

Первичная обработка выборочных данных

8.1. Основы теории

Научное исследование состоит из нескольких этапов.

Первый — **обнаружение проблемы**, либо какого-либо неисследованного объекта или явления.

Второй этап — **планирование наблюдения или эксперимента**, сбора данных для дальнейшей обработки.

Третий этап — **сбор данных**, за которым следует четвёртый — **обработка данных**. Заметим, что методы обработки данных необходимо подобрать на этапе планирования эксперимента. Они естественным образом следуют из природы собираемых данных. После обработки данных следует пятый этап — **формулирование выводов**.

Рассмотрим более пристально второй этап, на котором необходимо определить, что именно (какую величину или величины) надо измерять, очертить множество объектов, которые надо будет измерить (подобрать так называемую генеральную совокупность исследуемых объектов), отобрать для исследования часть этого множества для непосредственного измерения (сделать выборку). Приведем здесь теоретические моменты, связанные с этим этапом.

Генеральная совокупность — множество тех элементов, которые нужно обследовать, и относительно которых необходимо сделать выводы (относительно распределения исследуемого параметра).

Выборочная совокупность или **выборка** — множество элементов, которые непосредственно проходят измерение.

Замечание 4. В математической статистике понятие генеральной совокупности не применяется, а термин «генеральная совокупность» используется как образ, как метафора, и фактически генеральная совокупность — это множество значений случайной величины X , связанной со случайным испытанием.

Выборкой объёма n в математической статистике называют последовательность $X_1, X_2, \dots, X_n$ попарно и в совокупности независимых и одинаково распределённых с X случайных величин.

Множество исследуемых объектов выступает как генератор случайных чисел, имеющий распределение величины X . В результате исследования мы получаем набор чисел — $X_1, X_2, \dots, X_n$, называемый реализацией выборки. Именно на его основе исследователю придется делать выводы о распределении случайной величины X и его характеристиках. В дальнейшем мы будем говорить о выборке, как о реализации выборки.

Основа выборки — список всех элементов генеральной совокупности. С помощью основы выборки и генератора равномерно

распределённых случайных чисел можно получить так называемую простую случайную выборку.

Выборку объема n будем называть **простой случайной выборкой**, если каждый элемент имеет одну и ту же вероятность быть выбранным в неё и каждый набор из n элементов имеет равную вероятность быть сформированным по сравнению с другими наборами такой же длины.

Если при формировании выборки каждый объект генеральной совокупности может быть выбран несколько раз, то такая выборка называется **выбором с возвращением** (выбранный элемент как бы возвращается в генеральную совокупность и поэтому получает возможность быть выбранным снова).

Если при формировании выборки каждый объект генеральной совокупности может быть выбран только один раз, то такая выборка называется **выбором без возвращения**. Заметим, что в этом случае элементы выборки становятся статистически зависимыми. Однако в случае, когда объём выборки существенно мал, по сравнению с объёмом генеральной совокупности, этой зависимостью можно пренебречь.

Выборка должна быть такова, чтобы на её основе были сделаны правильные выводы о распределении изучаемого параметра. Это свойство выборки называется репрезентативностью.

Репрезентативность — это свойство выборки воспроизводить в себе основные характеристики генеральной совокупности. Данное свойство позволяет экстраполировать результаты исследования, полученные на ограниченном числе наблюдений, на всю генеральную совокупность, т.е. делать заключения о целом на основании части [2]. Сформировать нерепрезентативную выборку, а затем провести её статистический анализ — главная опасность выборочного метода.

Нерепрезентативные выборки называют **смещенными**.

Кроме указанной опасности у исследователя есть опасность неправильного определения генеральной совокупности, а именно, получить **неоднородную** генеральную совокупность, то есть такую генеральную совокупность, в которой имеются **несколько распределений**. Иными словами, она состоит из нескольких подмножеств, каждое из которых имеет свое распределение. В дальнейших рассуждениях и при решении задач мы будем предполагать, что исследуемая генеральная совокупность однородна. Далее нам надо затронуть вопросы, связанные с процессом измерения исследуемых параметров.

Измерение — это приписывание числовых форм объектам или

событиям в соответствии с определенными правилами [3].

Измерение — это процедура, с помощью которой измеряемый объект сравнивается с некоторым эталоном и получает числовое выражение в определенном масштабе или шкале [5]

Шкала — это множество элементов (числовых или нечисловых), которым ставятся в соответствие измеряемые объекты.

Существует четыре вида шкал:

1. Шкала наименований (или номинальная).

Шкальные значения зачастую представляют собой обычные слова, названия, имена (категории). Например, такими являются ответы {да, нет, не знаю}, результаты голосования {за, против, воздержался}. В шкале наименований значения нельзя упорядочить, нельзя сказать, какое значение больше, а какое меньше. Следовательно, в данных шкалах **порядок не определён**. Кроме того, нельзя ответить на вопросы, во сколько раз одно значение больше или меньше другого, или на сколько шкальных единиц одно значение больше или меньше другого. Отсюда следует, что в номинальных шкалах **не определены никакие арифметические операции**.

Методов обработки номинальных данных немного. Так, мы можем подсчитать, сколько раз в результате эксперимента появлялось то или иное значение (вычислить частоту значения), вычислить долевое (процентное) соотношение значений. Также можно, при измерении нескольких групп объектов, сравнивать частоту появления тех или иных значений, например, с помощью так называемого критерия Пирсона.

2. Шкала порядка. Если значения номинальной шкалы каким-либо образом упорядочить, присвоить значениям **ранги**, то мы получим шкалу порядка. Например, к такому виду шкал относятся упорядоченные по алфавиту фамилии студентов, всевозможные рейтинги (баллы ЕГЭ и ОГЭ, рейтинги спортсменов). Следовательно, в данных шкалах **определён порядок**. Однако арифметические операции **не определены**, несмотря на то, что шкальные значения, ранги, могут быть записаны натуральными числами.

Методы обработки порядковых данных аналогичны методам обработки номинальных данных.

3. Шкалы разностей.

В шкалах разностей, кроме **наличия порядка**, **осмыслена операция вычитания** и можно ответить на вопрос, **на сколько** единиц шкалы одно шкальное значение отличается от другого. Именно поэтому такого рода шкалы именуются шкалами разностей. Также можно к одному шкальному значению прибавить некоторое количество

единиц шкалы (не обязательно целое) и получить другое шкальное значение (**осмыслена операция сложения**).

К шкалам разностей относятся: шкалы температур по Цельсию, Фаренгейту, Реомюру; временная шкала, оперирующая понятиями «до новой эры», «новой эры».

Значения, измеренные в шкале разностей, могут быть изображены с помощью вещественной числовой прямой. Наличие порядка и арифметических операций сложения и вычитания позволяет сказать, что множества значений изучаемых в теории вероятностей случайных величин являются множествами значений шкал разностей.

Именно в таких шкалах можно вычислять различного рода средние величины. Отсюда следует, что для наблюдений, измеренных в шкалах разностей применимы практически все методы статистических исследований: построение статистических распределений, гистограмм, функций распределения, сравнение распределений, регрессионно-корреляционный анализ. Исключение составляет вычисление среднего геометрического и коэффициента вариации.

4. Шкалы отношений.

В шкалах отношений, кроме **наличия порядка**, и **осмысленных операций сложения и вычитания** также можно ответить на вопрос, **во сколько раз** одно шкальное значение отличается от другого, **осмыслена операция деления** (и обратная к ней — умножение шкального значения на вещественное число и получение другого шкального значения). Именно поэтому такого рода шкалы именуются шкалами отношений. Шкалы отношений имеют точку абсолютного нуля, в отличие от шкал разностей, в которых положение нуля, вообще говоря, выбирается произвольно.

К шкалам разностей относятся: шкала температур по Кельвину; шкала измерения длительности процессов; шкалы длины, площади, объёма; шкала массы.

К данным, измеренным в шкалах отношений, применимы любые процедуры статистических исследований.

Исследователь должен понимать, в какой шкале проводится измерение изучаемых параметров, и, следовательно, соответственным образом подбирать методы статистической обработки собранных данных.

Далее затронем задачу **систематизации и табличного представления** собранных данных.

После того, как данные собраны, их нужно поместить в таблицу.

Таблица должна создаваться по типу базы данных, а именно: каждый столбец содержит однотипные данные, а каждая строка содержит информацию об одном измеренном объекте.

Перечислим несколько случаев, отличающихся количеством групп

объектов и количеством измеряемых у одного объекта параметров. Для каждого случая приведем внешний вид таблицы, в которой собираются данные.

1. Одна группа, один измеряемый параметр. Таблица должна содержать ключевой столбец, в котором будут порядковые номера измеренных объектов (например, школьников), может (но не обязана) содержать столбец с наименованиями измеренных объектов (например, фамилиями и инициалами школьников), и столбец, в котором записаны значения исследуемого параметра (например, рост школьника в см).

№	Ф.И.О.	Рост
1	Иванов И.И.	167
2	Петров П.П.	155
3	Сидоров С.С.	170

2. Одна группа, несколько измеряемых параметров. Таблица должна содержать ключевой столбец, в котором будут порядковые номера измеренных объектов (например, школьников), может (но не обязана) содержать столбец с наименованиями измеренных объектов (например, фамилиями и инициалами школьников), и столбцы, в которых записаны значения исследуемых параметров (например, рост школьника в см, вес в кг и цвет волос, выраженный словом).

№	Ф.И.О.	Рост	Масса	Цвет
1	Иванов И.И.	167	62.53	Блондин
2	Петров П.П.	155	41.20	Брюнет
3	Сидоров С.С.	170	68.50	Шатен

145

3. Несколько групп, один измеряемый параметр.

Таблица должна содержать ключевой столбец, в котором будут порядковые номера измеренных объектов (например, школьников), столбец, по которому ведется разбиение испытуемых по группам и который содержит номера групп (классов, например), может (но не обязана) содержать столбец с наименованиями измеренных объектов (например, фамилиями и инициалами школьников), и столбец, в котором записаны значения исследуемого параметра (например, рост школьника в см).

№	Класс	Ф.И.О.	Рост
---	-------	--------	------

1	7А	Иванов И.И.	167
2	6Б	Петров П.П.	155
3	7А	Сидоров	170
2	6Б	С.С. Попова П.П.	149

4. Несколько групп, несколько измеряемых параметров.

Таблица должна содержать ключевой столбец, в котором будут порядковые номера измеренных объектов (например, школьников), столбец с номерами групп (классов, например), может (но не обязана) содержать столбец с наименованиями измеренных объектов (например, фамилиями и инициалами школьников), и столбцы, в которых записаны значения исследуемых параметров (например, рост школьника в см, вес в кг и цвет волос, выраженный словом).

№	Клас с	Ф.И.О.	Рост	Масса	Цвет
				62.53 0	
1	7А	Иванов И.И.	167	41.20	Блондин
2	6Б	Петров П.П.	155	0	Брюнет
3	6Б	Сидоров С.С.	170	68.50	Шатен
4	7А	Попова П.П.	161	0	Рыжий
				52.35 0	

Напоследок осветим вопросы, связанные с первичной обработкой данных, иначе называемой **разведывательным анализом** или **описательной статистикой**.

Задачей описательной статистики является получение так называемого выборочного статистического распределения, нахождение выборочной функции распределения, вычисление числовых характеристик выборочного распределения, а также визуализация данных.

Описательная статистика призвана натолкнуть исследователя на предположения о виде распределения изучаемой величины (генерального распределения). К примеру, только взглянув на выборочные значения, можно сделать предположение о типе распределения: если значения практически не повторяются, то это может говорить о том, что распределение непрерывно; если же есть часто повторяющиеся значения, то это может служить в пользу дискретности распределения.

Перечислим основные выборочные характеристики, разыскиваемые в процессе первичного анализа выборочных данных.

Объём выборки — количество элементов в выборке. Обозначим его буквой n .

Частота того или иного выборочного значения — количество повторений значения в выборке. Обозначение: p_i — частота значения x_i . Если всего k различных выборочных значений, то i находится в пределах от 1 до k .

Относительная частота того или иного выборочного значения — отношение частоты этого значения к объёму выборки. Обозначение: $p_i = p_i/n$ — относительная частота значения x_i .

Выборочная функция распределения определяется так:

для любого вещественного x

$F_{\text{выб}} = \text{сумма частот выборочных значений, меньших } x / \text{объём выборки}$

Среднее выборочное — выборочный аналог математического ожидания, среднее арифметическое значений выборки, вычисляется по формуле $x_v = \sum_1^k x_i p_i$, если в выборке k различных значений.

В случае выборки, полученной из непрерывного распределения, возможна операция **группирования выборки**. Для этого, промежуток значений выборки от минимального до максимального надо разделить на несколько (равных) частей (**частичных интервалов**) и вычислить количество попаданий выборочных значений в тот или иной частичный интервал. При группировании выборки мы от непрерывного распределения переходим к дискретному, в котором выборочными значениями являются середины интервалов m_i , а частотами p_i являются вычисленные количества попаданий выборочных значений в тот или иной интервал.

Соответственно вычисляются относительные частоты p_i . Тогда среднее выборочное для группированной выборки:

$x_v = \sum_1^k m_i p_i$, Здесь k — количество частичных интервалов.

Смещённая выборочная дисперсия — аналог дисперсии, характеризует квадрат разброса значений выборки около среднего выборочного и вычисляется по формулам:

$$D_{\text{выб.}} = \sum_1^k (x - x_{\text{выб.}})^2$$

Аналогична формула для группированной выборки, только x_i заменяется на m_i .

Несмещённая выборочная дисперсия вычисляется по формуле:

$$s^2 = \frac{n-1}{n} D_{\text{выб.}}$$

Выборочное смещённое среднее квадратическое отклонение:

$\sigma_{\text{выб}} = \sqrt{D_{\text{выб}}}$; выборочное несмещённое среднее квадратическое отклонение: $s = \sqrt{s^2}$.

Мода — выборочное значение, которое имеет наибольшую частоту, либо середина интервала, в котором сосредоточена наибольшая частота (последнее в случае группированной выборки).

Вариационный ряд — выборка, упорядоченная по возрастанию:

$$x_1 \leq x_2 \leq \dots \leq x_n$$

Значение x_i называется **i -й порядковой статистикой**.

Кроме того, можно сказать, что если мы имеем порядковую статистику x_i , то она является **выборочным квантилем порядка p_i** .

Отсюда следует, что не для всякого порядка p можно отыскать выборочный квантиль.

Размах выборки — разность между максимальным и минимальным элементами выборки. Обозначение: $w = x_{\max} - x_{\min}$

Выборочная медиана — число m , делящее вариационный ряд пополам. Заметим, что если мода, медиана и среднее выборочное примерно равны, то это может натолкнуть на предположение о том, что генеральное распределение — нормальное.