HCMC University of Technology
Dung Nguyen

Probability and Statistics

Confidence Intervals

Population vs. Sample

- A population is a collection of objects, items, humans/animals about which information is sought.
- A sample is a part of the population that is observed.
- A parameter is a numerical characteristic of a population, e.g. Vietnamese unemployment rate.
- A statistic is a numerical function of the sampled data, used to estimate an unknown parameter,
  e.g., unemployment rate in a sample.

The Sample Mean

Definition (Population Mean)

The population mean, denoted by μ, is the average of all x values in the entire population.

Definition (Sample Mean)

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}$$

In this class we will work with both the population mean μ and the sample mean $\overline{x}$. Do not confuse them!

The Sample Median

- List the data values in order from smallest to largest
  - the median is the middle value in the list
  - it divides the list into two equal parts.
- the process of determining the median
  - When n is odd: the sample median is the single middle value.
  - When n is even: there are two middle values in the ordered list, and we average these two middle values to obtain the sample median.
- Mean and median can be very different. The median is more robust to outliers.

The Sample Variance and Sample Standard Deviation

Definition

- *The sample variance, denoted by s^2, is used to approximate the population variance σ^2*

$$s^2 = \frac{\sum (x_i - \overline{x})^2}{n-1} = \frac{S_{xx}}{n-1}$$

s is called the sample standard deviation.

- *If the population is relatively small then we use*

$$\hat{s}^2 = s^2 \cdot \frac{N-n}{N-1}$$

to approximate σ^2, and $\frac{N-n}{N-1}$ is called the finite population correction factor.

The Sample Proportion

Relative frequency estimate of p is k/n.

The estimated value of $p \in [0,1]$.

Example (1)

5023 Heads are observed on 10000 tosses. The relative frequency estimate of p is 0.5023

Is it possible that actually $p = 0.5$ instead?

Is it possible that actually $p = 0.51$?

Interval Estimates

An interval estimate estimates the value of p as being in an interval (a, b) or $[a, b]$

Example (2)

5023 Heads are observed on 10000 tosses.

An interval estimate is of the form

- $0.4973 < p < 0.5073$
- $0.5013 \leq p \leq 0.5033$

The length of the interval is a crucial parameter of the estimate.

Confidence Interval

How sure are we that the unknown value of p actually is in the interval specified?

- $[0,1]$: 100% confident.
- Smaller intervals: lesser degree of confidence.
- "$0.4973 < p < 0.5073$" vs. "$0.5013 \leq p \leq 0.5033$".

Confidence Interval and level

- $(X_1, \ldots, X_n)$ is a random sample from a distribution that depends on a parameter θ
- A confidence interval for θ:
$$S_1 \leq \theta \leq S_2,$$
where S_1 and S_2 are
  - computed from the sample data.
  - called the lower- and upper- confidence limits
- The confidence level:
$$\gamma = P_\theta(S_1 \leq \theta \leq S_2).$$
- Wide interval $\Longleftrightarrow$ high confidence level

Confidence level and Significance level

- A confidence level (γ) is a measure of the degree of reliability of the interval.
- A significance level (α) is the probability we allow ourselves to be wrong when we are estimating a parameter with a confidence interval.
$$\gamma + \alpha = 1$$

One-Sided Confidence Intervals

Definition (Left-Sided Confidence Intervals/Limits)

- *Let S_1 be a statistic: for all values of θ,*
$$P(S_1 < \theta) = \gamma$$
- *(S_1, ∞) is called*
  - *a one-sided coefficient γ CI for θ or*
  - *a one-sided 100γ percent CI for θ.*
- *S_1 is called*
  - *a coefficient γ lower confidence limit for θ or*
  - *a 100γ percent lower confidence limit for θ.*

One-Sided Confidence Intervals

Definition (Right-Sided Confidence Intervals/Limits)

- *Let S_2 be a statistic: for all values of θ,*
$$P(\theta < S_2) = \gamma$$
- *$(-\infty, S_2)$ is called*
  - *a one-sided coefficient γ CI for θ or*
  - *a one-sided 100γ percent CI for θ.*
- *S_2 is called*
  - *a coefficient γ lower confidence limit for θ or*
  - *a 100γ percent lower confidence limit for θ.*

Normal Population + Known σ

Theorem

If $X_1,\dots,X_n$ are iid $\sim N(\mu,\sigma^2)$, then

$$\frac{\sqrt{n}(\widehat{\mu}-\mu)}{\sigma} \sim N(0,1).$$

CI of population mean

If $X_1,\dots,X_n$ are iid $\sim N(\mu,\sigma^2)$ and $\alpha = 1-\gamma$, where γ is the confidence level, then the confidence interval of the population mean is

$$\mu = \widehat{\mu} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

Sample size

Let $\mathrm{MOE} = \frac{\sigma}{\sqrt{n}} \cdot z_{\alpha/2}$. Then $\mathrm{MOE} \le \epsilon \iff n \ge \left(\frac{\sigma \cdot z_{\alpha/2}}{\epsilon}\right)^2$.

Example 3 – Pit Stop

In auto racing, a pit stop is where a racing vehicle stops for new tires, fuel, repairs, and other mechanical adjustments. The efficiency of a pit crew that makes these adjustments can affect the outcome of a race. A random sample of 32 pit stop times has a sample mean of 12.9 seconds. Assume that the population distribution is normal and the population standard deviation is 0.19 second.

ⓐ Construct a 99% confidence interval for the mean pit stop time.

ⓑ How many observations must be collected to ensure that the radius of the 99% CI is at most 0.01?

Solution

$$12.9 \pm 2.58 \cdot \frac{0.19}{\sqrt{32}} = 12.9 \pm 0.087 \quad \text{and} \quad n \ge 2395.198.$$

One-Sided Confidence Interval (Normal Population + Known σ)

- A $100(1-\alpha)\%$ upper-confidence bound for μ is
$$\mu \le \widehat{\mu} + z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}.$$
- A $100(1-\alpha)\%$ lower-confidence bound for μ is
$$\mu \ge \widehat{\mu} - z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}.$$

Example 4 – Pit Stop

In auto racing, a pit stop is where a racing vehicle stops for new tires, fuel, repairs, and other mechanical adjustments. The efficiency of a pit crew that makes these adjustments can affect the outcome of a race. A random sample of 32 pit stop times has a sample mean of 12.9 seconds. Assume that the population distribution is normal and the population standard deviation is 0.19 second. **Construct an upper, one-sided 95% confidence interval for the population mean.**

Normal Population + Unknown σ

Theorem

If $X_1, \ldots, X_n$ are i.i.d. $\sim N(\mu, \sigma^2)$, then

$$\frac{\widehat{\mu} - \mu}{s/\sqrt{n}} \sim t_{n-1} \quad \text{and} \quad \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

CI of the population mean

If $X_1, \ldots, X_n$ are i.i.d. $\sim N(\mu, \sigma^2)$ then

$$\mu = \widehat{\mu} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

CI of the population variance

Choose c_1 and c_2 so that the area in each tail of χ^2_{n-1} distribution is $\alpha/2$. The γ-confidence interval for the unknown variance σ^2: $\frac{(n-1)s^2}{c_2} \le \sigma^2 \le \frac{(n-1)s^2}{c_1}$.

Example 5 – Tread Depth

11 randomly selected automobiles were stopped, and the tread depth of the right front tire was measured. The mean was 0.32 inch, and the standard deviation was 0.08 inch. Find the 95% confidence interval of the mean depth and its variance. Assume that the variable is approximately normally distributed.

Solution

$$\mu = 0.32 \pm 2.228 \cdot \frac{0.08}{\sqrt{11}} \implies \mu = 0.32 \pm 0.05.$$

Example 6 – Point of inflammation of Diesel oil

Five independent measurements of the point of inflammation of Diesel oil gave the values (in F)

144 147 146 144 142

Assuming normality, determine a 99% confidence interval for the mean.

Solution

Required values: $\widehat{\mu} = 144.6, s = 1.949$. *Thus*

$$\mu = 144.6 \pm 4.604 \cdot \frac{1.949}{\sqrt{5}} = 144.6 \pm 4.014$$

CI of the population variance

- Choose c_1 and c_2 so that the area in each tail of χ^2_{n-1} distribution is $\alpha/2$. Then the γ-confidence interval for the unknown variance σ^2 is
$$\frac{(n-1)s^2}{c_2} \le \sigma^2 \le \frac{(n-1)s^2}{c_1}$$
- Choose c_1 and c_2 so that the area in each tail of χ^2_{n-1} distribution is α. The γ lower and upper confidence bounds on σ^2 are
$$\sigma^2 \ge \frac{(n-1)s^2}{c_2}$$
and
$$\sigma^2 \le \frac{(n-1)s^2}{c_1}$$

Example 7 –

An automatic filling machine is used to fill bottles with liquid detergent. A random sample of 20 bottles results in a sample variance of fill volume of $s^2 = 0.01532$. Assume that the fill volume is approximately normal. Compute a 95% upper confidence bound.

Solution

$$\sigma^2 \leq \frac{(20-1)0.0153}{10.117} = 0.0287,$$

and

$$\sigma \leq 0.17.$$

Large Sample Size

Theorem

If $X_1, \ldots, X_n$ are i.i.d. then

$$\frac{\widehat{\mu} - \mu}{s/\sqrt{n}} \simeq N(0,1)$$

CI of population mean – Large sample size

If $X_1, \ldots, X_n$ are i.i.d. and n is large then

$$\mu \approx \widehat{\mu} \pm z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}.$$

Click for video

Example 8 –

A random sample of 110 lighting flashes in a region resulted in a sample average radar echo duration of 0.81 s and a sample standard deviation 0.34 s. Calculate a 99% (two-sided) CI for the true average echo duration.

Example 9 –

A sample of fish was selected from Florida lakes, and mercury concentration in the muscle tissue was measured (ppm).

1.230	1.330	0.040	0.044	0.490	0.190	0.830	0.810
0.490	1.160	0.050	0.150	1.080	0.980	0.630	0.560
0.590	0.340	0.340	0.840	0.280	0.340	0.750	0.870
0.180	0.190	0.040	0.490	0.100	0.210	0.860	0.520
0.940	0.400	0.430	0.250				

Find an approximate 95% CI on μ.

Solution

$n = 36, \overline{x} = 0.5284, s^2 = 0.1361, s = 0.3690, z_{0.025} = 1.96$. *Then the CI*

$$0.5284 \pm 1.96\frac{0.3690}{\sqrt{36}} = 0.5284 \pm 0.1205 = [0.4079, 0.6490]$$

Population Proportion

Corollary

Let $X \sim B(n,p)$ and assume $np \geq 10, nq \geq 10$. Then

$$\frac{\hat{p}-p}{\sqrt{pq/n}} \simeq \mathrm{N}(0,1)$$

An approximate $100\gamma\%$ confidence interval for p is

$$p \approx \hat{p} \pm z_{\alpha/2} \cdot \frac{\sqrt{\hat{p}\ \hat{q}}}{\sqrt{n}}$$

The approximate $100\gamma\%$ lower and upper confidence bounds are

$$p \gtrsim \hat{p} - z_{\alpha} \cdot \frac{\sqrt{\hat{p}\ \hat{q}}}{\sqrt{n}}$$

and

$$p \lesssim \hat{p} + z_{\alpha} \cdot \frac{\sqrt{\hat{p}\ \hat{q}}}{\sqrt{n}}$$

respectively.

Example 10 – Population Proportion

An article reported that in $n = 45$ trials in a particular laboratory, 16 resulted in ignition of a particular type of substrate by a lighted cigarette. Let p denote the long-run proportion of all such trials that would result in ignition. Find a point estimate for p and the confidence interval for p with a confidence level of about 95%.

Solution

A point estimate for p is $\hat{p} = 16/45 = 0.36$. The confidence interval for p is

$$0.36 \pm 1.96\sqrt{0.36 \cdot 0.64/45} = 0.36 \pm 0.14.$$

Summary

CI for the mean

Population distribution

Normal

Known σ^2 $\boxed{\mathrm{MOE} = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}}$

Unknown σ^2 $\boxed{\mathrm{MOE} = t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}}$

Any $n \gg 1$ $\boxed{\mathrm{MOE} = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}}$