MATH 503: Mathematical Statistics

Lecture 8: Confidence Intervals, and Bayesian Statistics

Reading: CB Sections 7.2.3, 9.1-9.2.1
(or HMC Sections 5.4, 7.1, 11.1-11.3)

Kimberly F. Sellers
Department of Mathematics and Statistics

Today's Topics

- Confidence intervals
  - "What is it?" and further introduction
  - Interpretation
  - Choosing a large enough sample size
- Bayesian Statistics
  - Subjective Probability
  - Prior and Posterior Distributions
  - Bayesian Point Estimation
  - Additional Terminology

What is a confidence interval?

An **interval estimate** of a real-valued parameter θ is any pair of functions $L(x_1, \ldots, x_n)$ and $U(x_1, \ldots, x_n)$ of a sample that satisfy $L(\boldsymbol{x}) \leq U(\boldsymbol{x}) \ \forall \boldsymbol{x} \in \chi$. If $X = x$ is observed, the inference $L(\boldsymbol{x}) \leq \theta \leq U(\boldsymbol{x})$ is made. The random interval $[L(\boldsymbol{x}), U(\boldsymbol{x})]$ is called an **interval estimator**.

What is a confidence interval? (cont.)

- An alternative to reporting a single estimate (i.e. the point estimate) for a parameter.
- Because of sampling variability, the point estimate is almost never exactly the correct value for the parameter
- Point estimates don’t tell us how close we are to the actual parameter
- Instead, confidence intervals are calculated intervals of plausible values for the parameter

Coverage Probabilities & Confidence Coefficients

- For an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ of a parameter θ, the coverage probability of $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is the probability that the random interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ covers the true parameter, θ, i.e. $\mathrm{P}_\theta(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{X})])$.
- For an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ of a parameter θ, the confidence coefficient of $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is the infimum of the coverage probabilities, i.e. $\inf_\theta \mathrm{P}_\theta(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{X})])$.

What do you need to compute a confidence interval?

- a sample of observations (for purposes of discussion, from a normal distribution)
- the sample estimate, $\hat{\theta}$, and its standard error
- confidence level, traditionally 95% (other popular choices are 90% or 99%; generally, its $100(1-\alpha)\%$); the higher the confidence level, the more strongly we believe the value of θ falls within the interval

Creating a confidence interval for θ

- Let θ_0 denote the true (unknown) value of θ
- Suppose $\bar{X}$ estimator of θ_0 such that

$$\sqrt{n}(\bar{X} - \theta_0) \xrightarrow{d} \mathrm{N}(0, \sigma^2),$$

- Assume σ^2 known

Confidence interval for θ (cont.)

- A $100(1-\alpha)\%$ confidence interval (interval estimator) for θ of a normal population when the value of σ is known is

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

- The interval estimate is determined by replacing $\bar{X}$ above with $\bar{x}$

Examples

- Confidence interval for μ when σ known:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Large-sample interval for μ when σ unknown:

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

Examples (cont.)

- Small-sample interval for μ when σ unknown:

$$\bar{x} \pm t_{\alpha/2,\, n-1} \frac{s}{\sqrt{n}}$$

- Large-sample interval for p:

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

Example

A random sample of 100 observations from a normally distributed population possesses a mean equal to 83.2. Assume $\sigma = 6.4$. Find a 95% confidence interval for μ.

Questions

- How does the confidence interval change when
  - sample size increases?
  - σ increases?
  - α increases (say, to $\alpha' > \alpha$)?

Example (cont.)

Would a 90% confidence interval calculated from the same sample be narrower or wider than the above interval? Why?

Interpreting a confidence interval

- Question: how do I interpret, say, a 95% confidence interval?
  - Bayesian perspective: a belief with 95% confidence that the parameter, θ, is in the stated interval. We can generalize this idea for any $100(1-\alpha)\%$ confidence interval, i.e. we are $100(1-\alpha)\%$ sure that parameter, θ, is contained in the interval in question.

Interpreting a confidence interval (cont.)

– Frequentist perspective:

95% of all samples would give an interval that includes the parameter, θ, and 5% of all samples would give an erroneous interval. More generally, $100(1-\alpha)\%$ of all samples would give an interval that includes the parameter, θ, and $100\alpha\%$ of all samples would give an erroneous interval.

Steps to constructing confidence intervals

1. Answer the following questions to determine which kind of confidence interval to create:
   - What parameter do I want to estimate?
   - What is my sample size? Large or small?
   - Is the population variance, σ, known or unknown?
2. Determine the appropriate value of α.
3. Find the corresponding critical value(s) (e.g. $z_{\alpha/2}$ or $t_{\alpha/2}$) depending on the responses to Step 1.
4. Use the appropriate formula to compute the confidence interval.

Example

The Gallup Organization conducts annual national surveys on home gardening. Results are published by the National Association for Gardening in National Gardening Survey. A random sample is taken of 25 households with vegetable gardens. The mean size of their gardens is 643 sq. ft. with standard deviation equal to 247 sq. ft.

a) Determine a 90% confidence interval for the mean size, μ, of all household vegetable gardens in the U.S.

b) Interpret your answer from part (a).

Example

The accounting firm Price Waterhouse periodically monitors the U.S. Postal Service's performance. One parameter of interest is the percentage of mail delivered on time. In a sample of $332{,}000$ mailed items, Price Waterhouse determined that $282{,}200$ items were delivered on time (Tampa Tribune, Mar. 26, 1995). Use this information to estimate with 99% confidence the true percentage of items delivered on time by the U.S. Postal Service. Interpret the results.

Confidence Intervals for Differences of Means

- Assume samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$ gathered independently s.t. sample sizes have relative order
- Large-sample interval for $\Delta = \mu_1 - \mu_2$ when σ_1, σ_2 known:

$$\bar{x} - \bar{y} \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

For σ_1, σ_2 unknown, replace with s_1, s_2, respectively

Note: this CI is approximate based on asymptotic distribution theory!

Confidence Intervals for Differences of Means

- $X_1, \ldots, X_{n_1} \sim N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_{n_2} \sim N(\mu_2, \sigma^2)$, independently drawn
- Exact confidence interval with common σ^2:

$$\bar{x} - \bar{y} \pm t_{\alpha/2, n-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

Confidence Interval for Difference of Proportions

- $X_1, \ldots, X_{n_1} \sim \text{Bern}(p_1)$; $Y_1, \ldots, Y_{n_2} \sim \text{Bern}(p_2)$, independently drawn

- Large-sample interval for $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Example

In order to compare the means of two populations, independent random samples of 400 observations are selected from each population, with the following results:

Sample 1	Sample 2
$\bar{x} = 5275$	$\bar{y} = 5240$
$s_1 = 150$	$s_2 = 200$

Use a 95% CI to estimate the difference between the population means, $\mu_1 - \mu_2$.

Sample Size Determination

- **Idea:** we know the level of confidence we want to have regarding an interval surrounding our unknown parameter of interest, θ, and we also have a specified error bound or width for the confidence interval in question.
- **Goal:** to know how large the sample size, n, needs to be in order to fulfill these specifications.

- **How do we do that?**

Example

The EPA wants to test a randomly selected sample of n water specimens and estimate the mean daily rate of pollution produced by a mining operation. If the EPA wants a 95% confidence interval estimate with a bound on the error of 1 milligram per liter (mg/L), how many water specimens are required in the sample? Assume prior knowledge indicates that pollution readings in water samples taken during a day are approximately normally distributed with a standard deviation equal to 5 mg/L.

Example

According to estimates made by the General Accounting Office, the Internal Revenue Service (IRS) answered 18.3 million telephone inquiries during a recent tax season, and 17% of the IRS offices provided answers that were wrong. These estimates were based on data collected from sample calls to numerous IRS offices. How many IRS offices should be randomly selected and contacted in order to estimate the proportion of IRS offices that fail to correctly answer questions about gift taxes with a 90% confidence interval of width $.06$?

What is subjective probability?

- A bet
- All probability axioms and associated results hold under this philosophical construct!

Useful Tools

- The **conditional probability** of A given that the event B has occurred is

$$P(A|B) = \frac{P(A \cap B)}{P(B)},$$

where $P(B) > 0$.

- Note: by cross-multiplying both sides, we get "the **multiplication rule**":

$$P(A \cap B) = \mathrm{P}(\,\mathrm{A} \mid B\,) \cdot P(B)$$

Useful Tools (cont.)

The Law of Total Probability

$$P(B) = P(B \mid A)P(A) + P(B \mid A')P(A')$$

Bayes' Theorem/ Bayes' Law

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A')P(A')}$$

Prior & Posterior Distributions

- $X|\theta \sim f(x|\theta), \theta \in \Omega$
- $\Theta \sim \pi(\theta)$; called prior distribution

- Suppose $X_1, X_2, \ldots, X_n$ random sample drawn from population indexed by θ with pdf $f(x|\theta)$. Then, sampling distribution is

$$f(\boldsymbol{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

and joint pdf of $\boldsymbol{X}' = (X_1, X_2, \ldots, X_n)$ and Θ is

$$f(\boldsymbol{x}, \theta) = f(\boldsymbol{x}|\theta)\pi(\theta)$$

Prior & Posterior Dists. (cont.)

- Conditional pdf of Θ, given sample $\boldsymbol{X}$:

$$\pi(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x}, \theta)}{m(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|\theta)\pi(\theta)}{m(\boldsymbol{x})},$$

  where $m(\boldsymbol{x}) = \int_{-\infty}^{\infty} f(\boldsymbol{x}|\theta)\pi(\theta)\mathrm{d}\theta$

- This is posterior pdf of θ given $\boldsymbol{x}$

- **What does this mean???**

Example

Let $X_i \mid \theta \sim \text{Poisson}(\theta)$ iid, and $\Theta \sim \text{Gamma}(\alpha, \beta)$ where α, β known. Find the posterior distribution of $\theta \mid \boldsymbol{x}$.

Notes

- $\pi(\theta \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \theta) \cdot \pi(\theta)$
- If a sufficient statistic $Y = u(\boldsymbol{X})$ exists, then $\pi(\theta \mid \boldsymbol{x}) \propto f(u(\boldsymbol{x}) \mid \theta) \cdot \pi(\theta)$, thus

$$\pi(\theta \mid y) \propto f(y \mid \theta) \cdot \pi(\theta)$$

Example

Let $X_i \mid \theta \sim \text{Bernoulli}(\theta)$ iid $i = 1, \ldots, n$, and $\Theta \sim \text{Beta}(\alpha, \beta)$ where α, β known. Find the posterior distribution of $\theta \mid \boldsymbol{x}$.

Bayesian Point Estimation: Definitions and Framework

- $X_1, X_2, \ldots, X_n$ iid with pdf $f(x;\theta), \theta \in \Omega$.
- $Y = u(X_1, X_2, \ldots, X_n)$: statistic on which point estimate of θ is determined
- $\delta(y)$: decision rule, determining value of point estimate at θ
  - Numerical point estimate is a decision
- $\mathcal{L}(\theta, \delta(y))$: loss function, reflecting severity of difference between θ and $\delta(y)$

Loss Functions

- Examples include:
  - Squared-error loss: $\mathcal{L}\big(\theta, \delta(y)\big) = [\theta - \delta(y)]^2$
  - Absolute-value error loss: $\mathcal{L}\big(\theta, \delta(y)\big) = |\theta - \delta(y)|$
  - "Goal-post" loss:

$$\mathcal{L}\big(\theta, \delta(y)\big) = \begin{cases} 0, & \text{if } |\theta - \delta(y)| \leq a \\ b, & \text{otherwise} \end{cases}$$

- Don't have to be symmetric
- Goal of point estimation: find $\delta(y)$ s.t. $\mathcal{L}(\theta, \delta(y))$ is minimum

Risk Function

- **Problem:** θ unknown, thus minimizing $\mathcal{L}\big(\theta, \delta(y)\big)$ impossible
- Introduce risk function:

$$R(\theta, \delta) = E\big[\mathcal{L}\big(\theta, \delta(y)\big)\big] = \int_{-\infty}^{\infty} \mathcal{L}\big(\theta, \delta(y)\big) f_Y(y; \theta) \mathrm{d}y$$

- **Goal**: select $\delta(y)$ s.t. $R(\theta, \delta)$ minimum for all $\theta \in \Omega$
  - **Problem:** $\delta(y)$ s.t. $R(\theta, \delta)$ minimum does not minimize $R(\theta', \delta)$
  - **Resolution:** restrict decision function to certain class, or order risk functions

Minimax Principle

- If the decision function given by $\delta_0(y)$ is such that, for all $\theta \in \Omega$,

$$\max_{\theta} R[\theta, \delta_0(y)] \leq \max_{\theta} R[\theta, \delta(y)]$$

for every other decision function $\delta(y)$, then $\delta_0(y)$ is called a minimax decision function.

Example

Let $X_1, X_2, \ldots, X_{25} \sim \mathrm{N}(\theta, 1)$ iid, $-\infty < \theta < \infty$, $Y = \bar{X}$, and $\mathcal{L}\big(\theta, \delta(y)\big) = [\theta - \delta(y)]^2$. Consider two decision functions $\delta_1(y) = y$, and $\delta_2(y) = 0$ for $-\infty < y < \infty$.

1. Find the associated risk function associated with each decision function.
2. Which is better if you:
   (a) restrict the choice of $\delta(y)$ to satisfy $\mathrm{E}[\delta(y)] = \theta$?
   (b) want $\delta(y)$ to satisfy the minimax principle?

Example

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a distribution that is $\text{Bin}(1, \theta)$, $0 \leq \theta \leq 1$. Let $Y = \sum_{i=1}^{n} X_i$, and let $\mathcal{L}\big(\theta, \delta(y)\big) = [\theta - \delta(y)]^2$. Consider decision functions of the form $\delta(y) = by$, where b does not depend upon y. Prove that $R(\theta, \delta) = b^2 n\theta(1-\theta) + (bn-1)^2\theta^2$.

Notes

- $Y = u(\boldsymbol{X})$ statistic $\Rightarrow \delta(Y)$ statistic
- If $\delta(Y)$ satisfies $\mathrm{E}[\delta(y)] = \theta$, $\mathcal{L}\big(\theta, \delta(y)\big) = [\theta - \delta(y)]^2$, and minimizes $R(\theta, \delta)$, then $\delta(y)$ is MVUE

How does this change with knowledge from random sample?

- Goal: determine point estimator for θ, i.e. select decision function δ s.t. $\delta(\boldsymbol{x})$ is predicted value of θ when computed $\boldsymbol{x}$ and posterior distribution $\pi(\theta \mid \boldsymbol{x})$ known
- In other words, choose $\delta(\boldsymbol{x})$ s.t. minimize

$$E\{\mathcal{L}(\Theta, \delta(\boldsymbol{x})) \mid \boldsymbol{X} = \boldsymbol{x}\} = \int_{-\infty}^{\infty} \mathcal{L}(\theta, \delta(\boldsymbol{x}))\pi(\theta \mid \boldsymbol{x})\mathrm{d}\theta \quad (1)$$

- $\delta(\boldsymbol{x})$ called a Bayes estimator of θ
- Can generalize to find Bayes estimator of $\ell(\theta)$ by replacing θ with $\ell(\theta)$ in $\mathcal{L}(\theta, \delta(\boldsymbol{x}))$ of Equation (1)

How does this change …? (con't.)

- $E\{\mathcal{L}(\Theta, \delta(\boldsymbol{x})) \mid \boldsymbol{X} = \boldsymbol{x}\}$ is r.v. that is a function of $\boldsymbol{X}$
- Expected risk =

$$\int_{-\infty}^{\infty} \underbrace{\left\{ \int_{-\infty}^{\infty} \mathcal{L}(\Theta, \delta(\boldsymbol{x}))\, \pi(\theta \mid \boldsymbol{x})\, \mathrm{d}\theta \right\}}_{E\{\mathcal{L}(\Theta, \delta(\boldsymbol{x})) \mid \boldsymbol{X} = \boldsymbol{x}\}} m(\boldsymbol{x})\, d\boldsymbol{x}$$

$$= \int_{-\infty}^{\infty} \underbrace{\left\{ \int_{-\infty}^{\infty} \mathcal{L}(\Theta, \delta(\boldsymbol{x}))\, f(\boldsymbol{x} \mid \theta)\, \mathrm{d}\boldsymbol{x} \right\}}_{R(\theta, \delta), \text{ for every } \theta \in \Theta} \pi(\theta)\, d\theta$$

- $\delta(\boldsymbol{x})$ minimizing $E\{\mathcal{L}(\Theta, \delta(\boldsymbol{x})) \mid \boldsymbol{X} = \boldsymbol{x}\}$ for every satisfactory $\boldsymbol{x}$ $\Rightarrow$ it minimizes expected risk

Estimators Associated with Loss Functions

- For squared-error loss function,

$$\mathcal{L}\big(\theta, \delta(y)\big) = [\theta - \delta(y)]^2,$$

the conditional distribution mean of Θ, given $\boldsymbol{X} = \boldsymbol{x}$, is Bayes estimator

- For absolute-value loss function,

$$\mathcal{L}\big(\theta, \delta(y)\big) = |\theta - \delta(y)|,$$

the conditional distribution median of Θ, given $\boldsymbol{X} = \boldsymbol{x}$, is Bayes estimator

Example

Let $X_i \mid \theta \sim \text{Binomial}(1,\theta)=\text{Bernoulli}(\theta)$ iid, $i = 1, \ldots, n$, and $\Theta \sim \text{Beta}(\alpha, \beta)$ where α, β known. Find the Bayes estimator of θ using a squared-error loss function.

Example

Let $X_i \mid \theta \sim \mathrm{N}(\theta, \sigma^2)$ iid, $i = 1, \ldots, n$, where σ^2 known, and $\Theta \sim N(\theta_0, \sigma_0^2)$ where θ_0, σ_0^2 known. Find the Bayes estimator of θ using a squared-error loss function.

Example

Let Y_1 and Y_2 be statistics that have trinomial distribution with parameters n, θ_1, and θ_2. Here θ_1 and θ_2 are observed values of the random variables Θ_1 and Θ_2, which have a Dirichlet distribution with known parameters α_1, α_2, and α_3. Show that the conditional distribution of Θ_1 and Θ_2 is Dirichlet and determine the conditional means $E(\Theta_1 \mid y_1, y_2)$ and $E(\Theta_2 \mid y_1, y_2)$.

Bayesian Terminology

- A class of prior pdfs for the family of distributions with pdfs $f(x \mid \theta)$, $\theta \in \Omega$ is said to be a conjugate family of distributions if the posterior pdf of the parameter is in the same family of distributions as the prior.

- Let $\boldsymbol{X}' = (X_1, \ldots, X_n)$ be a random sample from the distribution with pdf $f(x \mid \theta)$. A prior $\pi(\theta) \geq 0$ for this family is said to be improper if it is not a pdf but the function $\pi(\theta \mid \boldsymbol{x}) \propto f(\boldsymbol{x} \mid \theta) \cdot \pi(\theta)$ can be made proper.

Bayesian Terminology (cont.)

- A noninformative prior is a prior that treats all values of θ uniformly equal
  - Note: Continuous noninformative priors are often improper
  - Example: $X \mid (\theta_1, \theta_2) \sim \mathrm{N}(\theta_1, \theta_2)$ and θ_1 has prior $h(\theta_1) = 1,\ -\infty < \theta_1 < \infty$