- Biểu đồ Quantile-normal: Đây là kỹ thuật EDA đơn biến phức tạp nhất. Nó được gọi là biểu đồ quantile-normal QN hoặc chính xác hơn là biểu đồ quantile-quantile QQ. Biểu đồ này được sử dụng để xem xét mức độ một mẫu cụ thể tuân theo phân phối lý thuyết cụ thể. Nó cho phép phát hiện ra nhưng sai lệch và chuẩn đoán được độ lệch và độ nhọn.

d) Đồ họa đa biến

Đồ hoạ đa biến sử dụng đồ hoạ để hiển thị mối quan hệ giữa hai hoặc nhiều tập dữ liệu. Cách sử dụng phổ biến nhất là biểu đồ thanh nhóm, với mỗi nhóm đại diện cho một cấp độ của trong các biến và mỗi thanh trong nhóm đại diện cho số lượng của biến khác. Các loại đồ hoạ đa biến phổ biến bao gồm:

- Biểu đồ phân tán: Đối với 2 biến số định lượng, kỹ thuật phân tích dữ liệu trực quan cơ bản là biểu đồ phân tán. Một biến trên trục x, một biến trên trục y và điểm nối cho mỗi trường hợp trong tập dữ liệu.
- Biểu đồ thời gian: Đây là biểu đồ đường của dữ liệu được vẽ theo thời gian.
- Biểu đồ nhiệt: Đây là biểu diễn đồ hoạ của dữ liệu, trong đó các giá trị được miêu tả bằng màu sắc.
- Biểu đồ đa biến: Đây là biểu đồ biểu diễn đồ hoạ của mối quan hệ giữa các yếu tố phản hồi
- Biểu đồ bong bóng: Đây là một phương tiện trực quan hoá dữ liệu hiển thị nhiều hình tròn (bong bóng) trong một đồ thị 2 chiều.

1.2. TÓM TẮT KẾT QUẢ THEO SUY DIỄN THỐNG KÊ

Tóm tắt các kết quả theo suy diễn thống kê như các tính toán về đặc trưng của dữ liệu mẫu, các bài toán ước lượng, bài toán kiểm định tham số, bài toán phân tích hệ số tương quan để tạo thành các mô dul phân tích thống kê. Những modul này kết với phân tích dữ liệu qua biểu đồ sẽ cho kết quả trực quan dữ liệu chính xác hơn. Ngoài ra, trong phần này các ví dụ minh họa sử dụng file dữ liệu Diem_TN, xem [1].

1.2.1. Thống kê mô tả

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

Lý thuyết	Hàm R
Số trung bình: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$	`mean(x)`
Phương sai: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$	`var(x)`
Độ lệch chuẩn: $s = \sqrt{s^2}$	`sd(x)`
Trị số thấp nhất	`min(x)`
Trị số cao nhất	`max(x)`
Toàn cự (range)	`range(x)`

Bảng 1.3: Các hàm tính thống kê mô tả cơ bản trong R.

Ví dụ 1.19. Xét file dữ liệu Diem_TN, để tìm giá trị trung bình, phương sai của điểm toán (T) chúng ta dùng lệnh đơn lẻ hoặc lệnh tổng quan:

```
# lệnh đơn lẻ
mean(Diem_TN$T)
sd(Diem_TN$T)
# lệnh tổng quan
Sumary(Diem_TN$T)
# kết quả hiển thị
> mean
[1] 7.22
>Sd
[2] 0.8230345
> summary(T)
```

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
4.80    6.65    7.20    7.22    8.15    9.00
```

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 = 6.65 có nghĩa là 25% học sinh có điểm toán bằng hoặc thấp hơn 6.65. Tương tự Third quartile = 8.15 có nghĩa là 75% học sinh có điểm toán bằng hoặc thấp hơn 8.15. Trung vị (Median) = 7.2 có nghĩa là 50% học sinh có điểm toán 7,2 trở xuống (hay 7.2 trở lên).

Ví dụ 1.20. Chúng ta có thể tóm tắt tổng quan toàn bộ file dữ liệu `Diem_TN`, như sau:

```
summary(Diem_TN)
# Kết quả hiển thị:
       ma          gioitinh              T
 Min.   : 1.00   Length:30          Min.   :4.80
 1st Qu.: 8.25   Class :character   1st Qu.:6.65
 Median :15.50   Mode  :character   Median :7.20
 Mean   :15.50                      Mean   :7.22
 3rd Qu.:22.75                      3rd Qu.:8.15
 Max.   :30.00                      Max.   :9.00
       A              V
 Min.   :4.60   Min.   :5.500
 1st Qu.:6.40   1st Qu.:6.500
 Median :7.20   Median :7.250
 Mean   :6.98   Mean   :7.167
 3rd Qu.:8.00   3rd Qu.:8.000
 Max.   :9.00   Max.   :8.500
```

R tính toán tất cả các biến số nào có thể tính toán được. Cột mã (tức mã số của đối tượng nghiên cứu) R cũng tính. Đối với các biến số mang tính phân loại như gioitinh thì R chỉ báo cáo tần số.

Ví dụ 1.21. Nếu chúng ta muốn kết quả cho từng nhóm nam và nữ riêng biệt, hàm `by()` trong R rất hữu dụng. Trong lệnh sau đây, chúng ta sử dụng R tóm lược dữ liệu Diem_TN theo gioitinh.

```
by(Diem_TN,gioitinh,summary)
# kết quả hiển thị:
Diem_TN$gioitinh: Nam
       ma         gioitinh               T
 Min.   : 1.0   Length:15          Min.   :6.000
 1st Qu.: 9.0   Class :character   1st Qu.:6.800
 Median :15.0   Mode  :character   Median :7.200
 Mean   :16.0                      Mean   :7.307
 3rd Qu.:23.5                      3rd Qu.:8.100
 Max.   :30.0                      Max.   :8.200
       A               V
 Min.   :4.600   Min.   :5.500
 1st Qu.:5.400   1st Qu.:6.500
 Median :6.400   Median :7.500
 Mean   :6.387   Mean   :7.033
 3rd Qu.:7.100   3rd Qu.:7.500
 Max.   :8.600   Max.   :8.000
----------------------------------------
Diem_TN$gioitinh: Nu
       ma       gioitinh              T
 Min.   : 2   Length:15          Min.   :4.800
 1st Qu.: 8   Class :character   1st Qu.:6.600
 Median :16   Mode  :character   Median :7.000
 Mean   :15                      Mean   :7.133
 3rd Qu.:22                      3rd Qu.:7.900
 Max.   :29                      Max.   :9.000
       A               V
 Min.   :4.800   Min.   :6.00
 1st Qu.:7.200   1st Qu.:6.75
 Median :8.000   Median :7.00
 Mean   :7.573   Mean   :7.30
 3rd Qu.:8.200   3rd Qu.:8.00
 Max.   :9.000   Max.   :8.50
```

Ví dụ 1.22. Xét file dữ liệu Diem_TN, nếu chúng ta muốn tính trung bình của một biến số như điểm toán (T) cho mỗi nhóm nam và nữ, hàm tapply trong R có thể dùng cho việc này:

```
tapply(Diem_TN$T, list(Diem_TN$gioitinh), mean)
# Kết quả hiển thị:
  Nam       Nu
7.306667 7.133333
```

Trong lệnh trên, T là biến số chúng ta cần tính, biến số phân nhóm là gioitinh, và chỉ số thống kê chúng ta muốn là trung bình (mean). Qua kết quả trên, chúng ta thấy điểm toán (T) trung bình của nam (7,3) cao hơn nữ (7,13).

1.2.2. Thống kê suy diễn trong các bài toán kiểm định

a) Trị số P-value

Trong nghiên cứu khoa học, ngoài những dữ kiện bằng số, biểu đồ và hình ảnh, con số mà chúng ta thường hay gặp nhất là trị số P (P-value). Do đó, trước khi nói đến các phương pháp phân tích thống kê bằng R, chúng ta cùng tìm hiểu về ý nghĩa của trị số này. Một giả thiết được xem là mang tính "khoa học" nếu giả thiết đó có khả năng "phản nghiệm". Theo Karl Popper, nhà triết học khoa học, đặc điểm duy nhất để có thể phân biệt giữa một lí thuyết khoa học thực thụ với ngụy khoa học (pseudoscience) là thuyết khoa học luôn có đặc tính có thể "bị bác bỏ" (hay bị phản bác – falsified) bằng những thực nghiệm đơn giản. Ông gọi đó là "khả năng phản nghiệm". Phép phản nghiệm là phương cách tiến hành những thực nghiệm không phải để xác minh mà để phê phán các lí thuyết khoa học và có thể coi đây như là một nền tảng cho khoa học thực thụ.

Có thể xem qui trình phản nghiệm là một cách học hỏi từ sai lầm. Khoa học phát triển cũng một phần lớn là do học hỏi từ sai lầm. Có thể xác định nghiên cứu khoa học như là một qui trình thử nghiệm giả thuyết, theo các bước sau đây:

- *Bước 1:* Nhà nghiên cứu cần phải định nghĩa một giả thuyết đảo (null hypothesis), tức là một giả thuyết ngược lại với những gì mà nhà nghiên cứu tin là sự thật.
- *Bước 2:* Nhà nghiên cứu cần phải định nghĩa một giả thuyết phụ (alternative hypothesis), tức là một giả thuyết mà nhà nghiên cứu nghĩ là sự thật, và điều cần được "chứng minh" bằng dữ kiện.
- *Bước 3:* Sau khi đã thu thập đầy đủ những dữ kiện liên quan, nhà nghiên cứu dùng một hay nhiều phương pháp thống kê để kiểm tra trong hai giả thuyết trên, giả thuyết nào được xem là khả dĩ. Cách kiểm tra này được tiến hành để trả lời câu hỏi: nếu giả thuyết đảo đúng, thì xác suất mà những dữ kiện thu thập được phù hợp với giả thuyết đảo là bao nhiêu. Giá trị của xác suất này thường được đề cập đến trong các báo cáo khoa học bằng kí hiệu "P value". Điều cần chú ý ở đây là nhà nghiên cứu không thử nghiệm giả thuyết khác, mà chỉ thử nghiệm giả thuyết đảo.
- *Bước 4:* Quyết định chấp nhận hay loại bỏ giả thuyết đảo bằng cách dựa vào giá trị xác suất trong bước thứ ba. Theo truyền thống nếu giá trị xác suất nhỏ hơn 5% thì nhà nghiên cứu sẵn sàng bác bỏ giả thuyết đảo. Tuy nhiên, nếu giá trị xác suất cao

hơn 5% thì nhà nghiên cứu chỉ có thể phát biểu rằng chưa có bằng chứng đầy đủ để bác bỏ giả thuyết đảo, và điều này không có nghĩa rằng giả thuyết đảo là đúng là sự thật.

- *Bước 5:* Nếu giả thuyết đảo bị bác bỏ, thì nhà nghiên cứu mặc nhiên thừa nhận giả thuyết phụ.

Chúng ta có thể tóm tắt tiến trình của một nghiên cứu (dựa vào trị số P) như sau:

- Đề ra một giả thuyết chính (H_1).
- Từ giả thuyết chính, đề ra một giả thuyết đảo (H_0).
- Tiến hành thu thập dữ kiện (D).
- Phân tích dữ kiện: tính toán xác suất D xảy ra nếu H_0 là sự thật. Nói theo ngôn ngữ toán xác suất, bước này xác định $P(D|H_0)$.

Vì thế, giá trị P có nghĩa là xác suất của dữ kiện D xảy ra nếu giả thuyết đảo H_0 là sự thật. Như vậy, giá trị P không trực tiếp cho chúng ta một ý niệm gì về sự thật của giả thuyết chính H_1; nó chỉ gián tiếp cung cấp bằng chứng để chúng ta chấp nhận giả thuyết chính và bác bỏ giả thuyết đảo.

b) Các loại sai lầm trong kiểm định giả thuyết

Sai lầm loại I: Nếu ta bác bỏ H_0 khi H_0 đúng thì sai lầm đó gọi là sai lầm loại I.

Sai lầm loại II: Nếu H_0 sai mà ta không bác bỏ H_0 thì sai lầm đó gọi là sai lầm loại II.

c) Kiểm định t (`t.test`)

Kiểm định t dựa vào giả thiết phân phối chuẩn. Có hai loại kiểm định t: kiểm định t cho một mẫu (one-sample t-test), và kiểm định t cho hai mẫu (two-sample t-test). Chúng ta sẽ minh họa hai kiểm định này qua số liệu của file Diem_TN

- ***Kiểm định giả thuyết cho kỳ vọng một mẫu***

Xét mẫu ngẫu nhiên $x_1, x_2, ..., x_n$ được chọn từ tổng thể có phân phối chuẩn (hoặc xấp xỉ chuẩn tức phân phối có dạng đối xứng) với kỳ vọng a và phương sai σ^2.

Giả thuyết H_0: $a = a_0$; Đối thuyết H_1: $\begin{cases} a \neq a_0 \\ a < a_0 \\ a > a_0 \end{cases}$ (Một trong 3 trường hợp)

Tính thống kê kiểm định: $t = \dfrac{\bar{x} - a_0}{s}.\sqrt{n}$.

Miền bác bỏ:

- Với H_1: $a \neq a_0$, bác bỏ H_0 nếu $t < -t_{1-\alpha/2}^{n-1}$ hoặc $t > t_{1-\alpha/2}^{n-1}$.
- Với H_1: $a < a_0$, bác bỏ H_0 nếu $t < -t_{1-\alpha}^{n-1}$.
- Với H_1: $a > a_0$, bác bỏ H_0 nếu $t > t_{1-\alpha}^{n-1}$.

Trong `R`, để tìm phân vị $t_{1-\alpha/2}^{n-1}$ sử dụng hàm `qt(1-alpha/2,n-1).`

Trong kết quả do `R` xuất ra, ta xác định có bác bỏ H_0 hay không thông qua P- giá trị.

Quy tắc: Khi P- giá trị bé hơn α thì bác bỏ H_0.

Khi cỡ mẫu n lớn, phân phối của thống kê t sẽ xấp xỉ phân phối chuẩn hóa N(0,1), khi đó giá trị tiêu chuẩn dùng để so sánh là $z_{1-\alpha/2}$ (dùng `qnorm(1-alpha/2)).`

Sử dụng hàm `t.test` để kiểm định theo cú pháp:

```
t.test(x,alternative= c("two.sided", "less", "greater"),mu = mu_0, conf.level = 0.95)
# Trong đó:
```

- x: véc tơ dữ liệu.
- alternative: xác định kiểm định là hai phía ("two.sided"), bên trái ("less") hay bên phải ("greater"), mặc định là two.sided.
- mu = mu_0: giá trị cần kiểm định.
- conf.level: xuất ra khoảng tin cậy với độ tin cậy tương ứng.

Ví dụ 1.23. Trong file dữ liệu Diem_TN, ta thấy điểm toán (T) trung bình của 30 học sinh là 7,22. Chúng ta sẽ kiểm định điểm toán của học sinh toàn trường có thực sự thấp hơn 8 hay không với mức ý nghĩa 5%? Đồng thời xác định khoảng tin cậy 95% cho điểm toán trung bình của toàn trường với câu lệnh"

```
t.test(Diem_TN$T, mu = 8, conf.level = 0.95)
Kết quả hiển thị:
One Sample t-test
data:  Diem_TN$T
t = -4.7092, df = 29, p-value = 5.689e-05
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 6.881241 7.558759
sample estimates:
mean of x
     7.22
```

Từ kết quả trên ta thu được"

- Thống kê kiểm định t = -4.7092, bậc tự do n – 1 = 29, p-value = 5.689e-05.
- Khoảng tin cậy 95%: $6.881241 \le a \le 7.558759$.
- Với mức ý nghĩa 5%, ta thấy p-value (rất thấp) < 0.05, do đó bác bỏ H_0 tức điểm trung bình toán của học sinh toàn trường thấp hơn 8.

Nếu sử dụng giá trị thống kê t = -4.7092, ta so sánh với $t_{1-\alpha/2}^{n-1} = t_{0,975}^{29} = 2.045$ (dùng lệnh `qt(0.975, 29)`) ta cũng có kết luận tương tự.

- ***Kiểm định giả thuyết cho tỷ lệ một mẫu***

Giả sử cần kiểm định tỷ lệ phần tử thỏa mãn tính chất A trong tổng thể. Khảo sát một cỡ mẫu n. Gọi m là tổng số phần tử thỏa mãn tính chất A trong n phần tử khảo sát, suy ra tỷ lệ mẫu: $f = \frac{m}{n}$. Giả thuyết cỡ mẫu khảo sát n phải tương đối lớn.

Giả thuyết: $H_0: p = p_0$; Đối thuyết: $H_1: \begin{cases} p \neq p_0 \\ p < p_0 \\ p > p_0 \end{cases}$ (Một trong 3 trường hợp).

Tính thống kê kiểm định: $u = \frac{f - p_0}{\sqrt{p_0.(1-p_0)}}.\sqrt{n}$.

Miền bác bỏ:

- Với H_1: $p \neq p_0$ bác bỏ H_0 nếu $u < -z_{1-\alpha/2}$ hoặc $u > z_{1-\alpha/2}$.
- Với H_1: $p < p_0$ bác bỏ H_0 nếu $u < -z_{1-\alpha}$.
- Với H_1: $p > p_0$ bác bỏ H_0 nếu $u > z_{1-\alpha}$.

Để tìm $z_{1-\alpha/2}$, sử dụng hàm qnorm(1-alpha/2).

Sử dụng hàm prop.test để kiểm định:

```
prop.test(m, n, p = p0, alternative = c("two.sided", "less", "greater"), conf.level = 0.95)
# trong đó:
```

- m: số phần tử thỏa mãn tính chất A trong n phần tử khảo sát.
- n: cỡ mẫu.
- alternative: xác định kiểm định là hai phía ("two.sided"), bên trái ("less") hay bên phải ("greater").
- p = p0: giá trị cần kiểm định.
- conf.level: xuất ra khoảng tin cậy với độ tin cậy tương ứng.

Ví dụ 1.24. Trong một cuộc bầu cử thị trưởng tại một thành phố, ứng cử viên A tin rằng có trên 50% người dân thành phố ủng hộ ông ta. Để kiểm định điều này, các chuyên gia thống kê chọn ngẫu nhiên 800 người dân trong thành phố, thấy có 448 người dân cho ý kiến ủng hộ ông A. Hãy nhận xét xem tuyên bố của ông A về tỷ lệ cử tri có đúng không với mức ý nghĩa 1%?

Ta có:

Cỡ mẫu khảo sát $n = 800$.

Số người dân ủng hộ ông A: $m = 448$.

Giả thuyết cần kiểm tra: $\begin{cases} H_0 : p = 0.5 \\ H_1 : p > 0.5 \end{cases}$

Trong đó p là tỷ lệ người dân thành phố ủng hộ ông A.

Sử dụng hàm `prop.test`:

```
> n=800;m=448
> prop.test(m,n,p=0.5,alternative="greater",conf.level=0.99)
Kết quả hiển thị:
1-sample proportions test with continuity
correction
data:  m out of n, null probability 0.5
X-squared = 11.281, df = 1, p-value = 0.0003915
alternative hypothesis: true p is greater than 0.5
99 percent confidence interval:
 0.5182781 1.0000000
sample estimates:
   p
0.56
```

Kết quả cho biết p-value = 0.0003915 <1% dẫn đến bác bỏ giả thuyết H_0, ta kết luận rằng tỷ lệ người dân ủng hộ ông A trong thành phố trên 50%.Khoảng tin cậy 99% cho tỷ lệ p là: $0.5182 \le p \le 1.0000$.

- ***Kiểm định trung bình hai mẫu***

Ví dụ 1.25. Xét dữ liệu Diem_TN, qua phân tích mô tả chúng ta thấy nam có điểm toán (T) trung bình (7.3) cao hơn nữ (7.1). Câu hỏi đặt ra là có phải thật sự điểm toán trung bình của nam và nữ khác nhau hay không.

Tính thống kê kiểm định: $t = \frac{\overline{x_2} - \overline{x_1}}{SED}$. Trong đó $\overline{x_1}$ và $\overline{x_2}$ là điểm toán trung bình của hai nhóm nam và nữ, và SED là độ lệch chuẩn của $(\overline{x_1} - \overline{x_2})$. SED có thể tính bằng công thức: $SED = \sqrt{SE_1^2 + SE_2^2}$ .Trong đó SE_1, SE_2 là sai số chuẩn (standard error) của hai nhóm nam và nữ. Theo lí thuyết xác suất, t tuân theo luật phân phối t với bậc tự do $n_1 + n_2 - 2$, trong đó n_1, 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(T~gioitinh)
# kết quả hiển thị:
        Welch Two Sample t-test
data:  T by gioitinh
t = 0.51659, df = 24.495, p-value = 0.6101
alternative hypothesis: true difference in means between group Nam and group Nu is not
equal to 0
95 percent confidence interval:
 -0.5184389  0.8651055
sample estimates:
mean in group Nam  mean in group Nu
         7.306667          7.133333
```

R trình bày các giá trị quan trọng trước hết:

```
t = 0.51659, df = 24.495, p-value = 0.6101
```

df là bậc tự do. Trị số p = 0.6101 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:
-0.5184389  0.8651055
```

Là khoảng tin cậy 95% về độ khác biệt giữa hai nhóm. Kết quả tính toán trên cho biết điểm toán trung bình của nữ có thể thấp hơn nam giới 0.52 hoặc cao hơn nam giới 0.86. 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(T~gioitinh, var.equal = TRUE)
Kết quả hiển thị:
        Two Sample t-test
data:  T by gioitinh
t = 0.51659, df = 28, p-value = 0.6095
alternative hypothesis: true difference in means between group Nam and group Nu is not
equal to 0
95 percent confidence interval:
 -0.5139801  0.8606468
sample estimates:
mean in group Nam  mean in group Nu
         7.306667          7.133333
```

Về mặt số liệu, 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.

- ***Kiểm định tỷ lệ hai mẫu***

Cho hai mẫu với số đối tượng n_1 và n_2, gọi số phần tử thỏa mãn tính chất A trong mẫu 1 là m_1, trong mẫu 2 là m_2. Do đó, chúng ta có thể tính được tỉ lệ tương ứng trong hai mẫu là p_1, 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:

$V_d = (\frac{1}{n_1} + \frac{1}{n_2})\mathrm{p}(1-\mathrm{p})$. Trong đó: $\mathrm{p} = \frac{m_1 + m_2}{n_1 + n_2}$; $z = d / V_d$ tuân theo luật phân phối chuẩn với trung bình 0 và phương sai 1.

Ví dụ 1.26. 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, nhóm B có 20 người gãy xương. Hỏi tỉ lệ gãy xương trong hai nhóm có bằng nhau (tức thuốc không có hiệu quả)?

Để kiểm định hai tỉ lệ này có thật sự khác nhau, chúng ta có thể sử dụng hàm `prop.test(x,n,π)` như sau:

```
> m<-c(7,20)
> n<-c(100,110)
> prop.test(m,n)
Kết quả hiển thị:
2-sample test for equality of proportions with continuity
        correction
data:  m out of n
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 là 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.

c) 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.

Ví dụ 1.26. Trong *Ví dụ 1.25.* chúng thấy trong file dữ liệu `Diem_TN` điểm toán (`T`) không có 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(T~gioitinh)
# Kết quả hiển thị:
Wilcoxon rank sum test with continuity correction
data:  T by gioitinh
W = 121, p-value = 0.7383
alternative hypothesis: true location shift is not equal to 0
```

Trị số p = 0.7383 cho thất quả thật độ khác biệt về điểm toán giữa hai nhóm nam và nữ không có ý nghĩa thống kê.

d) So sánh phương sai (`var.test`)

Ví dụ 1.28. Sử dụng file dữ liệu `Diem_TN`, để kiểm định phương sai điểm toán (`T`) giữa hai nhóm nam và nữ có khác nhau không, ta dùng câu lệnh sau:

```
> var.test(T~gioitinh)
Kết quả hiển thị:
F test to compare two variances
data:  T by gioitinh
F = 0.45106, num df = 14, denom df = 14, p-value = 0.1485
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1514355 1.3435331
sample estimates:
ratio of variances
         0.4510638
```

Kết quả trên cho thấy độ khác biệt về phương sai giữa hai nhóm là 0.45 lần. Trị số p = 0.1485 cho thấy phương sai giữa hai nhóm khác nhau không có ý nghĩa thống kê.

e) Thủ tục kiểm định shapiro.test về phân phối chuẩn

Để kiểm định một luật phân phối mẫu xem liệu có tuân theo luật chuẩn hay không, chúng ta có thể sử dụng hàm shapiro.test có cấu trúc như sau:

```
shapiro.test(x)

trong đó: x: là dữ liệu mẫu
```

Ví dụ 1.29. Theo dõi năng suất café tại một nông trường thu được số liệu sau:

Năng suất (tạ/ha)	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22
Số vườn	15	26	25	30	26	21	24	20	13

Với mức ý nghĩa 0,05 có thể kết luận năng suất café ở nông trường này tuân theo luật phân phối chuẩn?

```
dl<-rep(c(5,7,9,11,13,15,17,19,21),times=c(15,26,30,26,21,24,20,13))
shapiro.test(dl)
Kết quả hiển thị:
Shapiro-Wilk normality test
data:  dl
W = 0.94668, p-value = 9.158e-07
```

Qua bảng kết quả hiện thị ta thấy trị số W = 0.94668và trị số p-value = 9.158e-07. Vì vậy, kết quả nhận được là qua mẫu cụ thể chưa thể khẳng định năng suất tuân theo quy luật chuẩn.

1.2.3. Thống kê suy diễn trong các bài toán phân tích tương quan

Hệ số tương quan (r) là một chỉ số thống kê đo lường mối liên hệ tương quan giữa hai biến số. Hệ số tương quan có giá trị từ -1 đến 1. Hệ số tương quan bằng 0 (hay gần 0) có nghĩa là hai biến số không có liên hệ gì với nhau; ngược lại nếu hệ số bằng -1 hay 1 có nghĩa là hai biến số có một mối liên hệ tuyệt đối. Nếu giá trị của hệ số tương quan là âm ($r < 0$) có nghĩa là hai biến tương quan nghịch (biến này tăng thì biến kia giảm và ngược lại); nếu giá trị hệ số tương quan là dương ($r > 0$) có nghĩa là hai biến tương quan thuận (hai biến cùng tăng hoặc cùng giảm).

Có nhiều hệ số tương quan trong thống kê, nhưng ở đây chúng ta sẽ trình bày 3 hệ số tương quan thông dụng nhất: hệ số tương quan Pearson r, Spearman ρ, và Kendall τ.

Trong tiểu mục này dữ liệu dùng để minh họa là file dữ liệu markettimng.csv tham khảo từ link: https://drive.google.com/drive/folders/1maNUAWyCcjXrU0m6hMgZNhjEI0jUI9Gu

```
library(readr)
marketing <- read_csv "marketing.csv"
head(marketing)
# kết quả hiển thị
```

```
1 youtube facebook newspaper sales
<dbl>   <dbl>    <dbl>     <dbl> <dbl>
1     1    201.    142.      145.   943.
2     2    156.    130.       62.7  856.
3     3    124.    188.      140.   965.
4     4    158.    187.      144.  1017.
5     5    158.    222.      116.  1115.
6     6    132.    182.      120.   932.
7     7    121.    214.      144.  1022.
8     8    108.     82.6     126.   650.
9     9    190.    173.      104.  1001.
10    10    117.    115.      133.   713.
# ... with 190 more rows
```

Bảng 1.4: Dữ liệu quan sát số lượt quảng cáo, (nguồn: internet).

a) Hệ số tương quan mẫu

- ***Hệ số tương quan Pearson***

Cho hai biến số x và y từ n mẫu, hệ số tương quan Pearson được tính bằng công thức sau đây:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Trong đó, $\bar{x}$ và $\bar{y}$ là giá trị trung bình của biến số x và y.

Để tính hệ số tương quan Pearson trong R, cú pháp như sau:

```
cor(data, method = "pearson")
```

Ví dụ 1.30. Sử dụng file dữ liệu markettimng.csv, ta sẽ tính hệ số tương quan pearson giữa các biến trong đó:

```
cor(marketing, method = "pearson")
# kết quả hiển thị:
               ...1     youtube    facebook    newspaper        sales
...1       1.00000000 -0.04977015 -0.03918551 -0.177473371 -0.047532656
youtube   -0.04977015  1.00000000  0.08401121  0.047806059  0.487083735
facebook  -0.03918551  0.08401121  1.00000000 -0.039579633  0.903092760
newspaper -0.17747337  0.04780606 -0.03957963  1.000000000 -0.002900308
sales     -0.04753266  0.48708374  0.90309276 -0.002900308  1.000000000
```

Qua kết quả trên ta thấy: giữa biến sales và facebook có tương quan rất mạnh 0.903, còn lại đều có mức tương quan yếu giữa các biến khác.

- ***Hệ số tương quan Spearman ρ***

Hệ số tương quan Pearson chỉ hợp lí nếu biến số x và y tuân theo luật phân phối chuẩn. Nếu x và y không tuân theo luật phân phối chuẩn, chúng ta phải sử dụng một hệ số tương quan khác tên là Spearman, một phương pháp phân tích phi tham số. Hệ số này được ước tính bằng cách biến đổi hai biến số x và y thành thứ bậc (rank), và xem độ tương quan giữa hai dãy số bậc. Do đó, hệ số còn có tên tiếng Anh là Spearman's Rank correlation.

Để tính hệ số tương quan spearman trong R, cú pháp như sau:

```
cor(data, method = " spearman ")
```

Ví dụ 1.31. Sử dụng file dữ liệu markettimng.csv, ta sẽ tính hệ số tương quan spearman giữa các biến trong đó:

```
cor(marketing, method = "spearman")
# kết quả hiển thị:
                ...1      youtube     facebook     newspaper         sales
...1        1.000000000 -0.007446197 -0.06512643 -0.1753965822 -0.0706097652
youtube    -0.007446197  1.000000000  0.06036389  0.0620604644  0.4422117186
facebook   -0.065126427  0.060363895  1.00000000 -0.0332461435  0.8994231601
newspaper  -0.175396582  0.062060464 -0.03324614  1.0000000000 -0.0004522618
sales      -0.070609765  0.442211719  0.89942316 -0.0004522618  1.0000000000
```

Kết quả phân tích cũng tương tự như trong Ví dụ 1.29.

- ***Hệ số tương quan Kendall τ***

Hệ số tương quan Kendall (cũng là một phương pháp phân tích phi tham số) được ước tính bằng cách tìm các cặp số (x, y)"song hành" với nhau. Một cặp (x, y) song hành ở đây được định nghĩa là hiệu (độ khác biệt) trên trục hoành có cùng dấu hiệu (dương hay âm) với hiệu trên trục tung. Nếu hai biến số x và y không có liên hệ với nhau, thì cặp số song hành bằng hay tương đương với cặp số không song hành.

Vì có nhiều cặp phải kiểm định, phương pháp tính toán hệ số tương quan Kendall đòi hỏi thời gian của máy tính khá cao. Tuy nhiên, nếu một dữ liệu dưới 5000 đối tượng thì một máy vi tính có thể tính toán khá dễ dàng.

Để tính hệ số tương quan Kendall trong R, cú pháp như sau:

```
cor(data, method = " kendall ")
```

Ví dụ 1.32. Sử dụng file dữ liệu markettimng.csv, ta sẽ tính hệ số tương quan kendall giữa các biến trong đó:

```
cor(marketing, method = "  kendall")
# kết quả hiển thị:
             ...1      youtube     facebook    newspaper        sales
...1       1.000000000 -0.005025631 -0.04532891 -0.11734553 -0.04874372
```

```
youtube    -0.005025631  1.000000000  0.03789516  0.04829995  0.30324656
facebook   -0.045328911  0.037895160  1.00000000 -0.02326925  0.72656918
newspaper  -0.117345529  0.048299952 -0.02326925  1.00000000 -0.00587984
sales      -0.048743719  0.303246559  0.72656918 -0.00587984  1.00000000
```

b) Kiểm định hệ số tương quan

Bên cạnh việc tính các giá trị tương quan mẫu, chúng ta cũng có thể kiểm định hệ số tương quan lý thuyết với giả thuyết kiểm định:

- H0: Không có tương quan (hệ số tương quan = 0).
- H1: Có tương quan.

Để tính kiểm định trong R, cú pháp như sau:

```
cor.test(nhân tố 1, nhân tố 2, method = c("pearson","spearman", "kendall"))
```

Trong đó:

Nhân tố 1, nhân tố 2 là 2 biến cần kiểm định tính tương quan.

methol được lựa chọn một trong 3 phương pháp tương ứng.

Ví dụ 1.33. Sử dụng file dữ liệu markettimng.csv, ta sẽ kiểm định tính tương quan giữa 2 biến sales và youtube:

```
> cor.test(marketing$youtube, marketing$sales)
#Kết quả hiển thị:
        Pearson's product-moment correlation
data:  marketing$youtube and marketing$sales
t = 7.8478, df = 198, p-value = 2.597e-13
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3735893 0.5862096
sample estimates:
      cor
0.4870837
```

Kết quả trên với trị số p-value = 2.597e-13, cho thấy mối liên hệ giữa doanh thu (sales) và chi phí quảng cáo qua youtube là có ý nghĩa thống kê.

1.3. CÁC BƯỚC TRỰC QUAN HÓA DỮ LIỆU

Khi trực quan hóa dữ liệu, công việc đầu tiên là lựa chọn loại biểu đồ để thể hiện dữ liệu đang phân tích. Công việc này không chỉ là chọn một loại hay một số loại biểu đồ, mà có thể là lựa chọn kết hợp các biểu đồ, kết hợp biểu đồ và bảng phân tích thống kê mô tả, khi đó hiệu quả mang lại sẽ tối ưu nhất.

Việc lựa chọn biểu đồ phù hợp để trực quan dữ liệu phụ thuộc vào nhiều yếu tố, bao gồm loại dữ liệu, mục tiêu truyền đạt thông điệp, số lượng và phân loại của các biến dữ liệu, cũng như sở thích cá nhân. Hình 1.32 dưới đây là một số hướng dẫn để chọn biểu đồ phù hợp: