THỐNG KÊ ỨNG DỤNG

Đỗ Lân

dolan@tlu.edu.vn
Đại học Thủy Lợi

Ngày 6 tháng 11 năm 2018

Nội dung môn học

1. Tổng quan về Thống kê
2. Thu thập dữ liệu
3. Tóm tắt và trình bày dữ liệu bằng bảng và đồ thị
4. Tóm tắt dữ liệu bằng các đại lượng thống kê mô tả
5. Xác suất căn bản và biến ngẫu nhiên
6. Phân phối của tham số mẫu và ước lượng tham số tổng thể
7. **Kiểm định giả thuyết về tham số một tổng thể**
8. Kiểm định giả thuyết về tham số hai tổng thể
9. Phân tích phương sai
10. Kiểm định phi tham số
11. Kiểm định chi - bình phương
12. Hồi quy đơn biến
13. Hồi quy đa biến

Phần VII

Kiểm định giả thuyết thống kê

Mục lục

1. Giả thuyết thống kê
2. Logic của bài toán kiểm định
3. Các loại sai lầm có thể mắc phải
4. Kiểm định giả thuyết so sánh trung bình một tổng thể với một số
   - Trường hợp tổng thể tuân theo phân phối chuẩn, biết phương sai
     - P-giá trị
   - Kiểm định trung bình tổng thể bất kì khi cỡ mẫu lớn
   - Kiểm định trung bình tổng thể khi tổng thể có phân phối chuẩn
   - Thực tế

Bài toán

Giả sử một báo cáo được công bố cho thấy thu nhập trung bình mỗi tháng của người lao động ở thành phố của bạn là 5 triệu đồng mỗi tháng. Đây là cơ sở để họ xác định mức đánh thuế thu nhập cao. Bạn muốn kiểm tra xem báo cáo trên có chấp nhận được không. Làm thế nào để bạn có thể làm được điều này?

Bài toán

Giả sử một báo cáo được công bố cho thấy thu nhập trung bình mỗi tháng của người lao động ở thành phố của bạn là 5 triệu đồng mỗi tháng. Đây là cơ sở để họ xác định mức đánh thuế thu nhập cao. Bạn muốn kiểm tra xem báo cáo trên có chấp nhận được không. Làm thế nào để bạn có thể làm được điều này?

Solution

1. *Xác định bạn muốn kiểm định cái gì?* $\rightarrow$ *Giả thuyết thống kê.*

Bài toán

Giả sử một báo cáo được công bố cho thấy thu nhập trung bình mỗi tháng của người lao động ở thành phố của bạn là 5 triệu đồng mỗi tháng. Đây là cơ sở để họ xác định mức đánh thuế thu nhập cao. Bạn muốn kiểm tra xem báo cáo trên có chấp nhận được không. Làm thế nào để bạn có thể làm được điều này?

Solution

1. *Xác định bạn muốn kiểm định cái gì?* $\rightarrow$ *Giả thuyết thống kê.*
2. *Chọn mẫu.*

Bài toán

Giả sử một báo cáo được công bố cho thấy thu nhập trung bình mỗi tháng của người lao động ở thành phố của bạn là 5 triệu đồng mỗi tháng. Đây là cơ sở để họ xác định mức đánh thuế thu nhập cao. Bạn muốn kiểm tra xem báo cáo trên có chấp nhận được không. Làm thế nào để bạn có thể làm được điều này?

Solution

1. *Xác định bạn muốn kiểm định cái gì?* $\rightarrow$ *Giả thuyết thống kê.*
2. *Chọn mẫu.*
3. *Đại lượng cần kiểm định tuân theo phân phối nào?*

Bài toán

Giả sử một báo cáo được công bố cho thấy thu nhập trung bình mỗi tháng của người lao động ở thành phố của bạn là 5 triệu đồng mỗi tháng. Đây là cơ sở để họ xác định mức đánh thuế thu nhập cao. Bạn muốn kiểm tra xem báo cáo trên có chấp nhận được không. Làm thế nào để bạn có thể làm được điều này?

Solution

1. *Xác định bạn muốn kiểm định cái gì?* $\rightarrow$ *Giả thuyết thống kê.*
2. *Chọn mẫu.*
3. *Đại lượng cần kiểm định tuân theo phân phối nào?*
4. *Với xác suất sai lầm cho trước, căn cứ trên phân phối trên tính toán để đưa ra mốc quyết định.*

Khái niệm

Giả thuyết thống kê *là một kết luận tạm thời được đưa ra về phân phối của tổng thể, về các tham số của tổng thể hoặc về tính độc lập của các đặc điểm trong tổng thể.*

Cặp giả thuyết

Giả thuyết thống kê luôn bao gồm: Giả thuyết không H_0 (null hypothesis) và giả thuyết đối H_1 (alternative hypothesis). Trong đó:

- Trong H_0 và H_1 phải có giả thuyết đang cần được kiểm định;

Cặp giả thuyết

Giả thuyết thống kê luôn bao gồm: Giả thuyết không H_0 (null hypothesis) và giả thuyết đối H_1 (alternative hypothesis). Trong đó:

- Trong H_0 và H_1 phải có giả thuyết đang cần được kiểm định;
- Trong H_0 luôn có một dấu bằng: $=, \leq, \geq$.

Cặp giả thuyết

Giả thuyết thống kê luôn bao gồm: Giả thuyết không H_0 (null hypothesis) và giả thuyết đối H_1 (alternative hypothesis). Trong đó:

- Trong H_0 và H_1 phải có giả thuyết đang cần được kiểm định;
- Trong H_0 luôn có một dấu bằng: $=, \leq, \geq$.
- Trong H_1 không có dấu bằng: $\neq, <, >$.

Example

Một người cho rằng tỉ lệ sinh viên TLU ra trường xin được việc ngay trong năm đầu tiên 60%. Bạn muốn kiểm định điều này, cặp giả thiết sẽ là gì?

Example

Một người cho rằng tỉ lệ sinh viên TLU ra trường xin được việc ngay trong năm đầu tiên 60%. Bạn muốn kiểm định điều này, cặp giả thiết sẽ là gì?

Solution

- H_0 : *Tỉ lệ sinh viên TLU ra trường xin được việc làm ngay trong năm đầu tiên là 0.6.*

Example

Một người cho rằng tỉ lệ sinh viên TLU ra trường xin được việc ngay trong năm đầu tiên 60%. Bạn muốn kiểm định điều này, cặp giả thiết sẽ là gì?

Solution

- H_0 : *Tỉ lệ sinh viên TLU ra trường xin được việc làm ngay trong năm đầu tiên là 0.6.*
- H_1 : *Tỉ lệ sinh viên TLU ra trường xin được việc làm ngay trong năm đầu tiên khác 0.6.*

Example

Một nhận định của báo chí cho rằng thu nhập của sinh viên mới ra trường là thấp hơn mức trung bình của toàn xã hội. Biết rằng năm 2017 lương trung bình của dân Việt Nam là 55 triệu đồng/năm. Bạn muốn kiểm định xem điều khẳng định trên có đúng không. Cặp giả thiết của bạn là gì?

Example

Một nhận định của báo chí cho rằng thu nhập của sinh viên mới ra trường là thấp hơn mức trung bình của toàn xã hội. Biết rằng năm 2017 lương trung bình của dân Việt Nam là 55 triệu đồng/năm. Bạn muốn kiểm định xem điều khẳng định trên có đúng không. Cặp giả thiết của bạn là gì?

Solution

Gọi lương trung bình của sinh viên mới ra trường là μ triệu đồng/ năm.

Example

Một nhận định của báo chí cho rằng thu nhập của sinh viên mới ra trường là thấp hơn mức trung bình của toàn xã hội. Biết rằng năm 2017 lương trung bình của dân Việt Nam là 55 triệu đồng/năm. Bạn muốn kiểm định xem điều khẳng định trên có đúng không. Cặp giả thiết của bạn là gì?

Solution

Gọi lương trung bình của sinh viên mới ra trường là μ triệu đồng/ năm.

- $H_0 : \mu \geq 45$.

Example

Một nhận định của báo chí cho rằng thu nhập của sinh viên mới ra trường là thấp hơn mức trung bình của toàn xã hội. Biết rằng năm 2017 lương trung bình của dân Việt Nam là 55 triệu đồng/năm. Bạn muốn kiểm định xem điều khẳng định trên có đúng không. Cặp giả thiết của bạn là gì?

Solution

Gọi lương trung bình của sinh viên mới ra trường là μ triệu đồng/ năm.

- $H_0 : \mu \geq 45$.
- $H_1 : \mu < 45$.

Example

Một nhận định của báo chí cho rằng thu nhập của sinh viên mới ra trường là thấp hơn mức trung bình của toàn xã hội. Biết rằng năm 2017 lương trung bình của dân Việt Nam là 55 triệu đồng/năm. Bạn muốn kiểm định xem điều khẳng định trên có đúng không. Cặp giả thiết của bạn là gì?

Solution

Gọi lương trung bình của sinh viên mới ra trường là μ triệu đồng/ năm.

- $H_0 : \mu \geq 45$.
- $H_1 : \mu < 45$.

Example

Một cửa hàng bán điện thoại di động muốn xác định xem, sau khi thực hiện chiến dịch quảng cáo có giúp họ gia tăng lượng khách không. Trước đây, số khách trung bình mỗi ngày của họ là 30. Hiện tại, sau một tháng, mỗi ngày họ có khoảng 35 khách đến xem hàng. Cặp giả thuyết kiểm định của họ như thế nào?

Example

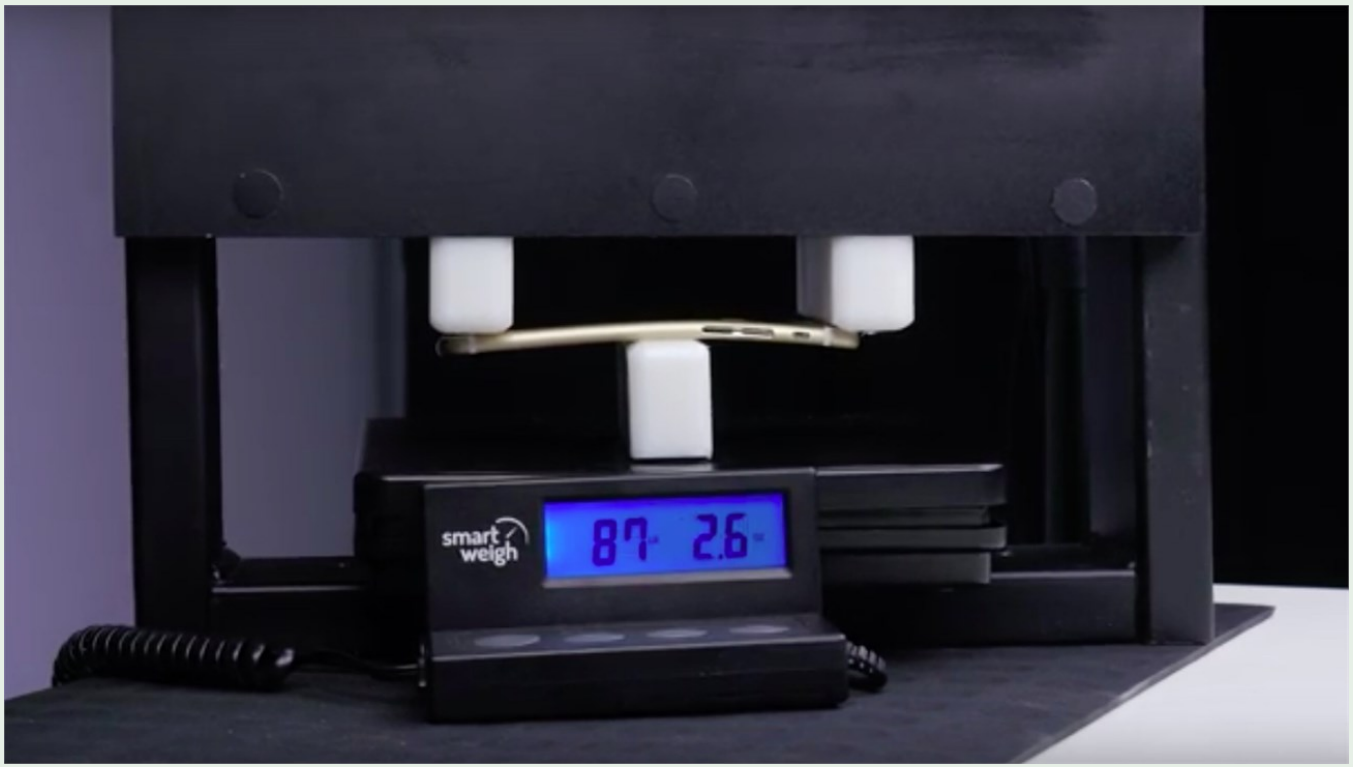
Một chủ hồ nuôi khi thương thảo bán lứa cá chim cho rằng lứa cá của họ có trọng lượng trung bình ít nhất là 2.5kg. Bạn là người đại diện công ty đến thu mua, và bạn muốn kiểm định điều ông chủ hồ nuôi nói, cặp giả thiết của bạn là gì?

Example

Một tổ chức vì quyền nữ giới cho rằng, ở Việt Nam, thu nhập trung bình của nữ giới thấp hơn so với nam giới. Bạn cảm thấy rằng không phải như vậy. Cặp giả thiết của bạn là gì?

Example

Một hãng sản suất điện thoại vừa cho ra đời một dòng điện thoại mới, họ nói rằng sức chịu lực trung bình của màn hình của dòng điện thoại này là ít nhất 90 pound.



Hãng cạnh tranh muốn kiểm định điều này. Cặp giả thiết của họ là gì?

Các tình huống

Gọi μ là trung bình của tổng thể, μ_0 là một số nào đấy thì có các tình huống sau:

Các tình huống

Gọi μ là trung bình của tổng thể, μ_0 là một số nào đấy thì có các tình huống sau:

1. $H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0.$

Các tình huống

Gọi μ là trung bình của tổng thể, μ_0 là một số nào đấy thì có các tình huống sau:

1. $H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0.$
2. $H_0 : \mu \leq \mu_0 \qquad H_1 : \mu > \mu_0.$

Các tình huống

Gọi μ là trung bình của tổng thể, μ_0 là một số nào đấy thì có các tình huống sau:

1. $H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0.$
2. $H_0 : \mu \leq \mu_0 \qquad H_1 : \mu > \mu_0.$
3. $H_0 : \mu = \mu_0 \qquad H_1 : \mu > \mu_0.$

Các tình huống

Gọi μ là trung bình của tổng thể, μ_0 là một số nào đấy thì có các tình huống sau:

1. $H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0.$
2. $H_0 : \mu \leq \mu_0 \qquad H_1 : \mu > \mu_0.$
3. $H_0 : \mu = \mu_0 \qquad H_1 : \mu > \mu_0.$
4. $H_0 : \mu \geq \mu_0 \qquad H_0 : \mu < \mu_0.$

Các tình huống

Gọi μ là trung bình của tổng thể, μ_0 là một số nào đấy thì có các tình huống sau:

1. $H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0.$
2. $H_0 : \mu \leq \mu_0 \qquad H_1 : \mu > \mu_0.$
3. $H_0 : \mu = \mu_0 \qquad H_1 : \mu > \mu_0.$
4. $H_0 : \mu \geq \mu_0 \qquad H_0 : \mu < \mu_0.$
5. $H_0 : \mu = \mu_0 \qquad H_0 : \mu < \mu_0.$

Các tình huống

Gọi μ là trung bình của tổng thể, μ_0 là một số nào đấy thì có các tình huống sau:

1. $H_0 : \mu = \mu_0 \qquad H_1 : \mu \neq \mu_0.$
2. $H_0 : \mu \leq \mu_0 \qquad H_1 : \mu > \mu_0.$
3. $H_0 : \mu = \mu_0 \qquad H_1 : \mu > \mu_0.$
4. $H_0 : \mu \geq \mu_0 \qquad H_0 : \mu < \mu_0.$
5. $H_0 : \mu = \mu_0 \qquad H_0 : \mu < \mu_0.$

Với những bài toán mà $H_1 : \mu > \mu_0$ ta gọi là bài toán bên phải, $H_1 : \mu < \mu_0$ ta gọi là bài toán bên trái, $H_1 : \mu \neq \mu_0$ gọi là bài toán 2 bên.

1. Giả thuyết thống kê
2. Logic của bài toán kiểm định
3. Các loại sai lầm có thể mắc phải
4. Kiểm định giả thuyết so sánh trung bình một tổng thể với một số
   - Trường hợp tổng thể tuân theo phân phối chuẩn, biết phương sai
     - P-giá trị
   - Kiểm định trung bình tổng thể bất kì khi cỡ mẫu lớn
   - Kiểm định trung bình tổng thể khi tổng thể có phân phối chuẩn
   - Thực tế

Bài toán

Giả sử một báo cáo được công bố cho thấy thu nhập trung bình mỗi tháng của người lao động ở một thành phố của bạn là 5 triệu đồng mỗi tháng. Đây là cơ sở để họ xác định mức đánh thuế thu nhập cao. Bạn muốn kiểm tra xem báo cáo trên có chấp nhận được không. Làm thế nào để bạn có thể kiểm định được điều này?

Bài toán

Giả sử một báo cáo được công bố cho thấy thu nhập trung bình mỗi tháng của người lao động ở một thành phố của bạn là 5 triệu đồng mỗi tháng. Đây là cơ sở để họ xác định mức đánh thuế thu nhập cao. Bạn muốn kiểm tra xem báo cáo trên có chấp nhận được không. Làm thế nào để bạn có thể kiểm định được điều này?

Solution

- H_0 : *Thu nhập trung bình của người lao động trong thành phố của bạn là 5 triệu mỗi tháng.*

Bài toán

Giả sử một báo cáo được công bố cho thấy thu nhập trung bình mỗi tháng của người lao động ở một thành phố của bạn là 5 triệu đồng mỗi tháng. Đây là cơ sở để họ xác định mức đánh thuế thu nhập cao. Bạn muốn kiểm tra xem báo cáo trên có chấp nhận được không. Làm thế nào để bạn có thể kiểm định được điều này?

Solution

- H_0 : *Thu nhập trung bình của người lao động trong thành phố của bạn là 5 triệu mỗi tháng.*
- H_1 : *Thu nhập trung bình của người lao động trong thành phố của bạn không phải là 5 triệu mỗi tháng.*

Logic của bài toán kiểm định

Chọn một mẫu ngẫu nhiên gồm những người lao động ở thành phố trên và tính lương trung bình của họ.

Logic của bài toán kiểm định

Chọn một mẫu ngẫu nhiên gồm những người lao động ở thành phố trên và tính lương trung bình của họ.
Giả sử giả thuyết H_0 là đúng, tức là thu nhập trung bình của người lao động ở thành phố của bạn là 5 triệu.

Logic của bài toán kiểm định

Chọn một mẫu ngẫu nhiên gồm những người lao động ở thành phố trên và tính lương trung bình của họ.
Giả sử giả thuyết H_0 là đúng, tức là thu nhập trung bình của người lao động ở thành phố của bạn là 5 triệu.

1. Thông tin trung bình mẫu liệu có chính xác là 5 triệu/tháng?

Logic của bài toán kiểm định

Chọn một mẫu ngẫu nhiên gồm những người lao động ở thành phố trên và tính lương trung bình của họ.
Giả sử giả thuyết H_0 là đúng, tức là thu nhập trung bình của người lao động ở thành phố của bạn là 5 triệu.

1. Thông tin trung bình mẫu liệu có chính xác là 5 triệu/tháng?
2. Trung bình mẫu tính ra là bao nhiêu thì bạn chấp nhận báo cáo trên có lí?

Logic của bài toán kiểm định

Chọn một mẫu ngẫu nhiên gồm những người lao động ở thành phố trên và tính lương trung bình của họ.
Giả sử giả thuyết H_0 là đúng, tức là thu nhập trung bình của người lao động ở thành phố của bạn là 5 triệu.

1. Thông tin trung bình mẫu liệu có chính xác là 5 triệu/tháng?
2. Trung bình mẫu tính ra là bao nhiêu thì bạn chấp nhận báo cáo trên có lí?
3. Khi nào bạn bác bỏ?

Logic của bài toán kiểm định

Chọn một mẫu ngẫu nhiên gồm những người lao động ở thành phố trên và tính lương trung bình của họ.
Giả sử giả thuyết H_0 là đúng, tức là thu nhập trung bình của người lao động ở thành phố của bạn là 5 triệu.

1. Thông tin trung bình mẫu liệu có chính xác là 5 triệu/tháng?
2. Trung bình mẫu tính ra là bao nhiêu thì bạn chấp nhận báo cáo trên có lí?
3. Khi nào bạn bác bỏ?

$\rightarrow$ cần một quy tắc quyết định có tính định lượng rõ ràng

1. Giả thuyết thống kê
2. Logic của bài toán kiểm định
3. Các loại sai lầm có thể mắc phải
4. Kiểm định giả thuyết so sánh trung bình một tổng thể với một số
   - Trường hợp tổng thể tuân theo phân phối chuẩn, biết phương sai
     - P-giá trị
   - Kiểm định trung bình tổng thể bất kì khi cỡ mẫu lớn
   - Kiểm định trung bình tổng thể khi tổng thể có phân phối chuẩn
   - Thực tế

Example

Giả sử tại một phiên tòa xét xử một bị cáo. Dù có nhiều bằng chứng gián tiếp bất lợi cho bị cáo nhưng tòa lại không có bằng chứng trực tiếp rõ ràng. Hãy thiết lập cặp giả thuyết và phân tích sai lầm có thể mắc phải khi tòa đưa ra kết luận.

Example

Giả sử tại một phiên tòa xét xử một bị cáo. Dù có nhiều bằng chứng gián tiếp bất lợi cho bị cáo nhưng tòa lại không có bằng chứng trực tiếp rõ ràng. Hãy thiết lập cặp giả thuyết và phân tích sai lầm có thể mắc phải khi tòa đưa ra kết luận.

Solution

H_0 : *Bị cáo vô tội.* H_1 : *Bị cáo có tội.*

Example

Giả sử tại một phiên tòa xét xử một bị cáo. Dù có nhiều bằng chứng gián tiếp bất lợi cho bị cáo nhưng tòa lại không có bằng chứng trực tiếp rõ ràng. Hãy thiết lập cặp giả thuyết và phân tích sai lầm có thể mắc phải khi tòa đưa ra kết luận.

Solution

H_0 : *Bị cáo vô tội.* H_1 : *Bị cáo có tội.*

- *Kết án bị cáo có tội nhưng thực ra anh ta vô tội.*

Example

Giả sử tại một phiên tòa xét xử một bị cáo. Dù có nhiều bằng chứng gián tiếp bất lợi cho bị cáo nhưng tòa lại không có bằng chứng trực tiếp rõ ràng. Hãy thiết lập cặp giả thuyết và phân tích sai lầm có thể mắc phải khi tòa đưa ra kết luận.

Solution

H_0 : *Bị cáo vô tội.* H_1 : *Bị cáo có tội.*

- *Kết án bị cáo có tội nhưng thực ra anh ta vô tội.Sai lầm kiểu này gọi là sai lầm loại I.*
- *Tha bổng cho bị cáo trong khi anh ta đã gây ra tội.*

Example

Giả sử tại một phiên tòa xét xử một bị cáo. Dù có nhiều bằng chứng gián tiếp bất lợi cho bị cáo nhưng tòa lại không có bằng chứng trực tiếp rõ ràng. Hãy thiết lập cặp giả thuyết và phân tích sai lầm có thể mắc phải khi tòa đưa ra kết luận.

Solution

H_0 : *Bị cáo vô tội.* H_1 : *Bị cáo có tội.*

- *Kết án bị cáo có tội nhưng thực ra anh ta vô tội.Sai lầm kiểu này gọi là sai lầm loại I.*
- *Tha bổng cho bị cáo trong khi anh ta đã gây ra tội. Sai lầm kiểu này gọi là sai lầm loại II.*

Definition

Sai lầm loại I: Xảy ra khi thực tế H_0 đúng nhưng ta lại bác bỏ nó. Xác suất để xảy ra sai lầm loại I kí hiệu là α và được gọi là mức ý nghĩa của phép kiểm định.

Definition

Sai lầm loại I: Xảy ra khi thực tế H_0 đúng nhưng ta lại bác bỏ nó. Xác suất để xảy ra sai lầm loại I kí hiệu là α và được gọi là mức ý nghĩa của phép kiểm định.

Sai lầm loại II: Xảy ra khi trong thực tế H_0 sai nhưng ta lại chấp nhận H_0. Xác suất để xảy ra sai lầm loại II kí hiệu là β. Khi đó $1-\beta$ là xác suất quyết định đúng, tức là ta bác bỏ H_0, giá trị này được gọi là năng lực của phép kiểm định.

Các tình huống có thể xảy ra

Quyết định	H_0 đúng	H_0 sai
Bác bỏ H_0	Sai lầm loại I xác suất $= \alpha$ α là mức ý nghĩa	Quyết định đúng xác suất $= 1 - \beta$ $1 - \beta$: năng lực của kiểm định
Không bác bỏ H_0	Quyết định đúng xác suất $= 1 - \alpha$	Sai lầm loại II xác suất $= \beta$

Các tình huống có thể xảy ra

Quyết định	H_0 đúng	H_0 sai
Bác bỏ H_0	Sai lầm loại I xác suất $= \alpha$ α là mức ý nghĩa	Quyết định đúng xác suất $= 1 - \beta$ $1 - \beta$: năng lực của kiểm định
Không bác bỏ H_0	Quyết định đúng xác suất $= 1 - \alpha$	Sai lầm loại II xác suất $= \beta$

$\rightarrow$ Cách duy nhất để cùng làm giảm xác suất của hai loại sai lầm α, β là tăng cỡ mẫu nghiên cứu.

Example

Trong ví dụ về phiên tòa xử tội bị cáo nói trên ta nói:

- Tại mức ý nghĩa 5% tòa bác bỏ H_0 có nghĩa là
  P(Bác bỏ $H_0|H_0$ đúng) $\leq 5\%$: xác suất bỏ tù oan là không quá 5%.
- Tại mức ý nghĩa 5% tòa chấp nhận H_0 có nghĩa là
  P(Bác bỏ $H_0|H_0$ đúng) $> 5\%$: Nếu bỏ tù bị cáo thì xác suất gây oan là quá 5%.

Example

Hãy xác định các sai lầm có thể mắc phải của tất cả các ví dụ đã xét. Trong mỗi ví dụ đó, "tại mức ý nghĩa 5% ta bác bỏ H_0" có nghĩa là gì? Còn câu: "Tại mức ý nghĩa α ta chấp nhận H_0" có nghĩa là gì?

Ôn tập

Câu hỏi

Chỉ số IQ của toàn xã hội có phân bố chuẩn với trung bình là 100, độ lệch chuẩn là 15. Một số người cho rằng những sinh viên học đại học thì IQ thường cao hơn mức chung của xã hội. Hãy đặt cặp giả thiết thống kê cho bài toán này. Các sai lầm có thể mắc phải là gì?

Solution

Gọi μ là chỉ số IQ trung bình của sinh viên đại học.
Ta đang muốn kiểm định: "những sinh viên học đại học thì IQ thường cao hơn mức chung của xã hội", tức là $\mu > 100$, đây là mệnh đề không có dấu "=" nên nó là H_1. Vậy

$$H_0 : \mu \leq 100 \qquad H_1 : \mu > 100$$

Câu hỏi

Một nhóm các phóng viên cho rằng các dụng cụ nhiệt kế được đưa đầu tư vào sử ở nhà hệ thống trường phổ thông là không chính xác. Họ lấy mẫu ngẫu nhiên những nhiệt kế một vài chục trường để đo nhiệt độ nước đang đun. Họ thấy rằng nhiều nhiệt kế chỉ số thấp hơn 100 khi nước đang sôi. Từ đó muốn chứng minh rằng trung bình chỉ số của các nhiệt kế thấp hơn 100 khi để trong nước đang sôi. Hãy thiết lập cặp giả thiết kiểm định của nhóm phóng viên.

Solution

Gọi μ là chỉ số trung bình của các nhiệt kế để trong môi trường nước sôi.

$$H_0 : \mu \geq 100 \qquad H_1 : \mu < 100$$

1. Giả thuyết thống kê
2. Logic của bài toán kiểm định
3. Các loại sai lầm có thể mắc phải
4. Kiểm định giả thuyết so sánh trung bình một tổng thể với một số
   - Trường hợp tổng thể tuân theo phân phối chuẩn, biết phương sai
     - P-giá trị
   - Kiểm định trung bình tổng thể bất kì khi cỡ mẫu lớn
   - Kiểm định trung bình tổng thể khi tổng thể có phân phối chuẩn
   - Thực tế

1. Giả thuyết thống kê
2. Logic của bài toán kiểm định
3. Các loại sai lầm có thể mắc phải
4. Kiểm định giả thuyết so sánh trung bình một tổng thể với một số
   - Trường hợp tổng thể tuân theo phân phối chuẩn, biết phương sai
     - P-giá trị
   - Kiểm định trung bình tổng thể bất kì khi cỡ mẫu lớn
   - Kiểm định trung bình tổng thể khi tổng thể có phân phối chuẩn
   - Thực tế

Nhắc lại

Theorem

Giả sử tổng thể có phân phối chuẩn trung bình μ (chưa biết), độ lệch chuẩn σ. Khi đó trung bình mẫu $\bar{X}$ của mẫu ngẫu nhiên cỡ n tuân theo $N(\mu, \frac{\sigma^2}{n})$

Như vậy

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Example

Một chủ hồ nuôi khi thương thảo bán lứa cá chim cho rằng lứa cá của họ có trọng lượng trung bình là 2.5kg. Công ty thu mua cá đã bắt ngẫu nhiên 20 con cá và cân nặng của chúng như sau (kg):

1.7 1.8 2.8 1.8 2.1 2.9 1.0 1.5 1.4 2.7 1.3 0.9 1.4 1.3 2.3 1.6 1.8 3.0 1.0 1.2

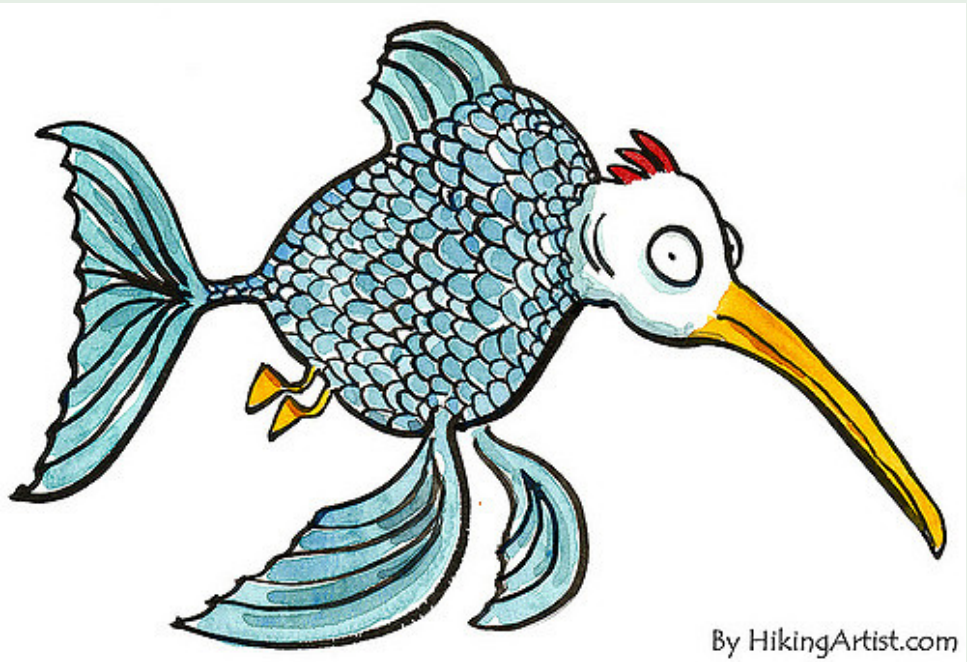
Giả sử rằng trọng lượng của tổng thể lứa cá có phân phối chuẩn với phương sai 0.36 (kg). Hỏi tại mức ý nghĩa 5% có thể chấp nhận khẳng định của chủ hồ nuôi không?

Example

Một chủ hồ nuôi khi thương thảo bán lứa cá chim cho rằng lứa cá của họ có trọng lượng trung bình là 2.5kg. Công ty thu mua cá đã bắt ngẫu nhiên 20 con cá và cân nặng của chúng như sau (kg):

1.7 1.8 2.8 1.8 2.1 2.9 1.0 1.5 1.4 2.7 1.3 0.9 1.4 1.3 2.3 1.6 1.8 3.0 1.0 1.2

Giả sử rằng trọng lượng của tổng thể lứa cá có phân phối chuẩn với phương sai 0.36 (kg). Hỏi tại mức ý nghĩa 5% có thể chấp nhận khẳng định của chủ hồ nuôi không?

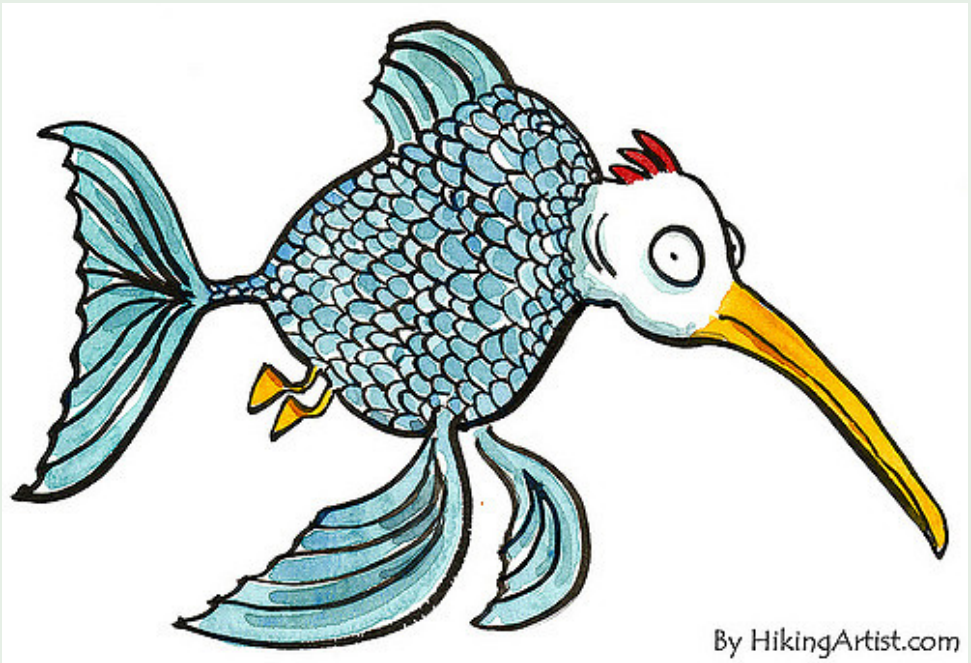


Example

Một chủ hồ nuôi khi thương thảo bán lứa cá chim cho rằng lứa cá của họ có trọng lượng trung bình là 2.5kg. Công ty thu mua cá đã bắt ngẫu nhiên 20 con cá và cân nặng của chúng như sau (kg):

1.7 1.8 2.8 1.8 2.1 2.9 1.0 1.5 1.4 2.7 1.3 0.9 1.4 1.3 2.3 1.6 1.8 3.0 1.0 1.2

Giả sử rằng trọng lượng của tổng thể lứa cá có phân phối chuẩn với phương sai 0.36 (kg). Hỏi tại mức ý nghĩa 5% có thể chấp nhận khẳng định của chủ hồ nuôi không?



Solution

- *Gọi μ là trọng lượng trung bình của lứa cá nói trên (kg). Ta có*

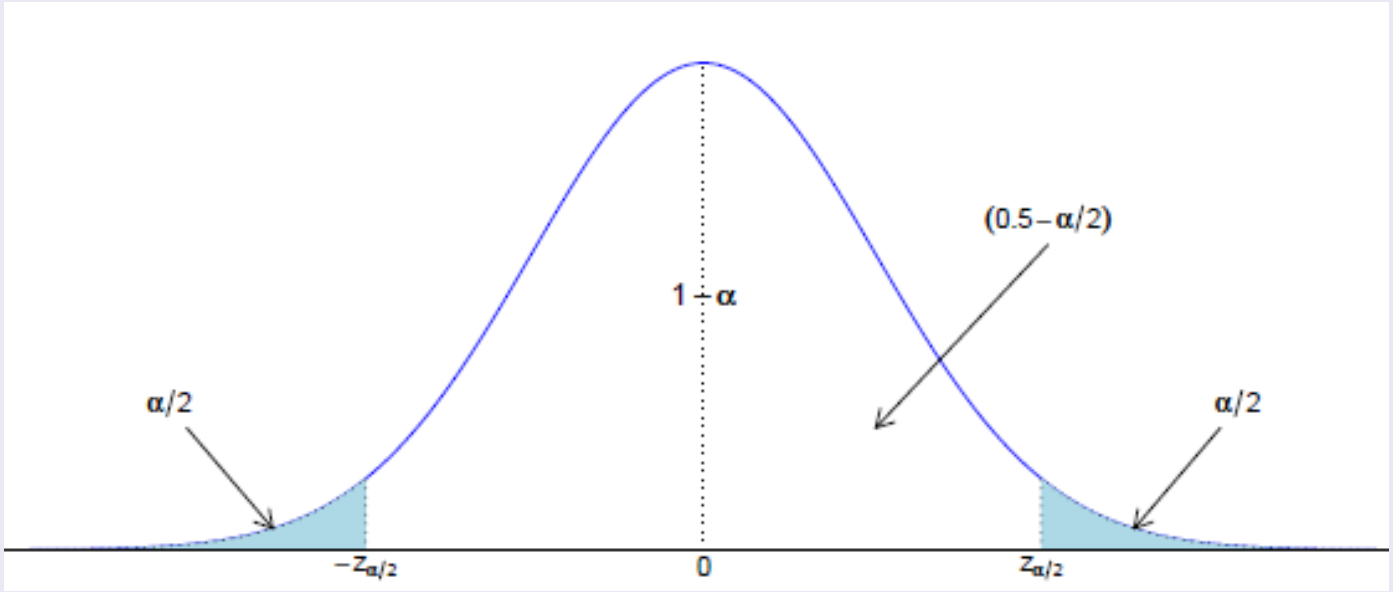
$$H_0 : \mu = 2.5 \qquad H_1 : \mu \neq 2.5$$

- *Gọi $\bar{X}$ là trọng lượng trung bình của một mẫu 20 con cá lấy ngẫu nhiên.*
  Nếu H_0 là đúng, thì $\mu = 2.5$ và theo như định lí trên, đại lượng

$$Z = \frac{\bar{X} - 2.5}{\sigma/\sqrt{n}} \sim N(0,1)$$

Theo đó, như đã biết ở phần ước lượng, xác suất để Z nói trên nằm trong khoảng $-z_{0.025}$ đến $z_{0.025}$ là 95%.

Solution



- *Bởi vậy, nếu Z tính ra từ công thức trên mà nằm ngoài khoảng $[-z_{0.025}; z_{0.025}]$ thì khi ta bác bỏ H_0, khả năng sai của ta là không quá 5%.*
  Ta có $\bar{x} = 1.775, n = 20, \sigma = 0.6$ do đó: $z = -5.4038$.
- *Ta đã biết,$[-z_{0.025}; z_{0.025}] = [-1.96; 1.96]$. Do vậy, với xác suất sai không quá 5%, ta bác bỏ H_0.*
- *Vậy, tại mức ý nghĩa 5%, khẳng định của chủ hồ nuôi là không đúng.*

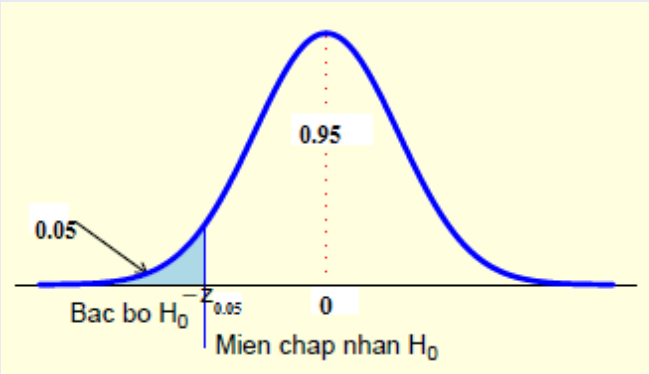
Example

Sau khi đã bác bỏ khẳng định của chủ hồ nuôi, khách hàng còn cho rằng trọng lượng trung bình của lứa cá này là dưới 1.8 (kg) và nếu điều này đúng, thì hợp đồng thu mua đã kí trước đó sẽ bị hủy. Tại mức ý nghĩa 5%, khẳng định của đại diện công ty có đúng không?

Solution

- $H_0 : \mu \geq 1.8 \qquad H_1 : \mu < 1.8$ *(bài toán bên trái)*
- *Ta có* $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

  Nên nếu H_0 *đúng, thì* $\mu \geq 1.8$ *nên* $z = \dfrac{\bar{X} - 1.8}{\sigma/\sqrt{n}}$ *có xu hướng* $\geq Z$*, vì thế, khả năng nó rơi vào miền* $[-z_{0.05}; +\infty]$ *chiếm tới 95%:*



  Khi giá trị z *tính ra từ mẫu ngẫu nhiên mà nhỏ hơn* $-z_{0.05} = -1.64$ *thì việc ta bác bỏ* H_0 *sẽ mắc phải xác suất sai lầm không quá* 5%*.*
- *Ta tính được* $z = -0.18634 > -1.64$*, do vậy, không thể bác bỏ* H_0 *nếu ta chỉ chấp nhận sai lầm* $\leq 5\%$*.*

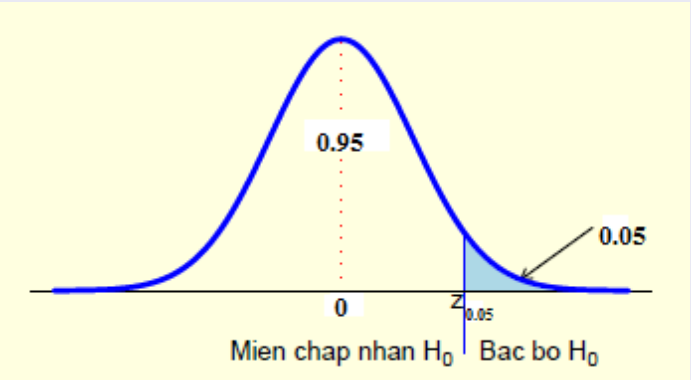
Example

Như vậy, mặc dù mẫu tính ra trung bình là $1.775 < 1.8$. Nhưng điều này không đảm bảo cho $\mu < 1.8$. Một câu hỏi tự nhiên được đặt ra, nếu ta muốn kiểm định xem, tại mức ý nghĩa 5%, liệu μ có lớn hơn 1.7 không thì ta sẽ làm thế nào?

Solution

- $H_0 : \mu \leq 1.7 \qquad H_1 : \mu > 1.7$ *(bài toán bên phải)*
- *Ta có* $Z = \dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

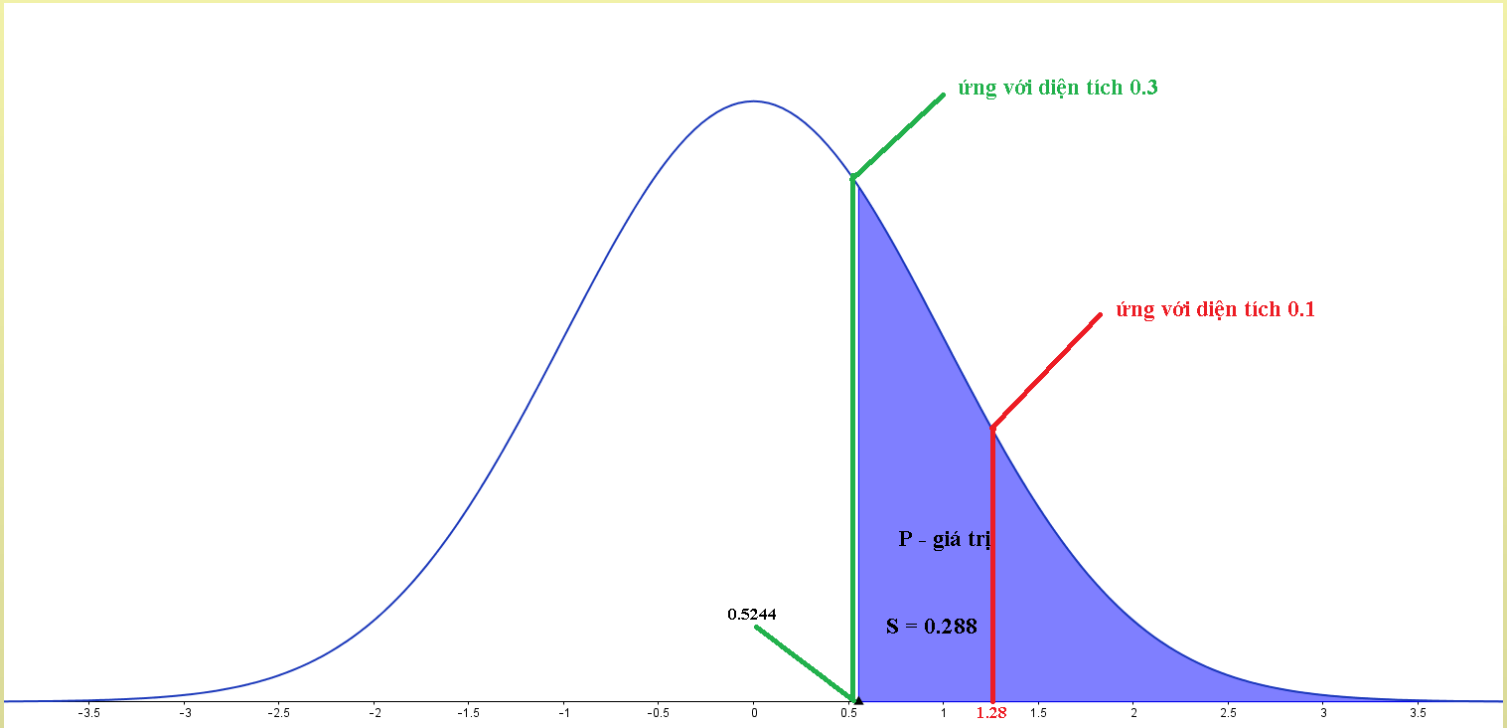
  Nên nếu H_0 *đúng, thì* $z = \dfrac{\bar{X} - 1.7}{\sigma/\sqrt{n}}$ *tính ra có xu hướng* $\leq Z$ *vì thế, khả năng nó rơi vào miền* $[-\infty; z_{0.05}]$ *chiếm tới 95%:*



  Bởi thế nếu giá trị z tính ra từ mẫu mà lớn hơn $z_{0.05} = 1.64$ *thì việc ta bác bỏ* H_0 *sẽ mắc phải xác suất sai lầm không quá 5%.*
- *Ta tính được* $z = 0.55902 < 1.64$*, do vậy, không thể bác bỏ* H_0 *tại mức ý nghĩa 5%.*

P - giá trị

- Trong bài toán cuối về trọng lượng đàn cá nói trên nếu mức ý nghĩa α mà ta chọn là 0.1 thì ta vẫn chấp nhận H_0, vì rằng khi đó vùng chấp nhận là từ $(-\infty; 1.28]$, miền này chứa $z = 0.55432$. Nhưng nếu chọn $\alpha = 0.3$ thì ta có miền chấp nhận là $(-\infty; 0.5244]$ và do đó ta bác bỏ H_0.



P - giá trị

- Như vậy, α càng lớn, khả năng chấp nhận H_0 càng giảm đi. Mốc của α mà chuyển từ quyết định chấp nhận sang bác bỏ H_0 được gọi là **P - giá trị**.

Định nghĩa

Giá trị α lớn nhất mà khi đó ta còn chấp nhận H_0 được gọi là **P - giá trị**. *Như vậy, khi α lớn hơn P - giá trị thì ta bác bỏ H_0.*

- Đối với bài toán bên phải, P - giá trị $= P(Z > z)$. Ta bác bỏ H_0 khi P - giá trị $< \alpha$.

Dùng P - giá trị để quyết định

Đối với bài toán ta đang xét ở đây, quy luật bác bỏ hay chấp nhận H_0 dựa trên thống kê z, đầu tiên ta tính giá trị kiểm định $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$, sau đó cách tính P - giá trị như sau:

- Nếu $H_1 : \mu > \mu_0$ thì p-giá trị $= P(Z > z_0)$;
- Nếu $H_1 : \mu < \mu_0$ thì p-giá trị $= P(Z < z_0)$;
- Nếu $H_1 : \mu \neq \mu_0$ thì p-giá trị $= 2P(Z > |z_0|)$.

Sau đó, bác bỏ H_0 khi và chỉ khi P - giá trị $< \alpha$.

Tổng kết quy luật bác bỏ và chấp nhận H_0

H_0	H_1	Giá trị thống kê z	Qui luật bác bỏ H_0	p-giá trị
$\mu \leq \mu_0$	$\mu > \mu_0$	$z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	$z > z_\alpha$	$P(Z > z)$
$\mu \geq \mu_0$	$\mu < \mu_0$	$z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	$z < -z_\alpha$	$P(Z < z)$
$\mu = \mu_0$	$\mu \neq \mu_0$	$z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	$\lvert z \rvert > z_{\alpha/2}$	$2P(Z > z)$

Các bước thực hiện kiểm định cụ thể như sau:

- Bước 1: Thiết lập cặp giả thuyết
- Bước 2: Tính giá trị thống kê
  - Tính giá trị thống kê: $z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$.
  - Tính giá trị tới hạn: z_α hoặc $z_{\alpha/2}$.
  - Hoặc, tính P - giá trị
- Bước 3: Kết luận tại mức ý nghĩa α theo quy tắc nói ở trên.
- Bước 4: Đưa ra những quyết định phù hợp trong kinh tế - xã hội.

Trên R

Trên R đã cài đặt toàn bộ các công thức trên thông qua hàm z.test có trong gói BSDA. Việc cài BSDA chỉ cần gõ câu lệnh:

$$install.packages("BSDA")$$

Sau đó chọn một mirror hiện ra trong danh sách (nếu có).
Sau khi cài xong, mỗi lần sử dụng đều phải khởi động BSDA qua lệnh:

$$library(BSDA)$$

$$\text{z.test(x=Véc tơ mẫu,sigma.x=}\sigma\text{, mu=}\mu_0\text{,alternative=...)}$$

Trong đó ... là "less" nếu là bài toán bên trái, "greater" nếu là bài toán bên phải, "two.side" nếu là bài toán hai bên.
Trong trường hợp chúng ta không có dữ liệu là một mẫu chi tiết mà chỉ biết $\bar{x}, n$ thì ta có thể dùng hàm sau

$$\text{zsum.test(mean.x = }\bar{x}\text{,n.x=}n\text{,sigma.x=}\sigma\text{, mu=}\mu_0\text{,alternative=...)}$$

Example

Với bài toán đầu tiên: $Ho : \mu = 2.5$ $H_1 : \mu \neq 2.5$

```
> TrongLuong=scan()
1: 1.7 1.8 2.8 1.8 2.1 2.9 1.0 1.5 1.4 2.7 1.3 0.9 1.4 1.3 2.
3 1.6 1.8 3.0 1.0 1.2
21:
Read 20 items
> z.test(TrongLuong,mu=2.5,sigma.x = 0.6,alternative = "t")

        One-sample z-Test

data:  TrongLuong
z = -5.4038, p-value = 6.523e-08
alternative hypothesis: true mean is not equal to 2.5
95 percent confidence interval:
 1.512043 2.037957
sample estimates:
mean of x
    1.775
```

Example (tiếp)

Từ kết quả đầu ra của câu lệnh, ta có:

- Nếu dùng P - giá trị, ta có ngay P - giá trị $= 6.523 \times 10^{-8} < \alpha = 0.05$ nên ta bác bỏ H_0.
- Nếu dùng giá trị tới hạn, hai bên, nên ta tính $z_{0.025} = qnorm(0.025, lower.tail = FALSE) = 1.959964$. Ta lưu ý giá trị thống kê $z = -5.4038 \notin [-1.959964; 1.959964]$ nên bác bỏ H_0.

Tóm lại, với việc chấp nhận sai lầm ở mức 5% ta có thể cho rằng lời khẳng định của chủ hồ nuôi là không đúng.

Thực hành

Câu hỏi

Trở lại với câu hỏi 1. Một nhóm sinh viên muốn kiểm định xem ý kiến nói trên có đúng với sinh viên TLU hay không. Họ tiến hành lấy mẫu hệ thống 20 sinh viên và test IQ của họ. Kết quả cho thấy trung bình chỉ số IQ là 110. Giả sử tổng thể chỉ số IQ của sinh viên TLU có phân bố chuẩn, độ lệch chuẩn là 15 giống toàn xã hội. Hãy kiểm định khẳng định nói trên tại mức ý nghĩa 5%, cho $z_{0.05} = 1.64, z_{0.025} = 1.96$.

Solution

Gọi m là chỉ số IQ trung bình của sinh viên TLU (Việc đặt cặp giả thuyết xem lại câu hỏi 1).

$$H_0 : \mu \leq 100 \qquad H_1 : \mu > 110$$

Do tổng thể số liệu được cho là có phân bố chuẩn, có giả thiết phương sai là 15^2 *nên ta dùng kiểm định z.*

Solution

```
> zsum.test(mean.x = 110,mu=100,sigma.x=15,n.x=20,alt = "g")

        One-sample z-Test

data:  Summarized x
z = 2.9814, p-value = 0.001435
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 104.483      NA
sample estimates:
mean of x
      110
```

Từ kết quả có giá trị kiểm định $z = 2.9814 > z_{0.05} = 1.96$ *ta suy ra bác bỏ* H_0.
Hoặc dùng P - giá trị $= 0.001435 < 0.05$ *nên bác bỏ* H_0.
Vậy, tại mức ý nghĩa 5%, ta có thể khẳng định chỉ số IQ trung bình của sinh viên TLU là cao hơn so với mặt bằng chung của xã hội.

1. Giả thuyết thống kê
2. Logic của bài toán kiểm định
3. Các loại sai lầm có thể mắc phải
4. Kiểm định giả thuyết so sánh trung bình một tổng thể với một số
   - Trường hợp tổng thể tuân theo phân phối chuẩn, biết phương sai
     - P-giá trị
   - Kiểm định trung bình tổng thể bất kì khi cỡ mẫu lớn
   - Kiểm định trung bình tổng thể khi tổng thể có phân phối chuẩn
   - Thực tế

Khi cỡ mẫu lớn

Theo định lí giới hạn trung tâm, khi cỡ mẫu $n \geq 30$, phương sai tổng thể chưa biết được thay bằng phương sai của mẫu, đại lượng $z = \dfrac{\overline{X} - \mu}{S_X/\sqrt{n}}$ xấp xỉ phân phối chuẩn hóa. Khi đó các qui trình kiểm định được thực hiện giống như trường hợp phương sai đã biết ở trên.

H_0	H_1	Giá trị thống kê z	Qui luật bác bỏ H_0	p-giá trị
$\mu \leq \mu_0$	$\mu > \mu_0$	$z = \dfrac{\overline{x} - \mu_0}{s_x/\sqrt{n}}$	$z > z_\alpha$	$P(Z > z)$
$\mu \geq \mu_0$	$\mu < \mu_0$		$z < -z_\alpha$	$P(Z < z)$
$\mu = \mu_0$	$\mu \neq \mu_0$		$\lvert z\rvert > z_{\alpha/2}$	$2P(Z > \lvert z\rvert)$

Example

Giả sử một điều tra với mẫu 100 người lao động của một thành phố cho thấy trung bình thu nhập của 100 người này là 4.8 triệu với độ lệch chuẩn mẫu là 1 triệu. Hỏi có thể chấp nhận báo cáo cho rằng lương trung bình của toàn bộ người lao động ở thành phố trên là ít nhất 5 triệu hay không? Chọn mức ý nghĩa 5%.

Solution

Gọi μ là trung bình lương của người lao động ở thành phố trên.

1. *Cặp giả thuyết:*
$$H_0 : \mu \geq 5 \qquad H_1 : \mu < 5$$
2. *Tính giá trị kiểm định* $z = \dfrac{4.8 - 5}{1/\sqrt{100}} = -2.$
3. *Giá trị tới hạn:* $-z_{0.05} = -1.64$. *Ta thấy* $z < -z_{0.05}$ *nên bác bỏ* H_0. *(Hoặc, nếu tính P - giá trị* $= P(Z < z) = P(Z < -2) = 0.02275 <$ *0.05 nên bác bỏ* H_0.*)*
4. *Vậy, tại mức ý nghĩa 5%, thu nhập trung bình của người lao động ở thành phố trên là thấp hơn 5 triệu*

Example

Vậy, trong ví dụ trên, liệu có thể coi thu nhập trung bình của người lao động ở thành phố trên là cao hơn 4.5 triệu không? Chọn mức ý nghĩa 5%.

Solution

1. *Cặp giả thuyết:*
$$H_0 : \mu \leq 4.5 \qquad H_1 : \mu > 4.5$$
2. *Tính giá trị kiểm định* $z = \dfrac{4.8 - 4.5}{1/\sqrt{100}} = 3.$
3. *Đây là bài toán bên phải nên giá trị tới hạn là:* $z_{0.05} = 1.64$. *Ta thấy* $z > z_{0.05}$ *nên bác bỏ* H_0.
(Hoặc, nếu tính P - giá trị =
$P(Z > z) = P(Z > 3) = pnorm(3, 0, 1, F) = 0.001349898 < 0.05$
nên bác bỏ H_0.*)*
4. *Vậy, tại mức ý nghĩa 5%, thu nhập trung bình của người lao động ở thành phố trên là hơn 4.5 triệu*

Trên R

Ta thấy rằng, để giải quyết bài toán trong trường hợp $n \geq 30$ này, thực chất ta chỉ thay đổi σ trong trường hợp đầu tiên bởi s_x. Do đó hàm kiểm định trong R có thể dùng z.test như phần trên, với sigma.x $= s_x$:

z.test(x=Véc tơ mẫu,sigma.x=s_x, mu=μ_0,alternative=. . .)

Câu hỏi

Dữ liệu ChiTieu2010.csv là mẫu điều tra ngẫu nhiên vài chục nghìn hộ gia đình ở nước ta. Từ đó hãy:

1. *Người ta cho rằng chi tiêu giáo dục trung bình một năm của các hộ gia đình nước ta vào thời điểm 2010 là ít nhất 250 (nghìn). Hãy kiểm định điều này với mức ý nghĩa 1%.*
2. *Kiểm định tại mức ý nghĩa 5% ý kiến cho rằng chi tiêu ăn uống trung bình một tháng của tổng thể các hộ gia đình sống ở khu vực thành thị (khu vực = 1) ở nước ta vào thời điểm đó là không vượt quá 1000 (nghìn).*

Solution

Gọi μ là trung bình chi tiêu cho giáo dục của các hộ ở nước ta trong năm 2010.

$$H_0 : \mu \geq 250 \qquad H_1 : \mu < 250$$

```
> sd(ChiTieuGiaoDucTrongNam)
[1] 720.7803
> z.test(ChiTieuGiaoDucTrongNam,sigma.x=720.7803,mu=250,alt="less")

        One-sample z-Test

data:  ChiTieuGiaoDucTrongNam
z = -1.3258, p-value = 0.09246
alternative hypothesis: true mean is less than 250
95 percent confidence interval:
       NA 252.3724
sample estimates:
mean of x
 240.1428
```

Với P - giá trị $= 0.09246 > 0.01$ nên ta chấp nhận H_0. Vậy, nếu muốn sai lầm không quá 1% thì ta không bác bỏ được khẳng định trên.

1. Giả thuyết thống kê
2. Logic của bài toán kiểm định
3. Các loại sai lầm có thể mắc phải
4. Kiểm định giả thuyết so sánh trung bình một tổng thể với một số
   - Trường hợp tổng thể tuân theo phân phối chuẩn, biết phương sai
     - P-giá trị
   - Kiểm định trung bình tổng thể bất kì khi cỡ mẫu lớn
   - Kiểm định trung bình tổng thể khi tổng thể có phân phối chuẩn
   - Thực tế

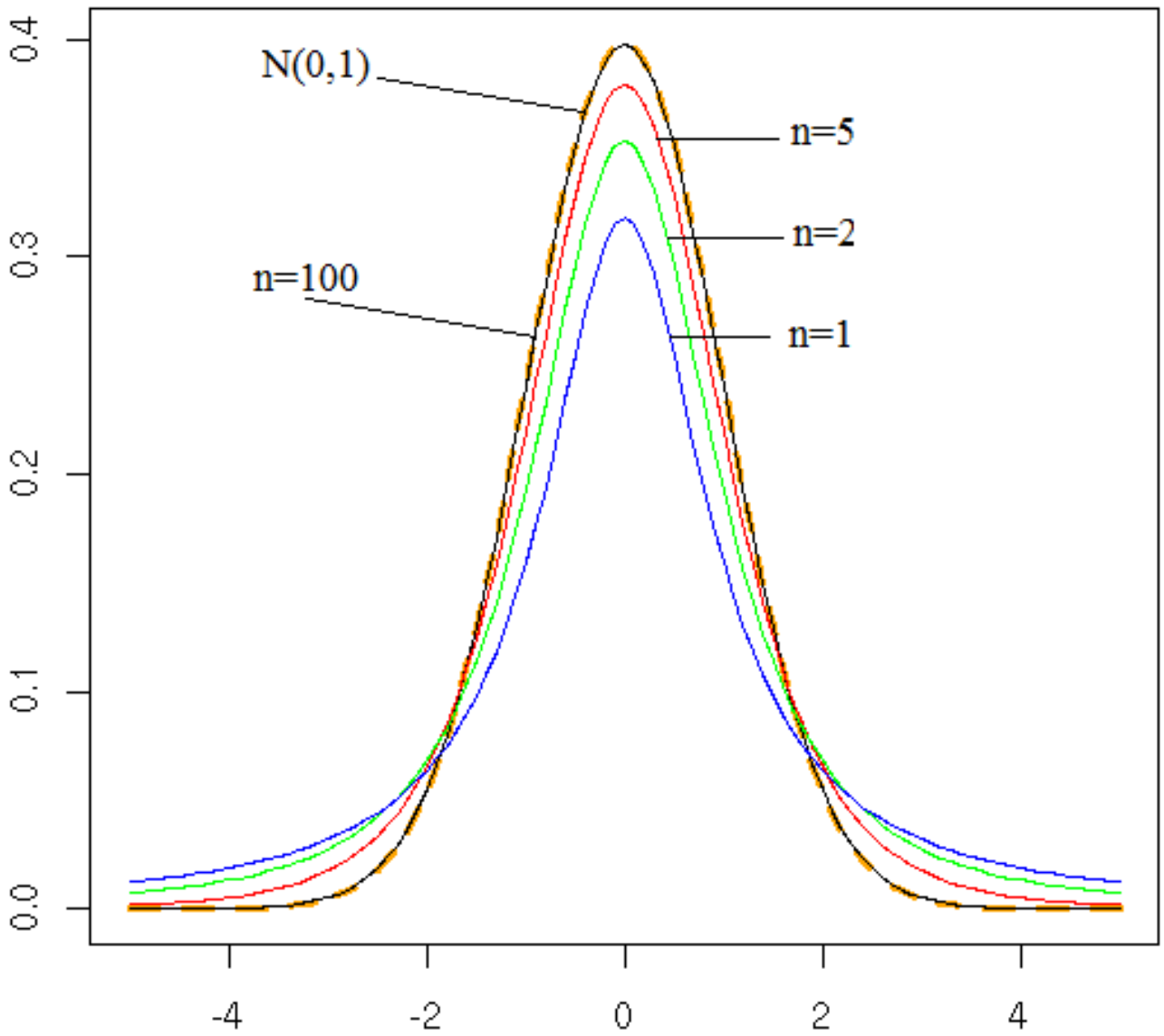
Khi tổng thể có phân phối chuẩn

Ta đã biết rằng, nếu tổng thể tuân theo phân phối chuẩn, khi chọn mẫu ngẫu nhiên và tính đại lượng

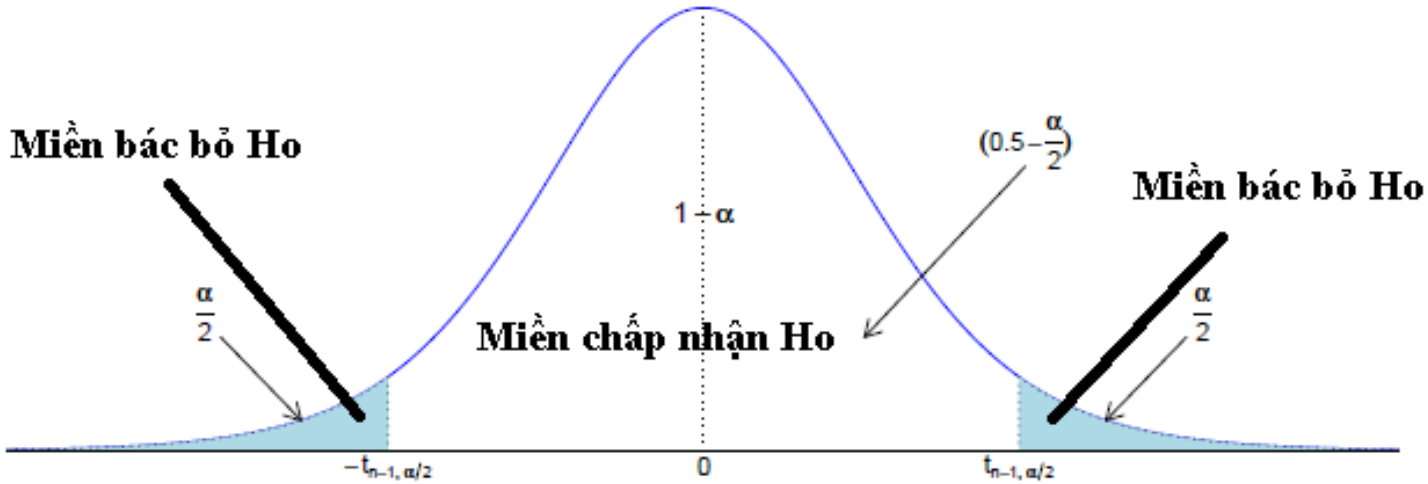
$$t = \frac{\overline{x} - \mu_0}{s_x/\sqrt{n}}$$

thì đại lượng này tuân theo phối Student .
Do đó để kiểm định cặp giả thiết về trung bình tổng thể khi tổng thể có phân phối chuẩn ta dùng phân phối Student, người ta gọi là kiểm định t.

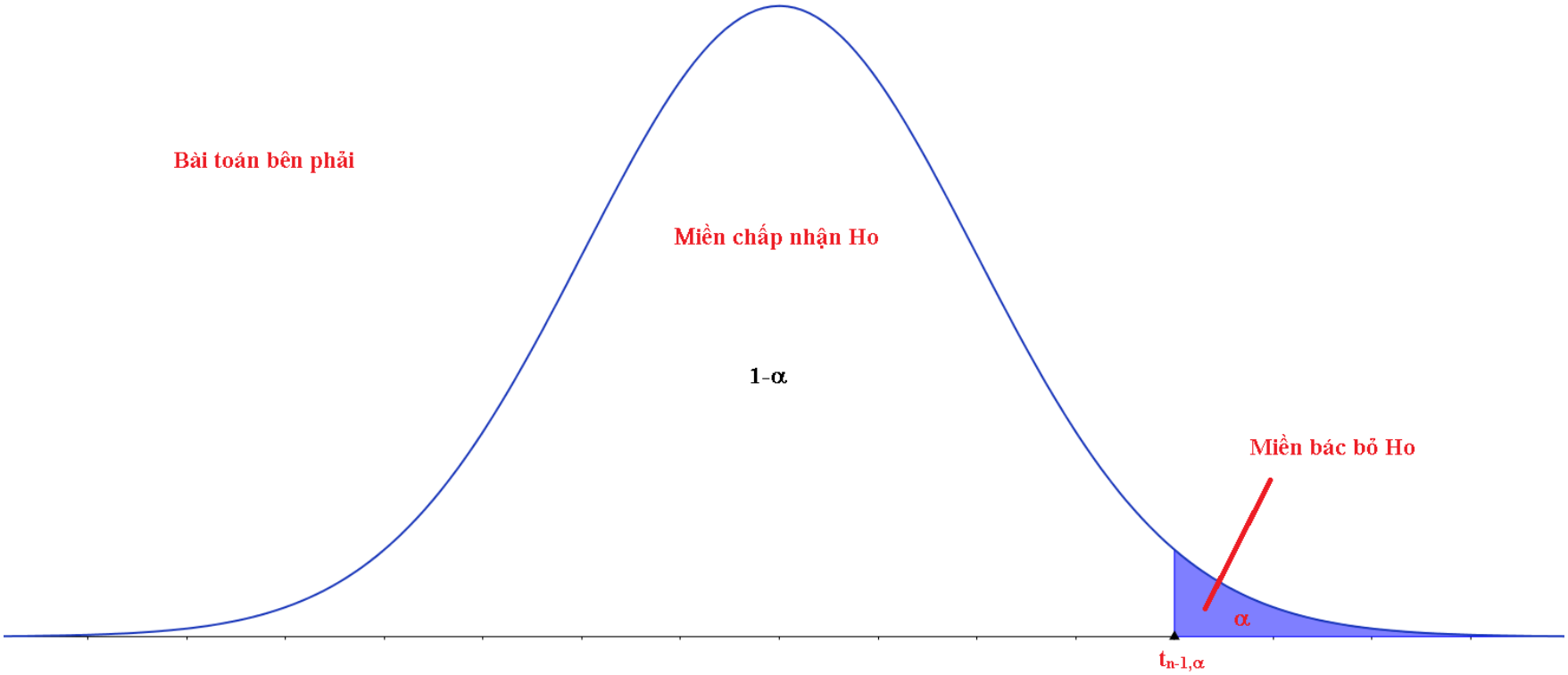
Phân phối t



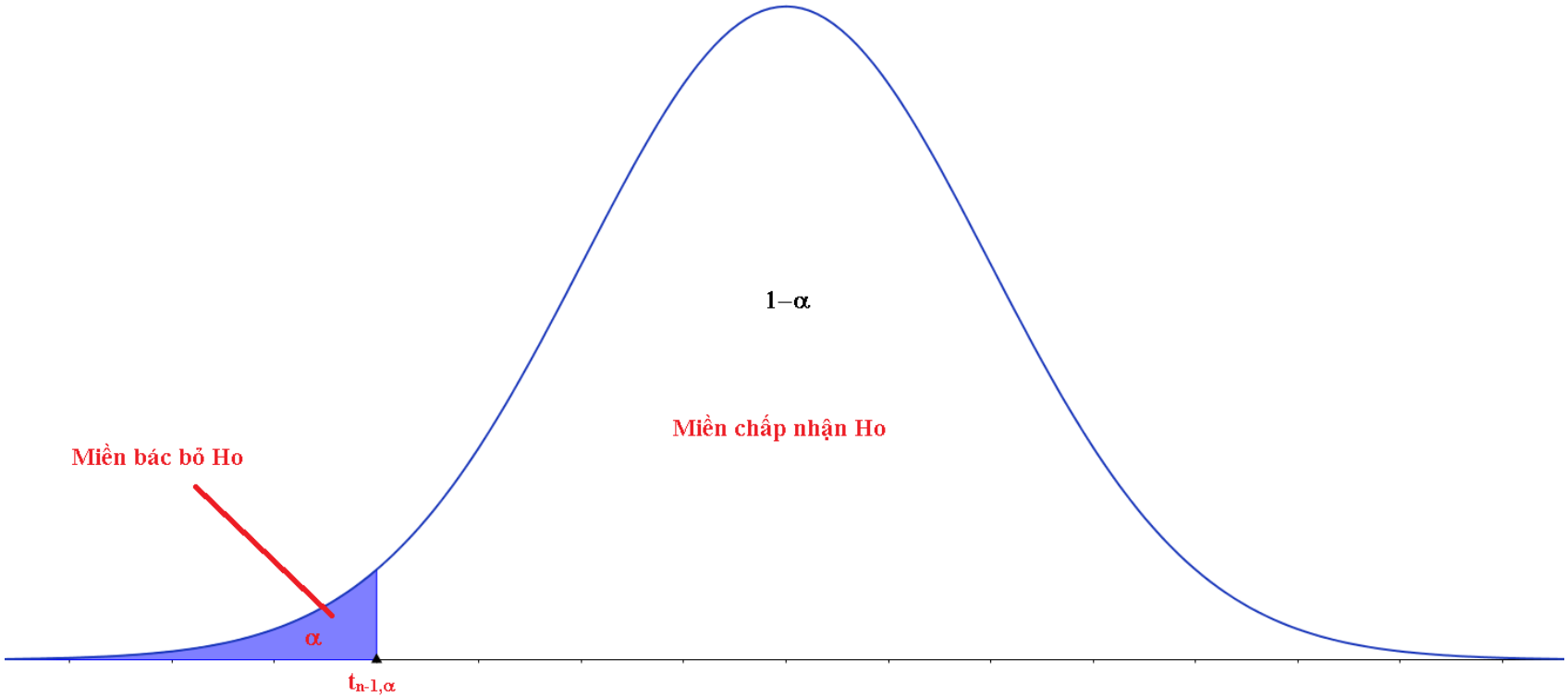
Bài toán hai bên



Bài toán bên phải



Bài toán bên trái



Khi tổng thể có phân phối chuẩn

H_0	H_1	Giá trị thống kê z	Qui luật bác bỏ H_0	p-giá trị
$\mu \leq \mu_0$	$\mu > \mu_0$	$t = \dfrac{\overline{x} - \mu_0}{s_x/\sqrt{n}}$	$t > t_{n-1,\alpha}$	$P(t_{n-1} > t)$
$\mu \geq \mu_0$	$\mu < \mu_0$		$t < -t_{n-1,\alpha}$	$P(t_{n-1} < t)$
$\mu = \mu_0$	$\mu \neq \mu_0$		$\lvert t\rvert > t_{n-1,\alpha/2}$	$2P(t_{n-1} > \lvert t\rvert)$

Trong phần mềm R ta có thể tính

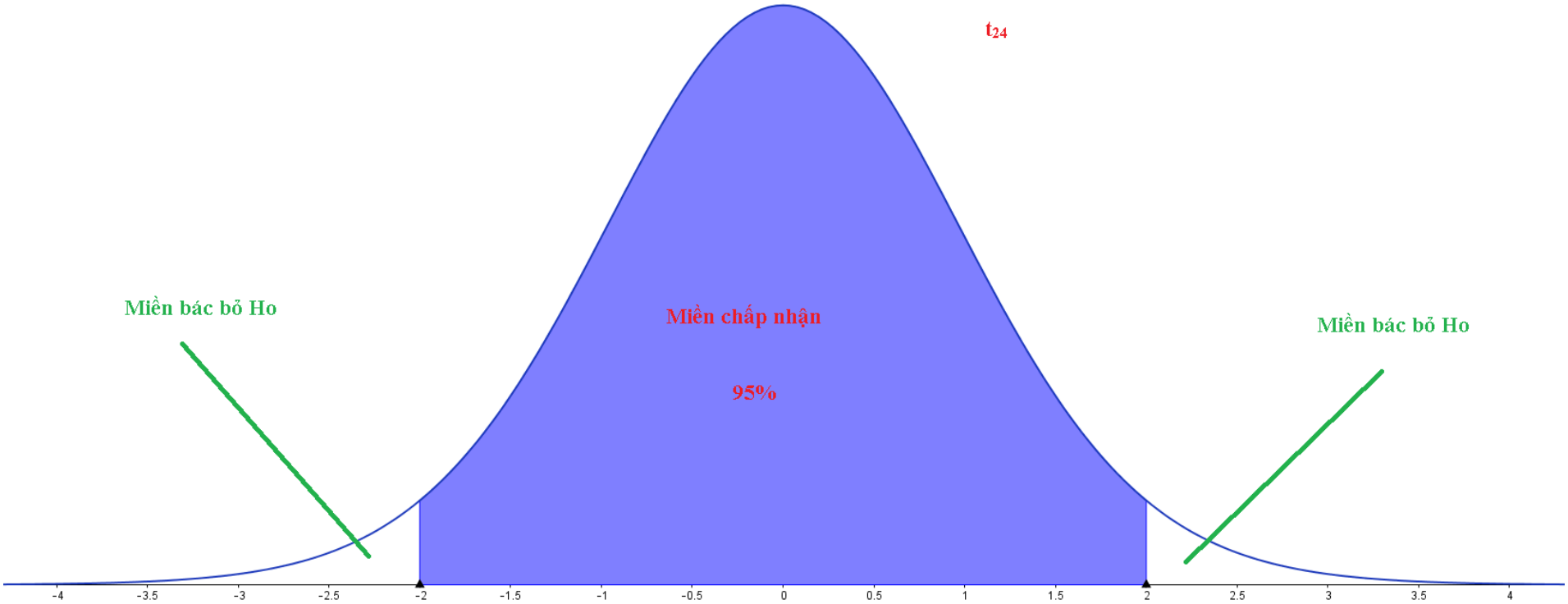
$$P(t_k < a) = pt(a, k); P(t_k > a) = pt(a, k, FALSE)$$

và

$$t_{k,\alpha} = qt(1 - \alpha, k)$$

Example

Trong những năm trước đây, giá cho thuê trung bình của cửa hàng ở một thành phố lớn vào khoảng 2 triệu/m^2. Một nhà đầu tư bất động sản muốn xác định xem con số này bây giờ có thay đổi không nên đã tiến hành thu thập một mẫu gồm 25 cửa hàng trong thành phố và thu được giá cho thuê trung bình là 2.2 triệu/m^2 với độ lệch chuẩn là 0.8 triệu. Tại mức ý nghĩa $\alpha = 5\%$ nhà đầu tư kết luận được điều gì nếu biết giá thuê một mét vuông cửa hàng ở thành phố này tuân theo phân phối chuẩn?


t_{24}
Miền bác bỏ Ho
Miền chấp nhận
95%
Miền bác bỏ Ho
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
3
3.5
4

Solution

Gọi μ là giá cho thuê trung bình tính trên mỗi m^2 của các cửa hàng thời điểm hiện tại.

1. *Cặp giả thuyết: $H_0 : \mu = 2 \qquad H_1 : \mu \neq 2.$*
2. *Do tổng thể phân phối chuẩn, chưa biết phương sai nên chọn thống kê là $t = \dfrac{\overline{x} - \mu_0}{s_x/\sqrt{n}}$ tuân theo phân phối Student với bậc tự do là 24.*
   Với $\overline{x} = 2.2, s_x = 0.8, n = 25, \mu = 2$ thay vào thống kê trên ta tính được $t = 1.25$.
   P - giá trị $= 2P(t_{24} > 1.25)$
   $= 2 * qt(1.25, df = 24, lower.tail = FALSE) = 0.2234.$
3. *Ta có P - giá trị $> \alpha = 5\%$ nên chấp nhận H_0.*
4. *Vậy ở mức ý nghĩa 5% ta chấp nhận giá thuê trung bình của các cửa hàng vẫn là 2 triệu/m^2.*

Trên R

Đối với kiểm định t, ta dùng hàm:

$$\text{t.test(x = Véc tơ mẫu, mu = } \mu_0\text{, alt = } \ldots\text{)}$$

Cách sử dụng tham số giả thuyết đối alt như đối với z.test.
Trong trường hợp không có dữ liệu gốc, ta cần có trung bình mẫu: $\bar{x}$, cỡ mẫu: n và thay vì dùng t.test ta dùng tsum.test

$$\text{tsum.test(mean.x=}\bar{x}\text{, n.x = } n\text{,mu=}\mu\text{,alt = } \ldots\text{)}$$

Example

Có ý kiến cho rằng các sinh viên đại học thì có chỉ số IQ cao hơn mức bình thường 100 của toàn xã hội. Một nhóm sinh viên muốn kiểm định xem ý kiến nói trên có đúng với sinh viên TLU hay không. Họ tiến hành lấy mẫu hệ thống 25 sinh viên và test IQ của họ. Kết quả như sau:

126	105	120	104	120	99	91	87	
116	119	103	117	115	115	143	138	109
108	126	127	126	107	70	105	121	

Giả sử tổng thể chỉ số IQ của sinh viên TLU có phân bố chuẩn. Hãy kiểm định khẳng định nói trên tại mức ý nghĩa 5%

Solution

Gọi μ là chỉ số IQ trung bình của sinh viên TLU.

$$H_0 : \mu \leq 100 \qquad \mu > 100$$

```
> IQ=scan()
1: 126 105 120 104 120  99  91  87 116 119 103 117 115 115 143
16: 138 109 108 126 127 126 107  70 105 121
26:
Read 25 items
> t.test(IQ,mu=100,alternative = "g")

        One Sample t-test

data:  IQ
t = 3.994, df = 24, p-value = 0.0002675
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 107.2484      Inf
sample estimates:
mean of x
   112.68
```

Solution (tiếp)

Với P - giá trị $= 0.0002675 < 0.05$ *nên ta bác bỏ* H_0.
Hoặc nếu dùng giá trị tới hạn, ta tính
$t_{24,0.05} = qt(0.05, df = 24, lower.tail = FALSE) = 1.710882$. *So sánh giá trị kiểm định* $t = 3.994 > t_{24,0.05}$ *nên bác bỏ* H_0.
Vậy, với sai lầm không quá 5% ta có thể cho rằng chỉ số IQ trung bình của sinh viên TLU là cao hơn mặt bằng chung.

Câu hỏi

Trở lại với câu hỏi 1. Một nhóm sinh viên muốn kiểm định xem ý kiến nói trên có đúng với sinh viên TLU hay không. Họ tiến hành lấy mẫu hệ thống 25 sinh viên và test IQ của họ. Kết quả cho thấy trung bình chỉ số IQ là 110, độ lệch chuẩn là 14. Giả sử tổng thể chỉ số IQ của sinh viên TLU có phân bố chuẩn. Hãy kiểm định khẳng định nói trên tại mức ý nghĩa 5%, cho $t_{24,0.05} = 1.71, t_{24,0.025} = 2.06$.

Solution

Gọi μ *là chỉ số IQ trung bình của sinh viên TLU.*

$$H_0 : \mu \leq 100 \qquad \mu > 100$$

```
> tsum.test(mean.x = 110,s.x=14,mu=100,n.x=25,alternative = "g")

        One-sample t-Test

data:  Summarized x
t = 3.5714, df = 24, p-value = 0.0007717
alternative hypothesis: true mean is greater than 100
95 percent confidence interval:
 105.2095       NA
sample estimates:
mean of x
      110
```

Ta có P - giá trị $= 0.0007717 < 0.05$ nên bác bỏ H_0.
Hoặc ta cũng có thể thấy do $t = 3.5714 > t_{24,0.05}$ nên bác bỏ H_0.
Vậy, với sai lầm không quá 5% ta có thể cho rằng chỉ số IQ trung bình của sinh viên TLU là cao hơn mặt bằng chung.

1. Giả thuyết thống kê
2. Logic của bài toán kiểm định
3. Các loại sai lầm có thể mắc phải
4. Kiểm định giả thuyết so sánh trung bình một tổng thể với một số
   - Trường hợp tổng thể tuân theo phân phối chuẩn, biết phương sai
     - P-giá trị
   - Kiểm định trung bình tổng thể bất kì khi cỡ mẫu lớn
   - Kiểm định trung bình tổng thể khi tổng thể có phân phối chuẩn
   - Thực tế

Trong thực tế chúng có thể ta dùng kiểm định t cho cả ba trường hợp đã xét. Lí do là vì:

1. Cỡ mẫu bất kì, có tổng thể phân bố chuẩn là ta có thể dùng kiểm định t.
2. Đối với trường hợp tổng thể phân bố chuẩn biết phương sai. Vì có điều kiện tổng thể phân bố chuẩn nên ta dùng được kiểm định t.
3. Đối với trường hợp tổng thể bất kì, cỡ mẫu lớn. Ta biết rằng $z = \dfrac{\bar{X} - \mu}{s_x/\sqrt{n}} \sim N(0,1)$. Nhưng ta cũng biết rằng khi cỡ mẫu lớn t_{n-1} cũng $\approx N(0,1)$. Bởi vậy nếu ta dùng $t = \dfrac{\bar{X} - \mu}{s_x/\sqrt{n}} \sim t_{n-1}$ thay cho $z \sim N(0,1)$ thì kết quả là gần như nhau. Xét ví dụ dưới đây để thấy rõ điều này.

So sánh t.test và z.test khi cỡ mẫu lớn

Câu hỏi

Người ta cho rằng chi tiêu giáo dục trung bình một năm của các hộ gia đình nước ta vào thời điểm 2010 là ít nhất 250 (nghìn). Từ dữ liệu ChiTieu2010.csv là mẫu điều tra ngẫu nhiên vài chục nghìn hộ gia đình ở nước ta trong năm 2010. Hãy kiểm định điều này với mức ý nghĩa 1%.

Solution

Gọi μ là trung bình chi tiêu cho giáo dục của các hộ ở nước ta trong năm 2010.

$$H_0 : \mu \geq 250 \qquad H_1 : \mu < 250$$

Ta so sánh kết quả khi dùng z.test và t.test:

```
> sd(ChiTieuGiaoDucTrongNam)
[1] 720.7803
> z.test(ChiTieuGiaoDucTrongNam,sigma.x=720.7803,mu=250,alt="less")

        One-sample z-Test

data:  ChiTieuGiaoDucTrongNam
z = -1.3258, p-value = 0.09246
alternative hypothesis: true mean is less than 250
95 percent confidence interval:
       NA 252.3724
sample estimates:
mean of x
 240.1428
```

```
> t.test(ChiTieuGiaoDucTrongNam,mu=250,alt="less")

        One Sample t-test

data:  ChiTieuGiaoDucTrongNam
t = -1.3258, df = 9397, p-value = 0.09247
alternative hypothesis: true mean is less than 250
95 percent confidence interval:
     -Inf 252.3736
sample estimates:
mean of x
 240.1428
```

Nhận xét rằng, P - giá trị khi dùng xấp xỉ theo z.test là 0.09246, còn khi dùng xấp xỉ theo t.test là 0.09247. Hai giá trị này có sự khác biệt rất nhỏ và sự khác biệt đó không hề ảnh hưởng đến việc chấp nhận hay bác bỏ H_0.

Kết luận và lưu ý

Khi kiểm định so sánh trung bình:

1. Nếu biết tổng thể có phân bố chuẩn ta dùng t.test.
2. Nếu cỡ mẫu lớn, không cần biết phân bố của tổng thể ta dùng t.test (hay gặp trong thực tế).
3. Khi biết phương sai tổng thể và tổng thể phân bố chuẩn: có thể dùng t.test hoặc z.test đều được, nhưng nên dùng z.test vì kết quả chính xác hơn (ít gặp trong thực tế).

Lưu ý, để đơn giản, ta sẽ bỏ qua các trường hợp phải dùng lệnh tsum.test, zsum.test trong thực hành bài tập, kiểm tra cũng như thi. Và do vậy, ta chỉ dùng t.test và z.test.