**Рекомендации для составления исследовательской базы данных**

Базу данных научного исследования рекомендуется вести в формате электронных таблиц Excel, так как такой формат является общедоступным и может быть импортирован в любые программы для проведения статистической обработки данных.

База данных представляет из себя таблицу, строки которой содержат данные объектов исследования, а столбцы – исследуемые признаки. Исследуемые признаки (столбцы таблицы) в терминах баз данных называются переменными или полями базы данных, а строки таблицы, содержащие данные каждого объекта исследования – записями.

Внимание! Заголовки базы данных рекомендуется заполнять стандартизированными терминами на английском языке.

**Обязательные элементы базы данных исследования**

Таблица данных обязательно должна иметь поле «ID» (идентификатор обследуемого), содержащее информацию, которая позволит однозначно определить каждого объекта: это может быть порядковый номер объекта, индивидуальный номер обследуемого, название растения, индентификатор клеточных культур и т.п.

Главное, чтобы ID был уникальным значением для каждого объекта!

Если дизайн исследования предусматривает деление обследуемых на группы, таблица данных должна содержать поле «Group».

Если исследование является мультицентровым, таблица данных должна содержать поле «Research Center»; если в работе принимают участие несколько исследователей, такая информация также должна быть отражена в соответствующем поле базы данных; если измерения одного количественного признака проводились при помощи нескольких вариантов оборудования, это тоже нужно зафиксировать в базе данных – все это позволит выявить систематические ошибки в данных.

Внимание: все перечисленные обязательные элементы базы данных должны быть заполнены по каждому объекту исследования!

Не забудьте перед заполнением таблицы на листе (Description of variables) указать внутренний номер исследования (Internal research number), который должен совпадать с номером, указанным в DMP!

**Требования к переменным**

- одна переменная (один признак исследуемого объекта) – один столбец в таблице данных;

- каждая переменная относится к одному типу данных – Nominal (номинальный), Ordinal (порядковый), Quantitative (количественные);

- количественная (Quantitative) информация в один столбец вносится строго с использованием одних единиц измерения и в одном формате с одинаковой точностью (числом знаков после запятой), при этом обозначения единиц измерения в ячейки таблицы не вносятся (они могут указываться в названии столбца и должны быть отражены в легенде). Все измеренные величины необходимо сохранять в количественном виде, группировки и преобразования можно будет провести дополнительно, создав для этого новые переменные.

|  |  |  |
| --- | --- | --- |
| **Correct:** |  | **Incorrect:** |
|  |  |  |  |  |
| Height, m |  | Height | Height | Height |
| 1,75 |  | 175 | 1,80 | От 170 до 180 |
| 1,64 |  | 1,64 | 1,64 m | <170 |
| 1,81 |  | 181,12 | 1 m 81 cm | >180 |

- Ordinal (порядковые) признаки следует представлять в виде цифровых обозначений, что позволит применить к ним подходящие методы статистической обработки.

|  |  |  |
| --- | --- | --- |
| **Correct:** |  | **Incorrect:** |
|  |  |  |  |  |  |  |  |  |
| Degree  | Intensity  | Stage  | Speed |  | Degree  | Intensity  | Stage  | Speed |
| 1 | 1 | 1 | 1 |  | I | + | A | Низкая |
| 2 | 2 | 2 | 2 |  | II | ++ | B | Средняя |
| 3 | 3 | 3 | 3 |  | III | +++ | C | Высокая |

- качественные (Nominal или Ordinal) признаки необходимо представлять единообразно во всей таблице (одна подгруппа - один способ обозначения).

|  |  |  |
| --- | --- | --- |
| **Correct:** |  | **Incorrect:** |
|  |  |  |
| Gender |  | Gender |
| m |  | м |
| f |  | ж |
| m |  | муж. |
| f |  | женский |

- переменным, обозначающим повторные измерения одного признака, рекомендуется давать одинаковые наименования с добавлением индекса, обозначающего время или этап исследования, например:

- weight\_0; weight\_1;

- temperature\_morning; temperature\_evening;

- BR\_1; BR\_2; BR\_3; BR- Breathing rate

- и т.д.

**Требования к заполнению информации об объектах исследования**

Обязательное правило: один объект наблюдения (пациент, животное, растение) – одна строка.

**Обозначение отсутствия информации**

1. В случае, если у отдельных обследуемых некоторые измерения не были выполнены, соответствующую ячейку таблицы данных необходимо оставить пустой. При этом следует помнить, что для качественной статистической обработки данных доля пропущенных значений в базе данных исследования не должна превышать 10%.

Внимание! С точки зрения дальнейшей статистической обработки данных грубой ошибкой является замена пропущенных значений нулями.

|  |  |  |
| --- | --- | --- |
| **Correct:** |  | **Incorrect:** |
|  |  |  |  |  |
| Full name | Height, m |  | Full name | Height |
| Ivanov I.I. | 1,75 |  | Ivanov I.I. | 1,75 |
| Petrov P.P. |   |  | Petrov P.P. |  0 |
| Sidorov S.S. |  1,86 |  | Sidorov S.S. |  1,86 |
| Vasiliev V.V. |  |  | Vasiliev V.V. | No |
| Dmitriev D.D. |  |  | Dmitriev D.D. | - |

1. В случае, когда измерение признака возможно только при определенных условиях, описание такого признака необходимо представить в виде двух переменных:
	1. Факт выполнения условия;
	2. Результат измерения (заполняется только для тех обследуемых, у которых условие оказалось выполненным; у остальных обследуемых на месте результата измерения в таблице данных остается пустая ячейка).

|  |  |  |
| --- | --- | --- |
| **Correct:** |  | **Incorrect:** |
|  |  |  |  |
| Focus of inflammation | The size of the focus, cm3 |  | The size of the focus, cm3 |
| Yes | 0,3 |  | 0,3 |
| No |   |  |  0 |
| Yes |  1,2 |  |  1,2 |

**Представление информации в файле**

Для удобства восприятия информации в таблице данных имена переменных (т.е. заголовки столбцов таблицы) лучше формулировать в виде кратких обозначений.

|  |  |  |
| --- | --- | --- |
| **Briefly:** |  | **Details:** |
|  |  |  |  |  |
| SBP\_0 | DBP\_0 |  | Systolic blood pressure at admission (mmHg) | Diastolic blood pressure at admission (mmHg) |
| 120 | 80 |  | 120 | 80 |
| 125 | 75 |  | 125 | 75 |
| 140 | 90 |  | 140 | 90 |

Отдельно нужно представить подробное описание переменных (Description of variables) с указанием полного наименования, единиц измерения (обязательно), точности измерения (обязательно), границ нормы и/или пределов изменения (желательно; это поможет проводить проверку данных на наличие ошибок). Для качественных и порядковых признаков нужно расшифровать каждый вариант обозначения.

Таким образом, файл базы данных должен содержать два листа:

* первый – Dataset (база данных);
* второй – Description of variables (описание переменных).

**Инструкция по работе с базой данных исследования**

1. Сформируйте перечень регистрируемых признаков. Обязательным для заполнения являются уже внесенные в базу данных признак:
	* ID (идентификатор обследуемого)

Далее в шапке таблицы указываются все измеряемы признаки объектов исследования.

Если исследуемыми объектами являются животные, то ID (идентификатором) может быть порядковый номер животного, код. Не указываются рост, дата рождения, указываются группа, дата включения в исследование, возраст (стадия развития), пол, вес.

В случаях, когда исследуемыми объектами являются растения, обязательными признаками для включения в базу данных являются:

* + ID (фракция, вид растения);
	+ Group;
	+ Сollection date;
	+ Part of the plant (flowers, root, etc.).

Если исследование состоит из данных анкетирования (удовлетворенность населения работой медицинской организации, качество оказываемой помощи), то в базе данных можно записать признаки:

* + Questionnaire number / code;
	+ Medical Organization / Institution code;
	+ Type of medical organization (если в исследовании были задействованы несколько МО);
	+ Date of questionnaire;
	+ Short names of questions from the questionnaire.

Остальные признаки вносятся самостоятельно: на листе «Description of variables» представляете название и подробную характеристику каждой переменной (здесь одна переменная – одна строка (пример см. рис. 1));



Рис. 1

на листе «Database» каждой переменной будет соответствовать столбец; названия переменных указываются в первой строке таблицы (см. рис. 2).



Рис. 2

2. При внесении в базу данных категориальных признаков (nominal, ordinal) для предотвращения ошибок рекомендуется устанавливать в соответствующем столбце проверку данных (в примере ниже показана проверка данных, которые будут вноситься в столбец «Gender»):

- выбираем меню «Данные» - «Проверка данных» (см. рис. 3):



Рис. 3

- в появившемся окне диалога устанавливаем тип данных «Список», в разделе «Источник» указываем варианты значений через точку с запятой без пробелов - «m; f» и нажимаем «Ок» (см. рис. 4):



Рис. 4

Теперь проверка данных будет выполняться всегда при внесении информации в столбец «Gender».

В Description of variables так же необходимо указать, как кодировались качественные признаки (nominal, ordinal). Например, степени тяжести заболевания: 1-легкая степень тяжести, 2- средняя степень тяжести, 3- тяжелая степень тяжести.

Если кодируются признаки Group, Gender, то необходимо это указать в описании переменных, чтобы в дальнейшем не получать затруднения при интерпретации результатов.

Количественные признаки тоже необходимо проверять сразу на этапе внесения информации в базу данных – например, установить допустимый формат (целое число) и разброс значений (от 0 до 250) признака Height (см. рис. 5):



Рис. 5

Попытка внести неверное (выходящее за пределы допустимого диапазона) значение признака пресекается сообщением «Введенное значение неверно» (см. рис. 6):



Рис. 6

Признаки, которые являются вторичными и могут быть рассчитаны на основе других признаков, нужно делать рассчитываемыми автоматически по формуле, например – значение индекса массы тела (Body mass index) можно рассчитать, зная рост (Height) и вес (Weight) (см. рис. 7):



Рис. 7

После установки всех ограничений, связанных с проверкой, можно начинать вносить данные в таблицу.

**Распространенные ошибки при составлении базы данных исследования:**

1. Не создается лист с описанием переменных. На дальнейших этапах исследователь может забыть, что он имел в виду под различной кодировкой признаков, заполнит ячейки некорректными значениями, что негативно отразится на результатах исследования.
2. Если в исследовании несколько баз данных, то каждую таблицу необходимо разместить на отдельном листе для удобного использования статистических программ.
3. Ошибки при планировании исследования (например, неправильно рассчитанный размер выборки).
4. Неправильное кодирование терминов на этапе обработки данных.
5. Использование неадекватных методов статистического анализа.