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. Kiểm định giả thuyết về tỉ lệ tổng thể

2. Kiểm định với phần mềm thống kê R

Câu hỏi

Giả sử một tờ báo nói rằng tỉ lệ xin được việc làm của TLU là 60%. Nếu ta muốn kiểm định điều này thì ta làm thế nào?

Câu hỏi

Giả sử một tờ báo nói rằng tỉ lệ xin được việc làm của TLU là 60%. Nếu ta muốn kiểm định điều này thì ta làm thế nào?

Solution

Để trả lời câu hỏi trên ta cần chọn ngẫu nhiên một mẫu sinh viên đã ra trường, giử sử n bạn, trong đó có X bạn có việc làm.
Cặp giả thuyết kiểm định là gì?

Solution

Gọi P là tỉ lệ xin được việc của sinh viên TLU.

$$H_0 : P = 0.6 \qquad\qquad H_1 : P \neq 0.6$$

Giả sử H_0 đúng, tức là $P = 0.6$. Khi đó $\hat{p} = \dfrac{X}{n}$ sẽ chủ yếu nằm trong khoảng nào?

Solution

Gọi P là tỉ lệ xin được việc của sinh viên TLU.

$$H_0 : P = 0.6 \qquad H_1 : P \neq 0.6$$

Giả sử H_0 đúng, tức là $P = 0.6$. Khi đó $\hat{p} = \dfrac{X}{n}$ sẽ chủ yếu nằm trong khoảng nào?
Với mức sai lầm $\alpha = 0.05$ cho trước, ta sẽ vạch ra miền chấp nhận và miền bác bỏ H_0 dựa trên phân phối của $\hat{p}$.

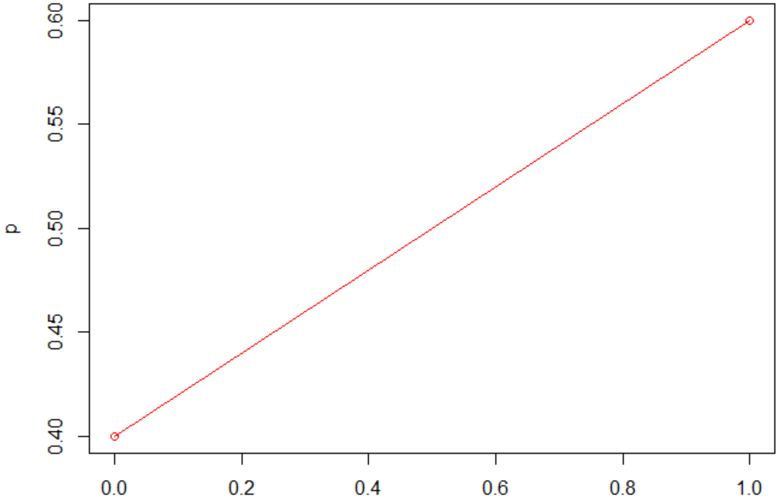
Solution

Gọi P là tỉ lệ xin được việc của sinh viên TLU.

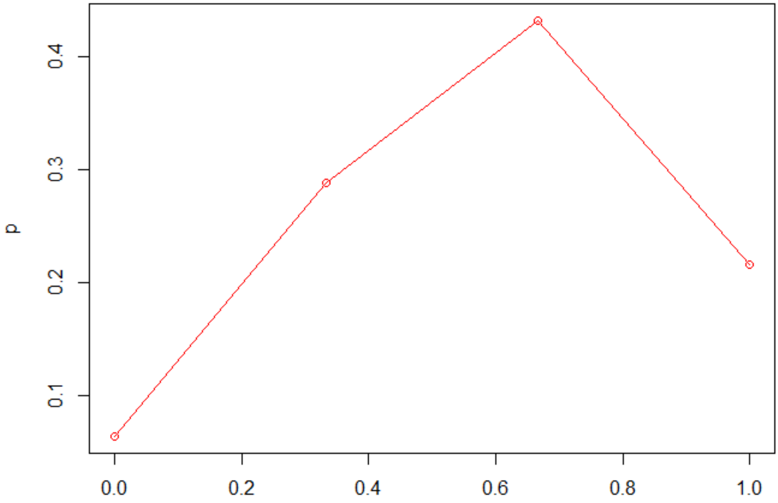
$$H_0 : P = 0.6 \qquad\qquad H_1 : P \neq 0.6$$

Giả sử H_0 đúng, tức là $P = 0.6$. Khi đó $\hat{p} = \dfrac{X}{n}$ sẽ chủ yếu nằm trong khoảng nào?
Với mức sai lầm $\alpha = 0.05$ cho trước, ta sẽ vạch ra miền chấp nhận và miền bác bỏ H_0 dựa trên phân phối của $\hat{p}$.
Phân phối của $\hat{p}$ như thế nào?

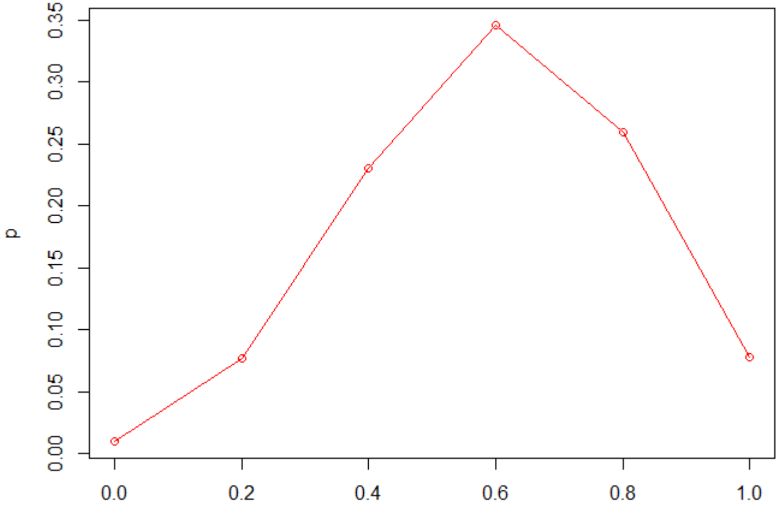
n=1



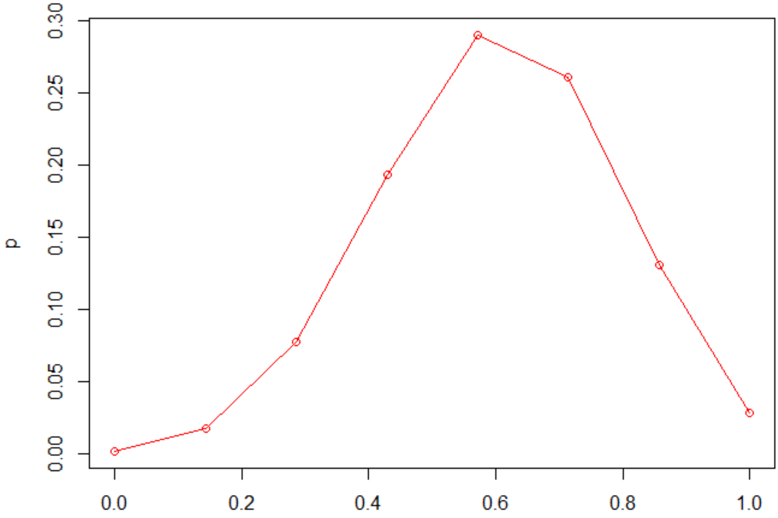
n=3



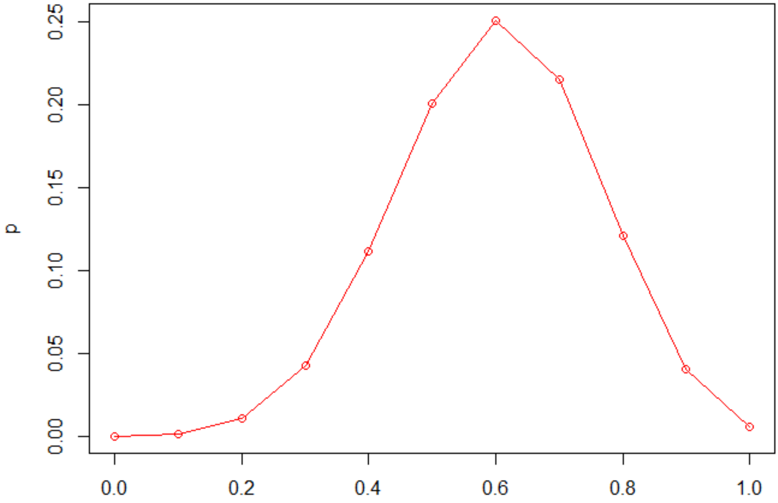
n=5



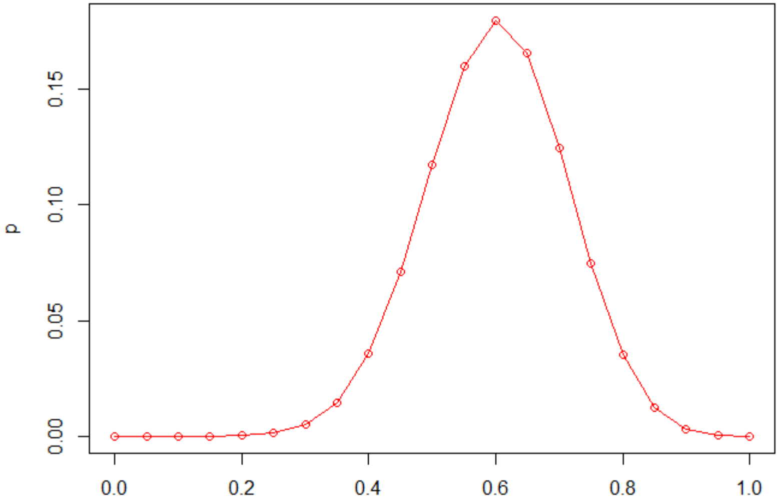
n=7



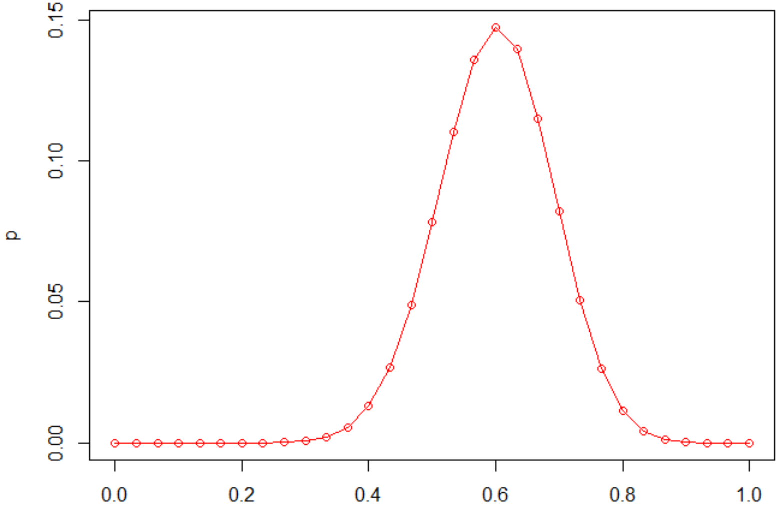
n=10



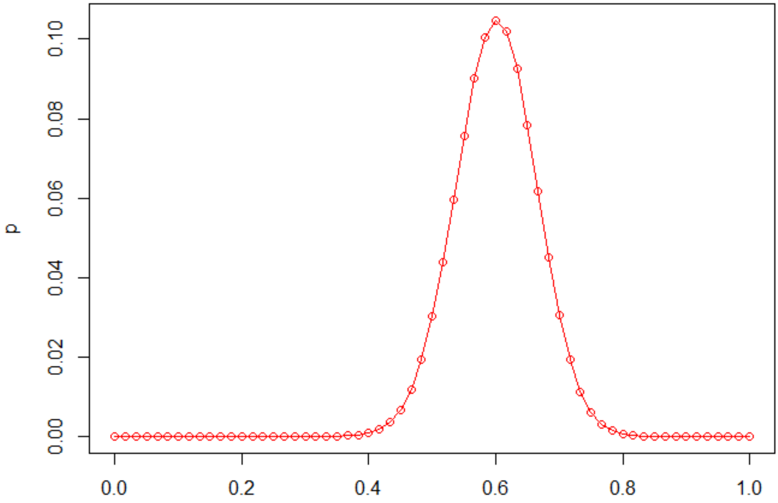
n=20



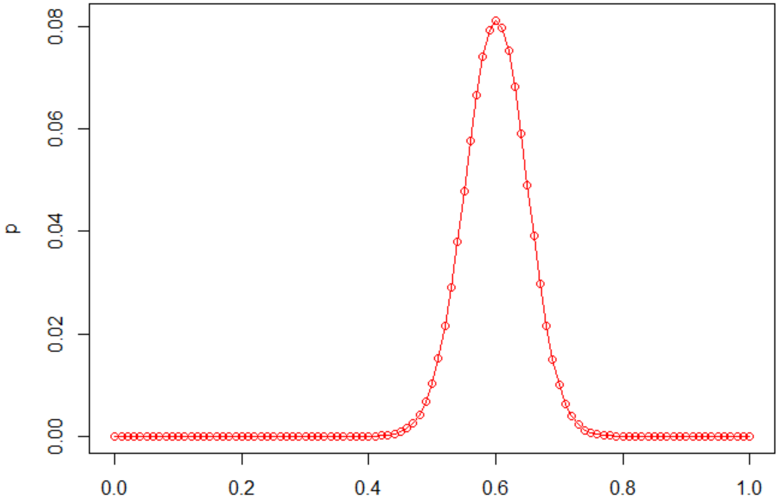
n=30



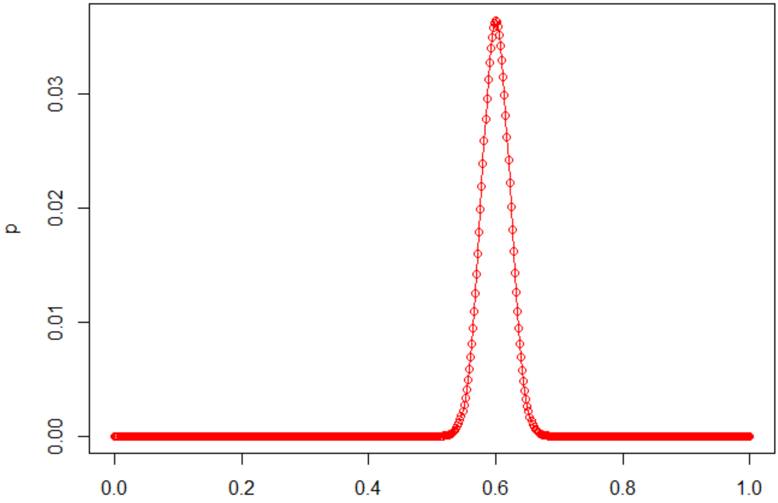
n=60



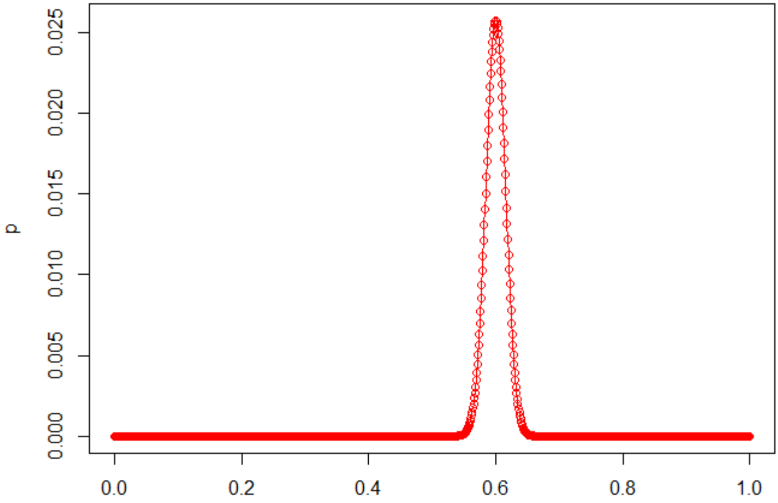
n=100



n=500



n=1000



Solution

Ta thấy khi n lớn lên, $\hat{p}$ *có phân bố xấp xỉ* $N(0.6, \dfrac{P(1-P)}{n})$ *hay*

$$z = \frac{P - \hat{p}}{\sqrt{0.6 * 0.4/n}} \approx N(0,1)$$

Từ đó khả năng $z \in [-z_{0.025}; z_{0.025}]$ *là* 95%. *Do vậy,*

- *Nếu* H_0 *đúng, rất ít khi có chuyện* $z \notin [-1.96; 1.96]$, *nên nếu z tính ra trên mẫu chọn ngẫu nhiên nằm ngoài khoảng trên ta bác bỏ* H_0 *mà sai lầm mắc phải là không quá 5%.*
- *Nếu* $z \in [-1.96; 1.96]$ *ta vẫn bác bỏ* H_0 *thì sai lầm cao hơn* 5%.

Solution

Ta thấy khi n lớn lên, $\hat{p}$ *có phân bố xấp xỉ* $N(0.6, \dfrac{P(1-P)}{n})$ *hay*

$$z = \frac{P - \hat{p}}{\sqrt{0.6 * 0.4/n}} \approx N(0,1)$$

Từ đó khả năng $z \in [-z_{0.025}; z_{0.025}]$ *là* 95%. *Do vậy,*

- *Nếu* H_0 *đúng, rất ít khi có chuyện* $z \notin [-1.96; 1.96]$, *nên nếu z tính ra trên mẫu chọn ngẫu nhiên nằm ngoài khoảng trên ta bác bỏ* H_0 *mà sai lầm mắc phải là không quá 5%.*
- *Nếu* $z \in [-1.96; 1.96]$ *ta vẫn bác bỏ* H_0 *thì sai lầm cao hơn* 5%.

Chẳng hạn, nếu điều tra 1000 cựu sinh viên TLU, thấy có 800 bạn đang có việc làm, thì ta chấp nhận H_0 *hay bác bỏ* H_0*?*

Khi cỡ mẫu lớn

Ta đã biết rằng khi cỡ mẫu đủ lớn (và thỏa mãn $np \geq 5, n(1-p) \geq 5$) thì đại lượng

$$z = \frac{P - p}{\sqrt{p(1-p)/n}}$$

xấp xỉ phân phối chuẩn hóa $Z = N(0,1)$. Nên ở đây ta sẽ dùng thống kê Z làm quy luật phân phối để kiểm định.

H_0	H_1	Giá trị thống kê z	Qui luật bác bỏ H_0	p-giá trị
$P \leq p_0$	$P > p_0$	$z = \dfrac{\hat{P} - p_0}{\sqrt{p_0(1-p_0)/n}}$	$z > z_\alpha$	$P(Z > z)$
$P \geq p_0$	$P < p_0$		$z < -z_\alpha$	$P(Z < z)$
$P = p_0$	$\mu \neq \mu_0$		$\|z\| > z_{\alpha/2}$	$2P(Z > \|z\|)$

Example

Một hãng sản xuất ti vi công bố rằng 95% số sản phẩm của họ không phải sửa chữa trong 5 năm đầu sử dụng. Một một kê ngẫu nhiên 200 gia đình sử dụng ti vi của hãng này cho thấy có 18 gia đình nói rằng họ đã phải sửa ti vi trong vòng 5 năm sử dụng. Có thể bác bỏ khẳng định của hãng ti vi trên không?

Solution

Gọi P là tỉ lệ ti vi phải sửa chữa trong 5 năm đầu sử dụng.

1. *Cặp giả thiết:* $H_0 : P = 0.95 \qquad H_1 : P \neq 0.95.$
2. *Giá trị kiểm định* $z = \dfrac{182/200 - 0.95}{\sqrt{0.95.(1-0.95)/200}} = -2.595543.$
3. *Giá trị tới hạn:* $-z_{0.025} = -1.96, z_{0.025} = 1.96.$ *Ta thấy* $z < -z_{0.025}$ *nên bác bỏ* H_0.
   Ta cũng có thể tính $P - value = 2P(Z > |z|) = 2.P(Z > 2.596)$ $= pnorm(2.595543, 0, 1, F) \approx 0.004722 < 0.05$ *nên bác bỏ* H_0.
4. *Vậy, tại mức ý nghĩa 5%, khẳng định của công ty là không đúng.*

1. Kiểm định giả thuyết về tỉ lệ tổng thể

2. Kiểm định với phần mềm thống kê R

Khi cần kiểm định:

1. về trung bình tổng thể:
   - có phân phối chuẩn, biết phương sai σ^2 ta dùng

$$z.test(x, mu = \mu_0, sigma.x = \sigma, alt = "less" / "two.sided" / "greater")$$

   - có phân phối chuẩn nhưng không biết phương sai:

$$t.test(x, mu = \mu_0, alt = "less" / "two.sided" / "greater")$$

   - không biết có phân phối chuẩn hay không, nhưng cỡ mẫu lớn hơn 30:

$$t.test(x, mu = \mu_0, alt = "less" / "two.sided" / "greater")$$

2. về tỉ lệ tổng thể

$$prop.test(x, n = \text{ cỡ mẫu }, p = p_0, alt, correct = ?)$$

   ở đó correct = FALSE nếu 5 $\leq$ số phần tử có dấu hiệu T $\leq n - 5$, và là TRUE trong trường hợp còn lại.

Trong đó $alt = "less"$ cho bài toán kiểm định bên trái, $alt = "two.sided"$ cho bài toán hai bên, $alt = "greater"$ cho bài toán bên phải.