

Math 4261: Mathematical Statistics I

Lecture 8: Quantile estimation II and Parameter estimation

References: HMC 4.4, 5.1–5.3

DeGroot–Schervish 7.8–7.9, 10.6–10.7

Wasserman, *All of Statistics*, 7.1–7.2

Rice 8.1–8.2; Many ideas and figures credited to Zhou Fan

Fall 2026

Outline

Statistical analysis of quantile estimators

A motivating example

Parameter estimation

Bias variance decomposition

Recap: Order statistics

Let $X_1, \dots, X_n$ be random variables. Their **order statistics** are

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}.$$

Suppose X_i 's i.i.d. with CDF F . To find the distribution of $X_{(k)}$, consider

► $N_x = \sum_{i=1}^n \mathbf{1}\{X_i \leq x\} \sim \text{Bin}(n, F(x)).$

► $\{X_{(k)} \leq x\} = \{N_x \geq k\}$

Distribution of k -th order statistic

$$\mathbb{P}[X_{(k)} \leq x] = \sum_{j=k}^n \binom{n}{j} F(x)^j [1 - F(x)]^{n-j}.$$

Recap: Estimating quantiles

Aside from understanding the population mean and variance, we may want to learn a **population cutoff**, such as the median or the 90th percentile.

Idea: estimate cutoff using the corresponding **ranked observation** in the sample. I.e. order statistics.

Example: Uniform(0, θ)

For $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} \text{Unif}(0, \theta)$ with $\theta > 0$, we estimated $q_1 = \theta$ by $X_{(n)}$. The sampling distribution is

$$\mathbb{P}[X_{(n)} \leq x] = \left(\frac{x}{\theta}\right)^n, \quad 0 \leq x \leq \theta.$$

This allows us to show that $X_{(n)}$ is consistent for θ .

Recap: Estimating quantiles

More generally, to estimate the population quantile for $0 < \alpha < 1$,¹

$$q_\alpha = F^{-1}(\alpha) = \inf\{x : F(x) \geq \alpha\},$$

we use

Plug-in quantile estimator

$$\hat{q}_{n,\alpha} = \inf\{x : \hat{F}_n(x) \geq \alpha\} = X_{(\lceil n\alpha \rceil)}.$$

► $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_i \leq x\}.$

► Scanning from the left, $\hat{q}_{n,\alpha}$ is the first place $\hat{F}_n$ reaches α .

E.g. For 1, 2, 2, 4, 7, 9, our sample median is $\hat{q}_{6,0.5} = X_{(3)} = 2$ (the lower median for even n).

Today: Is this estimator consistent? What can we say about its sampling distribution?

¹The $\alpha = 0, 1$ cases may require separate treatment. 

Outline

Statistical analysis of quantile estimators

A motivating example

Parameter estimation

Bias variance decomposition

Statistical properties of the empirical CDF

For fixed x , the empirical CDF $\widehat{F}_n(x)$ is an **average of i.i.d. Bernoulli** variables with success probability $F(x)$ so that

$$\mathbb{E}[\widehat{F}_n(x)] = F(x), \quad \text{Var}[\widehat{F}_n(x)] = \frac{F(x)[1 - F(x)]}{n}.$$

Consistency and Asymptotic normality

As $n \rightarrow \infty$,

$$\widehat{F}_n(x) \xrightarrow{p} F(x).$$

If $0 < F(x) < 1$, then

$$\sqrt{n}[\widehat{F}_n(x) - F(x)] \xrightarrow{d} \mathcal{N}(0, F(x)[1 - F(x)]).$$

These follow directly from LLN and CLT.

Consistency of sample quantiles

Fix $0 < \alpha < 1$ and let $q_\alpha = F^{-1}(\alpha)$.

Quantile consistency

Suppose $X_1, \dots, X_n$ are i.i.d. and, for every $\epsilon > 0$,

$$F(q_\alpha - \epsilon) < \alpha < F(q_\alpha + \epsilon).$$

Then the empirical-inverse sample quantile satisfies

$$\hat{q}_{n,\alpha} \xrightarrow{p} q_\alpha.$$

- ▶ The strict inequalities let the LLN keep the empirical CDF below α on the left and above α on the right, **trapping its first crossing** between the two thresholds.
- ▶ A continuous, positive density near q_α is one sufficient condition.

Why this condition?

Suppose $X_1, \dots, X_n$ are i.i.d., with $\mathbb{P}[X_i = 2] = \mathbb{P}[X_i = 4] = 1/2$.

Under our convention, the population median is $q_{0.5} = 2$: the CDF first reaches 0.5 at 2, then stays flat until 4.

Two possible samples of size 3:

Sample	Sample median $\hat{q}_{3,0.5}$
2, 2, 4	2
2, 4, 4	4

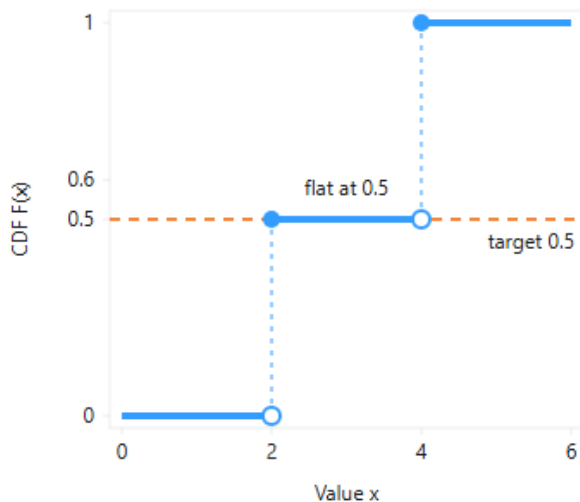
For any odd n , the sample median is whichever value occurs more often. By symmetry, each outcome has probability $1/2$.

Even with 1,000,001 observations, there is still a 50% chance that the sample median is 4.

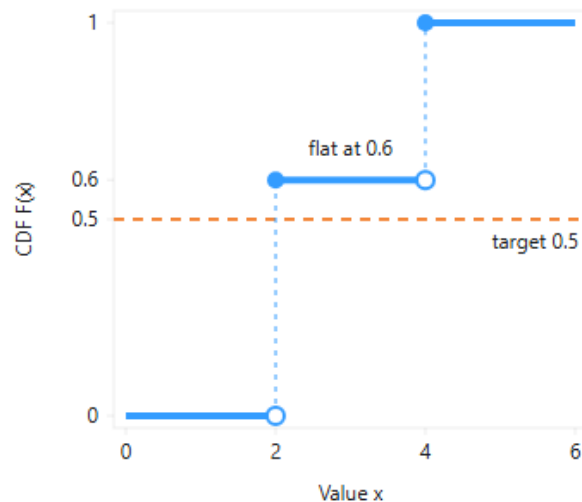
The chance of being 2 units from the target does not vanish, hence the sample median is **not consistent** for $q_{0.5} = 2$.

Why this condition?

50% at 2, 50% at 4



60% at 2, 40% at 4



On the right: $\mathbb{P}[X_i = 2] = 0.6$, $\mathbb{P}[X_i = 4] = 0.4$, and $q_{0.5} = 2$.

By the LLN, the fraction of twos approaches $0.6 > 0.5$. Thus the sample median equals 2 with probability tending to one: **consistency holds**.

Proof: Bracket the quantile with two thresholds

Fix $\epsilon > 0$, and write $a = q_\alpha - \epsilon$ and $b = q_\alpha + \epsilon$. The assumption gives $F(a) < \alpha < F(b)$.

1. **Apply the LLN** at the two fixed endpoints.

$$\widehat{F}_n(a) \xrightarrow{p} F(a) < \alpha, \quad \widehat{F}_n(b) \xrightarrow{p} F(b) > \alpha.$$

The limits are strictly below and above α . Thus, with probability tending to one, $\widehat{F}_n(a) < \alpha < \widehat{F}_n(b)$.

2. **Locate the first crossing.** On the event $\widehat{F}_n(a) < \alpha < \widehat{F}_n(b)$, the empirical CDF has not reached α at a , but has passed it by b . On every such sample, monotonicity gives

$$a < \widehat{q}_{n,\alpha} \leq b, \quad \text{so} \quad |\widehat{q}_{n,\alpha} - q_\alpha| \leq \epsilon.$$

Step 1 says this event has probability tending to one. Therefore,

$$\mathbb{P}[|\widehat{q}_{n,\alpha} - q_\alpha| \leq \epsilon] \longrightarrow 1.$$

This holds for every $\epsilon > 0$, proving consistency.

A CLT for sample quantiles

Fix $0 < \alpha < 1$. Suppose F has a continuous density f near q_α , and that $f(q_\alpha) > 0$.

Asymptotic normality of a sample quantile

For an i.i.d. sample,

$$\sqrt{n}(\hat{q}_{n,\alpha} - q_\alpha) \xrightarrow{d} \mathcal{N}\left(0, \frac{\alpha(1-\alpha)}{f(q_\alpha)^2}\right).$$

Thus a large-sample approximation is

$$\hat{q}_{n,\alpha} \approx \mathcal{N}\left(q_\alpha, \frac{\alpha(1-\alpha)}{n f(q_\alpha)^2}\right).$$

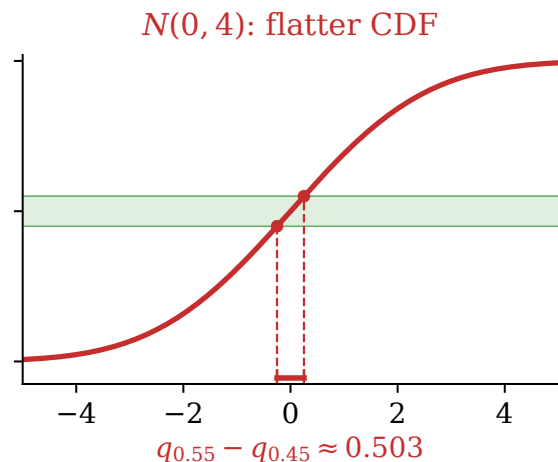
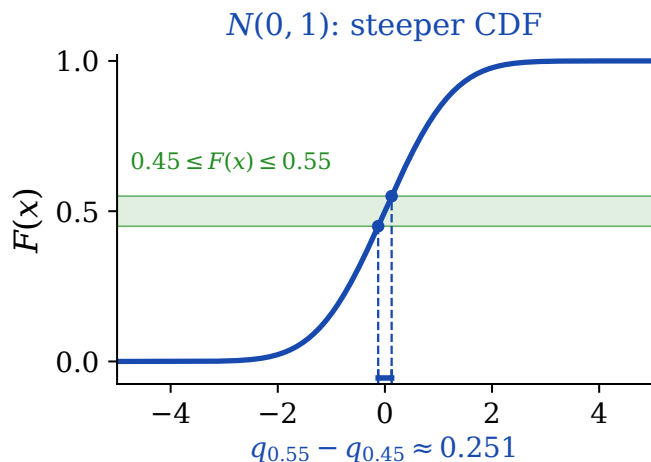
More observations n improve accuracy. So does a higher density of observations near the quantile (f is larger).

To use this approximation, we may need an estimate for $f(q_\alpha)$.

Why does the density appear in the denominator?

Based on a Taylor expansion,

$$\underbrace{\hat{q}_{n,\alpha} - q_\alpha}_{\text{quantile error (horizontal)}} \approx -\frac{1}{f(q_\alpha)} \underbrace{[\hat{F}_n(q_\alpha) - \alpha]}_{\text{CDF error (vertical)}}.$$



When the CDF is flatter near q_α , a small error in the estimated CDF can move its first α crossing further left or right. Thus the quantile estimate is less precise.

Example: Gaussian median

Let $X_1, \dots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$, with $\sigma > 0$. The population mean and median are both μ , and

$$q_{0.5} = \mu, \quad f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}.$$

The **quantile CLT** gives

$$\sqrt{n}(\hat{q}_{n,0.5} - \mu) \xrightarrow{d} \mathcal{N}\left(0, \frac{\pi\sigma^2}{2}\right).$$

Compare the fluctuations of these estimators about μ :

$$\begin{array}{l|l} \text{Sample mean} & \sigma/\sqrt{n} \\ \text{Sample median} & \sqrt{\pi/2} \cdot \sigma/\sqrt{n} \end{array}$$

So both consistently estimate μ . But the median's asymptotic standard error is about 1.25 times the mean's.

Example: Gaussian median

How accurate is the sample median for μ ?

Suppose $n = 100$ and $X_i \sim \mathcal{N}(\mu, 1)$ independently. The sample median is approximately

$$\hat{q}_{100,0.5} \approx \mathcal{N}\left(\mu, \frac{\pi}{200}\right).$$

With Φ the standard Gaussian CDF,

$$\begin{aligned}\mathbb{P}\left[|\hat{q}_{100,0.5} - \mu| \leq 0.1\right] &\approx 2\Phi\left(\frac{0.1}{\sqrt{\pi/200}}\right) - 1 \\ &\approx \mathbf{0.575}.\end{aligned}$$

For comparison, the sample mean is slightly better. It has the exact law $\bar{X}_{100} \sim \mathcal{N}(\mu, 1/100)$, giving

$$\mathbb{P}\left[|\bar{X}_{100} - \mu| \leq 0.1\right] = 2\Phi(1) - 1 \approx \mathbf{0.683}.$$

Outline

Statistical analysis of quantile estimators

A motivating example

Parameter estimation

Bias variance decomposition

Motivation

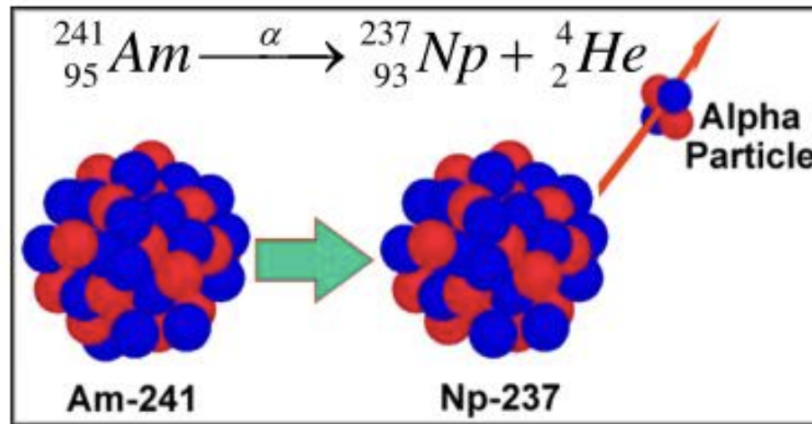
The goal of **parameter estimation** is to *fit* a probability distribution to the data.

Advantages

- ▶ Prediction: allows us to predict where a new data point may lie.
Example: hydrologists use stochastic models of rainfall in simulations to plan uses of water resources.
- ▶ Scientific theory may suggest a type of probability law appropriate for modeling. The parameters may be of direct interest to the scientific investigation.
- ▶ Many distributions, such as $\text{Pois}(\lambda)$ and $\mathcal{N}(\mu, \sigma^2)$, depend on only a **small** number of parameters which we can fit to the data.

Example: Alpha particle emissions

- ▶ Radioactive sources emits alpha particles
- ▶ Number of emissions in one unit of time (say 10 seconds) fluctuates seemingly randomly.
- ▶ Poisson model for radioactive decay appropriate because
 - ▶ Emitted particles come from a very large number of independent sources: atoms. n very large.
 - ▶ Particles emitted rarely (p is small)



Example: Alpha particle emissions

Berkson 1966 studied NIST data of particle emissions of **americium-241**. The following are counts of particles emitted in 1207 intervals each of 10 seconds. The mean emissions rate was **0.8392 emissions per second**.

# particles emitted	# intervals
0-2	18
3	28
4	56
5	105
6	126
7	146
8	164
9	161
10	123
11	101
12	74
13	53
14	23
15	15
16	9
17+	5
	1207

Fitting a Poisson

View above as 1207 independent realizations of Poisson r.v.'s.
Each having PMF

$$\mathbb{P}[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}$$

The average number of emissions in a 10 second interval is 8.392.

Since $\mathbb{E}[\text{Pois}(\lambda)] = \lambda$, take

$$\hat{\lambda} = 8.392$$

How well does it fit?

# particles emitted	# intervals	Expected # under i.i.d. $X_i \sim \text{Pois}(8.392)$
0-2	18	12.2
3	28	27.0
4	56	56.5
5	105	94.9
6	126	132.7
7	146	159.1
8	164	166.9
9	161	155.6
10	123	130.6
11	101	99.7
12	74	69.7
13	53	45.0
14	23	27.0
15	15	15.1
16	9	7.9
17+	5	7.1
	1207	1207

$\leftarrow E[\text{Bin}(1207, p_3)]$,
 where
 $p_3 = P[\text{Pois}(8.392) = 3]$

Qualitatively fits well. Quantitatively? Will discuss later.

Outline

Statistical analysis of quantile estimators

A motivating example

Parameter estimation

Bias variance decomposition

Choosing the model

Our choice of model may depend on many factors, including:

- ▶ What the data values represent. (Are they discrete or continuous measurements? Can they be negative?)
- ▶ Our understanding of the generative process for the data.
- ▶ Exploratory analysis and visual examination of the data.
- ▶ Considerations of computational time and cost.
- ▶ Considerations of how many parameters we can accurately learn given the amount of data that we have.
- ▶ Considerations of predictive accuracy, if the model is to make predictions on new unseen examples.

In the next few weeks, we will study the question: for

$$X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x \mid \theta),$$

how can we estimate the unknown parameter θ ?

Parameter estimation

Framework

- ▶ Regard observed data as realizations of r.v.'s $X_1, \dots, X_n$ that are (usually) i.i.d. Each X_i has **distribution**

$$f(x|\theta) \quad (\text{either PMF or PDF})$$

- ▶ $\theta \in \Omega \subseteq \mathbb{R}^k$ is a **parameter vector**. The **parameter space** Ω is the set of allowable values.

$$\text{Poisson: } \Omega = (0, \infty),$$

$$\text{Normal: } \theta = (\mu, \sigma^2), \quad \Omega = \mathbb{R} \times (0, \infty).$$

- ▶ **Joint distribution** of $X_1, \dots, X_n$ is

$$f(x_1|\theta) \cdot f(x_2|\theta) \cdots f(x_n|\theta)$$

- ▶ An **estimator** $\hat{\theta}$ is a function of the data

$$\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$$

Therefore, $\hat{\theta}$ is a random variable, with randomness arising from joint distribution of the X_i 's.

Example: Gaussian data

in practice we don't know it

Suppose $X_1, \dots, X_n$ are i.i.d. $\mathcal{N}(10, 3^2)$. Write the parameter vector and their joint PDF.

Param. vector $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} 10 \\ 3^2 \end{bmatrix}$

Each X_i has dist. (PDF)

$$f(x_i | \theta) = \frac{1}{\sqrt{2\pi} \cdot 3} \exp \left\{ -\frac{1}{2 \cdot 3^2} (x_i - 10)^2 \right\}$$

Their joint dist has PDF

$$\prod_{i=1}^n f(x_i | \theta)$$

Perspective: Statistics vs Probability

Probability

Suppose $X_1, \dots, X_n \sim \mathcal{N}(10, 3^2)$ i.i.d.

Q: What does $X_1, \dots, X_n$ typically look like?

$$\text{E.g. } \mathbb{P}[|\bar{X} - 10| > \varepsilon] \text{ ?}$$

Statistics

Suppose that $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. We observe

$$X_1, X_2, \dots, X_n = 9.7, 11.3, 9.32, \dots, 12.7$$

Q: Can we tell that $\mu = 10, \sigma = 3$?

Outline

Statistical analysis of quantile estimators

A motivating example

Parameter estimation

Bias variance decomposition

Bias and variance

Consider a parameter $\theta \in \mathbb{R}$. Any estimator $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ is a statistic and has variability due to the randomness of the data $X_1, \dots, X_n$.

If $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} f(x \mid \theta)$ with true parameter θ , we can measure the accuracy of $\hat{\theta}$ via its bias and variance:

- ▶ The **bias** of $\hat{\theta}$ is $\mathbb{E}_\theta[\hat{\theta}] - \theta = \mathbb{E}_\theta[\hat{\theta}(X_1, \dots, X_n)] - \theta$.
- ▶ The **variance** of $\hat{\theta}$

$$\text{Var}_\theta[\hat{\theta}] = \text{Var}_\theta[\hat{\theta}(X_1, \dots, X_n)].$$

The **standard error** of $\hat{\theta}$ is the standard deviation $\sqrt{\text{Var}_\theta[\hat{\theta}]}$.

Here $\mathbb{E}_\theta$ and Var_θ are computed assuming the true model under θ .

Unbiasedness

Definition

Let Ω be the parameter space. Say that $\hat{\theta}$ is **unbiased** for θ if

$$\mathbb{E}_{\theta}[\hat{\theta}] = \theta \quad \text{for every } \theta \in \Omega.$$

- ▶ For instance, $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} N(\theta, 1)$ with $\Omega = \mathbb{R}$. The sample mean is indeed unbiased:

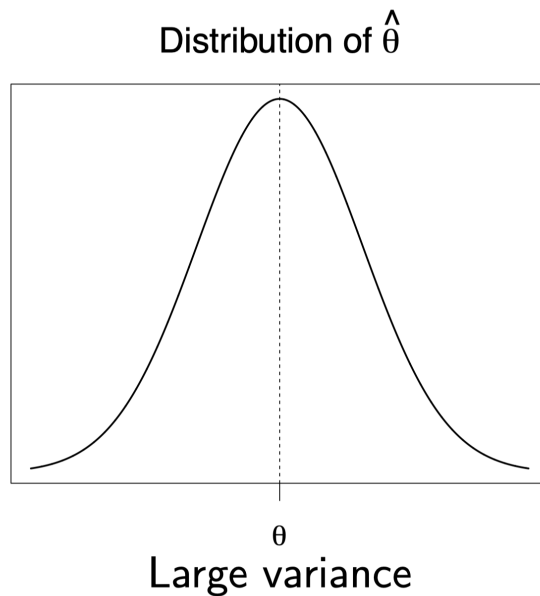
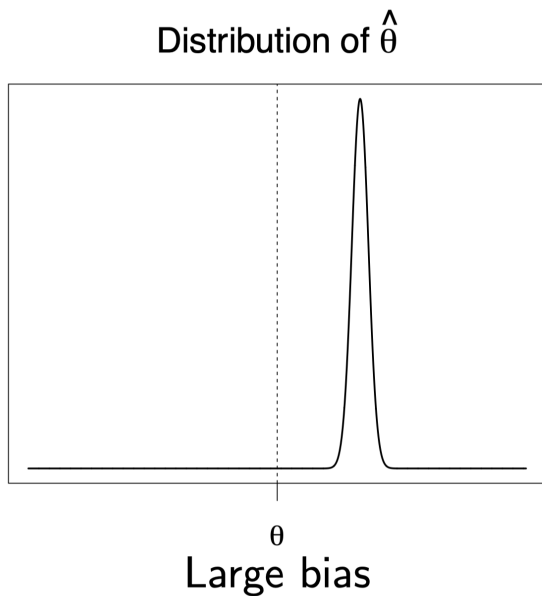
$$\mathbb{E}_{\theta}[\bar{X}_n] = \theta \quad \text{for every } \theta \in \mathbb{R}.$$

- ▶ Equality at just **one particular** θ is **not** enough. For instance $T_n = \bar{X}_n/2$ satisfies

$$\mathbb{E}_{\theta}[T_n] = \theta/2$$

which equals θ only when $\theta = 0$. But it is not an unbiased estimator for θ (over the whole parameter space).

Bias and variance



Bias measures how **close** the average value of $\hat{\theta}$ is to the true parameter θ . Variance measures how **variable** this estimate $\hat{\theta}$ is around its average value.

Example: Gaussian variance

Recall the empirical variance estimator for σ^2 :

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

We will show next that this estimator is **biased**:

$$\mathbb{E}[\hat{\sigma}^2] = \sigma^2 \left(\frac{n-1}{n} \right)$$

- ▶ But being unbiased is not the only criteria to be a good estimator.
- ▶ It is quite often the case that **biased estimators perform better** depending on what you mean by ‘better’.

Example: Gaussian variance

$$\begin{aligned}\text{Var } X &= \mathbb{E}[(X - \mathbb{E}X)^2] \\ &= \mathbb{E}[X^2] - (\mathbb{E}X)^2\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2\right] \\ &= \dots = \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^n X_i^2\right) - \bar{X}^2\right]\end{aligned}$$

$$\begin{aligned}\mathbb{E}[Z^2] &= \text{Var } Z \\ &\quad + (\mathbb{E}Z)^2\end{aligned}$$

$$= \frac{1}{n} \sum_{i=1}^n \underbrace{\mathbb{E}[X_i^2]}_{\sigma^2 + \mu^2} - \underbrace{\mathbb{E}[\bar{X}^2]}_{\frac{\sigma^2}{n} + \mu^2}$$

$$\begin{aligned}&= \cancel{\sigma^2} + \cancel{\mu^2} - \left(\frac{\sigma^2}{n} + \cancel{\mu^2}\right) = \sigma^2 \left(1 - \frac{1}{n}\right) \\ &= \sigma^2 \left(\frac{n-1}{n}\right)\end{aligned}$$