Оценка статистической мощности t-критерия Стьюдента

В этом сообщении я покажу, как можно рассчитать статистическую мощность t-критерия Стьюдента средствами R. Начнем с краткого рассмотрения сути проблемы.

Статистическая мощность (реже "чувствительность") (англ. statistical power) - это вероятность того, что тот или иной статистический критерий правильно отклонит неверную нулевую гипотезу. Иными словами - это способность критерия обнаружить различия там, где они действительно существуют. Обычно процесс проверки статистической гипотезы включает следующие шаги:

Формулировка собственно проверяемой нулевой гипотезы. Например, в случае двухвыборочного критерия Стьюдента она состоит в том, что обе выборки происходят из нормально распределенных генеральных совокупностей с одинаковыми средними значениями (подробнее см. здесь).

Выбор подходящего статистического критерия для проверки нулевой гипотезы. Вычисление значения этого критерия по имеющимся выборочным данным.

Определение критического значения критерия, исходя из желаемого уровня статистической значимости α и свойств теоретического распределения этого критерия.

Проверка того, превышает ли рассчитанный по выборочным данным критерий критическое значение. Если такое превышение не наблюдается, делают вывод о том, что нулевая гипотеза верна.

Не обнаружив различий, исследователь, к сожалению, часто делает вывод об их действительном отсутствии. Однако проблема заключается в том, что такой вывод не всегда будет верным, т.к. сравнивая случайные выборки исследователь, как известно, неизбежно рискует совершить одну из двух типов ошибок:

Ошибка первого типа (= "первого рода"): отклонение верной нулевой гипотезы. Риск совершить такую ошибку равен выбранному уровню значимости (например, α=0.05).

Ошибка второго типа (= "второго рода"): сохранение неверной нулевой гипотезы. Вероятность ошибочно сохранить неверную нулевую гипотезу обозначают буквой β.

Таким образом, прежде чем сделать вывод об остутствии различий, исследователь должен выяснить, была ли мощность использованного статистического критерия достаточной для их обнаружения. Уровень β-риска тесно связан с 1) величиной различий между выборками (т.н. "величиной эффекта"), 2) числом наблюдений и 3) разбросом данных. Наиболее важным является число наблюдений: чем больше размер выборок, тем выше мощность теста. При "достаточно" больших выборках даже небольшие различия окажутся статистически значимыми. И наоборот - при малых выборках даже большие различия выявить будет трудно. Зная эти закономерности, мы можем заранее (т.е. до проведения исследования) определить минимальный размер выборок, необходимый для выявления эффекта.

На практике обычно (но не всегда) приемлемой считается мощность теста, равная или превышающая 80% (что соответствует β-риску в 20%). Этот уровень является следствием т.н. "соотношения 1 к 4" (англ. "one-to-four trade-off") между уровнями α-риска и β-риска: если принять уровень значимости α=0.05, тогда β=0.05×4=0.20 и мощность критерия составит Π=1−0.20=0.80.

При проведении вычислений, связанных с мощностью критерия Стьюдента, нам придестя оперировать следующими параметрами:

Величина эффекта, которую мы хотим выявить в ходе исследования ("дельта", т.е. разница между средними значениями сравниваемых выборок).

Стандартное отклонение (предполагается, что оно статистически не различается в сравниваемых выборках).

Уровень значимости, α.

Мощность критерия (обычно выражается в %).

Объем выборки

Источник: lifecity.com.ua

В качестве примера предположим, что мы планируем проведение эксперимента для установления эффекта температуры на индивидуальный вес водного жука. В эксперименте будут задействованы два температурных режима и, соответственно, установление эффекта температуры мы будем проверять, сравнивая средние значения веса жуков из двух экспериментальных групп при помощи двустороннего критерия Стьюдента (двусторонний вариант выбран потому, что до проведения эксперимента мы не знаем, каков именно эффект окажет повышение температуры - повышение или понижение веса жуков). Проверяемая в данном случае нулевая гипотеза состоит в том, что температура не оказывает никакого влияния на вес жуков.

Допустим, что минимальная разница в среднем весе жуков, которую мы хотим выявить в ходе эксперимента, составляет 3 мг. При уровне значимости α=0.05 желаемая мощность теста должна составить 80%. Вопрос заключается в том, сколько животных мы должны задействовать в эксперименте для того, чтобы перечисленные условия были выполнены. Как следует из приведенного выше списка, для определения оптимального размера выборки нам необходимо знать стандартное отклонение веса изучаемого вида жуков. К сожалению, до проведения эксперимента мы не можем точно оценить этот параметр. Вариантов решения этой проблемы два: 1) основываясь на своем экспертном мнении, исследователь может дать примерную оценку стандартного отклонения; 2) можно попытаться найти соответствующие литературные данные. Предположим, что мы воспользовались вторым вариантом и выяснили, что стандартное отклонение веса для изучаемого вида жуков составляет 1.8 мг.

Теперь у нас есть вся необходимая информация для расчета минимального объема выборки. В R соответствующие вычисления можно выполнить при помощи базовой функции power.t.test():

power.t.test(delta = 3.0,

sd = 1.8,

sig.level = 0.05,

power = 0.8)

Two-sample t test power calculation

n = 6.76095

delta = 3

sd = 1.8

sig.level = 0.05

power = 0.8

alternative = two.sided

NOTE: n is number in \*each\* group

В приведенной выше команде delta - минимальная величина эффекта, которую мы хотим обнаружить в ходе эксперимента, sd - стандартное отклонение веса жуков (по литературным данным), sig.level - уровень значимости, а power - мощность t-критерия. В результатах вычислений программа еще раз перечисляет имеющиеся исходные параметры, а также сообщает n - рассчитанный минимальный размер каждой выборки для обнаружения желаемого эффекта при этих параметрах (округлив, получаем 7 жуков в каждой экспериментальной группе). Кроме того, программа напоминает нам, что вычисления были выполнены для двустроннего критерия Стьюдента (alternative = two.sided) и что параметр n соответствует числу наблюдений в каждой группе (n is number in \*each\* group).

Зная число наблюдений, величину эффекта, стандартное отклонение и уровень значимости мы можем рассчитать мощность теста. Например:

power.t.test(n = 15,

delta = 3.0,

sd = 1.8,

sig.level = 0.05)

Two-sample t test power calculation

n = 15

delta = 3

sd = 1.8

sig.level = 0.05

power = 0.9927162

alternative = two.sided

NOTE: n is number in \*each\* group

Как видим, при n = 15, delta = 3, sd = 1.8 и sig.level = 0.05 мощность критерия составит 99%.

При необходимости выполнить вычисления для парного критерия Стьюдента, в вызов функции power.t.test() достаточно добавить аргумент type = "paired":

power.t.test(delta = 3.0,

sd = 1.8,

sig.level = 0.05,

power = 0.8,

type = "paired")

Paired t test power calculation

n = 5.04919

delta = 3

sd = 1.8

sig.level = 0.05

power = 0.8

alternative = two.sided

NOTE: n is number of \*pairs\*, sd is std.dev. of \*differences\* within pairs

Как видим, в случае с зависимыми выборками минимальный размер выборок, необходимый для выявления заданной величины эффекта, несколько меньше, чем в случае с независимыми выборками (в рассматриваемом примере - 5 против 7 жуков в каждой группе).

Наконец, при необходимости выполнить одновыборочный t-тест аргументу type следует присвоить значение "one.sample":

power.t.test(delta = 3.0,

sd = 1.8,

sig.level = 0.05,

power = 0.8,

type = "one.sample")

One-sample t test power calculation

n = 5.04919

delta = 3

sd = 1.8

sig.level = 0.05

power = 0.8

alternative = two.sided

В заключение следует еще раз подчеркнуть, что если в ходе выполнения t-теста проверяемая нулевая гипотеза не отклонена, это не значит, что эффект в действительности остутствует. Возможно, объемы выборок были просто недостаточно велики для этого. Поэтому рекомендуется приводить в отчетах (статьях, презентациях, и т.п.) не только результаты статистического теста как такового, но еще и информацию о его мощности. Это позволит читателям отчета сформировать четкое представление о том, насколько полученные данные поддерживают сделанные по результатам анализа выводы.