1 Discrete Mathematics

2 Probability

Bold text *anywhere* signifies truth for both discrete and continuous RVs. If bold but not explicit, change $p(x)$ to $f(x)$ or integral to a sum, v.v.

Counting

- Experiment 1 has n outcomes, another has m. This gives $n \cdot m$ outcomes. N repitions of Experiment 1 gives n^N outcomes.
- Permutations: how many ways to order a set. $n!$, or if we count repeated items as indistinguishable, $\frac{n!}{r_1! \cdot \dots \cdot r_i!}$ where r_i is the number of times the number i was repeated.
- $P_{n,k} = \frac{n!}{(n-k)!}$ ($\binom{n}{k}$ but order matters), $\binom{n}{k} = \frac{n!}{k!(n-k)!}$
- The number of ways to choose something OR something else is addition. Choosing two things together (AND) is multiplication. If we are to find 'at least one (up to 3)', it would be the number of ways to choose 1 OR 2 OR 3.

Basic Probability

- $\bigcup_i E_i$ means at least one E_i and $\bigcap_i E_i$ means all of the E_i's.
- Mutex events (disjoint) can't happen at the same time. If $E \subseteq F$, E can't happen without F happening. $E^c = \Omega - E$.
- Axioms: for countably many *mutex* events, $P(\bigcup E_i) = \sum_i P(E_i)$. If not mutex, this equality is replaced with $\leq$.
- Elementary events are those such that all events have probability $\frac{1}{|\Omega|}$.
- $P(F - E) = P(F \cap E^c) = P(F) - P(F \cap E)$. This is the set difference law.
- De Morgan's: $(E \cap F)^c = E^c \cup F^c$ (flip intersection/union and take complement).
- $P(E|F) = \frac{P(E \cap F)}{P(F)}$. All axioms work fine, just add the 'given' part to each.
- Multiplication rule: $P(E_1 \cap E_2 \cap \dots \cap E_n) = P(E_1) \cdot P(E_2|E_1) \cdot P(E_3|E_1 \cap E_2) \cdots P(E_n|E_1 \cap \dots \cap E_{n-1})$.
- *Events* are independent $\iff P(F|E) = P(F) \iff P(E|F) = P(E) \iff P(E \cap F) = P(E) \cdot P(F)$. **Random variables** are thus independent $\iff p_{Y|X}(y|x) = p_Y(y) \iff p_{X|Y}(x|y) = p_X(x) \iff p(x,y) = p_X(x) \cdot p_Y(y)$ for all (x, y).
- Three or more events are mutually independent if the above multiplication rule applies to all pairs, and all of them together.
- Murphy's Law: As $n \to \infty$ with fixed probability p, $P(\text{all } n \text{ experiments succeed}) = p^n \to 0$, $P(\text{at least one succeeds}) = 1 - P(\text{each one fails}) = 1 - (1-p)^n \to 1$, $P(\text{exactly } k \text{ succeed}) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$.

Discrete Random Variables

- PMF: $p_X(x) = P(X = x)$, CDF: $F_X(x) = P(X \leq x) = \sum_{x_i \leq x} p(x_i)$. **CDF is non-decreasing.**
- $\mathbb{E}(X) = \sum_i x_i \cdot p(x_i)$. Needn't be a possible value. **Moments**: nth moment: $\mathbb{E}(X^n)$. Absolute moment: $\mathbb{E}(|X|^n)$.
- $\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. $\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}$.
- **Chebyshev Inequality**: for any constant $k \geq 1$, the probability that X is more than k standard deviations away from the mean is no more than $\frac{1}{k^2}$. $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$.
- Binomial: there are n independent trials, each succeeding with probability p. X counts the number of successes: $X \sim \mathrm{Bin}(n, p)$. If $n = 1$, this is Bernoulli where $P(X = 1) = p$ is the probability of the 'only' success. $p(x) = P(\text{number of successes is x}) = \binom{n}{x} p^x (1-p)^{n-x}$, $\mathbb{E}(X) = np$, $\mathrm{Var}(X) = np(1-p)$.
- Poisson: $X \sim \mathrm{Pois}(\lambda \in \mathbb{R}^+)$ if $p(x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}$, $\mathbb{E}(X) = \mathrm{Var}(X) = \lambda$. Poisson measures the probability of a number of events happening in a space, based on an average (λ). For example, number of calls per hour, or number of typos on a page. $\mathrm{Bin}(n, p) \approx \mathrm{Pois}(np)$ for large n and small p.
- Geometric: $X \sim \mathrm{Geom}(p)$ counts the number of independent trials repeated until we get a success (with probability p), with no memory. $p(x) = p(1-p)^{x-1}$, $\mathbb{E}(X) = \frac{1}{p}$, $\mathrm{Var}(X) = \frac{1-p}{p^2}$.

Continuous Random Variables

- The PDF is defined as being the function $f(x)$ such that its integral, from a to b, gives the probability $P(a \leq x \leq b)$, and from $-\infty$ to ∞ gives 1. The derivative of the CDF is the PDF.
- CDF: $F(x) = P(X \leq x) = \int_{-\infty}^{x} f(y)\,dy$. There will usually be a lower bound for Y, such that any values of Y less than this lower bound will have a 0 probability, meaning we don't have to compute an improper integral.
- $P(X > a) = 1 - F(a)$, $P(a \leq X \leq b) = F(b) - F(a)$.
- Percentiles: the 75th percentile is the value $\eta_{0.75}$ s.t. $P(X < \eta) = 0.75$. Thus we can calculate it with the inverse of the CDF: $\eta_{0.5} = F^{-1}(0.5)$, which also happens to be the *median*, sometimes written just η.
- $\mathbb{E}(X) = \int_{-\infty}^{\infty} x \cdot f(x)\,dx = \mu_X$, $\mathbb{E}(h(X)) = \int_{-\infty}^{\infty} h(x) \cdot f(x)\,dx$.
- $\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mu^2$, $\mathrm{SD}(X) = \sigma_X = \sqrt{\mathrm{Var}(X)}$.
- Uniform: $X \sim U(a, b)$ if $f(x) = \frac{1}{b-a}$ if $a \leq x \leq b$, $f(x) = 0$ otherwise. Equal probabilities for anything between a and b, otherwise 0.
- No real need for CDF. Use rectangle intuition: the height is $\frac{1}{b-a}$ and the width would be the amount 'along' the rectangle you would be. $P(a \leq X \leq c) = (c - a) \cdot \frac{1}{b-a} = P(X \leq c)$ if the probability is equal between a and b. If the density is high, the CDF's graph is steep.
- Normal (Gaussian): $X \sim N(\mu, \sigma)$ if $f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-(x-\mu)^2/(2\sigma^2)}$. μ is the centre while σ measures how widely spread it is. Height, sheep producing wool, etc., where many random factors are involved.
- About $\frac{2}{3}$ of probability mass is within one SD of the mean, 95% within two.
- Standard normal if $\mu = 0$ and $\sigma = 1$. $f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} \cdot e^{-z^2/2}$.
- Standardising: If $X \sim N(\mu, \sigma)$, $Z = \frac{X - \mu}{\sigma} \mid Z \sim N(0, 1)$. $P(X \leq a) = P(Z \leq \frac{a-\mu}{\sigma}) = \Phi(\frac{a-\mu}{\sigma})$, $\eta_p = \mu + \Phi^{-1}(p) \cdot \sigma$. Go opposite way in table for inverse.
- Approximating binomial: find $\mu = np$ and $\sigma = \sqrt{npq}$ where $q = (1 - p)$. Use these two values as parameters for normal. Thus $P(X \leq x) = \Phi(\frac{x + 0.5 - np}{\sqrt{npq}})$. Adequate if $np, nq \geq 10$. We add 0.5 for continuity correction.
- Exponential: how long until something happens, with no memory: $X \sim \mathrm{Exp}(\lambda)$ if $f(x; \lambda) = \lambda e^{-\lambda x}$ if $x > 0$, 0 otherwise. $F(x; \lambda) = 1 - e^{-\lambda x}$ if $x > 0$, 0 otherwise.
- $\mathbb{E}(X) = \frac{1}{\lambda} = \mathrm{SD}(X)$. $\mathrm{Var}(X) = \frac{1}{\lambda^2}$.
- Relation to Poisson: Poisson counts the number of arrivals each minute, while exponential counts the time between arrivals at a drive-through.

Transformations of Random Variables

- **Linearity of** $\mathbb{E}$: expectation of the sum of RVs is the sum of their expectations. **Rescaling**: $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$, $\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$, $\mathrm{SD}(aX + b) = |a| \mathrm{SD}(X)$.
- $\mathbb{E}(h(X)) = \int_{-\infty}^{\infty} h(x) \cdot f(x)\,dx$, $E(h(X)) = \sum_i h(x_i) \cdot p(x_i)$.
- $\mathrm{Cov}(aX + b, cY + d) = ac\,\mathrm{Cov}(X, Y)$, $\mathrm{Cov}(aX + bY + c, Z) = a\,\mathrm{Cov}(X, Z) + b\,\mathrm{Cov}(Y, Z)$.
- **Transformations of RVs are themselves RVs**: if a transformation of a random variable $Y = g(X)$ is monotonically increasing, like radius to area, then CDF is $F_Y(y) = F_X(g^{-1}(y))$. If monotonically decreasing, like speed to time, then $F_Y(y) = 1 - F_X(g^{-1}(y))$ since the inequality is flipped.
- PDF is $f_Y(y) = f_X(g^{-1}(y)) \cdot |\frac{d}{dy} g^{-1}(y)|$, where the derivative accounts for the change in width of the curve.

Joint Probability Distributions

Discrete

- JPMF of X and Y is $p(x, y)$ defined for every pair s.t. $p(x, y) = P(X = x \wedge Y = y)$. For an event $A \subset \mathbb{R} \times \mathbb{R}$, the probability of any of its outcomes occuring is $P((X, Y) \in A) = \sum_{(x,y) \in A} p(x, y)$.
- Marginal probabilities allow you to calculate individual variables' PMFs from their JPMF: $p_X(x) = \sum_y p(x, y)$, $p_Y(y) = \sum_x p(x, y)$. If given a table, add up the values in each column or row.
- $\mathbb{E}(g(X, Y)) = \sum_x \sum_y g(x, y) \cdot p(x, y)$.

Continuous

- Imagine a rectangle $A = \{(x, y) \mid a \leq x \leq b, c \leq y \leq d\}$. The probability $P((X, Y) \in A) = P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b (\int_c^d f(x, y)\,dy)\,dx = \int_c^d (\int_a^b f(x, y)\,dx)\,dy$. This is the probability that the random variables lie in the same rectangle together.
- Marginal probabilities: $f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,dy$, $f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx$. If given a support, like $a \leq X \leq b$ and $c \leq Y \leq d$, use the bounds of the *variable you're integrating with respect to* (which avoids improper integrals).
- $\mathbb{E}(g(X, Y)) = \int_{-\infty}^{\infty} (\int_{-\infty}^{\infty} g(x, y) \cdot f(x, y)\,dx)\,dy$.

Variance, Sums and Combinations

- **Linearity of expectation applies. Multiplying expectations of different random variables applies only if they are independent:** $\mathbb{E}(h(X, Y)) = \mathbb{E}(g_1(X) \cdot g_2(Y)) = \mathbb{E}(g_1(X)) \cdot \mathbb{E}(g_2(Y))$, and $h(X, Y) = g_1(X) \cdot g_2(X)$.
- **Covariance**: how much do X and Y vary together? If $\mathrm{Cov}(X, Y) > 0$, they vary together. If 0, they do not vary together. Vary in opposition otherwise. X and Y independent $\implies \mathrm{Cov}(X, Y) = 0$, but not the other way round.
- **Definition of** Cov: $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X) = \mathbb{E}(XY) - \mu_X \cdot \mu_Y$, $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$.
- **Correlation coefficient**: similar to standard deviation. $\rho_{X,Y} = \mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}$. If $\rho > 0$, positively correlated. Not linearly correlated or negatively correlated otherwise.
- **Properties of** Corr: $\mathrm{Corr}(X, Y) = \mathrm{Corr}(Y, X)$, $\mathrm{Corr(X, X)} = 1$ $\mathrm{Corr}(X, Y) = 0 \iff \mathbb{E}(XY) = \mu_X \cdot \mu_Y$, $\mathrm{Corr}(X, Y) = 1 \iff Y = aX + b \mid a > 0$, $\mathrm{Corr}(X, Y) = -1 \iff Y = aX + b \mid a < 0$. Correlation has the range $[-1, 1]$.
- **Linear combinations** of expectations are trivial by the linearity of expectation. For variance: $\mathrm{Var}(a_1X_1 + \dots + a_nX_n + b) = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathrm{Cov}(X_i, X_j)$, $\mathrm{Var}(aX + bY) = a^2 \mathrm{Var}(X) + b^2 \mathrm{Var}(Y) + 2ab \cdot \mathrm{Cov}(X, Y)$.
- **With independence**: $\mathrm{Var}(a_1X_1 + \dots + a_nX_n + b) = a_1^2 \mathrm{Var}(X_1) + \dots + a_n^2 \mathrm{Var}(X_n)$, $\mathrm{SD}(\cdots) = \sqrt{a_1^2\sigma_1^2 + \dots + a_n^2\sigma_n^2}$, $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) = \mathrm{Var}(X - Y)$.
- **Sums of RVs**: Let $W = X + Y$. $f_W(w) = \int_{-\infty}^{\infty} f(x, w - x)\,dx$. If X and Y are independent, can integrate product of marginals. This is convolution: $f_W = f_X * f_Y$. If discrete, sum instead.
- **Sums of standard distributions**: If $X_1 \cdots X_n$ are independent Poisson RVs with means $\mu_1 \cdots \mu_n$, their sum is also Poisson with $\mu = \mu_1 + \dots + \mu_n$. Similar for normal distributions, with $\sigma = \sqrt{\sigma_1^2 + \dots + \sigma_n^2}$.

Conditional and Limit Distributions

Conditional

- **The probability that** $Y = y$ given we know $X = x$ is $P_{Y|X}(y|x) = \frac{p(x,y)}{p_X(x)}$.
- **Conditional mean**: $\mu_{Y|X=x} = \mathbb{E}(Y|X = x) = \sum_y y \cdot p_{Y|X}(y|x)$, $\mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y|x)\,dy$. For a function $h(y)$, replace y with $h(y)$.
- **Conditional variance**: $\sigma_{Y|X=x}^2 = \mathrm{Var}(Y|X = x) = \mathbb{E}((Y - \mu_{Y|X=x})^2|X = x) = \mathbb{E}(Y^2|X = x) - \mu_{Y|X=x}^2$.
- **Law of total expectation/variance:** $\mathbb{E}(Y|X)$ and $\mathrm{Var}(Y|X)$ are themselves random variables with their own mean and variance: $\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X))$, $\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}(Y|X)) + \mathbb{E}(\mathrm{Var}(Y|X))$.

Limit

- If we have a **random sample**, a set of RVs with size n s.t. each RV is independent and identically distributed, the sample total $T = \sum_{i=1}^n X_i$. The sample mean $\bar{X} = \frac{T}{n}$. These are themselves RVs.
- **Properties** of T: $\mathbb{E}(T) = n \cdot \mu$, where μ is the underlying mean of the distribution and σ is the SD. $\mathrm{Var}(T) = n\sigma^2$, $\mathrm{SD}(T) = \sqrt{n} \cdot \sigma$. If the X_i are normally distributed, so is T.
- **Properties** of $\bar{X}$: $\mathbb{E}(\mu)$, $\mathrm{Var}(\bar{(}X) = \frac{\sigma}{n})$, $\mathrm{SD}(\bar{X}) = \frac{\sigma}{\sqrt{n}}$. If the X_i are normal, so is $\bar{X}$.
- **Law of large numbers**: as $n \to \infty$, the following are true: $\bar{X}$ converges to μ; $\mathbb{E}((\bar{X} - \mu)^2) \to 0$; $P((\bar{X} - \mu) \geq \epsilon) \to 0$ for $\epsilon > 0$.
- **Central limit theorem**: summing distributions leads to a narrower and narrower distribution; with large n, the less skewed the distribution gets, thus approaching normal. The peak of the resulting normal distrib. will be around the same mean μ.
- **For IID RVs**, $\lim_{n\to\infty} \bar{X} \sim N(n\mu, \frac{\sigma}{\sqrt{n}})$, and $\lim_{n\to\infty} T \sim N(n\mu, \sqrt{n} \cdot \sigma)$.
- We say that they are **asymptotically normal**.