

Симуляции, ЗБЧ и ЦПТ



План

- Основные распределения, встречающиеся в математической статистике



План

- Основные распределения, встречающиеся в математической статистике
- Закон больших чисел и Центральная предельная теорема



План

- Основные распределения, встречающиеся в математической статистике
- Закон больших чисел и Центральная предельная теорема
- Какими бывают сходимости случайных величин



План

- Основные распределения, встречающиеся в математической статистике
- Закон больших чисел и Центральная предельная теорема
- Какими бывают сходимости случайных величин
- Метод Монте-Карло и симуляции



План

- Основные распределения, встречающиеся в математической статистике
- Закон больших чисел и Центральная предельная теорема
- Какими бывают сходимости случайных величин
- Метод Монте-Карло и симуляции
- Квантильное преобразование



Распределения, связанные с нормальным



Нормальное распределение

Обозначение:

$$X \sim N(\mu, \sigma^2)$$

Плотность:

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

Характеристики:

$$\mathbb{E}(X) = \mu, \text{Var}(X) = \sigma^2$$



Нормальное распределение

Обозначение:

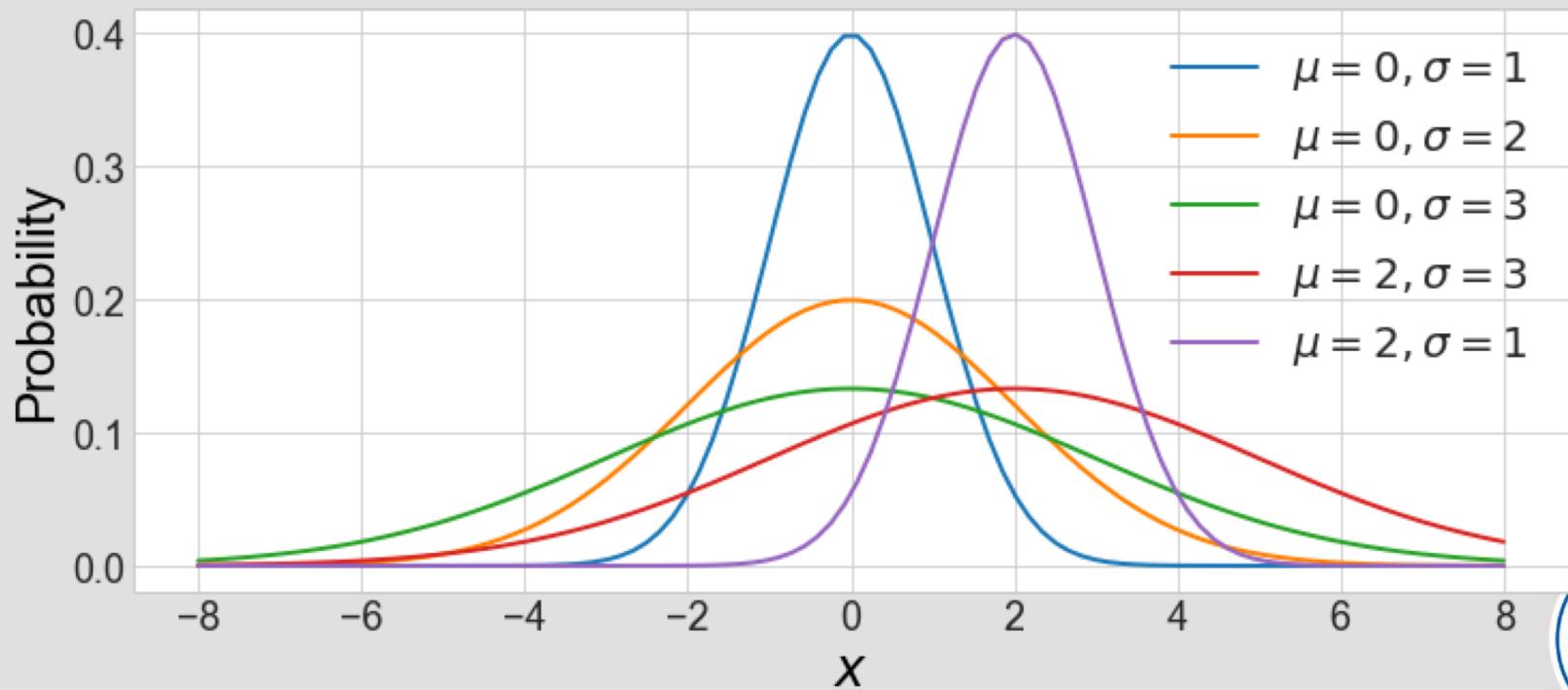
$$X \sim N(\mu, \sigma^2)$$

Плотность:

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

Характеристики:

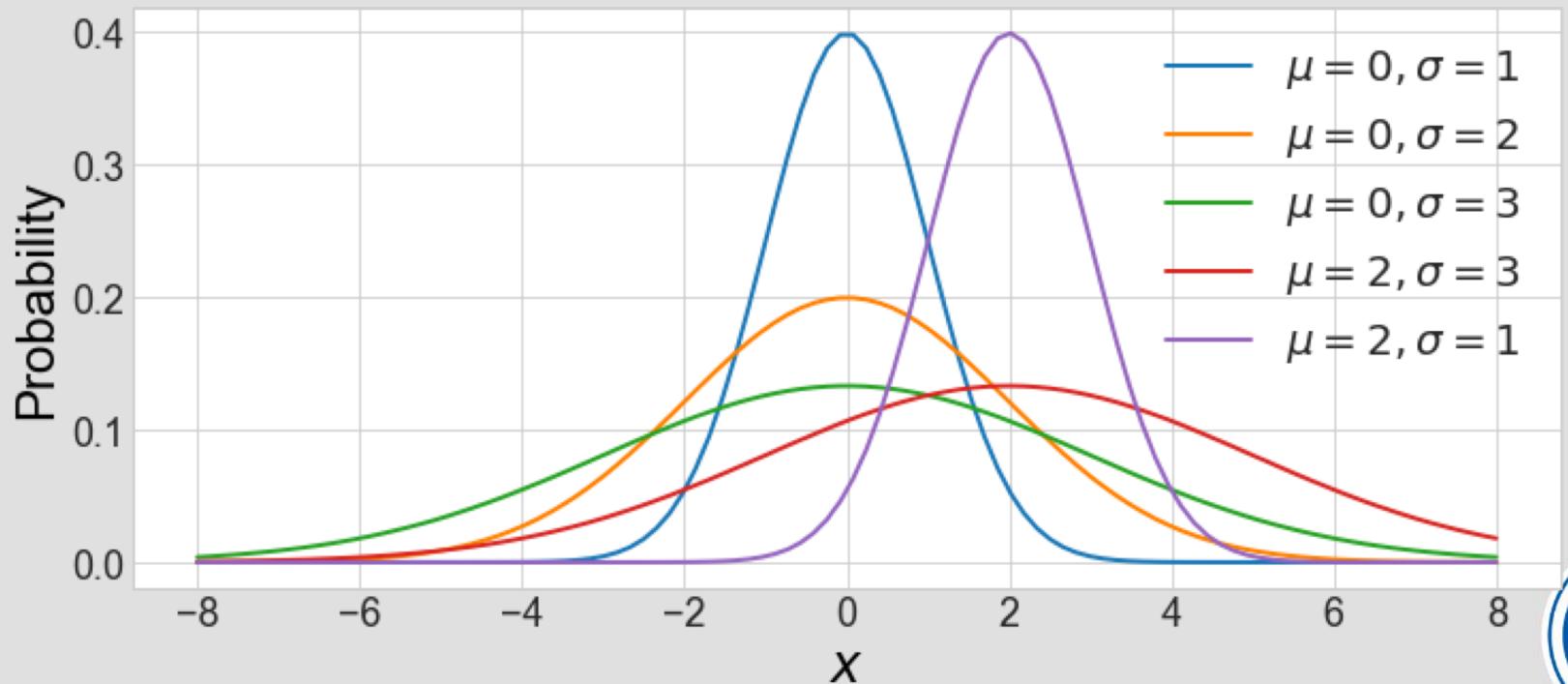
$$\mathbb{E}(X) = \mu, \text{Var}(X) = \sigma^2$$



Нормальное распределение

Когда возникает на практике: по центральной предельной теореме (изучим её на этой неделе) любое среднее имеет асимптотически-нормальное распределение.

На практике мы часто будем работать со средними.



Распределение хи-квадрат

Случайные величины $X_1, \dots, X_k \sim iid N(0,1)$.



Распределение хи-квадрат

Случайные величины $X_1, \dots, X_k \sim iid N(0,1)$.

Случайная величина $Y = X_1^2 + \dots + X_k^2 \sim \chi_k^2$ имеет “хи-квадрат” распределение с k степенями свободы



Распределение хи-квадрат

Случайные величины $X_1, \dots, X_k \sim iid N(0,1)$.

Случайная величина $Y = X_1^2 + \dots + X_k^2 \sim \chi_k^2$ имеет “хи-квадрат” распределение с k степенями свободы

- *iid* расшифровывается как *identically independently distributed* (независимы и одинаково распределены)



Распределение хи-квадрат

Случайные величины $X_1, \dots, X_k \sim iid N(0,1)$.

Случайная величина $Y = X_1^2 + \dots + X_k^2 \sim \chi_k^2$ имеет “хи-квадрат” распределение с k степенями свободы

 Когда возникает
на практике:

$$\hat{\sigma}^2 = \overline{x^2} - \bar{x}^2$$

► *iid* расшифровывается как *identically independently distributed* (независимы и одинаково распределены)



Распределение хи-квадрат

Случайные величины $X_1, \dots, X_k \sim iid N(0,1)$.

Случайная величина $Y = X_1^2 + \dots + X_k^2 \sim \chi_k^2$ имеет “хи-квадрат” распределение с k степенями свободы

- ✓ Когда возникает на практике:

$$\hat{\sigma}^2 = \bar{x^2} - \bar{x}^2$$

- Если выборка пришла из $N(0,1)$, величина $\bar{x^2}$ будет иметь “хи-квадрат” распределение



Распределение хи-квадрат

Случайные величины $X_1, \dots, X_k \sim iid N(0,1)$.

Случайная величина $Y = X_1^2 + \dots + X_k^2 \sim \chi_k^2$ имеет “хи-квадрат” распределение с k степенями свободы

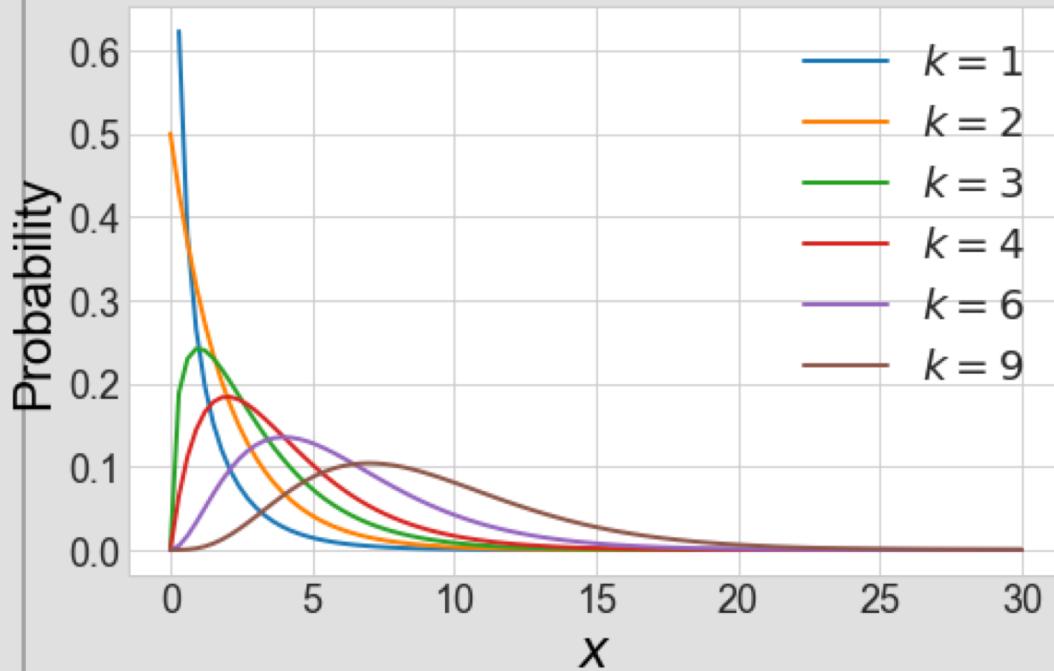
- ✓ Когда возникает на практике:

$$\hat{\sigma}^2 = \bar{x^2} - \bar{x}^2$$

- Если выборка пришла из $N(0,1)$, величина $\bar{x^2}$ будет иметь “хи-квадрат” распределение
- Для выборочной дисперсии тоже можно получить “хи-квадрат” распределение



Распределение хи-квадрат



Плотность:

$$f(x) = \frac{1}{\frac{k}{2^2} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot x^{\frac{k}{2}-1} \cdot e^{-\frac{x}{2}}, x \geq 0$$

$$X_1, \dots, X_k \sim iid N(0,1)$$

$$Y = X_1^2 + \dots + X_k^2 \sim \chi_k^2$$

Из-за квадратов
принимает только
положительные
значения

Характеристики:

$$\mathbb{E}(X) = k$$

$$\text{Var}(X) = 2k$$



Степени свободы

Число степеней свободы – количество элементов варьирования, которые могут принимать произвольные значения, не изменяющие заданных характеристик.



Степени свободы

Число степеней свободы – количество элементов варьирования, которые могут принимать произвольные значения, не изменяющие заданных характеристик.

Пример 1: Дано 7 чисел со средним 5. Нужно подобрать другие 7 чисел со средним 5.

Произвольно можем выбрать только 6 чисел. Число степеней свободы здесь равно $7 - 1 = 6$.



Степени свободы

Число степеней свободы – количество элементов варьирования, которые могут принимать произвольные значения, не изменяющие заданных характеристик.

Пример 2: При вычислении дисперсии по выборке из n наблюдений число степеней свободы равно $n - 1$, так как 1 степень свободы используется при расчёте среднего.



Распределение Стьюдента

Независимые случайные величины $X_0 \sim N(0,1)$, $Y \sim \chi^2_k$.



Распределение Стьюдента

Независимые случайные величины $X_0 \sim N(0,1)$, $Y \sim \chi^2_k$.

Тогда случайная величина

$$Z = \frac{X_0}{\sqrt{Y/k}} \sim t(k)$$

имеет распределение Стьюдента с k степенями свободы.



Распределение Стьюдента

Независимые случайные величины $X_0 \sim N(0,1)$, $Y \sim \chi^2_k$.

Тогда случайная величина

$$Z = \frac{X_0}{\sqrt{Y/k}} \sim t(k)$$

имеет распределение Стьюдента с k степенями свободы.

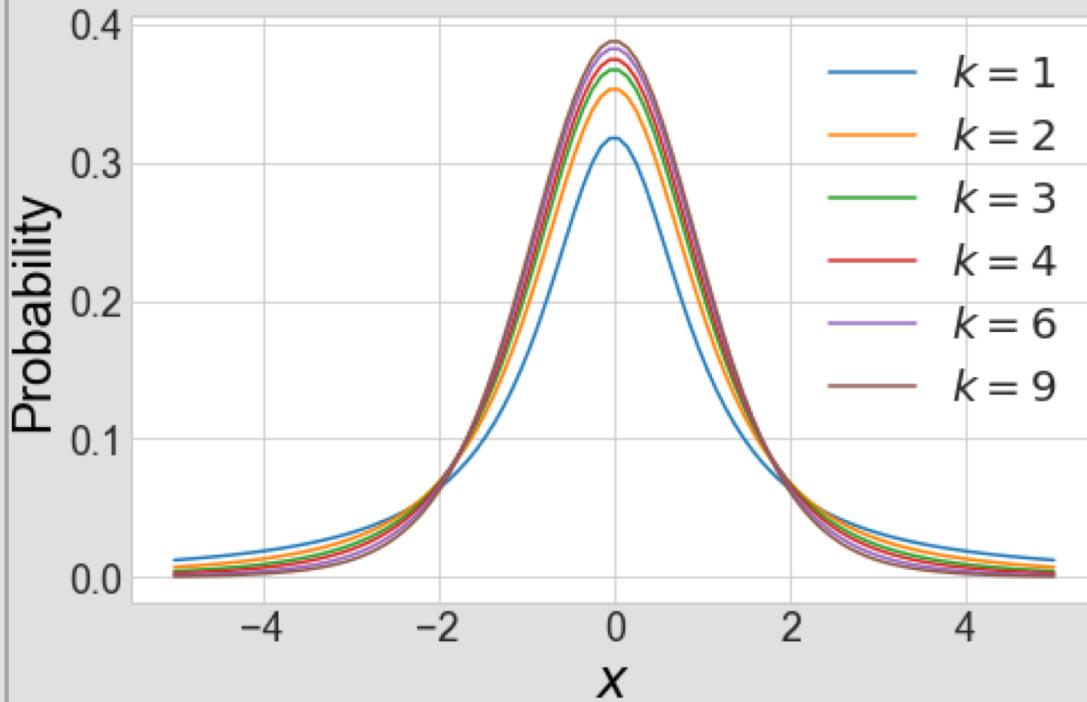


Когда возникает на практике:

Мы будем часто встречаться с выражением $\frac{\bar{x}}{\sqrt{\frac{\hat{\sigma}^2}{n}}}$,
имеющим распределение Стьюдента



Распределение Стьюдента



Плотность:

$$f(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}}$$

$$X_0 \sim N(0,1), Y \sim \chi_k^2,$$

$$Z = \frac{X_0}{\sqrt{Y/k}} \sim t(k)$$

Характеристики:

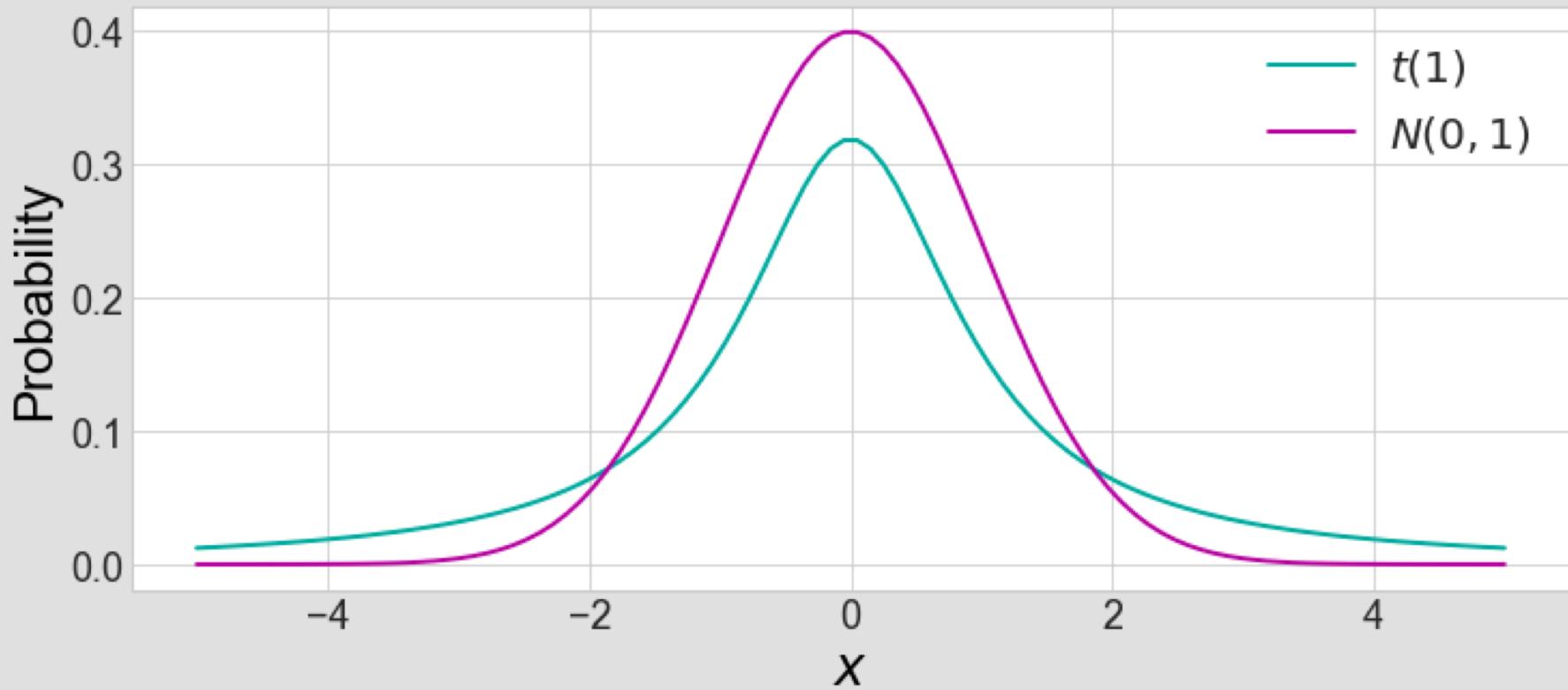
$$\mathbb{E}(Z) = 0$$

$$\text{Var}(Z) = \frac{k}{k-2}, k > 2$$



Тяжёлые хвосты

Распределение Стьюдента обладает более тяжёлыми хвостами, нежели нормальное



Распределение Фишера

Независимые случайные величины $X \sim \chi^2_k$, $Y \sim \chi^2_m$.



Распределение Фишера

Независимые случайные величины $X \sim \chi^2_k$, $Y \sim \chi^2_m$.

Случайная величина

$$Z = \frac{\sqrt{X/k}}{\sqrt{Y/m}} \sim F(k, m)$$

имеет распределение Фишера с k, m степенями свободы.



Распределение Фишера

Независимые случайные величины $X \sim \chi^2_k$, $Y \sim \chi^2_m$.

Случайная величина

$$Z = \frac{\sqrt{X/k}}{\sqrt{Y/m}} \sim F(k, m)$$

имеет распределение Фишера с k, m степенями свободы.

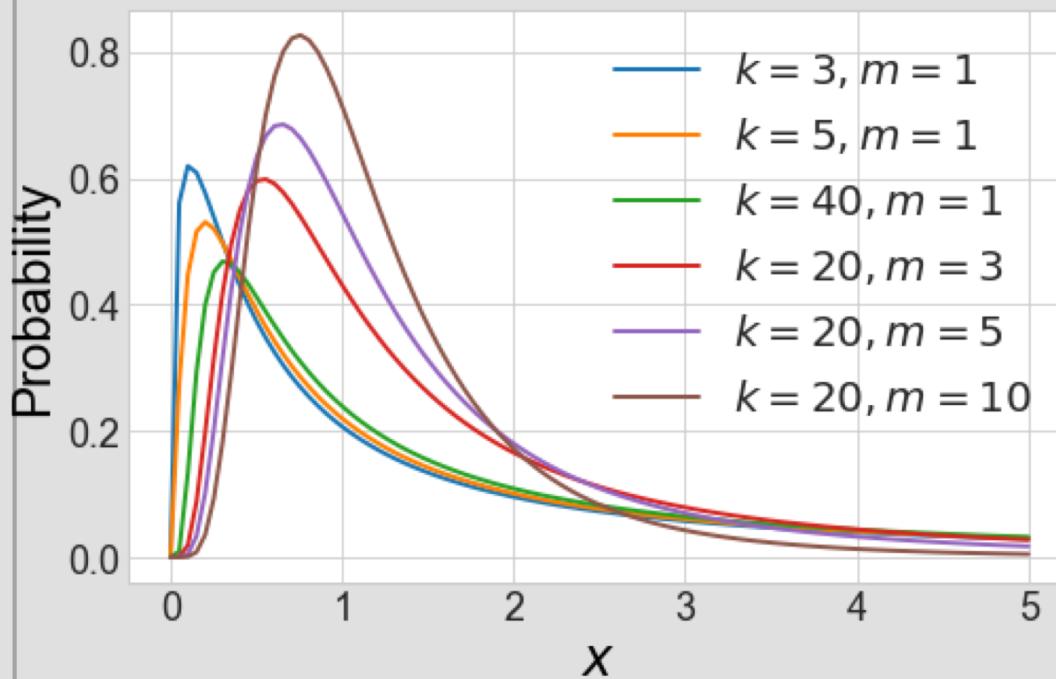


Когда возникает на практике:

Встречается при сравнении дисперсий.
Чтобы сравнить их между собой, одну
дисперсию делят на вторую.



Распределение Фишера



Характеристики:

$$\mathbb{E}(X) = \frac{m}{m-2}, m > 2$$

$$\text{Var}(X) = \frac{2m^2(k+m-2)}{n(m-2)^2(m-4)}$$

$$X \sim \chi_k^2, Y \sim \chi_m^2$$

$$Z = \frac{\sqrt{X/k}}{\sqrt{Y/m}} \sim F(k, m)$$

Из-за квадратов
принимает только
положительные значения

Плотность:

Очень громоздкая



Резюме

- Распределения хи-квадрат, Стьюдента, Фишера часто встречаются на практике при анализе нормально распределённых выборок
- В будущем мы часто будем обращаться к ним за помощью



Закон больших чисел (ЗБЧ)



Как устроен мир



Сундук – различные процессы порождения данных. Теория вероятностей изучает этот сундук. В реальности мы не видим его.



Как устроен мир



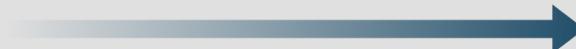
X

Сундук – различные процессы порождения данных. Теория вероятностей изучает этот сундук. В реальности мы не видим его.

Сундук порождает выборки. Математическая статистика изучает их и пытается восстановить внутренности сундука.



Как устроен мир



X

Сундук – различные процессы порождения данных. Теория вероятностей изучает этот сундук. В реальности мы не видим его.

Сундук порождает выборки. Математическая статистика изучает их, и пытается восстановить внутренности сундука.

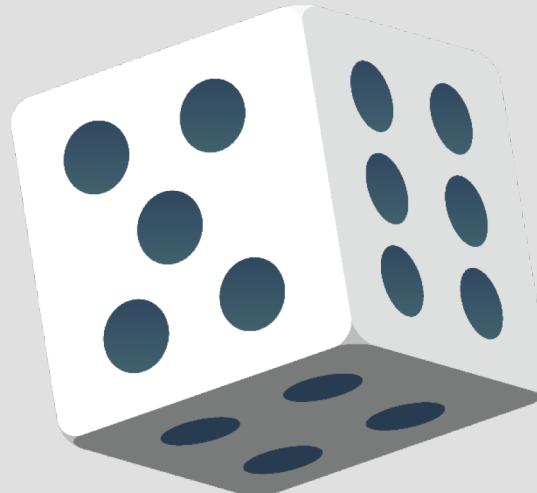
В этом нам помогает несколько теорем: ЗБЧ, ЦПТ и др.



Закон больших чисел (ЗБЧ)

ЗБЧ говорит, что среднее арифметическое большого числа похожих случайных величин “стабилизируется” с ростом их числа

Пример: Игровая кость



Закон больших чисел (ЗБЧ)

ЗБЧ говорит, что среднее арифметическое большого числа похожих случайных величин “стабилизируется” с ростом их числа



Слабая форма ЗБЧ (Чебышёв)

Теорема:

Пусть X_1, \dots, X_n попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $Var(X_1) < \infty$ тогда:



Слабая форма ЗБЧ (Чебышёв)

Теорема:

Пусть X_1, \dots, X_n попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $\text{Var}(X_1) < \infty$ тогда:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1)$$



Слабая форма ЗБЧ (Чебышёв)

Теорема:

Пусть X_1, \dots, X_n попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $\text{Var}(X_1) < \infty$ тогда:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1)$$

Среднее сходится по вероятности к математическому ожиданию при $n \rightarrow \infty$



Слабая форма ЗБЧ (Чебышёв)

Простым языком:

- Среднее арифметическое большого числа похожих случайных величин “стабилизируется” с ростом их числа



Слабая форма ЗБЧ (Чебышёв)

Простым языком:

- Среднее арифметическое большого числа похожих случайных величин “стабилизируется” с ростом их числа
- Среднее для бесконечного числа случайных величин неслучайно



Слабая форма ЗБЧ (Чебышёв)

Простым языком:

- Среднее арифметическое большого числа похожих случайных величин “стабилизируется” с ростом их числа
- Среднее для бесконечного числа случайных величин неслучайно
- Если у нас есть страховая фирма, мы можем заработать немного денег (самая простая формулировка)



Страховка

Вероятность того, что на машину во дворе упадёт дерево составляет **0.01**. Страховка в год стоит **100** рублей. В случае падения клиенту выплачивается **11000** рублей. Какой будет средняя прибыль компании с одной страховки?



Страховка

Вероятность того, что на машину во дворе упадёт дерево составляет **0.01**. Страховка в год стоит **100** рублей. В случае падения клиенту выплачивается **11000** рублей. Какой будет средняя прибыль компании с одной страховки?

X_i – прибыль с одного человека

\bar{X} – средняя прибыль компании



Страховка

Вероятность того, что на машину во дворе упадёт дерево составляет **0.01**. Страховка в год стоит **100** рублей. В случае падения клиенту выплачивается **11000** рублей. Какой будет средняя прибыль компании с одной страховки?

X_i – прибыль с одного человека

\bar{X} – средняя прибыль компании



Страховка

Вероятность того, что на машину во дворе упадёт дерево составляет **0.01**. Страховка в год стоит **100** рублей. В случае падения клиенту выплачивается **11000** рублей. Какой будет средняя прибыль компании с одной страховки?

X_i – прибыль с одного человека

\bar{X} – средняя прибыль компании



$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1) = 100 \cdot 0.99 - 10900 \cdot 0.01 = -10$$



Страховка

Вероятность того, что на машине в саду упадёт дерево составляет 0.01. Страховка защищает от 100 рублей. В случае падения клиенту выплачиваются 10000 рублей. Какой будет средняя прибыль компании от такой страховки?



Денег мы
не получим

X_i – прибыль с одного человека

\bar{X} – средняя прибыль компании

X_i	100	-10900
$\mathbb{P}(X_i = k)$	0.99	0.01

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1) = 100 \cdot 0.99 - 10900 \cdot 0.01 = -10$$



Вопрос про больницы

- Есть две больницы: большая и маленькая.



Вопрос про больницы

- Есть две больницы: большая и маленькая.
- В обеих принимают роды. Выяснилось, что в одной из них оценка вероятности появления мальчика составила 0.7.



Вопрос про больницы

- Есть две больницы: большая и маленькая.
- В обеих принимают роды. Выяснилось, что в одной из них оценка вероятности появления мальчика составила 0.7.
- В какой больнице это скорее всего произошло и почему?



Вопрос про больницы

Скорее всего это произошло в маленькой больнице.
При малых объемах выборки вероятность отклониться
от 0.5 больше. Именно об этом говорит нам ЗБЧ.



depositphotos.com



Некорректная работа при малых числах

- Данные часто поступают на обработку в агрегированной форме (по городам, по людям, по статьям из газет)

► <http://nsmn1.uh.edu/dgraur/niv/TheMostDangerousEquation.pdf>



Некорректная работа при малых числах

- Данные часто поступают на обработку в агрегированной форме (по городам, по людям, по статьям из газет)
- Для субъектов с маленьким числом наблюдений ЗБЧ не работает (города с маленьким населением)

► <http://nsmn1.uh.edu/dgraur/niv/TheMostDangerousEquation.pdf>



Некорректная работа при малых числах

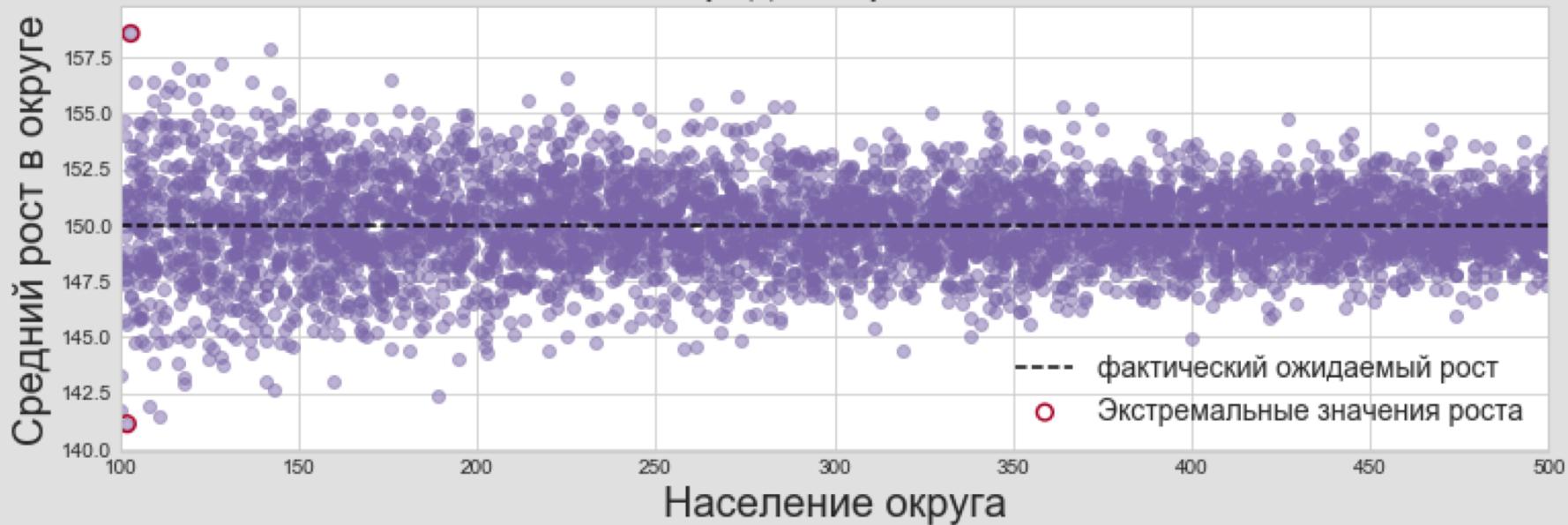
- Данные часто поступают на обработку в агрегированной форме (по городам, по людям, по статьям из газет)
- Для субъектов с маленьким числом наблюдений ЗБЧ не работает (города с маленьким населением)
- Среднее значение при маленьких выборках плохо отражает фактическое математическое ожидание

► <http://nsmn1.uh.edu/dgraur/niv/TheMostDangerousEquation.pdf>



Некорректная работа при малых числах

Зависимость среднего роста от населения



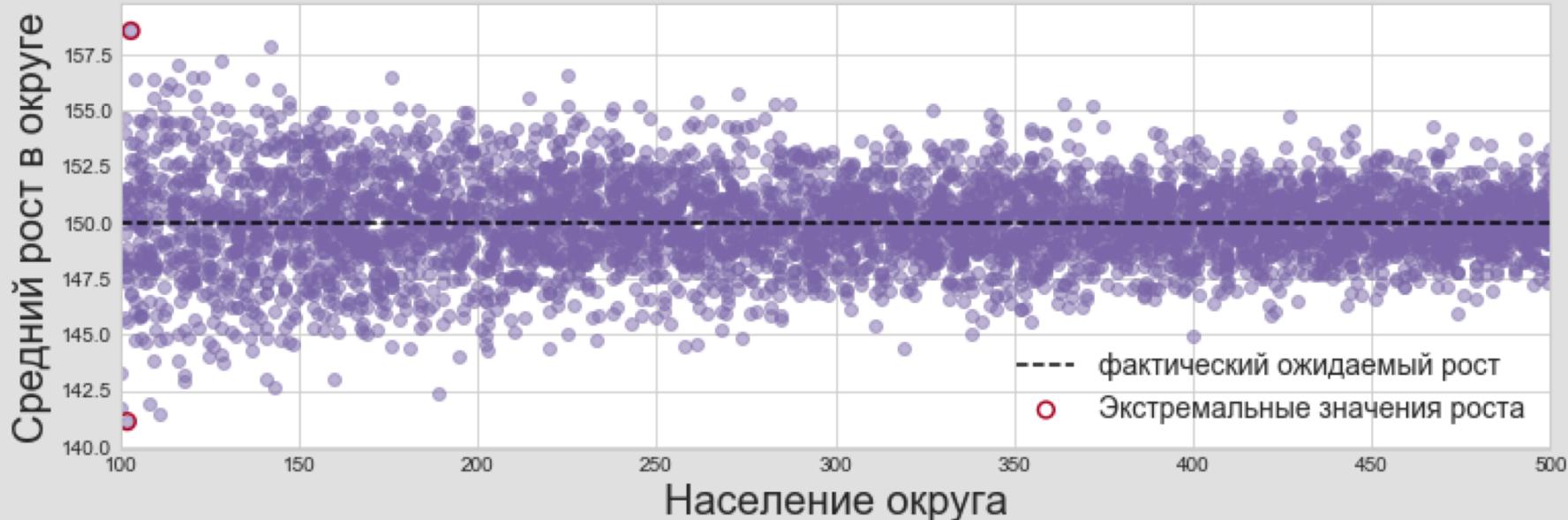
Если игнорировать размер округа, можно сказать, что округа с самым низким и высоким населением были верно обведены кружками

► <http://nsmn1.uh.edu/dgraur/niv/TheMostDangerousEquation.pdf>



Некорректная работа при малых числах

Зависимость среднего роста от населения



Если игнорировать размер округа, можно сказать, что округа с самым низким и высоким населением были верно обведены кружками

Это неверно, так как средний рост в этих районах считался по маленьким выборкам

► <http://nsmn1.uh.edu/dgraur/niv/TheMostDangerousEquation.pdf>



Резюме

ЗБЧ говорит, что при больших выборках и отсутствии аномалий среднее, рассчитанное по выборке, оказывается близким к теоретическому математическому ожиданию



Сходимость по вероятности



Слабая форма ЗБЧ (Чебышёв)

Теорема:

Пусть X_1, \dots, X_n попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $\text{Var}(X_1) < \infty$ тогда:

$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1)$$

Среднее сходится по вероятности к математическому ожиданию при $n \rightarrow \infty$



Сходимость по вероятности

Последовательность случайных величин X_1, \dots, X_n, \dots
сходится по вероятности к случайной величине X , если

$$\forall \varepsilon > 0 \quad \mathbb{P}(|X_n - X| < \varepsilon) \rightarrow 1 \text{ при } n \rightarrow \infty$$



Сходимость по вероятности

Последовательность случайных величин X_1, \dots, X_n, \dots
сходится по вероятности к случайной величине X , если

$$\forall \varepsilon > 0 \quad \mathbb{P}(|X_n - X| < \varepsilon) \rightarrow 1 \text{ при } n \rightarrow \infty$$



Сходимость по вероятности

Последовательность случайных величин X_1, \dots, X_n, \dots
сходится по вероятности к случайной величине X , если

$$\forall \varepsilon > 0 \quad \mathbb{P}(|X_n - X| < \varepsilon) \rightarrow 1 \text{ при } n \rightarrow \infty$$

То есть:

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$$



Сходимость по вероятности

Последовательность случайных величин X_1, \dots, X_n, \dots
сходится по вероятности к случайной величине X , если

$$\forall \varepsilon > 0 \quad \mathbb{P}(|X_n - X| < \varepsilon) \rightarrow 1 \text{ при } n \rightarrow \infty$$

То есть:

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$$



Обычно пишут:

$$X_n \xrightarrow{p} X \text{ при } n \rightarrow \infty \quad \text{либо} \quad \operatorname{plim}_{n \rightarrow \infty} X_n = X$$



Свойства сходимости по вероятности

Можно выносить константу за знак предела:

$$\operatorname{plim}_{n \rightarrow \infty} (c \cdot X_n) = c \cdot \operatorname{plim}_{n \rightarrow \infty} X_n, \quad c \in \mathbb{R}$$



Свойства сходимости по вероятности

Можно выносить константу за знак предела:

$$\operatorname{plim}_{n \rightarrow \infty} (c \cdot X_n) = c \cdot \operatorname{plim}_{n \rightarrow \infty} X_n, \quad c \in \mathbb{R}$$

Предел суммы – сумма пределов:

$$\operatorname{plim}_{n \rightarrow \infty} (X_n + Y_n) = \operatorname{plim}_{n \rightarrow \infty} X_n + \operatorname{plim}_{n \rightarrow \infty} Y_n$$



Свойства сходимости по вероятности

Можно выносить константу за знак предела:

$$\operatorname{plim}_{n \rightarrow \infty} (c \cdot X_n) = c \cdot \operatorname{plim}_{n \rightarrow \infty} X_n, \quad c \in \mathbb{R}$$

Предел суммы – сумма пределов:

$$\operatorname{plim}_{n \rightarrow \infty} (X_n + Y_n) = \operatorname{plim}_{n \rightarrow \infty} X_n + \operatorname{plim}_{n \rightarrow \infty} Y_n$$

Предел произведения – произведение пределов:

$$\operatorname{plim}_{n \rightarrow \infty} (X_n \cdot Y_n) = \operatorname{plim}_{n \rightarrow \infty} X_n \cdot \operatorname{plim}_{n \rightarrow \infty} Y_n$$



Свойства сходимости по вероятности

Можно выносить константу за знак предела:

$$\operatorname{plim}_{n \rightarrow \infty} (c \cdot X_n) = c \cdot \operatorname{plim}_{n \rightarrow \infty} X_n, \quad c \in \mathbb{R}$$

Предел суммы – сумма пределов:

$$\operatorname{plim}_{n \rightarrow \infty} (X_n + Y_n) = \operatorname{plim}_{n \rightarrow \infty} X_n + \operatorname{plim}_{n \rightarrow \infty} Y_n$$

Предел произведения – произведение пределов:

$$\operatorname{plim}_{n \rightarrow \infty} (X_n \cdot Y_n) = \operatorname{plim}_{n \rightarrow \infty} X_n \cdot \operatorname{plim}_{n \rightarrow \infty} Y_n$$

Сходимость не портится из-за непрерывных функций

$$\operatorname{plim}_{n \rightarrow \infty} g(X_n) = g(\operatorname{plim}_{n \rightarrow \infty} X_n), \quad g(t) \text{ – непрерывная}$$



Резюме

В слабой форме ЗБЧ среднее сходится к математическому ожиданию по вероятности

Для сходимости по вероятности верны такие же арифметические свойства, как и для обычных пределов



Центральная предельная теорема (ЦПТ)



Центральная предельная теорема (ЦПТ)

ЦПТ говорит, что сумма довольно большого числа случайных величин имеет распределение близкое к нормальному



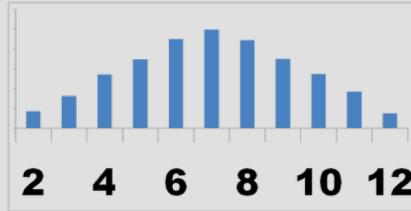
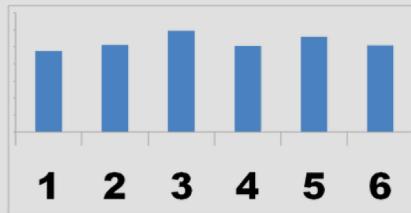
Центральная предельная теорема (ЦПТ)

ЦПТ говорит, что сумма довольно большого числа случайных величин имеет распределение близкое к нормальному



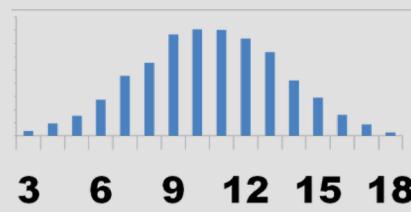
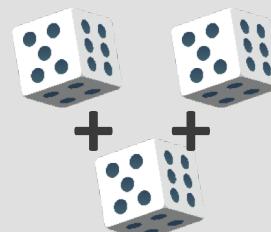
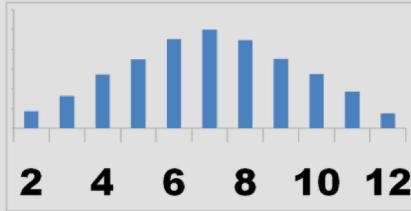
Центральная предельная теорема (ЦПТ)

ЦПТ говорит, что сумма довольно большого числа случайных величин имеет распределение близкое к нормальному



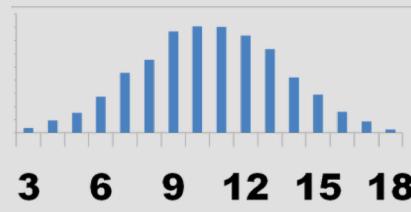
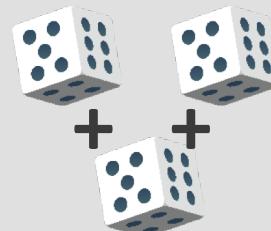
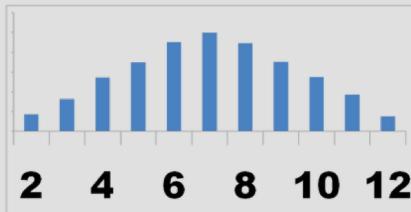
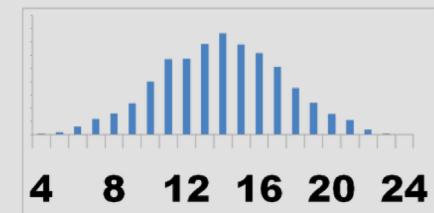
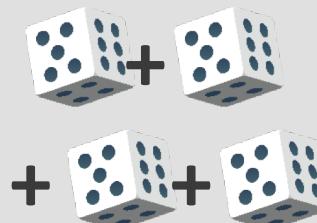
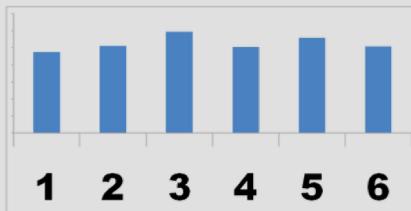
Центральная предельная теорема (ЦПТ)

ЦПТ говорит, что сумма довольно большого числа случайных величин имеет распределение близкое к нормальному



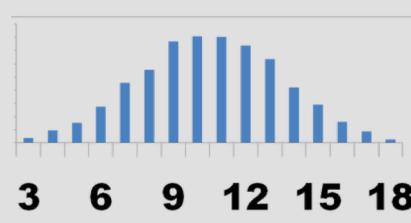
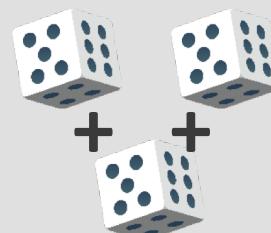
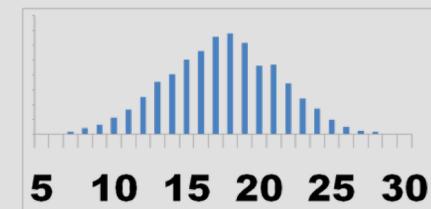
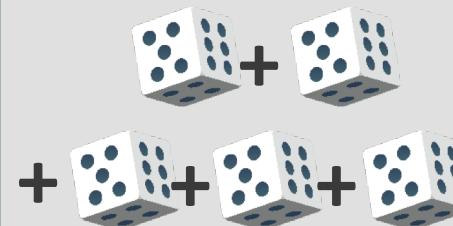
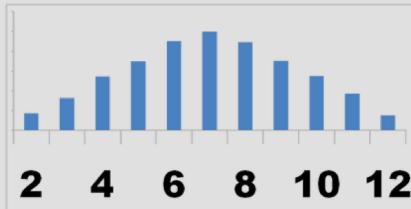
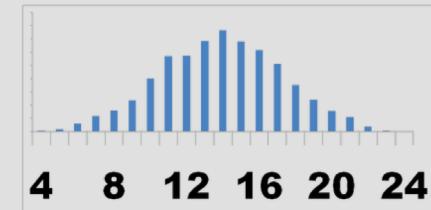
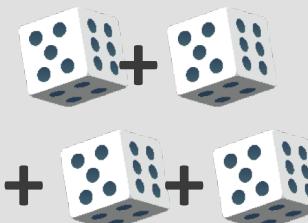
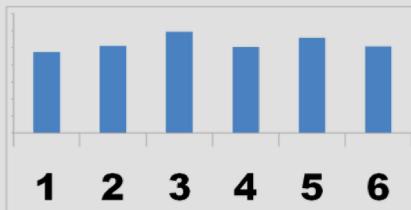
Центральная предельная теорема (ЦПТ)

ЦПТ говорит, что сумма довольно большого числа случайных величин имеет распределение близкое к нормальному



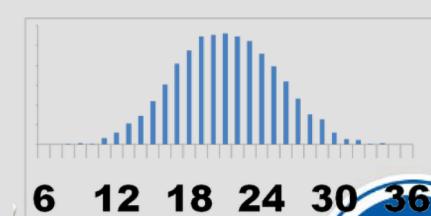
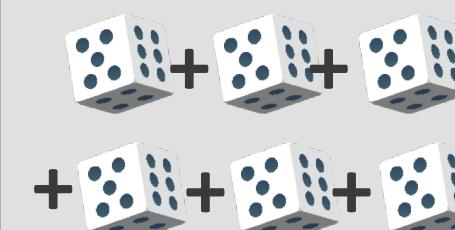
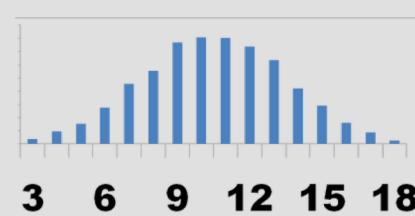
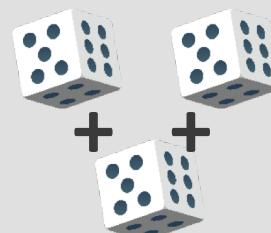
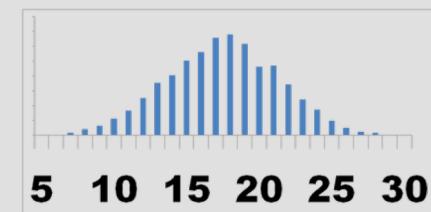
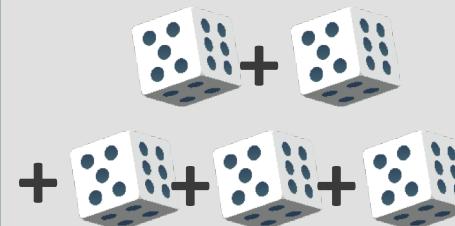
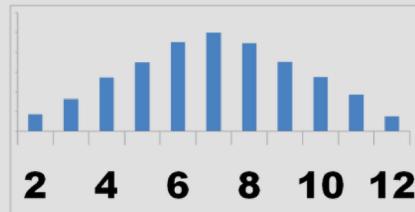
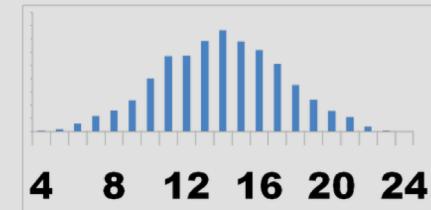
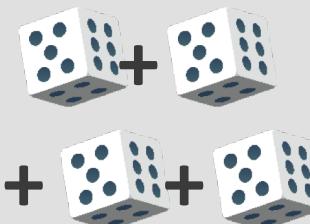
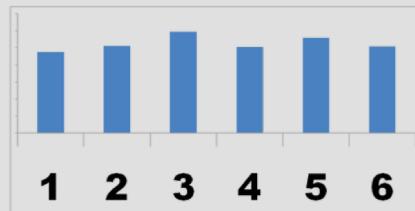
Центральная предельная теорема (ЦПТ)

ЦПТ говорит, что сумма довольно большого числа случайных величин имеет распределение близкое к нормальному



Центральная предельная теорема (ЦПТ)

ЦПТ говорит, что сумма довольно большого числа случайных величин имеет распределение близкое к нормальному



Центральная предельная теорема

Теорема:

Пусть X_1, \dots, X_n попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $\text{Var}(X_1) < \infty$ тогда:



Центральная предельная теорема

Теорема:

Пусть X_1, \dots, X_n попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $\text{Var}(X_1) < \infty$ тогда:

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{d} N\left(\mathbb{E}(X_1), \frac{\text{Var}(X_1)}{n}\right)$$



Центральная предельная теорема

Теорема:

Пусть X_1, \dots, X_n попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $\text{Var}(X_1) < \infty$ тогда:

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{d} N\left(\mathbb{E}(X_1), \frac{\text{Var}(X_1)}{n}\right)$$



Иногда пишут:

либо:

$$\frac{\bar{X}_n - \mathbb{E}(X_1)}{\sqrt{\frac{\text{Var}(X_1)}{n}}} \xrightarrow{d} N(0,1) \quad \sqrt{n} \cdot \frac{\bar{X}_n - \mathbb{E}(X_1)}{sd(X_1)} \xrightarrow{d} N(0,1)$$



Центральная предельная теорема

Простым языком:

- Сумма достаточно большого числа случайных величин имеет распределение близкое к нормальному



Центральная предельная теорема

Простым языком:

- Сумма достаточно большого числа случайных величин имеет распределение близкое к нормальному
- Есть очень большое количество формулировок ЦПТ с разными условиями



Центральная предельная теорема

Простым языком:

- Сумма достаточно большого числа случайных величин имеет распределение близкое к нормальному
- Есть очень большое количество формулировок ЦПТ с разными условиями
- Главное, чтобы случайные величины были похожи друг на друга и не было такого, что одна из них резко выделяется на фоне остальных



Центральная предельная теорема

$X =$

X — время прихода Миши на первую пару



Центральная предельная теорема

X_1 – на Мишу прыгнул кот, и он проснулся пораньше

$X =$ 

X – время прихода Миши на первую пару



Центральная предельная теорема

X_1 – на Мишу прыгнул кот, и он проснулся пораньше

X_2 – готовил завтрак, убежало молоко, задержался убрать

$$X = \text{ + } \begin{array}{c} \text{Cheshire Cat face} \\ + \\ \text{Pitcher of milk} \\ + \\ \text{Four-leaf clover} \end{array}$$

X – время прихода Миши на первую пару



Центральная предельная теорема

X_1 – на Мишу прыгнул кот, и он проснулся пораньше

X_2 – готовил завтрак, убежало молоко, задержался убрать

X_3 – автобус приехал пораньше



X – время прихода Миши на первую пару

Центральная предельная теорема

X_1 – на Мишу прыгнул кот, и он проснулся пораньше

X_2 – готовил завтрак, убежало молоко, задержался убрать

X_3 – автобус приехал пораньше

X_4 – из-за аварии попали в пробку



X – время прихода Миши на первую пару

Центральная предельная теорема

X_1 – на Мишу прыгнул кот, и он проснулся пораньше

X_2 – готовил завтрак, убежало молоко, задержался убрать

X_3 – автобус приехал пораньше

X_4 – из-за аварии попали в пробку

...

$$X = \text{} + \text{} + \text{} + \text{} + \dots$$

X – время прихода Миши на первую пару



Центральная предельная теорема

- X – время прихода Миши на первую пару
- Распределение близко к нормальному



Центральная предельная теорема

- X – время прихода Миши на первую пару
- Распределение близко к нормальному
- Если одна из случайных величин резко выделяется на фоне остальных, нормальность ломается, появляются **тяжёлые хвосты**



Крайнеземье и средиземье



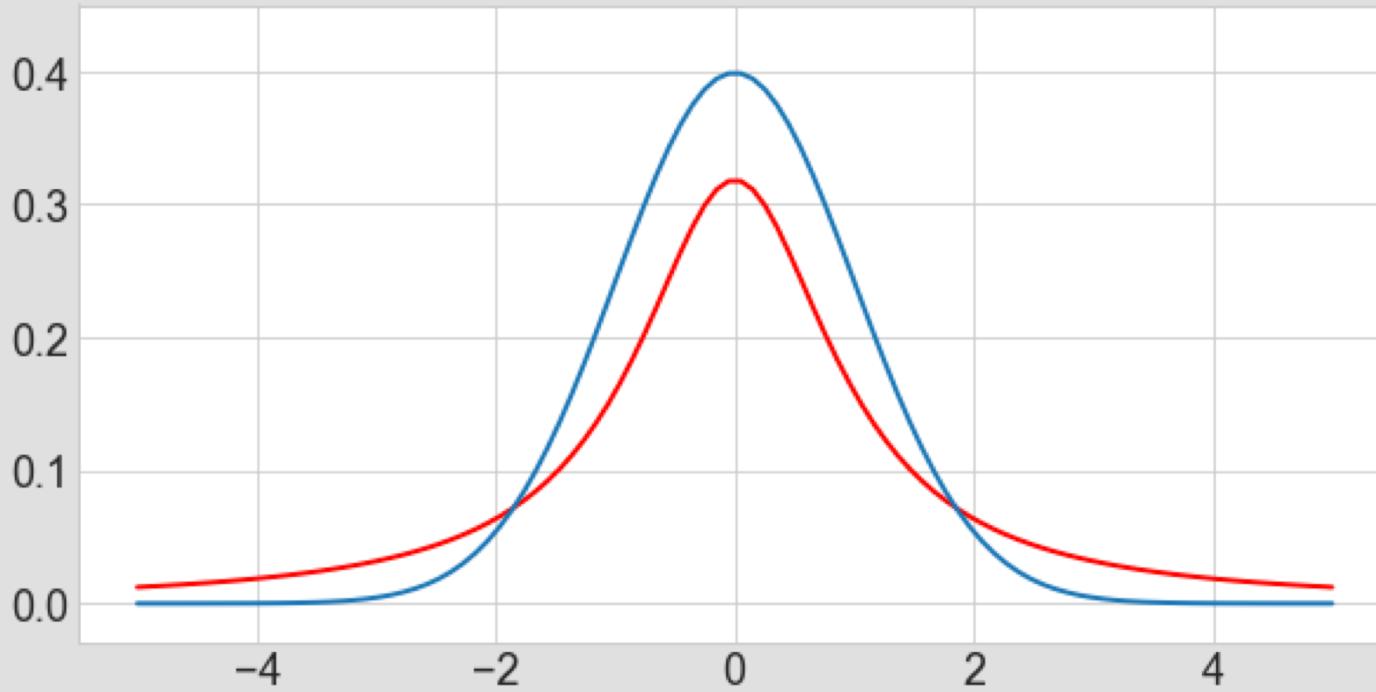
Крайнеземье и средиземье



ЦПТ и ЗБЧ работают в Средиземье



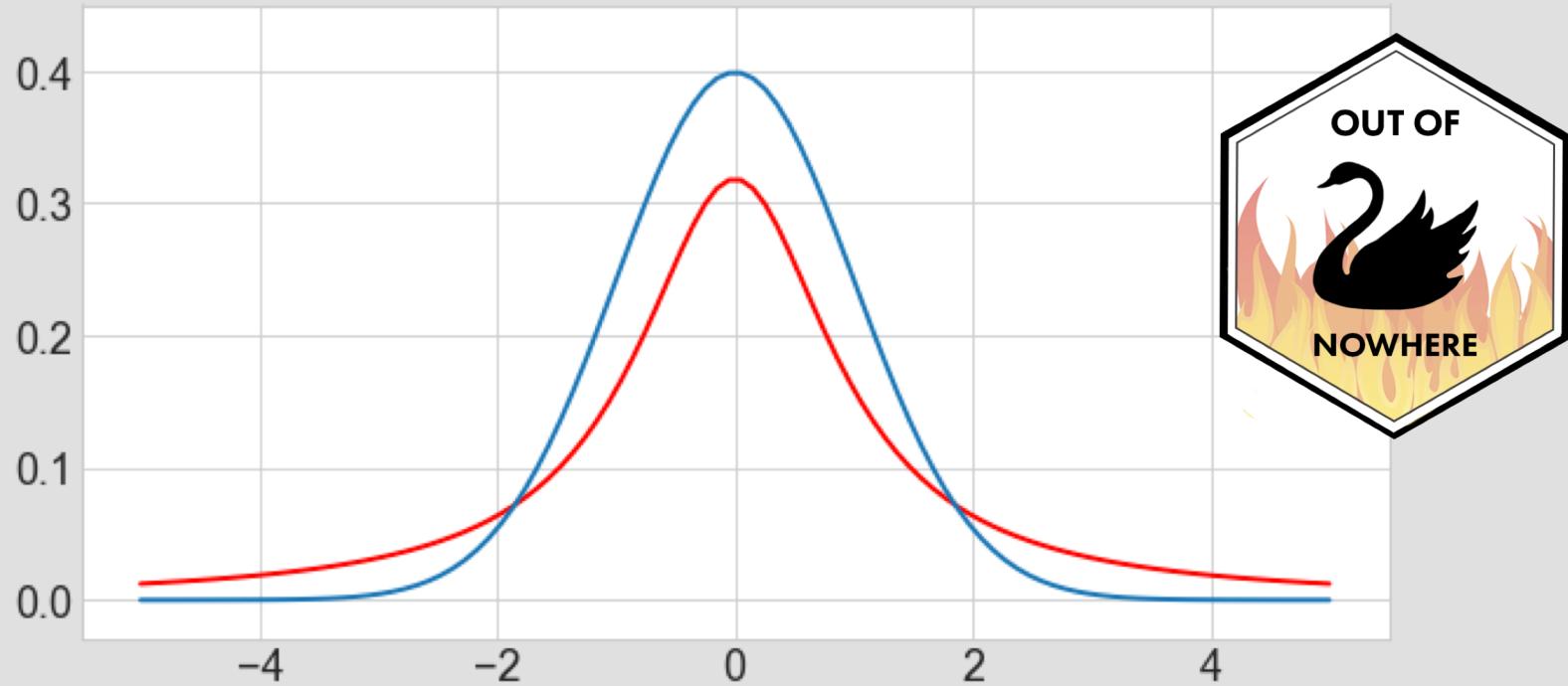
Крайнеземье и средиземье



- Хвосты красного распределения тяжёлые
- Под ними сосредоточена большая вероятностная масса
- События из-под них более вероятны



Крайнеземье и средиземье



- Статистика недооценивает тяжесть хвостов из-за того, что события из них встречаются редко



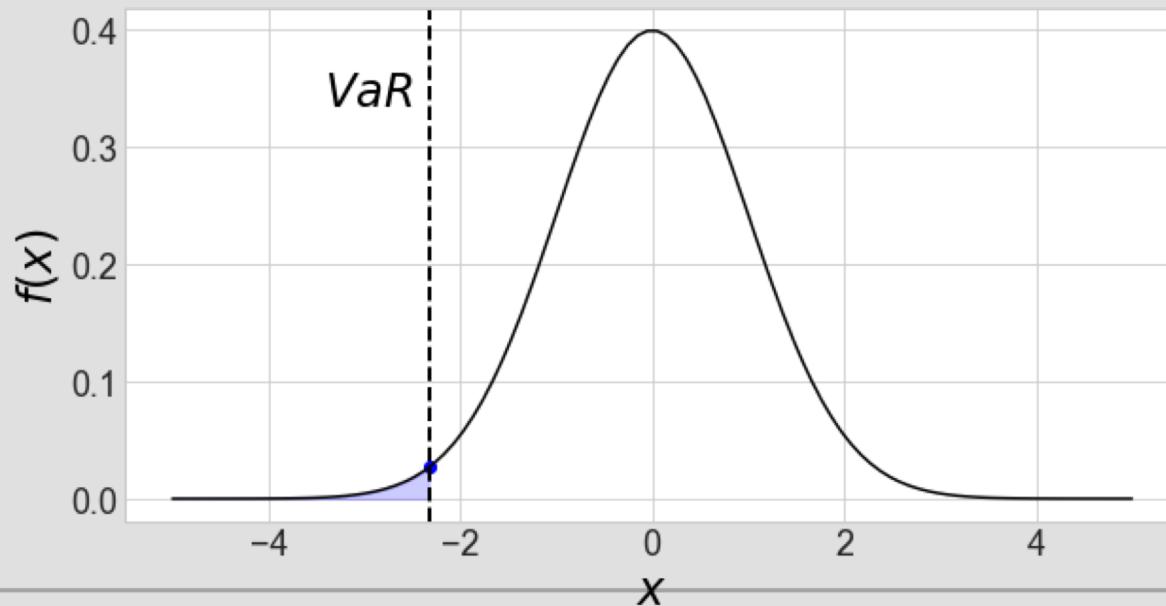
Тяжёлые хвосты и финансы

- Важно понимать, сколько денег мы потеряем в самом плохом случае



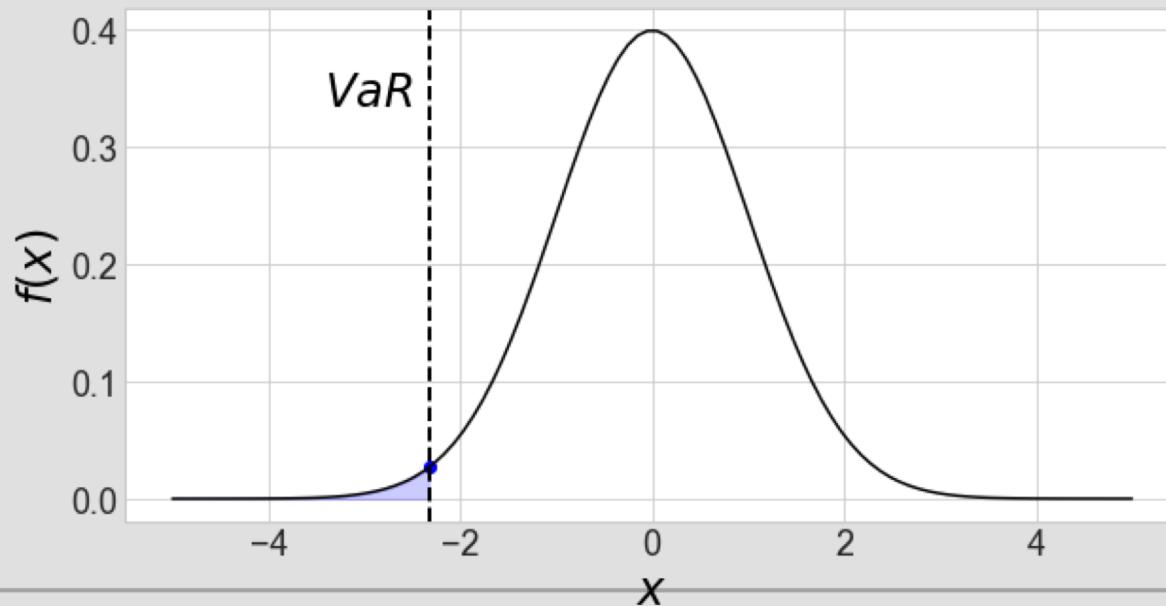
Тяжёлые хвосты и финансы

- Важно понимать, сколько денег мы потеряем в самом плохом случае
- Пытаются смоделировать 5% квантиль распределения доходностей, VaR – Value at risk.



Тяжёлые хвосты и финансы

- Важно понимать, сколько денег мы потеряем в самом плохом случае
- Пытаются смоделировать 5% квантиль распределения доходностей, VaR – Value at risk.
- Не нужно уметь хорошо моделировать всё распределение доходностей, достаточно уметь моделировать левый хвост



Тяжёлые хвосты и финансы

- Распределение доходностей чаще всего отличается от нормального, его хвосты оказываются тяжёлыми



Тяжёлые хвосты и финансы

- Распределение доходностей чаще всего отличается от нормального, его хвосты оказываются тяжёлыми
- Сложно набрать достаточное количество статистики, чтобы адекватно оценить с какой вероятностью произойдёт катастрофа (катастрофы очень редки)



Тяжёлые хвосты и финансы

- Распределение доходностей чаще всего отличается от нормального, его хвосты оказываются тяжёлыми
- Сложно набрать достаточное количество статистики, чтобы адекватно оценить с какой вероятностью произойдёт катастрофа (катастрофы очень редки)
- Оценки всегда занижены

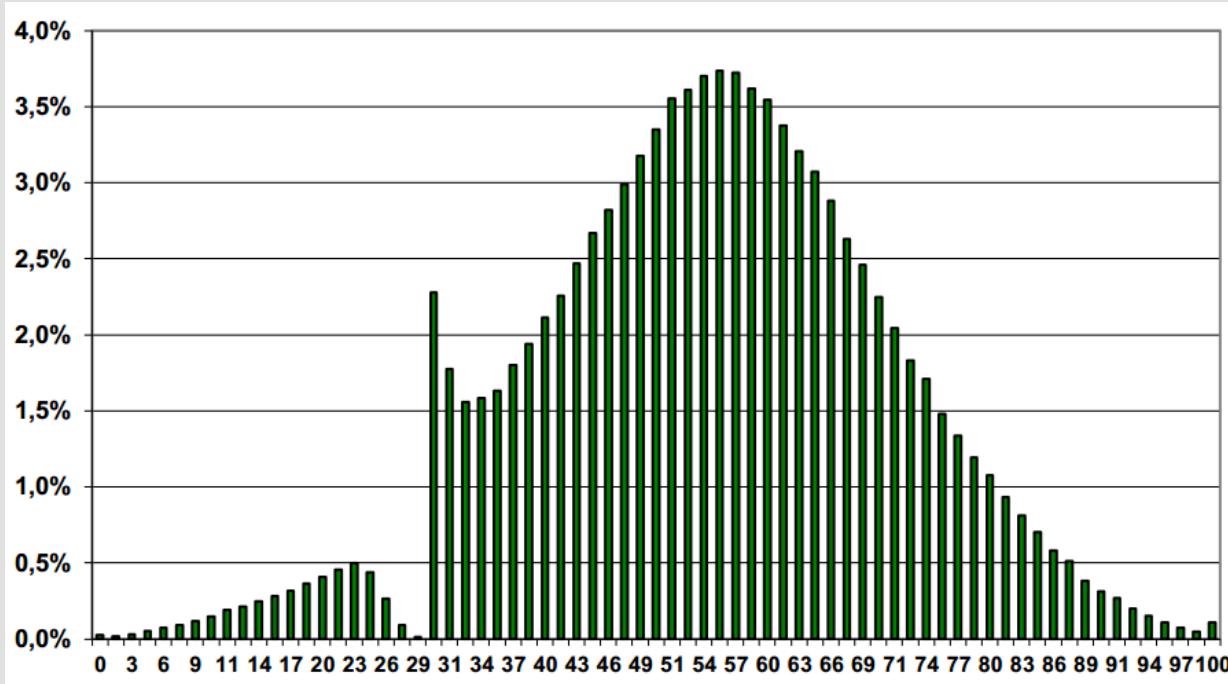


Тяжёлые хвосты и финансы

- Распределение доходностей чаще всего отличается от нормального, его хвосты оказываются тяжёлыми
- Сложно набрать достаточное количество статистики, чтобы адекватно оценить с какой вероятностью произойдёт катастрофа (катастрофы очень редки)
- Оценки всегда занижены
- Нужны специальные методы для работы с Крайнеземьем и тяжёлыми хвостами



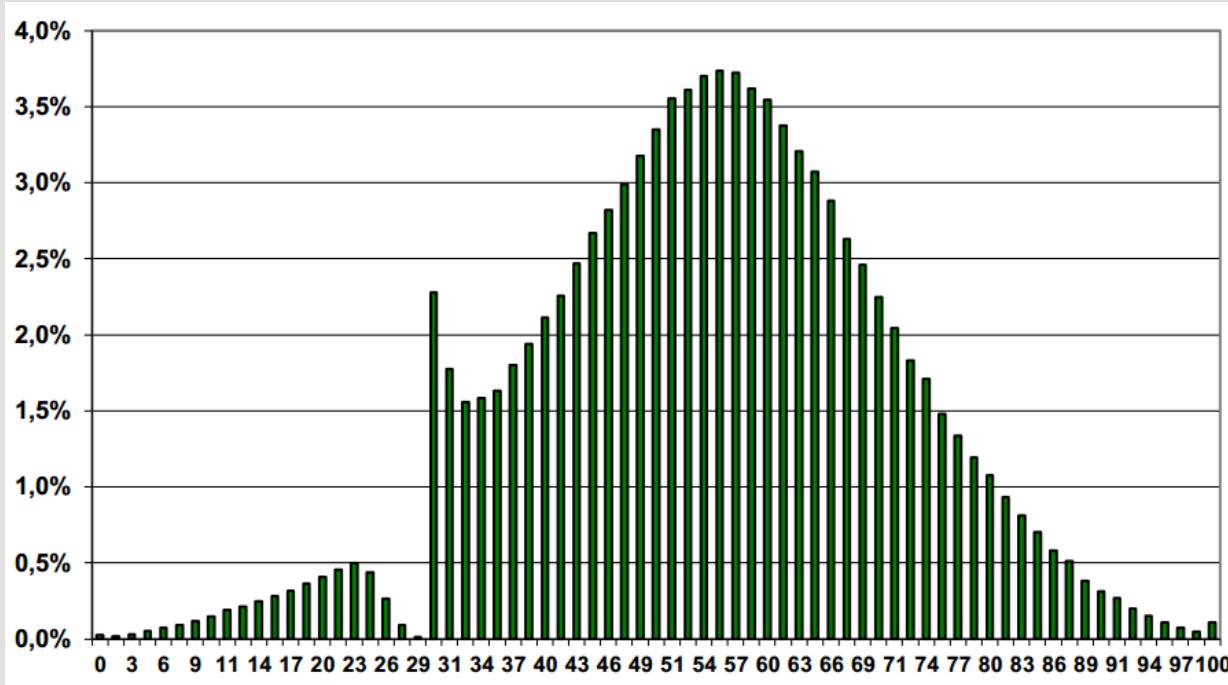
Польское ЕГЭ



- Результаты экзамена, скорее, относятся к Средиземью
 - Что не так с распределением результатов экзамена?
- https://www.reddit.com/r/poland/comments/ber86s/distribution_of_final_exam_scores_in_poland/



Польское ЕГЭ



- Подозрительный пик в районе проходного балла (30)
 - Подозрительный пик на 100 баллах
- https://www.reddit.com/r/poland/comments/ber86s/distribution_of_final_exam_scores_in_poland/



Центральная предельная теорема

Теорема:

Пусть X_1, \dots, X_n попарно независимые и одинаково распределённые случайные величины с конечной дисперсией, $\text{Var}(X_1) < \infty$ тогда:

$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{d} N\left(\mathbb{E}(X_1), \frac{\text{Var}(X_1)}{n}\right)$$



Буква d над стрелкой означает сходимость по распределению



Сходимость по распределению

Последовательность случайных величин X_1, \dots, X_n, \dots сходится по распределению к случайной величине X , если

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x),$$



Сходимость по распределению

Последовательность случайных величин X_1, \dots, X_n, \dots сходится по распределению к случайной величине X , если

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x),$$

то есть последовательность функций распределения $F_{X_n}(x)$ сходится к функции $F_X(x)$ во всех точках x , где $F_X(x)$ непрерывна.



Сходимость по распределению

Последовательность случайных величин X_1, \dots, X_n, \dots сходится по распределению к случайной величине X , если

$$\lim_{n \rightarrow \infty} F_{X_n}(x) = F_X(x),$$

то есть последовательность функций распределения $F_{X_n}(x)$ сходится к функции $F_X(x)$ во всех точках x , где $F_X(x)$ непрерывна.



Обычно пишут:

$$X_n \xrightarrow{d} X \text{ при } n \rightarrow \infty$$

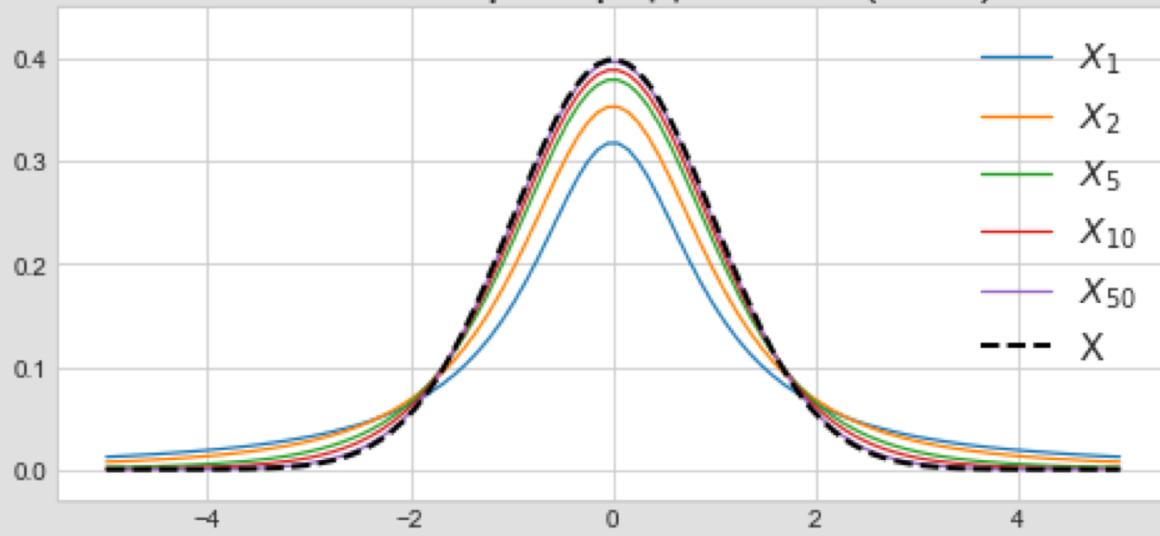
либо:

$$X_n \xrightarrow{F} X \text{ при } n \rightarrow \infty$$

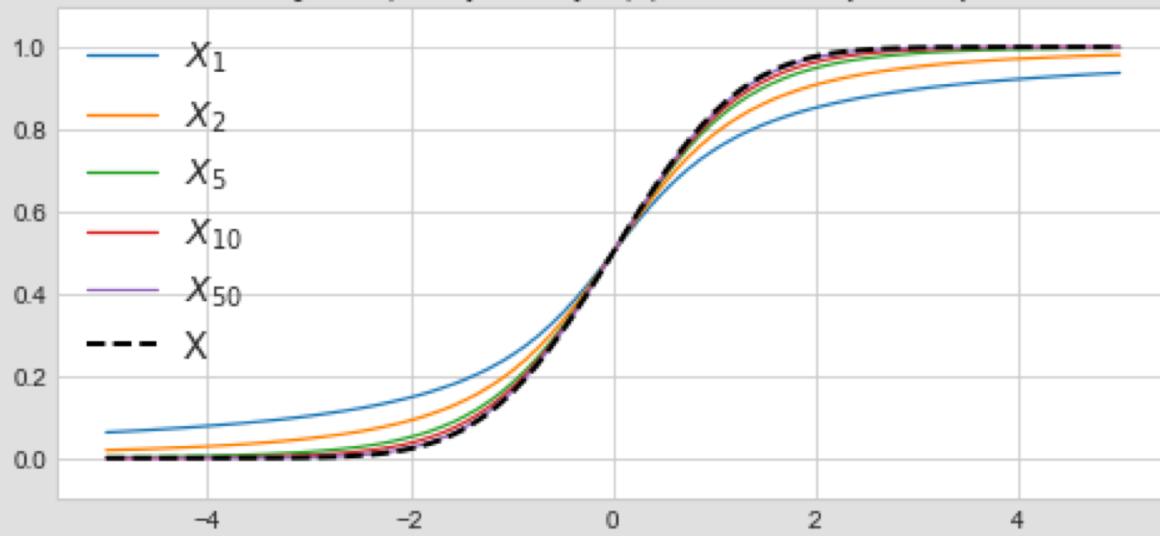


Сходимость по распределению

Плотность распределения (PDF)



Функция распределения (CDF)



ЗБЧ vs ЦПТ (две теоремы о среднем)

ЗБЧ:
$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1)$$

ЦПТ:
$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{d} N\left(\mathbb{E}(X_1), \frac{Var(X_1)}{n}\right)$$



ЗБЧ vs ЦПТ (две теоремы о среднем)

ЗБЧ: $\frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1)$

ЦПТ: $\frac{X_1 + \dots + X_n}{n} \xrightarrow{d} N\left(\mathbb{E}(X_1), \frac{Var(X_1)}{n}\right)$

ЗБЧ: одно среднее, посчитанное по выборке размера n .

При росте n среднее стабилизируется около математического ожидания



ЗБЧ vs ЦПТ (две теоремы о среднем)

ЗБЧ:
$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{p} \mathbb{E}(X_1)$$

ЦПТ:
$$\frac{X_1 + \dots + X_n}{n} \xrightarrow{d} N\left(\mathbb{E}(X_1), \frac{Var(X_1)}{n}\right)$$

ЗБЧ: одно среднее, посчитанное по выборке размера n .

При росте n среднее стабилизируется около математического ожидания

ЦПТ: много средних, посчитанных по разным выборкам размера n . При росте n распределение всё больше похоже на нормальное, оно всё компактнее вокруг математического ожидания



Резюме

ЦПТ говорит, что при больших выборках и отсутствии аномалий мы можем аппроксимировать распределение среднего нормальным распределением

В случае, если какие-то случайные величины сильно выделяются на фоне остальных, мы имеем дело с тяжёлыми хвостами

Тяжёлые хвосты часто встречаются в финансах и требуют к себе отдельного статистического подхода



Сходимости случайных величин



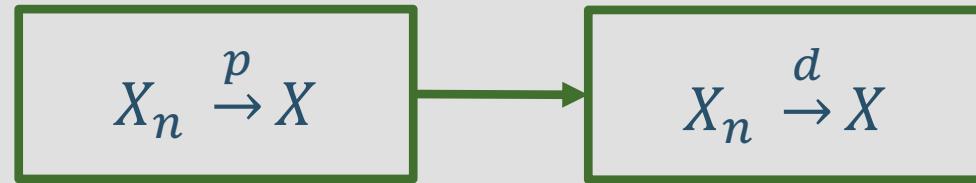
Виды сходимостей

$$X_n \xrightarrow{d} X$$

По распределению
(самая слабая)



Виды сходимостей

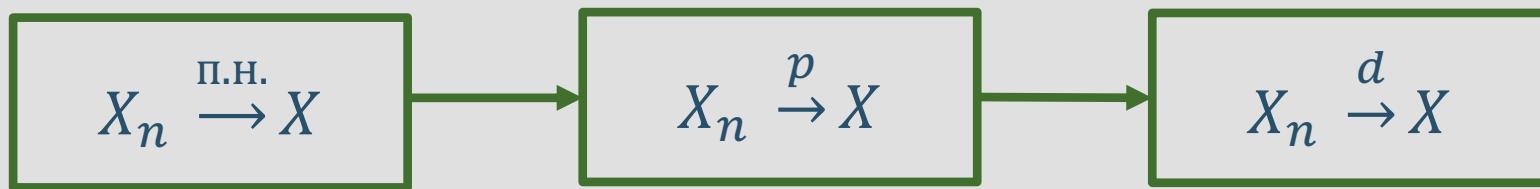


По вероятности

По распределению
(самая слабая)



Виды сходимостей



Почти наверное
(с вероятностью
единица)

По вероятности

По распределению
(самая слабая)

Сходимость «почти наверное» самая сильная из трёх,
сходимость «по распределению» самая слабая



Сходимость почти наверное

Последовательность случайных величин X_1, \dots, X_n, \dots
сходится почти наверное (с вероятностью единица)
к случайной величине X , если

$$\mathbb{P}(\lim_{n \rightarrow \infty} X_n = X) = 1$$

то есть у последовательности есть предел с вероятностью 1.



Сходимость почти наверное

Последовательность случайных величин X_1, \dots, X_n, \dots
сходится почти наверное (с вероятностью единица)
к случайной величине X , если

$$\mathbb{P}(\lim_{n \rightarrow \infty} X_n = X) = 1$$

то есть у последовательности есть предел с вероятностью 1.



Обычно пишут:

либо:

$$X_n \xrightarrow{\text{п.н.}} X \text{ при } n \rightarrow \infty$$

$$X_n \xrightarrow{a.s.} X \text{ при } n \rightarrow \infty$$



Сходимость почти наверное

Последовательность случайных величин X_1, \dots, X_n, \dots
сходится почти наверное (с вероятностью единица)
к случайной величине X , если

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1$$

то есть у последовательности есть предел с вероятностью 1.



В сходимости по вероятности речь
шла о пределе вероятностей:

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$$

Тут речь о вероятности предела.



Сходимость почти наверное

Кубик подкидывают **один раз**

$$X_n = \begin{cases} \frac{n}{n+1}, & \text{если чётное число} \\ (-1)^n, & \text{если нечётное число} \end{cases}$$

С какой вероятностью сходится последовательность X_n ?



Сходимость почти наверное

Кубик подкидывают **один раз**

$$X_n = \begin{cases} \frac{n}{n+1}, & \text{если чётное число} \\ (-1)^n, & \text{если нечётное число} \end{cases}$$

С какой вероятностью сходится последовательность X_n ?

Последовательность $\frac{n}{n+1}$ сходится к 1. Последовательность $(-1)^n$ расходится. Подкидывая кубик, мы получаем сходимость последовательности с вероятностью 0.5



Сходимость почти наверное

Кубик подкидывают **один раз**

$$X_n = \begin{cases} \frac{n}{n+1}, & \text{если чётное число} \\ (-1)^n, & \text{если нечётное число} \end{cases}$$

С какой вероятностью сходится последовательность X_n ?

Последовательность $\frac{n}{n+1}$ сходится к 1. Последовательность $(-1)^n$ расходится. Подкидывая кубик мы получаем сходимость последовательности с вероятностью 0.5

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} X_n = X\right) = 0.5$$

Если бы вероятность была бы 1, была бы сходимость почти наверное.



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

Сходится ли
последовательность
почти наверное?



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

Сходится ли
последовательность
почти наверное?

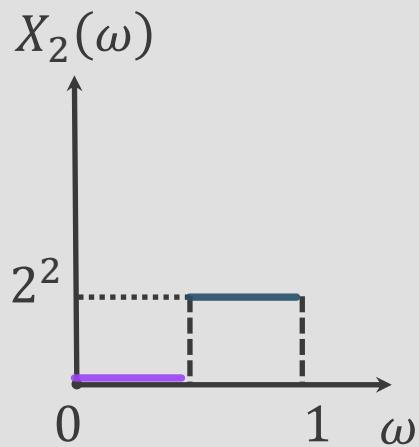
- ❗ Ответ зависит от того, как именно устроена наша
случайная величина



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

Сходится ли
последовательность
почти наверное?



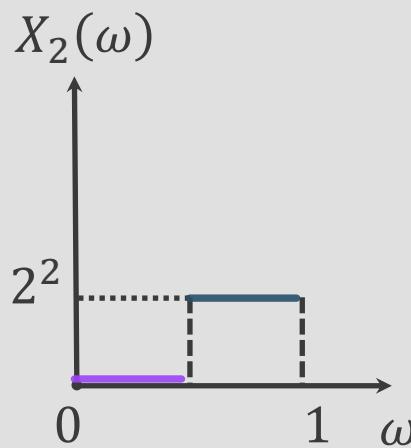
$$n = 2$$



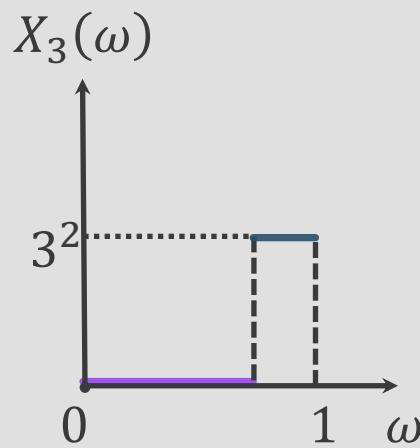
Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

Сходится ли
последовательность
почти наверное?



$$n = 2$$



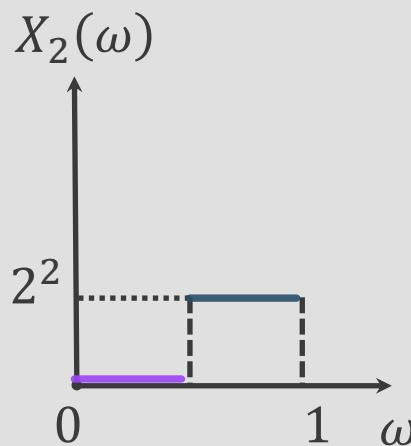
$$n = 3$$



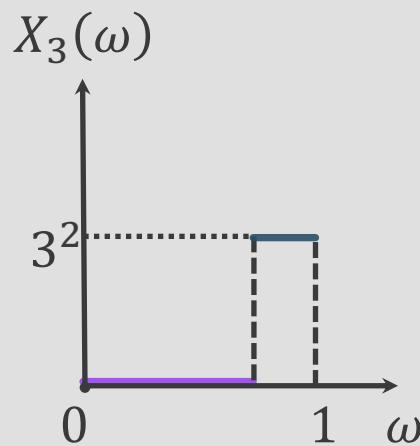
Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

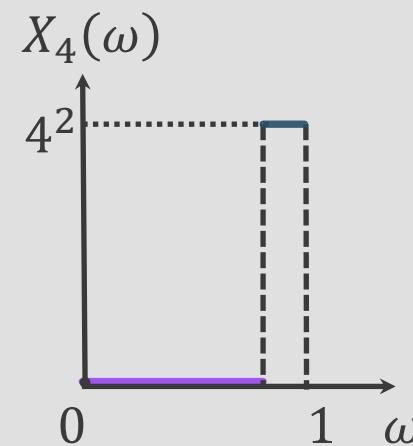
Сходится ли
последовательность
почти наверное?



$$n = 2$$



$$n = 3$$



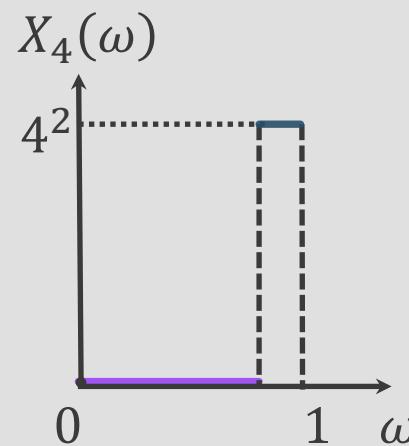
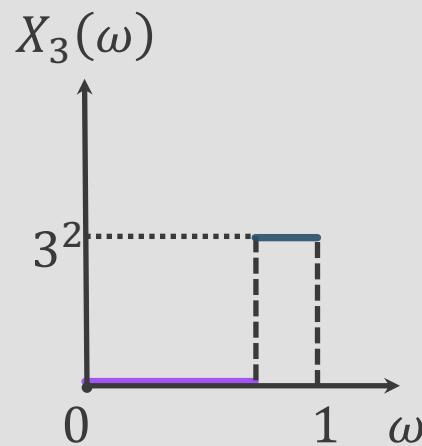
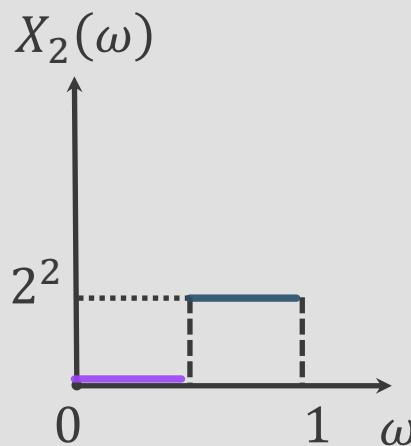
$$n = 4$$



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

Сходится ли
последовательность
почти наверное?



$n = 2$

$n = 3$

$n = 4$

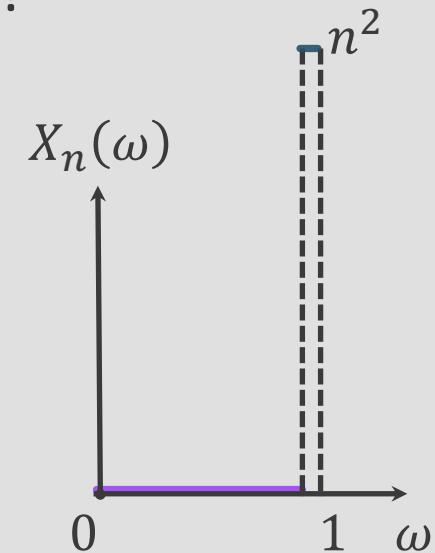
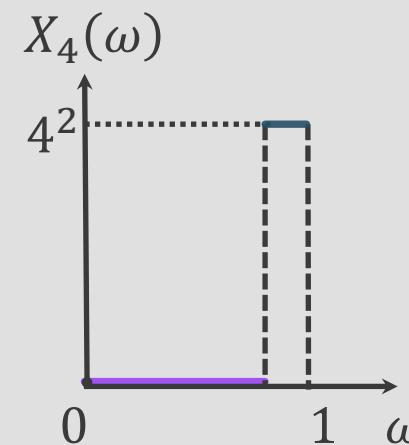
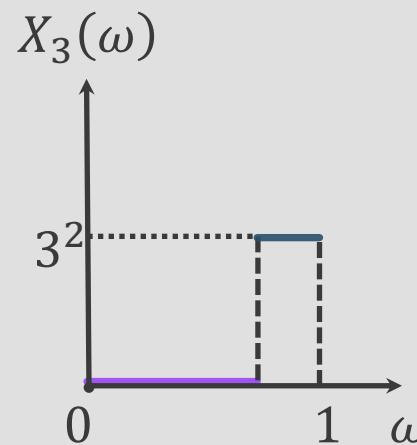
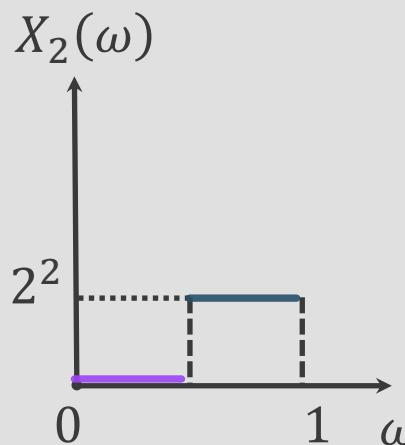
...



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

Сходится ли
последовательность
почти наверное?



$n = 2$

$n = 3$

$n = 4$

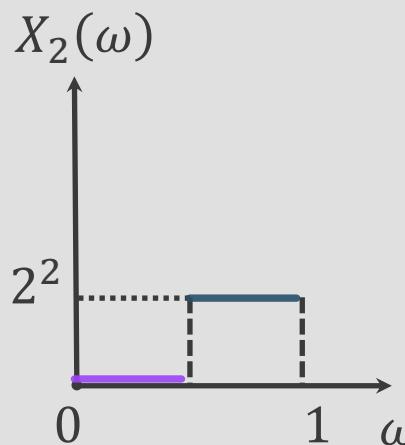
...



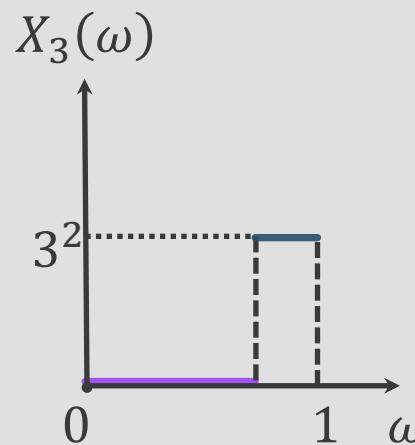
Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

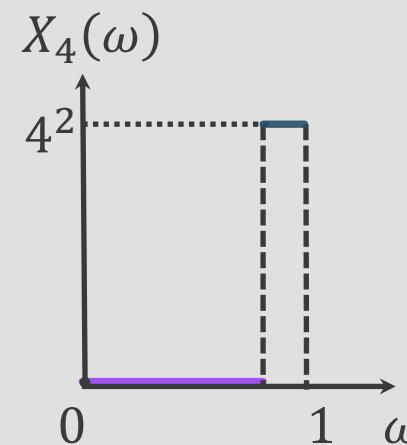
Сходится ли
последовательность
почти наверное?



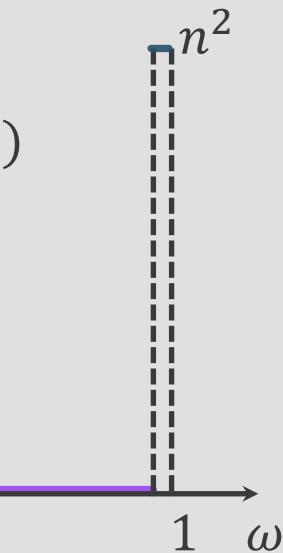
$$n = 2$$



$$n = 3$$



$$n = 4$$



...

- ❗ Отрезок, где случайная величина равна n^2 становится всё меньше, нельзя выделить расходящуюся подпоследовательность



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

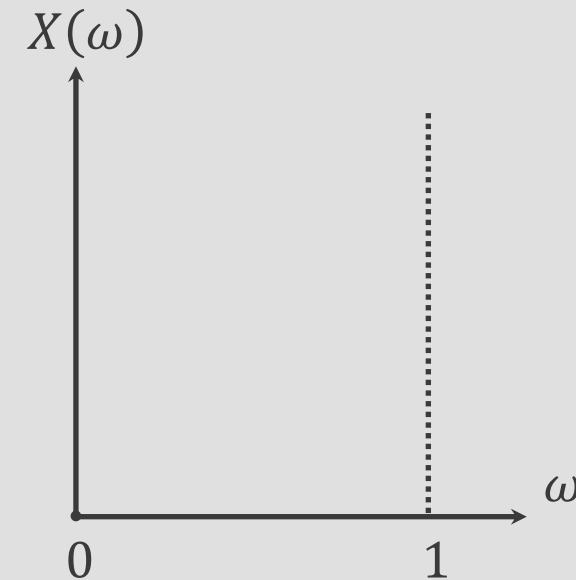
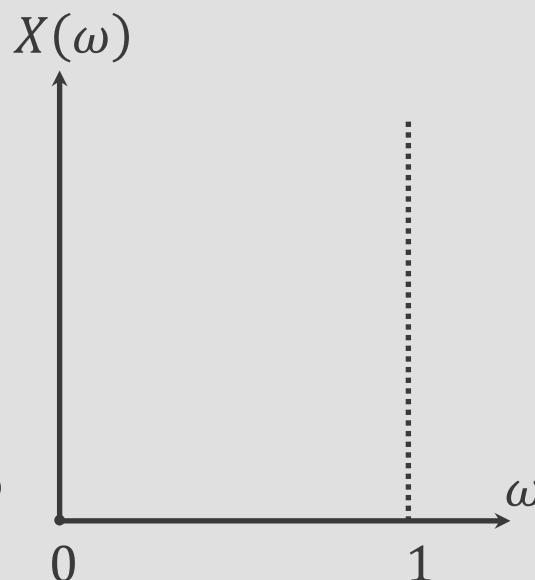
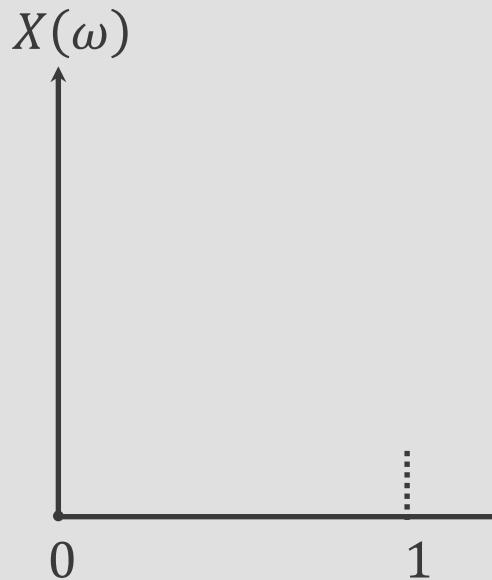
Сходится ли
последовательность
почти наверное?



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

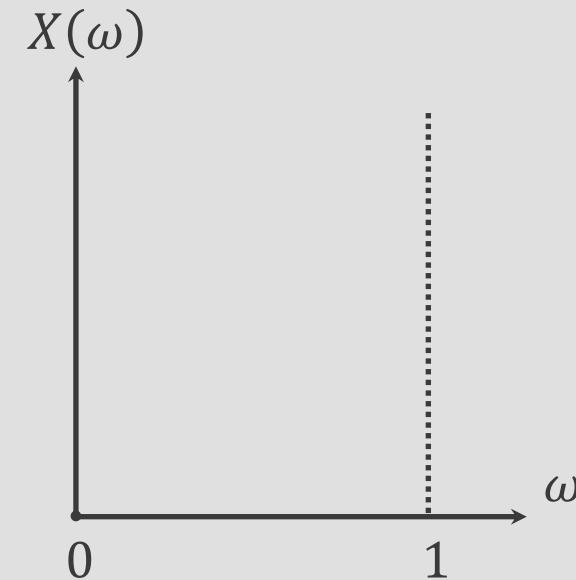
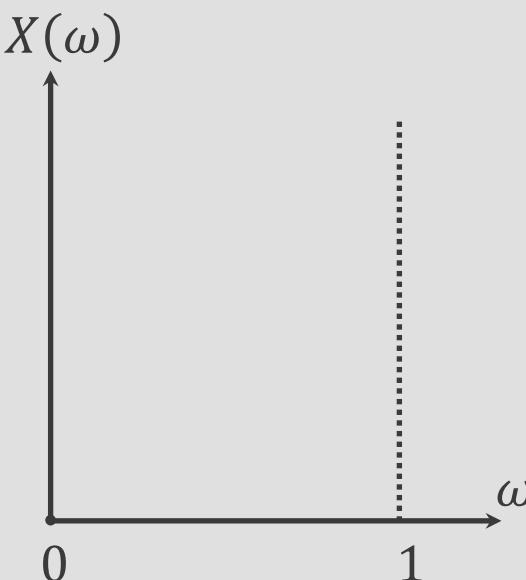
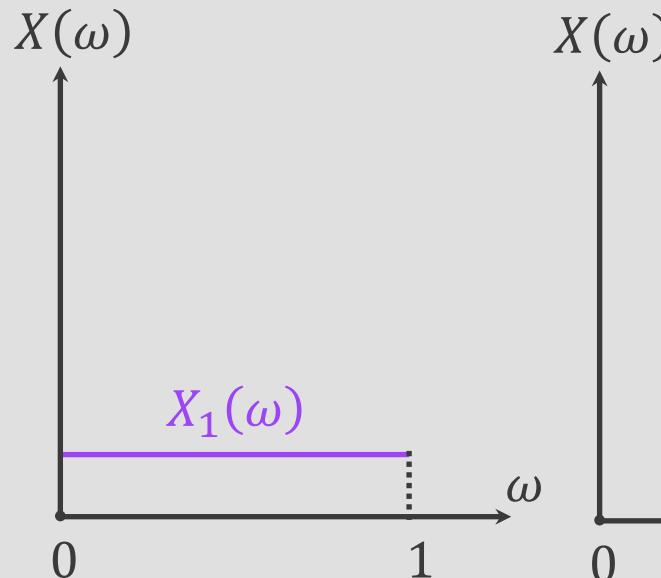
Сходится ли
последовательность
почти наверное?



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

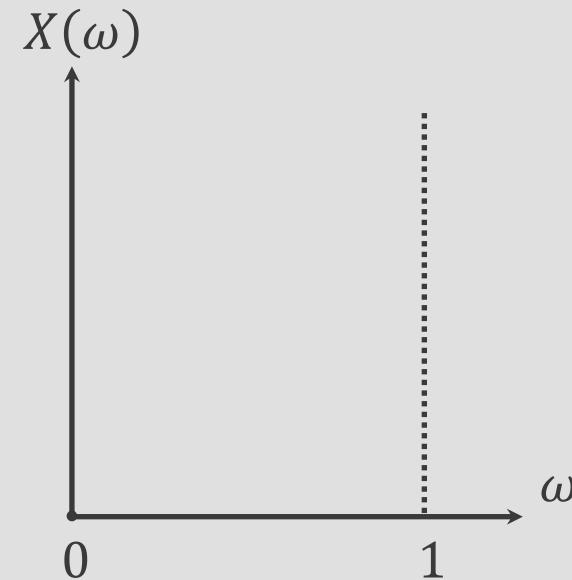
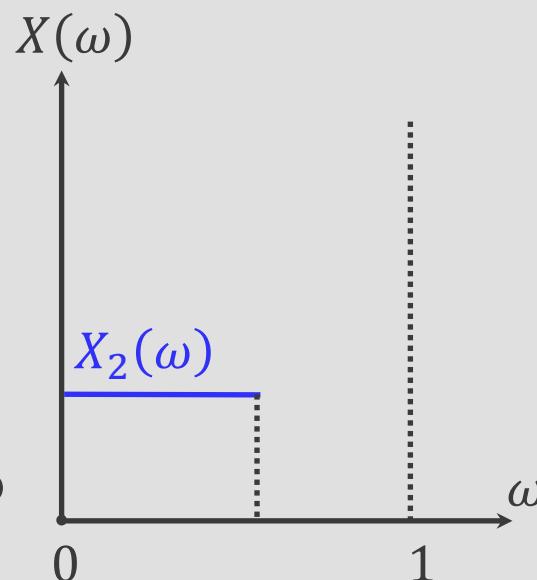
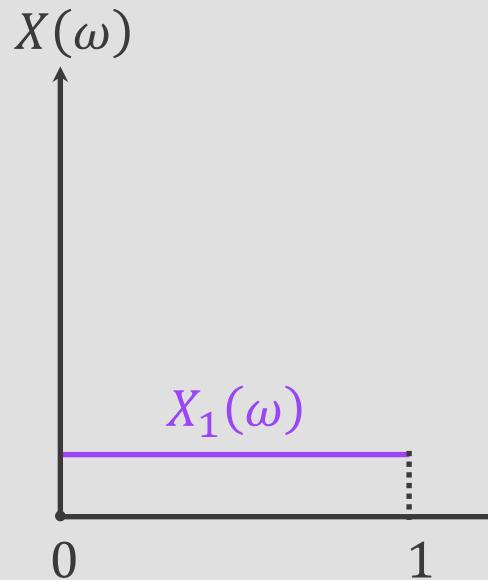
Сходится ли
последовательность
почти наверное?



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

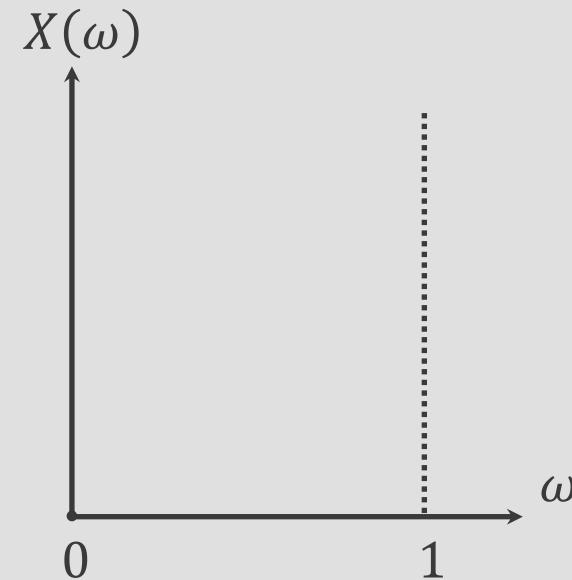
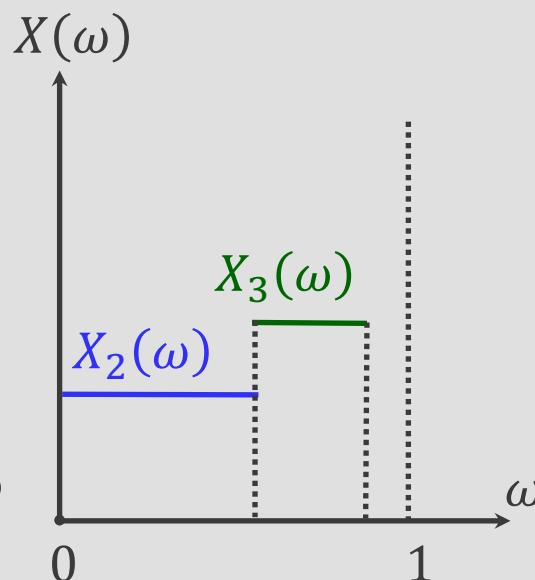
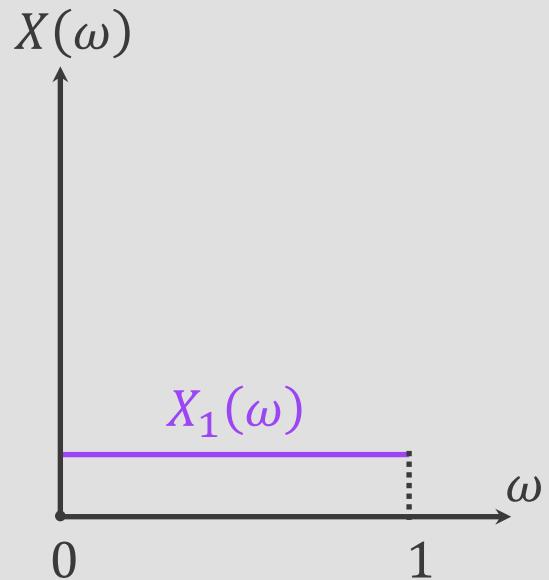
Сходится ли
последовательность
почти наверное?



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

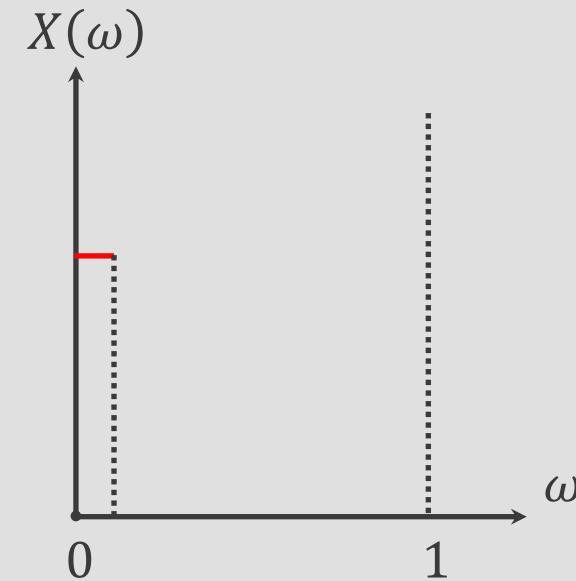
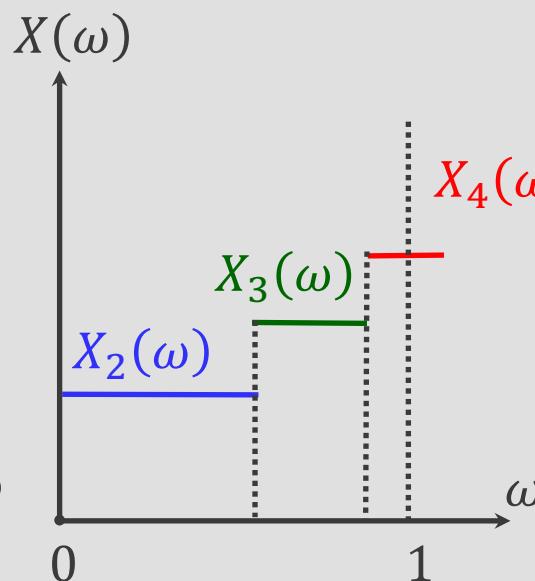
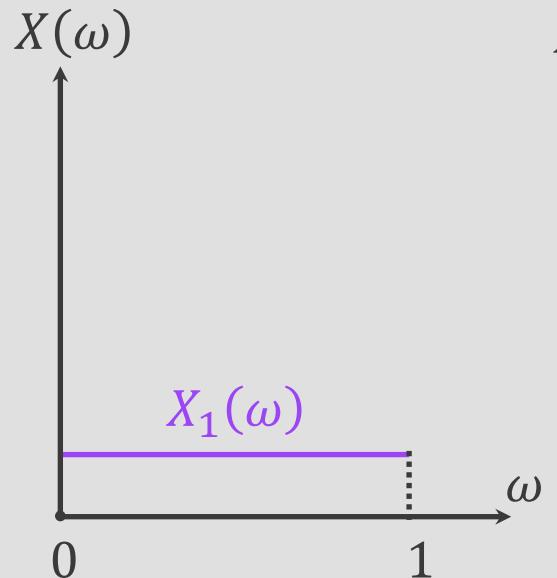
Сходится ли
последовательность
почти наверное?



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

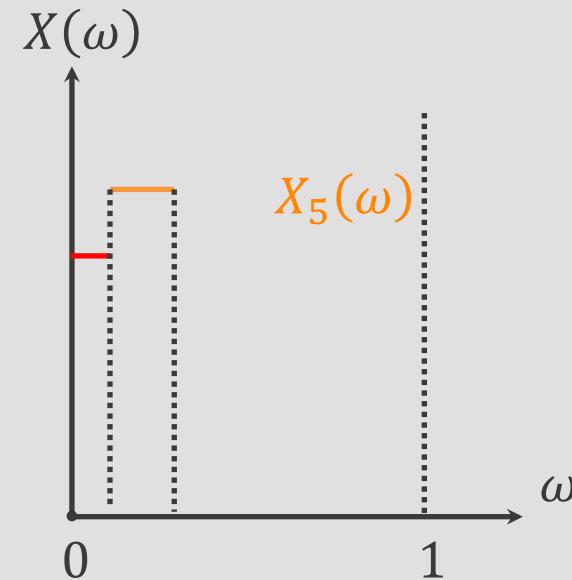
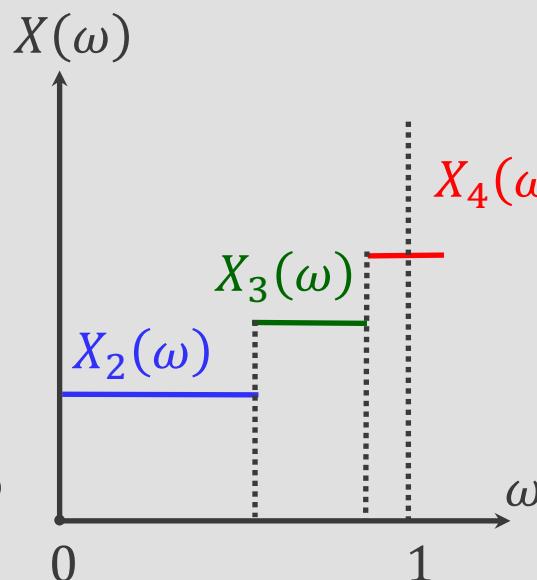
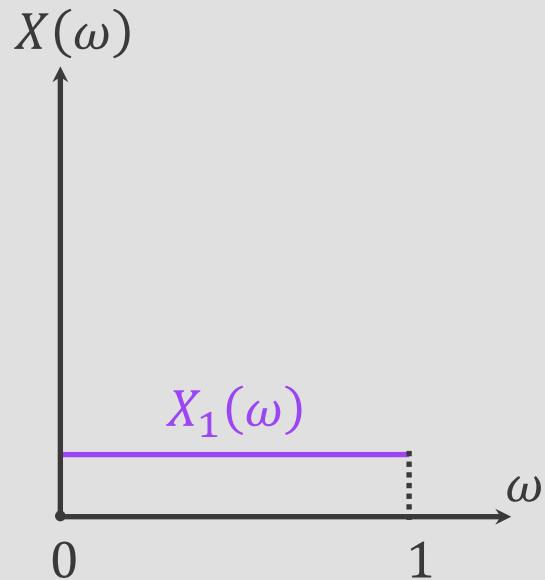
Сходится ли
последовательность
почти наверное?



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

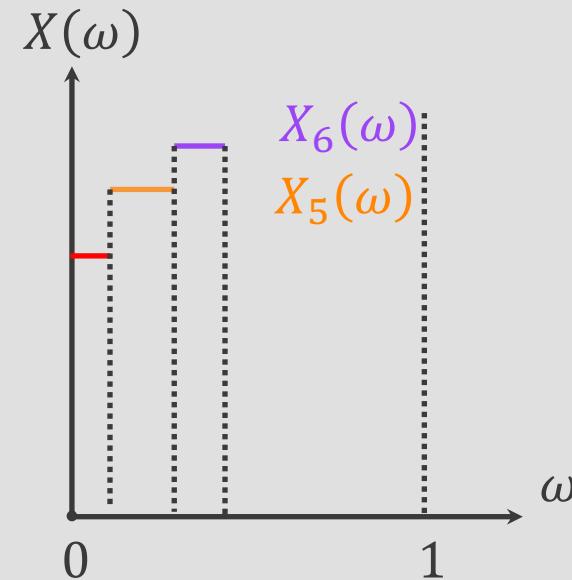
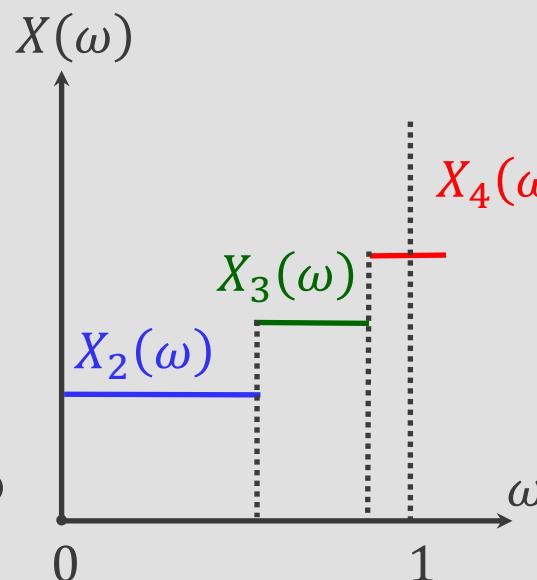
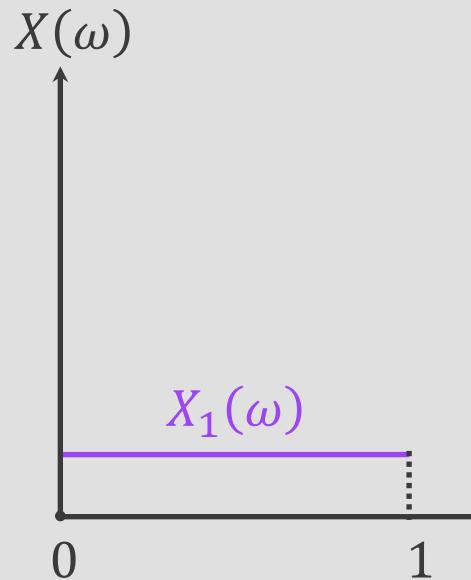
Сходится ли
последовательность
почти наверное?



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

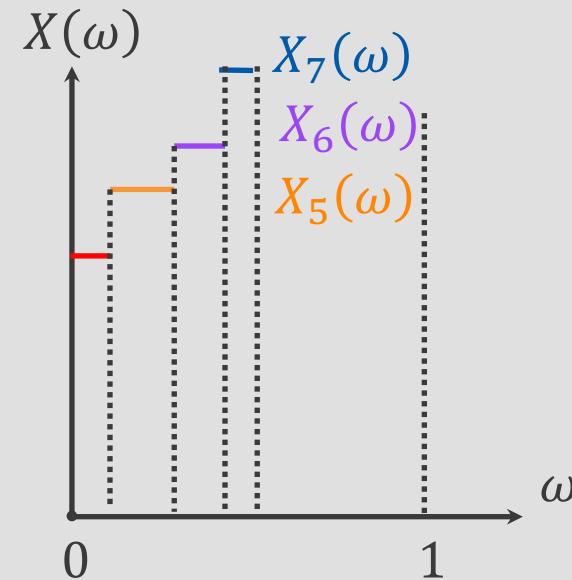
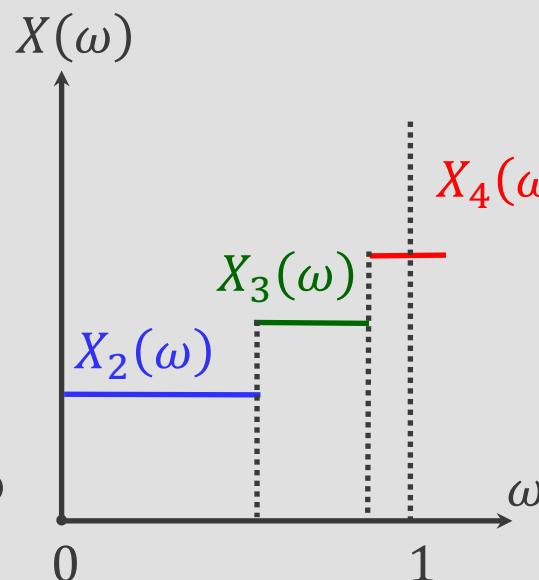
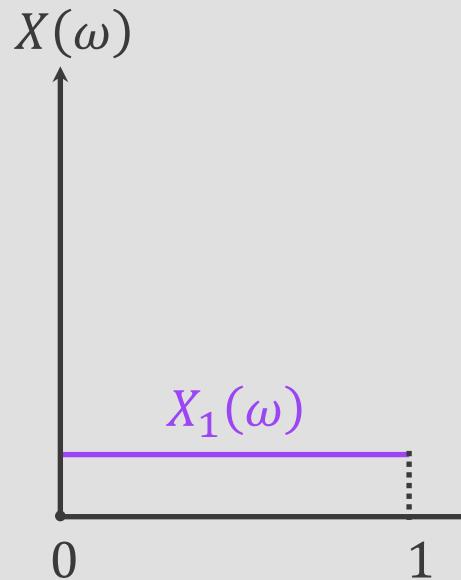
Сходится ли
последовательность
почти наверное?



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

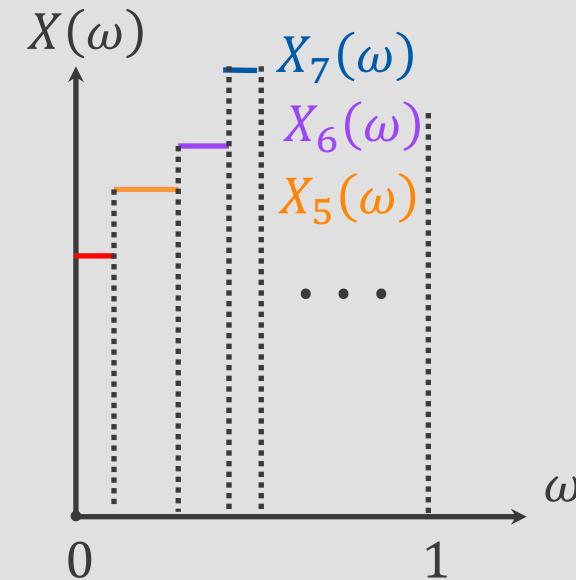
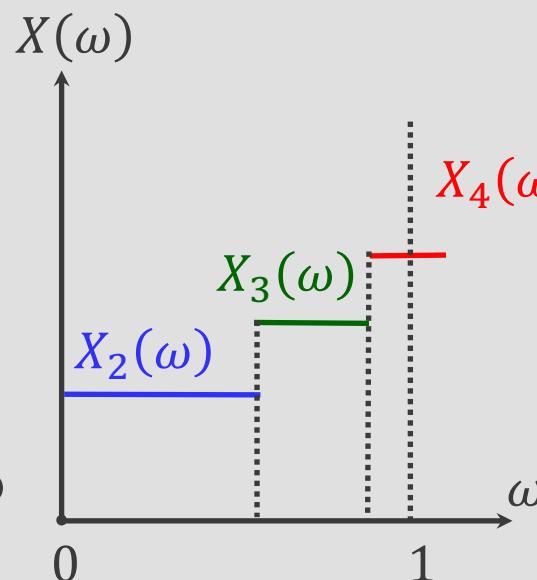
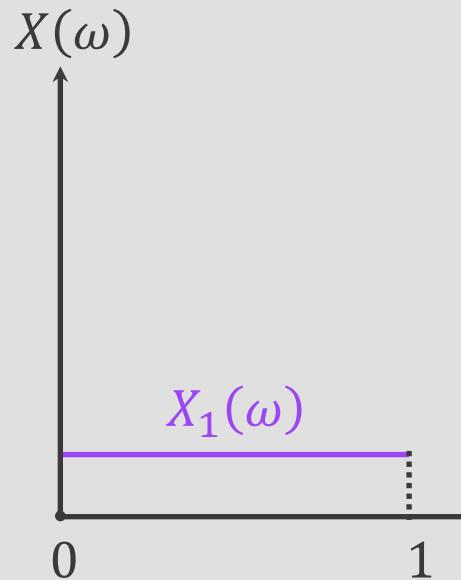
Сходится ли
последовательность
почти наверное?



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

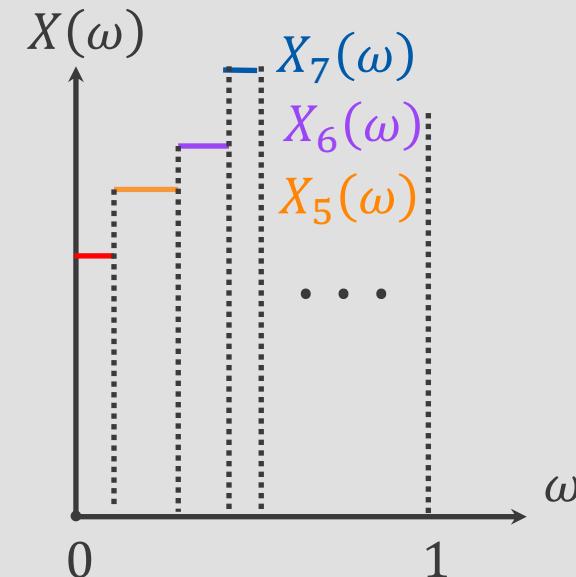
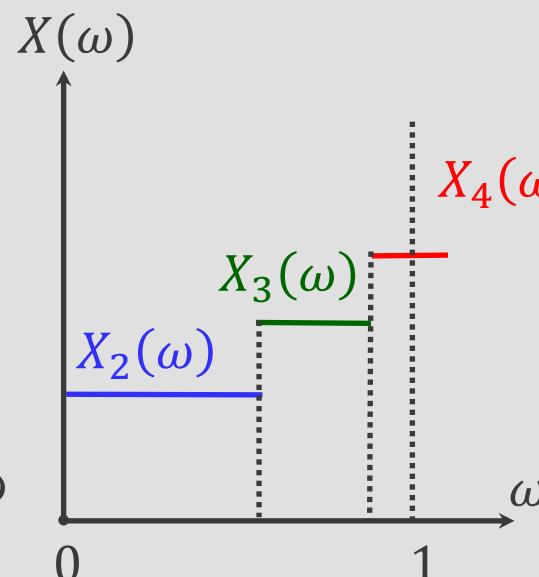
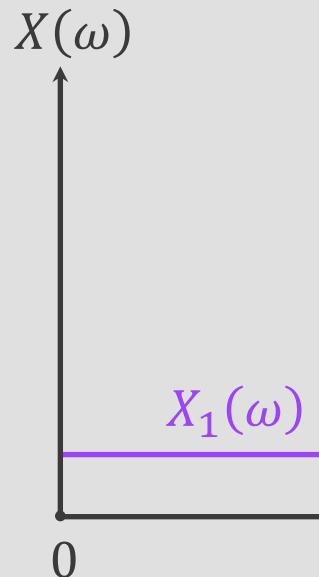
Сходится ли
последовательность
почти наверное?



Пример сходимости почти наверное

X_n	0	n^2
$\mathbb{P}(X_n = k)$	$1 - \frac{1}{n}$	$\frac{1}{n}$

Сходится ли
последовательность
почти наверное?



Отрезок, где случайная величина равна n^2 бегает по всему отрезку $[0; 1]$, каждый кусочек отрезка попадает в эту зону бесконечное число раз, есть $X_{n_k} \rightarrow \infty$



Сильная форма ЗБЧ (Колмогоров)

Теорема:

Пусть X_1, \dots, X_n последовательность независимых и одинаково распределённых случайных величин с $\mathbb{E}(|X_1|) < \infty$, тогда:

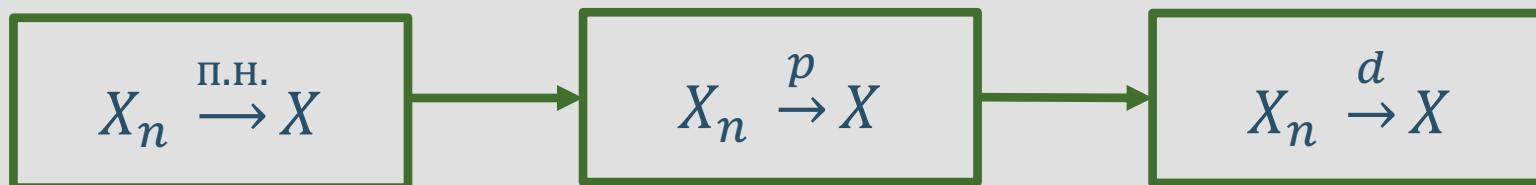
$$\bar{X} = \frac{X_1 + \dots + X_n}{n} \xrightarrow{\text{п.н.}} \mathbb{E}(X_1)$$

Среднее сходится почти наверное к математическому ожиданию при $n \rightarrow \infty$



Резюме

В теории вероятностей изучаются четыре типа сходимостей: по вероятности, по распределению, почти наверное и в среднем



Почти наверное
(с вероятностью
единица)

По вероятности

По распределению
(самая слабая)



Откуда компьютер берёт случайности



Изобретаем велосипед

Предположим, что мы с вами только что изобрели компьютер, и нам надо научить его генерировать случайные числа. Как бы вы поступили?



Изобретаем велосипед

Идея! Согласно квантовой теории, невозможно узнать наверняка когда произойдёт радиоактивный распад. Давайте положим в компьютер немножечко урана.



Кадр из мультипликационного сериала «Симпсоны» / Автор Мэтт Грейнинг, 20th Century Fox Television, Grace Films



Изобретаем велосипед

Идея! Действия человека непредсказуемы. Будем собирать те промежутки времени, которые проходят между нажатиями кнопок на клавиатуре. Это поможет генерировать случайные числа.



Фильм Прибытие (2016)



Изобретаем велосипед

Идея! Давайте использовать непредсказуемые шумы в атмосфере.



► <https://www.random.org/>

Irishtimes.com



Изобретаем велосипед

- Это всё довольно дорого
- Обычно используют псевдослучайные алгоритмы



Изобретаем велосипед

- Это всё довольно дорого
- Обычно используют псевдослучайные алгоритмы

Пример: последовательность цифр в числе пи довольно непредсказуема. Давайте окажемся в каком-то месте числа пи и начиная с него начнём генерацию.



Изобретаем велосипед

- Вся псевдослучайность зависит от начального значения
(не очень надёжный алгоритм)



Изобретаем велосипед

- Вся псевдослучайность зависит от начального значения
(не очень надёжный алгоритм)

Пример: вихрь Мерсена
(основан на простых числах, более надёжный)

- Некоторые алгоритмы держат в секрете



Генерация распределений

- Легче всего научиться генерировать равномерное распределение
- Остальные распределения можно сгенерировать из него. В этом помогает ещё одна базовая теорема.



Квантильное преобразование

Теорема:

Пусть функция распределения $F_X(x)$ непрерывна, тогда случайная величина $Y = F(X)$ имеет равномерное распределение на отрезке $[0; 1]$



Квантильное преобразование

Теорема:

Пусть функция распределения $F_X(x)$ непрерывна, тогда случайная величина $Y = F(X)$ имеет равномерное распределение на отрезке $[0; 1]$

Доказательство:

$$F_Y(y) = P(Y \leq y)$$



Квантильное преобразование

Теорема:

Пусть функция распределения $F_X(x)$ непрерывна, тогда случайная величина $Y = F(X)$ имеет равномерное распределение на отрезке $[0; 1]$

Доказательство:

$$F_Y(y) = P(Y \leq y) = P(F(X) \leq y)$$



Квантильное преобразование

Теорема:

Пусть функция распределения $F_X(x)$ непрерывна, тогда случайная величина $Y = F(X)$ имеет равномерное распределение на отрезке $[0; 1]$

Доказательство:

$$\begin{aligned}F_Y(y) &= P(Y \leq y) = P(F(X) \leq y) = \\&= P(X \leq F^{-1}(y))\end{aligned}$$



Квантильное преобразование

Теорема:

Пусть функция распределения $F_X(x)$ непрерывна, тогда случайная величина $Y = F(X)$ имеет равномерное распределение на отрезке $[0; 1]$

Доказательство:

$$\begin{aligned}F_Y(y) &= P(Y \leq y) = P(F(X) \leq y) = \\&= P(X \leq F^{-1}(y)) = F_X(F^{-1}(y))\end{aligned}$$



Квантильное преобразование

Теорема:

Пусть функция распределения $F_X(x)$ непрерывна, тогда случайная величина $Y = F(X)$ имеет равномерное распределение на отрезке $[0; 1]$

Доказательство:

$$\begin{aligned}F_Y(y) &= P(Y \leq y) = P(F(X) \leq y) = \\&= P\left(X \leq F^{-1}(y)\right) = F_X\left(F^{-1}(y)\right) = y\end{aligned}$$



Квантильное преобразование

Теорема:

Пусть функция распределения $F_X(x)$ непрерывна, тогда случайная величина $Y = F(X)$ имеет равномерное распределение на отрезке $[0; 1]$

Доказательство:

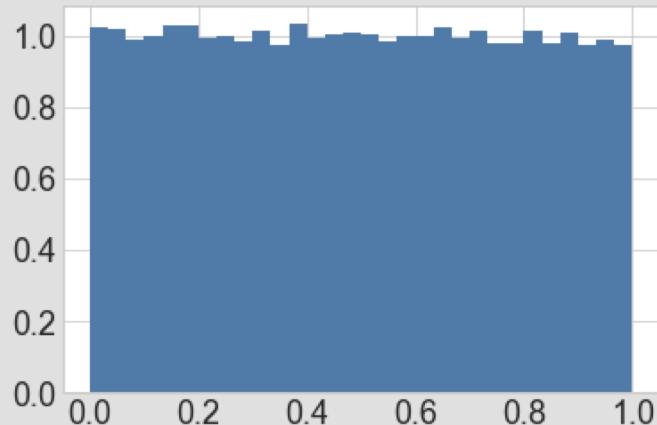
$$\begin{aligned}F_Y(y) &= P(Y \leq y) = P(F(X) \leq y) = \\&= P\left(X \leq F^{-1}(y)\right) = F_X\left(F^{-1}(y)\right) = y\end{aligned}$$

Функция распределения $F_Y(y) = y$ соответствует равномерному распределению на отрезке $[0; 1]$



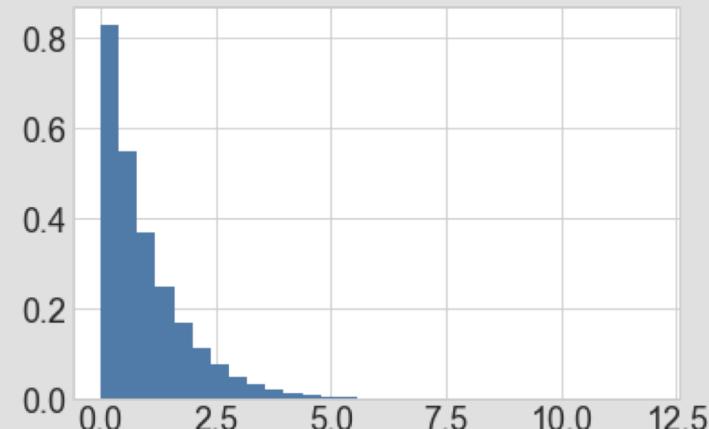
Квантильное преобразование

С помощью квантильного преобразования мы можем получить из равномерной случайной величины любую другую:



$$Y \sim U[0; 1]$$

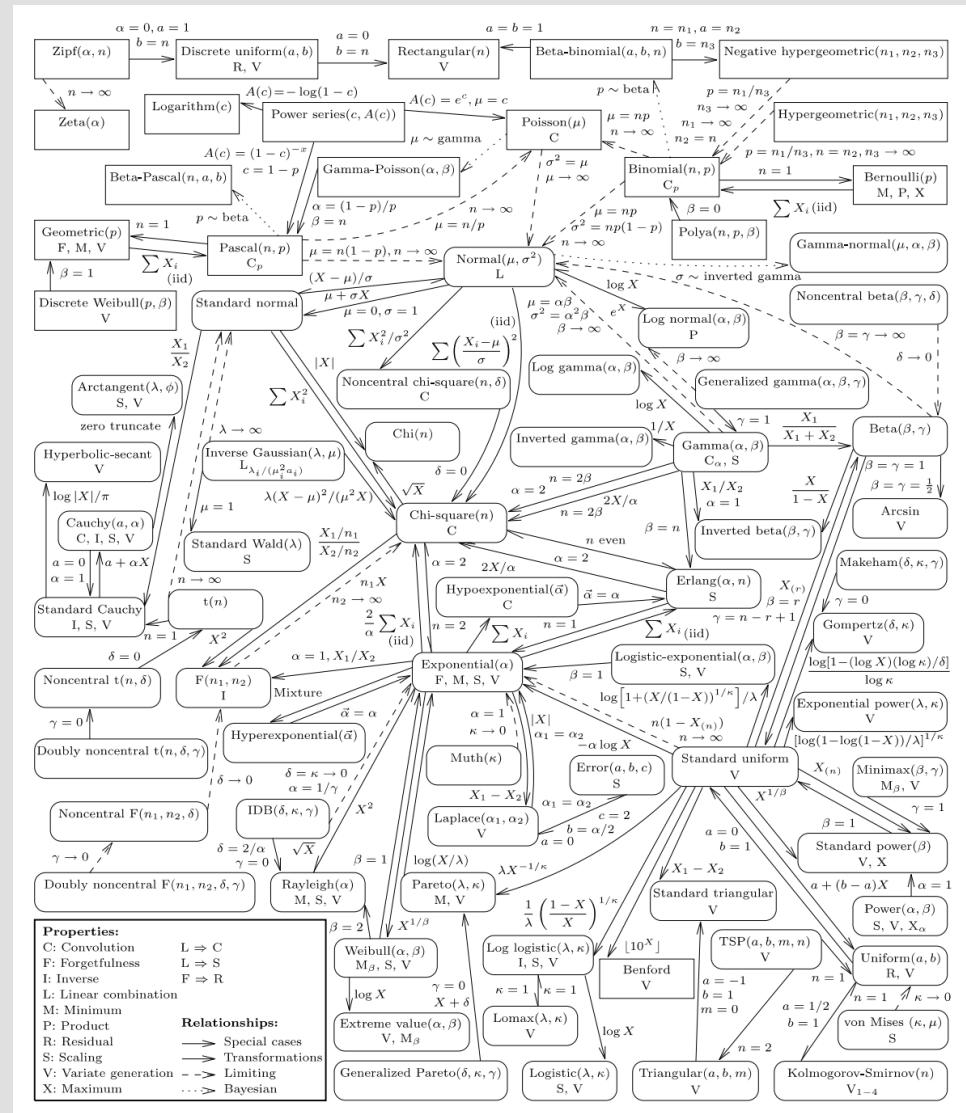
$$x = F^{-1}(y)$$



$$X \sim F(x)$$



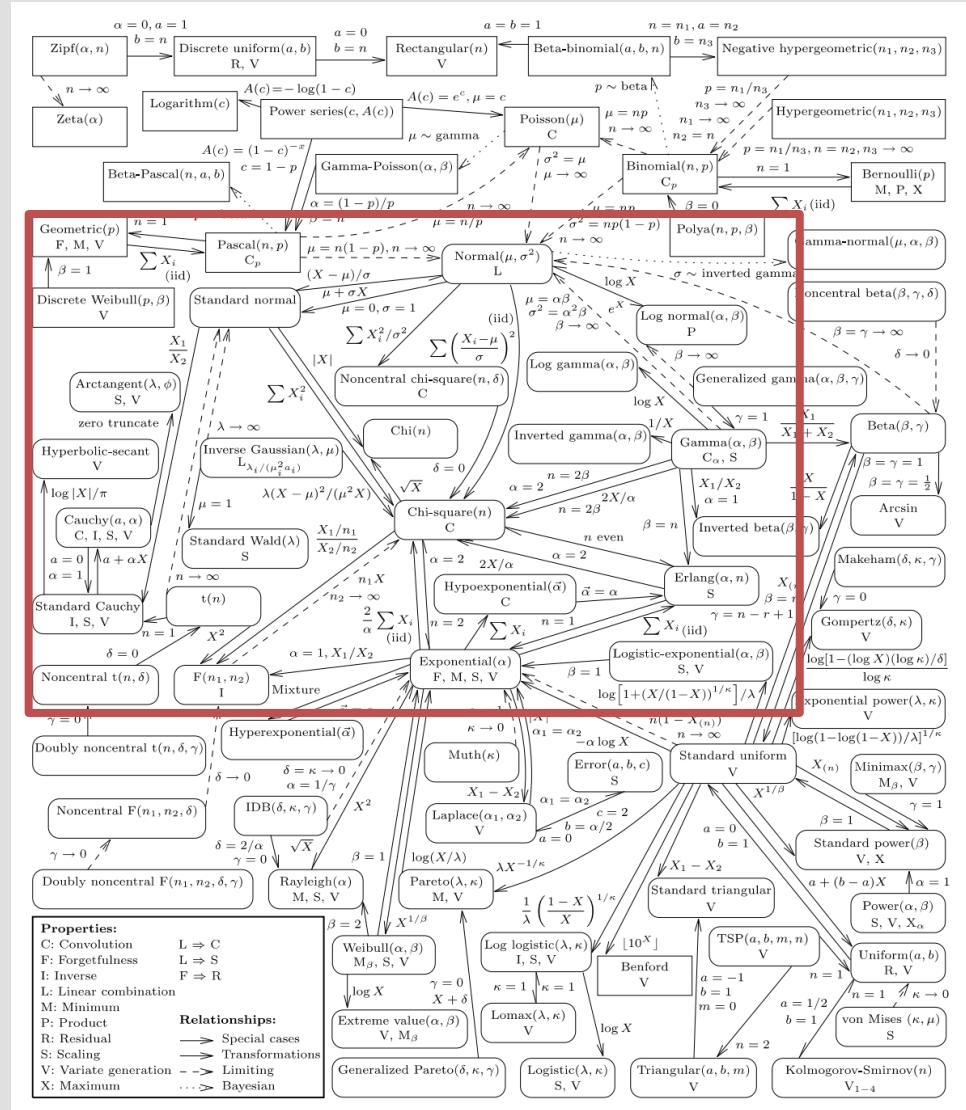
Всё переплетено



► <http://www.math.wm.edu/~leemis/2008amstat.pdf>



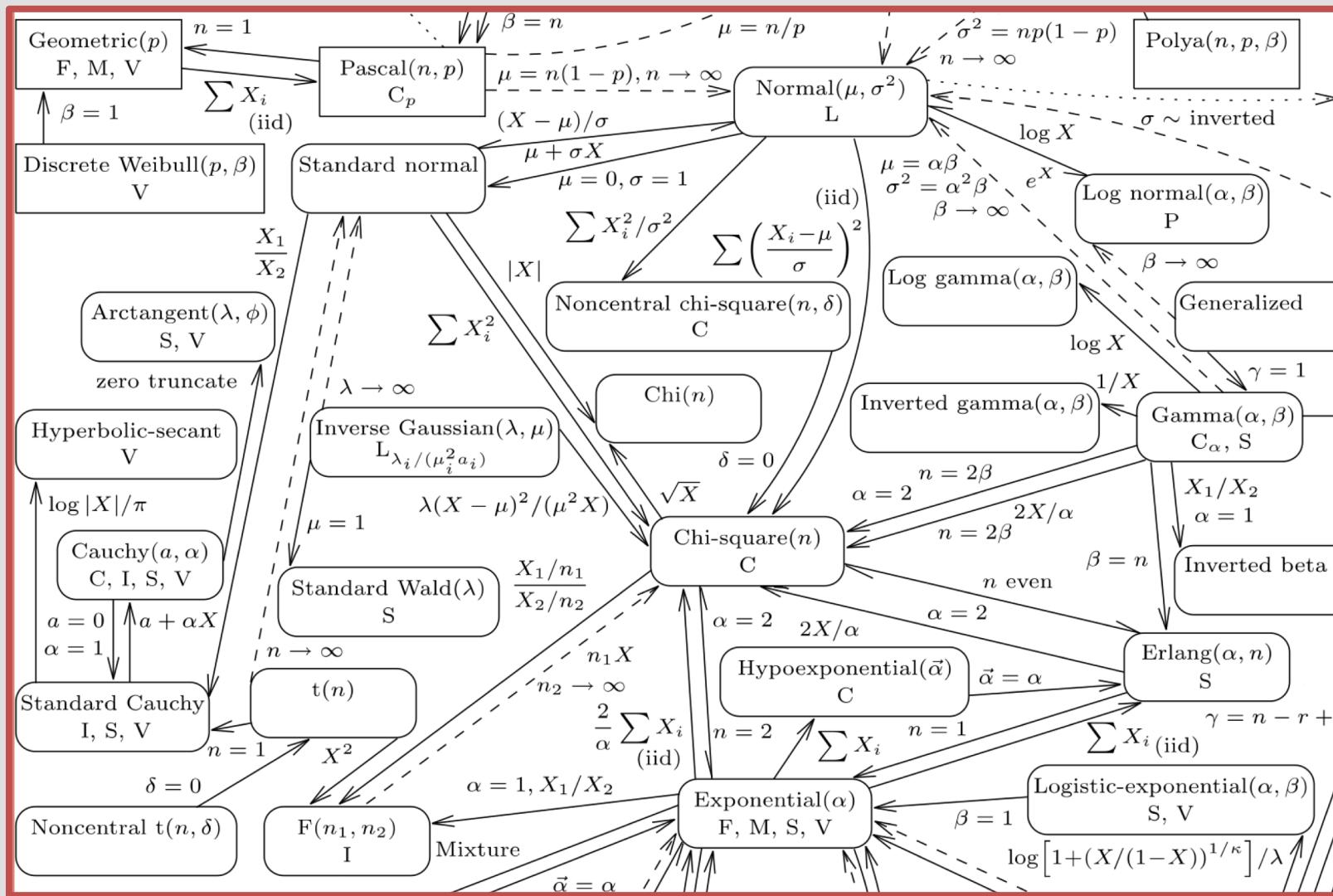
Всё переплетено



► <http://www.math.wm.edu/~leemis/2008amstat.pdf>



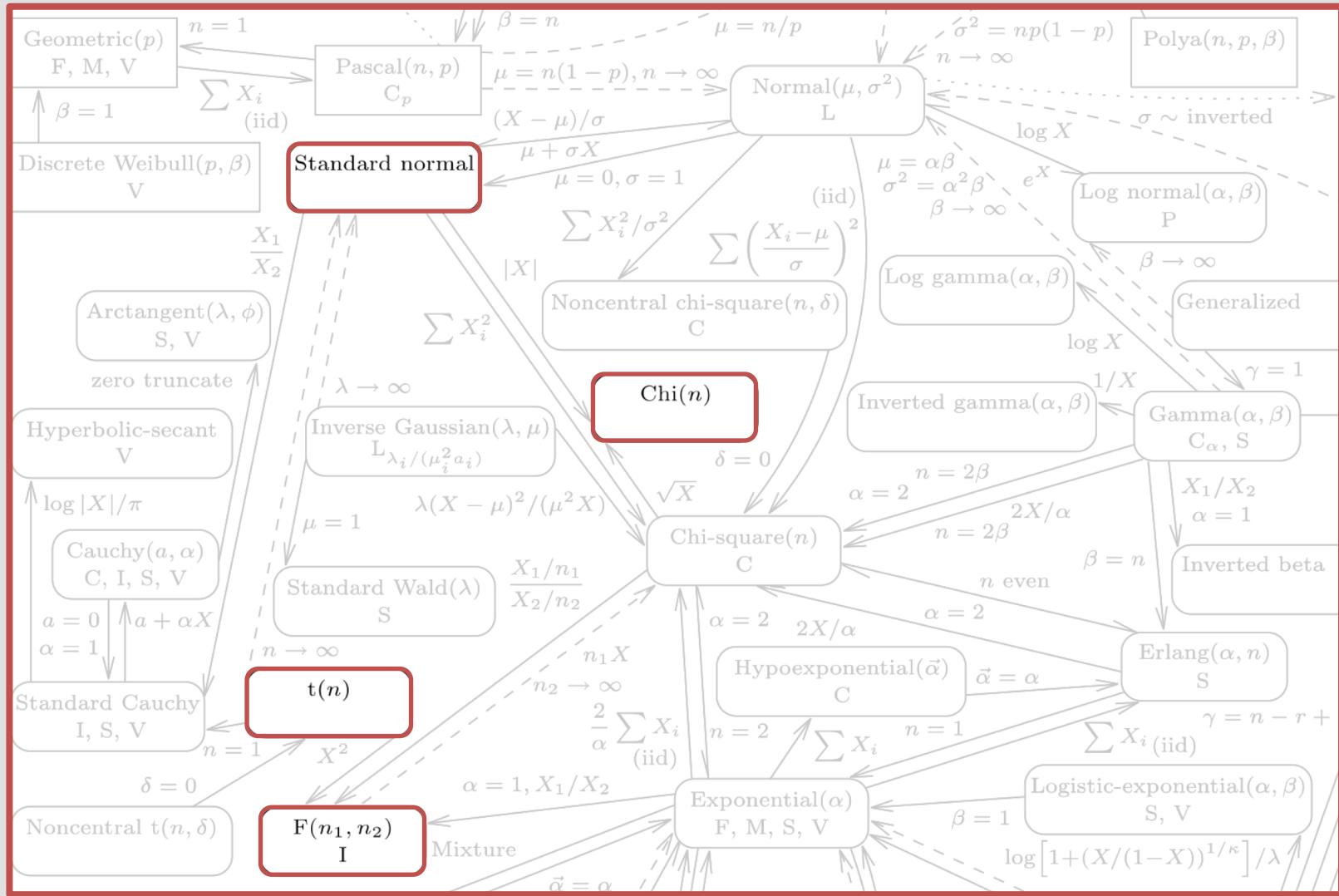
Всё переплетено



► <http://www.math.wm.edu/~leemis/2008amstat.pdf>



Всё переплетено



► <http://www.math.wm.edu/~leemis/2008amstat.pdf>



Резюме

- Квантильное преобразование помогает сгенерировать из равномерной случайной величины другие
- ЗБЧ и метод Монте-Карло помогают с помощью симуляций искать характеристики различных распределений
- Генерации не заменяют аналитических выкладок, так как они неэффективны, а также встречаются ситуации, где провести их очень сложно

