Hashing, using balls and bins with the union bound.

- Run hash function, result is bucket number. Trying to avoid collisions.
- m keys (balls), n buckets (bins). More than 1 ball per bin is a collision.
- $k = \binom{k}{2}$ pairs among keys. A_i means pair i has collision.
- Pr[at least 1 collision] $\leq \sum_{i=1}^{k} Pr[A_i] = k\frac{1}{n} = \frac{m(m-1)}{2n} \approx \frac{m^2}{2n}$. This is upper bound, so set this $\leq$ allowed percentage of collisions.

Random Variables:

- A **random variable** X on a sample space Ω is a function that assigns to each sample point $\omega \in \Omega$ a real number $X(\omega)$.
- Discrete random variables take values in a range that is finite or countably infinite.
- Since a random variable is defined on a probability space, we can calculate these probabilities given the probabilities of the sample points.
- Let a be any number in the range of a random variable X. Then the set $\{\omega \in \Omega : X(\omega) = a\}$ is an event in the sample space.
- The **distribution** of a discrete random variable X is the collection of values $\{(a, Pr[X = a]) : a \in \mathscr{A}\}$
- The collection of events $X = a, a \in \mathscr{A}$, satisfy two important properties:
  - any two events $X = a_1$ and $X = a_2$ with $a_1 \neq a_2$ are disjoint
  - the union of all these events is equal to the entire sample space Ω

Concepts:

- definition of distribution
- definition of random variable

linarity of expectation

- $E(X + Y) = E(X) + E(Y)$
- $E(cX) = cE(X)$

Properties and Methods:

- How to calculate stuff, linearity, etc.

Variance:

- "Spread" of distances from the mean.
- $Var(X) = E((X - \mu)^2) = E(X^2) - \mu^2$
- Standard deviation, $\sigma = sqrtVar(X)$
- $Var(cX) = c^2 Var(X)$
- $Var(X + Y) = Var(X) + Var(Y)$ (independent)
- $E(XY) = E(X)E(Y)$ (independent)
- $Var(X + Y) = Var(X) + Var(Y) + 2(E(XY) - E(X)E(Y))$

If X is uniform random variable {1..n} with prob 1/n, then:

- $E(X) = \frac{n+1}{2}$
- $\mathrm{Var}(X) = \frac{n^2-1}{12}$
- $\sigma(X) = \sqrt{\frac{b^2-1}{12}}$
- $Var(X_i) = E(X_i^2) - E(X_i)^2$

Binomial Distributions:

- $E(X) = np$
- $Var(X_i) = E(X_i^2) - E(X_i)^2 = p - p^2 \rightarrow Var(X) = np(1-p)$
- $P[X = i] = \binom{n}{i} p^i (1-p)^{n-i}$

Formulas

- Chebyshev (random var X with $E(X) = \mu$, $\alpha > 0$): $Pr[|X - \mu| \geq \alpha] \leq \frac{Var(X)}{\alpha^2}$
- Chebyshev part 2: (random var X with $E(X) = \mu$, $\sigma = \sqrt{Var(X)}$): $Pr[|X - \mu| \geq \beta\sigma] \leq \frac{1}{\beta^2}$
- Markov's inequality ($X > 0$, $E(X) = \mu$, $\alpha > 0$): $Pr[X \geq \alpha] \leq \frac{E(X)}{\alpha}$)

Sampling:

- p proportion, with $\epsilon = 0.1$ and confidence 95%

Geometric Distribution

- Flip n coins, stop after first heads. P($heads$) = p.
- $P[X = i] = (1-p)^i p$
- $E[X] = \frac{1}{p}$
- $Var(X) = \frac{1-p}{p^2}$

Continuous distribution

- integral = 1
- $E[X] = \int_{-\infty}^{\infty} XP(X)dx = \sum XP(X)$
- $P[X_0 < X < X_1] = \int_{X_0}^{X_1} XP(X)dx$
- $Var(x) = E[(X - \mu)^2] = E[X^2] - \mu^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$

Poisson

- $P[X] = 0$ iff $x < 0$ AND $P[x] = \lambda e^{-\lambda x}$ for $x \geq 0$
- $E[X] = \lambda$
- $P[X = i] = \frac{\lambda^i e^{-\lambda}}{i!}$
- $Var(x) = \lambda$

Probability Density Function

- $f(x) = 0$ for $x < 0$ and $x > \ell$
- $f(x) = \frac{1}{\ell}$ for $0 \leq x \leq \ell$
- $E(X) = \int_{-\infty}^{\infty} xf(x)dx = \frac{\ell}{2}$
- $Var(X) = \frac{\ell^2}{12}$

Formulas

- Joint Density function: $Pr[a \leq X \leq b, c \leq Y \leq d] = \int_c^d \int_a^b f(x, y)dxdy$
- probability per unit area: $Pr[x \leq X \leq x + \delta, y \leq Y \leq y + \delta] = \int_y^{y+\delta} \int_x^{x+\delta} f(u, v)dudv \approx \delta^2 f(x, y)$
- Independent variables x, y, then $f(x, y) = f(x)f(y)$

Exponential Distribution

- $f(x) = \lambda e^{-\lambda x}$ for $x > 0$, otherwise 0.
- $E(x) = \frac{1}{\lambda}$
- $Var(x) = \frac{1}{\lambda^2}$
- $P[X \geq C] = \int_C^\infty \lambda e^{-\lambda x} dx$

Normal Distribution

- Parameters σ and μ, centered at $x = \mu$, standard deviation is σ
- standard normal: $\mu = 0$ and $\sigma = 1$

- $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$
- $E(x) = \mu$
- $Var(x) = \sigma^2$
- $P[X \geq C] = \int_C^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} dx$ (just integrate the interval)
- Note: 68% of the data lies within 1σ from the mean, 95% within 2σ, and 99.7% within 3σ

Central Limit Theorem

- $A_n = \frac{\sqrt{n}(A_n - \mu)}{\sigma} = \frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$
- $\sigma_\mu = \frac{\sigma}{\sqrt{n}}$. plug in to normal pdf and intergate to get intervals.

How to Lie with Statistics:

- **Simpsons paradox:** A paradox in which a trend that appears in different groups of data disappears when these groups are combined. Ie, if more of a group applies to a harder department, even though individual department rates may be fair, overall trend seems discriminatory.

Infinity and Countability

- **Cardinality:** In order to determine whether two sets have the same cardinality, we need to demonstrate a *bijection* f between the two sets.

  We say that a set S is **countable** if there is a bijection between S and $\mathbb{N}$ or some subset of $\mathbb{N}$. Thus any finite set S is countable (since there is a bijection between S and the subset $\{0, 1, 2, \ldots, m-1\}$, where $m =\mid S \mid$ is the size of S).

  If there is a one-to-one function $f : A \to B$, then the cardinality of A is less than or equal to that of B. Now to show that the cardinality of A and B are the same we can show that $\mid A \mid \leq \mid B \mid$ and $\mid B \mid \leq \mid A \mid$.
- **Cantors Diagonalization:** a mathematical proof that there are infinite sets which cannot be put into one-to-one correspondence with the infinite set of natural numbers

  Let S be any set. Then the power set of S, denoted by $\mathscr{P}(S)$, is the set of all subsets of S. More formally, it is defined as: $\mathscr{P}(S) = \{T : T \subseteq S\}$.

Sets

- $\mathbb{Z}$ - integers (countable. f(x)=2x, f(-x)=-2x+1)
- $\mathbb{N}$ - natural (countable)
- $\mathbb{Q}$ - rational (countable - N/N. also spiral method)
- $\mathbb{R}$ = real (uncountable - always can find avg(a,b))
- C - complex. (?)
- Union of countable and uncountable must yield uncountable set.
- injection (1:1) - $\forall x, y \in A, x \neq y \to f(x) \neq f(y)$
- bijection (onto) - $\forall y \in B, \exists x \in A$ such that $f(x) = y$
- surjection (bijection) - both one to one and onto, definition of inverse.

The Liars Paradox: The barber proclaims: I shave all and only those men who do not shave themselves." It seems reasonable then to ask the question: Does the barber shave himself? If the barber does not shave himself, then according to what he announced, he shaves himself. If the barber does shave himself, then according to his statement he does not shave himself!

The Halting Problem: Given the description of a program and its input, we would like to know if the program ever halts when it is executed on the given input.

$$TestHalt(P, I) = \begin{cases} \text{"yes"} & \text{if program P halts on input I} \\ \text{"no"} & \text{if program P loops on input I} \end{cases}$$

Proof: Define the program

Turing(P)
  if TestHalt(P,P) = "yes" then loop forever
  else halt

So if the program P when given P as input halts, then Turing loops forever; otherwise, Turing halts.

TestHalt cannot exist, so it is impossible for a program to check if any general program halts.

Calculus Refresher
Integration

- Basics: $\int kxdx = kx^2 + C$, $\int e^u = e^u + C$
- By parts: $\int udv = uv - \int vdu$. Pick u and dv, find du and v, solve.

Differentiation

- Basic idea: Bring exponent down and decrement. $\frac{\partial}{\partial x} 3x^2 = 6x$, $\frac{\partial}{\partial x} x = 1$, $\frac{\partial}{\partial x} e^x = e^x$, $\frac{\partial}{\partial x} a^x = a^x ln(a)$, $\frac{\partial}{\partial x} sin(x) = cos(x)$, $cos \to -sin$

1. Relax.
2. You will do GREAT!
3. The "A" is yours.

CS70 Final Study Guide, by Ivan "Vania" Smirnov