



Mean values in historical studies

O. D. Dashkovskaya¹

¹P. G. Demidov Yaroslavl State University, Yaroslavl, Russian Federation

DOI: 10.18255/1996-5648-2024-1-60-69

Research article
Full text in Russian

The article examines the possibilities of using averages in historical researches. The basic principles of the application of these mathematical indicators are considered, the main errors of researchers arising from non-compliance with these rules are identified. On the basis of historical material, such averages as arithmetic mean, mode and median are analyzed in detail, their calculation in simple and non-standard cases (if statistical indicators are presented in the form of groupings, closed or open intervals). The possibilities of the «Microsoft Excel» program for calculating averages are demonstrated. Finally, the indicators of variation are characterized, allowing to reflect the individual values of the units of the statistical population.

Keywords: historical research; measurement; statistical totality; arithmetic mean; mode; median; interval; indicators of variation

INFORMATION ABOUT AUTHORS

Dashkovskaya, Olga D. | E-mail: dashod@mail.ru
Cand. Sc. (History), Associate Professor



Средние величины в исторических исследованиях

О. Д. Дашковская¹

¹Ярославский государственный университет им. П. Г. Демидова, Ярославль, Российская Федерация

DOI: 10.18255/1996-5648-2024-1-60-69
УДК 930.2 +311.14

Научная статья
Полный текст на русском языке

В статье изучаются возможности использования средних величин в исторических исследованиях. Рассматриваются основные принципы применения данных математических показателей, выявляются основные ошибки исследователей, возникающие при несоблюдении указанных правил. На основе исторического материала подробно проанализированы такие средние величины, как средняя арифметическая, мода и медиана, их расчет в простых и нестандартных случаях (если статистические показатели представлены в виде группировок, закрытых или открытых интервалов). Продемонстрированы возможности программы «Microsoft Excel» при вычислении средних величин. Наконец, охарактеризованы показатели вариации, позволяющие отразить индивидуальные значения единиц статистической совокупности.

Ключевые слова: историческое исследование; измерение; статистическая совокупность; средняя арифметическая; мода; медиана; интервал; показатели вариации

ИНФОРМАЦИЯ ОБ АВТОРАХ

Дашковская, Ольга Дмитриевна | E-mail: dashod@mail.ru
Кандидат исторических наук, доцент

Математические методы уже достаточно давно используются отечественными и зарубежными историками. Исследователи активно применяют их при изучении социально-экономической истории, исторической демографии, археологии, при анализе статических источников и нарративов [1]. Обработка материала с помощью математического инструментария позволяет вывести исторические исследования на новый количественный и качественный уровень. Использование методов систематизации и моделирования в исторической науке уже было охарактеризовано нами в пре-

дыдущих статьях, опубликованных в «Вестнике Ярославского государственного университета им. П. Г. Демидова. Серия Гуманитарные науки» [2–3].

Цель данной работы – проанализировать возможности применения одного из самых распространенных математических показателей, встречающихся в практике историков, – средних величин. Их важнейшая роль заключается в обобщении, при котором одна усредненная величина заменяет разнообразие отдельные значения какого-либо признака.

Средние величины применялись еще в государствах Древнего мира при налогообложении (усреднялись доходы различных категорий населения). В дореволюционной России материалы очень многих обследований представляли собой тома с усредненными данными о землевладении, прибыли и прочих показателях имущественного обеспечения сословий.

В современной практике средние величины встречаются повсеместно: средняя заработная плата, средние показатели рождаемости, средние температуры воздуха за определенный период и т. п. Такие усреднения уже давно стали основой для различных анекдотов, типа: «Начальник ест мясо, а я капусту, в среднем мы едим голубцы».

Избежать подобных ошибок позволяет соблюдение ряда общих принципов применения средних величин:

1. Средние величины всегда должны высчитываться по качественно однородной совокупности. Например, нельзя вычислять средний уровень заработной платы на предприятии по общим данным о доходах дирекции, управленческих кадров и простых рабочих;

2. При вычислении средних величин необходимо использовать максимально возможное число единиц наблюдения. Некорректно приводить показатели урожайности хлебов в Российской империи, основанные на материалах нескольких уездов;

3. Необходимо вычислять не только общие средние величины, но и сопровождать их показателями по отдельным группам. Так, при расчете среднемесячной заработной платы в современной России нужно приводить сведения и по отдельным субъектам РФ, по видам экономической деятельности и пр.;

4. Вычисления должны производиться на основании использования одних и тех же единиц измерения. Например, при определении сборов зерна до 1890-х гг. применялись исключительно меры объема. Даже посевная площадь измерялась в единицах объема – четвертях. Между тем весовое наполнение четверти существенно отличалось в различных губерниях, и у каждой культуры (ржи, пшеницы, овса, ячменя) четверть была «своя», что зависело от веса самих зерновых.

До середины XIX века даже единица земельной площади десятина могла высчитываться по-разному. Кроме традиционной казенной десятины, составлявшей 2400 кв. саженой, встречалась и хозяйственная, или вла-

дельческая (3200 кв. сажений). Перед нахождением средних необходимо привести единицы измерения к одному масштабу.

Для характеристики средних величин необходимо дать пояснение нескольким употребляемым в статистике терминам. Всё множество объектов, имеющих общий признак, с которыми работает исследователь, называется генеральной (статистической) совокупностью.

Варианты (переменные) – это объективные характеристики всего, что можно измерить: заработная плата, количество рабочих на предприятии, размер землевладения, возраст и т. д. Частота варианта – число повторений какого-либо значения переменной. Например, исследователь анализирует личные дела рабочих промышленного предприятия, в которых указана их заработная плата. В данном случае эти документы являются статистической совокупностью. Зарботная плата – признак, а ее конкретные значения – варианты (или переменные). Допустим, в 1913 г. заработная плата 33 рабочих фабрики составляла 22 рубля (у остальных были бóльшие или меньшие оклады). Таким образом, частота признака «заработная плата» при переменной «22» равна «33».

Наиболее часто в исторической науке используется *средняя арифметическая*. Она изучается в школе и слово «арифметическая» часто опускается исследователями, которые просто упоминают о вычислении «среднего размера заработной платы», «средней урожайности картофеля». Ее формула проста и определяется как сумма всех значений признака, поделенная на их количество:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_n}{n},$$

где X_1, X_2, X_3, X_n – варианты признака,
 n – число единиц наблюдения.

Пример 1. В 1915 г. в семи губерниях Центрально-промышленного района урожайность озимой ржи составляла (единица измерения – сам): Владимирская – 4,7; Костромская – 5,4; Ярославская – 5,5; Московская – 5,0; Тверская – 5,7; Калужская – 4,4; Нижегородская – 5,7 [4, с. 75–75]. Необходимо выявить среднюю урожайность озимой ржи за указанный год:

$$\bar{X} = \frac{4,7 + 5,4 + 5,5 + 5,0 + 5,7 + 4,4 + 5,7}{7} = 5,2 \text{ сам}$$

Итак, средняя урожайность озимой ржи в 1915 г. в Центрально-промышленном районе составила 5,2 сам.

Иногда историкам необходимо работать с групповыми статистическими данными. В этом случае необходимо вычислять *взвешенную среднеарифметическую величину*. В формуле средневзвешенной в числителе находится сумма произведений значения каждого признака

и частоты повторения данного признака, а в знаменателе – сумма всех частот признаков:

$$\bar{X} = \frac{\sum XR}{\sum R},$$

где X – варианты признака, R – частота варианта.

Данная формула применяется в том случае, если варианты генеральной совокупности встречаются неодинаковое число раз.

Пример 2. Дана заработная плата рабочих предприятия за месяц (табл. 1).

Таблица 1

Зарботная плата рабочих, тыс. руб.

Зарплата одного рабочего	20	25	30	32	35
Количество рабочих	6	4	10	5	5

Требуется найти средневзвешенную зарплату рабочих.

В данном случае зарплата – признак, его конкретные значения – варианты признака (первая строка), а число повторений одного и того же признака – частота варианта (вторая строка). Подставляем данные в указанную формулу средневзвешенной величины.

$$\bar{X} = \frac{20*6+25*4+30*10+32*5+35*5}{6+4+10+5+5} = 28,5 \text{ тыс. руб.}$$

Итак, средняя заработная плата работников составляет 28,5 тыс. руб.

Иногда статистические данные, которыми пользуется историк, представлены в виде интервальных рядов. Последние могут содержать как равные, так и неравные интервалы. Интервал считается «равным», если его ширина одинакова во всех диапазонах. Различают также закрытые и открытые интервалы. В закрытых всегда указаны и верхние, и нижние границы (например, от 0 до 10, от 10 до 20,...от 90 до 100). В данном случае при расчете средней арифметической в формулу подставляются середины интервалов – 5, 15,...95).

Однако на практике встречаются и открытые интервалы. У них указана только одна граница: у первого – верхняя, у последнего – нижняя. Открытый интервал может быть задан формулировками «до 18 лет», «от 100 десятин» и пр.

Пример 3. Показан доход причтов¹ Ярославской епархии в середине XIX века (табл. 2).

¹ Причт – это священно- и церковнослужители одного прихода/церкви.

Доходы причтов Ярославской епархии в 1863 г., руб. [5, л. 102об.—103об.]

Размер дохода	до 200	200–400	400–600	600–800	800 и более
Количество причтов	139	450	225	115	42

В случае открытого интервала середина определяется методом аналогий. Если все остальные диапазоны имеют одинаковую ширину, можно предположить, что размерность открытого интервала такая же. То есть у первого интервала середина это число 100, у последнего – 900. Остальные вычисления производятся по формуле из второго примера.

$$\bar{X} = \frac{100 \cdot 139 + 300 \cdot 450 + 500 \cdot 225 + 700 \cdot 115 + 900 \cdot 42}{139 + 450 + 225 + 115 + 42} \approx 391 \text{ руб.}$$

Итак, средний доход одного причта Ярославской епархии в 1863 году составлял ≈ 391 руб.

Для изучения внутреннего строения изучаемой исследователем совокупности используются мода и медиана. Эти показатели еще именуются структурными средними величинами.

Модой (M_0) называют такое значение признака, которое наиболее часто встречается в статистической совокупности. Эту величину активно используют для анализа спроса на отдельные товары, при изучении ценообразования. Например, маркетинговый отдел обувной фабрики интересуется наибольшим спросом на конкретные размеры обуви. Вполне понятно, что вычисление средней арифметической тут неуместно, тем более что полученное число может оказаться дробным. А вот применение такой величины, как мода, вполне логично.

Если анализируемая совокупность представляет последовательность чисел, то мода будет значением, которое встречается чаще всего. Например, в ряду 5, 3, 6, 3, 5, 3, 4 модой является число 3. Оно повторяется чаще остальных.

Вычислять моду в простой группировке не требуется, так как очевидно, что она определяется как вариант признака с наибольшей частотой. Во втором примере мода равняется 30, именно данной заработной плате соответствует наибольшее количество получающих ее рабочих – 10 человек.

Расчет моды в интервальном ряду – это достаточно сложная процедура, так как сначала требуется выяснить интервал с наибольшей частотой. Он именуется модальным классом. Потом для вычисления моды (M_0) используется следующая формула:

$$M_0 = X_0 + K \frac{P_2 - P_1}{2P_2 - P_1 - P_3},$$

при этом

X_0 – нижняя граница модального интервала;

K – ширина интервала;

P_1 – частота интервала, находящегося до модального;

P_2 – частота модального интервала;

P_3 – частота интервала, находящегося после модального.

Рассчитаем (конечно, условно и исключительно для демонстрации расчетов) моду на основе примера 3. Модальным интервалом будет являться диапазон 200–400, так как именно он содержит наибольшую частоту – 450.

$$M_0 = 200 + 200 * \frac{450 - 139}{2 * 450 - 139 - 225} = 316.$$

Это значит, что доход причта составлял 316 руб.

Медиана (Me) – это значение признака, расположенное в середине совокупности, если упорядочить последнюю по возрастанию. Медиана разделяет все имеющиеся варианты на две одинаковые части: в первой находятся все значения признака меньше медианы, во второй – те, что больше нее. Она позволяет получить максимально точные результаты, наиболее близкие к истинно среднему значению, поэтому ее считают представительной средней величиной.

Особенно логично использовать медиану, если изучаемая совокупность неоднородна. Например, при расчете средней заработной платы по стране разумнее вычислять именно медиану, а не среднюю арифметическую. Многим, например, не нравится, что при зарплате 20 тыс. руб. у обычного работника и 200 тыс. руб. – у директора средний уровень зарплаты равен 110 тыс. руб. Получается, что у большинства заработная плата оказывается ниже средней по стране/региону.

Если данных немного, то медиану можно определить без дополнительных вычислений. Для этого все имеющиеся значения располагаются в порядке возрастания и находится середина. Например, даны числа 5, 2, 7, 1, 14. Расположим их по возрастанию: 1, 2, 5, 7, 14. Число 5 и будет являться медианой. Если количество чисел четное, то необходимо сложить 2 числа, располагающихся посередине, и разделить на 2. Полученный итог и будет являться медианой.

Для интервальной группировки медиана вычисляется с помощью следующей формулы:

$$M_e = X_0 + K \frac{\sum_2^P - \sum_{m-1}}{P_m},$$

где

X_0 – нижняя граница медианного интервала;

K – величина медианного интервала;

\sum^P – сумма частот;

Σ_{m-1} – сумма частот интервалов, предшествующих медианному (накопленная частота в интервале, предшествующем медианному);

P_m – частота медианного интервала.

Рассчитаем медиану, используя цифровой материал третьего примера. Первоначально необходимо найти медианный интервал, для этого продублируем нашу таблицу (табл. 3).

Таблица 3

Доходы причтов Ярославской епархии в 1863 г., руб.

Размер дохода	до 200	200–400	400–600	600–800	800 и более
Количество причтов	139	450	225	115	42
Накопленные частоты	139	589	814	929	971

В таблицу была добавлена дополнительная строка, в которой отражены накопленные частоты (139+450; 139+450+225 и т. д.). Она необходима для вычисления медианного интервала. В качестве последнего выбирают тот диапазон, накопленная частота которого превышает половину всей суммы частот или равняется ей. Таким образом, половина суммы всех частот вычисляется на основе второй строки:

$$(139+450+225+115+42)/2 = 485,5.$$

На основании третьей строки мы видим, что медианным является интервал 200–400, так как $485,5 > 450$.

Затем находим значение медианы, подставив в формулу необходимые цифровые данные:

$$M_c = 200 + 200 \frac{485,5 - 139}{450} = 354.$$

Проведенные вычисления показывают, что одна половина церковных причтов имеет доход до 354 руб., а вторая половина – свыше этой суммы.

Итак, на основе примера 3 мы видим, что показатели, полученные при вычислении средневзвешенной, моды и медианы несколько отличаются (391, 316 и 354 руб. соответственно). Эти различия вызваны асимметричным распределением. На среднюю арифметическую оказывает влияние каждая переменная, именно поэтому данная величина смещается в сторону максимальных значений признака. Разумнее вычислять моду по отношению к качественным признакам. В третьем же примере применение медианы является более уместным, так как в значении признаков отмечаются существенные колебания, а границы крайних интервалов не определены.

Если данных много, то средние величины логично вычислять с помощью программы «Microsoft Excel», предназначенной для работы с электронны-

ми таблицами. Для этого необходимо зайти в раздел «Формулы» – «Вставить функцию» и через «Поиск функции» или в категории «Статистические» выбрать «СРЕДНЕЕ АРИФМЕТИЧЕСКОЕ», «МОДА» или «МЕДИАНА».

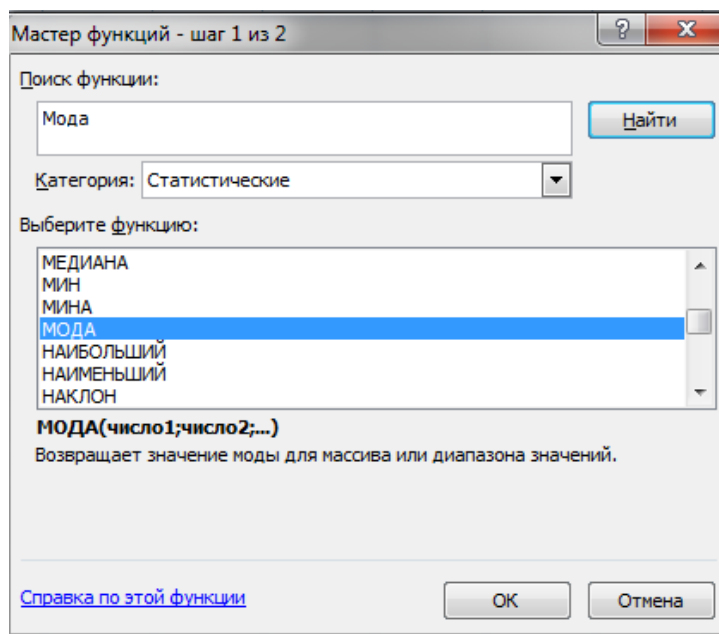


Рис. 1. Статистические функции в программе «MS Excel»

Затем активируется ячейка для расчета, выделяется необходимый диапазон данных и нажимается «ОК». Результат будет отображен в ячейке, в которой находился курсор мыши. К сожалению, для расчета средних величин по интервальным рядам соответствующей функции в MS Excel нет. Поэтому придется использовать приведенные выше формулы.

Средние величины используются, если необходимо дать обобщающую характеристику совокупности в целом, сгладить индивидуальные различия единиц. Подчеркнуть их и помочь понять сущность изучаемого явления историкам помогают разнообразные показатели вариации.

Вариация представляет собой изменяемость величины признака у отдельных единиц генеральной совокупности.

Упрощенным показателем является размах вариации (R), представляющий разность между наибольшим и наименьшими значениями признака:

$$R = X_{\max} - X_{\min}$$

где X_{\max} – максимальное значение признака, X_{\min} – минимальное значение признака в статистической совокупности.

В примере 2 наибольшим значением является заработная плата 35 тыс. руб., наименьшим – 20 тыс. руб., значит, размах вариации равен

15. В то же время размах вариации считается грубым показателем, так как учитывает лишь крайние значения признака.

Историки используют более представительные показатели вариации в своих работах. Так, знаменитый советский исследователь И. Д. Ковальченко применял среднее квадратическое отклонение. Первоначально на основании средней арифметической он выяснил, что в среднем размер дореформенного надела у барщинных и оброчных крестьян Симбирской губернии был практически одинаковым. Однако вычисление среднего квадратического отклонения показало, что в среде оброчных крестьян различия в размере наделов были в полтора раза больше, чем в среде барщинных крестьян [1, с. 95–96].

При современном уровне развития информационных технологий среднее квадратическое отклонение, коэффициент осцилляции, дисперсию и прочие показатели вариации логично находить с помощью статистических функций пакета «Microsoft Excel».

Итак, средние величины используются историками как обобщающие показатели, характеризующие общие свойства статистической совокупности. Они позволяют выявить закономерные черты, свойственные изучаемым объектам, показывают тенденции развития процессов и явлений, нивелируя случайные индивидуальные отклонения. Если же исследователям необходимо подчеркнуть индивидуальные различия единиц совокупности, они применяют показатели вариации.

Ссылки

1. Количественные методы в исторических исследованиях: учеб. пособие для студ. вузов, обуч-ся по спец. «История» / под. ред. И. Д. Ковальченко. М.: Высшая школа, 1984. 384 с.
2. Дашковская О. Д. Математическое моделирование в исторических исследованиях: задачи, особенности и возможности применения метода // Вестник Ярославского государственного университета им. П. Г. Демидова. Серия Гуманитарные науки. 2020. № 2. С. 48–55. EDN CDMOQE.
3. Дашковская О. Д. Методы систематизации количественной информации в исторических исследованиях // Вестник Ярославского государственного университета им. П. Г. Демидова. Серия Гуманитарные науки. 2021. Т. 15, № 4. С. 506–515. DOI 10.18255/1996-5648-2021-4-506-515.
4. Растяжников В. Г., Дерюгина И. В. Урожайность хлебов в России, 1795–2007. М.: ИВ РАН, 2009. 192 с.
5. Российский государственный исторический архив. Ф. 804. Оп. 1- III. Д. 463. 110 л.