STAT 390 A
Statistical Methods in Engineering and Science
Week 7 Lectures – Part 1 – Spring 2023
Statistical Models and Point Estimation

Alexander Giessing

Department of Statistics

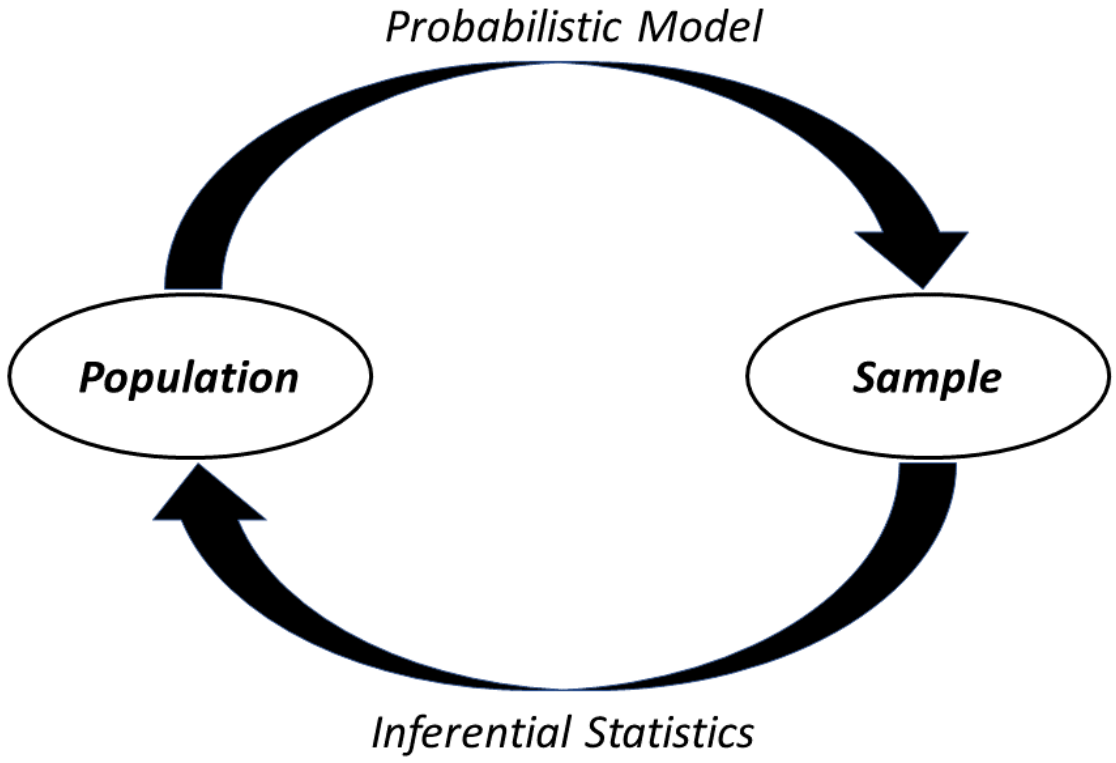
University of Washington

May 5, 2023

Outline

1. Statistical Models
2. Basic Concepts of Estimation
3. Comparing Different Estimators

Probability and Statistics



- **Probabilistic models** describe how samples are drawn from a population.
- **Inferential statistics** uses random samples and probabilistic models to learn about properties of a population.

Example: Defective items in a shipment

Suppose that there is a shipment of N flashlights. An unknown number θN for $0 < \theta < 1$ of the flashlights are defective. It is too expensive to examine all of the flashlights. How can we get information about θ, the percentage of defective flashlights?

- **Idea:** Draw a random sample of n flashlights with out replacement and let X denote the number of defective lights in this random sample. We know that $X \sim HyperGeo(N, \theta N, n)$, i.e.

$$P(X = k) = \frac{\binom{\theta N}{k}\binom{N-\theta N}{n-k}}{\binom{N}{n}},$$

  for $\max\{0, n - N + \theta N\} \leq k \leq \min\{\theta N, n\}$.

- **Differences to a probability model:**
  - The number θN is unknown; instead of a single pmf we have specified a "family" of pmfs $\{HyperGeo(N, \theta N, n) : 0 < \theta < 1\}$.
  - We are not interested in using this model to compute probabilities; instead, we want to learn about θ.

Example: Measurement errors

An experimenter makes n independent determinations of the value of a physical constant μ. Her measurements are subject to random fluctuations (errors) and the data can be thought of as μ plus/ minus some random errors. How can we get information about the "true" value μ?

- **Idea:** Let $X_1, X_2, \ldots X_n$ be the n measurements of μ. Then,

$$X_i = \mu + \varepsilon_i \qquad 1 \leq i \leq n,$$

  where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are the random measurement errors.

- To complete the description of the (joint) distribution of the X_i's we need to place assumptions on the (joint) distribution of the ε_i's, e.g.

$$\varepsilon_i \sim_{iid} N(0, \sigma^2) \qquad 1 \leq i \leq n.$$

- **Differences to a probability model:**
  - The numbers μ and σ^2 are unknown; instead of a single pdf we have specified a "family" of pdfs $\{N(\mu, \sigma^2) : \mu \in \mathbb{R},\ \sigma^2 > 0\}$.
  - We primarily interested in using the data to learn about μ.

Statistical model for repeated measurements

Statistical model for repeated measurements

A data set consisting of values $x_1, \ldots, x_n$ of repeated measurements of the same quantity is modeled as the realization of a random sample $X_1, \ldots, X_n$. The model may include a partial specification of the probability distribution of each X_i, $1 \leq i \leq n$.

- **Parametric models**: population pmf/ pdf is specified except for a parameter θ.
- **Nonparametric models:** population pmf/ pdf is unspecified, e.g. in the case of measurement errors it is often sufficient to assume the following:
  - The distr. of the ε_i's is independent of μ.
  - The ε_i's are independent and identically distributed.
  - The distr. of ε_i is continuous and symmetric about 0.

  (While these three assumptions characterize certain aspects of the distribution, they do not specify a pmf/ pdf indexed by a parameter.)

Statistical model for repeated measurements (Cont.)

- The probability distribution of each X_i is also called the **model distribution**.
- In parametric statistical models the model distribution is indexed by a **model parameter**.

  In above examples, model parameters are the percentage of defective flashlights and the pair (μ, σ^2).
- The unique distribution from which the sample actually originates is called the **"true" distribution** and the model parameter that indexes the "true" distribution is called the **"true" parameter**.

  I put "true" in quotation marks because it does not refer to something in the real world, but only to a distribution (or parameter) in the statistical model, which is merely an approximation of a real situation.
- The i.i.d. assumption on the data set (aka realization of a random sample) is strong and there exist situations in which this assumption is violated. In this course, however, we restrict ourselves to this scenario.

Example: Statistical model for a data set of coin tosses

Suppose that we have a data set with the recordings of 10 coin tosses. What is an appropriate statistical model and corresponding model distribution for this data set?

- The result of each coin toss can be modeled by a Bernoulli random variable taking values 1 (for heads) and 0 (for tails) with probability p and $1-p$, respectively. The probability p is the **model parameter**.
- It is reasonable to assume that the coin tosses did not influence each other. We therefore can assume that the 10 tosses in our data set are realizations of a random sample $X_1, \ldots X_{10}$ from a $Ber(p)$.
- If it is known that the **coin is fair**, we know that $p = 1/2$. In this case, the model distribution is fully specified and coincides with the "true" $Ber(1/2)$ distribution.
- If the **coin is possibly unfair**, then the "true" model distribution is $Ber(p)$ with a particular value for p, unknown to us. In this case, (one possible) **feature of interest** of the model distribution would be the probability p.

How to use statistical models

- Statistical models are only approximations to the real world; they provide a framework within which we can learn about the world via probabilistic tools.
- Statistical models allow us to ask specific questions about features of the model distribution (mean, variance, etc.).
- To effectively use statistical models, we need to answer the following questions:
  - Which feature of the model distribution represents the quantity of interest?

    In the two introductory examples, the model features of interest are the percentage of defective flashlights θ and the mean μ.
  - How do we use our data set to determine a value for a feature of the model distribution?
  - Which procedure for determining a value for a feature of a model distribution is best?
  - Which model distribution fits our data set best (aka model selection)?

Outline

1 Statistical Models

2 Basic Concepts of Estimation

3 Comparing Different Estimators

Statistics, estimators, and estimates

(Sample) Statistic

Let $X_1, \ldots X_n$ be a random sample. A (sample) statistic is function $T = h(X_1, \ldots X_n)$ that depends on the random sample $X_1, \ldots, X_n$ only.

- Note: A statistic is a random variable.

Estimator

Let $X_1, \ldots X_n$ be a random sample and θ be an (unknown) parameter. A statistic $T = h(X_1, \ldots, X_n)$ used to estimate θ is called an estimator for θ.

- The word "estimator" refers to the method or procedure for estimation.

Estimate

Let $x_1, \ldots, x_n$ be a realization of a random sample $X_1, \ldots, X_n$ and $T = h(X_1, \ldots X_n)$ an estimator for θ. Then $t = h(x_1, \ldots, x_n)$ is called an estimate of θ.

- An estimate is a realization of an estimator; it depends on the particular data set $x_1, \ldots, x_n$.

Example: Statistic or not?

Let $X_1, \ldots X_n$ be a random sample. Let θ be a unknown parameter of interest. Which of the following are statistics?

- $T_1(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i$.
  - Yes.
- $T_2(X_1, \ldots, X_n) = X_1$.
  - Yes.
- $T_3(X_1, \ldots, X_n) = X_1^2 + X_2 \times X_3$.
  - Yes.
- $T_4(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i - \theta$.
  - No, because T_4 depends on the unknown θ.

Example: An estimator for the mean

Let $X_1, \ldots X_n$ be a random sample. Suppose we are interested in the mean μ of the population from which the random sample was drawn. What is a sensible estimator for μ?

- An intuitive choice is the sample average,

$$T(X_1, \ldots, X_n) = \bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- How can we justify that T is indeed a sensible estimator based on what we have learnt in the first part of this course?

- Recall that if the X_i's have finite mean and variance, then by the WLLN, for all $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|\bar{X}_n - \mu| > \varepsilon) = 0.$$

  Thus, for large sample sizes n, the probability that the estimator T will return a value far away from μ is exceedingly small.

  $\Longrightarrow$ T is a reasonable estimator for μ (but not the only one!)

Example: An estimator for the CDF

Let $X_1, \dots X_n$ be a random sample. Suppose we are interested in the cdf F from which the random sample was drawn. How can we estimate F?

- Define the empirical (cumulative) distribution function as

$$F_n(a) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \in (-\infty, a]\} \qquad \text{for } a \in \mathbb{R},$$

  where $x \mapsto \mathbf{1}\{x \in C\}$ is the indicator function of set C, i.e. $= 1$ if $x \in C$ and $= 0$ if $x \notin C$.

- Note that the random variables $Y_1, \dots, Y_n$ defined as $Y_i = \mathbf{1}\{X_i \in (-\infty, a]\}$ are i.i.d. Bernoulli with parameter

$$p = P(Y_i = 1) = P(X_i \leq a) = F(a).$$

  Also, $\mathrm{E}[Y_i] = F(a)$, $\mathrm{Var}(Y_i) = F(a)\big(1 - F(a)\big)$, and $\mathrm{E}[F_n(a)] = F(a)$.

- Thus, by the WLLN applied to $Y_1, \dots Y_n$, for all $a \in \mathbb{R}$ and $\varepsilon > 0$,

$$\lim_{n \to \infty} P(|F_n(a) - F(a)| > \varepsilon) = 0.$$

  $\implies$ $F_n(a)$ is a reasonable estimator for $F(a)$.

Example: Choosing between different estimators

Let's consider the dielectric breakdown voltage for pieces of epoxy resin. Suppose that we have collected the following 20 observations:

24.46 25.61 26.25 26.42 26.66 27.15 27.31 27.54 27.74 27.94
27.98 28.04 28.28 28.49 28.50 28.87 29.11 29.13 29.50 30.88.

What would be reasonable estimate for the mean breakdown voltage μ for pieces of epoxy resin?

- $T_1(X_1, \ldots, X_n) = \bar{X}_n \implies t_1 = 555.86/20 = 27.793.$
- $T_2(X_1, \ldots, X_n) = X_6 \implies t_2 = 27.31.$
- $T_3(X_1, \ldots, X_n) = X_{\text{median}} \implies t_3 = (27.94 + 27.98)/2 = 27.960.$
- $T_4(X_1, \ldots, X_n) = \frac{1}{2}\left(\min_{1 \le i \le n} X_i + \max_{1 \le i \le n} X_i\right) \implies t_4 = 27.670.$

The estimates are all similar ... What can we do?

Example: Choosing between different estimators (Cont.)

Let's consider the dielectric breakdown voltage for pieces of epoxy resin. Suppose that we have collected the following 20 observations:

24.46 25.61 26.25 26.42 26.66 27.15 27.31 27.54 27.74 27.94
27.98 28.04 28.28 28.49 28.50 28.87 29.11 29.13 29.50 30.88.

- **Question:** Given a realization $x_1, \ldots x_n$, which estimate t_1, t_2, t_3, t_4, with $t_k = T_k(x_1, \ldots, x_n)$, is best?
  - Impossible to answer; question is ill-posed. If "best" means closest to μ, then the answer depends on the specific data set $x_1, \ldots x_n$. Therefore, this question will have different answers for different data sets.
- **Question:** Which estimator (aka procedure) T_1, T_2, T_3, T_4 is best?
  - Can be answered. T_1, T_2, T_3, T_4 are random variables, we can therefore analyze/ compare the associated pmfs/ pdfs and cdfs and make statements that will hold for a random sample $X_1, \ldots, X_n$ not just a specific data set $x_1, \ldots, x_n$.

Sampling distribution of an estimator

Sampling Distribution

Let $T = h(X_1, X_2, \ldots, X_n)$ be an estimator based on the random sample $X_1, X_2, \ldots, X_n$. The probability distribution of T is called the sampling distribution of T.

- The sampling distribution of an estimator can be very complicated if the estimator is a non-linear function of a random sample of size n. There are two remedies:
  - It is often sufficient to analyze the expected value and the variance of estimators in order to choose between competing ones.
  - If the sample size is large, one can apply the CLT, i.e. for large sample sizes the sampling distribution of an estimator can be approximated by a normal distribution.
  - If the sample size is small, then one can use the bootstrap method.

Outline

1 Statistical Models

2 Basic Concepts of Estimation

3 Comparing Different Estimators

Bias of an estimator

Unbiasedness

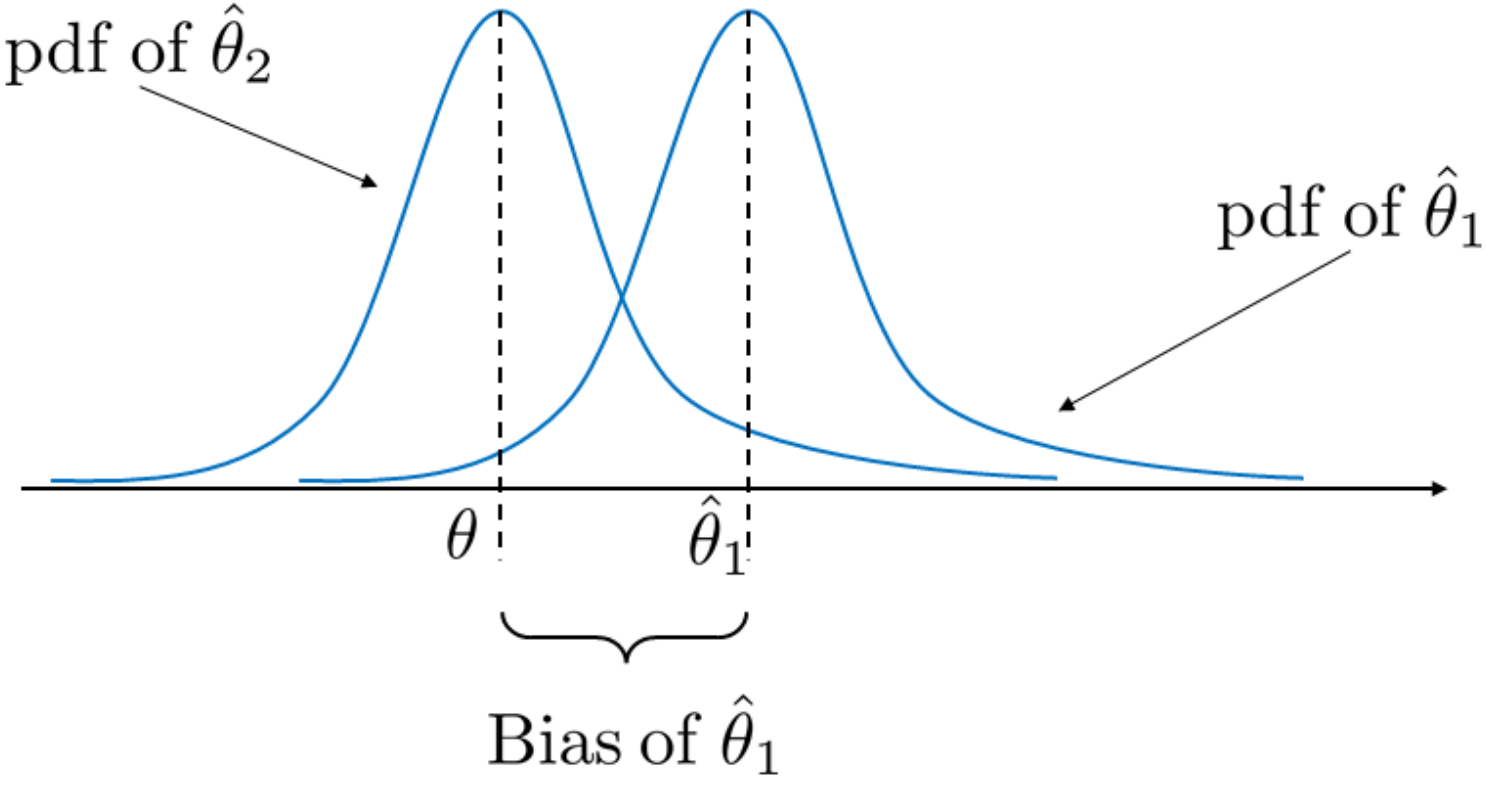
An estimator T is called an unbiased estimator for the parameter θ if $\mathrm{E}[T] = \theta$ irrespective of the value of θ.

- The difference $\mathrm{E}[T] - \theta$ is called the **bias** of T. If this difference is nonzero, then T is called biased.
- An estimator T for a parameter θ is unbiased if its probability distribution (i.e. its sampling distribution) is centered at the true value of the parameter θ.
- When choosing among several different estimators for a parameter θ, it is best to select the one that is unbiased.

  (Just like an experimenter who prefers a measurement device that is accurate over one with a systematic error (i.e. bias).)

Bias of an estimator (Cont.)

Suppose $\hat{\theta}_1, \hat{\theta}_2$ are estimators for a parameter θ. Below sketch shows a hypothetical case in which the sampling distribution of both estimators is symmetric with $\hat{\theta}_1$ biased and $\hat{\theta}_2$ unbiased.



Unbiased estimators for mean and variance

Unbiased estimators for mean and variance

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with finite expectation μ and finite variance σ^2. Then,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an unbiased estimator for μ and

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

is an unbiased estimator for σ^2.

- Two of the most common and simplest nonparametric estimators.
- Unbiasedness does not carry over to nonlinear functions, i.e. S_n is not unbiased for σ.

Unbiased estimators for mean and variance (Cont.)

We've already shown that $\mathrm{E}[\bar{X}_n] = \mu$. *We now show that* $\mathrm{E}[S_n^2] = \sigma^2$.

$$\begin{aligned}
\mathrm{E}[S_n^2] &= \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\mu)^2 + 2(X_i-\mu)(\mu-\bar{X}_n) + (\mu-\bar{X}_n)^2\right] \\
&= \frac{n}{n-1}\sigma^2 + \frac{n}{n-1}\mathrm{E}\left[(\mu-\bar{X}_n)^2\right] - \frac{2}{n-1}\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\mu)(\bar{X}_n-\mu)\right] \\
&\overset{(a)}{=} \frac{n+1}{n-1}\sigma^2 - \frac{2}{n-1}\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\mu)(\bar{X}_n-\mu)\right] \\
&= \frac{n+1}{n-1}\sigma^2 - \frac{2}{n-1}\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\mu)\left(\frac{1}{n}\sum_{j\neq i}(X_j-\mu)\right)\right] \\
&\quad - \frac{2}{n-1}\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\mu)\frac{1}{n}(X_i-\mu)\right] \\
&\overset{(b)}{=} \frac{n+1}{n-1}\sigma^2 - \frac{2}{n-1}\sigma^2 = \sigma^2.
\end{aligned}$$

where (a) and (b) hold because $X_1, \ldots, X_n$ are iid (verify at home!)

Standard deviation of an estimator (aka standard error)

Standard error

The standard deviation (SD) of an estimator $T = h(X_1, \ldots, X_n)$ based on a random sample $X_1, \ldots X_n$ is called the standard error (SE) of T, i.e.

$$SE(T) := \sqrt{\mathrm{E}[(T - \mathrm{E}[T])^2]}.$$

- Note/ Recall: The standard error of the sample mean $\bar{X}_n$ of random variables with variance σ^2 is $SD(\bar{X}_n) = \sigma/\sqrt{n}$.

Efficiency

Let T_1 and T_2 be two unbiased estimators for the same parameter θ. Then estimator T_2 is called more efficient than estimator T_1 if

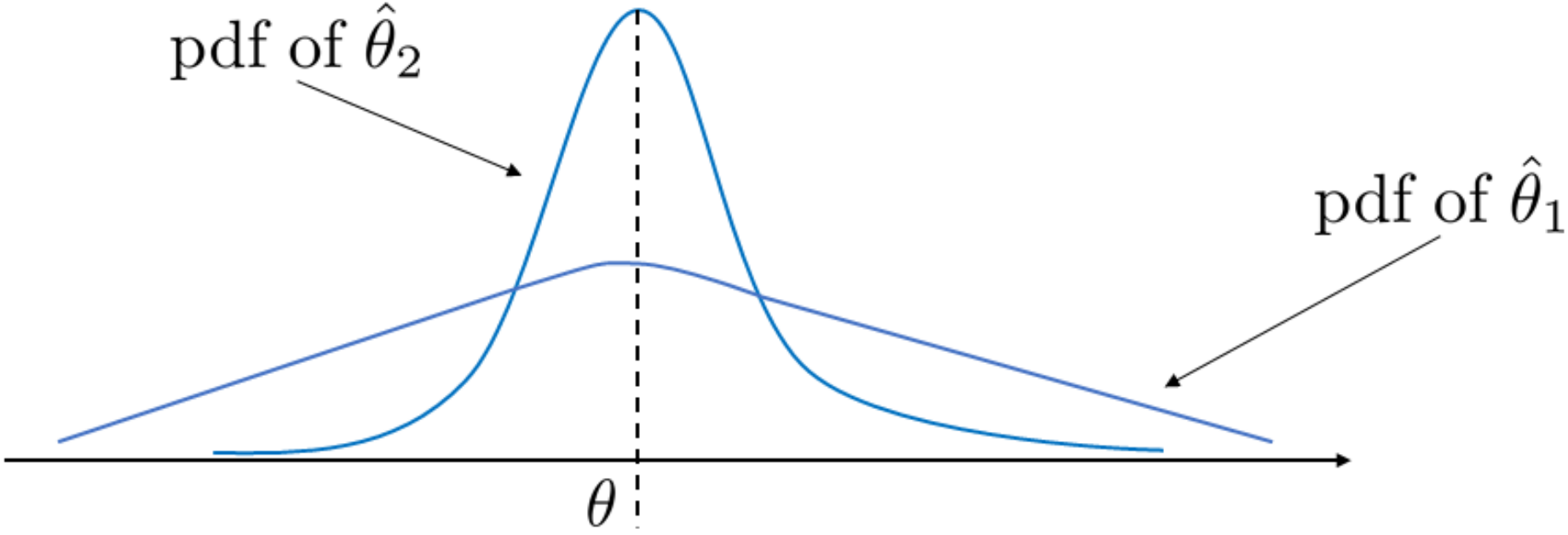
$$\mathrm{Var}(T_2) < \mathrm{Var}(T_1),$$

irrespective of the value of θ.

- When choosing among several unbiased estimators for a parameter θ, it is best to select the one that is most efficient.

Standard deviation of an estimator (Cont.)

Suppose $\hat{\theta}_1, \hat{\theta}_2$ are unbiased estimators for a parameter θ. Below sketch shows a hypothetical case in which $\hat{\theta}_2$ is more efficient than $\hat{\theta}_1$.



Example: Finding the more efficient estimator

Reconsider the data set on 20 measurements of dielectric breakdown voltage for pieces of epoxy resin:

24.46 25.61 26.25 26.42 26.66 27.15 27.31 27.54 27.74 27.94
27.98 28.04 28.28 28.49 28.50 28.87 29.11 29.13 29.50 30.88.

- Let's consider the following estimators of the mean μ:
  - $T_1(X_1, \ldots, X_n) = \bar{X}_n \implies t_1 = 555.86/20 = 27.793.$
  - $T_2(X_1, \ldots, X_n) = X_6 \implies t_2 = 27.31.$

- Both estimators are unbiased, i.e. $\mathrm{E}[T_1] = \mathrm{E}[T_2] = \mu$; however, T_1 is more efficient than T_2 since

$$\mathrm{Var}(T_1) = \frac{\sigma^2}{n} \leq \sigma^2 = \mathrm{Var}(T_2).$$

- Note: We do not use the data set to determine bias and efficiency since these are properties of the estimators (not the estimates!).

Mean squared error of an estimator

Most estimators are not unbiased. How can we compare biased estimators?

Mean Squared Error (MSE)

Let T be an estimator for a parameter θ. The mean squared error (MSE) of T is defined as $\mathrm{MSE}(T) := \mathrm{E}[(T-\theta)^2]$.

- When choosing among several (un)biased estimators for a parameter θ, it is best to select the one that has the smallest MSE.
- **Bias-Variance-Decomposition:**

$$\mathrm{MSE}(T) = \mathrm{E}[(T-\mathrm{E}[T])^2] + \big(\mathrm{E}[T]-\theta\big)^2 = \mathrm{Var}(T) + (\mathrm{Bias}(T))^2 .$$

(Verify this decomposition yourself!)

Example: MSE of two estimators for the mean

Suppose that an object is weighed 3 times with (potential) data X_1, X_2, X_3. We denote the unknown mean and standard deviation of the X_i's by μ and σ. Consider the following two estimators:

$$\widehat{\mu}_1 = (X_1 + X_2 + X_3)/3 \qquad \textit{and} \qquad \widehat{\mu}_2 = (X_1 + 2X_2 + X_3)/4.$$

- Why would we ever want to use an estimator such as $\widehat{\mu}_2$?

  (Suppose that the measurements are taken by three different lab assistants, and the principle investigator knows that lab assistant no. 2 is the most diligent among the three. Hence, the principle investigator decides to put more weight (quite literally!) on the second measurement . . .)

- Both estimators are unbiased $\mathrm{E}[\widehat{\mu}_1] = \mathrm{E}[\widehat{\mu}_2] = \mu$ but the variances are

$$\mathrm{Var}(\widehat{\mu}_1) = \sigma^2/3 \qquad \text{and} \qquad \mathrm{Var}(\widehat{\mu}_2) = (\sigma^2 + 4\sigma^2 + \sigma^2)/4^2 = 3\sigma^2/8.$$

- Thus, we compute

$$\mathrm{MSE}(\widehat{\mu}_1) = \sigma^2/3 \leq 3\sigma^2/8 = \mathrm{MSE}(\widehat{\mu}_2).$$