

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

Ngày phát: 8h20, 02/12/2021

Hạn nộp: 8h20, 04/12/2021

Học viên được yêu cầu chỉ nộp bản điện tử**Câu 1.** Suy luận thống kê liên quan đến tỷ lệ tổng thể khi mẫu nhỏ

Tỷ lệ xảy ra biến chứng bình quân của các ca phẫu thuật hiến ghép thận là 9%. Một vị bác sỹ thử nghiệm một phương pháp mới, theo đó chỉ để xảy ra 4 trường hợp biến chứng trong tổng số 56 ca hiến ghép được thực hiện theo phương pháp này.

a) Anh/Chị hãy kiểm định giả thuyết của vị bác sỹ trên cho rằng phương pháp mới giúp giảm tỷ lệ biến chứng sau phẫu thuật. Các anh chị trình bày các cấu phần của bài kiểm định, bao gồm:

- Giả thuyết trống (H_0) và giả thuyết thay thế (H_a);
- Trị thống kê của bài kiểm định, giả sử rằng phân phối mẫu của tỷ lệ xảy ra biến chứng theo một phân phối chuẩn. Theo Anh/Chị giả sử này có xác đáng không?
- P-value của bài kiểm định;
- Kết luận của bài kiểm định tương ứng với mức kiểm định (alpha) do Anh/Chị tự chọn.

b) Giả sử rằng phân phối mẫu của tỷ lệ xảy ra biến chứng không tuân theo phân phối chuẩn. và giả sử rằng chúng ta có được ghi nhận về các trường hợp biến chứng trong 1,000,000 ca phẫu thuật trước đó (xem file **complications.csv**). Anh/Chị hãy:

- Xây dựng phân phối mẫu của tỷ lệ xảy ra biến chứng cho bài kiểm định ở câu (a), với giả định rằng 1,000,000 ca phẫu thuật được ghi nhận trước đó là tổng thể. Lưu ý, khi xây dựng phân phối mẫu, Anh/Chị nêu rõ số lần lấy mẫu và số quan sát trong từng mẫu;
- Tính trị p-value của bài kiểm định từ phân phối mẫu này;
- Kết luận về bài kiểm định dựa trên mức kiểm định (alpha) trong câu (a). Anh/Chị có nhận xét gì về kết quả của bài kiểm định ở hai câu (a) và (b)?

c) Một nhóm nghiên cứu áp dụng phương pháp mới này với 37 ca hiến ghép thận với một trong những mục đích nhằm kiểm chứng kết quả của vị bác sỹ trên. H_0 quan sát thấy có 3 trường hợp biến chứng và cho rằng tỷ lệ xảy ra biến chứng trong thử nghiệm của vị bác sỹ nọ là quá thấp.

Anh/Chị hãy kiểm định giả thuyết cho rằng tỷ lệ xảy ra biến chứng trong thử nghiệm của nhóm nghiên cứu cao hơn tỷ lệ xảy ra biến chứng trong thử nghiệm của vị bác sỹ trên. Anh/Chị hãy:

- Xây dựng giả thuyết trống (H_0) và giả thuyết thay thế (H_a) của bài kiểm định;
- Xây dựng phân phối mẫu của sự khác nhau về tỷ lệ xảy ra biến chứng giữa hai thử nghiệm, với giả định rằng 1,000,000 ca phẫu thuật được ghi nhận trước đó (trong câu (b)) là tổng thể. Anh/chị cần nêu rõ số lần lấy mẫu và số quan sát trong từng mẫu của hai thử nghiệm;
- Tính trị p-value của bài kiểm định từ phân phối mẫu này;
- Kết luận về bài kiểm định tương ứng với mức kiểm định do Anh/Chị tự chọn.

Lưu ý: Các số liệu trên được giả lập với mục đích phục vụ bài thi và hoàn toàn không có giá trị tham khảo y khoa.

Câu 2. Phân tích khám phá dữ liệu trong file **bankExam.csv**. Cụ thể trong khoảng 3000 từ (dao động trong khoảng $\pm 5\%$), bao gồm đồ thị (mỗi đồ thị tương đương 250 từ), Anh/Chị hãy trình bày các đặc điểm của bộ dữ liệu dựa theo các yêu cầu cụ thể dưới đây:

- a) Anh/Chị vẽ biểu đồ histogram và boxplots cho các biến ‘age’ và ‘balance’, biểu đồ cột cho các biến ‘job’, ‘marital’, và ‘education’. Dựa vào các biểu đồ này và các trị thống kê cơ bản, Anh/Chị nhận định ngắn gọn về các cột số liệu này;
- b) Anh/chị thay thế các số liệu bị mất trong cột ‘age’ (được đánh dấu bằng giá trị -1) bằng tuổi trung bình của nhóm quan sát có cùng giá trị ‘job’ với hàng có số liệu ‘age’ bị mất (làm tròn tuổi trung bình đến số nguyên gần nhất);
- c) Chúng ta có đủ bằng chứng để kết luận giá trị trung bình của cột ‘age’ trước và sau khi thực hiện thay thế giá trị bị mất có khác nhau không? Anh/Chị hãy thực hiện bài kiểm định tương ứng. Những giả thiết cần có để thực hiện được bài kiểm định này là gì? (Giá trị trung bình của cột ‘age’ trước khi thực hiện thay thế giá trị bị mất được tính dựa trên các giá trị đã có);
- d) Anh/chị thay thế các giá trị bị mất trong cột ‘education’ (được đánh dấu bằng ký tự ‘-’) dựa trên tần suất xuất hiện của các giá trị hiện hữu. Anh/Chị nhận định ngắn gọn về cột dữ liệu này sau khi thực hiện thay thế dữ liệu bị mất (ví dụ, tỷ lệ của các giá trị trong cột có thay đổi không và mức độ thế nào);
- e) Anh/chị xác định các điểm dữ liệu ngoại lệ có thể có trong cột ‘balance’. Mô tả phương pháp Anh/Chị dùng để xác định các điểm ngoại lệ và nhận định ngắn gọn về các điểm dữ liệu này;
- f) Kiểm định giả thuyết cho rằng có sự tương quan giữa hai biến ‘education’ và ‘job’. Anh/Chị trình bày chi tiết các bước của bài kiểm định, tự chọn mức kiểm định và đưa ra kết luận tương ứng.

Lưu ý: Anh/Chị chỉ cần thực hiện thay thế dữ liệu bị mất 1 lần (single imputation). Anh/Chị có thể dùng công cụ tính toán phù hợp, ví dụ Excel, R, Python, SPSS, Stata hoặc một phần mềm khác. Tuy nhiên, Anh/Chị cần trình bày các bước phân tích và kết quả đầy đủ và súc tích. Hạn chế đưa các hàm hoặc các dòng code trực tiếp vào báo cáo. Thay vào đó, Anh/Chị mô tả chức năng, thuộc tính, các giả định và các thông số đầu vào và các kết quả của các hàm hoặc các dòng code đó.

---HẾT---