Statistical Methods for Data Science
Lesson 13 - Unbiased estimators. Efficiency and MSE

Salvatore Ruggieri

Department of Computer Science
University of Pisa
salvatore.ruggieri@unipi.it
A dataset \( x_1, \ldots, x_n \) consists of repeated measurements of a phenomenon we are interested in understanding.

- E.g., measurement of the speed of light

We model a dataset as the realization of a random sample.

A random sample is a collection of i.i.d. random variables \( X_1, \ldots, X_n \sim F(\alpha) \), where \( F() \) is the distribution and \( \alpha \) its parameter(s).

Challenging questions:

- How to determine \( E[X], \ Var(X) \), or other functions of \( X \)?
- How to determine \( \alpha \), assuming to know the form of \( F \)?
- How to determine both \( F \) and \( \alpha \)?
An example

What is an estimate of the true speed of light?

\[ x_1 = 850, \text{ or } \min x_i, \text{ or } \max x_i, \text{ or } \bar{x}_n = 852.4 \]
An example

- Speed of light dataset as realization of
  \[ X_i = c + \epsilon_i \]
  where \( \epsilon_i \) is measurement error with \( E[\epsilon_i] = 0 \) and \( \text{Var}(\epsilon_i) = \sigma^2 \)

- We are then interested in \( E[X_i] = c \)

- How to estimate?
- Use some info. For \( X = X_1 \):
  \[
  E[X] = E[X_1] = c \quad \text{Var}(X) = \text{Var}(X_1) = \sigma^2
  \]

- Use all info. For \( \bar{X}_n = (X_1 + \ldots + X_n)/n \):
  \[
  E[\bar{X}_n] = c \quad \text{Var}(\bar{X}_n) = \frac{\text{Var}(X_1)}{n} = \frac{\sigma^2}{n}
  \]
  Hence, for \( n \to \infty \), \( \text{Var}(\bar{X}_n) \to 0 \)
An estimate \( t \) is a value that obtained as a function \( h() \) over a dataset \( x_1, \ldots, x_n \):

\[
t = h(x_1, \ldots, x_n)
\]

- \( t = \bar{x}_n = 852.4 \) is an estimate of the speed of light
- \( t = x_1 = 850 \) is another estimate
- Since \( x_1, \ldots, x_n \) are modelled as realizations of \( X_1, \ldots, X_n \), estimates are realizations of the corresponding sample statistics \( h(X_1, \ldots, X_n) \)

An estimate \( t = h(x_1, \ldots, x_n) \) is a realization of the random variable:

\[
T = h(X_1, \ldots, X_n)
\]

The random variable \( T \) is called an estimator.

- \( T = \bar{X}_n = (X_1 + \ldots, X_n)/n \) is an estimator of the speed of light
- \( T = X_1 \) is another estimator
Parameter estimation

- The probability distribution of an estimator $T$ is called the *sampling distribution* of $T$.
- The standard deviation of the sampling distribution is called the *standard error* (SE).

**Unbiased estimator**

An estimator $T = h(X_1, \ldots, X_n)$ of some parameter $\theta$ is *unbiased* if:

$$E[T] = \theta$$

If the difference $E[T] - \theta$, called the *bias* of $T$, is non-zero, $T$ is called a *biased* estimator.

- $E[T] > \theta$ is a positive bias, $E[T] < \theta$ is a negative bias.
- Sometimes, $T$ is written as $\hat{\theta}$, e.g., $\hat{\mu}$ denotes an estimator of $\mu$.
- *When is an estimator better than another one?*
- *Is there a best possible estimator?*
When is an estimator better than another one?

Efficiency of unbiased estimators

Let $T_1$ and $T_2$ be unbiased estimators of the same parameter $\theta$. The estimator $T_2$ is more efficient than $T_1$ if:

$$\text{Var}(T_2) < \text{Var}(T_1)$$

- The relative efficiency of $T_2$ w.r.t. $T_1$ is $\frac{\text{Var}(T_1)}{\text{Var}(T_2)}$
- Speed of light example:
  - $E[X_1] = E[X_2] = \ldots = E[\bar{X}_n] = c$, i.e., all unbiased estimators
  - The mean is more efficient than a single value

$$\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n} < \sigma^2 = \text{Var}(X_1) \quad \frac{\text{Var}(X_1)}{\text{Var}(\bar{X}_n)} = n$$
Unbiased estimators for expectation and variance

Suppose \(X_1, X_2, \ldots, X_n\) is a random sample from a distribution with finite expectation \(\mu\) and finite variance \(\sigma^2\). Then

\[
\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}
\]

is an unbiased estimator for \(\mu\) and

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

is an unbiased estimator for \(\sigma^2\).

- Estimates: sample mean \(\bar{x}_n\) and sample variance \(s_n^2\) (see previous lesson)
- \(E[\bar{X}_n] = (E[X_1] + \ldots + E[X_n])/n = \mu\) and, by CLT, \(\bar{X}_n \sim N(\mu, \sigma^2/n)\)
- Why division by \(n-1\) in \(S_n^2\)? [Bessel’s correction]
\[ E[S_n^2] = \sigma \]

1. \[ E[X_i - \bar{X}_n] = E[X_i] - E[\bar{X}_n] = \mu - \mu = 0 \]
2. \[ \text{Var}(X_i - \bar{X}_n) = E[(X_i - \bar{X}_n)^2] - E[X_i - \bar{X}_n]^2 = E[(X_i - \bar{X}_n)^2] \quad \text{[by (1)]} \]
3. \[ X_i - \bar{X}_n = X_i - \frac{1}{n} \sum_{j=1}^{n} X_j = X_i - \frac{1}{n} X_i - \frac{1}{n} \sum_{j=1, j \neq i}^{n} X_j = \frac{n-1}{n} X_i - \frac{1}{n} \sum_{j=1, j \neq i}^{n} X_j \]
4. \[ \text{From (3):} \]
\[ \text{Var}(X_i - \bar{X}_n) = \frac{(n-1)^2}{n^2}\sigma^2 + \frac{1}{n^2}(n-1)\sigma^2 = \frac{n-1}{n} \sigma^2 \]

- Therefore:
\[ E[S_n^2] = \frac{1}{n-1} \sum_{i=1}^{n} E[(X_i - \bar{X}_n)^2] = \frac{1}{n-1} \sum_{i=1}^{n} \text{Var}(X_i - \bar{X}_n) = \frac{1}{n-1} n \frac{n-1}{n} \sigma^2 = \sigma^2 \]

- For normal distribution of \( X_i \)'s, \( S_n^2 \sim \text{Gam}(n-1, \sigma^2) \) and \( \text{Var}(S_n^2) = \frac{2\sigma^4}{n-1} \)
- In general, \( \text{Var}(S_n^2) \to 0 \) when \( n \to \infty \)
Degree of freedom

- For the estimator $V_n^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$:
  
  $$E[V_n^2] = E[\frac{n-1}{n} S_n^2] = \frac{n-1}{n} \sigma^2$$

- Hence, $E[V_n^2] - \sigma^2 = -\sigma^2/n$ [Negative bias]

- $V_n^2$ is asymptotically unbiased, i.e., $E[V_n^2] \to \sigma^2$ when $n \to \infty$

- Intuition on dividing by $n - 1$
  - $S_n^2$ uses in its definition $\bar{X}_n$
  - Thus, they are not independent
    - $S_n^2$ can be computed from $n - 1$ r.v. and the mean $\bar{X}_n$ (the $n$-th r.v. is implied)

- The degrees of freedom for an estimate is the number of values minus the number of parameters already estimated

- Assume that $\mu$ is known. Show that $\frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2$ is unbiased [Prove it]
Unbiasedness does not carry over

- $E[S_n^2] = \sigma^2$ implies $E[S_n] = \sigma$?
- Since $g(x) = x^2$ is convex, by Jensen’s inequality:
  \[
  \sigma^2 = E[S_n^2] = E[g(S_n)] > g(E[S_n]) = E[S_n]^2
  \]
  which implies $E[S_n] < \sigma$ [Negative bias]
- In general, if $T$ unbiased for $\theta$ does not imply $g(T)$ unbiased for $g(\theta)$
- A non-parametric (i.e., distribution free) unbiased estimator of $\sigma$ does not exist
Estimators for the median and quantiles

- \( T = Med(X_1, \ldots, X_n) \), for \( X_i \) with density function \( f(x) \)
- Let \( m \) be the true median, i.e., \( F(m) = 0.5 \):
  
  \[
  \text{for } n \to \infty, \quad T \sim N(m, \frac{1}{4nf(m)^2})
  \]

  and then for \( n \to \infty \):
  
  \[
  E[Med(X_1, \ldots, X_n)] = m
  \]

- \( T = Quantile_p(X_1, \ldots, X_n) \), for \( X_i \) with density function \( f(x) \)
- Let \( p \) quantile be the true quantile, i.e., \( F(q) = p \):
  
  \[
  \text{for } n \to \infty, \quad T \sim N(q, \frac{p(1-p)}{nf(q)^2})
  \]

  and then for \( n \to \infty \):
  
  \[
  E[Quantile_p(X_1, \ldots, X_n)] = q
  \]
Estimator for MAD

- Median of absolute deviations (MAD):
  \[ T = \text{MAD}(X_1, \ldots, X_n) = \text{Med}(|X_1 - \text{Med}(X_1, \ldots, X_n)|, \ldots, |X_n - \text{Med}(X_1, \ldots, X_n)|) \]
  - For \( X \sim F \), the population MAD is \( Md = G^{-1}(0.5) \) where \( |X - F^{-1}(0.5)| \sim G \)
  - For \( F \) symmetric, \( Md = F^{-1}(0.75) - F^{-1}(0.5) \).
  - \( Md \) is a more robust measure of scale than standard deviation

- Under mild assumptions: \( [\text{CLT for MAD}] \)
  \[ \text{for } n \to \infty, T \sim N(Md, \frac{\sigma_1^2}{n}) \]
  where \( \sigma_1 \) is defined in terms of \( Md, F^{-1}(0.5), F() \).
  - Then, for \( n \to \infty \):
    \[ E[\text{MAD}(X_1, \ldots, X_n)] = Md \]
Estimators for correlation

- **Pearson’s $r$ estimator**:

  \[
  r = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}
  \]

  \[
  \rho = \frac{E[(X - \mu_X) \cdot (Y - \mu_Y)]}{\sigma_X \cdot \sigma_Y}
  \]

  - **Fisher transformation** $F(r) = \text{arctanh}(r) = \frac{1}{2} \log \frac{1+r}{1-r}$
  - Transform a skewed sample into a normalized format
  - If $X, Y$ have a bivariate normal distribution:

    \[
    F(r) \sim N(\text{arctanh}(\rho), \frac{1}{n-3})
    \]

    Hence:

    \[
    \text{tanh}(E[F(r)]) = \rho
    \]

- **Same for Spearman’s correlation** (as it is a special case of Pearson’s)
Estimators for correlation

- Kendall’s $\tau_a$ estimator:

$$\tau_{xy} = \frac{2 \sum_{i<j} sgn(X_i - X_j) \cdot sgn(Y_i - Y_j)}{n \cdot (n - 1)}$$

$$\theta = E[sgn(X_1 - X_2) \cdot sgn(Y_1 - Y_2)]$$

- For $n > 10$, the sampling distribution is well approximated as:

$$\tau_{xy} \sim N(\theta, \frac{2(2n + 5)}{9n(n - 1)})$$

Hence:

$$E[\tau_{xy}] = \theta$$

See R script
Example: estimating the probability of zero arrivals

- \(X_1, \ldots, X_n\), for \(n = 30\), observations:
  
  \[X_i = \text{no of arrivals (of a packet, of a call, etc.) in a minute}\]

- \(X_i \sim \text{Pois}(\mu)\), where \(p(k) = P(X = k) = \frac{\mu^k}{k!} e^{-\mu}\) \([E[X] = \mu]\)

- We want to estimate \(p_0 = p(0)\), probability of zero arrivals

- Frequentist-based estimator \(S\):
  
  \[S = \frac{|\{i | X_i = 0\}|}{n}\]

  - Takes values \(0/30, 1/30, \ldots, 30/30\) ... may not exactly be \(p_0\)
  - \(S = Y/n\) where \(Y = I_{X_1=0} + \ldots + I_{X_n=0} \sim \text{Bin}(n, p_0)\)
  - Hence, \(E[S] = \frac{1}{n} E[Y] = \frac{n}{n} p_0 = p_0\) \([S \text{ is unbiased}]\)
Example: estimating the probability of zero arrivals

- Since \( p_0 = p(0) = e^{-\mu} \), we devise a mean-based estimator \( T \):

\[
T = e^{-\bar{X}_n}
\]

- By Jensen’s inequality:

\[
E[T] = E[e^{-\bar{X}_n}] > e^{-E[\bar{X}_n]} = e^{-\mu} = p_0
\]

Hence \( T \) is biased!

- \( T = e^{-Z/n} \) where \( Z = X_1 + \ldots + X_n \) is the sum of \( \text{Poi}(\mu) \)'s, hence \( Z \sim \text{Poi}(n \cdot \mu) \)

\[
E[T] = \sum_{k=0}^{\infty} e^{-\frac{k}{n}} \frac{(n\mu)^k}{k!} e^{-n\mu} = e^{-n\mu(1-e^{-1/n})} \rightarrow e^{-\mu} = p_0 \text{ for } n \rightarrow \infty
\]

Hence \( T \) is asymptotically unbiased!

[Exercise 19.9]

See R script
Example: estimating the probability of zero arrivals

- Let's look at the variances:

\[ Var(S) = \frac{1}{n} Var(Y) = \frac{np_0(1 - p_0)}{n^2} = \frac{p_0(1 - p_0)}{n} \rightarrow 0 \text{ for } n \rightarrow \infty \]

\[ Var(T) = E[T^2] - E[T]^2 = \ldots \text{ exercise} \ldots \rightarrow 0 \text{ for } n \rightarrow \infty \]

See R script
MSE: Mean Squared Error of an estimator

What if one estimator is unbiased and the other is biased but with a smaller variance?

\[ \text{MSE}(T) = E[(T - \theta)^2] \]

- An estimator \( T_1 \) performs better than \( T_2 \) if \( \text{MSE}(T_1) < \text{MSE}(T_2) \)
- Note that:

\[
\begin{align*}
\text{MSE}(T) &= E[(T - E[T] + E[T] - \theta)^2] \\
&= E[(T - E[T])^2] + (E[T] - \theta)^2 + 2E[T - E[T]](E[T] - \theta) \\
&= \text{Var}(T) + (E[T] - \theta)^2
\end{align*}
\]

- \( E[T] - \theta \) is called the bias of the estimator
- Hence, \( \text{MSE} = \text{Var} + \text{Bias}^2 \)
- A biased estimator with a small variance may be better than an unbiased one with a large variance!

See R script