§ 26. Нормальные и биномиальные распределения. Законы больших чисел

Распространённость нормальных распределений среди всех других распределений обосновывается *центральной предельной теоремой* и её вариантами, основные из которых были получены российскими математиками П. Л. Чебышёвым, А. А. Марковым, А. М. Ляпуновым в конце XIX — начале XX века.

В описательном виде эта теорема утверждает следующее.

> Распределение суммы n независимых случайных величин, удовлетворяющих некоторым ограничениям[1], становится при неограниченном возрастании n всё более похожим на нормальное распределение.

При этом сами эти ограничения носят весьма общий характер и выполняются в очень многих случаях. Ситуация в целом поразительная: слагаемые могут быть почти что любыми, а вот их сумма (при неограниченном увеличении числа слагаемых) распределена практически по нормальному закону. Грубо говоря, и тут наблюдается одно из проявлений *закона больших чисел*.

Мы ограничимся одним частным, но весьма важным, случаем применимости центральной предельной теоремы. А именно, схемой Бернулли и биномиальным распределением вероятности.

Напомним, что схема Бернулли (или испытания Бернулли) — это серия из нескольких независимых повторений одного и того же испытания с двумя исходами «успех» и «неудача». Вероятность наступления ровно k «успехов» в n испытаниях Бернулли находится по формуле

$$P_n(k) = C_n^k \cdot p^k \cdot q^{n-k}$$

Где p и $q = 1 - p$ — вероятности, соответственно, «успеха» и «неудачи» в одном испытании.

Вероятности $P_n(k)$ распределены по *биномиальному з*акону распределения. А именно, число k «успехов» принимает значения 0, 1, 2,..., n со следующими вероятностями:

[1] Например, условиям Ляпунова.

Число «успехов»	0	1	…	k	…	$n-1$	n
Вероятность	q^n	$C_n^1 pq^{n-1}$	…	$C_n^k p^k q^{n-k}$	…	$C_n^{n-1} p^{n-1} q$	p^n

Это распределение вероятности называют биномиальным, потому что числа из второй строки этой таблицы соответственно равны слагаемым в биноме Ньютона:

$$1 = (q+p)^n = C_n^0 q^n p^0 + C_n^1 q^{n-1} p^1 + C_n^2 q^{n-2} p^2 + \ldots + \\ + C_n^k q^{n-k} p^k + \ldots + C_n^{n-1} q^1 p^{n-1} + C_n^n q^0 p^n.$$

Подчеркнём, что независимость здесь — ключевой момент: при последовательном и независимом проведении, скажем, двух испытаний вероятность того, что в первом из них наступит событие A, а во втором — событие B, равна произведению $P(A) \cdot P(B)$ вероятностей этих событий. Без условия независимости не получатся ни формула Бернулли, ни биномиальное распределение.

Пример 1 Вероятность того, что биатлонист попадёт на рубеже во все пять мишеней оценивается в 80 %. Оценить вероятность того, что в индивидуальной гонке[1] биатлонист:

а) ни разу не промахнётся;

б) промахнётся только на одном рубеже;

в) промахнётся на половине рубежей;

г) не промахнётся только на одном рубеже.

Решение. В данном случае $n = 4$ — число рубежей, $p = 0{,}8$ — вероятность «успеха», т. е. вероятность попадания во все мишени на рубеже, $q = 0{,}2$ — вероятность «неудачи».

а) Здесь $k = 4$ и по формуле Бернулли получаем:

$$P_n(k) = P_4(4) = C_4^4 \cdot p^4 \cdot q^0 = p^4 = 0{,}8^4 = 0{,}4096;$$ примерно 41 %.

б) Промахнуться только на одном рубеже — это значит, что в четырёх повторениях мы имеем 3 «успеха» и одну «неудачу». Значит,

$$P_n(k) = P_4(3) = C_4^3 \cdot p^3 \cdot q^1 = 4p^3 q = 4 \cdot 0{,}8^3 \cdot 0{,}2 = 0{,}4096.$$

Ответ совпал с ответом из пункта а), но это случайность: такими оказались числовые данные.

[1] В этой гонке четыре стрелковых рубежа.

в) В этом случаем имеем 2 «успеха» и 2 «неудачи», значит,

$$P_n(k) = P_4(2) = C_4^2 \cdot p^2 \cdot q^2 = 6(pq)^2 = 6 \cdot 0{,}16^2 = 0{,}1536.$$

г) $P_n(k) = P_4(1) = C_4^1 \cdot p^1 \cdot q^3 = 4 \cdot 0{,}8 \cdot 0{,}2^3 = 0{,}0256.$

Абсолютная точность ответа, который даёт формула Бернулли, есть её несомненное достоинство. Но, к сожалению, ответ получается крайне сложным для практического применения. Представьте, что предстоит вычислить, скажем, $\frac{20!}{11! \cdot 9!} \cdot 0{,}23^{11} \cdot 0{,}77^9$ или $\frac{100!}{27! \cdot 73!} \cdot 0{,}7^{27} \cdot 0{,}3^{73}$.

Значит, для реальных приложений при больших n необходимы способы *приближённых вычислений* в формуле Бернулли. Такие способы есть и, пожалуй, основной из них как раз и связан с гауссовой функцией $y = \varphi(x)$ и функцией Лапласа $y = \Phi(x)$, которые как раз и появились в науке (XVIII в.) как ответ на вопрос о приближённых вычислениях в формуле Бернулли.

Для серии из n испытаний Бернулли обозначим k_1 число «успехов» в первом испытании, k_2 — число «успехов» во втором испытании, …, k_n — число «успехов» в последнем испытании. Тогда сумма $k_1 + k_2 + \ldots + k_n$ — это общее число k «успехов » во всей серии. Каждое из слагаемых k_1, k_2, … , k_n — это дискретная случайная величина, принимающая значения 0 или 1. Таблицы распределения у всех слагаемых одинаковы:

Число «успехов»	0	1
Вероятность	q	p

У всех слагаемых одно и то же математическое ожидание, равное p, и одна и та же дисперсия, равная pq, а у их суммы $k = k_1 + k_2 + \ldots + k_n$ математическое ожидание и дисперсия равны, соответственно np и npq. Оказывается, что для независимых слагаемых k_1, k_2, … , k_n выполнены ограничения из центральной предельной теоремы. Значит, по этой теореме распределение суммы $k = k_1 + k_2 + \ldots + k_n$ при неограниченном возрастании n приближается к нормальному закону.

Для чисел $a < b$ вероятность $P_n(a \leqslant k \leqslant b)$ того, что количество k «успехов» в n испытаниях Бернулли не меньше a и не больше b вычисляют по формуле

$$P_n(a \leqslant k \leqslant b) \approx \Phi\left(\frac{b - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - np}{\sqrt{npq}}\right).$$

Напомним, что $y = \Phi(x)$ — функция Лапласа, таблицу значений которой можно найти в *Приложении*. Вопрос о том, начиная с каких именно n, это приближение даёт приемлемую точность, весьма сложен. Обычно ограничиваются рекомендацией о проверке справедливости неравенства $npq > 10$. В частности, при $p = 0{,}5 = q$ оно равносильно неравенству $n > 40$, а при $p = 0{,}2$; $q = 0{,}8$ — неравенству $n > 62{,}5$. Коротко: приемлемая точность приближения достигается при нескольких десятках, сотнях (или тысячах) повторений испытания.

Пример 2 Политика поддерживают в среднем 60 % населения. Какова вероятность того, что из 600 случайно опрошенных людей этого политика поддерживают:

а) от 360 до 380 человек; б) не более 350 человек.

Решение. Проводят $n = 600$ повторений одного и того же опыта — получение ответа «да» или «нет» на вопрос о поддержке политика. Предполагается, что это независимые повторения и схема Бернулли тут применима. По условию $p = 0{,}6$, $q = 0{,}4$, т. е. $np = 360$, $npq = 144 > 10$, $\sqrt{npq} = 12$.

а) Здесь $a = 360$, $b = 380$ и

$$P_{600}(360 \leqslant k \leqslant 380) \approx \Phi\left(\frac{380 - 360}{12}\right) - \Phi(0) \approx \Phi(1{,}67) \approx 0{,}4525.$$

б) Здесь указана только оценка сверху $k \leqslant 350$, $b = 350$. В качестве нижней границы всегда можно взять $a = 0$ и рассмотреть вероятность $P_{600}(0 \leqslant k \leqslant 350)$. Она приближённо равна

$$\Phi\left(\frac{350 - 360}{12}\right) - \Phi\left(\frac{0 - 360}{12}\right) \approx \Phi(-0{,}83) - \Phi(-30).$$

Но в таблице для функции Лапласа $y = \Phi(x)$ нет отрицательных значений x. Как быть? Напомним, что функцию Лапласа $y = \Phi(x)$ для отрицательных аргументов доопределяют по нечётности, т. е. $\Phi(-x) = -\Phi(x)$ (см. § 25).

Получаем

$$P_{600}(0 \leqslant k \leqslant 350) \approx \Phi(-0{,}83) - \Phi(-30) = \Phi(30) - \Phi(0{,}83) \approx$$
$$\approx 0{,}5 - 0{,}2967 = 0{,}2033.$$

Пример 3 Вероятность рождения мальчика примем равной 50 %. Найти вероятность того, что среди 400 новорожденных будет:

а) ровно 200 мальчиков;

б) ровно 205 мальчиков.

Решение. Во-первых, и без вычислений ясно, что вероятность наступления точного равенства крайне невелика. Во-вторых, если приравнять $P_{400}(k = 200) = P_{400}(200 \leqslant k \leqslant 200)$ и попытаться применить формулу $P_n(a \leqslant k \leqslant b) \approx \Phi\left(\dfrac{b - np}{\sqrt{npq}}\right) - \Phi\left(\dfrac{a - np}{\sqrt{npq}}\right)$, то в правой части приближённого равенства получится 0, так как $b = a$. В-третьих, необходимо отметить, что для вероятностей $P_n(k = d)$ есть способ подсчёта, основанный на таблице значений гауссовой функции $y = \varphi(x)$, а не функции $y = \Phi(x)$. Мы его рассматривать не будем, а покажем, как можно обойтись только функцией $y = \Phi(x)$, вводя так называемую *поправку на непрерывность*.

Для целого числа d вероятность $P_n(k = d)$ того, что количество k «успехов» в n испытаниях Бернулли равно d вычисляют по формуле

$$P_n(k = d) \approx \Phi\left(\frac{d + 0{,}5 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{d - 0{,}5 - np}{\sqrt{npq}}\right).$$

а) По этой формуле получаем

$$P_{400}(k = 200) \approx \Phi\left(\frac{200 + 0{,}5 - 200}{10}\right) - \Phi\left(\frac{200 - 0{,}5 - 200}{10}\right) =$$
$$= 2\Phi(0{,}05) = 2 \cdot 0{,}0199 \approx 0{,}04.$$

б) Ещё раз используем ту же формулу:

$$P_{400}(k = 205) \approx \Phi\left(\frac{205{,}5 - 200}{10}\right) - \Phi\left(\frac{204{,}5 - 200}{10}\right) =$$
$$= \Phi(0{,}55) - \Phi(0{,}45) \approx 0{,}2088 - 0{,}1736 \approx 0{,}0252.$$

Следствиями центральной предельной теоремы (теорем) являются различные варианты закона больших чисел. В учебниках для

9—10-х классов мы уже рассказывали о явлении *статистической устойчивости*. Оно заключается в практическом совпадении теоретического и статистического определений вероятности случайного события.

Допустим, что имеется возможность неограниченно повторять некоторое испытание и наблюдать наступление (или не наступление) события A, вероятность $P(A)$ которого нам неизвестна, а известные нам теоретические сведения никак не дают ответа. Как найти $P(A)$? Рассмотрим наше испытание, как испытание Бернулли, а его повторения, как схему Бернулли. «Успех» — наступление события A, вероятность p «успеха» равна $P(A)$. Проведём $n = 10, 20, 30, \ldots, 100, \ldots$ повторений и всякий раз будем вычислять частоту $\frac{k}{n}$ наступления события A, т. е. частоту наступления «успеха». Закон больших чисел гарантирует нам, что чем больше будет проведено повторений, тем точнее будет приближение $\frac{k}{n} \approx p$: вероятность заметной ошибки в этом приближении стремится к нулю с ростом n.

Закон больших чисел (в форме Бернулли). При неограниченном увеличении n частота $\frac{k}{n}$ наступления «успеха» в n испытаниях Бернулли практически совпадает с вероятностью p «успеха» в одном испытании.

Для обоснования ЗБЧ используем формулы приближённых вычислений с помощью функции Лапласа $y = \Phi(x)$. Оценим вероятность того, что отклонение $\left|\frac{k}{n} - p\right|$ окажется не больше некоторого фиксированного числа $t > 0$. Неравенство $\left|\frac{k}{n} - p\right| < t$ заменим равносильным ему двойным неравенством $-tn \leqslant k - np \leqslant tn$ или $np - tn \leqslant k \leqslant np + tn$. Имеем

$$P_n\left(\left|\frac{k}{n} - p\right| \leqslant t\right) = P_n(np - tn \leqslant k \leqslant np + tn) \approx$$

$$\approx 2\Phi\left(\frac{tn}{\sqrt{npq}}\right) = 2\Phi\left(\frac{t}{\sqrt{pq}} \cdot \sqrt{n}\right).$$

В произведении $\frac{t}{\sqrt{pq}} \cdot \sqrt{n}$ сомножитель $\frac{t}{\sqrt{pq}}$ постоянен, а сомножитель $\sqrt{n}$ неограниченно возрастает. Значит, возрастает само произведение, и как только оно станет больше, скажем 3, то значение $2\Phi\left(\frac{t}{\sqrt{pq}} \cdot \sqrt{n}\right)$ превысит 0,997. А если $\frac{t}{\sqrt{pq}} \cdot \sqrt{n}$ станет больше 5, то значение $2\Phi\left(\frac{t}{\sqrt{pq}} \cdot \sqrt{n}\right)$ превысит 0,999999. Значит, событие «$\left|\frac{k}{n} - p\right| \leqslant t$» становится практически достоверным при неограниченном увеличении n: $\lim\limits_{n\to\infty} P_n\left(\left|\frac{k}{n} - p\right| \leqslant t\right) = 1$.

Пример 4 Насколько большим должно быть число n повторений испытания Бернулли для того, чтобы с вероятностью более 95 % можно было бы утверждать, что погрешность приближения $\frac{k}{n} \approx p$ не превышает 0,05, если вероятность p «успеха» равна 0,2?

Решение. По условию $\left|\frac{k}{n} - p\right| \leqslant 0{,}05$ или $|k - np| \leqslant 0{,}05n$, т. е. число k «успехов» должно находиться в пределах от $np - 0{,}05n$ до $np + 0{,}05n$. Вероятность этого события находим приближённо, с учётом $p = 0{,}2$, $q = 0{,}8$, $\sqrt{pq} = 0{,}4$:

$$P_n(np - 0{,}05n \leqslant k \leqslant np + 0{,}05n) \approx$$
$$\approx \Phi\left(\frac{0{,}05n}{\sqrt{npq}}\right) - \Phi\left(-\frac{0{,}05n}{\sqrt{npq}}\right) = 2\Phi\left(\frac{0{,}05n}{\sqrt{npq}}\right) = 2\Phi\left(\frac{\sqrt{n}}{8}\right).$$

По условию требуется, чтобы $2\Phi\left(\frac{\sqrt{n}}{8}\right) \geqslant 0{,}95$, $\Phi\left(\frac{\sqrt{n}}{8}\right) \geqslant 0{,}475$. По таблице в Приложении находим, что $0{,}475 \approx \Phi(1{,}96)$. Так как функция Φ возрастает, то получаем

$$\frac{\sqrt{n}}{8} \geqslant 1{,}96, \quad \sqrt{n} \geqslant 15{,}68, \quad n \geqslant 245{,}8624.$$

Ответ: не менее 245 повторений.

Для реальных приложений абсолютная точность ответа в предыдущем примере представляется излишней. Вполне можно было бы сказать, что следует провести, не менее 250 или даже 300 повторений.

Закон больших чисел является основой для *выборочного метода*, необходимого в различных социологических и статистических исследованиях. Например, для того чтобы оценить реальный рейтинг телеканала вовсе не нужно опрашивать всех владельцев телевизоров (их — десятки миллионов). При допустимой вероятности ошибки в несколько процентов можно сделать такой опрос лишь *выборочно*, для достаточно большого числа независимо опрашиваемых респондентов. Оказывается, для этого вполне достаточно взять выборку примерно в две тысячи человек.

Великий русский математик Пафнутий Львович Чебышёв доказал, что закон больших чисел имеет место не только для числа «успехов» в испытаниях Бернулли, но и, в значительно более общем виде, для последовательностей случайных величин. Открытие П. Л. Чебышёва основано на замечательном неравенстве, носящем его имя. Это неравенство верно для любой случайной величины S, у которой есть математическое ожидание $M(S)$ и дисперсия $D(S)$. Оно оценивает вероятность отклонения $|S - M(S)|$ случайной величины S от своего математического ожидания $M(S)$ через величину дисперсии $D(S)$.

Неравенство Чебышёва

$$P\big(|S - M(S)| \leqslant t\big) \geqslant 1 - \frac{D(S)}{t^2}$$

Например, если $t = 3\sigma = 3\sqrt{D}$, то $P\big(|S - M(S)| \leqslant 3\sigma\big) \geqslant 1 - \frac{1}{9}$. Получается, что с вероятностью не менее чем 88,8 %, совершенно *произвольная* с. в., а не только число «успехов», отклоняется от своего математического ожидания не более чем на 3σ (так называемое правило «трёх сигм»).

ЗБЧ (в форме Чебышёва) обосновывает хорошо известное экспериментальное правило «среднего». Например, пусть нужно измерить какой-то числовой показатель физического объекта. Его измеряют в первый раз, получают S_1. Затем в тех же условиях независимо измеряют его во второй раз, получают S_2 и т. д. Каждый раз возможны погрешности и результаты S_1, S_2, S_3, … измерений несколько отличаются друг от друга. Но если взять их среднее арифметическое

$\frac{S_1 + S_2 + \ldots + S_n}{n}$, то по ЗБЧ результат получится более надежным: с ростом n практически несомненно, что отклонение $\frac{S_1 + S_2 + \ldots + S_n}{n}$ от истинного значения M не превысит заданной точности t.

Упражнения

На столе стоят одинаковые по виду коробки. Среди них 5 пустых, 3 с призом и 2 с сюрпризом.

26.1. Наудачу выбирают две коробки. Какова вероятность того, что:
а) они пустые;
б) они обе не пустые;
в) одна из них пустая, а другая нет;
г) одна из них пустая, а другая — с сюрпризом;
д) они обе с призом;
е) в них нет сюрприза?

26.2. Наудачу выбирают три коробки. Какова вероятность того, что:
а) все они пустые;
б) все они не пустые;
в) одна из них пустая, одна с призом и одна с сюрпризом;
г) две из них пустые, а одна с призом;
д) одна пустая и две с сюрпризом;
е) в них нет сюрприза?

26.3. В каждом из пунктов а)—е) определите значения n, k, p, q и по формуле Бернулли выпишите (без вычислений) выражение для вероятности появления ровно:
а) 6 «орлов» при 16 бросаниях монеты;
б) 4 «решек» при 24 бросаниях монеты;
в) 99 нечётных чисел при 199 независимых выборах целых чисел от 0 до 9;
г) 50 чисел, кратных трём, при 500 независимых выборах целых чисел от 0 до 9.
д) 170 «неудач» при опросе 700 человек, произвольно называющих день недели, если считать «удачными» днями субботу и воскресенье.
е) 573 «удачных» из 735 бросаний кубика; «удачей» считаем выпадение 5 или 6 очков.

26.4. В приведённых формулах для подсчёта по формуле Бернулли есть пропуски (отмечены знаком ?). Заполните эти пропуски, если известно, что «успех» не менее вероятен чем «неудача».

а) $P_{60}(?) = C_?^6 \cdot 0{,}4^? \cdot ?^?$;
б) $P_?(?) = C_{34}^? \cdot 0{,}7^? \cdot ?^{28}$;
в) $P_?(5) = C_?^? \cdot 0{,}5^{55}$;
г) $P_{100}(10) = C_?^? \cdot 0{,}9^? \cdot ?^?$;
д) $P_?(?) = C_{33}^? \cdot ?^? \cdot 0{,}4^{22}$;
е) $P_{50}(?) = C_?^{45} \cdot 0{,}6^? \cdot ?^?$

После успешного прохождения первого уровня компьютерной игры на мониторе запускается лотерейный барабан, в котором 2 белых и 8 чёрных одинаковых по размеру шаров. Если выпадает белый шар, то игрок переходит сразу на третий уровень и это — «успех». Если выпадает чёрный шар, то игрок переходит на второй уровень и это — «неудача». Требуется применить формулу $P_n(a \leqslant k \leqslant b) \approx \Phi\left(\frac{b - np}{\sqrt{npq}}\right) - \Phi\left(\frac{a - np}{\sqrt{npq}}\right)$ для нахождения вероятности того, что число k «успехов» не менее, чем a, и не более, чем b.

26.5. Найдите:
а) вероятность p «успеха»;
б) вероятность q «неудачи»;
в) среднее np для $n = 10\,000$ повторений;
г) среднее квадратическое $\sqrt{npq}$;
д) границы a и b при отклонении k от np не более, чем на 40;
е) вероятность $P_n(a \leqslant k \leqslant b)$.

26.6. Для $n = 2500$ повторений найдите:
а) среднее np;
б) среднее квадратическое $\sqrt{npq}$;
в) вероятность $P_n(k \leqslant 700)$;
г) вероятность $P_n(k \leqslant 250)$;
д) вероятность $P_n(500 \leqslant k)$;
е) вероятность $P_n(450 \leqslant k \leqslant 550)$.

Вероятность рождения мальчика примем равной 50 %.

26.7. Найдите вероятность того, что девочек среди 10 000 новорождённых будет:
а) не более 4000;
б) не более 5000;
в) не более 6000;
г) от 4900 до 5000;
д) от 4950 до 5050;
е) от 4800 до 5100.

26.8. Найдите вероятность того, что мальчиков среди 100 новорождённых будет ровно:

а) 50; б) 51; в) 52;
г) 53; д) 60; е) 45.

В каждом из 150 контейнеров лежат 10 одинаковых по виду коробок, в 4 из которых — синие новогодние шары, а 6 — красные. Из каждого контейнера наудачу выбирают одну коробку.

26.9. Найдите вероятность того, что в 150 выбранных контейнерах синих шаров будет:

а) меньше 40; в) не больше 60; д) от 54 до 66;
б) меньше 80; г) не меньше 60; е) от 48 до 72.

26.10. По формуле $P_n(k = d) \approx \Phi\left(\frac{d + 0{,}5 - np}{\sqrt{npq}}\right) - \Phi\left(\frac{d - 0{,}5 - np}{\sqrt{npq}}\right)$ найдите вероятность того, что в 150 выбранных контейнерах синих шаров будет ровно:

а) 6; в) 136; д) 64;
б) 36; г) 60; е) 52.

Упражнения для повторения

26.11. Найдите точки экстремума функции:

а) $y = 0{,}2x^5 + 4x^3 - 13x$;
б) $y = e^{5-x}(x^2 - 8x + 8)$;
в) $y = x^2 - 3x + 3\ln(x + 1)$;
г) $y = -\frac{1}{3}x^3 + x^2 + 3x - 11$;
д) $y = e^{x-1}(x^2 - 3x + 3)$;
е) $y = x^2 + 5x + 14\ln(3 - x)$.

26.12. Найдите наименьшее и наибольшее значения функции на указанном отрезке:

а) $y = (10 - x)e^{x-9}$, $[8; 10]$;
б) $y = 4x - \ln(x + 2)^4$, $[-1; 0]$;
в) $y = \cos 3x - 3x + \pi$, $\left[-\frac{\pi}{3}; 0\right]$;
г) $y = (x + 3)e^{2-x}$, $[-3; 2]$;
д) $y = \ln(x + 4)^3 - 3x$, $(-4; 0]$;
е) $y = 4x - \sin 4x + 1$, $\left[-\frac{\pi}{4}; 0\right]$.