Chapter 6: Point Estimation

Curtis Miller

2018-06-13

Introduction

Karl Pearson, perhaps the first mathematical statistician, proposed the modern view that the objective of science is to estimate the parameters of a probability distribution that generates datasets [Salsburg, 2002]. Statistics has come a long way since Karl Pearson's methods, and in this chapter (where we finally leave our study of probability behind to dive into statistics) we see how to compute estimates for distribution parameters.

Initially there are many statistics competing to estimate some quantity; for example, both the sample mean and sample median could estimate the parameter μ of the Normal distribution. In the first section, we see general principles used to evaluate estimators. In the second section, we see methods for generating estimates.

Section 1: Some General Concepts of Point Estimation

There are many parameters we may try to estimate, such as

- μ from the distribution EXP(μ)
- μ and σ from the Normal distribution $N(\mu, \sigma)$
- α and β from the Weibull distribution WEI(α, β)
- And others

We want to discuss parameters and estimators using a general language. Let θ be a parameter, and $\hat{\theta}$ is an estimator for θ. Often the notation $\hat{\theta}$ refers to both a random variable and a specific point estimate.[1] We call $\hat{\theta}$ a **point estimator** for θ; we use the point estimator to compute a **point estimate**, a single plausible value for θ.

Examples of point estimators and the parameters they estimate include:

[1] I've said that usually capital letters refer to random variables; in this case, we would use $\hat{\Theta}$ to refer to the random version of the estimator, and $\hat{\theta}$ to refer to a specific number computed from an observed, no-longer-random dataset. However, this is not conventional; writers are lazy and don't like writing $\hat{\Theta}$, preferring $\hat{\theta}$ instead. Readers can usually tell whether the writer is referring to a random number or a computed number. As I said, capital letters *usually* refer to random variables; this is one of the (many) exceptions.

An estimator $\hat{\theta}$ is an **unbiased estimator** for θ if

---

Example 1

Show that the sample mean $\bar{X}$ computed from iid data is an unbiased estimator for the population mean μ.

Example 2

Suppose that $X_1, \ldots, X_n$ is an iid sample from a Bernoulli distribution with parameter p. Show that the sample proportion is an unbiased estimator for p.

Example 3

Show that the estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ computed from an iid sample $X_1, \ldots, X_n$ with $\text{Var}(X_1) = \sigma^2$ is *not* an unbiased estimator for σ^2.

Example 4

Show that the sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ computed from an iid sample $X_1, \ldots, X_n$ with $\text{Var}(X_1) = \sigma^2$ is an unbiased estimator for σ^2.[2]

[2] It's tempting to think that the sample standard deviation $S = \sqrt{S^2}$ is an unbiased estimator for σ, but this is *not* the case; S is a biased estimator for σ, with a tendency to underestimate the true σ. However, S is justified by other criteria. In fact, estimation of σ presents a good case study in why unbiasedness, as a criterion for good estimators, may be overrated (see Wikipedia [2018]).

Example 5

Suppose $X_1, \ldots, X_n$ is an iid sample from an exponential distribution with mean μ. Recall that the rate parameter of an exponential distribution is $\lambda = \frac{1}{\mu}$. Show that the estimator $\hat{\lambda} = \frac{1}{\bar{X}}$ is *not* an unbiased estimator for μ.

Suppose we want to estimate μ for Normally distributed data. $\bar{X}$ is an unbiased estimator for μ. So is $\tilde{X}$. In fact, X_1 is an unbiased estimator for μ since $\mathbb{E}[X_1] = \mu$. The last estimator is clearly silly, but not because of the unbiasedness criterion. Instead, the last estimator violates the **minimum variance** criterion, which states that the standard error (the standard deviation of $\hat{\theta}$, referred to as $\sigma_{\hat{\theta}}$) should be as small as possible, if not the smallest of all possible estimators. In this case, of the estimators I just mentioned, $\bar{X}$ has the smallest variance, and X_1 the largest. In fact, $\bar{X}$ is the **minimum variance unbiased estimator (MVUE)** for μ in this context, having the smallest variance of any unbiased estimator of μ. Likewise, $\hat{p}$ is the MVUE for p, when the data was drawn from a Bernoulli distribution with parameter p.

The minimum variance and unbiasedness criteria are not necessarily in agreement; there may be an estimator that has a smaller variance than all unbiased estimators and is close to the true value of θ when sample sizes are large. We may relax the unbiasedness criterion and instead require **consistency**, which says that a law of large numbers applies to the estimator; that is, $\hat{\theta}_n \to \theta$ in some sense as n grows (with $\hat{\theta}_n$ being an estimator for θ computed from n data points). The only estimator mentioned so far that isn't consistent is X_1; the rest (including the sample standard deviation) are consistent estimators.

Sometimes an estimator performs well in some circumstances but poorly in others; for example, $\bar{X}$ estimates the location of a distribution well when data is drawn from a Normal distribution but poorly when computed from data drawn from a distribution with heavy tails, such as the Laplace or Cauchy distributions. We call an estimator **robust** when the estimator performs well in multiple scenarios. Trimmed means, for example, as seen as robust estimators for the location of a distribution.

The **standard error** of an estimator is defined below:

The standard error can depend on unknown parameters. In that case, we may report an **estimated standard error**, where estimates for the unknown parameters are used in those parameters' place. Estimates of standard errors are often reported with point estimates to give a sense of how accurate the point estimate is. We will see how standard errors are often used to compute plausible regions for the location of θ in Chapter 7.

Example 6

Suppose $\text{Var}(X_1) = \sigma^2$ and the dataset $X_1, \ldots, X_n$ is an iid dataset. What is the standard error of $\bar{X}$? Use this to give estimates of standard errors for data drawn from Normal, exponential, and Poisson distributions.

Example 7

Suppose $X_1, \ldots, X_n$ is a Bernoulli dataset. What is the standard error of $\hat{p}$? What is an upper bound on the standard error? What is an estimate of the standard error?

Bootstrapping is a computer intensive technique for computing the standard errors of estimates. Bootstrap estimates of standard errors are often robust and allow us to obtain estimates when formulas for those errors would be intractable.

Suppose that $x_1, \ldots, x_n$ is a sample of iid data drawn from a distribution with pdf $f(x;\theta)$. The bootstrap procedure works as follows:

1. Estimate θ with $\hat{\theta}$.
2. Choose a large number B.
3. Generate B samples of data $X_{b1}, \ldots, X_{bn}$ $(1 \leq b \leq B)$ from the distribution with pdf $f(x;\hat{\theta})$, and from each of them compute $\hat{\theta}^*_b$, the estimate of θ using $x_{b1}, \ldots, x_{bn}$; you should now have a collection of data $\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B$
4. Compute $\bar{\theta}^* = \frac{1}{B}\sum_{b=1}^{B} \hat{\theta}^*_b$; this is the bootstrap estimate of θ
5. Compute $\hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^*_b - \bar{\theta}^*\right)^2}$; this is the bootstrap standard error estimate

---

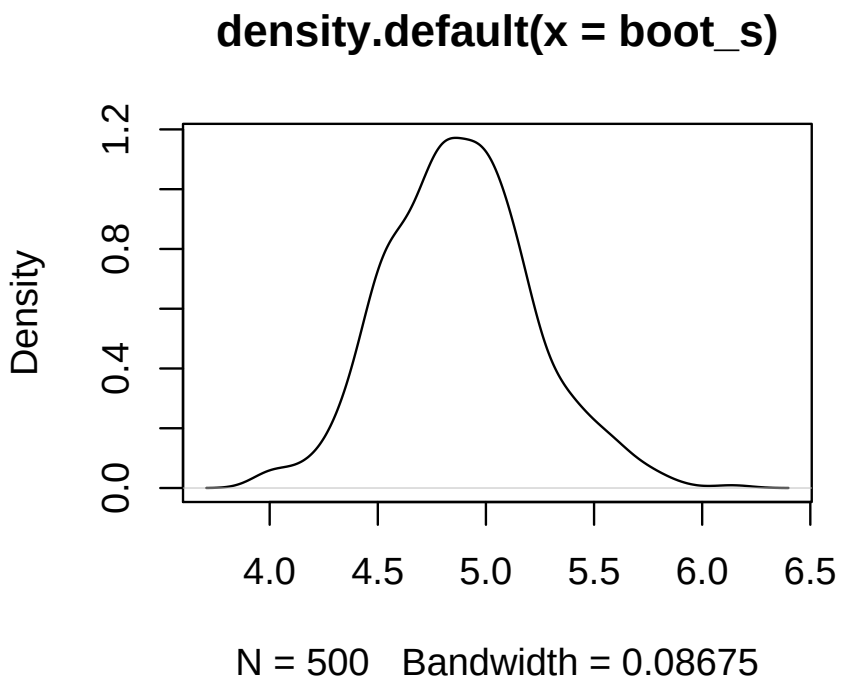
Example 8

In this example I demonstrate how to estimate the standard error of the estimate of the sample standard deviation computed from Normally distributed data.

```
n <- 100  # Our sample size is 100
B <- 500  # The bootstrap sample size is 500
dat <- rnorm(n, mean = 10, sd = 5)  # Our dataset
(s <- sd(dat))  # Estimated standard deviation
```

```
## [1] 4.891413
```

```
boot_s <- replicate(B, {
  boot_dat <- rnorm(n, mean(dat), s)
  sd(boot_dat)
})
plot(density(boot_s))
```




```r
mean(boot_s)  # The bootstrap estimator of s
```

```
## [1] 4.880043
```

```r
sd(boot_s)  # The bootstrap-estimated standard error of s
```

```
## [1] 0.3379874
```

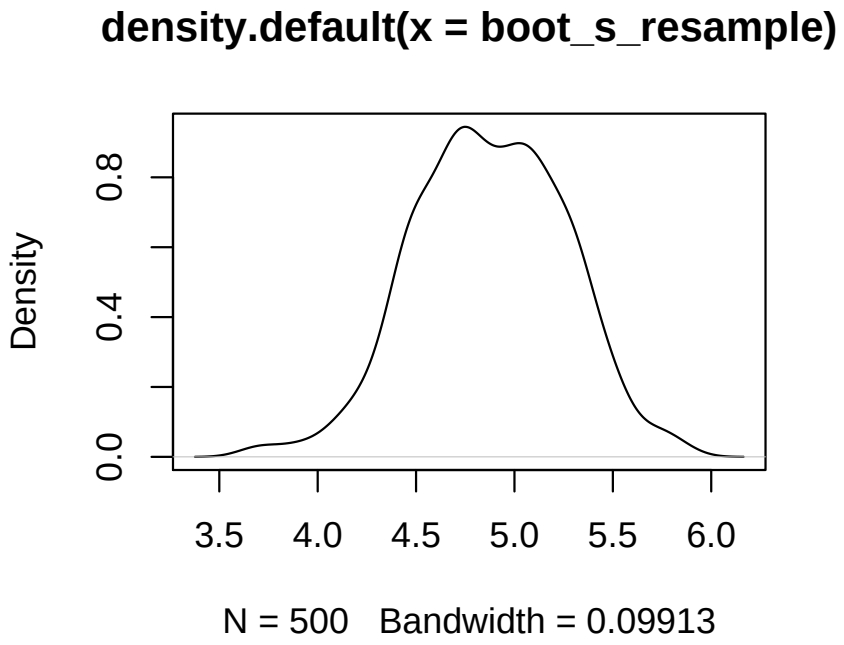
If we don't want to assume that the data came from a particular sample, we can sample instead from the data itself, doing so with replacement. When doing this, we are said to be sampling from the empirical cdf, or empirical distribution, of the data; that is, we are sampling from the distribution we observed, which serves as an estimate of the population distribution that generated the data.

Example 9

This example demonstrates obtaining a bootstrap estimate of the standard error of the standard deviation without assuming that the data was drawn from a particular distribution, using the resampling technique.

```r
boot_s_resample <- replicate(B, {
  boot_dat <- sample(dat, n, replace = TRUE)
```

```
  sd(boot_dat)
})
plot(density(boot_s_resample))
```




```
mean(boot_s_resample)
```

```
## [1] 4.873325
```

```
sd(boot_s_resample)
```

```
## [1] 0.3817121
```

Section 2: Methods of Point Estimation

Assume again that $X_1, \ldots, X_n$ is an iid sample from some distribution. $\mathbb{E}\left[X_1^k\right]$ is called the k^{th} **population moment**, and $\frac{1}{n}\sum_{i=1}^n X_i^k$ is called the k^{th} **sample moment**. As an example, $\bar{X}$ is the first sample moment and $\mathbb{E}[X_1] = \mu$ is the first population moment.[3] A sample moment is an unbiased estimator for the corresponding population moment.

[3] σ^2 is related to the second sample moment but isn't the second moment itself. The same goes for S^2 and sample moments.

Suppose a distribution has $\theta_1, \ldots, \theta_K$ parameters we wish to estimate. Below is the **method of moments estimation** procedure:

1. Compute the first K population moments, $m_1, \ldots, m_K$ in terms of the unknown parameters $\theta_1, \ldots, \theta_K$
2. Solve for $\theta_1, \ldots, \theta_K$ so they are expressed in terms of $m_1, \ldots, m_K$

3. Replace $m_1, \ldots, m_K$ with $M_1, \ldots, M_K$, the first K sample moments; the resulting expressions are $\hat{\theta}_1, \ldots, \hat{\theta}_K$, the **method of moments estimators (MMEs)** for the desired parameters.

Method of moments estimation produces consistent estimators for desired parameters using an intuitive procedure. There is no guarantee the estimators are unbiased (in fact they likely are not unbiased) and they usually are not minimum-variance estimators. In fact, in the context the estimators were computed, there likely is an estimator that is consistent and with a smaller variance than the MMEs That said, method of moment estimators are often robust and more tractable than other estimators while being easy to compute.[4]

[4] Method of moments estimation is often used in economics due to their simplicity and robustness.

Example 10

What is the method of moments estimator for the population variance?

Example 11

What is the MME for the rate parameter $\lambda = \frac{1}{\mu}$ for an exponential distribution?

Example 12

Let $X_1, \ldots, X_n$ be an iid sample from the distribution UNIF$(a-b, a+b)$. What are the MMEs for a and b?

Example 13

Consider a shifted exponential distribution that depends on two parameters μ and γ such that $X_1 - \gamma \sim \text{EXP}(\mu)$. What are the MMEs for μ and γ?

To illustrate the principle of the next estimation method, suppose I flip a coin and record whether I get heads or not. The coin could be a fair coin or a biased coin, where the probability of getting heads is $p = .9$. When I flip the coin and observe an outcome, how will I decide which coin was flipped?

Consider the following table:

After flipping the coin and observing the outcome, I look to the table to see what the probability of that outcome was under each scenario of coin choice. The maximum likelihood principle says that I should choose the coin that maximizes these probabilities.

Let $X_1, \ldots, X_n$ have the joint pmf/pdf $f(x_1, \ldots, x_n; \theta_1, \ldots, \theta_K)$. When $x_1, \ldots, x_n$ are the observed values of the dataset, this function is called the **likelihood function** when it is regarded as a function of $\theta_1, \ldots, \theta_K$, as expressed below:

When the random variables $X_1, \ldots, X_n$ are iid, the likelihood function is

The **maximum likelihood estimators (MLEs)** $\hat{\theta}_1, \ldots, \hat{\theta}_K$ are the values that maximize the likelihood function. They are interpreted as the most likely values of the parameters given the data we saw, in that we were most likely to see the values of the data if those were the parameters.

Usually the likelihood function is hard to maximize on its own, so instead we maximize the log-likelihood function

Since $\ln(x)$ is an increasing function, both functions have the same

maxima.

Example 14

Consider an iid dataset of Bernoulli data. What is the maximum likelihood estimator of the sample proportion p?

Example 15

Consider an iid dataset drawn from the EXP(μ) distribution. Find the MLE for μ.

Example 17

Consider an iid dataset drawn from the $N(\mu, \sigma^2)$ distribution. Find the MLE of μ and σ^2.

Example 18

Consider an iid dataset drawn from the UNIF$(0, \theta)$ distribution. Find the MLE of θ.

MLEs are consistent estimators and are either minimum variance or almost minimum variance, with these properties improving as the sample size grows large. Additionally, the MLE of a function of parameters $h(\theta_1, \ldots, \theta_K)$ is the value of that function when applied to the MLEs $h(\hat{\theta}_1, \ldots, \hat{\theta}_K)$.

Example 19

Expanding on Example 15, find the MLE of the rate parameter $\lambda = \frac{1}{\mu}$ of an exponential distribution.

Example 20

Expanding on Example 20, find the MLE of the standard deviation σ of a Normal distribution.

Maximum likelihood estimation is an example of a general approach to parameter estimation, where a "good" estimate is an estimate that optimizes some objective function. MLEs maximize the likelihood function. Least-squares estimators minimizes the sum of square errors, $\sum_{i=1}^{n} \left(x_i - \hat{x}_i(\hat{\theta}_1, \ldots, \hat{\theta}_K)\right)^2$ (with $\hat{x}_i(\hat{\theta}_1, \ldots, \hat{\theta}_K)$ being the predicted value of x_i based on the parameter estimates), and least absolute deviation estimators minimize $\sum_{i=1}^{n} \left|x_i - \hat{x}_i(\hat{\theta}_1, \ldots, \hat{\theta}_K)\right|$. M-estimators maximize $\sum_{i=1}^{n} \rho(x_i; \hat{\theta}_1, \ldots, \hat{\theta}_K)$, where the "objective function" ρ is chosen to give the resulting estimator desired robustness properties.

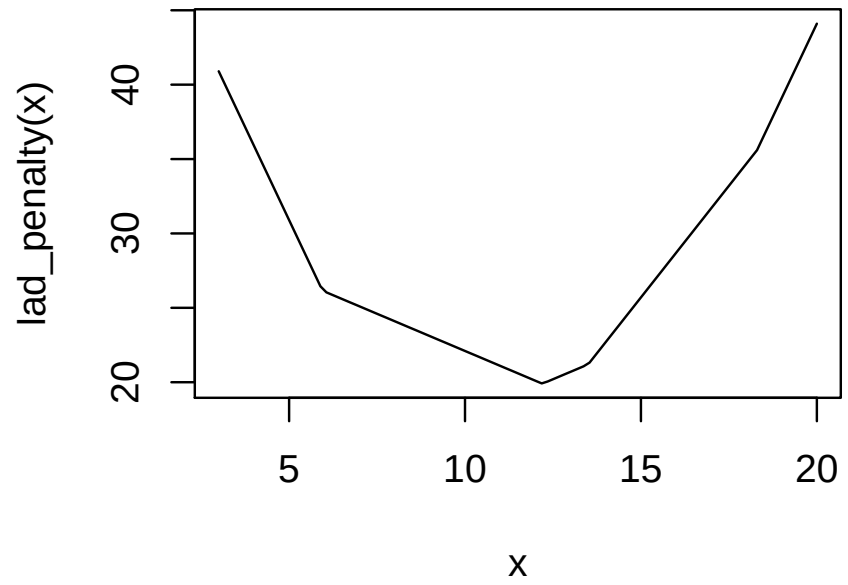
Example 21

Consider the following dataset:

```
x <- c(12.2, 18.3, 6.0, 5.9, 13.5)
```

The predicted value $\hat{\mu}$ of the data is the least absolute deviation estimator. Find the value of the estimator.

```
lad_penalty <- function(mu) {sum(abs(x - mu))}
lad_penalty <- Vectorize(lad_penalty)
curve(lad_penalty(x), 3, 20)
```



```
optim(0, lad_penalty)
```

```
## Warning in optim(0, lad_penalty): one-dimensional optimization by Nelder-Mead is unreliable:
## use "Brent" or optimize() directly
```

```
## $par
## [1] 12.3
##
## $value
## [1] 20
##
## $counts
## function gradient
##       24       NA
##
## $convergence
## [1] 0
##
## $message
## NULL
```

```
median(x)
```

```
## [1] 12.2
```

References

D. Salsburg. *The Lady Tasting Tea: How Statistics Revolutionized Science in the Twentieth Century*. Henry Holt and Company, 2002. ISBN 9780805071344. URL https://books.google.com/books?id=klbYy-C72ksC.

Wikipedia. Unbiased estimation of standard deviation — Wikipedia, the free encyclopedia, 2018. URL https://en.wikipedia.org/w/index.php?title=Unbiased_estimation_of_standard_deviation&oldid=823365997. [Online; accessed 12-June-2018].