Module 3 (Estimators and their distributions)

Selvakkadunko Selvaratnam

University of Toronto

Introduction

Statistical methods are particularly useful for studying, analyzing, and learning about populations of experimental units.

Experimental unit

An experimental (or observational) unit is an object (e.g., person, thing, transaction, or event) about which we collect data.

Population

A population is a set of all units (usually people, objects, transactions, or events) that we are interested in studying.
For example, populations may include

1. all employed workers in the United States,
2. all registered voters in California,
3. everyone who is afflicted with AIDS,
4. all the cars produced last year by a particular company.

Introduction

Variable

A variable is a characteristic or property of an individual experimental (or observational) unit in the population.
For example, we may be interested in the variables age, gender, and number of years of education of the people currently unemployed in the United States.

Measurement

Measurement is the process we use to assign numbers to variables of individual population units.

Sample

A sample is a subset of the units of a population.
For example, instead of polling all 145 million registered voters in the United States during a presidential election year, a pollster might select and question a sample of just 1,500 voters.

Introduction

Statistical inference

A statistical inference is an estimate, prediction, or some other generalization about a population based on information contained in a sample.

Observed sample

We consider the random variables $Y_1, Y_2, \cdots, Y_n$ observed in a random sample from a population of interest. The random variables $Y_1, Y_2, \cdots, Y_n$ are independent and have the same distribution that has a population mean μ and population variance σ^2. Suppose that a random sample of n observations $y_1, y_2, \cdots, y_n$ collected from a target population. We are interested to estimate the population mean μ.

Sample mean

The sample mean can be calculated by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Introduction

Sample variance

The sample variance can be computed by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Population parameter

A population parameter is an unknown numerical value. We are interested to conduct a statistical inference about the population parameter. For instance, the population mean μ and variance σ^2 are the examples of population parameters.

Statistic

A statistic is a function of the observable random variables in a sample and known constants. For instance, the sample mean $\bar{y}$ and the sample variance s^2 are the examples for statistic.

Point Estimation

Estimator

An **estimator** is a rule, often expressed as a formula, that tells how to calculate the value of an estimate based on the measurements contained in a sample.
For example, the sample mean

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

is one possible point estimator of the population mean μ.

The Bias and Mean Square Error of Point Estimators

- Let $\hat{\theta}$ be a point estimator for a parameter θ. Then $\hat{\theta}$ is an **unbiased estimator** if $E(\hat{\theta}) = \theta$. If $E(\hat{\theta}) \neq \theta$, $\hat{\theta}$ is said to be **biased**.
- The **bias** of a point estimator $\hat{\theta}$ is given by $B(\hat{\theta}) = E(\hat{\theta}) - \theta$.
- The **mean square error** of a point estimator $\hat{\theta}$ is

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E[(\hat{\theta} - \theta)^2] \\ &= V(\hat{\theta}) + [B(\hat{\theta})]^2. \end{aligned}$$

Interval estimator

Confidence Intervals

- An **interval estimator** is a rule specifying the method for using the sample measurements to calculate two numbers that form the endpoints of the interval. Ideally, the resulting interval will have two properties: First, it will contain the target parameter θ; second, it will be relatively narrow.
- Interval estimators are commonly called **confidence intervals**. The upper and lower endpoints of a confidence interval are called the **upper and lower confidence limits,** respectively. Suppose that $\hat{\theta}_L$ and $\hat{\theta}_U$ are the (random) lower and upper confidence limits, respectively, for a parameter θ. Then, if

$$P(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U) \quad = \quad 1 - \alpha,$$

  the probability $(1 - \alpha)$ is the **confidence coefficient**.
- The endpoints for a $100(1 - \alpha)\%$ confidence interval for θ are given by

$$\hat{\theta}_L = \hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}} \text{ and } \hat{\theta}_U = \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}}.$$

  where $\sigma_{\hat{\theta}} = SE(\hat{\theta})$ and SE represents the standard error. Also, $SE(\hat{\theta})$ can be obtained by estimating the standard deviation of $\hat{\theta}$.

Examples

e.g 1:

Let $y_1, y_2, \cdots, y_n$ be a random sample with $E(y_i) = \mu$ and $V(y_i) = \sigma^2$. Show that $S^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is a biased estimator for σ^2 and that $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is an unbiased estimator for σ^2.

e.g 2:

A sample of $n = 1000$ voters, randomly selected from a city, showed $y = 560$ in favor of candidate Jones. Estimate p, the fraction of voters in the population favoring Jones, and place a 2-standard-error bound on the error of estimation.

The sampling distribution of the sample mean $\bar{y}$

The main objective of an experiment is to conduct the statistical inference about the population mean μ. We do not know any information about the population. But, we have a sample information. Then, how we can connect the sample information to population? A sampling distribution is a bridge between the sample information and population.

e.g 3:

Consider the experiment that is toss a fair die. Let y be the number of dots showing on the up face and suppose the die is tossed two times. Find

(a) the sampling distribution of the sample mean $\bar{y}$,

(b) the population mean of this sampling distribution $\mu_{\bar{y}}$, and

(c) the population variance of this sampling distribution $\sigma^2_{\bar{y}}$.

The sampling distribution of the sample mean $\bar{y}$

Properties of the Sampling Distribution of $\bar{y}$

A random sample of size n has been chosen from any population with mean μ and standard deviation σ. Let $y_1, y_2, \cdots, y_n$ are the sample values.

1. The mean of the sampling distribution $\bar{y}$, $\mu_{\bar{y}} = E(\bar{y}) = \mu$.
2. The standard deviation of the sampling distribution $\bar{y}$, $\sigma_{\bar{y}} = \sigma/\sqrt{n}$ and is called to as the **standard error of the mean**.
   We found the mean and the standard deviation of $\bar{y}$. Next, **what is the distribution of** $\bar{y}$?

The sampling distribution of the sample mean $\bar{y}$

Introduction

The central limit theorem is a refinement of the law of large numbers. For a large number n of independent identically distributed random variables $y_1, y_2, \cdots, y_n$, with finite variance, the average $\bar{y}$ approximately has a normal distribution, no matter what the distribution of they_i is.

Central Limit Theorem

Let $y_1, y_2, \cdots, y_n$ be independent identically distributed random variables with $E(y_i) = \mu$ and $V(y_i) = \sigma^2 < \infty$. Define

$$Z_n = \frac{(\bar{y} - \mu)}{\sigma/\sqrt{n}}.$$

Then Z_n follows the standard normal distribution for a large sample size n. That is, $Z_n \sim N(0, 1)$ for a large n.

The sampling distribution of the sample mean $\bar{y}$

e.g 4:

Achievement test scores of all high school seniors in a state have mean 60 and variance 64. A random sample of $n = 100$ students from one large high school had a mean score of 58. Is there evidence to suggest that this high school is inferior? (Calculate the probability that the sample mean is at most 58 when $n = 100$.)

e.g 5:

The service times for customers coming through a checkout counter in a retail store are independent random variables with mean 1.5 minutes and variance 1.0. Approximate the probability that 100 customers can be served in less than 2 hours of total service time.

The sampling distribution of the sample mean $\bar{y}$

e.g 6:

A bottling machine can be regulated so that it discharges an average of μ ounces per bottle. It has been observed that the amount of fill dispensed by the machine is normally distributed with $\sigma = 1.0$ ounce. A sample of $n = 9$ filled bottles is randomly selected from the output of the machine on a given day (all bottled with the same machine setting), and the ounces of fill are measured for each. Find the probability that the sample mean will be within 0.3 ounce of the true mean μ for the chosen machine setting.

The Sampling Distribution of the Sample Proportion

COROLLARY

Consider an event A in the sample space of some experiment with $p = P(A)$. Let y = the number of times A occurs when the experiment is repeated n independent times, and define the sample proportion $\hat{p} = \frac{y}{n}$, then

(i) $E(\hat{p}) = p$

(ii) $V(\hat{p}) = \frac{p(1-p)}{n}$ and $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

(iii) As n increases, the distribution of $\hat{p}$ approaches a normal distribution. In practice, $\hat{p}$ is approximately normal, provided that $np \geq 10$ and $n(1-p) \geq 10$.

e.g 7:

Suppose that 60% of all the city's voters favor the candidate. In a random sample of 100 voters, what is the probability that fewer than half are in favor of the candidate?

GOSSET'S THEOREM

t-distribution

If $y_1, y_2, \cdots, y_n$ is a random sample from a $N(\mu, \sigma)$ distribution, then a random variable

$$\frac{\bar{y} - \mu}{s/\sqrt{n}}$$

has the t distribution with $(n-1)$ degrees of freedom, t_{n-1}.

e.g 8:

The tensile strength for a type of wire is normally distributed with unknown mean μ and unknown variance σ^2. Six pieces of wire were randomly selected from a large roll; Y_i, the tensile strength for portion i, is measured for $i = 1, 2, \cdots, 6$. The population mean μ and variance σ^2 can be estimated by $\bar{y}$ and s^2, respectively. Because $\sigma^2_{\bar{y}} = \sigma^2/n$, it follows that $\sigma^2_{\bar{y}}$ can be estimated by s^2/n. Find the appropriate probability that $\bar{y}$ will be within $2s/\sqrt{n}$ of the true population mean μ.

Chi-squared Distribution

Result

Let $y_1, y_2, \cdots, y_n$ be a random sample from a normal distribution with mean μ and variance σ^2. Then

$$\frac{(n-1)s^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

has a χ^2 Distribution with $(n-1)$ degrees of freedom (df).

e.g 9

In e.g 6, the ounces of fill from the bottling machine are assumed to have a normal distribution with $\sigma^2 = 1$. Suppose that we plan to select a random sample ten bottles and measure the amount of fill in each bottle. If these ten observations are used to calculate s^2, it might be useful to specify an interval of values that will include s^2 with a high probability. Find numbers b_1 and b_2 such that

$$P(b_1 \leq s^2 \leq b_2) = 0.90.$$

F-Distribution

Results

Let W_1 and W_2 be **independent** χ^2- distributed random variables with ν_1 and ν_2 df, respectively. Then

$$F = \frac{W_1/\nu 1}{W_2/\nu_2}$$

is said to have an F distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom

F-Distribution

e.g 10

If we take independent samples of size $n_1 = 6$ and $n_2 = 10$ from two normal populations with equal variances, find the number b such that

$$P\left(\frac{s_1^2}{s_2^2} \leq b\right) = 0.95.$$