MATH 321: Mathematical Statistics

Lecture 4: Point Estimation

Chapter 6: Point Estimation (5.8 and 6.4)

Introduction

The process, where we have been

EDA
↓
Model
↓
Parameters

- Suppose we were given a dataset and went through the EDA where we created lots of summary statistics, histograms, box plots, etc. By this point, we would have a good "feel" for the data.
- If we were focusing on determining (to the best of our ability) the population distribution that a variable came from, we would have used the shape of the sample distribution to guide our selection of a few potential models to test with q–q plots.

  Usually we are not trying to see if the data come from a particular distribution, but rather from a parametric family of distributions (such as the normal, uniform, or exponential, etc.). We are usually forced into this situation because we don't know the parameters.
- Suppose we find a good model, what next? Typically, the next step may be to estimate the parameters, which is what this section is all about.

  This section / topic can be divided into two parts:

  1. Evaluating estimators.
  2. Methods of finding estimators.
- In general, these two activities are intertwined. Often the methods of evaluating estimators will suggest new ones. We will focus mainly on finding estimators.

Point estimation

- Rationale
  - The rational behind point estimation is quite simple. When sampling from a population described by a pdf or pmf $f(x \mid \theta)$, knowledge of θ yields knowledge of the entire distribution.

    Hence, it is natural to seek a method of finding a good estimator of the point θ. For example, if we assume that the population is normally distributed and we know μ and σ^2, then we know everything about the distribution.
  - It may also be the case that some function of θ, say $\tau(\theta)$, is of interest. The second method described in this section can be used to obtain estimation of $\tau(\theta)$.

- Point estimator
  - Definition: A **point estimator** is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic is a point estimator.
  - Notes about definition:
    - Makes no mention of any correspondence between the estimator and the parameter to be estimated.

      If this were a part of the definition, it would restrict the available set of estimators.

      So, any statistic → We could use the sample variance as a point estimator for the population mean, but it would be a bad estimator because we get no insight about $\theta = \mu$
    - Also, there is no mention in the definition of the range of the statistic $W(X_1, \ldots, X_n)$.

      While, in principle, the range of the statistic should coincide with that of the parameter, this is not always the case. For example, if we need $\mu > 0$ but get $\bar{x} = -5$ based on the observed data, this is bad...
  - So, at this point, we want to be careful not to eliminate any candidates from consideration.

- Estimator vs Estimate
  - An **estimator** is a function of the sample; so it is a random variable (i.e. because it is a function of *iid* random variables $X_1, \ldots, X_n$).
  - An **estimate** is the realized value of an estimator that is obtained when the sample is actually taken; so it is just a number (because it is a function of the realized values $x_1, \ldots, x_n$).

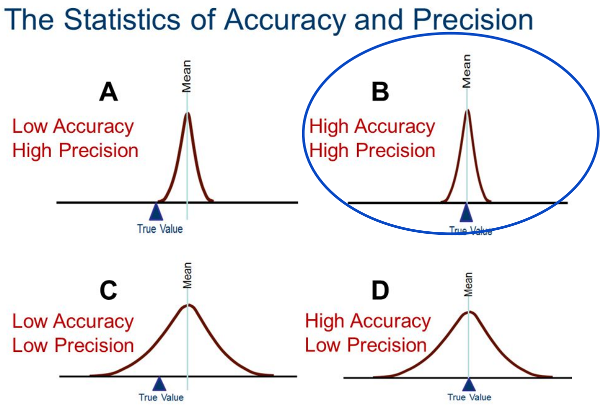
e.g) $\bar{X} = \frac{1}{n} \Sigma X_i$ $x = \{1, 2, 3\}$

$\bar{x} = 2$

Evaluating estimators

Introduction

- There are two ways that we will evaluate estimators. In other words, there are two criteria we apply to determine how [redacted] "good" estimator is.



- Some estimators will be good at one aspect and poor in another, so there is often a tradeoff between accuracy and precision.
- Now we will formalize the theoretical ideas of accuracy and precision.

Unbiasedness

- This criteria deals with the location of the sampling distribution of a statistic.
- Definition: Let $X_1, \ldots, X_n$ be a random sample from X and let θ be a parameter of the pdf (or pmf).

  If $W(X_1, \ldots, X_n)$ is some function of $X_1, \ldots, X_n$ and $E[W(X_1, \ldots, X_n)] = \theta$, then $W(X_1, \ldots, X_n)$ is an **unbiased estimator** of θ. Otherwise it is said to be **biased**.

$\theta \overset{\checkmark}{=} E(\hat{\theta})$

- Specific examples
  - If $\mu = E(X) = \theta$ is a parameter of the pdf (or pmf) of X, then $E(\bar{X}) = \mu = \theta$ and thus $\bar{X}$ is always an unbiased estimator of μ.

    Ex) For $X \sim$ Poisson$\,(\lambda)$: $\lambda = \mu \implies E(\bar{X}) = \mu \overset{\checkmark}{=} \lambda$ unbiased
  - If $\sigma^2 = V(X) = \theta$ is a parameter of the pdf (or pmf) of X, then $E(S^2) = \sigma^2 = \theta$ and thus the sample variance S^2 is always an unbiased estimator of σ^2.

    Ex) If $X \sim$ Normal$\,(\mu, \sigma^2)$:

$\hat{\sigma}^2_{MLE} = V = \frac{n}{n-1} S^2 \longrightarrow E(V) = E\left(\frac{n}{n-1} S^2\right) = \frac{n}{n-1}\sigma^2 \neq \sigma^2$

$\implies$ population variance V is biased for σ^2

Consistency

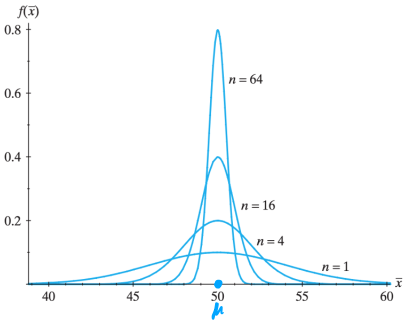
- This criteria deals with the variance of the sampling distribution of a statistic. Before we can formalize this, we need to learn another concept called convergence in probability and some associated theorems.

Convergence in probability idea

- The idea
  - When studying the mean $\bar{X}$ of a random sample of size n from a distribution with mean μ and variance $\sigma^2 > 0$, we saw that is a random variable with the following properties

$$E(\bar{X}) = \mu \qquad \text{and} \qquad V(\bar{X}) = \frac{\sigma^2}{n}$$

  - Thus, as the sample size n increases, the variance of $\bar{X}$ decreases.



pdfs of means of samples from $N(50, 16)$

$\bar{X}$ $X_i \overset{iid}{\sim}$

  - We can see that as n increases, the probability becomes concentrated in a small interval centered at μ.
  - That is, as n increases, $\bar{X}$ tends to converge to μ, or $(\bar{X} - \mu)$ tends to converge to 0 in a probability sense.

ex) $\bar{x}_1 \sim N(\mu, \frac{\sigma^2}{1})$

$\bar{x}_2 \sim N(\mu, \frac{\sigma^2}{2})$

$\vdots$

$\bar{x}_n \sim N(\mu, \frac{\sigma^2}{n})$

$\vdots$

as $n \to \infty$ $V(\bar{x}) \to 0$

- Convergence in statistics
  - Convergence in statistics is very different from that in mathematics.
  - In mathematics, a sequence of **constants** $a_1, a_2, \ldots$ converges to a constant:

$$\lim_{n\to\infty} a_n = a \qquad \text{ex. } \lim_{n\to\infty} \frac{1}{n} = 0$$

  - But in statistics, a sequence of **random variables** $X_1, X_2, \ldots$ converges to a random variable:

$$\lim_{n\to\infty} X_n = X$$

    (Note: it can also converge to a constant, depending on the situation.)
  - Three types of convergence in statistics:
    1. Convergence in probability.
    2. Almost sure convergence.
    3. Convergence in distribution.

    We will focus on number one, mention number three and ignore number two.

Convergence in probability definition

- This type of convergence is one of the weaker types and, hence, is usually quite easy to verify.
- Definition: A sequence of random variables, $Y_1, Y_2, \ldots,$ **converges in probability** to a random variable Y if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|Y_n - Y| \geq \epsilon) = 0 \quad \text{or, equivalently,} \quad \lim_{n\to\infty} P(|Y_n - Y| < \epsilon) = 1$$

complement

Breakdown of definition

- Notation
  - $Y_1, Y_2, \ldots$ represent statistics that depend on the subscript (i.e. functions of a random sample). More specifically, Y_n is a statistic defined with the original *iid* variables $X_1, \ldots, X_n$.
  - So $Y_n = T(X_1, \ldots, X_n)$.

    For example, if Y_n is the sample mean, then

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

  - So the distribution of Y_n changes as the subscript changes and it converges to some limiting distribution as n becomes large.

- Understanding $\lim_{n\to\infty} P(|Y_n - Y| < \epsilon)$
  - For a given (fixed) n, is $P(|Y_n - Y| < \epsilon)$ is just a regular probability; so it is just a constant.
  - So $P(|Y_n - Y| < \epsilon) = a_n$ and consequently

    $$\lim_{n\to\infty} P(|Y_n - Y| < \epsilon) = \lim_{n\to\infty} a_n$$

    * This is the convergence that we are familiar with in mathematics (more specifically in real analysis).
    * Rigorously, $\lim_{n\to\infty}$ can only be used with a sequence of constants and cannot be used with a sequence of random variables.

      It doesn't make sense to find the limit of a random variable (we can't find the pattern if each number is random and has a pattern of its own).
    * But probability is a constant number that we can find a limit of. Thus, using $\lim_{n\to\infty} P(|Y_n - Y| < \epsilon)$ notation makes sense.
  - So, because Y_n is a random variable, we cannot find its limit directly, but we can find its limit in probability or distribution.

- Interpretations
  - The event $|Y_n - Y|$ is the difference between Y_n and Y.
  - $P(|Y_n - Y| < \epsilon)$ is the probability that the difference between Y_n and Y is smaller than ϵ.

    In definition, "for every $\epsilon > 0$" means that we can pick any really tiny number (e.g. $\frac{1}{100000000}$).
  - Putting it all together: $\lim_{n\to\infty} P(|Y_n - Y| < \epsilon) = 1$

    Even though we choose a really tiny ϵ, the probability that the difference between Y_n and Y is less than the small number converges to one as n goes to ∞.
  - In other words, the probability that there is no difference between Y_n and Y goes to one as n approaches ∞.

    $$\lim_{n\to\infty} P(\text{no difference } Y_n + Y) = 1$$

    Thus, we can conclude that Y_n converges to Y **in probability.**

- Correct notation (note that we need the "in probability" part in all of these):
  - $Y_n \xrightarrow{p} Y$.
  - $Y_n \to Y$ in probability.
  - $\lim_{n\to\infty} Y_n = Y$ in probability.

(Weak) Law of Large Numbers (WLLN)

- Theorem
  - Frequently, statisticians are concerned with situations in which the limiting random variable is a constant and the random variables in the sequence are sample means (of some sort). The most famous result of this type is the following.
  - **WLLN** Theorem: Let $X_1, X_2, \ldots$ be *iid* random variable with $E(X_i) = \mu$ and $V(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for every $\epsilon > 0$,

    $$\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1 \iff \lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$$

    that is, $\bar{X}_n$ converges in probability to μ (notation: $\bar{X} \xrightarrow{p} \mu$).

- Notes about WLLN
  - Comparison
    - Convergence in probability definition: $\lim_{n\to\infty} P(|\underset{\bar{X}_n}{\underset{\downarrow}{Y_n}} - \underset{\mu}{\underset{\downarrow}{Y}}| \geq \epsilon) = 0$

      WLLN:
  - Summary of theorem:
    - The Weak Law of Large Numbers (WLLN) quite elegantly states that, under general conditions, the sample mean approaches the population mean as $n \to \infty$.

      This is because the probability associated with the distribution of $\bar{X}$ becomes concentrated in an arbitrarily small interval centered at μ as n increases.
    - ☆ Needed conditions: *iid* random variables and the first and second moments (i.e. the mean and a finite variance). And do not need any distributional assumption.
  - Consistency
    - The property summarized by the WLLN, that is a sequence of the "same" sample quantity approaches a constant as $n \to \infty$, is known as **consistency**.
    - Showing consistency is the same as showing convergence in probability.

      Thus, it can be said that $\bar{X}_n$ is a consistent estimator of μ.

$\bar{X}_n \xrightarrow{p} \mu \iff \bar{X}$ is a consistent estimator of μ

- WLLN for transformations of X (extension of theorem).
    - ∗ Additionally, it can be used for statistic of the form $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} g(X)$, where $g(X)$ is non negative transformation that still has a mean and a finite variance.
    - ∗ So, now instead of converging to μ, $\bar{X}_n$ converges to $E[g(X)]$ (in probability).

$\overline{g(X)} \xrightarrow{p} E(g(x))$, ex) $\frac{1}{n}\sum X_i^2 \xrightarrow{p} E(x^2) = \sigma^2 + \mu^2$

- Proof of WLLN
    - Want to show: $\lim_{n\to\infty} P(|\bar{X}_n - \mu| \geq \epsilon) = 0$

$\rightarrow \quad P(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{E[|\bar{X}_n - \mu|]}{\varepsilon}$

< by chebyshev's theorem >

$\Rightarrow \quad P((\bar{X}_n - \mu)^2 \geq \varepsilon^2) \leq \frac{E[(\bar{X}_n - \mu)^2]}{\varepsilon^2}$

$= \frac{V(\bar{x})}{\varepsilon^2}$

< V(X) = second central moment >

$E(x-\mu)^2 \rightarrow$ now $\bar{x}$

$= \frac{\sigma^2/n}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}$

< don't know expected value of absolute value, so square >

Complement

$\rightarrow \quad \lim_{n\to\infty} P(|\bar{x}_n - \varepsilon| \geq \varepsilon) = \lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} \overset{\checkmark}{=} 0 \quad \Longleftrightarrow \quad \bar{x}_n \xrightarrow{p} \mu$

- Application of WLLN
    - Example: Suppose we are planning a poll to figure out which is better, R or Excel. Let

$$X_i = \begin{cases} 1 & \text{if R} \\ 0 & \text{if Excel} \end{cases} \sim \text{Bernoulli}(p)$$

and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. $\longrightarrow E(\bar{x}_n) = \mu = p \; + \; V(\bar{x}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n}$

chebyshev's Inequality

$\rightarrow P(|\bar{x}_n - \mu| < \varepsilon) \geq 1 - \frac{V(\bar{x})}{\varepsilon^2}$

1. If $n = 400$, find a lower bound on $P(|\bar{X}_{400} - p| < 0.05)$.

↳ unknown

$P(\underbrace{|\bar{x}_{400} - p|}_{\text{within}} < \underbrace{0.05}_{\text{Margin of Error}}) \geq 1 - \frac{p(1-p)}{400(0.05)^2}$

$= 1 - p(1-p) \longrightarrow 1 - 0.5(0.5) = 0.75$

use p=0.5

2. If $n = 400$ and $p = 7/10$, find a lower bound on $P(|\bar{X}_{400} - 0.70| < 0.05)$.

↳ known

$P(|\bar{x}_{400} - 0.7| < 0.05) \geq 1 - \frac{0.7(0.3)}{400(0.05)^2}$

$\downarrow \quad = 1 - 0.21 = 0.79$

3. If $n = 500$ and $p = 7/10$, find a lower bound on $P(|\bar{X}_{500} - 0.70| < 0.05)$.

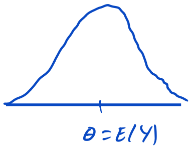
$P(|\bar{x}_{500} - 0.7| < 0.05) \geq 1 - \frac{0.7(0.3)}{500(0.05)^2}$

$\downarrow \quad = 1 - 0.168 = 0.837$

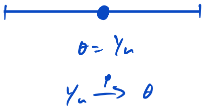
Summary of consistency and unbiasedness

- Comparison of unbiasedness vs consistency.
  - Unbiasedness → This tells us the mean of a statistic, regardless of n. So we can drop the inference on n. To be unbiased, the expected value of the statistic must equal the parameter of interest. $E(Y_n) = \theta \longrightarrow E(Y) = \theta$

    for all n



  - Consistency → This is all about the limit of the random variable as $n \to \infty$. If a statistic is consistent, then as $n \to \infty$, there is no variation in what the statistic converges to; the entire distribution converges to a constant.



- Examples of the difference (shown through counter examples)
  1. Let $Y_n \sim \mathrm{N}\,(\mu, \sigma^2)$.

     $E(Y_n) = \mu$, but $Y_n \xrightarrow{p} N(\mu, \sigma^2) \neq \mu$

     doesn't depend on n $\Rightarrow$ keeps variation

     unbiased ✓ consistent ✗

     So it still has some variation, whereas a constant has no variation (it is always the same).

     $E(Y_n) = \mu$ ____does not imply____ $Y_n \xrightarrow{p} Y$.

  2. Now let $Y_n \sim \mathrm{N}\,(\mu + \frac{1}{n}, \frac{\sigma^2}{n})$

     As $n \to \infty$: $E(Y_n) = \mu$ $\Rightarrow$ $Y_n \xrightarrow{p} \mu$, but for a fixed n $E(Y_n) = \mu + \frac{1}{n} \neq \mu$

     $V(Y_n) = 0$

     consistent ✓ unbiased ✗

     So, the mean of the distribution converges to μ and the variance disappears.

     $Y_n \xrightarrow{p} \mu$ ____does not imply____ $E(Y_n) = \mu$.

$\cancel{A}$ unbiased $\not\Leftrightarrow$ consistent

Return to methods of finding estimators

- Now that we have covered how to evaluate estimators, we can look at how to find estimators.
- In many cases, there will be an obvious or a natural candidate for a point estimator of a particular parameter. For example:
  - Population mean $\mu \to$ $\bar{X}$, Median, midrange, etc
  - If $X \sim \text{Uniform}\,(0, \theta) \to$ $\max(X_1, \dots, X_n) = X_{(n)}$
  - If $X \sim \text{Gamma}\,(\text{shape } \alpha, \text{rate } \beta) \to$??
- For more complicated models, intuition may not work and can often have bad results (e.g. gamma(α, β), there is no obvious estimators for the shape and scale parameters).
- Therefore, it is useful to have some techniques (more methodical ways of estimating parameters) that will at least give us some reasonable candidates for consideration.

  These still must be evaluated before their worth is established. Ideally, point estimators will provide insight and information about the unknown parameter θ.
- Now we will go into two methods for finding estimators.

Method of moments

Method of Moments (MME)

- The method of moments is a very simple procedure for finding an estimator for one or more population parameters.
- Types of moments:
  - $\boldsymbol{k}^{\text{th}}$ **(population) moment** of the distribution (about the origin)

    $$\mu'_k = E(X^k)$$
  - The corresponding **sample moment** is the average

    $$m'_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

e.g) $k = 1$

pop mean

$\to \mu'_1 = E(X) = \mu$

$\mu'_2 = E(X^2) = \sigma^2 + \mu^2 \neq \mu^2$

Sample mean

$\to m'_1 = \frac{1}{n}\sum X_i = \bar{X}$

$\to m'_2 = \frac{1}{n}\sum X_i^2 \neq \bar{X}^2$

- The method of moments logic
  - Based on the intuitively appealing idea that sample moments should provide good estimates of the corresponding population moments.
  - ☆ Population moments $\mu'_1, \dots, \mu'_k$ are usually functions of the population parameters, so we can equate corresponding population and sample moments and solve for the desired estimators.

population moment

equate sample moment

Solve for θ

$$\mu'_k = g(\theta) = m'_k \Rightarrow \theta = g^{-1}(m'_k)$$

- Official statement of **Method of Moments**:

  Choose as estimates those values of the parameters that are solutions of the equations $\mu_k' = m_k'$, for $k = 1, 2 \ldots, t$, where t is the number of parameters to be estimated.

- Steps to find MME:

  1. Write $E(X^k)$ as a function of the parameters of interest.

     Note: Might have to do some integration or summation to get $E(X^k)$.

     Example: If $X \sim$ Normal $(\mu, 1) \rightarrow$ $\mu_1' = E(X) = \mu$

  2. Then estimate the parameter of interest by equating the population moment with the sample moment and solving for the parameter.

     Example continued: $\dfrac{\mu_1' = m_1'}{\mu = \bar{x}} \Longrightarrow \hat{\mu}_{MME} = \bar{X}$

Examples

1. Let $X_1, \ldots, X_n$ be a random sample from Uniform$(0, \theta)$, where θ is unknown. Use the method of moments to estimate the parameter θ.

① $\mu_1' = E(X) = \dfrac{a+b}{2} = \dfrac{0+\theta}{2} = \dfrac{\theta}{2}$

$m_1' = \bar{X}$

② $\dfrac{\theta}{2} = \bar{X} \Longrightarrow \hat{\theta}_{MME} = 2\bar{X}$

2. Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Exp (λ). Find the method of moments estimator for λ.

① $\mu_1' = E(X) = \dfrac{1}{\lambda}$

$m_1' = \bar{X}$

② $\dfrac{1}{\lambda} = \bar{x} \Longrightarrow \hat{\lambda}_{MME} = \dfrac{1}{\bar{x}} \longrightarrow \dfrac{n}{\Sigma x_i}$

$V(X) = \dfrac{1}{\lambda^2}$ $\dfrac{1}{\lambda^2} = V$ $\dfrac{1}{V} = \lambda^2$ $\sqrt{\cdot} = \sqrt{\dfrac{n}{\Sigma(x_i - \bar{x})^2}}$

3. Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Uniform $(-\theta, \theta)$. Find the MME for θ. Recall $E(X) = \frac{a+b}{2}$ and $V(X) = \frac{(b-a)^2}{12}$ if $X \sim$ Uniform (a, b).

① $\mu_1' = E(X) = \dfrac{-\theta+\theta}{2} = 0$ $\longrightarrow$ No information about θ

$\mu_2' = E(X^2) = V(X) + (E(X))^2$

$= \dfrac{(\theta+\theta)^2}{12} + \theta^2$

$= \dfrac{1}{3}\theta^2$

$m_2' = \dfrac{1}{n}\Sigma x_i^2$

$\dfrac{1}{3}\theta^2 = \dfrac{1}{n}\Sigma x_i^2$

$\hat{\theta}_{MME} = \sqrt{\dfrac{3}{n}\Sigma x_i^2}$

?

$\dfrac{1}{3}\theta^2 = V = \sqrt{\dfrac{3}{n}\Sigma x_i^2 - \bar{x}^2}$

$\sqrt{3V}$

use central moments

$E(X) = \mu \longrightarrow \hat{\mu}_{MME} = \bar{x}$

$V(X) = \sigma^2 \longrightarrow \hat{\sigma}^2_{MME} = v$ ✓

$\downarrow = \Sigma(x_i - \bar{x})^2$

4. Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Normal(μ, σ^2). Find the MMEs for μ and σ^2.

   Note: There are two unknown parameters. So we will have to setup and solve a system of equations.

① $\mu_1' = E(X) = \mu$

$\mu_2' = E(X^2) = V(X) + (E(X))^2$

$\downarrow \quad = \sigma^2 + \mu^2$

$m_1' = \bar{x}$

$m_2' = \frac{1}{n} \Sigma x_i^2$

② $\mu = \bar{x} \Rightarrow \hat{\mu}_{MME} = \bar{x}$

$\sigma^2 + \mu^2 = \frac{1}{n} \Sigma x_i^2$

$\sigma^2 = \frac{1}{n} \Sigma x_i^2 - \boxed{\mu}^2$

$\hat{\sigma}^2_{MME} = \frac{1}{n} \Sigma x_i^2 - \bar{x}^2$

$\downarrow = \frac{1}{n} \Sigma (x_i - \bar{x})^2 = v$

Two forms

population variance

$\downarrow$

$= \frac{1}{n} \Sigma (x_i^2 - 2\bar{x} x_i + \bar{x}^2)$

$= \frac{1}{n} \Sigma x_i^2 - \underbrace{2\bar{x} \underbrace{\tfrac{1}{n} \Sigma x_i}_{\bar{x}}}_{2\bar{x}^2} + \underbrace{\frac{1}{n} \underbrace{\Sigma \bar{x}^2}_{n\bar{x}^2}}_{\bar{x}^2}$

✓ $= \frac{1}{n} \Sigma x_i - \bar{x}^2$

Show equivalent

Summary of method of moments

- Pros
  - Simple to find and fairly intuitive (simply matching the properties of a sample to that of the population distribution).
  - Nonparametric method.

    So it works without the distributional information about the population (think back to the normal MME example, those results for estimators of $E(X)$ and $V(X)$ are true for regardless of what we start with.

    Example: Suppose $X_i \overset{iid}{\sim}$ Gamma(α, β), then $\bar{X}$ is an estimator for α/β and v is an estimator for α/β^2.

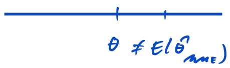
    This means that we don't have to assume the population distribution, which is a useful property when the population distribution information is unclear.
  - Consistent estimators most of the time.

    Sample moments are consistent estimators of the corresponding population moments (can show with the (Weak) Law of Large Numbers).
- Cons

  $m_k' = \frac{1}{n} \Sigma x_i^k = \overline{g(x)} \xrightarrow{p} E(g(x)) \Rightarrow$ Consistent

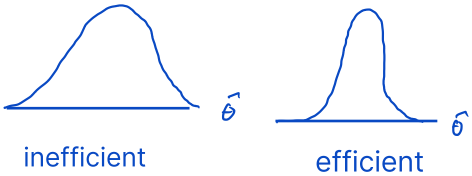
  - Nonparametric method.

    Because of this, MME information is only based on the data and doesn't give us any information about that relationship with the parameter of interest.
  - Often biased (so the center of the distribution of the estimator doesn't line up with θ).
  - May be inefficient (i.e. large variance of the distribution of $\hat{\theta}$).



$\theta \neq E(\hat{\theta}_{MME})$

Biased



inefficient

efficient

Maximum likelihood estimation

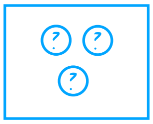
Context

- We just saw one way to get estimators, but noted that there are some disadvantages of that method. One reason for this is that it is a very general method because it is non-parametric. So how can we improve upon it?
- ☆ Parametric statistics: Assume the distribution of X and estimate the parameters that determine the distribution.

  Example: Know $X \sim \text{Normal}$, estimate μ and σ^2.
- The method of maximum likelihood is, by far, the most popular technique for deriving estimators.

Motivating (conceptual) example

- Suppose that we are confronted with a box that contains three balls. We know that each of the balls may be red or white, but we do not know the total number of either color. However, we are allowed to randomly sample two of the balls without replacement.
- If our random sample yields two red balls, what would be a good estimate of the total number of red balls in the box?



→ K = number of red out of 3

know k = 2 or 3

$$\left.\begin{aligned} \rightarrow \quad P(X=2 \mid K=2) &= \frac{\binom{2}{2}\binom{1}{0}}{\binom{3}{2}} = \frac{1}{3} \\ P(X=2 \mid K=3) &= \frac{\binom{3}{2}\binom{0}{0}}{\binom{3}{2}} = \boxed{1} \end{aligned}\right\rangle$$

more likely ⇒ K =3

Likelihood function

- **Parameter space** definition: Given pdf (or pmf) $f(x \mid \theta_1, \dots, \theta_k)$ the set of all possible values for $\theta_1, \dots, \theta_k$ is known as the parameter space.

  We denote the parameter space with Θ (capital "theta").
- Examples:

  If $X \sim \text{Normal}\,(\mu, \sigma^2) \rightarrow \Theta =$ $\{(\mu, \sigma^2) : -\infty < \mu < \infty,\ \sigma > 0\}$

  If $X \sim \text{Poisson}\,(\lambda) \rightarrow \Theta =$ $\{\lambda : \lambda > 0\}$

- Review: Joint pdf of $X_1, \ldots, X_n$ (if X_i's are continuous, *iid* random variables) is given by

  $$f(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$ < product of marginals >

  Pdf $f(x_i)$ is a function of X_i given the parameters $\boldsymbol{\theta}$: $f(X_i \mid \boldsymbol{\theta})$.

- **Likelihood function** definition: Let $f(\mathbf{x} \mid \boldsymbol{\theta})$ denote the joint pdf or pmf of the sample $\boldsymbol{X} = (X_1, \ldots, X_n)$. Then, given that $\mathbf{X} = \mathbf{x}$ is observed, the function of θ defined by

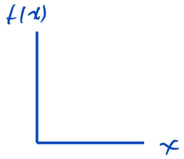
  $$L(\boldsymbol{\theta} \mid \mathbf{x}) = f(\mathbf{x} \mid \boldsymbol{\theta})$$

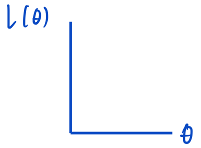
  is called the likelihood function.

- Notes about the likelihood function

  – The only distinction between the likelihood function and the joint pdf or pmf is which variable is considered fixed and which is varying.

  In other words, the likelihood function is the same thing as the joint density of the data, but from a different point of view (i.e. different information is known).





  * For the joint density of the data, θ is fixed, while $\boldsymbol{X}$ can vary.

    This is used to answer probability questions: we know the parameters and want to figure out the sample.

  * For the likelihood function, $\mathbf{X}$ is fixed, while θ can vary.

    This is used to answer statistics questions: we have data and want to figure out the most likely parameter value.

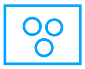
  – Because both $\mathbf{x}$ and θ are in the formula, this gives us information about the **relationship** between the data and the parameter.

- We can find the likelihood function with $L(\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$

- Comparing likelihood functions (this is what exactly we did with the colored balls example!)

  – If $\mathbf{X}$ is a discrete random vector, then $L(\theta \mid \mathbf{x}) = P_\theta(\mathbf{X} = \mathbf{x})$. If we compare the likelihood at two parameter points and find that

  $$P_{\theta_1}(\mathbf{X} = \mathbf{x}) = L(\theta_1 \mid \boldsymbol{x}) > L(\theta_2 \mid \mathbf{x}) = P_{\theta_2}(\boldsymbol{X} = \boldsymbol{x})$$



  $$P(X=2 \mid k=3) = L(k=3) > L(k=2) = P(X=2 \mid k=2)$$

  ✱ Likelihoods, just for a sample of n=1

  then we interpret this as follows:

  – The sample $\mathbf{x}$ we actually observed is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$ (with the same data).

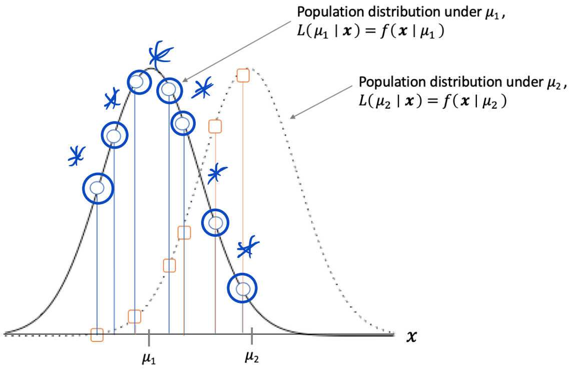
  This can be interpreted as saying that θ_1 is a more plausible value for the true value of θ than θ_2.

Maximum likelihood estimation (MLE) definition and concept

- Definition: For each sample point $\mathbf{x}$, let $\hat{\theta}(\boldsymbol{x})$ be a parameter value at which $L(\theta \mid \mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A **maximum likelihood estimator (MLE)** of the parameter θ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.

- Notes about the definition
  - Intuitively, the MLE is a reasonable choice for an estimator. The MLE is the parameter point for which the observed sample is most likely.
  - In general, the MLE is a good point estimator, possessing some of the optimality properties such as consistency.

- MLE conceptualized



  - The likelihood function is the product of the density curve heights at the observed xs.
  - So for this example, $\underline{\mu_1}$ is a more plausible value for μ.

How to find an MLE

- Example: Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Exp}(\lambda)$. Find the maximum likelihood estimator for λ. How do we do this?

Step 1

- Start with the likelihood function:

$$\rightarrow \quad L(\lambda / x) = \prod_{i=1}^{n} f(x_i \mid \lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \Sigma x_i} \quad , \quad \lambda > 0$$

- Then it's an optimization problem: To find the maximum of a function, we use calculus and derivatives.

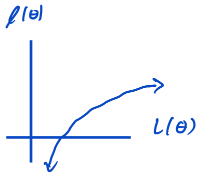
  If the likelihood function is differentiable, we can solve for the points at which the first derivatives equals zero:

$$L'(\theta \mid \mathbf{x}) = \frac{d}{d\theta} L(\theta \mid \mathbf{x}) = 0$$

Step 1

- Log likelihood
  - It is easier to work with the natural logarithm of $L(\theta \mid \mathbf{x})$ than it is to work with $L(\theta \mid \mathbf{x})$ directly. This is known as the **log likelihood**:

$$\ell(\theta) = \ell(\theta \mid \mathbf{x}) = \ln[L(\theta \mid \mathbf{x})] \longrightarrow \ell'(\theta) = 0$$



  - This transformation is valid since the natural log is a strictly increasing function on $(0, \infty)$ (so it's a one-to-one function), which means it's equivalent to maximize the natural log of the likelihood function.

Continuing example:

$$\rightarrow \ell(\lambda) = \ln[L(\lambda)] = \ln\left[\lambda^n e^{-\lambda \sum x_i}\right] = \ln(\lambda^n) + \ln\left(e^{-\lambda \sum x_i}\right) = n\ln(\lambda) - \lambda \sum x_i$$

step 2

$$\rightarrow \ell'(\lambda) = \frac{d}{d\lambda}\left[n\ln(\lambda) - \lambda \sum x_i\right] = \frac{n}{\lambda} - \sum x_i$$

$$\rightarrow 0 = \frac{n}{\lambda} - \sum x_i \implies \hat{\lambda} = \frac{n}{\sum x_i} = \frac{1}{\bar{x}}$$

Step 3

At this point, the solution $\hat{\theta}$ is only a **possible candidate** for the MLE of (θ).

First derivative being zero is only a necessary condition, but not a sufficient condition because points at may be local or global minimum / maximum, or inflection points.

- So we have to check the second derivatives at $\hat{\theta}$ to ensure they are global maximum:

$$L''(\theta \mid \mathbf{x}) = \frac{d^2}{d\theta^2} L(\theta \mid \mathbf{x}) \qquad \rightarrow \qquad L''(\hat{\theta} \mid \mathbf{x}) < 0 \longrightarrow \ell''(\theta), \quad \ell''(\theta) \overset{?}{<} 0$$

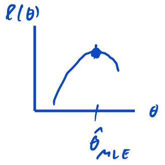
If this is true, then we know that we have found $\hat{\theta}_{MLE}$. Continuing example:

$$\rightarrow \ell''(\lambda) = \frac{d}{d\lambda}\ell'(\lambda) = -\frac{n}{\lambda^2}$$

$$\rightarrow \ell''\left(\frac{1}{\bar{x}}\right) = \frac{-n}{(1/\bar{x})^2} \overset{\checkmark}{<} 0 \implies \hat{\lambda}_{MLE} = \frac{1}{\bar{x}}$$

- Summary
  - Simply put, the likelihood function is hill shaped with the highest point at the MLE.
  - Note that the likelihood function not always differentiable, which adds in some extra complexity when finding the MLE.

    When this is the case, we can try to numerical maximization.



- The process that was just demonstrated was for univariate θ. It is the same process for a vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, except we have to work with partial derivatives.

Steps to find MLEs

1. Write the likelihood function (i.e. joint density function) and the log-likelihood,

$$L(\theta \mid \mathbf{x}) = \prod_{i=1}^{n} f(\mathbf{x} \mid \theta) \qquad \rightarrow \qquad \ell(\theta) = \ln[L(\theta \mid \mathbf{x})]$$

2. Optimize the log-likelihood function by taking the derivatives with respect to the parameter of interest.

   Set to zero and solve for the parameter of interest.

$$\ell'(\theta) = \frac{d}{d\theta}\ell(\theta) = 0 \qquad \rightarrow \qquad \hat{\theta} = \text{potential MLE}$$

3. Verify that the global maximum of the log-likelihood function occurs at $\theta = \hat{\theta}$.

   Find the second derivative of the log-likelihood function, then plug in $\hat{\theta}$ and see if less than zero.

$$\ell''(\theta) = \frac{d^2}{d\theta^2}\,\ell(\theta) \qquad \rightarrow \qquad \ell''(\hat{\theta}) \overset{?}{<} 0$$

   If so, then we have $\hat{\theta}_{MLE}$.

Examples

1. Let $X_1, \ldots, X_n \overset{iid}{\sim}$ Geometric (p). Find the maximum likelihood estimator for p.

   (a) Find the likelihood function and log-likelihood function for p.

$$\rightarrow L(p \mid x) = \prod_{i=1}^{n} f(x_i \mid p) = \prod_{i=1}^{n} (1-p)^{x_i - 1} p = (1-p)^{\sum x_i - n} p^n$$

$$\rightarrow \ell(p) = \ln[L(p \mid x)] = (\sum x_i - n)\ln(1-p) + n\ln(p)$$

   (b) Optimize the log-likelihood function and solve for $\hat{p}$.

$$\rightarrow \ell'(p) = \frac{d}{dp}[(\ldots)] = \frac{-(\sum x_i - n)}{1-p} + \frac{n}{p} = \frac{(n - \sum x_i)p}{(1-p)p} + \frac{n(1-p)}{p(1-p)} = \frac{n}{p(1-p)}\left[\left(1 - \frac{\sum x_i}{n}p\right) + 1 - p\right]$$

$$\rightarrow 0 = (1 - \bar{x})p + (1-p)$$

$$= \cancel{p} - p\bar{x} + 1 - \cancel{p} \implies \hat{p} = \frac{1}{\bar{x}}$$

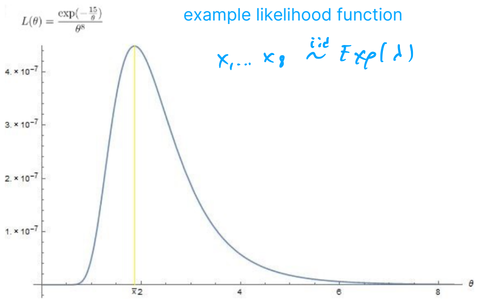
   (c) Perform second derivative test to confirm if $\hat{p}$ is the MLE for p.

$$\rightarrow \ell''(p) = \frac{d}{dp}[\ell'(p)] = \frac{-(\sum x_i - n)}{(1-p)^2} - \frac{n}{p^2} = -\left[\frac{\sum x_i - n}{(1-p)^2} + \frac{n}{p^2}\right]$$

$\geq 0 \rightarrow \sum x_i \geq n$ b/c $x_i = 1, 2, \ldots$

$$\rightarrow \ell''\left(\tfrac{1}{\bar{x}}\right) = -\left[\frac{\overset{\geq 0}{\boxed{\sum x_i - n}}}{\underset{>0}{\boxed{(1 - (\frac{1}{\bar{x}}))^2}}} + \underset{>0}{\boxed{\frac{n}{(\frac{1}{\bar{x}})^2}}}\right] \overset{\checkmark}{<} 0 \implies \hat{p}_{MLE} = \frac{1}{\bar{x}}$$

☆ easier to not simplify and consider each term separately ±



example likelihood function

$x_1 \ldots x_8 \overset{iid}{\sim} Exp(\lambda)$

2. Let $X_1, \ldots, X_n \overset{iid}{\sim} \text{Normal}\,(\mu, \sigma^2)$. Find the MLEs for μ and σ^2.

   Note: Trying to find MLEs for two parameters, so will have to take partial derivatives.

$$1) \rightarrow L(\mu, \sigma^2 | x) = \prod_{i=1}^{n} f(x_i | \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2}$$

$$\downarrow \quad = (2\pi\sigma^2)^{-n/2} e^{-\frac{1}{2\sigma^2}\Sigma(x_i - \mu)^2}$$

$$\rightarrow \ell(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\Sigma(x_i - \mu)^2$$

$$2) \rightarrow \frac{\partial \ell(\mu, \sigma^2)}{\partial \mu} = \frac{1}{\sigma^2}\Sigma(x_i - \mu) \longrightarrow 0 = \frac{1}{\sigma^2}\Sigma(x_i - \mu)$$

$$\downarrow \; = \Sigma x_i - n\mu$$

$$\hat{\mu} = \frac{\Sigma x_i}{n} = \bar{x}$$

$$\rightarrow \frac{\partial \ell(\mu, \sigma^2)}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\Sigma(x_i - \mu)^2 \longrightarrow 0 = \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\Sigma(x_i - \mu)^2$$

$$\frac{n}{\cancel{2\sigma^2}} = \frac{1}{\cancel{2}(\sigma^2)^{\cancel{2}}}\Sigma(x_i - \mu)^2 \Longrightarrow \hat{\sigma}^2 = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

$$\downarrow \; = v$$

treat as one "thing"
(not something squared)

When working with partial derivatives, the second derivative test checks to see if the determinant of the matrix of the second partial derivatives (called the Hessian matrix) is less than zero.

For this scenario, the solutions do provide a maximum.

$$\Longrightarrow \hat{\mu}_{MLE} = \bar{x} \quad +$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n}\Sigma(x_i - \bar{x})^2 = v$$

Finding MLEs for functions of parameters

- We mentioned this before in the overview of point estimation that we be interested in some function of θ, say $\tau(\theta)$,

  A useful property of MLE is know as the invariance property of MLE.

- **(Invariance property of MLEs)**: If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

- Notes about this property
  - This type of theorem usually only holds with continuous functions, but this one works with ANY function. So we don't have to check any conditions.
  - This means if we want to find the MLE for $\tau(\theta)$:
    1. Find the MLE of θ.
    2. Simply apply the invariance property to get the MLE of $\tau(\theta)$.

- Example: Let $X_1, \dots, X_n \overset{iid}{\sim}$ Geometric (p).
  Find the MLE for for $V(X) = \dfrac{1-p}{p^2} = \tau(p)$.

→ 1) $\hat{p}_{MLE} = \frac{1}{\bar{x}}$

→ 2) $\hat{\tau(p)}_{MLE} = \tau(\hat{p}_{MLE}) = \dfrac{1-\frac{1}{\bar{x}}}{(\frac{1}{\bar{x}})^2}$

Miscellaneous notes about MLEs

- Optimal properties
  - Results shows that under general conditions, MLEs are consistent estimators of their parameters and asymptotically efficient (small variance in the limiting distribution (think: convergence in distribution).
  - This means it is a method of finding an estimator that guarantees optimal properties, asymptotically.
- Maximation
  - The possibility of maximizing $L(\theta \mid \mathbf{x})$ is one of the most important features of MLEs.
  - Example: Let $X \sim$ Gamma (α, β). Find the MLEs for α and β.

$$l(\alpha, \beta \mid x) = \prod_{i=1}^{n} f(x_i \mid \alpha, \beta) = \prod_{i=1}^{n} \frac{\beta^{\alpha}}{\Gamma(\alpha)} x_i^{\alpha-1} e^{-\beta x_i} = \frac{\beta^{n\alpha}}{\boxed{(\Gamma(\alpha))^{n}}} (x_1 \cdot \ldots \cdot x_n)^{\alpha-1} e^{-\beta \Sigma x_i}$$

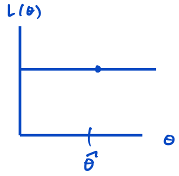
difficult derivative

- Turn to **numerical maximization**, which simply put essentially means plugging in a ton of numbers and seeing when the result is the largest.

  If a model (likelihood) can be written down, then there is some hope of maximizing it numerically and hence finding the MLEs of the parameters.

  When this is done, there is still always the question of whether a local or global maximum has been found.

- Numerical sensitivity
  - When we use numerical methods, we have to pay careful attention to a potential problem of **numerical sensitivity**. That is, how sensitive is the estimate to small changes in the data?
  - This situation arises when the MLE cannot be solved for explicitly (i.e. there is no closed form solution, perhaps because the derivative doesn't exist). This occurrence happens when the likelihood function is very flat in the neighborhood of its maximum or when there is not a finite max.



  - Example: The MLEs of n and p (both unknown) in binomial sampling can be highly unstable. Five realizations of a Binomial(n, p) experiment are observed.

    The first data set is (16, 18, 22, 25, **27**) and the MLE of n is $\hat{n} = \mathbf{99}$.

    The second data set is (16, 18, 22, 25, **28**) and the MLE of n is $\hat{n} = \mathbf{190}$.
  - So it is often wise to spend a little extra time investigating the stability of the solution.
  - If the MLE can be solved for explicitly, this is usually not a problem.