Доверительные интервалы для доли и дисперсии

План

- Доверительный интервал для доли
- Доверительный интервал для дисперсии
- Статистический бутстреппинг

Доверительный интервал для доли

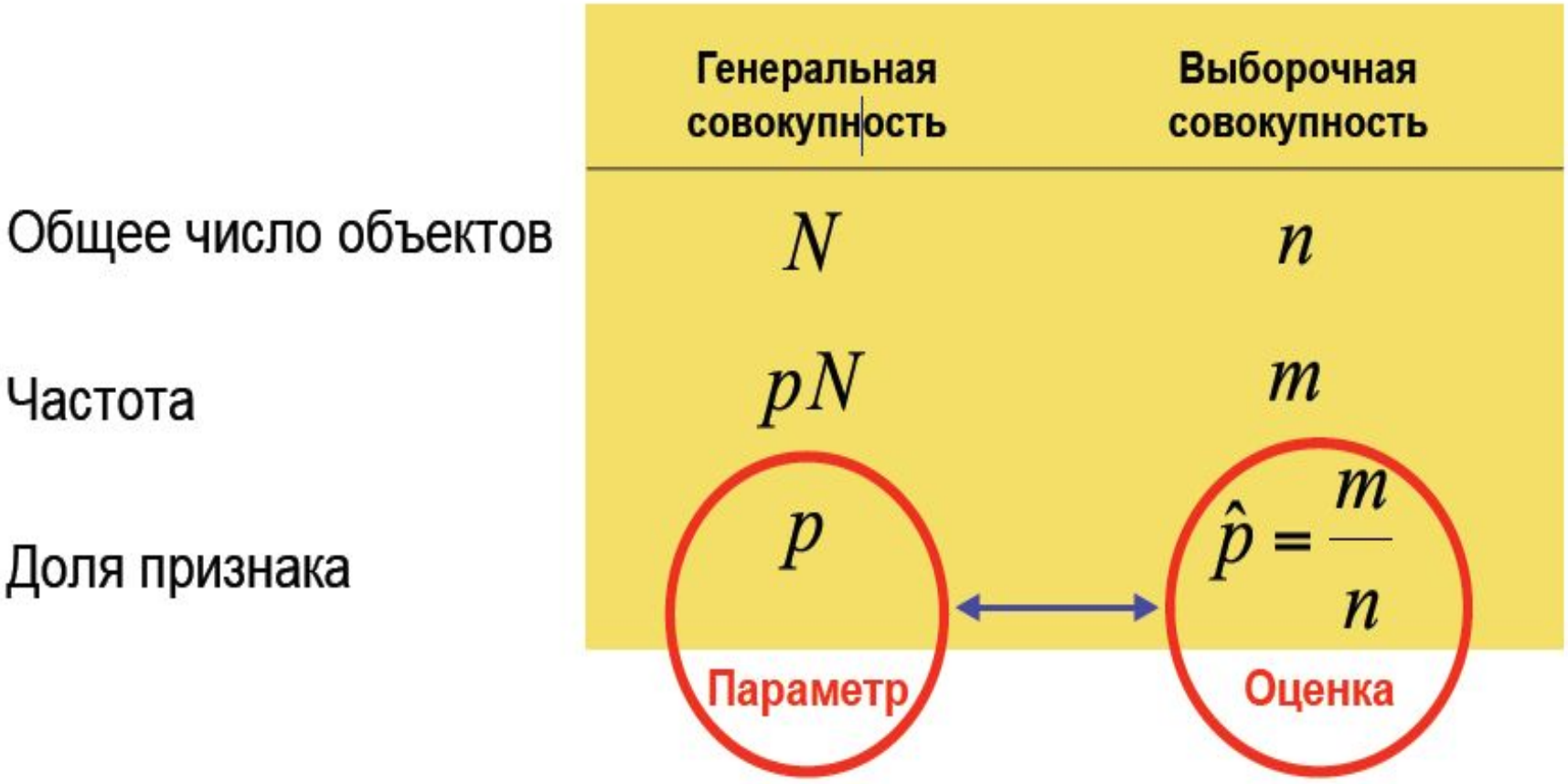
- Описание проблемы
- ДИ
- Алгоритм
- Пример

Оценка доли признака

Задача состоит в построении доверительной оценки для
генеральной доли, если известно значение выборочной доли.

Пример. Среди 500 резюме кандидатов на работу няни
оказалось 60 принадлежащих мужчинам. Если считать, что
выборка репрезентативна, то требуется построить 90%-ый
доверительный интервал для фактической доли мужчин,
устраивающихся на работу нянями.

Оценка доли признака



Формальное описание проблемы

Цель. Оценить долю признака в генеральной совокупности.

Что мы имеем. Имеем случайную выборку объема n из генеральной совокупности. По выборке вычислена доля признака. Выполнены условия $np \geq 5$ и $n(1 - p) \geq 5$.

Требуется. Построить доверительный интервал для доли:

$$p - E < p < p + E$$

Доверительный интервал для доли

Доля значений признака в генеральной совокупности с надежностью $1 - \alpha/2$ находится в доверительном интервале:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Обозначение: $\hat{q} = 1 - \hat{p}$

Последовательность действий

Шаг 1. По выборке вычислить долю признака.

Шаг 2. По таблице нормального распределения найти z-значение для доверительной вероятности $1 - \alpha$.

Шаг 3. Вычислить точность интервальной оценки по формуле:

$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$$

Последовательность действий

Шаг 4. Подставить полученные значения в формулу для доверительного интервала:

$$p - E < p < p + E$$

Шаг 5. Написать ответ.

Пример. Выборы мэра

В ходе проведенного опроса 829 жителей города выяснилось, что 417 опрошенных (51,5%) предполагают поддержать на предстоящих выборах кандидатуру действующего мера.

Местная многотиражка поспешила заявить, что более половины жителей города поддерживают перевыборы действующего мера на следующий срок.

Построить доверительный интервал для доли генеральной совокупности и проверить утверждение корреспондента.

Решение

Шаг 1. По условию, доля признака в выборке составила:

$$\hat{p} = 0{,}515$$

$$\hat{q} = 0{,}485$$

$$n = 829$$

Шаг 2. Для доверительной вероятности $1 - \alpha = 0{,}95$ по таблице нормального закона находим z-значение:

$$z_{\alpha/2} = 1{,}96$$

Решение

Шаг 3. Вычисляем точность интервальной оценки:

$$E = 1{,}96 \cdot \sqrt{\frac{0{,}515 \cdot 0{,}485}{829}} = 0{,}034$$

Шаг 4. Подставляем полученные значения в формулу для доверительного интервала:

$$0{,}515 - 0{,}034 < p < 0{,}515 + 0{,}034$$

Шаг 5. Ответ:

$$48{,}1\% < p < 54{,}9\%$$

Пример. Мужчины-няни

Среди 500 резюме кандидатов на работу няни оказалось 60 принадлежащих мужчинам.

Найти 90%-ый доверительный интервал для фактической доли мужчин, устраивающихся работать нянями.

Решение

Шаг 1. По условию, доля признака в выборке составила:

$$\hat{p} = 0{,}12$$

$$\hat{q} = 0{,}88$$

$$n = 500$$

Шаг 2. Для доверительной вероятности $1 - \alpha = 0{,}90$ по таблице нормального закона находим z-значение:

$$z_{\alpha/2} = 1{,}65$$

Решение

Шаг 3. Вычисляем точность интервальной оценки:

$$E = 1{,}65 \cdot \sqrt{\frac{0{,}12 \cdot 0{,}88}{500}} = 0{,}024$$

Шаг 4. Подставляем полученные значения в формулу для доверительного интервала:

$$0{,}12 - 0{,}024 < p < 0{,}12 + 0{,}024$$

Шаг 5. Ответ:

$$9{,}6\% < p < 14{,}4\%$$

Объем выборки для оценки доли

Минимальный объем выборки, требуемый для интервального оценивания генеральной доли, находится по формуле:

$$n = \hat{p}\hat{q} \cdot \left(\frac{z_{\alpha/2}}{E} \right)^2$$

При необходимости следует округлить n, чтобы получить целое число.

Важное замечание

Если оценка для доли неизвестна, минимальный объем находят по формуле:

$$n = 0{,}25 \cdot \left(\frac{z_{\alpha/2}}{E} \right)^2$$

Пример. У кого есть дома компьютер?

Исследователь хочет с 95%-ой вероятностью оценить количество людей, у которых дома имеется персональный компьютер.

По данным предыдущего исследования у 40% опрошенных есть дома компьютер.

Исследователь не хочет ошибиться больше,

чем на 2% по сравнению с генеральной долей.

Найти минимальный размер выборки.

Решение

Поскольку 1 - α = 0,95, то z-значение равно 1,96. E = 0,02.

Подставляем в формулу и вычисляем:

$$n = \hat{p}\hat{q} \cdot \left(\frac{z_{\alpha/2}}{E} \right)^2 =$$

$$= 0{,}40 \cdot 0{,}60 \cdot \left(\frac{1{,}96}{0{,}02} \right)^2 = 2304{,}96$$

Ответ. Нужно опросить 2305 людей.

Доверительный интервал для дисперсии

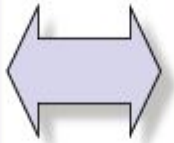
- Описание проблемы
- Доверительный интервал
- Алгоритм
- Пример

Оценка для генеральной дисперсии

Задача состоит в построении интервальной оценки генеральной дисперсии на основе выборочной дисперсии.

Генеральная дисперсия

$$\sigma^2$$



Выборочная дисперсия

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

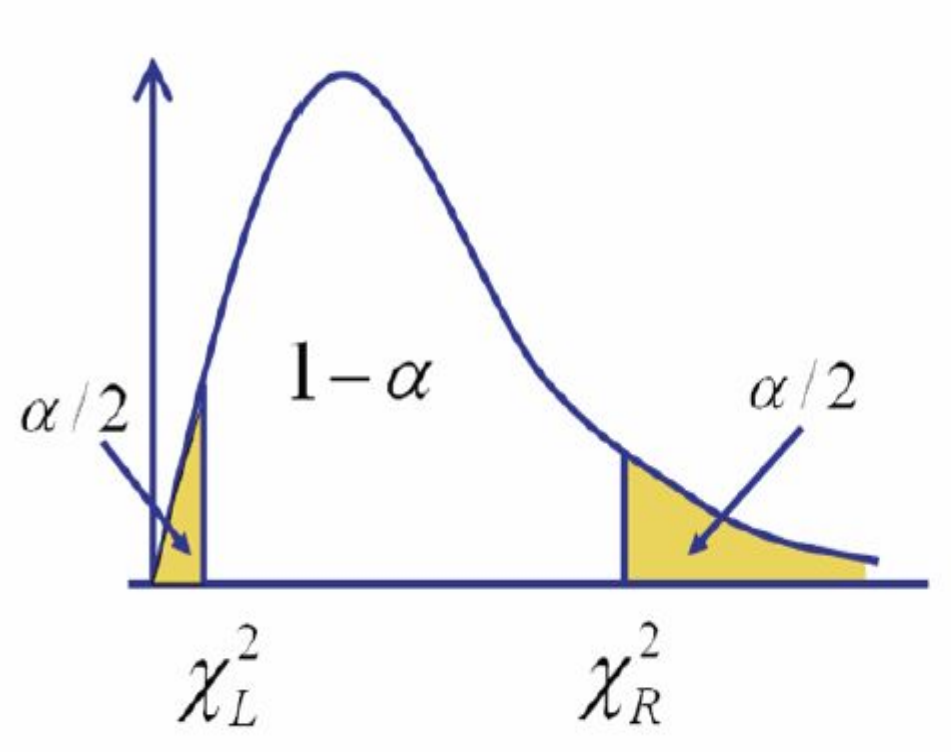
Доверительный интервал для дисперсии

Доверительный интервал для дисперсии находится по формуле:

$$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$$

Значения хи-квадрат

Значения χ^2_L и χ^2_R находятся по таблицам хи-квадрат распределения, исходя из следующих условий:



$$P(\chi^2 > \chi_L) = 1 - \frac{\alpha}{2}$$

$$P(\chi^2 > \chi_R) = \frac{\alpha}{2}$$

Оценка стандартного отклонения

Доверительный интервал для стандартного отклонения находится по следующей формуле:

$$\frac{\sqrt{n-1}\cdot s}{\chi_R} < \sigma < \frac{\sqrt{n-1}\cdot s}{\chi_L}$$

Последовательность действий

Шаг 1. По выборке вычислить дисперсию.

Шаг 2. По таблице найти два хи-квадрат значения χ^2_L и χ^2_R для доверительной вероятности $1 - \alpha$ и числа степеней свободы $df = n - 1$.

Шаг 3. Подставить полученные значения в формулу:

$$\frac{(n-1)s^2}{\chi^2_R} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_L}$$

Шаг 4. Написать ответ.

Пример. Оценка для дисперсии

Из нормально распределенной генеральной совокупности сделана выборка из 10 элементов. Выборочная дисперсия оказалась равна 28,2.

Требуется оценить дисперсию генеральной совокупности (построить доверительный интервал).

Доверительную вероятность выберем на уровне 90%.

Последовательность действий

Шаг 1. По выборке объема 10 вычислена дисперсия 28,2.

Шаг 2. Число степеней свободы df = n – 1 = 9. Поскольку доверительная вероятность равна 90%, по таблице хи-квадрат распределения находим значения $\chi^2_L = 3{,}325$ и $\chi^2_R = 16{,}919$.

Шаг 3. Подставим полученные значения в формулу:

$$\frac{(10-1)\cdot 28{,}2}{16{,}919} < \sigma^2 < \frac{(10-1)\cdot 28{,}2}{3{,}325}$$

Шаг 4. Ответ:

$$15{,}0 < \sigma^2 < 76{,}3$$

Оценка для стандартного отклонения

Если дисперсия находится в доверительном интервале:

$$15{,}0 < \sigma^2 < 76{,}3$$

То стандартное отклонение можно оценить следующим образом:

$$\sqrt{15{,}0} < \sigma < \sqrt{76{,}3}$$

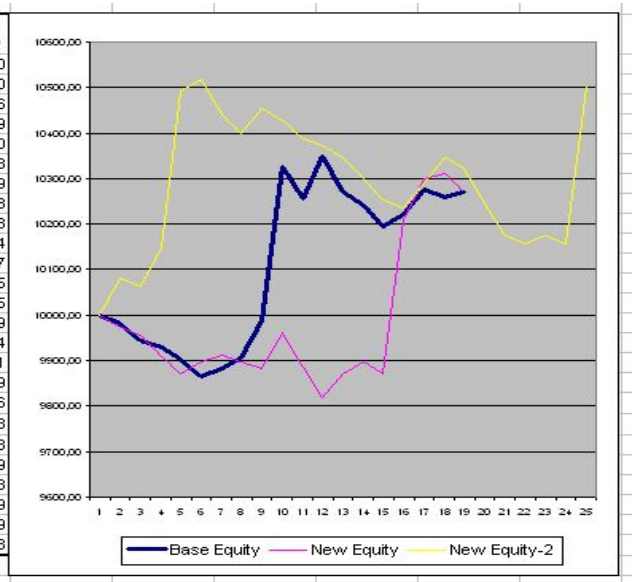
$$3{,}87 < \sigma < 8{,}73$$

Статистический бутстреппинг

- компьютерный метод определения статистик вероятностных распределений. Основан на многократном генерировании выборок методом Монте-Карло на базе данных обучающей выборки

Позволяет просто и быстро оценивать самые разные статистики (доверительные интервалы, дисперсию, корреляцию и так

№ Сделки	Результаты сделок в %%	Base Equity	Случайные числа	Выборка "без возвращения"	New Equity	Случайные числа	Выборка "с возвращением"	New Equity-2
		10000,00			10000,00			10000,00
1	-0,17	9983,00	0,411261215	-0,28	9972,00	0,428420518	0,82	10082,00
2	-0,39	9944,07	0,946500676	-0,17	9955,05	0,054145221	-0,17	10064,86
3	-0,16	9928,16	0,389427422	-0,45	9910,25	0,430213317	0,82	10147,39
4	-0,24	9904,33	0,904438491	-0,39	9871,60	0,50220282	3,4	10492,40
5	-0,39	9865,70	0,293166656	0,25	9896,28	0,833952213	0,25	10518,63
6	0,16	9881,49	0,815372834	0,16	9912,11	0,653749519	-0,76	10438,69
7	0,25	9906,19	0,045847387	-0,15	9897,24	0,107047182	-0,39	10397,98
8	0,82	9987,42	0,904222071	-0,16	9881,41	0,903284013	0,54	10454,13
9	3,4	10326,99	0,717001424	0,82	9962,44	0,215237584	-0,24	10429,04
10	-0,69	10255,74	0,502512969	-0,76	9886,72	0,088959513	-0,39	10388,37
11	0,9	10348,04	0,603913429	-0,69	9818,50	0,121159106	-0,16	10371,75
12	-0,76	10269,39	0,120196198	0,54	9871,52	0,221160886	-0,24	10346,85
13	-0,28	10240,64	0,742579112	0,25	9896,20	0,812957347	-0,45	10300,29
14	-0,45	10194,56	0,901660356	-0,24	9872,45	0,787930051	-0,45	10253,94
15	0,25	10220,04	0,609093728	3,4	10208,12	0,043479139	-0,17	10236,51
16	0,54	10275,23	0,519431979	0,9	10299,99	0,88359346	0,54	10291,79
17	-0,15	10259,82	0,008521943	0,1	10310,29	0,918644572	0,54	10347,36
18	0,1	**10270,08**	0,828671963	-0,39	**10270,08**	0,226100933	-0,24	10322,53
			ИНДЕКС($J$19:$J$36;РАНГ(M36;$M$19:$M$36))			0,704197638	-0,76	10244,08
						0,582988344	-0,69	10173,39
						0,997303657	-0,15	10158,13
						0,330318518	0,16	10174,39
						0,036849849	-0,17	10157,09
						0,518515738	3,4	10502,43
						ИНДЕКС($J$19:$J$36;ЦЕЛОЕ(1+Q42*($I$36-1)))		



Методы размножения выборок (бутстреп-методы)

- предложен в 1977 г. Б.Эфроном из Станфордского университета (США)

- "bootstrap” - кожаные петельки на задниках ботинок

- “lift himself by his bootstraps” - “вытащить себя из болота за ушки на задниках ботинок” , “выбиться в люди благодаря собственным усилиям”

Принцип статистического бутстрэппинга

- имитировать многократное получение выборки из генеральной совокупности, используя данные из имеющейся у нас выборки.

Предположим, что мы исследуем высоту людей во всем мире. Мы не можем измерить всех людей, а вместо этого выбираем лишь малую часть. Пусть в нашей выборке N людей. Мы можем посчитать среднее значение. Но для того, чтобы рассуждать о доверительном интервале роста населения, нам нужно некоторое представление о вариабельности среднего.

Используя наши исходные данные о росте N различных людей, составляем новую выборку, также размера N. Это новая выборка взята из исходной случайным образом так, что мы каждый раз случайным образом выбираем из N имеющихся значений). У такой выборки будет другое среднее.

Сделав такую выборку много раз (возможно, 1000 или 10000 раз), каждый раз вычисляя среднее, мы получаем гистограмму распределения, которая может ответить на вопросы о доверительном интервале.