Интервальные оценки в статистике

Когда мы говорим о точечных оценках, мы даём **одно число** в качестве приближённого значения параметра генеральной совокупности. Однако такие оценки **не учитывают неопределённость выборки** и могут сильно варьироваться.

Интервальная оценка (confidence interval, доверительный интервал) — это более информативный способ оценки, который даёт **диапазон значений**, в котором с некоторой вероятностью находится истинное значение параметра.

1. **Определение интервальной оценки**

Интервальная оценка параметра θ — это два числа $[\theta_L, \theta_U]$, такие что:

$$P(\theta_L \leq \theta \leq \theta_U) = 1 - \alpha.$$

Где:

- $[\theta_L, \theta_U]$ — **доверительный интервал** (confidence interval, CI),
- $1 - \alpha$ — **доверительная вероятность (доверительный уровень)**, обычно 95% или 99%,
- α — уровень значимости, вероятность ошибки.

Если $1 - \alpha = 0.95$, это означает, что **в 95% случаев, если мы повторим эксперимент, истинное значение параметра попадёт в этот интервал**.

Хи-квадрат распределение ($\chi 2$-распределение)

1. Формулировка хи-квадрат распределения

Хи-квадрат распределение ($\chi 2$-распределение) — это семейство распределений, используемых в статистике, особенно в проверке гипотез, анализе дисперсии и оценке разброса данных.

Определение:

Пусть $Y_1, Y_2, ..., Y_n$ — независимые стандартные нормальные случайные величины, то есть:

$$Y_i \sim N(0,1), \quad i = 1, 2, ..., n.$$

Тогда **случайная величина**, определённая как сумма квадратов этих величин,

$$\chi_n^2 = \sum_{i=1}^{n} Y_i^2,$$

имеет **хи-квадрат распределение с n степенями свободы** (обозначается $\chi^2(n)$).

Характеристическая функция

Характеристическая функция случайной величины — это один из ключевых инструментов в теории вероятностей, который полностью определяет её распределение и позволяет доказывать теоремы о предельных распределениях (например, центральную предельную теорему).

1. Определение

Пусть X — случайная величина. **Характеристическая функция** $\varphi_X(t)$ определяется как **математическое ожидание** комплексной экспоненты:

$$\varphi_X(t) = E[e^{itX}], \quad t \in \mathbb{R}.$$

где i — мнимая единица ($i^2 = -1$).

◆ **Альтернативная запись через интеграл:**

Если у X есть плотность вероятности $f_X(x)$, то:

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} f_X(x)\, dx.$$

◆ **Если X дискретная:**

Если X принимает значения x_k с вероятностями p_k, то:

$$\varphi_X(t) = \sum_k e^{itx_k} p_k.$$

2. Свойства характеристической функции

✅ **1. Характеристическая функция однозначно определяет распределение случайной величины** (то есть если две случайные величины имеют одинаковые характеристические функции, то они имеют одинаковые распределения).

✅ **2.** $\varphi_X(0) = 1$
Так как $e^0 = 1$, получаем:

$$\varphi_X(0) = E[e^0] = E[1] = 1.$$

✅ **3.** $|\varphi_X(t)| \leq 1$
Так как модуль экспоненты $|e^{itX}| = 1$, характеристическая функция не превосходит 1 по модулю.

✅ **4. Связь с моментами**
Если у X есть математическое ожидание и дисперсия, то:

$$\varphi'_X(0) = iE[X], \quad \varphi''_X(0) = -E[X^2].$$

✅ **5. Связь с функцией распределения**
Если $F_X(x)$ — функция распределения случайной величины X, то:

$$\varphi_X(t) = \int_{-\infty}^{\infty} e^{itx} dF_X(x).$$

✅ **6. Характеристическая функция суммы независимых величин**
Если X и Y независимы, то:

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

3. Примеры характеристических функций

1️⃣ Характеристическая функция нормального распределения $N(\mu, \sigma^2)$

Если $X \sim N(\mu, \sigma^2)$, то:

$$\varphi_X(t) = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}.$$

2️⃣ Характеристическая функция стандартного нормального распределения $N(0, 1)$

$$\varphi_X(t) = e^{-\frac{t^2}{2}}.$$

3 Характеристическая функция равномерного распределения $U(a, b)$

Если $X \sim U(a, b)$, то:

$$\varphi_X(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}.$$

---

4 Характеристическая функция хи-квадрат распределения χ_n^2

Если $X \sim \chi_n^2$, то:

$$\varphi_X(t) = (1 - 2it)^{-n/2}, \quad \text{для } t < \frac{1}{2}.$$

4. Применение характеристической функции

Доказательство Центральной предельной теоремы (ЦПТ) Показывает, что характеристическая функция суммы независимых случайных величин приближается к характеристической функции нормального распределения.

Доказательство распределения хи-квадрат и t-Стьюдента Используется для нахождения предельных распределений.

Преобразование Фурье в теории вероятностей Характеристическая функция — это обобщённое преобразование Фурье, которое позволяет переходить между функцией распределения и её спектральным представлением.

Вернемся к Хи-квадрат:

2. Найдём характеристическую функцию χ2

Плотность стандартного нормального распределения:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}.$$

Теперь найдём характеристическую функцию (функцию Лапласа) суммы квадратов нормальных величин:

$$M_{\chi_n^2}(t) = E[e^{t\chi_n^2}].$$

Так как $Y_1, Y_2, \ldots, Y_n$ независимы:

$$M_{\chi_n^2}(t) = E\left[e^{t\sum Y_i^2}\right] = \prod_{i=1}^{n} E[e^{tY_i^2}].$$

Найдём характеристическую функцию одной стандартной нормальной величины:

$$M_{Y^2}(t) = \int_{-\infty}^{\infty} e^{ty^2} \cdot \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

Это интеграл формы:

$$\int_{-\infty}^{\infty} e^{-\frac{1}{2}(y^2 - 2ty^2)} dy.$$

После замены $\sigma^2 = \frac{1}{1-2t}$, получаем:

$$M_{Y^2}(t) = (1 - 2t)^{-1/2}, \quad t < \frac{1}{2}.$$

Так как все Y_i независимы, их сумма даёт:

$$M_{\chi_n^2}(t) = (1 - 2t)^{-n/2}, \quad t < \frac{1}{2}.$$

Эта характеристическая функция совпадает с характеристической функцией хи-квадрат распределения:

$$\chi_n^2 \sim \text{Chi-Square}(n).$$

3. Вычислим плотность вероятности χ2

Рассмотрим плотность одной случайной величины $Z = Y^2$, где $Y \sim N(0, 1)$.

Преобразуем:

$$P(Z \leq x) = P(Y^2 \leq x) = P(-\sqrt{x} \leq Y \leq \sqrt{x}).$$

Дифференцируя по x, получаем плотность случайной величины Z:

$$f_Z(x) = \frac{1}{\sqrt{2\pi}} e^{-x/2} \cdot \frac{1}{\sqrt{x}}.$$

Теперь рассмотрим сумму n таких величин. Можно показать, что плотность распределения:

$$f_{\chi_n^2}(x) = \frac{x^{n/2-1} e^{-x/2}}{2^{n/2} \Gamma(n/2)}, \quad x > 0.$$

где $\Gamma(n/2)$ — гамма-функция:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt.$$

Это и есть плотность хи-квадрат распределения.

4. Числовые характеристики

Математическое ожидание (Среднее значение)

Математическое ожидание (среднее) хи-квадрат распределения равно:

$$E[\chi_n^2] = n.$$

Доказательство:

Хи-квадрат распределение определяется как сумма квадратов n независимых стандартных нормальных величин:

$$\chi_n^2 = \sum_{i=1}^{n} Y_i^2, \quad \text{где } Y_i \sim N(0,1).$$

Для стандартного нормального распределения известно, что:

$$E[Y_i^2] = \mathrm{Var}(Y_i) + (E[Y_i])^2 = 1 + 0 = 1.$$

Так как сумма математических ожиданий равна сумме индивидуальных ожиданий:

$$E[\chi_n^2] = \sum_{i=1}^{n} E[Y_i^2] = \sum_{i=1}^{n} 1 = n.$$

Дисперсия

Дисперсия хи-квадрат распределения определяется как:

$$\mathrm{Var}(\chi_n^2) = 2n.$$

Доказательство:

Так как $Y_i^2 \sim \chi_1^2$ (хи-квадрат распределение с 1 степенью свободы), его дисперсия равна:

$$\mathrm{Var}(Y_i^2) = 2.$$

Так как сумма независимых случайных величин приводит к сложению их дисперсий:

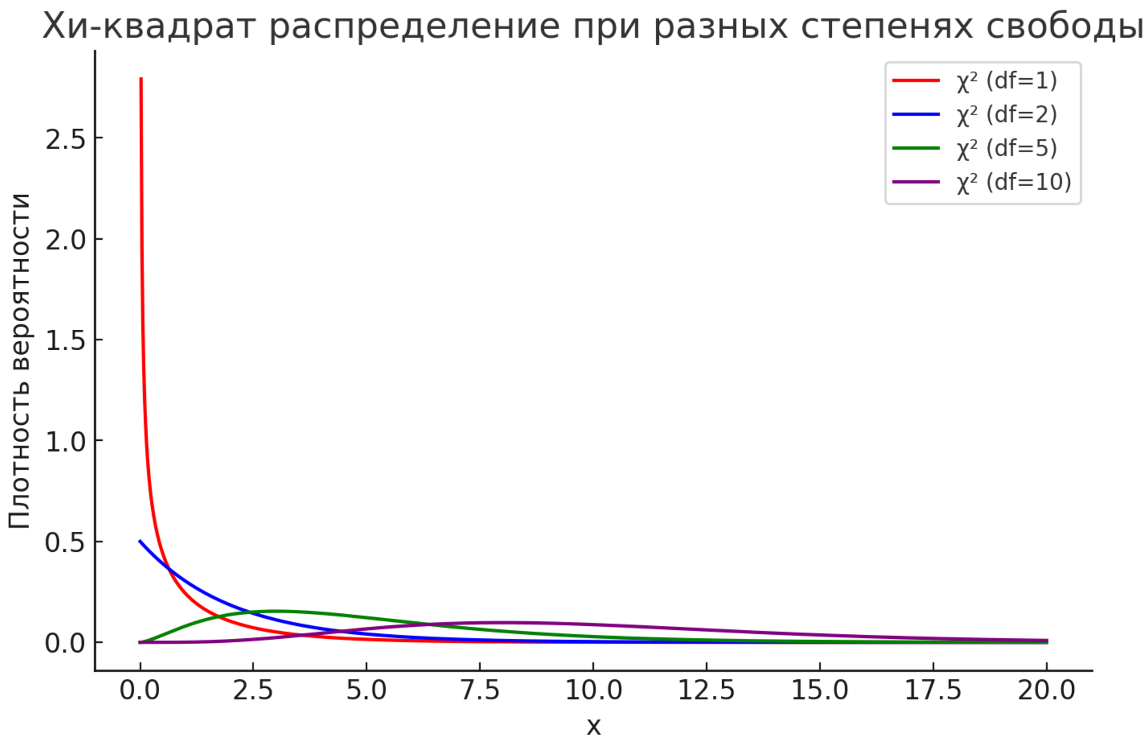
$$\mathrm{Var}(\chi_n^2) = \sum_{i=1}^{n} \mathrm{Var}(Y_i^2) = \sum_{i=1}^{n} 2 = 2n.$$

5. Где применяется Хи-квадрат?

1. Проверка гипотез (критерий хи-квадрат Пирсона)

2. Доверительные интервалы

6. Визуализация хи-квадрат распределения



На графике показано, как ведёт себя хи-квадрат распределение при разном числе степеней свободы:

- При df=1 (красная кривая): распределение сильно скошено вправо.
- При df=2 (синяя кривая): появляется небольшой пик, но распределение всё ещё несимметрично.
- При df=5 (зелёная кривая): пик становится более выраженным, и распределение начинает приближаться к нормальному.
- При df=10 (фиолетовая кривая): распределение становится более гладким и начинает напоминать нормальное.

💡 **Вывод:**

- Чем больше степеней свободы n, тем ближе $\chi^2(n)$ к нормальному распределению с параметрами $N(n, 2n)$.
- Малые n приводят к скошенному распределению, что важно учитывать в статистических тестах.

Хи-квадрат распределение активно используется в проверке гипотез, оценке дисперсии и анализе независимости категориальных данных.

Распределение Стьюдента (t-распределение)

Распределение Стьюдента (или **t-распределение Стьюдента**) — это важное распределение в математической статистике, используемое при работе с небольшими выборками, когда дисперсия генеральной совокупности неизвестна. Оно играет ключевую роль в **t-тестах** и построении **доверительных интервалов**.

1. Формулировка распределения Стьюдента

Распределение t-Стъюдента определяется следующим образом:

$$t = \frac{Y_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2}}$$

имеет **распределение Стьюдента с n степенями свободы**, используя свойства стандартного нормального распределения и хи-квадрат распределения.

1. Анализ слагаемых в формуле

Пусть $Y_0, Y_1, \ldots, Y_n$ — независимые стандартные нормальные случайные величины, то есть:

$$Y_i \sim N(0,1), \quad i = 0, 1, \ldots, n.$$

Рассмотрим отдельно числитель и знаменатель:

- **Числитель:** $Y_0 \sim N(0,1)$.
- **Знаменатель:**

$$\sum_{i=1}^{n} Y_i^2.$$

Мы знаем, что сумма квадратов n независимых стандартных нормальных величин подчиняется **хи-квадрат распределению** с n степенями свободы:

$$\sum_{i=1}^{n} Y_i^2 \sim \chi_n^2.$$

2. Преобразование знаменателя

Преобразуем знаменатель:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n} Y_i^2} = \sqrt{\frac{\chi_n^2}{n}}.$$

Таким образом, выражение для случайной величины t можно записать как:

$$t = \frac{Y_0}{\sqrt{\frac{\chi_n^2}{n}}}.$$

3. Использование определения t-распределения

Сравним это с классическим определением распределения Стьюдента:

$$T = \frac{Z}{\sqrt{\chi_n^2/n}} \sim t_n,$$

где:

- $Z \sim N(0,1)$ — стандартная нормальная случайная величина,
- χ_n^2 — независимая случайная величина, имеющая хи-квадрат распределение с n степенями свободы.

Так как:

- $Y_0 \sim N(0,1)$ соответствует Z,
- $\sum\limits_{i=1}^{n} Y_i^2 \sim \chi_n^2$,

то выражение

$$t = \frac{Y_0}{\sqrt{\frac{\chi_n^2}{n}}}$$

подчиняется распределению Стьюдента с n степенями свободы:

$$t \sim t_n.$$

4. Числовые характеристики распределения Стьюдента t_n

Распределение Стьюдента (t_n, или t-распределение) имеет **ключевые числовые характеристики**, которые зависят от **числа степеней свободы** n. Это важно, так как при малых n распределение значительно отличается от нормального, а при больших n приближается к $N(0,1)$.

4.1. Математическое ожидание (Среднее значение)

$$E[T] = \begin{cases} 0, & \text{если } n > 1, \\ \text{не определено}, & \text{если } n \leq 1. \end{cases}$$

🔷 **Интерпретация:**

- При $n > 1$ математическое ожидание равно **нулю**, поскольку распределение симметрично относительно 0.
- При $n \leq 1$ математическое ожидание **не существует**, так как интеграл не сходится.

4.2. Дисперсия

$$\mathrm{Var}(T) = \begin{cases} \frac{n}{n-2}, & \text{если } n > 2, \\ \infty, & \text{если } n \leq 2. \end{cases}$$

🔷 **Интерпретация:**

- Дисперсия распределения Стьюдента **существует только при** $n > 2$.
- Если $n = 1$ (распределение Коши), дисперсия **не существует**.
- Если $n = 2$, математическое ожидание существует, но **дисперсия не существует**.

Почему распределение назвали "Стьюдента"?

🔷 В 1908 году Уильям Госсет (William Sealy Gosset) опубликовал статью о t-распределении в журнале "Biometrika" под псевдонимом "Student".

🔷 Он работал в пивоваренной компании Guinness и занимался анализом небольших выборок.

🔷 Guinness запрещала сотрудникам публиковать научные работы, поэтому Госсет взял псевдоним "Student" (студент).

🔷 В результате t-распределение стали называть распределением Стьюдента.