Школа Data analyst

Занятие 12

Статистический анализ

Тема 2

Disclaimer

Все формулировки далее нестрогие, за более строгими определениями обращайтесь к специализированной литературе

План занятия

- Оценка распределения по выборке
- Важные характеристики распределений
- Важные статистики
- Доверительные интервалы
- Центральная предельная теорема
- Статистический вывод

Зачем оценивать распределение по выборке?

- Чтобы получить некое представление о мире
- Чтобы делать базовые прогнозы на основе полученных распределений (при этом имея численное, а не качественное выражение)
- Часто нас не интересуют супер точные результаты, да мы и не можем учесть такое кол-во переменных у себя в голове
- Некоторые алгоритмы машинного обучения (например, Байесовские алгоритмы классификации), основываются на знании априорных (предопределенных, доопытных) вероятностях классов

Оценка распределения по выборке

В общем виде задача формулируется так:

“Требуется оценить плотность распределения p(x) по выборке независимых случайных векторов, распределенных по этому закону p(x).”

Оценка распределения по выборке

Выборка случайной величины X :

$X^n = (X_1, \dots, X_n)$, $\quad n$ – объем выборки

X^n - независимы и распределены одинаково (i.i.d.)

$T(X^n)$ - статистика, функция от выборки, возвращающая какое-то число

* independent and identically-distributed

Оценка распределения по выборке

Воспоминание:

Распределение дискретной случайной величины задается функцией вероятности:

$$X \in \ \Omega = \{\omega_1, \omega_2, \omega_3, \dots\}, \qquad P(X = \omega_k) = p_k$$

$$\bar{p}_k = \frac{1}{n}\sum_{i=1}^{n}[Xi = \omega_k]$$

Оценка распределения по выборке

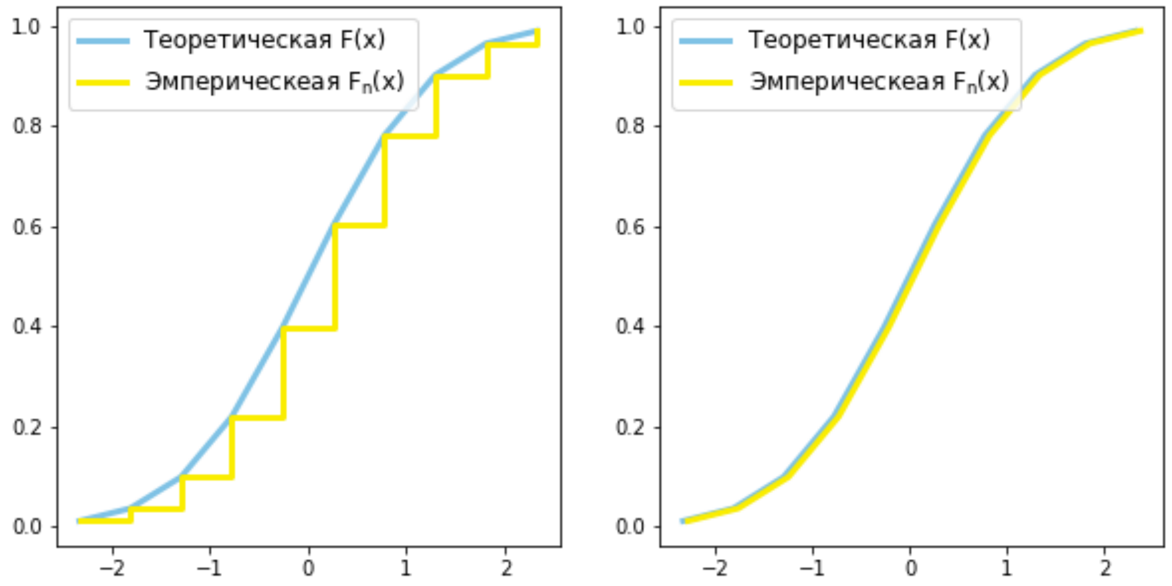
Для дискретной выборки функцию вероятности можно оценить частотами событий

Но что делать для непрерывной случайной величины?

$$\boldsymbol{X \sim F(x)}$$

Оценка распределения по выборке

$F_n(x) = \frac{1}{n}\sum_{i=1}^{n}[X_i \leq x]$ - *эмпирическая функция распределения*



$n \to \infty$

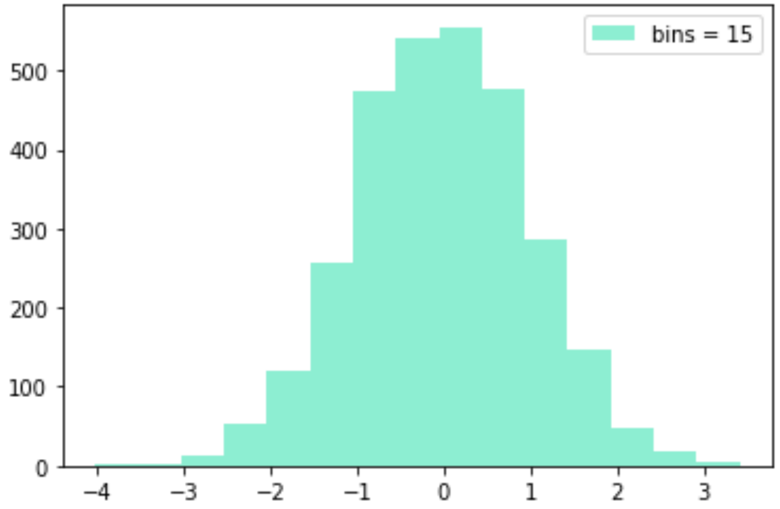
Оценка распределения по выборке

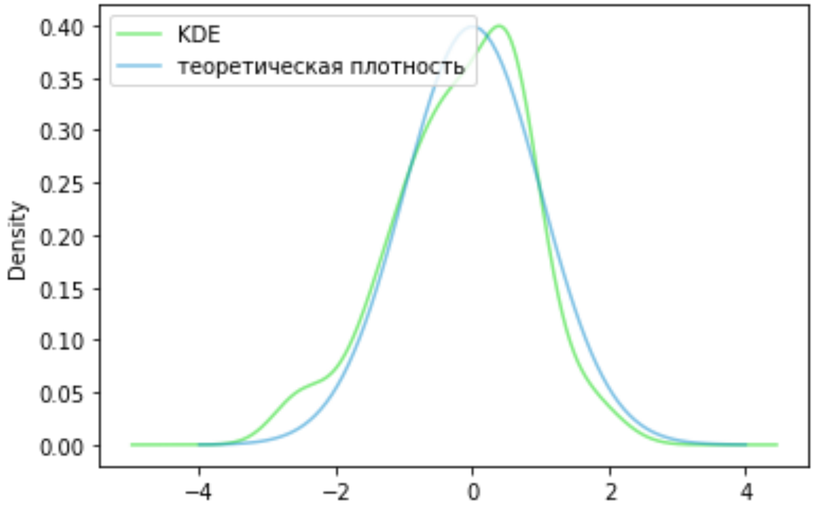
$X \sim f(x)$

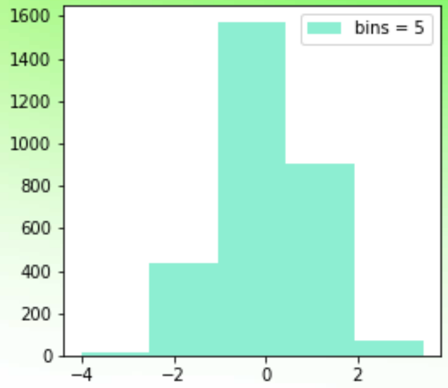
$$f(x):\ \int_a^b f(x)dx = P(a \le x \le b)$$

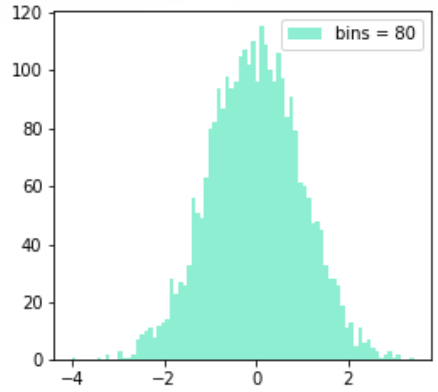
Формула Стерджесса

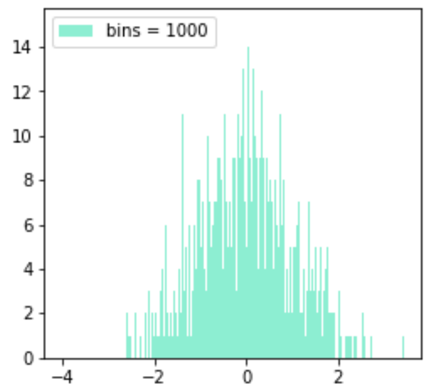
$$bins\ = 1 + [log_2 N]$$
$$bins\ = 1 + 3.322\ lg\ N$$



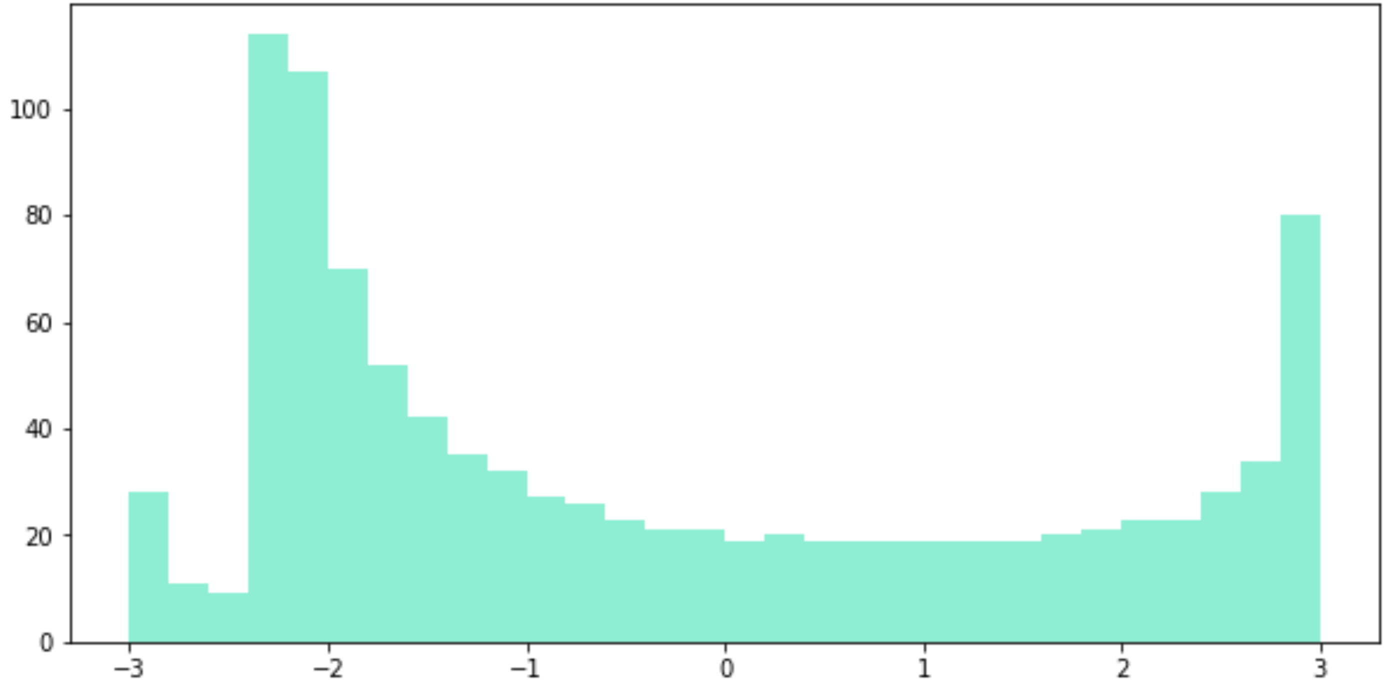








Оценка распределения по выборке



Colab? Colab!

Характеристики и статистики

Математическое ожидание

$$\mathbb{E}X = \begin{cases} \sum_i \omega_i p_i & \text{Дискретный случай} \\ \int_{-\infty}^{+\infty} x f(x) dx & \text{Непрерывный случай} \end{cases}$$

Важные характеристики распределений

Дисперсия и среднеквадратическое отклонение

$\mathbb{D}X = \mathbb{E}\left((X - \mathbb{E}X)^2\right)$ Дисперсия*

$\sigma = \sqrt{\mathbb{D}X}$ Стандартное отклонение

$IQR = X_{0.75} - X_{0.25}$ Интерквартильный размах

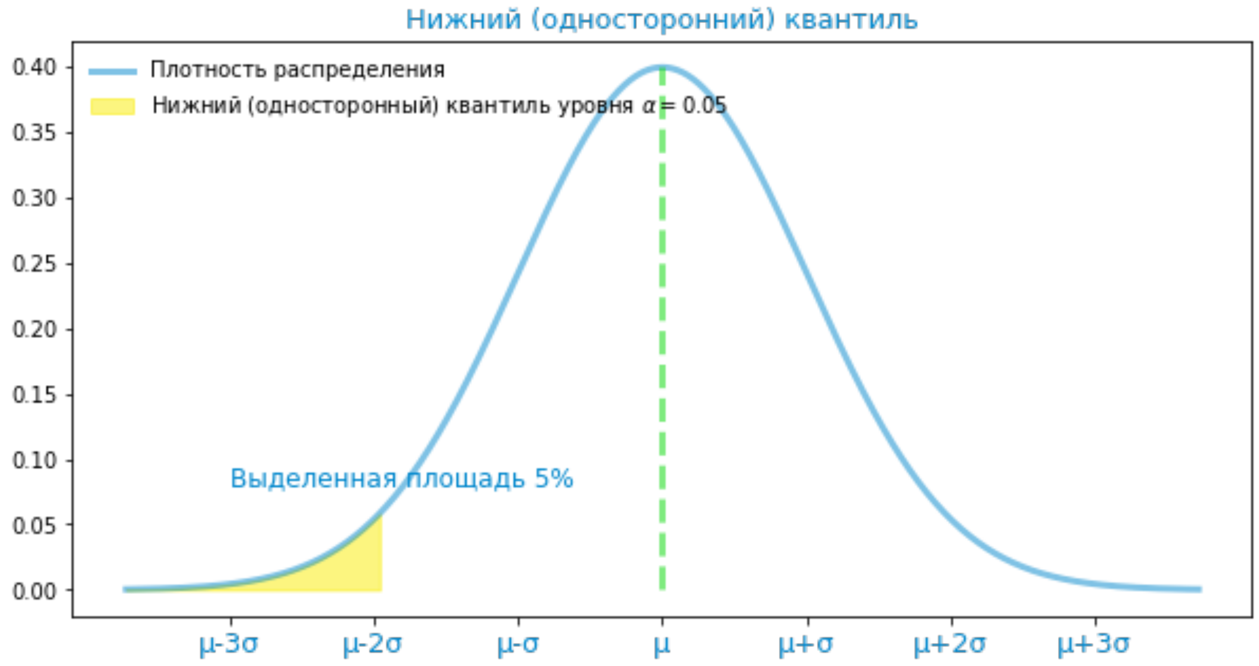
* Средний квадрат отклонения от среднего значения $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$ $n \to \infty$

Важные характеристики распределений

X_α Квантиль порядка α ∈ (0, 1):

$$\mathbb{P}(X \leq X_\alpha) \geq \alpha$$

$$\mathbb{P}(X \geq X_\alpha) > 1 - \alpha$$

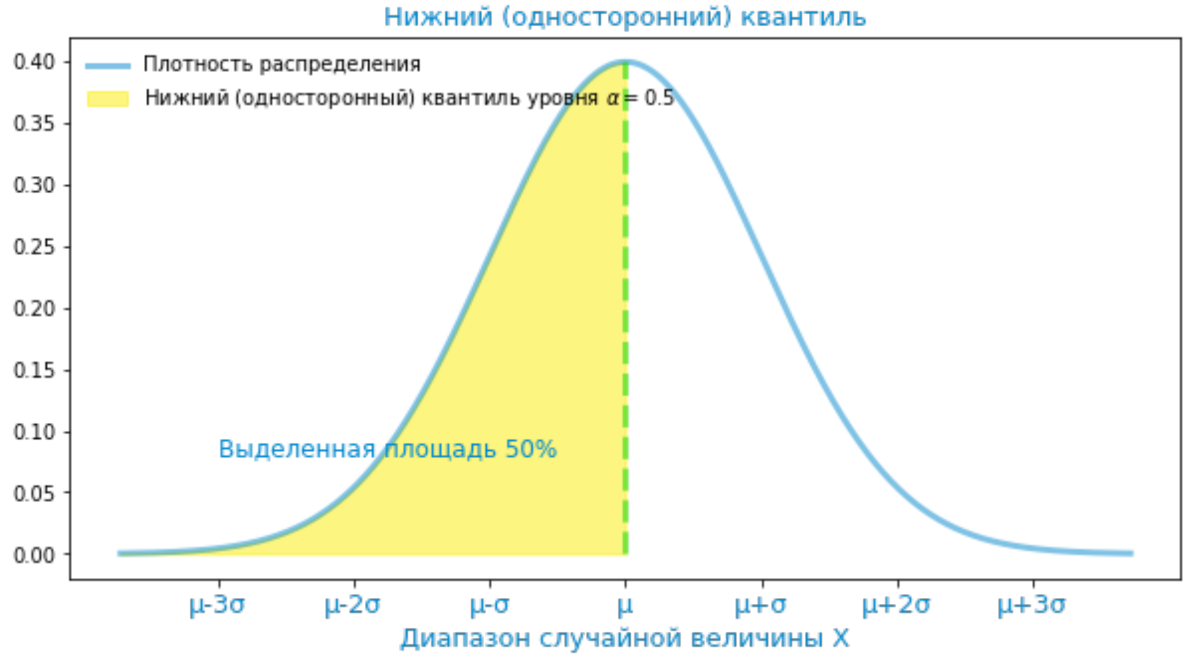


Важные характеристики распределений

Медиана — квантиль с α = 0.5. Т.е. элементы выборки с одинаковой вероятностью попадают по обе стороны медианы.

$$\mathbb{P}(X \leq X_\alpha) \geq 0.5$$

$$\mathbb{P}(X \geq X_\alpha) > 0.5$$

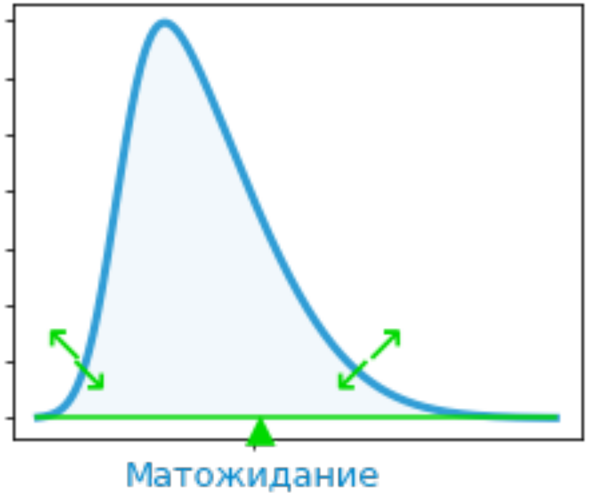


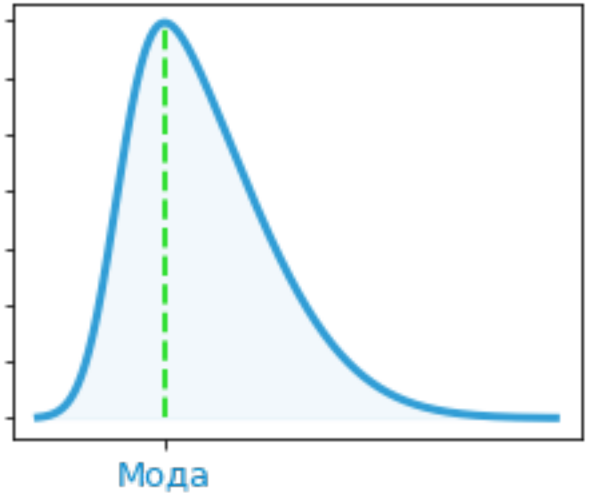
Важные характеристики распределений

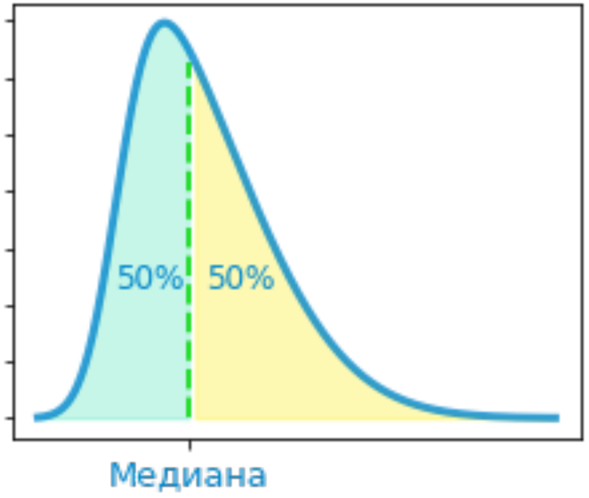
Мода — «наиболее вероятное» (частое) значение случайной величины

$$modeX = \begin{cases} \underset{i}{argmax}\, p_i \quad , & X - \text{дискретна} \\ \underset{x}{argmax}\, f(x), & X - \text{непрерывна} \end{cases}$$

Важные характеристики распределений







Важные характеристики распределений

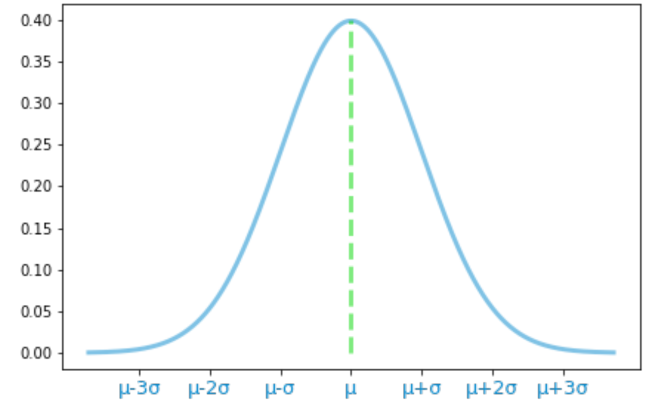
$$X \sim N(\mu, \sigma^2) \implies \mathbb{E}X = \mu = mode\ X = med\ X$$

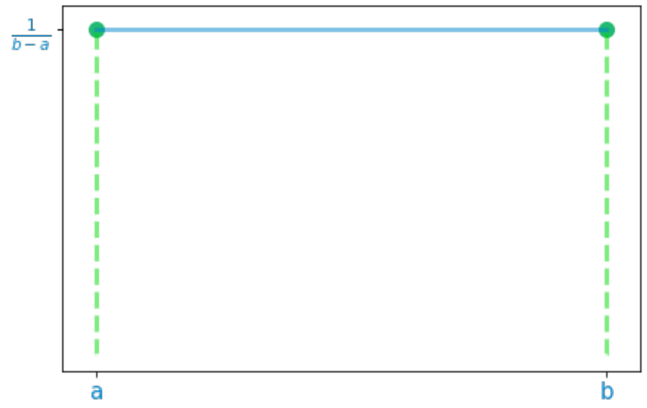
$$\mathbb{D}X = \sigma^2$$

$$X \sim U(a, b) \implies \mathbb{E}X = med\ X = \frac{a+b}{2}$$

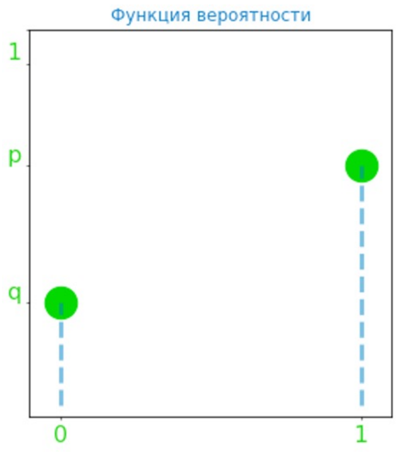
$mode\ X$ – не определена (любое число на отрезке $[a, b]$)

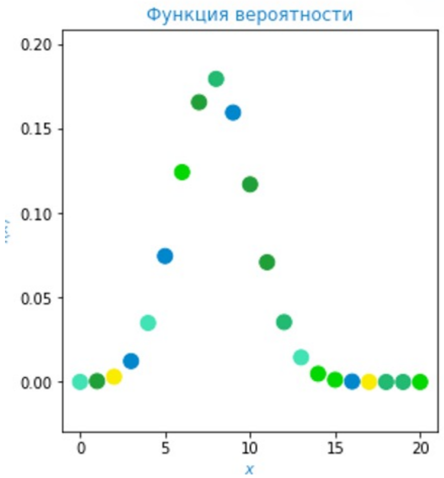
$$\mathbb{D}X = \frac{(b-a)^2}{12}$$

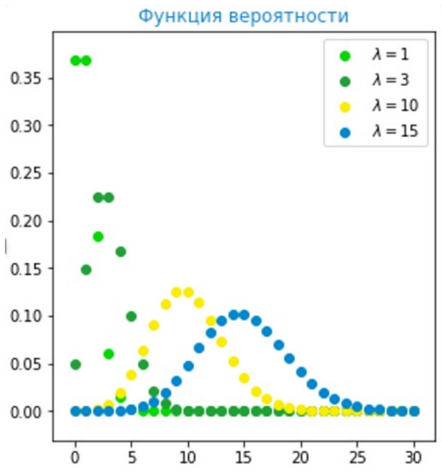




Важные характеристики распределений







$X \sim Ber(p) \implies \mathbb{E}X = p$

$\mathbb{D}X$= pq

$X \sim Binom(n, p) \implies \mathbb{E}X = np$

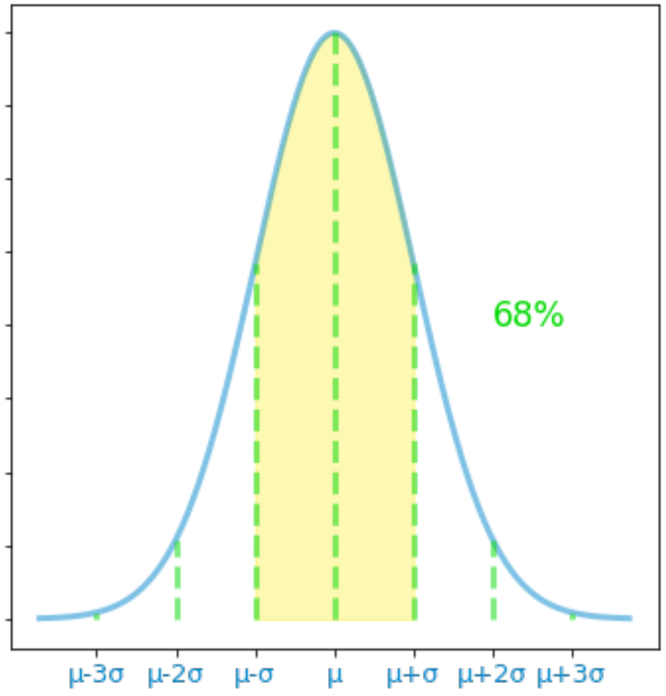
$\mathbb{D}X$= npq

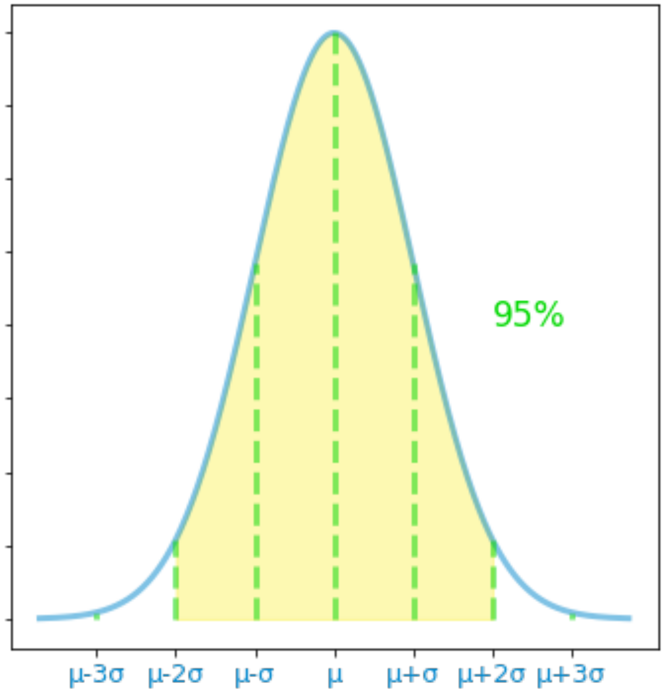
$X \sim Pois(\lambda) \implies \mathbb{E}X = \lambda$

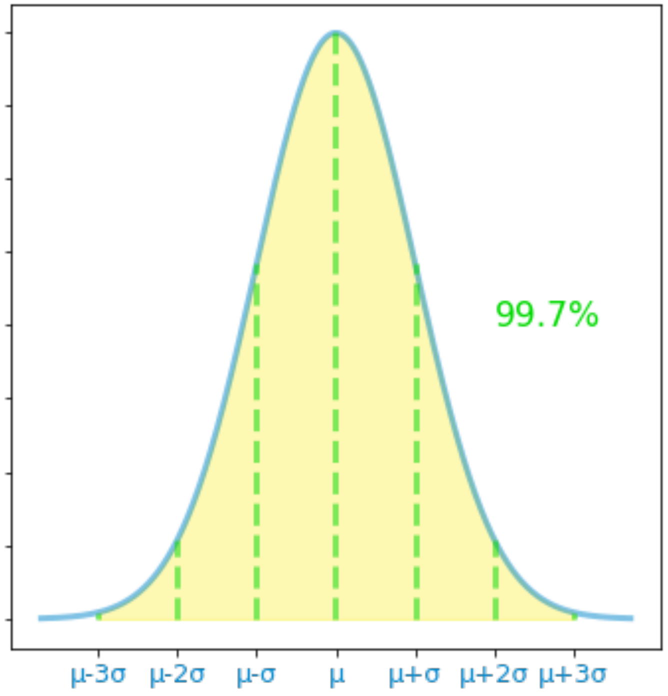
$\mathbb{D}X = \lambda$

При больших λ X $\implies$ N(λ, λ)

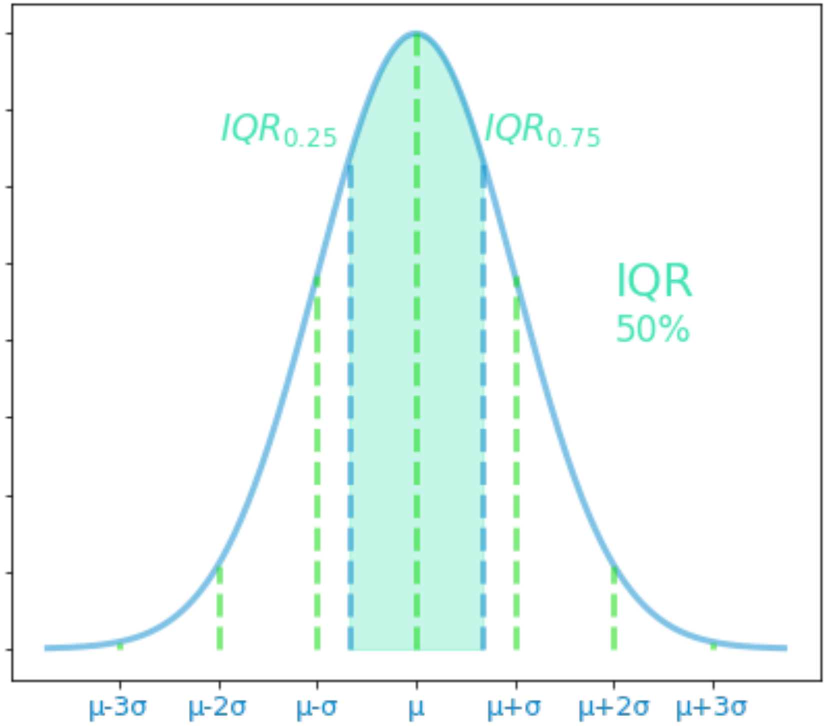
Важные характеристики распределений

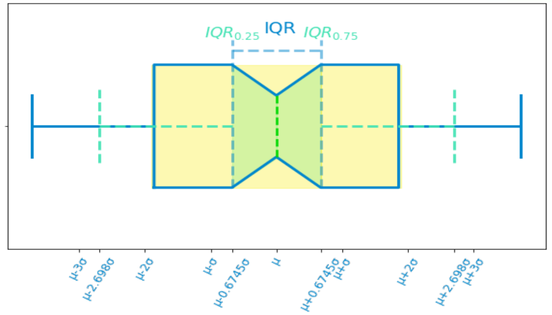


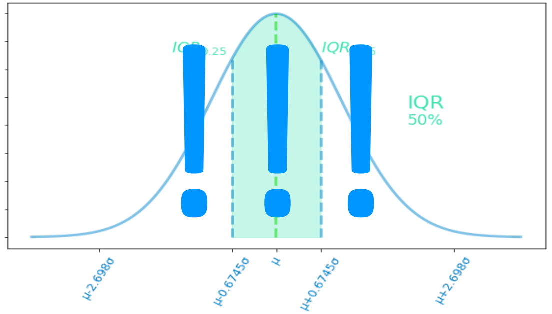


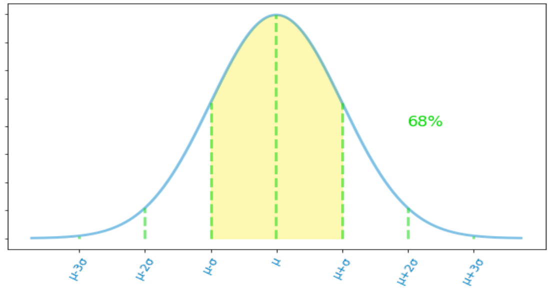


Важные характеристики распределений









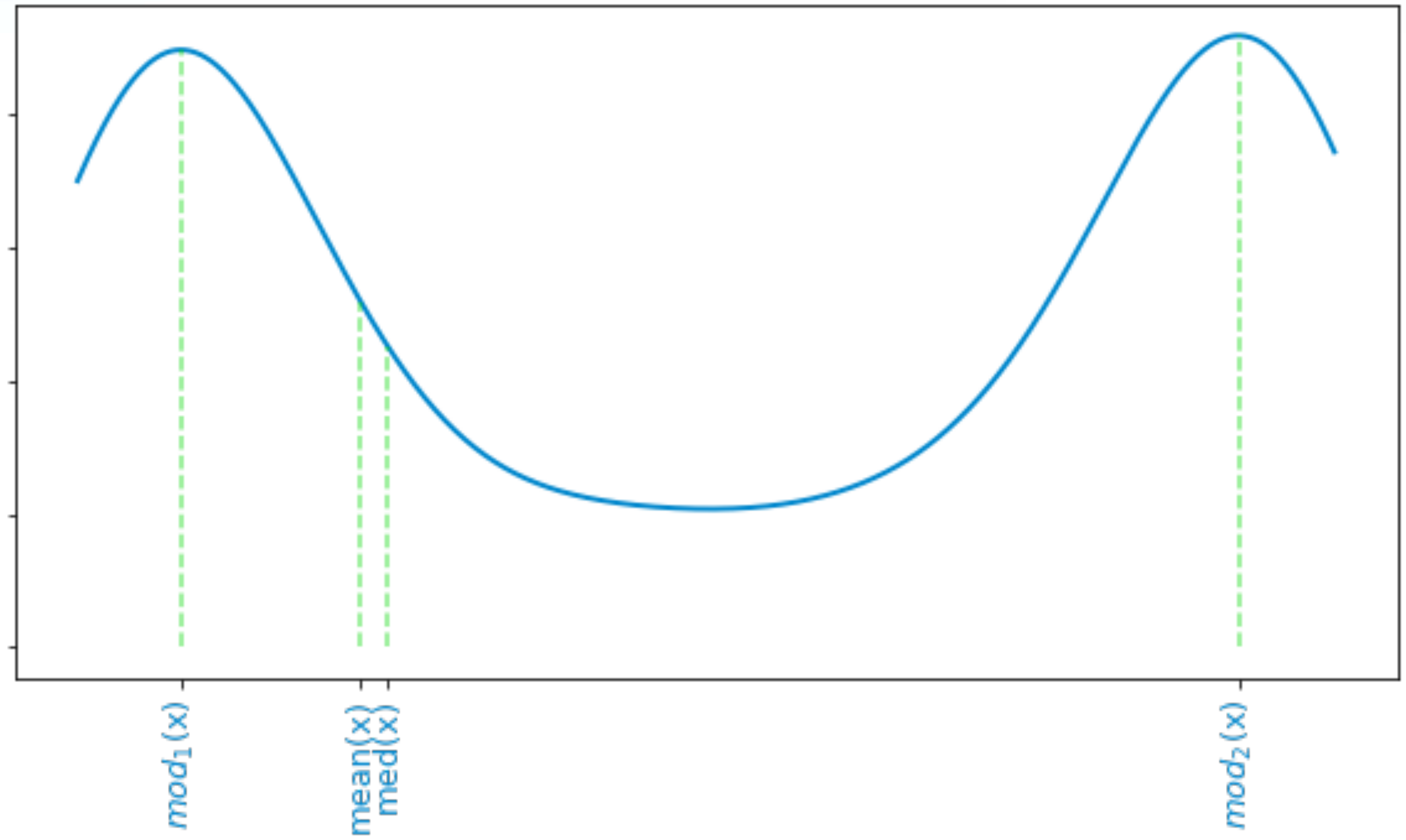
Важные характеристики распределений

Why not to trust statistics?

pandas.DataFrame.describe ???
scipy.stats???
numpy???

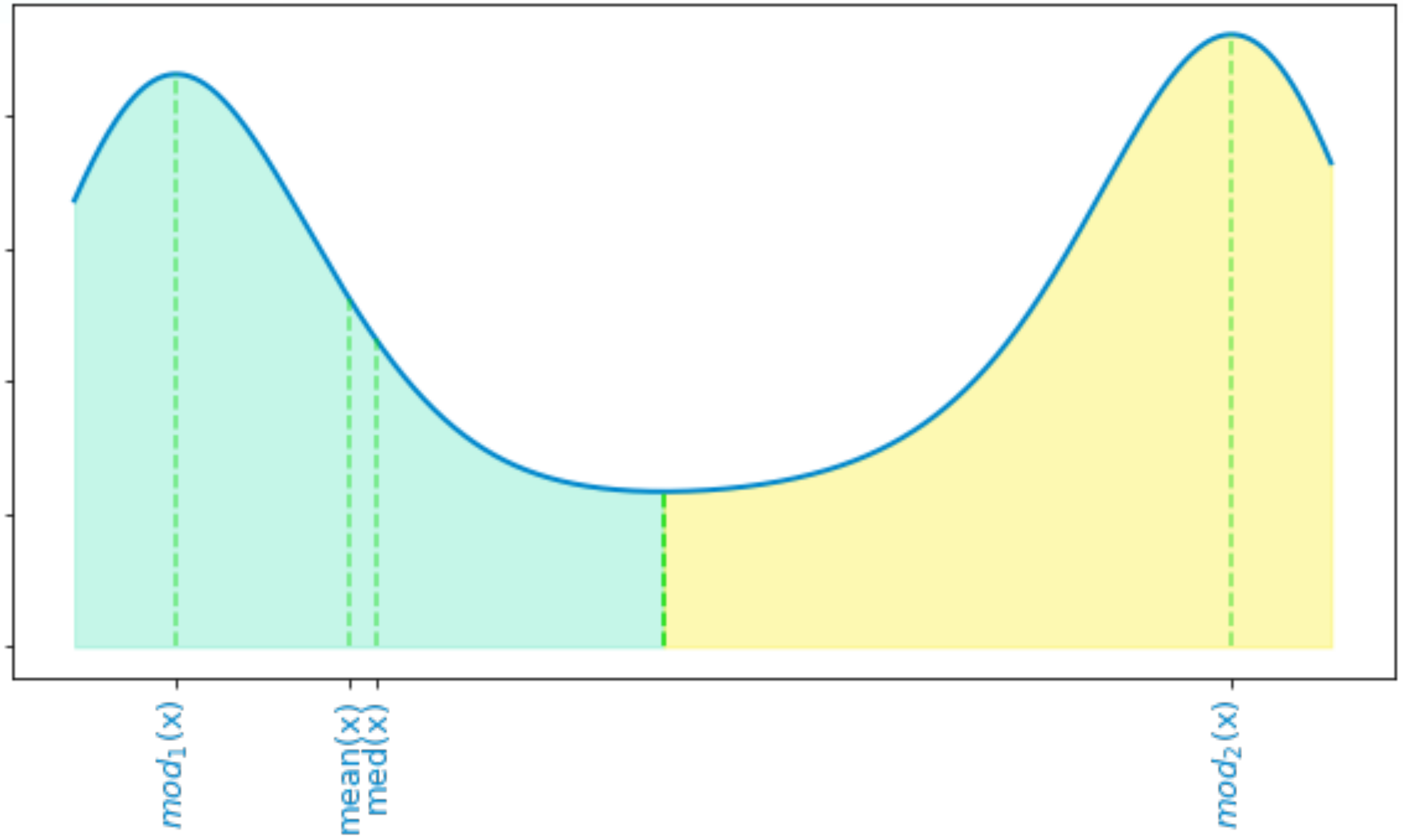
count	1.000000e+03	
mean	1.446043e-14	
std	2.150399e+00	
min	-3.719016e+00	
25%	-1.859508e+00	**= min/2 !?** ***(-0.6745σ)***
50%	1.421085e-14	**≈ mean**
75%	1.859508e+00	**= max/2 !?** ***(+0.6745σ)***
max	3.719016e	

Важные характеристики распределений



Как пользоваться?

Важные характеристики распределений



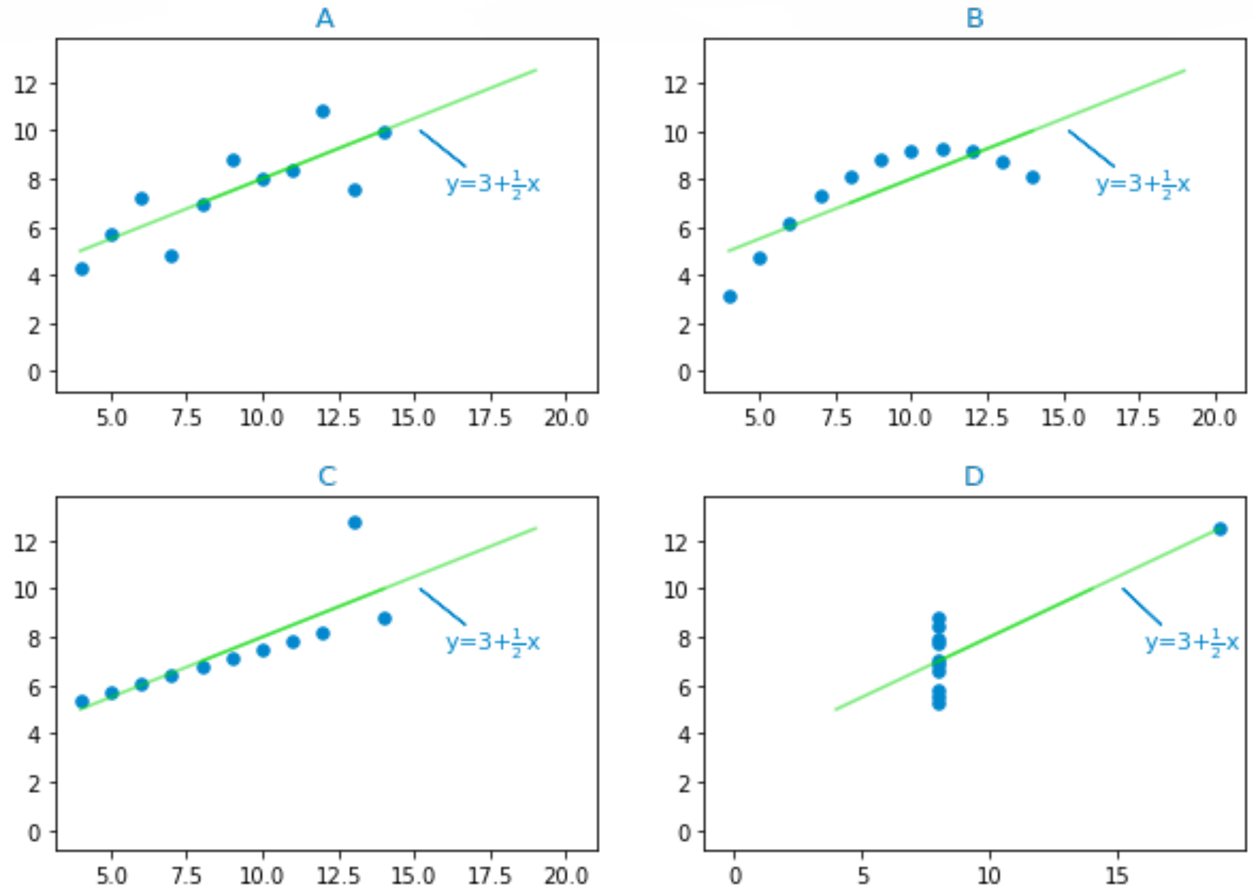
Сегментировать!

Важные характеристики распределений

Квартет Анскомбе	A	B	C	D
Среднее значение x	9.00	9.00	9.00	9.00
Дисперсия x	11.00	11.00	11.00	11.00
Среднее значение y	7.50	7.50	7.50	7.50
Дисперсия y	4.21	4.21	4.21	4.21
R^2	0.67	0.67	0.67	0.67
Прямая линейной регрессии	y = 3 + 0.5 x	y = 3 + 0.5 x	y = 3 + 0.5 x	y = 3 + 0.5 x

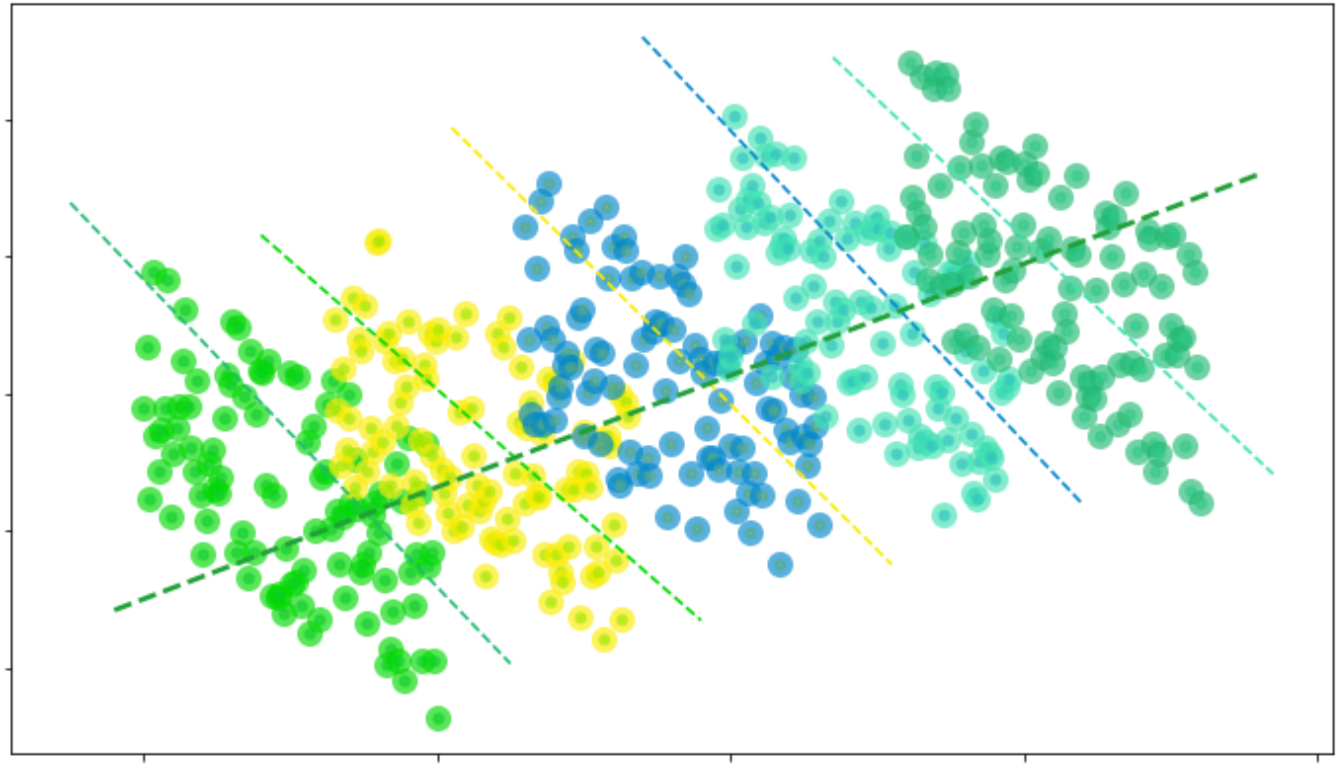
Важные характеристики распределений

Квартет Анскомбе



Важные характеристики распределений

Парадокс Симпсона



Важные характеристики распределений

Всегда смотрите на гистограммы/функции вероятности!

Характеристики могут привести в заблуждение!

Но, если мы работаем в одном и том же пространстве событий, то характеристики помогают сравнивать различные подходы, показатели, препараты, продукты...

Важные характеристики распределений

Выборочное среднее считается по такой формуле

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Важные характеристики распределений

Выборочная медиана, необходимо просто отсортировать выборку и выбрать значение в середине

$$X_n = (X_1, X_2, \dots, X_n)$$

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

$$med = \begin{cases} X_{(k+1)}, n = 2k + 1 \\ \dfrac{X_{(k)} + X_{(k+1)}}{2}, n = 2_k \end{cases}$$

Важные характеристики распределений

Выборочная дисперсия считается по такой формуле, деляя на n-1, а не на n мы получаем так называемую “несмещенную оценку” — это точечная оценка, математическое ожидание которой равно оцениваемому параметру.

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Colab? Colab!

Предсказательный интервал

Зная как распределена случайная величина X, мы можем понять в каком диапазоне она скорее всего окажется

$$\mathbb{P}(X_{\frac{\alpha}{2}} \leq X \leq X_{1-\frac{\alpha}{2}})$$

Доверительные интервалы

$$\mathbb{P}(C_L \leq \theta \leq CR) \geq 1 - \alpha$$

Левый предел

Оцениваемый параметр

Правый предел

Уровень доверия

Технология

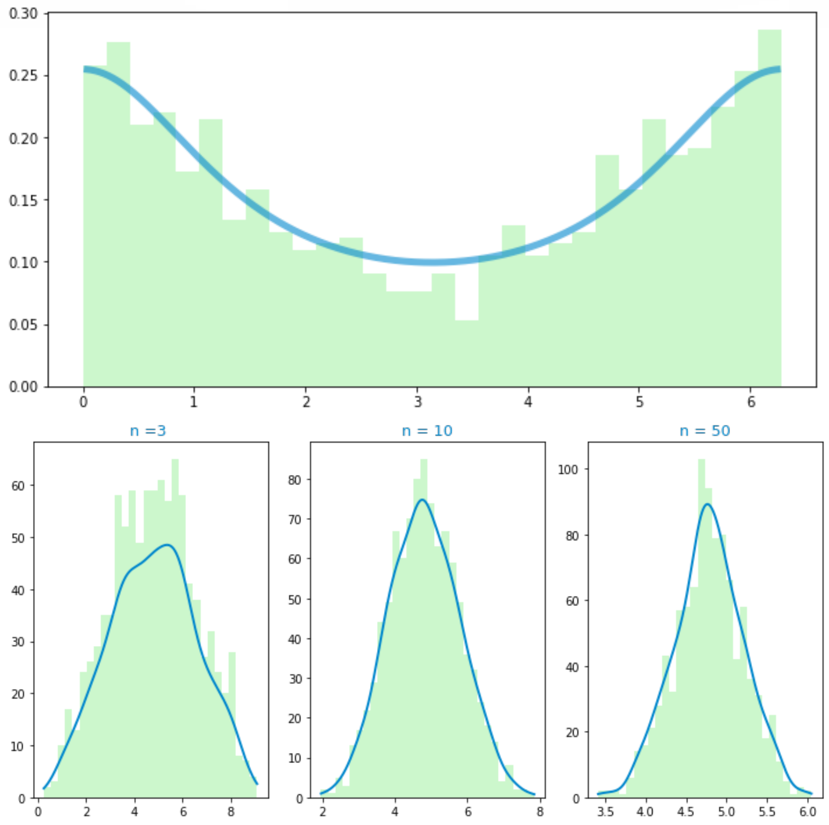
Зная как распределена статистика, мы, используя алгебраические преобразования можем понять в каком диапазоне будет изменяться неизвестный параметр. Диапазон при этом чаще всего задается квантилями распределений и другими статистиками по выборке

Центральная предельная теорема

https://habr.com/ru/post/471198/

http://datascientist.one/central-limit-theorem/

https://www.youtube.com/watch?v=InXimz8zikc



Центральная предельная теорема

Распределение выборочного среднего набора независимых одинаково распределенных случайных величин хорошо приближается нормальным распределением:

$$\bar{X}_n \approx \sim \mathcal{N}\left(\mathbb{E}X, \frac{\mathbb{D}X}{n}\right)$$

Центральная предельная теорема

Предсказательный интервал для $\bar{X}$

$$P(\mu - z_{1\,-\,\alpha/_2}\frac{\sigma}{\sqrt{n}} \leq \overline{X_{\mathrm{n}}} \leq \mu + z_{1\,-\,\alpha/_2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Доверительный интервал для μ

$$P(\overline{X_{\mathrm{n}}} - z_{1\,-\,\alpha/_2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X_{\mathrm{n}}} + z_{1\,-\,\alpha/_2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

СБЕР

Центральная предельная теорема

Доверительный интервал для $\mathbb{E}X$

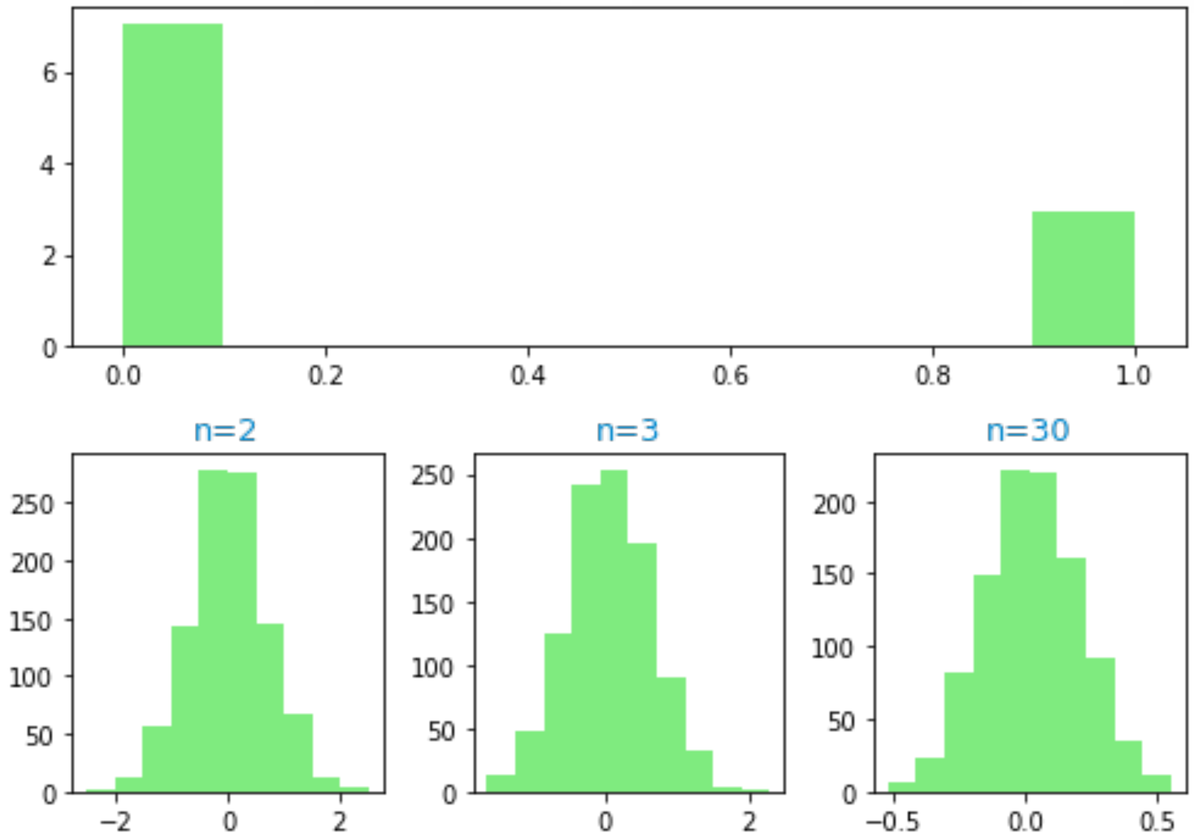
$$P(\overline{X_{\mathrm{n}}} - z_{1\,-\,{}^{\alpha}/_{2}}\sqrt{\frac{\mathbb{D}X}{n}} \le \mathbb{E}X \le \overline{X_{\mathrm{n}}} + z_{1\,-\,{}^{\alpha}/_{2}}\sqrt{\frac{\mathbb{D}X}{n}}) = 1 - \alpha$$

Центральная предельная теорема

Распределение Бернулли

$X \sim Ber(p)$

$P(X = 1) = p$
$P(X = 0) = 1 - p = q$



Центральная предельная теорема

Распределение Бернулли

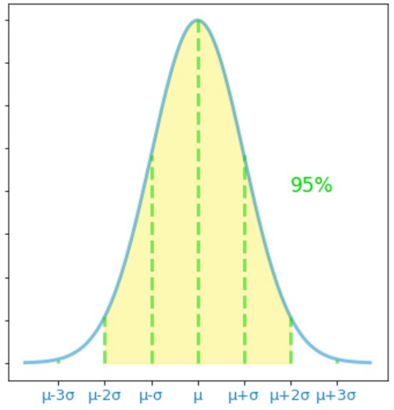
$$\bar{p} \approx \sim N(\mathbb{E}X, \frac{\mathbb{D}X}{n})$$

$X \sim Ber(p) \Longrightarrow \mathbb{E}X = p, \mathbb{D}X = p(1-p)$

$$P(\bar{p} - 2\sqrt{\frac{\bar{p}(1-\bar{p})}{n}} \leq \bar{p} \leq \bar{p} + 2\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}) \approx 95\%$$

$$\bar{p} \approx \sim N(p, \frac{p(1-p)}{n})$$

$$\sigma = \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$$



Пример:

Эксперимент 1: 30 подбрасываний, $\bar{p} = 0.467$

95% доверительный интервал: **0.285…0.649**

Эксперимент 2: 700 подбрасываний, $\bar{p} = 0.556$

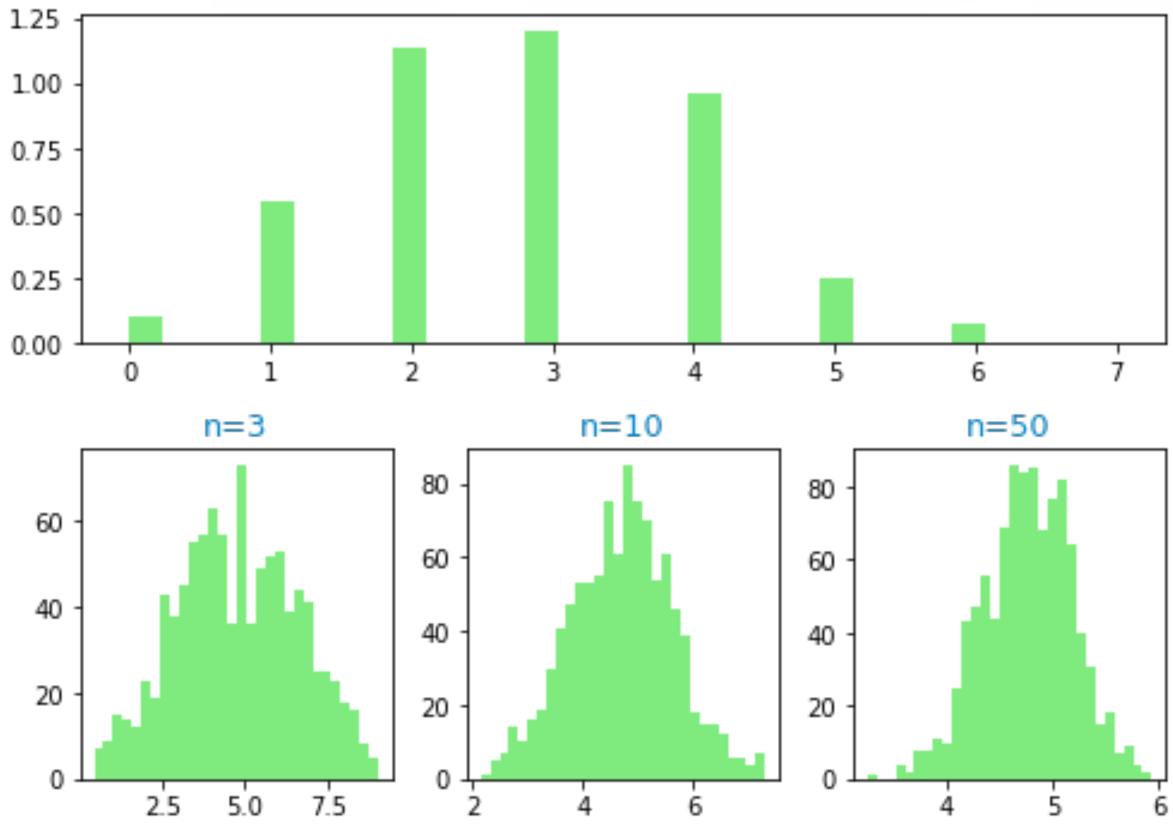
95% доверительный интервал: **0.519…0.594**

Центральная предельная теорема

Биномиальное распределение

$P(X = n) = p^n$ *

$P(X = k) = C_n^k p^k (1 - p)^{n-k}$ **



* вероятность попасть n-раз

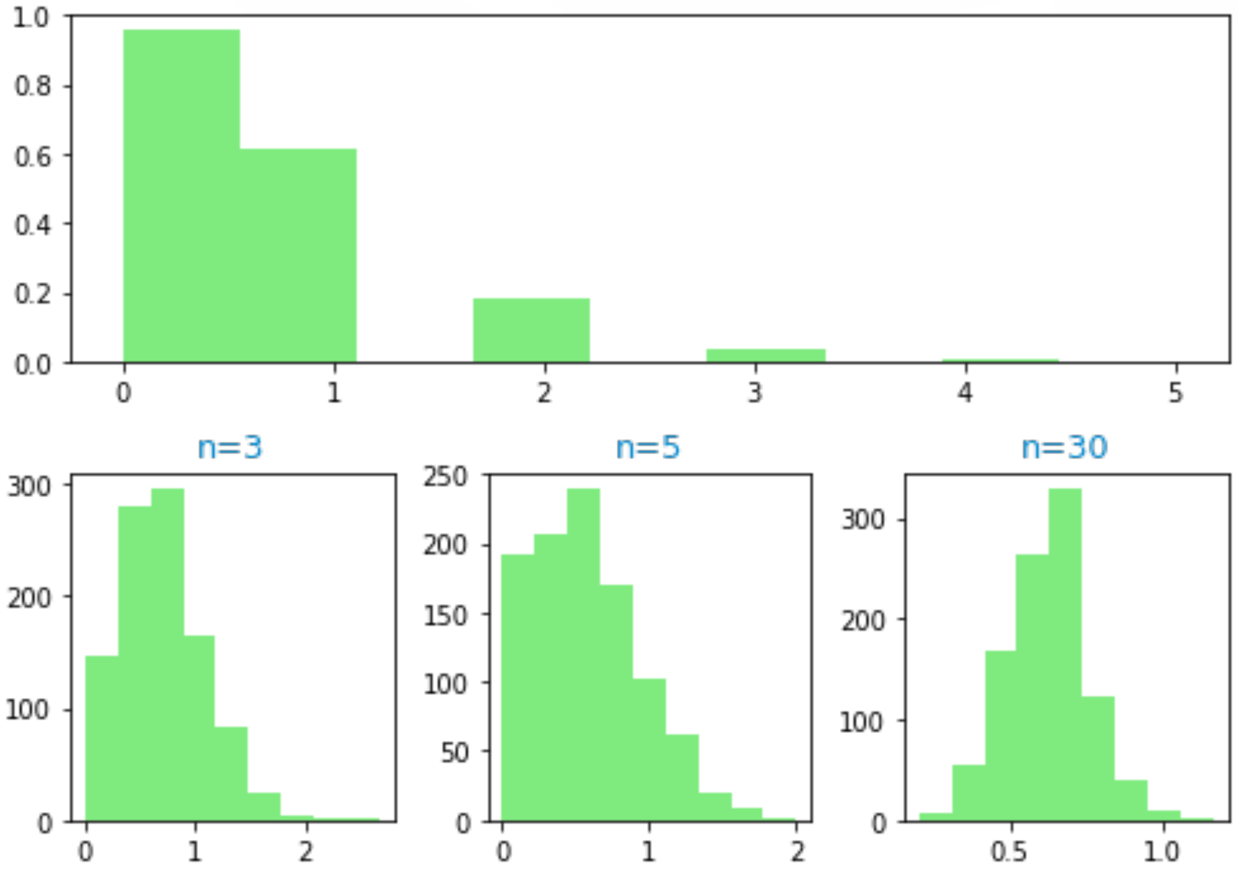
** вероятность попасть k-раз из n

Центральная предельная теорема

Биноминальное распределение

$P(X = n) = p^n$ *

$P(X = k) = C_n^k p^k (1 - p)^{n-k}.$ **

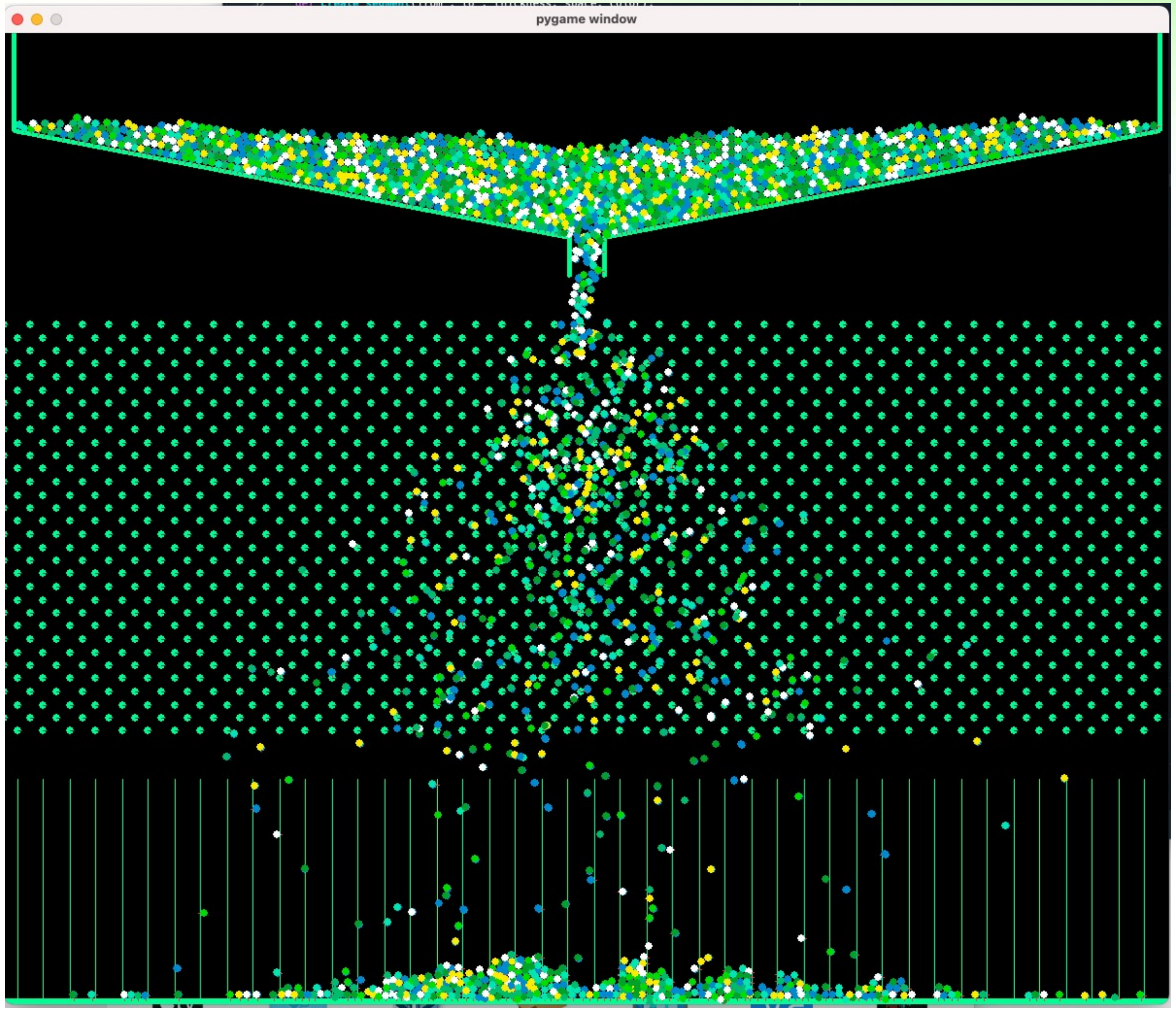


* вероятность попасть n-раз
** вероятность попасть k-раз из n

ЦПТ. Машина Гальтона

$2^0 = 1$
$2^1 = 2$
$2^2 = 4$
$2^3 = 6$
$2^4 = 16$
$2^5 = 32$

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1



Colab? Colab!

Резюме

- Узнали как оценивать распределение по выборке
- Рассмотрели важные характеристики распределений
- Посмотрели на важные статистики
- Узнали что такое Центральная предельная теорема и чем она может быть полезна

Обратная связь

?

Спасибо за внимание!