THỐNG KÊ ỨNG DỤNG

ĐỖ LÂN

dolan@tlu.edu.vn
Đại học Thủy Lợi

Ngày 14 tháng 9 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

Phần VI

Phân phối của tỷ lệ mẫu và ước lượng khoảng cho tỷ lệ tổng thể

Mục lục

Biến ngẫu nhiên có phân phối nhị thức

Định nghĩa

Phép thử nhị thức *là phép thử có các đặc điểm sau:*

Biến ngẫu nhiên có phân phối nhị thức

Định nghĩa

Phép thử nhị thức *là phép thử có các đặc điểm sau:*

- *Phép thử bao gồm n thử nghiệm giống hệt nhau.*

Biến ngẫu nhiên có phân phối nhị thức

Định nghĩa

Phép thử nhị thức *là phép thử có các đặc điểm sau:*

- *Phép thử bao gồm n thử nghiệm giống hệt nhau.*
- *Mỗi thử nghiệm này chỉ có kết quả là "thành công" hoặc "thất bại".*

Biến ngẫu nhiên có phân phối nhị thức

Định nghĩa

Phép thử nhị thức *là phép thử có các đặc điểm sau:*

- *Phép thử bao gồm n thử nghiệm giống hệt nhau.*
- *Mỗi thử nghiệm này chỉ có kết quả là "thành công" hoặc "thất bại".*
- *Xác suất "thành công" trong mỗi phép thử đều là p (và xác suất "thất bại" là* $q = 1 - p$*).*

Biến ngẫu nhiên có phân phối nhị thức

Định nghĩa

Phép thử nhị thức *là phép thử có các đặc điểm sau:*

- *Phép thử bao gồm n thử nghiệm giống hệt nhau.*
- *Mỗi thử nghiệm này chỉ có kết quả là "thành công" hoặc "thất bại".*
- *Xác suất "thành công" trong mỗi phép thử đều là p (và xác suất "thất bại" là $q = 1 - p$).*
- *Các thử nghiệm là độc lập lẫn nhau (kết quả của thử nghiệm này không ảnh hưởng tới kết quả của thử nghiệm khác).*

Định lí

Gọi $X =$ số lần "thành công" trong n thử nghiệm của phép thử nhị thức. thì X được gọi là biến ngẫu nhiên có phân phối nhị thức (biến ngẫu nhiên nhị thức), ký hiệu $X \sim B(n, p)$. Khi đó, xác suất để có k lần thành công trong n thử nghiệm là

$$P(X = k) = C_n^k p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

Chú ý

Nếu ta gọi Y là số lần thất bại thì ta cũng có $Y \sim B(n, q)$ với $q = 1 - p$.

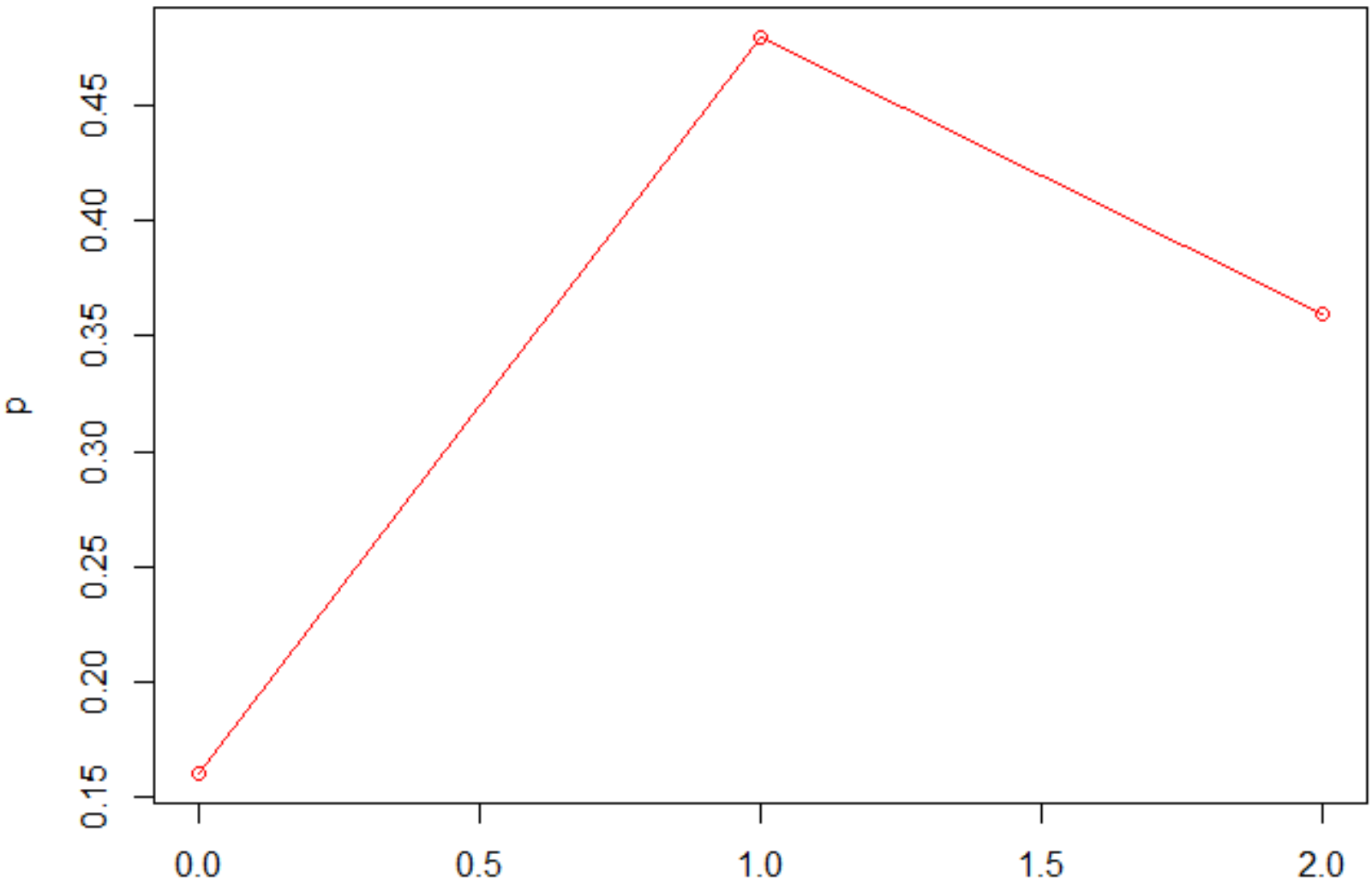
Example

Giả sử tỉ lệ thanh niên có việc làm ở một nước là 60 %. Chọn ngẫu nhiên n người và gọi X là số người có việc làm trong số n người đó. Hãy lập bảng phân phối xác suất cho X khi:

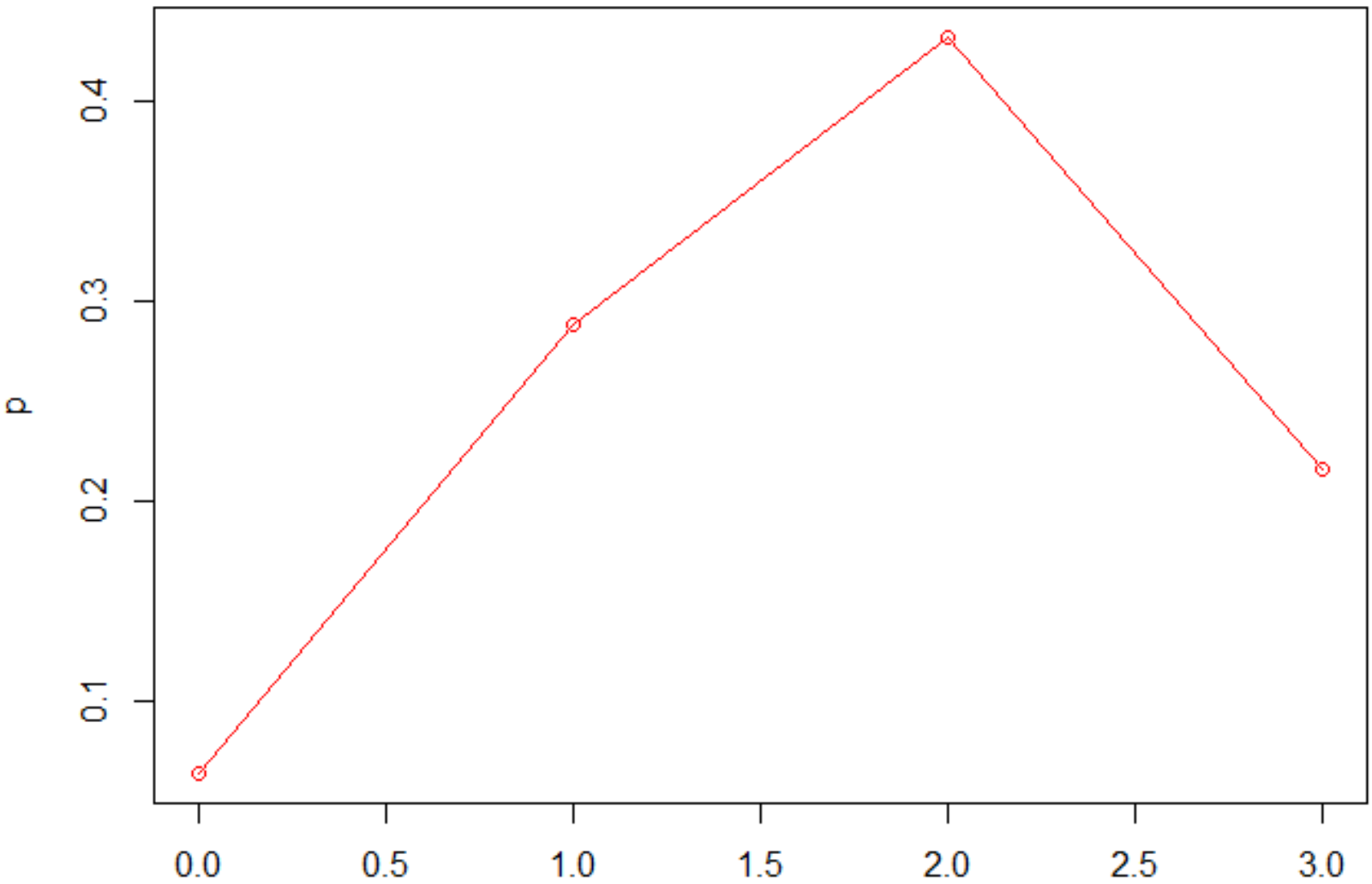
1. $n = 1$
2. $n = 3$
3. $n = 5$
4. $n = 7$

Sau đó vẽ biểu đồ xác suất cho nó, dạng tán xạ.

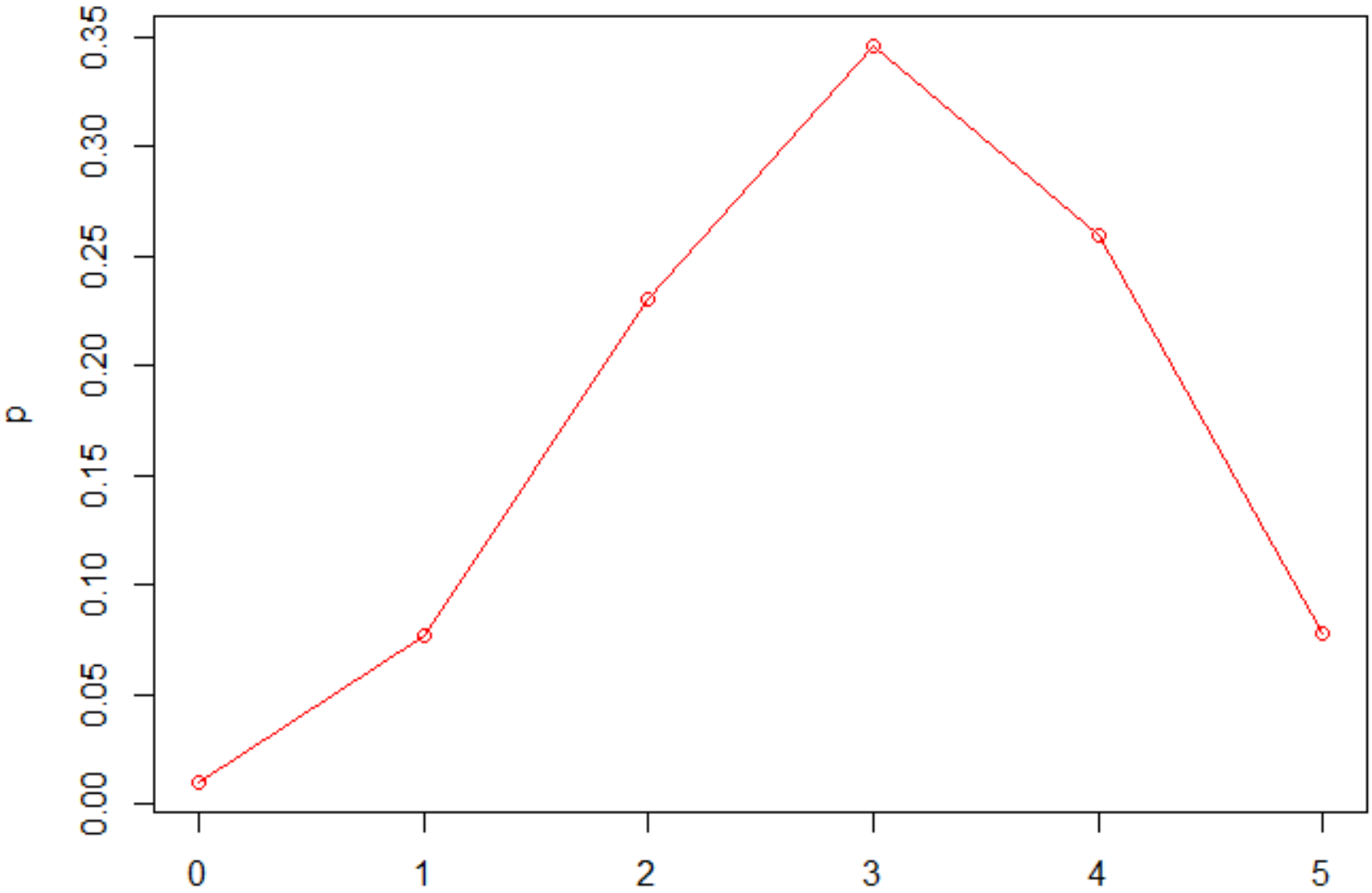
n=3



n=5



n=7



Câu hỏi dẫn nhập

Giả sử ta cần ước lượng tỉ lệ có việc làm sinh viên đã tốt nghiệp TLU.

1. *Ta làm thế nào để có điều đó?*

Câu hỏi dẫn nhập

Giả sử ta cần ước lượng tỉ lệ có việc làm sinh viên đã tốt nghiệp TLU.

1. *Ta làm thế nào để có điều đó?*
2. *Ta có thể điều tra hết lượng sinh viên đã tốt nghiệp không?*

Câu hỏi dẫn nhập

Giả sử ta cần ước lượng tỉ lệ có việc làm sinh viên đã tốt nghiệp TLU.

1. *Ta làm thế nào để có điều đó?*
2. *Ta có thể điều tra hết lượng sinh viên đã tốt nghiệp không?*
   $\longrightarrow$ *Phải chọn mẫu.*
3. *Từ mẫu đã chọn, giả sử có n người và có k bạn có việc làm. Vậy tỉ lệ có việc làm là bao nhiêu?*

Câu hỏi dẫn nhập

Giả sử ta cần ước lượng tỉ lệ có việc làm sinh viên đã tốt nghiệp TLU.

1. *Ta làm thế nào để có điều đó?*
2. *Ta có thể điều tra hết lượng sinh viên đã tốt nghiệp không?*
   $\longrightarrow$ *Phải chọn mẫu.*
3. *Từ mẫu đã chọn, giả sử có n người và có k bạn có việc làm. Vậy tỉ lệ có việc làm là bao nhiêu?Con số đó có mức độ chính xác cỡ nào?*

Câu hỏi dẫn nhập

Giả sử ta cần ước lượng tỉ lệ có việc làm sinh viên đã tốt nghiệp TLU.

1. *Ta làm thế nào để có điều đó?*
2. *Ta có thể điều tra hết lượng sinh viên đã tốt nghiệp không?*
   $\longrightarrow$ *Phải chọn mẫu.*
3. *Từ mẫu đã chọn, giả sử có n người và có k bạn có việc làm. Vậy tỉ lệ có việc làm là bao nhiêu?Con số đó có mức độ chính xác cỡ nào?*

Mục tiêu: ước lượng tỉ lệ của một dấu hiệu nào đó với độ chính xác kiểm soát được.

Câu hỏi dẫn nhập

Giả sử ta cần ước lượng tỉ lệ có việc làm sinh viên đã tốt nghiệp TLU.

1. *Ta làm thế nào để có điều đó?*
2. *Ta có thể điều tra hết lượng sinh viên đã tốt nghiệp không?*
   $\longrightarrow$ *Phải chọn mẫu.*
3. *Từ mẫu đã chọn, giả sử có n người và có k bạn có việc làm. Vậy tỉ lệ có việc làm là bao nhiêu?Con số đó có mức độ chính xác cỡ nào?*

Mục tiêu: ước lượng tỉ lệ của một dấu hiệu nào đó với độ chính xác kiểm soát được.
Hướng giải quyết:

1. *Tìm hiểu về phân phối của tỉ lệ mẫu* $\hat{p} = \dfrac{k}{n}$.
2. *Đưa ra được công thức tính đơn giản cho tỉ lệ tổng thể với xác suất đúng cho trước.*

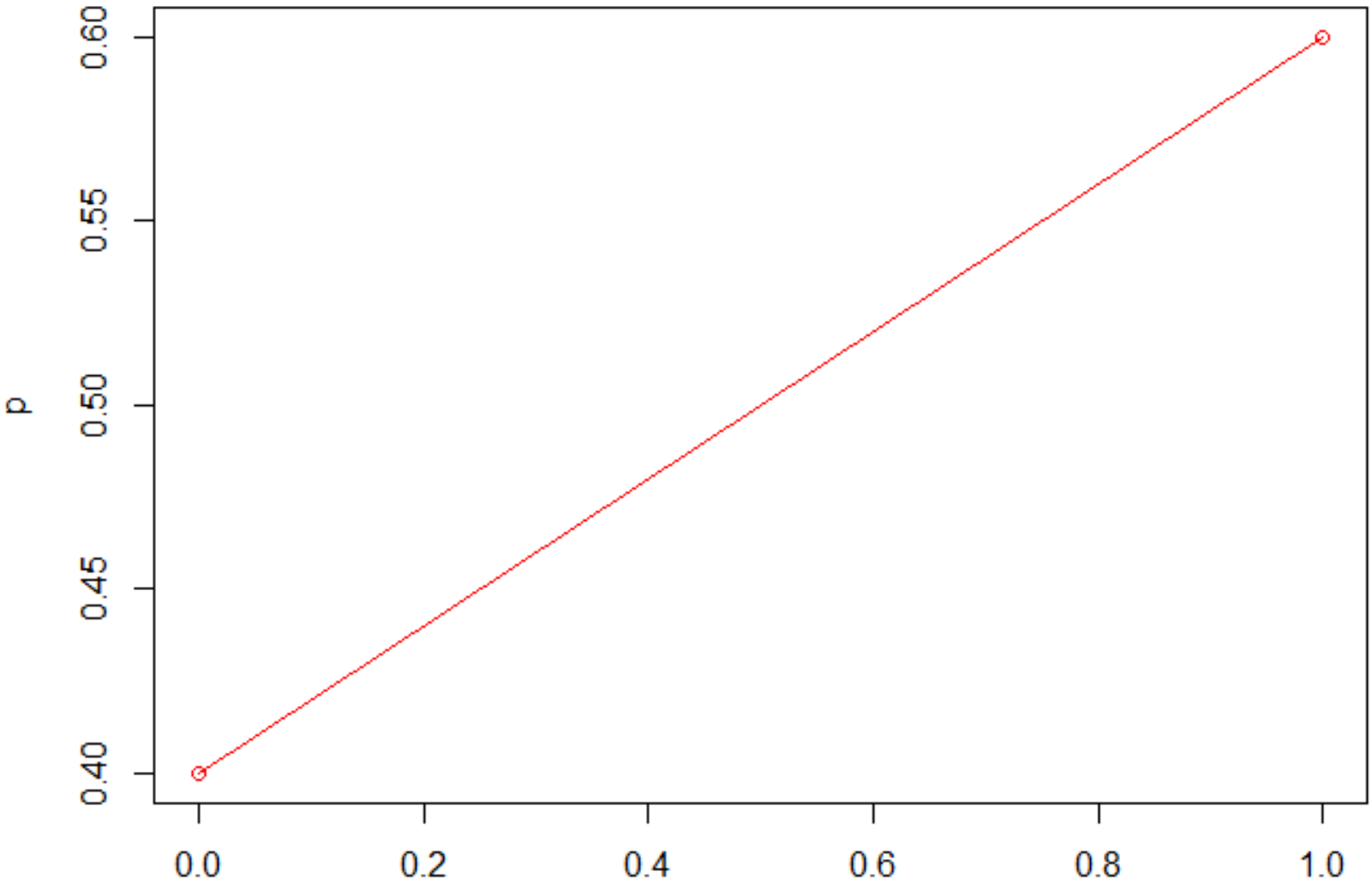
Example

Giả sử tỉ lệ sinh viên ra trường có việc làm của một trường đại học A là 60 %. Chọn ngẫu nhiên n người và gọi X là số người có việc làm trong số n người đó, $\hat{p} = \frac{X}{n}$. Hãy lập bảng phân phối xác suất cho $\hat{p}$ khi:

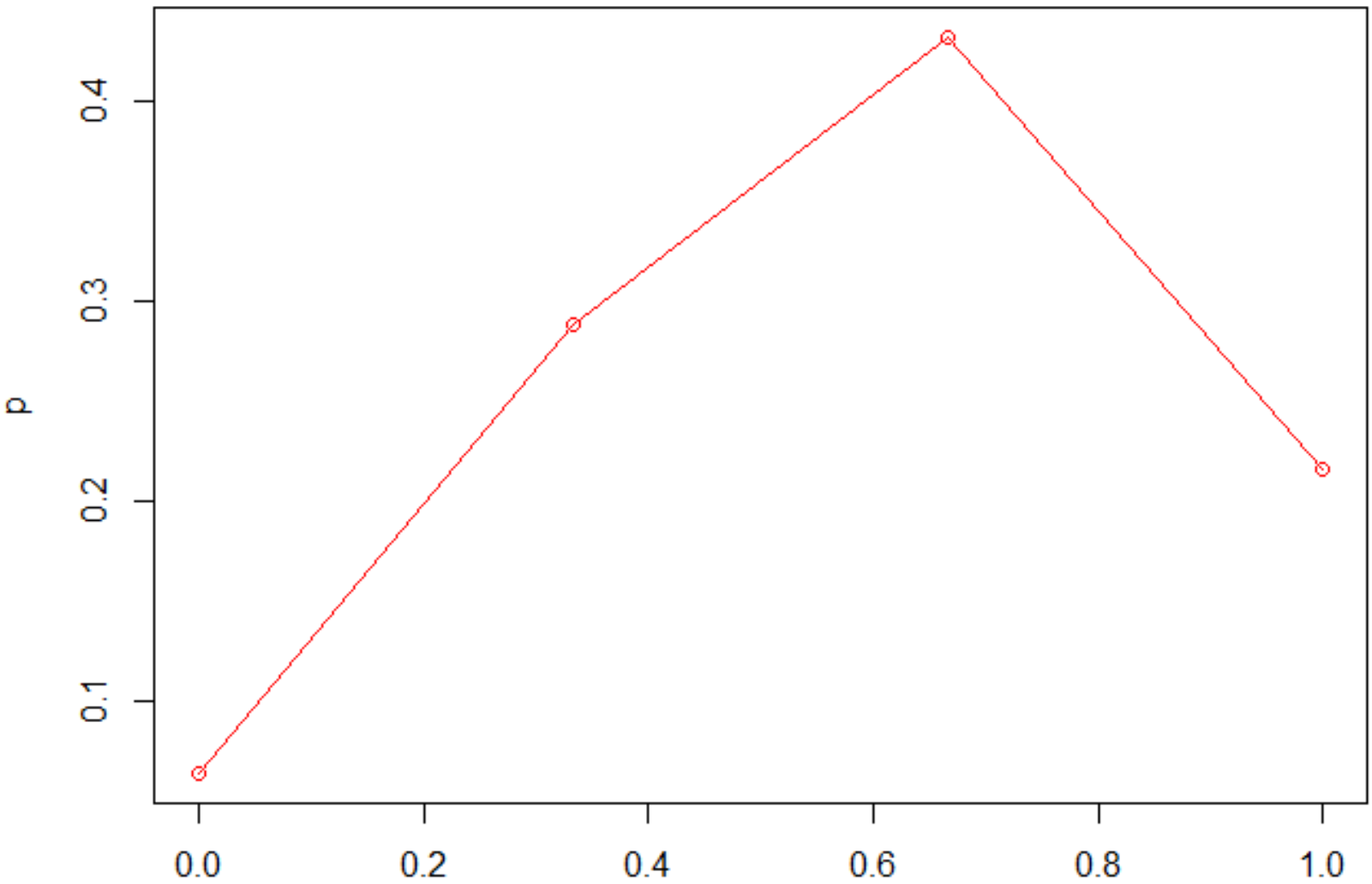
1. $n = 1$
2. $n = 3$
3. $n = 5$
4. $n = 7$

Sau đó vẽ biểu đồ xác suất cho nó, dạng tán xạ.
Thử cho trường hợp n lớn hơn, $n = 100, 1000 \ldots$ với sự trợ giúp của R.
Hãy đưa ra kết luận về quy luật phân phối của $\hat{p}$ về hình dáng , sự tập trung.

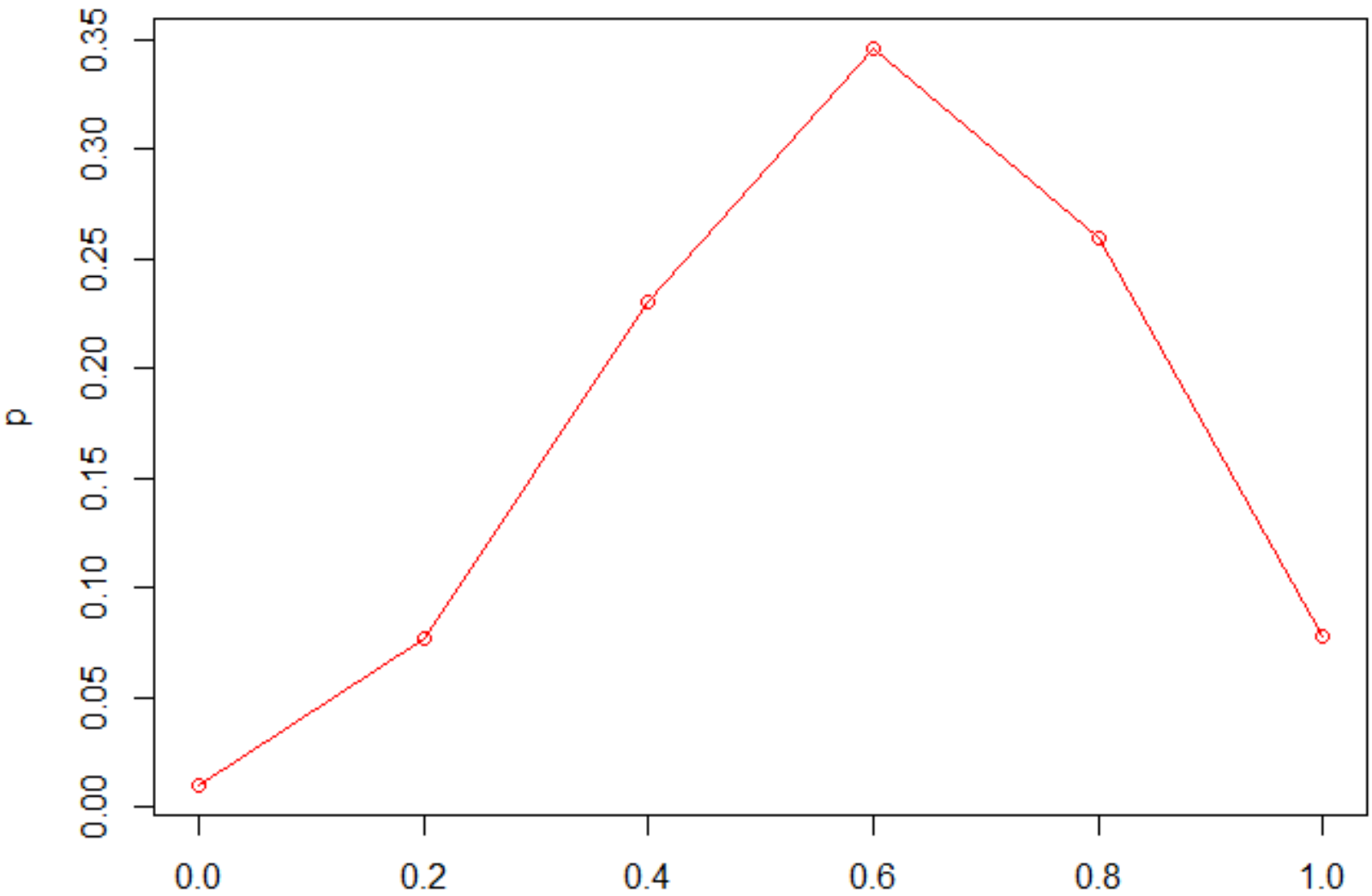
n=1



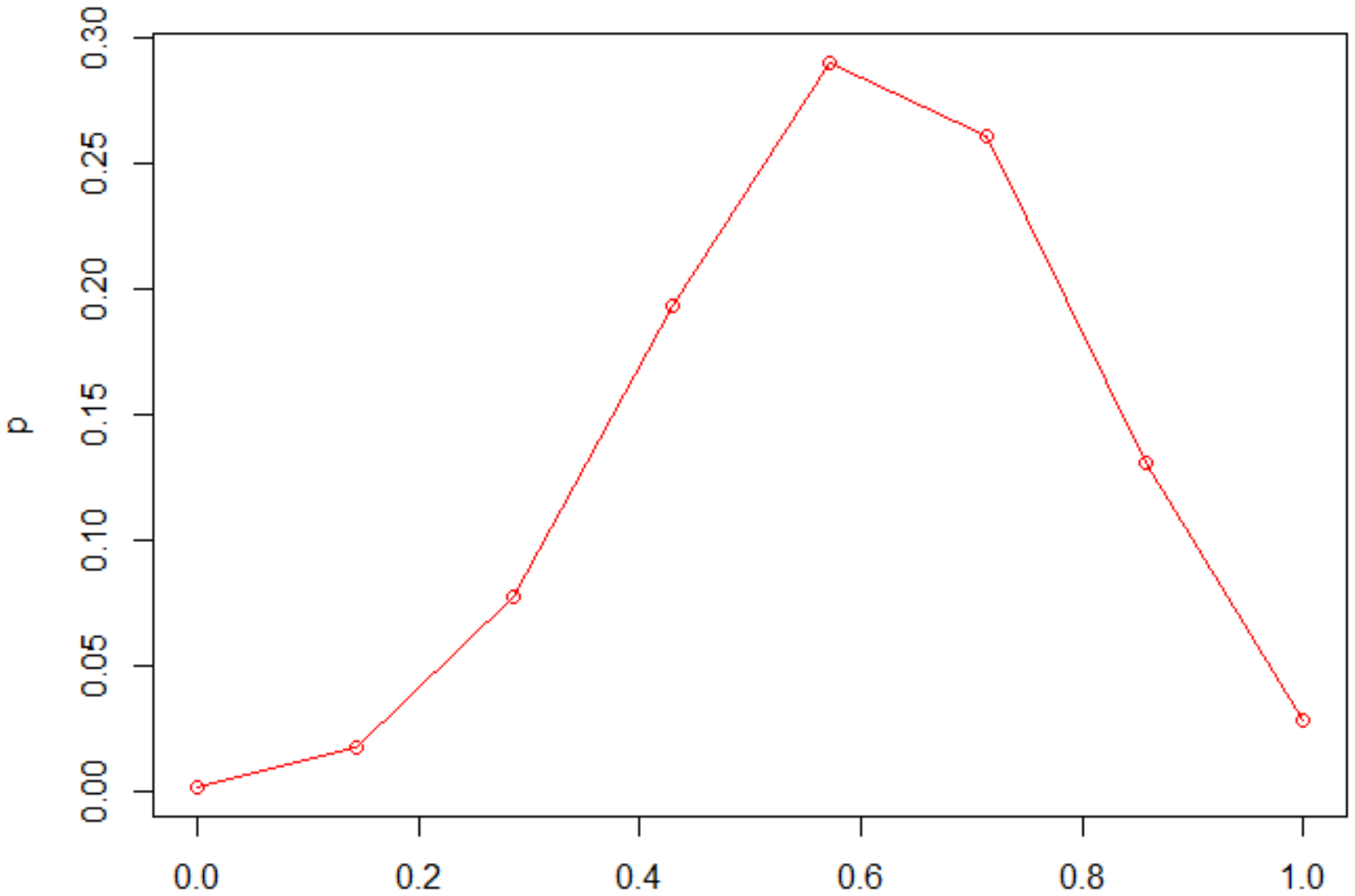
n=3



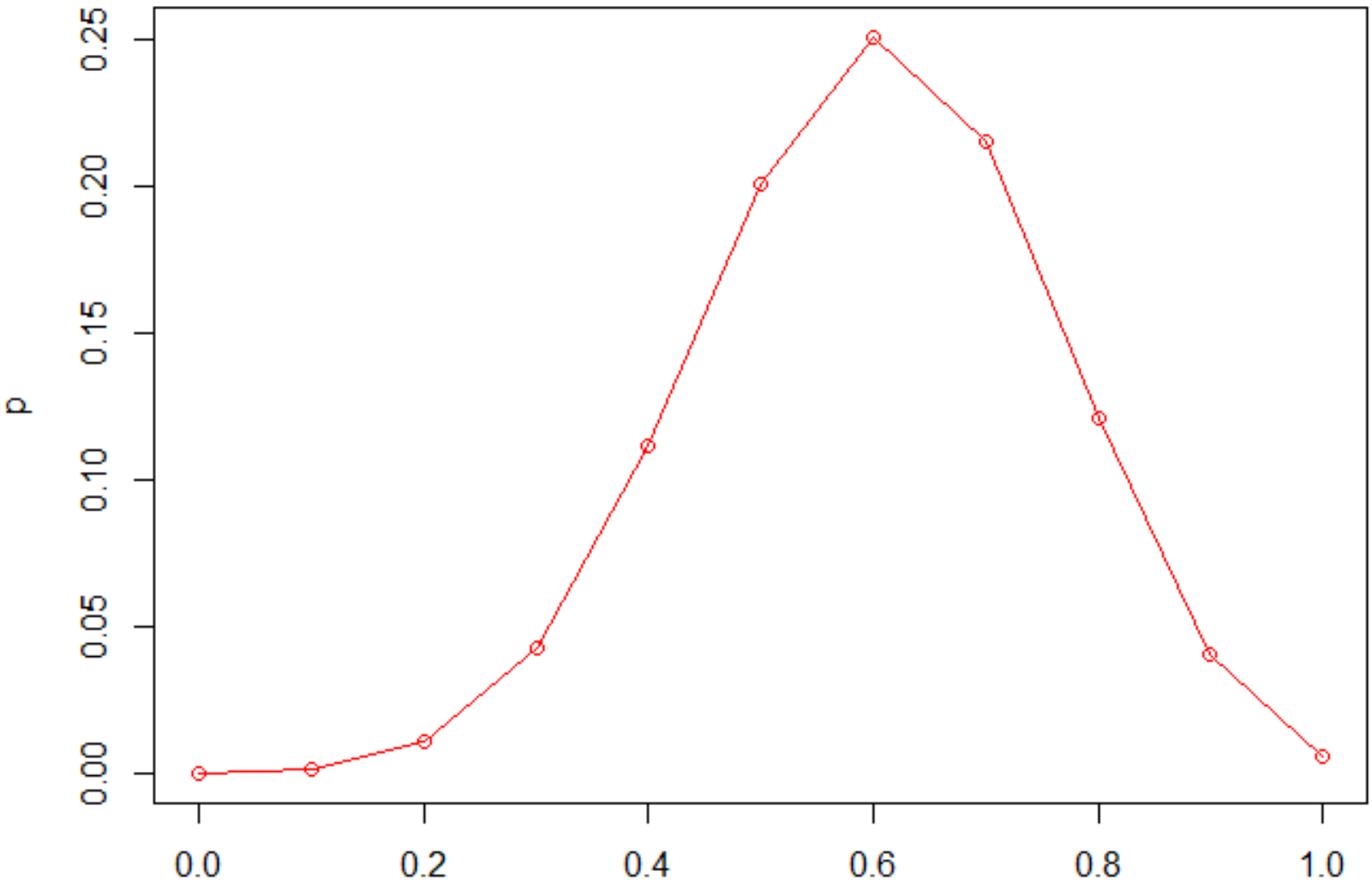
n=5



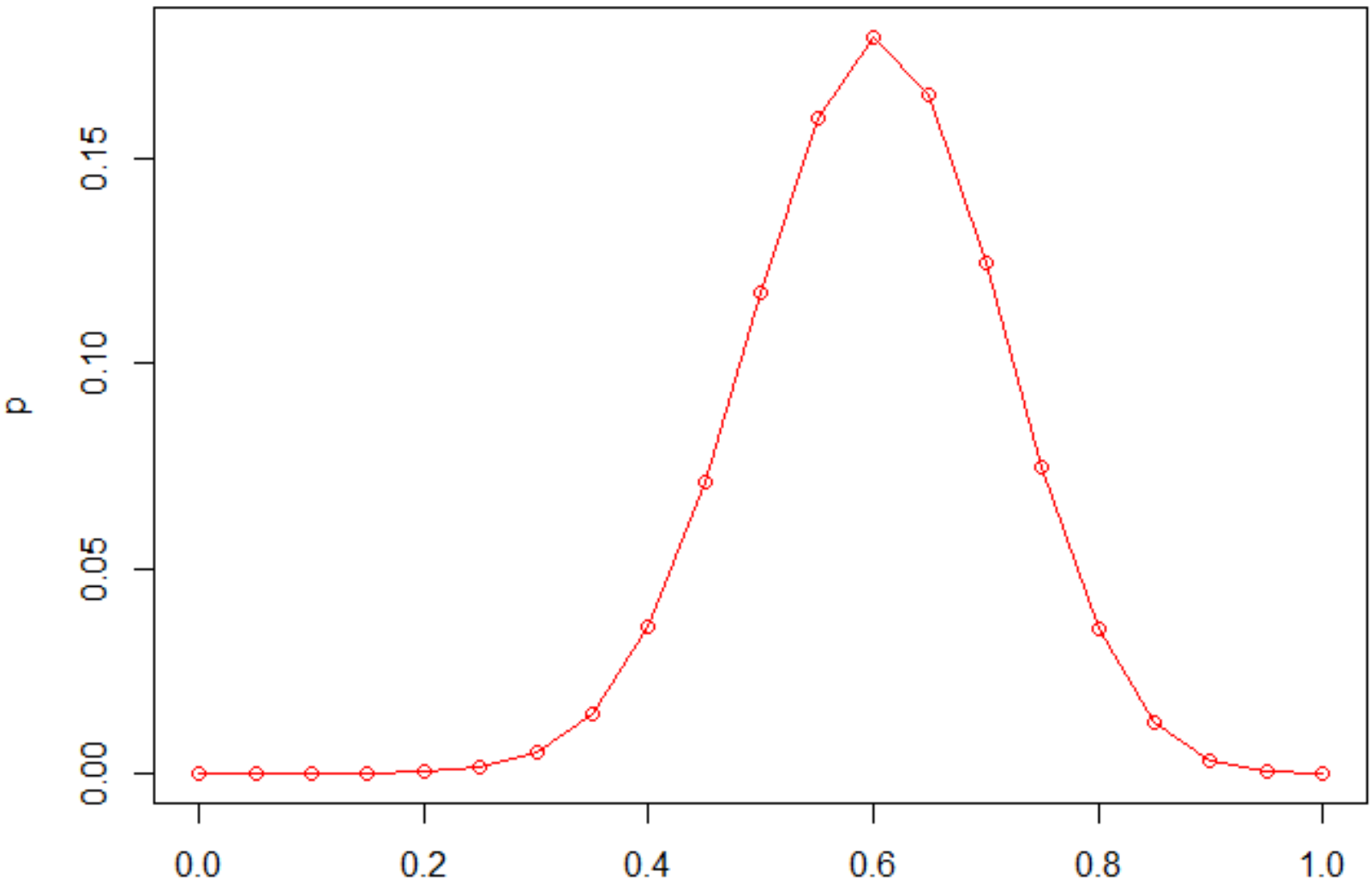
n=7



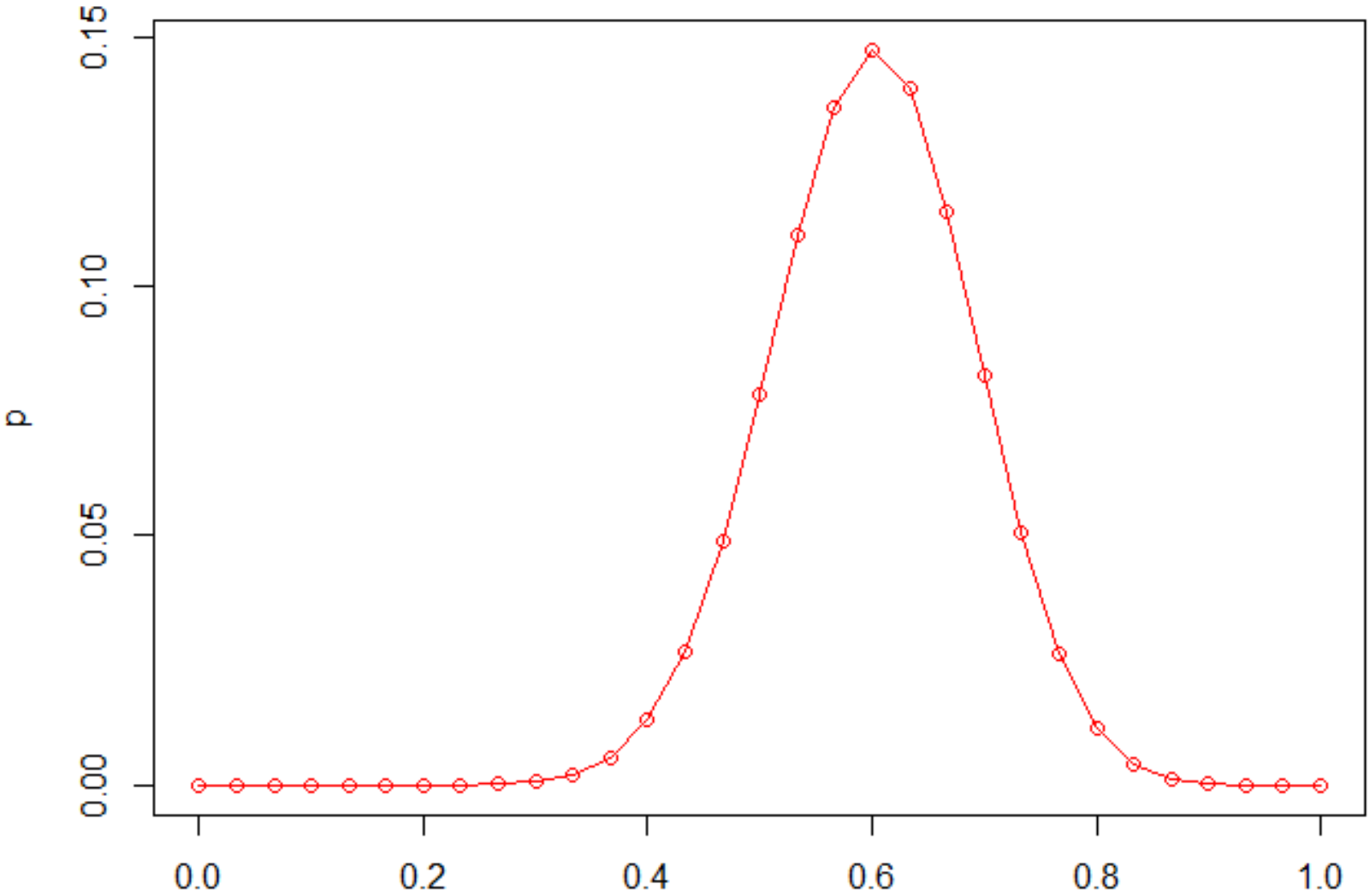
n=10



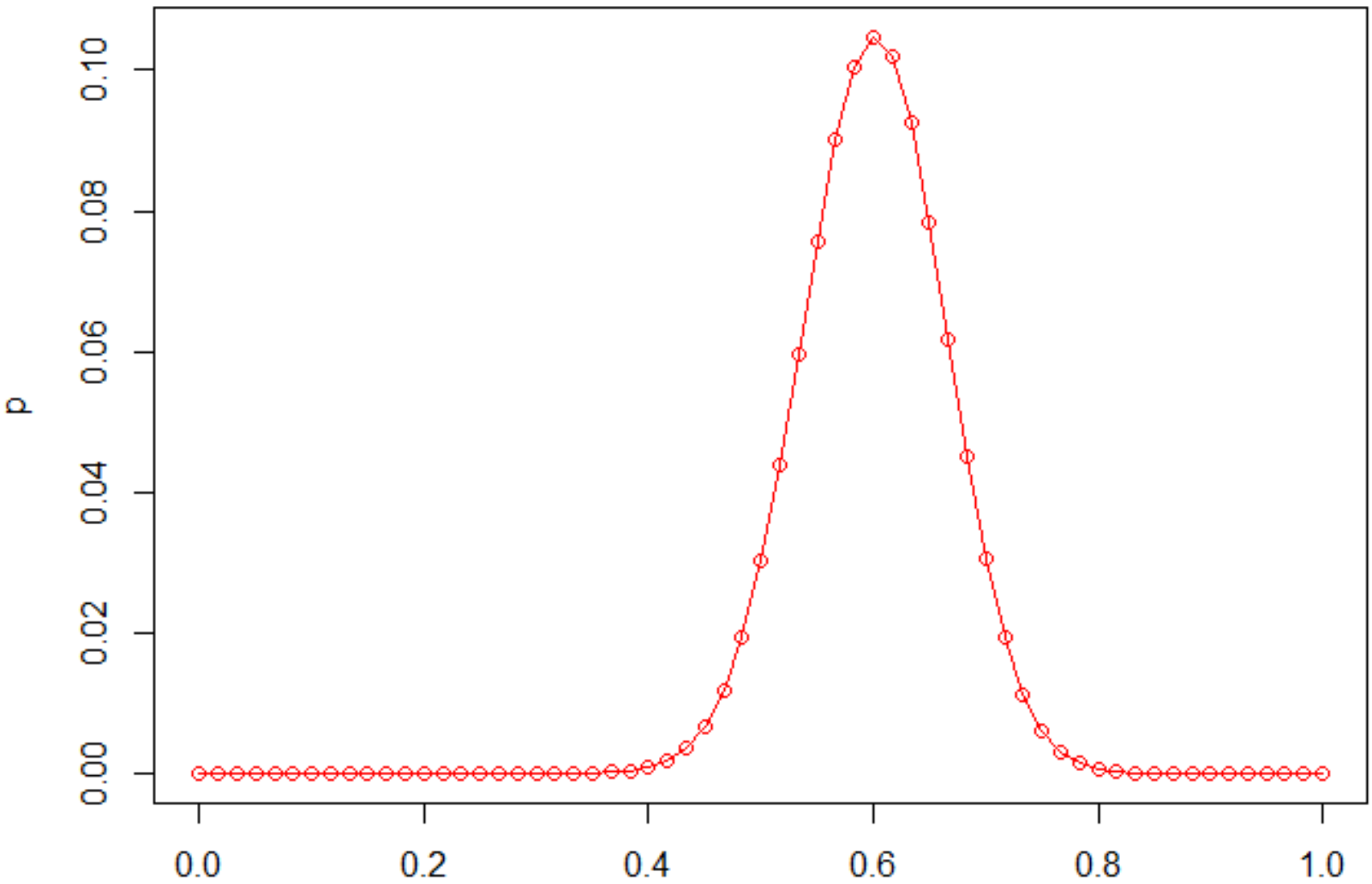
n=20



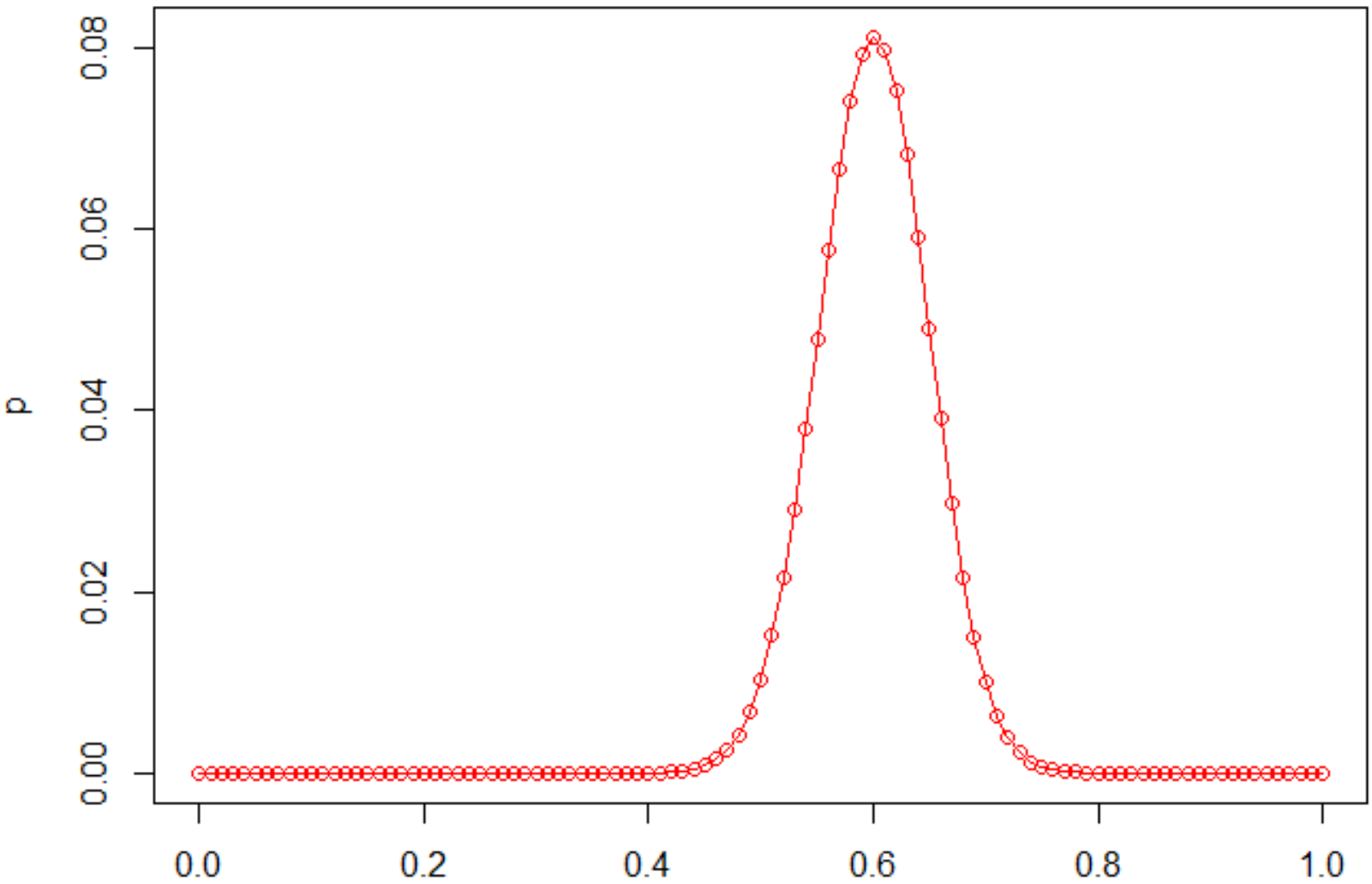
n=30



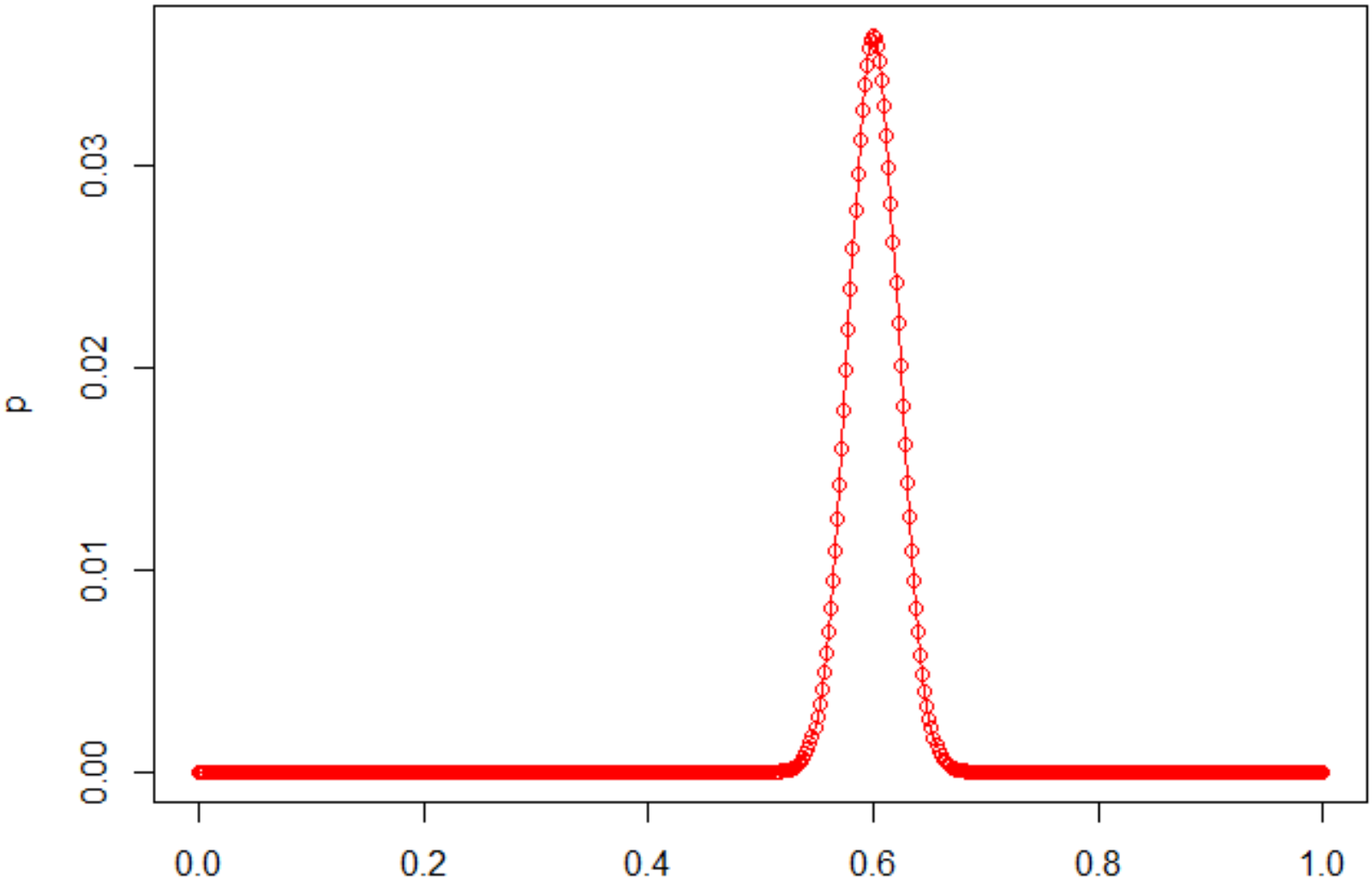
n=60



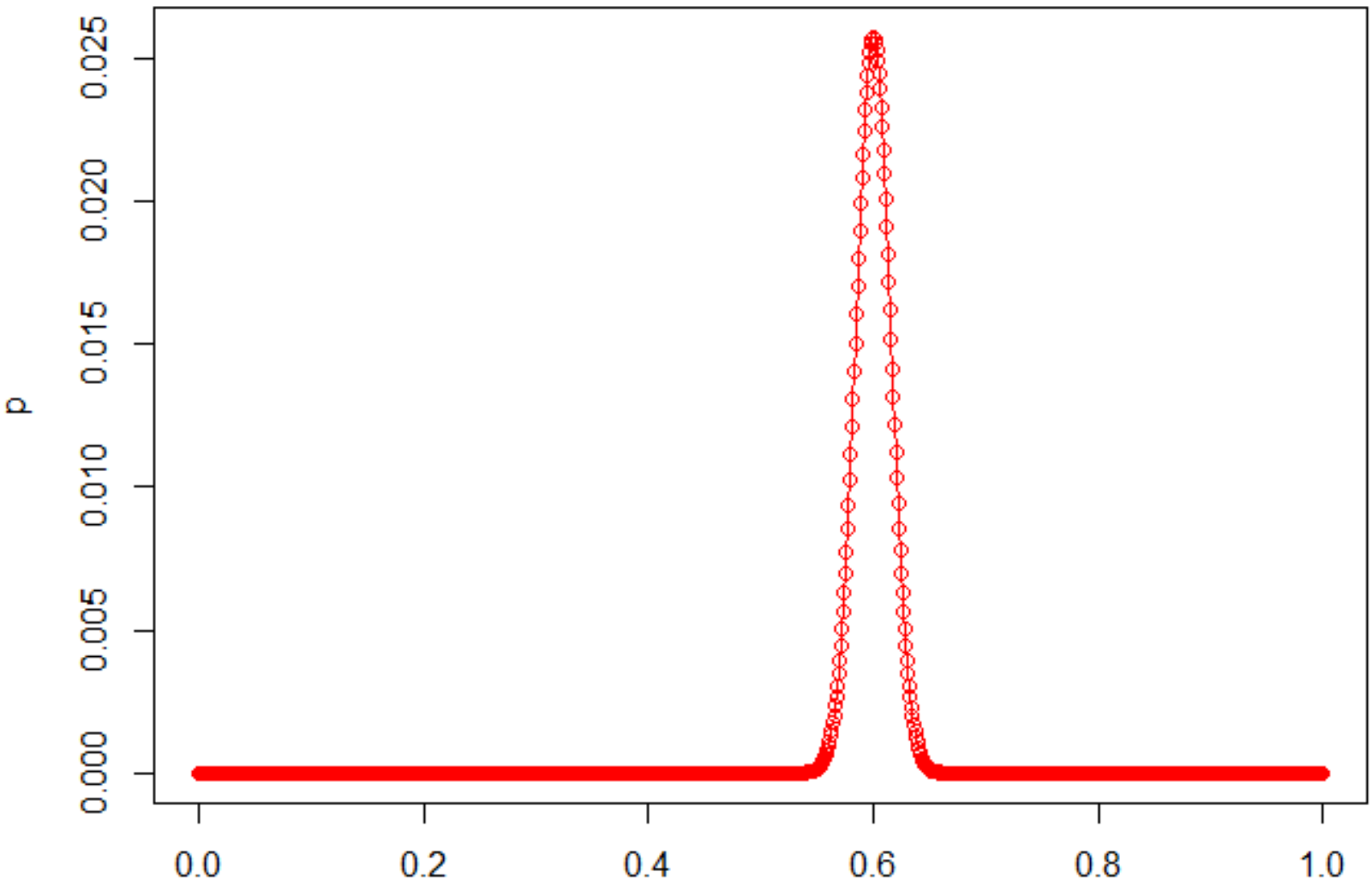
n=100



n=500



n=1000



Theorem

Với p là tỉ lệ của tổng thể có dấu hiệu T.
Khi cỡ mẫu n đủ lớn và thỏa mãn $5 \leq np \leq n-5$ thì tỉ lệ mẫu $\hat{p}$ xấp xỉ phân phối chuẩn $N(p, \frac{p(1-p)}{n})$. Do đó đại lượng

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n})}} \sim N(0,1)$$

1. Biến ngẫu nhiên nhị thức

2. Ước lượng khoảng cho tỉ lệ tổng thể
   - Ước lượng khoảng tin cậy 95%

3. Ước lượng điểm cho tỉ lệ tổng thể
   - Ước lượng khoảng tin cậy $1 - \alpha$ cho tỷ lệ tổng thể

1. Biến ngẫu nhiên nhị thức

2. Ước lượng khoảng cho tỉ lệ tổng thể
   - Ước lượng khoảng tin cậy 95%

3. Ước lượng điểm cho tỉ lệ tổng thể
   - Ước lượng khoảng tin cậy $1 - \alpha$ cho tỷ lệ tổng thể

Bài toán

Giả sử ta cần ước lượng tỉ lệ P các đối tượng có dấu hiệu T trong tổng thể. Ta cần thiết lập một cách ước lượng đạt độ chính xác cho trước bằng cách nào?

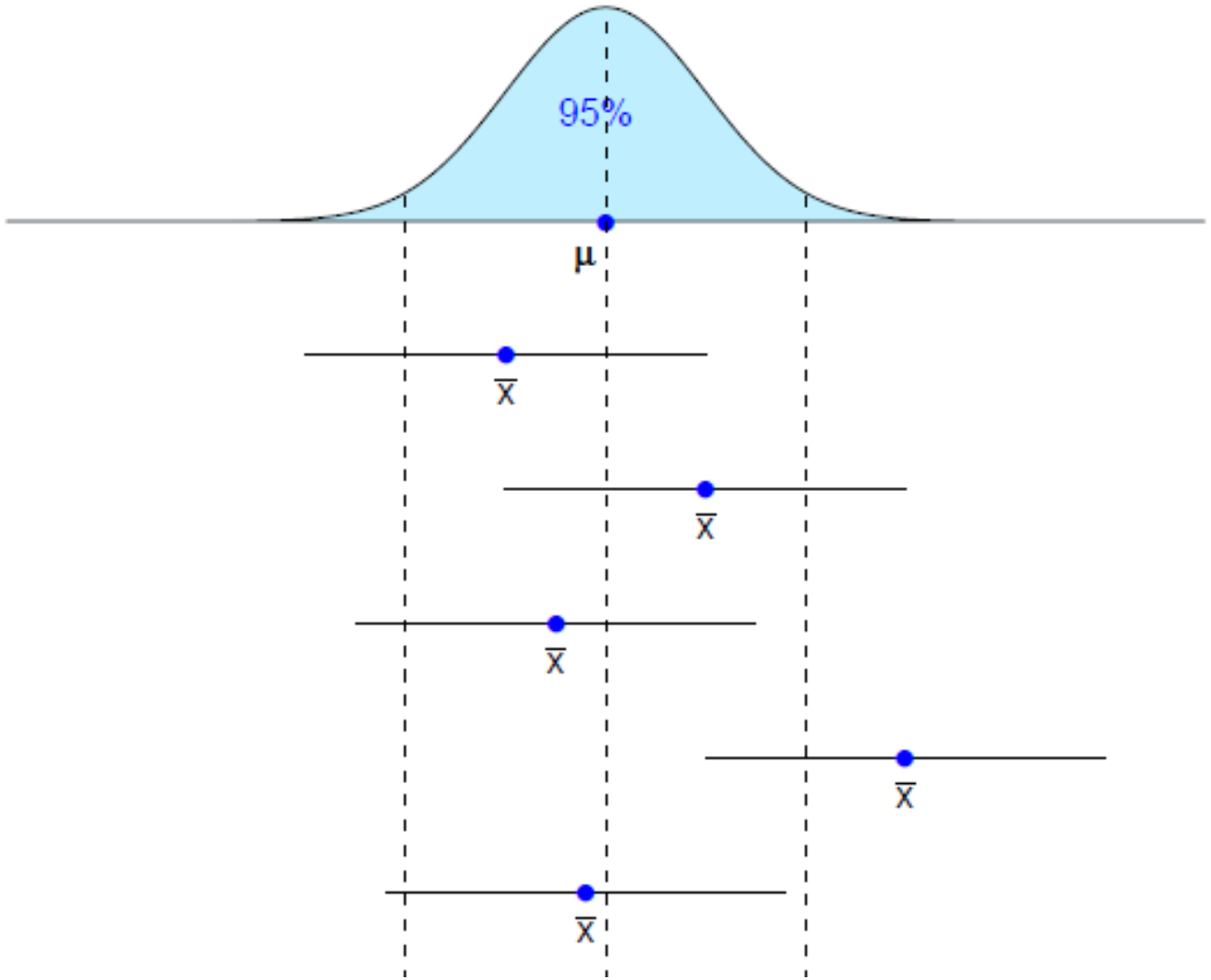
Bài toán

Giả sử ta cần ước lượng tỉ lệ P các đối tượng có dấu hiệu T trong tổng thể. Ta cần thiết lập một cách ước lượng đạt độ chính xác cho trước bằng cách nào?

Khái niệm

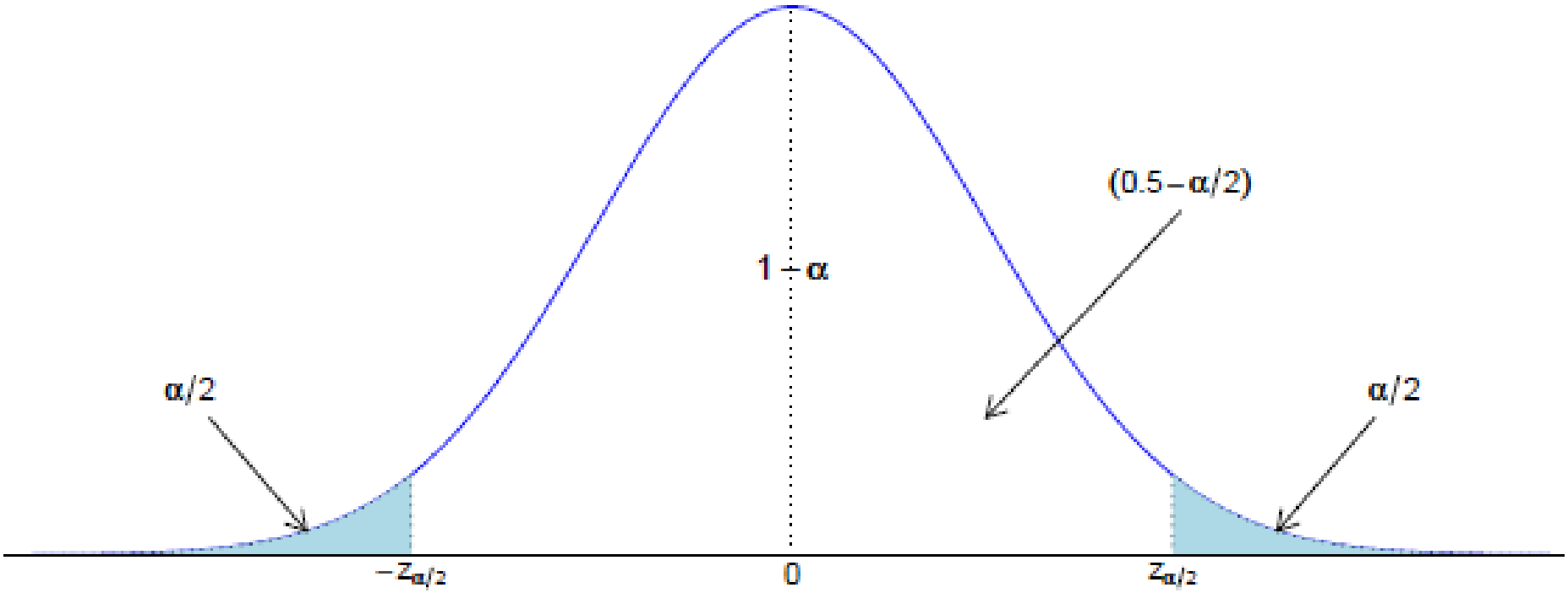
Cách ước lượng khoảng với độ tin cậy 95% cho P là một quy tắc mà:

- *từ mỗi mẫu điều tra cho ta một khoảng ước lượng*
- *cứ làm 100 lần như thế thì được 100 khoảng nhưng có tới 95 khoảng chứa P, còn lại thì không chứa P.*



Hình: Một số khoảng ước lượng 95%

Ta sẽ kí hiệu z_α là điểm trên trục hoành sao cho $P(Z > z_\alpha) = \alpha$, nó là điểm sao cho phần diện tích bôi màu của hình trên bằng α.



Theorem

Trong R,

$$z_k = qnorm(k, mean = 0, sd = 1, lower.tail = FALSE)$$

Do đó: $z_{\alpha/2} = qnorm(\alpha/2, mean = 0, sd = 1, lower.tail = FALSE)$

Định lí

Từ quy luật phân phối xác suất của tỉ lệ mẫu $\hat{p}$ ta có một cách để lập ra khoảng ước lượng với độ tin cậy 95% như sau:

1. *Chọn mẫu cỡ n đủ lớn, kiểm tra điều kiện* $5 \leq np \leq n - 5$

Định lí

Từ quy luật phân phối xác suất của tỉ lệ mẫu $\hat{p}$ ta có một cách để lập ra khoảng ước lượng với độ tin cậy 95% như sau:

1. *Chọn mẫu cỡ n đủ lớn, kiểm tra điều kiện* $5 \leq np \leq n - 5$
2. *Tìm hai điểm trên phân phối z để giới hạn khoảng xác xuất 95%. Hai điểm này đối xứng nhau, kí hiệu điểm dương là* $z_{0.025}$

Định lí

Từ quy luật phân phối xác suất của tỉ lệ mẫu $\hat{p}$ ta có một cách để lập ra khoảng ước lượng với độ tin cậy 95% như sau:

1. *Chọn mẫu cỡ n đủ lớn, kiểm tra điều kiện $5 \leq np \leq n - 5$*
2. *Tìm hai điểm trên phân phối z để giới hạn khoảng xác xuất 95%. Hai điểm này đối xứng nhau, kí hiệu điểm dương là $z_{0.025}$*
3. *Đặt điều kiện: $-z_{0.025} < \dfrac{\hat{p} - p}{\sqrt{\dfrac{p(1-p)}{n}}} < z_{0.025}$. Điều này tương đương*

   với:

$$-z_{0.025}\sqrt{\frac{p(1-p)}{n}} + \hat{p} < p < \hat{p} + z_{0.025}\sqrt{\frac{p(1-p)}{n}}$$

Định lí

Từ quy luật phân phối xác suất của tỉ lệ mẫu $\hat{p}$ ta có một cách để lập ra khoảng ước lượng với độ tin cậy 95% như sau:

1. *Chọn mẫu cỡ n đủ lớn, kiểm tra điều kiện $5 \leq np \leq n-5$*
2. *Tìm hai điểm trên phân phối z để giới hạn khoảng xác xuất 95%. Hai điểm này đối xứng nhau, kí hiệu điểm dương là $z_{0.025}$*
3. *Đặt điều kiện: $-z_{0.025} < \dfrac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} < z_{0.025}$. Điều này tương đương*

   với:

$$-z_{0.025}\sqrt{\frac{p(1-p)}{n}} + \hat{p} < p < \hat{p} + z_{0.025}\sqrt{\frac{p(1-p)}{n}}$$

? *Trong việc sử dụng công thức trên vào ước lượng cho P, có thành phần nào là ta chưa biết?*

Định lí

Từ quy luật phân phối xác suất của tỉ lệ mẫu $\hat{p}$ ta có một cách để lập ra khoảng ước lượng với độ tin cậy 95% như sau:

1. *Chọn mẫu cỡ n đủ lớn, kiểm tra điều kiện $5 \leq np \leq n-5$*
2. *Tìm hai điểm trên phân phối z để giới hạn khoảng xác xuất 95%. Hai điểm này đối xứng nhau, kí hiệu điểm dương là $z_{0.025}$*
3. *Đặt điều kiện: $-z_{0.025} < \dfrac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} < z_{0.025}$. Điều này tương đương*

   với:

$$-z_{0.025}\sqrt{\frac{p(1-p)}{n}} + \hat{p} < p < \hat{p} + z_{0.025}\sqrt{\frac{p(1-p)}{n}}$$

? *Trong việc sử dụng công thức trên vào ước lượng cho P, có thành phần nào là ta chưa biết? Ta ước lượng nó bằng cách nào?*

1. Biến ngẫu nhiên nhị thức

2. Ước lượng khoảng cho tỉ lệ tổng thể
   - Ước lượng khoảng tin cậy 95%

3. Ước lượng điểm cho tỉ lệ tổng thể
   - Ước lượng khoảng tin cậy $1 - \alpha$ cho tỷ lệ tổng thể

Definition

Hàm ước lượng điểm cho P là hàm số mà với mỗi một mẫu được chọn, hàm ước lượng đó cho ta kết là một con số $\hat{p}$ ước lượng cho P.

Definition

Hàm ước lượng điểm cho P là hàm số mà với mỗi một mẫu được chọn, hàm ước lượng đó cho ta kết là một con số $\hat{p}$ ước lượng cho P.

Definition

Với ước lượng điểm cho tỉ lệ P, người ta thường cố gắng tìm hàm ước lượng f đạt được tiêu chuẩn: không chệch, hiệu quả, vững, đầy đủ. Ở đây ta tập trung vào 2 tiêu chuẩn:

1. f được gọi là không chệch nếu $E(f) = P$.
2. f được gọi là hiệu quả nếu $V(f)$ là nhỏ nhất trong tất cả các f thỏa mãn $E(f) = P$.

Example

Giả sử ta có một tổng thể bạn gồm 4 người có lương như sau:

$$5 \qquad 7 \qquad 10 \qquad 8$$

Giả sử ta muốn ước lượng tỉ lệ P những người có lương trên 6 của tổng thể trên nhưng ta không biết lương của cả 4 người, ta cần ước lượng nó qua việc chọn mẫu. Giả sử ta chọn mẫu ngẫu nhiên 3 người.

1. Liệt kê tất cả các trường hợp có thể xảy ra của mẫu.
2. Tính tỉ lệ mẫu $f = \frac{k}{n}$ trong đó, k là số người có lương trong n người được chọn ra.
3. Tính trung bình của tất cả các f nói trên so sánh với P.
4. Vậy f có đạt được tiêu chuẩn không chệch không?

Theorem

Với $\hat{p} = \frac{k}{n}$. *Ta có:*

- $E(\hat{p}) = P$.
- $V(\hat{p}) = \frac{P(1-P)}{n} \leq V(f)\ \forall f:\ E(f) = P$

Có nghĩa là: $\hat{p}$ *là ước lượng không chệch và vững cho* P.

$\Longrightarrow$ Trong công thức ước lượng khoảng của p, ta thay phần p chưa biết ở hai đầu mút bởi $\hat{p}$.

Định lí

Từ quy luật phân phối xác suất của tỉ lệ mẫu $\hat{p}$ ta có một cách để lập ra khoảng ước lượng với độ tin cậy 95% như sau:

1. *Chọn mẫu cỡ n đủ lớn, tính $\hat{p}$, kiểm tra điều kiện $5 \leq n\hat{p} \leq n - 5$*

Định lí

Từ quy luật phân phối xác suất của tỉ lệ mẫu $\hat{p}$ ta có một cách để lập ra khoảng ước lượng với độ tin cậy 95% như sau:

1. *Chọn mẫu cỡ n đủ lớn, tính $\hat{p}$, kiểm tra điều kiện $5 \leq n\hat{p} \leq n-5$*
2. *Khi đó khẳng định sau đúng với xác suất 95%*

$$-z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + \hat{p} < p < \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Định lí

Từ quy luật phân phối xác suất của tỉ lệ mẫu $\hat{p}$ ta có một cách để lập ra khoảng ước lượng với độ tin cậy 95% như sau:

1. *Chọn mẫu cỡ n đủ lớn, tính $\hat{p}$, kiểm tra điều kiện $5 \leq n\hat{p} \leq n-5$*
2. *Khi đó khẳng định sau đúng với xác suất 95%*

$$-z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} + \hat{p} < p < \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

trong đó $z_{0.025} = 1.96$.

Example

Một cuộc điều tra về tỉ lệ các công ty sẵn lòng cho sinh viên đến thực tập tại một thành phố cho thấy có 100 trong số 150 công ty đồng ý. Hãy ước lượng khoảng tin cậy 95% cho tỉ lệ các công ty trong thành phố trên sẵn sàng nhận sinh viên đến thực tập.

Solution

Gọi P là tỉ lệ các công ty sẵn lòng cho sinh viên đến thực tập.

1. *Ta có* $\hat{P} = 100/150 = 2/3$.
2. *Ta có* $5 \leq 100 \leq 150$ *nên khoảng ước lượng 95% cho P là:*

$$[-1.96\sqrt{\frac{2/3(1-2/3)}{150}} + \frac{2}{3}; 1.96\sqrt{\frac{2/3(1-2/3)}{150}} + \frac{2}{3}]$$

hay

...

1. Biến ngẫu nhiên nhị thức

2. Ước lượng khoảng cho tỉ lệ tổng thể
   - Ước lượng khoảng tin cậy 95%

3. Ước lượng điểm cho tỉ lệ tổng thể
   - Ước lượng khoảng tin cậy $1 - \alpha$ cho tỷ lệ tổng thể

Định nghĩa

Một cách ước lượng với độ tin cậy $1-\alpha$ *cho tỉ lệ tổng thể là quy tắc mà theo cách làm đó, mỗi mẫu cho ta một khoảng ước lượng, và cứ 100 lần làm tương tự nhau thì kết luận của ta đúng* $100(1-\alpha)$ *lần.*

Theorem

Gọi p là tỉ lệ số phần tử của một tổng thể có một đặc tính nào đó cần được ước lượng. Theo định lí giới hạn trung tâm ta có khi cỡ mẫu n lớn và $np \geq 5, n(1-p) \geq 5$, *khoảng tin cậy* $100(1-\alpha)\%$ *cho tỉ lệ tổng thể p là:*

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad (1)$$

trong đó $\hat{p}$ *là tỷ lệ mẫu, n là cỡ mẫu.*

Example

Để ước lượng số cá cùng loài trong một hồ nuôi, người ta bắt 500 con cá và đánh dấu, sau đó thả lại vào hồ. Một thời gian sau bắt ngẫu nhiên 500 con cá thấy có 60 con có đánh dấu. Cho $z_{0.05} \approx 1.64$. Hãy:

1. Hãy cho một ước lượng điểm và một khoảng ước lượng với độ tin cậy 90% về tỉ lệ cá được đánh dấu trong hồ.
2. Từ kết quả trên suy diễn ra số lượng cá trong hồ.

Example

Giả sử một nhóm sinh viên khảo sát về tình hình xin được việc làm của sinh viên TLU trong một năm đầu ra trường cho thấy có 324 sinh viên đã tìm được việc ngay trong năm đầu, trong khi đó 150 bạn khác cho biết trong một năm đầu họ không xin được việc. Hãy ước lượng khoảng tin cậy 95 % cho tỉ lệ sinh viên TLU ra trường kiếm được việc ngay trong một năm đầu tiên. Cho $z_{0.025} \approx 1.96$.