Отношение сигнал/шум в астрофизических наблюдениях

Лекция 2

Статистические основы астрономических измерений. Случайные величины, среднее значение, дисперсия, стандартное отклонение, погрешности измерений

КазНУ имени аль-Фараби

Цель и план лекции

Цель

Изучение статистических понятий, необходимых для оценки шума в астрофизических данных, а также способов вычисления средних значений, дисперсий и распространения погрешностей.

1. Введение
2. Случайные величины и их характеристики
3. Основные распределения
4. Распространение погрешностей
5. Оценка параметров
6. Проверка гипотез и доверительные интервалы
7. Особенности данных с большими и малыми выборками

1 Введение

- Результат астрономического измерения является случайной величиной. Даже при неизменном объекте и условиях наблюдения повторные экспозиции дают различные значения, поскольку фотоны приходят случайно, электроника вносит флуктуации, а атмосфера изменяется.
- Статистика в астрономии имеет особенности. Выборки часто малы, данные неоднородны, объёмы измерений могут быть огромными, а наблюдаемые величины могут быть ограничены порогом чувствительности.

2 Случайные величины и их характеристики

Случайная величина X описывается функцией распределения $F(x) = P(X \le x)$ или плотностью вероятности $p(x)$.

Математическое ожидание

Определяется как $\mu = E[X] = \int x\, p(x)\, dx$ и указывает центр распределения.

Дисперсия

Определяется как $\sigma^2 = E[(X - \mu)^2]$ и измеряет разброс значений.

Стандартное отклонение

Квадратный корень из дисперсии, выраженный в единицах измеряемой величины.

Асимметрия и эксцесс

Третий и четвёртый нормированные моменты показывают форму распределения.

Ковариация

Определяется как $\mathrm{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$ и показывает связь величин.

2 Случайные величины и их характеристики (продолжение)

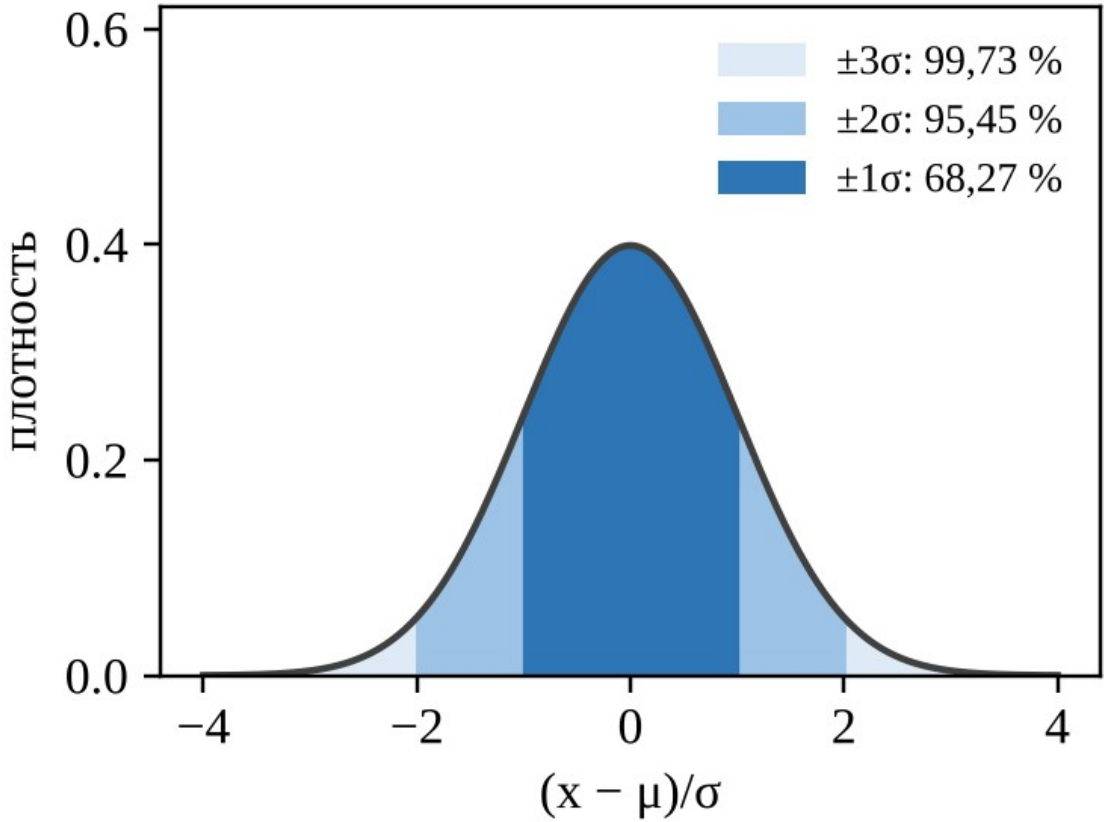
- По выборке $x_1, \ldots, x_N$ вычисляются оценки:

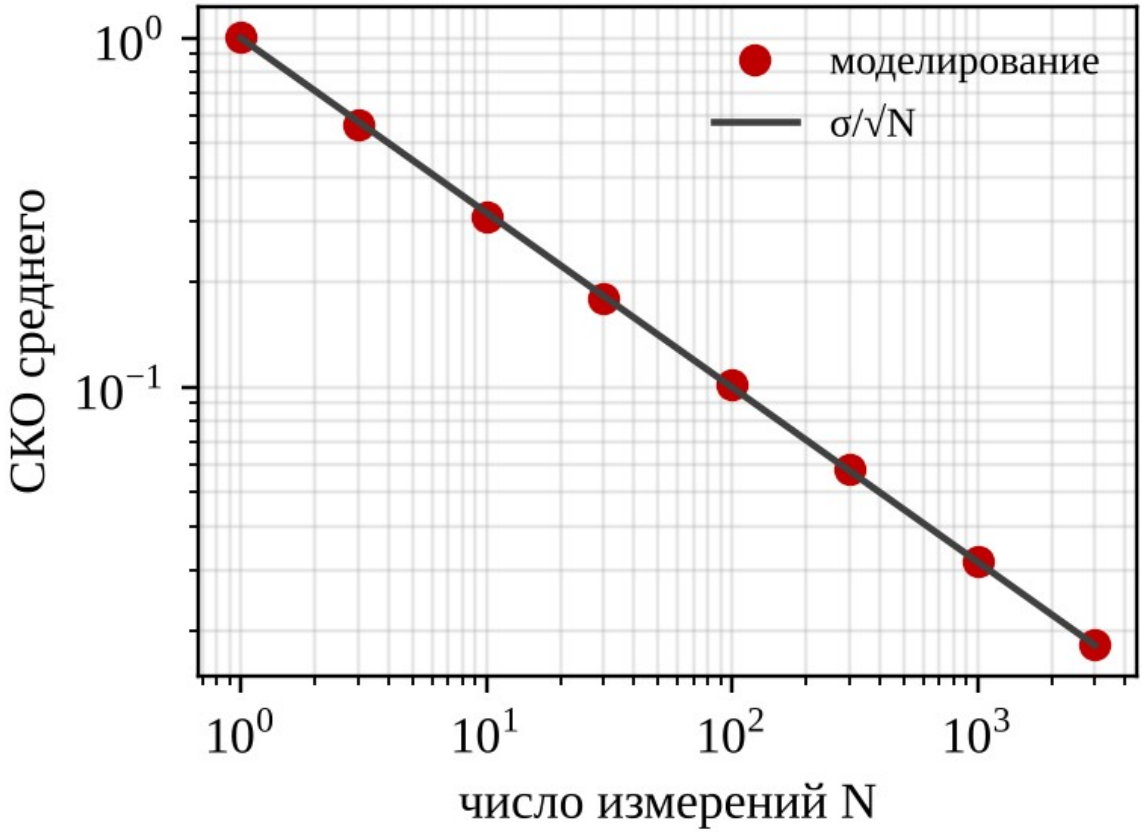
$$\bar{x} = (1/N)\,\Sigma\, x_i, \quad s^2 = (1/(N-1))\,\Sigma\,(x_i - \bar{x})^2.$$

- Знаменатель N−1 (поправка Бесселя) учитывает использование выборочного среднего вместо неизвестного математического ожидания.

$$\sigma_{\bar{x}} = \sigma / \sqrt{N}.$$

2 Случайные величины и их характеристики (продолжение)





Нормальное распределение и уменьшение ошибки среднего с увеличением числа измерений

3 Основные распределения

Нормальное (гауссово) распределение

Определяется средним и дисперсией. Его особое значение обусловлено центральной предельной теоремой: сумма большого числа независимых величин с конечной дисперсией приближается к нормальному закону.

Распределение Пуассона

Описывает число независимых редких событий, например фотонов, зарегистрированных за время экспозиции.

Хи-квадрат

Сумма квадратов независимых нормальных величин; применяется для проверки согласия модели с данными.

В астрономических данных также встречаются цензурированные данные: например, для некоторых объектов указан только верхний предел потока.

3 Основные распределения (продолжение)

Распределение Стьюдента

Используется, когда дисперсия неизвестна и оценена по малой выборке; его хвосты тяжелее, чем у нормального распределения.

Распределения с тяжёлыми хвостами

Например, смесь нормальных и распределение Лапласа, которые хорошо описывают данные с выбросами.

В астрономических данных также встречаются цензурированные данные: например, для некоторых объектов указан только верхний предел потока.

4 Распространение погрешностей

- Если величина $y = f(x_1, \ldots, x_n)$ зависит от независимых измеряемых величин, то в линейном приближении:

$$\sigma_y^2 = \sum (\partial f/\partial x_i)^2 \, \sigma_i^2.$$

- Если величины коррелированы, необходимо добавить слагаемые вида $2(\partial f/\partial x_i)(\partial f/\partial x_j)\,\mathrm{Cov}(x_i, x_j)$.
- для суммы и разности абсолютные дисперсии складываются;
- для произведения и частного складываются квадраты относительных погрешностей;
- для величины $m = -2{,}5 \lg F$ погрешность составляет $\sigma_m = 1{,}0857\, \sigma_F/F$.

4 Распространение погрешностей (продолжение)

- Важно, что линейное приближение применимо лишь при малых относительных ошибках.

5 Оценка параметров

- Метод максимального правдоподобия выбирает значения параметров, при которых вероятность наблюдаемых данных наибольшая.

$$\chi^2 = \sum (y_i - m(x_i; \theta))^2 / \sigma_i^2.$$

- Взвешенное среднее имеет весовые коэффициенты $w_i = 1/\sigma_i^2$, и его погрешность равна:

$$\bar{x}_w = \sum w_i x_i / \sum w_i, \quad \sigma_w = (\sum w_i)^{-1/2}.$$

- Медиана устойчива к выбросам, но для нормальных данных её дисперсия в π/2 раз больше дисперсии среднего.

5 Оценка параметров (продолжение)

- В байесовском подходе параметры рассматриваются как случайные величины, а результат выражается апостериорным распределением:

$$p(\theta \mid d) \propto p(d \mid \theta)\, p(\theta).$$

- Такой подход является естественным для задач с небольшим количеством фотонов, при наличии физических ограничений (поток не может быть отрицательным) и для объединения информации из разных наблюдений.

5 Оценки центра и разброса при наличии выбросов

Листинг 2.1.

```
import numpy as np

rng = np.random.default_rng(2)
x = rng.normal(1000.0, 30.0, 200)
y = x.copy()
y[:4] += 3000                    # космические частицы

def mad_sigma(a):
    return 1.4826 * np.median(np.abs(a - np.median(a)))

for name, a in (("Чистые данные", x), ("С выбросами", y)):
    print(f"{name:14} ср. {a.mean():7.1f} мед. {np.median(a):7.1f} "
          f"СКО {a.std(ddof=1):6.1f} MAD {mad_sigma(a):5.1f}")

sig = np.array([10.0, 20.0, 50.0])
val = np.array([100.0, 104.0, 130.0])
w = 1 / sig**2
print(f"Взвешенное среднее: {(w * val).sum() / w.sum():.2f} "
      f"+- {w.sum()**-0.5:.2f}")
print(f"Обычное среднее: {val.mean():.2f}")
```

5 Проверка распространения погрешностей

Листинг 2.2.

```
F0, s0 = 1000.0, 100.0
F = rng.normal(F0, s0, 500000)
m = -2.5 * np.log10(F[F > 0])
print(f"Монте-Карло, S/N=10: sigma_m = {m.std():.4f}")
print(f"Формула:             sigma_m = {1.0857 * s0 / F0:.4f}")

F2 = rng.normal(100.0, 50.0, 500000)
good = F2 > 0
m2 = -2.5 * np.log10(F2[good])
print(f"Монте-Карло, S/N=2: sigma_m = {m2.std():.4f}, "
      f"формула {1.0857 * 50 / 100:.4f}")
print(f"Отрицательных значений: {(~good).mean():.3f}")
```

```
Чистые данные  ср.   999.8 мед.   999.2 СКО   28.9 MAD  31.4
С выбросами    ср.  1059.8 мед.  1001.4 СКО  418.7 MAD  32.1
Взвешенное среднее: 101.71 +- 8.80
Обычное среднее: 111.33
```

6 Проверка гипотез и доверительные интервалы

- Доверительный интервал указывает диапазон, накрывающий истинное значение с заданной вероятностью при многократном повторении эксперимента.
- Проверка гипотез даёт p-значение – вероятность получить столь же экстремальный результат при условии справедливости нулевой гипотезы.
- В настоящее время для оценки погрешностей широко применяются методы повторной выборки.

7 Особенности данных с большими и малыми выборками

- В современных каталогах содержатся миллиарды измерений, и даже незначительное различие может иметь формально высокую статистическую значимость.
- При малых выборках справедлива обратная проблема: оценка дисперсии неустойчива.
- Важным случаем является наличие цензурирования: когда объект не обнаружен, его значение ограничено сверху.
- В задачах с множеством параметров следует использовать полную ковариационную матрицу.
- Перед применением любой формулы рекомендуется проверить её на модельных данных: задать известное значение параметра, добавить шум и повторить вычисление много раз.

8 Разбор примера

1. Пусть измерены пять значений потока: 98, 102, 101, 99, 100.
2. Среднее равно 100.
3. Отклонения равны −2, 2, 1, −1, 0, а сумма их квадратов – 10.
4. Выборочная дисперсия $s^2 = 10/4 = 2{,}5$, $s \approx 1{,}58$.
5. Стандартная ошибка среднего равна $1{,}58/\sqrt{5} \approx 0{,}71$.
6. Следовательно, результат записывается как $100{,}0 \pm 0{,}7$.
7. Если поток $F = 1000 \pm 50$, то относительная ошибка равна 0,05, а погрешность величины составляет $\sigma_m = 1{,}0857 \cdot 0{,}05 \approx 0{,}054^m$.

8 Разбор примера (продолжение)

8 Для отношения двух потоков с ошибками по 2 % относительная ошибка равна $\sqrt{(0{,}02^2 + 0{,}02^2)} \approx 2{,}8\ \%$.

9 Вопросы для обсуждения на занятии

Выбор оценки

В каких случаях следует отдать предпочтение медиане, а в каких – взвешенному среднему?

Погрешности

Что можно сказать о данных, если приведённое значение χ^2 составляет 0,2 или 5?

Интерпретация

Почему p-значение часто интерпретируют неверно?

Контрольные вопросы

1. Что такое случайная величина? Приведите примеры из наблюдательной астрономии.
2. Запишите определения среднего, дисперсии и стандартной ошибки среднего.
3. Какие распределения используются в астрономических данных?
4. Запишите формулу распространения погрешностей. Когда она неприменима?
5. Как связаны погрешности потока и звёздной величины?
6. В чём состоит сущность метода максимального правдоподобия?
7. Какие устойчивые оценки применяются при наличии выбросов?
8. Чем доверительный интервал отличается от байесовского достоверного интервала?

Список литературы

1. Bevington P.R., Robinson D.K. Data Reduction and Error Analysis for the Physical Sciences. – 3rd ed. – New York: McGraw-Hill, 2003.

2. Wall J.V., Jenkins C.R. Practical Statistics for Astronomers. – 2nd ed. – Cambridge: Cambridge University Press, 2012.

3. Ivezić Ž., Connolly A.J., VanderPlas J.T., Gray A. Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data. – Updated ed. – Princeton: Princeton University Press, 2020.

4. Hogg D.W., Bovy J., Lang D. Data analysis recipes: Fitting a model to data. – arXiv:1008.4686, 2010.

5. Data, Noise and Signal (for Astronomers): lecture notes. – Caltech, с. 2016. [источник указан в силлабусе]