Лабораторная работа по математической статистике

Проверка статистических гипотез

I. Проверка параметрических гипотез

В файле Mall_Customers.csv содержатся данные опроса покупателей торгового центра (источник: https://www.kaggle.com/datasets/shwetabh123/mall-customers). В таблице:

- CustomerID - индивидуальный номер респондента;
- Genre - пол;
- Age - возраст;
- Annual Income (k$) - годовой доход;
- Spending Score (1-100) - рейтинг расходов (процент от доходов, который тратятся на покупку товаров в торговых центрах)

1. Проверьте утверждение, что « Мужчины и женщины 50% своего дохода тратят в супермаркетах» .

2. Проверьте утверждение, что «Доход у мужчин больше, чем у женщин» с уровнем значимости 5%.

3. Проверьте гипотезу, что неоднородность возрастов женщин такая же как у мужчин, т. е. что дисперсии возраста женщин равна дисперсии возраста мужчин.

4. Выяснить, какие функции для проверки параметрических гипотез есть python-пакете statsmodels.stats. В возможных случаях сравнить полученные в пункта 1-3 результаты с предсказаниями найденных функций.

* Проверку проводить с уровнем значимости 5%.
** В пунктах 1-3 проверку гипотезы провести с помощью *всех* рассмотренных на лекции статистических критериев, см. табл. 1 (если для этого понадобятся истинные значения математического ожидания и дисперсии, то предположить, что они равны оценочным значениям).

II. Проверка непараметрических гипотез

Срок эксплуатации турбореактивного двигателя (характеристика надежности) определяется как сумма сроков службы его лопаток.

$$T = \sum_{i=1}^{N} t_i$$

Предполагая, что срок службы t_i каждой из $N = 100$ лопаток описывается экспоненциальным законом с параметром интенсивности выхода из строя, равном $\lambda = 2$. Все t_i являются независимыми. Определите:

Какой закон распределения будет иметь надежность всего двигателя? Обоснуйте вывод. Убедитесь в правильности ответа.
Для этого:
1) Постройте переменную T (для генерации выборки t_i можно использовать произвольное количество элементов, но желательно больше 100).
2) Вычислите параметры предполагаемого закона распределения T и сравните их с теоретическими.

3) Постройте гистограммы относительных частот и эмпирическую функцию распределения для T.
4) Проверьте гипотезу о предполагаемом виде закона распределения.

Форма сдачи лабораторной работы — устная. Знать теорию по проверке параметрических гипотез.

Табл. 1

№ п/п	Гипотеза H_0	Предположения	Тест MS Excel	Статистика критерия K и ее распределение $f_{H_0}(k)$	Область принятия гипотезы H_0		
					для двусторонней критической области	для правосторонней критической области	для левосторонней критической области
1	$m = m_0$, m_0 задано	дисперсия σ^2 известна	ZTECT (*массив*; m_0; *сигма*)	$U = \dfrac{m^* - m_0}{\sigma/\sqrt{n}}$, $U \in N(0;1)$	$\lvert U_{\text{набл}} \rvert < u_{1-\alpha/2}$	$U_{\text{набл}} < u_{1-\alpha}$	$U_{\text{набл}} > u_{\alpha}$
2		дисперсия σ^2 неизвестна	ZTECT (*массив*; m_0)	$T = \dfrac{m^* - m_0}{s/\sqrt{n}}$, $T \in St_{n-1}$	$\lvert T_{\text{набл}} \rvert < t_{1-\alpha/2,\,n-1}$	$T_{\text{набл}} < t_{1-\alpha,\,n-1}$	$T_{\text{набл}} > t_{\alpha,\,n-1}$
3	$m_1 = m_2$	дисперсии σ_1^2 и σ_2^2 известны	Двухвыборочный *Z*-тест для средних	$U = \dfrac{m_1^* - m_2^*}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$, $U \in N(0;1)$	$\lvert U_{\text{набл}} \rvert < u_{1-\alpha/2}$	$U_{\text{набл}} < u_{1-\alpha}$	$U_{\text{набл}} > u_{\alpha}$
4		σ_1^2 и σ_2^2 неизвестны, но принята гипотеза о их равенстве	TTECT (*массив*1; *массив*2; *хвосты*; 2)	$T = \dfrac{m_1^* - m_2^*}{\tilde{\sigma}\sqrt{1/n_1 + 1/n_2}}$, $\tilde{\sigma} = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-1}}$, $T \in St_{n_1+n_2-2}$	$\lvert T_{\text{набл}} \rvert < t_{1-\alpha/2,\,n_1+n_2-2}$	$T_{\text{набл}} < t_{1-\alpha,\,n_1+n_2-2}$	$T_{\text{набл}} > t_{\alpha,\,n_1+n_2-2}$
5	$m_1 = m_2$	σ_1^2 и σ_2^2 неизвестны, гипотеза о их равенстве отклонена	TTECT (*массив*1; *массив*2; *хвосты*; 3)	$T = \dfrac{m_1^* - m_2^*}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$, $T \in St_k$, $k = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1-1} + \dfrac{(s_2^2/n_2)^2}{n_2-1}}$	$\lvert T_{\text{набл}} \rvert < t_{1-\alpha/2,\,k}$	$T_{\text{набл}} < t_{1-\alpha,\,k}$	$T_{\text{набл}} > t_{\alpha,\,k}$

№ п/п	Гипотеза H_0	Предположения	Тест MS Excel	Статистика критерия K и ее распределение $f_{H_0}(k)$	Область принятия гипотезы H_0		
					для двусторонней критической области	для правосторонней критической области	для левосторонней критической области
1	$\sigma^2 = \sigma_0^2$ σ_0 задано	m известно $\sigma_B \equiv \equiv \dfrac{1}{n}\sum_{j=1}^{n}(x_j - m)^2$		$\chi^2 = \dfrac{n\sigma_B^2}{\sigma_0^2}$, $\chi^2 \in \chi_n^2$	$\chi^2_{\alpha/2,\,n} < \chi_{\text{набл}}$, $\chi_{\text{набл}} < \chi^2_{1-\alpha/2,\,n}$	$\chi^2_{\text{набл}} < \chi^2_{1-\alpha,\,n}$	$\chi^2_{\text{набл}} > \chi^2_{\alpha,\,n}$
2		m не известно,		$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$, $\chi^2 \in \chi^2_{n-1}$	$\chi^2_{\alpha/2,\,n} < \chi_{\text{набл}}$, $\chi_{\text{набл}} < \chi^2_{1-\alpha/2,\,n}$	$\chi^2_{\text{набл}} < \chi^2_{1-\alpha,\,n-1}$	$\chi^2_{\text{набл}} > \chi^2_{\alpha,\,n-1}$
3	$\sigma_1^2 = \sigma_2^2$	m_1 и m_2 известны, $\sigma_{iB} \equiv \dfrac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij} - m_i)^2$, $i = 1,2$.		$F = \dfrac{\sigma_{1B}^2}{\sigma_{2B}^2}$, $\sigma_{1B} > \sigma_{2B}$, $F \in F_{n_1,n_2}$	$F_{\text{набл}} < f_{1-\alpha/2,\,n_1,n_2}$	$H_1: \sigma_1^2 > \sigma_2^2$ $F_{\text{набл}} < f_{1-\alpha,\,n_1,n_2}$	
4		m_1 и m_2 неизвестны	Двухвыборочный *F*-тест для дисперсии	$F = \dfrac{s_1^2}{s_2^2}$, $s_1 > s_2$, $F \in F_{n_1-1,\,n_2-1}$	$F_{\text{набл}} < f_{1-\alpha/2,\,n_1-1,\,n_2-1}$	$H_1: \sigma_1^2 > \sigma_2^2$ $F_{\text{набл}} < f_{1-\alpha,\,n_1-1,\,n_2-1}$	