Теория вероятностей и математическая статистика

Вебинар 5

Проверка статистических гипотез. P-значения. Доверительные интервалы

Проверка статистических гипотез

Статистические гипотезы

Статистическая гипотеза — предположение о виде распределения и свойствах случайной величины, которое можно подтвердить или опровергнуть на основании имеющихся данных.

Например,

1. Гипотеза: математическое ожидание случайной величины равно 10.
2. Гипотеза: случайная величина имеет нормальное распределение.
3. Гипотеза: две случайные величины имеют одинаковое математическое ожидание.

Нулевая и альтернативная гипотезы

Проверяемую гипотезу принято называть *нулевой* и обозначать H_0.

Пример. Имеется станок, изготавливающий шарики для подшипников, который настроен делать шарики с диаметром 1 мм. На основании выборки из значений диаметров таких шариков мы можем проверить, правильно ли станок откалиброван (т.е. делает ли он такие шарики, которые он настроен делать).

В таком случае в качестве нулевой гипотезы H_0 берётся гипотеза о том, что математическое ожидание диаметра шарика равно 1 мм.

Нулевая и альтернативная гипотезы

Проверяемую гипотезу принято называть *нулевой* и обозначать H_0.

Пример. Имеется станок, изготавливающий шарики для подшипников, который настроен делать шарики с диаметром 1 мм. На основании выборки из значений диаметров таких шариков мы можем проверить, правильно ли станок откалиброван (т.е. делает ли он такие шарики, которые он настроен делать).

В таком случае в качестве нулевой гипотезы H_0 берётся гипотеза о том, что математическое ожидание диаметра шарика равно 1 мм.

Параллельно с нулевой гипотезой рассматривается противоречащая ей гипотеза H_1, называемая *альтернативной* или *конкурирующей*.

В нашем примере в качестве альтернативной гипотезы будет выступать гипотеза о том, что математическое ожидание диаметра шарика не равно 1 мм.

Альтернативные гипотезы

В зависимости от задачи альтернативные гипотезы бывают *левосторонние*, *правосторонние* или *двухсторонние*.

Например, в примере выше альтернативная гипотеза двухсторонняя, поскольку в соответствии с ней математическое ожидание может быть как больше 1, так и меньше.

Односторонние гипотезы, возникают, например, при проверке нулевой гипотезы о равенстве математических ожиданий. В таких случаях используются функции, измеряющие *уровень отличия* двух значений (подробнее на занятии 7), которые равны 0, если нулевая гипотеза верна, и больше 0, если гипотеза неверна.

В любом случае альтернативная гипотеза представляет собой *дополнение к нулевой* (т.е. хотя бы одна из них всегда верна).

Проверка статистических гипотез

Проверяя нулевую гипотезу, мы по выборке считаем некоторое значение (зависит от вида гипотезы) и сравниваем его с теоретическим.

При проверке данной гипотезы наша задача установить, какое отклонение полученного значения от теоретического мы можем принять как допустимое (т. е. *незначимое*), а какое отклонение уже нельзя списать на случайность (*значимое*).

Если отклонение значимо, то гипотеза H_0 отвергается в пользу альтернативной. Если же отклонение не является значимым, то гипотеза H_0 остаётся в силе.

Этапы проверки гипотез

1. Формулируются нулевая и альтернативная гипотезы.
2. Задаётся некоторая *статистика* (функция от выборки) $S(X)$, которая в условиях справедливости нулевой гипотезы H_0 имеет *известное распределение* (в частности, известна её функция распределения $F_S(x) = P(S < x)$).
3. Фиксируется *уровень значимости* α — допустимая для данной задачи *вероятность ошибки первого рода* (чаще всего 0.01, 0.05 или 0.1).
4. Определяется *критическая область* Ω_α, такая, что $P(S \in \Omega_\alpha | H_0) = \alpha$.
5. Проводится *статистический тест*: для конкретной выборки X считается значение $S(X)$, и если оно принадлежит Ω_α, то заключаем, что данные противоречат гипотезе H_0, и принимается гипотеза H_1.

Выбор статистики S

Выбор статистики S зависит от того, какого рода гипотеза проверяется, какое распределение имеется и что нам известно.

Например, если проверяется гипотеза относительно математического ожидания нормально распределённой случайной величины с *известной* дисперсией, то используется т.н. *Z-статистика*:

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}},$$

где X — выборка, $\overline{X}$ — выборочное среднее, μ — утверждаемое гипотезой H_0 значение математического ожидания, σ — известное среднее квадратическое отклонение, n — число элементов в выборке.

В предположении вероности гипотезы H_0 статистика Z имеет *стандартное нормальное распределение* (т.е. нормальное распределение с параметрами $\mu = 0$, $\sigma = 1$).

Выбор статистики S

Выбор статистики S зависит от того, какого рода гипотеза проверяется, какое распределение имеется и что нам известно.

Например, если проверяется гипотеза относительно математического ожидания нормально распределённой случайной величины с *известной* дисперсией, то используется т.н. *Z-статистика*:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

где X — выборка, $\overline{X}$ — выборочное среднее, μ — утверждаемое гипотезой H_0 значение математического ожидания, σ — известное среднее квадратическое отклонение, n — число элементов в выборке.

В предположении вероности гипотезы H_0 статистика Z имеет *стандартное нормальное распределение* (т.е. нормальное распределение с параметрами $\mu = 0$, $\sigma = 1$).

Замечание. Условие известной дисперсии, конечно, звучит крайне странно с точки зрения реальных примеров, где нам дана только выборка и всё. Но бывает и такое!

Выбор статистики S

Если же дисперсия неизвестна, используется несколько другая *t-статистика*:

$$t = \frac{\overline{X} - \mu}{\sigma_X / \sqrt{n}},$$

где σ_X — несмещённая оценка среднего квадратического отклонения.

В предположении вероности гипотезы H_0 t-статистика имеет *распределение Стьюдента* или *t-распределение* с параметром $df = n - 1$.

Выбор статистики S

Если же дисперсия неизвестна, используется несколько другая *t-статистика*:

$$t = \frac{\overline{X} - \mu}{\sigma_X / \sqrt{n}},$$

где σ_X — несмещённая оценка среднего квадратического отклонения.

В предположении вероности гипотезы H_0 t-статистика имеет *распределение Стьюдента* или *t-распределение* с параметром $df = n - 1$.

Замечание. Если известна дисперсия, то известно и среднее квадратическое отклонение. Наоборот, если известно среднее квадратическое отклонение, то известная и дисперсия.

Выбор статистики S

Для проверки гипотез о дисперсии нормального распределения используют квантили *распределения хи-квадрат*.

Для большинства других распределений соответствующие статистики *не разработаны*, поэтому по-честному применять такие статистические тесты для них нельзя. В таких случаях с можно обойти проблему с помощью Центральной предельной теоремы, однако, тогда результаты будут иметь некоторую погрешность.

Ошибки первого и второго рода

Ошибки первого и второго рода возникают в задачах, в которых требуется определить, произошло какое-то событие или нет.

Ошибка первого рода (т.н. *false positive*) соответствует ситуации, когда было определено, что событие произошло, тогда как реально оно не произошло.

Ошибка второго рода (т.н. *false negative*) — обратная ситуация: мы определили, что событие не произошло, а реально оно произошло.

Ошибки первого и второго рода

Например, на входе в каждый аэропорт стоит рамка-металлодетектор. Её задача — выявлять людей, которые пытаются пронести в аэропорт что-то опасное.

В этом случае:

- ошибка первого рода: рамка сработала на человеке, который ничего опасного не несёт,
- ошибка второго рода: рамка не сработала на человеке, который несёт что-то опасное.

Ошибки первого и второго рода

Например, на входе в каждый аэропорт стоит рамка-металлодетектор. Её задача — выявлять людей, которые пытаются пронести в аэропорт что-то опасное.

В этом случае:

- ошибка первого рода: рамка сработала на человеке, который ничего опасного не несёт,
- ошибка второго рода: рамка не сработала на человеке, который несёт что-то опасное.

Как правило, между вероятностями ошибок первого и второго рода *приходится балансировать*, т.е. уменьшение одной вероятности приводит к увеличению другой.

Например, в примере с рамкой вероятность ошибки первого рода будет невероятно высокой, поскольку ошибки второго рода в этом случае недопустимы.

Уровень значимости

В контексте проверки статистических гипотез под *событием* мы понимаем то, что *нулевая гипотеза была отвергнута*.

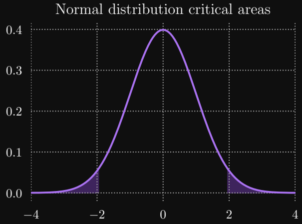
Уровень значимости — это вероятность ошибки первого рода, т.е. вероятность отвергнуть верную нулевую гипотезу.

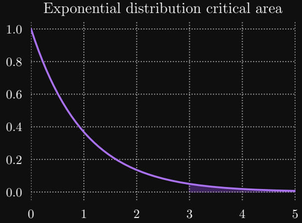
Низкие значения уровня значимости α делают проводимый тест *более осторожным*: вероятность ошибки первого рода мала, значит, если по результатам теста мы отвергаем нулевую гипотезу, то мы можем быть более уверены в том, что она и правда неверна.

Однако, вместе с тем повышается и вероятность не отвергнуть ложную нулевую гипотезу. Т.е. если нулевая гипотеза не была отвергнута, то далеко не факт, что она и впрямь верна.

Критическая область

Под *критической областью* случайной величины X, как правило, понимают её «хвосты», т.е. области, в которые чаще всего случайная величина не попадает.





Например, для нормального распределения справедливо правило двух сигм: вероятность попасть в интервал от $\mu - 2\sigma$ до $\mu + 2\sigma$ равна 0.95. Следовательно, можно рассмотреть критическую область $(-\infty, \mu - 2\sigma) \cup (\mu + 2\sigma, \infty)$, соответствующую вероятности 0.05.

Такая критическая область является *двухсторонней*, поскольку содержит два «хвоста».

Если же случайная величина имеет только один «хвост» (например, экспоненциальное распределение), то и критическая область будет *односторонней*.

Критическая область

При проведении статистического теста мы строим критическую область для статистики $S(X)$, построенной ранее. Мы можем это сделать, потому что знаем распределение этой статистики, в частности, её функцию распределения $F_S(x)$.

Кроме того, к этому моменту мы уже зафиксировали уровень значимости α. Это значение является вероятностью попасть в критическую область.

Итак, в зависимости от количества «хвостов» у распределения статистики $S(X)$, критическая область может быть одного из следующих видов:

- Левосторонняя область: $\Omega_\alpha = (-\infty, t_\alpha)$.
- Правосторонняя область: $\Omega_\alpha = (t_{1-\alpha}, \infty)$.
- Двусторонняя область: $\Omega_\alpha = \left(-\infty, t_{\alpha/2}\right) \cup \left(t_{1-\alpha/2}, \infty\right)$.

Здесь t_β обозначает квантиль порядка β, т.е. $F_S(t_\beta) = \beta$.

Статистический тест

Итак, осталось лишь провести статистический тест. У нас есть:

1. зафиксированная нами статистика $S(X)$,
2. построенная нами критическая область Ω_α.

Считаем значение статистики от нашей выборки. Если это значение попадает в критическую область, то заключаем, что данные противоречат нулевой гипотезе, и её следует отвергнуть.

Если значение в критическую область не попало, то заключаем, что данные нулевой гипотезе не противоречат.

Статистический тест

Итак, осталось лишь провести статистический тест. У нас есть:

1. зафиксированная нами статистика $S(X)$,
2. построенная нами критическая область Ω_α.

Считаем значение статистики от нашей выборки. Если это значение попадает в критическую область, то заключаем, что данные противоречат нулевой гипотезе, и её следует отвергнуть.

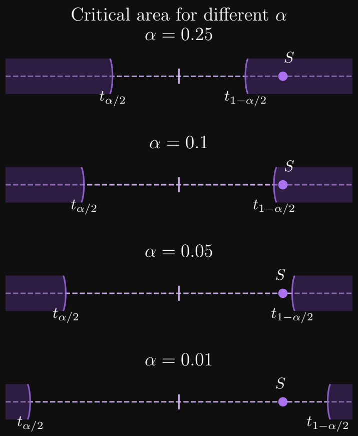
Если значение в критическую область не попало, то заключаем, что данные нулевой гипотезе не противоречат.

Замечание. То, что данные не противоречат нулевой гипотезе, не означает, что она верна.

P-значения

P-значения позволяют получить результат проверки статистических гипотез сразу для многих уровней значимости.

Как мы теперь знаем, уменьшение уровня значимости приводит к расширению границ критической области: чем меньше уровень значимости, тем сложнее попасть в критическую область и, как следствие, отвергнуть нулевую гипотезу. *P-значение* представляет собой наибольшее значение уровня значимости α, при котором гипотезу можно принять, т.е. при котором значение статистики, посчитанной по выборке, ещё не попадает в критическую область.



P-значение — это такое значение α, при котором значение статистики попадает ровно на границу критической области.

Пусть $F_S(x)$ — функция распределения рассматриваемой статистики, а t_β — квантиль порядка β для этого распределения. Как считать P-значение:

1. Для правосторонней области $\Omega_\alpha = (t_{1-\alpha}, \infty)$ имеем условие $t_{1-\alpha} = S$, откуда
$$P_r = 1 - F_S(S)$$

P-значения

P-значение — это такое значение α, при котором значение статистики попадает ровно на границу критической области.

Пусть $F_S(x)$ — функция распределения рассматриваемой статистики, а t_β — квантиль порядка β для этого распределения. Как считать P-значение:

1. Для правосторонней области $\Omega_\alpha = (t_{1-\alpha}, \infty)$ имеем условие $t_{1-\alpha} = S$, откуда
$$P_r = 1 - F_S(S)$$

2. Для левосторонней области $\Omega_\alpha = (-\infty, t_\alpha)$, условие $t_\alpha = S$, откуда
$$P_l = F_S(S)$$

P-значения

P-значение — это такое значение α, при котором значение статистики попадает ровно на границу критической области.

Пусть $F_S(x)$ — функция распределения рассматриваемой статистики, а t_β — квантиль порядка β для этого распределения. Как считать P-значение:

1. Для правосторонней области $\Omega_\alpha = (t_{1-\alpha}, \infty)$ имеем условие $t_{1-\alpha} = S$, откуда
$$P_r = 1 - F_S(S)$$
2. Для левосторонней области $\Omega_\alpha = (-\infty, t_\alpha)$, условие $t_\alpha = S$, откуда
$$P_l = F_S(S)$$
3. Для двухсторонней области $\Omega_\alpha = \left(-\infty, t_{\alpha/2}\right) \cup \left(t_{1-\alpha/2}, \infty\right)$ нужна комбинация двух:
$$P = 2 \cdot \min(P_l, P_r)$$

Р-значения

На практике использовать Р-значения можно так: если выбранный нами уровень значимости меньше Р-значения, полученного в результате теста, то гипотезу можно принять. В противном случае, гипотезу следует отвергнуть.

Доверительные интервалы

Доверительные интервалы

Ранее мы познакомились со способами оценивать параметры распределения по выборке. Всё это были *точечные* оценки, т.е. мы оценивали параметр каким-то единственным числом.

Минус таких оценок в том, что у нас нет возможности понять, насколько хорошей такая оценка получилась (т.е. насколько близко мы оказались к реальному значению оцениваемого параметра).

Доверительные интервалы

Ранее мы познакомились со способами оценивать параметры распределения по выборке. Всё это были *точечные* оценки, т.е. мы оценивали параметр каким-то единственным числом.

Минус таких оценок в том, что у нас нет возможности понять, насколько хорошей такая оценка получилась (т.е. насколько близко мы оказались к реальному значению оцениваемого параметра).

Для исправления этого недостатка используют доверительные интервалы. *Доверительный интервал* — это интервал, который с некоторой вероятностью (заданной заранее) содержит значение оцениваемого параметра.

Более строго: пусть задано число p, называемое *уровнем доверия* или *доверительной вероятностью*. Доверительным интервалом для параметра θ называется пара статистик L и U, таких, что

$$P(L \leq \theta \leq U) = p$$

Доверительные интервалы

Доверительные интервалы используют ту же математическую базу, что и статистические тесты.

Например, пусть дана выборка X из *нормально распределённой* случайной величины с *известной дисперсией* σ^2, и требуется построить доверительный интервал для математического ожидания μ с доверительной вероятностью p. Мы знаем, что в этом случае статистика

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

имеет стандартное нормальное распределение.

Доверительные интервалы

Обозначим $\alpha = 1 - p$. Можно убедиться в том, что

$$P\left(t_{\alpha/2} \leq Z \leq t_{1-\alpha/2}\right) = p,$$

где t_β — квантиль порядка β для стандартного нормального распределения. Подставляя сюда Z, получаем

$$P\left(t_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq t_{1-\alpha/2}\right) = p$$

$$P\left(t_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq t_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = p$$

$$\boxed{P\left(\overline{X} + t_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = p}$$

В случае *неизвестной дисперсии* мы используем статистику

$$t = \frac{\overline{X} - \mu}{\sigma_X / \sqrt{n}},$$

где σ_X — выборочное среднее квадратическое отклонение. Эта статистика имеет распределение Стьюдента, поэтому

$$P\left(t_{\alpha/2,\, n-1} \leq t \leq t_{1-\alpha/2,\, n-1}\right) = p,$$

где $t_{\beta,\, n-1}$ — квантиль порядка β для распределения Стьюдента с параметром $df = n - 1$. Аналогичным способом получаем доверительный интервал:

$$P\left(\overline{X} + t_{\alpha/2,\, n-1} \cdot \frac{\sigma_X}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{1-\alpha/2,\, n-1} \cdot \frac{\sigma_X}{\sqrt{n}}\right) = p$$

Доверительные интервалы

Для построения *доверительных интервалов для дисперсии* пользуются распределением хи-квадрат.

Как и в случае со статистическими тестами, для большинства иных распределений точные доверительные интервалы построить нельзя, но можно построить приблизительные с помощью Центральной предельной теоремы.

На следующем занятии

Взаимосвязь величин. Показатели корреляции.
Корреляционный анализ. Проверка на нормальность