Лекция 8

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ОСНОВЫ СТАТИСТИКИ

Критерии выбора оценок. Построение доверительного интервала

Здравствуйте!
В ходе изучения материалы вы будете ознакомлены со следующими понятиями:

- смещенная и несмещенная оценка
- доверительные интервал
- научитесь строить доверительные интервалы с определенным уровнем статистической значимости.

Целью этого направления статистики является оценка значений или параметров совокупности из статистики, вычисленной из выборок. Вы уже знакомы с опросами общественного мнения и предвари тельными прогнозами, наиболее распространенными из этих методов. Стандартная процедура оценки значений совокупности заключается в построении доверительного интервала, математического утверждения, которое гласит, что параметр находится в пределах определенного интервала или диапазона значений. Например, доверительный интервал может говорить, что 68%±3% или, между 65% и 71% – американцы одобряют высшую меру наказания, к примеру, В СМИ обычно подчеркивается центральное значение интервала (68% в данном случае), но важно понять, что параметр населения (процент всех американцев, одобряющих смертную казнь) может быть где угодно в интервале между 65% и 71%. Рассмотрим критерии выбора оценки: смещенность и эффективность. Процедуры оценки основаны на выборочной статистике. Оценка данных может быть выбрана в соответствии с двумя критериями: смещенность и эффективность. Оценки должны основываться на выборочных статистических данных, которые являются объективными и относительно эффективными. Каждый из этих критериев рассмотрим отдельно. Начнем с понятия выборочная несмещенная дисперсия.
Опишем данные, кто сталкивался с более-менее серьезным статистическим анализом, наверняка слышал термин «несмещенная дисперсия». Некоторые даже знают, чем расчет такой дисперсии отличается от обычной. Мы об этом уже говорили, в расчетах мы используем деление не на n, а на n-1. Думаю, будет интересно узнать, в чем различие и, собственно, зачем это нужно.
Из названия «выборочная несмещенная дисперсия» видно, что она как-то связана с выборкой. Действительно, выборочная дисперсия рассчитывается по выборке данных.

С точки зрения охвата объекта исследования, статистический анализ можно разделить на два вида: сплошной и выборочный. Сплошной статанализ предполагает изучение генеральной совокупности данных, то есть всего явления во всем его многообразии без распространения выводов на другие элементы, не входящие в анализируемую совокупность. Из названия данного типа явствует, что наблюдению подвергаются тотально все элементы. Результат анализа распространяется на всю генеральную совокупность без каких-либо допущений и поправок на ошибку. Данный тип статистического исследования является наиболее полным и точным, так как дополнительные знания почерпнуть уже неоткуда – информация собрана со всех элементов объекта исследования. Это бесспорный плюс. Отличным примером сплошного наблюдения является перепись населения. Более практичный пример сплошного наблюдения – опрос жителей многоэтажного дома на предмет строительства на существующей детской площадки парковки для автомобиля. Опрашиваются все, результат дает вполне однозначный ответ об отношении жителей к данному вопросу. Ошибки в выводах маловероятны.
Как бы там ни было, у сплошного наблюдения есть отрицательное качество: на организацию и проведение исследования могут потребоваться значительные ресурсы. Одно дело взять пробу из партии товаров, другое – проверять всю партию. Одно дело опросить тысячу прохожих на улице, совсем другое – организовать перепись населения.

В противовес сплошному придумали выборочное наблюдение. Название метода точно отражает его суть: из генеральной совокупности отбирается и анализируется только часть данных, а

выводы распространяют на всю генеральную совокупность. Отбор данных происходит таким образом, чтобы выборка была репрезентативной, то есть, сохранила внутреннюю структуру и закономерности генеральной совокупности. Если это условие не соблюдено, то дальнейший анализ во многом теряет смысл.

Сам анализ выборочных данных происходит так же, как и при сплошном наблюдении (рассчитываются различные показатели, делаются прогнозы и т.д.), только с поправкой на ошибку. Это значит, что рассчитывая тот или иной показатель, мы понимаем, что при повторной выборке его значение всегда будет иным. К примеру, провели опрос общественного мнения об отношении к кандидатам в президенты. Опрос показал, что за кандидата N желают проголосовать 60% опрошенных. Если провести еще один такой же опрос, даже в том же месте, то результат будет отличаться. То есть, взяв первое значение 60%, следует понимать, что с той или иной вероятностью оно могло быть, скажем, и 55%, и 65%. Точность и разброс выборочных показателей зависят от характера данных. Имеем то, что средняя величина постоянно меняется и для решения проблемы предлагается увеличить выборку. Большая выборка, бесспорно, дает более надежные результаты, чем маленькая, но даже в этом случае ошибка сохранится, только станет меньше. А иногда и выбора нет, приходится иметь дело с маленькими выборками.
У выборочного наблюдения есть один существенный плюс и один минус, однако по сравнению со сплошным наблюдением крайности меняются местами. Плюс заключается в том, что для проведения выборочного обследования требуется гораздо меньше ресурсов. Минус – в том, что выборочное наблюдение всегда ошибочно. Поэтому основная задача проведения выборочного наблюдения – добиться максимальной точности при приемлемых затратах на его проведение. Перейдем к понятию выборочная несмещенная дисперсия.

Дисперсия, как и доля или средняя арифметическая, также меняет свое значение от выборки к выборке, но здесь есть интересная особенность. Дисперсия ведь рассчитывается от средней величины, а она в свою очередь тоже рассчитывается по выборке, то есть является ошибочной. Как же это обстоятельство влияет на саму дисперсию? Если бы мы знали истинную среднюю величину (по генеральной совокупности), то ошибка дисперсии была бы связана только с нерепрезентативностью, то есть с тем, что данные в выборке оказались бы ближе или дальше от средней, чем в целом по генеральной совокупности. При этом при многократном повторении данные стремились бы к своему реальному расположению относительно средней. Выборочный показатель, который при многократном повторении выборки стремится к своему теоретическому значению, называется несмещенной оценкой. Почему оценкой? Потому что мы не знаем реальное значение показателя (по генеральной совокупности), и с помощью выборочного наблюдения пытаемся его оценить. Оценка показателя – это есть его характеристика, рассчитанная по выборке. Примером из жизни могут служить оценки в школе. Учитель же не может влезть в мозг школьника и измерить объем знаний. Школьнику задаются вопросы, задачи, на основе чего оцениваются его знания (производится как бы выборочное наблюдение). Как и в эконометрике, оценка знаний школьника может быть ошибочна, что многие знают по себе. Почему-то только каждый считает, что его оценки занижают. Правда, учителя считают, что оценки завышают. Теперь смотрим внимательно на выборочную среднюю. Выборочная средняя – это несмещенная оценка математического ожидания, так как средняя из выборочных средних стремится к своему теоретическому значению по генеральной совокупности. Где она расположена? Средняя всегда находится в центре значений, по которым рассчитана – на то она и средняя. А раз выборочная средняя находится в центре выборки, то из этого следует, что сумма квадратов расстояний от каждого значения выборки до выборочной средней всегда меньше, чем до любой другой точки, в том числе и до генеральной средней. Это ключевой момент. А раз так, то дисперсия в каждой выборке будет занижена. Средняя из заниженных дисперсий тоже даст заниженное значение. То

есть при многократном повторении эксперимента выборочная дисперсия не будет стремиться к своему истинному значению (как выборочная средняя), а будет смещена относительно истинного значения по генеральной совокупности. Отклонение выборочной средней от генеральной показано на рисунке 8.1.

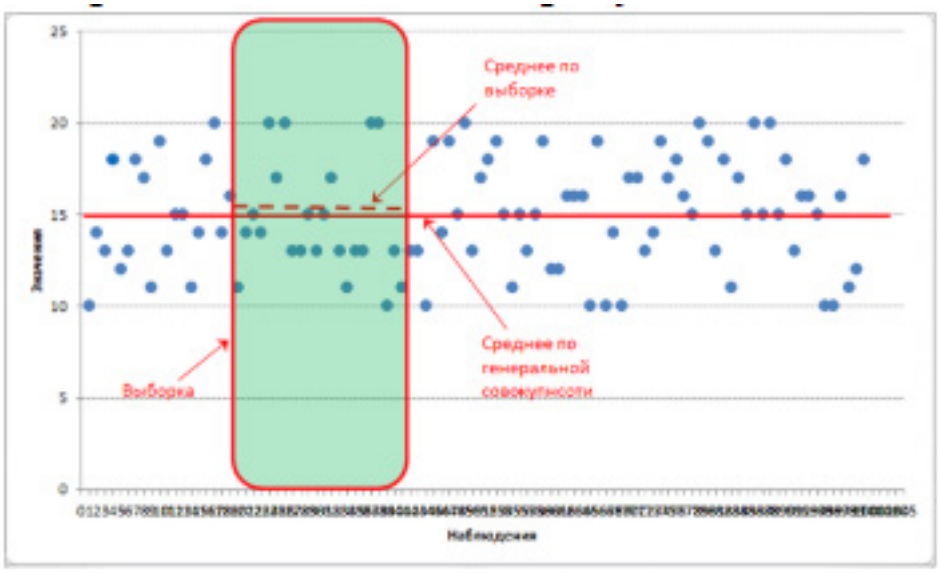

Рисунок 8.1. Отклонение выборочной средней от генеральной.

Несмещенность оценки – одна из важных характеристик статистического показателя. Смещенная оценка показателя заранее говорит о тенденции к ошибке. Поэтому показатели стараются оценивать таким образом, чтобы их оценки были несмещенными (как у средней арифметической). Для того, чтобы решить проблему смещенности оценки выборочной дисперсии в ее расчет вносят корректировку – умножают на N/(N-1), либо сразу при расчете в знаменатель ставят не N, а N-1. Получается так.

Выборочная смещенная дисперсия, расчитывается следующим образом, как мы видим в знаменателе отношения выборка N $s^2 = \frac{\sum (X_i - \overline{X})^2}{N}$

В формуле рассчета выборочной несмещенной дисперсии N-1: $s_0^2 = \frac{\sum (X_i - \overline{X})^2}{N-1}$

Под выборочной дисперсией понимают, как правило, именно несмещенный вариант.

Так, если мы возьмем выборку из 11 наблюдений, то 11/10 – это 10% относительной разницы. При 21 наблюдениях, отличие сокращается до 5%, при 31 наблюдении – до 3,3%, при 51 – до 2%, при 101 – до 1%. Короче, при достаточно большой выборке данных (50 и выше наблюдений) относительная разница между смещенной и несмещенной дисперсией практически исчезает. Оценка параметра, когда с ростом выборки его отклонение от теоретического значения уменьшается, называется асимптотически несмещенной оценкой. При переходе к среднему квадратическому отклонению по выборке (оценка среднеквадратического отклонения, равная квадратному корню из выборочной дисперсии) разница становится еще меньше. Таким образом,

эффект смещенной дисперсии проявляется в небольших выборках. В больших выборках можно использовать генеральную дисперсию, что как бы не усложняет и не упрощает жизнь. Вручную сейчас конечно, никто не считает. Все легко посчитать в Excel или SPSS. Но понимать различие в терминологии и в сути показателей все же следует.

Еще одно не менее важное понятие – эффективность оценки. Эффективной называется статистическая оценка, которая при одних и тех же объемах выборки имеет наименьшую дисперсию. В некоторых случаях становится интересным поведение оценки при неограниченном увеличении объема выборки. Чем меньший среднеквадратичное отклонение выборочного распределения, тем больше объединение в кластеры и выше эффективность. Среднеквадратичное отклонение выборочного распределения или средней квадратической ошибки среднего, равно среднеквадратичному отклонению, разделенному на квадратный корень N $\left(\sigma_{\bar{X}} = \sigma/\sqrt{N}\right)$

Мы можем повысить эффективность (или уменьшить среднеквадратичное отклонение выборочного распределения) увеличивая объем выборки. Рассмотрим пример, используя расчеты по двум выборка представленным в следующей таблице 8.1.

Таблица 8.1. Среднеквадратичное отклонение выборочного распределения двух выборок, при σ=500$

	Образец 1	Образец 2
Средний образец	$\overline{X}_1 = \$45.000$	$\overline{X}_2 = \$45.000$
Объем выборки	$N_1 = 100$	$N_2 = 1000$
Среднеквадратичное отклонение выборочного распределения	$\sigma_{\bar{X}_1} = \$50.00$	$\sigma_{\bar{X}_2} = \$15.81$

Для выборке первой, среднеквадратичное отклонение выборочного распределения всех возможных средств с N=100 составило бы 50.00$:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{500}{\sqrt{100}} = \frac{500}{10} = 50.00$$

Для второй выборки, среднеквадратичное отклонение с N=1000, показывает намного меньшую покупательскую способность 15.81$:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}} = \frac{500}{\sqrt{1000}} = \frac{500}{31{,}62} = 15.81$$

Переходим к рассмотрению доверительного интервала.
Доверительные интервалы один из типов интервальных оценок, используемых в статистике, которые рассчитываются для заданного уровня значимости. Они позволяют сделать утверждение, что истинное значение неизвестного статистического параметра генеральной совокупности

находится в полученном диапазоне значений с вероятностью, которая задана выбранным уровнем статистической значимости.
Когда известна вариация генеральной совокупности данных, для расчета доверительных пределов (граничных точек доверительного интервала) может быть использована Z-оценка, использование z-оценки позволит построить не только более узкий доверительный интервал, но и получить более надежные оценки математического ожидания и среднеквадратического (стандартного) отклонения (σ), поскольку Z-оценка основывается на нормальном распределении.

Для определения граничных точек доверительного интервала, при условии что известно среднеквадратическое отклонение генеральной совокупности данных, используется следующая формула

$$L = -Z_{\alpha/2}\frac{\sigma}{\sqrt{N}} \qquad U = \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{N}}$$

где X– математическое ожидание выборки, α – уровень статистической значимости, Zα/2 – Z-оценка для уровня статистической значимости α/2, σ – среднеквадратическое отклонение генеральной совокупности, N – количество наблюдений в выборке.

При этом, N является стандартной ошибкой. Немного о новом для нас термине уровень значимости. Уровень значимости в статистике является важным показателем, отражающим степень уверенности в точности, истинности полученных (прогнозируемых) данных. Понятие широко применяется в различных сферах: от проведения социологических исследований, до статистического тестирования научных гипотез. В научной практике уровень значимости выбирается перед сбором данных и, как правило, его коэффициент составляет 0,05 (5 %). Для систем, где крайне важны точные значения, этот показатель может составлять 0,01 (1 %) и менее. При построении доверительного интервала уровень значимости делим пополам так как определяем верхнюю и нижнюю границы интервала

Таким образом, доверительный интервал для уровня статистической значимости α можно записать в виде

$$\overline{X} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{N}}$$

Предположим, что размер выборки насчитывает 25 наблюдений, математическое ожидание выборки равняется 15, а среднеквадратическое отклонение генеральной совокупности составляет 8. Для уровня значимости α=5% Z-оценка равна Zα/2=1,96. В этом случае нижняя и верхняя граница доверительного интервала составят 11,864 и 18.136

$$L = 15 - 1.96\frac{8}{\sqrt{25}} = 11.864 \qquad L = 15 + 1.96\frac{8}{\sqrt{25}} = 18.136$$

А сам доверительный интервал может быть записан в виде
15 ± 3,136.

Таким образом, мы можем утверждать, что с вероятностью 95% математическое ожидание генеральной совокупности попадет в диапазон от 11,864 до 18,136.

Методы сужения доверительного интервала.

Допустим, что диапазон [11,864; 18,136] является слишком широким для целей нашего исследования. Уменьшить диапазон доверительного интервала можно двумя способами.
Снизить уровень статистической значимости α.
Увеличить объем выборки.
Снизив уровень статистической значимости до α=10%, мы получим Z-оценку равную Zα/2=1,64.
В этом случае нижняя и верхняя граница интервала составят 12,376 и 17,624

$$L = 15 - 1.64\frac{8}{\sqrt{25}} = 12.376$$

$$L = 15 + 1.64\frac{8}{\sqrt{25}} = 17.624$$

А сам доверительный интервал может быть записан в виде 15 ± 2,624
В этом случае, мы можем сделать предположение, что с вероятностью 90% математическое ожидание генеральной совокупности попадет в диапазон [12,376; 17,624].
Если мы хотим не снижать уровень статистической значимости α, то единственной альтернативой остается увеличение объема выборки. Увеличив ее до 144 наблюдений, получим следующие значения доверительных пределов

$$L = 15 - 1.96\frac{8}{\sqrt{144}} = 13.693$$

$$L = 15 + 1.96\frac{8}{\sqrt{144}} = 16.307$$

Сам доверительный интервал станет иметь следующий вид 15 ± 1,307
Таким образом, сужение доверительного интервала без снижения уровня статистической значимости возможно только лишь за счет увеличения объема выборки. Если увеличение объема выборки не представляется возможным, то сужение доверительного интервала может достигаться исключительно за счет снижения уровня статистической значимости.
Для удобства расчетов можно использовать таблицу Z-оценки различных уровней значимости.

Таблица 8.2. Z-оценка различных уровней значимости

Уровень значимости	Значимость α	α/2	Z оценка
90%	0.10	0.05	±1.65
95%	0.05	0.025	±1.96
99%	0.01	0.005	±2.58
99.9%	0.001	0.0005	±3.32
99.99%	0.0001	0.00005	±3.90

Выводы

1.Формула генеральной дисперсии в выборке дает смещенную оценку.

2. При большом объеме выборки (от 100 наблюдений) разница между смещенной и несмещенной дисперсиями практически исчезает.