

Mathematical Statistics

Anna Janicka

Lecture VIII, 8.04.2019

CONFIDENCE INTERVALS

Plan for Today

Interval estimation – confidence intervals, different models



Interval estimation – confidence intervals

- We do not provide a single value estimate, but rather a lower and an upper bound for the estimate (the true value will fit into these bounds with given probability)
- We estimate with given precision



Confidence interval

Let $g(\theta)$ be a function of unknown parameter θ , and let $\bar{g} = \bar{g}(X_1, X_2, \dots, X_n)$ and $\underline{g} = \underline{g}(X_1, X_2, \dots, X_n)$ be statistics

Then, $[\underline{g}, \bar{g}]$ is a **confidence interval** for $g(\theta)$ with a confidence level $1-\alpha$, if for any θ

$$P_{\theta}(\underline{g}(X_1, X_2, \dots, X_n) \leq g(\theta) \leq \bar{g}(X_1, X_2, \dots, X_n)) \geq 1 - \alpha$$



Confidence intervals – use and interpretation

- Typically: α is a small number, for example $1-\alpha = 0,95$ or $1-\alpha = 0,99$
- The condition from the definition means: the random interval $[\underline{g}, \bar{g}]$ includes the unknown value $g(\theta)$ with given (high) probability.
- If we calculate the *realization* of the confidence interval (e.g. $\underline{g} = 1, \bar{g} = 3$) then we CAN'T say that the unknown parameter is included in the range with probability $1-\alpha$ anymore!

the parameter is either in the interval or not – the event is not random, it is just something we don't know.



Confidence intervals – construction

- The confidence interval depends on the underlying probability distribution
- Usually, normal samples are considered (the distribution most frequently observed in nature)



Confidence intervals – construction cont.

- Convenient method: we look for random variables which depend on sample data and parameter values, but whose *distributions* do not depend on unknown parameters (*pivotal method*)
- If $U = U(X_1, X_2, \dots, X_n, \theta)$ is such a function, then we look for confidence intervals $[a, b]$ such that

$$P_{\theta}(a \leq U \leq b) \geq 1 - \alpha$$

- Usually we look for „symmetric” CI

$$P_{\theta}(U < a) \leq \frac{\alpha}{2}, \quad P_{\theta}(U > b) \leq \frac{\alpha}{2}$$



Most commonly used models

- Model I (normal): CI for the mean, variance known
- Model II (normal): CI for the mean, variance unknown
- Model II (normal): CI for the variance
- Model III (asymptotic): CI for the mean
- Model IV (asymptotic): CI for the fraction
- Asymptotic model: CI based on MLE



CI for the mean – Model I

Normal model: X_1, X_2, \dots, X_n are an IID sample from $N(\mu, \sigma^2)$, σ^2 is **known**.

The CI for μ , for a confidence level $1-\alpha$:

$$\left[\bar{X} - u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

where $u_{1-\alpha/2}$ is a quantile of rank $1-\alpha/2$ for the $N(0,1)$ distribution




CI for the mean – Model I, justification:

Point estimate for μ : $MLE(\mu) = \bar{X}$

We know the distribution of \bar{X} :

$$\bar{X} \sim N(\mu, \sigma^2/n), \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

distribution does not depend on μ 

We want: a CI symmetric around the point estimate (the distribution of the normalized average is symmetric around 0). We have:

$$P_{\mu} \left(\left| \sqrt{n}(\bar{X} - \mu) / \sigma \right| \leq u \right) = \Phi(u) - \Phi(-u) = 2\Phi(u) - 1$$
$$= 1 - \alpha$$

$$\text{so } u = u_{1-\alpha/2}$$



CI for the mean – Model I, properties

- Error: $d = u_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$
- Length of CI: $2d$
- Sample size allowing to obtain a given *precision* (error) d :

$$n \geq \frac{\sigma^2 u_{1-\alpha/2}^2}{d^2}$$



CI Model I – example phrasing

In a survey of food expenditures for $n=400$ randomly chosen respondents, the average weekly amount spent on fruit amounted to \$30. ***From previous research, we know that the variance of fruit expenditures is equal to 5.*** Assuming that food expenditures are distributed normally, find a 95% CI for the average weekly amount spent.



CI for the mean – Model II

Normal model: X_1, X_2, \dots, X_n are an IID sample from $N(\mu, \sigma^2)$, σ^2 is **unknown**.

The CI for μ , for a confidence level $1-\alpha$:

$$\left[\bar{X} - t_{1-\alpha/2}(n-1) \frac{S}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}(n-1) \frac{S}{\sqrt{n}} \right]$$

where $t_{1-\alpha/2}(n-1)$ is a quantile of rank $1-\alpha/2$ for a t -Student distribution with $n-1$ degrees of freedom $t(n-1)$, and $S = \sqrt{S^2}$ for the unbiased variance estimator S^2 .



CI for the mean – Model II, justification:

Point estimate for μ : $MLE(\mu) = \bar{X}$

We know the distribution of \bar{X} :

$$\bar{X} \sim N(\mu, \sigma^2/n), \quad \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

We want: a CI symmetric around the point estimate (the distribution of T is symmetric around 0). We have:

$$P_{\mu, \sigma} \left(\left| \sqrt{n}(\bar{X} - \mu) / S \right| \leq t \right) = 1 - \alpha$$

$$\text{so } t = t_{1-\alpha/2}(n-1)$$



CI for the mean – Model II, properties

□ Error:
$$d = t_{1-\alpha/2}(n-1) \frac{S}{\sqrt{n}}$$

□ Length of CI: $2d$

□ Sample size allowing to obtain a given *precision* (error) d :

to be determined on the base of the so-called Stein's two-stage procedure – we need a preliminary assessment of the variance first



Stein's two-stage procedure

1. We collect a preliminary sample X_1, X_2, \dots, X_{n_0} and estimate the variance:

$$S_0^2 = \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} (X_i - \bar{X}_0)^2$$

2. We check whether the sample fulfills the given condition: we calculate $k = \frac{S_0^2 [t_{1-\alpha/2}(n_0 - 1)]^2}{d^2}$

- a) if $n_0 \geq k$ then we take the CI

$$\left[\bar{X}_0 - t_{1-\alpha/2}(n_0 - 1) \frac{S_0}{\sqrt{n_0}}, \bar{X}_0 + t_{1-\alpha/2}(n_0 - 1) \frac{S_0}{\sqrt{n_0}} \right]$$

- b) if $n_0 < k$ then we choose $n \geq k$ and draw an additional sample of $X_{n_0+1}, X_{n_0+2}, \dots, X_n$. We calculate the mean for the *whole* sample X_1, X_2, \dots, X_n , and take the CI

$$\left[\bar{X} - t_{1-\alpha/2}(n_0 - 1) \frac{S_0}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2}(n_0 - 1) \frac{S_0}{\sqrt{n}} \right]$$



CI Model II – example phrasing

In a survey of food expenditures for $n=400$ randomly chosen respondents, the average weekly amount spent on fruit amounted to \$30, ***and the variance of fruit expenditures amounted to 5***. Assuming that food expenditures are distributed normally, find a 95% CI for the average weekly amount spent.



CI for the variance – Model II

Normal model: X_1, X_2, \dots, X_n are an IID sample from $N(\mu, \sigma^2)$

CI for σ^2 , for a confidence level $1-\alpha$:

$$\left[\frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)}, \frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)} \right]$$

where $\chi_{\alpha/2}^2(n-1)$ and $\chi_{1-\alpha/2}^2(n-1)$ are quantiles of rank $\alpha/2$ and $1-\alpha/2$, respectively, for a chi-squared distribution with $n-1$ degrees of freedom



CI for the variance – Model II, justification

Point estimate for σ^2 : S^2

We know the distr.: $U = \frac{(n-1)}{\sigma^2} S^2 \sim \chi^2(n-1)$

The chi-squared distribution is not symmetric. We want a „symmetric” CI, i.e. we look for $[a, b]$ such that

$$P_{\sigma^2}(U < a) = \frac{\alpha}{2}, \quad P_{\sigma^2}(U > b) = \frac{\alpha}{2}$$

so

$$a = \chi_{\alpha/2}^2(n-1) \text{ and } b = \chi_{1-\alpha/2}^2(n-1)$$



CI for the mean – Model III

Asymptotic model: X_1, X_2, \dots, X_n are an IID sample from a distr. with mean (μ) and variance, n – large.

Approximate CI for μ , for a confidence level $1-\alpha$:

$$\left[\bar{X} - u_{1-\alpha/2} \frac{S}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2} \frac{S}{\sqrt{n}} \right]$$

where $u_{1-\alpha/2}$ is a quantile of rank $1-\alpha/2$ from the $N(0,1)$ distribution, $S = \sqrt{S^2}$ for the unbiased estimator of the variance S^2 .

Justification: from CLT, when $n \rightarrow \infty$ we have

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{D} N(0,1)$$



CI for the fraction – Model IV

Asymptotic model: X_1, X_2, \dots, X_n are an IID sample from a two-point distribution, n – large.

$$P_p(X = 1) = p = 1 - P_p(X = 0)$$

Approximate CI for p , for a confidence level $1-\alpha$:

$$\left[\hat{p} - u_{1-\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{p} + u_{1-\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \right]$$

where $u_{1-\alpha/2}$ is a quantile of rank $1-\alpha/2$ from the $N(0,1)$ distribution



CI for the fraction – Model IV, justification

The point estimate for the fraction p :

$$\hat{p} = MLE(p) = \bar{X}$$

We know the asymptotic distribution: from CLT, when $n \rightarrow \infty$, we have

$$U = \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})}} \sqrt{n} \xrightarrow{D} N(0,1)$$

Using U , just like in model I, we get the formula.



CI for the fraction – Model IV, properties

- Assessment error: $d = u_{1-\alpha/2} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$
- Sample size allowing to obtain a given *precision* (error) d :

$$n \geq \frac{\hat{p}(1-\hat{p})u_{1-\alpha/2}^2}{d^2}$$

if we do not know anything about p , we need to consider the worst scenario

where $p=1/2$:
$$n \geq \frac{u_{1-\alpha/2}^2}{4d^2}$$



CI on the base of the MLE – Asymptotic model

Asymptotic model: X_1, X_2, \dots, X_n are an IID sample from a distr. with unknown parameter θ , n – large.

If $\hat{\theta} = MLE(\theta)$ is asymptotically normal with an asymptotic variance equal to $1/I_1(\theta)$, i.e.

$$(\hat{\theta} - \theta)\sqrt{n} \xrightarrow{D} N(0, 1/I_1(\theta))$$

and if $I(\hat{\theta}) = MLE(I(\theta))$ is consistent:

$$(\hat{\theta} - \theta)\sqrt{nl(\hat{\theta})} \xrightarrow{D} N(0,1)$$

Approximate CI for θ , for a confidence level $1-\alpha$:

$$\left[\hat{\theta} - u_{1-\alpha/2} \frac{1}{\sqrt{nl_1(\hat{\theta})}}, \hat{\theta} + u_{1-\alpha/2} \frac{1}{\sqrt{nl_1(\hat{\theta})}} \right]$$

where $u_{1-\alpha/2}$ is a quantile of rank $1-\alpha/2$ from $N(0,1)$



CI on the base of the MLE – Asymptotic model, general case

Asymptotic model: X_1, X_2, \dots, X_n are an IID sample from a distr. with unknown parameter θ , n – large.

If $g(\hat{\theta}) = g(MLE(\theta))$ is asymptotically normal with an asymptotic variance equal to $(g'(\theta))^2 / I_1(\theta)$, i.e.

$$(\hat{\theta} - \theta)\sqrt{n} \xrightarrow{D} N(0, (g'(\theta))^2 / I_1(\theta))$$

and if $I(\hat{\theta}) = MLE(I(\theta))$ is consistent:

$$(\hat{\theta} - \theta)\sqrt{nl(\hat{\theta})} \xrightarrow{D} N(0,1)$$

Approximate CI for $g(\theta)$, for a confidence level $1-\alpha$:

$$\left[g(\hat{\theta}) - u_{1-\alpha/2} \frac{|g'(\hat{\theta})|}{\sqrt{nl_1(\hat{\theta})}}, g(\hat{\theta}) + u_{1-\alpha/2} \frac{|g'(\hat{\theta})|}{\sqrt{nl_1(\hat{\theta})}} \right]$$

where $u_{1-\alpha/2}$ is a quantile of rank $1-\alpha/2$ from $N(0,1)$



CI on the base of the MLE – Example

Let X_1, X_2, \dots, X_n be an IID sample from a Poisson distr. with unknown parameter θ , n – large.

$\hat{\theta} = MLE(\theta) = \bar{X}$ is asymptotically normal (CLT) with an asymptotic variance equal to $1/I_1(\theta) = \theta$

$\hat{I}(\theta) = 1/\hat{\theta}$ behaves well.

Approximate CI for θ , for a confidence level $1-\alpha$:

$$\left[\bar{X} - u_{1-\alpha/2} \frac{\sqrt{\bar{X}}}{\sqrt{n}}, \bar{X} + u_{1-\alpha/2} \frac{\sqrt{\bar{X}}}{\sqrt{n}} \right]$$

where $u_{1-\alpha/2}$ is a quantile of rank $1-\alpha/2$ from $N(0,1)$

For example, if for $n=900$ we had $\bar{X} = 4$, then the 90% CI for θ would be

$$\approx \left[4 - 1.645 \sqrt{4/900}, 4 + 1.645 \sqrt{4/900} \right] \approx [3.89, 4.11]$$



CI on the base of the MLE – Example cont.

If we wanted to approximate the probability of the outcome = 0, we would look for $g(\theta) = e^{-\theta}$

$$g(\hat{\theta}) = g(MLE(\theta)) = e^{-\bar{X}}$$

And the *approximate* CI for $g(\theta)$, for a confidence level $1-\alpha$:

$$\left[e^{-\bar{X}} - u_{1-\alpha/2} \frac{\sqrt{\bar{X}}}{\sqrt{n}} e^{-\bar{X}}, e^{-\bar{X}} + u_{1-\alpha/2} \frac{\sqrt{\bar{X}}}{\sqrt{n}} e^{-\bar{X}} \right]$$

where $u_{1-\alpha/2}$ is a quantile of rank $1-\alpha/2$ from $N(0,1)$

For example, if for $n=900$ we had $\bar{X} = 4$, then the 90% CI for $g(\theta)$ would be

$$\approx \left[e^{-4} - 1.645 \sqrt{4/900} e^{-4}, e^{-4} + 1.645 \sqrt{4/900} e^{-4} \right] \approx [0.016, 0.020]$$





WARSAW UNIVERSITY
Faculty of Economic Sciences