Studentsourcing - aggregating and re-using data from a practical cell biology course

Joachim Goedhart[1]

[1]Molecular Cytology, SILS - University of Amsterdam, The Netherlands,

Corresponding author: Joachim Goedhart, j.goedhart@uva.nl

Abstract

Data that is generated by students during a course is often lost as there is no central, structured storage of those data. The loss of data prevents its re-use. To enable re-use of the data, I present an approach to collect, aggregate and re-use data that is generated by students in a practical course on cell biology. The course runs annually and I have recorded the data that was generated by students over 3 years. Two use-cases illustrate how this data can be aggregated and re-used either for the scientific record or for teaching. As the data is obtained by different students, in different groups, over different years, it is an excellent opportunity to discuss experimental design and modern data visualization methods such as the superplot. The first use-case shows how central data storage provides a unique opportunity to get precise quantitative data due to the large sample size. The second use-case demonstrates how the data can be presented as an online, interactive dashboard, providing real-time data of the measurements. Both use cases illustrate how data can be effectively aggregated and re-used.

Plain Language Summary

Data acquired by students has value and in this work we present ways to collect and re-use that data.

Introduction

Teaching practical skills in a lab course is a crucial part of education in biology, biomedical science, and life sciences. In these lab courses data is generated, reported and interpreted, much like *real* experimental lab work. However, students use their data just for their own lab report and the data is not centrally stored or aggregated. As a consequence, most of the data that is gathered in a lab course is lost. Yet, these data are potentially useful. Especially for larger course, an impressive amount of data under well-controlled conditions can be generated. Therefore, by collecting and aggregating the data of multiple students over multiple years, one can easily gather a large dataset with high numbers of independent observations, or as defined by Lazic et al. (2018), biological replicates and experimental units.

Microscopy is an essential tool in cell biology. The use of microscopes to observe cells and organisms has changed from a qualitative, descriptive approach, into a quantitative method (Renz, 2013; Senft et al., 2023; Wait et al., 2020; Waters, 2009). The development of digital cameras and image analysis software has catalyzed this transition (Carpenter, 2007). Therefore, experiments that use microscopes are often followed by bioimage analysis to extract quantitative information from the data. To teach these skills, we combine a basic course on microscopy in a course on cell biology with teaching image processing and analysis in ImageJ/FIJI (Schneider et al., 2012). In a typical year, over one hundred students are enrolled in this course and therefore, a substantial amount of data is generated in the course.

We decided to collect the data that was generated by the students in the lab course over several years and store the measurement results in a central location. The data by itself can be valuable for the scientific community as precise estimates with good statistics can be obtained. Moreover, the data are a starting point to discuss data visualization, experimental design and how experimental design affects the statistics and interpretation of data. Here, we report the methods to collect, process and visualize the data and demonstrate some of the ways the data can be re-used.

Methods

For full reproducibility, this document is written using Quarto (Posit, https://quarto.org/), and the source code of the manuscript and the notebooks, and the data are availble in a repository: https://github.com/JoachimGoedhart/MS-StudentSourcing

Use case 1

Sample preparation and measurements

HeLa cells are cultured according to standard procedures and seeded 1 or 2 days before the treatment on 12 mm diameter glass coverslips. HeLa cells are incubated with 10 µM EdU for 30 minutes at 37 ° C. The cells are fixed with 4% formaldehyde in PBS and permeabilised with 0.1% Triton X-100 in PBS. Click chemistry is performed with 9 µM Cy3-azide and 2 mM $CuSO_4$. To start the reaction, 20 mg/ml ascorbate (final concentration) is added and the solution is used immediately to stain the cells. After 30 minutes, the cells are washed 3x with PBS and the sample is incubated with 0.1 µg/ml DAPI for 5 minutes. Samples are mounted in Mowiol and used for observation with fluorescence microscopy. Images of at least 100 cells are acquired with the DAPI and TRITC filters sets. The nuclei in both channels are counted by hand, or in an automated way by segmentation and 'particle analysis' in imageJ to calculate the percentage of cells that are positive for Cy3 fluorescence, reflecting cells in the S-phase.

Data collection

The data of the measurements is collected on voluntary basis through a Google Form, an example of which is shown in *figure XX*. The data that is recorded is the group (A/B/C/D), the percentage of cells in the S-phase for two methods, i.e. manual and using ImageJ/FIJI. The form is easy to set up and the data is collected in Google Sheets, yielding four columns; Timestamp, Group, and two columns with percentages of S-phase determined by the two methods.

Data processing & visualization

The data that is in the Google Sheet can be downloaded and read into R (R Core Team, 2022) as a CSV. All subsequent processing and data visualization is done with `{rmarkdown}` (Allaire et al., 2022). The R Markdown file and a rendered version as HTML are available on Github: https://github.com/JoachimGoedhart/PreciSeR. The cleaning of the data consists of removing empty cells, changing the column names, conversion to a tidy format, forcing the data into a 'numeric' type and filtering for sensible values (anything outside the generous range of 0-100 will be removed).

Use case 2

Sample preparation and measurements

A buccal swab is used to harvest cheek cells by scraping ~5 times over the inside of the cheek. The tip of the sample collector is dipped into an eppendorf tube with 40 µl PBS, and the cells are transferred to an object slide by touching the slide with the tip. Next, 10 µl of 0.1% methyleneblue solution is added and the sample is enclosed by a square coverslip (22 x 22 mm). The sample is used immediately to observe the cells with a Leica microscope, equipped with a Lumenera camera. A 20x or 40x objective is used to observe and image the cells. A separate image of a micrometer (Electron Microscopy Sciences 6804208, Stage Micrometer S8, Horizontal Scale, 1 mm Length) is acquired at the same magnification. The images are processed in ImageJ/FIJI and the dimensions of the images are calibrated with the micrometer image. The line tool is used to measure the diameter of the cells (the longest axis).

Data collection

The data of the measurements is collected through a Google Form, an example of which is shown in *figure XX*. The data that is recorded by the form is the group (A/B/C/D), the size measurements of the cheek cells and the size measurements of the nucleus. The data is aggregated in a Google Sheet which has four columns with data on Timestamp, Group, size of cells, size of nuclei. When correctly uploaded, the two columns with the size data have comma separated values of 10 measurements.

Data processing

The data that is in the Google Sheet can be downloaded and read into R as a CSV. All subsequent processing and data visualisation and presentation in dashboard style is done in R. The code is available on Github: https://github.com/JoachimGoedhart/CellSizeR. The cleaning of the data consist of removing empty cells, changing the column names, listing all individual measurements in a single row, forcing the data into a 'numeric' type and filtering for sensible values (anything outside the generous range of 0-1000 will be removed). A detailed protocol that explains the processing is available as protocol 10 (Goedhart, 2022).

Data visualisation

A dashboard is composed in R Markdown with the `{flexdashboard}` package. The code is available here: https://github.com/JoachimGoedhart/CellSizeR and the live dashboard is available online: https://amsterdamstudygroup.shinyapps.io/CellSizeR/

Results

Use case 1: Determination of the percentage of cells in S-phase

The aim of the experiment is is to determine the number of cells, as percentage, that are in the S-phase. To this end, students stain cells that are treated with EdU and they use these samples to quantify the percentage of cells in the S-phase in two ways (manual and semi-automated). The results are uploaded via a Google Form. The collected data can be analysed in multiple ways and here we used it to compare the two analysis methods and, secondly, to obtain an estimate for the percentage of S-phase cells. The data on the two analysis methods, manual and automated, is paired and can be visualised by a doplot in which the pairs of the data are connected (Figure 1). The slopes of the lines vary a lot, whereas the average values per year between the two methods is similar. This implies that there can be substantial differences between the two methods, with roughly a similar number of cases where the automated analysis over- or underestimates the percentage, relative to the manual analysis.

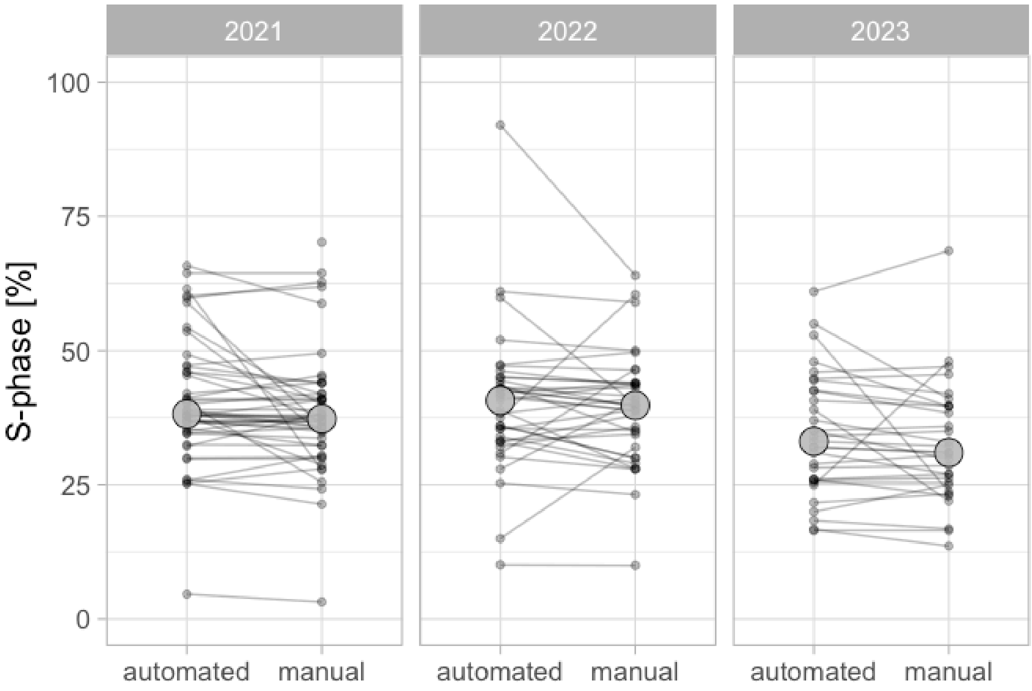

Figure 1: Comparing the manual and automated analysis. Source: The percentage of cells in the S-phase

There is increasing attention on effects of experimental design on data analysis and visualization. The recently proposed superplot to distinguish biological and technical replicates is an intuitive and straightforward way to communicate the design (Lord et al., 2020). The data on S-phase consists of both technical and biological replicates and is therefore ideally suited to explain the importance of correctly identifying the independent measurements. Here, we treat the data from each group as biological replicate, and the measurements within each group as a technical replicate. The reason is that a group of students all stain cells that are from the same passage number and treated at the same time and is therefore a technical replicate. On the other hand, different groups stain different passages of cells and so we treat these as independent observations. When the data is plotted for each individual technical replicate (Figure 2), it can be observed that we received multiple submissions per group, leading to a precise measurement per group. The median values range from 23% to 44%. The average value of the independent observations is 36.7% [N=12, 95%CI = 33.0%-40.3%].

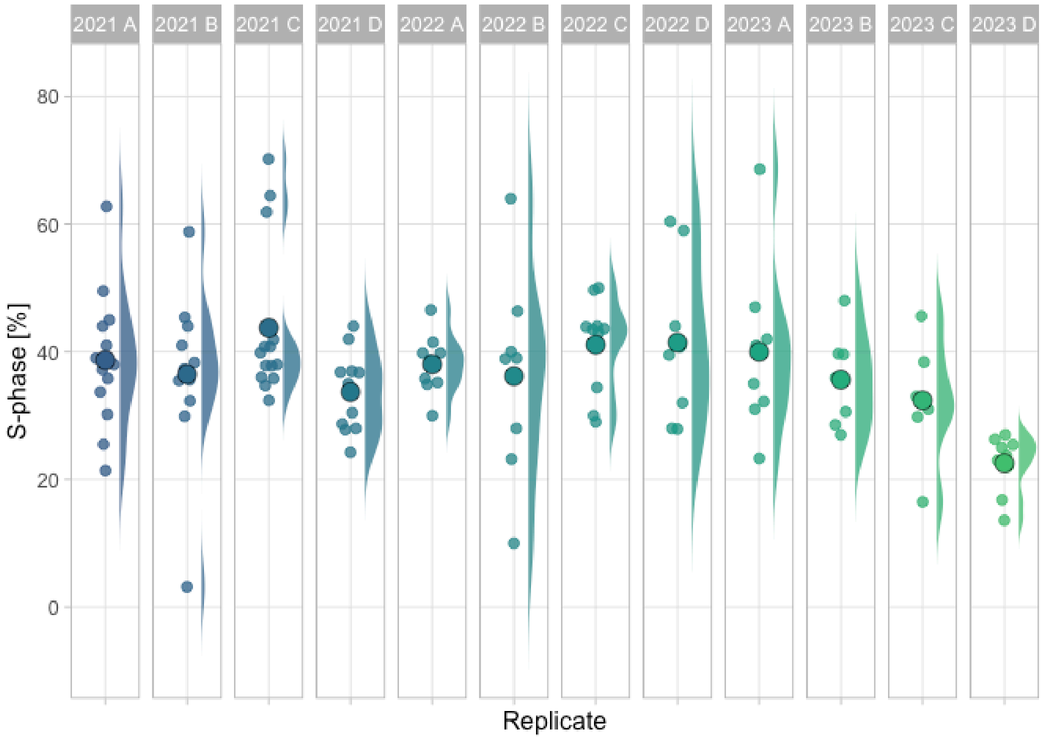

Figure 2: Data on the percentage of cells in e S-phase based on manual analysis. Each group & year defines an independent observations and is shown as dotplot and the distribution. The larger dot reflects the median value. Source: The percentage of cells in the S-phase

Use case 2: Comparing new results with historical data

The aim of the experiment is to determine the average size (diameter) of a human cheek cell and nucleus. To this end, the students acquire images of their own, stained cheek cells and measure the size of the cell and its nucleus. At least 10 measurements are made and the data are uploaded with a Google form. Each sample is an independent observation as it originates from a unique human specimen. The main learning objective for the students is to carry out the cell size measurements based on the images that are acquired. One important aspect to get to the right solution is the correct calibration of the field of view with a micro ruler. When the calibration is done incorrectly, this will affect the accuracy of the measurement, usually by an order of magnitude. To evaluate their results, the students can compare their own data with the historical data that is displayed on an online, interactive dashboard: `https://amsterdamstudygroup.shinyapps.io/CellSizeR/`. On the dashboard, users can select the data from all measurements, or from a single year. A histogram visualizes the distribution of individual data for both the cell and the nucleus and the bins of the histogram can be adjusted. Since the sizes vary susbtantially, the data can be shown on a log-scale as well (Figure 3). The dashboard shows the same graph as depicted in Figure 3, but it is interactive in the sense that by hovering over the plots, the exact values of the data can be inferred as well. The dashboard also shows the data for the 4 different groups and the size distribution of the cells by violin plots.

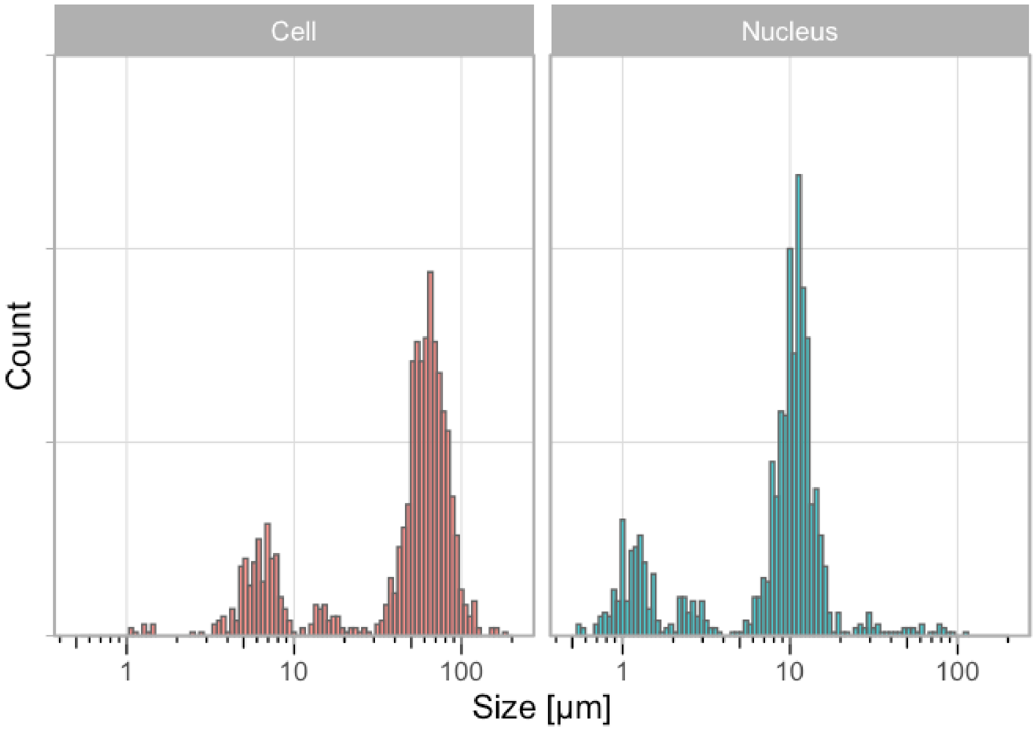

Figure 3: Distribution of the measured size of human cheek cells and their nucleus. Data from three years. Source: Summarizing the size of cells

The dashboard clearly shows a multimodal distribution of the results, which is caused by a set of incorrect measurements due to incorrect calibration. Still, if we assume that the majority of measurements is correct, it is possible for the students to make a fair comparison and discuss their results in the context of the historical data.

Conclusion

Data that is generated in courses is often recorded by individual students or groups of students in reports. However, it can be valuable and interesting to collect and use these data. Here, I present a flexible and open-source approach to collect and display data from a large group of students and over several years. A combination of Google Forms/Sheets for data collection and R for data processing and visualization is used. In the first use case, a R Markdown template is used to process and visualize the data. In the second use case, the data is displayed in dashboard style. The code is available and can be used as a starting point for the processing and visualization of other datasets.

Collecting and reusing the data has a number of advantageous aspects. First, a high number of measurements increases the precision of the measurement and therefore allows us to obtain precise numbers. Second, the historical data can be shared with the students and they can interpret and discuss their results in light of the existing data. Third, the obtained data serves as material that can be used to teach data manipulation, statistics and data visualization which is a fundamental aspect of science (Sailem et al., 2016). The use cases described in this paper deal with these aspects.

The aggregation of the data inevitably leads to a discussion on experimental design, as this is important to establish whether measurements are independent or not. This aspect of experimental design has received attention over the last years (Aarts et al., 2015; Eisner, 2021; Sikkel et al., 2017) and it is valuable to teach this aspect of data analysis and visualization. Although I have not implemented this yet, I think that having students participate in the data aggregation, creates a very practical opportunity to teach experimental design and the identification of biological units (Lazic et al., 2018).

In conclusion, I feel it is valuable to collect data from practical courses and here we report one way to achieve that. I hope that serves as a starting point for others that want to collect, store and use data from large groups of students.

Acknowledgments

A blog post by Garrick Aden-Buie was very helpful in the initial phase of this project. Many thanks to the people involved in Quarto, which was used to write and shape this paper. Most importantly, I'd like to thank all students involved in the course that have generously shared their data, making this a successful project.

References

Aarts, E., Dolan, C. V., Verhage, M., & Sluis, S. van der. (2015). Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. *BMC Neuroscience*, *16*(1). https://doi.org/10.1186/s12868-015-0228-5

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., et al. (2022). *Rmarkdown: Dynamic documents for r.* Retrieved from https://github.com/rstudio/rmarkdown

Carpenter, A. E. (2007). Software opens the door to quantitative imaging. *Nature Methods*, *4*(2), 120–121. https://doi.org/10.1038/nmeth0207-120

Eisner, D. A. (2021). Pseudoreplication in physiology: More means less. *Journal of General Physiology*, *153*(2). https://doi.org/10.1085/jgp.202012826

Goedhart, J. (2022). *DataViz protocols - an introduction to data visualization protocols for wet lab scientists.* Zenodo. https://doi.org/10.5281/ZENODO.7257808

Lazic, S. E., Clarke-Williams, C. J., & Munafò, M. R. (2018). What exactly is 'N' in cell culture and animal experiments? *PLOS Biology*, *16*(4), e2005282. https://doi.org/10.1371/journal.pbio.2005282

Lord, S. J., Velle, K. B., Mullins, R. D., & Fritz-Laylin, L. K. (2020). SuperPlots: Communicating reproducibility and variability in cell biology. *Journal of Cell Biology*, *219*(6). https://doi.org/10.1083/jcb.202001064

R Core Team. (2022). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Renz, M. (2013). Fluorescence microscopy-A historical and technical perspective. *Cytometry Part A*, *83*(9), 767–779. https://doi.org/10.1002/cyto.a.22295

Sailem, H. Z., Cooper, S., & Bakal, C. (2016). Visualizing quantitative microscopy data: History and challenges. *Critical Reviews in Biochemistry and Molecular Biology*, *51*(2), 96–101. https://doi.org/10.3109/10409238.2016.1146222

Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675. https://doi.org/10.1038/nmeth.2089

Senft, R. A., Diaz-Rohrer, B., Colarusso, P., Swift, L., Jamali, N., Jambor, H., et al. (2023). A biologist's guide to planning and performing quantitative bioimaging experiments. *PLOS Biology*, *21*(6), e3002167. https://doi.org/10.1371/journal.pbio.3002167

Sikkel, M. B., Francis, D. P., Howard, J., Gordon, F., Rowlands, C., Peters, N. S., et al. (2017). Hierarchical statistical techniques are necessary to draw reliable conclusions from analysis of isolated cardiomyocyte studies. *Cardiovascular Research*, *113*(14), 1743–1752. https://doi.org/10.1093/cvr/cvx151

Wait, E. C., Reiche, M. A., & Chew, T.-L. (2020). Hypothesis-driven quantitative fluorescence microscopy – the importance of reverse-thinking in experimental design. *Journal of Cell Science*, *133*(21). https://doi.org/10.1242/jcs.250027

Waters, J. C. (2009). Accuracy and precision in quantitative fluorescence microscopy. *Journal of Cell Biology*, *185*(7), 1135–1148. https://doi.org/10.1083/jcb.200903097