Basic Statistics

Dr. Sudipta Das

Department of Computer Science,
Ramakrishna Mission Vivekananda Educational & Research Institute

Outline I

1. Statistical Inference
   - Introduction
   - Interval Estimation

Chapter 8: Statistical Inference

Statistical Inference I

- The main objective in any statistical enquiry is the properties of one or more population.
    - However, the population(s) is (are) usually unknown to us, and we simply have a sample from the population (or, a sample from each of the given populations)

Statistical Inference II

- Statistical Inference:-
  Given the properties of the sample (or, of the samples), to infer about those of the population(s) is the problem of statistical inference
  - It is analogous to the inductive logic, the only difference being that the induction is achieved under probabilistic framework
    - Probability comes due to random sampling
  - It is a process of going over from the known sample to unknown population.

Statistical set-up of the problem of inference

- Let $(X_1, X_2, \dots, X_n)$ be a random sample of size n drawn from a population (discrete/continuous) with p.m.f/p.d.f $f(\underline{x}; \underline{\theta}) = f_{\underline{\theta}}(\underline{x})$, where $\underline{\theta}$ is the unknown parameter(s) of interest.
    - Our problem is to infer about $\underline{\theta}$
- Let Θ be the set of all possible values of θ
    - Θ is called the parameter space
- Note:
    - In the problem of statistical inference, Θ is known, although θ is unknown.
    - Example 1: $X_1, X_2, \dots, X_n \sim$ Bernoulli(p)
        - $\theta = p$ is unknown, $\Theta = [0, 1]$ is known
    - Example 2: $X_1, X_2, \dots, X_n \sim$ Normal(μ, σ)
        - $\underline{\theta} = [\mu, \sigma]'$ is unknown, $\Theta = (-\infty, \infty) \times (0, \infty)$ is known
        - $\theta = \mu$ is unknown, $\Theta = (-\infty, \infty)$ is known
        - $\theta = \sigma$ is unknown, $\Theta = (0, \infty)$ is known

- Statistical Inference
  1. Estimation
     - i Point Estimation
     - ii Interval Estimation
  2. Hypothesis-testing

1. Estimation:-
   Here, we have **no idea** about the true value of θ and the problem is **to estimate** the likely value of θ on the basis of the random sample $(X_1, X_2, \ldots, X_n)$ drawn from the population

i Point Estimation:-

Here, we estimate θ by a **single** value (i.e., by a point)

- Let $T = T(X_1, X_2, \ldots, X_n)$ be a statistic which is used to estimate the parameter θ, is called an **estimator** of θ
- For the observed sample $(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$, the observed value of the estimator, namely,

$$t = T(x_1, x_2, \ldots, x_n)$$

is called an estimate of θ

Statistical Inference VII

ii Interval Estimation:-

Here, we estimate θ by an **interval** of values

- Let $T_1 = T_1(X_1, X_2, \ldots, X_n)$ and $T_2 = T_2(X_1, X_2, \ldots, X_n)$ be two statistics such that

$$P[T_1 \leq \theta \leq T_2] = 1 - \alpha,$$

  where α is a pre-assigned small quantity. Usually, we take $\alpha = 0.05$ or 0.01 etc.
- If $\alpha = 0.05$, then $P[T_1 \leq \theta \leq T_2] = 0.95$. Hence, the observed values of $[T_1, T_2] = [t_1, t_2]$, say, is called a 95% confidence interval of θ

2 Hypothesis-testing:-
Here we have some idea about the true value of θ, in the form of a hypothesis, say, $\theta = \theta_0$,
- Our problem is **to judge or test** the validity/ feasibility/ tenability of the given hypothesis $\theta = \theta_0$ on the basis of random sample of the population

Chapter 8b: Interval Estimation

Interval Estimation I

- We estimate the parameter θ by an **interval** of values
  - Let
    $$L = T_1(X_1, X_2, \ldots, X_n) \text{ and } U = T_2(X_1, X_2, \ldots, X_n)$$
    be two statistics such that
    $$P[L \leq \theta \leq U] = 1 - \alpha,$$
    where α is a pre-assigned small quantity. Usually, we take $\alpha = 0.05$ or 0.01 etc.
  - The number $1 - \alpha$ is called the confidence coefficient
    - and the limits U and L, are called the upper and lower confidence limits, respectively.

Interval Estimation II

- The interval (L, U) is referred to as a $(1-\alpha)100\%$ confidence interval (CI) of the parameter θ.
- For example, if $\alpha = 0.05$, then $P[L \leq \theta \leq U] = 0.95$.
  - Hence, the observed values of $[L, U] = [l, u]$, (say), is called a 95% confidence interval of θ.

Interval Estimation III

- An interval (L, U) should have two properties:
  - $P(L < \theta < U)$ is high, that is, the true parameter θ is in (L, U) with high probability, and
  - the length of the interval (L, U) should be relatively narrow on the average.
- In summary,
  - Interval estimation goes a step beyond point estimation by providing, in addition to the estimating interval (L, U), a measure of one's confidence in the accuracy of the estimate.

Interval Estimation I

- Preliminary Notations
  Let $(x_1, \ldots, x_n)$ be the random sample drawn from the population of interest.
  1. n : Sample size
  2. $\bar{x}$: Sample mean (Calculated)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

  3. s^2 : Sample variance (Calculated)

$$s^2 = \frac{1}{n-1}\left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2\right]$$

  4. $\hat{p}$: Sample proportion (Calculated)

$$\hat{p} = \frac{\text{Total No. of sample units having the specific characteristic}}{n}$$

Interval Estimation II

Estimating population mean (μ)

- Pivotal quantity (Point Estimate): Sample mean ($\bar{x}$)
- Intervals:
  - $\left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$, where $z_{1-\frac{\alpha}{2}}$ is the $(1-\alpha/2)^{th}$ quantile from the standard normal distribution
    - Under the assumption: Population is Normal and variance (σ^2) is known
  - $\left[\bar{x} - t_{1-\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}, \bar{x} + t_{1-\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}\right]$, where $t_{1-\frac{\alpha}{2},n-1}$ is the $(1-\alpha/2)^{th}$ quantile from a $t-$ distribution with $n-1$ d.f.
    - Under the assumption: Population is Normal and variance is unknown
  - $\left[\bar{x} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$, where $z_{1-\frac{\alpha}{2}}$ is the $(1-\alpha/2)^{th}$ quantile from the standard normal distribution
    - Under the assumption: Sample size (n) is large

Interval Estimation III

- Example 6.2.3; (Page 302)
- Example 6.3.1; (Page 311)

Estimating population variance (σ^2)

- Pivotal quantity (Point Estimate): Sample variance (s^2)
- Intervals:
    - $\left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right]$, where $\chi^2_{1-\frac{\alpha}{2},n-1}$ and $\chi^2_{\frac{\alpha}{2},n-1}$ are the $(1-\alpha/2)^{th}$ and $(\alpha/2)^{th}$ quantiles from a χ^2- distribution with $(n-1)$ d.f., respectively.
        - Under the assumption: Population is Normal
- Example 6.4.1; (Page 317)

Interval Estimation V

Estimating population proportion (p)

- Pivotal quantity (Point Estimate): Sample proportion ($\hat{p}$)
- Intervals:
  - $\left[\hat{p} - z_{1-\frac{\alpha}{2}} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + z_{1-\frac{\alpha}{2}} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right]$, where $z_{1-\frac{\alpha}{2}}$ is the $(1-\alpha/2)^{th}$ quantile from the standard normal distribution
    - Under the assumption: Sample size (n) is large
      Both $np > 5$ and $n(1-p) > 5$
- Example 6.2.4; (Page 303)

Interval Estimation VI

- Width of a $(1-\alpha)100\%$ CI, for the true proportion (p)

$$b = 2z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

  - Note: - $b = 2z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{n}}$.

- Margin of error at $(1-\alpha)100\%$ CI, for the true proportion (p)

$$100d\%,$$

where

$$d = \frac{\max b}{2} = \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}$$

- Note: - Width and margin of error reduce as sample size increases.

Interval Estimation VII

Sample size selection without pilot study

- To estimate p at level $(1-\alpha)$ to within d (given) units of its true value
  - $\frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}} \leq d \Rightarrow n \geq \frac{z^2_{1-\frac{\alpha}{2}}}{4d^2}$
- Thus,

$$n = \left\lceil \frac{z^2_{1-\frac{\alpha}{2}}}{4d^2} \right\rceil$$

Interval Estimation VIII

Sample size selection after pilot study

- Sometimes, we may have an initial estimate $\tilde{p}$ of the parameter p from a pilot study or simulation.
- In this case, $d = \frac{b}{2} = z_{1-\frac{\alpha}{2}} \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$ and

$$n = \left\lceil \frac{z^2_{1-\frac{\alpha}{2}} \tilde{p}(1-\tilde{p})}{d^2} \right\rceil .$$

Example:

- Suppose that a local TV station in a city wants to conduct a survey to estimate support for the president's policies on economy within 3% error with 95% confidence.
  - The number of people should be surveyed by the station, if they have no information on the support level is

$$n = \left\lceil \frac{z^2_{1-\frac{\alpha}{2}}}{4d^2} \right\rceil = \left\lceil \frac{1.96^2}{4 \times 0.03^2} \right\rceil = 1068$$

  (b) Suppose they have an initial estimate that 70% of the people in the city support the economic policies of the president. Then, the number of people should be surveyed by the station is

$$n = \left\lceil \frac{z^2_{1-\frac{\alpha}{2}} \tilde{p}(1-\tilde{p})}{d^2} \right\rceil = \left\lceil \frac{1.96^2 \times 0.7 \times (1-0.7)}{0.03^2} \right\rceil = 897$$

Interval Estimation X

Interval Estimation

To Estimate	Assumptions	Interval
Mean	Population is Normal Variance (σ^2) is known	$\left[\bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right]$
	Population is Normal Variance is unknown	$\left[\bar{x} + t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}, \bar{x} - t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}\right]$
	Sample size (n) is large	$\left[\bar{x} + z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{x} - z_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}\right]$
Proportion	$np > 5$ as well as $n(1-p) > 5$	$\left[\hat{p} + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$
Variance	Population is Normal	$\left[\frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}, \frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2},n-1}}\right]$

Interval Estimation XI

- $(1-\alpha)\times 100$: Confidence level in percentage
- $z_{\frac{\alpha}{2}}$: $\alpha/2^{th}$ quantile from standard normal distribution

$$z_{\frac{\alpha}{2}} = qnorm\left(\frac{\alpha}{2}, mean = 0, sd = 1\right)$$

- $t_{\frac{\alpha}{2},n-1}$: $\alpha/2^{th}$ quantile from $t-$distribution with n-1 $d.f.$

$$t_{\frac{\alpha}{2},n-1} = qt\left(\frac{\alpha}{2}, df = n-1\right)$$

- $\chi^2_{\frac{\alpha}{2},n-1}$: $\alpha/2^{th}$ quantile from χ^2-distribution with n-1 $d.f.$

$$\chi^2_{\frac{\alpha}{2},n-1} = qchisq\left(\frac{\alpha}{2}, df = n-1\right)$$

- $\chi^2_{1-\frac{\alpha}{2},n-1}$: $(1-\alpha/2)^{th}$ quantile from χ^2-distribution with n-1 $d.f.$

$$\chi^2_{1-\frac{\alpha}{2},n-1} = qchisq\left(1-\frac{\alpha}{2}, df = n-1\right)$$