2.8 Статистический смысл выборочных показателей

Если произвести большое число выборок равного объема из генеральной совокупности, то для каждой выборки мы получим свои значения показателей (средних значений, дисперсий и т. д.), которые, например, для среднего значения признака X образуют ряд:

$$\bar{x}_1, \bar{x}_2, \bar{x}_3 \ldots .$$

Теперь, если число выборок устремить к бесконечности, то получится кривая частот, которая представляет собой кривую выборочного распределения.

Таким образом выборочные показатели являются случайными величинами.

При некоторых достаточно общих предположениях о распределении в генеральной совокупности (конечность средних и ограниченность дисперсии), выборочное распределение является нормальным, а его параметры совпадают с параметрами распределения изучаемого вариационного признака в генеральной совокупности.

Сделанные выше утверждения являются основой применения выборочного метода для изучения социально-экономических явлений.

Это замечание важно потому, что эконометрист всегда имеет дело с выборочной совокупностью.

Пусть из генеральной совокупности отобрана случайная выборка $X_1, X_2, X_3 \ldots X_n$.

Следует найти наилучшую оценку для генеральной средней.

Оценкой случайной величины X называется некоторая функция

$$\tilde{x}_n = \tilde{x}(x_1, x_2, \ldots x_n).$$

В частности, если речь идет о среднем значении, то в качестве оценки можно выбрать выражение

$$\tilde{x}_n = \bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

В качестве оценки среднего значения можно взять и полусумму максимального и минимального значений. Какая оценка является наилучшей?

Назвать наилучшей ту оценку, которая наиболее близка к истинному значению параметра невозможно, так как оценка является случайной величиной.

О качестве оценки следует судить не по ее индивидуальному значению, а по распределению ее значений в большом числе испытаний. Чем меньше рассеяние случайной величины относительно истинного значения, тем лучше оценка.

Оценка параметра X называется несмещенной, если математическое ожидание оценки равно ее истинному значению при любом объеме выборки $M(\tilde{x}_n) = \bar{x}_0$.

В противном случае оценка называется смещенной.

Оценка параметра X называется состоятельной, если она удовлетворяет закону больших чисел

$$\lim_{n\to\infty} P\left(|\tilde{x}_n - \bar{x}_0| < \varepsilon\right) = 1$$

и при увеличении объема выборки оценка приближается к истинному значению (в качестве случайной величины здесь взято среднее значение).

Несмещенная оценка называется эффективной, если она обладает наименьшей дисперсией.

Используемые оценки не всегда являются эффективными, поскольку для эффективной оценки формулы могут оказаться слишком сложными.

2.9. Свойства выборочной средней и дисперсии

Выборочная средняя является несмещенной оценкой генеральной средней.

Доказательство . Пусть выборочная средняя определяется формулой

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

Будем рассматривать $x_1, x_2, \ldots, x_n$

как случайные величины. Эти случайные величины имеют одинаковые параметры распределения (дисперсию и среднее значение).

Докажем, что математическое ожидание выборочной средней равно генеральной средней.

Действительно, из определения математического ожидания

$$M(\bar{x}_в) = M\left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right) = \bar{x}_0.$$

Поскольку каждая из величин $x_1, x_2, \ldots, x_n$

имеет то же распределения, что и случайная величина X в генеральной совокупности, то математическое ожидание

$$M(x_1) = M(x_2) = \ldots = M(x_n) = \bar{x}_0.$$

Отсюда сразу получаем

$$M(\bar{x}_в) = M(x) = \bar{x}_0.$$

Найдем дисперсию выборочной средней.

Будем рассматривать выборочные средние как случайные величины. Найдем дисперсию среднего арифметического одинаково распределенных случайных величин x_i

$$s_{\bar{x}}^2 = D\left(\frac{x_1 + x_2 + \ldots + x_n}{n}\right) =$$

$$= \frac{D(x_1) + D(x_2) + \ldots + D(x_n)}{n^2} = \frac{\sigma_0^2}{n}.$$

В этой формуле буквой D обозначена дисперсия аргумента, σ_0^2 дисперсия в генеральной совокупности.

Среднее квадратическое отклонение выборочных средних, которое обозначено буквой $s_{\bar{x}}$,

можно использовать для оценки по порядку величины отклонение выборочной средней от генеральной средней.

$$\bar{x} - \bar{x}_0 \approx \pm s_{\bar{x}} = \pm \frac{\sigma_0}{\sqrt{n}}.$$

При этом ошибка конкретной выборки может принимать различные значения, и она зависит от объема выборки и среднего квадратического отклонения в генеральной совокупности.

2.10. Оценка генеральной дисперсии по выборочной

Очень часто дисперсия в генеральной совокупности является неизвестной величиной и ее нужно оценить по выборочной дисперсии.

Если в качестве оценки генеральной дисперсии взять значение выборочной дисперсии, то такая оценка получается смещенной и дает заниженное значение генеральной дисперсии, приводя к систематической ошибке.

Поэтому на практике в качестве оценки генеральной дисперсии используют исправленную выборочную дисперсию σ^2, математическое ожидание которой равно генеральной дисперсии:

$$\sigma_0^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sigma^2 n}{n-1}.$$

При больших объемах выборки исправленная дисперсия несущественно отличается от выборочной. Доказательство этой формулы можно найти в учебниках по мат. статистике.

2.11. Доверительный интервал и доверительная вероятность

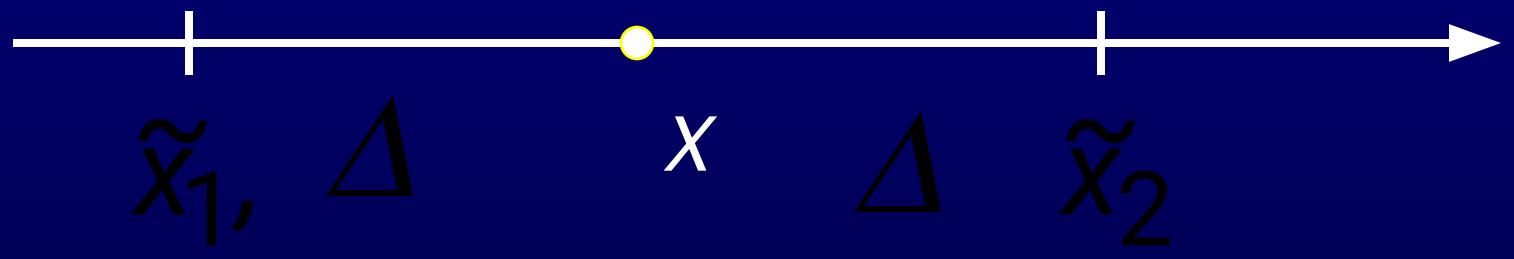
До сих пор оценку параметров генеральной совокупности мы производили одним числом. Такая оценка называется точечной.

В ряде задач нужно не только найти для параметра подходящую численную оценку, но и указать интервал значений параметра, который с заданной вероятностью «накроет» неизвестное значение параметра в генеральной совокупности.

Такая оценка параметра называется интервальной.

Определение

Интервальной оценкой $\tilde{x}$ параметра X называется числовой интервал $(\tilde{x}_1, \tilde{x}_2)$, который с заданной вероятностью γ накрывает неизвестное значение параметра X. Важно отметить, что $\tilde{x}_1$, и $\tilde{x}_2$ определяются по выборочному наблюдению.



Построим доверительный интервал для генеральной средней в случае большой повторной выборки (n велико).

Нас интересует ошибка конкретной выборки. Поэтому введем понятие нормированного отклонения, обозначив его буквой *t:*

$$t = \frac{\bar{x} - \bar{x}_0}{s_{\bar{x}}}.$$

Эта величина подчиняется распределению Стьюдента с числом степеней свободы k=n-1, где n - объем выборки.

Ошибки репрезентативности выборочного обследования избежать нельзя, но можно потребовать, чтобы вероятность отклонения выборочной средней от генеральной средней :

$$\Delta = \bar{x} - \bar{x}_0 = t s_{\bar{x}}$$

была допустимой для данного исследования.

Вероятность, которая принимается при расчете выборочной характеристики, называется доверительной вероятностью.

Для определения величины интервала, который с заданной с заданной доверительной γ вероятностью накроет среднее значение $\bar{x}_0$ мы должны потребовать выполнение равенства

$$P(|\bar{x}-\bar{x}_0|<\Delta)=\gamma,$$ где $P(|\bar{x}-\bar{x}_0|<\Delta)-$ вероятность того, что модуль отклонения

$$|\bar{x}-\bar{x}_0|<\Delta.$$

Или иначе $$P(|t|<\Delta/s_{\bar{x}})=\gamma,$$

Зная величину γ по таблице распределения Стьюдента или с помощью функции Excel СТЬЮДРАСПОБР(q;k), q =(1-P); где P - доверительная вероятность, находим критическое значение величины t.

Сказанное выше легко может быть проиллюстрировано на графике (см. след. слайд).

К определению критического значения статистики Стьюдента

Задача

При обследовании выработки 1000 рабочих цеха в отчетном году по сравнению с предыдущим по схеме собственно - случайной выборки было отобрано 100 рабочих (полученные данные изображены на след. слайде).

Определить:

а) вероятность того, что средняя выработка рабочих цеха отличается от средней выборочной не более чем на 1%;

б) границы в которых с вероятностью 0,95 заключена средняя выработка рабочих цеха.

Данные о выработке рабочих в отчетном году. 2

Выработка в отчетном году в % к предыдущему	Число рабочих
94,0 - 100,0	3
100,0 - 106,0	7
106,0 - 112,0	11
112,0 - 118,0	20
118,0 - 124,0	28
124,0 - 130,0	19
130,0 - 136,0	10
136,0 - 142,0	2
Всего	100

3 Решение

Найдем вначале среднее и дисперсию используя электронные таблицы.

Интервалы	Середина интервалов X	Часто ты f	X*f	(X-Xср)^2
94,0 - 100,0	97	3	291	1478,5
100,0 - 106,0	103	7	721	1837,1
106,0 - 112,0	109	11	1199	1144,4
112,0 - 118,0	115	20	2300	352,8
118,0 - 124,0	121	28	3388	90,72
124,0 - 130,0	127	19	2413	1156
130,0 - 136,0	133	10	1330	1904,4
136,0 - 142,0	139	2	278	784,08
			11920	8748
Срзнач=119,2%	Дисперсия =87,48		119,2	87,48

Найдем среднеквадратическую ошибку выборки для средней:

$$s_{\bar{x}} = \sqrt{\frac{\sigma_0^2}{n}} = \sqrt{\frac{n \cdot \sigma^2}{(n-1) \cdot n}} =$$

$$= \sqrt{\frac{87{,}48}{99}} = 0{,}94\%.$$

Искомую доверительную вероятность найдем из условия (Δ = 1 %), k=7

$$P(|\bar{x} - \bar{x}_0| < \Delta) = P(|t| < \Delta / s_{\bar{x}})$$

$$= 0{,}7.$$

Таким образом, вероятность того, что выборочная средняя отличается от генеральной не более чем на 1% равна 0, 7. Можно сказать, что в 70 случаях из 100 произведенное выборочное исследование даст ошибку определения средней производительности труда для всего цеха не более чем 1%.

Найдем границы в которых с вероятностью 0,95 будет находиться средняя выработка рабочих цеха. Опять используем условие

$$P(|t| < \Delta / s_{\bar{x}}) = 0{,}95.$$

Из таблиц для распределения Стьюдента, находим значение аргумента t. Это значение равно 2,3. Поэтому

$$\Delta / s_{\bar{x}} = 2{,}3; \quad \Delta = 2{,}3 * \sqrt{87{,}48/99} = 2{,}16\%.$$

Таким образом, генеральная средняя будет с вероятностью 0,95 находиться в интервале

$$119{,}2\% - 2{,}16\% < \bar{x}_0 < 119{,}2\% + 2{,}16\%.$$