TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 17, Số 8 (2020): 1382-1397
ISSN: 1859-3100
HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 17, No. 8 (2020): 1382-1397
Website: *http://journal.hcmue.edu.vn*

Bài báo nghiên cứu

THAM SỐ ĐO ĐỘ PHÂN TÁN TRONG THỐNG KÊ: KIẾN THỨC CỦA SINH VIÊN SƯ PHẠM TOÁN VÀ VẤN ĐỀ ĐẶT RA CHO CÔNG TÁC ĐÀO TẠO GIÁO VIÊN

Lê Thị Hoài Châu
Trường Đại học Văn Hiến, Việt Nam
Tác giả liên hệ: Lê Thị Hoài Châu – Email: chaulth@vhu.edu.vn
Ngày nhận bài: 11-3-2020; ngày nhận bài sửa: 31-3-2020; ngày duyệt đăng: 24-8-2020

TÓM TẮT

Nghiên cứu này hướng đến việc tìm hiểu kiến thức của giáo viên Toán tương lai về các tham số đo độ phân tán. Hai mươi lăm sinh viên sư phạm ngành Toán đã được đặt trước những tình huống đòi hỏi phải nắm nghĩa của loại tham số này. Các tình huống đưa ra cho sinh viên được thiết kế trên cơ sở một số công trình nghiên cứu khó khăn của người học trong việc hiểu và sử dụng tham số đo độ phân tán. Ứng xử của sinh viên cho thấy việc dạy học thống kê chú trọng vào áp dụng máy móc kĩ thuật tính toán đã khiến người học không nắm nghĩa của các tham số và không làm chủ ngôn ngữ thống kê, ở đây là biểu đồ. Kết quả thu được từ nghiên cứu của chúng tôi là điểm tựa cho việc nhìn lại chương trình đào tạo giáo viên Toán về dạy học Thống kê.

Từ khóa: kiến thức của giáo viên; tham số đo độ phân tán; độ lệch tuyệt đối trung bình; độ lệch chuẩn

1. Đặt vấn đề

1.1. Sự cần thiết của đào tạo về Thống kê

Lí thuyết Xác suất – Thống kê (XS-TK) không chỉ dành cho các nhà toán học. Đây là lĩnh vực khoa học quan trọng, tác động vào nhiều mặt của cuộc sống, nhiều hoạt động của mỗi công dân. Thế giới công nghiệp, kinh tế, hay bảo hiểm đều lệ thuộc nhiều vào các luật xác suất (XS). Vật lí về bản chất là XS trong tự nhiên. Nền tảng của sinh học, di truyền học và y học cũng thế. Ngay cả tính thoả đáng của nhiều quyết sách xã hội cũng phải được xem xét dựa vào các kiến thức về XS-TK. Tuy nhiên, tác động của XS không lộ rõ trước mắt mọi người, mà nó thường ẩn phía sau các dữ liệu thống kê (TK), là cái hiện diện ở mọi lĩnh vực của xã hội: kinh tế, giáo dục, an toàn thực phẩm, y tế, môi trường… Trong thời đại mà công nghệ ngày càng trở nên quan trọng và thông tin đến từ khắp nơi trên thế giới, việc sử dụng dữ liệu TK đang phát triển nhanh chóng. Mỗi công dân cần phải biết đưa ra những quyết định hay chính kiến xác đáng trước nguồn dữ liệu khổng lồ được cung cấp hàng ngày qua các phương tiện truyền thông.

Điều đó cho thấy sự cần thiết phải đưa những kiến thức cơ bản về TK vào chương trình giảng dạy ngay từ bậc phổ thông. Từ hơn nửa thế kỉ trước, nhiều nước có nền giáo dục tiên tiến đã ý thức được sự cần thiết này. Chẳng hạn, vào thời kì đó ở Pháp người ta đã nhận thấy:

Cite this article as: Le Thi Hoai Chau (2020). Dispersal parameter in statistics: Knowledge of mathematics student teachers and some issues for teacher education. *Ho Chi Minh City University of Education Journal of Science, 17*(8), 1382-1397.

> Việc không được đào tạo về Thống kê ở các trường trung học và nhiều ngành của giáo dục đại học dẫn đến những thái độ xã hội lệch lạc. (...) Dù kết quả thống kê được cung cấp hằng ngày qua các phương tiện truyền thông đại chúng, người đọc và người nghe cũng không đủ kiến thức để phân tích một cách thoả đáng. (...) Sự bất lực này càng đáng lo ngại hơn khi mà Thống kê, giống như mọi khoa học khác, đang rất phát triển. Người dùng, khách hàng và công dân cần phải chế ngự thông tin, và do đó phải biết các quy tắc, các xu hướng giải thích có thể có. (…) Nhưng điều đó rất ít khi được thực hiện. Không nghi ngờ gì cả, sự yếu kém của đào tạo về Thống kê ở Pháp là một trở ngại lớn cho vấn đề phát triển kinh tế và thực thi quyền công dân.
>
> (Régnier, J-C., 2012, p.22)

Năm 1959, một Hội thảo của Tổ chức Hợp tác Kinh tế châu Âu (Organisation Européenne de Coopération Économique) tiến hành tại Pháp, được dành riêng để luận bàn về vấn đề dạy học Toán, đã ủng hộ quan điểm đưa TK vào bậc trung học, và do đó cũng phải đưa nó vào chương trình đào tạo giáo viên:

> Thống kê – một nhánh của toán học ứng dụng, là phần chủ yếu của quá trình ra quyết định theo tinh thần của "phương pháp khoa học", và do đó việc sử dụng nó đang được gia tăng trong nhiều lĩnh vực (…) cũng như trong khoa học về hành vi của con người. Hơn nữa, phải thừa nhận rằng lập luận thống kê đang ngày càng trở nên quan trọng trong các hoạt động cộng đồng. Những kiến thức cơ bản về các tính toán xác suất và thống kê nên là một phần của chương trình giáo dục trung học mới, và các bài giảng chuẩn bị cho những môn học này nên được đưa vào chương trình giảng dạy (...) của các cơ sở đào tạo giáo viên.
>
> (Trích theo Régnier J-C., 2012, p.22)

Ở Việt Nam, thì phải đợi đến cuộc cải cách giáo dục thực hiện theo hình thức cuốn chiếu kéo dài 12 năm, bắt đầu ở lớp 1 từ năm học 1980-1981, những kiến thức ban đầu về XS-TK mới được đưa vào một cách đáng kể và tương đối có hệ thống trong môn Toán dạy ở bậc trung học. Điều này thể hiện ý muốn hội nhập với giáo dục thế giới theo quan điểm tăng cường tính ứng dụng thực tiễn của toán học dạy trong nhà trường. Tuy nhiên, trên thực tế thì chủ đề TK thường không được chú trọng đúng mức trong thực hành dạy học của giáo viên Toán ở trung học phổ thông. Giải thích hiện tượng này, nhiều giáo viên cho rằng lí do là TK chưa bao giờ xuất hiện trong các đề thi cuối cấp trung học và tuyển sinh đại học.

1.2. Nhu cầu nhìn lại công tác đào tạo giáo viên Toán

Ở vị trí của người tham gia công tác đào tạo, chúng tôi tự hỏi: Liệu sự vắng mặt trong các đề thi có phải là lí do duy nhất gây nên hiện tượng coi nhẹ dạy học TK? Hay thực ra hiện tượng ấy vừa là nguyên nhân, vừa là hệ quả của sự yếu kém về đào tạo TK trong nhà trường. Là nguyên nhân, hiện tượng coi nhẹ dạy học TK tạo nên "một trở ngại lớn (…) cho việc thực thi quyền công dân" như ghi nhận mà các nhà giáo dục Pháp đã bày tỏ cách đây từ sáu thập kỉ. Là hệ quả, vì việc không có thói quen vận dụng TK vào đời sống công dân lại dẫn đến chỗ không coi trọng đúng mức tầm quan trọng của khoa học này trong đào tạo ở bậc phổ thông cũng như đại học. Đó là một vòng luẩn quẩn, mà theo chúng tôi, muốn thoát khỏi thì trước hết, ngoài việc tác động vào quan điểm của các nhà quản lí, lập chương trình, tác giả sách giáo khoa, chúng ta không thể không bắt đầu từ công tác đào tạo giáo viên.

Nghiên cứu trình bày ở bài báo này nằm trong bối cảnh Việt Nam chuẩn bị triển khai chương trình giáo dục phổ thông mới do Bộ Giáo dục và Đào tạo công bố cuối năm 2018. Một trong những mục tiêu của chương trình đó là giúp cho học sinh (HS) "có đủ năng lực tối thiểu để tự tìm hiểu những vấn đề liên quan đến toán học trong suốt cuộc đời" (Ministry of Education and Training, 2018, p.6). Nội dung dạy học cốt lõi được xây dựng "xoay quanh

ba mạch kiến thức: Số, Đại số và Một số yếu tố giải tích; Hình học và Đo lường; Thống kê và Xác suất" (Ministry of Education and Training, 2018, p.16). Khác với các chương trình trước, mạch kiến thức thứ ba được coi trọng. Đối với phần TK, chương trình 2018 xác định:

> Thống kê (…) là một thành phần bắt buộc của giáo dục toán học trong nhà trường, góp phần tăng cường tính ứng dụng và giá trị thiết thực của giáo dục toán học. Thống kê (…) tạo cho học sinh khả năng nhận thức và phân tích các thông tin được thể hiện dưới nhiều hình thức khác nhau, (…) hình thành sự hiểu biết về vai trò của Thống kê như là một nguồn thông tin quan trọng về mặt xã hội, biết áp dụng tư duy thống kê để phân tích dữ liệu. Từ đó, nâng cao sự hiểu biết và phương pháp nghiên cứu thế giới hiện đại cho học sinh.
>
> (Ministry of Education and Training, 2018, p.16).

Liệu giáo viên Toán có đáp ứng được những đòi hỏi của chương trình 2018?

1.3. Tri thức được lựa chọn cho nghiên cứu: Tham số đo độ phân tán của dãy dữ liệu

Nhằm tìm câu trả lời, chúng tôi chọn *tham số đo độ phân tán của dãy dữ liệu,* để nghiên cứu kiến thức có ở sinh viên (SV) sư phạm ngành Toán – những giáo viên tương lai. Trong phần dưới, để ngắn gọn chúng tôi sẽ gọi *tham số phân tán* thay cho *tham số đo độ phân tán của dãy dữ liệu*.

Có hai lí do khiến chúng tôi đưa ra sự lựa chọn này.

Lí do thứ nhất là ghi nhận của chúng tôi về xu hướng dành sự chú ý cho các tham số đo độ tập trung – còn gọi là đo xu hướng hội tụ (đặc biệt là số trung bình (*mean*), mốt (*mode*)) trong thực hành xã hội (ví dụ như người ta thường nói về tuổi thọ trung bình, thu nhập trung bình, năng suất trung bình, điểm trung bình, loại xe được ưa chuộng nhất…). So với số trung bình nói riêng, tham số đo xu hướng hội tụ nói chung, thì các tham số phân tán của dãy dữ liệu dường như ít được sử dụng trong những phân tích TK thường gặp trên các phương tiện truyền thông đại chúng.

Thế nhưng, thực ra thì giữa xu hướng hội tụ với độ phân tán của dãy dữ liệu có mối liên hệ khăng khít. Việc sử dụng riêng rẽ một tham số hội tụ nhiều khi chẳng nói lên được điều gì chính xác về dãy dữ liệu. Chẳng hạn, số trung bình san bằng mọi sự chênh lệch về các giá trị của dữ liệu, không cho biết dãy dữ liệu phân tán hay tập trung quanh nó thế nào. Vì thế, thiếu phân tích về độ phân tán thì người ta không đủ cơ sở để khẳng định số trung bình có là thước đo thoả đáng hay không cho xu hướng tập trung của dãy dữ liệu được xem xét. Một cách tổng quát, các tham số đo xu hướng hội tụ chỉ thực sự có ích khi nó được giải thích trong mối quan hệ với độ phân tán của dãy dữ liệu. Tương tự, việc phân tích độ phân tán của dãy dữ liệu cũng không thể tách rời khỏi vấn đề xem xét xu hướng tập trung, như những gì toát lên từ công thức tính các tham số phân tán mà chúng tôi sẽ chỉ ra ở dưới.

Như vậy, để mô tả một hiện tượng quan sát được thì nghiên cứu độ phân tán của phân phối dữ liệu cũng quan trọng như việc xem xét xu hướng hội tụ của nó. Vì thế mà ngày nay hai trong số các nội dung chính của chương trình TK giảng dạy ở nhà trường là phân tích xu hướng tập trung và độ phân tán của dãy dữ liệu. Tuy thế, chúng ta cũng không cần quên rằng thực ra thì tầm quan trọng của tham số phân tán mới được thừa nhận cách đây không lâu trong các chương trình áp dụng ở bậc phổ thông. Trước thế kỉ XXI, nghiên cứu về dạy học TK đã tập trung rất nhiều vào vấn đề quan sát xu hướng hội tụ của dãy dữ liệu (theo Reading, & Shaughnessy, 2004).

Lí do thứ hai là ghi nhận của chúng tôi qua quan sát thực hành dạy học TK ở trường phổ thông, theo đó thì mục đích nhắm đến là vận dụng công thức để tính giá trị các tham số

(số trung bình, phương sai (*variance*), độ lệch chuẩn (*standard deviation*)…). Thực ra, nói về hiện tượng này thì Việt Nam không phải là trường hợp đặc biệt. Nhiều công trình, chẳng hạn như của Bakker (2004), Watson (2007) đã cho thấy là dạy học khái niệm số trung bình chủ yếu chú trọng vào kĩ thuật tính toán. Bàn về tác động của xu hướng dạy học ấy, Gattuso (1997) đã lưu ý rằng nó không đảm bảo việc hiểu rõ khái niệm đối với HS.

Cũng giống như số trung bình, xu hướng dạy học chú trọng vào áp dụng máy móc thuật toán để tính giá trị các tham số phân tán không đảm bảo việc hiểu và khả năng sử dụng chúng trong phân tích TK. Điều này đã được Makar và Confrey (2005) chỉ ra trong nghiên cứu của mình, theo đó thì một số giáo viên toán có thể đề cập khái niệm độ lệch chuẩn thông qua kĩ thuật tính nó, nhưng không thể giải thích ý nghĩa cho kết quả thu được. Thế nhưng, hiểu một kiến thức toán học không chỉ là biết *làm như thế nào*, mà còn phải trả lời được câu hỏi *tại sao – tại sao lại cần đến nó? tại sao lại làm như vậy?* Để nhấn mạnh việc hiểu nghĩa của các khái niệm, Boyé và Comairas (2002), cũng đã viết: "dạy học Thống kê không thể chỉ quy về việc học các công thức và áp dụng chúng" (p. 37). Điều đó lại càng đúng khi mà sự phát triển của công nghệ ngày nay đã giải phóng cho con người khỏi việc ghi nhớ công thức và thực hiện các kĩ thuật tính toán.

Nếu muốn thúc đẩy sự phát triển tư duy TK ở HS thì việc làm cho họ hiểu nghĩa của tham số phân tán dường như không thể bỏ qua. Do đó, cần phải tìm hiểu kiến thức của giáo viên về chủ đề này, bởi chính họ là người tổ chức dạy học bằng cách xây dựng những tình huống tạo thuận lợi cho việc học.

1.4. Phương pháp luận nghiên cứu

Kiến thức của giáo viên liên quan đến hai phương diện: toán học và sư phạm. Về toán học, trước hết chúng tôi sẽ làm rõ các đặc trưng của tham số phân tán. Dựa vào đó chúng tôi sẽ quan sát xem các giáo viên tương lai đạt được những gì. Về sư phạm: chúng tôi muốn tìm hiểu ứng xử của họ trước những sai lầm của HS. Cụ thể hơn, từ việc xác định rõ đặc trưng của tham số phân tán và kế thừa những nghiên cứu đã có về khó khăn trong thực hành dạy học, chúng tôi sẽ xây dựng vài tình huống cho phép làm bộc lộ kiến thức toán học và sư phạm của SV về loại tham số này. Các tình huống được thiết kế trên cơ sở một số công trình nghiên cứu khó khăn của người học trong việc hiểu và sử dụng tham số phân tán. Một phân tích chương trình đào tạo giáo viên sẽ được thực hiện sau đó, nhằm giải thích sự khiếm khuyết trong kiến thức toán học và ứng xử sư phạm mà chúng tôi quan sát được ở SV.

2. Về các tham số đo độ phân tán của dãy dữ liệu

Đặc trưng biến thiên của một biến TK được đánh giá chủ yếu qua các tham số đo độ phân tán của dãy giá trị. Các tham số này "cho phép mô tả tập hợp dữ liệu liên quan đến một biến cụ thể, thông qua việc cung cấp một dấu hiệu về sự biến thiên của các giá trị trong tập hợp" (Dodge, 1993, tr. 225). Cụ thể hơn, chúng cho biết dãy dữ liệu được phân bổ ra sao xung quanh các giá trị trung tâm.

Để đo độ phân tán của phân phối dữ liệu, tham số đầu tiên có thể nghĩ đến là *biên độ* (range), còn gọi là *khoảng biến thiên*. Lí do là sự đơn giản trong tính toán biên độ (hiệu giữa giá trị lớn nhất với giá trị nhỏ nhất của dãy dữ liệu) và cả ở sự đơn giản trong giải thích (khoảng bé nhất chứa tất cả các giá trị của dãy). Tuy nhiên, nếu được sử dụng một mình thì biên độ chỉ là một phương tiện rất hạn chế, không đủ đại diện cho mức độ biến thiên của dữ liệu, vì nó không tính đến các giá trị nằm giữa của biến TK và ảnh hưởng của tần số, tần suất

mỗi giá trị. Trong trường hợp giá trị lớn nhất hay nhỏ nhất có tính "ngoại lai" (quá cách xa số trung bình và gần nó có rất ít giá trị khác) thì khoảng biến thiên lại càng không mang lại mấy thông tin về phân phối dữ liệu. Thậm chí, do các giá trị tiêu biểu của dữ liệu không được tính đến nên biên độ có thể làm méo mó hình ảnh về phân phối.

Nhằm hạn chế nhược điểm của biên độ, người ta tách bỏ những giá trị ở gần hai cực của phân phối, sau khi chia dữ liệu (đã được sắp xếp theo thứ tự tăng dần) thành những lớp có tần số bằng nhau (hay gần như bằng nhau). Phương pháp này dẫn đến khái niệm *phân vị* (quantiles). Các phân vị thường dùng là *tứ phân vị* (quartiles), *thập phân vị* (deciles), *bách phân vị* (percentiles), ứng với việc chia dữ liệu thành 4, 10 hay 100 lớp. Chẳng hạn, với một tứ phân vị, người ta chia dữ liệu thành 4 lớp có tần suất bằng nhau, rồi bỏ đi 25% giá trị bé nhất (thuộc lớp đầu tiên) và 25% giá trị lớn nhất (thuộc lớp thứ 4), chỉ xét độ phân tán của 50% dữ liệu còn lại (bằng cách xét biên độ của 50% dữ liệu đó). Các khoảng phân vị cho phép đo độ phân tán của dữ liệu quanh số trung vị. Phương pháp này chỉ hạn chế chứ không loại bỏ hoàn toàn được yếu điểm của việc dùng biên độ.

Sự ra đời của *độ lệch tuyệt đối trung bình* (mean absolute deviation) – nhiều khi được gọi đơn giản là *độ lệch trung bình* (mean deviation) và *độ lệch chuẩn* chính là để giải quyết điểm yếu của biên độ. Hai tham số này mang trong chúng những thông tin về sự biến thiên các giá trị của biến TK, có tính đến ***tất cả*** các dữ liệu (không chỉ hai giá trị lớn nhất, nhỏ nhất) và cho phép nhận ra độ phân tán của chúng so với các tham số trung tâm của phân phối. Giải thích cho nhận định này, ta chỉ cần nhìn công thức tính độ lệch tuyệt đối trung bình và độ lệch chuẩn nêu ở dưới đây.

Xét dãy dữ liệu về một biến TK có k giá trị $x_1, x_2, \ldots, x_k$, trong đó tần số xuất hiện giá trị x_i là n_i ($i = \overline{1,k}$). Đặt $N = \sum_{i=1}^{k} n_i$. Giả sử $\overline{x}$ là số trung bình của dãy dữ liệu. Khi đó:

Độ lệch tuyệt đối trung bình là: $\frac{1}{N} \sum_{i=1}^{k} n_i |x_i - \overline{x}|$ (1)

Phương sai là: $s^2 = \frac{1}{N} \sum_{i=1}^{k} n_i \, (x_i - \overline{x})^2$ (2)

Độ lệch chuẩn là: $s = \sqrt{s^2} = \sqrt{\frac{1}{N} \sum_{i=1}^{k} n_i \, (x_i - \overline{x})^2}$ (3)

Nói một cách chính xác thì công thức (1) cho phép tính độ lệch tuyệt đối trung bình so với số trung bình. Người ta cũng có thể tính độ lệch tuyệt đối trung bình so với trung vị (gọi là độ lệch tuyệt đối trung vị) M_e bằng công thức tương tự:

$\frac{1}{N} \sum_{i=1}^{k} n_i |x_i - M_e|$ (1')

Đối với trường hợp biến có quá nhiều giá trị và liên tục thì người ta ghép chúng theo từng lớp. Nếu một dãy dữ liệu được ghép thành các lớp $[a_1,b_1)$, $[a_2,b_2)$, … $[a_k,b_k)$ trong đó lớp $[a_i,b_i)$ có tần số n_i ($i = \overline{1,k}$) thì công thức tính độ lệch tuyệt đối trung bình, phương sai, độ lệch chuẩn cũng giống như trên, nhưng x_i được thay bởi c_i - trung bình cộng của a_i và b_i.

Theo bốn công thức trên, độ phân tán của dãy dữ liệu được đánh giá dựa trên độ lệch (hiệu) giữa mỗi giá trị TK so với một tham số hội tụ (như số trung bình, trung vị) và ***tất cả các giá trị cùng với tần số (tần suất) của nó*** đều được tính đến. Độ lệch tuyệt đối trung bình được tính qua trung bình của các giá trị tuyệt đối của độ lệch so với số trung bình. Việc dùng giá trị tuyệt đối là cần thiết, vì các độ lệch âm cân bằng các độ lệch dương và do đó tổng các độ lệch so với số trung bình luôn bằng 0. Ta có thể xem là công thức tính độ lệch tuyệt đối trung bình nêu trên phản ánh một cách tự nhiên ý tưởng về khoảng cách qua giá trị tuyệt đối.

Như vậy, độ lệch tuyệt đối trung bình là số đo mang lại một dấu hiệu rõ ràng về sự phân tán của phân phối dữ liệu. Tuy nhiên, việc phải dùng trị tuyệt đối để tránh các độ lệch âm lại gây nên những khó khăn đối với nhiều xử lí dữ liệu TK. Trong khi đó, để tính độ lệch chuẩn, chỉ cần lấy căn bậc hai của phương sai – được định nghĩa là trung bình các bình phương của các độ lệch (so với số trung bình). Cũng như độ lệch tuyệt đối trung bình, tính toán phương sai và độ lệch chuẩn cho phép tránh các độ lệch âm. Nhưng, do dựa trên bình phương các độ lệch, phương sai có sự bất tiện ở chỗ nó không có cùng đơn vị với giá trị của phân phối. Việc lấy căn bậc hai của phương sai nhằm mục đích quay lại với đơn vị ban đầu. Độ lệch chuẩn hay độ lệch tuyệt đối trung bình càng bé nghĩa là phân phối của dãy dữ liệu càng tập trung xung quanh số trung bình (hay trung vị, trong trường hợp độ lệch tuyệt đối trung vị).

Bốn công thức tính nêu trên cho thấy mối liên hệ gắn bó giữa độ lệch tuyệt đối trung bình, độ lệch tuyệt đối trung vị và độ lệch chuẩn với hai tham số hội tụ (số trung bình, trung vị) mà người ta không thể bỏ qua trong các phân tích TK. Lúc này, những cặp tham số có thể chọn để phân tích các phân phối dữ liệu là độ lệch chuẩn và số trung bình, độ lệch tuyệt đối trung bình và số trung bình, độ lệch tuyệt đối trung vị và trung vị[2].

Biên độ, phân vị, độ lệch tuyệt đối trung bình, độ lệch chuẩn thuộc nhóm tham số phân tán tuyệt đối, chỉ có thể sử dụng để xem xét các biến TK cùng loại (và do đó gắn với cùng một đơn vị). Trong trường hợp muốn so sánh độ phân tán của các biến khác loại, người ta phải dùng *hệ số biến thiên* (coefficient of variation) – một tham số phân tán tương đối. Hệ số biến thiên được tính qua tỉ số giữa giá trị của một tham số phân tán tuyệt đối với giá trị của tham số hội tụ đi cùng cặp với nó (chẳng hạn như tỉ số giữa độ lệch chuẩn với số trung bình). Cánh tính này khiến hệ số biến thiên không còn gắn với đơn vị và người ta thường viết nó ở dạng phần trăm. Ở bài báo này chúng tôi chỉ tập trung xem xét nhóm tham số phân tán tuyệt đối, đặc biệt là độ lệch tuyệt đối trung bình và độ lệch chuẩn.

Bốn công thức trên cũng cho thấy là việc hiểu và sử dụng độ lệch chuẩn hay độ lệch tuyệt đối trung bình thâu tóm trong nó nhiều khái niệm TK, như số trung bình, trung vị, độ lệch và mật độ tương đối quanh số trung bình/trung vị. Để hiểu độ lệch chuẩn và độ lệch tuyệt đối trung bình thì cần phải nhận thức được rằng mỗi phân phối dữ liệu kết hợp trong nó tất cả những khái niệm vừa nói. Điều đó giải thích khó khăn mà giáo viên gặp phải trong dạy học.

Khó khăn lại càng tăng lên khi người ta phải làm việc với độ lệch chuẩn và độ lệch tuyệt đối trung bình thông qua các biểu đồ, đồ thị TK. Theo Garfield và Ben-Zvi (2005), việc có thể nhận ra cách thức biểu hiện sự phân tán của dữ liệu trong những biểu đồ TK khác nhau là một yếu tố quan trọng. Nó chứng tỏ sự nắm vững các khái niệm độ lệch chuẩn, độ lệch tuyệt đối trung bình. Thế nhưng biểu đồ lại có thể tạo ra trở ngại cho việc nhận biết đó, vì nó "khuyến khích" sự xuất hiện những quan niệm sai lầm sinh ra từ cái nhìn trực quan. Điều này đặc biệt hay xảy ra khi người ta làm việc với các *histogram*[3]. Nghiên cứu của Delmas và Liu (2005), Meletiou-Mavrotheris và Lee (2005), Cooper và Shore (2010) đã chỉ

[2] Ngoài ra, người ta cũng có thể chọn cặp phân vị và trung vị.

[3] Được gọi là *biểu đồ tổ chức* trong nhiều giáo trình TK tiếng Việt và là *biểu đồ tần số (tần suất) hình cột* trong các sách giáo khoa Đại số lớp 10 hiện hành, ví dụ như Tran Van Hao et al, 2011, (p.115). Trong phần còn lại của bài báo, chúng tôi sẽ dùng từ *histogram* khi không cần phân biệt đối tượng đang nói đến là *biểu đồ tần số* hay *biểu đồ tần suất hình cột.*

ra một số chiến lược sai có thể được sử dụng khi người ta xem xét sự phân tán dữ liệu từ các histogram. Ví dụ, thay vì căn cứ vào mật độ của dữ liệu xung quanh số trung bình thì có một quan niệm sai lầm cho rằng đặc tính biến thiên của dữ liệu thể hiện ở sự thay đổi chiều cao các dải chữ nhật. Quan niệm sai lầm đó dẫn đến kết luận là chiều cao các dải thay đổi càng nhiều thì mức độ biến thiên của dữ liệu càng lớn. Cũng từ đó mà người ta cho rằng độ lệch chuẩn và độ lệch tuyệt đối trung bình sẽ nhỏ nếu dãy dữ liệu ứng với một biểu đồ gồm những dải có chiều cao gần giống nhau.

Như vậy, có những khó khăn trong việc đọc các histogram, giải thích sự phân tán của phân phối dữ liệu trong sự kết hợp với số trung bình. Theo Garfield và Ben-Zvi (2005), những khó khăn này có thể càng tăng lên khi mà trong thực tế dạy học ở trung học thì vấn đề phân tích biểu đồ dường như bị bỏ quên. Dạy học TK ở bậc học này thường chú ý đến cách vẽ các biểu đồ khác nhau chứ không đặt trọng tâm vào nhiệm vụ phát triển một sự hiểu biết đầy đủ về mối quan hệ giữa các khái niệm TK, đặc biệt là giữa trung tâm của phân phối và sự phân tán dữ liệu được biểu diễn trên biểu đồ.

3. Nghiên cứu kiến thức của sinh viên sư phạm về độ lệch tuyệt đối trung bình và độ lệch chuẩn

Những phân tích trên khiến chúng tôi thấy sự cần thiết phải tìm hiểu kiến thức của SV sư phạm về các tham số đo độ phân tán. Trong khuôn khổ nghiên cứu này, chúng tôi giới hạn đối tượng tri thức ở độ lệch tuyệt đối trung bình và độ lệch chuẩn.

Một nghiên cứu thực nghiệm đã được chúng tôi thực hiện với 25 SV cuối năm thứ ba Khoá 41 của Khoa Toán – Tin, Trường Đại học Sư phạm Thành phố Hồ Chí Minh. Vào thời điểm tiến hành thực nghiệm, SV đã trải qua đợt thực tập thứ nhất của chương trình đào tạo, do đó đã có một hiểu biết nhất định về thực tế dạy học ở trung học phổ thông. Họ cũng vừa mới kết thúc tất cả các học phần thuộc khối đào tạo nghề trong chương trình (như *Lí luận dạy học đại cương, Lí luận dạy học Hình học, Lí luận dạy học Đại số và Giải tích, Ứng dụng công nghệ thông tin trong dạy học…*). Họ cũng đã hoàn thành hai học phần dành cho Lí thuyết *XS-TK* được đặt trong khối kiến thức cơ sở ngành.

3.1. Ba bài toán được sử dụng

Chúng tôi thiết kế thực nghiệm trên nền của ba bài toán được giới thiệu dưới đây.

Bài toán 1.

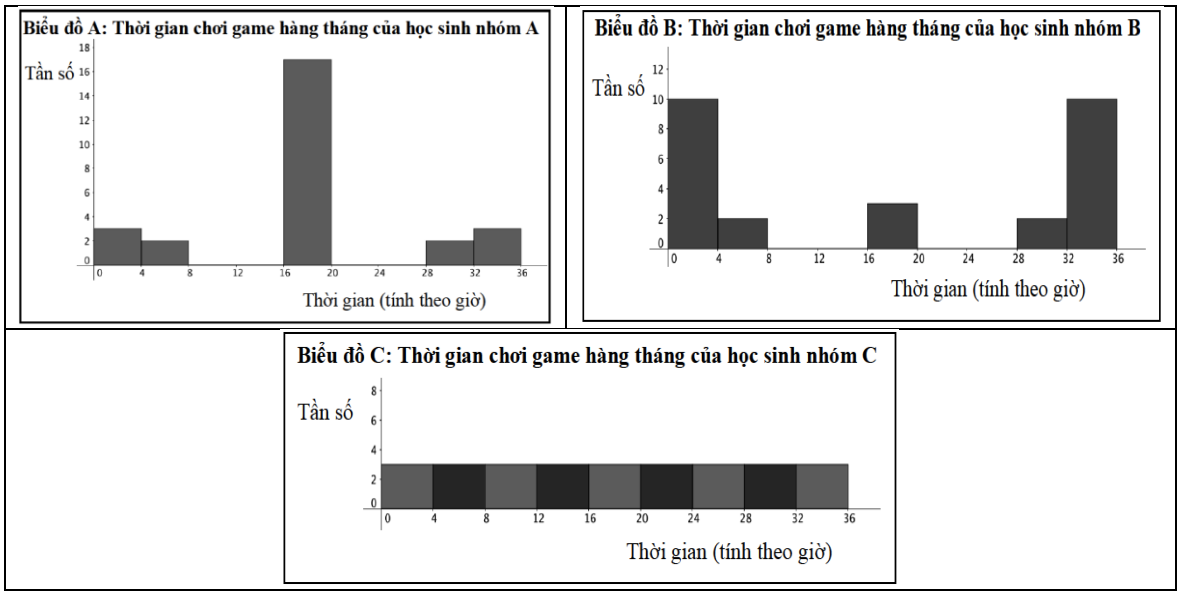
Liam muốn một ngày nào đó trở thành tay đua ô tô và đang tìm kiếm một loại xe đua tốt nhất. Anh tham khảo số liệu thống kê về số km mà mỗi xe đi được trước khi có sự cố bất chợt xảy ra. Đối với hai loại xe đua hiệu Dana và Toyo, Liam tìm thấy kết quả dưới đây:

Số km đã đi trước khi bị xảy ra sự cố của một số xe thuộc loại Dana:
65, 84, 87, 91, 97, 103, 107, 109, 114, 118, 122, 133, 136, 137, 142, 188, 192, 195, 195
Số km đã đi trước khi xảy ra sự cố của một số xe thuộc loại Toyo:
85, 89, 91, 102, 106, 106, 115, 115, 117, 120, 121, 129, 133, 136, 143, 165, 170, 170, 197

Bạn hãy đưa ra lời khuyên cho Liam: loại xe nào đáng tin cậy hơn cho cuộc đua 100 km? Giải thích câu trả lời của bạn.

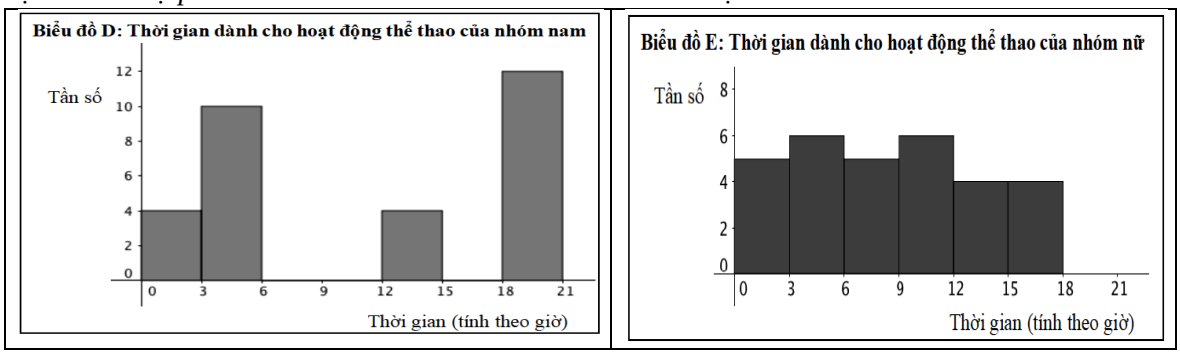
Bài toán 2.

Một giáo viên thu thập thông tin về thời gian chơi game hằng tháng của ba nhóm học sinh lớp 8, mỗi nhóm có 27 em. Số liệu do giáo viên đó thu thập được trình bày qua ba biểu đồ dưới đây. Theo bạn, đối với ba phân phối dữ liệu liên quan đến các nhóm A, B, C, phân phối nào có độ lệch chuẩn lớn nhất? phân phối nào có độ lệch chuẩn nhỏ nhất? Giải thích câu trả lời của bạn.



Bài toán 3.

Một người thu thập dữ liệu về thời gian dành cho hoạt động thể thao hàng tháng của SV khoa Toán – Tin. Người đó phân đối tượng điều tra thành hai nhóm theo giới tính, mỗi nhóm có 30 SV. Số liệu do người đó thu thập được trình bày bằng các biểu đồ D và E dưới đây. Theo bạn, phân phối dữ liệu nào có độ phân tán lớn hơn? Giải thích câu trả lời của bạn.



3.2. Dàn dựng và phân tích các bài toán

Thực nghiệm được chúng tôi chia làm bốn pha.

Pha 1. SV làm việc cá nhân với bài toán 1 trong 20 phút. Họ có thể sử dụng máy tính cầm tay để tính toán nếu cần. Đối với bài toán này, câu hỏi không nhắc đến tham số TK nào. Nhiệm vụ của SV là chọn những tham số mà họ cho là thích hợp để tìm câu trả lời. Chúng tôi muốn tìm hiểu xem liệu họ có huy động các tham số đo độ phân tán của phân phối dữ liệu khi giải quyết vấn đề hay không, và sự huy động đó có được đặt trong mối liên kết với số trung bình, trung vị hay không. Dữ liệu được tóm tắt qua Bảng 1 dưới đây:

Bảng 1. *Các tham số tính được từ hai bảng phân phối dữ liệu*

Loại xe	Số trung bình	Trung vị	Độ lệch tuyệt đối trung bình	Độ lệch tuyệt đối trung vị	Độ lệch chuẩn
Dana	127,10 km	118 km	31,7 km	30,68	38,25
Toyo	126,84 km	120 km	24,03 km	23	29,79

Chúng tôi dự kiến ba loại câu trả lời. Loại thứ nhất chỉ quan tâm đến các tham số đo xu hướng hội tụ (ở đây là số trung bình) của hai dãy dữ liệu và sẽ chọn hãng Dana. Loại thứ

hai chỉ quan tâm đến độ phân tán, cho rằng nên chọn Toyto vì phân phối dữ liệu ứng với nó có độ lệch tuyệt đối trung bình (hoặc độ lệch tuyệt đối trung vị) hay/và độ lệch chuẩn bé hơn. Hai loại câu trả lời này đều thể hiện quan niệm xem xét các tham số hội tụ và phân tán trong sự tách biệt. Loại thứ ba đã tính đến mối quan hệ giữa chúng, chọn Toyo, với lập luận là tuy số trung bình của hai phân phối dữ liệu gần như nhau, nhưng phân phối thứ hai có độ lệch tuyệt đối trung bình bé hơn và hiệu giữa số trung bình với độ lệch tuyệt đối trung bình vẫn lớn hơn 100. Điều này không xảy ra đối với phân phối dữ liệu thứ nhất. Thuộc loại thứ ba, người ta cũng có thể giải thích tương tự cho việc chọn hãng Toyo với một trong các cặp:

- độ lệch chuẩn và số trung bình,
- độ lệch tuyệt đối trung vị và trung vị.

Cách lựa chọn cặp cuối cùng mang lại thuận lợi trong tính toán vì không phải làm việc với số thập phân. Ngoài ra, để so sánh độ phân tán người ta cũng có thể dùng hệ số biến thiên (chẳng hạn như lập tỉ số giữa độ lệch chuẩn và số trung bình). Tuy nhiên, vì câu hỏi không phải là so sánh độ phân tán của hai phân phối dữ liệu, mà là "loại xe nào đáng tin cậy hơn cho cuộc đua 100 km", nên câu trả lời tốt nhất chính là xét hiệu của hai tham số trong cặp số trung bình và độ lệch tuyệt đối trung bình rồi so sánh với 100.

Pha 2. SV nghiên cứu cá nhân để tìm câu trả lời cho Bài toán 2. Thời gian làm việc của họ là 15 phút. Chúng tôi giới hạn thời gian tương đối ngắn, nhằm hạn chế việc sử dụng máy tính cầm tay.

Khác với Bài toán 1, câu hỏi của Bài toán 2 đã nêu tường minh nhiệm vụ tìm hiểu độ lệch chuẩn của các phân phối được biểu diễn bởi biểu đồ tần số ghép lớp. Vấn đề là so sánh độ lệch chuẩn của ba phân phối dữ liệu được cho trên ba biểu đồ. Để giảm bớt sự phức tạp trong nhiệm vụ giao cho SV, chúng tôi không lấy những biểu đồ có các lớp ghép không đều nhau. Như vậy, tần số của mỗi lớp ghép có thể được biểu diễn qua chiều cao hình chữ nhật ứng với lớp đó. Cả ba phân phối dữ liệu cho trên biểu đồ đều có số trung bình là 18. Điều này có thể đọc được dễ dàng với ghi nhận về sự đối xứng của mỗi biểu đồ qua đường thẳng $x = 18$. Việc tạo thuận lợi cho vấn đề xác định số trung bình trên biểu đồ nhằm tìm hiểu xem SV có để ý đến mối quan hệ giữa nó với độ lệch chuẩn hay không. Mức độ tập trung của dữ liệu quanh số trung bình được đánh giá qua tổng chiều cao những hình chữ nhật "đứng gần" hoặc chứa giá trị trung bình.

Câu trả lời chính xác ở đây là: độ lệch chuẩn của dãy dữ liệu có giá trị bé nhất ở biểu đồ *A* và lớn nhất ở biểu đồ *B*. Căn cứ là trên biểu đồ *A* thì các giá trị của biến tập trung phần lớn quanh số trung bình, trong khi đó thì biểu đồ *B* lại có tính chất ngược lại – phần lớn giá trị nằm cách xa số trung bình. Sự tập trung của dữ liệu quanh số trung bình ở biểu đồ *C* ít hơn biểu đồ *A*, nhưng vẫn nhiều hơn biểu đồ *B*, nên độ phân tán của *C* bé hơn của *B* và lớn hơn của *A*. Điều đó có nghĩa là $S_A < S_C < S_B$.

Pha 3. Vẫn với Bài toán 2, nhưng SV làm việc nhóm (trong 15 phút) với câu trả lời sai của hai HS giả định. Tình huống đưa ra cho SV là:

> *Trả lời cho câu hỏi nêu trong Bài toán 2, có hai học sinh lớp 10 đi đến hai kết luận khác nhau cho nhóm C.*
>
> *Học sinh thứ nhất khẳng định rằng dữ liệu liên quan đến nhóm C có độ lệch chuẩn lớn nhất. Em này lập luận là:* "biểu đồ C có nhiều hình chữ nhật nhất, và đó là dấu hiệu cho thấy số liệu thống kê biến thiên nhiều nhất".
>
> *Học sinh thứ hai lại khẳng định là biểu đồ C biểu diễn một phân phối dữ liệu có độ lệch chuẩn*

bé nhất. Lập luận của học sinh này là: "các hình chữ nhật trong biểu đồ C có chiều cao đồng nhất. Điều đó có nghĩa là dãy dữ liệu biến thiên ít nhất".

Theo bạn, ai đúng? Bạn sẽ nói gì với những học sinh này?

Việc lựa chọn hai lập luận sai của HS được hình thành từ các công trình của Cooper và Shore (2010), Delmas và Liu (2005). Các tác giả này đã chỉ ra rằng nhiều HS có quan niệm sai lầm khi phân tích độ phân tán của một phân phối dữ liệu được biểu diễn bằng histogram. Ở đây, HS thứ nhất quan niệm rằng độ phân tán thể hiện ở số hình chữ nhật. Đây không phải là quan niệm đúng, vì việc quan sát số hình chữ nhật không được đặt trong mối liên hệ với tham số hội tụ nào. Số hình chữ nhật nhiều không phải là dấu hiệu chứng tỏ độ phân tán lớn. Nếu theo đuổi logic lập luận này, HS sẽ không thể tìm ra nhóm có độ phân tán nhỏ nhất (bởi hai biểu đồ *A, B* có số hình chữ nhật như nhau). HS thứ hai thì chịu ảnh hưởng của quan niệm cho rằng chiều cao của các hình chữ nhật không thay đổi là dấu hiệu chứng tỏ dãy dữ liệu ít phân tán. Như vậy là HS này đã nhìn vào sự biến thiên của tần số (hay tần suất) chứ không phải sự biến thiên các giá trị của phân phối.

Tình huống đặt ra để tìm hiểu xem những giáo viên tương lai tham gia thực nghiệm sẽ xử trí ra sao trước sai lầm của HS. Lưu ý rằng, như chúng tôi đã nói trên, ứng xử của giáo viên phụ thuộc vào cả kiến thức sư phạm lẫn kiến thức toán học của họ.

Pha 4. SV tiếp tục làm việc nhóm trong 15 phút để giải Bài toán 3.

Với bài toán này, chúng tôi muốn tiếp tục tìm hiểu xem SV có nhận ra được rằng độ phân tán phải được xem xét trong mối tương quan với các tham số hội tụ hay không. Các lớp ghép vẫn có cùng độ dài, nhưng khác với Bài toán 2, hai biểu đồ cho ở đây không đối xứng nên có thể người ta không nhìn thấy ngay số trung bình. Câu hỏi không yêu cầu rõ so sánh tham số nào, chỉ nêu chung chung là độ phân tán. Người trả lời có thể nghĩ đến biên độ (ở đây là khác nhau, không như trường hợp ba biểu đồ *A, B, C*). Nhưng, như chúng tôi đã phân tích ở trên, tham số này không mang lại một hình ảnh chính xác về phân phối dữ liệu. Ngoài ra, giống như bài toán 2, việc chọn cặp "độ lệch tuyệt đối trung vị và trung vị" ở đây cũng không thuận tiện, vì không dễ dàng để tìm trung vị trong trường hợp dữ liệu ghép lớp, đặc biệt là khi nó được cho bằng biểu đồ. Độ lệch tuyệt đối trung bình hay độ lệch chuẩn vẫn là các tham số cần ưu tiên để tìm câu trả lời.

Có thể dễ dàng tìm số trung bình của hai phân phối, nhưng độ lệch tuyệt đối trung bình và độ lệch chuẩn thì đòi hỏi những tính toán phức tạp hơn, đặc biệt là trong trường hợp phân phối biểu thị bởi biểu đồ *E*. Vì thế, tính giá trị cụ thể của độ lệch tuyệt đối trung bình (hay độ lệch chuẩn) không phải là chiến lược tối ưu. Để nhanh chóng tìm ra câu trả lời chỉ cần quan sát hai biểu đồ. Trước hết, có thể đánh giá gần đúng số trung bình của mỗi phân phối dữ liệu, sau đó căn cứ vào mật độ dữ liệu quanh số trung bình - thể hiện ở số hình chữ nhật "đứng gần" số trung bình và tổng các chiều cao của chúng. Cụ thể: đối với biểu đồ *D,* có thể ước lượng số trung bình nằm "gần bên trái giá trị 12"; sát số trung bình chỉ có một hình chữ nhật có chiều cao bé hơn rất nhiều so với tổng chiều cao ba hình còn lại nằm cách xa về hai phía. Trong khi đó, với biểu đồ *E* thì từ quan sát các hình chữ nhật sẽ ước lượng được số trung bình của phân phối dữ liệu nằm "gần bên trái số 9" và quanh đó tập trung nhiều dữ liệu. Suy ra độ phân tán của phân phối dữ liệu ứng với nhóm *E* bé hơn so với nhóm *D*.

3.3. Phân tích kết quả thu được

Ghi nhận đầu tiên là khái niệm *độ lệch tuyệt đối trung bình* không được SV nào sử dụng để phân tích sự phân tán của những phân phối dữ liệu cho trong các Bài toán 1 và 3. Hiện tượng

đó có thể được giải thích bởi sự vắng mặt của nó trong chương TK trình bày ở các Sách giáo khoa Đại số 10 (ví dụ như Tran et al., 2011)), dù Sách giáo viên đi kèm có nói đến.

Đối với *Bài toán 1*

Trong 25 SV tham gia thực nghiệm, có:

- 11 chọn loại xe Toyo. Sự lựa chọn dựa trên số trung bình của phân phối dữ liệu;
- 8 chọn loại xe Dana do căn cứ vào cả hai loại tham số đo sự tập trung và phân tán. Lập luận của họ là: hai phân phối dữ liệu có số trung bình gần như nhau, nhưng phân phối ứng với Dana có độ lệch chuẩn bé hơn.
- 6 chọn Toyo. Lập luận đưa ra là: Dãy phân phối ứng với nó chỉ có 3/19 (gần 16%) giá trị nằm dưới 100, trong khi đó ứng với Dana có đến 5/19 (hơn 26%). Vậy loại xe Toyo đáng tin hơn cho cuộc đua dài 100 *km*.

Ta thấy chỉ có 8 SV tính đến cả độ tập trung lẫn độ phân tán của phân phối dữ liệu. Tuy nhiên, họ chỉ so sánh hai số trung bình và so sánh hai độ lệch chuẩn trong sự tách biệt. Chúng tôi không tìm thấy ở họ một giải thích tường minh dựa trên hiệu giữa số trung bình với độ lệch chuẩn của cùng một phân phối, hoặc dựa vào hệ số biến thiên. Như chúng tôi đã nói, việc tìm hiểu độ lệch chuẩn đòi hỏi phải tính đến số trung bình. Phải chăng họ chỉ tính toán độ lệch chuẩn theo công thức đã biết mà không tính đến ý nghĩa của các tham số.

Có 6 SV chỉ để ý số dữ liệu nằm dưới giá trị 100. Họ đã không hề phân tích các giá trị còn lại của hai phân phối TK thu được. Ta có thể đưa ra hai phân phối được lựa chọn phù hợp như một phản ví dụ để bác bỏ lập luận này. Đặc biệt, khi hai phân phối có cùng số giá trị nằm dưới 100 thì lập luận của họ không cho phép tìm câu trả lời.

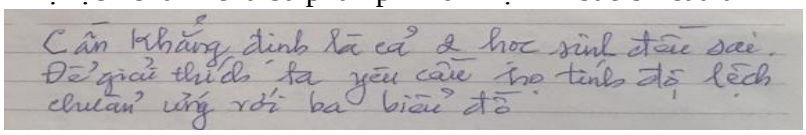
Đối với *Bài toán 2*

Quan sát kết quả thu được ở Pha 2, chúng tôi thấy 23/25 SV cho câu trả lời đúng, nhưng chiến lược tìm câu trả lời của họ khác nhau.

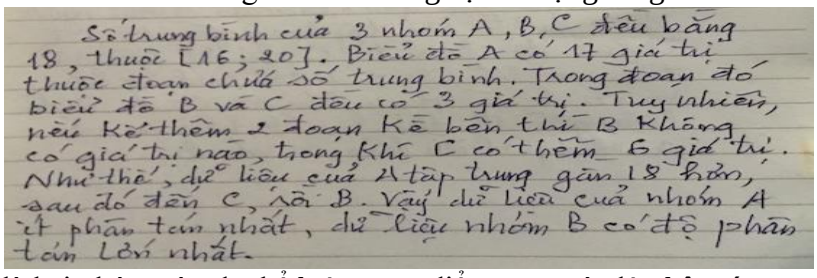
- 16/25 SV vẫn tính độ lệch chuẩn, dù dữ liệu cho bằng biểu đồ. Kết quả họ đưa ra là: $S_A \approx 8{,}85$; $S_B \approx 14{,}52$; $S_C \approx 9{,}52$. Tất nhiên, trước hết họ phải chuyển dữ liệu cho trên biểu đồ về dạng bảng, rồi tính giá trị trung bình và sau đó là độ lệch chuẩn theo công thức đã biết. Việc tính số trung bình là do sự xuất hiện của nó trong công thức tính độ lệch chuẩn chứ không hẳn là SV đã chú ý đến quan hệ giữa hai loại tham số TK. Trong số 16 SV này, 2 người chưa đi đến đáp số cuối cùng.
- 9/25 để ý đến tính đối xứng của cả ba biểu đồ nên không cần tính toán mà có ngay giá trị trung bình của ba phân phối: $\overline{x_A} = \overline{x_B} = \overline{x_C} = 18$.
- Trong số 9 SV này, 5 người tiếp tục tính độ lệch chuẩn theo công thức và đi đến kết quả: $S_A < S_C < S_B$. Như vậy, phần lớn SV (21/25) vẫn tính độ lệch chuẩn theo công thức để tìm câu trả lời.
- Chỉ có 4 trong 9 SV sau khi ước lượng $\overline{x_A} = \overline{x_B} = \overline{x_C} = 18$ đã tiếp tục quan sát biểu đồ (không thực hiện tính toán), nhận xét về mật độ các giá trị của từng phân phối quanh số trung bình để kết luận $S_A < S_C < S_B$. Cả 4 SV này đều thuộc nhóm 8 người đã tính đến cả số trung bình lẫn độ lệch chuẩn khi giải Bài toán 1. Bốn người còn lại trong số 8 người đó vẫn trở về với việc tính độ lệch chuẩn theo công thức.

Ở Pha 3, SV được chia ngẫu nhiên thành sáu nhóm (4-5 người). Do đã tìm ra kết quả đúng trong Pha 2 nên sáu nhóm đều thấy ngay là cả hai HS đều sai. Tuy nhiên, đối với câu

hỏi "*bạn sẽ nói gì với những HS này"* thì ứng xử của họ khác nhau. 4/6 nhóm đã viện dẫn đến việc tính ba độ lệch chuẩn của ba phân phối dữ liệu để bác bỏ câu trả lời sai của HS:


Cần khẳng định là cả 2 học sinh đều sai. Để giải thích ta yêu cầu họ tính độ lệch chuẩn ứng với ba biểu đồ

Còn lại 2 nhóm đưa ra cách giải thích không dựa sử dụng công thức tính toán. Chẳng hạn:


Số trung bình của 3 nhóm A, B, C đều bằng 18, thuộc [16; 20]. Biểu đồ A có 17 giá trị thuộc đoạn chứa số trung bình. Trong đoạn đó biểu đồ B và C đều có 3 giá trị. Tuy nhiên, nếu kể thêm 2 đoạn kề bên thì B không có giá trị nào, trong khi C có thêm 6 giá trị. Như thế, dữ liệu của A tập trung gần 18 hơn, sau đó đến C, rồi B. Vậy dữ liệu của nhóm A ít phân tán nhất, dữ liệu nhóm B có độ phân tán lớn nhất.

Ta thấy là hai nhóm này đã thể hiện quan điểm xem xét độ phân tán trong mối quan hệ với số trung bình và lập luận dựa trên biểu đồ chứ không phải trên các tính toán độ lệch chuẩn.

Không nhóm nào trong s nhóm chỉ ra nguồn gốc sai lầm của HS thứ nhất. HS này quan niệm độ phân tán thể hiện ở số hình chữ nhật. Như đã nói, số hình chữ nhật nhiều không phải là dấu hiệu chứng tỏ độ phân tán lớn. Hơn nữa, không nhóm nào dùng câu trả lời do chính HS đề nghị để chỉ ra sự bất cập của lập luận khi xét các biểu đồ *A, B*. Cũng không có nhóm nào chỉ ra quan niệm sai lầm của HS thứ hai: sự thay đổi chiều cao các hình chữ nhật phản ánh sự biến thiên của tần số (hay tần suất, nếu là biểu đồ tần suất ghép lớp) chứ không phải là sự biến thiên của các giá trị của phân phối.

Đối với *Bài toán 3*

Nếu như với Bài toán 2 có 9/25 SV ở Pha 2 và 2/6 nhóm ở Pha 3 tìm thấy số trung bình của ba phân phối dữ liệu mà không cần sử dụng công thức, thì với Bài toán 3 đã không có nhóm nào làm được điều đó. Cả 6/6 nhóm đều áp dụng thuật toán tính số trung bình bằng công thức. Có thể giải thích hiện tượng này bởi khó khăn của việc xác định số trung bình trên các biểu đồ không đối xứng như *D* và *E*. Lí do thứ hai nằm ở chỗ vấn đề ước lượng gần đúng các tham số TK không hề được đề cập trong dạy học ở Việt Nam. Đó là một biểu hiện về tính "thực tiễn hình thức" của dạy học TK. Thực ra, đối với Bài toán 3 thì chỉ cần ước lượng số trung bình, chỉ ra lớp ghép chứa nó là đủ, không nhất thiết phải tính giá trị cụ thể.

Sau khi tính được $\overline{x_D} = 11{,}3;\ \overline{x_E} = 8{,}5$, có 3 nhóm tiếp tục chiến lược tính toán theo công thức để xác định độ lệch chuẩn ($S_D \approx 7{,}47;\ S_E \approx 4{,}6$) và đưa ra câu trả lời đúng. Ba nhóm còn lại lập luận trên biểu đồ và cũng có cùng câu trả lời. Trong số 3 nhóm này có 2 nhóm trước đó, ở Pha 3, đã từng sử dụng biểu đồ để tìm cách bác bỏ câu trả lời sai của hai HS giả định. Nhóm còn lại trong ba nhóm đó từ bỏ chiến lược tính toán độ lệch chuẩn bằng công thức đã theo đuổi ở Pha 3. Dường như họ đã chịu ảnh hưởng của chiến lược sử dụng biểu đồ do hai HS giả định đưa ra.

Xét biểu đồ D: Có 12 giá trị nằm ở đoạn cuối bên phải. Số giá trị nằm ở đoạn mút phía trái là 4. Số trung bình $\overline{x_D} = 11{,}3$ *thuộc đoạn [9; 12]. Không có giá trị nào thuộc đoạn này. Kề đoạn đó chỉ có 4 giá trị. Chứng tỏ các dữ liệu không tập trung ở giữa.*

Tương tự, xét biểu đồ E : Số trung bình $\overline{x_E} = 8{,}5$ *thuộc đoạn [6; 9]. Có 5 giá trị nào thuộc đoạn này. Tính thêm cả hai đoạn kế bên về hai phía thì có 17 giá trị. Như vậy dữ liệu tập trung ở giữa nhiều hơn so với biểu đồ D. Chứng tỏ dữ liệu ở biểu đồ D phân tán hơn.*

Một nhóm còn nói thêm: "Khoảng cách giữa hai mút ở biểu đồ D là 21, ở biểu đồ E là 18 (nhỏ hơn 21)". Về bản chất, nhóm này đang nói đến "biên độ". Giống như "độ lệch tuyệt đối trung bình", khái niệm "biên độ" không được đưa vào Sách giáo khoa Đại số 10 hiện hành.

4. Kết luận và bàn luận

Ngày nay TK đã tạo thành một ngôn ngữ mà mỗi công dân cần nắm vững. Vị trí của TK trong xã hội buộc chúng ta phải suy nghĩ đến vấn đề dạy học môn học này với mục tiêu đào tạo lớp công dân tương lai. Vấn đề đối với mỗi công dân không chỉ ở chỗ tạo ra các dữ liệu TK, mà thường gặp hơn lại là biết giải thích kết quả nhận được, biết đưa ra một đánh giá cá nhân, một ý kiến phản biện theo phương pháp khoa học trên những dữ liệu, những biểu đồ TK gặp trong cuộc sống hàng ngày. Như vậy, nếu muốn phát triển tư duy TK ở HS thì cần phải làm cho họ có những hiểu biết cơ sở để giải thích các dữ liệu TK.

Độ phân tán cũng quan trọng như xu hướng hội tụ trong việc mô tả một hiện tượng quan sát được. Việc nhận ra đặc trưng phân tán của dữ liệu cho bằng biểu đồ lại càng quan trọng khi giáo viên muốn bồi dưỡng ở HS năng lực giải quyết các vấn đề thực tiễn, khả năng thực thi vai trò công dân một cách khoa học trong những tình huống TK.

Nghiên cứu của chúng tôi cho thấy nhiều SV ưu tiên cho số trung bình trong phân tích dữ liệu (Bài toán 1). Khi cần xem xét độ phân tán của phân phối dữ liệu thì họ đã không huy động độ lệch tuyệt đối trung bình, chỉ sử dụng độ lệch chuẩn. Điều đó khiến họ mất đi những thuận lợi trong việc phân tích độ phân tán của dữ liệu cho bằng histogram (bởi vì độ lệch tuyệt đối trung bình thể hiện rõ tư tưởng của khoảng cách, là cái dễ nhận ra trên các hệ trục). Như vậy, việc bổ sung khái niệm độ lệch tuyệt đối trung bình vào chương trình đào tạo từ bậc phổ thông là cần thiết.

Những SV tham gia thực nghiệm thường ưu tiên cho việc tính độ lệch chuẩn theo công thức, ngay cả khi dữ liệu được cho bằng biểu đồ. Họ phải chuyển thông tin cho trên biểu đồ về dạng bảng để áp dụng công thức. Dường như hiểu biết của họ về độ lệch chuẩn bị giới hạn vào các công thức tính toán. Đó là hệ quả của xu hướng xem dạy học TK như một sự áp dụng các thuật toán. Về xu hướng này, Duperret (2001) đã nhấn mạnh: "nếu dạy học TK quy về việc áp dụng máy móc các công thức, không dựa trên việc hiểu nghĩa khái niệm, thì người ta sẽ phải đặt ra câu hỏi về lợi ích của việc dạy học này" (p.9).

Nguyên nhân khiến SV lúng túng khi làm việc với các histogram chính là việc không tính đến nghĩa của các tham số phân tán – thể hiện qua mối quan hệ của chúng với tham số hội tụ. Dường như họ có khó khăn khi cần nghiên cứu sự biến thiên của dữ liệu bằng cách xấp xỉ chúng với trung tâm của phân phối. Việc thiếu thói quen ước lượng xấp xỉ các tham số làm khó khăn của họ càng lớn hơn, dù tình huống chúng tôi đưa ra đã giới hạn ở những phân phối dữ liệu ghép theo các lớp có độ dài bằng nhau, thậm chí biểu đồ có tính đối xứng như ở Bài toán 2.

Có thể hình dung là sự lúng túng sẽ càng trầm trọng hơn nếu biểu đồ ứng với các lớp ghép không đều nhau. Lúc đó, không thể tham chiếu vào sự biến thiên về chiều cao của dãy hình chữ nhật để phân tích các tham số TK nữa. Về vấn đề này, chúng tôi đã có kết quả nghiên cứu của tác giả Tang Minh Dung (2009):

> Sinh viên vẫn đang nhầm lẫn giữa đặc trưng chiều cao (…) và đặc trưng diện tích của biểu đồ tổ chức. (…) Các giáo viên tương lai chưa được trang bị đầy đủ kiến thức để có thể giải thích về tính hợp thức và tính trung thực của dạng đồ thị thống kê này. (…) Việc rèn luyện tư duy

> thống kê (…) cho sinh viên sư phạm còn khiếm khuyết. (…) Họ chưa thể đọc và thể hiện đúng theo quy tắc toán học các thông tin tần số-tần suất trên biểu đồ tổ chức. (Tang, 2009, p.59)

Ứng xử sư phạm của SV hiển nhiên phụ thuộc nhiều vào kiến thức họ có về tri thức đang bàn đến. Điều đó thể hiện rất rõ trong Pha 3: phần lớn SV yêu cầu HS dùng công thức tính độ lệch chuẩn để bác bỏ kết quả sai lầm mà các em đưa ra ban đầu. Như vậy, SV không tính đến quan niệm sai ẩn phía sau câu trả lời của HS, chỉ tìm cách bác bỏ nó bằng các kết quả thu được từ tính toán.

Như chúng tôi đã nói ở phần đầu của bài báo, đào tạo về TK không được coi trọng đúng mức, thường bị bỏ qua trong thực tế dạy học ở bậc phổ thông. Đối với các loại biểu đồ thì người ta chỉ giới thiệu cách vẽ và sau đó yêu cầu HS vẽ một loại biểu đồ nào đó. Dữ liệu cho sẵn ở dạng bảng, HS không phải tổ chức lại, cũng không được giao nhiệm vụ chọn loại tham số cần tính hay loại biểu đồ phù hợp cần vẽ. Họ cũng không phải đọc biểu đồ, phân tích dữ liệu cho trên biểu đồ, ước lượng các tham số của phân phối dữ liệu cho bằng biểu đồ.

Đó là thể chế dạy học TK ở bậc phổ thông. Còn thể chế đào tạo giáo viên Toán ở các trường sư phạm thì sao?

Chương trình đào tạo giáo viên Toán của các trường đại học sư phạm ở Việt Nam được phân thành ba nhóm:

- Nhóm các môn chung: gồm những học phần về triết học, ngoại ngữ, tâm lí học, giáo dục học…
- Nhóm các môn Toán cơ bản: mồm một số học phần thuộc các chuyên ngành Đại số, Giải tích, Hình học, Toán ứng dụng, trong đó có XS-TK.
- Nhóm các môn chuyên ngành: gồm các học phần về phương pháp giảng dạy Toán và ứng dụng công nghệ thông tin trong dạy học Toán.

Ngoài ra SV còn có hai đợt thực tập (làm quen với thực tiễn dạy học và thực hành các nhiệm vụ của một giáo viên).

Nhóm thứ hai trang bị cho SV một số lí thuyết toán học thuần túy. Nhóm thứ ba bàn về các nguyên tắc, mục đích, phương pháp dạy học Toán, các tình huống dạy học điển hình (như dạy định lí, dạy khái niệm, dạy giải bài tập) và những lưu ý trong dạy học một số chủ đề cụ thể (như hàm số, phương trình, bất phương trình, vectơ...). Một cấu trúc chương trình như vậy có thể xem là hợp lí. Vấn đề là nội dung cụ thể của các học phần.

Chúng tôi sẽ phân tích sơ bộ chương trình áp dụng ở Khoa Toán – Tin của Trường Đại học Sư phạm Thành phố Hồ Chí Minh để tìm các yếu tố giải thích cho kết quả thực nghiệm trình bày ở phần trên. Trong chương trình này, các nội dung về TK được nghiên cứu ở hai học phần bắt buộc có tên gọi *XS-TK 1, XS-TK 2* (thuộc nhóm các môn toán cơ bản) và việc dạy học chúng được bàn đến trong học phần *Phương pháp dạy học Đại số - Giải tích* (thuộc nhóm các môn chuyên ngành).

Đề cương chi tiết học phần *XS-TK 1* chỉ rõ sáu chương: *Không gian XS; Đại lượng ngẫu nhiên và phân phối XS; Các đặc trưng của biến ngẫu nhiên và một số định lí quan trọng trong lí thuyết XS; Ước lượng tham số; Kiểm định giả thiết TK; Hồi quy và tương quan*. Học phần *XS-TK 2* đưa vào 4 chương: *XS – độ đo tiêu chuẩn; Quá trình ngẫu nhiên; Các mô hình hồi quy; TK ứng dụng trong giáo dục*. Như vậy, liên quan trực tiếp đến tri thức mà chúng tôi lựa chọn trong bài báo này, chỉ có chương "Ước lượng tham số". Trong chương này SV được nghiên cứu "ước lượng điểm", "khoảng ước lượng", "khoảng ước lượng cho số trung bình", "khoảng ước lượng cho phương sai". Tuy nhiên, vấn đề ở đây không phải là

ước lượng trung bình, phương sai của mẫu, mà là tính toán các giá trị trên mẫu để đưa ra ước lượng cho tổng thể. Ở đây, người ta đưa vào các phương pháp và công thức ước lượng. Người ta chỉ giới thiệu những kiến thức mang tính hàn lâm, dành cho các chuyên gia về TK. Nhìn lại hai giáo trình thường được sử dụng cho dạy học, chúng tôi thấy tuyệt nhiên không hề xuất hiện các loại biểu đồ và những tình huống phổ thông của đời thường.

Học phần Phương pháp dạy học Đại số – Giải tích nghiên cứu 7 chương, bàn về nhiều nội dung dạy học ở trung học phổ thông (phương trình, hàm số, giải tích…). Các nội dung về *XS-TK* có mặt ở chương cuối cùng – chương "Dạy học mạch toán ứng dụng". Chương này đề cập đến 4 vấn đề:

- Mạch toán ứng dụng ở trường phổ thông;
- Dạy học những yếu tố của phương pháp số;
- Dạy học một số yếu tố của lí thuyết tối ưu;
- Dạy học một số yếu tố của TK, XS.

SV được làm việc với vấn đề thứ tư trong 4 tiết – 2 tiết lí thuyết, 2 tiết bài tập. Thời lượng này hiển nhiên là không đủ để bàn một cách sâu sắc về vấn đề dạy học những tri thức về XS-TK được đưa vào chương trình phổ thông.

Có thể nói rằng đào tạo giáo viên Toán về dạy học TK chiếm một vị trí thứ yếu, mặc dù khoa học TK ngày càng có vai trò quan trọng trong thực tế cũng như trong chương trình giáo dục phổ thông.

Kết quả thu được từ nghiên cứu của chúng tôi đặt ra những câu hỏi cần tính đến trong việc đào tạo giáo viên về dạy học TK nói chung, dạy học các tham số phân tán nói riêng. Kiến thức và khó khăn của SV trong việc làm chủ ngôn ngữ thống kê cho thấy cần phải suy nghĩ về nội dung đào tạo cũng như cách thức bồi dưỡng nghiệp vụ cho đội ngũ giáo viên tương lai. Những vấn đề gắn với đời thường và những kiểu tình huống làm việc với quan niệm sai lầm của HS mà chúng tôi sử dụng trong bài báo này có thể tạo nên điểm khởi đầu cho suy nghĩ đó.

❖ ***Tuyên bố về quyền lợi:*** *Tác giả xác nhận hoàn toàn không có xung đột về quyền lợi.*

TÀI LIỆU THAM KHẢO

Bakker, A. (2004). Reasoning about shape as a pattern in variability. *Statistics Education Research Journal*, *3*(2), 64-83.

Boyé, A. et Comairas, M.-C. (2002). Moyenne, médiane, écart type: quelques regards sur l'histoire pour éclairer l'enseignement des statistiques. *Repères-IREM*, *48*, 27-39.

Cooper, L., & Shore, F. (2010). The effects of data and graph type on concepts and visualizations of variability. *Journal of statistics education*, *18*(2), 1-16.

Delmas, R., & Liu, Y. (2005). Exploring students' conceptions of the standard deviation. *Statistics Education Research Journal*, *4*(1), 55-82.

Dodge, Y. (1993). *Statistique: dictionnaire encyclopédique*, Université de Neuchâtel, Suisse.

Duperret, J.-C. (2001). Des statistiques à la pensée statistique. *Publication IREM,* Université de Montpellier II.

Garfield, J., & Ben-Zvi, D. (2005). A framework for teaching and assessing reasoning about variability. *Statistics Education Research Journal*, *4*(1), 92-99

Gattuso, L. (1997). La moyenne, un concept évident? *Bulletin AMQ*, *37*(3), 10-19.

Makar, K., & Confrey, J. (2005). Variation-talk: Articulating meaning in statistics. *Statistics Education Research Journal*, *4*(1), 27-54.

Meletiou-Mavrotheris, M., & Lee, C. (2005). Exploring introductory statistics students' understanding of variation in histograms. *Proceedings of the 4th Congress of the European Society for Research in Mathematics Education*, Sant Feliu de Guíxols, Spain.

Ministry of Education and Training (2018). *Chuong trinh giao duc pho thong mon Toan* [*Mathematics General Education Curriculum*]. Hanoi.

Reading, C., & Shaughnessy, J. M. (2004), Reasoning about variation. In Ben-Zvi and J. Garfield (dir.), *The challenge of developing statistical literacy, reasoning and thinking*, 201-226. Dordrecht: Kluwer Academic Publishers.

Régnier, J-C. (2012). Enseignement et apprentissage de la statistique: Entre un art pédagogique et une didactique scientifique. *Statistique et Enseignement*, *3*(1), 19-36.

Tang, M. D. (2009). *Day hoc thong ke va van de dao tao giao vien* [*Teaching statistics and teacher training*]. Master Thesis of Education, Ho Chi Minh City University of Education.

Tran, V. H., Vu, T., Doan, M. C., Do, M. H., & Nguyen, T. T. (2000). *Dai so 10* [*Algebra 10*]. Vietnam education Publishing House.

Watson, J. M. (2007), The role of cognitive conflict in developing students' understanding of average. *Educational Studies in Mathematics*, *65*, 21-47.

DISPERSAL PARAMETER IN STATISTICS: KNOWLEDGE OF MATHEMATICS STUDENT TEACHERS AND SOME ISSUES FOR TEACHER EDUCATION

Le Thi Hoai Chau

Van Hien University, Vietnam

Corresponding author: Le Thi Hoai Chau – Email: chaulth@vhu.edu.vn

Received: March 11, 2020; Revised: March 31, 2020; Accepted: August 24, 2020

ABSTRACT

The research presented in this article aims to explore the knowledge of mathematics student teachers on the parameters of dispersion. Twenty five mathematics student teachers were exposed to the situations which require the mastery of the meaning of these parameters. The development of the situations proposed to the students was based on certain works dealing with learning difficulties for the understanding and the use of the dispersion parameters. The behaviors of the students observed show that the teaching of statistics which tends to focus on mechanically applying calculation techniques possibly contributes to the observed limited understanding of the meaning of parameters and the non-mastery of the languages of statistics (the histograms in this case). The results suggest the necessity to review the training of mathematics teachers in statistics.

Keywords: teacher knowledge; dispersion parameters; mean absolute deviation; standard deviation