Прикладная статистика

Слайды к лекции 1

17 января 2023

Сергей Александрович Спирин

sspirin@hse.ru

Структура отчётности по курсу

- Три контрольных работы по 1/15
- Три домашних задания: два по 1/10 и одно 1/5
- Экзамен 2/5

Статистика

- Описательная статистика
- Индуктивная статистика
- Математическая статистика

Описательная статистика

- Характеристики числовой совокупности:
  - среднее, квантили, квартили, процентили, децили, медиана
  - среднее квадратичное отклонение, стандартное отклонение
- Связанные наборы чисел: корреляция (обычная и ранговая)
- Графическое представление числовой совокупности:
  - гистограмма (и полигон)
  - ящик с усами
  - график оценки плотности распределения и скрипичная диаграмма
  - выборочная функция распределения (ВФР)
- Графическое представление двух и более совокупностей:
  - совместная гистограмма
  - несколько ящиков с усами
  - наложенные графики ВФР
- Графическое представление связанных совокупностей
  - scatter plot (диаграмма рассеяния, точечная диаграмма)
  - двумерная гистограмма

Индуктивная статистика

… она же просто «статистика»

Генеральная совокупность → выборка

По выборке судим о свойствах генеральной совокупности.

Важный момент: представительность выборки

Два раздела: оценка параметров и проверка гипотез.

Оценка параметров бывает точечная и интервальная.

Математическая статистика

Генеральную совокупность заменяем распределением случайной величины.

Вместо выборки рассматриваем набор одинаково распределённых случайных величин.

В результате и оценка параметров, и проверка гипотез формулируются в виде строгих математических задач.

Характеристики числовой совокупности

Имеем совокупность чисел $X_1, \ldots, X_n$

Среднее $\overline{X} = X_1 + \ldots + X_n / n$

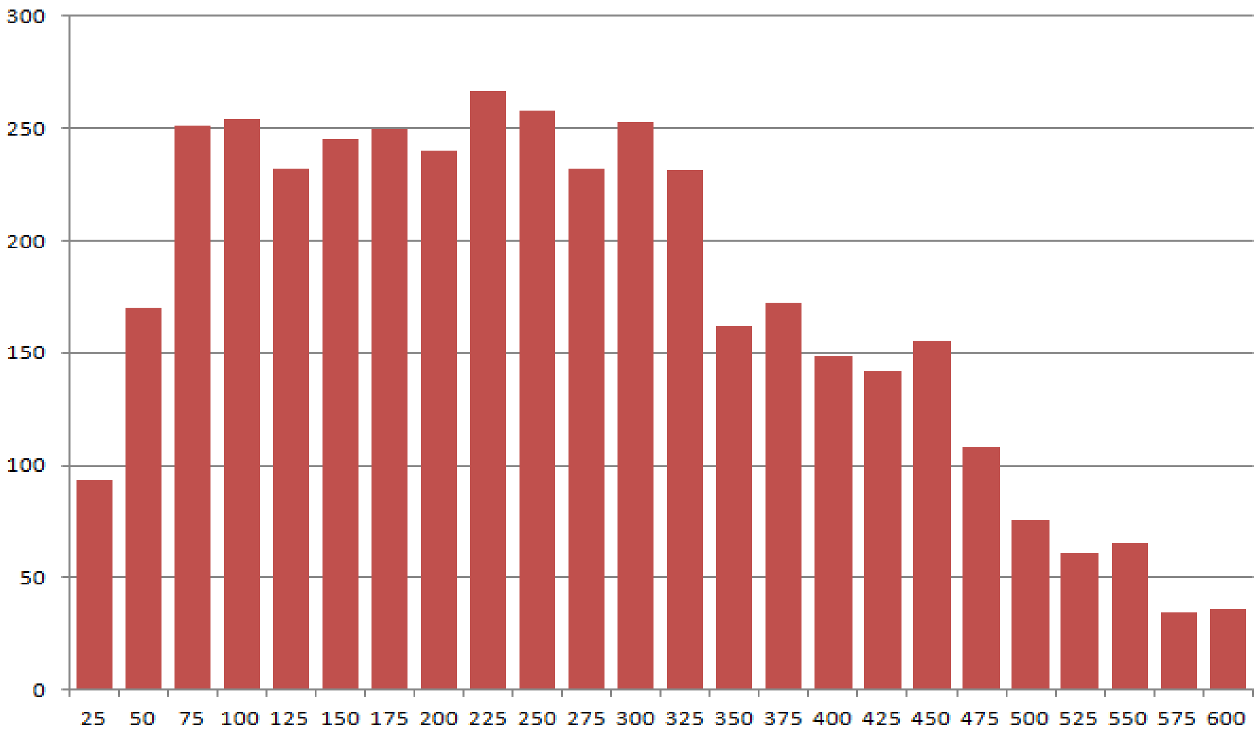
Медиана (если все X_i разные) — такое число, что количество X_i, меньших медианы, равно количеству X_i, больших медианы.
В общем случае надо упорядочить X_i по возрастанию и при нечётном n взять в упорядоченном списке элемент с номером $(n+1)/2$, а при чётном n — среднее элементов с номерами $n/2$ и $n/2 + 1$

Процентиль , соответствующий проценту k, в идеале — такое число, что $k\%$ данных меньше процентиля. Поскольку такое число бывает далеко не всегда, то за процентиль принимается элемент упорядоченного списка с номером $[kn/100] + 1$.
Квантиль — то же, что процентиль, но не для процента, а для доли (например, 0,7-квантиль).

Децили — это процентили 10%, 20%, …, 90%. **Кварти́ли** — это процентили 25% (нижний квартиль) и 75% (верхний квартиль). **Межквартильный размах** — разность между верхним и нижним квартилями.

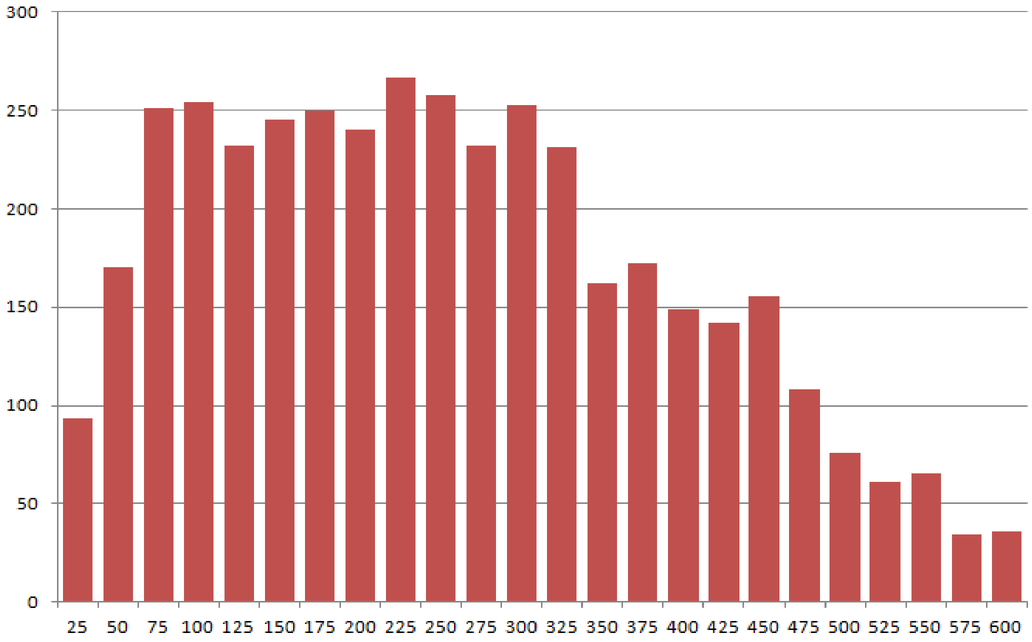
Среднее квадратичное отклонение – это корень квадратный из среднего квадрата отклонения от среднего, то есть из среднего величин $(X_i - \overline{X})^2$

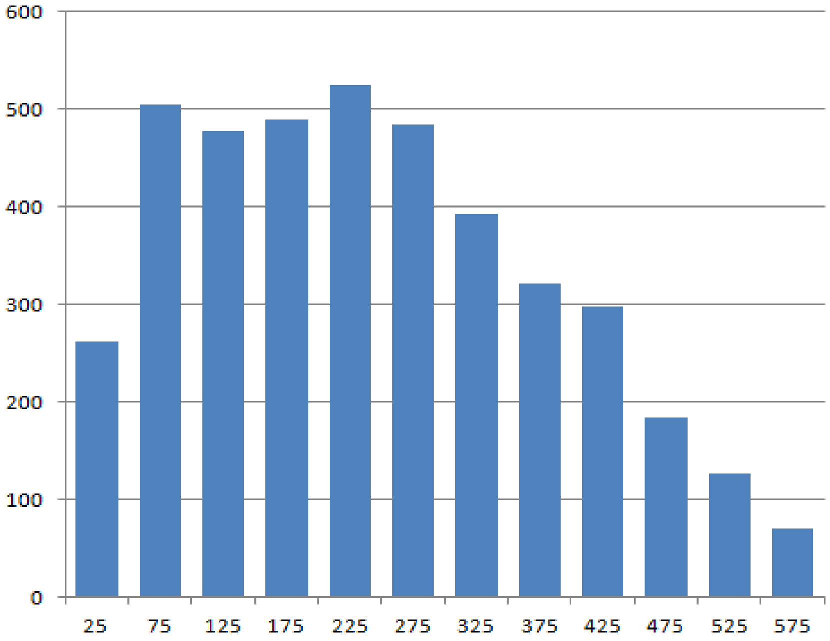
Гистограмма



Длины белков *E.coli*

Шаг гистограммы





Выбор шага гистограммы

Histogram - Wikipedia

en.wikipedia.org/wiki/Histogram#Number_of_bins_and_width

Number of bins and width [edit]

There is no "best" number of bins, and different bin sizes can reveal different features of the data. Grouping data is at least as old as Graunt's work in the 17th century, but no systematic guidelines were given[11] until Sturges' work in 1926.[12]

Using wider bins where the density of the underlying data points is low reduces noise due to sampling randomness; using narrower bins where the density is high (so the signal drowns the noise) gives greater precision to the density estimation. Thus varying the bin-width within a histogram can be beneficial. Nonetheless, equal-width bins are widely used.

Some theoreticians have attempted to determine an optimal number of bins, but these methods generally make strong assumptions about the shape of the distribution. Depending on the actual data distribution and the goals of the analysis, different bin widths may be appropriate, so experimentation is usually needed to determine an appropriate width. There are, however, various useful guidelines and rules of thumb.[13]

The number of bins *k* can be assigned directly or can be calculated from a suggested bin width *h* as:

$$k = \left\lceil \frac{\max x - \min x}{h} \right\rceil .$$

The braces indicate the ceiling function.

Square-root choice [edit]

$$k = \lceil \sqrt{n} \rceil$$

which takes the square root of the number of data points in the sample (used by Excel histograms and many other) and rounds to the next integer.[14]

Sturges' formula [edit]

Sturges' formula[12] is derived from a binomial distribution and implicitly assumes an approximately normal distribution.

$$k = \lceil \log_2 n \rceil + 1,$$

Sturges' formula implicitly bases bin sizes on the range of the data, and can perform poorly if $n < 30$, because the number of bins will be small—less than seven—and unlikely to show trends in the data well. On the other extreme, Sturges' formula may overestimate bin width for very large datasets, resulting in oversmoothed histograms.[15] It may also perform poorly if the data are not normally distributed.

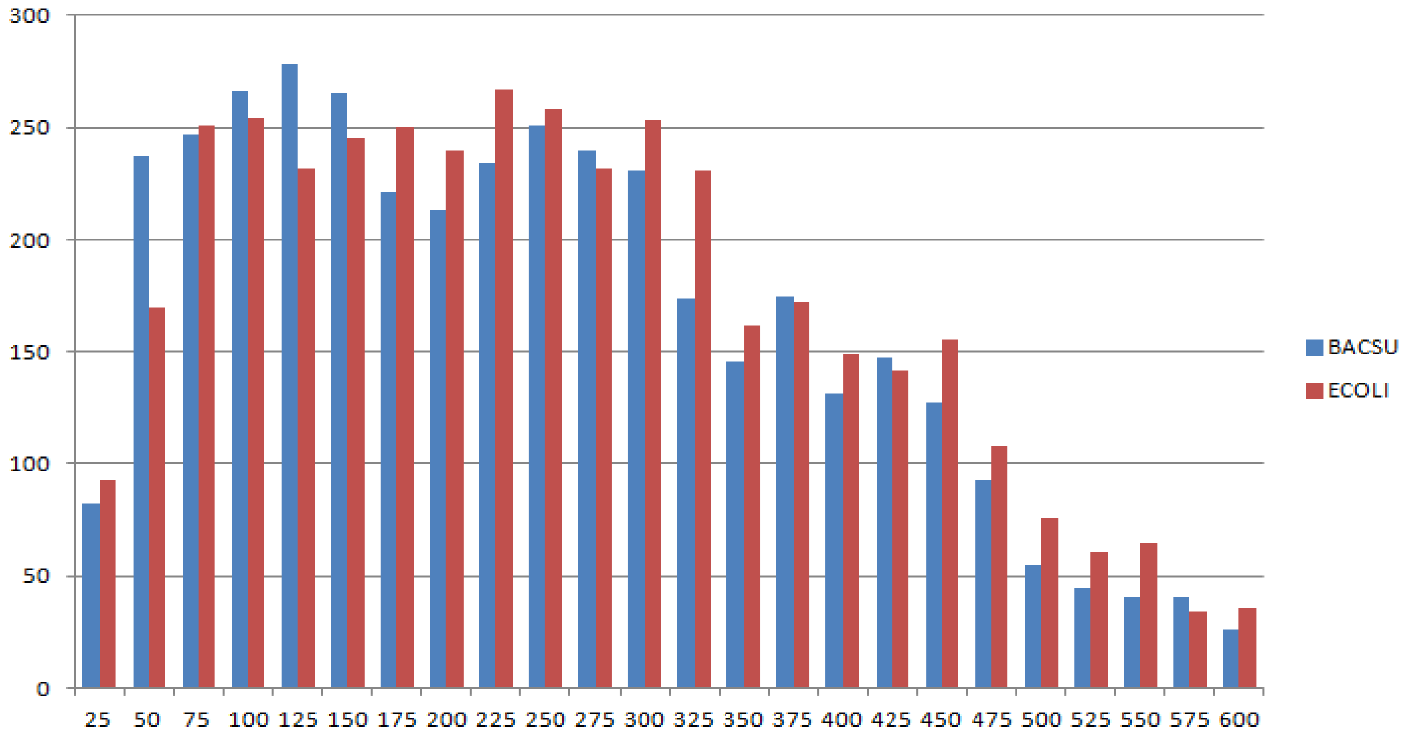
When compared to Scott's rule and the Terrell-Scott rule, two other widely accepted formulas for histogram bins, the output of Sturges' formula is closest when $n \approx 100$.[15]

Rice Rule [edit]

$$k = \lceil 2\sqrt[3]{n} \rceil ,$$

The Rice Rule [16] is presented as a simple alternative to Sturges' rule.

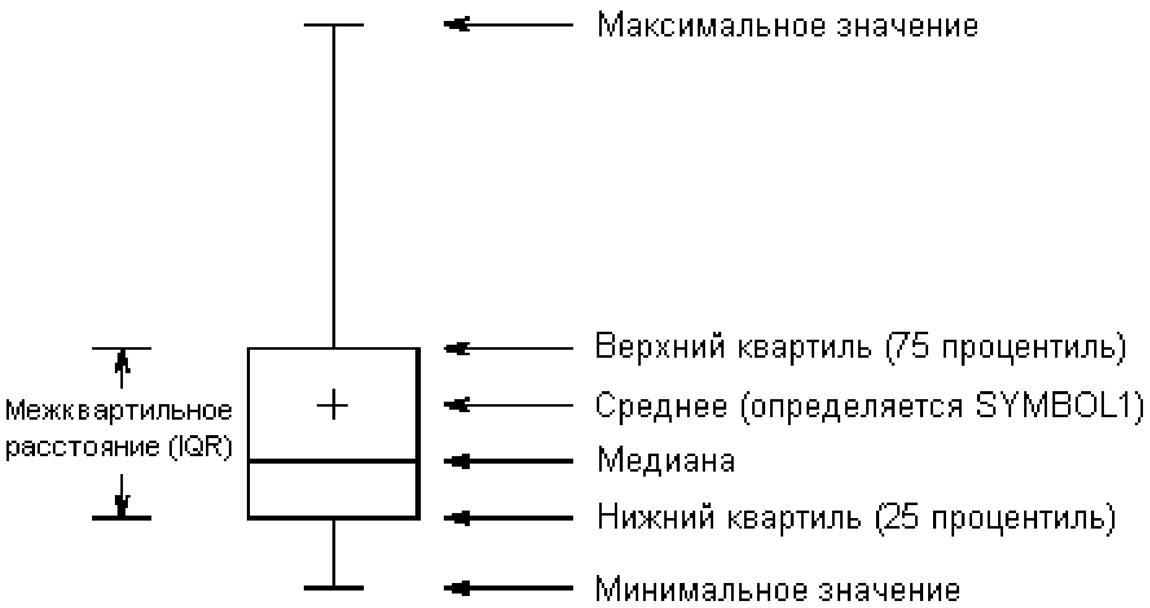
Совместная гистограмма



Длины белков *E.coli* и *B.subtilis*

Ящик с усами

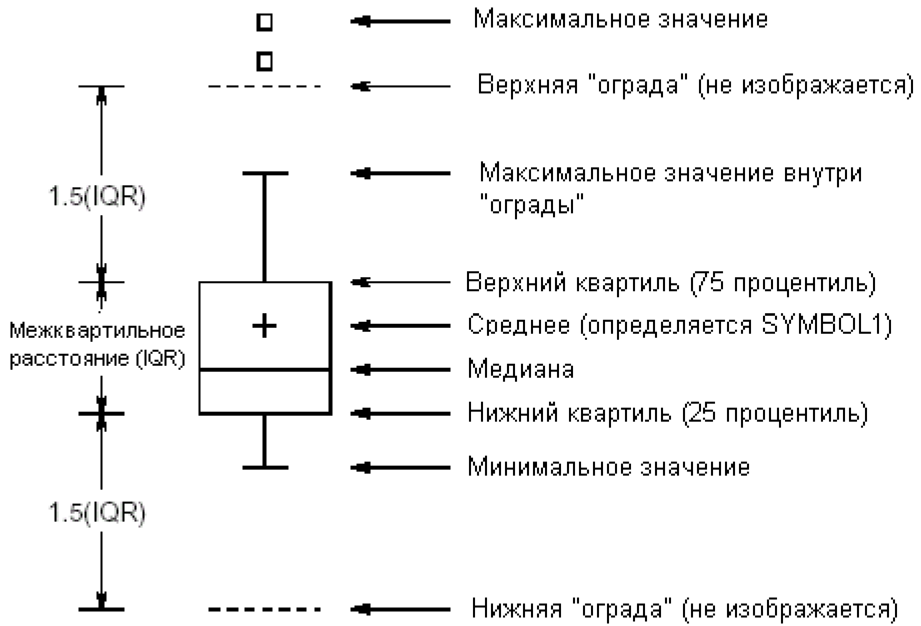
Стандартный вариант



http://pubhealth.spb.ru/SASDIST/Image384.gif

Ящик с усами

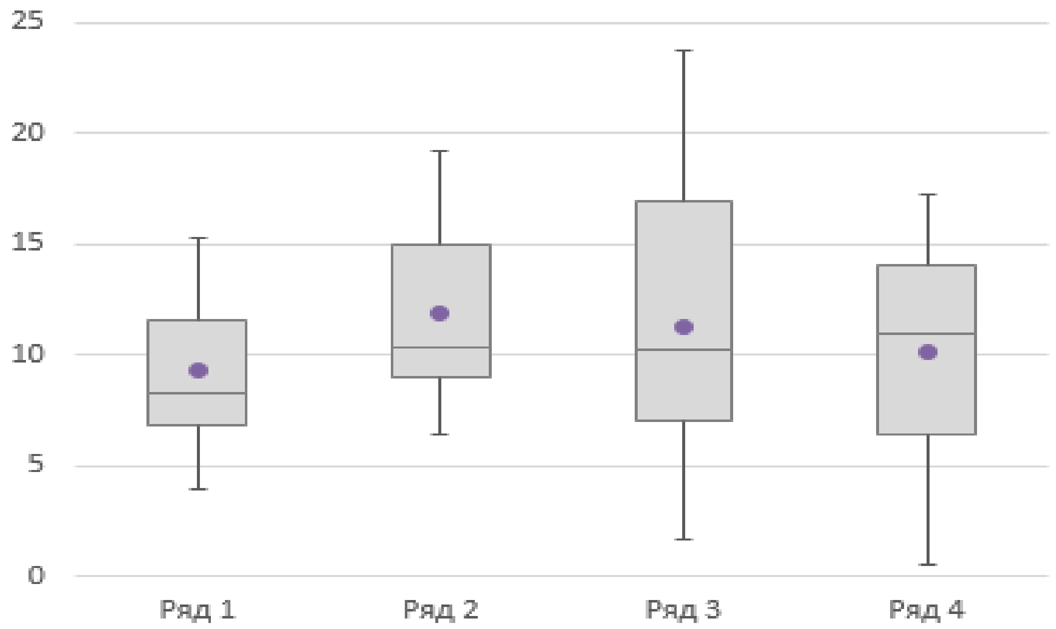
Вариант «с выбросами», очень популярный



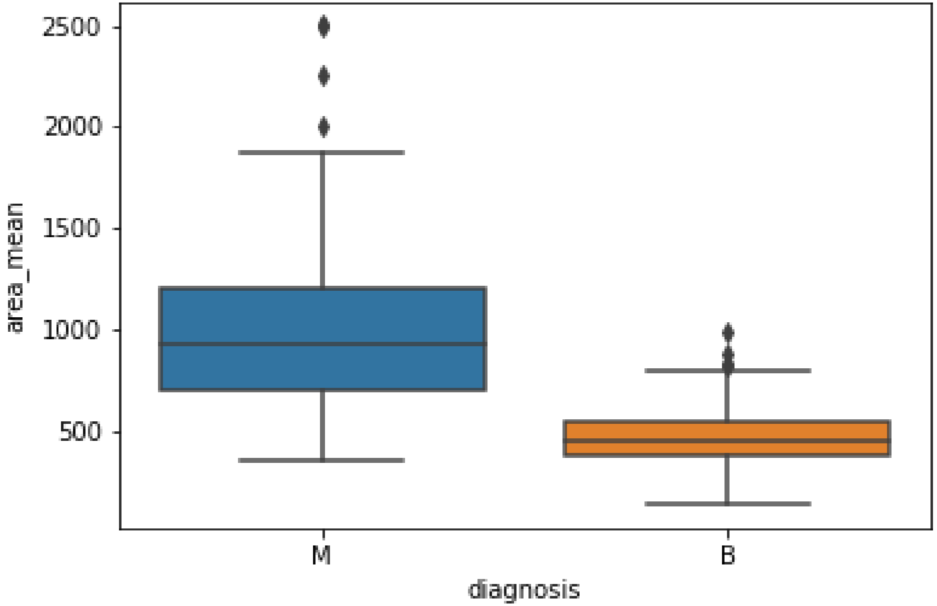
http://pubhealth.spb.ru/SASDIST/Image385.gif

Значения за «оградами» называются «выбросы» и изображаются отдельными точками (или другими символами)

Несколько ящиков с усами: сравнение выборок



Несколько ящиков с усами: сравнение выборок



https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Выдача FastQC

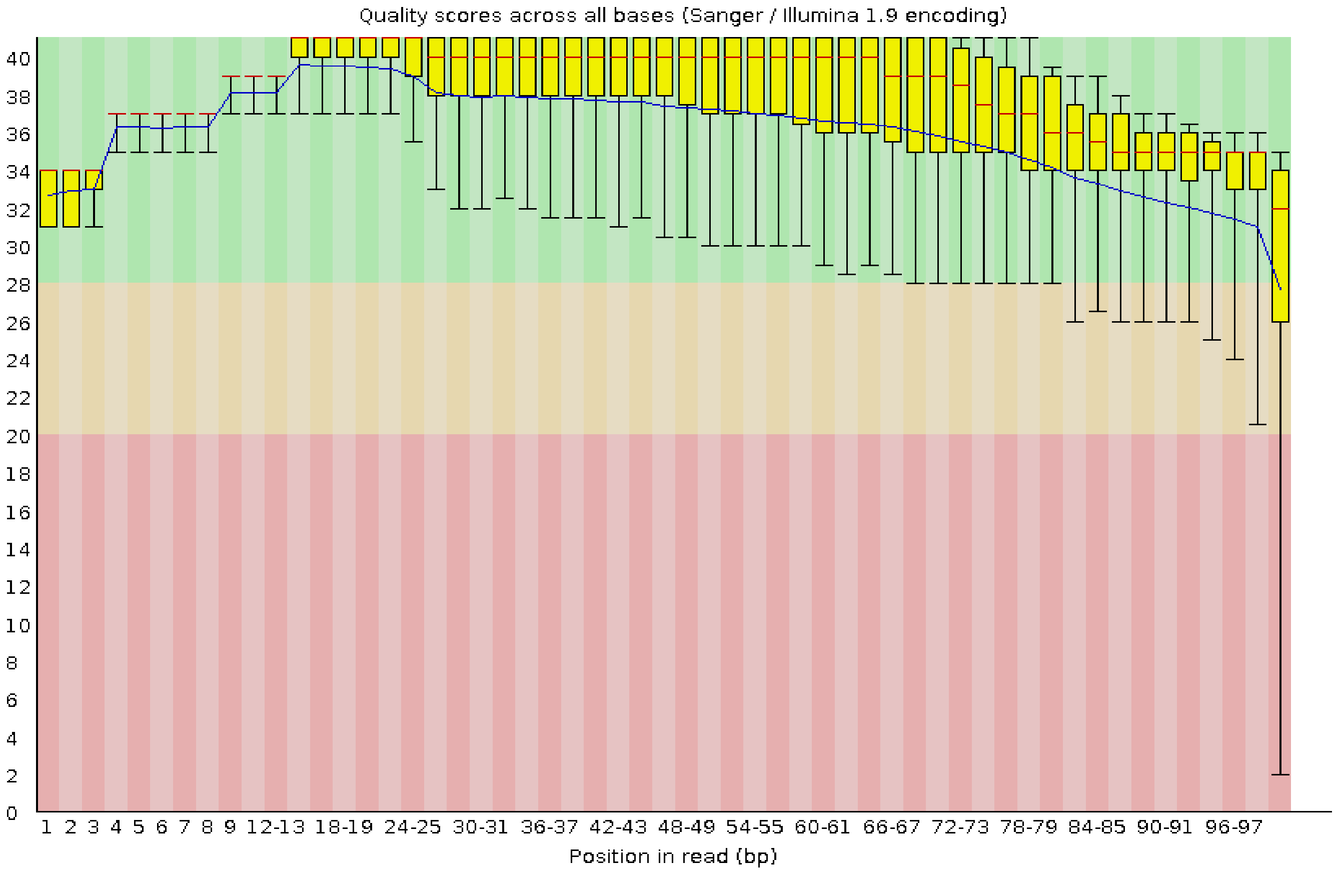
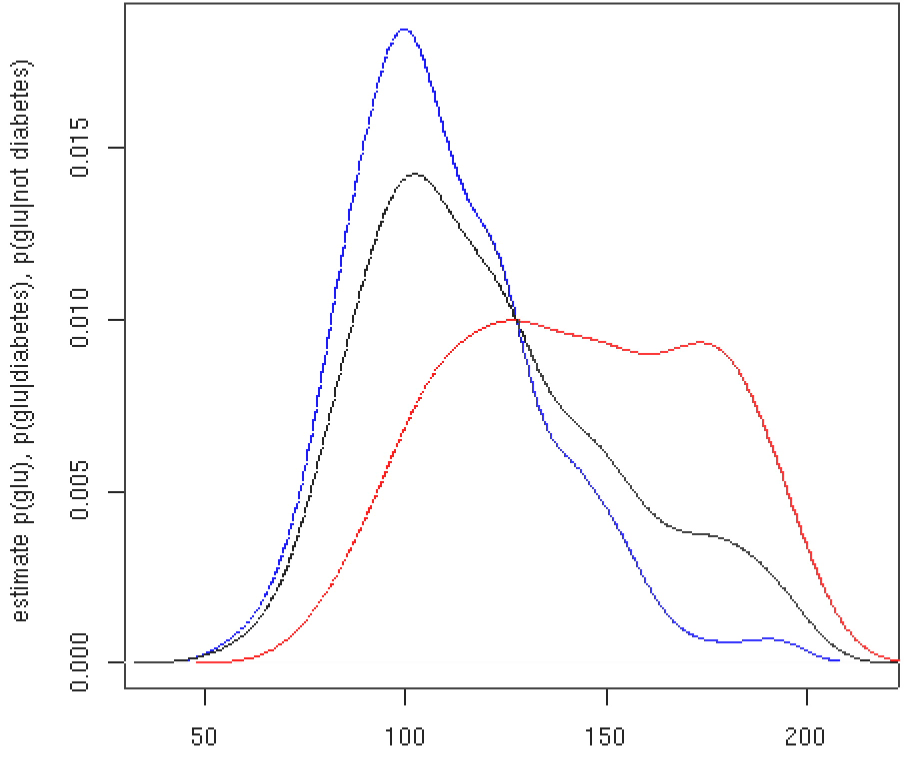
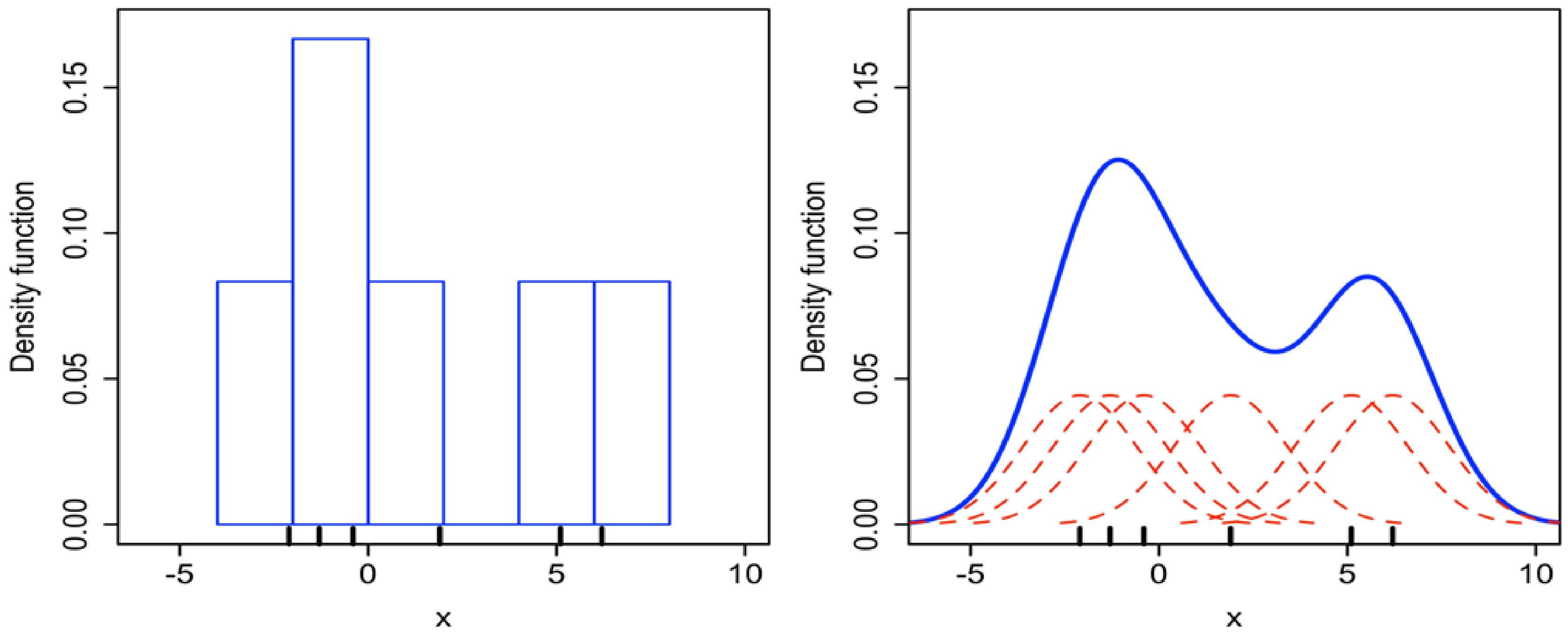


График оценки плотности распределения



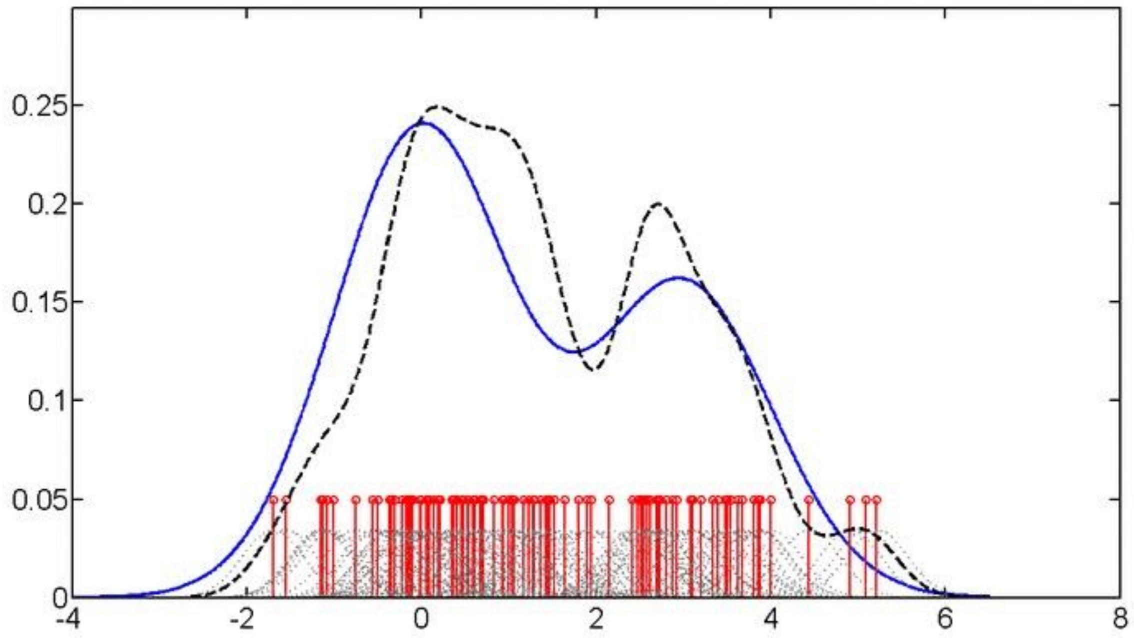
CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=643919

График оценки плотности распределения



Слева — гистограмма, справа — график оценки плотности распределения.
Т.н. «ядра» показаны красными штриховыми линиями.
Оценка плотности равна сумме значений ядер.
By Drleft at English Wikipedia, CC BY-SA 3.0,
https://commons.wikimedia.org/w/index.php?curid=57332968

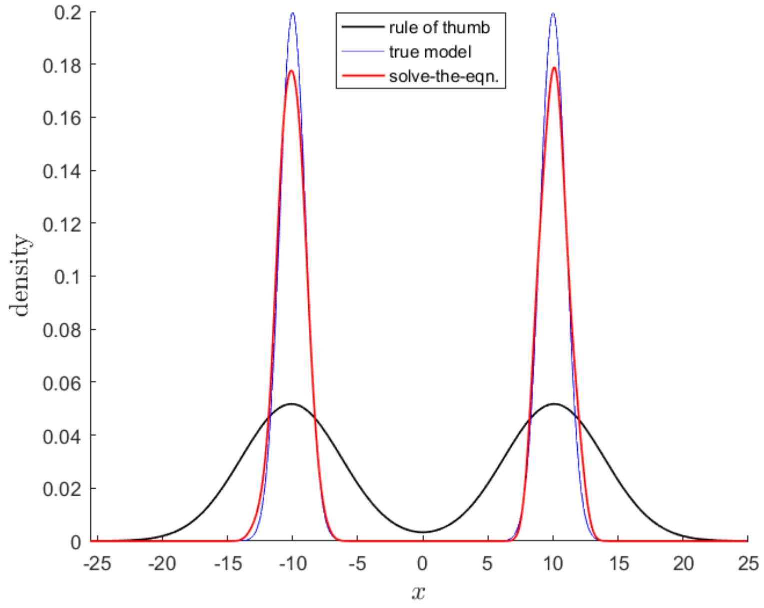
График оценки плотности распределения



Реальная плотность — синяя линия, оценка по случайной выборке — чёрная штриховая.

https://upload.wikimedia.org/wikipedia/commons/5/5c/KernelDensityGaussianAnimated.gif

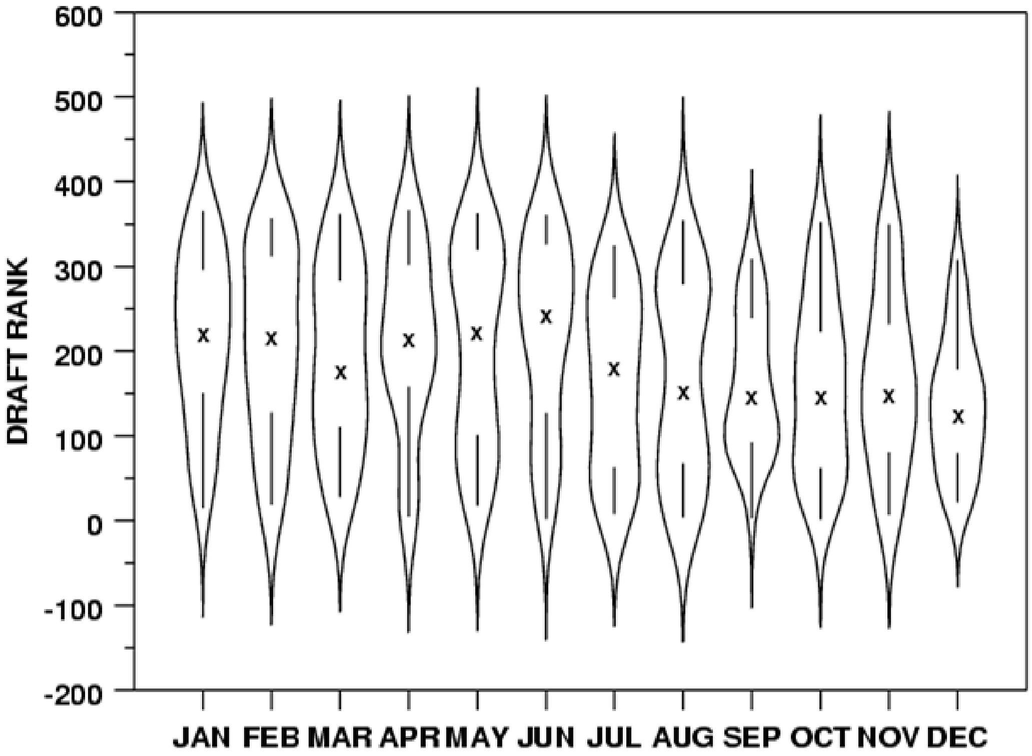
График оценки плотности распределения: роль выбора ядра



Реальная плотность — синяя линия, красная и чёрная — две оценки по одной и той же выборке, но с разными ядрами (здесь оба ядра — гауссовы, но различаются т. н. «полосой пропускания», т. е. шириной гауссианы).

By Kernel estimator - This diagram was created with MATLAB., CC0, https://commons.wikimedia.org/w/index.php?curid=73892722

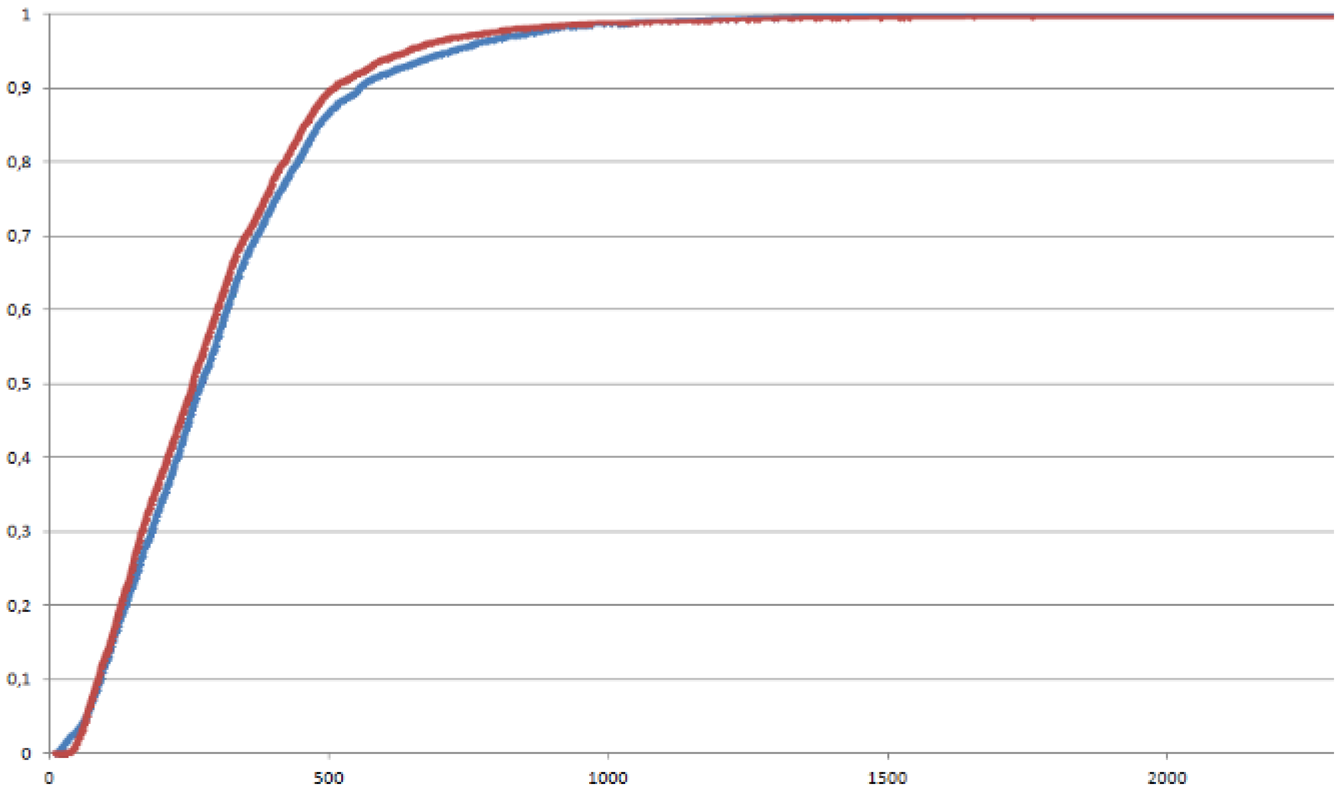
Скрипичная диаграмма



Скрипичная диаграмма (viloin plot) — как бы гибрид ящика с усами и графика оценки плотности

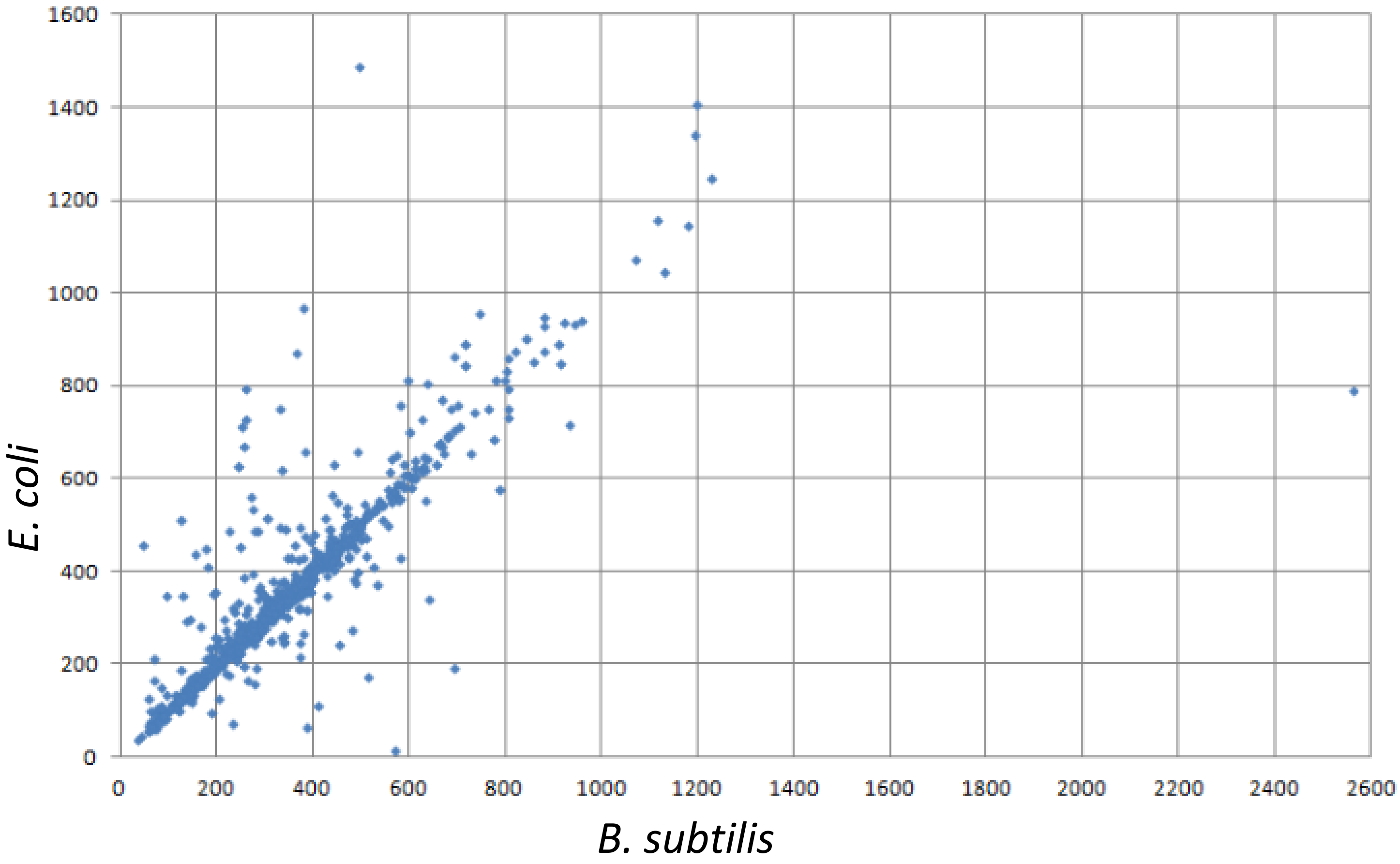
By itl.nist.gov - VIOLIN PLOT at www.itl.nist.gov, Public Domain, https://commons.wikimedia.org/w/index.php?curid=4992563

Эмпирические функции распределения



Значение функции равно доле чисел, меньших данного

Пары значений: точечная диаграмма



Пары значений: характеристики

- Ковариация
- Коэффициент корреляции (Пирсона)
- Коэффициент ранговой корреляции (Спирмена)

$$\mathbf{r}_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum(X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{X})^2 \sum(Y - \bar{Y})^2}}$$

Теория вероятностей

Случайные события
Случайные величины

Случайное событие происходит с некоторой вероятностью $0 \le P \le 1$

Случайная величина принимает числовые значения. То, что с.в. попала в какое-то числовое множество, является случайным событием и имеет вероятность.

Теория вероятностей: определения

Для каждой ситуации имеется некоторое «пространство элементарных событий» — множество, каждый элемент которого имеет вероятность.

Случайное событие — подмножество пространства элементарных событий. Каждое случайное событие имеет вероятность (равное сумме вероятностей элементов).
При этом элементарное событие — тоже частный случай события (одноэлементное множество).

Случайная величина — функция на пространстве элементарных событий.

Например, в ситуации броска двух игральных костей пространство элементарных событий состоит из всех возможных исходов (всевозможные пары очков).
Случайной величиной является, например, сумма очков на двух костях.
Пример случайного события — эта сумма равна 7.

Независимость и условная вероятность

Пусть A и B — два случайных события.

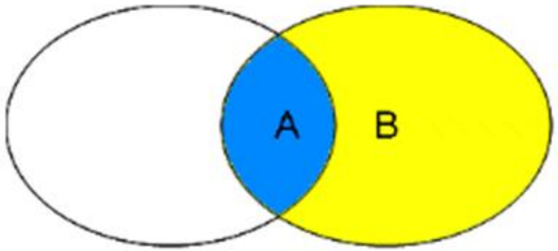
Событие, которое состоит в том, что произошли оба события A и B, обозначается $A \cdot B$ или $A \cap B$ (пересечение)

Событие, которое состоит в том, что произошло хотя бы одно из событий A или B, обозначается $A+B$ или $A \cup B$ (объединение)

A и B называются **независимыми**, если $P(A \cdot B) = P(A)P(B)$

Условная вероятность $P(A|B)$ определяется так: $P(A|B) = P(A \cdot B)/P(B)$

A и B независимы $\Leftrightarrow P(A|B) = P(A)$



Случайные величины ξ и η независимы, если для любых чисел a и b события $\xi = a$ и $\eta = b$ независимы.

Формула полной вероятности

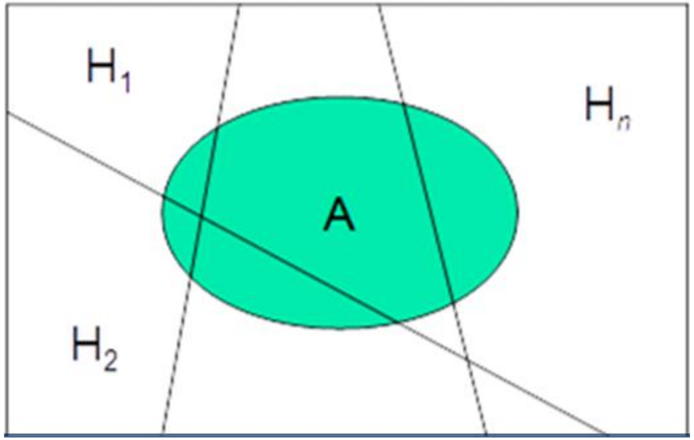
События H_1, H_2, …, H_n образуют **полную систему взаимоисключающих событий**, если

$P(H_i \cdot H_j) = 0$ для любых $i \neq j$ и $P(H_1 + H_2 + \ldots H_n) = 1$

(одно из событий обязательно происходят, и никакие два не могут произойти вместе) .

Формула полной вероятности:

$P(A) = P(A|H_1)P(H_1) + P(A|H_2)P(H_2) + \ldots + P(A|H_n)P(H_n)$



Формула Байеса

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

- P(A) – априорная (a priori) вероятность
- P(A|B) – апостериорная (a posteriori) вероятность

Задача

Предположим, что определенный тест на наркозависимость обладает 99% чувствительностью и 98% специфичностью, то есть тест даёт положительный результат для 99% потребителей наркотиков, и даёт отрицательный результат для 98% не-потребителей наркотиков.
Предположим, некая корпорация узнала, что 0,5% её сотрудников используют наркотики и решает проверить своих сотрудников на наркозависимость.
Для некоторого сотрудника тест дал положительный результат. Какова вероятность того, что этот сотрудник на самом деле употребляет наркотики?

Решение

A = «данный сотрудник потребляет наркотики»

B = «тест даёт положительный результат»

Известно, что:

- $P(A) = 0,005$
- $P(B|A) = 0,99$
- $P(B|\text{не } A) = 1 – 0,98 = 0,02$

Требуется посчитать $P(A|B)$. Применяем формулу Байеса:

$$P(A|B) = P(B|A) \cdot P(A) / P(B)$$

Чтобы посчитать $P(B)$, применяем формулу полной вероятности:

$P(B) = P(B|A) \cdot P(A) + P(B|\text{ не } A) \cdot P(\text{не } A) = 0,99 \cdot 0,005 + 0,02 \cdot 0,995 = 0,00495 + 0.0199 = 0,02485$

Получаем $P(A|B) = 0,99 \cdot 0,005 / 0,02485 = 0,00495 / 0,02485 = 0,199 \approx$ **20%**

Распределения случайных величин

- Дискретные
- Непрерывные

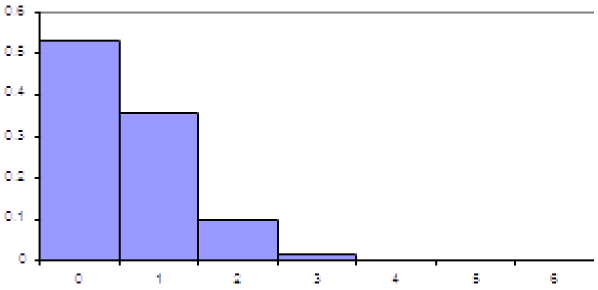
Каждая случайная величина имеет распределение.
Дискретная с.в. может принимать значения из какого-то конечного или бесконечного множества значений, каждое с ненулевой вероятностью.
Распределение дискретной с.в. задаётся набором вероятностей значений.

Биномиальное распределение

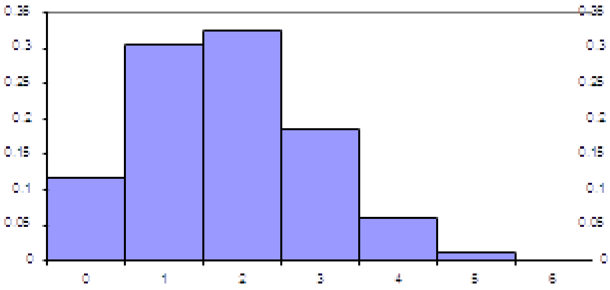
Биномиально распределённая величина = число успехов в n независимых испытаниях; параметр распределения p = вероятность успеха в одном испытании

$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

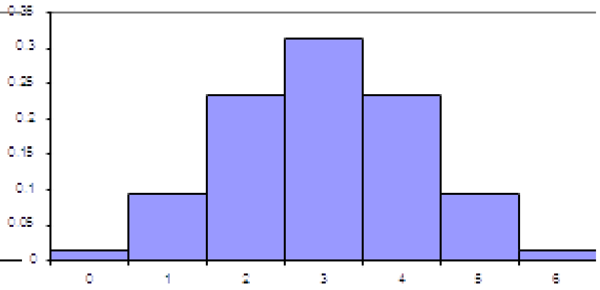
Вероятности 0, 1, …, 6 успехов при шести независимых испытаниях



p=0,1



p=0,3



p=0,5

Биномиальное распределение

Какова вероятность выбросить 5 или 6 шестёрок при шести бросках кости?

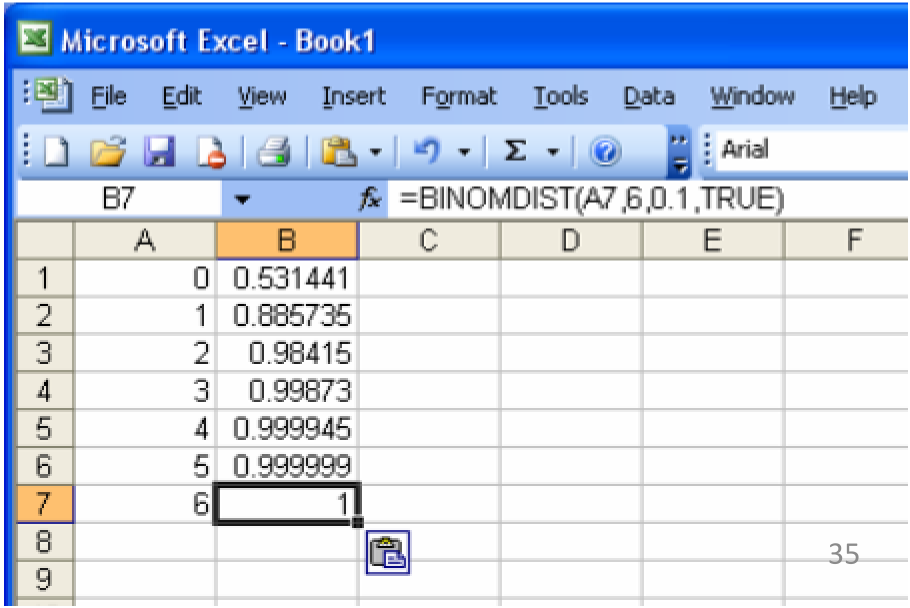
Интегральная вероятность

ξ = случайная величина

Функция распределения $F(x) = P(\xi \leq x)$

Большинство программных средств обработки данных содержит средства вычисления интегральной вероятности

Например, в Excel нужно указать функции BINOMDIST как последний аргумент TRUE, чтобы получить именно интегральную вероятность $P(\xi \leq x)$, а не обычную $P(\xi = x)$



Математическое ожидание

ξ	x_1	x_2	…	…	x_n
p	p_1	p_2	…	…	p_n

$E(\xi) = \Sigma\, x_i\, p_i =$ не случайная величина

(другое обозначение Mξ)

ξ	0	1
P	1/2	1/2

$E(\xi) = 0 \cdot 1/2 + 1 \cdot 1/2 = 1/2$

η	0	1
P	1/3	2/3

$E(\eta) = 0 \cdot 1/3 + 1 \cdot 2/3 = 2/3$

ξ+η	0	1	2
P	1/6	1/2	1/3

$E(\xi+\eta) = 1 \cdot 1/2 + 2 \cdot 1/3 = E\,(X) + E\,(Y)$

Дисперсия

$Var(\xi) = E\,[\,(\xi - E(\xi))^2\,] = E\,(\xi^2) - (E\,(\xi))^2$

другое обозначение $D\xi$

ξ	0	1
p	1/3	2/3

$E(\xi)=2/3$

$\xi - E(\xi)$	$-2/3$	1/3
p	1/3	2/3

$E(\xi - E(\xi)) = -\,2/9+2/9 = 0$

$(\xi - E(\xi))^2$	4/9	1/9
p	1/3	2/3

$Var(\xi) = 4/9 \cdot 1/3 + 1/9 \cdot 2/3 = 2/9$

ξ^2	0	1
P	1/3	2/3

$E(\xi^2)=2/3$

$Var(\xi)=E\,(\xi^2) - E^2(\xi)=2/3 - 4/9 = 2/9$

Математическое ожидание и дисперсия

ξ = случайная величина

- $E(\xi+\eta) = E(\xi) + E(\eta)$
- $E(c\xi) = cE(\xi)$
- $E(c) = c$
- если ξ и η независимы, то $E(\xi\ \eta) = E(\xi)E(\eta)$
- $Var(\xi) = E(\xi^2) - E^2(\xi)$
- $Var(c\xi)=c^2Var(\xi)$
- если ξ и η независимы, то $Var(\xi+\eta) = Var(\xi)+Var(\eta)$
- для общего случая $Var(\xi+\eta) = Var(\xi) + Var(\eta) + 2Cov(\xi,\eta)$

Упражнения

- Используя свойства $E(\xi)$, докажите, что
  - $Var(\xi) = E[\ (\xi - E(\xi))^2\ ] = E(\xi^2) - (E(\xi))^2$
  - $Var(\xi + \eta) = Var(\xi) + Var(\eta) + 2Cov(\xi, \eta)$

  где $Cov(\xi, \eta) = E[\ (\xi - E(\xi)) \cdot (\eta - E(\eta))\ ]$
  - $Cov(\xi, \eta) = E(\xi \cdot \eta) - E(\xi) \cdot E(\eta)$

- Приведите пример ξ и η таких, что ξ и η зависимы, но $Cov(\xi, \eta) = 0$

Биномиальное распределение

$$\Pr(K = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

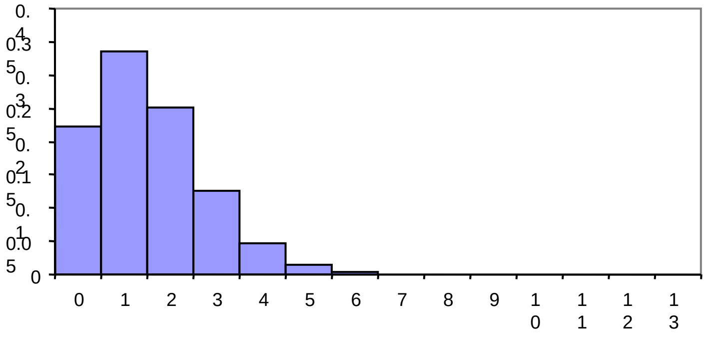
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

E(K) = np
Var(K) = $np(1–p)$

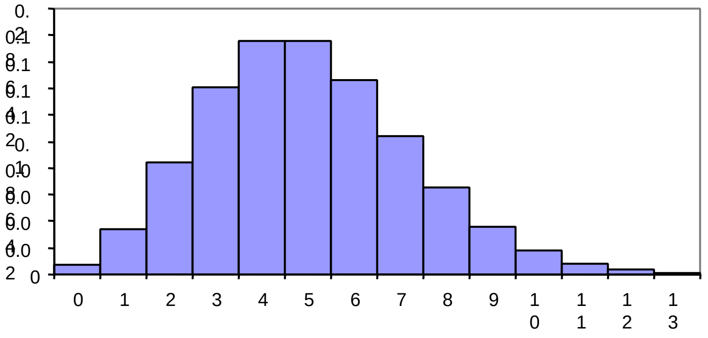
Распределение Пуассона

Случайная величина, распределённая по Пуассону = число (достаточно редких) событий за (достаточно большой) промежуток времени или в (достаточно большой) области пространства.
Имеет один параметр: λ — среднее число событий.

Вероятность наблюдать ровно *k* событий: $f(k;\lambda) = \dfrac{\lambda^k e^{-\lambda}}{k!}$

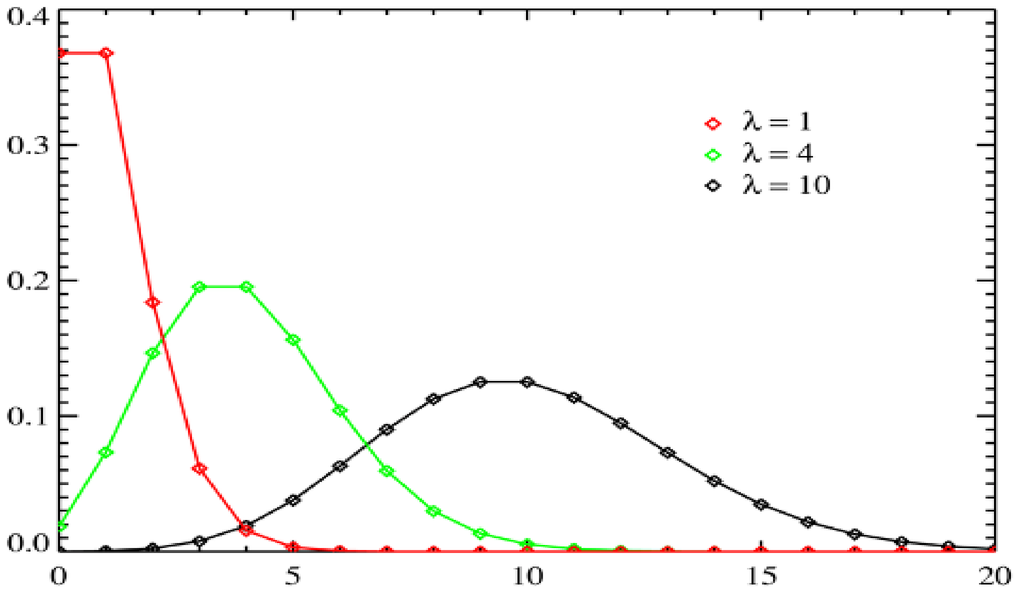


λ=1,5



λ=5

Распределение Пуассона



$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Если ξ — случайная величина, распределённая по Пуассону с параметром λ, то E(ξ) = λ и Var(ξ) = λ
(для распределения Пуассона мат. ожидание равно дисперсии)

Сборка чтений на геном

Пусть длина чтения 100, размер генома 1 млн п.н. и мы получили 50 000 чтений. Значит, среднее покрытие = 5. Хватит ли этого, чтобы собрать весь геном?

Сборка чтений на геном

Пусть длина чтения 100, размер генома 1 млн п.н. и мы получили 50 000 чтений. Значит, среднее покрытие = 5. Хватит ли этого, чтобы собрать весь геном?

Количество чтений, покрывающих данный нуклеотид, распределено по Пуассону:

$$P(k) = \exp(-\lambda)\, \lambda^k / k!$$

где k – число чтений, λ – среднее покрытие (в нашем случае $\lambda = 5$).

Значит, вероятность того, что на нуклеотид нее попадёт **ни одного** чтения, равна $P(0) = \exp(-\lambda)$. При $\lambda = 5$ эта вероятность равна $1/\exp(5) \approx 1/148$.

Сборка чтений на геном

Пусть длина чтения 100, размер генома 1 млн п.н. и мы получили 50 000 чтений. Значит, среднее покрытие = 5. Хватит ли этого, чтобы собрать весь геном?

Ответ: вряд ли. Чтения ложатся случайно, примерно каждый 150-ый нуклеотид ими не покроется. То есть почти наверняка более 6 000 нуклеотидов не будет покрыто, и при самой идеальной сборке получится не целый геном, а много кусков, разделённых непокрытыми участками.

При таком размере генома нужно не менее чем 15-кратное среднее покрытие, чтобы можно было рассчитывать собрать геном полностью.