Методы детектирования аномалий. Лекция 4: Вероятностные и статистические модели

Иван Шанин
ivan.shanin@gmail.com

ИПИ РАН

11.03.2019

Вероятность экстремальных значений

неравенство Маркова

Пусть X — случайная величина, принимающая только неотрицательные значения. Тогда для любой константы α, большей $\mathbb{E}(X)$ верно следующее:

$$P(X > \alpha) \leq \frac{1}{\alpha}\mathbb{E}(X)$$

- Данная оценка не требует знания распределения, в следствие чего считается грубой.

Доказательство:

$$\mathbb{E}(X) = \int_x xp(x)dx = \int_{0\leq x\leq\alpha} xp(x)dx + \int_{x>\alpha} xp(x)dx$$

$$\geq \int_{x>\alpha} xp(x)dx \geq \int_{x>\alpha} \alpha p(x)dx$$

■

Вероятность отклонения от матожидания

неравенство Чебышева

Пусть X — случайная величина, без ограничений. Тогда для любой константы $\alpha > 0$ верно следующее:

$$P\left(|X - \mathbb{E}(X)| > \alpha\right) \leq \frac{1}{\alpha^2}\mathbb{V}(X)$$

- Данная оценка следует из неравенства Маркова, но не требует условия неотрицательности

Доказательство:

$$|X - \mathbb{E}(X)| > \alpha \iff (X - \mathbb{E}(X))^2 > \alpha^2$$

Если взять случайную величину $Y = (X - \mathbb{E}(X))^2$ и применить к ней неравенство Маркова с константой α^2, то получается в точности требуемое, так как по определению дисперсии $\mathbb{E}(Y) = \mathbb{V}(X)$ ■

Суммы независимых случайных величин

Рассмотрим несколько примеров:

- **Отбор игроков в команду.** Есть M игроков, каждый из которых сыграл N игр, и для каждой игры известна количественная характеристика X игры каждого игрока. Тогда совокупную характеристику качества игры за N игр следует рассматривать как $X = \sum_{i=1}^{N} X_i$. Как выделить аномальных игроков?
- **Поведение покупателей.** В заданный день покупатель i приходит в магазин с вероятностью p_i, с какой вероятностью в этот день придет не менее N покупателей?
- **Контроль качества производства.** На конвейере производятся детали с вероятностью брака p. Как по N образцам определить соответствие производства норме?

Сумма бернуллиевских случайных величин

- **Поведение покупателей.** В заданный день покупатель i приходит в магазин с вероятностью p_i, с какой вероятностью в этот день придет не менее N покупателей?

неравенство Чернова

Пусть X — случайная величина, являющаяся суммой N независимых бернуллиевских случайных величин, каждая из которых принимает значение 1 с вероятностью p_i. Тогда для любой константы $\delta \in (0, 1)$, верно следующее:

$$P\left(X < (1-\delta)\cdot \mathbb{E}(X)\right) < e^{-\mathbb{E}(X)\cdot\delta^2/2}$$

$$P\left(X \geq (1+\delta)\cdot \mathbb{E}(X)\right) < e^{-\mathbb{E}(X)\cdot\delta^2/4}$$

[TODO: Добавить скетч доказательства]

Сумма независимых случайных величин

- **Отбор игроков в команду.** Есть M игроков, каждый из которых сыграл N игр, и для каждой игры известна количественная характеристика X игры каждого игрока. Тогда совокупную характеристику качества игры за N игр следует рассматривать как $X = \sum_{i=1}^{N} X_i$. Как выделить аномальных игроков?

неравенство Хефдинга

Пусть X — случайная величина, являющаяся суммой N независимых случайных величин X_i, каждая из которых каким-то образом распределена на отрезке $[l_i, u_i]$. Тогда для любой константы $\theta > 0$ верно следующее:

$$P(X - \mathbb{E}(X) > \theta) \leq e^{-\frac{2\theta^2}{\sum_{i=1}^{N}(u_i - l_i)^2}}$$

$$P(\mathbb{E}(X) - X > \theta) \leq e^{-\frac{2\theta^2}{\sum_{i=1}^{N}(u_i - l_i)^2}}$$

[TODO: Добавить скетч доказательства]

Сумма независимых одинаково распределенных случайных величин

- **Контроль качества производства.** На конвейере производятся детали с вероятностью брака p. Как по N образцам определить соответствие производства норме?

Центральная предельная теорема

Сумма N независимых и одинаково распределенных случайных величин с матожиданием μ и стандартным отклонением σ^2 при достаточно большом N сходится к случайной величине с матожиданием μN и стандартным отклонением $\sigma\sqrt{N}$

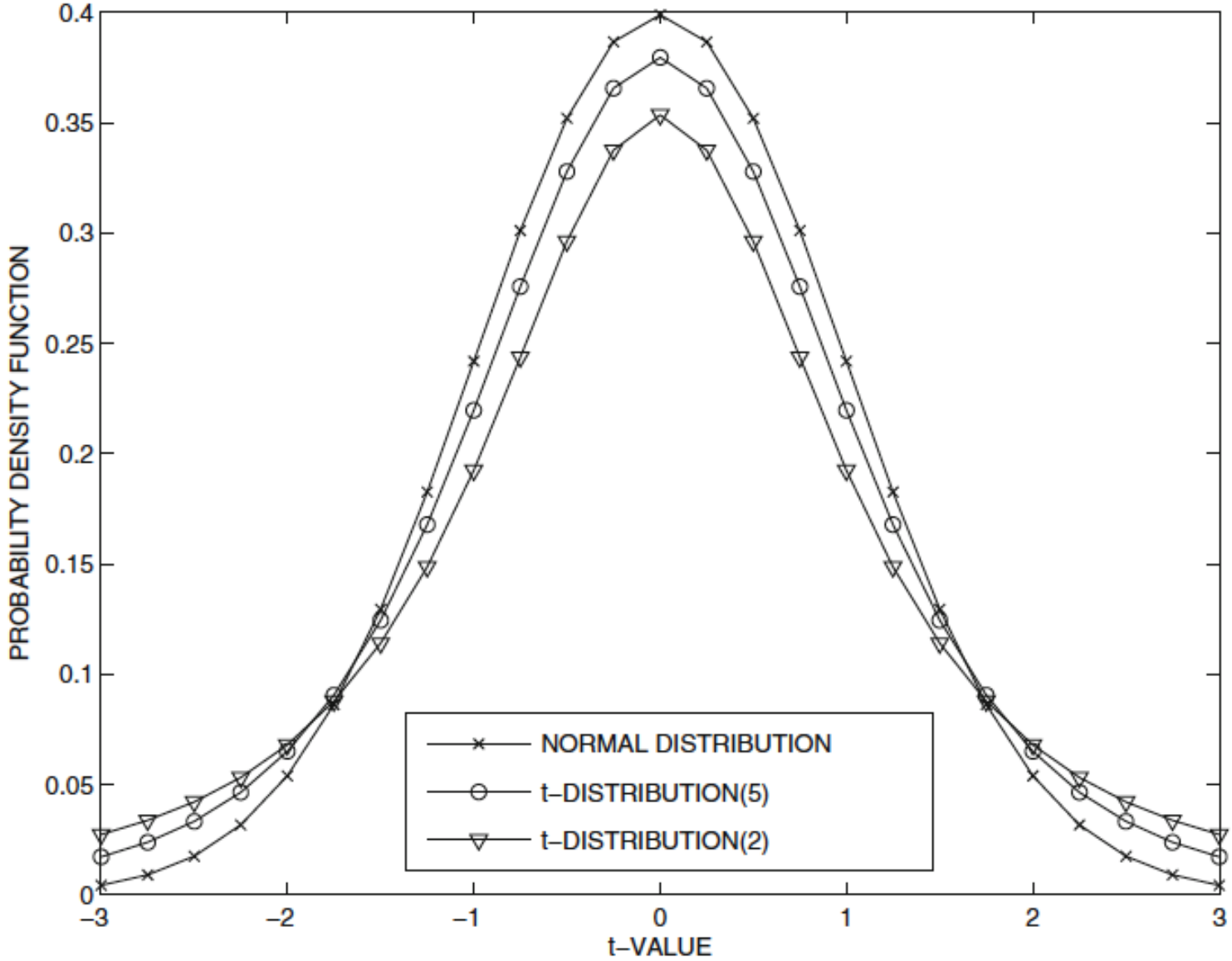
Проверка статистических гипотез

- Если исследуемое распределение является нормальным:

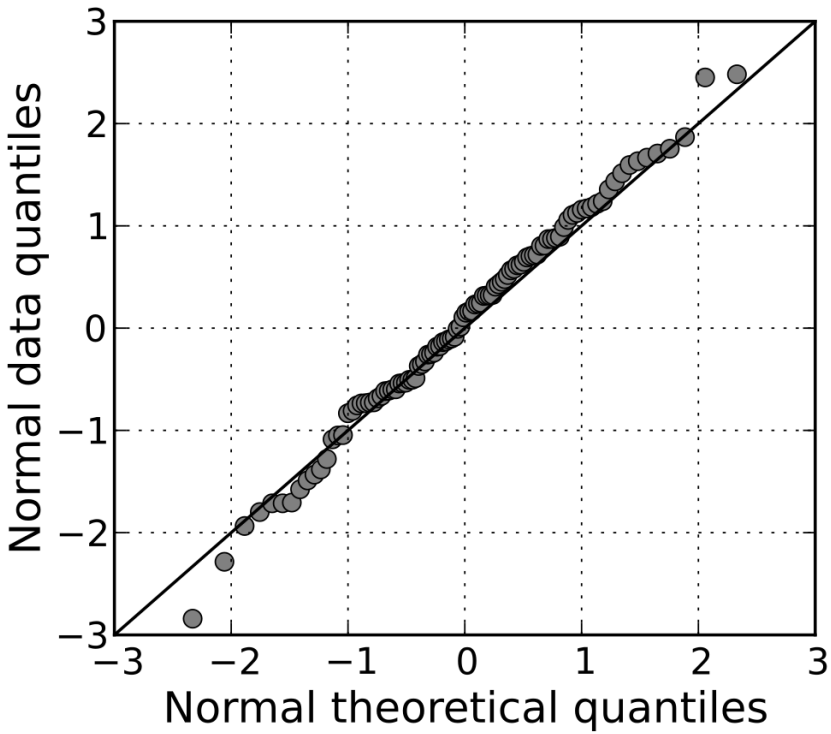
$$p(x) = \frac{1}{\sigma\sqrt{2 \cdot \pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

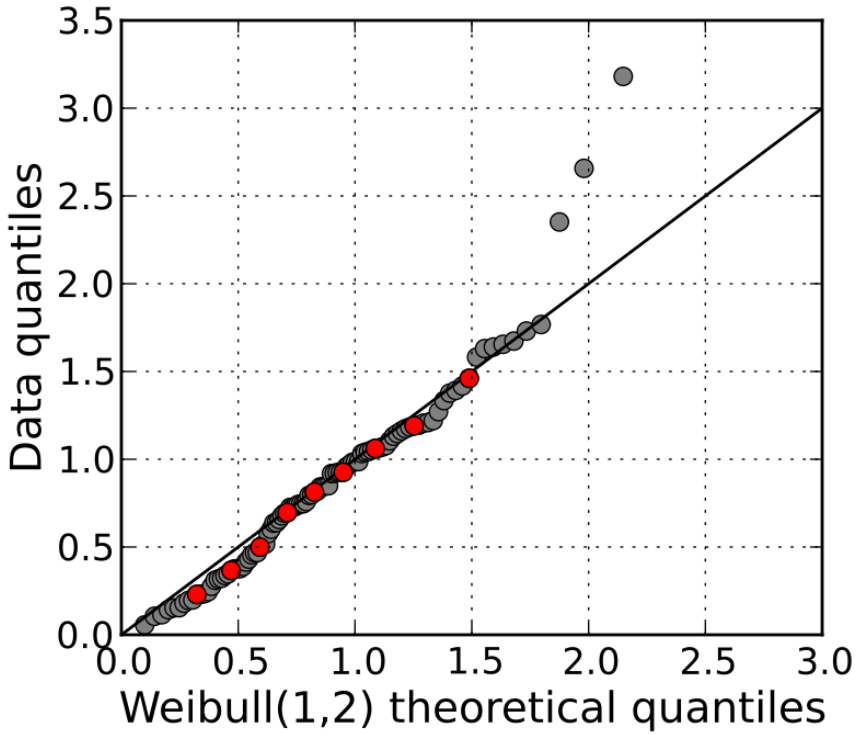
  то можно применять z-преобразование и использовать z-test.
- Если набор данных невелик настолько, то оценки матожидания и дисперсии будут некорректны. В этом случае следует применять критерий Стьюдента (t-test)
- Если интерес представляют не данные, а квадраты их отклонений, то следует использовать критерий χ^2

Распределение Стьюдента

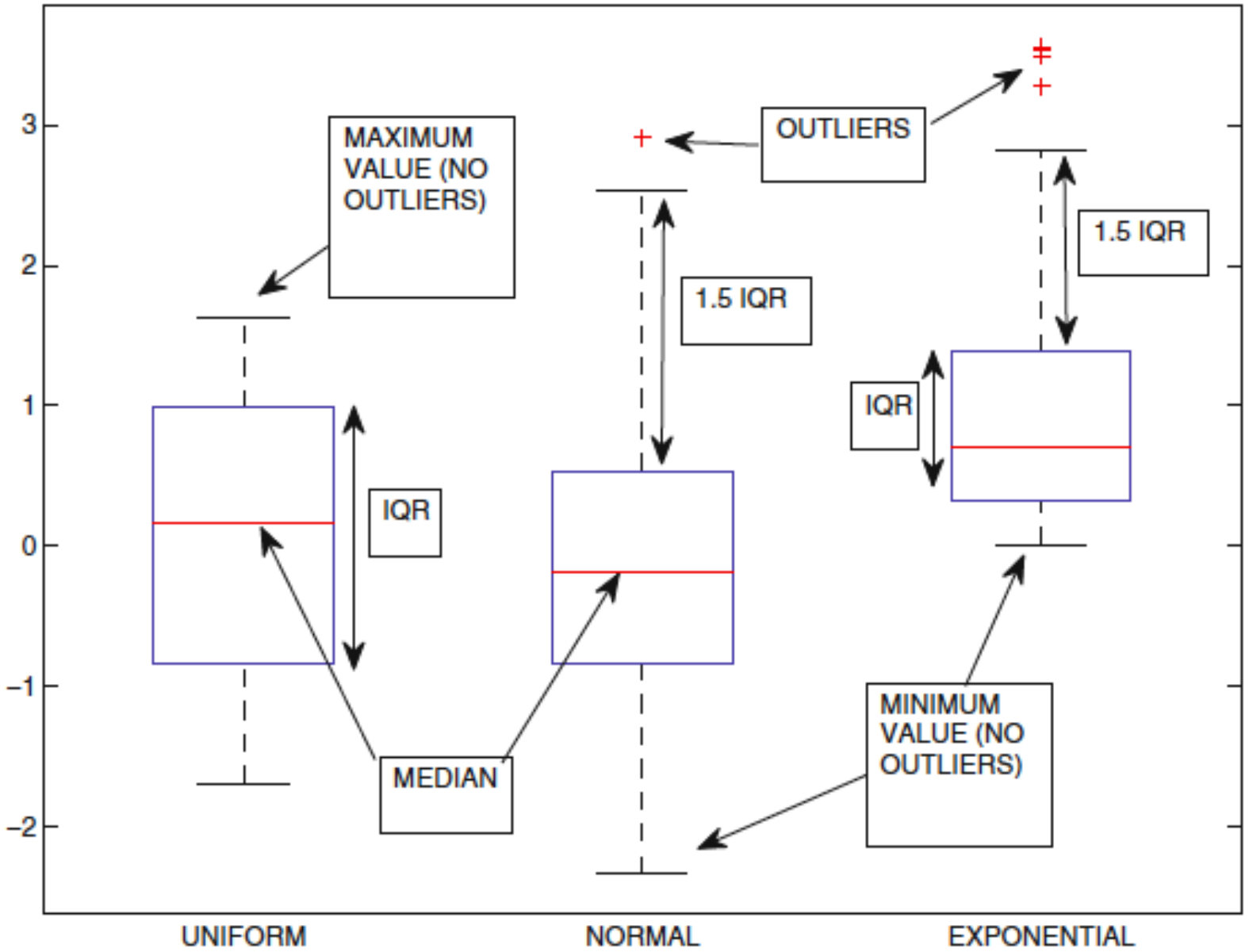


QQ-plot

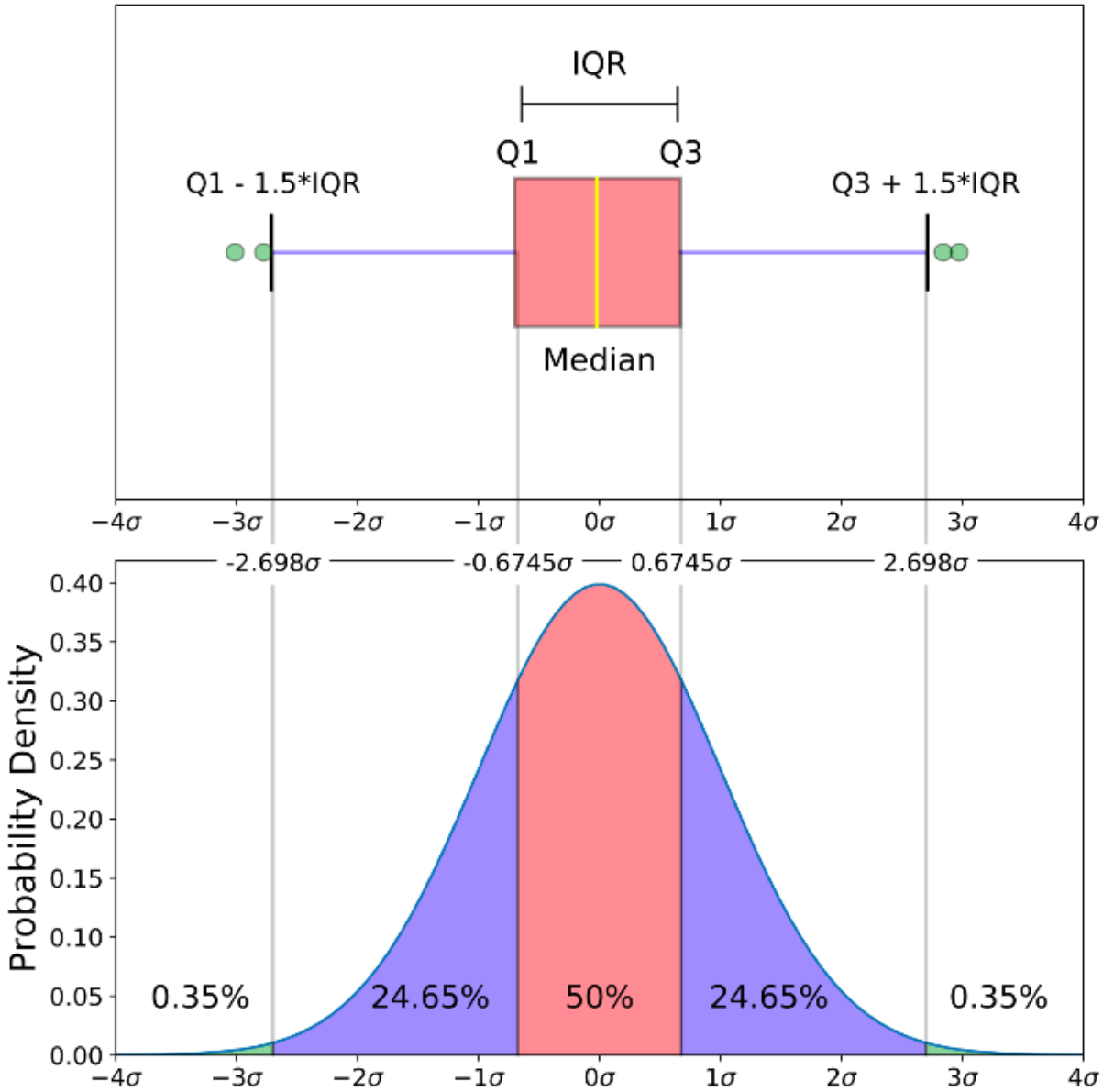

Normal data quantiles
Normal theoretical quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3


Data quantiles
Weibull(1,2) theoretical quantiles
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0

Box plot


MAXIMUM VALUE (NO OUTLIERS)
OUTLIERS
1.5 IQR
1.5 IQR
IQR
IQR
MEDIAN
MINIMUM VALUE (NO OUTLIERS)
3
2
1
0
−1
−2
UNIFORM
NORMAL
EXPONENTIAL

Box plot


IQR
Q1
Q3
Q1 - 1.5*IQR
Q3 + 1.5*IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
-2.698σ
-0.6745σ
0.6745σ
2.698σ
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.35%
24.65%
50%
24.65%
0.35%