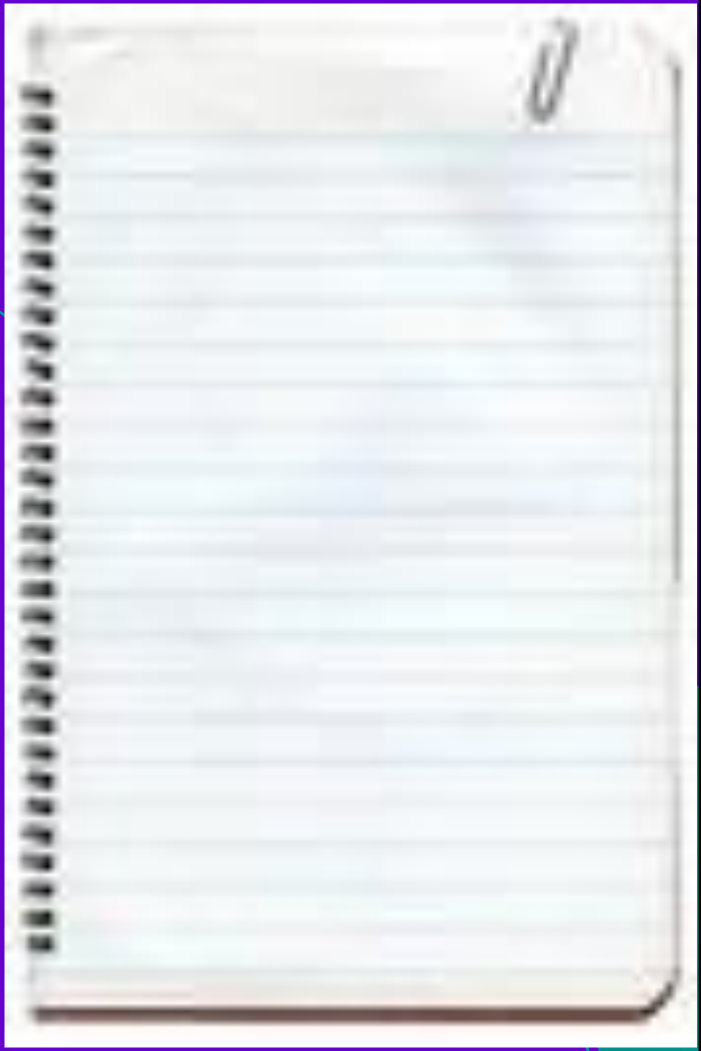
Общая теория статистики

Тема

Выборочное наблюдение



Найдите: медиану стажа и среднюю зарплату для коллектива из 5 человек

Номер	Стаж, лет	Зарплата, т.р.
1	5	№ в списке
2	12	8
3	№ в списке	15
4	20	22
5	25	19

Определение выборочного наблюдения

- Выборочное наблюдение — это способ несплошного статистического наблюдения, при котором обследуются не все единицы изучаемой (генеральной*)* совокупности, а лишь часть ее (выборка*)*, отобранная по определенным правилам (научно) и обеспечивающая получение данных, характеризующих совокупность в целом.

Причины применения:

- Экономия
- Невозможность проведения сплошного исследования

Основные обозначения

- **N – объем, численность, число единиц ГС**
- **n – объем ВС**

$$\bar{x} - генеральная\_средняя;$$

$$w_{ген} - доля\_единиц, обладающих$$

$$признаком\_в\_ГС$$

Основная идея выборочного метода состоит в том, что в результате обследования части совокупности можно судить с определенной

Для того, чтобы выборочная совокупность давала объективные результаты, она должна быть репрезентативной (каждая единица генеральной совокупности должна иметь равную возможность попасть в выборку). Тогда с увеличением объема выборки характеристики выборочной совокупности будут приближаться к характеристикам генеральной совокупности.

Теоретической основой выборки являются теоремы закона больших чисел (Чебышева, Ляпунова, Бернулли и др.)

Задачи выборочного метода

- Определение доверительного интервала, в котором находится характеристика генеральной совокупности
- Определение минимального объема выборки
- Определение доверительной вероятности того, что разность между характеристиками выборочной и генеральной совокупностей не превзойдет наперед заданного числа

Пример. Имеются данные о зарплате рабочих в у. е.

Группы по з/пл.	ГС - человек	Из них попали в выборку
10-13	100	5
13-16	150	10
16-19	400	30
19-22	200	45
22-25	150	10
Итого	**1000**	**100**

Сходство ГС и ВС

- Из теорем Чебышева, Ляпунова и закона больших чисел следует:

Хотя каждая выборочная средняя отличается от генеральной, среднее значение по ним равно генеральной:

$$\frac{\sum \tilde{x}}{n} = \overline{x}$$

ВЫВОД.

Реально наблюдаемая совокупность объектов, статистически представленная рядом наблюдений $x_1, x_2,\ldots, x_n$ случайной величины X, является *выборкой*, а гипотетически существующая (домысливаемая) – *генеральной совокупностью*.

В основе решения задач на выборочный метод лежат формулы предельных ошибок выборки

$$\Delta = t\mu$$

Обозначения

t - число, связанное с вероятностью через табл. закона распределения

μ - средняя ошибка выборки

Δ - предельная ошибка

Ошибки выборки

$$|\bar{x} - \bar{x}_0| = \varepsilon_{\bar{x}}$$ - ошибка средней

$\bar{x}_0$ - генеральная средняя

$$|W - W_0| = \varepsilon_w$$ - ошибка доли

W_0 - генеральная доля

Характеристики выборочной совокупности

$$\overline{x} = \frac{\sum x_i \cdot n_i}{n}$$ - выборочная средняя

$$\sigma^2 = \frac{\sum \left(x_i - \overline{x}\right)^2 \cdot n_i}{n}$$ - выборочная дисперсия

$$W = \frac{m}{n}$$ - выборочная доля

Объем выборки

- Число наблюдений ***n***, образующих выборку, называется ***объемом выборки***. Если объем выборки ***n*** достаточно велик ($n \to \infty$), выборка считается ***большой***, в противном случае она называется выборкой ***ограниченного объема***.

Малой считается выборка,

в которую входит

меньше 30 единиц.

Рассмотрим особенности малой выборки.

Если мы работаем с обычной выборкой, то используется таблица «Интеграла вероятностей закона нормального распределения».

В случае малой выборки необходимо пользоваться таблицей «Распределение Стьюдента», при этом число степеней свободы равно:

$$\boxed{K = n - 1}$$

Объем выборки

Выборка считается ***малой***, если при измерении одномерной случайной величины *X* объем выборки не превышает 30 (n <= 30), а при измерении одновременно нескольких (k) признаков в многомерном пространстве отношение n к k не превышает 10 ($n/k < 10$).

Условия проведения выборки

Выборка будет представлять всю совокупность с приемлемой точностью при выполнении двух условий.

Условия проведения выборки

Во-первых, она должна быть достаточно ***многочисленной****, чтобы в ней могли проявиться закономерности, существующие в генеральной совокупности.*

Условия проведения выборки

Во-вторых, элементы выборки должны быть отобраны ***объективно****, независимо от воли исследователя, чтобы каждый из них имел одинаковые шансы быть отобранным или чтобы эти шансы были известны исследователю.*

характеристика выборочного наблюдения

Генеральная совокупность может быть конечной (число наблюдений $N = const$) или бесконечной ($N = \infty$), а выборка из генеральной совокупности – это всегда результат ограниченного ряда n наблюдений.

Способы отбора

- По виду различают индивидуальный, групповой и комбинированный отбор. При **индивидуальном** отборе в выборочную совокупность отбираются отдельные единицы генеральной совокупности, при **групповом отборе** – группы единиц, а **комбинированный отбор** предполагает сочетание группового и индивидуального отбора.

Виды и схемы отбора

Процесс образования выборочной совокупности называется **отбором**. Он осуществляется в порядке беспристрастного, случайного отбора единиц из генеральной совокупности.

Существуют пять основных ***способов отбора***

Простой случайный отбор

при котором *n* объектов случайно извлекаются из генеральной совокупности *N* объектов (например с помощью таблицы или датчика случайных чисел), причем каждая из возможных выборок имеют равную вероятность. Такие выборки называются *собственно-случайными*.

Случайная выборка

- Случайная выборка - основа всех других способов отбора.
- Случайная выборка осуществляется методом жеребьевки: все единицы совокупности нумеруются, номера записываются на карточки, а потом отбираются.
- **Отбор может быть повторным и бесповторным**.

Формулы предельных ошибок выборки

	Повторный отбор	Бесповторный отбор
Для средней	$t \cdot \sqrt{\frac{\sigma^2}{n}}$	$t \cdot \sqrt{\frac{\sigma^2}{n} \cdot (1 - \frac{n}{N})}$
Для доли	$t \cdot \sqrt{\frac{w \cdot (1 - w)}{n}}$	$t \cdot \sqrt{\frac{w \cdot (1 - w) \cdot (1 - \frac{n}{N})}{n}}$

Обозначения:

- σ^2 - выборочная дисперсия;
- W - выборочная доля;
- n - объем выборочной совокупности;
- N - объем генеральной совокупности;
- t - число, связанное с вероятностью, которая берется из таблицы интеграла вероятностей закона распределения.

Пример

Для определения среднего срока службы изделий было обследовано 250 изделий. При этом средний срок службы был установлен на уровне 41,9 месяца. Среднее квадратическое отклонение равно 6,2 месяцам.

С вероятностью 0,9973 определить, в каких пределах находится средний срок службы всех изделий

Решение:

- P=0,9973, t=3 (из таблицы интеграла вероятностей закона нормального распределения).
-

$$\Delta_{\bar{x}} = 3 \cdot \sqrt{\frac{6{,}2^2}{250}} = 1{,}2\text{мес}$$

$$\bar{x} - \Delta_{\bar{x}} \le \overline{x_0} \le \bar{x} + \Delta_{\bar{x}}$$

$$41{,}9 - 1{,}2 \le \overline{x_0} \le 41{,}9 + 1{,}2$$

$$40{,}7\,мес \le \overline{x_0} \le 43{,}1\text{мес}$$

Пример

- Определить вероятность того, что предельная ошибка средней службы не превысит 1 месяц.

Решение:

$$\Delta_{\bar{x}} = t \cdot \sqrt{\frac{\sigma^2}{n}}$$

$$t = \frac{\Delta_{\bar{x}}}{\sqrt{\frac{\sigma^2}{n}}} = \frac{1}{\sqrt{\frac{6{,}2^2}{250}}} = 2{,}55$$

$$P = 0{,}9892$$

Пример

Определение минимального объема выборки.

Сколько следует прохронометрировать операций, чтобы с вероятностью 0,9973 можно было бы утверждать, что разность между средней продолжительностью операций в выборочной и генеральной совокупности не превысит 1 секунды, если по результатам предыдущего испытания установлено, что средняя продолжительность операции равна 30 секундам, а среднее квадратическое отклонение равно 7 секундам?

Решение :

$$n = ?$$

$$\Delta = 1$$

$$n = \frac{t^2 \cdot \sigma^2}{\Delta^2_{\bar{x}}} = \frac{3^2 \cdot 7^2}{1^2} = 441$$

Ответ: нужно прохронометрировать не менее 441 операции.

Простой отбор с помощью регулярной процедуры

осуществляется с применением механической составляющей (номера квартиры, даты, дня недели, буквы алфавита) и полученные таким способом выборки называются *механическими*.

Стратифицированный отбор

заключается в том, что генеральная совокупность объема N подразделяется на части совокупности или слои (страты) объема N_1, N_2, … , N_r, так что $N_1 + N_2 + \ldots + N_r = N$.

<u>**Стратифицированный** отбор</u>

Страты - однородные объекты с точки зрения статистических характеристик (например, население по возрасту делится на две страты – в трудоспособном и нетрудоспособном возрасте; банки – по размеру капитала). В этом случае выборки называются *стратифицированными* (*расслоенными*, *типическими*, *районированными*).

Серийный отбор

- Приемы **серийного** отбора используются для формирования *серийных или гнездовых выборок*. Они удобны в том случае, если необходимо обследовать сразу "блок" или серию объектов (например, партию товара, продукцию определенной серии или предприятия территориально-административной единицы).

Вся совокупность делится на серии, после чего механическим или собственно случайным способом отбирается некоторое количество серий. Все единицы совокупности, входящие в отобранные серии, подвергаются сплошному контролю.

Метод отбора Выборка	Повторный		Бесповторный	
	для средней	для доли	для средней	для доли
Серийная (гнездовая)	$t\sqrt{\frac{\delta_x^{\,2}}{r}}$	$t\sqrt{\frac{\delta_w^{\,2}}{r}}$	$t\sqrt{\frac{\delta_x^{\,2}}{r}\left(1-\frac{r}{R}\right)}$	$t\sqrt{\frac{\delta_w^{\,2}}{r}\left(1-\frac{r}{R}\right)}$

r – количество отобранных серий

R – общее число серий

$$\delta^2_{\bar{x}} = \frac{\Sigma(\bar{x}_i - \bar{x})^2}{r} \qquad \bar{x} = \frac{\Sigma \bar{x}_i}{r}$$

$\delta^2_{\bar{x}}$ - межсерийная дисперсия

$$\delta^2_W = \frac{\Sigma(W_i - \bar{W})^2}{r} \qquad \bar{W} = \frac{\Sigma W_i}{r}$$

δ^2_W - межсерийная выборочная дисперсия для доли

W_i - доля изучаемого признака в i-той группе

$\bar{W}$ - средняя выборочная доля изучаемого признака

Пример:

На предприятии 10 бригад. Изучается производительность труда. Отбираются 2 бригады. Средняя производительность труда 1-й бригады – 4,6 тонны, а 2-й – 3 тонны. С вероятностью 0,9973 определить пределы в кот. будет находиться средняя производительность труда рабочих данного предприятия.

$t = 3$

$$\bar{x} = \frac{4{,}6+3}{2} = 3{,}8m.$$

$$\delta_{\bar{x}}^2 = \frac{(4{,}6-3{,}8)^2 + (3-3{,}8)^2}{2} = 0{,}64$$

$$\Delta_{\bar{x}} = 3 \cdot \sqrt{\frac{0{,}64}{2} \cdot (1 - \frac{2}{10})} = 1{,}5m.$$

$$3{,}8 - 1{,}5 \le \bar{x}_0 \le 3{,}8 + 1{,}5$$

$$2{,}3 \le \bar{x}_0 \le 5{,}3$$

ОТВЕТ: $2{,}3 \le \bar{x}_0 \le 5{,}3$

Типическая выборка

способ проведения типической выборки:

1. вся совокупность делится на типические группы

пример

население → сельское

население → городское

2. из каждой типической группы отбирается некоторое количество единиц

Отбор может быть как ***пропорциональным*** объёму типических групп, так и ***непропорциональным***

Объем выборки

При отборе, пропорциональном объему типических групп, число наблюдений по каждой группе определяется по формуле:

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

n_i -объем выборки из i -й типической группы.

n -общий объем выборки.

N_i -объем i -й типической группы в генеральной совокупности.

N -объем генеральной совокупности.

Типическая выборка: формулы

Выборка \ Метод отбора	Повторный		Бесповторный	
	Для средней	Для доли	Для средней	Для доли
Типическая (при отборе пропорциональном объему групп)	$t \cdot \sqrt{\frac{\overline{\sigma_i^2}}{n}}$	$t \cdot \sqrt{\frac{\overline{w_i \cdot (1 - w_i)}}{n}}$	$t \cdot \sqrt{\frac{\overline{\sigma_i^2}}{n} \cdot (1 - \frac{n}{N})}$	$t \cdot \sqrt{\frac{\overline{w_i \cdot (1 - w_i)}}{n} \cdot (1 - \frac{n}{N})}$

Типическая выборка: пример

Задача. Определим средний возраст мужчин, вступающих в брак, произведя 5%-ю типическую выборку:

соц. группа	число мужчин n_i	средний возраст $\bar{x}_i$	ср. кв. отклонение σ_i	доля мужчин, вступающих во второй брак. w_i
рабочие	60	24	5	0.10
служащие	40	27	8	0.20

С вероятностью 0,954 определить

1) пределы, в которых будет находиться средний возраст мужчин, вступающих в брак
2) долю мужчин, вступающих в брак во второй раз.

Типическая выборка: пример

Решение. 1) Средний возраст вступления в брак мужчин находится в пределах

$$\bar{x} - \Delta_{\bar{x}} \le \bar{x}_0 \le \bar{x} + \Delta_{\bar{x}}$$

$$\bar{x} = \frac{\sum \bar{x}_i n_i}{\sum n_i} = \frac{24 \cdot 60 + 27 \cdot 40}{100} = 25{,}2\ \textit{года};$$

$$\overline{\sigma_i^2} = \frac{\sum \sigma_i^2 n_i}{\sum n_i} = \frac{25 \cdot 60 + 64 \cdot 40}{100} = 40{,}6$$

Решение примера типической выборки

$$\Delta_{\bar{x}} = 2 \cdot \sqrt{\frac{40,6}{100}(1-0,05)} = 1,2\ \textit{года}$$

Таким образом, с вероятностью 0,954 можно утвердить, что средний возраст мужчин, вступающих в брак, принимает значения **25,2 ± 1,2** года,
или $24 \le \bar{x} \le 26,4$

Типическая выборка: пример

Решение. 2) Доля мужчин, вступающих в брак во второй раз, находится в пределах

$$\overline{w} - \Delta_w \le w_0 \le \overline{w} + \Delta_w$$

$$\overline{w} = \frac{\sum w_i n_i}{\sum n_i} = \frac{0{,}1 \cdot 60 + 0{,}2 \cdot 40}{100} = 0{,}14 = 14\%$$

$$\overline{w_i(1 - w_i)} = \frac{\sum w_i(1 - w_i)n_i}{\sum n_i} = \frac{(0{,}1 \cdot 0{,}9) \cdot 60 + (0{,}2 \cdot 0{,}8) \cdot 40}{100} = 0{,}098$$

Вывод по примеру типической выборки

$$\Delta_w = 2 \cdot \sqrt{\frac{0,098}{100}(1-0,05)} = 0,06 = 6\%$$

Таким образом, с вероятностью 0,954 можно утверждать, что доля мужчин, вступающих в брак во второй раз, принимает значения **14% ± 6%**, или

$$8\% \leq p \leq 20\%$$

Комбинированный (ступенчатый) отбор

может сочетать в себе сразу несколько способов отбора (например, стратифицированный и случайный или случайный и механический); такая выборка называется *комбинированной*.

Методы отбора

По ***методу*** отбора различают *повторную и бесповторную* выборку. *Бесповторным* называется отбор, при котором попавшая в выборку единица не возвращается в исходную совокупность и в дальнейшем выборе не участвует; при этом численность единиц генеральной совокупности *N* сокращается в процессе отбора.

При **повторном отборе** попавшая в выборку единица после регистрации возвращается в генеральную совокупность и таким образом сохраняет равную возможность наряду с другими единицами быть использованной в дальнейшей процедуре отбора; при этом численность единиц генеральной совокупности N остается неизменной (метод в социально-экономических исследованиях применяется редко). Однако, при большом N ($N \to \infty$) формулы для бесповторного отбора приближаются к аналогичным для повторного отбора и практически чаще используются последние ($N = const$).

Механическая выборка

При механической выборке вся совокупность делится на группы по числу единиц, которые должны войти в выборку, после чего из каждой группы отбирается 1 единица. Таким образом механическая выборка является бесповторной. Для механической выборки применяются формулы собственно-случайного, бесповторного отбора

Механическая выборка применяется в случаях, когда генеральная совокупность каким либо образом , упорядочена , т.е. имеется определенная последовательность в расположении единиц (табельные номера работников, избирательные списки, телефонные номера респондентов, номера домов и квартир и т.п.)

Характеристики генеральной и выборочной совокупности

- В основе статистических выводов проведенного исследования лежит распределение случайной величины X, наблюдаемые же значения $(x_1, x_2, \ldots, x_n)$ называются **реализациями** случайной величины X (n – объем выборки).

Характеристики генеральной и выборочной совокупности

Распределение случайной величины *X* в генеральной совокупности носит теоретический, идеальный характер, а ее выборочный аналог является *эмпирическим* распределением.

Выборочная доля

- **Отношение числа единиц, обладающих данным признаком или данным его значением m, к общему числу единиц выборочной совокупности** n **называется <u>выборочной долей</u>** w:
- $w = m/n$.

Пример

В партии товара, содержащей 10 тыс. штук, при 4% выборке доля выборки k_n в абсолютной величине составляет 400 шт. ($n = N \times 0{,}04$); если же в этой выборке обнаружено 12 бракованных изделий, то выборочная доля брака w составит 0,03 ($w = 12/400 = 0{,}03$ или 3%).

$$\sigma^2 = \frac{\sum (x_i - \bar{x})^2 \bullet n_i}{n}$$

W_0-генеральная доля

$$W_0 = \frac{M}{N}$$

W – выборочная доля

$$W = \frac{m}{n}$$

Ошибка выборочного наблюдения

Поскольку выборочная совокупность отлична от генеральной, то возникают **ошибки выборки**. При сплошном и выборочном наблюдении могут произойти ошибки двух видов: регистрации и репрезентативности.

Ошибка выборочного наблюдения

Ошибки *регистрации* могут иметь *случайный* и *систематический* характер. *Случайные* ошибки складываются из множества различных неконтролируемых причин, носят непреднамеренный характер и обычно по совокупности уравновешивают друг друга (например, изменения показателей прибора при температурных колебаниях или магнитных бурях).

Ошибка выборочного наблюдения

Систематические ошибки тенденциозны, так как нарушают правила отбора объектов в выборку (например, отклонения в измерениях при изменении настройки измерительного прибора или отбор каждой четвертой квартиры при 25% выборке в доме с четырьмя квартирами на лестничной площадке).

Теоремы закона больших чисел устанавливают связь между предельной ошибкой выборки, гарантированной с определенной вероятностью, числом (t) и средней ошибкой выборки (μ)

При оценке результатов малой выборки величина генеральной дисперсии в расчетах не используется. Для определения возможных пределов ошибки пользуются так называемым **критерием Стьюдента**, определяемым по формуле

$$t = \frac{\tilde{x} - \bar{x}}{\mu_{MB}}$$

где $\mu_{MB} = \frac{\sigma}{\sqrt{n-1}}$ - мера случайных колебаний выборочной средней в малой выборке.

Теорема Ляпунова

- А.М. Ляпунов доказал, что распределение выборочных средних(а следовательно, и их отклонений от генеральной средней) при достаточно большом числе независимых наблюдений приближенно нормально при условии, что генеральная совокупность обладает конечной средней и ограниченной дисперсией.

Теорема Ляпунова

- Математически теорему Ляпунова можно записать так:

$$P\left\{\left|\overline{x}-\tilde{x}\right|\le \Delta_x\right\}=\frac{1}{\sqrt{2\pi}}\cdot\int_{-t}^{t}e^{-\frac{t^2}{2}}dt=F(t)$$

где $\Delta_x = t\sqrt{\frac{\sigma^2}{n}}$

- π=3,14(математическая постоянная);

Δ_x - предельная ошибка выборки, которая дает возможность выяснить, в каких пределах находится величина генеральной средней.

Ошибка выборочного наблюдения

Параметры эмпирического распределения $\bar{x}$ и s^2 являются случайными величинами, следовательно, ошибки выборки также являются случайными величинами, могут принимать для разных выборок разные значения и поэтому принято вычислять *среднюю ошибку*.

Средняя ошибка выборки

$$m = \sqrt{\frac{\sigma^2}{n}}$$

Средняя ошибка выборки

выражает среднее квадратическое отклонение выборочной средней от математического ожидания. Эта величина при соблюдении принципа случайного отбора зависит прежде всего от объема выборки n и от степени колеблемости признака: чем больше n и чем меньше вариация признака (следовательно, и значение σ^2), тем меньше величина средней ошибки выборки m.

Необходимый объем выборки

ФОРМУЛЫ РАСЧЁТА НЕОБХОДИМОЙ ЧИСЛЕННОСТИ ВЫБОРКИ ДЛЯ СОБСТВЕННО-СЛУЧАЙНОГО ОТБОРА

μ	Собственно-случайный отбор	
	повторный	бесповторный
Для средней	$\frac{t^2\sigma^2}{\Delta^2}$	$\frac{t^2\sigma^2 N}{N\Delta^2 + t^2\sigma^2}$
Для доли	$\frac{t^2 w(1-w)}{\Delta^2}$	$\frac{t^2 N w(1-w)}{N\Delta^2 + t^2 w(1-w)}$

Задача

В городе 2000 семей. Предполагается провести выборочное обследование методом случайной бесповторной выборки для нахождения среднего размера семьи.

Определить необходимую численность выборки при условии, что с вероятностью 0,954 ошибка выборки не превысит 1 человека при среднем квадратическом отклонении 3 человека.

Формула

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

Решение

$$n = \frac{t^2 \cdot \sigma^2 \cdot N}{N \cdot \Delta_{\bar{x}}^2 + t^2 \cdot \sigma^2}$$

Исходные данные

данные

$$N = 2000$$

$$P = 0,954$$

$$t = 2$$

$$\sigma = 3$$

$$\Delta_{\overline{x}} = 1$$

Ответ

$$n = \frac{4 \cdot 9 \cdot 2000}{2000 \cdot 1 + 4 \cdot 9} = 36$$

Необходимо обследовать не менее 36 семей.