Evolution of a Qualifying Examination from a Timed Closed-Book Format to an Open-Book Collaborative Take-Home Format: A Case Study and Commentary

Gregory Samsa[1,*]

[1]Department of Biostatistics and Bioinformatics, Duke University Medical Center, 11084 Hock Plaza, Durham NC 27710, UK

*Correspondence: Department of Biostatistics and Bioinformatics, Duke University Medical Center, 11084 Hock Plaza, Durham NC 27710, UK. Tel: 44-919-613-5212. E-mail: Greg.Samsa@duke.edu

Received: August 4, 2020 Accepted: September 23, 2020 Online Published: February 15, 2021

doi:10.5430/jct.v10n1p47 URL: https://doi.org/10.5430/jct.v10n1p47

Abstract

Objective: Our master's program in biostatistics requires a qualifying examination (QE). A curriculum review led us to question whether to replace a closed-book format with an open-book one. Our goal was to improve the QE.

Methods: This is a case study and commentary, where we describe the evolution of the QE, both in its goals and its content. The result was a week-long, open-book, collaborative, take-home examination structured around the analysis of two types of studies commonly encountered in biostatistical practice. Our evaluation of the revised format includes its fairness, student performance, and student feedback.

Results: The new format has a number of advantages: (1) it has a specific educational goal; (2) it provides sufficient time for students to produce their best work; (3) it encourages students to review elements of the first-year curriculum as needed; and (4) it can be administered remotely, even during a pandemic. Potential concerns pertaining to cheating and rigor can be adequately addressed. The results of our evaluation of the examination have been encouraging. The QE is intended to be a "fair" examination that covers important material which is beneficial to students, and does so in a way that is transparent and puts everyone in a position to perform their best work.

Conclusions: An examination using this format has much to recommend it. When designing an examination, it is important to (a) match its format with clearly specified educational goals; and (b) distinguish between the distinct constructs of difficulty and rigor.

Keywords: biostatistics, evaluation methods, open-book examination

1. Introduction

A meeting of the teaching faculty is taking place, and the topic is the results of the qualifying examination allowing students to continue in the doctoral program. Despite attempts to pay careful attention one of the participants is instead ruminating upon the motivations of his favorite character from the ancient television show Gilligan's Island. More specifically: that, despite pretending to be unable to do so, The Professor (who can build almost anything with coconuts and bamboo) could fix the boat and leave the island any time he likes, but prefers not to do so. Not only is he enjoying a paid vacation, and perhaps doing a bit of field research for a manuscript on cultural anthropology, but he is also escaping from tedious meetings of the teaching faculty.

This reverie is interrupted by a question from one faculty member to another: “So, when you were writing your dissertation did you ever lock yourself in your office, hide all the books, and then take the next five hours to write?”

"Of course not, why do you ask?” was the reply.

"Well, you are proposing to drop a student from our doctoral program, who we all enthusiastically recruited, who we all determined was well prepared, and who passed her first-year courses, because she failed an intentionally difficult 5-hour closed-book examination on those same courses, and the rationale that you offer is that her exam results demonstrate that she probably won't be able to successfully write a doctoral thesis. Wouldn't it be better, instead, to just let her find a thesis advisor and try to write one? That would provide direct evidence, rather than the results of an examination which you just stipulated has little connection with its stated goal?"

That point seemed obvious, although not to everyone in the room. Eventually we did the right thing by the student. But, it did lead us to ask what purpose the examination in question actually serves. Moreover, our master's program requires a similar examination and, furthermore, was undergoing a comprehensive curriculum review. This report is a case study describing the results of that review, the changes to the examination which resulted, and some insights which we hope might be helpful to others. In the spirit of The Professor's research on cultural anthropology, we combine scholarly discussion of issues of educational pedagogy with a more speculative commentary on the social and organizational contexts which motivate those who design and administer such examinations.

2. Background

2.1 The Program

Our program offers a Masters of Biostatistics (MB) degree. It is a 2-year program, with students taking core classes during their first year and then differentiating during year 2. A key element of the second-year curriculum is a master's project. Typically, one quarter to one third of students decide to immediately enter doctoral programs, with the remainder entering the work force.

The Qualifying Examination (QE) is taken between the first and second years. It has been a 5-hour, closed-book, closed-note examination, with two chances to pass (June and August). Although each case is considered on its own merits, students who fail the examination twice are at significant risk for dismissal.

3. Evolution of the Qualifying Examination

3.1 Initial Goals of the Qualifying Examination

As historical background, the initial rationale for the QE combined the observation that doctoral programs typically have similar examinations (not all master's programs do) with the notion that a substantial QE would be a component of an academically rigorous curriculum, a notion which was important to both departmental leadership and much of the teaching faculty. In effect, what we had done was to simply assume, without deep analysis, that the MB program ought to have a QE, and moreover that its structure ought to be similar to the QE for a typical doctoral program. It could be argued that the initial iteration of the QE did not have a formal educational "goal", as rigor in of itself is not a skill, competency or attribute which a student should possess (Ippersiel & el Atia, 2014).

Upon reflection, and without entering the debate about the degree to which the QE within our doctoral program was performing as intended, there was a strong argument that the QE for the MB program should be different from the QE for the doctoral program. The main goal of this latter examination was to assess the likelihood that the student would successfully complete a dissertation. On the other hand, most of our MB students opt for a "master's project" rather than a "master's thesis". More specifically, the thesis option tends to be selected by those who plan to immediately proceed to doctoral work, and for the present purposes is analogous to a mini-dissertation. The ideal master's project, on the other hand, results in a collaborative manuscript with biomedical researchers, and would among others demonstrate to potential employers that the student has mastered the biostatistical and team science skills necessary to be a valued contributor within their work environment. Because the two examinations have different purposes, it is reasonable to ask whether they should have different formats as well.

3.2 Changes to the Qualifying Examination Goals

Over time, the rationale for the QE evolved into (a) to encourage students to review the first-year course material; and (b) to humanely remove from the program those students experiencing profound academic difficulties. Regarding the latter, the time limit for completing a dissertation provides an analogy. If extenuating circumstances are present and optimism about eventually completing the dissertation remains then an exception can be granted, but otherwise the deadline takes precedence and the process proceeds as automatically and painlessly as possible for all concerned, including the student.

The first major change to the QE recognized that, in order to identify significantly at-risk students, more basic questions would suffice. Indeed, directions to the instructors were changed to the effect of "write questions such that if a student can't do X you would doubt their ability to successfully function as a biostatistician". A secondary benefit was in streamlining the process of grading, since students need not be precisely ranked according to performance but, instead, only that low outliers be identified. In addition to the revised QE, we strengthened procedures for identifying at-risk students as early as possible and developing personalized remediation plans. Nevertheless, in large part because the QE was associated with the risk of dismissal, students continued to find it to be quite stressful.

Reconsidering the first element of the above rationale, reviewing the first-year course material was neither a skill, a competency or an attribute, and thus not an educational goal, although "demonstrating mastery of the first-year course material" was. Indeed, perhaps demonstrating mastery was what was intended when faculty members advocated for the importance of reviewing first-year course material, since review can help assist in developing mastery, and since one of the advantages of high-stakes closed-book examinations is that students might tend to study for them more extensively than is the case for their open-book counterparts (Durning *et al*, 2016). Nevertheless, demonstrating mastery is a relatively non-specific construct. Ultimately, we decided that the QE should either be linked to a specific and sufficiently important educational goal or dropped.

3.3 Current Version of the Qualifying Examination

Ultimately, we revised the first educational goal to be “integrating material from the first-year courses, and demonstrating adequate mastery of that material in an applied setting which would provide practice in the skills required by the master’s project and beyond”. This goal is consistent with the program's overall mission of training excellent collaborative biostatisticians and thus demonstrably important. Moreover, it specifically refers to a crucial element of the second-year curriculum: namely, the master's project. Indeed, as stated it is analogous to the goal of the QE for our doctoral program, with the PhD dissertation replaced with the master's project.

Recognizing the importance of matching goals and evaluation methods, reconceptualizing the goals of the QE led to replacing a closed-book examination with an open-book take-home examination. Students are given a week to complete the examination, are welcome to discuss it among themselves and with their instructors, but are required to submit individual narrative answers. Students are informed that they might also be interviewed about their narrative – for example, in order to clarify their answers -- and also that overly similar responses would be met with skepticism. In the spirit of using the QE as a teaching tool (i.e., in addition to an evaluative tool), the examination is followed by a debriefing session, where the first-year instructors discuss their observations about student responses and emphasize key principles which the examination is intended to illustrate. In the spirit of remediation, struggling students also discuss problematic answers with their instructors.

Appendix 1 contains the exam questions for the first iteration of the "new and improved" version of the QE. The questions are organized according to two study designs commonly encountered in practice: a randomized trial and a genomic study. Within these broad categories, questions are interleaved from three of the four main threads of the first-year curriculum: namely, mathematical statistics, applied data analysis, and practice of biostatistics (e.g., study design, biology and communication). Each question includes analysis of a dataset, thus exercising the fourth thread of the curriculum (i.e., computer programming) as well.

4. Evaluation of the Qualifying Examination

Our evaluation of the QE used two metrics: (1) quality of student responses; and (2) student feedback. For the first metric, our rationale was that high-quality responses suggested that the QE was structured in such a way as to facilitate students being able to perform their best work. For the second metric, our rationale was that we wanted the QE to be perceived as a positive and relatively stress-free educational experience.

In brief, the results of the evaluation were encouraging. Student answers were generally correct, and the quality of their written communication (a programmatic focus) generally good. Moreover, the questions contained sufficient nuance to help assess which concepts were understood superficially rather than deeply -- information which was shared with students during the debriefing session and will also considered by their instructors as they work to improve the delivery of course content.

Students' overall evaluation of the QE on a scale of 0 (terrible) to 100 (awesome) had a mean of 79 (and a median of 80). Some selected responses to the question about which aspects of the QE worked well (paraphrased) were:

- I liked that it was collaborative, as it helped me get different perspectives and allowed me to feel more confident about my answers.

- It felt similar to the work I'm currently doing in my internship, and should provide good preparation for what will be expected of us after graduation.

• I liked that the questions were bigger scale and incorporated many concepts and thinking critically about them.

• I really enjoyed that it felt practical and not like an exam.

• Because it utilized knowledge from all the first-year courses, it encouraged me to review their material.

Some selected comments to the question about areas for improvement (paraphrased) were:

• The QE was rather long, which was problematic for those with summer internships.

• The applied courses were better represented than the theoretical ones.

• Some of the questions were inconsistent with how the material was presented in class and thus difficult to interpret.

• Some material on the QE (e.g., validation) was not deeply covered during classwork.

Considering the free text comments, those pertaining to the components of the QE which worked well tended to support the conclusion that students understood and supported its overall goal. Most areas for improvement pertained to implementation issues such as clarifying the questions and better matching the format of the questions with how the material was presented during class. The decision to include some material not deeply covered during class was intentional, recognizing that this is typical of actual practice, and potential concerns about this point might be addressed in the QE's preamble. The relative lack of emphasis on traditional "theory questions" was partly induced by the fact that the answers to almost any such question can be found online. The workaround was to include some non-standard questions, and also to emphasize communication through explaining theoretical results and how they apply to actual data analyses, one hope being that if a student can explain a theoretical idea with sufficient clarity they will also be to be able to apply that idea in practice.

5. Discussion

We have described the evolution of a qualifying examination from a timed closed-book format to an open-book take-home collaborative format. We believe that the current version of the QE has a number of advantages. It has a specific educational goal, and a format which is consistent with that goal. It allows students sufficient time to produce their best work, and also to review elements of the first-year curriculum as needed. It can be administered remotely, even during a pandemic. The results of our evaluation of the QE have been encouraging to date.

There is an expanding literature on the relative merits of open-book versus closed-book examinations, although few of the citations involve high-stakes testing, testing within mathematically-related disciplines such as biostatistics, or the explicit allowance for collaboration. (Durning *et al,* 2016) provides a recent systematic review (including an extensive set of references) covering the topics of exam preparation, test anxiety, exam performance, psychometrics and logistics, testing effects, and public perception. They concluded that neither approach to testing is uniformly superior to the other, and a consistent theme within this literature is the importance of matching evaluation format with educational goals. Indeed, not only should evaluation be in alignment with educational goals, but with teaching methods as well (Ioannidou, 1997). (Durning *et al*, 2016) noted that while some argue that closed-book examinations promote superficial memorization, others assert that they induce deeper study because of the need to commit information into memory. Moreover, "expert performance is closely-tied to rich well-organized content knowledge of a subject" (i.e., internalized proficiency), and thus some memorization will always be required. These authors also noted that cheating is a greater risk for open-book examinations, and that increased anxiety is a greater risk for their closed-book counterparts. Finally, they noted that both students and instructors would benefit from training in open-book examinations (e.g., for students: how to study and take such an exam; for instructors: how to design and grade such an exam).

Here, our educational goal (i.e., integrating material from the first-year courses, and demonstrating adequate mastery of that material in an applied setting which would provide practice in the skills required by the master's project and beyond) pertains to higher-order thinking and application of course content rather than recall of specific facts, and thus the open-book format seems particularly apt. Indeed, implicit in the notion of matching examination format with educational goals is that the educational goals are sound and well specified. Perhaps more obvious in retrospect than it was in real time, one of our programmatic challenges turned out to be getting the goals right and, indeed, as our educational goals evolved toward greater specificity the QE correspondingly evolved toward becoming better aligned with its goals.

Considering organizational dynamics, we speculate that part of the difficulty in establishing pedagogically appropriate goals was a lack of appreciation of their importance among instructors whose primary training was in mathematics and statistics rather than education. Among such instructors the temptation to immediately focus on course content rather than beginning with the end in mind was quite seductive and, despite our best intentions, this is essentially what transpired.

Two potential concerns about this new format, mentioned both in the literature and in meetings of the teaching faculty, pertain to cheating and rigor. Regarding cheating, given that collaboration is allowed, students could potentially plagiarize the words of their peers. At least some degree of protection is provided by software which assesses the degree of similarity between documents. The primary protection, however, is provided by the possibility of a follow-up interview – if a student cannot adequately explain their work then the level of skepticism is high and, conversely, if they can explain their work they have learned something in the process, even if they have also borrowed from others. Not to mention that the alternative approach – namely, attempting to proctor a week-long exam remotely – runs the risk of becoming oppressive and unsuccessful. An additional protection might be provided by providing students with slightly different versions of the dataset to be analyzed (although we did not do so), with the expectation that the results of the data analysis should be substantially similar but not identical.

Ultimately, our approach to the question of cheating relied on a number of considerations. First, and especially given the QE was administered during a pandemic with students located off campus, we did not feel that we could realistically prevent collaboration and thus ought to place all students on an equal footing in that respect. Second, allowing collaboration increased the level of authenticity of the QE, since collaboration is typically a crucial element of actual biostatistical practice. Moreover, by encouraging students to discuss the QE with instructors in addition to their classmates, our hope was that students would gain increased confidence about their answers and thus feel less incentive to borrow inappropriately from others. Finally, by embedding the possibility of an interview we intended to signal to students that they would ultimately be held responsible for their own work.

Regarding rigor, the literature raises the possibility that open-book examinations might be considered to be less rigorous than their closed-book counterparts, especially in high-stakes situations where the exam in question is used to document the test-taker's formal credentials (Durning *et al*, 2016). As might be anticipated from the history of the QE, the importance of rigor was particularly prominent in meetings of the teaching faculty. Although the rationale for maximum rigor was considered to be so self-evident as to not require elucidation, we speculate that the proponents of maximum rigor implicitly began with the sound premise that expert performance is closely-tied to a rich well-organized critical mass of content knowledge of biostatistics which is immediately retrievable from memory, and then essentially assumed that "more knowledge must be better than less", and so the more information which is immediately retrievable from memory the better. Unfortunately, the idea of storing large amounts of information within memory is inconsistent with how biostatistics is actually practiced (even by the advocates of maximum rigor) and, moreover, gratuitous educational requirements are fundamentally unfair, not even to mention the possibility that they might place students from non-traditional backgrounds at especially increased risk.

Although not providing the same type of qualitative data as being marooned on a tropical island, the workings of an educational program might also be considered to be a case study in cultural anthropology (among others). In that spirit, we might note that within our particular environment rigor is such an emotionally powerful construct that, as a practical matter, it is next to impossible to argue against in the abstract, nor is it likely that the less rigorous of two alternatives will be chosen when the choice is framed as less rigor versus more. But, what can sometimes be done is to reframe the question in terms of precisely which goals are we trying to achieve, and then how to best do so, with that “best” approach also being declared to be “rigorous”. If, in so doing, we reengineer our educational programs to be more effective and humane, then so much the better.

Above and beyond considering what the revised format of the QE is intended <u>not</u> to achieve (i.e., gratuitous rigor, facilitating cheating), the revision process led us to reflect on what we <u>are</u> ultimately trying to accomplish. One way to describe our intentions is operationally – that is, that we are trying to design a QE that meets the educational goal of “integrating material from the first-year courses, and demonstrating adequate mastery of that material in an applied setting which would provide practice in the skills required by the master’s project and beyond”. Another way to describe our intentions is conceptually, through three types of fairness we want to strive toward. One is educational process fairness, where we strive to enable students to live up to their potential, regardless of differing educational backgrounds and learning styles. Another is cultural and linguistic fairness, where we strive to present and test in a way that is sensitive to the many social, cultural and linguistic backgrounds of our students. Yet another is goal fairness, where we strive to present, test, and more generally challenge our students intellectually in ways that will benefit them in their chosen path. To paraphrase: a “fair” examination covers important material which is beneficial to students, and does so in a way that is transparent and puts everyone in a position to perform their best work. In doing so, it is important to clearly convey both the goals to be achieved by the examination and the methodology for achieving a fair assessment -- for example, the student comment that the QE was good preparation for what will be expected after graduation was particularly encouraging.

To summarize: we have described the evolution of a qualifying examination from a traditional closed-book format to an open-book take-home collaborative format. Our intention was that the QE would become "a learning process through knowledge transfer as well as exercise in thinking skills" (Theophilides & Koutselini, 2000). One possible implication for the reader is the importance of matching examination format with teaching goals, which in turn implies that goals are concretely specified as a skill, performance or an attribute. Another possible implication is the observation -- perhaps counterintuitive within a mathematically-based discipline which values mathematical and logical rigor -- that more rigor isn't always better.

The cast of Gilligan's Island believed themselves to be embarking on no more than a "three-hour tour", albeit with unintended consequences. Upon reflection, our previous iterations of the QE were essentially a "five-hour tour", with similarly unintended consequences. There is much to be said for simply fixing the boat and leaving the island.

References

Durning, S. J., Dong, T., Ratcliffe, T., Schuwirth, L., Artino, A. R., Boulet, J. R., & Eva, K. (2016). Comparing open-book and closed-book examinations: A systematic review. *Academic Medicine, 91*, 583-499. https://doi.org/10.1097/ACM.0000000000000977

Ioannidou, M. K. (1997). Testing and life-long learning: Open-book and closed-book examination in a university course. *Studies in Educational Evaluation, 23*(2), 131-139. https://doi.org/10.1016/S0191-491X(97)00008-4

Ippersiel, D., & el Atia, S. (2014). Assessing graduate attributes: Building a criteria-based competency model. *International Journal of Higher Education, 3*(3), 27-38. https://doi.org/10.5430/ijhe.v3n3p27

Theophilides, C., & Koutselini, M. (2000). Study behavior in his closed-book and the open-book examination: A comparative analysis. *Educational Research and Evaluation: An International Journal on Theory and Practice, 6*(4), 379-393. https://doi.org/10.1076/edre.6.4.379.6932

Appendix 1: Masters Qualifying Exam

General directions:

This examination is intended to assess how well you have integrated material from your first-year classes, and also to provide practice in applying that knowledge to two common study designs: a randomized trial and a genomic study. There is a significant emphasis on communication and interpretation.

This examination is take-home, open-book, open-note, open-technology (e.g., web searches are allowed). You are welcome to discuss the content among yourselves, with your instructors, etc. Our intention is to approximate the conditions under which you will be practicing biostatistics, and also to induce as little stress as possible. There are two exceptions. First, although you may discuss R and SAS code among yourselves, each of you should perform the analyses independently. Second, each of you is responsible for writing up the results on your own, and answers which are overly similar will be looked upon with skepticism.

We might choose to schedule follow-up interviews about some or all of the exam content. Reasons for doing so could include providing an opportunity to clarify your answers, and to support a critical review of the effectiveness of the exam questions. We plan to follow up the exam with an anonymous survey, and would appreciate your assessment of the extent to which the exam was fair, appropriate, and allowed you to perform your best work.

You will be analyzing two datasets: RCT and GENOMIC, both of which have an Excel version and a SAS version. The versions are effectively identical.

Your answers are due in 7 days.

Question 1: randomized trial

A pain clinic regularly records patients' self-rating of pain, using a single question, where respondents use a slider bar to report their pain from 0 (no pain) to 100 (worst possible pain). Responses can be treated as continuously scaled and considered to be normally distributed. The minimum clinically important difference is believed to be 5 units. Denote this pain scale by Y.

To be eligible to participate in this clinic, patients must have suffered chronic pain for a significant period of time. Clinicians report that, for such patients, the value of Y varies over time around a central value, which varies across patients. In symbols: a simple statistical model of this phenomenon denotes the observation at time "j" for patient "i" as Y_{ij}, where Y_{ij}=M+C_i+E_{ij}, with M denoting the overall mean pain score for the population, C_i denoting patient-specific differences in central values, and E_{ij} a random error term. C_i and E_{ij} are both normally distributed with mean 0 and standard deviations σ_p and σ_e, respectively. All the error terms are independent of the others.

Investigators are testing a new drug for chronic pain, and plan to randomize patients to either this new drug N or the usual drug U. To be eligible, the patient must report a pain score of greater than 65 at baseline (i.e., Y0>65). Patients are then randomized receive either N or U, and then they report pain scores 1 month later. Denote these latter pain scores by Y1. Y1 is the primary outcome variable.

The dataset RCT contains the following variables:

- X1 = {N, U}, the study group
- Y0 = pain score at baseline
- Y1 = pain score after receiving the drug (primary outcome)
- X2 = {no, yes}, indicating the presence of a genetic polymorphism affecting the biological pathway through which the new drug is believed to operate
- X3 = {low, medium, high}, a summary of various psychological variables believed to have an impact on perception of pain

Part 1: A statistically inexperienced investigator knows enough to recognize that it would be useful to use the data to assess the assumption of normality, but wrongly believes that the thing they should do is to plot the values of Y1 for the entire population and see how close this is to a bell-shaped curve. Explain, in plain English, why this approach is incorrect and what the investigator should do instead.

Part 2: Explain, in plain English, what a statistical interaction is. Based on the information above, explain which interaction is more likely if the new drug, N, really works: an interaction of X1*X2 or an interaction of X1*X3.

Part 3: A typical statistical analysis plan for a randomized trial begins with a descriptive analysis of the cohort and the outcome, followed by an unadjusted assessment of the efficacy of the intervention, then proceeds to an adjusted assessment of the efficacy of the intervention, and then proceeds to an assessment of the consistency of the intervention effect across subgroups. Perform these analyses, and then prepare a report including statistical methods and results. Ideally, an investigator would be able to cut and paste from this report directly into a manuscript, so organize your work into the following table of contents: (1) descriptive analyses; (2) unadjusted analyses; (3) adjusted analyses; and (4) subgroup analyses. Be sure to address magnitude, precision, statistical significance and clinical significance.

Part 4: Each of the statistical tests in part 3 can be defined as a comparison between a full model and a reduced model. Prepare a table which lists the full and reduced models for each of these statistical tests. For example, one row of the table would be the unadjusted assessment of efficacy, for which the full model contains X1 as a single predictor and the reduced model has no predictors at all.

Part 5: Explain, in plain English, why including Y0 in the model as a covariate is statistically beneficial.

Part 6: Is the genetic polymorphism helpful, harmful, or neither?

Part 7: A dataset from an observational study has exactly the same set of variables as RCT. The only difference is that the treatment isn't assigned randomly, and instead is based on clinical judgment and other considerations. In plain English, define confounding, and explain why a randomized trial is unlikely to show confounding whereas confounding is quite possible in an observational study.

Part 8: Assume that an observational study gives rise to a dataset with the same structure as RCT. Explain, in plain English, how you would go about assessing confounding.

Part 9: Investigators sometimes confuse confounding and interaction. Provide a formal statistical definition of each, and then a plain English definition of each. Finally, in plain English explain why these two concepts are different.

Question 2: genomic study

The dataset GENOMIC (n=2,000) has samples from 2,000 subjects and the following variables:

- Y = phenotype

- X1-X1000 = expression results for 1,000 genes
- Set = 'train' for 1,000 samples in the training set and 'valid' for 1,000 samples in the validation set.

Y and X1-X1000 have been rescaled to have mean 0 and standard deviation 1 (approximately). As an analyst, you have been asked to identify which of the 1,000 genes are the most promising (i.e., are most strongly associated with the phenotype) and which, if any, rise to the level of statistical significance.

Parts 1-12 of this question use the entire dataset (i.e., all 2,000 samples). Parts 13-16 treat the training set and the validation set separately.

Part 1: Using 1,000 bi-variate analyses (e.g., X1 versus Y, X2 versus Y, … X1000 versus Y), determine which of the 1,000 predictors are statistically significant using an alpha (i.e., the type 1 error rate) of 0.05.

Part 2: In plain English, what is the “multiple testing problem”? In other words, why might some of the statistically significant results in part 1 be false positives?

Part 3: In non-technical terms, the “global null hypothesis” states that none of the 1,000 predictors matter. State the global null hypothesis in statistical terms (i.e., as a statistical hypothesis).

Part 4: Suppose that the global null hypothesis is true. Using alpha=0.05, how many statistically significant results would you expect in part 1?

Part 5: Suppose that the global null hypothesis is true. Let X denote the number of statistically significant results from part 1. What is the distribution of X? Include its name, the mean, the standard deviation, and its support set. (You don't have to derive the mean and standard deviation, and can simply look this up if you prefer.)

Part 6: Using an exact distribution, perform a statistical test of the global null hypothesis using X from part 5.

Part 7: Using a normal approximation, perform a statistical test of the global null hypothesis using an appropriately transformed version of X from part 5. (You will need the mean and standard deviation of X from part 5.)

Part 8: An investigator might ask whether or not the statistical tests in parts 6 and 7 are well-powered to detect cases where a small number of genes matter. How would you reply? (There is no need to perform a formal power calculation – it is sufficient to apply logic.)

Part 9: Let X denote the p-values from the statistical tests in part 1. (The answer to part 1 generated 1,000 realizations of X.) Suppose that the global null hypothesis is true, what is the distribution of X? Include its name, the mean, the standard deviation, and the support set.

Part 10: Using X from part 9, how could a Q-Q plot be used to assess this global null hypothesis? What pattern would you expect to see if this global null hypothesis is true? What pattern would you expect to see if this global null hypothesis is false?

Part 11: Create the Q-Q plot in part 10, and describe your conclusions.

Part 12: Reassess the results of part 1 using a more appropriate type 1 error rate. Which (if any) of the 1,000 predictors are statistically significant using this more stringent alpha? Justify your choice of alpha.

The next set of analyses uses the first 1,000 samples as a training set and the next 1,000 samples as a validation set.

Part 13: The investigators don’t want to miss any potentially important predictors, would like for you to select the value of alpha in the training set to be relatively liberal value while still accounting (to a degree) for the impact of multiple comparisons. What value of alpha would you choose? There is no single correct answer to this question, but you should justify your choice.

Part 14: For the training set, using 1,000 bi-variate analyses (e.g., X1 versus Y, X2 versus Y, … X1000 versus Y), determine which of the 1,000 predictors are statistically significant (i.e., using the value of alpha that you selected in part 13).

Part 15: For the validation set, determine which (if any) of the statistically significant predictors from part 14 are statistically significant in the validation set. Use an appropriately chosen value of alpha (which might be different from the one used in the training set). Justify that choice of alpha.

Part 16: In plain English, explain the results of parts 13 through 15. Your explanation should include the logic behind model validation.

Part 17: Which types of genetic effects will the approach in parts 1-16 tend to find and which types of genetic effects will this approach tend to miss? (This is a question about biology, not about statistics.)

Copyrights