

XÂY DỰNG VÀ SỬ DỤNG BỘ DỮ LIỆU MÔ PHÒNG BẰNG NGÔN NGỮ LẬP TRÌNH PYTHON TRONG GIẢNG DẠY THỐNG KÊ MÔ TẢ

Nguyễn Hải Tuấn¹, Phùng Chí Thiện^{2*}

¹*Viện Vệ sinh dịch tễ Trung ương*

²*Trung tâm mô phỏng lâm sàng, Trường Đại học Y Dược Hải Phòng*

TÓM TẮT

Nghiên cứu được thực hiện nhằm xây dựng và sử dụng bộ dữ liệu mô phỏng trong đào tạo kỹ năng phân tích số liệu cho sinh viên và cán bộ y tế. Bộ số liệu mô phỏng về giới tính, cân nặng và tình trạng sống nội trú của 150 sinh viên được tạo bằng ngôn ngữ Python với các tham số có kiểm soát. Kết quả cho thấy phương pháp mô phỏng giúp trực quan hóa sinh động lý thuyết sai số thống qua biểu đồ hình bia, đồng thời cho phép thực hành trọn vẹn quy trình thống kê mô tả (trung bình, trung vị, yếu vị và khoảng tứ phân vị) và kiểm định giả thuyết (χ^2 , t-test, tỷ số chênh - Odds Ratio). Việc sử dụng dữ liệu mô phỏng giúp khắc phục khó khăn thực hiện nguyên tắc bảo mật dữ liệu thực và tạo ra kịch bản giảng dạy phù hợp, góp phần nâng cao năng lực phân tích dữ liệu trong giám sát và đáp ứng với dịch bệnh của mạng lưới y tế dự phòng.

Từ khóa: Dữ liệu mô phỏng; ngôn ngữ lập trình Python; thống kê mô tả; đào tạo thống kê

I. ĐẶT VẤN ĐỀ

Trong lĩnh vực y tế công cộng, phân tích dữ liệu là kỹ năng cốt lõi để thực hiện giám sát dịch tễ học giúp đánh giá thực trạng sức khỏe cộng đồng và đưa ra các quyết định dựa trên bằng chứng [1]. Việc nắm vững kiến thức thống kê cơ bản là cần thiết để cán bộ y tế có thể phiên giải chính xác các chỉ số sức khỏe [2]. Tuy nhiên, công tác đào tạo hiện nay đang gặp nhiều thách thức khi sử dụng dữ liệu thực tế. Dữ liệu thực tế yêu cầu bảo mật thông tin cá nhân, là rào cản trong việc chia sẻ rộng rãi [3]. Bên cạnh đó, số liệu thực tế thường không có các đặc điểm điển hình để minh họa rõ nét các hiện tượng thống kê như sai số đo lường hoặc sự vi phạm các giả định phân phối chuẩn [4]. Việc sử dụng dữ liệu mô phỏng đã được sử dụng khá rộng rãi trong giảng dạy một số lĩnh vực như y học cơ sở, lâm sàng, y tế công cộng, nghiên cứu thực nghiệm phương pháp trong

dịch tễ học và thống kê y sinh [5, 6].

Lợi ích lớn nhất của dữ liệu mô phỏng là tính trực quan, nguồn mở, khả năng tái lập và đặc biệt là có thể kiểm soát hoàn toàn các tham số như trung bình, độ lệch chuẩn và sai số hệ thống [5]. Cách tiếp cận này cho phép giảng viên thiết kế các kịch bản tình huống thống kê cho mục đích sư phạm rõ ràng, từ đó góp phần nâng cao kỹ năng thực hành phân tích số liệu trong điều tra và giám sát dịch tễ học.

Để khắc phục khó khăn trong việc sử dụng dữ liệu thực tế cho công tác đào tạo và hỗ trợ hoạt động giảng dạy kỹ năng phân tích số liệu dịch tễ học, nghiên cứu này được thực hiện với mục tiêu xây dựng và thử nghiệm sử dụng bộ dữ liệu mô phỏng minh họa kỹ thuật giảng dạy thống kê mô tả và thống kê suy luận cơ bản.

II. PHƯƠNG PHÁP NGHIÊN CỨU

*Tác giả: Phùng Chí Thiện

Đơn vị công tác: Trường Đại học Y Dược Hải Phòng

Điện thoại: +84 902085996

Email: pcthien@hpmu.edu.vn

Ngày nhận bài: 04/03/2026

Ngày phản biện: 13/03/2026

Ngày đăng bài: 12/05/2026

2.1 Đối tượng nghiên cứu

Bộ dữ liệu mô phỏng gồm 150 sinh viên Y khoa năm thứ 4 (gồm 60 nam và 90 nữ).

2.2 Địa điểm và thời gian nghiên cứu

Nghiên cứu được thực hiện tại Trung tâm Mô phỏng Lâm sàng, Trường Đại học Y Dược Hải Phòng. Thời gian thực hiện nghiên cứu từ tháng 9 đến tháng 12 năm 2025.

2.3 Thiết kế nghiên cứu

Nghiên cứu phương pháp luận sử dụng dữ liệu mô phỏng. Nghiên cứu thử nghiệm bộ dữ liệu mô phỏng kịch bản giảng dạy thống kê mô tả và suy luận.

2.4 Cỡ mẫu nghiên cứu

Cỡ mẫu cho mô hình mô phỏng thực nghiệm gồm 150 sinh viên Y khoa năm thứ 4, trong đó có 60 nam và 90 nữ, dựa trên tỷ lệ nữ thường cao hơn nam sinh viên theo học ngành y [7]. Thiết lập kích thước dữ liệu mô phỏng có chủ đích nhằm đáp ứng mục tiêu giảng dạy thống kê mô tả và suy luận cơ bản, bảo đảm đủ số quan sát cho: (1) ước lượng trung bình và độ phân tán ổn định; (2) kiểm định giả thuyết so sánh hai trung bình độc lập; (3) phân tích bảng 2 x 2: ước lượng tỷ số chênh – Odds Ratio và sai số chuẩn (Standard Error – SE) của $\ln(OR)$ ($OR = 2,5$ và $SE = 0,35$ được chọn có chủ đích). Với cấu trúc bảng 2 x 2 (giới tính x tình trạng nội trú), mỗi ô có giá trị kỳ vọng ≥ 5 , đáp ứng điều kiện áp dụng kiểm định χ^2 và ước lượng OR với sai số cho trước SE.

Để đáp ứng các giả định trên công thức tính cỡ mẫu dựa trên cách tính sai số chuẩn của $\ln(OR)$, với OR kỳ vọng bằng 2,5 và $SE = 0,35$.

$$SE[\ln(OR)] = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Trong đó, a: số đối tượng nam sống nội trú; b: số đối tượng nam sống ngoại trú; c: số đối tượng nữ sống nội trú; d: số đối tượng nữ sống ngoại trú. Giả sử $P_0 = 0,5$ với $OR = 2,5$ thì $P_1 =$

0,7143 tính theo công thức $OR = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$; n là số mẫu cho mỗi nhóm:

$$n = \frac{\frac{1}{p_1} + \frac{1}{1-p_1} + \frac{1}{p_0} + \frac{1}{1-p_0}}{SE^2[\ln(OR)]}$$

Cỡ mẫu tính được theo công thức là $N = 2n = 146$, chúng tôi làm tròn là 150. Phân phối với tỷ số giới nữ/nam là 1,5 sẽ tương ứng 60 nam và 90 nữ sinh viên. Đây là cỡ mẫu nhỏ nhất cần thiết cho giả định về $OR = 2,5$ và $SE = 0,35$ và đồng thời đáp ứng giả định phân phối chuẩn của của cân nặng ở từng giới và giả định của kiểm định χ^2 .

2.5 Phương pháp chọn mẫu

Nghiên cứu thiết lập cấu trúc dữ liệu mô phỏng gồm 60 sinh viên nam và 90 sinh viên nữ. Do đây là dữ liệu mô phỏng, quá trình chọn mẫu thực chất là quá trình sinh số liệu ngẫu nhiên có kiểm soát tham số.

2.6 Biến số nghiên cứu

Bộ dữ liệu mô phỏng gồm các biến: Giới tính (Nam/Nữ), cân nặng (kg) là biến định lượng liên tục và tình trạng nội trú (Nội trú/ Ngoại trú). Trong kịch bản mô phỏng sai số đo lường, các biến bổ sung gồm giá trị quan sát, sai số hệ thống (bias, kg), độ lệch chuẩn của sai số ngẫu nhiên (SD), sai số quan sát và giá trị đo được (kg). Các chỉ số thống kê được xác định từ dữ liệu bao gồm trung bình, trung vị, yếu vị, độ lệch chuẩn và khoảng tứ phân vị.

2.7 Phương pháp thu thập thông tin

Quy trình khởi tạo dữ liệu

Dữ liệu được mô phỏng bằng ngôn ngữ Python (phiên bản 3.12) trên 150 sinh viên năm thứ 4, gồm 60 nam và 90 nữ. Cân nặng được sinh ngẫu nhiên theo phân phối chuẩn riêng cho từng giới, sau đó bổ sung các kịch bản sai số đo lường với mức bias và độ lệch chuẩn khác nhau để minh họa độ đúng và độ chụm. Đồng thời, biến nội trú được gán theo

giới với tần số định trước để tạo OR mục tiêu bằng 2,5 với $SE = 0,35$ phục vụ cho phân tích suy luận, cụ thể các tham số như sau: Cân nặng được mô phỏng theo phân phối chuẩn với Nam (trung bình $\mu = 65$, độ lệch chuẩn $\sigma = 8$) và Nữ ($\mu = 52$, $\sigma = 6$). Kịch bản mô phỏng sai số (Lý thuyết đo lường) gồm bốn kịch bản sai số được thiết kế dựa trên sự thay đổi của giá trị trung bình kỳ vọng của sai số (Bias) và độ lệch chuẩn (SD) của sai số ngẫu nhiên: (1) Chùm cao - Đúng cao; (2) Chùm thấp - Đúng cao; (3) Chùm cao - Đúng thấp; (4) Chùm thấp - Đúng thấp. Mỗi kịch bản được tạo ra gồm 25 quan sát ngẫu nhiên tương ứng với các tình huống: độ đúng cao và độ chụm cao ($bias = 0$ và $SD_{\text{sai số}} = 0,8$); độ đúng cao và độ chụm thấp ($bias = 0$ và $SD_{\text{sai số}} = 4,0$); độ đúng thấp và độ chụm cao ($bias = 4,0$ và $SD_{\text{sai số}} = 0,5$) và độ đúng thấp và độ chụm thấp ($bias = 4,0$ và $SD_{\text{sai số}} = 4,0$). Biểu “tình trạng nội trú - sống trong ký túc xá của trường hay ở ngoài” được thiết lập có chủ đích để có mối liên quan với giới tính (Odds Ratio dự kiến xấp xỉ bằng 2,5) và thực hiện kiểm định thống kê χ^2 có $p < \alpha$, với $\alpha = 0,05$.

2.8 Xử lý và phân tích số liệu

Dữ liệu mô phỏng được tạo và phân tích bằng Jupyter Notebook với các thư viện Pandas, Matplotlib và Scipy. Các kỹ thuật phân tích minh họa cho giảng dạy bao gồm: trực

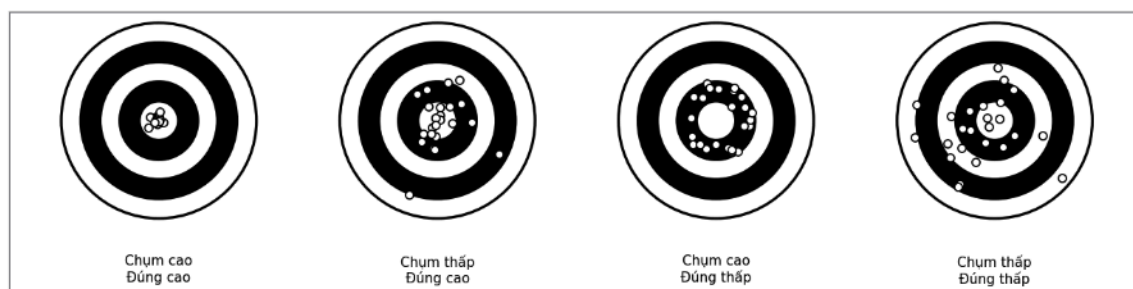
quan hóa số liệu bằng Biểu đồ bia bắn (Bull's eye plot), biểu đồ hộp (Boxplot) và biểu đồ cột (Histogram). Thống kê mô tả sử dụng giá trị trung bình - Mean, trung vị - Median, yếu vị - Mode, khoảng tứ phân vị - IQR và SD. Thống kê suy luận được thực hiện với kiểm định t-test, kiểm định χ^2 và tính toán Odds Ratio (OR) kèm khoảng tin cậy 95%. Nghiên cứu sử dụng mức ý nghĩa $\alpha = 0,05$.

2.9 Đạo đức nghiên cứu

Mô phỏng số liệu là hoạt động thường xuyên phục vụ huấn luyện kỹ năng cho sinh viên y khoa tại Trung tâm Mô phỏng Lâm sàng, Trường Đại học Y Dược Hải Phòng. Trung tâm được thành lập theo Nghị quyết số 29/NQ-HĐT-YDHP ngày 18 tháng 10 năm 2021. Tất cả dữ liệu mô phỏng tuân thủ các nguyên tắc đạo đức, bảo mật trong nghiên cứu y sinh và quy định về chia sẻ thông tin. Để đảm bảo tính minh bạch, khả năng tái lập và dễ tiếp cận học liệu, toàn bộ mã nguồn sử dụng để mô phỏng và phân tích số liệu được chia sẻ công khai trên nền tảng GitHub (https://github.com/nghait/basic_stat).

III. KẾT QUẢ

3.1 Trực quan hóa lý thuyết đo lường sai số qua biểu đồ hình bia

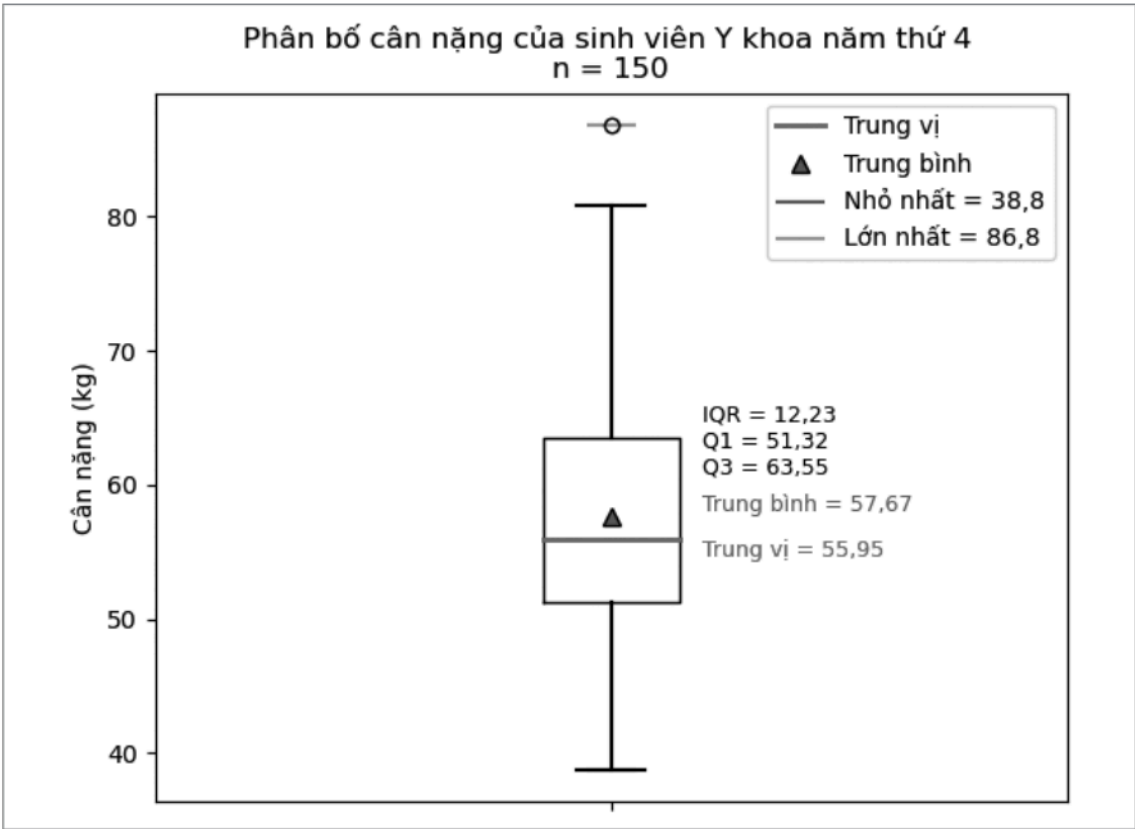


Hình 1. Mô phỏng 4 kịch bản sai số đo lường kết hợp giữa độ đúng và độ chụm

Hình ảnh mô phỏng 4 tình huống về sai số đo lường, mỗi kịch bản được tạo ra gồm 25 quan sát ngẫu nhiên. Tâm bia đại diện cho sai số hệ thống bằng 0 kg, từ trái qua phải: độ đúng cao và độ chụm cao ($bias = 0$ và $SD_{\text{sai số}} = 0,8$); độ đúng

cao và độ chụm thấp ($bias = 0$ và $SD_{\text{sai số}} = 4,0$); độ đúng thấp và độ chụm cao ($bias = 4,0$ và $SD_{\text{sai số}} = 0,5$) và độ đúng thấp và độ chụm thấp ($bias = 4,0$ và $SD_{\text{sai số}} = 4,0$).

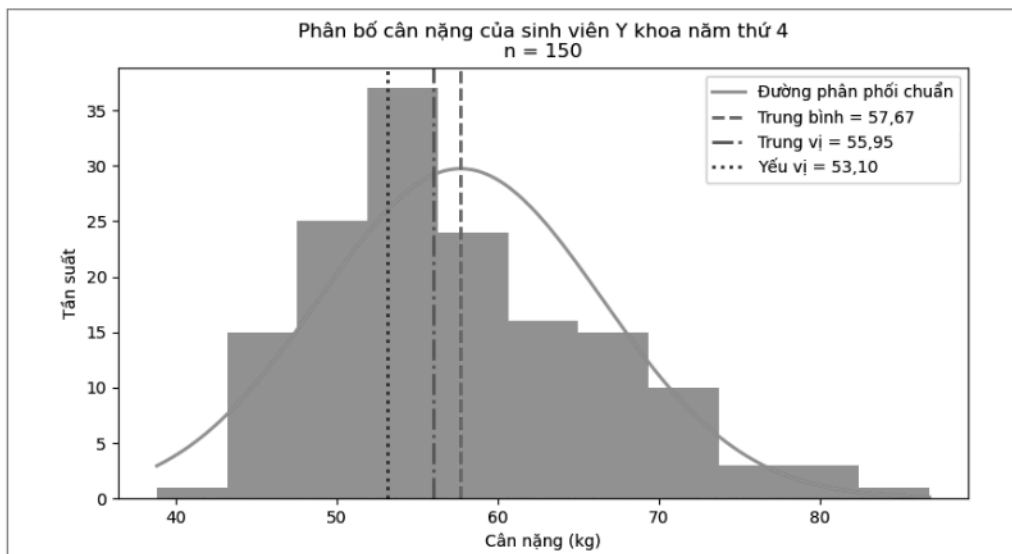
3.2 Đặc điểm phân phối cân nặng của sinh viên



Hình 2. Biểu đồ hộp phân bố cân nặng của sinh viên Y khoa năm thứ 4 (n = 150)

Phân bố cân nặng của sinh viên Y khoa năm thứ 4 cho thấy 75 (50%) sinh viên có cân nặng nằm trong khoảng từ 51,32 kg (Q1) đến 63,55

kg (Q3). Biểu đồ ghi nhận sự xuất hiện của một giá trị ngoại lệ ở mức cân nặng cao phía trên.



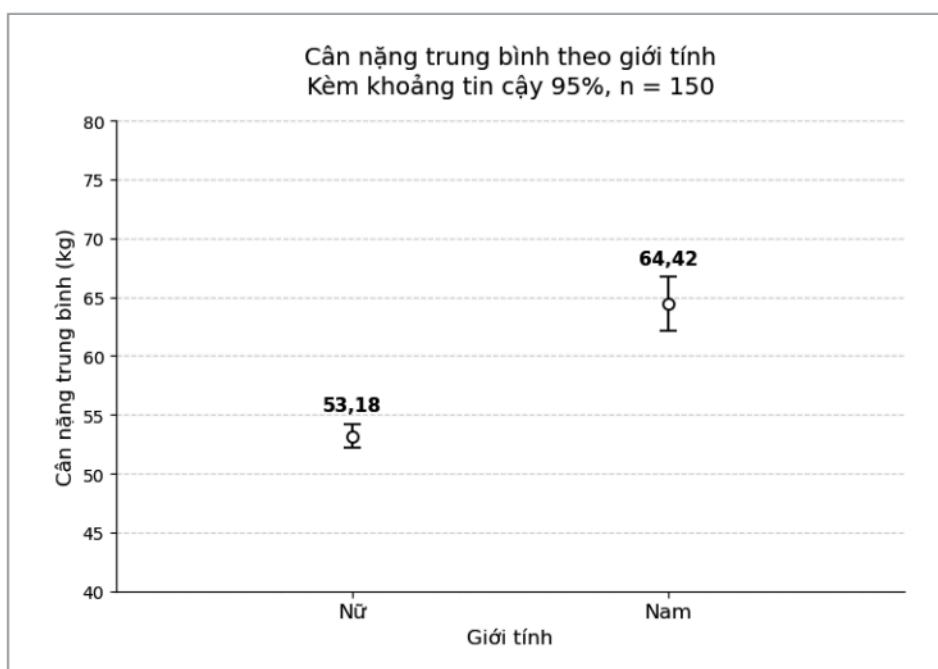
Hình 3. Biểu đồ cột phân bố cân nặng của sinh viên Y khoa năm thứ 4 (n = 150)

Biểu đồ cột (Histogram) kết hợp đường phân phối chuẩn cho thấy dữ liệu có độ lệch nhẹ về phía phải. Giá trị cân nặng trung bình $\pm$ ĐLC là $57,67 \pm 8,78$ kg, cao hơn so với trung vị (55,95 kg; Q1 – Q3: 51,32 – 63,55) và yếu vị (53,10 kg). Đa số sinh viên có cân nặng tập trung trong khoảng từ 45 kg đến 65 kg. Kết quả

kiểm định giả thuyết phân phối chuẩn Shapiro–Wilk cho thấy cân nặng không tuân theo phân phối chuẩn khi phân tích trên toàn bộ mẫu ($p = 0,0013$).

3.3. Phân tích mối liên quan biến định tính

So sánh cân nặng trung bình theo giới



Hình 4. Cân nặng trung bình-KTC 95% của sinh viên Y khoa năm thứ 4 theo giới tính (n = 150)

Kết quả kiểm tra giả định phân phối chuẩn của cân nặng theo giới tính bằng kiểm định Shapiro-Wilk cho thấy cân nặng ở nhóm sinh viên nam và sinh viên nữ đều tuân theo luật phân phối chuẩn với giá trị p tương ứng là 0,6317 và 0,3715. Kiểm định t-test độc lập cho thấy cân nặng trung bình của sinh viên nam cao hơn có

ý nghĩa thống kê so với sinh viên nữ (64,42 kg, KTC 95%: 62,15 - 66,69 so với 53,18 kg, KTC 95%: 52,16 - 54,19; $p < 0,0001$).

Mối liên quan giữa giới tính và tình trạng sống trong ký túc xá

Bảng 1. Tình trạng sinh viên sống trong ký túc xá của trường theo giới tính

Giới	Nội trú	Ngoại trú	Cộng
Nam	40 (66,7) [50]	20 (33,3) [28,6]	60 (100)
Nữ	40 (44,4) [50]	50 (55,6) [71,4]	90 (100)
Cộng	80 (100)	70(100)	150 (100)

Kết quả kiểm định χ^2 độc lập cho thấy có mối liên quan có ý nghĩa thống kê giữa giới tính và tình trạng sống nội trú trong ký túc xá của sinh viên ($p = 0,0075$). Sinh viên nam có khả năng sống nội trú cao gấp 2,5 lần so với sinh viên nữ (OR = 2,50; KTC 95%: 1,27 – 4,93).

IV. BÀN LUẬN

Nghiên cứu đã xây dựng thành công bộ dữ liệu mô phỏng có kiểm soát tham số. Đây có thể là một giải pháp khả thi, nhanh chóng, ít tốn kém và có thể áp dụng trong giảng dạy thống kê y sinh cơ bản. Bộ dữ liệu mô phỏng cho phép minh họa rõ ràng lý thuyết sai số đo lường, thực hành đầy đủ quy trình thống kê mô tả và suy luận, đồng thời bảo đảm tính minh bạch và khả năng tái lập phương pháp và kết quả nghiên cứu.

Trước hết, việc mô phỏng bốn kịch bản sai số thông qua biểu đồ hình bia đã trực quan hóa rõ ràng sự khác biệt giữa độ đúng (trueness)

và độ chụm (precision) (Hình 1). Khi thay đổi độc lập sai số hệ thống (bias) và độ lệch chuẩn của sai số ngẫu nhiên, dữ liệu thể hiện rõ các tình huống “đúng cao - chụm cao”, “đúng cao - chụm thấp”, “đúng thấp - chụm cao” và “đúng thấp - chụm thấp” (Hình 1). Cách tiếp cận này giúp người học tiếp cận một cách trực quan lý thuyết sai số [8].

Biểu đồ hình hộp biểu diễn trực quan phân bố cân nặng của nhóm sinh viên gồm các chỉ số trung vị, khoảng tứ phân vị, giá trị nhỏ nhất, giá trị lớn nhất và giá trị ngoại lệ (Hình 2). Biểu đồ hình cột (Hình 3) phân bố cân nặng của 150 sinh viên cho thấy phân bố lệch phải nhẹ, phân bố không tuân theo phân phối chuẩn khi xét toàn bộ mẫu (Shapiro - Wilk $p = 0,0013$, Hình 3), nhưng phân phối chuẩn trong từng nhóm giới tính (Hình 4). Kết quả này do cấu trúc quần thể (ở đây là sự khác biệt về cân nặng theo giới tính) ảnh hưởng đến hình dạng phân phối tổng thể. Việc nhận diện giá trị ngoại lệ và đánh giá phân phối dữ liệu là bước cần thiết trước khi tiến hành phân tích suy luận (Hình 4

và Bảng 1). Kiểm tra phân phối dữ liệu trước khi thiết lập đường cơ bản hoặc ngưỡng cảnh báo (ví dụ trong vùng giá trị trung bình $\pm 2SD$, thường không bao gồm giai đoạn dịch bùng phát) và phát hiện giá trị bất thường khi dịch bùng phát có ý nghĩa quan trọng trong giám sát dịch tễ học [9].

Bộ dữ liệu mô phỏng cho phép thực hành đầy đủ quy trình từ kiểm tra giả định đến lựa chọn kiểm định thống kê phù hợp và diễn giải kết quả. Thực hiện kiểm định t-test là lựa chọn phù hợp sau khi kiểm tra giả định phân phối chuẩn trong từng nhóm bằng kiểm định Shapiro-Wilk (Hình 4). Đồng thời, phân tích mối liên quan giữa giới tính và tình trạng sống nội trú ($OR = 2,50$; $KTC\ 95\%: 1,27 - 4,93$; $p = 0,0075$) (Bảng 1). Giá trị p cho biết mối liên quan quan sát được có đạt mức ý nghĩa thống kê hay không, hay nói một cách khác là ít có khả năng xuất hiện do ngẫu nhiên. Bên cạnh đó, $KTC\ 95\%$ của OR cung cấp nhiều thông tin hơn giá trị p đơn thuần, nó cho phép đánh giá đồng thời độ lớn, chiều hướng kết hợp và độ chính xác của ước lượng. Vì vậy, trong trình bày kết quả, cần ưu tiên báo cáo OR kèm khoảng tin cậy 95% , trong khi p-value nên được xem là thông tin bổ sung [10].

Một ưu điểm nổi bật của dữ liệu mô phỏng trong nghiên cứu là tính minh bạch và khả năng tái lập thông qua mã nguồn mở và hạt giống ngẫu nhiên [11]. Việc sử dụng ngôn ngữ Python cho phép kiểm soát hoàn toàn trung bình, độ lệch chuẩn và sự kết hợp giữa các biến, đồng thời dễ dàng điều chỉnh mức độ phức tạp của tình huống thống kê. Dữ liệu thực tế thường bị hạn chế chia sẻ bởi yêu cầu bảo mật và quyền riêng tư, tốn kém thời gian và nguồn lực hơn trong thu thập. Dữ liệu mô phỏng có thể được chia sẻ công khai trên nền tảng nguồn mở (GitHub), thuận lợi cho chuẩn hóa học liệu và triển khai đào tạo diện rộng. Bộ dữ liệu mô phỏng không chỉ có giá trị trong giảng dạy trên lớp mà còn có tiềm năng ứng dụng trong đào tạo thực hành dịch tễ học và y tế công cộng

thông qua việc tái hiện các tình huống gần với bối cảnh thực địa. Cách tiếp cận này giúp người học chuyển dần từ lý thuyết sang thực hành phân tích số liệu cộng đồng một cách an toàn, có hệ thống và phù hợp với yêu cầu bảo mật dữ liệu thực tế. Đây có thể là giải pháp bền vững góp phần hỗ trợ giảng dạy y tế công cộng. Nghiên cứu phù hợp với xu hướng ứng dụng dữ liệu mô phỏng trong đào tạo và nghiên cứu y học, đặc biệt trong bối cảnh chuyển đổi số và tăng cường năng lực phân tích dữ liệu tại Việt Nam [12, 13].

Nghiên cứu của chúng tôi có một số hạn chế. Hạn chế đầu tiên của phương pháp này đòi hỏi giảng viên và người học có kiến thức cơ bản về ngôn ngữ Python, trình độ phân tích và thống kê cơ bản. Hạn chế thứ hai là dữ liệu mô phỏng mới dừng ở so sánh giá trị trung bình và một mối liên quan nhị phân đơn giản, chưa phản ánh đầy đủ tính phức tạp của dữ liệu dịch tễ học thực tế. Phạm vi của nghiên cứu dừng lại ở việc tạo ra bộ dữ liệu mô phỏng đơn giản và thử nghiệm sử dụng trong phân tích cơ bản.

V. KẾT LUẬN

Nghiên cứu đã xây dựng thành công bộ dữ liệu mô phỏng giúp minh họa rõ ràng lý thuyết sai số đo lường và đặc điểm phân phối không chuẩn do cấu trúc quần thể, cho phép thực hành thống kê mô tả và suy luận cơ bản. Cách tiếp cận dựa trên dữ liệu mô phỏng là một giải pháp khả thi, linh hoạt và khả năng tái lập cao. Dữ liệu mô phỏng có thể ứng dụng trong đào tạo y tế công cộng và y học dự phòng.

TÀI LIỆU THAM KHẢO

1. Sanaeifar E, Houshmand E, Moghri J, et al. Requirements for evidence-based management competency in healthcare: a scoping review. *Front Public Heal*. 2025; 13: 1490454. doi:<https://doi.org/10.3389/fpubh.2025.1490454>.
2. Oster RA, Enders FT. The Importance of Statistical Competencies for Medical Research

- Learners. J Stat Educ. 2018; 26 (2): 137 - 142. doi:<https://doi.org/10.1080/10691898.2018.1484674>.
3. Mosquera L, El EK, Ding L, et al. A method for generating synthetic longitudinal health data. BMC Med Res Methodol. 2023; 23 (1): 67. doi:<https://doi.org/10.1186/s12874-023-01869-w>.
4. Maurer K, Lock D. Comparison of Learning Outcomes for Simulation-based and Traditional Inference Curricula in a Designed Educational Experiment. Technol Innov Stat Educ. 2016; 9 (1). doi:<https://doi.org/10.5070/T591026161>.
5. ELEK, Mosquera L, Fang X, et al. An evaluation of the replicability of analyses using synthetic health data. Sci Rep. 2024; 14 (1): 6978. doi:<https://doi.org/10.1038/s41598-024-57207-7>.
6. Tsao SF, Sharma K, Noor H, et al. Health Synthetic Data to Enable Health Learning System and Innovation: A Scoping Review. Stud Health Technol Inform. 2023; 302: 53 - 57. doi:<https://doi.org/10.3233/shti230063>.
7. Trần Văn Thiện, Lương Thị Phương Trang, Đỗ Thị Thu Cúc, cộng sự. Chất lượng giấc ngủ của sinh viên Trường Đại học Y Dược - Đại học Quốc gia Hà Nội học tập tại cơ sở Hòa Lạc năm học 2022 - 2023. Tạp chí Y học Cộng đồng. 2024; 65 (4): 177 - 185. doi:<https://doi.org/10.52163/yhc.v65i4.1213>.
8. Aschengrau A, Seage GR. Essentials of Epidemiology in Public Health. Jones Bartlett Publ. 2020; 4: 267 - 315.
9. Rotejanaprasert C, Areechokchai D, Maude RJ. Two-step spatiotemporal anomaly detection corrected for lag reporting time with application to real-time dengue surveillance in Thailand. BMC Med Res Methodol. 2024; 24 (1): 10. doi:<https://doi.org/10.1186/s12874-024-02141-5>.
10. Boscardin CK, Sewell JL, Tolsgaard MG, et al. How to Use and Report on p-values. Perspect Med Educ. 2024; 13 (1): 250 - 254. doi:<https://doi.org/10.5334/pme.1324>.
11. Mathur MB, Fox MP. Toward Open and Reproducible Epidemiology. Am J Epidemiol. 2023; 192 (4): 658 - 664. doi:<https://doi.org/10.1093/aje/kwad007>.
12. Cox V, O'driscoll M, Imai N, et al. Estimating dengue transmission intensity from serological data: A comparative analysis using mixture and catalytic models. PLoS Negl Trop Dis. 2022; 16 (7): e0010592. doi:<https://doi.org/10.1371/journal.pntd.0010592>.
13. Olson Z, Staples JA, Mock C, et al. Helmet regulation in Vietnam: impact on health, equity and medical impoverishment. Inj Prev. 2016; 22 (4): 233 - 238. doi:<https://doi.org/10.1136/injuryprev-2015-041650>.

BUILD AND USE SIMULATED DATASET WITH PYTHON PROGRAMMING LANGUAGE FOR TEACHING DESCRIPTIVE STATISTICS

Nguyen Hai Tuan¹, Phung Chi Thien²

¹National Institute of Hygiene and Epidemiology

²Center for Clinical Simulation, Hai Phong University of Medicine and Pharmacy

This study was conducted to build and use a simulated dataset in training data analytic skills for students and health staff. A simulated dataset containing gender, weight and the residency status of 150 medical students was generated using the Python programming language with controlled parameters. The results demonstrate that the simulation method provides a vivid visualization of error theory through bull's eye plots, while simultaneously allowing for the complete practice of descriptive statistics procedures (mean, median, mode, and

interquartile range) and hypothesis testing (Chi-square, t-test, Odds Ratio). The use of simulated data helps overcome barriers related to the privacy of real-world data and creates tailored teaching scenarios, contributing to the enhancement of data analysis capacities for surveillance and response within the preventive medicine network.

Keywords: Simulated dataset; Python programming language; descriptive statistics; teaching statistics