of θ

$$\hat{\theta} = \frac{1}{n}\sum_i q(x_i)$$

  - If sample mean of q(X) is used as the point estimator of θ, the estimate will be satisfactory at least for n large

3.2. Parameter estimation

- Interval Estimates
  - Definition
    - Interval estimate of parameter θ is an interval (θ_1, θ_2), the end points of which are function θ_1=g_1(X), θ_2=g_2(X) of the observation vector X.
    - Random interval (θ_1, θ_2) is an interval estimator of θ.
    - If $P\{\theta_1 < \theta < \theta_2\} < \gamma$, (2)
      (θ_1, θ_2) is γ confidence interval of θ.
    - The constant γ - confidence interval of the estimate and the difference $\delta = 1 - \gamma$ is confidence level
    - The objective of interval estimation is determination of functions g_1(X) and g_2(X) so as to minimize the length $(\theta_2 - \theta_1)$ subject to constrain (2)

3.2. Parameter estimation

- Mean estimate
  - R.v X with mean value η
  - The point estimate of mean value

$$\bar{x} = \frac{1}{n}\sum_i x_i$$

  - The interval estimator of mean value
    - Normality assumption of $\bar{X}$.
  - Known variance
    - Suppose that the variance σ^2 of x is known.
    - z_u – the u percentile of the standard normal density, we have:

$$P\{\eta - z_{1-\delta/2}\frac{\sigma}{\sqrt{n}} < \bar{x} < \eta + z_{1-\delta/2}\frac{\sigma}{\sqrt{n}}\} = G(z_{1-\delta/2}) - G(-z_{1-\delta/2}) = 1 - \frac{\delta}{2} - \frac{\delta}{2}$$

3.2. Parameter estimation

- Confidence coefficient γ:
  - η is in the interval $\bar{x} \pm z_{1-\delta/2}\sigma/\sqrt{n}$

$$P\{\bar{x} - z_{1-\delta/2}\frac{\sigma}{\sqrt{n}} < \eta < \bar{x} + z_{1-\delta/2}\frac{\sigma}{\sqrt{n}}\} = 1 - \delta = \gamma$$

- Determination of a confidence coefficient for η:
  - Observe the sample x_i của x
  - Form their average $\bar{x}$.
  - Select a number γ=1-δ
  - Find the standard percentile z_u for u=1-δ/2.
  - Form the interval $\bar{x} \pm z_u\sigma/\sqrt{n}$.
- If the discrete type r.v provided that n is large, this also holds.

3.2. Parameter estimation

- The choice of the confidence interval γ is dictated by two conflicting requirements:
  - If $\gamma \approx 1$, the estimate is reliable but the size $2z_u\sigma/\sqrt{n}$ of the confidence interval is large.
  - If γ is reduced, z_u is reduced, but the estimate is less reliable.
  - The final choice is a compromise bazed on the applications.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

21

3.2. Parameter estimation

- Tchebycheff inequality
  - Suppose that the distribution of $\overline{X}$ is unknown.
  - From Tchebycheff inequality:

$$P\{|x-\eta| \geq \varepsilon\} \leq \sigma^2/\varepsilon^2$$

  - Substitute X by $\overline{X}$, σ by $\sigma/\sqrt{n}$ and set $\varepsilon = \sigma/n\delta$
  - We have:

$$P\{\overline{x} - \frac{\sigma}{\sqrt{n\delta}} < \eta < \overline{x} + \frac{\sigma}{\sqrt{n}}\} > 1 - \delta = \gamma$$

  - This shows that, the exact γ confidence interval of η is contained in the interval ($\overline{X} \pm \sigma\sqrt{n\delta}$).

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

22

3.2. Parameter estimation

- Unknown variance σ^2
  - To estimate η:
    - Sample variance is unbiased estimator of variance σ^2.
    - It tens to σ^2 when $n \to \infty$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

    - For n large, we can use approximation $s \approx \sigma$
    - Confidence interval:

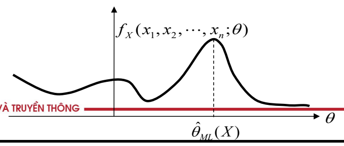
$$\overline{x} - z_{1-\delta/2}\frac{s}{\sqrt{n}} < \eta < \overline{x} + z_{1-\delta/2}\frac{s}{\sqrt{n}}$$

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

23

3.2. Parameter estimation

- Maximum likelihood estimation
  - R.v X has density $f(x, \theta)$.
  - Estimate θ in terms of a single observation of the r.v X.
  - Assume that the joint p.d.f of $X_1, \ldots, X_n$ given by $f_X(x_1, \ldots, x_n; \theta)$ depends on θ.
  - Observations $x_1, \ldots, x_n$ are given. The value of θ that maximizes f_X is the most likely value for θ.
  - This value is chosen as the ML estimate $\hat{\theta}_{ML}(X)$ for θ



VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

24

3.2. Parameter estimation

- Given $X_1 = x_1, \ldots, X_n = x_n$,
- The likelihood function $f_X(x_1, \ldots, x_n;\theta)$,
- Determination of the ML estimate by:

$$\sup_{\hat{\theta}_{ML}} f_X(x_1, x_2, \cdots, x_n; \theta)$$

- Or $L(x_1, x_2, \cdots, x_n; \theta) = \log f_X(x_1, x_2, \cdots, x_n; \theta).$
- If $L(x_1, \ldots, x_n;\theta)$ is differentiable and a supremum $\theta\hat{}_{ML}$ exists, then following equation must be satisfied:

$$\left.\frac{\partial \log f_X(x_1, x_2, \cdots, x_n; \theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{ML}} = 0.$$

3.2. Parameter estimation

- Example
  - $X_i = \theta + w_i$, i=1,...,n: n observations
    - θ - unknown parameter
    - w_i – n independent norrmal r.v with $\mu_i=0$ and variance σ^2.
    - ML estimate of θ ?
  - Solution:
    - Likelihood function
    - Each X_i is normal r.v with mean θ and variance σ^2

$$f_X(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f_{X_i}(x_i; \theta).$$

$$f_{X_i}(x_i; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\theta)^2/2\sigma^2}.$$

3.2. Parameter estimation

- Likelihood function.

$$f_X(x_1, x_2, \cdots, x_n; \theta) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{i=1}^{n}(x_i-\theta)^2/2\sigma^2}.$$

- Log-likelihood function:

$$L(X;\theta) = \ln f_X(x_1, x_2, \cdots, x_n; \theta) = \frac{n}{2}\ln(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(x_i-\theta)^2}{2\sigma^2},$$

- ML requirement:

$$\left.\frac{\partial \ln f_X(x_1, x_2, \cdots, x_n; \theta)}{\partial \theta}\right|_{\theta=\hat{\theta}_{ML}} = 2\sum_{i=1}^{n}\left.\frac{(x_i-\theta)}{2\sigma^2}\right|_{\theta=\hat{\theta}_{ML}} = 0,$$

- And we have

$$\hat{\theta}_{ML}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

3.2. Parameter estimation

- ML estimator is a r.v with expected value:

$$E[\hat{\theta}_{ML}(x)] = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \theta,$$

- This estimator is unbiased estimator for θ.
- The variance of the estimator:

$$Var(\hat{\theta}_{ML}) = E[(\hat{\theta}_{ML} - \theta)^2] = \frac{1}{n^2} E\left\{\left(\sum_{i=1}^{n} X_i - \theta\right)^2\right\}$$

$$= \frac{1}{n^2}\left\{\sum_{i=1}^{n} E(X_i - \theta)^2 + \underbrace{\sum_{i=1}^{n}\sum_{j=1,\, i\neq j}^{n} E(X_i - \theta)(X_j - \theta)}_{0}\right\}$$

3.2. Parameter estimation

- We have

$$Var(\hat{\theta}_{ML}) = \frac{1}{n^2}\sum_{i=1}^{n} E(X_i-\theta)^2 = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

- When n→∞,

$$Var(\hat{\theta}_{ML}) \to 0$$

The estimator is consistent

29

3.2. Parameter estimation

- Best Unbiased Estimator:
  - From last example of estimating mean, we have an unbiased estimator for θ with variance σ^2/n.
  - It is possible that, for a given *n*, there may be other unbiased estimators to this problem with even lower variances.
  - Question: In a given scenario, is it possible to determine the lowest possible value for the variance of *any* unbiased estimator?
    - A theorem by Cramer and Rao (Rao 1945; Cramer 1948) gives a complete answer to this problem.

30

3.2. Parameter estimation

- **Cramer - Rao Bound**:
  - Variance of any unbiased estimator $\hat{\theta}$ based on observations for θ must satisfy the lower bound

$$X_1 = x_1, \cdots, X_n = x_n$$

$$Var(\hat{\theta}) \geq \frac{1}{E\left(\frac{\partial \ln f_X(x_1,x_2,\cdots,x_n;\theta)}{\partial\theta}\right)^2} = \frac{-1}{E\left(\frac{\partial^2 \ln f_X(x_1,x_2,\cdots,x_n;\theta)}{\partial\theta^2}\right)}.$$

  - This important result states that the right side of the inequality acts as a lower bound on the variance of *all* unbiased estimator for θ, provided their joint p.d.f satisfies certain regularity restrictions.

31

3.2. Parameter estimation

- Any unbiased estimator whose variance coincides with that in inequality above, must be the best.
- Such estimates are known as *efficient* estimators.
- Example
  - Let examine whether $\hat{\theta}_{ML}$ for mean represents an efficient estimator. We have:

$$\left(\frac{\partial \ln f_X(x_1,x_2,\cdots,x_n;\theta)}{\partial\theta}\right)^2 = \frac{1}{\sigma^4}\left(\sum_{i=1}^{n}(X_i-\theta)\right)^2;$$

- and

$$E\left(\frac{\partial \ln f_X(x_1,x_2,\cdots,x_n;\theta)}{\partial\theta}\right)^2 = \frac{1}{\sigma^4}\left\{\sum_{i=1}^{n}E[(X_i-\theta)^2] + \sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}E[(X_i-\theta)(X_j-\theta)]\right\}$$
$$= \frac{1}{\sigma^4}\sum_{i=1}^{n}\sigma^2 = \frac{n}{\sigma^2},$$

- After substitution this into Cramer-Rao inequality, we obtain the Cramer - Rao lower bound for this problem to be σ^2/n

32

3.3. Hypothesis Testing

- Hypothesis test
  - Statistical hypothesis:
    - Assumption about the values of parameters of a statistical models
    - Hypothesis testing is process for establishing the validity of a hypothesis.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

33

3.3. Hypothesis Testing

- Problem
  - R.v X has known distribution function $F(x, \theta)$ depending on parameter θ.
  - Test assumption $\theta = \theta_0$ against $\theta \neq \theta_0$
  - Hypothesis $\theta = \theta_0$ – null hypothesis H_0
  - Hypothesis $\theta \neq \theta_0$ – alternative hypothesis H_1
  - $\theta \in \Theta_I$.
  - Simple hypothesis: Θ_I consists of single points
    - Otherwise – composite
  - Null hypothesis is simple in most cases.
- Hypohesis testing: whether observations reject null hypothesis ?

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

34

3.3. Hypothesis Testing

- Decision regions:
  - Based on observed sample x of X.
  - Suppose that under hypothesis H_0, the density $f(x, \theta_0)$ of the sample x is negligible in a certain region D_c of the sample space, taking significant values only in complement $\bar{D}_c$ of D_c.
  - This is resonable to reject H_0 if x in D_c and to accept H_0 if x is in $\bar{D}_c$.
  - The set D_c is called the critical region of the test and $\bar{D}_c$ is called the region of acceptance H_0.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

35

3.3. Hypothesis Testing

- Example
  - Experiment of fair coin tossing.
    - Toss a coin n times
    - The heads show k times
    - If k << n/2, so the coin is not fair
    - If $k \approx n/2$, so we can accept H_0.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

36

3.3. Hypothesis Testing

- Type of error, which may be occured depending on location of x.
  - First, suppose H_0 is true, if x $\in D_c$, we reject H_0 even though it is true.
    - Error type 1.
    - α - significance level of the test – the probability for the such an error

$$\alpha = P\{x \in D_c \mid H_0\}$$

    - The difference

$$1-\alpha = P\{x \notin D_c \mid H_0\}$$

Equals the probability that we accept H_0 when true.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

37

3.3. Hypothesis Testing

- Second, suppose that H_0 is false.
  - If $x \notin D_c$, we accept H_0 even though it is false.
  - Error type 2.
  - The probability for such an error is denoted by function $\beta(\theta)$, where θ is called the operating characteristics of the test.

$$\beta(\theta) = P\{x \notin D_c \mid H_1\}$$

  - The difference 1-$\beta(\theta)$ is the probability that we reject hypothesis H_0 when false.
  - $P(\theta)$ – power of the test:

$$P(\theta) = 1-\beta(\theta) = P\{x \notin D_c \mid H_1\}$$

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

38

3.3. Hypothesis Testing

- Critical region
  - The region D_c is chosen so as to keep the probabilities of both types of errors are small.
  - The selection of the region D_c proceeds as follows:
    - Assign value to typer I error α and search for region D_c of the sample space so as to minimize type II error probability for specific θ.
      - If the resulting value $\beta(\theta)$ is too large, increase α to its tolerable value.
      - If $\beta(\theta)$ still too large, increase the number *n* of samples.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

39

3.3. Hypothesis Testing

- The test is called most powerful if $\beta(\theta)$ is minimum.
  - In general, the critical region of a most powerful test depends on θ.
    - If it is the same for every $\theta \in \Theta_I$, the test is uniformly most powerful.
    - Such a test does not always exist.
    - The determination of the critical region of the most powerful test involve a search in the n-dimentional sample space.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

40

3.3. Hypothesis Testing

- Test statistic
  - Prior to any experiment, we select a function q = g(X) of a sample vector X.
  - We find a set R_c of the real line where under the hypothesis H_0, the density of q is negligible.
  - We reject H_0 if the value q=g(X) of q is in R_c.
  - R_c is the critical region of the test.
  - The r.v q is the test statistic.
  - In the selection of function g(X), we are giuded by the point estimate of θ.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

41

3.3. Hypothesis Testing

- Distribution
  - Hypothesis: the distribution function F(x) of a r.v X equals a given function $F_0(x)$.
    - H_0: $F(x) \equiv F_0(x)$;
    - H_1: $F(X) \neq F_0(x)$.
  - Kolmogoroff-Smirnov test
  - Anderson-Zanderling test
  - Chi-square χ^2 test

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

42

3.3. Hypothesis Testing

- Kolmogoroff-Smirnov test
  - Form the random process F^(x) as in the estimation problem and use as the test statistic the r.v

    $q = \max_x | F\hat{}(x) - F_0(x) |$ (1)
  - For a specific ζ, the function F^(x) is the empirical estimate of F(x) and it tends to F(x) as n tends to ∞. From this, it follows that:

  E{F^(x)} = F(x) and
  - It shows that, for large n, $\hat{F}(x) \underset{n\to\infty}{\rightarrow} F(x)$
    - q is close to 0 if H_0 is true and
    - It is close to $F(x) - F_0(x)$ if H_1 is true

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

43

3.3. Hypothesis Testing

- Conclusion: we must reject H_0 if q is larger than some constant c.
  - Constant c is determined in terms of the significance level $\alpha=P\{q>c|H_0\}$ and the distribution q.
  - Under hypothesis H_0, the test statistic q equals r.v w in equation $w=\max_x|F\hat{}(x)-F(x)|$.
  - Using Kolmogoroff approximation, we obtain:

    $$\alpha = P\{q > c \mid H_0\} = 1 - e^{-2nc^2}$$
- K-S test procedure:
  - Form the empirical estimate F^(x) of F(x)
  - Determine $q = \max_x|F\hat{}(x)-F_0(x)|$
  - Accept H_0 if and only if $q > \sqrt{-\frac{1}{2n}\ln\left(\frac{\alpha}{2}\right)}$
  - The resulting error type II error probability is reasonably small only if n large

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

44

3.3. Hypothesis Testing

- Chi-square χ^2 test:
  - Test procedures using the Chi-Square distribution
    - Goodness of Fit Test:
      - How close are sample results to the expected results ?
    - Test of Independence:
      - Are two variables of interest independent of each other ?

3.3. Hypothesis Testing

- Goodness of Fit) Test
  - Set up the null and alternative hypotheses.
  - Select a random sample and record the observed frequency, fi , for each of the k categories.
  - Assuming H0 is true, compute the expected frequency, ei , in each category by multiplying the category probability by the sample size.

3.3. Hypothesis Testing

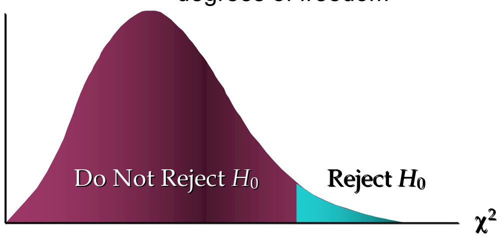
- Compute the value of the test statistic

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i}$$

  - Where
    - fi = observed frequency for category I
    - ei = expected frequency for category I
    - k = number of categories
  - Note: The test statistic has a chi-square distribution with k – 1 df provided that the expected frequencies are 5 or more for all categories.

3.3. Hypothesis Testing

- Rejection rules:
  - p-value approach: Reject H_0 if p-value < α
  - Critical value approach: Reject H_0 if $\chi^2 \geq \chi^2_\alpha$
  - where α is the significance level and there are k - 1 degrees of freedom




SOICT
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY
Thank you for
your attentions!
soict.hust.edu.vn/
fb.com/groups/soict

49