Практическое занятие № 6. Статистические гипотезы. Непараметрические методы

Теоретические сведения

Статистическими гипотезами называются утверждения о распределениях или количественных признаках в генеральных совокупностях, выдвигаемые и проверяемые на основе выборочных данных.

Выдвинутая гипотеза H_0 называется *нулевой* или *основной*, а противоречащая ей H_1 — *альтернативной*.

Пример гипотез. Пусть $H_0 : \theta = \theta_1$, тогда возможны следующие альтернативные варианты $H_1 : \theta < \theta_0$, $H_1 : \theta > \theta_0$, $H_1 : \theta \neq \theta_0$.

Если основная гипотеза H_0 отвергается, то делается вывод, что выборочные наблюдения противоречат основной гипотезе. Если же H_0 принимается, то выборочные данные могли быть получены из генеральной совокупности со свойствами, указанными в H_0, что не означает, что генеральная совокупность действительно имеет эти свойства (обычно используется формулировка – отвергать гипотезу нет оснований).

При принятии или отклонении гипотезы возможны 4 различные ситуации.

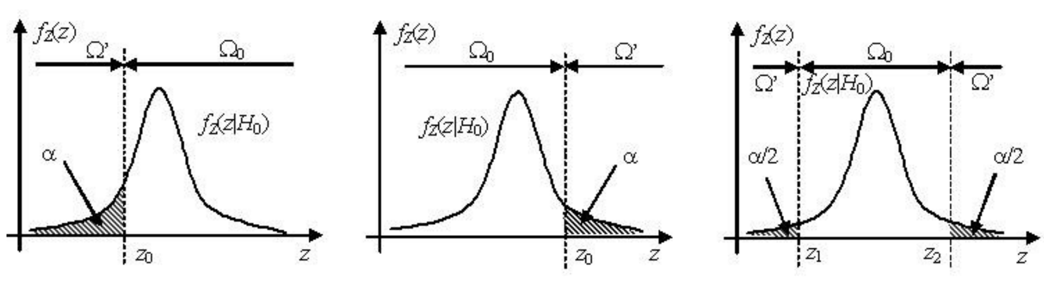
1. Ошибка 1-го рода: отклонена правильная гипотеза. Вероятность такого исхода равна α.
2. Ошибка 2-го рода: Принята неправильная гипотеза. Вероятность такого исхода β.
3. Принята правильная гипотеза α−1.
4. Отклонена неправильная гипотеза β−1.

При проверке гипотез, прежде всего, задаются вероятностью совершения ошибки 1-го рода α, которая называется *уровнем значимости гипотезы*. Чем выше α, тем выше требования к основной гипотезе. Уровень значимости принимают: 0.05, 0.01, 0.001. Сам критерий для заданного уровня значимости α выбирается так, чтобы вероятность ошибки 2-го рода β была минимальной. Вероятность отклонения неправильной основной гипотезы (1–β) называется *мощностью критерия*. Из всех возможных критериев с заданным уровнем значимости α выбирается наиболее мощный.

Таким образом, вся область значений делится на критическую область, где H_0 отвергается, и область принятия гипотезы, где отвергать H_0 нет оснований. Уровень значимости α определяет ширину критической области. *Критическая область* – это область маловероятных значений статистики критерия в хвостах распределения.

Существует 3 вида критических областей: правосторонняя, левосторонняя и двусторонняя. Если допустить, что основная гипотеза верна, то вероятность попадания критерия в критическую область есть вероятность ошибки 1-го рода α. В случае двусторонней критической области площади каждого из хвостов, как правило, выбираются равными.

То есть в условиях сформулированной гипотезы необходимо выбрать для нее альтернативную и вид статистики (выбирается из справочников или документации), для выбранной статистики построены виды распределений (тоже есть в справочниках). Заданный уровень α определяет вид критической области. Напомним, что площадь под кривой функции распределения плотности вероятностей равна единице, соответственно, мы ищем области в хвостах распределения, площадь которых равна половине α. Например, при $\alpha = 0{,}1$ при двусторонней критической области, площадь будет равна 1% $(=(1 - 0{,}1) \times 100\%)$. Таким образом, можем использовать перцентиль, или квантиль, рассчитанный для каждого распределения (соответствующие методы есть в каждой статистической функции Python).



Типы критических областей: левосторонняя, правосторонняя, двусторонняя

Сформулируем в общем виде метод проверки статистических гипотез, который состоит из 7 шагов.

1) Формирование гипотез H_0 и альтернативной гипотезы H_1.

2) Выбрать уровень значимости α при принятии гипотезы H_0.

3) Выбрать статистику критерия Z для проверки гипотезы H_0. (Для большинства встречающихся на практике статистических гипотез H_0 статистики Z найдены, нужно найти ее в справочниках и учебниках!).

4) Посмотреть закон распределения $f_Z(z \mid H_0)$ статистики Z при условии истинности гипотезы H_0 (законы распределения известны заранее, в классических учебниках по статистике представлены в виде таблиц).

5) Построить область допустимых значений Ω_0 и критическую область Ω_1.

6) Вычислить выборочное значение статистики критерия z на основе имеющихся выборочных наблюдений.

7) Принять статистическое решение, используя решающее правило: если выборочное значение статистики критерия $z \in \Omega_0$, то основная гипотеза H_0 принимается, если выборочное значение статистики критерия $z \in \Omega_1$, то основная гипотеза H_0 отвергается в пользу альтернативной гипотезы H_1.

Функции проверки статистических гипотез в Python как правило, возвращают выборочное значение z статистики и p-value. Здесь p-value – это площадь под графиком функции плотности распределения статистики критерия, расположенная левее/правее выборочного значения статистики критерия z. То есть для конкретной выборки считается z и считается площадь закрашенной области для этой выборки. Иногда удобно использовать это значение, соответственно, если площадь (напомним, что это вероятность) меньше, чем заданное α, то гипотеза отвергается.

Приведенному выше методу соответствуют все виды проверки гипотез. Приведем некоторые из них в табл. 1

Пусть заданы следующие условия. Пусть имеется выборка $X = \{x_1, .., x_n\}$ размерности n, из нормально распределенной генеральной совокупности с распределением $N(\mu, \sigma)$ (Или задано аналогичным образом две выборки X_1 и X_2 из нормально распределенной генеральной совокупности с параметрами математического ожидания μ_1, μ_2 и дисперсиями σ_1, σ_2). В этих условиях справедливы следующие статистики.

Таблица 1. Проверка гипотез о параметрах нормально распределенной генеральной совокупности и нормально распределенных величинах

№	Формулировка нулевой гипотезы	Название/Вид H_0	Статистика	Закон распределения
1.				
2.	Гипотеза о значении математического ожидания при известной дисперсии	One-sampled z-test $H_0 : \mu = \mu_0$; $\sigma = \text{const}$	$Z = \dfrac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$	Нормальный $N(0, 1)$
3.	Гипотеза о значении математического ожидания при неизвестной дисперсии	One-sample t-test $H_0 : \mu = \mu_0$	$Z = \dfrac{\bar{X} - \mu_0}{S / \sqrt{n}}$	Стьюдента с n–1 степенями свободы $T(n-1)$
4.	Гипотеза о значении дисперсии при известном математическом ожидании	Chi-squared test $H_0 : \sigma = \sigma_0$; $\mu = \text{const}$	$Z = \dfrac{nS_0^2}{\sigma_0^2}$	Хи-квадрат $\chi^2(n)$
5.	Гипотеза об оценки дисперсии при известном математическом ожидании	Chi-squared test $H_0 : \sigma = S_0$; $\mu = \text{const}$	$S_0^2 = \dfrac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2$	Хи-квадрат $\dfrac{\sigma^2}{n}\chi^2(n)$
6.	Гипотеза о значении дисперсии при неизвестном математическом ожидании	Chi-squared test $H_0 : \sigma = \sigma_0$	$Z = \dfrac{(n-1)S^2}{\sigma_0^2}$	Хи-квадрат с n–1 степенями свободы $\chi^2(n-1)$
7.	Гипотеза о равенстве математических ожиданий при известных	Two-sample z-test) $H_0 : \mu_1 = \mu_2$; $\sigma_1, \sigma_2 = \text{const}$	$Z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_2^2 / n}}$	Стандартизир. нормальное $N(0, 1)$

	дисперсиях			
8.	Гипотеза о равенстве дисперсий при известных математических ожиданиях	Two-sample F-test $H_0 : \sigma_1 = \sigma_2$ $\mu_1, \mu_2 - \text{const}$	$F_0 = \frac{S_{01}^2 / \sigma_1^2}{S_{02}^2 / \sigma_2^2}$	Распределение Фишера $F(n_1, n_2)$
9.	Гипотеза о равенстве дисперсий при неизвестных математических ожиданиях	Two-sample F-test). $H_0 : \sigma_1 = \sigma_2$	$F_0 = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$	Распределение Фишера $F(n_1 - 1, n_2 - 1)$
10.	Гипотеза о равенстве математических ожиданий при неизвестных равных дисперсиях	Two-sample unpooled t-test $H_0 : \mu_1 = \mu_2$ $\sigma_1 = \sigma_2$	$Z = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{1/n_1 + 1/n_2}}$, $S_p = \sqrt{\frac{S_1^2 + S_2}{2}}$	распределение Стьюдента с n_1+n_2-2 степенями свободы $T(n_1 + n_2 - 2)$

Реализация в Colab

Проиллюстрируем алгоритм на примере вычисления статистики для гипотезы ***о равенстве математических ожиданий при неизвестных равных дисперсиях,*** если заданный уровень значимости равен 0.2; объем двух выборок – 100; математическое ожидание равно 10; дисперсия – 5.

Нулевая гипотеза:

$$H_0 : \mu_1 = \mu_2 \text{ при неизвестных } \sigma_1 = \sigma_2 .$$

Для проверки гипотезы используется статистика

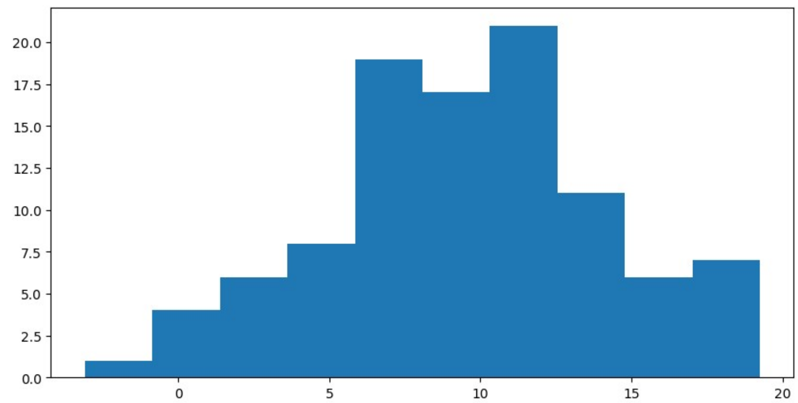
$$, Z = \frac{\bar{X}_1 - \bar{X}_2}{S_p \sqrt{1/n_1 + 1/n_2}}, \quad S_p = \sqrt{\frac{S_1^2 + S_2}{2}}$$

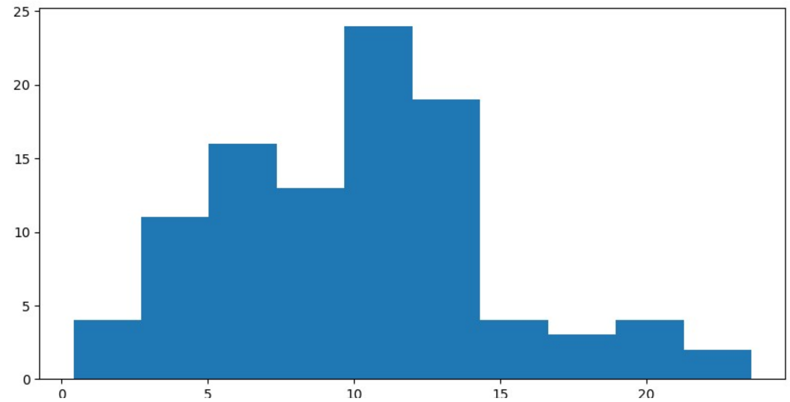
имеющая распределение Стьюдента с $(n_1 + n_2 – 2)$ степенями свободы:

$$P(Z) = T(n_1 + n_2 - 2).$$

В англоязычной литературе – Two-sample unpooled t-test.

```
import numpy as np
from scipy import stats
import matplotlib.pyplot as plt
import seaborn as sns
from numpy.ma.core import sqrt
from tqdm.notebook import tqdm

#исходные данные
mu = 10
sigma = 5
alpha_value = 0.3

rv_first = stats.norm(loc=mu, scale=sigma)
rv_second = stats.norm(loc=mu, scale=sigma)

number_of_samples = 100
margin = 0.0001

np.random.seed(42) # Инициализация генератора случайных чисел

size = number_of_samples
sample_first = rv_first.rvs(size=size) # Генерация выборки из 30 зн
ачений на основе первого распределения
sample_second = rv_second.rvs(size=size) # Генерация выборки из 30
значений на основе второго распределения

# Вывод гистограммы для первой выборки
plt.figure(figsize=(10, 5))
plt.hist(sample_first, bins=10)
plt.show()

# Вывод гистограммы для второй выборки
plt.figure(figsize=(10, 5))
plt.hist(sample_second, bins=10)
plt.show()
```

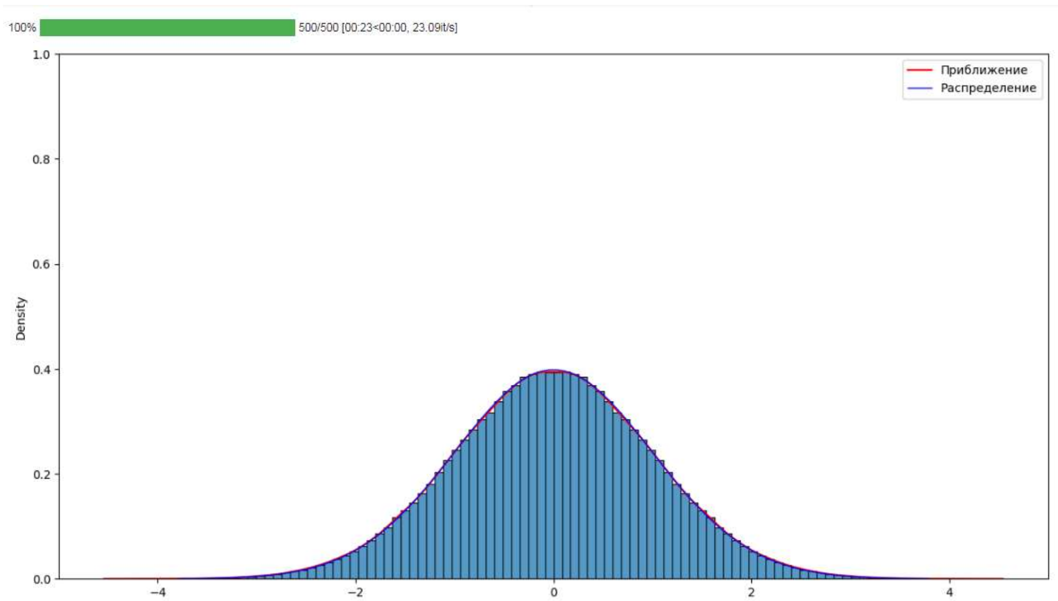




Сравнение экспериментальной и теоретической функций распределения заданной статистики

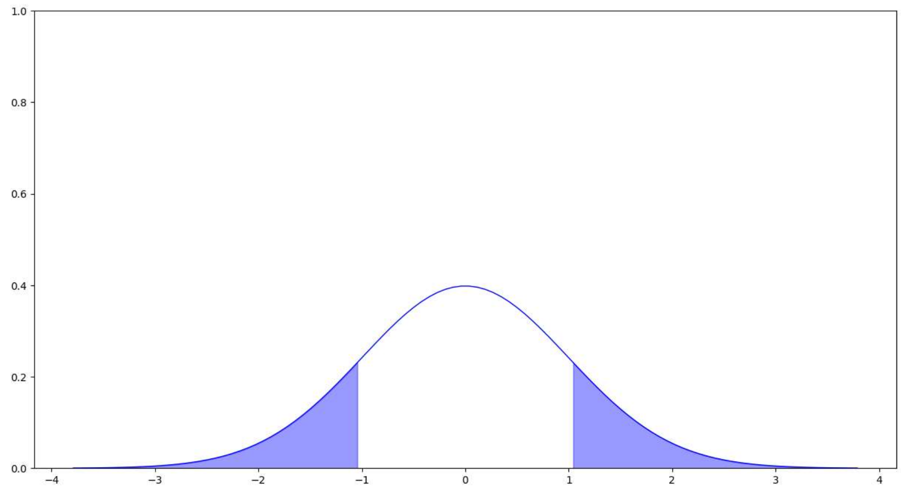
```
#задаем стастику в соответствии с задачей (гипотезой)
# a,b - выборки, n - мощность

def z_value_dm(a, b, n):
  z_val = (np.mean(a) -
  np.mean(b)) / (sqrt((np.var(a) + np.var(b)) / 2) * sqrt(2/n))
  return z_val

#Построим двумя способами распределение, которое соответствует стат
истике Z.
#Первый способ -
 строим экспериментально распределение Z, для этого генерируем 500
выборок мощностью 30 (n = 30), вычислим Z и построим ее гистограмму
#Второй способ -
 сразу строим теоретическое (известное) распределение Стьюдента T(n
_1+n_2 - 2). Здесь - n_1=n_2=30

fig, ax = plt.subplots(nrows=1, ncols=1, figsize = (15, 8))
plt.ylim((0,1))

n = 30
m = 500
```

```
z_stat = 0
data = []
for i in range(m):
  data.append(rv_first.rvs(size=n))

# Считаем распределение первм способом
calculated_stats = []
for i, a in tqdm(enumerate(data), total=m):
  for j, b in enumerate(data):
    if i != j:
       z_stat = z_value_dm(a, b, n)
       calculated_stats.append(z_stat)
sns.histplot(calculated_stats, ax=ax, bins=100, stat='density')
sns.kdeplot(calculated_stats, ax=ax, color='r', label='Приближение'
)


# Строим распределение вторым способом как Стуюдента, функция "t"
rv_theoretical = stats.t(df=(len(sample_first) + len(sample_second)
 - 2))

line_x = np.linspace(rv_theoretical.ppf(margin), rv_theoretical.ppf
(1 - margin), number_of_samples)
sns.lineplot(x=line_x, y=rv_theoretical.pdf(line_x), color='b', lw=
1, ax=ax, label='Распределение')

plt.legend()
plt.show()
```




```
# Покажем уровень значимости на функции распределения
# Закрашенная часть гистограммы в хвостах распределения -
 и есть критическая область, соответствующая alpha
```

```
# Если полученное на основании выборочных значений -
 Z_выборочное попадает в закрашеную область, гипотеза отклоняется
# в противном случае - отклонять нет оснований
# Здесь закрашенная область находит через метод ppf (Percent point
function) - квартили (процентили), обратная операция к методу cdf

fig, ax = plt.subplots(nrows=1, ncols=1, figsize = (15, 8))
plt.ylim((0,1))

n = number_of_samples
rv_theoretical = stats.t(df=(len(sample_first) + len(sample_second)
 - 2))

local_alpha_value = (alpha_value / 2)
left_vline_position = rv_theoretical.ppf(local_alpha_value)
right_vline_position = rv_theoretical.ppf(1 - local_alpha_value)

line_x = np.linspace(rv_theoretical.ppf(margin), rv_theoretical.ppf
(1 - margin), number_of_samples)
sns.lineplot(x=line_x, y=rv_theoretical.pdf(line_x), color='b', lw=
1, ax=ax)

x = np.linspace(rv_theoretical.ppf(margin), rv_theoretical.ppf(1 -
 margin), number_of_samples*10)
x_range = x[x<=left_vline_position]
ax.fill_between(x_range, rv_theoretical.pdf(x_range), np.zeros(len(
x_range)), color='b', alpha=0.4)

x = np.linspace(rv_theoretical.ppf(margin), rv_theoretical.ppf(1 -
 margin), number_of_samples*10)
x_range = x[x>=right_vline_position]
ax.fill_between(x_range, rv_theoretical.pdf(x_range), np.zeros(len(
x_range)), color='b', alpha=0.4)

plt.show()
```




```
def z_value_dm(a, b, n):
  z_val = (np.mean(a) -
  np.mean(b)) / (sqrt((np.var(a) + np.var(b)) / 2) * sqrt(2/n))
  return z_val


# здесь находим теоретическую кривую как с использование T-теста
rv_theoretical = stats.t(df=(len(sample_first) + len(sample_second)
 - 2))

z_stat, p = stats.ttest_ind(sample_first, sample_second)

# Считаем выборочную статистику, должна быть примерна равна теорети
ческой
z_value_d = z_value_dm(sample_first, sample_second, 100)

#  (np.mean(sample_first) -
  np.mean(sample_second)) / (sqrt((np.var(sample_first) + np.var(sa
mple_second)) / 2) * sqrt(2/n))


print("Z =", z_stat)
print("Z =", z_value_d)
print("P_лев_критическое =", left_vline_position)
print("P_прав_критическое= ", right_vline_position)
```

Таким образом, рассчитанная двумя способами z не попадает в критическую область, поэтому отвергать гипотезу нет оснований.

Полезные функции

Название	Описание
scipy.stats.ttest_ind	Т-критерий для гипотезы о том, что две независимые выборки имеют одинаковые матожидания и одинаковые неизвестные дисперсии
scipy.stats.ttest_rel	Т-критерий для двух связанных выборок при гипотезе о том, что две связанные или повторяющиеся выборки имеют одинаковые значения матожиданий
scipy.stats.chisquare	Тест хи-квадрат проверяет нулевую гипотезу о том, что данные имеют заданные параметры
scipy.stats.ttest_ind_from_stats	Т-критерий для средних значений двух независимых выборок при гипотезе о том, что две независимые выборки имеют одинаковые значений матожиданий
scipy.stats.ttest_1samp	Т-критерий для среднего значения при гипотезе о том, что матожидание выборки независимых наблюдений равно заданному генеральному среднему
scipy.stats.t	Т-распределение Стьюдента
scipy.stats.f	F-распределение Фишера
scipy.stats.chi2	Хи-квадрат распределение Пирсона

Варианты заданий.

Цель занятия: познакомиться с принципами решения задач проверки статистических гипотез. Понять, что такое нулевая гипотеза, статистика, метод проверки гипотез.

Студент получает вариант, состоящий из 3 цифр. Первая цифра – вид гипотезы из 2-го столбца, вторая и третья цифры – параметры выборки 1, параметры выборки 2, четвертая цифра – уровень значимости.

Необходимо:

1) построить экспериментальный и теоретический вид распределения, используемого для проверки гипотезы;

2) сгенерировать 2 выборки в соответствии с вариантами (размеры выборок – 25); проверить гипотезу в соответствии с заданным уровнем значимости (по вариантам).

№	Гипотеза	Выборка 1	Выборка 2	α
1.	Гипотеза о значении математического ожидания при известной дисперсии	$N(0,2)$	Лапласа(0,2)	0,2
2.	Гипотеза о значении математического ожидания при неизвестной дисперсии	$N(1,2)$	Лапласа(1,2)	0,1
3.	Гипотеза о значении дисперсии при известном математическом ожидании	$N(2,2)$	$N(0,2)$	0,05
4.	Гипотеза об оценки дисперсии при известном математическом ожидании	$N(10,2)$	$N(1,2)$	0,001
5.	Гипотеза о значении дисперсии при неизвестном математическом ожидании	$N(3,2)$	$N(3,2)$	0,25
6.	Гипотеза о равенстве математических ожиданий при известных дисперсиях	$N(4,2)$	$N(4,2)$	0,115
7.	Гипотеза о равенстве дисперсий при известных математических ожиданиях	$N(5,2)$	$N(5,2)$	0,12
8.	Гипотеза о равенстве дисперсий при неизвестных математических ожиданиях	$N(2,2)$	Вейбулла(2,2)	0,005
9.	Гипотеза о равенстве математических ожиданий при неизвестных равных дисперсиях	$N(4,2)$	Вейбулла (4,2)	0,3