Lecture - Confidence Interval

Dr. Arabin Kumar Dey

Assistant Professor
Department of Mathematics
Indian Institute of Technology Guwahati

August 27 - 28, 2013

Outline

Outline

Statistical Inference

- Populations and samples
- Sampling distributions

- Statistical inference is "the attempt to reach a conclusion concerning all members of a class from observations of only some of them."
- A population is a collection of observations - A parameter is a numerical descriptor of a population
- A sample is a part or subset of a population - A statistic is a numerical descriptor of the sample

Population Vs Sample

Polulation

- population size = N
- μ = mean, a measure of center
- σ^2 = variance, a measure of dispersion
- σ = standard deviation

Sample from the population is used to calculate sample estimates (statistics) that approximate population parameters.

- sample size = n
- $\bar{X}$ = sample mean
- s^2 = sample variance
- s = sample standard deviation.

- Usually μ is unknown and we would like to estimate it
- We use $\bar{X}$ to estimate μ
- We know the sampling distribution of $\bar{X}$.

Definition: Sampling distribution The distribution of all possible values of some statistic, computed from samples of the same size randomly drawn from the same population, is called the **sampling distribution** of that statistic

When sampling from a normally distributed population

- $\bar{X}$ will be normally distributed
- The mean of the distribution of X is equal to the true mean μ of the population from which the samples were drawn
- The variance of the distribution is $\frac{\sigma^2}{n}$, where σ^2 is the variance of the population and n is the sample size
- We can write: $X \sim N(\mu, \frac{\sigma^2}{n})$

When sampling from a population whose distribution is **not normal** and the sample size is **large**, use the **Central Limit Theorem**.

Central Limit Theorem

Given a population of any distribution with mean, μ, and variance, σ^2, the sampling distribution of $\bar{X}$, computed from samples of size n from this population, will be approximately $N(\mu, \frac{\sigma^2}{n})$ when the sample size is large

- In general, this applies when $n \geq 25$
- The approximation of normality becomes as better as n increases.

What if a random variable has a Binomial distribution ?

- First, recall that a Binomial variable is just the sum of n Bernoulli variable: $S_n = \sum_{i=1}^{n} X_i$
- Notation: $S_n \sim Binomial(n, p)$
  $X_i \sim Bernoulli(p) = Binomial(1, p) \text{for} i = 1, \cdots, n$
- In this case, we want to estimate p by $\hat{p}$ where
  $\hat{p} = \frac{S_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$
- $\hat{p}$ is just a sample mean.
- So we can use the central limit theorem when n is large.

Binomial CLT

- For a Bernoulli variable $\mu = \text{mean} = p$
  $\sigma^2 = \text{variance} = p(1-p)$
-
  $$\bar{X} \approx N(\mu, \frac{\sigma^2}{n})$$
  as before.
- Equivalently,
  $$\hat{p} \approx N(p, \frac{p(1-p)}{n})$$

Distribution of Differences

Often we are interested in detecting a difference between two populations

- Differences in average income by neighborhood
- Differences in disease cure rates by age

Population 1 : Sample of size n_1 from population Size $= N_1$ Mean $= \mu_{X_1} = \mu_1$ Mean $= \mu_1$ Standard deviation $= \sqrt{\frac{\sigma_1^2}{n_1}} = \sigma_{\bar{X}_1}$
Standard deviation $= \sigma_1$

Population 2 : Sample of size n_2 from population Size $= N_2$ Mean $= \mu_{\bar{X}_2} = \mu_2$ Mean $= \mu_2$ Standard deviation $= \sqrt{\frac{\sigma_2^2}{n_2}} = \sigma_{\bar{X}_2}$
Standard deviation $= \sigma_2$

Distribution of Differences : CLT results

Now by CLT, for large n,

- $\bar{X}_1 \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$
- $\bar{X}_2 \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$
- $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

Difference in Proportion

We're done if the underlying variable is continuous. What if the underlying variable is Binomial?

- Then $\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$ is replaced by
- $N(\mu_1 - \mu_2, \frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2})$

Summary of Sampling Distributions

	Sampling Distribution	
Statistic	Mean	Variance
$\bar{X}$	μ	$\frac{\sigma^2}{n}$
$\bar{X}_1 - \bar{X}_2$	μ_1 - μ_2	$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$
$\hat{p}$	p	$\frac{pq}{n}$
$n\hat{p}$	np	npq
$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

What do we mean by Estimation

Point Estimation :

- An *estimator* of a population parameter: a statistic (i.e., $\bar{X}$, p)
- An *estimate* of a population parameter: the value of the estimator for a particular sample
    - From a sample of 100 infants, sample mean birth weight was $\bar{X} = 3012$ grams
    - From a sample of 100 Vitamin A treated girls, 2 died so $\hat{p} = \frac{2}{100} = 0.02$

Interval Estimate A point estimate plus an interval that expresses the uncertainty or variability associated with the estimate
$100(1 - \alpha)\%$ Confidence interval:
estimate $\pm$ (critical value of z or t) $\times$ (standard error)

Example

Confidence interval for the population mean Plugging in the values, we get

$$\bar{X} \pm z_{\alpha/2} \times \sigma_{\bar{X}} = [L, U]$$

Note: The $z_{\alpha/2}$ is the value such that under a standard normal curve the area under the curve that is larger than $z_{\alpha/2}$ is $\alpha/2$ and the area under the curve that is less than $-z_{\alpha}/2$ is $\alpha/2$

Derivation of Confidence Interval (CI) for the mean

We get the 100(1 - α)% confidence interval for μ by taking:

- $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$, in later slides, we show $z_{\alpha/2}$ is the most rational choice.
- $P(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma_{\bar{X}}} \leq z_{\alpha/2}) = 1 - \alpha$
- $P(-z_{\alpha/2} \cdot \sigma_{\bar{X}} \leq \bar{X} - \mu \leq z_{\alpha/2} \cdot \sigma_{\bar{X}}) = 1 - \alpha$

After some algebra:

$$P(\bar{X} - z_{\alpha/2} \cdot \sigma_{\bar{X}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \sigma_{\bar{X}}) = 1 - \alpha$$

$$P(L \leq \mu \leq U) = 1 - \alpha$$

Summary: CI for mean

A 100(1 - α)% confidence interval for μ, the population mean, is given by the interval estimate

$$\bar{X} \pm z_{(\alpha/2)} \cdot \frac{\sigma}{\sqrt{n}}$$

when the population variance σ^2 is known.
In this class, we'll always use 100(1 - α)% $=$ 95% confidence intervals, but you might sometimes see 90% or 99% CI in the literature.

Observations

- As sample size increases, width of confidence interval gets shorter.
- Width of the confidence interval decreases, as standard deviation σ decreases.
- Confidence level increases as width of the confidence interval increases.
- There should be some trade off between confidence level and width of the confidence interval. **Our strategy would be finding the shortest CI so that we can attain a desired confidence level. [discussed in later slides]**

Interpretation of the CI for μ

- Before the data are observed, the probability is at least $(1 - \alpha)$ that $[L, U]$ will contain μ, the population parameter
- In repeated sampling from a normally distributed population, 100(1 - α)% of all intervals of the form above will include the the population mean μ.

Coverage Probability

- Simulated probability that the constructed interval will include true parameter μ or in repeated sampling, it is the percentage of all constructed intervals that will include the true parameter μ.
- Coverage probability plays an important role in determining the sample size in case of asymptotic confidence intervals.

CI with shortest length

Problem is: for a given confidence coefficient $(1-\alpha)$, find the CI with the shortest length.

Example : $X_1, X_2, X_3, \cdots, X_n \sim$ i.i.d. $N(\mu, \sigma^2)$ with σ^2 known.

Let's take $Z = \sqrt{n}\frac{(\bar{X}_n-\mu)}{\sigma}$ is pivotal, therefore any (a, b) satisfying

$$P(a \le Z_n \le b) = \Phi(b) - \Phi(a) = 1-\alpha \qquad (1.1)$$

yields a corresponding $(1-\alpha)$-CI for μ :

$$\{\mu : \bar{X}_n - b\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n - a\frac{\sigma}{\sqrt{n}}\}$$

Now we want to choose (a, b) so that $b-a$ is the shortest length possible, for a given confidence coefficient $(1-\alpha)$.

Taking derivative $L = b - a$ with respect to a, we get
$\frac{dL}{da} = \frac{db}{da} - 1 = 0$
Also derivative of (1.1), we get $\phi(b)\frac{db}{da} - \phi(a) = 0$
Therefore, $\phi(b) = \phi(a)$ implies $b = -a$ and The symmetric solution is

$$\begin{aligned} 1 - \alpha &= \Phi(-a) - \Phi(a) = 1 - 2\Phi(a) \\ \implies & a = \Phi^{-1}(\frac{\alpha}{2}) \end{aligned}$$

Unknown Variance Assumption

- Sampling from a normally distributed population with population variance unknown
- We can make use of the sample variance s^2 Now we construct the confidence interval as:
  - $\bar{X} \pm z_{(\alpha/2)} \cdot s_X$ when n is "large"
  - $\bar{X} \pm t_{(\alpha/2,n-1)} \cdot s_X$ when n is "small"
- Estimate σ^2 with s^2 Here, $s_X = \frac{\sigma}{\sqrt{n}}$ and $t_{\alpha/2}$ has $n-1$ degrees of freedom
- The distribution of $\bar{X}$ is not quite normal, so we need the t-distribution