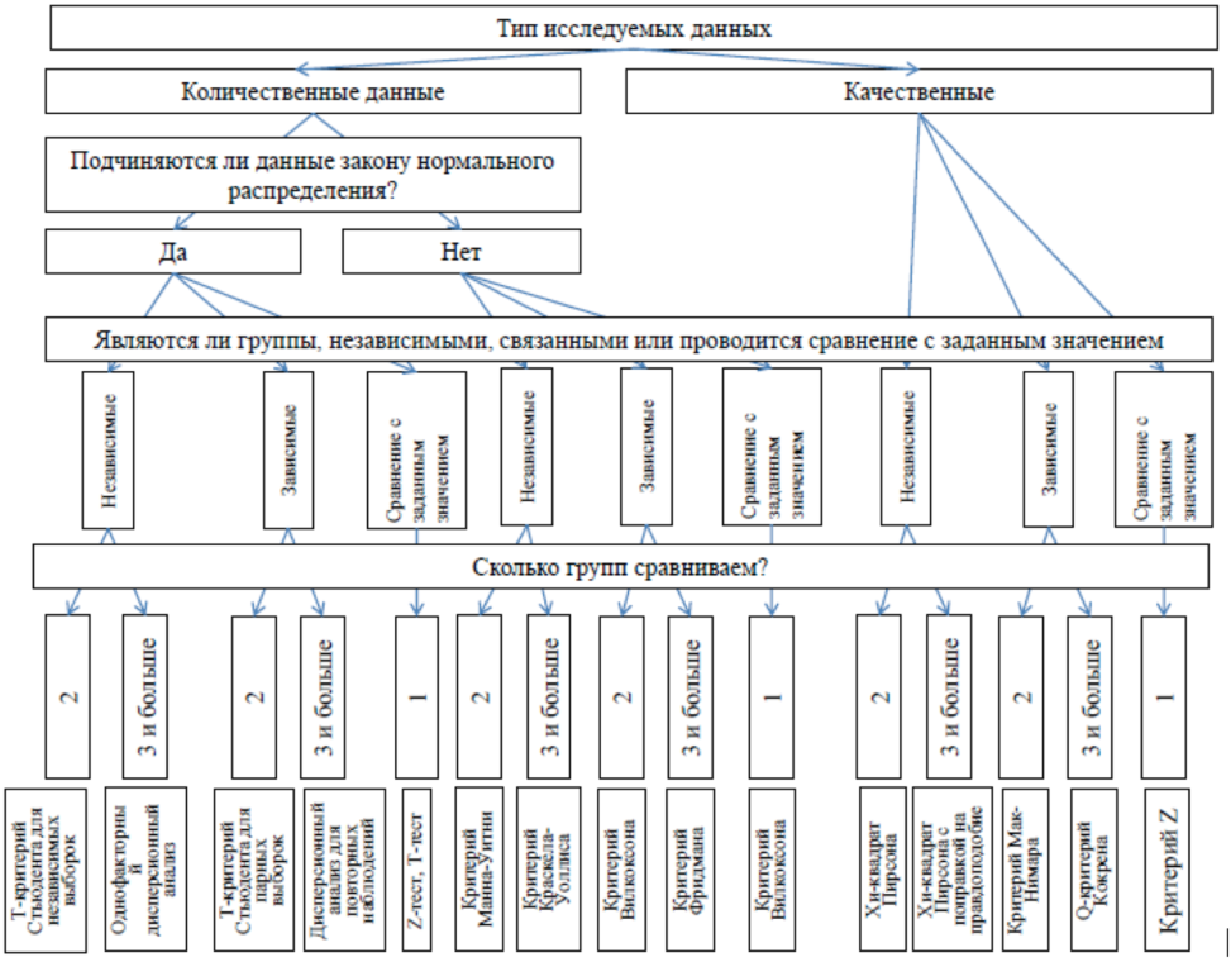


Тип исследуемых данных
Количественные данные
Качественные
Подчиняются ли данные закону нормального распределения?
Да
Нет
Являются ли группы, независимыми, связанными или проводится сравнение с заданным значением
Независимые
Зависимые
Сравнение с заданным значением
Независимые
Зависимые
Сравнение с заданным значением
Независимые
Зависимые
Сравнение с заданным значением
Сколько групп сравниваем?
2
3 и больше
2
3 и больше
1
2
3 и больше
2
3 и больше
1
2
3 и больше
2
3 и больше
1
Т-критерий Стьюдента для независимых выборок
Однофакторный дисперсионный анализ
Т-критерий Стьюдента для парных выборок
Дисперсионный анализ для повторных наблюдений
Z-тест, Т-тест
Критерий Манна-Уитни
Критерий Краскела-Уоллиса
Критерий Вилкоксона
Критерий Фридмана
Критерий Вилкоксона
Хи-квадрат Пирсона
Хи-квадрат Пирсона с поправкой на правдоподобие
Критерий Мак-Нимара
Q-критерий Кокрена
Критерий Z

Выборочная совокупность — множество всех объектов, отобранных случайно из генеральной совокупности для изучения.

Нулевая гипотеза (Н0)— гипотеза о сходстве

Альтернативная гипотеза, конкурирующая, (Н1)– гипотеза о различиях

Критерии согласия

Критерии согласия. Проверка предположения о том, что исследуемая случайная величина подчиняется предполагаемому закону распределения.

Тесты нормальности

1) **Графические.**
2) **Параметрические.**

t-критерий Стьюдента

Критерий Фишера

Критерий отношения правдоподобия

Критерий Романовского

3) Непараметрические.

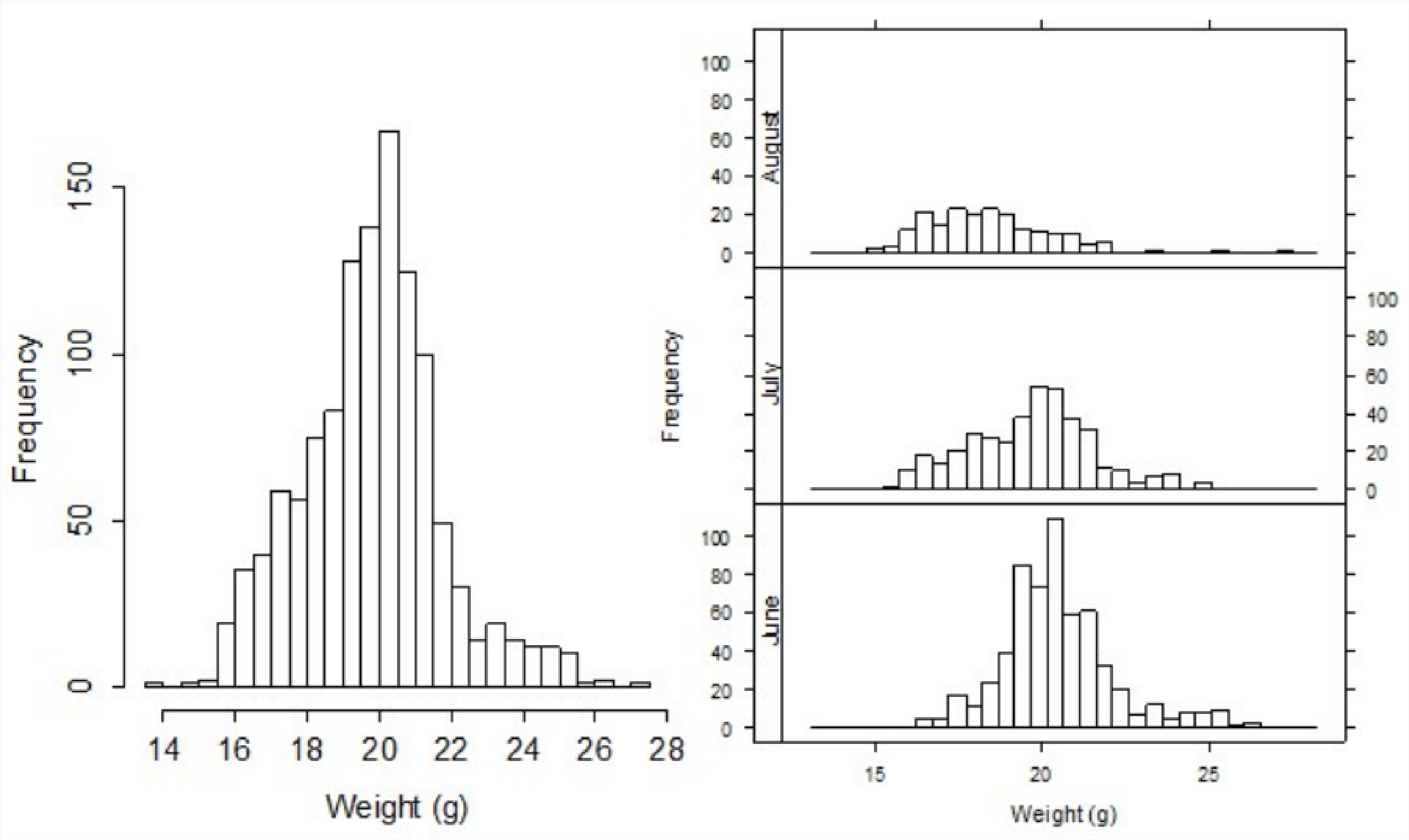
Q-критерий Розенбаума

U-критерий Манна — Уитни

Критерий Уилкоксона

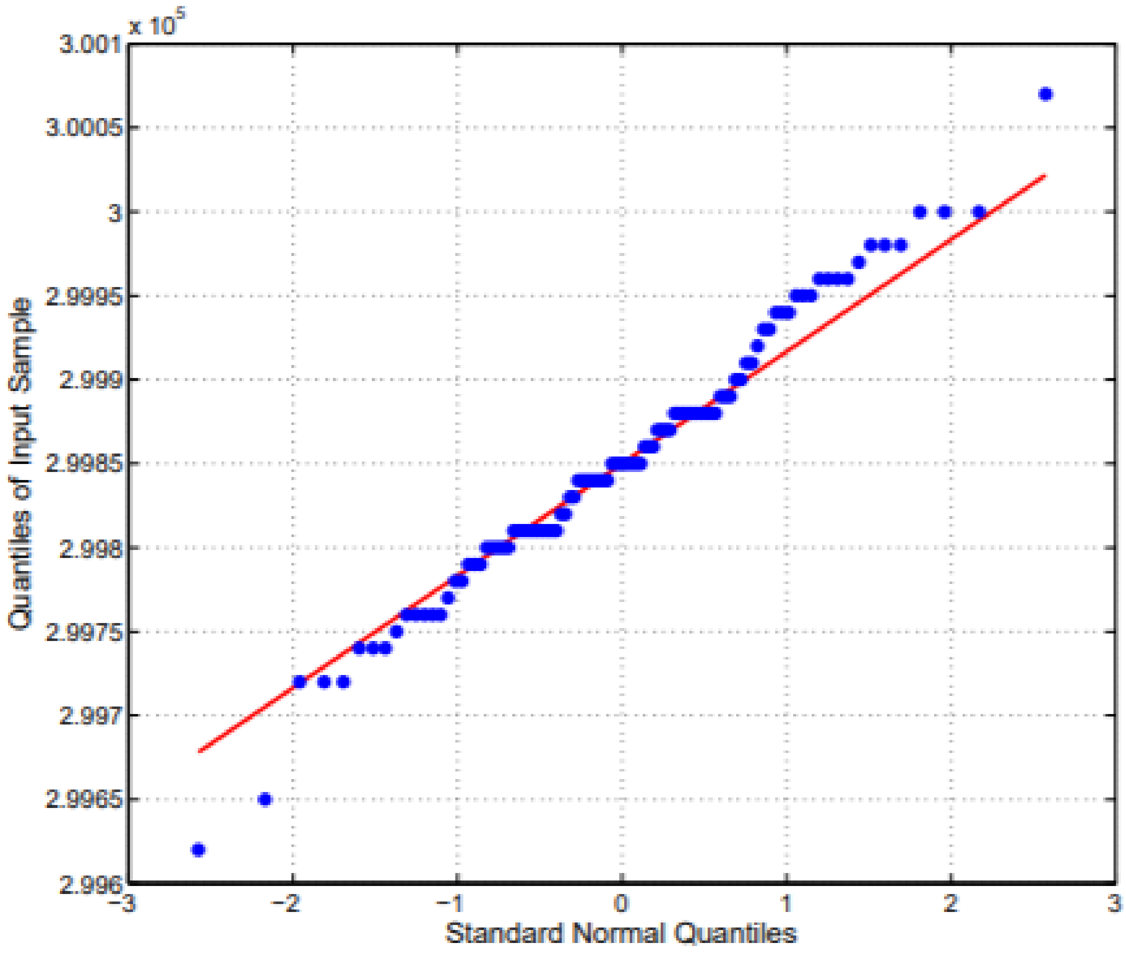
Критерий Пирсона

Критерий Колмогорова — Смирнова


Frequency
Weight (g)
August
July
June
Frequency
Weight (g)

Тест по квантильной диаграмме

Визуальный метод проверки согласия выборки и распределения — q-q plot (для нормального распределения называется также normal probability plot)



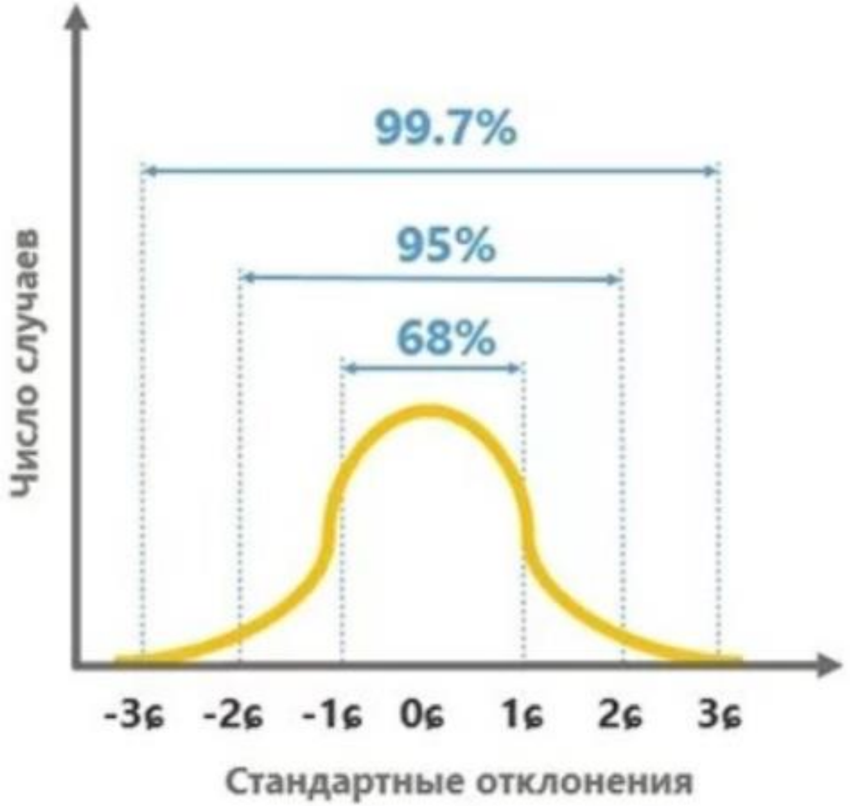
 # QQ Plot

#одинаковые "случайные" числа
 seed the random number generator
seed(1)

# generate univariate observations
data = 5 * randn(100) + 50

# q-q plot
qqplot(data, line='s')
pyplot.show()

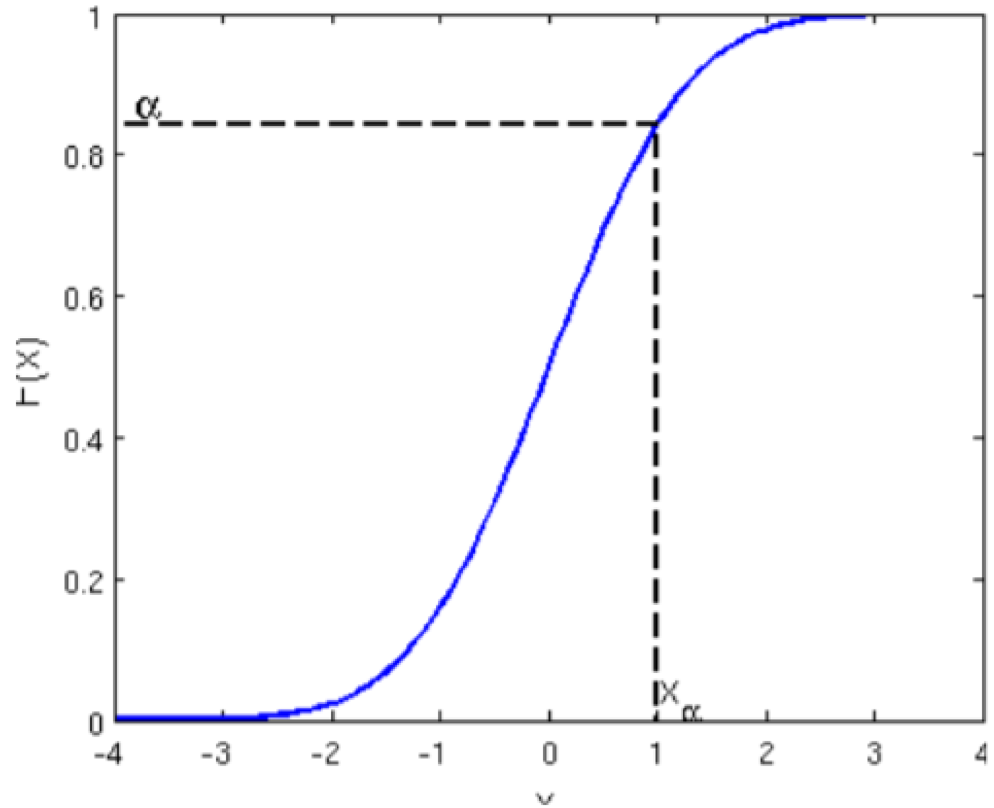
Доверительный интервал — интервал, в котором лежит 95% данных.

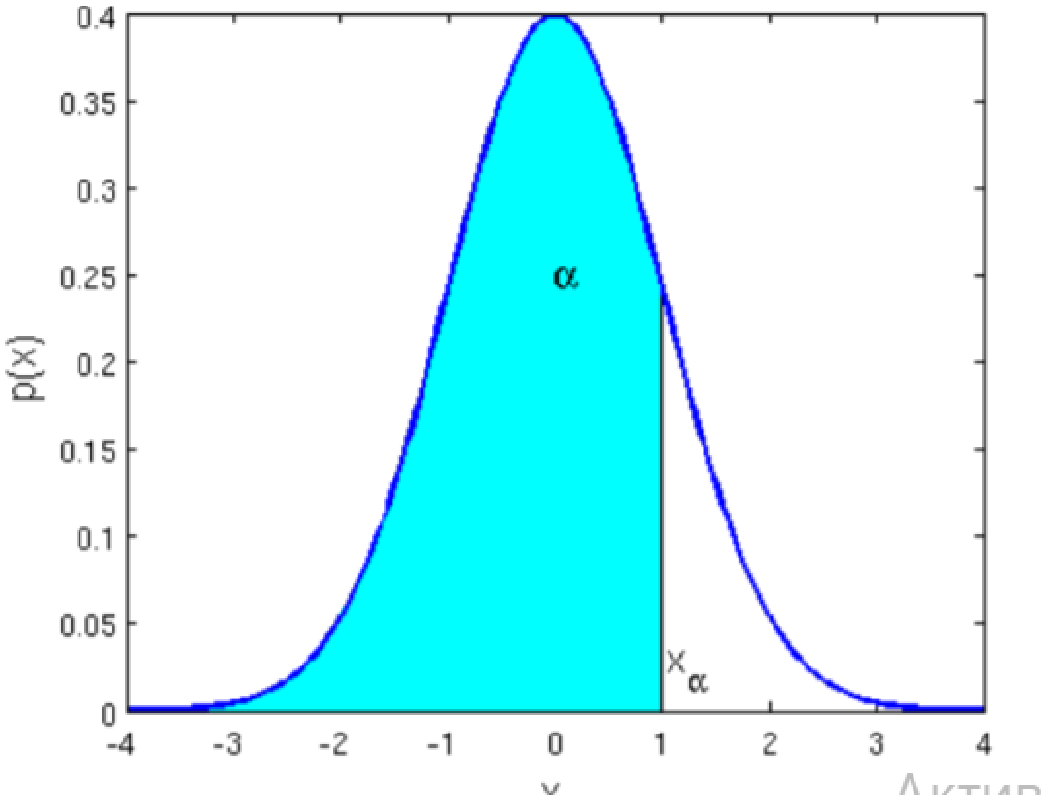


Квантили

Квантиль (α-квантиль) x_α– число, такое, что заданная случайная величина превышает его лишь с фиксированной вероятностью $(1-\alpha)$, т.е. $P(X \le x_\alpha) = \alpha$

Квантиль рассчитывается по уравнению: $F(x_\alpha) = \alpha$



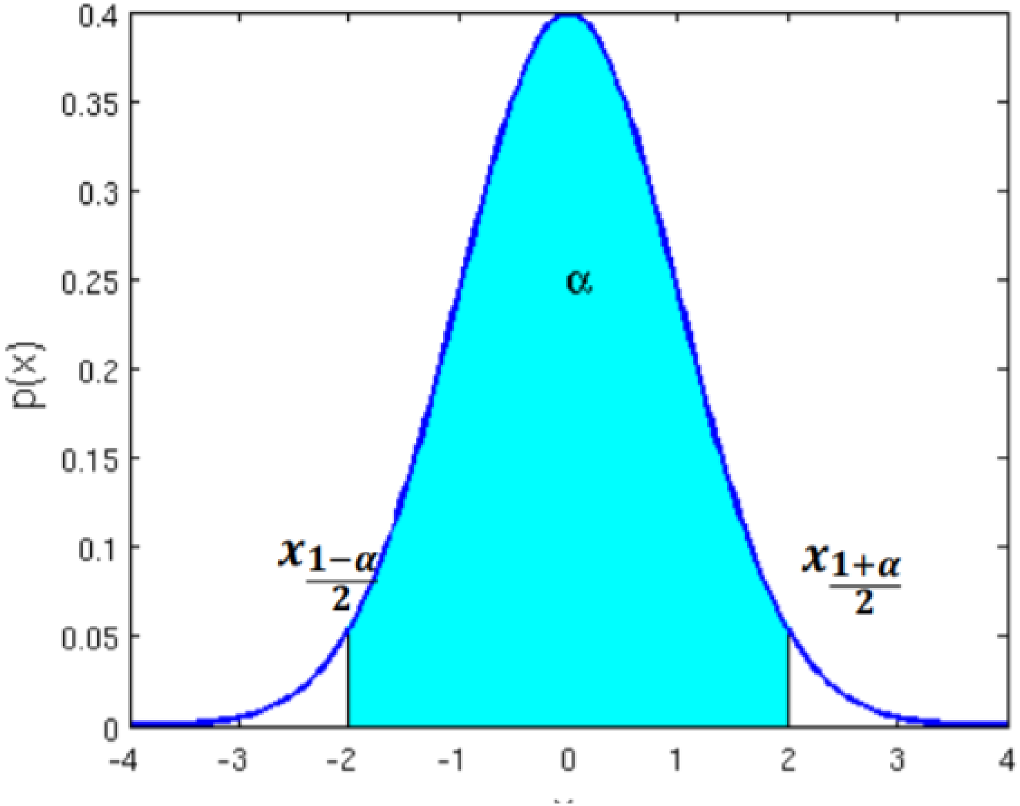


Двухсторонний квантиль

Определение

$$P\left(x_{\frac{1-\alpha}{2}} \le X \le x_{\frac{1+\alpha}{2}}\right) = \alpha$$

$$F\left(x_{\frac{1+\alpha}{2}}\right) - F\left(x_{\frac{1-\alpha}{2}}\right) = \alpha$$



Случай симметричного распределения

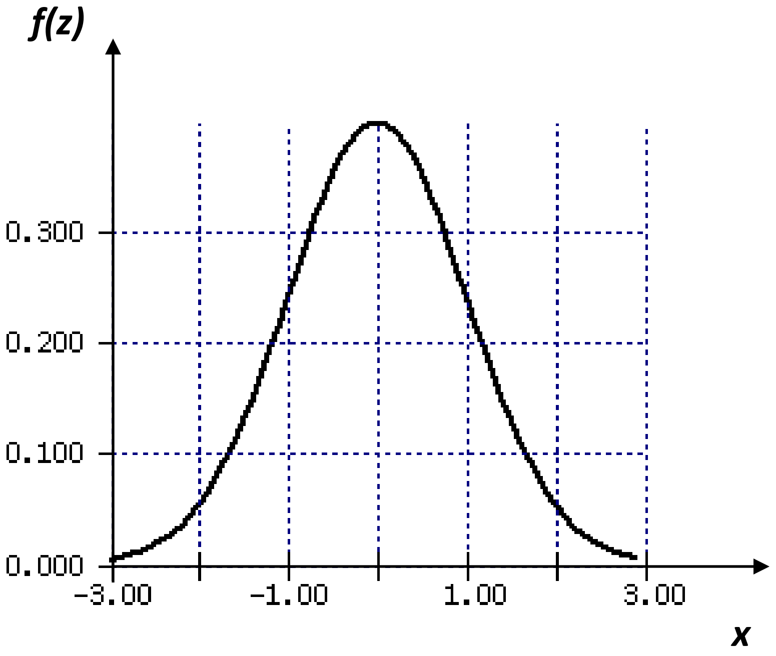
$$x_{\frac{1+\alpha}{2}} = -x_{\frac{1-\alpha}{2}}$$

Пример: $\alpha = 0.95$

$$\frac{1+\alpha}{2} = \frac{1+0.95}{2} = 0.975$$

$$\frac{1-\alpha}{2} = \frac{1-0.95}{2} = 0.025$$

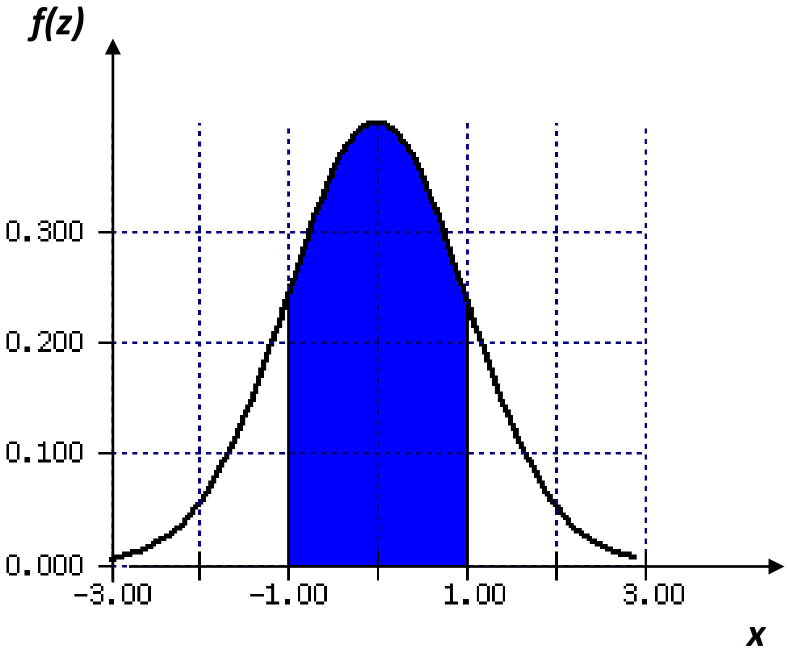
У нормального распределения есть три стандартных числа:



Стандартные распределения и их квантили

У нормального распределения есть три стандартных числа:

Вероятность попадания *x* в интервал [μ-1σ; μ+1σ] равна ≈ 68%.

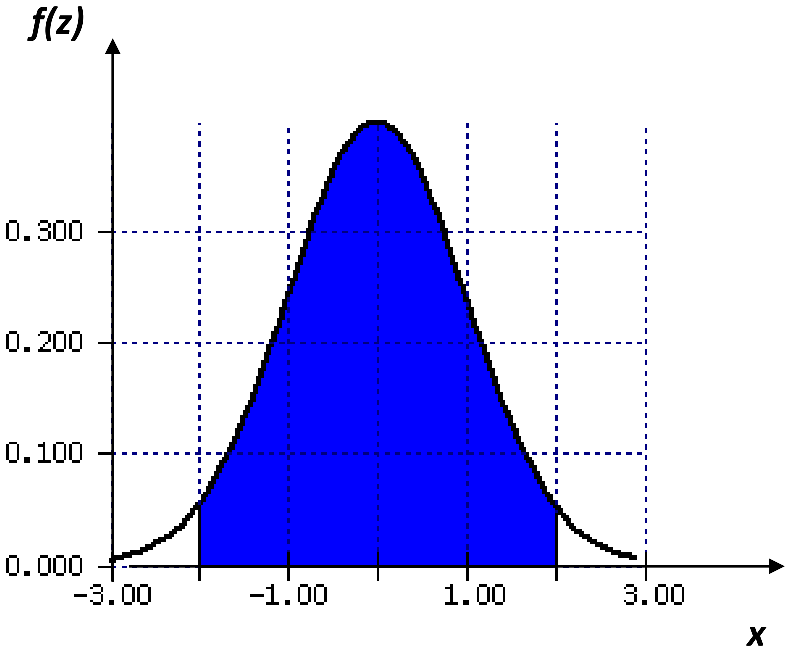


Стандартные распределения и их квантили

У нормального распределения есть три стандартных числа:

Вероятность попадания *x* в интервал [μ-1σ; μ+1σ] равна ≈ 68%.

Вероятность попадания *x* в интервал [μ-2σ; μ+2σ] равна ≈ 95%.



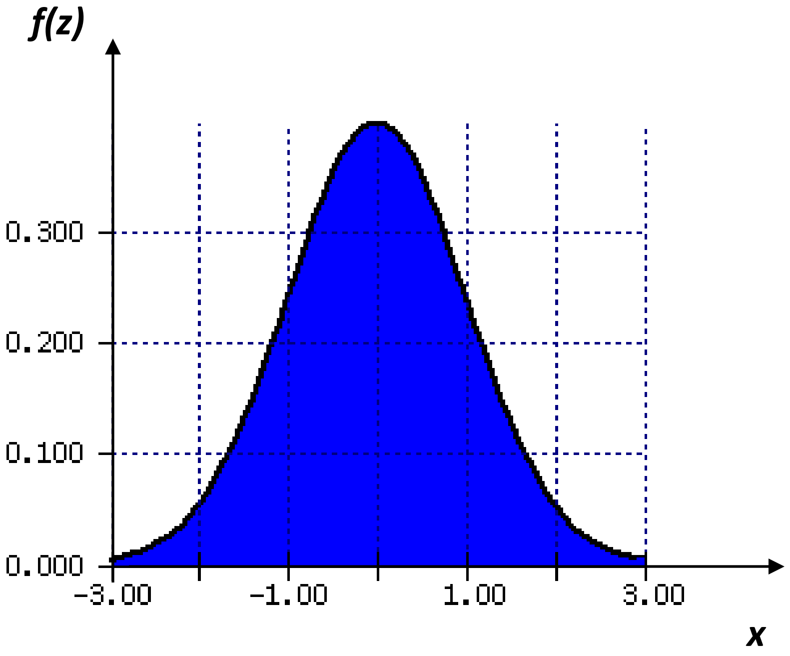
Стандартные распределения и их квантили

У нормального распределения есть три стандартных числа:

Вероятность попадания x в интервал $[\mu-1\sigma; \mu+1\sigma]$ равна $\approx$ 68%.

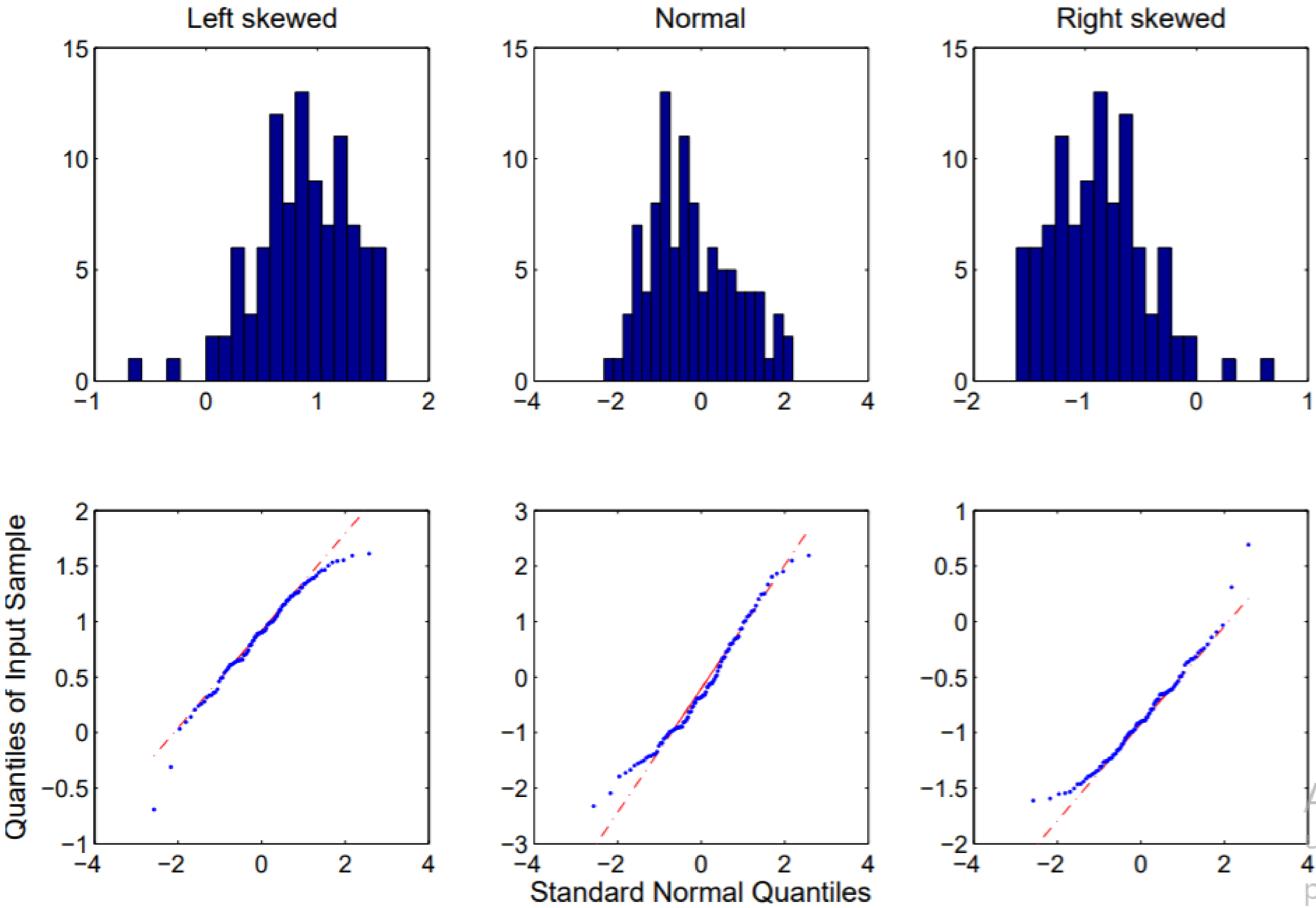
Вероятность попадания x в интервал $[\mu-2\sigma; \mu+2\sigma]$ равна $\approx$ 95%.

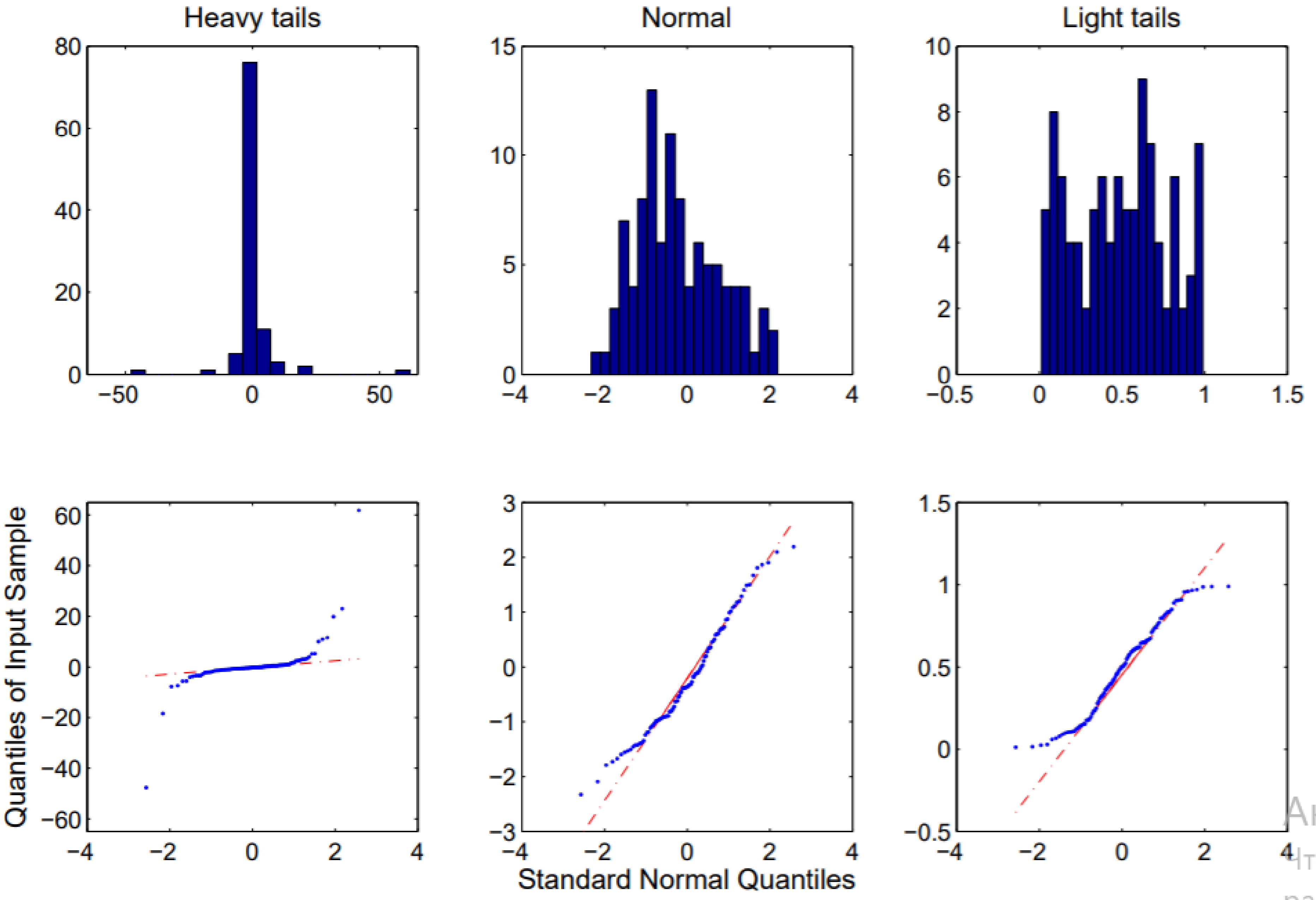
Вероятность попадания x в интервал $[\mu-3\sigma; \mu+3\sigma]$ равна $\approx$ 99,7%.



Таким образом, на отрезке $[-3\sigma, 3\sigma]$ находятся почти все значения. Это и есть так называемое **правило "трех сигм"**.

Например, пусть имеется выборка наблюдений за ежедневными продажами в магазине. Значения наблюдений распределены по нормальному закону со средним значением 150000 руб. и среднеквадратическим отклонением 20000 руб. Тогда в соответствии с правилом 3-х сигм продажи ниже, чем 150 000 - 20 000 x 3 = 90 000, и выше, чем 150 000 + 20 000 x 3 = 210 000, являются практически невозможными событиями. Фактически это означает, что рассматривать данные объемы продаж как потенциально возможные не имеет смысла.


Left skewed
Normal
Right skewed
Quantiles of Input Sample
Standard Normal Quantiles


Heavy tails
Normal
Light tails
Quantiles of Input Sample
Standard Normal Quantiles

Существенные отклонения

- 1. Наличие выбросов в данных.
- 2. Явная асимметрия гистограммы.
- 3. Очень сильное отклонение формы гистограммы от колоколообразной формы.

Существенные отклонения от нормальности – отвергаем гипотезу?

- 1. Наличие выбросов в данных.
- 2. Явная асимметрия гистограммы.
- 3. Очень сильное отклонение формы гистограммы от колоколообразной формы.

Рекомендуется

- Строго - к присутствию выбросов,
- Снисходительно - к отклонениям от симметрии.
- Отношение к колоколообразной форме гистограммы зависит от числа наблюдений. Если имеется меньше 30 наблюдений, то принимаем, если число наблюдений находится между 30 и 150, мы относимся к отклонениям снисходительно, если имеется больше 150 наблюдений – строго.

Способы

- Выбросы — удаляем (осторожно!)
- Асимметрия — преобразуем данные (например, логарифмируем, или преобразование Бокса-Кокса)
- Бимодальность — разбиваем выборку на подвыборки

- **очень маленькие выборки**: любой критерий может пропустить отклонения от нормальности, графические методы бесполезны;
- **очень большие выборки**: любой критерий может выявлять небольшие статистически, но не практически значимые отклонения от нормальности; значительная часть методов, предполагающих нормальность, демонстрируют устойчивость к отклонениям;
- **выбросы**: сильно влияют на выборочные коэффициенты асимметрии и эксцесса;

Благодаря центральной предельной теореме и удобству вывода критериев для нормально распределённых выборок методы, основанные на предположении о нормальности данных, наиболее широко распространены.

- Перед использованием методов, предполагающих нормальность, стоит проверить нормальность.
- Если принять предположение о нормальности, то можно применять более мощные критерии. Зачастую они также чувствительны к небольшим отклонениям от нормальности.
- Если гипотеза нормальности отвергается, следует использовать непараметрические методы.

Проверка гипотезы о нормальном распределении. Критерий Пирсона.

- Используя критерий Пирсона, при уровне значимости $a = 0{,}05$ проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с эмпирическим распределением выборки объема n .

H0: исследуемый признак X объектов генеральной совокупности распределен нормально.
H1: это не так.

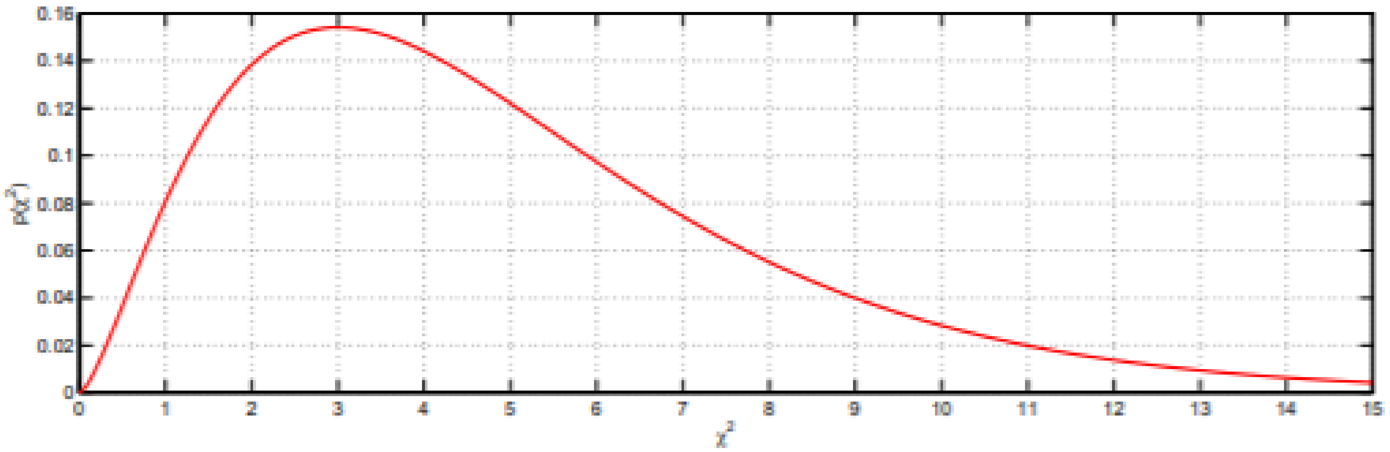
Принимается некоторый уровень значимости α.
Требуется указать критерий, по которому можно было бы решить, принимать или отвергать гипотезу *H0*.

Критерий согласия Пирсона (хи-квадрат)

выборка: $X^n = (X_1, \ldots, X_n)$;
нулевая гипотеза: $H_0\colon X \sim N(\mu, \sigma^2)$;
альтернатива: $H_1\colon H_0$ неверна;
статистика: $\chi^2(X^n) = \sum\limits_{i=1}^{K} \frac{(n_i - np_i)^2}{np_i}$;

$$\chi^2(X^n) \sim \begin{cases} \chi^2_{K-1}, & \mu, \sigma \text{ заданы}, \\ \chi^2_{K-3}, & \mu, \sigma \text{ оцениваются} \end{cases} \text{ при } H_0;$$

$[a_i, a_{i+1}], i = 1, \ldots, K$ — интервалы гистограммы,
n_i — число элементов выборки в $[a_i, a_{i+1}]$,
$p_i = F(a_{i+1}) - F(a_i)$ — вероятность попадания в i-й интервал при H_0.



Недостатки:

- разбиение на интервалы неоднозначно;
- требует больших выборок ($np_i > 5$ в 80% ячеек).

Ограничения критерия следующие.

1. Объем выборки должен быть достаточно большим: $n > 30$. При $n < 30$ критерий χ^2 дает весьма приближенные значения. Точность критерия повышается при больших значениях n.
2. Теоретическая частота для каждой ячейки таблицы не должна быть меньше 5. Это означает, что если число классов M задано заранее и не может быть изменено, то применять метод χ^2, не накопив определенного минимального числа наблюдений, нельзя. Если, например, проверяются предположения о том, что частота заболеваний гриппом неравномерно распределяются по 7 дням недели, то потребуется исследование $5 \cdot 7 = 35$ случаев для анализа. Таким образом, если количество классов M задано заранее ($M=7$), как в данном случае, минимальное число наблюдений (n_{min}) определяется по формуле: $n_{min} = 5 \cdot M = 35$.

Тест Шапиро–Уилка

- H0: что случайная величина, выборка X которой известна, распределена по нормальному закону.
- H1:закон распределения не является нормальным
- при небольших объемах выборки <2000 (сортировка)
- с увеличением количества наблюдений достоверность его снижается

```
from scipy.stats import shapiro
data1 = .... ,
p = shapiro(data)
```

Критерий согласия Андерсона-Дарлинга

- Классический непараметрический критерий согласия Андерсона-Дарлинга предназначен для проверки простых гипотез о принадлежности выборки некоторому закону распределения с известными параметрами. В этом случае распределение статистики критерия не зависит от закона, с которым проверяется согласие: критерий обладает свойством "свободы от распределения".
- H0: образец имеет гауссово распределение.
- H1: образец не имеет гауссовского распределения.
- его можно использовать при малых выборках, $n \leq 25$.

Код Python

```
from scipy.stats import anderson
data1 = ....
result = anderson(data)
```

ЦПТ

- Согласно центральной предельной теореме достаточно большая сумма сравнительно малых случайных величин ведет себя как нормальная случайная величина.

Влияние на машинное обучение

- Центральная предельная теорема имеет важные значения в прикладном машинном обучении.
- Теорема действительно дает информацию о решении линейных алгоритмов, таких как линейная регрессия, но не экзотических методов, таких как искусственные нейронные сети, которые решаются с использованием методов численной оптимизации. Вместо этого мы должны использовать эксперименты для наблюдения и записи поведения алгоритмов и использовать статистические методы для интерпретации их результатов.

Доверительные интервалы

- После того, как мы подготовили окончательную модель, мы можем сделать вывод о том, насколько искусной будет модель на практике.
- Представление этой неопределенности называется доверительным интервалом.
- Мы можем разработать несколько независимых (или близких к независимым) оценок точности модели, чтобы получить совокупность оценок навыков кандидатов. Среднее значение этих оценок навыка будет оценкой (с ошибкой) истинной базовой оценки навыка модели по проблеме.
- Зная, что среднее значение выборки будет частью гауссовского распределения из центральной предельной теоремы, мы можем использовать знания о распределении Гаусса для оценки вероятности среднего значения выборки на основе размера выборки и вычисления интервала желаемой достоверности вокруг модели.

подбрасывание одного игрального кубика - > Нескольких кубиков

```
import random
n =100
mas = []
for i in range(n): mas.append(random.randint(1,6))

pl.hist(mas, bins=[0.5,1.5,2.5,3.5,4.5,5.5,6.5])

pl.show()
```

Закон больших чисел

При неограниченном увеличении числа испытаний средние величины стремятся к некоторым постоянным.

Следствие 1 из закона больших чисел – о сходимости по вероятности среднего арифметического одинаково распределенных случайных величин к их математическому ожиданию.

Следствие 2 из закона больших чисел – теорема Бернулли о сходимости по вероятности относительной частоты к вероятности

- Как известно, нельзя заранее уверенно предвидеть, какое из возможных значений примет случайная величина в итоге испытания; это зависит от многих случайных причин, учесть которые невозможно. Казалось бы, поскольку о каждой случайной величине мы располагаем в этом смысле весьма скромными сведениями, то вряд ли можно установить закономерности поведения и суммы достаточно большого числа случайных величин.
- На самом деле это не так. Оказывается, что при некоторых сравнительно широких условиях суммарное поведение достаточно большого числа случайных величин почти утрачивает случайный характер и становится закономерным.

Закон больших чисел

- Закон больших чисел подтверждает, что выборка становится более представительной для населения по мере увеличения его размера.

- Закон больших чисел является другой отличной теоремой от статистики. Проще в том, что он утверждает, что по мере увеличения размера выборки, чем точнее оценка, то среднее значение выборки будет иметь среднее значение по совокупности.
- Центральная предельная теорема ничего не говорит о единственном образце среднего; напротив, он более широкий и что-то говорит о форме или распределении выборочных средств.
- Закон больших чисел интуитивно понятен. Именно поэтому мы считаем, что сбор большего количества данных приведет к более репрезентативной выборке наблюдений из области Теорема подтверждает эту интуицию.
- Центральная предельная теорема не интуитивна. Вместо этого мы можем использовать этот вывод, чтобы претендовать на примерные средства.

СПАСИБО ЗА ВНИМАНИЕ