


ОСНОВЫ СТАТИСТИКИ

Инференциальная (параметрическая)
статистика. Понятие выборочной
совокупности и оценка





Здравствуйте!

В ходе изучения данного материала вы ознакомитесь с материалом, который позволит вам:

1. Различать виды выборок и их объем.
2. Строить выборочные распределения.
3. Применять центральную предельную теорему при проведении исследований.
4. Применять основные обозначения, для изучения дополнительной литературы в области статистики

Статистический анализ и оценка, является фундаментальным исследованием. Оценка может быть использована в следующих областях:

- Изменения в ценностях и отношениях (например, поддержка смертной казни или мнения о гей-браке) в Соединенных Штатах за эти годы.
- Уровни счастья и благополучия в странах по всему миру.
- Эффективность лекарств или других методов лечения заболеваний.
- Обращение кандидатов на политические должности в различных слоях голосующего населения (например, среди женщин).
- Немедленная реакция общественности на спорные вопросы, такие как новые законы об иммиграции или здравоохранении.

Одной из целей исследований в области социальных наук является проверка наших теорий с использованием множества разных людей, групп, обществ и исторических эпох. Однако основная проблема в исследованиях в области социальных наук заключается в том, что наиболее подходящие группы населения для наших тестов очень большие. Как мы можем проверить наши теории, если мы не можем получить доступ ко всему населению?

Введем ряд понятий

Выборка или выборочная совокупность — множество случаев (испытуемых, объектов, событий, образцов), с помощью определённой процедуры выбранных из генеральной совокупности для участия в исследовании.

Характеристики выборки:

- Качественная характеристика выборки – кого именно мы выбираем и какие способы построения выборки мы для этого используем.
- Количественная характеристика выборки – сколько случаев выбираем, другими словами объём выборки.

Необходимость выборки:

1. Объект исследования очень обширный. Например, потребители продукции глобальной компании – огромное количество территориально разбросанных рынков.
2. Существует необходимость в сборе первичной информации.

Объём выборки – число случаев, включённых в выборочную совокупность. Из статистических соображений рекомендуется, чтобы число случаев составляло не менее 30–35.

Выборка может рассматриваться в качестве репрезентативной (Репрезентативность определяет, насколько возможно обобщать результаты исследования с привлечением определённой выборки на всю генеральную совокупность, из которой она была собрана) или нерепрезентативной.



Выборки делятся на два типа:

- вероятностные
- невероятностные

Простая вероятностная выборка. Использование такой выборки основывается на предположении, что каждый респондент с равной долей вероятности может попасть в выборку. На основе списка генеральной совокупности составляются карточки с номерами респондентов. Они помещаются в колоду, перемешиваются и из них наугад вынимается карточка, записывается номер, потом возвращается обратно. Далее процедура повторяется столько раз, какой объем выборки нам необходим. Минус: повторение единиц отбора.

Процедура построения простой случайной выборки включает в себя следующие шаги:

- 1) необходимо получить полный список членов генеральной совокупности и пронумеровать этот список. Такой список, напомним, называется основой выборки;
- 2) определить предполагаемый объем выборки, то есть ожидаемое число опрошенных;
- 3) извлечь из таблицы случайных чисел столько чисел, сколько нам требуется выборочных единиц. Если в выборке должно оказаться 100 человек, из таблицы берут 100 случайных чисел. Эти случайные числа могут генерироваться компьютерной программой.
- 4) выбрать из списка-основы те наблюдения, номера которых соответствуют выписанным случайным числам

Простая случайная выборка имеет очевидные преимущества. Этот метод крайне прост для понимания. Результаты исследования можно распространять на изучаемую совокупность. Большинство подходов к получению статистических выводов предусматривают сбор информации с помощью простой случайной выборки. Однако метод простой случайной выборки имеет как минимум четыре существенных ограничения:

- 1) зачастую сложно создать основу выборочного наблюдения, которая позволила бы провести простую случайную выборку.
- 2) результатом применения простой случайной выборки может стать большая совокупность, либо совокупность, распределенная по большой географической территории, что значительно увеличивает время и стоимость сбора данных.
- 3) результаты применения простой случайной выборки часто характеризуются низкой точностью и большей стандартной ошибкой, чем результаты применения других вероятностных методов.

Хотя выборки, полученные простым случайным отбором, в среднем адекватно представляют генеральную совокупность, некоторые из них крайне некорректно представляют изучаемую совокупность. Вероятность этого особенно велика при небольшом объеме выборки.

1. Простая бесповторная выборка. Процедура построения выборки такая же, только карточки с номерами респондентов не возвращаются обратно в колоду.
2. Систематическая вероятностная выборка. Является упрощенным вариантом простой вероятностной выборки. На основе списка генеральной совокупности через определённый интервал (K) отбираются респонденты. Величина K определяется случайно. Наиболее достоверный результат достигается при однородной генеральной совокупности, иначе возможны совпадение величины шага и каких-то внутренних циклических закономерностей выборки (смещение выборки). Минусы: такие же как и в простой вероятностной выборке.
3. Серийная (гнездовая) выборка. Единицы отбора представляют собой статистические серии (семья, школа, бригада и т. п.). Отобранные элементы подвергаются сплошному обследованию. Отбор статистических единиц может быть организован по типу случайной или систематической выборки. Минус: Возможность большей однородности, чем в генеральной совокупности.



4. Районированная выборка. В случае неоднородной генеральной совокупности, прежде, чем использовать вероятностную выборку с любой техникой отбора, рекомендуется разделить генеральную совокупность на однородные части, такая выборка называется районированной. Группами районирования могут выступать как естественные образования (например, районы города), так и любой признак, заложенный в основу исследования. Признак, на основе которого осуществляется разделение, называется признаком расслоения и районирования.

5. «Удобная» выборка. Процедура «удобной» выборки состоит в установлении контактов с «удобными» единицами выборки — с группой студентов, спортивной командой, с друзьями и соседями. Если необходимо получить информацию о реакции людей на новую концепцию, такая выборка вполне обоснована. «Удобную» выборку часто используют для предварительного тестирования анкет

Невероятностная выборка. Отбор в такой выборке осуществляется не по принципам случайности, а по субъективным критериям – доступности, типичности, равного представительства и т.д.

1. Квотная выборка – выборка строится как модель, которая воспроизводит структуру генеральной совокупности в виде квот (пропорций) изучаемых признаков. Число элементов выборки с различным сочетанием изучаемых признаков определяется с таким расчётом, чтобы оно соответствовало их доле (пропорции) в генеральной совокупности. Так, например, если генеральная совокупность у нас представлена 5000 человек, из них 2000 женщин и 3000 мужчин, тогда в квотной выборке у нас будут 20 женщин и 30 мужчин, либо 200 женщин и 300 мужчин. Квотированные выборки чаще всего основываются на демографических критериях: пол, возраст, регион, доход, образование и прочих. Минусы: обычно такие выборки нерепрезентативны, т.к. нельзя учесть сразу несколько социальных параметров. Плюсы: легкодоступный материал.

2. Метод снежного кома. Выборка строится следующим образом. У каждого респондента, начиная с первого, просят контакты его друзей, коллег, знакомых, которые подходили бы под условия отбора и могли бы принять участие в исследовании. Таким образом, за исключением первого шага, выборка формируется с участием самих объектов исследования. Метод часто применяется, когда необходимо найти и опросить труднодоступные группы респондентов (например, респондентов, имеющих высокий доход, респондентов, принадлежащих к одной профессиональной группе, респондентов, имеющих какие-либо схожие хобби/увлечения и т.д.)

3. Стихийная выборка – выборка так называемого «первого встречного». Часто используется в теле- и радиоопросах. Размер и состав стихийных выборок заранее не известен, и определяется только одним параметром – активностью респондентов. Минусы: невозможно установить какую генеральную совокупность представляют опрошенные, и как следствие – невозможность определить репрезентативность.

4. Маршрутный опрос – часто используется, если единицей изучения является семья. На карте населённого пункта, в котором будет производиться опрос, нумеруются все улицы. С помощью таблицы (генератора) случайных чисел отбираются большие числа. Каждое большое число рассматривается как состоящее из 3-х компонентов: номер улицы (2-3 первых числа), номер дома, номер квартиры. Например, число 14832: 14 – это номер улицы на карте, 8 – номер дома, 32 – номер квартиры.

5. Районированная выборка с отбором типичных объектов. Если после районирования из каждой группы отбирается типичный объект, т.е. объект, который по большинству изучаемых в исследовании характеристик приближается к средним показателям, такая выборка называется районированной с отбором типичных объектов.

В статистике мы связываем информацию из выборки с совокупностью с распределением выборки: теоретическим вероятностным распределением статистики для всех возможных выборок определенного размера (N). То есть распределение выборки является распределением статистики



(например, среднего или пропорции), основанной на каждой мыслимой комбинации случаев из популяции. Важным моментом в распределении выборки является то, что его характеристики основаны на законах вероятности, а не на эмпирической информации, и очень хорошо известны. На самом деле, распределение выборки является центральным понятием в статистике, и длительное изучение его характеристик, безусловно, соответствует порядку трех распределений, используемых в статистической статистике. Все приложения статистики перемещаются между выборкой и совокупностью посредством распределения выборки. Таким образом, три отдельных распределения переменной участвуют в каждом приложении статистики:

1. Распределение населения или переменная, которая, хотя и эмпирическая, неизвестна. Накапливать информацию о населении или делать выводы для него - единственная цель статистики.
2. Распределение выборки по переменной, которая является неэмпирической или теоретической. Из-за законов вероятности многое известно об этом распределении. В частности, можно определить форму, центральную тенденцию и дисперсию распределения, и, следовательно, распределение можно адекватно охарактеризовать.
3. Выборочное распределение переменной, которое является эмпирическим (то есть существует в реальности) и известно. Форма, центральная тенденция и дисперсия переменной могут быть установлены для образца. Помните, однако, что информация из выборки важна в первую очередь, поскольку она позволяет исследователю узнать о населении.

Полезность распределения выборки подразумевается его определением. Поскольку он включает в себя статистику всех возможных результатов выборки, распределение выборки позволяет нам оценить вероятность любого конкретного результата выборки, процесс, который будет занимать наше внимание во время проведения исследований.

Построение выборочного распределения. Распределение выборки является теоретическим, что означает, что оно никогда не строится. Однако, чтобы лучше понять структуру и функцию распределения, давайте рассмотрим пример того, как его можно построить. Предположим, мы хотели собрать некоторую информацию о возрасте определенного сообщества из 10 000 человек. Мы формируем выборку из 100 жителей, спрашиваем у всех 100 респондентов их возраст и используем эти индивидуальные оценки для расчета среднего возраста 27 лет. Этот показатель отмечен на графике на рисунке 7.1.

Обратите внимание, что эта выборка является одной из бесчисленных возможных комбинаций из 100 человек, взятых из этой 10 000 человек, а статистика (среднее из 27) является одним из 'миллионов возможных результатов выборки.

Теперь замените первые 100 респондентов, нарисуйте другую выборку того же размера ($N=100$) и снова вычислите средний возраст. Предположим, что среднее значение для второй выборки равно 30, и отметьте этот результат выборки на рисунке 7.1.

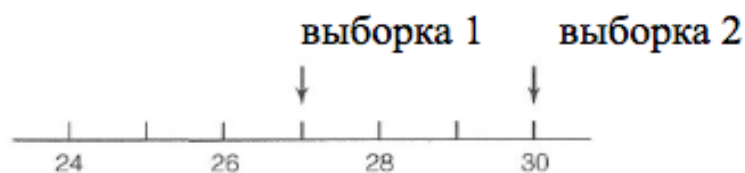


Рисунок 7.1 Построение выборочного распределения

это еще одна из бесчисленных комбинаций из 100 человек, взятых из этой 10 000 человек, а выборочное среднее 30 - это еще одна из миллионов возможных статистических выборок.



Замените этих респондентов и нарисуйте еще одну выборку, вычислите и отметьте среднее значение, замените эту третью выборку и нарисуйте четвертую выборку, непрерывно продолжая эти операции бесконечное число раз, вычисляя и не указывая среднее значение для каждой выборки. Теперь попытайтесь представить, как будет выглядеть рисунок 7.1. после того, как были собраны десятки тысяч отдельных образцов, и для каждого образца было рассчитано среднее значение. Какую форму, среднее значение и стандартное отклонение будет иметь это распределение выборочных средних, несмотря на то, что мы собрали все возможные комбинации 100 респондентов из 10 000 населения?

Во-первых, мы знаем, что каждая выборка будет, по крайней мере, немного отличаться от любой другой выборки, поскольку очень маловероятно, что мы будем отбирать точно одинаковые 100 человек дважды. Поскольку каждая выборка будет уникальной, среднее значение каждой выборки будет, по меньшей мере, различаться по значению.

Мы также знаем, что не все выборки будут репрезентативными. Например, если мы продолжим брать выборки из 100 человек достаточно долго, мы в конечном итоге выберем выборку, которая будет включать только самых маленьких жителей. Такая выборка будет иметь среднее значение намного ниже, чем истинное среднее значение популяции. Аналогичным образом, некоторые из наших выборок будут включать только пожилых людей и будут иметь значения, которые намного выше, чем среднее значение для населения. Однако здравый смысл подсказывает, что такие нерепрезентативные выборки будут редкими и что большинство выборочных средств будут сгруппированы вокруг истинного значения населения.

Чтобы проиллюстрировать это далее, предположим, что мы каким-то образом узнали, что истинный средний возраст населения в 10000 человек - 30 лет. Поскольку, как мы только что увидели, большая часть выборочных средних также будет приблизительно 30, выборочное распределение этих выборочных средних должно достигать пика в 30. Некоторые из выборок будут нерепрезентативными и их средства будут «не попадать в цель», но частота таких промахов должна уменьшаться по мере того, как мы отдаляемся от среднего значения населения 30. То есть, распределение должно падать к основанию, когда мы отходим от значения совокупности - выборочные средние значения 29 или 31 должны быть общими. Значения 20 или 40 должны быть редкими. Выборки являются случайными, поэтому их средние значения должны пропустить одинаковое число раз по обе стороны от значения совокупности, и поэтому само распределение должно быть примерно симметричным. Другими словами, распределение выборки всех возможных средних значений выборки должно быть приблизительно нормальным, и должно напоминать распределение, представленное на рисунке 7.2.

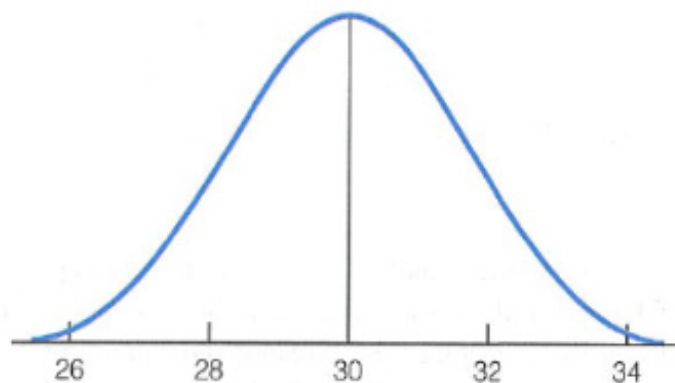


Рисунок 7.2. Выборочное распределение



ве теоремы. Эти общие идеи о форме распределения выборки и другая очень важная информация о центральной тенденции и дисперсии изложены в двух теоремах. Прежде чем исследовать теоремы, нам нужно рассмотреть некоторые символы, которые мы будем использовать. Напомним, что символом среднего значения для выборки является \bar{X} , а μ является символом для среднего населения. Для измерения дисперсии мы используем S для обозначения стандартного отклонения выборки σ (строчная греческая буква sigma) для обозначения стандартного отклонения популяции. Первая из этих теорем гласит: Если повторные случайные выборки размера N взяты из нормальной популяции со средним значением μ и стандартным отклонением σ , то распределение выборки средних значений выборки будет нормальным, со средним значением μ и стандартным отклонением σ/N

Интерпретация: если мы начнем с признака, который обычно распределяется по совокупности (например, оценки IQ) и возьмем бесконечное количество случайных выборок равного размера из этой совокупности, то распределение выборки средних значений выборки будет нормальным. Если мы знаем, что переменная нормально распределена среди населения, мы можем предположить, что выборочное распределение будет нормальным. Однако теорема говорит нам больше, чем форма распределения выборки. Он также определяет его среднее значение и стандартное отклонение. Фактически, это говорит о том, что среднее значение распределения выборки будет точно таким же значением, что и среднее значение по населению. То есть, если мы знаем, что средний IQ всей популяции равен 100, то мы знаем, что среднее любого распределения выборки из средних значений IQ выборки также равно 100. Почему именно эти два средства будут иметь одинаковое значение, не может быть полностью объяснено на этом уровне. Что касается дисперсии, то в теореме говорится, что стандартное отклонение распределения выборки, также называемое стандартной ошибкой среднего, будет равно стандартному отклонению популяции, деленному на квадратный корень из N (символически: σ/N). Если среднее значение и стандартное отклонение нормально распределенной переменной в совокупности известны, теорема позволяет нам вычислить среднее и стандартное отклонение выборочного распределения. Таким образом, мы будем знать точно столько же о распределении выборки (форма, центральная тенденция и дисперсия), сколько мы когда-либо знали о любом эмпирическом распределении. Первая теорема требует, чтобы переменная была нормально распределена в популяции. Что происходит, когда распределение рассматриваемой переменной неизвестно или известно, что оно не имеет нормальной формы (например, доход, который всегда имеет положительный сдвиг). Эти случаи (на самом деле очень распространенные) охватываются второй теоремой, называемой центральной предельной теоремой:

Если повторные случайные выборки размера N взяты из любой популяции со средним μ и стандартным отклонением σ , тогда, когда N станет большим, распределение выборки средних значений будет приближаться к норме, при этом среднее μ и стандартное отклонение будет равным σ/N

Интерпретация: Распределение выборки средних значений выборки станет нормальным по мере того, как размер выборки будет увеличиваться и изменяться, даже если переменная обычно не распределяется по совокупности. Когда N велико, среднее значение распределения выборки будет равно среднему значению для популяции, а его стандартное отклонение (или стандартная ошибка среднего) будет равно σ/N

Центральная предельная теорема важна, потому что она устраняет условие, что переменная должна быть нормально распределена в популяции. Когда размер выборки велик, мы можем предположить, что распределение выборки имеет нормальную форму со средним значением, равным среднему значению для популяции, и стандартным отклонением, равным σ/N . Таким образом, даже если мы работаем с переменной, которая, как известно, имеет асимметричное распределение (например, доход), мы все равно можем принять нормальное распределение выборки.



Оставшаяся проблема, конечно, состоит в том, чтобы определить, что подразумевается под большой выборкой. Хорошее эмпирическое правило заключается в том, что если размер выборки N равен 100 или более, применяется Центральная предельная теорема, и вы можете предположить, что распределение выборки статистики выборки имеет нормальную форму. Когда N меньше 100, вы должны иметь веские доказательства нормального распределения популяции, прежде чем можно будет предполагать, что распределение выборки является нормальным. Таким образом, если выборка превышает 100, мы можем предположить, что распределение выборки будет иметь нормальную форму.

Символы и терминология

В следующих лекциях мы будем работать с тремя совершенно разными дистрибутивами. Кроме того, нас будут интересовать несколько различных видов распределений выборки, включая выборочное распределение выборочных средних и выборочное распределение пропорций выборки. Чтобы различать эти различные распределения, мы будем использовать символы. Для быстрого ознакомления в Таблице 7.1 представлены символы, которые будут использоваться для распределения выборки. По сути, выборочное распределение обозначается греческими буквами, которые подписываются в соответствии с выборочной статистикой, представляющей интерес.

	Средний	Среднеквадратичное отклонение	Пропорция
1. Образцы	\bar{X}	s	P_s
2. Население	μ	σ	P_{μ}
3. Выборочные распределения из средств	$\mu_{\bar{X}}$	$\sigma_{\bar{X}}$	
из пропорций	μ_p	σ_p	

Таблица 7.1. Символы и среднеквадратичных отклонений трех распределений

Обратите внимание, что среднее значение и стандартное отклонение выборки обозначены английскими буквами (X и S), а среднее и стандартное отклонение популяции обозначены эквивалентами греческих букв (μ и σ). Значения, рассчитанные для выборок, обозначаются как P_s , в то время, как P_{μ} для популяций. Символами для выборочного распределения являются греческие буквы с англо-буквенными индексами. Среднее и стандартное отклонение выборочного распределения средних значений для выборки представляют собой $\mu_{\bar{X}}$ и $\sigma_{\bar{X}}$. Среднее и стандартное отклонение выборочного распределения значений выборки составляет P и P .



Выводы:

Существуют различные выборки со своими характеристиками

Распределение выборки, центральное понятие в логической статистике, является теоретическим распределением всех возможных результатов выборки. Поскольку его общая форма, среднее значение и стандартное отклонение известны (при условиях, указанных в двух теоремах), распределение выборки может быть адекватно охарактеризовано и использовано исследователями.

Две теоремы, которые были введены, утверждают, что когда интересующая переменная обычно распределяется в популяции или когда размер выборки велик, распределение выборки будет нормальным по форме, среднее будет равно среднему по популяции, и его стандартное отклонение (или стандартная ошибка) будет равно стандартному отклонению популяции, деленному на квадратный корень из N .