Outline

Balls in Bins.

- Birthday.
- Coupon Collector.
- Load balancing.

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

Tail Sum for Expectation.

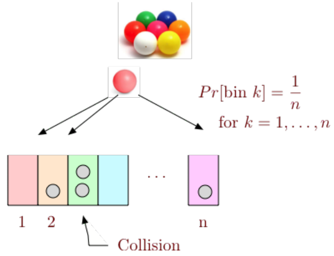
Regression (optional.)

Balls in bins

One throws m balls into $n > m$ bins.

Balls in bins

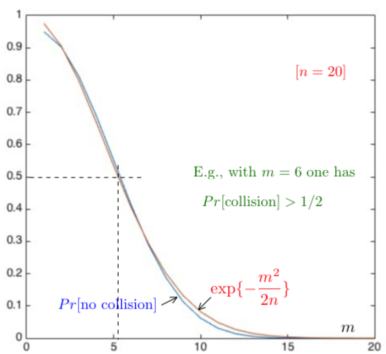
One throws m balls into $n > m$ bins.



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

Balls in bins

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



Balls in bins

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

In particular, $Pr[\text{no collision}] \approx 1/2$ for $m^2/(2n) \approx \ln(2)$, i.e.,

$$m \approx \sqrt{2\ln(2)n} \approx 1.2\sqrt{n}.$$

E.g., $1.2\sqrt{20} \approx 5.4$.

Roughly, $Pr[\text{collision}] \approx 1/2$ for $m = \sqrt{n}$. ($e^{-0.5} \approx 0.6$.)

The Calculation.

A_i = no collision when ith ball is placed in a bin.

$Pr[A_i | A_{i-1} \cap \cdots \cap A_1] = (1 - \frac{i-1}{n})$.

no collision = $A_1 \cap \cdots \cap A_m$.

Product rule:
$Pr[A_1 \cap \cdots \cap A_m] = Pr[A_1]Pr[A_2|A_1] \cdots Pr[A_m|A_1 \cap \cdots \cap A_{m-1}]$

$$\Rightarrow Pr[\text{no collision}] = \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{m-1}{n}\right).$$

Hence,

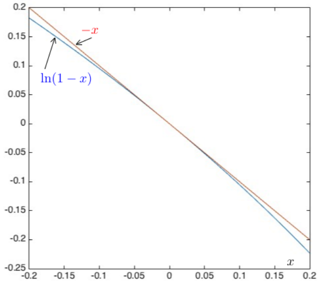
$$\ln(Pr[\text{no collision}]) = \sum_{k=1}^{m-1} \ln(1 - \frac{k}{n}) \approx \sum_{k=1}^{m-1} (-\frac{k}{n})^{(*)}$$

$$= -\frac{1}{n}\frac{m(m-1)}{2}^{(\dagger)} \approx -\frac{m^2}{2n}$$

$^{(*)}$ We used $\ln(1 - \varepsilon) \approx -\varepsilon$ for $|\varepsilon| \ll 1$.

$^{(\dagger)}$ $1 + 2 + \cdots + m - 1 = (m-1)m/2$.

Approximation



$$\exp\{-x\} = 1 - x + \frac{1}{2!}x^2 + \cdots \approx 1 - x, \text{ for } |x| \ll 1.$$

Hence, $-x \approx \ln(1-x)$ for $|x| \ll 1$.

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

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

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

Using linearity of expectation.

Experiment: m balls into n bins uniformly at random.

Random Variable:
X = Number of collisions between pairs of balls.
or number of pairs i and j where ball i and ball j are in same bin.
$X_{ij} = 1\{\text{balls } i,j \text{ in same bin}\}$
$X = \sum_{ij} X_{ij}$

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

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

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

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

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

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

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: $b \geq 2.9\ln(m) + 9$.

Proof:

Let $n = 2^b$ be the number of checksums.
We know $Pr[\text{no collision}] \approx \exp\{-m^2/(2n)\} \approx 1 - m^2/(2n)$. Hence,

$$\begin{aligned} &Pr[\text{no collision}] \approx 1 - 10^{-3} \Leftrightarrow m^2/(2n) \approx 10^{-3} \\ &\Leftrightarrow 2n \approx m^2 10^3 \Leftrightarrow 2^{b+1} \approx m^2 2^{10} \\ &\Leftrightarrow b + 1 \approx 10 + 2\log_2(m) \approx 10 + 2.9\ln(m). \end{aligned}$$

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

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) $Pr[\text{miss any one of the items}] \leq ne^{-\frac{m}{n}}$.

Coupon Collector Problem: Analysis.

Event A_m = 'fail to get Brian Wilson in m cereal boxes'

Fail the first time: $(1 - \frac{1}{n})$
Fail the second time: $(1 - \frac{1}{n})$
And so on ... for m times. Hence,

$$\begin{aligned} Pr[A_m] &= (1 - \frac{1}{n}) \times \cdots \times (1 - \frac{1}{n}) \\ &= (1 - \frac{1}{n})^m \\ ln(Pr[A_m]) &= m\ln(1 - \frac{1}{n}) \approx m \times (-\frac{1}{n}) \\ Pr[A_m] &\approx \exp\{-\frac{m}{n}\}. \end{aligned}$$

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

Collect all cards?

Experiment: Choose m cards at random with replacement.

Events: E_k = 'fail to get player k' , for k = 1, . . . , n

Probability of failing to get at least one of these n players:

$$p := Pr[E_1 \cup E_2 \cdots \cup E_n]$$

How does one estimate p? Union Bound:

$$p = Pr[E_1 \cup E_2 \cdots \cup E_n] \leq Pr[E_1] + Pr[E_2] \cdots Pr[E_n].$$

$$Pr[E_k] \approx e^{-\frac{m}{n}}, k = 1, \ldots, n.$$

Plug in and get

$$p \leq ne^{-\frac{m}{n}}.$$

Collect all cards?

Thus,

$$Pr[\text{missing at least one card}] \leq ne^{-\frac{m}{n}}.$$

Hence,

$$Pr[\text{missing at least one card}] \leq p \text{ when } m \geq n\ln(\frac{n}{p}).$$

To get $p = 1/2$, set $m = n\ln(2n)$.

$(p \leq ne^{-\frac{m}{n}} \leq ne^{-\ln(n/p)} \leq n(\frac{p}{n}) \leq p.)$

E.g., $n = 10^2 \Rightarrow m = 530; n = 10^3 \Rightarrow m = 7600$.

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.

$Pr[\text{“get second coupon”}|\text{“got } \cancel{\text{milk}} \text{ first coupon”}] = \frac{n-1}{n}$

$E[X_2]$? Geometric ! ! ! $\Longrightarrow E[X_2] = \frac{1}{p} = \frac{1}{\frac{n-1}{n}} = \frac{n}{n-1}$.

$Pr[\text{“getting } i\text{th coupon}|\text{“got } i-1\text{rst coupons”}] = \frac{n-(i-1)}{n} = \frac{n-i+1}{n}$

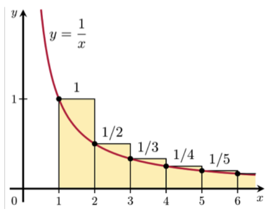
$E[X_i] = \frac{1}{p} = \frac{n}{n-i+1}, i = 1, 2, \ldots, n.$

$$\begin{aligned} E[X] &= E[X_1] + \cdots + E[X_n] = \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} \\ &= n(1 + \frac{1}{2} + \cdots + \frac{1}{n}) =: nH(n) \approx n(\ln n + \gamma) \end{aligned}$$

Review: Harmonic sum

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

.



A good approximation is

$$H(n) \approx \ln(n) + \gamma \text{ where } \gamma \approx 0.58 \text{ (Euler-Mascheroni constant).}$$

Simplest..

Load balance: m balls in n bins.

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 $\geq 1 - \delta$?

$\delta = \frac{1}{n^c}$ for today. c is 1 or 2.

Balls in bins.

For each of n balls, choose random bin: X_i balls in bin i.

$Pr[X_i \geq k] \leq \sum_{S \subseteq [n], |S| = k} Pr[\text{balls in } S \text{ chooses bin } i]$

From Union Bound: $Pr[\cup_i A_i] \leq \sum_i Pr[A_i]$

$Pr[\text{balls in } S \text{ chooses bin } i] = (\frac{1}{n})^k$ and $\binom{n}{k}$ subsets S.

$$\begin{aligned} \Pr[X_i \geq k] &\leq \binom{n}{k}\left(\frac{1}{n}\right)^k \\ &\leq \frac{n^k}{k!}\left(\frac{1}{n}\right)^k = \frac{1}{k!} \end{aligned}$$

Choose k, so that $Pr[X_i \geq k] \leq \frac{1}{n^2}$.

$Pr[\text{any } X_i \geq k] \leq n \times \frac{1}{n^2} = \frac{1}{n} \rightarrow$ max load $\leq k$ w.p. $\geq 1 - \frac{1}{n}$

Solving for k

$Pr[X_i \geq k] \leq \frac{1}{k!} \leq 1/n^2$?

What is upper bound on max-load k?

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

$k! \geq n^2$ for $k = 2e\log n$

(Recall $k! \geq (\frac{k}{e})^k$.)

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

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

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

$k^k \to n^c$.

Actually Max load is $\Theta(\log n / \log\log n)$ w.h.p.

(W.h.p. - means with probability at least $1 - O(1/n^c)$ for today.)

Better than variance based methods...

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

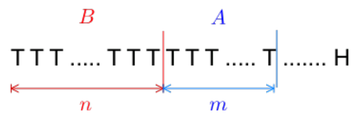
$$Pr[X > n+m | X > n] = Pr[X > m], m, n \geq 0.$$

Proof:

$$\begin{aligned} Pr[X > n+m | X > n] &= \frac{Pr[X > n+m \text{ and } X > n]}{Pr[X > n]} \\ &= \frac{Pr[X > n+m]}{Pr[X > n]} \\ &= \frac{(1-p)^{n+m}}{(1-p)^n} = (1-p)^m \\ &= Pr[X > m]. \end{aligned}$$

Geometric Distribution: Memoryless - Interpretation

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



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

The coin is memoryless, therefore, so is X.

Expected Value of Integer RV

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

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

Proof: One has

$$\begin{aligned} E[X] &= \sum_{i=1}^{\infty} i \times Pr[X = i] \\ &= \sum_{i=1}^{\infty} i\{Pr[X \geq i] - Pr[X \geq i+1]\} \\ &= \sum_{i=1}^{\infty} \{i \times Pr[X \geq i] - i \times Pr[X \geq i+1]\} \\ &= \sum_{i=1}^{\infty} \{i \times Pr[X \geq i] - (i-1) \times Pr[X \geq i]\} \\ &= \sum_{i=1}^{\infty} Pr[X \geq i]. \end{aligned}$$

$\square$

Geometric Distribution: Yet another look

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

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

If $X = G(p)$, then $Pr[X \geq i] = Pr[X > i-1] = (1-p)^{i-1}$.

Hence,

$$E[X] = \sum_{i=1}^{\infty} (1-p)^{i-1} = \sum_{i=0}^{\infty} (1-p)^i = \frac{1}{1-(1-p)} = \frac{1}{p}.$$

Sum of Poisson Random Variables.

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

Poission? 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 λ/n.

Now:
arrival for X happens with probability λ/n
arrival for Y happens with probability μ/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 λ^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 $a = E[Y]$.

Proof:

Let $\hat{Y} := Y - E[Y]$.
Then, $E[\hat{Y}] = E[Y - E[Y]] = E[Y] - E[Y] = 0$.
So, $E[\hat{Y}c] = 0, \forall c$. Now,

$$\begin{aligned} E[(Y-a)^2] &= E[(Y - E[Y] + E[Y] - a)^2] \\ &= E[(\hat{Y} + c)^2] \text{ with } c = E[Y] - a \\ &= E[\hat{Y}^2 + 2\hat{Y}c + c^2] = E[\hat{Y}^2] + 2E[\hat{Y}c] + c^2 \\ &= E[\hat{Y}^2] + 0 + c^2 \geq E[\hat{Y}^2]. \end{aligned}$$

Hence, $E[(Y-a)^2] \geq E[(Y - E[Y])^2], \forall a$. $\square$

Linear Regression: Preamble

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: $g(X) = a + bX$.

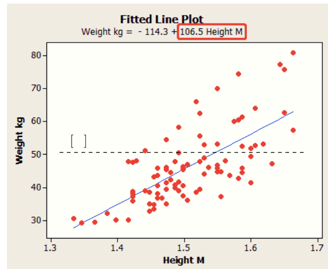
What is the best linear function? That is our next topic.

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

Linear Regression: Motivation

Example 1: 100 people.

Let (X_n, Y_n) = (height, weight) of person n, for $n = 1, \ldots, 100$:



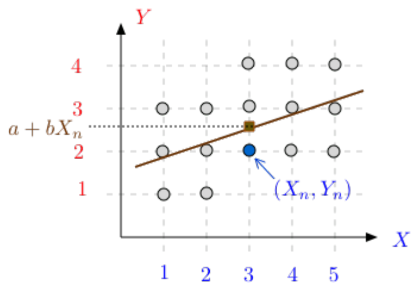
The blue line is $Y = -114.3 + 106.5X$. (X in meters, Y in kg.)

Best linear fit: Linear Regression.

Motivation

Example 2: 15 people.

We look at two attributes: (X_n, Y_n) of person n, for $n = 1, \ldots, 15$:



The line $Y = a + bX$ is the linear regression.

LLSE

$LLSE[Y|X]$ - best guess for Y given X.
Theorem
Consider two RVs X, Y with a given distribution $Pr[X = x, Y = y]$. Then,

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

Proof 1:
$Y - \hat{Y} = (Y - E[Y]) - \frac{cov(X,Y)}{var[X]}(X - E[X])$. $E[Y - \hat{Y}] = 0$ 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: $\hat{Y} = \alpha + \beta X$ for some α, β, so $\exists c, d$ s.t. $\hat{Y} - a - bX = c + dX$.
Then, $E[(Y - \hat{Y})(\hat{Y} - a - bX)] = 0, \forall a, b$. Now,

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

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

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

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

because $E[(Y - \hat{Y})E[X]] = 0$.

Now,

$$\begin{aligned} &E[(Y - \hat{Y})(X - E[X])] \\ &= E[(Y - E[Y])(X - E[X])] - \frac{cov(X,Y)}{var[X]} E[(X - E[X])(X - E[X])] \\ &=^{(*)} cov(X,Y) - \frac{cov(X,Y)}{var[X]} var[X] = 0. \quad \square \end{aligned}$$

$^{(*)}$ Recall that $cov(X,Y) = E[(X - E[X])(Y - E[Y])]$ and $var[X] = E[(X - E[X])^2]$.

Discrete Probability.

Probability Space: Ω, $Pr : \Omega \to [0,1]$, $\sum_{\omega \in \Omega} Pr(w) = 1$.

Events: $A \subset \Omega$.

Simple Total Probability: $Pr[B] = Pr[A \cap B] + Pr[\overline{A} \cap B]$.

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: $X : \Omega \to R$.

Distribution: $Pr[X = a] = \sum_{\omega : X(\omega) = a} Pr(\omega)$

X and Y independent $\iff$ all associated events are independent.
Expectation: $E[X] = \sum_a aPr[X = a] = \sum_{\omega \in \Omega} X(\omega)Pr(\omega)$.
Linearity: $E[X + Y] = E[X] + E[Y]$.

Variance: $Var(X) = E[(X - E[X])^2] = E[X^2] - (E(X))^2$
For independent X, Y, $Var(X + Y) = Var(X) + Var(Y)$.
Also: $Var(cX) = c^2 Var(X)$ and $Var(X + b) = Var(X)$.

Poisson: $X \sim P(\lambda)$ $\quad Pr[X = i] = e^{-\lambda} \frac{\lambda^i}{i!}$.
$E(X) = \lambda$, $Var(X) = \lambda$.
Binomial: $X \sim B(n, p)$ $\quad Pr[X = i] = \binom{n}{i} p^i (1 - p)^{n-i}$
$E(X) = np$, $Var(X) = np(1 - p)$
Uniform: $X \sim U\{1, \ldots, n\}$ $\quad \forall i \in [1, n], Pr[X = i] = \frac{1}{n}$.
$E[X] = \frac{n+1}{2}$, $Var(X) = \frac{n^2-1}{12}$.
Geometric: $X \sim G(p)$ $\quad Pr[X = i] = (1 - p)^{i-1} p$
$E(X) = \frac{1}{p}$, $Var(X) = \frac{1-p}{p^2}$

Note: Probability Mass Function $\equiv$ Distribution.