7 - Sử dụng R cho tính toán xác suất

7.1 Hoán vị (permutation)

Chúng ta biết 3! = 3.2.1 = 6, và 0!=1. Nói chung, công thức tính số hoán vị cho một số n là: $n! = n(n-1)(n-2)(n-3)-...-1$. Trong R cách tính này rất đơn giản với lệnh `prod()` như sau:

Tìm 3!

```
> prod(3:1)
[1] 6
```

Tìm 10!

```
> prod(10:1)
[1] 3628800
```

Tìm 10.9.8.7.6.5.4

```
> prod(10:4)
[1] 604800
```

Tìm (10.9.8.7.6.5.4) / (40.39.38.37.36)

```
> prod(10:4) / prod(40:36)
[1] 0.007659481
```

7.2 Tổ hợp (combination)

$$C_n^k = \frac{n!}{k!(n-k)!}$$

Tổ hợp tính bằng hàm choose(n,k) Thí dụ choose(5,2) = 10

7.3 Biến ngẫu nhiên và hàm phân phối

Khi nói đến "phân phối" (hay distribution) là đề cập đến các giá trị mà biến có thể có. Các *hàm phân phối* (distribution function) là hàm mô tả các biến đó một cách hệ thống. "Có hệ thống" ở đây có nghĩa là theo một mô hình toán học cụ thể với những thông số cho trước. Trong xác suất thống kê có khá nhiều hàm phân phối, chúng ta sẽ em xét qua một số hàm quan trọng nhất và thông dụng nhất: đó là phân phối nhị phân, phân phối Poisson, và phân phối chuẩn. Trong mỗi luật phân phối, có 4 loại hàm quan trọng mà chúng ta cần biết:

hàm mật độ xác suất (probability density distribution);
hàm phân phối tích lũy (cumulative probability distribution);
hàm định bậc (quantile); và
hàm mô phỏng (simulation).

R có những hàm đinh sẵn có thể ứng dụng cho tính toán xác suất. Tên mỗi hàm được gọi bằng một tiếp đầu ngữ để chỉ loại hàm phân phối, và viết tắt tên của hàm đó. Các tiếp đầu ngữ là `d` (chỉ distribution hay xác suất), `p` (chỉ cumulative probability, xác suất tích lũy),

`q` (chỉ định bậc hay quantile), và `r` (chỉ random hay số ngẫu nhiên). Các tên viết tắt là `norm` (normal, phân phối chuẩn), `binom` (binomial , phân phối nhị phân), `pois` (Poisson, phân phối Poisson), v.v… Bảng sau đây tóm tắt các hàm và thông số cho từng hàm:

pphoi	**Mật độ**	**Tích lũy**	**Định bậc**	**Mô phỏng**
Chuẩn	dnorm(x, mean,sd)	pnorm(q, mean, sd)	qnorm(p, mean, sd)	rnorm(n, mean, sd)
Nhị phân	dbinom(k, n, p)	pbinom(q, n, p)	qbinom (p, n, p)	rbinom(k, n, prob)
Poisson	dpois(k, lambda)	ppois(q, lambda)	qpois(p, lambda)	rpois(n, lambda)
Uniform	dunif(x,min,max)	punif(q, min, max)	qunif(p, min, max)	runif(n, min, max)
Negative binomial	dnbinom(x, k, p)	pnbinom(q, k, p)	qnbinom (p,k,prob)	rbinom(n, n, prob)
Beta	dbeta(x, shape1,shape2)	pbeta(q,shape1, shape2)	qbeta(p,shape1, shape2)	rbeta(n,shape1, shape2)
Gamma	dgamma(x, shape, rate, scale)	gamma(q,shape ,rate,scale)	qgamma(p,shape , rate, scale)	rgamma(n, shape, rate, scale)
Geometric	dgeom(x, p)	pgeom(q, p)	qgeom(p, prob)	rgeom(n, prob)
Exponentia l	dexp(x, rate)	pexp(q, rate)	qexp(p, rate)	rexp(n, rate)
Weibull	dnorm(x, mean, sd)	pnorm(q, mean, sd)	qnorm(p, mean, sd)	rnorm(n, mean, sd)
Cauchy	dcauchy(x, location, scale)	pcauchy(q, location, scale)	qcauchy(p, location, scale)	rcauchy(n, location, scale)
F	df(x, df1, df2)	pf(q, df1, df2)	qf(p, df1, df2)	rf(n, df1, df2)
T	dt(x, df)	pt(q, df)	qt(p, df)	rt(n, df)
Chi-squared	dchisq(x, df)	pchi(q, df)	qchisq(p, df)	rchisq(n, df)

Chú thích: Trong bảng trên, df = degrees of freedome (bậc tự do);prob = probability (xác suất); n = sample size (số lượng mẫu). Các thông số khác có thể tham khảo thêm cho từng luật phân phối. Riêng các luật phân phối F, t, Chi-squared còn có một thông số khác nữa là non-centrality parameter (ncp) được cho số 0. Tuy nhiên người sử dụng có thể cho một thông số khác thích hợp, nếu cần.

7.3.1 Hàm phân phối nhị phân (Binomial distribution)

Như tên gọi, hàm phân phối nhị phân chỉ có hai giá trị: nam / nữ, sống / chết, có / không, v.v… Hàm nhị phân được phát biểu bằng định lí như sau: Nếu một thử nghiệm được tiến hành n lần, mỗi lần cho ra kết quả hoặc là thành công hoặc là thất bại, và gồm xác suất thành công được biết trước là p, thì xác suất có k lần thử nghiệm thành công là: $P(k \mid n, p) = C_k^n p^k (1-p)^{n-k}$, trong đó $k = 0, 1, 2, \ldots, n$. Trong R, có hàm `dbinom(k, n, p)` có thể giúp chúng ta tính công thức $P(k \mid n, p) = C_k^n p^k (1-p)^{n-k}$ một cách nhanh chóng. Trong trường hợp trên, chúng ta chỉ cần đơn giản lệnh:

```
> dbinom(2, 3, 0.60)
[1] 0.432
```

Ví dụ 2: Hàm nhị phân tích lũy (Cumulative Binomial probability distribution). Xác suất thuốc chống loãng xương có hiệu nghiệm là khoảng 70% (tức là $p = 0.70$). Nếu chúng ta điều trị 10 bệnh nhân, xác suất có tối thiểu 8 bệnh nhân với kết quả tích cực là bao nhiêu? Nói cách khác, nếu gọi X là số bệnh nhân được điều trị thành công, chúng ta cần tìm $P(X \geq 8) = ?$ Để trả lời câu hỏi này, chúng ta sử dụng hàm `pbinom(k, n, p)`. Xin nhắc lại rằng hàm `pbinom(k, n, p)` cho chúng ta $P(X \leq k)$. Do đó, $P(X \geq 8) = 1 - P(X \leq 7)$. Thành ra, đáp số bằng R cho câu hỏi là:

```
> 1-pbinom(7, 10, 0.70)
[1] 0.3827828
```

Ví dụ 3: Mô phỏng hàm nhị phân: Biết rằng trong một quần thể dân số có khoảng 20% người mắc bệnh cao huyết áp; nếu chúng ta tiến hành chọn mẫu 1000 lần, mỗi lần chọn 20 người trong quần thể đó một cách ngẫu nhiên, sự phân phối số bệnh nhân cao huyết áp sẽ như thế nào? Để trả lời câu hỏi này, chúng ta có thể ứng dụng hàm `rbinom (n, k, p)` trong R với những thông số như sau:

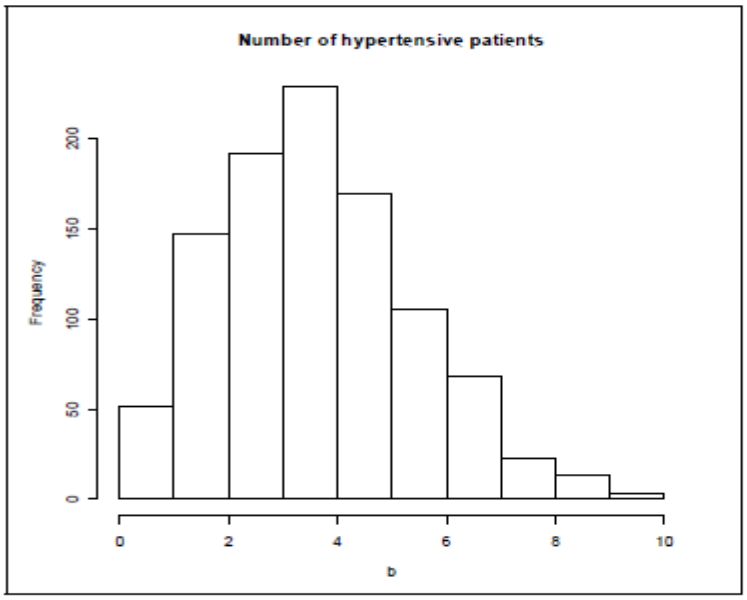
```
> b <- rbinom(1000, 20, 0.20)
```

Trong lệnh trên, kết quả mô phỏng được tạm thời chứa trong đối tượng tên là `b`. Để biết `b` có gì, chúng ta đếm bằng lệnh `table`:

```
> table(b)
b
  0   1   2   3   4   5   6   7   8   9  10
  6  45 147 192 229 169 105  68  23  13   3
```

Dòng số liệu thứ nhất (0, 5, 6, …, 10) là số bệnh nhân mắc bệnh cao huyết áp trong số 20 người mà chúng ta chọn. Dòng số liệu thứ hai cho chúng ta biết số lần chọn mẫu trong 1000 lần xảy ra. Do đó, có 6 mẫu không có bệnh nhân cao huyết áp nào, 45 mẫu với chỉ 1 bệnh nhân cao huyết áp, v.v… Có lẽ cách để hiểu là vẽ đồ thị các tần số trên bằng lệnh `hist` như sau:

```
> hist(b, main="Number of hypertensive patients")
```



7.3.2 Hàm phân phối Poisson (Poisson distribution)

Hàm phân phối Poisson, nói chung, rất giống với hàm nhị phân, ngoại trừ thông số *p* thường rất nhỏ và *n* thường rất lớn. Vì thế, hàm Poisson thường được sử dụng để mô tả các biến số rất hiếm xảy ra (như số người mắc ung thư trong một dân số chẳng hạn). Hàm Poisson còn được ứng dụng khá nhiều và thành công trong các nghiên cứu kĩ thuật và thị trường như số lượng khách hàng đến một nhà hàng mỗi giờ.

Ví dụ 4: **Hàm mật độ Poisson (Poisson density probability function)**. Qua theo dõi nhiều tháng, người ta biết được tỉ lệ đánh sai chính tả của một thư kí đánh máy. Tính trung bình cứ khoảng 2.000 chữ thì thư kí đánh sai 1 chữ. Hỏi xác suất mà thư kí đánh sai chính tả 2 chữ, hơn 2 chữ là bao nhiêu?

Vì tần số khá thấp, chúng ta có thể giả định rằng biến số "sai chính tả" (tạm đặt tên là biến số *X*) là một hàm ngẫu nhiên theo luật phân phối Poisson. Ở đây, chúng ta có tỉ lệ sai chính tả trung bình là 1 $\lambda = 1$). Luật phân phối Poisson phát biểu rằng xác suất mà $X = k$, với điều kiện tỉ lệ trung bình λ

$$p(X = k) = e^{-\lambda} \lambda^k /k!$$

Do đó, đáp số cho câu hỏi trên là: $e^{-1}/2! = 0{,}1839$
tính bằng R một cách nhanh chóng hơn bằng hàm `dpois` như sau:

```
> dpois(2, 1)
[1] 0.1839397
```

Chúng ta cũng có thể tính xác suất sai 1 chữ, và xác suất không sai chữ nào:

```
> dpois(1, 1)
[1] 0.3678794

>  dpois(0,  1)
[1] 0.3678794
> dpois(2,1)
```

Chú ý trong hàm trên, chúng ta chỉ đơn giản cung cấp thông số $k = 2$ và $\lambda = 1$. Trên đây là xác suất mà thư kí đánh sai chính tả đúng 2 chữ. Nhưng xác suất mà thư kí đánh sai chính tả hơn 2 chữ (tức 3, 4, 5, … chữ) có thể ước tính bằng:

$$P(X > 2) = P(X = 3) + P(X = 4) + P(X = 5) + ...$$

$$= 1 - (X \leq 2) = 1 - 0.3678 - 0.3678 - 0.1839 = 0.08$$

Bằng R, chúng ta có thể tính như sau:

```
# P(X≤2) > ppois(2, 1) [1] 0.9196986
# 1-P(X≤2)
> 1-ppois(2, 1) [1] 0.0803014
```

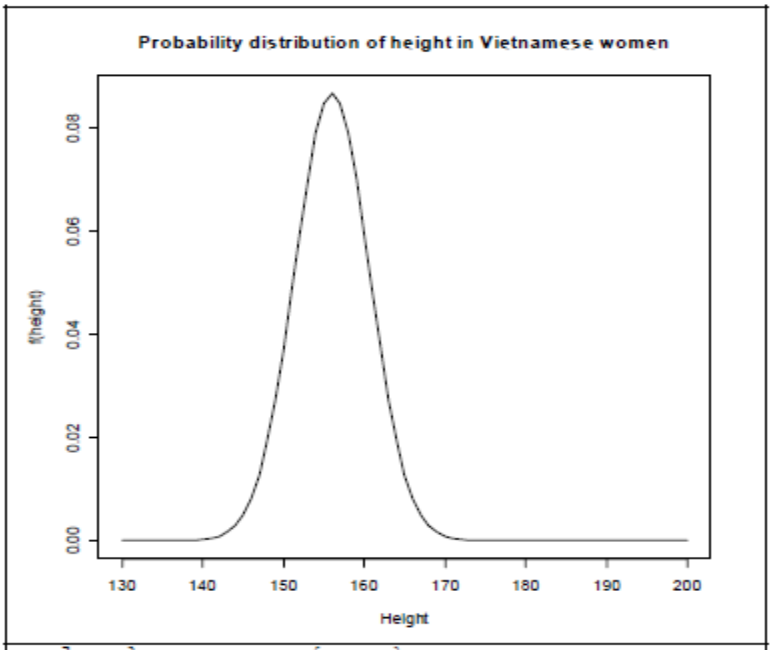
7.3.3 Hàm phân phối chuẩn (Normal distribution)

Hai luật phân phối mà chúng ta vừa xem xét trên đây thuộc vào nhóm phân phối áp dụng cho các biến số phi liên tục (discrete distributions), mà trong đó biến số có những giá trị theo bậc thứ hay thể loại. Đối với các biến số liên tục, có vài luật phân phối thích hợp khác, mà quan trọng nhất là phân phối chuẩn. Phân phối chuẩn là nền tảng quan trọng nhất của phân tích thống kê. Có thể nói hầu hết lí thuyết thống kê được xây dựng trên nền tảng của phân phối chuẩn.

Hàm mật độ phân phối chuẩn có dạng:

$$P(X = x \mid \mu, \sigma^2) = f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2} \right]$$

Ví dụ 5: **Hàm mật độ phân phối chuẩn (Normal density probability function)**. Chiều cao trung bình hiện nay ở phụ nữ Việt Nam là 156 cm, với độ lệch chuẩn là 4.6 cm



Biểu đồ 2. Phân phối chiều cao ở phụ nữ Việt Nam với trung bình 156 cm và độ lệch chuẩn 4.6 cm. Trụng hoành là chiều cao và trục tung là xác suất cho mỗi chiều cao.

Biểu đồ trên được vẽ bằng hai lệnh sau đây. Lệnh đầu tiên nhằm tạo ra một biến số `height` có giá trị 130, 131, 132, …, 200 cm. Lệnh thứ hai là vẽ biểu đồ với điều kiện trung bình là 156 cm và độ lệch chuẩn là 4.6 cm.

```
> height <- seq(130, 200, 1)
> plot(height, dnorm(height, 156, 4.6),
       type="l",
       ylab="f(height)",
       xlab="Height",
       main="Probability distribution of height in Vietnamese women")
```

Với hai thông số trên (và biểu đồ), chúng ta có thể ước tính xác suất cho bất cứ chiều cao nào. Chẳng hạn như xác suất một phụ nữ Việt Nam có chiều cao 160 cm là:

$$P(X = 160 \mid \mu=156, \sigma=4.6) = \frac{1}{4.6\sqrt{2\times 3.1416}} \exp\left[-\frac{(160-156)^2}{2(4.6)^2}\right]$$
$$= 0.0594$$

Hàm `dnorm(x, mean, sd)` trong R có thể tính toán xác suất này cho chúng ta một cách gọn nhẹ:

```
> dnorm(160, mean=156, sd=4.6)
[1] 0.05942343
```

Hàm xác suất chuẩn tích lũy (cumulative normal probability function).

Vì hiều cao là một biến số liên tục, trong thực tế chúng ta ít khi nào muốn tìm xác suất cho một giá trị cụ thể *x*, mà thường tìm xác suất cho một khoảng giá trị *a* đến *b*. Chẳng hạn như chúng ta muốn biết xác suất chiều cao từ 150 đến 160 cm (tức là $P(150 \le X \le 160)$, hay xác suất chiều cao thấp hơn 145 cm, tức $P(X < 145)$. Để tìm đáp số các câu hỏi như thế, chúng ta cần đến hàm xác suất chuẩn tích lũy, được định nghĩa như sau:

$$P(a \le X \le b) = \int_a^b f(x)dx$$

Thành ra, $P(160 \le X \le 150)$ chính là diện tích tính từ trục hoành = 150 đến 160 của **biểu đồ 2**. Trong R có hàm `pnorm(x, mean, sd)` dùng để tính xác suất tích lũy cho một phân phối chuẩn rất có ích.

$$\texttt{pnorm(a, mean, sd)} = \int_{-\infty}^{a} f(x)dx = P(X \le a \mid \text{mean, sd})$$

Xác suất chiều cao phụ nữ Việt nam thấp hơn hay bằng 150 cm

```
> pnorm(150, 156, 4.6)
[1] 0.0960575
```

Khoảng 9,6% phụ nữ Việt nam thấp hơn hay bằng 150 cm

Hay xác suất chiều cao phụ nữ Việt Nam bằng hoặc cao hơn 165 cm là:

```
> 1-pnorm(164, 156, 4.6)
[1] 0.04100591
```

Nói cách khác, chỉ có khoảng 4.1% phụ nữ Việt Nam có chiều cao bằng hay cao hơn 165 cm.

Ví dụ 6: Ứng dụng luật phân phối chuẩn: Trong một quần thể, chúng ta biết áp suất máu trung bình là 100 mmHg và độ lệch chuẩn là 13 mmHg, hỏi: có bao nhiêu ngươi trong quần thể này có áp suất máu bằng hoặc cao hơn 120 mmHg? Câu trả lời bằng R là:

```
> 1-pnorm(120, mean=100, sd=13)
[1] 0.0619679
```

Tức khoảng 6.2% người trong quần thể này có áp suất máu bằng hoặc cao hơn 120 mmHg.

7.3.4 Hàm phân phối chuẩn chuẩn hóa (Standardized Normal distribution)

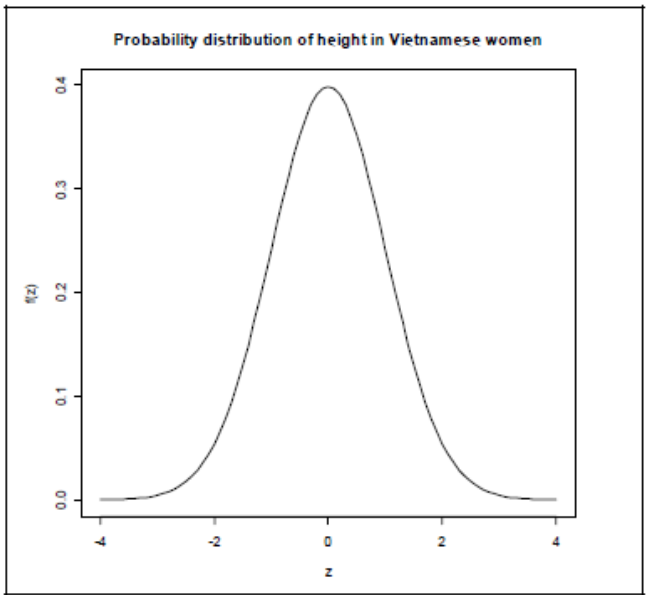
Một biến X tuân theo luật phân phối chuẩn với trung bình bình μ và phương sai σ^2 thường được viết tắt là:

$$X \sim N(\mu, \sigma^2)$$

Ở đây μ và σ^2 tùy thuộc vào đơn vị đo lường của biến số. Chẳng hạn như chiều cao được tính bằng cm (hay m), huyết áp được đo bằng mmHg, tuổi được đo bằng năm, v.v... cho nên đôi khi mô tả một biến số bằng đơn vị gốc rất khó so sánh. Một cách đơn giản hơn là chuẩn hóa (standardized) X sao cho số trung bình là 0 và phương sai là 1. Sau vài thao tác số học, có thể chứng minh dễ dàng rằng, cách biến đổi X để đáp ứng điều kiện trên là:

$$Z = \frac{X - \mu}{\sigma}$$

Biểu đồ phân phối chiều cao của phụ nữ Việt Nam có thể mô tả bằng một đơn vị mới, đó là chỉ số *z* như sau:



Biểu đồ 3. Phân phối chuẩn hóa chiều cao ở phụ nữ Việt Nam.

Biểu đồ trên được vẽ bằng hai lệnh sau đây:

```
> height <- seq(-4, 4, 0.1)
> plot(height, dnorm(height, 0, 1),
       type="l",
       ylab="f(z)",
       xlab="z",
       main="Probability distribution of height in Vietnamese women")
```

Với phân phối chuẩn chuẩn hoá, chúng ta có một tiện lợi là có thể dùng nó để mô tả và so sánh mật độ phân phối của bất cứ biến nào, vì tất cả đều được chuyển sang chỉ số z.

Trong biểu đồ trên, trục tung là xác suất z và trục hoành là biến số z. Chúng ta có thể tính toán xác suất z nhỏ hơn một hằng số (constant) nào đó dê dàng bằng R. Ví dụ, chúng ta muốn tìm $P(z \leq -1.96) = ?$ cho một phân phối mà trung bình là 0 và độ lệch chuẩn là 1.

```
> pnorm(-1.96, mean=0, sd=1)
[1] 0.02499790
```

Hay $P(z \leq 1.96) = ?$

```
> pnorm(1.96, mean=0, sd=1)
[1] 0.9750021
```

Do đó, $P(-1.96 < z < 1.96)$ chính là:

```
> pnorm(1.96) - pnorm(-1.96)
[1] 0.9500042
```

Nói cách khác, xác suất 95% là z nằm giữa -1.96 và 1.96. (Chú ý trong lệnh trên không cung cấp `mean=0, sd=1`, bởi vì trong thực tế, `pnorm` giá trị mặc định (default value) của thông số `mean` là 0 và `sd` là 1).

Ví dụ 5 (tiếp tục). Xin nhắc lại để tiện việc theo dõi, chiều cao trung bình ở phụ nữ Việt Nam là 156 cm và độ lệch chuẩn là 4.6 cm. Do đó, một phụ nữ có chiều cao 170 cm cũng có nghĩa là z = (170 – 156) / 4.6 = 3.04 độ lệch chuẩn, và ti lệ các phụ nữ Việt Nam có chiều cao cao hơn 170 cm là rất thấp, chỉ khoảng 0.1%.

```
> 1-pnorm(3.04)
[1] 0.001182891
```

Tìm định lượng (quantile) của một phân phối chuẩn. Đôi khi chúng ta cần làm một tính toán đảo ngược. Chẳng hạn như chúng ta muốn biết: nếu xác suất *Z* nhỏ hơn một hằng số *z* nào đó cho trước bằng *p*, thì *z* là bao nhiêu? Diễn tả theo kí hiệu xác suất, chúng ta muốn tìm *z* trong nếu:

$$P(Z < z) = p$$

Để trả lời câu hỏi này, chúng ta sử dụng hàm `qnorm(p, mean=, sd=)`.

Ví dụ 7: Biết rằng $Z \sim N(0, 1)$ và nếu $P(Z < z) = 0.95$, chúng ta muốn tìm *z*.

```
> qnorm(0.95, mean=0, sd=1)
[1] 1.644854
```

Hay $P(Z < z) = 0.975$ cho phân phối chuẩn với trung bình 0 và độ lệch chuẩn 1:

```
> qnorm(0.975, mean=0, sd=1)
 [1] 1.959964
```

8. Biểu đồ

Trong ngôn ngữ R có rất nhiều cách để thiết kế một biểu đồ gọn và đẹp. Phần lớn những hàm để thiết kế biểu đồ có sẵn trong R, nhưng một số loại biểu đồ tinh vi và phức tạp khác có thể thiết kế bằng các package chuyên dụng như `lattice` hay `trellis` có thể tải từ website của R. Trong chương này tôi sẽ chỉ cách vẽ các biểu đồ thông dụng bằng cách sử dụng các hàm phổ biến trong R.

8.1 Số liệu cho phân tích biểu đồ

Sau khi đã biết qua môi trường và những lựa chọn để thiết kế một biểu đồ, bây giờ chúng ta có thể sử dụng một số hàm thông dụng để vẽ các biểu đồ cho số liệu. Theo tôi, biểu đồ có thể chia thành 2 loại chính: biểu đồ dùng để mô tả một biến số và biểu đồ về mối liên hệ giữa hai hay nhiều biến số. Tất nhiên, biến số có thể là liên tục hay không liên tục, cho nên, trong thực tế, chúng ta có 4 loại biểu đồ. Trong phần sau đây, tôi sẽ điểm qua các loại biểu đồ, từ đơn giản đến phức tạp.

Có lẽ cách tốt nhất để tìm hiểu cách vẽ đồ thị bằng R là bằng một dữ liệu thực tế. Tôi sẽ quay lại **ví dụ 2** (phần 4.2). Trong ví dụ đó, chúng ta có dữ liệu gồm 8 cột (hay biến số): `id, sex, age, bmi, hdl, ldl, tc,` và `tg`. (Chú ý, `id` là mã số của 50 đối tượng nghiên cứu; `sex` là giới tính (nam hay nữ); `age` là độ tuổi; `bmi` là tỉ số trọng lương; `hdl` là high density cholesterol; `ldl` là low density cholesterol; `tc` là tổng số - total – cholesterol; và `tg` triglycerides). Dữ liệu được chứa trong directory directory `c:/works/insulin` dưới tên `chol.txt`. Trước khi vẽ đồ thị, chúng ta

bắt đầu bằng cách nhập dữ liệu này vào R.

```
> setwd("c:/works/stats")
> cong <- read.table("chol.txt", header=TRUE, na.strings=".")
> attach(cong)
```

Hay để tiện việc theo dõi tôi sẽ nhập các dữ liệu đó bằng các lệnh sau đây:

```
sex <- c("Nam", "Nu", "Nu","Nam","Nam", "Nu","Nam","Nam","Nam", "Nu",
         "Nu","Nam", "Nu","Nam","Nam", "Nu", "Nu", "Nu", "Nu", "Nu",
         "Nu", "Nu", "Nu", "Nu","Nam","Nam", "Nu","Nam", "Nu", "Nu",
         "Nu","Nam","Nam", "Nu", "Nu","Nam", "Nu","Nam", "Nu", "Nu",
         "Nam", "Nu","Nam","Nam","Nam", "Nu","Nam","Nam", "Nu", "Nu")

age <- c(57, 64, 60, 65, 47, 65, 76, 61, 59, 57,
         63, 51, 60, 42, 64, 49, 44, 45, 80, 48,
         61, 45, 70, 51, 63, 54, 57, 70, 47, 60,
         60, 50, 60, 55, 74, 48, 46, 49, 69, 72,
         51, 58, 60, 45, 63, 52, 64, 45, 64, 62)

bmi <- c( 17, 18, 18, 18, 18, 18, 19, 19, 19, 19, 20, 20, 20, 20, 20,
          20, 21, 21, 21, 21, 21, 21, 21, 21, 22, 22, 22, 22, 22, 22,
          22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 23, 23, 24, 24, 24,
          24, 24, 24, 25, 25)

hdl <- c(5.000,4.380,3.360,5.920,6.250,4.150,0.737,7.170,6.942,5.000,
         4.217,4.823,3.750,1.904,6.900,0.633,5.530,6.625,5.960,3.800,
         5.375,3.360,5.000,2.608,4.130,5.000,6.235,3.600,5.625,5.360,
         6.580,7.545,6.440,6.170,5.270,3.220,5.400,6.300,9.110,7.750,
         6.200,7.050,6.300,5.450,5.000,3.360,7.170,7.880,7.360,7.750)

ldl <- c(2.0,  3.0,  3.0,  4.0,  2.1,  3.0,  3.0,  3.0,  3.0, 2.0,
         5.0,  1.3,  1.2,  0.7,  4.0,  4.1,  4.3,  4.0,  4.3, 4.0,
         3.1,  3.0,  1.7,  2.0,  2.1,  4.0,  4.1,  4.0,  4.2, 4.2,
         4.4,  4.3,  2.3,  6.0,  3.0,  3.0,  2.6,  4.4,  4.3, 4.0,
         3.0,  4.1,  4.4,  2.8,  3.0,  2.0,  1.0,  4.0,  4.6,  4.0)

tc <-c (4.0, 3.5, 4.7, 7.7, 5.0, 4.2, 5.9, 6.1, 5.9, 4.0,
        6.2, 4.1, 3.0, 4.0, 6.9, 5.7, 5.7, 5.3, 7.1, 3.8,
        4.3, 4.8, 4.0, 3.0, 3.1, 5.3, 5.3, 5.4, 4.5, 5.9,
        5.6, 8.3, 5.8, 7.6, 5.8, 3.1, 5.4, 6.3, 8.2, 6.2,
        6.2, 6.7, 6.3, 6.0, 4.0, 3.7, 6.1, 6.7, 8.1, 6.2)

tg <- c(1.1, 2.1, 0.8, 1.1, 2.1, 1.5, 2.6, 1.5, 5.4, 1.9,
        1.7, 1.0, 1.6, 1.1, 1.5, 1.0, 2.7, 3.9, 3.0, 3.1,
        2.2, 2.7, 1.1, 0.7, 1.0, 1.7, 2.9, 2.5, 6.2, 1.3,
        3.3, 3.0, 1.0, 1.4, 2.5, 0.7, 2.4, 2.4, 1.4, 2.7,
        2.4, 3.3, 2.0, 2.6, 1.8, 1.2, 1.9, 3.3, 4.0, 2.5)

cong <- data.frame(sex, age, bmi, hdl, ldl, tc, tg)
```

8.2 Biểu đồ cho một biến số rời rạc (discrete variable): `barplot`

Biến `sex` trong dữ liệu trên có hai giá trị (`nam` và `nu`), tức là một biến không liên tục. Chúng ta muốn biết tần số của giới tính (bao nhiêu nam và bao nhiêu nữ) và vẽ một biểu đồ đơn giản. Để thực hiện ý định này, trước hết, chúng ta cần dùng hàm table để biết tần số:

```
> sex.freq <- table(sex)
>
sex.freq
sex
```

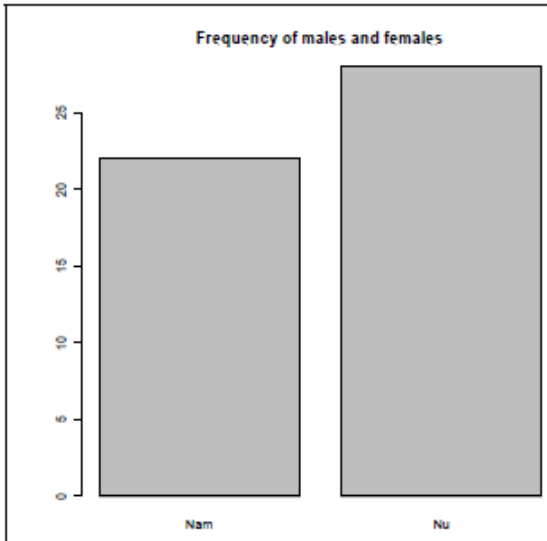
```
Nam  Nu
 22  28
```

Có 22 nam và 28 nữa trong nghiên cứu. Sau đó dùng hàm `barplot` để thể hiện tần số này như sau:

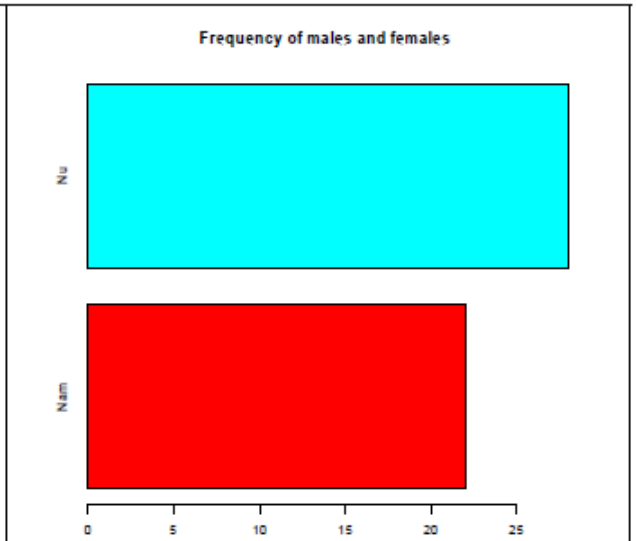
```
> barplot(sex.freq, main="Frequency of males and females")
```

Biểu trên cũng có thể có được bằng một lệnh đơn giản hơn (**Biểu đồ 8a**):

> barplot(table(sex), main="Frequency of males and females")



Biểu đồ 8a. Tần số giới tính thể hiện bằng cột số.



Biểu đồ 8b. Tần số giới tính thể hiện bằng dòng số.

Thay vì thể hiện tần số nam và nữ bằng 2 cột, chúng ta có thể thể hiện bằng hai dòng bằng thông số horiz = TRUE, như sau (xem kết quả trong **Biểu đồ 6b**):

> barplot(sex.freq,
horiz = TRUE,
col = rainbow(length(sex.freq)),
main="Frequency of males and females")

8.3 Biểu đồ cho hai biến số rời rạc (discrete variable): `barplot`

Age là một biến số liên tục. Chúng ta có thể chia bệnh nhân thành nhiều nhóm dựa vào độ tuổi. Hàm `cut` có chức năng "cắt" một biến liên tục thành nhiều nhóm rời rạc. Chẳng hạn như:

```
> ageg <- cut(age, 3)
> table(ageg)
ageg
  (42,54.7] (54.7,67.3]   (67.3,80]
         19          24           7
```

Có hiệu quả chia biến `age` thành 3 nhóm. Tần số của ba nhóm này là: 42 tuổi đến 54.7 tuổi thành nhóm 1, 54.7 đến 67.3 thành nhóm 2, và 67.3 đến 80 tuổi thành nhóm 3. Nhóm 1 có 19 bệnh nhân, nhóm 2 và 3 có 24 và 7 bệnh nhân.

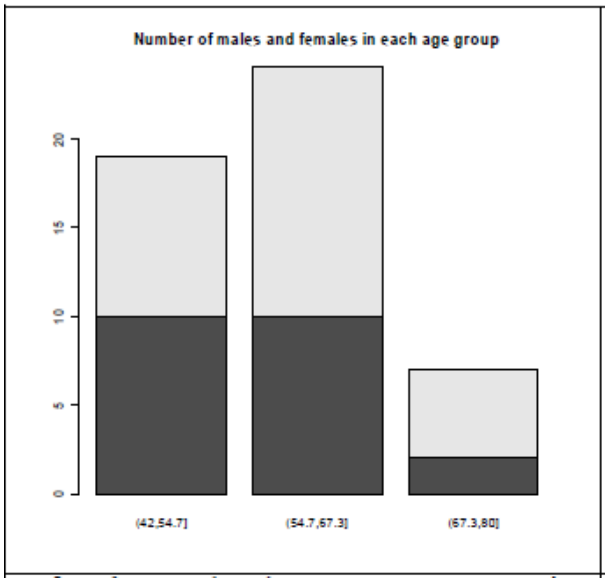
Bây giờ chúng ta muốn biết có bao nhiêu bệnh nhân trong từng độ tuổi và từng giới tính

bằng lệnh `table`:

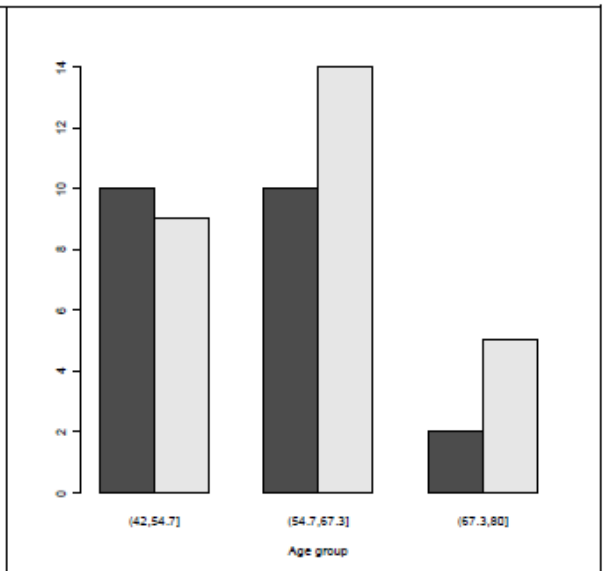
```
> age.sex <- table(sex, ageg)
> age.sex
      ageg
sex    (42,54.7] (54.7,67.3] (67.3,80]
  Nam         10          10         2
  Nu           9          14         5
```

Kết quả trên cho thấy chúng ta có 10 bệnh nhân nam và 9 nữ trong nhóm tuổi thứ nhất, 10 nam và 14 nữa trong nhóm tuổi thứ hai, v.v… Để thể hiện tần số của hai biến này, chúng ta vẫn dùng `barplot`:

```
> barplot(age.sex, main="Number of males and females in each age
   group")
```



Biểu đồ 9a. Tần số giới tính và nhóm tuổi thể hiện bằng cột số.



Biểu đồ 9b. Tần số giới tính và nhóm tuổi thể hiện bằng hai dòng số.

Trong **Biểu đồ 9a**, mỗi cột là cho một độ tuổi, và phần đậm của cột là nữ, và phần màu nhạt là tần số của nam giới. Thay vì thể hiện tần số nam nữ trong một cột, chúng ta cũng có thể thể hiện bằng 2 cột với `beside=T` như sau (**Biểu đồ 9b**):

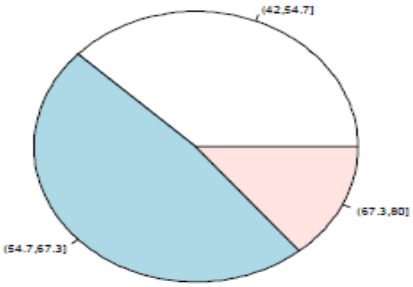
```
barplot(age.sex, beside=TRUE, xlab="Age group")
```

8.4 Biểu đồ hình tròn

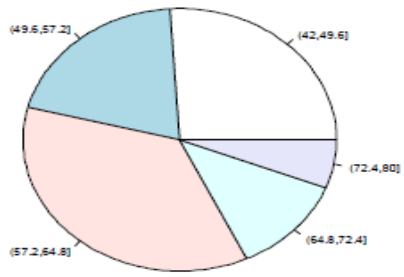
Tần số một biến rời rạc cũng có thể thể hiện bằng biểu đồ hình tròn. Ví dụ sau đây vẽ biểu đồ tần số của độ tuổi. **Biểu đồ 10a** là 3 nhóm độ tuổi, và **Biểu đồ 10b** là biểu đồ tần số cho 5 nhóm tuổi:

> pie(table(ageg))

Thí dụ pie(table(cut(age,5)))



Biểu đồ 10a. Tần số cho 3 nhóm tuổi



Biểu đồ 10b. Tần số cho 5 nhóm tuổi

8.5 Biểu đồ cho một biến số liên tục: stripchart và hist

8.5.1 Stripchart

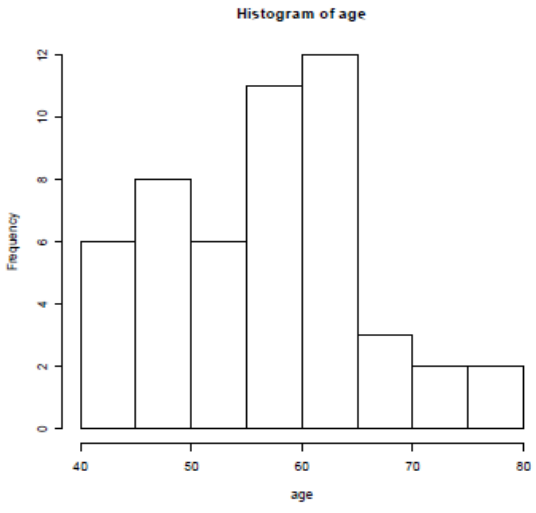
Biểu đồ strip cho chúng ta thấy tính liên tục của một biến số. Chẳng hạn như chúng ta muốn tìm hiểu tính liên tục của triglyceride (tg), hàm stripchart() sẽ giúp trong mục tiêu này:

> stripchart(tg, main="Strip chart for triglycerides", xlab="mg/L")

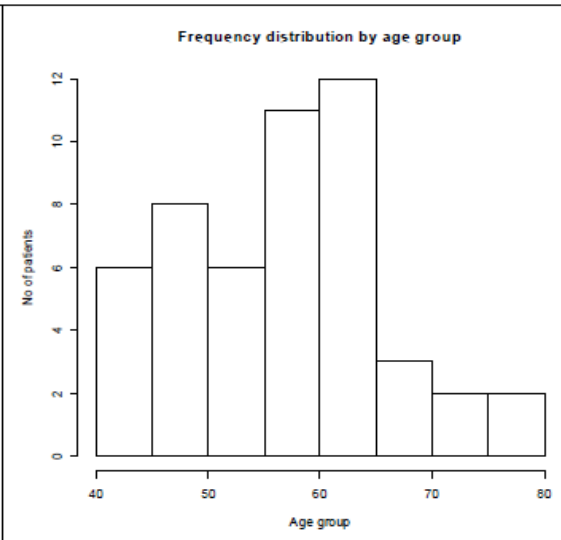
8.5.2 Histogram

Age là một biến số liên tục. Để vẽ biểu đồ tần số của biến số `age`, chúng ta chỉ đơn giản lệnh `hist(age)`. Như đã đề cập trên, chúng ta có thể cải tiến đồ thị này bằng cách cho thêm tựa đề chính (`main`) và tựa đề của trục hoành (`xlab`) và trục tung (`ylab`):

```
> hist(age)
> hist(age, main="Frequency distribution by age group",
xlab="Age group", ylab="No of patients")
```



Biểu đồ 11a. Trục tung là số bệnh nhân (đối tượng nghiên cứu) và trục hoành là độ tuổi. Chẳng hạn như tuổi 40 đến 45 có 6 bệnh nhân, từ 70 đến 80 tuổi có 4 bệnh nhân.

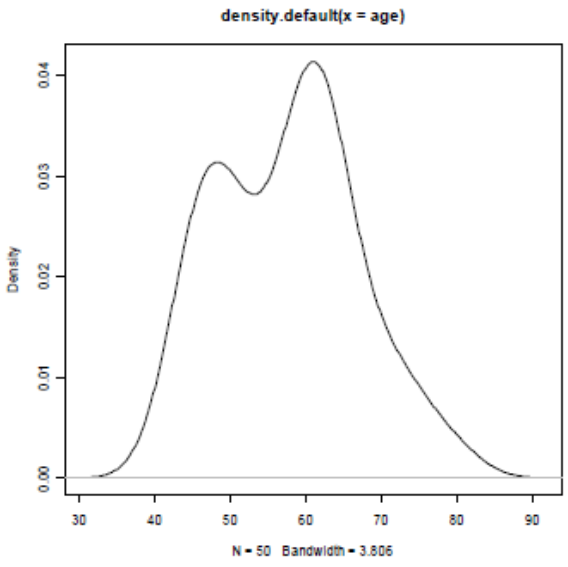


Biểu đồ 11b. Thêm tên biểu đồ và tên của trục trung và trục hoành bằng `xlab` và `ylab`.

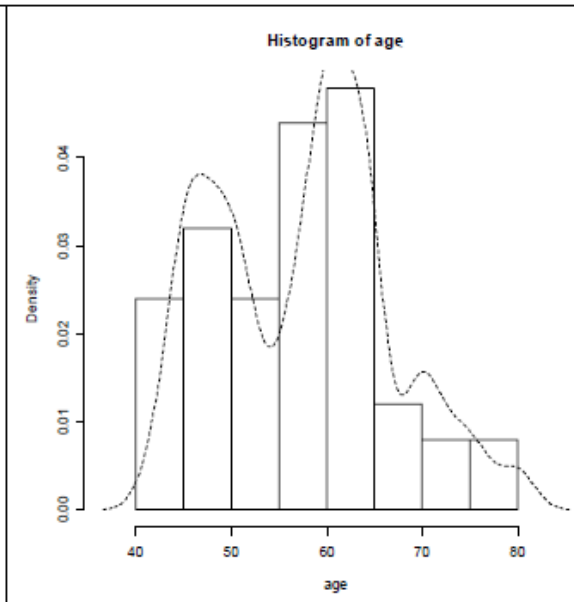
Chúng ta cũng có thể biến đổi biểu đồ thành một đồ thị phân phối xác suất bằng hàm

`plot(density)` như sau (kết quả trong **Biểu đồ 12a**):

```
> plot(density(age),add=TRUE)
```



Biểu đồ 12a. Xác suất phân phối mật độ cho biến age (độ tuổi).



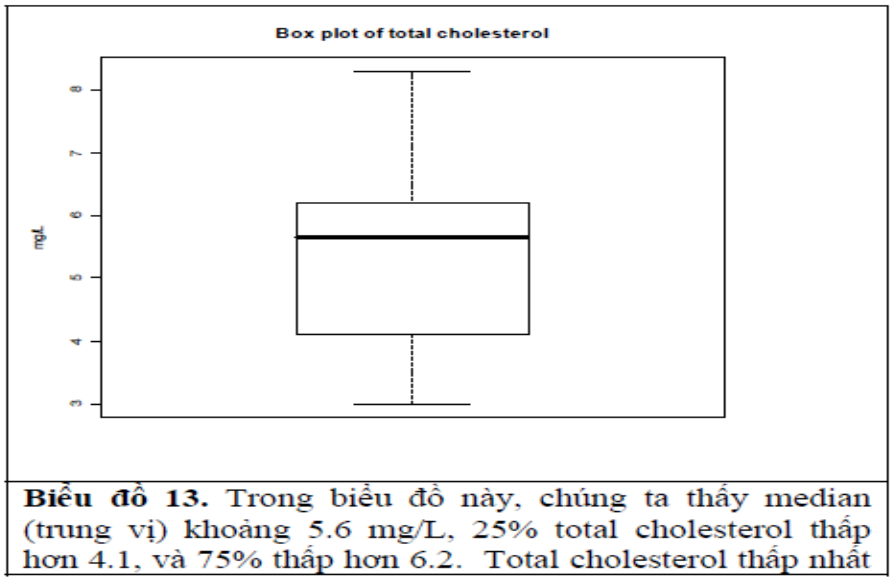
Biểu đồ 12b. Xác suất phân phối mật độ cho biến age (độ tuổi) với nhiều interquartile.

Chúng ta có thể vẽ hai đồ thị chồng lên bằng cách dùng hàm interquartile như sau (kết quả xem **Biểu đồ 12b**):

8.6 Biểu đồ hộp (`boxplot`)

Để vẽ biểu đồ hộp của biến số `tc`, chúng ta chỉ đơn giản lệnh:

```
> boxplot(tc, main="Box plot of total cholesterol", ylab="mg/L")
```



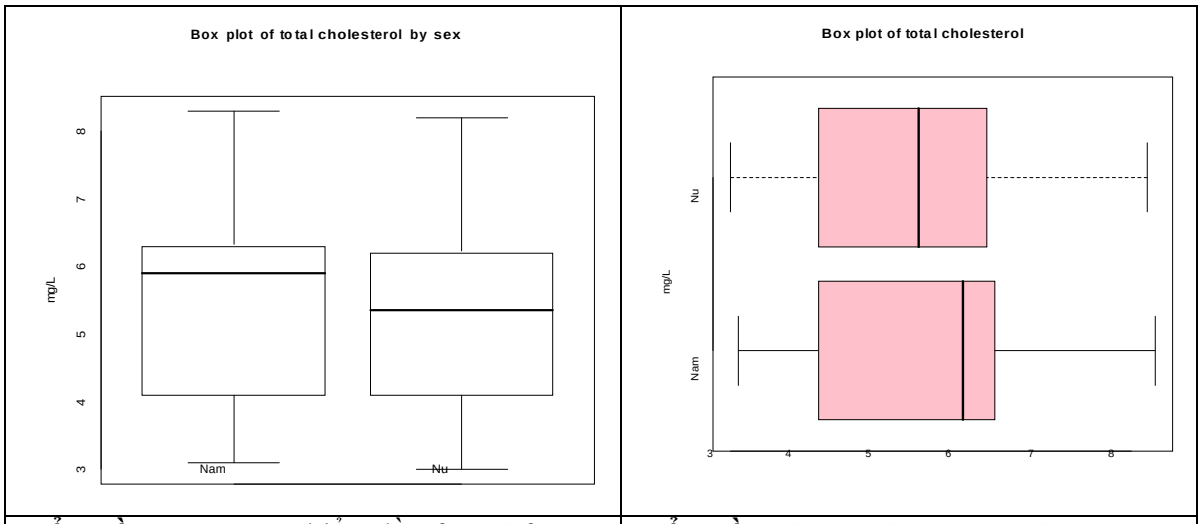
Biểu đồ 13. Trong biểu đồ này, chúng ta thấy median (trung vị) khoảng 5.6 mg/L, 25% total cholesterol thấp hơn 4.1, và 75% thấp hơn 6.2. Total cholesterol thấp nhất

Trong biểu đồ sau đây, chúng ta so sánh `tc` giữa hai nhóm nam và nữ:

```
> boxplot(tc ~ sex, main="Box plot of total cholestrol by
sex", ylab="mg/L")
```

Kết quả trình bày trong **Biểu đồ 14a**. Chúng ta có thể biến đổ giao diện của đồ thị bằng cách dùng thông số `horizontal=TRUE` và thay đổi màu bằng thông số `col` như sau (**Biểu đồ 14b)**:

```
> boxplot(tc~sex, horizontal=TRUE, main="Box plot of
total cholesterol", ylab="mg/L", col = "pink")
```



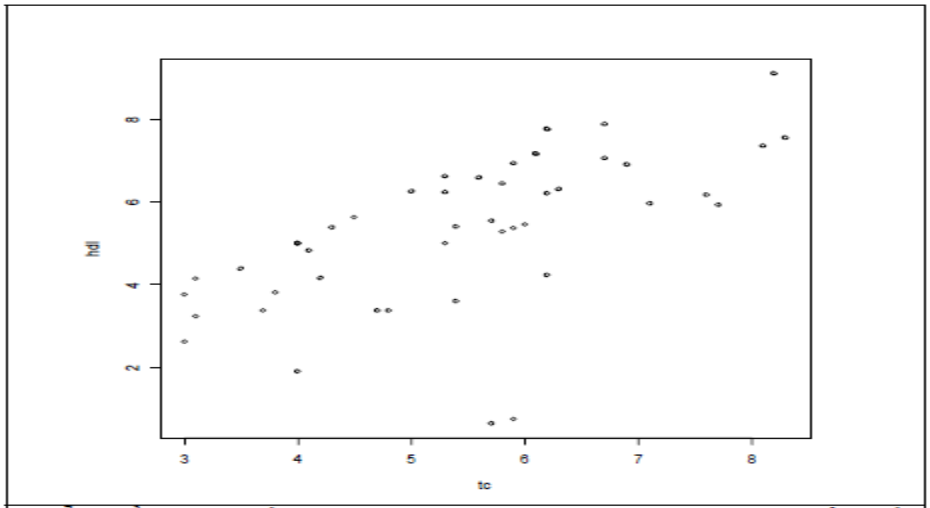
Biểu đồ 14a. Trong biểu đồ này, chúng ta thấy trung vị của total cholesterol ở nữ giới thấp hơn nam giới, nhưng độ dao động giữa hai nhóm không khác nhau bao nhiêu.

Biểu đồ 14b. Total cholesterol cho từng giới tính, với màu sắc và hình hộp nằm ngang.

8.7 Phân tích biểu đồ cho hai biến liên tục

8.7.1 Biểu đồ tán xạ (scatter plot)

Để tìm hiểu mối liên hệ giữa hai biến, chúng ta dùng biểu đồ tán xạ. Để vẽ biểu đồ tán xạ về mối liên hệ giữa biến số `tc` và `hdl`, chúng ta sử dụng hàm `plot`. Thông số thứ nhất của hàm plot là trục hoành (x-axis) và thông số thứ 2 là trục tung. Để tìm hiểu mối liên hệ giữa `tc` và `hdl` chúng ta đơn giản lệnh: > plot(tc,hdl)



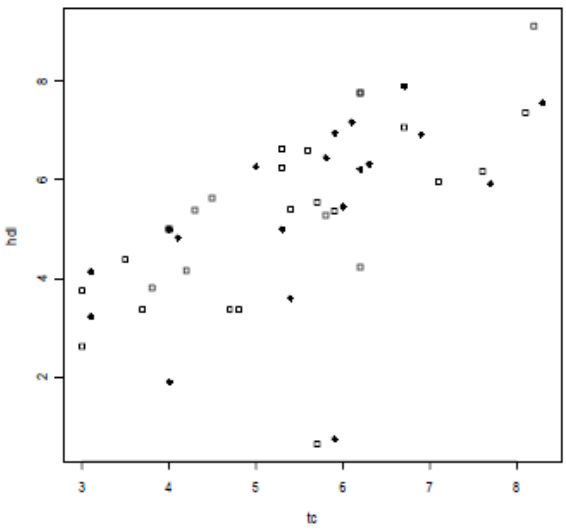
Biểu đồ 15. Mối liên hệ giữa `tc` và `hdl`. Trong biểu đồ này, chúng ta vẽ biến số `hdl` trên trục tung và `tc` trên trục hoành.

Chúng ta muốn phân biệt giới tính (nam và nữ) trong biểu đồ trên. Để vẽ biểu đồ đó, chúng ta phải dùng đến hàm ifelse. Trong lệnh sau đây, nếu sex==”Nam” thì vẽ kí tự số 16 (ô tròn), nếu không nam thì vẽ kí tự số 22 (tức ô vuông):

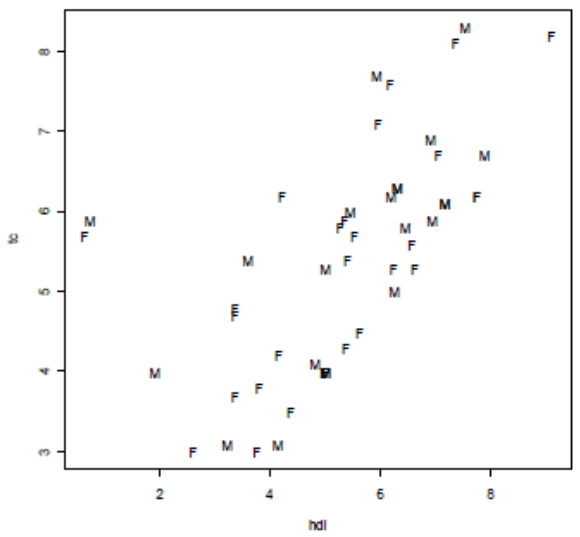
> plot(hdl, tc, pch=ifelse(sex=="Nam", 16, 22))

Kết quả là **Biểu đồ 16a**. Chúng ta cũng có thể thay kí tư thành “M” (nam) và “F” nữ(xem **Biểu đồ 16b**):

> plot(hdl, tc, pch=ifelse(sex=="Nam", “M”, “F”))



Biểu đồ 16a. Mối liên hệ giữa tc và hdl theo từng giới tính được thể hiện bằng hai kí hiệu dấu.



Biểu đồ 16a. Mối liên hệ giữa tc và hdl theo từng giới tính được thể hiện bằng hai kí tự.

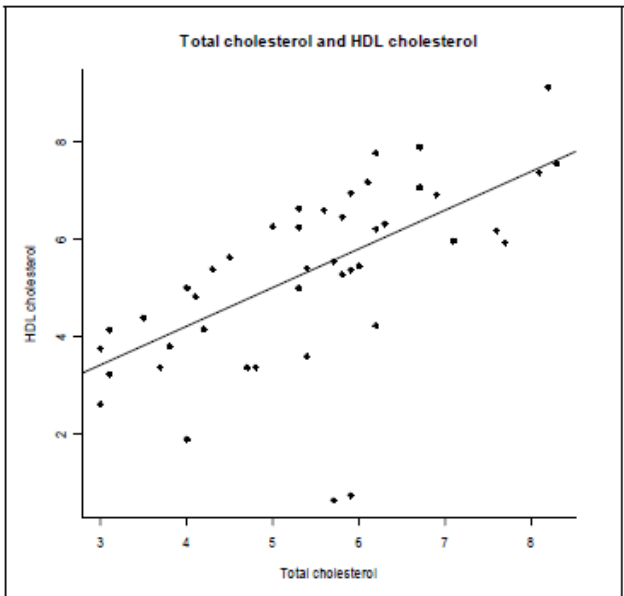
Chúng ta cũng có thể vẽ một đường biểu diễn hồi qui tuyến tính (regression line) qua các điểm trên bằng cách tiếp tục ra các lệnh sau đây:

```
> plot(hdl ~ tc, pch=16, main="Total cholesterol and HDL
cholesterol", xlab="Total cholesterol", ylab="HDL cholesterol",
bty=”l”)
> reg <- lm(hdl ~ tc)
> abline(reg)
```

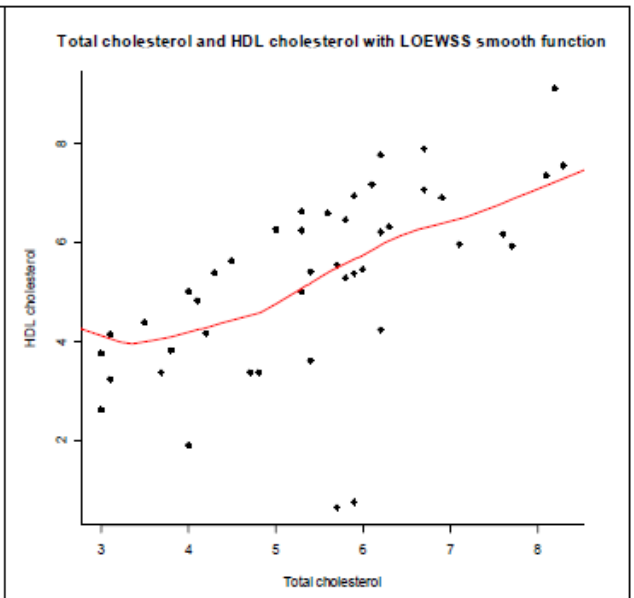
Kết quả là **Biểu đồ 17a** dưới đây. Chúng ta cũng có thể dùng hàm trơn (smooth function) để biểu diễn mối liên hệ giữa hai biến số. Đồ thị sau đây sử dụng `lowess` (một hàm thông thường nhất) trong việc “làm trơn” số liệu `tc` và `hdl` **(Biểu đồ 17b).**

```
>plot(hdl ~tc,pch=16,
       main="Total cholesterol and HDL cholesterol with LOEWSS
smooth function",
         xlab="Total cholesterol", ylab="HDL cholesterol",
                              bty="l")

> lines(lowess(hdl, tc, f=2/3, iter=3), col="red")
```



Biểu đồ 17a. Trong lệnh trên, `reg<-lm(hdl~tc)` có nghĩa là tìm phương trình liên hệ giữa `hdl` và `tc` bằng linear model (`lm`) và đ8ạt kết quả vào đối tượng `reg`. Lệnh thứ hai `abline(reg)` yêu cầu R vẽ đường thẳng từ phương trình trong `reg`



Biểu đồ 17b. Thay vì dùng abline, chúng ta dùng hàm lowess để thể hiện mối liên hệ giữa `tc` và `hdl`.

Bạn đọc có thể thí nghiệm với nhiều thông số `f=1/2, f=2/5`, hay thậm chí `f=1/10` sẽ thấy đồ thị biến đổi một cách "thú vị".

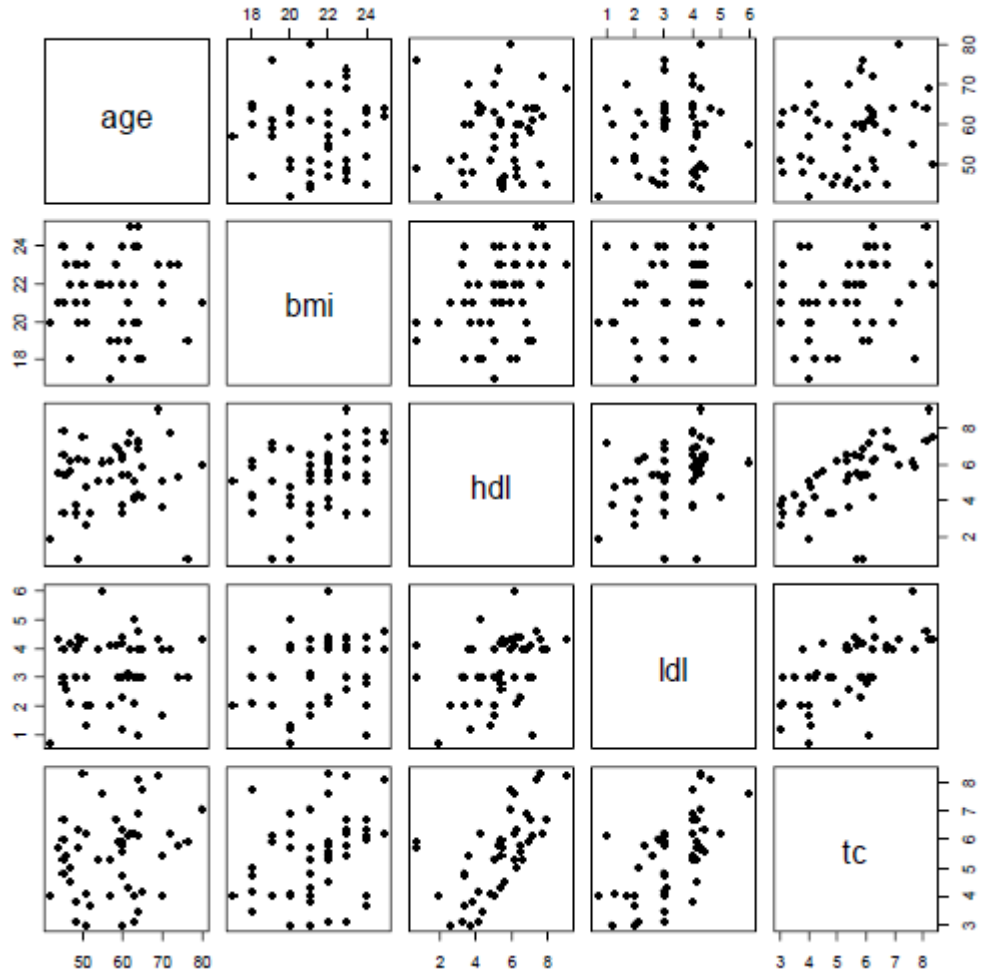
8.8 Phân tích Biểu đồ cho nhiều biến: pairs

Chúng ta có thể tìm hiểu mối liên hệ giữa các biến số như age, bmi, hdl, ldl và tc bằng cách dùng lệnh pairs. Nhưng trước hết, chúng ta phải đưa các biến số này vào một data.frame chỉ gồm những biến số có thể vẽ được, và sau đó sử dụng hàm pairs trong R.

> lipid <- data.frame(age,bmi,hdl,ldl,tc)

> pairs(lipid, pch=16)

Kết quả sẽ là:



8.9 Biểu đồ với sai số chuẩn (standard error)

Trong biểu đồ sau đây, chúng ta có 5 nhóm (biến số x được mô phỏng chứ không phải số liệu thật), và mỗi nhóm có giá trị trung bình mean, và độ tin cậy 95% (lcl và ucl).

Thông thường lcl=mean-1.96*SE và ucl = mean+1.96*SE (SE là sai số chuẩn). Chúng ta muốn vẽ biểu đồ cho 5 nhóm với sai số chuẩn đó. Các lệnh và hàm sau đây sẽ cần thiết:

```
> group <- c(1,2,3,4,5)
> mean <- c(1.1, 2.3, 3.0, 3.9, 5.1)
> lcl <- c(0.9, 1.8, 2.7, 3.8, 5.0)
  > ucl <- c(1.3, 2.4, 3.5, 4.1, 5.3)
> plot(group, mean, ylim=range(c(lcl, ucl)))
   > arrows(group, ucl, group, lcl, length=0.5, angle=90, code=3)
```