Bài Kiểm tra giữa kỳ

Môn: Các phương pháp thống kê hiên đại trong xã hội học
Học viên: Vũ Minh Hưng

1. **Xác định phân phối của dữ liệu (likelihood):**
   - Gọi Y là đại lượng ngẫu nhiên biểu diễn số lượng trẻ em sinh ra bởi phụ nữ khảo sát trong bộ dữ liệu.
   - Gọi y_i là giá trị của Y_i ứng với số lượng trẻ em được sinh ra bởi người phụ thứ i trong bộ dữ liệu. $i = 1, 2, 3, \dots n$
   - Vì số lượng trẻ em sinh ra tuân theo phân phối Poisson với tham số θ, ký hiệu $dpois(y, \theta)$. Ta có công thức tổng quát của phân phối Poisson sẽ là:
     $$Pr(Y = y|\theta) = dpois(y, \theta) = \frac{\theta^y e^{-\theta}}{y!}, với y \in 0,1,2, \dots$$
   - Áp dụng cho dữ liệu khảo sát, ta có phân phối của dữ liệu sinh sẽ được tính như sau:
     $$\Pr(Y_1 = y_1, Y_2 = y_2 \dots, Y_n = y_n|\theta) = \prod_{i=1}^{n} p(y_i|\theta) = \prod_{i=1}^{n} \frac{1}{y_i!}\theta^{y_i}e^{-\theta}$$
     $$= \theta^{\Sigma y_i}e^{-n\theta} \times c_1(y_1, y_2, \dots y_n)$$

2. **Xác định phân phối hậu nghiệm của tỷ lệ sinh khi phân phối tiên nghiệm là:**
   a. **Phân phối đều:**
   b. **Phân phối gamma:**
   - Từ quy tắc Bayes, ta có công thức sau:
     $$p(\theta|y_1, y_2, \dots, y_n) = \frac{p(y_1, y_2, \dots, y_n|\theta) \times p(\theta)}{p(y_1, y_2, y_3, \dots, y_n)}$$
     Trong đó:
     + $p(\theta|y_1, y_2, \dots, y_n)$ là phân phối của hậu nghiệm $\theta|Y$
     + $p(\theta)$ là phân phối của tiên nghiệm θ
     + $p(y_1, y_2, \dots, y_n|\theta)$ là phân phối likelihood của dữ liệu
     + $p(y_1, y_2, y_3 \dots, y_n)$ là xác suất đồng thời của các quan sát trong tập dữ liệu.
     - Vì $p(y_1, y_2 \dots, y_n)$ là một giá trị cụ thể $c_2(y_1, y_2, \dots, y_n)$, nên suy ra phân phối hậu nghiệm sẽ tỉ lệ thuận với tích của likelihood và tiên nghiệm.
       $$p(\theta|y_1, y_2 \dots, y_n) \propto p(y_1, y_2, \dots, y_n|\theta) \times p(\theta)$$

   a. Nếu phân phối tiên nghiệm là phân phối đều: $\theta \sim U(a, b)$
     $$\mathrm{p}(\theta) = \mathrm{duniform}(\theta, \mathrm{a}, \mathrm{b}) = \begin{cases} 1/(b-a) \; nếu \; a \le \theta \le \; b \\ 0 \; nếu \; ngược \; lại \end{cases}$$
     Vì ngoài khoảng $[a, b]$,xác suất = 0, nên chỉ xét trong khoảng $[a, b]$, ta có kết quả phân phối hậu nghiệm của $\theta|Y$ như sau:

$$p(\theta|y_1, y_2, \dots, y_n) = \frac{p(y_1, y_2, \dots, y_n|\theta) \times p(\theta)}{p(y_1, y_2, y_3, \dots, y_n)}$$
$$= \frac{c_1(y_1, y_2, \dots, y_n)\theta^{\Sigma y_i} e^{-n\theta} \times \left(\frac{1}{b-a}\right)}{c_2(y_1, y_2, \dots, y_n)}$$
$$= \frac{\theta^{\Sigma y_i} e^{-n\theta}}{b-a} \times c_3(y_1, y_2, \dots, y_n)$$
$$= \theta^{\Sigma y_i} e^{-n\theta} \times c_5(y_1, y_2, \dots, y_n, b, a)$$

b. Nếu phân phối tiên nghiêm là phân phối gamma: $\theta \sim \Gamma(\alpha, \beta)$

$$p(\theta) = dgamma(\theta, \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{a-1} e^{-b\theta}, với\ i\ \theta, \alpha, \beta > 0$$

Kết quả phân phối hậu nghiệm như sau:

$$p(\theta|y_1, y_2, \dots, y_n) = \frac{p(y_1, y_2, \dots, y_n|\theta) \times p(\theta)}{p(y_1, y_2, y_3, \dots, y_n)}$$
$$= \frac{c_1(y_1, y_2, \dots, y_n)\theta^{\Sigma y_i} e^{-n\theta} \times \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{a-1} e^{-b\theta}}{c_2(y_1, y_2, \dots, y_n)}$$
$$= \theta^{\Sigma y_i} e^{-n\theta} \times \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{a-1} e^{-b\theta} \times c_3(y_1, y_2, \dots, y_n)$$
$$= \theta^{\Sigma y_i + \alpha - 1} e^{-(n+b)\theta} \times c_4(y_1, y_2, \dots, y_n, \alpha, \beta)$$

3. Nêu đặc điểm các phân phối hậu nghiệm tìm được:

- Từ cả hai kết quả ở câu 2, khi mà tiên nghiệm có phân phối đều hoặc phân phối gamma, thì kết quả của phân phối hậu nghiệm đều có dạng của phân phối gamma
- Tiên nghiệm θ có phân phối đều, thì hậu nghiệm sẽ có phân phối như sau:

$$p(\theta|y_1, y_2, \dots, y_n) = \theta^{\Sigma y_i} e^{-n\theta} \times c_5(y_1, y_2, \dots, y_n, b, a)$$
$$\rightarrow (\theta|y_1, y_2, \dots, y_n) \sim \Gamma(\sum y_i + 1, n)$$

- Tiên nghiệm θ có phân phối gamma, thì hậu nghiệm sẽ có phân phối như sau:

$$p(\theta|y_1, y_2, \dots, y_n) = \theta^{\Sigma y_i + \alpha - 1} e^{-(n+b)\theta} \times c_4(y_1, y_2, \dots, y_n, \alpha, \beta)$$
$$\rightarrow (\theta|y_1, y_2, \dots, y_n) \sim \Gamma(\sum y_i + a, n + b)$$

4. Ước lượng tỷ lệ sinh của 2 nhóm phụ nữ:

- Các phân tích và tính toán sử dụng ngôn ngữ Python.
- Đọc dữ liệu từ file "gss.rdata"

```
import pyreadr

result = pyreadr.read_r('./gss.rdata') # also works for Rds
print(result.keys())
```

```
odict_keys(['gss'])
```

```
df = result["gss"]
```

```
df.head(5)
```

	YEAR	WRKSTAT	MARITAL	AGEWED	CHILDS	AGE	EDUC	PAEDUC	MAEDUC	DEGREE	...	ALIKE3	ALIKE4	ALIKE5	ALIKE6	ALIKE7	ALIKE8	AGEWEI
0	1972	1	5	NaN	0	23	16	10	NaN	3	...	NaN	NaN	NaN	NaN	NaN	NaN	
1	1972	5	1	21	5	70	10	8	8	0	...	NaN	NaN	NaN	NaN	NaN	NaN	
2	1972	2	1	20	4	48	12	8	8	1	...	NaN	NaN	NaN	NaN	NaN	NaN	
3	1972	1	1	24	0	27	17	16	12	3	...	NaN	NaN	NaN	NaN	NaN	NaN	
4	1972	7	1	22	2	61	12	8	8	1	...	NaN	NaN	NaN	NaN	NaN	NaN	

5 rows × 36 columns

- Lọc dữ liệu theo “AGE”, giới tính là “FEMALE” và năm trong thập niên 90s, chỉ giữ lại các cột “CHILDS” và “DEGREE”, sau đó loại bỏ dữ liệu bị thiếu:

```
condition = (df['YEAR'] >= 1990) & (df['YEAR'] < 2000) & (df['AGE'] == 40)& (df['FEMALE'] == 1)
filtered_df = df.loc[condition][['CHILDS', 'DEGREE']].dropna()
```

```
filtered_df.describe()
```

	CHILDS	DEGREE
count	155	155
unique	7	5
top	2	1
freq	51	80

- Chia dữ liệu làm hai nhóm, nhóm có trình độ trên phổ thông “df_high_degree” và nhóm còn lại “df_low_degree”, sau đó lọc chỉ giữ lại đại lượng ngẫu nhiên “CHILDS” là đại lượng đang xét.

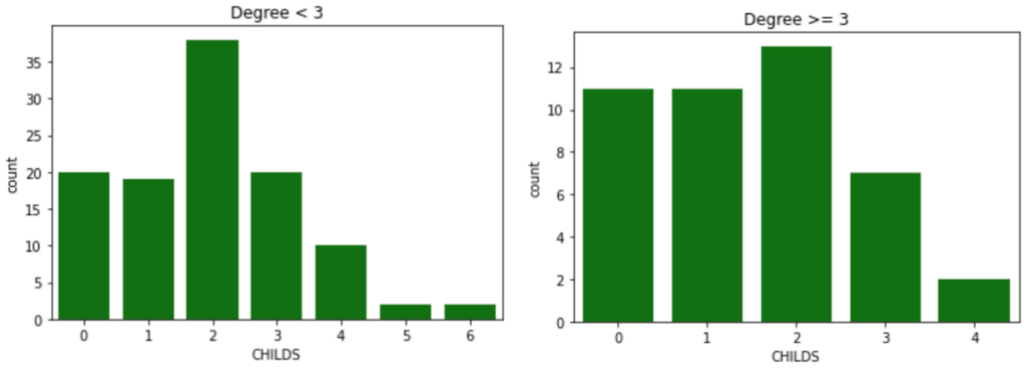
```
df_low_degree = df_y.loc[(df_y['DEGREE'] < 3)][['CHILDS']]
df_low_degree.describe()
```

	CHILDS
count	111
unique	7
top	2
freq	38

```
df_high_degree = df_y.loc[(df_y['DEGREE'] >= 3)][['CHILDS']]
df_high_degree.describe()
```

	CHILDS
count	44
unique	5
top	2
freq	13

- Xem phân bố tỷ lệ sinh của hai nhóm phụ nữ:



- Ước lượng tỷ lệ sinh của hai nhóm phụ nữ chính là ước lượng kỳ vọng của hai nhóm. Vì số trẻ sinh ra có phân phối Poisson nên kỳ vọng chính là kỳ vọng hậu nghiệm $\theta|Y$ tương ứng trên hai nhóm dữ liệu df_low_degree và df_high_degree.
- Giả thiết tiên nghiệm θ có phân phối Gamma(a, b), sử dụng kết quả của câu 3, ta có phân bố hậu nghiệm là phân bố Gamma với các tham số sau α, β tính như sau:

$$\alpha = \sum y_i + a\,, \beta = n + b$$

Với n là kích thước mẫu, y_i là số trẻ em sinh ra của quan sát thứ i

⇨ Giá trị kỳ vọng của phân bố hậu nghiệm:

$$E[\theta|y_1, y_2, \dots, y_n] = \frac{\alpha}{\beta} = \frac{\sum y_i + a}{n+b}$$

- Chọn $a = 2, b = 1$.
- Tỷ lệ sinh của nhóm phụ nữ có trình độ dưới phổ thông và trên phổ thông sẽ có kết quả như sau:

```
# duoi pho thong
n1 = df_low_degree.count()
y1_sum = df_low_degree['CHILDS'].sum()
print(n1, y1_sum)
```

```
CHILDS    111
dtype: int64 217
```

```
# tren pho thong
n2 = df_high_degree.count()
y2_sum = df_high_degree['CHILDS'].sum()
print(n2, y2_sum)
```

```
CHILDS    44
dtype: int64 66
```

```
a = 2
b =1
```

```
# duoi pho thong
e1 = (y1_sum + a)/(n1 + b)

# tren pho thong
e2 = (y2_sum + a)/(n2 + b)

print("Ky vong cua hau nghiem ung voi phu nu co trinh do duoi pho thong: ", e1)
print("Ky vong cua hau nghiem ung voi phu nu co trinh do tren pho thong: ", e2)
```

```
Ky vong cua hau nghiem ung voi phu nu co trinh do duoi pho thong:  CHILDS    1.955357
dtype: float64
Ky vong cua hau nghiem ung voi phu nu co trinh do tren pho thong:  CHILDS    1.511111
dtype: float64
```

- Kết luận:
  - Nhóm phụ nữ dưới trình độ phổ thông có phân phối sau:
    $$\theta|Y_1 \sim Gamma(217+2, 111+1) = Gamma(219, 112)$$
  - Nhóm phụ nữ trình độ trên phổ thông có phân phối sau:
    $$\theta|Y_1 \sim Gamma(66+2, 44+1) = Gamma(68, 45)$$

5. **So sánh tỷ lệ sinh của 2 nhóm và đánh giá sự khác biệt (nếu có):**
   - Ta tiến hành bài toán kiểm định:
   - Mức ý nghĩa lấy mặc định $\alpha = 0.05$
   - Giả thiết: $H_0: \mu_1 = \mu_2$, đối thiết: $H_1: \mu_1 \neq \mu_2$
   - Sử dụng T-test với hai mẫu dữ liệu tương ứng với hai nhóm phụ nữ (DEGREE < 3, và DEGREE >= 3).

```
from scipy.stats import ttest_ind
```

```
stat, p = ttest_ind(df_low_degree['CHILDS'].to_list(), df_high_degree['CHILDS'].to_list(), axis=0)
```

```
print(stat, p)
```

```
1.9304175709985183 0.055404660390320604
```

   - Ta thấy rằng $p - value = 0.0554 > 0.05 = \alpha$, như vậy chưa đủ căn cứ để bác bỏ giả thiết H_0. Như vậy, với mức ý nghĩa 5% thì coi hai nhóm phụ nữ có tỷ lệ sinh bằng nhau.