

Topic 13: Method of Moments*

October 25, 2011

1 Introduction

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

Let M_1, M_2, \dots 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.

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

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, \dots 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 a one dimensional parameter 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

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

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.

*© 2011 Joseph C. Watkins

2 The Procedure

More generally, for independent random variables X_1, X_2, \dots 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

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

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

$$\mu_m = EX^m = k_m(\theta). \quad (1)$$

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 (1) for the first d moments,

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

obtaining d equations in d unknowns.

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

$$\theta_1 = g_1(\mu_1, \mu_2, \dots, \mu_d), \quad \theta_2 = g_2(\mu_1, \mu_2, \dots, \mu_d), \quad \dots, \quad \theta_d = g_d(\mu_1, \mu_2, \dots, \mu_d). \quad (2)$$

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

$$\bar{x}, \bar{x^2}, \dots, \bar{x^d}.$$

Using the law of large numbers, we have that $\mu_m \approx \bar{x^m}$.

- **Step 4.** We replace the distributional moments μ_m by the sample moments $\bar{x^m}$, then the solutions in (2) give us formulas for the **method of moment estimators** $(\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_d)$. For the data, these estimates are

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

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

3 Examples

Example 1. Let X_1, X_2, \dots, 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,

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

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 estimate 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 $\hat{\beta}$, we compute

$$g_1'(\mu) = -\frac{1}{(\mu - 1)^2}, \quad \text{giving} \quad g_1' \left(\frac{\beta}{\beta - 1} \right) = -\frac{1}{\left(\frac{\beta}{\beta - 1} - 1 \right)^2} = -\frac{(\beta - 1)^2}{(\beta - (\beta - 1))^2} = -(\beta - 1)^2$$

and find that $\hat{\beta}$ has mean approximately equal to β and variance

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

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

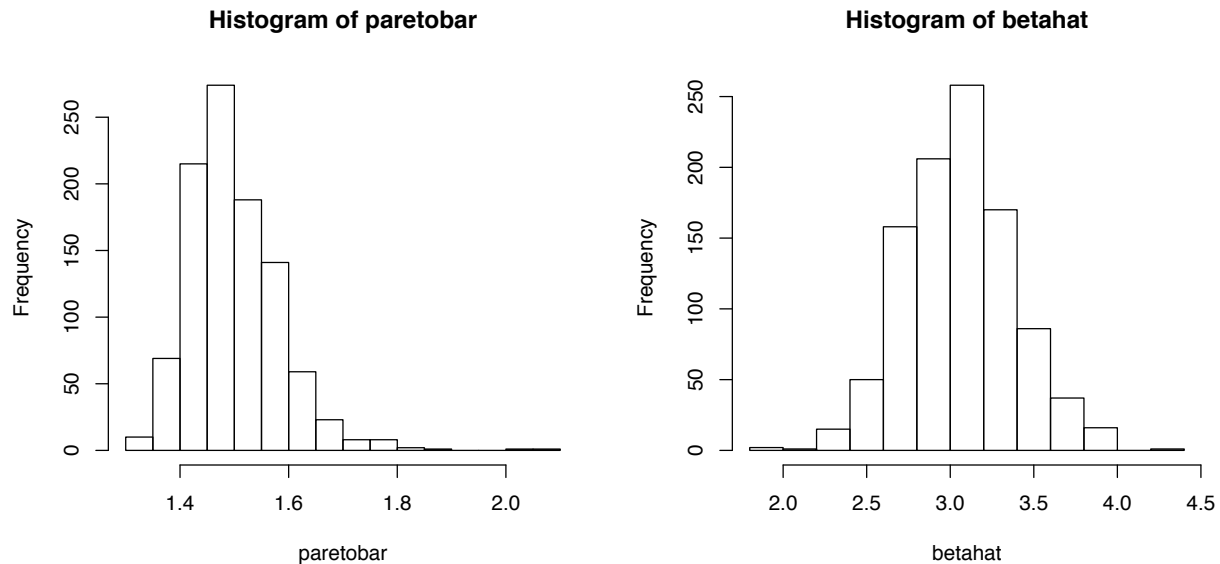
$$\sigma_{\hat{\beta}}^2 \approx \frac{3 \cdot 2^2}{100 \cdot 1} = \frac{12}{100} = \frac{3}{25}, \quad \sigma_{\hat{\beta}} \approx \frac{\sqrt{3}}{5} = 0.346.$$

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

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

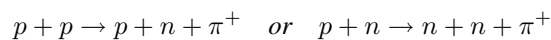
Note that if U_i are uniform random variables on the interval $[0, 1]$ then so are $V_i = 1 - U_i$. Consequently, $1/\sqrt[3]{V_1}, 1/\sqrt[3]{V_2}, \dots$ have the appropriate Pareto distribution.

```
> paretobar<-rep(0,1000)
> 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
```

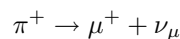


The sample mean for the estimate for β at 3.053 is close to the simulated value of 3. In this example, the estimator $\hat{\beta}$ is **biased upward**. In other words, on average the estimate is greater than the parameter, i. e., $E_{\beta}\hat{\beta} > \beta$. The sample standard deviation value of 0.320 is close to the value 0.346 estimated by the delta method.

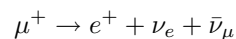
Exercise 2. The **muon** is an elementary particle with an electric charge of -1 and a **spin** (an intrinsic angular momentum) of $1/2$. It is an unstable subatomic particle with a mean lifetime of $2.2 \mu\text{s}$. Muons have a mass of about 200 times the mass of an electron. Since the muon's charge 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 with the nuclei of a production target produces positive pions (π^+) under one of two possible reactions.



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



where ν_{μ} is the symbol of a **muon neutrino**. Neutrinos have been recently reported as traveling faster than the speed of light. The decay of a muon into a **positron** (e^+), an **electron neutrino** (ν_e), and a **muon antineutrino** ($\bar{\nu}_{\mu}$)



has a distribution angle t with density given by

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

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 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,

We will use a method of moments strategy to estimate N . 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 $X = X_1 + X_2 + \dots + X_k$ and the expected number of tagged individuals is

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

Thus,

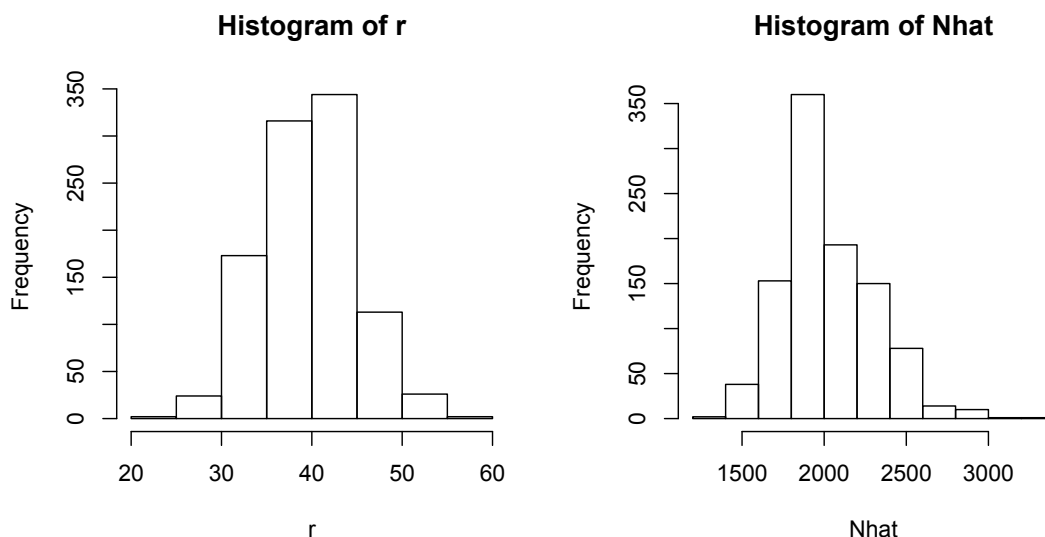
$$N = \frac{kt}{\mu}.$$

Now in this case, we are only averaging one outcome, namely, we are estimating μ , the mean number of 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.

```
> r<-rep(0,1000)
> fish<-c(rep(1,200),rep(0,1800))
> for(j in 1:1000){r[j]<-sum(sample(fish,400))}
> Nhat<-200*400/r
> 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 4. Use the delta method to estimate $\text{Var}(\hat{N})$ and $\sigma_{\hat{N}}$. Apply this to the simulated sample and to the Deep Cove Creek data.

Example 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 of newly arising mutations 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, we will need to determine the first two moments.

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

Thus, for step 1, we find that

$$\mu_1 = k_1(\alpha, \beta) = \frac{\alpha}{\beta}, \quad \mu_2 = k_2(\alpha, \beta) = \frac{\alpha(1 + \alpha)}{\beta^2}.$$

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

$$\begin{aligned} \mu_2 - \mu_1^2 &= \frac{\alpha}{\beta^2}, \\ \frac{\mu_1}{\mu_2 - \mu_1^2} &= \frac{\alpha/\beta}{\alpha/\beta^2} = \beta, \end{aligned}$$

and

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

So set

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

to obtain estimators

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

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 <- rep(0,1000)
> x2bar <- rep(0,1000)
> 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] 0.2599513
> sd(alphahat)
[1] 0.06628046
> mean(betahat)
[1] 6.331623
> sd(betahat)
[1] 2.046528
> hist(alphahat,probability=TRUE)
```

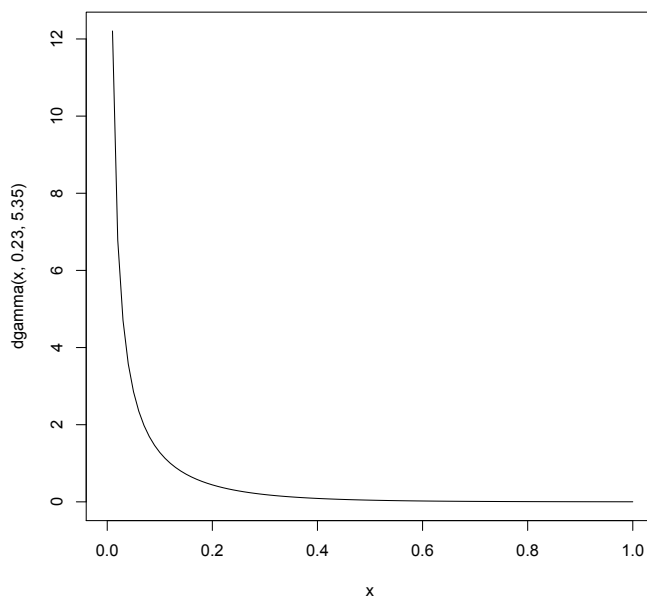
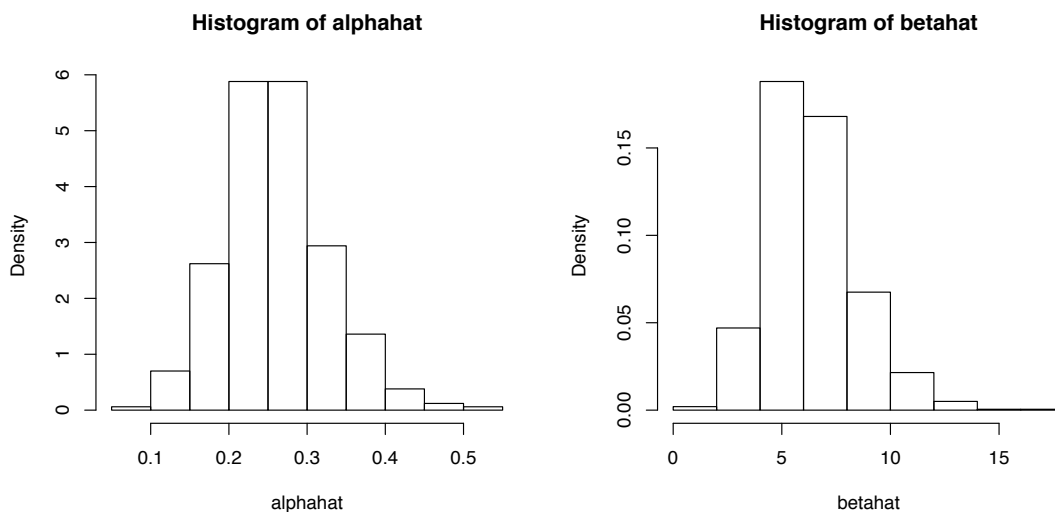


Figure 1: The density of a $\Gamma(0.23, 5.35)$ random variable.

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



As we see, the variance in the estimate of β is quite large. We will revisit this example using another estimation method in the hopes of reducing this variance.

4 Answers to Selected Exercises

2. Let T be the random variable that is the angle between the positron trajectory and the μ^+ -spin

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

Thus, $\alpha = 3\mu/3$. 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 mean of the square of the observations.

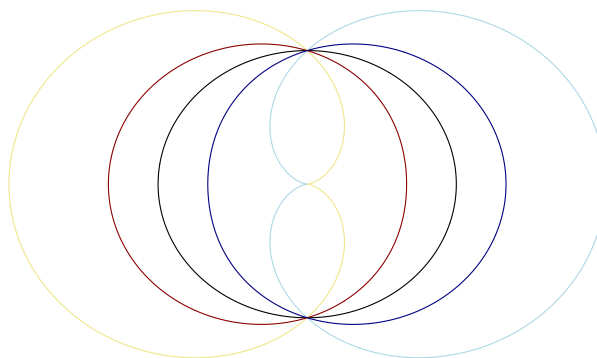


Figure 2: Densities $f(t|\alpha)$ for the values of $\alpha = -1$ (yellow), $-1/3$ (red), 0 (black), $1/3$ (blue), 1 (light blue).

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

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

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

The variance of \hat{N}

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

Now if we replace μ_X by its estimate r we obtain

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

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 $\sigma_{\hat{N}} = 6373.4$.