Stat 113 (Introduction to Mathematical Statistics)

Lesson 4.1 Point and Confidence Interval Estimation

Learning Outcomes

At the end of the lesson, students must be able to

1. Articulate the principle of statistical inference;
2. Differentiate point and interval estimation;
3. Explain the basic properties of estimators; and
4. Calculate point and interval estimates of common population parameters.

Introduction

The real power of statistics comes from applying the concepts of probability to situations where you have data but not necessarily the whole population. The results, called statistical inference, give you probability statements about the population of interest based on that set of data.

Statistical inference is the act of generalizing from the data (**sample**) to a larger phenomenon (**population**) with calculated degree of certainty. The act of generalizing and deriving statistical judgments is the process of inference.

There are two types of statistical inferences: *Estimation* and *Statistical Tests*.

Estimation uses the information from the sample to estimate (or predict) the parameter of interest.

For instance, using the result of a poll about the president's current approval rating to estimate (or predict) his or her true current approval rating nationwide.

Statistical tests use information from the sample to determine whether a certain statement about the parameter of interest is true. Statistical tests are also referred to as *hypothesis tests*.

For instance, suppose a news station claims that the President's current approval rating is more than 75%. We want to determine whether that statement is supported by the poll data.

The two common forms of estimation are *point* and *interval* estimation.

In point estimation we use a single number as estimate of the unknown parameter. An example of a point estimate is the sample mean or the sample proportion.

In interval estimation we provide lower and upper limits of the plausible values that the parameter may assume with some degree of confidence.

Before going any further it is very important to understand the difference between the following 3 concepts: *parameter*, *estimator*, and *estimate*:

The **parameter** is the unknown value of interest. It is an unknown number. Examples of parameters include:

a. Success probability (p) in a binomial distribution

b. The mean (μ) of the normal distribution

c. The standard deviation (σ) of the normal distribution

The **estimator** is a random variable that can take different values depending on the sample. This is not a number, but a random variable. Examples of a point estimators are:

a. Sample mean: $\overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$

b. Sample proportion: $\hat{p} = \frac{\sum_{i=1}^{n} Y_i}{n}$, where $Y_i = 1,$ if a "success" and $Y_i = 0,$ if a "failure"

c. Sample standard deviation: $s = \sqrt{\frac{\sum_{i=1}^{2} (Y_i - \overline{Y})^2}{n-1}}$

The **estimate** is the particular value that the estimator takes in the sample we have at hand. Given the sample, this is a known number.

Example 4.1.1

Suppose we are interested in the mean of a random variable Y which follows a $N(\mu, 1)$ distribution. Consider the following **random** observations Y_1, Y_2, Y_3, Y_4, Y_5 (data): -0.4336, 0.3426, 3.5784, 2.7694, -1.3499.

SOLUTION

Parameter: μ

Estimator: $\overline{Y} = \frac{Y_1+Y_2+Y_3+Y_4+Y_5}{5}$

Estimate: $\overline{y} = 0.98138$

Remarks:

1. Point estimates are calculated from data; parameters are not.
2. Point estimates vary from sample to sample; parameters do not.
3. Point estimators are random variables; parameters are constants.

There are more than one estimator for a parameter of interest. We must study the properties of these estimators in order to be able to choose. In determining what makes a good estimator, there are two key features:

1. The center of the sampling distribution for the estimate is the same as that of the population. When this property is true, the estimate is said to be **unbiased**. The most often-used measure of the center is the mean.
2. The estimate has the smallest standard error when compared to other estimators (**relative efficiency**). For example, in the normal distribution, the mean and median are essentially the same. However, the standard error of the median is about 1.25 times that of the standard error of the mean. We know the standard error of the mean is $\frac{\sigma}{\sqrt{n}}$. Therefore in a normal distribution, the SE(median) is about 1.25 times $\frac{\sigma}{\sqrt{n}}$. This is why the mean is a better estimator than the median when the data is normal (or approximately normal).

More generally, let θ denote a parameter. Then an estimator $\hat{\theta}$ is said to be unbiased for θ if

$$E(\hat{\theta}) = \theta \tag{1}$$

Any estimator $\hat{\theta}$ which does not satisfy (1) is said to be biased for θ.

Now, suppose we have two unbiased estimators for the parameter θ, say, $\hat{\theta}_1$ and $\hat{\theta}_2$. That is, $E(\hat{\theta}_1) = \theta,$ and $E(\hat{\theta}_2) = \theta$.

Which estimator shall we choose?

We would prefer the estimator with smaller variance as in this way the estimates will be more concentrated around their mean, which is the true value of the parameter.

We say that an unbiased estimator is **efficient** among unbiased estimators if it attains the smallest variance.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be unbiased estimators for θ. We say $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if

$$Var(\hat{\theta}_1) < Var(\hat{\theta}_2)$$

The ratio of these variances is referred to as **relative efficiency**. The relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is

$$RE(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} \tag{2}$$

We interpret the value of (2) as follows:

- If $RE(\hat{\theta}_1, \hat{\theta}_2) > 1$, then $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$
- If $RE(\hat{\theta}_1, \hat{\theta}_2) < 1$, then $\hat{\theta}_1$ is less efficient than $\hat{\theta}_2$
- If $RE(\hat{\theta}_1, \hat{\theta}_2) = 1$, then $\hat{\theta}_1$ is as efficient as $\hat{\theta}_2$

Example 4.1.2

Let Y_1, Y_2, Y_3, Y_4 be a random sample from the distribution of a random variable Y with mean μ and variance 8. Three researchers would like to estimate the mean μ using three different estimators.

- Researcher 1: $\hat{\mu}_1 = \overline{Y}$
- Researcher 2: $\hat{\mu}_2 = 0.1Y_1 + 0.1Y_2 + 0.1Y_3 + 0.7Y_4$
- Researcher 3: $\hat{\mu}_3 = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$

a. Which of these estimators are unbiased?

b. Which of these estimators is the most efficient?

SOLUTION

a. To show that $\hat{\mu}_1$ is unbiased for μ, we must show that $E(\hat{\mu}_1) = \mu$. Now,

$$\begin{aligned} E[\hat{\mu}_1] &= E[\overline{Y}] \\ &= E\left[\frac{Y_1 + Y_2 + Y_3 + Y_4}{4}\right] \\ &= \frac{1}{4}[E(Y_1) + E(Y_2) + E(Y_3) + E(Y_4)] \\ &= \frac{1}{4}[\mu + \mu + \mu + \mu] \\ &= \frac{1}{4}[4\mu] \\ &= \mu \end{aligned}$$

Thus, $\hat{\mu}_1$ is unbiased for μ.

In a similar manner it can be shown that $\hat{\mu}_2$ and $\hat{\mu}_3$ are unbiased for μ. [VERIFY!]

b. The variances of the three estimators are:

$$\begin{aligned} Var(\hat{\mu}_1) &= Var\left[\frac{Y_1 + Y_2 + Y_3 + Y_4}{4}\right] \\ &= \left(\frac{1}{4}\right)^2 [Var(Y_1) + Var(Y_2) + Var(Y_3) + Var(Y_4)] \\ &= \frac{1}{16}[8 + 8 + 8 + 8] \\ &= 2 \end{aligned}$$

Notice that, in general, $Var(\overline{Y}) = \frac{\sigma^2}{n}$. Hence, the variance of $\hat{\mu}_1 = \overline{Y}$ is equal to $\frac{8}{4} = 2$.

The variances of the other estimators are: [VERIFY!]

$$\begin{aligned} Var(\hat{\mu}_2) &= 4.16 \\ Var(\hat{\mu}_3) &= 2.54 \end{aligned}$$

Hence, $Var(\hat{\mu}_1) < Var(\hat{\mu}_3) < Var(\hat{\mu}_2)$. The relative efficiency is given below:

$$RE(\hat{\mu}_1, \hat{\mu}_2) = \frac{Var(\hat{\mu}_2)}{Var(\hat{\mu}_1)} = \frac{4.16}{2} = 2.08 > 1 \implies \hat{\mu}_1 \text{ is more efficient than } \hat{\mu}_2$$

$$RE(\hat{\mu}_1, \hat{\mu}_3) = \frac{Var(\hat{\mu}_3)}{Var(\hat{\mu}_1)} = \frac{1.54}{2} = 0.77 < 1 \implies \hat{\mu}_1 \text{ is less efficient than } \hat{\mu}_3$$

$$RE(\hat{\mu}_2, \hat{\mu}_3) = \frac{Var(\hat{\mu}_3)}{Var(\hat{\mu}_2)} = \frac{1.54}{4.16} \approx 0.37 < 1 \implies \hat{\mu}_2 \text{ is less efficient than } \hat{\mu}_3$$

Some Common Unbiased Point Estimators

Target Parameter θ	Sample Size(s)	Point Estimator $\hat{\theta}$	$E(\hat{\theta})$	Standard Error $\sigma_{\hat{\theta}}$
μ	n	$\overline{Y}$	μ	$\frac{\sigma}{\sqrt{n}}$
p	n	$\hat{p} = \frac{Y}{n}$	p	$\sqrt{\frac{pq}{n}}$
$\mu_1 - \mu_2$	n_1 and n_2	$\overline{Y}_1 - \overline{Y}_2$	$\mu_1 - \mu_2$	$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}^{*\dagger}$
$p_1 - p_2$	n_1 and n_2	$\hat{p}_1 - \hat{p}_2$	$p_1 - p_2$	$\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}^{\dagger}$

*σ_1^2 and σ_2^2 are the variances of populations 1 and 2, respectively.

†The two samples are assumed to be independent.

The unbiased estimator for the population variance σ^2 is the sample variance

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$$

Example 4.1.3:

An auditor randomly samples 20 accounts receivable from among the 500 such accounts of a client's firm. The auditor lists the amount of each account and checks to see if the underlying documents comply with stated procedures. The data are recorded in the accompanying table (amounts are in dollars, Y = yes, and N = no).

Account	Amount	Compliance	Account	Amount	Compliance
1	278	Y	11	188	N
2	192	Y	12	212	N
3	310	Y	13	92	Y
4	94	N	14	56	Y
5	86	Y	15	142	Y
6	335	Y	16	37	Y
7	310	N	17	186	N
8	290	Y	18	221	Y
9	221	Y	19	219	N
10	168	Y	20	305	Y

a. Estimate the true mean amount per account.

b. Calculate the standard error of the estimate in (a).

c. Estimate the true proportion of compliant accounts.

d. Compute the standard error of the estimate in (c).

SOLUTION

a. Based on the discussion, the sample mean $\overline{Y}$ is the unbiased point estimator of the population mean μ.

$$\begin{aligned}\overline{Y} &= \frac{\sum_{i=1}^{20} Y_i}{20} \\ &= \frac{3942}{20} \\ &= 197.1\end{aligned}$$

b. The estimated standard error of the sample mean is

$$\begin{aligned}SE(\overline{Y}) &= \frac{s}{\sqrt{n}} \\ &= \frac{90.85726}{\sqrt{20}} \\ &= 20.2163\end{aligned}$$

c. The sample proportion $\hat{p}$ is the unbiased point estimator of the true proportion p.

$$\begin{aligned}\hat{p} &= \frac{\text{total number of compliant accounts}}{20} \\ &= \frac{14}{20} \\ &= 0.7\end{aligned}$$

d. The estimated standard error of $\hat{p}$ is

$$\begin{aligned}SE(\hat{p}) &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ &= \sqrt{\frac{0.70(1-0.70)}{20}} \\ &= 0.1024695\end{aligned}$$

Example 4.1.4

A study was conducted to compare the mean number of police emergency calls per 8-hour shift in two districts of a large city. Samples of 100 8-hour shifts were randomly selected from the police records for each of the two regions, and the number of emergency calls was recorded for each shift. The sample statistics are given in the following table.

	Region	
	1	2
Sample size	100	100
Sample mean	2.4	3.1
Sample variance	1.44	2.64

a. Estimate the difference between the true mean number of police emergency calls in the two districts.

b. Calculate the estimated standard error of the estimate in (a).

SOLUTION: Left as a classroom exercise!

Example 4.1.5

In a survey among college students, 300 of 500 fraternity men favored a certain proposition whereas 64 of 100 non-fraternity men favored it.

a. Estimate the difference in the true proportions favoring the proposition.

b. Calculate the estimated standard error of the estimate in (a).

SOLUTION: Left as a classroom exercise!

Confidence Intervals

Recall that point estimates, such as the sample proportion ($\hat{p}$), the sample mean ($\overline{Y}$), and the sample standard deviation (s) depend on the particular sample selected.

When we use the sample mean ($\overline{Y}$) to estimate the population mean μ, can we be confident that ($\overline{Y}$) is close to μ? Do we have any idea as to how close the sample statistic is to the population parameter?

Rather than using just a point estimate, we could find an interval (or range) of values that we can be really confident contains the actual unknown population parameter. For example, we could find lower (L) and upper (U) values between which we can be really confident the population mean falls:

$$L < \mu < U$$

An interval of such values is called a confidence interval. Each interval has a confidence coefficient $(1 - \alpha)$ or a confidence level $((1 - \alpha) \times 100\%)$.

Typical confidence coefficients are 0.90, 0.95, and 0.99, with corresponding confidence levels 90%, 95%, and 99%. For example, upon calculating a confidence interval for a mean with a confidence level of, say 95%, we can say:

"We can be 95% confident that the population mean falls between L and U"

The general format of a confidence interval is

$$estimate \ \pm \ c \times SE(estimate)$$

where the multiplier c depends on the confidence level and the sampling distribution of the estimate, and $SE(estimate)$ is the standard error of the estimate and is simply equal to $\sqrt{Var(estimate)}$.

The expression $c \times SE(estimate)$ is commonly referred to as the Margin of Error (ME).

Confidence Interval for the Population Proportion

If $np > 5$ and $n(1-p) > 5$, then the sample proportion $\hat{p}$ has an approximate normal distribution with mean equal to the true proportion p and variance $\frac{p(1-p)}{n}$. Under these conditions (large sample), a $(1-\alpha)\%$ confidence interval for p is given by

$$\hat{p} \pm z_{\alpha/2} \times SE(\hat{p})$$

where the multiplier $z_{\alpha/2}$ is the $\alpha/2$ upper-quantile of the standard normal distribution which can be directly obtained from the Z table or using the $qnorm()$ function in R. The quantity $SE(\hat{p})$ is equal to $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$. Since p is unknown, we use the estimated standard error given by

$$\hat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example 4.1.6

A random sample of 1500 Filipinos were asked whether they approve or disapprove of the plan of the Philippine government to re-apply for membership of the International Criminal Court (ICC). Of the 1500 surveyed, 660 respond with "approve". Calculate a 95% confidence interval for the overall proportion of Filipinos who approve the plan.

SOLUTION

Here we have $\hat{p} = \frac{660}{1500} = 0.44$. Also, $\alpha = 0.05$, hence, $z_{\alpha/2} = 1.96$. But before we construct the confidence interval let us check the conditions first:

$$n\hat{p} = 1500 \times 0.44 = 660 > 5$$
$$n(1-\hat{p}) = 1500 \times (1 - 0.44) = 840 > 5$$

So the two conditions are met and we can now proceed to the construction of the confidence interval.

$$\begin{aligned}
\hat{p} \pm z_{\alpha/2} SE(\hat{p}) &\implies 0.44 \pm 1.96 \times \sqrt{\frac{0.44(1-0.44)}{1500}} \\
&\implies (0.44 \pm 1.96 \times 0.0128) \\
&\implies (0.44 - 0.0251,\ 0.44 + 0.0251) \\
&\implies (0.4149,\ 0.4651)
\end{aligned}$$

Therefore, we are 95% confident that the true proportion of Filipinos who approve the plan to re-apply for ICC membership is between 0.4149 and 0.4651.

Confidence Interval for the Population Mean

Case 1: σ is known

Suppose we have a random sample of size n from a normal distribution with mean μ and variance σ^2. It can be shown that the sample mean $\overline{y}$ has a normal distribution with mean μ and variance σ^2/n.

A $(1-\alpha)\%$ confidence interval for μ is given by

$$\overline{y} \pm z_{\alpha/2} \times SE(\overline{y})$$

where $SE(\overline{y}) = \frac{\sigma}{\sqrt{n}}$.

Note that the only condition for the above confidence interval is that the population distribution is normal.

Case 2: σ is unknown

Now, if we have a random sample of size $n > 30$ from a certain distribution (not necessarily normal) with mean μ and variance σ^2. Then by the Central Limit Theorem, the sample mean $\overline{y}$ has an approximate normal distribution with mean μ and variance σ^2/n.

In many practical cases, the population variance σ^2 is not known. We estimate σ using the sample standard deviation s. Then a $(1-\alpha)\%$ confidence interval for μ is given by

$$\overline{y} \pm z_{\alpha/2} \times SE(\overline{y})$$

where $SE(\overline{y}) = \frac{s}{\sqrt{n}}$.

Meanwhile, if the sample size $n < 30$, then a $(1-\alpha)\%$ confidence interval for μ is given by

$$\overline{y} \pm t_{\alpha/2,\, n-1} \times SE(\overline{y})$$

where $t_{\alpha/2,\, n-1}$ is the $\alpha/2$ upper quantile of the t distribution with $n-1$ degrees of freedom and $SE(\overline{y}) = \frac{s}{\sqrt{n}}$. The quantity $t_{\alpha/2,\, n-1}$ can be obtained from the T table or using the R function $qt()$.

Example 4.1.7:

Suppose we are interested in the average waiting time at a local bank. You take a random sample of 50 bank clients who visit the bank over the past week. From this sample, the mean waiting time was 30 minutes and the standard deviation was 20 minutes. Find a 95% confidence interval for the average waiting time for the bank.

SOLUTION

Is the population data normal? We don't know. However, the sample size is 50 which exceeds our minimum requirement of 30.

So a 95% confidence interval for μ is given by

$$\begin{aligned} \overline{y} \pm z_{\alpha/2} \times SE(\overline{y}) &\implies 30 \pm 1.96 \times \frac{20}{\sqrt{50}} \\ &\implies (30 \pm 5.54) \\ &\implies (30 - 5.54,\ 30 + 5.54) \\ &\implies (24.46,\ 35.54) \end{aligned}$$

Therefore, we are 95% confident that the true mean waiting time at the local bank is from 24.46 minutes to 35.54 minutes.

Example 4.1.8

Suppose in Example 4.1.7, $n = 25$, then a 95% confidence interval for μ is given by

$$\begin{aligned} \overline{y} \pm t_{\alpha/2,\, n-1} \times SE(\overline{y}) &\implies 30 \pm 2.064 \times \frac{20}{\sqrt{25}} \\ &\implies (30 \pm 8.256) \\ &\implies (30 - 8.256,\ 30 + 8.256) \\ &\implies (21.744,\ 38.256) \end{aligned}$$

Note that $t_{\alpha/2,\, n-1} = t_{0.025,\, 24} = 2.064$.

Remarks:

1. Notice that the 95% confidence interval for the mean μ is wider for a small sample (Example 4.1.8) than a large sample (Example 4.1.7).
2. Confidence intervals for the difference between two means $\mu_1 - \mu_2$and between two proportions $p_1 - p_2$ are constructed in a similar manner as in Examples 4.1.6 thru 4.1.8.

Example 4.1.9

Consider the data in Example 4.1.4. A 90% confidence interval for the difference in the mean number of emergency calls between the two regions is constructed as follows:

$$\begin{aligned}(\overline{y}_1 - \overline{y}_2) \pm z_{\alpha/2} \times SE(\overline{y}_1 - \overline{y}_2) &\implies (2.4 - 3.1) \pm 1.645 \times \sqrt{\frac{1.44}{100} + \frac{2.64}{100}} \\ &\implies -0.7 \pm 0.33 \\ &\implies (-1.03,\ -0.37)\end{aligned}$$

Therefore, we are 90% confident that the true difference between the mean number of emergency call for the two regions is between -1.03 and -0.37.