Выборки

Определение

Выборка

— подмножество генеральной совокупности

Определение

Семплирование

— способ получения выборки

Важно

В зависимости от семплирования выводы о генеральной совокупности по выборке могут очень сильно отличаться

- Пусть у нас есть генеральная совокупность
- Из нее мы можем выделить множество различных выборок
- Семплирование - способ получения выборки
- "какая выборка вероятнее всего получится, если формировать ее определенным образом"

> **Семплирование** - правило, которое для каждой выборки возвращает вероятность ее получения при определенном алгоритме выбора

Определение

Семплирование как распределение

— вероятностное распределение на множестве всех возможных выборок фиксированного размера из генеральной совокупности.

Таким образом, первое определение это **алгоритм формирования выборки**, а второе - **результат его применения**

⚠ Важно

если ты проводишь семплирование, ты можешь предсказать, **насколько вероятно** получить ту или иную выборку, но **не можешь точно сказать, какая именно окажется в твоих руках**.

Однако есть и неслучайное семплирование, когда итоговая выборка определяется однозначно.

Систематическая ошибка семплирования

Определение

Систематическая ошибка приближения среднего (смещенная оценка)

— разность между матожиданием выборочного среднего и средним в генеральной совокупности

$$\text{Bias} = \mathbb{E}(\overline{X}) - \mathbb{E}X$$

Если $\mathbb{E}(\overline{X}) = \mathbb{E}X$, систематической ошибки нет.

Пример

Истинное распределение роста людей — нормальное (μ=175см), но мы измеряем только спортсменов-баскетболистов, чей средний рост выше. Оценка среднего будет **систематически завышена**.

⚠ Важно

Отсутствие систематической ошибки (несмещённость) — это хорошо, но **не гарантирует точности оценки**, если при этом велика **случайная ошибка (дисперсия)**. В таких случаях оценка может быть «в среднем правильной», но на практике сильно варьироваться от выборки к выборке.

Пример

Вы хотите купить арбуз весом около 5 кг. У вас есть два друга, которые помогают оценить вес — **на глаз или на ручных весах**.

- **Анна** — никогда не обманывает, но очень неровно оценивает вес:
- В среднем она правильна (не завышает и не занижает),
- Но может сказать «4 кг», когда на самом деле 6 кг, или «7 кг» при 3 кг. → Её оценки **без систематической ошибки**, но с **большой дисперсией**.
- **Борис** — немного консервативен:
- Всегда говорит «на пару сотен грамм меньше», чем на самом деле,
- Зато почти всегда **попадает в предел ±100 грамм**. → Его оценки **имеют небольшое смещение**, но **очень малую дисперсию**.

Вычисление в Python методом Монте-Карло (см. далее):

```
means = [df.sample(n=10)['value'].mean() for _ in range(10_000)]
bias = np.mean(means) - population_mean
```

Определение

Маргинальная ошибка семплирования (margin of error)

— мера точности оценки, указывает насколько может отличатся выборочное среднее от среднего по генеральной совокупности с заданной вероятностью

$$P(|\overline{X} - \mathbb{E}X| \leq ME) = p,$$

где ME - маргинальная ошибка, p - уровень доверия

Пример

Маргинальная ошибка 3 при уровне доверия 95% означает, что истинная доля в генеральной совокупности с вероятностью 95% лежит между $\mu - 3$ и $\mu + 3$

Вычисление в Python методом Монте-Карло:

```
mc_means = np.array([df.sample(100)['spend'].mean() for _ in range(5000)])
me_95 = np.percentile(np.abs(mc_means - true_mean), 95)
print(f"Margin of error (95%): {me_95:.2f}")
```

Виды систематических ошибок

- **СО выжившего** - в выборке оказывается много наблюдений из одной группы объектов, которую условно называют «выжившие»
- **Черри-пикинг** - в выборку взяты те наблюдения, которые доказывают выводы исследователя, зато проигнорированы те, которые опровергают
- **СО самоотбора** - в выборку взяли тех, кто сам пожелал войти в неё
- **Систематическая ошибка недостаточного охвата отдельных частей генеральной совокупности (undercoverage bias)** - некоторые сегменты генеральной совокупности в выборке практически не представлены или представлены в меньшей пропорции, чем в генеральной совокупности. Если у таких объектов какая-то характеристика выше, чем у представленных, то выборочное значение будет систематически ниже, чем в генеральной совокупности.

Виды семплирования

- простое случайное семплирование
- случайное семплирование с возвращением
- стратифицированное случайное семплирование

Простое случайное семплирование

Определение

> **Простое случайное семплирование (SRS)**
>
> — семплирование, при котором все выборки из n наблюдений с учетом порядка и без возвращений равновероятны

Учет порядка нужен, чтобы быть уверенным в равновероятности выборки, он используется только на этапе формирования выборок.

Алгоритм

> **Получение SRS**
>
> 1. По исходной генеральной совокупности строим генеральную совокупность всех возможных выборок без повторения из n наблюдений с учётом порядка
> 2. Из этого набора равновероятно выбираем одну выборку

В Python:

```
df.sample(n)
```

Случайное семплирование с возвращением

Определение

Случайное семплирование с возвращением (RS)

— семплирование, при котором все выборки из n наблюдений с учетом порядка и возможным повторением элементов равновероятны

- алгоритм построения аналогичный

В Python:

```
df.sample(n, replace=True)
```

Лемма

Сравнение семплирования с возвращением и простого случайного

Если размер выборки намного меньше размера генеральной совокупности (эмпирически - если $n < 0{,}01 \cdot N$), то семплирование с возвращением и простое случайное семплирование приблизительно одинаковы

Объяснение

Пусть мы выбираем без возвращения один элемент за другим из огромной генеральной совокупности. Формально в ней становится меньше элементов, но из-за большого размера их остаётся всё ещё много: генеральная совокупность почти не меняется. Поэтому нет существенной разницы: возвращать или не возвращать очередной элемент выборки в генеральную совокупность.

Случайная выборка со стратификацией

Определение

Страта

— конечное подмножество генеральной совокупности, соответствующее одному из уровней фактора стратификации

Определение

Стратификация

— представление генеральной совокупности в виде дизъюнктного объединения страт

Определение

Фактор стратификации

— характеристика наблюдений, по значениям которой происходит разбиение на страты.

Интерпретация

Страта — это **однородная подгруппа (слой) генеральной совокупности**, выделенная по определённому признаку, который считается значимым для анализа.

Зачем это нужно?

1. **Уменьшить дисперсию оценок** (внутри страты объекты однороднее),
2. **Обеспечить представительность** (гарантировать, что каждый тип учтён),
3. **Позволить отдельный анализ по группам** (например, сравнить эффективность по классам техники).

Определение

Случайное семплирование со стратификацией (stratified random sampling)

— семплирование, при котором равновероятны все выборки из n наблюдений с учётом порядка без повторения элементов, для которых доля каждой страты в выборке совпадает с долей этой страты в генеральной совокупности (с точностью до округления)

Алгоритм

Получение стратифицированной выборки

1. Для каждой страты вычисляем требуемое количество наблюдений из этой страты в выборке $n_i = n \cdot P(Stratum_i)$
2. Из каждой страты получаем подвыборку нужного размера простым случайным семплированием.
3. Объединяем подвыборки в выборку.

В Python:

```
def stratified_sample(df, strata_col, n):
    # Пропорциональная стратификация: сколько строк брать из каждой страты
    def sample_group(x):
        # Доля страты в общем объёме
        group_size = len(x)
        # Пропорциональный размер выборки из этой страты
        sample_size = int(round(group_size / len(df) * n))
        # Берём минимум, чтобы избежать проблем с малыми группами
        return x.sample(min(sample_size, group_size))

    result = df.groupby(strata_col).apply(sample_group)
    return result.reset_index(drop=True)
```

Важно

- Если n_i целые, то СО нет, однако если они дробные, то приходится округлять до целых.
- Если округлять вниз, то количество наблюдений будет меньше требуемого, если вверх, то больше
- Так или иначе появится некоторая СО, однако маргинальная ошибка все равно уменьшится по сравнению с SRS

> Стратифицированное семплирование заметно точнее простого случайного, если разброс значений внутри страт заметно меньше разброса среди всех значений в генеральной совокупности

Метод Монте-Карло

Определение

Эксперимент Монте-Карло

1. случайная генерация одной выборки в соответствии с семплированием
2. проверка, случилось ли в выборке событие или вычисление описательной статистики
3. сохранение результата для этой выборки.

Определение

Метод Монте-Карло

— это проведение M случайных экспериментов Монте-Карло и дальнейшая обработка их результатов

В Python с SRS:

```
np.random.seed(42)
M = 10_000
count = 0
for _ in range(M):
    df_sample = df.sample(n=2) # SRS
    if df_sample['Рост'].mean().astype(int) == 170: # событие случилось
        count = count + 1 # считаем количество
print(count / M)
```

Цель метода Монте-Карло

Аппроксимировать теоретические вероятности и распределения с помощью частоты на большом числе симуляций.

Лемма

Точность метода Монте-Карло

Эмпирическое правило: если увеличить в 100 раз размер выборки и она всё ещё останется маленькой ($10n < 0{,}01 \cdot N$), то точность возрастёт в 10 раз, но для больших выборок увеличение размера в k раз повышает точность больше чем в $\sqrt{k}$ раз.

Flashcards

tags: #flashcardsSTAT

Что такое Выборка?
%
— подмножество генеральной совокупности

Что такое Семплирование?
%
— способ получения выборки

Что такое Семплирование как распределение?
%
— вероятностное распределение на множестве всех возможных выборок фиксированного размера из генеральной совокупности.

Систематическая ошибка семплирования

Что такое Систематическая ошибка приближения среднего (смещенная оценка)?
%
— разность между матожиданием выборочного среднего и средним в генеральной совокупности

> $$\text{Bias} = \mathbb{E}(\overline{X}) - \mathbb{E}X$$
>
> Если $\mathbb{E}(\overline{X}) = \mathbb{E}X$, систематической ошибки нет.

Что такое Маргинальная ошибка семплирования (margin of error)?
%
— мера точности оценки, указывает насколько может отличатся выборочное среднее от среднего по генеральной совокупности с заданной вероятностью

$$P(|\overline{X} - \mathbb{E}X| \leq ME) = p,$$

где ME - маргинальная ошибка, p - уровень доверия

Виды семплирования

Что такое Простое случайное семплирование (SRS)?
%
— семплирование, при котором все выборки из n наблюдений с учетом порядка и без возвращений равновероятны

Что такое Получение SRS?
%

1. По исходной генеральной совокупности строим генеральную совокупность всех возможных выборок без повторения из n наблюдений с учётом порядка
2. Из этого набора равновероятно выбираем одну выборку

Что такое Случайное семплирование с возвращением (RS)?
%
— семплирование, при котором все выборки из n наблюдений с учетом порядка и возможным повторением элементов равновероятны

- алгоритм построения аналогичный

Сформулируйте лемму: Сравнение семплирования с возвращением и простого случайного
%
Если размер выборки намного меньше размера генеральной совокупности (эмпирически - если n < 0,01 · N), то семплирование с возвращением и простое случайное семплирование приблизительно одинаковы

Что такое Страта?
%
— конечное подмножество генеральной совокупности, соответствующее одному из уровней фактора стратификации

Что такое Стратификация?
%
— представление генеральной совокупности в виде дизъюнктного объединения страт

Что такое Фактор стратификации?
%
— характеристика наблюдений, по значениям которой происходит разбиение на страты.

Что такое Случайное семплирование со стратификацией (stratified random sampling)?
%
— семплирование, при котором равновероятны все выборки из n наблюдений с учётом порядка без повторения элементов, для которых доля каждой страты в выборке совпадает с долей этой страты в генеральной совокупности (с точностью до округления)

Что такое Получение стратифицированной выборки?
%

1. Для каждой страты вычисляем требуемое количество наблюдений из этой страты в выборке $n_i = n \cdot P(Stratum_i)$
2. Из каждой страты получаем подвыборку нужного размера простым случайным семплированием.
3. Объединяем подвыборки в выборку.

Метод Монте-Карло

Что такое Эксперимент Монте-Карло?
%

1. случайная генерация одной выборки в соответствии с семплированием
2. проверка, случилось ли в выборке событие или вычисление описательной статистики

3. сохранение результата для этой выборки.

Что такое Метод Монте-Карло?
%
— это проведение M случайных экспериментов Монте-Карло и дальнейшая обработка их результатов

Сформулируйте лемму: Точность метода Монте-Карло
%
Эмпирическое правило: если увеличить в 100 раз размер выборки и она всё ещё останется маленькой ($10n < 0{,}01 \cdot N$), то точность возрастёт в 10 раз, но для больших выборок увеличение размера в k раз повышает точность больше чем в $\sqrt{k}$ раз.

Что такое неслучайное семплирование?
%
— семплирование, при котором одна выборка имеет вероятность 100%, все остальные — 0%.

Почему undercoverage bias не зависит от размера выборки?
%
— потому что он вызван **отсутствием или недоучётом целых сегментов** генеральной совокупности. Даже огромная выборка будет смещённой, если она не включает важную группу.

Как влияет размер выборки на точность Монте-Карло?
%
— увеличение числа симуляций M повышает точность оценки ~ на $\sqrt{M}$. Чтобы ошибка уменьшилась в 10 раз, нужно увеличить M в 100 раз.

Как получить простую случайную выборку (SRS) в pandas?
%

```
df.sample(n=100)
```

🔷 По умолчанию `replace=False` и `random_state` не фиксирован.
✅ Для воспроизводимости добавьте `random_state`:

```
df.sample(n=100, random_state=42)
```

Как получить случайную выборку с возвращением (RS)?
%

```
df.sample(n=100, replace=True, random_state=42)
```

🔷 `replace=True` позволяет одному наблюдению попасть в выборку несколько раз.

Как реализовать стратифицированное семплирование по столбцу `strata_col`?
%

```python
def stratified_sample(df, strata_col, n_total):
    return (
        df.groupby(strata_col)
          .sample(frac=n_total/len(df), replace=False)
          .sample(frac=1, random_state=42)  # перемешать в конце
    )
```

🔷 `frac=n_total/len(df)` — доля от каждой страты.
🔷 Второй `.sample(frac=1)` — перемешивает объединённую выборку.
⚠️ При округлении может быть небольшая ошибка покрытия.

Как провести эксперимент Монте-Карло для оценки среднего?
%

```python
np.random.seed(42)
M = 10_000
mc_means = []

for _ in range(M):
    sample = df.sample(n=200)
    mc_means.append(sample['value'].mean())

# Теперь можно анализировать распределение mc_means
```

✅ Цель: оценить поведение выборочного среднего при многократном повторении выборки.

Как вычислить маргинальную ошибку (margin of error) методом Монте-Карло?
%

```python
true_mean = df['value'].mean()  # или известно из условия
errors = np.abs(np.array(mc_means) - true_mean)
me_90 = np.percentile(errors, 90)  # ME с вероятностью 90%
print(f"Margin of error (90%): {me_90:.2f}")
```

📌 Это эмпирическая оценка:
$P(|\overline{X} - \mathbb{E}X| \leq ME) = 0.9$

Как проверить, есть ли систематическая ошибка у семплирования?
%

```python
bias = np.mean(mc_means) - true_mean
print(f"Bias: {bias:.3f}")
```

✅ Если `bias ≈ 0` — систематической ошибки нет.
🔷 Полезно сравнивать для разных методов (SRS, RS, стратификация).

Как написать стратифицированную выборку с контролем размера подвыборки?
%

```
def stratified_sample_exact(df, strata_col, n_total):
    props = df[strata_col].value_counts(normalize=True)
    n_strata = (props * n_total).round().astype(int)

    # Корректировка суммы
    while n_strata.sum() != n_total:
        diff = n_total - n_strata.sum()
        idx = np.random.choice(n_strata.index)
        n_strata[idx] += np.sign(diff)

    return df.groupby(strata_col, group_keys=False).apply(
        lambda x: x.sample(n=n_strata[x.name], random_state=42)
    )
```