Тема: Понятие генеральной совокупности и выборки.

1. Изучить теоретический материал. Рассмотреть примеры.

Важнейшими понятиями математической статистики являются понятия генеральной совокупности и выборочной совокупности. Рассмотрим эти понятия на простых примерах.

Пример 1

Студент выполняет лабораторную работу по определению коэффициента вязкости жидкости методом Стокса.

Экспериментальная часть этой работы состоит в том, что в высокий цилиндрический сосуд с жидкостью сбрасывается достаточно маленький и тяжёлый шарик, после чего замеряется время его погружения.

Время погружения шарика зависит от множества случайных факторов: прямоты рук экспериментатора, погрешности измерения времени, хаотичного движения молекул жидкости и т.д., вплоть до влияния Луны. Поэтому эксперимент целесообразно провести 5-10 раз. Предположим, что в результате 5 опытов получены следующие результаты (в секундах): $t_1 = 6{,}9, \quad t_2 = 6{,}7, \quad t_3 = 7, \quad t_4 = 7{,}2, \quad t_5 = 6{,}8$

Что произошло? Студент собрал ***первичные*** *(ещё не обработанные)* статистические данные. Они *эмпирические* (взяты непосредственно из опыта), носят случайный характер.

Полученные экспериментальные значения называются ***вариантами (вариАнта*** *(существительное женского рода) – в статистике означает отдельно взятое эмпирическое значение*., а их совокупность – ***вариационным рядом***. Почему так? Потому что полученные значения *варьируются* под воздействием случайных факторов.

Далее студент должен обработать полученные данные. Во-первых, посмотреть, а нет ли среди полученных значений *варианты*, которая сильно отличается от всех остальных? Наличие такого значения сигнализирует о том, что соответствующий опыт проведён неудачно и его следует исключить из рассмотрения.

Нет, все значения достаточно близкИ друг к другу, и теперь напрашивается вычислить среднюю величину – разделить сумму значений на их $n = 5$ количество:

$$\bar{t} = \frac{\sum_{i=1}^{5} t_i}{n} = \frac{t_1 + t_2 + t_3 + t_4 + t_5}{n} = \frac{6{,}9 + 6{,}7 + 7 + 7{,}2 + 6{,}8}{5} = \frac{34{,}6}{5} = 6{,}92$$ секунды.

Это значение называют ***простой средней*** или ***средним арифметическим***. Его стандартно обозначают с чёрточкой наверху.

(математический значок $\sum$ *означает суммирование, а переменная* i *играет роль «счётчика»; в данном случае* i *изменяется от 1 до 5*.

Если есть сомнения на счёт точности, то лучше не полениться и провести 10 опытов, что, кстати, удобнее в плане вычислений (на 10 делить проще). И, разумеется, полученный результат будет надёжнее, чем в 1-м случае.

Статические данные обработаны, осталось сделать выводы. А именно, с помощью значения $\bar{t}$ вычислить коэффициент вязкости жидкости.

Пример 2

Студенческая группа сдала зачет по математике со следующими результатами:

x_i	2	3	4	5
N_i	5	10	7	3

Требуется определить среднюю успеваемость группы

Сбором статистических данных здесь занимался преподаватель, и обратите внимание на их характер: они эмпирические, массовые и отчасти случайные. Кому-то повезло с вопросом, кому-то нет, кто-то что-то вспомнил, забыл, списал, прогулял и так далее.

Как нетрудно понять, роль ***вариант*** x_i здесь играют полученные оценки, а N_i – это соответствующие ***частоты*** – количество студентов, которые получили ту или иную оценку. Подсчитаем общую численность группы:

$$N = \sum_{i=1}^{4} N_i = N_1 + N_2 + N_3 + N_4 = 5 + 10 + 7 + 3 = 25$$ человек,

исследуемое множество называют ***статистической совокупностью***, а количество его элементов – ***объёмом*** совокупности.

Теперь обратим внимание на следующую вещь: двоечников и отличников у нас мало, а нормальных студентов много. И возникает вопрос: как вычислить «справедливую» среднюю оценку по всей совокупности? Решение напрашивается – с помощью так называемой ***средневзвешенной средней***:

$$\bar{x} = \frac{\sum_{i=1}^{4} x_i N_i}{\sum_{i=1}^{4} N_i} = \frac{x_1 N_1 + x_2 N_2 + x_3 N_3 + x_4 N_4}{N} = \frac{2 \cdot 5 + 3 \cdot 10 + 4 \cdot 7 + 5 \cdot 3}{25} =$$

$$= \frac{10 + 30 + 28 + 15}{25} = \frac{83}{25} = 3{,}32$$ – средняя успеваемость по группе.

Давайте проанализируем их **принципиальные отличия** этих двух примеров:

1) В первом примере проводится статистическое исследование *количественной* величины (времени), а во втором «оцифровывается» и анализируется *качественный* признак (успеваемость).

2) В первом случае исследуемая величина *непрерывна*, и, строго говоря, все полученные значения **различны** (отличаются хоть какими-то миллисекундами). Во втором случае варианты *дискретны*, т.е. представляют собой отдельно взятые изолированные значения. Следует заметить, что они не обязаны быть целыми, так, например, можно ввести в рассмотрение оценки 2,5; 3,5 и 4,5. И у дискретной величины, как правило, есть *неоднократно* встречающиеся (одинаковые) варианты, так, например, «пятёрка» встретилась 3 раза.

3) В первом примере речь идёт о ***выборке*** значений. Что это значит? Это значит, что шарик можно сбрасывать в воду гораздо бОльшее и теоретически вообще бесконечное количество раз. Таким образом, проведённые 5 опытов есть, по сути, *выборка*, которую называют ***выборочной совокупностью***. При этом соответствующее среднее значение принято называть ***выборочной средней***.

Второй пример отличен тем, что в нём исследуется ВСЯ совокупность, и поэтому её называют ***генеральной совокупностью***, а соответствующее среднее значение – ***генеральной средней***. Но такая ситуация редкость. Редко когда удаётся исследовать всю совокупность.

Рассмотрим **основной метод математической статистики**:

Пример 3

Некоторый статистик пошёл на базу исследовать помидоры. Требуется определить среднюю массу помидора и среднюю долю первосортных помидоров.

Разбираемся в ситуации. Очевидно, что на базе находится очень и очень много помидоров, обозначим их общее количество через N. Это *генеральная совокупность*. Для того чтобы решить задачу, можно взвесить каждый овощ: $x_1, x_2, x_3, \ldots, x_N$ (в граммах, например) и вычислить *генеральную среднюю*:

$\overline{x}_г = \frac{x_1 + x_2 + x_3 + \ldots + x_N}{N}$ – среднюю массу помидора.

Но это долго и трудно.

Поэтому для оценки параметров генеральной совокупности целесообразно использовать ***выборочный метод***. Его суть состоит в том, что из генеральной совокупности достаточно выбрать n объектов, которые хорошо характеризуют всю совокупность. Это «хорошо» называют ***представительностью*** или, как говорят, ***репрезентативностью*** выборки.

Что нужно для того, чтобы обеспечить репрезентативность?

Ну, во-первых, выборка должна быть достаточно велика, помидоров так 500-1000 точно, что уже вполне по силам даже одному человеку.

Во-вторых, отбор следует осуществлять *равномерно* – из каждого ящика.

В-третьих, отбор должен быть *случайным*. Для этого используются разные приёмы, и самый простой здесь – это выбор «вслепую» из случайно выбранного места ящика, обязательно с разной глубины (а то мало ли, что поставщик там мог спрятать).

И, в-четвёртых, есть и другие факторы, которые могут быть менее очевидны. В частности, важно знать, а *однородна* ли генеральная совокупность? Так, если помидоры поступили от разных поставщиков, то каждую партию полезно исследовать по отдельности (сделать несколько выборок).

Итак, пусть статистик по всем правилам выбрал n помидоров, и теперь дело за малым – взвесить каждый овощ: $x_1, x_2, x_3, \ldots, x_n$ (граммы) и вычислить *выборочную среднюю*:

$\overline{x}_в = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$ – среднюю массу помидора в выборке.

При этом очевидно, что чем больше ***объем n выборочной совокупности***, тем полученное значение будет точнее приближать генеральную среднюю $\overline{x}_г$.

Если начать увеличивать выборку в два, три и бОльшее количество раз, то будут получаться выборочные средние, которые мало отличаются от уже рассчитанного значения $\overline{x}_в$. Это установлено эмпирически, в результате огромного количества реально проведённых исследований.

Таким образом, **нет никакого практического смысла** тратить силы, время, деньги, нервы на исследование бОльшей выборки и тем более, всей генеральной совокупности.

Вторая часть задачи. Определим вместе со статистиком среднюю долю высококачественных помидоров на базе.

В отличие от первого этапа, здесь мы исследуем уже *качественный* признак, для которого, тем не менее, можно сформулировать чёткие критерии. Пусть первосортный помидор – это красный, спелый, без видимых дефектов, массой выше среднего.

Совершенно понятно, что генеральная совокупность содержит K таких помидоров, и существует точное значение:

$\omega_г = \frac{K}{N}$ – ***генеральная доля*** первосортных помидоров.

Но по причине трудозатратности и нецелесообразности полного исследования, достаточно подсчитать количество k таких овощей в выборке и вычислить:

$\omega_в = \frac{k}{n}$ – ***выборочную долю***, которая будет весьма близка к истинному значению $\omega_г$. Но это только, напомню, при условии грамотно организованной и проведённой выборки.

Доля, как вы догадываетесь, может принимать значение от 0 до 1, и иногда её домножают на 100, чтобы выразить этот показатель в процентах.

В качестве разминки предлагаю вам задачу с тремя пунктами различного уровня сложности. Проверьте наличие инструментов под рукой и свои навыки вычислений (Эксель ~~вечной живой~~ **по-прежнему тут**):

Пример 4

а) Урожайность картофеля по трём областям за год составила 147, 145, 155 ц/га (центнеров с га). Требуется вычислить среднюю урожайность.

Используем простую среднюю:

$\overline{x} = \frac{147+145+155}{3} = \frac{447}{3} = 149$ *ц/га – в среднем по трём областям.*

б) Известны следующие данные по трём областям:

Область	Общая посевная площадь, тыс. га	Урожайность, ц/га
А	139,80	147
Б	102,34	145
В	63,29	155

Требуется вычислить среднюю урожайность.

Обратите внимание, что здесь урожайность, скажем, по 3-й области велика, но её посевная площадь мала. Поэтому урожайность уместно «взвесить» по площадям.

Используем средневзвешенную (по площади) среднюю:

$\overline{x} = \frac{147 \cdot 139{,}8 + 145 \cdot 102{,}34 + 155 \cdot 63{,}29}{139{,}8 + 102{,}34 + 63{,}29} = \frac{20550{,}6 + 14839{,}3 + 9809{,}95}{305{,}43} =$

$= \frac{45199{,}85}{305{,}43} \approx 147{,}99$ *ц/га в среднем по трём областям.*

в) вычислить среднюю урожайность по следующим данным:

Область	Валовой сбор картофеля, тыс. тонн	Урожайность, ц/га
А	2055	147
Б	1484	145
В	981	155

Здесь урожайность тоже следует переоценить через посевную площадь, используя формулу Посевная площадь = Валовой сбор / Урожайность:

$$\overline{x} = \frac{2055+1484+981}{\frac{2055}{147}+\frac{1484}{145}+\frac{981}{155}} \approx \frac{2055+1484+981}{13{,}980+10{,}234+6{,}329} = \frac{4520}{30{,}543} \approx 147{,}9876 \approx 147{,}99$$

ц/га в среднем по трём областям. Такой вид средней иногда называют ***средней гармонической*****.**

И в заключение урока систематизируем самое важное:

Математическая статистика – это наука, изучающая методы сбора и обработки статистической информации для получения научных и практических выводов.

Основным методом математической статистики является **выборочный метод**, его суть состоит в исследовании представительной *выборочной совокупности* – для достоверной характеристики совокупности генеральной. Данный метод экономит временнЫе, трудовые и материальные затраты, поскольку исследование всей совокупности зачастую затруднено или невозможно.