Практическая работа № 8. Сравнение характеристик распределения двух генеральных совокупностей параметрическими методами

8.1. О содержании и задачах практической работы

В этой практической работе будем работать с парами выборок, принадлежащими генеральным совокупностям с однотипным законом распределения (нормальным или биномиальным). Нас будет интересовать вопрос о близости параметров распределений генеральных совокупностей.

Обсудим две задачи.

1. Сравнение математических ожиданий и сравнение дисперсий двух нормально распределенных генеральных совокупностей по выборкам.

2. Сравнение параметров двух биномиальных распределений по результатам двух серий испытаний.

О первой задаче. Можно выделить несколько практических задач, которые сводятся к проверке гипотез о равенстве параметров генеральных совокупностей. В частности, такая задача возникает при исследовании влияния, которое оказывает изменение некоторого фактора на измеряемую величину. Например, если измерения проводятся на двух приборах и нужно исследовать влияние фактора «прибора» на результаты измерений.

На данном этапе рассмотрим методы, которые «работают» только в условиях, когда известен тип закона распределения генеральных совокупностей, которым принадлежат опытные данные (мы ограничимся случаями нормального и биномиального распределений).

При изучении близости параметров законов двух генеральных совокупностей будем использовать два подхода. Первый подход предполагает следование стандартной схеме проверки гипотез (она применялась нами в практической работе № 6 для проверки гипотез о равенстве параметра распределения эталонному значению). Второй подход основан на

применении доверительных интервалов. Подходы взаимосвязаны, при этом одностороннему критерию значимости соответствует односторонний доверительный интервал, а двустороннему – двусторонний доверительный интервал.

Когда проверка гипотезы строится на использовании доверительных интервалов, нужно иметь в виду следующее. Если проверяется гипотеза $H_{0:}: m_1 = m_2$, то рассматривается доверительный интервал для разности $m_1 - m_2$. Гипотеза принимается, если он накрывает нулевое значение. Если проверяется гипотеза $H_{0:}: \sigma_1^2 = \sigma_2^2$, то рассматривается доверительный интервал для отношения дисперсий $\frac{\sigma_1^2}{\sigma_2^2}$. Гипотеза принимается, если этот интервал накрывает значение, равное единице.

Выбор статистики при проверке гипотезы о равенстве математических ожиданий зависит от ситуации с дисперсиями генеральных совокупностей (равны они или нет). Поэтому сравнение средних всегда предваряется проверкой гипотезы о равенстве дисперсий (см. пп. 3 и 4 раздела 8.2).

О второй задаче. При проверке гипотез о равенстве вероятностей событий в двух сериях испытаний важно учитывать характеристики выборок, поскольку они влияют на корректный выбор статистики критерия (см. п. 5 раздела 8.2).

8.2. Математические понятия и утверждения: краткая информация и ссылки на источники

1. Доверительные интервалы для разности математических ожиданий и отношений дисперсий нормально распределенных генеральных совокупностей [1, с. 174 – 177; 2, с. 142 – 143].

2. Критерий проверки гипотезы о равенстве дисперсий по критерию Фишера [1, с. 188 – 190].

Критерий используется в предположении, что выборки $x_1, x_2, ..., x_{n_1}$ и $y_1, y_2, ..., y_{n_2}$ относятся к независимым нормально распределенным случайным величинам X и Y.

Основная гипотеза $H_0: \sigma_1^2 = \sigma_2^2$.

Альтернативная гипотеза: $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Используется статистика $F = \dfrac{S_1^2}{S_2^2}$ (отношение несмещенных оценок дисперсий случайных величин X и Y, причем $s_1^2 > s_2^2$). При основной гипотезе статистика F имеет распределение Фишера $F \sim F\left(n_1 - 1, n_2 - 1\right)$.

Критическая область при заданном уровне значимости α определяется неравенством $F > f_{1-\alpha/2}\left(n_1 - 1, n_2 - 1\right)$, где $f_{1-\alpha/2}\left(n_1 - 1, n_2 - 1\right)$ – квантиль распределения Фишера.

3. Критерий проверка гипотезы о равенстве математических ожиданий при неизвестных дисперсиях в условии принятия гипотезы о равенстве дисперсий.

Критерий используется в предположении, что выборки $x_1, x_2, ..., x_{n_1}$ и $y_1, y_2, ..., y_{n_2}$ относятся к независимым нормально распределенным случайным величинам X и Y и принята гипотеза об одинаковости дисперсий.

Основная гипотеза $H_0 :\ m_1 = m_2$.

Альтернативная гипотеза H_1 формулируется в одном из трех вариантов: $m_1 < m_2$, $m_1 \neq m_2$, $m_1 > m_2$.

Используется статистика $T = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$

(S_1^2, S_2^2 – несмещенные оценки дисперсий случайных величин X и Y). При основной гипотезе статистика T имеет распределение Стьюдента $T \sim St\left(n_1 + n_2 - 2\right)$.

Критическая область определяется стандартно в соответствии с видом альтернативы: $V_{\text{кр}} = \{T < t_\alpha\}$, $V_{\text{кр}} = \{T < t_{\frac{\alpha}{2}}\} \cup \{T > t_{1-\frac{\alpha}{2}}\}$, $V_{\text{кр}} = \{T > t_{1-\alpha}\}$, где t_α – квантиль распределения Стьюдента.

Если найдено p – значение, то при значениях уровня значимости, удовлетворяющих неравенству $\alpha > p(x_1, x_2, ..., x_n)$, основная гипотеза отклоняется, в противном случае – принимается.

4. Проверка гипотезы о равенстве математических при неизвестных дисперсиях при условии отклонения гипотезы о равенстве дисперсий (критерий Уэлча).

Критерий используется в предположении, что выборки $x_1, x_2, ..., x_{n_1}$ и $y_1, y_2, ..., y_{n_2}$ относятся к независимым нормально распределенным случайным величинам X и Y и принята гипотеза о различии дисперсий.

Основная гипотеза $H_0: m_1 = m_2$.

Альтернативная гипотеза H_1 формулируется в одном из трех вариантов: $m_1 < m_2$, $m_1 \neq m_2$, $m_1 > m_2$.

Используется статистика $T = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$ (S_1^2, S_2^2 – несмещенные оценки дисперсий случайных величин X и Y). При основной гипотезе статистика T имеет распределение Стьюдента $T \sim St(k)$, где рассчитывается по формуле $k = \dfrac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\dfrac{\left(S_1^2/n_1\right)^2}{n_1 - 1} + \dfrac{\left(S_2^2/n_2\right)^2}{n_2 - 1}}$ (с округлением до целого).

Критическая область определяется стандартно в соответствии с видом альтернативы.

Если найдено p – значение, то для всех значений уровня значимости, таких, что $\alpha \le p(x_1, x_2, ..., x_n)$, основная гипотеза принимается, при всех $\alpha > p(x_1, x_2, ..., x_n)$ основная гипотеза отклоняется.

5. Проверка гипотезы о равенстве вероятностей событий [2, с. 270 – 272, 296 – 297].

Рассмотрим два критерия, которые используются в предположении, что выборки $x_1, x_2, ..., x_{n_1}$ и $y_1, y_2, ..., y_{n_2}$ относятся к независимым случайным величинам X и Y, распределенным по биномиальному закону.

Пусть некоторое событие A в серии из n_1 испытаний произошло k_1 раз, а в другой серии из n_2 испытаний произошло k_2 раза.

1) Пусть соблюдены условия: сумма (n_1+n_2) – «большая» и наименьшая из величин $\frac{(k_1+k_2)n_1}{n_1+n_2}$, $\frac{(k_1+k_2)n_2}{n_1+n_2}$, $\frac{(n-k_1-k_2)n_1}{n_1+n_2}$, $\frac{(n-k_1-k_2)n_2}{n_1+n_2}$ больше 5.

Основная гипотеза $H_0: p_1 = p_2$.

Альтернативная гипотеза H_1 формулируется в одном из трех вариантов: $p_1 < p_2$, $p_1 \neq p_2$, $p_1 > p_2$.

Используется статистика $Z = \frac{\frac{k_1}{n_1} - \frac{k_2}{n_2}}{\sqrt{h(1-h)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$ ($h = \frac{k_1+k_2}{n_1+n_2}$).

При основной гипотезе статистика Z имеет распределение, близкое к стандартизированному нормальному распределению.

Критическая область определяется стандартно в соответствии с видом альтернативы.

Если найдено p – значение, то при значениях уровня значимости, удовлетворяющих неравенству $\alpha > p(x_1, x_2, ..., x_n)$, основная гипотеза отклоняется, в противном случае – принимается.

2) Пусть соблюдены условия: $n_1+n_2 > 20$ и наименьшая из величин $\frac{(k_1+k_2)n_1}{n_1+n_2}$, $\frac{(k_1+k_2)n_2}{n_1+n_2}$, $\frac{(n-k_1-k_2)n_1}{n_1+n_2}$, $\frac{(n-k_1-k_2)n_2}{n_1+n_2}$ больше 3.

Основная гипотеза $H_0: p_1 = p_2$.

Альтернативная гипотеза $H_1: p_1 \neq p_2$.

Используется статистика $Z = \frac{(n_1+n_2)(k_1(n_2-k_2)-k_2(n_1-k_1))^2}{n_1 \cdot n_2 \cdot (k_1+k_2)(n_1+n_2-(k_1+k_2))}$. При основной гипотезе статистика Z имеет распределение, близкое к распределению хи-квадрат с одной степенью свободы.

Критическая область определяется неравенством $Z > \chi^2_{1-\alpha}(1)$.

8.3. Библиотечные инструменты языка программирования Python

Загрузка основных модулей

```
import numpy as np
import scipy.stats as sts
import matplotlib.pyplot as plt
%matplotlib inline
```

Проверка гипотезы о равенстве математических ожиданий при неизвестных дисперсиях в условии принятия гипотезы о равенстве дисперсий (критерий изложен в п. 3 подраздела 8.2)

Функция `ttest_ind(x_1, x_2, axis=0, equal_var=True, nan_policy='propagate')` модуля `scipy.stats`

Возвращает выборочное значение статистики $T = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\left(\dfrac{1}{n_1}+\dfrac{1}{n_2}\right)}}$ и достигаемый уровень значимости p-значение.

Параметры: `x_1` и `x_2` – выборки; `nan_policy` – задает способ обработки пропущенных значений.

Проверка гипотезы о равенстве математических ожиданий при неизвестных дисперсиях при условии отклонения гипотезы о равенстве дисперсий (критерий изложен в п. 4 подраздела 8.2)

Функция `ttest_ind(x_1, x_2, axis=0, equal_var=False, nan_policy='propagate')` модуля `scipy.stats`

Возвращает выборочное значение статистики $T = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{s_1^2}{n_1}+\dfrac{s_2^2}{n_2}}}$ и достигаемый уровень значимости p-значение.

Параметры: `x_1` и `x_2` – выборки; `nan_policy` – задает способ обработки пропущенных значений.

8.4. Примеры для совместного обсуждения

Пример 1. Проведем эксперимент. Выдадим двум студентам группы по игральному кубику и попросим каждого из них подбросить свой кубик 50 раз, после чего записать результат – число выпадений 5 или 6 (выпадение этих чисел назовем событием A).

Проверим гипотезу о равенстве вероятностей события A в двух сериях испытаний. При каких уровнях значимости гипотеза будет принята?

Пример 2. В файле «Данные_8» приведены выборки 1, 2 и 3, принадлежащие нормально распределенным генеральным совокупностям.

1) Методом доверительных интервалов для каждой пары выборок 1 и 2, 2 и 3, 1 и 3 проверим гипотезу о равенстве дисперсий (доверительную вероятность возьмем равной 95%).

2) Для тех пар выборок, для которых гипотеза о равенстве дисперсий была принята, проверим гипотезу о равенстве математических ожиданий. При каких значениях доверительной вероятности проверка даст положительный результат (не будет противоречить опытным данным)?

8.5. Задания для самостоятельного выполнения

Задание 1. В файле «Данные 8_1» приведены данные успеваемости студентов 1-го курса по четырем математическим дисциплинам по 100 балльной шкале (нормированные к 1). Иван и Петр, входящие в число этих студентов, поспорили насчет того, по каким дисциплинам у их однокурсников хороших оценок больше, а по каким меньше. Иван думает, что в этом плане между дисциплинами особой разницы нет, а Петр настаивает, что есть. Подумав, они решили, что для объективного разрешения спора им надо, опираясь на методы математической статистики, сопоставить вероятности получения не менее M баллов в разных парах дисциплин (а именно, проверить гипотезы о равенстве вероятностей получения не менее M баллов по этим дисциплинам (см. п. 5 подраздела 8.2)). Ваша задача – помочь Ивану и Петру, взяв на себя анализ пары дисциплин (в соответствии с вашим вариантом). Нужно найти диапазоны значений M, при которых на уровне значимости $\alpha = 0{,}05$ данные подтверждают мнение Ивана, и диапазоны значений M, при которых на уровне значимости $\alpha = 0{,}05$, данные подтверждают мнение Петра.

Замечание. Данные реальные и требуют предварительной обработки.

Вариант 1: ОМА и МА
Вариант 2: ОМА и АиГ
Вариант 3: ОМА и ДУ
Вариант 4: МА и АиГ
Вариант 5: МА и ДУ
Вариант 6: АиГ и ДУ

Задание 2. 1) Сгенерируйте две выборки A и B различных объемов из нормально распределенных генеральных совокупностей с параметрами m_A, σ_A и m_B, σ_B (возьмите значения m_A и m_B близкими, но не равными, а σ_A и σ_B выберите такими, чтобы близки (но не равны) были их квадраты). Проверьте гипотезу о равенстве дисперсий по критерию Фишера (см. п. 2 подраздела 8.2) на уровне значимости 0,05. Сформулируйте выводы.

2) Для выборок A и B проверьте гипотезу о равенстве математических ожиданий. Укажите диапазон значений α, при которых гипотеза принимается.