Конспект по теме "Проверка гипотез"

Случайная выборка и выборочное среднее

Логика проведения статистической проверки гипотез немного другая, по сравнению с механизмами в теории вероятностей. Прежде всего, мы будем судить о большом объёме данных, **генеральной совокупности**, по части — **выборке**.

Для анализа необязательно загружать все данные, достаточно взять небольшую, но **репрезентативную**, представляющую всю генеральную совокупность, часть данных. Самый простой способ добиться репрезентативности — взять **случайную выборку**. Из всего датасета генератором случайных чисел отбирают случайные элементы. По ним будут судить обо всей генеральной совокупности.

Она может состоять из нескольких неравных по размеру частей, сильно отличающихся по исследуемому параметру. Тогда есть смысл взять пропорциональные случайные выборки из этих частей, и потом соединить между собой. Получается **стратифицированная выборка**, более репрезентативная, чем просто случайная. Она так называется, потому что мы разбили генеральную совокупность на **страты** — группы, объединённые общим признаком. Случайные выборки получают уже из них.

По выборке судят о генеральной совокупности — точнее об её статистических параметрах. Обычно достаточно оценить среднее и дисперсию, чтобы сделать выводы о **равенстве или неравенстве средних значений** исследуемых совокупностей. Нас будет интересовать именно такая постановка задачи.

Что можно сказать о среднем и дисперсии генеральной совокупности по среднему и дисперсии, посчитанным на выборке, или **выборочному среднему** и **выборочной дисперсии**? Почти всё, при условии, что выборка достаточно велика.

Одна из формулировок центральной предельной теоремы звучит так: если в выборке достаточно наблюдений, **выборочное распределение** выборочного среднего из любой генеральной совокупности распределено

нормально вокруг среднего этой генеральной совокупности. *«Любая генеральная совокупность»* означает, что сама генеральная совокупность может быть распределена как угодно. Датасет из средних значений выборок всё равно будет нормально распределён вокруг среднего всей генеральной совокупности.

Стандартное отклонение выборочного среднего от настоящего среднего генеральной совокупности называется **стандартной ошибкой** и находится по формуле:

$$E.S.E. = \frac{S}{\sqrt{n}}$$

E.S.E. — оценённая стандартная ошибка. «Оценённая» — имея только выборку, мы не знаем точную ошибку и *оцениваем* её исходя из имеющихся данных.

S — оценка стандартного отклонения генеральной совокупности.

n — размер выборки. Раз корень из *n* стоит в знаменателе, стандартная ошибка уменьшается с увеличением размера выборки.

Формулирование гипотез

Никакие экспериментально полученные данные никогда не *подтвердят* какую-либо гипотезу. Это наше фундаментальное ограничение. Данные могут лишь не противоречить ей или, наоборот, показывать крайне маловероятные результаты (при условии, что гипотеза верна). Но и в том, и в другом случае нет оснований утверждать, что выдвинутая гипотеза *доказана*.

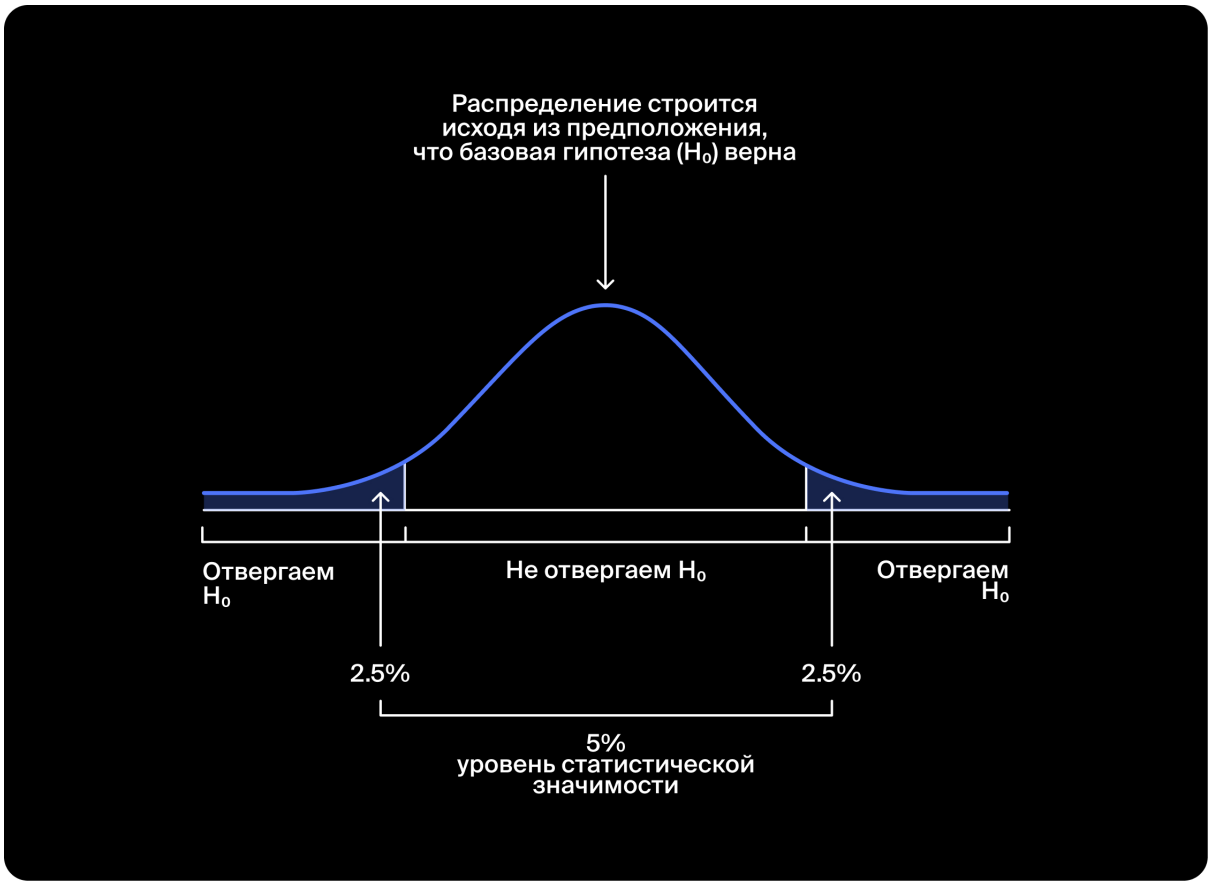
Допустим, данные гипотезе не противоречат, тогда мы её *не отвергаем*. Если же мы приходим к выводу, что получить такие данные в рамках этой гипотезы вряд ли возможно, у нас появляется основание отбросить эту гипотезу.

Типичные статистические гипотезы касаются средних значений генеральных совокупностей и звучат так:

- среднее генеральной совокупности равно конкретному значению;
- средние двух генеральных совокупностей равны между собой.

Алгоритм проверки статистических гипотез всегда начинается с формулирования гипотез. Сначала формулируется **нулевая гипотеза H_0**. Например, «среднее рассматриваемой генеральной совокупности равно A», где *A* — некоторое число. Исходя из H_0 формулируется **альтернативная гипотеза H_1**. Для этой H_0 она звучит как «среднее генеральной совокупности не равно A». H_0 всегда формулируется так, чтобы использовать знак равенства.

Построим распределение на предположении, что гипотеза H_0 верна. В нашем случае это будет нормальное распределение вокруг интересующего нас параметра — среднего. Дисперсия, или стандартное отклонение, оценивается по данным выборки.



Для нормального распределения вероятность попасть в тот или иной интервал равна площади графика над этим интервалом. В районе среднего значения и в некотором диапазоне вокруг него будут значения, которые весьма вероятно получить случайно.

Как определить, где мы ещё не отвергаем нулевую гипотезу, а где пора? Критические значения задаются выбранным уровнем значимости

проверки гипотезы. **Уровень значимости** — это суммарная вероятность того, что измеренное эмпирически значение окажется далеко от среднего. Если наблюдаемое на выборке значение попадает в эту зону, вероятность такого события при верной нулевой гипотезе признаётся слишком малым, значит, у нас есть основание отвергнуть нулевую гипотезу. Когда значение попадает в зону «Не отвергаем H_0», то оснований отвергать нулевую гипотезу нет. Считаем, что эмпирически полученные данные не противоречат нулевой гипотезе.

В Python существует метод, который просто возвращает **статистику разности** между средним и тем значением, с которым вы его сравниваете. Главное — уровень значимости, на котором они находятся друг от друга — **p-value**.

Статистика разности — это количество стандартных отклонений между сравниваемыми значениями, если оба распределения привести к стандартному нормальному распределению со средним 0 и стандартным отклонением 1. По этой цифре сложно сориентироваться.

Есть смысл принимать решение о принятии или отвержении нулевой гипотезы по ***p-value***. Это вероятность получить наблюдаемый результат при условии, что нулевая гипотеза верна. Общепринятые пороговые значения — 5% и 1%. Окончательное решение, какой порог считать достаточным, всегда остаётся за аналитиком.

Для проверки гипотезы о равенстве среднего генеральной совокупности некоторому значению, можно использовать метод `scipy.stats.ttest_1samp()`. Параметры метода: `array` — массив, содержащий выборку; `popmean` — предполагаемое среднее, на равенство которому мы делаем тест. После выполнения метод вернёт статистику разности между popmean и выборочным средним из array, а также уровень значимости:

```
from scipy import stats as st

interested_value = 120

results = st.ttest_1samp(
    array,
    interested_value)

print('p-значение: ', results.pvalue)
```

Гипотеза о равенстве средних двух генеральных совокупностей

Когда генеральных совокупностей две, бывает нужно сопоставить их средние. Чтобы проверить гипотезу о равенстве среднего двух генеральных совокупностей по взятым из них выборкам, примените метод `scipy.stats.ttest_ind()`. Методу передают параметры: `array1`, `array2` — массивы, содержащие выборки; `equal_var` — **необязательный** параметр, задающий считать ли равными дисперсии выборок. Если есть основание полагать, что выборки взяты из схожих по параметрам совокупностей, тогда укажите `equal_var = True`, и дисперсия каждой выборки будет оценена по объединенному датасету из двух выборок, а не для каждой по отдельности по значениям в ней самой. Это позволяет получить более точные результаты, но только в том случае, если считать примерно равными дисперсии генеральных совокупностей, из которых взяты выборки. Иначе нужно указать `equal_var = False`; по умолчанию он задан как `equal_var = True` (если вообще его не указывать).

```
from scipy import stats as st

sample_1 = [...]
sample_2 = [...]

results = st.ttest_ind(
    sample_1,
    sample_2)

print('p-значение: ', results.pvalue)
```

Гипотеза о равенстве средних для зависимых (парных) выборок

Когда генеральная совокупность одна, полезно понять, равно ли себе среднее этой совокупности до и после изменения. **Парная выборка** означает, что мы измеряем некоторую переменную для одних и тех же единиц. Чтобы проверить гипотезу о равенстве средних двух

генеральных совокупностей для зависимых (парных) выборок в Python, применим функцию `scipy.stats.ttest_rel()`:

```
from scipy import stats as st

before = [...]
after = [...]

results = st.ttest_rel(
    before,
    after)

print('p-значение: ', results.pvalue)
```