РАСПРОСТРАНЕННЫЕ ОШИБКИ ПРИ ЗАПОЛНЕНИИ ЭЛЕКТРОННЫХ ТАБЛИЦ И КАК ИХ ИЗБЕЖАТЬ

DOI: 10.37586/2686-8636-1-2021-105-109

УДК: 001.891.5:51-76

Лысенков С.Н.[1, 2]

[1] ФГАОУ ВО РНИМУ им. Н.И. Пирогова Минздрава России, Российский геронтологический научно-клинический центр, Москва, Россия
[2] МГУ имени М.В. Ломоносова, биологический факультет, Москва, Россия

Резюме

Ошибки при проведении исследований могут возникать не только на этапе планирования, сбора данных и их статистической обработки, но и на этапе заполнения электронных таблиц. В большинстве случаев они могут быть легко устранены, в противном случае возможно искажение результатов. Эти ошибки можно разделить на неверный формат и ошибки ввода. В первом случае особенно опасным является смешение отсутствия данных и нулевого значения признака, что может приводить к систематическим ошибкам: завышению (если нулевые значения заполнены как отсутствие данных) или занижению (если отсутствие данных заполнено как нулевые значения) среднего значения или распространенности признака. Ошибки ввода чаще всего случайны, и их эффект снижается при увеличении выборки, но для корректного анализа они также должны быть максимально устранены. В статье приводится алгоритм действий, позволяющий найти ошибки ввода и исправить их до начала статистической обработки.

Ключевые слова: медицинская статистика; методические указания; электронные таблицы; ошибки в исследованиях

Для цитирования: Лысенков С.Н. Распространенные ошибки при заполнении электронных таблиц и как их избежать. *Российский журнал гериатрической медицины*. 2021; 1(5): 105–109. DOI: 10.37586/2686-8636-1-2021-105-109

COMMON ERRORS IN FILLING IN SPREADSHEETS AND HOW TO AVOID THEM

Lysenkov S.N.[1,2]

[1] Pirogov Russian National Research Medical University, Russian Gerontology Research and Clinical Centre, Moscow, Russia
[2] M.V. Lomonosov Moscow State University, Biological Faculty, Moscow, Russia

Abstract

Errors during research can occur not only at the stage of planning, data collection and statistical analysis, but also at the stage of filling in spreadsheets. In most cases they can be easily corrected, otherwise results may be distorted. These errors can be categorized into invalid format and input errors. In the first case, the mixing of the absence of data and the zero values is especially dangerous, since it can lead to systematic errors: overestimation (if the zero values are filled in as the absence of data) or underestimation (if the absence of data is filled in as zero values) of the mean or the prevalence. Input errors are most often random, and their effect decreases with increasing sample size, but for the better analysis they should also be corrected as much as possible. The article provides an algorithm that allows you to find input errors and correct them before statistical analysis.
Key words: medical statistics; practical guides; spreadsheets; research errors

Keywords: medical statistics; practical guides; spreadsheets; research errors

For citation: Lysenkov S.N. Common errors in filling in spreadsheets and how to avoid them. *Russian Journal of Geriatric Medicine*. 2021; 1(5): 105–109. DOI: 10.37586/2686-8636-1-2021-105-109

Статистическая обработка данных давно стала неотъемлемой частью медицинских исследований. Существуют публикации, которые описывают распространенные ошибки в медико-биологической статистике, но в основном это ошибки обработки и представления данных [1–3]. Еще одна проблема — это собственно качество исходных данных [4–6]. Но даже если данные собраны добросовестно, электронные таблицы довольно часто заполнены некорректно, что затрудняет их обработку и иногда может также приводить к ошибочным результатам.

Ошибки при заполнении баз данных можно условно разделить на две группы: неверный формат данных и ошибки ввода.

В первом случае данные введены так, что их интерпретация затруднена. В большинстве случаев неверный формат относительно легко меняется на верный, но на это тратится время, которое может уходить в том числе и на сверку с исходными данными, не всегда доступными для статистика.

Самая частая (и при этом самая серьезная) ошибка этого рода — это обозначение отсутствия данных нулем. Проблема тут состоит в том, что статистические программы воспринимают ноль как значение и оперируют им, как числом. Отсутствие данных может возникать либо из-за того, что признак неприменим к пациенту (возраст менопаузы у мужчины, возраст выхода на пенсию у работающего, тест рисования часов у слепого и т.п.), либо из-за того, что по той или иной причине он не был измерен (отказ пациента, нехватка времени и т.п.).

Особо опасно это смешение в том случае, когда нулевые значения возможны. Нулевой рост или артериальное давление — очевидные ошибки, и их легко выявить. Но нулевые значения могут быть осмысленными в некоторых лабораторных показателях или шкалах. В случае, когда нулем и единицей зашифрованы отсутствие или наличие бинарного признака (например, определенного заболевания), большое число нулей на месте пропущенных значений приведет к существенной недооценке распространенности признака.

Например, из десяти человек положительно на вопрос о курении ответили четыре человека (их в таблице обозначили единицей), еще четыре ответили отрицательно (обозначили нулём), а два человека по той или иной причине не отвечали на вопрос, но в таблице эти ячейки также заполнили нулями. Тогда истинная распространенность курения в выборке $\rho = 4/(4 + 4) = 0.5$, а доля пропущенных значений $m = 2/10 = 0.2$. Расчет же распространенности курения по таблице, где пропущенные значения обозначены нулём, даст $\rho^* = 4/10 = 0.4$. В общем случае, если истинную распространенность признака в выборке обозначить ρ (в приведенном примере это 0.5), а долю, которую составляют пропущенные значения, обозначенные нулем, от всего массива данных, как m (в приведенном примере 0.2), то заниженная из-за неправильного формата данных оценка распространенности составит $\rho^* = \rho(1 - m)$.

На рисунке 1 видно, что разница оказывается более существенной при высоких значениях истинной распространенности и высокой доле пропущенных значений. Второй вариант, конечно, менее распространен, но также должен быть принят во внимание.

Обратный пример — когда пустой ячейкой обозначают отсутствие признака. Так как в больших базах неизбежны пропущенные значения, то при интерпретации всех пустых ячеек как нулей распространенность признака будет переоценена. Вернемся к примеру с курением. Пусть пять человек из десяти ответили положительно на вопрос о курении, а пять — отрицательно, но при заполнении таблицы у двух из них соответствующие ячейки оставили пустыми. Тогда истинная распространенность курения будет $\rho = 5/10 = 0.5$. Но расчет по таблице даст значение $\rho^* = 5/8 = 0.625$. В общем случае, если обозначить истинную распространенность признака как ρ, а долю исти ных нулей, обозначенных пропуском данных, как k (в приведенном примере $k = 2/5 = 0.4$), то определенная по таким данным распространенность признака будет $\rho/(1 + \rho k - k)$. Рисунок 2 показывает, что в этом случае наиболее серьезной проблема становится при не очень высокой истинной распространенности и высокой доле нулей, записанных как отсутствие данных.

Важно, что обе рассмотренные ошибки относятся к систематическим, то есть всегда смещают результат в сторону завышения или занижения,

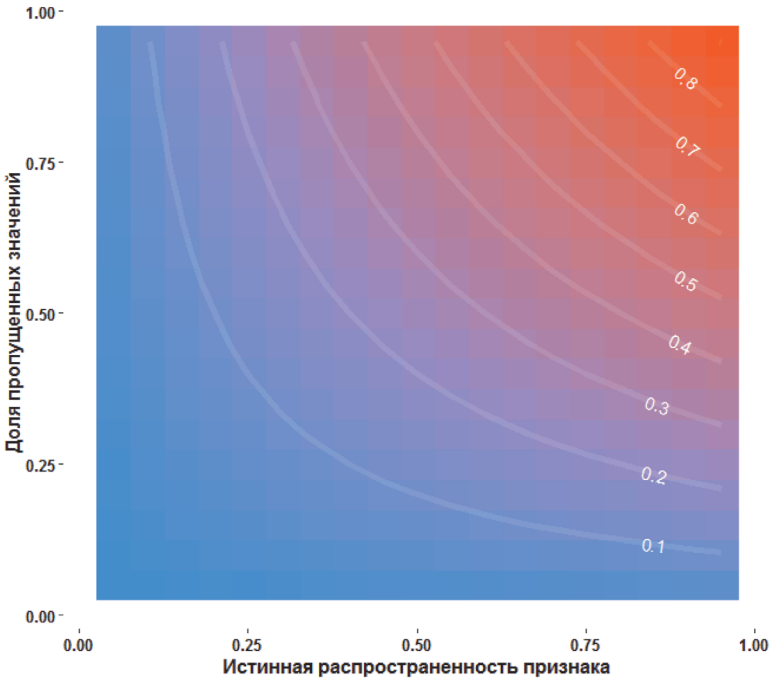

Рис. 1. Разность истинной распространенности бинарного признака и посчитанной по таблице, в которой пропущенные значения обозначены нулями, как и реальное отсутствие признака. Указаны изолинии разности.

и этот эффект не уменьшается при увеличении выборки.

Другая возможная ошибка, связанная с неправильным форматом, — это использование разных единиц измерения в одной переменной. Так может случиться, если в таблицу сведены показатели, полученные в разных лабораториях. При этом важно отметить, что в правильно составленной таблице в самих ячейках с данными не должно быть единиц измерения, так как в этом случае ни одна из программ не воспримет его как число.

Не распознаются как число и те ячейки, в которые внесены какие-то дополнительные комментарии, помимо значения, или более сложные значения, например, диапазоны значений (часты в лабораторных анализах числа клеток) или же артериальное давление, заполненное через косую черту. В последнем случае следует делать отдельные столбцы для систолического и диастолического давления: это позволит быстрее посчитать и производные показатели (пульсовое давление, долю людей с превышением нормы и т.д.). Диапазоны значений для анализа надо заменять тем или иным значением (нижней или верхней границей, или серединой диапазона) — каким именно, зависит от задач исследования.

Второй вариант некорректного заполнения таблиц — ошибки ввода — неизбежен при работе с большими объемами данных. Часть из них так и останутся невыявленными, если не выходят за область допустимых значений (например, систолическое артериальное давление записано как 150 вместо 140). Однако в этом случае такие ошибки следует трактовать как случайные, а не систематические, поэтому при достаточно большой выборке завышения и занижения значений будут взаимно уничтожаться.

В других случаях ошибки ввода могут быть выявлены. Но их исправление часто требует обращения к исходным данным, которые могут быть недоступны статистику, поэтому, на мой взгляд, первичную проверку на них имеет смысл проводить тем, кто заполняет базы данных. Ниже приведены простые приемы, позволяющие найти ошибки ввода.

После того, как все данные внесены в таблицу, в каждом столбце, содержащем числовые данные, надо посчитать минимум и максимум, общее число числовых данных и общее число значений. Для этого используются функции (в скобках приведены их названия, используемые в англоязычных вариантах электронных таблиц): МИН (MIN), МАКС (MAX), СЧЁТ (COUNT) и СЧЁТЗ (COUNTA).

Минимум и максимум позволяют найти значения, выбивающиеся за пределы допустимых. Например, в столбце, который должен содержать только 0 и 1, из-за промахов по клавишам могут возникать 10, 2 или какие-то иные числа.

Функция СЧЁТ считает число ячеек, содержащих числовые данные. Соответственно, если есть ошибки ввода, когда введено не число (примеры из моего опыта: «1,!», «5е»), то эта функция выдаст меньшее значение, чем функция СЧЁТЗ, считающая число непустых ячеек.

В таблице 1 приведен демонстрационный пример небольшой таблицы данных о росте и поле десяти человек, содержащей почти все возможные ошибки, а также результат применения вышеназванных функций к столбцам с данными. Рекомендуется перенести эти данные

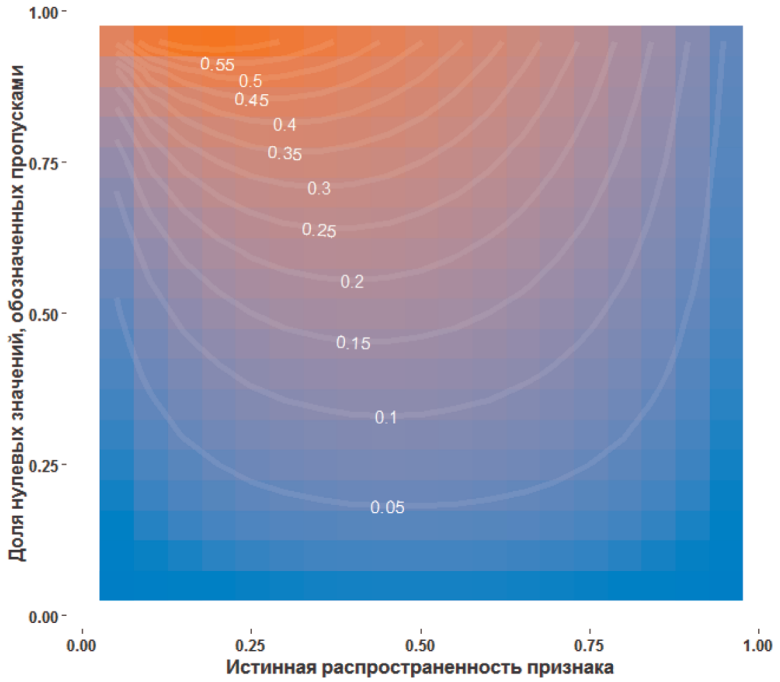

Рис. 2. Разность распространенности бинарного признака, посчитанной по таблице, в которой часть ячеек с нулевым значением (отсутствие признака) обозначена пропуском данных, и истинной распространенности признака в выборке. Указаны изолинии разности.

Таблица 1.

Образец таблицы с данными, содержащей ошибки формата и ошибки ввода. Последние четыре строки содержат результат применения функций МИН, МАКС, СЧЁТЗ и СЧЁТ, позволяющий выявить ошибки

Номер пациента	Рост, см	Пол (ж — 0, м — 1)
1	170	1
2	181	!
3	0	0
4	180	1
5	!76	0
6	177	1
7	262	0
8	16	2
9	183	10
10	164	0
Минимум	0	0
Максимум	262	10
Значений	10	10
Чисел	9	9

Таблица 2.

Результат описанного в тексте статьи исправления данных, приведённых в таблице 1. Последние четыре строки содержат результат применения функций МИН, МАКС, СЧЁТЗ и СЧЁТ

Номер пациента	Рост, см	Пол (ж — 0, м — 1)
1	170	1
2	181	1
3		0
4	180	1
5	176	0
6	177	1
7	162	0
8	160	1
9	183	0
10	164	0
Минимум	160	0
Максимум	183	1
Значений	9	10
Чисел	9	10

в электронный вид, чтобы было удобнее отслеживать поиск ошибок.

Вначале займёмся столбцом с данными о росте. Во-первых, в нем «запредельные» значения минимума и максимума — 0 и 262. Скорее всего, нулевым значением тут обозначено отсутствие данных, поэтому удалим это значение (для поиска этих значений удобно использовать встроенный в электронные таблицы фильтр данных). Теперь минимальное значение — 16 — тоже явная ошибка ввода. Но устранить ее без обращения к исходным данным сложно: это может быть и 160, и 166, и 176. Для простоты заменим это значение на 160. Максимальное значение (262) тоже, скорее всего, является ошибкой ввода, так как хотя такой рост и меньше максимально зарегистрированного у человека, представляется крайне маловероятным, что это не ошибка (тем более, что согласно таблице это женщина), поэтому лучше свериться с исходными данными. Предположим, что это ошибочный ввод числа 162.

Во-вторых, непустых ячеек в диапазоне больше, чем ячеек с числовыми данными (10 и 9 соответственно до устранения ошибочного нулевого значения, 9 и 8 после такового). Просмотр значений (его также можно провести с помощью фильтра; нечисловые данные идут в нем после числовых) показывает, что проблема в пятом пациенте — там стоит «!76». Скорее всего, это ошибочный ввод числа «176».

Переходим к столбцу с данными о поле. Во-первых, снова есть значение, выходящее за область допустимых значений — 10 — у девятого пациента. Исправить эту ошибку без обращения к исходным данным невозможно. Предположим, что это все же была женщина, и заменим значение нулем. Но максимум снова выходит за возможные пределы — теперь это значение «2» у восьмого пациента. Скорее всего, это мужчина (заполнявший таблицу «промахнулся» по клавише). Исправим значение, хотя в реальности лучше свериться с исходными данными.

Во-вторых, снова есть одно нечисловое значение — это восклицательный знак у второго пациента. Скорее всего, это снова ошибка во вводе единицы (был зажат Shift).

Таблица 2 показывает окончательный результат исправления данных.

В заключение обсуждения ошибок, связанных с заполнением таблиц, хотелось бы напомнить важный принцип: «мусор на входе — мусор на выходе» [7]. Он в полной мере применим и к статистической обработке данных. Даже самые передовые статистические методы не помогут получить верные выводы, если исходный материал для них содержит много ошибок.

СПИСОК ЛИТЕРАТУРЫ

1. Lang T. Twenty Statistical Errors Even YOU Can Find in Biomedical Research Articles. Croat Med J. 2004; 45 (4): 361–370

2. Worthy G. Statistical analysis and reporting: common errors found during peer review and how to avoid them. Swiss Med Wkly. 2015; 145: w14076. DOI: 10.4414/smw.2015.14076

3. Lee S.S. Avoiding negative reviewer comments: common statistical errors in anesthesia journals. Korean Journal of Anesthesiology. 2016; 69(3): 219–226. DOI: 10.4097/kjae.2016.69.3.219

4. Munyisia E.N., Reid D., Yu P. Accuracy of outpatient service data for activity-based funding in New South Wales, Australia. Health Inf Manag J. 2017; 46(2): 78–86. DOI: 10.1177/1833358316678957

5. Fararouei M., Marzban M., Shahraki G. Completeness of cancer registry data in a small Iranian province: a capture–recapture approach. Health Inf Manag J. 2017; 46(2): 96–100. DOI: 10.1177/1833358316668605

6. Kilkenny M.F., Robinson K.M. «Data quality: «Garbage in – garbage out». Health Inf Manag J. 2018; 47(3): 103-105. DOI: 10.1177/1833358318774357

7. Vogt W.P. Garbage In, Garbage Out. (n.d.). In: Dictionary of Statistics & Methodology. 2005. P. 158. DOI: 10.4135/9781412983907.n809