DASC7011 Statistical Inference for Data Science

Chapter 1 Estimation and Hypothesis Test — A Review

Department of Statistics and Actuarial Science
Department of Computer Science
The University of Hong Kong

August 2024

Contents

§1.1 An Introduction

- Data science is an interdisciplinary academic field that uses *statistics*, scientific *computing*, scientific methods, *processes*, *algorithms* and *systems* to extract or **extrapolate knowledge and insights** from noisy, structured, and unstructured data.[1]
- Data science combines *math* and *statistics*, specialized *programming*, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to **uncover** actionable **insights** hidden in an organization's data.[2]
- The ability to take data – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it ······[3]

[1] Wikipedia
[2] IBM
[3] School of Information, UC Berkeley

Subjects in data science

- Mathematics – scientific computations and methods such as approximations and optimizations.
- Computer programming – processes, systems, implementation.
- Statistics – ideas, reasoning.
  - What to compute?
  - What and how to uncover?
  - **Soul of data science**.
- Algorithms are frequently mentioned as integrated tools in data science. An **algorithm** is an unambiguous specification/rule of how to solve a class of problems, such as *calculation*, *data processing*, *automated reasoning*, etc. [1]

Example statistical ideas

- *Averaging*: using averages to estimate expectations, such as method-of-moments estimation (MME), GMM, ...
- *The most accurate*: minimizing certain loss functions (*inaccuracy*), such as least-squares estimation (LSE), least-absolute deviations, ...
- *The most possible*: maximizing the possibility, a typical example is the maximum likelihood estimation (MLE),
- *Learning/Updating*: Bayesian inference (getting more and more accurate or possible),
- *Logically reasoning*: question answering systems used in Intelligent Humanoid Robot,
- ......

Statistics and Statistical inference

- **Statistics** (from German: *Statistik*, orig. "*description of a state, a country*") is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data. [1]
- **Statistical inference** is the process of using data (sample) analysis (algorithms) to deduce properties of an underlying statistical model (population).
  - Estimation
  - Prediction/Forecast
  - Hypothesis testing
  - Model selection
  - Reasoning
  - ······

Statistical inference in AI and DS

- Consider an example of predicting stock prices via various ML/DL models: CNN, RF, Logistic Regression, etc.
- Output probabilities of future trend: Very Weak (VW), Weak (W), Neutral (N), Strong (S), and Very strong (VS).
- Processes of estimation and/or prediction are usually black boxes.
- Our concerns:
  - Methods/Criteria of estimation/optimization.
  - Accuracy of predictions: confidence intervals, or testing results.
  - Decisions based on predicted probabilities – comparison: hypothesis testing.
  - Model selection.
  - ······

§1.2 Probabilistic Convergence

- Various types of probabilistic convergence are frequently used in statistical inferences such as estimation and hypothesis testing.
- The following topics are briefly introduced in this section.
  - Convergence in distribution
  - Convergence in probability
  - Almost sure convergence
  - Convergence in mean
  - Law of large numbers
  - Central limit theorems
- Through out this section, let $X \sim F$ be a random variable with a cumulative distribution function (cdf) $F(x)$, and $\{X_n : n \geq 1\}$ be a sequence of random variables with cdfs $F_n(x)$, respectively.

Convergence in distribution

1.2.1. Convergence in distribution

- Convergence in distribution is in some sense the weakest type of probabilistic convergence.
- If, for any x at which $F(\cdot)$ is continuous,

$$\lim_{n\to\infty} F_n(x) = F(x), \tag{1.1}$$

  we say that $\{X_n\}$ **converge to X in distribution**, and denote this as $X_n \xrightarrow{\mathrm{d}} X$ (as $n \to \infty$).
- It is noticeable that $X_n \xrightarrow{\mathrm{d}} X$ does NOT imply that X_n converges to X in values or in probability. An *obvious* example is that $\{X, X_n : n \geq 1\} \overset{i.i.d.}{\sim} N(0,1)$.

1.2.2. Convergence in probability

- If, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(|X_n - X| > \varepsilon\big) = 0, \tag{1.2}$$

  we say we say that $\{X_n\}$ **converge to X in probability**, and denote this as $\operatorname*{plim}_{n\to\infty} X_n = X$ or $X_n \xrightarrow{\mathrm{P}} X$.

- **Theorem 1.1**: If $X_n \xrightarrow{\mathrm{P}} X$, then $X_n \xrightarrow{\mathrm{d}} X$.
  (*Students may find a proof of this online, e.g., in Wikipedia.*)

- However, convergence in probability does NOT ensure X_n converges to X in values either.

Convergence in probability

- **Example 1.1**: Consider random variables on interval $(0, 1]$ (with Lebesgue measure as the probabilistic measure).
- Let $X(t) \equiv 0$ for all $t \in (0, 1]$, and

$$
\begin{aligned}
&X_1(t) = 1(0 < t \le 1/2], \qquad X_2(t) = 1(1/2 < t \le 1],\\
&X_3(t) = 2 \cdot 1(0 < t \le 1/2^2], \qquad X_4(t) = 2 \cdot 1(1/2^2 < t \le 2/2^2],\\
&X_5(t) = 2 \cdot 1(2/2^2 < t \le 3/2^2], \qquad X_6(t) = 2 \cdot 1(3/2^2 < t \le 1],\\
&X_7(t) = 2^2 \cdot 1(0 < t \le 1/2^3], \qquad X_8(t) = 2^2 \cdot 1(1/2^3 < t \le 2/2^3],\\
&X_9(t) = 2^2 \cdot 1(2/2^3 < t \le 3/2^3], \qquad \cdots
\end{aligned}
$$

- Apparently, $X_n \xrightarrow{\text{P}} X$ because for any $0 < \varepsilon < 1$, $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(X_n > 0) = 2^{-k}$ for some k, and $k \uparrow \infty$.
- However, $X_n \not\to X$ in values because for any $0 < t_0 \le 1$, there are infinite m's and n's such that $X_m(t_0) = 0$ and $X_n(t_0) > 0$.$\square$

1.2.3. Almost sure convergence

- If

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) = 1, \tag{1.3}$$

  we say that $\{X_n\}$ **converge to X almost surely**, and denote this as "$\lim X_n = X$ a.s.", or $X_n \xrightarrow{\text{a.s.}} X$.

- **Fatou's Lemma**: Almost sure convergence implies convergence in probability, and hence implies convergence in distribution.
- <u>The converses are not true</u>.

1.2.4. Convergence in mean

- Recall that the *distance* $|X_n - X|$ converges to zero in probability if $X_n \xrightarrow{\mathrm{P}} X$. Another way to define convergence in terms of distances is considering the expected distance.

- If

$$\lim_{n\to\infty} \mathbb{E}\big(|X_n - X|^r\big) = 0 \tag{1.4}$$

  for some $r \geq 1$, we say that $\{X_n\}$ converges **in the r-th mean** or **in the L^r norm** to X, and denote this as $X_n \xrightarrow{L^r} X$.

- The most common choice is $r = 2$, in which case it is also called the L^2 **convergence** or **mean-square convergence**.

Convergence in mean

- **Theorem 1.2**: Let $1 \le r \le s$. If $X_n \xrightarrow{L^s} X$, then $X_n \xrightarrow{L^r} X$.
- **Theorem 1.3**: $X_n \xrightarrow{L^r} X$ implies $X_n \xrightarrow{\mathrm{P}} X$. The converse is not true. (Cf. Example 1.1.)
- **Theorem 1.4**: $X_n \xrightarrow{L^r} X$ implies $\mathbb{E}(|X_n|^r) \to \mathbb{E}(|X|^r)$. The converse is not true. (Cf. the "obvious example".)
- A frequently made mistake is: treating $\mathbb{E}(X_n) \to \mathbb{E}(X)$ as the definition of convergence in mean or L^1 convergence.

1.2.5. Law of large numbers (LLN)

- **Weak Law of Large Numbers**: Let $\{X_i : i \geq 1\}$ be a sequence of i.i.d. random variables with finite mean μ, and $\overline{X}_n \widehat{=} \frac{1}{n}\sum_{i=1}^{n} X_i$. Then,

$$\overline{X}_n \xrightarrow{\text{P}} \mu, \qquad \text{as } n \to \infty. \tag{1.5}$$

- **Strong Law of Large Numbers**: Let $\{X_i : i \geq 1\}$ be a sequence of i.i.d. random variables with finite mean μ, and $\overline{X}_n \widehat{=} \frac{1}{n}\sum_{i=1}^{n} X_i$. Then,

$$\overline{X}_n \xrightarrow{\text{a.s.}} \mu, \qquad \text{as } n \to \infty. \tag{1.6}$$

- **Remark**: $\{X_i\}$ being i.i.d. with finite mean is the sufficient, but not necessary, condition for the convergence of $\overline{X}_n$.

1.2.6. Central limit theorem

- There are two frequently used versions of central limit theorems (CLT), one for independent and identically distributed (i.i.d.) sequence, and another for independent sequence.

- **Lindeberg-Lévy CLT**: Suppose $\{X_i : i \geq 1\}$ is a sequence of *i.i.d.* random variables with $\mathbb{E}(X_i) = \mu < \infty$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then, as n approaches infinity, the random variables $\sqrt{n}(\overline{X}_n - \mu)$ converge in distribution to a normal $N(0, \sigma^2)$:

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{\mathrm{d}} N(0, \sigma^2). \tag{1.7}$$

- The convergence in Eq. (1.7) is sometimes rewritten as

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathrm{d}} N(0, 1).$$

Central limit theorem

- **Lyapunov CLT**: Suppose $\{X_i : i \geq 1\}$ is a sequence of *independent* random variables, each with finite mean μ_i and variance σ_i^2. Define

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2, \qquad n \geq 1.$$

If for some $\delta > 0$, *Lyapunov's condition*

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^{n} \mathbb{E}\left[|X_i - \mu_i|^{2+\delta}\right] = 0$$

is satisfied, then, as n approaches infinity, the *standardized sum* of $(\overline{X}_i - \mu_i)$ converge in distribution to standard normal:

$$\frac{1}{s_n} \sum_{i=1}^{n} (X_i - \mu_i) \xrightarrow{\mathrm{d}} N(0,1). \tag{1.8}$$

§1.3 Estimation

- **Estimation** is the process of finding an **estimate** or approximation of a character which is a value (generally fixed/not random) that is usable for some purpose. For example,
    - (estimate) the rate of people aged over 65 in HK in (by the end of) year 2023,
    - (estimate) the quantitative relationship between salary and graduating GPA (maybe more) of MDASC graduates.
- A **prediction** or **forecast** is a *statement* (value, accuracy, etc.) about an event (usually random) in the future or under certain (new) situations/conditions. For example,
    - (predict) the rate of people aged over 65 in HK in year 2024,
    - (predict) the salary of a MDASC graduate with graduating GPA 3.7 (a specific condition).

Estimation and prediction

- Estimation is often done by *sampling*, which is counting a small number of *representatives*, and projecting that number onto a larger *population*.
  - For the rate of elders, we may calculate the rate among a group of representative HK residences, and then claim that the rate for all people in HK is *around the estimated one*.
  - For the salary, we may postulate a quantitative model, collect data from some MDASC graduates, and the estimate (calculate/count) the model using certain statistical methods (implementations of statistical ideas).
- Prediction is usually done upon estimation – predict the event based on certain estimated results/models.

Population and Sample

- In statistics, **population** is a set of random items or events which is of interest for some question or experiment, denoted as a random variable/vector X.

- A *statistical model* $\mathcal{M}$ is usually hypothesized for a population X, denoted as $X \sim \mathcal{M}$.

- A **sample** is a set of *representatives* selected (or collected) from a statistical population $X \sim \mathcal{M}$ by a defined procedure, denoted as $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ or $\{X_i\}_{i=1}^n$. Each *individual* X_i follows the same model $\mathcal{M}$.

- If individuals in a sample $\boldsymbol{X}$ are i.i.d., then we call $\boldsymbol{X}$ a **simple random sample**.

- The numerical values for a sample $\boldsymbol{X}$ are called *observations* or *realizations*, denoted as $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ or $\{x_i\}_{i=1}^n$.

Estimate, estimator and estimated value

- Let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a sample from a population $X \sim \mathcal{M}$, and $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ be observations. An **estimator**, or **estimate**, is a certain function (needn't in explicit functional form) of the sample $\boldsymbol{X}$, without any unknowns.
  - For example, the rate of elder people r can be estimated by

$$\widehat{r} = \frac{\text{number of people aged above 65 in the sample}}{\text{number of people in the sample}}.$$

    Here, the proportion function is an estimator.

- A general estimator is usually denoted as $T = T(\boldsymbol{X})$.
- The **estimated value** of an estimator $T = T(\boldsymbol{X})$ is its numerical value evaluated at $\boldsymbol{X} = \boldsymbol{x}$, denoted as $T(\boldsymbol{x})$.
  - The estimated value of the rate can be 30% for one sample, and 33% for another sample.

Properties of Estimators

- Let $\boldsymbol{X}$ be a sample of population X, and $\theta \in \Omega$ be a quantity of the population to be estimated.
- An estimator $\widehat{\theta} = T(\boldsymbol{X})$ of θ is said to be **unbiased** if

$$\mathbb{E}_\theta(\widehat{\theta}) = \theta, \qquad \underline{\text{for all } \theta \in \Omega}. \tag{1.9}$$

- The **bias** of an estimator $\widehat{\theta} = T(\boldsymbol{X})$ is defined as

$$\text{bias}(\widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta}) - \theta, \tag{1.10}$$

which is a function of θ (depends on the true value of θ).

- The **mean squared error** (**MSE**) of an estimator $\widehat{\theta} = T(\boldsymbol{X})$ is defined as

$$\text{MSE}_\theta(\widehat{\theta}) = \mathbb{E}_\theta(\widehat{\theta} - \theta)^2. \tag{1.11}$$

Properties of Estimators

- (*Some math*) By definitions (1.10) and (1.11), we have the following decomposition of MSE (we drop the sub-fix θ for simplicity):

$$
\begin{aligned}
\mathrm{MSE}(\widehat{\theta}) &= \mathbb{E}\big[\widehat{\theta} - \mathbb{E}(\widehat{\theta}) + \mathbb{E}(\widehat{\theta}) - \theta\big]^2 \\
&= \mathbb{E}\big[\widehat{\theta} - \mathbb{E}(\widehat{\theta})\big]^2 + \mathbb{E}\big[\mathbb{E}(\widehat{\theta}) - \theta\big]^2 \\
&\quad +2\mathbb{E}\Big[\big(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\big)\big(\mathbb{E}(\widehat{\theta}) - \theta\big)\Big] \\
&= \mathrm{Var}(\widehat{\theta}) + \big[\mathrm{bias}(\widehat{\theta})\big]^2.
\end{aligned}
\tag{1.12}
$$

The last equations holds because

$$
\begin{aligned}
&\mathbb{E}\Big[\big(\widehat{\theta} - \mathbb{E}(\widehat{\theta})\big)\big(\mathbb{E}(\widehat{\theta}) - \theta\big)\Big] \\
&= \mathbb{E}\Big[\widehat{\theta}\cdot\mathbb{E}(\widehat{\theta}) - \widehat{\theta}\cdot\theta - \big[\mathbb{E}(\widehat{\theta})\big]^2 + \mathbb{E}(\widehat{\theta})\cdot\theta\Big] \\
&= \big[\mathbb{E}(\widehat{\theta})\big]^2 - \mathbb{E}(\widehat{\theta})\cdot\theta - \big[\mathbb{E}(\widehat{\theta})\big]^2 + \mathbb{E}(\widehat{\theta})\cdot\theta \\
&= 0.
\end{aligned}
$$

Properties of Estimators

- Let $\widehat{\theta}_i = T_i(\boldsymbol{X})$, $i = 1, 2$, be two estimators of θ *based on the same sample* $\boldsymbol{X}$. If

$$\mathrm{MSE}_\theta(\widehat{\theta}_1) \le \mathrm{MSE}_\theta(\widehat{\theta}_2) \qquad \underline{\text{for all } \theta \in \Omega},$$

then, we say that $\widehat{\theta}_1$ is *uniformly* better than $\widehat{\theta}_2$.

- Let $\widehat{\theta}_i = T_i(\boldsymbol{X})$, $i = 1, 2$, be two *unbiased* estimators of θ *based on the same sample* $\boldsymbol{X}$. If

$$\mathrm{Var}_\theta(\widehat{\theta}_1) \le \mathrm{Var}_\theta(\widehat{\theta}_2) \qquad \underline{\text{for all } \theta \in \Omega},$$

then, we say that $\widehat{\theta}_1$ is (uniformly) **more efficient** than $\widehat{\theta}_2$.

- The most efficient estimator (if exist), which is the unbiased estimator with the minimal variance, is called the **MVUE**.

Large-sample Properties of Estimators

- Let $\boldsymbol{X}_n = \{X_1, \cdots, X_n\}$ be a sample from a population X, and $\widehat{\theta}_n = T(\boldsymbol{X}_n)$ be an estimator of an unknown population character $\theta \in \Omega$. The suffix n is added to emphasize that it depends on the *sample size n*.

- If, as n increases and tends to infinity,

$$\lim_{n\to\infty} \mathbb{E}_\theta(\widehat{\theta}_n) = \theta, \qquad \underline{\text{for all } \theta \in \Omega}, \tag{1.13}$$

  then we say that $\widehat{\theta}_n$ (more correctly, $\widehat{\theta} = T(\cdot)$) is **asymptotically unbiased**.

- $\widehat{\theta}_n$ (or, $\widehat{\theta} = T(\cdot)$) is said to be a **consistent** estimator of θ, if

$$\operatorname*{plim}_{n\to\infty} \widehat{\theta}_n = \theta, \qquad \underline{\text{for all } \theta \in \Omega}, \tag{1.14}$$

  where "plim" stands for *converges in probability*.

Credible Interval

- Let $\boldsymbol{X}$ be a sample from a population $X \sim \mathcal{M}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Omega$.
- A random and *convex* subset (a region) Ω, $C(\boldsymbol{X}) \subset \Omega$, is called a **credible/confidence region** at the **confidence level** $1-\alpha$ $(0 < \alpha < 1)$ if

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{\theta} \in C(\boldsymbol{X})) = 1-\alpha, \qquad \underline{\text{for all } \boldsymbol{\theta} \in \Omega}. \tag{1.15}$$

- If $\boldsymbol{\theta} = \theta$ is a scalar character, and $C(\boldsymbol{X})$ has the form of an interval $(\widehat{\theta}_1,\ \widehat{\theta}_2)$, we call it a **credible/confidence interval** (**CI**). Moreover, we call
  - $\widehat{\theta}_1$ the *lower credible/confidence limit* (*lcl*) or *lower credible/confidence bound*, and
  - $\widehat{\theta}_2$ the *upper credible/confidence limit* (*ucl*).

Credible Interval

- **Example 1.2**: Let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a random sample from a (unidimensional) population X. Define

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

- Assume $X \sim N(\mu, \sigma^2)$, where μ is unknown but σ^2 is known.
- Since $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ is standard normal, for any $\alpha_1 > 0$ and $\alpha_2 > 0$ with $\alpha_1 + \alpha_2 = \alpha < 1$,

$$\mathbb{P}(-Z_{\alpha_1} < Z < Z_{\alpha_2}) = 1 - \alpha_1 - \alpha_2 = 1 - \alpha,$$

  where Z_α is the upper α-quantile of the standard normal distribution.

Credible Interval

- Mathematically,

$$
\begin{aligned}
& -Z_{\alpha_1} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha_2} \\
\Leftrightarrow \qquad & \overline{X} - Z_{\alpha_2}\sigma/\sqrt{n} < \mu < \overline{X} + Z_{\alpha_1}\sigma/\sqrt{n}.
\end{aligned}
$$

- Therefore, the following intervals are possible CIs of μ at the $1 - \alpha$ confidence level,

$$
(\widehat{\mu}_1, \widehat{\mu}_2) = (\overline{X} - Z_{\alpha_2}\sigma/\sqrt{n}, \overline{X} + Z_{\alpha_1}\sigma/\sqrt{n}). \tag{1.16}
$$

- Among all the CIs in (1.16), the *optimal* one is that with $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$, in the sense that it has the smallest width (the most accurate/precise). $\square$

Pivotal quantity

- The crucial step in Example 1.2 is finding the quantity Z, whose value depends on both the sample and the characteristic of interest, θ, but whose distribution is (approximately) known. Such a quantity is called a **pivotal quantity** for θ.

- **Lemma 1.1**: Let X be a random variable with cumulated distribution function (cdf) $F(x)$. Define

$$U = -2\log F(X), \qquad V = -2\log[1 - F(x)]. \tag{1.17}$$

Both U and V have a $\chi^2(2)$ distribution.

- *Proof*. (For U only) Observe that for any $x > 0$,

$$\begin{aligned}\mathbb{P}(U \le x) &= \mathbb{P}[F(X) \ge \exp(-x/2)] = \mathbb{P}[X \ge F^{-1}(\exp(-x/2))] \\ &= 1 - F[F^{-1}(\exp(-x/2))] = 1 - \exp(-x/2).\end{aligned}$$

Hence, U has a cdf of a $\chi^2(2)$ distribution as required. $\square$

Pivotal quantity

- Lemma 1.1 has an immediate, and very important, application.
- Suppose we have a random sample $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ from a population $X \sim F(x;\theta)$. Define for each $1 \leq i \leq n$ that

$$U_i = -2\log[F(X_i;\theta)], \qquad V_i = -2\log[1 - F(X_i;\theta)].$$

Then, $\{U_i\} \overset{i.i.d.}{\sim} \chi^2(2)$, $\{V_i\} \overset{i.i.d.}{\sim} \chi^2(2)$. Hence,

$$Q_1(\boldsymbol{X};\theta) = \sum_{i=1}^{n} U_i \sim \chi^2(2n)$$

and

$$Q_2(\boldsymbol{X};\theta) = \sum_{i=1}^{n} V_i \sim \chi^2(2n)$$

are two pivotal quantities for θ.

- **Example 1.3**: Let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a random sample from an exponential population $X \sim \mathcal{E}(x;\lambda)$, that is, $F(x;\lambda) = 1 - e^{-\lambda x}$ for all $x \geq 0$. Hence,

$$Q_2(\boldsymbol{X};\lambda) = -2\sum_{i=1}^{n} \log[1 - F(X_i)] = 2n\lambda\overline{X} \sim \chi^2(2n).$$

At a $(1-\alpha)$ confidence level,

$$\mathbb{P}\{\chi^2_{1-\alpha/2}(2n) < Q_2(\boldsymbol{X};\lambda) < \chi^2_{\alpha/2}(2n)\} = 1 - \alpha.$$

Therefore,

$$\left(\frac{\chi^2_{1-\alpha/2}(2n)}{2n\overline{X}}, \frac{\chi^2_{\alpha/2}(2n)}{2n\overline{X}}\right)$$

is a $(1-\alpha)$-credible interval of λ. □

§1.4 Hypothesis Testing

- A statistical **hypothesis test** or **hypothesis testing** is a method of statistical inference (or the process) used to decide whether the data (sample) sufficiently support a particular statement (hypothesis) about the population.
- Let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a random sample from a population (a model) $X \sim \mathcal{M}(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Omega$.
- For illustration, suppose we are testing a hypothesis on the population parameter (character) $\boldsymbol{\theta}$.

Steps of Hypothesis Testing

(1) **Postulate a pair of hypotheses**, a *null hypothesis* $H_0: \boldsymbol{\theta} \in \Omega_0 \subset \Omega$, and an *alternative hypothesis* $H_a: \boldsymbol{\theta} \in \Omega_a \subset \Omega$, exclusive to Ω_0, i.e., $\Omega_0 \cap \Omega_a = \emptyset$.

(2) **Design a test statistic**: a *pivotal quantity* for $\boldsymbol{\theta}$, $T = T(\boldsymbol{X}; \boldsymbol{\theta})$, whose distribution is *conditionally* known when the null hypothesis H_0 is true or marginally true.

(3) **Formulating**: choosing a *significance level* (or, *level of significance*) α which is a small probability, and finding a *rejection region* R_α (a rule) such that

$$\mathbb{P}(T(\boldsymbol{X}; \boldsymbol{\theta}) \in R_\alpha \,|\, H_0 \text{ is true}) \leq \alpha. \tag{1.18}$$

When H_0 is true, $T(\boldsymbol{X}; \boldsymbol{\theta})$ *very unlikely* falls in R_α.

(4) **Draw conclusion** based on observations $\boldsymbol{x}$: reject H_0 at the α-th level when $T(\boldsymbol{x}; \boldsymbol{\theta}) \in R_\alpha$, not reject otherwise.

Hypothesis Testing

- Under the null hypothesis H_0 (marginally), the probability of $T(\boldsymbol{X};\boldsymbol{\theta})$ taking values *more extreme/unlike than*, or *equally extreme as* $T(\boldsymbol{x};\boldsymbol{\theta})$ is called the ***p*-value** of the test.
- The exact form of p-value depends on the form of rejection region R_α. Some commonly used definitions of p-value for a scalar test statistic $T(\boldsymbol{X};\boldsymbol{\theta})$, are summarized in the following table, where c_α is called the *critical value*.

Form of R_α	Definition of p-value
$\{T(\boldsymbol{x};\boldsymbol{\theta}) \geq c_\alpha\}$	$\mathbb{P}_{H_0}\{T(\boldsymbol{X};\boldsymbol{\theta}) \geq T(\boldsymbol{x};\boldsymbol{\theta})\}$
$\{T(\boldsymbol{x};\boldsymbol{\theta}) \leq c_\alpha\}$	$\mathbb{P}_{H_0}\{T(\boldsymbol{X};\boldsymbol{\theta}) \leq T(\boldsymbol{x};\boldsymbol{\theta})\}$
$\{\lvert T(\boldsymbol{x};\boldsymbol{\theta})\rvert \geq c_\alpha\}$	$\mathbb{P}_{H_0}\{\lvert T(\boldsymbol{X};\boldsymbol{\theta})\rvert \geq T(\boldsymbol{x};\boldsymbol{\theta})\}$

Table 1.1: Commonly used definitions of p-value for a scalar test statistic.

Hypothesis Testing

- Two types of error (probability):
  - **Type I error**: reject H_0 when it is true.
  - **Type II error**: not reject (accept) H_0 when it is false.
- When a specific $\boldsymbol{\theta}_a \in \Omega_a$ is true, the probability that we (correctly) reject H_0,

$$p(\boldsymbol{\theta}_a) = \mathbb{P}(T(\boldsymbol{X};\boldsymbol{\theta}) \in R_\alpha \,|\, \boldsymbol{\theta} = \boldsymbol{\theta}_a) \tag{1.19}$$

  is called the **power** (function) of the test.
- Keying "`??test`" in R, you may find a number of available built-in functions including the key word "test", in various libraries, for or related to different types of hypothesis testing.

- **Example 1.4**: Let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a random sample from a normal population $X \sim N(\mu, \sigma^2)$, where σ^2 is known while $\mu \in \mathbb{R}$ is unknown.

- Test hypotheses $H_0: \ \mu \geq \mu_0$ against $H_1: \ \mu < \mu_0$.

- Test statistic and its null distribution:

$$Z = T(\boldsymbol{X}) = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \text{when } \mu = \mu_0 \, (\in \Omega_0).$$

- Given a significance level α, the rejection region has a form of $R_\alpha = (-\infty, c)$ for some *critical value* c.

- Criteria: $\mathbb{P}(Z < c \,|\, \mu = \mu_0) = \alpha \Longrightarrow c = -Z_\alpha = Z_{1-\alpha}$.

- Suppose that $\sigma = 2$, $\mu_0 = 5$, $\alpha = 0.05$, and $\overline{x} = 4.8$ based on a set of observations $\boldsymbol{x} = \{x_1, \cdots, x_{100}\}$. We have

$$z = T(\boldsymbol{x}) = \frac{4.8 - 5}{2/\sqrt{100}} = -1 > -Z_{0.05} = -1.645.$$

  Therefore, at the 5% significance level, the null hypothesis H_0 can not be rejected *based on observations* $\boldsymbol{x}$.

- The observed test statistic z has a p-value

$$\mathbb{P}(Z < z \,|\, \mu = 5) = \Phi(-1) = 0.1587.$$

  where $\Phi(\cdot)$ is the cdf of the standard normal distribution. Since the p-value is greater than $\alpha = 0.05$, H_0 can not be rejected.

Hypothesis Testing

- For any $\Omega_a \ni \mu_a < \mu_0$, the power function of the test is

$$\begin{aligned} p(\mu_a) &= \mathbb{P}\left(\frac{\overline{X}-\mu_0}{\sigma/\sqrt{n}} < -1.645 \,\middle|\, \mu = \mu_a\right) \\ &= \mathbb{P}\left(\frac{\overline{X}-\mu_a}{\sigma/\sqrt{n}} + \frac{\mu_a-\mu_0}{\sigma/\sqrt{n}} < -1.645 \,\middle|\, \mu = \mu_a\right) \\ &= \Phi\left(-1.645 - \frac{\mu_a-\mu_0}{\sigma/\sqrt{n}}\right). \end{aligned}$$

- For example, when $\mu_a = 4.5$ and/or 4,

$$\begin{aligned} p(4.5) &= \Phi\left(-1.645 - \frac{4.5-5}{2/\sqrt{100}}\right) = \Phi(0.855) = 0.8037, \\ p(4) &= \Phi(3.355) = 0.9996. \end{aligned}$$

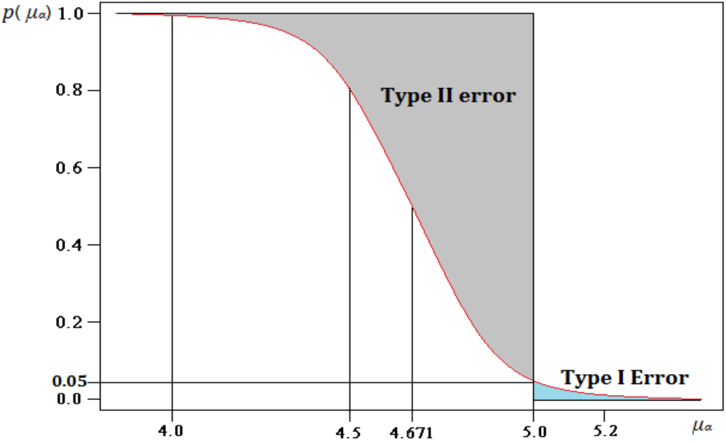

Figure 1.1: Power function of the test that a normal sample of size 100 with variance 4 has the mean value greater than 5. □

Duality between Estimation and Test

- There usually exists a kind of duality between estimation and hypothesis testing. We use the following example for illustration.

- **Example 1.5**: Let $\boldsymbol{X} = \{X_1, \cdots, X_n\}$ be a random sample from a normal population $X \sim N(\mu, \sigma^2)$, where σ^2 is known while $\mu \in \mathbb{R}$ is unknown.

- Consider the following credible intervals at the $(1-\alpha)$ level of confidence:

Type	Credible interval
Both bounds	$(\overline{x} - Z_{\alpha/2}\sigma/\sqrt{n}, \overline{x} + Z_{\alpha/2}\sigma/\sqrt{n})$
Upper bound only	$(-\infty, \overline{x} + Z_{\alpha}\sigma/\sqrt{n})$
Lower bound only	$(\overline{x} - Z_{\alpha}\sigma/\sqrt{n}, +\infty)$

Table 1.2

Duality between Estimation and Test

- For hypothesis testing $H_0: \ \mu = \mu_0$ at the significance level α.

H_a	Not reject H_0 if (non-rejection region)
$\mu \neq \mu_0$	$\overline{x} \in (\mu_0 - Z_{\alpha/2}\sigma/\sqrt{n},\ \mu_0 + Z_{\alpha/2}\sigma/\sqrt{n})$
$\mu < \mu_0$	$\overline{x} \in (\mu_0 - Z_{\alpha}\sigma/\sqrt{n},\ +\infty)$
$\mu > \mu_0$	$\overline{x} \in (-\infty,\ \mu_0 + Z_{\alpha}\sigma/\sqrt{n})$

Table 1.3

- This can be rewritten into the following.

H_a	Not reject H_0 if (non-rejection region)
$\mu \neq \mu_0$	$\mu_0 \in (\overline{x} - Z_{\alpha/2}\sigma/\sqrt{n},\ \overline{x} + Z_{\alpha/2}\sigma/\sqrt{n})$
$\mu < \mu_0$	$\mu_0 \in (-\infty,\ \overline{x} + Z_{\alpha}\sigma/\sqrt{n})$
$\mu > \mu_0$	$\mu_0 \in (\overline{x} - Z_{\alpha}\sigma/\sqrt{n},\ +\infty)$

Table 1.4

Duality between Estimation and Test

- A kind of duality can be easily seen among these three tables, especially between tables 1.2 and 1.4.
- We can easily draw conclusion for a hypothesis testing based on the corresponding credible interval.
  - For example, we had the 95% confidence interval of μ as $(5.14, 5.92)$, we are not able to reject $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$ at the 5% significance level if $\mu_0 = 5.5$ since $5.5 \in ((5.14, 5.92)$. However, we shall reject H_0 if $\mu_0 = 6$ because $6 \notin ((5.14, 5.92)$.
- On the other hand, we may also able to create the credible interval based on corresponding tests. For example, <u>the right-sided (lower bound only) credible interval is the collection of all μ_0's such that $H_0 : \mu \leq \mu_0$ is NOT rejected in favor of $H_a : \mu > \mu_0$</u>. $\square$

Interpretations of the duality

- The duality can be also interpreted or understood as following:
  - A both bounds credible interval of some character θ means θ, with a high probability, is neither too large nor too small. Consequently, when θ_0 falls in the credible interval, $H_0 : \theta = \theta_0$ cannot be rejected vs $\underline{H_a : \theta \neq \theta_0}$, which stands for "$\underline{\theta \text{ is significantly larger or smaller than } \theta_0}$".
  - Similarly, a lower bound only credible interval means θ is not too small. Consequently, $H_0 : \theta = \theta_0$ cannot be rejected vs $\underline{H_a : \theta < \theta_0}$ if θ_0 falls in the credible interval.
  - An upper bound only credible interval means θ is not too large. Consequently, $H_0 : \theta = \theta_0$ cannot be rejected vs $\underline{H_a : \theta > \theta_0}$ if θ_0 falls in the credible interval.

§1.5 Frequentism in Practice

- Consider some characteristic $\theta = \theta(X)$ of a population $X \sim F(x)$. Suppose that, based on a sample $\boldsymbol{X}$ and its realization $\boldsymbol{x}$, we have an estimator $\widehat{\Theta} = T(\boldsymbol{X})$ (regarded as a rule/an algorithm), and an (observed) estimate $\widehat{\theta} = T(\boldsymbol{x})$ (a realization).
- **Frequentist inference** or **frequentism**: the accuracy of $\widehat{\Theta}$ (or $\widehat{\theta}$) is defined as the *probabilistic accuracy* of $\widehat{\Theta}$ as a random estimator of θ.
- The randomness of $\widehat{\Theta}$ can be understood as "*an infinite sequence of future trails/samples* $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \cdots$".
- *Bias* and *standard error* are familiar examples of frequentism:

$$\begin{aligned} \operatorname{bias}(\widehat{\Theta}) &= \mathbb{E}_F(\widehat{\Theta}) - \theta = \mu_\Theta - \theta, \\ \operatorname{se}(\widehat{\Theta}) &= \operatorname{sd}(\widehat{\Theta}) = \sqrt{\mathbb{E}_F(\widehat{\Theta} - \mu_\Theta)^2}. \end{aligned}$$

Frequentism in Practice

- Frequentism needs the distribution of the statistics $\widehat{\Theta}$, $F_{\widehat{\Theta}}$. Or, at least, estimates of the bias/mean and the standard error.
- Nevertheless, some practical frequentism principles are usually used when $F_{\widehat{\theta}}$ is unknown, or (estimates of) the bias/mean and standard error, are unavailable (not clear, not trivial).

1. **The plug-in principle**: estimate bias($\widehat{\Theta}$) and/or se($\widehat{\Theta}$) by plugging known estimates, e.g., sample mean $\overline{X}$ for population mean μ and/or sample variance S^2 for population variance σ^2, if μ and/or σ^2 appear in the formula(e) of bias($\widehat{\Theta}$) and/or se($\widehat{\Theta}$).

- **Example 1.6**: Consider the sample variance S^2 of a normal sample $\boldsymbol{X} = \{X_i : 1 \le i \le n\} \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. Suppose we want to estimate the standard error of S^2, se(S^2).

Frequentism in Practice

- It is well known that $\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, with mean $\mathbb{E}(\chi^2) = n-1$, and variance $\text{Var}(\chi^2) = 2(n-1)$.

- Since $S^2 = \frac{\sigma^2}{n-1}\chi^2$, we have

$$\begin{aligned}
\mathbb{E}(S^2) &= \mathbb{E}(\chi^2) \cdot \frac{\sigma^2}{n-1} = \sigma^2, \qquad \text{(unbiased)} \\
\text{Var}(S^2) &= \text{Var}(\chi^2) \cdot \left(\frac{\sigma^2}{n-1}\right)^2 = \frac{2\sigma^4}{n-1}, \qquad \text{and} \\
\text{se}(S^2) &= \sqrt{\frac{2}{n-1}}\sigma^2.
\end{aligned}$$

- Replacing σ^2 with S^2 in the last equation gives an estimate of the standard error, that is,

$$\widehat{\text{se}}(S^2) = \sqrt{\frac{2}{n-1}}S^2. \tag{1.20}$$

$\square$

Frequentism in Practice

2. **Taylor approximations**. Suppose that $\widehat{\eta} = f(\widehat{\theta})$ is a known function of an estimator $\widehat{\theta}$, for which we are able to do statistical inferences. Then the (usually linear) Taylor approximation $d\widehat{\eta} \approx f'(\widehat{\theta})(d\widehat{\theta})$ provides us a way to do inference(s) for $\widehat{\eta}$, especially estimating the standard error, thinking of $f'(\widehat{\theta})$ as a constant.
   - This method (linear approximation) is sometimes referred to as the "delta-method" or "delta-approximation".

- **Example 1.7**: Suppose we want to estimate the standard error of the sample standard deviation S (not S^2) of the normal sample $\boldsymbol{X}$ in Example 1.5.
- The exact distribution of S could be very complicated, or even *unknown*.

Frequentism in Practice

- Making use of the Taylor approximation for the square root function: $\Delta\sqrt{x} \approx \frac{1}{2\sqrt{x}}\Delta x$, we have

$$\Delta S \approx \frac{1}{2\sqrt{S^2}}\Delta S^2 = \frac{\Delta S^2}{2S},$$
$$\mathrm{se}(S) \approx \frac{\mathrm{se}(S^2)}{2S}.$$

- Plugging the estimated standard error of S^2 in Eq. (1.20) gives the following approximated estimated standard error of S:

$$\widehat{\mathrm{se}}(S) \approx \frac{\widehat{\mathrm{se}}(S^2)}{2S} = \sqrt{\frac{1}{2(n-1)}}S. \tag{1.21}$$

$\square$

Frequentism in Practice

3. **Parametric families and maximum likelihood theory**. Theoretical expressions for the standard error of a MLE. Will be introduced in Chapter 4.

4. **Simulation and the bootstrap**. Simulate $\boldsymbol{X}^{(b)}$ and $T(\boldsymbol{X}^{(b)})$, $1 \leq b \leq B$, to get the e.s.e. $\widehat{\mathrm{se}}(\widehat{\theta})$. Will be introduced in Chapters 6 and 7.

5. **Pivotal statistics**. Distribution of $\widehat{\Theta}$ does not depend on the underlying population distribution F (e.g., difference between pairwise data), and the theoretical distribution of $\widehat{\Theta}$ applies exactly to $\widehat{\theta}$.