Статистический анализ данных. Спецкурс.

Лекция 1. Базовые понятия теории вероятностей.

Ботанический сад-институт ДВО РАН

Кислов Д.Е.

6 марта 2017 г.

Структура и задачи курса

- формирование общих представлений о методах теории вероятностей и математической статистики;
- знакомство с математическими методами статистики;
- изучение многофункциональных программных сред для решения практических задач;

Важные исторические даты

- Формирование понятия случайного явления;
  - Боэций С. (Рим, ок. 520 г):
    "случайность — не подлинное явление, а результат скрещения независимых друг от друга процессов, каждый из которых имеет вполне определенную причину";
- Предпосылки возникновения теории вероятностей (X-XI век);
- Попытки систематического изложения теории (Д. Кардано (1501-1576), Н. Тарталья (1499–1557));

Важные исторические даты

- Х. Гюйгенс (1629–1695): Первая книга по теории вероятностей – "О расчетах в азартной игре". Введено понятие среднего значения — математического ожидания;
- Зарождение статистики: Джон Граунт (1620–1674); Вильям Петти (1623–1687). "Естественные и политические наблюдения над бюллетенями смертности"(Граунт, 1662), "Политическая арифметика"(Петти, 1676);

Рис.: Х. Гюйгенс

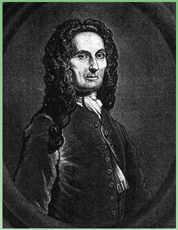


Рис.: А. Муавр

Рис.: Я. Бернулли

Важные исторические даты

- И. Ньютон (1642–1727) — Я. Бернулли (1654–1705); "Искусство предположений"(Бернулли, 1713);
- Абрахам де Муавр (1667–1754): "Учение о случаях"(Муавр, 1733); П.С. Лаплас (1749 – 1827);
- Т. Байес (1702–1761) "Формула Байеса"(Байес, 1763);
- Теория ошибок (конец XVIII) — К.Ф. Гаусс (1777-1855);
- А.Н. Колмогоров (1903–1987) — Аксиоматизация теории вероятностей (Колмогоров, 1933);
- К. Пирсон (1857–1936) — Критерий χ^2; Р.А. Фишер (1890–1962) — метод максимального правдоподобия; Е. Нейман (1894–1977) — статистическая проверка гипотез;

Предпосылки теории вероятностей: комбинаторные задачи

Задача 1

Кодовый замок состоит из 10 кнопок, а открывается при одновременном нажатии 2 кнопок. Охарактеризовать численно его надежность.

Задача 2

Какова вероятность из цифр 1, 3, 5, 7, 9 сложить заданное пятизначное число?

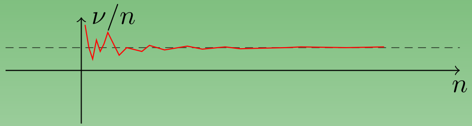
Задача 3

В селе 2500 жителей. Каждый из них примерно 6 раз в месяц (30 дней) ездит в город, выбирая дни поездок по случайным, независящим от других мотивам. Рассчитать минимальную вместительность поезда, обеспечивающую его переполнение не чаще одного раза за 100 дней.

Если ν – число осуществлений некоторого события, то $\frac{\nu}{n}$ – его частота реализации (появления).

ОСНОВНОЕ МОДЕЛЬНОЕ ПОЛОЖЕНИЕ

Частота появления события при многократном повторении эксперимента должна проявлять устойчивость: осуществляя колебания, она должна стремиться к определенному значению.



Что такое вероятность?

Под термином "вероятность" события будем понимать некоторое число, характеризующее частоту его реализации при многократном повторении эксперимента.

Теория вероятностей (ТВ) и математическая статистика (МС)

Задача ТВ

Построение математических моделей случайных явлений, *проявляющих свойство статистической устойчивости*.

Задача МС

Формирование выводов на основе данных опыта и представлений теории вероятностей.

Задача

Симметричную монету подбросили 100 раз, из которых 42 раза выпала «решка» и 58 — «орел». Построены 2 модели этого явления: 1) $P(«орел»)=1/2$, $P(«решка»)=1/2$; 2) $P(«орел»)=2/3$, $P(«решка»)=1/3$. Какую модель следует выбрать?

Таблица: Соответствие вероятностных и теоретико-множественных представлений

Множественное понятие	Понятие теории вероятностей
1. Множество A пусто ($A = \emptyset$).	1. Событие A невозможно.
2. $A \bigcap B = \emptyset$.	2. Два события несовместны.
3. $A_1 \bigcap A_2 \bigcap \ldots \bigcap A_k = X$.	3. Событие X состоит в одновременном наступлении событий $A_1, A_2, \ldots, A_k$.
4. $A_1 \bigcup A_2 \bigcup \ldots \bigcup A_k = X$.	4. Событие X состоит в наступлении одного из событий $A_1, A_2, \ldots, A_k$.
5. Дополнительное множество к A ($\overline{A}$).	5. Событие состоит в ненаступлении события A (в этом случае говорят, что наступило противоположное A событие).
6. $B \subseteq A$.	6. Наступление события B влечет наступление события A.
7. $A = \Omega$.	7. Событие A достоверно.

Аксиомы теории вероятностей (А.Н. Колмогоров)

Вероятностная модель явления построена, если:

- Задано множество элементарных исходов эксперимента (Ω) и возможных событий ($\mathcal{F}$);
- Каждому событию $A \in \mathcal{F}$ сопоставляется действительное число $P(A) \in [0, 1]$, именуемое его вероятностью;
- $P(\Omega) = 1$;
- Для любых двух событий A и B, таких что $A \cap B = \emptyset$, выполнено $P(A \cup B) = P(A) + P(B)$;

Аксиомы теории вероятностей (А.Н. Колмогоров)

Вероятностная модель явления построена, если:

- Задано множество элементарных исходов эксперимента (Ω) и возможных событий ($\mathcal{F}$);
- Каждому событию $A \in \mathcal{F}$ сопоставляется действительное число $P(A) \in [0, 1]$, именуемое его вероятностью;
- $P(\Omega) = 1$;
- Для любых двух событий A и B, таких что $A \cap B = \emptyset$, выполнено $P(A \cup B) = P(A) + P(B)$;

Пример

В эксперименте с подбрасыванием монеты: $\Omega =${«Орел», «Решка»}; В качестве $\mathcal{F}$ можно выбрать {Ω, «Орел», «Решка», $\emptyset$}, или $\mathcal{F}$={Ω, $\emptyset$}, и положить: P(«Орел»)=p, P(«Решка»)=$1 - p$, $p < 1$.

Схема Бернулли

Задача 4

Эксперимент состоит в n–кратном повторении опыта с двумя исходами. Вероятность «успеха» в опыте равна p, вероятность «неудачи» — q $(p+q=1)$. Определить вероятность k успехов при выполнении эксперимента.

Решение

Рассмотрим событие, состоящее в том, что первые k испытаний окончились «успехом», а остальные $n-k$ — «неудачей». Вероятность такого события — $p^k \cdot q^{n-k}$. Общее число подобных событий в эксперименте, отличающихся порядком «успехов» и «неудач» равно количеству k–элементных подмножеств n–элементного множества, т. е. C_n^k. Следовательно, искомая вероятность определяется выражением: $C_n^k p^k q^{n-k}$. $C_n^k = n!/(k!(n-k)!)$

Задача 5

При посадке тиса приживаемость составляет 10%. Какова вероятность, что из 10 посаженных образцов приживется хотя бы один.

Типовые задачи

Задача 6

При посадке тиса приживаемость составляет 10%. Какова вероятность, что из 10 посаженных образцов приживется хотя бы один.

Решение

- $(1/10)^{10}$ – не приживется ни один, $1-(1/10)^{10}$ – приживется хотябы один.

Типовые задачи

Задача 7

При посадке тиса приживаемость составляет 10%. Какова вероятность, что из 10 посаженных образцов приживется хотя бы один.

Решение

- $(1/10)^{10}$ – не приживется ни один, $1 - (1/10)^{10}$ – приживется хотябы один.
- $\sum_{k=1}^{10} C_{10}^{k} (1/10)^{k} (9/10)^{n-k}$

Типовые задачи

Задача 8

При посадке тиса приживаемость составляет 10%. Какова вероятность, что из 10 посаженных образцов приживется хотя бы один.

Решение

- $(1/10)^{10}$ – не приживется ни один, $1-(1/10)^{10}$ – приживется хотябы один.
- $\sum\limits_{k=1}^{10} C_{10}^{k}(1/10)^{k}(9/10)^{n-k}$

Задача 9

При высаживании непикированной рассады помидоров только 80% растений приживаются. Найдите вероятность того, что из десяти посаженных кустов приживется не менее 7.

Задача 10

Приживаемость саженцев составляет в среднем 30%. Каков должен быть минимальный объём посадок, чтобы можно было гарантировать выживаемость 50 экземпляров с доверительной вероятностью не меньшей 90%?

Схема Бернулли. Планирование эксперимента.

Задача 11

Приживаемость саженцев составляет в среднем 30%. Каков должен быть минимальный объём посадок, чтобы можно было гарантировать выживаемость 50 экземпляров с доверительной вероятностью не меньшей 90%?

Решение

$$\min_N \sum_{k=50}^{N} C_n^k 0.3^k 0.7^{n-k} \geq 0.9$$

Задача 12

Приживаемость саженцев составляет в среднем 30%. Каков должен быть минимальный объём посадок, чтобы можно было гарантировать выживаемость 50 экземпляров с доверительной вероятностью не меньшей 90%?

Решение

$$\min_{N} \sum_{k=50}^{N} C_n^k 0.3^k 0.7^{n-k} \geq 0.9$$

Решить такую задачу можно численно, используя язык программирования.

Предельная теорема в схеме Бернулли

Теорема Муавра–Лапласа

Вероятность того, что число успехов в схеме Бернулли (с вероятностью «успеха» p и вероятностью «неудачи» $q = 1 - p$ находится в интервале $[k_1,\ k_2]$ равна:

$$P(n; k_1, k_2) \approx \Phi(\frac{k_2 - np}{\sqrt{npq}}) - \Phi(\frac{k_1 - np}{\sqrt{npq}})$$

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_0^x e^{-t^2/2} dt$$

Эксперимент $\omega_i \in \Omega$:

$$\omega_1 \to x_1 = \psi(\omega_1)$$

$$\omega_2 \to x_2 = \psi(\omega_2)$$

$$\omega_3 \to x_3 = \psi(\omega_3)$$

$$\vdots$$

$$\omega_n \to x_n = \psi(\omega_n)$$

Определение

Если $P(\omega_i) = p_i$ определены, то преобразование ψ совместно с p_i и Ω определяют дискретную случайную величину ξ: $P(\xi = x_i) = p_i$. Множество $\{(x_i, p_i)\}$ образуют закон распределения случайной величины ξ.

Пример

Количество очков, выпавшее при подбрасывании игральной кости, — дискретная случайная величина.

Функция распределения случайной величины

Определение

Функция $F_\xi(x) = P(\xi < x)$, где $x \in \mathbb{R}$, называется функцией распределения случайной величины ξ.

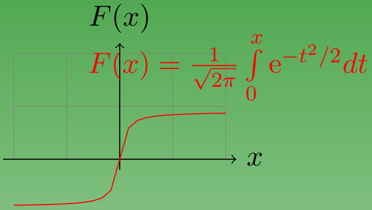
Замечание

Случайная величина называется *распределенной непрерывно*, если соответствующая функция распределения является *непрерывной*.

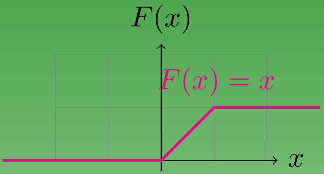
Замечание

Для случайных величин, имеющих дискретное распределение, функция распределения терпит разрывы. Аксиомы теории вероятностей определяют основные свойства функции распределения.

Примеры непрерывных и разрывной функций распределения

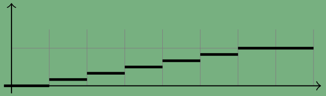


Стандартная нормальная функция распределения.



Функция распределения случайной величины равномерно распределенной на интервале [0,1].

Разрывная функция распределения



Функция распределения случайной величины — числа очков при подбрасывании кубика.

Свойства функций распределений

Утверждение

Если $F_\xi(x)$ — функция распределения случайной величины ξ, то

- $F_\xi(x)$ – неубывающая функция;
- $\lim\limits_{x\to-\infty} F_\xi(x) = F(-\infty) = 0$;
- $\lim\limits_{x\to+\infty} F_\xi(x) = F(+\infty) = 1$;
- $P(a \le \xi < b) = F_\xi(b) - F_\xi(a), a, b \in \mathbb{R}$.

Плотность распределения случайной величины

Определение

Плотностью распределения ($f_\xi(x)$) случайной величины ξ называется производная по x(если таковая существует) от функции распределения $F_\xi(x)$:

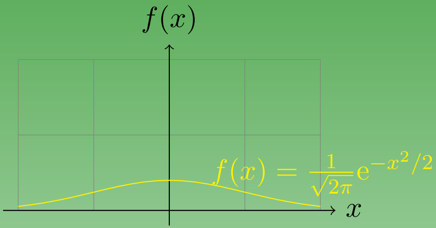
$$f_\xi(x) = \frac{dF_\xi(x)}{dx}.$$

Замечание

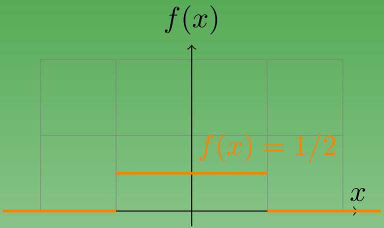
Плотность распределения — скорость изменения вероятности события $\{\xi < x\}$, зависящая от x.

Замечание

Плотность распределения в точке x характеризует вероятность принадлежности случайной величины бесконечно малому интервалу, содержащему x.



Плотность стандартного нормального распределения.



Плотность случайной величины ξ, распределенной равномерно на интервале [-1, 1].

Свойства плотности распределения

Утверждение

Если $f_\xi(x)$ плотность распределения случайной величины ξ, то

- $F_\xi(x) = \int\limits_{-\infty}^{x} f_\xi(\tau)d\tau$;
- $\lim\limits_{x\to\pm\infty} f_\xi(x) = 0$;
- $\int\limits_{-\infty}^{\infty} f_\xi(x) = 1$;
- $P(a \le \xi < b) = \int\limits_{a}^{b} f_\xi(x)dx, a, b \in \mathbb{R}$.

Математическое ожидание и дисперсия

Определение

Дисперсией случайной величины ξ называется число $\mathrm{D}\xi = \mathrm{E}(\xi - \mathrm{E}\xi)^2$, или в случае дискретного распределения — $\mathrm{D}\xi = \sum\limits_i p_i(x_i - \sum\limits_j p_j x_j)^2$; в случае непрерывного распределения с $f_\xi(x)$ — $\mathrm{D}\xi = \int\limits_{-\infty}^{\infty} (x - \mathrm{E}\xi)^2 f_\xi(x)dx$, $\mathrm{E}\xi = \int\limits_{-\infty}^{\infty} x f_\xi(x)dx$.

Свойства математического ожидания

- $\mathrm{E}(\alpha\xi_1 + \beta\xi_2 + \gamma) = \alpha\mathrm{E}\xi_1 + \beta\mathrm{E}\xi_2 + \gamma$;
- $\mathrm{E}\xi_1\xi_2 = \mathrm{E}\xi_1\mathrm{E}\xi_2$, если ξ_1 и ξ_2 независимы;
- $|\mathrm{E}\xi| \leq \mathrm{E}|\xi|$,

где ξ_1, ξ_2, ξ некоторые случайные величины, для которых определено математическое ожидание; α, β, γ — произвольные действительные числа.

Математическое ожидание и дисперсия

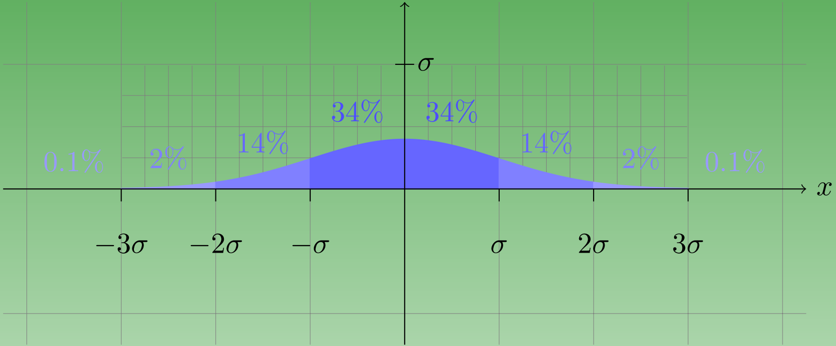
Свойства дисперсии

- $\mathrm{D}(\xi_1+\xi_2)=\mathrm{D}\xi_1+\mathrm{D}\xi_2$, если ξ_1, ξ_2 являются независимыми;
- $\mathrm{D}\alpha=0, \alpha\in\mathbb{R}$; $\mathrm{D}(\alpha\xi)=\alpha^2\mathrm{D}\xi$;

Характеристики некоторых распределений

- $\xi\in\mathbf{U}_{[0,1]}$ — равномерное распределение на интервале $[0,1]$: $\mathrm{E}\xi=1/2$, $\mathrm{D}\xi=1/12$;
- $\xi\Subset f_\xi(x)=\dfrac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2}$ — нормальное распределение: $\mathrm{E}\xi=0$, $\mathrm{D}\xi=1$; ($\xi\in\mathcal{N}(0,1)$, ξ имеет стандартное нормальное распределение);
- $\eta=\sigma\xi+a$, где $a,\sigma\in\mathbb{R}; \xi\in\mathcal{N}(0,1)$, то η имеет нормальное распределение: $\mathrm{E}\xi=a$, $\mathrm{D}\xi=\sigma^2$; $(\xi\in\mathcal{N}(a,\sigma^2))$.

$$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-x^2}{2\sigma^2}\right)$$



Несмещенное нормальное распределение с дисперсией σ^2.

Многомерное нормальное распределение

Определение

Пусть $\Upsilon = (\eta_1, \dots, \eta_n) \in \mathcal{N}(0,1)$ – независимы в совокупности случайные величины; тогда $\Xi = A\Upsilon + b$, где $A = A_{n\times n}$, $\dim b = n$, имеет многомерное нормальное распределение.

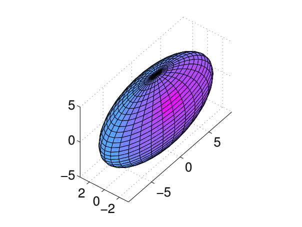
Утверждение

Если матрица A невырождена, то Ξ имеет плотность

$$f_\Xi(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(A^T A)}} \mathrm{e}^{-(x-b)^T (A^T A)^{-1}(x-b)/2}.$$

Геометрическая интерпретация

Пример многомерного нормального распределения с диагональной матрицей A;



Неравенство П.Л. Чебышева

Утверждение

Если для случайной величины ξ определена дисперсия, то

$$P(|\xi - \mathrm{E}\xi| \geq \varepsilon) \leq \frac{\mathrm{D}\xi}{\varepsilon^2},$$

выполнено для любого $\varepsilon > 0$.

Доказательство

Пусть случайная величина ξ непрерывно распределена и, кроме того, $\xi > 0$. Дадим оценку вероятности $P(\xi \geq \varepsilon)$. Рассмотрим выражение для математического ожидания случайной величины ξ.

$$\mathrm{E}\xi = \int\limits_0^\infty x f_\xi(x)dx = \underbrace{\int\limits_0^\varepsilon x f_\xi(x)dx}_{\geq 0} + \int\limits_\varepsilon^\infty x f_\xi(x)dx \geq \int\limits_\varepsilon^\infty x f_\xi(x)dx \geq \varepsilon P(\xi \geq \varepsilon),$$

Доказательство

откуда следует, что

$$P(\xi \geq \varepsilon) \leq \frac{\mathrm{E}\xi}{\varepsilon}. \qquad (*)$$

Положим $\eta = (\xi - \mathrm{E}\xi)^2$, тогда $\mathrm{E}\eta = \mathrm{D}\xi$ и неравенство Чебышева является следствием доказанного неравенства (*).

Задача 13. (применение неравенства Чебышева)

Требуется оценить число необходимых измерений листовой пластинки, чтобы с заданной точностью охарактеризовать ее среднюю длину. Каждое измерение длины представляет собой $l_i = l + \varepsilon_i$, где l – неизвестное среднее, ε_i – случайная величина с нулевым средним и дисперсией σ. Сколько измерений необходимо провести, чтобы выполнялись следующие условия по точности:

$$P(|\hat{l} - l| \geq \varepsilon) \leq \gamma,$$

где $\hat{l} = \frac{1}{N}\sum\limits_i l_i$ – оценка среднего, γ и ε – заданные показатели точности.

Неравенство П.Л. Чебышева

Утверждение

Если для случайной величины ξ определена дисперсия, то

$$P(|\xi - \mathrm{E}\xi| \geq \varepsilon) \leq \frac{\mathrm{D}\xi}{\varepsilon^2},$$

выполнено для любого $\varepsilon>0$.

Решение.

Пусть $\varepsilon = 0.1$, $\gamma = 0.01$, $\sigma^2 = 0.5$. Из неравенства Чебышева, полагая $\gamma = \dfrac{n\sigma^2}{n^2\varepsilon^2}$, получим:

$$l_i = l + \varepsilon_i, \hat{l} = \tfrac{1}{N}\sum_i l_i$$

$$P\left(\left|\tfrac{1}{n}\sum_i \varepsilon_i\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \leq \gamma$$

$$n \geq \frac{\sigma^2}{\gamma\varepsilon^2} = \frac{0.5}{0.01 \cdot 0.01} = 5000$$

Неравенство П.Л. Чебышева

Утверждение

Если для случайной величины ξ определена дисперсия, то

$$P(|\xi - \mathrm{E}\xi| \geq \varepsilon) \leq \frac{\mathrm{D}\xi}{\varepsilon^2},$$

выполнено для любого $\varepsilon>0$.

Решение.

Пусть $\varepsilon = 0.1$, $\gamma = 0.01$, $\sigma^2 = 0.5$. Из неравенства Чебышева, полагая $\gamma = \frac{n\sigma^2}{n^2\varepsilon^2}$, получим:

$$l_i = l + \varepsilon_i, \hat{l} = \tfrac{1}{N}\sum_i l_i$$

$$P\left(\left|\tfrac{1}{n}\sum_i \varepsilon_i\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2} \leq \gamma$$

$$n \geq \frac{\sigma^2}{\gamma\varepsilon^2} = \frac{0.5}{0.01 \cdot 0.01} = 5000$$

$n \geq 5000$ – грубая оценка, ее почти всегда можно улучшить

Центральная предельная теорема (Ц.П.Т.)

Утверждение

Пусть $\{\xi_i\}$ последовательность независимых одинаково распределенных случайных величин, причем $\mathrm{E}\xi_i = a$ и $\mathrm{D}\xi_i = \sigma^2 \in (0, \infty)$. Если $\zeta_n = \dfrac{-an + \sum\limits_{i=1}^{n} \xi_i}{\sqrt{n}\sigma}$, то

$$P(\zeta_n < x) \to \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} \mathrm{e}^{-t^2/2} dt,$$

равномерно по x при $n \to \infty$

Задача 14.

Дан набор измерений параметра a: $a_i = a + \xi_i$, $i = \overline{1, n}$. Погрешности измерений ξ_i являются независимыми одинаково распределенными случайными величинами, имеющими нулевое математическое ожидание и дисперсию σ. В качестве оценки параметра a принимается среднее арифметическое $\{a_i\}$: $\hat{a} = \dfrac{\sum\limits_i a_i}{n}$. Сколько измерений необходимо провести, чтобы достичь заданной точности ε с вероятностью выхода за пределы точности γ.

Решение

Неравенство Чебышева

$$P(\left|\frac{\sum\limits_i \xi_i}{n}\right| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \equiv \gamma,$$

$n \geq \dfrac{\gamma^{-1}\sigma^2}{\varepsilon^2}$.

При $\sigma = 1$, $\varepsilon = 0.01$, $\gamma = 0.01$: $n \geq 10^6$.

Ц.П.Т.

$$P\left(\left|\frac{\sum\limits_i \xi_i}{n}\right| < \varepsilon\right) \approx \frac{2}{\sqrt{2\pi}} \int\limits_0^{\frac{\varepsilon\sqrt{n}}{\sigma}} e^{-t^2/2}dt \equiv$$

$1 - \gamma$.

При тех же условиях: $n \geq 2223$.

Вычислительные среды для анализа данных

- R, `http://r-project.org` + Packages (`https://cran.r-project.org/`)
- Python, `http://python.org`
  - Pandas, `http://pandas.pydata.org`
  - SciPy+NumPy, `http://scipy.org`
  - Matplotlib, `http://matplotlib.org/`
  - Scikit-learn, `http://scikit-learn.org/`
  - ... Packages
- Statistica, `http://statsoft.com`
- MatLab, `http://www.mathworks.org`

Интерактивные среды

RStudio Server

www.rstudio.com

Jupiter Notebook

ipython.org/notebook.html

Информационные ресурсы

R:

- С.Э Мастицкий, В.К. Шитиков Статстический анализ и визуализация данных с помощью R, 2014.
- http://www.statmethods.net/

Python:

- У. Маккинни Python и анализ данных, 2015
- М. Лутц Программирование на Python, 2011
- Т. Сегран Программируем коллективный разум, 2008
- http://python.org