

Закон больших чисел (ЗБЧ)



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.

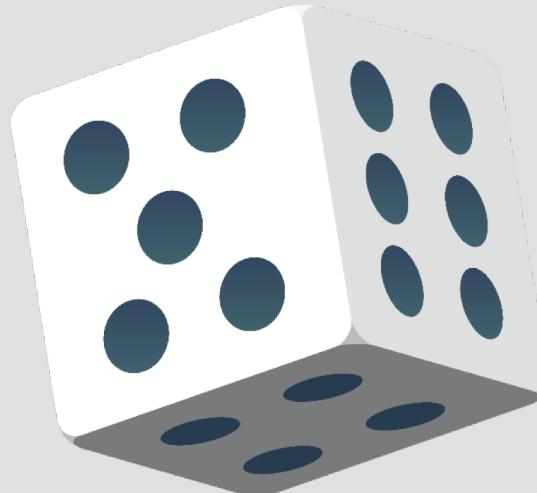
В этом нам помогает несколько теорем: ЗБЧ, ЦПТ и др.



Закон больших чисел (ЗБЧ)

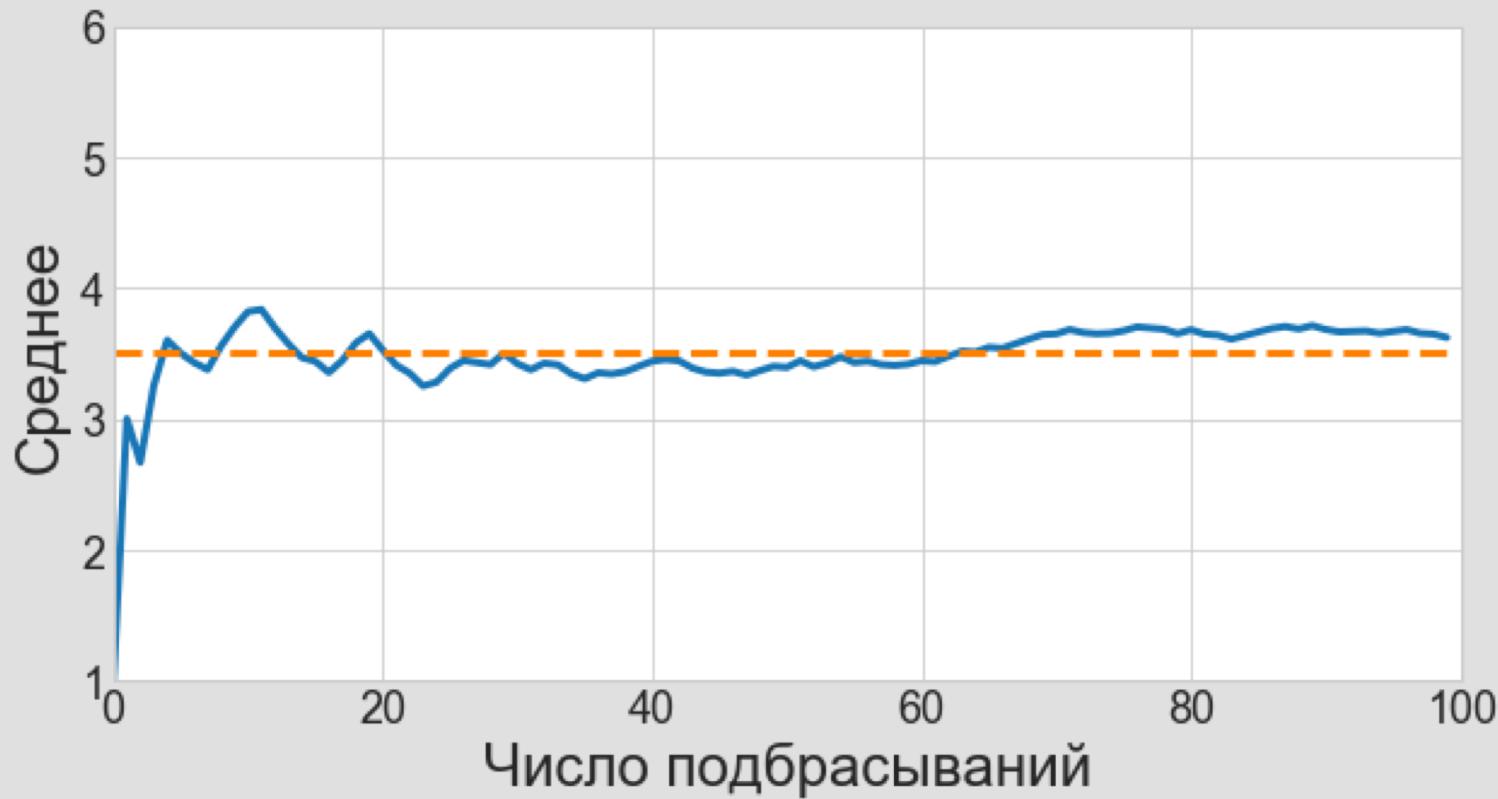
ЗБЧ говорит, что среднее арифметическое большого числа похожих случайных величин “стабилизируется” с ростом их числа

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



Резюме

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



Центральная предельная теорема (ЦПТ)



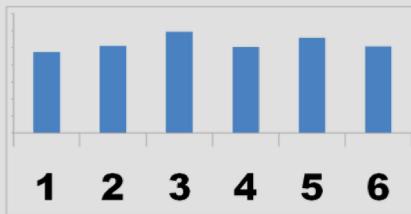
Центральная предельная теорема (ЦПТ)

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



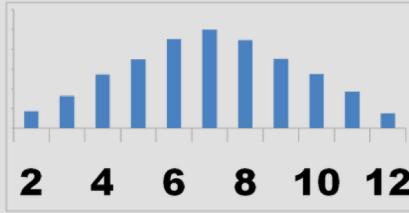
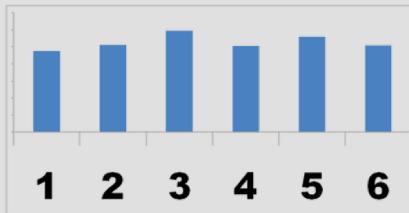
Центральная предельная теорема (ЦПТ)

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



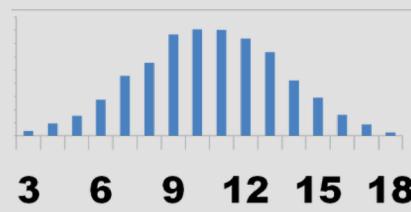
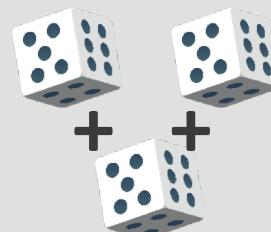
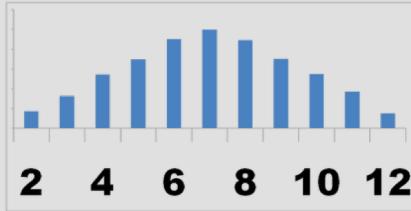
Центральная предельная теорема (ЦПТ)

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



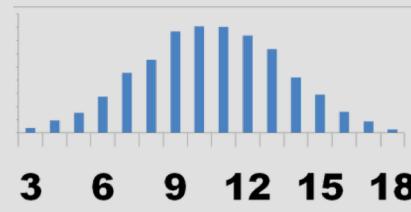
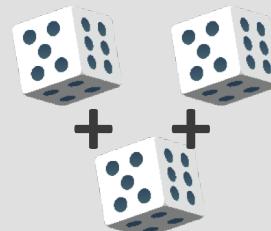
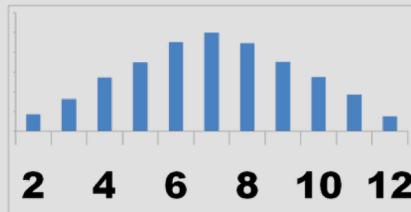
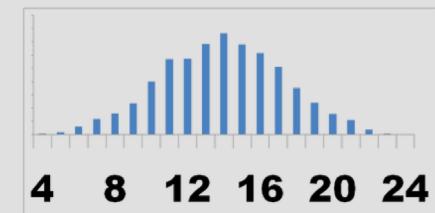
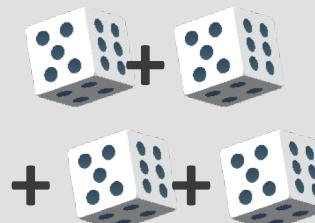
Центральная предельная теорема (ЦПТ)

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



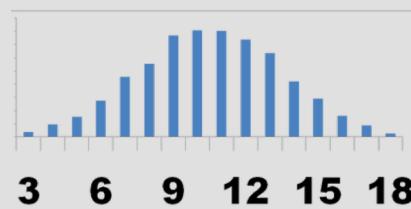
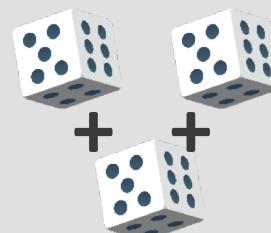
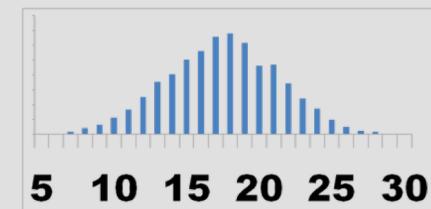
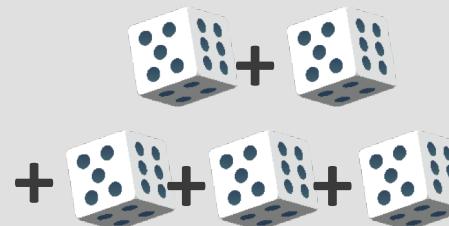
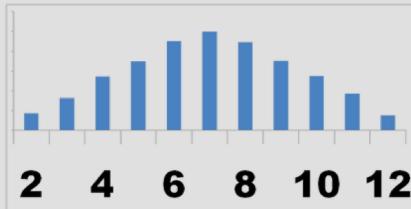
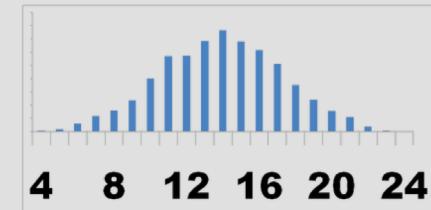
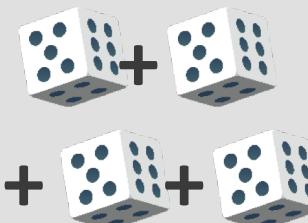
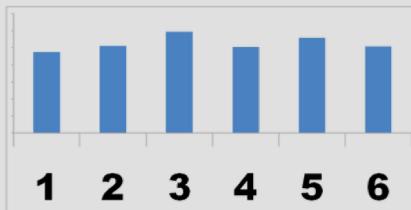
Центральная предельная теорема (ЦПТ)

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



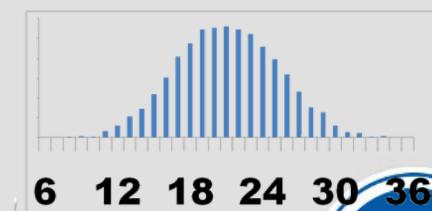
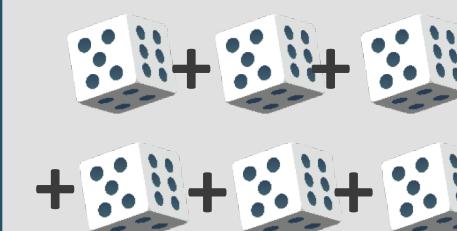
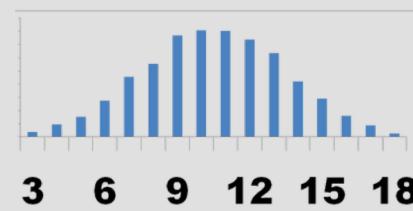
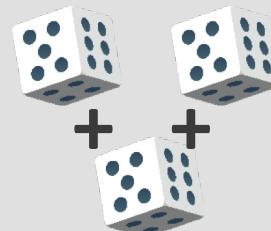
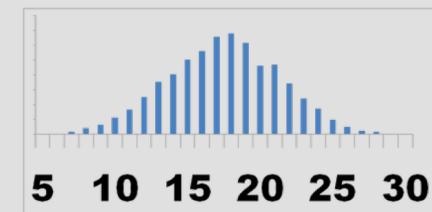
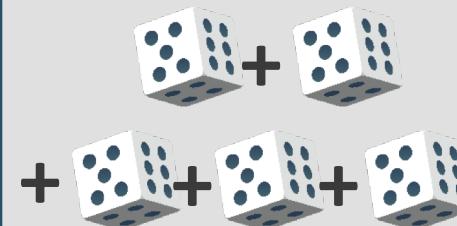
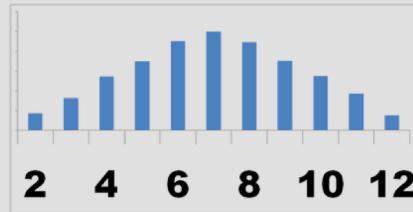
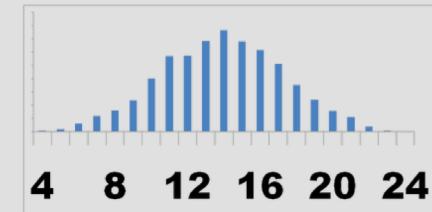
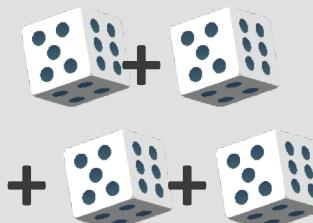
Центральная предельная теорема (ЦПТ)

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



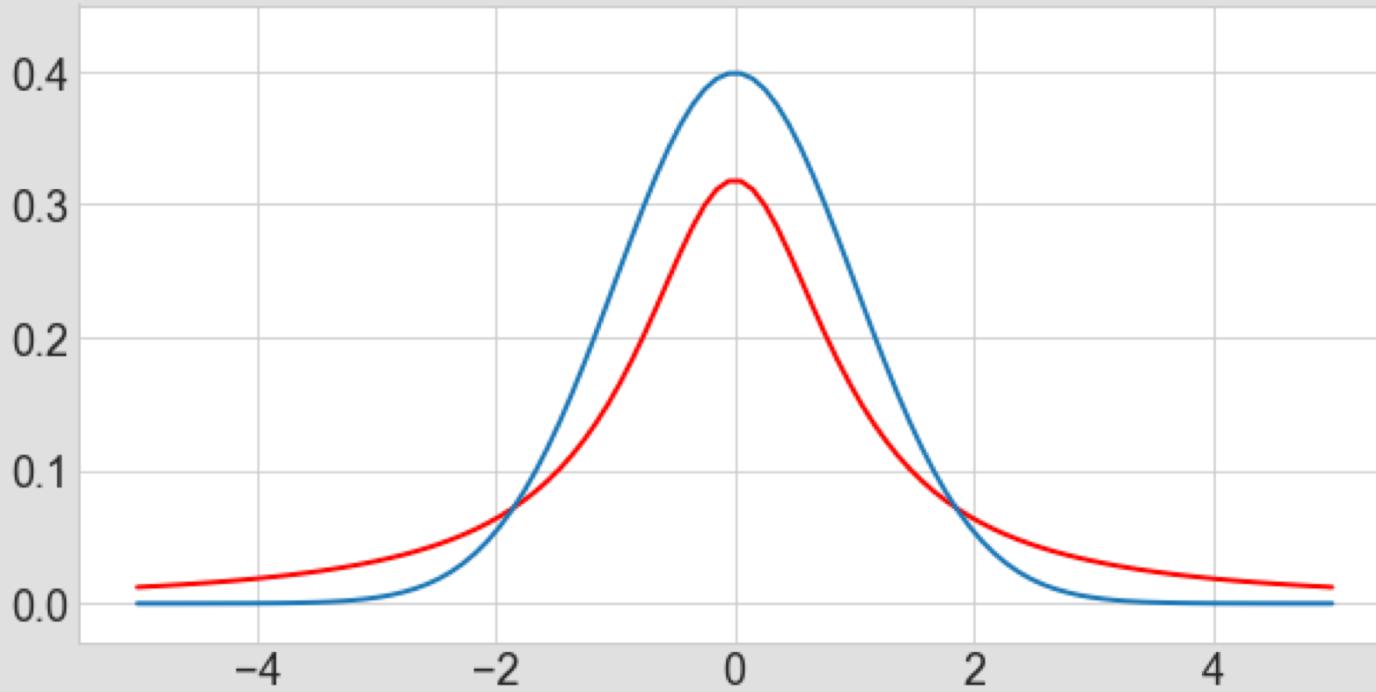
Крайнеземье и средиземье



ЦПТ и ЗБЧ работают в Средиземье



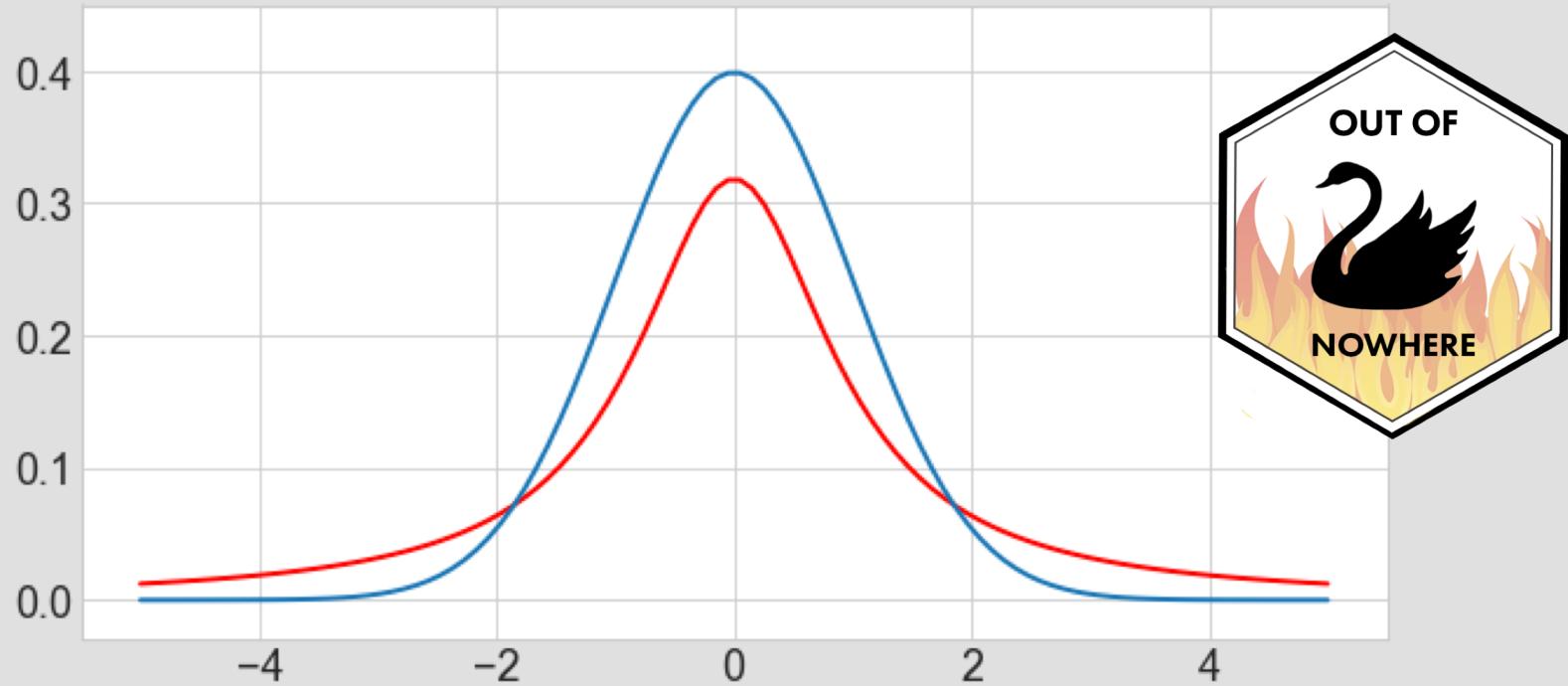
Крайнеземье и средиземье



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



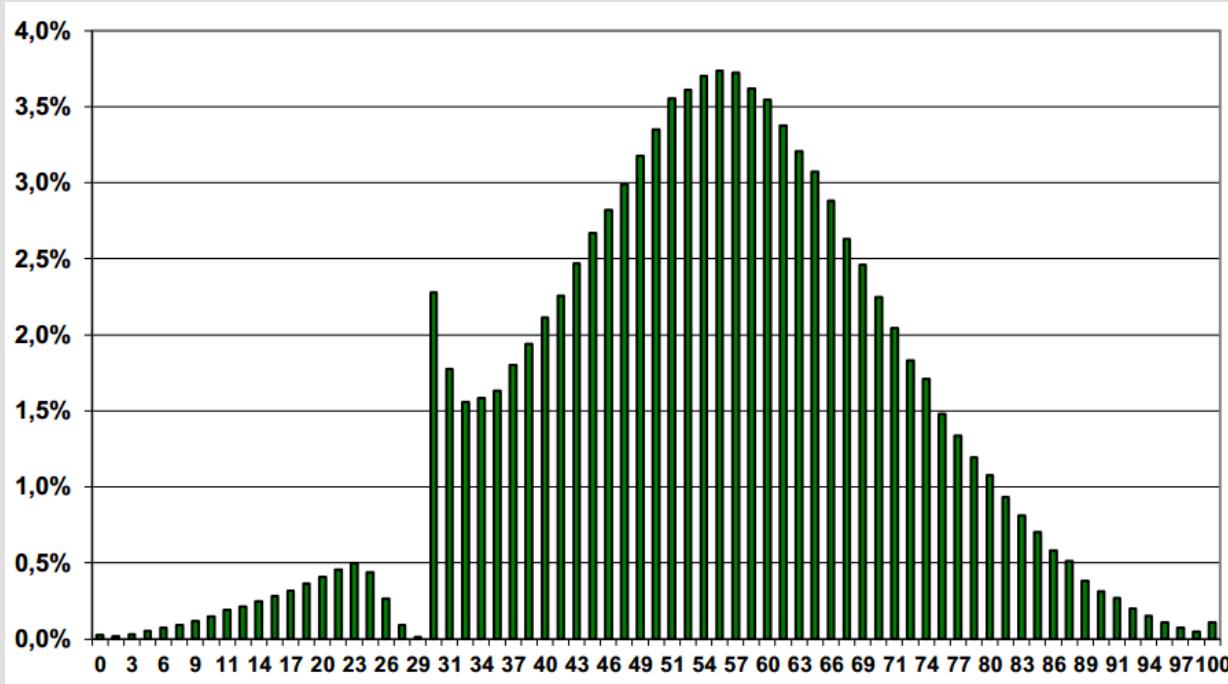
Крайнеземье и средиземье



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



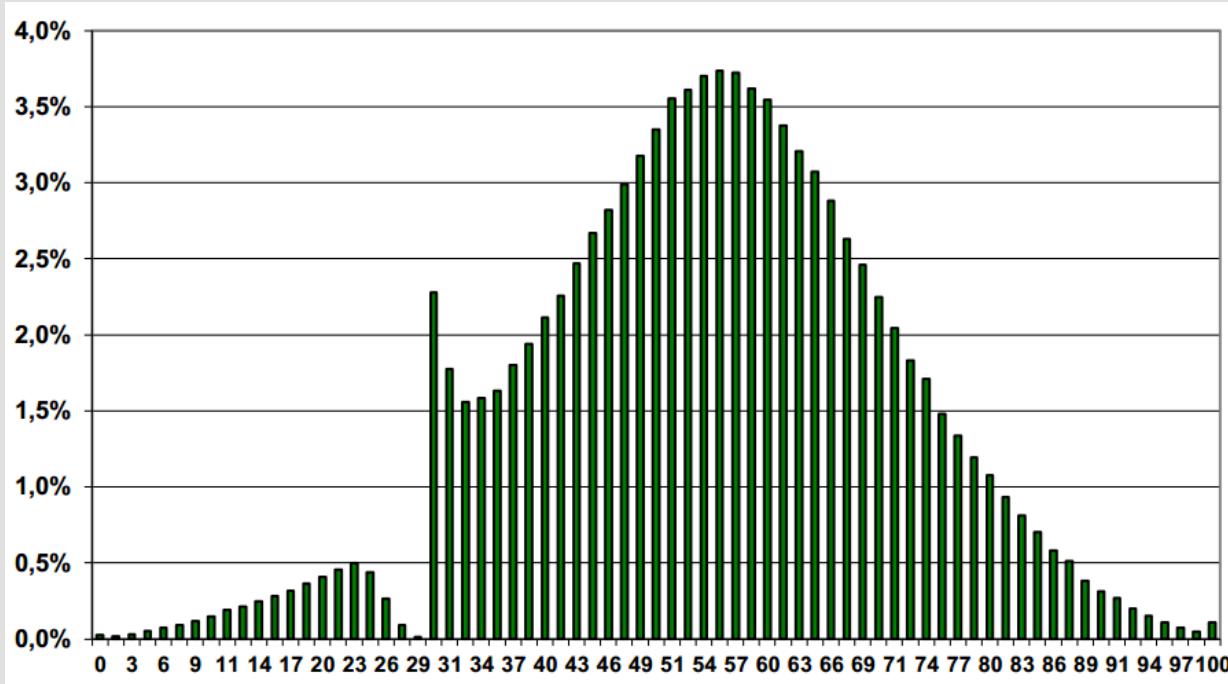
Польское ЕГЭ



- 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/



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



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:

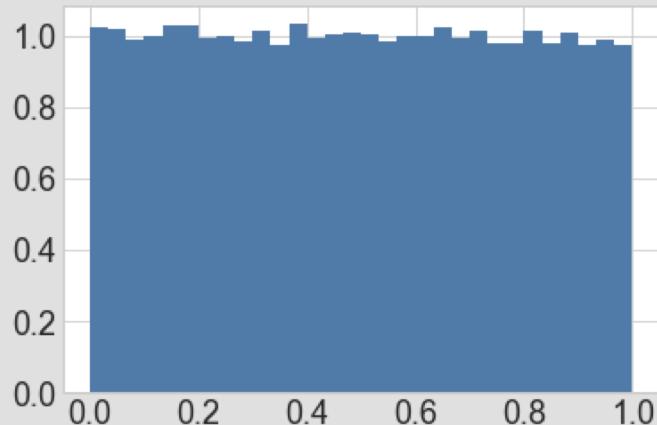
$$\begin{aligned}F_Y(y) &= P(Y \leq y) = P(F(X) \leq y) = \\&= P\left(X \leq F^{-1}(y)\right) = F_X\left(F^{-1}(y)\right) = y\end{aligned}$$

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



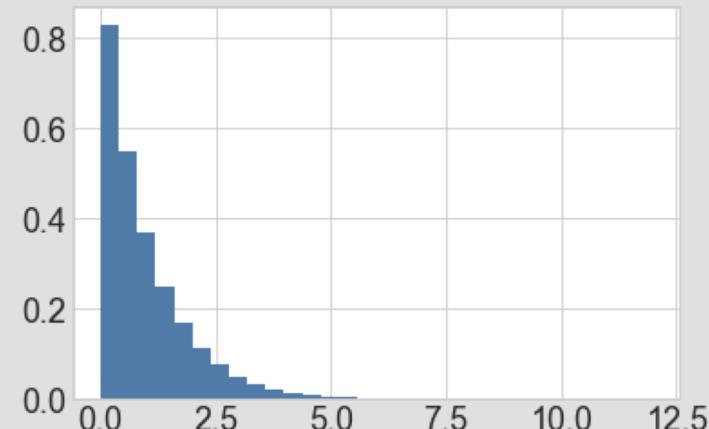
Квантильное преобразование

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



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

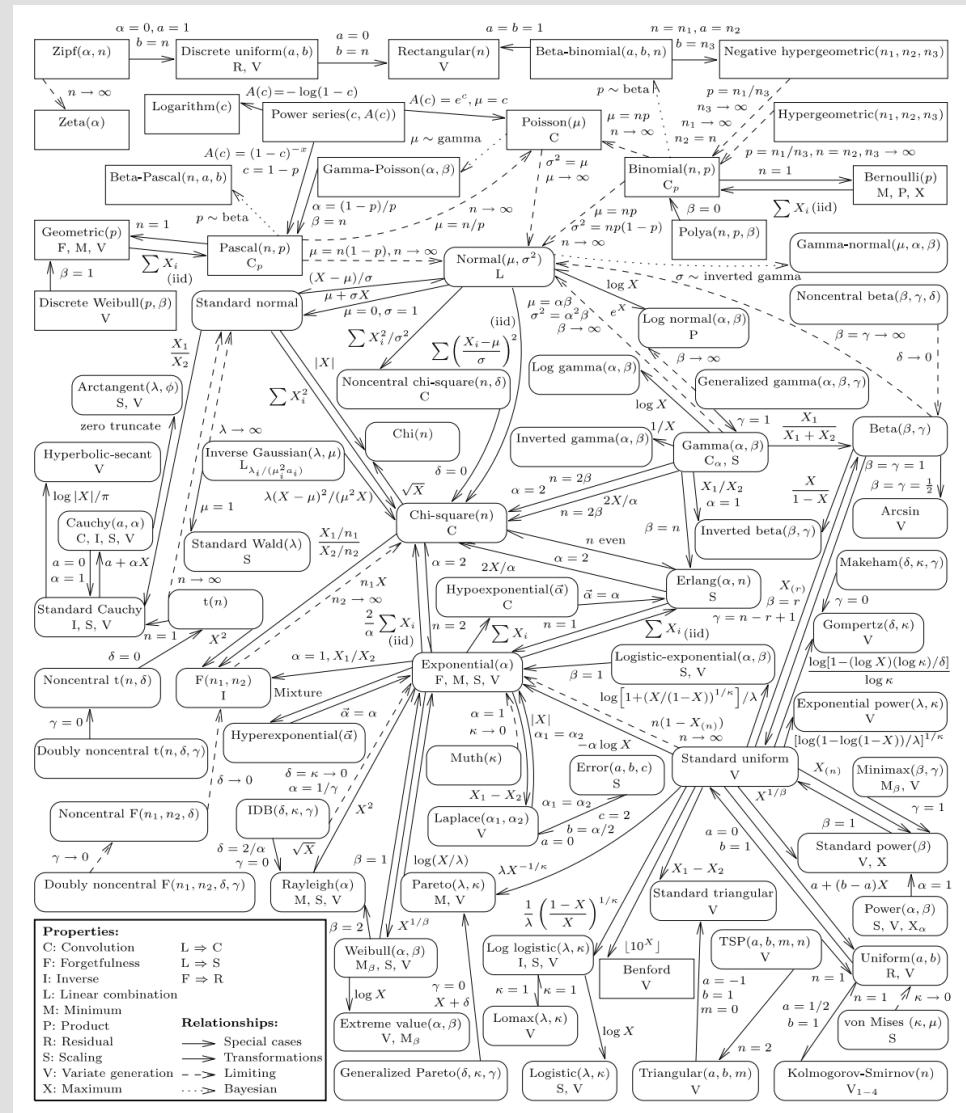
$$x = F^{-1}(y)$$



$$X \sim F(x)$$



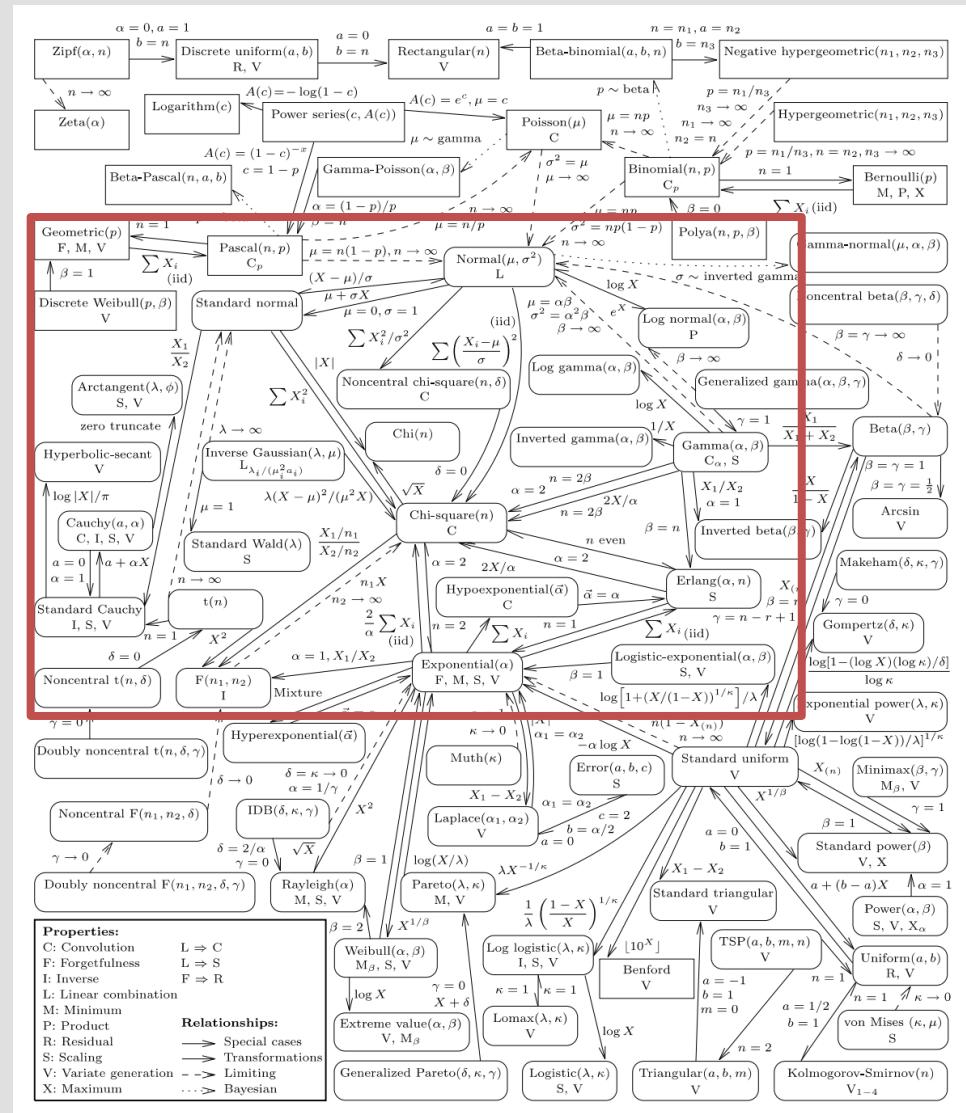
Всё переплетено



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



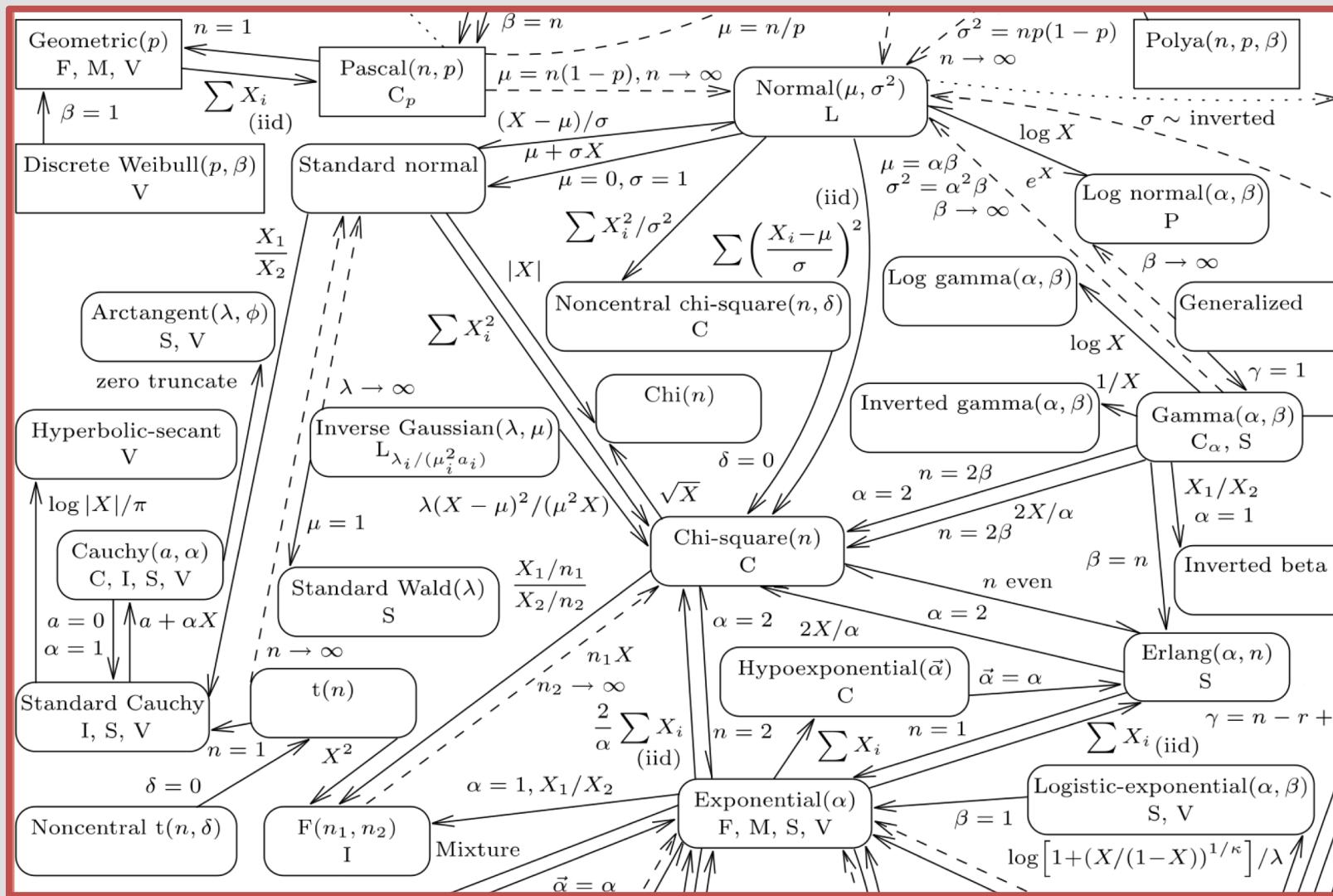
Всё переплетено



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



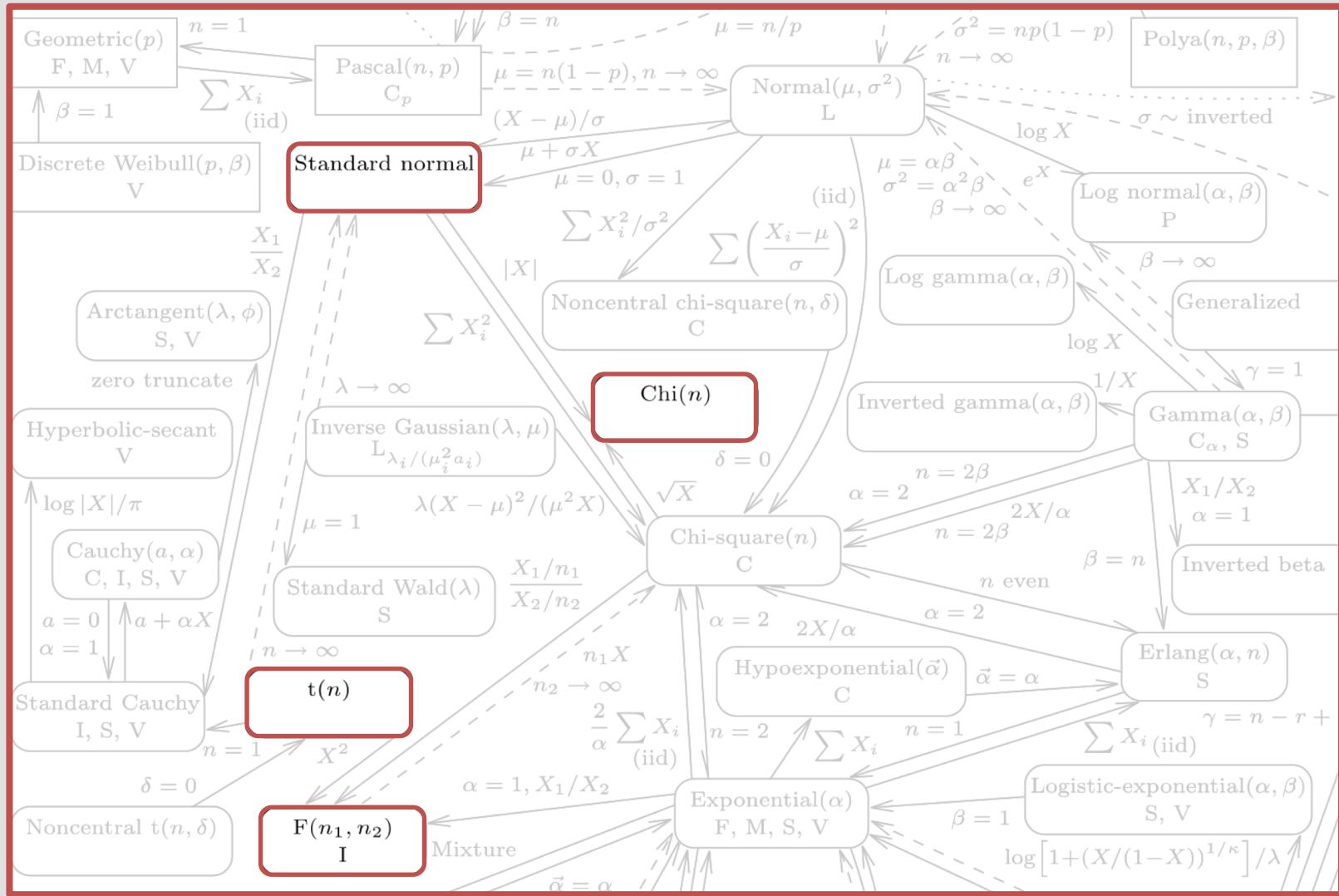
Всё переплетено



► <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

