Выборочные распределения

Основной целью анализа данных являются статистические выводы, т.е. применение выборочных показателей для оценки параметров генеральной совокупности. Статистические выводы относятся к генеральным совокупностям, а не к выборкам из них. Например, социологи изучают результаты выборочных обследований только для того, чтобы оценить шансы кандидатов получить голоса из всей генеральной совокупности избирателей в целом. Выборочное среднее, полученное при обследовании конкретной выборки, само по себе интереса не представляет. [1]

На практике из генеральной совокупности извлекается выборка заранее установленного объема. Элементы, принадлежащие данной выборке, выбираются случайным образом, например, с помощью датчика случайных чисел. Распределения выборочных параметров называют выборочными.

Выборочное распределение средних значений

Ранее мы рассмотрели несколько оценок математического ожидания распределения. Чаще всего для этого используется арифметическое среднее. Это наилучшая оценка математического ожидания, если распределение является нормальным.

Арифметическое среднее называется **несмещенным**, поскольку *среднее значение всех выборочных средних (при заданном объеме выборки n) равно математическому ожиданию генеральной совокупности*. Продемонстрируем это свойство на примере. Предположим, что генеральная совокупность машинисток в секретариате компании состоит из четырех сотрудниц. Каждую из них попросили напечатать один и тот же текст. Количество опечаток, сделанных каждой машинисткой: Энн – $X_1 = 3$, Кэт – $X_2 = 2$, Карла – $X_3 = 1$, Ширли – $X_4 = 4$. Распределение ошибок приведено на рис. 1.

Рис. 1. Количество опечаток, сделанных четырьмя машинистками

Скачать заметку в формате Word или pdf, примеры в формате Excel2013

Математическим ожиданием генеральной совокупности называется сумма всех значений совокупности, деленная на ее объем:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$

где μ – математическое ожидание генеральной совокупности, N – объем генеральной совокупности, X_i— i-й элемент генеральной совокупности.

Стандартным отклонением генеральной совокупности называется корень квадратный из ее дисперсии:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

Таким образом, в нашем примере:

$$\mu = \frac{3+2+1+4}{4} = 2{,}5 \text{ опечатки}$$

$$\sigma = \sqrt{\frac{(3-2{,}5)^2 + (2-2{,}5)^2 + (1-2{,}5)^2 + (4-2{,}5)^2}{4}} = 1{,}12 \text{ опечатки}$$

Если из этой генеральной совокупности необходимо извлечь с возвращением выборку, состоящую из двух машинисток, возникает 16 вариантов выбора ($N^m = 4^2 = 16$, подробнее см. *Первое правило счета* в заметке Основные понятия теории вероятностей). Эти варианты приведены в таблице (рис. 2). Если усреднить все 16 средних значений, мы получим величину $\mu_{\bar{x}}$, равную математическому ожиданию генеральной совокупности μ, т.е. числу 2,5.

Выборка	Машинистки	Количество опечаток	Среднее значение
1	Энн, Энн	3, 3	$\bar{X}_1 = 3$
2	Энн, Кэт	3, 2	$\bar{X}_2 = 2,5$
3	Энн, Карла	3, 1	$\bar{X}_3 = 2$
4	Энн, Ширли	3, 4	$\bar{X}_4 = 3,5$
5	Кэт, Энн	2, 3	$\bar{X}_5 = 2,5$
6	Кэт, Кэт	2, 2	$\bar{X}_6 = 2$
7	Кэт, Карла	2, 1	$\bar{X}_7 = 1,5$
8	Кэт, Ширли	2, 4	$\bar{X}_8 = 3$
9	Карла, Энн	1, 3	$\bar{X}_9 = 2$
10	Карла, Кэт	1, 2	$\bar{X}_{10} = 1,5$
11	Карла, Карла	1, 1	$\bar{X}_{11} = 1$
12	Карла, Ширли	1, 4	$\bar{X}_{12} = 2,5$
13	Ширли, Энн	4, 3	$\bar{X}_{13} = 3,5$
14	Ширли, Кэт	4, 2	$\bar{X}_{14} = 3$
15	Ширли, Карла	4, 1	$\bar{X}_{15} = 2,5$
16	Ширли, Ширли	4, 4	$\bar{X}_{16} = 4$
			$\mu_{\bar{X}} = 2,5$

Рис. 2. Все возможные варианты выбора двух машинисток из четырех

Итак, среднее значение всех выборочных средних $\mu_{\bar{x}}$ равно математическому ожиданию генеральной совокупности. Следовательно, хотя нам неизвестно, насколько хорошо конкретное выборочное среднее аппроксимирует математическое ожидание

генеральной совокупности, среднее значение всех выборочных средних совпадает с математическим ожиданием генеральной совокупности.

Стандартная ошибка среднего. На рис. 3 приведено выборочное распределение среднего количества ошибок, сделанных машинистками, образующих все 16 возможных выборок, полученных путем случайного выбора с возвращением. Как видим, колебание выборочных средних вокруг математического ожидания генеральной совокупности меньше, чем колебание исходных данных. Этот факт непосредственно следует из закона больших чисел. Исходная генеральная совокупность может содержать числа, которые являются как очень большими, так и очень маленькими. Однако, если экстремальное значение попадет в выборку, ее влияние на среднее значение будет ослаблено, поскольку оно будет просуммировано со всеми остальными элементами выборки. При увеличении объема выборки влияние экстремальных значений ослабевает, поскольку в усреднении принимает участие все большее количество элементов.

Рис. 3. Распределение выборочных средних из таблицы (рис. 2); по оси абсцисс отложены значения выборочных средних X_i, по оси ординат – частота встречаемости этих значений; например, на рис. 2 среднее значение X_i = 2,5 встретилось 4 раза, т.е. точке с абсциссой 2,5 соответствует ордината 4

Диапазон изменения выборочных средних описывается их

стандартным отклонением. Эта величина называется стандартной ошибкой среднего и обозначается как $\sigma_{\bar{X}}$. Стандартная ошибка среднего равна стандартному отклонению генеральной совокупности σ, деленному на квадратный корень из объема выборки n:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Следовательно, при возрастании объема выборки n стандартная ошибка среднего уменьшается со скоростью, пропорциональной квадратному корню из n. Эту формулу можно применять для аппроксимации стандартной ошибки среднего, если выборки извлекаются из генеральной совокупности без возвращения при условии, что каждая выборка содержит не более 5% элементов всей генеральной совокупности. Проиллюстрируем это свойство следующим примером. Если из нескольких тысяч коробок случайным образом извлекается без возвращения выборка из 25 коробок, в нее попадет не более 5% элементов всей генеральной совокупности. Вычислите стандартную ошибку среднего, если стандартное отклонение веса коробки равно 15 г. Подставим в формулу значения n = 25 и σ =15, получаем стандартную ошибку среднего:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{25}} = \frac{15}{5} = 3$$

Обратите внимание на то, что изменчивость выборочных средних намного меньше, чем изменчивость исходных данных (т.е. $\sigma_{\bar{X}}$ = 3 намного меньше, чем σ = 15).

Выборки из нормально распределенных генеральных совокупностей. Введя понятие выборочных распределений и дав определение стандартной ошибки среднего, мы можем ответить на вопрос, как распределены выборочные средние $\bar{X}$. Можно доказать, что если выборки извлекаются с возвращением из нормально распределенной генеральной совокупности, математическое

ожидание которого равно μ, а стандартное отклонение — σ, то выборочное распределение средних также является нормальным при любом объеме выборок n, причем $\mu_{\bar{X}} = \mu$, а стандартная ошибка — $\sigma_{\bar{X}}$.

В наиболее простом варианте, когда объем каждой выборки равен единице, каждое выборочное среднее равно единственному элементу выборки:

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

Следовательно, если генеральная совокупность является нормально распределенной, причем ее математическое ожидание равно μ, а стандартное отклонение — σ, то выборочное распределение средних также является нормальным при $n = 1$, причем $\mu_{\bar{X}} = \mu$, а стандартная ошибка $\sigma_{\bar{X}} = \sigma/\sqrt{1} = \sigma$. Обратите внимание на то, что при увеличении объема выборок выборочное распределение средних остается нормальным, причем $\mu_{\bar{X}} = \mu$. Однако увеличение объема выборки приводит к уменьшению стандартной ошибки среднего, поэтому чем больше становится выборка, тем ближе становятся выборочные средние к математическому ожиданию генеральной совокупности.

В этом можно убедиться, проанализировав рис. 4. На нем изображены выборочные распределения среднего, построенные по 500 выборкам с объемами $n = 1, 2, 4, 8, 16$ и 32, случайным образом извлеченным из нормально распределенной генеральной совокупности. Полигоны, изображенные на рис. 4, свидетельствуют от том, что выборочное распределение средних является лишь приближенно нормальным. Однако по мере возрастания объема выборок выборочные средние становятся ближе к математическому ожиданию генеральной совокупности.

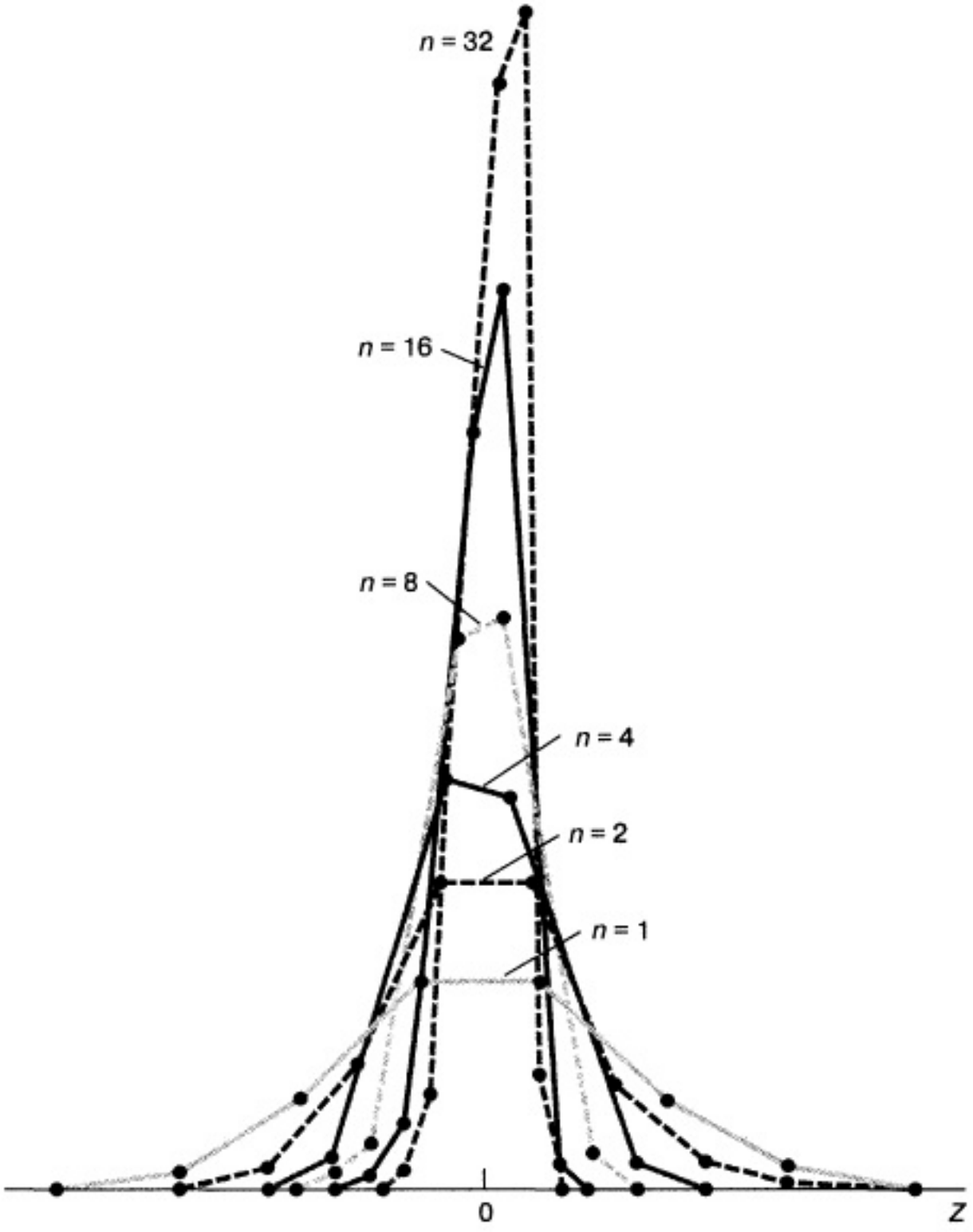


Рис. 4. Выборочные распределения средних, построенные по 500 выборкам с объемами n = 1, 2, 4, 8, 16 и 32, извлеченным из нормально распределенной генеральной совокупности

Чтобы глубже разобраться в понятии выборочного распределения, вернемся к примеру с коробками. Предположим, что упаковочная машина, заполняющая 368-граммовые коробки, настроена так, что количество кукурузных хлопьев, засыпанных в мешки, распределено нормально, причем среднее значение распределения равно 368 г. Измерения показали, что стандартное отклонение веса коробок равно 15 г. Допустим, что из многих тысяч коробок, заполненных за день, наугад выбираются 25 коробок и вычисляется их средний вес. Следует ли ожидать, что выборочный средний вес окажется равным

368 г? А может быть, он будет равен 200 г или 365 г?

Выборка является миниатюрной моделью генеральной совокупности, поэтому если исходная генеральная совокупность распределена нормально, выборка из нее должна быть приближенно нормальной. Следовательно, если мат. ожидание генеральной совокупности равно 368 г, выборочное среднее также должно быть близким к 368 г. Продолжая наши рассуждения, зададимся вопросом, как вычислить вероятность того, что выборочное среднее, полученное для выборки объемом $n = 25$, окажется меньше 365 г. Из свойств нормального распределения следует, что площадь, отсекаемая каждым значением случайной величины X от фигуры, ограниченной гауссовой кривой, можно вычислить, преобразовав стандартизованную нормальную случайную величину Z:

$$Z = (X - \mu)/\sigma$$

Для расчетов в Excel преобразование не требуется. Просто воспользуйтесь функцией =НОРМ.РАСП() (рис. 5).

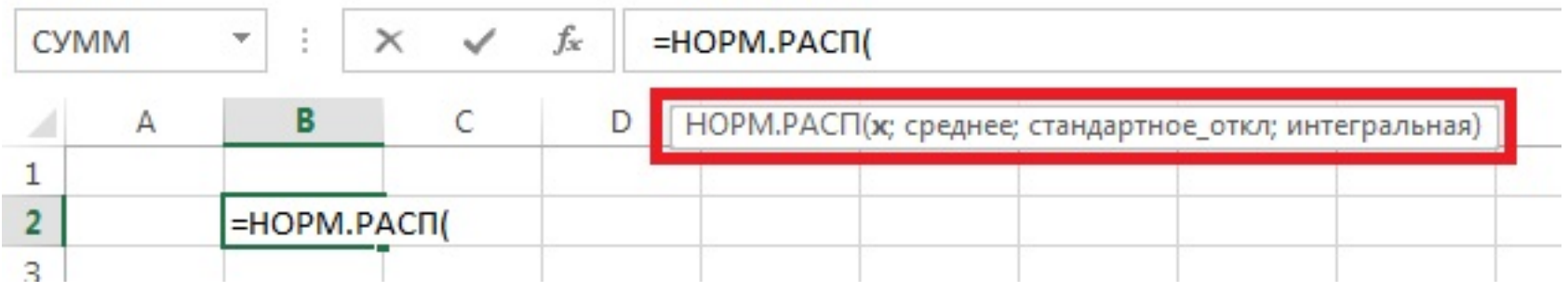


Рис. 5. Нормальная функция распределения; параметры: *x* – нормальная случайная величина, *μ* – среднее, *σ* – стандартное_откл, *интегральная* = ИСТИНА

Подставляя в приведенную выше формулу величину $\bar{X}$ вместо X, величину $\mu_{\bar{X}}$ вместо μ и величину $\sigma_{\bar{X}}$ вместо σ, получаем:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Обратите внимание на то, что благодаря несмещенности величина $\mu_{\bar{X}}$ всегда равна μ. Таким образом, значение величины Z, соответствующее вероятности того, что выборочное среднее, полученное для выборки объемом n = 25, окажется меньше 365 г, равна:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{365 - 368}{\frac{15}{\sqrt{25}}} = \frac{-3}{3} = -1{,}00$$

а вероятность, соответствующая значению Z = –1, равна P(Z=–1) = 0,1587

Следовательно, выборочное среднее 15,87% всех возможных выборок, имеющих объем n = 25, не превосходит 365 г. Это не значит, что вес 15,87% элементов выборок не превосходит 365 г. Долю таких элементов можно вычислить по следующей формуле:

$$Z = \frac{X - \mu}{\sigma} = \frac{365 - 368}{15} = \frac{-3}{15} = -0{,}20$$

а вероятность, соответствующая значению Z = –0,2, равна P(Z=–0,2) = 0,4207

Следовательно, в каждой выборке, имеющей объем n = 25, вес 42,07% коробок не превосходит 365 г. Это можно объяснить тем, что каждая выборка состоит из 25 разных значений, некоторые из которых велики, а некоторые — малы. Процедура усреднения ослабляет влияние отдельных элементов, особенно при увеличении объема выборки. Таким образом, вероятность того, что выборочное среднее, вычисленное по выборке, состоящей из 25 коробок, будет значительно отличаться от математического ожидания генеральной совокупности, меньше вероятности, что вес отдельных элементов значительно отличается от этого значения.

В Excel задачу можно решить с помощью одной формулы (рис. 6).

C5 | =НОРМ.РАСП(C1;C2;C3/КОРЕНЬ(C4);ИСТИНА)

	A	B	C	D
1	Выборочное среднее	$\bar{x}$	365	г
2	Математическое ожидание генеральной совокупности	μ	368	г
3	Стандартное отклонение генеральной совокупности	σ	15	г
4	Объем выборки	n	25	штук
5	Вероятность того, что в выборке из 25 коробок среднее выборочное будет меньше 365 г	$P(\bar{x} = 365)$	15,87%	

Рис. 6. Вероятность того, что выборочное среднее, полученное для выборки объемом *n* = 25, окажется меньше 365 г.

Иногда необходимо найти интервал, в котором лежит фиксированная часть элементов выборки или выборочных средних. В этом случае необходимо вычислить расстояние от математического ожидания генеральной совокупности, которому соответствует заданная площадь фигуры, ограниченной гауссовой кривой. Воспользуемся формулой

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Преобразовав ее получим, что величину $\bar{X}$ можно вычислить по формуле:

$$\bar{X} = \mu + Z\frac{\sigma}{\sqrt{n}}$$

Рассмотрим несколько примеров.

Пример 1. Влияние объема выборки n на стандартное отклонение выборочного среднего. Увеличим объем выборки с 25 до 100. Как изменится стандартное отклонение выборочного среднего?
Решение. Если *n* = 100, то

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1{,}5$$

Обратите внимание на то, что четырехкратное увеличение объема

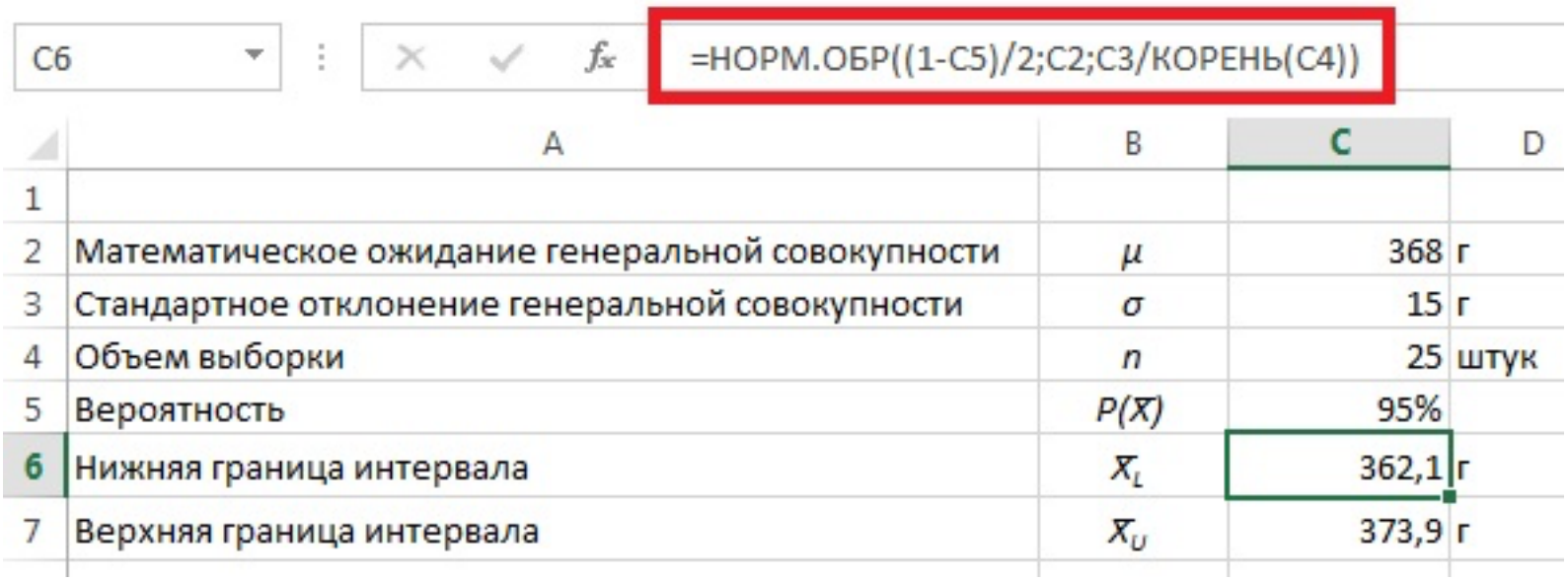
выборки приводит к уменьшению стандартного отклонения выборочного среднего вдвое — с 3 г до 1,5 г. Это значит, что, извлекая из генеральной совокупности выборки большего объема, мы обнаружим меньшую изменчивость выборочного среднего.

Пример 2. Влияние объема выборки n на концентрацию средних значений в выборочном распределении. Увеличим объем выборки с 25 до 100. Как изменится вероятность того, что выборочное среднее, полученное для выборки объемом *n* = 25, окажется меньше 365 г? Решение. Используя функцию =НОРМ.РАСП() получаем: P(n = 100) = 2,28% (рис. 7). Следовательно, в каждой выборке, имеющей объем *n* = 100, вес 2,28% коробок не превосходит 365 г. Напомним, что для выборок, имеющих объем *n* = 25, эта вероятность была равна 15,87%.

	A	B	C	D
1	Выборочное среднее	$\bar{x}$	365	г
2	Математическое ожидание генеральной совокупности	μ	368	г
3	Стандартное отклонение генеральной совокупности	σ	15	г
4	Объем выборки	n	100	штук
5	Вероятность того, что в выборке из 25 коробок среднее выборочное будет меньше 365 г	P($\bar{X}$ = 365)	2,28%	

Рис. 7. Вероятность того, что выборочное среднее, полученное для выборки объемом n = 100, окажется меньше 365 г.

Пример 3. Определение интервала, содержащего заданную часть выборочных средних. Найдите интервал, в котором лежат 95% всех выборочных средних, вычисленных по выборкам, состоящим из 25 коробок с кукурузными хлопьями. Решение. Интервал, содержащий 95% всех выборочных средних, вычисленных по выборкам, имеющим объем *n* = 25, делится на две равные части. Первая часть лежит слева от математического ожидания генеральной совокупности, а вторая — справа. Для расчета нижней $\bar{X}_L$ и верхней $\bar{X}_U$ границ интервалов воспользуемся функцией =НОРМ.ОБР(), возвращающая обратное нормальное распределение (рис. 8).



	A	B	C	D
1				
2	Математическое ожидание генеральной совокупности	μ	368	г
3	Стандартное отклонение генеральной совокупности	σ	15	г
4	Объем выборки	n	25	штук
5	Вероятность	$P(X)$	95%	
6	Нижняя граница интервала	$\bar{X}_L$	362,1	г
7	Верхняя граница интервала	$\bar{X}_U$	373,9	г

Рис. 8. Функция =НОРМ.ОБР(), возвращающая обратное нормальное распределение

Следовательно, 95% всех выборочных средних, вычисленных по выборкам, имеющим объем n = 25, лежат в интервале от 362,1 г до 373,9 г.

Выборки из генеральных совокупностей, распределения которых отличаются от нормального. До сих пор мы рассматривали выборочное распределение средних для нормально распределенной генеральной совокупности. Однако во многих ситуациях распределение генеральной совокупности либо неизвестно, либо заведомо отличается от нормального. Таким образом, следует рассмотреть выборочное распределение средних для генеральной совокупности, распределение которой отличается от нормального. Этот анализ приводит нас к основной теореме статистики:

Центральная предельная теорема утверждает, что при достаточно большом объеме выборок выборочное распределение средних можно аппроксимировать нормальным распределением. Это свойство не зависит от вида распределения генеральной совокупности.

Какой объем выборок следует считать «достаточно большим»? Этот вопрос изучался во многих статистических исследованиях. Как правило, для подавляющего большинства генеральных

совокупностей выборочное распределение средних становится приближенно нормальным при n = 30. Однако, если известно, что распределение генеральной совокупности является колоколообразным, эту теорему можно применять и для меньшего объема выборок. Если же распределение генеральной совокупности обладает сильной асимметрией или имеет несколько мод, объем выборок следует увеличить.

Применение центральной предельной теоремы к различным генеральным совокупностям проиллюстрировано на рис. 9.

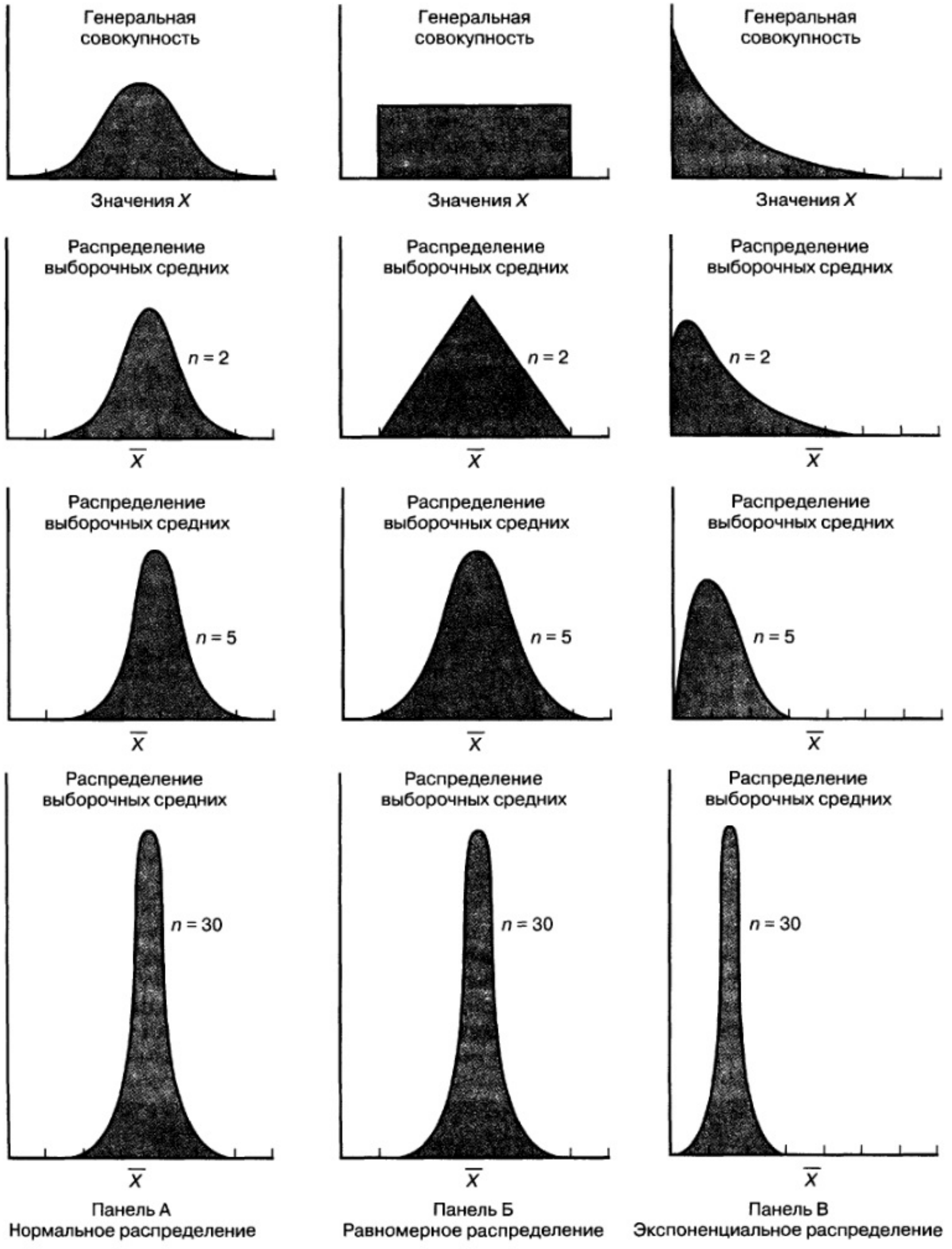


Рис. 9. Выборочное распределение средних для разных генеральных совокупностей при объемах выборок n = 2, 5 и 30

Панель А: выборочное распределение средних, построенное для генеральной совокупности, имеющей нормальное распределение.

Как указывалось выше, если генеральная совокупность является нормально распределенной, выборочное распределение средних также является нормальным, независимо от объема выборок. При увеличении объема выборок изменчивость выборочных средних уменьшается. Панель Б: выборочное распределение средних, построенное для генеральной совокупности, имеющей равномерное распределение. При n = 5 выборочное распределение средних является приближенно нормальным. При n = 30 выборочное распределение средних становится практически нормальным. Панель В: выборочное распределение средних, построенное для генеральной совокупности, имеющей экспоненциальное распределение. Это распределение имеет ярко выраженную положительную асимметрию. При n = 2 асимметрия выборочного распределения средних сохраняется, но выражена слабее. При n = 5 выборочное распределение средних становится почти симметричным со слабой положительной асимметрией. При n = 30 выборочное распределение средних становится приближенно нормальным. В любом случае среднее выборочных средних всегда совпадает с математическим ожиданием генеральной совокупности, а его изменчивость при увеличении объема выборок уменьшается.

Свойства выборочного распределения средних:

- Если объем выборок превышает 30, выборочное распределение средних для большинства генеральных совокупностей является приближенно нормальным.
- Если генеральная совокупность распределена симметрично, выборочное распределение средних становится приближенно нормальным уже при n = 15.
- Если генеральная совокупность является нормально распределенной, выборочное распределение средних является нормальным при любом объеме выборок.

Генерирование выборок в Excel

Чтобы создать массив случайных нормально распределенных чисел (например, стандартизованных, т.е. имеющих $\mu = 0$ и $\sigma = 1$), подставим в функцию =НОРМ.СТ.ОБР(вероятность) в качестве параметра *вероятность* генератор случайных чисел СЛЧИС() (рис. 10). Этот генератор создает случайные числа в диапазоне от 0 (включая) до 1 (не включая). Случайные числа расположились в столбце *A*. Чтобы совокупность была «внушительной» формулой заполнены 1000 строк (при каждом изменении на листе формулы пересчитываются; кроме того, можно принудительно пересчитать формулы, нажав F9). Глядя на числа в столбце *A* трудно определить, что они нормально распределены. Чтобы эта закономерность стала видна, я создал сводную таблицу, а затем сгруппировал данные в столбце *C* по диапазонам, от –3 до 3 с шагом 0,4. В столбце *D* я отразил количество чисел, попадающих в диапазон (если ранее вы не сталкивались с такими настройками сводной таблицы, рекомендую почитать Изменение настраиваемого вычисления для поля в отчете сводных таблиц).

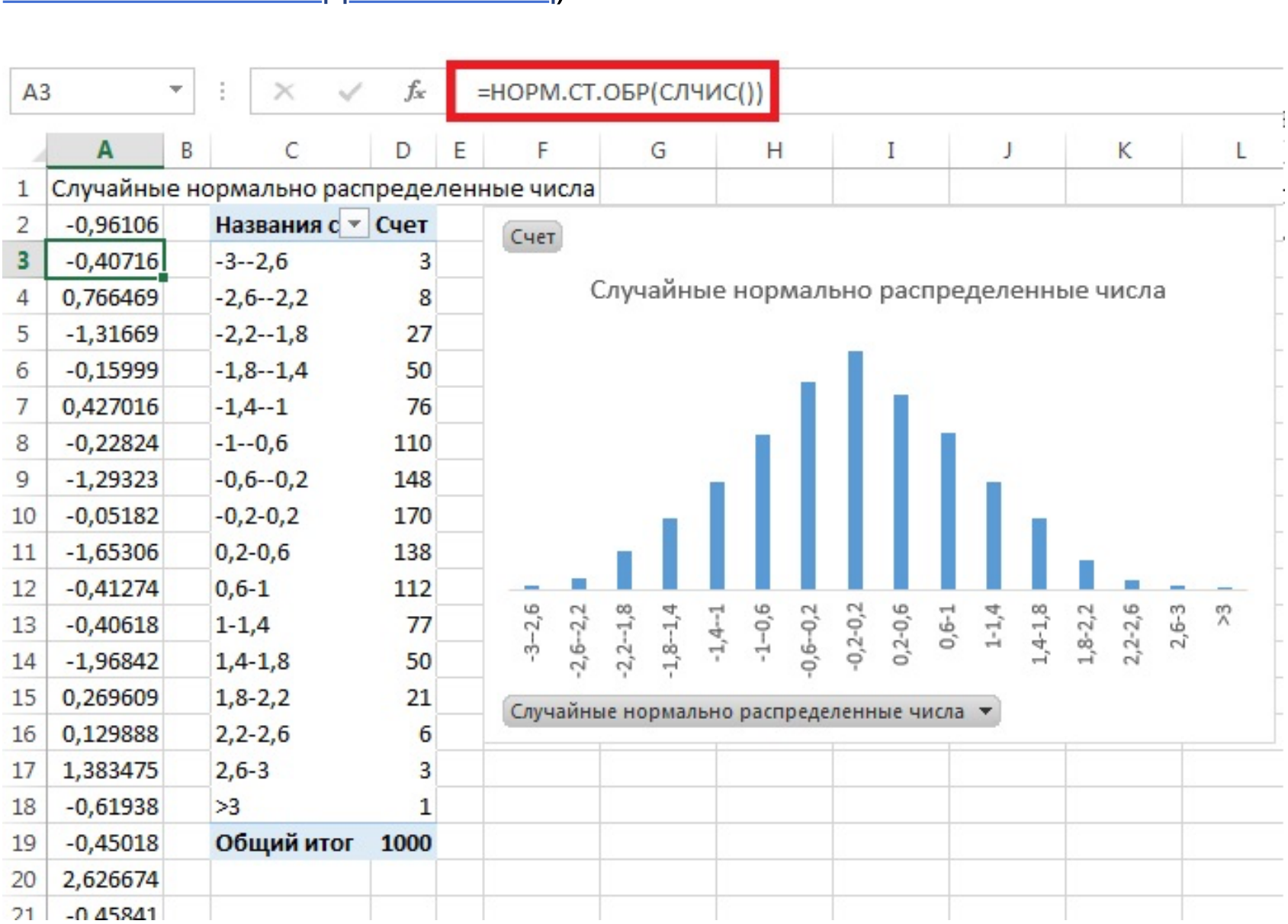


Рис. 10. Сгенерированный массив случайных нормально

распределенных чисел (μ = 0 и σ = 1)

Аналогичного результата можно добиться, с помощью надстройки *Пакет анализа*. Выберите закладку *Данные* → область *Анализ* → *Анализ данных* → *Генератор случайных чисел*. (Если у вас не установлена надстройка *Пакет анализа* см. описание после рис. 5 заметки Представление числовых данных в виде таблиц и диаграмм.) Параметры открывшегося окна *Генератор случайных чисел* изменяются в зависимости от выбранного типа распределения. В поле *Число переменных* указывается необходимое количество столбцов, а в ноле *Число случайных* чисел — необходимое количество строк. Например, если нужно получить 200 случайных чисел, расположенных в 10 столбцах и 20 строках, введите в эти поля соответственно числа 10 и 20.

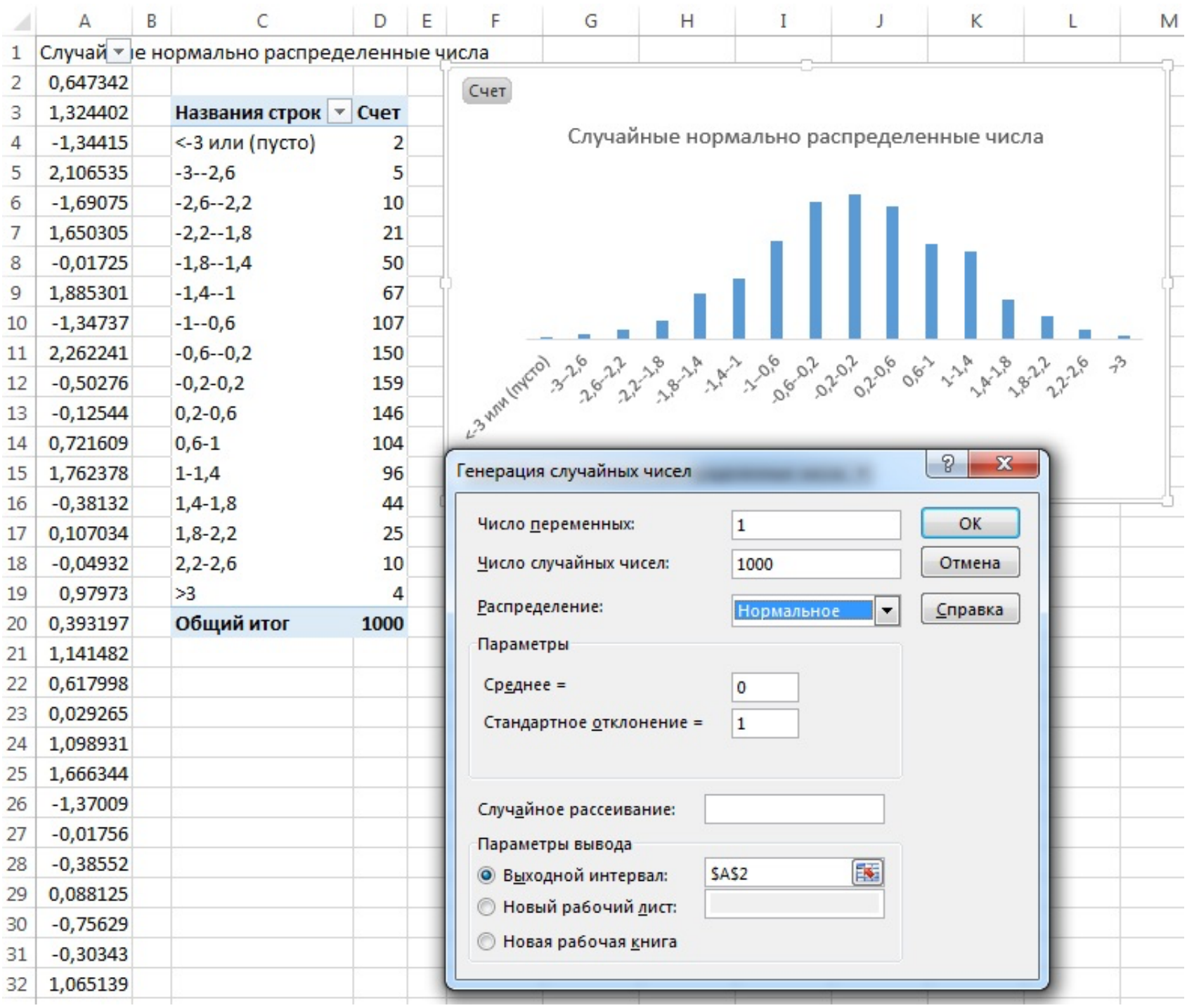


Рис. 11. Генерирование случайных чисел с помощью надстройки *Пакет анализа*

Поле *Случайное рассеивание* позволяет задать начальное значение, которое будет использовано программой в алгоритме генерации случайных чисел. Обычно это поле оставляют пустым. Однако, если необходимо генерировать одинаковые последовательности случайных чисел, задайте рассеивание в диапазоне от 1 до 32 767 (допускаются только целые числа). Из раскрывающегося списка *Распределение* можно выбрать одну из перечисленных ниже опций.

- *Равномерное*. Генерируется последовательность равномерно распределенных случайных чисел в заданном интервале. Необходимо указать верхнюю и нижнюю границы интервала.
- *Нормальное*. Генерируется последовательность случайных чисел, соответствующих нормальному распределению. Задается среднее значение и стандартное отклонение.
- *Бернулли*. Генерируется последовательность случайных чисел, принимающих только значение 0 или 1, в зависимости от заданной вероятности успеха.
- *Биномиальное*. Генерируется последовательность случайных чисел, соответствующих распределению Бернулли для некоторого числа попыток, с заданной вероятностью успеха.
- *Пуассона*. Генерируется последовательность случайных чисел, соответствующих распределению Пуассона. Это распределение характеризует дискретные события, произошедшие в интервале времени, где вероятность одного события пропорциональна размеру интервала. Параметр Лямбда — это ожидаемое количество событий в интервале. В распределении Пуассона Лямбда равняется среднему, которое совпадает с дисперсией. Подробнее см. Распределение Пуассона
- *Модельное*. Эта опция на самом деле не генерирует случайных чисел. Вместо этого она повторяет последовательность чисел в заданном порядке.

- *Дискретное*. Эта опция позволяет определить вероятность, характеризующую выбираемые значения. Для нее требуется входной диапазон, состоящий из двух столбцов: в первом столбце содержатся значения, а во втором — вероятности каждого значения. Сумма вероятностей во втором столбце должна равняться 1.

Полученное в результате работы *Пакета анализа* случайные числа в диапазоне A2:A1001, я проанализировал также как и выше с помощью сводной таблицы и гистограммы. Надо отметить, что качество сгененрированных чисел мне не понравилось. В частности, за пределами диапазона ±3σ наблюдалось 6 чисел (теоретически их должно быть порядка трех), а одно даже приняло значение
–9,5, что уж совсем невероятно...

Выборочное распределение долей

При анализе категорийных данных, принимающих одно из двух значений — мужчина или женщина, любит / не любит и т.д., — результаты часто обозначают единицами (да) и нулями (нет). Среднее значение, вычисленное по выборке, состоящей из *n* таких элементов, равно количеству единиц, деленному на *n*. Например, из пяти респондентов три человека предпочитают торговую марку А, а двое — торговую марку Б. Следовательно, выборка состоит из трех единиц и двух нулей. Суммируя элементы выборки и деля сумму на пять, получаем, что доля поклонников торговой марки А в данной выборке равна 0,6. Таким образом, для категорийных данных выборочное среднее нулей и единиц представляет собой выборочную долю p_S некоторой характеристики, которой обладают элементы выборки.

Выборочная доля признака:

p_S = X / n = количество объектов, имеющих указанную характеристику / размер выборки

Выборочная доля признака p_S имеет особое свойство: она принимает значения от 0 до 1. Если все элементы выборки обладают одинаковыми характеристиками, то каждому из них присваивается единица, а выборочная доля признака также становится равной единице. Если только половина элементов выборки обладает интересующим нас свойством, им приписываются единицы, а остальные обозначаются нулями. В этом случае выборочная доля признака p_S равна 0,5. Если ни один элемент выборки не обладает интересующим нас свойством, им приписываются нули. В этом случае выборочная доля признака p_S равна нулю.

В то время как выборочное среднее является несмещенной оценкой математического ожидания генеральной совокупности статистика p_S является несмещенной оценкой доли признака p в генеральной совокупности. По аналогии с распределением выборочных средних можно ввести стандартную ошибку доли признака:

$$\sigma_{p_S} = \sqrt{\frac{p(1-p)}{n}}$$

Если выборка извлекается из конечной генеральной совокупности без возвращения, выборочное распределение доли признака подчиняется биномиальному закону (см. Биноминальное распределение). Однако, если значения np и $n(1-p)$ больше 4, это распределение можно аппроксимировать нормальным. При статистическом анализе долей признака объем выборки играет очень важную роль. Следовательно, во многих ситуациях для оценки выборочного распределения доли признака можно использовать нормальное распределение. Таким образом, в формуле

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

величину $\bar{X}$ можно заменить величиной p_S, величину μ — величиной p, а величину $\sigma/\sqrt{n}$ – величиной

$$\sqrt{\frac{p(1-p)}{n}}$$

Разность между выборочной долей признака и долей признака в генеральной совокупности:

$$Z = \frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}}$$

Проиллюстрируем выборочное распределение доли признака следующим примером. Предположим, что менеджер местного отделения банка выяснил, что 40% всех вкладчиков имеют в банке несколько счетов. Если создать выборку из 200 вкладчиков, то можно вычислить вероятность того, что выборочная доля вкладчиков, имеющих несколько счетов, не превосходит 0,3. Поскольку np = 200*0,4 = 80 > 5 и $n(1-p)$ = 200×0,6 = 120 > 5, выборочное распределение доли вкладчиков практически совпадает с нормальным. Применим только что полученную формулу:

$$Z = \frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0,30 - 0,40}{\sqrt{\frac{0,40 \times 0,60}{200}}} = \frac{-0,10}{\sqrt{\frac{0,24}{200}}} = \frac{-0,10}{0,0346} = -2,89$$

Значению Z = –2,89 соответствует $p(Z)$ = 0,0019. Следовательно, вероятность того, что доля вкладчиков, имеющих несколько счетов, не превосходит 0,3, равна 0,19%, т.е. крайне мала.

Выборки из конечных генеральных совокупностей

Центральная предельная теорема, а также формулы для вычисления стандартной ошибки среднего и стандартной ошибки доли признака основаны на предположении, что выборки извлекаются из генеральной совокупности с возвращением. Однако практически во всех статистических исследованиях выборки извлекаются из генеральных совокупностей конечного объема N без возвращения. Если объем выборок n достаточно велик по сравнению с объемом

генеральной совокупности *N* (т.е. выборка содержит более 5% элементов генеральной совокупности), так что *n/N* > 0,05, то при вычислении стандартной ошибки среднего и стандартной ошибки доли признака следует учитывать *поправочный коэффициент для конечной генеральной совокупности* (fpc — finite population correction factor). Эта поправка вычисляется по формуле:

$$fpc = \sqrt{\frac{N-n}{N-1}}$$

где *n* — объем выборки, а *N* — объем генеральной совокупности.

Таким образом, формулы для вычисления стандартной ошибки среднего и стандартной ошибки доли признака принимают следующий вид:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$\sigma_{p_i} = \sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}$$

Анализ формулы для вычисления поправочного коэффициента для конечной генеральной совокупности (6.19) показывает, что ее числитель всегда меньше знаменателя, поскольку число *n* всегда больше единицы. Следовательно, поправочный коэффициент для конечной генеральной совокупности меньше единицы. Поскольку этот коэффициент умножается на стандартную ошибку, скорректированная стандартная ошибка уменьшается. Таким образом, с учетом поправочного коэффициента для конечной генеральной совокупности мы получаем более точные оценки.

Например, предположим, что банк обслуживает 1000 клиентов, причем 400 из них имеют больше одного счета. Используя поправочный коэффициент для конечной генеральной совокупности, определите вероятность извлечь выборку, состоящую из 200 клиентов, в которой доля клиентов, имеющих несколько

банковских счетов меньше 0,30.

Решение. Разность между выборочной долей признака и долей признака в генеральной совокупности (при n = 200):

$$Z = \frac{p_s - p}{\sqrt{\frac{p(1-p)}{n}}\sqrt{\frac{N-n}{N-1}}} = \frac{0,30-0,40}{\sqrt{\frac{0,40\times0,60}{200}}\sqrt{\frac{1\,000-200}{1\,000-1}}} = -3,23$$

Значению Z = –3,23 соответствует вероятность $p(Z)$ = 0,00062. Следовательно, вероятность того, что доля вкладчиков, имеющих несколько счетов, не превосходит 0,3, равна 0,06%. Учет поправочного коэффициента для конечной генеральной совокупности втрое уменьшил вероятность (см. аналогичный пример выше, где $p(Z)$ = 0,19%).

Предыдущая заметка Равномерное и экспоненциальное распределения

Следующая заметка Построение доверительного интервала для математического ожидания генеральной совокупности

К оглавлению Статистика для менеджеров с использованием Microsoft Excel

[1] Используются материалы книги Левин и др. Статистика для менеджеров. – М.: Вильямс, 2004. – с. 385–415