Probability Theory

Probability space ≡ sample space Ω with probability $Pr[w]$ gives us:
$0 \le Pr[w] \le 1$ for $w \in \Omega$, $\sum_{w \in \Omega} Pr[w] = 1$

Uniform Probability
For event A, $Pr[A] = \sum_{w \in A} Pr[w]$ and when uniform, $Pr[A] = |A|/|\Omega|$
Nonuniform probability assigns weights to events, string of tosses -> # heads

Conditional Probability
For $A, B \in \Omega$, $P[B] > 0$, $P[A|B] = \frac{P[A \cap B]}{P[B]}$
Bayes' Rule $P[A|B] = \frac{P[B|A]P[A]}{P[B]}$
Sum out to obtain $P[B]$
Total P Rule: $P(B) = P(A, B) + P(\bar{A}, B)$

Counting

First Rule: If object result of k successive choices, n_1 first, n_2 second given first, ..., $n_1 \cdot n_2 \cdot n_3 \cdot \ldots \cdot n_k$ looks like tree pictorally

Second Rule: If object is result of succession of choices where order does not matter, can choose as if it does then map by a k to 1 function. $\binom{n}{k} = \frac{n!}{(n-k)!k!}$

Sampling w/ replacement => multiple trials w/ same outcome

Zeroth Rule: If set A can be placed in one-to-one correspondance w/ set B, then $|A| = |B|$.

Stars and Bars: Partition collection of n objects into k separate "bins" $\binom{n+k-1}{n} = \binom{n+k-1}{k-1}$ binary string of items and bars

Combinatorial Proofs
Use counting arguments/"stories" to prove statements like: $\binom{n}{k+1} = \binom{n-1}{k} + \binom{n-2}{k} + \cdots + \binom{k}{k}$
or $\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$

Probability Counting Examples

Stirling's Approximation: $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$
$\binom{n}{r} = \frac{n!}{r!(n-r)!} \approx \frac{1}{\sqrt{2\pi}\sqrt{nq(1-q)}}\left(\frac{1}{q^q(1-q)^{(1-q)}}\right)^n$

Balls and Bins: Given n bins, m balls
$P[\text{bin 1 empty}] = \left(\frac{n-1}{n}\right)^m$, $P[\text{bin 1} \ge 1] = 1 - P[\text{bin 1 empty}]$

Birthday Paradox: $P[\text{no two have same birthday}] = \frac{365 \times 364 \times 363 \times \ldots \times (365-n+1)}{365^n}$

Binom RV: $P[A_r] = \binom{n}{r} p^r (1-p)^r$, Stirling's approximation shows $P(\text{exactly } 1/2)$ decreasing as $1/\sqrt{n}$

Probability of Combinations of Events

For disjoint events A_i, $Pr[\bigcup_{i=1}^n A_i] = \sum_{i=1}^n Pr[A_i]$ (finite additivity)

Two events A and B are independent if $P[A \cap B] = P[A]P[B]$ and similarly $P[A|B] = P[A]$

Mutual independence: $Pr[\bigcap_{i \in I} A_i] = \prod_{i \in I} Pr[A_i]$ for every subset I of $\{1 \ldots n\}$

Pairwise independence: same but only with pair subsets

Intersections of Events (AND)
(Product Rule) $Pr[\bigcap_{i=1}^n A_i] = \prod_{i=1}^n Pr[A_i | \bigcap_{j=1}^{i-1} A_j]$
Serves as an alternative to counting approach

Union of Events (OR)
Inclusion/Exclusion: $Pr[\bigcup_{i=1}^n A_i] = \sum_{i=1}^n Pr[A_i] - \sum_{\{i,j\}} Pr[A_i \cap A_j] + \sum Pr[A_i \cap A_j \cap A_k]$
Visualize as over/undercounting Venn diagram $- \ldots \pm Pr[\bigcap_{i=1}^n A_i]$
Disjoint events: If all events A_i are disjoint, then $Pr[\bigcup_{i=1}^n A_i] = \sum_{i=1}^n Pr[A_i]$
Union bound: $Pr[\bigcup_{i=1}^n A_i] \le \sum_{i=1}^n Pr[A_i]$ (adding P_i's usually will overestimate)

Random Variables

Random variable X on a sample space Ω is a function that assigns to each sample point $w \in \Omega$ some real number $X(w)$.

The distribution of discrete random variable X is collection of values $\{(a, Pr[X=a]) : a \in A\}$, A = set of all possible values taken on by X.
Any two events $X = a_1$, $X = a_2$ w/ $a_1 \ne a_2$ are disjoint.
The union of these events ≡ Ω

Probability mass function $P[X=i] = Pr[\{w \in \Omega | X(w) = i\}]$
$0 \le P_X(i) \le 1$, $\sum_i P_X(i) = 1$
Given joint PMF $P_{X,Y}(i,j)$ can sum $\sum_j P_{X,Y}(i,j) = P_X(i)$
Binomial Distribution: $P[X=i] = \binom{n}{i} p^i (1-p)^{n-i}$
$Pr[X \ge n] = \sum_{i=n}^{n+k} \binom{n+k}{i} (1-p)^i p^{n+k-i}$ (packets getting through)

Independent Random Variables -> $P[X=a, Y=b] = P[X=a]P[Y=b]$

Expectation
$E(X) = \sum_{a \in A} a \times Pr[X=a]$ (summed over all values)

Linearity of Expectation: $E(X+Y) = E(X) + E(Y)$, $E(cX) = cE(X)$
Proof: $E(X) = \sum_{a \in A} a \times Pr[X=a] = \sum_{w \in \Omega} X(w) \times Pr(w)$
$E(X+Y) = \sum_{w \in \Omega} (X+Y)(w) \times Pr[w] = \sum_{w \in \Omega} (X(w) + Y(w)) \times Pr[w]$
$= \sum_{w \in \Omega} X(w) \times Pr[w] + \sum_{w \in \Omega} Y(w) \times Pr[w] = E[X] + E[Y]$

If independent X and Y, we have: $E(XY) = E(X)E(Y)$
$E(XY) = \sum_a \sum_b ab \times Pr[X=a, Y=b]$
$= \sum_a \sum_b ab \times Pr[X=a] \times Pr[Y=b]$
$= \sum_a a Pr[X=a] \sum_b b Pr[X=b]$
$= \left(\sum_a a Pr[X=a]\right)\left(\sum_b b Pr[X=b]\right)$
$= E(X) \times E(Y)$

Variance

For random variable X with $E(X) = \mu$, variance $Var(X) = E((X-\mu)^2)$
$Var(X) = E[X^2] - E[X]^2$ by $E(X^2 - 2X\mu + \mu^2) = E[X^2] - E[2X\mu] + E[\mu^2]$
$= E[X^2] - 2\mu^2 + \mu^2 = E[X^2] - E[X]^2$

For independent random variables: $Var(cX) = c^2 Var(X)$, $Var(X+Y) = Var(X) + Var(Y)$
$Var(cX) = E[cX^2] + (E[cX])^2 = c^2 E[X^2] + (cE[X])^2 = c^2(E[X^2] - E[X]^2)$
$Var(X+Y) = E[(X+Y)^2] - E[X+Y]^2 = E[X^2 + 2XY + Y^2] - (E[X] + E[Y])^2$
$= E[X^2] + 2E[XY] + E[Y^2] - E[X]^2 - 2E[X]E[Y] - E[Y]^2$
$= (E[X^2] - E[X]^2) + (E[Y^2] - E[Y]^2) + 2E[XY] - 2E[X]E[Y]$
$= Var(X) + Var(Y)$

Bounds

Name	Assumptions	Bound/Approx	When to Use	
Markov	$X \ge 0$	$P(X \ge \alpha) \le \frac{E[X]}{\alpha}$	To prove other bounds, or when you have nothing else.	
Chebyshev	a) $E[X] = \mu$, $Var(X) = \sigma^2$; b) X_i iid, pairwise uncorrelated, $A_n = \frac{1}{n}\sum_{i=1}^n X_i$	$P(\|X-\mu\| \ge \epsilon) \le \frac{Var(X)}{\epsilon^2}$; $P(\|A_n - \mu\| \ge \epsilon) \le \frac{Var(Y)}{n\epsilon^2}$	To bound two sided tails when you have variance and nothing else.	Can optimize, like choose worst case variance
Hoeffding	X_i iid, $E[X_i] = \mu$ bounded, $X_i \in [a,b]$ ✓	$P(A_n \ge \mu + \epsilon) \le e^{-n\left(\frac{2\epsilon^2}{(b-a)^2}\right)}$	Bounds a one-sided tail; use when you have bounded rv and independence	want exponential bound but don't need it to be tight
CLT	X_i independent, $E[X_i] = \mu$, $Var(X_i) = \sigma^2$	$P(Z_n \in [a,b]) \approx \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$; $Z_n = \frac{\sum X_i - n\mu}{\sqrt{n\sigma^2}}$	When an approximation is good enough, but you want the probability "correct" not worse	
Chernoff	X_i iid, $E[e^{sX}]$ exists for $s>0$, $X_i \sim Bern(p)$	$P(A_n \ge a) \le e^{-n\Phi_X(a)}$; $\Phi_X = \max_{s>0}(sa - \ln E[e^{sX}])$ almost always need to use calculus; $\Phi_X(a) = D(a \| p)$	one sided tail; Exponentially small and tight in exponent	(Last resort)

Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be iid random variables with common expectation $\mu = E(X)$. Define $A_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then, for any $\alpha > 0$, we have $Pr[|A_n - \mu| \ge \alpha] \to 0$ as $n \to \infty$

Proof: Let $Var(X_i) = \sigma^2$
$Pr[|A_n - \mu| \ge \alpha] \le \frac{Var(A_n)}{\alpha^2} = \frac{\sigma^2}{n\alpha^2} \to 0$ as $n \to \infty$

Bernoulli: $\to E[e^{sX_i}] = \sum Pr(X_i = x) e^{sX_i}$
$\Phi_{X_i}(a) = a \ln\frac{a}{p} + (1-a)\ln\frac{1-a}{1-p} = pe^s + (1-p)(1)$

Distributions

Geometric distribution
A random variable X with distribution $P[X=i] = (1-p)^{i-1}p$
$E[X] = \sum_{i=1}^{\infty} Pr[X \ge i] \to Pr[X \ge i] = (1-p)^{i-1}$
and we get $\frac{1}{1-(1-p)} = \frac{1}{p}$
$Var(X) = \frac{1-p}{p^2}$ $\quad E[X^2] - E[X]^2$
$E[X^2] = p + (E[X^2] + 2E[X] + 1)(1-p)$
$E[X^2](1 - (1-p)) = p + (2/p + 1)(1-p)$
$p(E[X^2]) = p + \left(\frac{2}{p} + 1\right)(1-p)$
$E[X^2] = 1 + \left(\frac{2}{p^2} + \frac{1}{p}\right)(1-p)$
$= \cancel{1} + \left(\frac{2}{p^2} + \frac{1}{p}\right) - \left(\frac{2}{p} + 1\right)$
$= \frac{2}{p^2} - \frac{1}{p}$
$E[X^2] - E[X]^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1}{p^2} - \frac{1}{p} = \frac{1-p}{p^2}$

Coupon Collectors
Each new coupon is essentially a geometric random variable.
Expect $\frac{n}{n}$ then $\frac{n-1}{n}$ .... to get
$E[X] = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \ldots = n\sum_{i=1}^n \frac{1}{i}$
$\sum_{i=1}^n \frac{1}{i} \approx \ln n + \gamma$, where $\gamma = 0.5772$

Poisson Distribution
$Pr[X=i] = \frac{\lambda^i}{i!} e^{-\lambda}$ for $i = 0, 1, 2, \ldots$

Chernoff derivation
$P(X \ge a) = P(e^{sX} \ge e^{sa}) \le E[e^{sX}]e^{-sa}$
$P(X \ge a) \le \min_{s>0}(E[e^{sX}]e^{-sa}) = e^{-\Phi_X(a)}$
$\Phi_X(a) = \max_{s \ge 0}(sa - \ln E[e^{sX}])$
$P(\sum X_i \ge na) \le e^{-\Phi_{\sum_{i=1}^n X_i}(na)}$ where
$\Phi_{\sum_{i=1}^n X_i}(na) = \max_{s>0}(sna - \ln E[e^{s\sum_{i=1}^n X_i}])$
but $e^{s\sum_{i=1}^n X_i} = \prod_{i=1}^n e^{sX_i} \Rightarrow \ln \prod_{i=1}^n E[e^{sX_i}]$
then $\Phi_{\sum_{i=1}^n X_i}(na) = n\max(sa - \ln E[e^{sX_i}])$
$Pr\left(\frac{1}{n}\sum_{i=1}^n X_i \ge a\right) \le e^{-n\Phi_X(a)}$