Brief Theory of Probability: Notes from MATH 431

Compiled by Harry Luo

Contents

1 Sample Spaces, collection of events, probability measure

- Sample space Ω: set of all possible outcomes of an experiment. Comes in n-tuples where n represents number of repeated trials.
- Collection of events $\mathcal{F}$: subset of state space to which we assign a probability.
- Probability measure: function that assigns a probability to each event.$P : F \to \mathbb{R}$.
  - Range is $[0, 1]$.
  - $P(\Omega) = 1$ and $P(\emptyset) = 0$
  - For pairwise disjoint events $A_1, A_2, \ldots$, $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$

---

2 Sampling: Uniform, Replacement, Order

- uniform sampling: each outcome is equally likely
- Binomial coeff

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1}$$

2.1 Replacement

- ex: sample K distinct marked balls from N balls in a box, **with** Replacement

$$\begin{gathered} \Omega = \{1, 2, 3, \ldots, N\}^K \\ \|\Omega\| = N^K \\ P(\text{none of the balls is marked 1}) = \frac{(N-1)^K}{N^K} \end{gathered} \tag{2}$$

- ex: sample K distinct marked balls from N balls in a box, **without** Replacement

$$\begin{gathered} \Omega = \{(i_1, i_2, \ldots, i_K) \mid i_1, \ldots, i_K \in \{1, 2, \ldots, N\}, \text{distinct} \\ \|\Omega\| = \binom{N-1}{K} \\ P(\text{none of the balls is marked 1}) = \frac{\binom{N-1}{K}}{\binom{N}{K}} = \frac{N-K}{N} \end{gathered} \tag{3}$$

2.2 Order

- order matters: $A_n^k = \frac{n!}{(n-k)!}$
- order doesn't matter: $\binom{n}{k} = C_n^k = \frac{n!}{k!(n-k)!}$

---

3 Infinite Sample Spaces

3.1 discrete

$$\Omega = \{\infty, 1, 2, \ldots\} \tag{4}$$

3.2 continuous

$$\begin{gathered} P([a', b']) = \frac{\text{length of } [a', b']}{\text{length of}[a, b]} \\ \text{single point, or sets of points: } P(\{x\}) = P(\cup_{i=1}^{\infty} \{x_i\}) = 0 \end{gathered} \tag{5}$$

- Complements: $P(A) = 1 - P\left(A^C\right)$

---

4 Conditioinal Probability, Law of Total Prob., Bayes' Theorem, Independence

4.1 Conditional prob.

$$P(A|B) = \frac{|A \cap B|}{|B|} \Rightarrow P(AB) = P(B)P(A|B) \tag{6}$$

(new sample space is B, total number of outcomes is $A \cap B$)

4.2 Law of total probability:

Given partitions $B_1, B_2, \ldots$ of Ω,

$$P(A) = \sum_i P(A|B_i)P(B_i) \tag{7}$$

4.3 Bayes' Theorem:

Given events A, B, P(A) and P(B) >0,

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} \tag{8}$$

Considering the law of total prob., the generalized form, when B_i are partitions, is given as:

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P\left(A|B_j\right)P\left(B_j\right)} \tag{9}$$

4.4 Independence:

$$P(AB) = P(A)P(B) \Leftrightarrow P(B|A) = P(B) \tag{10}$$

Note: By virtue of conventions, we write $A \cap B$ as AB in Probability.
If A,B,C,D are independent, it follows that $P(ABCD) = P(A)P(B)P(C)P(D)$; however, the inverse is not always true.

- Independence of Random Variables (messy as hell...)

Given 2 random variables

$$\begin{gathered} X_1 \in \{x_{11}, x_{12}, x_{13}, \ldots, x_{1m}\} \\ X_2 \in \{x_{21}, x_{22}, x_{23}, \ldots, x_{2n}\} \\ \text{Random variables X\_1 and X\_2 are independent} \Leftrightarrow \\ P\left(X_1 = x_{1i}, X_2 = x_{2j}\right) = P(X_1 = x_{1i})P\left(X_2 = x_{2j}\right) \end{gathered} \tag{11}$$

Need to check n*m equations to verify independence.

4.5 Conditional Independence:

For events $A_1, A_2, \ldots, A_n, B$, any set of events in A: A_{i1}, A_{i2}, A_{i3}, they are conditionally independent given B if

$$P(A_{i1}A_{i2}A_{i3}|B) = P(A_{i1}|B) * P(A_{i2}|B) * P(A_{i3}|B) \tag{12}$$

5 Independent Trials, Distributions

5.1 Bernoulli dirtribution:

a single trial, with success probability p, and failure probability 1-p. Prameter being the success probability.

$$X \sim \text{Ber}(p) \Rightarrow P(X = x) = p^x * (1 - p)^{1-x}, x \in \{0, 1\} \quad (13)$$

5.2 Binomial Distribution:

multiple independent Bernoulli trials, with success probability p, and failure probability 1-p. Parameters being the number of trials n and the success probability p.

$$X \sim \text{Bin}(n, p) \Rightarrow P(X = k) = \binom{n}{k} p^k * (1 - p)^{n-k}, k \in \{0, 1, ..., n\} \quad (14)$$

5.3 Geometric distribution:

multiple independent Bernoulli trials with success probability p, while stoping the experiment at the first success.

$$X \sim \text{Geom}(p) = p * (1 - p)^{k-1}, k \in \{1, 2, ...\} \quad (15)$$

5.4 Hypergeometric distribution:

There are N objects of type A, and $N_A - N$ objects of type B. Pick n objects without replacement. Denote number of A objects we picked as k. Parameters are N, N_A, n.

$$P(X = k) = \frac{\binom{N_A}{k}\binom{N-N_A}{n-k}}{\binom{N}{n}} \quad (16)$$

choose k from N_A, choose n-k from N-N_A, divide by total number of ways to choose n from N

6 Random Variables

Properties of Random Variables	
Discrete	Continuous
Probability mass function $p_X(k) = P(X = k)$	Probability density function $f_X(x)$
$P(X \in B) = \sum_{k: k \in B} p_X(k)$	$P(X \in B) = \int_B f_X(x) dx$
Cumulative distribution function $F_X(a) = P(X \leq a)$	
$F_X(a) = \sum_{k: k \leq a} p_X(k)$ F_X is a step function.	$F_X(a) = \int_{-\infty}^{a} f(x)\, dx$ F_X is a continuous function.
$P(X < a) = \lim_{t \to a^-} F(t) = F(a-)$ $P(X = a) = F(a) - \lim_{t \to a^-} F(t) = F(a) - F(a-)$	
$E(X) = \sum_k k\, p_X(k)$	$E(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx$
$E(aX + b) = aE[X] + b$	
$E[g(X)] = \sum_k g(k)\, p_X(k)$	$E[g(X)] = \int_{-\infty}^{\infty} g(x) f(x)\, dx$
$\text{Var}(X) = E\left[(X - E[X])^2\right] = E[X^2] - (E[X])^2$	
$\text{Var}(aX + b) = a^2\, \text{Var}(X)$	

6.1 Discrete random variable

Discrete random variables are random variables that can take on a countable number of values. It comes naturally from discrete, finite or infinitly countable sample spaces. (As briefly discussed in Section 3.1)

For $A = \{k_1, k_2, \ldots, \}$ s.t. random variable$X \in A$, or $P(X \in A) = 1$, X is a random variable, with possible values $k_1, k_2, \ldots$ and $P(X = k_n) > 0$

6.1.1 Probability Mass Function (pmf)

The PMF is a function that defines the probability distribution for a discrete random variable. It gives the probability of the random variable taking on each possible value. The PMF, denoted as

$$p_X(k) = P(X = k) \text{ , where } k \text{ are possible values of X} \tag{17}$$

It is a function of k, and

$$p_X : S \to [0, 1], \tag{18}$$

where:

S is the support set, i.e., the set of all possible values that the discrete random variable X can take. [0, 1] represents the range of the function, as probabilities are always between 0 and 1. For each value k in the support set S, the PMF assigns a probability p_X(k), which represents the likelihood of the random variable X taking the value k.

The PMF satisfies the following properties:

Non-negativity: $p_{X(k)} \geq 0$ for all k in S.
Total probability: $\sum_k p_{X(k)} = 1$ where the sum is taken over all k in S.

Example: For a fair six-sided die, the PMF would be $P(X = x) = \frac{1}{6}$ for $x = 1, 2, 3, 4, 5, 6$. Or more elegantly,

$$p_X(k) = \frac{1}{6}, \text{for every } k \in \{1, 2, 3, 4, 5, 6\} \tag{19}$$

6.2 continuous Random Variables

Not rigorously defined in this class, but a continuous random variable is one that can take on any value in a range. The probability of a continuous random variable taking on a specific value is 0. It came natually from continuous sample spaces.The probability is assigned to intervals of values, and they are assigned by the **probability density function**.

6.2.1 Probability Density Function (pdf)

continuous r.v are defined in this class by having a probability density function. A random variable X is continuous if there exists a function f(x) such that

$$\begin{gathered} \int_{-\infty}^{\infty} f(x) \,\mathrm{d}x = 1, f(x) > 0 \text{ everywhere} \\ \text{and } P(X \leq b) = \int_{-\infty}^{b} f(x) \,\mathrm{d}x \Leftrightarrow P(a \leq X \leq b) = \int_{a}^{b} f(x) \,\mathrm{d}x \end{gathered} \tag{20}$$

6.2.2 Cumulative Distribution Function (cdf)

cdf of a r.v. is defined as

$$F(x) = P(X \leq x) \tag{21}$$

and it follows that

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \tag{22}$$

- Continuous r.v.

  it looks suspiciously like an indefinite integral, and when we are dealing with continuous r.v., it is.

  $$\boxed{F(s) = P(X \leq s) = \int_{-\infty}^{s} f(x) \,\mathrm{d}x}$$

  Recall the fundamental theorm of calculus,

  $$F'(x) = f(x), \tag{23}$$

  so the pdf is the derivative of the cdf.

- Discrete r.v.

  pmf and cdf is connected by

  $$F(x) = P(X \leq s) = \sum_{k \leq x} p_{X(k)} \tag{24}$$

where the sum is taken over all k such that $k \leq x$.
In english, the cdf is the sum of the pmf up to the value x, or "compound probability thus far"

If the cdf graph is stepped (piecewise constant), it is a discrete r.v. If it is continuous except at several points, it is a continuous r.v.

6.3 Expectation and Variance

6.3.1 Expectation

1. Exp of discrete r.v. is defined as

$$E(X) = \sum_k kP(X = k) \tag{25}$$

where the sum is taken over all possible values of X. It is the weighted average of the possible values of X, where the weights are given by the probabilities.

Expectation is a linear operator, i.e.

$$E(aX + b) = aE(X) + b \tag{26}$$

for any constants a and b.

- exp of **Bernoulli** r.v. is

$$E(X) = p \tag{27}$$

where p is the probability of success.

- exp of **binomial** r.v. is

$$E(X) = np \tag{28}$$

where n is the number of trials and p is the probability of success.

- exp of **geometric** r.v. is

$$E(X) = \frac{1}{p} \tag{29}$$

where p is the probability of success.

1. Exp of continuous r.v. is defined as

$$E(X) = \int_{-\infty}^{\infty} x f(x) \,\mathrm{d}x \tag{30}$$

where the integral is taken over the entire range of possible values of X. It is the weighted average of the possible values of X, where the weights are given by the probability density function.

- exp of **uniform** r.v. is

$$E(X) = \frac{a + b}{2} \tag{31}$$

where a and b are the lower and upper bounds of the interval.

6.3.2 Expectation of a function of a random variable

When we have a function of a random variable, we can find the expectation of that function by applying the function to each possible value of the random variable and taking the weighted average of the results.

- if X is a discrete r.v. with pmf p_X(k), and g is a function of X, then

$$E(g(X)) = \sum_k g(k) p_{X(k)} \tag{32}$$

- if X is a continuous r.v. with pdf f(x), and g is a function of X, then

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f(x) \,\mathrm{d}x \tag{33}$$

6.3.3 Moments, and moment generating function

1. The **nth moment** of the random variable X is the expectation $E(X^n)$.
   - X as discrete r.v. with pmf p_X(k), the nth moment is

   $$E(X^n) = \sum_k k^n p_{X(k)} \tag{34}$$

   - X as continuous r.v. with pdf f(x), the nth moment is

   $$E(X^n) = \int_{-\infty}^{\infty} x^n f(x) \,\mathrm{d}x \tag{35}$$

2. The **moment generating function** of a
   - discrete random variable X is defined as

   $$M_X(t) = E(e^{tX}) = \sum_k e^{tk} p_{X(k)} \tag{36}$$

   - continuous random variable X is defined as

   $$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x) \,\mathrm{d}x \tag{37}$$

   It is a function of t.

   We can easily find the nth moment of X by taking the nth derivative of the moment generating function with respect to t and evaluating it at t = 0. i.e.

   $$E(X^n) = \frac{\mathrm{d}}{\mathrm{d}t} M_X(t = 0) \tag{38}$$

6.3.4 Variance

The variance of a random variable X is a measure of how much the values of X vary around the mean. It is defined as the expectation of the squared deviation of X from its mean. i.e.

$$\sigma^2 = \mathrm{Var}(X) = E\Big((X - E(X))^2\Big) \tag{39}$$

alternatively,

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2 \tag{40}$$

Variance is not a linear operator, i.e.

$$\mathrm{Var}(aX + b) = a^2 \ \mathrm{Var}(X) \tag{41}$$

for any constants a and b.

1. variance of bournoli r.v. is

$$p(1-p) \tag{42}$$

2. variance of binomial r.v. is

$$np(1-p) \tag{43}$$

3. variance of geometric r.v. is

$$\frac{1-p}{p^2} \tag{44}$$

4. variance of uniform r.v. is

$$\frac{(b-a)^2}{12} \tag{45}$$

7 continuous Distribution

Based on different pdf, we have different behaviors of random variables. We call them distributions.

7.1 Uniform Distribution

r.v. X has the uniform distribution on the interval [a,b] if its pdf is

$$f(x) = \begin{cases} \frac{1}{b-a} \text{ for } a \leq x \leq b \\ 0 \text{ otherwise} \end{cases} \tag{46}$$

7.2 Normal (Gaussian) Distribution

7.2.1 standard normal distribution

r.v. Z has the Standard normal distribution if its pdf is

$$f(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{47}$$

where z is the standard normal r.v. and phi is the standard normal pdf. It's abbrieviated as $Z \sim N(0, 1)$ where 0 is the mean and 1 is the variance.

- The **cdf** of the standard normal distribution is denoted as

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} \varphi(z) \, \mathrm{d}z \tag{48}$$

  Check for table for values of $\Phi(z)$

7.2.2 normal distribution (generalized)

two parameters: the mean μ and the variance σ^2 . The pdf of a normal distribution is given by the formula:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{49}$$

abbrieviated as $X \sim N(\mu, \sigma^2)$

- Linearity of normal distribution

  If $X \sim N(\mu, \sigma^2), Y = aX + b$, then $Y \sim N(a\mu + b, a^2\sigma^2)$

- **normalization of normal distribution** For $X \sim N(\mu, \sigma^2)$, we can standardize it to $Z \sim N(0,1)$ by $Z = \frac{X-\mu}{\sigma}$

8 Approximations of Binomial Distribution

Recall: **Binomial distribution** is the distribution of the *number of successes* of n independent Bernoulli trials. It has two parameters: the number of trials n and the probability of success p.

Depending on the probability of success p and the number of trials n, the binomial distribution can be approximated by the normal distribution or the Poisson distribution.

8.1 Central limit theorem (approximation with normal distribution)

If n is large and p is not too close to 0 or 1, the binomial distribution can be approximated by the normal distribution.

For $S_n \sim \text{ Bin}(n, p)$; $E(S_n) = np$, Var $(S_n) = \sigma^2 = np(1-p)$,

$$\lim_{n \to \infty} P\left(a \leq \frac{S_n - \mu}{\sigma} \leq b\right) = \int_a^b \varphi(x) \, \mathrm{d}x = \Phi(b) - \Phi(a) \tag{50}$$

where phi is the standard normal pdf. This is the central limit theorem, which states that the binomial random variables approaches a normal distribution when $np(1-p) > 10$.

8.1.1 continuity correction

$$P(a \leq S_n \leq b) = P(a - 0.5 \leq S_n \leq b + 0.5) \tag{51}$$

where S Bin(n,p) and a,b are integers. Useful when a,b are close, and np(1-p) is not large.

8.1.2 Law of large numbers

For

$$\begin{gathered} S_n \sim \text{ Bin}(n, p) \ ; \ E(S_n) = np, E\left(\frac{S_n}{n}\right) = p \\ P\left(|\frac{S_n}{n} - p| < \varepsilon\right) \to 1 \text{ as } n \to \infty \end{gathered} \tag{52}$$

In English, this is saying that, as n is large, the frequency of success in n trials will converge to the probability of success p.

8.1.3 Confidence interval

In most cases, if real probability of success is unknown, we can use the Law of large number to

1. approximate p
2. find confidence interval $(\hat{p} - \varepsilon, \hat{p} + \varepsilon)$ (know how accurate the approximation is.) Connecting law of large number with CLT, we can proof that

$$P(|\hat{p} - p| < \varepsilon) \geq 2\Phi(2\varepsilon\sqrt{n}) - 1 \tag{53}$$

   where, $2\Phi(2\varepsilon\sqrt{n}) - 1$ is the confidence level, i.e. how confident we are that the real probability is in the interval.

8.2 Poisson Distribution

8.2.1 Poisson r.v.