**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 là đại lượng ngẫu nhiên biểu diễn số lượng trẻ em sinh ra bởi phụ nữ khảo sát trong bộ dữ liệu.
* Gọi là giá trị của ứng với số lượng trẻ em được sinh ra bởi người phụ thứ trong bộ dữ liệu.
* Vì số lượng trẻ em sinh ra tuân theo phân phối Poisson với tham số , ký hiệu . Ta có công thức tổng quát của phân phối Poisson sẽ là:

* Áp dụng cho dữ liệu khảo sát, ta có phân phối của dữ liệu sinh sẽ được tính như sau:

1. **Xác định phân phối hậu nghiệm của tỷ lệ sinh khi phân phối tiên nghiệm là:**
   1. **Phân phối đều:**
   2. **Phân phối gamma:**

* Từ quy tắc Bayes, ta có công thức sau:

Trong đó:

+ là phân phối của hậu nghiệm

+ là phân phối của tiên nghiệm

+ là phân phối likelihood của dữ liệu

+ là xác suất đồng thời của các quan sát trong tập dữ liệu.

* + Vì là một giá trị cụ thể , nên suy ra phân phối hậu nghiệm sẽ tỉ lệ thuận với tích của likelihood và tiên nghiệm.

1. Nếu phân phối tiên nghiệm là phân phối đều:

Vì ngoài khoảng ,xác suất = 0, nên chỉ xét trong khoảng , ta có kết quả phân phối hậu nghiệm của như sau:

b. Nếu phân phối tiên nghiêm là phân phối gamma:

Kết quả phân phối hậu nghiệm như sau:

1. **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 có phân phối đều, thì hậu nghiệm sẽ có phân phối như sau:
* Tiên nghiệm có phân phối gamma, thì hậu nghiệm sẽ có phân phối như sau:

1. **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”

A screenshot of a computer

Description automatically generated

* 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:

A screenshot of a computer

Description automatically generated

* 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.

A screenshot of a computer

Description automatically generated

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

A graph of a degree

Description automatically generatedA graph of a degree

Description automatically generated

* 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 tương ứng trên hai nhóm dữ liệu df\_low\_degree và df\_high\_degree.
* Giả thiết tiên nghiệm 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 tính như sau:

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

* Giá trị kỳ vọng của phân bố hậu nghiệm:

* Chọn .
* Tỷ lệ sinh của nhóm phụ nữ có trình độ dưới phổ thông và trên phổ thông sẽ có kết quả như sau:

A screenshot of a computer

Description automatically generated

* Kết luận:
  + Nhóm phụ nữ dưới trình độ phổ thông có phân phối sau:
  + Nhóm phụ nữ trình độ trên phổ thông có phân phối sau:

1. **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
* Giả thiết: , đối thiết:
* Sử dụng T-test với hai mẫu dữ liệu tương ứng với hai nhóm phụ nữ (DEGREE < 3, và DEGREE >= 3).

A screenshot of a computer

Description automatically generated

* Ta thấy rằng , như vậy chưa đủ căn cứ để bác bỏ giả thiết . Như vậy, với mức ý nghĩa thì coi hai nhóm phụ nữ có tỷ lệ sinh bằng nhau.