В изложении доцента БГТУ Сахарова В.Ю.

# Построение интервальных оценок

Полученные оценки математического ожидания $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ и дисперсии $s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2$ для нормально распределённой случайной величины автоматически являются оценками параметров этой случайной величины. Так происходит потому что в формуле плотности вероятности нормально распределённой случайной величины $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-a)^2}{2\sigma^2}}$, параметр *a* как раз и является математическим ожиданием, а параметр $\sigma^2$ равен дисперсии. Выписанные оценки принято называть «точечными». Обратим внимание, что в силу парадокса нулевой вероятности вероятность того, что значение непрерывной случайной величины, в частности точечной оценки, будет равно конкретному числу, в частности оцениваемому параметру, равно нулю. Понятно, что всё равно на практике за неимением других придётся использовать приближённые равенства $a \approx \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$ и $\sigma^2 \approx s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)^2$, но возникает вопрос о точности этих приближённых равенств. На языке теории вероятностей это означает, что для хотя бы каким-то образом выбранных $\varepsilon_1>0$ и $\varepsilon_2>0$ нужно определить вероятности: $P\left(\bar{x} - \varepsilon_1 < a < \bar{x} + \varepsilon_2\right)$ и $P\left(s_1^2 - \varepsilon_1 < \sigma^2 < s_1^2 + \varepsilon_2\right)$.

Для построения интервальной оценки математического ожидания нормально распределённой случайной величины предлагается использовать четвёртый пункт леммы Фишера (принимаемый здесь без доказательства): случайная величина, равная $S_{n-1} = \frac{\sqrt{n}\left(\bar{x} - a\right)}{s_1}$ имеет распределение Стьюдента с $(n-1)$ степенями свободы.

Найдём

$$P\left(\bar{x} - \varepsilon_1 < a < \bar{x} + \varepsilon_2\right) = P\left(-\varepsilon_1 < a - \bar{x} < \varepsilon_2\right) = P\left(\varepsilon_1 > \bar{x} - a > -\varepsilon_2\right) = P\left(-\varepsilon_2 < \bar{x} - a < \varepsilon_1\right) =$$

$$= P\left(-\frac{\varepsilon_2\sqrt{n}}{s_1} < \frac{\left(\bar{x} - a\right)\sqrt{n}}{s_1} < \frac{\varepsilon_1\sqrt{n}}{s_1}\right) = P\left(-\frac{\varepsilon_2\sqrt{n}}{s_1} < S_{n-1} < \frac{\varepsilon_1\sqrt{n}}{s_1}\right).$$

По выведенной формуле можно находить вероятность того, что интервал $\left(\bar{x} - \varepsilon_1; \bar{x} + \varepsilon_2\right)$ будет содержать истинное значение *a*. Такие вероятность и интервал называются «доверительными». Отметим, что на практике обычно стоит задача, обратная к только что решённой: требуется построить доверительный интервал, зная доверительную вероятность. То есть, выбирается достаточно большая вероятность (по мнению исследователя равная вероятности практически достоверного события) и требуется построить интервал практически гарантировано содержащий истинное значение *a*. Очевидно, что чем больше будет эта вероятность, тем длиннее окажется доверительный интервал. Обозначим доверительную вероятность числом $\beta$. Тогда $\beta = P\left(-\frac{\varepsilon_2\sqrt{n}}{s_1} < S_{n-1} < \frac{\varepsilon_1\sqrt{n}}{s_1}\right)$.

В программе Microsoft Office Excel встроена функция СТЬЮДРАСПОБР(вероятность; число степеней свободы), с помощью которой можно находить значения функции, обратной к функции распределения Стьюдента. Согласно тексту справки она возвращает такое значение $t(x,n)$, для которого верно равенство $P(|S_n|>t)=x$. Заметим, что при такой постановке $t>0$, $0<x<1$. Событие $(|S_n|>t)$ почти противоположно событию $(|S_n|<t)=(-t<S_n<t)$. Поэтому сумма их вероятностей равна единице $P(|S_n|>t)+P(-t<S_n<t)=1$. Откуда $P(-t<S_n<t)=1-P(|S_n|>t)$ или $P(|S_n|>t)=1-P(-t<S_n<t)$. Слово «почти» означает, что оставшееся возможное событие $(|S_n|=t)$ имеет нулевую вероятность в силу парадокса нулевой вероятности.

Получается, что с помощью Microsoft Office Excel удобно находить границы доверительного интервала, если предположить, что они являются противоположными по знаку. Для нас это означает, что удобно положить $\varepsilon_1=\varepsilon_2=\varepsilon$ и получать доверительный интервал симметричным. Итак,

$$\beta = P\left(-\frac{\varepsilon\sqrt{n}}{s_1} < S_{n-1} < \frac{\varepsilon\sqrt{n}}{s_1}\right)$$

$$1-\beta = 1-P\left(-\frac{\varepsilon\sqrt{n}}{s_1} < S_{n-1} < \frac{\varepsilon\sqrt{n}}{s_1}\right)$$

$$1-\beta = P\left(|S_{n-1}| > \frac{\varepsilon\sqrt{n}}{s_1}\right)$$

Теперь с помощью функции СТЬЮДРАСПОБР находим число

$$\frac{\varepsilon\sqrt{n}}{s_1} = t(1-\beta, n-1)$$

А теперь выражаем

$$\varepsilon = \frac{s_1}{\sqrt{n}} \cdot t(1-\beta, n-1)$$

Найден доверительный интервал $\left(\bar{x}-\varepsilon;\bar{x}+\varepsilon\right)=\left(\bar{x}-\frac{s_1}{\sqrt{n}}\cdot t(1-\beta,n-1);\bar{x}+\frac{s_1}{\sqrt{n}}\cdot t(1-\beta,n-1)\right)$, содержащий параметр $a$, с доверительной вероятностью β.

# Часть 2

Для построения интервальной оценки дисперсии нормально распределённой случайной величины необходимо использовать третий пункт леммы Фишера (принимаемый здесь без доказательства): случайная величина, равная $\chi^2_{n-1} = \frac{(n-1)s_1^2}{\sigma^2}$ имеет распределение хи-квадрат (Пирсона) с $(n-1)$ степенями свободы.

Приравняем выбранную доверительную вероятность β вероятности того, что интервал $(s_1^2-\varepsilon_1; s_1^2+\varepsilon_2)$ будет содержать дисперсию $\sigma^2$. То есть $\beta = P\left(s_1^2-\varepsilon_1 < \sigma^2 < s_1^2+\varepsilon_2\right)$. Распределение хи-квадрат в отличие от распределения Стьюдента не обладает свойствами симметрии и поэтому арифметически невозможно построить доверительный интервал, симметричный относительно точечной оценки $s_1^2$. Чтобы преодолеть создавшееся затруднение, числа $\varepsilon_1$ и $\varepsilon_2$ подбирают так, чтобы вероятности попадания дисперсии $\sigma^2$ в оставшиеся «хвосты», то есть промежутки $(-\infty; s_1^2-\varepsilon_1)$ и $(s_1^2+\varepsilon_2;+\infty)$ были бы одинаковыми: $P(\sigma^2 \le s_1^2-\varepsilon_1) = P(\sigma^2 \ge s_1^2+\varepsilon_2)$. Отметим, что $\beta = P(s_1^2-\varepsilon_1 < \sigma^2 < s_1^2+\varepsilon_2) = 1-P(\sigma^2 \le s_1^2-\varepsilon_1) - P(\sigma^2 \ge s_1^2+\varepsilon_2)$. Решив эту систему двух

линейных уравнений с двумя неизвестными, получим $P(\sigma^2 \le s_1^2 - \varepsilon_1) = P(\sigma^2 \ge s_1^2 + \varepsilon_2) = \frac{1-\beta}{2}$. Преобразуем по очереди эти два выражения для использования функции, встроенной в Microsoft Office Excel:

$$P(\sigma^2 \le s_1^2 - \varepsilon_1) = \frac{1-\beta}{2}$$

$$P\left(\frac{\sigma^2}{s_1^2} \le \frac{s_1^2 - \varepsilon_1}{s_1^2}\right) = \frac{1-\beta}{2}$$

$$P\left(\frac{s_1^2}{\sigma^2} \ge \frac{s_1^2}{s_1^2 - \varepsilon_1}\right) = \frac{1-\beta}{2}$$

$$P\left(\frac{(n-1)s_1^2}{\sigma^2} \ge \frac{(n-1)s_1^2}{s_1^2 - \varepsilon_1}\right) = \frac{1-\beta}{2}$$

$$P\left(\chi_{n-1}^2 \ge \frac{(n-1)s_1^2}{s_1^2 - \varepsilon_1}\right) = \frac{1-\beta}{2}$$

Согласно справке Microsoft Office Excel функция ХИ2ОБР(вероятность;число степеней свободы), возвращает такое число $H(\alpha,n)$, для которого верно равенство $P\left(\chi_n^2 > H\right) = \alpha$. Разумеется строгость или нестрогость неравенства не имеет никакого значения в силу парадокса нулевой вероятности.

$$\frac{(n-1)s_1^2}{s_1^2 - \varepsilon_1} = H\left(\frac{1-\beta}{2}, n-1\right)$$

$$s_1^2 - \varepsilon_1 = \frac{(n-1)s_1^2}{H\left(\frac{1-\beta}{2}, n-1\right)}$$

Найдена левая граница доверительного интервала. Не многим сложнее находится правая граница.

$$P(\sigma^2 \ge s_1^2 + \varepsilon_2) = \frac{1-\beta}{2}$$

$$P\left(\frac{\sigma^2}{s_1^2} \ge \frac{s_1^2 + \varepsilon_2}{s_1^2}\right) = \frac{1-\beta}{2}$$

$$P\left(\frac{s_1^2}{\sigma^2} \le \frac{s_1^2}{s_1^2 + \varepsilon_2}\right) = \frac{1-\beta}{2}$$

$$P\left(\frac{(n-1)s_1^2}{\sigma^2} \le \frac{(n-1)s_1^2}{s_1^2 + \varepsilon_2}\right) = \frac{1-\beta}{2}$$

$$P\left(\chi_{n-1}^2 \le \frac{(n-1)s_1^2}{s_1^2 + \varepsilon_2}\right) = \frac{1-\beta}{2}$$

Для использования функции ХИ2ОБР неравенство под знаком вероятности должно быть в другую сторону. Поэтому найдём вероятность противоположного события.

$$P\left(\chi_{n-1}^2 > \frac{(n-1)s_1^2}{s_1^2 + \varepsilon_2}\right) = 1 - \frac{1-\beta}{2}$$

$$P\left(\chi_{n-1}^2 > \frac{(n-1)s_1^2}{s_1^2 + \varepsilon_2}\right) = \frac{2-1+\beta}{2}$$

$$P\left(\chi^2_{n-1} > \frac{(n-1)s_1^2}{s_1^2+\varepsilon_2}\right) = \frac{1+\beta}{2}$$

Теперь с помощью Microsoft Office Excel вычисляем

$$\frac{(n-1)s_1^2}{s_1^2+\varepsilon_2} = H\left(\frac{1+\beta}{2}, n-1\right)$$

$$s_1^2+\varepsilon_2 = \frac{(n-1)s_1^2}{H\left(\frac{1+\beta}{2}, n-1\right)}$$

Найден доверительный интервал $\left(s_1^2-\varepsilon_1; s_1^2+\varepsilon_2\right) = \left(\frac{(n-1)s_1^2}{H\left(\frac{1-\beta}{2}, n-1\right)}; \frac{(n-1)s_1^2}{H\left(\frac{1+\beta}{2}, n-1\right)}\right)$,

содержащий параметр $\sigma^2$, с доверительной вероятностью $\beta$.