Quick Reference of Probability

Hao Zhang

National Key Laboratory for Novel Software Technology, Nanjing University, China
zhangh0214@gmail.com

Abstract

In this note, we give a quick reference of some of the key concepts of counting, probablisitcs, and statistics. Besides, a section of balls and bins model is provided as an exercise. For further reading, you may consult [1, 2, 3, 4, 5, 6].

1 Mathematics Basis

1.1 Propositions

Definition 1 (Implies $P \Rightarrow Q$)**.** True exactly when P is false or Q is true, i.e., $P \Rightarrow Q \coloneqq \neg P \vee Q$.

Lemma 1. $\exists x, \forall y.\ P(x, y) \Rightarrow \forall y, \exists x.\ P(x, y)$.

Lemma 2 (The Well Ordering Principle)**.** *Every nonempty set of nonnegative integers has a smallest element.*

1.2 Sets and Binary Relations

Lemma 3. *The followings are properties of sets.*

- $S \subseteq A \wedge T \subseteq B \Rightarrow S \times T \subseteq A \times B$.
- ***DeMorgan's Law***. $\overline{A \cup B} = \bar{A} \cap \bar{B}$, $\overline{A \cap B} = \bar{A} \cup \bar{B}$.

Definition 2 (Binary Relations)**.** A binary relation $R : A \to B$ is a subset of $A \times B$. A is called the **domain**, and B is called **codomain**. The **image** of a set $S \subseteq A$ is the subset of B that is related to some element in S. The **range** is the image of A. There are kinds of relations.

- **Function**: ≤ 1 arrow out.
- **Total**: ≥ 1 arrow out.
- **Total function**: $= 1$ arrow out.
- **One-to-one/Injective**: ≤ 1 arrow in.
- **Onto/Surjective**: ≥ 1 arrow in.
- **One-to-one correspondence/Bijective**: $= 1$ arrow out and $= 1$ arrow in.

Definition 3 (Countable Infinite)**.** An infinite set that can be put into a one-to-one correspondence with the natural numbers $\mathbb{N}$; otherwise, it is **uncountable**.

2 Counting

2.1 Summations

Lemma 4. *The followings are some well-known series.*

- ***Telescoping series***. $\sum_{i=1}^{n-1} (a_{i+1} - a_i) = a_n - a_1$.
- $\sum_{i=1}^{n-1} \frac{1}{i(i+1)} = 1 - \frac{1}{n}$.
- ***Arithmetic series***. $\sum_{i=1}^{n} i = \frac{n(n+1)}{2} = \Theta(n^2)$.
- ***Sum of squares***. $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} = \Theta(n^3)$.
- ***Sum of cubes***. $\sum_{i=1}^{n} i^3 = \frac{n^2(n+1)^2}{4} = \Theta(n^4)$.
- ***Geometric series***. $\sum_{i=0}^{n} x^i = \frac{1-x^{n+1}}{1-x}$ *if* $x \neq 1$.
- ***Infinite geometric series***. $\sum_{i=0}^{\infty} x^i = \frac{1}{1-x}$ *if* $|x| < 1$.
- ***Harmonic series***. $\sum_{i=1}^{n} \frac{1}{i} = \log n + \Theta(1)$.

Lemma 5 (Integration Bound)**.** *Let* $f : \mathbb{R}_+ \to R_+$ *be a weakly increasing function,*

$$\int_1^n f(x)\,\mathrm{d}x + f(1) \leq \sum_{i=1}^{n} f(i) \leq \int_1^n f(x)\,\mathrm{d}x + f(n)\,. \quad (1)$$

Lemma 6 (Stirling's Formula)**.** $\forall n \geq 1.\ n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$.

Lemma 7 (Maclaurin's Theorem)**.** $f(x) = \sum_{i=0}^{\infty} \frac{1}{i!} \frac{\mathrm{d}^i f(0)}{\mathrm{d}x^i} x^i$.

Lemma 8. *The followings are some well-known series.*

- $\frac{1}{1-x} = \sum_{i=0}^{\infty} x^i$.
- $\exp x = \sum_{i=0}^{\infty} \frac{1}{i!} x^i$.
- $\exp cx = \sum_{i=0}^{\infty} \frac{c^n}{i!} x^i$.
- $\log(1 - x) = -\sum_{i=1}^{\infty} \frac{1}{i} x^i$.

2.2 Cardinality Rules

Lemma 9 (Rule of Sum)**.** *If we have* n_1 *ways to perform action 1 and* n_2 *ways to perform action 2, and we cannot do both at the same time, then there are* $n_1 + n_2$ *choose one of the actions.*

Lemma 10 (Rule of Product)**.** *If there are* n_1 *ways to perform action 1 and then by* n_2 *ways to perform action 2, then there are* $n_1 \cdot n_2$ *ways to perform action 1 followed by action 2.*

Lemma 11 (Pigenhole Principle)**.** *If $|A| > k|B|$, then for every total function $f : A \to B$, there exists $k+1$ different elements of set A that are mapped to the same element of B.*

Lemma 12 (Inclusion-Exclusion Principle)**.** $|A \cup B| = |A| + |B| - |A \cap B|$.

Definition 4 (Ordinary Generating Function)**.** The ordinary generating funciton $F(x)$ for the sequence $(f_i)_{i=0}^{\infty}$ is the power series $F(x) \coloneqq \sum_{i=0}^{\infty} f_i x^i$.

Lemma 13 (Convolution Rule)**.** *Let $F(x)$ be the generating function for selection items from a set A, and $G(x)$ for the set B. A are B are disjoint. The generating function for selecting items from the union $A \cup B$ is $F(x)G(x)$.*

2.3 Permutations and Combinations

Definition 5 (Permutation)**.** A permutation of a finite set is an ordered sequence of all the elements in that set, wich each element appearing exactly once.

Theorem 14. *The number of permutations of k elements out of a set of n elements is $\frac{n!}{(n-k)!}$. Specifically, the number of permutations of a set of n elements is $n!$.*

Definition 6 (Combination)**.** A combination of a finite set is a subset of that set.

Theorem 15. *The number of combinations of k elements out of a set of n elements is $\binom{n}{k} \coloneqq \frac{n!}{k!(n-k)!}$. This is also called the **binomial coefficient**. The followings are properties of binomial coefficient.*

- ***Binomial expansion**. $(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^{n-k} b^k$. In particular, $2^n = \sum_{k=0}^{n} \binom{n}{k}$.*
- ***Binomail bounds**. $\left(\frac{n}{k}\right)^k \le \binom{n}{k} \le \left(\frac{en}{k}\right)^k$.*
- ***Pascal's triangle identity**. $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$.*

Theorem 16. *The number of ways of depositing n distinct objects into r distinct bins, with k_1 objects in the first bin, k_2 objects in the second bin, and so on is*

3 Probability Basis

3.1 Terminologies

Definition 7 (Probability)**.** Mathematical language for quantifying uncertainty. In probability, there are a few logically self-contained rules. The random process is fully known, and the objective is to find the probability of a certain outcome. There is only one correct answer logically follow these rules.

Definition 8 (Statistics)**.** Using data to infer the distribution that generated the data. In statistics, the outcome is known, and we apply probability to make inference (illuminate the unknown random process) from experimental data. There is no single correct answer.

Definition 9 (Experiment)**.** A repeatable procedure with well-defined possible outcomes.

Definition 10 (Sample Space Ω)**.** Set of all possible **outcomes** ω of an experiment.

Definition 11 (Event E)**.** Subset of the sample space $E \subseteq \Omega$.

Definition 12 (Probability Function Pr)**.** A function $\Pr : \Omega \to [0,1]$ giving the probability for each outcome. Besides, Pr must also satisfies $\sum_{\omega \in \Omega} \Pr(\omega) = 1$.

Definition 13 (Probability of an Event)**.** The probability of an event E is the sum of the probabilities of all the outcomes in E, i.e., $\Pr(E) \coloneqq \sum_{\omega \in E} \Pr(\omega)$. Suppose Ω is finite and if each outcome is equally likely, then $\Pr(E) = \frac{|E|}{|\Omega|}$.

Lemma 17. *For two events E_1 and E_2, $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2) - \Pr(E_1 \cap E_2)$. Specifically, if E_1 and E_2 are disjoint, $\Pr(E_1 \cup E_2) = \Pr(E_1) + \Pr(E_2)$.*

3.2 Conditional Probability

Definition 14 (Conditional Probability $\Pr(A \mid B)$)**.** Probability of an event A given that another event B has occurred. If $\Pr(B) \neq 0, \Pr(A \mid B) \coloneqq \frac{\Pr(A \cap B)}{\Pr(B)}$. Furthermore, $\Pr(A \cap B) = \Pr(A \mid B)\Pr(B)$ is called the **multiplication rule**.

Trees are a great way to organize the underlying process into a sequence of actions. Each level of the tree shows the outcomes of one action. Conditional probabilities are written along the branches, so the probability of getting to any node is the product of the probabilities along the path to get there.

Theorem 18 (Law of Total Probability)**.** *Suppose $B_1, B_2, \dots, B_n$ are a partition of the sample space, then for any event A,*

$$\Pr(A) = \sum_{i=1}^{n} \Pr(A \cap B_i) = \sum_{i=1}^{n} \Pr(A \mid B_i)\Pr(B_i) \tag{2}$$

Definition 15 (Independent $A \perp B$)**.** Two event A and B are independent if knowing that B occurred does not change the probability that A occurred.

$$A \perp B \Leftrightarrow \begin{cases} \Pr(A \mid B) = \Pr(A) & \text{if } \Pr(B) \neq 0; \\ \Pr(B \mid A) = \Pr(B) & \text{if } \Pr(A) \neq 0. \end{cases} \tag{3}$$

In other words, $A \perp B \Leftrightarrow \Pr(A \cap B) = \Pr(A)\Pr(B)$. Note that disjoint events with positive probability are not independent.

Theorem 19 (Bayes' Theorem)**.** *For two events A and B,*

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A)} = \frac{\Pr(A \mid B)\Pr(B)}{\sum_i \Pr(A \mid B_i)\Pr(B_i)}. \tag{4}$$

Table 1: Distribution functions of random variables. We allow x to be any number. If x is a value that X never takes, then $p(x) = 0$.

	Range	Discrete	Continuous
PMF	$\mathbb{R} \to [0,1]$	$p(x) \coloneqq \Pr(X = x)$	-
PDF	$\mathbb{R} \to [0,\infty)$	-	$\Pr(a \le X \le b) = \int_a^b p(x)\,\mathrm{d}x$
CDF	$\mathbb{R} \to [0,1]$	$\Pr(X \le x) = \sum_{u=-\infty}^{x} p(u)$	$\Pr(X \le x) = \int_{-\infty}^{x} p(u)\,\mathrm{d}u$
Joint PMF	$\mathbb{R} \times \mathbb{R} \to [0,1]$	$p(x,y) \coloneqq \Pr(X = x \wedge Y = y)$	-
Joint PDF	$\mathbb{R} \times \mathbb{R} \to [0,\infty)$	-	$\Pr(a \le X \le b \wedge c \le Y \le d) = \int_c^d \int_a^b p(x,y)\,\mathrm{d}x\,\mathrm{d}y$
Joint CDF	$\mathbb{R} \times \mathbb{R} \to [0,1]$	$\sum_{u=-\infty}^{x} \sum_{v=-\infty}^{y} p(u,v)$	$\int_{-\infty}^{x} \int_{-\infty}^{y} p(u,v)\,\mathrm{d}u\,\mathrm{d}v$
Marginal PMF	$\mathbb{R} \to [0,1]$	$p(x) \coloneqq \sum_{y=-\infty}^{\infty} p(x,y)$	-
Marginal PDF	$\mathbb{R} \to [0,\infty)$	-	$p(x) \coloneqq \int_{-\infty}^{\infty} p(x,y)\,\mathrm{d}y$
Marginal CDF	$\mathbb{R} \to [0,1]$	$F(x) \coloneqq \lim_{y\to\infty} F(x,y)$	$F(x) \coloneqq \lim_{y\to\infty} F(x,y)$

Table 2: Properties of discrete expectation and variance.

	Expectation	Variance	Covariance
Notation	$\mathbb{E}[X], \mu$	$\operatorname{var} X, \sigma^2$	$\operatorname{cov}(X,Y)$
Meaning	Measure of central tendency	Measure of spread around the center	How much two random variables vary together
Definition (discrete)	$\sum_x xp(x)$	$\mathbb{E}[(X-\mu)^2] = \sum_x (x-\mu)^2 p(x)$	$\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$
Definition (continuous)	$\int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x$	$\mathbb{E}[(X-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,\mathrm{d}x$	$\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$
Alternative	-	$\operatorname{var} X = \mathbb{E}[X^2] - \mu^2$	$\operatorname{cov}(X,Y) = \mathbb{E}[XY] - \mu_X \mu_Y$
Scale and shift	$\mathbb{E}[aX+b] = a\,\mathbb{E}[X] + b$	$\operatorname{var}(aX+b) = a^2 \operatorname{var} X$	$\operatorname{cov}(aX+b, cY+d) = ac \operatorname{cov}(X,Y)$
Linearity	$\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$	$\operatorname{var}(X+Y) = \operatorname{var} X + \operatorname{var} Y + 2\operatorname{cov}(X,Y)$	$\operatorname{cov}(X_1+X_2, Y) = \operatorname{cov}(X_1,Y) + \operatorname{cov}(X_2,Y)$
Function of X	$\mathbb{E}[f(X)] = \sum_x f(x)p(x)$	-	-

4 Random Variables

4.1 Terminologies

Definition 16 (Discrete Random Variable X)**.** A discrete random variable is a function $X : \operatorname{pow}\Omega \to \mathbb{R}$. For any value x, we write $X = x$ to mean the event $\{\omega \in \Omega \mid X(\omega) = x\}$, so $\Pr(X = x) \coloneqq \Pr(\{\omega \in \Omega \mid X(\omega) = x\})$.

Table 1 summaries distribution functions of random variables.

Lemma 20 (Relationship Between CDF and PMF/PDF)**.**

$$p(x) = \begin{cases} F(x) - F(x-1) & \textit{if } X \textit{ is discrete}\,; \\ \dfrac{\mathrm{d}F(x)}{\mathrm{d}x} & \textit{if } X \textit{ is continuous}\,. \end{cases} \tag{5}$$

$$p(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}\,. \tag{6}$$

Table 2 summaries properties of expectation, variance, and covariance.

Lemma 21. *If $X_1, X_2, \dots, X_n$ are independent,*

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i]\,, \tag{7}$$

$$\operatorname{var}\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 \operatorname{var} X_i\,. \tag{8}$$

Definition 17 (Correlation ρ)**.** The correlation between X and Y is $\rho \coloneqq \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \in [-1,1]$. Furthermore,

$$\rho = 1 \Leftrightarrow \exists a > 0, b.\ Y = aX + b\,, \tag{9}$$

$$\rho = -1 \Leftrightarrow \exists a < 0, b.\ Y = aX + b\,. \tag{10}$$

Definition 18 (Independent $X \perp Y$)**.** Jointly-distributed random variables X and Y are independent if any event defined by X is independent of any event defined by Y. This is equivalent to $F(X,Y) = F(x)F(y)$ or $p(x,y) = p(x)p(y)$.

Lemma 22. *$X \perp Y \Rightarrow \operatorname{cov}(X,Y) = 0$. However, the converse if false, since $\operatorname{cov}(X,Y)$ only measures the linear relationship between X and Y.*

Definition 19 (Quantile)**.** The q-th quantile of X is

$$F^{-1}(q) \coloneqq \operatorname*{arg\,min}_x F(x) \le q = \operatorname*{arg\,min}_x \Pr(X \le x) \le q\,. \tag{11}$$

In particular, we call $F^{-1}\left(\frac{1}{4}\right)$ the first quantile, $F^{-1}\left(\frac{1}{2}\right)$ the median, and $F^{-1}\left(\frac{3}{4}\right)$ the third quantile.

4.2 Common Distributions

Table 3 summaries common distributions.

Table 3: Common discrete distributions.

Distribution	Model	Notation	PMF/PDF	CDF	Domain	Expect.	Variance
Bernoulli	One trial resulting in success/failure	$\mathrm{Ber}(p)$	$p^x(1-p)^{1-x}$	-	$\{0,1\}$	p	$p(1-p)$
Binomial	# successes in n independent $\mathrm{Ber}(p)$ trials	$\mathrm{Bin}(n,p)$	$\binom{n}{x}p^x(1-p)^{1-x}$	-	$\{0,\dots,n\}$	np	$np(1-p)$
Geometric	# failures before the first success of indep. $\mathrm{Ber}(p)$ trials	$\mathrm{Geo}(p)$	$p(1-p)^x$	-	$\mathbb{N}$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$
Poisson	Count for rare events like radio. decay and traf. accidents	$\mathrm{Poi}(\lambda)$	$\exp(-\lambda)\frac{\lambda^x}{x!}$	-	$\{0,1,\dots\}$	λ	λ
Uniform	Situation when all the outcomes are equally likely	$\mathrm{Unif}(a,b)$	$\frac{1}{b-a+1}$	-	$\{a,\dots,b\}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
Gaussian	Measurement error, averages of lots of data, etc.	$\mathcal{N}(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$		$\mathbb{R}$	μ	σ^2
Exponential	Lifetime of elect. comp./waiting time between rare events	$\mathrm{Exp}(\beta)$	$\beta\exp(-\beta x)$	$1-\exp(-\beta x)$	$[0,\infty)$	$\frac{1}{\beta}$	$\frac{1}{\beta^2}$
Uniform	Situation when all the outcomes are equally likely	$\mathrm{Unif}(a,b)$	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
Pareto	Prob. that an event occurs varies according to power law	$\mathrm{Pareto}(m,\alpha)$	$\frac{\alpha m^\alpha}{x^{\alpha+1}}$	$1-\frac{m^\alpha}{x^\alpha}$	$[m,\infty)$		

Table 4: Common inequalities, where $\bar{X}$ is the sample mean, and μ and σ^2 are the mean and variance, respectively.

Name	Assumption	Parameter	Inequality
Gaussian tail inequality	$X\sim\mathcal{N}(0,1)$	$\varepsilon>0$	$\Pr(\lvert X\rvert>\varepsilon)\le\frac{2}{\varepsilon}\exp\left(-\frac{\varepsilon^2}{2}\right).$
Gaussian tail inequality	$X_1,X_2,\dots,X_m$ are iid drawn from $\mathcal{N}(0,1)$	$\varepsilon>0$	$\Pr(\lvert\bar{X}\rvert>\varepsilon)\le\frac{2}{\varepsilon\sqrt{m}}\exp\left(-\frac{\varepsilon^2 m}{2}\right)$
Markov's inequality	X is a non-negative	$\varepsilon>0$	$\Pr(X>\varepsilon)\le\frac{\mu}{\varepsilon}$
Chebyshev inequality	-	$\varepsilon>0$	$\Pr(\lvert X-\mu\rvert\ge\varepsilon)\le\frac{\sigma^2}{\varepsilon^2}$
Hoeffding's inequality	$\mu_i=0$ and $a_i\le X_i\le b_i$	$\varepsilon>0,t>0$	$\Pr\left(\sum_{i=1}^m X_i\ge\varepsilon\right)\le\exp(-t\varepsilon)\prod_{i=1}^m\exp\frac{t^2(b_i-a_i)^2}{8}$
Hoeffding's inequality	$X_1,X_2,\dots,X_m$ iid drawn from $\mathrm{Ber}(p)$	$\varepsilon>0$	$\Pr(\lvert\bar{X}-p\rvert>\varepsilon)\le 2\exp(-2\varepsilon^2 m)$

Lemma 23. *Geometric distribution and exponential distribution are memoryless,*

$$\Pr(X=x+x_0\mid X\ge x_0)=\Pr(X=x)\,. \tag{12}$$

Definition 20 (Multivariate Gaussian Distribution)**.** $\boldsymbol{X}\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ if

$$p(\boldsymbol{x})=\frac{1}{\sqrt{(2\pi)^d\det\boldsymbol{\Sigma}}}\exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{13}$$

Definition 21 (Multinomial Distribution)**.** $\boldsymbol{X}\sim\mathrm{Mult}(n,\boldsymbol{p})$ if

$$p(\boldsymbol{x})=\binom{n}{x_1,x_2,\dots,x_K}\prod_{i=1}^K p_i^{x_i}\,. \tag{14}$$

It models a situation where drawing n balls from an urn with K differnet colors. p_i is the probability of drawing color i, and x_i is the count of the number of balls for color i. $\binom{n}{x_1,x_2,\dots,x_r}\coloneqq\frac{n!}{\prod_{i=1}^K x_i!}$ is also called the **multinomial coefficient**.

Lemma 24 (Multinomail Expansion)**.**

$$\left(\sum_{i=1}^r a_i\right)^n=\sum_{x_1+\cdots+x_r=n}\binom{n}{x_1,x_2,\dots,x_r}\prod a_i^{x_i}\,. \tag{15}$$

5 Inequalities and Convergence

5.1 Inequalities

Table 4 summaries common inequalities.

Theorem 25 (Cauchy-Schwarz Inequality)**.**

$$\mathbb{E}[\lvert XY\rvert]\le\sqrt{\mathbb{E}[X^2]\,\mathbb{E}[Y^2]}\,. \tag{16}$$

Theorem 26 (Jensen's Inequality)**.** *If f is convex,*

$$\mathbb{E}[f(X)]\ge f(\mathbb{E}[X])\,. \tag{17}$$

5.2 Convergence of Random Variables

Theorem 27 (Law of Large Numbers)**.** *Suppose $X_1,X_2,\dots,X_m,\dots$ are i.i.d. random variables with mean μ and variance σ^2. Let $\bar{X}_m\coloneqq\frac{1}{m}\sum_{i=1}^m X_i$ be the sample mean. Then*

$$\forall\varepsilon>0.\ \lim_{m\to\infty}\Pr(\lvert\bar{X}_m-\mu\rvert<\varepsilon)=1\,. \tag{18}$$

In other words, the sample mean is very close to the true mean of the distribution with high probability.

Corollary 28 (Law of Large Numbers for Histograms)**.** *With high probability the density histogram of a large number of samples from a distribution is a good approximation of the graph of the underlying PDF p.*

Theorem 29 (Central Limit Theorem)**.** *Suppose $X_1, X_2, \ldots, X_m, \ldots$ are i.i.d. random variables with mean μ and variance σ^2. Let $\bar{X}_m := \frac{1}{m}\sum_{i=1}^{m} X_i$ be the sample mean. Then for large m,*

$$\bar{X}_m \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{m}\right), \tag{19}$$

$$\frac{\bar{X}_m - \mu}{\sigma/\sqrt{m}} \sim \mathcal{N}(0,1)\,. \tag{20}$$

6 Standard Model: Balls and Bins

6.1 Distinct Balls

Consider the process that we random toss n distinct balls into k bins.

Lemma 30. *Suppose the order within a bin does not matter. The number of ways of placing the balls into bins is k^n.*

Proof. There are k ways to place each of the n balls, respectively. □

Lemma 31. *Suppse the balls in each bin are ordered. The number of ways of placing the balls into bins is $\frac{(n+k-1)!}{(k-1)!}$.*

Proof. The number of ways of arranging n balls and $k-1$ sticks (as separators) in a row is $(n+k-1)!$. Since sticks are indistinguishable, $(k-1)!$ of the arrangements are duplicated. Thus the answer is $\frac{(n+k-1)!}{(k-1)!}$. □

Lemma 32 (Birthday Paradox)**.** *The probability that at least two balls are tossed into the same bin is $\geq 1 - \exp\left(-\frac{n(n-1)}{2k}\right)$.*

Proof. By looking at the complementary event. The probability that all balls are tossed into different bins is

$$\begin{aligned}
\frac{\binom{k}{n}n!}{k^n} &= \frac{k(k-1)(k-2)\cdots(k-n+1)}{k^n} \\
&= (1-\frac{0}{k})(1-\frac{1}{k})(1-\frac{2}{k})\cdots(1-\frac{n-1}{k}) \\
&< \exp(0)\exp(-\frac{1}{k})\exp(-\frac{2}{k})\cdots\exp(-\frac{n-1}{k}) \\
&= \exp\left(-\sum_{i=0}^{n-1}\frac{i}{k}\right) \\
&= \exp\left(-\frac{n(n-1)}{2k}\right)\,. \qquad (21)
\end{aligned}$$

When $n = \sqrt{2k}$, it approximates $\frac{1}{\mathrm{e}}$. □

Lemma 33. *The expected number of pairs of balls within the the same bin is $\frac{n(n-1)}{2k}$.*

Proof. Let $X_{ij} = \mathbb{I}(\text{ball } i \text{ and } j \text{ are tossed into the same bin})$. Then

$$\mathbb{E}\left[\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_{ij}\right] = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\mathbb{E}[X_{ij}] = \binom{n}{2}\frac{1}{k} = \frac{n(n-1)}{2k}\,. \tag{22}$$

□

6.2 Identical Balls

Consider the process that we randomly toss n identical balls into k bins.

Lemma 34. *The number of ways of placing the balls into bins is $\binom{n+k-1}{n}$.*

Proof. If balls are distinct, there are $\frac{(n+k-1)!}{(k-1)!}$ ways. Since balls are indistinguishable, $n!$ of the arrangements are duplicated. Thus the answer is $\frac{(n+k-1)!}{(k-1)!n!} = \binom{n+k-1}{n}$. □

Lemma 35. *Suppose no bin may contain more than one ball, so that $n \leq k$. The number of ways of placing the balls is $\binom{K}{n}$.*

Proof. Select n out of the k bins to put balls in. □

Lemma 36. *Suppose no may be left empty. Assumging that $n \geq k$, the number of ways of placing the balls is $\binom{n-1}{k-1}$.*

Proof. First, we put a ball in each bin. Then we use the result of Lemma 34. $\binom{n-k+k-1}{n-k} = \binom{n-1}{n-k} = \binom{n-1}{n-1-n+k} = \binom{n-1}{k-1}$. □

Lemma 37. *The expect number of balls fall in a given bin is $\frac{n}{k}$.*

Proof. It follows a binomial distribution $\mathrm{Bin}(n, \frac{1}{k})$. □

Lemma 38. *The expected number of balls we toss until a given bin contains a ball is k.*

Proof. It follows a geometric distribution $\mathrm{Geo}(\frac{1}{k})$. □

Lemma 39 (Coupon Collector's Problem)**.** *The expected number of balls we toss until every bin contains at least one ball is $\sim k\log k$.*

Proof. Partition the tosses into k stages. The i-th stage contains of the tosses after tossing a ball into the empty $(i-1)$-th bin until tossing a ball into the empty i-th bin. During the i-th stage, $i-1$ bins contain balls and $k-i+1$ bins are empty. It follows a geometric distribution $\mathrm{Geo}(\frac{k-i+1}{k})$. The expected number of tosses is $\frac{k}{k-i+1}$. The overall expected number of tosses is

$$\sum_{i=1}^{k}\frac{k}{k-i+1} = k\sum_{i=1}^{k}\frac{1}{i} \sim k\log k\,. \tag{23}$$

□

7 Statistics Basis

7.1 Terminologies

Definition 22 (Point Estimation $\hat{\theta}$)**.** Provide a single "best guess" of some quantity θ of interest. We denote the **bias** of $\hat{\theta}$ as $\mathbb{E}[\hat{\theta}] - \theta$. We say that $\hat{\theta}$ is **unbiased** if $\mathbb{E}[\hat{\theta}] = \theta$, and **consistent** if $\forall \epsilon > 0$. $\lim_{m \to \infty} \Pr(|\hat{\theta} - \theta| < \epsilon) = 1$. We denote the **mean square error** of $\hat{\theta}$ as $\mathbb{E}[(\hat{\theta} - \theta)^2]$.

Definition 23 (Sample Mean $\bar{X}$)**.** $\bar{X} := \frac{1}{m} \sum_{i=1}^m X_i$. Sample mean is unbiased $\mathbb{E}[\bar{X}] = \mu$, and var $\bar{X} = \frac{\sigma^2}{m}$.

Definition 24 (Sample Variance s^2)**.** $s^2 := \frac{1}{m-1} \sum_{i=1}^m (X_i - \bar{X})^2$. Sample variance is unbiased $\mathbb{E}[s^2] = \sigma^2$.

Definition 25 (Empirical CDF $\hat{F}$)**.** Suppose $X_1, X_2, \dots, X_m$ are iid drawn from a distribution, the empirical CDF $\hat{F}$ puts mass $\frac{1}{m}$ at each data point X_i, $\hat{F}(x) = \frac{1}{m} \sum_{i=1}^m \mathbb{I}(X_i \leq x)$. $\hat{F}$ is an unbiased and consistent estimation of the true CDF F.

7.2 Parametric Inference

Theorem 40 (Maximum Likelihood Estimate, MLE)**.** *Let $X_1, X_2, \dots, X_m$ be iid with pdf p, the maximum likelihood estimate for a parameter distribution is the parameter for which the data is most likely,*

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{m} p(X_i; \theta) = \arg\max_{\theta} \log \sum_{i=1}^{m} p(X_i; \theta) \,. \qquad (24)$$

MLE is a point estimation and it is asymptotically unbiased and has asymptotically minimal variance.

Theorem 41 (Bayes Estimate)**.** *Bayes estimate regards θ as random,*

$$p(\theta \mid D) = \frac{p(D \mid \theta) p(\theta)}{p(D)} \,. \qquad (25)$$

We can then compute a point estimation from the posterior, e.g.,

$$\hat{\theta} = \mathbb{E}[\theta \mid D] = \int \theta p(\theta \mid D) \, \mathrm{d}\theta \,. \qquad (26)$$

Definition 26 (Frequentists)**.** Frequentists say that probability measures the frequency of various outcomes of an experiment. For example, saying a fair coin has a 50% probability of heads means that if we toss it many times then we expect about half the tosses to land heads.

Definition 27 (Bayesians)**.** Bayesians say that probability is an abstract concept that measures a state of knowledge or a degree of belief in a given proposition. In practice, Bayesians do not assign a single value for the probability of a coin coming up heads. Rather they consider a range of values each with its own probability of being true.

References

[1] G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002. 1

[2] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms, 3rd Edition*. MIT Press, 2009. 1

[3] R. Hogg, J. McKean, and A. Craig. *Introduction to Mathematical Statistics. 7th Edition*. Pearson Press, 2012. 1

[4] R. J. Larsen and M. L. Marx. *Introduction to Mathematical Statistics and Its Applications: Pearson New International Edition*. Pearson Higher Ed, 2013. 1

[5] E. Lehman, T. Leighton, and A. Meyer. *Mathematics for Computer Science*. MIT, 2017. 1

[6] L. Wasserman. *All of statistics: A Concise Course in Statistical Inference*. Springer Science & Business Media, 2013. 1