Probability

https://math.mit.edu/~sheffield/fall2023math600.html

帽子

- n people toss hats into a bin, randomly shuffle, return one hat to each person. Find probability nobody gets own hat.
- Inclusion-exclusion. Let E_i be the event that ith person gets own hat.
- What is $P(E_{i_1}E_{i_2}\dots E_{i_r})$?
- Answer: $\frac{(n-r)!}{n!}$.
- There are $\binom{n}{r}$ terms like that in the inclusion exclusion sum. What is $\binom{n}{r}\frac{(n-r)!}{n!}$?
- Answer: $\frac{1}{r!}$.
- $P(\cup_{i=1}^n E_i) = 1 - \frac{1}{2!} + \frac{1}{3!} - \frac{1}{4!} + \dots \pm \frac{1}{n!}$
- $1 - P(\cup_{i=1}^n E_i) = 1 - 1 + \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \dots \pm \frac{1}{n!} \approx 1/e \approx .36788$

Condition Prob:

- Definition: $P(E|F) = P(EF)/P(F)$.
- Call $P(E|F)$ the "conditional probability of E given F" or "probability of E conditioned on F".
- Nice fact: $P(E_1E_2E_3\dots E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1E_2)\dots P(E_n|E_1\dots E_{n-1})$
- Useful when we think about multi-step experiments.

-

$$\begin{aligned} P(E) &= P(EF) + P(EF^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \end{aligned}$$

独立事件:

- Say E and F are **independent** if $P(EF) = P(E)P(F)$.
- Equivalent statement: $P(E|F) = P(E)$. Also equivalent: $P(F|E) = P(F)$.

Bernoulli

- Toss fair coin n times. (Tosses are independent.) What is the probability of k heads?
- Answer: $\binom{n}{k}/2^n$.
- What if coin has p probability to be heads?
- Answer: $\binom{n}{k}p^k(1-p)^{n-k}$.
- Writing $q = 1 - p$, we can write this as $\binom{n}{k}p^k q^{n-k}$
- Can use binomial theorem to show probabilities sum to one:
- $1 = 1^n = (p+q)^n = \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k}$.
- Number of heads is **binomial random variable with parameters** (n, p).

np, npq

从伯努利到泊松分布: $np = \lambda$

- Let λ be some moderate-sized number. Say $\lambda = 2$ or $\lambda = 3$. Let n be a huge number, say $n = 10^6$.
- Suppose I have a coin that comes on heads with probability λ/n and I toss it n times.
- How many heads do I expect to see?
- Answer: $np = \lambda$.
- Let k be some moderate sized number (say $k = 4$). What is the probability that I see exactly k heads?
- Binomial formula:
$\binom{n}{k}p^k(1-p)^{n-k} = \frac{n(n-1)(n-2)\ldots(n-k+1)}{k!}p^k(1-p)^{n-k}$.
- This is approximately $\frac{\lambda^k}{k!}(1-p)^{n-k} \approx \frac{\lambda^k}{k!}e^{-\lambda}$.
- A **Poisson random variable** X with parameter λ satisfies $P\{X = k\} = \frac{\lambda^k}{k!}e^{-\lambda}$ for integer $k \geq 0$.

均值和方差都是λ

泊松点过程:

- A Poisson point process is a random function $N(t)$ called a Poisson process of rate λ.
- For each $t > s \geq 0$, the value $N(t) - N(s)$ describes the number of events occurring in the time interval (s, t) and is Poisson with rate $(t - s)\lambda$.
- The numbers of events occurring in disjoint intervals are independent random variables.
- Probability to see zero events in first t time units is $e^{-\lambda t}$.
- Let T_k be time elapsed, since the previous event, until the kth event occurs. Then the T_k are independent random variables, each of which is exponential with parameter λ.

Geometric distribution:

- Let X be a geometric with parameter p, i.e., $P\{X = k\} = (1 - p)^{k-1}p = q^{k-1}p$ for $k \geq 1$.
- What is $E[X]$?
- By definition $E[X] = \sum_{k=1}^{\infty} q^{k-1}pk$.
- There's a trick to computing sums like this.
- Note $E[X - 1] = \sum_{k=1}^{\infty} q^{k-1}p(k - 1)$. Setting $j = k - 1$, we have $E[X - 1] = q\sum_{j=0}^{\infty} q^{j-1}pj = qE[X]$.
- Kind of makes sense. $X - 1$ is "number of extra tosses after first." Given first coin heads (probability p), $X - 1$ is 0. Given first coin tails (probability q), conditional law of $X - 1$ is geometric with parameter p. In latter case, conditional expectation of $X - 1$ is same as a priori expectation of X.
- Thus $E[X] - 1 = E[X - 1] = p \cdot 0 + qE[X] = qE[X]$ and solving for $E[X]$ gives $E[X] = 1/(1 - q) = 1/p$.

$1/p, q/p^2$

Negative binomial random variable with parameters (r, p):

Consider an infinite sequence of independent tosses of a coin that comes up heads with probability p. Let X be such that the rth heads is on the Xth toss.

$P(X = k) = \binom{k-1}{r-1}p^r q^{k-r}$

Write $X = X_1 + X_2 + \ldots + X_r$ where X_k is number of tosses (following (k − 1)th head) required to get kth head. Each X_k is geometric with parameter p.

$rp, rq/p^2$

- **Binomial** (S_n — number of heads in n tosses), **geometric** (steps required to obtain one heads), **negative binomial** (steps required to obtain n heads).
- **Standard normal** approximates law of $\frac{S_n - E[S_n]}{\mathrm{SD}(S_n)}$. Here $E[S_n] = np$ and $\mathrm{SD}(S_n) = \sqrt{\mathrm{Var}(S_n)} = \sqrt{npq}$ where $q = 1 - p$.
- **Poisson** is limit of binomial as $n \to \infty$ when $p = \lambda/n$.
- **Poisson point process**: toss one λ/n coin during each length $1/n$ time increment, take $n \to \infty$ limit.
- **Exponential**: time till first event in λ Poisson point process.
- **Gamma distribution**: time till nth event in λ Poisson point process.

- **Sum of two independent binomial random variables** with parameters (n_1, p) and (n_2, p) is itself binomial $(n_1 + n_2, p)$.
- **Sum of n independent geometric random variables** with parameter p is negative binomial with parameter (n, p).
- **Expectation of geometric random variable** with parameter p is $1/p$.
- **Expectation of binomial random variable** with parameters (n, p) is np.
- **Variance of binomial random variable** with parameters (n, p) is $np(1 - p) = npq$.

- **Sum of n independent exponential random variables** each with parameter λ is gamma with parameters (n, λ).
- **Memoryless properties:** given that exponential random variable X is greater than $T > 0$, the conditional law of $X - T$ is the same as the original law of X.
- Write $p = \lambda/n$. **Poisson random variable expectation** is $\lim_{n\to\infty} np = \lim_{n\to\infty} n\frac{\lambda}{n} = \lambda$. **Variance** is $\lim_{n\to\infty} np(1-p) = \lim_{n\to\infty} n(1-\lambda/n)\lambda/n = \lambda$.
- **Sum of λ_1 Poisson and independent λ_2 Poisson** is a $\lambda_1 + \lambda_2$ Poisson.
- **Times between successive events** in λ Poisson process are independent exponentials with parameter λ.
- **Minimum of independent exponentials** with parameters λ_1 and λ_2 is itself exponential with parameter $\lambda_1 + \lambda_2$.

- **DeMoivre-Laplace limit theorem (special case of central limit theorem):**

$$\lim_{n\to\infty} P\{a \le \frac{S_n - np}{\sqrt{npq}} \le b\} \to \Phi(b) - \Phi(a).$$

- This is $\Phi(b) - \Phi(a) = P\{a \le X \le b\}$ when X is a standard normal random variable.

- Toss a million fair coins. Approximate the probability that I get more than $501,000$ heads.
- Answer: well, $\sqrt{npq} = \sqrt{10^6 \times .5 \times .5} = 500$. So we're asking for probability to be over two SDs above mean. This is approximately $1 - \Phi(2) = \Phi(-2)$.
- Roll 60000 dice. Expect to see 10000 sixes. What's the probability to see more than 9800?
- Here $\sqrt{npq} = \sqrt{60000 \times \frac{1}{6} \times \frac{5}{6}} \approx 91.28$.
- And $200/91.28 \approx 2.19$. Answer is about $1 - \Phi(-2.19)$.

- Say X is a (standard) **normal random variable** if $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.
- Mean zero and variance one.
- The random variable $Y = \sigma X + \mu$ has variance σ^2 and expectation μ.
- Y is said to be normal with parameters μ and σ^2. Its density function is $f_Y(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$.
- Function $\Phi(a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2} dx$ can't be computed explicitly.
- Values: $\Phi(-3) \approx .0013$, $\Phi(-2) \approx .023$ and $\Phi(-1) \approx .159$.
- Rule of thumb: "two thirds of time within one SD of mean, 95 percent of time within 2 SDs of mean."

Gamma Distribution

- Say X is an **exponential random variable of parameter** λ when its probability distribution function is $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$ (and $f(x) = 0$ if $x < 0$).
- For $a > 0$ have

$$F_X(a) = \int_0^a f(x)dx = \int_0^a \lambda e^{-\lambda x} dx = -e^{-\lambda x}\Big|_0^a = 1 - e^{-\lambda a}.$$

- Thus $P\{X < a\} = 1 - e^{-\lambda a}$ and $P\{X > a\} = e^{-\lambda a}$.
- Formula $P\{X > a\} = e^{-\lambda a}$ is very important in practice.
- Repeated integration by parts gives $E[X^n] = n!/\lambda^n$.
- If $\lambda = 1$, then $E[X^n] = n!$. Value $\Gamma(n) := E[X^{n-1}]$ defined for real $n > 0$ and $\Gamma(n) = (n-1)!$.

- Say that random variable X has gamma distribution with parameters (α, λ) if $f_X(x) = \begin{cases} \frac{(\lambda x)^{\alpha-1} e^{-\lambda x} \lambda}{\Gamma(\alpha)} & x \geq 0 \\ 0 & x < 0 \end{cases}$.
- Same as exponential distribution when $\alpha = 1$. Otherwise, multiply by $x^{\alpha-1}$ and divide by $\Gamma(\alpha)$. The fact that $\Gamma(\alpha)$ is what you need to divide by to make the total integral one just follows from the definition of Γ.
- Waiting time interpretation makes sense only for integer α, but distribution is defined for general positive α.

Uniform distribution:

- Suppose X is a random variable with probability density function $f(x) = \begin{cases} \frac{1}{\beta-\alpha} & x \in [\alpha, \beta] \\ 0 & x \notin [\alpha, \beta]. \end{cases}$
- Then $E[X] = \frac{\alpha+\beta}{2}$.
- And $\mathrm{Var}[X] = \mathrm{Var}[(\beta - \alpha)Y + \alpha] = \mathrm{Var}[(\beta - \alpha)Y] = (\beta - \alpha)^2 \mathrm{Var}[Y] = (\beta - \alpha)^2/12$.

Independent

- We say X and Y are independent if for any two (measurable) sets A and B of real numbers we have

$$P\{X \in A, Y \in B\} = P\{X \in A\}P\{Y \in B\}.$$

- When X and Y are discrete random variables, they are independent if $P\{X = x, Y = y\} = P\{X = x\}P\{Y = y\}$ for all x and y for which $P\{X = x\}$ and $P\{Y = y\}$ are non-zero.
- When X and Y are continuous, they are independent if $f(x, y) = f_X(x)f_Y(y)$.

Sum two independent variables:

- Say we have independent random variables X and Y and we know their density functions f_X and f_Y.
- Now let's try to find $F_{X+Y}(a) = P\{X + Y \leq a\}$.
- This is the integral over $\{(x, y) : x + y \leq a\}$ of $f(x, y) = f_X(x)f_Y(y)$. Thus,
- $$P\{X + Y \leq a\} = \int_{-\infty}^{\infty} \int_{-\infty}^{a-y} f_X(x)f_Y(y)dxdy$$
  $$= \int_{-\infty}^{\infty} F_X(a - y)f_Y(y)dy.$$
- Differentiating both sides gives $f_{X+Y}(a) = \frac{d}{da} \int_{-\infty}^{\infty} F_X(a-y)f_Y(y)dy = \int_{-\infty}^{\infty} f_X(a-y)f_Y(y)dy$.
- Latter formula makes some intuitive sense. We're integrating over the set of x, y pairs that add up to a.

Order statistics

- Consider i.i.d random variables $X_1, X_2, \ldots, X_n$ with continuous probability density f.
- Let $Y_1 < Y_2 < Y_3 \ldots < Y_n$ be list obtained by *sorting* the X_j.
- In particular, $Y_1 = \min\{X_1, \ldots, X_n\}$ and $Y_n = \max\{X_1, \ldots, X_n\}$ is the maximum.
- What is the joint probability density of the Y_i?
- Answer: $f(x_1, x_2, \ldots, x_n) = n! \prod_{i=1}^{n} f(x_i)$ if $x_1 < x_2 \ldots < x_n$, zero otherwise.
- Let $\sigma : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ be the permutation such that $X_j = Y_{\sigma(j)}$
- Are σ and the vector $(Y_1, \ldots, Y_n)$ independent of each other?
- Yes.

Expectation:

- For both discrete and continuous random variables X and Y we have $E[X+Y] = E[X] + E[Y]$.
- In both discrete and continuous settings, $E[aX] = aE[X]$ when a is a constant. And $E[\sum a_i X_i] = \sum a_i E[X_i]$.
- But what about that delightful "area under $1 - F_X$" formula for the expectation?
- When X is non-negative with probability one, do we always have $E[X] = \int_0^\infty P\{X > x\}$, in both discrete and continuous settings?
- Define $g(y)$ so that $1 - F_X(g(y)) = y$. (Draw horizontal line at height y and look where it hits graph of $1 - F_X$.)
- Choose Y uniformly on $[0,1]$ and note that $g(Y)$ has the same probability distribution as X.
- So $E[X] = E[g(Y)] = \int_0^1 g(y)dy$, which is indeed the area under the graph of $1 - F_X$.

Covariance:

- **General statement of bilinearity of covariance:**

$$\mathrm{Cov}(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \mathrm{Cov}(X_i, Y_j).$$

- Special case:

$$\mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2 \sum_{(i,j):i<j} \mathrm{Cov}(X_i, X_j).$$

Conditional Expectation:

- Can think of $E[X|Y]$ as a function of the random variable Y. When $Y = y$ it takes the value $E[X|Y = y]$.
- So $E[X|Y]$ is itself a random variable. It happens to depend only on the value of Y.
- Thinking of $E[X|Y]$ as a random variable, we can ask what *its* expectation is. What is $E[E[X|Y]]$?
- **Very useful fact:** $E[E[X|Y]] = E[X]$.
- In words: what you expect to expect X to be *after learning* Y is same as what you *now* expect X to be.
- Proof in discrete case:
  $E[X|Y = y] = \sum_x xP\{X = x|Y = y\} = \sum_x x\frac{p(x,y)}{p_Y(y)}$.
- Recall that, in general, $E[g(Y)] = \sum_y p_Y(y)g(y)$.
- $E[E[X|Y = y]] = \sum_y p_Y(y) \sum_x x\frac{p(x,y)}{p_Y(y)} = \sum_x \sum_y p(x, y)x = E[X]$.

Conditional Variance:

- Definition:
  $\mathrm{Var}(X|Y) = E\left[(X - E[X|Y])^2|Y\right] = E\left[X^2 - E[X|Y]^2|Y\right]$.
- $\mathrm{Var}(X|Y)$ is a random variable that depends on Y. It is the variance of X in the conditional distribution for X given Y.
- Note $E[\mathrm{Var}(X|Y)] = E[E[X^2|Y]] - E[E[X|Y]^2|Y] = E[X^2] - E[E[X|Y]^2]$.
- If we subtract $E[X]^2$ from first term and add equivalent value $E[E[X|Y]]^2$ to the second, RHS becomes $\mathrm{Var}[X] - \mathrm{Var}[E[X|Y]]$, which implies following:
- **Useful fact:** $\mathrm{Var}(X) = \mathrm{Var}(E[X|Y]) + E[\mathrm{Var}(X|Y)]$.
- One can discover X in two stages: first sample Y from marginal and compute $E[X|Y]$, then sample X from distribution given Y value.
- Above fact breaks variance into two parts, corresponding to these two stages.

Moment Generating Function:

- Let X be a random variable and $M(t) = E[e^{tX}]$.
- Then $M'(0) = E[X]$ and $M''(0) = E[X^2]$. Generally, nth derivative of M at zero is $E[X^n]$.
- Let X and Y be independent random variables and $Z = X + Y$.
- Write the moment generating functions as $M_X(t) = E[e^{tX}]$ and $M_Y(t) = E[e^{tY}]$ and $M_Z(t) = E[e^{tZ}]$.
- If you knew M_X and M_Y, could you compute M_Z?
- By independence, $M_Z(t) = E[e^{t(X+Y)}] = E[e^{tX}e^{tY}] = E[e^{tX}]E[e^{tY}] = M_X(t)M_Y(t)$ for all t.

- We showed that if $Z = X + Y$ and X and Y are independent, then $M_Z(t) = M_X(t)M_Y(t)$
- If $X_1 \ldots X_n$ are i.i.d. copies of X and $Z = X_1 + \ldots + X_n$ then what is M_Z?
- Answer: M_X^n. Follows by repeatedly applying formula above.
- This a big reason for studying moment generating functions. It helps us understand what happens when we sum up a lot of independent copies of the same random variable.
- If $Z = aX$ then $M_Z(t) = E[e^{tZ}] = E[e^{taX}] = M_X(at)$.
- If $Z = X + b$ then $M_Z(t) = E[e^{tZ}] = E[e^{tX+bt}] = e^{bt}M_X(t)$.

- If X is binomial with parameters (p, n) then $M_X(t) = (pe^t + 1 - p)^n$.
- If X is Poisson with parameter $\lambda > 0$ then $M_X(t) = \exp[\lambda(e^t - 1)]$.
- If X is normal with mean 0, variance 1, then $M_X(t) = e^{t^2/2}$.
- If X is normal with mean μ, variance σ^2, then $M_X(t) = e^{\sigma^2 t^2/2+\mu t}$.
- If X is exponential with parameter $\lambda > 0$ then $M_X(t) = \frac{\lambda}{\lambda - t}$.

Characteristic functions

- The **characteristic function** of X is defined by $\phi(t) = \phi_X(t) := E[e^{itX}]$. Like $M(t)$ except with i thrown in.
- Recall that by definition $e^{it} = \cos(t) + i\sin(t)$.
- Characteristic functions are similar to moment generating functions in some ways.
- For example, $\phi_{X+Y} = \phi_X\phi_Y$, just as $M_{X+Y} = M_X M_Y$.
- And $\phi_{aX}(t) = \phi_X(at)$ just as $M_{aX}(t) = M_X(at)$.
- And if X has an mth moment then $E[X^m] = i^m\phi_X^{(m)}(0)$.
- But characteristic functions have a distinct advantage: they are always well defined for all t even if f_X decays slowly.

Cauchy Distribution

- A standard **Cauchy random variable** is a random real number with probability density $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.
- There is a "spinning flashlight" interpretation. Put a flashlight at $(0,1)$, spin it to a uniformly random angle in $[-\pi/2, \pi/2]$, and consider point X where light beam hits the x-axis.
- $F_X(x) = P\{X \le x\} = P\{\tan\theta \le x\} = P\{\theta \le \tan^{-1}x\} = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}x$.
- Find $f_X(x) = \frac{d}{dx}F(x) = \frac{1}{\pi}\frac{1}{1+x^2}$.

Beta Distribution

- Two part experiment: first let p be uniform random variable $[0,1]$, then let X be binomial (n,p) (number of heads when we toss n p-coins).
- **Given** that $X = a-1$ and $n - X = b-1$ the conditional law of p is called the β distribution.
- The density function is a constant (that doesn't depend on x) times $x^{a-1}(1-x)^{b-1}$.
- That is $f(x) = \frac{1}{B(a,b)}x^{a-1}(1-x)^{b-1}$ on $[0,1]$, where $B(a,b)$ is constant chosen to make integral one. Can show $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.
- Turns out that $E[X] = \frac{a}{a+b}$ and the mode of X is $\frac{(a-1)}{(a-1)+(b-1)}$.

Central Limit Theorem

- Let X_i be an i.i.d. sequence of random variables with finite mean μ and variance σ^2.
- Write $S_n = \sum_{i=1}^n X_i$. So $E[S_n] = n\mu$ and $\mathrm{Var}[S_n] = n\sigma^2$ and $\mathrm{SD}[S_n] = \sigma\sqrt{n}$.
- Write $B_n = \frac{X_1+X_2+\ldots+X_n-n\mu}{\sigma\sqrt{n}}$. Then B_n is the difference between S_n and its expectation, measured in standard deviation units.
- **Central limit theorem:**

$$\lim_{n\to\infty} P\{a \le B_n \le b\} \to \Phi(b) - \Phi(a).$$

Law of Large Numbers

- Suppose X_i are i.i.d. random variables with mean μ.
- Then the value $A_n := \frac{X_1+X_2+\ldots+X_n}{n}$ is called the *empirical average* of the first n trials.
- Intuition: when n is large, A_n is typically close to μ.
- Recall: **weak law of large numbers** states that for all $\epsilon > 0$ we have $\lim_{n\to\infty} P\{|A_n - \mu| > \epsilon\} = 0$.
- The **strong law of large numbers** states that with probability one $\lim_{n\to\infty} A_n = \mu$.
- It is called "strong" because it implies the weak law of large numbers. But it takes a bit of thought to see why this is the case.

Markov chain

- Consider a sequence of random variables $X_0, X_1, X_2, \ldots$ each taking values in the same state space, which for now we take to be a finite set that we label by $\{0, 1, \ldots, M\}$.
- Interpret X_n as state of the system at time n.
- Sequence is called a **Markov chain** if we have a fixed collection of numbers P_{ij} (one for each pair $i, j \in \{0, 1, \ldots, M\}$) such that whenever the system is in state i, there is probability P_{ij} that system will next be in state j.
- Precisely, $P\{X_{n+1} = j | X_n = i, X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0\} = P_{ij}$.
- Kind of an "almost memoryless" property. Probability distribution for next state depends only on the current state (and not on the rest of the state history).

- Say Markov chain is **ergodic** if some power of the transition matrix has all non-zero entries.
- Turns out that if chain has this property, then $\pi_j := \lim_{n\to\infty} P_{ij}^{(n)}$ exists and the π_j are the unique non-negative solutions of $\pi_j = \sum_{k=0}^{M} \pi_k P_{kj}$ that sum to one.
- This means that the row vector

$$\pi = \begin{pmatrix} \pi_0 & \pi_1 & \ldots & \pi_M \end{pmatrix}$$

is a left eigenvector of A with eigenvalue 1, i.e., $\pi A = \pi$.
- We call π the *stationary distribution* of the Markov chain.

Markov and Chebyshev Inequalities

Markov's Inequality

If X is any nonnegative random variable, then

$$P(X \geq a) \leq \frac{EX}{a}, \qquad \text{for any } a > 0.$$

Chebyshev's Inequality

If X is any random variable, then for any $b > 0$ we have

$$P\big(|X - EX| \geq b\big) \leq \frac{Var(X)}{b^2}.$$