Week 9 – Unbiased estimation - Efficiency, MSE

Juan Pablo Lewinger

Last Class-CLT

$X_1, \ldots, X_n$ i.i.d. $E[X_i] = \mu$, $Var[X_i] = \sigma^2$ then:

Sum form: $S_n = X_1 + \ldots + X_n$

$$Z_n = \frac{S_n - E[S_n]}{\sqrt{Var[S_n]}} = \frac{S_n - n\mu}{\sqrt{n}\sigma} \xrightarrow{D} N(0,1)$$

Sample mean form: $\overline{X}_n = \frac{X_1 + \ldots + X_n}{n}$

$$Z_n = \frac{\overline{X}_n - E[\overline{X}_n]}{\sqrt{Var[\overline{X}_n]}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1)$$

The symbol $\xrightarrow{D}$ denotes convergence in distribution. It means that as $n \to \infty$ the cumulative distribution function of $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$ gets closer and closer to the cdf of a standard normal.

Formally, if $F_n(x) = F_{Z_n}(x)$ denotes the cdf of Z_n, then $F_n(x) \to \Phi(x)$ for every $x \in \mathbb{R}$, where $\Phi(x)$ is the cdf of a $N(0,1)$ RV.

Last Class-Basic statistical model

- Want to learn characteristics of a population:
  - Income of Los Angeles residents
  - Blood pressure of patients form LA county hospital
  - Vaping among teenagers in California
  - Genotypes at a gene among individuals with Hispanic ancestry
- We model the distribution of the characteristic as a random variable X (e.g. height, blood pressure, vaping (yes vs. no), Genotypes (0,1,2)))
- To learn about X we collect a random sample: $X_1, X_2, \ldots, X_n$ from $F_X(x)$, the distribution of X
- $X_1, X_2, \ldots, X_n$ have the same probability distribution and are mutually independent.
- The distribution $F_X(x)$ of X_i is unknown or only partially known.
- E.g. $X_i \sim N(\mu, \sigma^2)$ with known σ^2 but unknown μ

Estimation

- We want to learn about a distribution in a population
- e.g. proportion of democrat voters in CA, average blood pressure among covid survivors aged 70+, vaping frequency among young adults in the US, rate of patient ER night admissions in LA county hospitals
- We take a sample of size n from the population and conceptualize it as a random sample $X_1, \ldots, X_n$ from a distribution $F_X(x)$ that is **totally or partially unknown to us**.
- We want to estimate specific characteristics or **parameters** of the underlying distribution:
- true mean $E[X_i] = \mu$ (e.g. blood pressure)
- the true variance, $Var[X_i] = \sigma^2$ (X_i e.g. blood pressure)
- or a probability like $P(X_i = 1)$ (e.g. X_i = 1 = if democrat voter; 0= if not democrat voter)
- or a rate (e.g. expected number of ER patients per hour) (X_i = number of patients within a period of 1-hour)

Estimation

Example 1: To estimate the unknown mean of a population μ (e.g. blood pressure)

- Natural to use the sample mean $\overline{X_n}$ to estimate μ because we know that for large n the sample mean will be close to the true mean.

Example 2. We can model the number of arrivals per hour at an ER unit as a $Poisson(\lambda)$. Suppose we count the arrivals during each on n (non-overlapping) 1-hour intervals to get the random sample $X_1, \dots, X_n$. Here λ is unknown and we want to estimate it.

- A natural estimate is also the sample mean $\overline{X_n}$ because $E[X_i] = \lambda$.
- But using the sample variance S_n^2 is also reasonable because $Var[X_i] = \lambda$
- Today we'll learn how to choose among different options for estimating a parameter of interest

Estimator vs. estimate

- Estimate: value t that only depends on the dataset $x_1, x_2, \dots, x_n$, i.e., t is some function of the dataset $t = h(x_1, x_2, \dots, x_n)$. Example: $t = \overline{x_n} = \frac{x_1 + \ldots + x_n}{n}$
- *An estimate is a number* (or a vector of numbers in more complex problems)
- Estimator: Let $t = h(x_1, x_2, \dots, x_n)$ be an estimate based on the dataset $x_1, x_2, \dots, x_n$. Then t is a realization of the random variable $T = h(X_1, X_2, \dots, X_n)$. The random variable T is called an estimator.
- *An estimator is a random variable*
- *An estimate is a realization of random variable*

Sampling distribution

Example: estimating the proportion p of LA teenagers that vape from a sample $X_1, \dots X_n$, $X_i \sim Bernoulli(p)$.

- X_i records whether the teenager vapes (yes=1 vs. no=0)
- The sample proportion $\hat{p}_n = \frac{S_n}{n}$ of vapers is a natural estimator of p ($S_n = X_1 + \dots + X_n$ is the total number of vapers in the sample), because by the LLN $\hat{p} \longrightarrow p$
- The sampling distribution is just the standard distribution (pdf, pmf, cdf) of the random variable we call the estimator.
- For example of vapers, the sampling distribution of $\hat{p}$ is the distribution of the random variable $\hat{p}_n = \frac{S_n}{n}$

Sampling distribution of the sample proportion

Simulating one study: $X_1, \ldots, X_n \sim Bernoulli(p)$

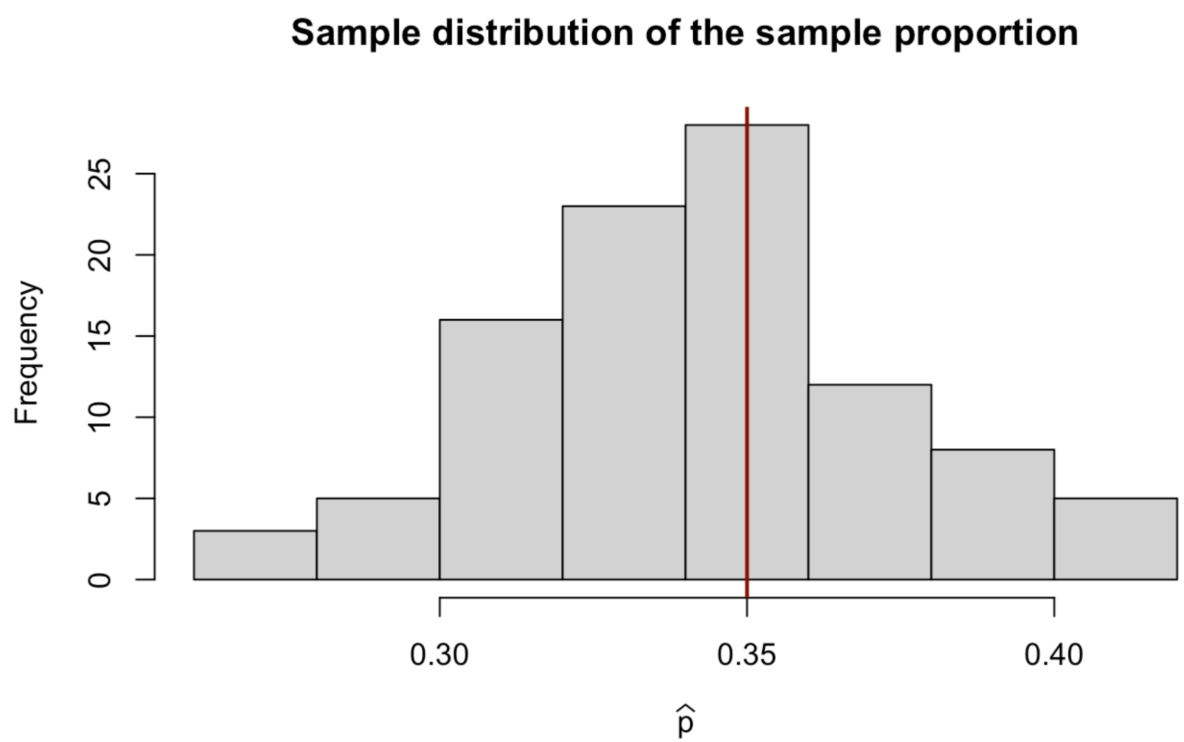
```
set.seed(2023)
n = 200 # sample size
p = 0.35 # True population frequency (eg. vaping among teenagers)
X = rbinom(n, size = 1, prob=p)
mean(X)
```

```
## [1] 0.37
```

Sampling distribution of the sample proportion

Simulating MULTIPLE studies

```
for (i in 1:nsims){
  X = rbinom(n, size = 1, prob=p)
  p_hat[i] = mean(X)
  }
```



Sampling distribution of the sample mean

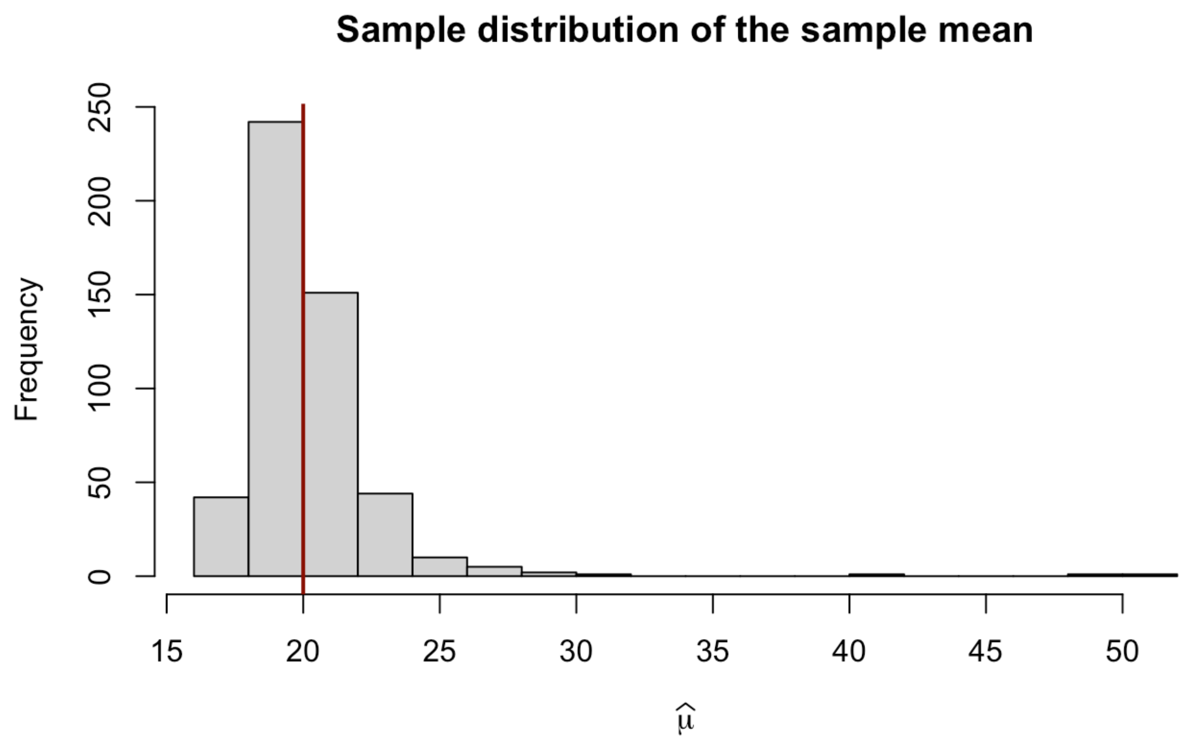
Simulating one study: $X_1, \ldots, X_n \sim Pareto(x_m, \alpha)$

```
library(EnvStats)
set.seed(2023)
n = 200 # sample size
loc = 10; alpha = 2 #E[X_i=20] e.g. income distribution in $1000
X = rpareto(n, location = loc, shape=alpha)
mean(X)
```

```
## [1] 21.14341
```

Sampling distribution of the sample mean

```
for (i in 1:nsims){
  X = rpareto(n, location = loc, shape=alpha)
  mu_hat[i] = mean(X)
  }
```



Unbiased Estimators

$X_1, \dots X_n$ a random sample from distribution $F(x|\theta)$, and θ a parameter of interest about F (e.g. mean)

- An estimator T of a parameter θ is unbiased if $E[T] = \theta$
- An unbiased estimator has no systematic tendency to produce estimates that are larger than or smaller than the target parameter θ
- The difference $E[T] - \theta$ is called the bias of the estimator T
- If $bias \neq 0$ the estimator is called biased

Sample Mean and sample variance

- $X_1, \dots X_n$ a random sample with mean $E[X_i] = \mu$ and $Var[X_i] = \sigma^2$
- The sample variance $\overline{X_n}$ is an unbiased estimator of the population mean μ.

$$E[\overline{X_n}] = \mu$$

- The sample mean $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}_n\right)^2$ is an unbiased estimator of the population variance σ^2

$$E[S_n^2] = \sigma^2$$

Unbiasedness is not preserved under general transformations

- In general, if T is an unbiased estimator of θ, $g(T)$ is not an unbiased estimate of $g(\theta)$
- Example:

  $\overline{X_n}$ is unbiased for $E[X_i] = \mu$

  But $\bar{X}_n{}^2$ is not unbiased for $E[X_i]^2 = \mu^2$, Because by Jensen's inequality $E[\bar{X}_n{}^2] > E[\bar{X}_n]^2 = \mu^2$
- Jensen's inequality: if $g(t)$ is a convex function, then $E[g(T)] \geq g(E[T])$. Equality holds only if $g(x) = at + b$
- If g linear $g(t) = at + b$, $E[g(T)] = E[aT + b] = aE[T] + b = g(E[T])$ so $g(T)$ is unbiased for $g(\theta)$.

Choosing among unbiased estimators. Example

$X_1, \ldots, X_n \overset{i.i.d}{\sim} U[0, \theta]$

$f_X(x) = \frac{1}{\theta} I_{[0,\theta]}(x)$

$E[X_i] = \frac{\theta}{2}$, so $E[\overline{X_n}] = \frac{\theta}{2}$

Then $\widehat{\theta} = 2\overline{X_n}$ is an unbiased estimator of θ

Intuitively a natural estimate for theta could also be based on $T = \max\{X_1, \ldots, X_n\}$

$f_T(t) = n\frac{t^{n-1}}{\theta^n} I_{[0,\theta]}(t)$

$E[T] = \frac{n}{n+1}\theta$ so $\widetilde{\theta} = \frac{n+1}{n} T$ is an unbiased estimator of θ

So, both $\widehat{\theta}$ and $\widetilde{\theta}$ are unbiased estimators of θ, which one is better?

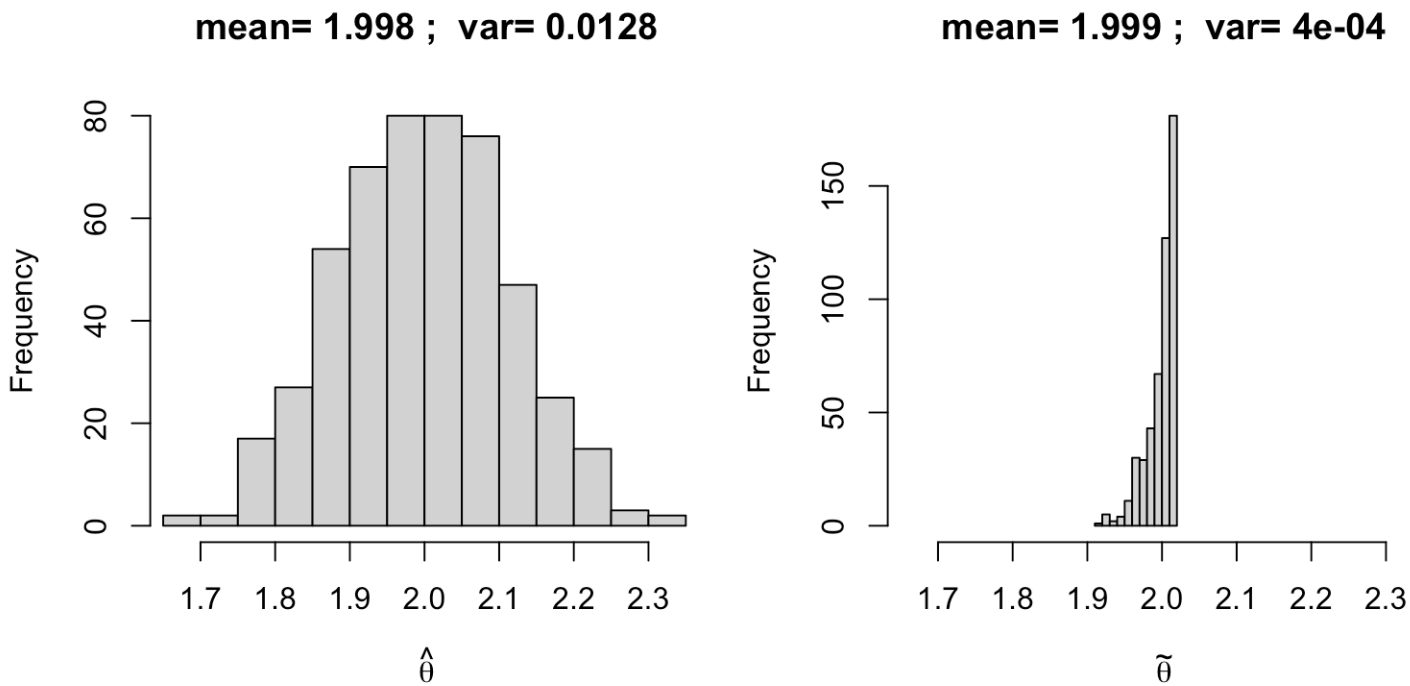
Choosing among unbiased estimators-Example.

Let’s perform a simulation to see how the two estimators behave:

```
set.seed(101)
theta = 2
n = 50
nsims=500
theta_hat = theta_tilde = numeric(nsims)

for (i in 1:nsims){
    x = runif(n=n, max=2)
    theta_hat[i] = 2 * mean(x)
    theta_tilde[i] = ((n+1)/n) * max(x)
}
```

Choosing among unbiased estimators-Example



Choosing among unbiased estimators-Example.

- We see empirically that $\widetilde{\theta}$ is much less variable than $\widehat{\theta}$
- But also theoretically, $Var[\widetilde{\theta}] = \frac{\theta^2}{n(n+2)}$ and $Var[\widehat{\theta}] = \frac{\theta^2}{3n}$
- So, $Var[\widetilde{\theta}] < Var[\widehat{\theta}]$, for $n \geq 2$ and goes much faster to zero!
- Additionally, $\widehat{\theta}$ can take values way over the true θ
- So, $\widetilde{\theta}$ is a much better estimate of θ than $\widehat{\theta}$

Mean square error

- Although unbiasedness is a desirable property, unbiased estimators do not always exist
- Even when they exist, requiring unbiadness maybe too stringent (i.e. there can be other good estimators that are biased)
- A general performance of an estimator can be judged by the way it spreads around the parameter to be estimated:

  If T is an estimator for a parameter θ, the mean squared error of T is the number:

  $MSE(T) = E[(T - \theta)^2]$.

- It's easy to show that $MSE(T) = Var(T) + Bias(T)^2$, where $Bias(T) = E[T] - \theta$
- A biased estimator with a small bias could be more useful than an unbiased estimator with a large variance.
- Better to use $MSE(T)$ to choose between estimators
- When $Bias(T) = 0$, $MSE(T) = Var(T)$

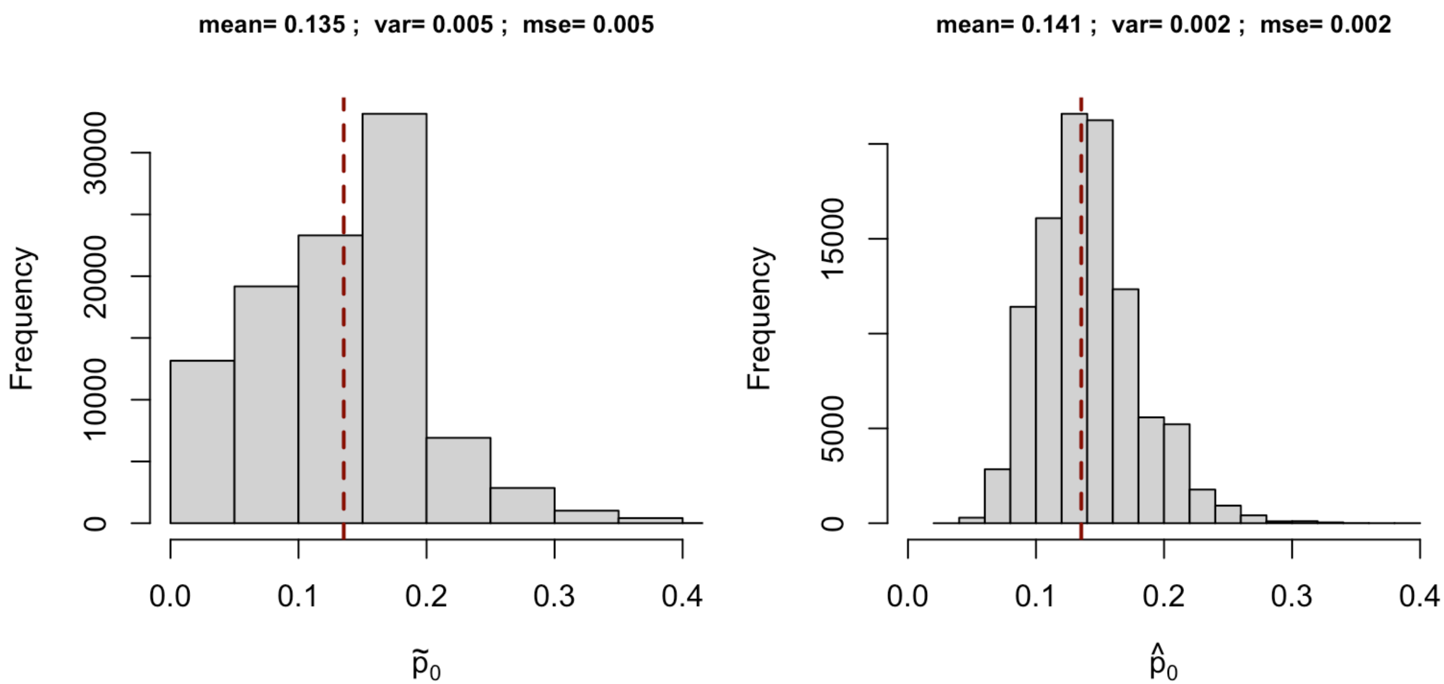
Mean square error - example

$X_1, \ldots, X_n \overset{i.i.d}{\sim} Poisson(\mu)$

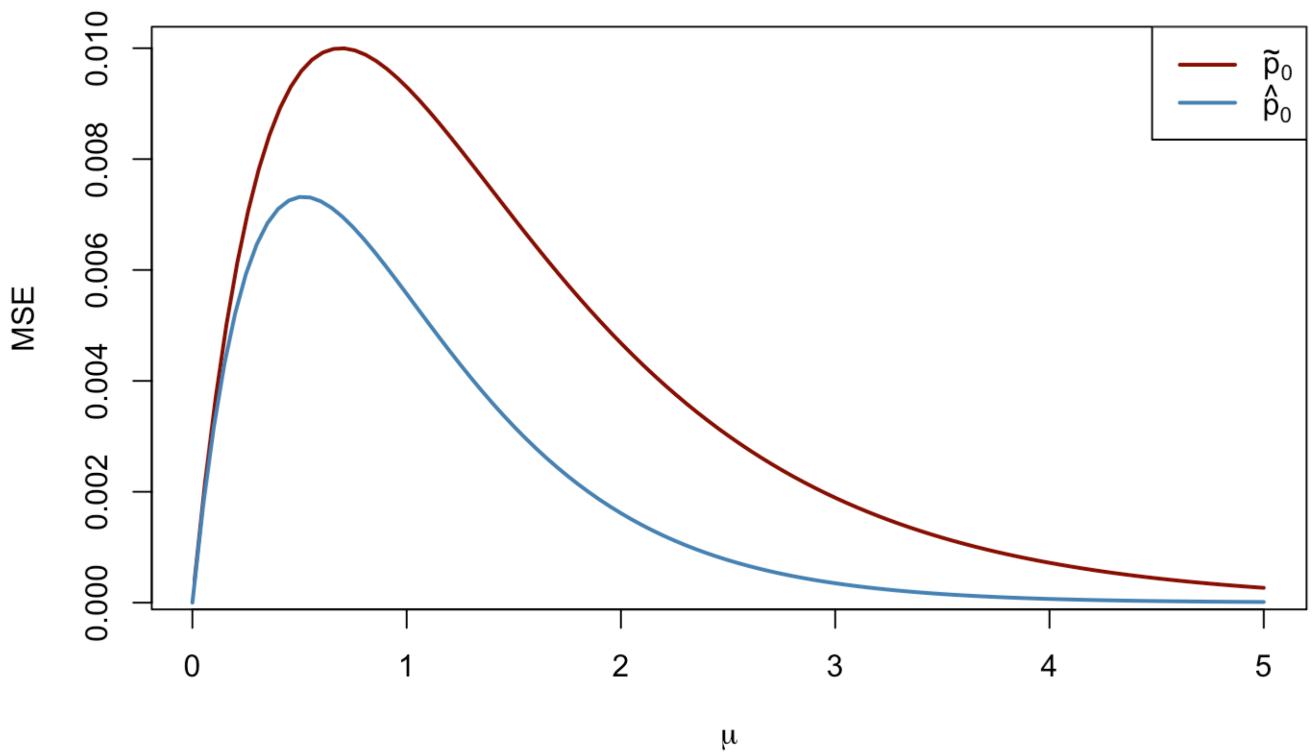
- Two candidate estimators for $p_0 = P(X_i = 0) = e^{-\mu}$:
  - $\widetilde{p}_0 = \frac{\text{number of } X_i = 0}{n}$
  - $\widehat{p}_0 = e^{-\overline{X_n}}$
- $\widetilde{p}_0$ is unbiased but $\widehat{p}_0$ is not? Which one is better?

Mean square error - example

- Simulation to see how the two estimators behave
- True $p_0 = e^{-2} = 0.135$



Mean square error - example



Next week

- Read *MIPS Ch 21*