# Note on Advanced Statistical Inference

Zepeng CHEN
The HK PolyU

*Date: February 25, 2023*

## 1 Common Families of Distributions

### 1.1 Exponential Family

**Definition 1.1 (Exponential Family (Casella and Berger, 2001, p. 111))**

*A family of pdfs or pmfs is called an exponential family if it can be expressed as*

$$f(x|\boldsymbol{\theta}) = h(x)c(\boldsymbol{\theta})exp\left(\sum_{i=1}^{k} w_i(\boldsymbol{\theta})t_i(x)\right).$$

**Definition 1.2 (Exponential Family (Farnia, 2023, Slide. 2))**

*Given a feature map $\phi : \mathcal{X} \to \mathbb{R}^m$ and an $m$-dimensional canonical parameter vector $\boldsymbol{\theta} \in \mathbb{R}^m$, an exponential family is defined as the set $\mathcal{P} = \{p_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \mathbb{R}^m\}$ where the density function $p_{\boldsymbol{\theta}}$ satisfies the following for a log-partition function $A : \mathbb{R}^m \to \mathbb{R}$:*

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = exp\left(\boldsymbol{\theta}^\top \phi(\mathbf{x}) - A(\boldsymbol{\theta})\right).$$

**Note on** *Many common families are exponential families. These include the continuous families—normal, gamma, and beta, and the discrete families—binomial, Poisson, and negative binomial. For example, define*[1]

$$h(x) = 1 \text{ for all } x;$$

$$c(\boldsymbol{\theta}) = c(\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\mu^2}{2\sigma^2}\right), \quad -\infty < \mu < \infty, \sigma > 0;$$

$$w_1(\mu, \sigma) = \frac{1}{\sigma^2}, \quad \sigma > 0; \quad w_2(\mu, \sigma) = \frac{\mu}{\sigma^2}, \sigma > 0;$$

$$t_1(x) = -x^2/2; \quad \text{and} \quad t_2(x) = x.$$

*Then*

$$f(x|\mu, \sigma^2) = h(x)c(\mu, \sigma)exp[w_1(\mu, \sigma)t_1(x) + w_2(\mu, \sigma)t_2(x)].$$

*With the help of indicator function, $f$ can be rewritten as*

$$h(x)c(\mu, \sigma)exp[w_1(\mu, \sigma)t_1(x) + w_2(\mu, \sigma)t_2(x)]I_{(-\infty,\infty)}(x).$$

[1] Sometimes we define parameters as $(\theta_1, \theta_2) = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$ (Wasserman, 2020, Lec. 12). Thus, we have sufficient statistics $(x, x^2)$.

**Lemma 1.1 (Log-partition Function (Farnia, 2023, Slide. 2))**

*The log-partition function $A : \mathbb{R}^m \to \mathbb{R}$ can be determined as:*

$$A(\boldsymbol{\theta}) = log\left(\sum_{\mathbf{x}\in\mathcal{X}} exp\left(\boldsymbol{\theta}^\top \phi(\mathbf{x})\right)\right).$$

**Proof** Because

$$\sum_{\mathbf{x}\in\mathcal{X}} p_{\boldsymbol{\theta}}(\mathbf{x}) = 1.$$

■

**Lemma 1.2**

*(i) The gradient of the log-partition function $A$ is the mean of random vector $\phi(\mathbf{x})$:*

$$\nabla A(\boldsymbol{\theta}) = \boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbb{E}_{X\sim p_{\boldsymbol{\theta}}}[\phi(\mathbf{x})].$$

*(ii) The Hessian of the log-partition function $A$ is the covariance matrix of random vector $\phi(\mathbf{x})$:*

$$H_A(\boldsymbol{\theta}) = \mathrm{Cov}_{X\sim p_{\boldsymbol{\theta}}}(\phi(\mathbf{x})).$$

**Proof**

(i) Because

$$\nabla A(\boldsymbol{\theta}) = \frac{\sum_{\mathbf{x}\in\mathcal{X}} e^{\boldsymbol{\theta}^\top\phi(\mathbf{x})}\phi(\mathbf{x})}{\sum_{\mathbf{x}\in\mathcal{X}} e^{\boldsymbol{\theta}^\top\phi(\mathbf{x})}} = \sum_{\mathbf{x}\in\mathcal{X}} \frac{e^{\boldsymbol{\theta}^\top\phi(\mathbf{x})}}{\sum_{\mathbf{x}'\in\mathcal{X}} e^{\boldsymbol{\theta}^\top\phi(\mathbf{x}')}}\phi(\mathbf{x}) = \sum_{\mathbf{x}\in\mathcal{X}} p_{\boldsymbol{\theta}}(\mathbf{x})\phi(\mathbf{x}).$$

(ii) Because (Wasserman, 2020, Lec. 12)

$$\frac{\partial^2 A(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} = \mathbb{E}[(\phi_i(\mathbf{x}) - \mathbb{E}[\phi_i(\mathbf{x})])(\phi_j(\mathbf{x}) - \mathbb{E}[\phi_j(\mathbf{x})])] = cov(\phi_i(\mathbf{x}), \phi_j(\mathbf{x})).$$

■

**Lemma 1.3**

*The log-partition function $A$ of an exponential family is a convex function.*

**Proof** From probability we know that a covaraince matrix is always positive semi-definte (PSD). Thus, the Hessian of $A$ is a PSD matrix, implying it is a convex function. ■

**Note on** *In other words, $\nabla A(\boldsymbol{\theta})$ is a monotone function of the canonical parameters $\boldsymbol{\theta}$, i.e.,*

$$\forall\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d : \quad (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)^\top \left(\boldsymbol{\mu}_{\boldsymbol{\theta}_2} - \boldsymbol{\mu}_{\boldsymbol{\theta}_1}\right) \geq 0.$$

*Moreover, under the assumption of invertible map, we have*

$$\boldsymbol{\theta} = (\nabla A)^{-1}(\boldsymbol{\mu}).$$

### 1.1.1 Gamma Distribution

**Definition 1.3 (Gamma Distribution)**

*An uncertain positive quantity $\theta$ has a gamma$(a, b)$ distribution if*

$$p(\theta) = \text{dgamma}(\theta, a, b) = \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}, \quad \textit{for } \theta, a, b > 0$$

$$\begin{aligned} E[\theta] &= \frac{a}{b} \\ Var[\theta] &= \frac{a}{b^2} \\ \text{mode}[\theta] &= \begin{cases} (a-1)/b & \textit{if } a > 1 \\ 0 & \textit{if } a \leq 1 \end{cases} \end{aligned} \tag{1}$$

From the Gamma Distribtion's density

$$1 = \int_0^\infty \frac{b^a}{\Gamma(a)}\theta^{a-1}e^{-b\theta}d\theta \quad \text{for any values } a, b > 0$$

We can obtain

$$\int_0^\infty \theta^{a-1}e^{-b\theta}d\theta = \frac{\Gamma(a)}{b^a} \text{ for any values } a, b > 0$$

### 1.1.2 Beta Distribution

**Definition 1.4 (Beta distribution)**

*An uncertain quantity $\theta$, known to be between 0 and 1, has a beta$(a, b)$ distribution if*[a]

$$p(\theta) = \text{dbeta}(\theta, a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \quad \textit{for } 0 \leq \theta \leq 1$$

*And beta distribution has follow properties:*

$$\begin{aligned} \textit{mode}[\theta] &= \frac{a-1}{(a-1)+(b-1)} \quad \textit{if } a > 1 \qquad\qquad b > 1; \\ E[\theta] &= \frac{a}{a+b} \\ Var[\theta] &= \frac{ab}{(a+b+1)(a+b)^2} = \frac{E[\theta]E[1-\theta]}{a+b+1} \\ \int_0^1 \theta^{a-1}(1-\theta)^{b-1}d\theta &= \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \end{aligned} \tag{2}$$

[a] Note that $\Gamma(x+1) = x!$ if x is a positive integer, and $\Gamma(1) = 1$.

### 1.1.3 chi-squared

## 1.2 Location-scale Family

# 2 Transformation

# 3 Point Estimation

## 3.1 Maximum Likelihood Method

**Definition 3.1 (Maximum Likelihood Estimator)**

*Given a parameterized family of distributions* $\{p_{\boldsymbol{\theta}} : \theta \in \mathbb{R}^d\}$*, the maximum liklihood estimator (MLE) of the model parameters from observed samples* $\mathbf{x}_1, \cdots, \mathbf{x}_n$ *will be*

$$\begin{aligned} \boldsymbol{\theta}^{MLE} &:= \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{x}_i) \\ &\iff \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \sum_{i=1}^{n} log p_{\boldsymbol{\theta}}(\mathbf{x}_i) \quad \textit{(log is monotontic.)} \end{aligned}$$

**Note on Product** *The basic idea of MLE is that we want to find a good estimate of the unknown parameter* $\boldsymbol{\theta}$ *which maximizes the probability or the likelihood of getting the data we observed, i.e.,*

$$\max_{\boldsymbol{\theta}} P(X_1 = x_1, \cdots, X_n = x_n) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(\mathbf{x}_i).$$

**Definition 3.2 (MLE for Exponential Family)**

*Given a exponential family of distributions* $\{p_{\boldsymbol{\theta}} : \theta \in \mathbb{R}^d\}$ *with canonical parameters* $\boldsymbol{\theta}$ *and log-partition function* $A(\boldsymbol{\theta})$*, the maximum liklihood estimator (MLE) of the model parameters from observed samples* $\mathbf{x}_1, \cdots, \mathbf{x}_n$ *will be*

$$\begin{aligned} \boldsymbol{\theta}^{MLE} &:= \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} (\frac{1}{n}\sum_{i=1}^{n} \phi(\mathbf{x}_i))^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) \\ &= \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \hat{\boldsymbol{\mu}}^\top \boldsymbol{\theta} - A(\boldsymbol{\theta}) \qquad \textit{(Let } \hat{\boldsymbol{\mu}} \textit{ denote the empirical mean)} \end{aligned}$$

**Lemma 3.1**

*The maximum likelihood problem for fitting canonical parameters of an exponential family is a convex optimization problem.*

**Proof** Obviously the objective function regarding $\boldsymbol{\theta}$ is concave. ■

**Corollary 3.1**

*Since the maximum likelihood problem for fitting canonical parameters of an exponential*

*family is a convex optimization problem, by the FOC, we have*

$$\boldsymbol{\theta}^{MLE} = (\nabla A)^{-1}(\hat{\boldsymbol{\mu}}).$$

*In addition, the mean parameter $\boldsymbol{\mu}_{\boldsymbol{\theta}^{MLE}}$ under the maximum likelihood estimator match the empirical mean $\hat{\boldsymbol{\mu}}$:*

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{\theta}^{MLE}} =& \nabla A(\boldsymbol{\theta}^{MLE}) \\ =& \hat{\boldsymbol{\mu}}\end{aligned}$$

**Note on** *For example,*

*Interestingly, this problem is not jointly convex in mean and variance. Though we can derive the optimal solution via sequential optimization. The key here is that the maximum over $\mu$ does not depend on $\sigma$, and for this maximum over $\mu$, there is again a unique optimal $\sigma$ (Bazzi, 2018; Kanti, 2018).*

**Theorem 3.1 (Central Limit Theorem for Canonical parameter)**

*Consider a sequence of independent random vectors $(\mathbf{x}_i)_{i=1}^{\infty}$ distributed as $p_{\boldsymbol{\theta}}$. Then, for the Maximum Liklikelihood canonical parameter $\boldsymbol{\theta}_n^{MLE}$ from $n$ samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$, the following holds*

$$\sqrt{n}\left(\boldsymbol{\theta}_n^{MLE} - \boldsymbol{\theta}^*\right) \xrightarrow{dist} \mathcal{N}\left(\mathbf{0}, \mathrm{Cov}_{\theta^*}^{-1}(\phi(\mathbf{x}))\right).$$

## 3.2 Method of Moments

**Definition 3.3 (Method of Moments Estimator)**

*Given a parameterized family of distributions $\{p_{\boldsymbol{\theta}} : \theta \in \mathbb{R}^d\}$, the method of moments estimator $\hat{\boldsymbol{\theta}}$ of the model parameters from observed samples $\mathbf{x}_1, \cdots, \mathbf{x}_n$ matches the empirical mean vector, i.e., $\hat{\boldsymbol{\theta}}$ satisfies*

$$\mathbb{E}_{\hat{\boldsymbol{\theta}}}[\phi(\mathbf{x})] = \frac{1}{n}\sum_{i=1}^{n}\phi(\mathbf{x}_i).$$

## 3.3 Connections

### 3.3.1 Method of Moments and MLE

**Proposition 3.1 (Equivalence of Method of Moments and MLE)**

*Given a parameterized family of distributions $\{p_{\boldsymbol{\theta}} : \theta \in \mathbb{R}^d\}$ with feature function $\phi$, the method of moments estimator with $\phi$-based moments results in the same estimator as maximum likelihood estimator.*

**Proof** Note that $\boldsymbol{\mu}_{\boldsymbol{\theta}^{\mathrm{MLE}}} = \hat{\boldsymbol{\mu}}$ by Corollary 3.1, and this coincides with the definition of the method of moments estimator. ■

# 4 Large-Sample Theory (Keener, 2010, Ch. 8)

This section focus on the behavior of certain sample as the sample size approaches infinity. Although the notion of infinity is unreachable in reality, it can provide us with some useful approximations for the finite-sample case.

## 4.1 Convergence in Probability and Weak Law of Large Numbers

**Definition 4.1**

*A sequence of random variables $Y_n$ converges in probability to a random variable $Y$ as $n \to \infty$, written $Y_n \xrightarrow{p} Y$, if for every $\epsilon > 0$,*

$$P\left(|Y_n - Y| \geq \epsilon\right) \to 0$$

*as $n \to \infty$.*

**Proposition 4.1**

*If $E(Y_n - Y)^2 \to 0$ as $n \to \infty$, then $Y_n \xrightarrow{p} Y$.*

**Proof** By Chebyshev's inequality, for any $\epsilon > 0$,

$$P\left(|Y_n - Y| \geq \epsilon\right) \leq \frac{E\left(Y_n - Y\right)^2}{\epsilon^2} \to 0.$$

■

**Theorem 4.1 (Weak law of large numbers)**

*Suppose $X_1, X_2, \ldots$ are i.i.d. with common mean $\mu$ and variance $\sigma^2$, and let $\bar{X}_n = \left(X_1 + \cdots + X_n\right)/n$, then $\bar{X}_n \xrightarrow{p} \mu$ as $n \to \infty$.*

**Proof** This theorem can be proved by Proposition 4.1,

$$E\left(\bar{X}_n - \mu\right)^2 = \mathrm{Var}\left(\bar{X}_n\right) = \sigma^2/n \to 0.$$

■

**Proposition 4.2**

*If $f$ is continuous at $c$ and if $Y_n \xrightarrow{p} c$, then $f\left(Y_n\right) \xrightarrow{p} f(c)$.*

**Note on** *That is, if the sequence $X_1, X_2, \ldots$ converges in probability to a random variable $X$ or to a constant $a$, if $h$ is continuous, we can make conclusions about the sequence of random variables $h(X_1), h(X_2), \ldots$ too.*

**Proof** Continuity means that given any $\epsilon > 0$, there exists $\delta_\epsilon > 0$ such that $|f(y) - f(c)| < \epsilon$ whenever $|y - c| < \delta_\epsilon$. Thus,

$$P\left(|Y_n - c| < \delta_\epsilon\right) \leq P\left(|f\left(Y_n\right) - f(c)| < \epsilon\right),$$

which implies

$$P\left(|f\left(Y_n\right) - f(c)| \geq \epsilon\right) \leq P\left(|Y_n - c| \geq \delta_\epsilon\right) \to 0.$$

■

**Definition 4.2 (Consistency)**

*A sequence of estimators $\delta_n$, $n \geq 1$, is consistent for $g(\theta)$ if for any $\theta \in \Omega$,*

$$\delta_n \xrightarrow{P_\theta} g(\theta)$$

*as $n \to \infty$, where $P_\theta$ is the underlying probability measure.*

## 4.2 Almost Sure Convergence and Strong Law of Large Numbers

**Definition 4.3**

*Random variables $Y_1, Y_2, \ldots$ defined on a common probability space converge almost surely to a random variable $Y$ on the same space if*

$$P\left(Y_n \to Y\right) = 1 \quad \text{or} \quad P\left(\lim_{n\to\infty} |Y_n - Y| < \epsilon\right) = 1.$$

**Note on Difference** *This type of convergence is stronger than convergence in probability, i.e., convergence almost surely implies convergence in probability. For example (Casella and Berger, 2001, p. 234), consider the sample space $S$ of the closed interval $[0, 1]$ with the uniform probability distribution. Define:*

$$X_1(s) = s + I_{[0,1]}(s), \quad X_2(s) = s + I_{[0,\frac{1}{2}]}(s), \quad X_3(s) = s + I_{[\frac{1}{2},1]}(s),$$

$$X_4(s) = s + I_{[0,\frac{1}{3}]}(s), \quad X_5(s) = s + I_{[\frac{1}{3},\frac{2}{3}]}(s), \quad X_6(s) = s + I_{[\frac{2}{3},1]}(s),$$

*etc. Let $X(s) = s$, and it is straightforward to see that $X_n$ converges to $X$ in probability. However, $X_n$ does not converge to $X$ almost surely. For every $s$, the value $X_n(s)$ alternates between the values $s$ and $s+1$ infinitely often. No pointwise convergence occus for this sequence.*

**Note on Example** *Consider the sample space $S$ of the closed interval $[0, 1]$ with the uniform probability distribution (Casella and Berger, 2001, p. 234). Define r.v. $X_n(s) = s + s^n$ and $X(s) = s$. For every $s \in [0, 1)$, $s^n \to 0$ as $n \to \infty$ and $X_n(s) \to s = X(s)$. However, $X_n(1)$ does not converge to $1 = X(1)$. But since the convergence occurs on the set $[0, 1)$ and $P([0, 1)) = 1$, $X_n$ converges to $X$ almost surely.*

**Theorem 4.2 (Strong Law of Large Numbers)**

*If $X_1, X_2, \ldots$ are i.i.d. with finite mean $\mu = EX_i$, and if $\bar{X}_n = (X_1 + \cdots + X_n)/n$, then $\bar{X}_n \to \mu$ almost surely as $n \to \infty$.*

**Note on Assumption** *Actually, both the weak and strong laws hold without the assumption of a finite variance. The only moment condition needed is that $E|X_i| < \infty$.*

## 4.3 Central Limit Theorem

## 4.4 Convergence in Distribution

**Definition 4.4 ($Y_n \Rightarrow Y$ or $Y_n \Rightarrow P_Y$)**

*A sequence of random variables $Y_n$, $n \geq 1$, with cdf $H_n$, converges in distribution (or law) to a random variable $Y$ witch cdf $H$ if*

$$H_n(y) \to H(y)$$

*as $n \to \infty$ whenever $H$ is continuous at $y$.*

**Note on Pointwise convergence at continuity points** *Note that pointwise convergence of the cdf only has to hold at continuity points of $H$. For example, suppose $Y_n = 1/n$, a degenerate r.v., and that $Y$ is always zero. Then*

$$H_n(y) = P\left(Y_n \leq y\right) = I\{1/n \leq y\}.$$

*If $y > 0$, $H_n(y) \to 1$ as $n \to \infty$, and if $y \leq 0$, then $H_n(y) = 0$. And $H_n(y) \to H(y)$ if $y \neq 0$. In this example, $Y_n \Rightarrow Y$, but the cdf $H_n(y)$ do not converge to $H(y)$ when $y = 0$, a discontinuity point of $H$.*

**Note on Difference** *This type of convergence is weaker than other types of convergence, i.e., convergence in distribution is implied by convergence in probability and almost sure convergence. And we have Theorem 4.3 and Theorem 4.4 to summarize their connections.*

**Note on Example** ***Maximum of uniforms (Casella and Berger, 2001, p. 235).*** *Suppose $X_1, \ldots$ are i.i.d. uniform (0,1) and $X_{(n)} = \max_{1 \leq i \leq n} X_i$. Then $X_{(n)}$ converges to 1 in probablity and $n(1 - X_{(n)})$ converges in distribution to an exponential (1) r.v..*

$$\begin{aligned} P\left(\left|X_{(n)} - 1\right| \geq \varepsilon\right) &= P\left(X_{(n)} \geq 1 + \varepsilon\right) + P\left(X_{(n)} \leq 1 - \varepsilon\right) \\ &= 0 + P\left(X_{(n)} \leq 1 - \varepsilon\right) \qquad = (1 - \varepsilon)^n \to 0 \quad \textit{(i.i.d.)} \end{aligned}$$

*And if we take $\varepsilon = t/n$, then we have*

$$P\left(n\left(1 - X_{(n)}\right) \leq t\right) \to 1 - e^{-t}.$$

**Theorem 4.3 ((Casella and Berger, 2001, p. 236))**

*If the sequence of r.v., $X_1, \ldots$ converges in probability to a random variable $X$, the sequence also converges in distribution to $X$.*

**Theorem 4.4 ((Casella and Berger, 2001, p. 236))**

*The sequence of r.v., $X_1, \ldots$ converges in probability to a constant $\mu$ iff the sequence also converges in distribution to $\mu$. That is, the statement*

$$P\left(|X_n - \mu| > \varepsilon\right) \to 0 \textit{ for every } \varepsilon > 0$$

*is equivalent to*

$$P(X_n \leq x) \to \begin{cases} 0 & \text{if } x < \mu \\ 1 & \text{if } x > \mu \end{cases}.$$

**Theorem 4.5**

*Convergence in distribution, $Y_n \Rightarrow Y$, holds iff $Ef(Y_n) \to Ef(Y)$ for all bounded continuous functions $f$.*

**Corollary 4.1**

*If $g$ is a continuous function and $Y_n \Rightarrow Y$, then*

$$g(Y_n) \Rightarrow g(Y).$$

## 4.5 Central Limit Theorem

**Theorem 4.6 (Central Limit Theorem)**

*Suppose $X_1, X_2, \ldots$ are i.i.d. with common mean $\mu$ and variance $\sigma^2$. Take $\bar{X}_n = (X_1 + \cdots + X_n)/n$. Then*

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \Rightarrow N\left(0, \sigma^2\right).$$

**Note on** *The importance of this theorem is that the assumption of finite variances leads to convergence to normality. However, it does not show how good the approximation is in general.*

**Note on** *The central limit theorem stated only provides direct information about distributions of averages. To discuss variables that are smooth functions of an average, i.e., $f(\bar{X}_n)$, the Taylor approximation motivates Proposition 4.3.*

**Theorem 4.7 (Slutsky's Theorem (Casella and Berger, 2001, p. 239))**

*If $X_n \to X$ in distribution and $Y_n \to a$, a constant, in probability, then*

- *$Y_n X_n \to aX$ in distribution.*
- *$X_n + Y_n \to X + a$ in distribution.*

**Theorem 4.8**

*If $Y_n \Rightarrow Y$, $A_n \xrightarrow{p} a$, and $B_n \xrightarrow{p} b$, then*

$$A_n + B_n Y_n \Rightarrow a + bY.$$

**Note on** *This theorem combines convergence in distribution with convergence in probability.*

## 4.6 The Delta Method

The previous section gives conditions under which a standardized random variable has a limit normal distribution. However, sometimes we are more interested in the distribution of some

function of the random variable rather than the random variable itself.

**Proposition 4.3 (Delta Method)**

*With the assumptions in the central limit theorem, if $f$ is differentiable at $\mu$, then*

$$\sqrt{n}\left(f\left(\bar{X}_n\right) - f(\mu)\right) \Rightarrow N\left(0, \left[f'(\mu)\right]^2 \sigma^2\right).$$

**Note on** *This use of Taylor's theorem to approximate distributions is called the delta method, and for the statistical application of Taylor's theorem, we are most concerned with the first-order Taylor series.*

**Note on** *This also means that if we use $f(\bar{X}_n)$ as an estimator of $f(\mu)$, we can say approximately,*

$$\mathrm{E}_\mu f(\bar{X}_n) \approx f(\mu)$$

$$\mathrm{Var}_\mu\, f(\bar{X}_n) \approx \left[f'(\mu)\right]^2 \sigma^2$$

**Note on Second-order Delta Method** *One concern is the possibility that $f'(\mu) = 0$, and this leads to the Second-order Delta Method (Casella and Berger, 2001, p. 244).*

# Bibliography


Bazzi, Ahmad (July 2018). *Answer to "Prove Neg. Log Likelihood for Gaussian Distribution Is Convex in Mean and Variance."*

Casella, George and Roger L. Berger (June 2001). *Statistical Inference*. 2nd edition. Australia ; Pacific Grove, CA: Cengage Learning. ISBN: 978-0-534-24312-8.

Farnia, Farzan (2023). *CSCI 5030 Machine Learning Theory*.

Kanti, John (Dec. 2018). *Proof That If All Vertices Have Degree at Least Two Then $G$ Contains a Cycle*. Forum Post.

Keener, Robert W. (Sept. 2010). *Theoretical Statistics: Topics for a Core Course*. 2010th edition. New York: Springer. ISBN: 978-0-387-93838-7.

Wasserman, Larry (2020). *36-705 Intermediate Statistics*.