RAD (2-3) Moments and Deviation

Occupancy problems

vi har m bolde vi assigner til n bins, boldene smides uafhængigt og fordeler sig i de n spande uniformt.

hvad er det maximale antal bolde i en spand

hvad er det foventet antal spande med k bolde

Proof

lad $m = n \geq 3$, og for $i = 1, \ldots, n$ lad X_i antallet af bolde i den i'te spand.

We study the following occupancy problem: find k så at, med høj sandsynlighed (sandsynlighed mindst $1 - 1/n$), ingen spand har mere end k bolde.

Definer et event:

Lad $\mathcal{E}_j(k)$ være eventet hvor spand j har mindst k bold (Med andre ord $X_j \geq k$). Vi kigger først på $\mathcal{E}_1(k)$, eventet hvor spand 1 har mindst k bolde. De andre spande er symmetriske

For et $i \in \{0, 1, \cdots, n\}$ og $k \geq 3$

sandsynligheden for at der er i bolde i spand 1:

Det her er binomial fordelingen. Det er fordi vi har gentagne bernoulli forsøg, enten er bolden i spanden eller ej. Her er $p = 1/n$ fordi at sandsynligheden for at en bold rammer i spand 1 er 1/n

Binomial koefficent: antallet af måder du kan vælge i bolde ud af de af n bolde.

$\left(\frac{1}{n}\right)^i$ er sandsynligheden for at vores i udvalgte bolde rammer i spanden

$\left(1-\frac{1}{n}\right)^{n-i}$ sandsynligheden for at de resterende bolde ryger ned i andre spande, så spand 2 til n.

$$\Pr\left[X_1 = i\right] = \binom{n}{i}\left(\frac{1}{n}\right)^i\left(1-\frac{1}{n}\right)^{n-i}$$

Vi smider den sidste faktor væk og får en øvre grænse da vi siger den er højst 1, så vi sætter den til 1.

$$\leq \binom{n}{i}\left(\frac{1}{n}\right)^i$$

propasition B. 2. 3. N'er går ud med hianden og vi får $(e/i)^i$

$$\leq \left(\frac{ne}{i}\right)^i\left(\frac{1}{n}\right)^i = \left(\frac{e}{i}\right)^i \Rightarrow$$

Sandsynligheden for at spand 1 får mindst k bolde, sandsynligheden for at den modtager k + k+1 + k+2 op til k+n bolde

$$\Pr\left[\mathcal{E}_1(k)\right] = Pr[X_1 \geq k] = \sum_{i=k}^{n} Pr[X_1 = i]$$

eventet hvor spand 1 har mindst k bolde

Øvre grænse på at spand 1 modtager mindst i bolde er $(e/i)^i$. så hvis vi summer fra k til n får vi en øvre grænse på at spand 1 får mindst k bolde.

$$\Pr\left[\mathcal{E}_1(k)\right] \leq \sum_{i=k}^{n}\left(\frac{e}{i}\right)^i$$

Nævner må altid være mindst k helt op til n, derfor når vi istedet bare vælger den til at være k får vi den mindste mulige nævner kan have. da alle led nu har en mindre værdi, resulterer det i at brøkken giver et større tal og derfor bliver summen større.

$$\leq \sum_{i=k}^{n} \left(\frac{e}{k}\right)^i$$

mange flere led, derfor er sum strengt større

$$< \sum_{i=k}^{\infty} \left(\frac{e}{k}\right)^i$$

Prøv at sæt k = 3 i begge udtryk, dertil bliver det klart at de er ens.

$$= \sum_{i=0}^{\infty} \left(\frac{e}{k}\right)^{k+i}$$

$(\frac{e}{k})^k$ er en konstant og kan trækkes ud.

$$= \left(\frac{e}{k}\right)^k \sum_{i=0}^{\infty} \left(\frac{e}{k}\right)^i$$

Summen er en geometrisk sum, og har udtrykket:

$$= \left(\frac{e}{k}\right)^k \left(\frac{1}{1-\frac{e}{k}}\right)$$

a=(e/k)

Lad $k^\star = \lceil 3 \frac{\ln n}{\ln \ln n} \rceil$ (viser ikke denne ulighed)

$$\left(\frac{e}{k^\star + 1}\right)^{k^\star+1} \left(\frac{1}{1-\frac{e}{k^\star+1}}\right) \leq n^{-2}$$

Markov's Inequality

Lad Y være en stokastisk variabel som kun tager positive værdier. Så for alle $t > 0$:

$$Pr[Y \geq t] \leq \frac{E[Y]}{t}$$

På samme måde, for k > 0 og hvis $E[Y] > 0$:

$$Pr[Y \geq kE[Y]] \leq \frac{1}{k}$$

Y kunne være køretid, så sandsynligheden for at den er langsom den er højst forventingen divideret med t.

proof

$$\mathbb{E}[Y] = \sum_{y} y \Pr[Y = y] \geq$$

sum af subset af y, subset er mindre, y er positiv og t > 0, sum er derfor mindre nu. værdi vi ikke inkluderer kan ikke være negative.

$$\sum_{y \geq t} y \Pr[Y = y] \geq$$

erstat y med t, vi ved at alle y værdier er større end t pga grænse i sum.

$$\sum_{y \geq t} t \Pr[Y = y] =$$

t er en konstant i den sum så tag den ud.

$$t \sum_{y \geq t} \Pr[Y = y] = t \Pr[Y \geq t]$$

sum er sandsynlighed for at y >= t.

Divider med t på begge sider:

$$\mathbb{E}[Y] \geq \sum_{y \geq t} t \Pr[Y = y] = t \Pr[Y \geq t]$$

derved, $\Pr[Y \geq t] \leq \frac{\mathbb{E}[Y]}{t}$.

Anden ulighed

sæt $t = k\mathbb{E}[Y] > 0$ i markov:

$$\Pr[Y \geq k\mathbb{E}[Y]] = \Pr[Y \geq t] \leq \frac{\mathbb{E}[Y]}{t} = \frac{\mathbb{E}[Y]}{k\mathbb{E}[Y]} = \frac{1}{k}$$

Chebyshev's Inequality

Given a random variable X with expectation $\mathbb{E}[X] = \mu_X$, define its variance $(\mathrm{Var}[X]$ or $\sigma_X^2)$ as $\sigma_X^2 := \mathbb{E}\left[(X - \mu_X)^2\right]$, and its standard deviation as $\sigma_X := \sqrt{\mathbb{E}\left[(X - \mu_X)^2\right]}$.

Theorem

Let X be a random variable with expectation μ_X and standard deviation $\sigma_X > 0$. Then for all $t > 0$:

$$\Pr\left[|X - \mu_X| \geq t\sigma_X\right] \leq \frac{1}{t^2}$$

Sandsynligheden for X's afvigelse (forskellen) fra dens expectation er mindst t standard afvigelser væk fra expectation, er mindre eller lig $\frac{1}{t^2}$.

Proof

Here we use that $k > 0$ and $\mathbb{E}[Y] = \sigma_X^2 > 0$ so that we can use the second version of Markov's inequality.
Let $k = t^2$ and $Y = (X - \mu_X)^2$.
Then $\sigma_X^2 = \mathbb{E}[Y]$

Since numerical signs are used we can take the power of 2 through the whole inequality and it's the same. This makes it positive.

$$
\begin{aligned}
\Pr\left[\left|X-\mu_X\right| \geq t\sigma_X\right] &= \Pr\left[\left(X-\mu_X\right)^2 \geq t^2\sigma_X^2\right] \\
&= \Pr[Y \geq k\mathbb{E}[Y]] \\
&\leq \frac{1}{k} \\
&= \frac{1}{t^2}
\end{aligned}
$$

Theorem

Let X be a random variable with expectation μ_X and standard deviation $\sigma_X > 0$. Then for all $t > 0$:

$$
\Pr\left[\left|X-\mu_X\right| \geq t\sigma_X\right] \leq \frac{1}{t^2}
$$

Proof

Letting $Y = \left(X-\mu_X\right)^2$,

$$
\begin{aligned}
\Pr\left[\left|X-\mu_X\right| \geq t\right] &= \Pr\left[\left(X-\mu_X\right)^2 \geq t^2\right] \\
&= \Pr\left[Y \geq t^2\right] \\
&\leq \frac{\mathbb{E}[Y]}{t^2} = \frac{\sigma_X^2}{t^2}
\end{aligned}
$$

where the inequality follows from the first version of Markov's inequality.

Theorem

Let X be a random variable with expectation μ_X and standard deviation $\sigma_X > 0$. Then for all $t > 0$:

$$
\Pr\left[\left|X-\mu_X\right| \geq t\right] \leq \frac{\sigma_X^2}{t^2} = \frac{\operatorname{Var}[X]}{t^2}
$$

Ekstra

Union bound

Union bound

For k events $\mathcal{E}_1, \ldots, \mathcal{E}_k$,

$$\Pr[\cup_{i=1}^{k} \mathcal{E}_i] \leq \sum_{i=1}^{k} \Pr[\mathcal{E}_i]$$

This is called a *union bound*; we will make use of this later in the course.

If the events are *disjoint* (i.e. $\mathcal{E}_i \cap \mathcal{E}_j = \emptyset$ for $i \neq j$),

$$\Pr[\cup_{i=1}^{k} \mathcal{E}_i] = \sum_{i=1}^{k} \Pr[\mathcal{E}_i]$$

Independent variables

A set of random variables $X_1, X_2, \ldots, X_n$ is (mutually) independent if for every subset $I \subseteq [1, n]$ and for any set of real values $\{x_i\}_{i\in 1}$,

$$\Pr\left[\cap_{i\in I} X_i = x_i\right] = \prod_{i\in I} \Pr\left[X_i = x_i\right]$$

This is similar to the definition on the previous examples
for events $\{X_i = x_i\}$ but has to hold for all choices of the values $\{x_i\}_{i\in 1}$

k-independent variables

A set of random variables $X_1, X_2, \ldots, X_n$ is k-independent if for any subset $I \subseteq [1, n]$ with $|I| \leq k$ and for any set of real values $\{x_i\}_{i\in l}$,

$$\Pr\left[\cap_{i\in I} X_i = x_i\right] = \prod_{i\in I} \Pr\left[X_i = x_i\right]$$

If $k = 2$, the random variables are said to be pairwise independent: for every distinct pair of indices (i, j) and any values a, b,

$$\Pr\left[X_i = a \wedge X_j = b\right] = \Pr\left[X_i = a\right] \cdot \Pr\left[X_j = b\right]$$

linearity of variance

Let $X_1, X_2, \ldots, X_m$ be pairwise independent random variables, and $X = \sum_{i=1}^{m} X_i$. We will show that the linearity of variance for independent variables is also true for pairwise independent variables. As follows from the course book the variance of X is given by

$$E[(X-\mu)^2] = E[(\sum_{i=1}^{m} X_i - \mu_i)^2]$$

where $\mu_i = E[X_i]$ and $\mu = \sum_{i=1}^{m} \mu_i$. Using linearity of expectation the expression can be expanded

$$E[(X-\mu)^2] = \sum_{i=1}^{m} E[(X_i-\mu_i)^2] + 2\sum_{i<j} E[(X_i-\mu_i)(X_j-\mu_j)].$$

Since all pairs X_i, X_j are pairwise independent, so are the pairs $(X_i - \mu_i), (X_j - \mu_j)$ and by (3.2), the expectation of the product can be replaced by the product of the expectations. Since $E[(X_i - \mu_i)] = E[X_i] - \mu_i = 0$, the latter summations vanishes and it follows that

$$E[(X-\mu)^2] = \sum_{i=1}^{m} E[(X_i-\mu_i)^2] = \sum_{i=1}^{m} \sigma_{X_i}^2.$$

Two-Point sampling

truly random numbers are hard to obtain. Two-point sampling is a way to take just two random independent values and turn them into many pairwise independent values.

Let p be prime, and let a, b be independent random variables uniformly chosen from

$\mathbb{Z}_p = \{0, \cdots, p-1\}$.
For $i = 0, 1, \cdot, p-1$, let

$$r_i = (a \cdot i + b) \; mod \; p$$

Then for any $i \neq j$ (mod p), r_i and r_j are independent and uniform in $\mathbb{Z}_p$

Thus, $r_0, r_1, \cdots, r_p - 1$ are pairwise independent.

Two-point Sampling, Application

Let $L \subseteq \Sigma^*$ be some language, and let p be a prime number.

A function $A : \Sigma^* \times \mathbb{Z}_p \to \{0, 1\}$ is an **RP** algorithm for L, if it runs in polynomial time for all inputs, and

If $x \in L$, then $A(x, r) = 1$ for at least half of all $r \in \mathbb{Z}_p$.

If $x \notin L$ then $A(x, r) = 0$ for all $r \in \mathbb{Z}_p$.

RP stands for "Randomized Polynomial" (time).

- Note that A takes a pair (x, r) as input. x is the problem instance and r is the random number in $\mathbb{Z}_p$ given to A.
- If $x \notin L$, A gives the correct output (0) for any choice of r. Otherwise, A gives the correct output (1) for at least half the choices of r.
- Put differently, we choose a random $r \in \mathbb{Z}_p$, and if $A(x, r) = 1$ then we know that $x \in L$. But if $A(x, r) = 0$ then either $x \notin L$ or we have chosen a bad r. The probability of such a *false negative* is at most $\frac{1}{2}$.
- For this reason, we call A a Monte Carlo algorithm with *one-sided error* (Section 1.5.2).

Antag at algoritme A bruger $\lg n$ tilfældige bits repræsenteret som et tal $r \in \{0, \ldots, n-1\}$ hvor n er et primtal. I følgende bruger vi notationen $A(x, r)$ for at beskrive outputtet af A på input x, hvor A vælger den tilfældige bitstreng r. Og lad os i fejlsandsynlighederne antage, at vores konkrete $x \in L$ så det korrekte svar er 1.

Algoritme 1 - $t \lg n$ random bits

Vælg t tal $r_0, \ldots, r_{t-1} \in [n]$ uafhængigt og uniformt tilfældigt. Beregn $A(x, r_0), \ldots, A(x, r_{t-1})$. Hvis vi en enkelt gang ser tallet 1 er det bevis på $x \in L$, ellers hvis vi \emph{alle} gange får 0 vælger vi det som output.

Så vil fejlsandsynligheden være $< \frac{1}{2}^t = 1/2^t$.

Problemet ved denne tilgang er, at vi skal vælge $t \lg n$ random bits. Hvis vi f.eks. vælger $t = 2$ skal vi bruge $2 \ln n$ random bits for en fejlsandsynlighed $< 1/4$.

Algoritme 2 - $2 \lg n$ random bits}

Vælg $a, b \in [n]$ uafhængigt og uniformt tilfældigt.

Da vi antager n er et primtal, så ved vi at såfremt vi lades $r_i = (a * i + b) \bmod n$, så vil r_i og r_j hvor $i \neq j$ være uniformt distribueret i $[n]$ og parvist uafhængige (kan blot antages, skal ikke bevises).

Igen beregner vi $A(x, r_0), \ldots, A(x, r_{t-1})$ og vælger 1 såfremt den optræder bare én gang, ellers 0.Algoritme 2 - $2 \lg n$ random bits}

Nu bruger vi kun $2 \lg n$ random bits.

Sandsynlighed for at algoritme 2 fejler

For $i = 0, \ldots, t-1$ lader vi $Y_i = A(x, r_i)$. Lad nu $Y = \sum_{i \in [t]} Y_i$.

Da kan vi beregne den forventede værdi:

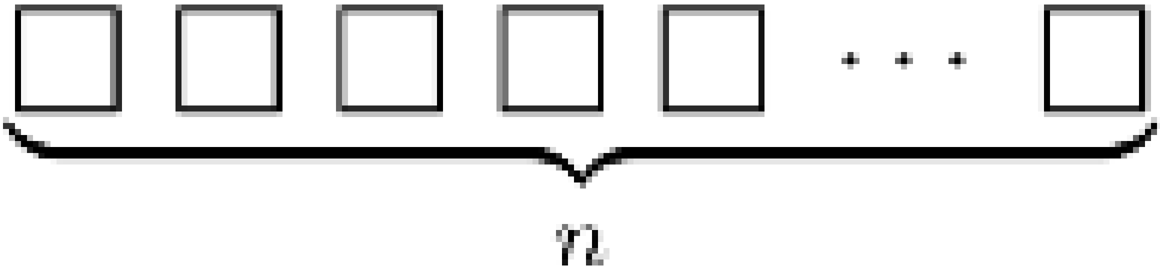
$$\mathbb{E}[Y] = \sum_{i \in [t]} \mathbb{E}[Y_i] = tp \geq \frac{t}{2}$$

Idet vi lader symbolet $p = \mathbb{P}[Y_i = 1] \geq \frac{1}{2}$.

sandsynligheden for den fejler $1/t$

Coupon collector

Betragt følgende eksperiment. Vi har n unikke unikke kupontyper:



I hver runde vælges en kupon-type uafhængigt og uniformt tilfældigt. Vi stopper når alle kupon-typer er valgt. Hvor mange runder vil der være i dette eksperiment?\

For at besvare dette skal vi først definere hvad en epoke er. For $i = 0, \ldots, n-1$ består den i'te epoke af de runder, der starter lige efter den i'te succes og slutter i runden med $(i+1)$'te succes, hvor en succes er defineret som at vælge en kupontype vi ikke har set før. Eksempelvis kunne vi have:

$$\underbrace{C_2,}_{\text{Epoke 0}} \underbrace{C_2, C_1,}_{\text{Epoke 1}} \underbrace{C_2, C_2, C_3,}_{\text{Epoke 2}} \ldots$$

For $i = 0, \ldots, n-1$ lader vi Y_i være længden af epoke i. Lad nu $Y = \sum_{i=0}^{n-1} Y_i$. Vi har, at sandsynligheden i den i'te epoke for at finde en ny kupon er antallet af ufundne kuponer $n-i$ over alle de forskellige kupontyper n:

$$p_i = \frac{n-i}{n}$$

Bruger vi, at dette er geometrisk distribueret får vi:

$$\mathbb{E}[Y_i] = \frac{1}{p_i} = \frac{n}{n-i}$$

Da kan vi beregne:

$$\mu_Y = \sum_{i=0}^{n-1} \mathbb{E}[Y_i] = \sum_{i=0}^{n-1} \frac{n}{n-i} = n \sum_{i=1}^{n} \frac{1}{i} = nH_n = n \ln n + \Theta(n) = O(n \ln n)$$