Homework week 2

Bùi Khánh Duy

2023-02-26

Homework

4.1 List all possible simple random samples of size n = 2 that can be selected from the population {0, 1, 2, 3, 4}. Calculate σ^2 for the population and $V(\bar{y})$ for the sample.

```
dataX <- c(0,1,2,3,4)
N <- length(dataX)
n <- 2
x <- t(combn(dataX, n))
```

```
##       [,1] [,2]
##  [1,]    0    1
##  [2,]    0    2
##  [3,]    0    3
##  [4,]    0    4
##  [5,]    1    2
##  [6,]    1    3
##  [7,]    1    4
##  [8,]    2    3
##  [9,]    2    4
## [10,]    3    4
```

```
Sample <- as.vector.data.frame(apply(x, MARGIN = 1, function(x) {
  paste(sprintf("{%d,%d}",x[1], x[2]))
}))
y_bar <- as.vector.data.frame(rowMeans(x))

data2 <- data.frame(Sample, y_bar)
```

Sample	y_bar
{0,1}	0.5
{0,2}	1.0
{0,3}	1.5
{0,4}	2.0
{1,2}	1.5
{1,3}	2.0

and

$$\sigma^2 = V(y) = E(y-\mu)^2 \text{with } \mu = E(\bar{y})$$

```
y_bar.mean <- mean(y_bar)
y_bar.mean
```

```
## [1] 2
```

```
sigma = mean((x - y_bar.mean)^2)
print(paste("sigma^2 = ",sigma))
```

```
## [1] "sigma^2 =  2"
```

$$V(\bar{y}) = E(\bar{y} - \mu)^2$$

```
v_y_bar <- mean((y_bar - y_bar.mean)^2)
v_y_bar
```

```
## [1] 0.75
```

```
sigma/n*(N-n)/(N-1)
```

```
## [1] 0.75
```

$$\Rightarrow V(\bar{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

4.2 For the simple random samples generated in Exercise 4.1, calculate s^2 for each sample. Show numerically that

$$E(s^2) = \frac{N}{N-1}\sigma^2$$

```
s_2 <- as.vector.data.frame(apply(x, MARGIN=1, function(x) {
  var(x)
}))
data3 <- data.frame(data2, s_2)
```

Sample	y_bar	s_2
{0,1}	0.5	0.5
{0,2}	1.0	2.0
{0,3}	1.5	4.5
{0,4}	2.0	8.0
{1,2}	1.5	0.5
{1,3}	2.0	2.0
{1,4}	2.5	4.5
{2,3}	2.5	0.5
{2,4}	3.0	2.0
{3,4}	3.5	0.5

```
s_2.mean <- mean(s_2)
s_2.mean
```

```
## [1] 2.5
```

```
N/(N-1)*sigma
```

```
## [1] 2.5
```

$$\Rightarrow E(s^2) = \frac{N}{N-1}\sigma^2$$

That all!

4.3 Suppose you want to estimate the number of weed clusters of a certain type in a field. What is the population, and what would you use for sampling units? How would you construct a frame? How would you select a simple random sample? If a sampling unit is an area such as a square yard, does the size chosen for a sampling unit affect the accuracy of the results? What considerations go into our choice of size of sampling unit?

- Tổng thể (population): Tất cả các cụm cỏ trên cánh đồng
- Đơn vị mẫu: Sử dụng đơn vị diện tích: mẫu, sào, mét vuông, v.v
- Khung mẫu: Tiến hành đo đạc toàn bộ cánh đồng để biết diện tích của tổng thể là bao nhiêu, từ đấy tiến hành chia nhỏ và đánh số để chọn mẫu
- Chọn mẫu ngẫu nhiên đơn giản: Tiến hành chọn ngẫu nhiên theo số thứ tự được đánh.
- Nếu kích thước chọn mẫu là 1 thước vuông (square yard) => đây là đơn vị phổ biến thường được dùng để đo đạc ruộng đất, và nó cũng đủ lớn để tiến hành nghiên cứu. Việc đơn vị mẫu lớn hơn có thể hiệu quả hơn trong khảo sát, nhưng khi đơn vị mẫu nhỏ hơn lại có thể cung cấp kết quả chính xác hơn. Cần phải cân nhắc giữa hiệu quả và độ chính xác, và kích thước đơn vị mẫu nên được chọn cẩn thận. Ở bài toán ruộng đất thì có thể ưu tiên độ hiệu quả hơn.
- Các yếu tố cần được xem xét khi lựa chọn kích thước của đơn vị mẫu:
    1. Tính đồng nhất: Nếu tổng thể là đồng nhất, có thể sử dụng các đơn vị mẫu nhỏ vì đã có thể bao quát được sự biến động của đơn vị quan sát.
    2. Tính đa dạng: Nếu tổng thể đa dạng, có thể cần sử dụng các đơn vị mẫu lớn hơn.
    3. Độ chính xác mẫu: Các đơn vị mẫu nhỏ hơn có thể cung cấp độ chính xác cao hơn, nhưng có thể đòi hỏi nhiều tài nguyên và thời gian để thực hiện chọn mẫu và phân tích.
    4. Chi phí: Các đơn vị mẫu lớn hơn có thể hiệu quả về chi phí hơn, vì cần ít công sức để lấy mẫu và phân tích hơn, nhưng có thể dẫn đến độ chính xác không cao.
    5. Phương pháp lấy mẫu: Lựa chọn kích thước đơn vị mẫu cũng có thể bị ảnh hưởng bởi phương pháp lấy mẫu. Ví dụ, lấy mẫu phân tầng có thể đòi hỏi đơn vị mẫu nhỏ hơn so với phương pháp lấy mẫu ngẫu nhiên đơn giản.

4.4 In which of the following situations can you reasonably generalize from the sample to the population?

a. You use your statistics class to get an estimate of the percentage of students in your school who study at least two hours a night. Có thể, bởi vì việc chọn mẫu là khả thi, có thể thực hiện và có đủ lớn để kết luận cho tổng thể.

b. You use the average annual income of the ambassadors to the United Nations to get an estimate of average per-capita income for the world as a whole. Không thể. Vì mẫu chọn được chỉ là 1 quốc gia Mỹ trong khi trên thế giới có gần 300 quốc gia.

c. In 1996, a Gallup poll sampled 235 U.S. residents ages 18 to 29, to estimate the per- centage of all U.S. residents ages 18 to 29 who favored cuts in social spending. Không biết, vì không đủ dữ kiện về cách chọn mẫu. Nhưng nếu dựa trên tỉ lệ cho tổng thể là tất cả cư dẫn Mỹ từ 18 đến 29 tuổi thì là không đủ.

4.5 Describe the type of sample selection bias that would result from each of these sampling methods.

a. A student wants to determine the average size of farms in a county in Iowa. He drops some rice randomly on a map of the county and uses the farms hit by grains of rice as the sample. Sai lệch chọn mẫu do khung chưa đủ. Vì cách chọn này không theo một hệ thống lấy mẫu hay khung mẫu cụ thể nào nên rất dễ dẫn đến việc bỏ sót mẫu.

b. In a study about whether valedictorians "succeed big in life," a professor "traveled across Illinois, attending high school graduations and selecting 81 students to participate. . . . He picked students from the most diverse communities possible, from little rural schools to rich

suburban schools near Chicago to city schools." Source: Michael Ryan, "Do Valedictorians Succeed Big in Life?" Parade Magazine, May 17, 1998, pages 14–15. Sai lệch chọn mẫu do mẫu không ngẫu nhiên. Giáo sư đã lựa chọn học sinh từ các đa dạng các nhóm, điều này có thể đưa ra sai lệch vào nghiên cứu. Ví dụ, học sinh từ một số nhóm có thể có cơ hội giáo dục hoặc hoàn cảnh gia đình khác nhau, ảnh hưởng đến việc thành công của họ trong tương lai.

c. To estimate the percentage of students who passed the first Advanced Placement Statistics exam, a teacher on an Internet discussion list for teachers of AP Statistics asked teachers on the list to report to him how many of their students took the test and how many passed. Sai lệch chọn mẫu do tự chọn. Việc các giáo viên có thể chọn trả lời hoặc không / trả lời đúng hoặc sai làm ảnh hưởng đến kết quả chọn mẫu.

d. To find the average length of string in a bag, a student reaches in, mixes up the strings, selects one, mixes them up again, selects another, and so on. Sai lệch chọn mẫu do lấy mẫu theo hệ thống. Mỗi lần lấy 1 sợi dây ra khỏi túi, xác suất của các sợi dây đã bị thay đổi, dẫn đến kết quả sai trong việc tính mẫu.

e. In 1984, Ann Landers conducted a poll on the marital happiness of women by asking women to write to her. Sai lệch chọn mẫu do phản hồi tự nguyện. Chỉ những phụ nữ quan tâm đến việc tham gia cuộc khảo sát mới viết thư cho Ann Landers, điều này có thể dẫn đến sự đại diện quá nhiều của phụ nữ có chung quan điểm về hạnh phúc hôn nhân.