Выборочный метод в статистике.

1. Понятие выборочного наблюдения.
2. Виды выборочного наблюдения.
3. Ошибки выборочного наблюдения.
4. Организация и практика применения выборочного метода наблюдения.

Выборочным называется такое наблюдение, при котором:

а) характеристика всей совокупности единиц дается по некоторой их части;

б) эта часть включает в себя определенное число единиц совокупности отобранных в случайном порядке;

в) характеристика совокупности лежит в известных удовлетворяющих исследователя пределах.

Одно из главных условий выборочного наблюдения заключается в том, что при отборе единиц совокупности для обследования **обеспечивается равная возможность попадания в отобранную часть** любой из единиц. Это достигается путём беспристрастного **строго случайного отбора**, организуемого по схемам предлагаемым математической статистикой.

Другим важным условием является то, что уже при организации выборочного наблюдения **устанавливается численность** или доля единиц совокупности, которая будет подвергнута обследованию.

Рассмотренные условия позволяют установить границы возможных ошибок и получить практически достоверные данные, поскольку эти ошибки могут быть учтены.

Совокупность явлений, из которой производится выбор части для непосредственного изучения называется **генеральной совокупностью (N),** отобранная часть – **выборочной совокупностью** (**n**).

При выборочном наблюдении имеют дело с двумя категориями обобщающих показателей:

- относительными величинами

- средними величинами.

Относительные величины применяют для сводной характеристики совокупностей **по альтернативному признаку**; такая характеристика дается **в виде доли** тех единиц совокупности, которые обладают интересующим признаком. Например, при изучении демографической структуры населения определяют долю состоящих в браке. Во всех подобных случаях мы будем иметь дело с обобщающим показателем в виде относительной доли единиц, составляющих какую-то часть всей совокупности.

Этот сводный показатель для генеральной совокупности называется **генеральной долей**, или **долей в генеральной совокупности (p),** а для выборочной совокупности – **выборочной долей,** или **частостью (w)**.

Задача, таким образом, заключается в том, чтобы на основе выборочной доли дать правильное представление о доле в генеральной совокупности.

Перед выборочным исследованием может также стоять задача измерения среднего значения варьирующего признака во всей совокупности. В этом случае среднее значение варьирующего признака во всей совокупности называется **генеральной средней** ($\bar{x}$), а среднее значение признака у единиц, которые подверглись выборочному наблюдению, **- выборочной средней** ($\bar{x}$). Здесь задача будет заключаться в том, чтобы на основе выборочной средней дать правильное представление о средней генеральной.

Между характеристиками выборочной совокупности и искомыми характеристиками генеральной совокупности, как правило, существует некоторое расхождение, которое называют **ошибкой.**

Общая величина возможной ошибки выборочной характеристики слагается из **ошибок регистрации и ошибок репрезентативности**.

Ошибки репрезентативности присущи только несплошным наблюдениям и представляют собой расхождение между величиной полученных по выборке показателей и величиной этих показателей, если бы они были получены при сплошном наблюдении, проведенном с той же степенью точности.

Ошибки репрезентативности могут быть

- систематическими

- случайными.

Систематические ошибки могут возникать в связи с особенностями принятой системы отбора и обработки данных наблюдения или в связи с нарушением установленных правил отбора.

Возникновение **случайных** ошибок репрезентативности объясняется недостаточно равномерным представлением в выборочной совокупности различных категорий единиц генеральной совокупности, в силу чего распределение отобранной совокупности единиц не вполне точно воспроизводит распределение единиц генеральной совокупности.

Классификация способов отбора единиц из генеральной совокупности.

Способы отбора единиц в выборочную совокупность классифицируются по различным признакам:

- число одновременно отбираемых единиц,
- метод отбора,
- число ступеней отбора,
- детализированность программы наблюдения.

По числу одновременно отбираемых единиц различаются индивидуальный, групповой и комбинированный отбор. При ***индивидуальном*** отборе выбираются единицы, ***групповом*** – качественно однородные группы или серии. ***Комбинированный*** отбор – сочетание индивидуального и группового.

Методы отбора

Для того чтобы выборка была репрезентативной, отбор единиц из генеральной совокупности должен быть соответствующим образом организован.

Исторически и логически первым сложился так называемый **собственно-случайный отбор**, то есть, отбор единиц из всей генеральной совокупности посредством жеребьевки или какого-либо иного способа (жеребьевки, таблица случайных чисел).

В практике выборочного наблюдения наиболее широко распространен **механический отбор**, который представляет собой последовательный отбор единиц через равные промежутки из определенного расположения их в генеральной совокупности. Промежутки определяются в соответствии с долей отбора (каждая пятая, десятая единица и т.д.). Принцип случайного отбора в механической выборке обеспечивается тем, что единицы генеральной совокупности располагаются в таком порядке, который не оказывает никакого влияния на поведение интересующего нас признака.

При этом расположение единиц генеральной совокупности в списке (или на месте наблюдения) может быть двояким

- упорядоченным или

- не упорядоченным относительно изучаемого признака.

Так, если нас интересует успеваемость студентов, то расположение их по алфавиту будет не упорядоченным, так как успеваемость никак не зависит от начальной буквы фамилии.

В статистической практике также часто применяется **районированный (типический, стратифицированный) отбор.**

Как правило, социально-экономические явления характеризуются большим разнообразием и не являются достаточно однородным в отношении изучаемых признаков. При наличии в составе генеральной совокупности различных типов явления с разными уровнями признаков надо так организовать выборку, чтобы обеспечить более равномерное представительство в выборочной совокупности различных частей (типов) явления.

Для этого общий список единиц генеральной совокупности в целом предварительно разбивается на отдельные списки, каждый из которых включает единицы, принадлежащие к одной однородной по определенному признаку группе (типу). В качестве типов (районов) могут быть взяты сложившиеся группы – республики, области, предприятия, цеха и т.д. или группы образованные специально (рыночные сегменты). Другими словами, **типическая выборка опирается на статистическую группировку** – по одному признаку или по комбинации нескольких. Из каждой выделенной группы в случайном порядке отбирается некоторое количество единиц.

Таким образом, при проведении типической выборки необходимо разбить общий объём выборки "**n**" между группами и определить число подлежащих наблюдению единиц в группе.

Это делается тремя способами:

1. наиболее часто применяется так называемое **пропорциональное размещение**, в этом случае количество отбираемых в выборку единиц пропорционально удельному весу данной группы в генеральной совокупности, при этом число наблюдений по группе определяется по формуле:

$$n_i = n \frac{N_i}{N}$$

где,

n_i - число наблюдений из i –й типической группы.

n - общий объём выборки,

N_i - объём i –й типической группы в ген. совокупности,

N - объём генеральной совокупности.

2. Возможен и другой вариант, когда из каждой группы отбирают одинаковое число единиц,

$$n_i = \frac{n}{k}$$

т.е. где k – число выделенных типических групп.

3. Третий вариант учитывает также и степень вариации признака в различных группах генеральной совокупности, а расчёт объёма выборки из каждой группы производится по формуле:

$$n_i = n\frac{N_i\sigma_i}{\sum N_i\sigma_i}$$

где σ_i - среднее квадратическое отклонение изучаемого признака в i–й группе.

Здесь пропорция отбора для групп с большой колеблемостью признака увеличивается, что в свою очередь приводит к соответствующему уменьшению возможной случайной ошибки в определении групповой средней.

Таким образом, при типическом отборе в выборку попадают представители всех типических групп, поэтому вероятность получить большую точность выборки здесь больше, чем при простой случайной выборке.

В практике выборочного наблюдения применяется **гнездовой (серийный, кластерный) отбор.**

В этом случае в случайном порядке отбираются не единицы, а целые гнёзда (серии) единиц совокупности, которые подвергаются сплошному обследованию.

Получающаяся в процессе этого отбора случайная ошибка выборки в подавляющем большинстве случаев больше, чем при любом другом способе отбора.

Особым видом выборочного наблюдения является **моментное наблюдение**. Суть его состоит в том, что на определенные моменты времени фиксируется наличие отдельных элементов изучаемого процесса.

Моментное наблюдение, в частности, применяется для изучения использования рабочего времени. В этих случаях в момент наблюдения фиксируется, находился ли работник (объект) в процессе работы или в простое. Моментное наблюдение охватывает всех работников фирмы (цеха) и в этом смысле является сплошным. Выборочное же оно потому, что охватывает не всё время работы цеха (смены), а лишь моменты, в которые осуществляется контроль.

Рассмотренные способы отбора, осуществляются путём **одноступенчатой** выборки.

Однако можно сформировать выборочную совокупность в два этапа:

сначала в случайном порядке выбираются подлежащие обследованию серии, а

затем из каждой отработанной серии в случайном порядке отбирается определённое количество единиц, подлежащих непосредственному наблюдению. **Ошибка** такой выборки будет зависеть от ошибки серийного отбора и от ошибки индивидуального отбора, т.е. такой отбор даёт, как правило, менее точные результаты, что объясняется возникновением ошибок репрезентативности на каждой ступени выборки.

При **многоступенчатом отборе** на всех ступенях, кроме последней, осуществляется только отбор, а наблюдение единиц производится только на последней ступени.

При многоступенчатой выборке единицы отбора на первых ступенях обычно представляют собой организационные единства единиц наблюдения и на разных ступенях применяются единицы отбора разных порядков.

Например, при текущем изучении бюджетов служащих единицей наблюдения является семья, формирование выборочной совокупности производится путём отбора сначала отраслей, потом предприятий, а затем лиц, работающих на предприятиях (членов семей).

Поэтому число ступеней отбора определяется числом типов единиц отбора, при этом на каждой последующей ступени единица отбора по своим масштабам уменьшаются и *только на конечной, единица отбора совпадает с единицей выборочной совокупности.*

Иногда в целях экономии средств удобно анализировать данные по некоторым интересующим нас признакам на основании изучения всех единиц выборочной совокупности, а по другим признакам – на основании части единиц выборочной совокупности, которые представляют подборку из единиц первоначальной выборки. Этот способ называют **двухфазным отбором.**

При наличии нескольких подвыборок можно говорить о **многофазном отборе**.

Многофазный отбор отличается от многоступенчатого тем, что при многофазном на каждой фазе пользуются одними и теми же единицами отбора, тогда как при многоступенчатом на разных ступенях применяются единицы отбора разных порядков.

Многофазным отбором пользуются в тех случаях, когда число единиц, необходимых для определения отдельных показателей с заданной точностью, весьма различно, как вследствие различий в степени колеблемости взаимосвязанных переменных, так и вследствие того, что для различных показателей требуется разная точность. Ошибки при многофазной выборке рассчитывают на каждой фазе отдельно.

Часто бывает целесообразно взять из изучаемой совокупности две или несколько независящих друг от друга выборок, применяя для получения каждой из них один и тот же способ отбора.

Такие выборки называют **взаимопроникающими**. Их преимущество в том, что они позволяют получить отдельные и независимые оценки тех или иных признаков изучаемой совокупности.

Все рассмотренные виды отбора (кроме механического) могут быть **повторными и бесповторными**.

Повторный – это такой отбор, при котором однажды попавшая в выборку единица генеральной совокупности при последующих испытаниях снова имеет возможность попасть в выборку.

При **бесповторном** отборе однажды попавшая в выборку единица не участвуют в последующих испытаниях.

При повторном отборе вероятность попасть в выборку для отдельной единицы совокупности в продолжение всего отбора не меняется, при бесповторном – эта вероятность изменяется после выбора каждой единицы.

Ошибки выборочного наблюдения.

Очевидно, что из генеральной совокупности можно сделать большое число одинаковых выборок, по которым расхождение фактической средней (или доли) с генеральной средней (или долей) будет случайным, так как каждая из выборок складывается под влиянием случайных факторов.

Можно предвидеть размеры этих расхождений. Фактические ошибки выборки могут быть оценены посредством сопоставления их со средними ошибками. Методами теории вероятностей установлено, что средняя ошибка выборки при изучении средних показателей определяется по формуле:

$$\mu_{\bar{x}} = \sqrt{\frac{\sigma_x^2}{n}}$$

где $\mu_{\bar{x}}$ - средняя ошибка выборки.

σ_x^2 - дисперсия признака x (варьирующего) в генеральной совокупности.

n - численность выборочной совокупности

При изучении долей признака (относительных показателей) формула средней ошибки имеет вид:

$$\mu_p = \sqrt{\frac{p(1-p)}{n}}$$

где,

μ_p - средняя ошибка доли;

p - доля признака в генеральной совокупности;

p(1 - p) - дисперсия доли изучаемого признака.

Средние ошибки выражаются в разных физических единицах, они различны по абсолютной величине. В статистике с целью сравнимости абсолютные величины представляются в относительном виде.

В теории выборки средние ошибки выражаются в известных стандартных единицах ***t*** (**коэффициент кратности ошибки, коэффициент доверия**), зависящий от вероятности с которой можно гарантировать, что предельная ошибка не превысит ***t*** – краткую среднюю ошибку.

Нас интересует количество баллов набранных студентами факультета на занятиях по предмету Х в семестре.
Предмет Х изучает 2000 человек.
Выборка - 200 чел.

Группы студентов по количеству полученных баллов, (x)	Число студентов, (f)	xf	$x - \bar{x}$	$(x - \bar{x})^2 f$	Число студентов в выборке
90	300	27000	-18	97200	24
100	600	60000	-8	38400	59
110	500	55000	+2	2000	52
120	400	48000	+12	57600	43
130	200	26000	+22	96800	22
Итого	2000	216000		292000	200

При правильно проведенной выборке в числе отображенных 200 студентов должны оказаться представители всех групп, численностью, соответствующей приблизительно 1/10 численности соответствующих групп в генеральной совокупности.

Рассчитаем среднее число баллов у 2000 студентов.

$$\bar{x} = \frac{\Sigma xf}{\Sigma f} = \frac{90 \cdot 300 + 100 \cdot 600 + 110 \cdot 550 + 120 \cdot 400 + 130 \cdot 200}{2000} = 108$$

Среднее число баллов по 200 студентам составит:

$$\tilde{x} = \frac{\Sigma xf}{\Sigma f} = \frac{90 \cdot 24 + 100 \cdot 59 + 110 \cdot 52 + 120 \cdot 43 + 130 \cdot 22}{200} = 109$$

Показатели по выборочной и генеральной совокупности могут совпадать лишь в редчайших случаях. Разница между ними при условии, что отбор в выборочную совокупность произведен правильно, и будет **случайной ошибкой выборки**. В нашем примере фактическая случайная ошибка = 1 балл (109-108).

При расчете относительных показателей ошибки репрезентативности представляют собой разность между долями одного признака в генеральной и выборочной совокупностях. Допустим, среди отобранных 200 студентов москвичей оказалось 90 чел., а среди 2000 – 940 чел.

Таким образом, доля москвичей в выборочной совокупности составляла 0,45 (90/200), а в генеральной совокупности – 0,47 (940/2000). Разность между 0,45 и 0,47 (0,02) является **фактической случайной ошибкой доли.**

Теперь воспользуемся приведёнными формулами и исчислим средние ошибки выборки по среднему баллу и доле москвичей среди 2000 студентов.

Вычислим дисперсию средней:

$$\sigma^2_{\bar{x}} = \frac{\Sigma(x-\bar{x})^2 f}{\Sigma f} = \frac{292\,000}{2\,000} = 146$$

таким образом

$$\mu_{\bar{x}} = \sqrt{\frac{\sigma^2_{\bar{x}}}{n}} = \sqrt{\frac{146}{200}} = \sqrt{0{,}73} \approx 0{,}85$$

Поскольку среди 2000 студентов 940 москвичей, то *p*=940/2000=0,47, откуда средняя ошибка доли будет:

$$\mu_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}47(1-0{,}47)}{200}} \approx 0{,}035 \quad \textit{или} \quad 3{,}5\%$$

Значит, если средний балл составил 109 , а средняя ошибка ($\mu_{\bar{x}}$) - 0,85 балла, то средний балл по генеральной совокупности можно ожидать в пределах $109 \pm 0{,}85$, то есть от 108,15 до 109,85.

Методами математической статистики и теории вероятностей доказано, что генеральная средняя не выйдет за пределы, равные величине одной средней ошибки, не во всех возможных выборках, а лишь в **6827** выборках из **10000**, то есть сформулированное положение об ожидаемых пределах можно утверждать лишь с вероятностью **0,6827**.

Величина вероятности (p) 0,6827 представляет собой величину интеграла вероятности Лапласа (Ф (t)) при t = 1.

Вероятность ***p*** иногда называют **доверительной вероятностью**, при обследовании общественных явлений её величину обычно принимают в пределах от **0,90** до **0,99**.

Итак, при вероятности 0,6827 (t=1) в 6827 выборках из 10000 фактическая ошибка не превысит 0,85 балла. По нашей же выборке из 200 студентов **фактическая ошибка** составила 1 балл, то есть эта выборка входит в число остальных 3173 выборок из 10000. Такая вероятность вряд ли может устроить. Чтобы повысить её, приходится расширять пределы возможных ошибок, увеличивая t. Например, при t=2 генеральная средняя не выйдет уже за пределы, равные двум средним ошибкам; вероятность этого утверждения повышается до 0,9545, а при t=3 вероятность становится 0,9973.

Это означает, что в 9545 из 10000 выборок средний балл по генеральной совокупности не выйдет за пределы

109 2 0,85 = 109 1,7, ±

а в 9973 выборках не выйдет за пределы

109 3 ± 0,85=109 2,55.

Результаты расчетов уже при вероятности 0,99 (t=2,58) можно считать практически достоверными.

Средняя ошибка, умноженная на принятый коэффициент доверия (t), носит название **предельной ошибки выборки** Δ.

$$\Delta_{\bar{x}} = t \cdot \mu_{\bar{x}}; \qquad \Delta_p = t \cdot \mu_p$$

Для величины t и p составлены подробные таблицы соответствующих вероятностей. Приведенные выше формулы относятся к повторному собственно-случайному отбору. При **бесповторном** отборе в подкоренное выражение вводится дополнительный множитель **(*1-n/N*),** где ***n/N*** – доля отобранных единиц в генеральной совокупности:

Следовательно для бесповторной выборки формулы предельных ошибок примут вид для средней:

$$\Delta_{\bar{x}} = t\sqrt{\frac{\sigma^2_{\bar{x}}}{n}\left(1-\frac{n}{N}\right)}$$

для доли

$$\Delta_{p} = t\sqrt{\frac{p(1-p)}{n}\left(1-\frac{n}{N}\right)}$$

Факторы, определяющие величину предельной ошибки:

1) **Вариация (колеблемость)** изучаемого признака $\sigma^2_{\bar{x}}$ и ***p*(*1-p*).** Размер ошибки прямо пропорционален величине колеблемости признаков.

2) **Вероятность**, с которой исследователь желает получить пределы ошибок. Чем выше заданная вероятность, тем больше коэффициент доверия и соответственно больше предельная ошибка выборки.

3) **Способ отбора единиц в выборочную совокупность** (повторный или бесповторный), так как общий объем выборки (*n*) всегда меньше объема генеральной совокупности (*N*), то дополнительный множитель **(*1-n/N*)** всегда меньше 1. Значит ошибка выборки, при бесповторном отборе всегда будет меньше, чем при повторном отборе. В то же время при сравнительно небольшом проценте выборки этот множитель близок к единице. Например, при 5 % - ой выборке он равен 0,95.

4) **Численность выборки**. Ошибка выборки зависит *в большей* степени от абсолютной численности выборки и в *меньшей* – от её относительной доли (процента выборки).

Предположим, что производится 225 наблюдений в первом случае из генеральной совокупности в 4500 единиц и во втором – из генеральной совокупности в 225000 единиц.

Пусть дисперсия в обоих случаях равна 25. Тогда в первом случае при 5 %-ом отборе ошибка выборки составит:

$$\mu=\sqrt{\frac{25}{225}(1-\frac{225}{4500})}=\sqrt{0{,}1045}=0{,}323$$

Во втором случае при 0,1 %-ом отборе $\left(\frac{225}{225\,000}\cdot 100\right)$ она будет равна:

$$\mu=\sqrt{\frac{25}{225}(1-\frac{225}{225\,000})}=\sqrt{0{,}10989}=0{,}331$$

Хотя во втором случае процент выборки уменьшился в 50 раз, ошибка выборки увеличилась очень незначительно, так как численность выборки не изменилась.

Предположим теперь, что численность выборки увеличили до 625 наблюдений при генеральной совокупности в 225000 единиц. В этом случае ошибка выборки будет равна:

$$\mu = \sqrt{\frac{25}{625}\left(1 - \frac{625}{225\,000}\right)} = \sqrt{0{,}03988} = 0{,}2$$

Таким образом, увеличив численность выборки в 2,8 раза при той же численности генеральной совокупности в 225000 единиц, мы снизили размеры ошибки более чем в 1,6 раза.

Ошибка выборки в этом случае будет также практически в 1,6 раза меньше, чем в первом случае, когда было отобрано 225 единиц из 4500, хотя там применялся 5%-й отбор, а здесь всего лишь около 0,3%-ый.

Любая формула предельной ошибки принципиально позволяет решать задачи трех видов:

1) определить предел возможной ошибки средней (доли), т.е. насколько может отклониться показатель выборочной совокупности от показателя в генеральной совокупности;

2) определить необходимую численность выборки, при которой пределы возможной ошибки не превысят некоторой заданной величины;

3) определить вероятность того, что в проведенной выборке ошибка будет заключаться в заданных пределах.

Решение той или иной из поставленных задач зависит от того, какие из переменных величин, входящих в формулу, известны, а какие нет. При этом σ , p, N во всех случаях являются величинами постоянными, так как они заданы действительностью.

Продолжим рассмотрение примера с выборкой 200 студентов из 2000 при условии, что отбор студентов произведен механическим способом (то есть он бесповторный). Пусть теперь на основании того, что нам известно надо с вероятностью 0,9545 определить, в каких пределах можно ожидать средний балл для 2000 студентов. Как установлено выше, средний балл в выборке составляет 109. Определить нужно предельную ошибку среднего балла ($\Delta_{\bar{x}}$)

Известно, что $n = 200$, $N = 2000$, $\sigma_{\bar{x}}^2 = 146$, $t = 2$ (при $p = 0{,}9545$), тогда

$$\Delta_{\bar{x}} = t\sqrt{\frac{\sigma_{\bar{x}}^2}{n}\left(1-\frac{n}{N}\right)} = 2\sqrt{\frac{146}{200}\left(1-\frac{200}{2000}\right)} \approx 1{,}62$$

Таким образом, с вероятностью 0,9545 можно утверждать, что по совокупности 2000 студентов средний балл будет находиться в пределах $109 \pm 1{,}62$.

Предположим, что предел $\pm$ 1,62 нас не устраивает. Можно подсчитать, какую численность выборки следует взять, чтобы предельная ошибка не превышала, например, 1 ($\sigma^2_{\bar{x}}$, ***N, t*** – остаются без изменений). Тогда из формулы $\Delta_{\bar{x}}$ находим общий объем выборки **(*n*)**.

$$n = \frac{t^2\sigma^2_{\bar{x}}N}{N\Delta^2_{\bar{x}} + t^2\sigma^2_{\bar{x}}} = \frac{4 \cdot 146 \cdot 2000}{2000 \cdot 1 + 4 \cdot 146} \approx 452 чел.$$

С другой стороны уменьшения предела ошибки с 1,62 до 1 можно добиться уменьшением *t* и связанной с ним вероятности (в этом случае без изменений остаются $\sigma_{\bar{x}}^2$, *n*, ***N***). Из формулы $\Delta_{\bar{x}}$ теперь находим ***t.***

$$t = \frac{\Delta_{\bar{x}}}{\sqrt{\frac{\sigma_{\bar{x}}^2}{n}(1-\frac{n}{N})}} = \frac{1}{\sqrt{\frac{146}{200}(1-\frac{200}{2000})}} = \frac{1}{0.81} \approx 1{,}23$$

Вероятность при этой величине t равна 0,7813. Это означает, что из 10000 выборок ошибка не превысит 1 в 7813 из них.

Далее в условиях механического отбора с вероятностью 0,9545 рассчитаем пределы, в каких должна оказаться в генеральной совокупности доля москвичей. В выборочной совокупности она составляет 0,45.

Нам известно, что $N = 2000$, $n = 200$, $p = 0,47$, при вероятности 0,9545 ($t = 2$). Находим предельную ошибку:

$$\Delta_p = t\sqrt{\frac{p(1-p)}{n}\left(1-\frac{n}{N}\right)} = 2\sqrt{\frac{0,47(1-0,47)}{200}\left(1-\frac{200}{2000}\right)} \approx 2 \cdot 0,0335 \approx 0,067$$

Это означает, что доля москвичей в генеральной совокупности находится в пределах 0,450± 0,067 (в процентах получится 45%± 6,7). Таким образом, с вероятностью 0,9545 можно утверждать, что удельный вес числа москвичей в генеральной совокупности находится в пределах от 38,3 до 51,7 %.

До сих пор при решении задачи на выборку 200 студентов из 2000 в формулах ошибок выборки показатели вариации брались по генеральной совокупности $\sigma^2_{\bar{x}}$ и $p\ (1 - p)$]. В принципе (по теории) это так и должно быть. Однако в процессе расчетов по выборке этими данными статистик не располагает (более того, выборка для того и проводится, чтобы определить показатель по генеральной совокупности). Поэтому **на практике вместо показателей вариации генеральной совокупности приходится пользоваться выборочными показателями вариации.** Определим их.

Ранее среднюю величину изучаемого признака в выборке мы обозначили через $\widetilde{x}$ теперь же долю признака в выборке обозначим через w. Тогда показатели вариации в выборочной совокупности получат выражение $\sigma^2_{\tilde{x}}$ и $\omega(1-\omega)$ а предельные ошибки соответственно $\Delta_{\tilde{x}}$ и Δ_{ω}

Численные различия средних и предельных ошибок выборки, рассчитанные по показателям вариации генеральной и выборочной совокупностей незначительны.

В математической статистике доказывается, что:

$$\sigma^2_{\bar{x}} = \sigma^2_{\tilde{x}}\left(\frac{n}{n-1}\right) \qquad и \qquad p(1-p) = w(1-w)\left(\frac{n}{n-1}\right)$$

В случае выборки большого размера поправочный коэффициент $n/(n-1)$ близок к 1 и им пренебрегают, и учитывают этот коэффициент лишь в выборках малого размера.

В нашем случае с 200 студентами мы имеем данные о вариации признаков по выборочной совокупности. Вычисленные по этим данным предельные ошибки составляют: $\Delta_{\tilde{x}} = 1{,}60, \quad \Delta_w = 0{,}0668$

Следовательно, пределы ошибок, исчисленные по вариации признаков в генеральной и выборочной совокупностях, различаются очень мало:

- по среднему числу посещений разница = 0,02 (1,62 – 1,60)
- по доле москвичей - 0,0002 (0,0670 – 0,0668).

Таким образом, пределы, в которых находится величина показателя по генеральной совокупности устанавливаются следующим образом: сначала находятся предельные ошибки, а затем эти ошибки прибавляются и вычитаются из выборочного показателя:

$$\bar{x} = \tilde{x} \pm \Delta_{\tilde{x}}; \qquad \tilde{x} - \Delta_{\tilde{x}} \le \bar{x} \le \tilde{x} + \Delta_{\tilde{x}}$$

$$p = w \pm \Delta_w; \qquad w - \Delta_w \le p \le w + \Delta_w$$

Так решаются задачи при собственно-случайном и механическом отборах.

Теория средних и предельных ошибок, рассмотренная выше справедлива для *обычных выборок достаточно большого объема*. Однако такие выборки не всегда возможны и необходимы.

Поэтому наряду с ними приходится пользоваться с так называемыми **малыми выборками** ($n<30$).

Первые работы в области теории малой выборки были сделаны английским статистиком В.С. Госсетом в 1908г. (псевдоним – "Стьюдент").

Формулы для определения предельных ошибок малой выборки для повторного отбора будут такими:

$$\Delta_{\tilde{x}\ м.в.} = t\sqrt{\frac{\sigma^2_{\tilde{x}}}{n}\left(\frac{n}{n-1}\right)} = t\sqrt{\frac{\sigma^2_{\tilde{x}}}{n-1}} \qquad \Delta_{w\ м.в.} = t\sqrt{\frac{w(1-w)}{n-1}}$$

Для оценки возможных пределов ошибки малой выборки пользуются так называемым отношением Стьюдента:

$$t = \frac{\tilde{x} - \bar{x}}{\mu_{\tilde{x}(м.в.)}} \text{ или } t = \frac{w-p}{\mu_{w(м.в.)}} \text{ где } \mu_{\tilde{x}(м.в)} = \sqrt{\frac{\sigma^2_{\tilde{x}}}{n-1}} \quad \mu_{w(м.в.)} = \sqrt{\frac{w(1-w)}{n-1}}$$

Исчисление ошибок малой выборки по данным формулам может быть удовлетворительным при условии, что распределение изучаемого признака в генеральной совокупности нормально или близко к нему.

Организация и практика применения выборочного метода наблюдения.

Организация выборочного наблюдения предполагает решение нескольких вопросов:

1. Определение единиц отбора.

2. Определение вида отбора.

3. Определение численности выборочной совокупности.

1. Определение единиц отбора.

Единицы для исследования отбираются из определенного круга явлений, составляющих "основу выборки". В качестве такой "основы" могут выступать списки отдельных лиц, домохозяйств, жилищ, планы городов, карты сельских районов, списки населенных пунктов и т.д. Единица отбора в выборке не должна быть меньше единицы наблюдения. Например, если в качестве единиц наблюдения выступают предприятия, то единицами отбора не могут быть бригады вида отбора.

2. Определение вида отбора.

Главный критерий при этом – величина ошибки и простота организации отбора.

3. **Определение численности выборочной совокупности.**

Как известно в формулах предельных ошибок лишь ***n***, Δ и ***t*** выступают переменными величинами. Однако и они по своей природе не одинаковы: Δ и ***t*** определяются природой изучаемого явления и задачами, стоящими перед исследованием, и лишь **n** является собственно неизвестной.

Действительно, приступая к выборке необходимо хотя бы ориентировочно знать допустимые пределы, в которых могут находиться возможные ошибки предстоящей выборки, а также степень вероятности, с которой эти пределы должны быть гарантированы. Исходя из требований исследования определяют и величину **t**: чем более достоверные данные мы хотим получить, тем большую величину **t** и связанную с ней вероятность необходимо задать. В социально-экономических исследованиях **t** берут обычно в пределах от 2 до 3, что соответствует вероятности от 0,954 до 0,997.

Далее расчет численности выборки сталкивается со следующими трудностями: хотя $\sigma^2_{\bar{x}}$ и $p(1 - p)$ или $\sigma^2_{\tilde{x}}$ и $w(1-w)$ заданы действительностью, но к началу выборочного наблюдения они не известны ни по генеральной, ни по выборочной совокупности. Нахождение этих величин при организации выборки является одной из труднейших задач. Как решается эта задача?

Вместо действительных $\sigma^2_{\tilde{x}}$ и $p(1 - p)$ в формулах предельных ошибок приходится использовать некие приближенные величины.

σ (*сигма*) приближённо можно определить следующими путями:

1) используется установленная ранее по данным какого-либо предыдущего наблюдения. Однако это целесообразно лишь в случае, когда за время, прошедшее после предыдущего наблюдения не произошло существенных изменений;

2) математическая статистика доказывает, что средняя (стандартная) ошибка может быть определена из формулы

$$t = \frac{\tilde{x} - \bar{x}}{\sigma}$$

при $t=1$ $\tilde{x} - \bar{x} = \sigma$

при $t=2$ $\tilde{x} - \bar{x} = 2\sigma$

при $t=3$ $\tilde{x} - \bar{x} = 3\sigma$

Так как вероятность при t=3 достигает 0,9973, то считается, что отклонение $\tilde{x}$ от $\bar{x}$ в пределах $\pm 3\sigma$ вполне гарантирует удовлетворительное решение подавляющего большинства задач *("правило трех сигм")*. Следовательно, весь размах **(R)** между **min** и **max** значениями признака может быть принят за 6σ Значит, если известен размах значений признака по изучаемому явлению, то σ можно принять за 1/6 часть его: σ=1/6R. Для большей гарантии за 1/5R;

3) практика показывает, что во многих явлениях колебания вариации происходят в промежутке от 25 до 35 %, то есть составляет примерно 1/4-1/3 Поэтому, если в совокупности с нормальной колеблемостью признака известно $\bar{x}$ то условно $\bar{x}$ можно принять за 1/4 или 1/3 σ

При установлении колеблемости доли, как и средней, в первую очередь надо попытаться найти ориентировочные данные о величине *p*. Если их нет, то берётся максимальная величина произведения ***p(1 - p)*** **равная 0,25.**

Теперь из формул пределов ошибок для собственно-случайных и механических отборов находим в общем виде численность выборочной совокупности.

При повторном отборе: для средней $$n = \frac{t^2 \sigma_{\bar{x}}^2}{\Delta_{\bar{x}}^2}$$

для доли $$n = \frac{t^2 p(1-p)}{\Delta_p^2}$$

При бесповторном: для средней $$n = \frac{t^2 \sigma_{\bar{x}}^2 N}{N\Delta_{\bar{x}}^2 + t^2 \sigma_{\bar{x}}^2}$$

для доли $$n = \frac{t^2 p(1-p) N}{N\Delta_p^2 + t^2 p(1-p)}$$