Lâm sàng thống kê

Phân phối chuẩn

Nguyễn Văn Tuấn

Tuần vừa qua tôi nhận được một câu hỏi rất căn bản, mà tôi thấy cần phải giải thích rõ ràng, vì đây là cơ sở cho những phân tích thống kê. Khi phụ trách mục này, tôi giả định bạn đọc đã biết qua vài điều căn bản về thống kê và xác suất, nhưng có lẽ giả định đó không đúng, vì theo câu hỏi của bạn đọc này, vẫn có nhiều người chưa học qua, hoặc đã học qua mà … không hiểu. Cũng giống như tôi ngày xưa, học qua thống kê mà không hiểu vì nó quá trừu tượng. Không dám đổ thừa thầy giải thích không rõ, nhưng có lẽ vì khi giảng thầy không đề cập đến ứng dụng nên học chỉ để học chứ chẳng biết để làm gì.

"Gởi anh Tuấn! Tôi là một bác sĩ già, nên không rành về thống kê gì cả, vì hồi xưa tôi không có học thống kê. Nhưng bây giờ làm nghiên cứu tôi mới thấy sự quan trọng của nó. Tôi tìm sách để tự học, nhưng đọc hoài vẫn không hiểu! Trong khi sắp "đầu hàng" tình cờ tôi vào trang nhà ykhoanet và đọc được tất cả những bài giảng của anh. Phải nói thật anh giảng hay lắm, quá rõ ràng, làm cho một bác sĩ già như tôi mà cũng hiểu được các khái niệm thống kê, và tôi thấy yêu cái môn học này! Có lẽ anh không biết rằng anh đã giúp cho tôi rất nhiều. Xin cám ơn anh.
Tôi rất mong đọc tiếp loạt bài giảng "lâm sàng thống kê" của anh. Nhân đây tôi muốn hỏi anh một câu nhỏ. Trong mấy bài vừa qua, anh nhắc đến "phân phối chuẩn" và con số 1,96 để tính khoảng tin cậy 95% rất nhiều lần. Vậy xin hỏi anh, con số 1,96 này đến từ đâu và phân phối chuẩn là phân phối gì? Xin cám ơn anh trước.
TVĐ"

Xin thành thật cám ơn bạn đọc TVĐ về những câu chữ đầy khích lệ. Viết ra mà có người đọc và theo dõi thì thật là quí lắm. Đó cũng là động cơ để tôi viết tiếp. Nhân dịp này, tôi muốn mượn câu hỏi để giải thích về một định luật phân phối trụ cột của thống kê học: đó là phân phối chuẩn.

Thú thật với các bạn, ngày xưa, mỗi lần nghe đến hai chữ "distribution" (phân phối) là tôi đã thấy lùng bùng trong đầu rồi, vì không biết nó có nghĩa là gì. Cái khổ của một sinh viên ngoại quốc như tôi (tức là trình độ tiếng Anh lúc đó còn kém, nhúc nhác) giữa đồng môn người bản xứ, tôi không dám hỏi thầy, sợ bị mắng là … dốt. Sau này, tôi mới nghiệm ra rằng biết được mình dốt là một điều cực kì có ích và cũng là một hạnh phúc. Cái dốt của tôi bắt đầu từ chữ distribution, mà tôi thấy chưa có sách giáo khoa nào giải thích cụ thể cả, hay giải thích theo kiểu toán học rất trừu tượng.

Để cụ thể hóa vấn đề, bạn đọc có thể làm một thí nghiệm (hay tưởng tượng một thí nghiệm) đơn giản như sau: chọn ngẫu nhiên 100 đồng nghiệp hay sinh viên, đo chiều cao của họ. Kết quả mà bạn đọc sẽ thu thập được có thể như sau:

```
176.1 176.0 160.6 158.4 165.3 158.0 155.3 164.2 157.2 159.0
167.7 155.6 165.1 170.0 167.4 166.4 162.3 167.1 154.0 159.3
164.5 171.5 151.9 166.0 166.9 162.0 152.5 147.6 163.6 163.5
172.2 165.8 172.4 162.0 149.6 159.9 157.0 154.6 162.3 171.2
171.1 162.0 158.6 164.4 176.6 159.5 149.9 164.0 162.2 162.0
167.3 156.1 162.5 158.4 156.8 167.8 168.7 164.6 170.6 165.2
168.9 166.2 155.3 157.9 167.4 171.8 170.2 178.7 171.7 171.5
164.0 171.7 162.7 155.8 161.4 163.4 148.3 160.9 156.1 165.6
157.9 166.8 157.2 158.8 162.7 157.1 165.9 162.7 176.7 172.1
157.0 160.8 165.2 161.8 163.8 164.2 174.7 158.2 162.3 168.9
```

Trước một "rừng" con số như thế, chúng ta phải làm gì? Câu hỏi đó còn tùy thuốc vào mục đích của nghiên cứu. Nhưng ở đây, chúng ta muốn mô tả chiều cao và huyết áp của 100 đối tượng. Trong văn chương, "mô tả" có nghĩa là dùng từ ngữ để nói đến những khía cạnh của một sự kiện mà trong tiếng Anh nó tóm gọn trong những chữ cái *W*: *what* (sự kiện gì), *when* (xảy ra ở đâu), *where* (xảy ra lúc nào), và khó hơn chút là *why* (tại sao sự kiện xảy ra). Trong khoa học, chúng ta cũng mô tả sự kiện với những khía cạnh đó, nhưng chúng ta sử dụng cả từ ngữ và con số. Vì mô tả bằng con số, chúng ta cần hỏi thêm những câu hỏi như "bao nhiêu" (*how many* hay *how much*) như: chiều cao thấp nhất và cao nhất là bao nhiêu, chiều cao trung bình bao nhiêu, độ dao động cao thấp bao nhiêu, v.v…

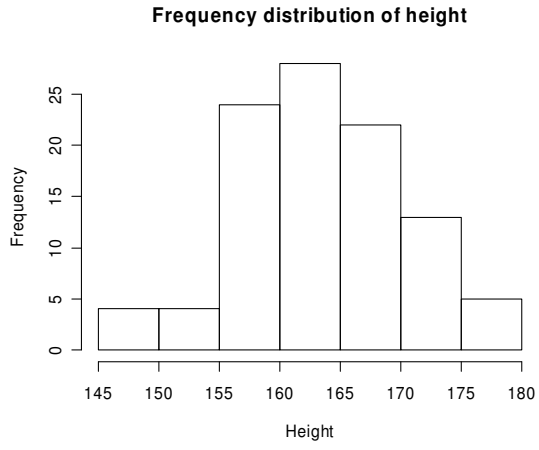
Với hàng trăm con số như thế, rất khó cảm nhận được vấn đề. Một cách khác tốt hơn là chúng ta sắp xếp số liệu từ thấp nhất đến cao nhất như sau:

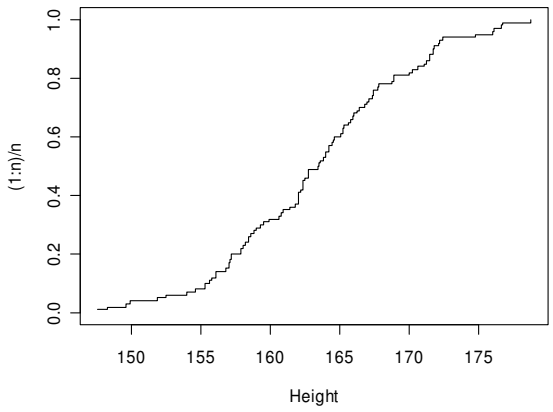
```
147.6 148.3 149.6 149.9 151.9 152.5 154.0 154.6 155.3 155.3
155.6 155.8 156.1 156.1 156.8 157.0 157.0 157.1 157.2 157.2
157.9 157.9 158.0 158.2 158.4 158.4 158.6 158.8 159.0 159.3
159.5 159.9 160.6 160.8 160.9 161.4 161.8 162.0 162.0 162.0
162.0 162.2 162.3 162.3 162.3 162.5 162.7 162.7 162.7 163.4
163.5 163.6 163.8 164.0 164.0 164.2 164.2 164.4 164.5 164.6
165.1 165.2 165.2 165.3 165.6 165.8 165.9 166.0 166.2 166.4
166.8 166.9 167.1 167.3 167.4 167.4 167.7 167.8 168.7 168.9
168.9 170.0 170.2 170.6 171.1 171.2 171.5 171.5 171.7 171.7
171.8 172.1 172.2 172.4 174.7 176.0 176.1 176.6 176.7 178.7
```

Cách sắp xếp này (tiếng Anh gọi là *sort*) cho chúng ta thấy người có chiều cao thấp nhất là 148.7 cm, và người cao nhất là 178.7 cm. Nhưng nếu nhìn kĩ, chúng ta cũng chú ý rằng phần lớn các đối tượng có chiều cao khoảng 160 đến 165 cm.

Đến đây thì câu hỏi đặt ra là có bao nhiêu đối tượng với mỗi chiều cao từ 160 đến 165 cm, và có bao nhiêu đối tượng có chiều cao thấp hơn hay cao hơn hai giá trị đó? Cố

nhiên, cách hay nhất là chúng ta đếm. Nhưng với máy tính, chúng ta có thể yêu cầu máy tính đếm và tốt hơn nữa là vẽ biểu đồ dưới đây.





Biểu đồ 1: (a) Mật độ phân phối của chiều cao, với trục tung là số đối tượng. (b) Biểu đồ bên phải là xác suất tích lũy (cumulative probability) của chiều cao.

Trong Biểu đồ trên (phía trái), trục tung là số đối tượng và trục hoành là chiều cao. Như bạn đọc có thể thấy, có 4 đối tượng với chiều cao từ 145 đến 150 cm, và từ 151 đến 155 cm. Tương tự, chỉ có 4 đối tượng có chiều cao từ 175 đến 180 cm. Đúng như cảm nhận ban đầu, đỉnh của biểu đồ là số đối tượng có chiều cao từ 160 đến 170 cm.

Biểu đồ bên phải thể hiện xác suất tích lũy chiều cao. Nhìn qua biểu đồ này, chúng ta có thể nói rằng khoảng 30% đối tượng có chiều cao thấp hơn 160 cm, và khoảng 80% đối tượng có chiều cao thấp hơn hay bằng 170 cm. Nói cách khác, số đối tượng có chiều cao từ 160 đến 170 cm chiếm khoảng 50% tổng số cỡ mẫu.

Do đó, nói đến "phân phối" là đề cập đến tần số khả dĩ (hay xác suất) của các giá trị chiều cao.

Về hình dạng, chúng ta dễ dàng thấy rằng sự phân phối chiều cao ở 100 đối tượng này giống như một hình chuông. Các phân phối có hình dạng này được gọi là "Normal distribution" (chữ N của normal viết hoa), hay phân phối bình thường. Nhưng vì tính cách chuẩn hóa của phân phối này, nên tôi tạm dịch là *phân phối chuẩn*. Để cho có vẻ khoa học và "trí thức" một chút (và làm cho nhiều người phải bức tóc gãi đầu), giới toán học thỉnh thoảng thêm chữ "luật" thành "luật phân phối"!

Phân phối bình thưởng còn được gọi là Gaussian distribution, bởi vì người phát hiện ra luật phân phối này là nhà toán học danh tiếng Carl F. Gauss (người Đức). Thật ra,

người đề cập đến luật phân phối này là nhà toán học người Pháp De Moivre, nhưng ông không phát triển thêm. Trong cuốn *Theorie Analytique des Probabilites,* Gauss phát triển các đặc điểm của luật phân phối chuẩn và chỉ ra rằng luật phân phối này phù hợp với các hiện tượng tự nhiên. Thật vậy, hầu hết các hiện tượng sinh học tự nhiên (như chiều cao, trọng lượng cơ thể, huyết áp, mật độ xương, v.v…) đều có thể mô tả bằng luật phân phối bình thưởng một cách chính xác. Chính vì thế mà luật phân phối chuẩn được ứng dụng cực kì rộng rãi trong khoa học thực nghiệm. Có thể nói không ngoa rằng phân phối chuẩn là nền tảng, là trụ cột của tất cả các phân tích thống kê. Không có luật phân phối này cũng có nghĩa là không có khoa học thống kê hiện đại.

Để hiểu rõ hơn tầm quan trọng của luật phân phối chuẩn, chúng ta cần ghi nhớ rằng trong nghiên cứu khoa học thực nghiệm, chúng ta không biết các thông số của một quần thể, mà chỉ sự vào các số liệu từ một hay nhiều mẫu để suy luận cho một quần thể. Cụ thể hơn, ở đây chúng ta không biết chiều cao trung bình của toàn thể người Việt là bao nhiêu, chúng ta chỉ biết chiều cao của 100 đối tượng vừa thu thập được, và chúng ta muốn sử dụng các số liệu này để suy luận cho toàn thể người Việt.

Do đó, trong bất cứ phân tích thống kê nào, chúng ta lúc nào nên nhớ và phân biệt giữa khái niệm quần thể (population) và mẫu (sample). Các chỉ số thống kê được ước tính từ mẫu gọi là *ước số* (*estimates*), và các chỉ số thống kê của quần thể chúng ta gọi là *thông số* (*parameters*). Thông thường các ước số được thể hiện bằng kí hiệu La Mã (như *m, s, t*), còn các thông số được kí hiệu bằng chữ Hi Lạp tương đương (như μ, σ, τ).

I. Phân phối chuẩn

Quay trở lại với vấn đề của chúng ta, một trong những câu hỏi mà có lẽ chúng ta muốn biết là: nếu một người đàn ông được chọn ngẫu nhiên, xác suất mà người đàn ông này có chiều cao bằng 160 cm là bao nhiêu. Hỏi cách khác (và theo ngôn ngữ không toán học), có bao nhiêu đàn ông người ở Việt Nam có chiều cao chính xác là 160 cm? Câu trả lời có thể dựa vào số liệu thu thập được. Chúng ta thấy chỉ có một người có chiều cao 159.9 cm (hay 160 cm), do đó xác suất là 1% (vì có mẫu chúng ta có là 100 người).

Nhưng vì chúng ta chọn mẫu ngẫu nhiên, cho nên con số này chưa chắc chính xác. Nếu chúng ta ngẫu nhiên chọn 100 người khác, có thể có hai người có chiều cao 160 cm, và do đó xác suất là 2%.

Thật ra, chúng ta cũng có thể đặt một câu hỏi chung như sau: nếu một đàn ông được chọn ngẫu nhiên, xác suất mà vị đàn ông này có chiều cao *x* cm là bao nhiêu? Hay, nói cách khác, có bao nhiêu phần trăm đàn ông Việt Nam với chiều cao *x* cm, trong đó *x* có thể là bất cứ giá trị chiều cao nào. Trong tình huống bất định của chọn mẫu như thế, luật phân phối chuẩn cung cấp cho chúng ta một mô hình toán học để trả lời câu hỏi này.

Gọi X là biến số chiều cao, μ là chiều cao trung bình của một quần thể, và σ là độ lệch chuẩn, câu hỏi trên có thể phát biểu bằng công thức toán học như sau:

$$P\left(X = x \mid \mu, \sigma^2\right) = ?$$

(Chú ý, P là viết tắt của chữ probability, tức xác suất; kí hiệu "|" có nghĩa là "given" hay "với điều kiện"). Do đó, kí hiệu trên có thể đọc như sau: xác suất mà $X = x$ với điều kiện chúng ta biết được μ và σ là bao nhiêu). Câu trả lời mà Gauss đã có sẵn cho chúng ta là:

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

Chú ý rằng công thức trên đôi khi cũng xuất hiện trong các sách giáo khoa với một hình thức khác: thay vì viết $P\left(X = x \mid \mu, \sigma^2\right)$, có tác giả viết khó hiểu hơn là *f(x)*! Tất nhiên, trong công thức trên $\pi = 3.1416\ldots$

Như có thể thấy qua công thức [1] trên đây, luật phân phối chuẩn được hoàn toàn xác định bởi 2 thông số: trung bình μ và độ lệch chuẩn σ. Nói cách khác, nếu chúng ta biết được 2 thông số này, chúng ta có thể ước tính xác suất cho bất cứ chiều cao nào. (Do đó chúng ta cần phải chọn mẫu (sample) nghiên cứu như thế nào để cho các ước số của mẫu nghiên cứu là rất sát với các thông số tương đương của quần thể. Phần này đã được đề cập chi tiết trong bài chọn mẫu nghiên cứu). Trong trường hợp của chúng ta, ước số cho μ và σ chính là số trung bình và độ lệch chuẩn của mẫu. Các ước số này là (các bạn có thể kiểm tra):

Trung bình: m = 163.3 cm
Độ lệch chuẩn: s = 6.6 cm

Thay thế các ước số này cho cho μ và σ, chúng ta có thể trả lời câu hỏi "có bao nhiêu đàn ông người ở Việt Nam có chiều cao chính xác là 160 cm":

$$P\left(X = 160\right) = \frac{1}{6.6\times\sqrt{2\times 3.1416}} \exp\left[-\frac{(160-163.3)^2}{2\times(6.6)^2}\right] = 0.0533$$

Theo đáp số này, chúng ta có thể đoán rằng có khoảng 5.3% đàn ông Việt Nam có chiều cao chính xác là 160 cm. Tuy cách tính thoạt đầu nhìn qua có vẻ khác phức tạp, nhưng với phần mềm R, chỉ một lệnh đơn giản `dnorm(160, mean=163.3, sd=6.6)` là chúng ta có ngay đáp số chính xác!

Tương tự, chúng ta có thể ước tính xác suất cho bất cứ chiều cao nào qua công thức [1]. Bảng sau đây trình bày một số xác suất cho chiều cao từ thấp đến cao.

Bảng 1. Xác suất chiều cao của đàn ông Việt Nam

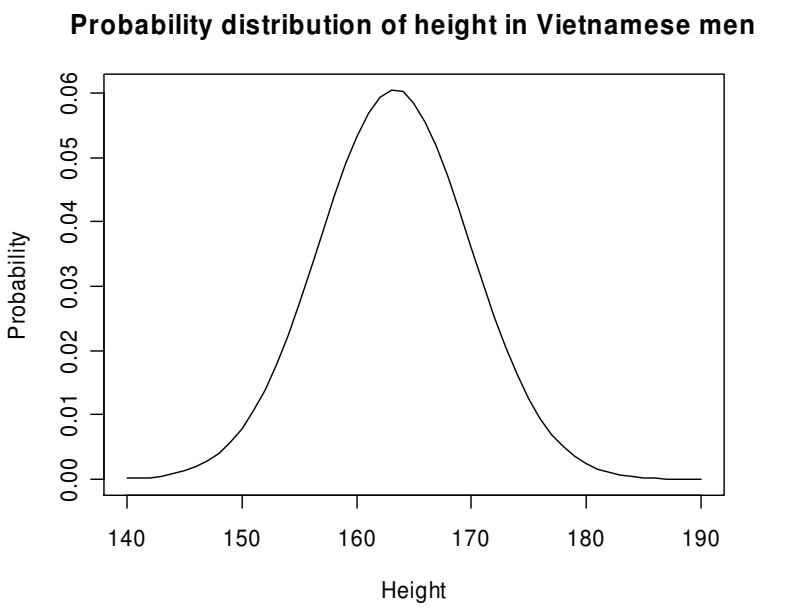
Chiều cao (cm)	Xác suất (tính bằng %)
140	0.0118
141	0.0200
142	0.0331
143	0.0533
144	0.0840
145	0.1290
146	0.1947
147	0.2863
148	0.4116
149	0.5781
150	0.7935
151	1.0645
152	1.3958
153	1.7886
154	2.2398
155	2.7412
156	3.2788
157	3.8327
158	4.3786
159	4.8887
160	5.3343

Chiều cao (cm)	Xác suất (tính bằng %)
161	5.6885
162	5.9285
163	6.0383
164	6.0107
165	5.8474
166	5.5594
167	5.1656
168	4.6908
169	4.1630
170	3.6107
171	3.0606
172	2.5354
173	2.0527
174	1.6242
175	1.2559
176	0.9491
177	0.7010
178	0.5060
179	0.3570
180	0.2461
181	0.1658

Nếu bạn đọc chịu khó cộng tất cả các xác suất này lại (thực ra không cần) thì tổng số sẽ là gần bằng 100%. Nói tóm lại, xác suất gần 100% là chiều cao của đàn ông Việt Nam dao động từ 140 đến 181 cm.

Giả dụ như nếu một đàn ông có chiều cao 200 cm, câu hỏi đặt ra là chiều cao này có "bất bình thường" hay không. Theo sự phân phối chiều cao như vừa mô tả (tức trung bình 163.3 cm và độ lệch chuẩn 6.6 cm), số đàn ông Việt Nam có chiều cao 200 cm chỉ 0.00000116 mà thôi.

Các xác suất trên đây cũng có thể thể hiện bằng một biểu đồ mà thuật ngữ tiếng Anh gọi là *probability density distribution (pdf)* mà tôi tạm dịch là *phân phối của mật độ xác suất*. Biểu đồ này như sau:



Biểu đồ 2. Mật độ xác suất chiều cao ở đàn ông Việt Nam với trung bình 163.3 cm và độ lệch chuẩn 6.6 cm.

Biểu đồ trên chính là luật phân phối chuẩn (theo công thức [1]). Tất nhiên, tổng diện tích dưới đường biểu diễn phải bằng 1 (hay 100%). Điều này có nghĩa là nếu chúng ta muốn ước tính xác suất cho bất cứ khoảng chiều cao nào. Ví dụ nếu chúng ta muốn biết có bao nhiêu đàn ông Việt Nam có chiều thấp hơn 150 cm, chúng ta chỉ cần tính diện tích mà trục hoành từ 150 cm hay thấp hơn dưới đường biểu diễn. Phát biểu theo ngôn ngữ toán học câu hỏi này là: $P(X < 150) = ?$ Hay nói chính xác hơn nữa:

$$P\left(X < 150 \mid \mu = 163.3, \sigma = 6.6\right) = ?$$

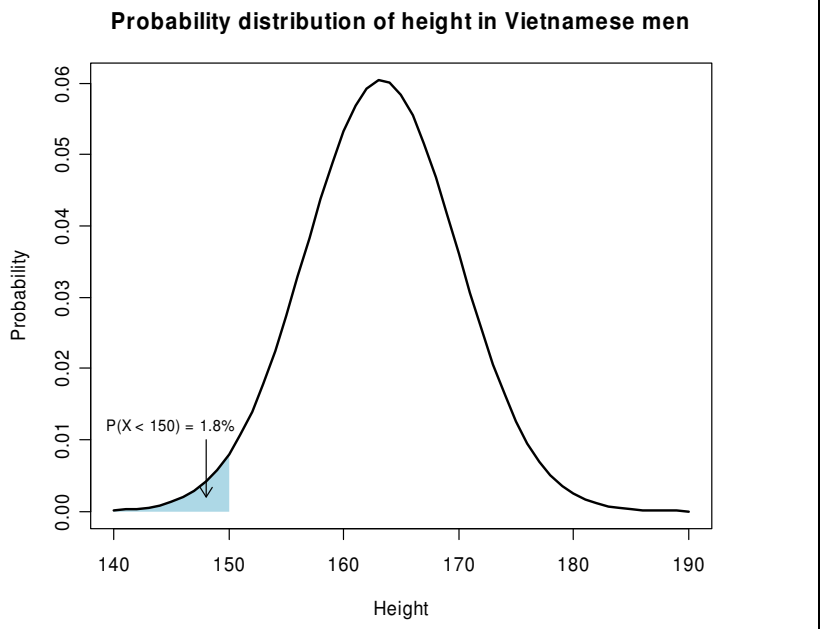
Cách tính đơn giản nhất là chúng ta cộng các xác suất chiều từ 140 đến 149 (Bảng 1 (**Bảng 1**): 0.0118 + 0.0200 + 0.0331 + …. + 0.5781 = 1.8%.

Tuy nhiên, có một cách tính nhanh hơn và "tinh vi" hơn là sử dụng tích phân. Bạn đọc nào còn nhớ tích phân thì câu trả lời cho câu hỏi này quá đơn giản: chỉ cần tính tích phân chiều cao từ 0 (thấp nhất) đến 159 cm:

$$P\left(X < 150\right) = \int_0^{149} f\left(x\right)dx$$

trong đó, $f(x)=\frac{1}{6.6\sqrt{2\pi}}\exp\left[-\frac{(x-163.3)^2}{2(6.6)^2}\right]$. Kết quả tất nhiên là 0.018. Bạn đọc không cần phải làm các tính toán tích phân phức tạp, vì phầm mềm R có một lệnh đơn giản để tính tích phân trên (tôi trình bày lệnh này trong phần chú thích ở phía cuối bài).

Biểu đồ dưới đây minh họa cho xác suất này bằng cách tô đậm diện tích dưới đường biểu diễn để bạn đọc có thể hiểu rõ hơn:



Biểu đồ 3. Diện tích dưới đường biểu diễn (màu xanh nhạt) cho chiều cao <150 cm là xác suất $P(X<150 \mid \mu=163.3, \sigma=6.6)$ = 0.018

Tương tự, chúng ta có thể ước tính xác suất cho bất cứ khoảng chiều cao nào giữa *a* và *b* theo công thức tích phân trên đây. Chẳng hạn như xác suất đàn ông Việt Nam có chiều cao từ 160 đến 170 cm là:

$$P(160 \le X \le 170) = \int_{160}^{170} f(x)dx$$

Hay một cách chung hơn:

$$P(a < X < b) = \int_{a}^{b} f(x)dx \qquad [2]$$

II. Phân phối chuẩn hóa – standardized normal distribution

Trong phần trên, chúng ta quan tâm đến việc phân tích chiều cao bằng cách ứng dụng luật phân phối chuẩn. Tuy nhiên, như đề cập trong phần đầu, luật phân phối chuẩn có thể ứng dụng cho rất nhiều hiện tượng tự nhiên. Nhưng các biến khác nhau về đơn vị đo lường, như chiều cao đo bằng cm, nhưng huyết áp đo bằng mmHg, nên chúng ta khó mà so sánh hai biến số này bởi vì chúng có đơn vị đo lường khác nhau, và có thể độ lệch chuẩn cũng khác nhau. Chẳng hạn như nếu một đối tượng có chiều cao là 175 cm và huyết áp là 120 mmHg, làm sao chúng ta biết các thông số cá nhân này cao hay thấp. Do đó, chúng ta cần phải có một cách chuẩn hóa luật phân phối sao cho chúng ta có thể so sánh các biến số này mà không cần biết đến đơn vị đo lường.

Một trong những cách chuẩn hóa đó là phân phối chuẩn hóa, mà có lẽ bạn đọc từng thấy đâu đó trong sách giáo khoa người ta gọi là *standardized normal distrubution.* Như thấy trong công thức [1], hai thông số trung bình và độ lệch chuẩn hoàn toàn xác định luật phân phối chuẩn, cho nên, một cách chuẩn hóa là hoán chuyển chiều cao (hay một biến số) sao cho chúng độc lập với đơn vị đo lường. Cách hoán chuyển này có tên là *z-transformation* hay hoán chuyển *z*. Kết quả của hoán chuyển là một *chỉ số z* (thuật ngữ tiếng Anh là *z-score*).

Trong ví dụ về chiều cao, z là độ khác biệt giữa chiều cao một cá nhân (kí hiệu là *x*) và chiều cao trung bình của quần thể chia cho độ lệch chuẩn. Nói cách khác:

$$z = \frac{x - \mu}{\sigma} \qquad [3]$$

Bởi vì *x,* μ và σ trong công thức trên đây đều có cùng đơn vị (cm), và cm chia cho cm thì không biến mới hoàn toàn độc lập với đơn vị đo lường. Thật ra, đơn vị của *z* bây giờ không còn là cm nữa, mà là độ lệch chuẩn. Xem kĩ công thức [3] trên chúng ta có thể rút ra vài nhận xét như sau:

- Nếu chiều cao của một cá nhân thấp hơn chiều cao trung bình của dân số (tức là $x < \mu$) chỉ số *z* sẽ âm. Chẳng hạn như nếu ông A có chiều cao 150 cm, thì chỉ số z của ông là $z = \frac{150 - 163.3}{6.6}$ = -2.01, tức là thấp hơn chiều cao của dân số khoảng 2 độ lệch chuẩn;

- Nếu $x = \mu$, chỉ số *z* sẽ là 0;

- Và nếu $x > \mu$, chỉ số *z* sẽ là số dương. Chẳng hạn như nếu chiều cao của một đối tượng là 175 cm, thì *z* = 1.77. Nói cách khác, chiều cao của đối tượng này cao hơn trung bình khoảng 1.8 độ lệch chuẩn.

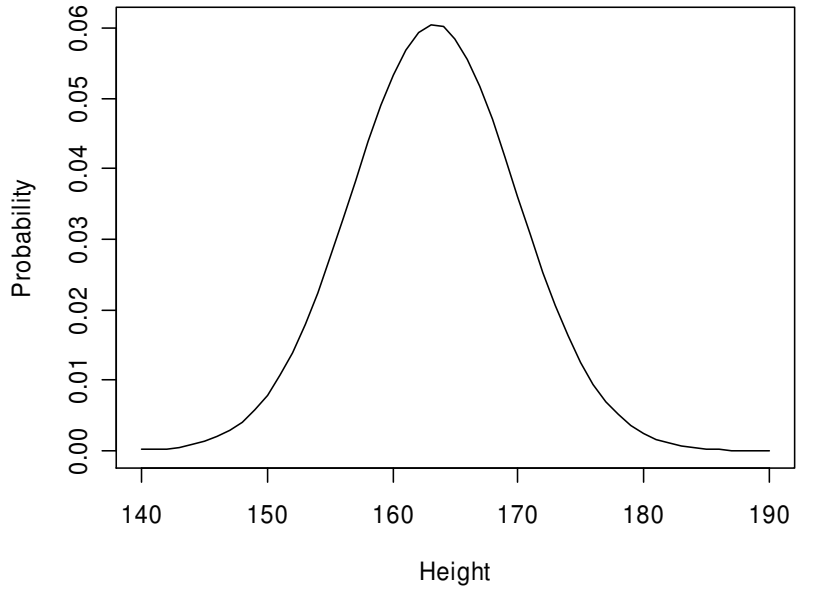
Như vậy, thay vì mô tả sự phân phối của chiều cao bằng đơn vị cm với hàm số [1], chúng ta mô tả bằng đơn vị độ lệch chuẩn hay chỉ số *z*. Chỉ số z bây giờ có số trung bình là $\mu = 0$ và độ lệch chuẩn là $\sigma = 1$. Nếu thay [3] vào [1], chúng ta có một hàm số mới và đơn giản hơn như sau:

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right] \qquad [4]$$

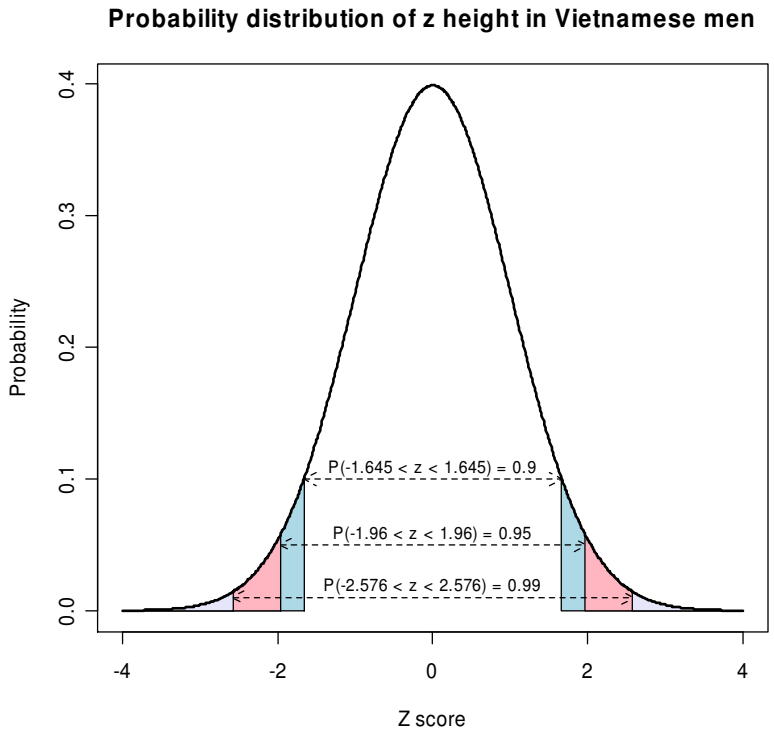
Và hàm số tích lũy [2] sẽ trở thành:

$$P(a < z < b) = \int_a^b f(z)dz = \int_a^b \frac{e^{-0.5z^2}}{\sqrt{2\pi}}dz \qquad [5]$$

Biểu đồ 4 dưới đây minh họa cho phân phối chiều cao tính bằng cm và bằng chỉ số z:



Biểu đồ 4a. Mật độ xác suất chiều cao ở đàn ông Việt Nam, mô tả bằng cm.



Biểu đồ 4b. Mật độ xác suất của phân phối chuẩn *f(z)*, với trung bình 0 và độ lệch chuẩn 1.

Cố nhiên, diện tích dưới đường biểu diễn của hàm số *f(z)* trong Biểu đồ 4b phải là khoảng 1. Nói cách khác, $P(-4 < z < 4) = \int_{-4}^{4} f(z)dz$; 1. Ngoài ra, phân phối chuẩn như mô tả qua Biểu đồ 4b còn hàm chứa một số thông tin có ích và thú vị:

- Xác suất mà $z \leq 1.96$ là 0.025 (tức 2.5%). Nói cách khác, diện tích dưới đường biểu diễn tính từ $z = -1.96$ hay thấp hơn là 0.025.

- Bởi vì phân phối chuẩn cân đối (symmetric), chúng ta cũng có thể nói (hay suy luận) rằng xác suất mà $z \geq 1.96$ cũng bằng 0.025.

- Như vậy, xác suất mà *z* nằm trong khoảng -1.96 và 1.96 là 1–0.025–0.025 = 0.95 (hay 95%). Nói cách khác, khoảng tin cậy 95% của *z* là -1.96 đến 1.96.

- Tương tự, chúng ta cũng có thể phát biểu (và bạn đọc có thể tự mình kiểm chứng) rằng xác suất mà *z* nằm trong khoảng -1.645 đến 1.645 là 90%. Xác suất mà *z* nằm trong khoảng -2.576 đến 2.576 là 99%. Xác suất mà *z* nằm trong khoảng -3.09 đến 3.09 là 99.9%.

Đến đây, chúng ta đã thấy hằng số 1.96, 1.64 hay 3.0 xuất phát từ đâu! Các hằng số này chẳng có gì bí mật cả: chúng là chỉ số *z* của phân phối chuẩn. Bảng sau đây sẽ cung cấp một số xác suất cho các chỉ số *z* thông dụng trong thống kê học và ứng dụng trong y khoa:

Bảng 2. Xác suất các giá trị *z*

z	-3.090	-2.326	-1.96	-1.645	-1.282	0	1.282	1.96	2.326	3.090
P(Z≤z)	0.001	0.01	0.025	0.05	0.10	0.50	0.90	0.975	0.99	0.999

III. Khoảng tin cậy 95%

Bây giờ chúng ta sẽ điểm qua vài ứng dụng luật phân phối chuẩn trong y khoa. Vì có quá nhiều ứng dụng, nên tôi chỉ tập trung vào những vấn đề liên quan đến những bài giảng của tôi, và một vấn đề mà chúng ta hay thấy là ước tính khoảng tin cậy 95% (thuật ngữ tiếng Anh là *95% confidence interval* hay có khi còn viết là *95% confidence limit*, thậm chí *95% credible interval*).

Trong nhiều nghiên cứu y học mang tính mô tả, chúng ta thường muốn phát triển một các tham chiếu (*reference range* hay có khi gọi không chính xác là *normal range*). Chẳng hạn như để phát triển các giá trị tham chiếu cho một biến số sinh hóa như calcium trong máu, chúng ta có thể ngẫu nhiên chọn một số đối tượng và đo nồng độ calcium trong máu, và sau đó tính khoảng tin cậy 95%. Khoảng tin cậy 95% này chính là các giá trị tham chiếu. Nếu nồng độ calcium trong máu của một cá nhân nằm ngoài khoảng tin cậy 95% thì chúng ta có thể (xin nhấn mạnh: "có thể") phát biểu rằng nồng độ của cá nhân này "bất bình thường".

Để ước tính khoảng tin cậy 95% (KTC95%), chúng ta chú ý mối liên hệ giữa *x* và *z* trong công thức [3]; vì $z = \frac{x-\mu}{\sigma}$, do đó:

$$x = \mu + z\sigma$$

Như đề cập trong phần trên, 95% giá trị của *z* nằm trong khoảng -1.96 đến +1.96, cho nên chúng ta cũng có thể nói rằng 95% giá trị của *x* nằm trong khoảng $\mu - 1.96\sigma$ và $\mu + 1.96\sigma$. Hay nói ngắn gọn hơn, 95% các giá trị *x* nằm trong khoảng:

$$x = \mu \pm 1.96\sigma \qquad [6]$$

Quay lại với ví dụ về chiều cao, chúng ta biết rằng số trung bình là 163.3 cm và độ lệch chuẩn là 6.6 cm. Do đó, chúng ta có thể suy luận rằng 95% đàn ông Việt Nam có chiều cao trong khoảng $163.3 \pm 1.96 \times 6.6 = 150.4$ cm đến 176.2 cm.

Tất nhiên, chúng ta cũng có thể ước tính xác suất 99% chiều cao đàn ông Việt Nam nằm trong khoảng $163.3 \pm 3 \times 6.6 = 143.5$ cm đến 183.1 cm. Do đó, nếu một đàn ông có chiều cao thấp hơn 143.5 cm, chúng ta có thể nói là "thấp", với xác suất dưới 0.5%!

Tùy theo vấn đề cụ thể, nhưng phần lớn các giá trị tham chiếu trong y khoa đều lấy khoảng tin cậy 95% làm chuẩn. Khi xác suất một chỉ số thống kê nằm ngoài khoảng tin cậy 95% được xem là "có ý nghĩa thống kê" (statistical significant).

IV. Kết luận

Qua bài này, hi vọng tôi đã giải thích phân phối chuẩn là gì, và hằng số 1.96 trong cách tính khoảng tin cậy 95% xuất phát từ đâu. Phân phối chuẩn đóng một vai trò thiết yếu trong khoa học thống kê. Hầu hết tất cả các suy luận thống kê đều dựa vào luật phân phối chuẩn để phát triển các kiểm định thống kê (statistical tests). Ngay cả các luật phân phối nhị phân hay phân phối Poisson (mà tôi sẽ bàn đến trong một bài khác) cũng có thể mô hình bằng luật phân phối chuẩn.

Như là một qui luật tự nhiên, rất nhiều biến số lâm sàng và khoa học thực nghiệm nói chung đều tuân theo luật phân phối chuẩn. Cũng có thể có một số biến số sinh hóa không tuân theo luật phân phối chuẩn, nhưng có thể hoán chuyển để chúng tuân theo luật phân phối chuẩn. Do đó, các phương pháp phân tích tham số (parametric methods) vẫn có thể áp dụng cho các biến loại này.

Các mã R sử dụng trong bài viết:

```
# Nhập dữ liệu về chiều cao và gọi biến là ht
# nguồn: mô phỏng

ht <- c(
176.1, 176.0, 160.6, 158.4, 165.3, 158.0, 155.3, 164.2, 157.2, 159.0,
167.7, 155.6, 165.1, 170.0, 167.4, 166.4, 162.3, 167.1, 154.0, 159.3,
164.5, 171.5, 151.9, 166.0, 166.9, 162.0, 152.5, 147.6, 163.6, 163.5,
172.2, 165.8, 172.4, 162.0, 149.6, 159.9, 157.0, 154.6, 162.3, 171.2,
171.1, 162.0, 158.6, 164.4, 176.6, 159.5, 149.9, 164.0, 162.2, 162.0,
167.3, 156.1, 162.5, 158.4, 156.8, 167.8, 168.7, 164.6, 170.6, 165.2,
168.9, 166.2, 155.3, 157.9, 167.4, 171.8, 170.2, 178.7, 171.7, 171.5,
164.0, 171.7, 162.7, 155.8, 161.4, 163.4, 148.3, 160.9, 156.1, 165.6,
157.9, 166.8, 157.2, 158.8, 162.7, 157.1, 165.9, 162.7, 176.7, 172.1,
157.0, 160.8, 165.2, 161.8, 163.8, 164.2, 174.7, 158.2, 162.3, 168.9)


# Sắp xếp số liệu chiều cao từ thấp đến cao

sort(ht)

# Vẽ biểu đồ mật 1a

hist(ht, breaks=10,
     xlab="Height", main="Frequency distribution of height")

# Vẽ biểu đồ mật 1b

n <- length(ht)
plot(sort(ht), (1:n)/n,
     type="s", ylim=c(0,1), xlab="Height")

plot(density(ht), main="Plot of density distribution of height",
xlab="Height")

# Tìm số trung bình và độ lệch chuẩn của chiều cao

mean(ht)
sd(ht)

# Ước tính xác suất chiều cao = 160 cm với trung bình=163.3 và sd=6.6

dnorm(160, mean=163.3, sd=6.6)

# Ước tính xác suất cho bảng 1

height <- seq(140, 181, 1)
dnorm(height, mean=163.3, sd=6.6)*100

# Vẽ biểu đồ 2
```

```
height <- seq(140, 190, 1)
plot(height, dnorm(height, 163.3, 6.6),
       type="l",
       ylab=”Probability”,
       xlab=”Height”,
       main="Probability distribution of height in Vietnamese men")
```

 # Ước tính xác suất chiều cao < 150 cm, $P(X<150)=\int_0^{149} f(x)dx$

```
pnorm(149, mean=163.3, sd=6.6)
```

 # Vẽ biểu đồ 3

```
height <- seq(140, 190, 1)
dht <- dnorm(height, 163.3, 6.6)
ht <- data.frame(z=height, ht=dht)
zc <- 150
plot(ht,
     type="n",
     ylab="Probability",
     xlab="Height",
     main="Probability distribution of height in Vietnamese men")

t <- subset(ht, z<= zc)
polygon(c(rev(t$z), t$z),
        c(rep(0, nrow(t)), t$ht), col="lightblue", border=NA)
lines(ht, lwd=2)
arrows(148,0.01,148,0.002, angle=30, length=0.1)
text(145,0.012, "P(X < 150) = 1.8%", cex=0.8)
```

 # Hoán chuyển sang z score và vẽ biểu đồ 4b

```
zheight <- seq(-4, 4, 0.01)
dzht <- dnorm(zheight, 0, 1)
zht <- data.frame(z=zheight, ht=dzht)

plot(zht,
       type="n",
       ylab="Probability",
       xlab="Z score",
       main="Probability distribution of z height in Vietnamese men")

z1 <- 1.65
z2 <- -1.65
z3 <- 1.96
z4 <- -1.96
z5 <- 2.58
z6 <- -2.58

t1 <- subset(zht, z>= z1)
polygon(c(rev(t1$z), t1$z),
        c(rep(0, nrow(t1)), t1$ht), col="lightblue")
```

```
t2 <- subset(zht, z<= z2)
polygon(c(rev(t2$z), t2$z),
        c(rep(0, nrow(t2)), t2$ht), col="lightblue")

t3 <- subset(zht, z>= z3)
polygon(c(rev(t3$z), t3$z),
        c(rep(0, nrow(t3)), t3$ht), col="lightpink")

t4 <- subset(zht, z<= z4)
polygon(c(rev(t4$z), t4$z),
        c(rep(0, nrow(t4)), t4$ht), col="lightpink")

t5 <- subset(zht, z>= z5)
polygon(c(rev(t5$z), t5$z),
        c(rep(0, nrow(t5)), t5$ht), col="lavender")

t6 <- subset(zht, z<= z6)
polygon(c(rev(t6$z), t6$z),
        c(rep(0, nrow(t6)), t6$ht), col="lavender")

lines(zht, lwd=2)
arrows(-1.65,0.1,1.65,0.1, angle=30, length=0.1, code=3, lty=2)
text(0,0.11, "P(-1.645 < z < 1.645) = 0.9", cex=0.8)

arrows(-1.96,0.05,1.96,0.05, angle=30, length=0.1, code=3, lty=2)
text(0,0.06, "P(-1.96 < z < 1.96) = 0.95", cex=0.8)

arrows(-2.58,0.01,2.58,0.01, angle=30, length=0.1, code=3, lty=2)
text(0,0.02, "P(-2.576 < z < 2.576) = 0.99", cex=0.8)

# Cho bài tập : nhập số liệu huyết áp của 100 đối tượng
# nguồn: nghiên cứu bệnh đái tháo đường TPHCM 2007.

bp <- c(
 90, 130,  120,  130,  100,  150,  100,  120,  100,  110,
110, 170,  110,  110,  120,  110,  110,  120,  110,   85,
110, 120,  120,  120,  110,  150,  120,  120,  120,  110,
130, 150,  150,  110,  140,  140,  120,  110,  120,  110,
150, 110,  120,  120,  130,  110,  110,  120,  120,  140,
100, 130,  130,  130,  140,  100,  110,  110,  110,  120,
130, 110,  130,  120,  150,  100,  120,  100,  120,  140,
120, 100,  100,  110,  140,  125,  100,  140,  110,  120,
120, 120,  150,  120,  110,  120,  150,  100,  110,  120,
160, 110,  110,  120,  120,  150,  120,  130,  160,   90)
```