Week 4: Estimators and their distributions

STA238: Probability, Statistics, and Data Analysis II

Luis Sierra Muntané

2026-01-27

Recap Estimators

- An **estimator** is a statistic we associate to a model parameter.
- Since it is a function of random variables, it is itself a random variable
- It is intended to approximate an unknown parameter of the data-generating distribution.

Given a sample $(X_1, X_2, \ldots, X_n)$, examples of estimators include:

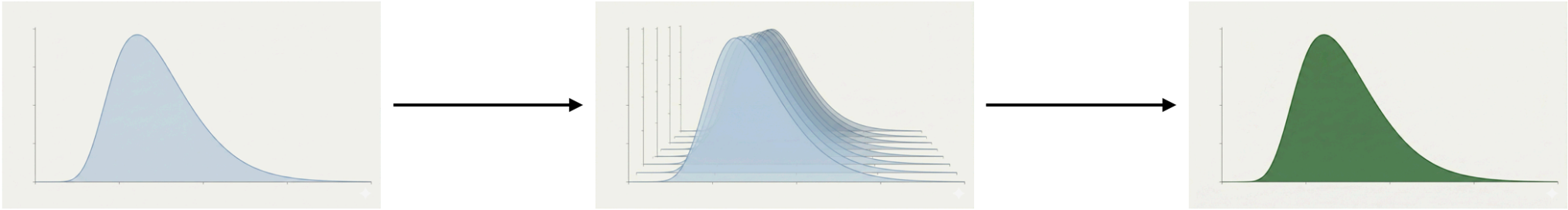
- The sample mean $\hat{\mu} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$, for μ.
- The sample median $\hat{\mu} = \text{Med}(X_1, X_2, \ldots, X_n)$, for μ.
- The trimmed mean for μ.
- The sample variance $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ for the variance σ^2.
- $\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ for the variance σ^2.

- An **estimate** is a realization of an estimator, and so is also associated to a model parameter.
- It is obtained *after* observing the data.

> Readings: - Chapter 14, 19, 20 MIPS - Chapter 6.1-6.2, Chapter 7.1 MMSA

Sampling distribution

- A sample from a population distribution $X_1, \ldots, X_n \sim F$ follows some distribution F.
- From the sample we calculate a statistic $h(X_1, \ldots, X_n)$, which is therefore a random variable.
- The distribution of the random variable $\hat{\theta} = h(X_1, \ldots, X_n)$ is known as the **sampling distribution** $F_{\hat{\theta}}$.



Sampling distribution II

Suppose we have a iid. samples $X_1, \ldots, X_n \sim \mathbf{Bern}(p)$.

We choose to track the statistic of the sample mean:
$\bar{X}_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$

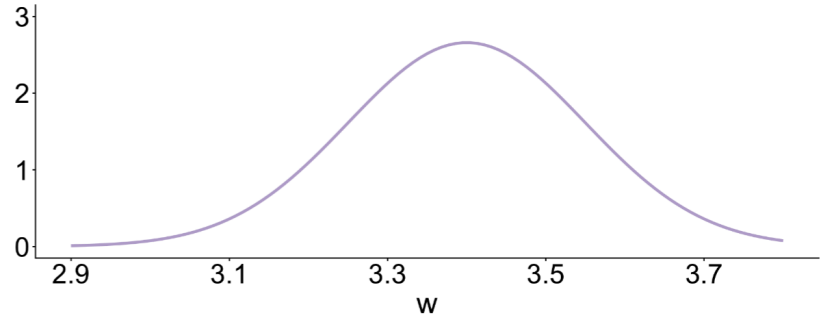
- For $n = 1$: $\bar{X}_1 = X_1 \sim \mathbf{Bern}(p)$,
- For $n = 2$: $\bar{X}_2 = \frac{1}{2}(X_1 + X_2) \implies \begin{cases} \mathbb{P}(\bar{X}_2 = 0) = 1/4, \\ \mathbb{P}(\bar{X}_2 = 1/2) = 1/2, \\ \mathbb{P}(\bar{X}_2 = 1) = 1/4. \end{cases}$
- $\bar{X}_n \sim \frac{1}{n}\mathbf{Bin}(n, p)$.

Example: Normal distribution

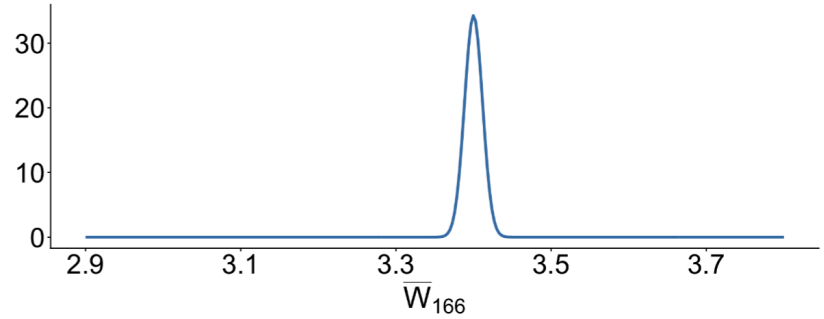
Assume that our data is iid. normally distributed $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$.

We can calculate the sampling distribution of $\bar{X}_n$ exactly:

- $\mathbb{E}[\bar{X}_n] =$
- $\mathbf{Var}(\bar{X}_n) =$



And by properties of the normal distribution, in fact we have $\bar{X}_n \sim \mathcal{N}(\quad, \quad)$

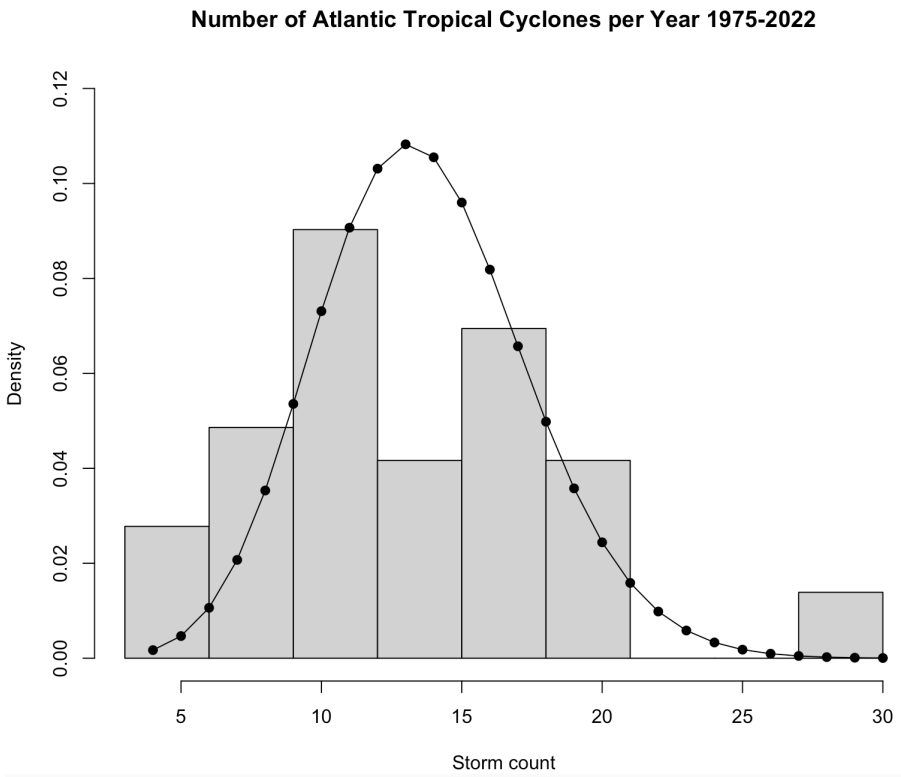


Example: Atlantic Tropical Cyclones

We can try to model yearly Atlantic cyclones using a Poisson distribution $Y_i \sim \text{Pois}(\lambda)$.

```r
data(storms)

# Count unique storms per year
storms_per_year <- tapply(
  storms$name,
  storms$year,
  function(x) length(unique(x))
)
```

We can then choose an estimator for λ.



Example: how many fish are there in a pond?

Mark and recapture method:

We want to estimate the number N of fish in a pond.

Capture k fish, mark them, and release them back into the pond. Then capture k more random fish. The number of recaptured fish that are marked is a random variable X.

Example: Mark and recapture

- What assumptions are we making if we assume that $k/N \approx X/k$?
- We can use $\hat{N} = k^2/X$ as an estimator for the number of total fish in the pond.
- The sampling distribution of $\hat{N}$ is then $\textbf{Hypergeometric}(N, k, k)$
- This means that $\mathbb{E}[\hat{N}] = k^2/(k^2/N) = N$ and $\textbf{Var}(\hat{N})$

$$\implies \hat{N} \textbf{ is an *unbiased* estimator for } N.$$

Desirable estimators:

We saw that an estimator $\hat{\theta}$ for a population parameter θ can theoretically be *any* well-defined sample statistic $h(X_1, \ldots, X_n)$, but what makes an estimator more desirable than another one?

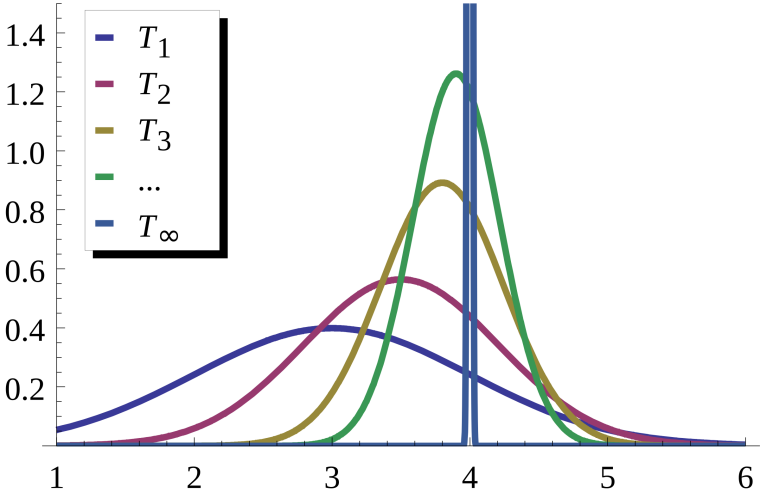
- Consistent: $\hat{\theta} \xrightarrow{n} \theta$ in the sense of the WLLN.
- Unbiased: $\mathbb{E}[\hat{\theta}] = \theta$.
- Low Variance: we might prefer $\hat{\theta}_1$ over $\hat{\theta}_2$ if $\mathbf{Var}(\hat{\theta}_1) < \mathbf{Var}(\hat{\theta}_2)$.

> These are all qualities that can be derived from the sampling distribution.

Consistency:

WLLN: the sample mean *converges in probability* $\bar{X}_n \xrightarrow{p} \mu$. Consistency can be defined for general estimators:

$$\lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \to 0.$$



Consistency II:

The WLLN essentially states that the sample mean is a consistent estimator for the population mean.

Assuming normally distributed data, is $\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ a consistent estimator for σ^2?

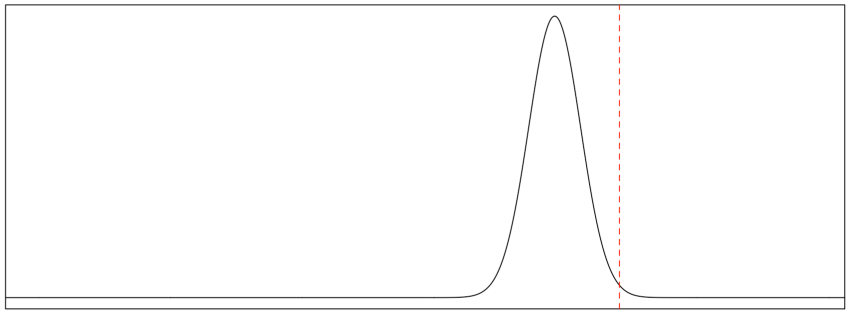
Proof.

Unbiasedness:

How far off will the estimator be from its true parameter value, on average.

We say an estimator $\hat{\theta}$ is *unbiased* if $\mathbb{E}[\hat{\theta}] = \theta$.

The *bias* of an estimator is then $\left|\mathbb{E}[\hat{\theta}] - \theta\right|$ (is this known in practice?).



Example: Standard deviation

What is wrong with $\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ as an estimator for σ^2?

By the law of total variance,

$$\mathbf{Var}(X) = \mathbb{E}[\mathbf{Var}(X|\bar{X})] + \mathbf{Var}(\mathbb{E}[X|\bar{X}])$$

Unbiasedness II:

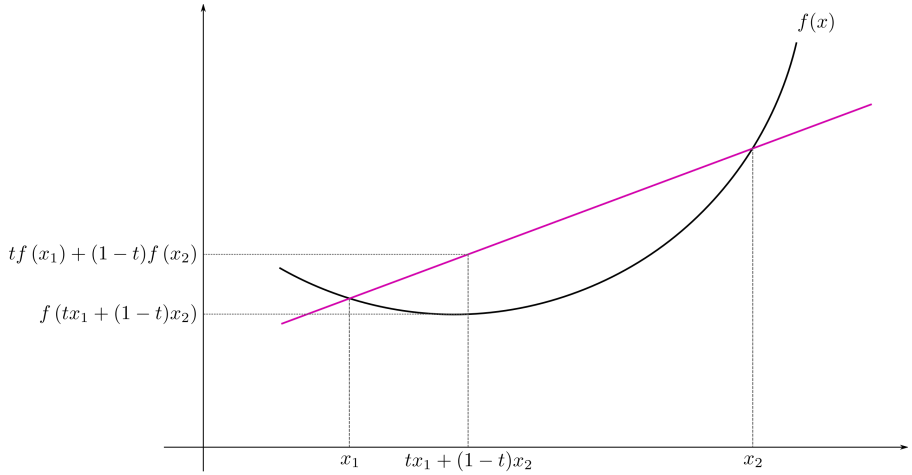
Suppose we now want to estimate a function of a parameter θ:

- Consider estimating θ^2, but note that $f(x) = x^2$ is a *convex* function.

For two points, if for any $t \in [0, 1]$ and any $x, y \in \mathbb{R}$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

then we say f is convex.



Jensen's inequality

- Expectations are convex combinations

$$\mathbb{E}[X] = \sum_{i=1}^{n} x\mathbb{P}(X = x) \qquad (\textbf{discrete case}).$$

- If f is convex, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$
- If f is concave, $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$
- $\implies$ $\sqrt{S^2}$ is a *biased* estimator of σ.

Example: Poisson probability

Suppose we model Atlantic tropical cyclones using $X_i \sim \mathbf{Pois}(\lambda)$ distribution, and an insurance company asks us to estimate the probability of there being no cyclones in the next 2 years:

$$\mathbb{P}(X = 0)^2 = e^{-2\lambda}.$$

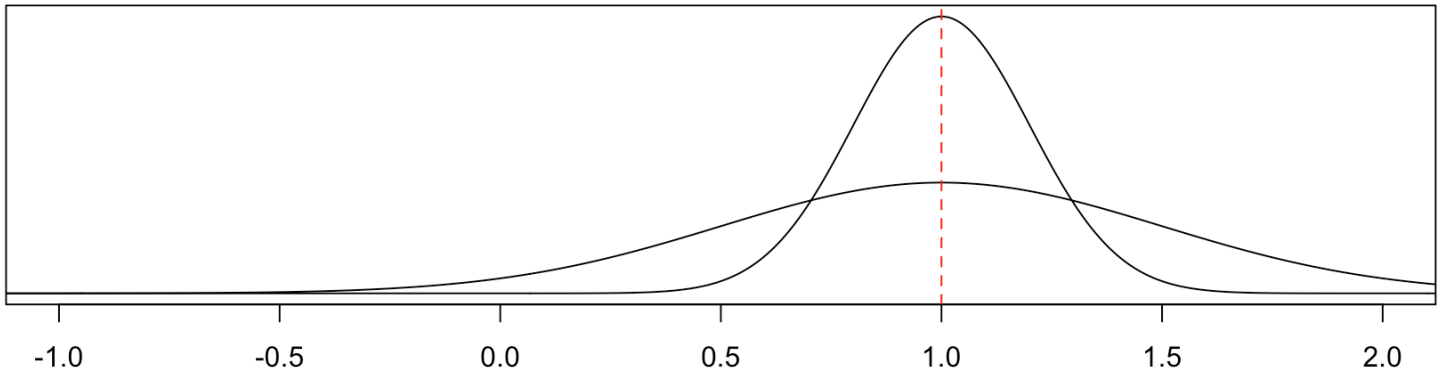
An unbiased estimator $T(X)$ would satisfy:

$$\mathbb{E}[T(X)] = \sum_{k \geq 0} T(k) \frac{\lambda^k e^{-\lambda}}{k!} = e^{-2\lambda}.$$

Efficiency of estimators

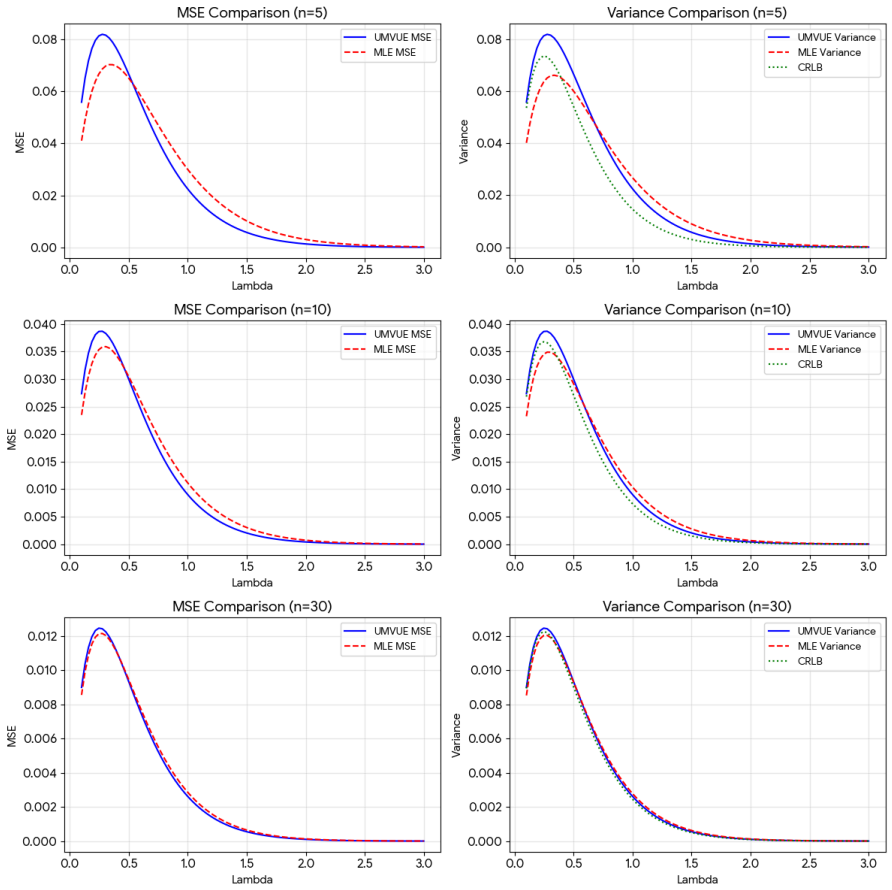
We say an estimator is *more efficient* if needs fewer input data or observations than a less efficient one to achieve a certain level of variance.

The *relative efficiency* between two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ is $\frac{\mathbf{Var}(\hat{\theta}_1)}{\mathbf{Var}(\hat{\theta}_2)}$.



Estimator $\hat{\theta}_2$ is called *more efficient* than estimator $\hat{\theta}_1$ if $\mathbf{Var}(\hat{\theta}_2) < \mathbf{Var}(\hat{\theta}_1)$, irrespective of the value of θ.

Efficiency of estimators II



Here we compare the variance of two estimators for $\mathbb{P}(X=0)^2 = e^{-2\lambda}$ from our Poisson example:

$$\hat{\theta}_1 = e^{-2\bar{X}},$$

$$\hat{\theta}_2 = \left(\frac{n-2}{n}\right)^{\sum_{i=1}^{n} X_i}.$$

> Next week: **efficient estimators** and the **Cramér-Rao Lower Bound**.
>
> **Spoiler**: we will be able to systematically find a class of estimators that are unbiased* and have the minimum variance possible (under certain conditions).

Assessing the quality of an estimator

We want our estimator $\hat{\theta}$ to be close to θ. If these are real numbers, then a common quality metric is the *Mean Squared Error*:

$$\begin{aligned}\mathbf{MSE}(\hat{\theta}, \theta) &= \mathbb{E}[(\hat{\theta} - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2] \\ &= \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + 2\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]](\mathbb{E}[\hat{\theta}] - \theta) + (\mathbb{E}[\hat{\theta}] - \theta)^2 \\ &= \mathbf{Var}(\hat{\theta}) + (\mathbb{E}[\hat{\theta}] - \theta)^2\end{aligned}$$

Example: Poisson Probability II

Suppose we again model Atlantic tropical cyclones using $X_i \sim \mathbf{Pois}(\lambda)$ and we now want to estimate the probability of there being no cyclones next year:

$$\mathbb{P}(X = 0) = e^{-\lambda},$$

for which we have two choices of estimators:

- $\hat{p}_U = \frac{\#\{X_i = 0\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i = 0)$
- $\hat{p}_B = e^{-\bar{X}}$

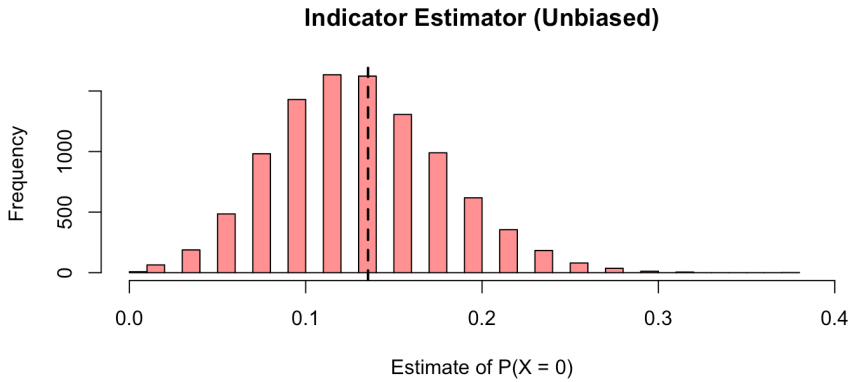
- Unbiased, $\mathbf{Var}(\hat{p}_U) = \frac{e^{-\lambda}(1-e^{-\lambda})}{n}$.
- Biased, $\mathbf{Var}(\hat{p}_B) = \frac{\lambda e^{-2\lambda}}{n}$.

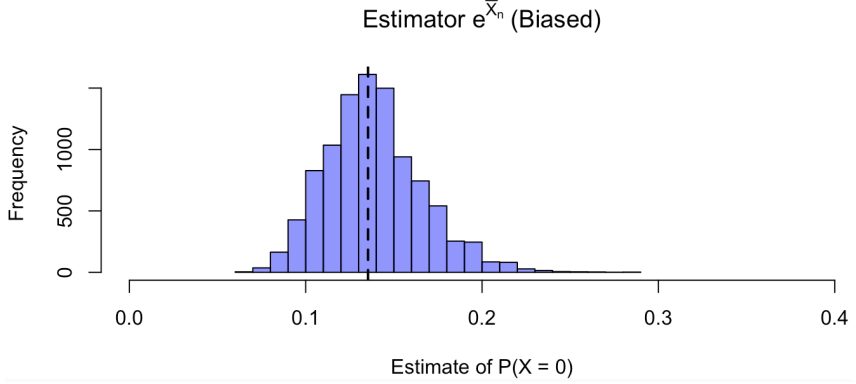
The relative efficiency of the two estimators is $\frac{\mathbf{Var}(\hat{p}_B)}{\mathbf{Var}(\hat{p}_U)} = \frac{\lambda e^{-\lambda}}{1-e^{-\lambda}}$.

Intuition: what happens if we observe a realization $(0, 0, 100)$?

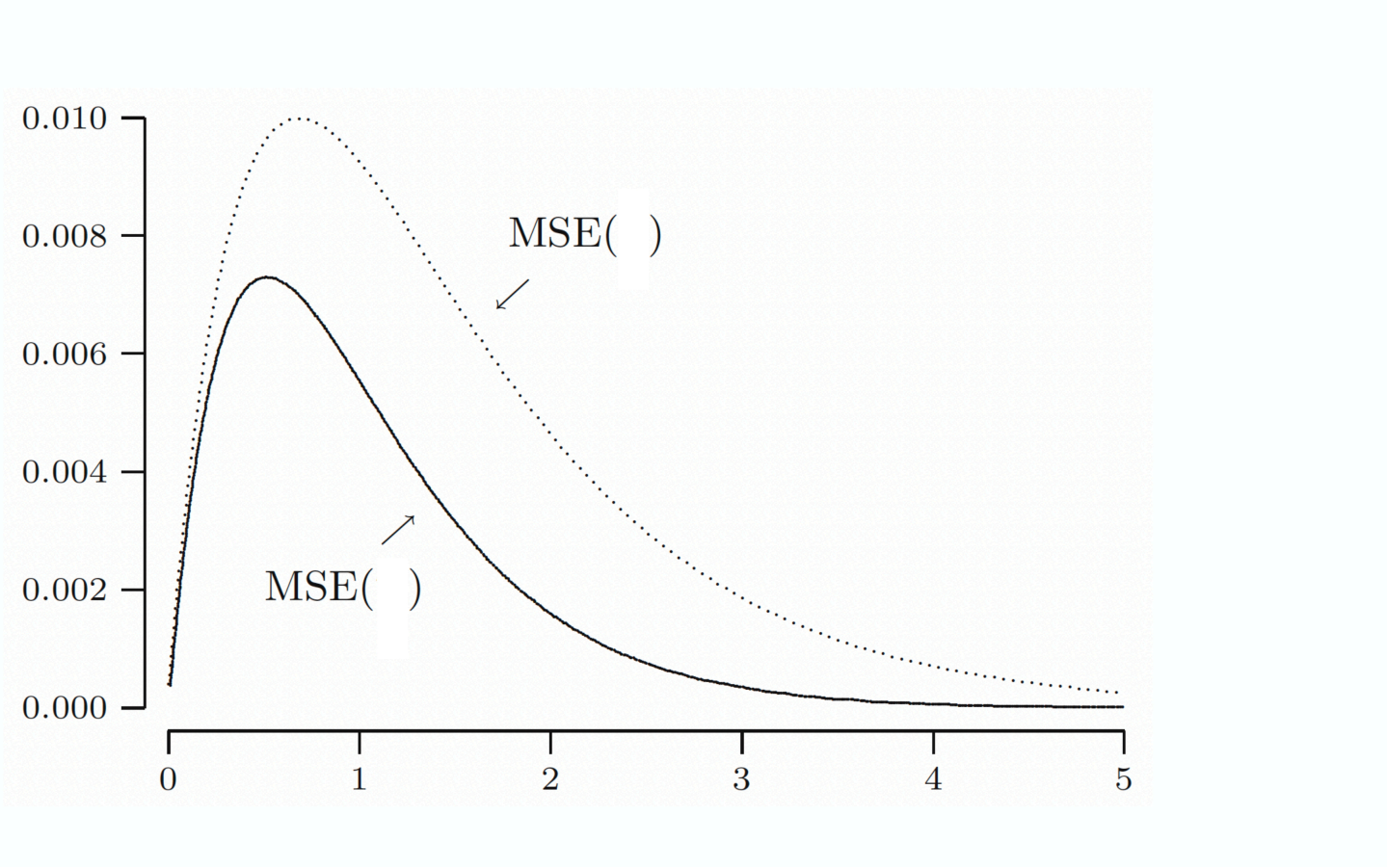
Example: Poisson Probability II

```r
set.seed(238)
n_sims <- 10000     # Number of experiments
n <- 50             # Sample size per experiment
lambda <- 2         # True lambda
true_p0 <- exp(-lambda) # True Probability P(X=0)

# Vectors to store our estimates
est_indicator <- numeric(n_sims)
est_mle <- numeric(n_sims)

# Simulation
for(i in 1:n_sims) {
  x <- rpois(n, lambda)
  # Estimator 1: The Indicator (Count zeros / n)
  est_indicator[i] <- mean(x == 0)
  # Estimator 2: The MLE (e^-mean)
  est_mle[i] <- exp(-mean(x))
}

calc_mse <- function(estimates, truth)
  { mean((estimates - truth)^2) }
```





Example: Poisson Probability MSE



MSE Comparison from Dekking et al. (MIPS) Fig. 20.3.

Recap Quiz

Let $X_i = T + Y_i$ be iid. readings from a thermometer. Here, T is the true temperature and Y_i is a random measurement error with $\mathbb{E}[Y_i] = b > 0$, where b is unknown. Let $X_1, \ldots, X_n$ be a random sample.

Which of the following would be a consistent estimator for T?

a. $\bar{X}_n$,

b. $\bar{X}_n - b$,

c. $X_1 - b$,

d. There are no consistent estimators for T.

Recap Quiz

Let $X_1, \ldots, X_n$ be a random sample with $\mathbb{E}[X_i] = \mu$, $\mathbb{E}|X_i| < \infty$. Consider the following estimator for μ:

$$\hat{\mu}_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n} X_i.$$

Which of the following statements is true regarding this estimator?

a. The WLLN does not apply here because the sum is divided by $n+1$ instead of n.

b. Since $\mathbb{E}[\hat{\mu}_{n+1}] \neq \mu$, the WLLN does not apply and $\hat{\mu}_{n+1}$ is not a consistent estimator for μ.

c. $\hat{\mu}_{n+1}$ is a consistent estimator only when $\mu = 0$.

d. $\hat{\mu}_{n+1}$ is a consistent estimator for any μ.

Recap Quiz

Which of the following statements about estimators is true:

a. Unbiasedness implies consistency

b. Consistency implies unbiasedness

c. Unbiasedness implies consistency and consistency implies unbiasedness

d. Neither implies the other.

Summary

- We have examined estimators and some of their properties: (consistency, unbiasedness, variance).
- We have learned to assess the quality of an estimator in terms of the Mean Squared Error.
- To assess the qualities of an estimator, one can study its sampling distribution.
- We've shown how to show whether an estimator is unbiased and how Jensen's Inequality can help with this task.
- We have seen how there can be a tradeoff between bias and variance of an estimator.