Отношение сигнал/шум в астрофизических наблюдениях

Лекция 9

Обнаружение слабых астрофизических сигналов. Порог обнаружения, предельная звёздная величина и статистическая значимость

КазНУ имени аль-Фараби

Цель и план лекции

Цель

Изучение статистических критериев обнаружения слабых источников, понятий порога обнаружения и предельной звёздной величины, а также способов корректной оценки значимости при множественных проверках.

1. Введение
2. Проверка гипотез при обнаружении
3. Значимость и множественные проверки
4. Обнаружение при пуассоновском шуме
5. Предельная звёздная величина и полнота
6. Согласованная фильтрация изображений
7. Заключительные замечания
8. Байесовский подход к обнаружению

1 Введение

- Обнаружение слабого источника – это принятие решения в условиях неопределённости.

2 Проверка гипотез при обнаружении

Формально рассматриваются две гипотезы: H_0 – наблюдается только шум, H_1 – присутствует сигнал.

Ошибка первого рода (ложная тревога)

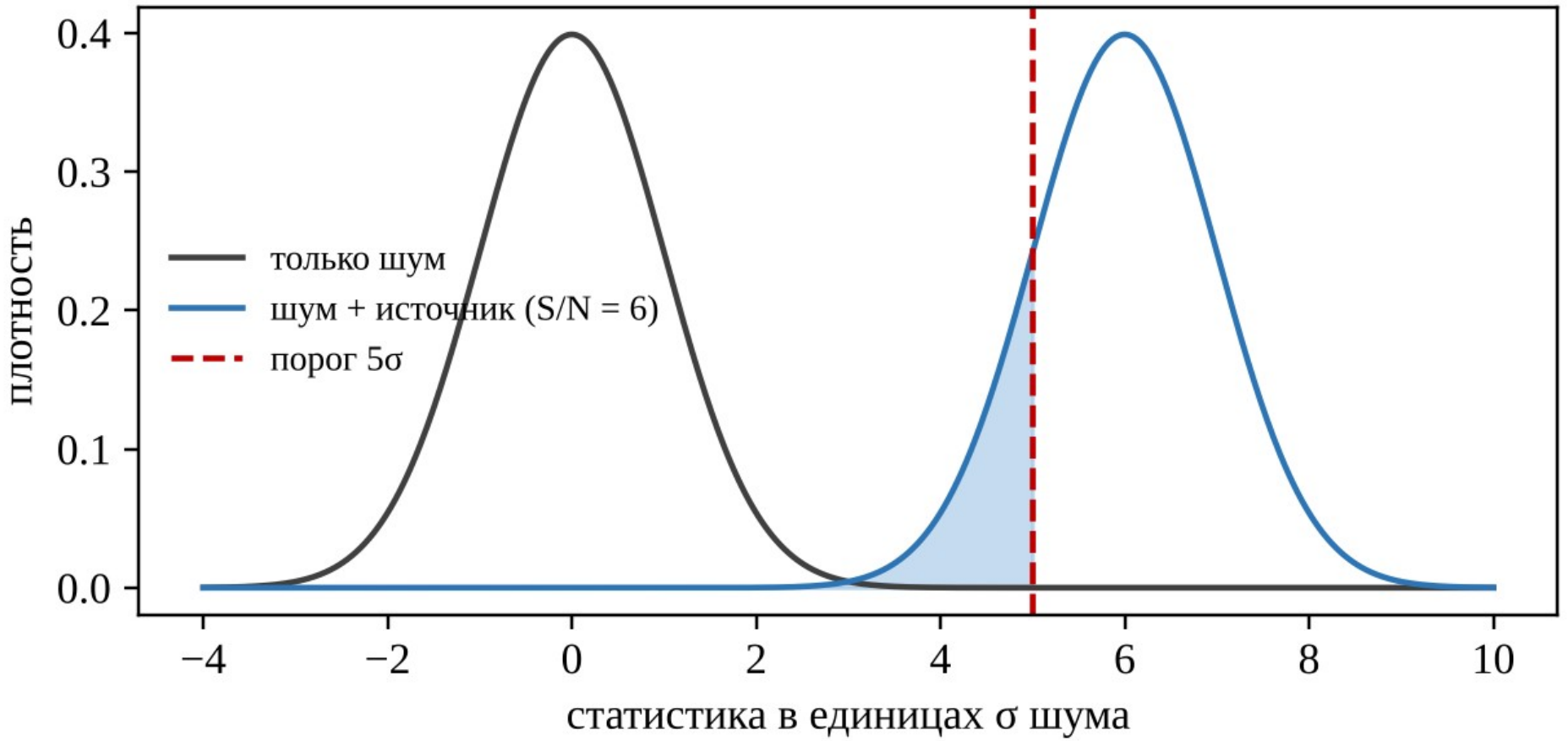
Принят сигнал, хотя его нет. Вероятность этого обозначается α.

Ошибка второго рода (пропуск)

Сигнал существует, но не обнаружен. Вероятность этого обозначается β, а 1 − β называется мощностью, или полнотой обнаружения.

Лемма Неймана–Пирсона показывает, что наибольшую мощность при заданном α обеспечивает отношение правдоподобия.

2 Проверка гипотез при обнаружении (продолжение)



Распределения статистики при отсутствии и наличии сигнала и порог 5σ

2 Проверка гипотез при обнаружении (продолжение)

- Если порог равен истинному значению S/N, полнота равна 50 %, поскольку шум с равной вероятностью увеличивает и уменьшает значение.

3 Значимость и множественные проверки

- Вероятность случайного превышения одним измерением уровня $k\sigma$ (односторонняя проверка) в нормальном приближении:
- 3σ – $1{,}35\cdot10^{-3}$;
- 5σ – $2{,}87\cdot10^{-7}$.
- Поскольку изображение содержит миллионы независимых элементов, ожидаемое число ложных срабатываний растёт пропорционально количеству проверок.
- Коррекция на количество проверок выполняется поправкой Бонферрони (α/N), а для каталогов широко применяется метод контроля доли ложных открытий (FDR, Бенджамини и Хохберг, 1995).

4 Обнаружение при пуассоновском шуме

- При малых числах отсчётов значимость обнаружения вычисляется по биномиальному распределению или формуле Ли и Ма (1983).

$$S = \sqrt{2}\,\{\, N_{on}\,\ln[\,(1+\alpha)/\alpha\cdot N_{on}/(N_{on}+N_{off})\,] + N_{off}\,\ln[\,(1+\alpha)\,N_{off}/(N_{on}+N_{off})\,]\,\}^{1/2}.$$

- Формула широко применяется в рентгеновской и гамма-астрономии, поскольку учитывает погрешность фона.

5 Предельная звёздная величина и полнота

- Предельная звёздная величина – это наибольшая величина, при которой объект обнаруживается при заданном S/N. Для фонового ограничения имеем:

$$m_{lim} = ZP - 2{,}5\,\lg(\,k\,\sigma\,\sqrt{(n_{eff})}\,/\,t\,),$$

- где σ – шум на пиксель за экспозицию t, n_{eff} – эффективное число пикселей, k – порог.

  Увеличение экспозиции в 4 раза улучшает предел примерно на $0{,}75^{m}$, поскольку S/N растёт в 2 раза.

- Полнота каталога определяется по искусственным звёздам: их вставляют в изображения, повторяют обработку и определяют долю обнаружений как функцию величины.

5 Предельная звёздная величина и полнота (продолжение)

- В статистических выборках также следует учитывать смещения отбора: смещение Малмквиста (в ограниченных по потоку выборках обнаруживаются в основном яркие объекты), смещение Эддингтона (случайные ошибки переносят большее число объектов из области слабых в более яркие, если их больше).

6 Согласованная фильтрация изображений

- Для обнаружения точечного источника изображение сворачивают с PSF. Статистика на каждом положении является оптимальной оценкой потока, делённой на её погрешность.

6 Значимость и ожидаемое число ложных источников

Листинг 9.1.

```
import numpy as np
from scipy import stats

Ntr = 4096**2 / 10
print(" k   p (односторонняя)   ложных на кадр")
for k in (3, 4, 5, 6):
    p = stats.norm.sf(k)
    print(f"{k:2} {p:18.3e} {p * Ntr:14.2f}")

for sn in (5, 6, 6.28, 7):
    print(f"S/N={sn}: полнота при пороге 5: "
          f"{stats.norm.sf(5 - sn):.3f}")

def li_ma(on, off, a):
    A = on * np.log((1 + a) / a * on / (on + off))
    B = off * np.log((1 + a) * off / (on + off))
    return np.sqrt(2 * (A + B))

print(f"Ли-Ма: on=50, off=100, a=0.25: S = {li_ma(50, 100, 0.25):.2f}")
print(f"Простая оценка: {(50 - 25) / np.sqrt(50 + 0.25**2 * 100):.2f}")
```

6 Результат: Значимость и ожидаемое число ложных источников

```
 k   p (односторонняя)   ложных на кадр
 3          1.350e-03         2264.75
 4          3.167e-05           53.14
 5          2.867e-07            0.48
 6          9.866e-10            0.00
S/N=5: полнота при пороге 5: 0.500
S/N=6: полнота при пороге 5: 0.841
S/N=6.28: полнота при пороге 5: 0.900
S/N=7: полнота при пороге 5: 0.977
Ли-Ма: on=50, off=100, a=0.25: S = 3.82
Простая оценка: 3.33
```

6 Предельная звёздная величина

Листинг 9.2.

```
ZP, nef, k = 25.0, 40, 5
sky, RN = 50.0, 4.0            # e-/c/пкс, e-
print("  t, c  m_lim")
for t in (10, 40, 160, 640):
    sig = np.sqrt(sky * t + RN**2)
    N = k * sig * np.sqrt(nef)
    print(f"{t:6} {ZP - 2.5 * np.log10(N / t):6.2f}")
```

```
  t, c  m_lim
    10  20.36
    40  21.12
   160  21.88
   640  22.63
```

7 Заключительные замечания

- Значимость нельзя определять после просмотра данных и выбора наиболее привлекательного участка.

8 Байесовский подход к обнаружению

- Частотный подход отвечает на вопрос о вероятности данных при отсутствии сигнала, тогда как исследователя обычно интересует вероятность наличия сигнала при имеющихся данных.

$$B_{10} = p(d \mid H_1) \,/\, p(d \mid H_0).$$

- Значение учитывает всю функцию правдоподобия, в том числе априорное распределение параметров, и естественно наказывает слишком сложные модели.

- Верхние пределы также требуют внимания. При необнаружении источника следует приводить ограничения на поток с указанием доверительного уровня, а не только значение S/N. Метод Фельдмана–Казинса автоматически обеспечивает переход от двустороннего интервала к верхнему пределу.

8 Байесовский подход к обнаружению (продолжение)

- Важным вопросом является учёт негауссовых хвостов шума. Значения вероятностей, связанные с уровнем 5σ, основаны на нормальном распределении.
- Значимость совместного обнаружения в нескольких фильтрах можно увеличить с помощью объединённого кадра: источники ищут на сумме значений $(S/N)^2$ по всем фильтрам.

9 Разбор примера

1. Найдём полноту при пороге 5σ и S/N = 6.
2. Шум равен σ = 1, поэтому вероятность превысить порог равна Φ(6 − 5) = Φ(1) = 0,841.
3. Следовательно, пропускается приблизительно 16 % таких источников.
4. Пусть в области источника зарегистрировано 50 событий, а в фоновой области вчетверо большего размера – 100 событий (α = 0,25).
5. Оценка фона равна 25, избыток составляет 25.
6. Упрощённая оценка 25/√(50 + 6,25) ≈ 3,33, а формула Ли–Ма даёт S ≈ 3,8.

10 Вопросы для обсуждения на занятии

Порог

Какой порог уместен для поиска кандидатов и какой – для заявления об открытии?

Каталог

Как влияет порог на полноту и загрязнение каталога?

Смещение

Как смещение Малмквиста может исказить физические выводы?

Контрольные вопросы

1. Какие ошибки возникают при обнаружении сигнала?
2. В чём заключается лемма Неймана–Пирсона?
3. Почему при большом количестве проверок нужно применять порог 5σ?
4. Как определяется полнота обнаружения при пороге 5σ?
5. В каких случаях применяют формулу Ли–Ма?
6. Как определяется предельная звёздная величина?
7. Что такое смещение Малмквиста и смещение Эддингтона?
8. Как определить полноту каталога?

Список литературы

1. Neyman J., Pearson E.S. On the problem of the most efficient tests of statistical hypotheses // Philosophical Transactions of the Royal Society A. – 1933. – Vol. 231. – P. 289–337.
2. Benjamini Y., Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing // Journal of the Royal Statistical Society B. – 1995. – Vol. 57, No. 1. – P. 289–300.
3. Li T.-P., Ma Y.-Q. Analysis methods for results in gamma-ray astronomy // The Astrophysical Journal. – 1983. – Vol. 272. – P. 317–324.
4. Feldman G.J., Cousins R.D. Unified approach to the classical statistical analysis of small signals // Physical Review D. – 1998. – Vol. 57. – P. 3873–3889.
5. Gehrels N. Confidence limits for small numbers of events in astrophysical data // The Astrophysical Journal. – 1986. – Vol. 303. – P. 336–346.
6. Bertin E., Arnouts S. SExtractor: Software for source extraction // Astronomy and Astrophysics Supplement Series. – 1996. – Vol. 117. – P. 393–404.
7. 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.