Probability and Mathematical Statistics in Data Science

Lecture 25: Section 9.3: Confidence Intervals: Method

Confidence Intervals

- A point estimate, because it is a single number, by itself provides no information about the precision and reliability of estimation

- An alternative to reporting a single sensible value for the parameter being estimated is to calculate and report an entire interval of plausible values – an *interval estimate* or *confidence interval* (*CI*).

- A confidence interval is always calculated by first selecting a *confidence level*, which is a measure of the degree of reliability of the interval.

Using $\bar{X}$ to estimate μ

- $\bar{X}$ is an unbiased estimator of μ
- If we also know that each of the X_k had SD σ, what can we say about $SD(\bar{X})$?

- What does the Central Limit theorem say about the sample mean?

- We will use the CLT and the sample mean to define a interval that will *cover* the true mean with a specified probability, say 95%

Illustration (σ *is Known*)

- Let's first consider a simple, somewhat unrealistic problem situation.
  1. We are interested in the population mean parameter μ.
  2. The population distribution is normal.
  3. The value of the population standard deviation σ is known. (**unlikely!**)

- Suppose we have a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with mean value μ and standard deviation σ. As we know, $\bar{X}$ also follows a normal distribution with mean value μ and standard error $\sigma/\sqrt{n}$. Thus, we could get a standard normal distribution by normalizing $\bar{X}$

.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Construction

- The smallest interval that contains 95% of the possible outcomes of Z is (-1.96, 1.96).

$$-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96$$

$$\Downarrow$$

$$-1.96 \cdot \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

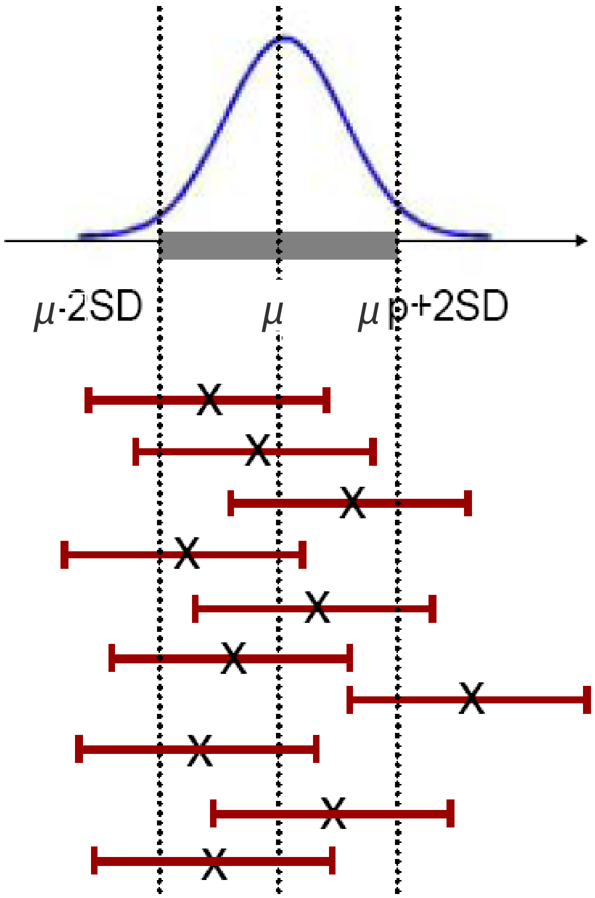
$$\Downarrow$$

$$\bar{X} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

Random Interval

- By constructing a confidence interval like this, we never be sure whether μ actually lies in our confidence interval.

  However, we know that about 95 out of 100 times intervals constructed using this method will capture the true parameter.



- Interpreted as: "*the probability is .95 that the random interval includes or covers the true value of μ.*"

Confidence Interval for the Population Mean

The 95% confidence interval for the population mean is calculated as follows:

$$\boxed{\bar{x} \pm z\,(\frac{\sigma}{\sqrt{n}})}$$

sample mean ± **2** x (standard error)

where the standard deviation of the mean is:

$$\boxed{se(\bar{x}) = \frac{\sigma}{\sqrt{n}}}$$

Example

- A population distribution is known to have an SD of 20. The average of an sample of 64 observations is 55. What is your 95% confidence interval for the population mean?

Kaggle Dataset – Heart Disease UCI

- This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them.

- Attribute Information:
  > 1. age
  > 2. sex
  > 3. chest pain type (4 values)
  > 4. resting blood pressure
  > 5. serum cholestoral in mg/dl
  > 6. fasting blood sugar > 120 mg/dl

source: https://www.kaggle.com/ronitf/heart-disease-uci

Kaggle Dataset – Heart Disease Dataset

- The dataset consists of 303 patients from the Cleveland Clinic, a non-profit academic medical center.

- We will analyze the measurements taken for Cholesterol level an blood pressure.

Q. What are the necessary assumptions and conditions that should be checked before constructing a confidence intervals for the population mean in this case?

Kaggle Dataset – Heart Disease Dataset

- **Random Sample**: Heart Disease patients from Cleveland Clinic. Not enough information given to decide whether the data was based on a random or an convenience sample

- **Independence**: The sample may not be random but (in this case) it is fair to assume the measurement values (cholesterol levels) are independent. **Why?**

- **Sample Size:** The sample size of 303 is greater than 40, so even if the population distribution of individual cholesterol levels is not normal, our analysis will still be valid.

Kaggle Dataset – Heart Disease UCI

Cholesterol Levels

Sample Size = 303 Sample Mean = 246.3

Sample Standard Deviation = 51.8

95% Confidence Interval: sample mean ± **2** x (standard error)

246.3 ± 2 * (51.8 / square root of 303)

= 246.3 ± 5.95

= [240.35, 252.25]

Q. What does the confidence interval mean?

Kaggle Dataset – Heart Disease UCI

Blood Pressure

Sample Size = 303 Sample Mean = 131.6 Sample Standard Deviation = 17.5

95% Confidence Interval: 131.6 ± 2 * (17.5 / square root of 303)

= 131.6 ± 1.0

= [130.6, 132.6]

Q. What does the 95% confidence interval mean?

Confidence levels

- The probability with which our *random* interval will cover the mean is called the confidence level.

- In reality (vs theory), we will have just one *realization* (observed value) of the sample mean (from our data sample), and we use that value to write down the realization of our random interval.

- What would we do differently if we wanted a 68% CI? 99.7% CI?

- What about an 90% CI? 99% CI?

Understanding the Confidence Level

- For a confidence level of 95%, ***we expect that about 95% of all such intervals will actually cover the true population value***.
- The remaining 5% will not. Confidence is in the *procedure* over the long run.

- 90% confidence level => multiplier = 1.645
- 95% confidence level => multiplier = 1.96
- 99% confidence level => multiplier = 2.576
- More confidence ⇔ Wider Interval (for the standard error)

Confidence Intervals for Population Proportions

If our sample data adheres to the assumptions and conditions for valid analysis, a confidence interval for the population proportion, p, can be calculated as follows:

$$\boxed{\hat{p} \pm z(se(\hat{p}))}$$

sample proportion ± **2** x (standard error)

where the standard error is:

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

Example

- We have a population of 1 million in a town. We take a SRS of size 400 and find that 22% of the sample is unemployed. Estimate the percentage of unemployed people in the town.