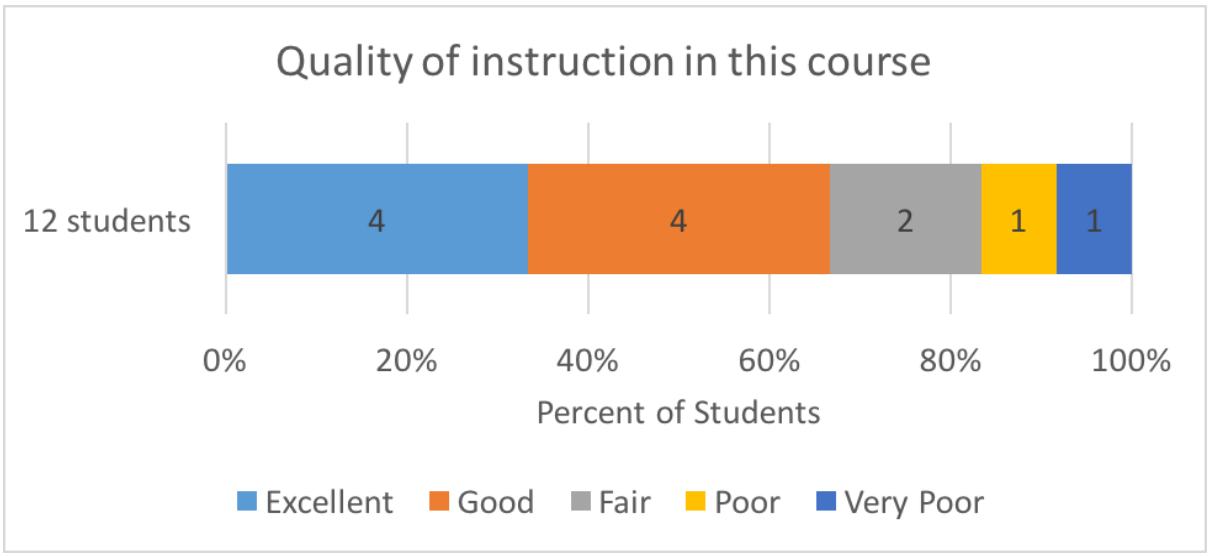
Stanford | Evaluation & Research

Stability of Course Evaluation Statistics

Variation in the mean of any course evaluation statistic is to be expected over time, as a different class of students takes and evaluates each offering of a particular course. The degree of expected variation can be theoretically estimated, as well as empirically examined, using recent course evaluation data. The average course evaluation scores can vary within a range of 0.6 points, while courses with smaller enrollments may expect scores to vary by a whole point or more over time. Understanding the margin of error in evaluation scores is crucial to interpreting course ratings.

Imagine a small seminar with 12 students. In response to the question "How would you rate the quality of instruction in this course?" students select an option from "Excellent" to "Very Poor." One possible distribution of responses is seen in this chart:



If we were to compute the traditional mean score for this class (assigning *Excellent* a score of 5, *Good* a score of 4, etc.) we would observe a mean of 3.75.

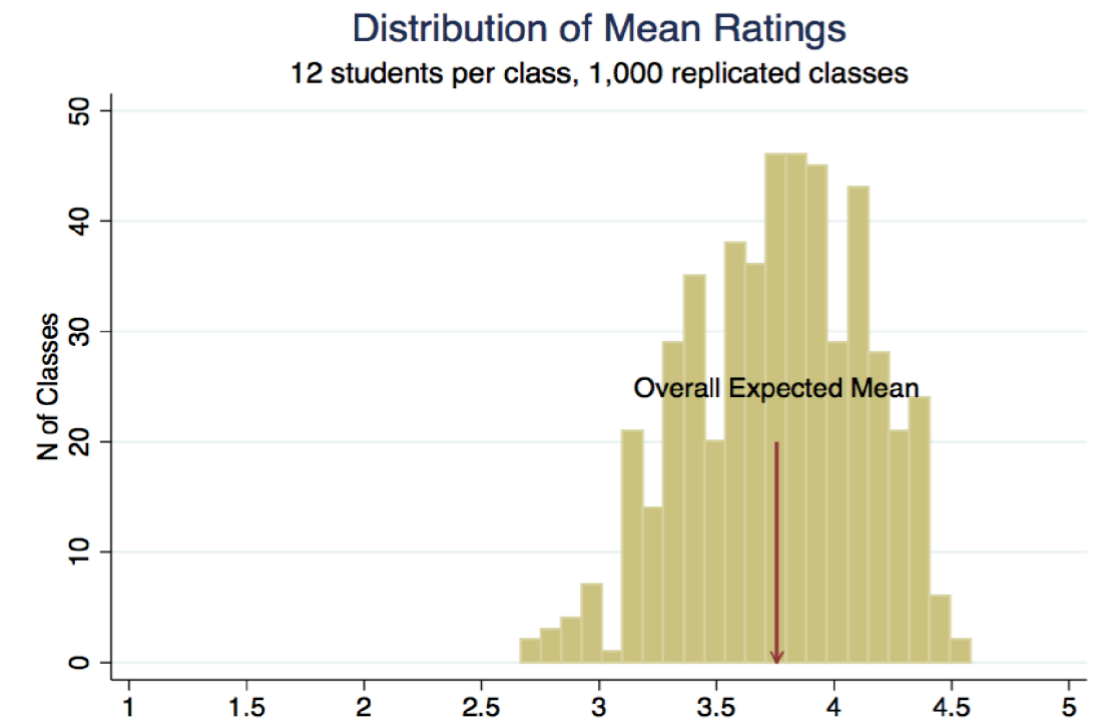
Now, imagine that the following quarter, this same instructor teaches the same seminar using exactly the same curriculum and materials. How would we expect the student evaluation scores to be distributed? Our intuition says they would probably look a lot like the graph above, but probably not *exactly* like the graph above. Perhaps 5 students would select *Excellent*, while only 1 would select *Fair*. In that scenario, the mean score would be 3.92, a difference of nearly two-tenths of a point from the previous quarter.

Because a very minor change in the evaluation distribution results in a two-tenths point change in the mean rating, we might ask just how much variation we ought to expect from class to class, when all else (teaching methods, student population, etc.) are held constant? That is, how much "wiggle room" can we expect to observe around a course's mean rating?

Variability of mean ratings

Conceptually, we can imagine a population of students distributed similarly to the above chart, where ⅓ would respond with *Excellent* were they to take this class, ⅓ would respond *Good*, and so on. What if we were to randomly

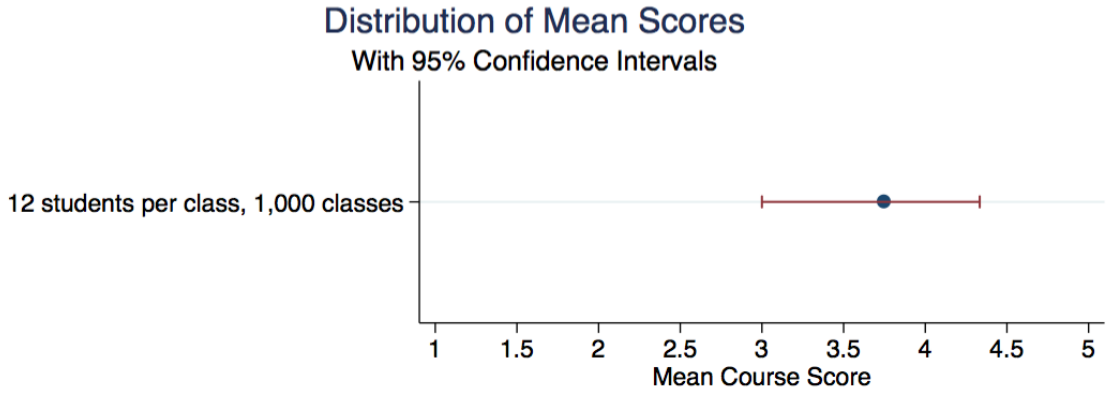
draw 12 students from this population, compute the mean rating, write that down, draw another sample of 12, and so on? What would the distribution of those mean ratings look like? The following charts illustrates this distribution.



If we sample 1,000 hypothetical classes of students, we might observe some classes where the mean rating falls below 3.0, as well as some classes where the rating tops 4.5. The center of the distribution is the expected mean of 3.75, but the mean scores span a range of more than 1.5 points on a scale of 1 to 5.

This *bootstrap* technique of repeated sampling to compute the plausible range of a statistic was pioneered by Bradley Efron at Stanford's Department of Statistics. We can compute what is known as a *confidence interval* (more colloquially known as a "margin of error"), or a range of values we would expect the mean rating to fall within under identical conditions.

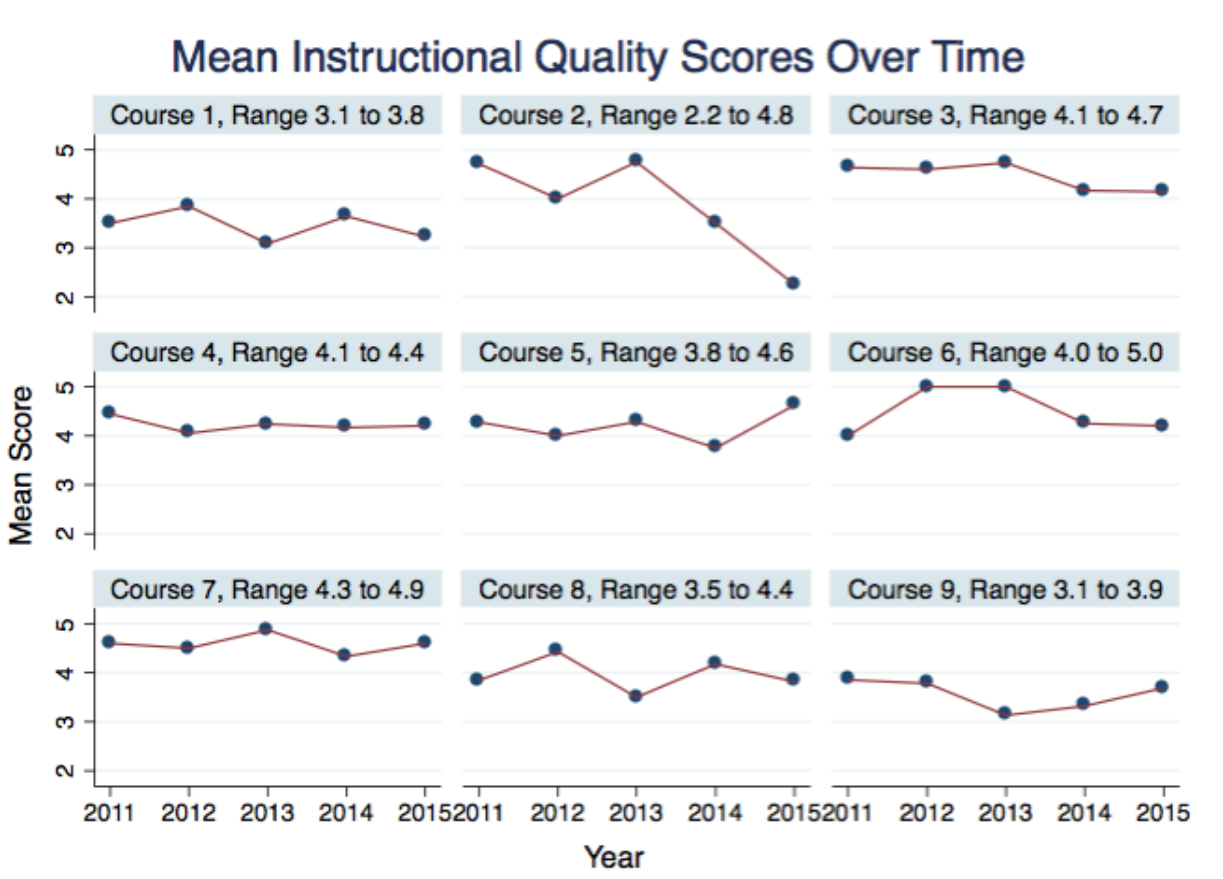
A customary representation for confidence intervals includes error bars, as shown in the following graph (next page). Here we see a dot at the observed mean value of 3.75, and a set of brackets delineating the likely range of this mean score under similar conditions.



The takeaway point is this: **while this particular course resulted in a mean rating of 3.75, this same course could have easily been rated as low as 3.0 or as high as 4.4**. This variability implies that we should be cautious when rank-ordering courses by a summary statistic.

Actual variability of course scores

To illustrate the actual variability of evaluation scores over time, a random sample of nine courses was drawn from the historical course evaluation data spanning AY 2010-11 through 2014-15. The selection was restricted to courses taught by the same instructor and the same quarter across all five years. Mean scores for the item asking students to rate the instructor's overall teaching on a scale of 1 to 5 were computed and plotted over time. The results appear in the chart below.



As you can see, the mean scores can vary quite dramatically. Of these nine courses, Course 4 shows the smallest variation (spanning a range of 0.3 points), while Course 2 shows the greatest variation (spanning a range of 2.6 points). These ranges are all similar to what would be theoretically predicted. Across **all** courses with two or more years of evaluation data, the average range for this mean score spanned 0.6 points. Subsequent inspection of the historical record also confirms another prediction—courses with larger enrollments tend to have more stable mean evaluation scores over time.

Accounting for variability

When using mean course evaluation scores for decision-making (e.g., deciding to renew an instructor's contract), it is important to remember that a single course's score could easily have been a few tenths of a point higher or lower simply due to the random sampling of students in that particular course. An increase or decrease of a few tenths of a point from year to year should not necessarily be interpreted as an actual increase or decrease in instructional performance.

For more information, please see our website at evals.stanford.edu

Contact Us:
518 Memorial Way, Stanford, CA. 94305 | course-evaluations@stanford.edu