Part I Exam Cheatsheets: Probability and Statistics Theory

Eric Ordoñez

April 2025

1 Common distributions

Distribution	Mean	Variance	PMF/PDF
$\mathrm{Unif}([a,b])$	$\frac{a+b}{2}$	$\frac{(b-a)(b-a+2)}{12}$	$\frac{1}{b-a+1}$
$\mathrm{Ber}(p)$	p	$p(1-p)$	$p^x(1-p)^{1-x}$
$\mathrm{Bin}(n,p)$	np	$np(1-p)$	$\binom{n}{k}p^k(1-p)^{n-k}$
$\mathrm{Geom}(p)$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$p(1-p)^{k-1}$
$\mathrm{Poiss}(\lambda)$	λ	λ	$\frac{\lambda^k e^{-\lambda}}{k!}$
$\mathrm{Unif}([a,b])$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{1}{b-a}$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
$\mathrm{Exp}(\lambda)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\lambda e^{-\lambda x}$
$\mathrm{Beta}(\alpha,\beta)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$
$\mathrm{Gamma}(\alpha,\beta)$	$\frac{\alpha}{\beta}$	$\frac{\alpha}{\beta^2}$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$

2 Conditioning, independence, and counting

2.1 Conditional probability

Multiplication rule: Given a countable set of events $\{A_i\}$,

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 \cap A_2)\cdots\mathbb{P}\left(A_n \,\middle|\, \bigcap_{i=1}^{n-1} A_i\right)$$

Law of total probability: Given a mutually exclusive, collectively exhaustive, and countable set of events $\{A_i\}$, $\mathbb{P}(B) = \sum_i \mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i)$.

2.2 Independence

A countable set of events $\{A_i\}$ is independent if $\mathbb{P}\left(\bigcap_{i\in S} A_i\right) = \prod_{i\in S}\mathbb{P}(A_i)$ for every subsets S of the enumeration of $\{A_i\}$.

Some facts about independence:

- A and B are independent iff $\mathbb{P}(A \mid B) = \mathbb{P}(A)$.
- If A and B are independent, so are A and B^c (and so are A^c and B^c).
- Independence implies pairwise independence, but not vice versa.
- Independence does not imply conditional independence, and vice versa.
- If X and Y are independent r.v.s, then $\mathbb{E}[g(X)h(Y)] = \mathbb{E}[g(X)]\,\mathbb{E}[h(Y)]$ for any functions g, h, and $\mathrm{Var}(X+Y) = \mathrm{Var}\,X + \mathrm{Var}\,Y$.

2.3 Counting

Number of…

- Permutations of n objects: $n!$
- k-permutations of n objects: $n!/(n-k)!$
- Combinations of k out of n objects: $n!/(k!(n-k)!)$
- Partitions of n objects into r groups with the i-th group having n_i objects: $n!/(n_1!n_2!\cdots n_r!)$

3 Random variables

3.1 Properties of expectation and variance

Law of iterated expectations: $\mathbb{E}[\mathbb{E}[X \mid Y]] = \mathbb{E}[X]$

Law of total expectation: $\mathbb{E}[X] = \int_Y \mathbb{E}[X \mid Y = y] f_Y(y)\,\mathrm{d}y$

Law of total variance: $\mathrm{Var}\,X = \mathbb{E}[\mathrm{Var}(X \mid Y)] + \mathrm{Var}\,\mathbb{E}[X \mid Y]$

3.2 Derived distributions

How to find the distribution of a function $Y = g(X)$ of a continuous r.v. X with known distribution f_X:

$$f_Y(y) = \frac{\mathrm{d}F_Y(y)}{\mathrm{d}y} = \frac{\mathrm{d}}{\mathrm{d}y}\mathbb{P}\left[g(X) \le y\right] = \frac{\mathrm{d}}{\mathrm{d}y}\int_{\{x \mid g(x) \le y\}} f_X(x)\,\mathrm{d}x$$

Two important cases:

- A linear transformation $Y = aX + b$:
$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$$
- A monotonic transformation $Y = g(X)$, where $h(y) = g^{-1}(y)$:
$$f_Y(y) = f_X\left(h(y)\right)\left|\frac{\mathrm{d}h(y)}{\mathrm{d}y}\right|$$

3.3 Sum of independent random variables

The PDF of the sum of two independent r.v.s is the *convolution* of their PDFs. If $Z = X + Y$, then $f_Z(z) = \int_{\mathbb{R}} f_X(x) f_Y(z-x)\,\mathrm{d}x$.

One application of this is that the sum of finitely many independent normal variables is normal: $\sum_{i=1}^n \mathcal{N}\left(\mu_i, \sigma_i^2\right) \sim \mathcal{N}\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right)$.

3.4 Correlation and covariance

The *correlation coefficient* measures the linear association between variables:

$$\rho_{XY} = \frac{1}{n}\sum_{i=1}^n \left(\frac{X_i - \overline{X}_n}{\sigma_X}\right)\left(\frac{Y_i - \overline{Y}_n}{\sigma_Y}\right) = \frac{\mathrm{Cov}\left(X, Y\right)}{\sigma_X \sigma_Y} \in [-1, 1]$$

Properties of covariance:

- $\mathrm{Cov}\left(aX + b, Y\right) = a\,\mathrm{Cov}\left(X, Y\right)$
- $\mathrm{Cov}\left(X, Y + Z\right) = \mathrm{Cov}\left(X, Y\right) + \mathrm{Cov}\left(Y, Z\right)$
- $\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}\,X_i + \sum_{\{(i,j) \mid i \ne j\}} \mathrm{Cov}\left(X_i, X_j\right)$

4 Stochastic processes

Start with a sequence of independent geometric (exponential) random variables (T_n) with common parameter p (λ). (Let these be the interarrival times). Then the sequence (Y_n) of arrival times is a Bernoulli (Poisson) process defined $Y_k = \sum_{i=1}^k T_i$.

If Bernoulli, the PMF of Y_k is the Pascal PMF of order k:

$$p_{Y_k}(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k} \qquad t = k, k+1, \ldots$$

If Poisson, the PDF of Y_k is the Erlang PDF of order k:

$$f_{Y_k}(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}$$

For a Bernoulli process with parameter p over n steps, the number of arrivals is $S \sim \mathrm{Bin}\left(n, p\right)$. For a Poisson process with rate λ over an interval of length τ, the number of arrivals is $N_\tau \sim \mathrm{Poiss}\left(\lambda\tau\right)$.

Splitting a Bernoulli (Poisson) process with parameter p (λ):

1. Keep with probability q and get a Bernoulli process with parameter pq.
2. Keep with probability p and get a Poisson process with rate λp.

Merging two independent Bernoulli (Poisson) processes with parameters p and q (λ_1 and λ_2), respectively:

1. Get a Bernoulli process with parameter $1 - (1-p)(1-q) = p + q - pq$.
2. Get a Poisson process with rate $\lambda^* = \lambda_1 + \lambda_2$, with arrival probabilities λ_1/λ^* and λ_2/λ^* of originating from the first and second process, respectively.

5 Convergence and limit theorems

5.1 Useful inequalities

Markov: For $X \ge 0$ with $\mathbb{E}\left[X\right] > 0$ and $t > 0$, $\mathbb{P}\left[X \ge t\right] \le \mathbb{E}\left[X\right]/t$.

Chebyshev: For X with $\mathbb{E}\left[X\right] < \infty$ and $t > 0$, $\mathbb{P}\left[\left|X - \mathbb{E}\left[X\right]\right| \ge t\right] \le (\mathrm{Var}\,X)/t^2$.

Hoeffding: Given $X_{i \in [n]} \overset{\text{i.i.d.}}{\sim} X$ that are a.s. bounded, i.e., there exist $a < b$ such that $\mathbb{P}\left[X_i \notin [a, b]\right] = 0$, then $\mathbb{P}\left[\left|\overline{X}_n - \mathbb{E}\left[X\right]\right| \ge \epsilon\right] \le 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$ for all $\epsilon > 0$.

5.2 Modes of convergence

Let (T_n) be a sequence of r.v.s and T another r.v., all in $\mathbb{R}$.

1. *Convergence almost surely:* $T_n \xrightarrow{\text{a.s.}} T \iff \mathbb{P}\left[\{\omega \mid T_n(\omega) \to T(\omega)\}\right] = 1$
2. *Convergence in probability:* $T_n \xrightarrow{\mathbb{P}} T \iff \mathbb{P}\left[|T_n - T| \ge \epsilon\right] \to 0$ for all $\epsilon > 0$
3. *Convergence in distribution:* $T_n \xrightarrow{\text{d}} T \iff \mathbb{E}\left[f(T_n)\right] \to \mathbb{E}\left[f(T)\right]$ for all continuous and bounded f

(1) implies (2) implies (3), but (3) implies (2) only if the limit T has a density: $T_n \xrightarrow{\mathrm{d}} T \implies \mathbb{P}[a \leq T_n \leq b] \to \mathbb{P}[a \leq T \leq b]$.

Continuous mapping theorem: Continuous functions preserve limits.

5.3 Limit theorems

Let $X_{i\in[n]} \overset{\text{i.i.d.}}{\sim} X$ with finite mean μ and sample mean $\overline{X}_n$.

- **Strong LLN:** $\overline{X}_n \xrightarrow{\text{a.s.}} \mu$, i.e., $\mathbb{P}\left[\lim_{n\to\infty} \overline{X}_n = \mu\right] = 1$.
- **Weak LLN:** If $\operatorname{Var} X < \infty$, then $\overline{X}_n \xrightarrow{\mathbb{P}} \mu$, i.e., $\mathbb{P}\left[\left|\overline{X}_n - \mu\right| \geq \epsilon\right] \to 0$ for all $\epsilon > 0$.

Central limit theorem: If, in addition, $\operatorname{Var} X = \sigma^2 < \infty$, then the sample mean is asymptotically normal, i.e., $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2)$.

Slutsky's theorem: Let (T_n) and (U_n) be sequences of r.v.s such that $T_n \xrightarrow{\mathrm{d}} T$ and $U_n \xrightarrow{\mathbb{P}} u \in \mathbb{R}$. Then

- $T_n + U_n \xrightarrow{\mathrm{d}} T + u$
- $T_n U_n \xrightarrow{\mathrm{d}} Tu$
- $\frac{T_n}{U_n} \xrightarrow{\mathrm{d}} \frac{T}{u}$ if $u \neq 0$.

6 Statistical inference

6.1 Models and estimation

For a statistical model $\left(E, \{\mathbb{P}_\theta\}_{\theta\in\Theta}\right)$:

- The model is *parametric* if $\Theta \subseteq \mathbb{R}^m$ and $\mathbb{P}_\theta$ is uniquely specified by θ.
- θ is *identifiable* if the map $\theta \mapsto \mathbb{P}_\theta$ is injective.

For an associated i.i.d. sample $X_{i\in[n]}$ drawn from a distribution $\mathbb{P}_\theta$:

- A *statistic* is any measurable function of the sample.
- An *estimator* of θ is a statistic whose expression does not depend on θ.
- An estimator $\hat{\theta}_n$...
  - is *weakly consistent* if $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.
  - is *asymptotically normal* if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \sigma^2)$, with *asymptotic variance* σ^2.
  - has *bias* equal to $\mathbb{E}[\hat{\theta}_n] - \theta$.
  - has *quadratic risk* equal to $\mathbb{E}[|\hat{\theta}_n - \theta|^2] = \text{variance} + \text{bias}^2$.

6.2 Delta method

Let (Z_n) be a sequence of r.v.s that are asymptotically normal around θ with variance σ^2. If the function g is continuously differentiable at θ, then $g(Z_n) \xrightarrow{\mathbb{P}} g(\theta)$ and $g(Z_n)$ is asymptotically normal around $g(\theta)$ with variance $g'(\theta)^2\sigma^2$.

7 Bayesian inference

Recall **Bayes' theorem:**

$$\mathbb{P}(A_i \mid B) = \frac{\mathbb{P}(A_i)\,\mathbb{P}(B \mid A_i)}{\sum_j \mathbb{P}(A_j)\,\mathbb{P}(B \mid A_j)} = \frac{\mathbb{P}(A)\,\mathbb{P}(B \mid A)}{\mathbb{P}(B)} \quad \text{if only one event } A$$

Let $\pi(\theta)$ and $\pi(\theta \mid X)$ be the prior and posterior distributions, respectively.

- **Bayes estimate:** $\hat{\theta}^{(\pi)} = \int_\Theta \mathrm{d}\pi(\theta \mid X)$
- **Maximum a posteriori estimate:** $\hat{\theta}^{\text{MAP}} = \operatorname{argmax}_{\theta\in\Theta} \pi(\theta \mid X)$.
- **Least mean squares estimate:** $\hat{\theta}^{\text{LMS}} = \mathbb{E}[\Theta \mid X = x]$.

Ways to evaluate a Bayesian estimator (can be unconditional or conditional):

- **Probability of error:** $\mathbb{P}[\hat{\theta} \neq \theta]$
- **Mean squared error:** $\mathbb{E}[(\hat{\theta} - \theta)^2]$

On prior and posterior distributions:

- If the PDF of X can be written $f(x) = ce^{-(\alpha x^2 + \beta x + \gamma)}$ with $\alpha > 0$, then X is normal with mean $-\beta/2\alpha$ and variance $1/2\alpha$.
- An *improper prior* is measurable, nonnegative, but not integrable.
- Example: Bernoulli experiment with a beta prior parameterized (α, β) has a beta posterior with updated parameters $\left(\alpha + \sum_{i=1}^n X_i, \beta + n - \sum_{i=1}^n X_i\right)$.
- Jeffreys prior: A *non-informative prior*, i.e., lacking prior information about a parameter, defined $\pi_{\mathcal{J}}(\theta) \propto \sqrt{\det I(\theta)}$.

8 Hypothesis testing

8.1 Confidence intervals

The *quantile* of order $1 - \alpha$ of a r.v. X is the number q_α such that $\mathbb{P}[X \leq q_\alpha] = 1 - \alpha$.

A *confidence interval* of (asymptotic) level $1 - \alpha$ for θ is any random (dependent upon the random sample) interval $\mathcal{I}$, whose boundaries do not depend on θ, such that $(\lim_{n\to\infty})\,\mathbb{P}[\mathcal{I} \ni \theta] \geq 1 - \alpha$ for all $\theta \in \Theta$.

8.2 Errors and p-values

The *p-value* is the smallest significance level at which H_0 is rejected.

- *Type I error:* Reject H_0 when H_0 is true.
- *Type II error:* Fail to reject H_0 when H_1 is true.
- *Significance level α:* $\mathbb{P}(\text{Type I error}) \leq \alpha$.
- *Power:* $1 - \mathbb{P}(\text{Type II error})$.

8.3 Wald test vs t-test

- The t-test requires the data to be Gaussian and can only be performed on expected values.
- The Wald test is asymptotic; the t-test can compute non-asymptotic p-values.
- For large sample sizes, the quantiles of the T distribution converge to those of the standard normal distribution.
- In general, the Wald test is more flexible and leads to lower p-values.

9 Methods of estimation

9.1 Maximum likelihood estimation

Minimize an estimate of the KL divergence between an observed distribution and a hypothesized distribution defined by a true parameter θ^*:

$$\mathrm{KL}\left(\mathbb{P}_\theta, \mathbb{P}_{\theta'}\right) = \int_E f_\theta(x) \log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right) \mathrm{d}x$$

Under some technical conditions, the MLE is a weakly consistent estimator for θ^*:

- θ^* is identifiable.
- θ^* is in the interior of Θ.
- The support of $\mathbb{P}_\theta$ does not depend on θ.

9.2 Fisher information

Define the log-likelihood for one observation as $\ell(\theta) = \log L(X, \theta)$ and assume ℓ is twice differentiable. Under some regularity conditions, the *Fisher information* is

$$I(\theta) = \operatorname{Var} \ell'(\theta) = -\mathbb{E}\left[\ell''(\theta)\right]$$

and, if $I(\theta) \neq 0$ in a neighborhood of θ^*, then the MLE is asymptotically normal with variance $I(\theta^*)^{-1}$.

Use it to construct the Wald test statistic for the MLE: $W = \sqrt{nI(\hat{\theta}^{\mathrm{MLE}})}(\hat{\theta}^{\mathrm{MLE}} - \theta^*)$.

9.3 M-estimation

Let $X_{i \in [n]}$ be i.i.d. with some unknown distribution $\mathbb{P}$ and associated parameter μ^* on a sample space E. An *M-estimator* $\hat{\mu}$ of μ^* is the minimizer of an estimator of a function $\mathcal{Q}(\mu)$ such that:

- $\mathcal{Q}(\mu) = \mathbb{E}\left[\rho(X, \mu)\right]$ for some function $\rho : E \times \mathcal{M} \to \mathbb{R}$, where $\mathcal{M}$ is the set of all possible values for μ^*.
- $\mathcal{Q}(\mu)$ attains a unique minimum at μ^*.

The goal is to find a loss function ρ that satisfies these properties. MLE is a special case of M-estimation where ρ is negative (log-)likelihood.

10 Linear regression

Solve $\min_\beta \|y - X\beta\|_2^2$ to get $\hat{\beta} = \left(X^\top X\right)^{-1} X^\top y$. If X is not full rank, regularize the objective by adding $\lambda\|\beta\|_p^2$ with hyperparameter $\lambda > 0$.

- If $p = 2$, this is ℓ_2 regularization that penalizes large values of β_j.
- If $p = 1$, this is ℓ_1 (lasso) regularization that prefers sparse β.

11 Generalized linear models

Relax the assumptions of linear regression: Assume that $Y|\, X = x$ is distributed according to some $\mathbb{P}$ and that $g(\mu(x)) = x^\top \beta$, where g is the *link function* and $\mu(x) = \mathbb{E}\left[Y|\, X = x\right]$ is the regression function.

k-parameter exponential family: A family of distributions $\left\{\mathbb{P}_\theta \,\middle|\, \theta \in \Theta \subset \mathbb{R}^k\right\}$ such that there exist real-valued functions $\eta_1, \eta_2, \dots, \eta_k$ and B of θ and $T_1, T_2, \dots, T_k$ and h of $y \in \mathbb{R}^q$ such that the density of $\mathbb{P}_\theta$ can be written

$$f_\theta(y) = \exp\left[\sum_{i=1}^k \eta_i(\theta) T_i(y) - B(\theta)\right] h(y)$$

The *canonical exponential family* ($k = 1$, $y \in \mathbb{R}$) for some known functions b and c is

$$f_\theta(y) = \exp\left[\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right]$$

If the *dispersion parameter* ϕ is known, then this is a one-parameter exponential family with θ the canonical parameter. It can be derived from log-likelihood that $\mathbb{E}\left[Y\right] = b'(\theta)$ and $\operatorname{Var} Y = b''(\theta)\phi$.

If g is monotone increasing and differentiable, then $\mu = g^{-1}\left(X^\top \beta\right)$. The *canonical link* is $g(\mu) = \theta = (b')^{-1}(\mu)$ for the canonical parameter θ.