

# В ПОМОЩЬ ИССЛЕДОВАТЕЛЮ

## ОСНОВНЫЕ ПОНЯТИЯ СТАТИСТИКИ

УДК: 311  
03.01.02 — биофизика  
Поступила 30.06.2020 г.

**А. П. Баврина**

ФГБОУ ВО «Приволжский исследовательский медицинский университет» Минздрава России, Нижний Новгород

Рассматриваются основные понятия статистики, которые позволяют исследователю правильно подойти к статистическому анализу полученных данных и адекватно интерпретировать результаты исследования. Особое внимание уделено основным ошибкам при планировании исследования, таким как ошибки отбора (ошибки репрезентативности и ошибки выборки) и ошибки регистрации. Предлагаются способы минимизации вышеперечисленных ошибок и расчет минимального допустимого объема выборки. Представлена детальная систематизация различных типов выборок, подробно разобраны их особенности. Изложены правила описания малых выборок, к использованию которых часто прибегают в медико-биологических исследованиях. Рассмотрены доступные методы импутирования — воссоздания на основе имеющихся данных наиболее вероятного пропущенного значения.

**Ключевые слова:** генеральная совокупность; вероятностная и невероятностная выборка; малая выборка; импутирование; уровень значимости; доверительная вероятность.

## BASIC CONCEPTS OF STATISTICS

**A. P. Bavrina**

Privolzhsky Research Medical University, Nizhny Novgorod

The basic concepts of statistics allowing the researcher to arrange a correct approach to the statistical analysis of the data obtained and to correct interpretation of the research results are considered. Particular attention is given to the major errors in study design, such as selection errors (errors of representativeness and sampling errors) and registration errors. The ways to minimize the above mentioned errors and the calculation of the minimum allowable sample size are proposed. A detailed systematization of various types of samples is presented, their features are analyzed in detail. The rules for describing small samples often being used in biomedical research are outlined. The available methods of imputing — recreating the most probable missing value based on the available data are considered.

**Key words:** general population; probabilistic and improbable sample; small sample; imputation; level of significance, confidence level.

Настоящая статья продолжает серию публикаций по применению медико-биологической статистики в научных исследованиях [1]. Цель данной серии — познакомить читателей с принципами и правилами представления медико-биологических данных в научных исследованиях, сформировать целостное представление о методах статистической обработки результатов медико-биологических исследований, а также предупредить возникновение ошибок при статистической обработке данных.

Прежде чем перейти непосредственно к аналитической статистике, необходимо уделить внимание важным базовым понятиям: их понимание поможет исследователю правильно подойти к статистическому анализу полученных данных и адекватно интерпретировать результаты.

До недавнего времени исследования планировались довольно редко, они просто «случались» [2].

Однако сейчас, в эру грантов и госзаданий, довольно часто научные исследования качественно планируются. Организации и дизайну научных исследований будет посвящена отдельная статья. Но основные понятия мы обсудим в данной публикации.

### ВВЕДЕНИЕ

Итак, описательная статистика описывает выборку, а аналитическая статистика на основе выявленных свойств переносит их на всю генеральную совокупность.

**Генеральная совокупность** — это теоретически бесконечно большая или приближающаяся к бесконечности совокупность всех элементов, которые могут быть к ней отнесены. Естественно, изучить все элементы генеральной совокупности не представляется возможным, поэтому для ее описания используют выборку.

**Выборка** — это случайно отобранные единицы генеральной совокупности, достаточные для того, чтобы в ней были выражены существенные черты изучаемого распределения.

Если описание генеральной совокупности выборкой является полным и корректным, то выборка называется **репрезентативной**. При этом одна и та же выборка может быть репрезентативной и нерепрезентативной для разных генеральных совокупностей.

Подобрать выборку, которая полностью повторяет все особенности генеральной совокупности, невозможно. Поэтому при организации исследования необходимо стремиться, чтобы выборка сохраняла наиболее существенные для данного исследования характеристики генеральной совокупности. Это приведет к наименьшей потере информации.

Кроме репрезентативности, каждая выборка характеризуется ошибкой, при этом ошибка репрезентативности и ошибка выборки — это разные явления, возникающие при отборе. Кроме ошибок отбора, выборки характеризуются ошибкой регистрации (рис. 1).

К сожалению, в современной литературе наблюдается путаница при их описании. Часто ошибка выборки подменяется ошибкой репрезентативности, и наоборот. Либо ошибка выборки представляется авторами как «более точная» ошибка репрезентативности, что является в корне неверным.

Выявление и минимизация каждого вида ошибок чрезвычайно важны для проведения любого научного исследования.

**ОШИБКИ ОТБОРА**

**Ошибка репрезентативности** — расхождение между величинами, которые получены при выборке показателей, и данными величинами генеральной совокупности. Ошибка репрезентативности возникает при недостаточно равномерном представлении в выборке различных категорий единиц генеральной совокупности (случайная ошибка репрезентативности) либо при нарушении установленных пра-

вил отбора (систематическая ошибка репрезентативности). Систематическая ошибка репрезентативности может быть преднамеренной и непреднамеренной и всегда имеет однонаправленное отклонение от истинных значений. Величина ошибки характеризует степень надежности результатов отбора. Таким образом, ошибка репрезентативности — это результат деятельности самого исследователя, не зависящий от объема выборки.

Первым следствием ошибок репрезентативности является формирование смещенной (нерепрезентативной) выборки, которая существенным образом отличается от генеральной совокупности. Результаты, полученные с использованием смещенной выборки, представляют значительно меньшую ценность для науки.

Вторым следствием ошибок репрезентативности является отсутствие сопоставимости сравниваемых групп по характеристикам, существенно влияющим на результаты исследования [2]. Например, несоответствие выборок по существенным признакам (пол, вес, возраст и т.д.), которое может обуславливать обнаружение статистически значимых различий между группами не по причине влияния нового метода лечения или диагностики, а из-за расхождения групп по данным характеристикам.

К сожалению, ошибки репрезентативности математически неизмеримы, если не известны значения изучаемого показателя для генеральной совокупности. В связи с этим они являются наиболее опасными и приводят к смещению результатов исследования в ту или иную сторону. Кроме того, ошибки репрезентативности неустранимы ввиду того, что выборочная совокупность никогда полностью не воспроизводит характеристики генеральной совокупности. Однако минимизировать возникновение ошибок репрезентативности возможно на этапе планирования исследования, путем подробного изучения характеристик генеральной совокупности и эффективной рандомизации единиц выборки.

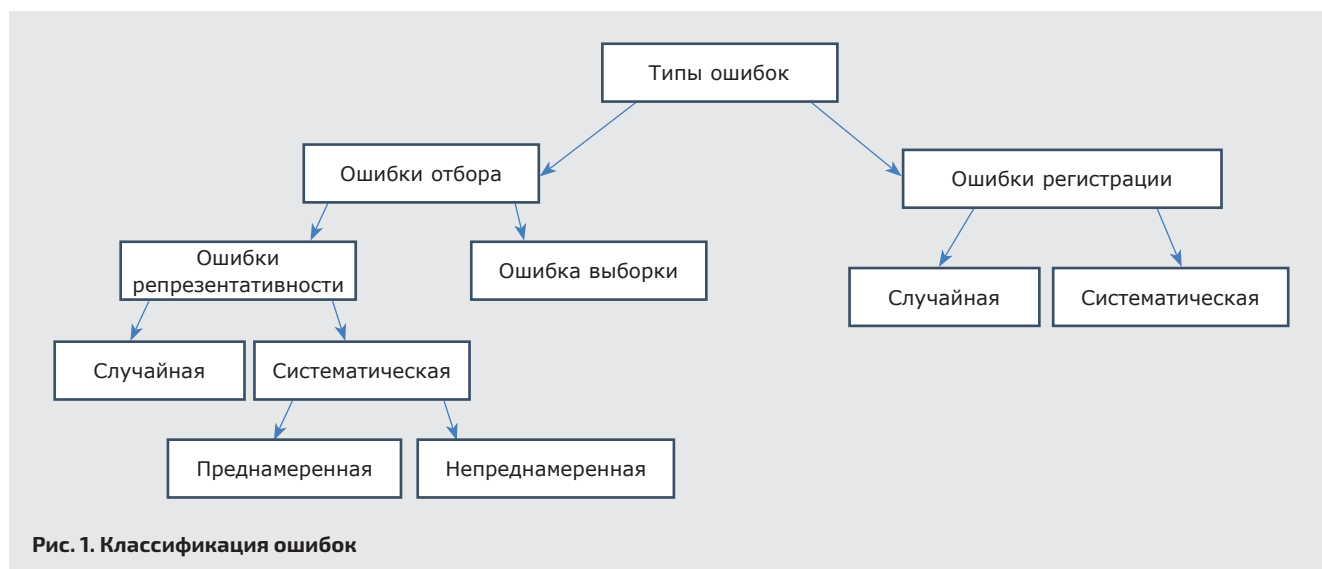


Рис. 1. Классификация ошибок

Ошибки репрезентативности выявляют также на этапе проведения статистического анализа — путем сравнения основных исходных характеристик исследуемых выборок, способных влиять на результаты исследования. Отсутствие статистически значимых различий между основными характеристиками сравниваемых выборок (например, пол, возраст и другие важные для исследования критерии) предполагает, что они однородны по данным показателям и выявленные преимущества нового вида лечения либо диагностической процедуры связаны именно с ними, а не являются следствием неоднородности выборок.

**Ошибка выборки** — отклонение средних характеристик выборочной совокупности от средних характеристик генеральной совокупности. Иными словами, всегда присутствует вероятность выхода средних значений исследуемого признака за пределы установленного доверительного интервала (в медико-биологических исследованиях — за пределы 95% доверительного интервала).

Ошибка выборки является случайной и всегда связана с ее **объемом** — числом наблюдений  $n$ , которые образуют выборку. Как правило, объем выборки  $n$  значительно меньше объема всей генеральной совокупности. При этом чем больше объем выборки, тем ниже случайная ошибка выборки.

Минимизировать случайную ошибку выборки воз-

можно путем **расчета минимального допустимого объема выборки**. К сожалению, этой теме в руководствах по статистической обработке результатов медико-биологических исследований не уделено должного внимания, хотя расчет предельных объемов выборок является обязательным условием для обеспечения статистической значимости результатов современных исследований. Кроме того, до настоящего момента в практике медицинских исследований, как российских, так и зарубежных, отсутствует четко установленная методология определения минимально необходимого объема выборки [3].

Существует несколько наиболее простых методов для определения минимального объема выборки, не требующих предварительных знаний о свойствах генеральной совокупности. К таким методам относят таблицы расчета предельного объема выборки по нескольким методикам (табл. 1–3). Данные таблицы изначально разрабатывались для социально-гигиенических исследований, что предполагает довольно большой минимальный объем выборок, однако они могут использоваться и при планировании клинических исследований.

Считается, что величина допускаемой ошибки в медицинских исследованиях должна быть не более 5%.

Метод В.И. Паниотто, в отличие от двух предыду-

Таблица 1

Определение минимального объема выборки по методу N. Фох [4]

| Предельная величина случайной ошибки выборки, % | Объем выборки |
|---|---------------|
| 10  | 88            |
| 5   | 350           |
| 3   | 971           |
| 2   | 2188          |
| 1   | 8750          |

Таблица 2

Определение минимального объема выборки по методу К.А. Отдельновой [5]

| Уровень значимости (p) | Уровень точности                                   |  |  |
|------------------------|--|--|--|
|                        | Ориентировочное знакомство (пилотное исследование) | Исследование средней точности (публикация статьи с последующим более глубоким изучением) | Исследование повышенной точности (диссертационное исследование с формированием окончательных заключений) |
| 0,05                   | 44   | 100  | 400  |
| 0,01                   | 100  | 225  | 900  |

Таблица 3

Определение минимального объема выборки по методу В.И. Паниотто [6]

| Объем генеральной совокупности | 500 | 1000 | 2000 | 3000 | 4000 | 5000 | 10000 | 100000 | ∞    |
|--------------------------------|-----|------|------|------|------|------|-------|--------|------|
| Объем выборки                  | 222 | 286  | 333  | 350  | 360  | 370  | 385   | 398    | 4000 |

Примечание: уровень значимости 0,05.

щих, учитывает предварительные данные — предположительный объем генеральной совокупности.

В настоящий момент некоторые статистические программы позволяют автоматически рассчитать необходимый объем выборки (например, IBM SPSS Statistics v. 26). Кроме того, при планировании исследования можно воспользоваться онлайн-калькуляторами, автоматически вычисляющими предельный объем выборки. Их использование требует знаний об объеме генеральной совокупности и приемлемой для исследования доверительной вероятности.

Применение вышеперечисленных таблиц относится к методам, не требующим предварительных знаний об изучаемом явлении. Кроме них, существует группа методик, требующих от исследователя знания предварительных данных. Использование данных математических методов дает меньшие минимальные объемы выборки и характеризуется большей точностью. Однако необходимым минимальный объем выборки с использованием точных методов рассчитывается достаточно сложно. При интересе читателей к данной тематике в будущих публикациях будет уделено внимание расчету объема выборок по специализированным формулам.

Исходя из вышесказанного, при планировании и проведении крупномасштабных современных научных исследований на высоком уровне расчет необходимого минимального объема выборки является обязательным и требует описания с указанием метода в разделе, посвященном дизайну исследования. Но следует отметить, что при планировании клинических исследований в рамках диссертационных работ набор столь больших групп часто является невыполнимым.

Кроме того, объем выборок чрезвычайно важен при дальнейшем использовании статистических критериев. Известно, чем больше объем выборки, тем чувствительнее критерий (тем выше его мощность). При достаточно больших выборках малейшее различие оказывается статистически значимым. И наоборот, если выборки малы, даже большие различия могут быть статистически незначимы. Зная эти закономерности, можно заранее определить численность выборок, необходимую для выявления эффекта [7].

Таким образом, ошибки репрезентативности обычно математически неизмеримы, не зависят от объема выборки, могут быть случайными и систематическими, зависят лишь от деятельности исследователя. Ошибки выборки, напротив, поддаются математическому описанию, напрямую зависят от объема выборки, носят случайный характер.

### **ОШИБКИ РЕГИСТРАЦИИ (ИЗМЕРЕНИЯ)**

Под ошибкой регистрации имеется в виду не инструментальная погрешность работы измерительной аппаратуры, а ошибки (в основном систематические), допускаемые исследователями в описании основных характеристик выборки. Например, ошибки регистрации возникают при различной

глубине исследования двух сравниваемых выборок, в частности, группы здоровых людей и группы людей, имеющих заболевание.

Обычно ошибки регистрации рождаются в результате непосредственной деятельности исследователя. В некоторых случаях причиной их появления являются сами испытуемые, которые могут непреднамеренно исказить некоторые характеристики. Такие ошибки регистрации являются случайными. Например, здоровые люди, входящие в группу сравнения, хуже помнят свои заболевания, чем пациенты, страдающие от какого-либо заболевания. Кроме того, испытуемые могут намеренно исказить некоторые данные, например, утаивать факт курения, а также снижать количество выкуриваемых ежедневно сигарет.

\*\*\*

Сведение к минимуму ошибок выборок является одной из основных целей статистического анализа данных, полученных в медико-биологических исследованиях [2]. Однако следует помнить, что никакая статистическая обработка данных не в состоянии устранить неизвестную ошибку отбора или регистрации, которая может возникать на любом этапе исследования.

### **КЛАССИФИКАЦИЯ ВЫБОРОК ПО СПОСОБУ ОТБОРА ЕДИНИЦ**

Существует несколько классификаций выборок.

1. По способу отбора единиц выборки.
2. По этапу отбора.
3. По объему.

Согласно классификации по способу отбора единиц (рис. 2), различают вероятностные и невероятностные выборки.

**Вероятностные выборки.** Отбор в них производится по принципам случайности, все субъективные критерии исключаются.

Разновидностями вероятностных выборок являются случайная, механическая, стратифицированная и серийная выборки.

**Случайная выборка** — та, где отбор единиц выборки производится непосредственно из всей массы единиц генеральной совокупности, причем каждая единица генеральной совокупности имеет известную и равную вероятность отбора.

Случайный отбор может быть проведен двумя способами:

1. Метод таблиц случайных чисел. На первом этапе каждой единице генеральной совокупности присваивается индивидуальный номер от 01 до  $n$ . Далее с помощью существующих таблиц равномерно распределенных случайных чисел выбирается их последовательность в нужном для выборки объеме. Кроме того, получить набор случайных чисел можно с помощью инструмента «Генератор случайных чисел», входящего в надстройку «Пакет анализа» Microsoft Excel. Затем с использованием полученного ряда случайных чисел отбираются единицы генеральной совокупности, имеющие те же уникальные номера.

2. Метод жеребьевки (лотереи). В научных исследованиях данный метод часто называют методом случайных конвертов. Как и в предыдущем случае, каждой единице генеральной совокупности присваивается уникальный номер. Далее эти номера записываются на индивидуальных карточках, помещаются в одинаковые конверты, они тщательно перемешиваются в корзине и вынимаются случайным образом. Недостаток этого метода заключается в трудоемкости: процесс подготовки карточек, их подписания, помещения в конверты, перетасовки — достаточно долгий и сложный, особенно если генеральная совокупность велика.

Многие авторы научных и диссертационных исследований указывают, что при формировании выборки использовался случайный отбор. Однако часто это не соответствует действительности. В большинстве клинических исследований применяются невероятные методы, к которым случайный отбор и случайная выборка не имеют никакого отношения. Уточним еще раз, что формирование вероятностной выборки (в том числе и случайной) подразумевает наличие полного списка всех элементов генеральной совокупности, из которого случайным образом отбираются единицы случайной выборки.

Таким образом, случайная выборка является этапом в научных исследованиях.

**Механическая** (систематическая) **выборка** — это разновидность случайной выборки, упорядоченная по какому-либо признаку. Генеральная совокупность разбивается по какому-либо нейтральному признаку (буква алфавита, год рождения, номер телефона и др.) на интервалы, из которых выбирается по одной единице выборки. Отправная точка определяется произвольно (первый элемент выбирается случайно), затем с шагом  $n$  выбирается каждый  $i$ -ый элемент. Для минимизации систематической ошибки репрезентативности из каждого интервала берется элемент, близкий к его середине. При анализе достаточно большой генеральной совокупности механический отбор равен случайному отбору по точности анализа.

**Стратифицированная** (типическая) **выборка** применяется для отбора элементов из неоднородной (статистически сложной) генеральной совокупности. Для формирования стратифицированной выборки генеральная совокупность разбивается на страты — качественно однотипные группы, объединенные по признакам, которые влияют на исследуемые показатели (например, разделение по профессиональной квалификации, стадии заболевания и т.д.). Далее в каждой страте отбор осуществляется с помощью случайного или механического метода.

**Серийная** (гнездовая, или кластерная) **выборка** — та, для которой из генеральной совокупности отбираются не отдельные элементы, а группы с равным объемом — кластеры (гнезда). Затем в выборку включаются либо все элементы кластера (тогда выборка называется одноступенчатой серийной), либо ис-



Рис. 2. Классификация выборок по способу отбора

пользуется случайный или механический отбор для каждого кластера (в этом случае выборка называется двухступенчатой серийной).

Может показаться, что стратифицированная и серийная выборки похожи по методологии отбора. Однако они обладают прямо противоположными свойствами. Отобранные кластеры в серийной выборке должны быть максимально однородны между собой, при этом элементы внутри кластера должны быть максимально разнородны. Напротив, сформированные страты в стратифицированной выборке должны быть максимально разнородны между собой, а элементы внутри каждой страты — максимально однородны.

Разновидностью серийной выборки является **территориальная выборка**. В данном случае в качестве кластеров выступают географические территории: округа, области, города, районы и т.д. Территориальная выборка может быть одноступенчатой, двухступенчатой или многоступенчатой в зависимости от количества этапов отбора.

Таким образом, все виды вероятностных выборок в клинических исследованиях практически не применимы и встречаются крайне редко, но в сфере общественного здоровья, эпидемиологии, социальнo-гигиенических исследованиях используются очень часто.

**Невероятностные выборки** — те, для которых отбор осуществляется не случайным образом, как в вероятностных выборках, а по субъективным критериям — доступности, типичности, целенаправленности и т.д. В медико-биологических исследованиях используется в большинстве случаев именно данный подход.

Невероятностные выборки также подразделяются на несколько подтипов.

**Квотная выборка** наиболее часто используется

в клинических исследованиях и характеризуется изначальным разбиением генеральной совокупности на непересекающиеся группы, подобно стратифицированной выборке. Например, разбиение на квоты может проводиться по возрасту, полу, наличию изучаемого заболевания и т.д. Затем из каждой квоты произвольно выбирается известное заранее количество элементов на основании предпочтений исследователя. Отбор элементов в квотную выборку не является случайным, вероятность каждого элемента попасть в выборку неравна. Требования квотной выборки могут быть жесткими (соответствие пропорций генеральной совокупности и выборки по всем сочетаниям признаков) и пониженными (соответствие пропорций по каждому параметру отдельно). Например, исследование подразумевает разбиение на группы по трем параметрам: пол (2 градации — мужчины и женщины), возраст (4 градации — 40–50, 51–60, 61–70 лет, старше 70 лет) и стадия заболевания (3 градации — 1, 2-я и 3-я стадия заболевания). При жестких требованиях в исследовании должно быть  $2 \cdot 4 \cdot 3 = 24$  группы; при пониженных требованиях:  $2 + 4 + 3 = 9$  групп. Обычно предпочтение отдается пониженным требованиям, так как основная цель использования квотной выборки — уменьшение количества элементов.

Считается, что в квотной выборке можно использовать меньшее количество элементов, так как у исследователя имеются известные параметры и характеристики генеральной совокупности (пол, возраст, частота встречаемости исследуемого заболевания и т.д.) и он может подобрать элементы выборки, опираясь на эти знания.

Однако это не так. Ф.Э. Шереги экспериментально доказал несправедливость заключения о возможности уменьшения количества элементов выборки при использовании квотного отбора [8]. На первом этапе Ф.Э. Шереги использовал случайный вероятностный отбор трехсот рабочих текстильной фабрики и математически доказал высокую репрезентативность случайной выборки. На следующем этапе он применил невероятностные квотные выборки, где провел разбиение на квоты по возрасту и национальности. Первая квотная выборка составила 200 элементов, вторая — 100. Оказалось, что в выборке из двухсот элементов 80% всех параметров имели ошибку не более 3%, при этом только 6% параметров давали ошибку более 5%. В выборке из ста элементов только 55% параметров имели ошибку менее 3%, а 25% параметров имели ошибку более 5%.

**Стихийная выборка** также часто применяется в медицинских исследованиях и характеризуется отбором наиболее доступных элементов. Примером стихийного отбора является добровольное заполнение анкет пациентами медицинских учреждений, а также прохождения электронных опросов. Объем и состав стихийной выборки до окончания предварительного этапа остаются неизвестными и не зависят от исследователя, определяются только активностью анкетированных людей.

Кроме того, при использовании стихийного отбора существуют две проблемы:

- проблема появления существенных ошибок регистрации, которые были описаны выше;
- проблема пропущенных значений.

**Проблема пропущенных значений** возникает, когда респондент пропустил вопрос по причине простого нежелания отвечать, либо респондент не помнит какие-либо важные для исследования факты. Причин бывает множество. Проблема пропущенных значений возникает не только при использовании анкет и опросов. Пропуски данных по различным причинам появляются при проведении клинических и экспериментальных исследований.

Проблема пропущенных значений решается несколькими способами:

удаление данного респондента/пациента/лабораторного животного из исследования целиком или удаление только пропущенных значений с сохранением остальных данных;

импутирование — воссоздание на основе остальных данных наиболее вероятного пропущенного значения.

Остановимся подробнее на существующих методах **импутирования**, которые позволят сохранить полученную исследователем с приложением больших усилий информацию. Методы импутирования могут применяться не только при стихийном отборе, но и при любых других его видах.

1. Метод безусловного среднего. Является самым простым методом импутирования и заключается в заполнении пропущенных значений средним арифметическим по группе. Однако использование данного метода значительно искажает данные.

2. Метод подбора внутри групп. Предполагает, что вся совокупность объектов разбивается на группы по определенному признаку, внутри каждой группы для заполнения пропусков используются только присутствующие в ней значения [9].

3. Метод подбора ближайшего соседа (метод Hot Deck). Заключается в подстановке вместо пропуска значения для наиболее близкого объекта с полной информацией [10]. Для обнаружения наиболее близких объектов может быть использован кластерный анализ, позволяющий определить расстояния между объектами.

4. Множественное импутирование (Multiple imputation). Подразумевает приписывание каждому пропуску нескольких потенциальных значений с целью отразить степень неопределенности [11]. Метод множественного импутирования является достаточно перспективным, так как лежит в основе заполнения пропущенных значений с использованием специализированного программного обеспечения, например IBM SPSS Statistics.

5. Метод Resampling. Применяется для заполнения пропусков в неполных данных, когда значения выбираются случайным образом из исходного множества данных. Затем строится регрессионное уравнение для предсказания отсутствующего зна-

чения. Процедура построения регрессионного моделирования повторяется несколько раз. После определенного количества повторений значения полученных регрессионных коэффициентов усредняют и получают окончательное решение, дающее максимальную точность прогноза пропущенного значения [9].

6. Регрессионное моделирование пропусков. На первом этапе по совокупности полных наблюдений строится регрессионная модель и оцениваются коэффициенты в уравнении, где в качестве зависимой переменной выступает целевая переменная, пропущенные значения по которой необходимо восстановить. На втором этапе по полученному уравнению, в которое подставляются известные значения независимых переменных предикторов, для каждого целевого объекта рассчитывается отсутствующее значение по зависимой целевой переменной. В случае интервальных переменных определяется конкретное значение, а для порядковых и номинальных переменных с некоторой вероятностью предсказывается категория, к которой должен быть отнесен объект [12].

Кроме вышеперечисленных методов в арсенале специалистов по статистике имеются и более сложные методы:

1. Метод максимизации ожиданий (Expectation maximization, или EM,— оценивание), основанный на определении функции условного математического ожидания логарифма полной функции правдоподобия при известном значении целевой переменной [13].

2. Метод Бартлетта, основанный на методе наименьших квадратов. На первом этапе происходит подстановка вместо пропусков начальных значений. На втором этапе проводится ковариационный анализ целевой переменной и дихотомического индикатора полноты наблюдения по целевой переменной [9].

3. Алгоритм Zet, заключающийся в подборе для каждого пропуска импутируемого значения из некоторой части совокупности полных наблюдений, называемой компонентной матрицей. Она состоит из компонентных строк и столбцов. Компонентность некоторой строки или объекта представляет собой величину, обратно пропорциональную декартовому расстоянию до целевой строки (неполного наблюдения с пропуском) в пространстве, оси которого заданы переменными — рассматриваемыми характеристиками объектов [14].

Таким образом, импутирование позволяет сохранить полученную информацию с использованием различных по сложности и по точности методов. Однако, если исследователем принято решение о заполнении пропусков, следует отдать предпочтение таким методам, как регрессионное моделирование пропусков, Resampling, множественное импутирование или методам подбора, основанным на расстояниях между объектами, а не простому заполнению пропусков средними значениями, так как кроме яв-

ных преимуществ импутирование имеет ряд недостатков:

использование для предсказания пропусков имеющихся полных данных искажает структуру результирующих данных (после импутирования), которая смещается в сторону структуры только полных наблюдений;

искусственная подстановка значений вносит в массив определенную долю искусственных данных, которые в свою очередь приводят к смещению значимости получаемых на их основе результатов [15].

**Выборка типичных случаев** — та, в которую отбираются единицы генеральной совокупности, обладающие средним (типичным) значением исследуемого признака. При этом возникает проблема выбора признака и определения его типичного значения: исследователь с целью обеспечения репрезентативности опирается на свои субъективные суждения.

Данный метод, основанный на индивидуальной оценке, не позволяет объективно оценить точность результатов исследования. Кроме того, полученные результаты нельзя статистически распределить на генеральную совокупность ввиду возникновения существенных систематических ошибок репрезентативности. Поэтому выборка типичных случаев применяется в медико-биологических исследованиях крайне редко. Предпочтение обычно отдается квотной выборке. Единственной целью использования выборки типичных случаев является иллюстрирование, иными словами, качественное описание объектов, о которых уже имеется достаточная информация.

**Метод «снежного кома»** обычно применяется для отбора экспертов, особенно по узкой проблеме. Данный метод в медико-биологических исследованиях практически не используется, поэтому подробно останавливаться на нем не будем. Суть метода состоит в том, что только первый шаг отбора совершается самим исследователем, а все остальные шаги осуществляются по рекомендациям отобранных на первом шаге экспертов (эксперты используют знания в своей узкой профессиональной области). Отбор заканчивается тогда, когда набрано достаточное для исследования количество экспертов или когда кандидатуры начинают повторяться. Метод получил свое название в связи с тем, что выборка шаг за шагом разрастается вширь, подобно снежному кому, катящемуся с горы.

#### КЛАССИФИКАЦИЯ ВЫБОРОК ПО ЭТАПАМ ОТБОРА

Данной классификации уже уделялось внимание при описании серийной выборки. По этапам отбора выборки подразделяются на одноступенчатые и многоступенчатые.

**Одноступенчатая выборка.** Ее особенность состоит в том, что после определения кластера или страты изучению подвергается каждая единица выделенной группы.

**Многоступенчатая выборка.** В ней, в отличие

от одноступенчатой выборки, изучению подвергаются не все единицы выделенных групп, а происходит последующий отбор отдельных единиц. Многоступенчатый отбор обычно используется в крупномасштабных исследованиях, где выборка формируется последовательно на двух и более иерархических уровнях.

### КЛАССИФИКАЦИЯ ВЫБОРОК ПО ОБЪЕМУ

Условно по объему элементов выборки делят на три типа: малые, средние и большие.

К малым относят выборки, объем которых не превышает 30 ед. ( $n \leq 30$ ).

Средняя выборка удовлетворяет условию  $30 < n \leq 200$  единиц.

Понятие большой выборки до конца не определено, но считается, что выборка является большой по объему, если количество ее элементов превышает 200 ( $n > 200$ ).

Особое внимание следует уделить описанию **малых выборок**, объем которых не превышает 30 ед. и может достигать до 4–5 элементов. Такие малые выборки в медико-биологических исследованиях встречаются довольно часто. Причины использования малых выборок может быть несколько: исследование пациентов с редким заболеванием или обладающим редким признаком; исследование очень дорогостоящее.

По возможности при планировании исследования следует исключать использование малых выборок. Но если этого не избежать, то результаты будут правильным представить в виде медианы и межквартильного интервала [1], которые легко рассчитываются с использованием статистических программ. Однако описание малых выборок не ограничено этими двумя описательными статистиками. Далее будет рассмотрен **метод описания малых выборок**, предложенный У. Госсетом (более известен научному сообществу по псевдониму Стьюдент).

При описании выборок от 10 до 30 ед. действуют следующие правила. Если выборка имеет малый объем, то заявление о том, что среднее значение признака в выборке совпадает со средним значением всей генеральной совокупности, будет некорректным. В данной ситуации гораздо правильнее использовать диапазон средних значений генеральной совокупности (доверительный интервал), в который с заданной вероятностью попадают средние значения выборки. То есть при описании малой выборки наиболее предпочтительными будут интервальные оценки, а не точечные. При этом нормальное распределение заменяется на распределение Стьюдента, которое имеет зависимость от степеней свободы ( $n - 1$ ).

Средняя ошибка малой выборки ( $\mu$  малой выборки) определяется по формуле:

$$\mu_{\text{малой выборки}} = \sqrt{\frac{\sigma^2}{n}}$$

При расчете дисперсии следует вычислять количество степеней свободы, как  $n - 1$ :

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

Предельная ошибка малой выборки ( $\Delta$  малой выборки) определяется по формуле:

$$\Delta_{\text{малой выборки}} = t \cdot \mu_{\text{малой выборки}}$$

Величина  $t$  вычисляется по таблице распределения Стьюдента (табл. 4) при числе степеней свободы  $n - 1$ .

Важно, что корректировка на число степеней свободы, равное  $n - 1$ , производится однократно при нахождении дисперсии или средней ошибки малой выборки. Кроме того, поправка  $n - 1$  необходима лишь при малых по численности выборках ( $n < 30$ ); при  $n > 30$  данная корректировка становится несущественна, приближаясь к нулю.

В качестве примера рассмотрим доверительный интервал, в котором находятся значения некоторой переменной, определенной в научном исследовании. Допустим, получены следующие результаты: 35, 33, 38, 34, 35, 36, 37, 35, 34, 33 ( $n = 10$ ).

Рассчитаем среднее значение:

$$\bar{x} = \frac{\sum x}{n} = \frac{350}{10} = 35$$

Поправку на  $n - 1$  произведем в формуле для расчета дисперсии малой выборки:

$$\begin{aligned} \sigma^2 &= \frac{\sum (x - \bar{x})^2}{n - 1} = \\ &= \frac{0^2 + 2^2 + 3^2 + 1^2 + 0^2 + 1^2 + 2^2 + 0^2 + 1^2 + 2^2}{9} = \\ &= \frac{4 + 9 + 1 + 1 + 4 + 1 + 4}{9} = \frac{24}{9} = 2,67 \end{aligned}$$

Рассчитаем среднюю ошибку малой выборки:

$$\mu_{\text{малой выборки}} = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{2,67}{10}} = 0,85$$

Определим предельную ошибку малой выборки (значение  $t$  находим по таблице распределения Стьюдента для уровня значимости 0,05):

$$\Delta_{\text{малой выборки}} = t \cdot \mu_{\text{малой выборки}} = 2,3 \cdot 0,85 = 1,95$$

Доверительный интервал (95% ДИ) будет иметь следующий вид:

$$33,05 (35 - 1,95) \leq \bar{x} \leq 36,95 (35 + 1,95).$$

Таким образом, с вероятностью 0,95 можно говорить, что среднее значение изучаемого признака находится в пределах от 33,05 до 36,95. Корректное представление в этом случае имеет вид: [95% ДИ 33,05–36,95].



Если объем выборки меньше 10 ед., то при ее описании следует указать весь перечень полученных значений, не прибегая к традиционным методам математической статистики, которые в данном случае являются неприменимыми.

Обращаем внимание, что стандартная ошибка среднего при описании данных может использоваться только в случае расчета доверительных интервалов (интервальная оценка). В случае представления данных в виде среднего значения (точечная оценка) в качестве показателя разброса может указываться только стандартное отклонение ( $\bar{x} \pm \sigma$ ) (см. предыдущую публикацию №2 (63)).

В качестве обоснования вышеизложенного правила приведем пример, представленный в книге С. Гланца [7].

«Рассмотрим пример, позволяющий почувствовать различие между стандартным отклонением и стандартной ошибкой среднего, а также уяснить, почему не следует пренебрегать стандартным отклонением. Положим, исследователь, обследовав выборку из 20 человек, пишет в статье, что средний сердечный выброс составлял 5,0 л/мин со стандартным отклонением 1 л/мин. Мы знаем, что 95% нормально распределенной совокупности попадает в интервал «среднее плюс-минус два стандартных отклонения». Тем самым, из статьи видно, что почти у всех обследованных сердечный индекс составил от 3 до 7 л/мин. Такие сведения весьма полезны, их легко использовать во врачебной практике.

Увы, приведенный пример далек от реальности. Скорее автор укажет не стандартное отклонение, а стандартную ошибку среднего. Тогда из статьи вы узнаете, что «сердечный выброс составил  $5,0 \pm 0,22$  л/мин». И если бы мы спутали стандартную ошибку среднего со стандартным отклонением, то пребывали бы в уверенности, что 95% совокупности заключено в интервал от 4,56 до 5,44 л/мин. На самом деле в этом интервале (с вероятностью 95%) находится среднее значение сердечного выброса».

#### УРОВЕНЬ ЗНАЧИМОСТИ

Далее остановимся еще на одном важном понятии — **уровне значимости** ( $p$ ). Понимание смысла  $p$  требует понимания логики проверки статистической гипотезы [7]. Определение уровня значимости можно сформулировать несколькими способами.

$P$  показывает степень статистической значимости выявленных различий между выборками, т.е. вероятность ошибочного заключения о существовании различий, которых в действительности нет.

$P$  — это вероятность того, что значение критерия окажется не меньше критического значения при условии справедливости нулевой гипотезы об отсутствии различий между группами.

$P$  — это вероятность справедливости нулевой гипотезы.

Часто можно услышать, что  $p$  — это вероятность ошибки. Однако такое определение несколько неточно, так как ошибки, связанные с неверным при-

Таблица 4  
Значения коэффициента  $t$  распределения Стьюдента

| n  | Уровень значимости ( $p$ ) |      |
|----|----------------------------|------|
|    | 0,05                       | 0,01 |
| 2  | 12,7                       | 63,7 |
| 3  | 4,3                        | 9,9  |
| 4  | 3,2                        | 5,8  |
| 5  | 2,8                        | 4,6  |
| 6  | 2,6                        | 4,0  |
| 7  | 2,4                        | 3,7  |
| 8  | 2,4                        | 3,6  |
| 9  | 2,3                        | 3,4  |
| 10 | 2,3                        | 3,3  |
| 11 | 2,2                        | 3,2  |
| 12 | 2,2                        | 3,1  |
| 13 | 2,2                        | 3,1  |
| 14 | 2,2                        | 3,0  |
| 15 | 2,1                        | 3,0  |
| 16 | 2,1                        | 2,9  |
| 17 | 2,1                        | 2,9  |
| 18 | 2,1                        | 2,9  |
| 19 | 2,1                        | 2,9  |
| 20 | 2,1                        | 2,9  |
| 21 | 2,1                        | 2,8  |
| 22 | 2,1                        | 2,8  |
| 23 | 2,1                        | 2,8  |
| 24 | 2,1                        | 2,8  |
| 25 | 2,1                        | 2,8  |
| 26 | 2,1                        | 2,8  |
| 27 | 2,1                        | 2,8  |
| 28 | 2,0                        | 2,8  |
| 29 | 2,0                        | 2,8  |
| 30 | 2,0                        | 2,8  |

нятием и отвержением нулевой гипотезы, подразделяются на два рода.

**Ошибка I рода** — это ошибочное отвержение исследователем нулевой гипотезы, иными словами, ошибочное заключение о существовании различий, которых в действительности нет. Вероятность именно этой ошибки оценивает  $p$ . Максимальная приемлемая вероятность ошибки I рода называется уровнем значимости [7]. В медико-биологических исследованиях критическое значение  $p \leq 0,05$ , т.е. 5%. Однако исследователем может быть принят и меньший уровень значимости, например, 0,01 или 0,001 (во втором случае различия будут считаться сверхзначимыми).

Таблица 5

Ошибки критериев

|   | Различия есть в действительности                       | Различий нет в действительности                   |
|---|--|---|
| Различия выявлены с помощью критерия    | Истинно положительный результат, $1 - \beta$           | Ложноположительный результат (ошибка I рода), $p$ |
| Различия не выявлены с помощью критерия | Ложноотрицательный результат (ошибка II рода), $\beta$ | Истинно отрицательный результат, $1 - p$          |



Рис. 3. Графическое отображение правила трех сигм

Возможна и противоположная ошибка — принять неверную нулевую гипотезу, т.е. не найти действительно существующее различие там, где оно есть. Это так называемая **ошибка II рода**. О вероятности этой ошибки  $p$  ничего не говорит. Ее вероятность обозначается буквой  $\beta$ . Именно  $\beta$  характеризует **мощность или чувствительность** критерия, которая равна  $1 - \beta$  (табл. 5).

**Мощность** — это вероятность (обычно выраженная в процентах) отклонить нулевую гипотезу, когда она действительно ложна, иными словами, это вероятность обнаружить реальные статистически значимые различия. Считается, что мощность должна быть не менее 70–80%.

Существует ряд факторов, которые непосредственно влияют на мощность критерия:

объем выборки — чем больше объем выборки, тем меньше  $\beta$ , и соответственно, выше мощность;

уровень значимости — чем меньше  $p$ , тем ниже мощность;

вариабельность наблюдений — чем больше отношение величины различий к стандартному отклонению, тем выше мощность.

Кроме уровня значимости, большое значение имеет доверительная вероятность, которая неразрывно связана с доверительным интервалом. Напомним, что **доверительный интервал** ( $\bar{x} - \Delta x; \bar{x} + \Delta x$ ) называют интервал, который с заданной вероятностью содержит истинное значение искомой величины. **Доверительная вероятность** — это вероят-

ность, что истинное значение измеряемой величины содержится внутри заданного доверительного интервала ( $\bar{x} - \Delta x; \bar{x} + \Delta x$ ). Доверительную вероятность можно выражать в процентах, в этом случае она будет называться надежностью, либо в абсолютных значениях. В медико-биологических исследованиях как минимально допустимый принят 95% доверительный интервал и, соответственно, минимально допустимое значение доверительной вероятности, равное 0,95.

Сущность описанных выше понятий может быть объяснена с помощью **правила трех сигм**.

Правило трех сигм гласит: если величина распределена нормально, то вероятность того, что случайная величина отклонится от своего математического ожидания более чем на три стандартных отклонения, практически равна нулю.

На рисунке 3 видно, что в пределах одного стандартного отклонения лежит 68,3% значений, принимаемых нормально распределенной случайной величиной. В пределах двух стандартных отклонений — уже 95,4%, а в пределах трех — 99,7%. Это означает, что вероятность того, что случайная величина примет значение, отклоняющееся от математического ожидания больше чем на три среднеквадратических отклонения, не превышает 0,28%, т.е. пренебрежимо мала.

Математически эти факты можно представить в виде следующих равенств:

$$P = (-\sigma < \bar{x} < +\sigma) = 0,683$$

$$P = (-2\sigma < \bar{x} < +2\sigma) = 0,954$$

$$P = (-3\sigma < \bar{x} < +3\sigma) = 0,997$$

Эти числовые значения (68, 95, 99,7%) получены из кумулятивной функции нормального распределения.

В медико-биологических исследованиях результат может считаться статистически значимым, если его доверительная вероятность имеет порядок двух сигм (95%).

**ЗАКЛЮЧЕНИЕ**

Таким образом, владение знаниями об основных понятиях статистики поможет подойти к рассмотрению основ статистического анализа результатов медико-биологических исследований. Следующая статья будет посвящена правилам использования параметрических и непараметрических критериев.

**Финансирование исследования и конфликт интересов.** Исследование не финансировалось каким-либо источником, и конфликты интересов, связанные с данным исследованием, отсутствуют.

**ЛИТЕРАТУРА/REFERENCES**

1. Баврина А. П. Современные правила использования методов описательной статистики в медико-биологических исследованиях. *Медицинский альманах* 2020; 2(63): 95–104. Bavrina A. P. Modern

rules for the use of descriptive statistics methods in biomedical research. *Medicinskij al'manah* 2020; 2(63): 95–104.

2. Реброва О.Ю. *Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA*. М: МедиаСфера; 2000. Rebrova O. Ju. *Statisticheskij analiz medicinskih dannyh. Primenenie paketa prikladnyh programm STATISTICA* [Statistical analysis of medical data. Using the STATISTICA application package]. Moscow: MediaSfera; 2000.

3. Наркевич А.Н., Виноградов К.А. Методы определения минимально необходимого объема выборки в медицинских исследованиях. *Социальные аспекты здоровья населения [сетевое издание]* 2019; 65(6): 10. URL: <http://vestnik.mednet.ru/content/view/1123/30/lang,ru/>, <https://doi.org/10.21045/2071-5021-2019-65-6-10>. Narkevich A.N., Vinogradov K.A. Methods for determining the minimum required sample size in medical research. *Social'nye aspekty zdorov'ya naseleniya [setevoe izdanie]* 2019; 65(6): 10. URL: <http://vestnik.mednet.ru/content/view/1123/30/lang,ru/>, <https://doi.org/10.21045/2071-5021-2019-65-6-10>.

4. Fox N., Hunn A., Mathers N. *Sampling and sample size calculation*. Yorkshire & the Humber: The NIHR RDS for the East Midlands; 2009.

5. Отдельнова К.А. Определение необходимого числа наблюдений в социально-гигиенических исследованиях. *Сб. трудов 2-го ММИ* 1980; 150(6): 18–22. Otdel'nova K.A. Determination of the required number of observations in social and hygienic research. *Sb. trudov 2-go MMI* 1980; 150(6): 18–22.

6. Паниотто В.И., Максименко В.С. *Количественные методы в социологических исследованиях: коллективная монография*. Киев: Наук. Думка; 1982. Paniotto V.I., Maksimenko V.S. *Kolichestvennye metody v sociologicheskikh issledovaniyah: kollektivnaya monografiya* [Quantitative methods in sociological research: collective monograph]. Kiev: Nauk. Dumka; 1982.

7. Гланц С. *Медико-биологическая статистика*. М: Практика; 1998. Glanc S. *Mediko-biologicheskaja statistika* [Biomedical statistics]. Moscow: Praktika; 1998.

8. Горшков М.К., Шереги Ф.Э. *Прикладная социология. Методология и методы*. М; 2012. Gorshkov M.K., Sheregi F.E. *Prikladnaya sociologiya. Metodologiya i metody* [Applied sociology. Methodology and methods]. Moscow; 2012.

9. Злоба Е., Яцкив И. Статистические методы восстановления

пропущенных данных. *Computer Modeling & New Technologies* 2004; 6(1): 51–61. Zloba E., Yackiv I. Statisticheskie metody voss-tanovleniya propuschennyh dannyh. *Computer Modeling & New Technologies* 2004; 6(1): 51–61.

10. Kalton G., Kasprzyk D. The treatment of missing survey data. *Survey Methodology* 1986; 12: 1–16.

11. Rubin D. B. *Multiple imputation for nonresponse in surveys*. New York: Willey; 1987.

12. Крыштановский А.О. *Анализ социологических данных с помощью пакета SPSS*. М: ГУВШЭ; 2006. Kryshchanovskij A. O. *Analiz sociologicheskikh dannyh s pomoschyu paketa SPSS* [Analysis of sociological data using the SPSS package]. Moscow: GUVShE; 2006.

13. Королев В.Ю. EM-алгоритм, его модификации и их применение к задаче разделения смесей вероятностных распределений. *Теоретический обзор*. М; 2007. Korolev V. Yu. *EM-algoritm, ego modifikacii i ih primenenie k zadache razdeleniya smesej veroyatnostnyh raspredelenij. Teoreticheskij obzor* [EM-algorithm, its modifications and their application to the problem of separation of mixtures of probability distributions. Theoretical review]. Moscow; 2007.

14. Снитюк В.Е. *Эволюционный метод восстановления пропусков в данных*. Сборник трудов VI-й Межд. конф. «Интеллектуальный анализ информации». Киев; 2006, с. 262–271. Snityuk V.E. *Evolucionnyj metod vosstanovleniya propuskov v dannyh*. Sbornik trudov VI-j Mezhd. konf. "Intellektualnyj analiz informacii" [An evolutionary method for recovering data gaps. Collection of works of VI<sup>th</sup> international conference "Intelligent analysis of information"]. Kiev; 2006, p. 262–271.

15. Rubin D. B. Multiple imputation after 18+ years. *J Am Stat Assoc* 1996; 91: 473–489.

#### ИНФОРМАЦИЯ ОБ АВТОРЕ:

**А. П. Баврина**, к. б. н., доцент кафедры медицинской физики и информатики, руководитель Центра биомедицинской статистики, организации исследований и цифровой медицины ФГБОУ ВО «Приволжский исследовательский медицинский университет» Минздрава России.

**Для контактов:** Баврина Анна Петровна,  
e-mail: annabavr@gmail.com