# Statistics - Lecture One

Charlotte Wickham
wickham@stat.berkeley.edu
http://www.stat.berkeley.edu/~wickham/

## Outline

1. Basic ideas about estimation
2. Method of Moments
3. Maximum Likelihood
4. Confidence Intervals
5. Introduction to the Bootstrap

These notes follow Rice [2007] very closely.

# Introduction

Our aim is to learn about a random process by observing a sample of outcomes. Generally, we have a sample $X_1, \ldots, X_n$ drawn at random and we want to learn about their underlying distribution. For example, we might know that the distribution is normal and we want to estimate the parameters for the mean and variance.

We can divide the type of models we might consider into two classes: parametric models and non-parametric models. A parametric model can de defined up to a finite dimensional parameter. Otherwise, it is considered a non-parametric model. For example, whenever we assume the random variable follows a particular probability distribution up to an unknown parameter we are considering a parametric model. In a nonparametric model the number and nature of the parameters is flexible and not fixed in advance. Here we will concentrate on estimation in the parametric case.

# 1 Estimation

Given a model that is defined up to a finite set of parameters and a random sample from the model $X_1, \ldots, X_n$ we wish to come up with a function of the sample that will estimate the parameters we don't know. Once we have decided on this estimator, $\theta_n = t(X_1, ..., X_n)$, it is only useful for inference if we know something about how it behaves. Since $\theta_n$ is a function of a random sample it itself is a random variable and its behavior will depend on the size of the sample, $n$.

Consider, for example, estimating the population mean of a normally distributed population (for illustrative purposes say $\mathcal{N}(10, 9)$). The most obvious estimate is to simply draw a sample and calculate the sample mean. If we repeat this process with a new sample we would expect to get a different estimate. The distribution that results from repeated sampling is called the sampling distribution of the estimate. Figure 1 illustrates 500 estimates of the population mean based on a sample of size 100. We can see that our estimates are generally centered around the true value of 10

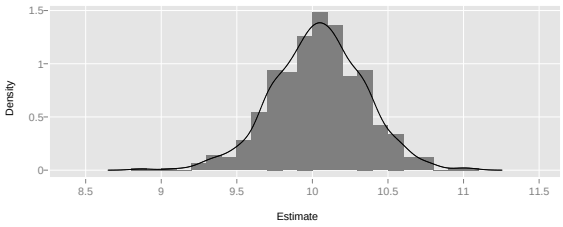


Figure 1: 500 estimates of the population mean based on a sample size 100

but there is some variation - maybe a standard deviation of about 1. These observations translate to more formal ideas: the expectation of the estimator and the standard error of the estimator. Of course, it is even nicer if we can also parameterize the sampling distribution.

So, what do we look for in a good estimator? Obviously, we want our estimate to be close to the true value. Also we want $\theta_n$ to behave in a nice way as the sample size, $n$, increases. If we take a large sample we would like the estimate to be more accurate than a small sample. If we go back to our previous example we can look at the difference in behavior if we use different sample sizes. Each histogram in figure 2 represents 500 estimates of the population mean for sample sizes of 5, 10, 25, 50 and 100. We can see that the standard deviation of the estimate is smaller as the sample size increases. Formally, this is embodied in the principle of **consistency**. A consistent estimator will converge to the true parameter value as the sample size increases. Our estimate $\bar{X}$

for the population mean of a normal seems very well behaved. Now, for some examples of not so well behaved estimators.

### Consistent but biased estimator

Here we estimate the variance of the normal distribution used above (i.e. the true variance is 9). We use the estimate,

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\,,$$

which happens to be the maximum likelihood estimate (to be discussed later). Histograms for 500 estimates based on different sample sizes are shown in Figure 3. For small sample sizes we can

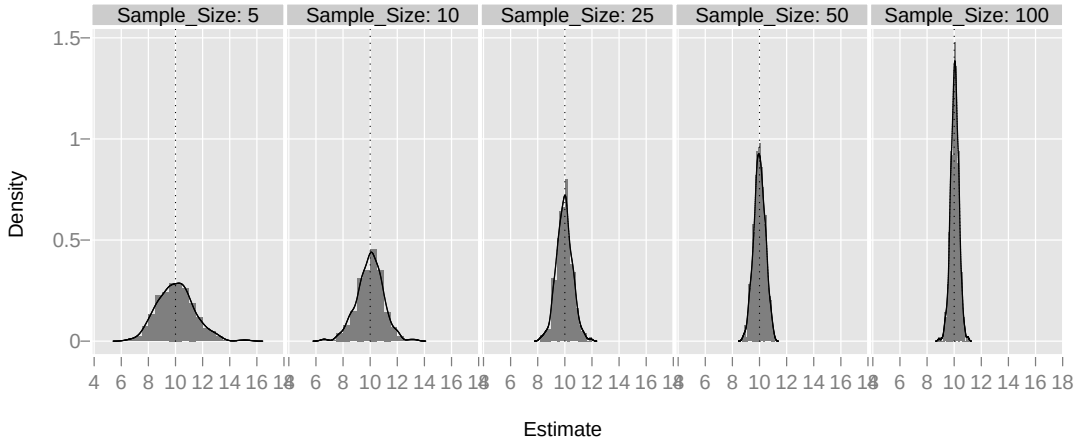


Figure 2: 500 estimates of the population mean based on sample sizes of 5, 10, 25, 50 and 100

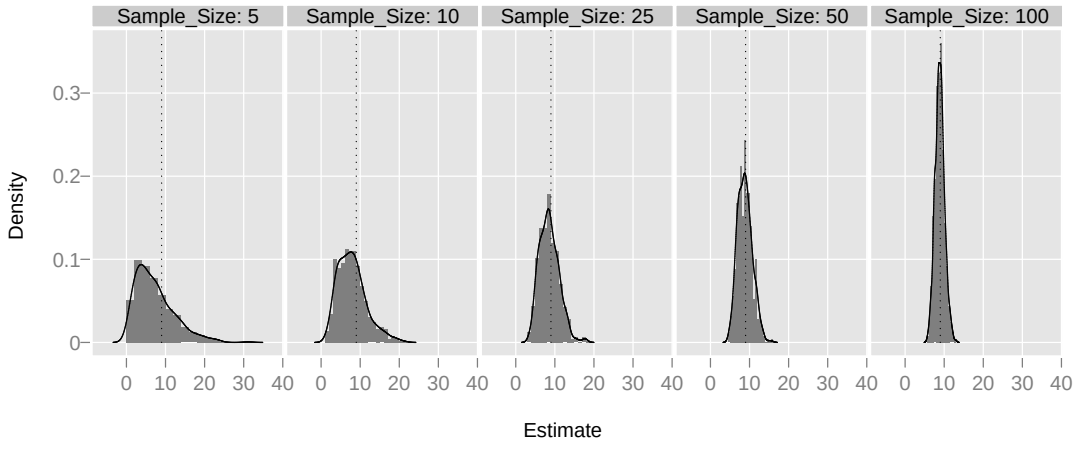


Figure 3: 500 maximum likelihood estimates of the population variance based on sample sizes of 5, 10, 25, 50 and 100

see that the estimates are not centered around 9. This is an example of a biased estimator. As the sample size increases the estimates move closer to being centered around 9. The estimate is said to be asymptotically unbiased. As sample size increases the estimates have smaller variance. This estimator is consistent.

### Inconsistent estimator

It is very easy to come up with inconsistent estimators. After all, any function can be called an estimator even if it clearly will not have nice properties. For example, we could use the sample median to estimate the population mean. If the underlying distribution is antisymmetric then this will clearly be a poor estimator. This is illustrated in Figure 4 where the underlying distribution is exponential with mean 1. We can see that although the estimates have decreasing variance they

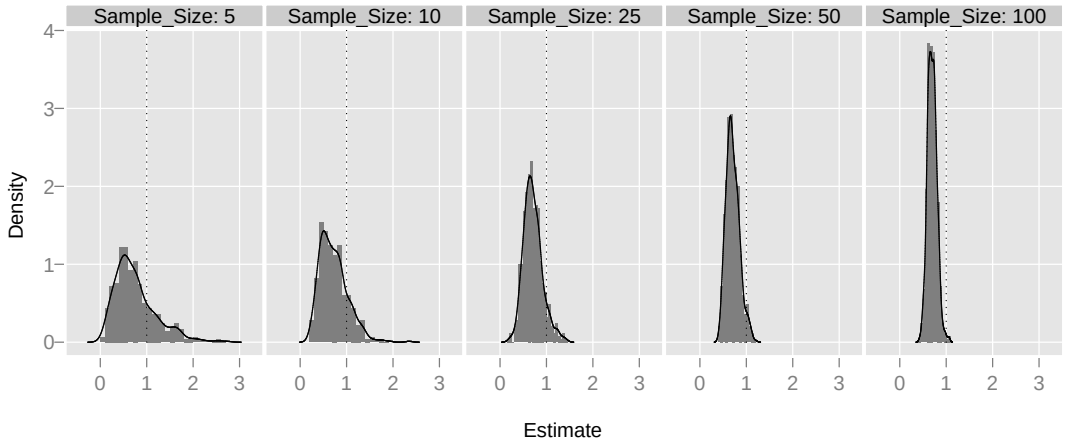


Figure 4: 500 estimates of the population mean of an exponential distribution using the sample median. Based on sample sizes of 5, 10, 25, 50 and 100

are not getting closer to the true value of 1.

## Formal Definitions

### Consistency

Let $\theta_n$ be an estimate of $\theta$ based on a sample of size $n$. $\theta_n$ is **consistent** in probability if $\theta_n$ converges in probability to $\theta$; that is for any $\epsilon < 0$

$$P(|\theta_n - \theta| > 0) \to 0 \text{ as } n \to \infty$$

### Bias

$\theta_n$ is **unbiased** if,

$$E(\theta_n) = \theta.$$

### Efficiency

An estimator is efficient if it has the lowest possible variance among all unbiased estimators

### Mean Squared Error

The quality of an estimator is generally judged by its mean squared error (MSE),

$$\text{MSE} = \text{variance} + \text{bias}^2.$$

## 2 Method of Moments

Let $X$ be random variable following some distribution. Then the $k$**th moment** of the distribution is defined as,

$$\mu_k = E(X^k).$$

For example $\mu_1 = E(X)$ and $\mu_2 = Var(X) + E(X)^2$.

The **sample moments** of observations $X_1, X_2, \ldots, X_n$ independent and identically distributed (iid) from some distribution are defined as,

$$\hat{\mu_k} = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

For example, $\hat{\mu_1} = \bar{X}$ is the familiar sample mean and $\hat{\mu_2} = \hat{\sigma}^2 + \bar{X}^2$ where $\hat{\sigma}$ is the standard deviation of the sample.

The method of moments estimator simply equates the moments of the distribution with the sample moments ($\mu_k = \hat{\mu_k}$) and solves for the unknown parameters. Note that this implies the distribution must have finite moments.

### Example - Poisson

Assume $X_1, \ldots, X_n$ are drawn iid from a Poisson distribution with mass function,

$$P(X = x) = \lambda^x e^{-\lambda}/x!, \quad x = 0, 1, 2, \ldots,$$

where $\lambda$ is an unknown parameter. Check that $E(X) = \lambda$. So, $\mu_1 = E(X) = \lambda = \bar{X} = \hat{\mu_1}$. Hence, the method of moments estimator of $\lambda$ is the sample mean.

### Example - Gamma

Assume $X_1, \ldots, X_n$ are drawn iid from a Gamma distribution with density,

$$f(x|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x \geq 0,$$

where $\lambda$ and $\alpha$ are unknown parameters. The first two moments of the gamma distribution are (check this yourself),

$$\mu_1 = \frac{\alpha}{\lambda}$$
$$\mu_2 = \frac{\alpha(\alpha+1)}{\lambda^2}.$$

From the first equation,

$$\lambda = \frac{\alpha}{\mu_1}.$$

Substituting this into the second equation gives,

$$\frac{\mu_2}{\mu_1^2} = \frac{\alpha+1}{\alpha},$$

or

$$\alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2}.$$

Then we have

$$\begin{aligned}\lambda &= \frac{\mu_1^2}{\mu_2 - \mu_1^2}\frac{1}{\mu_1} \\ &= \frac{\mu_1}{\mu_2 - \mu_1^2}.\end{aligned}$$

We substitute in the sample analogs of the moments and find the method of moments estimators are,

$$\begin{aligned}\hat{\lambda} &= \frac{\bar{X}}{\hat{\sigma^2}} \\ \hat{\alpha} &= \frac{\bar{X}^2}{\hat{\sigma^2}}.\end{aligned}$$

### Properties of the method of moments estimator

Nice properties:

- it is consistent
- sometimes easier to calculate than maximum likelihood estimates

Not so nice properties:

- sometimes not sufficient
  Sufficiency has a formal definition but intuitively it means that all the data that is relevant to estimating the parameter of interest is used.
- sometimes gives estimates outside the parameter space

## 3 Maximum Likelihood

Let $X_1, X_2, \ldots, X_n$ be a random vector of observations with joint density function $f(x_1, ...x_n|\theta)$. Then the **likelihood** of $\theta$ as a function of the observed values, $X_i = x_i$, is defined as,

$$lik(\theta) = f(x_1, \ldots, x_n|\theta).$$

The **maximum likelihood estimate** (mle) of the parameter $\theta$ is the value of $\theta$ that maximizes the likelihood function. In general, it is easier to maximize the natural log of the likelihood. In the case that the $X_i$ are iid the log likelihood is generally of the form,

$$l(\theta) = \log f(x|\theta) = \sum_{i=1}^{n} f(x_i|\theta).$$

**Example - Poisson**

Assume the same conditions as the example in the previous section. So, the log likelihood is,

$$l(\lambda) = \sum_{i=1}^{n} (x_i \log \lambda - \lambda - \log x_i!)$$

To find the maximum we set the first derivative to zero,

$$\begin{aligned} l'(\lambda) &= \sum_{i=1}^{n} \left( \frac{x_i}{\lambda} - 1 \right) \\ &= \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n = 0 \end{aligned}$$

Solving for $\lambda$ we find that the mle is,

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{X}$$

Note that this agrees with the method of moments estimator.

**Example - Gamma**

Again, assume the same conditions as the Gamma example in the previous section. The log likelihood is,

$$l(\alpha, \lambda) = \sum_{i=1}^{n} (-\log \Gamma(\alpha) + \alpha \log \lambda + (\alpha - 1) \log x_i - \lambda x_i)$$

In this case we have two parameters so we take the partial derivatives and set them both to zero.

$$\begin{aligned} \frac{\partial l}{\partial \alpha} &= \sum_{i=1}^{n} \left( -\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \log \lambda + \log x_i \right) \\ &= \sum_{i=1}^{n} \log x_i + n \log \lambda - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0 \\ \frac{\partial l}{\partial \lambda} &= \sum_{i=1}^{n} \left( \frac{\alpha}{\lambda} - x_i \right) \\ &= \frac{n\alpha}{\lambda} - \sum_{i=1}^{n} x_i = 0 \end{aligned}$$

This second equality gives the mle for $\lambda$ as,

$$\hat{\lambda} = \frac{\hat{\alpha}}{\bar{X}}.$$

Substituting this into the first equation we find that the mle for $\alpha$ must satisfy,

$$n \log \hat{\alpha} - n \log \bar{X} + \sum_{i=1}^{n} \log x_i - n \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})} = 0.$$

This equation needs to be solved by numerical means. Note however this is a different estimate to that given by the method of moments.

### Properties of maximum likelihood estimation

Nice properties:

- consistent
- very widely applicable
- unaffected by monotonic transformations of the data
- MLE of a function of the parameters, is that function of the MLE
- theory provides large sample properties
- asymptotically efficient estimators

Not so nice properties:

- may be slightly biased
- parametric model required and must adequately describe statistical process generating data
- can be computationally demanding
- fails in some case, eg. if too many nuisance parameters

(More on the asymptotic properties of the mle in the second Stats lecture).

## 4 Confidence Intervals

Confidence intervals are an intuitive way to present information on an estimate and its uncertainty. Formally an $100(1-\alpha)\%$ confidence interval says that the interval is expected to contain the true value of the parameter $100(1-\alpha)\%$ of the time. Since a confidence interval is calculated from observations of a random variable it is itself a random variable. Each time a sample is taken (or an experiment performed) we would expect to get a different interval. If we calculate 95% confidence interval each time we would expect about 1 in 20 of the intervals to miss the true parameter. Figure 5 illustrates this. Fifty samples of size thirty are drawn at random from a population with mean 15. For each sample a 95% confidence interval for the population mean is calculated. These fifty intervals are plotted as vertical lines. We can see 4 intervals missed the true value.

#### Example - Population mean

Say we want an $100(1-\alpha)\%$ confidence interval for the population mean, $\mu$. By Central Limit Theorem $\sqrt{n}(\bar{X}-\mu)/\sigma$ is distributed $\mathcal{N}(0,1)$. Let $z(\alpha)$ be the value such that the area to the right of $z(\alpha)$ is $\alpha$. By symmetry of the normal $z(1-\alpha) = -z(\alpha)$. So,

$$P\left(-z(\alpha/2) \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z(\alpha)\right) = 1-\alpha$$

Rearranging this we find

$$P\left(\bar{X} - z(\alpha/2)\sigma/\sqrt{n} \leq \mu \leq \bar{X} + \sqrt{n}z(\alpha/2)\sigma/\sqrt{n}\right) = 1-\alpha$$

So a $(1-\alpha)\%$ confidence interval for $\mu$ is $(\bar{X} - z(\alpha/2)\sigma/\sqrt{n}, \bar{X} + z(\alpha/2)\sigma/\sqrt{n})$. Of course in practice we don't know the value of $\sigma$ so an estimate is generally plugged in instead.

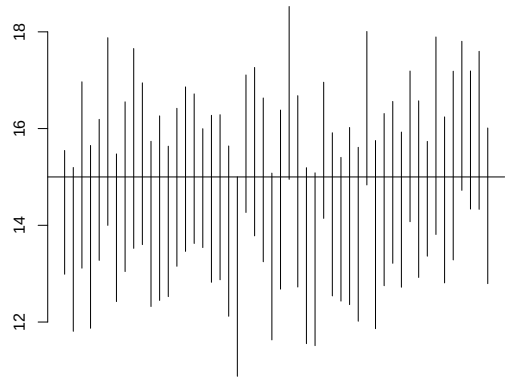


Figure 5: Fifty 95% confidence intervals for the population mean calculated from samples of size thirty. The horizontal line is the true population mean.

# 5 Bootstrap

What if we don't know the properties of our estimator? The idea of the bootstrap is to use simulation on the data we have to give us an estimate of the sampling distribution of the estimator.

## General Idea

Repeatedly resample the data with replacement and calculate the estimate each time. The distribution of these bootstrap estimates approximates the sampling distribution of the estimate.

## Details

Assume we have a sample $X_1, \ldots, X_n$ that have some known distribution up to parameter $\theta$. Also we have an estimator, $\hat{\theta_n} = t(X_1, \ldots, X_n)$, of $\theta$. Now, we resample from the data **with** replacement. We have a new sample $X_1^*, \ldots, X_n^*$ of size $n$ that could contain any of the original sample datapoints any number of times. This resampled data is considered one bootstrap replicate. We calculate,

$$\hat{\theta_n^*} = t(X_1^*, \ldots, X_n^*) = \text{ an estimate of } \theta \text{ from a bootstrap replicate.}$$

The theory behind the bootstrap says that the distribution of $(\hat{\theta} - \theta)$ can be approximated by the distribution of $(\hat{\theta^*} - \hat{\theta})$. This method is probably best explained through an example. Below we use the bootstrap to investigate the sampling distribution of the method of moments estimator for the parameters in a gamma distribution.

### Example - Method of moments for gamma

We have a sample of size 30 from a Gamma distribution with unknown parameters $\lambda$ and $\alpha$. The data is as follows.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.038 | 0.009 | 1.190 | 0.312 | 0.015 | 0.036 | 0.865 | 0.425 | 0.269 | 0.098 |
| 0.498 | 0.511 | 0.000 | 0.159 | 0.437 | 1.141 | 0.002 | 0.111 | 0.013 | 0.150 |
| 0.076 | 0.786 | 0.985 | 0.344 | 0.668 | 0.018 | 0.344 | 0.085 | 0.460 | 0.622 |
| 0.082 | 0.944 | 0.028 | 0.035 | 2.018 | 2.293 | 0.123 | 2.604 | 0.017 | 0.014 |
| 0.108 | 0.442 | 3.316 | 0.679 | 0.407 | 0.120 | 0.654 | 0.338 | 1.434 | 0.508 |

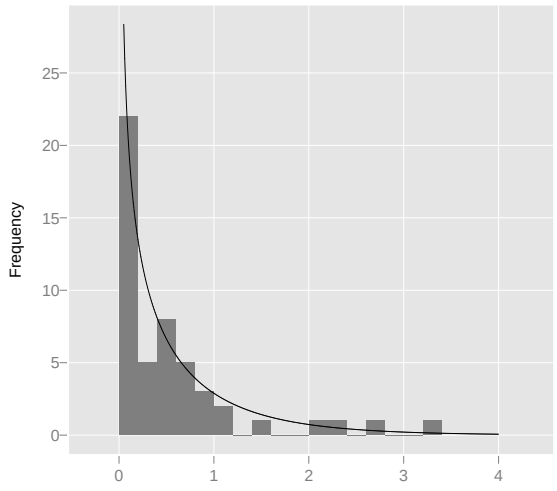


Figure 6: Histogram of the observations along with density of the estimated gamma distribution.

The sample statistics are $\bar{X} = 0.537$ and $\hat{\sigma^2} = 0.496$. The method of moments estimators are,

$$\hat{\lambda} = \frac{\bar{X}}{\hat{\sigma^2}} = 1.082$$
$$\hat{\alpha} = \frac{\bar{X}^2}{\hat{\sigma^2}} = 0.581.$$

Figure 6 shows a histogram of the observed data along with the fitted density using the method of moments estimator.

Now to the bootstrap. We take 1000 new samples from the original data. For example, the first sample is (sorted), This has a sample mean of $\bar{X}^* = 0.66$ and sample variance of $\hat{\sigma^2}^* = 0.801$.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0.002 | 0.014 | 0.028 | 0.038 | 0.108 | 0.120 | 0.338 | 0.460 | 0.668 | 0.944 |
| 0.013 | 0.015 | 0.035 | 0.076 | 0.108 | 0.123 | 0.344 | 0.460 | 0.668 | 0.985 |
| 0.013 | 0.015 | 0.035 | 0.082 | 0.108 | 0.123 | 0.407 | 0.654 | 0.679 | 1.434 |
| 0.013 | 0.017 | 0.036 | 0.098 | 0.108 | 0.159 | 0.442 | 0.654 | 0.786 | 1.434 |
| 0.014 | 0.018 | 0.038 | 0.108 | 0.111 | 0.269 | 0.460 | 0.668 | 0.944 | 1.434 |

This gives us method of moment estimators of $\hat{\lambda}^* = 0.825$ and $\hat{\alpha}^* = 0.545$. We repeat this process for the remaining 999 samples and plot the estimates in a histogram (figure 7). We can use this

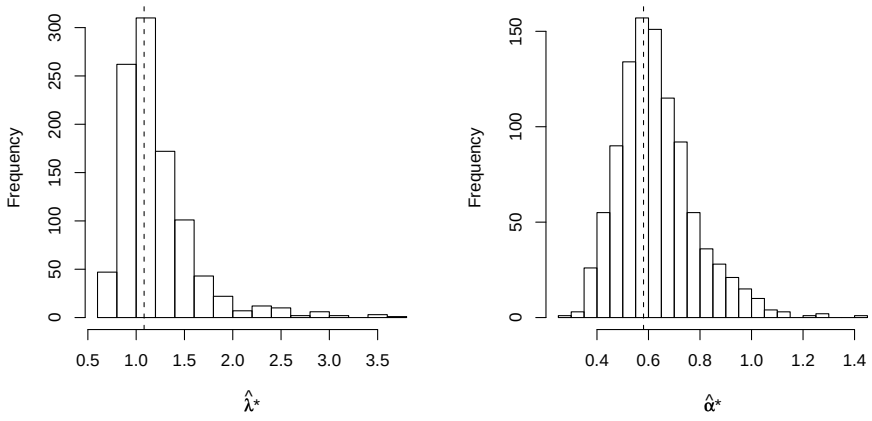


Figure 7: Histograms for the estimates of $\lambda$ and $\alpha$ from 1000 bootstrap replicates. The dotted lines correspond to the estimates from the original sample

histogram as an approximation to the sampling distribution. We can see that the distribution for both parameters is right skewed. We can estimate the bias in the estimates by comparing the average of the bootstrap estimates with the initial estimate.

$$\hat{\lambda}^*_{\text{ave}} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\lambda}^*_i = 1.208$$

$$\hat{\alpha}^*_{\text{ave}} = \frac{1}{1000}\sum_{i=1}^{1000}\hat{\alpha}^*_i = 0.631$$

Comparing these to the initial estimate we can see that the method of moments estimator might be slightly positively biased. We can also use the bootstrap replicates to estimate the standard error of the estimates. The estimates for the sampling standard errors are simply the standard deviation of the bootstrap estimates.

$$\widehat{\text{se}_\lambda} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\lambda}^*_i - \hat{\lambda}^*_{\text{ave}}\right)^2} = 0.374$$

$$\widehat{\text{se}_\alpha} = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\hat{\alpha}^*_i - \hat{\alpha}^*_{\text{ave}}\right)^2} = 0.15$$

## Approximate Confidence Intervals

We can also use the bootstrap to find approximate confidence intervals. The bootstrap procedure is the same. Once the bootstrap distribution is obtained we can simply take the 2.5 and 97.5 percentiles as the upper and lower values of the confidence interval.

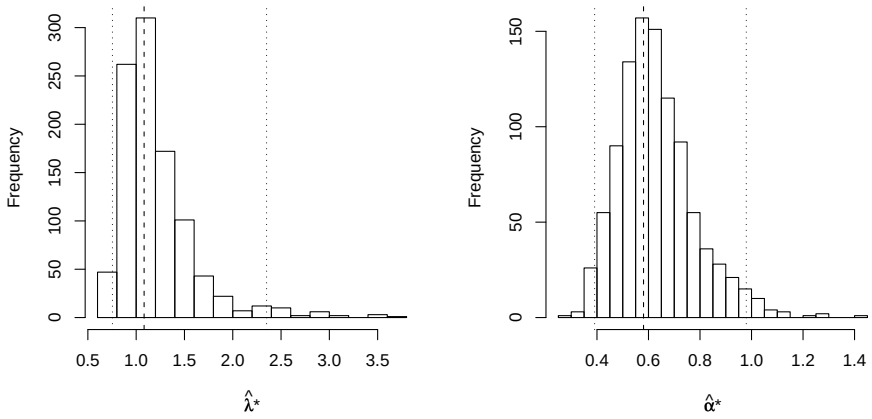


Figure 8: Histograms for the estimates of $\lambda$ and $\alpha$ from 1000 bootstrap replicates. The dotted lines correspond to the estimates from the original sample

**Example - Gamma**

We find a 95% confidence interval for each parameter simply by finding the corresponding quantiles from the approximate sampling distributions found previously. These are plotted as dotted lines in Figure 8. The intervals are,

$$\alpha : (0.404, 0.982)$$
$$\lambda : (0.770, 2.195)$$

In words we would say: we estimate the true value of $\alpha$ to be between 0.404 and 0.982 with 95% confidence. Similarly for $\lambda$.

# References

John A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, Belmont, CA, third edition, 2007.