LAB 4: ƯỚC LƯỢNG

Nội dung:

1. Mô phỏng dữ liệu
2. Ước lượng điểm
3. Ước lượng khoảng
4. Bài tập

1. Mô phỏng dữ liệu:

- Dữ liệu thực tế không phải khi nào cũng có sẵn, để có dữ liệu để thực hiện thống kê đòi hỏi nhiều thời gian và công sức trong việc thu nhập và tiền xử lý dữ liệu. Để thuận lợi cho việc học tập và nghiên cứu, ta có thể tạo một bộ dữ liệu mô phỏng theo mong muốn của mình. Hiện nay, có rất nhiều công cụ cho phép ta mô phỏng dữ liệu, module <np.random> là một trong những công cụ hữu ích.

- Một số chức năng mà module <np.random> có thể cung cấp:

- Tạo một số ngẫu nhiên trong khoảng (0, 1):

random()

- Tạo hạt giống ngẫu nhiên: nhằm mục đích có thể tạo bộ dữ liệu mô phỏng giống lần trước. Giả sử, bạn tạo một bộ dữ liệu để xử lý bằng cách sử dụng hàm random(). Sau đó một người khác cũng lặp lại cách làm của bạn, tuy nhiên khi sử dụng hàm random() thì được bộ dữ liệu khác với của bạn dẫn đến kết quả xử lý có thể khác nhau nên không thể so sánh được. Để giải quyết vấn đề này, bạn có thể tạo một hạt giống ngẫu nhiên là một số nguyên bất kỳ trước khi thực hiện mô phỏng dữ liệu, trường hợp tái mô phỏng lại bộ dữ liệu cũ, chỉ cần phát sinh đúng hạt giống ban đầu.

seed()

- **Ví dụ:** mô phỏng tung đồng xu 4 lần

```python
import numpy as np

#tạo hạt giống ngẫu nhiên là 10
np.random.seed(10)

#in hai số ngẫu nhiên trong khoảng (0,1)
print(np.random.rand())
print(np.random.rand())

#phát sinh 4 số ngẫu nhiên trong khoảng (0,1)
a=np.random.random(size=4)
print(a)

#tạo mô phỏng 4 lần tung đồng xu với (True: Sấp; False: Ngửa)
coin_sample=a<0.5
print(coin_sample)
```

✓ 0.0s

```
0.771320643266746
0.0207519493594015
[0.63364823 0.74880388 0.49850701 0.22479665]
[False False  True  True]
```

2. Ước lượng điểm:

- ***Ví dụ 2.1:*** Sử dụng ước lượng điểm để ước lượng tham số của quần thể

```
import numpy as np
#Khởi tạo một quần thể cho trước thể hiện chiều cao (cm) của 5 người
small_pop=np.array([156, 152, 160, 165, 170])
print(small_pop)
mean_small_pop=np.mean(small_pop)
print('Chiều cao trung bình của Quần thể: {}'.format(mean_small_pop))
#lấy ngẫu nhiên một mẫu có kích thước là 4 và tính chiều cao trung bình
# và so sánh với giá trị của quần thể
np.random.seed(24)
sample1=np.random.choice(small_pop, size=4, replace=False)
sample1_mean=np.mean(sample1)
print('Mẫu ngẫu nhiên 1:', sample1)
print('Chiều cao trung bình của mẫu 1: {}'.format(sample1_mean))
print('Sai số ước lượng: {}'.format(abs(mean_small_pop-sample1_mean)))
```

✓ 0.0s

```
[156 152 160 165 170]
Chiều cao trung bình của Quần thể: 160.6
Mẫu ngẫu nhiên 1: [170 152 156 165]
Chiều cao trung bình của mẫu 1: 160.75
Sai số ước lượng: 0.15000000000000568
```

Nhận xét: Sai số ước lượng là: 0.15 cm. Ta có thể chấp nhận được với bài toán đo chiều cao.

- ***Ví dụ 2.2:*** Để cho việc đo sai số khách quan, ta thử lặp lại việc lấy mẫu trên 10 lần, và tính sai số ước lượng.

```
mean_array=np.empty(10)
for i in range(10):
    random_sample=np.random.choice(small_pop, size=4, replace=True)
    random_sample_mean=np.mean(random_sample)
    mean_array[i]=random_sample_mean
print('Chiều cao trung bình của 10 mẫu thu được:', mean_array)
print('Sai số ước lượng: {}'.format(abs(np.mean(mean_array)-mean_small_pop)))
```

✓ 0.0s

```
Chiều cao trung bình của 10 mẫu thu được: [158.25 161.5  160.5  166.25 158.25 158.25 158.25 158.   162.75 166.25]
Sai số ước lượng: 0.22499999999999432
```

Nhận xét: Ta nhận thấy, khi thực hiện việc lấy mẫu nhiều lần, sai số trung bình có nhỏ hơn so với ví dụ trên.

- ***Ví dụ 2.3:*** Minh hoạ ảnh hưởng của cỡ mẫu đến độ chính xác của ước lượng

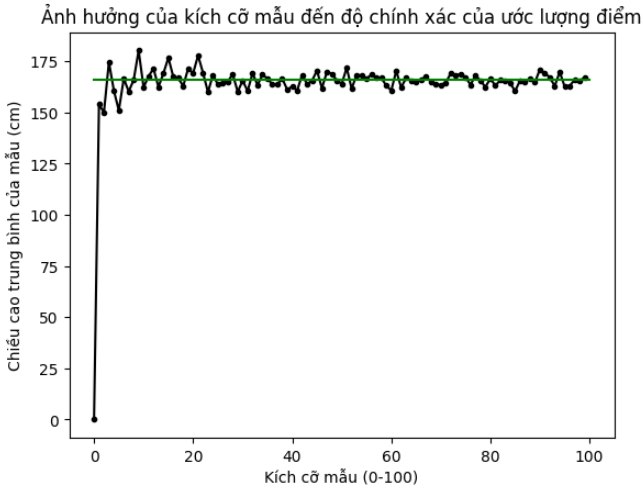
+ Để rõ ràng ta sẽ tạo một quần thể mới gồm 100 cá thể: MEDUIUM_POP

+ Lần lượt lấy mẫu với kích cỡ khác nhau sample_size = 1, 2, 3, ... và tính trung bình mẫu: mean_array

+ Trực quan bằng đồ thị

```
import matplotlib.pyplot as plt
MEDIUM_POP=np.random.randint(130,200,size=100)
mean_of_MEDIUM_POP=np.mean(MEDIUM_POP)
mean_array=np.empty(100)
mean_array[0]=0

for sample_size in range(1,100):
    temp=np.random.choice(MEDIUM_POP, size=sample_size)
    mean_array[sample_size]=np.mean(temp)

x=np.arange(100)
_=plt.plot(x, mean_array, marker='.', color='black')

_=plt.xlabel('Kích cỡ mẫu (0-100)')
_=plt.ylabel('Chiều cao trung bình của mẫu (cm)')
_=plt.title('Ảnh hưởng của kích cỡ mẫu đến độ chính xác của ước lượng điểm')

xx=np.array([0,100])
yy=np.empty(2)
yy[0]=yy[1]=mean_of_MEDIUM_POP

_=plt.plot(xx,yy,color='green')
print(mean_array)
plt.show()
```

```
[  0.         154.         150.         174.66666667 160.75
 151.         166.5        160.28571429 165.875      180.22222222
 162.3        167.63636364 171.5        162.46153846 169.07142857
 176.73333333 167.4375     166.82352941 162.88888889 171.52631579
 169.15       177.85714286 169.31818182 159.95652174 168.33333333
 163.64       164.53846154 165.18518519 168.67857143 160.10344828
 165.56666667 160.48387097 169.15625    163.42424242 168.44117647
 166.6        163.75       163.78378378 166.57894737 161.1025641
 162.75       160.73170732 168.35714286 163.88372093 165.5
 170.48888889 161.84782609 169.80851064 168.9375     165.67346939
 163.98       171.88235294 161.65384615 168.28301887 168.12962963
 166.56363636 168.48214286 167.07017544 166.89655172 163.06779661
 160.73333333 170.32786885 162.16129032 166.95238095 165.65625
 164.92307692 166.1969697  167.43283582 165.02941176 163.5942029
 163.57142857 164.53521127 169.13888889 168.05479452 168.82432432
 167.32       163.47368421 168.18181818 165.51282051 162.46835443
 166.5875     163.12345679 165.73170732 165.55421687 164.29761905
 160.47058824 165.65116279 164.75862069 166.67045455 164.70786517
 171.07777778 169.25274725 167.13043478 162.88172043 169.84042553
 162.63157895 162.97916667 165.86597938 165.26530612 167.13131313]
```



- *Nhận xét:* Qua ví dụ trên ta có thể nhận thấy kích thước mẫu có liên quan đến độ chính xác của ước lượng điểm

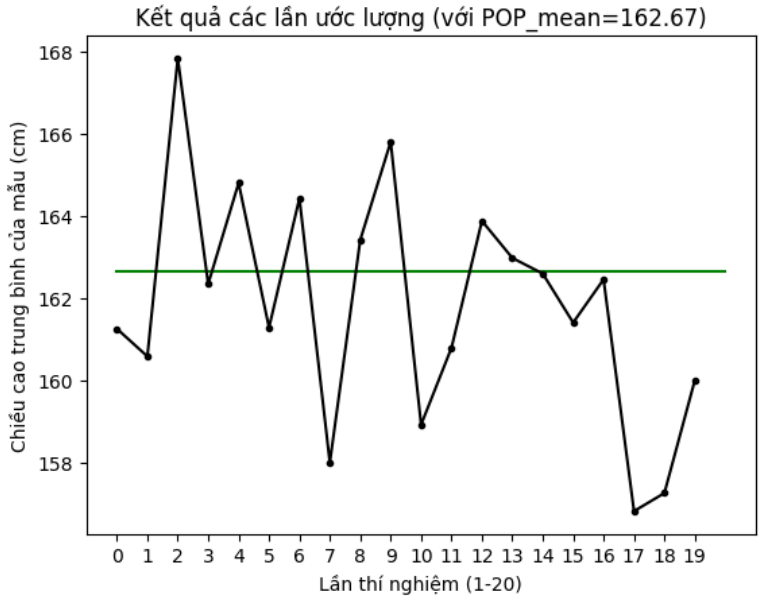
- *Kết luận:* Để tăng độ chính xác của ước lượng ta có thể tăng kích thước của quần thể.

3. Ước lượng khoảng:

- ***Ví dụ 3.1:*** Để tăng độ chính xác ta sử dụng ước lượng khoảng thay thế cho ước lượng điểm.

+Trước tiên, ta thực hiện ước lượng điểm 20 lần với cỡ mẫu mỗi lần là 40 để xem xét kết quả của mỗi lần ước lượng

```
MEDIUM_POP=np.random.randint(130,200,size=100)
mean_of_MEDIUM_POP=np.mean(MEDIUM_POP)
np.random.seed(24)
estimate_times=20
mean_array=np.empty(estimate_times)
for i in range(estimate_times):
    temp=np.random.choice(MEDIUM_POP, size=100)
    mean_array[i]=np.mean(temp)
#vẽ giá trị thực tế (mean của quần thể)
_=plt.plot(np.asarray([0,estimate_times]), np.asarray([mean_of_MEDIUM_POP,mean_of_MEDIUM_POP]), color='green')

#Vẽ các kết quả ước lượng
x=np.arange(20)
_=plt.plot(x, mean_array, marker='.', color='black')
_=plt.xticks(np.arange(0,estimate_times, step=1))
_=plt.xlabel('Lần thí nghiệm (1-20)')
_=plt.ylabel('Chiều cao trung bình của mẫu (cm)')
_=plt.title('Kết quả các lần ước lượng (với POP_mean={})'.format(mean_of_MEDIUM_POP))
plt.show()
```



- ***Nhận xét:*** Kết quả ước lượng nằm dao động xung quanh giá trị thực tế, có kết quả gần giá trị thực tế, nhưng cũng có kết quả rất xa.

+ Hạn chế của ước lượng điểm: kết quả mỗi lần khác nhau, và có kết quả có sai số rất lớn

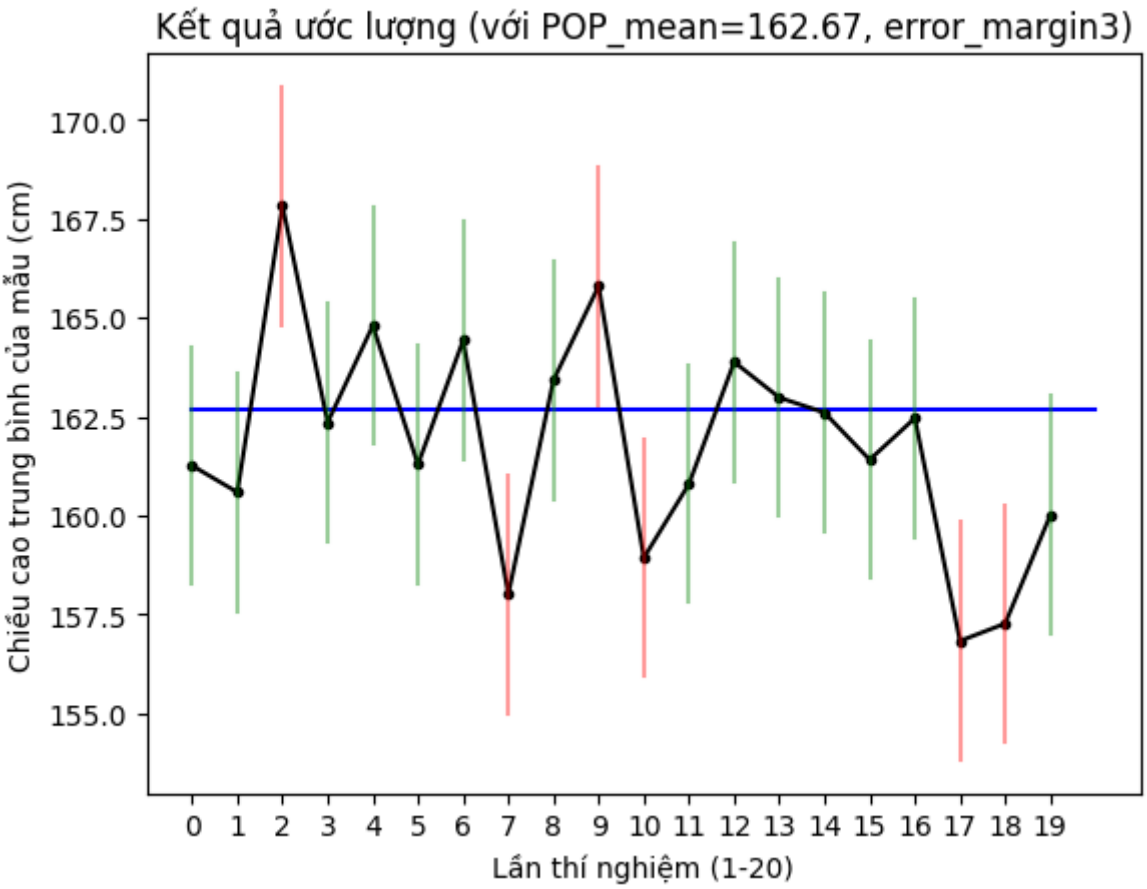
+ Để tăng độ chính xác của ước lượng thay vì dùng điểm ước lượng ta dùng một khoảng ước lượng.

- ***Ví dụ 3.2:*** Ta vẽ lại biểu đồ trên nhưng thay vì dùng điểm ước lượng ta dùng khoảng ước lượng.

+ Giả sử ta cho phép biên độ lỗi là error_margin = 3cm, như vậy khoảng ước lượng sẽ là [point_estimate - 3, point_estimate + 3]

+ Sau đó ta tỷ lệ chính xác của ước lượng bằng cách tìm tỷ lệ phần trăm kết quả ước lượng đúng

```
import matplotlib.pyplot as plt
#Vẽ giá trị thực tế (mean của quần thể)
_=plt.plot(np.asarray([0,estimate_times]), np.asarray([mean_of_MEDIUM_POP,mean_of_MEDIUM_POP]), color='blue')
#Vẽ các kết quả ước lượng
x=np.arange(20)
_=plt.plot(x, mean_array, marker='.', color='black')
#Vẽ khoảng ước lượng với biên độ lỗi là 3cm
error_margin=3
true_result=0
for i in range(estimate_times):
    xx=np.array([i,i])
    lower=mean_array[i]-error_margin
    upper=mean_array[i]+error_margin
    yy=np.array([lower,upper])
    if (mean_of_MEDIUM_POP<=upper and mean_of_MEDIUM_POP>=lower):
        _=plt.plot(xx,yy, color='green', alpha=0.4)
        true_result+=1
    else:
        _=plt.plot(xx,yy, color='red', alpha=0.4)

_=plt.xticks(np.arange(0,estimate_times, step=1))
_=plt.xlabel('Lần thí nghiệm (1-20)')
_=plt.ylabel('Chiều cao trung bình của mẫu (cm)')
_=plt.title('Kết quả ước lượng (với POP_mean={}, error_margin{})'.format(mean_of_MEDIUM_POP,error_margin))
plt.show()
print('Tỷ lệ chính xác của ước lượng:{}'.format(true_result/estimate_times))
```



Tỷ lệ chính xác của ước lượng: 0.65

Nhận xét:

+ Kết quả ước lượng khoảng vẫn có thể sai (không chứa giá trị thực tế)

+ Để tăng độ chính xác có thể tăng độ rộng của khoảng ước lượng (tăng error_margin) --> Xem như bài tập

+ Nếu khoảng ước lượng quá rộng thì ta khó tìm giá trị thực sự trong khoảng ấy. Vậy thì độ rộng của khoảng ước lượng bao nhiêu là đủ?

- ***Ví dụ 3.3:*** Như vậy, để thực hiện bài toán ước lượng, ta phải cân nhắc giữa 2 yếu tố là: độ rộng khoảng và độ chính xác của ước lượng

+ Giả sử ta chấp nhận giảm độ chính xác của ước lượng xuống 95% để đổi lấy một khoảng ước lượng bé hơn

+ Điều này có nghĩa là xác suất có được một khoảng ước lượng chứa giá trị thực tế là 95%

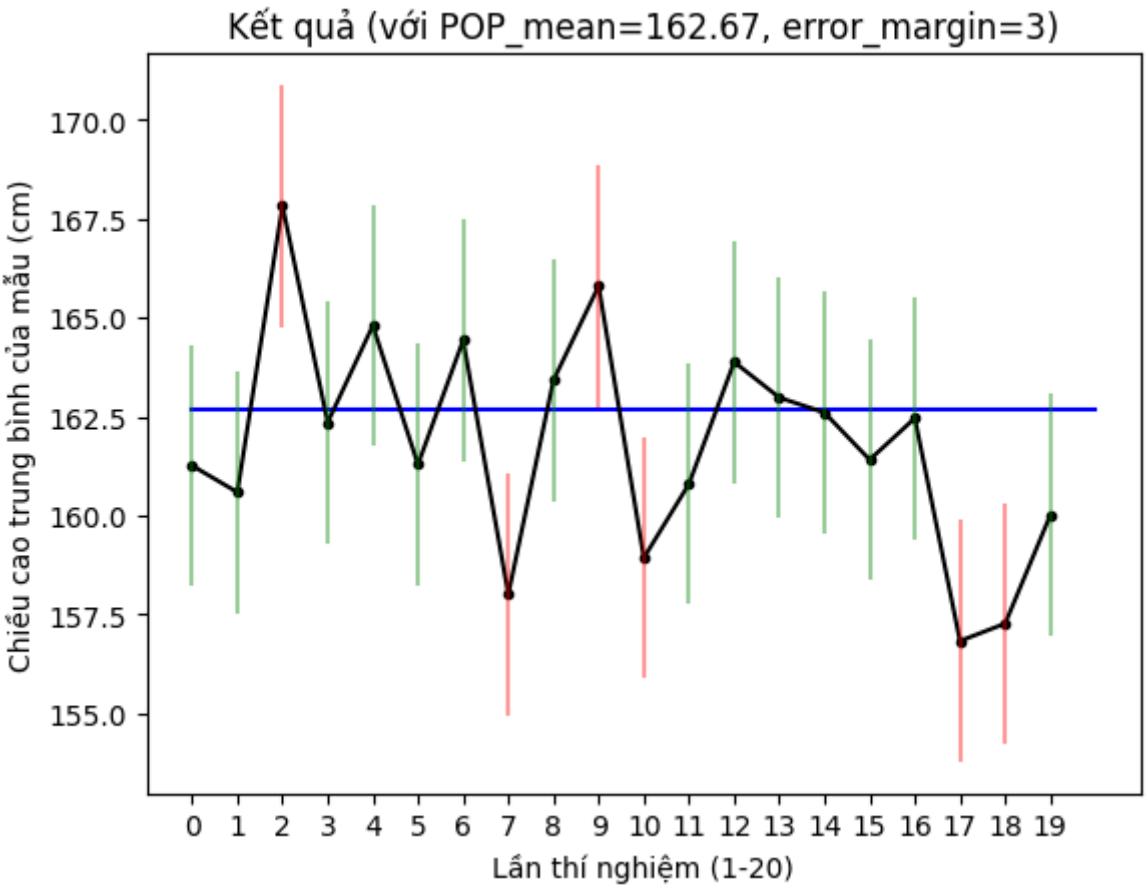
+ Để làm điều này ta sử dụng định lý Giới Hạn Trung Tâm

+ Với độ tin cậy của ước lượng là 95%, ta sẽ tìm biên độ lỗi dựa vào một chọn ngẫu nhiên một mẫu có cùng kích cỡ

```
from scipy import stats
#Tạo ngẫu nhiên một mẫu có kích thước 40 từ quần thể MEDIUM_POP
sample2=np.random.choice(MEDIUM_POP, size=40)
mean_sample2=np.mean(sample2)
#Chuẩn bị các thông số
n=sample2.size
degree_of_freedom=n-1
confidence_level=0.95
t_score=stats.t.ppf(confidence_level, degree_of_freedom)
standard_error=sample2.std()/np.sqrt(n)
print('Biên độ lỗi với độ tin cậy là 95%:{}'.format(error_margin))
```

- ***Ví dụ 3.4:*** Chạy lại Ví dụ 3.2 với biên độ lỗi mới, và kiểm tra tỷ lệ ước lượng chính xác

```
#Vẽ giá trị thực tế (mean của quần thể)
_=plt.plot(np.asarray([0,estimate_times]), np.asarray([mean_of_MEDIUM_POP,mean_of_MEDIUM_POP]), color='blue')
#Vẽ các kết quả ước lượng
x=np.arange(20)
_=plt.plot(x,mean_array,marker='.',color='black')
#Vẽ khoảng ước lượng với biên độ lỗi mới
true_result=0
for i in range(estimate_times):
    xx=np.asarray([i,i])
    lower=mean_array[i]-error_margin
    upper=mean_array[i]+error_margin
    yy=np.asarray([lower, upper])
    if (mean_of_MEDIUM_POP<=upper and mean_of_MEDIUM_POP>=lower):
        _=plt.plot(xx,yy,color='green',alpha=0.4)
        true_result=true_result+1
    else:
        _=plt.plot(xx,yy,color='red', alpha=0.4)
_=plt.xticks(np.arange(0,estimate_times,step=1))
_=plt.xlabel('Lần thí nghiệm (1-20)')
_=plt.ylabel('Chiều cao trung bình của mẫu (cm)')
_=plt.title('Kết quả (với POP_mean={}, error_margin={})'.format(mean_of_MEDIUM_POP, error_margin))
plt.show()
print('Tỷ lệ chính xác của ước lượng:{}'.format(true_result/estimate_times))
```



Nhận xét: Qua ví dụ trên ta thấy với biên độ lỗi được tính từ độ tin cậy mong muốn là 95%. Ta được tỷ lệ chính xác của ước lượng cũng là 95% (với số lần thực hiện là 20)

Bạn hãy thử tăng số lần thực hiện lên 100 hay 1000 lần xem tỷ lệ này còn chính xác không?Tỷ lệ chính xác của ước lượng:0.7.

4. Bài tập

Bài 1. Green M&M Candies liên quan đến Dataset 18 trong file excel. Tìm tỉ lệ mẫu của M&M có màu xanh lá. Sử dụng kết quả để xây dựng 1 ước lượng khoảng tin cậy 95% của % quần thể M&M có màu xanh lá. Có phải kết quả này có nhất quán với tỉ lệ 16% được báo cáo bởi nhà sản xuất kẹo. Tại sao nhất quán và tại sao không?

Bài 2. Freshman Weight Gain liên quan đến Dataset 3 trong file excel

a. Dựa vào kết quả của mẫu, tìm ước lượng điểm tốt nhất của tỉ lệ phần trăm các sinh viên cao đẳng tăng cân trong năm thứ 1.

b. Xây dựng ước lượng khoảng tin cậy 95% về tỉ lệ phần trăm các sinh viên cao đẳng tăng cân trong năm thứ 1.
c. Giả sử rằng bạn là nhà báo, viết phát biểu mô tả kết quả trên bao gồm các thông tin liên quan.

Bài 3. Lượng mưa ở Boston: liên quan đến Dataset 14 trong file excel, và quan tâm đến các ngày với các giá trị lượng mưa khác nhau từ 0 đến các ngày có mưa có giá trị ượng mưa lớn hơn 0. Xây dựng ước lượng khoảng tin cậy 95% cho tỉ lệ mưa trong các ngày Thứ Tư và xây dựng ước lượng khoảng tin cậy 95% cho tỉ lệ mưa trong các ngày Chủ Nhật. So sánh kết quả. Có phải lượng mưa xuất hiện ở các ngày này nhiều hơn so với các ngày khác hay không?

Bài 4. Bình chọn phim: liên quan đến Dataset 19 trong file excel. Tìm tỉ lệ phim với tỉ lệ bình chọn là R. Sử dụng tỉ lệ đó để xây dựng ước lượng khoảng tin cậy 95% cho tỉ lệ các phim với kết quả bình chọn là R. Giả sử rằng các phim trên đã liệt kê trong file được lấy mẫu theo phương pháp lấy mẫu ngẫu nhiên đơn giản, chúng ta có thể kết luận rằng hầu như các phim có tỉ lệ bình chọn khác R không? Tại sao có hoặc tại sao không?

Bài 5. Tổng số tiền phim: liên quan đến Dataset 9 trong file excel. Xây dựng ước tính khoảng thời gian tin cậy 95% của tổng số tiền trung bình cho quần thể của tất cả các phim. Giả định rằng độ lệch chuẩn của quần thể được biết là 100 triệu đô la.

Bài 6. Điểm đánh giá tín dụng FICO: liên quan đến Dataset 24 trong file excel. Xây dựng ước lượng khoảng tin cậy 99% của điểm FICO trung bình cho quần thể. Giả sử độ lệch chuẩn của quần thể là 92.2

Bài 7. Nicotine trong thuốc lá: Nicotine trong thuốc lá: liên quan đến Dataset 4 trong file excel. Giả định rằng các mẫu là các mẫu ngẫu nhiên đơn giản thu được từ các quần thể có phân phối chuẩn.

a. Xây dựng ước lượng khoảng tin cậy 95% lượng nicotin trung bình trong thuốc lá có kích thước vừa (cỡ king), không lọc, không menthol, và không ánh sáng.
b. Xây dựng ước lượng khoảng tin cậy 95% lượng nicotin trung bình trong thuốc lá có chiều dài 100 mm, được lọc, không menthol và không ánh sáng.

c. So sánh kết quả. Bộ lọc trên thuốc lá có vẻ hiệu quả không?

Bài 8. Nhịp tim: Một bác sĩ muốn phát triển các tiêu chí để xác định xem bệnh nhân có nhịp tim không bình thường, và cô ấy muốn xác định liệu có sự khác biệt đáng kể giữa nam và nữ. Sử dụng nhịp tim mẫu trong Dataset 1.

a. Xây dựng ước lượng khoảng tin cậy 95% của nhịp tim trung bình cho nam.
b. Xây dựng ước tính khoảng tin cậy 95% của nhịp tim trung bình cho nữ.
c. So sánh các kết quả trước đó. Chúng ta có thể kết luận rằng trung bình quần thể cho nam và nữ có khác nhau không? Tại sao có hay tại sao không?