Outline

Balls in Bins.
Birthday.
Coupon Collector.
Load balancing.

Geometric Distribution: Memoryless property.
Poission Distribution: Sum of two Poission is Poission.

Tail Sum for Expectation.
Regression (optional.)

Balls in bins

One throws \( m \) balls into \( n > m \) bins.

Theorem:
\[
\Pr[\text{no collision}] \approx \exp\left(-\frac{m^2}{2n}\right), \text{ for large enough } n.
\]

In particular,
\[
\Pr[\text{no collision}] \approx \frac{1}{2} \text{ for } m = \sqrt{n}. (e^{-0.5} \approx 0.6.)
\]

The Calculation.

\( A_i = \text{no collision when } i\text{th ball is placed in a bin.} \)
\[
\Pr[A_i|A_{i-1} \cap \cdots \cap A_1] = \left(1 - \frac{i-1}{n}\right).
\]

no collision = \( A_1 \cap \cdots \cap A_m \).

Product rule:
\[
\Pr[A_1 \cap \cdots \cap A_m] = \Pr[A_1]\Pr[A_2|A_1] \cdots \Pr[A_m|A_1 \cap \cdots \cap A_{m-1}]
\]
\[
= \Pr[\text{no collision}] = \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{m-1}{n}\right).
\]

Hence,
\[
\ln(\Pr[\text{no collision}]) = \sum_{k=1}^{m-1} \ln\left(1 - \frac{k}{n}\right) \approx \sum_{k=1}^{m-1} \left(-\frac{k}{n}\right) \quad (\ast)
\]
\[
= -\frac{m(m-1)}{2n} \quad (\ast)\text{ We used } \ln(1 - \varepsilon) \approx -\varepsilon \text{ for } |\varepsilon| < 1.
\]

(\ast) 1 + 2 + \cdots + m - 1 = (m-1)m/2.
Today's your birthday, it's my birthday too.

Probability that $m$ people all have different birthdays?

With $n = 365$, one finds

$Pr[\text{collision}] \approx 1/2$ if $m \approx 1.2\sqrt{365} \approx 23$.

If $m = 60$, we find that

$Pr[\text{no collision}] = \exp \left( -\frac{m^2}{2n} \right) = \exp \left( -\frac{60^2}{2 \times 365} \right) \approx 0.007$.

If $m = 366$, then $Pr[\text{no collision}] = 0$. (No approximation here!)

Using linearity of expectation.

Event: $m$ balls into $n$ bins uniformly at random.

Random Variable:

$X_i$ = Number of collisions between pairs of balls.

or number of pairs $i$ and $j$ where ball $i$ and ball $j$ are in same bin.

$X = \sum_i X_i$

$X = \sum_i X_i$

$E[ X_i ] = Pr[ \text{balls } i, j \text{ in same bin} ] = \frac{1}{n}$.

Ball $i$ in some bin, ball $j$ chooses that bin with probability $1/n$.

$E[ X ] = \frac{mn(m-1)}{2^n}$.

For $m = \sqrt{n}$, $E[ X ] = 1/2$.

Markov: $Pr[ X \geq c ] \leq \frac{E[ X ]}{c}$.

$Pr[ X \geq 1 ] \leq \frac{E[ X ]}{1} = 1/2$.

Coupon Collector Problem: Analysis.

Event $A_m = \text{‘fail to get Brian Wilson in } m \text{ cereal boxes’}$

Fail the first time: $(1 - \frac{1}{n})$

Fail the second time: $(1 - \frac{1}{n})$

Fail the third time: $(1 - \frac{1}{n})$

And so on ... for $m$ times. Hence,

$Pr[ A_m ] = (1 - \frac{1}{n}) \times \cdots \times (1 - \frac{1}{n})$

$= (1 - \frac{1}{n})^m$

$ln(Pr[ A_m ]) = m \times (1 - \frac{1}{n})$

$Pr[ A_m ] = \exp \left( -\frac{m}{n} \right)$.

For $n_{\rho_0} = \frac{1}{e}$, we need around $n \ln 2 \approx 0.69n$ boxes.

Approximation

Consider a set of $m$ files.

Each file has a checksum of $b$ bits.

How large should $b$ be for $Pr[\text{share a checksum}] \leq 10^{-3}$?

Claim: $b \geq 2.9 \ln(m)$.

Proof:

Let $n = 2^b$ be the number of checksums.

We know $Pr[\text{no collision}] = \exp \left( -\frac{m^2}{2n} \right) = 1 - \frac{m^2}{2n}$.

Hence,

$Pr[\text{no collision}] = 1 - 10^{-3} \Rightarrow \frac{m^2}{2n} \approx 10^{-3}$

$\Rightarrow 2n \approx m^2 10^3 \Rightarrow 2^{b+1} \approx m^2 2^{10}$

$\Rightarrow b + 1 \approx 10 + 2 \log_2(m) \approx 10 + 2.9 \ln(m)$.

Note: $\log_2(x) = \log_2(e) \ln(x) = 1.44 \ln(x)$.

Checksums!

Consider a set of $m$ files.

Each file has a checksum of $b$ bits.

How large should $b$ be for $Pr[\text{share a checksum}] \leq 10^{-3}$?

Claim: $b \geq 2.9 \ln(m)$.

Proof:

Let $n = 2^b$ be the number of checksums.

We know $Pr[\text{no collision}] = \exp \left( -\frac{m^2}{2n} \right) = 1 - \frac{m^2}{2n}$.

Hence,

$Pr[\text{no collision}] = 1 - 10^{-3} \Rightarrow \frac{m^2}{2n} \approx 10^{-3}$

$\Rightarrow 2n \approx m^2 10^3 \Rightarrow 2^{b+1} \approx m^2 2^{10}$

$\Rightarrow b + 1 \approx 10 + 2 \log_2(m) \approx 10 + 2.9 \ln(m)$.

Note: $\log_2(x) = \log_2(e) \ln(x) = 1.44 \ln(x)$.

Coupon Collector Problem.

There are $n$ different baseball cards.

(Brian Wilson, Jackie Robinson, Roger Hornsby, ...)

One random baseball card in each cereal box.

Theorem: If you buy $m$ boxes,

(a) $Pr[\text{miss one specific item}] \approx e^{-\frac{m}{n}}$

(b) $Pr[\text{miss any one of the items}] \leq ne^{-\frac{m}{n}}$.
Collect all cards?

Experiment: Choose $m$ cards at random with replacement.
Events: $E_i = \text{fail to get player } k$, for $k = 1, \ldots, n$
Probability of failing to get at least one of these $n$ players:
$$p := \Pr[E_1 \cup E_2 \cup \cdots \cup E_n]$$
How does one estimate $p$? Union Bound:
$$p = \Pr[E_1 \cup E_2 \cup \cdots \cup E_n] \leq \Pr[E_1] + \Pr[E_2] + \cdots + \Pr[E_n].$$
Plug in and get
$$\Pr[X_i] = e^{-\frac{q}{n}}, k = 1, \ldots, n.$$ 

Review: Harmonic sum

$$H(n) = 1 + \frac{1}{2} + \cdots + \frac{1}{n} = \int_1^n \frac{1}{x} \, dx = \ln(n).$$

A good approximation is
$$H(n) \approx \ln(n) + \gamma$$

where $\gamma \approx 0.58$ (Euler-Mascheroni constant).

Collect all cards?

Thus,
$$\Pr[\text{missing at least one card}] \leq ne^{-\frac{q}{n}}.$$ 
Hence,
$$\Pr[\text{missing at least one card}] \leq p$$
when $m \geq n\ln\left(\frac{n}{p}\right)$.

To get $p = 1/2$, set $m = n\ln(2n)$.
($p \leq ne^{-\frac{q}{n}} \leq ne^{-n\ln(p)} \leq n\left(\frac{q}{n}\right) \leq p$)
E.g., $n = 10^2 \Rightarrow m = 530$; $n = 10^3 \Rightarrow m = 7600$.

Collect all cards?

Time to collect coupons

$X$-time to get $n$ coupons.
$X_1$ - time to get first coupon. Note: $X_1 = 1$. $E(X_1) = 1$.
$X_2$ - time to get second coupon after getting first.
$Pr[\text{get second coupon} | \text{got milk first coupon}] = \frac{2}{n}$.
$E[X_2]$: Geometric!!! $\implies E[X_2] = \frac{1}{\frac{1}{2}} = \frac{n}{2}$.
$Pr[\text{getting ith coupon} | \text{got } i-1 \text{ first coupons}] = \frac{n-1}{n-1} = \frac{n-1}{n}$
$E[X_i] = \frac{1}{\frac{n}{2}}, i = 1, 2, \ldots, n$.

$$E[X] = E[X_1] + \cdots + E[X_n] = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} = n\left(\frac{1}{2} + \cdots + \frac{1}{n}\right) = nH(n) \approx n(\ln n + \gamma)$$

Balls in bins.

For each of $n$ balls, choose random bin: $X_i$ balls in bin $i$.
For simplicity: $n$ balls in $n$ bins.
Round robin: load 1 !
Centralized! Not so good.
Uniformly at random? Average load 1.
Max load?
$n$. Uh Oh!
Max load with probability $\geq 1 - \delta$?
$\delta = \frac{1}{3}$ for today. $c$ is 1 or 2.
Solving for $k$

\[ \Pr[X \geq k] \leq \frac{1}{n^2} \]

What is upper bound on max-load $k$?

**Lemma:** Max load is $\Theta(\log n)$ with probability $\geq 1 - \frac{1}{n}$.

(Recall $k! \geq \left( \frac{1}{2} \right)^k$)

\[ \implies \frac{1}{k!} \leq \left( \frac{1}{2} \right)^k \leq \left( \frac{1}{2\log n} \right)^k \]

If $\log n \geq 1$, then $k = 2\log n$ suffices.

Also: $k = \Theta(\log n / \log \log n)$ suffices as well.

$k^k \to n^k$.

Actually Max load is $\Theta(\log n / \log \log n)$ w.h.p. (W.h.p. - means with probability at least $1 - O(1/n^2)$ for today.)

Better than variance based methods...

**Expected Value of Integer RV**

**Theorem:** For a r.v. $X$ that takes values in $\{0, 1, 2, \ldots\}$, one has

\[ E[X] = \sum_{i=1}^{\infty} i \Pr[X \geq i]. \]

**Proof:** One has

\[ E[X] = \sum_{i=1}^{\infty} i \times \Pr[X = i] = \sum_{i=1}^{\infty} \sum_{j=i}^{\infty} j \times \Pr[X = j] \]

\[ = \sum_{i=1}^{\infty} \left( i \times \Pr[X \geq i] - (i-1) \times \Pr[X \geq i] \right) \]

\[ = \sum_{i=1}^{\infty} (i \times \Pr[X \geq i] - (i-1) \times \Pr[X \geq i]) \]

\[ = \sum_{i=1}^{\infty} \Pr[X \geq i]. \]

**Geometric Distribution: Memoryless**

Let $X$ be $G(p)$. Then, for $n \geq 0$,

\[ \Pr[X > n] = \Pr[\text{first } n \text{ flips are } T] = (1-p)^n. \]

**Theorem**

\[ \Pr[X > n + m|X > n] = \Pr[X > m].m.n \geq 0. \]

**Proof:**

\[ \Pr[X > n + m|X > n] = \frac{\Pr[X > n + m \text{ and } X > n]}{\Pr[X > n]} \]

\[ = \frac{\Pr[X > n + m]}{\Pr[X > n]} \]

\[ = \frac{(1-p)^{n+m}}{(1-p)^n} \]

\[ = (1-p)^m \]

\[ = \Pr[X > m]. \]

**Geometric Distribution: Memoryless - Interpretation**

\[ \Pr[X > n + m|X > n] = \Pr[A|B] = \Pr[A] = \Pr[X > m]. \]

The coin is memoryless, therefore, so is $X$.

**Sum of Poisson Random Variables.**

For $X = P(\lambda)$ and $Y = P(\mu)$, what is $X + Y$?

Poisson? Yes.

What parameter? $\lambda + \mu$.

Why?

$P(\lambda)$ is limit $n \to \infty$ of $B(n, \lambda/n)$.

Recall Derivation:

break interval into $n$ intervals and each has arrival with probability $\lambda/n$.

Now:

arrival for $X$ happens with probability $\lambda/n$
arrival for $Y$ happens with probability $\mu/n$

So, we get limit $n \to \infty$ is $B(n, (\lambda + \mu)/n)$.

Details: both could arrive with probability $\lambda \mu / n^2$.

But this goes to zero as $n \to \infty$.

(Like $\lambda^2 / n^2$ in previous derivation)
Linear Regression: Preamble

The "best" guess about Y, if we know only the distribution of Y, is $E[Y]$. If "best" is Mean Squared Error, more precisely, the value of $a$ that minimizes $E[(Y-a)^2]$ is $a = E[Y]$.

Thus, if we want to guess the value of Y, we choose $E[Y]$.

Now assume we make some observation $X$ related to Y.

How do we use that observation to improve our guess about Y? The idea is to use a function $g(X)$ of the observation to estimate Y.

The simplest function $g(X)$ is a constant that does not depend of X. The next simplest function is linear: $g(X) = a + bX$.

What is the best linear function? That is our next topic.

A bit later, we will consider a general function $g(X)$.

LLSE

Theorem

Consider two RVs $X, Y$ with a given distribution $Pr[X = x, Y = y]$.

Proof 1:

$E[Y | X] = \hat{Y} = E[Y] + \frac{cov(X,Y)}{var(X)} (X - E[X])$.

$Y - \hat{Y} = (Y - E[Y]) - \frac{cov(X,Y)}{var(X)} (X - E[X])$, $E[Y - \hat{Y}] = 0$ by linearity.

Also, $E[(Y - \hat{Y})X] = 0$, after a bit of algebra. (next slide)

Combine Brown inequalities: $E[(Y - \hat{Y})(c + dX)] = 0$ for any $c, d$.

Since $\hat{Y} = a + bX$ for some $a, b$, so $\hat{Y} - a = bX - c + dX$.

Then, $E[(Y - \hat{Y})(a - bX)] = 0$, $a, b$. Now,

$E[(Y - a - bX)^2] = E[(Y - \hat{Y} - a - bX)^2]$

$= E[(Y - \hat{Y})^2] + E[(\hat{Y} - a - bX)^2] + 0 \geq E[(Y - \hat{Y})^2]$

This shows that $E[(Y - \hat{Y})^2] \leq E[(Y - a - bX)^2]$, for all $(a, b)$.

Thus $\hat{Y}$ is the LLSE.

A Bit of Algebra

$Y - \hat{Y} = (Y - E[Y]) - \frac{cov(X,Y)}{var(X)} (X - E[X])$.

Hence, $E[Y - \hat{Y}] = 0$. We want to show that $E[(Y - \hat{Y})X] = 0$.

Note that $E[(Y - \hat{Y})X] = E(Y - \hat{Y})(X - E[X])$.

$E[(Y - \hat{Y})(X - E[X])] = E[(X - E[X])(Y - E[Y]) - \frac{cov(X,Y)}{var(X)} (X - E[X]) (X - E[X])]$

$= \frac{cov(X,Y)}{var(X)} = 0$. (Recall that $cov(X,Y) = E[(X - E[X])(Y - E[Y])])$ and $var[X] = E[(X - E[X])^2]$. Thus $\hat{Y}$ is the LLSE.
Discrete Probability.

Probability Space: \( \Omega, \Pr : \Omega \to [0,1], \sum_{w \in \Omega} \Pr(w) = 1. \)
Events: \( A \subseteq \Omega. \)
Simple Total Probability: \( \Pr[B] = \Pr[A \cap B] + \Pr[A^c \cap B]. \)
Conditional Probability: \( \Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}. \)
Simple Product Rule: \( \Pr[A \cap B] = \Pr[A|B] \Pr[B]. \)
Bayes Rule: \( \Pr[A|B] = \frac{\Pr[B|A] \Pr[B]}{\Pr[B]}. \)

Inference:
Have one of two coins. Flip coin, which coin do you have?
Got positive test result. What is probability you have disease?

Random Variables

Random Variables: \( X : \Omega \to \mathbb{R}. \)
Distribution: \( \Pr[X = a] = \sum_{\omega : X(\omega) = a} \Pr(\omega) \)
\( X \) and \( Y \) independent \( \iff \) all associated events are independent.
Expectation: \( E[X] = \sum_a \Pr[X = a] = \sum_{\omega \in \Omega} X(\omega) \Pr(\omega). \)
Linearity: \( E[X + Y] = E[X] + E[Y]. \)
Variance: \( \text{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2 \)
For independent \( X, Y, \) \( \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y). \)
Also: \( \text{Var}(cX) = c^2 \text{Var}(X) \) and \( \text{Var}(X + b) = \text{Var}(X). \)

Poisson: \( X \sim P(\lambda), \Pr[X = i] = e^{-\lambda} \frac{\lambda^i}{i!}. \)
\( E(X) = \lambda, \ Var(X) = \lambda. \)

Binomial: \( X \sim B(n,p), \Pr[X = i] = \binom{n}{i} p^i (1-p)^{n-i}. \)
\( E(X) = np, \ Var(X) = np(1-p). \)

Uniform: \( X \sim U\{1, \ldots, n\}, \forall i \in [1,n], \Pr[X = i] = \frac{1}{n}. \)
\( E[X] = \frac{n+1}{2}, \ Var(X) = \frac{n^2-1}{12}. \)

Geometric: \( X \sim G(p), \Pr[X = i] = (1-p)^{i-1} p. \)
\( E(X) = \frac{1}{p}, \ Var(X) = \frac{1-p}{p^2}. \)

Note: Probability Mass Function \( \equiv \) Distribution.