Avoiding common problems with statistical analysis of biological experiments using a simple nested data simulator

by Veronika V. Alexandrova^{1*}, Mikhail N. Anisimov^{1,2*}, Igor A. Eltsov^{3*}, Anastasia P. Kilina^{2*}, Iuliia N. Lopanskaia^{1*}, Liubov O. Makarova^{3*}, Maxim A. Vovchenko^{3*} and Nikita B. Gudimchuk^{1,2,4#}

1. Center for Theoretical Problems of Physicochemical Pharmacology, Russian Academy of Sciences, Moscow, Russia;
2. Physics Department, Lomonosov Moscow State University, Moscow, Russia;
3. Moscow Institute of Physics and Technology, Dolgoprudny, Russia;
4. Dmitry Rogachev National Research Center of Pediatric Hematology, Oncology and Immunology, Moscow, Russia.

Correspondence: gudimchuk@phys.msu.ru

* these authors contributed equally to this work

Received: 11.03.2021

Accepted: 29.03.2021

Published: 31.03.2021

DOI: 10.52455/sbpr.01.202101013

Abstract

Despite an extensive literature on statistical methods and their proper application to biological data, incorrect analyses remain a critical and widely spread problem in research papers. Inherently hierarchical (nested, clustered) structure of biological measurements is often erroneously neglected, leading to pseudo-replication and false positive results. This, in turn, complicates the correct assessment of statistical power and impairs optimal planning of experiments. In order to attract more attention to this problem and to illustrate the importance of direct account for the nested structure of biological data, in this article we present a simple open-source simulator of two-level normally distributed stochastic data. By defining 'true' mean values and 'true' intra- and inter-cluster variances of the simulated data, users of the simulator can test various scenarios, appreciate the importance of using correct multi-level analysis and the danger of neglecting the information about the data structure. Here we apply our nested data simulator to highlight some commonly arising mistakes with data analysis and propose a workflow, in which our simulator could be employed to correctly compare two nested groups of experimental data and to optimally plan new experiments in order to increase statistical power when necessary.

Keywords: nested data, statistical analysis, p-value, false positive, false negative, statistical power, simulated data, intra-cluster correlation

Introduction

Biological experiments often produce 'clustered' or 'nested' data. By these terms, we mean a hierarchical grouping of individual measurements via a certain principle, so that the data points in each cluster are not completely independent from each other (Fig. 1). Imagine a scenario when some measurement is repeated on several experimental days. Despite the experimentalist's effort to reproduce exactly the same conditions every time, in practice, the actual conditions may still slightly vary from day to day. Therefore, data points collected on a particular day may correlate with each other more than with points from other days. Such grouping of data by day is one very common example of hierarchical data structure in biological experiments. Another common example is a scenario when multiple data points are collected from a smaller number of animals, patients, cells, etc. In this case, measurements from the same animal/patient/cell may not be completely independent, so they form clusters with slightly shifted mean values.

Unfortunately, it is not uncommon for researchers to ignore the nested structure of their experimental data. In this case, a researcher naively pools all data points collected under a certain test condition assuming (often erroneously!) that they are independent. The pooled data are then usually processed altogether, using some standard statistical criteria and methods such as t-test, regression analysis, ANOVA etc., which are precisely applicable only for independent values. Considering statistically non-independent data as statistically independent creates a problem of pseudo-replication [1]. This artificially increases the sample size and leads to irreproducible or meaningless results. Due to pseudo-replication, statistical analysis

Figure 1. Schematic of a typical biological experiment design, generating nested data

tests an incorrect hypothesis, rather than the one, which the researcher really intends to test [2]. This problem has been described in several excellent reviews [2, 3, 4]. A hallmark of pseudo-replication is an overly small p-value reported by the test, even in spite of substantial overlaps of data points from the compared groups [5].

Besides the problem of false positives, incorrect statistical treatment of data and ignoring their nested structure may also increase the probability of false negative results. For example, one might naively expect that just adding more data from each cell/animal/patient/day should boost the statistical power of the analysis. This is not necessarily the case with clustered data. Therefore, careful analysis of data clustering is very important for both optimal planning of experiments and for solid statistical evaluation of biological measurements [1, 3].

Correct treatment of nested biological data requires multi-level statistical analysis [6]. Various software packages have been developed to implement it, including a commercial tool Prism GraphPad [7] and free programs, such as InVivoStat [8], developed specifically to plan and analyze experiments with animals. In the multi-level analysis, data are compared at the level of cluster averages while maintaining the difference between the variance within clusters and the variance between clusters. This approach explains parts of the variation of the experimental effect and provides the greatest statistical power. The multi-level analysis not only ensures correct statistical interpretation of the results and thus correct conclusions but can also provide unique information on the collected research data that cannot be obtained when standard statistical methods are used on either individual observations or summary statistics [5, 6].

Here we present a simple open-source data simulator to illustrate common problems arising in the statistical analysis of clustered biological data. We show how false assumptions about data independence can lead to incorrect assessment of the statistical significance of the difference between

the compared groups and how this result depends on the extent of intra-cluster correlation (ICC) of the data. Using our simulator, we also demonstrate how the statistical power of analysis changes depending on the number of clusters and the number of elements in them, and propose an algorithm for processing multi-level data and planning optimal experimental measurements.

Methods

Generation of stochastic nested data

We developed a program, which generates two normally distributed stochastic variables: 'control' (x) and 'experiment' (y), each containing N data clusters with n observations per cluster. Random variables, representing 'control' and 'experimental' values are generated as follows:

$$\begin{aligned} x(i, j) &= x_{mean}^{intra}(i) + \sigma^{intra} R(0, 1), \\ y(i, j) &= y_{mean}^{intra}(i) + \sigma^{intra} R(0, 1), \end{aligned} \quad \text{Eq. 1}$$

where i is the cluster number from 1 to N ; j is the number of the observation within the cluster from 1 to n ; $x_{mean}^{intra}(i)$ is the mean value of 'control' observations in the i -th cluster; $y_{mean}^{intra}(i)$ is the mean value of 'experimental' observations in the i -th cluster; σ_{intra}^2 is the intra-cluster variance; $R(0, 1)$ is a random normally distributed number with a mean of 0 and a standard deviation of 1.

Intra-cluster means, $x_{mean}^{intra}(i)$ and $y_{mean}^{intra}(i)$, are defined as follows:

$$\begin{aligned} x_{mean}^{intra}(i) &= x^{true} + \sigma^{inter} R(0, 1), \\ y_{mean}^{intra}(i) &= y^{true} + \sigma^{inter} R(0, 1), \end{aligned} \quad \text{Eq. 2}$$

where x^{true} and y^{true} are 'true' mean values of the 'control' and 'experiment'; σ_{inter}^2 is the inter-cluster variance; $R(0, 1)$ is a random normally distributed number with a mean of 0 and a standard deviation of 1.

Statistical comparison of 'Control' and 'Experimental' groups

The simulator script processes the stochastic datasets, using three methods to test the null hypothesis about the quality of the means in 'control' and 'experiment'.

Method 1: n observations from all N clusters are combined together to form two aggregated pools of values: 'control' and 'experiment'. The aggregated pools are then compared using a standard unpaired two-tailed t-test, based on the score t_1 :

$$t_1 = \frac{|\langle x \rangle - \langle y \rangle|}{s \sqrt{\frac{2}{M}}}, \quad \text{Eq. 3}$$

where triangular brackets denote averaging, $M = N \cdot n$ is the total dataset size and s is the pooled standard deviation defined as:

$$s = \sqrt{\frac{(M-1) \cdot SD_x^2 + (M-1) \cdot SD_y^2}{2M-2}}, \quad \text{Eq. 4}$$

where SD_x and SD_y are the standard deviations of the observations in the pooled 'control' and 'experimental' groups, respectively.

Next, the p-value is computed from t-distribution with $h = 2M - 2$ degrees of freedom at 0.05 significance level.

Method 2: observations in each cluster are averaged. Mean values from each cluster are used as inputs for a standard unpaired t-test, based on the score t_2 :

$$t_2 = \frac{|\langle x \rangle - \langle y \rangle|}{s_{mean} \sqrt{\frac{2}{N}}}, \quad \text{Eq. 5}$$

here triangular brackets denote averaging, N is the number of clusters and s_{mean} is the combined standard deviation of the cluster means:

$$s_{mean} = \sqrt{\frac{(N-1) \cdot SD_{xmean}^2 + (N-1) \cdot SD_{ymean}^2}{2N-2}}, \quad \text{Eq. 6}$$

where SD_{xmean} and SD_{ymean} are the standard deviations of the intra-cluster means in 'control' and 'experimental' groups, respectively.

The p-value is found as in the first method, except that the degrees of freedom are now reduced to $h = 2N - 2$.

Method 3: the nested structure of the datasets is taken into account by computing an adjusted unpaired two-tailed t-test statistics [10]:

$$t_3 = c \cdot t_1, \quad \text{Eq. 7}$$

where c is the correction factor for t-distribution:

$$c = \sqrt{\frac{(M-2) - 2(n-1) \cdot ICC}{(M-2)(1 + (n-1) \cdot ICC)}}, \quad \text{Eq. 8}$$

where M is the total number of data points, n is the number of observations per cluster, ICC is the intra-cluster correlation, computed as:

$$ICC = \frac{\sigma_{inter}^2}{\sigma_{inter}^2 + \sigma_{intra}^2}, \quad \text{Eq. 9}$$

The adjusted p-value is now calculated from t_3 -distribution from Eq.6 with h degrees of freedom computed as follows:

$$h = \frac{((M-2) - 2(n-1) \cdot ICC)^2}{(M-2)(1 - ICC)^2 + n(M-2n) \cdot ICC^2 + 2(M-2n) \cdot ICC \cdot (1 - ICC)}. \quad \text{Eq. 10}$$

Statistical power analysis

Provided that the true 'control' and 'experiment' values are different, the script also calculates statistical power, achieved by each method at a given set of simulation parameters. For this purpose, the stochastic simulation is repeated 1000 times. At the significance level of 0.05, the null-hypothesis about equality of the means is rejected or accepted based on computed p-values in each of the 1000 runs. The power is determined as the percentage of cases when false negative outcome is not obtained.

Code availability

The nested data simulator is available from https://github.com/juliaLopanskaia/nested_data_simulator.

Results

Do not pool your data if they are intra-cluster-correlated

We would like to start by using our random nested data simulator to illustrate the importance of accounting for data clustering. Suppose it is necessary to conduct a study of some quantity for the equality of its means under two different conditions ('control' and 'experiment'; here and everywhere below the null hypothesis - the means are equal). For this purpose, we will generate 50 simulated 'measurements' for each of the three clusters. In the first example scenario, we will generate the data in such a way that the true means in each condition are equal so that there is no difference between 'control' and 'experiment'. As described in Methods, our nested data simulator populates each cluster with normally distributed random data, which are generated to have a user-defined mean and intra-cluster variance, σ_{intra}^2 . The user also defines the inter-cluster variance, σ_{inter}^2 , describing the extent of scatter of the mean values among the clusters.

Depending on the variances between and within

clusters, the observations can be more or less independent from each other. Intra-cluster correlation (ICC) serves as a convenient metric for the degree of relative similarity of observations from the same cluster (Eq. 9).

Let us consider two simulated cases: (1) the simulated data have a weak relative correlation within clusters (ICC = 0.01, Fig 2A) and (2) the simulated data have a strong relative correlation within clusters

(ICC = 0.2, Fig 2B).

In order to illustrate some common pitfalls with statistical analysis, we can evaluate three ways of processing these data: 1) by naively combining observations from all clusters per condition and using an unpaired t-test to compare the two conditions; 2) by calculating means for each cluster and processing them with an unpaired t-test; 3) by using an adjusted unpaired t-test, which takes into account the multi-

Figure 2. Analysis of an example of simulated nested data with equal means in 'control' and 'experimental' groups. The number of observations in a cluster 50, the number of clusters 3. In all cases, the true means in 'control' and 'experiment' are equal (1.0 a.u.). (A) Weak intra-cluster correlation (ICC = 0.01). Intra-cluster variance: 0.3 a.u., inter-cluster variance: 0.03 a.u. Different colors mark different clusters. Numbers above the plots represent three p-values, computed as described in Methods. Central lines and error bars show mean and s.e.m. (B) Strong intra-cluster correlation (ICC = 0.2). Intra-cluster variance: 0.3 a.u., inter-cluster variance: 0.15 a.u. Ignoring data clustering in this case leads to a false positive result. Different colors mark different clusters. Numbers above the plots represent three p-values, computed as described in Methods. Central lines and error bars show mean and s.e.m. (C) Dependence of the probability of obtaining a false positive result on the ICC when pooling all data (red points), using the means for each cluster (blue points) and applying adjusted t-test (green points). It can be seen that 'incorrect' processing (pooling data from all clusters) significantly increases the likelihood of a false positive result if the clustering is strong. At the same time, the use of means for each cluster or the adjusted t-test significantly reduces the likelihood of obtaining a false positive result.

Figure 3. Analysis of statistical power of experiments with different intra-cluster correlation, numbers of observations and clusters. In all cases, the true means in 'control' and 'experiment' are different and equal to 1.0 and 0.8 a.u., correspondingly. Different colors mark different clusters. Numbers above the plots represent p-values and powers, computed as described in Methods. Central lines and error bars show mean and s.e.m. (A) Weak intra-cluster correlation (ICC = 0.01). (B) Strong intra-cluster correlation (ICC = 0.2). The number of measurements per each cluster 5, the number of clusters 3. (C-F) Statistics increased by adding measurements within each cluster (C, D), or clusters (E-F). False negative cases are shown in B and D, associated with low power. In all cases (C-F) the displayed power is slightly less when using processing the means for each cluster and an unmodified t-test compared to using the adjusted t-test. Data processing (A-F) was carried out using mean values for each cluster and a t-test based on per-cluster means or using an adjusted t-test.

level structure of the data [10].

In case if the observations are almost independent ($ICC = 0.01$, Fig 2A), all three ways to statistically process the data return similarly high p-values. This indicates that the likelihood that the null-hypothesis about equality of the means is correct is very high. This is the right answer, as we know that the 'true' means have indeed been user-defined to be equal.

In the case of a strong intra-cluster correlation ($ICC = 0.2$, Fig 2B), however, the three statistical procedures lead to different outcomes. Specifically, the most naive approach of pooling all data points together and conducting a standard unpaired t-test produces an extremely tiny p-value. This would be erroneously interpreted as an indication that 'control' and 'experiment' are significantly different. Obviously, this is not the case in our simulated scenario, so this type of analysis exemplifies how disregarding the nested data structure can lead to a false positive result. The second statistical procedure, that is averaging all the observations within each cluster and then using the intra-cluster means to conduct a t-test, performs better and does not lead to incorrect rejection of the null hypothesis. Finally, the last of the considered statistical procedures also supports the null-hypothesis of equal means.

To further illustrate the importance of taking data clustering into account, we plotted the dependence of the probability of obtaining a false positive result on the extent of the intra-cluster correlation ICC (Fig. 2C). This graph shows that the naive approach of pooling all data together dramatically increases the probability of false positives, while two other statistical treatments produce more accurate results in this scenario. The heatmap in Fig. S1 shows how the probability of obtaining a false positive result depends on the number of clusters and the number of observations per cluster.

[How to properly plan your experiment and adjust data processing for detecting a true difference between compared groups?](#)

We hope it is clear from our simulated example, described in the previous section, that combining measurements from all clusters into one pool and treating them as independent observations likely leads to false positive results unless the intra-cluster correlation is lower than 0.01. Therefore, it is highly recommended to recognize the nested structure of the data and average data within clusters to conduct statistical analysis on their means or use an adjusted unpaired t-test.

But if the latter two methods produce comparable rates of false positive results, is there really any preference to use one or the other? To answer this question, let us now consider another simulated scenario, in which both the 'control' and 'experiment' are represented by three clusters of data; but the

'true' means in 'control' and 'experiment' are slightly different (Fig. 3A, B). Which type of statistical analysis will provide a lower probability of a false negative result?

We will start with an example, in which the amount of data is limited: only $n = 5$ observations per cluster, and only $N = 3$ clusters per condition. By running our nested data simulator repeatedly with fixed user-defined parameters, it is easy to calculate the probability of obtaining a false negative result, β . The corresponding statistical power is defined as $(1 - \beta) \cdot 100\%$. From the simulations, it becomes clear that the power is consistently higher when the same data are evaluated using adjusted t-test, compared to the t-test based on per-cluster means (Fig. 3A, B). This small, but reproducible gain in statistical power allows detecting small differences between the 'control' and 'experiment' more reliably with the former method. Note that the statistical power is lower with both methods in case the data have higher ICC.

The probability of obtaining a false negative result depends on both the number of measurements in the cluster and the number of clusters (Fig. S2). Thus, to increase the sensitivity of data processing, the number of measurements in each cluster and/or the number of clusters should be increased. In practice, however, it is usually difficult to increase the number of clusters (i.e. number of organisms/patients/cells/experimental days). Therefore, for reducing the probability of obtaining a false negative result, it is easier (and sometimes more expedient) to increase the number of measurements within the cluster. Fig. 3C, D shows the examples, in which the number of observations per cluster is increased 10-fold, compared to Fig. 3A, B. As one can see, in this case, the statistical power is indeed increased, but the gain is higher when the ICC is low (Fig. 3C), while with higher ICC increasing statistics within each cluster does not have such a dramatic effect (Fig. 3D). Adding more clusters usually is a more effective way to enhance statistical power, as demonstrated by the simulated examples in Fig. 3E, F, in which the number of clusters is increased 5-fold relative to the case considered in Fig. 3A, B. In practice, researchers should make a decision whether to add more observations per cluster or more clusters based on the understanding of the associated expected gain in power (Fig. S2), and other factors, such as time, cost and effort investments, required to collect each type of data.

In any scenario, complete description of the outcomes of statistical analysis is essential for correct interpretation of the study. Some of the best practices of communicating the results of statistical comparisons include data visualization in the form of SuperPlots [5], and reporting a full set of details about the statistical analysis, such as naming the null hypothesis, which was tested; reporting which

Figure 4. Proposed workflow for optimal planning of experiment and data processing.

statistical method was used; what was the sample size; which type of p-value (one-sided vs. two-sided) was computed. Besides the statistical significance, the substantive significance of the comparison (the effect size) should also be reported [11].

How to process nested data properly?

In this final section, we propose a tentative experimental workflow incorporating our nested data simulator (Fig. 4). It is applicable when the experimenter aims to compare mean values of some quantity under two experimental conditions and the associated measurements generate nested data, such as observations from different cells/organisms/patients/experimental days. We suggest that the study should be carried out as follows. First, a pilot set of measurements should be made to produce two minimal datasets ($N = 3$ clusters, $n = 10$ -20 observations per cluster). With these data in hand, one can estimate the mean values, the variances in each cluster and the variances of the means between clusters.

Note that for simplicity we only consider the case, when the data are normally distributed and the variances in control and experimental groups are not significantly different. This is often the case with real data, justifying the application of the t-test. If the data do not follow a normal distribution, nonparametric tests, such as Mann-Whitney-Wilcoxon U -test, can be applied [12]. Taking the means and variances from the pilot experiment as input parameters for simulations, our nested data simulator can be used to estimate the statistical power of the pilot experiment. A reasonable and conventionally used tradeoff value for statistical power is 80% [13].

If the estimated power is higher than 80%, the pilot experiment is considered sufficient to detect the difference between the 'control' and 'experiment', so no more experiments are needed. The user can just employ the p-values from adjusted t-test, reported by the simulator, to reject or accept the null-hypothesis about the equality of the means. If the p-value is lower than some threshold, α , the 'experiment' is statistically different from the 'control' with the level of significance $1 - \alpha$.

If the estimated power of the pilot experiment is lower than 80%, we suggest using the simulator to predict, how many new clusters (N) (i.e. animals/cells/patients/experimental days) or how many observations per cluster (n) one should add to increase the power to the required level. This can be done by simply running the nested data simulator with the means and variances estimated from the pilot experiment, but increasing the number of clusters and observations within clusters until one gets a satisfactory outcome. After a good combination is found, the additional measurements should be carried out and the data added to the datasets. The new means and variances should then be estimated in 'control' and 'experimental' datasets and the cycle of the workflow repeated (Fig. 4). This workflow is expected to maximize efficiency, high statistical power and ensure a proper estimation of the significance of the difference of the means in 'control' and 'experimental' conditions.

Conclusion

One of the strengths of a simple simulator like ours is the ability to clearly illustrate how incorrect implicit or explicit assumptions about data independence can

lead to misuse of statistical procedures, producing false positive or false negative results. These common pitfalls become especially evident because the 'true' answers are known in the simulated data. In our experience, this kind of exercise with the nested data simulator helps training more intuitive thinking about statistical methods. But besides just illustration of widely-spread mistakes, the simulator can be directly incorporated into experimental workflow to help researchers plan and correctly analyze their experiments. We hope that by using our simple nested data simulator here, we have managed to attract attention to the old and important problem of statistical analysis of hierarchical biological data and convince the readers about the importance of accounting for data structure.

Acknowledgements

This work was supported by the RF President's grant # MK-1869.2020.4.

Author contributions

V.V.A., M.N.A., I.A.E., A.P.K., I.N.L., L.O.M., M.A.V. performed research and wrote the initial draft of the manuscript, N.B.G., designed research, acquired funding and edited the manuscript.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. S. H. Hurlbert, 'Pseudoreplication and the Design of Ecological Field Experiments', *Ecological Monographs*, vol. 54, is. 2, pp. 187–211, Jun. 1984, doi:10.2307/1942661
2. S. E. Lazic, 'The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis?', *BMC Neurosci.*, vol. 11, p. 5, Jan. 2010, doi: 10.1186/1471-2202-11-5.
3. S. E. Lazic, C. J. Clarke-Williams, and M. R. Munafò, 'What exactly is "N" in cell culture and animal experiments?', *PLoS Biol.*, vol. 16, no. 4, Apr. 2018, doi: 10.1371/journal.pbio.2005282.
4. J. P. Dorman, 'The effect of clustering on statistical tests: an illustration using classroom environment data', *Educational Psychology*, vol. 28, no. 5, pp. 583–595, Aug. 2008, doi: 10.1080/01443410801954201.
5. Lord SJ, Velle KB, Mullins RD, Fritz-Laylin LK. 'SuperPlots: Communicating reproducibility and variability in cell biology.', *J Cell Biol.* 2020 Jun 1, 219(6):e202001064, doi: 10.1083/jcb.202001064.
6. E. Aarts, C. V. Dolan, M. Verhage, and S. van der Sluis, 'Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives', *BMC Neurosci.*, vol. 16, Dec. 2015, doi: 10.1186/s12868-015-0228-5.
7. <https://www.graphpad.com/scientific-software/prism/>
8. S. T. Bate, R. A. Clark, and S. C. Stanford, 'Using InVivoStat to perform the statistical analysis of experiments', *J. Psychopharmacol. Oxf. Engl.*, vol. 31, no. 6, pp. 644–652, Jun. 2017, doi: 10.1177/0269881116682213.
9. J. K. Haseman and M. D. Hogan, 'Selection of the experimental unit in teratology studies', *Teratology*, vol. 12, no. 2, pp. 165–171, Oct. 1975, doi: 10.1002/tera.1420120209.
10. L. V. Hedges, 'Correcting a Significance Test for Clustering', *J. Educ. Behav. Stat.*, vol. 32, no. 2, pp. 151–179, Jun. 2007, doi: 10.3102/1076998606298040.
11. G. M. Sullivan and R. Feinn, 'Using Effect Size—or Why the P Value Is Not Enough', *J Grad Med Educ*, vol. 4, no. 3, pp. 279–282, Sep. 2012, doi: 10.4300/JGME-D-12-00156.1
12. B. Rosner and D. Grove, 'Use of the Mann–Whitney U-test for clustered data', *Statistics in Medicine*, vol. 18, no. 11, pp. 1387–1400, 1999, doi: [https://doi.org/10.1002/\(SICI\)1097-0258\(19990615\)18:11<1387::AID-SIM126>3.0.CO;2-V](https://doi.org/10.1002/(SICI)1097-0258(19990615)18:11<1387::AID-SIM126>3.0.CO;2-V).
13. J. Cohen and B. B. Wolman, 'Handbook of clinical psychology' McGraw-Hill New York, pp. 95–121, 1965.