

CÁC PHƯƠNG PHÁP ĐỊNH LƯỢNG 1**BÀI THI GIỮA KỲ**

Ngày phát: 05/12/2023

Hạn nộp: 8h20, 10/12/2023

Bài làm được yêu cầu chỉ nộp bản điện tử trên Microsoft Teams**PHẦN A: CÂU HỎI LÝ THUYẾT (30đ)**

Hãy trả lời ngắn gọn các câu hỏi sau đây:

Câu 1: Cung cấp một ví dụ cho (1a) biến ngẫu nhiên rời rạc và (1b) biến ngẫu nhiên liên tục. Phân phối xác suất của hai biến này theo bạn là như thế nào?

Câu 2: Biến ngẫu nhiên X có quy luật phân phối là một tam giác đều trong khoảng 0 đến 100. Hãy vẽ minh họa và tính toán hàm phân phối p.d.f. và c.d.f. của biến X . Sau đó, tính xác suất của X nằm trong khoảng từ 20 đến 40.

Câu 3(*): Khảo sát nhiều học viên FSPPM các khoá khi học các môn ‘Các phương pháp định lượng 1’ và ‘Luật và chính sách công’ thì thời gian làm bài tuân theo phân phối chuẩn với các hệ số được thể hiện theo bảng sau đây:

Đơn vị: giờ	Các PP định lượng 1 (X)		Luật và chính sách công (Y)	
	Giá trị kì vọng	Độ lệch chuẩn	Giá trị kì vọng	Độ lệch chuẩn
Thời gian chuẩn bị	$E(X_1) = 2$	$\sigma_{X_1} = 1$	$E(Y_1) = 3$	$\sigma_{Y_1} = 3$
Thời gian trình bày	$E(X_2) = 1$	$\sigma_{X_1} = 1$	$E(Y_2) = 2$	$\sigma_{Y_2} = 2$

Thời gian chuẩn bị và thời gian trình bày độc lập với nhau. Hệ số tương quan giữa thời gian làm bài hai môn học là 0.8. Dựa vào các thông tin được cung cấp, hãy tính toán xác suất để học viên dùng không quá 12 giờ để hoàn thành 2 bài tập hai môn (giả định rằng học viên không làm song song 2 môn cùng một lúc).

Gọi ý: $\text{Var}(X \pm Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$

Câu 4: Giải thích sự khác biệt giữa giá trị thu thập thực tế cho một mẫu n quan sát $(x_1, x_2, \dots, x_n)$ và trừu tượng hóa các giá trị đó dưới dạng mẫu ngẫu nhiên cỡ mẫu n . Bạn hãy dùng một ví dụ đơn giản để minh họa.

Câu 5: Giải thích ngắn gọn chúng ta cần làm gì khi xây dựng một ước lượng điểm từ một mẫu ngẫu nhiên cỡ n . Tại sao ước lượng điểm lại là một biến ngẫu nhiên?

Câu 6: Khảo sát thị trường lao động tại TP. Hồ Chí Minh cho thấy phân phối mức lương theo tháng có giá trị kì vọng là 10 triệu VNĐ và độ lệch chuẩn là 4 triệu VNĐ. Giả định phân phối chuẩn là phân phối tổng thể. Hãy tính:

- Xác suất một người có mức lương lớn hơn 20 triệu một tháng.
- Thu thập một mẫu 25 người từ một địa phương chưa biết, cho thấy mức lương trung bình mẫu là 8 triệu. Với mức ý nghĩa 5%, hãy sử dụng kiểm định giả thuyết thống kê xem mẫu này có đặc điểm giá trị kì vọng giống như tổng thể TP. Hồ Chí Minh hay không?
- (*) Kết quả câu (a) và (b) thay đổi thế nào nếu phân phối thu nhập là hàm loga chuẩn (log-normal).

Gợi ý: bạn sử dụng hàm `LOGNORM.DIST()` để tìm giá trị hàm c.d.f. cho phân phối log-normal. Khi phân phối log-normal X có phương sai và độ lệch chuẩn μ_X và σ_X^2 , thì giá trị cần

lựa chọn cho hàm `LOGNORM.DIST( ,  $\mu, \sigma$ , TRUE)` là
$$\begin{cases} \mu = \log\left(\frac{\mu_X^2}{\sqrt{\mu_X^2 + \sigma_X^2}}\right) \\ \sigma^2 = \log\left(1 + \frac{\sigma_X^2}{\mu_X^2}\right) \end{cases}$$

PHẦN B: PHÂN TÍCH DỮ LIỆU SỬ DỤNG EXCEL (70đ)

Các bạn được cung cấp một phần bộ dữ liệu Điều tra mức sống hộ dân cư Việt Nam (VHLSS) năm 2014¹. Dữ liệu và mô tả biến được cung cấp ở file *Midterm.xlsx*. Bạn có thể sử dụng bất kì phần mềm nào theo độ thông thạo của bạn để trả lời những câu hỏi sau đây:.

Câu 1: Kiểm tra xem mẫu thu thập có giá trị khuyết (missing values) không? Đếm số lượng missing values và bỏ các quan sát đó ra khỏi mẫu (Gợi ý: sử dụng chức năng ‘Go to’ của Excel và chọn). Kiểm tra xem mẫu có giá trị bất thường (outliers) cho biến thu nhập wage hay không? Nếu có hãy đề xuất phương án xử lý.

Câu 2: Có ý kiến cho rằng, đặc điểm về kinh tế - xã hội giữa hai vùng Đông Nam Bộ và Đồng Bằng Sông Cửu Long (ĐBSCL) không giống nhau. Hãy sử dụng thống kê mô tả từ các biến trong bộ dữ liệu để làm rõ ý kiến trên. Bạn trình bày như một bài phân tích ngắn với giới hạn 1,000 chữ (bao gồm bảng biểu).

Câu 3: Vẽ biểu đồ histogram cho biến thu nhập wage cho 13 tỉnh ĐBSCL. Có cách nào để chuyển biến này thành một biến có xấp xỉ phân phối chuẩn hay không?

Câu 3b (*): So sánh độ nhọn và độ nghiêng của biến mới này với con số tiêu chuẩn để xác minh rõ hơn nhận định của bạn.

Câu 4: Dùng lệnh `ranbetween()` tương tự như file demo đã học để chọn một mẫu ngẫu nhiên với cỡ mẫu 20 từ biến *wage*. Thực hiện điều này liên tục để tạo ra 200 mẫu. Với mỗi mẫu hãy tính

¹ Bộ dữ liệu này đã được tinh chỉnh cho mục đích sư phạm, không phù hợp cho các mục đích khác

giá trị trung bình và phương sai. Vẽ đồ thị histogram cho 200 trung bình và phương sai mẫu này. Bạn hãy tăng cỡ mẫu lên và mô tả xu hướng thay đổi của đồ thị histogram, từ đó liên hệ với kiến thức đã học.

Câu 5: Ở độ tin cậy 95%, hãy ước lượng khoảng cho trung bình tổng thể thu nhập của tỉnh 13 tỉnh ĐBSCL. Với mức ý nghĩa 1%, kiểm định giả thuyết thống kê số năm học trung bình của 13 tỉnh ĐBSCL có bằng trung bình tổng thể cả nước là 8.5 năm hay không?

Câu 6: Có ý kiến cho rằng chi phí lao động khu vực FDI ở giữa ba tỉnh giáp ranh TP.HCM, bao gồm Bình Dương, Đồng Nai và Long An, có sự khác biệt đáng kể. Bạn hãy chọn phương pháp phù hợp với bộ dữ liệu để làm rõ ý kiến trên (nêu rõ giả định bạn sử dụng để thực hiện và không cần kiểm tra giả định).

Câu 7: Tính hệ số tương quan giữa số năm đi học và thu nhập cho quan sát ở 13 tỉnh ĐBSCL. Hệ số tương quan giúp bạn có kết luận gì?

Câu 7b (*). Vẽ biểu đồ scatter plot giữa độ tuổi và thu nhập cho các quan sát ở TP.HCM. Bạn có nhận xét gì về mối quan hệ giữa hai biến này?

Ghi chú: Mỗi câu hỏi (*) ở cả hai phần đều có giá trị 5 điểm.

---HẾT---