Памятка по проверке статистических гипотез через построение критической области и вычисление p-value

Тамбовцева А.А.

Небольшое вступление

Чтобы понимать, откуда при вычислении наблюдаемых значений статистики критерия в формулах возникают некоторые выражения, полезно вспомнить факты, вытекающие из центральной предельной теоремы:

1. Если мы рассматриваем выборки достаточно большого размера n ($n \geqslant 30$) из генеральной совокупности со средним a и стандартным отклонением σ, выборочное среднее имеет нормальное распределение со средним a и стандартным отклонением $\frac{\sigma}{\sqrt{n}}$.
2. Если мы рассматриваем выборки достаточно большого размера n из генеральной совокупности, имеющей бинарное распределение[1] с параметром p, то:
   - число успехов S_n описывается нормальным распределением с математическим ожиданием np и стандартным отклонением $\sqrt{npq}$.
   - доля успехов $\hat{p}$ описывается нормальным распределением с математическим ожиданием p и стандартным отклонением $\sqrt{\frac{pq}{n}}$.

Алгоритм проверки статистической гипотезы

1. Сформулировать нулевую и альтернативную гипотезы.

Гипотезы формулируем для параметров распределения генеральной совокупности (зачем строить гипотезы о конкретной выборке, которая у нас есть, если про неё мы и так всё знаем со 100%-ной уверенностью, любую характеристику можно просто посчитать).

Примеры гипотез для доли (p_0 – конкретное число, например, 0.5):

$$H_0 : p = p_0$$

$H_1 : p \neq p_0$ (двусторонняя) $H_1 : p > p_0$ (правосторонняя) $H_1 : p < p_0$ (левосторонняя)

Примеры гипотез для среднего (a – конкретное число, например, 10):

$$H_0 : \mu = a$$

$H_1 : \mu \neq a$ (двусторонняя) $H_1 : \mu > a$ (правосторонняя) $H_1 : \mu < a$ (левосторонняя)

Если нам нужно зафиксировать одностороннюю альтернативу и направление мы выбираем, исходя из данных, то:

- правосторонняя альтернатива выбирается, если выборочная оценка больше значения, указанного в нулевой гипотезе ($\hat{p} > p_0$ или $\bar{x} > a$);

[1] Набор из успехов и неудач, закодированных как 0 и 1, где p – вероятность успеха

- левосторонняя альтернатива выбирается, если выборочная оценка меньше значения, указанного в нулевой гипотезе ($\hat{p} < p_0$ или $\bar{x} < a$).

2. Выбрать подходящий статистический критерий для проверки нулевой гипотезы и посчитать наблюдаемое значение статистики критерия.

Выбор критерия зависит от задачи. Так, для проверки гипотезы о равенстве доли числу мы используем z-критерий для доли, для проверки гипотезы о равенстве среднего числу – критерий Стьюдента для одной выборки.

Проверка гипотезы о доле

Рассмотрим логику вычисления наблюдаемого значения статистики на примере доли.

Задача, которую мы решаем при проверке гипотезы – оценить разницу между значением параметра распределения, заявленным в нулевой гипотезе (вероятность p_0, среднее a) и значением его оценки, которое мы получаем по выборке (выборочная доля $\hat{p_0}$, выборочное среднее $\bar{x}$).

Если мы будем просто сравнивать два числа как есть, мы упустим из вида тот факт, что значения выборочных оценок могут изменяться от выборки к выборке. Например, один исследователь сформирует одну выборку из 100 человек и выяснит, что доля оптимистов среди них равна 0.6, другой сформирует другую выборку объёма 100 и получит долю оптимистов равную 0.4. И теоретически таких вариантов может быть бесконечно много. Как поступить? Учесть эту изменчивость и оценивать разницу между $\hat{p}$ и p_0 или между $\bar{x}$ и a относительно этой изменчивости.

Выборочные оценки являются случайными величинами. Какие характеристики отвечают за изменчивость, вариативность случайной величины? Дисперсия и стандартное отклонение. Выберем стандартное отклонение как более интерпретируемый показатель и определим его для доли в предположении, что нулевая гипотеза верна:

$$\sqrt{\frac{p_0 q_0}{n}}$$

Теперь учтём этот показатель изменчивости и оценим разницу между значением вероятности из нулевой гипотезы p_0 и значением доли по выборке $\hat{p}$. Получим наблюдаемое значение статистики критерия:

$$z_{\text{набл}} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

К тому же результату можно прийти и другим способом – сравнивать не доли, а абсолютное число успехов, полученное на выборке S_n, и ожидаемое число успехов в случае, если H_0 верна[2]:

$$z_{\text{набл}} = \frac{S_n - np_0}{\sqrt{np_0 q_0}}.$$

[2] Формула ниже эквивалентна предыдущей – поделите числитель и знаменатель во второй формуле на n и увидите, что получится первая формула. А при делении числителя и знаменателя на одно и то же число дробь не меняется.

Проверка гипотезы о среднем

Здесь логика та же: сравниваем среднее значение, полученное по выборке, и значение в нулевой гипотезе, а затем полученную разность делим на показатель изменчивости среднего:

$$t_{\text{набл}} = \frac{\bar{x} - a}{\frac{s}{\sqrt{n}}}$$

3. Проверить, насколько полученное наблюдаемое значение статистики критерия типично при условии, что нулевая гипотеза верна.

Сделать это можно двумя способами: через построение критической области и через вычисление p-value.

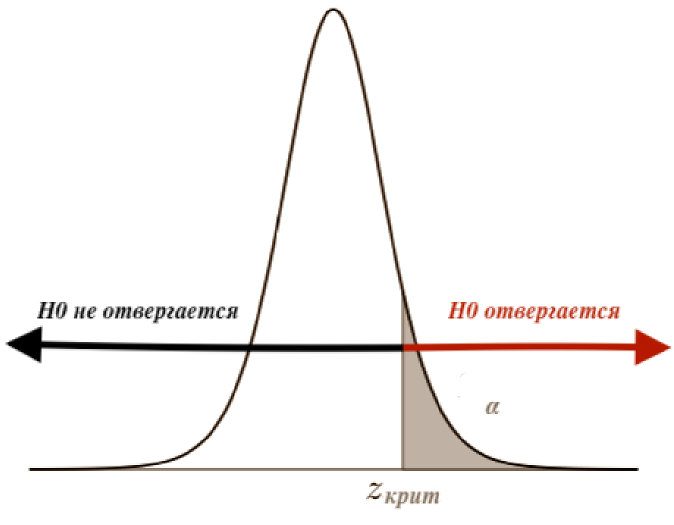
Через построение критической области

Критическая область – область отвержения нулевой гипотезы, то есть область, в которой лежат значения статистики критерия, нетипичные при условии, что нулевая гипотеза верна.

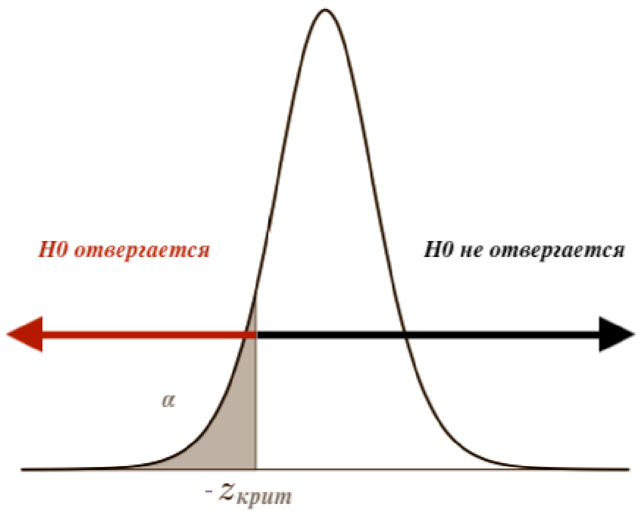
Вид критической области зависит от типа альтернативной гипотезы и выбранного уровня значимости α (уровень значимости – вероятность отвергнуть нулевую гипотезу при условии, что она верна, то есть вероятность некоторой ошибки).

Критические области в случае проверки гипотезы о доле:

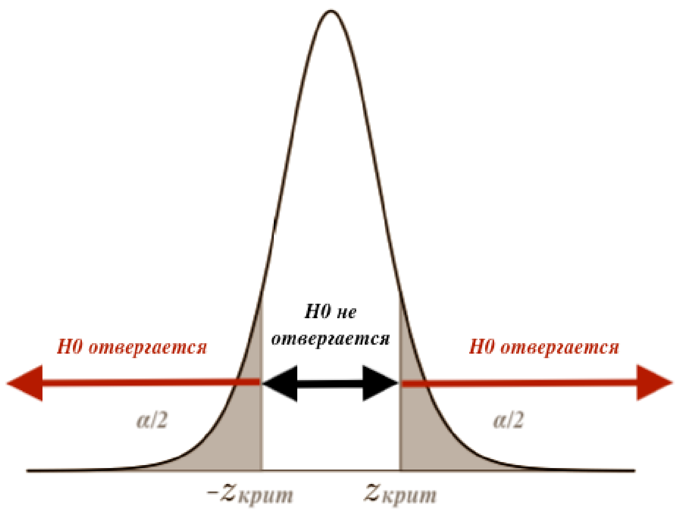
- Если альтернативная гипотеза правосторонняя, то область нетипичных значений лежит справа. Критическое значение статистики, отделяет область не-отвержения нулевой гипотезы от критической области: $z_{\text{крит}} = z_{1-\alpha}$.



- Если альтернативная гипотеза левосторонняя, то область нетипичных значений лежит слева. Критическое значение статистики, отделяет область не-отвержения нулевой гипотезы от критической области: $z_{\text{крит}} = -z_{1-\alpha}$.



- Если альтернативная гипотеза двусторонняя, то область нетипичных значений состоит из двух симметричных «хвостов» распределения:



  Критические значение статистики, отделяют область не-отвержения нулевой гипотезы от критической области: $z_{\text{крит}} = z_{1-\frac{\alpha}{2}}$ и $-z_{\text{крит}} = -z_{1-\frac{\alpha}{2}}$

Критические области в случае проверки гипотезы о среднем (уже без графиков, они похожи на предыдущие, распределение Стьюдента тоже симметрично относительно 0):

- Если альтернативная гипотеза правосторонняя, то область нетипичных значений лежит справа. Критическое значение статистики, отделяет область не-отвержения нулевой гипотезы от критической области: $t_{\text{крит}} = t_{1-\alpha,\ \text{df}=n-1}$.

- Если альтернативная гипотеза левосторонняя, то область нетипичных значений лежит слева. Критическое значение статистики, отделяет область не-отвержения нулевой гипотезы от критической области: $t_{\text{крит}} = -t_{1-\alpha,\ \text{df}=n-1}$.

- Если альтернативная гипотеза двусторонняя, то область нетипичных значений состоит из двух симметричных «хвостов» распределения. Критические значение статистики, отделяют область не-отвержения нулевой гипотезы от критической области: $t_{\text{крит}} = t_{1-\frac{\alpha}{2},\,\text{df}=n-1}$ и $-t_{\text{крит}} = t_{1-\frac{\alpha}{2},\,\text{df}=n-1}$.

Итог: если наблюдаемое значение статистики попадает в критическую область, то нулевая гипотеза отвергается, если нет – не отвергается.

Через p-value

Значение p-value – это вероятность получить значение статистики критерия равное наблюдаемому или более нетипичное по сравнению с наблюдаемым при условии, что нулевая гипотеза верна. Более неформально, p-value – это «жизнеспособность» нулевой гипотезы, которую мы оцениваем по имеющимся данным.

Рассмотрим пример. У нас есть знакомый, который вдруг почувствовал себя нехорошо. Наша нулевая гипотеза заключается в том, что знакомый не болен, а просто переутомился перед сессией. Мы собираем различные данные, например, температуру тела, давление, другие симптомы, которые можно измерить количественно. По результатам анализа этих данных мы делаем вывод о том, что знакомый просто переутомился. Насколько мы оказались правы? Вероятность того, что значения показателей о здоровье, которые мы получили, действительно объясняются тем, что наш знакомый не болен, и есть p-value.

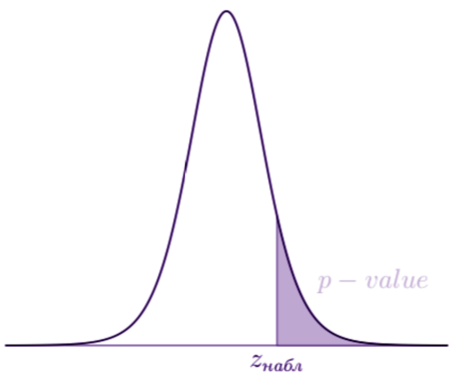
Что нам дают эти сведения? Во-первых, понимание, что p-value – это вероятность, причем не простая, а условная. Поэтому значения p-value всегда будут принадлежать интервалу $[0, 1]$. Во-вторых, так как p-value – это вероятность того, что нулевая гипотеза жизнеспособна, чем выше p-value, тем лучше, если мы хотим, чтобы нулевая гипотеза не была отвергнута.

Расчет p-value зависит от того, какого типа альтернативная гипотеза. Рассмотрим все случаи на примере z-статистики, которая имеет стандартное нормальное распределение.

Вычисление p-value в случае проверки гипотезы о доле:

- Если альтернативная гипотеза правосторонняя, то область нетипичных значений лежит справа, поэтому нас интересует площадь «хвоста» справа от наблюдаемого значения:

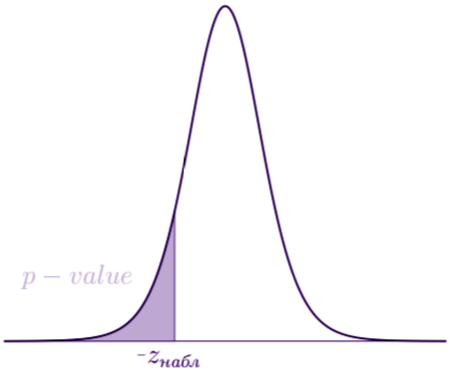
$$\text{p-value} = \mathrm{P}(z \geqslant z_{\text{набл}}) = 1 - \mathrm{P}(z < z_{\text{набл}}) = 1 - \Phi(z_{\text{набл}})$$



- Если альтернативная гипотеза левосторонняя, то область нетипичных значений лежит слева, поэтому нас интересует площадь «хвоста» слева от наблюдаемого значения:

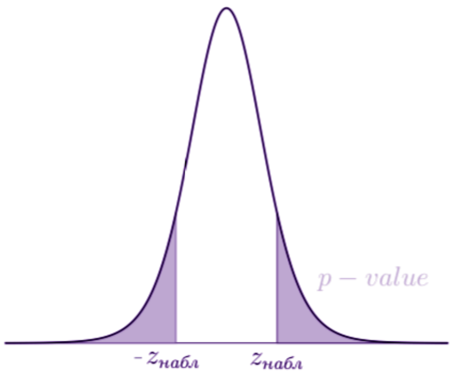
$$\text{p-value} = \mathrm{P}(z \leqslant -z_{\text{набл}}) = 1 - \Phi(z_{\text{набл}})$$

Почему это так? Левостороннюю альтернативу мы выбираем в случае, когда есть основания считать, что истинное значение параметра будет меньше значения, указанного в нулевой гипотезе. Когда такие основания появляются? Когда выборочная оценка параметра (например, доля, посчитанная по выборке) меньше значения, зафикисированного в нулевой гипотезе. Все ещё более нетипичные значения статистики находятся в левом «хвосте»:



- Если альтернативная гипотеза двусторонняя, то область нетипичных значений состоит из двух симметричных «хвостов» распределения:

$$\text{p-value} = \mathrm{P}(|z| \geqslant z_{\textit{набл}}) = 2\mathrm{P}(z \geqslant z_{\textit{набл}}) = 2(1 - \Phi(z_{\textit{набл}}))$$



Как по p-value определить, есть ли основания отвергнуть нулевую гипотезу? Тут важно сначала зафиксировать уровень значимости α, а потом уже делать выводы. Уровень значимости α – это вероятность отвергнуть нулевую гипотезу при условии, что она верна. P-value – это минимальный уровень значимости, на котором нулевая гипотеза может быть отвергнута. Соответственно, если p-value меньше нашего фиксированного уровня значимости, на котором мы проверяем гипотезу, то нулевую гипотезу следует отвергнуть, если более – то отвергать нулевую гипотезу оснований нет. Если вдруг получилось, что p-value совпало с уровнем значимости (а это бывает довольно редко), то в таких случаях поступают на усмотрение исследователя.

Итак, получается:

- p-value $< \alpha \Rightarrow H_0$ отвергаем на уровне значимости α, на имеющихся данных
- p-value $> \alpha \Rightarrow H_0$ не отвергаем на уровне значимости α, на имеющихся данных

Важно: в выводе относительно отвержения / не-отвержения нулевой гипотезы необходимо указывать уровень значимости, так как от этого зависит результат. Так, например, в случае, если p-value равно 0.02, у нас есть основания отвергнуть нулевую гипотезу на уровне значимости 5% ($0.02 < 0.05$), и нет оснований отвергнуть ее на уровне значимости 1% ($0.02 > 0.01$).

4. Сделать статистический и содержательный вывод.

Пример статистического вывода: на имеющихся данных, на уровне значимости 5% (уровне доверия 95%) есть основания/нет оснований отвергнуть нулевую гипотезу в пользу альтернативы.

Пример содержательного вывода: среднее значение роста женщин не равно 168 см.

NB: Важно всегда указывать уровень значимости (уровень доверия), на котором проверяется гипотеза, так как без этого уточнения выводы о нулевой гипотезе не имеют большого смысла: на одном уровне значимости гипотеза может быть отвергнута, а при выборе другого уровня значимости – нет. Желательно также прописывать, что выводы делаются на имеющихся данных, так как мы можем отвечать только за те результаты, которые получили по той выборке / выборкам, которые у нас есть, а не за «истинность» выводов вообще.

Важно! По результатам проверки статистической гипотезы мы никогда не делаем вывод о том, что нулевая гипотеза верна / должна быть принята. Вопрос об истинности нулевой гипотезы – содержательный вопрос, и если он и проверяется статистически, то с помощью более продвинутых методов и в рамках специально продуманного дизайна исследования. Всё, что мы можем решить по итогам проверки: отвергнуть нулевую гипотезу или нет. Как из того, что события не независимы, автоматически не следует их зависимость, так и из того, что нулевая гипотеза не отвергается, не следует, что она принимается.