Introduction to Confidence Intervals in Statistics using a Probability Perspective

Yunhao Cao, Yufan Mao

April 2022

Abstract

This document is a mini-project for the STAT 134 class and will give a brief guide to what is the concept of confidence interval and how to use it as a tool for statistical analysis of discrete data.

Contents

1 Introduction To Confidence Interval Concepts **1**
1.1 Summary of Confidence Interval Analysis 2

2 Demonstration of Confidence Interval **3**

3 Reference **3**

4 Contribution **3**

1 Introduction To Confidence Interval Concepts

Let us consider the following scenario:

We have sampled lots of real-world Covid-19 data from our healthcare system, that includes information about people that got infected by Covid-19 (with sensitive data removed), and the features of the data is age, and now we want to estimate the *real, underlying mean* of the people's age that got infected, how should we achieve our goal?

We will now define some notations. First of all, we will denote the feature vector of collected samples as

$$\vec{x} = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}^\top$$

, where x_i denotes the age of the i-th sample.

Now let's estimate the *real* mean of our data, **assuming that the data comes from a normal distribution**. Note that the real underlying distribution might actually be a binomial distribution, or exponential distribution, but we have to *make an assumption for us to start guessing the parameters of our guessed distribution.*

Side Note: We also chose normal distributions for a reason, we're estimating the underlying mean of our observed sample points, and the **Central Limit Theorem** states that the mean of any observed distribution, as the number of samples goes to infinity, will converge into a standard normal distribution. With introduction of CLT, the concept of confidence interval can also be proved using normal distributions.

For us to complete the estimation, we need to compute the mean, μ_x, and the variance, σ_x^2 of our individual data points.

From what we learned in **Law of Large Numbers**, we know that as we sample i.i.d. variables, sampling enough data would make the mean of those discrete observed values reach the actual mean (or expectation) of the underlying distribution, expressed as

$$\hat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$P(|\hat{\mu}_x - \mu_x| > \epsilon) \to 1$ as $n \to \infty$, no matter how small ϵ is

So therefore we can compute our *estimated* mean of the normal distribution as

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

However, one problem that emerged in the process is that we simply cannot gather infinite amount of data, so then there is a possibility that our estimated parameter is wrong(actually the chance that the underlying mean equals to exactly our estimated mean is infinitesimally small). That's why we want to use the concept of **confidence interval** to make sure that our estimated range of mean has a big enough probability for it to be accurate.

The Chebychev's Inequality is useful in estimating the confidence interval of a parameter guess, the inequality is written out as

$$P(|X - \mu_x| \geq c) \leq \frac{\sigma_x^2}{c^2}$$

Therefore we see that we can specify a probability γ that our estimated range of parameter $\mu_x \in (\hat{x} - \epsilon, \hat{x} + \epsilon)$ contains the actual parameter using this inequality:

$$P(\mu_x \in (\hat{x} - \epsilon, \hat{x} + \epsilon)) = P(|\hat{x} - \mu_x| \geq \epsilon) \leq \frac{\sigma_{\hat{x}}^2}{\epsilon^2} = \gamma$$

Ad therefore we obtain that in order for us to be at least $(100\gamma)\%$ sure that the true parameter μ_x lays in the range of $(\hat{x} - \epsilon, \hat{x} + \epsilon)$, we have to solve ϵ through the following equation:

$$\epsilon^2 = \frac{\sigma_{\hat{x}}^2}{\gamma}, \epsilon = \sqrt{\frac{\sigma_{\hat{x}}^2}{\gamma}}$$

However, the problem is that we have zero knowledge about the standard deviation of the distribution, and therefore we must estimate it.

$$\hat{\sigma}_{\hat{x}}^2 = \frac{1}{n}\hat{\sigma}_x^2 = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \hat{x})^2$$

Since the standard deviation is estimated, we therefore cannot say that the **probability** of the underlying distribution parameter is laying in our confidence interval is γ, instead we only say that we are $(100\gamma)\%$ **confident** that the actual value is lying inside our interval.

So therefore,

$$\epsilon = \sqrt{\frac{\sigma_{\hat{x}}^2}{\gamma}} \approx \sqrt{\frac{\hat{\sigma}_{\hat{x}}^2}{\gamma}} = \frac{\hat{\sigma}_x}{\sqrt{n\gamma}}$$

We see that the value of ϵ, is inversely correlated with the square root of n, and γ, therefore we can narrow the confidence interval by sampling more data if the required confidence value is unchanged.

1.1 Summary of Confidence Interval Analysis

Notation:

- γ - confidence value, $\in (0, 1)$
- $x_1, x_2, \ldots, x_n$ - collected samples
- $\hat{x}$ estimated mean (single numeric value)
- $(\hat{x} - \epsilon, \hat{x} + \epsilon)$ estimated confidence interval, part of the solution
- σ_x^2 - variance (standard deviation squared) of sample distribution
- $\hat{\sigma}_x^2$ - estimated variance of sample distribution
- $\sigma_{\hat{x}}^2$ - variance of mean of n discrete samples
- $\hat{\sigma}_{\hat{x}}^2$ - estimated variance of mean of n discrete samples

Question: We have collected data $x_1, x_2, \ldots, x_n$, we want to compute the underlying mean and standard deviation of our collected samples, with the range of mean having a confidence of γ, how should I calculate the estimated mean range and standard deviation?

Calculations:

$$\hat{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\sigma_{\hat{x}}^2 \approx \hat{\sigma}_{\hat{x}}^2 = \frac{1}{n}\hat{\sigma}_x^2 = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \hat{x})^2$$

$$\epsilon = \sqrt{\frac{\sigma_{\hat{x}}^2}{\gamma}} \approx \sqrt{\frac{\hat{\sigma}_{\hat{x}}^2}{\gamma}} = \frac{\hat{\sigma}_x}{\sqrt{n\gamma}}$$

Therefore, we are $(100\gamma)\%$ confident that the actual mean of the distribution for discrete samples $x_1, x_2, \ldots, x_n$ lays in the range $(\hat{x} - \epsilon, \hat{x} + \epsilon)$.

2 Demonstration of Confidence Interval

Suppose a quality inspection department needs to extract n bags of products from a factory (n is greater than 50), we want to observe the weight of each bag, and estimate the average weight of each bag of products in the factory (assuming a 95% confidence) . The recordings for the weights of each bag is

$$50, 55, 60, 64, 38, 40, 66, 39, 50, 52$$
$$66, 71, 29, 58, 47, 73, 17, 49, 20, 50$$
$$50, 18, 29, 50, 73, 98, 30, 49, 50, 18$$
$$55, 65, 42, 87, 29, 91, 56, 75, 32, 40$$

We record our observations as follows: $X_1, X_2, X_3 \ldots, x_n$. We can obtain the sample mean as $\bar{x} = 50.775$.

Since the population variance is unknown, we will follow the steps and estimate the sample variance as $\bar{\sigma}_{\bar{x}}^2 = \frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2 = 9.5406$, $\bar{\sigma}_{\bar{x}} = 3.0888$

Since confidence interval is 95%, $\gamma = 0.95$

Using the formula in section 1, $\epsilon = \frac{\bar{\sigma}_{\bar{x}}}{\sqrt{n\gamma}} = \frac{3.0888}{40\times 0.95} = 0.08128$

Therefore, we are 95% confident that the mean of the weights produced by the factory would be between 50.775 ± 0.08128.

3 Reference

1. Yunhao Cao's Notes for STAT 134
2. stat88.org's online textbook, section 9.2, 9.3, and 9.4
3. Jim Pitman, "Probability". 1993, Springer-Verlag New York.

4 Contribution

Section	Contributor
1	Yunhao for Writing & researching Proof from Chebyshevś Inequity Yufan for researching proof from CLT and Normal Dist.
2	Yufan Mao