Online-Journals.org

iJAC International Journal of Advanced Corporate Learning

iJAC | eISSN: 1867-5565 | Vol. 18 No. 3 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijac.v18i3.53197

TLIC PAPER

Let Computing Be a Powerful Tool for Teaching Conceptual Biostatistics to Public Health Students

Qi Zheng(✉)

Texas A&M School of Public Health, College Station, Texas, USA

qzheng@tamu.edu

ABSTRACT

Over a decade ago some statistics educators were excited at the prospect of using computing as an effective tool for teaching conceptual knowledge of statistics, but few concrete examples of this novel teaching approach have emerged. Here I describe a novel method of treating computing as a tool for teaching conceptual biostatistical knowledge to public health students. By this new approach instructors translate the task of learning conceptual biostatistical knowledge into a process of doing carefully designed computing exercises that are reliant on elementary mathematics and first principles thinking. Infusing basic computer coding knowledge into the task of learning biostatistics, instructors succeed in bringing conceptual biostatistical knowledge into the zone of proximal development for public health students, catalyzing a beneficial synergistic effect in teaching biostatistics and computing. I demonstrate the feasibility of the new approach by citing unedited student work examples drawn from an undergraduate biostatistics course where students used the command-line R computing environment as their course software.

KEYWORDS

Public health, zone of proximal development, computational thinking, constructivism, likelihood function

1 IS THERE A COMPUTING GENE?

At the 2024 Learning Ideas Conference I articulated a possibly novel proposition that computing ability is universal [1]. Because this proposition underpinned my distinctive method of using computing to teach conceptual biostatistics to public health students, the curiosity among the audience was palpable, and the questions raised by the audience were encouraging and inspiring. When I later presented that proposition at the 2024 Texas Conference on Student Success, some attendees enthusiastically expressed their appreciation of the relevance of my proposition to today's education. One veteran educator seemed to thrill to read on the poster board my claim that "every student has an innate inclination to enjoy the beauty of

Zheng, Q. (2025). Let Computing Be a Powerful Tool for Teaching Conceptual Biostatistics to Public Health Students. *International Journal of Advanced Corporate Learning (iJAC)*, 18(3), pp. 103–118. https://doi.org/10.3991/ijac.v18i3.53197

Article submitted 2024-11-05. Revision uploaded 2024-12-22. Final acceptance 2025-01-05.

computing." It is helpful that I mention at the outset what propelled me to advocate the concept of a computing gene.

As I reported previously [1], my proposition was inspired by a better-known proposition in mathematics education that math ability is universal, which was only recently made known to the public by Tyre [2]. My proposition concerning universal computing ability has several corollaries that parallel those of the proposition concerning universal math ability. For example, Colleran [3] argues that all humans "have an innate, natural ability to do mathematics to the highest level: we are all endowed with a 'maths gene'." This argument can be traced to Butterworth [4], who coined the term "number module" to refer to this hypothetical math gene. Regardless whether this mathematical gene will eventually be biologically provable, the idea of a math gene has no doubt helped usher in a far-reaching "math revolution" in American middle schools and high schools, which was publicized by Tyre's absorbing account appearing in The Atlantic magazine [2] in 2016. I would venture to hypothesize that the mathematical gene has a counterpart called a computing gene. Thus, all students have an innate ability to learn computing to high levels. Another important corollary of my proposition is that computing ability is not always easy to spot. The parallel corollary concerning mathematical ability is well phrased by Borovik and Gardiner [5] as follows. "Mathematical abilities in a child are often dormant and remain unnoticed both by the child and his or her teachers." I believe that the same can be said about computing abilities in many a public health student.

Despite a myriad of intensive investigations into human's innate mathematical abilities, the existence of a math gene has been supported so far only by empirical evidence. Small wonder that the dream of scientists being able to edit the mathematical gene for the remediation of defects in that gene could come true only in the far-distant future, according to the renowned neuropsychologist Butterworth and colleagues [6]. However, the math revolution did not wait for the location of the mathematical gene to be pinpointed and the sequence of the gene determined. The math revolution arrived unannounced, catalyzed by large numbers of enthusiastic math teachers who believed that mathematical abilities were innate. Furthermore, holding the unshakable belief that everyone can learn mathematics to high levels, Boaler and colleagues [7] spread the math revolution into the arena of online teaching by developing an open online math course that has been taken by over 160,000 participants. By contrast, research into the computing gene is scant. However, the idea of "computational thinking for everyone," as advocated by Wing [8], and the growing recognition among educators that computational thinking is a fundamental competency for an informed citizen, as pointed out by Grover and Pea [9], suggest that many educators now believe, at least subconsciously, that computing ability is universal.

Lack of solid genetic evidence for universal computing ability should not be an excuse for inaction, which is an inspiring lesson one can learning from the math revolution. The increasing awareness of the importance of computational thinking is conducive to a movement at public health schools to weave computing into biostatistics as a tool for teaching conceptual knowledge. The work I discuss here exemplifies one way to reach that goal, which may help address the crippling shortage of public health professionals.

2 RESPONDING TO A PUBLIC HEALTH CRISIS

In 2008, a group of prominent public health leaders issued an urgent call to confront the public health workforce crisis [10]. Eleven years later, the Covid-19 pandemic exacerbated the crisis, prompting another group of outstanding public

health researchers to sound a clarion call to action to expand the public health workforce. According to these researchers, the minimum number of additional full time equivalent public health workers that must be hired to allow local and state public health departments to deliver adequate core public health services was estimated at 80,000 [11]. An obvious yet important way to ameliorate the crippling shortage of public health professionals is to increase public health educational capacity [10]. However, the multifaceted task of training 21st-century public health professionals has a uniquely daunting dimension. The ever-increasing flood of public health data has created a demand for an unprecedented increase in both breadth and depth of biostatistical knowledge in public health curricula [12], as well as a correlative increase in computational skills that were unheard-of two decades ago [13].

Clearly, the public health workforce is multitiered in terms of the extent to which public health professionals engage in active research. As Brearley et al. [14] rightly emphasized, some public health professionals will not do research and just need to keep up with the literature in their fields. However, society increasingly needs public health professionals who are conversant with complex biostatistical methods and related advanced computing technologies.

To public health graduate students, e.g., students pursuing MPH or DrPH degrees, biostatistics and computing are two intertwined themes forming an integral part of their curricula. Despite repeated calls for action to update and enrich these two themes in public health graduate curricula [13], [15], concrete and actionable schemes for curriculum change have rarely been proposed. A noteworthy recent effort towards this direction [14] illustrates how exciting and challenging curriculum change at public health school can be. That effort gave rise to an innovative biostatistical literacy course aiming at public health graduate students and health sciences professionals whose work does not involve research. This paper addresses a different yet increasingly important segment of public health students who will analyze their own data and conduct their own public health research. A biostatistical literacy course is far from sufficient for these students because they need a solid conceptual understanding of key statistical ideas to allow them to apply a myriad biostatistical procedures properly. In this article, I give an overview and the rationale of a computing based pedagogical approach that is conducive to teaching conceptual biostatistical knowledge to public health students,

3 CONCEPTUAL BIOSTATISTICAL KNOWLEDGE FOR PUBLIC HEALTH STUDENTS

Educators may have different opinions about what constitutes conceptual biostatistical knowledge for public health students. My view on this issue is in agreement with the idea of helping students build procedural fluency buttressed by a clear conceptual understanding, which was advocated by education researchers within the context of teaching advanced placement (AP) statistics to aspiring high school students [16]. In the present context, conceptual knowledge should directly help public health students to apply statistical procedures and interpret the results with confidence. Public health students should not be distracted by information that is not essential for building procedural fluence. Thus, for example, how to construct the likelihood function for a small logistic regression model should be taught, but how to maximize that likelihood function algorithmically should be omitted. We should never lose sight of our overarching goal: we aim to train public health students to be informed consumers of biostatistics in their public health practice and research, but we should not hope to elevate these students to the level of a professional statistician who engage in statistical methodology research [17].

I find it helpful to loosely divide the desired conceptual biostatistical knowledge into four groups. To begin with, consider the well-known t-test formula that students encounter in their first biostatistics courses:

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{1}$$

Here x_1 and x_2 are two random samples drawn from two independent normal populations. The sample sizes are n_1 and n_2, respectively. The quantity t is used to test whether the two population means are equal when the two population variances may be assumed to be equal.

I call the first group of knowledge easy topics. For example, the two sample means $\overline{x_1}$ and $\overline{x_2}$ in equation (1) belong to this group because students can understand them easily. The second group consists of claims. For example, when the two population means are equal, the t statistic in equation (1) follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom. This statement about the t statistic is a claim. The third group consists of procedural knowledge. For example, showing students how to manually compute the t statistic using real-world or fictitious data is a process of imparting procedural knowledge. Showing students how to perform a t test using a statistical software package is also a process of imparting procedural knowledge. The last group consists of conceptual knowledge, which helps students better understand claims or procedural knowledge. For example, instructors are conveying conceptual knowledge when they explain why a large magnitude of the t statistic constitutes statistical evidence against the null hypothesis that the two population means are equal. In this simplistic example, conveying conceptual knowledge seems an easy task. Under the null hypothesis, the sample mean difference $d = \overline{x_1} - \overline{x_2}$ should fluctuate around zero because the mean of d is zero. If the magnitude of d is excessively large the null hypothesis is less likely to be true than the alternative hypothesis that the two population means are unequal. Because the sample means can be changed by choosing different measuring units, e.g., in inches or in centimeters, the sample difference d must be normalized by the quantity in the denominator in equation (1) to get a meaningful test statistic.

In the above trivial example, conceptual knowledge can be taught easily. However, in general, teaching conceptual knowledge to public health students poses unique challenges that have received only scant attention so far. Consider extending the t test to the comparison of three or more normal population means. The test of interest is the F test in the one-way analysis of variance procedure [18, pp. 554–556]. To help students acquire a conceptual understanding of the F test, the instructor should discuss the sample means for each population and the grand mean. The instructor should then focus introduce a weighted sum of the squared distances between each population means and the grand mean, Students then learn that the magnitude of this weighted sum, commonly denoted by SS_B, plays a role similar to that of the mean difference d in the two sample t test. Students finally see that SS_B should be normalized by an estimate of the common variance within each population called SS_W by analogy with the two-sample t test. These instructional process helps students see why the following F test makes sense:

$$F = \frac{SS_B / (a-1)}{SS_W / (N-a)} \tag{2}$$

Here, a is the number of populations sampled, and N is the total number od samples. Under the null hypothesis that all population means are the same, the above

F statistic obeys an F distribution with $a - 1$ and $N - a$ degrees of freedom. However, to digest the whole idea, students need to work on numeric examples to undergo a process of knowledge construction, which was recently advocated by Zheng [19] from a constructivism perspective [20]. Students can work on such an example by hand or by a calculator as found in many popular introductory statistics textbooks. However, students can understand the concept of the F test more efficiently by writing computer code to work on the same kind of numeric problems, because this approach takes away the burden of tedious arithmetic to allow students to focus more attention on building a conceptual understanding of the F test. Figure 2 shows how an undergraduate student underwent this process, which will be discussed later.

In the above F test example my computational approach is helpful but not essential. However, most biostatistical procedures important to public health students are based on statistical models that students find difficult to understand conceptually. To help students understand these models I designed similar computational exercises to allow students to focus their attention on the likelihood function. Students are given a small data set and a particular type of model to be considered, such as the logistic regression model [21] or the Poisson regression model [22]. Students then figure out the log likelihood function by first principles following an in-class example, then code the log likelihood function using a popular computer language, and finally compute the maximum value of the log likelihood function by setting the parameters to their maximum likelihood estimates that are determined by a prestigious statistical software tool. Students often need to debug their code to make the maximum value of the log likelihood function they fine agree with that produced by the statistical package. I have given numerous examples of using this computational approach to teach conceptual biostatistical knowledge to public health students [17, 19, 21, 22]. I now provide further information about the rationale of this approach in hope of inspiring other educators to consider computing as an efficient tool to convey conceptual biostatistical knowledge to public health students.

4 THE RATIONALE FOR TAKING A COMPUTING APPROACH

The spectrum of today's public health students' mathematical ability ranged from hating mathematics to enjoying it [19]. While haters and aficionados of mathematics are not common among these students, a great majority of them are mathematically unprepared for learning statistical conceptual knowledge solely by abstract mathematical reasoning. For example, learning the all-important concept of the likelihood function would be an extremely difficult task for most public health students if the instructors rely only on mathematical reasoning. As one may rightly argue, the same may be true for public health students' computing ability. Indeed, some students had never used a text editor before they enrolled in my biostatistics classes. Most of these students have never used a command-line environment to perform a computing task before entering a biostatistics class. It is worth noting that math educators increasingly see the detrimental effect of a prevailing yet false worldview that some students have an innate ability to do mathematics while others are innately incapable of doing mathematics. It is increasingly accepted that the wide range of mathematical proficiency is the result of the amount of time spent doing and practicing mathematics, as Colleran [3] put it. I would here share a parallel idea: it is detrimental to divide students into those who are innately capable of doing computing and those who are innately incapable of doing computing. Every student has an innate inclination to enjoy the beauty of computing. Many public health

students are relatively weak in mathematics and computing because they have not been given much opportunity to develop their mathematical and computing minds.

Thus, accepting the existence of both a math gene and a computing gene, a biostatistics instructor is faced with a difficult choice between two distinct ways of teaching conceptual biostatistical knowledge: instructors may consider weaving mathematics or computing into their biostatistics courses. However, as I previously emphasized [19], one can make an important distinction between most public health students' mathematical ability and their computing ability from the perspective of the zone of proximal development (ZPD). The ZPD as proposed by Vygotsky [23, p. 84] was originally a concept in child psychology, but it has been successfully applied in education research and practice. The ZPD theory distinguishes a student's actual development level and their next level of performance achievable with the help of a teacher. Learning tasks in a student's ZPD are those that they can accomplish only when aided by a teacher. Using Figure 1, I here explain my computational pedagogical approach by borrowing concepts from the ZPD theory.

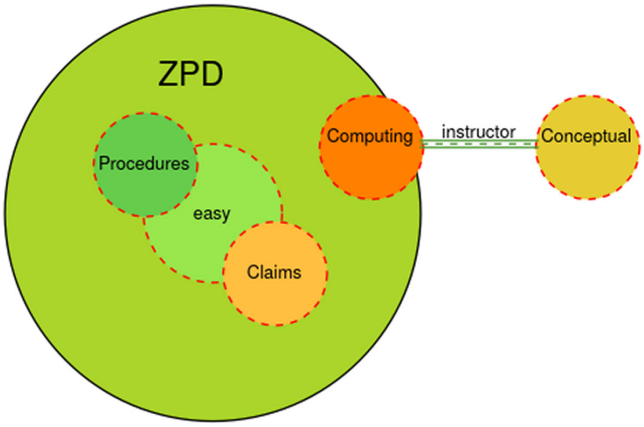

Fig. 1. Diagram showing how computing can help students acquire conceptual biostatistical knowledge

The small circle at the center of the largest circle represents learning tasks that students can accomplish unaided, and I call these tasks easy ones. As mentioned earlier in the *t* test example, computing a sample average is an easy task. The area between the "easy" circle and the largest circle, labeled ZPD, represents tasks that are extremely difficult or impossible for a student to accomplish without the aid of a teacher. For example, part of the procedural knowledge can be within the easy zone, and the rest inside the ZPD. The area outside the ZPD represents learning tasks that students cannot accomplish even with the help of a competent teacher, as the learning curve would be far too steep for most students. Conceptual knowledge as taught to statistics majors would be outside the ZPD for most public health students. However, as demonstrated in my previous reports [9, 10], an instructor can assemble a set of computing skills that are mostly within the ZPD for most public health students. The instructor can then design computational exercises that allow students to learn and use the computing skills to build conceptual knowledge. In doing so, the instructor plays the role of a bridge between a student's ZPD and conceptual knowledge that is otherwise outside their ZPD.

As the examples in my previous report [21] clearly show, the method I used to bridge the gap computing skills within the ZPD and difficult biostatistical concepts bears striking resemblance to the innovative use of subgoal labeled worked examples by Margulieux et al. [24] to teach an introductory Java programming course.

First, I discuss a worked example that breaks the learning of a concept into several small coding tasks. Then, students solve a similar practice problem. Finally, students solidify their knowledge by redoing the worked example and practice problem in preparation for exams. While students in the study of Margulieux et al. [24] focused on programming per se, students in my classes strived to assimilate conceptual biostatistical knowledge by solving well-designed coding problems.

To highlight this important distinction, I here revisit an example from [21]. The data in this example is adapted from a real-world experiment that was conducted to explore the effects of a pesticide on the death of beetles. I use the simplified data given in Table 1 to help students focus on important concepts.

Table 1. Data adapted from a pesticide experiment

Concentration	10.8	11.6	12.1	12.6	13.1	13.5
Death	15	24	26	24	29	29
Group size	50	49	50	50	50	49

Students learn easily to write simple SAS [25] code to fit a logistic model to the data, and they obtain the maximum likelihood estimates: intercept = –4.81 and concentration = 0.39. It is still relatively easy for students to understand that this model fitting exercise leads to the following empirical model:

$$\text{Prob(a beetle dies at concentration } C) = \frac{1}{1+e^{-(-4.81+0.39C)}}$$

So far, the learning process occurs within the ZPD of most public health students. To help students develop a conceptual understanding of the above logistic model, I teach students to think in the following fashion. Let the unknown intercept and concentration parameters be denoted by a and b, respectively. The probability of 15 beetles dying out of 50 beetles exposed to the pesticide at the concentration of 10.8 should be proportional to the expression

$$\left(\frac{1}{1+e^{-(a+10.8b)}}\right)^{15}\left(\frac{1}{1+e^{(a+10.8b)}}\right)^{35}$$

Using basic properties of the logarithmic function from high school algebra, students then write the logarithm of the above expression as

$$-15\log(1+\exp(-(a+1.08\,b))) - 35\log(1+\exp(a+1.08\,b))$$

Students now can see clearly that this is the first component of the log likelihood function that is the sum of six such components. Imitating an in-class example, students translate these ideas into a function of a and b, which is implemented in SAS, and test the function by assigning random values to a and b. Figure 2 shows a student's work resulting from this learning process. As Figure 3 shows, students can then see whether their definition of the log likelihood function is the same as the one hidden inside SAS by computing the value of their own log likelihood function using the maximum likelihood estimates of a and b supplied by a standard procedure of SAS. In Figure 3, the student triumphantly finds that the maximum value of her own log likelihood function agrees with the corresponding value presented to her by a standard SAS procedure.

```
/* PHEB 609 HW2 */
proc fcmp outlib=work.hw2.a;
function like1(a,b);
l1=-15*log(1+exp(-(a+10.8*b)))-35*log(1+exp(a+10.8*b));
l2=-24*log(1+exp(-(a+11.6*b)))-25*log(1+exp(a+11.6*b));
l3=-26*log(1+exp(-(a+12.1*b)))-24*log(1+exp(a+12.1*b));
l4=-24*log(1+exp(-(a+12.6*b)))-26*log(1+exp(a+12.6*b));
l5=-29*log(1+exp(-(a+13.1*b)))-21*log(1+exp(a+13.1*b));
l6=-29*log(1+exp(-(a+13.5*b)))-20*log(1+exp(a+13.5*b));
return (l1+l2+l3+l4+l5+l6);
endsub;
run;
options cmplib=work.hw2;

data beetle;
a=2; b=3; lik=like1(a,b); output;
run;

proc print data=beetle;
run;
```

The SAS System

Obs	a	b	lik
1	2	3	-5795.3

Fig. 2. A student's work showing how the log likelihood function is coded in SAS for the beetle example

After completing the above computational learning task, students develop a deeper understanding of the meaning the maximum likelihood estimates *a* and *b* and the precise model from which these estimates are derived. Students acquire this kind of conceptual knowledge by a hands-on meaning-making process without the burden of learning new mathematics, and they acquired coding skills as a byproduct.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates: Log Likelihood	Intercept and Covariates: Full Log Likelihood
AIC	415.062	408.038	32.245
SC	418.759	415.432	39.640
-2 Log L	413.062	404.038	28.245

-2log(L)=404.038, hence log (L)=l= **- 202.019**

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.8097	1.6210	8.8037	0.0030
conc	1	0.3893	0.1315	8.7618	0.0031

Confirmed:

```
data beetle;
a=-4.81; b=0.39; lik=like1(a,b); output;
run;
proc print data=beetle;
run;
```

(-4.81, 0.39)= -202.021

Obs	a	b	lik
1	-4.81	0.39	-202.021

Fig. 3. The same student as in Figure 2 showing her work comparing her own maximized value of the log likelihood function with that produced by SAS software. The hand-drawn arrows are added by this author

5 AN EXAMPLE OF FEASIBILITY: WEAVING THE R LANGUAGE INTO A PUBLIC HEALTH UNDERGRADUATE BIOSTATISTICS COURSE

Writing computer code is an integral part of my approach to teaching conceptual biostatistics. As the above example shows, this new approach engages students in a meaning-making process to assimilate the conceptual knowledge that would otherwise be inaccessible to them. Understandably, some educators may worry that weaving code writing into a biostatistics course in the way I advocated could distract students from learning biostatistics. In my experience, students can learn basic coding skills while simultaneously learning procedural and conceptual biostatistical knowledge. I put this idea into practice since 2004 when I started teaching an introductory biostatistics course called PHEB 602 to my school's graduate students. I adopted SAS software as the sole computing tool for the course, and students readily regarded learning basic coding skills as part of their biostatistics education. By the second homework assignment students were able to write SAS code to find sample percentiles (see Figure 4). More examples used in my PHEB 602 classes can be found in [17].

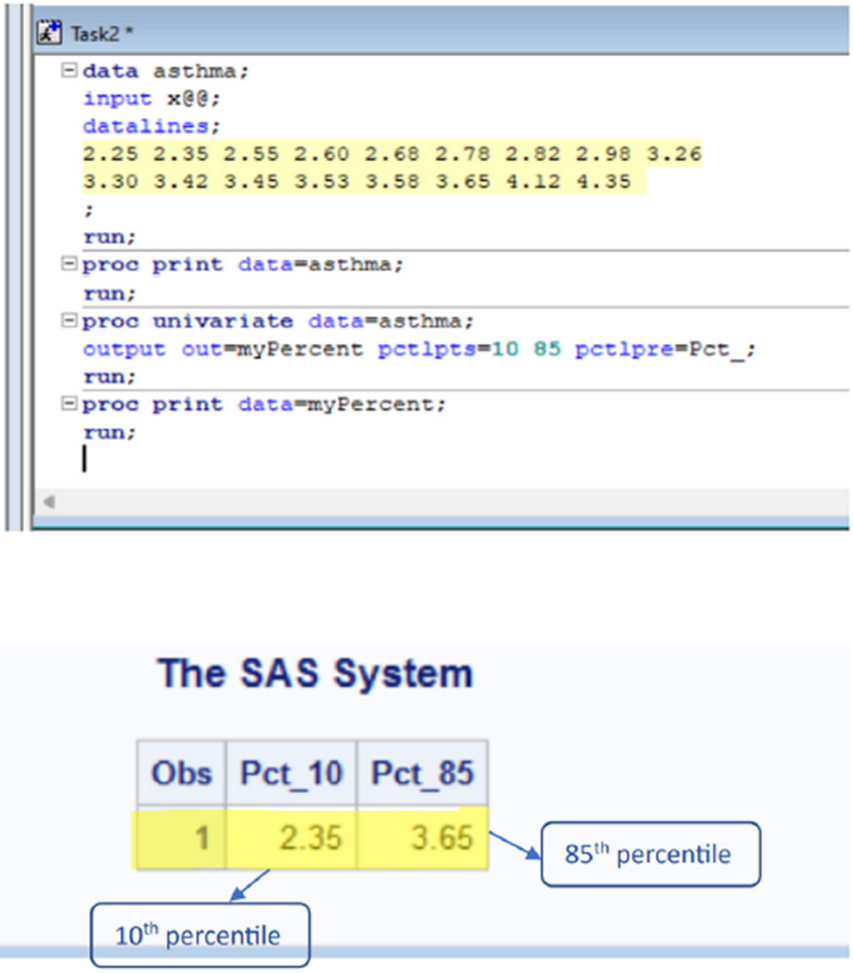

Fig. 4. Work of a PHEB 602 student showing how to find sample percentiles from data using SAS software

To demonstrate further the feasibility of my computing approach, I offer a few more cogent examples drawn from a biostatistics course for undergraduate public

health students, in which I adopted the R command-line computing environment [26] as the course software. Note that Gerbing [27] has previously designed innovative course content for an introductory statistics course for business students, in which students also used the R command-line environment [26] as their course software. In Gerbing's class, students successfully learned to execute command-lines to perform statistical computations, instead of relying on Excel worksheets. I offered my course called PHLT 315 In the fall 2023 semester. However, my approach differed from that of Gerbing in an important way. Gerbing's students interacted with the R computing environment mainly through a pedagogical R package that Gerbing wrote to reduce his students' cognitive burden of transitioning into the command-line R environment from Excel worksheets. While there were clear advantages of shielding beginning students from the overwhelming syntactic complexity of the command-line R environment by a pedagogical package, I found it more effective to introduce my students gently but directly to the genuine R environment at the start.

Undergraduate public health majors at my school take PHLT 314 and PHLT 315 for their biostatistics education. Students in my PHLT 315 class took PHLT 314 using Excel as their course software. However, they transitioned to R with little difficulty in my PHLT 315 class. The famous two-sample t-test given in equation (1) acted as a springboard for this transition. Students could easily call the R function t.test to conduct a two-sample t test, and they could also easily verify their results by executing a short sequence of R commands to perform the calculations prescribed by equation (1). (See Figures 5 and 6)

```
a=c (87, 86, 92, 68, 92, 78, 103, 87, 89, 90, 69)

b=c (110, 105, 63, 86, 82, 75, 96, 109, 84, 113, 76, 93)

>t.test(a,b,var.equal=T)


Two Sample t-test

data:  a and b

t = -0.96174, df = 21, p-value = 0.3471

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -17.249130   6.340039

sample estimates:

mean of x mean of y

 85.54545  91.00000
```

Conclusion: **We should not reject the null hypothesis because 0 falls between the 95% confidence interval of -17.25 and 6.34.**

Fig. 5. A student's work showing how she performs a two-sample t-test and draws a conclusion about the null hypothesis

```
> x=c(87, 86, 92, 68, 92, 78, 103, 87, 89, 90, 69)
> y=c(110, 105, 63, 86, 82, 75, 96, 109, 84, 113, 76, 93)
> xbar=mean(x)
> ybar=mean(y)
> c(xbar, ybar)
[1] 85.54545 91.00000
> sx2=var(x)
> sy2=var(y)
> c(sx2,sy2)
[1] 106.2727 255.8182
> sp2=(10*sx2+11*sy2)/(11+12-2)
> sp=sqrt(sp2)
> c(sp2, sp)
[1] 184.60606  13.58698
> myt=(xbar-ybar)/sp/sqrt(1/11+1/12)
> pval=pt(q=myt, df=11+12-2)*2
> c(myt, pval)
[1] -0.9617421  0.3471233
```

Fig. 6. A student's work showing how she performs a two-sample t test by following the definition given in equation (1). She highlights the t-value and the p-value that agree with their counterparts given by the R function t.test. (See Figure 5)

The above simple example illustrates an approach that I utilized to introduce students to the R command-line environment incrementally to minimize the chance of students being distracted from learning biostatistics. This approach is especially conducive to teaching conceptual biostatistics, which I further illustrate by the example of how students in my class learned the *F* test in the one-way analysis of variance procedure via computing exercises involving a data set consisting of 4 groups. Students first drew boxplots to visualize the data. (See Figure 7) Afterward, as Figure 8 shows, students put data from the four groups into y1, ..., y4 and computes the group means and the grand mean. Next, students calculated the between-group sum of squares SS_B and the within-group sum of squares SS_W according to their definitions. In the third step, students computed the *F* statistic according to equation (2) and obtained the corresponding P value. Finally, students invoked the R function *aov* to check their results.

```
> smoker.df=data.frame(treat=rep(c('nonsmoker','former','light','heavy'),each=7), fev=c(0.56,
0.85, 0.68, 0.96, 0.72, 0.92, 0.87, 0.58, 0.85, 0.72, 0.93, 0.86, 0.76, 0.65, 0.48, 0.76, 0.60, 0.82,
0.72, 0.60, 0.51, 0.27, 0.34, 0.62, 0.47, 0.70, 0.60, 0.54) )

> smoker.df$treat=factor(smoker.df$treat,levels=c('nonsmoker','former','light','heavy'))

> boxplot(fev~treat, data=smoker.df)
```

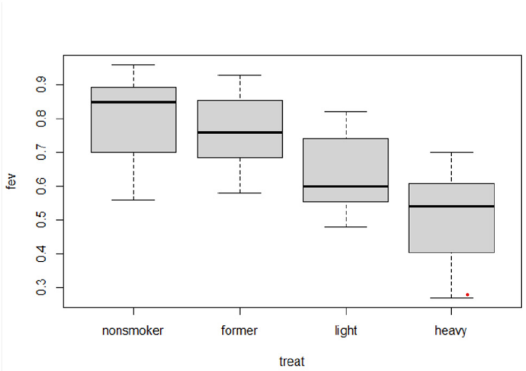

Fig. 7. Graph drawn by a student before performing an analysis of variance test on the data

Despite the limited R programming knowledge that I could cover in a two credit-hour undergraduate course that focus on biostatistics, students had exciting opportunities to see how their R knowledge could be applied to real-world research problems. Consider the following homework problem.

In the 1990s a prestigious medical journal published a study conducted to compare the efficacy of standard calcium supplement with that of calcitriol in treating postmenopausal osteoporosis. Some patients were forced to withdraw from the study by severe side effects. The data you need is tabulated below.

Treatment	Withdrawal		Total
	yes	no	
Calcitriol	27	287	314
Calcium	20	288	308
Total	47	575	622

Students were asked to find the relative risk (RR) and its 95% confidence interval by following standard formulas given by the textbook and by calling a function in the epidemiology R package epitools. Figure 9 shows a student's work.

The ability to execute correct R commands to accomplish a desired computational task as demonstrated by students in Gerbing's introductory statistics classes and in my PHLT 315 class (see Figures 5–9) supports my conviction that computing ability is universal. I infer that learning to use R commands falls within the ZPD of most undergraduate students. I further infer that the task of writing a small computer program such as the ones shown in Figures 2 and 4 falls within the ZPD of a typical public health student. The examples from my introductory biostatistics classes [17], categorical data analysis classes [21, 22], and longitudinal data analysis classes [19] support my claim, and hence the feasibility of my innovative computing approach to teaching conceptual biostatistical knowledge to public health students.

6 CONCLUDING REMARKS

My computing approach to teaching biostatistics shares important features with well-known pedagogical paradigms such as constructivism [8] and computational thinking [18]. What distinguishes my computing approach is its unique, computing-centered way to engage students in a mathematically gentle, computationally refreshing meaning-making process to develop students' conceptual understanding of biostatistics. In my teaching practice I have found this approach effective in teaching introductory biostatistics, categorial data analysis, longitudinal data analysis, and survival data analysis.

In championing the concept of innate mathematical ability, Colleran [3] proposed providing a more egalitarian mathematics education. In the same spirit, accepting the idea of innate computing ability, I propose democratizing conceptual biostatistical knowledge among public health students. The teaching approach I presented here suggests a new avenue not only for accomplishing that goal, but also for spreading computational thinking among public health students. In particular, I hope my work will raise awareness that computing is not just for data analysis as it can also be a powerful tool for teaching conceptual knowledge of biostatistics. In a sense, my novel approach provides a unique example of an idea advocated

by Nolan and Lang [28] that "statistical computing could be taught in a way that offers an alternative approach to teaching statistical concepts (both elementary and advanced)."

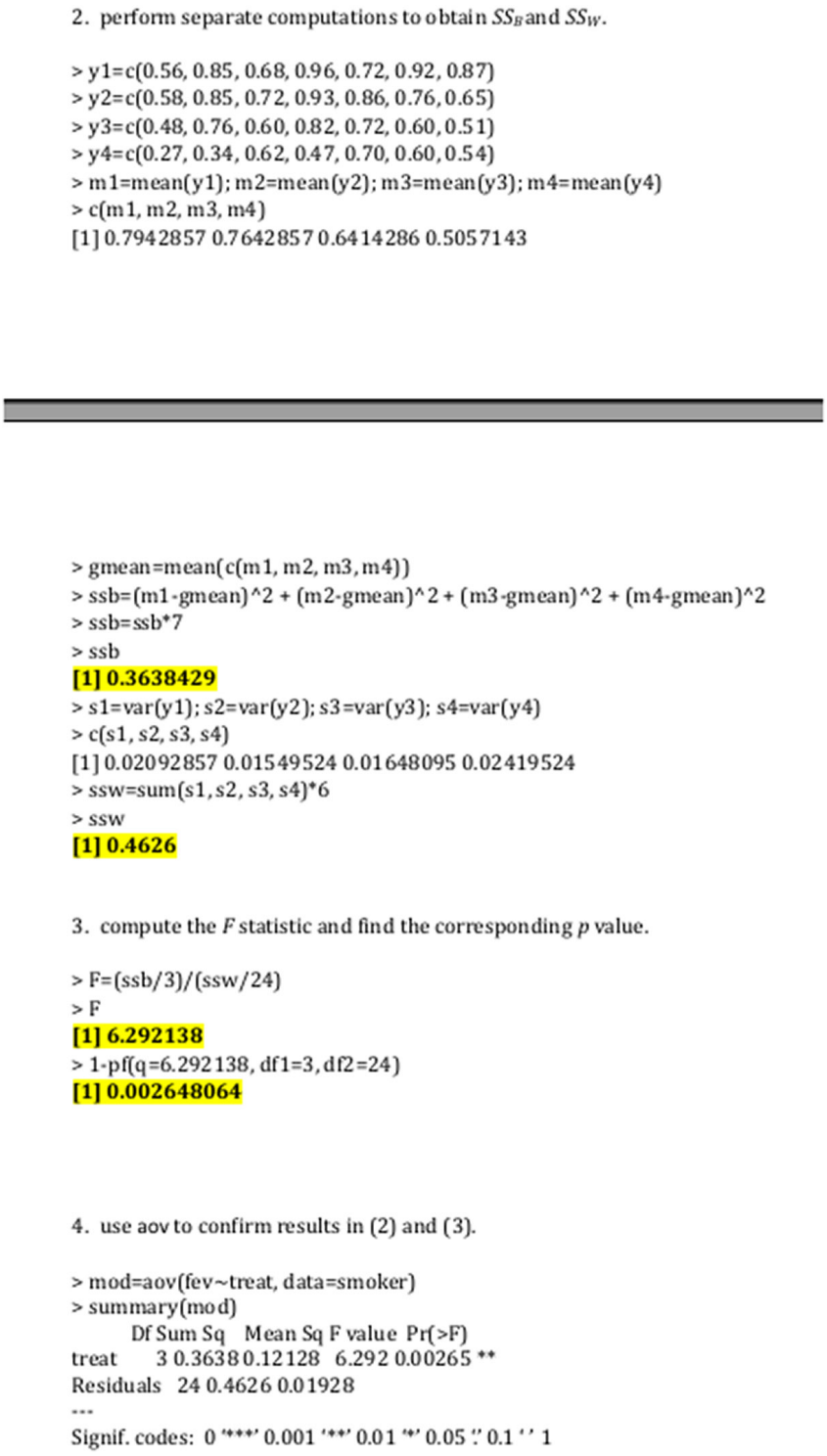

Fig. 8. Work by an undergraduate student in my PHLT 315 class showing how the *F* statistic is computed using the R computing language. The student first obtains SS_B and SS_W and computes the *F* statistic according to equation (2). She then finds the P-value. Finally, she compares her results with the output from the R function *aov*

```
> p1=27/314
> p1
[1] 0.08598726
> p2=20/308
> p2
[1] 0.06493506
> rr=p1/p2
> rr
[1] 1.324204
```

```
> center=log(rr)
> center
[1] 0.2808114
> se=sqrt((1/27)-(1/314)+(1/20)-(1/308))
> se
[1] 0.2839112
> low=center-1.96*se
> up=center+1.96*se
> c(low, up)
[1] -0.2756546  0.8372773
> c(exp(low), exp(up))
[1] 0.7590751 2.3100689
```

```
> tab=matrix(c(288, 20, 287, 27), byrow=T, ncol=2)
> tab
     [,1] [,2]
[1,]  288   20
[2,]  287   27
> riskratio(tab,conf.level=0.95)
$data
          Outcome
Predictor  Disease1 Disease2 Total
  Exposed1     288      20   308
  Exposed2     287      27   314
  Total        575      47   622
$measure
          risk ratio with 95% C.I.
Predictor  estimate    lower    upper
  Exposed1 1.000000       NA       NA
  Exposed2 1.324204 0.7590829 2.310045
```

Fig. 9. Upper panels: A student in my PHLT 315 class calculates the relative risk and its 95% confidence interval by following standard formulas; lower panel: the same student invokes the *risk ratio* function in the epitools R package to verify her results

7 REFERENCES

[1] Q. Zheng, "Bringing conceptual knowledge of biostatistics into the zone of proximal development via integrated computing exercise," in *Creative Approaches to Technology-Enhanced Learning for the Workplace and Higher Education. TLIC 2024*, in Lecture Notes in Networks and Systems, D. Guralnick *et al.*, Eds., 2024, pp. 1–11. https://link.springer.com/chapter/10.1007/978-3-031-73427-4_29

[2] P. Tyre, "The math revolution," The Atlantic, March 2016. https://www.theatlantic.com/magazine/archive/2016/03/the-math-revolution/426855/

[3] N. Colleran, "Exploring the genealogy of the concept of 'innate mathematical ability' and its potential for an egalitarian approach to mathematics education," *Adults Learning Mathematica: An International Journal*, vol. 13, no. 1, pp. 25–37, 2018. https://files.eric.ed.gov/fulltext/EJ1192152.pdf
[4] B. Butterworth, *The Mathematical Brain*. London: Macmillan, 1999.
[5] A. V. Borovik and T. Gardiner, "Mathematical abilities and mathematical skills," MIMS EPrint, p. 109, 2007. https://eprints.maths.manchester.ac.uk/839/
[6] B. Butterworth, C. R. Gallistel, and G. Vallortigara, "Introduction: The origins of numerical abilities," *Phil. Trans. R. Soc. B*, vol. 373, p. 20160507, 2018. https://doi.org/10.1098/rstb.2016.0507
[7] J. Boaler, J. A. Dieckmann, G. Pérez-Núñez, K. L. Sun, and C. Williams, "Changing students minds and achievement in mathematics: The impact of a free online student course," *Frontiers in Education*, vol. 3, p. 26, 2018. https://doi.org/10.3389/feduc.2018.00026
[8] J. M. Wing, "Computational thinking and thinking about computing," *Philosophical Transactions of the Royal Society A*, vol. 366, pp. 3717–3725, 2008. https://doi.org/10.1098/rsta.2008.0118
[9] S. Grover and R. Pea, "Computational thinking: A competency whose time has come," in *Computer Science Education Perspective on Teaching and Learning in School*, S. Sentance, E. Barendsen, and C. Schulte, Eds., Bloomsbury Academic, 2018, pp. 19–38. https://doi.org/10.5040/9781350057142.ch-003
[10] L. Rosenstock *et al.*, "Confronting the public health workforce crisis: ASPH statement on the public health workforce," *Public Health Reports*, vol. 12, no. 3, pp. 395–398, 2008. https://www.jstor.org/stable/20723357
[11] J. P. Leider *et al.*, "Staffing up and sustaining the public health workforce," *Journal of Public Health Management and Practice*, vol. 29, no. 3, pp. E100–E107, 2022. https://doi.org/10.1097/PHH.0000000000001614
[12] J. C. Karran, E. E. M. Moodie, and M. P. Wallace, "Statistical method use in public health research," *Scandinavian Journal of Public Health*, vol. 43, no. 7, pp. 776–782, 2015. https://doi.org/10.1177/1403494815592735
[13] S. Kunkle, G. Christie, D. Yach, and A. M. El-Sayed, "The importance of computer science for public health training: An opportunity and call to action," *JMIR Public Health Surveillance*, vol. 2, no. 1, p. e10, 2016. https://doi.org/10.2196/publichealth.5018
[14] A. M. Brearley, K. W. Rott, and L. J. Le, "A biostatistical literacy course: Teaching medical and public health professionals to read and interpret statistics in the published literature," *Journal of Statistics and Data Science Education*, vol. 31, no. 3, pp. 286–294, 2023. https://doi.org/10.1080/26939169.2023.2165987
[15] M. J. Hayat, A. Powell, T. Johnson, and B. L. Cadwell, "Statistical methods used in the public health literature and implications for training of public health professionals," *PLoS ONE*, vol. 12, no. 6, p. e0179032, 2017. https://doi.org/10.1371/journal.pone.0179032
[16] B. Conway IV, W. G. Martin, M. Strutchens, M. Kraska, and H. Huang, "The statistical reasoning learning environment: A comparison of students' statistical reasoning ability," *Journal of Statistics and Data Science Education*, vol. 27, no. 3, pp. 171–187, 2019. https://doi.org/10.1080/10691898.2019.1647008
[17] Q. Zheng, "Let computational thinking permeate biostatistics education of public health students," in *The 6th International Conference on Distance Education and Learning*, Association for Computing Machinery, 2021, pp. 283–288. https://doi.org/10.1145/3474995.3475043
[18] B. Rosner, *Fundamentals of Biostatistics*, 8th ed. Boston, MA: Cengage Learning, 2016.
[19] Q. Zheng, "Integrating computational thinking into a longitudinal data analysis course for public health students," *Discover Education*, vol. 1, p. 15, 2022. https://doi.org/10.1007/s44217-022-00015-w

[20] E. von Glaserfeld, "A constructivist approach to teaching," in *Constructivism in Education*, L. Steffe and J. Gale, Eds., The University of Georgia, Atlanta: Erlbaum, 1995, pp. 3–16.
[21] Q. Zheng, "Let master of public health students experience statistical reasoning," *Athens Journal of Health and Medical Sciences*, vol. 7, no. 1, pp. 47–62, 2020. https://doi.org/10.30958/ajhms.7-1-4
[22] Q. Zheng, "Improving the teaching of biostatistics in an online master degree program in epidemiology," in *Proceedings of the 5th International Conference on Distance Education and Learning*, Association for Computing Machinery, 2020, pp. 89–93. https://doi.org/10.1145/3402569.3402582
[23] L. S. Vogotsky, *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press, 1978.
[24] L. E. Margulieux, B. B. Morrison, and A. Decker, "Reducing withdrawal and failure rates in introductory programming with subgoal labeled worked examples," *International Journal of STEM Education*, vol. 7, p. 19, 2020. https://doi.org/10.1186/s40594-020-00222-7
[25] SAS Institute Inc. SAS/STAT Software, Version 9.4. Cary, NC, 2016.
[26] R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2023. https://www.R-project.org
[27] D. W. Gerbing, "Enhancement of the command-line environment for use in the introductory statistics course and beyond," *Journal of Statistics and Data Science Education*, vol. 29, no. 3, pp. 251–266, 2021. https://doi.org/10.1080/26939169.2021.1999871
[28] N. Nolan and D. T. Lang, "Computing in the statistics curricula," *The American Statistician*, vol. 94, no. 2, pp. 97–107, 2010. https://doi.org/10.1198/tast.2010.09132

8 AUTHOR

Qi Zheng is with the Texas A&M School of Public Health, College Station, Texas 77843, USA (E-mail: qzheng@tamu.edu).