Master Program in *Data Science and Business Informatics*

Statistics for Data Science

Lesson 18 - Unbiased estimators. Efficiency and MSE

Salvatore Ruggieri

Department of Computer Science
University of Pisa, Italy
salvatore.ruggieri@unipi.it

Statistical model for repeated measurement

- 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

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 α its parameter(s).

- Challenging questions/inferences on a population given a sample:
  - How to determine $E[X]$, $Var(X)$, or other functions of X?
  - How to determine α, assuming to know the form of F?
  - How to determine both F and α?

An example

Table 17.1. Michelson data on the speed of light.

850	740	900	1070	930	850	950	980	980	880
1000	980	930	650	760	810	1000	1000	960	960
960	940	960	940	880	800	850	880	900	840
830	790	810	880	880	830	800	790	760	800
880	880	880	860	720	720	620	860	970	950
880	910	850	870	840	840	850	840	840	840
890	810	810	820	800	770	760	740	750	760
910	920	890	860	880	720	840	850	850	780
890	840	780	810	760	810	790	810	820	850
870	870	810	740	810	940	950	800	810	870

- What is an estimate of the true speed of light (estimand)?

$$x_1 = 850, \text{ or } \mathit{min}\ x_i, \text{ or } \mathit{max}\ x_i, \text{ or } \bar{x}_n = 852.4 \ ?$$

An example

- Speed of light dataset as realization of

$$X_i = c + \epsilon_i$$

  where ϵ_i is measurement error with $E[\epsilon_i] = 0$ and $Var(\epsilon_i) = \sigma^2$
- We are then interested in $E[X_i] = c$
- How to estimate it?
- Use some data. For X_1:

$$E[X_1] = c \qquad Var(X_1) = \sigma^2$$

- Use all data. For $\bar{X}_n = (X_1 + \ldots + X_n)/n$:

$$E[\bar{X}_n] = c \qquad Var(\bar{X}_n) = \frac{Var(X_1)}{n} = \frac{\sigma^2}{n}$$

  Hence, for $n \to \infty$, $Var(\bar{X}_n) \to 0$

Estimate

Estimand and estimate

An *estimand* θ is an unknown parameter of a distribution $F()$.
An *estimate* t of θ 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 (estimand) $\quad t = x_1 = 850$ 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 $h(X_1, \ldots, X_n)$

Statistics and estimator

A *statistics* is a function of $h(X_1, \ldots, X_n)$ of r.v.'s.
An *estimator* of a parameter θ is a statistics $T_n = h(X_1, \ldots, X_n)$ intended to provide information about θ.

- An estimate $t = h(x_1, \ldots, x_n)$ is a realization of the estimator $T_n = h(X_1, \ldots, X_n)$
- $T_n = \bar{X}_n = (X_1 + \ldots, X_n)/n$ is an estimator of $\mu \quad T_n = X_1$ is another estimator

Unbiased estimator

- The probability distribution of an estimator T is called the *sampling distribution* of T

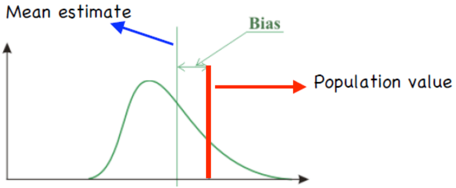
Unbiased estimator

An estimator $T_n = h(X_1, \ldots, X_n)$ of a parameter θ (estimand) is *unbiased* if:

$$E[T_n] = \theta$$

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

- $E[T_n] > \theta$ is a positive bias, $E[T_n] < \theta$ is a negative bias
- **Asymptotically unbiased:** $\lim_{n\to\infty} E[T_n] = \theta$
- Sometimes, T_n written as $\hat{\theta}$, e.g., $\hat{\mu}$ estimator of μ



On $E[T]$

- Random sample i.i.d. $X_1, \ldots, X_n \sim F(\alpha)$
- $E[T] = E[h(X_1, \ldots, X_n)]$ over the joint distribution $\prod_{i=1}^{n} F(\alpha) = F(\alpha)^n$
- E.g., for $F()$ continuous with d.f. $f()$

$$E[T] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(x_1, \ldots, x_n) f(x_1) \ldots f(x_n) dx_1, \ldots, dx_n$$

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 θ. The estimator T_2 is *more efficient* than T_1 if:

$$Var(T_2) < Var(T_1)$$

- The *relative efficiency* of T_2 w.r.t. T_1 is ${}^{Var(T_1)}/{}_{Var(T_2)}$
- 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

$$Var(\bar{X}_n) = \sigma^2/n < \sigma^2 = Var(X_1) \qquad \frac{Var(X_1)}{Var(\bar{X}_n)} = n$$

- The standard deviation of the sampling distribution is called the *standard error* (SE)
  - The SE of the mean estimator $\bar{X}_n$ is $\sigma/\sqrt{n}$

Unbiased estimators for expectation and variance

UNBIASED ESTIMATORS FOR EXPECTATION AND VARIANCE. Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a distribution with finite expectation μ and finite variance σ^2. Then

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

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.

- Estimates: sample mean $\bar{x}_n$ and sample variance s_n^2
- $E[\bar{X}_n] = (E[X_1] + \ldots + E[X_n])/n = \mu$ and, by CLT, $Var(\bar{X}_n) \to 0$ for $n \to \infty$
- Why division by $n-1$ in S_n^2? *[Bessel's correction]*

$E[S_n^2] = \sigma^2$ and Bessel's correction

(1) $E[X_i - \bar{X}_n] = E[X_i] - E[\bar{X}_n] = \mu - \mu = 0$

(2) $Var(X_i - \bar{X}_n) = E[(X_i - \bar{X}_n)^2] - E[X_i - \bar{X}_n]^2 = E[(X_i - \bar{X}_n)^2]$ *[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) From (3):

$$Var(X_i - \bar{X}_n) = \frac{(n-1)^2}{n^2}\sigma^2 + \frac{1}{n^2}(n-1)\sigma^2 = \frac{n-1}{n}\sigma^2$$

- Therefore:

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

- **In general**: $Var(S_n^2) = \frac{1}{n}(\mu_4 - \frac{n-3}{n-1}\sigma^4) \to 0$ for $n \to \infty$

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, $(X_i - \bar{X}_n)$'s 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 observations n minus the number of parameters already estimated
- Assume that μ is known. Show that $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ is unbiased **[Exercise at home]**

Unbiasedness does not carry over (no functional invariance)

- $E[S_n^2] = \sigma^2$ 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 θ does not imply $g(T)$ unbiased for $g(\theta)$
  - But it holds for $g()$ linear transformation!
- A non-parametric (i.e., distribution free) unbiased estimator of σ **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., $F(m) = 0.5$: **[CLT for medians]**

$$\text{for } n \to \infty, T \sim N(m, \frac{1}{4nf(m)^2})$$

and then for $n \to \infty$:

$$E[Med(X_1, \ldots, X_n)] = m$$

- $T = q_{X_1,\ldots,X_n}(p)$, for X_i with density function $f(x)$
- Let q_p be the true p-quantile, i.e., $F(q_p) = p$: **[CLT for quantiles]**

$$\text{for } n \to \infty, T \sim N(q_p, \frac{p(1-p)}{nf(q_p)^2})$$

and then for $n \to \infty$:

$$E[q_{X_1,\ldots,X_n}(p)] = q_p$$

See R script

Estimator for MAD

- Median of absolute deviations (MAD):

$$T = MAD(X_1, \dots, X_n) = Med(|X_1 - Med(X_1, \dots, X_n)|, \dots, |X_n - Med(X_1, \dots, X_n)|)$$

  - 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: **[CLT for MADs]**

$$\text{for } n \to \infty, T \sim N(Md, \frac{\sigma_1^2}{n})$$

where σ_1 is defined in terms of $Md, F^{-1}(0.5), F()$, and then for $n \to \infty$:

$$E[MAD(X_1, \dots, X_n)] = Md$$

Estimators for correlation

- Pearson's r estimator:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \cdot \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad \rho = \frac{Cov(X, Y)}{\sigma_X \cdot \sigma_Y}$$

  - The sampling distribution of the estimator is highly skewed!
  - **Fisher transformation** $FisherZ(r) = \frac{1}{2} \log \frac{1+r}{1-r}$
  - Transform a skewed sample into a normalized format.
  - If X, Y have a bivariate normal distribution:

$$FisherZ(r) \sim N(FisherZ(\rho), \frac{1}{n-3})$$

  Hence:

$$FisherZ^{-1}(E[FisherZ(r)]) = \rho$$

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

Estimators for correlation

- Kendall's τ_a estimator:

$$\tau_{xy} = \frac{2\sum_{i<j} sgn(X_i - X_j) \cdot sgn(Y_i - Y_j)}{n \cdot (n-1)} \qquad \theta = E_{X_1,X_2 \sim F_X, Y_1,Y_2 \sim F_Y}[sgn(X_1 - X_2) \cdot sgn(Y_1 - Y_2)]$$

  - 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:

$$X_i = \text{ number of arrivals (of a packet, of a call, etc.) in a minute}$$

- $X_i \sim Pois(\mu)$, 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:

$$S = \frac{|\{i \mid X_i = 0\}|}{n}$$

  - Takes values $0/30, 1/30, \ldots, 30/30$ ...may not exactly be p_0
  - $S = Y/n$ where $Y = \mathbb{1}_{X_1=0} + \ldots + \mathbb{1}_{X_n=0} \sim Bin(n, p_0)$
  - Hence, $E[S] = \frac{1}{n} E[Y] = \frac{n}{n} p_0 = p_0$ *[S 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 $Poi(\mu)$'s, hence $Z \sim Poi(n \cdot \mu)$

    Prove it *by doing [T, Exercise 11.2]*

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

    - since $\sum_{k=0}^{\infty} \frac{x^k}{k!} = e^x$ and $\lim_{n\to\infty} n(1 - e^{-1/n}) = 1$

    Hence T is asymptotically unbiased!

See R script

Example: estimating the probability of zero arrivals

- Let's look at the variances:

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

$$Var(T) = E[T^2] - E[T]^2 = \ldots \textbf{ exercise at home } \ldots \to 0 \text{ for } n \to \infty$$

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?

MSE

The Mean Squared Error of an estimator T for a parameter θ is defined as:

$$MSE(T) = E[(T - \theta)^2]$$

- An estimator T_1 performs better than T_2 if $MSE(T_1) < MSE(T_2)$
- Note that:

$$MSE(T) = E[(T - E[T] + E[T] - \theta)^2] =$$
$$= E[(T - E[T])^2] + (E[T] - \theta)^2 + 2E[T - E[T]](E[T] - \theta) = Var(T) + (E[T] - \theta)^2$$

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

See R script

Best estimators

Consistent estimator

An estimator T_n is a squared error consistent estimator if:

$$\lim_{n \to \infty} MSE(T_n) = 0$$

- Hence, for $n \to \infty$, both *Bias* and *Var* converge to 0
- $\bar{X}_n$ is a squared error consistent estimator of μ
- What if there is no consistent estimator or if there are more than once?

MVUE

An unbiased estimator T_n is a Minimum Variance Unbiased Estimators (MVUE) if:

$$Var(T_n) \leq Var(S_n)$$

for all unbiased estimators S_n.

- **Corollary.** $MSE(T_n) \leq MSE(S_n)$
- $\bar{X}_n$ is a MVUE of μ if $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ [proof in the next lesson]