Sum of random variables

Mean and variance of the sum of random variables

In general

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \mathbb{E}(X_i) \\
Var\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} Var(X_i) + \sum_{i \neq j} Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} Var(X_i) + 2 \sum_{1 \leq i < j \leq n} Cov(X_i, X_j)
\end{aligned}$$

If X_i are independent

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \mathbb{E}(X_i) \\
Var\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} Var(X_i)
\end{aligned}$$

If X_i are iid

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} \mathbb{E}(X_i) = n\mathbb{E}(X_1) \\
Var\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} Var(X_i) = nVar(X_1)
\end{aligned}$$

Mean and variance of the weighted sum of random variables

In general

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n} a_i \mathbb{E}(X_i) \\
Var\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n} a_i^2 Var(X_i) + \sum_{i \neq j} a_i a_j Cov(X_i, X_j) \\
&= \sum_{i=1}^{n} Var(X_i) + 2 \sum_{1 \leq i < j \leq n} a_i a_j Cov(X_i, X_j)
\end{aligned}$$

If X_i are independent

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n} a_i \mathbb{E}(X_i) \\
Var\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n} a_i^2 Var(X_i)
\end{aligned}$$

If X_i are iid

$$\begin{aligned}
\mathbb{E}\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n} a_i \mathbb{E}(X_i) = \left(\sum_{i=1}^{n} a_i\right) \mathbb{E}(X_1) \\
Var\left(\sum_{i=1}^{n} a_i X_i\right) &= \sum_{i=1}^{n} a_i^2 Var(X_i) = \left(\sum_{i=1}^{n} a_i^2\right) Var(X_1)
\end{aligned}$$

Mean and variance of the weighted sum of random variables - Matrix form

$$S = \sum_{i=1}^{n} a_i X_i$$

where

μ_i	mean of X_i
σ^2	variance of X_i
σ_{ij}	covariance between X_i and X_j
ρ_{ij}	correlation between X_i and X_j

Mean

$$\mathbb{E}S = \sum_{i=1}^{n} a_i \mathbb{E}X_i = \sum_{i=1}^{n} a_i \mu_i$$

Variance

$$\begin{aligned}
Var(S) &= Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{n} a_j X_j\right) \\
&= \sum_{i=j} a_i a_j Cov\left(X_i, X_j\right) + \sum_{i \neq j} a_i a_j Cov\left(X_i, X_j\right) \\
&= \sum_{i=1}^{n} a_i^2 Var\left(X_i\right) + 2 \sum_{1 \leq i < j \leq n} a_i a_j Cov\left(X_i, X_j\right) \\
&= \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2 \sum_{1 \leq i < j \leq n} a_i a_j \sigma_{ij} \\
&= \sum_{i=1}^{n} a_i \sigma_i^2 + 2 \sum_{1 \leq i < j \leq n} a_i a_j \rho_{ij} \sigma_i \sigma_j
\end{aligned}$$

Or in matrix form

$$Var(S) = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \underbrace{\begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix}}_{\Sigma \quad \text{Covariance matrix}} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$$

Example - Binomial distribution via Bernoulli distribution

Flip a p-coin n times independently and count the number S of heads. Let A_i be the event that the i-th p-coin lands on head and let 1_{A_i} be its indicator.

$$1_{A_i} \text{ iid } B(p) \quad \Rightarrow \quad S = \sum_{i=1}^{n} 1_{A_i} \sim B(n,p)$$

Mean

$$\mathbb{E}S = \sum_{i=1}^{n} \mathbb{E}1_{A_i} = np$$

Variance

$$Var(S) = \sum_{i=1}^{n} Var(1_{A_i}) = npq$$

Example - Negative binomial distribution via geometric distribution

Flip a p-coin until we have the r-th head and count the number S of flips. Let X_i be the number of flips to have the i-th head after I have the $(i-1)$-th head.

$$X_i \text{ iid } Geo(p) \quad \Rightarrow \quad S = \sum_{i=1}^{r} X_i \sim NB(r,p)$$

Mean

$$\mathbb{E}S = \sum_{i=1}^{r} \mathbb{E}X_i = \frac{r}{p}$$

Variance

$$Var(S) = \sum_{i=1}^{r} Var(X_i) = \frac{rq}{p^2}$$

Example - Roll the dice 1000 times

Roll the dice 1000 times. Each time we gain the face value of the roll if we have odd and loose the face value if even.

$$D_i = \begin{cases} 1 & \text{with probability } \frac{1}{6} \\ -2 & \text{with probability } \frac{1}{6} \\ 3 & \text{with probability } \frac{1}{6} \\ -4 & \text{with probability } \frac{1}{6} \\ 5 & \text{with probability } \frac{1}{6} \\ -6 & \text{with probability } \frac{1}{6} \end{cases}$$

To be fair we get 0.5 as an extra for each game.

$$X_i = D_i + 0.5 \quad \text{iid}$$

Let S be the total P&L after the 1000 games. Then, S can be represented in terms of X_i:

$$S = \sum_{i=1}^{1000} X_i$$

where

$$\begin{aligned} &(1) \quad \mathbb{E}D_i = -0.5, \quad Var(D_i) = 14.9167 \\ &(2) \quad \mathbb{E}X_i = 0, \quad\quad Var(X_i) = 14.9167 \end{aligned}$$

Mean

$$\mathbb{E}S = \sum_{i=1}^{1000} \mathbb{E}X_i = 0$$

Variance

$$Var(S) = \sum_{i=1}^{1000} Var(X_i) = 14916.7$$

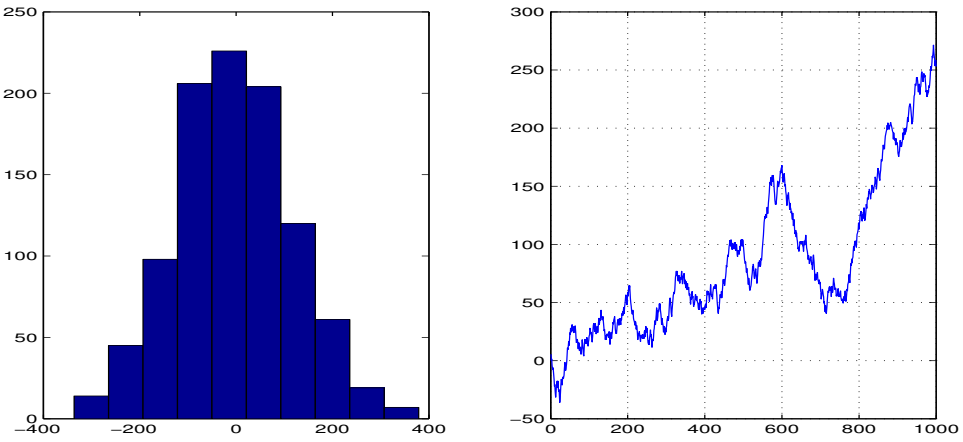

Figure 1: Histogram (left) of the total Profit & Loss after the 1000 games and a sample path (right) of the cumulative P&L of the 1000 games.

```
clear all; close all; clc; rng('default')

NumSimu=1000; % Number of simulation
NumRolling=1000; % Number of rolling for each simulation

% Random samples
x=rand(NumRolling,NumSimu);
dice=zeros(NumRolling,NumSimu);
dice(x<=1/6)=1;
dice(1/6<x&x<=2/6)=2;
dice(2/6<x&x<=3/6)=3;
dice(3/6<x&x<=4/6)=4;
dice(4/6<x&x<=5/6)=5;
dice(5/6<x)=6;

increment=zeros(NumRolling,NumSimu);
increment(dice==1)=1;
increment(dice==2)=-2;
increment(dice==3)=3;
increment(dice==4)=-4;
increment(dice==5)=5;
increment(dice==6)=-6;
increment=increment+0.5; % P&L for each game
Sn=cumsum(increment); % Cumulative P&L

subplot(1,2,1); z=Sn(end,:); hist(z) % Histogram of total P&L
subplot(1,2,2); w=Sn(:,1); plot(1:NumRolling,w); grid on % A sample path
```

Example - Coupon collector problem

To collect all the n toys offered by Mc Donald I start eating the happy meal hamburger. Whenever I order the happy meal, I get a toy randomly among the n different type of toys. Let T_n be the minimum number of the happy meal hamburgers that I have to order to collect all the n different type of toys. Let τ_i be the minimum number of the happy meals that I have to eat to collect the i-th new toy after I get the $(i-1)$-th new toy. Then, T_n can be represented in terms of τ_i:

$$T_n = \sum_{i=1}^{n} \tau_i$$

where

$$\begin{aligned} &(1) \quad \tau_i \sim Geo(\frac{N-(i-1)}{N}) \\ &(2) \quad \tau_i \text{ independent} \end{aligned}$$

Mean

$$\mathbb{E}T_n = \sum_{i=1}^{n} \mathbb{E}\tau_i = n\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right) \sim n \log n$$

Variance

$$Var(T_n) = \sum_{i=1}^{n} Var(\tau_i) = O(n^2)$$

$$Var(T_n) = \sum_{k=1}^{n} \frac{1-(k/n)}{(k/n)^2} = \left(\sum_{k=1}^{n} \frac{1}{k^2}\right) n^2 - \left(\sum_{k=1}^{n} \frac{1}{k}\right) n \approx \frac{\pi^2}{6} n^2 - n \log n$$

Unbiased estimator

Population parameter

$$\theta$$

Samples from population

$$X_1, X_2, \cdots, X_n \quad \text{(approximately) iid}$$

Statistic

A statistic is any function $f(X_1, X_2, \cdots, X_n)$ of samples.

Estimator

An estimator of θ is a statistic $f(X_1, X_2, \cdots, X_n)$ used to estimate θ.

Unbiased estimator

An estimator $f(X_1, X_2, \cdots, X_n)$ of θ is unbiased if

$$\mathbb{E}f(X_1, X_2, \cdots, X_n) = \theta$$

Example - Unbiased estimation of mean and variance

Let X_i be n iid samples from a distribution with unknown mean μ and variance σ^2. Then,

Sample mean	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$	Unbiased estimator of μ
Sample variance	$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$	Unbiased estimator of σ^2

$$\mathbb{E}\bar{X} = \frac{\sum_{i=1}^n \mathbb{E}X_i}{n} = \mu$$

$$Var(\bar{X}) = \frac{1}{n^2} Var\left(\sum_{i=1}^n X_i\right) = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{\sigma^2}{n}$$

$$(X_i - \bar{X})^2 = ((X_i - \mu) - (\bar{X} - \mu))^2 = (X_i - \mu)^2 + (\bar{X} - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu)$$

$$\Rightarrow \quad \mathbb{E}(X_i - \bar{X})^2 = \sigma^2 + \frac{\sigma^2}{n} - 2\mathbb{E}(X_i - \mu)(\bar{X} - \mu)$$

$$\begin{aligned}
\mathbb{E}(X_i - \mu)(\bar{X} - \mu) &= \mathbb{E}(X_i - \mu)\left(\frac{\sum_{j=1}^n (X_j - \mu)}{n}\right) \\
&= \mathbb{E}(X_i - \mu)\left(\frac{\sum_{j\neq i} (X_j - \mu)}{n} + \frac{X_i - \mu}{n}\right) \\
&= \frac{1}{n}\mathbb{E}(X_i - \mu)^2 = \frac{\sigma^2}{n}
\end{aligned}$$

$$\Rightarrow \quad \mathbb{E}(X_i - \bar{X})^2 = \sigma^2 + \frac{\sigma^2}{n} - 2\mathbb{E}(X_i - \mu)(\bar{X} - \mu) = \frac{n-1}{n}\sigma^2$$

$$\mathbb{E}S^2 = \frac{\sum_{i=1}^n \mathbb{E}(X_i - \bar{X})^2}{n-1} = \sigma^2$$

Example - Number of pairs with same birthday

There are n people in the class. Each choose one's birthday independently and uniformly over the 365 days. Let S_n be the number of pairs with same birthday. Let A_{ij} be the event that i and j share the common birthday and let $1_{A_{ij}}$ be its indicator.

$$S_n = \sum_{1 \le i < j \le n} 1_{A_{ij}} \quad \text{is not } B(m,p),\ m = \binom{n}{2},\ p = 1/365$$

where

(1) $1_{A_{ij}} \sim B(p)$

(2) $1_{A_{ij}}$ not independent (but pairwise independent)

Mean

$$\mathbb{E}S_n = \sum_{1 \le i < j \le n} \mathbb{E}1_{A_{ij}} = \binom{n}{2} \cdot \frac{1}{365}$$

Variance

$$Var(S_n) = \sum_{1 \le i < j \le n} Var(1_{A_{ij}}) = \binom{n}{2} \cdot \frac{1}{365} \cdot \frac{364}{365}$$

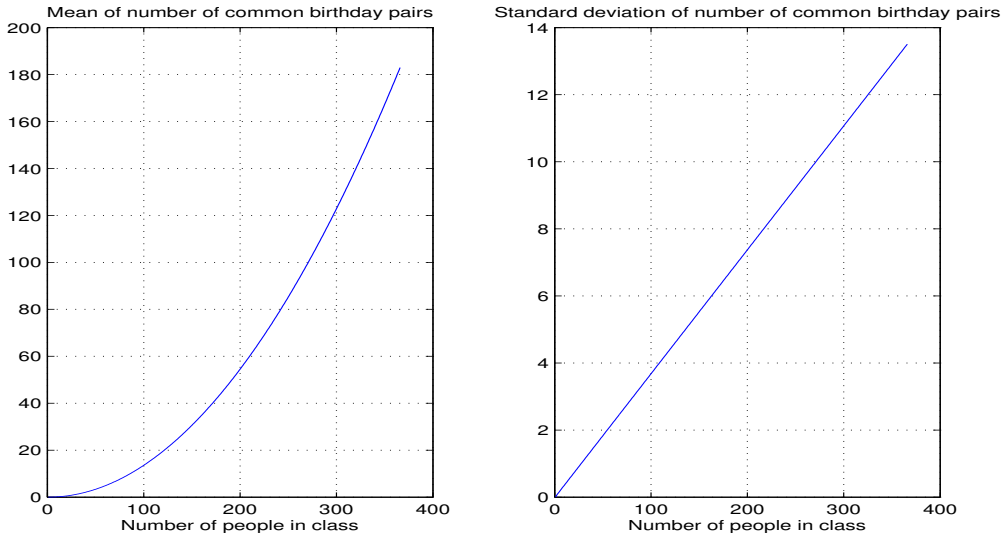

Figure 2: Mean and standard deviation of the number of common birthday pairs

```
clear all; close all; clc;

n=1:366; % Number of people

mu=binomial(n,2)/365;
v2=mu*364/365;
v=sqrt(v2);

subplot(1,2,1)
plot(n,mu); grid on
xlabel('Number of people in class');
title('Mean of number of common birthday pairs')

subplot(1,2,2)
plot(n,v); grid on
xlabel('Number of people in class');
title('Standard deviation of number of common birthday pairs')
```

Example - Number of empty bins

There are n balls and 365 bins. Each ball choose one's bin independently and uniformly over the 365 bins and each ball moves into the chosen bin. Let S_n be the number of overall empty bins. Let A_i be the event that the ith bin is empty and let 1_{A_i} be its indicator.

$$S_n = \sum_{i=1}^{365} 1_{A_i} \quad \text{is not } B(m,p),\ m = 365,\ p = (364/365)^n$$

where

$$\begin{array}{ll} (1) & 1_{A_i} \sim B(p) \\ (2) & 1_{A_i} \text{ not independent} \end{array}$$

Mean

$$\mathbb{E}S_n = \sum_{i=1}^{365} \mathbb{E}1_{A_i} = 365p$$

Variance

$$Var(S_n) = 365Var(1_{A_1}) + 2\binom{365}{2}Cov(1_{A_1}, 1_{A_2})$$

where

$$\begin{aligned} Var(1_{A_1}) &= pq \\ Cov(1_{A_1}, 1_{A_2}) &= P(A_1A_2) - P(A_1)^2 = \left(\frac{363}{365}\right)^n - p^2 \end{aligned}$$

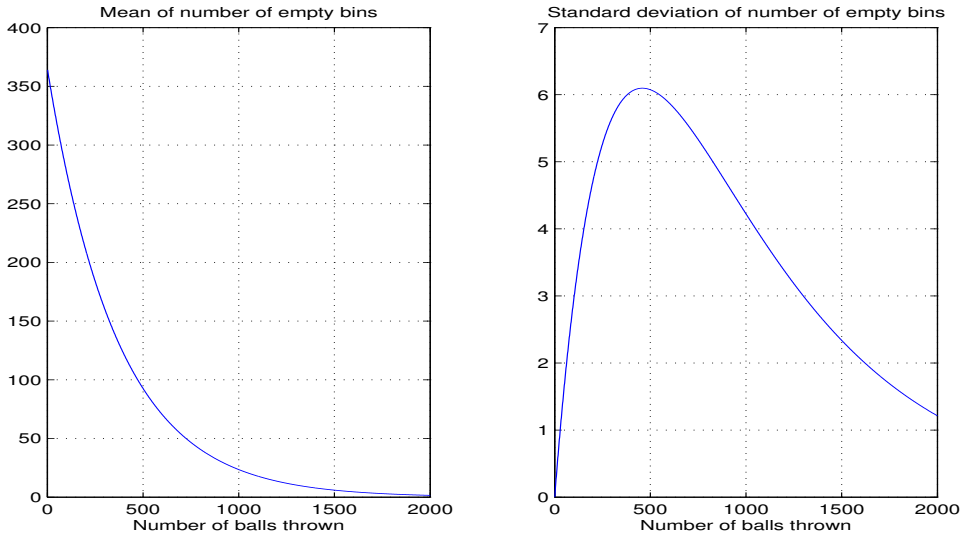

Figure 3: Mean and standard deviation of the number of empty bins

```
clear all; close all; clc;

n=1:2000; % Number of balls
M=365; % Number of bins

p1=((M-1)/M).^n; % Probability that a particular bin is empty
p2=((M-2)/M).^n; % Probability that two particular bins are empty
mu=M*p1;
v2=M.*p1.*(1-p1)+2*binomial(M,2).*(p2-p1.^2);
v=sqrt(v2);

subplot(1,2,1)
plot(n,mu); grid on
xlabel('Number of balls thrown');
title('Mean of number of empty bins')

subplot(1,2,2)
plot(n,v); grid on
xlabel('Number of balls thrown');
title('Standard deviation of number of empty bins')
```

Example - Number of stops

There are n people in the elevator at the basement. Each one choose one's stop independently and uniformly over the 365 floors. Let X_n be the number of overall stops. We can relate the number X_n of overall stops and the number S_n of empty bins:

$$\begin{aligned} \text{Person} &\leftrightarrow \text{Ball} \\ \text{choose} &\leftrightarrow \text{choose} \\ \text{Floor} &\leftrightarrow \text{Bin} \\ \text{to move} &\leftrightarrow \text{to move} \end{aligned}$$

$$\begin{aligned} \text{Stops} &\leftrightarrow \text{Bins with balls} \\ \text{Non-stop floors} &\leftrightarrow \text{Empty bins} \end{aligned}$$

$$X_n \quad \text{Number of stops} \quad \leftrightarrow \quad 365 - S_n \quad \text{Number of empty bins}$$

Mean

$$\mathbb{E}X_n = 365 - \mathbb{E}S_n$$

Variance

$$Var(X_n) = Var(S_n)$$

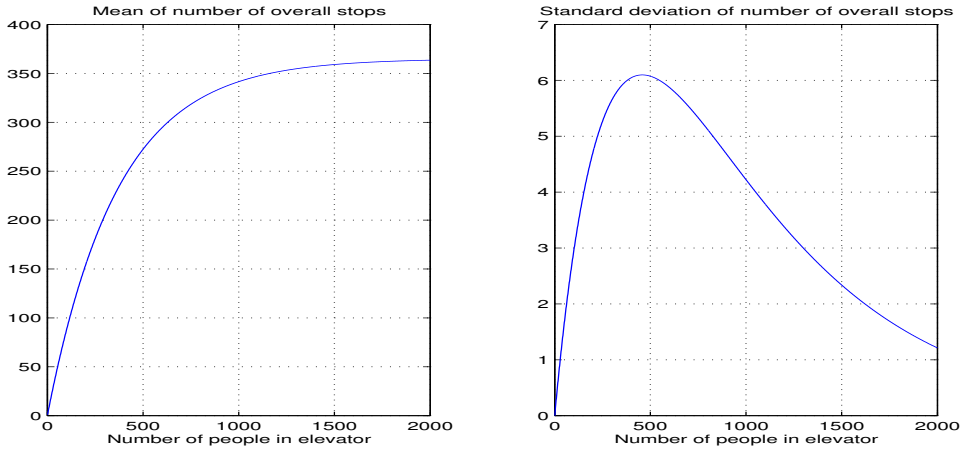

Figure 4: Mean and standard deviation of the number of overall stops

```
clear all; close all; clc;

n=1:2000; % Number of balls
M=365; % Number of bins

p1=((M-1)/M).^n; % Probability that a particular bin is empty
p2=((M-2)/M).^n; % Probability that two particular bins are empty
mu=365-M*p1;
v2=M.*p1.*(1-p1)+2*binomial(M,2).*(p2-p1.^2);
v=sqrt(v2);

subplot(1,2,1)
plot(n,mu); grid on
xlabel('Number of people in elevator');
title('Mean of number of overall stops')

subplot(1,2,2)
plot(n,v); grid on
xlabel('Number of people in elevator');
title('Standard deviation of number of overall stops')
```

Hypergeometric distribution $H(n, m, M)$

n	Number of samples (balls drawn) with/without replacement
m	Number of balls with marker H in the bin
M	Number of balls in the bin

Take n balls with or without replacement and count the number X or Y of heads.

With replacement	$X \sim B(n, p)$,	Mean np,	Variance npq
Without replacement	$Y \sim H(n, m, M)$,	Mean np,	Variance $npqf^2$

where

$$\begin{aligned} p &= \frac{m}{M} \\ f &= \sqrt{\frac{M-n}{M-1}} \end{aligned}$$

Hypergeometric distribution $H(n, m, M)$

Take n balls without replacement and count the number Y of heads. Let A_i be the event that the i-th chosen ball is H and let 1_{A_i} be its indicator. Then, Y can be represented in terms of 1_{A_i}:

$$Y = \sum_{i=1}^{n} 1_{A_i} \sim H(n, m, M)$$

where

(1) $1_{A_i} \sim B(p)$

(2) 1_{A_i} not independent

Mean

$$\mathbb{E}Y = \sum_{i=1}^{n} \mathbb{E}1_{A_i} = np$$

Variance

$$Var(Y) = npqf^2$$

$$\mathbb{P}(A_iA_j) = \mathbb{P}(A_i)\mathbb{P}(A_j|A_i) = \frac{m}{M}\cdot\frac{m-1}{M-1} = p\cdot\frac{m-1}{M-1} = p^2\cdot\frac{(m-1)/m}{(M-1)/M}$$

$$\begin{aligned}
Var(Y) &= \sum_{i=1}^{n} Var(1_{A_i}) + 2\sum_{1\le i<j\le n} Cov(1_{A_i}, 1_{A_j}) \\
&= nVar(1_{A_1}) + 2\binom{n}{2} Cov(1_{A_1}, 1_{A_2}) \\
&= npq + 2\binom{n}{2}\left(p^2\cdot\frac{(m-1)/m}{(M-1)/M} - p^2\right) \\
&= npq + n(n-1)\left(p^2\cdot\frac{(m-1)/m}{(M-1)/M} - p^2\right) \\
&= np\left[q + (n-1)\left(\frac{(m-1)/m}{(M-1)/M} - 1\right)p\right] \\
&= np\left[q + (n-1)\frac{-M+m}{m(M-1)}p\right] \\
&= np\left[q + (n-1)\frac{-qM}{m(M-1)}\frac{m}{M}\right] \\
&= npq\left[1 + (n-1)\frac{-M}{m(M-1)}\frac{m}{M}\right] \\
&= npq\left[\frac{M-n}{M-1}\right] \\
&= npqf^2
\end{aligned}$$

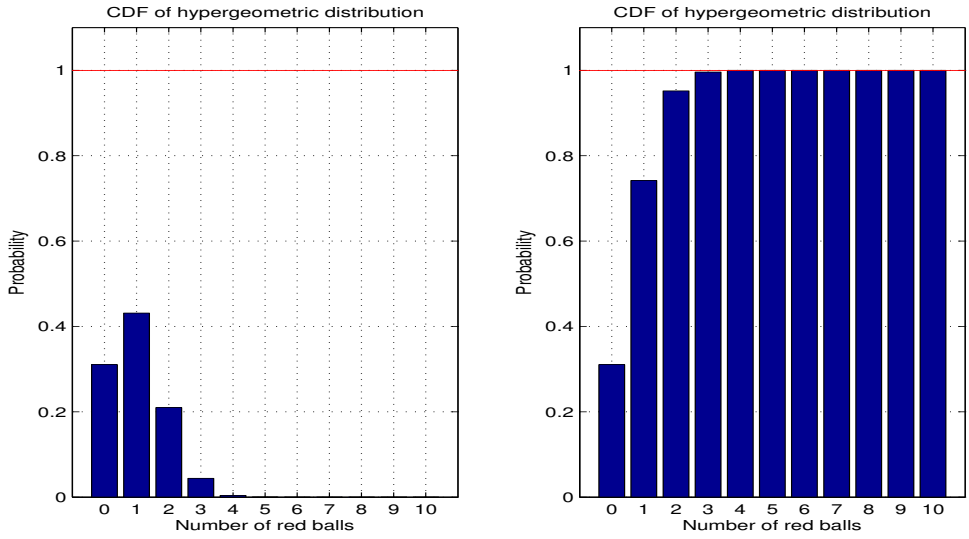

Figure 5: PMF and CDF of hypergeometric distribution $H(10, 5, 50)$

```
clear all; close all; clc;

n=10; % Number of balls drawn
m=5; % Number of red balls in bin
M=50; % Number of all balls in bin
p=m/M;

i=0:n; % Possible outcomes
pmf=binomial(m,i).*binomial(M-m,n-i)./binomial(M,n);
cdf=cumsum(pmf);

subplot(1,2,1)
bar(i,pmf); hold on; grid on;
plot(-1:0.01:n+1,1,'-r')
axis([-1 n+1 0 1.1])
xlabel('Number of red balls'); ylabel('Probability');
title('CDF of hypergeometric distribution');

subplot(1,2,2)
bar(i,cdf); hold on; grid on;
plot(-1:0.01:n+1,1,'-r')
axis([-1 n+1 0 1.1])
xlabel('Number of red balls'); ylabel('Probability');
title('CDF of hypergeometric distribution');
```

Example - Number of aces in hands

We takes five cards out of 52 cards and count the number X of aces in hands. Compute the mean and variance of X.

$n = 5$	Number of balls drawn without replacement
$m = 4$	Number of balls with marker H in the bin
$M = 52$	Number of balls in the bin

$$\begin{aligned} p &= \frac{m}{M} = \frac{4}{52} \\ f &= \sqrt{\frac{M-n}{M-1}} = \sqrt{\frac{52-5}{52-1}} \end{aligned}$$

$$\begin{aligned} X &\sim H(n, m, M) = H(5, 4, 52) \\ \mathbb{E}X &= np = 5 * \tfrac{4}{52} \\ Var(X) &= npqf^2 = 5 * \tfrac{4}{52} * \left(1 - \tfrac{4}{52}\right) * \tfrac{52-5}{52-1} \end{aligned}$$