CEE 260/MIE 273: Probability and Statistics in Civil Engineering

Lecture M4a: Point Estimates, Sampling Variability and Central Limit Theorem

Jimi Oke

UMass Amherst

College of Engineering

October 14, 2025

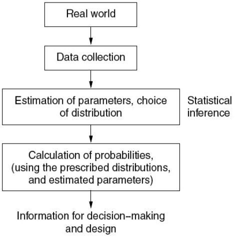
Outline

1. Statistical inference
2. Point estimation
3. Method of moments
4. Variability and CLT
5. Outlook

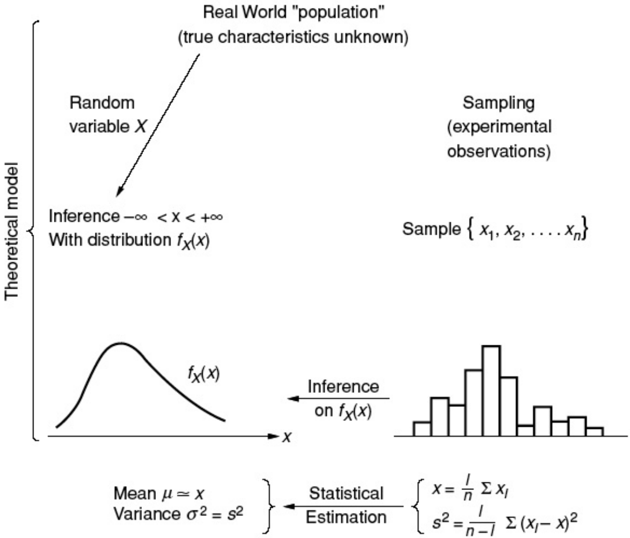
Statistical inference

To develop probabilistic models from observational data, we need to *estimate* the statistical parameters and probabilities of the distributions.

- In most applications, the true population is unknown
- Estimates are obtained from representative **samples**



Role of sampling in statistical inference



Statistical inference

This module (M4) covers concepts in statistical inference:

- Point estimates and sampling variability (M4a; today)
- Confidence intervals for a proportion (M4b)
- Hypothesis testing for a proportion (M4c)

Point estimates

Definition

A **point estimate** of a parameter θ (e.g. proportion p, or mean value μ) is a single number that can be regarded as a sensible value for θ and is obtained by computing the value of a suitable statistic from given sample data. The selected statistic is the **point estimator** of θ.

Notation

- $\hat{\Theta}$: point estimator (pronounced *theta hat*)
- $\hat{\theta}$: point estimate

$$\hat{\theta} = \theta + \text{ estimation error} \tag{1}$$

- A hat can be placed on the actual statistic estimated for clarity, e.g.

$$\hat{p} = \overline{X}$$

Properties of point estimators

Desired properties of a point estimator:

- Unbiasedness
- Consistency
- Efficiency
- Sufficiency

Desired properties of point estimators

Unbiasedness

An estimator is *unbiased* if its expected value is equal to the true value of the parameter it estimates:

$$\mathbb{E}(\hat{\theta}) = \theta \quad \text{(if } \hat{\theta} \text{ is unbiased)} \tag{2}$$

Thus, the **bias** is given by:

$$\text{Bias}_{\hat{\theta}} = \mathbb{E}(\hat{\theta}) - \theta \tag{3}$$

Consistency

An estimator is consistent if $\hat{\theta} \rightarrow \theta$ as $n \rightarrow \infty$,i.e. the estimation error should decrease with increasing sample size.

Desired properties of point estimators

Efficiency

The efficiency of an estimator is defined by how small its variance is.

Sufficiency

A sufficient estimator uses all the relevant information in a given sample in its estimation.

In many applications, efficiency (low variance) and unbiasedness (low bias) are the most important properties of an estimator.

Bias vs. variance

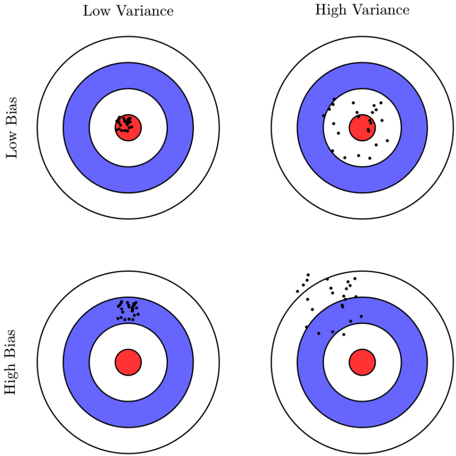


Image source: https://tex.stackexchange.com/a/307285/2269

Sample moments

- The moments of a random variable are its key descriptors.
- Parameters of the distribution of a random variable are usually related to the first and second moments (mean and variance, respectively)

Given a sample $x_1, x_2, \ldots, x_n$, the point estimates of the population mean μ and variance σ^2 are:

Sample mean

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{4}$$

Sample variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{5}$$

Unbiasedness of s^2

From Equation (5), you can show (as an exercise) that:

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right] \tag{6}$$

You may be wondering why the sample variance is not just the average of the sum of squared deviations from the sample mean. But

$$s^2 = E\left(\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right) \quad = \quad \sigma^2 \tag{7}$$

$$\hat{\sigma}^2 = E\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right) \quad = \quad \frac{n-1}{n}\sigma^2 \tag{8}$$

The second estimator is biased $(-\frac{\sigma^2}{n})$ and underestimates σ^2.

Sample mean and variance

Example 1: Elastic modulus of alloys

The elastic modulus (GPa) of a sample of alloy specimens from a die-casting process is:

$$X = 44.2, 43.9, 44.7, 44.2, 44.0, 43.8, 44.6, 43.1$$

(a) Estimate the population mean using the estimator $\overline{x}$ (sample mean)

(b) Estimate the population variance using the estimator s^2 (sample variance)

(c) Now, estimate the variance replacing the denominator $(n-1)$ with n in the estimator s^2. What do you notice?

Sample mean and variance

Example 1: Elastic modulus of alloys (cont.)

$X = 44.2, 43.9, 44.7, 44.2, 44.0, 43.8, 44.6, 43.1$

(a) $\hat{\mu} = \overline{x} = \frac{1}{8}\sum_{i=1}^{8} x_i \approx \boxed{44.063}$

(b) $s^2 = \frac{1}{7}\left[\sum_{i=1}^{8} x_i^2 - 8(44.063^2)\right] \approx \boxed{0.251}$

(c) Biased estimate of σ^2:

$\hat{\sigma}^2 = \frac{1}{8}\left[\sum_{i=1}^{8} x_i^2 - 8(44.063^2)\right] = \frac{7}{8}(0.251) = \boxed{0.220}$

$\hat{\sigma}^2$ underestimates σ^2 by 0.031 squared units.

Variability of a point estimate

Example 2: Solar energy expansion

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest. Develop a simulation to investigate how the sample proportion $\hat{p}$ behaves compared to the true population proportion p:

- **(a)** Create a set of a large number of entries (e.g. 300 million) where 88% are in support and 12% are not.
- **(b)** Sample $n = 1000$ entries without replacement
- **(c)** Plot the histogram of the sampling distribution of $\hat{p}$
- **(d)** Compute the sample mean $x_{\hat{p}}$
- **(e)** Compute the standard deviation $s_{\hat{p}}$ (called the **standard error** $SE_{\hat{p}}$).
- **(f)** Investigate what happens as n increases.

The Central Limit Theorem (CLT)

Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. If n is sufficiently large, then the sample mean $\overline{X}$ has approximately a normal distribution with $\mu_{\overline{X}} = \mu$ and $\sigma^2_{\overline{X}} = \sigma^2/n$; and the sample total ($S_n = X_1 + X_2 + \ldots + X_n$) has approximately a normal distribution with

$$\mu_S = n\mu \tag{9}$$

$$\sigma_S^2 = n\sigma^2 \tag{10}$$

Implications:

- The sum of a **large number** of random components approaches a **normal/Gaussian distribution**
- The product of large number of random components approaches the lognormal distribution

Central limit theorem (cont.)

Sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \tag{11}$$

Sum of sample observations

$$S_n = X_1 + X_2 + \cdots + X_n \tag{12}$$

If n is sufficiently large for **any** sample:

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \tag{13}$$

$$S_n \sim \mathcal{N}(n\mu, n\sigma^2) \tag{14}$$

Note that the quantity $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ is also known as the **sampling error** (SE) or the **standard error of the mean** (SEM)

Sample proportion and the CLT

If the observations in a given sample are a Bernoulli sequence with a constant proportion (or probability) p, then if n is large, the sample proportion $\hat{p}$ follows a normal distribution (according to the CLT):

$$\hat{p} \sim \mathcal{N}(\mu_{\hat{p}}, SE_{\hat{p}}^2) = \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \tag{15}$$

where

$$\begin{aligned} \text{Sample mean proportion: } \mu_{\hat{p}} &= p \\ \text{Sampling error/standard error of } \hat{p}\text{: } SE_{\hat{p}} &= \sqrt{\frac{p(1-p)}{n}} \end{aligned}$$

One rule of thumb for determining whether n is large enough is to check that both np and $n(1-p)$ are ≥ 10 (also known as the success-failure condition).

CLT application: sample proportion

Example 3: Solar energy expansion (CLT)

Suppose the proportion of American adults who support the expansion of solar energy is $p = 0.88$, which is our parameter of interest. If we were to take a poll of 1000 American adults on this topic, the estimate would not be perfect, but how close might we expect the sample proportion in the poll would be to 88%?

(a) According to the CLT, what is the distribution of $\hat{p}$?

(b) According to the CLT, what are $\mu_{\hat{p}}$ and $SE_{\hat{p}}$, respectively?

CLT application: sample proportion (cont.)

Example 3: Solar energy expansion (CLT)

(a) First, we note that the response of each American adult in the entire population is part of a Bernoulli sequence with $p = 0.88$. According to the CLT, the distribution of $\hat{p}$ (the mean proportion based on the sample) is normal/Gaussian. We can denote this as:

$$\hat{p} \sim \mathcal{N}\left(p, \frac{\sigma^2}{n}\right) \text{ OR } \mathcal{N}\left(\mu_p, \frac{\sigma_p^2}{n}\right) \tag{16}$$

(b) $\mu_{\hat{p}}$ denotes the mean estimate of p, which is 0.88 (according to the CLT, the mean of the sample is the population mean if n is large).
$SE_{\hat{p}}$ denotes the sampling error, which is the the square root of the variance of the sample mean: $\sqrt{\sigma^2/n}$. Given that the sample is governed by the Binomial distribution with $\sigma^2 = p(1-p)$. Thus:

$$SE_{\hat{p}}^2 = \frac{\sigma^2}{n} = \frac{p(1-p)}{n} = \frac{0.88(0.12)}{1000}$$

$$\sqrt{\frac{0.88(0.12)}{\ldots}}$$

Success-failure condition

In the case of a proportion p, the CLT holds only if:

- The observations are independent (i.e. random)
- The sample size n is **sufficiently large**

The second condition is typically observed via the **success-failure condition**, i.e.:

$$np \geq 10 \tag{17}$$
$$n(1-p) \geq 10 \tag{18}$$

Another application of the CLT

Example 4: Mean batch weight

A certain brand of cement is shipped in batches of 40 bags. Previous records indicate the weight of a randomly selected bag of this brand has a mean of 2.5 kg and an SD of 0.1 kg. The exact distribution is unknown.

(a) What is the mean weight of one batch of this brand of cement?

(b) If the shipping company charges an overweight fee if a batch exceeds the mean batch weight by more than 1 kg, what is the probability that a batch will be charged?

Another application of the CLT

Example 4: Mean batch weight (cont.)

Let B be the total weight of one batch.

(a) The mean weight of one batch is thus

$$\mu_B = 40 \times 2.5 = 100 \text{ kg} \tag{19}$$

(b) By the CLT, B is approximately normal with $\mu_B = 100$ and $\sigma_B^2 = 40(0.1)^2$.The probability a batch will be charged is:

$$\begin{aligned} P(B > 101) &= 1 - \Phi\left(\frac{101 - 100}{0.1\sqrt{40}}\right) \\ &= 1 - \Phi(1.581) \\ &= 1 - 0.9431 \approx \boxed{5.69\%} \end{aligned}$$

Summary

- Desired properties of point estimates: unbiasedness and efficiency
- Distribution of sample proportions (or other parameters) is called a sampling distribution
- When n is sufficiently large and observations are independent, the sample proportion (or other parameter) follows a normal distribution
- The success-failure condition can be used to determine if n is large enough for the CLT to hold (for a sample proportion)

Simulation

Simulation is the process of representing (modeling) a hypothetical process using a compute in order to compute or evaluate outcomes absent of real-life experimentation.

Uses of simulation

- Estimating probabilities
- Estimating statistical parameters for a given population, such as mean, median and variance
- Finding the bias (difference between true and estimated value) of a given parameter
- Synthesizing data
- Conducting experiments

Example 4: Estimating probability via simulation

A motorist is driving at the posted maximum speed along a stretch of road. The probability that the motorist approaches a red light at the first intersection (A) is 0.4. The probability that the motorist encounters a red light *again* at the next intersection (B) is 0.7. If the motorist does not encounter a red light at the first intersection (A), then the probability the motorist encounters a red light at the second intersection (B) is 0.2. If in a certain instance the motorist is observed to have encountered a red light at intersection B, what is the probability the motorist encountered a red light at intersection A?

(a) First use Bayes' Theorem to compute the required probability. Recall:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})} \tag{20}$$

(b) Perform a simulation to estimate this probability.

Example 4: Estimating probability via simulation (cont.)

Solution

(a) From the problem, we know that $P(A) = 0.4$, $P(B|A) = 0.7$, $P(B|\overline{A}) = 0.2$. Thus,

$$\begin{aligned}
P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|\overline{A})P(\overline{A})} \\
&= \frac{0.7(0.4)}{0.7(0.4) + (0.2)(1-0.4)} \\
&= \frac{0.28}{0.28 + 0.12} \\
&= \boxed{0.7}
\end{aligned}$$

(b) For the simulation, see `M4_matlab_example4.m`