

Карточка курса:

- Сергей Александрович Спириин
sspirin@hse.ru
 - Будет ссылка на диск
 - 3 кр по 1/15
 - 3 дз 2 по 1/10 + 1 1/5
 - экзамен 2/5
 - первая кр по теме лекции 1
 - м/б больше, чем 3 раза кр
- ex = "example" (пример)

① Статистика

описательная

- числовые кар-ки
- сегодня об этом

индуктивная = "простая статистика"

- Э генеральная совокупность и выборка
- по выборке делаем о ген. совов.
- её мот. и кол. в индуктивной статистике

математическая

② Разделы индуктивной статистики

- Оценка параметров
 - оценить среднее (точная оценка) ex
 - интервальная оценка
- Проверка гипотез
 - у совокупностей одинаковые средние (ex)

③ Мат. стат.

- Э случайная величина с распределением

④ Нужно знать теорию вероятностей для этого курса (след. лекция будет по теорверу)

Лекция 1

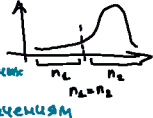
⑤ Кар-ки чисел

X_1, \dots, X_n - набор чисел

- среднее $\bar{X} = (X_1 + \dots + X_n) / n$

- медиана

→ при четном числе данных медиана = двум значениям



→ не средний доход чел-ка, а доход среднего чел-ка

- процентиля → медиана - 50й процентиль
- квантили
- децимы, квартили, межквартирный размах
- среднее квадратичное отклонение

⑥ Графические кар-ки

- гистограмма + совместная гистограмма

25 - карман от 0 до 25
50 - карман от 25 до 50

→ шаг м/б переменным

→ как выбрать карман и шаг?

! главная проблема

- при неправильном шаге можно не заметить полимодальности данных

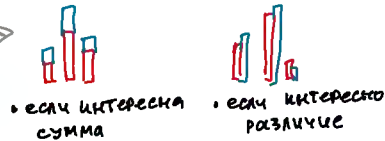
! творческая задача

- м/б стоит перейти к логарифмической шкале?
- где больше видно различий?

→ карман = бин ("bin")

- есть несколько способов выбрать число бинов:

$$k = \lceil \sqrt{n} \rceil, \\ k = \lfloor \log_2 n \rfloor + 1 \text{ и т.д.}$$



- ящик с усами (бокс-плот)

→ для сравнения нескольких наборов (2 и более)

→ разные типы, надо добавлять подписи (описание) для статьи

→ может лучше для логарифмов сделать? проверить!

- график оценки плотности распределения (kde plot)

→ сглаженная гистограмма

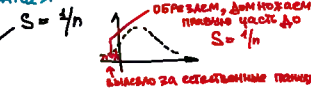
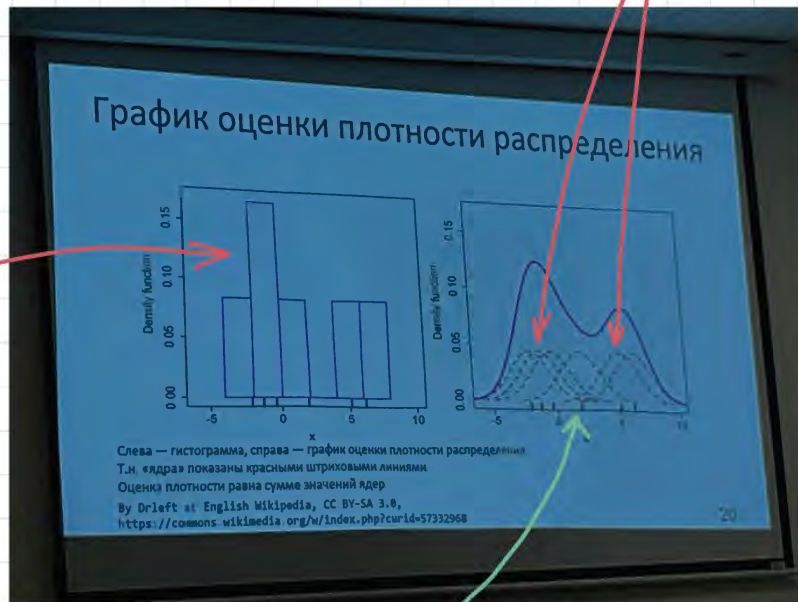


График оценки плотности ядра


- альтернатива гистограммам
- более наглядной

сумма столбцов
равна
единице



сумма площадей под каждой кривой = 1

В штолфовые каро водятся шот,
а в градных распространение
лучше водятся ширину едра.

→ Сложение ядер (гауссовых)


• скрипичная диаграмма

→ широта, ящика с усами
и плотности

→ имеет преимущества
того и другого

→ симметричная

• эмпирическая функция
распределения

→ y - доля чисел, меньше
данного x

→ нет потери информации

→ не оч. информативно,
но иногда полезно

• точечная диаграмма (scatter plot)

→ для пары значений
(сравнение длин белков)

→ две числовые хар-ки одного
и того же объекта
(одинак. белки у разных бактерий)

Лекция 1

7 Теория вероятностей

→ Случайное событие $0 \leq P \leq 1$

→ Случайная величина - число

→ $A \cdot B = A \cap B$ (пересечение), "и"

→ $A+B = A \cup B$ (объединение), "или"

→ независимые события:

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

• вероятность выпадения 6 на
двух кубиках

→ условная вероятность:

$$P(A|B) = P(A \cdot B) / P(B) - \text{завис. события}$$

$$P(A|B) = P(A) - \text{для независимых событий}$$

→ ξ -ки, η -ята

8 Формула полной вероятности

$$P(A) = P(A|H_1) \cdot P(H_1) + \dots + P(A|H_n) \cdot P(H_n)$$



9 Формула Байеса

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

↑
"Рот А при условии В"

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

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

10 Задача

A - употребляет, \bar{A} - не употр.

B - тест "+"

$$P(A) = 0,005$$

$$P(A|B) = ?$$

Решение:

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

$$P(B|A) = 0,99$$

$$\rightarrow P(B) = \underbrace{P(B|A) \cdot P(A)}_{0,99 \cdot 0,005} + \underbrace{P(B|\bar{A}) \cdot P(\bar{A})}_{(1-0,99) \cdot 1-0,005}$$

$$P(B) = 0,00495 + 0,0199 = 0,02485$$

$$P(A|B) = \frac{0,99 \cdot 0,005}{0,02485} \approx 0,2$$

↑
80% "рез. - ложные"

КР на след. занятии после перерыва:

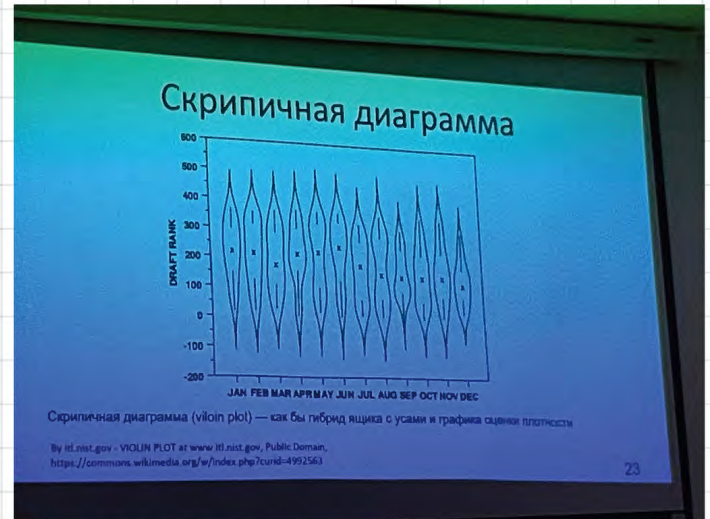
• дана гистограмма, в каких
пределах медиана и квартиль?



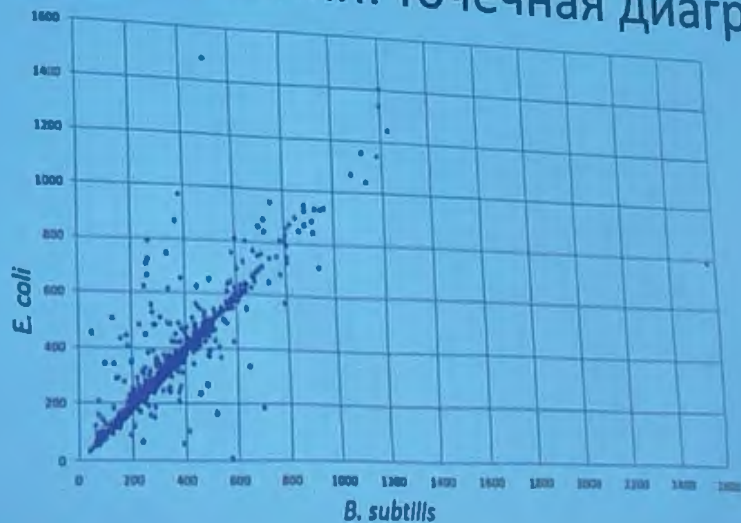
• задача на формулу
Байеса

• задача на распределение
Пуассона и на
биномиально P

В скрипичной диаграмме (violin plot) изображается плотность распределения (в этом примере — по сравнению с усреднением с усреднением).



Пары значений: точечная диаграмма



Пары значений: характеристики

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

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

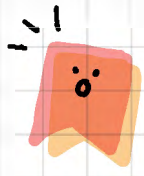
среднеквадратичное
отклонение

Задача

Предположим, что определенный тест на наркозависимость обладает 99% чувствительностью и 98% специфичностью, то есть тест даёт положительный результат для 99% потребителей наркотиков, и даёт отрицательный результат для 98% не-потребителей наркотиков.

Предположим, некая корпорация узнала, что 0,5% её сотрудников используют наркотики и решает проверить своих сотрудников на наркозависимость.

Для некоторого сотрудника тест дал положительный результат. Какова вероятность того, что этот сотрудник на самом деле употребляет наркотики?



Про случайные события ...
Решение задачи:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)} = \frac{0,005 \cdot 0,99}{0,02485} \approx 0,199 \approx 20\%$$

$$\begin{aligned} P(A) &= P(A|B) \cdot P(B) + P(A|\bar{B}) \cdot P(\bar{B}) = \\ &= 0,99 \cdot 0,005 + 0,02 \cdot 0,995 = \\ &= 0,02485 \end{aligned}$$

11) Распределение случайных величин

↓
дискретные непрерывные

- Биномиальное распред.

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

! ФОРМУЛА!

→ показывает сколько раз произошло то или иное событие

15) Дисперсия

→ $Var()$ или $\Phi()$

$$\begin{aligned} \rightarrow Var(\xi) &= E[(\xi - E(\xi))^2] = \\ &= E(\xi^2) - (E(\xi))^2 \end{aligned}$$

→ насколько разбросана величина

12) Интегральная вероятность

?

13) Задача

$= \text{BINOMDIST}(2, 6, 5/6, \text{TRUE})$
решение в EXCEL

14) Математическое ожидание

• E - "estimate" or M - "mean"

• $\xi + \eta$ - сложение случайных величин

• мат. ожидание суммы случайных величин

сумме = мат. ожиданий

$$E(\xi + \eta) = E(\xi) + E(\eta)$$

→ среднее значение по большому числу испытаний

→ монетки: мат. ожидание = 0.5
кубики: м.о. = 3.5

Лекция 1

16) Свойства

→ Cov - ковариация

Математическое ожидание и дисперсия

ξ = случайная величина

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

* упражнения для умных, для КР не надо

- (17) Ковариация
 → у независ. величин = 0
 → у завис. иногда = 0

- (18) Свойства биномиального распределения

$$E(k) = np$$

$$\text{Var}(k) = np(1-p)$$

- (19) Считаем мат. ожидание:

ξ	0	1	значение
	$1-p$	p	вероятность

$$E(\xi) = p$$

Дисперсия:

$\xi - E(\xi)$	$-p$	$1-p$
	$1-p$	p

$$\text{Var}(\xi) = E((\xi - E(\xi))^2) =$$

$$= (1-p) \cdot p^2 + p(1-p)^2 = p(1-p)$$

Лекция 1

→ поймали 100 рпв, там 30 паразитов
 заражённость = $\frac{3}{10}$
 в каких пределах это число может меняться?
 • считаем дисперсию, а потом доверительный интервал

- (20) Распределение Пуассона → бином. распредел. с бесконечным числом испытаний

- Сколько на нас вечером съедет комаров?
 → Сколько растений в каждом м² на лугу?

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

$$e \approx 2,718$$

- только целые неотриц. значения
 → Дисперсия = мат. ожиданию

$$E(\xi) = \text{Var}(\xi) = \lambda$$

- (21) Задача

- покрытие = 5, т.е. каждый нуклеотид покрыт в среднем пятью рчдами

$$P(k) = \frac{5^k \cdot e^{-5}}{k!}$$
 ↑
 число рчдов

$$P(0) = e^{-5}$$

 → на практике хотя бы 3 должно быть покрытие
 → E.coli геном \approx 4 млн п.н.
 → тут нужно покрытие 15
 → $1/148$ - доля нуклеотидов, не покрытых чтением вообще

Пример:

Бросили монетку
10 раз. Считаем вероят-
ность выпад. орлов.

Биномиальное распределение

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

$$Pr(K = k) = \binom{n}{k} p^k \underbrace{(1-p)^{n-k}}_{\text{вероятность неудач в остальных испытаниях}} \quad \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

вероятность успехов

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



$p=0,1$



$p=0,3$



$p=0,5$

Биномиальное распределение

Какова вероятность выбросить не менее 5 шестёрок при шести бросках кости?

$$P(K=5) + P(K=6) = \binom{6}{5} \left(\frac{1}{6}\right)^5 \cdot \frac{5}{6} + \binom{6}{6} \cdot \left(\frac{1}{6}\right)^6$$

т.е. выпадает либо 5 шестёрок,
либо 6 шестёрок

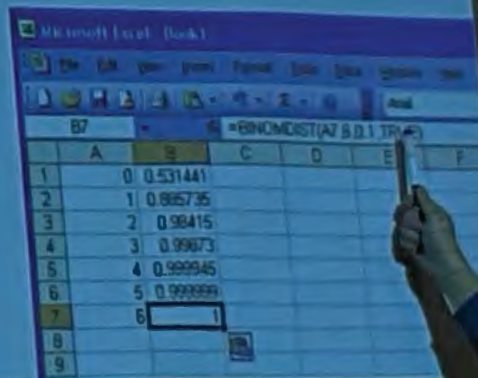
Интегральная вероятность

ξ = случайная величина

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

Большинство программных средств обработки данных содержит средства вычисления интегральной вероятности

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



Microsoft Excel (Book1)

Formula bar: `=BINOMDIST(A7:B7, 5, 0.1, TRUE)`

	A	B	C	D	E	F
1	0	0.531441				
2	1	0.865735				
3	2	0.98415				
4	3	0.99873				
5	4	0.999945				
6	5	0.999999				
7	6					
8						
9						

Математическое ожидание

- то, к чему будет стремиться среднее значение

ξ	x_1	x_2	x_n
p	p_1	p_2	p_n

$E(\xi) = \sum x_i p_i =$ не случайная величина
(другое обозначение $M\xi$)

Пример: два броска монеты
(орёл - 1, решка - 0).
Мат. ожид. = $\frac{1+0}{2} = \frac{1}{2}$.

ξ	0	1
p	1/2	1/2

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

η	0	1
p	1/3	2/3

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

$\xi + \eta$	0	1	2
p	1/6	1/2	1/3

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

Дисперсия

$$\text{Var}(\xi) = E[(\xi - E(\xi))^2] = E(\xi^2) - (E(\xi))^2$$

другое обозначение $D\xi$

ξ	0	1
p	1/3	2/3

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

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

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

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

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

ξ^2	0	1
p	1/3	2/3

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

$$\text{Var}(\xi) = E(\xi^2) - E^2(\xi) = 2/3 - 4/9 = 2/9$$

- это мат.
ожидаемые
квадрата
отклонения
от мат.
ожидаемых

Математическое ожидание и дисперсия

ξ — случайная величина

- $E(\xi + \eta) = E(\xi) + E(\eta)$

- $E(c\xi) = cE(\xi)$

- $E(c) = c \leftarrow \text{мат. ожидание константы } (c) = c$

- если ξ и η независимы, то $E(\xi \eta) = E(\xi)E(\eta)$

- $\text{Var}(\xi) = E(\xi^2) - E^2(\xi)$

- $\text{Var}(c\xi) = c^2 \text{Var}(\xi)$

- если ξ и η независимы, то $\text{Var}(\xi + \eta) = \text{Var}(\xi) + \text{Var}(\eta)$

- для общего случая $\text{Var}(\xi + \eta) = \text{Var}(\xi) + \text{Var}(\eta) + 2\text{Cov}(\xi, \eta)$

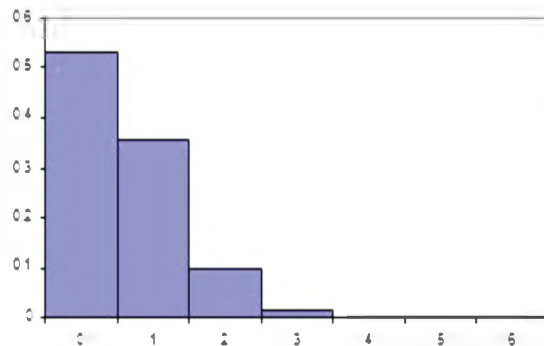
- $\text{Var}(\xi - \eta) = \text{Var}(\xi) + \text{Var}(-\eta) = \text{Var}(\xi) + \text{Var}(\eta)$

Биномиальное распределение

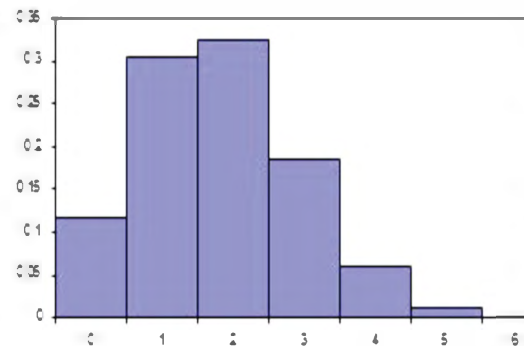
Биномиально распределённая величина = число успехов в n независимых испытаниях; параметр распределения p = вероятность успеха в одном испытании

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

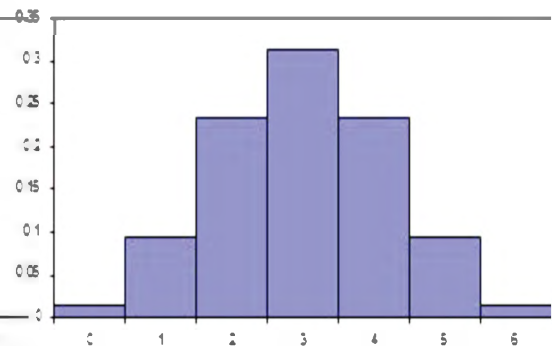
Вероятности 0, 1, ..., 6 успехов при шести независимых испытаниях



$p=0,1$



$p=0,3$



$p=0,5$

Биномиальное распределение

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

$$E(K) = np$$

$$\text{Var}(K) = np(1-p)$$

Распределение Пуассона

Случайная величина, распределённая по Пуассону = число (достаточно редких) событий за (достаточно большой) промежуток времени или в (достаточно большой) области пространства.

Имеет один параметр: λ — среднее число событий.

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

Вероятность наблюдать ровно k событий:

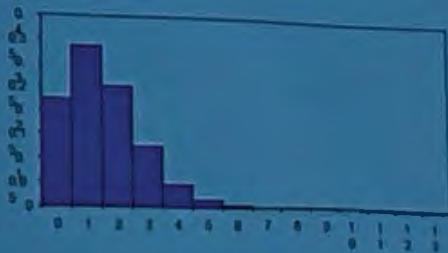
Если ξ — случайная величина, распределённая по Пуассону с параметром λ , то $E(\xi) = \lambda$ и $\text{Var}(\xi) = \lambda$ (для распределения Пуассона мат. ожидание равно дисперсии)

Распределение Пуассона

Случайная величина, распределённая по Пуассону = число (достаточно редких) событий за (достаточно большой) промежуток времени или в (достаточно большой) области пространства.

Имеет один параметр: λ — среднее число событий.

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



$\lambda=1,5$



$\lambda=5$

Распределение Пуассона:

- Сколько комаров укусит тебя за минуту?
- Сколько лампочек перегорит за сутки?

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

k - число событий

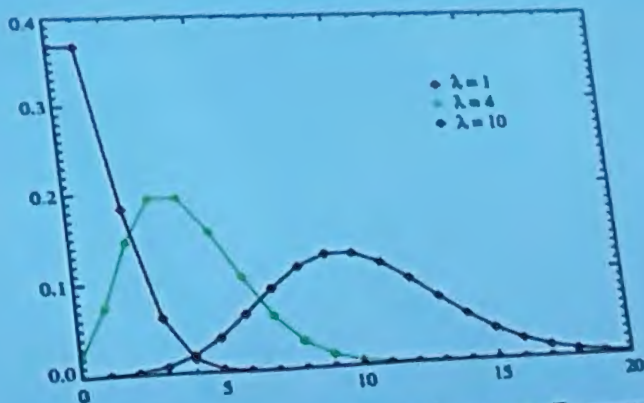
λ - среднее
число событий

Это я придумала. Возможно, это правильный пример.



Котик выиграл 10 котятней,
но, в сферике, точно выигрывает
7. Какова вероятность сегодня
побить свой рекорд?

Распределение Пуассона



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

Если ξ — случайная величина, распределённая по Пуассону с параметром λ ,
то $E(\xi) = \lambda$ и $\text{Var}(\xi) = \lambda$
(для распределения Пуассона мат. ожидание равно дисперсии)

! Дисперсия распредел.
Пуассона = λ

Сборка чтений на геном

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

$$P(K=0) = \frac{1 \cdot e^{-5}}{1} \approx \frac{1}{148}$$

$$P(K=1)$$

...

$$P(K=k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!}$$

← вероятность того, что
на 1 нуклеотид не попадет
ни одного чтения

k - число чтений
(покрытие)