Statistics Cheatsheet

Hongyi Guo*

October 21, 2020

1 Probabilty & Distribution

Definition 1 (Population)**.** The entire group of individuals about which we want to learn something about.

Definition 2 (Sample)**.** Subset of the population from which the information is actually obtained.

Definition 3 (Statistics)**.** A numerical characteristic of the sample, a random variable.

Remark 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 5 (SD, SE)**.** SD is the standard deviation, while SE is the standard deviation of the sampling distribution, e.g., $SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$, $SE(\bar{X}) = \frac{\widehat{\sigma}}{\sqrt{n}} = \frac{s}{n}$.

Definition 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 $f_X(s) = \frac{\mathrm{d}}{\mathrm{d}x}F_X(s)$.

*Northwestern University; hongyiguo2025@u.northwestern.edu

Definition 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.).

Definition 8 (Point)**.**

Remark 9. The cdf, pmf/pdf, mle of various distributions.

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

Definition 11 (Bias, Var)**.** As follows,

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

Definition 12 (Consistent)**.** $\widehat{\theta}$ is consistent if for any $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\{|\widehat{\theta} - \theta| \ge \epsilon\} = 0$.

2.2 Pivitol

Theorem 13. The CI of $\bar{X} - \bar{Y}$ is

2.3 MLE

Definition 14 (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.,

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

Remark 15. For normal distribution,

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

Definition 16 (MLE Principle)**.** Choose estimator of $\boldsymbol{\theta}$ to maximize $L(\boldsymbol{\theta};\mathbf{x})$: $\widehat{\boldsymbol{\theta}} \equiv \operatorname{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta};\mathbf{x})$.

Remark 17. 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 18 (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.

Theorem 19 (Rao-Cramer Lower Bound)**.** Let $X_1,\dots,X_n$ be i.i.d. with common pdf $f(x;\theta)$ for $\theta \in \Omega$. Assume that the regularity conditions hold. Let $Y = u(X_1, X_2, \dots, X_n$ be a statistic with mean $E(Y) = E[u(X_1,X_2,\dots,X_n)] = k(\theta)$. Then

$$\operatorname{Var}(Y) \geq \frac{[k'(\theta)]^2}{nI(\theta)}.$$

2.4 Method of Moments

Lemma 20 (Shannon's Lemma)**.** Uniqueness.

3 Find confidence regions and critical regions

Definition 21 (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 22 (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 23 (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 24 (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 25. Fisher information relates to how accurately we can identify $\boldsymbol{\theta}$. $\mathbf{I}(\boldsymbol{\theta})$ is inversely proportional to the variance of the MLE.

Theorem 26. 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 27. Let $\boldsymbol{g}$ be a transformation $\boldsymbol{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \ldots, g_k(\boldsymbol{\theta}))^\top$ s.t. $1 \leq k \leq p$ 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, $\mathbf{I}(\boldsymbol{\eta}) = [\boldsymbol{B}\mathbf{I}^{-1}(\boldsymbol{\theta})\boldsymbol{B}^\top]$.

Remark 28. 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 29 (Find <u>approximate</u> 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 30 (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

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

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

Remark 31. 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 $\mathbf{D} = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_k\}$. 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 32. 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 33. Test with asymptotic distribution can better control α-risk.