Лекция 8 23.04.2012

СТАТИСТИКА

Выборки и выборочные статистики

Одними из основных понятий математической статистики являются понятия генеральной совокупности и выборки.

Генеральная совокупность. Содержательно генеральную совокупность можно понимать как множество всех потенциально возможных наблюдений в рамках изучаемой задачи. Если, например, исследуется распределение дохода в определенном обществе, то генеральная совокупность − это множество всех возможных значений дохода членов этого общества. Если речь идет о выборе из двух альтернатив (кандидат от демократической или республиканской партий), то генеральная совокупность− это количество всех избирателей (или их доля во всем обществе), голосующих за соответствующего кандидата.

Более формально, генеральная совокупность отождествляется с некоторой случайной величиной *X*, которая соответствует изучаемому явлению. Так, в примере с доходом *X* − это доход наугад выбранного человека из генеральной совокупности или, что равносильно, распределение дохода в изучаемой совокупности. В примере с выборами соответствующая случайная величина − это бернуллиевская случайная величина ε.

Будем обозначать $E(X) = m_X,\ V(X) = \sigma_X^2$

Выборка. Выборка объема *n* − это совокупность наблюдений $x_1,\ x_2,...,x_n$, взятых из генеральной совокупности с помощью процедуры простого случайного выбора, согласно которой любой набор объема *n* имеет одинаковый шанс попасть в выборку. С формальной точки зрения каждый элемент выборки x_i считается случайной величиной, имеющей такое же распределение, что и генеральная совокупность *X*, и эти величины $x_1,\ x_2,...,x_n$ независимы. В частности $E(x_i) = m_X,\ V(x_i) = \sigma_X^2$ для каждого $i,\ i = 1,...,n$.

Выборочные характеристики. Основные выборочные характеристики − это

выборочное среднее $\overline{x} = \dfrac{x_1 + x_2 + ... + x_n}{n} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$

и

выборочная дисперсия $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

Подчеркнем, что выборочные характеристики при нашем подходе являются случайными величинами. Из свойств математического ожидания следует, что

$$E(\overline{x}) = m_X,\ V(\overline{x}) = \frac{\sigma_X^2}{n}.$$

Можно также доказать, что

$$E(s^2) = \sigma_X^2.$$

Из закона больших чисел следует, что $\overline{x} \to m_X,\ n \to \infty$, т.е. при большом числе наблюдений выборочное среднее приближается к среднему генеральной совокупности.

Центральная предельная теорема утверждает, что при больших *n* распределение выборочного среднего близко к нормальному с параметрами m_X, $\frac{\sigma_X^2}{n}$.

Частным случаем является соотношение между популяционной и выборочной пропорцией. Пусть в генеральной совокупности пропорция объектов, обладающих некоторым признаком (голосующих за кандидата демократической партии, поддерживающих закон об ограничении курения и т.п.) равна π. Предположим, что в выборке объема *n* указанным признаком обладает *m* объектов. Величина $p = \frac{m}{n}$ называется выборочной пропорцией. Выборочная пропорция является частным случаем выборочного среднего, $E(p) = \pi, V(p) = \frac{\pi(1-\pi)}{n}$.

Статистическое оценивание параметров

1. Точечное оценивание

Предполагается, что распределение генеральной совокупности зависит от некоторого параметра (или совокупности параметров) θ. Например, среднее значение m_X, популяционная пропорция π и т.п. Требуется по выборке $x_1, x_2, ..., x_n$ *оценить* параметр θ, т.е. построить функцию $\hat{\theta} = g(x_1, ..., x_n)$. Объектом изучения служит *метод оценивания* (по-английски, *estimator*), т.е. функция g, а не конкретное значение $g(x_1, ..., x_n)$ для конкретных значений наблюдений $x_1, x_2, ..., x_n$ (*estimate*). Функция $\hat{\theta} = g(x_1, ..., x_n)$ называется точечной оценкой параметра θ. Поскольку наблюдения $x_1, x_2, ..., x_n$ трактуются как случайные величины, то и оценка $\hat{\theta} = g(x_1, ..., x_n)$ также является случайной величиной.

Возникает естественная задача, какой же метод оценивания выбрать, Для этого обычно формулируют свойства, которым должен удовлетворять «хороший» метод оценивания.

Свойства оценок

1. Метод оценивания $\hat{\theta} = g(x_1, ..., x_n)$ называется несмещенным, если $E(\hat{\theta}) = \theta$. Иными словами, несмещенность содержательно означает, что многократное применение этого метода не дает систематической ошибки. Пример: 1) выборочное среднее $\bar{x}$ как оценка популяционного среднего m_X является несмещенной; 2) выборочная дисперсия s^2 является несмещенной оценкой популяционной дисперсии σ_X^2. В общем случае величина $bias \equiv E(\hat{\theta}) - \theta$ называется смещением оценки $\hat{\theta}$.

2. Метод оценивания $\hat{\theta} = g(x_1, ..., x_n)$ называется состоятельным, если $\hat{\theta} \to \theta$ при $n \to \infty$, т.е. с ростом числа наблюдений оценка, полученная с помощью этого метода, сходится к оцениваемому параметру θ. Пример: выборочное среднее $\bar{x}$ как оценка популяционного среднего m_X является состоятельной − это следует из закона больших чисел.

Если оценка $\hat{\theta}$ является несмещенной, то естественной мерой ее точности является ее дисперсия $V(\hat{\theta})$, т.е. из двух несмещенных оценок предпочтительнее та, у которой дисперсия меньше. В общем случае мерой точности оценки $\hat{\theta}$ служит среднеквадратичная ошибка $MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2$. Можно показать, что $MSE(\hat{\theta}) = V(\hat{\theta}) + bias^2$.

Таким образом, обычно ограничиваются классом состоятельных и, если возможно, несмещенных оценок, и в этом классе пытаются найти наиболее эффективную оценку. Пример: можно показать (и это не совсем простой математический результат), что

выборочное среднее $\overline{x}$ является наиболее эффективной оценкой теоретического (популяционного) среднего m_X *для нормальной генеральной совокупности X.*

2. Интервальное оценивание (доверительные интервалы)

Наличие точечной оценки $\hat{\theta}$ параметра θ не позволяет в общем случае «локализовать» положение этого параметра. Поэтому естественной является следующий вопрос: нельзя ли найти такой интервал $I=(a,b)$, зависящий от наблюдений $x_1, x_2,...,x_n$, который «накрывает» параметр θ с достаточно большой вероятностью.

Определение. Интервал $I=(a,b)=(a(x_1, x_2,...,x_n), b(x_1, x_2,...,x_n))$ называется доверительным интервалом для параметра θ с уровнем доверия $1-\alpha$ (или $100(1-\alpha)\%$), если

$$P(\theta \in I) \geq 1-\alpha .$$

Обычно рассматривают 90%-ные ($\alpha=0.1$), 95%-ные ($\alpha=0.05$), 99%-ные ($\alpha=0.01$) доверительные интервалы.

Не существует универсального метода построения доверительных интервалов, однако есть некоторые приёмы, позволяющие для находить такие интервалы для определённых классов задач.

Пример 1. Пусть генеральная совокупность является нормальной, $X \sim N(m,\sigma^2)$, и пусть $x_1, x_2,...,x_n$ – выборка из этой генеральной совокупности. Можно доказать (соответствующее утверждение называется леммой Фишера), что величина

$$t=\frac{(\overline{x}-m)\sqrt{n}}{s}$$

имеет распределение Стьюдента (t–распределение) с $n-1$ степенями свободы. Пусть задан уровень доверия $1-\alpha$, обозначим $t_{\alpha/2}(n-1)$ $100(\alpha/2)\%$-ную точку этого распределения. Тогда по определению получаем:

$$1-\alpha = P(|t|<t_{\alpha/2}(n-1)) = P\left(\frac{|(\overline{x}-m)\sqrt{n}|}{s}<t_{\alpha/2}(n-1)\right).$$

Разрешая внутреннее неравенство относительно *m*, получаем:

$$P\left(\overline{x}-t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}<m<\overline{x}+t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right)=1-\alpha .$$

Но это и значит, что интервал $\left(\overline{x}-t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}},\ \overline{x}+t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}\right)$ является доверительным интервалом для *m* с уровнем доверия $1-\alpha$. Часто используют обозначение

$$m=\overline{x}\pm t_{\alpha/2}(n-1)\frac{s}{\sqrt{n}}.$$