

Генеральная совокупность Случайная выборка и выборка Дизайн исследования

С.В. Курочкин, ИСПИ ВлГУ

Совокупности объектов

*Вся совокупность однородных объектов, обладающих некоторым признаком, интересующим исследователя, называется **генеральной совокупностью**.*

*Совокупность объектов, случайно отобранных из генеральной совокупности для исследования, называется **выборочной совокупностью или выборкой**.*

Повторной называют выборку, при составлении которой отобранный объект после исследования (перед отбором следующего) возвращается в генеральную совокупность.

Бесповторной называют выборку, при составлении которой отобранный объект после исследования в генеральную совокупность не возвращается.

Для проведения статистического анализа данных обычно используют бесповторную выборку.

Совокупности объектов

Для проведения статистического анализа данных обычно используют бесповторную выборку.

Выборочная совокупность позволяет оценить интересующий признак генеральной совокупности и обладает определенными свойствами:

- 1) она должна быть **репрезентативной** (представительной), т.е. выборка должна правильно представлять свойства и пропорции генеральной совокупности;
- 2) выборка должна быть **случайной** (и только тогда она будет репрезентативной), т.е. каждый объект выборки отобран случайно из генеральной совокупности и все объекты имеют одинаковую вероятность попасть в выборку.

*Число объектов совокупности называется **объемом** совокупности*

Способы представления статистических данных

Отношение частоты к объему выборки называют **относительной частотой**. Сумма всех относительных частот равна единице.

Если полученные в ходе наблюдения или эксперимента значения признака записывают в последовательности измерений, то получают **простой статистический ряд**.

Если полученные в ходе наблюдения или эксперимента значения признака записывают в порядке возрастания, то получают **вариационный ряд**.

Дискретным (точечным) статистическим распределением называется перечень вариантов и соответствующих им частот или относительных частот.

Статистическое распределение, заданное в виде последовательности интервалов и соответствующих им частот m_i встречаемости вариант в интервале, называется **интервальным (непрерывным) статистическим распределением**. Частота m_i встречаемости вариант в интервале равна сумме частот вариант, попавших в этот интервал.

Способы представления статистических данных

Дискретное статистическое распределение графически можно представить в виде *полигона частот (относительных частот)*.

Полигоном частот называют ломаную, отрезки которой соединяют точки $(x_1; m_1), (x_2; m_2), \dots, (x_k; m_k)$.

Интервальное статистическое распределение графически представляют в виде *гистограммы*.

Гистограммой называют совокупность смежных прямоугольников, расположенных на одной прямой, основания которых одинаковы и равны ширине интервала, а высоты равны отношению частоты или относительной частоты к ширине интервала. Отношение m_i/a или p_i^*/a называют соответственно **плотностью частоты** или **плотностью относительной частоты**.

Оценки параметров

- **Статистическая оценка неизвестного параметра теоретического распределения** – это функция от наблюдаемых случайных величин.

$\bar{\theta}_i$ называется **несмещенной оценкой** для параметра θ_i , если ее выборочное математическое ожидание равно оцениваемому параметру генеральной совокупности.

$$M(\bar{\theta}_i) = \theta_i; \quad M(\bar{\theta}_i) - \theta_i = \varphi_i,$$

где φ_i - это смещение оценки.

Смещенная оценка – это оценка параметра, чье математическое ожидание не равно оцениваемому параметру, т.е.

$$M(\bar{\theta}_i) \neq \theta_i \quad \text{или} \quad \varphi_i \neq 0$$

$\bar{\theta}_i$ является **состоятельной оценкой** для параметра θ_i , если она удовлетворяет закону больших чисел.

Точечные оценки параметров генеральной совокупности

Параметр	Оценка	Дисперсия для повторной выборки	Дисперсия для бесповторной выборки
μ	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	$\frac{\sigma^2}{n}$	$\frac{\sigma^2}{n} \left(1 - \frac{n}{N} \right)$
σ^2	$\overline{S^2} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	-	-
p	$\frac{m}{n}$	$\frac{p(1-p)}{n}$	$\frac{p(1-p)}{n} \left(1 - \frac{n}{N} \right)$

При больших n ($n > 30$) неизвестные параметры в формулах для дисперсии можно заменить на их выборочные значения без особой потери точности

Метод максимального правдоподобия

- Для непрерывной случайной величины *функцией правдоподобия* называется плотность вероятности совместного появления результатов выборки $x = (x_1 x_2 \dots x_n)$:

$$L(x_1, x_2, \dots, x_n, \theta) = P(x_1, \theta) \times P(x_2, \theta) \times \dots \times P(x_n, \theta)$$

Для дискретной случайной величины в этой формуле $P(x_i, \theta)$ означает вероятность появления значений x_i . В качестве оценки неизвестного параметра θ принимается такое значение $\bar{\theta}_n$, которое максимизирует функцию L .

Функция L является *мерой правдоподобия* полученных наблюдений x_1, x_2, \dots, x_n . Поэтому естественно выбирать оценку $\bar{\theta}_n$ так, чтобы имеющиеся наблюдения были наиболее правдоподобны.

Метод максимального правдоподобия

- Для непрерывной случайной величины *функцией правдоподобия* называется плотность вероятности совместного появления результатов выборки $x = (x_1 x_2 \dots x_n)$:

$$L(x_1, x_2, \dots, x_n, \theta) = P(x_1, \theta) \times P(x_2, \theta) \times \dots \times P(x_n, \theta)$$

Для дискретной случайной величины в этой формуле $P(x_i, \theta)$ означает вероятность появления значений x_i . В качестве оценки неизвестного параметра θ принимается такое значение $\bar{\theta}_n$, которое максимизирует функцию L .

Функция L является *мерой правдоподобия* полученных наблюдений x_1, x_2, \dots, x_n . Поэтому естественно выбирать оценку $\bar{\theta}_n$ так, чтобы имеющиеся наблюдения были наиболее правдоподобны.

Интервальное оценивание

Точечная оценка параметра – это оценка, которая характеризуется одним конкретным числом (например, математическим ожиданием, дисперсией, средним квадратичным отклонением и т.д.). Точечные оценки параметров генеральной совокупности могут быть приняты в качестве ориентировочных, первоначальных результатов обработки выборочных данных. Их основной недостаток заключается в том, что неизвестно, с какой точностью оценивается параметр.

Интервальные оценки позволяют построить с заданной вероятностью интервал, в котором находится оцениваемый параметр генеральной совокупности. Таким образом, интервальные оценки характеризуются двумя числами – концами интервалов.

Интервальное оценивание

Надежность [доверительная вероятность] оценки θ по $\bar{\theta}$ - это вероятность γ , с которой осуществляется неравенство

$$|\theta - \bar{\theta}| < \Delta$$

Вероятность того, что интервал $(\bar{\theta} - \Delta; \bar{\theta} + \Delta)$ включает в себе [покрывает] неизвестный параметр θ , равна γ :

$$P(\bar{\theta} - \Delta < \theta < \bar{\theta} + \Delta) = \gamma$$

Доверительный интервал – это интервал $(\bar{\theta} - \Delta; \bar{\theta} + \Delta)$, который покрывает оцениваемый параметр θ с заданной надежностью γ .

Границы $\bar{\theta} - \Delta$ и $\bar{\theta} + \Delta$ называются **доверительными границами** интервала. Они определяются на основе выборочных данных и являются функциями от случайных величин x_1, x_2, \dots, x_n , и, следовательно, сами являются случайными величинами.

Доверительный интервал

Доверительный интервал уровня надежности γ для генеральной средней a имеет вид:

$$\bar{x} - \Delta < a < \bar{x} + \Delta$$

где Δ - предельная ошибка, зависящая от доверительной вероятности γ .

При $n > 30$ для повторной выборки

$$\Delta = t \frac{\bar{S}}{\sqrt{n}}$$

для бесповторной выборки

$$\Delta = t \frac{\bar{S}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$$

Здесь t определяется из условия

$$\Phi(t) = \gamma$$

где $\Phi(t)$ - функция Лапласа.

Доверительный интервал

Параметр	Оценка	Предельная ошибка выборки			
		Повторная выборка		Бесповторная выборка	
		$n > 30$	$n \leq 30$	$n > 30$	$n \leq 30$
a	$\bar{x} = mean(x)$	$\Delta = t \frac{\bar{S}}{\sqrt{n}}$	$\Delta = t_{n-1} \frac{\bar{S}}{\sqrt{n}}$	$\Delta = t \frac{\bar{S}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$	$\Delta = t_{n-1} \frac{\bar{S}}{\sqrt{n}} \sqrt{1 - \frac{n}{N}}$
P	$w = \frac{m}{n}$	$\Delta = t \sqrt{\frac{w(1-w)}{n}}$		$\Delta = t \sqrt{\frac{w(1-w)}{n}} \left(1 - \frac{n}{N}\right)$	
		$t = qnorm\left(1 - \frac{\alpha}{2}, 0, 1\right) \quad t_{n-1} = qt\left(1 - \frac{\alpha}{2}, n - 1\right) \quad \bar{S} = \sqrt{S^2}$			

Доверительный интервал для дисперсии и объем выборки

Если $x = (x_1 \ x_2 \ \dots \ x_n)$ - n -выборка из нормальной совокупности $N(a, \sigma)$, где a и σ неизвестны, то статистика

$$\psi = \frac{n\overline{S^2}}{\sigma^2}$$

имеет распределение хи-квадрат с $n-1$ степенями свободы. Тогда доверительный интервал

$$P(\chi_{n-1}^2 > z_1) = \frac{1+\gamma}{2} = 1 - \frac{\alpha}{2}; \quad P(\chi_{n-1}^2 > z_2) = \frac{1-\gamma}{2} = \frac{\alpha}{2}$$

Параметр	Объем выборки	
	Повторная выборка	Бесповторная выборка
a	$n = \frac{t^2 \overline{S^2}}{\Delta^2}$	$n = \frac{N t^2 \overline{S^2}}{N \Delta^2 + t^2 \overline{S^2}}$
P	$n = \frac{w(1-w)t^2}{\Delta^2}$	$n = \frac{N t^2 w(1-w)}{N \Delta^2 + t^2 w(1-w)}$

Дизайн исследования

Дизайн **исследования** — это стратегия ответа на вопрос исследования с использованием эмпирических данных.

План исследования означает:

- Общие цели исследования и подход
- Тип дизайна исследования
- Методы выборки или критерии отбора субъектов
- Методы сбора данных
- Процедуры сбора данных
- Методы анализа данных

Виды подходов

Качественный подход	Количественный подход
<ul style="list-style-type: none">• Понимание субъективных переживаний, убеждений и концепций• Получите глубокие знания о конкретном контексте или культуре• Исследование малоизученных проблем и генерация новых идей	<ul style="list-style-type: none">• Измерение различных типов переменных и описание частоты, средних значений и корреляции• Проверка гипотез о взаимосвязях между переменными• Проверка эффективности нового лечения, программы или продукта

Практические и этические соображения при планировании исследования

Помимо научных соображений, при планировании исследования необходимо мыслить практически. Если в исследовании участвуют люди или животные, то необходимо учитывать этику исследования.

- Сколько времени дано на сбор данных и подготовку исследования?
- Имеется ли доступ к нужным данным (например, отправившись в определенное место или связавшись с определенными людьми)?
- Обладаете ли вы необходимыми исследовательскими навыками (например, статистический анализ или методы проведения интервью)?
- Потребуется ли этическое одобрение ?

На каждом этапе процесса разработки исследования убедитесь, что выбранный дизайн исследования практически осуществим.

Типы дизайна исследования

Тип дизайна	Назначение и характеристики
Экспериментальный	Используется для проверки причинно-следственных связей Включает в себя манипулирование независимой переменной и измерение ее влияния на зависимую переменную. Субъекты случайным образом распределяются по группам Обычно проводится в контролируемой среде (например, в лаборатории)
Квазиэкспериментальный	Используется для проверки причинно-следственных связей Аналогично экспериментальному плану, но без случайного распределения Часто включает сравнение результатов ранее существовавших групп Часто проводится в естественной среде (более высокая экологическая достоверность)
Корреляционный	Используется для проверки того, связаны ли переменные (и насколько сильно) Переменные измеряются без влияния на них
Описательный	Используется для описания характеристик, средних значений, тенденций и т. д. Переменные измеряются без влияния на них

- **Экспериментальные и квазиэкспериментальные** схемы позволяют проверить причинно-следственные связи
- **Описательные и корреляционные** планы позволяют измерять переменные и описывать отношения между ними.

Типы качественных исследований

Тип дизайна	Назначение и характеристики
Тематическое исследование	<p>Подробное изучение конкретного предмета (например, места, события, организации и т. д.).</p> <p>Данные могут быть собраны с использованием различных источников и методов.</p> <p>Ориентирован на целостное понимание дела</p>
Этнография	<p>Детальное изучение культуры конкретного сообщества или группы.</p> <p>Данные собираются путем продолжительного погружения и тщательного наблюдения.</p> <p>Основное внимание уделяется описанию и интерпретации убеждений, условностей, социальной динамики и т. д.</p>
Обоснованная теория	<p>Направлен на индуктивное развитие теории путем систематического анализа качественных данных</p>
Феноменология	<p>Направлена на понимание явления или события путем описания жизненного опыта участников</p>

Методы отбора проб

Вероятностная выборка	Невероятностная выборка
Образец отбирается случайным образом В основном используется в количественных исследованиях Позволяет делать убедительные статистические выводы о населении	Образец выбран неслучайным образом Используется как в качественных, так и в количественных исследованиях Легче достичь, но выше риск предвзятости исследования

- Вероятностная выборка является наиболее статистически достоверным вариантом, но ее часто трудно реализовать, если вы не имеете дело с очень небольшой и доступной совокупностью.
- По практическим причинам во многих исследованиях используется невероятностная выборка, но важно знать об ограничениях и тщательно учитывать потенциальные систематические ошибки. Вы всегда должны прилагать усилия для сбора максимально репрезентативной выборки населения.

Методы сбора данных

Методы сбора данных — это способы прямого измерения переменных и сбора информации.

Методы обследования

Анкеты	Интервью
<p>Чаще встречается в количественных исследованиях</p> <p>Может распространяться онлайн, по телефону, по почте или лично</p> <p>Обычно предлагают закрытые вопросы с ограниченными вариантами</p> <p>Непротиворечивые данные могут быть получены от многих людей</p>	<p>Чаще встречается в качественных исследованиях</p> <p>Проводится исследователем лично, по телефону или онлайн</p> <p>Обычно позволяют участникам отвечать своими словами</p> <p>Идеи можно подробно изучить в небольшой группе (например, в фокус-группе)</p>

Методы сбора данных

Методы наблюдения

Количественное наблюдение	Качественное наблюдение
Систематический подсчет или измерение Категории и критерии определяются заранее	Ведение подробных заметок и написание подробных описаний Все соответствующие наблюдения могут быть записаны

Другие методы сбора данных

Область деятельности	Примеры методов сбора данных
Медиа и коммуникации	Сбор выборки текстов (например, выступлений, статей или постов в социальных сетях) для сбора данных о культурных нормах и нарративах
Психология	Использование таких технологий, как нейровизуализация, отслеживание взгляда или компьютерные задачи для сбора данных о таких вещах, как внимание, эмоциональная реакция или время реакции
Образование	Использование тестов или заданий для сбора данных о знаниях и навыках
Физические науки	Использование научных инструментов для сбора данных о таких вещах, как вес, артериальное давление или химический состав

Стратегия анализа данных

В количественных исследованиях используется какая-либо форма статистического анализа. С помощью статистики можно обобщать выборочные данные, делать оценки и проверять гипотезы.

Используя описательную статистику можно обобщить выборочные данные с точки зрения:

- Распределение данных (например, частота каждой оценки в тесте)
- Центральная **тенденция** данных (например, среднее значение для описания среднего балла)
- Изменчивость данных (например, стандартное отклонение для описания степени разброса оценок)
- Конкретные расчеты, которые вы можете сделать, зависят от уровня измерения ваших переменных.

Стратегия анализа данных

Используя логическую статистику, вы можете:

- Сделать оценку показателя на основе ваших выборочных данных.
- Проверить гипотезы о связи между переменными.

Регрессионные и корреляционные тесты ищут ассоциации между двумя или более переменными, а сравнительные тесты (такие как t — тесты и ANOVA) ищут различия в результатах разных групп.

В качественных исследованиях данные обычно очень насыщены информацией и идеями. Вместо того, чтобы подводить итоги в цифрах, необходимо детально просмотреть данные, интерпретировать их значения, выявить закономерности и выделить те части, которые наиболее важны для исследовательского вопроса.

Стратегия анализа данных

Двумя наиболее распространенными подходами к этому являются тематический анализ и анализ дискурса.

Подход	Характеристики
Тематический анализ	Ориентирован на содержание данных Включает кодирование и организацию данных для определения ключевых тем
Анализ речи	Сосредоточены на размещении данных в контексте Включает анализ различных уровней общения (язык, структура, тон и т. д.)

Стратегия анализа данных

Двумя наиболее распространенными подходами к этому являются тематический анализ и анализ дискурса.

Подход	Характеристики
Тематический анализ	Ориентирован на содержание данных Включает кодирование и организацию данных для определения ключевых тем
Анализ речи	Сосредоточены на размещении данных в контексте Включает анализ различных уровней общения (язык, структура, тон и т. д.)