Выборки: генеральная и выборочная. Как правильно сделать вывод

Академия Яндекса

Различия генеральной и выборочной совокупности

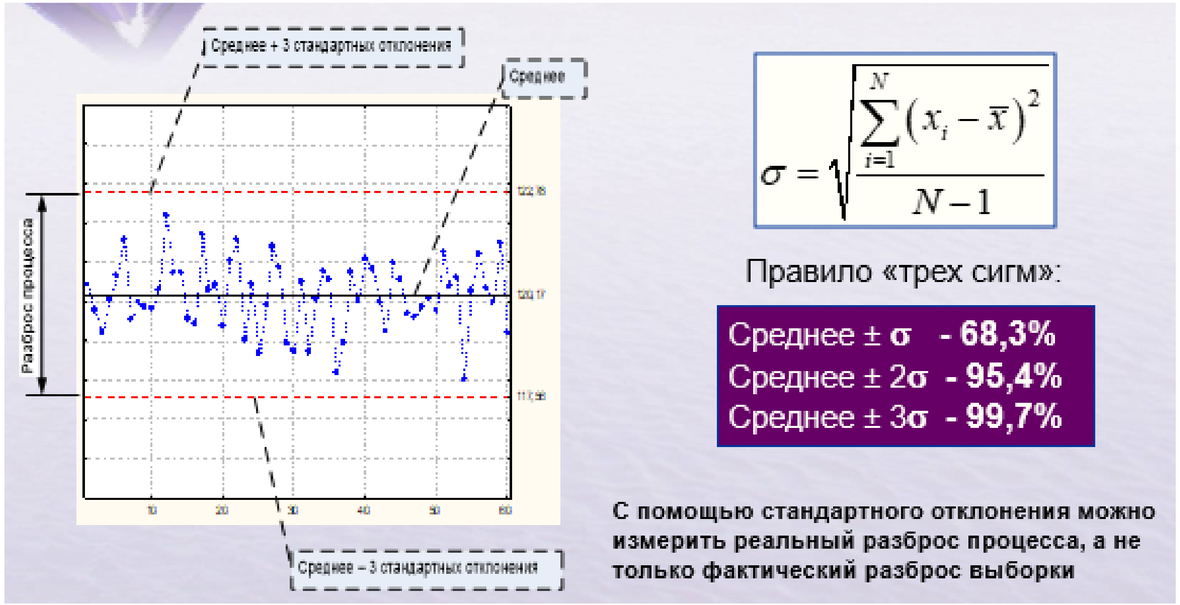
В прошлом уроке мы с вами много занимались тем, что делали выводы/суждения о том, как один фактор влияет на другой. Например, пример с ремонтом космолётов. У нас в таблице было около 60 случаев ремонта различных кораблей. И анализируя их - делали выводы.

Успешность управленческих решений напрямую зависит от того, на сколько можно доверять сделанным выводам о корреляции между факторами по данным о 60 случаев ремонта. А если взять другие 60 ремонтов - сделанные выводы сохранятся? Или будут иными?

Вот в прошлый раз мы выявляли связь между «временем_ремонта» и «налёта_от последнего_ремонта». Она равна 0.63.

На сколько можно доверять этому числу 0.63 при другой выборке данных, т.е. если будут ремонты у других космолётов?

Золотое правило 3-х сигм



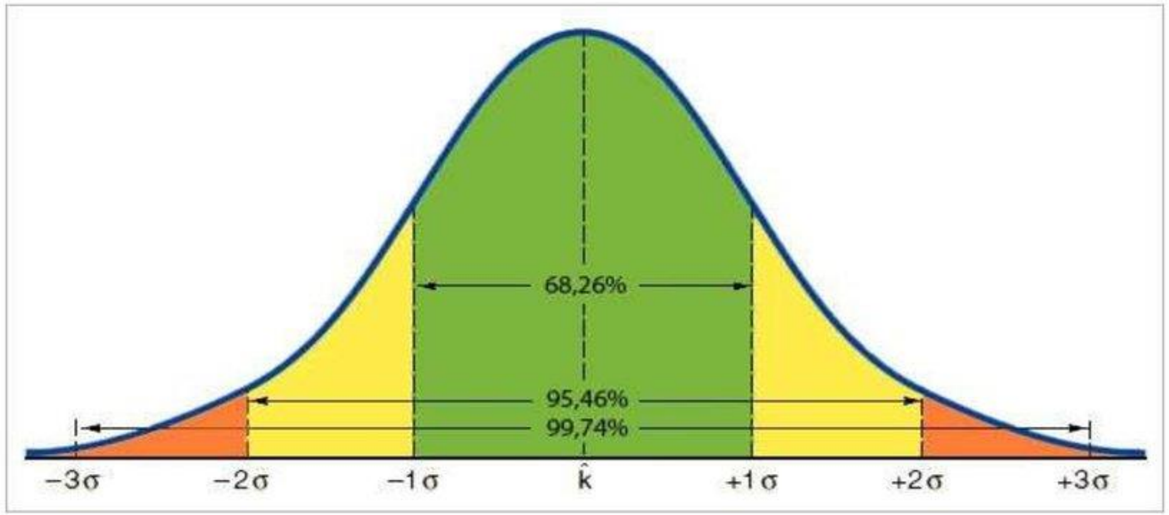
$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$$

Правило «трех сигм»:

Среднее ± σ - 68,3%
Среднее ± 2σ - 95,4%
Среднее ± 3σ - 99,7%

С помощью стандартного отклонения можно измерить реальный разброс процесса, а не только фактический разброс выборки

Правило трех сигм



Различия генеральной и выборочной совокупности

Генеральная совокупность — совокупность всех возможных наблюдений, относительно которых предполагается делать выводы при анализе данных.

Выборка — часть генеральной совокупности, которая была охвачена сбором данных.

Как мы понимаем обследовать все объекты какой-либо совокупности данных (все возможные ремонты всех существующих космолётов) вряд ли удастся. Всегда приходится довольствоваться лишь изучением части генеральной совокупности – частичной выборки данных. И вот такую выборку, отражающую свойства всей совокупности принято называть представительной.

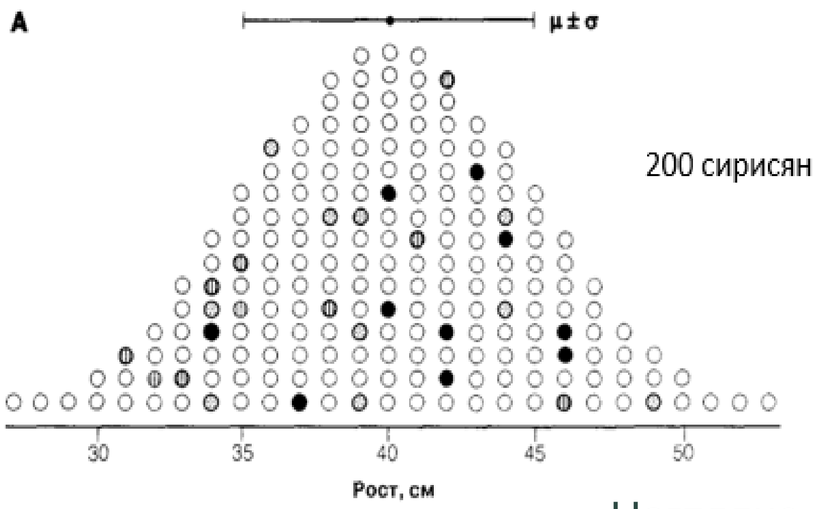
Т.е. для того, чтобы оценить любое явление, не обязательно изучать все объекты (генеральную совокупность). Для оценки здоровья человека не нужно анализировать всю кровь, достаточно небольшой пробирки.

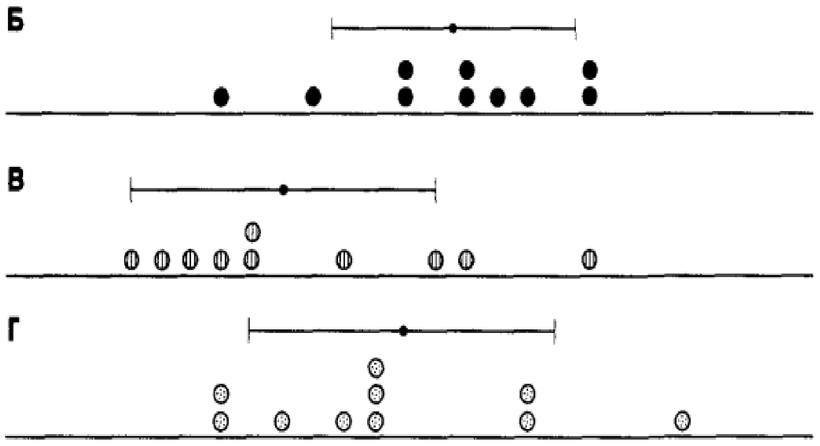
Чтобы понять суждение жителей страны можно не опрашивать всех (например в нашей стране – около 150 млн человек), а ограничиться несколькими тысячами. При этом, если всё верно сделать, то оценка не сильно потеряет в точности.

Выборки

Дальше мы вместе увидим, как оценивать степень доверия к данным и выводам на основе выборок.

Но сначала возвратимся к исследованиям «Стражей галактики». Как вы помните, они исследовали рост сирисян. Всего на планете Сирис 200 жителей. Это есть генеральная совокупность.
Нам повезло, редко можно исследовать генеральную совокупность:)

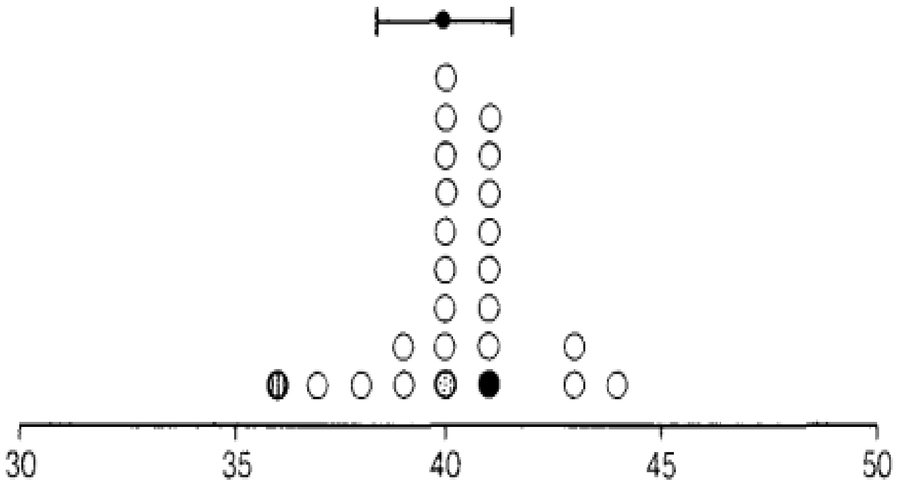




Наглядно видно, что три случайные выборки дают три разных средних и три разных стандартных отклонений. Причём видим также, что они не совпадают с тем средним, что мы с вами считали по генеральной совокупности (среднее было 40, а стандартное отклонение – 5). Получается то меньше 40, до более 40…

Выборки

А теперь предположим, что сможем сделать вот таких вот замеров по 10 жителей ещё больше. Можно посчитать, что число таких разных выборок может более, чем 10 в десятой степени. Но мы возьмём только 20-30 таких выборок. Посчитаем по ним среднее и построим распределение средних значений выборок:



Если бы была возможность, то мы бы взяли таких выборок много-много тысяч и увидели бы, что распределение будет стремиться к нормальному. Причём, чем больше выборок, тем точнее будет среднее значение средних по выборкам приближаться к истинному среднему (оно у нас 40). А разброс - снижаться.

Вот этот разброс средних значений носит название стандартной ошибкой среднего. И принято считать вот по такой формуле:

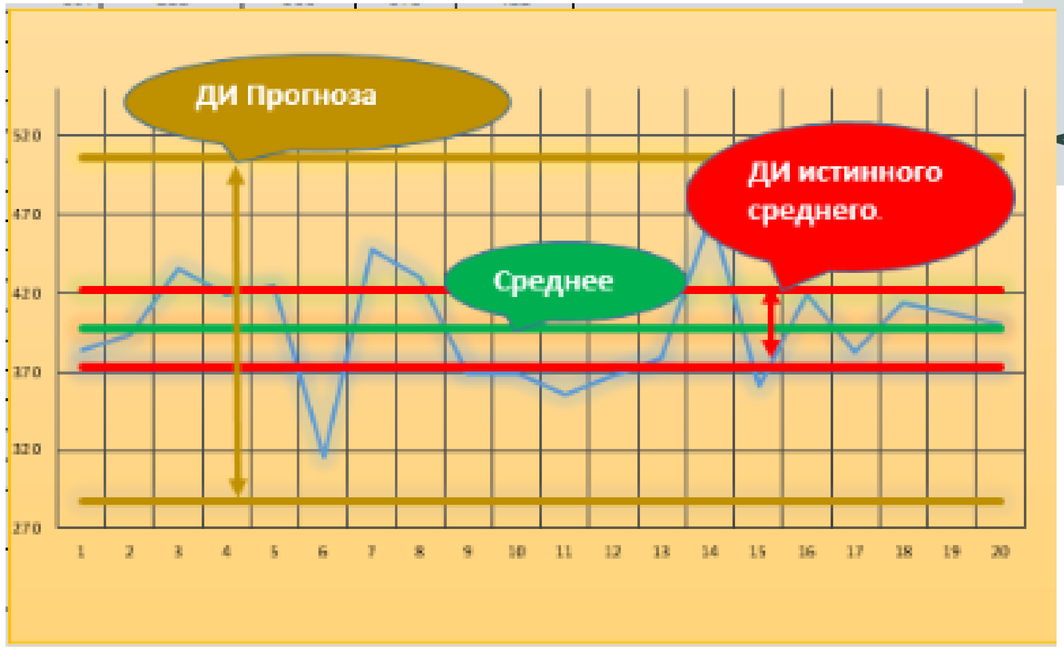
$$S_{\bar{X}} = \frac{S}{\sqrt{n}},$$

S – стандартное отклонение

Стандартная ошибка среднего – позволяет оценить степень доверия расчёта среднего по выборке. А ещё дальше – позволяет оценивать степень доверия к выводам при проверке гипотез. Например, при сравнении средних значений у двух групп. В следующем уроке это увидим.

Разница в формулах. Поправка Бесселя

Наименование характеристики	Для генеральной совокупности	Для выборки
Количество элементов	N	n
Частота	M	m
Частость (доля)	$p = \frac{M}{N}$	$w = \frac{m}{n}$
Среднее	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Дисперсия	$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Стандартное отклонение	$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}}$	$S_x = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$



Методы расчёта выборок

Если суп хорошо перемешать,
то достаточно одной ложки,
чтобы сделать вывод о вкусе всей кастрюли

На практике могут встречаться разного типа задачи.

1) На заводе изготавливаются гайки. И контролируется внутренний диаметр гайки (там где резьба). Гаек выпускается сотнями тысяч. Не контролировать же все 500 тысяч за смену. Ни у кого сил не хватит. Сколько нужно взять (т.е. какую выборку) чтобы судить о всей генеральной совокупности?

2) Или такая задача. Маркетинговая компания получила заказ на проведение социологического исследования с целью выявить долю курящих лиц в населении города. Для этого сотрудники компании будут задавать прохожим один вопрос: «Вы курите?». Возможных вариантов ответа, таким образом, только два: «Да» и «Нет». Нужно опросить всех жителей? Конечно же, только часть. Т.е. долю жителей. Какую?

Как считать размер выборок

Для оценки среднего значения

$$n = \left(\frac{z * \sigma}{H}\right)^2$$

Для оценки доли

$$n = \left(\frac{z\sqrt{p(1-p)}}{H}\right)^2$$

, где

z – доверительный уровень (степень доверия). Он показывает, насколько вы уверены, что фактическое среднее значение будет находиться в пределах выбранной вами погрешности. Большинство исследователей выбирают уровни уверенности 90%, 95% или 99%. Заданный вами уровень доверия затем соответствует Z-score, или постоянному значению, которое необходимо для уравнения размера выборки. Вот Z-коэффициенты для этих трех наиболее распространенных уровней доверия:

α (%)	60	70	80	85	90	95	97	99	99,7
z	0,84	1,03	1,29	1,44	1,65	1,96	2,18	2,58	3,0

р/σ – стандарт отклонения/ уровень дисперсии (разброса), который мы ожидаем от собранной информации.

H - предел погрешности. Определяет, сколько места для ошибки мы готовы допустить. Другими словами, какую разницу допускаем между средним значением выборки и средним значением генеральной совокупности. Т.е. для первой задач - это не что иное, как «стандартная ошибка среднего».

Как считать размер выборок

У нас процесс производства гаек уже давнейший. О нём мы можем сказать много. А значит, нам известна и дисперсия реального производственного процесса σ = 0,6. Z возьмём 1.96 (z = 1,96) – 95% уровень доверия (вспомнить золотое провило 3-х сигм). А теперь H – предел погрешности или по другому – стандартная ошибка среднего. Какую разницу между средним значением выборки и истинным средним мы допускаем. Возьмём H (стандартную ошибку) = 0.01, т.е. H = 0.01

$$n = \left(\frac{Z * \sigma}{H}\right) ** 2 = \left(\frac{1.96 * 0.6}{0.01}\right)^2 = 13830$$

Нужно взять для проверки 13830 гайки. В этом случае, с вероятностью 0.95 (выбранное Z), разница между истинным средним генеральной совокупности и средним по выборки будет не более, чем 0.01.

Почти 14000 – это большая работа.
А что если нам не надо такая сильное совпадение между выборочной и генеральной средней?

Как считать размер выборок

У нас процесс производства гаек уже давнейший. О нём мы можем сказать много. А значит, нам известна и дисперсия реального производственного процесса σ = 0,6. Z возьмём 1.96 (z = 1,96) – 95% уровень доверия (вспомнить золотое правило 3-х сигм). А теперь H – предел погрешности или по другому – стандартная ошибка среднего. Какую разницу между средним значением выборки и истинным средним мы допускаем. Возьмём H (стандартную ошибку) = 0.01, т.е. H = 0.01

$$n = \left(\frac{Z * \sigma}{H}\right) ** 2 = \left(\frac{1.96 * 0.6}{0.01}\right)^2 = 13830$$

Нужно взять для проверки 13830 гайки. В этом случае, с вероятностью 0.95 (выбранное Z), разница между истинным средним генеральной совокупности и средним по выборки будет не более, чем 0.01.

Почти 14000 – это большая работа.
А что если нам не надо такая сильное совпадение между выборочной и генеральной средней?

Стандартная ошибка среднего	0,01	0,05	0,1
	13830	553	138

Если допустим расхождение между выборочным и истинным средним на уровне 0,05, то выборка сразу уменьшается до 553.

Т.е. видим, что размер выборки существенно связан с тем уровнем доверия к данным, который мы хотим заложить в исследование.

Как считать размер выборок. Доли

Уровень доверительности такой же z = 1,96. Ещё раз об уровне доверия: можно сказать, что он выражает вероятность того, что респонденты генеральной совокупности ответят так же, как и представители анализируемой выборки.

Изучая ранее население города, считаем, что половина респондентов может ответить на вопрос о том, курят ли они — «Да». Тогда p = 0,5 (вариация p),. Следовательно находим q = 1 – p = 1 – 0,5 = 0,5. Исходя из требуемой заказчиком точности☺, допустимую ошибку принимаем за 15%, то есть H = 0,15.

(на практике: мы готовы согласиться с тем, что 15% выборки будут лишними, например. Предположим, стоимость опроса одного человека составляет 1000 рублей. Выборка 10000 человек. Тогда 15% умножить на 1000 получим 1,5 млн. – много это или мало)) Но опираться будем на заказчика.)

$$n = \left(\frac{Z * \sqrt{p * (1-p)}}{H}\right) ** 2 = \left(\frac{1.96 * 0.5}{0.15}\right)^2 = 43$$

Допустимая ошибка	0,05	0,1	0,15
	384	96	43

Следовательно, для проведения исследования с заданными параметрами (**уровень доверия, допустимая ошибка, изменчивость признака**) компании необходимо опросить **96** человек.

Академия Яндекса

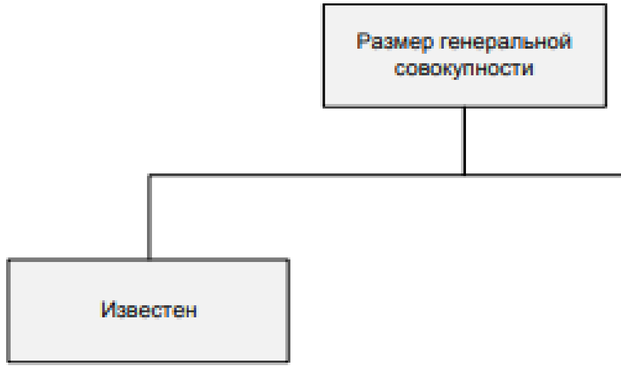
Итоги

Итого, можно сделать такой вывод: размер выборки зависит от следующих составляющих:

- **изменчивость признака** (чем больше разброс (**р/б**) имеют показания, тем больше наблюдений нужно, чтобы это уловить)
- **размер ошибки** (чем меньшие погрешности хотим (стандартная ошибка/), тем больше наблюдений необходимо)
- **уровень доверия** (уровень вероятности (z) при который мы готовы принять выводы)

Цель определения размера выборки для описательного анализа

- Для количественных шкал – оценить среднее значение признака $(\bar{x})$
- Для бинарных шкал – оценить долю признака (p)



- 7897 клиентов компании
- 220 кошек питомника
- 20 000 торговых точек, осуществлявших продажу продуктов питания в 2008 году

- Все возможные клиенты компании
- Все возможные кошки
- Все возможные торговые точки, осуществлявшие продажу продуктов питания

При определении размера выборки *n* рекомендуется скорректировать его на известный объем генеральной совокупности *N*:

$$n^* = \frac{n \cdot N}{N + n - 1}$$

Академия Яндекса

Интересные с вами выводы сделали.

Но не айс, пока ещё.

А что ещё-то не так?

Качество выборки

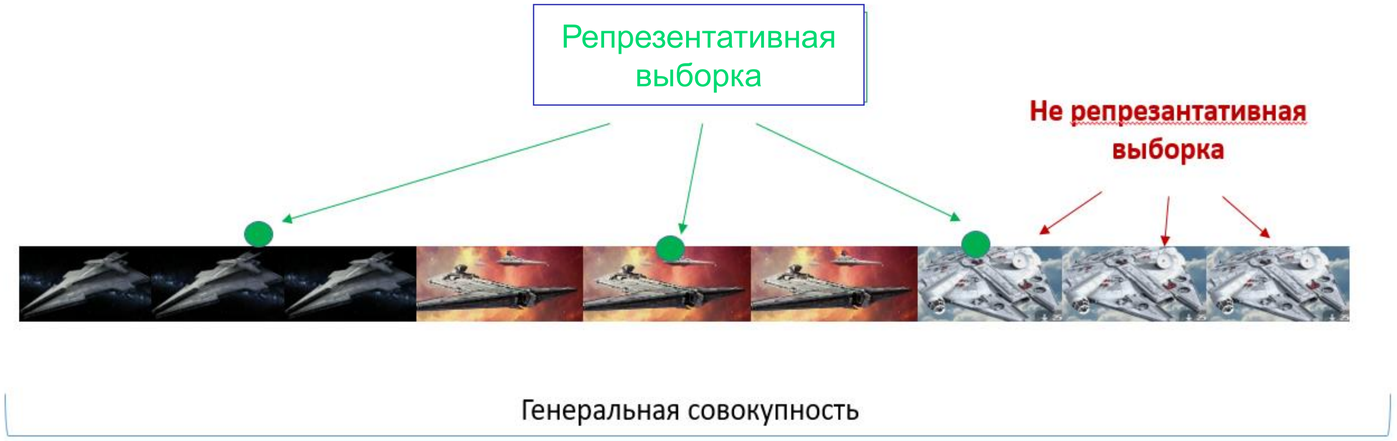
Так, как качество вывода по выборкам можно распространять (экстраполировать) на всё генеральную совокупность только в том случае, если сама выборка качественная. Или, как принято называть в сообществе – репрезентативная.

Репрезентативность — это степень соответствия характеристик выборки характеристикам генеральной совокупности.

Репрезентативная выборка

Методы создания репрезентативных выборок

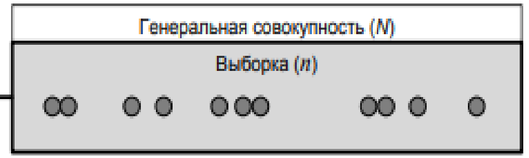
Идеальная выборка — это когда каждый космолёт имеет равную вероятность попасть в число исследуемых. Полностью случайный отбор трудно достижим (да это и очень дорого), но к нему нужно стремиться.



Нужно понимать, что сам метод сбора данных может деформировать выборку (а значит и конечные управленческие решения могут быть ложными). Например, социологический опрос может быть «однобоким»: онлайн опросы отсекают пенсионеров, опрос по стационарным телефонам исключают экономических активных людей. Или представьте, как будут различаться рейтинги, если провести опрос на в «Фэйсбуке» и в районной газете «здоровье после 60».

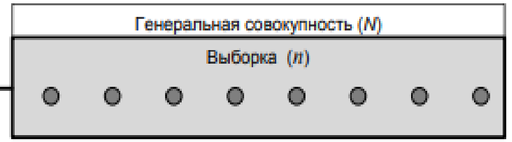
Методы создания репрезентативных выборок

Простой случайный отбор



Наблюдения отбираются случайным образом. Имея список объектов генеральной совокупности, можно, на основе генератора случайных чисел, отобрать необходимое число наблюдений.

Систематический отбор

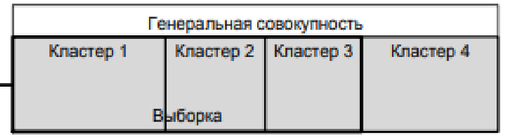


Отбирается каждое i-ое наблюдение, начиная с первого.
Шаг i определяется по формуле: $i = \frac{N}{n}$

Первое наблюдение отбираются случайным образом (от 1 до i).

Одноступенчатый кластерный отбор

Генеральная совокупность состоит из r кластеров, каждый из которых отражает свойства всей генеральной совокупности, является небольшой моделью генеральной совокупности.



В выборку входят кластеры, выбранные с помощью простого случайного отбора. Из отобранных кластеров в выборку включаются все наблюдения.

Расслоенный отбор

Генеральная совокупность состоит из k слоев (групп), средние значения изучаемого признака в которых существенно различаются.

Равномерный

Из каждого слоя случайным образом отбирается одинаковое количество наблюдений.

Пропорциональный

Выборки из каждого слоя, сформированные случайным образом, пропорциональны размерам слоев генеральной совокупности.

Оптимальный

Выборки из каждого слоя, сформированные случайным образом, пропорциональны разбросу в каждом слое генеральной совокупности (чем больше разброс, тем больше размер слоя).

Академия Яндекса

Ещё одна задачка на размер выборки:

Мы работаем на стороне розничного продавца. И нам интересно узнать, сколько наших клиентов купили товар после просмотра сайта в определенный день.

Учитывая, что посещаемость нашего сайта в среднем составляет **5 000** человек в день, надо определить размер выборки клиентов, которых стоит отслеживать с доверительной вероятностью **(z) 95 %** с погрешностью **(p) 5 %**.

Ещё одна задачка на размер выборки:

Мы работаем на стороне розничного продавца. И нам интересно узнать, сколько наших клиентов купили товар после просмотра сайта в определенный день.

Учитывая, что посещаемость нашего сайта в среднем составляет 5 000 человек в день, надо определить размер выборки клиентов, которых стоит отслеживать с доверительной вероятностью (z) 95 % с погрешностью (H) 5 %.

Итого

- количество посетителей, N = 5 000 (генеральная совокупность)
- Критическое значение при уровне достоверности 95%, Z = 1,96.
- Уровень ошибки (H) = 5% или 0,05
- А т.к. текущий коэффициент конверсии неизвестен, примем p = 0,5 (он даёт максимальное значение в произведении p*q).

Тогда:

$$n = \left(\frac{1.96 * 0.5}{0.05}\right)^2 = 384$$

Ещё одна задачка на размер выборки:

Мы работаем на стороне розничного продавца. И нам интересно узнать, сколько наших клиентов купили товар после просмотра сайта в определенный день.

Учитывая, что посещаемость нашего сайта в среднем составляет 5 000 человек в день и коэффициент конверсии на исторических данных составляет 3%, надо определить размер выборки клиентов, которых стоит отслеживать с доверительной вероятностью (z) 95 % с погрешностью (p) 5 %.

Итого

- количество посетителей, N = 5 000 (генеральная совокупность)
- Критическое значение при уровне достоверности 95%, Z = 1,96.
- Уровень ошибки = 5% или 0,05
- А т.к. текущий коэффициент конверсии неизвестен, примем p = 0,5 (он даёт максимальное значение в произведении p*q).

Тогда:

$$n = \left(\frac{1.96 * 0.5}{0.05}\right)^2 = 384$$

$$n' = \frac{n * N}{N + n - 1} = 357$$

Академия Яндекса

Мы молодцы!

Совсем итог уже

Вид отбора	Оценка разброса (дисперсии)		Расчет размера выборки
	для среднего	для доли	
Простой случайный **Систематический**	$\sigma^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}$	$\sigma^2 = p \cdot (1-p)$	общий размер выборки: $n = \frac{\sigma^2}{SE^2}$, $n' = \frac{n \cdot N}{N+n-1}$
Расслоенный	разброс каждого слоя: $\sigma_i^2 = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x_i}\right)^2}{n_i - 1}$, $i = 1,\ldots,k$ средний разброс по всем слоям: $\overline{\sigma^2} = \frac{\sum_{i=1}^{k}\left(\sigma_i^2 \cdot (n_i - 1)\right)}{n-k}$	разброс каждого слоя: $\sigma_i^2 = p_i \cdot (1-p_i)$, $i = 1,\ldots,k$ средний разброс по всем слоям: $\overline{\sigma^2} = \frac{\sum_{i=1}^{k}\left(\sigma_i^2 \cdot (n_i - 1)\right)}{n-k}$	общий размер выборки: $n = \frac{\overline{\sigma^2}}{SE^2}$, $n' = \frac{n \cdot N}{N+n-1}$
Равномерный			размер выборки из каждого слоя одинаковый: $n_i = \frac{n}{k}$
Пропорциональный			размер выборки из каждого слоя пропорционален размеру слоя в генеральной совокупности: $n_i = n\frac{N_i}{N}$
Оптимальный			размер выборки из каждого слоя пропорционален разбросу каждого слоя: $n_i = n\frac{N_i \cdot \sigma_i}{\sum_{i=1}^{k}\left(N_i \sigma_i\right)}$
Кластерный	$\sigma^2 = \frac{\sum_{i=1}^{r}\left(\overline{x}_i - \overline{x}\right)^2 n_i}{\sum_{i=1}^{r} n_i}$	$\sigma^2 = \frac{\sum_{i=1}^{r}\left(p_i - p\right)^2 n_i}{\sum_{i=1}^{r} n_i}$	общий размер выборки (число кластеров): $r = \frac{\sigma^2}{SE^2}$, $r' = \frac{r \cdot R}{N+R-1}$

Обозначения:

n – размер выборки, N – размер генеральной совокупности,

$\sigma\left(\sigma^2\right)$ – стандартное отклонение (дисперсия) выборки,

k – количество слоев в генеральной совокупности (выборке),

n_i – количество значений в i-ом слое выборки или в i-ом кластере, N_i – размер слоя в генеральной совокупности,

$\sigma_i\left(\sigma_i^2\right)$ – стандартное отклонение (дисперсия) i-го слоя выборки,

r – число отобранных из генеральной совокупности кластеров, R - число кластеров в генеральной совокупности,

$\overline{x_i}$ – среднее значение признака в i-ом слое или i-ом кластере, p_i – доля признака в i-ом слое или i-ом кластере.

Что нужно будет для задачек

Для оценки среднего значения

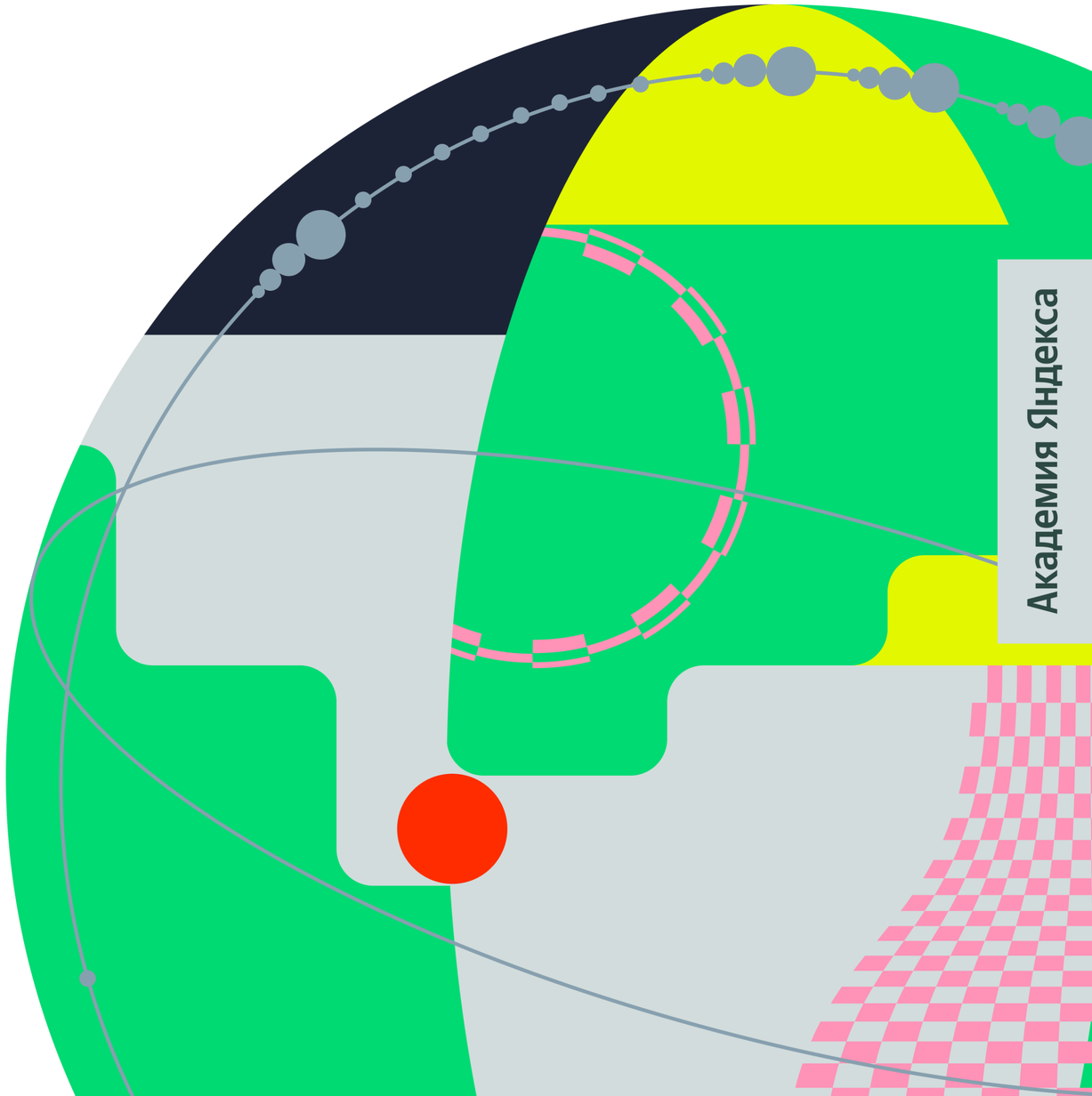
$$n = \left(\frac{z * \sigma}{H}\right)^2$$

Для оценки доли

$$n = \left(\frac{z\sqrt{p(1-p)}}{H}\right)^2$$

$$n' = \frac{n * N}{N + n - 1}$$

Спасибо за внимание



Академия Яндекса позволяет школьникам и студентам освоить востребованные IT-профессии по программам, разработанным экспертами компании