Материал взят из пособия Миссаров М.Д. Введение в теорию вероятностей: Учебное пособие/ М.Д. Миссаров - Казань: Изд-во Казан.ун-та,2019-126с

3.4. Совместные функции распределения нескольких случайных величин.

В большом количестве приложений теории вероятностей возникают задачи, в которых мы наблюдаем одновременно сразу несколько различных случайных величин. В дальнейшем мы в основном будем рассматривать случай, когда имеются две случайные величины, заданные на одном и том же вероятностном пространстве $(\Omega, \mathcal{A}, P)$.

Пусть, например, случайная величина ξ принимает значения $a_1, \ldots, a_n$, а случайная величина η принимает значения $b_1, \ldots, b_m$. Нас может заинтересовать вопрос о том, с какой вероятностью $\xi = a_i, \quad \eta = b_j$. Пусть

$$p_{ij} = P\left(\left\{\omega \in \Omega : \xi(\omega) = a_i, \eta(\omega) = b_j\right\}\right). \qquad (3.23)$$

Здесь выражение $\left\{\omega \in \Omega : \xi(\omega) = a_i, \eta(\omega) = b_j\right\}$ означает пересечение событий $\{\omega: \xi(\omega) = a_i\}$ и $\left\{\omega : \eta(\omega) = b_j\right\}$. В дальнейшем для упрощения записи мы будем обозначать пересечение этих событий как $\left(\xi = a_i, \eta = b_j\right)$, $p_{ij} = P\left(\xi = a_i, \eta = b_j\right)$.

Определение 3.14. Набор вероятностей

$$\left\{p_{ij}, \ i = 1, \ldots, n, \ \ j = 1, \ldots, m\right\}$$

называется совместным законом распределения вероятностей случайных величин ξ и η.

Эти вероятности можно записать в прямоугольную таблицу размера $n \times m$, в которой строки нумеруются индексом i , а столбцы – индексом j , так, что ячейке с номерами i и j приписано число p_{ij} . Такая таблица называется таблицей совместного распределения вероятностей случайных величин ξ и η .

Заметим, что

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij} = 1 ,$$

так как события $\left(\xi = a_i, \eta = b_j\right)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ образуют полную группу событий.

Зная совместный закон распределения вероятностей ξ и η, легко вычислить одномерные распределения вероятностей случайных величин ξ и η. Например,

$$\sum_{j=1}^{m} p_{ij} = P(\xi = a_i), \qquad i = 1, \ldots, n . \qquad (3.24)$$

Действительно, пусть $A_i = \{\xi = a_i\}$, $i = 1, \ldots, n$, $B_j = \left\{\eta = b_j\right\}$, $j = 1, \ldots, m$. Группа событий A_i , $i = 1, \ldots, n$, так же как и группа B_j , $j = 1, \ldots, m$ образуют полные группы событий:

$$\sum_{j=1}^{m} p_{ij} = \sum_{j=1}^{m} P(A_i B_j) = P\left(\bigcup_{j=1}^{m} A_i B_j\right) = P\left(A_i \left(\bigcup_{j=1}^{m} B_j\right)\right) =$$

$$= P(A_i \Omega) = P(A_i) = P(\xi = a_i) .$$

Таким же образом доказывается, что

$$\sum_{i=1}^{n} p_{ij} = P\left(\eta = b_j\right), \qquad j = 1, \ldots, m \qquad (3.25)$$

Предположим теперь, что у нас имеется n произвольных случайных величин $\xi_1, \ldots, \xi_n$, определенных на одном и том же вероятностном пространстве.

Определение 3.15. Совместной функцией распределения случайных величин $\xi_1, \ldots, \xi_n$ называется функция $F_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n)$, которая определяется формулой:

$$F_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n) = P(\xi_1 < x_1, \ldots, \xi_n < x_n)\,. \qquad (3.26)$$

Здесь $x_1, \ldots, x_n$ – аргументы совместной функции распределения, которые мы можем объединить в один вектор $x = (x_1, \ldots, x_n)$, $x \in \mathbb{R}^n$. Событие $(\xi_1 < x_1, \ldots, \xi_n < x_n)$ в формуле (3.26) понимается как пересечение событий $\{\omega: \xi_1(\omega) < x_1\}, \ldots, \{\omega: \xi_n(\omega) < x_n\}$.

Функция $F_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n)$ является обобщением одномерной функции распределения на случай n случайных величин. В дальнейшем мы будем опускать обозначение индексов $\xi_1, \ldots, \xi_n$ в записи функции $F_{\xi_1,\ldots,\xi_n}(x_1, \ldots, x_n)$ в том случае, когда это не вызывает недоразумений.

Приведем свойства совместной функции распределения:

1) $0 \le F(x_1, \ldots, x_n) \le 1$ для любых значений $x_1, \ldots, x_n$.
2) Монотонность по любой переменной: если $x_i > x'_i$, то
$$F(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) \ge F(x_1, \ldots, x_{i-1}, x'_i, x_{i+1}, \ldots, x_n).$$
3) Для любой переменной x_i, $i = 1, \ldots, n$
$$\lim_{x_i \to -\infty} F(x_1, \ldots, x_i, \ldots, x_n) \to 0\,.$$

Это свойство следует из свойства непрерывности вероятности и того факта, что событие $\{\xi_i < -\infty\}$ является невозможным.

4) Для любой переменной x_i, $i = 1, \ldots, n$
$$\lim_{x_i \to +\infty} F(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) = F_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n),$$

где

$$F_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) =$$
$$= P(\xi_1 < x_1, \ldots, \xi_{i-1} < x_{i-1}, \xi_{i+1} < x_{i+1}, \ldots, \xi_n < x_n).$$

Это свойство также следует из свойства непрерывности вероятности и того, что событие $\{\xi_i < \infty\}$ является достоверным. Свойство 4 показывает, что в предельном переходе $x_i \to +\infty$ мы получаем совместную функцию распределения для величин $\xi_1, \ldots, \xi_{i-1}, \xi_{i+1}, \ldots, \xi_n$ (величина ξ_i удалена из общего набора случайных величин).

Зная совместную функцию распределения, мы можем вычислить вероятности $P((\xi_1, \ldots, \xi_n) \in A)$, где A – множество в n-мерном пространстве $\mathbb{R}^n$. Докажем это утверждение в двумерном случае для множества $A = [a_1, b_1) \times [a_2, b_2) \subset \mathbb{R}^2$.

Теорема 3.2. Пусть совместное распределение случайных величин ξ_1 и ξ_2 описывается двухмерной функцией распределения $F(x_1, x_2)$. Пусть $a_1 < b_1$, $a_2 < b_2$. Тогда

$$P(a_1 \le \xi_1 < b_1,\, a_2 \le \xi_2 < b_2) =$$
$$= F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2). \qquad (3.27)$$

Доказательство. Пусть

$$A_1 = \{\omega: \xi_1(\omega) < a_1\},\ B_1 = \{\omega: \xi_1(\omega) < b_1\},$$
$$A_2 = \{\omega: \xi_2(\omega) < a_2\},\ B_2 = \{\omega: \xi_2(\omega) < b_2\}.$$

Заметим, что $A_1 \subset B_1$, $A_2 \subset B_2$. Тогда

$$P(a_1 \le \xi_1 < b_1,\, a_2 \le \xi_2 < b_2) =$$
$$= P(B_1 B_2 (A_1 \cup A_2)) = P(B_1 B_2) - P(B_1 B_2 (A_1 \cup A_2))\,.$$

В свою очередь

$$P(B_1 B_2 (A_1 \cup A_2)) = P(B_1 B_2 A_1) + P(B_1 B_2 A_2) - P(B_1 B_2 A_1 A_2) =$$
$$= P(A_1 B_2) + P(B_1 A_2) - P(A_1 A_2)\,.$$

Таким образом,

$$P(a_1 \le \xi_1 < b_1,\, a_2 \le \xi_2 < b_2) =$$
$$= P(B_1 B_2) - P(A_1 B_2) - P(B_1 A_2) + P(A_1 A_2) =$$
$$= F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)\,,$$

что и требовалось доказать.

Можно показать, что из счетного множества прямоугольников вида $[a_1, b_1) \times [a_2, b_2) \subset \mathbb{R}^2$ можно конструировать произвольные измеримые множества в $\mathbb{R}^2$.

Понятие плотности распределения вероятностей также обобщается на n -мерный случай.

Определение 3.16. Пусть справедливо представление совместной функции распределения в виде:

$$F(x_1, \dots, x_n) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_n} p(y_1, \dots, y_n)\, dy_1 \dots dy_n \qquad (3.28)$$

для некоторой функции $p(y_1, \dots, y_n)$. В этом случае функцию $p(y_1, \dots, y_n)$ называют совместной плотностью распределения случайных величин $\xi_1, \dots, \xi_n$.

Справедливы следующие свойства совместной плотности распределения:

1. Условие нормировки:

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} p(x_1, \dots, x_n)\, dx_1 \dots dx_n = 1 .$$

2.

$$p(x_1, \dots, x_n) = \frac{\partial}{\partial x_1} \dots \frac{\partial}{\partial x_n} F(x_1, \dots, x_n) ,$$

в тех точках $(x_1, \dots, x_n)$, в которых функция $F(x_1, \dots, x_n)$ дифференцируема.

3. Для произвольного $i \in \{1, \dots, n\}$ функция

$$p_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = \int_{-\infty}^{\infty} p(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n)\, dx_i$$

является совместной плотностью распределения случайных величин $\xi_1, \dots, \xi_{i-1}, \xi_{i+1}, \dots, \xi_n$.

Свойства 1 и 2 легко следуют из определений. Поясним свойство 3. Пусть $F_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)$ – совместная функция распределения случайных величин $\xi_1, \dots, \xi_{i-1}, \xi_{i+1}, \dots, \xi_n$ (величина ξ_i отсутствует в этом списке). Тогда из свойства 4 совместной функции распределения и формулы (3.28) следует, что

$$F_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) =$$
$$= \int_{-\infty}^{x_1} dy_1 \dots \int_{-\infty}^{x_{i-1}} dy_{i-1} \int_{-\infty}^{\infty} dy_i \int_{-\infty}^{x_{i+1}} dy_{i+1} \dots \int_{-\infty}^{x_n} p(y_1, \dots, y_n)\, dy_i. \quad (3.29)$$

Предполагая, что все операции дифференцирования корректны, получаем

$$p_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) =$$
$$= \frac{\partial}{\partial x_1} \dots \frac{\partial}{\partial x_{i-1}} \frac{\partial}{\partial x_{i+1}} \dots \frac{\partial}{\partial x_n} F_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) =$$
$$= \int_{-\infty}^{\infty} p(x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_n)\, dy_i .$$

4.4.Ковариация и коэффициент корреляции случайных величин.

Пусть даны две случайные величины ξ и η. Заметим, что математические ожидания и дисперсии $E\,\xi,\ D\,\xi$ и $E\,\eta,\ D\,\eta$ характеризуют поведение этих величин по отдельности, но не несут информации о том, как эти величины связаны между собой.

Определение 4.5. Пусть ξ и η – случайные величины, $E\,\xi < \infty$, $E\,\eta < \infty$. Ковариацией величин ξ и η называется число

$$Cov\,(\xi, \eta) = E\,(\xi - E\,\xi)(\eta - E\,\eta) \qquad (4.14)$$

(при условии, что математическое ожидание в 4.14 существует).

Свойства ковариации:

Свойство 1.

$$Cov\,(\xi,\eta) = E\,\xi\eta - E\,\xi\,E\,\eta\,. \qquad (4.15)$$

Действительно,

$$Cov\,(\xi,\eta) = E\,(\xi - E\,\xi)(\eta - E\,\eta) =$$
$$= E\,[\xi\eta - \xi(E\,\eta) - \eta(E\,\xi) + E\,\xi\,E\,\eta] =$$
$$= E\,\xi\eta - E\,\xi\,E\,\eta - E\,\xi\,E\,\eta + E\,\xi\,E\,\eta = E\,\xi\eta - E\,\xi\,E\,\eta$$

(здесь используется то, что $E\,\xi$ и $E\,\eta$ являются константами).

Свойство 2. Линейность операции ковариации:

$$Cov\,(\alpha\xi + \beta\eta,\zeta) = \alpha\,Cov\,(\xi,\zeta) + \beta\,Cov\,(\eta,\zeta)\,.$$

Доказательство следует из свойства линейности операции математического ожидания.

Свойство 3. Симметричность:

$$Cov\,(\xi,\eta) = Cov\,(\eta,\xi)\,.$$

Свойство 4.

Если ξ и η – независимые случайные величины, то

$$Cov\,(\xi,\eta) = 0\,.$$

Так как для независимых случайных величин $E\,\xi\eta = E\,\xi\,E\,\eta$, то утверждение следует из (4.15).

Свойство 5.

$$Cov\,(\xi,\xi) = D\,\xi\,.$$

Теорема 4.2. Пусть $\xi_1,\xi_2,\dots,\xi_n$ – случайные величины с конечными дисперсиями $D\,\xi_i < \infty$, $i = 1,\dots,n$. Тогда

$$D\left(\sum_{i=1}^{n}\xi_i\right) = \sum_{i=1}^{n} D\,\xi_i + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} Cov\,(\xi_i,\xi_j). \qquad (4.16)$$

Доказательство. Пользуясь свойствами линейности, симметрии и свойством 5 ковариации, мы можем написать

$$D\left(\sum_{i=1}^{n}\xi_i\right) = Cov\left(\sum_{i=1}^{n}\xi_i\,,\sum_{j=1}^{n}\xi_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov\,(\xi_i,\xi_j) =$$
$$= \sum_{i=1}^{n} Cov\,(\xi_i,\xi_i) + \sum_{i=1}^{n}\sum_{j\ne i} Cov\,(\xi_i,\xi_j) =$$
$$= \sum_{i=1}^{n} D\,\xi_i + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n} Cov\,(\xi_i,\xi_j)\,.$$

Случайные величины ξ и η называются некоррелированными, если $Cov\,(\xi,\eta) = 0$. Если величины $\xi_1,\xi_2,\dots,\xi_n$ являются попарно некоррелированными, то

$$D\left(\sum_{i=1}^{n}\xi_i\right) = \sum_{i=1}^{n} D\,\xi_i\,.$$

Определение 4.6. Пусть ξ и η – случайные величины, $E\,\xi < \infty$, $E\,\eta < \infty$, $D\,\xi < \infty$, $D\,\eta < \infty$. Коэффициентом корреляции величин ξ и η называется число

$$\rho(\xi,\eta) = \frac{Cov\,(\xi,\eta)}{\sqrt{D\,\xi\,D\,\eta}}\,. \qquad (4.17)$$

В дальнейшем нам понадобится следующее утверждение:

Лемма. (Неравенство Коши-Буняковского). Пусть ξ и η – случайные величины такие, что $E\,\xi^2 < \infty$, $E\,\eta^2 < \infty$. Тогда $E\,\xi\eta$ существует и

$$|E\,\xi\eta| \le (E\,\xi^2\,E\,\eta^2)^{1/2}\,. \qquad (4.18)$$

Равенство в формуле (4.18) достигается тогда и только тогда, когда существуют константы α и β (не равные нулю одновременно) такие, что $\alpha\xi + \beta\eta = 0$.

Доказательство. Если $E\,\xi^2 = 0$, то $P(\xi = 0) = 1$. В этом случае $P(\xi\eta = 0) = 1$. В этом случае, $E\,\xi\eta = 0$ и неравенство (4.18) выполнено. То же верно и в случае, когда $E\,\eta^2 = 0$.

Пусть $0 < E\,\xi^2 < \infty$, $0 < E\,\eta^2 < \infty$. Для любых α и β $E(\alpha\xi \pm \beta\eta)^2 \geq 0$. Отсюда

$$E(\alpha\xi \pm \beta\eta)^2 = \alpha^2 E\,\xi^2 + \beta^2 E\,\eta^2 \pm 2\alpha\beta\, E\,\xi\eta \geq 0 \,. \quad (4.19)$$

Пусть $\alpha = (E\,\xi^2)^{1/2}$, $\beta = (E\,\eta^2)^{1/2}$. Тогда из (4.19) следует, что

$$-(E\,\xi^2\, E\,\eta^2)^{1/2} \leq E\,\xi\eta \leq (E\,\xi^2\, E\,\eta^2)^{1/2} \,,$$

т.е.

$$|E\,\xi\eta| \leq (E\,\xi^2\, E\,\eta^2)^{1/2} \,.$$

Свойства коэффициента корреляции:

Свойство 1. Пусть ξ и η – произвольные случайные величины, $\xi' = \alpha_0 + \alpha_1\xi$, $\eta' = \beta_0 + \beta_1\eta$ – линейные преобразования этих величин. Тогда

$$\rho(\xi',\eta') = sign\,(\alpha_1\beta_1)\rho(\xi,\eta)\,.$$

Здесь $sign\,x$ – знак числа x .

Действительно,

$$Cov\,(\xi',\eta') = Cov\,(\alpha_0 + \alpha_1\xi, \beta_0 + \beta_1\eta) = \alpha_1\beta_1 Cov\,(\xi,\eta)\,,$$

так как ковариация любой случайной величины с константой равна 0. Далее,

$$D\,\xi' = D(\alpha_0 + \alpha_1\xi) = \alpha_1^2\, D\,\xi\,, \qquad D\,\eta' = D(\beta_0 + \beta_1\eta) = \beta_1^2\, D\,\eta\,.$$

Следовательно,

$$\rho(\xi',\eta') = \frac{Cov\,(\xi',\eta')}{\sqrt{D\,\xi' D\,\eta'}} = \frac{\alpha_1\beta_1 Cov\,(\xi,\eta)}{|\alpha_1\beta_1|\sqrt{D\,\xi\, D\,\eta}} = sign\,(\alpha_1\beta_1)\rho(\xi,\eta)\,.$$

Свойство 2. $|\rho(\xi,\eta)| \leq 1$,
$|\rho(\xi,\eta)| = 1$ тогда и только тогда, когда величина η является линейным преобразованием величины ξ : $\eta = \alpha\xi + \beta$. Если $\rho(\xi,\eta) = +1$, то $\alpha > 0$, если $\rho(\xi,\eta) = -1$, то $\alpha < 0$.

Доказательство. Коэффициент корреляции величин ξ и η определен только в том случае, когда $D\,\xi \neq 0$ и $D\,\eta \neq 0$. Пусть

$$\xi_1 = \frac{\xi}{\sqrt{D\,\xi}}\,, \quad \eta_1 = \frac{\eta}{\sqrt{D\,\eta}}\,.$$

Тогда

$$D(\xi_1 \pm \eta_1) = D\,\xi_1 + D\,\eta_1 \pm 2Cov\,(\xi_1,\eta_1) =$$
$$= \frac{D\,\xi}{D\,\xi} + \frac{D\,\eta}{D\,\eta} \pm 2\frac{Cov\,(\xi,\eta)}{\sqrt{D\,\xi\, D\,\eta}} = 2\big(1 \pm \rho(\xi,\eta)\big)\,.$$

Так как дисперсия любой величины неотрицательна, то

$$-1 \leq \rho(\xi,\eta) \leq 1\,.$$

Если $\rho(\xi,\eta) = \pm 1$, то $D(\xi_1 \mp \eta_1) = 0$. Следовательно, $\xi_1 \mp \eta_1 = E\,\xi_1 \mp E\,\eta_1$. Отсюда

$$\eta_1 = \pm\xi_1 \mp E\,\xi_1 + E\,\eta_1$$

или

$$\eta = \pm\frac{\sqrt{D\,\eta}}{\sqrt{D\,\xi}}\,\xi + E\,\eta \mp \frac{\sqrt{D\,\eta}}{\sqrt{D\,\xi}}\,E\,\xi\,.$$

Таким образом, $\eta = \alpha\xi + \beta$, где $\alpha = \pm\frac{\sqrt{D\,\eta}}{\sqrt{D\,\xi}}$, $\beta = E\,\eta \mp \frac{\sqrt{D\,\eta}}{\sqrt{D\,\xi}}\,E\,\xi$.

Обратно, пусть $\eta = \alpha\xi + \beta$. Тогда по свойству 1 коэффициента корреляции

$$\rho(\xi,\eta) = sign\,\alpha \cdot \rho(\xi,\xi) = sign\,\alpha\,,$$

так как $\rho(\xi,\xi) = 1$. Заметим, что если $\rho(\xi,\eta) = -1$, то $\alpha < 0$, если же $\rho(\xi,\eta) = +1$, то $\alpha > 0$.

Свойство 3. Если ξ и η – независимые случайные величины, то $\rho(\xi,\eta) = 0$.

Доказательство следует из того, что для независимых величин ξ и η $Cov\,(\xi,\eta)=0$ (свойство 4 ковариации). Обратное, вообще говоря, неверно.

Пример. Пусть величина ξ принимает три значения $-1,\ 0,\ 1$ с одинаковыми вероятностями и пусть $\eta=\xi^2$. Тогда $P(\eta=0)=1/3$, $P(\eta=1)=2/3$. Легко видеть, что случайные величины ξ и η зависимы между собой. Например,

$$P(\xi=0,\eta=1)=0\,,\quad P(\xi=0)P(\eta=1)=\frac{2}{9}\,,$$

т.е. $P(\xi=0,\eta=1)\neq P(\xi=0)P(\eta=1)$. Так как $E\,\xi\eta=E\,\xi^3=E\,\xi=0$ и $E\,\xi\,E\,\eta=0$, то $Cov\,(\xi,\eta)=\ E\,\xi\eta-E\,\xi\,E\,\eta=0$, т.е. величины ξ и η не коррелированы между собой.

5.3. Центральная предельная теорема. Задача о расчете страховой премии.

Пусть $\xi_1,\xi_2,\ldots$ – последовательность независимых одинаково распределенных случайных величин, имеющих конечное математическое ожидание и дисперсию: $E\,\xi_i=m<\infty\,,\quad D\,\xi_i=\sigma^2<\infty\,,\quad i=1,2,\ldots$. Рассмотрим сумму этих величин

$$\eta_n=\xi_1+\xi_2+\ \ldots+\ \xi_n\,.$$

Так как

$$E\,\eta_n=E(\xi_1+\xi_2+\ \ldots+\ \xi_n)=nm\to\infty\quad \text{при }\ n\to\infty\,,$$
$$D\,\eta_n=D(\xi_1+\xi_2+\ \ldots+\ \xi_n)=n\sigma^2\to\infty\quad \text{при }\ n\to\infty\,,$$

то говорить о существовании предела у последовательности величин η_n , $n=1,2,\ldots$ не приходится. Рассмотрим линейное преобразование величины η_n:

$$\zeta_n=\frac{\eta_n-E\,\eta_n}{\sqrt{D\,\eta_n}}=\frac{\eta_n-nm}{\sqrt{n}\sigma}\,. \qquad (5.16)$$

Числовые характеристики величин $\zeta_n,\ n=1,2,\ldots$ являются фиксированными числами:

$$E\,\zeta_n=\frac{1}{\sqrt{D\,\eta_n}}E(\eta_n-E\,\eta_n)=0\,,$$
$$D\,\zeta_n=\frac{1}{D\,\eta_n}D(\eta_n-E\,\eta_n)=\frac{D\,\eta_n}{D\,\eta_n}=1\,.$$

Преобразование (5.16) называется преобразованием стандартизации, поскольку оно «масштабирует» величину η_n в новой «системе координат», в которой среднее значение и дисперсия принимают «стандартные» значения. Удивительным фактом является то, что стандартизованные величины ζ_n имеют предел при $n\to\infty$, и этот предел не зависит от распределения величин $\xi_1,\xi_2,\ldots$.

Теорема 5.5. (Центральная предельная теорема). Пусть $\xi_1,\xi_2,\ldots$ – последовательность независимых одинаково распределенных случайных величин, $E\,\xi_i=m<\infty\,,\quad D\,\xi_i=\sigma^2<\infty\,,\quad i=1,2,\ldots$. Пусть $F_n(x)$ – функции распределения случайных величин ζ_n , $F_n(x)=\mathrm{P}(\zeta_n<x)$,

$$\zeta_n=\frac{\xi_1+\ \ldots+\ \xi_n-nm}{\sqrt{n}\sigma}\ ,\quad n=1,2,\ldots\,.$$

Тогда

$$F_n(x)\to\Phi(x)=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{x}exp\left\{-\frac{y^2}{2}\right\}dy\quad \text{при }\ n\to\infty\,. \qquad (5.17)$$

Другими словами, последовательность случайных величин сходится к случайной величине $\zeta\sim N(0,1)$ по распределению: $\zeta_n\xrightarrow{F}\zeta$ при $n\to\infty$.

Мы не приводим доказательство этой важной теоремы, поскольку оно является достаточно сложным для нашего курса. Эта теорема позволяет использовать так называемое «гауссовское приближение» в задачах, в которых нужно оценивать распределение сумм большого количества случайных величин. Статистики советуют использовать такое приближение, когда количество слагаемых достаточно велико (например, больше 40).

Покажем, как гауссовское приближение может быть использовано в задаче вычисления страховой премии.

Пусть в портфеле страховой компании находится n однотипных договоров. Предположим, что страховые случаи происходят независимо друг от друга и вероятность того, что страховой случай произойдет по любому договору, равна p. Пусть η_i обозначает случайную величину – индикатор того, что по i -му договору произойдет страховой случай: $P(\eta_i = 1) = p$,
$P(\eta_i = 0) = q = 1 - p$. Величина η_i имеет распределение Бернулли:

$$E\,\eta_i = 0 \cdot q + 1 \cdot p = p\,,\;\; D\,\eta_i = E\,\eta_i{}^2 - (E\,\eta_i)^2 = p - p^2 = pq\,.$$

Обозначим возможный ущерб по i -му договору как ξ_i ,
$i = 1, \dots, n$. Мы предполагаем, что величины ξ_i, $i = 1, \dots, n$ являются независимыми, одинаково распределенными случайными величинами, $E\,\xi_i = m$, $D\,\xi_i = \sigma^2$. Обозначим страховую выплату по i -му договору как ζ_i . Легко видеть, что
$\zeta_i = \eta_i \xi_i$, $i = 1,2, \dots, n$. Действительно, если страхового случая не будет, то $\eta_i = 0$ и компании не придется производить выплату. Если же страховой случай произойдет, то $\eta_i = 1$, и компании придется возместить ущерб в полном объеме: $\zeta_i = \xi_i$.

Мы будем предполагать независимость всех величин η_i , ξ_i , $i = 1,2, \dots, n$ между собой. Такое предположение является естественным, так как сам факт наступления страхового случая не связан с размером ущерба.

Вычислим математические ожидания и дисперсии величин ζ_i , $i = 1, \dots, n$. В силу независимости случайных величин ξ_i и η_i , $i = 1, \dots, n$

$$E\,\zeta_i = E\,\xi_i \eta_i = E\,\xi_i E\,\eta_i = mp\,, \qquad (5.18)$$

$$\begin{aligned} D\,\zeta_i &= E\,\zeta_i{}^2 - (E\,\zeta_i)^2 = E\,\xi_i{}^2 \eta_i{}^2 - (E\,\xi_i \eta_i)^2 = \\ &= E\,\xi_i{}^2 E\,\eta_i{}^2 - (E\,\xi_i E\,\eta_i)^2 = \\ &= (m^2 + \sigma^2) \cdot p - (mp)^2 = p(m^2 q + \sigma^2)\,. \end{aligned} \qquad (5.19)$$

Общий объем выплат страховой компании обозначим S_n :

$$S_n = \zeta_1 + \dots + \zeta_n\,.$$

Пользуясь формулами (5.18) и (5.19), мы можем вычислить среднее значение и дисперсию величины S_n :

$$E\,S_n = nmp\,,\quad D\,S_n = np(m^2 q + \sigma^2). \qquad (5.20)$$

Вычисление распределения вероятностей величины S_n в общем случае является весьма сложной задачей. Поскольку S_n является суммой независимых одинаково распределенных случайных величин, воспользуемся центральной предельной теоремой. Можно утверждать, что при достаточно больших n стандартизованная величина S_n приближенно имеет нормальное распределение:

$$P\left(\frac{S_n - E\,S_n}{\sqrt{D\,S_n}} < x\right) \approx \Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} exp\left\{-\frac{y^2}{2}\right\} dy\,.$$

Зададимся вопросом: сколько денег надо брать с клиентов, чтобы собранных денег хватило на покрытие суммарного ущерба? Среднее значение суммарного ущерба равно $E\,S_n = nmp$, и, если мы возьмем с каждого владельца страхового полиса страховую премию $P_0 = {}^{E\,S_n}\!/_{n} = mp$, то собранная сумма будет равна $E\,S_n = nmp$. Какова вероятность того, что собранная сумма покроет суммарный ущерб?

$$P(S_n < E\,S_n) = P(S_n - E\,S_n < 0) = P\left(\frac{S_n - E\,S_n}{\sqrt{D\,S_n}} < 0\right) \approx$$

$$\approx \Phi(0) = 0.5\,.$$

Таким образом, вероятность того, что собранная сумма покроет суммарный ущерб, равна всего лишь 0.5.

Какой должна быть величина K для того, чтобы она покрыла суммарный ущерб с вероятностью, большей, чем заданное число γ (например, $\gamma = 0{,}9$ или $\gamma = 0{,}99$) ?

$$P(S_n < K) = P(S_n - E\,S_n < K - E\,S_n) =$$
$$= P\left(\frac{S_n - E\,S_n}{\sqrt{D\,S_n}} < \frac{K - E\,S_n}{\sqrt{D\,S_n}}\right) \approx \Phi\left(\frac{K - E\,S_n}{\sqrt{D\,S_n}}\right) = \gamma\,. \qquad (5.21)$$

Решение уравнения $F(x) = \gamma$ для некоторого распределения F называется γ-квантилем распределения F . В нашем случае $\Phi(x)$ – функция распределения стандартной гауссовской величины. Обозначим γ -квантиль распределения $\Phi(x)$ как $u(\gamma)$. Например, $u(0{,}9) = 1.282$, $u(0{,}99) = 2.326$.

Из уравнения (5.21) следует, что

$$\frac{K - E\,S_n}{\sqrt{D\,S_n}} = u(\gamma)\,,$$

откуда

$$K = ES_n + u(\gamma)\sqrt{D\,S_n}\ . \qquad (5.22)$$

Таким образом, для того, чтобы с вероятностью γ выполнить обязательства, страховая компания должна собрать сумму, заданную формулой (5.22). Так как все договоры однотипные, то страховые премии должны быть одинаковыми:

$$P = \frac{K}{n} = \frac{E\,S_n}{n} + u(\gamma)\frac{\sqrt{D\,S_n}}{n}\ .$$

Обозначим

$$P_0 = \frac{E\,S_n}{n}\,, \quad P_r = u(\gamma)\frac{\sqrt{D\,S_n}}{n}\ ,$$

где P_0 называется основной частью премии, а P_r – рисковой надбавкой. Из формулы (5.20) следует, что

$$P_0 = \frac{E\,S_n}{n} = mp\,,$$
$$P_r = u(\gamma)\frac{\sqrt{D\,S_n}}{n} = u(\gamma)\frac{\sqrt{p(m^2q + \sigma^2)}}{\sqrt{n}}\ .$$

Мы видим, что основная часть премии определяется математическим ожиданием ущерба по одному договору и вероятностью страхового случая. Рисковая надбавка зависит еще от дисперсии индивидуального ущерба и количества договоров в портфеле. Видно, что при увеличении количества договоров рисковая надбавка стремится к 0 .

6.1 Статистические модели. Задача точечного оценивания.

Исследование любого сложного объекта приводит к задаче изучения тех или иных числовых или качественных (категорных) характеристик этого объекта. Многие характеристики моделируются как случайные величины.

Пусть X – некоторая интересующая нас случайная величина. Проводится серия независимых одинаковых экспериментов, в каждом из которых измеряется значение величины X. На математическом языке эта серия экспериментов моделируется в виде последовательности независимых одинаково распределенных случайных величин $X_1, \ldots, X_n$, каждая из которых имеет то же распределение, что и величина X. Набор случайных величин $(X_1, \ldots, X_n)$ называют случайной выборкой. Любой возможный набор значений $(x_1, \ldots, x_n)$ случайной выборки $X_1, \ldots, X_n$ называется выборкой независимых наблюдений случайной величины X, n называется объемом выборки. Множество B всех возможных выборок называется выборочным пространством.

Распределение вероятностей величины X, вообще говоря, не известно. Если же мы по каким-то соображениям предполагаем, что закон распределения величины X принадлежит некоторому семейству вероятностных распределений $\mathcal{P}$, то в таком случае пара $(B,\mathcal{P})$ называется статистической моделью. Если семейство $\mathcal{P}$ может быть представлено как семейство распределений определенного вида, зависящего от параметра θ, то такую модель называют параметрической статистической моделью: $\mathcal{P} = \{P_\theta;\ \theta \in \Theta\}$.

Пример. X имеет нормальное распределение с параметрами μ и σ^2 : $X \sim N(\mu\,,\sigma^2)$. В этом случае $\theta = (\mu\,,\sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}^+$.

Пример. X имеет распределение Бернулли, $P(X=0) = 1-p$, $P(X=1) = p$. В этом случае параметром θ является вероятность p.

Задачей точечного оценивания является получение оценки неизвестного значения параметра θ (в виде числа). Естественно, что полученная оценка должна учитывать те наблюдения (выборку), которые стали известны на момент оценивания.

Определение 6.1. Оценкой (или статистикой) параметра θ называется произвольная функция $\hat{\theta}$ от случайной выборки $(X_1, \dots, X_n)$: $\hat{\theta} = \hat{\theta}\,(X_1, \dots, X_n)$.

Мы видим, что как функция от случайных величин, оценка $\hat{\theta}$ также является случайной величиной. Если в результате независимых наблюдений мы получаем выборку $(x_1, \dots, x_n)$, то $\hat{\theta}(x_1, \dots, x_n)$ будет числом, являющимся оценкой параметра θ.

Поскольку реализации (выборки) будут отличаться друг от друга в разных сериях наблюдений, то, вообще говоря, мы будем получать разные числовые оценки для θ. Статистик (оценок) существует очень много, и возникает вопрос о том, какие статистики можно считать «хорошими». Например, хотелось бы, чтобы в « среднем» статистика давала ответ, близкий к правильному.

Определение 6.2. Оценка $\hat{\theta}\,(X_1, \dots, X_n)$ называется несмещенной оценкой параметра θ, если для всех значений $\theta \in \Theta$

$$E_\theta \hat{\theta}\,(X_1, \dots, X_n) = \theta.$$

Пусть, например, случайная величина X имеет распределение с математическим ожиданием $m = EX$. Если m не известно, то эту величину можно рассматривать как параметр. Тогда выборочное среднее $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ является несмещенной оценкой для m. Но легко видеть, что любая статистика вида $\alpha_1 X_1 + \dots + \alpha_n X_n$, где $\alpha_1 + \dots + \alpha_n = 1$, также является несмещенной оценкой m.

Таким образом, несмещенных статистик также очень много. Как определить, какая из двух несмещенных оценок предпочтительней? Если есть две несмещенные статистики $\hat{\theta}_1$ и $\hat{\theta}_2$ и $D_\theta\,\hat{\theta}_1 \le D_\theta\,\hat{\theta}_2$ при всех θ, то говорят, что статистика $\hat{\theta}_1$ более эффективна, чем статистика $\hat{\theta}_2$ (здесь $D_\theta\,\hat{\theta} = E_\theta(\hat{\theta} - \theta)^2$). В этом случае первая статистика в «средне-квадратичном» более точна, чем вторая.

Определение 6.3. Несмещенная статистика $\hat{\theta}_n$ параметра θ называется эффективной, если она обладает наименьшей дисперсией в классе всех несмещенных статистик для этого параметра.

Метод максимального правдоподобия.

Обсудим теперь наиболее известный метод получения точечных оценок. Пусть независимые одинаково распределенные случайные величины $X_1, \dots, X_n$ имеют закон распределения $p(x;\theta)$, где $\theta = (\theta_1, \dots, \theta_m)$ – вектор параметров этого закона. Здесь под $p(x;\theta)$ мы понимаем плотность распределения случайных величин X_i том случае, если они абсолютно непрерывны, или же вероятность $p(x;\theta) = P\{X_i = x\}$ в случае дискретных величин X_i. Из независимости случайных величин $X_1, \dots, X_n$ следует, что закон распределения случайного вектора $(X_1, \dots, X_n)$ имеет вид:

$$p(x_1, \dots, x_n;\theta) = p(x_1;\theta) \dots p(x_n;\theta).$$

Определение 6.4. Функция $L(\theta; x) = p(x_1; \theta) \dots p(x_n; \theta)$, рассматриваемая как функция от θ при фиксированных значениях элементов выборки $x = (x_1, \dots x_n)$ называется функцией правдоподобия.

Определение 6.5. Метод максимального правдоподобия предлагает в качестве оценки параметра θ выбрать такое значение $\hat{\theta}_n = \hat{\theta}_n(x_1, \dots x_n)$, при котором функция правдоподобия принимает наибольшее значение:

$$L(\hat{\theta}_n; x) = \max_\theta L(\theta; x) = \max_\theta p(x_1; \theta) \dots p(x_n; \theta).$$

Такая статистика $\hat{\theta}_n$ называется оценкой максимального правдоподобия (о.м.п.).

Разумность метода максимального правдоподобия следует из того факта, что при заданной выборке $x = (x_1, \dots x_n)$ мы выбираем то значение θ, при котором появление выборки x наиболее вероятно. Фактически $L(\theta; x)$ является мерой правдоподобности наблюдения выборки x при значении θ. Иногда проще найти максимум не самой функции L, а ее логарифма $ln\, L$, поскольку максимум обеих функций достигается при одном и том же значении.

Пример. Пусть $X_1, \dots, X_n$ является случайной выборкой из нормального распределения $N(\mu; \sigma^2)$. В этом случае $\theta = (\mu; \sigma^2)$,

$$L(\theta; x) = \frac{1}{(\sqrt{2\pi}\sigma^2)^{n/2}}\, exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\},$$

$$ln\, L(\theta; x) = -\frac{n}{2}\ ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.$$

Дифференцируя $ln\, L(\theta)$ по μ и σ^2, мы получаем следующие оценки максимального правдоподобия (о. м. п.):

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}\,, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Не всегда удается свести задачу нахождения максимума функции правдоподобия к стандартной задаче математического анализа. Приведем два примера.

Пример. Пусть $X_1, \dots, X_n$ – случайная выборка из равномерного распределения $U[0, \theta]$. Тогда

$$L(\theta; x) = \begin{cases} \theta^{-n}, & \text{если } x_i \in [0, \theta],\ i = 1, \dots, n \\ 0, & \text{в противном случае} \end{cases}.$$

Если $max_i\, x_i \le \theta$, то $L(\theta; x) = \theta^{-n}$, если же $max_i\, x_i > \theta$, то $L(\theta; x) = 0$. Для достижения максимума $L(\theta; x)$ надо выбрать наименьшее значение θ, удовлетворяющее условию $max_i\, x_i \le \theta$. Таким образом, о.м.п. является $\hat{\theta} = max_i\, x_i$.

Пример. Предположим, что мы хотим оценить количество рыб в озере. Для этого сходили на рыбалку, поймали k рыб, пометили их и отпустили обратно. Через достаточное время вернулись на озеро, поймали l рыб, из которых x рыб оказались мечеными. Обозначим количество рыб в озере через n. Это число и является неизвестным параметром в этой задаче. Количество меченых рыб X во втором улове имеет гипергеометрическое распределение. Функция правдоподобия равна

$$L(n; x) = p(x; n) = \frac{C_k^x \cdot C_{n-k}^{l-x}}{C_n^l}\,,$$

В данном случае n является целочисленным параметром. Так как

$$\frac{L(n; x)}{L(n-1; x)} = \frac{(n-k)(n-l)}{n(n-k-l+x)},$$

то легко видеть, что последовательность $L(n; x)$ монотонно растет по n до тех пор, пока $n < \frac{kl}{x}$, далее она начинает убывать. Значит, о.м.п. в данной модели является целой частью числа $\frac{kl}{x}$: $\hat{n} = \left[\frac{kl}{x}\right]$.

6.2. Задача интервального оценивания.

Даже несмещенная точечная оценка $\hat{\theta}_n(x)$ параметра θ является приближенной оценкой этого параметра. Точечная оценка не несет информации о точности и достоверности оценивания. Интервальное или доверительное оценивание позволяет отвечать на такие вопросы.

Определение 6.6. Пусть $0 \le \gamma \le 1$. Интервальной оценкой параметра θ называется числовой интервал $\left(\hat{\theta}'_n, \hat{\theta}''_n\right)$ такой, что

$$P\left(\hat{\theta}'_n(x) \le \theta \le \hat{\theta}''_n(x)\right) \ge \gamma . \qquad (6.1)$$

Здесь $X = (X_1, \dots, X_n)$ – случайная выборка. Интервал $\left(\hat{\theta}'_n, \hat{\theta}''_n\right)$ называют доверительным интервалом, γ называют доверительной вероятностью или уровнем доверия. Ясно, что, как правило, интерес представляют значения γ, близкие к 1. Типичным выбором для γ являются значения γ=0,9 , γ=0,95 или γ=0,99 .

Условие (6.1) легко интерпретировать, используя теорему о том, что частота определенного события A в серии независимых испытаний стремится к вероятности этого события. Если у нас имеется N выборок $x^1 = (x_1^1, \dots, x_n^1)$, ... , $x^N = (x_1^N, \dots, x_n^N)$, то интервалы $\left(\hat{\theta}'_n(x^i),\ \hat{\theta}''_n(x^i)\right)$, $i = 1, \dots, N$ будут накрывать значения θ с частотой, превышающей значение γ при $N \to \infty$.

Обсудим вопрос о построении доверительного интервала для среднего нормального распределения.

Предположим, что среднее нормального распределения не известно, а дисперсия известна. Этот случай не совсем реалистичен, но он наиболее прост с математической точки зрения.

Если $X_1, \dots, X_n$ является случайной выборкой из распределения $N(\mu, \sigma^2)$, то как известно, выборочное среднее $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ также имеет нормальное распределение, $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.

Стандартизованная случайная величина $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ имеет стандартное нормальное распределение $N(0,1)$.

Критическим значением z_p распределения $N(0,1)$ называется число, удовлетворяющее условию

$$P(Z > z_p) = p , \quad 0 \le p \le 1 .$$

Заметим, что z_p фактически является $(1-p)$-квантилем $N(0,1)$ - распределения: $P(Z \le z_p) = 1 - p$. Из симметрии $N(0,1)$ – распределения следует, что

$$P(Z \le -z_p) = P(Z \ge z_p) = p.$$

Найдем симметричный доверительный интервал для μ с доверительной вероятностью γ. Пусть $\alpha = 1 - \gamma$. Найдем значение c, такое, что $P(-c \le Z \le c) = \gamma$:

$$P(-c \le z \le c) =$$
$$= 1 - P(Z > c) - P(Z < -c) = 1 - 2P(Z > c) = \gamma.$$

Отсюда

$$P(Z > c) = \frac{1-\gamma}{2} = \frac{\alpha}{2}.$$

Следовательно, $c = z_{\alpha/2}$. Значит,

$$P\left(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le z_{\alpha/2}\right) = P\left(\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \le \mu \le \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right) = \gamma .$$

Таким образом, интервал $\left(\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right)$ является доверительным интервалом для μ с доверительной вероятностью γ. Из таблиц нормального распределения можно увидеть, что при γ=0,9 $z_{(1-\gamma)/2} = z_{0,05} = 1{,}645$, при γ=0,95 $z_{(1-\gamma)/2} = z_{0,025} = 1{,}96$, при γ=0,99 $z_{(1-\gamma)/2} = z_{0,05} = 2{,}58$.

Предположим, что $X_1, \dots, X_n$ является случайной выборкой из негауссовского распределения со средним μ и дисперсией σ^2. Если объем выборки n достаточно велик, то по центральной предельной теореме величина

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

имеет приближенно стандартное нормальное распределение. Тогда

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma}\sqrt{n} < z_{\alpha/2}\right) \approx 1 - \alpha\,.$$

Если по каким-то причинам σ нам известно, то интервал $(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n})$ будет доверительным для μ с уровнем доверия $1-\alpha$. Если же стандартное отклонение не известно, то мы можем заменить его оценкой $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$. В этом случае интервал $(\bar{X} - z_{\alpha/2}S/\sqrt{n}$,
$\bar{X} + z_{\alpha/2}S/\sqrt{n})$ будет накрывать среднее μ приблизительно с вероятностью $1-\alpha$. Конечно, такое рассуждение справедливо в том случае, когда начинает «действовать» центральная предельная теорема. Статистики советуют пользоваться таким приближением при $n > 40$.

Важным применением такого рассуждения является задача о построении доверительного интервала для «доли» или «вероятности успеха». Предположим, что мы следим за некоторым событием A и хотим оценить вероятность его наступления. Например, событие состоит в том, что случайно выбранный избиратель проголосует за определенного кандидата в президенты или за определенное решение в сфере политики или экономики. Пусть $p = P(A) -$ вероятность события A, и p является в нашем случае неизвестным параметром. Введем случайную величину X, которая принимает значение 1 в случае наступления события A и 0 в противном случае. В результате n независимых наблюдений мы получаем случайную выборку $X_1, \dots, X_n$ с распределением вида

$$P(X_i = x_i) = p^{x_i}(1-p)^{1-x_i}\ , \qquad x_i = 0{,}1\,.$$

Функция правдоподобия имеет вид

$$L(p;x) = p^{x_1}(1-p)^{1-x_1}\ \dots\ p^{x_n}(1-p)^{1-x_n} = p^{x_1+\dots+x_n}(1-p)^{n-(x_1+\dots+x_n)}\,.$$

Логарифмируя эту функцию и дифференцируя по p, мы получим о.м.п. для p :

$$\hat{P} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i\,.$$

Легко видеть, что оценка является несмещенной, и можно доказать , что она является эффективной. Дисперсия этой оценки равна

$$D_p\hat{P} = \frac{1}{n}D_pX_1 = \frac{1}{n}p(1-p).$$

Величина $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ будет иметь приближенно стандартное гауссовское распределение и, следовательно,

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \approx 1 - \alpha\,. \qquad (6.2)$$

Так как значение p в знаменателе нам не известно, то мы можем заменить его на оценку $\hat{p}$ и получить следующий приближенный доверительный интервал для p :

$$P\big(\ \hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\ \big) \approx 1-\alpha\,.$$

6.3. Проверка гипотез.

Статистической гипотезой называется утверждение о каких-то числовых характеристиках (параметрах) или свойствах изучаемого распределения или о виде самого распределения в целом. Примерами таких предположений могут быть следующие утверждения:

1) Доля студентов факультета ИВМ и ИТ, устраивающихся на работу по специальности, равна 0,8.
2) Проведение рекламной компании увеличит объем продаж на 10%.
3) Средняя продолжительность жизни женщин на 10 лет больше средней продолжительности жизни мужчин.

Такие гипотезы постоянно возникают в различных областях жизни, бизнеса, политики и т.д.

Предположим, что есть основная (нулевая) гипотеза H_0, и необходимо найти критерий, который, в зависимости от результатов наблюдений, позволяет отклонять или не отклонять эту гипотезу. Часто формулируется альтернативная гипотеза H_1, которая находится в противоречии с гипотезой H_0.

Гипотеза называется простой, если она полностью определяет распределение вероятностей изучаемой величины. Например, в рамках параметрической статистической модели гипотеза вида $H_0 : \theta = \theta_0$ однозначно определяет распределение $p(x;\theta)$ и поэтому является простой. Гипотеза вида $H_0 : \theta < \theta_0$ является сложной.

Предположим, что основная и альтернативная гипотезы являются простыми: $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$. Мы должны сформулировать критерий, который бы позволил по выборке $(X_1, \dots, X_n)$ определить, надо ли нам отвергнуть гипотезу H_0 и, следовательно, принять H_1 или же не отвергать гипотезу H_0 и отвергнуть H_1. Здесь можно провести аналогию с «презумпцией невиновности» в юриспруденции. Если нет серьезных доказательств в виновности человека, он считается невиновным.

Пусть B – выборочное пространство. Выделим в нем подмножество $B_1 \subset B$, которое назовем критической областью. Если выборка $x = (x_1, \dots, x_n)$ попадает в B_1, то мы отвергаем «нулевую» гипотезу H_0 и примем альтернативную H_1. Если же выборка x попадет в область $B_0 = B \backslash B_1$, то мы примем (не отвергнем) H_0 и отвергнем H_1.

В результате применения такого критерия могут быть сделаны ошибки двух типов. Ошибка первого рода возникает тогда, когда мы отвергаем основную гипотезу, в то время как она верна. Ошибка второго рода возникает тогда, когда мы отвергаем альтернативную гипотезу, в то время как она верна.

Вероятность ошибки первого рода равна вероятности области B_1, вычисленной по распределению $P(X;\theta_0)$: $\alpha = P(B_1/H_0)$. Вероятность ошибки второго рода равна вероятности области B_0, вычисленной по распределению $P(X;\theta_1)$: $\beta = P(B_0/H_1)$.

Ясно, что хотелось бы иметь такую критическую область B_1, для которой вероятности ошибок обоих родов были бы наименьшими. Но заметим, что $\beta = 1 - P(B_1/H_1)$. Уменьшение ошибки первого можно проводить за счет уменьшения области B_1. Но тогда будет уменьшаться величина $P(B_1/H_1)$, и, значит, увеличиваться вероятность ошибки второго рода.

Статистики предлагают следующий компромисс: зафиксировать вероятность ошибки первого рода α, и среди всех областей B_1 с такой ошибкой искать область с наименьшей ошибкой второго рода.

Величину $\alpha = P(B_1/H_0)$ называют еще размером критической области B_1, а величину $1-\beta = P(B_1/H_1)$ называют мощностью критической области. Отсюда возникает такое понятие:

Определение 6.7. Оптимальной критической областью размера α называется область, имеющая наибольшую мощность среди всех областей размера α.

Статистики придумали методы, позволяющие находить оптимальные критические области.

Ошибка первого рода обычно менее приемлема, чем ошибка второго рода. Поэтому вероятность ошибки первого рода заранее фиксируется на заданном уровне. Типичные значения $\alpha = 0{,}1$, $\alpha = 0{,}05$, $\alpha = 0{,}01$.

Рассмотрим задачу проверки гипотезы о среднем значении нормального распределения $N(\mu, \sigma^2)$ при известном значении σ^2. Пусть нулевая гипотеза $H_0: \mu = \mu_0$, альтернативная гипотеза $H_1: \mu > \mu_0$. Заметим, что гипотеза H_1 -- сложная. Вспомним статистику Z, которую мы изучали при построении доверительного интервала для μ:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

Здравый смысл говорит о том, что если $\bar{X}$ существенно больше μ_0, то, по-видимому, следует отказаться от гипотезы H_0, в противном случае не следует от нее отказываться. Поэтому естественно определить критическую область с помощью неравенства $\bar{Z} = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \geq$ c. Заметим, что эта область имеет вид полупространства R^n:

$$V = \left\{(x_1, \dots, x_n) : \sum_{i=1}^{n} x_i \geq c'\right\}, \quad c' = n\mu_0 + c\sigma\sqrt{n}\ .$$

Вероятность ошибки первого рода

$$\alpha = P\{Z \geq c | Z \sim N(0,1)\} = 1 - \Phi(\text{c}).$$

Отсюда видно, что критическое z_α, которое определялось соотношением $P(Z \geq z_\alpha) = \alpha$, задает искомое значение c.

Если альтернативная гипотеза имеет вид $H_1: \mu < \mu_0$, то естественно искать критическую область в виде $z \leq c$, где c уже является отрицательным числом. Если верна гипотеза H_0, то $Z \sim N(0,1)$ и так как $P(Z \leq -z_\alpha) = \alpha$, то критическая область с заданным уровнем значимости α может быть выбрана как $V = \{x: z \leq -z_\alpha\}$.

В случае, когда $H_1: \mu \neq \mu_0$, то критическую область естественно искать в виде $V = \{x: |z| \geq c\}$. Если $Z \sim N(0,1)$, то

$$P(|Z| \geq c) = P(Z \leq -c) + P(Z \geq c),$$

и при $c = z_{\alpha/2}$ мы получаем критерий с уровнем значимости $\alpha: \alpha = P\left(|Z| \geq z_{\alpha/2}\right)$.

Вычислим вероятность ошибки в том случае, когда мы зафиксируем альтернативное значение $\mu: \mu = \mu_1$. Пусть, например, $\mu_1 > \mu_0$. Если в качестве критической области мы выберем $V = \{x: z \geq z_\alpha\}$, то вероятность ошибки второго рода будет равна

$$\beta(\mu_1) = P(z < z_\alpha \mid \mu = \mu_1) = P\left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_\alpha \mid \mu = \mu_1\right) =$$
$$= P\left(\bar{X} \leq \mu_0 + \frac{\sigma z_\alpha}{\sqrt{n}} | \mu = \mu_1\right).$$

Если $\mu = \mu_1$, то $\bar{X} \sim N\left(\mu_1, \frac{\sigma^2}{n}\right)$. Поэтому

$$\beta(\mu_1) = P\left(\bar{X} \leq \mu_0 + \frac{\sigma z_\alpha}{\sqrt{n}} \mid \mu = \mu_1\right) =$$
$$= P\left(\frac{\bar{X} - \mu_1}{\sigma/\sqrt{n}} \leq z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} \ |\mu = \mu_1\right) = \Phi\left(z_\alpha + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}}\right).$$

Видно, что вероятность ошибки второго рода уменьшается с ростом μ_1.