

9. Phân tích thống kê mô tả

9.1 Thống kê mô tả (descriptive statistics, summary)

Để minh họa cho việc áp dụng R vào thống kê mô tả, tôi sẽ sử dụng một dữ liệu nghiên cứu có tên là `igfdata`. Trong nghiên cứu này, ngoài các chỉ số liên quan đến giới tính, độ tuổi, trọng lượng và chiều cao, chúng tôi đo lường các hormone liên quan đến tình trạng tăng trưởng như `igfi`, `igfbp3`, `als`, và các markers liên quan đến sự chuyển hóa của xương `pinp`, `ictp` và `p3np`. Có 100 đối tượng nghiên cứu. Dữ liệu này được chứa trong directory `c:\works\stats`. Trước hết, chúng ta cần phải nhập dữ liệu vào R với những lệnh sau đây (các câu chữ theo sau dấu `#` là những chú thích để bạn đọc theo dõi):

```
> options(width=100)
# chuyển directory
> setwd("c:/works/stats")

# đọc dữ liệu vào R
> igfdata <- read.table("igf.txt", header=TRUE, na.strings=".")
> attach(igfdata)

# xem xét các cột số trong dữ liệu
> names(igfdata)
[1] "id"      "sex"     "age"     "weight"  "height"  "ethnicity"
[7] "igfi"    "igfbp3"  "als"     "pinp"    "ictp"    "p3np"

> igfdata
  id  sex age weight height ethnicity  igfi  igfbp3  als  pinp  ictp  p3np
1   1 Female 15   42   162   Asian 189.000 4.00000 323.667 353.970 11.2867 8.3367
2   2  Male 16   44   160 Caucasian 160.000 3.75000 333.750 375.885 10.4300 6.7450
3   3 Female 15   43   157   Asian 146.833 3.43333 248.333 199.507  8.3633 12.5000
4   4 Female 15   42   155   Asian 185.500 3.40000 251.000 483.607 13.3300 14.2767
5   5 Female 16   47   167   Asian 192.333 4.23333 322.000 105.430  7.9233  4.5033
6   6 Female 25   45   160   Asian 110.000 3.50000 284.667  76.487  4.9833  4.9367
7   7 Female 19   45   161   Asian 157.000 3.20000 274.000  75.880  6.3500  5.3200
8   8 Female 18   43   153   Asian 146.000 3.40000 303.000  86.360  7.3700  4.6700
9   9 Female 15   41   149   Asian 197.667 3.56667 308.500 254.803 11.8700  6.8200
10  10 Female 24   45   157  African 148.000 3.40000 273.000  44.720  3.7400  6.1600
...
...
97  97 Female 17   54   168 Caucasian 204.667 4.96667 441.333  64.130  5.1600  4.4367
98  98  Male 18   55   169   Asian 178.667 3.86667 273.000 185.913  7.5267  8.8333
99  99 Female 18   48   151   Asian 237.000 3.46667 324.333 105.127  5.9867  5.6600
100 100  Male 15   54   168   Asian 130.000 2.70000 259.333 325.840 10.2767  6.5933
```

Trên đây chỉ là một phần số liệu trong số 100 đối tượng.

Cho một biến số $x_1, x_2, x_3, \dots, x_n$ chúng ta có thể tính toán một số chỉ số thống kê mô tả như sau:

Lý thuyết	Hàm R
Số trung bình: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	<code>mean(x)</code>
Phương sai: $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	<code>var(x)</code>
Độ lệch chuẩn: $s = \sqrt{s^2}$	<code>sd(x)</code>
Sai số chuẩn (standard error): $SE = \frac{s}{\sqrt{n}}$	Không có
Trị số thấp nhất	<code>min(x)</code>
Trị số cao nhất	<code>max(x)</code>
Toàn cự (range)	<code>range(x)</code>

Để tìm giá trị trung bình của độ tuổi, chúng ta chỉ đơn giản lệnh:

```
> mean(age)
[1] 19.17
```

Hay phương sai và độ lệch chuẩn của tuổi:

```
> var(age)
[1] 15.33444

> sd(age)
[1] 3.915922
```

Tuy nhiên, R có lệnh `summary` có thể cho chúng ta tất cả thông tin thống kê về một biến số:

```
> summary(age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 13.00   16.00   19.00   19.17   21.25   34.00
```

Nói chung, kết quả này đơn giản và các viết tắt cũng có thể dễ hiểu. Chú ý, trong kết quả trên, có hai chỉ số “1st Qu” và “3rd Qu” có nghĩa là first quartile (tương đương với vị trí 25%) và third quartile (tương đương với vị trí 75%) của một biến số. First quartile = 16 có nghĩa là 25% đối tượng nghiên cứu có độ tuổi bằng hoặc nhỏ hơn 16 tuổi. Tương tự, Third quartile = 34 có nghĩa là 75% đối tượng có độ tuổi bằng hoặc thấp hơn 34 tuổi. Tất nhiên số trung vị (median) 19 cũng có nghĩa là 50% đối tượng có độ tuổi 19 trở xuống (hay 19 tuổi trở lên).

R không có hàm tính sai số chuẩn, và trong hàm `summary`, R cũng không cung cấp độ lệch chuẩn. Để có các số này, chúng ta có thể tự viết một hàm đơn giản (hãy gọi là `desc`) như sau:

```
desc <- function(x)
{
  av <- mean(x)
  sd <- sd(x)
  se <- sd/sqrt(length(x))
  c(MEAN=av, SD=sd, SE=se)
}
```

Và có thể gọi hàm này để tính bất cứ biến nào chúng ta muốn, như tính biến `als` sau đây:

```
> desc(als)
      MEAN      SD      SE
301.841120  58.987189  5.898719
```

Thí dụ muốn có các thống kê mô tả đối với `igdata`

```
> summary(igfdata)
      id      sex      age      weight      height      ethnicity
Min.   : 1.00  Female:69  Min.   :13.00  Min.   :41.00  Min.   :149.0  African   : 8
1st Qu.:25.75  Male  :31   1st Qu.:16.00  1st Qu.:47.00  1st Qu.:157.0  Asian     :60
Median :50.50                Median :19.00  Median :50.00  Median :162.0  Caucasian:30
Mean   :50.50                Mean   :19.17  Mean   :49.91  Mean   :163.1  Others    : 2
3rd Qu.:75.25                3rd Qu.:21.25  3rd Qu.:53.00  3rd Qu.:168.0
Max.   :100.00               Max.   :34.00  Max.   :60.00  Max.   :196.0

      igfi      igfbp3      als      pinp      ictp
Min.   : 85.71  Min.   :2.000  Min.   :192.7  Min.   : 26.74  Min.   : 2.697
1st Qu.:137.17  1st Qu.:3.292  1st Qu.:256.8  1st Qu.: 68.10  1st Qu.: 4.878
Median :161.50  Median :3.550  Median :292.5  Median :103.26  Median : 6.338
Mean   :165.59  Mean   :3.617  Mean   :301.8  Mean   :167.17  Mean   : 7.420
3rd Qu.:186.46  3rd Qu.:3.875  3rd Qu.:331.2  3rd Qu.:196.45  3rd Qu.: 8.423
Max.   :427.00  Max.   :5.233  Max.   :471.7  Max.   :742.68  Max.   :21.237

      p3np
Min.   : 2.343
1st Qu.: 4.433
Median : 5.445
Mean   : 6.341
3rd Qu.: 7.150
Max.   :16.303
```

```
> by(igfdata, sex, summary)
```

sex: Female

id	sex	age	weight	height
Min. : 1.0	Female: 69	Min. : 13.00	Min. : 41.00	Min. : 149.0
1st Qu.: 21.0	Male : 0	1st Qu.: 17.00	1st Qu.: 47.00	1st Qu.: 156.0
Median : 47.0		Median : 19.00	Median : 50.00	Median : 162.0
Mean : 48.2		Mean : 19.59	Mean : 49.35	Mean : 161.9
3rd Qu.: 75.0		3rd Qu.: 22.00	3rd Qu.: 52.00	3rd Qu.: 166.0
Max. : 99.0		Max. : 34.00	Max. : 60.00	Max. : 196.0

ethnicity	igfi	igfbp3	als
African : 4	Min. : 85.71	Min. : 2.767	Min. : 204.3
Asian : 43	1st Qu.: 136.67	1st Qu.: 3.333	1st Qu.: 263.8
Caucasian: 22	Median : 163.33	Median : 3.567	Median : 302.7
Others : 0	Mean : 167.97	Mean : 3.695	Mean : 311.5
	3rd Qu.: 186.17	3rd Qu.: 3.933	3rd Qu.: 361.7
	Max. : 427.00	Max. : 5.233	Max. : 471.7

pinp	ictp	p3np
Min. : 26.74	Min. : 2.697	Min. : 2.343
1st Qu.: 62.75	1st Qu.: 4.717	1st Qu.: 4.337
Median : 78.50	Median : 5.537	Median : 5.143
Mean : 108.74	Mean : 6.183	Mean : 5.643
3rd Qu.: 115.26	3rd Qu.: 7.320	3rd Qu.: 6.143
Max. : 502.05	Max. : 13.633	Max. : 14.420

sex: Male

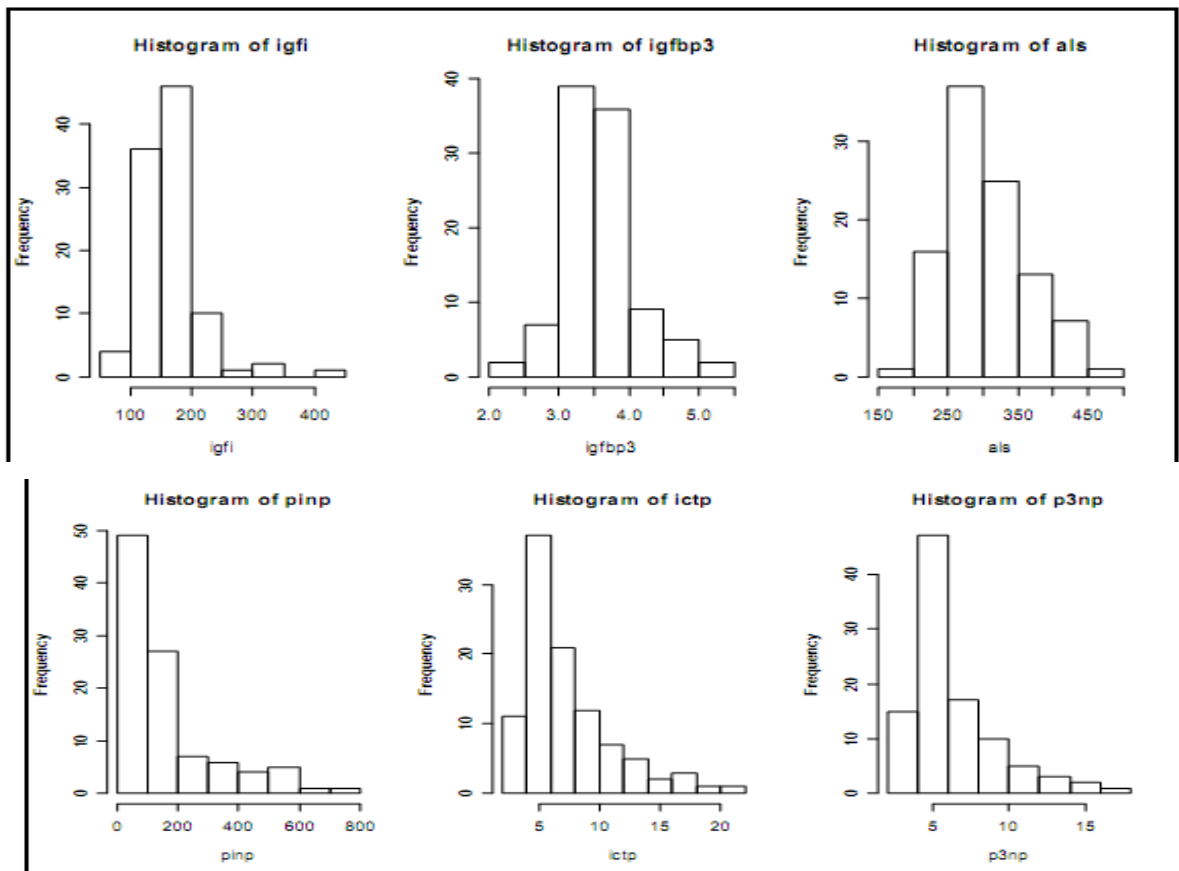
id	sex	age	weight	height
Min. : 2.00	Female: 0	Min. : 14.00	Min. : 44.00	Min. : 155.0
1st Qu.: 34.50	Male : 31	1st Qu.: 15.00	1st Qu.: 48.50	1st Qu.: 161.5
Median : 56.00		Median : 17.00	Median : 51.00	Median : 164.0
Mean : 55.61		Mean : 18.23	Mean : 51.16	Mean : 165.6
3rd Qu.: 75.00		3rd Qu.: 20.00	3rd Qu.: 53.50	3rd Qu.: 169.0
Max. : 100.00		Max. : 27.00	Max. : 59.00	Max. : 191.0

ethnicity	igfi	igfbp3	als
African : 4	Min. : 94.67	Min. : 2.000	Min. : 192.7
Asian : 17	1st Qu.: 138.67	1st Qu.: 3.183	1st Qu.: 249.8
Caucasian: 8	Median : 160.00	Median : 3.500	Median : 276.0
Others : 2	Mean : 160.29	Mean : 3.443	Mean : 280.2
	3rd Qu.: 183.00	3rd Qu.: 3.775	3rd Qu.: 311.3
	Max. : 274.00	Max. : 4.500	Max. : 388.7

pinp	ictp	p3np
Min. : 56.28	Min. : 3.650	Min. : 3.390
1st Qu.: 135.07	1st Qu.: 6.900	1st Qu.: 5.375
Median : 245.92	Median : 9.513	Median : 7.140
Mean : 297.21	Mean : 10.173	Mean : 7.895
3rd Qu.: 450.38	3rd Qu.: 13.517	3rd Qu.: 10.010
Max. : 742.68	Max. : 21.237	Max. : 16.303

Để xem qua phân phối của các hormones và chỉ số sinh hóa cùng một lúc, chúng ta có thể vẽ đồ thị cho tất cả 6 biến số. Trước hết, chia màn ảnh thành 6 cửa sổ (với 2 dòng và 3 cột); sau đó lần lượt vẽ:

```
> op <- par(mfrow=c(2,3))
> hist(igfi)
> hist(igfbp3)
> hist(als)
> hist(pinp)
> hist(ictp)
> hist(p3np)
```



9.2 Thống kê mô tả theo từng nhóm

Nếu chúng ta muốn tính trung bình của một biến số như `igfi` cho mỗi nhóm nam và nữ giới, hàm `tapply` trong R có thể dùng cho việc này:

```
> tapply(igfi, list(sex), mean)
      Female      Male 
167.9741 160.2903
```

Trong lệnh trên, `igfi` là biến số chúng ta cần tính, biến số phân nhóm là `sex`, và chỉ số thống kê chúng ta muốn là trung bình (`mean`). Qua kết quả trên, chúng ta thấy số trung bình của `igfi` cho nữ giới (167.97) cao hơn nam giới (160.29).

Nhưng nếu chúng ta muốn tính cho từng giới tính và sắc tộc, chúng ta chỉ cần thêm một biến số trong hàm `list`:

```
> tapply(igfi, list(ethnicity, sex), mean)
      Female      Male 
African  145.1252 120.9168
```

```
Asian      165.6589 160.4999
Caucasian 176.6536 169.4790
Others           NA 200.5000
```

Trong kết quả trên, NA có nghĩa là “not available”, tức không có số liệu cho phụ nữ trong các sắc tộc “others”.

9.3.1 Kiểm định t một mẫu

Ví dụ 10. Qua phân tích trên, chúng ta thấy tuổi trung bình của 100 đối tượng trong nghiên cứu này là 19.17 tuổi. Chẳng hạn như trong quần thể này, trước đây chúng ta biết rằng tuổi trung bình là 30 tuổi. Vấn đề đặt ra là có phải mẫu mà chúng ta có được có đại diện cho quần thể hay không. Nói cách khác, chúng ta muốn biết giá trị trung bình 19.17 có thật sự khác với giá trị trung bình 30 hay không.

Để trả lời câu hỏi này, chúng ta sử dụng kiểm định t. Theo lý thuyết thống kê, kiểm định t được định nghĩa bằng công thức sau đây:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Trong đó, $\bar{x}$ là giá trị trung bình của mẫu, μ là trung bình theo giả thiết (trong trường hợp này, 30), s là độ lệch chuẩn, và n là số lượng mẫu (100). Nếu giá trị t cao hơn giá trị lý thuyết theo phân phối t ở một tiêu chuẩn có ý nghĩa như 5% chẳng hạn thì chúng ta có lý do để phát biểu khác biệt có ý nghĩa thống kê. Giá trị này cho mẫu 100 có thể tính toán bằng hàm `qt` của R như sau:

```
> qt(0.95, 100)
[1] 1.660234
```

Nhưng có một cách tính toán nhanh gọn hơn để trả lời câu hỏi trên, bằng cách dùng hàm `t.test` như sau:

```
> t.test(age, mu=30)

One Sample t-test

data:  age
t = -27.6563, df = 99, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 30
95 percent confidence interval:
 18.39300 19.94700
sample estimates:
mean of x
 19.17
```

Trong lệnh trên `age` là biến số chúng ta cần kiểm định, và `mu=30` là giá trị giả thiết. R trình bày trị số $t = -27.66$, với 99 bậc tự do, và trị số $p < 2.2e-16$ (tức rất thấp). R cũng cho biết độ tin cậy 95% của `age` là từ 18.4 tuổi đến 19.9 tuổi (30 tuổi nằm quá ngoài khoảng tin cậy này). Nói cách khác, chúng ta có lý do để phát biểu rằng độ tuổi trung bình trong mẫu này thật sự thấp hơn độ tuổi trung bình của quần thể.

9.3.2 Kiểm định t hai mẫu

Ví dụ 11. Qua phân tích mô tả trên (phần `summary`) chúng ta thấy phụ nữ có độ hormone `igfi` cao hơn nam giới (167.97 và 160.29). Câu hỏi đặt ra là có phải thật sự đó là một khác biệt có hệ thống hay do các yếu tố ngẫu nhiên gây nên. Trả lời câu hỏi này, chúng ta cần xem xét mức độ khác biệt trung bình giữa hai nhóm và độ lệch chuẩn của độ khác biệt.

$$t = \frac{\bar{x}_2 - \bar{x}_1}{SED}$$

Trong đó $\bar{x}_1$ và $\bar{x}_2$ là số trung bình của hai nhóm nam và nữ, và SED là độ lệch chuẩn của $(\bar{x}_1 - \bar{x}_2)$. Thực ra, SED có thể ước tính bằng công thức:

$$SED = \sqrt{SE_1^2 + SE_2^2}$$

Trong đó SE_1 và SE_2 là sai số chuẩn (standard error) của hai nhóm nam và nữ. Theo lý thuyết xác suất, t tuân theo luật phân phối t với bậc tự do $n_1 + n_2 - 2$, trong đó n_1 và n_2 là số mẫu của hai nhóm. Chúng ta có thể dùng R để trả lời câu hỏi trên bằng hàm `t.test` như sau:

```
> t.test(igfi ~ sex)

Welch Two Sample t-test

data:  igfi by sex
t = 0.8412, df = 88.329, p-value = 0.4025
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.46855  25.83627
sample estimates:
mean in group Female  mean in group Male
    167.9741         160.2903
```

R trình bày các giá trị quan trọng trước hết:

```
t = 0.8412, df = 88.329, p-value = 0.4025
```

df là bậc tự do. Trị số $p = 0.4025$ cho thấy mức độ khác biệt giữa hai nhóm nam và nữ không có ý nghĩa thống kê (vì cao hơn 0.05 hay 5%).

```
95 percent confidence interval:
-10.46855 25.83627
```

là khoảng tin cậy 95% về độ khác biệt giữa hai nhóm. Kết quả tính toán trên cho biết độ igf ở nữ giới có thể thấp hơn nam giới 10.5 ng/L hoặc cao hơn nam giới khoảng 25.8 ng/L. Vì độ khác biệt quá lớn và đó là thêm bằng chứng cho thấy không có khác biệt có ý nghĩa thống kê giữa hai nhóm.

Kiểm định trên dựa vào giả thiết hai nhóm nam và nữ có khác phương sai. Nếu chúng ta có lí do để cho rằng hai nhóm có cùng phương sai, chúng ta chỉ thay đổi một thông số trong hàm `t` với `var.equal=TRUE` như sau:

```
> t.test(igfi ~ sex, var.equal=TRUE)

Two Sample t-test

data: igfi by sex
t = 0.7071, df = 98, p-value = 0.4812
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-13.88137 29.24909
sample estimates:
mean in group Female    mean in group Male
      167.9741           160.2903
```

Về mặt số, kết quả phân tích trên có khác chút ít so với kết quả phân tích dựa vào giả định hai phương sai khác nhau, nhưng trị số p cũng đi đến một kết luận rằng độ khác biệt giữa hai nhóm không có ý nghĩa thống kê.

9.4 Kiểm định Wilcoxon cho hai mẫu (`wilcox.test`)

Kiểm định t dựa vào giả thiết là phân phối của một biến phải tuân theo luật phân phối chuẩn. Nếu giả định này không đúng, kết quả của kiểm định t có thể không hợp lí (valid). Để kiểm định phân phối của `igfi`, chúng ta có thể dùng hàm `shapiro.test` như sau:

```
> shapiro.test(igfi)

Shapiro-Wilk normality test
```

```
data: igfi
W = 0.8528, p-value = 1.504e-08
```

Trị số p nhỏ hơn 0.05 rất nhiều, cho nên chúng ta có thể nói rằng phân phối của *igfi* không tuân theo luật phân phối chuẩn. Trong trường hợp này, việc so sánh giữa hai nhóm có thể dựa vào phương pháp phi tham số (non-parametric) có tên là kiểm định Wilcoxon, vì kiểm định này (không như kiểm định t) không tùy thuộc vào giả định phân phối chuẩn.

```
> wilcox.test(igfi ~ sex)

Wilcoxon rank sum test with continuity correction

data: igfi by sex
W = 1125, p-value = 0.6819
alternative hypothesis: true mu is not equal to 0
```

Trị số $p = 0.682$ cho thấy quả thật độ khác biệt về *igfi* giữa hai nhóm nam và nữ không có ý nghĩa thống kê. Kết luận này cũng không khác với kết quả phân tích bằng kiểm định t .

9.5 Kiểm định t cho các biến số theo cặp (paired t -test, t .test)

Kiểm định t vừa trình bày trên là cho các nghiên cứu gồm hai nhóm độc lập nhau (như giữa hai nhóm nam và nữ), nhưng không thể ứng dụng cho các nghiên cứu mà một nhóm đối tượng được theo dõi theo thời gian. Tôi tạm gọi các nghiên cứu này là nghiên cứu theo cặp. Trong các nghiên cứu này, chúng ta cần sử dụng một kiểm định t có tên là paired t -test.

Ví dụ 12. Một nhóm bệnh nhân gồm 10 người được điều trị bằng một thuốc nhằm giảm huyết áp. Huyết áp của bệnh nhân được đo lúc khởi đầu nghiên cứu (lúc chưa điều trị), và sau khi điều trị. Số liệu huyết áp của 10 bệnh nhân như sau:

Trước khi điều trị (x_0)	180, 140, 160, 160, 220, 185, 145, 160, 160, 170
Sau khi điều trị (x_1)	170, 145, 145, 125, 205, 185, 150, 150, 145, 155

Câu hỏi đặt ra là độ biến chuyển huyết áp trên có đủ để kết luận rằng thuốc điều trị có hiệu quả giảm áp huyết. Để trả lời câu hỏi này, chúng ta dùng kiểm định t cho từng cặp như sau:

```
> # nhập dữ kiện
> before <- c(180, 140, 160, 160, 220, 185, 145, 160, 160, 170)
> after <- c(170, 145, 145, 125, 205, 185, 150, 150, 145, 155)
> bp <- data.frame(before, after)

> # kiểm định t
> t.test(before, after, paired=TRUE)
```

Paired t-test

```
data: before and after
t = 2.7924, df = 9, p-value = 0.02097
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 1.993901 19.006099
sample estimates:
mean of the differences
      10.5
```

Kết quả trên cho thấy sau khi điều trị áp suất máu giảm 10.5 mmHg, và khoảng tin cậy 95% là từ 2.0 mmHg đến 19 mmHg, với trị số $p = 0.0209$. Như vậy, chúng ta có bằng chứng để phát biểu rằng mức độ giảm huyết áp có ý nghĩa thống kê.

Chú ý nếu chúng ta phân tích sai bằng kiểm định thống kê cho hai nhóm độc lập dưới đây thì trị số $p = 0.32$ cho biết mức độ giảm áp suất không có ý nghĩa thống kê!

```
> t.test(before, after)
```

Welch Two Sample t-test

```
data: before and after
t = 1.0208, df = 17.998, p-value = 0.3209
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-11.11065 32.11065
sample estimates:
mean of x mean of y
  168.0    157.5
```

9.6 Kiểm định Wilcoxon cho các biến số theo cặp (`wilcox.test`)

Thay vì dùng kiểm định t cho từng cặp, chúng ta cũng có thể sử dụng hàm `wilcox.test` cho cùng mục đích:

```
> wilcox.test(before, after, paired=TRUE)
```

Wilcoxon signed rank test with continuity correction

```
data: before and after
V = 42, p-value = 0.02291
alternative hypothesis: true mu is not equal to 0
```

Kết quả trên một lần nữa khẳng định rằng độ giảm áp suất máu có ý nghĩa thống kê với trị số ($p=0.023$) chẳng khác mấy so với kiểm định t cho từng cặp.

9.7 Tần số (frequency)

Hàm `table` trong R có chức năng cho chúng ta biết về tần số của một biến số mang tính phân loại như `sex` và `ethnicity`.

```
> table(sex)
sex
Female    Male
     69     31

> table(ethnicity)
ethnicity
African    Asian Caucasian    Others
      8      60         30        2
```

Một bảng thống kê 2 chiều:

```
> table(sex, ethnicity)
sex      ethnicity
sex      African Asian Caucasian Others
Female      4     43         22      0
Male        4     17         8       2
```

Chú ý trong các bảng thống kê trên, hàm `table` không cung cấp cho chúng ta số phần trăm. Để tính số phần trăm, chúng ta cần đến hàm `prop.table` và cách sử dụng có thể minh họa như sau:

```
# tạo ra một object tên là freq để chứa kết quả tần số
> freq <- table(sex, ethnicity)

# kiểm tra kết quả
> freq
sex      ethnicity
sex      African Asian Caucasian Others
Female      4     43         22      0
Male        4     17         8       2

# dùng hàm margin.table để xem kết quả
> margin.table(freq, 1)
sex
Female    Male
     69     31

> margin.table(freq, 2)
ethnicity
African    Asian Caucasian    Others
```

```

      8      60      30      2

# tính phần trăm bằng hàm prop.table
> prop.table(freq, 1)
      ethnicity
sex      African      Asian      Caucasian      Others
Female 0.05797101 0.62318841 0.31884058 0.00000000
Male   0.12903226 0.54838710 0.25806452 0.06451613

```

Trong bảng thống kê trên, `prop.table` tính tỉ lệ sắc tộc cho từng giới tính. Chẳng hạn như ở nữ giới (female), 5.8% là người Phi châu, 62.3% là người Á châu, 31.8% là người Tây phương da trắng. Tổng cộng là 100%. Tương tự, ở nam giới tỉ lệ người Phi châu là 12.9%, Á châu là 54.8%, v.v...

```

# tính phần trăm bằng hàm prop.table
> prop.table(freq, 2)
      ethnicity
sex      African      Asian      Caucasian      Others
Female 0.5000000 0.7166667 0.7333333 0.0000000
Male   0.5000000 0.2833333 0.2666667 1.0000000

```

```

# tính phần trăm cho toàn bộ bảng
> freq/sum(freq)
      ethnicity
sex      African      Asian      Caucasian      Others
Female    0.04    0.43         0.22    0.00
Male      0.04    0.17         0.08    0.02

```

9.8 Kiểm định tỉ lệ (proportion test, `prop.test`, `binom.test`)

Kiểm định một tỉ lệ thường dựa vào giả định phân phối nhị phân (binomial distribution). Với một số mẫu n và tỉ lệ p , và nếu n lớn (tức hơn 50 chẳng hạn), thì phân phối nhị phân có thể tương đương với phân phối chuẩn với số trung bình np và phương sai $np(1-p)$. Gọi x là số biến cố mà chúng ta quan tâm, kiểm định giả thiết $p = \pi$ có thể sử dụng thống kê sau đây:

$$z = \frac{x - n\pi}{\sqrt{n\pi(1-\pi)}}$$

Ở đây, z tuân theo luật phân phối chuẩn với trung bình 0 và phương sai 1. Cũng có thể nói z^2 tuân theo luật phân phối Chi bình phương với bậc tự do bằng 1.

Ví dụ 13. Trong nghiên cứu trên, chúng ta thấy có 69 nữ và 31 nam. Như vậy tỉ lệ nữ là 0.69 (hay 69%). Để kiểm định xem tỉ lệ này có thật sự khác với tỉ lệ 0.5 hay không, chúng ta có thể sử dụng hàm `prop.test(x, n,  $\pi$ )` như sau:

```
> prop.test(69, 100, 0.50)

1-sample proportions test with continuity correction

data: 69 out of 100, null probability 0.5
X-squared = 13.69, df = 1, p-value = 0.0002156
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.5885509 0.7766330
sample estimates:
      p 
0.69
```

Trong kết quả trên, `prop.test` ước tính tỉ lệ nữ giới là 0.69, và khoảng tin cậy 95% là 0.588 đến 0.776. Giá trị Chi bình phương là 13.69, với trị số $p = 0.00216$. Như vậy, nghiên cứu này có tỉ lệ nữ cao hơn 50%.

Một cách tính chính xác hơn kiểm định tỉ lệ là kiểm định nhị phân `binom.test(x, n,  $\pi$ )` như sau:

```
> binom.test(69, 100, 0.50)

Exact binomial test

data: 69 and 100
number of successes = 69, number of trials = 100, p-value = 0.0001831
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5896854 0.7787112
sample estimates:
probability of success
      0.69
```

Nói chung, kết quả của kiểm định nhị phân không khác gì so với kiểm định Chi bình phương, với trị số $p = 0.00018$, chúng ta càng có bằng chứng để kết luận rằng tỉ lệ nữ giới trong nghiên cứu này thật sự cao hơn 50%.

9.9 So sánh hai tỉ lệ (`prop.test`, `binom.test`)

Phương pháp so sánh hai tỉ lệ có thể khai triển trực tiếp từ lí thuyết kiểm định một tỉ lệ vừa trình bày trên. Cho hai mẫu với số đối tượng n_1 và n_2 , và số biến cố là x_1 và x_2 . Do đó, chúng ta có thể ước tính hai tỉ lệ p_1 và p_2 . Lí thuyết xác suất cho phép chúng ta phát biểu rằng độ khác biệt giữa hai mẫu $d = p_1 - p_2$ tuân theo luật phân phối chuẩn với số trung bình 0 và phương sai bằng:

$$V_d = \left(\frac{1}{n_1} + \frac{1}{n_2} \right) p(1-p)$$

Trong đó:

$$p = \frac{x_1 + x_2}{n_1 + n_2}$$

Thành ra, $z = d/V_d$ tuân theo luật phân phối chuẩn với trung bình 0 và phương sai 1. Nói cách khác, z^2 tuân theo luật phân phối Chi bình phương với bậc tự do bằng 1. Do đó, chúng ta cũng có thể sử dụng `prop.test` để kiểm định hai tỉ lệ.

Ví dụ 14. Một nghiên cứu được tiến hành so sánh hiệu quả của thuốc chống gãy xương. Bệnh nhân được chia thành hai nhóm: nhóm A được điều trị gồm có 100 bệnh nhân, và nhóm B không được điều trị gồm 110 bệnh nhân. Sau thời gian 12 tháng theo dõi, nhóm A có 7 người bị gãy xương, và nhóm B có 20 người gãy xương. Vấn đề đặt ra là tỉ lệ gãy xương trong hai nhóm này bằng nhau (tức thuốc không có hiệu quả)? Để kiểm định xem hai tỉ lệ này có thật sự khác nhau, chúng ta có thể sử dụng hàm

`prop.test(x, n, pi)` như sau:

```
> fracture <- c(7, 20)
> total <- c(100, 110)
> prop.test(fracture, total)

      2-sample test for equality of proportions with continuity
correction

data:  fracture out of total
X-squared = 4.8901, df = 1, p-value = 0.02701
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.20908963 -0.01454673
sample estimates:
   prop 1    prop 2 
0.0700000 0.1818182
```

Kết quả phân tích trên cho thấy tỉ lệ gãy xương trong nhóm 1 là 0.07 và nhóm 2 là 0.18. Phân tích trên còn cho thấy xác suất 95% rằng độ khác biệt giữa hai nhóm có thể 0.01 đến 0.20 (tức 1 đến 20%). Với trị số $p = 0.027$, chúng ta có thể nói rằng tỉ lệ gãy xương trong nhóm A quả thật thấp hơn nhóm B.

9.10 So sánh nhiều tỉ lệ (`prop.test`, `chisq.test`)

Kiểm định `prop.test` còn có thể sử dụng để kiểm định nhiều tỉ lệ cùng một lúc. Trong nghiên cứu trên, chúng ta có 4 nhóm sắc tộc và tần số cho từng giới tính như sau:

```
> table(sex, ethnicity)
```

	ethnicity			
sex	African	Asian	Caucasian	Others
Female	4	43	22	0
Male	4	17	8	2

Chúng ta muốn biết tỉ lệ nữ giới giữa 4 nhóm sắc tộc có khác nhau hay không, và để trả lời câu hỏi này, chúng ta lại dùng `prop.test` như sau:

```
> female <- c( 4, 43, 22, 0)
> total <- c(8, 60, 30, 2)
> prop.test(female, total)

      4-sample test for equality of proportions without continuity
      correction

data:  female out of total
X-squared = 6.2646, df = 3, p-value = 0.09942
alternative hypothesis: two.sided
sample estimates:
   prop 1    prop 2    prop 3    prop 4 
0.5000000 0.7166667 0.7333333 0.0000000 

Warning message:
Chi-squared approximation may be incorrect in: prop.test(female, total)
```

Tuy tỉ lệ nữ giới giữa các nhóm có vẻ khác nhau lớn (73% trong nhóm 3 (người da trắng) so với 50% trong nhóm 1 (Phi châu) và 71.7% trong nhóm Á châu, nhưng kiểm định Chi bình phương cho biết trên phương diện thống kê, các tỉ lệ này không khác nhau, vì trị số $p = 0.099$.

9.10.1 Kiểm định Chi bình phương (Chi squared test, `chisq.test`)

Thật ra, kiểm định Chi bình phương còn có thể tính toán bằng hàm `chisq.test` như sau:

```
> chisq.test(sex, ethnicity)

      Pearson's Chi-squared test

data:  sex and ethnicity
X-squared = 6.2646, df = 3, p-value = 0.09942

Warning message:
Chi-squared approximation may be incorrect in: chisq.test(sex,
ethnicity)
```

Kết quả này hoàn toàn giống với kết quả từ hàm `prop.test`.

9.10.2 Kiểm định Fisher (Fisher's exact test, `fisher.test`)

Trong kiểm định Chi bình phương trên, chúng ta chú ý cảnh báo:

```
"Warning message:
Chi-squared approximation may be incorrect in: prop.test(female, total)"
```

Vì trong nhóm 4, không có nữ giới cho nên tỉ lệ là 0%. Hơn nữa, trong nhóm này chỉ có 2 đối tượng. Vì số lượng đối tượng quá nhỏ, cho nên các ước tính thống kê có thể không đáng tin cậy. Một phương pháp khác có thể áp dụng cho các nghiên cứu với tần số thấp như trên là kiểm định `fisher` (còn gọi là Fisher's exact test). Bạn đọc có thể tham khảo lý thuyết đằng sau kiểm định `fisher` để hiểu rõ hơn về logic của phương pháp này, nhưng ở đây, chúng ta chỉ quan tâm đến cách dùng R để tính toán kiểm định này. Chúng ta chỉ đơn giản lệnh:

```
> fisher.test(sex, ethnicity)

      Fisher's Exact Test for Count Data

data:  sex and ethnicity
p-value = 0.1048
alternative hypothesis: two.sided
```

Chú ý trị số p từ kiểm định Fisher là 0.1048, tức rất gần với trị số p của kiểm định Chi bình phương. Cho nên, chúng ta có thêm bằng chứng để khẳng định rằng tỉ lệ nữ giới giữa các sắc tộc không khác nhau một cách đáng kể.