STOCHASTIC METHODS IN WATER RESOURCES

Unit 1: Introduction to probability and statistics

Lecture 4a: Model estimation and testing

Luis Alejandro Morales, Ph.D.

Universidad Nacional de Colombia
Department of Civil and Agriculture Engineering

August 25, 2025

Generalities

- Statistical inference deals with statistical estimations based on a sample from the population.
- Some definitions:
  - **Population**: consist of all possible observations of a process (e.g. air temperature at certain location). Some of the observations in the in the population may not have any physical sense, perhabs, due to sensor errors.
  - **Sample**: is a subset of the population (e.g. instantaneous daily streamflow for a certain period in a station). A random sample is thus a sample that is representative of the population.
  - **Random variables**: is a real-valued function defined on a sample space. Wheather a random variable is discrete or continuous depends on how the sample space is defined.
  - **Statistic**: is a function of the observations that is quantifiable and does not contain any unknown parameter. Note that a statistic is also a random variable that provides an estimation.
  - **Estimator**: is the method or rule of estimation. For instance, the sample mean $\bar{X}$ is a point estimator of the population mean μ.
  - **Estimate**: is the value yielded by the estimator.
- Suppose that the population of variable X follows a Normal distribution and the distribution parameters θ are unknown. Thus, a random sample of X of size n.
- Parameters θ can be described by a number or a range; this last include an uncertainty.

Properties of estimators

Unbiasedness

Given a sample of observations of a random variable X, $X_1, X_2, \cdots, X_n$, the objective is to estimate the value of the parameter θ, which is also a random variable. The idea is to seek for a estimator of θ that get as closest as possible to the true value. This estimator will produce statistics that are distributed following a certain pdf. Following this, a **point estimator** $\hat{\theta}$ is an **unbiased estimator** of the population parameter θ if $E[\hat{\theta}] = \theta$. If the estimator is biased, the bias $= E[\hat{\theta}] - \theta$.

Mean of the sample mean

Let us show that the sample mean $\bar{X}$ is an unbiased estimator of μ. If $\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$, then $E[\bar{X}] = \frac{1}{n} nE[X_i] = \frac{1}{n}(n\mu) = \mu$.

Note that many estimator are biased and methods such as jackknife and bootstrap are used to reduce the bias. Ideal estimator must also have the following properties.

Consistency

An estimator $\hat{\theta}_n$, based on a sample size n, is a **consistent** estimator of θ if for any positive number ε $\lim_{n\to\infty} Pr[|\hat{\theta}_n - \theta| \leq \varepsilon] = 1$

Minimum variance

A part for looking for an unbiased estimator, it is also desirable to seek for an **minimum variance estimator**. The minimum variance unbiased estimator is the estimator with the smallest variance out of all unbiased estimator

Properties of estimators

Efficiency

Efficiency is the relative measure of the variance of the sampling distribution, with the efficiency increasing as the variance decrease. In terms of the mean square error (mse) an estimator with the minimum *mse* among all possible unbiased estimators is called and **efficient estimator**. If A is an estimator of θ, the *mse* is:

$$\begin{aligned} E\left[(A-\theta)^2\right] &= E\left[((A-E[A])-(\theta-E[A]))^2\right] \\ &= E[(A-E[A])^2]+(\theta-E[A])^2 \\ &= Var[A]-bias^2 \end{aligned}$$

The *mse* of an estimator can be used as a relative measure of efficiency when comparing two or more estimator.

Sufficiency

So far, properties such as unbiasedness, consistency and the minimum *mse* are key to select the most suitable estimators. Thus, a **sufficient estimator** provides as much information as possible about a sample of observations of a random variable X. Let a sample $X_1, X_2, \cdots, X_n$ be drawn randomly from population having probability distribution with unknown parameters θ. Thus, the statistic $T = f(X_1, X_2, \cdots, X_n)$ is said to be sufficient to estimate θ if the distribution of $X_1, X_2, \cdots, X_n$ conditional to the statistics T is independent of θ.

Median and mean

The median, taken as a measure of mean density or central tendency, does not contain all the information in a sample. The median is the middle value of the sample; if any other value is changed the mean changes but the median is unaltered.

Estimation of confidence intervals

The uncertainty of point estimates can be quantified by the relative variances or mean square errors of the estimators. Because of this uncertainty, the next step of inference is **interval estimation**. Two numbers, say, a and b, that are expected to include within their range an unknown parameter θ in a specified percentage of cases after repeated experimentation under identical conditions. That is, in place of one statistic that estimates θ, we find a range specified by two statistics, which includes it at a given level of probability. The end points a and b of this range are known as **confidence limits**, and the interval (a, b) is known as the **confidence interval**. We do not have the precision as for a point estimator but we have confidence (without absolute certainty) that the assertion is right.

Confidence interval estimation of the mean when the standard deviation is known

Let C_l and C_u be the lower and upper confidence limits that include an unknown but invariable parameter θ. Although there is some uncertainty associated with this statement, we will be right in a proportion, say, $1-\alpha$, of the samples taken from the same population, on average, when we make the assertion that the given interval includes θ. Thus we can say, by adopting the long-run frequency interpretation of probability,

$$Pr[C_l \leq \theta \leq C_u] = 1 - \alpha$$

where θ is a constant, but the estimator $\hat{\theta}$ and the confident limits C_l and C_u are random variables. Recall that examples of $\hat{\theta}$ are $\bar{X}$ and $\bar{S}^2$. The probability $(1-\alpha)$ is known as the confidence level or confidence coefficient. It often takes values such as 0.99, 0.95, and 0.90. The confidence limits C_l and C_u depend on the sampling distribution of θ. The standard deviation of the statistic $\hat{\theta}$ called its standard error.

Estimation of confidence intervals

Confidence interval estimation of the mean when the standard deviation is known

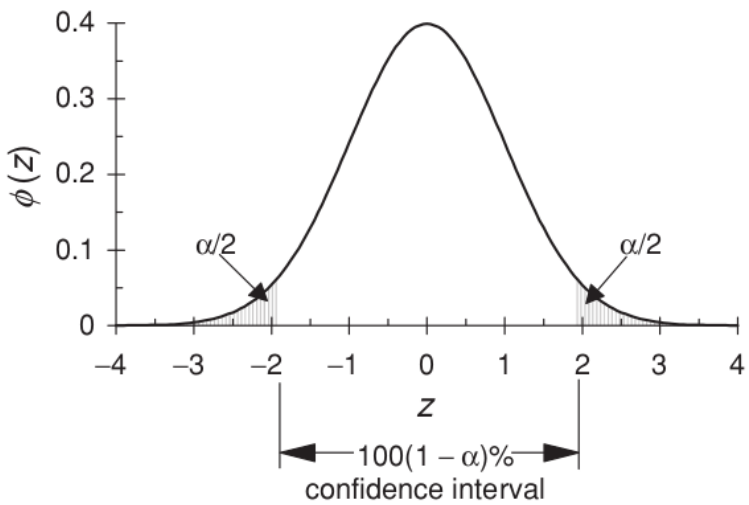
In some situations, we may require one-sided confidence limits. The lower and upper one-sided confidence limits are specified, respectively, by

$$Pr[C_l \leq \theta] = 1 - \alpha \quad \text{for } 0 < \alpha < 1$$

and

$$Pr[\theta \leq C_u] = 1 - \alpha \quad \text{for } 0 < \alpha < 1$$

Let $\bar{X}$ be the mean of a random sample of size n drawn from a population with known standard deviation σ. The $100(1-\alpha)$ percent central two-sided confidence interval for the population mean μ is $\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$, where $z_{\alpha/2}$ is a standard normal variate that is exceeded with probability $\alpha/2$, $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ $N \sim (0, 1)$. The one-sided upper and lower $100(1-\alpha)$ percent confidence limits for the population mean μ are, $\bar{X} + z_\alpha \frac{\sigma}{\sqrt{n}}$ and $\bar{X} - z_\alpha \frac{\sigma}{\sqrt{n}}$, respectively.



Estimation of confidence intervals

Confidence interval estimation of the mean when the standard deviation is unknown

Quite often in practice the mean and standard deviation are both unknown. Under such conditions we must modify our approach. We assume normality of the variate X as before, but the consequences of a nonnormal distribution are minor for small to moderate departures from normality, if the sample size is large, say, $n > 30$. In this situation we apply the Student's t distribution.