Improving Statistical Literacy for Physician Scientists: Sampling Distributions

Jesse D. Troy[1,*], Caroline Falvey[1], Suzanne Angermeier[1], Gina-Maria Pomann[1], Steven C. Grambow[1], Megan L. Neely[1] & Gregory P. Samsa[1]

[1]Department of Biostatistics and Bioinformatics, School of Medicine, Duke University, Durham, NC, USA

*Correspondence: Department of Biostatistics and Bioinformatics, School of Medicine, Duke University, Durham, NC, USA. Tel: 1-919-668-2932. E-mail: jesse.troy@duke.edu

Received: January 29, 2024 Accepted: September 11, 2024 Online Published: November 15, 2024

doi:10.5430/jct.v13n5p371 URL: https://doi.org/10.5430/jct.v13n5p371

Abstract

This is the first in a multi-part series on teaching statistical inference to physician-scientists training to work as members of interdisciplinary scientific teams. This unique student audience has greater scientific sophistication than a typical statistics student but less background in mathematics and computer programming, which presents challenges for traditional approaches to teaching statistics. Here, we illustrate an innovative approach to teaching sampling distributions to physician-scientists. Sampling distributions are a fundamental element of statistical inference; they are a building block of confidence intervals and hypothesis tests which are vital tools for performing clinical research. As such, it is essential that physician-scientists have a strong foundation in sampling distributions. Key elements of our innovative approach include the use of a running example, delivery of content in small pieces to reduce cognitive burden, preceding formulae with pictures, combining static and dynamic content using an R Shiny app, and use of self-graded quizzes to provide immediate feedback. The resulting course module can be reused in multiple contexts, including as part of self-directed, asynchronous learning or by incorporation into a traditional or flipped classroom setting.

Keywords: constructivism, sampling distributions, statistical education, statistical inference

1. Introduction

We teach statistics in multiple settings at an academic medical center, including a master's degree in biostatistics (Neely et al., 2022; Troy, Granek, et al., 2022), which is a professional training program for future biostatisticians and the Clinical Research Training Program (CRTP), which is a professional Master of Health Sciences degree program for physician-scientists who are medical fellows or early-career faculty members (Armstrong et al., 2009; G. Samsa & Oddone, 1994; Wilkinson & Oddone, 2002). CRTP students are training to engage in interdisciplinary team science, including interpreting statistical analyses performed by other team members, such as biostatisticians (G. P. Samsa, 2018). As such, CRTP students require instruction in statistics but not at the same depth as biostatistics graduate students (Enders et al., 2017). Importantly, CRTP students have a background in biological sciences, are practitioners in subspecialty areas of medicine, and typically lack the rigorous mathematical preparation of graduate students in biostatistics. Therefore, our goals and educational approach to teaching classical statistical inference to CRTP students differ from those of biostatistics graduate students in our master's program. In particular, the pedagogic objectives for CRTP students place reduced emphasis on the mathematical basis for statistical methods or how to implement statistical analyses and more emphasis on the appropriateness of the statistical techniques selected for a given research problem and how to interpret results (Garfield & Ben-Zvi, 2007). Indeed, as discussed later, CRTP students are exemplars of a much broader audience of non-statitician scientists spanning the biological and life sciences, and our approach is also intended to generalize to this broader audience. For now, we will focus on CRTP students.

Designing effective curricula for CRTP students requires navigating two related pedagogic challenges. The first challenge is how to modify the traditional biostatistics curriculum, which heavily relies upon mathematics and derivations (Troy, McCormack, et al., 2022), to better fit a target audience whose prior training, in comparison with traditional students of statistics, did not focus in these areas. The second challenge is determining how much students

should be taught to "do" in addition to "interpret". Even if students will subsequently delegate statistical analyses to others, some level of active, hands-on engagement will (1) help clarify their understanding as statistical content is being learned; and (2) place them in a better position to eventually take on the role of first and/or senior author, in discharging their responsibility to verify that the data and findings are accurate. Determining how to organize this active engagement isn't trivial since students' relative weakness in mathematics and computer programming complicates some of the usual approaches. We illustrate here a specific example of how this might be achieved using sampling distributions as a use case. Sampling distributions are a key concept in statistical inference as multiple other concepts directly build upon them, e.g., confidence intervals and hypothesis testing. Understanding these fundamental concepts in statistical inference has been identified as a core competency for medical researchers (Enders et al., 2017).

We begin our discussion of this topic by defining sampling distributions and highlighting the foundational nature of understanding this topic as a prerequisite to learning the classical approaches to statistical inference, which include interval estimation and hypothesis testing. We then explain why sampling distributions are challenging to teach (Findley & Lyford, 2019), after which we present an innovative approach to teaching sampling distributions that addresses these challenges. Finally, we comment on how the pedagogic principles illustrated by this use case can be applied more generally.

2. Sampling Distributions and their Connection to Statistical Inference

From the frequentist perspective, classical statistical inference fundamentally relies on a thought experiment: what would happen if a study were repeated multiple times (Lindley, 2000)? For example, suppose that the goal of a study is to estimate the mean low-density lipoprotein (LDL) level in a population. This might be done, for example, to assess whether the population is at risk for cardiovascular disease, which could be indicated by a mean LDL level of 140 mg/dL or higher (*Cholesterol Numbers and What They Mean*, n.d.). Suppose a random sample of 100 individuals is taken from the population, a mean LDL is obtained from that sample, and a standard deviation quantifies the variation in LDL values across individuals. In practice, the analyst does not have the luxury of observing the results of multiple studies and must make an estimate of the population mean LDL based on this single sample alone. However, if the analyst had performed multiple studies, each with a sample size of 100, they could have calculated the mean from the first study, $\overline{x}_1$, the mean from the second study, $\overline{x}_2$, and so forth. Each of these means could become a single point which is plotted as part of a histogram that describes the distribution of sample means across the multiple studies. This histogram is a visual representation of a sampling distribution: more precisely, the sampling distribution of the means derived from multiple studies, each with a sample size of 100.

If the histogram representing this sampling distribution is very tight (i.e., with a small standard deviation), then all sample means will be very similar. Accordingly, the sample mean observed in the study will be very close to the (unknown) mean LDL in the population under study. Similarly, if the histogram representing this sampling distribution is very spread out, then the mean observed in the performed study provides little information about the population mean. It turns out that the width (or variance) of the sampling distribution is related to both the sample size and the standard deviation in the population. This is how the sampling distribution links a single sample with inference about the characteristics of the population as a whole. Indeed, our instructional materials are organized around visual representations of this concept.

Understanding sampling distributions is a prerequisite for learning about statistical inference generally (Kula & Koçer, 2020). As an example, we discuss the connection between sampling distributions and two related constructs: interval estimation and hypothesis testing. In the above LDL example, the mean LDL of the population is unknown, and the analytic task is to obtain an estimate. The sample mean (i.e., the mean of the LDL measurements for the 100 individuals who were studied) provides a natural "point estimate" of the unknown population "parameter". A confidence interval provides a range of plausible values. Point estimates, while helpful, are not the only important information contained in the sample about the population. If the point estimate in question—the sample mean LDL—came from a sampling distribution where all the sample means were clustered close together, we might guess that the point estimate is more precise than if it had come from a sampling distribution where there is more variability in the sampling means. This leads to the related idea of a confidence interval, which is typically reported along with the point estimate to communicate the amount of certainty (or lack thereof) in the point estimate.

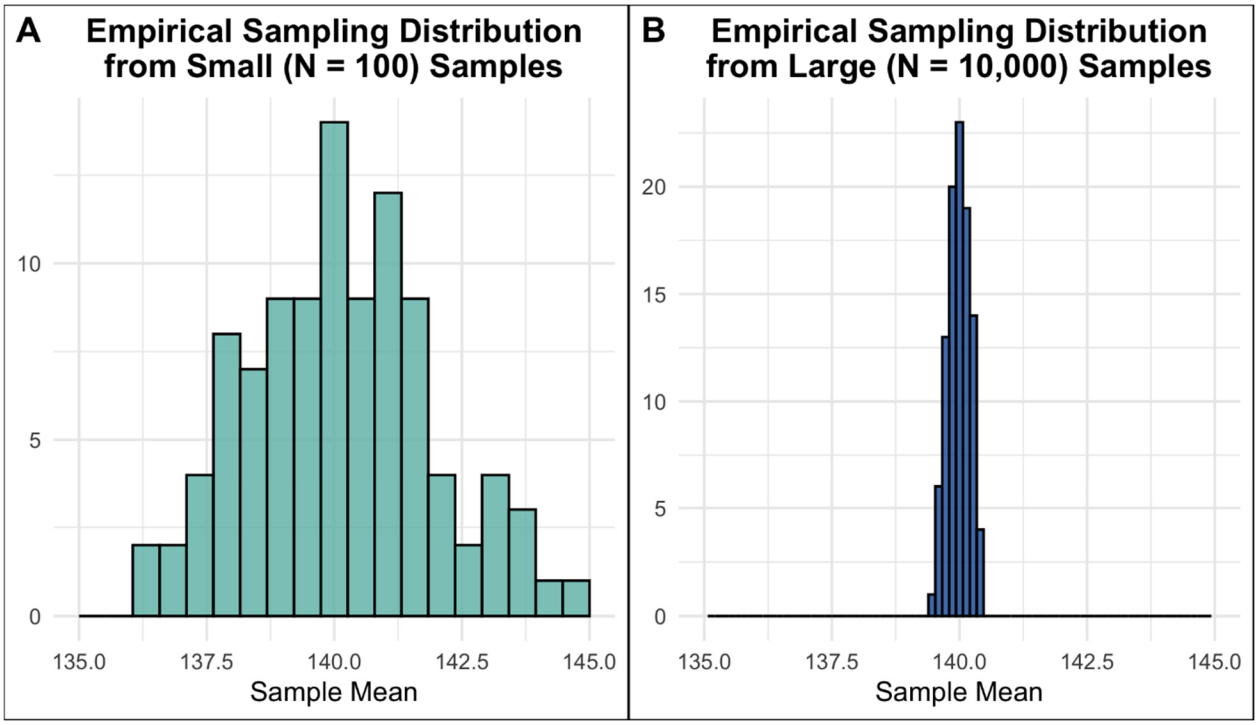


Figure 1. Empirical Sampling Distributions from Small (N=100) and Large (N=10,000) Samples of the Same Population

Description: Both plots show histograms of the sample mean from repeated samples of a normally distributed population with a mean 140 and standard deviation of 20. Panel (A) shows a simulation of the sampling distribution when the sample size is 100. Panel (B) shows a simulation of the sampling distribution for the sample mean when the sample size is 10,000. In panel (B), there is less variation in the sample means than in panel (A).

A firm grasp of sampling distributions is also required to interpret hypothesis tests. A hypothesis test is a formal decision-making procedure to assess whether the data are consistent with a pre-specified mean (or means). For example, an investigator might ask whether the samples are consistent with a population mean of 140 or less. If the sample mean is near 140 (or below), this hypothesis is supported, whereas if the sample mean is far above 140, this hypothesis is not. Again, the variability in the sampling distribution for the sample mean is a key factor. Thus, understanding the behavior of the repeated samples from the same population is a prerequisite for understanding a commonly used statistical procedure.

Although this is not an exhaustive list of reasons students will benefit from understanding sampling distributions, the examples we give here illustrate the central role of this construct in understanding statistical analyses. Importantly, sampling distributions must be understood by the biostatistician and the non-statistician team scientist working with the biostatistician. Thus, the sampling distribution represents an intersection between the mental maps for biostatistics graduate students and CRTP students. However, as discussed below, certain challenges are inherent to teaching sampling distributions for the CRTP student audience. Furthermore, we argue that the details of what the CRTP students must know are a subset of the details that biostatistics graduate students know.

3. Challenges in Teaching Sampling Distributions and the Inadequacy of the Traditional Instructional Approach for Non-Statistician Scientists

One challenge of teaching sampling distributions is that they are based on a thought experiment and are not directly observed. Another challenge is that they touch upon several related concepts, not all of which have been mastered (or necessarily encountered) when sampling distributions are introduced. For example, the above presentation utilizes two distinct standard deviations (i.e., representing the variability of individual patients and the variability of the means of repeated samples), and students often confuse the two (or fail to recognize the distinction). The result is a

tangled thicket of ideas, none of which are necessarily challenging when considered individually but risk using undefined terms and notions, especially at the start. To extend the metaphor, the first step into this thicket is the most dangerous and requires special care. Moreover, the optimal first step might depend on the audience.

The default masters-level curriculum in a program for biostatisticians is fundamentally derived from a mathematically rigorous PhD-facing curriculum -- beginning with that curriculum and then removing some of the especially esoteric elements (Troy, McCormack, et al., 2022). The default curriculum in a program such as CRTP removes even more elements. In both cases, the intention is to remove elements of the curriculum for which students are poorly prepared and are not relevant to their learning needs. For example, CRTP students are typically not facile with mathematical proofs, and most will rely on quatitative partners to perform statistical analyses during their careers in clinical research.

At face value, removing theoretical elements from statistical curriculum developed for PhD-facing biostatistical programs seems reasonable. First, it removes content CRTP students are not prepared for and most likely will never need. Second, the instructor has a place to start: namely, a traditional course with which they are already familiar. However, developing a curriculum in this manner has some pedagogic flaws. Much like a game of Jenga, removing too many blocks from an existing curriculum may result in a curriculum with a shaky foundation (Figure 2). Its demarcation point is an existing curriculum developed from premises that don't necessarily apply to the student population under consideration. The result isn't tailored toward the characteristics of the student audience, including their learning goals. Dropping elements from an existing curriculum can cause that curriculum to lose coherence. We propose instead that CRTP students deserve a new curriculum built from the ground up. By doing so, we can effectively communicate to non-statistician scientists the basic principles of statistical thinking in a way that fits well with their objectives of solving problems in clinical medicine.

Figure 2. A Jenga Tower Collapsing

Description: Jenga is a game in which players attempt to remove blocks from a tower without collapsing the tower. See https://www.jenga.com/ for more information. This image was generated by the authors of this article using DALL-E, https://openai.com/research/dall-e.

4. Goals and Pedagogic Philosophy for A Revised Approach to Teaching Sampling Distributions for Non-Statistician Scientists

Our instructional goals for CRTP students are based on scientific tasks that utilize them as building blocks. More specifically, we want CRTP students to understand sampling distributions in sufficient detail to (1) interpret confidence intervals; and (2) interpret hypothesis tests. Each goal can be framed as a task related to writing a manuscript for which their collaborating statistician has provided input, and the clinical investigator's task is to use this information properly.

An additional instructional goal is to familiarize CRTP students with some of the mental maps used by the biostatisticians with whom they will collaborate (Pomann et al., 2020). Some of these mental maps include:

- Statistical inference involves a thought experiment where a study is repeated many times. The result in a sampling distribution.
- Confidence intervals and hypothesis tests are similar, as they quantify which values of an underlying population parameter are sufficiently consistent with the observed data.
- Many things in statistics have a normal distribution, including sampling distributions (under many circumstances).

- Statistical inference requires a handoff between deduction and induction.

Describing the mental maps used in statistics--even though biostatisticians primarily execute while clinical investigators primarily interpret--is (among others) intended to assist in communication across disciplines.

5. Approach to Teaching Sampling Distributions

Our teaching materials are shown in Appendix 1 and include static content developed in R Markdown (Grayson et al., 2022) and dynamic content developed in R Shiny (Arnholt, 2019). The dynamic content can be accessed here: https://github.com/megan-neely/JCT_Teaching_Sampling_Distributions_RShinyApp. The teaching materials are amenable for use in a traditional classroom or deployment as part of a self-paced, asynchronous, online learning module. The material on sampling distributions illustrates a number of pedagogic features, as discussed below.

5.1 Use of a Running Example

The lesson is taught using the example of estimating the mean LDL cholesterol level of the patient population served by a single medical clinic. This example is accessible to the student audience because it uses a common health outcome measure with a defined clinical significance threshold that indicates the risk of cardiovascular disease.

5.2 Delivery of Content in Small Pieces to Reduce Cognitive Burden

Sampling distributions are a high-level construct that requires a base knowledge of several fundamental statistical concepts, including: 1) a basic understanding of probability, 2) the concepts of random variables and population distributions, and 3) measures of central tendency and variation. Our lesson on sampling distributions begins with these prerequisite concepts and builds slowly toward an intuitive understanding of sampling distributions that ends with the mathematical details expressed at an appropriate level for the CRTP student audience.

The lesson includes 29 distinct items (Table 1), which are meant to be followed in order. The items build in complexity from start to finish, where each step requires only minimal cognitive engagement. As concepts are introduced in items 1-7, the static content relies primarily on visual presentations with limited amounts of text and little to no mathematical symbols. As the complexity of the material increases, additional text accompanies graphical images as with items 2-12 and 13-15. Mathematical notation is brought to the forefront only in the later parts of the lesson, in items 16-26.

Table 1. Summary of the Lesson on Sampling Distributions

Items	Topics Covered	Static Content Presentation Methods
1-7	• Definition of population distributions • Introduction to the concepts of central tendency and variation • Visual impact of changing the mean and standard deviation	• Visual
2-12	• Repeated sampling from the population	• Visual
13-15	• Distribution of sample means vs. the distribution of the population	• Visual • Brief textual summary of items 1-14
16-26	• Reiteration of concepts covered in items 1-15 but using symbols	• Using the language of mathematics
27-29	• Summary linking sampling distributions to other topics in statistical inference	• Text only

Description: This table summarizes the sampling distribution lesson illustrated in detail in Appendix A.

5.3 Pictures First, Formulae Second

Recognizing that CRTP students do not necessarily find symbol manipulation helpful, we begin with pictures and only when the idea behind the concept has been illustrated, add formulae.

5.4 Static Content is Combined with Dynamic Content, Allowing Students to Actively Engage with the Material

The dynamic content in the R Shiny app mimics the static content described in Table 1 and shown in Appendix 1. However, the R Shiny app allows students to change features of the graphical presentations that are fixed in the static content. For example, in item 2 the static content shows how the population changes shape as the standard deviation

increases from 1 to 2. The dynamic version allows the student to modify the standard deviation to any non-zero value and observe immediate changes to the shape of the distribution. Importantly, students can use the static and dynamic content simultaneously or in lockstep according to their learning styles.

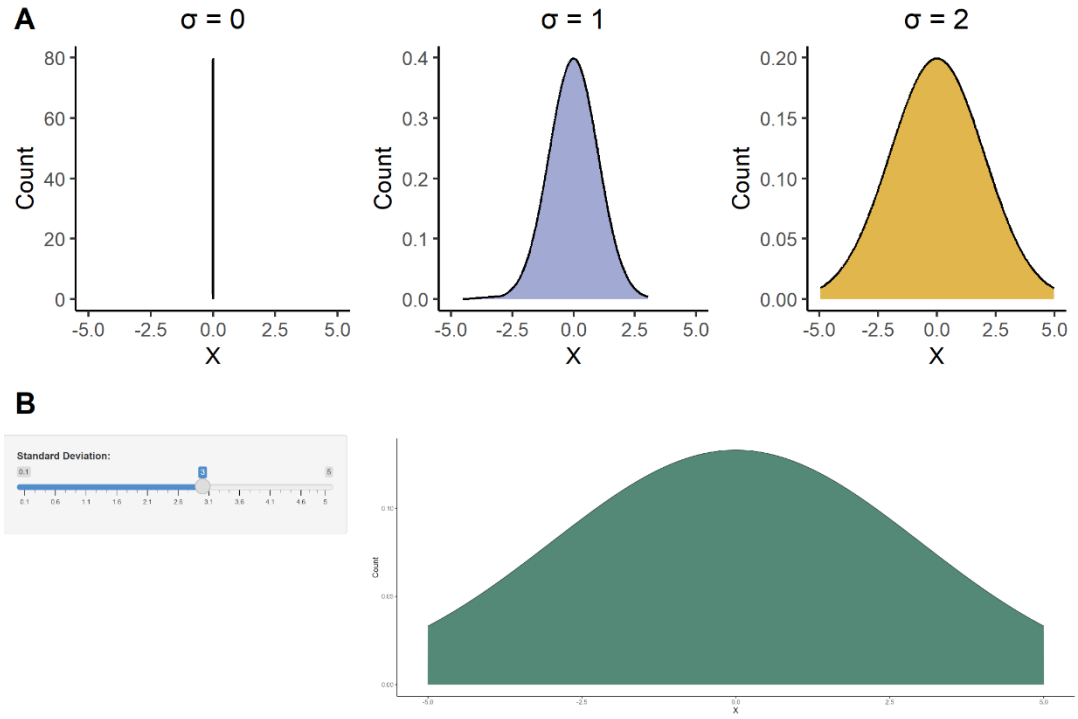


Figure 3. Example Alignment Between Static and Dynamic Content

Description: An example of static content showing how in Panel (A), the student can see three examples of normal distributions with different standard deviations. The corresponding dynamic content from the R Shiny app is shown in panel (B). The app gives the student the ability to change the standard deviation using a slider widget and visualize how the distribution changes shape in response.

5.5 Automatically Graded Quizzes Provide Immediate Feedback about Mastery of the Content

Table 2. Example Quiz Questions and Answers

	Example Question	Answer	Commentary
1.	What is the impact on the mean and standard error of the sampling distribution of decreasing the standard deviation of the underlying distribution?	The mean is unchanged, the standard error is decreased.	This is one of two ways to increase the precision of the sampling distribution (see question 2).
2.	What is the impact on the mean and standard error of the sampling distribution of increasing the sample size?	The mean is unchanged, the standard error is decreased.	This is the other way to increase the precision of the sampling distribution.
3.	Which decreases with increasing sample size: the standard deviation of the underlying variable or the standard error of the sampling distribution?	The standard error of the sampling distribution.	This highlights that σ is a fixed characteristic of the variable being studied, and thus is unaffected by sample size.

The quizzes are designed to be narrow in scope and to have low cognitive burden, similar to how the material is

presented (see Table 1) and are intended to be completed as students progress through the material. By testing knowledge as they go, students can gain confidence that they have understood the material and identify points where assistance from an instructor could be helpful. Examples are shown in Table 2.

6. Discussion

We have described an approach for teaching sampling distributions, a topic which is (1) key to statistical inference; and (2) often difficult for clinical audiences such as CRTP students. A conceptual challenge pertains to the level of abstraction since sampling distributions are based on a thought experiment and not directly observed. A pedagogic challenge is that sampling distributions are part of a thicket of related ideas, and a thoughtful first step into this thicket is required. Another pedagogic challenge pertains to how to best tailor the presentation to CRTP students, for whom neither mathematical formulae nor computer programming are relative areas of strength.

Educational practice increasingly relies upon tutorial modules as a supplement to traditional lectures, e.g., using a flipped classroom format (Sohrabi & Iraj, 2016). An advantage for students of the materials in Appendix A is that they can be accessed multiple times at the students' convenience. An advantage for instructors is scalability. For example, the same module can be offered to multiple students and, once developed, can be used in multiple classes. Regardless of whether instructors are available to answer questions offline, a key to success is for the module to be intuitive in real time -- a question delayed can become a question unasked. Key aspects supporting intuition for these modules include transparent organization and the liberal use of pictures.

Indeed, although it might initially seem to be a small thing, we begin with pictures rather than formulae. Content is organized around developing intuition through pictures, and only then following up with formulae once the concept in question is understood. This contrasts with the traditional presentation in a biostatistics course that begins with and is organized around, formulae. Consistent with the tenets of constructivism (Biggs, 1996), intuition around the picture is enhanced by active engagement. For example, the dynamic content encourages the student to manipulate inputs and observe their impact on the outputs, and static content provides one example of this and serves as a springboard for students to engage with the dynamic content on their own. Also consistent with these tenets, the underlying statistical principles are translated into plain English as well as explicit mental maps of how statisticians conceptualize these topics (Wilkerson-Jerde & Wilensky, 2011; Wilkerson, 2008). Thus, even if CRTP students do not implement these tasks in practice during their future careers in clinical research, we hope that the introduction of these mental maps will help to establish a common language with their statistical colleagues, which in turn will help to support effective interdisciplinary team science (Enders et al., 2017). One example is the ability to properly interpret a confidence interval (Enders et al., 2017). CRTP students need not understand how to produce a confidence interval but an understanding of the interval rooted in statistical thinking holds promise for improving interdisciplinary communications.

Another feature of our approach is that it is consistent with the "see one, do one, teach one" paradigm by which CRTP students typically encounter clinical content (Ayub, 2022) during their medical training. In the clinical context, the "see one, do one" component involves observing and then practicing clinical skills on patients of increasing complexity. Hands-on examples to demonstrate mastery of "doing" are especially crucial for this group of students (e.g., a habit of mind developed during medical school). "Teach one" is shorthand for deeper reflection, creating generalizable knowledge, extending from individual patients to populations, etc.; in essence, for building their mental map of the topic being mastered. By analogy, "see one" corresponds to our emphasis on pictures, "do one" corresponds to our hands-on examples, and "teach one" corresponds to explicit emphasis on the underlying mental maps.

It is, of course, natural to ask whether this approach to teaching statistics should be limited to an audience with the characteristics of CRTP students -- in other words, student populations with less tolerance for symbol manipulation and expertise in computer programming (and greater scientific sophistication) than a typical student of statistics. We argue that it need not be and, indeed, research suggests that even professional mathematicians initially learn through pictures, analogy, trial and error, etc., to describe the underlying ideas and only place the results into rigorous symbolic formalism as a final step (Wilkerson-Jerde & Wilensky, 2011; Wilkerson, 2008). To carry the Jenga analogy forward, we argue that one could build a set of bespoke Jenga towers, one for each student population, where each tower has a common foundation built out of the blocks described in this article. Additional levels could be added to towers for student populations that need more advanced training. Building Jenga towers in this way would have the added benefit of providing all student populations, who may interact one day in their careers in interdisclipinary team, with the same foundation and, thus, the same language with which to discuss these topics. For

example, to extend this presentation to graduate students in biostatistics training programs, these modules might be used as an initial introduction and supplemented with (1) having students program examples and (2) a more extensive translation into symbols and general statistical concepts. As an example of the latter, the module on sampling distributions relies on the "Central Limit Theorem," which, among others, provides the rationale for assuming that, under general conditions, the sampling distribution of sample means approaches normality as the sample size increases. For biostatisticians in training, the Central Limit Theorem can be stated more precisely and perhaps even illustrated with a rigorous proof. To extend this presentation to other biomedical scientists, the same structure could be maintained and the LDL example replaced with another that is more specific to their discipline. Alternatively, the LDL example could be retained, as it is conceptually straightforward and ought to be accessible to a broad audience, and supplemented by a discipline-specific illustration.

7. Conclusion

Effective communication is key to the successful practice of interdisciplinary team science (Pomann et al., 2020). The concept of overlapping mental maps can provide a common understanding that enables this effective communication (G. P. Samsa, 2018; Troy et al., 2021). However, finding the intersection of mental maps can be difficult when the parties who must communicate have different technical backgrounds and scientific training. The approach we propose here for teaching sampling distributions to non-statisticians could be considered an approach to establishing that critical intersection between mental maps for clinical investigators and biostatisticians. Our hope is that this intersection of mental maps will give CRTP students an intuitive understanding of statistics that they can apply in their collaborative work, particularly with respect to interepretation of research results and design of research studies.

References

Armstrong, A. Y., DeCherney, A., Leppert, P., Rebar, R., & Maddox, Y. T. (2009). Keeping clinicians in clinical research: the Clinical Research/Reproductive Scientist Training Program. *Fertility and Sterility*, *91*(3), 664-666. https://doi.org/10.1016/j.fertnstert.2008.10.029

Arnholt, A. T. (2019). Using a Shiny app to teach the concept of power. *Teaching Statistics*, *41*(3), 79-84. https://doi.org/10.1111/test.12186

Ayub, S. M. (2022). “See one, do one, teach one”: Balancing patient care and surgical training in an emergency trauma department. *Journal of Global Health*, *12*. https://doi.org/10.7189/JOGH.12.03051

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347-364. https://doi.org/10.1007/BF00138871

Cholesterol Numbers and What They Mean. (n.d.). Cleveland Clinic. Retrieved January 11, 2024, from https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean

Enders, F. T., Lindsell, C. J., Welty, L. J., Benn, E. K. T., Perkins, S. M., Mayo, M. S., Rahbar, M. H., Kidwell, K. M., Thurston, S. W., Spratt, H., Grambow, S. C., Larson, J., Carter, R. E., Pollock, B. H., & Oster, R. A. (2017). Statistical competencies for medical research learners: What is fundamental? *Journal of Clinical and Translational Science*, *1*(3), 146-152. https://doi.org/10.1017/cts.2016.31

Findley, K., & Lyford, A. (2019). Investigating Students’ Reasoning Abut Sampling Distributions Through a Resources Perspective. *Statistics Education Research Journal*, *18*(1), 26-45. https://doi.org/10.52041/serj.v18i1.148

Garfield, J., & Ben-Zvi, D. (2007). How Students Learn Statistics Revisited: A Current Review of Research on Teaching and Learning Statistics. *International Statistical Review*, *75*(3), 372-396. https://www.jstor.org/stable/41509878

Grayson, K. L., Hilliker, A. K., & Wares, J. R. (2022). R Markdown as a dynamic interface for teaching: Modules from math and biology classrooms. *Mathematical Biosciences*, *349*, 108844. https://doi.org/10.1016/j.mbs.2022.108844

Kula, F., & Koçer, R. G. (2020). Why is it difficult to understand statistical inference? Reflections on the opposing directions of construction and application of inference framework. *Teaching Mathematics and Its Applications*, *39*(4), 248-265. https://doi.org/10.1093/teamat/hrz014

Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society Series D: The Statistician*,

49(3), 293-337. https://doi.org/10.1111/1467-9884.00238

Neely, M. L., Troy, J. D., Gschwind, G., Pomann, G.-M., Grambow, S. C., & Samsa, G. P. (2022). Preorientation Curriculum: An Approach for Preparing Students with Heterogenous Backgrounds for Training in a Master of Biostatistics Program. *Journal of Curriculum and Teaching*, *11*(4), 77-85.

Pomann, G.-M., Boulware, L. E., Cayetano, S. M., Desai, M., Enders, F. T., Gallis, J. A., Gelfond, J., Grambow, S. C., Hanlon, A. L., Hendrix, A., Kulkarni, P., Lapidus, J., Lee, H.-J., Mahnken, J. D., McKeel, J. P., Moen, R., Oster, R. A., Peskoe, S., Samsa, G., … Thomas, S. M. (2020). Methods for training collaborative biostatisticians. *Journal of Clinical and Translational Science*, *5*(1), e26. https://doi.org/10.1017/cts.2020.518

Samsa, G. P. (2018). A Day in the Professional Life of a Collaborative Biostatistician Deconstructed: Implications for Curriculum Design. *Journal of Curriculum and Teaching, 7*(1), 20-31. https://doi.org/10.5430/jct.v7n1p20

Samsa, G., & Oddone, E. Z. (1994). Integrating scientific writing into a statistics curriculum: A course in statistically based scientific writing. *American Statistician*, *48*(2). https://doi.org/10.1080/00031305.1994.10476037

Sohrabi, B., & Iraj, H. (2016). Implementing flipped classroom using digital media: A comparison of two demographically different groups perceptions. *Computers in Human Behavior*, *60,* 514-524. https://doi.org/10.1016/j.chb.2016.02.056

Troy, J. D., Granek, J., Samsa, G. P., Pomann, G.-M., Updike, S., Grambow, S. C., & Neely, M. L. (2022). A Course in Biology and Communication Skills for Master of Biostatistics Students. *Journal of Curriculum and Teaching*, *11*(4), 120-138. https://doi.org/10.5430/jct.v11n4p120

Troy, J. D., McCormack, K., Grambow, S. C., Pomann, G.-M., & Samsa, G. P. (2022). Redesign of a First-Year Theory Course Sequence in Biostatistics. *Journal of Curriculum and Teaching*, *11*(8), 1-12. https://doi.org/10.5430/jct.v11n8p1

Troy, J. D., Neely, M. L., Grambow, S. C., & Samsa, G. P. (2021). The Biomedical Research Pyramid: A Model for the Practice of Biostatistics. *Journal of Curriculum and Teaching*, *10*(1), 10-17. https://doi.org/10.5430/jct.v10n1p10

Wilkerson, M. (2008). How do mathematicians learn mathematics? In *Proceedings of the Joint Meeting of PME 32 and PME-NA XXX* (Vol. 4).

Wilkerson-Jerde, M. H., & Wilensky, U. J. (2011). How do mathematicians learn math?: Resources and acts for constructing and understanding mathematics. *Educational Studies in Mathematics*, *78*(1), 21-43. https://doi.org/10.1007/s10649-011-9306-5

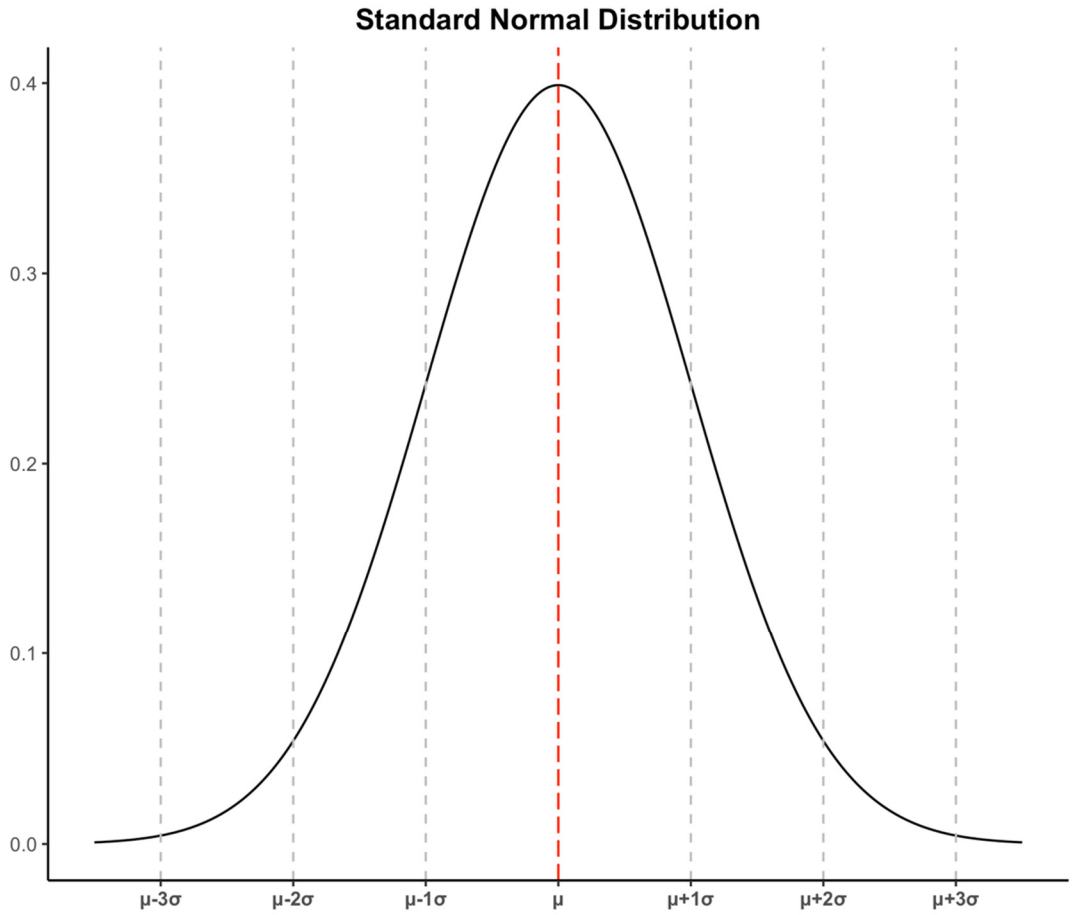
Wilkinson, W. E., & Oddone, E. Z. (2002). Training physicians for careers in clinical research: A tailored educational experience. *Nature Biotechnology*, *20*(1), 99-100. https://doi.org/10.1038/nbt0102-99

Appendix A

Teaching Sampling Distributions

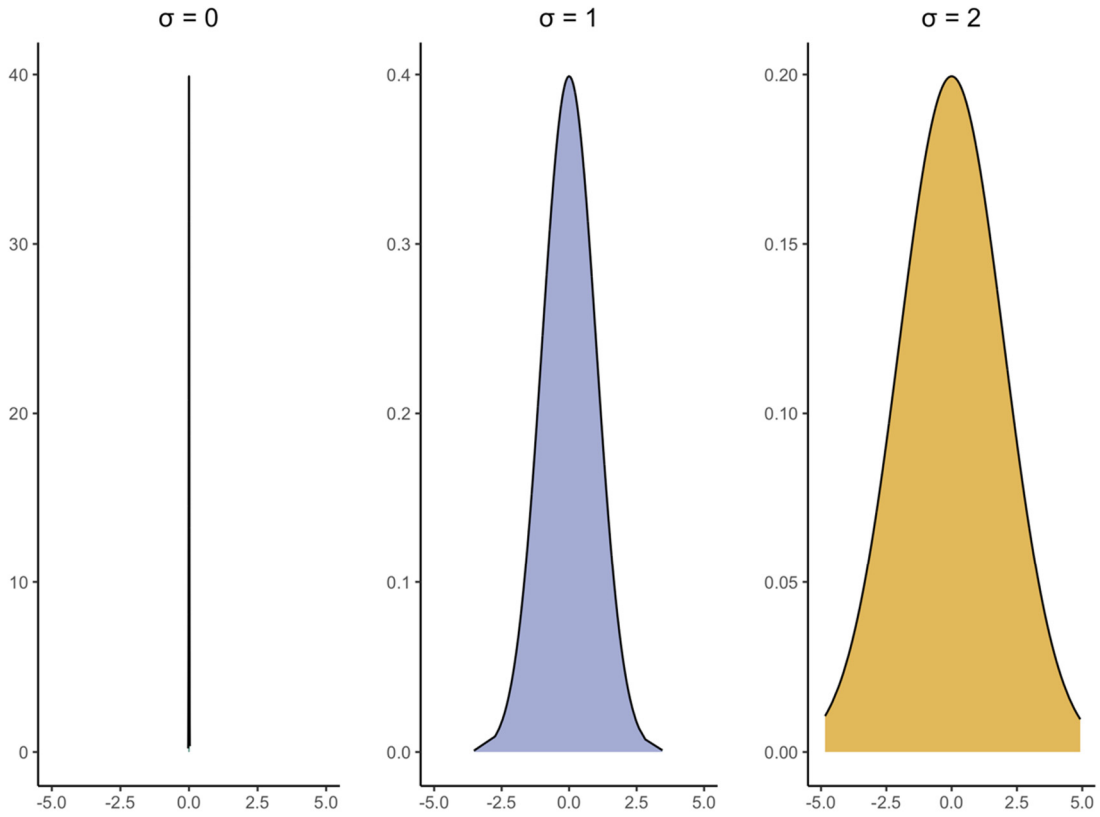
Please refer to the Symbol Key at the end of this section as you read through these notes.

1. Central tendency and variability are fixed characteristics of a population.
2. The mean often quantifies central tendency, μ, and the standard deviation often quantifies the variability, σ.



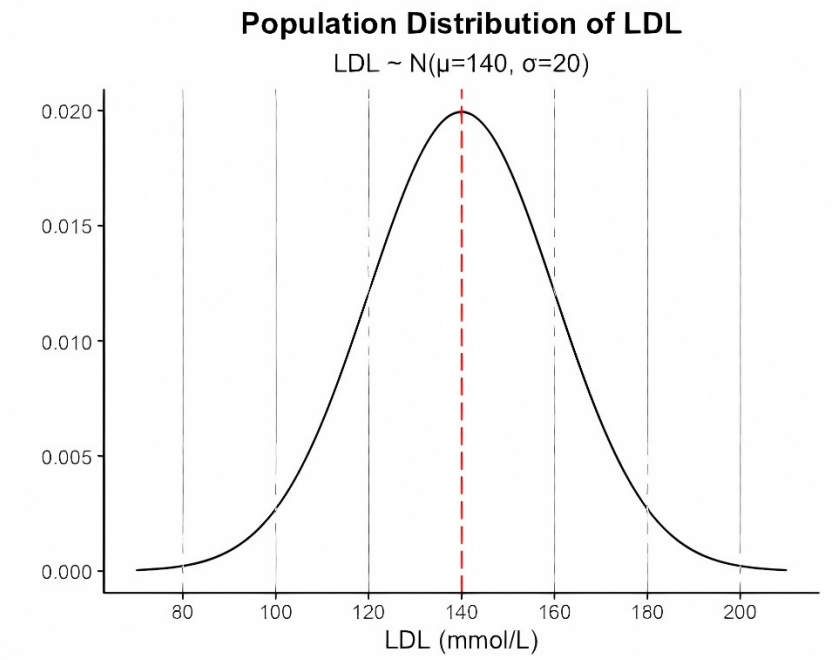
The X axis of this plot shows the population values in terms of how far they are from the mean (1, 2 or 3 standard deviations, as indicated by the gray dashed lines). The Y axis shows the relative frequency of different values in the population. The mean of the population is indicated by the red dashed line in the center of the distribution. Because of the symmetry in this population, the mean corresponds to the middle, or median: the demarcation point where half the population has smaller (or larger) values. Most of the population has values within 3 standard deviations of the mean (because most of the area under the curve falls between these values).

3. As an intuitive demonstration that σ quantifies variability, plugging identical LDL values into the formula for σ leads to a value of 0 — any other data will lead to positive values of σ (the formula for σ is available in #14 below). In short, larger values of σ represent greater amounts of variation. The following graphical depiction illustrates this. In these pictures the population mean is $\mu = 0$.



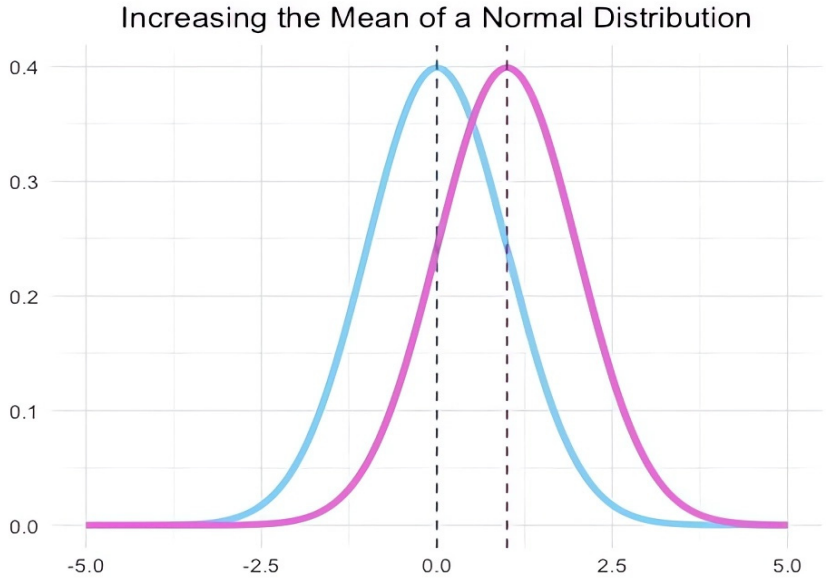
The X axis on these plots shows population values. The mean of the population is 0. The Y axis shows the relative frequency of each value in the population. The left-most plot illustrates that when all population values are the same there is no variability and therefore $\sigma = 0$. The other two plots show that larger values of σ indicate more variability in the population; i.e., the curve becomes wider as σ increases

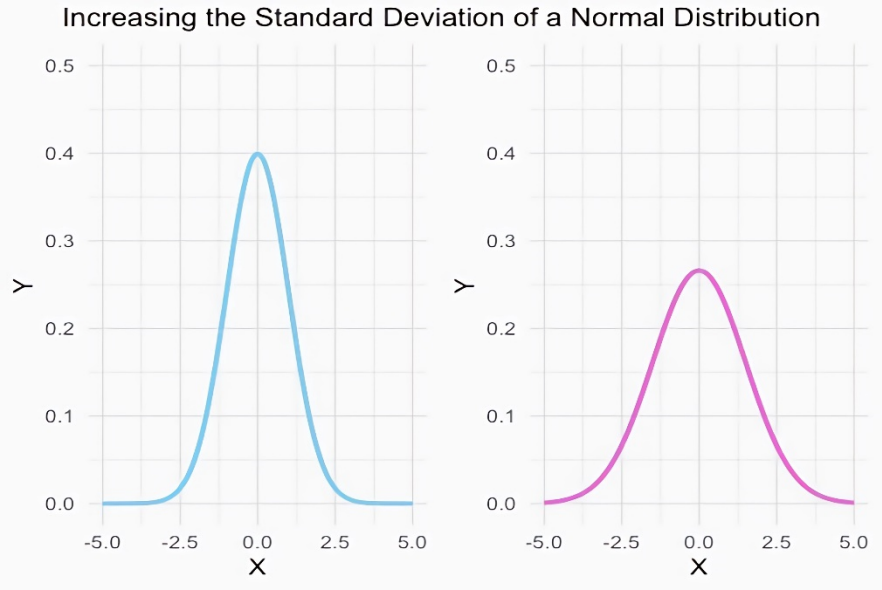
4. Because they are fixed characteristics of a population, neither μ or σ are related to sample size. For example, the LDL values of a population of patients might have $\mu = 140$ and $\sigma = 20$. Another fixed characteristic of a population is its shape, which is often operationalized through a statistical distribution. For example, although this is a simplifying assumption because of the lack of a perfect bell shape, in some circumstances, it might be reasonable to assume that LDL has a normal (i.e., bell-shaped) distribution. For our purposes, we will use histograms to visualize our sample and smooth curves (i.e., density curves) to visualize our population.



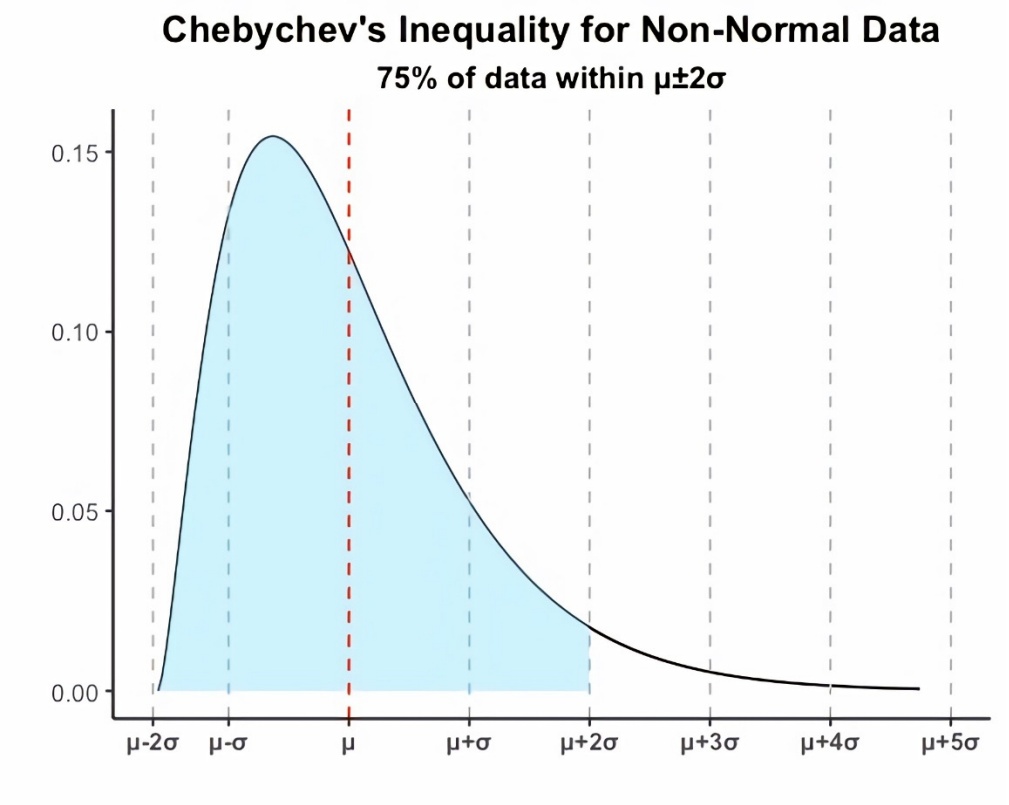
This plot shows the population distribution of LDL cholesterol. The population has a mean of 140, as indicated by the red line. The standard deviation is 20. Most of the population has LDL cholesterol that is between +/- 3 standard deviations from the mean, i.e., 140-3*20=80 and 140+3*20=200.

5. Graphically, increasing the mean of a normal distribution shifts it to the right, but without changing the amount of spread (1st plot). Increasing the standard deviation of a normal distribution increases the amount of spread, but without changing its center (2nd plot).

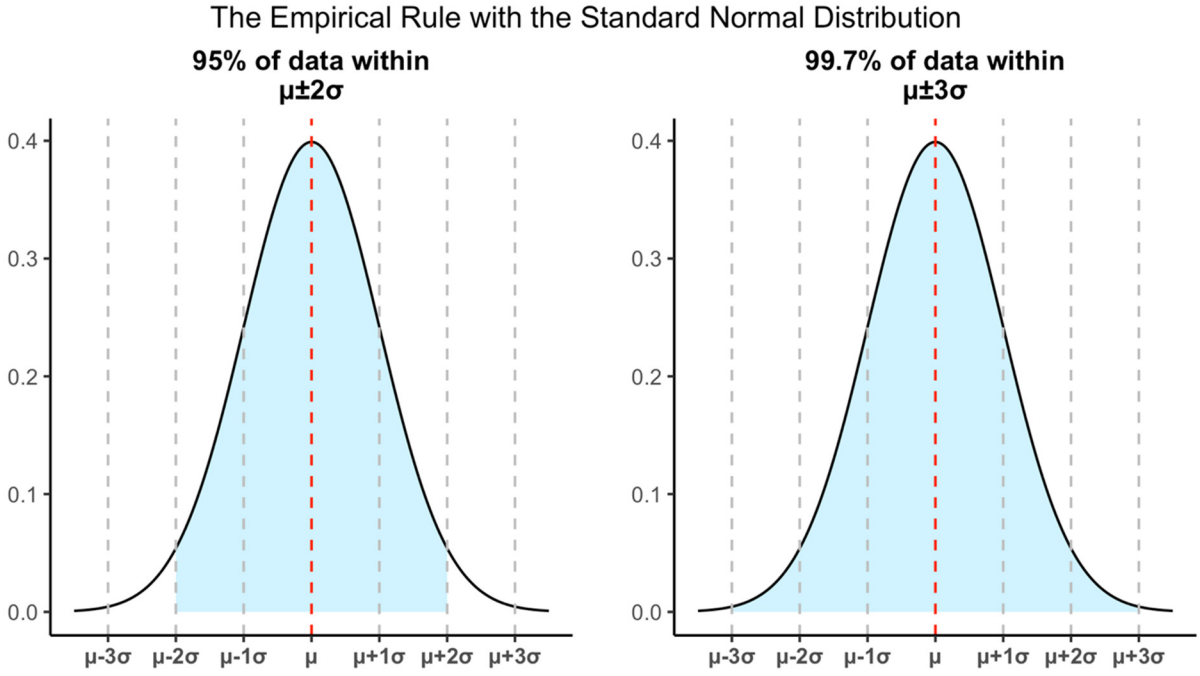




6. One reason we care about σ is that, in conjunction with μ, it helps to predict the likely range of LDL values you are likely to see in your clinic. Based on something called Chebychev's inequality, statisticians know that in any distribution—regardless of shape—that at least 75% of the values will fall within 2 standard deviations of the mean (i.e., between $\mu - 2\sigma$ and $\mu + 2\sigma$) and at least 89% of values will fall within 3 standard deviations of the mean ($\mu - 3\sigma$ and $\mu + 3\sigma$). So, for example, if $\mu = 140$ and $\sigma = 20$, then we know that at least 75% of LDL values should fall between 100 and 180 and at least 89% of values should fall between 80 and 200.

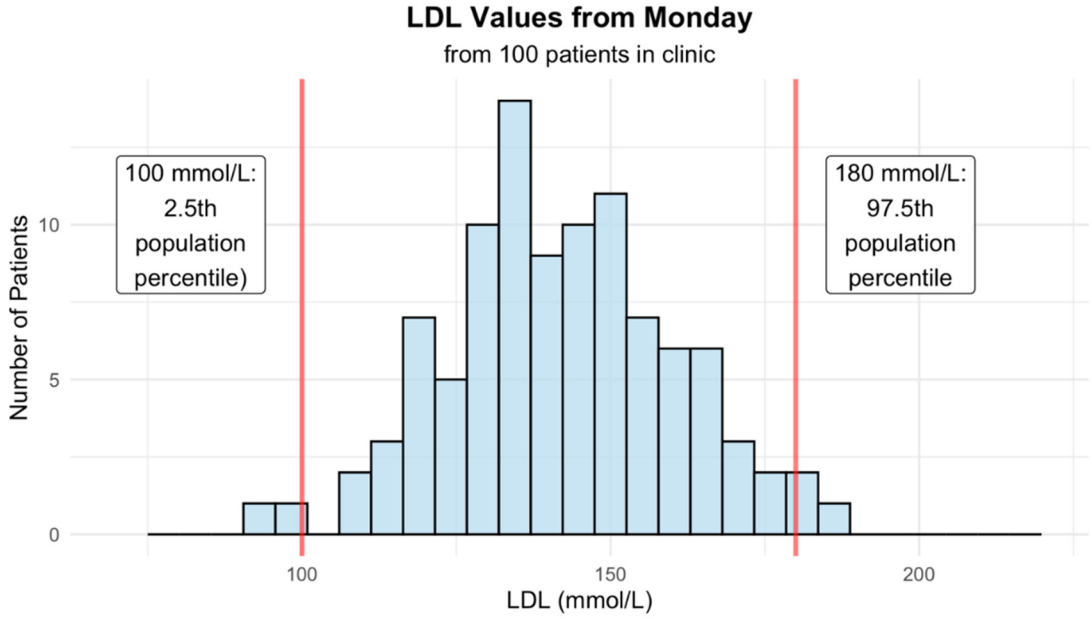


7. This range is even tighter if the distribution is normal (i.e., bell-shaped): 95% of observations fall within 2 standard deviations of the mean (actually, 1.96 standard deviations but we rounded to 2 here for simplicity) and >99% of observations fall within 3 standard deviations. So, for example, if $\mu = 140$ and $\sigma = 20$, then 95% of LDL values for your patients fall between 100 and 180.

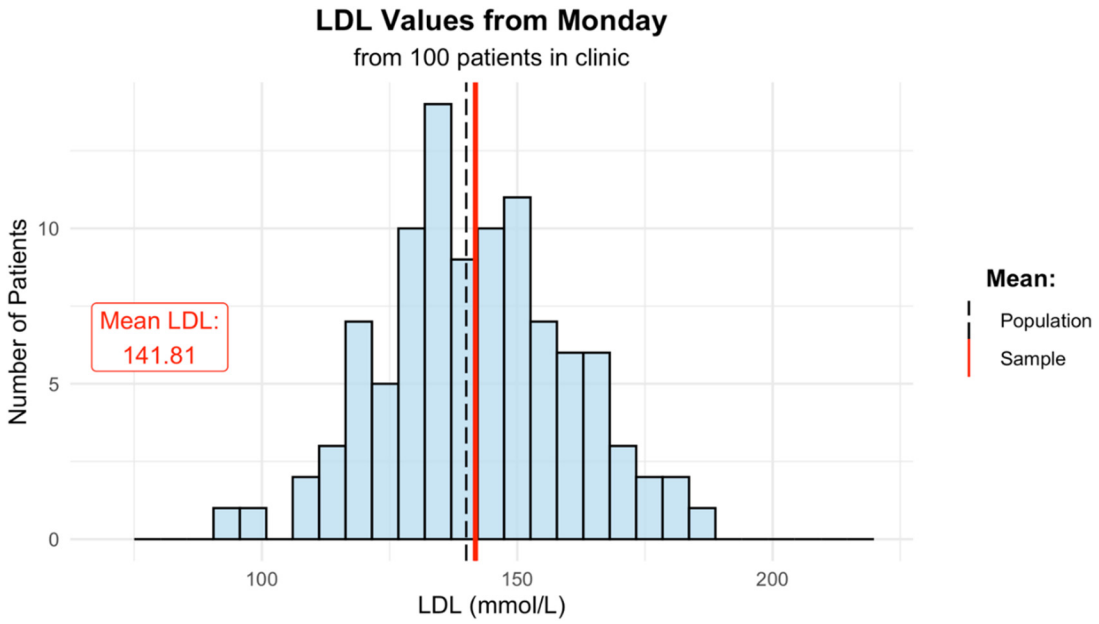


8. Suppose that LDL is normal with $\mu = 140$ and $\sigma = 20$, and that your clinic sees 100 patients per day. On Monday, you expect that approximately 95% of patients will have LDL values between 100 and 180. For any patient, though, the LDL value cannot be predicted ahead of time.

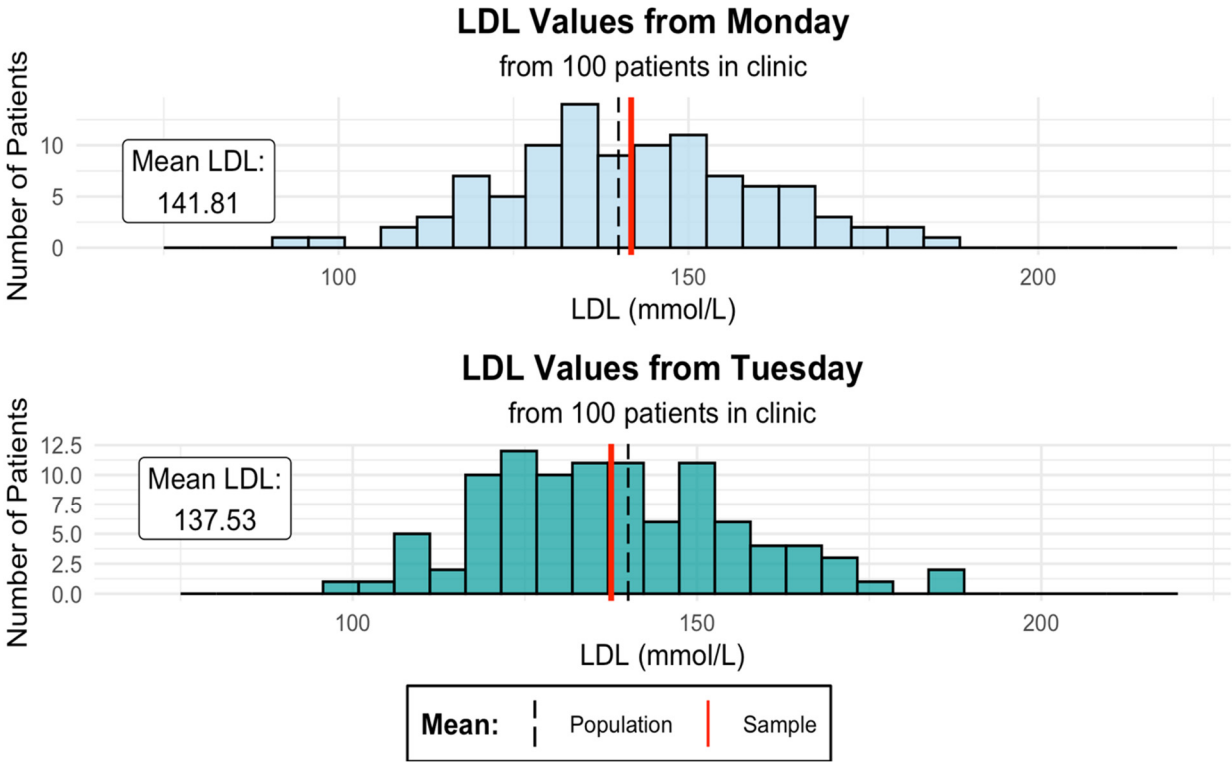
   In the plot below, the X axis shows the LDL cholesterol of the 100 patients who attended the clinic on Monday. The height of the bars corresponds to the frequency of each value of LDL cholesterol in the sample. The red lines illustrates that 95% of the LDL cholesterol values are between the 2.5th and 97.5th percentiles of the distribution.



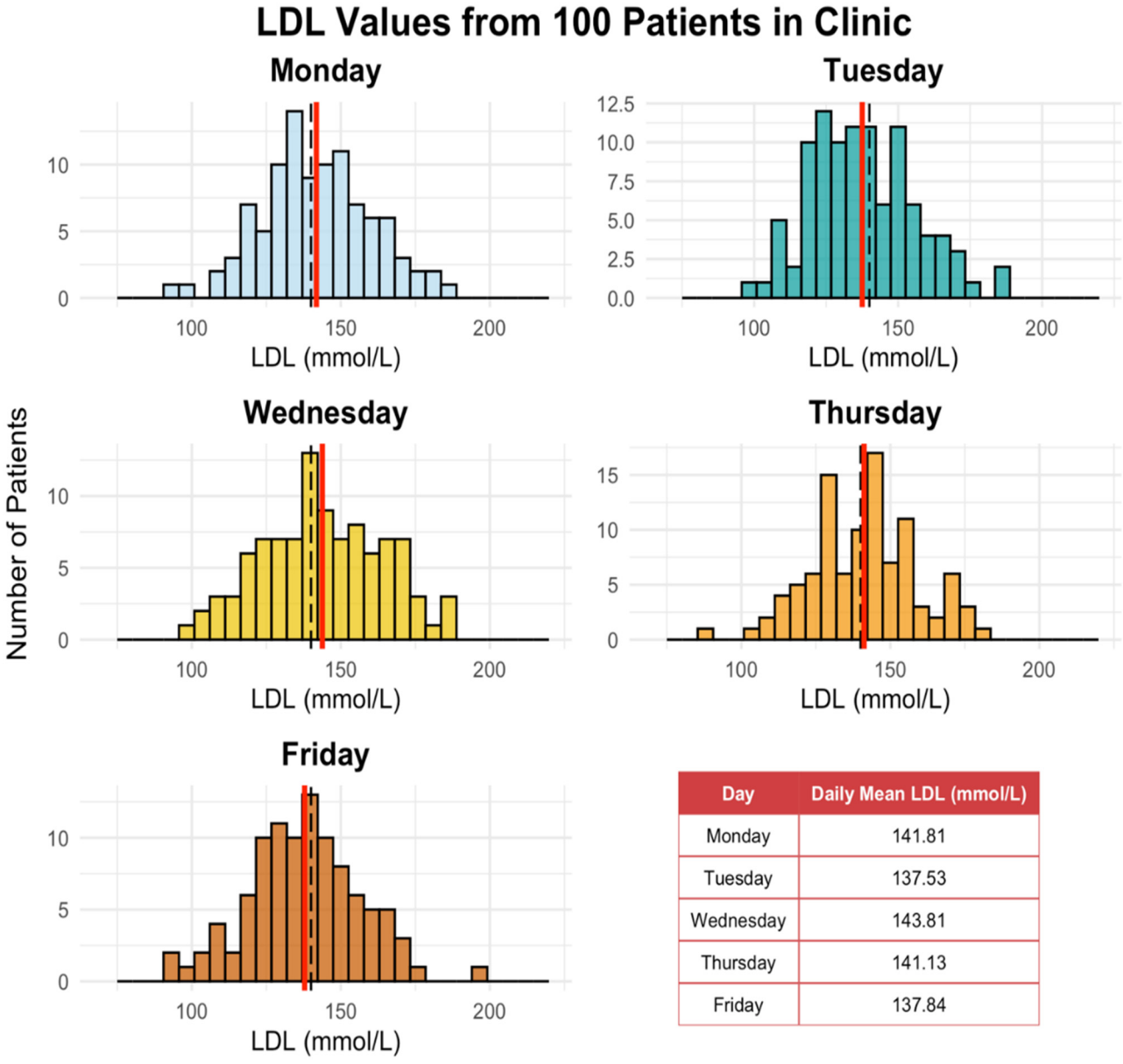
9. You are not only interested in the characteristics of patients, but also in the average (i.e., mean) LDL value for any day. Perhaps calculating mean LDL is part of your standard clinic reporting – for example, to verify that your clinic population is at high risk for cardiovascular disease. At the close of business on Monday, you take the 100 LDL values, add them, then divide by 100 to calculate the mean LDL for the day. It ought to be near the population mean of 140 (annotated below with the black dotted line), but probably will not equal 140 exactly. In our visual below, the sample mean is marked by the red line.



10. You can do the same thing for the 100 patients seen on Tuesday. The mean LDL value for Tuesday will also probably be near 140, but not equal to 140 exactly, nor will it exactly equal Monday's value. As above, the population mean is marked by the black dotted line while the sample mean is marked by the red line.

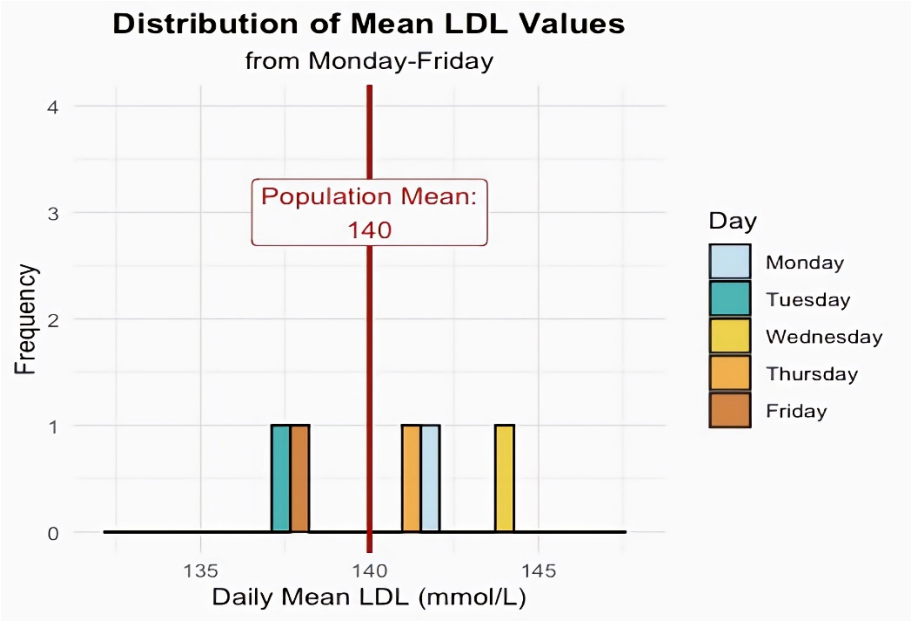


11. This process could be continued across multiple days. The empirical results could be summarized by a histogram, the first entry to which is Monday's mean LDL value, the second entry to which is Tuesday's mean LDL value, etc. As above, the population mean is marked by the black dotted line while the sample mean is marked by the red line.

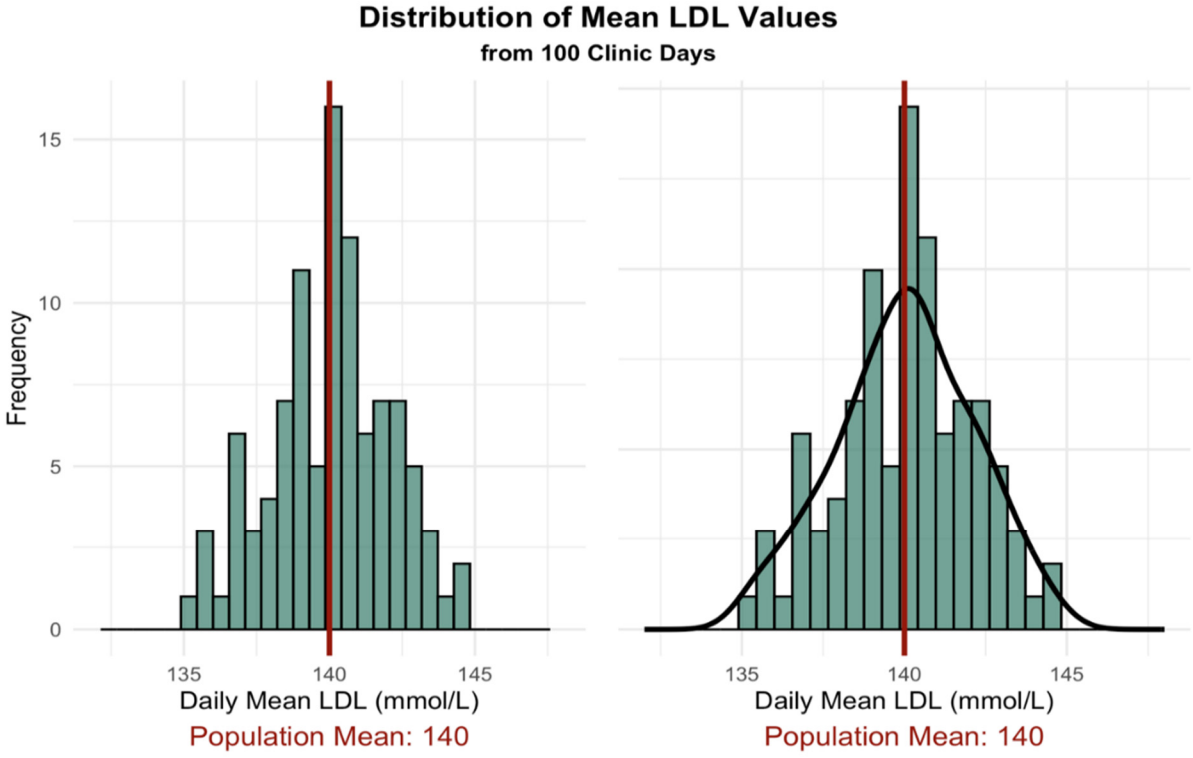


Day	Daily Mean LDL (mmol/L)
Monday	141.81
Tuesday	137.53
Wednesday	143.81
Thursday	141.13
Friday	137.84

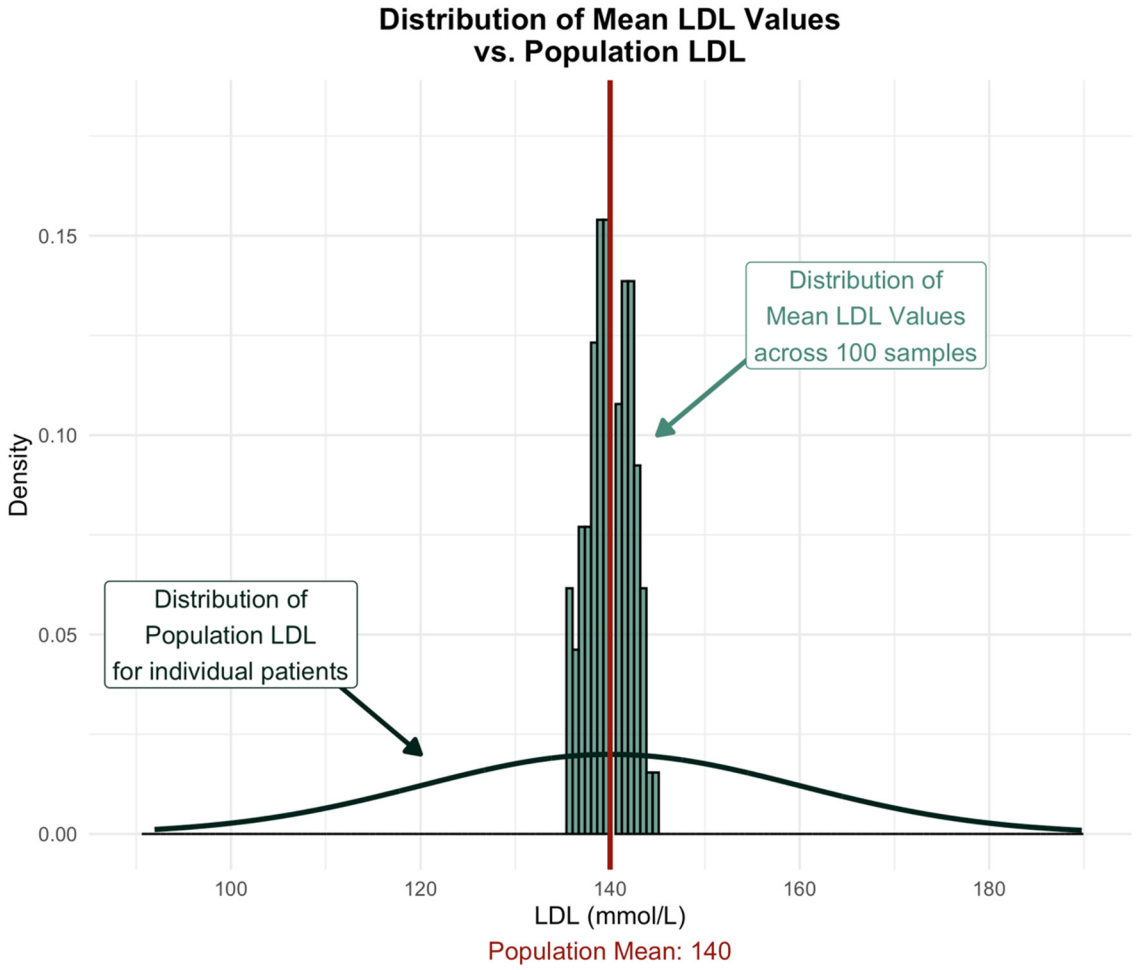
12. It is intuitively reasonable (and true) that the distribution of daily mean LDL values should be centered around 140. The histogram of mean LDL values, the first five values of which are generated from Monday through Friday, is a visual representation of a "sampling distribution."



13. The left plot illustrates the histogram of the sample means, each of size 100, for 100 clinic days. This histogram is an empirical estimate of the sampling distribution of sample means. We use the phrase 'empirical estimate' because the sampling distribution is a theoretical concept that can't be directly observed. The plot on the right overlays the theoretical sampling distribution (i.e., the smooth curve) on the empirical estimate of the sampling distribution.



14. The below plot combines the above two plots to highlight that there are two different distributions being considered: the distribution of individual LDL values (i.e., X, the distribution of population LDL) and the distribution of sample mean LDL values (each of size 100, i.e., $\overline{X}_{100}$). To recapitulate, the distribution of individual LDL values uses the patient as the unit of analysis, whereas the distribution of sample mean LDL values uses the sample of 100 LDL values as the unit of analysis. The distribution of sample mean LDL values has less spread than the distribution of individual LDL values.



15. We previously described the distribution of patient-specific LDL values. The unit of reporting (and analysis) was the LDL value for an individual patient. For example, we used information about this distribution to predict how variable LDL values will be across patients, the ranges within which LDL values are likely to fall, etc. Because LDL values for specific patients can't be precisely predicted ahead of time, patient-specific LDL values are a "random variable." We can denote this random variable using X. Random variables have distributions, such as the bell-shaped normal distribution assumed here. Random variables also have parameters such as the mean, μ, and the standard deviation, σ. A parameter is a characteristic of a distribution, such as its center and spread.

    The parameters μ and σ refer to the mean and spread of the distribution of LDL values in the population. As such, they aren't directly observable to the analyst. What is observable is the sample mean and sample standard deviation, which are denoted by $\overline{X}$ and S, respectively. Sometimes, for clarity, we will add the sample size to our notation. For example, we would use $\overline{X}_{100}$ and S_{100} to emphasize that what we've observed is based on a sample size of 100. See the symbol key at the end of this text for further clarification on notation.

    We also described the distribution of mean LDL values for samples (of size 100) taken at different clinic days. The unit of reporting (and analysis) is the observed mean LDL (of a sample of size 100) for a clinic day. The mean LDL values for specific clinic days can't be precisely predicted ahead of time, and so these are also a

random variable. As a random variable, it has a distribution, a population mean, and a population standard deviation. To emphasize that the unit of analysis is the clinic day rather than the individual patient, we term it a "sampling distribution."

In summary, we have described two distributions which are distinct but related: (1) the (underlying) distribution of patient-specific LDL values; and (2) the distribution of the mean LDL values of samples of 100 patients (i.e., the sampling distribution). Next, we will describe some notation which allows us to speak precisely about these distributions — for example, to distinguish between the mean of the underlying distribution and the mean of the sampling distribution, to distinguish between the standard deviation of the underlying distribution and the standard deviation of the sampling distribution, etc.

16. The mean LDL value for any day is:

$$\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

Since you see 100 patients per day, $n = 100$, and this formula becomes:

$$\overline{X}_{100} = \frac{X_1 + X_2 + \ldots + X_{99} + X_{100}}{100}$$

The standard deviation for a particular clinic day is:

$$S = \sqrt{\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2}{n-1}}$$

As above, for a clinic day where you see 100 patients, $n = 100$ and the formula for sample standard deviation is:

$$S = \sqrt{\frac{\left(X_1 - \overline{X}_{100}\right)^2 + \left(X_2 - \overline{X}_{100}\right)^2 + \ldots + \left(X_{100} - \overline{X}_{100}\right)^2}{99}}$$

Here, $\overline{X}_n$ estimates μ and S estimates σ. Both $\overline{X}_n$ and S can be calculated using data from a single clinic day. Additionally, $\overline{X}_n$ and S can also be used to estimate the parameters of the sampling distribution, as described below.

17. Much of statistical inference is concerned with properties of the sampling distribution of $\overline{X}_{100}$ — especially, how to use $\overline{X}_{100}$ to estimate the unknown population mean, μ. As previously mentioned, since $\overline{X}_{100}$ will differ from one clinic day to another, it is a random variable, and, thus, has its own population mean and standard deviation. We'll denote this by $\mu_{\overline{X}_{100}}$ and $\sigma_{\overline{X}_{100}}$.

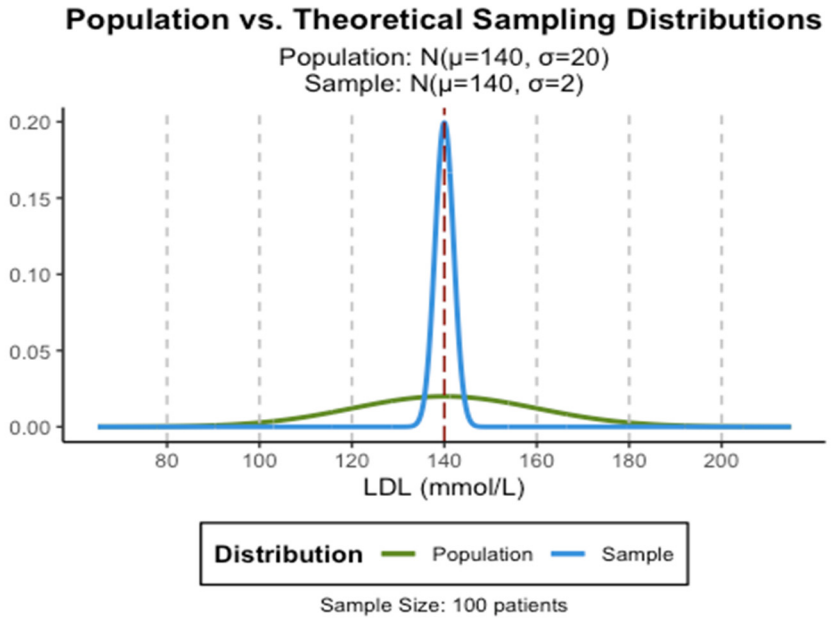
18. Moreover, to apply the results from #16 above to our study with $n = 100$, as the number of patients per clinic day increases, $\sigma_{\overline{X}_{100}}$ decreases. In fact, $\sigma_{\overline{X}_{100}} = \frac{\sigma}{10}$, where 10 is the square root of 100, the number of clinic patients per day. However, since σ is an unknown population parameter, we can use S to estimate the standard deviation of individual LDL values as in 16 above. Therefore, we have:

$$\sigma_n = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}} \Longrightarrow \sigma_{\overline{X}_{100}} = \sqrt{\frac{\sigma^2}{100}} = \frac{\sigma}{10}$$

$$\Longrightarrow \sigma_{\overline{X}_{100}} = \sqrt{\frac{S^2}{100}} = \frac{S}{10}$$

This formula provides the crucial link from a single clinic day to the sampling distribution. Our best guess about $\mu_{\overline{X}_{100}}$ is $\overline{X}$. Our best guess about $\sigma_{\overline{X}_{100}}$ is $S/\sqrt{n}$. $\overline{X}$, S, and n can all be observed using data from a single clinic day, and so we can use data from a single clinic day to make an educated guess about the parameters of the sampling distribution.

19. It also turns out that the mean LDL per clinic day (calculated using 100 patients per day) has a normal shape, and so it can be fully described as normal with mean 140 and standard deviation 2. Because any day's mean LDL can't be precisely predicted ahead of time, it is a random variable. Moreover, it turns out that this sampling distribution (i.e., of the means of repeated samples, each taking 100 patients within a single clinic day) has a normal shape, and so we can apply the result that 95% of daily values will fall within $\mu_{\overline{X}_{100}} \pm 2\sigma_{\overline{X}_{100}}$.
20. In fact, for 95% of days, the mean LDL per clinic day should fall within 136 and 144.
21. To recapitulate, what if you only had data from Monday and want to derive the above sampling distribution? On first blush, this seems impossible, since all you have is a single data point. Suppose that value is 139, which we anticipate will be near but not exactly equal to the unknown mean value of $\mu_{LDL} = 140$. We can also calculate a standard deviation from the 100 patients, which estimates σ_{LDL}. Suppose that value happens to be 21, which we anticipate will be near but not exactly equal to the unknown value of $\sigma_{LDL} = 20$.
22. This is where statistical theory comes to the rescue with 3 crucial results.
23. The first crucial result was mentioned before – namely that $\mu = \mu_{\overline{X}_{100}}$. So, based on the data from Monday, we estimate $\mu_{\overline{X}_{100}}$ to be 139.
24. The second crucial result is that $S_{\overline{X}_{100}} = \frac{\sigma}{\sqrt{100}}$, where 100 is the sample size. So, based on the data form Monday, we estimate that $S_{\overline{X}_{100}} = \frac{21}{\sqrt{100}} = 2.1$
25. The third critical result is that, under general conditions, sample means based on 100 patients per day have a normal distribution.
26. Putting these 3 results together: the true sampling distribution in question is normal with mean 140 and standard error 2.0.



In our example above, we observed that LDL values from a single sample size of 100 is normally distributed with mean 139 and standard deviation of 21 (i.e., $N(139,21)$).

From this, we conclude using the above results that the estimated sampling distribution for a sample of size 100 is normal with mean 139 and standard error 2.1 (i.e., $N(139,2.1)$), which we obtained from our observed LDL values.

We will use the theoretical sampling distribution, $N(140,2)$, to obtain a point estimate and measure around μ, the true population mean LDL value (rather than $\mu_{\overline{X}_{100}}$). We can estimate a confidence interval for μ using our observed sampling distribution:

$$139 \pm z * (2.1)$$

where z is the critical value determined by our confidence level.

27. At this point, the concept of a sampling distribution should be clear to the student, and, indeed, we are but a single step from introducing related concepts such as estimation, confidence intervals, p-values, etc.
28. To help connect this presentation with more mathematically-based treatments of this information, we will close by noting that we've embedded some mathematical assumptions into our analysis. For example, we implicitly assumed that LDL values for one patient are unrelated to LDL values for other patients, and that each patient is similar, in that they represent an independent sample from an underlying population of LDL values. In other words, we've assumed that the LDL values for individual patients are independently and identically distributed. We've also assumed that the mean LDL values for each day were similarly representative — that is, independently and identically distributed. Indeed, a single clinic day was used as a metaphor for the more general construct of taking a single sample.
29. Roughly speaking, the CLT states that if the distribution of LDL values (i.e., across individual patients) is normal (bell-shaped), then the sampling distribution of sample means will be normal as well, with the mean and standard deviation as previously provided. Moreover, if the distribution of LDL values isn't normal, as the sample size (i.e., the number of clinic patients per day) increases, the shape of the sampling distribution will more and more closely approximate normality, with the mean and standard deviation as above. The speed at which this occurs depends on both the shape of the underlying distribution and the sample size — for LDL, $n = 30$ is probably sufficient.

Symbol Key:

	Our Example			
	Sample Size (n) = 100 patients			
	X	$\overline{X}_{100}$	**Actual Value by CLT:**	**Estimated By:**
Unit of Analysis	patient	sample	–	–
Population Mean	μ	$\mu_{\overline{X}_{100}}$	$\mu_{\overline{X}_{100}} = \mu$	$\overline{X}_{100}$
Population Standard Deviation	σ	$\sigma_{\overline{X}_{100}}$	$\sigma_{\overline{X}_{100}} = \frac{\sigma}{\sqrt{100}}$	$\frac{S}{\sqrt{100}}$
Sample Mean	$\overline{X}_{100}$	$\overline{X}_{\overline{X}_{100}}$	–	–
Sample Standard Deviation	S	$S_{\overline{X}_{100}}$	–	–
Distribution	any	Normal (by CLT)	–	–

Note: the sample mean and sample standard deviation of $\overline{X}_{100}$ are not used in our presentation above, but they have been included here for completeness

Acknowledgments

Not applicable.

Authors contributions

Not applicable.

Funding

Gina-Maria Pomann's support of this project was made possible (in part) by Grant Number UL1TR002553 from the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health (NIH), and NIH Roadmap for Medical Research. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NCATS or NIH.

Competing interests

Not applicable.

Informed consent

Not applicable.

Ethics approval

The Publication Ethics Committee of the Sciedu Press.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.