Some discrete probability distributions

- Binomial distribution (二項分布) $Bin(n,p)$. $Bin(n,p)$ is the distribution of the number of successes in $n$ trials. It is assumed that the trials are independent with the same success probability $p$.

- The PMF of $Bin(n,p)$. Suppose that $X$'s distribution is $Bin(n,p)$. Let $p_X$ be the PMF of $X$, then

$$p_X(x) = \begin{cases} C_x^n p^x (1-p)^{n-x}; & \text{if } x \in \{0, \ldots, n\}, \\ 0 & \text{otherwise.} \end{cases}$$

  where

$$C_x^n = \frac{n!}{x!(n-x)!}$$

  is the number of ways of choosing $x$ items from $n$ different items. The text uses the notation ${}_nC_x$ instead of $C_x^n$.

- Some notation.
  - $X \sim Bin(n,p)$ means the distribution of $X$ is $Bin(n,p)$.
    * In general, $Bin(n,p)$ can be replaced by another distribution $\mathcal{D}$ and $X \sim \mathcal{D}$ means the distribution of $X$ is $\mathcal{D}$.
  - $P(Bin(n,p) \in A)$ is $P(X \in A)$, where $X \sim Bin(n,p)$.
    * In general, $Bin(n,p)$ can be replaced by another distribution $\mathcal{D}$ and $P(\mathcal{D} \in A)$ is $P(X \in A)$, where $X \sim \mathcal{D}$.
  - $X \sim Y$ means $X$ and $Y$ have the same distribution, where $X$ and $Y$ are random variables.

- $Bin(1,p)$ is known as the Bernoulli distribution with success probability $p$, denoted by $Ber(p)$. If $X_1$, ..., $X_n$ are IID and $X_i \sim Ber(p)$ for each $i$, then $\sum_{i=1}^n X_i \sim Bin(n,p)$.

- The mean and variance of $Bin(n,p)$. Suppose that $X \sim Bin(n,p)$. Then

$$E(X) = np \text{ and } Var(X) = np(1-p).$$

  - Note that the variance is small when $p$ is close to 0 or 1.

- Example 1. Suppose a coin is tossed twice and the probability of getting a tail in each toss is $p$. Let $X$ be the total number of tails in the two tosses. Find the PMF of $X$, $E(X)$ and $Var(X)$.

- Hypergeometric distribution (超幾何分布). Suppose that we have a group of $N$ items, of which $S$ items are good and $N-S$ items are defective. Choose a sample of $n$ items from the group and let $X$ be the number of good items in the sample. Then the distribution of $X$, denoted by $H(N,S,n)$, is a hypergeometric distribution.

- Let $p_X$ be the PMF of the hypergeometric distribution $H(N,S,n)$, then

$$p_X(x) = \frac{C_x^S C_{n-x}^{N-S}}{C_n^N}$$

  if $x \in \{0, 1, \ldots, n\}$, $x \le S$ and $n - x \le N - S$, and $p_X(x) = 0$ otherwise.

- Example 2.
  - Number of club members = 50, of which 40 are above 30 years old. Choose 5 members at random to form a committee.

  What is the probability that exactly 4 of the 5 selected for the committee are above 30 years old?

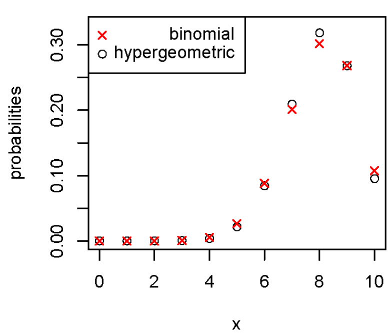
  Sol.

$$P(H(50, 40, 5) = 4) = \frac{C_4^{40} C_{5-4}^{50-40}}{C_5^{50}} = \frac{2193360}{5085024} = 0.4313372.$$

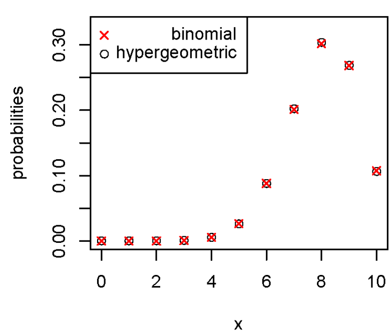
- The PMF of $H(N, S, n)$ can be approximated by the PMF of $Bin(n, p)$ when $S/N \approx p$ and $N$ is large comparing to $n$.
- Compute binomial probabilities using R.
  - `choose(n,x)` gives $C_x^n$.
- PMF comparison between $H(N, 0.8N, 10)$ and $Bin(10, 0.8)$ for $N = 100$ and $N = 1000$.
  - R-codes for $N = 100$.

```
N <- 100; S <- 0.8*N; n <- 10
x <- 0:10
pmf.hyper <- function(x, N,S,n){ choose(S,x)*choose(N-S,n-x)/choose(N,n) }
pmf.binom <- function(x, n, p){  choose(n,x)*(p^x)*( (1-p)^(n-x) ) }
plot(x, pmf.hyper(x, N,S,n), ylab="probabilities", type="p")
points(x, pmf.binom(x, n, 0.8), col="red", pch=4)
```

  - PMF plots




(a) $N = 100$




(b) $N = 1000$

- Poisson distribution $Poisson(\mu)$.

  - $Poisson(\mu)$ is the limit of $Bin(n, \mu/n)$ as $n \to \infty$. Suppose that $X$ is the number of occurrence of a (rare) event during a time period with $E(X) = \mu$. If the distribution of $X$ can be approximated well by $Bin(n, \mu/n)$ for large $n$, then it is reasonable to assume $X \sim Poisson(\mu)$.

- $Poisson(\mu)$ PMF. Suppose that $X \sim Poisson(\mu)$, then the PMF of $X$ is

$$p_X(x) = \frac{e^{-\mu}\mu^x}{x!} \text{ for } x \in \{0, 1, \ldots\}$$

  where

$$e = \lim_{y\to\infty}\left(1 + \frac{1}{y}\right)^y \approx 2.718.$$

- Facts about $e$:

$$e^\lambda = \lim_{y\to\infty}\left(1 + \frac{\lambda}{y}\right)^y \text{ and } e^\lambda = \sum_{x=0}^{\infty}\frac{\lambda^x}{x!}.$$

- Compute Poisson probabilities using R.

  - `factorial(n)` gives $n!$.
  - `exp(-0.3)` gives $e^{-0.3}$.
  - `exp(c(-0.3, 0.4))` gives the vector $(e^{-0.3}, e^{0.4})$.

- Example 3. 小明每天會帶狗出門散步. 根據以往經驗, 在散步途中, 平均每10小時會遇到3隻其他的狗. 今天小明打算帶狗出門散步1小時. 令$X$爲小明今天散步遇到的其他狗數. 提出適合$X$的分布並根據這分布計算小明今天散步沒有遇到其他狗的機率. 計算機率可使用R指令`exp(c(-0.1, -0.2, -0.3))`執行結果, 執行結果如下

```
> exp(c(-0.1, -0.2, -0.3))
[1] 0.9048374 0.8187308 0.7408182
```

  Sol. $X$ 分布爲離散型, 可能值爲 0, 1, 2, .... 在常用的 $Bin(n, p)$, $H(N, S, n)$, $Poisson(\mu)$ 幾種分布中, 以$Poisson(\mu)$較適合作爲$X$的分布, 因爲$Poisson(\mu)$的可能值和$X$的可能值一致. 另外, 我們預期散步1小時會遇到$3/10 = 0.3$隻狗, 所以可假設$\mu = E(X) = 0.3$. 小明今天散步沒有遇到其他狗的機率爲

$$P(Poisson(0.3) = 0) = \frac{e^{-0.3}(0.3)^0}{0!} = e^{-0.3} \approx 0.7408182.$$

- PMF comparison between $Poisson(0.3)$ and $Bin(n, 0.3/n)$ for $n = 100$.

  - R-codes

```
n <- 100; x <- 0:n
pmf.binom <- function(x, n, p){  choose(n,x)*(p^x)*( (1-p)^(n-x) ) }
pmf.poisson <- function(x, mu){ exp(-mu)*(mu^x)/factorial(x) }
plot(x, pmf.binom(x, n, .3/n), ylab="probabilities", type="p")
points(x, pmf.poisson(x, 0.3), col="red", pch=4)
```

– 執行以上程式, 可發現$Poisson(0.3)$和$Bin(n, 0.3/n)$的PMF差不多

- The mean and variance of $Poisson(\mu)$ are both equal to $\mu$. The results agree with the fact that $Poisson(\mu)$ can be approximated using $Bin(n, \mu/n)$ if $n$ is large.
- Compute the mean of $Poisson(\mu)$ using its PMF. For $X \sim Poisson(\mu)$,

$$\begin{aligned} E(X) &= \sum_{x=0}^{\infty} x e^{-\mu} \frac{\mu^x}{x!} \\ &= \sum_{x=1}^{\infty} x e^{-\mu} \frac{\mu^x}{x!} \\ &= \sum_{x=1}^{\infty} e^{-\mu} \frac{\mu \cdot \mu^{x-1}}{(x-1)!} \\ &= \mu \underbrace{\sum_{y=0}^{\infty} e^{-\mu} \frac{\mu^y}{y!}}_{=1} = \mu. \end{aligned}$$

- For $X \sim Poisson(\mu)$, it can be shown that $E(X(X-1)) = \mu^2$, so

  $$Var(X) = E(X^2) - (E(X))^2 = E(X(X-1)) + E(X) - (E(X))^2 = \mu^2 + \mu - \mu^2 = \mu.$$

  The verification of $E(X(X-1)) = \mu^2$ is left as an exercise.