Задание

1. Ознакомиться с теоретический материалом темы, законспектировать.

2. Фотоотчет и сообщение присылать на электронную почту

С уважением, Хвастова Светлана Ивановна

!!! Если возникнут вопросы обращаться по телефону 0721389311 (ватсап). Электронная почта: xvsviv@rambler.ru

Тема лекции: «**Точечные оценки неизвестных параметров распределения**»

Цели: объяснить понятие точечных оценок неизвестных параметров распределения, выучить основные понятия темы, рассмотреть решение типовых задач.

Содержание

1 Основные понятия

2 Способы построения оценок

2.1 Метод наибольшего правдоподобия.

2.2 Метод моментов.

2.3 Метод наименьших квадратов.

2.4 Байесовский подход к получению оценок.

Основные понятияПолучив статистические оценки параметров распределения (выборочное среднее, выборочную дисперсию и т.д.), нужно убедиться, что они в достаточной степени служат приближением соответствующих характеристик генеральной совокупности. Определим требования, которые должны при этом выполняться.

Пусть Θ* - статистическая оценка неизвестного параметра Θ теоретического распределения. Извлечем из генеральной совокупности несколько выборок одного и того же объема *n* и вычислим для каждой из них оценку параметра Θ: $\Theta_1^*, \Theta_2^*, ..., \Theta_k^*$. Тогда оценку Θ* можно рассматривать как случайную величину, принимающую возможные значения $\Theta_1^*, \Theta_2^*, ..., \Theta_k^*$. Если математическое ожидание Θ* не равно оцениваемому параметру, мы будем получать при вычислении оценок систематические ошибки одного знака (с

избытком, если $M(\Theta^*) > \Theta$, и с недостатком, если $M(\Theta^*) < \Theta$). Следовательно, необходимым условием отсутствия систематических ошибок является требование $M(\Theta^*) = \Theta$.

Определение 1. Статистическая оценка Θ^* называется **несмещенной**, если ее математическое ожидание равно оцениваемому параметру Θ при любом объеме выборки:

$$M(\Theta^*) = \Theta. \qquad (1)$$

Смещенной называют оценку, математическое ожидание которой не равно оцениваемому параметру.

Однако несмещенность не является достаточным условием хорошего приближения к истин-ному значению оцениваемого параметра. Если при этом возможные значения Θ^* могут значительно отклоняться от среднего значения, то есть дисперсия Θ^* велика, то значение, найденное по данным одной выборки, может значительно отличаться от оцениваемого параметра. Следовательно, требуется наложить ограничения на дисперсию.

Определение 2. Статистическая оценка называется **эффективной**, если она при заданном объеме выборки n имеет наименьшую возможную дисперсию.

При рассмотрении выборок большого объема к статистическим оценкам предъявляется еще и требование состоятельности.

Определение 3. **Состоятельной** называется статистическая оценка, которая при $n \to \infty$ стремится по вероятности к оцениваемому параметру (если эта оценка несмещенная, то она будет состоятельной, если при $n \to \infty$ ее дисперсия стремится к 0).

Убедимся, что $\bar{x}_B$ представляет собой несмещенную оценку математического ожидания $M(X)$.

Будем рассматривать $\bar{x}_B$ как случайную величину, а $x_1, x_2, \ldots, x_n$, то есть значения исследуемой случайной величины, составляющие выборку, – как независимые, одинаково распределенные случайные величины $X_1, X_2, \ldots, X_n$,

имеющие математическое ожидание a. Из свойств математического ожидания следует, что

$$M(\overline{X}_B) = M\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) = a.$$

Но, поскольку каждая из величин X_1, X_2,…, X_n имеет такое же распределение, что и генеральная совокупность, $a = M(X)$, то есть $M(\overline{X}_B) = M(X)$, что и требовалось доказать. Выборочное среднее является не только несмещенной, но и состоятельной оценкой математического ожидания. Если предположить, что X_1, X_2,…, X_n имеют ограниченные дисперсии, то из теоремы Чебышева следует, что их среднее арифметическое, то есть $\overline{X}_B$, при увеличении n стремится по вероятности к математическому ожиданию a каждой их величин, то есть к $M(X)$. Следовательно, выборочное среднее есть состоятельная оценка математического ожидания.

В отличие от выборочного среднего, выборочная дисперсия является смещенной оценкой дисперсии генеральной совокупности. Можно доказать, что

$$M(D_B) = \frac{n-1}{n} D_Г, \qquad (2)$$

где $D_Г$ – истинное значение дисперсии генеральной совокупности. Можно предложить другую оценку дисперсии – **исправленную дисперсию s^2**, вычисляемую по формуле

$$s^2 = \frac{n}{n-1} D_B = \frac{\sum_{i=1}^{k} n_i (x_i - \overline{x}_B)^2}{n-1}. \qquad (3)$$

Такая оценка будет являться несмещенной. Ей соответствует **исправленное среднее квадратическое отклонение**

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k} n_i (x_i - \overline{x}_B)^2}{n-1}}. \qquad (4)$$

Определение 17.4. Оценка некоторого признака называется **асимптотически несмещенной**, если для выборки x_1, x_2, …, x_n

$$\lim_{n\to\infty}\frac{x_1+x_2+\ldots+x_n}{n}=X\,, \qquad (5)$$

где X – истинное значение исследуемой величины.

2 Способы построения оценок.

2.1 Метод наибольшего правдоподобия.

Пусть X – дискретная случайная величина, которая в результате n испытаний приняла значения x_1, x_2, …, x_n. Предположим, что нам известен закон распределения этой величины, определяемый параметром Θ, но неизвестно численное значение этого параметра. Найдем его точечную оценку.

Пусть $p(x_i, \Theta)$ – вероятность того, что в результате испытания величина X примет значение x_i. Назовем **функцией правдоподобия** дискретной случайной величины X функцию аргумента Θ, определяемую по формуле:

$$L\,(x_1, x_2, \ldots, x_n; \Theta) = p(x_1,\Theta)p(x_2,\Theta)\ldots p(x_n,\Theta).$$

Тогда в качестве точечной оценки параметра Θ принимают такое его значение $\Theta^* = \Theta(x_1, x_2, \ldots, x_n)$, при котором функция правдоподобия достигает максимума. Оценку Θ* называют **оценкой наибольшего правдоподобия**.

Поскольку функции L и $\ln L$ достигают максимума при одном и том же значении Θ, удобнее искать максимум $\ln L$ – **логарифмической функции правдоподобия**. Для этого нужно:

1) найти производную $\frac{d\ln L}{d\Theta}$;

2) приравнять ее нулю (получим так называемое *уравнение правдоподобия*) и найти критическую точку;

3) найти вторую производную $\frac{d^2\ln L}{d\Theta^2}$; если она отрицательна в критической точке, то это – точка максимума.

Достоинства метода наибольшего правдоподобия: полученные оценки состоятельны (хотя могут быть смещенными), распределены асимптотически нормально при больших значениях n и имеют наименьшую дисперсию по

сравнению с другими асимптотически нормальными оценками; если для оцениваемого параметра Θ существует эффективная оценка Θ*, то уравнение правдоподобия имеет единственное решение Θ*; метод наиболее полно использует данные выборки и поэтому особенно полезен в случае малых выборок.

Недостаток метода наибольшего правдоподобия: сложность вычислений.

Для непрерывной случайной величины с известным видом плотности распределения $f(x)$ и неизвестным параметром Θ функция правдоподобия имеет вид:

$$L\,(x_1, x_2, \ldots, x_n; \Theta) = f(x_1,\Theta)f(x_2,\Theta)\ldots f(x_n,\Theta).$$

Оценка наибольшего правдоподобия неизвестного параметра проводится так же, как для дискретной случайной величины.

2.2 Метод моментов.

Метод моментов основан на том, что начальные и центральные эмпирические моменты являются состоятельными оценками соответственно начальных и центральных теоретических моментов, поэтому можно приравнять теоретические моменты соответствующим эмпирическим моментам того же порядка.

Если задан вид плотности распределения $f(x, \Theta)$, определяемой одним неизвестным параметром Θ, то для оценки этого параметра достаточно иметь одно уравнение. Например, можно приравнять начальные моменты первого порядка:

$$\bar{x}_B = M(X) = \int_{-\infty}^{\infty} x f(x;\Theta)dx = \varphi(\Theta),$$

получив тем самым уравнение для определения Θ. Его решение Θ* будет точечной оценкой параметра, которая является функцией от выборочного среднего и, следовательно, и от вариант выборки:

$$\Theta = \psi\ (x_1, x_2, \ldots, x_n).$$

Если известный вид плотности распределения $f(x, \Theta_1, \Theta_2)$ определяется двумя неизвестными параметрами Θ_1 и Θ_2, то требуется составить два уравнения, например

$$\nu_1 = M_1,\ \ \mu_2 = m_2.$$

Отсюда $\begin{cases} M(X) = \bar{x}_B \\ D(X) = D_B \end{cases}$ - система двух уравнений с двумя неизвестными Θ_1 и Θ_2. Ее решениями будут точечные оценки Θ_1* и Θ_2* - функции вариант выборки:

$$\Theta_1 = \psi_1\ (x_1, x_2, \ldots, x_n),$$

$$\Theta_2 = \psi_2(x_1, x_2, \ldots, x_n).$$

2.3 Метод наименьших квадратов.

Если требуется оценить зависимость величин y и x, причем известен вид связывающей их функции, но неизвестны значения входящих в нее коэффициентов, их величины можно оценить по имеющейся выборке с помощью метода наименьших квадратов. Для этого функция $y = \varphi\ (x)$ выбирается так, чтобы сумма квадратов отклонений наблюдаемых значений $y_1, y_2, \ldots, y_n$ от $\varphi(x_i)$ была минимальной:

$$\sum_{i=1}^{n} (y_i - \varphi(x_i))^2 = \min.$$

При этом требуется найти стационарную точку функции $\varphi(x;\ a,\ b,\ c\ldots)$, то есть решить систему:

$$\begin{cases} \sum_{i=1}^{n} (y_i - \varphi(x_i; a, b, c...)) \left(\frac{\partial \varphi}{\partial a} \right)_i = 0 \\ \sum_{i=1}^{n} (y_i - \varphi(x_i; a, b, c...)) \left(\frac{\partial \varphi}{\partial b} \right)_i = 0 \\ \sum_{i=1}^{n} (y_i - \varphi(x_i; a, b, c...)) \left(\frac{\partial \varphi}{\partial c} \right)_i = 0 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \end{cases}$$

(решение, конечно, возможно только в случае, когда известен конкретный вид функции φ).

Рассмотрим в качестве примера подбор параметров линейной функции методом наименьших квадратов.

Для того, чтобы оценить параметры a и b в функции $y = ax + b$, найдем $\left(\frac{\partial\varphi}{\partial a}\right)_i = x_i; \left(\frac{\partial\varphi}{\partial b}\right)_i = 1.$

Тогда $\begin{cases} \sum_{i=1}^{n}(y_i - (ax_i + b))x_i = 0 \\ \sum_{i=1}^{n}(y_i - (ax_i + b)) = 0 \end{cases}$.

Отсюда $\begin{cases} \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i - bn = 0 \end{cases}$.

Разделив оба полученных уравнения на n и вспомнив определения эмпирических моментов, можно получить выражения для a и b в виде:

$$a = \frac{(K_{xy})_B}{(D_x)_B}, \quad b = \bar{y}_B - \frac{(K_{xy})_B}{(D_x)_B}\bar{x}_B.$$

Следовательно, связь между x и y можно задать в виде:

$$y - \bar{y}_B = \frac{(K_{xy})_B}{(D_x)_B}(x - \bar{x}_B).$$

2.4 Байесовский подход к получению оценок.

Пусть (Y, X) – случайный вектор, для которого известна плотность $p(y|x)$ условного распреде-ления Y при каждом значении $X = x$. Если в результате эксперимента получены лишь значения Y, а соответствующие значения X неизвестны, то для оценки некоторой заданной функции $\varphi(x)$ в качестве ее приближенного значения предлагается искать условное математическое ожидание $M(\varphi(x)|Y)$, вычисляемое по формуле:

$\psi(Y)=\dfrac{\int\varphi(x)p(Y\mid x)p(x)d\mu(x)}{q(Y)}$, где $q(y)=\int p(y\mid x)p(x)d\mu(x)$, $p(x)$ – плотность безусловного распределения X, $q(y)$ – плотность безусловного распределения Y. Задача может быть решена только тогда, когда известна $p(x)$. Иногда, однако, удается построить состоятельную оценку для $q(y)$, зависящую только от полученных в выборке значений Y.

3 Решение типовых задач

<u>Пример 1.</u> Найти выборочную среднюю $\overline{x_в}$, выборочную дисперсию $D_в$, выборочное среднее квадратическое отклонение $\delta_в$ и исправленную выборочную дисперсию S^2 по данному распределению выборки.

x_i	1	2	5	6
n_i	2	3	4	1

<u>Решение.</u> Объём выборки n=$\sum_{i=1}^{4} n_i = 10$; $\overline{x_в}=\frac{1}{10}(2\cdot1+3\cdot2+4\cdot5+1\cdot6)=3{,}4.$

Для вычисления выборочной дисперсии используем формулу

$$D_в=\overline{x^2}-(\bar{x})^2=\frac{\sum_{i=1}^{k} n_i x_i^2}{n}-\left(\frac{\sum_{i=1}^{k} n_i x_i}{n}\right)^2:$$

$$D_в=\frac{2\cdot1^2+3\cdot2^2+4\cdot5^2+1\cdot6^2}{10}-3{,}4^2=15-11{,}56=3{,}44;$$

$$\delta_в=\sqrt{3{,}44}\approx1{,}85;$$

$$S^2=\frac{n}{n-1}D_в=\frac{10}{9}\cdot3.44\approx3{,}82;$$

$$S=\sqrt{3{,}82\approx1{,}95}.$$

<u>Пример 2.</u> Найти выборочную среднюю $\overline{x_в}$, выборочную дисперсию $D_в$ и выборочное среднее квадратичное отклонение $\delta_в$ по данному распределению выборки.

x_i	83	85	87	89	91	93	95	97	99	101
n_i	6	7	12	15	30	10	8	6	4	2

Решение. Объём выборки $n=\sum n_i=100$. Так как варианты равноотстоящие, то перейдём к условным вариантам $\mathbf{u_i}$; h=2.

Пусть c=91 (варианта 91 расположена примерно в середине вариационного ряда).

Тогда $u_i = \frac{x_i - 91}{2}$. Для упрощения вычислений составим расчётную таблицу.

x_i	n_i	u_i	$n_i u_i$	$n_i u_i^2$	$n_i (u_i+1)^2$
83	6	-4	-24	96	54
85	7	-3	-21	63	28
87	12	-2	-24	48	12
89	15	-1	-15	15	0
91	30	0	0	0	30
93	10	1	10	10	40
95	8	2	16	32	72
97	6	3	18	54	96
99	4	4	16	64	100
101	2	5	10	50	72
	n=100	$\sum n_i u_i$ = -14	$\sum n_i u_i^2$ =432		$\sum n_i (u_i+1)^2$ =504

Для контроля вычислений пользуемся тождеством

$$\sum_{i=1}^{k} n_i (u_i+1)^2 = \sum_{i=1}^{k} n_i u_i^2 + 2\sum_{i=1}^{k} n_i u_i + n.$$

В данном случае $\sum n_i (u_i+1)^2$ =504;

$$\sum_{i=1}^{k} n_i (u_i+1)^2 = \sum_{i=1}^{k} n_i u_i^2 + 2\sum_{i=1}^{k} n_i u_i + n = 432-28+100=504.$$

Совпадение контрольных сумм свидетельствует о правильности вычислений.

$$\overline{u_в} = \frac{1}{n}\sum n_i u_i = -0{,}14;$$

$$D_в(\mathrm{u}) = \frac{1}{n}\sum n_i u_i^2 - \left(\overline{u_в}\right)^2 = 4{,}32-0{,}0196=4{,}30.$$

Используя формулы

$$\overline{x_в} = h \cdot \overline{u_в} + c;$$
$$D_в(X) = h^2 \cdot D_в(u),$$

находим

$\overline{x_в}$ =-0,14·2+91=90,72;

$D_в$ (X)=4,30·4=17,20;

$\delta_в$ (X)= $\sqrt{17,20} \approx$ 4,15.

Контрольные вопросы:

1. Что называется параметром распределения *a* случайной величины?
2. Точечной оценкой параметра распределения случайной величины называется?
3. Оценка параметра называется состоятельной, если…?
4. Виды оценок неизвестных параметров. Охарактеризуйте их.
5. Требования несмещенности.
6. Перечислите и охарактеризуйте способы построения оценок