В отличие от обычной задачи теории вероятностей (мы готовимся к эксперименту и оцениваем вероятности исходов), в статистической постановке задачи мы, как правило, уже получили результаты и хотим сделать правдоподобные утверждения об эксперименте, который был поставлен.

Рассмотрим сначала простейшую ситуацию. У нас есть некоторая случайная величина ξ, для которой существуют $M\xi$, $D\xi$. В эксперименте измеряются n независимых случайных величин, распределённых так же, как и ξ. Мы хотим (хотя бы приблизительно) узнать, каково математическое ожидание ξ.

Естественное построение: просто рассмотрим среднее из всех полученных значений.

Теперь определим, что мы должны построить в общем случае, и как проверить осмысленность построения.

Определение 1. Пусть у нас есть семейство распределений вероятностей на наборе случайных величин $\overline{\xi}$, и есть некоторая характеристика a, определённая для каждого распределения. *Оценкой* на a называется функция $\theta : \overline{\xi} \mapsto a'$, которая по любому результату эксперимента выдаёт предполагаемое (не обязательно, разумеется, точное) значение характеристики.

При этом при заданном распределении значение оценки на исходе является случайной величиной --- мы просто смотрим на набор случайных величин, и применяем к ним заранее фиксированную функцию. Каждая случайная величина была функцией на исходах, поэтому и результат применения функции, задающей метод оценки, является функцией на исходах.

Кроме того, характеристика может быть определена только для распределений из семейства распределений. Например, можно рассматривать равномерные распределения на отрезках и пытаться выяснить, где расположены концы отрезков.

Определение 2. Оценка θ на характеристику a называется *несмещённой*, если при любом фиксированном распределении вероятностей (из рассматриваемого семейства) математическое ожидание случайной величины $\theta(\overline{\xi})$ равно a.

В частности, если есть случайная величина, то просто её значение (без каких бы то ни было независимых повторов) является несмещённой оценкой математического ожидания этой случайной величины.

Но ясно, что нам хочется получить не оценку, которая с равной вероятностью намного больше или намного меньше истинного значения, а оценку, которая будет действительно лежать близко к оцениваемому значению. Определить ``близко" для фиксированных величин труднее, чем для последовательности, потому что для последовательности можно потребовать сходимости.

Определение 3. Последовательность случайных величин ξ_n *сходится по вероятности* к случайной величине η, если $\forall \varepsilon > 0\, P(|\xi_n - \eta| > \varepsilon) \to 0)$.

Замечание 1. Обратите внимание, что из этого не следует, что для какого-либо исхода ω последовательность значений $\xi_n(\omega)$ стремится к $\eta(\omega)$. Например, можно взять в качестве вероятностного пространства (множества исходов) отрезок единичной длины с равномерным распределением (вероятность интервала равна его длине), в качестве η взять константу 0.

Последовательность ξ_n построим вместе со вспомогательной последовательностью x_n. Положим $x_0 = 0$, $x_n = x + n - 1 + \frac{1}{n} \mod 1$. Величину же ξ_n определим как 1 на отрезке от x_{n-1} до x_n (если $x_n < x_{n-1}$, то имеется

в виду $[x_{n-1};1]\cup[0;x_n]$ и 0 вне него. Ясно, что $P(|\xi_n-\eta|>\varepsilon)=\frac{1}{n}\to 0$, но так как $\sum\frac{1}{n}$ расходится, то на каждом исходе бесконечно много раз будет принято значение 1.

Определение 4. Пусть есть семейство распределений вероятностей для значений случайной величины, для каждого из которых определена какая-то характеристика a. Последовательность оценок $\theta_n:\mathbb{R}^n\to\mathbb{R}$ называется *состоятельной*, если для любого распределения из семейства последовательность случайных величин $\theta_n(\xi_1,\ldots,\xi_n)$, где ξ_k независимые и имеют данное распределение, сходится по вероятности к константе a.

Докажем, что среднее имеющихся значений --- состоятельная последовательность оценок для математического ожидания (при конечных математическом ожидании и дисперсии). Действительно, в этих условиях применим закон больших чисел (дисперсия среднего падает с ростом количества усредняемых величин, математическое ожидание остаётся постоянным, и можно применить оценку Чебышёва на вероятность отклонений от математического ожидания, существенно больших дисперсии).

Теперь посмотрим, как оценивать дисперсию (предполагая, например, что математические ожидания $|\xi|^n$ конечны).

Попробуем простейший путь --- найдём среднее, а потом возьмём средний квадрат отклонения от него. Будет ли такая оценка несмещённой?

Среднее --- это $\mu=\frac{1}{n}\sum\xi_k$. Сначала заметим, что $\frac{1}{n}M\sum(\xi_k-\mu)^2=M(\xi_1-\mu)^2$, так как математическое ожидание каждого слагаемого одно и то же из симметрии. Мы можем считать, что на самом деле $M\xi_k=0$ (при этом μ может равняться или не равняться 0), так как вычитание $M\xi_k$ из ξ_k не изменит распределение $\xi_k-\mu$ (μ тоже уменьшится на $M\xi_k$). Раскроем скобки, вынесем сложение за знак математического ожидания: $M(\xi_1-\mu)^2=M\xi_1^2-2M\mu\xi_1+M\mu^2$. Заметим, что при независимых ξ_k и $M\xi_k=0$ мы знаем, что $M\xi_i\xi_j=0$ и $M\xi_i^2=D\xi_i$. Поэтому $M\xi_1^2-2M\mu\xi_1+M\mu^2=D\xi_1-2\frac{1}{n}(D\xi_1+0+\ldots+0)+\frac{1}{n^2}\sum\limits_{k=1}^{n}D\xi_k=D\xi_1(1-\frac{2}{n}+\frac{n}{n^2})=\frac{n-1}{n}D\xi_1$. Наша оценка не является несмещённой, но станет такой при умножении на $\frac{n}{n-1}$, то есть несмещённой является оценка

$$\frac{1}{n-1}M\sum(\xi_k-\mu)^2$$

Можно объяснить и то, почему исходная оценка оказалась смещённой. С одной стороны, можно сказать, что квадрат несмещённой оценки не будет несмещённой оценкой на квадрат. С другой стороны, понятна причина этого: μ смещено в сторону, в которую отклонились значения (в совокупности), поэтому ожидание квадрата отклонения от него меньше, чем относительно истинного ожидания.

Доказательство состоятельности этой оценки мы опустим, для него надо просто доказать, что усредняется линейное по n количество величин с ограниченной дисперсией.

Какие есть методы построения оценки в общем случае?

Пусть у нас есть семейство случайных величин $\xi(a,\omega)$ (a --- параметр, ω --- исход) с плотностью вероятности $f_a(x)$. Если у нас есть выборка $x_1,\ldots,x_n$, то можно посмотреть, при каком значении параметра a плотность вероятности в окрестности реально наблюдённой выборки (то есть $\prod f_a(x_k)$) будет максимальна. Эта оценка называется оценкой максимального правдоподобия. Она не обязана быть несмещённой, так как рассматривает только максимум плотности вероятности, а не всё распределение. Например, можно представить себе лампочку, которая перегорает либо в первую секунду после включения (как обычно и бывает), либо в какой-то последующий момент, причём

плотность вероятности уже намного меньше и монотонно падает. Если мы знаем, когда перегорела лампочка, то оценка максимального правдоподобия на момент включения - тогда же, хотя несмещённая оценка предскажет чуть более раннее включение.

Можно привести и пример, в котором всё посчитается. Пусть плотность вероятности ξ при параметре a равна e^{x-a}. Так как плотность вероятности для любого фиксированного набора значений $x_1, \ldots, x_n$ увеличивается при увеличении параметра a, пока все x_k больше a, оценка максимального правдоподобия --- это $\min_k x_k$. С другой стороны, $M\xi = \int\limits_0^\infty xe^{-x}dx = -\int\limits_0^\infty xde^{-x} = -xe^{-x}\,|_0^\infty + \int\limits_0^\infty e^{-x}dx = -0 + 1 = 1$, и несмещённой оценкой через математическое ожидание будет $\frac{1}{n}\sum\limits_k x_k - 1$.

Оценка максимального правдоподобия несмещённой не будет. Действительно, пусть a задано. Для удобства расчётов будем рассматривать $a = 0$. Математическое ожидание оценки максимального правдоподобия не может не быть положительным, так как это усреднение положительных чисел.

При этом можно оценить, насколько оценка отклоняется от истинного значения при больших n. Оценивать, как и в случае с дисперсией, будем средний квадрат отклонения. Для несмещённой оценки мы получим просто дисперсию, которая убывает как $\frac{1}{n}$. Для оценки максимального правдоподобия придётся сначала найти плотность. Если плотность равна e^{-x}, то вероятность превышения значения x (уже вероятность, а не плотность!) будет равна e^{-x}. Вероятность превышения минимумом из n независимых повторений значения x равна вероятности, что все повторы превысят x, то есть e^{-nx}. Ожидание квадрата отклонения равно $\int\limits_0^\infty x^2e^{-nx}dx = -\int\limits_0^\infty x^2de^{-nx} \overset{t=nx}{=} -\frac{1}{n^2}\int\limits_0^\infty t^2de^{-t}$. Вычислять значение интеграла мы не будем, так как это просто константа. Мы видим, что ожидаемый квадрат отклонения для оценки максимального правдоподобия убывает намного быстрее, чем для оценки по среднему. Можно себе представлять узкий интервал с высокой плотностью чуть справа от истинного значения, который имеет меньшее значение среднего квадрата отклонения, чем приблизительно нормальное распределение с большой дисперсией и правильным средним.

В частности, если мы считаем, что шум ненастроенного прибора имеет всегда нормальное распределение, то, подбирая зависимость показаний от реального значения по методу наименьших квадратов, мы строим оценку максимального правдоподобия. Максимум $\prod e^{-(\Delta x)^2} = exp(-\sum(\Delta x)^2)$ достигается там же, где и минимум $\sum(\Delta x)^2$.

Рассматривают также та называемую *Байесову* оценку. Её применение требует, чтобы у нас заранее были представления о вероятностях разных параметров a. Отличие самой оценки в том, что в максимизируемое произведение добавляют множитель $g(a)$ --- плотность вероятности для самого параметра. Оценка максимального правдоподобия --- это частный случай Байесовой оценки при $g(a) = const$.

В конце определим ещё одно понятие --- доверительный интервал.

Пусть есть семейство случайных величин $\xi(a, \omega)$. Мы говорим, что θ, по выборке размера n возвращающая интервал, задаёт оценку с доверительной вероятностью ε, если при любом фиксированном a вероятность получения выборки, на которой θ укажет интервал, не содержащий a, меньше ε. Обратите внимание, что эта вероятность не имеет никакого прямого формального отношения к достоверности оценки для данной выборки --- это только свойство метода.

В частности, когда при социологическом опросе говорят, что статистическая погрешность составляет 3%, имеют в виду (предположим, что опрос проведён честно), что из 1000 опросов с такой методологией отклонение доли опрошенных, ответивших ``да", от доли людей, которые ответили бы ``да" на вопрос, превышает 3% (в среднем) в 3

опросах.

Из неумышленных ошибок в опросах стоит отметить неверность предположения о случайной независимой выборке --- часто готовность человека ответить на вопрос связана с тем, какой ответ он даст. Например, можно ожидать, что при телефонном опросе по случайным номерам многие из тех, кто считает, что использование автоматических звонков для рекламы должно наказываться закрытием организации, бросят трубку, не выслушав вопрос.