**Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)**

Факультет «Информатика и системы управления»

Кафедра «Теоретическая информатика и компьютерные технологии»

**Лабораторная работа № 3**

**по курсу «Моделирование»**

ЗАДАЧА ПРОВЕРКИ СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ С ИСПОЛЬЗОВАНИЕМ КРИТЕРИЯ СОГЛАСИЯ КОЛМОГОРОВА-СМИРНОВА

Студент: Яровикова А. С.

Группа: ИУ9-81Б

Преподаватель: Домрачева А. Б.

Москва, 2024

# **ЦЕЛЬ И ПОСТАНОВКА ЗАДАЧИ**

**Цель**

Целью данной работы является получение опыта постановки статистической гипотезы и ее проверки на основе критерия согласия Колмогорова-Смирнова.

**Постановка задачи**

Подбираются две выборки размерами ненаблюдаемых одновременно показателей . Выборки разделяются на рабочие (обучающие) и контрольные (валидационные) в отношении 0.75 : 0.25.

Необходимо:

1. Найти выборочное среднее для обеих выборок.
2. Найти выборочную дисперсию для обеих выборок.
3. Построить выборочные функции распределения для обеих выборок
4. Найти оптимальные значения параметров  .
5. Предполагая, что законы распределения выборок имеют вид: , и имея значения , сравнить выборочные функции распределения  с гипотетическими функциями распределения .
6. С помощью критерия согласия Колмогорова-Смирнова проверить гипотезу о законе распределения
7. Найти значения – выборочные дисперсии выборок, .
8. Подтвердить гипотезу о функции связи , где .

# **ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ**

Задача проверки статистической гипотезы является ключевым аспектом статистического анализа данных. Она заключается в формулировании гипотезы о предполагаемом значении параметра или законе распределения и последующей проверке этой гипотезы на основе наблюдаемых данных.

Целью проверки является принятие или отвержение статистической гипотезы на основе степени убедительности представленных данных.

Важными понятиями в задаче проверки статистической гипотезы является статистический критерии проверки гипотез. **Статистические критерии** — это совокупность правил, позволяющих на основе параметров выборки принять или отвергнуть основную гипотезу.

Имеется общий принцип построения статистического критерия. Сначала задают некоторую функцию являющейся статистикой критерия, которая зависит от результатов эксперимента. Множество всех возможных значений S разбивают на два подмножества: (принятие основной гипотезы) и (критическое множество). Если конкретное значение статистики попадает в , то основную гипотезу принимают, иначе отвергают (в пользу альтернативной или переформулируют задачу).

Если закон распределения известен изначально, как в поставленной перед нами задачей, то задача ставится в узком смысле, полученные статистические выводы будут достаточно точны.

**Критерии согласия** показывают, насколько предположение о законе распределения соответствует экспериментальным данным. В них гипотеза определяет закон распределения полностью, либо с точностью до небольшого числа параметров.

В общем случае критерий согласия выглядит так: пусть имеется выборка размера , теоретическая функция распределения , гипотетическая — . — выборочная, соответствующая гипотетической. Тогда основная гипотеза — это = , где это некоторое множество.

Если верна, то → при n →

Примером критерия согласия является **критерий согласия Колмогорова—Смирнова**. Критерий основан на метрике Его основная идея заключается в следующем: выбирается , где пороговое значение с заданным уровнем значимости ; все, что вне Критерий согласия выполняется, если выполняется следующее неравенство: .

Таким образом, правильное формулирование и проведение проверки статистической гипотезы позволяет принимать обоснованные решения на основе данных и выводов статистического анализа.

# **ПРАКТИЧЕСКАЯ РЕАЛИЗАЦИЯ**

В данном задаче были использованы выборки мужских и женских доходов из набора данных по покупке беговых дорожек для фитнеса. Всего было 104 и 76 записей в выборках для мужчин и женщин соответственно. Данные были поделены на обучающие и контрольные выборки – 78 и 26 значений для мужских доходов и 57 и 19 – для женских, соответственно.

Исходный код лабораторной работы представлен ниже.

Импортирование необходимых библиотек:

**import** **pandas** **as** **pd**

**import** **numpy** **as** **np**

**import** **matplotlib.pyplot** **as** **plt**

**from** **scipy.optimize** **import** minimize

Загрузка датасета и извлечение нужного столбца:

# Загрузка датасета из CSV файла

dataset = pd.read\_csv('CardioGoodFitness.csv')

m\_dataset = dataset[dataset['Gender'] == 'Male']

f\_dataset = dataset[dataset['Gender'] == 'Female']

print("Количество записей в выборках: ", len(m\_dataset), len(f\_dataset))

# Вывод выборок

print("**\n**Выборка со значением Male в колонке Gender:")

print(m\_dataset)

print("**\n**Выборка со значением Female в колонке Gender:")

print(f\_dataset)

# Извлечение необходимого столбца

sample1 = m\_dataset['Income']

sample2 = f\_dataset['Income']

Вычисление выборочного среднего, выборочной дисперсии и построение выборочных функций распределения для обеих выборок:

# Вычисление выборочного среднего

mean1 = sample1.mean()

mean2 = sample2.mean()

# Вычисление выборочной дисперсии

variance1 = sample1.var(ddof=**1**)

variance2 = sample2.var(ddof=**1**)

# Вычисление выборочной функции распределения

cdf1 = np.cumsum(sample1) / sample1.sum()

cdf2 = np.cumsum(sample2) / sample2.sum()

# Вывод результатов

print("**\n**Выборочное среднее выборки 1:", mean1)

print("Выборочное среднее выборки 2:", mean2)

print("Выборочная дисперсия выборки 1:", variance1)

print("Выборочная дисперсия выборки 2:", variance2)

# plt.plot(np.sort(sample1), cdf1, marker='.', linestyle='none')

# Построение выборочной функции распределения выборки 1

plt.hist(sample1, bins=len(sorted(set(sample1))), density=**True**, cumulative=**True**, histtype='step', linewidth=**1.5**)

plt.xlabel('Значение')

plt.ylabel('Вероятность')

plt.title('Выборочная функция распределения мужских доходов')

plt.grid(**True**)

plt.show()

# plt.plot(np.sort(sample2), cdf2, marker='.', linestyle='none')

# Построение выборочной функции распределения выборки 2

plt.hist(sample2, bins=len(sorted(set(sample2))), density=**True**, cumulative=**True**, histtype='step', linewidth=**1.5**, color='red')

plt.xlabel('Значение')

plt.ylabel('Вероятность')

plt.title('Выборочная функция распределения женских доходов')

plt.grid(**True**)

plt.show()

Разделение выборок на обучающие и валидационные выборки:

# Размеры массивов для обучения и валидации

n1 = len(sample1)

n2 = len(sample2)

print("Количество записей в выборках: ", n1, n2)

train\_size1 = int(n1 \* **0.75**) # 75% для обучения

val\_size1 = n1 - train\_size1 # 25% для валидации

train\_size2 = int(n2 \* **0.75**) # 75% для обучения

val\_size2 = n1 - train\_size2 # 25% для валидации

# Разделение на обучающую и валидационную выборки

train\_sample1 = sample1[:train\_size1]

valid\_sample1 = sample1[train\_size1:]

train\_sample2 = sample2[:train\_size2]

valid\_sample2 = sample2[train\_size2:]

print("train1:", len(train\_sample1), " valid1:", len(valid\_sample1))

print("train2:",len(train\_sample2), " valid2:", len(valid\_sample2))

Вычисление параметров :

**def** **distr\_func**(samples, x):

**return** np.sum(samples <= x) / len(samples)

**def** **distr**(alpha, beta, x):

**return** **1** - np.exp(-alpha \* x \*\* beta)

# Метод наименьших квадратов

# использование минимизации суммы квадратов разностей между подсчитанными и наблюдаемыми значениями.

**def** **squared\_sum**(params, samples):

alpha, beta = params

# гипотетическое значение

hyp\_values = list(map(**lambda** x: distr(alpha, beta, x), xs))

# целевое значение

target\_values = list(map(**lambda** x: distr\_func(samples, x), xs))

**return** np.sum((np.array(hyp\_values) - np.array(target\_values))\*\***2**)

scale = **100000.0**

lst1 = [x/scale **for** x **in** train\_sample1]

lst2 = [x/scale **for** x **in** train\_sample2]

samples = lst1

xs = np.arange(min(samples), max(samples), (max(samples) - min(samples)) / **100**)

values = list(map(**lambda** x: distr\_func(samples, x), xs))

initial\_guess = [**0**, **0**]

result = minimize(squared\_sum, initial\_guess, args=(samples,))

alpha1, beta1 = result.x

print("Оценка alpha1:", alpha1)

print("Оценка beta1:", beta1)

samples = lst2

xs = np.arange(min(samples), max(samples), (max(samples) - min(samples)) / **100**)

values = list(map(**lambda** x: distr\_func(samples, x), xs))

initial\_guess = [**0**, **0**]

result = minimize(squared\_sum, initial\_guess, args=(samples,))

alpha2, beta2 = result.x

print("Оценка alpha2:", alpha2)

print("Оценка beta2:", beta2)

Построение выборочной функции распределения с нормированными значениями показателей для обеих выборок:

# распределение Вейбулла

**def** **F**(x, a, b):

**return** **1**- np.exp(-a \* x\*\*b)

unique\_values1 = sorted(set(lst1))

unique\_values2 = sorted(set(lst2))

plt.hist(lst1, bins=len(unique\_values1), density=**True**, cumulative=**True**, histtype='step', linewidth=**1.5**)

plt.xlabel('Нормированное Значение')

plt.ylabel('Вероятность')

plt.title('Выборочная функция распределения мужских доходов')

plt.grid(**True**)

plt.show()

# unique\_values2 = sorted(set(lst2))

plt.hist(lst2, bins=len(unique\_values2), density=**True**, cumulative=**True**, histtype='step', linewidth=**1.5**)

plt.xlabel('Нормированное Значение')

plt.ylabel('Вероятность')

plt.title('Выборочная функция распределения женских доходов')

plt.grid(**True**)

plt.show()

Построение графиков выборочной функции распределения и кривой гипотетического закона распределения для обеих выборок:

res1 = [F(x, alpha1, beta1) **for** x **in** unique\_values1]

plt.hist(lst1, bins=len(unique\_values1), density=**True**, cumulative=**True**, histtype='step', linewidth=**1.5**)

plt.plot(unique\_values1, res1)

plt.xlabel('Нормированное Значение')

plt.ylabel('Вероятность')

plt.title('Мужские доходы')

plt.grid(**True**)

plt.show()

res2 = [F(x, alpha2, beta2) **for** x **in** unique\_values2]

plt.hist(lst1, bins=len(unique\_values2), density=**True**, cumulative=**True**, histtype='step', linewidth=**1.5**)

plt.plot(unique\_values2, res2)

plt.xlabel('Нормированное Значение')

plt.ylabel('Вероятность')

plt.title('Женские доходы')

plt.grid(**True**)

plt.show()

Критерий согласия Колмогорова-Смирнова:

**def** **kolmogorov**(y1, y2, sample\_len):

eps = **0.01**

Dn = max(abs(y1[i] - y2[i]) **for** i **in** range(len(y1)))

betta = **0.01**

Db = **1** / np.sqrt(sample\_len) \* np.sqrt(-**0.5**\*np.log(betta))

print(Dn, Db)

**return** (Dn - Db) <= eps

print("Критерий Колмогорова для выборки 1:", kolmogorov(list(set(lst1)), res1, len(res1)))

print("Критерий Колмогорова для выборки 2:", kolmogorov(list(set(lst2)), res2, len(res2)))

Вычисление и проверка гипотезы о функции связи , где :

s1 = valid\_sample1.var(ddof=**1**)

s2 = valid\_sample2.var(ddof=**1**)

betta = (s1/s2)\*\*(**1**/**2**)

# betta = (s1\*\*2/s2\*\*2)\*\*(1/2)

L1 = np.mean(np.log(valid\_sample1))

L2 = np.mean(np.log(valid\_sample1))

alphaa = np.exp(L1 - betta\*L2)

print("alpha^: ", alphaa, " beta^:", betta)

print(alpha1/alpha2)

print(beta1/beta2)

# **РЕЗУЛЬТАТЫ**

Графики выборочных функций распределения и результаты подсчета параметров представлены на рисунках ниже.



Рисунок 1 - Выборочная функция распределения мужских доходов



Рисунок 2 - Выборочная функция распределения женских доходов

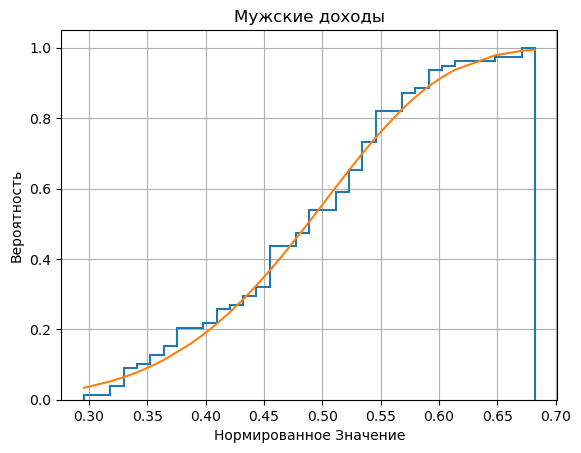


Рисунок 3 - Выборочная функция распределения и гипотетическая функция распределения мужских доходов

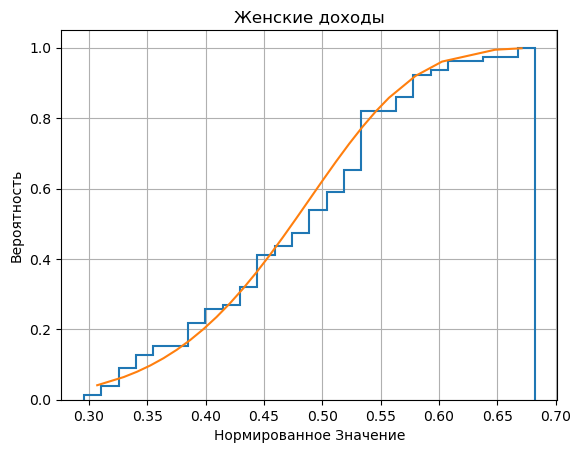


Рисунок 4 - Выборочная функция распределения и гипотетическая функция распределения женских доходов

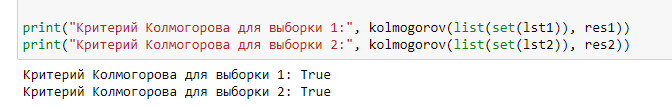


Рисунок 5 - Проверка критерия согласия Колмогорова-Смирнова для двух выборок

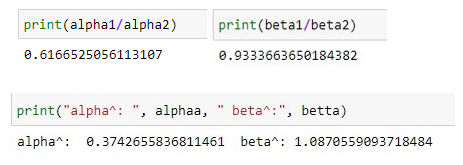


Рисунок 6 - Проверка гипотезы о функции связи

# **ВЫВОДЫ**

В ходе выполнения лабораторной работы была изучено применение критерия согласия Колмогорова-Смирнова для проверки статистической гипотезы. Данныйкритерий позволяет оценить соответствие эмпирической функции распределения данных теоретической функции распределения.

Результаты работы позволили сделать вывод о том, что гипотеза о законе распределения данных (ненаблюдаемых одновременно показателей – доходов мужчины и женщин) подтверждена и соответствует экспериментальным данным.

В заключении хочется отметить, что при использовании критерия согласия Колмогорова-Смирнова ключевыми аспектами являются: правильный выбор уровня значимости или вероятности ошибки второго рода (когда основная гипотеза не верна, но ее приняли) и интерпретация результатов проверки гипотезы. При этом стоит помнить о том, что состоятельность критерия Колмогорова-Смирнова ниже состоятельности, например интегрального критерия согласия. Поэтому проверка сложной гипотезы, например подтверждение наличия некоторого распределения при неизвестном требует рассмотрения модифицированных статистик, которые так же учитывают параметры распределения.