Теория вероятностей и математическая статистика

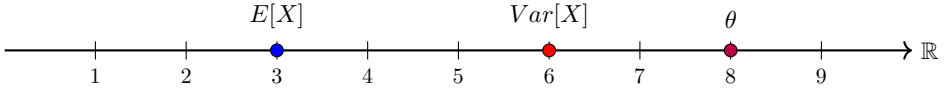
Выборочные распределения. Точечные оценки.

Глеб Карпов

ФКН ВШЭ

Введение в статистику: визуализация

Идеальная ситуация: мы знаем параметры и/или характеристики



Реальная ситуация: параметры и/или характеристики неизвестны

Значения будто скрыты от нас туманом

- Параметры и характеристики случайной величины — числа, **точки** на числовой оси. Семейство методов для угадывания этих значений обычно называется **точечной оценкой**.

Случайная выборка и её реализация

- Собранные данные обычно называются выборкой, но в статистике мы одновременно имеем дело с двумя различными типами выборок.
- Случайная выборка — это вектор (коллекция, набор, совокупность) независимых и одинаково распределенных (i.i.d.) случайных величин:

$$\mathcal{X} = (X_1, X_2, \ldots, X_n), \quad f_{X_i}(x) = f_{X_j}(x), \ \forall i, j \in [1, n], \forall x.$$

- Реализация случайной выборки — это набор наблюдений из случайной выборки $\mathcal{X}$, набор конкретных чисел:

$$x = (x_1, x_2, \ldots, x_n).$$

- Генеральная совокупность (популяция) — полное множество объектов, обладающих интересующим признаком, несущих реализацию интересующей нас случайной величины. Извлекая наблюдения из генеральной совокупности, мы можем сформировать реализацию выборки.

Выборочные распределения

- Пусть $\mathcal{X} = (X_1, \dots, X_n)$ — случайная выборка со средним $\mu = E[X_i]$ и дисперсией $\sigma^2 = Var[X_i] < \infty$.
- Возможная статистика — это, например, рассмотренная ранее сумма всех элементов $S_n = \sum_{i=1}^{n} X_i$. Её характеристики: $E[S_n] = n\mu$, $Var[S_n] = n\sigma^2$
- Одна из самых важных статистик — выборочное среднее:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Характеристики выборочного среднего $\bar{X}$:
  - $E[\bar{X}] = E\left[\frac{X_1}{n} + \dots + \frac{X_n}{n}\right] = \frac{1}{n}E\left[X_1 + \dots + X_n\right] = \frac{n\mu}{n} = \mu$
  - $\mathrm{Var}\left[\bar{X}\right] = \mathrm{Var}\left[\frac{X_1}{n} + \dots + \frac{X_n}{n}\right] = \frac{1}{n^2}\mathrm{Var}\left[X_1 + \dots + X_n\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$
- При выполнении условий ЦПТ ($n \geq 30$) можем заявлять, что:

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Распределение X^2 (хи-квадрат)

- Пусть $Z_1, \dots, Z_k$ — независимые стандартные нормальные случайные величины: $Z_i \sim \mathcal{N}(0, 1)$.
- Определим новую случайную величину:

$$\chi^2(k) = \sum_{i=1}^{k} Z_i^2$$

- Распределение такой случайной величины называется **распределением хи-квадрат**.
- Это распределение играет важную роль в статистических процедурах.

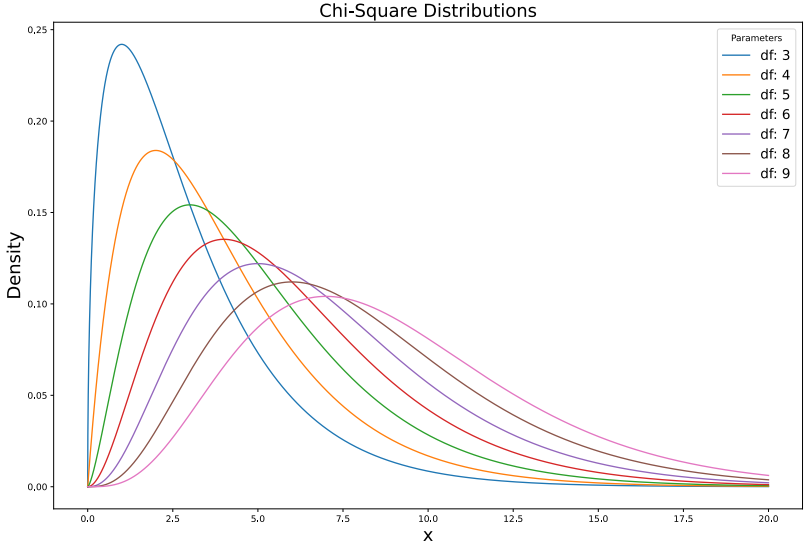
Степени свободы

- Функция плотности распределения хи-квадрат **критично зависит** от числа слагаемых k в сумме.
- Число слагаемых k называется **степенями свободы** (degrees of freedom, df).
- Правильное обозначение: $\chi^2(k)$ — читается как "хи-квадрат распределение с k степенями свободы".

$$W \sim N(\mu, \sigma^2)$$

$$X^2 \sim \chi^2(K)$$

Функция плотности распределения хи-квадрат



Математическое ожидание и дисперсия распределения хи-квадрат

- Если $Y \sim \chi^2(k)$, то:

$$E[Y] = k, \quad \mathsf{Var}(Y) = 2k$$

- Доказательство для $E[Y]$:

  Поскольку $Y = \sum_{i=1}^{k} Z_i^2$, где $Z_i \sim \mathcal{N}(0,1)$ независимы:

$$E[Y] = \sum_{i=1}^{k} E[Z_i^2] = \sum_{i=1}^{k} \left(\mathsf{Var}(Z_i) + (E[Z_i])^2\right) = \sum_{i=1}^{k} (1+0) = k$$

Свойство суммы отклонений от среднего

Theorem

Пусть $X_1, \dots, X_k$ — случайная выборка из некоторой генеральной совокупности. Тогда:

$$\sum_{i=1}^{k} (X_i - \bar{X}) = 0$$

где $\bar{X} = \frac{1}{k} \sum_{i=1}^{k} X_i$ — выборочное среднее.

$$\sum_{i=1}^{k} X_i - \sum_{i=1}^{k} \bar{X} = \sum_{i=1}^{k} X_i - k\bar{X} =$$

$$= \sum_{i=1}^{k} X_i - \frac{\sum X_i}{k} \cdot k = 0$$

$\bar{X}$

Свойство суммы отклонений от среднего

Proof

$$\sum_{i=1}^{k}\left(X_i - \bar{X}\right) = \sum_{i=1}^{k}\left(X_i\right) - k\bar{X}$$

$$= \sum_{i=1}^{k} X_i - k\frac{\sum_{i=1}^{k} X_i}{k}$$

$$= \sum_{i=1}^{k} X_i - \sum_{i=1}^{k} X_i = 0$$

Теорема о $k\bar{Z}^2$

Theorem

Пусть $Z_1, \dots, Z_k$ — случайная выборка из стандартного нормального распределения, т.е. $Z_i \sim \mathcal{N}(0, 1)$. Тогда случайная величина $k\bar{Z}^2$ имеет распределение хи-квадрат с одной степенью свободы:

$$k\bar{Z}^2 \sim \chi^2(1)$$

Теорема о $k\bar{Z}^2$

Proof

- Характеристики:

$$E[\bar{Z}] = \frac{1}{k}\sum_{i=1}^{k} E[Z_i] = 0, \quad \mathsf{Var}[\bar{Z}] = \frac{1}{k^2}\sum_{i=1}^{k} \mathsf{Var}[Z_i] = \frac{1}{k^2}k = \frac{1}{k}$$

- По свойству устойчивости, как сумма независимых нормальных случайных величин, $\bar{Z} \sim \mathcal{N}\left(0, \frac{1}{k}\right)$
- Далее рассмотрим случайную величину $\sqrt{k}\bar{Z}$:

$$E[\sqrt{k}\bar{Z}] = 0, \quad \mathsf{Var}(\sqrt{k}\bar{Z}) = k\mathsf{Var}(\bar{Z}) = k\frac{1}{k} = 1$$

- Значит, $\sqrt{k}\bar{Z} \sim \mathcal{N}(0,1)$ — имеет стандартное нормальное распределение. Поэтому одна такая величина в квадрате будет распределена как хи-квадрат:

$$(\sqrt{k}\bar{Z})^2 = k\bar{Z}^2 \sim \chi^2(1)$$

Разложение суммы квадратов

- Рассмотрим преобразование:

$$\sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{k} \left[(Z_i - \bar{Z} + \bar{Z})^2 \right] = \sum_{i=1}^{k} \left[\Big((Z_i - \bar{Z}) + \bar{Z} \Big)^2 \right]$$

$$= \sum_{i=1}^{k} \left[(Z_i - \bar{Z})^2 + 2 (Z_i - \bar{Z}) \bar{Z} + \bar{Z}^2 \right]$$

$$= \sum_{i=1}^{k} (Z_i - \bar{Z})^2 + 2\bar{Z} \sum_{i=1}^{k} (Z_i - \bar{Z}) + k\bar{Z}^2$$

- По свойству суммы отклонений: $\sum\limits_{i=1}^{k} (Z_i - \bar{Z}) = 0$
- Получаем:

$$\sum_{i=1}^{k} Z_i^2 = \sum_{i=1}^{k} (Z_i - \bar{Z})^2 + k\bar{Z}^2$$

Распределение суммы квадратов отклонений

- Из разложения:

$$\underbrace{\sum_{i=1}^{k} Z_i^2}_{k \text{ степеней свободы}} = \underbrace{\sum_{i=1}^{k} (Z_i - \bar{Z})^2}_{(k-1) \text{ степеней свободы}} + \underbrace{k\bar{Z}^2}_{1 \text{ степень свободы}}$$

- Известно:
  - $\sum_{i=1}^{k} Z_i^2 \sim \chi^2(k)$, и $k\bar{Z}^2 \sim \chi^2(1)$
- В итоге получаем, что оставшееся слагаемое имеет $(k-1)$ степень свободы:

$$\sum_{i=1}^{k} (Z_i - \bar{Z})^2 \sim \chi^2(k-1)$$

- **Ключевой результат**: Сумма квадратов отклонений от выборочного среднего имеет распределение хи-квадрат с $(k-1)$ степенями свободы. Это фундаментальный результат для статистических тестов и доверительных интервалов.

Выборочная дисперсия

- Пусть $X_1, \dots, X_n$ — случайная выборка из нормального распределения: $X_i \sim \mathcal{N}(\mu, \sigma^2)$.
- **Выборочная дисперсия**:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

- Выборочная дисперсия используется для оценки неизвестной дисперсии генеральной совокупности σ^2.

Связь выборочной дисперсии с хи-квадрат

- Нормализуем наблюдения: $Z_i = \frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0, 1)$
- Тогда выборочное среднее нормализованных величин: $\bar{Z} = \frac{\bar{X} - \mu}{\sigma}$
- Из предыдущих результатов:

$$\sum_{i=1}^{n} (Z_i - \bar{Z})^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} - \frac{\bar{X} - \mu}{\sigma} \right)^2 = \boxed{\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 \sim \chi^2(n-1)}$$

Распределение выборочной дисперсии

- Подставляя определение выборочной дисперсии:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

- **Важный результат**:

$$\boxed{\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)}$$

- Это означает, что выборочная дисперсия имеет распределение хи-квадрат с $(n-1)$ степенями свободы.
- **Математическое ожидание и дисперсия**:

$$E\left[\frac{(n-1)s^2}{\sigma^2}\right] = n-1, \quad \mathsf{Var}\left[\frac{(n-1)s^2}{\sigma^2}\right] = 2(n-1)$$

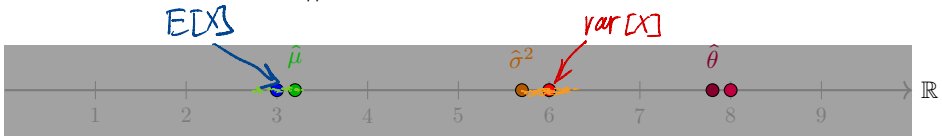
- **Следствия для выборочной дисперсии**:

$$E[s^2] = \sigma^2, \quad \mathsf{Var}(s^2) = \frac{2\sigma^4}{n-1}$$

Цель точечной оценки

- На основе реализации случайной выборки $x_1, x_2, \ldots, x_n$ получить **предположения** $\hat{\theta}$ о значениях скрытых в тумане реальности параметров.
- Идея состоит в том, чтобы посчитать значение оценки на реальных имеющихся данных, и чтобы полученное число было бы как можно ближе к истинному значению параметра.
- Следуя аналогии, мы хотим найти затерянные в тумане точки, путём их угадывания специальным способом, с помощью функции **оценки**.

Значения точечных оценок $\hat{\mu}$, $\hat{\sigma}^2$ и $\hat{\theta}$
"попадают" близко к истинным значениям



$\hat{\mu} = \bar{X}$

$\sigma^2 = s^2$

Уже известные точечные оценки

Пусть у нас есть случайная выборка $\mathcal{X} = (X_1, X_2, \dots, X_n)$ (Независимые, одинаково распределенные) с $\mu \equiv E[X_i]$, $\sigma^2 \equiv Var[X_i]$.

1. Выборочное среднее $\bar{X}$

- Определение: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
- Характеристики: $E[\bar{X}] = \mu$ (несмещенная оценка), $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$
- Распределение:
  - Если $X_i \sim \mathcal{N}(\mu, \sigma^2)$, то $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$
  - По ЦПТ: при больших n выполняется $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

2. Выборочная дисперсия S^2

- Определение: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- Характеристики: $E[S^2] = \sigma^2$ (несмещенная оценка), $\text{Var}(S^2) = \frac{2\sigma^4}{n-1}$
- Распределение: если $X_i \sim \mathcal{N}(\mu, \sigma^2)$, то:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$