# Contents

1 Lecture 1: Introduction 2

1.1 Incomplete Repertory of Tasks 2

1.2 Decision-theoretic Framework 2

2 Lecture 2: Evaluation of Statistical Procedures I 2

2.1 How to compare $\delta$? 2

2.2 Comparing risk function I: Bayes risk 3

2.3 Bayes theorem 3

2.4 Bayes risk revisited 3

3 Lecture 3: Location Estimation, Bayes Rules for Parametric Models 4

3.1 Prerequisites & Bayes risk revisited 4

3.2 Estimation in parametric models 4

4 Lecture 4: Evaluation of Statistical Procedures II 5

4.1 Comparing risk function II: minimax 5

4.2 Connection between minimax and Bayes 5

4.3 Admissibility 6

4.4 Unbiasedness 6

5 Lecture 5: Building Statistical Procedure I 7

5.1 Sufficient statistics 7

5.2 Complete statistics 9

5.3 Cramér-Rao bound 10

6 Lecture 6: Building Statistical Procedure II 13

6.1 Substitution principle 13

6.2 Consistency 15

6.3 Asymptotic normality 15

6.4 Maximum likelihood estimate 16

7 Lecture 7: Estimating the precision of estimates 16

7.1 Bootstrap 16

7.2 Delta method 18

8 Lecture 8: Confidence interval 18

8.1 Bayesian confidence/probability intervals 18

8.2 General confidence intervals 18

9 Lecture 9: Hypothesis testing 20

9.1 Setup 20

9.2 Testing evaluation 20

9.3 $p$-value 21
1 Lecture 1: Introduction

1.1 Incomplete Repertory of Tasks

- Estimation. E.g. estimating someone’s weight.
- Testing. E.g. testing whether a treatment will work.
- Classification. E.g. email spam filter.
- Ranking...

1.2 Decision-theoretic Framework

- **Data** $x \in X$, an outcome of random element $X$, a point in the sample space $X$.
- **Action space** $A$, the space of decisions
  - For classification, $A$ is finite with at least two elements.
  - For testing, two possible elements: accept/reject.
- **Decision rule** $\delta$, procedure, any (possibly randomized) function. $\delta : X \mapsto A$.
- **Model** $P$, from which $X$ is drawn, an element of some collection of distributions $\mathcal{P}$.
  - Parametric model $\mathcal{P} = \{P_\theta\}$, with $\theta$ in some space $\Theta$ (say $\mathbb{R}^n$).
- **Loss function** $l(\delta(x), P)$, the loss incurred when action $a = \delta(x)$ is chosen, and $X$ is from $P$. Usually $l \geq 0$.

2 Lecture 2: Evaluation of Statistical Procedures I

2.1 How to compare $\delta$?

- If $a = \delta(x)$ is randomized, then first average the loss over all possible $a$:
  $$\bar{l}(\delta(x), P) = E_a(l(\delta(x), P)).$$
- Compare based on risk:
  $$r_\delta(P) = E_{x \sim P} [l(\delta(x), P)] = \int l(\delta(x), P) dP(x).$$
  - If $\delta$ is randomized, then replace $l$ by $\bar{l}$ first.
  - It depends on $P$. 

Estimation of the mean of normal.  
\( X \sim \mathcal{N}(\theta, 1) \), to estimate \( \theta \) with quadratic loss \( l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2 \). Given an observation \( X \), consider two estimators:

\[
\hat{\delta}(X) = \hat{\theta}(X) = X; \quad \tilde{\delta}(X) = \tilde{\theta}(X) = 0.
\]

Their respective risks are given by

\[
r_{\hat{\delta}}(P) = E[(X - \theta)^2] = Var(X) = 1 \]
\[
r_{\tilde{\delta}}(P) = E[(0 - \theta)^2] = E(\theta^2) = \theta^2
\]

Therefore, none of \( \hat{\delta}(x) \) and \( \tilde{\theta}(x) \) is dominant, because the risks depend on \( \theta \), that is, \( P \), the distribution of \( X \).

2.2 Comparing risk function I: Bayes risk

- Prior distribution \( \Pi \) of \( P \) over distribution space \( \mathcal{P} \).
  - In parametric cases, prior \( \Pi \) of \( \theta \) over its space \( \Theta \).
  - Bayes inference, given \( X \), we can update our belief on \( P \)
    
    \[
    Pr(P|X) = \frac{Pr(X|P)Pr(P)}{Pr(X)} \propto Pr(X|P)\Pi(P)
    \]

- **Bayes risk** is defined by
  
  \[
  R^\Pi_\delta = E_{P \sim \Pi}[r_\delta(P)] = \int r_\delta(P)d\Pi(P).
  \]

- It only depends on the decision rule \( \delta \) and the prior \( \Pi \).

2.3 Bayes theorem

- Suppose that \( f_{U,V}(u,v) \) is a joint density of random elements \( U \) and \( V \). The (marginal) density of \( V \) is

  \[
  f_V(v) = \int f_{U,V}(u,v)du.
  \]

  The conditional density of \( U \) given \( V \) is

  \[
  f_{U|V}(u|v) = \frac{f_{U,V}(u,v)}{f_V(v)} = \frac{f_{U,V}(u,v)}{\int f_{U,V}(u,v)du}.
  \]

  The **Bayes theorem** states

  \[
  f_{V|U}(v|u) = \frac{f_{U,V}(u,v)}{f_U(u)} = \frac{f_{U,V}(u,v)}{\int f_{U,V}(u,v)dv} = \frac{f_{U,V}(u|v)f_V(v)}{\int f_{U|V}(u|v)f_V(v)dv}.
  \]

2.4 Bayes risk revisited

- Let the **posterior risk** be

  \[
  R^\Pi(\delta(X)|P) = E_{P \sim \Pi}[l(\delta(X), P)|X].
  \]

- Bayes risk can be computed via posterior distribution

  \[
  R^\Pi_\delta = E_{P \sim \Pi}[r_\delta(P)] = \int r_\delta(P)d\Pi(P) = \int r_\delta(p)f_P(p)dp
  \]
  
  \[
  = E_{P \sim \Pi}[E_{X \sim P}[l(\delta(X), P)|X]] = \int \left( \int l(\delta(x), P)f_{X|P}(x|p)dx \right)f_P(p)dp
  \]
  
  \[
  = E_{(X,P) \sim (P,\Pi)}[l(\delta(X), P)] = \int \int l(\delta(x), P)f_{X,P}(x,p)dx dp
  \]
  
  \[
  = E_{X \sim P}[E_{P \sim \Pi}[l(\delta(X), P)|X]] = \int \left( \int l(\delta(x), P)f_{P|X}(p|x)dp \right)f_X(x)dx
  \]
  
  \[
  = E_{X \sim P}[R^\Pi(\delta(X)|P)] = \int R^\Pi(\delta(X)|P)f_X(x)dx.
  \]

This will be favourable when posterior distribution \( f_{P|X}(p, x) \) is easily accessible.
3 Lecture 3: Location Estimation, Bayes Rules for Parametric Models

3.1 Prerequisites & Bayes risk revisited

- Assume that $X \sim Q$ and we are interested in estimating some characteristic quantity of the distribution $Q$, say $\theta(Q)$, where $\theta(\cdot)$ is a functional.

- Some characteristic quantities of the distribution $Q$
  - Mean: $\theta(Q) = \int x dQ(x)$. Not always exists (e.g. Cauchy).
  - Median: $\theta(Q)$ satisfies
    \[ \Pr(X \leq \theta(Q)) \geq \frac{1}{2}, \Pr(X \geq \theta(Q)) \geq \frac{1}{2} \]
  - Quantile: For $\tau \in (0, 1)$, $\theta_r(Q)$ satisfies
    \[ \Pr(X \leq \theta_r(Q)) \geq \tau, \Pr(X \geq \theta_r(Q)) \geq 1 - \tau. \]
    When $\tau = \frac{1}{2}$, it is median.

- Evaluation of estimation quality: loss
  - Quadratic loss: $l^{(2)}(a, Q) = (a - \theta(Q))^2$, then $r^{(2)}_\delta(Q) = E^{(2)}_{X \sim Q}[(\delta(X) - \theta(Q))^2]$.
  - Absolute loss: $l^{(1)}(a, Q) = |a - \theta(Q)|$, then $r^{(1)}_\delta(Q) = E^{(1)}_{X \sim Q}(|\delta(X) - \theta(Q)|)$.
  - 0-1 loss: $l^{(0)}(a, Q) = I(a \neq \theta(Q))$, then
    \[
    r^{(0)}_\delta(Q) = E^{(0)}_{X \sim Q}[I(\delta(X) \neq \theta(Q))] \\
    = \Pr(\delta(X) \neq \theta(Q)) \cdot I(\delta(X) \neq \theta(Q)) + \Pr(\delta(X) = \theta(Q)) \cdot I(\delta(X) \neq \theta(Q)) \\
    = \Pr(\delta(X) \neq \theta(Q)) \cdot 1 + \Pr(\delta(X) = \theta(Q)) \cdot 0 \\
    = \Pr(\delta(X) \neq \theta(Q)).
    \]
    The second equation is because $X$ can choose two types of values, those $\delta(X) \neq \theta(Q)$ and those $\delta(X) = \theta(Q)$.

- Bayes approach
  - Get the posterior distribution given $X$.
    * Know the posterior distribution of $Q$, then compute $\theta(Q)$.
    * Or know the posterior distribution of $\theta$ directly.
  - Observe the loss function in question and determine its corresponding characteristic value.
    * For $l^{(2)}$, the solution is the mean of posterior distribution.
    * For $l^{(1)}$, the solution is the median of posterior distribution.
    * For $l^{(0)}$, the solution is the mode of posterior distribution.

3.2 Estimation in parametric models

- Assume we have $n$ independent variables $X_i(i = 1, \cdots, n)$ jointly from distribution $P$, which is determined by $Q$, the identical distribution of each individual $X_i$. Further assume that $\theta(Q)$ is one-to-one map.

- To estimate the quantity $\theta(P)$ given data, we need its posterior distribution. (No loss function for the moment)
Posterior of normal.

\( X_i \sim \mathcal{N}(\mu, \sigma) \), to estimate \( \mu \) given \( X_i = x_i \). Assume normal prior for \( \mu \):
\( \mu \sim \mathcal{N}(\mu_0, \sigma_0) \). That is,

\[
\pi(\mu) = \frac{1}{\sigma_0 \sqrt{2\pi}} \exp\left(-\frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)
\]

\[
q(x|\mu) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)
\]

The posterior of \( \mu \) given data \( x_1, \cdots, x_n \) is

\[
f(\mu) \propto \prod_{i=1}^{n} q(x_i|\mu) \pi(\mu)
\]

\[
= \exp\left(-\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right)
\]

\[
= \exp\left(-\frac{1}{2} \left( \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2} \right) \mu^2 + \left( \frac{n\bar{x}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right) \mu \right).
\]

That is,

\[
\mathcal{N}\left( \frac{\bar{x} \sigma^2 + \mu_0}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}} \right).
\]

In general, if a random variable \( X \) has a density of the form \( K \exp(ax^2 + bx + c) \), then

- \( a < 0 \), otherwise the integral will not converge to 1;
- The density can be expressed as

\[
K \exp(ax^2 + bx + c) = K \exp\left( \frac{c}{a} - \frac{b^2}{4a^2} \right) \exp\left[ a \left( x - \left( -\frac{b}{2a} \right) \right)^2 \right],
\]

which is the density of a normal distribution with \( \mu = -\frac{b}{2a} \) and \( \sigma^2 = -\frac{1}{2a} \);
- \( c \) is free and \( K = K(a, b, c) \) is a normalizing positive constant.

4 Lecture 4: Evaluation of Statistical Procedures II

4.1 Comparing risk function II: minimax

- The minimax risk ("worst case")

\[
\bar{R}_\delta = \sup_{P \in \mathcal{P}} r_\delta(P).
\]

- Minimax rule is the rule that minimize minimax risk.

4.2 Connection between minimax and Bayes

- Suppose \( \delta^\Pi \) is the Bayes rule for some prior \( \Pi \), i.e., \( R^\Pi_{\delta^\Pi} = \inf_{\delta^\Pi} R^\Pi_{\delta} \) and suppose that for all \( P \), \( r_{\delta^\Pi}(P) \leq R^\Pi_{\delta^\Pi} \), then \( \delta^\Pi \) is minimax (and \( \Pi \) is called a least favourable prior).

  - Proof: If \( \delta^\Pi \) were not minimax, then there would exist \( \delta \) such that

    \[
    \sup_{P} r_{\delta}(P) < \sup_{P} r_{\delta^\Pi}(P) \leq R^\Pi_{\delta^\Pi}.
    \]

As the average never exceeds sup, and the average of a constant is that constant, we would have a contradiction with the assumptions:

\[
R^\Pi_{\delta} = E_{P \in \mathcal{P}}[r_{\delta}(P)] \leq \sup_{P} r_{\delta}(P) < \sup_{P} r_{\delta^\Pi}(P) \leq R^\Pi_{\delta^\Pi}.
\]

- If \( \delta \) is the Bayes rule with respect to some prior \( \Pi \), and if it has constant risk, \( r_{\delta}(P) = c \) for all \( P \), then \( \delta \) is minimax.
In fact, \( r_{\delta n}(P) = R^{\Pi}_{\delta n} = c \) in such cases.

### 4.3 Admissibility

- \( \delta \) is **admissible** if there is no \( \tilde{\delta} \) such that \( r_{\tilde{\delta}}(P) \leq r_{\delta}(P) \) with strict inequality \(<\) at least for one \( P \).

- Connection to Bayes rule: if \( \delta^\Pi \) is the unique Bayes rule with respect to a prior \( \Pi \), then \( \delta^\Pi \) is admissible.

  \[
  R^{\Pi}_{\delta} = E_{P \sim \Pi}[r_{\delta}(P)] \leq E_{P \sim \Pi}[r_{\delta n}(P)] = R^{\Pi}_{\delta n}.
  \]

  Not necessarily the middle inequality is \(<\), despite the strict inequality for at least one \( P \) (because difference at single point may not influence their integral; but if \( P \) is discrete or continuous, then strict inequality will hold); however, the result is a contradiction with uniqueness.

- Connection to minimax rule: if \( \delta \) has constant risk and is admissible, then it is minimax.

  \[
  \text{Proof: Let} \; \delta_c \text{ be an admissible rule with constant risk} \; c, \text{ i.e., } r_{\delta_c}(P) = c, \forall P \in \mathcal{P} \\text{. Because} \; \delta_c \text{ is admissible, for any other rule} \; \delta, \text{ there exists} \; P_0 \in \mathcal{P}, \text{ such that} \]

  \[
  r_{\delta_c}(P_0) \leq r_{\delta}(P_0). \]

  Now we prove the claim by contradiction. Assume that \( \delta_c \) is not minimax, then there exists \( \delta \), such that

  \[
  \sup_{P \in \mathcal{P}} r_{\delta}(P) < \sup_{P \in \mathcal{P}} r_{\delta_c}(P). \]

  Since the supremum should be larger than or equal to the risk at any specific \( P \), we have

  \[
  r_{\delta}(P_0) \leq \sup_{P \in \mathcal{P}} r_{\delta}(P) \]

  To combine, we have

  \[
  c = r_{\delta_c}(P_0) \leq r_{\delta}(P_0) \leq \sup_{P \in \mathcal{P}} r_{\delta}(P) < \sup_{P \in \mathcal{P}} r_{\delta_c}(P) = c. \]

  A contradiction. Therefore, the assumption is false. \( \delta_c \) is minimax.

\[ \Delta \]

**James-Stein estimator.**

\( X_i \sim \mathcal{N}(\theta_i, 1), i = 1, \ldots, p, \) to estimate \( \theta_i \) with quadratic loss \( l(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2 \).

The natural estimator is \( \hat{\theta}_i = X_i \). It is admissible for \( p = 1, 2 \), but not for \( p > 3 \); in that case, the James-Stein estimator

\[
\hat{\theta}^{JS}_i = \left(1 - \frac{p - 2}{\sum_{i=1}^{p} X_i^2}\right)^+ X_i
\]

has smaller risk. However, James-Stein estimator is not admissible either.

### 4.4 Unbiasedness

- \( \delta \) is unbiased with respect to a loss \( l \) if for every \( P \)

  \[
  r_{\delta}(P) = E_{X \sim P}[l(\delta(X), P)] \leq E_{X \sim P}[l(\delta(X), Q)], \text{ for all} \; Q. \]

  That is, \( E_{X \sim P}[l(\delta(X), Q)] \) is minimized at \( Q^* = P \).

- When \( P \) is parametrized, \( \delta \) is unbiased with respect to a loss \( l \) if for every \( \theta \)

  \[
  E_{X \sim P}[l(\delta(X), \theta)] \leq E_{X \sim P}[l(\delta(X), \tilde{\theta})], \text{ for all} \; \tilde{\theta}. \]
5 Lecture 5: Building Statistical Procedure I

5.1 Sufficient statistics

- A statistic is a function $T : \mathcal{X} \mapsto \mathbb{R}$.
- A statistic $T$ is called **sufficient** for the model $\mathcal{P}$, if the conditional distribution of the data $X$ given the value of $T(X) = T(x)$ does not depend on $\mathcal{P}$.

### Sufficient statistic for binomial distribution.

Suppose $X \in \{0, 1\}^n$ with i.i.d. entries, where $P(X_i = 1) = p$, $\forall i$. Then $T(X) = X^T 1$ is sufficient:

$$
P(X|X^T 1 = s) = \frac{P(X, X^T 1 = s)}{P(X^T 1 = s)}
= \begin{cases} 
\binom{n}{s} p^s (1-p)^{n-s} & \text{if } X^T 1 = s \\
0 & \text{if } X^T 1 \neq s
\end{cases}
$$

which does not depend on $p$.

- If $T(\cdot)$ is a sufficient statistic for $\mathcal{P}$ and $S$ is a one-to-one function, then $S(T(\cdot))$ is also a sufficient statistic for $\mathcal{P}$.
- A sufficient statistic which is a function of every other sufficient statistic is called **minimal sufficient**.
  - May not exist.
  - In the binomial example, $T(X) = X$ is not, but $T(X) = X^T 1$ is.

- Let $\Pi$ be a prior distribution on $\mathcal{P}$. A statistic $T(\cdot)$ is called **Bayes sufficient** for $\Pi$, if the posterior distribution of $P$ given $X = x$ is the same as the posterior distribution of $P$ given $T(X) = T(x)$, for all $x$.
- (Kolmogorov) If $T(X)$ is sufficient for $\mathcal{P}$, it is Bayes sufficient for every $\Pi$.
  - The converse is also true, but not in general.

- (Rao-Blackwell) **Construct decision rule from sufficient statistics.** Suppose that the loss function is convex for fixed $P$:

$$
l(\alpha_1 a_1 + \alpha_2 a_2, P) \leq \alpha_1 l(a_1, P) + \alpha_2 l(a_2, P)
$$

where $\alpha_1, \alpha_2 \geq 0$ and $\alpha_1 + \alpha_2 = 1$. If $T(X)$ is sufficient for $\mathcal{P}$ and $\delta$ is a decision rule, then the decision rule $\delta^*(X) = E_{\delta(X)}(\delta(X)|T(X))$ has uniformly smaller risk:

$$
r_{\delta^*}(P) \leq r_{\delta}(P), \forall P.
$$
Also, if $\delta$ is unbiased, so is $\delta^*$.  

Rao-Blackwell for binomial distribution. Suppose $X \in \{0,1\}^n$ with independent and entries, where $P(X_i = 1) = p, \forall i$. Then $T(X) = X^T 1$ is sufficient for $p$. Consider $\delta(X) = X_1$ (estimating $p$, unbiased). Then

$$
\delta^*(X) = E_{\delta(X)}[\delta(X)|T(X) = T(x)] \\
= E_{X_i}(X_1|X^T 1 = s) \\
= 0 \cdot P(X_1 = 0|X^T 1 = s) + 1 \cdot P(X_1|X^T 1 = s) \\
= P(X_1 = 1, X^T 1 = s) \\
= \frac{p \cdot \binom{n-1}{s-1} p^{s-1} (1-p)^{(n-1)-(s-1)}}{P(X^T 1 = s)} \\
= \frac{s}{n} = \frac{X^T 1}{n}.
$$

It is unbiased with respect to quadratic loss:

$$
E_{X \sim p}\left[\frac{X^T 1}{n}\right] = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = p.
$$

Thus, its risk is the variance (see bias-variance decomposition):

$$
r_{\delta^*}(P) = Var\left(\frac{X^T 1}{n}\right) = \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2} \cdot n \cdot p(1-p) = \frac{p(1-p)}{n}.
$$

It has uniformly smaller risk than $\delta(X)$ for any $p$:

$$
r_{\delta^*}(P) = E_{X \sim p}\left[l\left(\frac{X^T 1}{n}, p\right)\right] \\
\leq E_{X \sim p}[l(X_1, p)] \\
= p \cdot (1-p)^2 + (1-p) \cdot (0-p)^2 \\
= p(1-p).
$$

• (Neyman-Savage) **Factorization criterion** for sufficient statistics. Suppose that $X$ has a density (or mass). $T$ is sufficient for $\theta$ iff there are $g$ and $h$ such that

$$
f(x|\theta) = g(T(x), \theta) h(x).
$$

- $T$ is sufficient for $\theta$ if and only if the following is true:

$$
T(x) = T(y) \Rightarrow f(x|\theta) = c(x, y) f(y|\theta).
$$

- $T$ is minimal sufficient for $\theta$ if and only if the following is true:

$$
T(x) = T(y) \Leftrightarrow f(x|\theta) = c(x, y) f(y|\theta).
$$
Neyman-Savage factorization criterion for binomial distribution.
Suppose $X \in \{0, 1\}^n$ with independent and entries, where $P(X_i = 1) = p, \forall i$. Then $T(X) = X^T 1$ is sufficient for $p$, since if $T(X) = T(x) = s$,

$$f(x|p) = p^s(1-p)^{n-s} = g(s,p)h(x),$$

where

$$g(s,p) = p^s(1-p)^{n-s}; h(x) = 1.$$

$T$ is minimal. Let

$$f(x|p) = p^{\sum x_i}(1-p)^{n-\sum x_i}; f(y|p) = p^{\sum y_i}(1-p)^{n-\sum y_i}.$$

$T$ is minimal because

$$T(x) = T(y) = s \iff f(x|p) = c(x,y)f(y|p),$$

where $c(x,y) = 1$.

Neyman-Savage factorization criterion for normal distribution.
Suppose $X_i \sim \mathcal{N}(\mu, \sigma^2), i = 1, \cdots, n$, $\sigma^2$ is known, to estimate $\mu$. Then $T(X) = X^T 1$ is sufficient for $\mu$, since if $T(X) = T(x) = s$,

$$f(x|p) = \frac{1}{\sigma^n(2\pi)^{n/2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right)$$

$$= \frac{1}{\sigma^n(2\pi)^{n/2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 \right) \cdot \exp \left( \frac{1}{2\sigma^2} (2\mu \sum_{i=1}^{n} x_i - n\mu^2) \right)$$

$$= h(x)g(s,p),$$

where

$$g(s,p) = \exp \left( \frac{1}{2\sigma^2} (2\mu s - n\mu^2) \right)$$

$$b(x) = \frac{1}{\sigma^n(2\pi)^{n/2}} \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} x_i^2 \right).$$

$T$ is minimal. To see this, set $c(x,y) = \exp \left( -\frac{1}{2\sigma^2} \sum (x_i^2 - y_i^2) \right)$.

5.2 Complete statistics

- Assume a parametric model $\{P_\theta\}$ and the quadratic loss function.
- A statistic $S$ is complete if for every function $g$, independent of $\theta$,

$$E_{X \sim P_\theta} [g(S(X))] = 0, \forall \theta \Rightarrow \Pr_{X \sim P_\theta} [g(S(X)) = 0] = 1, \forall \theta.$$ 

Roughly speaking, if the expectation with respect to all $\theta$ is 0, then $g$ is identically zero.

Complete statistic for binomial distribution.
Suppose $X \in \{0, 1\}^n$ with independent and entries, where $P(X_i = 1) = p, \forall i$. Then $T(X) = X^T 1$ is complete for $p$: if

$$E[g(T(X))] = \sum_{k=0}^{n} g(k) \Pr(T(X) = k) = \sum_{k=0}^{n} g(k) \binom{n}{k} p^k (1-p)^{n-k},$$

equals to zero for all $p \in [0, 1]$, then $g(k) = 0$ for all $k$, because $E[g(T(X))]$ is a polynomial of $p$. 
(Lehmann-Scheffé) Any unbiased estimator based (only) on a complete, sufficient statistic is minimum-variance unbiased estimator. That is, it has the smallest variance (= MSE for unbiased), for all \( \theta \), among all unbiased estimators of \( \theta \).

5.3 Cramér-Rao bound

• In this part we only consider regular models, whose support \( \{ x \mid f(x; \theta) > 0 \} \) does not depend on \( \theta \). Also assumed is that we may interchange integration and differentiation.

• The score function is

\[
 s(x; \theta) = \frac{\partial}{\partial \theta} \log f(x; \theta) = \frac{\partial f(x; \theta)}{f(x; \theta)} .
\]

- Note that

\[
 E_{X \sim P_{\theta}}[s(X; \theta)] = \int \frac{\partial f(x; \theta)}{f(x; \theta)} f(x; \theta) dx
 = \frac{\partial}{\partial \theta} f(x; \theta) dx
 = \frac{\partial}{\partial \theta} \int f(x; \theta) dx
 = \frac{\partial}{\partial \theta} 1 = 0 .
\]

- When \( X \) consists of independent r.v.s then

\[
 s(x; \theta) = \frac{\partial}{\partial \theta} \log f(x; \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \log g(x_i; \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log g(x_i; \theta)
\]

• The Fisher information is defined by

\[
 I(\theta) = Var_{X \sim P_{\theta}}[s(X; \theta)]
 = E_{X \sim P_{\theta}}[s^2(X; \theta)]
 = \int \left( \frac{\partial f(x; \theta)}{f(x; \theta)} \right)^2 f(x; \theta) dx
 = \int \frac{(\partial f(x; \theta))^2}{f(x; \theta)} dx
\]

- Another way to compute \( I(\theta) \) via second derivative. First note that

\[
 \frac{\partial^2}{\partial^2 \theta} \log f = \frac{\partial}{\partial \theta} \frac{f'}{f} = \frac{f'' - f' f'}{f^2} .
\]

Also,

\[
 E_{X \sim P_{\theta}} \left( \frac{f''}{f} \right) = \int \frac{f''}{f} dx = \int f'' dx = 0 \text{ as } \int f dx = 1 .
\]

To combine, we have

\[
 E \left( -\frac{\partial^2}{\partial^2 \theta} \log f \right) = E \left( \frac{f''' - f' f''}{f^2} \right)
 = E \left( \frac{f'}{f} \right)^2 - E \left( \frac{f''}{f} \right)
 = E(s^2) - 0 = I(\theta)
\]

- When \( X \) consists of independent r.v.s then

\[
 I(\theta) = Var_{X \sim P_{\theta}}[s(X; \theta)]
 = Var_{X \sim P_{\theta}} \left[ \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log g(x_i; \theta) \right]
 = \sum_{i=1}^{n} Var_{X \sim P_{\theta}} \left[ \frac{\partial}{\partial \theta} \log g(x_i; \theta) \right]
 = \sum_{i=1}^{n} \int \frac{(\partial g(x_i; \theta))^2}{g(x_i; \theta)} dx_i
\]
If all \( g_i \) are identical, then
\[
I(\theta) = n \int \left( \frac{\hat{g} g(y; \theta)}{g(y; \theta)} \right)^2 dy
\]

- **Cramér-Rao inequality** provides a lower bound on the variance of any statistic \( U(X) \). Consider the covariance of \( s(X; \theta) \) and \( U(X) \). By Cauchy-Schwartz inequality
\[
[Cov_{X \sim P_\theta}(s(X; \theta), U(X))]^2 \leq \text{Var}_{X \sim P_\theta}(s(X; \theta)) \cdot \text{Var}_{X \sim P_\theta}(U(X))
\]
\[
= I(\theta) \cdot \text{Var}_{X \sim P_\theta}(U(X))
\]

To compute \( \text{Cov}_{X \sim P_\theta}(s(X; \theta), U(X)) \) (note Eq.(5.1)):
\[
\text{Cov}_{X \sim P_\theta}(s(X; \theta), U(X)) = E_{X \sim P_\theta}[s(X; \theta)U(X)] - E_{X \sim P_\theta}(s(X; \theta))E_{X \sim P_\theta}[U(X)]
\]
\[
= E_{X \sim P_\theta}[s(X; \theta)U(X)]
\]
\[
= \int \frac{\hat{f}}{f(x; \theta)} U(x)f(x; \theta)dx
\]
\[
= \int \frac{\hat{f}}{\theta} f(x; \theta)U(x)dx
\]
\[
= \frac{\hat{g}}{\theta} \int U(x)f(x; \theta)dx = \frac{\hat{g}}{\theta} E_{X \sim P_\theta}[U(X)].
\]

Therefore, **Cramér-Rao lower bound** gives
\[
\text{Var}_{X \sim P_\theta}(U(X)) \geq \left( \frac{\hat{g} E_{X \sim P_\theta}[U(X)]}{I(\theta)} \right)^2
\]

- When \( U(X) \) is unbiased w.r.t. quadratic loss:
\[
E_{X \sim P_\theta}(U(X)) = \theta.
\]

Because \( \frac{\hat{g}}{\theta} E_{X \sim P_\theta}[U(X)] = \frac{\hat{g}}{\theta} = 1 \), we have
\[
\text{Var}_{X \sim P_\theta}(U(X)) \geq \frac{1}{I(\theta)}.
\]

\[\Box\]

**Cramér-Rao lower bound for binomial distribution.**

Suppose \( X \in \{0,1\}^n \) with independent and entries, where \( P(X_i = 1) = p, \forall i. \)
\( g(x_i; p) = p^x_i (1-p)^{1-x_i} \). To compute the Fisher information
\[
I(p) = n \times E_X, \left[ \frac{\partial}{\partial \theta} \log g(x_i; p) \right]^2
\]
\[
= n \cdot E_X \left[ \left( \frac{x_i}{p} - 1 - x_i \right)^2 \right]
\]
\[
= \frac{n}{p^2(1-p)^2} E_X[(x_i - p)^2]
\]
\[
= \frac{n}{p^2(1-p)^2} \text{Var}_X(x_i) = \frac{n}{p(1-p)}.
\]

Therefore, any unbiased estimator of \( p \) must have a variance (which equals its mean square error) greater than
\[
\frac{1}{I(p)} = \frac{p(1-p)}{n}.
\]

Now let’s compute the variance (also MSE) of an unbiased estimator \( T(X) = \frac{\sum x_i}{n} \):
\[
\text{Var}_X(T(X)) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}_X(x_i) = \frac{p(1-p)}{n}.
\]

Therefore, this estimator makes the Cramér-Rao lower bound tight.

- When Cramér-Rao an equality? The inequality is the result of Cauchy-Schwartz.

Therefore, if the score function has the form
\[
s(x; \theta) = \frac{\partial}{\partial \theta} \log f(x; \theta) = c(\theta) + d(\theta)U(x),
\]
equality will hold. Then
\[ f(x; \theta) = \exp(\eta(\theta)U(X) - a(\theta) + g(x)). \]

That is, exponential family preserves equality in Cramér-Rao. For the exponential family, if we define \( \eta = \eta(\theta) \) as a new parameter, then \( a(\theta) = b(\eta) \). For the density, we have
\[ \int f(x; \eta)dx = \int \exp(\eta U(X) - b(\eta) + g(x))dx = 1. \]

Differentiating in \( \eta \) on both sides (assuming we can interchange integration and differentiation), we have
\[
0 = \int \exp(\eta U(X) - b(\eta) + g(x))(U(x) - b'(\eta))dx \\
= \int U(x) \exp(\eta U(X) - b(\eta) + g(x))dx - b'(\eta) \exp(\eta U(X) - b(\eta) + g(x))dx \\
= E_X(U(X)) - b'(\eta).
\]

Therefore, we have \( E_X(U(x)) = b'(\eta) \). Similarly, we have \( Var_X(U(X)) = b''(\eta) \).

### Cramér-Rao for exponential distribution.

Exponential distribution is specified by
\[
f(x; \lambda) = \begin{cases} 
\lambda e^{-\lambda x} & \text{when } x \geq 0 \\
0 & \text{otherwise}
\end{cases}
\]

where \( \lambda > 0 \) is the parameter. Note that it can be also expressed as
\[
f(x; \theta) = \begin{cases} 
\frac{1}{\theta} e^{\frac{1}{\theta} x} & \text{when } x \geq 0 \\
0 & \text{otherwise}
\end{cases}
\]

for \( \theta > 0 \). With this new parametrization, we can see that \( U(X) = X \) is unbiased for \( \theta \), with least variance (MSE) one can have because of Cramér-Rao inequality.

Note that if we use old parametrization with \( \lambda \), then \( U(X) = -X \) and
\[
\lambda e^{-\lambda x} = e^{-\lambda x - (-\ln \lambda)}.
\]

Thus \( b'(\lambda) = \frac{1}{\lambda} \), so is \( E_X(U(X)) = E_X(-X) = -E_X(X) = \frac{1}{\lambda} \).

(Side note: having an estimator, \( \hat{\theta} \) of \( \theta \), with some properties does not mean that \( \hat{\theta} \) is the estimator of \( g(\theta) \) with the same properties (unless \( g \) is very simple - say, a linear function). For instance, an unbiased estimator for \( \sigma^2 \) may not be unbiased for \( \sigma \)).

### Cramér-Rao for binomial distribution: revisited.

Suppose \( X \in \{0,1\}^n \) with independent and entries, where \( P(X_i = 1) = p, \forall i \).
\[ g(x_i; p) = p^{x_i}(1-p)^{1-x_i}. \]

The joint distribution is
\[
f(x; p) = p^\sum x_i (1-p)^{n-\sum x_i} \\
= \left( \frac{p}{1-p} \right)^{\sum x_i} (1-p)^n \\
= \exp \left( \ln \frac{p}{1-p} \sum x_i + n \ln(1-p) \right).
\]

Let \( \eta = \ln \frac{p}{1-p} \), we have
\[
f(x; p) = \exp \left( \eta \sum x_i - n \ln(1+e^\eta) \right)
\]

Therefore, it is also a member of exponential family. As a result, Cramér-Rao bound is sharp (as we already shown).
6 Lecture 6: Building Statistical Procedure II

6.1 Substitution principle

- Let \( x = (x_1, x_2, \ldots, x_n) \), which is a realization of \( X = (X_1, X_2, \ldots, X_n) \), where \( X_i \) is a r.v. with distribution \( P \). We want to estimate \( \theta(P) \), some characteristic quantity of \( P \). Typically, \( X_i \) are independent, but it is not absolutely necessary; some permutational invariance (exchangeability) is enough.

- An empirical distribution by \( x \) is the discrete distribution that assigns probability \( \frac{1}{n} \) to every point \( x_i \), denoted \( P_x \):

\[
P_x(E) = \frac{1}{n} \text{card}\{i|x_i \in E\}.
\]

- The substitution principle states that to estimate \( \theta(P) \), replace \( P \) by \( P_x \).

#### Moment estimation.

Suppose we want to estimate the \( k \)th moment

\[
\theta(P) = \int z^k dP(z), \quad k = 1, 2, \ldots.
\]

The resultant estimator with substitution principle is

\[
\hat{\theta}(P_x) = \int z^k dP_x(z) = \frac{1}{n} \sum_{i=1}^{n} x_i^k.
\]

#### Variance estimation.

Suppose we want to estimate the variance

\[
\theta(P) = \int \left( z - \int udP(u) \right)^2 dP(z).
\]

The resultant estimator with substitution principle is

\[
\hat{\theta}(P_x) = \frac{1}{n} \sum_{i=1}^{n} \left( x_i - \frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2.
\]
**Linear regression.**

Consider the linear regression model:

\[ Y = \alpha X + \beta + U, \quad (6.1) \]

where \( X \) is input variable and \( Y \) is output variable (jointly with \( U \) from some distribution), \( \alpha, \beta \) are the parameters of the model, and \( U \) is an error term independent of \( X \) (so they are also uncorrelated, \( E(XU) = 0 \)) with \( E(U) = 0 \). Taking expectation on both sides, we have

\[ E(Y) = \alpha E(X) + \beta. \quad (6.2) \]

Moreover, we can multiply Eq.(6.1) by \( X \), then take expectation:

\[ E(XY) = \alpha E(X^2) + \beta E(X). \quad (6.3) \]

With Eq.(6.2) and Eq.(6.3), we can solve \( \alpha, \beta \) as

\[
\alpha = \frac{E(XY) - E(X)E(Y)}{E(X^2) - (E(X))^2} = \frac{Cov(X,Y)}{Var(X)} \\
\beta = E(Y) - \alpha E(X).
\]

By substitution principle, all expectations (variance/covariance) can be computed from sample \( \{(x_1,y_1), (x_2,y_2), \cdots, (x_n,y_n)\} \):

\[
\hat{\alpha} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
\hat{\beta} = \bar{y} - \hat{\alpha}\bar{x}.
\]

**Quantile estimation.**

For \( \tau \in (0,1) \), suppose we are going to estimate the quantile \( q_\tau \) such that

\[ P((-\infty, q_\tau]) \geq \tau; P([q_\tau, +\infty)) \geq 1 - \tau. \]

We can see quantile in a different way. Define “check function” as

\[
\rho_\tau(z) = \begin{cases} 
|z| + (2\tau - 1)z & \text{for } z > 0 \\
2(\tau - 1)z & \text{for } z \leq 0
\end{cases}
\]

Then

\[ q_\tau = \arg\min_c E[\rho_\tau(Z - c)] = \int \rho_\tau(z - c)dP(z). \]

Therefore, given \( \tau \in (0,1) \), to estimate \( q_\tau \), find the minimizer \( c^* \) of

\[ \int \rho_\tau(z - c)dP(z) = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(x_i - c). \]

For instance, if \( \tau = 0.5 \), we are trying to estimate the median, then

\[
\text{card}\{i|x_i \leq c^*\} \geq \frac{n}{2}; \text{card}\{i|x_i \geq c^*\} \geq \frac{n}{2}.
\]

That is, \( c^* \) is just the sample median.

- Can be used to estimate non-parametric model. If we want to estimate the accumulative distribution \( F(z) = P((-\infty, z]) \), by substitution principle, we have

\[ F_n(z) = \mathbb{P}_n(z) = \frac{1}{n} \text{card}\{i|x_i \leq z\}, \]

which is essentially a step function.

\[ - E(\mathbb{P}_n(z)) = F(z). \]

\[ - \text{Var}(\mathbb{P}_n(z)) = \frac{F(z)(1-F(z))}{n}. \]
• F(z) → F(z) as n → ∞ (in probability, almost surely).
• sup_z |F(z) − F(z)| → 0 as n → ∞ (in probability, almost surely).

6.2 Consistency

• If the estimator \( \hat{\theta}_n \) converges to the target/estimated quantity \( \theta \) as \( n \to \infty \), where convergence is determined by
  
  – Convergence in probability
    \[ \Pr(|\hat{\theta} − \theta| ≥ \epsilon) → 0 \text{ for every } \epsilon > 0. \]
  
  – Almost surely (with probability 1) convergence
    \[ \Pr(|\hat{\theta} − \theta| → 0) = 1. \]
  
  – Convergence in some mean sense
    \[ E(|\hat{\theta} − \theta|^p) → 0. \]

then we say \( \hat{\theta}_n \) is consistent.

\[ \text{Consistent estimator for mean and variance.} \]
Assume that \( X_i \) are iid r.v.s and the mean \( \mu = E(X_i) \) exists. Then \( \bar{X}_n \) is a consistent estimator for \( \mu \). By a law of large numbers (have different versions),

\[ \bar{X}_n \xrightarrow{p} \mu \]
as \( n \to \infty \).

Now further assume that the variance \( \sigma^2 = \text{Var}(X_i) \) exists. Consider the quantity

\[ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i − \mu) = \sqrt{n}(\bar{X}_n − \mu). \]

A central limit theorem (again, many versions) states that

\[ \frac{\bar{X}_n − E(\bar{X}_n)}{\sqrt{\text{Var}(\bar{X}_n)}} = \frac{\bar{X}_n − \mu}{\sqrt{\frac{\sigma^2}{n}}} \]

converges in distribution to the standard normal distribution \( N(0, 1) \). It follows that \( \sqrt{n}(\bar{X}_n − \mu) \) converges in distribution to \( N(0, \sigma^2) \).

6.3 Asymptotic normality

• If the estimator, \( \hat{\theta}_n \), of \( \theta \), has the property that \( \sqrt{n}(\hat{\theta}_n − \theta) \) converges in distribution to \( N(0, \sigma^2) \), then we call that estimator asymptotically normal with asymptotic variance \( \sigma^2 \).

  – The smaller \( \sigma^2 \) is, the better (more accurate).

• To compare two asymptotically normal estimator \( \hat{\theta} \) and \( \tilde{\theta} \), with

\[ \sqrt{n}(\hat{\theta} − \theta) \overset{d}{\to} Z \sim N(0, \sigma^2); \quad \sqrt{n}(\tilde{\theta} − \theta) \overset{d}{\to} Z \sim N(0, \tilde{\sigma}^2). \]

The asymptotic relative efficiency (ARE) of \( \tilde{\theta} \) to \( \hat{\theta} \) is defined as

\[ \text{ARE}(\tilde{\theta}, \hat{\theta}) = \frac{\sigma^2}{\tilde{\sigma}^2}. \]
ARE of sample mean versus sample median.

Assume that \( X_i \) are iid r.v.s whose mean and median are both \( \mu \).

- Suppose that the variance of \( X_i \) is \( \sigma^2 \); from the central limit theorem, we know that \( \sigma^2 \) is the asymptotic variance of the sample mean.

- Suppose that the common density, \( f \), of \( X_i \) exists and is positive at \( \mu \). For the sample median, Kolmogorov proved that under these assumptions, it is asymptotically normal with the asymptotic variance \( \frac{1}{4f(\mu)^2} \).

For instance, if the distribution of \( X_i \) is normal, then the asymptotic variance of sample mean is \( \sigma^2 \), while the asymptotic variance of sample median is

\[
\frac{1}{4f(\mu)^2} = \frac{\pi}{2} \sigma^2.
\]

Therefore,

\[
ARE(\mu_{\text{median}}, \mu_{\text{mean}}) = \frac{\sigma^2}{\frac{\pi}{2} \sigma^2} = \frac{2}{\pi} \approx 0.6366,
\]

which means sample mean \( \mu_{\text{mean}} \) is more efficient.

For any unimodal \( f \) (only has one mode), the ratio is \( \geq 1/3 \) and there are \( f \) with \( > 1 \), i.e., the sample median is more efficient (\( t \) distribution with 3 or 4 degrees of freedom, for instances).

### 6.4 Maximum likelihood estimate

- **Likelihood**
  \[
  L(\theta) = f(x; \theta).
  \]
  If we have independent r.v.s, then
  \[
  L(\theta) = f(x; \theta) = \prod_{i=1}^{n} g(x_i; \theta).
  \]

- **Maximum likelihood estimate** is given by
  \[
  \hat{\theta} = \text{argmax}_\theta L(\theta).
  \]
  Usually take the logarithm when we have independent r.v.s.

- Suppose that \( \hat{\theta}_n \) are maximum likelihood estimators of \( \theta \), from iid sample where the distribution of \( X_i \) is specified by \( \theta \). Then typically,
  - Maximum likelihood estimators are consistent (in probability): \( \hat{\theta}_n \xrightarrow{P} \theta \).
  - They are asymptotically normal, and asymptotically efficient:
    \[
    \sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} Z \sim N \left( 0, \frac{1}{I(\theta)} \right),
    \]
    where \( I(\theta) \) is the Fisher information for one observation from the family parametrized by \( \theta \). (Note that the Fisher information for the whole sample \( X_1, X_2, \ldots, X_n \) is \( nI(\theta) \).)

### 7 Lecture 7: Estimating the precision of estimates

#### 7.1 Bootstrap

- We care about how accurate our prediction is.
Standard error: a canonical example.

Consider an example where \( X_1, \ldots, X_n \) are i.i.d. r.v.s with the same distribution \( P \) with mean \( \mu \) and variance \( \sigma^2 \). We are estimating the mean \( \mu \) by \( \hat{X} \). It is known that \( \text{Var}(\hat{X}) = \frac{\sigma^2}{n} \). So the standard error of sample mean is given by

\[
se_{X \sim P} (\hat{X}) = \frac{\sigma}{\sqrt{n}}.
\]

However, we don’t know \( \sigma \), so we can estimate this by substitution principle, we have

\[
se_{X \sim P} (\hat{X}) = \frac{1}{\sqrt{n}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{X})^2}.
\]

(Sometimes, \( n - 1 \) is preferable.)

Unlike the example above, standard error (standard deviation) of estimator \( \hat{\theta} \) may not be calculated in closed form. So we can estimate its standard error via bootstrap.

- Generate \( B \) bootstrap samples of size \( n \) (sampling from \( P_x \) with replacement.)
- We estimate \( se_{P_x} (\hat{\theta}) \) by the standard derivative of \( \hat{\theta}^* \), the estimation from bootstrap sample:

\[
se_{P_x} (\hat{\theta}) \approx \sqrt{\frac{1}{B} \sum_{i=1}^{B} (\hat{\theta}_i^* - \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^*)^2}.
\]

(7.1)

- In theory, there are \( \binom{2n-1}{n} \) distinct bootstrap samples of size \( n \) (place \( n - 1 \) boards in between \( n \) balls). However, their probabilities are different. The probability of a bootstrap sample in which \( x_i \) appears \( k_i \) times, with \( k_1 > 0 \) and \( k_1 + k_2 + \cdots + k_n = n \) is

\[
\frac{n!}{k_1!k_2!\cdots k_n!}.
\]

The most probable sample is the one with \( k_i = 1 \) - the original one.

• Bias correction via bootstrap.
- The bias of \( \hat{\theta} \) is

\[
b_{X \sim P} (\hat{\theta}) = E_{X \sim P} (\hat{\theta}) - \theta.
\]

If it is known, then we can use

\[
\tilde{\theta} = \hat{\theta} - b_{X \sim P} (\hat{\theta})
\]

as a “corrected” estimate: \( E_{X \sim P} (\tilde{\theta}) = E_{X \sim P} (\hat{\theta}) - E_{X \sim P} (b_{X \sim P} (\hat{\theta})) = [\theta + b_{X \sim P} (\hat{\theta})] - b_{X \sim P} (\hat{\theta}) = \theta \).
- When \( b_{X \sim P} (\hat{\theta}) \) is unknown, we estimate it by

\[
b_{X \sim P_x} (\hat{\theta}) = E_{X \sim P_x} (\hat{\theta}) - \tilde{\theta},
\]

where \( E_{X \sim P_x} (\hat{\theta}) \) can be estimated by bootstrap \( \frac{1}{B} \sum_{i=1}^{B} \hat{\theta}_i^* \). That is, we can correct bias by

\[
\tilde{\theta} = 2\hat{\theta} - E_{X \sim P_x} (\hat{\theta}).
\]

• Parametric bootstrap
- Non-parametric bootstrap is to substitute \( P \) by \( P_x \).
- Parametric bootstrap assumes that the distribution \( P \) comes from a model \( \{P_\theta\}_{\theta \in \Theta} \), and substitutes \( P_\theta \) for \( P \). In Monte Carlo approximation, it means that we do not draw random samples from \( P_x \), but from \( P_\theta \) instead.
7.2 Delta method

• Suppose we have an asymptotic normality theorem for \( \hat{\theta} = \hat{\theta}_n \) (for example, CLT with \( \hat{\theta} = X \)):

\[
\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2),
\]

then we have

\[
\hat{\theta}_n \sim N(\theta, \frac{\sigma^2}{n}).
\]

where \( \sim \) is “approximately distributed as”. \( \sigma \) can be known or estimated, then we can use \( \sigma/\sqrt{n} \) as standard error of \( \hat{\theta}_n \).

• Sometimes we care about \( g(\theta) \) instead of \( \theta \) itself. Then we may estimate \( g(\theta) \) by \( g(\hat{\theta}) \) (MLE works for instance). If \( g \) is differentiable (which implies continuous) at \( \theta \) and \( g'(\theta) \neq 0 \), then

\[
\sqrt{n} \frac{g(\hat{\theta}_n) - g(\theta)}{\sigma |g'(\theta)|} \xrightarrow{d} N(0, 1),
\]

and then

\[
g(\hat{\theta}_n) \sim N \left( g(\theta), \frac{\sigma^2 |g'(\theta)|^2}{n} \right),
\]

which implies the standard error of \( g(\hat{\theta}_n) \) is \( \frac{\sigma |g'(\theta)|}{\sqrt{n}} \).

• \( \theta \) is unknown, so \( g'(\theta) \) is also unknown. If \( \hat{\theta}_n \) is consistent (\( \hat{\theta} \xrightarrow{p} \theta \)), and \( g' \) is continuous (at \( \theta \)), then we have (by Slutsky’s Theorem)

\[
\sqrt{n} \frac{g(\hat{\theta}_n) - g(\theta)}{\sigma |g'(\theta)|} \xrightarrow{d} N(0, 1),
\]

and then

\[
g(\hat{\theta}_n) \sim N \left( g(\theta), \frac{\sigma^2 |g'(\hat{\theta})|^2}{n} \right).
\]

8 Lecture 8: Confidence interval

8.1 Bayesian confidence/probability intervals

• Bayesian approach: everything is in posterior distribution.

• Percentile method.

  – Take two quantiles, \( q_\beta \) and \( q_{1-\gamma} \), set \( \beta, \gamma \) such that

\[
\Pr(q_\beta \leq \theta \leq q_{1-\gamma}) = 1 - \alpha.
\]

Usually, \( \beta = \gamma = \alpha/2 \).

  – HPD (highest posterior density). With posterior density \( f_{\theta|x}(u) \), find \( c \) such that the region is \( E = \{u | f_{\theta|x}(u) \geq c \} \), where

\[
\Pr(\theta \in E) = \int_{E} f_{\theta|x}(u) du = 1 - \alpha \quad \text{(or} \geq 1 - \alpha \text{if discrete}).
\]

It is the shortest interval if \( f_{\theta|x}(u) \) is unimodal.

8.2 General confidence intervals

• Main idea: find the distribution of the estimates.
Normal observations: unknown $\mu$, known $\sigma$.
Consider an example where $X_1, \cdots, X_n$ are i.i.d. r.v.s with $N(\mu, \sigma^2)$. We are estimating the mean $\mu$ by $\bar{X}$. We know that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$, so $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Then
\[
1 - \alpha = \Pr\left[\left|\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right| \leq z_{\alpha/2}\right] = \Pr\left[\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2} + \frac{\sigma}{\sqrt{n}} \theta\right],
\]
where $z_{\alpha/2}$ is the $\alpha/2$-quantile of standard normal $N(0,1)$.

Generally, if $\theta$ is (approximately) $N(\theta, (se(\theta))^2)$, then
\[
\Pr\left[\hat{\theta} - se(\hat{\theta})z_{\alpha/2} \leq \theta \leq \hat{\theta} + se(\hat{\theta})z_{\alpha/2}\right] = 1 - \alpha.
\]

- **Bootstrap confidence intervals** (normal case). If $se(\hat{\theta})$ is unknown ($\sigma$ is unknown), then we can estimate it via bootstrap (7.1). This works if $\theta$ is (approximately) normal.

- **Bootstrap “percentile” confidence intervals** (normal case). We can estimate the end points $\hat{\theta} \pm se(\hat{\theta})z_{\alpha/2}$ directly by bootstrap estimates $\hat{\theta}_{\alpha/2}^*, \hat{\theta}_{1-\alpha/2}^*$. Recall that we have $B$ bootstrap sample estimates $\hat{\theta}^*$. $\hat{\theta}_{\alpha/2}^*$ corresponds to the $\alpha/2$ quantile of these $B$ estimates.

- **Bootstrap pivotal confidence intervals**. We can estimate the $\alpha/2$ and $1-\alpha/2$ quantiles ($q_{\alpha/2}$ and $q_{1-\alpha/2}$) of $\theta - \hat{\theta}$, by $\hat{\theta}_{\alpha/2}^* - \hat{\theta}, \hat{\theta}_{1-\alpha/2}^* - \hat{\theta}$. Then
\[
1 - \alpha = \Pr[q_{\alpha/2} \leq \theta \leq q_{1-\alpha/2}] = \Pr[\hat{\theta} - q_{1-\alpha/2} \leq \theta \leq \hat{\theta} - q_{\alpha/2}] \approx \Pr[\hat{\theta} - (\hat{\theta}_{1-\alpha/2}^* - \hat{\theta}) \leq \theta \leq \hat{\theta} - (\hat{\theta}_{\alpha/2}^* - \hat{\theta})] = \Pr[2\hat{\theta} - \hat{\theta}_{1-\alpha/2}^* \leq \theta \leq 2\hat{\theta} - \hat{\theta}_{\alpha/2}^*]
\]

Normal observations: unknown $\mu$, unknown $\sigma$.
Consider an example where $X_1, \cdots, X_n$ are i.i.d. r.v.s with $N(\mu, \sigma^2)$. We are estimating the mean $\mu$ by $\bar{X}$. Let
\[
s^2 = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n}(X_i - \bar{X})^2}.
\]
We know that
\[
Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)
\]
\[
\chi^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)
\]
\[
t = \frac{Z}{\sqrt{\chi^2/(n-1)}} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \sqrt{n} \frac{\bar{X} - \mu}{s} \sim \frac{N(0,1)}{\sqrt{\chi^2/(n-1)}} = t(n-1),
\]
since $Z$ and $\chi^2$ are independent. Then
\[
1 - \alpha = \Pr\left[|t| \leq t_{\alpha/2}(n-1)\right] = \Pr\left[\frac{\bar{X} - s}{\sqrt{n}}t_{\alpha/2}(n-1) \leq \mu \leq \frac{\bar{X} + s}{\sqrt{n}}t_{\alpha/2}(n-1)\right],
\]
where $t_{\alpha/2}(n-1)$ is the $\alpha/2$-quantile of $t(n-1)$, $t$ distribution with $(n-1)$ degree of freedom.
9 Lecture 9: Hypothesis testing

9.1 Setup

- Null hypothesis set \( \mathcal{P}_0 \); alternative hypothesis set \( \mathcal{P}_A \). (\( \Theta_0 \) and \( \Theta_A \) if parametric).
- \( \mathcal{P}_0 \cap \mathcal{P}_0 = \emptyset \); \( \mathcal{P}_0 \cup \mathcal{P}_0 = \mathcal{P} \).
- Rejection region \( \mathcal{R} \subseteq \mathcal{X} \): if data \( \mathcal{X} \) falls into \( \mathcal{R} \), then reject null hypothesis; accept null hypothesis if \( \mathcal{X} \notin \mathcal{X} \setminus \mathcal{R} \).
- Errors

<table>
<thead>
<tr>
<th>Truth</th>
<th>Decision</th>
<th>Accept ( \mathcal{H}_0 )</th>
<th>Reject ( \mathcal{H}_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mathcal{H}_0 )</td>
<td>Correct</td>
<td>Type I Error</td>
<td></td>
</tr>
<tr>
<td>( \mathcal{H}_A )</td>
<td>Type II Error</td>
<td>Correct</td>
<td></td>
</tr>
</tbody>
</table>

9.2 Testing evaluation

- Power function, level, size
  - The power function is defined as
    \[
    \beta(P) = \Pr(X \in \mathcal{R}),
    \]
    where \( X \sim P \) and \( \mathcal{R} \) is calculated based on the testing method. Note that
    \[
    \beta(P) = \begin{cases} 
    \Pr(\text{Type I error}) & \text{if } P \in \mathcal{P}_0 \\
    1 - \Pr(\text{Type II error}) & \text{if } P \in \mathcal{P}_A 
    \end{cases}
    \]
    We say that a test is powerful if \( \beta(P) \) is “large” for \( P \in \mathcal{P}_A \).
  - Given \( 0 \leq \alpha \leq 1 \), a test is of (significance) level \( \alpha \) if \( \sup_{P \in \mathcal{P}_0} \beta(P) \leq \alpha \).
  - Given \( 0 \leq \alpha \leq 1 \), a test is of size \( \alpha \) if \( \sup_{P \in \mathcal{P}_0} \beta(P) = \alpha \).
- Most powerful test. A test at level \( \alpha \) that has higher or equal power than all other tests at level \( \alpha \) for all \( P \in \mathcal{P}_A \) is called uniformly most powerful at level \( \alpha \).
- Neyman-Pearson lemma. To test one simple hypothesis \( \mathcal{P}_0 \) against one simple alternative hypothesis \( \mathcal{P}_A \). Assuming they can be represented by density \( f_0(x) \), \( f_A(x) \), respectively. On the basis of observed \( x \), the (uniformly) most powerful test exists and is
  \[
  \text{reject } \mathcal{H}_0 \text{ if } \frac{f_A(x)}{f_0(x)} \geq c,
  \]
  where \( c \) is set so that
  \[
  \mathcal{P}_0 [x \in \mathcal{R}] = \mathcal{P}_0 \left[ \frac{f_A(x)}{f_0(x)} \geq c \right] = \alpha.
  \]
  - Randomized version:
    \[
    \text{reject } \mathcal{H}_0 \text{ if } \frac{f_A(x)}{f_0(x)} > c,
    \]
    \[
    \text{reject } \mathcal{H}_0 \text{ with probability } d \text{ if } \frac{f_A(x)}{f_0(x)} = c,
    \]
    \[
    \text{accept } \mathcal{H}_0 \text{ if } \frac{f_A(x)}{f_0(x)} < c,
    \]
    where \( c \) and \( d \in [0,1] \) are set so that
    \[
    \mathcal{P}_0 [x \in \mathcal{R}] = \mathcal{P}_0 \left[ \frac{f_A(x)}{f_0(x)} > c \right] + \mathcal{P}_0 \left[ \frac{f_A(x)}{f_0(x)} = c \right] \cdot d = \alpha.
    \]
9.3 \textit{p-value}

- Suppose we have nested rejection regions \( R_{\alpha_1} \subseteq R_{\alpha_2} \) whenever \( \alpha_1 \leq \alpha_2 \). Given the observed data \( x \), the observed significance level (or p-value) is defined as \( p(x) = \inf\{\alpha | x \in R_\alpha\} \).

10 Lecture 10: Multiple testing

Suppose we have \( K \) tests, \( k = 1, 2, \cdots, K \); testing \( H_{0k} : \mathcal{P}_{0k} \) against \( H_{Ak} : \mathcal{P}_{Ak} \) with rejection region \( R_k \).

10.1 Union-intersection test

- Testing \( H_0 : \mathcal{P}_0 = \bigcap_k \mathcal{P}_{0k} \) against \( H_A : \mathcal{P}_A = (\bigcap_k \mathcal{P}_{0k})^c = \bigcup_k \mathcal{P}_{Ak} \).
- Rejection region is \( R = \bigcup_k R_k \).
- Union bound: \( P_0(x \in R) = P_0(x \in \bigcup_k R_k) \leq \sum_k P_0(x \in R_k) \).

10.2 Intersection-union test

- Testing \( H_0 : \mathcal{P}_0 = \bigcup_k \mathcal{P}_{0k} \) against \( H_A : \mathcal{P}_A = (\bigcup_k \mathcal{P}_{0k})^c = \bigcap_k \mathcal{P}_{Ak} \).
- Rejection region is \( R = \bigcap_k R_k \).

10.3 Controlling family-wise error rate

\textbf{Family-wise error rate (FWER)}: the probability of committing \textit{at least} one error of the first kind. We want to bound it as

\[ \text{FWER} = P_0 \left( X \in \bigcup_{k=1}^K R_k \right) \leq \alpha. \]

- \textbf{Bonferroni method}: reject all null hypotheses whose \( p \)-value \( p_k \) is smaller than \( \alpha/K \).
  
  By union bound,

  \[ \text{FWER} = P_0(x \in R) = P_0 \left( X \in \bigcup_{k=1}^K R_k \right) \leq \sum_{k=1}^K P_0(x \in R_{k-1}) = \sum_{k=1}^K \frac{\alpha}{K} = \alpha. \]

- \textbf{Holm method}.
Order p-values as \( p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(K)} \).

- If \( \frac{\alpha}{K} \leq p_{(1)} \), then accept all null hypotheses and stop; otherwise reject \( H_{0(1)} \) and continue.
- If \( \frac{\alpha}{K-1} \leq p_{(2)} \), then accept all remaining null hypotheses and stop; otherwise reject \( H_{0(2)} \) and continue.
- \( \cdots \)
- If \( \frac{\alpha}{1} \leq p_{(K)} \), then accept \( H_{0(K)} \) and stop; otherwise reject \( H_{0(K)} \) and stop.

10.4 Controlling false discovery rate

Rejecting null hypothesis when it is true means “false discovery” (Type I error).

- **False discovery proportion (FDP)** is defined as
  \[
  \text{FDP} = \frac{\# \text{ of false discoveries}}{\# \text{ of all discoveries}},
  \]
  where the \# is counted from \( K \) tests.

- **False discovery rate (FDR)** is defined as the expectation of FDP, i.e.,
  \[
  \text{FDR} = E(\text{FDP}).
  \]
  We want to control FDR as \( \text{FDR} \leq \alpha \).

- **Benjamini and Hochberg method.**
  -Order p-values as \( p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(K)} \).
  -Let \( l_i = \frac{\alpha}{KC_K} \), where
    \[
    C_K = \begin{cases} 
    1 & \text{if tests are independent} \\
    \sum_{i=1}^{K} \frac{1}{i} & \text{otherwise}
    \end{cases}
    \]
  -Let \( r = \max\{i | p_{(i)} < l_i \} \).
  -Set \( t = p_{(r)} \) as the Benjamini-Hochberg rejection threshold. Reject all null hypotheses whose \( p_k \leq t \).

11 Lecture 11: Hypothesis testing, practical procedures

11.1 Wald test

- \( \hat{\theta} \) is an estimator of \( \theta \). To test \( H_0 : \theta = \theta_0 \), against the alternative \( H_A : \theta \neq \theta_0 \), \( \frac{\hat{\theta} - \theta_0}{se(\hat{\theta})} \) is a good indicator of discrepancy.

- Suppose \( \hat{\theta} \) is (approximately) normal:
  \[
  \sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2) \implies \hat{\theta} \sim N(\theta, (se(\hat{\theta}))^2) \implies \frac{\hat{\theta} - \theta}{se(\hat{\theta})} \sim N(0, 1).
  \]

- If \( se(\hat{\theta}) \) is unknown (because \( \theta \) is unknown), then we can estimate it by
  \[
  se(\hat{\theta}) \approx \hat{se}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})}, \text{ or } \sigma^2 \approx \hat{\sigma}^2.
  \]

We reject \( H_0 \) if \( \frac{\hat{\theta} - \theta_0}{\sqrt{\text{Var}(\hat{\theta})}} \) is too large or too small. Equivalently, reject \( H_0 \) if
\[
\frac{(\hat{\theta} - \theta_0)^2}{\text{Var}(\hat{\theta})} \sim \chi^2(1)
\]
is too large.
• In multidimensional case, with (approximately) normality,
\[ \sqrt{n} (\hat{\theta} - \theta) \sim N(0, V), \]
where \( V_{p \times p} \) is variance matrix. Then the Wald test becomes reject \( H_0 \) if
\[ n(\hat{\theta} - \theta_0)^T V^{-1}(\hat{\theta} - \theta_0) \sim \chi^2(p). \]
If \( V \) unknown, estimate it as \( \hat{V} = V(\hat{\theta}) \) or \( \hat{V} = V(\theta_0) \).

• If \( \hat{\theta} \) is an \textbf{MLE} of \( \theta \), then
\[ V(\theta) = I^{-1}(\theta), \]
where \( I(\theta) \) is the Fisher information matrix for \textit{ONE} observation.

11.2 \textbf{Likelihood ratio test}
Consider parametric model and its hypotheses \( H_0 : \theta \in \Theta_0 \) and \( H_A : \theta \in \Theta_A \).

• From Neyman-Pearson lemma, the optimal test is based on
\[ \frac{f_A(x)}{f_0(x)} = \frac{L(\theta_A)}{L(\theta_0)} \geq c. \]
Or equivalently
\[ \log L(\theta_A) - \log L(\theta_0) = l(\theta_A) - l(\theta_0) \geq c. \]

• To extend this to multiple hypotheses case, the \textbf{likelihood ratio test statistic} is defined as: reject \( H_0 \) if
\[ \sup_{\theta \in \Theta_A} L(\theta) \geq c. \]
Or alternatively, reject \( H_0 \) if
\[ \sup_{\theta \in \Theta_0} L(\theta) \geq c. \]

• Let \( \hat{\theta} \) and \( \hat{\theta}_0 \) be unconstrained and constrained MLE, respectively. Then the test in logarithm form: reject \( H_0 \) if
\[ 2(l(\hat{\theta}) - l(\hat{\theta}_0)) \geq c, \]
where the 2 is to ensure that the statistic has the approximate distribution \( \chi^2(p) \), where \( p \) is the number of restrictions imposed by the null hypothesis.

11.3 \textbf{Rao score test via Lagrange multipliers}
If the null hypothesis is interpreted as a restriction on parameters: \( H_0 : g(\theta) = 0 \), and the alternative is again \( H_A : g(\theta) \neq 0 \), then following the idea of Neyman-Pearson, we can check the magnitude of Lagrange multiplier as an indicator of how much the constraint is violated.

• Consider maximizing \( l(\theta) - \lambda g(\theta) \). Setting the derivative (in \( \theta \)) to zero, we have
\[ \lambda(\theta) = \frac{f'(\theta)}{g'(\theta)} = \frac{f'(x; \theta)}{f(x; \theta)} \frac{1}{g'(\theta)}. \]
We reject null if \( \left| \frac{\lambda(\theta)}{\lambda(\Theta)} \right| \) or \( \frac{\chi^2(\theta)}{\text{Var}(\lambda(\theta))} \) is too large.

• When \( g(\theta) = \theta - \theta_0 \), \( g'(\theta) = 1 \), \( \text{Var}(\lambda(\theta_0)) = nI(\theta_0) \). Then
\[ \frac{\lambda(\theta_0)}{\sqrt{nI(\theta_0)}} \sim N(0, 1), \frac{\lambda^2(\theta_0)}{nI(\theta_0)} \sim \chi^2(1). \]
Quantiles can be applied to find rejection region.
3 Score test for Binomial.

\[ X \sim Bin(n, p). \] To compute the score function and Fisher information:

\[ \lambda(p) = s(p) = \frac{n(\hat{p} - p)}{p(1 - p)} \] \[ nI(p) = \frac{n}{p(1 - p)}. \]

Therefore,

\[ Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1), \]

which is equivalent to Wald test with \( se(\hat{p}|p) \) estimated as \( se(\hat{p}|p_0) \) instead of \( se(\hat{p}|\hat{p}) \).

![Figure 1: Illustration of Wald, LRT and Rao tests.](image)

11.4 Bayes factor

To interpret Neyman-Pearson in Bayesian formula, consider averaging instead of maximization: reject \( H_0 \) if

\[ \frac{\int_{\Theta_A} L(\theta)\pi_A(\theta)d\theta}{\int_{\Theta_0} L(\theta)\pi_0(\theta)d\theta} \geq c, \]

where \( \pi_A \) and \( \pi_0 \) are priors over \( \Theta_A \) and \( \Theta_0 \).