Министерство образования и науки 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, искусственный интеллект и системы управления

КАФЕДРА Теоретическая информатика и компьютерные технологии

Домашняя работа

по курсу «Моделирование»

«Проверка статистических гипотез»

Студент группы ИУ9-82Б Виленский С. Д.

Преподаватель Домрачева А. Б.

Москва, 2025 г.

ЦЕЛЬ

Приобретение навыка формулирования статистических гипотез и их проверки.

ПОСТАНОВКА ЗАДАЧИ

Проверить гипотезу о возможности описания стохастической зависимости между переменными двупараметрической степенной функцией связи. Для оценки параметров использовать модифицированный метод моментов, для проверки гипотез - критерий Колмогорова-Смирнова.

ИЗУЧЕНИЕ ПРЕДМЕТНОЙ ОБЛАСТИ

Предметной областью и объектами исследования являются женщины с различными параметрами функционирования организма и целевым бинарным параметром Outcome, значение которого требуется предсказать. Множество признаков включает в себя:

- количество беременностей,
- концентрация глюкозы в плазме через 2 часа при пероральном тесте на толерантность к глюкозе,
- диастолическое артериальное давление (мм рт. ст.),
- толщина кожной складки трицепса (мм),
- 2-часовой уровень инсулина в сыворотке (мкЕд/мл),
- индекс массы тела (вес в кг/(рост в м)^2),
- диабет родословная функция (функция, которая оценивает вероятность развития диабета на основе семейного анамнеза),
- возраст (лет).

Поскольку в данной задаче для цензурирования данных и их преобразования требуется специалист в предметной области, преобразований выборок данных произведено не было.

ФОРМУЛИРОВКА СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

Назовем статистической гипотезой H в данной задаче утверждение о том, что стохастическую зависимость между выборками значений концентрации глюкозы и уровня инсулина можно описать двупараметрической степенной функцией связи.

Таким образом основной гипотезой H0 будет утверждение о возможности описания зависимости степенной функцией согласно критерию Колмагорова-Смирнова с использованием таблицы пороговых значений (рисунок 1), а альтернативной гипотезой H1 будет утверждение о невозможности.

Степень свободы	Проверка единичной выборки *			Проверка двух выборок **	
N	$D_{0,10}$	$D_{0,05}$	$D_{0,01}$	$D_{0,05}$	$D_{0,01}$
15	0,304	0,338	0,404	0,533	0,600
16	0,295	0,328	0,392	0,500	0,625
17	0,286	0,318	0,381	0,471	0,588
18	0,278	0,309	0,371	0,500	0,556
19	0,272	0,301	0,363	0,474	0,526
20	0,264	0,294	0,356	0,450	0,550
25	0,240	0,270	0,320	0,400	0,480
30	0,220	0,240	0,290	0,370	0,430
35	0,210	0,230	0,270	0,340	0,390
Более 35	$\frac{1,22}{\sqrt{N}}$	$\frac{1,36}{\sqrt{N}}$	$\frac{1,63}{\sqrt{N}}$	$1,36\sqrt{\frac{N_1+N_2}{N_1N_2}}$	$1,63\sqrt{\frac{N_1+N_2}{N_1N_2}}$

* Применяется для оценки степени близости выборочных значений к теоретическому распределению.
N – объем выборки.
** Применяется для определения принадлежности двух выборок объемами N_1 и N_2 одному и тому же распределению. При малых размерах выборки $N = N_1 = N_2$.

Рисунок 1 − Таблица оценок пороговых значений для статистики Кломогорова-Смирнова

Для сравнения были использованы две выборки и проверяем их принадлежность к одному и тому же распределению. Используется следующая формула: $glucose = \alpha * insulin^{\beta}$, где α и β — гиперпараметры, glucose и insulin — значения выборок.

Для нахождения гиперпараметров были применены следующие формулы:

- $insulin^{*} = \frac{1}{n_{insulin}} \sum_{i=1}^{n_{insulin}} ln(insulin_i)$;

- $glucose^{*} = \frac{1}{n_{glucose}} \sum_{i=1}^{n_{glucose}} ln(glucose_i)$;
- $S^2_{insulin} = \frac{1}{n_{insulin}} \sum_{i=1}^{n_{insulin}} (ln(insulin_i) - insulin^{*})^2$;
- $S^2_{glucose} = \frac{1}{n_{glucose}} \sum_{i=1}^{n_{glucose}} (ln(glucose_i) - glucose^{*})^2$;
- $\hat{\beta} = \sqrt{S^2_{insulin}/S^2_{glucose}}$;
- $\hat{\alpha} = e^{insulin^{*} - \hat{\beta} * glucose^{*}}$.

ОЦЕНКА ПО СТАТИСТИКЕ КОЛМОГОРОВА-СМИРНОВА

Статистика Колмогорова-Смирнова описывается следующей формулой: $D_n = sup_n|G(x) - G_n(x)| <= D_\beta$. В первую очередь требуется проверить более слабый критерий $D_\beta = 1.36\sqrt{\frac{N_{insulin}+N_{glucose}}{N_{insulin}N_{glucose}}}$.

Был реализован алгоритм проверки гипотезы (листинг 1) и вычислены следующие значения:

- $\hat{\alpha} = 0.975$;
- $\hat{\beta} = 0.215$;
- $\hat{glicose} = \alpha * insulin^{\hat{\beta}}$.

Плотности распределений выборок изображены на диагонали матрицы рассеяний на рисунке 1.

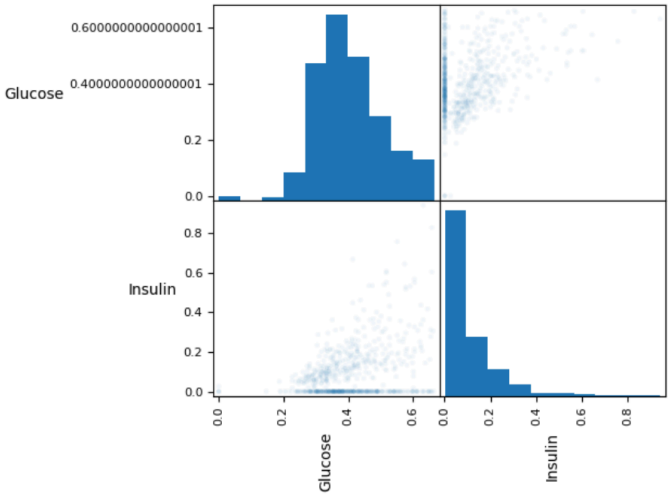

Рисунок 1 − Матрица рассеяний выборок данных

На диагонали данной матрицы, на которой изображены плотности распределения значений видно, что среди значений уровня инсулина большая часть значений равны нулю, что повлияло на точность построенной модели.

Графики функций распределения значений глюкозы и вычисленных значений глюкозы через значения инсулина через степенную функцию изображены на рисунке 2.

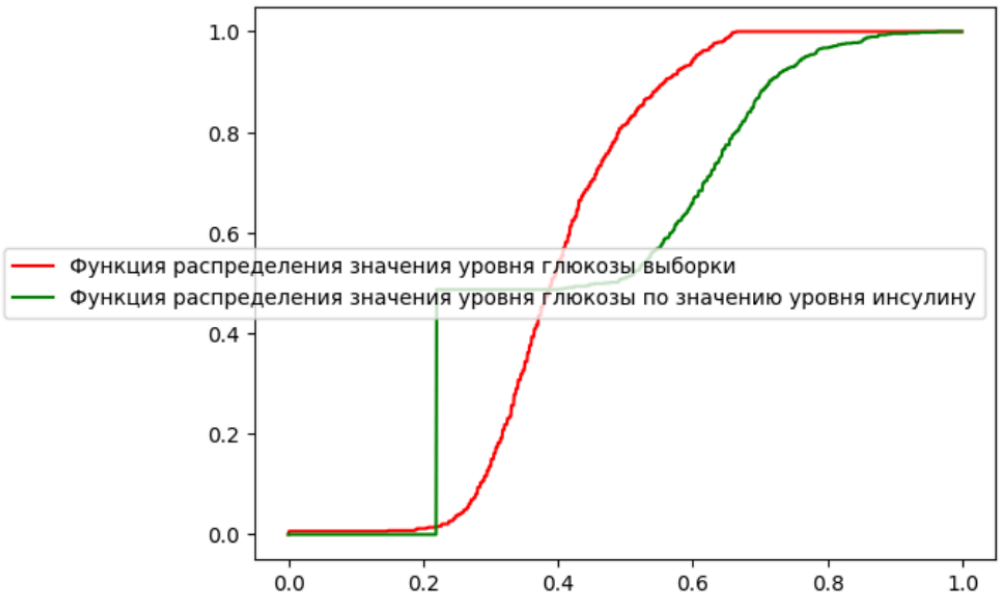

Рисунок 2 − Графики функций распределения

Из них видно, что из-за большого количества нулей в исходной выборке значений инсулина практически половина равны нулю, нет явной возможности построить достаточно точную степенную функцию сопоставления двух выборок.

Вычислительно это подтверждается критерием статистики Колмогорова-Смирнова: $D_n = 0.471; D_\beta = 0.083$. Следовательно гипотеза H0 отвергается и принимается альтернативная гипотеза H1.

Объясняется это тем, что половина от всех значений уровня инсулина равны нулю, что дало данный результат.

ФРАГМЕНТЫ ИСХОДНОГО КОДА

Листинг 1 – Программная реализация проведения случайных экспериментов согласно описанной модели марковской цепи

```
df_target = df

n = len(df_target)

n_glucose = len(df_target['Glucose'])
n_insulin = len(df_target['Insulin'])

avg_ln_glucose = 1 / n_glucose * sum(np.log(gl_val) for gl_val in df_target['Glucose'])
avg_ln_insulin = 1 / n_insulin * sum(np.log(ins_val) for ins_val in df_target['Insulin'])

S_2_glucose = 1 / n_glucose * sum((np.log(gl_val) - avg_ln_glucose) ** 2 for gl_val in df_target['Glucose'])
S_2_insulin = 1 / n_insulin * sum((np.log(ins_val) - avg_ln_insulin) ** 2 for ins_val in df_target['Insulin'])

beta_ = (S_2_glucose / S_2_insulin) ** .5
alpha_ = np.exp(avg_ln_glucose - beta_ * avg_ln_insulin)
```

```
fact_distribution_glucose = lambda x: sum(df['Glucose'] <= x) / len(df)
pred_distribution_glucose = lambda x: sum((alpha_ * df['Insulin'] ** beta_) <= x) / len(df)
```

ВЫВОДЫ

В результате выполнения домашнего задания была проверена и отвергнута гипотеза возможности описания стохастической зависимости между переменными уровня глюкозы и инсулина двупараметрической степенной функцией связи в исходном датасете без предобработки данных.

Вызвано это наличием чрезмерно большого количества одинаковых значений в выборке инсулина. Однако поскольку было поставлено предположение о том, что данные предобработаны и предоставлены специалистом предметной области, мы интерпретируем отвержение нулевой гипотезы и принятие альтернативной как невозможность описания стахостической зависимости между переменными уровня глюкозы и инсулина двухпараметрической степенной функцией связи.