

**CÁC PHƯƠNG PHÁP PHÂN TÍCH ĐỊNH LƯỢNG****BÀI TẬP 1****THỐNG KÊ MÔ TẢ VÀ XÁC SUẤT**

Ngày phát bài: 23/10/2018

Hạn nộp bài: 8h20, 06/11/2018

**Bài làm được yêu cầu nộp cả bản cứng và bản điện tử trên kho học liệu mở**

Tập tin “*income-yoeduc.xlsx*” chứa thông tin liên quan đến thu nhập, tình trạng giáo dục, kinh nghiệm làm việc của hơn 7.500 quan sát (những người có thu nhập trong năm) đến từ các tỉnh thành trong cả nước. Thông tin được rút trích từ bộ dữ liệu khảo sát mức sống hộ gia đình – VHLSS 2010 và có một số điều chỉnh, biến đổi để phù hợp với mục đích giảng dạy và thực hành. Các biến cụ thể được cung cấp bao gồm:

province:	tỉnh/thành phố
region:	vùng địa lý
yoeduc:	số năm đi học [tính từ lớp 1] (năm)
yoexper:	số năm kinh nghiệm (năm)
income:	thu nhập trong năm (nghìn đồng)
gender:	giới tính [1 – nữ; 2 – nam]
married:	tình trạng hôn nhân [1 – đã kết hôn; 0 – chưa kết hôn hoặc đã ly hôn]
publicschool:	được đào tạo từ đại học công lập [1 – công lập; 0 – tư thục hoặc chưa học đại học]
public:	làm việc trong khu vực công [1 – khu vực công; 0 – khu vực khác hoặc không đi làm]
foreign:	làm việc trong khu vực FDI [1 – khu vực FDI; 0 – khu vực khác hoặc không đi làm]
official:	làm việc trong khu vực chính thức [1 – khu vực chính thức; 0 – khu vực phi chính thức hoặc không đi làm]

Học viên được yêu cầu sử dụng thông tin cung cấp để bình luận về mối quan hệ giữa thu nhập với số năm đi học và kinh nghiệm công tác của người dân cả nước. Sau đây là các yêu cầu cụ thể.

**Câu 1:** Dựa vào thông tin được cung cấp, anh/chị hãy phân loại các biến được cung cấp theo: biến định lượng hay định tính, biến liên tục hay rời rạc, loại thang đo cụ thể của từng biến? Dữ liệu cung cấp là dữ liệu chéo hay dữ liệu thời gian?

**Câu 2:** Hãy thực hiện tính toán các trị thống kê mô tả cơ bản và bình luận về thu nhập của người dân cả nước? Theo anh/chị, phân phối xác suất của thu nhập lệch xiên về bên nào, giải thích? Tính toán và giải thích ý nghĩa của Kurtosis?

- Câu 3:** Hãy vẽ đồ thị tần suất phân theo từng khoảng thu nhập thay đổi (1 triệu đồng). Bình luận đồ thị và so sánh kết quả đánh giá độ lệch xiên từ các trị thống kê mẫu đã tính trước đó?
- Câu 4:** Lập bảng số phân vị thu nhập theo các mức 10%; 50%; 90%; 95%; 99%. Kết luận chung của anh chị từ các kết quả đã tính tại Câu 1 – 4 về thu nhập của người dân cả nước?
- Câu 5:** Hãy tính hệ số tương quan giữa thu nhập, số năm đi học và số năm kinh nghiệm. Kết luận của anh/chị là gì từ kết quả tính toán?
- Câu 6:** Hãy kiểm tra lại mối quan hệ giữa số năm đi học với thu nhập và giữa số năm kinh nghiệm với thu nhập bằng phương pháp đồ thị scatter. Bình luận của anh chị và so sánh với kết quả ở Câu 5?
- Câu 7:** Anh chị hãy tính toán các chỉ tiêu cơ bản về: Trung bình, Giá trị nhỏ nhất, Giá trị lớn nhất, Khoảng tin cậy, Độ lệch chuẩn của thu nhập phân theo từng vùng địa lý. Bình luận của anh/chị về thu nhập giữa các vùng miền? Hãy so sánh độ biến thiên trong thu nhập của người dân tại vùng Duyên hải Miền Trung và vùng Tây Nguyên?
- Câu 8:** Anh chị hãy so sánh thu nhập bình quân phân theo các nhóm biến: gender, married, publicschool, public, foreign, official? Bình luận về kết quả tính toán và tỷ lệ tham gia của người nhân trong từng nhóm thuộc tính?
- Câu 9:** Hãy tổng hợp tất cả các tính toán và thông tin có được từ Câu 1 – 8 hoặc thông tin anh/chị khai thác thêm từ bộ dữ liệu cung cấp. Hãy truyền tải ngắn gọn các thông điệp về mối quan hệ giữa thu nhập với số năm đi học và số năm kinh nghiệm, cũng như thu nhập phân theo vùng miền và các thuộc tính khác? *[Tối đa không quá 1 trang A4 cho phần trình bày]*
- Câu 10:** Giả định rằng một người có số năm đi học từ 1 - 5 năm sẽ có trình độ tiểu học, từ 6 – 9 năm sẽ có trình độ trung học cơ sở, từ 10 – 12 năm sẽ có trình độ THPT, và từ 13 năm trở lên có trình độ trên ĐH-CD; và những người có số năm đi học bằng 0 là không đi học. Hãy lập bảng tần suất 2 chiều phân theo vùng địa lý và các cấp trình độ học vấn như trên, và trả lời các câu hỏi cụ thể sau đây:
- Xác suất để chọn một người mà người này không đi học?
  - Xác suất để chọn một người có trình độ từ THPT trở lên?
  - Xác suất để chọn một người vừa có trình độ từ CD-ĐH trở lên vừa đến từ khu vực Tây nguyên?
  - Xác suất để chọn một người trong vùng Tây Nguyên mà người này là không đi học?
  - Xác suất để chọn một người trong nhóm không đi học mà người này đến từ vùng ĐBSH?
- Câu 11:** Kết quả tính toán cho thấy vùng ĐBSCL là vùng có tỷ lệ không đi học cao nhất (1,6%). Để làm rõ nguyên nhân vì sao có hiện trạng trên, bạn muốn chọn mẫu ngẫu nhiên để thực hiện phỏng vấn sâu một số người đã tham gia khảo sát trước đó. Bạn cần phỏng vấn cả những người không đi học và những người có đi học, có cả nam và nữ.
- Hãy tính xác suất để một người được chọn trong vùng ĐBSCL thỏa mãn các điều kiện sau:
- Người được chọn không đi học?
  - Người được chọn vừa là nam, vừa không đi học?

- c. Người nữ được chọn là người không đi học?
- d. Người không đi học được chọn là người nam?

**---HẾT---**