STA261 Lecture 3 — 2017-07-12

Neil Montgomery

Last edited: 2017-07-12 19:07

parameter estimation

Basic problem

We have a probability model for something, e.g.

- heights have a $N(\mu, \sigma^2)$ distribution

Basic problem

We have a probability model for something, e.g.

- heights have a $N(\mu, \sigma^2)$ distribution
- failure times have an $Exp(\lambda)$ distribution

Basic problem

We have a probability model for something, e.g.

- heights have a $N(\mu, \sigma^2)$ distribution
- failure times have an $Exp(\lambda)$ distribution
- some thing is random with mean μ and variance σ^2

Basic problem

We have a probability model for something, e.g.

- heights have a $N(\mu, \sigma^2)$ distribution
- failure times have an $Exp(\lambda)$ distribution
- some thing is random with mean μ and variance σ^2

Basic problem

We have a probability model for something, e.g.

- heights have a $N(\mu, \sigma^2)$ distribution
- failure times have an $Exp(\lambda)$ distribution
- some thing is random with mean μ and variance σ^2

But we don't know the parameter values, or sometimes even the underlying distribution.

Basic problem

We have a probability model for something, e.g.

- heights have a $N(\mu, \sigma^2)$ distribution
- failure times have an $Exp(\lambda)$ distribution
- some thing is random with mean μ and variance σ^2

But we don't know the parameter values, or sometimes even the underlying distribution.

So we plan to gather a sample $X_1, \ldots, X_n$ i.i.d. from the underlying distribution. Then what?

point estimation

We treat population parameters as constants (as opposed to *Bayesian statistics...*)

The goal is to use an estimator $\hat{\theta}$, which is any function of a sample (i.e. a "statistic") to estimate the value of a parameter θ, which could be a vector.

point estimation

We treat population parameters as constants (as opposed to *Bayesian statistics...*)

The goal is to use an estimator $\hat{\theta}$, which is any function of a sample (i.e. a “statistic”) to estimate the value of a parameter θ, which could be a vector.

e.g. Sample is from $N(\mu, 3)$. We want to estimate μ.

point estimation

We treat population parameters as constants (as opposed to *Bayesian statistics...*)

The goal is to use an estimator $\hat{\theta}$, which is any function of a sample (i.e. a "statistic") to estimate the value of a parameter θ, which could be a vector.

e.g. Sample is from $N(\mu, 3)$. We want to estimate μ.

e.g. Sample is from Bernoulli(p). We want to estimate p.

point estimation

We treat population parameters as constants (as opposed to *Bayesian statistics...*)

The goal is to use an estimator $\hat{\theta}$, which is any function of a sample (i.e. a "statistic") to estimate the value of a parameter θ, which could be a vector.

e.g. Sample is from $N(\mu, 3)$. We want to estimate μ.

e.g. Sample is from Bernoulli(p). We want to estimate p.

e.g. Sample is from $N(\mu, \sigma^2)$. We want to estimate (μ, σ).

open questions about estimators

What are some desirable properties that an estimator might have?

How do I figure out which estimator to use, from first principles?

desirable properties

desirable properties

Unbiasedness: we would like the estimator to be, on average, equal to the true value. This is called "unbiased".

desirable properties

Unbiasedness: we would like the estimator to be, on average, equal to the true value. This is called "unbiased".

Error: we would like the estimator to be as close as possible to the true value.

desirable properties

Unbiasedness: we would like the estimator to be, on average, equal to the true value. This is called "unbiased".

Error: we would like the estimator to be as close as possible to the true value.

Consistency: we would like the estimator to get closer (in probability) to the "true" value as the sample size gets larger.

desirable properties

Unbiasedness: we would like the estimator to be, on average, equal to the true value. This is called "unbiased".

Error: we would like the estimator to be as close as possible to the true value.

Consistency: we would like the estimator to get closer (in probability) to the "true" value as the sample size gets larger.

Invariance: if $\hat{\theta}$ is a good estimator for θ, we would like $g(\hat{\theta})$ to be a good estimator for $g(\theta)$ (g invertible).

desirable properties

Unbiasedness: we would like the estimator to be, on average, equal to the true value. This is called "unbiased".

Error: we would like the estimator to be as close as possible to the true value.

Consistency: we would like the estimator to get closer (in probability) to the "true" value as the sample size gets larger.

Invariance: if $\hat{\theta}$ is a good estimator for θ, we would like $g(\hat{\theta})$ to be a good estimator for $g(\theta)$ (g invertible).

Sufficiency: we would like the estimator to contain all possible information about the parameter that is present in the sample.

desirable properties

Unbiasedness: we would like the estimator to be, on average, equal to the true value. This is called "unbiased".

Error: we would like the estimator to be as close as possible to the true value.

Consistency: we would like the estimator to get closer (in probability) to the "true" value as the sample size gets larger.

Invariance: if $\hat{\theta}$ is a good estimator for θ, we would like $g(\hat{\theta})$ to be a good estimator for $g(\theta)$ (g invertible).

Sufficiency: we would like the estimator to contain all possible information about the parameter that is present in the sample.

Large sample properties: we would like to know how the estimator behaves when the sample size is large.

bias

It is desirable to be correct on average. We say $\hat{\theta}$ is unbiased for θ when $E(\hat{\theta}) = \theta$. The bias of an estimator is $B(\hat{\theta}) = \hat{\theta} - \theta$.

bias

It is desirable to be correct on average. We say $\hat{\theta}$ is unbiased for θ when $E(\hat{\theta}) = \theta$. The bias of an estimator is $B(\hat{\theta}) = \hat{\theta} - \theta$.

Example 3.0: $\overline{X}$ is always unbiased for the mean μ of any population.

bias

It is desirable to be correct on average. We say $\hat{\theta}$ is unbiased for θ when $E(\hat{\theta}) = \theta$. The bias of an estimator is $B(\hat{\theta}) = \hat{\theta} - \theta$.

Example 3.0: $\overline{X}$ is always unbiased for the mean μ of any population.

Example 3.1: S^2 is unbiased for σ^2 when the popuation is $N(\mu, \sigma^2)$.

bias

It is desirable to be correct on average. We say $\hat{\theta}$ is unbiased for θ when $E(\hat{\theta}) = \theta$. The bias of an estimator is $B(\hat{\theta}) = \hat{\theta} - \theta$.

Example 3.0: $\overline{X}$ is always unbiased for the mean μ of any population.

Example 3.1: S^2 is unbiased for σ^2 when the popuation is $N(\mu, \sigma^2)$.

Example 3.2: Suppose $X_1, \ldots, X_n$ is i.i.d. Exponential with rate λ. The mean of an Exp(λ) is $1/\lambda$. What to do???

mean square error

The **mean square error** of an estimator is:

$$MSE(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right)$$

mean square error

The **mean square error** of an estimator is:

$$MSE(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right)$$

Proposition 3.3: $MSE(\hat{\theta}) = B(\hat{\theta})^2 + \text{Var}\left(\hat{\theta}\right)$, so when $\hat{\theta}$ is unbiased, $MSE(\hat{\theta}) = \text{Var}\left(\hat{\theta}\right)$.

mean square error

The **mean square error** of an estimator is:

$$MSE(\hat{\theta}) = E\Big((\hat{\theta} - \theta)^2\Big)$$

Proposition 3.3: $MSE(\hat{\theta}) = B(\hat{\theta})^2 + \text{Var}\Big(\hat{\theta}\Big)$, so when $\hat{\theta}$ is unbiased, $MSE(\hat{\theta}) = \text{Var}\Big(\hat{\theta}\Big)$.

A common criteria used to say an estimator is the "best" is to say, among all unbiased estimators, it has the smallest variance. (Major goal of this course.)

unbiased with smaller variance

Example 3.4: Consider estimating μ from a $N(\mu, \sigma^2)$ population using a sample $X_1, \ldots, X_n$. These are some unbiased estimators for μ: $\overline{X}$, X_1, $(X_1 + X_n)/2$. Compare their variances.

unbiased with smaller variance

Example 3.4: Consider estimating μ from a $N(\mu, \sigma^2)$ population using a sample $X_1, \ldots, X_n$. These are some unbiased estimators for μ: $\overline{X}$, X_1, $(X_1 + X_n)/2$. Compare their variances.

It turns out $\overline{X}$ has the smallest variance among all unbiased estimators. (To be proven.)

consistency

This is a minimally good property to have.

To set things up, let's say $\hat{\theta}_n$ is an estimator for θ from a sample $X_1, \ldots, X_n$.

consistency

This is a minimally good property to have.

To set things up, let's say $\hat{\theta}_n$ is an estimator for θ from a sample $X_1, \ldots, X_n$.

By consistency* I mean:

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

consistency

This is a minimally good property to have.

To set things up, let's say $\hat{\theta}_n$ is an estimator for θ from a sample $X_1, \ldots, X_n$.

By consistency* I mean:

$$\lim_{n \to \infty} P(|\hat{\theta}_n - \theta| > \varepsilon) = 0$$

Recall (or welcome to. . .) Chebyshev's inequality:

$$P(|Y - E(Y)| > \varepsilon) \leqslant \frac{\text{Var}(Y)}{\varepsilon^2}$$

consistency examples

consistency examples

Theorem 3.5: Given a sample $X_1, \ldots, X_n$ i.i.d. from any distribution with mean μ and variance σ^2, $\overline{X}_n$, is a consistent estimator for μ. In addition, if the distribution is normal, S_n^2 is consistent for σ^2.

consistency examples

Theorem 3.5: Given a sample $X_1, \ldots, X_n$ i.i.d. from any distribution with mean μ and variance σ^2, $\overline{X}_n$, is a consistent estimator for μ. In addition, if the distribution is normal, S_n^2 is consistent for σ^2.

Note 3.6: Actually S_n^2 is also consistent under the original conditions of the theorem (tedious to show).

consistency examples

Theorem 3.5: Given a sample $X_1, \ldots, X_n$ i.i.d. from any distribution with mean μ and variance σ^2, $\overline{X}_n$, is a consistent estimator for μ. In addition, if the distribution is normal, S_n^2 is consistent for σ^2.

Note 3.6: Actually S_n^2 is also consistent under the original conditions of the theorem (tedious to show).

Note 3.7: An unbiased estimator $\hat{\theta}_n$ for θ, with variance that converges to 0, is consistent for θ.

consistency examples

Theorem 3.5: Given a sample $X_1, \ldots, X_n$ i.i.d. from any distribution with mean μ and variance σ^2, $\overline{X}_n$, is a consistent estimator for μ. In addition, if the distribution is normal, S_n^2 is consistent for σ^2.

Note 3.6: Actually S_n^2 is also consistent under the original conditions of the theorem (tedious to show).

Note 3.7: An unbiased estimator $\hat{\theta}_n$ for θ, with variance that converges to 0, is consistent for θ.

Non-example 3.8: The stupid estimators for μ from before, X_1 and $(X_1 + X_n)/2$, are not consistent.

techniques for finding parameter estimators

method of moments

Think back to the example where we dreamed up an estimator for the rate λ from an $\mathrm{Exp}(\lambda)$ distribution.

method of moments

Think back to the example where we dreamed up an estimator for the rate λ from an $\text{Exp}(\lambda)$ distribution.

We made a correspondence between the parameter and the first moment (AKA the mean), plugged the sample average in place of the mean, and solved for the parameter value.

method of moments

Think back to the example where we dreamed up an estimator for the rate λ from an $\mathrm{Exp}(\lambda)$ distribution.

We made a correspondence between the parameter and the first moment (AKA the mean), plugged the sample average in place of the mean, and solved for the parameter value.

That's the method of moments, in a nutshell.

method of moments

The k^{th} moment of X (the "underlying population") is $E(X^k)$ (if it exists).

Definition: the k^{th} *sample moment* of a sample $X_1, \ldots, X_n$ i.i.d. with same distribution as X is:

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

The method of moments: Express the parameter(s) of the distribution as function(s) of moment(s), invert the functions, and replace moments with sample moments.

method of moments examples

Example 3.9: Bernoulli(p) distribution.

Example 3.10: Estimate μ and σ from a $N(\mu, \sigma^2)$ distribution.

Example 3.11: Estimate η from a Weibull$(2, \eta)$ distribution.

value of method of moments as a technique

Method of moments estimators are useful because:

1. They are consistent under some mild conditions.

value of method of moments as a technique

Method of moments estimators are useful because:

1. They are consistent under some mild conditions.
2. They might be the only estimators available (i.e. other techniques we haven't seen yet don't work.)