Topic 6. Confidence Intervals for Population Means

Cheng Peng

Contents

1 Introduction

We will discuss the general framework of confidence intervals of population means and proportions. Before start the discussion of the new topic, we review fundamental tools for the confidence interval in the following video: sampling distributions of sample means and proportions. [**44 minutes**] .

In the next few sections, we learn how to **estimate the population parameters** such as mean, standard deviation, and proportion from a random sample and build a confidence interval (also called interval estimate) to show how good the estimate is, and finally, we should be able to interpret the estimate.

2 A General Framework

This section dedicates the basic framework to estimating population means using confidence intervals.

2.1 Some Technical Terms

The following is a list of several basic terms to use when discussing the estimate of population parameters.

- An **estimate** is *a specific value* or *range of values* obtained from a random sample that is used to approximate a population parameter.

- A **point estimate** is a single value (obtained from the random sample) that is used to approximate a population parameter.

Example 1. The sample mean, denoted by $\bar{x}$, is the best point estimate of the population mean μ.

Example 2. The sample variance, denoted by s^2 is the best point estimate of the population variance σ^2.

- The **bias of the point estimate** is equal to the difference between the estimate and the true parameter. For example, the bias of the sample mean is $\bar{x} - \mu$. The bias of an estimate measures *the accuracy of the estimate*.

- Since the point estimate of a parameter is obtained from an underlying random sample, it is a random variable. The variance of the point estimate measures the precision of the point estimate of the corresponding population parameter.

- **The goodness of an estimate** – *no bias and small variance.*

2.2   Issues of A Point Estimate and the Logic for Developing A Solution

For a point estimate, we can only say that it is close to the true population parameter. Three questions remain to answer:

1). How close is "close"?

2). How confident is your statement?

3). How precise is the estimate?

We use the following artificial example to explain the above issues about a **point estimate** of the population mean.

Example 3. Assume that the true mean height of WCU student population is 69 inches. To estimate the true mean, we randomly select a sample of 50 students and measure their heights to get a random sample 50 heights. Assume that the sample average of 68.5 inches.

Now the question: **Is 68.5 close to 69?**

Apparently, we cannot provide an objective answer to the above question with a single sample mean without any additional information.

To address the issue, let's pretend sampling is free. We could take as many samples of the same size using the same protocol as we wish. Then we get many sample means. For example, pretending we obtained 100 sample means from 100 random samples taken from the WCU student population with the same size of 50.

```
68.2 68.3 68.3 68.4 68.4 68.5 68.6 68.6 68.6 68.6 68.6 68.7 68.7 68.7 68.7 68.7
68.7 68.7 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8 68.8
68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 68.9 69.0 69.0 69.0 69.0
69.0 69.0 69.0 69.0 69.0 69.0 69.0 69.1 69.1 69.1 69.1 69.1 69.1 69.1 69.1 69.1
69.1 69.1 69.1 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.2 69.3 69.3 69.3
69.3 69.3 69.3 69.4 69.4 69.4 69.5 69.5 69.5 69.5 69.5 69.5 69.6 69.6 69.6 69.6
69.7 69.7 69.7 69.9
```

With these sample means, we make a histogram in the following
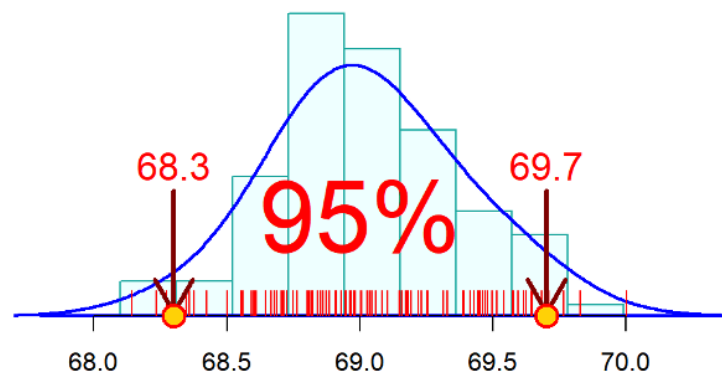
We use the definition of percentile to find 2.5% and 97.5% percentiles to be 68.3, and 69.7 respectively (shown in the above figure). This means that 95% of the sample means are within interval: [68.3, 69.7]. Clearly, 69 is inside the interval.

With the above interval, address the aforementioned three issues:

- **All values inside the interval are considered to be close to the true mean of 69.**

- The interval contains 95% of the sample means. If a sample mean is in the interval, we are 95% confident that the sample mean is close to the true mean of 69.

- The width of the interval $(69.7 - 68.3 = 1.4)$ reflects the precision of the estimate.

Therefore, this interval contains all desired information, but what is more important is that **the interval was constructed from the distribution of sample means**.

Definition 1: The area of $0.95 = 95\%$ of the middle region in the above figure is called ** 95% Confidence Level**.

Definition 2: *The above interval is called the 95% confidence interval of the average height of WCU student population.* Furthermore, **68.3** is called lower confidence limit (LCL) and **69.7** is called upper confidence limit (UCL).

The above interval was found based on the histogram of 100 sample means. In real-world applications, taking multiple samples could be very costly. **The new and practical question is** *how to find an interval similar to the above one without taking multiple samples.*

2.3   General Framework of Confidence Intervals

The key information needed to find the confidence interval of a population mean (μ) is to know the distribution of the sample mean $(\bar{X})$. The distribution of $\bar{X}$ is also called the sampling distribution of the sample mean.

We have discussed the sampling distribution of sample $\bar{X}$ in the previous note.

- If the sample size $n > 30$, by the CLT, the distribution of the sample mean $(\bar{X})$ is *approximately* normal with the mean and standard deviation specified below

$$\bar{X} \to N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- If the population is normally distributed as $N(\mu, \sigma)$, the

$$\bar{X} \to N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

regardless of the sample size.

- For a binary population with success probability p, the sampling distribution of sample proportion $(\hat{p})$ is

$$\hat{p} \to N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

Thought Process of Obtaining A Confidence Interval of Population Mean

With the above sampling distribution, we can find the confidence interval without taking multiple samples. Next, we use the sampling distribution based on the CLT to illustrate the logic for obtaining the confidence interval of the population mean (μ).

- when $n > 30$,

$$\bar{X} \to N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$
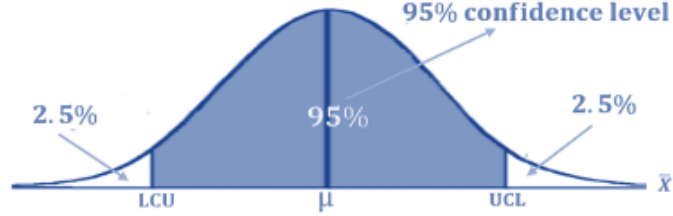
For the given `95% confidence level` (the area of the middle region in the above figure), we cannot find the `confidence limits LCL and UCL` directly from the normal table directly since $\bar{X}$.

- However, we can use z-score transformation to transform the sampling distribution of $\bar{X}$ to the standard normal distribution in the following.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \to N(0, 1)$$

When the area of 0.95 of the middle symmetric region in the above standard normal curve is given, we can find the 2.5% and 97.5% percentiles (-1.96 and 1.96 respectively) from the normal table directly.

Definition 3: The percentiles ± 1.96 are called **critical values (CV)** of 95% confidence level for the standard normal distribution. The two critical values are symmetric with respect to the origin!

- Relationship between Critical Values and Confidence Limits

The critical values corresponding to the 95% confidence level are ± 1.96. Using the z-score transformation, we have the following relationship.

$$-1.96 = \frac{LCL - \mu}{\sigma/\sqrt{n}}$$

and

$$1.96 = \frac{UCL - \mu}{\sigma/\sqrt{n}}$$

We can solve the confidence limits

$$LCL = \mu - 1.96 \times \frac{\sigma}{\sqrt{n}}, \qquad UCL = \mu + 1.96 \times \frac{\sigma}{\sqrt{n}}$$

- If we replace μ and σ with sample mean $\bar{X}$ and sample standard deviation s, then LCL and UCL will be completely dependent on the sample data. Then 95% confidence interval can be written as

$$(LCL, UCL) = (\bar{X} - 1.96 \times \frac{s}{\sqrt{n}}, \quad \bar{X} + 1.96 \times \frac{s}{\sqrt{n}})$$

- Since the confidence interval is constructed based on the random sample, it is random. We can interpret the confidence interval: **there is a 95% chance that the confidence interval [68.3, 69.7] contains the true population mean (μ).**

The above thought process is summarized in the following chart
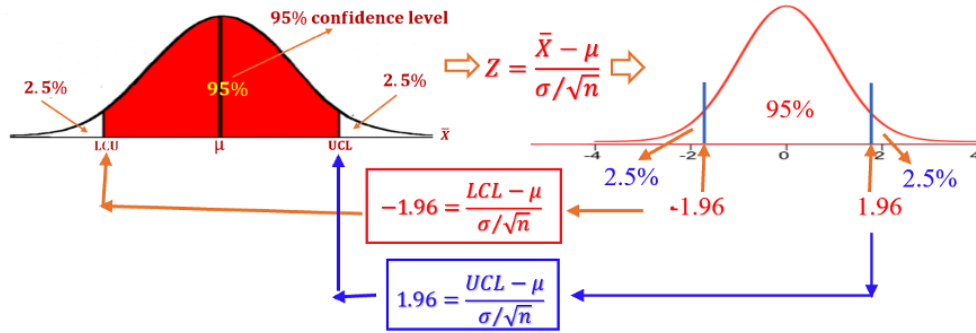
Example 4. Suppose we want to estimate, with 95% confidence level, the mean (average) length of all walleye fingerlings in a fish hatchery pond. A random sample of 100 fingerlings was selected. The average length is 7.5 inches and standard deviation is 2.3 inches. That is, $\bar{x} = 7.5$, $s = 2.3$, and $n = 100$.

Solution: Since $n = 100 > 30$, by the central limit theorem, the sample mean is approximately distributed. The 95% critical values are ± 1.96 (using normal table). Based on the above discussion, the lower and upper confidence limits are

$$(LCL, UCL) = (7.5 - 1.96 \times 2.3/\sqrt{100}, \quad 7.5 + 1.96 \times 2.3/\sqrt{100}) = (7.05, \ 7.95)$$

That is, interval $(7.05, \ 7.95)$ has a 95% chance of containing the true population mean length of all walleye fingerlings in a fish hatchery pond.

3 Formal Steps For Constructing C.I.

We now formulate the steps for constructing a CI based on the above framework and introduce a few new concepts.

Step 1: Identify the confidence level. If it is NOT given, our default confidence level $1 - \alpha = 0.95$ should be used in this class. **Note:**

Step 2: Based on the confidence level to find the critical value from the normal table.

Step 3: Evaluate the margin of error, denoted by E and defined to be

$$E = CV \times \frac{s}{\sqrt{n}}$$

5

Step 4: Write out the confidence interval explicitly in the following form

$$\bar{x} \pm E = (\bar{x} - E,\ \bar{x} + E)$$

Step 5: Interpretation of the confidence interval: Two versions of interpretations

`Version #1: We are 95% confident that the interval contains the population mean`

`Version #2: There is a 95% chance that the interval contains the population means.`

Some Remarks: The following important general facts of confidence intervals can be visualized using **IntroStatsApps: Normal Confidence Interval** at: https://wcu-peng.shinyapps.io/NormalCI4MeanProp/

- By convention, we use $1 - \alpha$ to denote the confidence level (i.e., the area of the middle symmetric region on the density curve). This means that α is the sum of the two tail areas. In other words, both left and right tail areas are equal to $\alpha/2$.
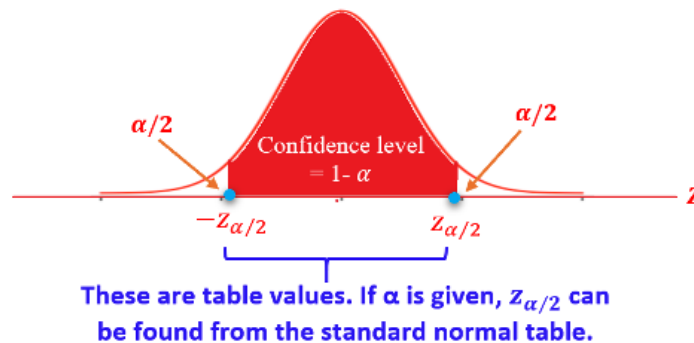
- The margin of error, E, is equal to half of the width of the interval.

- Since sample size, n, is in the denominator of E, as sample size increases, E decreases. This implies that as the sample size increases, the interval gets narrower (See the following formula).

- As the level of confidence increases, the critical value (CV) increases. This implies that the width of the confidence interval increases as the confidence level increases (see the above formula for this relationship).

Example 5. Assume that we collect 81 measurements of the iron-solution index of tin-plate specimens, designed to measure the corrosion resistance of tin-plated steel. Assume that the sample mean is 57 and the sample standard deviation is 15. Construct a 95% confidence interval of the iron-solution index.

Solution We will follow the 5-step procedure to construct the confidence interval.

Step 1: The confidence level $1 - \alpha = 0.95$, this means that $\alpha/2 = 0.025$.

Step 2: The critical values are $CV = Z_{0.025} = 1.96$.

Step 3: The margin of error
$$E = CV \times \frac{s}{\sqrt{n}} = 1.96 \times \frac{15}{\sqrt{81}} = 3.27.$$

Step 4: The explicit expression of the 95% CI is given by
$$57 \pm E = (57 - 3.27, \ 57 + 3.27) = (53.73, \ 60.27).$$

Step 5: The interpretation of CI. We are 95% confident that the interval $(53.73, \ 60.27)$ contains the true population mean of the iron-solution index.

The following video outlines the above discussion with a example [**44 minutes**] .

4 Use of Technology

I also created an interactive application to generate confidence intervals for population mean and proportion based on the assumption of large samples. The app is at: https://wcu-peng.shinyapps.io/NormalCI4MeanProp/

The following is the screenshot of the solution of Example 5 generated by the app.

IntroStatsApps: Normal C.I.s for Population Mean (μ) and Proportion (p)

1. CI for μ or p?

◉ Population Mean (μ)
○ Population Proportion (p)

2. Sample Statistics

Sample Mean: $\bar{X}$

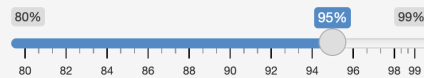| 57 |
| --- |

Standard Deviation: s or σ_0

15

3. Sample Size and Confidence Level

Sample Size: n

81

Confidence Level: α

80%   95%   99%

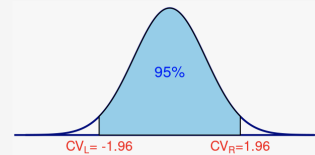80   82   84   86   88   90   92   94   96   98 99

Report bugs to C. Peng

Steps for Constructing A C.I.

Step 1. The given confidence level.
conf.level $= 1 - \alpha = 95\%$

Step 2. CV on the standard normal density curve.
$CV = Z_{\alpha/2} = 1.960$

95%

$CV_L = -1.96$ $CV_R = 1.96$

Step 3: Margin of Error

$$E = CV \times \frac{s}{\sqrt{n}} = 1.96 \times \frac{15}{\sqrt{81}} = 3.267$$

Step 4: Expression of Confidence Interval

$$(\bar{X} - E, \bar{X} + E) = (57 - 3.267, 57 + 3.267) = (53.733, 60.267)$$

Step 5: Interpretation of Confidence Interval

There is a 95 % chance that the confidence interval (53.733 , 60.267) contains the true population mean.