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 $[0, 1]$.
- Axioms
  - $P(\Omega) = 1$ and $P(\emptyset) = 0$
  - For pairwise disjoint events $A_1, A_2, ...,$ $P(A_1 \cup A_2 \cup ...) = P(A_1) + P(A_2) + ...$

---

2 Sampling

2.1 Uniform sampling

If the sample space Ω has finitely many elements and each outcome is equally likely, then for any event $A \subset \Omega$ we have

$$P(A) = \frac{\#A}{\#\Omega} \tag{1}$$

where # means the "cardinality" of the set.

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

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{2}$$

2.2 Sampling with Replacement, order matters

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

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

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

$$\begin{gathered} \Omega = \{(i_1, i_2, ..., i_K) \mid i_1, ..., i_K \in \{1, 2, ..., N\}, \text{distinct} \\ \|\Omega\| = \binom{N-1}{K} \\ P(\text{none of the balls is marked 1}) = \frac{\binom{N-1}{K}}{\binom{N}{K}} = \frac{N-K}{N} \end{gathered} \tag{4}$$

2.3 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

$$\Omega = \{\infty, 1, 2, ...\} \tag{5}$$

3.2 continuous

$$P([a',b']) = \frac{\text{length of } [a',b']}{\text{length of}[a,b]} \tag{6}$$
$$\text{single point, or sets of points: } P(\{x\}) = P(\cup_{i=1}^{\infty}\{x_i\}) = 0$$

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

---

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

4.1 Conditional prob.

$$P(A|B) = \frac{|A \cap B|}{|B|} \Rightarrow P(AB) = P(B)P(A|B) \tag{7}$$

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

4.2 Law of total probability:

Given partitions $B_1, B_2, ...$ of Ω,

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

4.3 Bayes' Theorem:

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

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} \tag{9}$$

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

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P\big(A|B_j\big)P\big(B_j\big)} \tag{10}$$

4.4 Independence:

$$P(AB) = P(A)P(B) \Leftrightarrow P(B|A) = P(B) \tag{11}$$

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

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

Need to check n*m equations to verify independence.

4.5 Conditional Independence:

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

$$P(A_{i1}A_{i2}A_{i3}|B) = P(A_{i1}|B) * P(A_{i2}|B) * P(A_{i3}|B) \tag{13}$$

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.

$$X\sim\text{Ber}(p) \Rightarrow P(X=x) = p^x * (1-p)^{1-x}, x \in \{0,1\} \tag{14}$$

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.

$$X\sim\text{Bin}(n,p) \Rightarrow P(X=k) = \binom{n}{k} p^k * (1-p)^{n-k}, k \in \{0,1,...,n\} \tag{15}$$

5.3 Geometric distribution:

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

$$X\sim\text{Geom}(p) = p * (1-p)^{k-1}, k \in \{1,2,...\} \tag{16}$$

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.

$$P(X=k) = \frac{\binom{N_A}{k}\binom{N-N_A}{n-k}}{\binom{N}{n}} \tag{17}$$

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.
$P(X < a) = \lim_{t\to a^-} F(t) = F(a-)$ $P(X = a) = F(a) - \lim_{t\to a^-} F(t) = F(a) - F(a-)$	
$E(X) = \sum_k k\,p_X(k)$	$E(X) = \int_{-\infty}^{\infty} x f(x)\,dx$
$E(aX + b) = aE[X] + b$	
$E[g(X)] = \sum_k g(k)\,p_X(k)$	$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)\,dx$
$\mathrm{Var}(X) = E\left[(X - E[X])^2\right] = E[X^2] - (E[X])^2$	
$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$	

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 sec.discreteSampleSpace)

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

$$p_X(k) = P(X = k) \text{ , where } k \text{ are possible values of X} \tag{18}$$

It is a function of k, and

$$p_X : S \to [0, 1], \tag{19}$$

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,

$$p_X(k) = \frac{1}{6}, \text{for every } k \in \{1, 2, 3, 4, 5, 6\} \tag{20}$$

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

$$\begin{gathered} \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1, f(x) > 0 \text{ everywhere} \\ \text{and } P(X \leq b) = \int_{-\infty}^{b} f(x)\,\mathrm{d}x \Leftrightarrow P(a \leq X \leq b) = \int_{a}^{b} f(x)\,\mathrm{d}x \end{gathered} \tag{21}$$

6.2.2 Cumulative Distribution Function (cdf)

cdf of a r.v. is defined as

$$F(x) = P(X \leq x) \tag{22}$$

and it follows that

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) \tag{23}$$

- 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,

  $$F'(x) = f(x), \tag{24}$$

  so the pdf is the derivative of the cdf.

- Discrete r.v.

  pmf and cdf is connected by

  $$F(x) = P(X \leq s) = \sum_{k \leq x} p_{X(k)} \tag{25}$$

  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

$$E(X) = \sum_k kP(X = k) \tag{26}$$

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 possible values.

Expectation is a linear operator, i.e.

$$E(aX + b) = aE(X) + b \tag{27}$$

for any constants a and b.

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

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

where p is the probability of success.

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

$$E(X) = np \tag{29}$$

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

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

$$E(X) = \frac{1}{p} \tag{30}$$

where p is the probability of success.

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

$$E(X) = \int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x \tag{31}$$

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

$$E(X) = \frac{a+b}{2} \tag{32}$$

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

$$E(g(X)) = \sum_k g(k)p_{X(k)} \tag{33}$$

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

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,\mathrm{d}x \tag{34}$$

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

$$E(X^n) = \sum_k k^n p_{X(k)} \tag{35}$$

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

$$E(X^n) = \int_{-\infty}^{\infty} x^n f(x)\,\mathrm{d}x \tag{36}$$

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

$$M_X(t) = E(e^{tX}) = \sum_k e^{tk} p_{X(k)} \tag{37}$$

   - continuous random variable X is defined as

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\,\mathrm{d}x \tag{38}$$

   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.

$$E(X^n) = \frac{\mathrm{d}^n}{\mathrm{d}t} M_X(t=0) \tag{39}$$

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.

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

alternatively,

$$\mathrm{Var}(X) = E(X^2) - (E(X))^2 \tag{41}$$

Variance is not a linear operator, i.e.

$$\mathrm{Var}(aX + b) = a^2\ \mathrm{Var}(X) \tag{42}$$

for any constants a and b.

1. variance of bournoli r.v. is

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

2. variance of binomial r.v. is

$$np(1-p) \tag{44}$$

3. variance of geometric r.v. is

$$\frac{1-p}{p^2} \tag{45}$$

4. variance of uniform r.v. is

$$\frac{(b-a)^2}{12} \tag{46}$$

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

$$f(x) = \begin{cases} \dfrac{1}{b-a} \text{ for } a \le x \le b \\ 0 \text{ otherwise} \end{cases} \tag{47}$$

7.2 Normal (Gaussian) Distribution

7.2.1 standard normal distribution

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

$$f(z) = \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \tag{48}$$

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

$$\Phi(z) = P(Z \le z) = \int_{-\infty}^{z} \varphi(z)\,\mathrm{d}z \tag{49}$$

  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:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \tag{50}$$

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

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - \mu}{\sigma} \le b\right) = \int_a^b \varphi(x)\,\mathrm{d}x = \Phi(b) - \Phi(a) \tag{51}$$

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

$$P(a \le S_n \le b) = P(a - 0.5 \le S_n \le b + 0.5) \tag{52}$$

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

$$S_n \sim \text{Bin}(n,p)\ ;\ E(S_n)=np, E\left(\frac{S_n}{n}\right)=p$$
$$P\left(|\frac{S_n}{n}-p|<\varepsilon\right)\to 1 \text{ as } n\to\infty \tag{53}$$

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

$$P(|\hat{p}-p|<\varepsilon)\geq 2\Phi\big(2\varepsilon\sqrt{n}\big)-1 \tag{54}$$

   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.

A discrete r.v. L has the Poisson distribution with parameter λ>0 if its pmf is

$$p_L(k)=e^{-\lambda}\frac{\lambda^k}{k!} \tag{55}$$

for k = 0, 1, 2, ...

- write $L\sim$ Poisson(λ)
- The mean and variance of a Poisson r.v. are both equal to λ.

8.2.2 Law of rare events

For $S_n\sim$ Bin$\left(n,\frac{\lambda}{n}\right)$, where $\frac{\lambda}{n}<1$, S_n follows the law of rare events,

$$\lim_{n\to\infty} P(S_n=k)=e^{-\lambda}\frac{\lambda^k}{k!} \tag{56}$$

The distribution Bin(n, $\frac{\lambda}{n}$) approaches Poisson(λ) distribution, where $E(S_n)=\lambda$

For a fixed n, to quantify the error in approximation, we have:

Let X Bin(n,p), and Y Poisson(λ), where $\lambda=np$

then for any subset

$$A\subseteq\{0,1,2,...,n\}, k\in A$$
$$|P(X=k)-P(Y=k)|\leq np^2 \tag{57}$$

if $np^2<1$, then the approximation is good, and that

$$P(X=k)\approx P(Y=k)=e^{-\lambda}\frac{\lambda^k}{k!} \tag{58}$$

8.3 Exponential Distribution

No mentioning where it comes from, but will be told when “can be modeled by exponential distribution” A continuous r.v. X has the exponential distribution with parameter λ>0 if its pdf is

$$f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ for } x \geq 0 \\ 0 \text{ otherwise} \end{cases} \tag{59}$$

Write $X\sim$ Exp(λ) The cdf is found by integrating the pdf,

$$F(x) = \begin{cases} 1 - e^{-\lambda x} \text{ for } x \geq 0 \\ 0 \text{ otherwise} \end{cases} \tag{60}$$

Notice the tail probability,

$$P(X > t) = e^{-\lambda t} \tag{61}$$

Expectations and variance are

$$E(X) = \frac{1}{\lambda}, \text{Var}(X) = \frac{1}{\lambda^2} \tag{62}$$

- EXp distribution is memoryless, i.e.

$$\begin{aligned} P(X > t + s \mid X > t) &= \frac{P(X > t + s, X > t)}{P(X > t)} \\ &= \frac{P(X > t + s)}{P(X > t)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} \\ &= e^{-\lambda s} \\ &= P(X > s) \end{aligned} \tag{63}$$

  for all s,t > 0

9 Joint Distribution

9.1 discrete joint distribution

- definition:

$$p(k_1, k_2, k_3) = P(X_1 = k_1, X_2 = k_2, X_3 = k_3) \tag{64}$$

  for r.v. $X_1 = k_1, X_2 = k_2, X_3 = k_3$

- expectation:

$$E(g(X_1, X_2, X_3)) = \sum_{k_1} \sum_{k_2} \sum_{k_3} g(k_1, k_2, k_3) p(k_1, k_2, k_3) \tag{65}$$

- marginal distribution:

$$p_1(k) = \sum_{k_2} \sum_{k_3} p(k, k_2, k_3) \tag{66}$$

- Multinomial distribution when looking for the probability of some independent events together, we can use the multinomial distribution.

$$P(X_1 = k_1, X_2 = k_2, X_3 = k_3) = \frac{n!}{k_1! k_2! k_3!} p_1^{k_1} p_2^{k_2} p_3^{k_3} \tag{67}$$

  abbreviate this as $(X_1, X_2, ..., X_r)\sim$ Multi$(n, r, p_1, p_2, ..., p_r)$

9.2 Continuous joint distribution

- definition:

$$P((X_1, X_2, X_3) \in A) = \int_A f(x_1, x_2, x_3) \, \mathrm{d}x_1 \, \mathrm{d}x_2 \, \mathrm{d}x_3 \tag{68}$$

for r.v. X_1, X_2, X_3 and set $A \in \mathfrak{R}$

- expectation:

$$E(g(X_1, X_2, X_3)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1, x_2, x_3) f(x_1, x_2, x_3)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\mathrm{d}x_3 \tag{69}$$

- marginal distribution:

$$f_1(x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y, z)\,\mathrm{d}y\,\mathrm{d}z \tag{70}$$

9.3 Independent joint random variables

- Necessary and sufficient Condition:
  - discrete

$$p(x_1, x_2) = p_1(x_1)p_2(x_2) \tag{71}$$

  - Continuous

$$f(x_1, x_2) = f_1(x_1)f_2(x_2) \tag{72}$$

- If two r.v. depend on different parameters, they are independent. i.e.

$$\begin{gathered} Y = f(X_1, X_2, X_3); \quad Z = g(X_4, X_5, X_6) \\ \Rightarrow Y \text{ and } Z \text{ are independent} \end{gathered} \tag{73}$$

10 Sums of independent r.v.& Symmetry

10.1 Convolution of two distributions

given two independent r.v. X and Y, the distribution of $Z = X + Y$ is the convolution of the distributions of X and Y.

1. when X and Y are both discrete, the pmf of $X + Y$ is given by

$$p_{X+Y}(n) = p_X * p_Y(n) = \sum_k p_{X(k)}p_{Y(n-k)} = \sum_k p_{X(n-k)}p_{Y(k)} \tag{74}$$

2. when X and Y are both continuous, the pdf of $X + Y$ is given by

$$f_{X+Y}(z) = f_X * f_Y(z) = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy = \int_{-\infty}^{\infty} f_X(z-y)f_Y(y)dy \tag{75}$$

- *example: convolution of geometric random variables*

  let X and Y be independent geometric random variables with the same suvvess parameter $p < 1$, find the distrbution of $Z = X + Y$.

  We know $p_X(k) = p_Y(k) = p(1-p)^{k-1}$ $k \geq 1$ r.v. $Z = X + Y$ takes on values $n = 2, 3, \ldots$. Via the convolution magic promised above, we have

$$\begin{aligned} P(X + Y = n) &= \sum_{k=-\infty}^{\infty} P(X = k)P(Y = n - k) \\ &= \sum_{k=1}^{n-1} p(X = k)P(Y = n - k) \\ &= \sum_{k=1}^{n-1} p(1-p)^{k-1}p(1-p)^{n-k-1} \\ &= \sum_{k=1}^{n-1} p^2(1-p)^{n-2} \\ &= (n-1)p^2(1-p)^{n-2} \end{aligned} \tag{76}$$

10.2 Negative binomial distribution

Coming off from the geometric distribution, we have the negative binomial distribution, which is the distribution of the number of trials needed to get r successes in a sequence of independent Bernoulli trials with success probability p. Its distribution, i.e. pmf, is given by

$$P(X=n)=\binom{-1}{k-1}p^k(1-p)^{n-k} \quad (n\geq k) \tag{77}$$

abbriviate this by $X\sim$ Negbin (k,p) , where the geometric is a special case with $k=1$.

10.3 Collection of normal distributed r.v.s

For $X_i\sim\mathcal{N}(\mu_i,\sigma_i^2)$, $X=\sum_i a_iX_i$, we know

$$\begin{gathered} X\sim\mathcal{N}(\mu,\sigma^2) \\ \text{where } \mu=\sum_i a_i\mu_i,\ \sigma^2=\sum_i a_i^2\sigma_i^2 \end{gathered} \tag{78}$$

in other words, the sum of normal distributed r.v.s is also normal distributed.

10.4 Exchangeable r.v.s

a sequence of r.v.s $X_1,X_2,X_3,...,X_n$ is **exchangable** if the following condition holds: for any permutation (k_1,k_2,k_3) of $(1,2,...,n)$, we have

$$(X_1,X_2,...,X_n)\overset{d}{=}\left(X_{k_1},X_{k_2},...,X_{k_n}\right) \tag{79}$$

- **How to check exchangability**

  "it just works" method: check if the r.v. are identically distributed, i.e. if marginal pdf or pmf is the same.

  Suppose $X_1,X_2,...,X_n$are discrete random variables with joint probability mass function p. Then these random variables are exchangeable if and only if p is a symmetric function.

  Suppose $X_1,X_2,...,X_n$ are jointly continuous random variables with joint density function f.Then these random variables are exchangeable if and only if f is a symmetric function.

  If the expectation is conserved under permutations of our set of r.v.s.

Importantly, if the r.v.s are independent and identically distributed, they are also exchangeable.

remarks:

1. r.v. denoting outcomes of sampling without replacement $X_1,X_2,...X_n$ are exchangeable.
2. For any function g dependent on , the r.v.s $g(X_1),g(X_2),...,g(X_n)$ are exchangeable.

10.5 Expectation and Varience of Multivariable r.v.

10.5.1 Expectation: linear

$$E[g_1(X_1)+g_2(X_2)+...+g_n(X_n)]=E[g_1(X_1)]+E[g_2(X_2)]+...+E[g_n(X_n)] \tag{80}$$

$$E[X_1+X_2+...+X_n]=E[X_1]+E[X_2]+...+E[X_n] \tag{81}$$

Expectation of a sum is always the sum of expectations.

10.5.2 Varience: sum of independent r.v., linear

$$\mathrm{Var}(X_1+X_2+...+X_n)=\mathrm{Var}(X_1)+\mathrm{Var}(X_2)+...+\mathrm{Var}(X_n) \tag{82}$$

10.5.3 The indicator method

- *example* We draw five cards from a deck of 52 without replaceement. Let X denote the number of Aces among the chosen cards. Find the expected value of X.

  Two ways to solve this:

1. Since order does not matter in our draw of 5, by argument of exchangability, we can construct the following inditator:

$$I_i = \begin{cases} 1 & \text{if the ith card is an ace} \\ 0 & \text{otherwise} \end{cases} \tag{83}$$

Since X is the number of Aces among our 5 cards, we have

$$X = I_1 + I_2 + I_3 + I_4 + I_5 \tag{84}$$

Recall the linearity of expectation, we can rephrase the expected value as

$$E[X] = E[I_1] + E[I_2] + E[I_3] + E[I_4] + E[I_5] \tag{85}$$

Since r.v. I_i are exchangeable, we have

$$E[I_1] = E[I_2] = E[I_3] = E[I_4] = E[I_5] \tag{86}$$

Equation 85 becomes

$$5 * E[I_1] = 5 * P(I_1 = 1) = 5 * \frac{4}{52} = \frac{5}{13} \tag{87}$$

2. We can also label the Aces in the total deck as 1,2,3,4, and have our indicators j_1, j_2, j_3, j_4 indicating if the ith Ace is in our draw or not. The number of Aces in our draw is then $X = j_1 + j_2 + j_3 + j_4$. By similar arguements of exchangability, we have $E[X] = 4E[J_1] = 4P(\text{one of the ace is among the 5 cards})$. Notice that

$$\begin{aligned} P(\text{one of the ace is among the 5 cards}) &= \frac{\binom{1}{1}, \binom{51}{4}}{\binom{52}{5}} = \frac{5}{52} \\ \Rightarrow E[X] &= \frac{5}{13} \end{aligned} \tag{88}$$

10.5.4 Expectation of multiple products

let X_1, X_2, X_3 be independend r.v., when for all function g_1, g_2, g_3

$$E\left[\prod_{i=1}^{3} g_{i(X_i)}\right] = \prod_{i=1}^{3} E\left[g_{i(X_i)}\right] \tag{89}$$

10.6 Moment generating function with sums of r.v.

For independent r.v. X, Y, and mgf $M_X(t)$, $M_Y(t)$,

$$M_{X+Y}(t) = M_X(t) M_Y(t) \tag{90}$$

10.7 Covariance and correlation

10.7.1 Covariance

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y] \tag{91}$$

- X&Y are
  - positively correlated if $\text{Cov}(X, Y) > 0$
  - negatively correlated if $\text{Cov}(X, Y) < 0$
  - uncorrelated if $\text{Cov}(X, Y) = 0$

10.7.2 Properties of Covariance

- $\text{COV}(X, Y) = \text{COV}\ (Y, X)$
- $\text{COV}\ (aX + b, Y) = a\ \text{COV}(X, Y)$
- for any r.v. X_i, Y_j and real numbers a_i, b_j:

$$\text{COV}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{m} a_i b_j\ \text{COV}\left(X_i, Y_j\right) \tag{92}$$

- Practically,

$$\begin{gathered}\operatorname{Cov}(Y_1+Y_2,Z)=\operatorname{Cov}(Y_1,Z)+\operatorname{Cov}(Y_2,Z)\\ \operatorname{Cov}(X,X)=\operatorname{Var}(X)\end{gathered} \tag{93}$$

10.7.3 Variance of sum of r.v.s

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right)=\sum_{i=1}^{n}\operatorname{Var}\left(X_i\right)+2\sum_{i\le i<j\le n}\operatorname{Cov}(X_i,X_j)) \tag{94}$$

For two r.v.s, this comes down to

$$\operatorname{Var}\left(X+Y\right)=\operatorname{Var}\left(x\right)+\operatorname{Var}\left(Y\right)+2\operatorname{Cov}(X,Y) \tag{95}$$

For three r.v.s, this is uglier...

$$\begin{gathered}\operatorname{Var}\left(X+Y+Z\right)\\ =\operatorname{Var}\left(X\right)+\operatorname{Var}\left(Y\right)+\operatorname{Var}\left(Z\right)+2\operatorname{Cov}(X,Y)+2\operatorname{Cov}(X,Z)+2\operatorname{Cov}(Y,Z)\end{gathered} \tag{96}$$

You dont want to compute this for four or more...

10.7.4 Correlation

$$\operatorname{Corr}(X,Y)=\frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}} \tag{97}$$

11 Tail bounds and limit theorems

11.1 Markov's inequality

For any non-negative r.v. X and any $a>0$, we have

$$P(X\ge a)\le\frac{E[X]}{a} \tag{98}$$

11.2 Chebyshev's inequality

For any r.v. X with finite mean and variance, and any $k>0$, we have

$$P(|X-E[X]|\ge k)\le\frac{\operatorname{Var}(X)}{k^2} \tag{99}$$

normally used to find $P(X\ge c+\mu)\le\frac{\sigma^2}{c^2}$ and $\quad P(X\le\mu-c)\le\frac{\sigma^2}{c^2}$

11.3 generalized Law of large numbers

For a sequence of iid r.v.s $X_1,X_2,...,X_n$ with finite mean $E[X_i]=\mu$ and finite variance $\operatorname{Var}\left[X_i\right]=\sigma^2$, letting $S_n=X_1+X_2+...+X_n$, for any $\varepsilon>0$, we have

$$\lim_{n\to\infty}P\left(|\frac{S_n}{n}-\mu|<\varepsilon\right)=1 \tag{100}$$

11.4 Generalized Central Limit Theorem

For a sequence of iid r.v.s $X_1,X_2,...,X_n$, where n is the sample size, with finite mean $E[X_i]=\mu$ and finite variance $\operatorname{Var}\left[X_i\right]=\sigma^2$, letting $S_n=X_1+X_2+...+X_n$, we have

$$\lim_{n\to\infty}P\left(a\le\frac{S_n-n\mu}{\sigma\sqrt{n}}\le b\right)=\Phi(b)-\Phi(a) \tag{101}$$

More practically, we use

$$P(S\ge k)=P\left(\frac{S_n-n\mu}{\sqrt{n\sigma^2}}\ge\frac{k-n\mu}{\sqrt{n\sigma^2}}\right)=1-\Phi\left(\frac{k-n\mu}{\sqrt{n\sigma^2}}\right) \tag{102}$$

12 Conditional distribution

A combination of conditional probability and marginal distribution.

12.1 Discrete conditional distribution

recall the conditional probability

$$P(A|B) = \frac{P(A \cap B)}{B}, \text{for } P(B) > 0 \tag{103}$$

When A is now a r.v., we have the conditional distribution

$$p_{X|B}(k) = P(X = k|B) = \frac{P(\{X = k\} \cap B)}{P(B)} \tag{104}$$

12.1.1 Conditional expectation of X, given event B

$$E[X|B] = \sum_k kP(X = k|B) \tag{105}$$

12.1.2 Unconditiond pmf of X

$$p_X(k) = \sum_{i=1}^{n} p_{X|B_i}(k)\ P(B_i) \tag{106}$$

- From Equation 105 and Equation 106 we get

$$E[X] = \sum_{i=1}^{n} E[X|B_i]P(B_i) \tag{107}$$

12.1.3 Conditioning on r.v.

When both X and Y are r.v.s, we can have the following two-variable function

$$p_{X|Y}(k|j) = P(X = k|Y = j) = \frac{P(\{X = k\}, \{Y = j\})}{P(Y = j)} = \frac{p_{X,Y}(k, j)}{p_Y(j)} \tag{108}$$

12.1.4 Conditional expectation of X, given Y=Y

$$E[X|Y = j] = \sum_k kP(X = k|Y = j) = \sum_k kp_{X|Y}(k|j) \tag{109}$$

12.1.5 Unconditioned pmf with 2 r.v.S

$$p_X(k) = \sum_i p_{X|Y}(k|j)\ p_Y(j) \tag{110}$$

- From this, we can derive the unconditioned expectation of X and Y

$$E[X] = \sum_k E[X|Y = j]\ p_Y(j) \tag{111}$$

12.1.6 Joint pmf with 2 r.v.s

$$p_{X,Y}(k, j) = p_{X|Y}(k|j)p_Y(j) = p_{Y|X}(j|k)p_X(k) \tag{112}$$

12.2 Continuous conditional distribution

For continuous r.v.s, with both X, Y random variables, we have the conditional pdf of X given Y = y as

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \tag{113}$$

12.2.1 Conditional probability and expectation

$$P(X \in A|Y = y) = \int_A f_{X|Y}(t|y)dt \tag{114}$$

The conditoinal expectation of $g(X)$

$$E[g(X)|Y=y] = \int_{-\infty}^{\infty} g(t)\ f_{X|Y}(t|y)dt \tag{115}$$

12.2.2 The unconditioned pdf and expectation of X

Given the conditional pdf $f_{X|Y}(x|y)$, we can derive the unconditioned pdf of X as

$$f_X(x) = \int_{-\infty}^{\infty} f_{X|Y}(x|y) f_Y(y) dy \tag{116}$$

$$E[g(X)] = \int_{-\infty}^{\infty} E[g(X)|Y=y]\ f_Y(y)\,\mathrm{d}y \tag{117}$$

12.3 Conditional expectation

12.3.1 conditional expectation as a r.v.

Let X and Y jointly continuous r.v., The conditional expectatino of X given Y is a new random variable dependent on Y $v(Y)$

$$v(Y) = E[X|Y=y] \tag{118}$$

12.3.2 Conditioning and independence

recall that

- Discrete r.v. two discrete r.v.s are only independent iff

$$p_{X,Y}(x,y) = p_X(x)p_Y(y) \tag{119}$$

- Continuous r.v. two continuous r.v.s are only independent iff

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) \tag{120}$$

now, If given pmf or pdf of X given Y, we now have the joint pmf

$$p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y) \tag{121}$$

and the joint pdf

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) \tag{122}$$

12.3.3 Independency of X and Y

discrete r.v. X and Y are independent iff

$$p_{X|Y}(x|y) = p_X(x) \tag{123}$$

continuous r.v. X and Y are independent iff

$$f_{X|Y}(x|y) = f_X(x) \tag{124}$$

12.4 Conditioning on the random variable

12.4.1 Conditioning X on y

for independent r.v. X and Y, we can condition on Y and have the conditional expectation of X given Y = y as

$$E[g(X)|Y=y] = E[g(X)] \quad \text{and } E[g(X)|Y=y] = E[g(X)] \tag{125}$$

12.4.2 COnditioning X on X

For a r.v. X, we can condition on X itself, and have the conditional expectation of X given X = x as

$$E[g(X)|X=x] = g(X) \tag{126}$$