常用分配查詢與生成模擬資料

- Textbook reading: Section 11.1.
- 分配符號
  - $Bin(n, p)$: binomial distribution with number of trials n and success probability p.
  - $N(\mu, \sigma^2)$: normal distribution with mean μ and variance σ^2.
  - For a random variable X and a distribution $\mathcal{D}$, $X \sim \mathcal{D}$ means the distribution of X is $\mathcal{D}$.
  - For a distribution $\mathcal{D}$, $P(\mathcal{D} \leq x)$ means $P(X \leq x)$, where $X \sim \mathcal{D}$.
- 分配$\mathcal{D}$的CDF (cumulative distribution function, 累積分布函數)。令

$$F(x) = P(\mathcal{D} \leq x), x \in (-\infty, \infty),$$

  則F稱爲分配$\mathcal{D}$的CDF.
- 分配$\mathcal{D}$的quantile. 令F爲$\mathcal{D}$的CDF, $q \in (0, 1)$. 一般情況下，$\mathcal{D}$的 q quantile 定義爲

$$\min\{x : F(x) \geq q\}.$$

  若F爲一對一且連續, 則$\mathcal{D}$的 q quantile 爲 $F^{-1}(q)$. 即若z 爲$\mathcal{D}$的 q quantile, 則

$$F(z) = P(\mathcal{D} \leq z) = q.$$

- 關於常用分配, 可用R查詢相關機率(CDF值)，quantile, PDF值，也可從給定的分配生成模擬資料。指令名稱分別爲 `pxxx`, `qxxx`, `dxxx`, `rxxx`, 其中`xxx` 爲分配名稱縮寫。例如常態分配名稱縮寫爲norm, 因此計算常態分配CDF值指令名稱分別爲 `pnorm`. 以 $N(0, 1.1^2)$爲例，
  - `pnorm(1.96, mean=0, sd=1.1)` 計算$P(N(0, 1.1^2) \leq 1.96)$.
  - `qnorm(0.95, mean=0, sd=1.1)` 計算 $N(0, 1.1^2)$的 0.95 quantile, 也就是使$P(N(0, 1.1^2) \leq z) = 0.95$的$z$值。
  - `dnorm(x, mean=0, sd=1.1)`計算$N(0, 1.1^2)$的 PDF (probability density function) 在`x` 的值，即

```
exp(-x^2/(2*1.1^2))/(sqrt(2*pi)*1.1)
```

  - `rnorm(100, mean=0, sd=1.1)` 生成100筆來自$N(0, 1.1^2)$的模擬資料
- Example 1.

  (a) 計算 $N(0, 1.1^2)$的 0.975 quantile.

  (b) 計算 $P(-1 < N(0, 1.1^2) \leq 1)$即

$$P(N(0, 1.1^2) \leq 1) - P(N(0, 1.1^2) \leq -1).$$

(c) 畫出 $N(0, 1.1^2)$的PDF (probability density function) 在$[-2, 2]$ 上的圖形。

(d) 畫出$N(0, 1.1^2)$的CDF在$[-2, 2]$ 上的圖形。

Sol.

```
#a
qnorm(0.975, mean=0, sd=1.1)

#b
pnorm(1, mean=0, sd=1.1)-pnorm(-1, mean=0, sd=1.1)

#c
f <- function(x){ return(dnorm(x, mean=0, sd=1.1)) }
curve(f, -2,2)

#d
F <- function(x){ return(pnorm(x, mean=0, sd=1.1)) }
curve(F, -2,2)
```

- Example 2. 生成10000筆來自$N(0, 1)$的模擬資料，計算資料落在區間$(-1, 1]$的比例並與$P(-1 < N(0, 1) \leq 1)$比較。

  Sol.

```
set.seed(1)
x <- rnorm(10000)
length(x[(x>-1)&(x<=1)])/length(x) #資料落在區間$(-1,1]$的比例
pnorm(1)-pnorm(-1)                 #P(-1 < N(0,1) < 1)
```

  資料落在區間$(-1, 1]$的比例接近$P(-1 < N(0, 1) \leq 1)$.

- `set.seed` 用於指定seed, 而 seed 決定亂數生成方式。指定同樣seed則生成同樣亂數。例如生成來自$N(0, 1)$的模擬資料二次時，若先執行`set.seed`, 則二次生成的模擬資料是一樣的。

```
set.seed(1)
rnorm(1)
set.seed(1)
rnorm(1)
```

  若不執行`set.seed`, 則二次生成的模擬資料是不一樣的

```
rnorm(1)
rnorm(1)
```

- Binomial distributions.

- `pbinom(4, 10, 0.3)`: 計算$P(Bin(10,0.3) \leq 4)$.
- `qbinom(0.95, 10, 0.3)`: 計算$Bin(10,0.3)$的 0.95 quantile.
- `dbinom(4, 10, 0.3)`: 計算 $P(Bin(10,0.3) = 4)$.
- `rbinom(100, 10, 0.3)`: 生成100筆來自$Bin(10,0.3)$的模擬資料

- Example 3. 假設X爲分配爲 $Bin(10,0.3)$的隨機變數. 計算 $P(X = 0)$, $P(X = 1)$, $P(X = 2)$ 以及 $P(X \leq 2)$.

  Sol.

```
dbinom(0, 10, 0.3)  #P(Bin(10,0.3)=0)
dbinom(1, 10, 0.3)  #P(Bin(10,0.3)=1)
dbinom(2, 10, 0.3)  #P(Bin(10,0.3)=2)
                    # or  dbinom(0:2, 10, 0.3)

pbinom(2, 10, 0.3)  #P(Bin(10,0.3)<=2), or sum(dbinom(0:2, 10, 0.3))
```

- 其它常用分配: 先用

```
help.search("Distribution$", package="stats", fields="title")
```

  列出所有分配再用 `help`查詢特定分配。

  Example 4. 假設X爲 random variable, X的分布爲自由度20的卡方分布 (chi-squared distribution with 20 degrees of freedom)。計算$P(X > 19.5)$.

  Sol. 執行

```
help.search("Distribution$", package="stats", fields="title")
```

  列出所有分配. 搜尋結果中有一筆是 chi-squared distribution:

```
stats::Chisquare       The (non-central) Chi-Squared Distribution
```

  執行

```
help(Chisquare)
```

  可查到 `dchisq`, `pchisq`, `qchisq`, `rchisq` 四個指令. 其中

```
 pchisq(19.5, 20)
```

  爲 $P(X \leq 19.5)$, 因此$P(X > 19.5)$爲

```
1-pchisq(19.5, 20)
```

- 若離散型分配只有有限多個可能值, 則可用 `sample` 指令生成模擬資料. 語法: `sample(x, n, replace=T, prob=p)`, 其中`x`爲分配可能值組成的向量, `p`爲對應之機率, `n`爲資料筆數.

  Example 5. 假設X爲 random variable, $P(X=3)=0.4$, $P(X=6)=0.3$, $P(X=8)=0.3$. 從X的分配生成1000筆模擬資料並計算每個可能值的出現比例.

```
y <- sample(c(3,6,8), 1000, replace=T, prob=c(0.4, 0.3,0.3))
length(y[y==3])/1000
length(y[y==6])/1000
length(y[y==8])/1000
```