Topic 13

Method of Moments

13.1 Introduction

Method of moments estimation is based solely on the law of large numbers, which we repeat here:

Let $M_1, M_2, \ldots$ be independent random variables having a common distribution possessing a mean μ_M. Then the sample means converge to the distributional mean as the number of observations increase.

$$\bar{M}_n = \frac{1}{n}\sum_{i=1}^{n} M_i \to \mu_M \quad \text{as } n \to \infty.$$

To show how the method of moments determines an estimator, we first consider the case of one parameter. We start with independent random variables $X_1, X_2, \ldots$ chosen according to the probability density $f_X(x|\theta)$ associated to an unknown parameter value θ. The common mean of the X_i, μ_X, is a function $k(\theta)$ of θ. For example, if the X_i are continuous random variables, then

$$\mu_X = \int_{-\infty}^{\infty} x f_X(x|\theta)\, dx = k(\theta).$$

The law of large numbers states that

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mu_X \quad \text{as } n \to \infty.$$

Thus, if the number of observations n is large, the distributional mean, $\mu = k(\theta)$, should be well approximated by the sample mean, i.e.,

$$\bar{X} \approx k(\theta).$$

This can be turned into an estimator $\hat{\theta}$ by setting

$$\bar{X} = k(\hat{\theta}).$$

and solving for $\hat{\theta}$.

We shall next describe the procedure in the case of a vector of parameters and then give several examples. We shall see that the delta method can be used to estimate the variance of method of moment estimators.

13.2 The Procedure

More generally, for independent random variables $X_1, X_2, \ldots$ chosen according to the probability distribution derived from the parameter value θ and m a real valued function, if $k(\theta) = E_\theta m(X_1)$, then

$$\frac{1}{n}\sum_{i=1}^{n} m(X_i) \to k(\theta) \quad \text{as } n \to \infty.$$

The **method of moments** results from the choices $m(x) = x^m$. Write

$$\mu_m = EX^m = k_m(\theta). \tag{13.1}$$

for the m-th moment.

Our estimation procedure follows from these 4 steps to link the sample moments to parameter estimates.

- **Step 1**. If the model has d parameters, we compute the functions k_m in equation (13.1) for the first d moments,

  $$\mu_1 = k_1(\theta_1, \theta_2 \ldots, \theta_d), \quad \mu_2 = k_2(\theta_1, \theta_2 \ldots, \theta_d), \quad \ldots, \quad \mu_d = k_d(\theta_1, \theta_2 \ldots, \theta_d),$$

  obtaining d equations in d unknowns.

- **Step 2**. We then solve for the d parameters as a function of the moments.

  $$\theta_1 = g_1(\mu_1, \mu_2, \cdots, \mu_d), \quad \theta_2 = g_2(\mu_1, \mu_2, \cdots, \mu_d), \quad \ldots, \quad \theta_d = g_d(\mu_1, \mu_2, \cdots, \mu_d). \tag{13.2}$$

- **Step 3.** Now, based on the data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, we compute the first d **sample moments**,

  $$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2, \quad \ldots, \quad \overline{x^d} = \frac{1}{n}\sum_{i=1}^{n} x_i^d.$$

  Using the law of large numbers, we have, for each moment, $m = 1, \ldots, d$, that $\mu_m \approx \overline{x^m}$.

  NB Sometimes, the *central moments* are more convenient. For the case of $d = 2$, the entails using

  $$m_1 \qquad \text{and} \qquad \sigma^2 = m_2 - m_1^2$$

  in place of m_1 and m_2.

- **Step 4**. We replace the distributional moments μ_m by the sample moments $\overline{x^m}$, then the solutions in (13.2) give us formulas for the **method of moment estimators** $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_d)$. For the data $\mathbf{x}$, these estimates are

  $$\hat{\theta}_1(\mathbf{x}) = g_1(\bar{x}, \overline{x^2}, \cdots, \overline{x^d}), \quad \hat{\theta}_2(\mathbf{x}) = g_2(\bar{x}, \overline{x^2}, \cdots, \overline{x^d}), \quad \ldots, \quad \hat{\theta}_d(\mathbf{x}) = g_d(\bar{x}, \overline{x^2}, \cdots, \overline{x^d}).$$

How this abstract description works in practice can be best seen through examples.

13.3 Examples

Example 13.1. *Let $X_1, X_2, \ldots, X_n$ be a simple random sample of Pareto random variables with density*

$$f_X(x|\beta) = \frac{\beta}{x^{\beta+1}}, \quad x > 1.$$

The cumulative distribution function is

$$F_X(x) = 1 - x^{-\beta}, \quad x > 1.$$

The mean and the variance are, respectively,

$$\mu = \frac{\beta}{\beta - 1}, \qquad \sigma^2 = \frac{\beta}{(\beta-1)^2(\beta-2)}.$$

In this situation, we have one parameter, namely β. Thus, in step 1, we will only need to determine the first moment

$$\mu_1 = \mu = k_1(\beta) = \frac{\beta}{\beta - 1}$$

to find the method of moments estimator $\hat{\beta}$ for β.

For step 2, we solve for β as a function of the mean μ.

$$\beta = g_1(\mu) = \frac{\mu}{\mu - 1}.$$

Consequently, a method of moments estimator for β is obtained by replacing the distributional mean μ by the sample mean $\bar{X}$.

$$\hat{\beta} = \frac{\bar{X}}{\bar{X} - 1}.$$

A good estimator should have a small variance . To use the delta method to estimate the variance of $\hat{\beta}$,

$$\sigma^2_{\hat{\beta}} \approx g_1'(\mu)^2 \frac{\sigma^2}{n}.$$

we compute

$$g_1'(\mu) = -\frac{1}{(\mu-1)^2}, \quad \textit{giving in terms of } \beta,$$

$$g_1'\left(\frac{\beta}{\beta-1}\right) = -\frac{1}{(\frac{\beta}{\beta-1} - 1)^2} = -\frac{(\beta-1)^2}{(\beta-(\beta-1))^2} = -(\beta-1)^2.$$

Thus, $\hat{\beta}$ has mean approximately equal to β and variance

$$\sigma^2_{\hat{\beta}} \approx g_1'(\mu)^2 \frac{\sigma^2}{n} = (\beta-1)^4 \frac{\beta}{n(\beta-1)^2(\beta-2)} = \frac{\beta(\beta-1)^2}{n(\beta-2)}$$

As a example, let's consider the case with $\beta = 3$ and $n = 100$. Then,

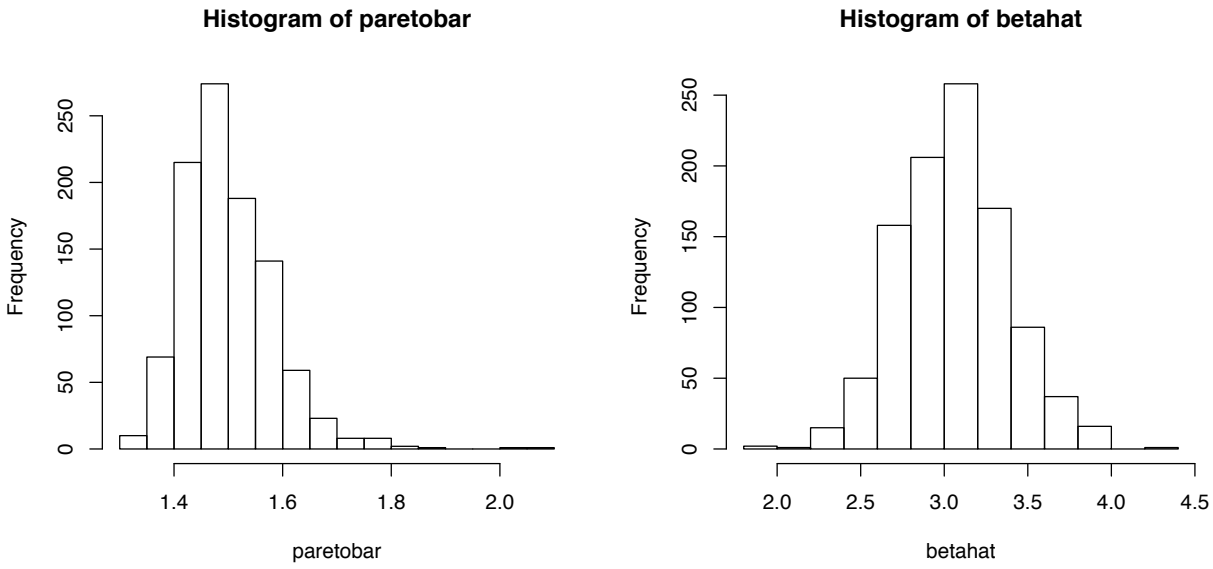
$$\sigma^2_{\hat{\beta}} \approx \frac{3 \cdot 2^2}{100 \cdot 1} = \frac{12}{100} = \frac{3}{25}, \quad \textit{and} \quad \sigma_{\hat{\beta}} \approx \frac{\sqrt{3}}{5} = 0.346.$$

To simulate this, we first need to simulate Pareto random variables. Recall that the probability transform states that if the X_i are independent Pareto random variables, then $U_i = F_X(X_i)$ are independent uniform random variables on the interval $[0,1]$. Thus, we can simulate X_i with $F_X^{-1}(U_i)$. If

$$u = F_X(x) = 1 - x^{-3}, \quad \textit{then} \quad x = (1-u)^{-1/3} = v^{-1/3}, \quad \textit{where } v = 1 - u.$$

Note that if U_i are uniform random variables on the interval $[0,1]$ then so are $V_i = 1 - U_i$. Consequently, $1/\sqrt[\beta]{V_1}, 1/\sqrt[\beta]{V_2}, \cdots$ have the appropriate Pareto distribution. ..

```
> paretobar<-numeric(1000)
> for (i in 1:1000){v<-runif(100);pareto<-1/v^(1/3);paretobar[i]<-mean(pareto)}
> hist(paretobar)
> betahat<-paretobar/(paretobar-1)
> hist(betahat)
> mean(betahat)
[1] 3.053254
> sd(betahat)
[1] 0.3200865
```



The sample mean for the estimate for β at 3.053 is close to the simulated value of 3. In this example, the estimator $\hat{\beta}$ is **biased upward**, *In other words, on average the estimate is greater than the parameter, i. e., $E_\beta\hat{\beta} > \beta$. The sample standard deviation value of 0.320 is close to the value 0.346 estimated by the delta method. When we examine unbiased estimators, we will learn that this bias could have been anticipated.*

Exercise 13.2. *The* **muon** *is an elementary particle with an electric charge of -1 and a* **spin** *(an intrinsic angular momentum) of 1/2. It is an unstable subatomic particle with a mean lifetime of 2.2 μs. Muons have a mass of about 200 times the mass of an electron. Since the muon's charge and spin are the same as the electron, a muon can be viewed as a much heavier version of the electron. The collision of an accelerated proton (p) beam having energy 600 MeV (million electron volts) with the nuclei of a production target produces positive pions (π^+) under one of two possible reactions.*

$$p + p \to p + n + \pi^+ \quad \textit{or} \quad p + n \to n + n + \pi^+$$

From the subsequent decay of the pions (mean lifetime 26.03 ns), positive muons (μ^+), are formed via the two body decay

$$\pi^+ \to \mu^+ + \nu_\mu$$

where ν_μ is the symbol of a **muon neutrino**. *The decay of a muon into a* **positron** *(e^+), an* **electron neutrino** *(ν_e), and a* **muon antineutrino** *($\bar{\nu}_\mu$)*

$$\mu^+ \to e^+ + \nu_e + \bar{\nu}_\mu$$

has a distribution angle t with density given by

$$f(t|\alpha) = \frac{1}{2\pi}(1 + \alpha\cos t), \quad 0 \le t \le 2\pi,$$

with t the angle between the positron trajectory and the μ^+-spin and **anisometry parameter** *$\alpha \in [-1/3, 1/3]$ depends the polarization of the muon beam and positron energy. Based on the measurement $t_1, \dots t_n$, give the method of moments estimate $\hat{\alpha}$ for α. (*Note: *In this case the mean is 0 for all values of α, so we will have to compute the second moment to obtain an estimator.)*

Example 13.3 (Lincoln-Peterson method of mark and recapture)**.** *The size of an animal population in a habitat of interest is an important question in conservation biology. However, because individuals are often too difficult to find,*

a census is not feasible. One estimation technique is to capture some of the animals, mark them and release them back into the wild to mix randomly with the population.

Some time later, a second capture from the population is made. In this case, some of the animals were not in the first capture and some, which are tagged, are recaptured. Let

- t *be the number captured and tagged,*
- k *be the number in the second capture,*
- r *be the number in the second capture that are tagged, and let*
- N *be the total population size.*

Thus, both t and k are under the control of the experimenter. The value of r is random and the populations size N is the parameter to be estimated. We will use a method of moments strategy to estimate N. First, note that we can guess the the estimate of N by considering two proportions.

$$\text{the proportion of the tagged fish in the second capture} \approx \text{the proportion of tagged fish in the population}$$
$$\frac{r}{k} \approx \frac{t}{N}$$

This can be solved for N to find $N \approx kt/r$. The advantage of obtaining this as a method of moments estimator is that we evaluate the precision of this estimator by determining, for example, its variance. To begin, let

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th individual in the second capture has a tag.} \\ 0 & \text{if the } i\text{-th individual in the second capture does not have a tag.} \end{cases}$$

The X_i are Bernoulli random variables with success probability

$$P\{X_i = 1\} = \frac{t}{N}.$$

They are not Bernoulli trials because the outcomes are not independent. We are sampling **without replacement**. *For example,*

$$P\{\text{the second individual is tagged}|\text{first individual is tagged}\} = \frac{t-1}{N-1}.$$

In words, we are saying that the probability model behind mark and recapture is one where the number recaptured is random and follows a **hypergeometric distribution**. *The number of tagged individuals is $X = X_1 + X_2 + \cdots + X_k$ and the expected number of tagged individuals is*

$$\mu = EX = EX_1 + EX_2 + \cdots + EX_k = \frac{t}{N} + \frac{t}{N} + \cdots + \frac{t}{N} = \frac{kt}{N}.$$

The proportion of tagged individuals, $\bar{X} = (X_1 + \cdots + X_k)/k$, has expected value

$$E\bar{X} = \frac{\mu}{k} = \frac{t}{N}.$$

Thus,

$$N = \frac{kt}{\mu}.$$

Now in this case, we are estimating μ, the mean number recaptured with r, the actual number recaptured. So, to obtain the estimate $\hat{N}$. we replace μ with the previous equation by r.

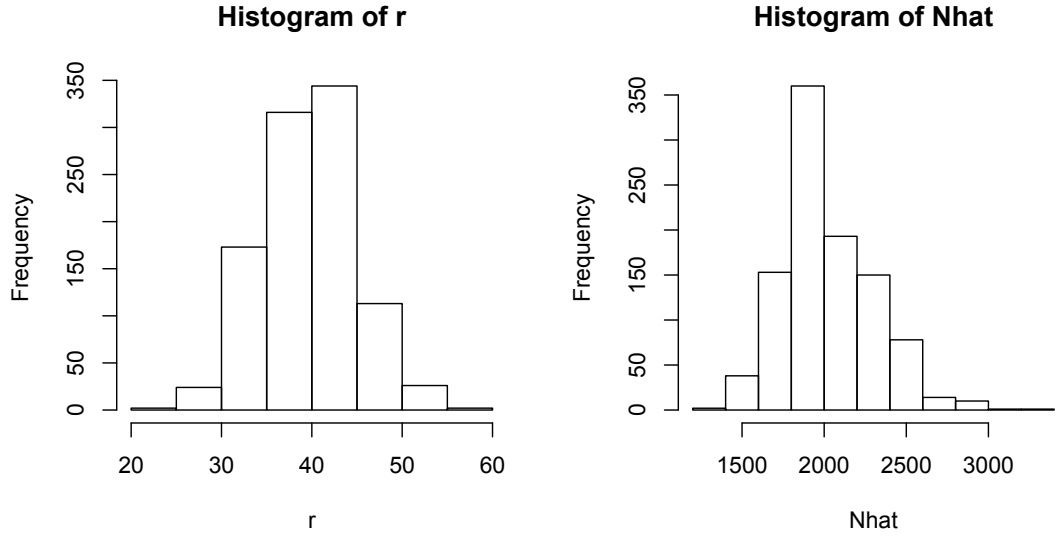
$$\hat{N} = \frac{kt}{r}$$

To simulate mark and capture, consider a population of 2000 fish, tag 200, and capture 400. We perform 1000 simulations of this experimental design. (The R *command* `replicate` *repeats a chosen number of times (here 1000) the stated expression and stores, in this case, in the vector* `r`*).*

```
> t<-200;k<-400;N<-2000
> fish<-c(rep(1,t),rep(0,N-t))
> r<-replicate(1000,sum(sample(fish,k)))
> Nhat<-k*t/r
```

The command `sample(fish,400)` *creates a vector of length 400 of zeros and ones for, respectively, untagged and tagged fish. Thus, the* `sum` *command gives the number of tagged fish in the simulation. This is repeated 1000 times and stored in the vector* `r`*. Let's look a summaries of* `r` *and the estimates* $\hat{N}$ *of the population.*

```
> mean(r)
[1] 40.09
> sd(r)
[1] 5.245705
> mean(Nhat)
[1] 2031.031
> sd(Nhat)
[1] 276.6233
```



To estimate the population of pink salmon in Deep Cove Creek in southeastern Alaska, 1709 fish were tagged. Of the 6375 carcasses that were examined, 138 were tagged. The estimate for the population size

$$\hat{N} = \frac{6375 \times 1709}{138} \approx 78948.$$

Exercise 13.4. *Use the delta method to estimate Var*$(\hat{N})$ *and* $\sigma_{\hat{N}}$*. Apply this to the simulated sample and to the Deep Cove Creek data.*

Example 13.5. *Fitness is a central concept in the theory of evolution. Relative fitness is quantified as the average number of surviving progeny of a particular genotype compared with average number of surviving progeny of competing genotypes after a single generation. Consequently, the distribution of fitness effects, that is, the distribution of fitness for newly arising mutations is a basic question in evolution. A basic understanding of the distribution of fitness effects is still in its early stages. Eyre-Walker (2006) examined one particular distribution of fitness effects, namely, deleterious amino acid changing mutations in humans. His approach used a gamma-family of random variables and gave the estimate of* $\hat{\alpha} = 0.23$ *and* $\hat{\beta} = 5.35$*.*

A $\Gamma(\alpha, \beta)$ random variable has mean α/β and variance α/β^2. Because we have two parameters, the method of moments methodology requires us, in step 1, to determine the first two moments.

$$E_{(\alpha,\beta)}X_1 = \frac{\alpha}{\beta} \quad \textit{and} \quad E_{(\alpha,\beta)}X_1^2 = \textit{Var}_{(\alpha,\beta)}(X_1) + E_{(\alpha,\beta)}[X_1]^2 = \frac{\alpha}{\beta^2} + \left(\frac{\alpha}{\beta}\right)^2 = \frac{\alpha(1+\alpha)}{\beta^2} = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2}.$$

Thus, for step 1, we find that

$$\mu_1 = k_1(\alpha, \beta) = \frac{\alpha}{\beta}, \quad \mu_2 = k_2(\alpha, \beta) = \frac{\alpha}{\beta^2} + \frac{\alpha^2}{\beta^2}.$$

For step 2, we solve for α and β. Note that

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

$$\frac{\mu_1}{\mu_2 - \mu_1^2} = \frac{\alpha/\beta}{\alpha/\beta^2} = \beta,$$

and

$$\mu_1 \cdot \frac{\mu_1}{\mu_2 - \mu_1^2} = \frac{\alpha}{\beta} \cdot \beta = \alpha, \quad \textit{or} \quad \alpha = \frac{\mu_1^2}{\mu_2 - \mu_1^2}.$$

So set

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \textit{and} \quad \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

to obtain estimators

$$\hat{\beta} = \frac{\bar{X}}{\overline{X^2} - (\bar{X})^2} = \frac{\bar{X}}{S^2} \quad \textit{and} \quad \hat{\alpha} = \hat{\beta}\bar{X} = \frac{(\bar{X})^2}{\overline{X^2} - (\bar{X})^2} = \frac{(\bar{X})^2}{S^2}.$$

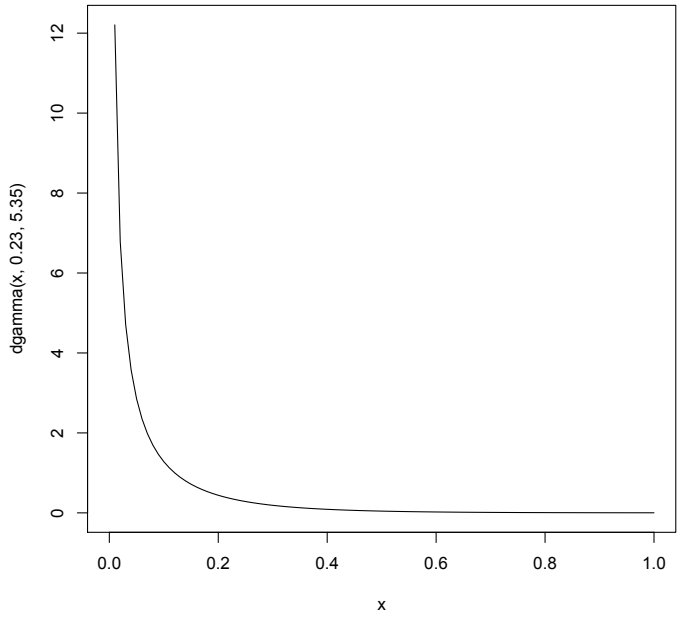


Figure 13.1: The density of a $\Gamma(0.23, 5.35)$ random variable.

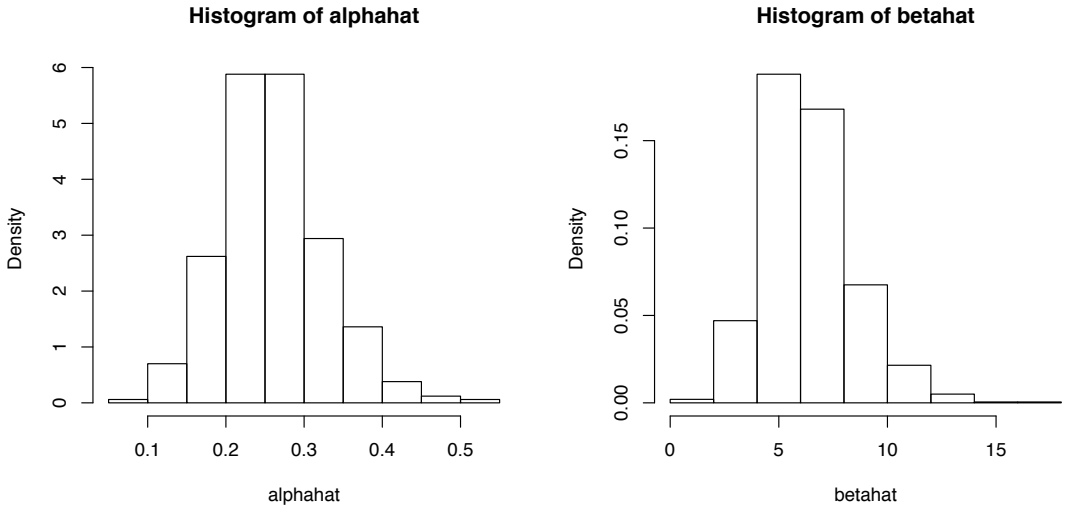
The result shows how using the sample variance can simplify the algebra in finding the method of moments estimator.

To investigate the method of moments on simulated data using R*, we consider 1000 repetitions of 100 independent observations of a* $\Gamma(0.23, 5.35)$ *random variable.*

```
> xbar <- numeric(1000)
> x2bar <- numeric(1000)
> for (i in 1:1000){x<-rgamma(100,0.23,5.35);xbar[i]<-mean(x);x2bar[i]<-mean(x^2)}
> betahat <- xbar/(x2bar-(xbar)^2)
> alphahat <- betahat*xbar
> mean(alphahat)
[1] 0.2599894
> sd(alphahat)
[1] 0.06672909
> mean(betahat)
[1] 6.315644
> sd(betahat)
[1] 2.203887
```

To obtain a sense of the distribution of the estimators $\hat{\alpha}$ *and* $\hat{\beta}$*, we give histograms.*

```
> hist(alphahat,probability=TRUE)
> hist(betahat,probability=TRUE)
```



As we see, the variance in the estimate of β *is quite large. We will revisit this example using maximum likelihood estimation in the hopes of reducing this variance. The use of the delta method is more difficult in this case because it must take into account the correlation between* $\bar{X}$ *and* $\overline{X^2}$ *for independent gamma random variables. Indeed, from the simulation, we have an estimate..*

```
> cor(xbar,x2bar)
[1] 0.8120864
```

Moreover, the two estimators $\hat{\alpha}$ *and* $\hat{\beta}$ *are fairly strongly positively correlated. Again, we can estimate this from the simulation.*

```
> cor(alphahat,betahat)
[1] 0.7606326
```

In particular, an estimate of $\hat{\alpha}$ *and* $\hat{\beta}$ *are likely to be overestimates or underestimates in tandem.*

13.4 Answers to Selected Exercises

13.2. Let T be the random variable that is the angle between the positron trajectory and the μ^+-spin. Then integrate by parts twice to obtain

$$\mu_2 = E_\alpha T^2 = \frac{1}{2\pi}\int_{-\pi}^{\pi} t^2(1+\alpha\cos t)dt = \frac{\pi^2}{3} - 2\alpha$$

Thus, $\alpha = (\mu_2 - \pi^2/3)/2$. This leads to the method of moments estimate

$$\hat{\alpha} = \frac{1}{2}\left(\overline{t^2} - \frac{\pi^2}{3}\right)$$

where $\overline{t^2}$ is the sample second moment.

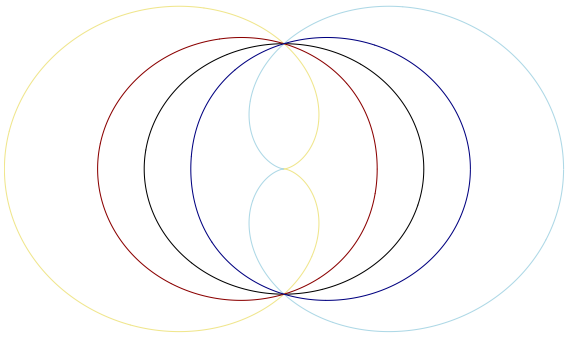


Figure 13.2: Densities $f(t|\alpha)$ for the values of $\alpha = -1$ (yellow). $-1/3$ (red), 0 (black), 1/3 (blue), 1 (light blue).

13.4. Let X be the random variable for the number of tagged fish. Then, X is a hypergeometric random variable with

$$\text{mean } \mu_X = \frac{kt}{N} \quad \text{and variance} \quad \sigma_X^2 = k\frac{t}{N}\frac{N-t}{N}\frac{N-k}{N-1}$$

$$N = g(\mu_X) = \frac{kt}{\mu_X}. \quad \text{Thus, } g'(\mu_X) = -\frac{kt}{\mu_X^2}.$$

The variance of $\hat{N}$

$$\begin{aligned}\text{Var}(\hat{N}) \approx g'(\mu)^2\sigma_X^2 &= \left(\frac{kt}{\mu_X^2}\right)^2 k\frac{t}{N}\frac{N-t}{N}\frac{N-k}{N-1} = \left(\frac{kt}{\mu_X^2}\right)^2 k\frac{t}{kt/\mu_X}\frac{kt/\mu_X - t}{kt/\mu_X}\frac{kt/\mu_X - k}{kt/\mu_X - 1}\\ &= \left(\frac{kt}{\mu_X^2}\right)^2 k\frac{\mu_X t}{kt}\frac{kt-\mu_X t}{kt}\frac{kt-k\mu_X}{kt-\mu_X} = \left(\frac{kt}{\mu_X^2}\right)^2 k\frac{\mu_X}{k}\frac{k-\mu_X}{k}\frac{k(t-\mu_X)}{kt-\mu_X}\\ &= \frac{k^2t^2}{\mu_X^3}\frac{(k-\mu_X)(t-\mu_X)}{kt-\mu_X}\end{aligned}$$

Now if we replace μ_X by its estimate r we obtain

$$\sigma_{\hat{N}}^2 \approx \frac{k^2t^2}{r^3}\frac{(k-r)(t-r)}{kt-r}.$$

For $t = 200, k = 400$ and $r = 40$, we have the estimate $\sigma_{\hat{N}} = 268.4$. This compares to the estimate of 276.6 from simulation.

For $t = 1709, k = 6375$ and $r = 138$, we have the estimate $\sigma_{\hat{N}} = 6373.4$.