# Статистики

## 1. Оценка распределения по выборке

Рассматривается выборка из случайной величины $X$:

$$X^n = (X_1, \dots, X_n),$$

где $n$ — объем выборки. Величины $X_1, X_2, \dots, X_n$ — независимые одинаково распределенные случайные величины ($i.i.d.$).

**Статистикой** $T(X^n)$ называется любая функция от данной выборки.

Далее будет рассмотрено, какие статистики используются для оценок по выборкам законов распределения случайных величин различных классов. Распределение *дискретной случайной величины* задается функцией вероятности:

$$X \in A = \{a_1, a_2, \dots\}, \quad P\big(X = a_k\big) = p_k.$$

Для выборки из такой случайной величины лучшей оценкой для вероятностей из функции вероятностей являются частоты соответствующих событий на выборке (по закону больших чисел):

$$\bar{p}_k = \frac{1}{n} \sum_{i=1}^{n} \big[X_i = a_k\big].$$

Если *непрерывная случайная величина* задается с помощью функции распределения, то ее можно оценить с помощью **эмпирической функции распределения**:

$$X \sim F(x), \quad F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \big[X_i \leqslant x\big].$$

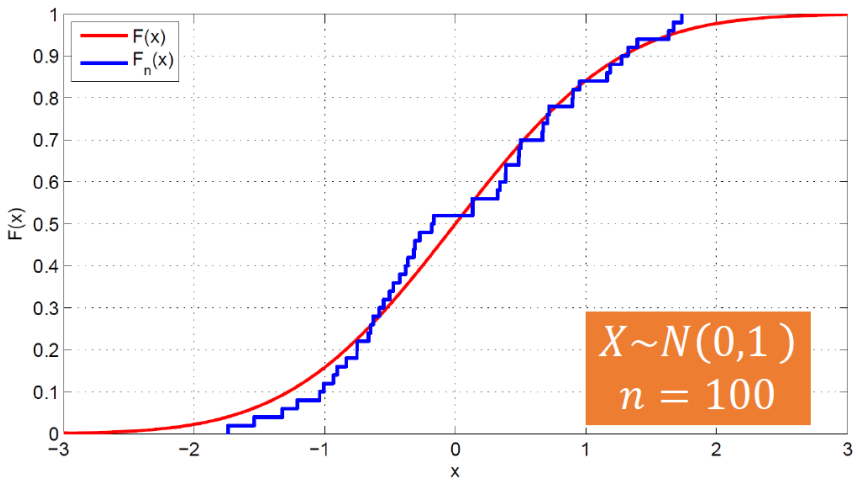


Рис. 1.

На рис. 1 красная линия соответствует теоретической функции стандартного нормального распределения (нормальное распределение со средним, равным нулю, и с дисперсией, равной 1). Синяя линия соответствует эмпирической функции распределения, построенной по выборке объема 100.

Непрерывные случайные величины также могут задаваться с помощью плотностей. Для оценки плотности можно разбить область определения случайной величины на интервалы одинаковой длины. Количество объектов выборки в каждом интервале будет пропорционально среднему значению плотности на нем. Именно так устроена **гистограмма**.

На гистограмме, приведенной на рис. 2, изображена продолжительность жизни крыс на строгой диете (в днях).

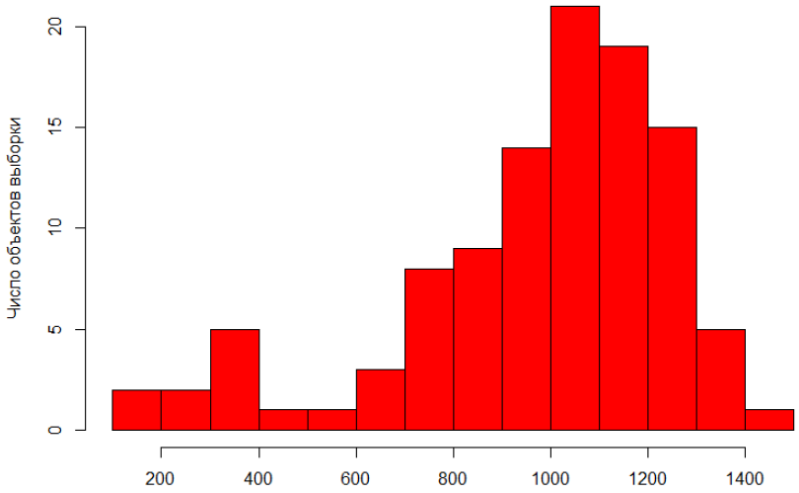


Рис. 2.

По такой гистограмме хорошо видны все особенности распределения данных: оно бимодально, основной пик приходится примерно на 1000 дней, но есть крысы, которые живут существенно меньше.

Важным аспектом работы с гистограммами является правильный выбор числа интервалов.

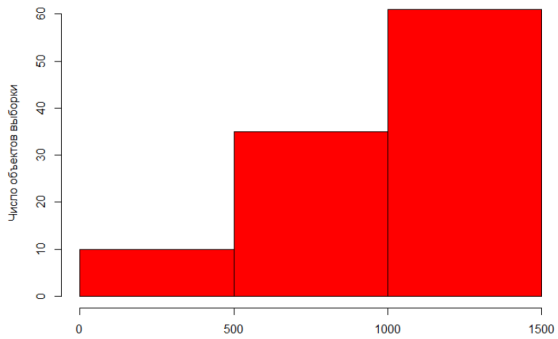


Рис. 3.

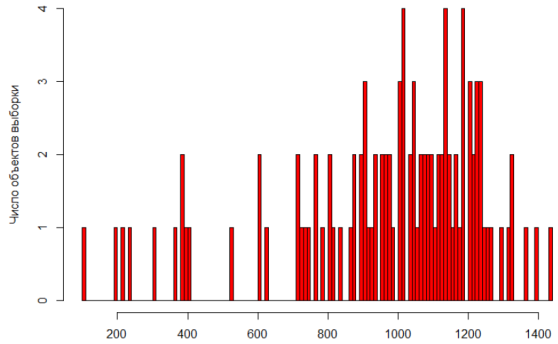


Рис. 4.

Если рассмотреть слишком мало интервалов, то они будут слишком большими, в результате гистограмма получится грубой (см. рис. 3). Аналогично в случае слишком большого количества интервалов — в большую часть из них не попадет ни одного объекта выборки (см. рис. 4). В обоих случаях построенные гистограммы не являются информативными.

Описанного недостатка лишены *гладкие оценки плотности* $f(x)$. Для построения такой оценки необходимо взять окно ширины $h$ и, двигая его по числовой оси, вычис-

лять в нем значение функции, называемой **ядром**. **Ядерная оценка плотности** имеет вид:

$$f_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right).$$

На рис. 5 показана оценка, построенная на тех же данных о продолжительности жизни крыс. Как и в случае гистограммы, на таком графике видны все особенности распределения данных.

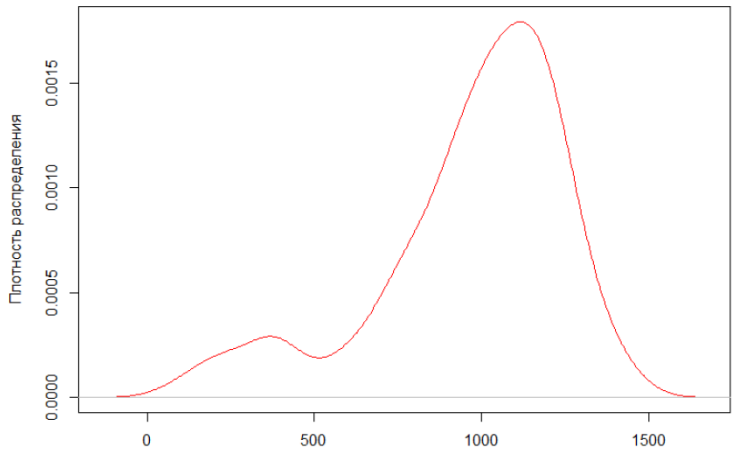


Рис. 5.

На рис. 6 продемонстрированы все виды оценок распределения для выборки из стандартного нормального распределения.

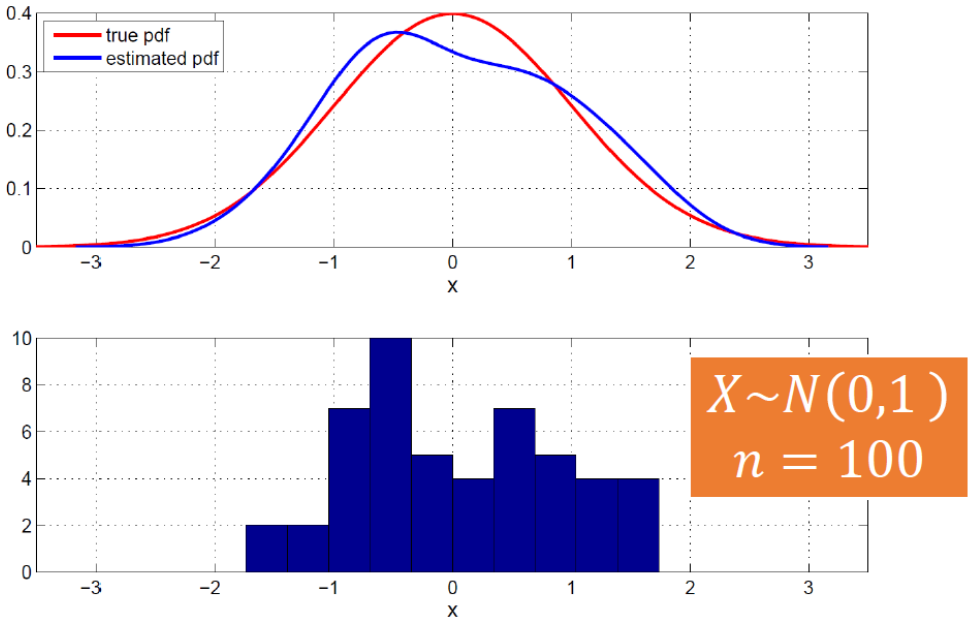


Рис. 6.

Необходимо отметить, что ни один из представленных способов оценки плотностей не является идеальным, так что рекомендуется использовать оба.

### 2. Важные характеристики распределений

Часто возникает необходимость оценить не всю функцию распределения, а некоторые ее параметры. Самым важным классом параметров распределения являются **средние**. Нестрогое определение можно сформулировать следующим образом: среднее — это значение, вокруг которого группируются все остальные.

Одним из вариантов уточнения данного определения является **матожидание**:

$$EX = \begin{cases} \sum\limits_i a_i p_i, & X \text{ — дискретна,} \\ \int\limits_{-\infty}^{+\infty} x\, f(x)\, dx, & X \text{ — непрерывна.} \end{cases}$$

Другой характеристикой среднего является медиана. Она определяется с помощью квантиля. **Квантилем** порядка $\alpha \in (0,1)$ называется величина $X_\alpha$ такая, что:

$$P\big(X \leqslant X_\alpha\big) \geqslant \alpha, \quad P\big(X \geqslant X_\alpha\big) \geqslant 1-\alpha.$$

**Медиана** — это квантиль порядка 0,5:

$$P\big(X \leqslant \operatorname{med} X\big) \geqslant 0{,}5, \quad P\big(X \geqslant \operatorname{med} X\big) \geqslant 0{,}5.$$

Еще одной характеристикой среднего является **мода** — самое вероятное значение случайной величины (в нестрогом смысле):

$$\operatorname{mode} X = \begin{cases} a_{\operatorname{argmax}\limits_i p_i}, & X \text{ — дискретна,} \\ \operatorname{argmax}\limits_x f(x), & X \text{ — непрерывна.} \end{cases}$$

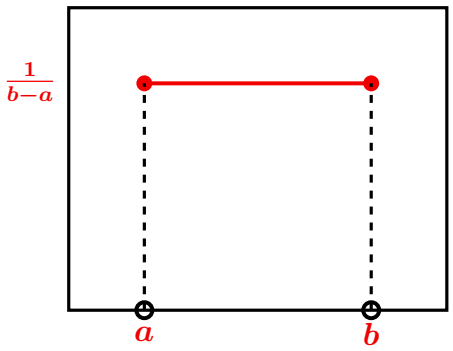


Рис. 7.

В случае нормально распределенной случайной величины ее матожидание, медиана и мода в точности совпадают:

$$X \sim N\big(\mu, \sigma^2\big) \quad \Rightarrow \quad EX = \operatorname{med} X = \operatorname{mode} X = \mu.$$

Если случайная величина $X$ равномерно распределена на отрезке [a,b], то ее матожидание и медиана совпадают:

$$X \sim U(a,b) \quad \Rightarrow \quad EX = \operatorname{med} X = \frac{a+b}{2}.$$

Мода такой случайной величины не определена, поскольку у плотности распределения нет максимума (см. рис. 7). Значит, модой в данном случае может быть любое число на интервале от $a$ до $b$.

В случае бимодального распределения мода приходится на максимум плотности, а медиана и матожидание смещены в сторону второго «горба», причем смещение матожидания больше, чем смещение медианы (см. рис. 8).


Рис. 8.

Следующая рассматриваемая группа параметров распределения — это параметры, характеризующие **разброс**, то есть то, насколько случайная величина концентрируется вокруг своего среднего значения. Одним из наиболее важных параметров здесь является **дисперсия**:

$$DX = E\Big((X - EX)^2\Big).$$

Часто используется величина $\sqrt{DX}$, называемая **среднеквадратическое отклонение**.

Еще одна характеристика разброса — **интерквантильный размах**:

$$\text{IQR} = X_{0,75} - X_{0,25}.$$

Если случайная величина $X$ распределена по закону Пуассона с параметром $\lambda$, то её дисперсия и матожидание совпадают:

$$X \sim \text{Pois}(\lambda) \quad \Rightarrow \quad DX = \lambda, \quad EX = \lambda.$$

Для нормально распределенной случайной величины дисперсия — это второй параметр распределения:

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad DX = \sigma^2.$$

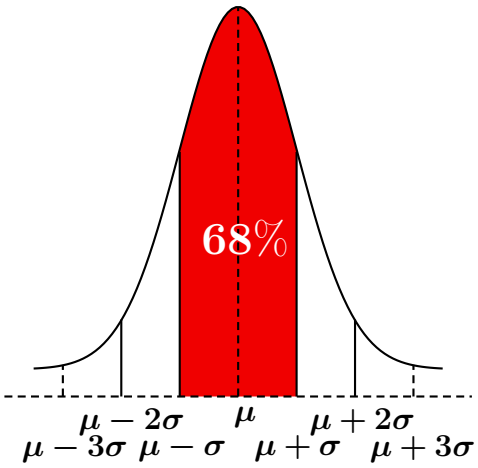


Рис. 9.

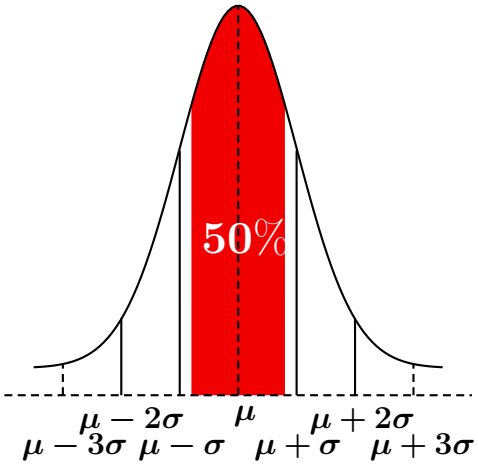


Рис. 10.

В отличие от характеристик среднего значения, характеристики разброса для нормального распределения не совпадают. Для иллюстрации можно отложить от среднего значения $\mu$ интервалы, соответствующие среднеквадратическим отклонениям $\sigma$. В интервале от $\mu-\sigma$ до $\mu+\sigma$ лежит 68% вероятностной массы нормального распределения (см. рис. 9). В интервал, соответствующий интерквартильному размаху вокруг среднего, попадает 50% случайной величины (см. рис. 10).

В интервал от $\mu-2\sigma$ до $\mu+2\sigma$ попадает примерно 95% вероятностной массы нормально распределенной случайной величины (см. рис. 11). Это часто используемое на практике *правило двух сигм*. Другой его вариант — *правило трех сигм* (см. рис. 12): в интервале от $\mu-3\sigma$ до $\mu+3\sigma$ случайная величина реализуется практически со стопроцентной вероятностью (99,7%).

### 3. Важные статистики

Оценка матожидания случайной величины — это **выборочное среднее**:

$$\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i.$$

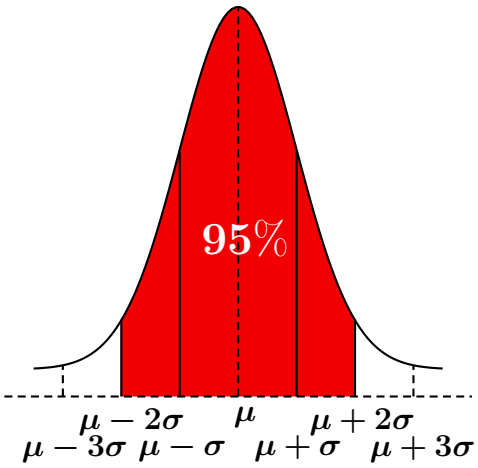


Рис. 11.

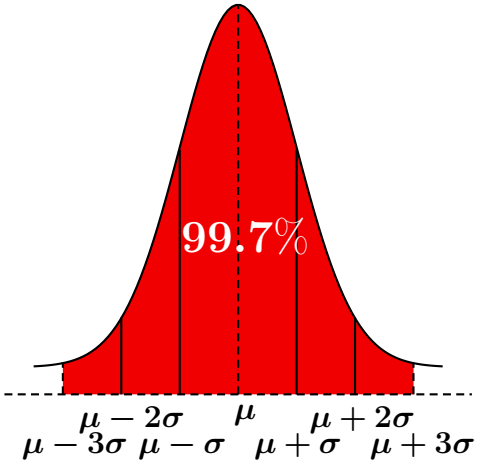


Рис. 12.

Для построения выборочной медианы необходимо составить из рассматриваемой выборки вариационный ряд:

$$X^n = \big(X_1, X_2, \ldots, X_n\big) \quad \Rightarrow \quad X_{(1)} \leqslant X_{(2)} \leqslant \ldots \leqslant X_{(n)}.$$

Элемент вариационного ряда $i$ называется *$i$-й порядковой статистикой*. **Выборочная медиана** является центральным элементом вариационного ряда:

$$m = \begin{cases} X_{(k)}, & n = 2k+1, \\ \dfrac{X_{(k)} + X_{(k+1)}}{2}, & n = 2k. \end{cases}$$

**Выборочная мода** оценивается по максимуму оценки плотности распределения.

Показателен следующий пример. Рассматривается выборка из 25 человек, для каждого из которых известен годовой доход. В выборке есть десять человек, годовой доход которых равен двум тысячам долларов, один человек с годовым доходом в три тысячи долларов, и так далее. Один человек получает сорока пять тысяч долларов в год. Среднее арифметическое годовых доходов на этой выборке — 5700 долларов. Здесь медиана составляет 3000 долларов, а мода — 2000.

Необходимо заметить, что все рассматриваемые величины называются «средними». Значит, для оптимистичного отчета по данной выборке можно воспользоваться средним арифметическим, а для пессимистичного — модой.

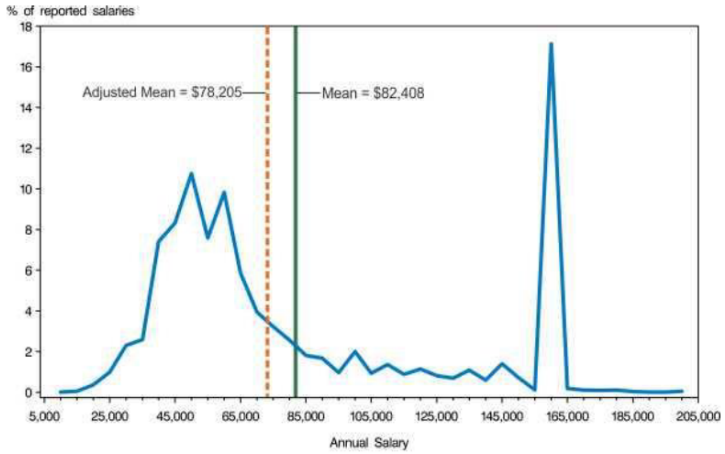


Рис. 13.

**Выборочная дисперсия** оценивает дисперсию и имеет следующий вид:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(X_i - \bar{X}\big)^2.$$

Для построения выборочной оценки интерквартильного размаха необходимо определить *выборочный квантиль* порядка $\alpha$: — это порядковая статистика, порядок которой равен целой части от $\alpha n$ ($X_{([\alpha n])}$). Тогда **выборочный интерквартильный размах** определяется следующим образом:

$$\mathrm{IQR}_n = X_{([0{,}75n])} - X_{([0{,}25n])}.$$

В качестве следующего примера будет рассмотрено распределение годового дохода членов американской ассоциации юристов (см. рис. 13). Имеются два выраженных пика — 168 тысяч долларов и 45 тысяч. Среднее значение, посчитанное по данной выборке, равно 82 тысячам долларов. Видно, что это значение несет очень мало информации о выборке, так как крайне мало людей получают именно такой доход.

Другой показательный пример — *квартет Энскомба* (см. рис. 14). Рассматриваются четыре искусственно сгенерированных пары выборок, характеристики которых в каждом из четырех случаев совпадают (равны выборочные средние и выборочные дисперсии). Однако, при рассмотрении диаграмм рассеяния по этим четырем парам выборок, видно, что в каждом из четырех случаев происходят абсолютно разные вещи (см. рис. 15). Таким образом, даже совокупность статистик не позволяет полностью понять данные, и рекомендуется всегда при анализе данных изучать графики, гистограммы и оценки плотности вместо одних лишь цифр.

### 4. Центральная предельная теорема

Рассмотрим случайную величину $X$ с функцией распределения $F(x)$. Пусть имеется ее выборка объема $n$:

$$X \sim F(x), \quad X^n = \big(X_1, X_2, \ldots, X_n\big),$$

| № | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| $\bar{x}$ | 9 | 9 | 9 | 9 |
| $S_x$ | 11 | 11 | 11 | 11 |
| $\bar{y}$ | 7.5 | 7.5 | 7.5 | 7.5 |
| $S_y$ | 4.13 | 4.13 | 4.13 | 4.13 |

Рис. 14.

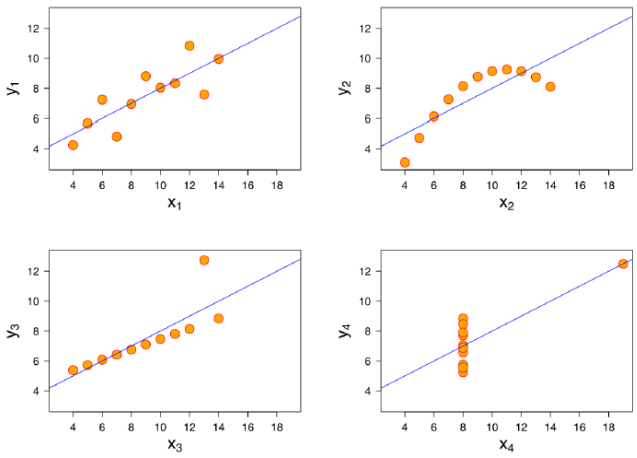


Рис. 15.

По выборке можно вычислить выборочное среднее:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Какое распределение имеет выборочное среднее, и как оно связано с исходным распределением?

Можно провести эксперимент. Берется случайная величина с распределением, показанным на рис. 16.

Из данной случайной величины можно взять выборку объема $n$ и посчитать по ней выборочное среднее. Данное действие необходимо повторить в рамках эксперимента достаточно много раз, чтобы затем построить гистограмму полученных выборочных средних. На рис. 17 приведена гистограмма, построенные по выборкам объема $n = 2$. По сравнению с исходной плотностью случайной величины, данная гистограмма выглядит более гладкой. С увеличением объёма выборки процесс сглаживания продолжается (см. рис. 18 для $n = 3$).

При объеме выборки $n = 5$ гистограмма становится унимодальной (см. рис. 19). Дальнейшее увеличение выборки не влияет на форму гистограммы, она лишь становится более узкой (см. рис. 20 для $n = 30$).

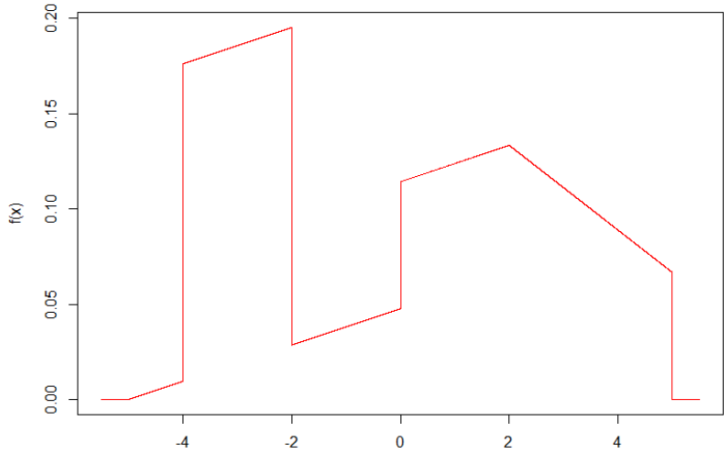


Рис. 16.

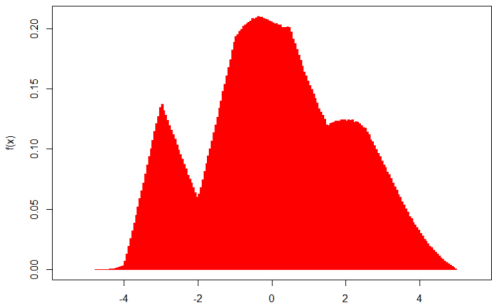


Рис. 17.

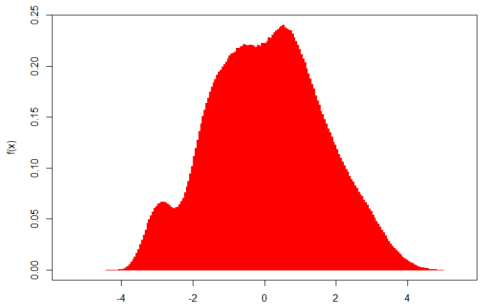


Рис. 18.

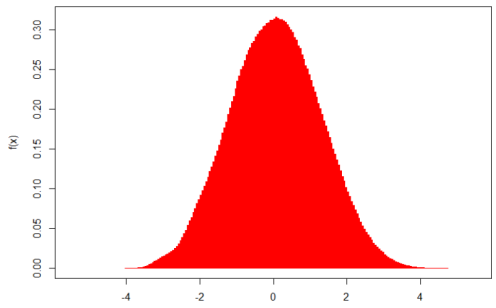


Рис. 19.

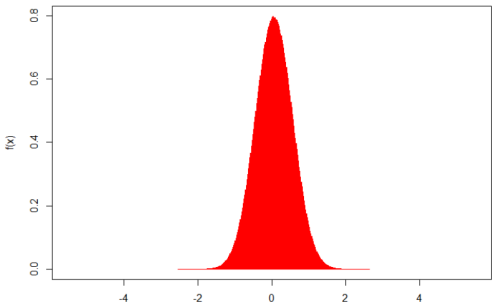


Рис. 20.

Можно заметить, что распределение выборочных средних достаточно хорошо описывается нормальным распределением, что является утверждением **центральной предельной теоремы**:

$$X \sim F(x), \quad X^n = (X_1, X_2, \ldots, X_n) \quad \Rightarrow \quad \bar{X}_n \approx\sim N\left(EX, \frac{DX}{n}\right).$$

С ростом $n$ точность нормальной аппроксимации увеличивается.

Полученный результат справедлив не только для непрерывных распределений, но и для дискретных. Это можно рассмотреть на примере биномиального распределения (см. рис. 21). Как и в предыдущем эксперименте, можно повторить данный несколько раз и построить гистограммы для различного объема выборок (см. рис. 22 для $n = 2$).

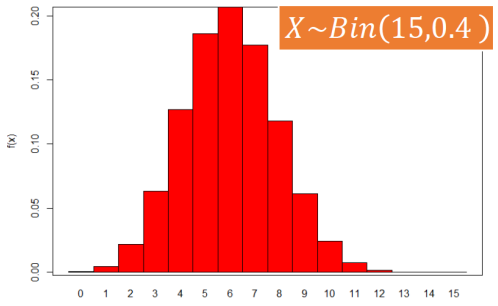


Рис. 21.

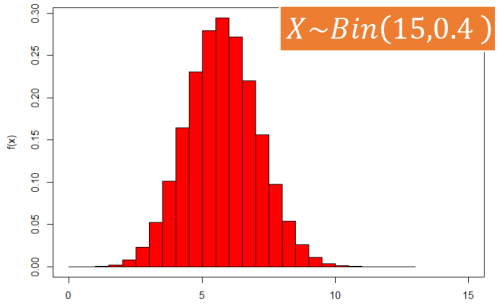


Рис. 22.

При увеличении объема выборок происходит то же, что и в предыдущем эксперименте — распределение становится все более гладким и все более похожим на нормальное (см. рис. 23 для $n = 5$ и рис. 24 для $n = 30$). Таким образом, центральная предельная теорема в данном случае работает.

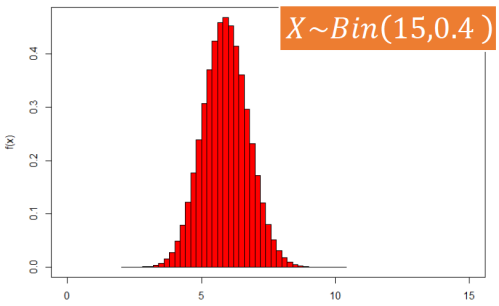


Рис. 23.

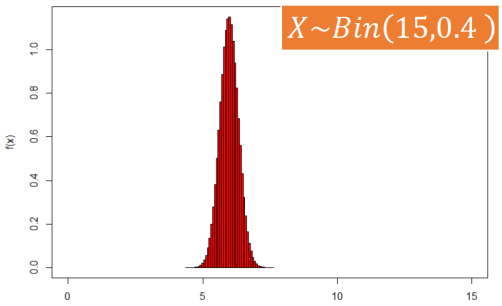


Рис. 24.

Проведём еще один эксперимент для биномиального распределения, на этот раз взяв $p = 0{,}01$. Функция вероятности данной случайной величины показана на рис. 25. На рис. 26 изображено распределение выборочных средних, построенных по выборкам объема $n = 2$.

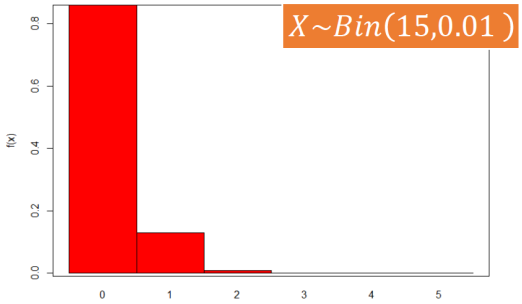


Рис. 25.

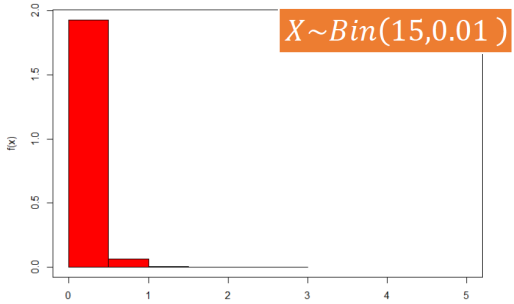


Рис. 26.

Исходное распределение не позволяет распределению выборочных средних быстро сходиться к нормальному (см. рис. 27 для $n = 5$ и рис. 28 для $n = 30$). Даже по выборке объема $n = 30$ гистограмма не может быть хорошо описана нормальным законом — даже относительно максимума данной гистограммы распределение несимметрично.

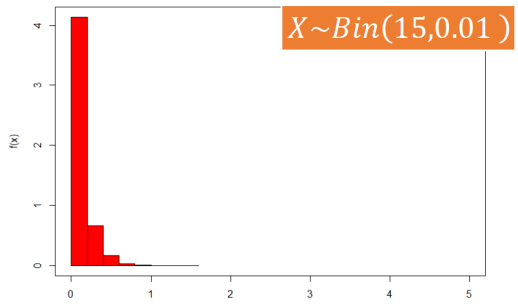


Рис. 27.

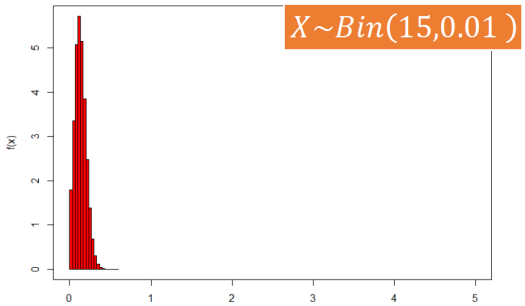


Рис. 28.

Центральная предельная теорема хорошо работает, если исходное распределение не слишком скошено. Существует эмпирическое правило: когда распределение $X$ не слишком скошено, распределение $X_n$ хорошо описывается нормальным при $n \geqslant 30$.

### 5. Доверительные интервалы

Имеется некий продукт, для которого известна его целевая аудитория. Необходимо узнать, насколько хорошо целевая аудитория знакома с данным продуктом. Введём следующую случайную величину:

$$X = \begin{cases} 1, & \text{член ЦА знает продукт,} \\ 0, & \text{не знает.} \end{cases}$$

Такая случайная величина имеет распределение Бернулли с параметром $p$ — *узнаваемостью продукта*:

$$X \sim \text{Ber}(p).$$

Измерить узнаваемость продукта можно с помощью опроса. Если в опросе $n$ участников, то на выходе получится выборка $X^n$, состоящая из 0 и 1. Оценкой узнаваемости по данной выборке будет выборочное среднее:

$$\bar{p}_n = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}_n.$$

Пусть по итогам первого опроса, в котором приняло участие 10 человек, оказалось, что 6 из них знакомы с продуктом:

$$n = 10, \quad \bar{p}_n = 0{,}6.$$

Второй опрос дал следующий результат:

$$n = 100, \quad \bar{p}_n = 0{,}44.$$

Необходимо определить, какая из двух полученных оценок лучше. Точность измеренной оценки определяется с помощью **доверительных интервалов**: пары статистик $C_L, C_U$ такой, что:

$$P\big(C_L \leqslant \theta \leqslant C_U\big) \geqslant 1 - \alpha.$$

Здесь $\theta$ — это *оцениваемый параметр*, $(1-\alpha)$ — *уровень доверия*, а $C_L$ и $C_U$ — *верхний* и *нижний доверительные пределы* (соответственно). При бесконечном повторении эксперимента в $100(1-\alpha)\%$ случаев этот интервал будет накрывать истинное значение параметра $\theta$.

Доверительные интервалы можно построить для оценок узнаваемости продукта из рассматриваемого примера. Оценки узнаваемости являются, по сути, выборочными средними. Следовательно, можно воспользоваться центральной предельной теоремой:

$$\bar{p}_n \approx\sim N\left(EX, \frac{DX}{n}\right).$$

Для случайной величины с распределением Бернулли известно, что:

$$X \sim \text{Ber}(p) \quad \Rightarrow \quad EX = p, \quad D = p(1-p),$$

тогда:

$$\bar{p}_n \approx\sim N\left(p, \frac{p(1-p)}{n}\right).$$

В правую часть данного выражения можно подставить $\bar{p}_n$ вместо $p$:

$$\bar{p}_n \approx\sim N\left(\bar{p}_n, \frac{\bar{p}_n(1-\bar{p}_n)}{n}\right).$$

Распределение стало полностью определенным. Далее необходимо воспользоваться правилом двух сигм:

$$\sigma = \sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}} \quad \Rightarrow \quad P\left(\bar{p}_n - 2\sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}} \leqslant p \leqslant \bar{p}_n + 2\sqrt{\frac{\bar{p}_n(1-\bar{p}_n)}{n}}\right) \approx 0{,}95.$$

Применение полученного выражения к двум опросам даст следующий результат. В первом опросе 95% доверительный интервал — $(0{,}29, 0{,}91)$, во втором — $(0{,}34, 0{,}54)$. Таким образом, доверительный интервал помогает в описании степени неуверенности в полученной оценке.

Доверительные интервалы не обязательно строить с помощью центральной предельной теоремы. Для конкретных распределений существуют более точные способы. Например, для распределения Бернулли наиболее точен метод Уилсона. Однако, именно центральная предельная теорема является универсальным средством построения доверительных интервалов — она работает вне зависимости от того, из какого распределения взята исходная выборка.