Статистические оценки

Алла Тамбовцева

24 апреля 2017

Статистическая оценка как случайная величина

Одна из важных, но не совсем тривиальных задач при знакомстве со статистическим оцениванием – осознать то, что статистическая оценка является случайной величиной, которая имеет свое распределение, математическое ожидание и дисперсию. Могут возникнуть два логичных вопроса: откуда берется случайность и почему у оценки есть свое распределение, если оценка – это просто одно число? Рассмотрим пример.

Представим, что нас интересует средний доход жителей Москвы в месяц. Обычно в опросах респондентам не предлагается указывать доход в виде конкретного числа, вместо этого в качестве вариантов ответа на вопрос используются интервалы или ответы вида "покупка большинства товаров длительного пользования (холодильник, телевизор) не вызывает трудностей, однако купить квартиру мы не можем", но предположим на время, что мы не очень сознательные исследователи.

Итак, у нас есть генеральная совокупность – все население Москвы. Конечно, опросить всех жителей и узнать их доход у нас не получится. Но мы можем взять выборку из этой совокупности – случайным образом выбрать 1000 жителей Москвы – и посчитать среднее арифметическое доходов людей в этой выборке. Полученное среднее арифметическое будет оценкой среднего дохода всех жителей Москвы.

Полученная оценка среднего – это единственное число. Почему это число случайно и почему у него должно быть распределение? На самом деле, когда речь идет о распределении статистической оценки, имеется в виду не конкретное значение, посчитанное по выборке, а целая совокупность таких значений. Наша выборка из 1000 человек – это всего лишь одна из всех возможных выборок, которые мы можем взять из населения Москвы. Мы взяли одну выборку из 1000 человек, другие исследователи взяли другую выборку из 1000 человек, третьи – третью и так далее. Так как доход людей во всех выборках разный, среднее по выборке у всех исследователей тоже будет получаться разным.

```
# среднее выборки 1, среднее выборки 2, среднее выборки 3
```

```
## [1] 385597.9 398932.9 405905.2
```

Если мы возьмем все возможные выборки одинакового размера n из генеральной совокупности (k выборок) и посчитаем среднее значение каждой выборки, то получим набор из k средних значений:

$$\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$$

Этот набор средних значений имеет свое распределение, математическое ожидание и дисперсию. Математическое ожидание оценки среднего – это среднее ожидаемое значение средних арифметических, посчитанных по всем возможным выборкам из генеральной совокупности, а дисперсия оценки среднего – дисперсия таких средних арифметических.

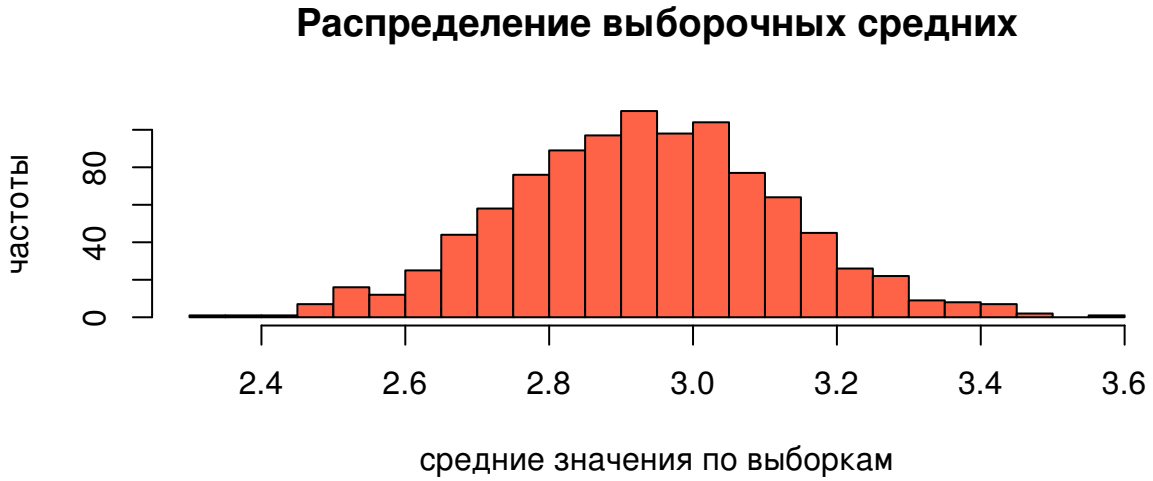
Отвлечемся от примера со средним доходом жителей Москвы и посмотрим на новую генеральную совокупность – набор из 1000 каких-то значений, которые распределены нормально со средним значением 3 и стандартным отклонением 2.

Возьмем и "надергаем" из нашей генеральной совокупности 1000 разных выборок по 100 наблюдений и посчитаем среднее арифметическое по каждой выборке (все возможные выборки брать не будем, их слишком много). Получим набор из 1000 средних значений по выборкам.

```
# первые несколько значений
```

[1] 2.82 3.30 2.96 3.09 3.09 2.85

Можем построить гистограмму – график, который будет показывать, сколько каких значений встречается в нашем наборе:



Как распределены средние значения по выборкам, мы увидели: самые часто встречающиеся значения находятся в окрестности 3, самые редкие – в окрестностях 2.4 и 3.6. А какое у них среднее значение и стандартное отклонение? Посчитаем.

```
# среднее значение и стандартное отклонение
```

[1] 2.93 0.19

Видно, что среднее значение средних арифметических по выборкам примерно равно среднему генеральной совокупности (3). Со стандартным отклонением все менее очевидно, но ясно, что полученные значения возникли неслучайно. Итак, верно следующее:

Если у нас есть генеральная совокупность со средним значением μ и стандартным отклонением σ, то после того, как мы извлечем из генеральной совокупности все возможные выборки размера n и посчитаем по ним средние значения, эти средние значения будут иметь распределение со средним значением μ и стандартным отклонением примерно $\frac{\sigma}{\sqrt{n}}$.

Проверим. Стандартное отклонение нашей генеральной совокупности было равно 2, брали выборки по 100 наблюдений. По формуле $\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{100}}$ получаем значение 0.2. А стандартное отклонение средних, полученное выше, равно 0.19. Значения примерно совпадают.

Может возникнуть вопрос: а как этот факт можно использовать на практике? При проведении исследований мы же не можем позволить себе взять все возможные выборки и посчитать по ним средние значения, выборка у нас, как правило, одна. На самом деле знание о среднем и стандартном отклонении распределения статистической оценки дает нам две важные практические вещи:

1) позволяет посчитать доверительный интервал для интересующего параметра (обсудим позже)

2) дает представление о соотношении точности оценки и размера выборки

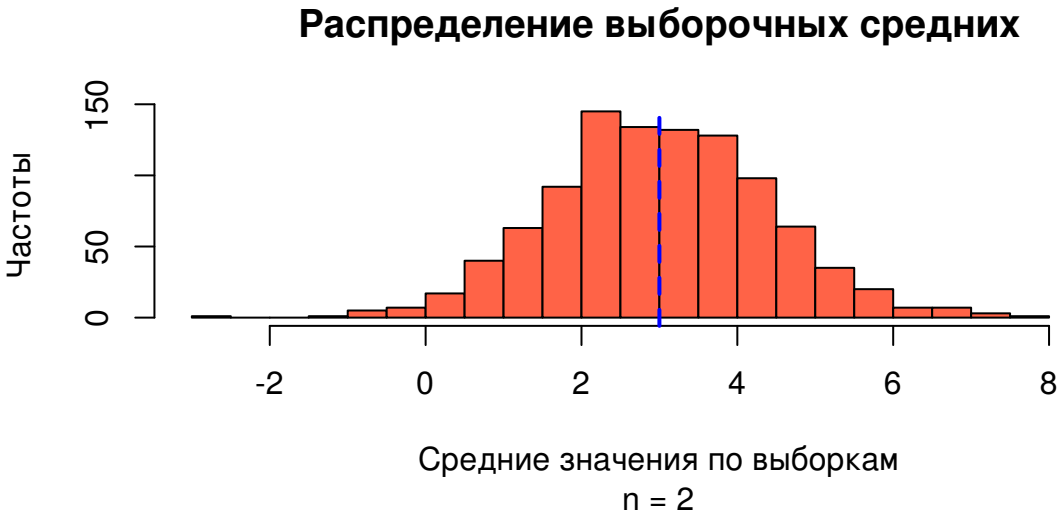
Мы знаем, что стандартное отклонение отвечает за то, насколько сильно значения разбросаны относительно среднего. Если мы хотим получить более точную оценку, то мы хотим, чтобы разброс ее возможных значений был небольшим, то есть, чтобы стандартное отклонение оценки было маленьким. Как этого добиться в случае среднего? Стандартное отклонение среднего равно $\frac{\sigma}{\sqrt{n}}$, причем σ – стандартное отклонение генеральной совокупности – не изменяется (оно одно), а размер выборки n мы можем менять. Получается, чтобы стандартное отклонение среднего было достаточно маленьким, нужно взять достаточно большую выборку.

Закон больших чисел

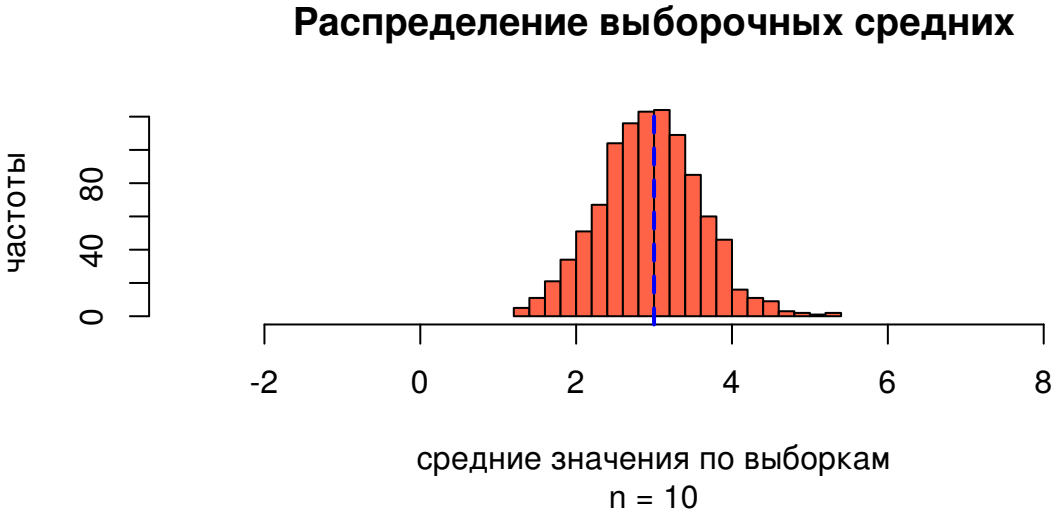
Согласно закону больших чисел, с увеличением размера выборки, среднее значение выборки становится ближе к среднему значению генеральной совокупности. Другими словами, чтобы получить более точную оценку среднего генеральной совокупности, нужно взять достаточно большую выборку.

Понаблюдаем за действием закона больших чисел на примере. Из генеральной совокупности, описанной выше (нормально распределенной со средним 3 и стандартным отклонением 2) будем брать выборки разного размера и смотреть, насколько сильно средние значения выборок сконцентрированы относительно среднего генерального совокупности.

Для начала возьмем 1000 выборок размера 2, посчитаем по ним средние значения и построем для них гистограмму:



Видно, что значения достаточно сильно разбросаны относительно среднего генеральной совокупности (отмечено пунктирной линией). А теперь возьмем и проделаем то же самое для выборок размера 10:



Заметно, что в данном случае средние значения уже более сконцентрированы вокруг среднего генеральной совокупности, график “более узкий”. Попробуем взять выборки большего размера, по 400 наблюдений:

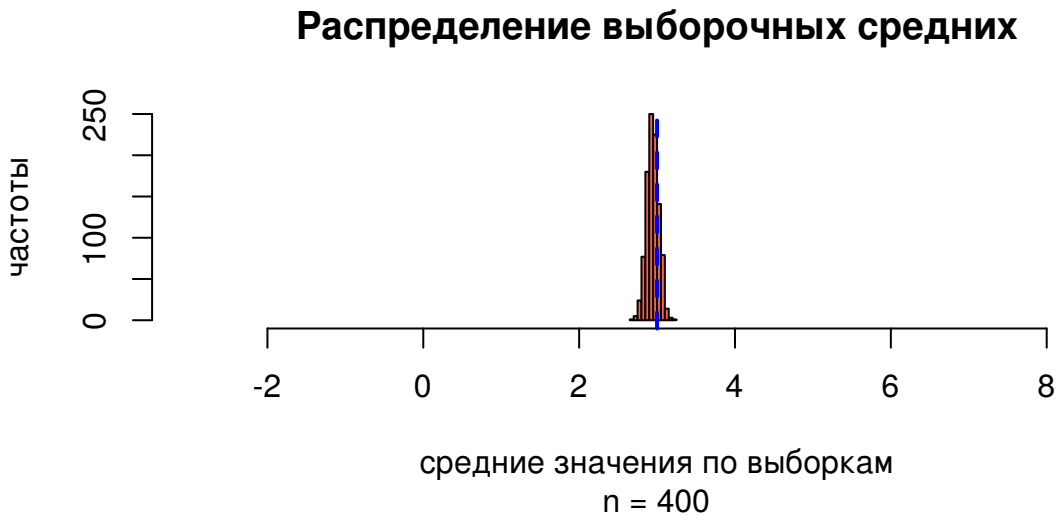
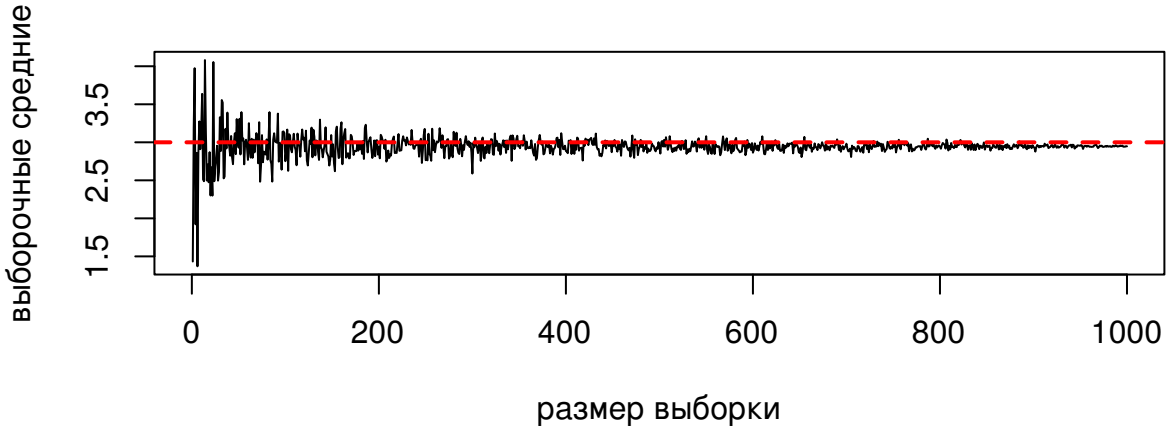


График стал "совсем узким", разброс средних значений по выборкам стал очень маленьким, то есть мы получили достаточно точные оценки среднего значения генеральной совокупности.

Можем проиллюстрировать то же самое несколько иначе. Отметим на графике среднее значение генеральной совокупности (красная пунктирная линия), а затем будем отмечать точками среднее значение выборок разного размера и соединять их линиями.



По такому графику тоже хорошо видно, что с ростом размера выборки, выборочное среднее все ближе становится к среднему генеральной совокупности.

Как связана точность оценки и размер выборки, мы разобрались. Теперь остался один вопрос: когда мы обсуждали распределение оценки среднего генеральной совокупности, мы зафиксировали, какое среднее и стандартное отклонение имеет эта оценка, но ничего не сказали о том, какое распределение она имеет. Ответ на этот вопрос нам может дать центральная предельная теорема.

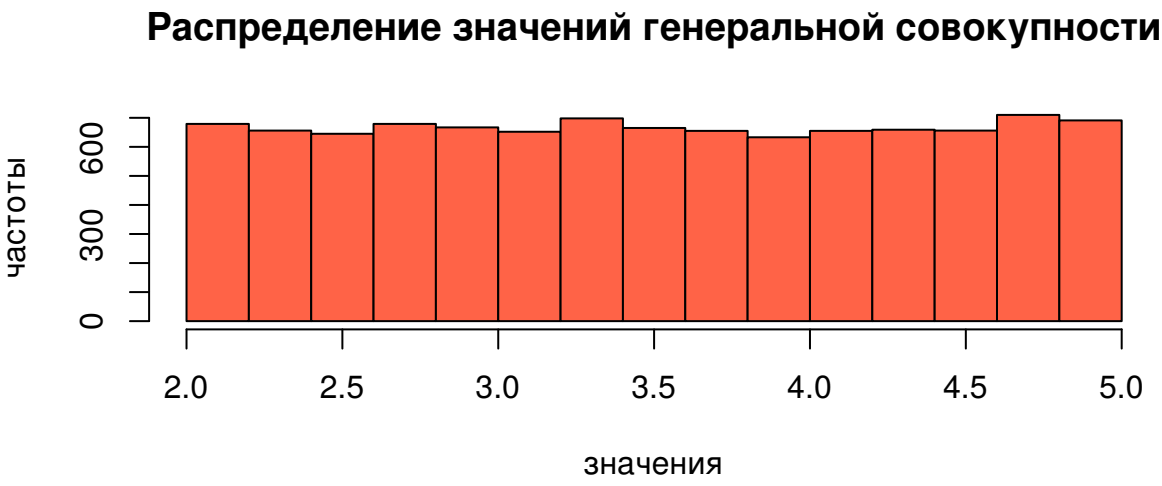
Центральная предельная теорема

Пусть у нас есть генеральная совокупность со средним значением μ и стандартным отклонением σ. Если мы извлечем из этой совокупности все возможные выборки достаточно большого размера n ($n \geq 30$) и посчитаем по каждой выборке среднее значение, то эти средние значения будут примерно нормально распределены со средним μ и стандартным отклонением $\frac{\sigma}{\sqrt{n}}$.

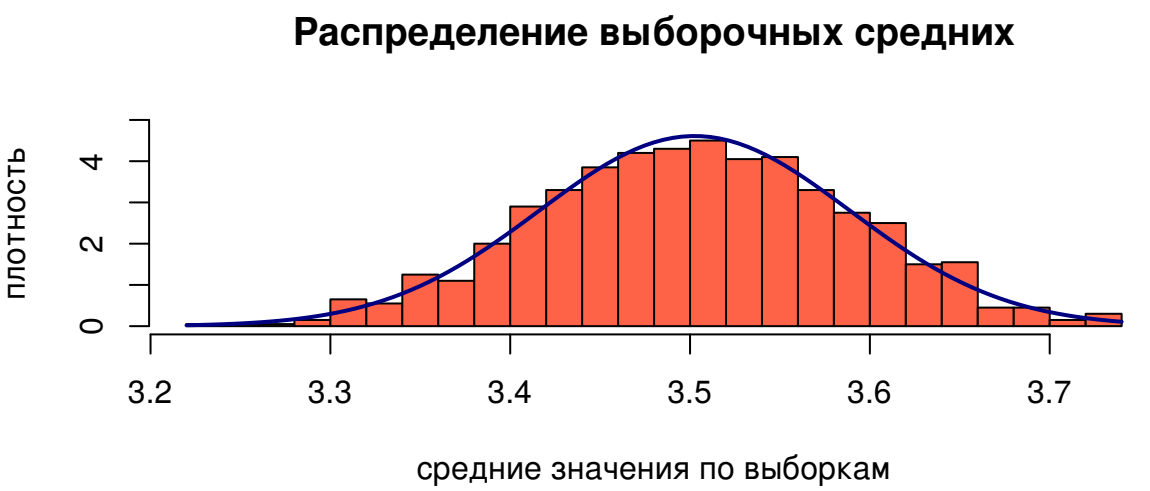
Итак, благодаря центральной предельной теореме, мы знаем, что распределение средних выборочных значений является нормальным. Важно то, что это выполняется вне зависимости от того, какое

распределение имеет генеральная совокупность. Даже если генеральная совокупность имеет какое-то неизвестное распределение, если мы возьмем много достаточно больших выборок из него и посчитаем их средние значения, эти средние будут распределены нормально. Рассмотрим пример.

Возьмем генеральную совокупность из 1000 элементов, имеющую равномерное распределение с минимальным значением 2 и максимальным значением 5. Среднее значение этого распределения примерно равно 3.5, стандартное отклонение примерно равно 0.87. Посмотрим на распределение значений генеральной совокупности:



По гистограмме видно, что значения генеральной совокупности распределены равномерно. Теперь возьмем из этого распределения 1000 выборок по 100 наблюдений, посчитаем средние значения по каждой выборке и посмотрим на их распределение:



Видно, что распределение выборочных средних уже не равномерное, а примерно нормальное. О чем нам и говорит центральная предельная теорема. В заключение проверим, чему равно среднее значение и стандартное отклонение выборочных средних.

```
# среднее и стандартное отклонение
```

```
## [1] 3.50 0.09
```

Полученное среднее значение совпадает со средним генеральной совокупности (3.5), а полученное стандартное отклонение примерно равно $\frac{0.87}{\sqrt{100}} = 0.087$.