Прикладная статистика

Слайды к лекции 2

24 января 2023

Сергей Александрович Спирин

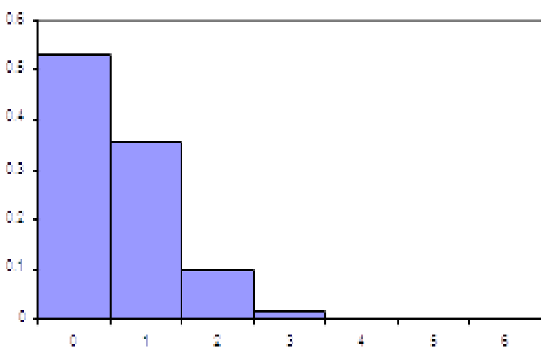
sspirin@hse.ru

Биномиальное распределение

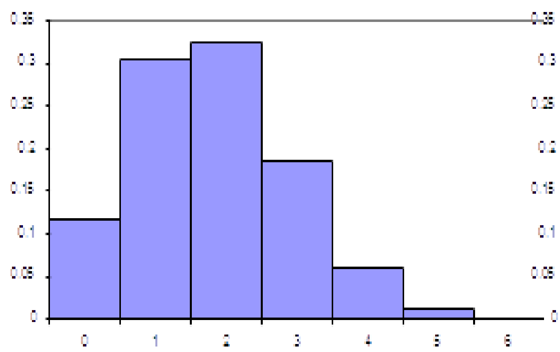
Биномиально распределённая величина = число успехов в *n* независимых испытаниях; параметр распределения *p* = вероятность успеха в одном испытании

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

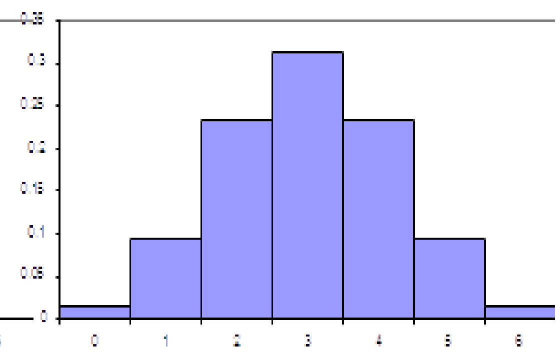
Вероятности 0, 1, …, 6 успехов при шести независимых испытаниях



p=0,1



p=0,3



p=0,5

Биномиальное распределение

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

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

E(K) = np

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.

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

Если ξ — случайная величина, распределённая по Пуассону с параметром λ, то E(ξ) = λ и Var(ξ) = λ
(для распределения Пуассона мат. ожидание равно дисперсии)

Пуассоновское приближение к биномиальному распределению

$\xi = \text{Binom}(n, p)$

n = число испытаний

p = вероятность одного успеха

$n \to \infty$

$p \to 0$

$np = \lambda = const$

$\xi = \text{Poisson}(\lambda)$

$\lambda = np$

Задача

Носителями редкого варианта некоторого гена является 0,001 популяции. В выборке из 3000 человек у $k = 7$ обнаружился редкий вариант. Насколько вероятно, что такое превышение наблюдаемого значения над ожидаемым вызвано случайными причинами?

Задача

Носителями редкого варианта некоторого гена является 0,001 популяции. В выборке из 3000 человек у k = 7 обнаружился редкий вариант. Насколько вероятно, что такое превышение наблюдаемого значения над ожидаемым вызвано случайными причинами?

Решение. Для такого большого размера выборки и такого маленького p количество носителей в случайной выборке распределено по Пуассону со средним pn = 3.
Тем самым нужно посчитать вероятность того, что распределённая по Пуассону со средним 3 случайная величина примет значение ≥ 7. Эта вероятность равна:

$$P(k \geq 7) = 1 - P(k < 7) =$$
$$= 1 - \big(P(k = 0) + P(k = 1) + P(k = 2) + P(k = 3) + P(k = 4) + P(k = 5) + P(k = 6)\big) =$$
$$= 1 - e^{-3} \cdot (1 + 3 + 3^2/2 + 3^3/6 + 3^4/24 + 3^5/120 + 3^6/720) =$$
$$= 1 - 0{,}0498 \cdot (1 + 3 + 4{,}5 + 4{,}5 + 3{,}375 + 2{,}025 + 1{,}0125) = 1 - 0{,}0498 \cdot 19{,}4125$$
$$\approx \mathbf{0{,}033}$$

Распределения случайных величин

- Дискретные
- Непрерывные

Непрерывно распределённая случайная величина может принимать любое числовое значение. При этом каждое конкретное значение принимается в вероятностью 0.

Непрерывные распределения задаются вероятностями попадания случайной величины в то или иное числовое множество (например, интервал).

Строгое формальное определение непрерывной случайной величины довольно сложно. На практике можно считать, что непрерывная с.в. — это функция на очень большом пространстве элементарных событий (настолько большом, что вероятность каждого элементарного события практически 0)

Функция распределения и плотность распределения

Для случайной величины ξ её функция распределения определяется так:

$F_\xi(x) = P(\xi \leq x)$.

Для дискретной случайной величины $F_\xi(x)$ — ступенчатая функция (кусочно-постоянная).

Для непрерывной случайной величины $F_\xi(x)$ — гладкая функция (т.е., у неё есть производная). Эта производная называется **плотностью вероятности**.

Плотность вероятности $p(x)$ обладает свойством: её интеграл по любому отрезку равен вероятности попадания с.в. в этот отрезок.

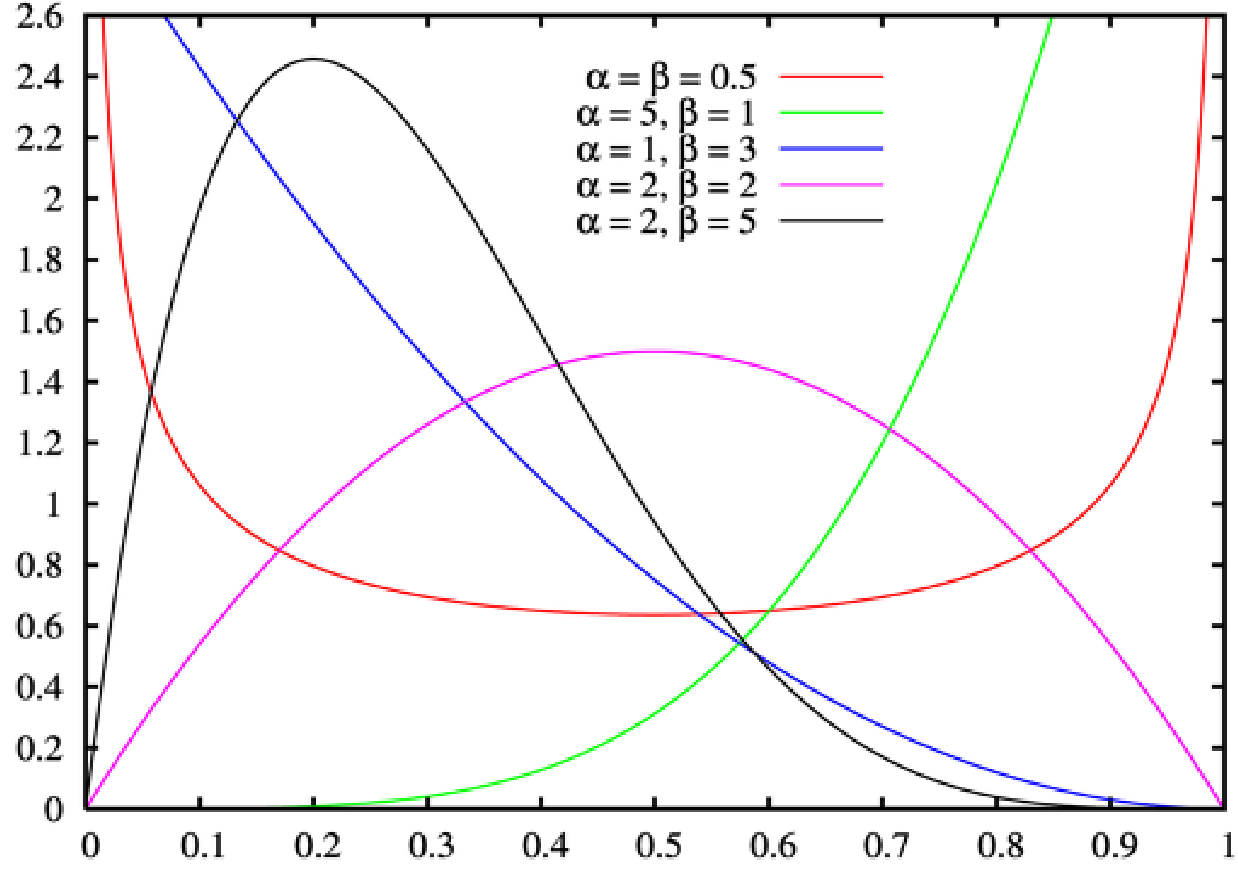
Плотность вероятности в точке x можно определить как вероятность попасть в маленькую окрестность x, делённую на длину этой окрестности.

Бета-распределение

Плотность

$$f_X(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

B(α,β) подбирается так, чтобы интеграл по отрезку [0,1] был равен 1

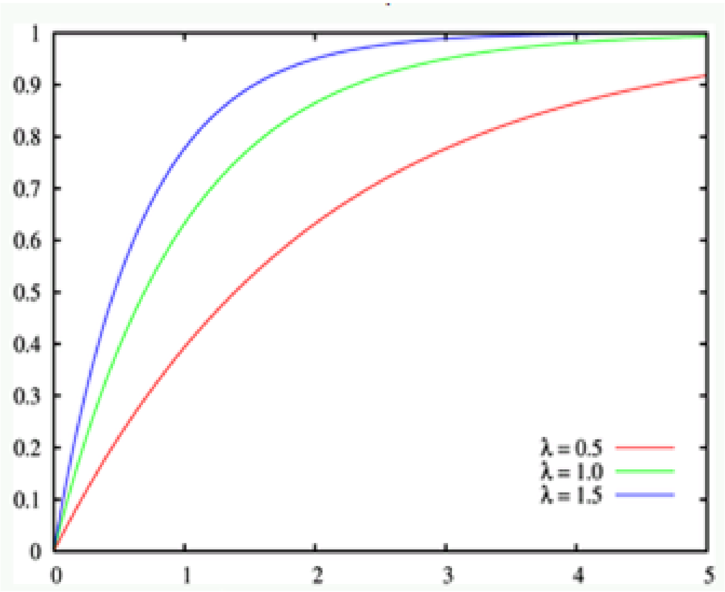


Beta distribution (плотность вероятности)

Равномерное распределение на отрезке [0, 1] – частный случай бета-распределения (при α=β=1)

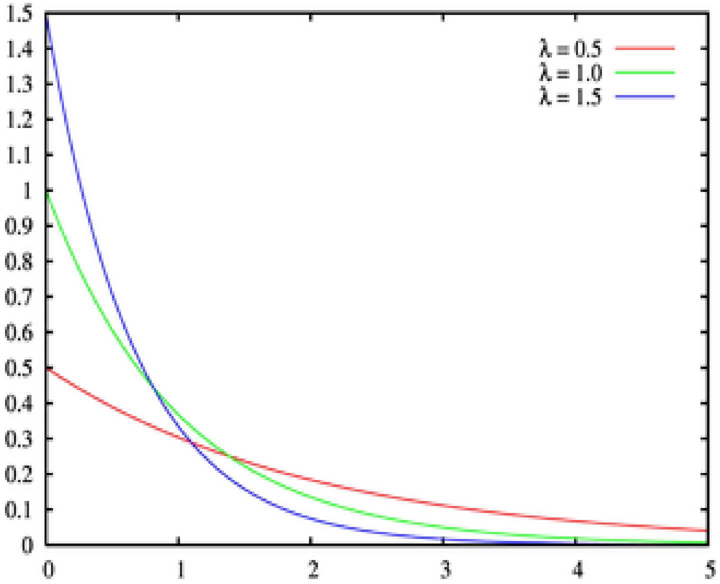
Экспоненциальное распределение

Функция распределения



$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & , x \geq 0 \\ 0 & , x < 0 \end{cases}$$

Плотность вероятности



$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

Экспоненциальное распределение

Экспоненциальное распределение возникает, когда случайная величина представляет собой время ожидания события, которое может равновероятно произойти в любой момент.
Например, экспоненциально распределено:
- время ожидания распада радиоактивного атома
- время ожидания поклёвки при рыбной ловле (а также время между поклёвками)
- время от выхода на улицу тёплым летним вечером до первого укуса комара (а также время между укусами)
- время жизни нестареющих животных (например, мелких птиц в природе)

Аналогично, если, например, на тропинке кое-где (нечасто) встречается подорожник, то расстояние между соседними растениями тоже распределено примерно экспоненциально.

Дискретный аналог экспоненциального распределения — **геометрическое распределение**: число испытаний до первого успеха, $P(n) = p(1-p)^n$, где p — вероятность успеха. Если p очень мало, а испытания происходят часто, то геометрическое распределение числа испытаний до первого успеха можно на практике заменить экспоненциальным распределением времени до первого успеха.

Математическое ожидание экспоненциально распределённой с параметром λ случайной величины равно $1/\lambda$, а дисперсия — $1/(\lambda^2)$.

Если время между событиями распределено экспоненциально, то число событий за заданный промежуток времени (например, число распадов атомов в секунду) будет распределено по Пуассону.

Задача

Предположим, что в геноме все буквы A, C, G, T равновероятны и не зависят от соседей. Какова вероятность того, что отрезок от заданного места генома до ближайшего сайта рестрикции CATG окажется длиннее 500 п.н.?

Задача

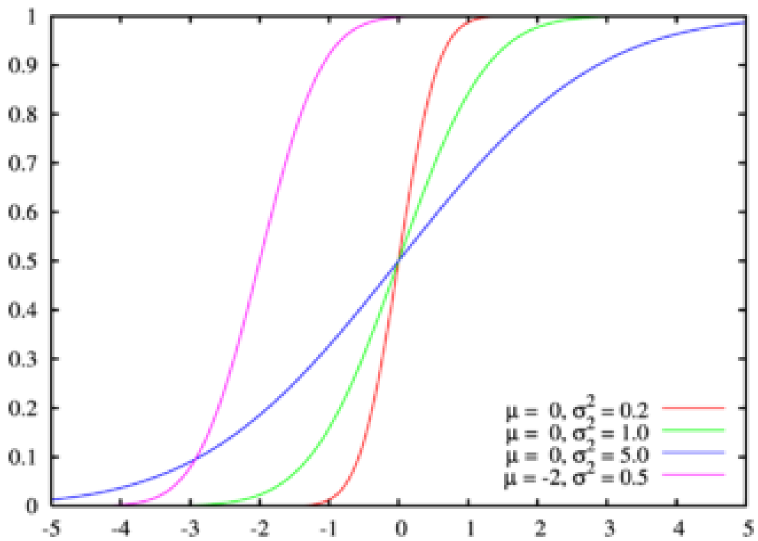
Предположим, что в геноме все буквы A, C, G, T равновероятны и не зависят от соседей. Какова вероятность того, что отрезок от заданного места генома до ближайшего сайта рестрикции CATG окажется длиннее 500 п.н.?

Решение. Пусть вероятности всех букв одинаковы (равны 1/4) и не зависят от соседних букв. Тогда вероятность увидеть в заданном месте заданное слово длины 4 («вероятность успеха») равна $(1/4)^4 = 1/(2^8) = 1/256$. Поэтому среднее расстояние от любого места до ближайшего сайта — 256 п.н. Сами эти расстояния распределены экспоненциально с $\lambda = 1/256$. Поэтому искомая вероятность:

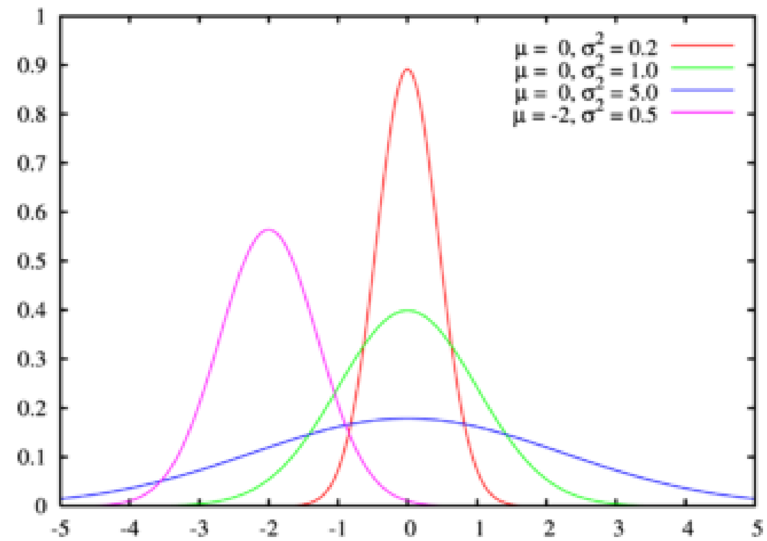
$P(\xi > 500) = 1 - P(\xi \leq 500) = 1 - (1 - e^{-500/256}) = e^{-500/256} \approx e^{-1{,}95} \approx 0{,}142$

Нормальное распределение

Функция распределения



Плотность вероятности



Формулы нет :(

Единственно возможная формула — интеграл от плотности.

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

Смысл параметров: μ — мат. ожидание, σ^2 — дисперсия.

σ часто называют стандартным отклонением

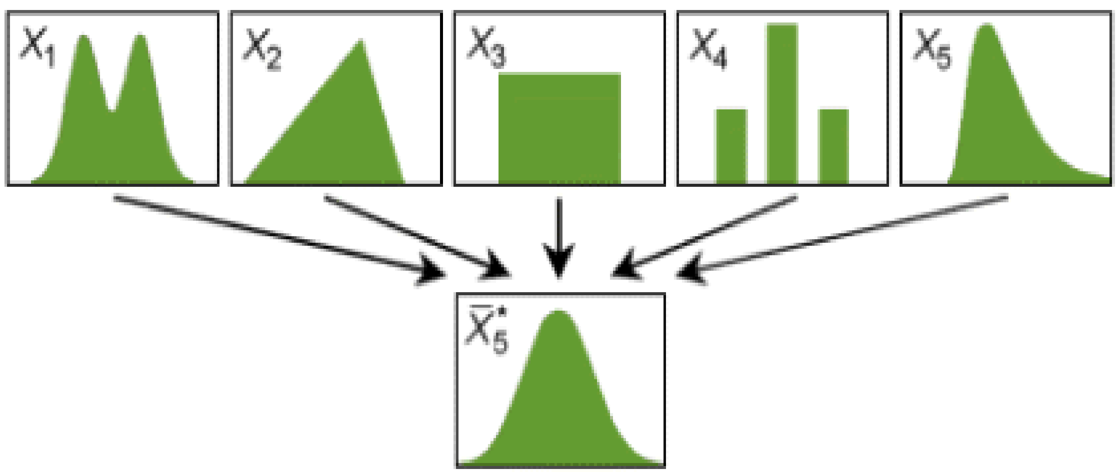
Нормальное распределение

Нормальное распределение возникает везде, где величина представляет собой сумму большого количества элементов, вносящих приблизительно одинаковый вклад (без сильного доминирования небольшого числа из них)
Например:

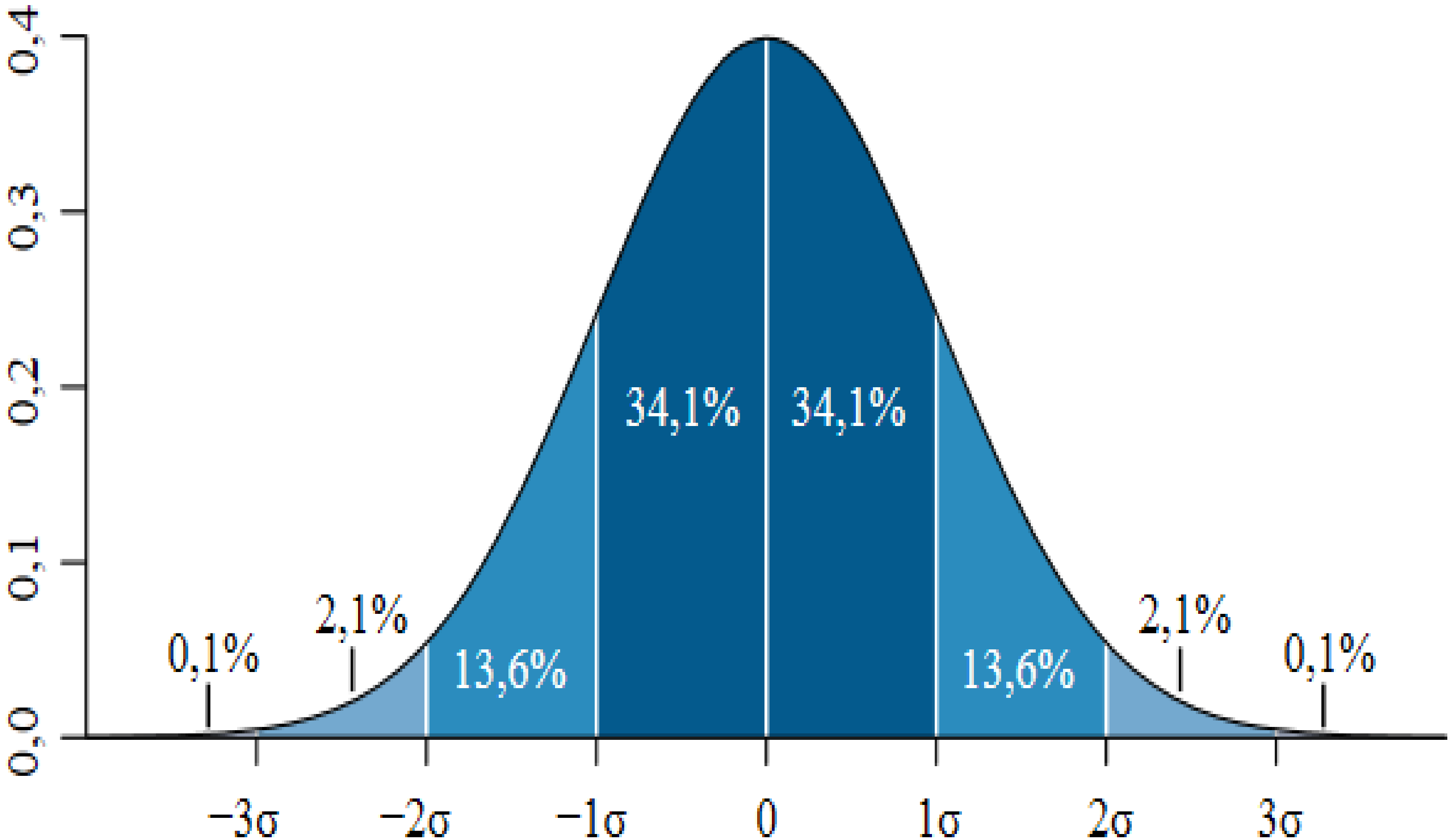
- длина тела животных (одной популяции, одного пола и возраста), как правило, распределена нормально
- ошибки измерений в большинстве экспериментов распределены нормально
- количество крупинок в 1 кг сахарного песка распределено нормально
- число выпадений «орла» при бросании монеты 1 млн. раз распределено нормально
- и т. д.

Центральная предельная теорема

- Сумма достаточно большого числа независимых одинаково распределённых случайных величин распределена приблизительно нормально независимо от исходного распределения
- То же верно и для по-разному распределённых случайных величин при некоторых условиях (малый разброс дисперсий слагаемых)



Нормальное распределение со средним 0



«Правило трёх сигм»: вероятность удалиться от среднего (в заранее заданную сторону) более чем на три стандартных отклонения – около одной тысячной

Таблица стандартного нормального распределения (μ=0, σ=1)

	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
-5	$2{,}9\cdot10^{-7}$	$4{,}8\cdot10^{-7}$	$7{,}9\cdot10^{-7}$	$1{,}3\cdot10^{-6}$	$2{,}1\cdot10^{-6}$	$3{,}4\cdot10^{-6}$	$5{,}4\cdot10^{-6}$	$8{,}5\cdot10^{-6}$	$1{,}3\cdot10^{-5}$	$2{,}1\cdot10^{-5}$
-4	$3{,}2\cdot10^{-5}$	$4{,}8\cdot10^{-5}$	$7{,}2\cdot10^{-5}$	$1{,}1\cdot10^{-4}$	$1{,}6\cdot10^{-4}$	$2{,}3\cdot10^{-4}$	$3{,}4\cdot10^{-4}$	$4{,}8\cdot10^{-4}$	$6{,}9\cdot10^{-4}$	$9{,}7\cdot10^{-4}$
-3	0,0013	0,0019	0,0026	0,0035	0,0047	0,0062	0,0082	0,011	0,014	0,018
-2	0,023	0,029	0,036	0,045	0,055	0,067	0,081	0,097	0,12	0,14
-1	0,16	0,18	0,21	0,24	0,27	0,31	0,34	0,38	0,42	0,46

По строкам — целая часть x, по столбцам — дробная.
В ячейках значения функции распределения $F(x)$ (то есть $P(X < x)$).

Стоит запомнить значения F для $x = -2$ и $x = -3$ и несколько значений x, прежде всего, что
$F(-1{,}65) = 2F(-1{,}96) = 0{,}05$

Другие нормальные распределения

- Z = N(0,1)
  - среднее = 0
  - дисперсия = 1
- X = N(μ, σ)
  - среднее = μ
  - дисперсия = σ^2
- $\boldsymbol{Z = (X - \mu)/\sigma}$

Если $X_1 = N(\mu_1, \sigma_1)$ и $X_2 = N(\mu_2, \sigma_2)$ (X_1 и X_2 независимы),
то $aX_1 + bX_2 = N(a\mu_1 + b\mu_2, (a^2\sigma_1{}^2 + b^2\sigma_2{}^2)^{1/2})$
То есть:
• сумма независимых нормальных с.в. нормальна;
• при сложении мат. ожидания и дисперсии складываются;
• при умножении на число мат. ожидание умножается на то же число, а дисперсия на квадрат этого числа.

Задача

- Завод производит стальные диски средним диаметром 2,5 см со стандартным отклонением в 0,2 мм. Какова вероятность, что случайно выбранный диск окажется шире 2,54 см?

Задача

- Завод производит стальные диски средним диаметром 2,5 см со стандартным отклонением в 0,2 мм. Какова вероятность, что случайно выбранный диск окажется шире 2,54 см?

Решение

Вычисляем Z = (2,54 – 2,5)/0,02 = 2

Смотрим в таблице $F(-2) = 0{,}023$

Нормальное приближение к биномиальному распределению

$X = \mathrm{Binom}(n, p)$

n = число испытаний

p = вероятность одного успеха

$X = N(\mu, \sigma)$

$\mu = np$

$\sigma^2 = np(1 - p)$

$n>40$
$np>5$
$n(1-p)>5$

Задача

Фирмы интересует, кто слушает спонсируемые ими радиопередачи. Некая радиостанция сообщила, что только 20% позвонивших в утреннее ток-шоу – мужчины. На этой неделе в программу позвонило 200 человек. Какова вероятность того, что среди позвонивших хотя бы 50 мужчин?

Задача

Безработица в Давилоне составляет 8,5% . Сделана случайная выборка из 100 работоспособных жителей Давилона. Оцените вероятность того, что выборка содержит по крайней мере 10 безработных.

Задача

Безработица в Давилоне составляет 8,5% . Сделана случайная выборка из 100 работоспособных жителей Давилона. Оцените вероятность того, что выборка содержит по крайней мере 10 безработных.

Попытка решения

Нужно оценить вероятность $P(X \geq 10) = P(X > 9)$

Число безработных распределено биномиально с параметрами

$p = 0{,}085$, $n = 100$

Мат. ожидание: $np = 8{,}5$

Дисперсия: $np(1 - p) = 7{,}78$

Сигма: корень квадратный из дисперсии, то есть 2,79

$Z = (X - 8{,}5)/2{,}79$

Задача

Безработица в Давилоне составляет 8,5% . Сделана случайная выборка из 100 работоспособных жителей Давилона. Оцените вероятность того, что выборка содержит по крайней мере 10 безработных.

Попытка решения

Нужно оценить вероятность $P(X \geq 10) = P(X > 9)$

Число безработных распределено биномиально с параметрами $p = 0{,}085$, $n = 100$

Матожидание: $np = 8{,}5$

Дисперсия: $np(1 - p) = 7{,}78$

Сигма: корень квадратный из дисперсии, то есть 2,79

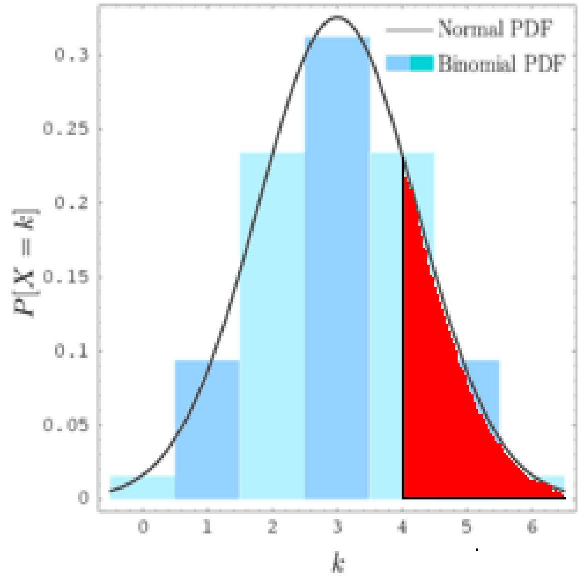
$Z = (X - 8{,}5)/2{,}79$

Проблема: что подставлять в формулу для Z: $X = 9$ или $X = 10$?

Поправка на непрерывность

$$P(X \leq x) = P(X < x + 1)$$

$$P(X \leq x) = P(X < x + 1) \approx P(Y \leq x + 1/2)$$



Здесь *X* распределено биномиально, *Y* — нормально.

Нормальное приближение – это только приближение!

Задача

Безработица в Давилоне составляет 8,5% . Сделана случайная выборка из 100 работоспособных жителей Давилона. Оцените вероятность того, что выборка содержит по крайней мере 10 безработных.

Решение

В формулу $Z = (X - 8{,}5)/2{,}79$ подставляем $X = 9{,}5$

Получаем $Z = 0{,}36$

$\mathrm{P}\ (Z > 0{,}36) = \mathrm{P}\ (Z < -0{,}36) \approx 0{,}36$