Statistics Cheatsheet

Hongyi Guo*

December 7, 2020

1 Probabilty & Distribution

Definition 1.1 (Population)**.** The entire group of individuals about which we want to learn something about.

Definition 1.2 (Sample)**.** Subset of the population from which the information is actually obtained.

Definition 1.3 (Statistics)**.** A numerical characteristic of the sample, a random variable.

Remark 1.4. Remarks on the difference between samples and population.

(a) Population mean $\mu = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(b) Population variance $\sigma^2 = \frac{1}{n}\sum_{i=1}(X_i - \mu)^2$.

(c) Sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(d) Sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Definition 1.5 (SD, SE)**.** SD is the standard deviation, while SE is the standard deviation of the sampling distribution, e.g., $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, $SE(\bar{X}) = \frac{\widehat{\sigma}}{\sqrt{n}} = \frac{s}{\sqrt{n}}$.

Definition 1.6 (Cdf, pmf, pdf)**.** Let X be an r.v., then its *cumulative distribution function* (cdf) is defined by $F_X(x)$, where

$$F_X(x) = \mathbb{P}_X((-\infty, x]) = \mathbb{P}(X \leq x).$$

For a discrete r.v. X, the *probability mass function* (pmf) of X is given by $p_X(x) = \mathbb{P}[X = x]$, for $x \in \mathcal{D}$. For a continuous r.v. X, the *probability density function* is given by $f_X(s) = \frac{\mathrm{d}}{\mathrm{d}x}F_X(s)$.

Definition 1.7 (Support)**.** The support of a r.v. X is the points x in the space of X that $p_X(x) > 0$ (discrete r.v.) or $f_X(x) > 0$ (continuous r.v.).

Remark 1.8. The cdf, pmf/pdf, mle of various distributions.

*Northwestern University; hongyiguo2025@u.northwestern.edu

Name	Support	pmf/pdf	Mean	Variance	MLE	Fisher
Bernouli	$(0,1)$	$f(x;p)=\begin{cases}p: & x=1\\ 1-p: & x=0\end{cases}$	p	$p(1-p)$	$\widehat{p}=\bar{x}$	$\frac{1}{p(1-p)}$
Uniform	$[a,b]$	$f(x;p)=\begin{cases}\frac{1}{b-a} & a\le x\le b\\ 0: & o.w.\end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\widehat{b}=\min\{x_1,x_2,\dots\}$ $\widehat{a}=\max\{x_1,x_2,\dots\}$	-
Normal	$(-\infty,\infty)$	$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$	μ	σ	$\widehat{\mu}=\bar{x}$ $\widehat{\sigma}=\sum_{i=1}^{n}(x_i-\bar{x})^2$	$\begin{bmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{2}{\sigma^2}\end{bmatrix}$

2 Finding Point Estimators and Test Statistics

2.1 Point estimator

Definition 2.1 (Point estimates)**.** Based on the data sample, come up with the best single guess $\widehat{\theta}$ for the unknown true parameter θ.

Definition 2.2 (Bias, Var)**.** As follows,

$$\text{Bias}(\widehat{\theta})=\mathbb{E}[\widehat{\theta}]-\theta,\quad \text{Var}(\widehat{\theta})=\mathbb{E}[\widehat{\theta}-\mathbb{E}[\widehat{\theta}]]^2.$$

Definition 2.3 (Mean square error)**.**

$$\text{MSE}(\widehat{\theta})\equiv\mathbb{E}[\widehat{\theta}-\theta]^2.$$

Remark 2.4 (Bias variance trade-off)**.**

$$\text{MSE}(\widehat{\theta})\equiv\mathbb{E}[(\widehat{\theta}-\theta)^2]=\mathbb{E}\left[(\widehat{\theta}-\mathbb{E}[\widehat{\theta}]+\mathbb{E}[\widehat{\theta}]-\theta)^2\right]=\underbrace{\mathbb{E}[(\widehat{\theta}-\mathbb{E}[\widehat{\theta}])^2]}_{\text{Var}[\widehat{\theta}]}+\underbrace{\mathbb{E}[(\mathbb{E}[\widehat{\theta}]-\theta)^2]}_{(\text{Bias}(\widehat{\theta}))^2}+0$$

Definition 2.5 (Consistent)**.** $\widehat{\theta}$ is consistent if for any $\epsilon>0$, $\lim_{n\to\infty}\mathbb{P}\{|\widehat{\theta}-\theta|\ge\epsilon\}=0$.

2.2 Hypothesis Test

Definition 2.6 (Hypothesis test)**.** Null hypothesis: $H_0:\boldsymbol{\theta}\in\omega_0$; alternative hypothesis: $\boldsymbol{\theta}\in\omega_1$; critical region: $C=\{\mathbf{x}:\text{reject }H_0\}$; size α: $\alpha=\max_{\theta\in\omega_0}P_\theta[\mathbf{X}\in C]$; power function: $\pi_C(\boldsymbol{\theta})=1-\beta_C(\boldsymbol{\theta})=P_\theta[\mathbf{X}\in C]$ for $\boldsymbol{\theta}\in\omega_1$.

2.3 Pivitol

Remark 2.7. If $X \sim N(\mu, \sigma)$, then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Remark 2.8. If $X \sim N(\mu, \sigma)$, then

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

Theorem 2.9 (Central limit theorem)**.**

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Theorem 2.10. The CI of $\bar{X} - \bar{Y}$ is

2.4 MLE

Definition 2.11 (Likelihood)**.** Likelihood function $L(\boldsymbol{\theta}; \mathbf{x}) \equiv f(\mathbf{x}; \boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$ for fixed $\mathbf{x}$. It's often simpler to maximize log-likelihood: $\ell(\boldsymbol{\theta}; \mathbf{x}) \equiv \log(L(\boldsymbol{\theta}; \mathbf{x}))$. When i.i.d.,

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{x}) &= f(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) \\ \ell(\boldsymbol{\theta}; \mathbf{x}) &= \log\left(\prod_{i=1}^{n} f(x_i; \boldsymbol{\theta})\right) = \sum_{i=1}^{n} \log(f(x_i; \boldsymbol{\theta})). \end{aligned}$$

Remark 2.12. For normal distribution,

$$\begin{aligned} L(\boldsymbol{\theta}; \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i; \boldsymbol{\theta}) &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\} \exp\left\{-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right\}. \end{aligned}$$

Definition 2.13 (MLE Principle)**.** Choose estimator of $\boldsymbol{\theta}$ to maximize $L(\boldsymbol{\theta}; \mathbf{x})$: $\widehat{\boldsymbol{\theta}} \equiv \text{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta}; \mathbf{x})$.

Remark 2.14. S^2 is an unbiased estimator of σ^2. S is a biased estimator of σ. $\widehat{\sigma}$ is an unbiased estimator of σ. $\widehat{\sigma}^2$ is a biased estimator of σ^2.

Assumption 2.15 (Regularity assumptions)**.** These assumptions are

i) The pdf's are distinct for different $\boldsymbol{\theta}$;

ii) The pdf's have common support for all $\boldsymbol{\theta}$;

iii) $\boldsymbol{\theta}_0$ is in the interior of Ω.

2.5 Method of Moments

Lemma 2.16 (Shannon's Lemma)**.** Uniqueness.

3 Find confidence regions and critical regions

Definition 3.1 (CR)**.** For each $\boldsymbol{\theta}_0 \in \Omega$, let $C(\boldsymbol{\theta}_0) \subset \mathbb{R}^n$ denote the critical region for a size-α test of $H_0 : \boldsymbol{\theta} = \theta_0$ and a suitable H_1, and for each $\mathbf{x}$,

$$R(\mathbf{x}) \equiv \{\boldsymbol{\theta}_0 : \mathbf{x} \notin C(\boldsymbol{\theta}_0)\} \subset \mathbb{R}^p.$$

Then $R(\mathbf{x})$ is a $1-\alpha$ confidence region for $\boldsymbol{\theta}$.

Remark 3.2 (CR)**.** A few remarks on CR.

1. CI is for scalar θ. CR is for $p > 1$ dimensional $\boldsymbol{\theta}$.
2. When adopting pivotal quantities to find CRs, in the scalar case, one-sided intervals are unique, but two-sided are not. Taking the shortest length interval is equivalent to requiring that θ has constant likelihood on the boundary. In $p > 1$ case, taking a minimum-volumn CR is equivalent to requiring that $\boldsymbol{\theta}$ has constant likelihood on the boundary.

3.1 Asymptotic distribution

Definition 3.3 (Score function)**.** The score function is defined as

$$\mathbf{S}(\boldsymbol{\theta}) \equiv \nabla \log f(x;\boldsymbol{\theta}) \equiv \left[\frac{\partial f(x;\boldsymbol{\theta})/\partial\theta_1}{f(x;\boldsymbol{\theta})}, \frac{\partial f(x;\boldsymbol{\theta})/\partial\theta_2}{f(x;\boldsymbol{\theta})}, \ldots, \frac{\partial f(x;\boldsymbol{\theta})/\partial\theta_p}{f(x;\boldsymbol{\theta})}\right]^\top$$

Definition 3.4 (Fisher Information, Hessian Matrix)**.** The fisher information is defined as

$$\mathbf{I}(\boldsymbol{\theta}) \equiv -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log(f(X;\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}^2}\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial \log f(X;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)\left(\frac{\partial \log f(X;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right)^\top\right] = \mathrm{Cov}_{\boldsymbol{\theta}}[\nabla \log f(X;\boldsymbol{\theta})] \geq 0.$$

The Hessian matrix is defined as

$$\mathbf{H}(\boldsymbol{\theta}) \equiv \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log(f(X;\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}^2}\right] = -\mathbf{I}(\boldsymbol{\theta}).$$

Remark 3.5. Fisher information relates to how accurately we can identify $\boldsymbol{\theta}$. $\mathbf{I}(\boldsymbol{\theta})$ is inversely proportional to the variance of the MLE.

Remark 3.6 (Joint Fisher information)**.** The joint fisher information of $\mathbf{X} = (X_1, \ldots, X_n)^\top$ is

$$\mathbf{I}_{\text{joint}}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \log(f(\mathbf{X};\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^2}\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial \log f(\mathbf{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log f(\mathbf{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)^\top\right] = \text{Cov}_{\boldsymbol{\theta}}[\nabla \log f(\mathbf{X};\boldsymbol{\theta})].$$

For $X_i \overset{\text{i.i.d.}}{\sim} f(x;\boldsymbol{\theta})$, $\mathbf{I}_{\text{joint}}(\boldsymbol{\theta}) = n\mathbf{I}(\boldsymbol{\theta})$.

Theorem 3.7. Let $X_1, \ldots, X_n$ be iid with pdf $f(x;\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Omega$. Assume the regularity conditions hold. Then

1. The likelihood function $\frac{\partial}{\partial \boldsymbol{\theta}} l(\boldsymbol{\theta}) = \mathbf{0}$ has a solution $\widehat{\boldsymbol{\theta}}_n$ s.t. $\widehat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}$.
2. For any sequence which satisfies (1),

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{D} N_p\left(\boldsymbol{\theta}, \frac{\mathbf{I}^{-1}(\boldsymbol{\theta})}{n}\right).$$

Theorem 3.8. Let $\boldsymbol{g}$ be a transformation $\boldsymbol{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_k(\boldsymbol{\theta}))^\top$ s.t. $1 \le k \le p$ and that the $k \times p$ matrix of a partial derivatives $\boldsymbol{B} = \left[\frac{\partial g_i}{\partial \theta_j}\right]$, $i = 1, \ldots, k$, $j = 1, \ldots, p$ has continuous elements and does not vanish in a neighborhood of $\boldsymbol{\theta}$. Let $\widehat{\boldsymbol{\eta}} = \boldsymbol{g}(\widehat{\boldsymbol{\theta}})$. Then $\widehat{\boldsymbol{\eta}}$ is the mle of $\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta})$, and

$$\widehat{\boldsymbol{\eta}} \xrightarrow{D} N_k\left(\boldsymbol{\eta}, \frac{\boldsymbol{B}\mathbf{I}^{-1}(\boldsymbol{\theta})\boldsymbol{B}^\top}{n}\right).$$

Hence, $\mathbf{I}(\boldsymbol{\eta}) = [\boldsymbol{B}\mathbf{I}^{-1}(\boldsymbol{\theta})\boldsymbol{B}^\top]^{-1}$.

Remark 3.9. If expectation is tractable, take it. Substitute $\widehat{\mathbf{I}} = \mathbf{I}(\widehat{\boldsymbol{\theta}})$ for $\mathbf{I}(\boldsymbol{\theta}_0)$ if needed. Otherwise, calculate observed Fisher info matrix.

Theorem 3.10 (Find approximate CRs or hyp tests)**.** Individual normal CI on θ_j:

$$\widehat{\theta}_j \xrightarrow{D} N\left(\theta_j, \frac{[\mathbf{I}^{-1}(\boldsymbol{\theta})]_{j,j}}{n}\right).$$

Then, $SD(\widehat{\theta}_j) = \sqrt{\frac{[\mathbf{I}^{-1}(\boldsymbol{\theta})]_{j,j}}{n}}$, $SE(\widehat{\theta}_j) = \sqrt{\frac{[\widehat{\mathbf{I}}^{-1}]_{j,j}}{n}}$, the approx. $1-\alpha$ CI is $\theta_j \in \widehat{\theta}_j \pm z_{\alpha/2} SE(\widehat{\theta}_j)$.

Theorem 3.11 (Find joint χ^2 CR on $\boldsymbol{\theta}$)**.** With the construction $[\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}]^\top [n\mathbf{I}(\boldsymbol{\theta})][\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}] \xrightarrow{D} \chi^2_p$, an approx. $1-\alpha$ joint CR is

$$\left\{\boldsymbol{\theta} : [\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}]^\top [n\widehat{\mathbf{I}}][\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}] \le \chi^2_{p,\alpha}\right\}.$$

The CR is an ellipsoid centered at $\widehat{\boldsymbol{\theta}}$.

Remark 3.12. A few remarks:

1. If $\mathbf{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $(\mathbf{X}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \chi^2_k$.
2. Assume $\boldsymbol{\Sigma}$ has orthonormal eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots$ and eigenvalues $\lambda_1, \lambda_2, \ldots$, and let $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k]$ and $\mathbf{D} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$. Then $\boldsymbol{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^\top = [\mathbf{V}\mathbf{D}^{1/2}][\mathbf{V}\mathbf{D}^{1/2}]^\top = \mathbf{A}\mathbf{A}^\top$.

3.2 Critical Regions from Asymptotic Dist.

Theorem 3.13. Assume that

1. $X_i : 1 = 1, 2, \ldots, n$ i.i.d., $X \sim f(x; \boldsymbol{\theta})$, $n \to \infty$, same regularity conditions as for asymptotic distribution of MLE.
2. $\omega_0 = \{\boldsymbol{\theta} \in \Omega : g_i(\boldsymbol{\theta}) = a_i, i = 1, 2, \ldots, q\}$ for some set of $q \leq p$ smooth independent functions $g_i(\cdot)$ and constants a_i, and ω_0 is in the interior of Ω. (ω_0 is a $p-q$ dimensional manifold)
3. $\widehat{\boldsymbol{\theta}}_0$ and $\widehat{\boldsymbol{\theta}}$ in the LRT are consistent MLE solutions.

Then, when H_0 is true:

$$-2\log\Lambda(\mathbf{X}) \xrightarrow{D} \chi^2_q.$$

We reject H_0 if $-2\log\Lambda(\mathbf{X}) > \chi^2_{q,\alpha}$, i.e., reject H_0 if $\Lambda(\mathbf{X}) < c$ with $c = \exp\left\{\frac{-\chi^2_{q,\alpha}}{2}\right\}$.

Remark 3.14. Test with asymptotic distribution can better control α-risk.

4 Optimality Properties and Limits of Performance

Theorem 4.1 (Rao-Cramer Lower Bound)**.** Consider any estimator $\widetilde{\boldsymbol{\theta}}$. Under mild regularity conditions, we have

$$\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}) \geq \dot{\boldsymbol{\mu}}_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}) \mathbf{I}^{-1}_{\text{joint}}(\boldsymbol{\theta}) \dot{\boldsymbol{\mu}}'_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}).$$

For any linear combo $\mathbf{a}'\widetilde{\boldsymbol{\theta}}$ of elements of $\widetilde{\boldsymbol{\theta}}$, a lower bound on its variance is

$$\text{Var}(\mathbf{a}'\widetilde{\boldsymbol{\theta}}) = \mathbf{a}'\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\theta})\mathbf{a} \geq \mathbf{a}' \left[\dot{\boldsymbol{\mu}}_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}) \mathbf{I}^{-1}_{\text{joint}}(\boldsymbol{\theta}) \dot{\boldsymbol{\mu}}'_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\theta})\right] \mathbf{a}.$$

Definition 4.2 (Efficient & asymptotically efficient)**.** An unbiased estimator $\widetilde{\boldsymbol{\theta}}$ is efficient if it achieves the C-R bound, i.e., $\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}) = \mathbf{I}^{-1}_{\text{joint}}(\boldsymbol{\theta})$. An estimator $\widetilde{\boldsymbol{\theta}}$ is asymptotically efficient if $\boldsymbol{\mu}_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}) \overset{n\to\infty}{\to} \boldsymbol{\theta}$, and $n\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{\theta}}}(\boldsymbol{\theta}) \overset{n\to\infty}{\to} n\mathbf{I}^{-1}_{\text{joint}}(\boldsymbol{\theta})$.

Remark 4.3. MLEs are asymptotically efficient.

Definition 4.4 (Sufficient statistic)**.** $t(\mathbf{X})$ is a sufficient statistic for θ if $\frac{f(\mathbf{x};\theta)}{f_T(t(\mathbf{x});\theta)}$ does not depend on θ.

Theorem 4.5 (Neyman factorization theorem)**.** $t(\mathbf{X})$ is a sufficient statistic for θ iff we can factor $f(\mathbf{x};\theta) = g_1(t(\mathbf{x});\theta)g_2(\mathbf{x})$.

Definition 4.6 (MSS)**.** A sufficient statistic $\mathbf{t}(\mathbf{X}) = [t_1(\mathbf{X}), t_2(\mathbf{X}), \ldots, t_m(\mathbf{X})]'$ is minimal if it cannot be reduced to smaller dimension while retaining sufficiency.

Remark 4.7 (Data reduction)**.** If $f(\mathbf{x};\boldsymbol{\theta}) = g_1(t(\mathbf{x});\boldsymbol{\theta})g_2(\mathbf{x})$, then MLE $\widehat{\boldsymbol{\theta}}$ and LRT $\Lambda(\mathbf{X}) = \frac{L(\widehat{\boldsymbol{\theta}}_0;\mathbf{X})}{L(\widehat{\boldsymbol{\theta}};\mathbf{X})} = \frac{g_1(\mathbf{t}(\mathbf{x});\widehat{\boldsymbol{\theta}}_0)}{g_1(\mathbf{t}(\mathbf{x});\widehat{\boldsymbol{\theta}})}$ only depend on $\mathbf{x}$ via $\mathbf{t}(\mathbf{x})$.

Theorem 4.8 (Rao-Blackwell Theorem)**.** For general $f(\mathbf{x};\theta)$, let $\widetilde{\theta}(\mathbf{X})$ be an unbiased estimator of θ and $\mathbf{T} = \mathbf{t}(\mathbf{X})$ a sufficient statistic. Then the new estimator $\widetilde{\widetilde{\theta}}(\mathbf{T}) \equiv E_\theta[\widetilde{\theta}(\mathbf{X}) \,|\, \mathbf{T}]$ is also unbiased, and $\mathrm{Var}_\theta(\widetilde{\widetilde{\theta}}) \leq \mathrm{Var}_\theta(\widetilde{\theta}(\mathbf{X}))$ with equality iff $\widetilde{\theta}(\mathbf{X})$ is a function of $\mathbf{t}(\mathbf{X})$ alone (then $\widetilde{\widetilde{\theta}}(\mathbf{T}) = \widetilde{\theta}(\mathbf{X})$).

Remark 4.9. $\widetilde{\widetilde{\theta}}(\mathbf{T}) \equiv E_\theta[\widetilde{\theta}(\mathbf{X}) \,|\, \mathbf{T}]$ is not a function of θ.

Theorem 4.10. Let Z and Y be any two scalar r.v.s. Then we have $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y \,|\, Z]]$ and the variance decomposition $\mathrm{Var}[Y] = \mathbb{E}[\mathrm{Var}[Y \,|\, Z]] + \mathrm{Var}[\mathbb{E}[Y \,|\, Z]]$.

Definition 4.11 (Most powerful test)**.** For simple hypotheses, a test C is a most powerful test of size-α if C is size-α, and for any other size-α test C', $\pi_C(\boldsymbol{\theta}_1) \geq \pi_{C'}(\boldsymbol{\theta}_1)$.

Theorem 4.12 (Neyman-Pearson Theorem)**.** The size-α LRT is most powerful for simple hypotheses.

Definition 4.13 (Uniformly most powerful)**.** For composite hypotheses, a test C is a uniformly most powerful (UMP) test of size-α if C is size-α, and for any other size-α test C', $\pi_C(\boldsymbol{\theta}) \geq \pi_{C'}(\boldsymbol{\theta}), \forall \boldsymbol{\theta} \in \omega_1$.

5 Prediction of Random Variables

Remark 5.1. For $\mathbf{X}$ and $\mathbf{Y}$ continuous r.v.s,

$$f_{\mathbf{Y}\,|\,\mathbf{X}}(\mathbf{y}\,|\,\mathbf{x}) = \frac{f_{\mathbf{X},\mathbf{Y}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{X}}(\mathbf{x})} = \frac{f_{\mathbf{X}\,|\,\mathbf{Y}}(\mathbf{x}\,|\,\mathbf{y})f_{\mathbf{Y}}(\mathbf{y})}{\int f_{\mathbf{X}\,|\,\mathbf{Y}}(\mathbf{x}\,|\,\mathbf{z})f_{\mathbf{Y}}(\mathbf{z})\mathrm{d}\mathbf{z}}.$$

For $\mathbf{X}$ continuous and $\mathbf{Y}$ discrete,

$$p_{\mathbf{Y}\,|\,\mathbf{X}}(\mathbf{y}\,|\,\mathbf{x}) = \frac{f_{\mathbf{X}\,|\,\mathbf{Y}}(\mathbf{x}\,|\,\mathbf{y})p_{\mathbf{Y}}(\mathbf{y})}{\sum_{\mathbf{z}} f_{\mathbf{X}\,|\,\mathbf{Y}}(\mathbf{x}\,|\,\mathbf{z})p_{\mathbf{Y}}(\mathbf{z})}.$$

Definition 5.2 (Mean square error)**.**

$$\text{MSE}(\widehat{\mathbf{Y}}(\mathbf{X})) \equiv \mathbb{E}\left\{\left[\mathbf{Y}-\widehat{\mathbf{Y}}(\mathbf{X})\right]\left[\mathbf{Y}-\widehat{\mathbf{Y}}(\mathbf{X})\right]'\right\} = \text{Cov}\left(\mathbf{Y}-\widehat{\mathbf{Y}}(\mathbf{X})\right),$$

where the last equality holds when $\mathbb{E}[\widehat{\mathbf{Y}}(\mathbf{X})] = \mathbb{E}[\mathbf{Y}]$.

Theorem 5.3. Consider any function $\mathbf{g}(\mathbf{x}) = [g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x})]'$ of $\mathbf{x}$ (e.g., any prediction $\widehat{\mathbf{Y}}(\mathbf{x}) = \mathbf{g}(\mathbf{x})$). Then,

i) $\text{MSE}(\boldsymbol{\mu}_{\mathbf{Y}\,|\,\mathbf{X}}(\mathbf{X})) \leq \text{MSE}(\mathbf{g}(\mathbf{X}))$ (i.e., the conditional mean is the MMSE predictor of $\mathbf{Y}$).

ii) $\mathbb{E}\{[\mathbf{Y}-\boldsymbol{\mu}_{\mathbf{Y}\,|\,\mathbf{X}}(\mathbf{X})]g_j(\mathbf{X})\} = \mathbf{0}_{m\times 1}, \forall j$ (i.e., the prediction error for the optimal predictor are uncorrelated with any function of $\mathbf{X}$).

Theorem 5.4. For jointly Gaussian $\begin{bmatrix}\mathbf{Y}\\ \mathbf{X}\end{bmatrix} \sim N_{m+n}\left(\begin{bmatrix}\boldsymbol{\mu}_{\mathbf{Y}}\\ \boldsymbol{\mu}_X\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{\mathbf{Y}} & \boldsymbol{\Sigma}_{\mathbf{YX}}\\ \boldsymbol{\Sigma}_{\mathbf{XY}} & \boldsymbol{\Sigma}_{\mathbf{X}}\end{bmatrix}\right)$,

$$\mathbf{Y}\,|\,\mathbf{X} \sim N_m\left(\boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}}), \boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}\right).$$

In particular, the MMSE predictor is $\widehat{\mathbf{Y}}(\mathbf{x}) = \boldsymbol{\mu}_{\mathbf{Y}\,|\,\mathbf{X}}(\mathbf{x}) = \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{X}-\boldsymbol{\mu}_{\mathbf{X}})$ with $\text{MSE}(\boldsymbol{\mu}_{\mathbf{Y}\,|\,\mathbf{X}}(\mathbf{X})) = \boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}$.

Theorem 5.5 (Gauss-Markov theorem for prediction)**.** Construct the linear MMSE predictor

$$\begin{aligned}\widehat{\mathbf{Y}}^*(\mathbf{x}) &= \boldsymbol{\mu}_{\mathbf{Y}} + \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{\mathbf{X}}),\\ \text{with}\quad \text{MSE}(\widehat{\mathbf{Y}}^*(\mathbf{x})) &= \boldsymbol{\Sigma}_{\mathbf{Y}} - \boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{YX}}\boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\Sigma}_{\mathbf{XY}}.\end{aligned}$$

Furthermore, the prediction error $\mathbf{Y}-\widehat{\mathbf{Y}}^*(\mathbf{X})$ are zero-mean and uncorrelated with any linear function of $\mathbf{X}$, i.e., $\mathbb{E}\{\mathbf{Y}-\widehat{\mathbf{Y}}^*(\mathbf{X})\} = \mathbf{0}_{m\times 1}$ and $\mathbb{E}\{[\mathbf{Y}-\widehat{\mathbf{Y}}^*(\mathbf{X})](\mathbf{a}'\mathbf{X}+b)\} = \mathbf{0}_{m\times 1}, \forall \mathbf{a}\in\mathbb{R}^n, b\in\mathbb{R}$.

6 Bayesian Statistical Inference

Definition 6.1 (Bayesian paradigm for statistical inference)**.** After conducting an experiment and observing an $\mathbf{X} = \mathbf{x}$, calculate a posterior distribution based on a prior distribution $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$:

$$f_{\boldsymbol{\Theta}\,|\,\mathbf{X}}(\boldsymbol{\theta}\,|\,\mathbf{x}) = \frac{f_{\mathbf{X}\,|\,\boldsymbol{\Theta}}(\mathbf{x}\,|\,\boldsymbol{\theta})f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{\int f_{\mathbf{X}\,|\,\boldsymbol{\Theta}}(\mathbf{x}\,|\,\boldsymbol{\eta})f_{\boldsymbol{\Theta}}(\boldsymbol{\eta})\mathrm{d}\boldsymbol{\eta}}.$$

Remark 6.2. For normal prior and likelihood, posterior is also normal: $\boldsymbol{\Theta}\,|\,\mathbf{X} \sim N(\gamma, \lambda^2)$. Here γ is a weighted average of the prior mean and $\bar{x}$, where the weights are the inverse of the prior variance and the conditional variance of $\bar{X}\,|\,\mu$.

Definition 6.3 (Maximum a posterior (MAP) estimator)**.** The MAP estimator is given by

$$\widehat{\boldsymbol{\theta}}(\mathbf{x}) \equiv \operatorname*{argmax}_{\boldsymbol{\theta}} f_{\boldsymbol{\Theta} \mid \mathbf{X}}(\boldsymbol{\theta} \mid \mathbf{x}) = \operatorname*{argmax}_{\boldsymbol{\theta}} f_{\mathbf{X} \mid \boldsymbol{\Theta}}(\mathbf{x} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}).$$

Remark 6.4. Compared to MLE

$$\widehat{\boldsymbol{\theta}}(\mathbf{x}) \equiv \operatorname*{argmax}_{\boldsymbol{\theta}} f_{\mathbf{X} \mid \boldsymbol{\Theta}}(\mathbf{x} \mid \boldsymbol{\theta}).$$

Definition 6.5 (Posterior mean estimator)**.** The posterior mean estimator is given by

$$\widehat{\boldsymbol{\theta}}(\mathbf{x}) \equiv \mathbb{E}[\boldsymbol{\Theta} \mid \mathbf{X} = \mathbf{x}] = \int \boldsymbol{\theta} f_{\boldsymbol{\Theta} \mid \mathbf{X}}(\boldsymbol{\theta} \mid \mathbf{x}) \mathrm{d}\boldsymbol{\theta} = \frac{\int \boldsymbol{\theta} f_{\mathbf{X} \mid \boldsymbol{\Theta}}(\mathbf{x} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}{\int f_{\mathbf{X} \mid \boldsymbol{\Theta}}(\mathbf{x} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}.$$

Definition 6.6 (Bayesian credible regions)**.** A region $R \subset \mathbb{R}^p$ is a $1-\alpha$ credible region for $\boldsymbol{\Theta}$ if $\mathbb{P}\{\boldsymbol{\Theta} \in R \mid \mathbf{X} = \mathbf{x}\} = \int_R f_{\boldsymbol{\Theta} \mid \mathbf{X}}(\boldsymbol{\theta} \mid \mathbf{x}) \mathrm{d}\boldsymbol{\theta} = 1 - \alpha$.

Definition 6.7 (Highest posterior density (HPD))**.** A $1-\alpha$ highest posterior density (HPD) credible region R is a $1-\alpha$ credible region s.t. $\forall \boldsymbol{\theta} \in R, \boldsymbol{\theta}' \notin R$, we have $f_{\boldsymbol{\Theta} \mid \mathbf{X}}(\boldsymbol{\theta} \mid \mathbf{x}) \geq f_{\boldsymbol{\Theta} \mid \mathbf{X}}(\boldsymbol{\theta}' \mid \mathbf{x})$.

Definition 6.8 (Bayesian hyp. tests)**.** For testing $H_0 : \boldsymbol{\theta} \in \omega_0$ and $H_1 : \boldsymbol{\theta} \in \omega_1$. Choose p and reject H_0 if $\mathbb{P}\{\boldsymbol{\Theta} \in \omega_1 \mid \mathbf{X} = \mathbf{x}\} = \int_{\omega_1} f_{\boldsymbol{\Theta} \mid \mathbf{X}}(\boldsymbol{\theta} \mid \mathbf{x}) \mathrm{d}\boldsymbol{\theta} > p$.

Definition 6.9 (Bayesian prediction intervals/regions)**.** Predict $\mathbf{Y}$ based on

$$f_{\mathbf{Y} \mid \mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \int f_{\mathbf{Y} \mid \mathbf{X}, \boldsymbol{\Theta}}(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) f_{\boldsymbol{\Theta} \mid \mathbf{X}}(\boldsymbol{\theta} \mid \mathbf{x}) \mathrm{d}\boldsymbol{\theta}.$$

The posterior MMSE predictor of $\mathbf{Y} \mid \mathbf{X} = \mathbf{x}$ is the posterior mean $\widehat{\mathbf{Y}}(\mathbf{x}) = \mathbb{E}[\mathbf{Y} \mid \mathbf{X}] = \int \mathbf{y} f_{\mathbf{Y} \mid \mathbf{X}}(\mathbf{y} \mid \mathbf{x}) \mathrm{d}\mathbf{y}$.

Definition 6.10 (Conjugate)**.** A family of priors $\{f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}; \boldsymbol{\gamma})\}$ is conjugate for a given likelihood $f_{\mathbf{X} \mid \boldsymbol{\Theta}}(\mathbf{x} \mid \boldsymbol{\theta})$ if the posterior $f_{\boldsymbol{\Theta} \mid \mathbf{X}}(\boldsymbol{\theta} \mid \mathbf{x})$ is in the same family.

Definition 6.11 (Noninformative prior, improper prior)**.** A noninformative prior for $\boldsymbol{\Theta}$ is $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = c$ (a constant) for all $\boldsymbol{\theta} \in \boldsymbol{\Omega}$. A noninformative prior is called an improper prior if $\int f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \neq 1$. This is OK if $f_{\boldsymbol{\Theta} \mid \mathbf{X}}(\boldsymbol{\theta} \mid \mathbf{x}) \propto f_{\mathbf{X} \mid \boldsymbol{\Theta}}(\mathbf{x} \mid \boldsymbol{\theta}) f_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is proper.

Remark 6.12. It's common to use a noninformative prior for $\log(\boldsymbol{\theta})$, which is equivalent to using $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \propto c/\theta$.

Definition 6.13 (Jeffrey's prior)**.** The Jeffrey's prior sets $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \propto |\mathbf{I}(\boldsymbol{\theta})|^{1/2}$, where $\mathbf{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}\left[\left(\frac{\partial \log f(X;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \log f(X;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)'\right]$ is the Fisher information matrix. The Jeffrey's prior is invariant under transformation.

Remark 6.14. When $f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = c$, the MAP of $\boldsymbol{\theta}$ is MLE.

Definition 6.15 (Decision rule)**.** A decision rule $\boldsymbol{\delta}(\mathbf{x})$ is a function that maps $\mathbf{x}$ to a decision usually regarding the parameters $\boldsymbol{\theta}$.

Definition 6.16 (Loss function)**.** A loss function $L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x}))$ represents the cost of deciding $\boldsymbol{\delta}(\mathbf{x})$ when true parameters are $\boldsymbol{\theta}$, e.g. the squared-error loss $L(\boldsymbol{\theta}, \widehat{\boldsymbol{\theta}}(\mathbf{x})) = [\widehat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}][\widehat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}]'$.

Definition 6.17 (Risk function)**.** The risk function is the expected loss w.r.t. $\mathbf{X}$ for fixed $\boldsymbol{\theta}$, e.g.,

$$R(\boldsymbol{\theta}, \boldsymbol{\delta}) \equiv \mathbb{E}[L(\boldsymbol{\Theta}, \boldsymbol{\delta}(\mathbf{X})) \,|\, \boldsymbol{\Theta} = \boldsymbol{\theta}] \equiv \int L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) f_{\mathbf{X} \,|\, \boldsymbol{\Theta}}(\mathbf{x} \,|\, \boldsymbol{\theta}) \mathrm{d}\mathbf{x}.$$

Definition 6.18 (Bayesian risk)**.**

$$R(\boldsymbol{\delta}) \equiv \mathbb{E}[L(\boldsymbol{\Theta}, \boldsymbol{\delta}(\mathbf{X}))] = \iint L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) f_{\mathbf{X}, \boldsymbol{\Theta}}(\mathbf{x}, \boldsymbol{\theta}) \mathrm{d}\mathbf{x} \mathrm{d}\boldsymbol{\theta} = \iint L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) f_{\mathbf{X} \,|\, \boldsymbol{\Theta}}(\mathbf{x} \,|\, \boldsymbol{\theta}) \mathrm{d}\mathbf{x} f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}.$$

Remark 6.19. We can write

$$R(\boldsymbol{\delta}) \equiv \mathbb{E}[L(\boldsymbol{\Theta}, \boldsymbol{\delta}(\mathbf{X}))] = \mathbb{E}\{\mathbb{E}[L(\boldsymbol{\Theta}, \boldsymbol{\delta}(\mathbf{X})) \,|\, \mathbf{X}]\} = \int \left\{ \int L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) f_{\boldsymbol{\Theta} \,|\, \mathbf{X}}(\boldsymbol{\theta} \,|\, \mathbf{x}) \mathrm{d}\boldsymbol{\theta} \right\} f_{\mathbf{X}}(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

So the optimal function $\boldsymbol{\delta}^*(\cdot)$ is the pointwise (in $\mathbf{x}$) solution to

$$\boldsymbol{\delta}^*(\mathbf{x}) \equiv \underset{\boldsymbol{\delta}(\mathbf{x})}{\operatorname{argmin}} \int L(\boldsymbol{\theta}, \boldsymbol{\delta}(\mathbf{x})) f_{\boldsymbol{\Theta} \,|\, \mathbf{X}}(\boldsymbol{\theta} \,|\, \mathbf{x}) \mathrm{d}\boldsymbol{\theta}.$$