6
Tính toán xác suất và mô phỏng (simulation)

Xác suất là nền tảng của phân tích thống kê. Tất cả các phương pháp phân tích số liệu và suy luận thống kê đều dựa vào lí thuyết xác suất. Lí thuyết xác suất quan tâm đến việc mô tả và thể hiện qui luật phân phối của một biến số ngẫu nhiên. "Mô tả" ở đây trong thực tế cũng có nghĩa đơn giản là đếm những trường hợp hay khả năng xảy ra của một hay nhiều biến. Chẳng hạn như khi chúng ta chọn ngẫu nhiên 2 đối tượng , và nếu 2 đối tượng này có thể được phân loại bằng hai đặc tính như giới tính và sở thích, thì vấn đề đặt ra là có bao nhiêu tất cả "phối hợp" giữa hai đặc tính này. Hay đối với một biến số liên tục như huyết áp, mô tả có nghĩa là tính toán các chỉ số thống kê của biến như trị số trung bình, trung vị, phương sai, độ lệch chuẩn, v.v… Từ những chỉ số mô tả, lí thuyết xác suất cung cấp cho chúng ta những mô hình để thiết lập các hàm phân phối cho các biến số đó. Trong chương này, tôi sẽ bàn qua hai lĩnh vực chính là phép đếm và các hàm phân phối.

6.1 Các phép đếm

6.1.1 Phép hoán vị (permutation).

Theo định nghĩa, hoán vị *n* phần tử là cách sắp xếp *n* phần tử theo một thứ tự định sẵn. Định nghĩa này thật là … khó hiểu, chẳng khác gì … đố! Có lẽ một ví dụ cụ thể sẽ làm rõ định nghĩa hơn. Hãy tưởng tượng một trung tâm cấp cứu có 3 bác sĩ (*x, y* và *z*), và có 3 bệnh nhân (*a, b* và *c*) đang ngồi chờ được khám bệnh. Cả ba bác sĩ đều có thể khám bất cứ bệnh nhân *a, b* hay *c*. Câu hỏi đặt ra là có bao nhiêu cách sắp xếp bác sĩ – bệnh nhân? Để trả lời câu hỏi này, chúng ta xem xét vài trường hợp sau đây:

- Bác sĩ *x* có 3 lựa chọn: khám bệnh nhân *a, b* hoặc *c*;
- Khi bác sĩ *x* đã chọn một bệnh nhân rồi, thì bác sĩ *y* có hai lựa chọn còn lại;
- Và sau cùng, khi 2 bác sĩ kia đã chọn, bác sĩ *z* chỉ còn 1 lựa chọn.
- Tổng cộng, chúng ta có 6 lựa chọn.

Một ví dụ khác, trong một buổi tiệc gồm 6 bạn, hỏi có bao nhiêu cách sắp xếp cách ngồi trong một bàn với 6 ghế? Qua cách lí giải của ví dụ trên, đáp số là: 6.5.4.3.2.1 = 720 cách. (Chú ý dấu "." có nghĩa là dấu nhân hay tích số). Và đây chính là phép đếm *hoán vị*.

Chúng ta biết rằng 3! = 3.2.1 = 6, và 0!=1. Nói chung, công thức tính hoán vị cho một số *n* là: $n! = n(n-1)(n-2)(n-3)\times...\times 1$. 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
```

6.1.2 Tổ hợp (combination).

Tổ hợp *n* phần tử chập *k* là mọi tập hợp con gồm *k* phần tử của tập hợp *n* phần tử. Định nghĩa này phải nói là rất khó hiểu và … rườm rà! Cách dễ hiểu nhất là qua một ví dụ như sau: Cho 3 người (hãy cho là *A, B,* và *C*) ứng viên vào 2 chức chủ tịch và phó chủ tịch, hỏi: có bao nhiêu cách để chọn 2 chức này trong số 3 người đó. Chúng ta có thể tưởng tượng có 2 ghế mà phải chọn 3 người:

Cách chọn	Chủ tịch	Phó chủ tịch
1	*A*	*B*
2	*B*	*A*
3	*A*	*C*
4	*C*	*A*
5	*B*	*C*
6	*C*	*B*

Như vậy có 6 cách chọn. Nhưng chú ý rằng cách chọn 1 và 2 trong thực tế chỉ là 1 cặp, và chúng ta chỉ có thể đếm là 1 (chứ không 2 được). Tương tự, 3 và 4, 5 và 6 cũng chỉ có thể đếm là 1 cặp. Tổng cộng, chúng ta có 3 cách chọn 3 người cho 2 chức vụ. Đáp số này được gọi là tổ hợp.

Thật ra tổng số lần chọn có thể tính bằng công thức sau đây:

$$\binom{3}{2} = \frac{3!}{2!(3-2)!} = \frac{6}{2} = 3 \text{ lần.}$$

Nói chung, số lần chọn *k* người từ *n* người là:

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

Công thức này cũng có khi viết là C_k^n thay vì $\binom{n}{k}$. Với R, phép tính này rất đơn giản bằng hàm `choose(n, k)`. Sau đây là vài ví dụ minh họa:

- Tìm $\binom{5}{2}$

```
> choose(5, 2)
[1] 10
```

- Tìm xác suất cặp *A* và *B* trong số 5 người được đắc cử vào hai chức vụ:

```
> 1/choose(5, 2)
[1] 0.1
```

6.2 Biến số ngẫu nhiên và hàm phân phối

Phần lớn phân tích thống kê dựa vào các luật phân phối xác suất để suy luận. Hai chữ "phân phối" (distribution) có lẽ cũng cần vài dòng giải thích ở đây. Nếu chúng ta chọn ngẫu nhiên 10 bạn trong một lớp học và ghi nhận chiều cao và giới tính của 10 bạn đó, chúng ta có thể có một dãy số liệu như sau:

	1	2	3	4	5	6	7	8	9	10
Giới tính	Nữ	Nữ	Nam	Nữ	Nữ	Nữ	Nam	Nam	Nữ	Nam
Chiều cao (cm)	156	160	175	145	165	158	170	167	178	155

Nếu tính gộp chung lại, chúng ta có 6 bạn gái và 4 bạn trai. Nói theo phần trăm, chúng ta có 60% nữ và 40% nam. Nói theo ngôn ngữ xác suất, xác suất nữ là 0.6 và nam là 0.4.

Về chiều cao, chúng ta có giá trị trung bình là 162.9 cm, với chiều cao thấp nhất là 155 cm và cao nhất là 178 cm.

Nói theo ngôn ngữ thống kê xác suất, biến số giới tính và chiều cao là hai *biến số ngẫu nhiên* (random variable). Ngẫu nhiên là vì chúng ta không đoán trước một cách chính xác các giá trị này, nhưng chỉ có thể đoán giá trị tập trung, giá trị trung bình, và độ dao động của chúng. Biến giới tính chỉ có hai "giá trị" (nam hay nữ), và được gọi là biến *không liên tục*, hay *biến rời rạc* (discrete variable), hay *biến thứ bậc* (categorical variable). Còn biến chiều cao có thể có bất cứ giá trị nào từ thấp đến cao, và do đó có tên là *biến liên tục* (continuous variable).

Khi nói đến "phân phối" (hay distribution) là đề cập đến các giá trị mà biến số có thể có. Các *hàm phân phối* (distribution function) là hàm nhằm mô tả các biến số đó một cách có hệ thống. "Có hệ thống" ở đây có nghĩa là theo mộ mô hình toán học cụ thể với những thông số cho trước. Trong xác suất thống kê có khá nhiều hàm phân phối, và ở đây chúng ta sẽ xem 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 sẵn trên 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:

Hàm phân phối	**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)`
Exponential	`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.

6.3 Các hàm phân phối xác suất (probability distribution function)

6.3.1 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, . . . , n. Để hiểu định lí đó rõ ràng hơn, chúng ta sẽ xem qua qua vài ví dụ sau đây.

Ví dụ 1: Hàm mật độ nhị phân (Binomial density probability function). Trong ví dụ trên, lớp học có 10 người, trong đó có 6 nữa. Nếu 3 bạn được chọn một cách ngẫu nhiên, xác suất mà chúng ta có 2 bạn nữ là bao nhiêu? Chúng ta có thể trả lời câu hỏi này một cách tương đối thủ công bằng cách xem xét tất cả các trường hợp có thể xảy ra. Mỗi lần chọn có 2 khả khăng (nam hay nữ), và 3 lần chọn, chúng ta có $2^3 = 8$ trường hợp như sau.

Bạn 1	**Bạn 2**	**Bạn 3**	**Xác suất**
Nam	Nam	Nam	(0.4)(0.4)(0.4) = 0.064
Nam	Nam	Nữ	(0.4)(0.4)(0.6) = 0.096
Nam	Nữ	Nam	(0.4)(0.6)(0.4) = 0.096
Nam	Nữ	Nữ	(0.4)(0.6)(0.6) = 0.144
Nữ	Nam	Nam	(0.6)(0.4)(0.4) = 0.096
Nữ	Nam	Nữ	(0.6)(0.4)(0.6) = 0.144
Nữ	Nữ	Nam	(0.6)(0.6)(0.4) = 0.144
Nữ	Nữ	Nữ	(0.6)(0.6)(0.6) = 0.216
Tất cả các trường hợp			1.000

Chứng ta biết trước rằng trong nhóm 10 học sinh có 6 nữ, và do đó, xác suất nữ là 0.60. (Nói cách khác, xác suất chọn một bạn nam là 0.4). Do đó, xác suất mà tất cả 3 bạn được chọn đều là nam giới là: 0.4 x 0.4 x 0.4 = 0.064. Trong bảng trên, chúng ta thấy có 3 trường hợp mà trong đó có 2 bạn gái: đó là trường hợp Nam-Nữ-Nữ, Nữ-Nữ-Nam, và Nữ-Nam-Nữ, cả 3 đều có xác suất 0.144. Thành ra, xác suất chọn đúng 2 bạn nữ trong số 3 bạn được chọn là 3x0.144= 0.432.

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:

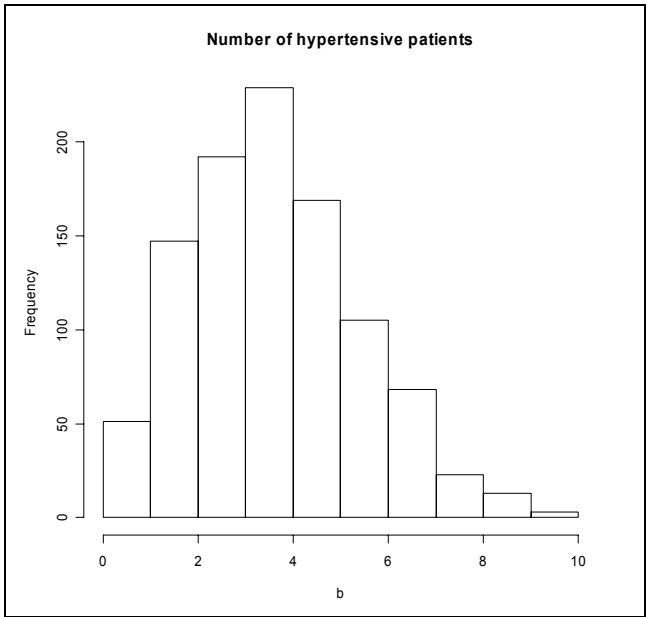
```
> b <- rbinom(1000, 20, 0.20)
```

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")
```



Biểu đồ 1. Phân phối số bệnh nhân cao huyết áp trong số 20 người được chọn ngẫu nhiên trong một quần thể gồm 20% bệnh nhân cao

huyết áp, và chọn mẫu được lặp lại 1000 lần.

Qua biểu đồ trên, chúng ta thấy xác suất có 4 bệnh nhân cao huyết áp (trong mỗi lần chọn mẫu 20 người) là cao nhất (22.9%). Điều này cũng có thể hiểu được, bởi vì tỉ lệ cao huyết áp là 20%, cho nên chúng ta kì vọng rằng trung bình 4 người trong số 20 người được chọn phải là cao huyết áp. Tuy nhiên, điều quan trọng mà biểu đồ trên thể hiện là có khi chúng ta quan sát đến 10 bệnh nhân cao huyết áp dù xác suất cho mẫu này rất thấp (chỉ 3/1000).

Ví dụ 4: Ứng dụng hàm phân phối nhị phân: Hai mươi khách hàng được mời uống hai loại bia A và B, và được hỏi họ thích bia nào. Kết quả cho thấy 16 người thích bia A. Vấn đề đặt ra là kết quả này có đủ để kết luận rằng bia A được nhiều người thích hơn bia B, hay là kết quả chỉ là do các yếu tố ngẫu nhiên gây nên?

Chúng ta bắt đầu giải quyết vấn đề bằng cách giả thiết rằng nếu không có khác nhau, thì xác suất p=0.50 thích bia A và q=0.5 thích bia B. Nếu giả thiết này đúng, thì xác suất mà chúng ta quan sát 16 người trong số 20 người thích bia A là bao nhiêu. Chúng ta có thể tính xác suất này bằng R rất đơn giản:

```
> 1- pbinom(15, 20, 0.5)
[1] 0.005908966
```

Đáp số là xác suất 0.005 hay 0.5%. Nói cách khác, nếu quả thật hai bia giống nhau thì xác suất mà 16/20 người thích bia A chỉ 0.5%. Tức là, chúng ta có bằng chứng cho thấy khả năng bia A quả thật được nhiều người thích hơn bia B, chứ không phải do yếu tố ngẫu nhiên. Chú ý, chúng ta dùng 15 (thay vì 16), là bởi vì $P(X \geq 16) = 1 - P(X \leq 15)$. Mà trong trường hợp ta đang bàn, $P(X \leq 15) =$ `pbinom(15, 20, 0.5)`.

6.3.2 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 5: **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, :

$$P(X = k \mid \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Do đó, đáp số cho câu hỏi trên là: $P(X = 2 \mid \lambda = 1) = \frac{e^{-2}1^2}{2!} = 0.1839$. Đáp số này có thể 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
```

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:

$$\begin{aligned} P(X > 2) &= P(X = 3) + P(X = 4) + P(X = 5) + \ldots \\ &= 1 - P(X \le 2) \\ &= 1 - 0.3678 - 0.3678 - 0.1839 \\ &= 0.08 \end{aligned}$$

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
```

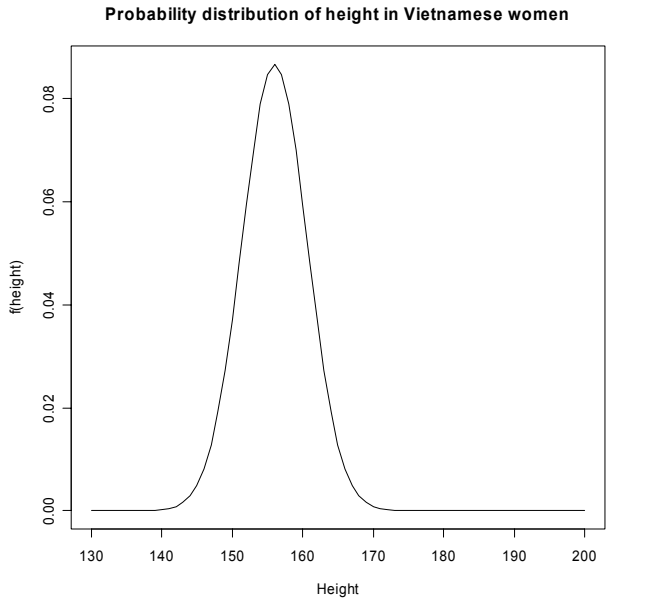
6.3.3 Hàm phân phối chuẩn (Normal distribution)

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 không ngoa rằng hầu hết lí thuyết thống kê được xây dựng trên nền tảng của phân phối chuẩn. Hàm mật độ phân phối chuẩn có hai thông số: trung bình μ và phương sai σ^2 (hay độ lệch chuẩn σ). Gọi X là một biến số (như chiều cao chẳng hạn), hàm mật độ phân phối chuẩn phát biểu rằng xác suất mà $X = x$ là:

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

Ví dụ 6: **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. Cũng biết rằng chiều cao này tuân theo luật phân phối chuẩn. Với hai thông số μ=156, σ=4.6, chúng ta có thể xây dựng một hàm phân phối chiều cao cho toàn bộ quần thể phụ nữ Việt Nam, và hàm này có hình dạng như sau:



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 chiều 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 $P(160 \le X \le 150)$, 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.

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

Chẳng hạn như xác suất chiều cao phụ nữ Việt Nam bằng hoặc thấp hơn 150 cm là 9.6%:

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

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 7: Ứng dụng luật phân phối chuẩn: Trong một quần thể, chúng ta biết rằng áp suất máu trung bình là 100 mmHg và độ lệch chuẩn là 13 mmHg, hỏi: có bao nhiêu người 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.

6.3.4 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:

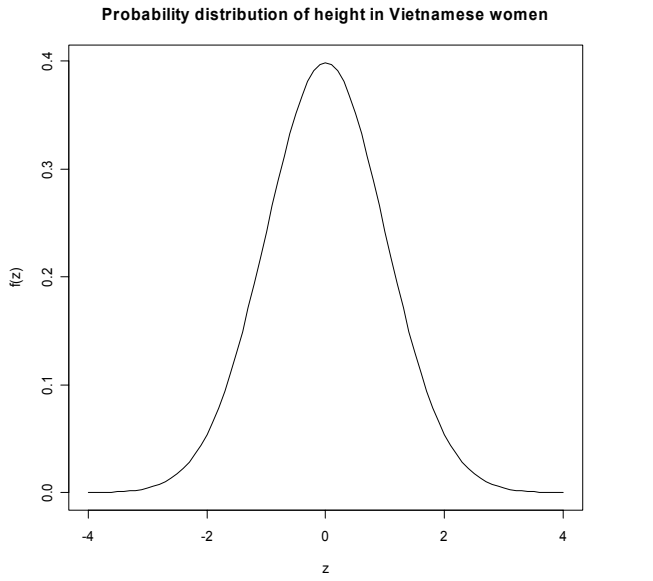
$$X \sim N(\mu, \sigma^2)$$

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:

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

Nói theo ngôn ngữ toán: nếu $X \sim N(\mu, \sigma^2)$, thì $(X - \mu)/\sigma^2 \sim N(0, 1)$. Như vậy qua công thức trên, Z thực chất là độ khác biệt giữa một số và trung bình tính bằng số độ lệch chuẩn. Nếu $Z = 0$, chúng ta biết rằng X bằng số trung bình μ. Nếu $Z = -1$, chúng ta biết rằng X thấp hơn μ đúng 1 độ lệch chuẩn. Tương tự, $Z = 2.5$, chúng ta biết rằng X cao hơn μ đúng 2.5 độ lệch chuẩn. v.v…

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 tôi 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 6 (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 8: 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
```

6.3.5 Hàm phân phối t, F và χ^2

Các hàm phân phối t, F và χ^2 trong thực tế là hàm của hàm phân phối chuẩn. Mối liên hệ và cách tính các hàm này có thể được mô tả bằng vài ghi chú sau đây:

- **Phân phối Chi bình phương (χ^2)**. Phân phối χ^2 xuất phát từ tổng bình phương của một biến phân phối chuẩn. Nếu nếu $x_i \sim N(0, 1)$, và gọi $u = \sum_{i=}^{n} x_i^2$, thì u tuân theo luật phân phối Chi bình phương với bậc tự do n (thường viết tắt là `df`). Nói theo ngôn ngữ toán, $u \sim \chi_n^2$.

Ví dụ 9: Tìm xác suất của một biến Chi bình phương, do đó, chỉ cần hai thông số *u* và *n*. Chẳng hạn như nếu chúng ta muốn tìm xác suất P(u=21, df=13), chỉ đơn giản dùng hàm `pchisq` như sau:

```
> dchisq(21, 13)
[1] 0.01977879
```

Tìm xác suất mà một biến số *u* nhỏ hơn 21 với bậc tự do 13 df. Tức là tìm P($u \leq$ 21 | df=13) = ?

```
> pchisq(21, 13)
[1] 0.9270714
```

Cũng có thể nói kết quả trên cho biết P($\chi^2_{13} < 21$) = 0.927.

Tìm quantile của một trị số *u* tương đương với 90% của một phân phối χ^2 với 15 bậc tự do:

```
> qchisq(0.95, 15)
[1] 24.99579
```

Nói cách khác, P($\chi^2_{15} < 24.99$) = 0.95.

Phi trung tâm (Non-centrality). Chú ý trong định nghĩa trên, phân phối χ^2 xuất phát từ tổng bình phương của một biến phân phối chuẩn có trung bình 0 và phương sai 1. Nhưng nếu một biến phân phối chuẩn có trung bình không phải là 0 và phương sai không phải là 1, thì chúng ta sẽ có một **phân phối Chi bình phương phi trung tâm**. Nếu $x_i \sim N(\mu_i, 1)$ và đặt $u = \sum_{i=1}^{n} x_i^2$, thì *u* tuân theo luật phân phối Chi bình phương phi trung tâm với bậc tự do *n* và thông số phi trung tâm (non-centrality parameter) λ như sau:

$$\lambda = \sum_{i=1}^{n} \mu_i^2$$

Và kí hiệu là $u \sim \chi^2_{n,\lambda}$. Có thể nói thêm rằng, trung bình của *u* là $n+\lambda$, và phương sai của *u* là $2(n+2\lambda)$.

Tìm xác suất mà *u* nhỏ hơn hoặc bằng 21, với điều kiện bậc tự do là 13 và thông số non-centrality bằng 5.4:

```
> pchisq(21, 13, 5.4)
[1] 0.6837649
```

Tức là, P($\chi^2_{13,5.4} < 21$) = 0.684.

Tìm quantile của một trị số tương đương với 50% của một phân phối χ^2 với 7 bậc tự do và thông số non-centrality bằng 3.

```
> qchisq(0.5, 7, 3)
[1] 9.180148
```

Do đó, P($\chi^2_{7,3}$ < 9.180148) = 0.50

- **Phân phối t (t distribution)**. Chúng ta vừa biết rằng nếu $X \sim N(\mu, s^2)$ thì thì$(X - \mu)/\sigma^2 \sim N(0, 1)$. Nhưng phát biểu đó đúng (hay chính xác) khi chúng ta biết phương sai σ^2. Trong thực tế, ít khi nào chúng ta biết chính xác phương sai, mà chỉ ước tính từ số liệu thực nghiệm. Trong trường hợp phương sai được ước tính từ số liệu nghiên cứu, và hãy gọi ước tính này là s^2, thì chúng ta có thể phát biểu rằng: $(X - \mu)/s^2 \sim t(0, v)$, trong đó v là bậc tự do.

  Ví dụ 10. Tìm xác suất mà x lớn hơn 1, trong biến theo luật phân phối t với 6 bậc tự do:

```
> 1-pt(1.1, 6)
[1] 0.1567481
```

  Tức là, $P(t_6 > 1.1) = 1 - P(t_6 < 1.1) = 0.157$.

  Tìm định lượng của một trị số tương đương với 95% của một phân phối t với 15 bậc tự do:

```
> qt(0.95, 15)
[1] 1.753050
```

  Nói cách khác, $P(t_{19} < 1.75035) = 0.95$.

- **Phân phối *F***. Tỉ số giữa hai biến số theo luật phân phối χ^2 có thể chứng minh là tuân theo luật phân phối F. Nói cách khác, nếu $u \sim \chi^2_n$ và $v \sim \chi^2_m$, thì $u/v \sim F_{n,m}$, trong đó n là bặc tự do tử số (numerator degrees of freedom) và m là bặc tự do mẫu số (denominator degrees of freedom).

  Ví dụ 11: Tìm xác suất mà một trị số F lớn hơn 3.24, biết rằng biến số đó tuân theo luật phân phối F với bậc tự do 3 và 15 df và thông số non-centrality 5:

```
> 1-pf(3.24, 3, 15, 5)
[1] 0.3558721
```

  Do đó, $P(F_{3, 15, 5} > 3.24) = 1 - P(F_{3, 15,5} \geq 3.24) = 0.355338$.

  Với bậc tự do 3 và 15, tìm C sao cho $P(F_{3, 15} > C) = 0.05$. Lời giải của R là:

```
> qf(1-0.05, 3, 15)
[1] 3.287382
```

Nói cách khác, $P(F_{3, 15} > 3.287382) = 1 - P(F_{3, 15} \geq 3.287382) = 1 - 0.95 = 0.05$

6.4 Mô phỏng (simulation)

Trong phân tích thống kê, đôi khi vì hạn chế số mẫu chúng ta khó có thể ước tính một cách chính xác các thông số, và trong trường hợp bất định đó, chúng ta cần đến mô phỏng để biết được độ dao động của một hay nhiều thông số. Mô phỏng thường dựa vào các luật phân phối. Đây là một lĩnh vực khá phức tạp mà tôi không có ý định trình bày đầy đủ trong chương này. Ở đây, tôi chỉ trình bày một số mô hình mô phỏng mang tính minh họa để bạn đọc có thể dựa vào đó mà phát triển thêm.

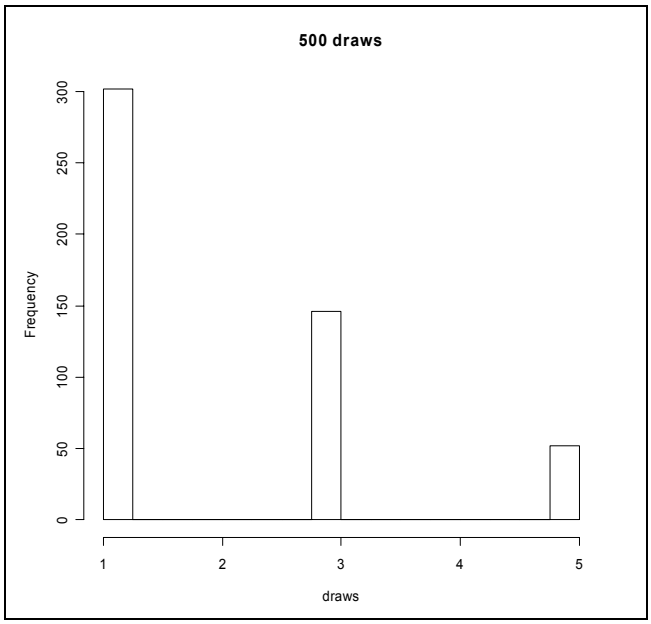
Ví dụ 11: Mô phỏng để chứng minh phương sai của số trung bình bằng phương sai chia cho *n* ($\text{var}\left(\bar{X}\right) = \sigma^2 / n$). Chúng ta sẽ xem một biến số không liên tục với giá trị 1, 3 và 5 với xác suất như sau:

x	P(*x*)
1	0.60
3	0.30
5	0.10

Qua số liệu này, chúng ta biết rằng giá trị trung bình là (1x0.60)+(3x0.30)+(5x0.10) = 2.0 và phương sai (bạn đọc có thể tự tính) là 1.8.

Bây giờ chúng ta sử dụng hai thông số này để thử mô phỏng 500 lần. Lệnh thứ nhất tạo ra 3 giá trị của *x*. Lệnh thứ hai nhập số xác suất cho từng giá trị của *x*. Lệnh `sample` yêu cầu R tạo nên 500 số ngẫu nhiên và cho vào đối tượng `draws`.

```
x <- c(1, 3, 5)
px <- c(0.6, 0.3, 0.1)
draws <- sample(x, size=500, replace=T, prob=px)
hist(draws, breaks=seq(1,5, by=0.25), main="1000 draws")
```



Từ luật phân phối xác suất chúng ta biết rằng tính trung bình sẽ có 60% lần có giá trị "1", 30% có giá trị "2", và 10% có giá trị "5". Do đó, chúng ta kì vọng sẽ quan sát 300, 150 và 50 lần cho mỗi giá trị. Biểu đồ trên cho thấy phân phối các giá trị này gần với giá trị mà chúng ta kì vọng. Ngoài ra, chúng ta cũng biết rằng phương sai của biến số này là khoảng 1.8. Bây giờ chúng ta kiểm tra xem có đúng như kì vọng hay không:

```
> var(draws)
[1] 1.835671
```

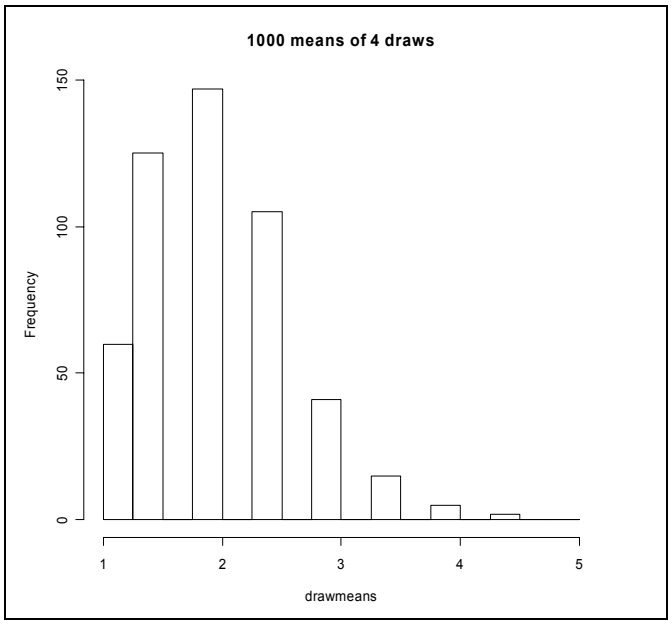
Kết quả trên cho thấy phương sai của 500 mẫu là 1.836, tức không xa mấy so với giá trị kì vọng.

Bây giờ chúng ta thử mô phỏng 500 giá trị trung bình $\overline{x}$ ($\overline{x}$ là số trung bình của 4 số liệu mô phỏng) từ quần thể trên:

```
> draws <- sample(x, size=4*500, replace=T, prob=px)
> draws = matrix(draws, 4)
> drawmeans = apply(draws, 2, mean)
```

Lệnh thứ nhất và thứ hai tạo nên đối tượng tên là `draws` với 4 dòng, mỗi dòng có 500 giá trị từ luật phân phối trên. Nói cách khác, chúng ta có 4*500 = 2000 số. 500 số cũng có nghĩa là 500 cột: 1 đến 500. Tức mỗi cột có 4 số. Lệnh thứ ba tìm trị số trung bình cho mỗi cột. Lệnh này sẽ cho ra 500 số trung bình và chứa trong đối tượng `drawmeans`. Biểu đồ sau đây cho thấy phân phối của 500 số trung bình:

```
> hist(drawmeans,breaks=seq(1,5,by=0.25), main="1000 means of 4 draws")
```



Chúng ta thấy rằng phương sai của phân phối này nhỏ hơn. Thật ra, phương sai của 500 số trung bình này là 0.45.

```
> var(drawmeans)
 [1] 0.4501112
```

Đây là giá trị tương đương với giá trị 0.45 mà chúng ta kì vọng từ công thức $\text{var}\left(\bar{X}\right) = \sigma^2 / 4 = 1.8/4 = 0.45$.

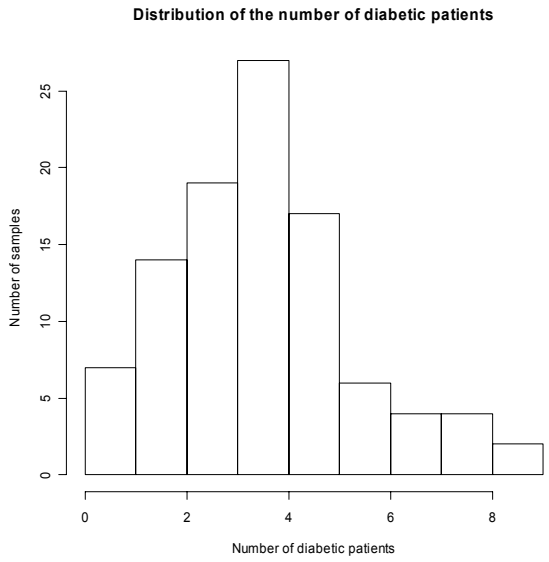
6.4.1 Mô phỏng phân phối nhị phân

Ví dụ 12: Mô phỏng mẫu từ một quần thể với luật phân phối nhị phân. Giả dụ chúng ta biết một quần thể có 20% người bị bệnh đái đường (xác suất p=0.2). Chúng ta muốn lấy mẫu từ quần thể này, mỗi mẫu có 20 đối tượng, và phương án chọn mẫu được lặp lại 100 lần:

```
> bin <- rbinom(100, 20, 0.2)
> bin
  [1] 4 4 5 3 2 2 3 2 5 4 3 6 7 3 4 4 1 5 3 5 3 4 4 5 1 4 4 4 4 3 2 4 2 2 5 4 5
 [38] 7 3 5 3 3 4 3 2 4 5 2 4 5 5 4 2 2 2 8 5 5 5 3 4 5 7 4 3 6 4 6 6 8 8 3 3 1
 [75] 1 4 4 2 3 9 7 4 4 0 0 8 6 9 3 1 4 5 6 4 5 3 2 4 3 2
```

Kết quả trên là số lần đầu, chúng ta sẽ có 4 người mắc bệnh; lần 2 cũng 4 người; lần 3 có 5 người mắc bệnh; v.v… kết quả này có thể tóm lược trong một biểu đồ như sau:

```
> hist(bin,
       xlab="Number of diabetic patients",
       ylab="Number of samples",
       main="Distribution of the number of diabetic patients")
```




```
> mean(bin)
[1] 3.97
```

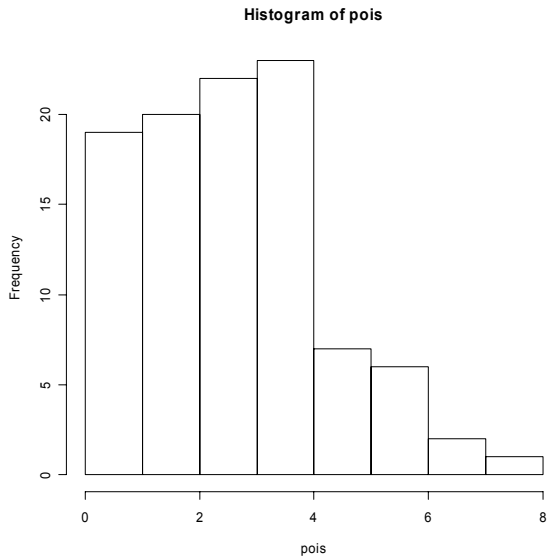
Đúng như chúng ta kì vọng, vì chọn mỗi lần 20 đối tượng và xác suất 20%, nên chúng ta tiên đoán trung bình sẽ có 4 bệnh nhân đái đường.

6.4.2 Mô phỏng phân phối Poisson

Ví dụ 13: Mô phỏng mẫu từ một quần thể với luật phân phối Poisson. Trong ví dụ sau đây, chúng ta mô phỏng 100 mẫu từ một quần thể tuân theo luật phân phối Poisson với trung bình λ=3:

```
> pois <- rpois(100, lambda=3)
> pois
> pois
  [1] 4 3 2 4 2 3 4 4 0 7 5 0 3 3 4 2 2 6 1 4 2 3 3 5 4 2 1 4 0 2 1 5 1 2 2 2 6
 [38] 1 3 6 3 3 5 4 3 2 2 5 3 3 3 1 4 7 3 4 3 2 6 1 4 1 0 5 2 2 2 3 6 8 4 4 1 4
 [75] 1 0 0 4 3 3 2 3 3 3 4 1 5 4 4 1 3 1 6 4 4 4 2 2 2 4
```

Và mật độ phân phối:



Phân phối Poisson và phân phối mũ. Trong ví dụ sau đây, chúng ta mô phỏng thời gian bệnh nhân đến một bệnh viện. Biết rằng bệnh nhân đến bệnh viện một cách ngẫu nhiên theo luật phân phối Poisson, với trung bình 15 bệnh nhân cho mỗi 150 phút. Có thể chứng minh dễ dàng rằng thời gian giữa hai bệnh nhân đến bệnh viện tuân theo luật phân phối mũ. Chúng ta muốn biết thời gian mà bệnh nhân ghé bệnh viện; do đó, chúng ta mô phỏng 15 thời gian giữa hai bệnh nhân từ luật phân phối mũ với tỉ lệ 15/150 = 0.1 mỗi phút. Các lệnh sau đây đáp ứng yêu cầu đó:

```
# Tạo thời gian đến bệnh viện
> appoint <- rexp(15, 0.1)
> times <- round(appoint,0)
> times
 [1] 37  5  8 10 24  5  1  7  8  6 12  6  3 25 15
```

6.4.3 Mô phỏng phân phối χ^2, t, F

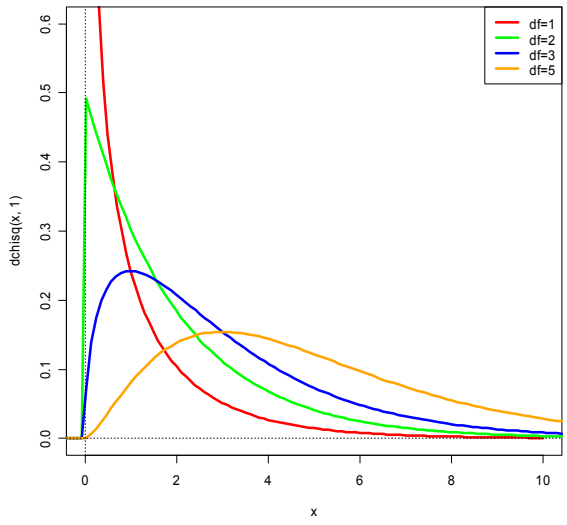
Cách mô phỏng trên đây còn có thể áp dụng cho các luật phân phối khác như nhị phân âm (negative binomial distribution với `rnbinom`), gamma (`rgamma`), beta (`rbeta`), Chi bình phương (`rchisq`), hàm mũ (`rexp`), t (`rt`), F (`rf`), v.v… Các thông số cho các hàm mô phỏng này có thể tìm trong phần đầu của chương.

Các lệnh sau đây sẽ minh họa các luật phân phối thông thường đó:

- Phân phối Chi bình phương với một số bậc tự do:

```
> curve(dchisq(x, 1), xlim=c(0,10), ylim=c(0,0.6), col="red", lwd=3)
> curve(dchisq(x, 2), add=T, col="green", lwd=3)
> curve(dchisq(x, 3), add=T, col="blue", lwd=3)
> curve(dchisq(x, 5), add=T, col="orange", lwd=3)
> abline(h=0, lty=3)
```

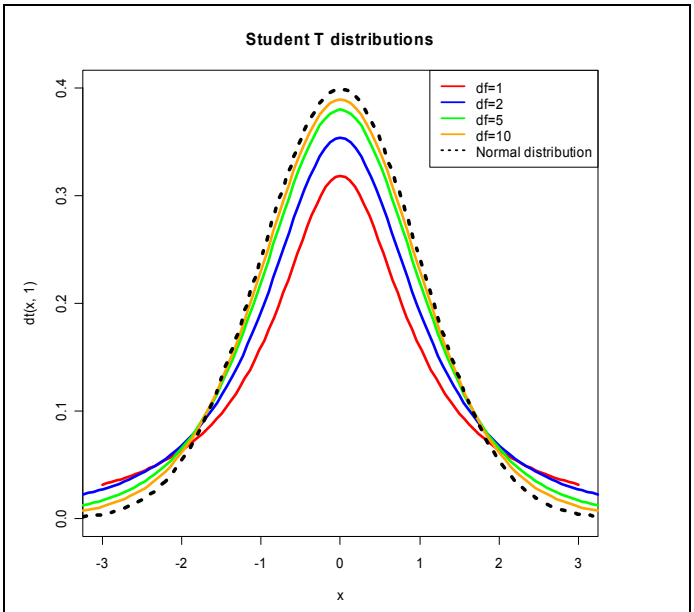
```
> abline(v=0, lty=3)
> legend(par("usr")[2], par("usr")[4],
         xjust=1,
         c("df=1", "df=2", "df=3", "df=5"), lwd=3, lty=1,
         col=c("red", "green", "blue", "orange"))
```



Biểu đồ 4. Phân phối Chi bình phương với bậc tự do =1, 2, 3, 5.

- Phân phối *t*:

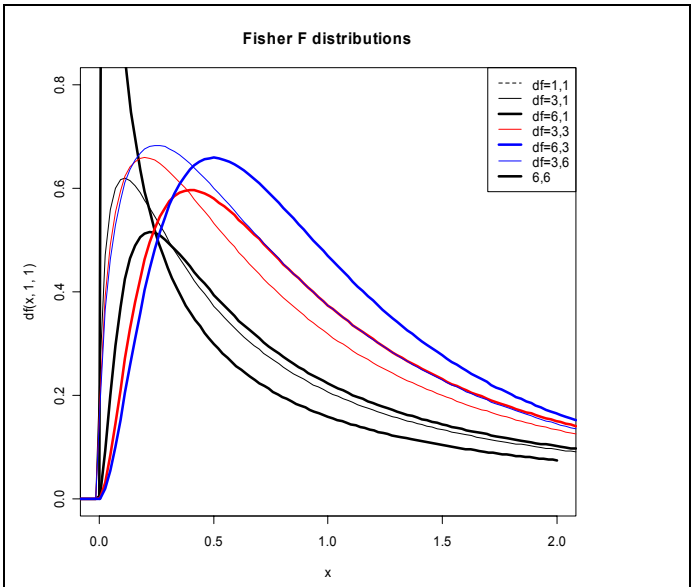
```
> curve(dt(x, 1), xlim=c(-3,3), ylim=c(0,0.4), col="red", lwd=3)
> curve(dt(x, 2), add=T, col="blue", lwd=3)
> curve(dt(x, 5), add=T, col="green", lwd=3)
> curve(dt(x, 10), add=T, col="orange", lwd=3)
> curve(dnorm(x), add=T, lwd=4, lty=3)
> title(main="Student T distributions")
> legend(par("usr")[2], par("usr")[4],
         xjust=1,
         c("df=1", "df=2", "df=5", "df=10", "Normal distribution"),
         lwd=c(2,2,2,2,2),
         lty=c(1,1,1,1,3),
         col=c("red", "blue", "green", "orange", par("fg")))
```



Biểu đồ 5. Phân phối t với bậc tự do =1, 2, 5, 10 so sánh với phân phối chuẩn.

- Phân phối *F*:

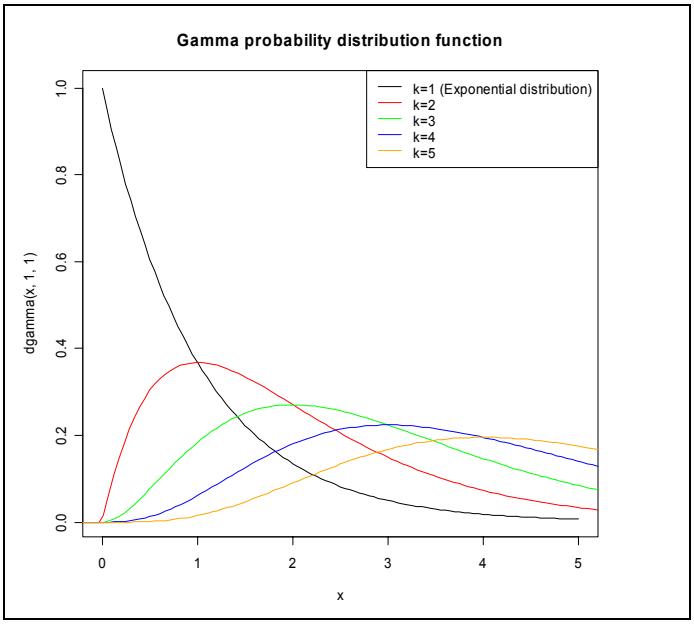
```
> curve(df(x,1,1), xlim=c(0,2), ylim=c(0,0.8), lwd=3)
> curve(df(x,3,1), add=T)
> curve(df(x,6,1), add=T, lwd=3)
> curve(df(x,3,3), add=T, col="red")
> curve(df(x,6,3), add=T, col="red", lwd=3)
> curve(df(x,3,6), add=T, col="blue")
> curve(df(x,6,6), add=T, col="blue", lwd=3)
> title(main="Fisher F distributions")
> legend(par("usr")[2], par("usr")[4],
         xjust=1,
         c("df=1,1", "df=3,1", "df=6,1", "df=3,3", "df=6,3",
           "df=3,6", df="6,6"),
         lwd=c(1,1,3,1,3,1,3),
         lty=c(2,1,1,1,1,1,1),
         col=c(par("fg"), par("fg"), par("fg"), "red", "blue", "blue"))
```



Biểu đồ 6. Phân phối F với nhiều bậc tự do khác nhau.

- Phân phối *gamma*:

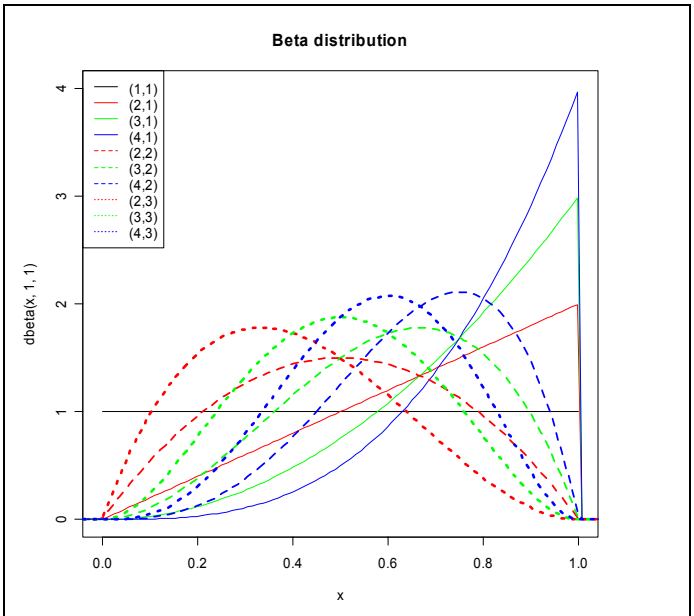
```
> curve( dgamma(x,1,1), xlim=c(0,5) )
> curve( dgamma(x,2,1), add=T, col='red' )
> curve( dgamma(x,3,1), add=T, col='green' )
> curve( dgamma(x,4,1), add=T, col='blue' )
> curve( dgamma(x,5,1), add=T, col='orange' )
> title(main="Gamma probability distribution function")
> legend(par('usr')[2], par('usr')[4], xjust=1,
       c('k=1 (Exponential distribution)', 'k=2', 'k=3', 'k=4', 'k=5'),
       lwd=1, lty=1,
       col=c(par('fg'), 'red', 'green', 'blue', 'orange') )
```



Biểu đồ 7. Phân phối Gamma với nhiều hình dạng.

- Phân phối *beta*:

```
> curve( dbeta(x,1,1), xlim=c(0,1), ylim=c(0,4) )
> curve( dbeta(x,2,1), add=T, col='red' )
> curve( dbeta(x,3,1), add=T, col='green' )
> curve( dbeta(x,4,1), add=T, col='blue' )
> curve( dbeta(x,2,2), add=T, lty=2, lwd=2, col='red' )
> curve( dbeta(x,3,2), add=T, lty=2, lwd=2, col='green' )
> curve( dbeta(x,4,2), add=T, lty=2, lwd=2, col='blue' )
> curve( dbeta(x,2,3), add=T, lty=3, lwd=3, col='red' )
> curve( dbeta(x,3,3), add=T, lty=3, lwd=3, col='green' )
> curve( dbeta(x,4,3), add=T, lty=3, lwd=3, col='blue' )
> title(main="Beta distribution")
> legend(par('usr')[1], par('usr')[4], xjust=0,
       c('(1,1)', '(2,1)', '(3,1)', '(4,1)',
         '(2,2)', '(3,2)', '(4,2)',
         '(2,3)', '(3,3)', '(4,3)' ),
       lwd=1, #c(1,1,1,1, 2,2,2, 3,3,3),
       lty=c(1,1,1,1, 2,2,2, 3,3,3),
       col=c(par('fg'), 'red', 'green', 'blue',
             'red', 'green', 'blue',
             'red', 'green', 'blue' ))
```

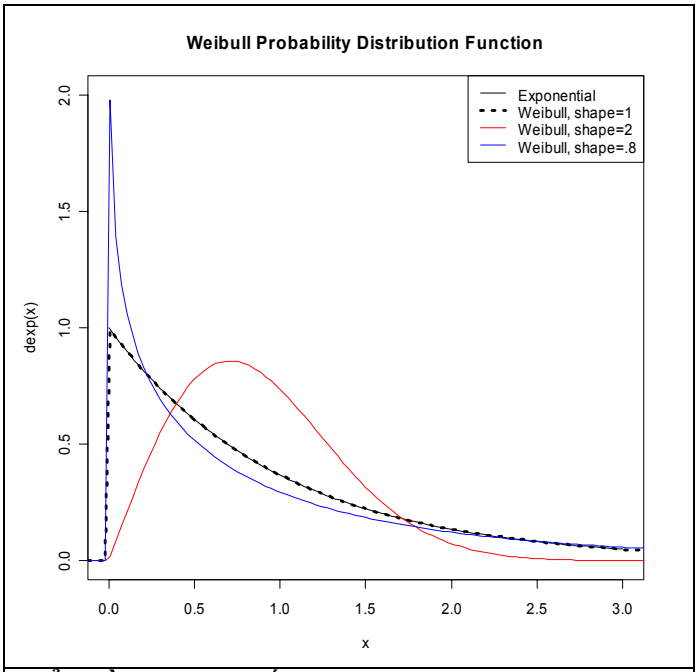


Biểu đồ 8. Phân phối beta với nhiều hình dạng.

- Phân phối *Weibull*:

```
> curve(dexp(x), xlim=c(0,3), ylim=c(0,2))
> curve(dweibull(x,1), lty=3, lwd=3, add=T)
> curve(dweibull(x,2), col='red', add=T)
```

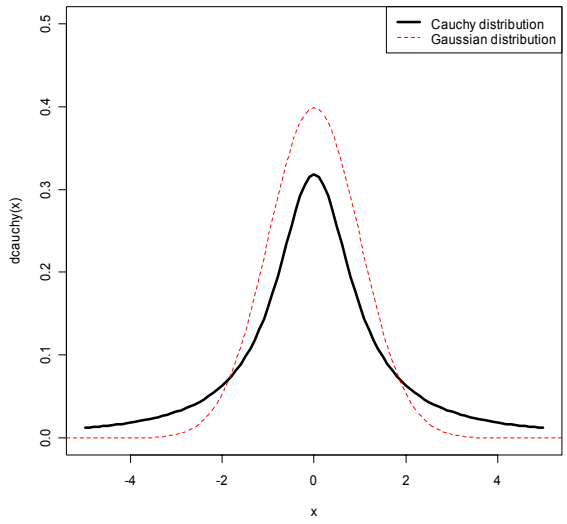
```
> curve(dweibull(x,.8), col='blue', add=T)
> title(main="Weibull Probability Distribution Function")
> legend(par('usr')[2], par('usr')[4], xjust=1,
        c('Exponential', 'Weibull, shape=1',
          'Weibull, shape=2', 'Weibull, shape=.8'),
        lwd=c(1,3,1,1),
        lty=c(1,3,1,1),
        col=c(par("fg"), par("fg"), 'red', 'blue'))
```



Biểu đồ 9. Phân phối Weibull.

- Phân phối *Cauchy*:

```
> curve(dcauchy(x),xlim=c(-5,5), ylim=c(0,.5), lwd=3)
> curve(dnorm(x), add=T, col='red', lty=2)
> legend(par('usr')[2], par('usr')[4], xjust=1,
        c('Cauchy distribution', 'Gaussian distribution'),
        lwd=c(3,1),
        lty=c(1,2),
        col=c(par("fg"), 'red'))
```



Biểu đồ 9. Phân phối Cauchy so sánh với phân phối chuẩn.

6.5 Chọn mẫu ngẫu nhiên (random sampling)

Trong xác suất và thống kê, lấy mẫu ngẫu nhiên rất quan trọng, vì nó đảm bảo tính hợp lí của các phương pháp phân tích và suy luận thống kê. Với R, chúng ta có thể lấy mẫu một mẫu ngẫu nhiên bằng cách sử dụng hàm `sample`.

Ví dụ: Chúng ta có một quần thể gồm 40 người (mã số 1, 2, 3, …, 40). Nếu chúng ta muốn chọn 5 đối tượng quần thể đó, ai sẽ là người được chọn? Chúng ta có thể dùng lệnh `sample()` để trả lời câu hỏi đó như sau:

```
> sample(1:40, 5)
 [1] 32 26  6 18  9
```

Kết quả trên cho biết đối tượng 32, 26, 8, 18 và 9 được chọn. Mỗi lần ra lệnh này, R sẽ chọn một mẫu khác, chứ không hoàn toàn giống như mẫu trên. Ví dụ:

```
> sample(1:40, 5)
[1]  5 22 35 19  4

> sample(1:40, 5)
[1] 24 26 12  6 22

> sample(1:40, 5)
[1] 22 38 11  6 18
```

v.v…

Trên đây là lệnh để chúng ta chọn mẫu ngẫu nhiên mà không thay thế (random sampling without replacement), tức là mỗi lần chọn mẫu, chúng ta không bỏ lại các mẫu đã chọn vào quần thể.

Nhưng nếu chúng ta muốn chọn mẫu thay thế (tức mỗi lần chọn ra một số đối tượng, chúng ta bỏ vào lại trong quần thể để chọn tiếp lần sau). Ví dụ, chúng ta muốn chọn 10 người từ một quần thể 50 người, bằng cách lấy mẫu với thay thế (random sampling with replacement), chúng ta chỉ cần thêm tham số `replace = TRUE`:

```
> sample(1:50, 10, replace=T)
 [1] 31 44  6  8 47 50 10 16 29 23
```

Hay ném một đồng xu 10 lần; mỗi lần, dĩ nhiên đồng xu có 2 kết quả H và T; và kết quả 10 lần có thể là:

```
> sample(c("H", "T"), 10, replace=T)
 [1] "H" "T" "H" "H" "H" "T" "H" "H" "T" "T"
```

Cũng có thể tưởng tượng chúng ta có 5 quả banh màu xanh (`X`) và 5 quả banh màu đỏ (`D`) trong một bao. Nếu chúng ta chọn 1 quả banh, ghi nhận màu, rồi để lại vào bao; rồi lại chọn 1 quả banh khác, ghi nhận màu, và bỏ vào bao lại. Cứ như thế, chúng ta chọn 20 lần, kết quả có thể là:

```
> sample(c("X", "D"), 20, replace=T)
 [1] "X" "D" "D" "D" "D" "D" "X" "X" "X" "X" "X" "D" "X" "X" "D" "X" "X" "X" "X"
[20] "D"
```

Ngoài ra, chúng ta còn có thể lấy mẫu với một xác suất cho trước. Trong hàm sau đây, chúng ta chọn 10 đối tượng từ dãy số 1 đến 5, nhưng xác suất không bằng nhau:

```
> sample(5, 10, prob=c(0.3, 0.4, 0.1, 0.1, 0.1), replace=T)
[1] 3 1 3 2 2 2 2 2 5 1
```

Đối tượng 1 được chọn 2 lần, đối tượng 2 được chọn 5 lần, đối tượng 3 được chọn 2 lần, v.v… Tuy không hoàn toàn phù hợp với xác suất 0.3, 0.4, 0.1 như cung cấp vì số mẫu còn nhỏ, nhưng cũng không quá xa với kì vọng.