STAT 512: Statistical Inference **Autumn 2022**

Quiz Session 9: Final Review

Yikun Zhang *December 7, 2022*

This note intends to give a brief review on lecture materials and highlight those important concepts/results in STAT 512. The review is by no means comprehensive and in order to excel at the final exam, a student is expected to master those fundamentals in the course instead of simply memorizing the key formulae or theorems.

Most parts of this note are selected from Professor Yen-Chi Chen's[1] and Professor Michael Perlman's lecture notes [Perlman, 2020].

1 Probability Distributions and Random Variables

Probability space: A probability space is written as $(\Omega, \mathcal{F}, \mathbb{P})$, where

1. Ω is the sample space;
2. $\mathcal{F}$ is a σ-algebra (also called σ-field);
3. $\mathbb{P}$ is a probability measure with $\mathbb{P}(\Omega) = 1$.

$\star$ Notes: You should be familiar with the definition of σ-algebra, properties of a probability measure (countable additivity, inclusion, complementation, monotone continuity, etc.).

Random variable: A *random variable* $X : \Omega \to \mathbb{R}$ is a (measurable) function satisfying

$$X^{-1}\left((-\infty, c]\right) := \{\omega \in \Omega : X(\omega) \leq c\} \in \mathcal{F} \quad \text{for all } c \in \mathbb{R}.$$

The probability that X takes on a value in a Borel set $B \subseteq \mathbb{R}$ is written as:

$$\mathbb{P}(X \in B) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in B\}\right).$$

Cumulative distribution function (CDF): The CDF $F : \mathbb{R} \to [0, 1]$ of a random variable X is defined as:

$$F(x) := \mathbb{P}(X \leq x) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \leq x\}\right).$$

Probability mass function (PMF) and probability density function (PDF):

- If the range $\mathcal{X} \subset \mathbb{R}$ of a random variable X is countable, it is called a *discrete* random variable, whose distribution can be characterized by the PMF as:

$$\mathbb{P}(X = x) = F(x) - \lim_{\epsilon \to 0^+} F(x - \epsilon) \quad \text{for all } x \in \mathcal{X}.$$

- If the range $\mathcal{X} \subseteq \mathbb{R}$ of a random variable X has an absolutely continuous CDF F, then we can describe its distribution through the PDF as:

$$p(x) = F'(x) = \frac{d}{dx} F(x).$$

  In this case, $F(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} p(u)\, du$.

[1] See http://faculty.washington.edu/yenchic/20A_stat512.html.

$\star$ Notes: You are expected to know the PMF or PDF of all the common distributions in Statistics; see Section 1.3 in Lecture 1 notes.

Conditional probability and distribution: For two events $A, B \in \mathcal{F}$, the conditional probability of A given B is given by

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)},$$

where the second equality follows from Bayes formula. Similarly, when both X and Y are continuous/discrete random variables, the conditional PDF/PMF of Y given $X = x$ is

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)} = \frac{p_{X|Y}(x|y) \cdot p_Y(y)}{p_X(x)},$$

where $p_X(x) = \int_{-\infty}^{\infty} p_{XY}(x,y)\, dy$ or $p_X(x) = \sum_y p_{XY}(x,y)$ is the marginal PDF or PMF of X.

Independence and conditional independence: Two events A and B are independent if

$$\mathbb{P}(A|B) = \mathbb{P}(A) \quad \text{or equivalently, } \mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

For three events A, B, C, we say that A and B are conditionally independent given C if

$$\mathbb{P}(A \cap B|C) = \mathbb{P}(A|C) \cdot \mathbb{P}(B|C).$$

The independence and conditional independence can be analogously defined for random variables X, Y, Z as:

- We say that X and Y are independent ($X \perp Y$) if

  $$F(x,y) = \mathbb{P}(X \le x, Y \le y) = \mathbb{P}(X \le x) \cdot \mathbb{P}(Y \le y).$$

  If X and Y have PDFs or PMFs, then the independence of X and Y can be equivalently defined as:

  $$p_{XY}(x,y) = p_X(x) \cdot p_Y(y),$$

  where p_X, p_Y are marginal PDFs or PMFs of X and Y.

- We say that X and Y are conditionally independent given Z (*i.e.*, $X \perp Y|Z$) if

  $$\mathbb{P}(X \le x, Y \le y|Z) = \mathbb{P}(X \le x|Z) \cdot \mathbb{P}(Y \le y|Z).$$

  Recall Theorem 1.1 and subsequent discussions in Lecture 1 notes for equivalently definitions and key properties of conditional independence.

2 Transforming continuous distributions

For a continuous random variable X with PDF $p_X(x)$ supported on $[a,b]$, the PDF of a transformed random variable $Y = f(X)$ by a strictly increasing function f is

$$p_Y(y) = \begin{cases} \frac{p_X\left(f^{-1}(y)\right)}{f'(f^{-1}(y))}, & f(a) \le y \le f(b), \\ 0, & \text{otherwise.} \end{cases}$$

For deriving the distribution $U = f(X,Y)$, which is a function of two (or more) random variables X, Y, one can start from its CDF as:

$$F_U(u) = \mathbb{P}\left(f(X,Y) \le u\right)$$

and determine the region $\{(X,Y) \in \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^2 : g(X,Y) \le u\}$. Or, one can introduce a second variable $V = h(X,Y)$, where the function h is chosen cleverly, so that it is relatively easy to find the joint distribution of (U,V) via the Jacobian method and then marginalize to find the distribution of U.

3 Expectation and Basic Asymptotic Theories

Expectation, variance, and covariance: For random variables X, Y, we define

- *expectation* (or mean): $\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot p_X(x)\,dx$ or $\sum_{x\in\mathcal{X}} x \cdot p_X(x)$.
- *variance*: $\mathrm{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))^2\right]$.
- *Covariance*: $\mathrm{Cov}(X,Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right]$.

⋆ Notes: You should be able to compute the expectations and variances of those common probability distributions in Statistics.

Moment generating function (MGF): The MGF of a random variable X is defined as:

$$M_X(t) = \mathbb{E}(e^{tX})$$

for some $t \in \mathbb{R}$. M_X may not exist for some or all $t \in \mathbb{R}$. When M_X exists in a neighborhood of 0, we have that

$$\mathbb{E}(X^j) = M_X^{(j)}(0) = \frac{d^j M_X(t)}{dt^j}\Big|_{t=0}.$$

For two random variables X, Y, if their MGFs exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then they have the same distributions; see Theorem 2.3.11 in Casella and Berger [2002]. For a sequence of random variables $X_i, i = 1, 2, ...$, if $\lim_{i\to\infty} M_{X_i}(t) = M_X(t)$ around a neighborhood of 0, then

$$\lim_{i\to\infty} F_{X_i}(x) = F_X(x)$$

for all x at which F_X is continuous; see Theorem 2.3.12 in Casella and Berger [2002].

The multivariate MGF for a random vector $X = (X_1, ..., X_d) \in \mathbb{R}^d$ is defined as:

$$M_X(t) = \mathbb{E}\left(e^{t^T X}\right)$$

with $t \in \mathbb{R}^d$. The MGF of a multivariate normal random vector $X \sim N_d(\mu, \Sigma)$ can be utilized to derive that

$$Z = AX + b \sim N_d\left(A\mu + b, A\Sigma A^T\right),$$

where $A \in \mathbb{R}^{d\times d}$ and $b \in \mathbb{R}^d$ are deterministic.

Convergence of random variables: We discuss four different convergences of a sequence $\{X_n\}_{n=1}^{\infty}$ of random variables:

- *Convergence in distribution*: $\lim\limits_{n\to\infty} F_n(x) = F(x)$, where the CDF of F is continuous at $x \in \mathbb{R}$ and $\{F_n\}_{n=1}^{\infty}$ are CDFs of $\{X_n\}_{n=1}^{\infty}$. We can write $X_n \xrightarrow{D} X$ or $X_n \rightsquigarrow X$.
- *Convergence in probability*: For any $\epsilon > 0$, $\lim\limits_{n\to\infty} \mathbb{P}\left(|X_n - X| > \epsilon\right) = 0$ and we can write $X_n \xrightarrow{P} X$.
- *Convergence in L^p-norm*: $\lim\limits_{n\to\infty} \mathbb{E}\left(|X_n - X|^p\right) = 0$, provided that the p-th absolute moments $\mathbb{E}|X_n|^p$ and $E|X|^p$ of $\{X_n\}_{n=1}^{\infty}$ and X exist.
- *Almost sure convergence*: $\mathbb{P}\left(\left\{\omega \in \Omega : \lim\limits_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$ and we can write $X_n \xrightarrow{a.s.} X$.

We prove the implications between the above convergences and provide counterexamples for which the converse directions do not hold in Quiz Session 3.

Markov's inequality: For a nonnegative random variables X, we have that

$$\mathbb{P}(X > \epsilon) \leq \frac{\mathbb{E}(X)}{\epsilon} \quad \text{for any } \epsilon > 0.$$

Chebyshev's inequality: For a random variable X with finite variance, we have that

$$\mathbb{P}\left(|X - \mathbb{E}(X)| > \epsilon\right) \leq \frac{\mathrm{Var}(X)}{\epsilon^2} \quad \text{for any } \epsilon > 0.$$

Weak Law of Large Numbers: Let $X_1, ..., X_n$ be independent and identically distributed (IID) random variables with $\mu = \mathbb{E}|X_1| < \infty$ and $\mathrm{Var}(X_1) < \infty$. The sample average converges in probability to μ, *i.e.*,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \stackrel{P}{\to} \mu.$$

The strong law of large number strengthens the convergence in probability to the almost sure convergence.

Central Limit Theorem: Let $X_1, ..., X_n$ be IID random variables with $\mu = \mathbb{E}|X_1| < \infty$ and $\sigma^2 = \mathrm{Var}(X_1) < \infty$. We also denote the sample average by $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then,

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \stackrel{D}{\to} Z,$$

where Z follows the standard normal distribution $N(0, 1)$.

$\star$ Notes: You should be familiar with the proofs of weak law of large numbers and central limit theorem.

Continuous mapping theorem: Let g be a continuous function and $\{X_n\}_{n=1}^{\infty}$ be a sequence of random variables.

- If $X_n \stackrel{D}{\to} X$, then $g(X_n) \stackrel{D}{\to} g(X)$;
- If $X_n \stackrel{P}{\to} X$, then $g(X_n) \stackrel{P}{\to} g(X)$;
- If $X_n \stackrel{a.s.}{\to} X$, then $g(X_n) \stackrel{a.s.}{\to} g(X)$.

Slutsky's theorem: Let $\{X_n\}_{n=1}^{\infty}$ and $\{Y_n\}_{n=1}^{\infty}$ be two sequences of random variables such that $X_n \stackrel{D}{\to} X$ and $Y_n \stackrel{P}{\to} c$, where X is a random variable and c is a constant. Then,

$$X_n + Y_n \stackrel{D}{\to} X + c, \quad X_nY_n \stackrel{D}{\to} cX, \quad \text{and} \quad \frac{X_n}{Y_n} \stackrel{D}{\to} \frac{X}{c} \text{ (when } c \neq 0).$$

Hoeffding's inequality: Let $X_1, ..., X_n \in [m, M]$ be IID random variables with $-\infty < m < M < \infty$ and $\bar{X}_n$ be their sample average. Then, for any $\epsilon > 0$,

$$\mathbb{P}\left(|\bar{X}_n - \mathbb{E}(\bar{X}_n)| \geq \epsilon\right) \leq 2\exp\left(-\frac{2n\epsilon^2}{(M-m)^2}\right).$$

It provides an improved concentration bound for $\bar{X}_n$ than the one derived from Chebyshev's inequality.

$\star$ Notes: You are encouraged to understand the proof and related examples about the concentration of mean in Lecture 3 notes.

4 Conditional Expectation

The conditional expectation of Y given X is the random variable $\mathbb{E}(Y|X)$ such that when $X = x$, its value is $\mathbb{E}(Y|X = x) = \int y \cdot p(y|x)\, dy$ or $\sum_y y \cdot p(y|x)$.

Law of total expectation: For any measurable function $g(x, y)$, we have that $\mathbb{E}\left[\mathbb{E}\left(g(X,Y)|X\right)\right] = \mathbb{E}\left[g(X,Y)\right]$. It gives rise to several applications:

- For any measurable functions $g(x), h(y)$, we have that $\mathbb{E}\left[g(X) \cdot h(Y)\right] = \mathbb{E}\left[g(X) \cdot \mathbb{E}\left(h(Y)|X\right)\right]$.
- For any measurable functions $g(x), h(y)$, we have that $\mathrm{Cov}\left(g(X), h(Y)\right) = \mathrm{Cov}\left(g(X), \mathbb{E}\left[h(Y)|X\right]\right)$.

Law of total variance: Given a random variable Y, we have that $\mathrm{Var}(Y) = \mathbb{E}\left[\mathrm{Var}(Y|X)\right] + \mathrm{Var}\left[\mathbb{E}(Y|X)\right]$.

$\star$ Notes: Both examples about missing data and survey sampling are instructive, and you are expected to fully understand them.

5 Correlation, Prediction, and Regression

Pearson's correlation coefficient: For two random variables X and Y, their (Pearson's) correlation coefficient is defined as:

$$\rho_{XY} = \mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}},$$

where $\rho_{XY} \in [-1, 1]$ by the Cauchy-Schwarz inequality; see Quiz Session 1 notes. It measures the *linear* relation between two random variables.

Mean-square error prediction: The regression function (or best predictor) $\mathbb{E}(Y|X = x) := m(x)$ of Y on X minimizes the mean square error $R(g) = \mathbb{E}\left[(Y - g(X))^2\right]$ among all possible functions for g.

$\star$ Notes: You should be able to derive those properties about the best predictor $\mathbb{E}(Y|X)$ and residual $Y - \mathbb{E}(Y|X)$.

Linear prediction: The linear regression function that minimizes the mean square error $R(\alpha, \beta) = \mathbb{E}\left[(Y - \alpha - \beta X)^2\right]$ is given by

$$\begin{aligned} m^*(x) &= \mathbb{E}(Y) + \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}\left[x - \mathbb{E}(X)\right] \\ &= \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X), \end{aligned}$$

where $\mu_X = \mathbb{E}(X), \mu_Y = \mathbb{E}(Y), \sigma_X^2 = \mathrm{Var}(X), \sigma_Y^2 = \mathrm{Var}(Y)$, and ρ_{XY} is the Pearson's correlation coefficient. In practice, these population quantities $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY}$ are estimated from a data sample $\{(X_1, Y_1), ..., (X_n, Y_n)\}$ as:

$$\widehat{\mu}_X = \frac{1}{n}\sum_{i=1}^n X_i := \bar{X}_n, \quad \widehat{\sigma}_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2, \quad \widehat{\mu}_Y = \frac{1}{n}\sum_{i=1}^n Y_i := \bar{Y}_n,$$

$$\widehat{\sigma}_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2, \quad \widehat{\rho}_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)^2}\sqrt{\sum_{i=1}^n (Y_i - \bar{Y}_n)^2}}.$$

$\star$ Notes: You should be familiar with the generalization of the above results for the univariate linear regression to the multivariate setting.

Classification: Our goal is to find a classifier that minimizes the risk $R(c) = \mathbb{E}\left[L(c(X), Y)\right]$ for a given loss function L. Under the 0-1 loss $L(u, v) = \mathbb{1}_{\{u \neq v\}}$, one can obtain the *Bayes classifier* as:

$$c_*(x) = \arg\max_{y \in \{0,1\}} \mathbb{P}(y|x) = \begin{cases} 0, & \text{if } \mathbb{P}(0|x) \geq \mathbb{P}(1|x), \\ 1, & \text{if } \mathbb{P}(1|x) > \mathbb{P}(0|x). \end{cases}$$

Note that the Bayes classifier only depends on the distribution of (X, Y) but not the class of classifiers (such as k-Nearest Neighbors, decision trees, etc.).

6 Estimators

The central topic of this section is to estimate the parameter (vector) $\theta \in \Theta \subset \mathbb{R}^k$ from IID data $X_1, ..., X_n$ that are sampled from the underlying (parametric) distribution $p(x; \theta)$.

Method of moment estimators: Let $m_j(\theta) = \mathbb{E}(X^j)$ for $j = 1, 2, ...$ Then, the method of moment estimator for $\theta = (\theta_1, ..., \theta_k)$ is obtained by solving the system of equations

$$\begin{cases} m_1(\theta) & = \frac{1}{n} \sum\limits_{i=1}^{n} X_i, \\ m_2(\theta) & = \frac{1}{n} \sum\limits_{i=1}^{n} X_i^2, \\ & \vdots \\ m_k(\theta) & = \frac{1}{n} \sum\limits_{i=1}^{n} X_i^k. \end{cases}$$

Maximum likelihood estimator (MLE): The MLE is defined as:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p(X_i; \theta) := \arg\max_{\theta \in \Theta} \ell_n(\theta),$$

where $\ell_n(\theta)$ is the log-likelihood function. Under the conditions of (d) in Theorem 7 in Quiz Session 1, the MLE solves the score equation, *i.e.*,

$$S_n(\widehat{\theta}_{MLE}) = 0,$$

where $S_n(\theta) = \sum\limits_{i=1}^{n} \frac{\partial}{\partial \theta} \log p(X_i; \theta)$. In addition, by the central limit theorem,

$$\sqrt{n} \left(\widehat{\theta}_{MLE} - \theta_0 \right) \xrightarrow{D} N_k \left(0, I(\theta_0)^{-1} \right),$$

where $I(\theta) = \mathbb{E}\left[\nabla_\theta \log p(X; \theta) \nabla_\theta \log p(X; \theta)^T\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta \partial \theta^T} \log p(X; \theta)\right]$ is the Fisher's information matrix.

Bayesian estimator: In the regime of Bayesian statistics, the parameter θ of interest is assumed to be generated from a *prior distribution* $\pi(\theta)$ with $\theta \in \Theta \subset \mathbb{R}^k$. The inference on θ is carried out through the *posterior distribution* defined by the Bayes formula as:

$$f(\theta | X_1, ..., X_n) = \frac{p(X_1, ..., X_n | \theta) \cdot \pi(\theta)}{p(X_1, ..., X_n)} \propto \underbrace{p(X_1, ..., X_n | \theta)}_{\text{likelihood}} \times \underbrace{\pi(\theta)}_{\text{prior}}.$$

The posterior distribution leads to (at least) two Bayesian estimators:

- *posterior mean*: $\widehat{\theta}_p = \mathbb{E}(\theta | X_1, ..., X_n) = \int \theta \cdot f(\theta | X_1, ..., X_n) \, d\theta$;

- *Maximum a posteriori (MAP)*: $\widehat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} f(\theta | X_1, ..., X_n)$.

Empirical risk minimization: Given a class of predictors $\mathcal{F}$, we seek to find the predictor $f^* \in \mathcal{F}$ that minimizes the risk function given a loss function L, *i.e.*,

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}\left[L(Y, f(X))\right].$$

Such predictor f^* has the best prediction performance among $\mathcal{F}$ under the loss function L. When the distribution of (X, Y) is unknown in practice, we pursue the estimator $\widehat{f} \in \mathcal{F}$ that minimizes the *empirical risk* function, *i.e.*,

$$\widehat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L\left(Y_i, f(X_i)\right).$$

7 Multinomial Distribution

The PMF of a multinomial random vector $X = (X_1, ..., X_k) \sim \text{Multinomial}(n; p_1, ..., p_k)$ is given by

$$\mathbb{P}(X_1 = x_1, ..., X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} \cdot p_1^{x_1} \cdots p_k^{x_k}.$$

Properties of the multinomial distribution:

- *Additional trials*: If $(X_1, ..., X_k) \sim \text{Multinomial}(n; p_1, ..., p_k)$ and $(Y_1, ..., Y_k) \sim \text{Multinomial}(m; p_1, ..., p_k)$ are independent, then

$$(X_1 + Y_1, ..., X_k + Y_k) \sim \text{Multinomial}(n + m; p_1, ..., p_k).$$

- *Combining cells*: If $(X_1, ..., X_4) \sim \text{Multinomial}(n; p_1, ..., p_4)$ and $Y_1 = X_1 + X_2, Y_2 = X_3 + X_4$, then

$$(Y_1, Y_2) \sim \text{Multinomial}(n; p_1 + p_2, p_3 + p_4).$$

- *Conditional distributions*: If $(X_1, ..., X_4) \sim \text{Multinomial}(n; p_1, ..., p_4)$ and $Y_1 = X_1 + X_2, Y_2 = X_3 + X_4$, then

$$(X_1, X_2) \perp (X_3, X_4) | (Y_1, Y_2)$$

  and

$$\begin{aligned}
(X_1, X_2) | X_1 + X_2 &\sim \text{Multinomial}\left(X_1 + X_2; \frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right), \\
(X_1, X_2) | X_3 + X_4 &\sim \text{Multinomial}\left(n - X_3 - X_4; \frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right), \\
(X_3, X_4) | X_3 + X_4 &\sim \text{Multinomial}\left(X_3 + X_4; \frac{p_3}{p_3 + p_4}, \frac{p_4}{p_3 + p_4}\right).
\end{aligned}$$

- *Covariance between cells*: If $(X_1, ..., X_k) \sim \text{Multinomial}(n; p_1, ..., p_k)$, then for $1 \leq i \neq j \leq k$,

$$X_i | X_j \sim \text{Binomial}\left(n - X_j, \frac{p_i}{1 - p_j}\right)$$

  so that $\text{Cov}(X_i, X_j) = -np_ip_j$.

Parameter estimation for a multinomial distribution: Given an observed random vector $X = (X_1, ..., X_k) \sim \text{Multinomial}(n; p_1, ..., p_k)$ with $\sum_{j=1}^{k} p_j = 1$, we derive the MLE of its parameter $(p_1, ..., p_k)$ using the Lagrangian multiplier:

- *Goal*: maximize the log-likelihood function $\ell_n(p_1, ..., p_k|X) = \sum_{j=1}^{k} X_j \log p_j + C_n$ under the constraint $\sum_{j=1}^{k} p_j = 1$, where $C_n = \log \frac{n!}{X_1! \cdots X_k!}$ is a quantity that is independent of $(p_1, ..., p_k)$ and $\sum_{j=1}^{k} X_k = n$.
- The *Lagrangian function* is defined as:

$$F(p_1, ..., p_k, \lambda) = \sum_{j=1}^{k} X_j \log p_j + \lambda \left(1 - \sum_{j=1}^{k} p_j\right).$$

  Differentiating this function with respect to $p_1, ..., p_k, \lambda$ and setting them to 0 yield that

$$\frac{\partial F}{\partial p_j} = \frac{X_j}{p_j} - \lambda = 0, j = 1, ..., k, \quad \frac{\partial F}{\partial \lambda} = 1 - \sum_{j=1}^{k} p_j = 0. \tag{1}$$

  Since the log-likelihood $\ell_n(p_1, ..., p_k|X)$ is concave and the parameter set $\left\{(p_1, ..., p_k) \in [0,1]^k : \sum_{j=1}^{k} p_j = 1\right\}$ is convex, we know that the solution to (1) is indeed the MLE, *i.e.*, $(\widehat{p}_{1,MLE}, ..., \widehat{p}_{k,MLE}) = \left(\frac{X_1}{n}, ..., \frac{X_k}{n}\right)$.

⋆ Notes: You are expected to fully understand the examples presented during the lectures.

Dirichlet distribution: The PDF of a Dirichlet distribution is

$$p(u_1, ..., u_k; \alpha_1, ..., \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} u_i^{\alpha_i - 1} \quad \text{with } \sum_{i=1}^{k} u_i = 1 \text{ and } u_i \geq 0,$$

where $B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k} \alpha\right)}$ and $\alpha_1, ..., \alpha_k \geq 0$. It is generally used as a prior distribution for the multinomial parameters $p_1, ..., p_k$, leading to the posterior distribution as:

$$\begin{aligned} f(p_1, ..., p_k|X) &\propto \frac{n!}{X_1! \cdots X_k!} \cdot p_1^{X_1} \cdots p_k^{X_k} \times \frac{1}{B(\alpha)} \cdot p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1} \\ &\propto p_1^{X_1 + \alpha_1 - 1} \cdots p_k^{X_k + \alpha_k - 1} \\ &\sim \text{Dirichlet}(X_1 + \alpha_1, ..., X_k + \alpha_k). \end{aligned}$$

The posterior mean estimator for $(p_1, ..., p_k)$ is

$$(\widehat{p}_{p,1}, ..., \widehat{p}_{p,k}) = \left(\frac{X_1 + \alpha_1}{\sum_{j=1}^{k} (X_j + \alpha_j)}, ..., \frac{X_k + \alpha_k}{\sum_{j=1}^{k} (X_j + \alpha_j)}\right),$$

and the MAP estimator for $(p_1, ..., p_k)$ is

$$(\widehat{p}_{MAP,1}, ..., \widehat{p}_{MAP,k}) = \left(\frac{X_1 + \alpha_1 - 1}{\sum_{j=1}^{k} (X_j + \alpha_j) - k}, ..., \frac{X_k + \alpha_k - 1}{\sum_{j=1}^{k} (X_j + \alpha_j) - k}\right).$$

⋆ Notes: You should be able to derive the MAP estimator for $(p_1, ..., p_k)$ using the Lagrangian multiplier.

8 Linear Models and the Multivariate Normal Distribution

Key concepts in linear algebra:

- *Matrix multiplication:* For two matrices $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times p}$, AB is a $m \times p$ matrix, whose (i,j)-entry is
$$[AB]_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$
for $1 \leq i \leq m$ and $1 \leq j \leq p$. In particular, for a vector $x \in \mathbb{R}^n$,
$$Ax = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} A_{1i} x_i \\ \sum_{i=1}^{n} A_{2i} x_i \\ \vdots \\ \sum_{i=1}^{n} A_{mi} x_i \end{pmatrix}.$$
The matrix multiplication on $\mathbb{R}^n$ is linear, *i.e.*, $A(ax+by) = aAx + bAy$ for any $x, y \in \mathbb{R}^n$ and $a, b \in \mathbb{R}$.

- *Spectral decomposition*: For a symmetric (square) matrix $A \in \mathbb{R}^{n\times n}$, *i.e.*, $A = A^T$, we can apply the spectral decomposition to it as:
$$A = U\Lambda U^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T,$$
where $U = [u_1, ..., u_n] \in \mathbb{R}^{n\times n}$ is an orthogonal matrix whose columns are eigenvectors of A.

- *Positive definite matrix*: A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is positive definite if $x^T Ax > 0$ for all $x \in \mathbb{R}^n$ with $x \neq 0$. It is positive semi-definite if $x^T Ax \geq 0$ for all $x \in \mathbb{R}^n$.

- *Inverse of a partitioned matrix and Schur complement:* If $A \in \mathbb{R}^{n\times n}$ is invertible (or nonsingular) and we partition A into blocks as:
$$A = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$
where $S_{ij} \in \mathbb{R}^{n_i \times n_j}$ with $i, j = 1, 2$ and $n = n_1 + n_2$, then the inverse of A can be calculated as:
$$A^{-1} = \begin{pmatrix} S_{11,2}^{-1} & -S_{11}^{-1} S_{12} S_{22,1} \\ -S_{22}^{-1} S_{21} S_{11,2}^{-1} & S_{22,1}^{-1} \end{pmatrix},$$
where $S_{11,2} = S_{11} - S_{12} S_{22}^{-1} S_{21}$ is called the Schur complement of S_{11} and $S_{22,1} = S_{22} - S_{21} S_{11}^{-1} S_{12}$ is called the Schur complement of S_{22}.

$\star$ Notes: You should be familiar with the rank, inverse, transpose, trace, determinant, eigenvalues, and eigenvector of a matrix. You are also expected to know the common types of matrices, such as identity, triangular, orthogonal, projection matrices, etc.

Jacobian method: Suppose that there is a smooth one-to-one (or bijective) mapping $T : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}^n$ with $y = T(x)$ for all $x \in \mathcal{X}$ (such mapping is also known as diffeomorphism). We define the Jacobian matrix as:
$$J_T(x) \equiv \left(\frac{\partial y}{\partial x}\right) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{n\times n},$$

and the Jacobian is $|\det(J_T(x))| = \left|\left(\frac{\partial y}{\partial x}\right)\right| = \left|\frac{\partial y}{\partial x}\right|$. Let $A, B \subset \mathbb{R}^n$ be two subsets such that $B = \{T(x) : x \in A\}$ and f be a real-valued integrable function on A. Then,

$$\int_A f(x)\,dx = \int_B f\left(T^{-1}(y)\right)\left|\frac{\partial x}{\partial y}\right| dy,$$

where $\left|\frac{\partial x}{\partial y}\right| = \left|\frac{\partial y}{\partial x}\right|^{-1}$. Assume that X is a random variable with its PDF p_X supported on A. Then, the PDF of $Y = T(X)$ is given by

$$p_Y(y) = p_X\left(T^{-1}(y)\right) \cdot \left|\frac{\partial x}{\partial y}\right| \cdot \mathbb{1}_B.$$

Covariance matrix: For a random vector $X \in \mathbb{R}^n$, its covariance matrix is defined as

$$\mathrm{Cov}(X) = \mathbb{E}\left[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^T\right] = \mathbb{E}\left(XX^T\right) - \mathbb{E}(X)\mathbb{E}(X)^T.$$

Given a deterministic matrix $A \in \mathbb{R}^{n\times n}$ and vector $b \in \mathbb{R}^n$, we have that $\mathrm{Cov}(AX + b) = A\mathrm{Cov}(X)A^T$.

Multivariate normal distribution: The PDF of a multivariate normal random vector $X \sim N_n(\mu, \Sigma)$ is given by

$$p(x) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right].$$

- *Linearity*: $Y = CX + b \sim N_m(A\mu + b, A\Sigma A^T)$ with $A \in \mathbb{R}^{m\times n}$ as a deterministic nonsingular matrix and $b \in \mathbb{R}^m$ as a deterministic vector, where $X \sim N_n(\mu, \Sigma)$.
- *Equivalence of independence and uncorrelation*: If X and Y are both multivariate normal random variables/vectors, then $X \perp Y \iff \mathrm{Cov}(X, Y) = 0$.
- *Normality of marginal and conditional distributions*: Given a multivariate normal random vector $X \sim N_n(\mu, \Sigma)$, we partition it into $X = (X_1, X_2)^T \in \mathbb{R}^n$, where $X_1 \in \mathbb{R}^{n_1}$ and $X_2 \in \mathbb{R}^{n_2}$ with $n = n_1 + n_2$. Then,

  $$X_1 \sim N_{n_1}(\mu_1, \Sigma_{11}), \quad X_2 \sim N_{n_1}(\mu_2, \Sigma_{22}), \quad \text{and} \quad X_1|X_2 \sim N_{n_1}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11,2}\right),$$

  where we partition μ and Σ as $\mu = (\mu_1, \mu_2)^T \in \mathbb{R}^n$ and $\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix} \in \mathbb{R}^{n\times n}$. Here, $\Sigma_{11,2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

$\star$ Notes: The properties about multivariate normal distributions are very important.

Chi-square distribution: If $Z_1, ..., Z_n$ are IID normal random variable $N(0, 1)$, then $W_n = \sum_{i=1}^n Z_i^2$ follows a χ^2-distribution with n degrees of freedom. We write $W_n \sim \chi^2_n$.

- If $X \sim N_n(\mu, \Sigma)$, then $(X-\mu)^T\Sigma^{-1}(X-\mu) \sim \chi^2_n$.
- Let $X \sim N_n(\mu, \boldsymbol{I}_n)$ and P be a projection matrix (*i.e.*, it is idempotent $P^2 = P$) with $\mathrm{rank}(P) = m < n$. Then, $(X-\mu)^T P(X-\mu) \sim \chi^2_m$.
- Given some IID normal random variables $X_1, ..., X_n \sim N(\mu, \sigma^2)$, we know that
  - $\bar{X}_n = \frac{1}{n}\sum\limits_{i=1}^n X_i$ and $S_n^2 = \frac{1}{n-1}\sum\limits_{i=1}^n (X_i - \bar{X}_n)^2$ are independent.
  - $\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and $\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$.

9 Order Statistics

Let $X_1, ..., X_n$ be IID random variables. The *order statistics* $X_{(1)} \leq \cdots \leq X_{(n)}$ are the ordered values of $X_1, ..., X_n$. The distribution (or PMF) of the order statistics when $X_1, ..., X_n$ are discrete random variables can be derived by enumerating all possible configurations of $X_1, ..., X_n$ that leads to $\{X_{(1)} = y_1, ..., X_{(n)} = y_n\}$.

Now, when $X_1, ..., X_n$ has PDF $p_X(x)$ and CDF $F_X(x)$,

- the PDF of $X_{(j)}$ is

$$p_{X_{(j)}}(y) = \frac{n!}{(n-j)!(j-1)!} \cdot F_X(y)^{j-1} \left[1 - F_X(y)\right]^{n-j} p_X(y);$$

- the joint PDF of $(X_{(j)}, X_{(k)})$ with $j < k$ is

$$p_{X_{(j)},X_{(k)}}(y,z) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!} \cdot F_X(y)^{j-1} \left[F_X(z) - F_X(y)\right]^{k-j-1} \left[1 - F_X(z)\right]^{n-k} p_X(y) \cdot p_X(z);$$

- the joint PDF of $\left(X_{(1)}, ..., X_{(n)}\right)$ is $p(y_1, ..., y_n) = n! \cdot p_X(y_1) \cdots p_X(y_n)$.

Order statistics of Uniform$[0,1]$**:** When $X_1, ..., X_n$ are IID uniform random variables on $[0,1]$, the j-th order statistic follows the Beta$(j, n-j+1)$ distribution.

10 Statistical Functional and Bootstrap

Empirical CDF: Given a random sample $\{X_1, ..., X_n\}$, the empirical CDF is defined as: $\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$. We know that for any fixed $x \in \mathbb{R}$,

$$\mathbb{E}\left[\widehat{F}_n(x)\right] = F(x), \quad \text{Var}(\widehat{F}_n(x)) = F(x)\left[1 - F(x)\right], \quad \widehat{F}_n(x) \xrightarrow{P} F(x),$$

and $\sqrt{n}\left(\widehat{F}_n(x) - F(x)\right) \xrightarrow{D} N\left(0, F(x)\left[1 - F(x)\right]\right)$.

Statistical functional[2]**:** When the functional T is smooth, the plug-in estimator $T(\widehat{F}_n)$ for the population statistical functional $T(F)$ is consistent, *i.e.*, $T(\widehat{F}_n) \xrightarrow{P} T(F)$.
$\star$ Notes: You should be familiar with those examples related to statistical functionals discussed in the lectures.

Delta Method: Let $\{X_n\}_{n=1}^{\infty}$ be a sequence of random vectors in $\mathbb{R}^k$ such that $\sqrt{n}(Y_n - \mu) \xrightarrow{D} N_k(0, \Sigma)$. If a function $f : \mathbb{R}^k \to \mathbb{R}$ is differentiable at $\mu \in \mathbb{R}^k$, then

$$\sqrt{n}\left[f(X_n) - f(\mu)\right] \xrightarrow{D} N_1\left(0, \nabla f(\mu)^T \Sigma \nabla f(\mu)\right).$$

Linear functional and influence function: Given a function $\omega : \mathbb{R}^k \to \mathbb{R}$, a linear functional can be written as $T_\omega(F) = \int \omega(x)\, dF(x)$, whose plug-in estimator is given by $T_\omega(\widehat{F}_n) = \sum_{i=1}^{n} \omega(X_i)$, where $X_1, ..., X_n \in \mathbb{R}^k$

[2] The interested student can refer to Professor Jon Wellner's note https://sites.stat.washington.edu/people/jaw/COURSES/580s/581/LECTNOTES/ch7.pdf for further studies.

are random observations from F. We define the influence function as $L_F(x) = \omega(x) - T_\omega(F)$. By the central limit theorem,

$$\sqrt{n}\left(T_\omega(\widehat{F}_n) - T_\omega(F)\right) \xrightarrow{D} N\left(0, \mathbb{V}_\omega(F)\right) \quad \text{with} \quad \mathbb{V}_\omega(F) = \int L_F^2(x)\, dF(x),$$

provided that $\int \omega(x)^2 dF(x) < \infty$.

Nonlinear functional: Given a point mass δ_x at point $x \in \mathbb{R}^k$, the influence function of a general statistical functional T_{target} is

$$L_F(x) = \lim_{\epsilon \to 0} \frac{T_{\text{target}}\left((1-\epsilon)F + \epsilon\delta_x\right) - T_{\text{target}}(F)}{\epsilon}.$$

Nonparametric bootstrap: Given a random sample $\mathcal{D} = \{X_1, ..., X_n\}$, we *sample with replacement* from $\mathcal{D}$ to obtain a bootstrap sample $\mathcal{D}^* = \{X_1^*, ..., X_n^*\}$. Such bootstrap process is generally repeated for B times to obtain B bootstrap samples $\mathcal{D}^{*(b)} = \left\{X_1^{*(b)}, ..., X_n^{*(b)}\right\}, b = 1, ..., B$. They can be utilized to quantify the variance $\text{Var}(S(\mathcal{D}))$ (or estimation error) of a statistic $S(\mathcal{D})$ that is constructed on the original sample $\mathcal{D}$ as:

$$\text{Var}\left(S(\mathcal{D})\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left[S(\mathcal{D}^{*(b)}) - \frac{1}{B}\sum_{b=1}^{B} S(\mathcal{D}^{*(b)})\right].$$

The bootstrap method is particularly useful when $\text{Var}(S(\mathcal{D}))$ has no analytical forms.

References

G. Casella and R. Berger. *Statistical Inference*. Duxbury advanced series. Thomson Learning, 2nd ed. edition, 2002.

M. Perlman. Probability and Mathematical Statistics I (STAT 512 Lecture Notes), 2020. URL https://sites.stat.washington.edu/people/mdperlma/STAT%20512%20MDP%20Notes.pdf.