Лекция 2 Оценки разновидности и свойства

Жигалов Августин

Data Engineer

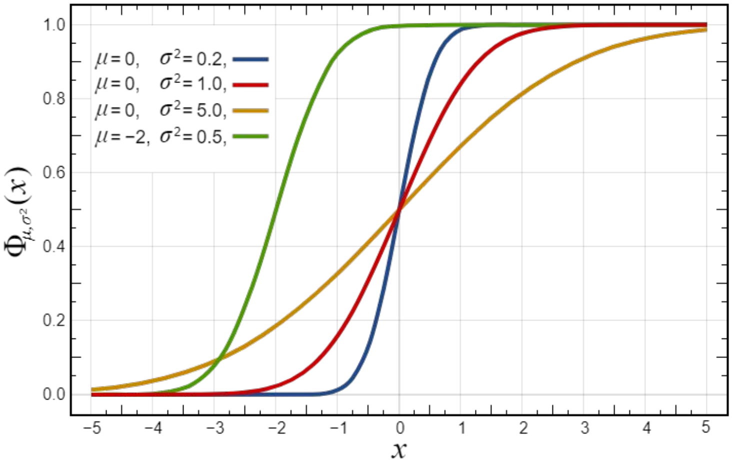
РСХБ

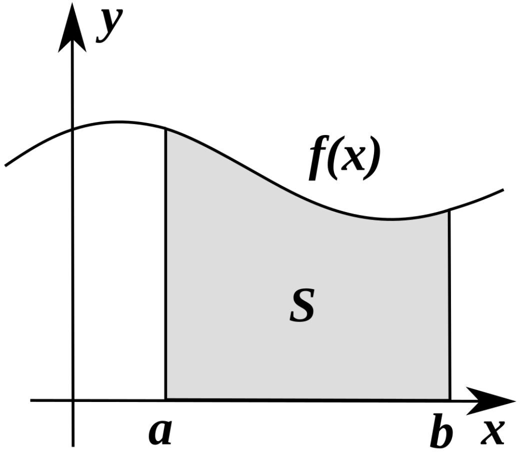
План лекции

- Распределения случайных величин
- Свойства оценок
  - Несмещенность
  - Состоятельность
  - Эффективность
- Интервальные оценки
- Домашнее задание

Распределения случайных величин

Фу́нкция распределе́ния в теории вероятностей — функция, характеризующая распределение случайной величины; вероятность того, что случайная величина X примет значение, меньшее или равное x
Функция плотности вероятности показывает распределение целевых значений





Какими бывают случайные величины

Распределение Бернулли

ИТМО

- Выпадение орла или решки при броске монеты

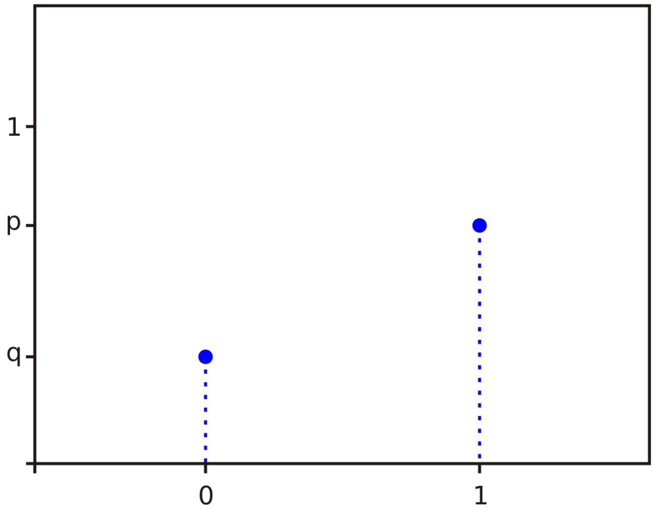
	решка	орел
X	0	1
P(X=k)	1-p	p

$$\mathbb{E}(X) = 1 \cdot p + 0 \cdot (1-p) = p$$

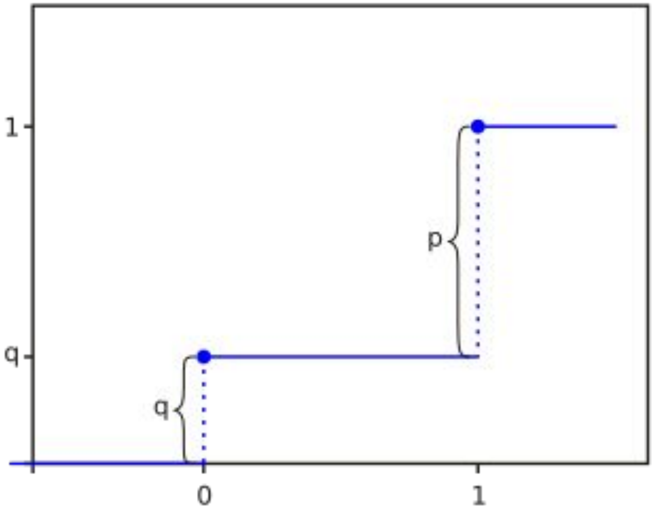
$$Var(X) = E(X^2) - E^2(X) = p - p^2 = p \cdot (1-p)$$

Распределение Бернулли

Распределение вероятностей



Функция распределения



Биномиальное распределение

- Число попаданий
  в баскетбольную корзину

Биномиальная случайная величина: $X \sim Bin(p, n)$

n – число испытаний

p – вероятность успеха

k принимает значения от 0 до n

Биномиальное распределение

ИТМО

$$Y_i \sim Bern(p)$$

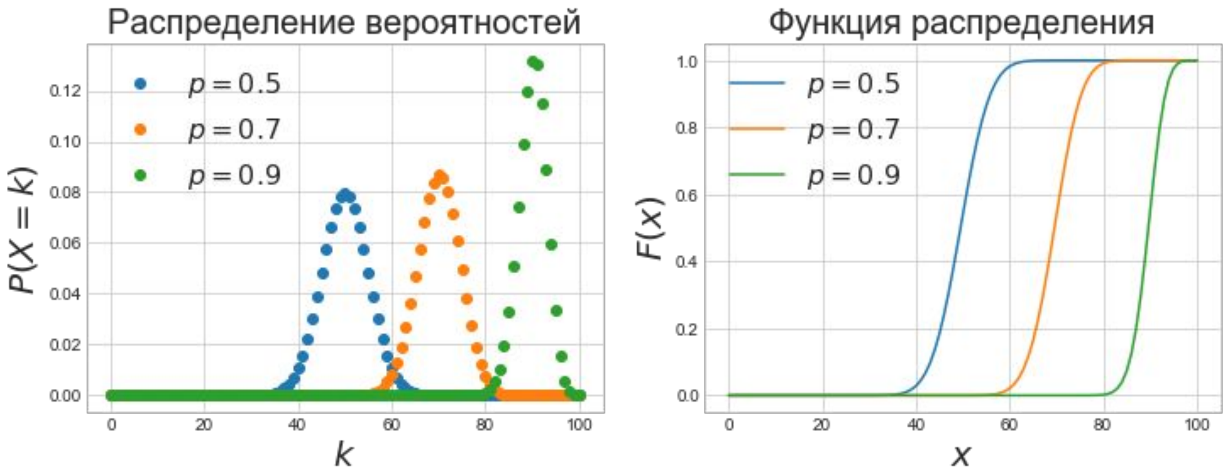
$$X = Y_1 + \cdots + Y_n$$

$$X \sim Bin(p, n)$$

$$\mathbb{E}(X) = n \cdot p$$

$$Var(X) = n \cdot p \cdot (1 - p)$$

$$\mathbb{P}(X = k) = C_n^k \cdot p^k (1 - p)^{n-k}$$



Биномиальное распределение

$$P(X=0)=C_4^0p^0q^4=\frac{4!}{0!\cdot 4!}(0.1)^0(0.9)^4=(0.9)^4=0.6561$$

$$P(X=1)=C_4^1pq^3=\frac{4!}{1!\cdot 3!}(0.1)(0.9)^3=4\cdot 0.1\cdot 0.729=0.2916$$

$$P(X=2)=C_4^2p^2q^2=\frac{4!}{2!\cdot 2!}(0.1)^2(0.9)^2=6\cdot 0.01\cdot 0.81=0.0486$$

$$P(X=3)=C_4^3p^3q=\frac{4!}{3!\cdot 1!}(0.1)^3(0.9)=4\cdot 0.001\cdot 0.9=0.0036$$

$$P(X=4)=C_4^4p^4q^0=\frac{4!}{0!\cdot 4!}(0.1)^4(0.9)^0=0.0001$$

Геометрическое распределение

- Номер броска, когда произошло первое попадание в корзину

Геометрическая случайная величина: $X \sim Geom(p)$

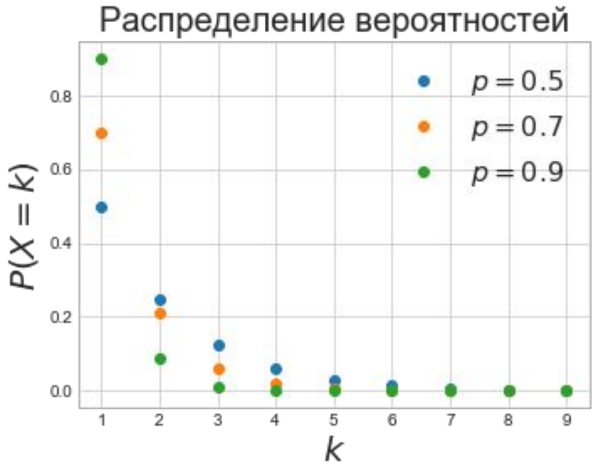
p – вероятность успеха

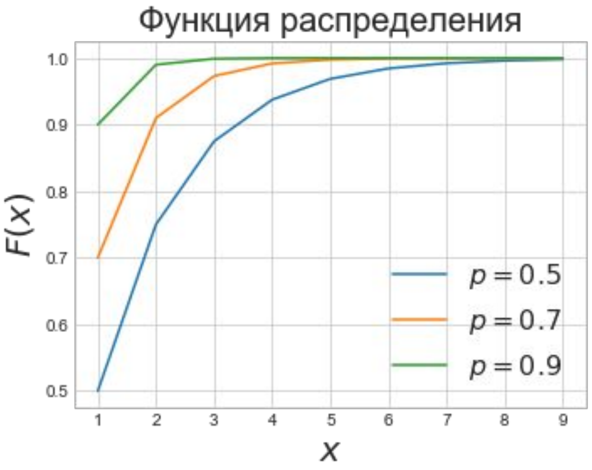
k принимает значения $1, 2, 3, \ldots$

$$\mathbb{P}(X = k) = p \cdot (1-p)^{k-1}$$

$$\mathbb{E}(X) = \frac{1}{p}$$

$$Var(X) = \frac{1-p}{p^2}$$

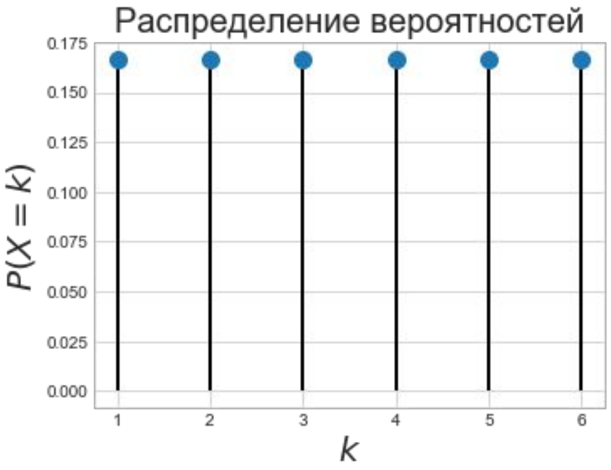


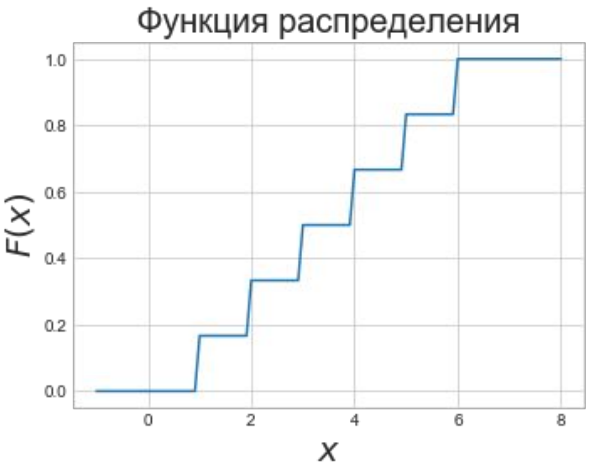


Произвольное дискретное распределение

- Подбрасывание игральной кости

X	1	2	3	4	5	6
P(X=k)	1/6	1/6	1/6	1/6	1/6	1/6





Распределение Пуассона

Распределение Пуассона помогает предсказывать **частоту редких событий**, таких как количество звонков или дефектов, когда мы знаем среднее количество этих событий на фиксированный интервал времени или пространства.
Если есть среднее число событий λ, распределение Пуассона отвечает на вопрос: "Какова вероятность того, что произойдёт ровно k таких событий?".
Пример задачи: В среднем на каждые 100 метров производственной линии обнаруживается 2 дефекта. Какова вероятность того, что на случайных 100 метрах будет обнаружен 1 дефект?

Распределение Пуассона

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

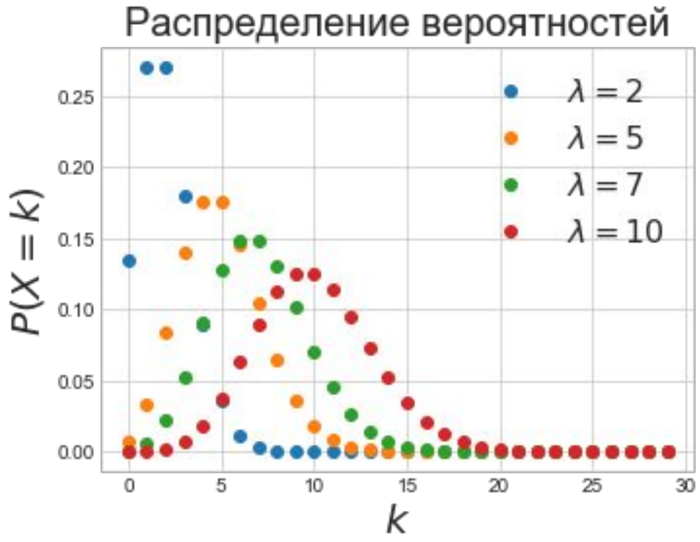
$$X \sim Poiss(\lambda)$$

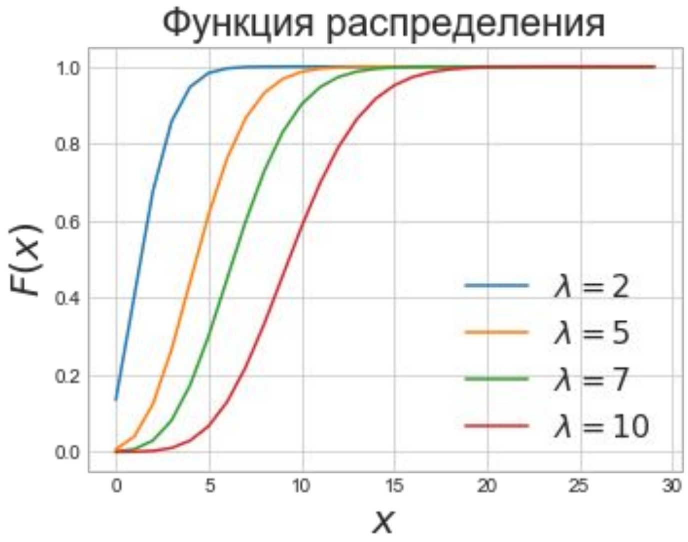
k - количество событий, которое произойдет

$\lambda = np$

$$Var(X) = \lambda$$

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





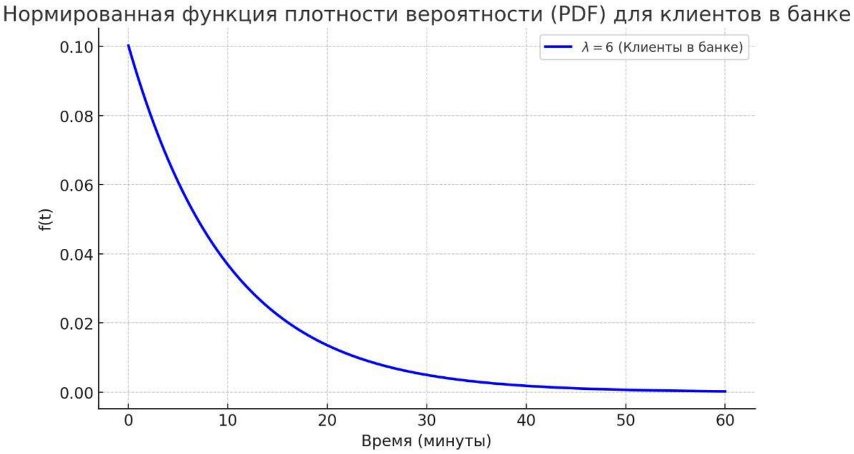
Экспоненциальное распределение

- Время до прихода нового человека
- Время до того как сядет телефон
- Время до того, как позвонит клиент

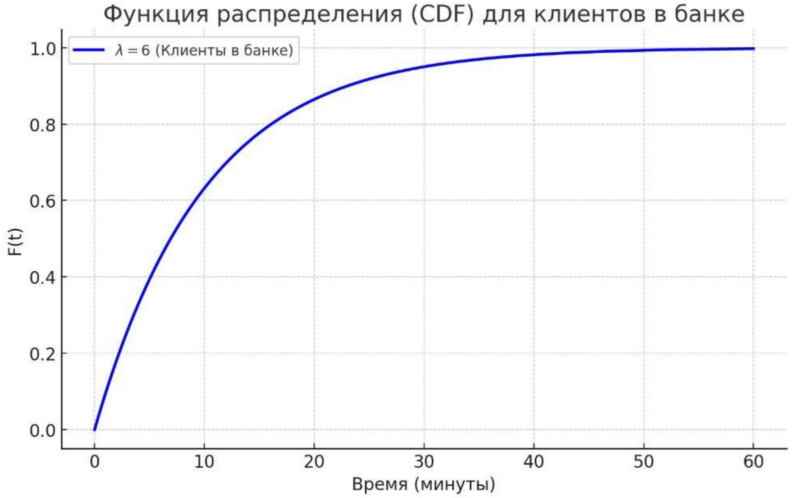
Экспоненциальное распределение

График показывает вероятность того, что клиент придет в точности через t минут. Максимальная вероятность в первые несколько минут, затем она резко падает. Например, вероятность того, что клиент придёт ровно через 5–10 минут, выше, чем вероятность ожидания 20–30 минут.

$$f_X(x) = \lambda \cdot e^{-\lambda \cdot x}, x \geq 0$$



$$F_X(x) = 1 - e^{-\lambda \cdot x}, x \geq 0$$



Равномерное распределение

Предположим, что некий автобус ходит с интервалом в 10 минут, и вы в случайный момент времени подошли к остановке. Какова вероятность того, что автобус подойдёт в течение 1 минуты? Очевидно, 1/10. А вероятность того, что придётся ждать 5 минут? Тоже . А вероятность того, что автобус придётся ждать 9 минут? Одна десятая!

Равномерное распределение

- Время ожидания автобуса

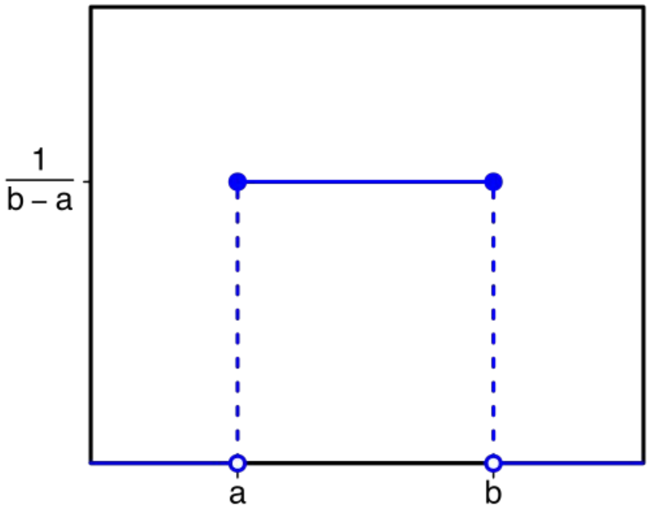
Равномерная случайная величина: $X \sim U[a;b]$

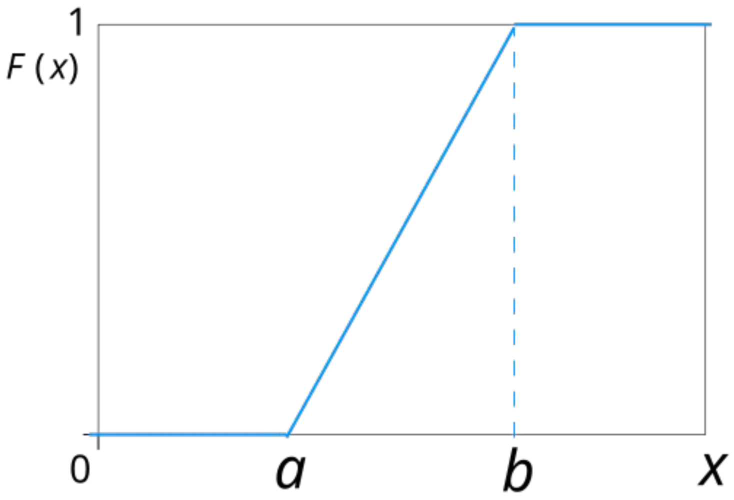
$$\mathbb{E}(X) = \frac{a+b}{2}$$

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

$$f_X(x) = \frac{1}{b-a}, x \in [a;b]$$

$$F_X(x) = \frac{x-a}{b-a}, x \in [a;b]$$

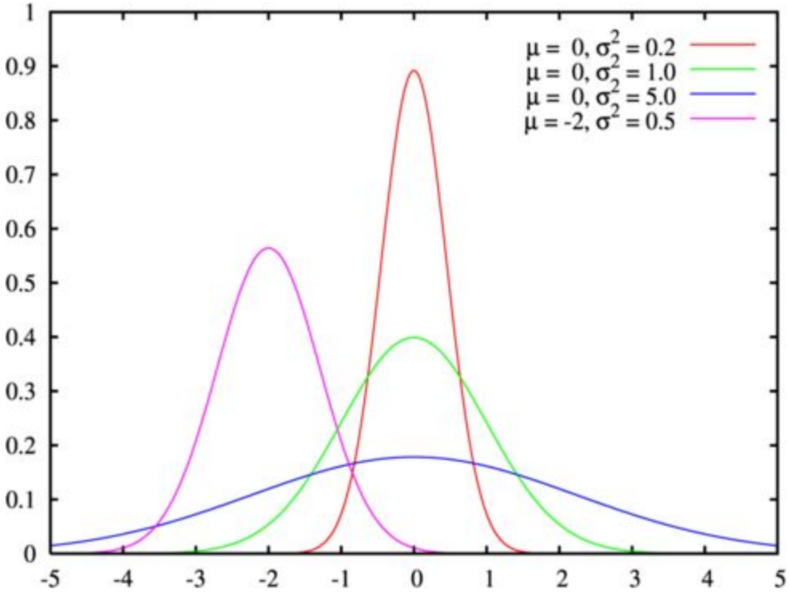


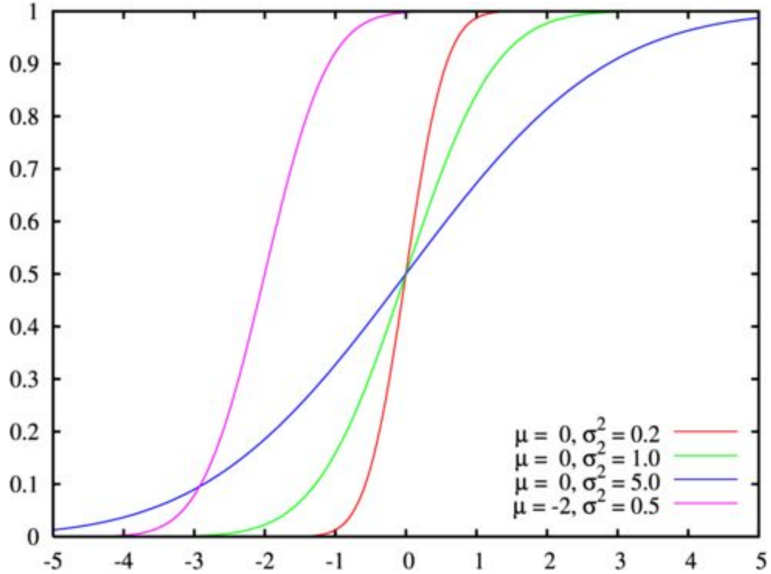


Нормальное распределение

- Погрешности измерений

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



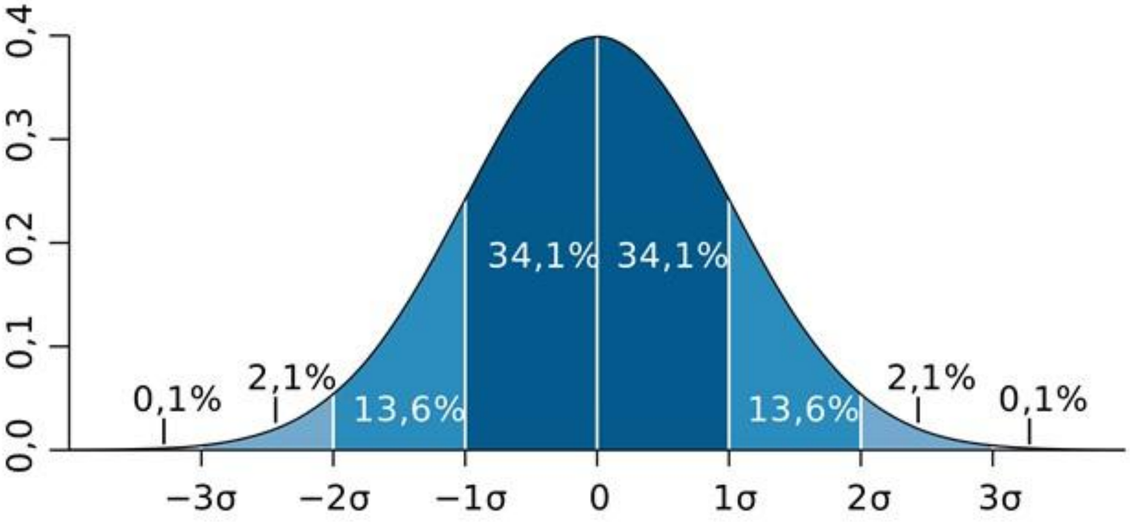


$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-a)^2}{2\sigma^2}}$$

$$F(x) = \int_{-\infty}^{x} f(x)\ dx$$

Нормальное распределение

Правило трёх сигм (3 σ) гласит: с крайне высокой вероятностью случайная величина не отклонится от своего среднего значения более, чем на 3 σ . Практически все значения нормально распределенной случайной величины лежат в интервале **(μ − 3 σ ; μ + 3 σ)** , где μ = E (ξ) — математическое ожидание случайной величины.



Случайная величина	**Распределение**
Пол ребенка	Bern(p)
Попадания в корзину	Binom(n,p)
Число бросков до первого попадания	Geom(p)
Число людей в очереди	Poiss(λ)
Подбрасывание кости	Дискретное
Время между событиями	Exp(λ)
Время до поломки часов	Exp(λ)
Время рождения ребенка	U[0;24]
Погрешность весов	N(0,σ^2)

Точечные оценки

ИТМО

Точечная оценка — это единственное числовое значение, полученное из выборочных данных, которое используется для приближения неизвестного параметра генеральной совокупности.

Примеры точечных оценок: выборочное среднее, выборочная дисперсия, относительная частота успехов, медиана выборки

Несмещённость

Оценка называется несмещённой, если её математическое ожидание равно оцениваемому параметру:

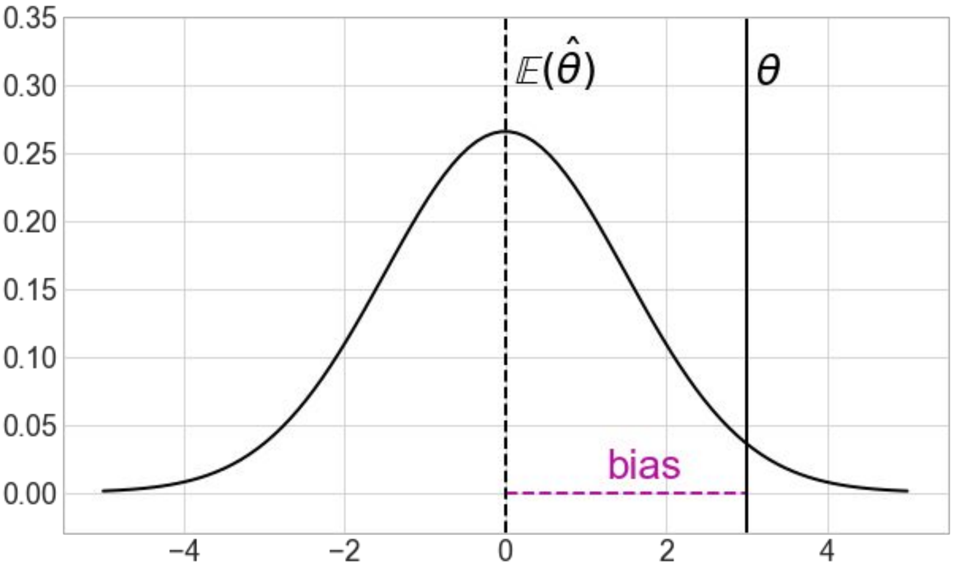
$$\mathbb{E}(\hat{\theta}) = \theta$$

Смещение оценки это разница между её математическим ожиданием и её реальным значением:

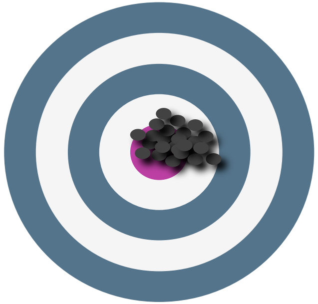
$$bias(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$$

Интуитивно: если при фиксированном *n* мы постоянно используем нашу оценку, в среднем мы не ошибаемся

Несмещённость



Оценка 1



Оценка 2

Встречались ли мы уже со смещением

Выборочная дисперсия и отклонение

ИТМО

Выборочная дисперсия

$$\hat{\sigma}^2 = \frac{(X_1 - \bar{x})^2 + \cdots (X_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{x})^2$$

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

Стандартное отклонение

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2}$$

лет $=\sqrt{\text{лет в квадрате}}$

Несмещенная выборочная дисперсия

$$s^2 = \frac{(X_1 - \bar{x})^2 + \cdots (X_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})^2$$

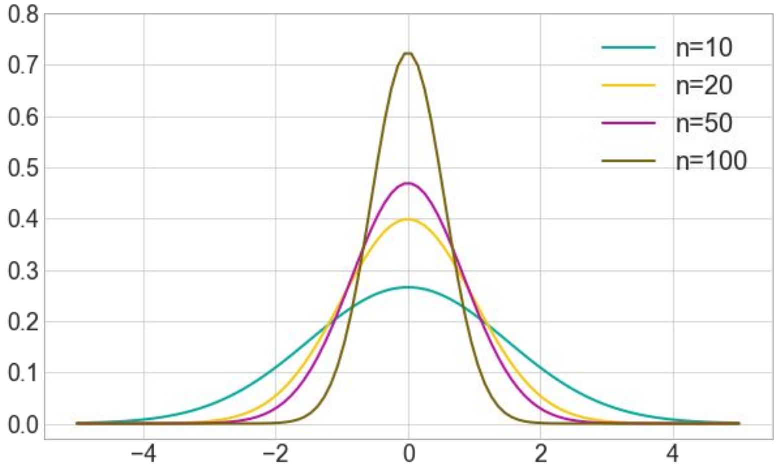
Состоятельность

Оценка называется состоятельной, если она сходится по вероятности к истинному значению параметра при $n \to \infty$

$$\hat{\theta} \xrightarrow{p} \theta$$

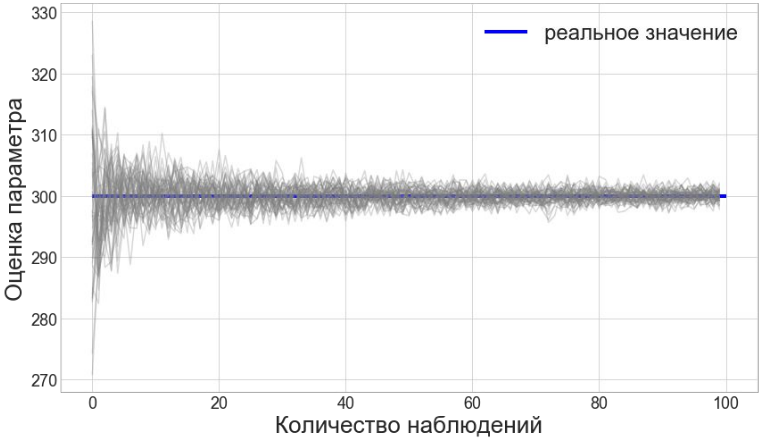
Интуитивно: чем больше наблюдений, тем мы ближе к истине

Состоятельность

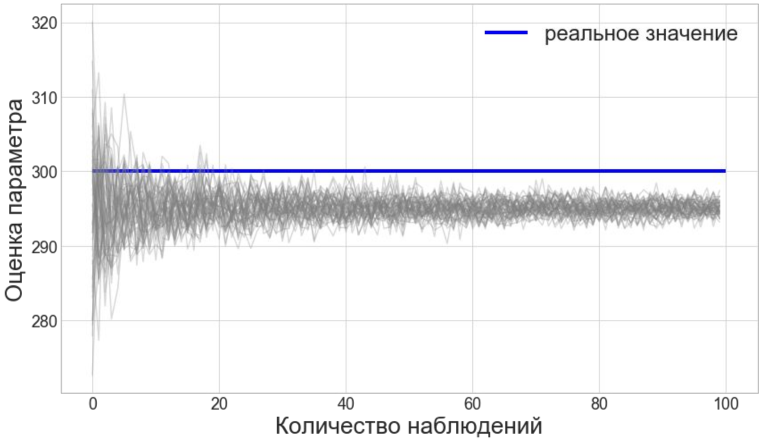


n = 10 n = 20 n = 50 n = 100

Состоятельность



Состоятельная оценка



Несостоятельная оценка

Сравнение оценок

Несмещённых и состоятельных оценок может оказаться несколько $\Rightarrow$ нужно научиться их сравнивать

Обычно оценки между собой сравнивают с помощью квадратичной ошибки:

$$MSE = \mathbb{E}(\hat{\theta} - \theta)^2$$

Для несмещённых оценок MSE совпадает с дисперсией оценки

Интуитивно: чем более предсказуема оценка, тем точнее прогноз (уже доверительный интервал)

Какие оценки стоит получить

В идеале хочется получить:

- несмещенную оценку – в среднем не ошибаться при фиксированном размере выборки
- состоятельную оценку – при большом числе наблюдений быть близко к реальности
- оценку с маленькой средней квадратичной ошибкой

Эффективность

ИТМО

Между смещением и разбросом можно искать компромисс, это позволяет уменьшить среднеквадратичную ошибку

$$MSE = \mathbb{E}\left(\hat{\theta} - \theta\right)^2 = Var\left(\hat{\theta}\right) + \mathrm{bias}^2(\hat{\theta})$$

В классе всех возможных оценок наилучшей в смысле среднеквадратического подхода не существует

Можно попробовать зафиксировать смещение и найти оценку с наименьшей дисперсией

Эффективность

Можно попробовать зафиксировать смещение и найти оценку с наименьшей дисперсией

Такая оценка называется эффективной в классе со смещением $\mathrm{bias}(\hat{\theta})$

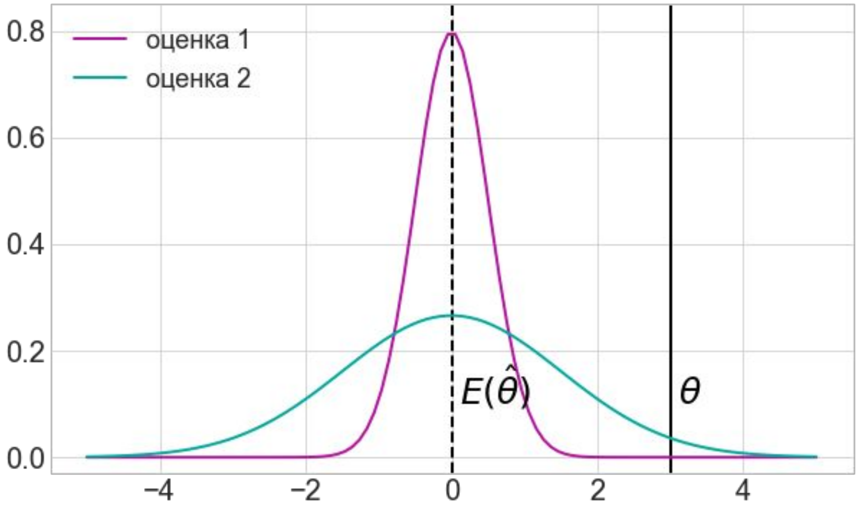
Нас будут интересовать несмещённые эффективные оценки

Интуитивно: эффективная оценка обладает самым узким доверительным интервалом в своём классе

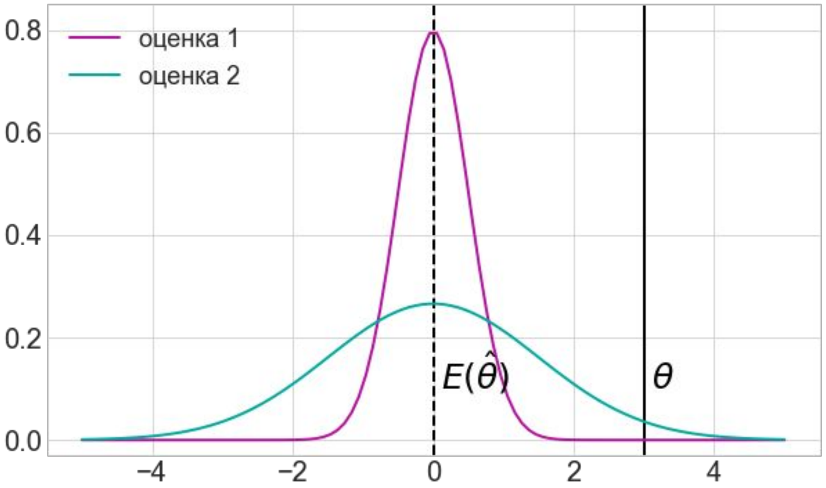
Эффективность

У оценок одинаковое смещение (класс), но при этом
у оценки 1 дисперсия меньше

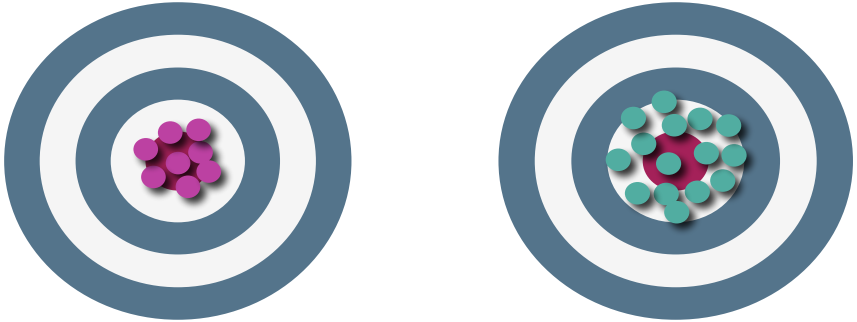
Если у оценки 1 самая маленькая дисперсия из всех существующих ⇒ она для нас самая предпочтительная



Эффективность



Оценка 1



Оценка 2

Интервальные оценки

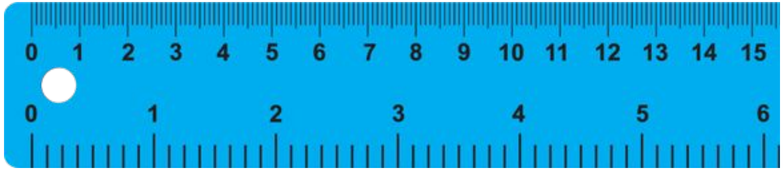
- Точечная оценка делается по случайной выборке ⇒ неопределённость
- Нужно делать выводы в каком-то диапазоне
- Доверительный интервал показывают, насколько мы уверены в точечной оценке

На практике пытаются построить наиболее короткий доверительный интервал

Интервальные оценки

ИТМО

Представьте, что вы биолог, исследующий популяцию лягушек в определенном водоеме. Ваша цель — определить **средний размер лягушек** в этой популяции. Однако у вас есть только **обычная школьная линейка**, которая имеет ограничения по точности измерений.



Интервальные оценки

Выборка: Вы поймали **50 лягушек** из водоема.
Процесс измерения:

- Каждую лягушку вы измеряете по длине тела от головы до хвоста.
- Из-за активности лягушек (лягушк может надуться) и ограничений линейки возможны **погрешности измерений**

Допустим, сумма всех измерений составляет 260 см.

Объем выборки n = 50

тогда среднее получится примерно 5.2

Интервальные оценки

Среднее значение 5.2 см дает нам представление о среднем размере лягушек в выборке.
Однако:

- Мы не знаем, насколько это среднее близко к **истинному среднему размеру** всей популяции лягушек в водоеме.
- **Погрешности измерений** и **вариация размеров лягушек** влияют на точность нашего среднего.

Учитывая погрешность линейки 0.1 см и погрешность измерения двигающихся лягушек 0.2 см получим доверительный интервал **5.2 +- 0.3 см**

Интервальные оценки

Доверительный интервал (ДИ) — это диапазон значений, в котором, с определенной степенью уверенности, находится истинное значение параметра (в нашем случае — среднего размера лягушек).
Уровень доверия обычно выбирается в 95% или 99%, что означает, что мы уверены в том, что истинное значение лежит внутри этого интервала с вероятностью 95% или 99%.
Представление точности:

- Среднее значение без интервала не показывает, насколько оно надежно.
- Доверительный интервал дает представление о **точности и надежности** оценки среднего.

Сколько лягушек сегодня было на слайдах?

Домашнее задание

Темы "Введение в МатСтат" и "Виды статистических оценок"

Общие положения:

- Макс кол-во баллов за ДЗ - 25 баллов

- Качество оформления и кода играет роль
Формат - ноутбук в Collab

В течение 2 недель присылаем ссылки менторам на ваши collab в зависимости от того, кто у вас ментор по табличке распределения. **(Дедлайн 21.10 до лекции)**

Завтра примерно в 10.00 в чат прилетит ссылка на коллаб копируем себе (важно) и в ячейках выполняем домашнее задание.

Всего **2 попытки** сдать

Еще вопросы?