Testing for over- and under-dispersion in physics degree outcomes.

Astra Sword
School of Physical Sciences, The Open University, Milton Keynes, England, United Kingdom, MK7 6AA

As the scale of quantitative data available to physics education researchers grows, it is imperative that we critically assess how well the assumptions behind "standard" statistical methods apply in our field. In the present work, I give a background on a common statistical assumption used to analyse proportion data, the binomial assumption; I discuss scenarios in which this assumption may break down in the context of education research; and test this assumption using a large population-level data set. This data set comprises academic outcomes (rate of 'good degrees') for all undergraduate physics degree programs that ran in the UK across the 2012/13–2018/19 period (26,960 students across 79 programs). I estimate dispersion parameters and their significance for each program in the data set and discuss the implications of the results for analysing proportion data in physics education research.
I. INTRODUCTION

As quantitative physics education research increasingly makes use of larger data sets [1] there is a great potential for producing broad, wide reaching results that can help inform policy and practice on a national, and even international, level. However, with the scale of the data, the impact of errors in analysis also grows. As such, special care must be taken with the application and interpretation of the statistical methods used to analyse these data:- with a critical view to understanding where the (often unspoken) assumptions that underlie these standard techniques break down.

This is particularly true for quantitative diversity, equity, and inclusion (DEI) research, where the principal research subjects of interest are people from socially disadvantaged—and consequently, often vulnerable—demographic groups. Within PER, there is a rich and growing literature addressing the breakdown of conceptual and statistical assumptions in PER-related DEI research addressing topics like the problems with using a binary framework for gender [2], omitted variable bias in the analysis of demographic gaps [3], and general issues of causal inference [4].

In this paper I contribute to this wider body of work by using a large national data set containing degree outcomes for UK physics graduates to test for violations of a particularly ubiquitous statistical assumption: the binomial assumption. Along with its generalised (multinomial) form, this assumption underpins a number of techniques used to analyse these data:- with a critical view to understanding where the (often unspoken) assumptions that underlie these standard techniques break down.

The research question for this study is: "For each degree program in the data, does the observed variance in the good degree rate match the expected variance under the binomial distribution?" See Section II. & III. for definitions.

In addition to making a general contribution to the literature, this work is part of a wider project using the aforementioned large national data set to understand what the demographic gaps in physics are, their causes, and what role physics-as-a-subject plays in their evolution. It is hoped that the results of this analysis will help inform similar projects in the future.

II. BACKGROUND

A. The Binomial Distribution

The binomial assumption is the assumption that a collection of binary response variables, for example \( Y_i \in \{\text{Failure}, \text{Success}\} \), are Bernoulli variables that are both (a) identically distributed and (b) independent.

\[
Y_i \sim \text{Bernoulli}(\pi)
\]

In other words, \( Y_i = 1 \) with a probability \( \pi \) and \( Y_i = 0 \) with probability of \( (1 - \pi) \). This is equivalent to declaring that your response variable is generated by a process equivalent to flipping a weighted coin that has probability \( \pi \) of getting heads for each observation \( i \).

In practice, we usually observe binary response variables (such as whether a student has passed a particular course) in clusters (i.e. cohorts). If we observe \( K_j \) successes per cluster of size \( m_j \), we can say \( K_j \) is independently identically distributed according to:

\[
K_j \sim \text{Binomial}(m_j, \pi)
\]

Both forms for the data (1) and (2) are strictly equivalent; though usually (2) tends to be the representation used when dealing with proportion data, such as the fraction of students passing a course. In this case, the expected pass rate for a cluster is \( \text{E}(Y_i) = \pi \) with a variance of \( \text{Var}(Y_i) = \pi (1 - \pi) \). Alternatively we could express these as the mean and variance of \( K_j \) instead; \( \text{E}(K_j) = m_j \pi \text{ and } \text{Var}(K_j) = m_j \pi (1 - \pi) \).

This brings us to the crux of the issue: the variance of binomially distributed data is uniquely determined by the mean. (Contrast this with normally distributed data, where the mean and variance can vary freely.) Having only one free parameter results in the fit between the binomial distribution and the real data being more fragile.

B. Relaxing the Binomial Assumptions

As noted, modeling the pass rate of a cohort of students as binomial is equivalent to declaring (a) that each student has exactly the same chance of passing and (b) that success of each student is independent of every other student. Both these seem unlikely, consider these hypothetical motivating arguments for why these assumptions may not apply:

(a) **Diverse chances of success.** We expect a student’s chance of success to depend on what they already know—experience suggests this is unlikely to be the same for every student, and so chance of success should be different for different students too.

(b) **Correlated outcomes.** Students do not study in isolation—in fact, they are typically encouraged to form study groups for working on problems. For assessments where students can share answers, success is likely to be correlated within each study group.
While these two arguments may not be universally applicable to every classroom, their plausibility suggests that we should expect student outcomes to not meet the assumptions of the binomial distribution at least some of the time.

1. Diverse chances of success

When the assumption of identical distribution is not met, the extra uncertainty in the chance of success $\pi$ leads to an overdispersed distribution [7], that is to say a distribution with larger variance than would be expected from a binomial model with the same mean value of $\pi$. In practice, this can be modeled by specifying an underlying distribution for $\pi$. The most common choice is the beta distribution which results in using a beta-binomial distribution to model the data [8]. Figure 1 shows an example overdispersion due to beta-distributed $\pi$. Note that allowing $\pi$ to vary always leads to a distribution that is overdispersed relative to the binomial [7].

2. Correlated outcomes

When the assumption of independence between outcomes is not met, there are two main possibilities. If outcomes are positively correlated—such as when students working together on a problem submit similar answers—this also leads to the data being overdispersed relative to the binomial [7]. However, if outcomes are negatively correlated—one student's success makes other students less likely to succeed—then the data will be underdispersed, that is exhibit less variance than the corresponding binomial distribution [7].

What could cause student outcomes to be negatively correlated with one another? While this is not something I have seen discussed in the wider literature in the context of underdispersion, I suggest at least two plausible mechanisms:

- **Shared, finite resources.** While students typically have access to the same lectures and materials, some academic resources are necessarily finite and must be shared with other students. Possible examples include informal one-to-one tuition from instructors; places on enrichment programs; or access to non-academic resources that may contribute to success such as affordable housing close to campus.

- **Post-hoc marking of assessments.** This is the adjustment of student scores and adjustment of grade boundaries for an assessment after it has taken place. Specific post-hoc marking practices are diverse, but may include adjusting the pass mark so the expected number of students pass [9]. As a consequence, each student that fails the assessment increases the chance other students will have their mark adjusted past the pass threshold post-hoc, leading to negatively correlated outcomes.

C. Implications for Statistical Analysis

It is important to note that under- or overdispersion of a outcome variable across any particular groupings within a data set is not necessarily a statistical issue in and of itself. Statistical techniques that rely on the binomial or multinomial assumptions, like those listed in the introduction, only require that those assumptions hold for the outcome variable conditioned on the regressors included in the model [10].

In practice, it is often impractical or impossible to take into account all the factors that may impact student outcomes. The extent of non-binomial variation in a data set can be summarised using the dispersion parameter:

$$\phi = \frac{\sigma_{\text{obs}}^2}{\sigma_{\text{exp}}^2}$$

Where $\sigma_{\text{obs}}^2$ is the variance observed in the data set and $\sigma_{\text{exp}}^2$ is the variance expected under a binomial assumption. If overdispersion ($\phi > 1$) is present, models using the binomial assumption will more likely to produce false-positives, while for underdispersion ($\phi < 1$), the false-negative rate will be higher. In both cases the true confidence interval (CI) can be estimated by multiplying the naïve CI obtained under the binomial assumption by $\sqrt{\phi}$ [8].

D. Further Reading

Several standard texts address overdispersion, albeit briefly. See [11] for a brief overview. For a more technical treatment [8][10] are good choices, with the former being focused exclusively on categorical statistics and the latter having broader coverage of statistical models. Where underdispersion is noted in these and similar texts, it is usually just to mention that it is rare. However, [12] does have a chapter introducing both over and underdispersion, as well as a chapter with material on under-dispersed Poisson models.


III. METHODS

If Section II. establishes plausible scenarios for non-binomial dispersion the question remains: is there empirically relevant over- or under-dispersion in academic outcomes in physics? To answer this question, I estimated the root dispersion parameters and their statistical significance for a large data set comprising all graduates of accredited physics degree programs that ran in the UK across the full 7 years spanning 2012/13 and 2018/19 (26,960 students; 42 institutions).

A. Context

In the UK, undergraduate degree programs typically focus on a single subject, chosen by the student at the start of their studies. On graduation, students are awarded a bachelor’s degree (BSc) with a classification determined by their academic performance. Of these, the highest are the first class (1st) and upper second (2:1) which are roughly equivalent to a 3.7+ GPA and a 3.3+ GPA respectively [13]. The percentage of graduates on a particular program or at an institution achieving a 1st or a 2:1 is known as the good degree rate and is used across the sector to measure academic success.

At most UK institutions delivering physics programs, high performing students often have the option to extend their program by a year to gain an "enhanced" degree, graduating with a master’s (MSc) rather than a "non-enhanced" bachelor’s level qualification. Broadly these two programs will be the same for the first three years at any particular institution, with the students being in a common cohort and having the opportunity to transfer from and to the enhanced program—though the exact details vary by institution.

Both enhanced and non-enhanced physics programs in the UK undergo accreditation by the Institute of Physics (IOP), which requires accredited programs to teach a standard set of topics [14].

B. Data Source

The data in this study is a subset of a larger data set procured from the Higher Education Statistical Agency in the UK (HESA) for the purposes of studying equity in physics degree outcomes. This larger data set includes all students that have studied first-degrees in the UK between 2012/13 and 2019/20, not just physics students, and comprises approximately 4 million students in total.

Due to the sensitivity of this data, my reporting is subject to a number of stipulations [15] including:

(a) Counts of students are rounded to the nearest 5.

(b) Individual higher education providers must not be identifiable.

A consequence of (b) is that I am unable to report the exact size of any of the specific physics programs investigated here as that information could be used to identify them.

C. Identification of Physics Degree Programs

I identified physics degree programs as those that met the following criteria:

1. The degree program is accredited by the IOP [14].

2. The degree program has "physics" in the title (i.e., is not a natural science degree).

3. The program’s accreditation is for the degree program as a whole, rather than just for a subset of students on the program that take specific modules.

I considered all programs that met these criteria at a particular institution and that led to the same level of qualification (i.e., enhanced vs non-enhanced) to belong to the same physics program.

D. Statistical Inference

While the dispersion parameter (3) is the standard way of measuring non-binomial dispersion, to aid interpretation I calculated the root dispersion parameter instead:

$$\sqrt{\phi} = \frac{\sigma^2_{\text{obs}}}{\sigma^2_{\text{exp}}} \approx \frac{\text{True CI}}{\text{Naïve CI}}$$

(4)

Where $\sigma^2_{\text{obs}}$ was the observed year-on-year variance calculated across the 7 years in the sample, and $\sigma^2_{\text{exp}}$ was the variance expected under a binomial assumption. In theory, for a program of fixed size each year, the expected variance in the good degree rate $\pi$ would simply be $\pi(1 - \pi)$ and thus $\sigma^2_{\text{exp}}$ could be calculated analytically. However, as the number of students graduating in any particular cohort varies year on year, a different approach was needed. For this analysis, I used the testDispersion function of the R library DHARMa [6] to estimate the distribution of expected variance via simulation. I used 1,000,000 iterations to estimate the distribution under the null hypothesis for each program.

Post-hoc Analysis. Given that this analysis involved a large number of significance tests, I used the statmodels [16] implementation of the Benjamin-Hochberg procedure to control the false discovery rate at 5% [17].

IV. RESULTS

a. Descriptive statistics. Across the 42 intuitions in the data set, there were 42 non-enhanced (BSc) and 37 enhanced (MSc) degree programs—this is in line with expectations as most institutions deliver both types. The average cohort size per year varied considerably for both non-enhanced ($M = 45, SD = 20$) and enhanced programs ($M = 50, SD = 40$), while the average good degree rate per year across the period varied moderately for both non-enhanced ($M = 57\%, SD = 15\%$) and enhanced programs ($M = 89\%, SD = 10\%$). The total number of graduates of non-enhanced and enhanced programs were similar, 13,465 and 13,495 graduates respectively.
FIG. 2. Swarm-and-box plot showing the empirical distribution of the root dispersion parameter for the good degree rate of enhanced and non-enhanced physics degree programs in the UK for the 2012/13 to 2018/19 period.

b. Root dispersion parameters. Figure 2 illustrates the distribution of estimated root dispersion parameters for both non-enhanced ($M = 1.10$, SD = 0.29) and enhanced degrees ($M = 1.02$, SD = 0.12), with those that were significant at the $\alpha = 0.05$ level highlighted.

c. Significance. Prior to adjusting the p-values via the Benjamin-Hochberg procedure, 16 programs showed significant non-binomial dispersion (Table I). Controlling the false discovery rate, only programs A and G were found to have significant non-binomial dispersion at the $\alpha = 0.05$ level; though a further 4 were 'near misses' at $p < 0.051$.

TABLE I. Root dispersion parameters, significance and adjusted significance for all programs with significant dispersion prior to the Benjamin-Hochberg post-hoc adjustment.

<table>
<thead>
<tr>
<th>Program</th>
<th>$\sqrt{\phi}$</th>
<th>p-value (unadjusted)</th>
<th>p-value (adjusted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Enhanced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1.90</td>
<td>0.00024</td>
<td>0.022</td>
</tr>
<tr>
<td>B</td>
<td>1.65</td>
<td>0.022</td>
<td>0.12</td>
</tr>
<tr>
<td>C</td>
<td>1.47</td>
<td>0.014</td>
<td>0.12</td>
</tr>
<tr>
<td>D</td>
<td>1.44</td>
<td>0.021</td>
<td>0.12</td>
</tr>
<tr>
<td>E</td>
<td>1.37</td>
<td>0.0027</td>
<td>0.051</td>
</tr>
<tr>
<td>F</td>
<td>1.34</td>
<td>0.021</td>
<td>0.12</td>
</tr>
<tr>
<td>G</td>
<td>1.33</td>
<td>0.0010</td>
<td>0.036</td>
</tr>
<tr>
<td>H</td>
<td>1.27</td>
<td>0.0099</td>
<td>0.08</td>
</tr>
<tr>
<td>I</td>
<td>0.63</td>
<td>0.035</td>
<td>0.17</td>
</tr>
<tr>
<td>J</td>
<td>0.41</td>
<td>0.0038</td>
<td>0.051</td>
</tr>
<tr>
<td>Enhanced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>1.37</td>
<td>0.0047</td>
<td>0.053</td>
</tr>
<tr>
<td>L</td>
<td>1.30</td>
<td>0.0026</td>
<td>0.051</td>
</tr>
<tr>
<td>M</td>
<td>1.20</td>
<td>0.0033</td>
<td>0.051</td>
</tr>
<tr>
<td>N</td>
<td>1.11</td>
<td>0.023</td>
<td>0.12</td>
</tr>
<tr>
<td>O</td>
<td>0.86</td>
<td>0.0072</td>
<td>0.08</td>
</tr>
<tr>
<td>P</td>
<td>0.83</td>
<td>0.016</td>
<td>0.11</td>
</tr>
</tbody>
</table>

VI. DISCUSSION

In the context of UK based education research, this data demonstrates the existence of physics programs with non-binomially distributed good degree rates: namely programs A and G showing clear statistical significance; with programs E, J, L and K being very close to significance. The root dispersion parameters cover a wide range, from 0.41 (Program J) to 1.9 (Program A). Together with Equation 4, these parameters suggest the true confidence intervals for these programs range from under half through to nearly twice the confidence intervals that would be calculated under the binomial assumption. These large departures from the binomial assumption have clear empirical relevance for how differences in good degree rate should be interpreted between these programs.

The presence of underdispersed ($\sqrt{\phi} < 1$) programs is particularly interesting given that underdispersion arises from the anti-correlation of outcomes between students. This may reveal that for these programs the mechanisms suggested for the anti-correlation in Section II. (e.g., resource scarcity, post-hoc marking) have a dominant presence in these programs compared to the mechanisms that lead to overdispersion (e.g., diverse success rates, students working together). Similarly, the stark contrast in the spread of root dispersion parameters between program types may be reflective of the higher entry requirements for enhanced programs, which leads to a narrower range of individual student success rates—thus reducing the extent of any overdispersion.

For PER as a field, the presence of statistically significant and empirically relevant non-binomial dispersion in this data set suggests that the validity of the binomial assumption cannot be taken for granted. I propose three recommendations for PER researchers to account for non-binomial dispersion going forward:

- **Interpret binomial tests with caution.** Consider reporting how much non-binomial dispersion would be needed to alter the interpretation of any associated p-values or confidence intervals.
- **Use dispersion-aware models.** Quasi-likelihood models can account for both over- and underdispersion. If dealing with overdispersion only, a parametric or mixed effects approach may be better. [8] [10]
- **Measure it!** Non-binomial dispersion is a feature, not a bug. Its size and direction is a clue about what is happening in the classroom. The dispersion parameter can be quickly estimated from goodness-of-fit statistics included the output of statistical packages (see [11]) but a simulation based approach is more robust [6].

VI. ACKNOWLEDGEMENTS

Special thanks to Sally Jordan, Annika Lohstroh, Rachel Hilliam, and Mark Jones for their encouragement and feedback on the draft of this paper. This project was funded by The Ogden Trust and The Open University.