Parameter Estimation

Statistics and Estimation

A random variable that is derived from a random sample $X_1, \dots, X_n$ of a population is said to be ***statistic***. Examples include

- any of the sample quartiles q_1, q_2, q_3,
- the sample maximum $\max\{X_1, \dots, X_n\}$,
- the sample mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

We would like to use a given sample statistic to ***estimate*** a population parameter.

For example, the sample mean $\overline{X}$ can be used to estimate the population mean μ.

Any statistic that is used in such way is then called an ***estimator*** and the value of the statistic a ***point estimate***.

Bias and Mean Square Error

We would like an estimator to have the following properties:

- The expected value of $\widehat{\theta}$ should be equal to θ,
- $\widehat{\theta}$ should have small variance for large sample sizes.

This motivates the following definition:

12.1. Definition. The difference

$$\theta - \mathsf{E}[\widehat{\theta}]$$

is called the ***bias*** of an estimator $\widehat{\theta}$ for a population parameter θ. If $\mathsf{E}[\widehat{\theta}] = \theta$, we say that $\widehat{\theta}$ is ***unbiased***.

The ***mean square error*** of $\widehat{\theta}$ is defined as

$$\mathsf{MSE}(\widehat{\theta}) := \mathsf{E}[(\widehat{\theta} - \theta)^2].$$

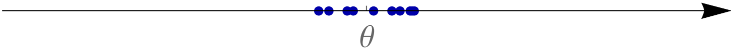
Quality of Estimators

The mean square error measures the overall quality of an estimator. We can write

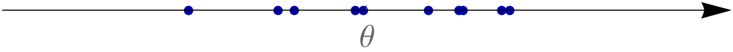
$$\begin{aligned}\mathsf{MSE}(\widehat{\theta}) &= \mathsf{E}[(\widehat{\theta} - \mathsf{E}[\widehat{\theta}])^2] + (\theta - \mathsf{E}(\widehat{\theta}))^2 \\ &= \mathsf{Var}\,\widehat{\theta} + (\mathsf{bias})^2.\end{aligned}$$

Hence variance can be just as important as bias for an estimator. In general, unbiased estimators are preferred but sometimes biased estimators are used.

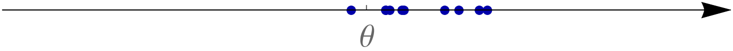
Simulation of 10 Estimates of θ



Unbiased, small variance



Unbiased, large variance



Biased, small variance

Biased, large variance

Sample Mean

12.2. Theorem. Let $X_1, \dots, X_n$ be a random sample of size n from a distribution with mean μ. The sample mean $\overline{X}$ is an unbiased estimator for μ.

Proof.
We simply insert the definition of the sample mean and use the properties of the expectation:

$$\begin{aligned}
\mathsf{E}[\overline{X}] &= \mathsf{E}[(X_1 + \cdots + X_n)/n] = \frac{1}{n}\mathsf{E}[X_1 + \cdots + X_n] \\
&= \frac{1}{n}(\mathsf{E}[X_1] + \cdots + \mathsf{E}[X_n]) = \frac{n\mu}{n} = \mu.
\end{aligned}$$

□

12.3. Theorem. Let $\overline{X}$ be the sample mean of a random sample of size n from a distribution with mean μ and variance σ^2. Then

$$\operatorname{Var}\overline{X} = \mathsf{E}[(\overline{X} - \mu)^2] = \frac{\sigma^2}{n}.$$

Sample Variance

Proof.
We simply insert the definition of the sample mean and use the properties of the variance:

$$\begin{aligned}\operatorname{Var}\overline{X} &= \operatorname{Var}((X_1+\cdots+X_n)/n) = \frac{1}{n^2}\operatorname{Var}(X_1+\cdots+X_n) \\ &= \frac{1}{n^2}(\operatorname{Var}X_1+\cdots+\operatorname{Var}X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.\end{aligned}$$

□

Thus $\overline{X}$ is both unbiased and has a variance that decreases with large n; it is a "nice" estimator, since we can make the mean square error $\operatorname{MSE}\overline{X}$ as small as desired by taking n large enough.

12.4. Definition. The standard deviation of $\overline{X}$ is given by $\sqrt{\operatorname{Var}\overline{X}} = \sigma/\sqrt{n}$ and is called the ***standard error of the mean***.

Medieval Standard Units of Measurement

Determination of the rood and the foot in Frankfurt, in J. Köbel, ***Geometrei. Von künstlichem Feldmessen und Absehen***, published in Frankfurt, 1570. Image via Wikimedia, http://commons.wikimedia.org/wiki/File:Determination_of_the_rute_and_the_feet_in_Frankfurt.png

Medieval Standard Units of Measurement

The picture on the previous slide is taken from the book "Geometrei. Von künstlichem Feldmessen und Absehen", published in Frankfurt at the beginning of the 16th century. An edition from 1570 is available online at http://books.google.de/books?id=80JSAAAAcAAJ&pg=PA1.

The book describes the recommended method for obtaining a measurement of "1 foot" (although it doesn't actually use the term):

> *Sixteen men, small and large, as they freely leave the church one after the other, are each to put in front of the other a shoe. This same length is and shall be a right and proper measuring rood. [...] Using a compass, this same measured rood is to be divided and distinguished into sixteen equal parts and shall forthwith be accepted and recognized as a right measuring rood for use in the field.*

Standard Error of the Mean and Sample Variance

Question. If the standard deviation of the size of shoes or feet in a population is σ, what is the standard deviation of 1/16 rood?

(1) σ
(2) $\sigma/2$
(3) $\sigma/4$
(4) $\sigma/16$

(In Japan, $\mu = 24.9$ cm and $\sigma = 1.05$ cm.)

The Method of Moments

General problem: How to find an estimator for a parameter of a distribution?

The ***method of moments*** was developed by Chebyshev and Pearson towards the end of the 19th century.

It is based on the fact that, given a random sample $X_1, \dots, X_n$ of a random variable X, for any integer $k \geq 1$,

$$\widehat{\mathrm{E}[X^k]} = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

is an unbiased estimator for the kth moment of X. (The proof is completely the same as for the sample mean.)

In other words, we have good estimators for the moments of a random variable.

The Method of Moments

The idea is now to express a parameter in terms of moments and then simply insert the estimators for these moments to obtain an estimator for the parameter.

Advantage: This is a simple method to obtain a basic estimator for a parameter.

Disadvantage: The estimators may not be unbiased and may yield non-sensical results in some cases.

For example, the variance of a random variable is $\operatorname{Var}[X] = \operatorname{E}[X^2] - \operatorname{E}[X]^2$, so we can set

$$\widehat{\sigma^2} = \widehat{\operatorname{E}[X^2]} - \widehat{\operatorname{E}[X]}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$$

$$= \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2$$

Estimator for the Variance

However, this estimator is not unbiased:

$$
\begin{aligned}
&\mathrm{E}\Big[\sum_{k=1}^{n}(X_k-\overline{X})^2\Big] \\
&= \mathrm{E}\Big[\sum_{k=1}^{n}(X_k-\mu+\mu-\overline{X})^2\Big] \\
&= \mathrm{E}\Big[\sum_{k=1}^{n}(X_k-\mu)^2-2(\overline{X}-\mu)\sum_{k=1}^{n}(X_k-\mu)+n(\mu-\overline{X})^2\Big] \\
&= \mathrm{E}\Big[\sum_{k=1}^{n}(X_k-\mu)^2-2(\overline{X}-\mu)\Big(\Big(\sum_{k=1}^{n}X_k\Big)-n\mu\Big)+n(\mu-\overline{X})^2\Big] \\
&= \mathrm{E}\Big[\sum_{k=1}^{n}(X_k-\mu)^2-2(\overline{X}-\mu)(n\overline{X}-n\mu)+n(\mu-\overline{X})^2\Big].
\end{aligned}
$$

Sample Variance

Simplifying, we have

$$
\begin{aligned}
\mathrm{E}\left[\sum_{k=1}^{n}(X_k - \overline{X})^2\right] &= \mathrm{E}\left[\sum_{k=1}^{n}(X_k - \mu)^2 - n(\overline{X} - \mu)^2\right] \\
&= \left(\sum_{k=1}^{n}\mathrm{E}\big[(X_k - \mu)^2\big] - n\,\mathrm{E}\big[(\overline{X} - \mu)^2\big]\right).
\end{aligned}
$$

We now use that

$$
\begin{aligned}
\mathrm{E}[(X_k - \mu)^2] &= \mathrm{Var}[X_k] = \sigma^2, \\
\mathrm{E}\big[(\overline{X} - \mu)^2\big] &= \mathrm{Var}[\overline{X}] = \sigma^2/n.
\end{aligned}
$$

Sample Variance

Then

$$\mathrm{E}\left[\sum_{k=1}^{n}(X_k - \overline{X})^2\right] = \left(\sum_{k=1}^{n}\sigma^2 - n\frac{\sigma^2}{n}\right) = (n-1)\sigma^2.$$

It follows that the estimator $\widehat{\sigma^2}$ obtained by the method of moments is biased:

$$\mathrm{E}\left[\frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2\right] = \frac{n-1}{n}\sigma^2,$$

therefore this estimator would tend to underestimate the true variance.

Instead, we will work with the unbiased ***sample variance***

$$S^2 := \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2.$$

Method of Maximum Likelihood

Fisher developed the following approach to finding estimators. Initial ideas go back to Gauß who used similar approaches on certyain problems.

Given a set of observations $x_1, \dots x_n$ from a random variable X, with parameter θ one finds the value of θ most likely to have produced these observations. This value becomes the estimate $\widehat{\theta}$.

In other words, we express the probability of obtaining $x_1, \dots x_n$ as a function of the parameter θ and then determine the value of θ that maximizes this probability.

Method of Maximum Likelihood

Let X_θ be a random variable with parameter θ and density f_{X_θ}. Given a random sample $(X_1, \dots, X_n)$ that yielded values $(x_1, \dots, x_n)$ we define the ***likelihood function*** L by

$$L(\theta) = \prod_{i=1}^{n} f_{X_\theta}(x_i).$$

If X_θ is a discrete random variable, then $L(\theta)$ is just the probability of obtaining the observed measurements:

$$P[X_1 = x_1 \text{ and } \dots \text{ and } X_n = x_n] = \prod_{i=1}^{n} P[X_i = x_i] = \prod_{i=1}^{n} f_{X_\theta}(x_i)$$

If X_θ is continuous, it represents the probability density.

We then maximize $L(\theta)$. The location of the maximum is then chosen to be the estimator $\widehat{\theta}$.

Estimating the Poisson Parameter

12.5. Example. Suppose it is known that X follows a Poisson distribution with parameter k and we wish to estimate k.

The density for X is given by $f_k(x) = \frac{e^{-k}k^x}{x!}$, $x \in \mathbb{N}$. Given a random sample $X_1, \dots, X_n$ the likelihood function is

$$L(k) = \prod_{i=1}^{n} f_k(x_i) = e^{-nk} \frac{k^{\sum x_i}}{\prod x_i!}.$$

To simplify our calculations, we take the logarithm:

$$\ln L(k) = -nk + \ln k \sum_{i=1}^{n} x_i - \ln \prod x_i!.$$

Maximizing $\ln L(k)$ will also maximize $L(k)$.

Estimating the Poisson Parameter

We take the first derivative and set it equal to zero:

$$\frac{d \ln L(k)}{dk} = -n + \frac{1}{k} \sum_{i=1}^{n} x_i = 0$$

so we find

$$\widehat{k} = \overline{x}.$$

This is not unexpected, since $\mu = k$ for the Poisson distribution and $\overline{X}$ is a good estimator for μ. In this case the maximum-likelihood estimator coincides with the method-of-moments estimator.

This method can be generalized to finding estimators for multiple parameters. In that case, the maximum of $L(\theta_1, \dots, \theta_j)$ is found with respect to all j variables.

Estimators with Mathematica

Both the method of moments and the method of maximum likelihood (as well as other methods) are available. Using the data from the previous section,

```
maxlike = FindDistributionParameters[Data,
  ExponentialDistribution[β],
  ParameterEstimator -> "MaximumLikelihood"]
```

$\{\beta \to 0.0087382\}$

```
mom = FindDistributionParameters[Data,
  ExponentialDistribution[β],
  ParameterEstimator -> "MethodOfMoments"]
```

$\{\beta \to 0.0087382\}$

Estimators with Mathematica