СМАДЗУ, Лекция 3.

Элементы математической статистики

Выборочный метод.

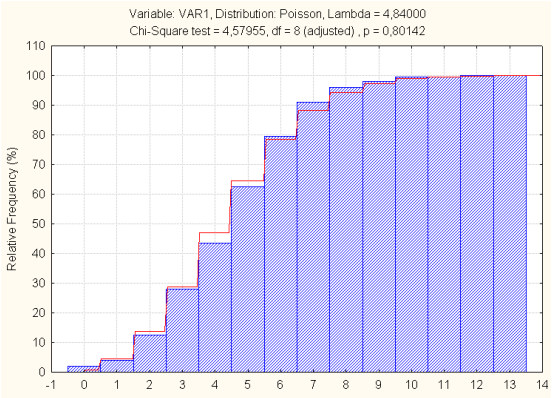
Пусть в одинаковых условиях проводится ***n*** опытов и пусть в результате этих опытов мы получили измерения x_1, x_2, …, x_n (в общем случае – это ***n*** случайных величин). Так как результаты каждого из этих опытов случайны, то каждую из этих случайных величин можно считать конкретной реализацией какой-то одной случайной величины X, которая называется **генеральной совокупностью**.

Принято говорить, что случайная последовательность x_1, x_2, …, x_n является **выборкой** из генеральной совокупности, число данных ***n*** – **объем выборки**, а величина $R = x_{max} - x_{min}$ – **размах выборки**.

Выборочные значения, записанные в порядке их регистрации обычно неудобны для дальнейшего анализа, поэтому проводят **первичную обработку выборки**. Это либо

1) ***Упорядочение*** применяемое обычно в случае малого объема выборки (*n~50÷100*). Чаще всего выборку упорядочивают по возрастанию $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Такая упорядоченная выборка называется **вариационным рядом**. На основе вариационного ряда можно построить **эмпирическую функцию распределения**:

$$F_n(y) = \begin{cases} 0, & \text{если } y \leqslant X_{(1)}, \\ \dfrac{k}{n}, & \text{если } X_{(k)} < y \leqslant X_{(k+1)}, \\ 1 & \text{при } y > X_{(n)}. \end{cases}$$



Легко видеть, что $F_n(y)$ есть кусочно постоянная неубывающая, непрерывная справа функция со скачками в точках $x_{(i)}$. Если все $x_{(i)}$ различны, то величина любого скачка равна *1/n*. Если какие-либо из $x_{(i)}$ совпадают, то

соответствующие скачки суммируются. Следует иметь в виду, что $F_n(y)$ является случайной функцией, т.к. она зависит от случайных значений выборки $x_1, x_2, \ldots, x_n$.

Порядковые статистики. Когда наблюдения в выборке располагаются в порядке вариационного ряда, каждое из упорядоченных значений – это значение случайной величины, **называемой порядковой статистикой**; k-е значение называется статистикой k-го порядка.

Центр распределения может быть оценен с помощью статистики n/2-порядка, называемой **медианой**. Выборочные квантили также вычисляются по порядковым статистикам.

Величина $R_n = x_{(n)} - x_{(1)}$ называется **размахом выборки** и может служить для оценки рассеяния распределения.

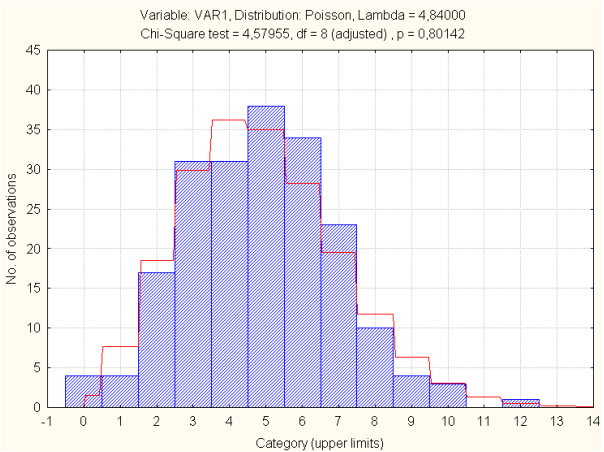
Значение эмпирической функции распределения состоит в том, что при большом числе наблюдений по ней со сколь угодно высокой точностью можно восстановить неизвестную генеральную функцию распределения $F(x)$. Как гласит известная **Теорема Гливенко:**

$$\sup_y |F_n(y) - F(y)| \to 0 \text{ при } n \to \infty$$

с вероятностью 1.

2) ***Группировка.*** При очень большом объеме выборки ее элементы объединяют в группы, представляя результаты опытов в виде **группированного статистического ряда**. Для этого интервал, содержащий все значения выборки, разбивают на m непересекающихся интервалов (удобнее разбивать на равные интервалы). При этом считается, что правая
граница интервала принадлежит следующему интервалу.

Для наглядного представления группированного статистического ряда используется специальная таблица – **гистограмма**, показывающая какое число n_i



элементов выборки попало в ***i***-й интервал (*i=1,2,…m*). Очевидно, что $\sum_{i=1}^{m} n_i = n$. Чтобы гистограмма каким-то образом приближала плотность вероятности распределения, к которому принадлежит генеральная совокупность, необходимо нормировать гистограмму, поделив каждое n_i на n, так что каждому интервалу будет соответствовать относительная частота $v_i = n_i/n$. Построенная гистограмма называется **гистограммой относительных частот.**

На рисунке вверху, наоборот, красным выделена **гистограмма теоретических частот**, полученных умножением вероятности попадания в интервал гистограммы p_i на величину n. Сумма квадратов разностей между реальными частотами n_i и их теоретическими значениями служит показателем близости экспериментальной выборки к теоретическому распределению.

Если рассматривать выборку $x_1, x_2, \ldots, x_n$ как некую механическую систему, в которой в точках с координатами x_i расставлены одинаковые веса 1/n, то величины центра тяжести этой системы и ее момент инерции определяют важные характеристики выборки: ее **выборочное среднее** $\bar{x} = \frac{1}{n}\sum_{i=1}^{m} x_i$ и **выборочную дисперсию** $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, т.е. меру рассеяния выборки по отношению к ее выборочному среднему. Корень квадратный из дисперсии называют **средним квадратическим отклонением**: $s = \sqrt{s^2}$.

Точечные оценки
(свойства и методы нахождения)

По выборке можно вычислить **точечные оценки** числовых параметров распределения, т.е. такие **функции от выборочных наблюдений**, значения которых хотя и **случайны**, но с определяемой степенью достоверности могут приниматься в качестве приближенных значений искомого параметра.

Рассмотрим несколько примеров точеных оценок. Пусть дана некоторая выборка: x_1, x_2, …, x_n. Тогда в качестве оценки центра распределения можно рассматривать величину выборочного среднего: $\bar{x} = \frac{1}{n}\sum_{i=1}^{m} x_i$. Для нормального распределения с параметрами $(0,\sigma^2)$ хорошей оценкой параметра σ может служить среднеквадратическое отклонение выборки s.

Однако по данной выборке для каждого из параметров распределения можно получить несколько разных оценок. Например, центр распределения можно оценить с помощью медианы - среднего члена вариационного ряда, т.е. величины $x_{((n+1)/2)}$, параметр рассеяния – с помощью размаха выборки $R=x_{(n)}-x_{(1)}$. Поэтому важно понять, какими критериями следует руководствоваться при сравнении оценок и выборе наилучшей из них.

В математической статистике известны следующие свойства точечных оценок:

- состоятельность;
- несмещенность;
- эффективность в данном классе оценок.

- Оценка $\hat{\theta}_n(X)$ называется **состоятельной** оценкой параметра θ, если с ростом объема выборки она сходится по вероятности к оцениваемому параметру, т.е. при $n\to\infty$

$$P\left(\left|\hat{\theta}_n(X)-\theta\right|>\varepsilon\right)\to 0, \forall\varepsilon$$

- Оценка $\hat{\theta}_n(X)$ называется **несмещенной** оценкой параметра θ, если для любого фиксированного n ее математическое ожидание равно значению этого параметра, т.е. $M\hat{\theta}_n(X)=\theta$. Для смещенной оценки величина **смещения** равна $b(\theta)=M\tilde{\theta}(X)-\theta$, а

среднеквадратическая ошибка смещенной оценки имеет вид $M(\tilde{\theta}-\theta)^2 = D\tilde{\theta} + b^2(\theta)$.

- Если в некотором классе несмещенных оценок параметра θ, имеющих конечную дисперсию, существует оценка $\hat{\theta}(X)$ такая, что ее дисперсия является минимальной в данном классе оценок, т.е. неравенство $D\hat{\theta}(X) \le D\tilde{\theta}(X)$ выполняется для всех оценок $\tilde{\theta}(X)$ из этого класса, то оценка $\hat{\theta}(X)$ называется **эффективной**.

Примеры применения этих критериев. Нетрудно проверить состоятельность и несмещенность обеих оценок центра распределения, введенных выше, как выборочного среднего $\bar{x}$, так и медианы ***med***. Однако сравнение дисперсий этих оценок показывает, что в случае нормального распределения $D\bar{x} / D_{med} \approx 0.6366$, т .е. выборочное среднее является более эффективной оценкой параметра центра распределения, чем медиана. Однако это не так в случае других распределений, например, для показательного распределения обе эти оценки $\bar{x}$ и ***med*** – одинаково эффективны, т.к. имеют одинаковые дисперсии, а в случае распределения Коши $\bar{x}$ вообще не может использоваться для оценки центра распределения и только медиана пригодна для этой цели.
Примером смещенной оценки может служить выборочная дисперсия $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$,

В самом деле, рассмотрим выборку x_1, x_2, …, x_n из распределения со средним $m=Mx_i$ и дисперсией $\sigma^2 = M(x_i - m)^2$ и посмотрим, будет ли математическое ожидание s^2 совпадать с σ^2, т.е. $Ms^2=\sigma^2$? Рассмотрим $Ms^2 = M[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2]$.

Прибавим и вычтем m внутри скобок, возведем полученное в квадрат и используем конечность суммы для

перестановки знаков математического ожидания и суммы. Получим

$$M\{\frac{1}{n}\sum_{i=1}^{n}[(x_i-m)-(\bar{x}-m)]^2\}=\frac{1}{n}\sum_{i=1}^{n}M(x_i-m)^2-M[\frac{2}{n}\sum_{i=1}^{n}(x_i-m)(\bar{x}-m)]+\frac{1}{n}\sum_{i=1}^{n}M[(\bar{x}-m)^2]$$

Очевидно, что первое слагаемое в правой части равно σ^2, а для вычисления второго воспользуемся тем, что $\frac{2}{n}\sum_{i=1}^{n}(x_i-m)(\bar{x}-m)=2(\bar{x}-m)\{\frac{1}{n}\sum_{i=1}^{n}x_i-\frac{1}{n}\sum_{i=1}^{n}m\}=2(\bar{x}-m)^2$. Поскольку величина $M(\bar{x}-m)^2$ - это дисперсия выборочного среднего, равная $D\bar{x}=\frac{\sigma^2}{n}$, то мы получаем, что сумма второго и третьего слагаемых равна $-\sigma^2/n$. Итак мы получили, что $Ms^2=\sigma^2-\sigma^2/n=\sigma^2(n-1)/n$. Величина отрицательного смещения $-\sigma^2/n$ убывает с объемом выборки. Такую смещенность оценки можно легко устранить если домножить выражение для s^2 на величину $n/(n-1)$, т.е. взять вместо s^2 уже несмещенную оценку $\tilde{s}^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$.

Методы получения оценок

1. Метод моментов

Это наиболее простой метод, заключающийся в применении выборочных моментов для получения оценок соответствующих параметров. Мы уже знакомы с оценками среднего $\bar{x}$ и дисперсии $s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$ по первому и второму выборочным моментам. Аналогично можно с помощью третьего и четвертого центрального выборочного момента получить оценки таких параметров распределения, как асимметрия и коэффициент эксцесса.

2.Метод максимального правдоподобия. Пусть ξ - случайная величина с плотностью $f(t;\theta)$, $\theta=(\theta_1,...,\theta_r)$, $\theta\in\Theta$. Совместная плотность наблюдений, рассматриваемая как функция от параметра, называется *функцией правдоподобия* (ФП) выборки. В нашем случае она имеет вид

$L(X_1,...,X_n;\theta)=\prod_{i=1}^{n} f(X_i;\theta)$. Основная идея метода состоит в том, что в качестве оценок параметров предлагается взять значения (из области Θ), доставляющие максимум ФП при данной выборке.

Оценкой максимального правдоподобия (ММП-оценкой) параметра θ называют такую функцию выборочных значений $\hat{\theta}(X)$, что для любой реализации *x* выборки *X* значения $\hat{\theta}(x)$ удовлетворяли условию $L(x;\hat{\theta})=\max L(x;\theta)$ по всем $\theta\in\Theta$. Нахождение такой оценки сводится к решению системы *r* уравнений $\frac{\partial L(x;\theta)}{\partial\theta_k}=0$, $k=1,...,r$. Поскольку функция $\ln L$ имеет максимум в той же точке, что и функция L, обычно систему уравнений правдоподобия записывают в виде $\frac{\partial \ln L(x;\theta)}{\partial\theta_k}=0$, $k=1,...,r$. Оценки, полученные методом максимального правдоподобия, состоятельны, асимптотически эффективны и асимптотически нормальны.

***Пример*.** Дана выборка X объема *n* из генеральной совокупности с нормальным $N(\mu,\sigma^2)$ законом распределения. Найти методом максимального правдоподобия точечные оценки параметров μ и σ^2.

Выпишем функцию правдоподобия выборки $L(X;\mu,\sigma^2)=\frac{1}{\sigma^n(2\pi)^{n/2}}\exp\left\{-\frac{\sum(X_i-\mu)^2}{2\sigma^2}\right\}$, и ее логарифмическую функцию правдоподобия $\ln L=-n\ln\sqrt{2\pi}-n\ln\sigma-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2$

Уравнения правдоподобия имеют вид $\begin{cases}\frac{\partial \ln L(x;\mu,\sigma^2)}{\partial\mu}\equiv\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0,\\ \frac{\partial \ln L(x;\mu,\sigma^2)}{\partial\sigma}\equiv-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i-\mu)^2=0.\end{cases}$

Решая систему уравнений относительно μ и σ^2, получим искомые оценки параметров $\hat{\mu}=\overline{X}$ $\hat{\sigma}^2=S^2$

, .

Примеры оценки параметров методом максимального правдоподобия

1. Распределение Пуассона $p_k = e^{-\lambda}\frac{\lambda^k}{k!}$

Выборка $k_1, k_2, \ldots k_n$

$$l = \ln\left(\prod_{i=1}^{n} p_{k_i}\right) = -\lambda n + \ln(\lambda)\sum_i k_i - \sum_i \ln(k_i!)$$

$$\frac{dl}{d\lambda} = -n + \frac{\sum_i k_i}{\lambda} = 0 \quad , \quad \text{откуда} \quad \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} k_i$$

2. Распределение Релея $f(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{2\sigma^2}}, 0 \le x < \infty$

Выборка $x_1, x_2, \ldots x_n$

$$l = \ln\left(\prod_{i=1}^{n} f(x_i)\right) = -2n\ln(\sigma) - \frac{1}{2\sigma^2}\sum_i x_i^2$$

$$\frac{dl}{d\sigma} = -\frac{2n}{\sigma} + \frac{2}{2\sigma^3}\sum_i x_i^2 = 0$$ т.к. $\sigma \neq 0$, умножаем на σ^3

$$-2n\sigma^2 + \sum_i x_i^2 = 0 \quad \text{откуда} \quad \hat{\sigma}^2 = \frac{1}{2n}\sum_{i=1}^{n} x_i^2$$

Распределения: равномерное и Коши – оценки из вариационного ряда!

Распределения: биномиальное и показательное – самостоятельный вывод.