

HKU STAT3902 Study Notes

Chiu Ka Long (Leo)*

Last Updated: 2024-12-29

This work is licensed under a [Creative Commons “Attribution 4.0 International”](#) license.



Contents

1	Empirical Probability Distributions	2
1.1	Sample of Observations	2
1.2	Empirical Cumulative Distribution Function	2
2	Convergence	3
2.1	Modes of Convergence	3
2.2	Some Theorems for Convergence	4
3	Point Estimation	5
3.1	Unbiasedness	6
3.2	Efficiency	6
3.3	Consistency	8
3.4	Sufficiency and Likelihood	8
3.5	Completeness	9
3.6	Cramer-Rao Inequality	9
3.7	Maximum Likelihood Estimation and Method of Moments	10
3.8	Asymptotic Properties of MLE	10
4	Interval Estimation	11
4.1	Confidence Intervals	11
4.2	Constructing Confidence Intervals	11
5	Hypothesis Testing	13
5.1	Hypothesis Tests	13
5.2	Quality of a Hypothesis Test	14
5.3	p -value	15
5.4	Uniformly Most Powerful Test	17
5.5	Likelihood Ratio Test	17
5.6	Tests Based on Asymptotic Theory	18
5.7	Goodness of Fit Test	19
5.8	Test of Independence	20

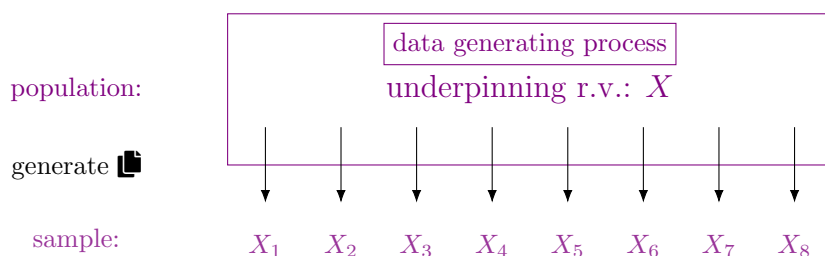
*email ✉: leockl@connect.hku.hk; personal website 🌐: <https://leochiukl.github.io>

1 Empirical Probability Distributions

- 1.0.1 After studying *probability theory* in STAT2901, we will focus on *statistical inference* in STAT3902: *inference based on observations*. In this section we will first focus on the *observations* to see how they can be “explored”.

1.1 Sample of Observations

- 1.1.1 Much of statistical inference is based on the idea of a *sample*. Consider a “population random variable” X with a cdf F . Then, a (random) **sample** of n (unrealized) observations from the population is a vector of n *random variables*: (X_1, \dots, X_n) where X_1, \dots, X_n are independent and identically distributed with the common cdf F , denoted by $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} F$.



[Intuition 💡: Population can be understood as a *data generating process*: Each observation in the sample is a (random) datum.]

[Note: The **sample size** of the sample (X_1, \dots, X_n) is its number of entries: n .]

- 1.1.2 A sample is *unrealized* in the sense that it contains random variables but not the actual observed *values*. Depending on the outcome $\omega \in \Omega$, the actual observed values can differ in multiple samplings (collections of samples).

In practice, we can only “see” 👁 a *realized* sample. However, for theoretical purpose, we often work with *unrealized* samples (consisting of random variables).

1.2 Empirical Cumulative Distribution Function

- 1.2.1 Given a sample (X_1, \dots, X_n) , the **empirical cumulative distribution function** (cdf) is

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\{X_i \leq x\}}.$$

Remarks:

- In statistical inference, notation with a hat $\hat{}$ usually *estimates/approximates* something.
- In words, it means

$$\hat{F}_n(x) = \frac{\text{no. of observations} \leq x}{\text{no. of observations}}.$$

- 1.2.2 Intuitively, the empirical cdf should be somewhat “close” to the actual cdf, especially if the number of observations n is large. This is justified by the following result:

Theorem 1.2.a (Glivenko-Cantelli theorem). Suppose that we have *infinitely many* i.i.d. (unrealized) observations X_1, X_2, \dots to be included in a sample. For any $n \in \mathbb{N}$, the vector (X_1, \dots, X_n) constitutes a sample (of n observations), and we denote the corresponding empirical cdf by \hat{F}_n . Then, the sequence of empirical cdfs $\{\hat{F}_n\}$ *converges uniformly* to the actual cdf F almost surely (i.e., with probability 1).

[Note: Roughly speaking, this suggests that $\hat{F}_n(x)$ is always within a very small region around $F(x)$ (“very close”) for any $x \in \mathbb{R}$ (with probability 1) when n is very large.]

2 Convergence

2.0.1 In statistical inference, sometimes we are interested in the *asymptotic* behaviour of the observations in the sample when the sample size $n \rightarrow \infty$. Thus, we are interested in studying different *modes* in which a sequence of random variables (or observations in the context of statistical inference) $\{X_n\}$ “converges”.

2.1 Modes of Convergence

2.1.1 Consider a sequence of r.v.’s $\{X_n\}$. Here we introduce *four modes* in which the sequence can converge:

- convergence in probability
- convergence in distribution
- almost sure convergence
- mean square convergence

2.1.2 The sequence $\{X_n\}$ **converges in probability** to a r.v. X (denoted by $\{X_n\} \xrightarrow{p} X$) if for any $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0.$$

[Note: Roughly, this means when n is large, it is of *very high probability* that X_n is *very close to* X (their difference does not exceed a specified constant ε).]

2.1.3 The sequence $\{X_n\}$ **converges in distribution** to a r.v. X (denoted by $\{X_n\} \xrightarrow{d} X$) if

$$\lim_{n \rightarrow \infty} \hat{F}_n(x) = F(x)$$

for any x at which the cdf F is continuous.

[Note: Roughly, this means when n is large, the empirical cdf $\hat{F}_n(x)$ is very close to $F(x)$ (for any x at which F is continuous).]

2.1.4 The sequence $\{X_n\}$ **converges almost surely** to a r.v. X (denoted by $\{X_n\} \xrightarrow{\text{a.s.}} X$) if

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} X_n = X\right) = \mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

[Note: Roughly, this means we are *very sure* that the sequence $\{X_n\}$ converges (“exactly”, not just “in probability”) to X .]

2.1.5 The sequence $\{X_n\}$ **converges in mean square** to a r.v. X (denoted by $\{X_n\} \xrightarrow{\text{m.s.}} X$) if

$$\lim_{n \rightarrow \infty} \mathbb{E}[|X_n - X|^2] = 0.$$

[Note: Roughly, this means when n is large, the mean squared distance between X_n and X is *very close to* zero.]

2.1.6 We have the following relationship between different modes of convergences:

$$\begin{array}{c} \{X_n\} \xrightarrow{\text{a.s.}} X \\ \searrow \\ \{X_n\} \xrightarrow{p} X \implies \{X_n\} \xrightarrow{d} X \\ \nearrow \\ \{X_n\} \xrightarrow{\text{m.s.}} X \end{array}$$

2.2 Some Theorems for Convergence

2.2.1 Here we will introduce several theorems related to different modes of convergence:

- Slutsky's theorem
- continuous mapping theorem
- weak law of large numbers
- strong law of large numbers
- central limit theorem (famous!)

2.2.2 Slutsky's theorem suggests arithmetic properties of sequences convergent in some modes.

Theorem 2.2.a (Slutsky's theorem). Let $\{X_n\}$ and $\{Y_n\}$ be two sequences of random variables. Suppose $\{X_n\} \xrightarrow{d} X$ and $\{Y_n\} \xrightarrow{p} c$ where X is a random variable and c is a constant. Then,

- (a) $\{X_n + Y_n\} \xrightarrow{d} X + c$
- (b) $\{X_n - Y_n\} \xrightarrow{d} X - c$
- (c) $\{X_n Y_n\} \xrightarrow{d} Xc$
- (d) $\{X_n/Y_n\} \xrightarrow{d} X/c$.

Also, the result still holds when *every* " \xrightarrow{d} " gets replaced by " \xrightarrow{p} ".

[Warning: Note that in the condition we have $\{Y_n\} \xrightarrow{p} c$ where c is a *constant*. When the constant gets replaced by a random variable, the result may not hold anymore.]

2.2.3 Continuous mapping theorem suggests that applying continuous function preserves convergence for some modes.

Theorem 2.2.b (Continuous mapping theorem). Let $\{X_n\}$ be a sequence of random variables and X be a random variable. Then, for any continuous function g ,

$$\{X_n\} \xrightarrow{*} X \implies \{g(X_n)\} \xrightarrow{*} g(X)$$

where " $*$ " is " d ", " p ", or " $a.s.$ " (the modes of convergence at LHS and RHS should be the same).

2.2.4 Weak and strong laws of large numbers (LLN) concern the asymptotic behaviour of a sequence of *sample means*: It connects *sample* mean and *theoretical* mean (mathematical expectation).

Theorem 2.2.c (Weak and strong laws of large numbers). Let $\{X_n\}$ be a sequence of i.i.d. random variables with *finite mean* μ . Define the **sample mean** \bar{X}_n by $\frac{1}{n} \sum_{i=1}^n X_i$, for any $n \in \mathbb{N}$. Then,

- (a) (weak LLN) $\{\bar{X}_n\} \xrightarrow{p} \mu$;
- (b) (strong LLN) $\{\bar{X}_n\} \xrightarrow{a.s.} \mu$.

[Note: Since almost sure convergence implies convergence in probability, strong LLN implies weak LLN, as one may expect. (But one can prove weak LLN without appealing to strong LLN.)]

2.2.5 The famous *central limit theorem* suggests that sample mean is approximately normally distributed when n is large under certain conditions \rightarrow *normal distribution* appears frequently in statistical inference.

Theorem 2.2.d (Central limit theorem). Let $\{X_n\}$ be a sequence of i.i.d. random variables with *finite mean* μ and *finite variance* σ^2 . Consider the sample mean $\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ for any $n \in \mathbb{N}$. Then,

$$\left\{ \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \right\} \xrightarrow{d} Z$$

where $Z \sim N(0, 1)$ is a standard normal random variable.

3 Point Estimation

3.0.1 Starting from here, we will discuss three major kinds of statistical inference:

- point estimation (section 3)
- interval estimation (section 4)
- hypothesis testing (section 5)

3.0.2 The process of **point estimation** is described as follows:

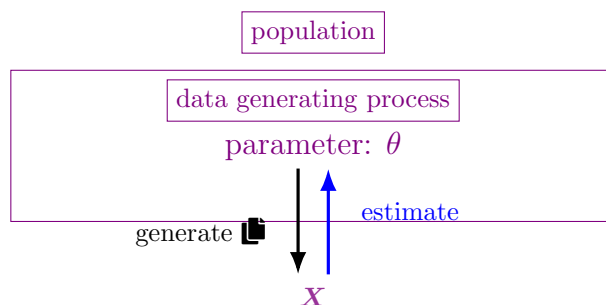
- Assume that the population (data generating process) is “controlled” by an unknown “population parameter” θ (i.e., the distribution of the underpinning r.v. is parameterized by θ).

[Note: In this notes we shall assume that θ is *deterministic*. Statistical inference under this assumption is known as **frequentist inference**. For statistical inference that assumes θ is an unknown *random variable*, it is called **Bayesian inference** (which will not be discussed here).]

- Given a sample $\mathbf{X} = (X_1, \dots, X_n)$ collected from the population, we then estimate the true parameter θ using an **estimator** (random variable) based on the sample \mathbf{X} , denoted by $\hat{\theta}$ (as a function of \mathbf{X}).

Remarks:

- More generally, we can estimate a *function* of true parameter θ : $\psi(\theta)$ by an estimator $T = T(\mathbf{X})$. (The upcoming terminologies/results apply analogously to this general case.)
- In general, any function of sample \mathbf{X} is called a **statistic**. So, $\hat{\theta}$ and T here are both statistics.



3.0.3 Here, we will discuss two common approaches to perform point estimation:

- maximum likelihood estimation (MLE)
- method of moments

But before that, let us first investigate what qualifies as a “good” estimator.

3.0.4 One major goal of estimation in statistics is to find an estimator with “high quality”. But very often we can analyze an estimator in different “angles”, and different estimators may be “better” in different aspects.

In the following, we will discuss multiple aspects where we can evaluate an estimator:

- unbiasedness
- efficiency
- consistency
- sufficiency
- completeness

Based on some of these aspects, we can characterize “optimal” estimator in certain sense.

3.1 Unbiasedness

3.1.1 Unbiasedness is related to the “average accuracy” of an estimator.

3.1.2 We first introduce the concept of *bias*: The **bias** of an estimator $\hat{\theta}$ (of θ) is

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

3.1.3 The estimator $\hat{\theta}$ is **unbiased** if $\text{Bias}(\hat{\theta}) = 0$, or $\mathbb{E}[\hat{\theta}] = \theta$.

[Note: Roughly speaking, for an *unbiased* estimator, its “average” value coincides with the true parameter $\theta \rightarrow$ desirable.]

3.1.4 As mentioned in [2.0.1], sometimes we are interested in the *asymptotic* case. So, even if an estimator is biased, we are interested in knowing whether it is unbiased *asymptotically*.

In the asymptotic case, we shall assume that there are infinitely many i.i.d. observations X_1, X_2, \dots to be included in a sample. Here, to be more explicit about the dependence of the estimator on the sample (X_1, \dots, X_n) , we express the estimator as $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$, for any $n \in \mathbb{N}$.

Then, the estimator $\hat{\theta}$ is **asymptotically unbiased** if

$$\lim_{n \rightarrow \infty} \text{Bias}(\hat{\theta}(X_1, \dots, X_n)) = 0,$$

or

$$\lim_{n \rightarrow \infty} \mathbb{E}[\hat{\theta}(X_1, \dots, X_n)] = \theta.$$

3.2 Efficiency

3.2.1 To motivate the notion of *efficiency*, we first consider the following result:

Proposition 3.2.a (Bias-variance decomposition). Given an estimator $\hat{\theta}$, its **mean squared error** (MSE) is $\mathbb{E}[(\hat{\theta} - \theta)^2]$, and we can write the MSE as:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2,$$

assuming all terms involved exist.

Proof: The key idea in the proof is to write

$$(\hat{\theta} - \theta)^2 = (\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta)^2 = (\hat{\theta} - \mathbb{E}[\hat{\theta}])^2 + 2(\hat{\theta} - \mathbb{E}[\hat{\theta}])(\mathbb{E}[\hat{\theta}] - \theta) + (\mathbb{E}[\hat{\theta}] - \theta)^2.$$

From this, we can express the MSE as

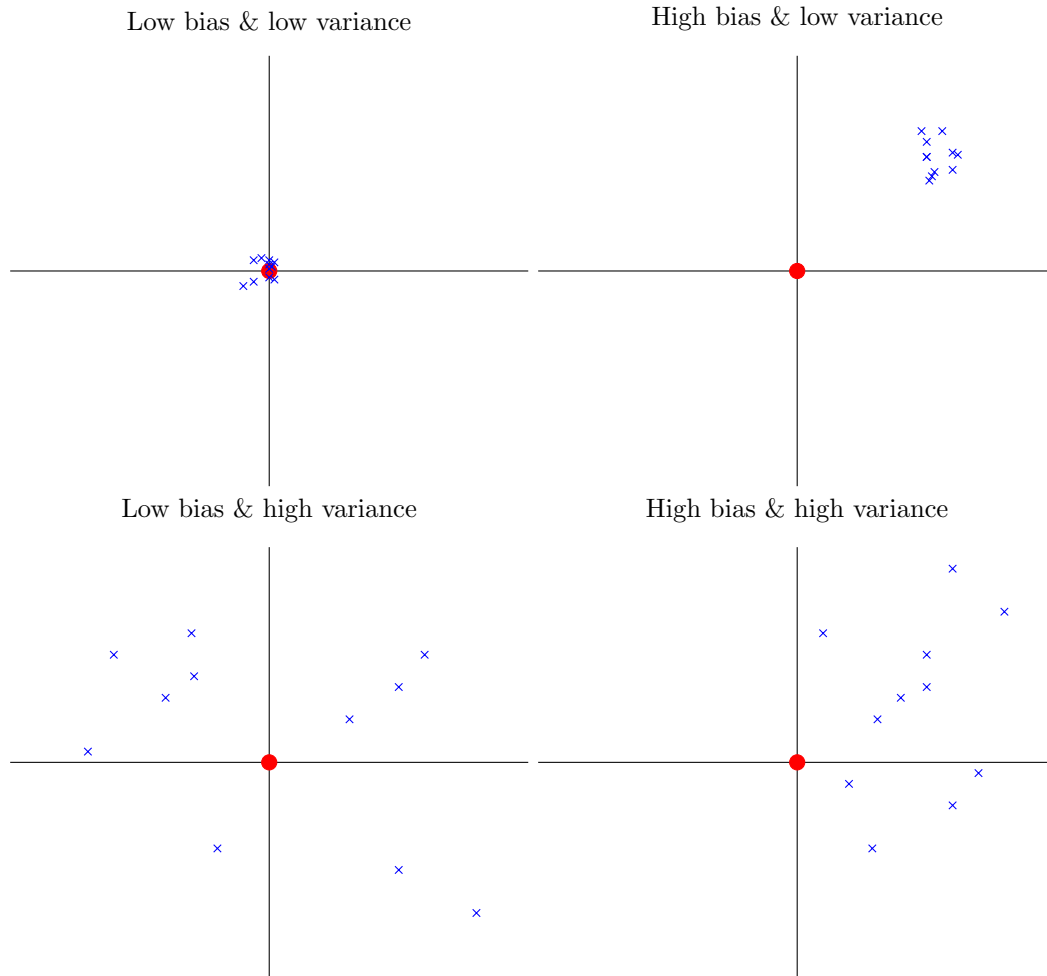
$$\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2] + 2 \underbrace{(\mathbb{E}[\hat{\theta}] - \theta)}_{\text{deterministic}} \underbrace{\mathbb{E}[\hat{\theta} - \mathbb{E}[\hat{\theta}]]}_{\mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}] = 0} + (\mathbb{E}[\hat{\theta}] - \theta)^2 = \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2.$$

□

3.2.2 Since MSE is measuring “average” squared distance between the estimator and true parameter θ , having a *lower* MSE is better (“closer” to the true parameter).

Now, from the bias-variance decomposition, we can observe that to achieve a low MSE, we need *both* low variance and low bias. This suggests that merely being unbiased is not enough: Having a *low* variance is important also!

[Note: Roughly speaking, variance measures “average” *precision* of $\hat{\theta}$: lower variance = higher precision. As one may expect intuitively, both (average) *accuracy* and *precision* matter for an estimator to be “good”.]



3.2.3 The concept of *efficiency* is related to the *variance* of estimator. Consider two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$. The **efficiency** of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is

$$\text{Eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_1)}.$$

When $\text{Eff}(\hat{\theta}_1, \hat{\theta}_2) > 1$ (i.e., $\text{Var}(\hat{\theta}_1) < \text{Var}(\hat{\theta}_2)$), the estimator $\hat{\theta}_1$ is said to be **relatively more efficient** (i.e., having a lower variance; “better” in efficiency) than $\hat{\theta}_2$.

3.2.4 When the two estimators in comparison are both unbiased, the one which is relatively more efficient (having a lower variance) would have a lower MSE \rightarrow better. Consequently, if we focus only on unbiased estimators, estimator with the *minimum variance* is *optimal* (in MSE sense).

More precisely, an estimator $\hat{\theta}$ (of θ) is called **uniformly minimum variance unbiased estimator** (UMVUE) if it has the minimum variance among *all* unbiased estimators of θ , for any possible value of θ .

Remarks:

- Note that the minimum variance requirement needs to be satisfied for any possible value of θ , not just some specific values \rightarrow hence “uniformly”.
- UMVUE characterizes the optimal *unbiased* estimator in MSE sense \rightarrow it is a very “good” estimator.
- It turns out that if UMVUE exists, then it must be *unique* (in almost sure sense, i.e., any two UMVUEs are equal with probability 1).

3.3 Consistency

3.3.1 The concept of *consistency* is again related to an asymptotic setting. Like before, we assume that there are infinitely many i.i.d. observations X_1, X_2, \dots to be included in a sample, and we express the estimator as $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$, for any $n \in \mathbb{N}$.

3.3.2 The estimator $\hat{\theta}$ is **consistent** if

$$\left\{ \hat{\theta}(X_1, \dots, X_n) \right\} \xrightarrow{P} \theta.$$

[Note: This roughly means the estimator can be very close to the true parameter θ with very high probability, when the sample size is sufficiently large.]

3.4 Sufficiency and Likelihood

3.4.1 Given a sample \mathbf{X} , a statistic $T = T(\mathbf{X})$ can *summarize* the information given by the sample, which can be convenient especially if the sample size is very large. However, some information may be *lost* during the summarization process.

3.4.2 For a “nice” statistic T , it should carry all “relevant” information, and this is related to the concept of *sufficiency*.

A statistic $T = T(\mathbf{X})$ is **sufficient** for θ if the conditional distribution of \mathbf{X} given T is *free of* θ .

[Note: Loosely speaking, this means T already carries all information contained in the sample \mathbf{X} that is relevant to θ (“sufficient” information). So when it is conditioned on, we already “completely know” $\theta \rightarrow$ no appearance of unknown θ in distribution anymore (free of θ).]

3.4.3 Although this definition is quite intuitive, it is not very friendly to be directly used for proving sufficiency. Usually, we use the *factorization criterion* for proving sufficiency instead. But before stating the result, let us first introduce the concept of *likelihood*, which will be used in the factorization criterion.

3.4.4 The **likelihood function** of θ given \mathbf{X} is given by

$$L_{\mathbf{X}}(\theta) = f(\mathbf{X}|\theta)$$

where $f(\cdot|\theta)$ is the joint probability function of \mathbf{X} with parameter θ .

Remarks:

- Here $L_{\mathbf{X}}(\theta)$ is *unrealized* and is a random variable. It takes the value $L_{\mathbf{x}}(\theta) = f(\mathbf{x}|\theta)$ (likelihood function of θ given $\mathbf{X} = \mathbf{x}$) when $\mathbf{X} = \mathbf{x}$ is realized.
- The likelihood function $L_{\mathbf{X}}(\theta)$ (function of θ) measures the “likelihood” of each θ being the *true* parameter for the population that generates the sample \mathbf{X} we have.
- The **log-likelihood function** is the natural log of likelihood function: $\ln L_{\mathbf{X}}(\theta)$.

3.4.5 The factorization criterion is as follows:

Theorem 3.4.a (Factorization criterion). A statistic $T = T(\mathbf{X})$ is sufficient for θ iff there exists a function g such that $L_{\mathbf{X}}(\theta) \propto g(T(\mathbf{X}), \theta)$ for any sample \mathbf{X} .

[Note: “ \propto ” means “is proportional to” (where sample \mathbf{X} is treated as “constant”). Here, $L_{\mathbf{X}}(\theta) \propto g(T(\mathbf{X}), \theta)$ means

$$L_{\mathbf{X}}(\theta) = g(T(\mathbf{X}), \theta)h(\mathbf{X})$$

for some function h (factorization).]

3.4.6 If a statistic T is sufficient for θ , then $u(T)$ is sufficient for θ also, where u is any *injective* function.

[Intuition 🧠: When the statistic T carries all “relevant” information, applying an *injective* function u to T should not lead to any loss in information, so $u(T)$ should also carry all “relevant” information.¹]

¹The same thing cannot be said when the function u is not injective. For example, if u is zero function (mapping every input to zero), then basically all information carried in T is lost.

3.4.7 An application of sufficient statistics is suggested as follows.

Theorem 3.4.b (Rao-Blackwell theorem). Let $\hat{\theta}$ be an unbiased estimator of θ , and $T = T(\mathbf{X})$ be a sufficient statistic of θ . Suppose that $\mathbb{E}[\hat{\theta}^2]$ is finite. Then, $\hat{\theta}_* = \mathbb{E}[\hat{\theta} | T]$ is also unbiased with

$$\text{Var}(\hat{\theta}_*) \leq \text{Var}(\hat{\theta}).$$

[Note: This means taking conditional expectation on *any* unbiased estimator given a *sufficient* statistic yields a better (or at least not worse) unbiased estimator.]

3.5 Completeness

3.5.1 Completeness is another concept that describes a statistic. A statistic $T = T(\mathbf{X})$ is **complete** for θ if for any function g free of θ ,

$$\mathbb{E}_\theta[g(T(\mathbf{X}))] = 0 \quad \forall \theta \implies \mathbb{P}_\theta(g(T(\mathbf{X})) = 0) = 1 \quad \forall \theta.$$

Remarks:

- $\mathbb{E}_\theta[\cdot]$ and $\mathbb{P}_\theta(\cdot)$ respectively denote expectation operator and probability measure where the parameter θ is used for calculations (true parameter is θ). These notations are used for emphasizing the parameter used in calculations.
- This means that an unbiased estimator of 0 (0 is a function of true parameter θ) *that is a function of the complete statistic T* can only possibly be a function that equals zero with probability 1 (“almost surely zero function”), for any true parameter θ .²

3.5.2 A main result that utilizes both *sufficiency* and *completeness* is the *Lehmann-Scheffé* theorem:

Theorem 3.5.a (Lehmann-Scheffé theorem). Let $T = T(\mathbf{X})$ be a *complete* and *sufficient* statistic of θ . Suppose that $g(T)$ is an unbiased estimator for $\psi(\theta)$ where g and ψ are functions. Then, $g(T)$ is the UMVUE of $\psi(\theta)$.

3.5.3 Given a complete and sufficient statistic T , we can set $g(T) = \mathbb{E}[Y|T]$ where Y is any unbiased estimator for $\psi(\theta)$, which is an unbiased estimator of $\psi(\theta)$ ³. Then, by Lehmann-Scheffé theorem, $\mathbb{E}[Y|T]$ is the UMVUE of $\psi(\theta)$.

[Note: This method always results in the unique UMVUE, regardless of the choice of Y .]

3.6 Cramer-Ráo Inequality

3.6.1 *Cramer-Ráo inequality* provides another method for finding UMVUE. Before stating the inequality, we first introduce some terminologies.

3.6.2 The **score function** is the partial derivative of log-likelihood function with respect to θ :

$$S_{\mathbf{X}}(\theta) = \frac{\partial \ln L_{\mathbf{X}}(\theta)}{\partial \theta}.$$

[Note: When score function takes extreme value, it indicates that the shape of log-likelihood function varies sharply for different θ (“likelihoods” for different θ have clear difference) \rightarrow easy to deduce θ based on a sample (the sample “carries” rich information about θ).]

3.6.3 The **Fisher information** is the variance of score function: $I(\theta) = \text{Var}(S_{\mathbf{X}}(\theta))$.

Remarks:

²The condition “free of θ ” is to ensure that applying the function g results in a *legitimate* estimator that does not use an *unknown* parameter!

³We have $\mathbb{E}[g(T)] = \mathbb{E}[\mathbb{E}[Y|T]] = \mathbb{E}[Y] = \psi(\theta)$.

- When the variance of score function is high, score function tends to take more extreme values \rightarrow the sample carries more information about $\theta \rightarrow$ Fisher information measures information about θ carried by sample.
- Usually $\mathbb{E}[S_{\mathbf{X}}(\theta)] = 0$, and we can write $I(\theta) = \mathbb{E}[(S_{\mathbf{X}}(\theta))^2]$ in this case.

3.6.4 Now we are ready to state the Cramér-Rao inequality:

Theorem 3.6.a (Cramér-Rao inequality). Let $\psi(\theta)$ be a differentiable function of θ , and $T = T(\mathbf{X})$ be an unbiased estimator of $\psi(\theta)$. Suppose that the support of common distribution of observations in \mathbf{X} is free of θ . Then,

$$\text{Var}(T) \geq \frac{\psi'(\theta)^2}{I(\theta)},$$

where the RHS is known as **Cramér-Rao lower bound** (CRLB).

[Note: This result suggests that an unbiased estimator T^* with variance being CRLB must be the UMVUE of $\psi(\theta)$ (since it has the least variance among all unbiased estimators), assuming the support is free of θ .]

3.7 Maximum Likelihood Estimation and Method of Moments

3.7.1 **Maximum likelihood estimator** (mle) $\hat{\theta}$ of θ is *maximizer* of the likelihood function $L_{\mathbf{X}}(\theta)$ (or equivalently, the log-likelihood function $\ln L_{\mathbf{X}}(\theta)$).

[Note: To find the mle, standard optimization technique can be used. In some “nice” cases, the mle is the solution to

$$\frac{d \ln L_{\mathbf{X}}(\theta)}{d\theta} = 0$$

when $\theta \in \mathbb{R}$, or

$$\frac{\partial \ln L_{\mathbf{X}}(\theta)}{\partial \theta_1} = \dots = \frac{\partial \ln L_{\mathbf{X}}(\theta)}{\partial \theta_k} = 0$$

when $\theta = (\theta_1, \dots, \theta_k) \in \mathbb{R}^k$.]

3.7.2 For method of moments, we first let μ_k be the k th moment of the random variable X underpinning the population: $\mathbb{E}[X^k]$, and m_k be the k th *sample* moment: $m_k = \frac{1}{n} \sum_{i=1}^n X_i^k$.

Now, assume $\theta = h(\mu_1, \dots, \mu_p)$ for some function h . Then, **method of moments estimator** (mme) of θ is

$$\hat{\theta} = h(m_1, \dots, m_p).$$

[Note: The method of moments estimator is *not* unique, since the function h is not unique (there can be multiple ways to express θ in terms of moments).]

3.8 Asymptotic Properties of MLE

3.8.1 It turns out that MLE possesses some nice *asymptotic* properties. So, we consider the asymptotic case here again. Likewise, suppose that there are infinitely many i.i.d. observations X_1, X_2, \dots to be included in a sample, and let $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$ be mle of θ based on the sample (X_1, \dots, X_n) (i.e., this sample is used in the underlying likelihood function).

3.8.2 Now, under certain regularity conditions,

- $\hat{\theta}$ is consistent for θ ;
- $\left\{ \frac{\hat{\theta}(X_1, \dots, X_n) - \theta}{\sqrt{1/I(\theta)}} \right\} \xrightarrow{d} Z$ where $Z \sim N(0, 1)$,

assuming $I(\theta)$ exists.

[Note: This suggests that when n is large, $\hat{\theta}(X_1, \dots, X_n)$ is *approximately* normally distributed with mean θ and variance $1/I(\theta)$ (CRLB) \rightarrow “approximately” UMVUE. This provides some theoretical support on using maximum likelihood estimation.]

4 Interval Estimation

- 4.0.1 In section 3, we focus on *point estimation*, which gives a *single value* as an estimate of true parameter θ . Here, we shall focus on *interval estimation* which gives an *interval* of “plausible” values as an estimate of true parameter θ . It can suggest the amount of *uncertainty* we have in the estimation: wider interval \rightarrow more uncertainties involved and less “precise”.

4.1 Confidence Intervals

- 4.1.1 The core concept in interval estimation is *confidence interval* (CI). Given a sample $\mathbf{X} = (X_1, \dots, X_n)$ collected from the population, a $100(1 - \alpha)\%$ **confidence interval for θ** is any (random) interval $\mathcal{I}(\mathbf{X})$ where for every θ ,

$$\mathbb{P}(\theta \in \mathcal{I}(\mathbf{X})) = 1 - \alpha,$$

with $\alpha \in [0, 1]$.

- 4.1.2 To *interpret* a $100(1 - \alpha)\%$ CI $\mathcal{I}(\mathbf{X})$, we consider a large number of realized confidence intervals $\mathcal{I}(\mathbf{x})$. Then, approximately $100(1 - \alpha)\%$ of them contains the true parameter θ .

- 4.1.3 Often the CI $\mathcal{I}(\mathbf{X})$ takes the form of $[L(\mathbf{X}), U(\mathbf{X})]$ where L and U are functions, and a CI of this form is called **two-sided**.

Sometimes, the CI $\mathcal{I}(\mathbf{X})$ is instead **one-sided**: taking the form of $[L(\mathbf{X}), \infty)$ or $(-\infty, U(\mathbf{X})]$.

4.2 Constructing Confidence Intervals

- 4.2.1 There are three typical approaches for constructing a CI:

- (a) using *pivotal quantity*
- (b) constructing approximated CI based on large sample
- (c) inverting hypothesis test (see section 5)

The last approach provides some theoretical support on the “goodness” of CI. Usually, when we have a “good” hypothesis test (more details to be discussed in section 5), inverting it would yield a “good” CI. Indeed, methods based on pivotal quantity often originate from inverting some “standard” hypothesis tests.

- 4.2.2 A **pivotal quantity** is a random variable $Q = Q(\mathbf{X}, \theta)$ whose distribution is free of θ . Using quantiles of Q , we can construct a $100(1 - \alpha)\%$ CI. For example, by writing

$$\ell \leq Q(\mathbf{X}, \theta) \leq u \iff L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})$$

where $\mathbb{P}(\ell \leq Q \leq u) = 1 - \alpha$, the interval $[L(\mathbf{X}), U(\mathbf{X})]$ serves as a $100(1 - \alpha)\%$ CI.

- 4.2.3 To construct approximated CI based on large sample, typically we utilize *central limit theorem*.

Consider a sample $\bar{X} = (X_1, \dots, X_n)$ where the common distribution has a finite mean μ and finite variance σ^2 . Here, we focus on estimating the parameter μ using a CI, and for the purpose of constructing CI, we shall replace σ by its mle $\hat{\sigma} = \hat{\sigma}(\mathbf{X})$ (as a further “approximation”).

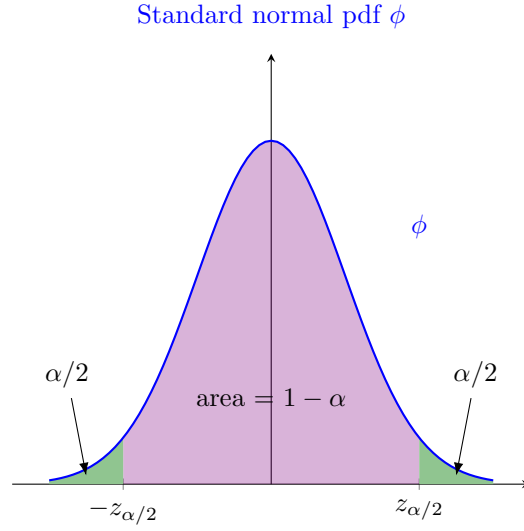
By central limit theorem,

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}} \underset{\text{approx.}}{\sim} N(0, 1)$$

when n is large. Hence, we can use this as an “approximated pivotal quantity”:

$$\mathbb{P}\left(-z_{\alpha/2} \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}} \leq z_{\alpha/2}\right) \approx 1 - \alpha,$$

where z_p is the p th upper quantile of standard normal distribution (i.e., $\mathbb{P}(Z > z_p) = p$ when $Z \sim N(0, 1)$).



Then, we have

$$-z_{\alpha/2} \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}} \leq z_{\alpha/2} \iff \mu \in \left[\bar{X}_n - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right].$$

So, an approximated $100(1 - \alpha)\%$ CI for μ is

$$\left[\bar{X}_n \pm z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right]$$

which is a shorthand for $\left[\bar{X}_n - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{X}_n + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right]$.

5 Hypothesis Testing

- 5.0.1 Instead of *estimating* the parameter θ , here we perform some *tests* on what the true parameter θ could and could not be.

5.1 Hypothesis Tests

- 5.1.1 A **hypothesis test** takes the following form:

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1$$

for some sets Θ_0 and Θ_1 . Here, H_0 is known as **null hypothesis** and H_1 is known as **alternative hypothesis**.

Each hypothesis postulates a range of possible values of true parameter θ , and the main aim of hypothesis testing is to *test* which hypothesis (range) is *more plausible*. [⚠ **Warning**: “More plausible” is not the same as “true”!]

[Note: In general, the intersection $\Theta_0 \cap \Theta_1$ may be *nonempty* and the union $\Theta_0 \cup \Theta_1$ may *not* be the whole parameter space (i.e., set of all possible values of θ). (But often the intersection is indeed empty and the union is indeed the whole parameter space.)]

- 5.1.2 To conduct a hypothesis test, we seek for empirical data evidence from the sample \mathbf{X} that supports **rejection** of null hypothesis H_0 *in favour of* alternative hypothesis H_1 . [⚠ **Warning**: Again this does not mean alternative hypothesis H_1 is *true*!]

[Note: When null hypothesis H_0 is *rejected* (i.e., values in Θ_0 are deemed “not plausible”), the values in $\Theta_1 \setminus \Theta_0$ automatically become “more plausible” (due to having less “competition”) \rightarrow alternative hypothesis H_1 becomes more favourable/plausible (as long as $\Theta_1 \setminus \Theta_0$ is nonempty, which is almost always the case).]

- 5.1.3 When there is “adequate” evidence from \mathbf{X} that supports the rejection, we *reject* H_0 in favour of H_1 (or “accept H_1 ”). On the other hand, when there is “inadequate” evidence from \mathbf{X} that supports the rejection, we *do not reject* H_0 (or “accept H_0 ”).

[⚠ **Warning**: Here, “reject” and “accept” do not carry the meaning of “declaring as false” or “declaring as true” (respectively). Rather, rejecting (accepting) a hypothesis here just means deeming the hypothesis as “not plausible” (“plausible”).]

- 5.1.4 More theoretically and mathematically, a hypothesis test can be described as a **test function** (or simply **test**) φ defined by

$$\varphi(\mathbf{X}) = \mathbb{P}_\theta(\text{reject } H_0 | \mathbf{X}).$$

[Note: This means that we reject H_0 with probability $\varphi(\mathbf{X})$ when the sample \mathbf{X} is observed (“unrealized” case), or we reject H_0 with probability $\varphi(\mathbf{x})$ when the sample $\mathbf{X} = \mathbf{x}$ is observed (“realized” case).

More specifically, when the “probability” is one (zero), it simply means we reject H_0 (do not reject H_0) and the probabilistic meaning is gone.]

- 5.1.5 A commonly seen form of test function (which is also our focus here) is given by

$$\varphi(\mathbf{X}) = \mathbf{1}_{\{\mathbf{X} \in \mathcal{C}\}}.$$

where $\mathbf{1}_{\{\cdot\}}$ is the indicator function, and \mathcal{C} is known as **critical region** (or **rejection region**).

Remarks:

- This means we reject H_0 iff $\mathbf{X} \in \mathcal{C}$.
- When $\mathbf{X} = \mathbf{x}$ is observed, the test function is given by $\varphi(\mathbf{x})$, and so in this context H_0 is rejected iff $\mathbf{x} \in \mathcal{C}$.

5.1.6 When we reject H_0 , sometimes we call the evidence as **statistically significant**. Note that “significant” do *not* carry the meaning of “importance” here. Instead, it means the evidence *signifies* that H_0 should be rejected (it is adequate for supporting the rejection).

[Note: Likewise, when an evidence is called **statistically insignificant**, it just means that the evidence does *not* signify that H_0 should be rejected.]

5.1.7 A common form of the critical region \mathcal{C} is

$$\mathcal{C} = \{\mathbf{X} : T(\mathbf{X}) > c\} \quad (\text{or simply } \{T(\mathbf{X}) > c\})$$

where $T = T(\mathbf{X})$ is a statistic (known as **test statistic**) and c is a constant (known as **critical value**). In this case, we reject H_0 iff $T(\mathbf{X}) > c$.

Remarks:

- The inequality “ $>$ ” can be replaced by “ $<$ ”, “ \leq ”, or “ \geq ”.
- When $\mathbf{X} = \mathbf{x}$ is observed, the critical region would be modified to $\{\mathbf{x} : T(\mathbf{x}) > c\}$ (similar for critical regions with other kinds of inequalities).

5.2 Quality of a Hypothesis Test

5.2.1 To assess the quality/“goodness” of a test φ , we mainly use two metrics: *size* and *power*. Before discussing them, let us first introduce a preliminary concept: *power function* (not to be confused with *power*!).

5.2.2 The **power function** of a test φ is

$$w_\varphi(\theta) = \mathbb{E}_\theta[\varphi(\mathbf{X})] = \mathbb{E}_\theta[\mathbb{P}_\theta(\text{reject } H_0 | \mathbf{X})] = \mathbb{P}_\theta(\text{reject } H_0)$$

where the last equality follows from law of total expectation.

5.2.3 When $\theta \in \Theta_0$ (H_0 is true), the power function is $w_\varphi(\theta) = \mathbb{P}_\theta(\text{reject true } H_0)$, and when $\theta \in \Theta_1 \setminus \Theta_0$ (assuming $\Theta_1 \setminus \Theta_0$ is nonempty), the power function is $w_\varphi(\theta) = 1 - \mathbb{P}_\theta(\text{not reject false } H_0)$.

The act of rejecting true H_0 is known as **type I error**, and the act of not rejecting false H_0 is known as **type II error**.

5.2.4 The **size** of a test φ is $\sup\{w_\varphi(\theta) : \theta \in \Theta_0\}$ (or $\sup_{\theta \in \Theta_0}\{w_\varphi(\theta)\}$), which can be loosely interpreted as the “worse” (“highest”) type I error probability. This is also known as the **significance level** of the test φ .

5.2.5 The **power** of a test φ at $\theta \in \Theta_1 \setminus \Theta_0$ is given by the output of power function at θ : $w_\varphi(\theta)$. This can be interpreted as $1 - \text{type II error probability}$.

5.2.6 The *smaller* the size and the *larger* the power at each $\theta \in \Theta_1 \setminus \Theta_0$ (i.e., type I and II errors are both smaller), the *better* the test. Then, it may seem that the “optimal” test is the one that simultaneously achieves the smallest size and the largest power at each $\theta \in \Theta_1 \setminus \Theta_0$.

5.2.7 Unfortunately, we cannot get “the best of both worlds” in general since often reducing the size would lead to reduction in power also (and increasing the power would lead to increase in size also).

[Intuition 💡: When we try to reduce the size, the test becomes more “conservative” (requiring *very strong* evidence for rejecting H_0) \rightarrow avoiding much type I error, but also raising type II error.

On the other hand, when we try to increase the power, the test becomes more “aggressive” (requiring rather weak evidence for rejecting H_0) \rightarrow rejecting H_0 frequently \rightarrow avoiding much type II error, but also raising type I error.]

- 5.2.8 Due to the constraint in [5.2.7], we usually seek a test φ whose size is (at most) a certain predetermined level α and whose power is *as large as possible* for every $\theta \in \Theta_1 \setminus \Theta_0 \rightarrow$ type I error probability has been controlled, and we try to reduce type II error as far as possible (or in other words, increase the power as far as possible).

[Note: Note that when we consider *power* of a test, we focus on $\Theta_1 \setminus \Theta_0$ rather than Θ_1 (just the parameter values appearing *exclusively* in H_1). Consequently, even if one modifies the alternative hypothesis to $H_1 : \theta \in \Theta_1 \setminus \Theta_0$, there is not much “material” effect, and often the previous results would still hold. Thus, we usually treat the tests before and after this modification as interchangeable.]


5.3 p -value

- 5.3.1 A frequently seen jargon in statistical inference is *p-value* (which is unfortunately a confusing concept to many people!).
- 5.3.2 In the discussion of p -value here, we shall focus on the typical case where the test function is $\varphi(\mathbf{X}) = \mathbf{1}_{\{\mathbf{X} \in \mathcal{C}\}}$.
- 5.3.3 When $\mathbf{X} = \mathbf{x}$ is observed, the *p-value*, denoted by $\text{pv}(\mathbf{x})$, is the “*smallest*” possible size for a test φ while retaining $\varphi(\mathbf{x}) = 1$ (i.e., the test suggests rejection of H_0).

[Note: The term “smallest” is not very accurate technically. More precisely, the p -value $\text{pv}(\mathbf{x})$ is the *infimum*

$$\inf\{\text{size of } \varphi : \varphi \text{ is a test with } \varphi(\mathbf{x}) = 1\}.$$

]

[Intuition : When $\mathbf{X} = \mathbf{x}$ is observed, the p -value $\text{pv}(\mathbf{x})$ indicates the size of the “most conservative test”⁴ we can use to reject H_0 based on \mathbf{x} .

Now, if the p -value $\text{pv}(\mathbf{x})$ is very small, it suggests that the size of the “most conservative test” is very small \rightarrow the “most conservative test” we can use to reject H_0 based on \mathbf{x} is really “very conservative” \rightarrow evidence contained in \mathbf{x} is really strong (even “very conservative” test rejects H_0).

Hence, the p -value $\text{pv}(\mathbf{x})$ also measures the “strength” of evidence contained in \mathbf{x} (smaller $\text{pv}(\mathbf{x}) \rightarrow$ stronger evidence contained).]

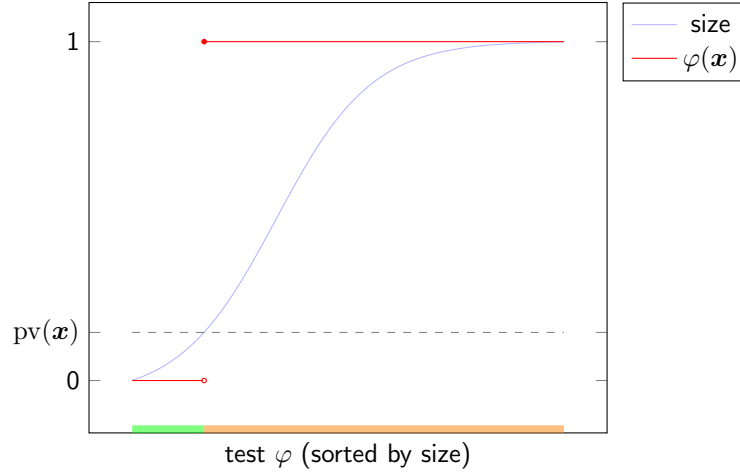
- 5.3.4 The main usage of p -value is to provide an alternative way to decide whether H_0 is rejected (as an alternative to checking whether $\varphi(\mathbf{x}) = 1$, i.e., whether $\mathbf{x} \in \mathcal{C}$), when $\mathbf{X} = \mathbf{x}$ is observed. The alternative method is as follows.

Proposition 5.3.a. Suppose that $\mathbf{X} = \mathbf{x}$ is observed. Focus on any collection of tests in which a test with higher size does not have smaller $\varphi(\mathbf{x})$. Then, the null hypothesis H_0 is rejected at significance level α ⁵ iff the p -value $\text{pv}(\mathbf{x}) \leq \alpha$.

Proof:

⁴Here, “more conservative” \leftrightarrow more “prudent” in rejecting H_0 (requiring “stronger” evidence) \leftrightarrow smaller size. So, “most conservative” \leftrightarrow “smallest size”.

⁵That is, the test is of significance level/size α .



The key observation is that for such collection of tests, the red function $\varphi \mapsto \varphi(\mathbf{x})$ above is “nondecreasing” (as test φ with higher size does not have smaller $\varphi(\mathbf{x})$).

Consequently, for any test with size $\alpha \geq \text{pv}(\mathbf{x})$ (orange highlighted), H_0 is rejected, and for any test with size $\alpha < \text{pv}(\mathbf{x})$ (green highlighted), H_0 is not rejected.

□

[Note: In most cases of practical interest, the collection of tests focused satisfies the condition imposed here, and thus this alternative method can be utilized. Indeed, this often serves as the main/standard approach for deciding whether H_0 is rejected when a computer can be used to calculate p -values.]

5.3.5 The following give some commonly seen/conventional formulas related to p -value:

(a) Focusing on tests having critical region of the form $\{T(\mathbf{X}) > c\}$:

$$\text{pv}(\mathbf{x}) = \inf_{c < T(\mathbf{x})} \sup_{\theta \in \Theta_0} \underbrace{\mathbb{E}_\theta[\varphi(\mathbf{X})]}_{\mathbb{P}_\theta(T(\mathbf{X}) > c)} = \sup_{\theta \in \Theta_0} \underbrace{\inf_{c < T(\mathbf{x})} \mathbb{P}_\theta(T(\mathbf{X}) > c)}_{\mathbb{P}_\theta(T(\mathbf{X}) \geq T(\mathbf{x}))} = \boxed{\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(\mathbf{X}) \geq T(\mathbf{x}))}.$$

(b) Focusing on tests having critical region of the form $\{T(\mathbf{X}) < c\}$:

$$\text{pv}(\mathbf{x}) = \inf_{c > T(\mathbf{x})} \sup_{\theta \in \Theta_0} \underbrace{\mathbb{E}_\theta[\varphi(\mathbf{X})]}_{\mathbb{P}_\theta(T(\mathbf{X}) < c)} = \sup_{\theta \in \Theta_0} \underbrace{\inf_{c > T(\mathbf{x})} \mathbb{P}_\theta(T(\mathbf{X}) < c)}_{\mathbb{P}_\theta(T(\mathbf{X}) \leq T(\mathbf{x}))} = \boxed{\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(T(\mathbf{X}) \leq T(\mathbf{x}))}.$$

(c) Focusing on tests having critical region of the form $\{|T(\mathbf{X})| > c\}$:

$$\text{pv}(\mathbf{x}) = \inf_{c < |T(\mathbf{x})|} \sup_{\theta \in \Theta_0} \underbrace{\mathbb{E}_\theta[\varphi(\mathbf{X})]}_{\mathbb{P}_\theta(|T(\mathbf{X})| > c)} = \sup_{\theta \in \Theta_0} \underbrace{\inf_{c < |T(\mathbf{x})|} \mathbb{P}_\theta(|T(\mathbf{X})| > c)}_{\mathbb{P}_\theta(|T(\mathbf{X})| \geq |T(\mathbf{x})|)} = \boxed{\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(|T(\mathbf{X})| \geq |T(\mathbf{x})|)}.$$

[Mnemonic 🧠: Replacing the “ c ” in the expression for critical region by $T(\mathbf{x})$ (or $|T(\mathbf{x})|$) gives the expression inside $\mathbb{P}_\theta(\cdot)$ in each of the formula.]

Remarks:

- Recall that $\sup_{\theta \in \Theta_0} \mathbb{E}_\theta[\varphi(\mathbf{X})]$ is the size of the test φ .
- Here we assume that “sup” and “inf” can be swapped without affecting the value (this holds under some nice cases, and this case is one of them).
- The same formulas still hold when we replace “ $>$ ” (“ $<$ ”) by “ \geq ” (“ \leq ”) in the critical regions.
- In general, even with these formulas, it is quite hard to compute p -values by hand. (But they are helpful for obtaining p -values by computer.)

5.4 Uniformly Most Powerful Test

5.4.1 Based on the idea in [5.2.8], we can characterize an *optimal* test in the sense of the notion in [5.2.8]. Such test is known as *uniformly most powerful test*.

5.4.2 A test φ_0 is **uniformly most powerful** (UMP) of size α if

- φ_0 is of size α ;
- the power function $w_{\varphi_0}(\theta) \geq w_{\varphi}(\theta)$ for any $\theta \in \Theta_1 \setminus \Theta_0$ and any test φ of size $\leq \alpha$.

Remarks:

- UMP test may not exist. But if it exists, then it is a desirable test.
- The word “uniformly” reflects that the inequality holds for every $\theta \in \Theta_1 \setminus \Theta_0$ (not just a particular θ). In case $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$ (with $\theta_0 \neq \theta_1$), we can drop the word “uniformly” and just call the test φ_0 **most powerful** (MP) of size α .

5.5 Likelihood Ratio Test

5.5.1 An important kind of hypothesis test which possesses some nice properties is *likelihood ratio test*. As its name suggests, it involves a “ratio of likelihood functions” (as test statistic).

5.5.2 For a hypothesis test with $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$, the **likelihood ratio** for this test given sample \mathbf{X} is

$$\Lambda_{\mathbf{X}}(H_0, H_1) = \frac{\sup_{\theta \in \Theta_1} L_{\mathbf{X}}(\theta)}{\sup_{\theta \in \Theta_0} L_{\mathbf{X}}(\theta)}$$

where $L_{\mathbf{X}}(\theta)$ is the likelihood function of θ given \mathbf{X} .

Remarks:

- The likelihood ratio measures the “strength of evidence” provided by \mathbf{X} against H_0 in favour of H_1 . (Higher value \rightarrow “stronger” evidence.)
- Roughly, the likelihood ratio means the “odds” of H_1 against H_0 : When it is higher, there is a higher “tendency” to reject H_0 in favour of H_1 .

5.5.3 After having the likelihood ratio, we can use it as test statistic to obtain a **likelihood ratio test** which is a test with critical region

$$\mathcal{C} = \{\mathbf{X} : \Lambda_{\mathbf{X}}(H_0, H_1) > c\}$$

where $c > 0$ is a constant.

5.5.4 As a special case, when $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta_1\}$, the critical region for likelihood ratio test takes the form

$$\mathcal{C} = \left\{ \mathbf{X} : \frac{L_{\mathbf{X}}(\theta_1)}{L_{\mathbf{X}}(\theta_0)} > c \right\}$$

where $c > 0$ is a constant. In this case, the likelihood ratio test is proven to be UMP by the following result.

Theorem 5.5.a (Neyman-Pearson lemma). The likelihood ratio test in this special case, with size α , is UMP of size α .

5.6 Tests Based on Asymptotic Theory

5.6.1 Suppose that there are infinitely many independent *but not necessarily identically distributed* random variables X_1, X_2, \dots to be included in a “sample” (technically due to the lack of identically distributed property, it is not really a sample).

Now, for any $n \in \mathbb{N}$, consider the “sample” (X_1, \dots, X_n) , and suppose that its joint probability function can be expressed as a product of marginal probability functions as follows:

$$p_1(x_1|\theta) \times \cdots \times p_n(x_n|\theta),$$

where $\theta \in \Theta \subseteq \mathbb{R}^r$ is a *vector*.

5.6.2 Then, consider the hypothesis test

$$H_0 : \theta \in \Theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta_1$$

where $\Theta_0 \subseteq \Theta_1 \subseteq \mathbb{R}^r$, and Θ_0 and Θ_1 have s and r *degrees of freedom* with $s < r$.

[Note: Informally, degree of freedoms is the number of “free” parameters available. For example, since Θ_0 has s degrees of freedom, although every vector in Θ_0 has $r > s$ entries,]

5.6.3 Let $\hat{\theta}_n$ and $\tilde{\theta}_n$ be the MLE of θ under H_1 (over $\Theta_1 \supseteq \Theta_0$) and the *constrained* MLE of θ under H_0 (only the values in Θ_0 are admissible) respectively. Then, the **generalized likelihood ratio** for this test given sample \mathbf{X} is

$$2 \ln \Lambda_{\mathbf{X}}(H_0, H_1) = 2 \ln \left[\frac{\sup_{\theta \in \Theta_1} L_{\mathbf{X}}(\theta)}{\sup_{\theta \in \Theta_0} L_{\mathbf{X}}(\theta)} \right] = 2 \ln \left[\frac{L_{\mathbf{X}}(\hat{\theta}_n)}{L_{\mathbf{X}}(\tilde{\theta}_n)} \right] = 2 \left[\sum_{i=1}^n \ln p_i(X_i|\hat{\theta}_n) + \sum_{i=1}^n \ln p_i(X_i|\tilde{\theta}_n) \right]$$

where $\Lambda_{\mathbf{X}}(H_0, H_1)$ is the likelihood ratio for the test given sample \mathbf{X} .

5.6.4 The following result provides a basis for using the generalized likelihood ratio as “asymptotic” test statistic.

Theorem 5.6.a (Wilks’ theorem). Under *regularity conditions*⁶ and if the null hypothesis H_0 is true (i.e, $\theta \in \Theta_0$),

$$\{2 \ln \Lambda_{\mathbf{X}}(H_0, H_1)\} \xrightarrow{d} \chi_{r-s}^2.$$

[Note: Recall that r and s are degrees of freedom for Θ_1 and Θ_0 respectively.]

5.6.5 Now, based on Wilks’ theorem, we can construct a **generalized likelihood ratio test** which has the critical region

$$\mathcal{C} = \{\mathbf{X} : 2 \ln \Lambda_{\mathbf{X}}(H_0, H_1) > \chi_{r-s}^2(\alpha)\}$$

where $\chi_{r-s}^2(\alpha)$ is the α th upper quantile of the distribution χ_{r-s}^2 . This test is *approximately* size α if n is large (we shall assume that the regularity conditions are satisfied).

5.6.6 Consider a special case where $\Theta_0 = \{\theta_0\}$. Then, we can write the hypothesis test as

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \in \Theta.$$

[Note: More commonly, we consider instead the hypothesis test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0$$

where the alternative hypothesis is modified to $H_1 : \theta \in \Theta_1 \setminus \{\theta_0\}$. As mentioned previously, these two tests can be regarded as interchangeable, so the idea of generalized likelihood ratio test still applies for this kind of test (by implicitly considering instead the hypothesis test above).]

⁶We shall omit the technical details here.

In this case, Θ_0 and Θ_1 have 0 and r degrees of freedom respectively. Consequently, by Wilks' theorem, if H_0 is true (i.e., $\theta = \theta_0$), then

$$2 \ln \Lambda_{\mathbf{X}}(H_0, H_1) = 2 \ln \left[\frac{L_{\mathbf{X}}(\hat{\theta}_n)}{L_{\mathbf{X}}(\theta_0)} \right] \xrightarrow{d} \chi_r^2$$

under regularity conditions.

5.7 Goodness of Fit Test

5.7.1 In this section, we study a special kind of hypothesis test which is designed for checking the *goodness of fit* for different probability models (how well different models “fit”/“match” with the observed sample).

5.7.2 Consider a *discrete* population random variable X with support $\{a_1, \dots, a_m\}$ (a_1, \dots, a_m are all distinct) and pmf given by

$$p(a_j) = p_j \quad \forall j = 1, \dots, m$$

where $p_j > 0$ for any $j = 1, \dots, m$ and $m \geq 2$ is an integer.

Different p_j 's correspond to different probability models.

5.7.3 Given a sample $\mathbf{X} = (X_1, \dots, X_n)$ from the population, for any $j = 1, \dots, m$, let O_j be the (observed) number of observations in the sample \mathbf{X} taking the value a_j , i.e.,

$$O_j = \sum_{i=1}^n \mathbf{1}_{\{X_i = a_j\}}.$$

Then, $(O_1, \dots, O_m) \sim \text{Multinomial}(n, p_1, \dots, p_m)$, i.e.,

$$\mathbb{P}(O_1 = n_1, \dots, O_m = n_m) = \frac{n!}{n_1! \dots n_m!} p_1^{n_1} \dots p_m^{n_m}$$

for any integers $n_1, \dots, n_m \geq 0$ with $n_1 + \dots + n_m = n$.

5.7.4 Now, we assess the goodness of fit of a prespecified probability model for the observed sample \mathbf{X} by performing the following hypothesis test

$$H_0 : p_j = p_{j0} \text{ for any } j = 1, \dots, m-1 \quad \text{vs.} \quad H_1 : p_j \neq p_{j0} \text{ for some } j = 1, \dots, m-1$$

where p_{j0} 's are some known constants.

Remarks:

- The $m-1$ parameters p_1, \dots, p_{m-1} (together with sample size n) are already sufficient for uniquely specifying the distribution of population random variable X (probability model), as from these we can deduce $p_m = 1 - p_1 - \dots - p_{m-1}$.
- Here the probability model to be assessed is specified in the null hypothesis H_0 .

5.7.5 To construct a critical region, we will again do this asymptotically. First, assume that H_0 is true. Then, we let:

- $p_{m0} = 1 - \sum_{j=1}^{m-1} p_{j0}$
- $E_j = np_j = np_{j0}$ (expected number of observations in the sample \mathbf{X} taking the value a_j) for any $j = 1, \dots, m$

Consider the random variable

$$Q_n = \sum_{j=1}^m \frac{(O_j - E_j)^2}{E_j}.$$

It turns out that when n is large,

$$Q_n \approx 2 \ln \Lambda_{\mathbf{X}}(H_0, H_1).$$

Hence, by Wilks' theorem, when n is large, *approximately*

$$\{Q_n\} \xrightarrow{d} \chi_{m-1}^2.$$

[Note: The degrees of freedom of Θ_0 and Θ_1 are 0 and $m - 1$ respectively in this case.]

Thus, when n is large,

$$Q_n \stackrel{\text{approx.}}{\sim} \chi_{m-1}^2.$$

5.7.6 Consequently, we can construct an *approximated* generalized likelihood ratio test of size α (approximately) by setting the critical region as

$$\mathcal{C} = \{\mathbf{X} : Q_n > \chi_{m-1}^2(\alpha)\}$$

to form the **chi-squared goodness of fit test**.

[Note: As a rule of thumb, we should have $E_j \geq 5$ for any $j = 1, \dots, m$ for this approximation to work "well enough".]

5.8 Test of Independence

5.8.1 In this section, we focus on another special kind of hypothesis test which tests *independence* of random variables.

5.8.2 Here, consider two populations with discrete population random variables X and Y . Suppose that the supports of X and Y are $\{a_1, \dots, a_r\}$ and $\{b_1, \dots, b_c\}$, where $r, c \geq 2$ are integers. Then, we perform a test on parameters sourcing from these two populations to deduce whether X and Y are independent.

5.8.3 To specify the joint distribution of X and Y , we can use the following table:

probability	b_1	\cdots	b_c	row sum
a_1	p_{11}	\cdots	p_{1c}	$p_{1\bullet}$
\vdots	\vdots	\ddots	\vdots	\vdots
a_r	p_{r1}	\cdots	p_{rc}	$p_{r\bullet}$
column sum	$p_{\bullet 1}$	\cdots	$p_{\bullet c}$	1

where p_{ij} is the probability $\mathbb{P}(X = a_i \cap Y = b_j)$, and $p_{\bullet j}$ and $p_{i\bullet}$ are the marginal probabilities $\mathbb{P}(Y = j)$ and $\mathbb{P}(X = i)$ respectively, for any $i = 1, \dots, r$ and $j = 1, \dots, m$.

5.8.4 Now consider the hypothesis test

$$H_0 : X \text{ and } Y \text{ are independent} \quad \text{vs.} \quad H_1 : X \text{ and } Y \text{ are not independent,}$$

or more precisely,

$$\begin{aligned} H_0 : p_{ij} &= p_{i\bullet} p_{\bullet j} \text{ for any } i = 1, \dots, r-1 \text{ and } j = 1, \dots, c-1 \\ \text{vs. } H_1 : p_{ij} &\neq p_{i\bullet} p_{\bullet j} \text{ for some } i = 1, \dots, r-1 \text{ and } j = 1, \dots, c-1. \end{aligned}$$

In this case, a sample for this test is formed from two samples $\mathbf{X} = (X_1, \dots, X_n)$ and $\mathbf{Y} = (Y_1, \dots, Y_n)$ sourcing from the two populations:

$$\mathbf{Z} = ((X_1, Y_1), \dots, (X_n, Y_n)).$$

5.8.5 Based on this sample \mathbf{Z} , we can get a table of *observed frequencies* (called **contingency table** or **cross-tabulation**):

observed frequency	b_1	\cdots	b_c	row sum
a_1	O_{11}	\cdots	O_{1c}	$O_{1\bullet}$
\vdots	\vdots	\ddots	\vdots	\vdots
a_r	O_{r1}	\cdots	O_{rc}	$O_{r\bullet}$
column sum	$O_{\bullet 1}$	\cdots	$O_{\bullet c}$	n

where O_{ij} is the (observed) number of observations in \mathbf{Z} taking the value (a_i, b_j) (observed frequency of (a_i, b_j)).

5.8.6 Now, let

$$Q_n = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where $E_{ij} = n \cdot \frac{O_{i\bullet}}{n} \cdot \frac{O_{\bullet j}}{n} = \frac{O_{i\bullet} O_{\bullet j}}{n}$ is the “expected” number of observations in \mathbf{Z} taking the value (a_i, b_j) when H_0 is true.

It turns out that

$$Q_n \approx 2 \ln \Lambda_{\mathbf{Z}}(H_0, H_1).$$

Hence, like before, we have

$$\{Q_n\} \xrightarrow{d} \chi_{(r-1)(c-1)}^2$$

approximately.

[Note: The degrees of freedom of Θ_0 and Θ_1 are $(r-1) + (c-1)$ and $rc-1$ respectively, and we have $(rc-1) - (r-1) - (c-1) = (r-1)(c-1)$.

For the degree of freedom of Θ_1 is $rc-1$, when $\theta \in \Theta_1$, we can freely set $rc-1$ p_{ij} 's, and then the remaining one can be deduced (as all the probabilities must sum to one).

For the degree of freedom of Θ_0 , under the null hypothesis H_0 , we can freely set the values of $p_{1\bullet}, \dots, p_{r-1,\bullet}$ and $p_{\bullet 1}, \dots, p_{\bullet, c-1}$ ($(r-1) + (c-1)$ of them), and then p_{ij} for any $i = 1, \dots, r-1, j = 1, \dots, c-1$ can be deduced.]

5.8.7 Thus, we can construct an *approximated* generalized likelihood ratio of size α (approximately) by setting the critical region as

$$\mathcal{C} = \{\mathbf{Z} : Q_n > \chi_{(r-1)(c-1)}^2(\alpha)\}$$

to form the **chi-squared test of independence**.

[Note: Similarly, as a rule of thumb, we should have $E_{ij} \geq 5$ for any $i = 1, \dots, r$ and $j = 1, \dots, c$ for this approximation to work “well enough”.]

References

- Bickel, P. J., & Doksum, K. A. (2015). *Mathematical statistics: Basic ideas and selected topics, volume I* (2nd ed.). CRC Press.
- Miller, I., & Miller, M. (2014). *John E. Freund's mathematical statistics with applications* (8th ed.). Pearson.
- Young, G. A., & Smith, R. L. (2005). *Essentials of statistical inference*. Cambridge University Press.

Concepts and Terminologies

$100(1 - \alpha)\%$ confidence interval for θ , 11
 p -value, 15

alternative hypothesis, 13
asymptotically unbiased, 6

Bayesian inference, 5
bias, 6

chi-squared goodness of fit test, 20
chi-squared test of independence, 21
complete, 9
consistent, 8
contingency table, 21
converges almost surely, 3
converges in distribution, 3
converges in mean square, 3
converges in probability, 3
Cramér-Rao lower bound, 10
critical region, 13
critical value, 14
cross-tabulation, 21

efficiency, 7
empirical cumulative distribution function, 2
estimator, 5

Fisher information, 9
frequentist inference, 5

generalized likelihood ratio, 18
generalized likelihood ratio test, 18

hypothesis test, 13

likelihood function, 8
likelihood ratio, 17
likelihood ratio test, 17
log-likelihood function, 8

Maximum likelihood estimator, 10
mean squared error, 6
method of moments estimator, 10
most powerful, 17

null hypothesis, 13

one-sided, 11

pivotal quantity, 11
point estimation, 5
power, 14
power function, 14

rejection region, 13
relatively more efficient, 7

sample, 2
sample mean, 4
sample size, 2
score function, 9
significance level, 14
size, 14
statistic, 5
statistically insignificant, 14
statistically significant, 14
sufficient, 8

test, 13
test function, 13
test statistic, 14
two-sided, 11
type I error, 14
type II error, 14

unbiased, 6
uniformly minimum variance unbiased estimator, 7
uniformly most powerful, 17

Results

Section 1

- theorem 1.2.a: Gilvenko-Cantelli theorem

Section 2

- [2.1.6]: implications between different modes of convergence
- theorem 2.2.a: Slutsky's theorem

- theorem 2.2.b: continuous mapping theorem
- theorem 2.2.c: weak and strong laws of large numbers
- theorem 2.2.d: central limit theorem

Section 3

- proposition 3.2.a: bias-variance decomposition
- theorem 3.4.a: factorization criterion
- theorem 3.4.b: Rao-Blackwell theorem
- theorem 3.5.a: Lehmann-Scheffé theorem
- theorem 3.6.a: Cramér-Rao inequality
- [3.8.2]: asymptotic properties of MLE

Section 5

- proposition 5.3.a: method for deciding whether H_0 is rejected based on p -value
- [5.3.5]: conventional formulas for p -values
- theorem 5.5.a: Neyman-Pearson lemma
- theorem 5.6.a: Wilks' theorem