Report I - Experiment Design in Computer Sciences

Matheus Silva de Lima[1]

Abstract

In this first report for the 2023 spring semester of Experiment Design in Computer Sciences, we analyze data from the Brazilian national high-school examination (Enem). One characteristic of this exam is that, in addition to other cheating prevention measures, the exam comes in 4 different question orders. In this report, we are interested in evaluating if the question order within the exam affect the overall performance of the examinee. In order to evaluate this, we analyze the statistics from the 2021 exam.

1. Introduction

In Brazil, the National High-School Exam (ENEM) is the main entrance exam for most Brazilian nationals who wish to continue to study into the higher education system. Every year, the exam mobilize millions of people, who takes the exam synchronously, in all states, during the spam of two days. In 2021, the exam had over 3.3 million people registered.

The exam is organized by the Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (National Institute for Studies and Research on Education Anísio Teixeira, free translation), an agency connected to the Ministry of Education. Every year this institute provides a public dataset with information about, between other things, the examenees performance and social-economic background. Information contained in this dataset is anonymous, and therefore it is not subjected to the General Personal Data Protection Law.

Since 2017, the exam takes place in two consecutive Sundays. In 2021, it was administrated on the 21 and 28 of November. On the first day, students have 5.5 hours to solve a "Foreign Languages" (English or Spanish) exam, "Human Sciences" exam, and write an essay. On the second day, students have 5 hours to solve a "Natural Sciences" exam, and a "Mathematics" exam. Each exam has 45 questions, with a total of 180 questions.

As an extra measure to prevent cheating within the exam, each of the four exams comes in four different question orders, shown by the exam color: blue, yellow, pink or gray/white. Questions are randomly shuffled for each color, and all four exam colors are equally distributed between students in no particular

[1] Computer Vision Laboratory, University of Tsukuba

Table 1: Number of Students per exam order. Students are evenly distributed between all exam orders.

	Languages	**Human Sci.**	**Natural Sci.**	**Mathematics**
Blue	517.132	517.132	517.658	517.658
Yellow	524.356	524.356	523.253	523.253
Pink	518.465	518.465	518.543	518.543
Gray	517.563	517.563	518.395	518.395

order. The number of students per exam color for the 2021 exam is shown in Table 1.

A natural question that can arise in this exam system is if the exam order affects the student's final score. To answer this question, we analyze the data from the 2021 exam, comparing statistics from all four exam orders.

The remain of this report is organized as follows: in Section 2 we explain the methodology of data processing and data analysis; in Section 3, we report the results of this analysis; and in Section 4, we conclude this report.

2. Methodology

In this research, we use the Microdados Enem 2021 dataset [1]. We restrict this analysis for candidates without disabilities who made the exams and essay on the 21 and 28 of November, 2021. The total amount of candidates who abide to these restrictions are 2.077.516 on the 11/21, and 2.077.849 on the 11/28.

We organize the data by subject area and exam order, and evaluate the influence of exam order in the candidate score by comparing the mean and standard deviation of all data distributions. We then normalize the distribution to get a probability distribution, and calculate the pairwise Kullback–Leibler (KL) divergence for all exam orders in $\log_2$. Here, the KL divergence is a measurement of information difference between distributions, and is measured in bits. We show this information by a confusion-like matrix.

Finally, we calculate the confidence intervals of the mean using a Normal distribution with a confidence score of 99.9%.

3. Results

The mean and standard deviation statistics can be seen o Tables 2. A graphic representation of this can be seen in Figure 1. The largest difference in mean seen in Table 2 is between between colors Pink and Yellow of the Mathematics exam, with a difference of 6.87 points.

The distribution of scores per student can be seen in Figure 2. Visually, the distributions look very similar, but as a comparison method we calculate the KL divergence between all pairs of normalized distribution per exam. Results are shown by the matrices on Figure 3.

Table 2: Exam mean score and standard deviation per color.

	Languages	**Human Sci.**	**Natural Sci.**	**Mathematics**
Blue	$\mu = 509.09$, $\sigma = 77.15$	$\mu = 526.65$, $\sigma = 92.95$	$\mu = 492.55$, $\sigma = 81.53$	$\mu = 538.47$, $\sigma = 109.26$
Yellow	$\mu = 506.31$, $\sigma = 77.08$	$\mu = 524.82$, $\sigma = 94.19$	$\mu = 494.04$, $\sigma = 80.13$	$\mu = 536.08$, $\sigma = 111.76$
Pink	$\mu = 507.43$, $\sigma = 76.99$	$\mu = 524.48$, $\sigma = 94.15$	$\mu = 494.95$, $\sigma = 80.78$	$\mu = 542.96$, $\sigma = 110.92$
Gray	$\mu = 505.64$, $\sigma = 78.42$	$\mu = 527.25$, $\sigma = 93.83$	$\mu = 493.52$, $\sigma = 80.94$	$\mu = 538.01$, $\sigma = 110.91$

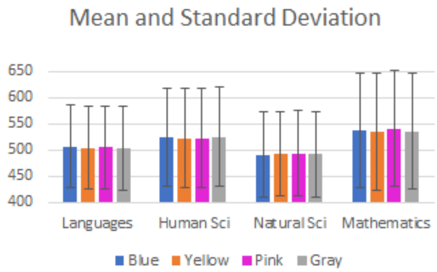


Figure 1: Mean and standard deviation visualization per exam.

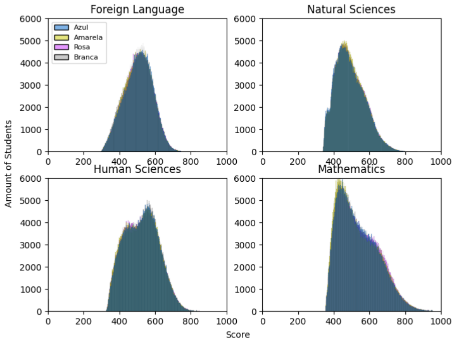
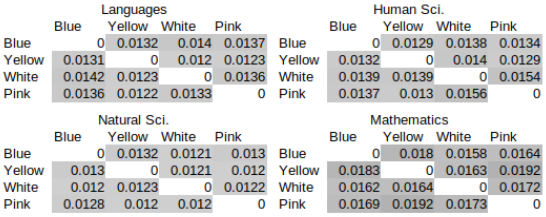


Figure 2: Distribution of scores per exam. It is clear that distributions are mostly overlap.



Languages

	Blue	Yellow	White	Pink
Blue	0	0.0132	0.014	0.0137
Yellow	0.0131	0	0.012	0.0123
White	0.0142	0.0123	0	0.0136
Pink	0.0136	0.0122	0.0133	0

Human Sci.

	Blue	Yellow	White	Pink
Blue	0	0.0129	0.0138	0.0134
Yellow	0.0132	0	0.014	0.0129
White	0.0139	0.0139	0	0.0154
Pink	0.0137	0.013	0.0156	0

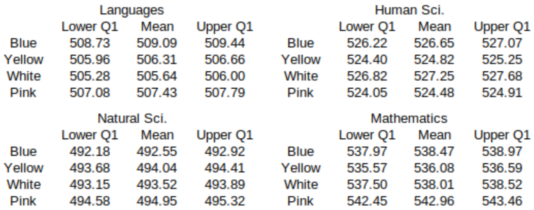
Natural Sci.

	Blue	Yellow	White	Pink
Blue	0	0.0132	0.0121	0.013
Yellow	0.013	0	0.0121	0.012
White	0.012	0.0123	0	0.0122
Pink	0.0128	0.012	0.012	0

Mathematics

	Blue	Yellow	White	Pink
Blue	0	0.018	0.0158	0.0164
Yellow	0.0183	0	0.0163	0.0192
White	0.0162	0.0164	0	0.0172
Pink	0.0169	0.0192	0.0173	0

Figure 3: KL divergence between exams distributions for each exam color (bits).



	Languages				Human Sci.		
	Lower Q1	Mean	Upper Q1		Lower Q1	Mean	Upper Q1
Blue	508.73	509.09	509.44	Blue	526.22	526.65	527.07
Yellow	505.96	506.31	506.66	Yellow	524.40	524.82	525.25
White	505.28	505.64	506.00	White	526.82	527.25	527.68
Pink	507.08	507.43	507.79	Pink	524.05	524.48	524.91

	Natural Sci.				Mathematics		
	Lower Q1	Mean	Upper Q1		Lower Q1	Mean	Upper Q1
Blue	492.18	492.55	492.92	Blue	537.97	538.47	538.97
Yellow	493.68	494.04	494.41	Yellow	535.57	536.08	536.59
White	493.15	493.52	493.89	White	537.50	538.01	538.52
Pink	494.58	494.95	495.32	Pink	542.45	542.96	543.46

Figure 4: Table of confidence intervals.

The largest information difference of distributions was also observed between exams Pink and Yellow of the Mathematics exam, that being 0.0192 bits.

Finally, the confidence interval of the mean are calculated assuming a Normal distribution and a confidence interval of 99.9%. This can be seen in Figure 4.

4. Discussion

Means and standard deviations of all exam colors are indeed very similar, with the largest difference observed being 6.87 points for the mean, or around 6% of the standard deviation. One could argue, however, that the KL divergence is a better metric for measuring the difference between data distributions. In this case, the largest difference observed was still smaller then 0.02 bits.

What was surprising, however, was the confidence intervals calculated for this dataset. In fact, many of then don't have overlaps, even for a confidence level of 99.9%. Perhaps this happened given the size of the dataset and the difference in means. Interestingly enough, the most affected exam was the mathematics one, having more than 2 times the mean variation observed in any of the others. This result can lead us to believe that the exam order can in fact have a slight affect on the students score, specially in the case of the mathematics exam, but since the exam is different every year and one cannot choose which exam to take, this cannot be exploited or easily controlled. Maybe a more in depth analysis could be done by, for example, evaluating the probability of getting a question right given the question before, and how similar are these questions.

References

[1] I. N. de Estudos e Pesquisas Educacionais Anísio Teixeira, Microdados do Enem 2021, [online], [citado 2022-05-25]. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem (2022).