Statistcs & Probability Cheat-sheet

Instructor: Dr. Razieh Nabi
Teaching assistant: Anna Guo

1 Probability rules

- **Joint distributions:** The joint cumulative distribution function (CDF) of random variables (r.v.s) X and Y is the function $F_{X,Y}$ given by

$$F_{X,Y}(x,y) = P(X \leq x, Y \leq y).$$

The joint probability mass function (PMF) of *discrete* r.v.s X and Y is the function $p_{X,Y}$ given by

$$p_{X,Y}(x,y) = P(X = x, Y = y).$$

The joint probability density function (PDF) of *continuous* r.v.s X and Y with joint CDF $F_{X,Y}$ is the function $f_{X,Y}$ given by

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y).$$

- **Marginalization:** For *discrete* r.v.s X and Y, the marginal PMF of X is

$$P(X = x) = \sum_y P(X = x, Y = y).$$

For *continuous* r.v.s X and Y with joint PDF $f_{X,Y}$, the marginal PDF of X is

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy.$$

- **Conditional distributions:** For *discrete* r.v.s X and Y, the conditional PMF of Y given $X = x$ is

$$P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}.$$

For *continuous* r.v.s X and Y with joint PDF $f_{X,Y}$, the conditional PDF of Y given $X = x$ is

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \forall x \text{ that } f_X(x) > 0$$

- **Bayes' theorem:**

$$P(X \mid Y) = \frac{P(X,Y)}{P(X)} = \frac{P(Y \mid X)P(X)}{P(Y)}$$

- **Independence of random variables:** Random variables $X_1, \ldots, X_n$ are independent if

$$P\left(X_1 \leq x_1, \ldots, X_n \leq x_n\right) = P\left(X_1 \leq x_1\right) \ldots P\left(X_n \leq x_n\right)$$

Go to page 1

- **Expectation:** The expected value (also called the expectation or mean) of a *discrete* r.v. X whose distinct possible values are $x_1, x_2, \cdots$ is defined by

$$\mathbb{E}(X) = \sum^{\infty} x_j P\left(X = x_j\right), \quad \text{or} \quad \mathbb{E}(X) = \sum_x \underbrace{x}_{\text{value}} \underbrace{P(X = x)}_{\text{PMF at } x} \text{ if the support is finite.}$$

The expected value of a *continuous* r.v. X with PDF f is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx.$$

Properties of Expectation:

1. Let g and h be functions of random variables X and Y (discrete or continuous) respectively, and let a and b be constants.

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$
$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$
$$\mathbb{E}\{ag(X) + bh(X)\} = a\mathbb{E}\{g(X)\} + b\mathbb{E}\{h(X)\}$$

2. IF X and Y be independent random variables, then

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$$
$$\mathbb{E}(g(X)h(Y)) = \mathbb{E}(g(X))\mathbb{E}(h(Y))$$

- **Covariance:** The covariance between r.v.s X and Y is

$$\text{Cov}(X, Y) = \mathbb{E}\{(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\} = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

Properties of Covariance:

1. $\text{Cov}(X, X) = \text{Var}(X)$
2. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$
3. $\text{Cov}(X, c) = 0$ for any constant c
4. $\text{Cov}(aX, Y) = a\,\text{Cov}(X, Y)$ for any constant a
5. $\text{Cov}(X + Y, Z + W) = \text{Cov}(X, Z) + \text{Cov}(X, W) + \text{Cov}(Y, Z) + \text{Cov}(Y, W)$
6. If X and Y are independent, then $\text{Cov}(X, Y) = 0$, but the reverse is not necessarily true (only true under normality assumption).

- **Variance:** The variance of r.v. X is

$$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mathbb{E}^2(X)$$

The square root of the variance is called the **standard deviation** (SD):

$$\text{SD}(X) = \sqrt{\text{Var}(X)}.$$

Properties of Variance:

Go to page 1

1. Let $g(X)$ be a function r.v. X (discrete or continuous), and let a and b be constants:

$$\operatorname{Var}(aX+b) = a^2 \operatorname{Var}(X).$$

$$\operatorname{Var}\{ag(X)+b\} = a^2 \operatorname{Var}\{g(X)\}.$$

2. For two r.v.s X and Y:

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y) + 2\operatorname{Cov}(X,Y).$$

   For n r.v.s $X_1, \ldots, X_n$:

$$\operatorname{Var}\left(X_1+\cdots+X_n\right) = \operatorname{Var}\left(X_1\right) + \cdots + \operatorname{Var}\left(X_n\right) + 2\sum_{i<j} \operatorname{Cov}\left(X_i, X_j\right)$$

3. If X and Y are independent, then

$$\operatorname{Var}(X+Y) = \operatorname{Var}(X) + \operatorname{Var}(Y).$$

   For n independent r.v.s $X_1, \ldots, X_n$

$$\operatorname{Var}\left(X_1+\cdots+X_n\right) = \operatorname{Var}\left(X_1\right) + \cdots + \operatorname{Var}\left(X_n\right)$$

- **Conditional expectation:**

$$\mathbb{E}(Y \mid X=x) = \sum_y y P(Y=y \mid X=x), \text{ if } Y \text{ is } \textit{discrete}$$

$$\mathbb{E}(Y \mid X=x) = \int_{-\infty}^{\infty} y f_{Y \mid X}(y \mid x) dy, \text{ if } Y \text{ is } \textit{continuous}.$$

- **Law of total Expectation/Tower rule/Adam's law:**

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X \mid Y))$$
$$\mathbb{E}(X \mid Y) = \mathbb{E}(\mathbb{E}(X \mid Z, Y) \mid Y)$$

Properties of conditional expectation:

1. If X and Y are independent, then

$$\mathbb{E}(Y \mid X) = \mathbb{E}(Y).$$

2. For any function h,

$$\mathbb{E}(h(X)Y \mid X) = h(X)\mathbb{E}(Y \mid X).$$

3. Linearity

$$E\left(Y_1+Y_2 \mid X\right) = E\left(Y_1 \mid X\right) + E\left(Y_2 \mid X\right).$$

4. Projection interpretation: For any function h, the random variable $Y - \mathbb{E}(Y \mid X)$ is uncorrelated with $h(X)$, i.e., $\operatorname{cov}(Y - \mathbb{E}[Y \mid X], h(X)) = 0$. Equivalently,

$$\mathbb{E}[(Y - \mathbb{E}(Y \mid X))h(X)] = 0.$$

Proof.
By applying the tower rule, we have

$$\begin{aligned}\mathbb{E}[(Y - \mathbb{E}(Y \mid X))h(X)] &= \mathbb{E}\left[\mathbb{E}((Y - \mathbb{E}(Y \mid X))h(X) \mid X)\right] \\ &= \mathbb{E}\left[h(X)\mathbb{E}(Y - \mathbb{E}(Y \mid X) \mid X)\right] \\ &= \mathbb{E}\left[h(X)(\mathbb{E}(Y \mid X) - \mathbb{E}(Y \mid X))\right] \\ &= 0\end{aligned}$$

- **Conditional variance:** The conditional variance of Y given X is

$$\operatorname{Var}(Y \mid X) = \mathbb{E}\left((Y - \mathbb{E}(Y \mid X))^2 \mid X\right) = \mathbb{E}\left(Y^2 \mid X\right) - \mathbb{E}^2(Y \mid X).$$

- **Law of total Variance/Eve's law:**

$$\operatorname{Var}(Y) = \mathbb{E}[\operatorname{Var}(Y \mid X)] + \operatorname{Var}(\mathbb{E}[Y \mid X])$$

2 Inference

- **Modes of Convergence:** Under probability measure space $(\Omega, \mathcal{A}, P)$.

1. Convergence almost surely: X_n is said to converge almost surely to X, denoted by $X_n \to_{a.s.} X$, if there exists a set $A \subset \Omega$ such that $P(A^c) = 0$ and for each $\omega \in A, X_n(\omega) \to X(\omega)$ in real space. Equivalently,

$$X_n \to_{a.s} X \Longleftrightarrow \forall \epsilon > 0. \lim_{n\to\infty} P\left(\sup_{m \geq n} \mid X_m - X \mid > \epsilon\right) = 0.$$

2. Convergence in probability: X_n is said to converge in probability to X, denoted by $X_n \to_p X$, if for every $\epsilon > 0$,

$$P\left(|X_n - X| > \epsilon\right) \to 0$$

3. Convergence in moments/means: For $X_n, X \in L_r(P)$, X_n is said to converge in r-th mean to X, denoted by $X_n \to_r X$ if

$$E\left(|X_n - X|^r\right) \to 0$$

($X \in L_r(P)$: $\mathbb{E}(|X|^r) < \inf$)

4. Convergence in distribution: X_n is said to converge in distribution to X, denoted by $X_n \to_d X$, if the distribution functions of X_n and X, denoted by F_n and F respectively, satisfy

$$F_n(x) \to F(x)$$

for each continuous point x of F.

- **Relationship among modes:**

1. $X_n \to_{a.s.} X \Longrightarrow X_n \to_p X$.
2. $X_n \to_p X \Longrightarrow X_{n_k} \to_{a.s.} X$ for some subsequence X_{n_k}.
3. $X_n \to_r X \Longrightarrow X_n \to_p X$.
4. $X_n \to_p X$ and $|X_n|^r$ is uniformly integrable $(\lim_{\lambda\to\infty} \sup_n E\{|X_n| I(|X_n| \geq \lambda)\} = 0) \Longrightarrow X_n \to_r X$.
5. $X_n \to_p X$ and $\limsup_n E|X_n|^r \leq E|X|^r \Longrightarrow X_n \to_r X$.
6. $X_n \to_r X \Longrightarrow X_n \to_{r'}, \forall\, 0 < r' < r$.
7. $X_n \to_p X \Longrightarrow X_n \to_d X$.
8. $X_n \to_p X$ if and only if for every subsequence $\{X_{n_k}\}$, there exists a further subsequence $\{X_{n_{k_l}}\}$ such that $X_{n_{k_l}} \to_{a.s.} X$.
9. $X_n \to_d c$, for some constant $c \Longrightarrow X_n \to_p c$.

- **Algebra of big O and small o:**

$O(\cdot)$ and $o(\cdot)$ in calculus: For two sequences of real numbers $\{a_n\}$ and $\{b_n\}$,

1. $a_n = O(b_n)$ if and only if $\exists\, C \in \mathbb{R}$, such that $|a_n| \leq C|b_n|, \forall\, n$.
2. $a_n = o(b_n)$ if and only if $a_n/b_n \to 0$ as $n \to \infty$.

$O(\cdot)$ and $o(\cdot)$ for random variables: Let $X_1, \cdots, X_n$ and $Y_1, \cdots, Y_n$ be random variables defined a probability space $(\Omega, \mathcal{A}, P)$.

1. $X_n = O(Y_n)$, a.s. if and only if $X_n(\omega) = O(Y_n(\omega))$, a.s. wrt P.
2. $X_n = o(Y_n)$, a.s. if and only if $X_n/Y_n \to_{a.s.} 0$.
3. $X_n = O_p(Y_n)$ if and only if,for any $\epsilon > 0$, there is a constant $C_\epsilon > 0$, such that

$$\sup_n P(|X_n| > C_\epsilon |Y_n|) < \epsilon$$

4. $X_n = o_p(Y_n)$ if and only if $X_n/Y_n \to_p 0$.

Properties of big O and small o:

1. $X_n = o_p(1) \Longrightarrow X_n = O_p(1)$.
2. $W_n = O_p(1)$, $X_n = O_p(1) \Longrightarrow W_n + X_n = O_p(1)$, $W_n X_n = O_p(1)$.
3. $W_n = O_p(1)$, $X_n = o_p(1) \Longrightarrow W_n + X_n = O_p(1)$, $W_n X_n = o_p(1)$.
4. $X_n = O_p(Y_n)$, $W_n = O_p(Z_n) \Longrightarrow W_n X_n = O_p(Y_n Z_n)$, $W_n + X_n = O_p(\max(Z_n, Y_n))$

- **Weak law of large numbers:** If $X_1, X_2, \dots, X_n$ are i.i.d with mean μ, then for sample mean $\bar{X}_n = \sum_{i=1}^n X_i/n$, we have $\bar{X}_n \to_p \mu$.

- **Strong law of large numbers:** If $X_1, X_2, \dots, X_n$ are i.i.d with mean μ, then $\bar{X}_n \to_{a.s.} \mu$.

- **Central limit theorem:** Suppose $\{X_1, X_2, \dots, X_n\}$ is a sequence of i.i.d. random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$ then as $n \to \infty$, $\bar{X} \to N(\mu, \frac{\sigma^2}{n})$

- **Score:** For r.v X with PDF $f(x;\theta)$. Score Z is defined as the partial derivative with respect to θ of the natural logarithm of the likelihood function:

$$Z = l' = \frac{\partial}{\partial\theta} \log f(X;\theta)$$

$$E(Z) = 0 \text{ and } Z \xrightarrow{d} N(0, I(\theta))$$

- **Fisher information:** The variance of the score is defined to be the Fisher information

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

- **Maximum likelihood estimation (MLE):** Consider a parametric model $f(x;\theta)$ where $\theta \in \mathbb{R}^k$. Suppose we have n i.i.d observations $X_1, \dots, X_n \overset{i.i.d}{\sim} f(x;\theta)$. MLE estimator, denoted by $\widehat{\theta}$, is constructed by maximizing the likelihood function $L(\theta)$ or equivalently the log-likelihood function $l(\theta)$

$$L(\theta) = \prod_{i=1}^n f(X_i;\theta), \qquad l(\theta) = \log(L(\theta)) = \sum_{i=1}^n \log(f(X_i;\theta)).$$

Properties of MLE estimators:

1. Consistency: $\hat{\theta} \to_p \theta$
2. Efficiency: it achieves the Cramer–Rao lower bound (discussed below) when the sample size tends to infinity.

$$\sqrt{n}(\hat{\theta} - \theta) \to_d \mathbb{N}(0, \frac{1}{\mathbb{I}(\theta)})$$

- **Delta method:** If a function $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable at $\theta \in \mathbb{R}$, and if

$$\sqrt{n}(\hat{\theta} - \theta) \to \mathbb{N}(0, v(\theta))$$

in distribution as $n \to \infty$ for some variance $v(\theta)$, then

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \to \mathbb{N}\left(0, g'(\theta)^2 v(\theta)\right)$$

Proof.
Perform a Taylor expansion of $g(\hat{\theta})$ around $\hat{\theta} = \theta$:

$$g(\hat{\theta}) \approx g(\theta) + (\hat{\theta} - \theta)g'(\theta).$$

Go to page 1

Rearranging yields

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \approx \sqrt{n}(\hat{\theta} - \theta)g'(\theta).$$

The result follows, because multiplying $\sqrt{n}(\hat{\theta} - \theta)$ by $g'(\theta)$ scales its variance by $g'(\theta)^2$.

- **Continuous mapping theorem:** Suppose that $X_n \to_{a.s.} X$ or $X_n \to_p X$ or $X_n \to_d X$. Then for any continuous function g, $g(X_n)$ converges to $g(X)$ almost surely, or in probability, or in distribution respectively.
- **Slutsky Theorem:** Suppose $X_n \to_d X$, $Y_n \to_p Y$ and $Z_n \to_p Z$ for some constant y and z. Then

$$Z_n X_n + Y_n \to_d zX + y.$$

- **Cramer-Rao lower bound:** Consider a parametric model $f(x;\theta)$ where $\theta \in \mathbb{R}$ is a single parameter. Let T be any unbiased estimator of θ based on data $X_1, \ldots, X_n \overset{i.i.d}{\sim} f(x;\theta)$. Then (under mild smoothness assumptions)

$$\text{Var}[T] \geq \frac{1}{n\mathbb{I}(\theta)}.$$

Proof.

$$Z = \frac{\partial}{\partial \theta} \log f\left(X_1, \ldots, X_n; \theta\right) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \log f\left(X_i; \theta\right).$$

Given that

$$\mathbb{E}\left[\frac{\partial}{\partial \theta} \log f\left(X_i; \theta\right)\right] = 0, \quad \text{Var}\left[\frac{\partial}{\partial \theta} \log f\left(X_i; \theta\right)\right] = \mathbb{I}(\theta).$$

(The score has mean 0, and variance given by the Fisher information.) Then

$$\mathbb{E}[Z] = 0, \quad \text{Var}[Z] = n\mathbb{I}(\theta).$$

Note that the correlation between Z and the estimator T is always between -1 and 1:

$$\text{Cov}[Z,T]^2 \leq \text{Var}[Z] \times \text{Var}[T] \leq n\mathbb{I}(\theta) \times \text{Var}[T]$$

Since T is unbiased,

$$\theta = \mathbb{E}[T] = \int T\left(x_1, \ldots, x_n\right) f\left(x_1, \ldots, x_n; \theta\right) dx_1 \ldots dx_n.$$

Differentiating both sides with respect to θ,

$$\begin{aligned} 1 &= \int T\left(x_1, \ldots, x_n\right) \frac{\partial}{\partial \theta} f\left(x_1, \ldots, x_n \mid \theta\right) dx_1 \ldots dx_n \\ &= \int T\left(x_1, \ldots, x_n\right) \left(\frac{\partial}{\partial \theta} \log f\left(x_1, \ldots, x_n \mid \theta\right)\right) f\left(x_1, \ldots, x_n \mid \theta\right) dx_1 \ldots dx_n = \mathbb{E}_\theta[TZ]. \end{aligned}$$

Since $\mathbb{E}[Z] = 0$, this implies

$$\mathrm{Cov}[T, Z] = \mathbb{E}\left[(T - \mathbb{E}T)(Z - \mathbb{E}Z)\right] = \mathbb{E}\left[T(Z - \mathbb{E}Z)\right] = \mathbb{E}[TZ] = 1,$$

so $\mathrm{Var}[T] \geq \frac{1}{n\mathbb{I}(\theta)}$ as desired.

Corollary.
For a parametric model $f(x;\theta)$ with a single parameter $\theta \in \mathbb{R}$, if T is any unbiased estimator of $g(\theta)$ based on data $X_1, \ldots, X_n \overset{i.i.d}{\sim} f(x;\theta)$, then (under mild smoothness assumptions)

$$\mathrm{Var}_\theta[T] \geq \frac{g'(\theta)^2}{n\mathbb{I}(\theta)}$$

Proof.
Similar as the previous proof but need Delta method additionally.

- **Ancillary statistics:** A statistics $S(X)$ whose distribution does not depend on the parameter θ is called an ancillary statistic. More precisely, a statistic $S(X)$ is ancillary for Θ if it's distribution is the same for all $\theta \in \Theta$.

Example (Location Family Ancillary Statistic): Let $X_1, \cdots, X_n$ be i.i.d. observations from a location parameter family with CDF $F(x-\theta), -\infty < \theta < \infty$. Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics from the sample. The range $R = X_{(n)} - X_{(1)}$ is always an ancillary statistic.
Proof.
Suppose $Z_1, \cdots, Z_n$ are i.i.d. observations from $F(x)$, with $X_1 = Z_1 + \theta, \cdots, X_n = Z_n + \theta$. It follows that the CDF of the range R is

$$\begin{aligned} F_R(r;\theta) = P_\theta(R \leq r) &= P_\theta\left(\max_i X_i - \min_i X_i \leq r\right) \\ &= P_\theta\left(\max_i Z_i - \min_i Z_i \leq r\right) \end{aligned}$$

The distribution of Z_i does not dependent on θ. Thus, the CDF of R does not depend on θ and hence R is ancillary.
Example: Let $X_1, \cdots, X_n$ be i.i.d observations from $\mathbb{N}(\mu, \sigma^2)$. Let

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$$

We know that

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

so that S^2 depends on σ^2 but not on μ. Therefore, S^2 is ancillary for

$$\Theta_1 = \left\{\left(\mu, \sigma^2\right) : \sigma^2 = \sigma_0^2\right\},$$

but is not ancillary for

$$\Theta_2 = \left\{\left(\mu, \sigma^2\right) : \sigma^2 > 0\right\}.$$

References

1. STATS5 325 by Dr. Rachel Fewster, The University of Auckland.
2. Blitzstein, J. K., & Hwang, J. (2019). Introduction to probability. Crc Press.
3. S&DS 410/610: Statistical Inference by Dr. Fan Zhou, Yale University.
4. S&DS 242/542: Theory of Statistics by Dr. Fan Zhou, Yale University.
5. BIOS 710: Probability Theory II by Dr. Limin Peng, Emory University.
6. STAT 205B: Classical Inference by Jizhou Kang, University of California, Santa Cruz.
7. Notes by Dr. Debdeep Pati, Texas A&M University.