Нижегородский государственный университет им. Н.И. Лобачевского

Институт информационных технологий, математики и механики

«Наглядный вероятностно-статистический анализ данных»

Лекция 2

Подготовка данных к статистическому анализу

Пройдакова Екатерина Вадимовна,
доцент кафедры ТВиАД ИИТММ

1. ГЕНЕРАЛЬНАЯ И ВЫБОРОЧНАЯ СОВОКУПНОСТИ

1.1. Понятие генеральной совокупности

Напомним, что в рамках данного курса мы рассматриваем экспериментальные данные которые получены в результате **достаточного числа наблюдений** за статистически устойчивым экспериментом $\boldsymbol{E}$ с некоторой неопределенностью в задании вероятностной функции $\mathbf{P}(\bullet)$.

В результате многократного проведения эксперимента $\boldsymbol{E}$ наблюдалось $\boldsymbol{K}$ его исходов из множества $\boldsymbol{\Gamma_0} = \{\boldsymbol{A_1}, \boldsymbol{A_2}, \ldots, \boldsymbol{A_K}\}$, причем среди $\boldsymbol{A_1}, \boldsymbol{A_2}, \ldots, \boldsymbol{A_K}$ **могут быть повторяющиеся**. Полагаем, что **части из этих** $\boldsymbol{K}$ объектов присуще значение некоторой характеристики ξ.

1.1. Понятие генеральной совокупности

Отберем из множества Γ_0 только **все такие элементы** $B_1, B_2, \ldots, B_N$, для которых **можно** путем измерений **определить значение указанной характеристики** ξ.

Совокупность $\Gamma = \{B_1, B_2, \ldots, B_N\}$, $\Gamma \subset \Gamma_0$ назовем **генеральной совокупностью** объема $N \leq K$.

Множества Γ и Γ_0 может быть как конечными, так и бесконечными (в теории).

1.1. Понятие генеральной совокупности

Пример генеральной совокупности. Пусть эксперимент заключается в случайном отборе шарообразной детали, изготавливаемой на некотором заводе и случайном выборе рабочего данного завода. В результате независимого проведения эксперимента **500** раз были отобраны детали $\mathbf{D_1, D_2, \ldots, D_{500}}$ и рабочие $\mathbf{R_1, R_2, \ldots, R_{500}}$.

Здесь $\mathbf{\Gamma_0} = \{\mathbf{D_1, R_1, D_2, R_2, \ldots, D_{500}, R_{500}}\}$, альтернативная запись множества имеет вид $\mathbf{\Gamma_0} = \{\mathbf{A_1, A_2, \ldots, A_{1000}}\}$, где $\mathbf{A_1} = \mathbf{D_1}$, $\mathbf{A_2} = \mathbf{R_1}$, $\mathbf{A_3} = \mathbf{D_2}$, $\mathbf{A_4} = \mathbf{R_2}, \ldots,$ $\mathbf{A_{999}} = \mathbf{D_{500}}$, $\mathbf{A_{1000}} = \mathbf{R_{500}}$.

Если исследуемая характеристика ξ - диаметр детали, то для нее генеральная совокупность имеет вид $\mathbf{\Gamma} = \{\mathbf{D_1, D_2, \ldots, D_{500}}\}$.

Пусть исследуемая характеристика $\xi = (\eta_1, \eta_2)$, где η_1 и η_1 - это возраст и стаж рабочего соответственно, то в этом случае $\mathbf{\Gamma} = \{\mathbf{R_1, R_2, \ldots, R_{500}}\}$.

1.2. Понятие выборочной совокупности

Для изучения интересующих нас характеристик генеральной совокупности Γ вовсе не обязательно изучать каждый ее элемент.

С этой целью формируют **выборочную совокупность** $\Gamma_в \subset \Gamma$. Считаем, что **объем** N генеральной совокупности Γ **велик** по сравнению с **объемом** n выборочной совокупности $\Gamma_в$.

Выборочная совокупность обязательно должна удовлетворять **условию *репрезентативности***, то есть, давать ***адекватное представление обо всей генеральной совокупности.***

Репрезентативности выборки можно **достичь** за счет осуществления **отбора элементов** из генеральной совокупности **случайно и независимым образом**.

1.2. Понятие выборочной совокупности

Задав номера всех N элементов генеральной совокупности Γ, можно **использовать датчик псевдослучайных чисел** для получения $n < N$ последовательных элементов $C_1, C_2, \ldots, C_n$ выборочной совокупности $\Gamma_в = \{C_1, C_2, \ldots, C_n\}$. Тогда выборка $\Gamma_в$ будет обладать **свойством репрезентативности.**

Пусть случайная величина ξ_i определяет количественную характеристику ξ элемента C_i. Тогда $\xi_1, \xi_2, \ldots, \xi_n$ являются независимыми, и одинаково распределенными случайными величинами. Они представляют собой n копий (клонов) случайной величины ξ.

Случайный **вектор** $(\xi_1, \xi_2, \ldots, \xi_n)$ называют **повторной выборкой**.

2. ПРЕДСТАВЛЕНИЕ ВЫБОРОЧНОЙ СОВОКУПНОСТИ

2.1. Способы представления выборочных значений, вариационный ряд

Вектор $(x_1, x_2, \ldots, x_n) \in X^n$ есть реализация повторной выборки $(\xi_1, \xi_2, ..., \xi_n)$. Значения $x_1, x_2, \ldots, x_n$ называют также выборочными значениями для случайных величин $\xi_1, \xi_2, ..., \xi_n$.

Удобно представить величины $x_1, x_2, \ldots, x_n$ в виде неубывающей последовательности $x_1^* \le x_2^* \le \ldots \le x_n^*$. Такая последовательность называется **вариационным рядом.**

Здесь x_1^* - **минимальный член** вариационного ряда,

x_n^* - **максимальный член** вариационного ряда.

2.1. Способы представления выборочных значений, статистический ряд

Если среди выборочных значений x_1, x_2, …, x_n встречаются одинаковые, то используют **статистический ряд:**

y_j	y_1	y_2	y_3	…	y_m
n_j	n_1	n_2	n_3	…	n_m

В статистическом ряде y_1, y_2, …, y_m представляют собой расположенные в порядке возрастания **различные значения выборки** $x_1, x_2, ..., x_n$.

Числа n_1, n_2, …, n_m означают количества выборочных значений, равных соответственно значениям y_1, y_2, …, y_m $(m \le n)$, причем

$$n_1 + n_2 + \ldots + n_m = n.$$

2.1. Способы представления выборочных значений, группированный статистический ряд

При больших объеме $\boldsymbol{n}$ и диапазоне или размахе выборки (от x_1^* до x_n^*) записи вариационного и статистического рядов громоздки.

Чтобы сделать запись компактной и обозримой для визуального анализа рассматривают **группированный статистический ряд (информационную совокупность)**.

При построении группированного статистического ряда весь диапазон выборки разбивают на $\boldsymbol{k}$ промежутков или разрядов:

$$\Pi_j = [a_{j-1}, a_j), \text{ где } j = \overline{1,k}, \quad a_0 \le x_1^*, \quad a_k \ge x_n^*, \quad a_0 < a_1 < a_2 < \ldots < a_k$$

2.1. Способы представления выборочных значений, группированный статистический ряд

Π_j	$[a_0, a_1)$	$[a_1, a_2)$	…	$[a_{j-1}, a_j)$	…	$[a_{k-1}, a_k)$
ν_j	ν_1	ν_2	…	ν_j	…	ν_k
p_j^*	$\frac{\nu_1}{n}$	$\frac{\nu_2}{n}$	…	$\frac{\nu_j}{n}$	…	$\frac{\nu_k}{n}$

Для каждого разряда Π_j находятся следующие величины:

$\boldsymbol{\nu_j}$ - абсолютная частота разряда с номером $\boldsymbol{j}$, равная количеству значений вариационного ряда, попавших в данный разряд;

$\boldsymbol{p_j^* = \nu_j/n}$ - относительная частота попадания наблюдений в $\boldsymbol{j}$-ый разряд. Очевидно $\boldsymbol{\nu_1 + \nu_2 + \ldots + \nu_k = n}$ и $\boldsymbol{p_1^* + p_2^* + \ldots + p_k^* = 1}$.

2.1. Способы представления выборочных значений, группированный статистический ряд

Если разряды $\Pi_j = [a_{j-1}, a_j),\ j = \overline{1,k}$ выбираются **равной ширины**, то их количество k можно определить по **формуле Стёрджеса**:

$$k \simeq \log_2 n + 1 = 1{,}44 \ln n + 1$$

где n - это объем выборки, значение k округляем до целого.

Ширина Δ разрядов Π_j вычисляется по формуле:

$$\Delta \simeq \frac{(x^*{}_n - x^*{}_1)}{(k-1)}.$$

Границы разрядов $\Pi_j,\ \ j = \overline{1,k}$ находятся следующим образом:

$$a_0 = x^*{}_1 - \frac{\Delta}{2}, \quad a_j = a_{j-1} + \Delta, \quad j = \overline{1,k}$$

2.2. Примеры способов представления выборочных значений

Пусть исследуется величина ξ – **масса клетчатки в граммах**, ежедневно потребляемая респондентом. В результате интернет-опроса 30 респондентов в возрасте 18-20 лет были получена следующая выборка.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
100	80	240	230	180	160	150	210	250	230

x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
260	215	270	175	145	190	320	300	390	235

x_{21}	x_{22}	x_{23}	x_{24}	x_{25}	x_{26}	x_{27}	x_{28}	x_{29}	x_{30}
160	195	180	215	220	210	60	120	130	250

2.2. Примеры способов представления выборочных значений

По исходной выборке строим **вариационный ряд:**

x_1^*	x_2^*	x_3^*	x_4^*	x_5^*	x_6^*	x_7^*	x_8^*	x_9^*	x_{10}^*
60	80	100	120	130	145	150	160	160	175
x_{11}^*	x_{12}^*	x_{13}^*	x_{14}^*	x_{15}^*	x_{16}^*	x_{17}^*	x_{18}^*	x_{19}^*	x_{20}^*
180	180	190	190	210	210	215	215	220	230
x_{21}^*	x_{22}^*	x_{23}^*	x_{24}^*	x_{25}^*	x_{26}^*	x_{27}^*	x_{28}^*	x_{29}^*	x_{30}^*
230	235	240	250	250	260	270	300	320	390

Минимальный член вариационного ряда $x_1^* = 60$.

Максимальный член вариационного ряда $x_{30}^* = 390$.

2.2. Примеры способов представления выборочных значений

Среди значений $x_1^*, x_2^*, ..., x_{30}^*$ **есть повторяющиеся**, поэтому далее будем строить **статистический ряд**.

Построенный по вариационному ряду **статистический ряд** содержит **22 различных значения**:

y_j	60	80	100	120	130	145	160	175	180	190	210
n_j	1	1	1	1	1	1	2	1	2	2	2

y_j	215	220	230	235	240	250	260	270	300	320	390
n_j	2	1	2	1	1	2	1	1	1	1	1

2.2. Примеры способов представления выборочных значений

Далее строим группированный статистический ряд (информационную совокупность). Выбираем разряды равной ширины, определяем их количество $\boldsymbol{k}$. В нашем примере объем выборки $\boldsymbol{n} = \mathbf{30}$, следовательно

$$k \simeq 1,44 \ln 30 + 1 \approx 6$$

Вычисляем ширину разряда: $\Delta = \dfrac{(390-60)}{(6-1)} = \dfrac{330}{5} = 66$

Находим границы разрядов $\Pi_j = [a_{j-1}, a_j)$, $j = \overline{1,6}$:

$$a_0 = 60 - \frac{66}{2} = 60 - 33 = 27$$

$$a_1 = a_0 + 66 = 27 + 66 = 93$$

$$a_2 = a_1 + 66 = 93 + 66 = 159$$

$$a_3 = a_2 + 66 = 159 + 66 = 225$$

$$a_4 = a_3 + 66 = 225 + 66 = 291$$

$$a_5 = a_4 + 66 = 291 + 66 = 357$$

$$a_6 = a_5 + 66 = 357 + 66 = 423$$

2.2. Примеры способов представления выборочных значений

Для каждого разряда $\Pi_j = [a_{j-1}, a_j)$, $j = \overline{1,6}$ находим величины $\boldsymbol{\nu_j}$ и $\boldsymbol{p_j^* = \nu_j/n}$. В итоге **группированный статистический ряд** примет следующий вид:

Π_j	$[27, 93)$	$[93, 159)$	$[159, 225)$	$[225, 291)$	$[291, 357)$	$[357, 423)$
ν_j	2	5	12	8	2	1
p_j^*	$\frac{2}{30}$	$\frac{5}{30}$	$\frac{12}{30}$	$\frac{8}{30}$	$\frac{2}{30}$	$\frac{1}{30}$

3. ИЗМЕРИТЕЛЬНЫЕ ШКАЛЫ

3.1. Шкалы измерения значений характеристики эксперимента

При определении значений характеристики ξ эксперимента $\boldsymbol{E}$ используются определенные единицы измерения и соответственно **шкала измерения**. **Тип шкалы**, в которой проведено измерение, влияет на **количество информации**, содержащееся в анализируемой характеристике. На рис. 2.1 представлены типы шкал.

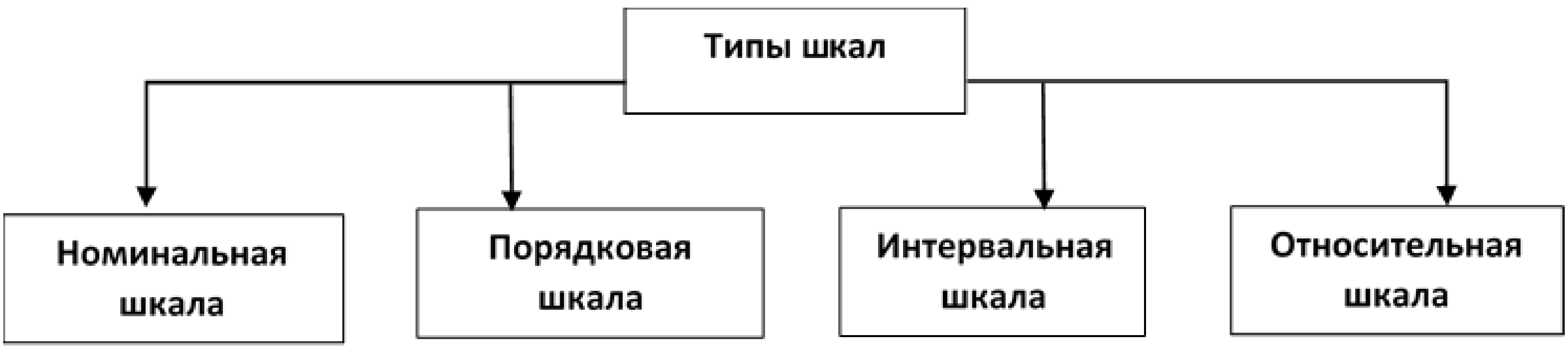


Рис.2.1.

3.1. Шкалы измерения значений характеристики эксперимента

Номинальная шкала - это шкала, классифицирующая по названию. Название же не измеряется количественно, оно лишь позволяет отличить один объект от другого, здесь **числа используются лишь как метки.**

По такой шкале *могут быть измерены*, например, ИНН (индивидуальный номер налогоплательщика), *раса, цвет волос, цвет глаз, и т.д.*

3.1. Шкалы измерения значений характеристики эксперимента

Простейший случай номинальной шкалы - **дихотомическая (логическая)** шкала, **состоящая всего лишь из двух делений.** Признак, который измеряется по дихотомической шкале, называется **альтернативным**.

Такой признак может принимать всего **два значения**, например: «есть признак» - «признак отсутствует»;

«проголосовал ЗА» - «проголосовал ПРОТИВ» и т.п.

3.1. Шкалы измерения значений характеристики эксперимента

Порядковая (ординальная, ранговая) шкала – это шкала, классифицирующая по принципу «больше – меньше».

Если в номинальной шкале было безразлично, в каком порядке расположены объекты, то в порядковой шкале они **образуют последовательность** от объекта «самое малое значение» к объекту «самое большое значение» (или наоборот).

В порядковой шкале должно быть не менее трех классов. От классов легко перейти к числам, если мы условимся считать, что низший класс получает ранг 1, средний класс - ранг 2, а высший класс - ранг 3, или наоборот.

3.1. Шкалы измерения значений характеристики эксперимента

Порядковая шкала используется практически **во всех областях человеческой деятельности.**

Например, *в минералогии* используется шкала Мооса, по которому минералы классифицируются согласно критерию твердости, в *медицине* - шкала стадий гипертонической болезни (по Мясникову), шкала степеней сердечной недостаточности (по Стражеско) и т.д.

Порядковая шкала также используется при *оценке качества продукции и услуг, определение сортность продукции, оценка мнения экспертов* и т.д.

3.1. Шкалы измерения значений характеристики эксперимента

Интервальная шкала - это шкала, сформированная по принципу «больше на определенное количество единиц» - «меньше на определенное количество единиц». **На такой шкале исследователь сам задает точку отсчета и выбирает единицу измерения**. По интервальной шкале, например, измеряют величину потенциальной энергии.

Допустимыми преобразованиями в шкале интервалов являются линейные преобразования. Например, температурные шкалы Цельсия и Фаренгейта связаны именно такой зависимостью: $\mathbf{C = (F - 32)/1,8}$. Здесь C - температура в градусах по шкале Цельсия, а F - температура по шкале Фаренгейта.

3.1. Шкалы измерения значений характеристики эксперимента

Относительная шкала или шкала отношений. Особенностью этой шкалы является наличие твердо фиксированного нуля, который означает полное отсутствие какого-либо свойства или признака. Если строго фиксировать начало отсчета, то любая интервальная шкала превращается в шкалу отношений.

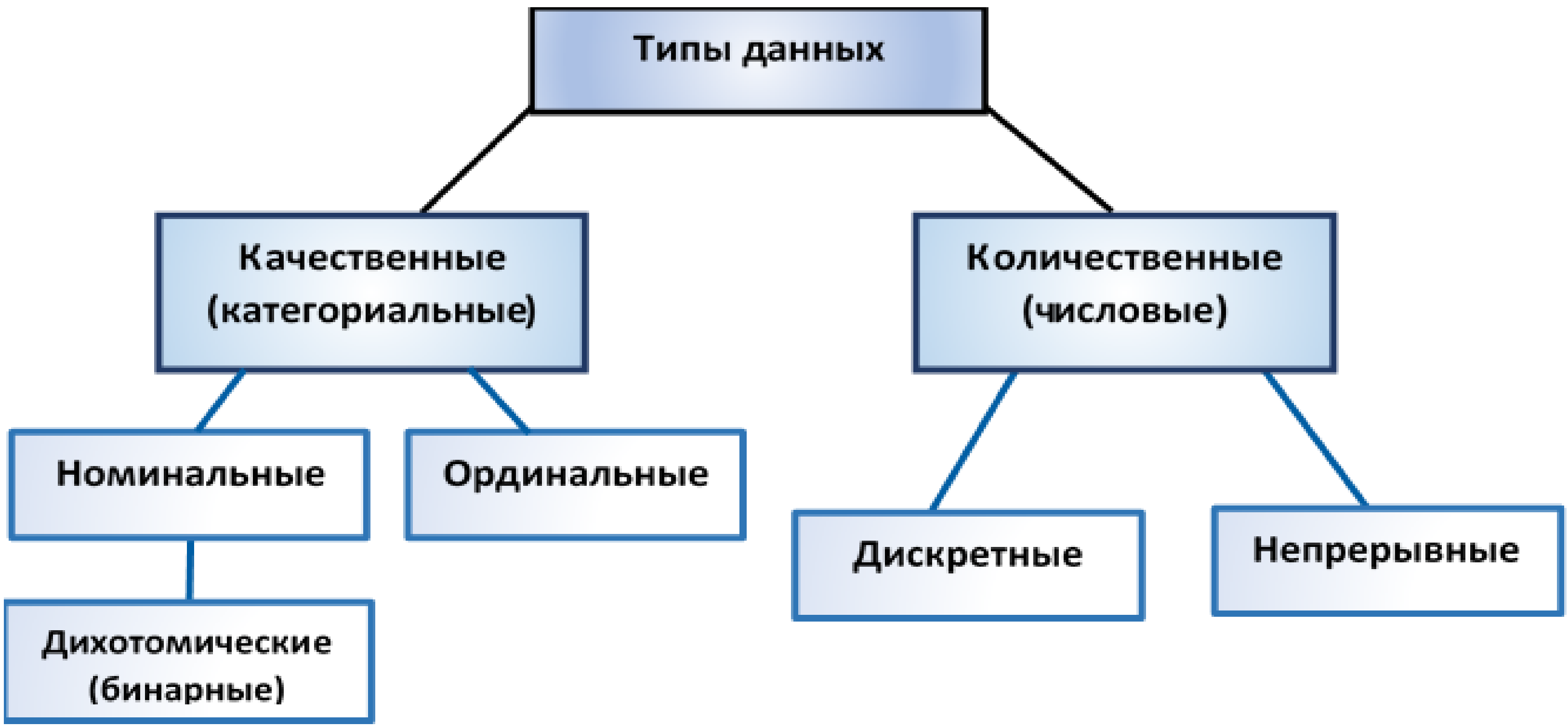
Относительная шкала является наиболее информативной и допускает любые математические операции.

По относительной шкале измерены большинство физических величин: масса, длина, заряд и т.д. Это самая распространенная шкала и в экономике.

4. ТИПЫ И ВИДЫ ДАННЫХ

4.1. Типы данных

При выборе статистического метода для анализа, предварительно **необходимо определить, какому типу относятся данные**. На это влияют в том числе и шкалы их измерений. Основные типы данных представлены ниже.



4.1. Типы данных

Качественные данные — это субъективная информация, которую нельзя измерить, здесь описываются характеристики, атрибуты, свойства, качества явления или объекта.

Номинальные данные — это качественные данные, в которых категории не упорядочиваются, а просто имеют названия. Например, семейное положение (замужем, вдова, не замужем и т. д.). Такого рода данные часто называют *категоризованными*, поскольку о каждом из рассматриваемых объектов известно, в какую из нескольких заранее заданных категорий он попадает.

4.1. Типы данных

Дихотомические (бинарные) данные — это номинальные данные, которые могут принимать одно из двух значений (0 или 1), т.е. являются результатами измерений значений альтернативного признака.

Ординальные (ранговые, порядковые) данные — качественные данные, в которых категории могут упорядочиваться, но интервал между значениями таких данных не может быть выражен количественно. Обычно они качественно отражают *условную степень выраженности какого-либо признака*. Например, стадии болезни (запушенная стадия, средняя, начальная стадия болезни или отсутствие болезни), и т.д.

4.1. Типы данных

Количественные данные — данные, которые имеет некоторую числовую величину (значение). Можно подразделить числовые данные на два типа.

Дискретные данные — такие количественные данные, при которых величина может принимать только определенные числовые значения из конечного множества.

Например, *число детей в семье; число вызовов "скорой помощи", поступающих в больницу; число отказов изделия; число клиентов, обратившихся в фирму за определенный промежуток времени, и т. д*.

4.1. Типы данных

Непрерывные данные — количественные данные, которые могут принимать любые действительные значения из некоторого промежутка. Как правило, *непрерывные данные предполагают большую точность.*

Например, *дальность полета снаряда с точностью до метра, урожайность культуры, выращенной в хозяйстве с точностью до кг, рост взрослого человека, с точностью до мм, фактическая масса буханки хлеба с точностью до мг и т. п.*

4.1. Типы данных

Приведенное выше **деление на типы данных достаточно условно**, поскольку существуют еще различные виды статистических данных, определяемые не только шкалой, но еще и **способами получения**.

Например **цензурированные** данные, **первичные** и **вторичные** данные.

При изучении выборки с несколькими упорядоченными характеристиками (возможно разного типа) мы *получаем вектор,* который также можно рассматривать как новый вид данных – **многомерные данные**.

4.2. Одномерные данные

Выборочные значения $x_1, x_2, \ldots, x_n$ – это **одномерные данные**, они характеризуют одну случайную величину ξ. В таблице 1 представлены одномерные данные, а именно изменение индекс потребительских цен на территории Нижегородской области по годам.

Таблица 1

ИНДЕКСЫ ПОТРЕБИТЕЛЬСКИХ ЦЕН ПО НИЖЕГОРОДСКОЙ ОБЛАСТИ

	2012	2013	2014	2015	2016	2017	2018
	Декабрь к декабрю предыдущего года						
Индекс потребительских цен	121,7	109,9	111,4	112,2	105,4	103,1	104,7

Источник: https://nizhstat.gks.ru/publication_collection

4.3. Многомерные данные

Многомерные данные содержат информацию о двух или более характеристиках для каждого элемента $\boldsymbol{B_i}$.

В дополнение к той информации, которую можно извлечь из одномерных данных, многомерные данные можно использовать для получения информации о том, существует ли простая зависимость между этими признаками, насколько они взаимосвязаны, можно ли предсказать значение одной переменной на основании значений остальных и т.д.

4.3. Многомерные данные

Примеры многомерных данных :

- характеристика работника некоторой фирмы с помощью нескольких показателей: заработная плата, пол, образование, стаж работы, категория работы и производительность труда;
- характеристика квартиры на рынке вторичного жилья в Нижнем Новгороде с помощью следующих показателей: стоимость квартиры, общая площадь, площадь кухни, удаленность от центра, этаж, материалы стен дома.

Таблицы 2 и 3 также содержат примеры многомерных данных.

4.3. Многомерные данные

Таблица 2

11. ЗАНЯТОЕ В ЭКОНОМИКЕ НАСЕЛЕНИЕ ЧАСТНЫХ ДОМОХОЗЯЙСТВ В ВОЗРАСТЕ 20 - 72 ЛЕТ ПО НАЛИЧИЮ УЧЕНОЙ СТЕПЕНИ ПО СУБЪЕКТАМ РОССИЙСКОЙ ФЕДЕРАЦИИ

	Занятое население с высшим и послевузовским профессиональным образованием	Указавшие наличие ученой степени	в том числе			Не указавшие наличие ученой степени
			кандидата наук	доктора наук	не имеющие ученой степени	
Приволжский федеральный округ	**3844491**	**3668805**	**61799**	**10220**	**3596786**	**175686**
Республика Башкортостан	436537	436476	8476	1396	426604	61
Республика Марий Эл	81208	80229	1096	131	79002	979
Республика Мордовия	116827	115318	2244	315	112759	1509
Республика Татарстан	557224	530272	10258	1735	518279	26952
Удмуртская Республика	194514	189502	2407	412	186683	5012
Чувашская Республика	150517	148718	2014	280	146424	1799
Пермский край	278339	267245	3910	712	262623	11094
Кировская область	148549	146729	1742	236	144751	1820
Нижегородская область	497215	460898	7996	1387	451515	36317
Оренбургская область	230475	225440	3098	490	221852	5035
Пензенская область	166649	161354	2439	324	158591	5295
Самарская область	502969	444741	7102	1289	436350	58228
Саратовская область	335384	317955	7030	1219	309706	17429
Ульяновская область	148084	143928	1987	294	141647	4156

Источник https://www.gks.ru

4.3. Многомерные данные

Таблица 3

ОСНОВНЫЕ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИЕ ХАРАКТЕРИСТИКИ НИЖЕГОРОДСКОЙ ОБЛАСТИ

ОСНОВНЫЕ СОЦИАЛЬНО-ЭКОНОМИЧЕСКИЕ ПОКАЗАТЕЛИ

	2013	2014	2015	2016	2017	2018
1. Численность населения (на конец года), тыс. человек	3307,6	3270,2	3260,3	3247,7	3234,8	3214,6
2. Естественный прирост, убыль (-) населения, тыс. человек	-23,0	-13,1	-10,6	-11,4	-13,6	-16,4
3. Миграционный прирост, убыль (-) населения, человек	3796	1782	702	-1134	595	-3731
4. Среднегодовая численность занятых, тыс. человек	1710,9	1677,7	1650,9	1644,9	1658,7	1633,1
5. Численность безработных (по методологии МОТ), тыс. человек	139,8	75,2	75,2	76,3	75,2	73,1
6. Численность пенсионеров, тыс. человек	1006,3	1031,6	1040,9	1045,6	1050,2	1054,1
7. Среднедушевые денежные доходы населения в месяц, руб.	16477,3	27048,6	30003,5	30057,2	30325,7	31408,0
8. Среднемесячная номинальная начисленная заработная плата работников организаций, руб.	16327,6	25497,1	26480,7	28399,0	30387,1	32949,3

4.3. Многомерные данные

Из многомерных данных можно без труда сформировать одномерные, просто выбрав интересующую нас характеристику.

Например из таблицы 3 выберем только «Среднедушевые денежные доходы населения в месяц» и получим одномерные данные (таблица 4)

Таблица 4

	2010	2014	2015	2016	2017	2018
Среднедушевые денежные доходы населения в месяц, руб.	16477,3	27048,6	30003,5	30057,2	30325,7	31408,0

5. ЗАКЛЮЧЕНИЕ

5. Заключение

- В математической статистике возникают такие понятия, как **генеральная совокупность Γ и выборочная совокупность $\Gamma в \subset \Gamma$.** Для выборочной совокупности обязательным является требование **репрезентативности.**
- Существует несколько способов представления значений выборочной совокупности, каждый способ удобен в своем случае.
- **Данные,** или элементы выборки, **могут иметь различный вид**, на который влияют в том числе и шкалы их измерений.
- **Правильно описать тип данных очень важно**, поскольку это влияет и на применяемые методы дальнейшего статистического анализа.

Литература

1. Федоткин М.А. Основы прикладной теории вероятностей и статистики. — М.: Высшая школа. 2006. - 168 с.

2. Практическая статистика для специалистов Data Science: Пер. с англ./ П. Брюс, Э. Брюс. — СПб.: БХВ-Петербург, 2018. — 304 с.

3. Лагутин М. Б. Наглядная математическая статистика: учебное пособие. — 2-е изд., испр. — М. : БИНОМ. Лаборатория знаний, 2009. — 472 с.