Лабораторная забота №3.

Проверка статистических гипотез о виде и параметрах распределения случайных величин.

Часть 1.

1. Используя данные Росстата, приведенные в файле данных CHISLO_DOCTORS.xlsx, найти значения случайного показателя X : **Число врачей на 10 тысяч человек населения –** в указанном в Вашем варианте Федеральном Округе в каждый год периода T:2005, 2010, 2015, 2019, 2020 и 2021 гг.
2. Визуализировать данные показателя X в каждый год периода T по указанному в Вашем варианте Федеральному Округу с помощью графиков и боксплотов.
3. Вычислить описательную статистику: среднее, стандартное отклонение, квартили, минимальное и максимальное значения показателя X в каждый год периода T по указанному в Вашем варианте Федеральному Округу.
4. Проверить, можно ли считать, что распределение случайной величины X в указанном Федеральном Округе в каждый год периода T подчинено нормальному закону распределения. Использовать для проверки тест Шапиро-Уилка (уровень значимости α указан в Вашем варианте).

Для дальнейшего исследования использовать только те года, для которых распределение случайного показателя X в указанном в Вашем варианте Федеральном Округе можно считать нормальным.

5. Выделить те года t_1-t_m с нормально распределенными значениями рядов данных Xt_1-Xt_m, где Xt_1- Xt_m имеют одинаковую дисперсию (уровень значимости взять равным α). Использовать для проверки нулевой гипотезы о равенстве дисперсий тесты Бартлетта и Левена.

Для выполнения п. 6-8 использовать ряды данных Xt_1- Xt_m **из этой группы**.

6. Проверить, можно ли считать, что среднее значение показателя X по данному Федеральному Округу в каждый год периода t_1-t_m значимо выше (ниже) общероссийского значения показателя X (уровень значимости взять равным α). Общероссийские значения показателя X найти в файле CHISLO_DOCTORS.xlsx. Использовать для проверки гипотезы о равенстве средних t-тест для одной выборки.
7. Проверить, можно ли считать, что различия между средними значениями показателя X по данному Федеральному Округу в какие-то два года из периода t_1-t_m незначимы, появились случайно (уровень значимости взять равным α). Использовать для проверки гипотезы о равенстве средних t-тест для двух выборок.
8. Проверить значимость отличий средних в выбранной группе (уровень значимости взять равным α). Использовать для проверки гипотезы о равенстве средних групп тест Тьюки и односторонний тест ANOVA.

Часть 2.

1. Проверить, можно ли считать, что распределение случайной величины X в указанном Федеральном Округе за весь период T подчинено нормальному закону распределения. Использовать для проверки следующие три критерия:

Хи-квадрат, Шапиро-Уилка и критерий Д'Агостино (уровень значимости α указан в Вашем варианте).

2. Смоделировать M выборок объемом *n* из значений случайной величины X, имеющей нормальное распределение с параметрами, указанными в Вашем варианте. На уровне значимости α проверить для каждой выборки гипотезу о нормальном законе распределения с помощью критериев Шапиро-Уилка и Д'Агостино. По результатам моделирования M выборок вычислить оценку вероятности совершить ошибку первого рода.

3. Смоделировать M выборок объемом *n* из значений случайной величины Y, имеющей указанное в Вашем варианте распределение. На уровне значимости α проверить для каждой выборки гипотезу о нормальном законе распределения с помощью критериев Шапиро-Уилка и Д'Агостино. Вычислить оценку вероятности не допустить ошибку второго рода. Какой из критериев при данной альтернативе является более мощным?

УКАЗАНИЯ.

Полную информацию о статистических критериях от команды разработчиков см.

https://docs.scipy.org/doc/scipy/reference/stats.html

Информацию о t-тестах см.

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_1samp.html#scipy.stats.ttest_1samp

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind_from_stats.html#scipy.stats.ttest_ind_from_stats

Информацию о тестах для проверки на нормальность см.

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html#scipy.stats.normaltest

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html#scipy.stats.shapiro

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.chisquare.html#scipy.stats.chisquare

Информацию о тестах для проверки выборок на равенство дисперсий см.

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.bartlett.html

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.levene.html#scipy.stats.levene

Информацию о тестах для проверки групп на равенство средних (ожидаемых) значений групп см.

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.tukey_hsd.html

https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.f_oneway.html#scipy.stats.f_oneway

Краткая информация по параметрическим гипотезам

Предполагается, что случайная величина X имеет нормальный закон распределения.

Таблица 1. Основные параметрические гипотезы для одной выборки.

H_0	Предпо-ложения	Статистика критерия	H_1	Область принятия H_0
$a = a_0$	σ^2 известно	$U = \frac{\bar{x} - a_0}{\sigma}\sqrt{n}$	$a > a_0$ $a < a_0$ $a \neq a_0$	$U < u_{кр}$, $\Phi_0(u_{кр}) = 1/2 - \alpha$ $U > -u_{кр}$, $\Phi_0(u_{кр}) = 1/2 - \alpha$ $\lvert U\rvert < u_{кр}$, $\Phi_0(u_{кр}) = (1-\alpha)/2$
	σ^2 не-известно	$T = \frac{\bar{x} - a_0}{s}\sqrt{n}$	$a > a_0$ $a < a_0$ $a \neq a_0$	$T < t_{кр}(\alpha, n-1)$ для односторонней области $T > -t_{кр}(\alpha, n-1)$ для односторонней области $\lvert T\rvert < t_{кр}(\alpha, n-1)$ для двусторонней области
$\sigma^2 = \sigma_0^2$	a не-известно	$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$	$\sigma^2 > \sigma_0^2$ $\sigma^2 < \sigma_0^2$ $\sigma^2 \neq \sigma_0^2$	$\chi^2 < \chi^2_{\alpha;n-1}$ $\chi^2 > \chi^2_{1-\alpha;n-1}$ $\chi^2_{1-\alpha/2;n-1} < \chi^2 < \chi^2_{\alpha/2;n-1}$
$p = p_0$	n порядка несколь-ких десятков или сотен	$U = \frac{w - p_0}{\sqrt{p_0(1-p_0)}}\sqrt{n}$, где $w = m/n$	$p > p_0$ $p < p_0$ $p \neq p_0$	$U < u_{кр}$, $\Phi_0(u_{кр}) = 1/2 - \alpha$ $U > -u_{кр}$, $\Phi_0(u_{кр}) = 1/2 - \alpha$ $\lvert U\rvert < u_{кр}$, $\Phi_0(u_{кр}) = (1-\alpha)/2$

Предполагается, что случайные величины X и Y являются независимыми и имеют нормальный закон распределения.

Таблица 2. Основные параметрические гипотезы для двух выборок.

H_0	Предположения	Статистика критерия	H_1	Область принятия H_0
$a_x = a_y$	σ_x^2 и σ_y^2 известны	$U = \dfrac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_x^2}{n} + \dfrac{\sigma_y^2}{m}}}$	$a_x > a_y$	$U < u_{кр}$, $\Phi_0(u_{кр}) = 1/2 - \alpha$
			$a_x < a_y$	$U > -u_{кр}$, $\Phi_0(u_{кр}) = 1/2 - \alpha$
			$a_x \neq a_y$	$\lvert U\rvert < u_{кр}$, $\Phi_0(u_{кр}) = (1 - \alpha)/2$
	σ_x^2 и σ_y^2 неизвестны, но равны	$T = \dfrac{\bar{x} - \bar{y}}{s\sqrt{\dfrac{1}{n} + \dfrac{1}{m}}}$, где $s^2 = \dfrac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2}$	$a_x > a_y$	$T < t_{кр}(\alpha, n + m - 2)$ для односторонней области
			$a_x < a_y$	$T > -t_{кр}(\alpha, n + m - 2)$ для односторонней области
			$a_x \neq a_y$	$\lvert T\rvert < t_{кр}(\alpha, n + m - 2)$ для двусторонней области
$\sigma_x^2 = \sigma_y^2$	a_x и a_y неизвестны	$F = \dfrac{s_x^2}{s_y^2}$, где $s_x^2 > s_y^2$	$\sigma_x^2 > \sigma_y^2$	$F < F_{кр}(\alpha, n - 1, m - 1)$
			$\sigma_x^2 \neq \sigma_y^2$	$F < F_{кр}(\alpha/2, n - 1, m - 1)$
$p_1 = p_2$	n_1 и n_2 порядка нескольких десятков или сотен	$U = \dfrac{w_1 - w_2}{\sqrt{w(1-w)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, где $w = \dfrac{m_1 + m_2}{n_1 + n_2}$	$p_1 > p_2$	$U < u_{кр}$, $\Phi_0(u_{кр}) = 1/2 - \alpha$
			$p_1 < p_2$	$U > -u_{кр}$, $\Phi_0(u_{кр}) = 1/2 - \alpha$
			$p_1 \neq p_2$	$\lvert U\rvert < u_{кр}$, $\Phi_0(u_{кр}) = (1 - \alpha)/2$

См. также критерии Колмогорова- Смирнова, Хи-квадрат, Шапиро- Уилка, Бартлета, Тьюки, Левена и др..

ВАРИАНТЫ ЗАДАНИЙ

Вариант	**Федеральный округ (ФО)**	**α**	***M***	$(a;\sigma^2)$	***n***	**Закон распределения случайной величины Y**
1	Центральный ФО	0,025	3000	**(1; 4)**	55	Хи-квадрат распределение с числом степеней свободы k =2 .
2	Южный ФО	0,01	4000	**(2; 9)**	65	Распределение Стьюдента с числом степеней свободы k =4.
3	Приволжский ФО	0,015	5000	**(-1; 5)**	60	Распределение Релея с модой, равной 16.
4	Уральский ФО	0,02	5500	**(-2; 25)**	70	F-распределение с числом степеней свободы $k_1=k_2=8$.
5	Сибирский ФО	0,03	6000	**(3; 9)**	75	Распределение Лапласа с параметром масштаба, равным 2, и параметром сдвига, равным 1.
6	Дальневосточный ФО	0,035	2500	**(- 3; 7)**	80	Распределение Стьюдента с числом степеней свободы k =3.
7	Северо-Западный федеральный округ	0,045	3500	**(0;4)**	50	Хи-квадрат распределение с числом степеней свободы k =5 .
8	Центральный ФО	0,035	4500	**(0; 1,21)**	45	Распределение Стьюдента с числом степеней свободы k =6.
9	Южный ФО	0,055	5500	**(1; 6)**	69	Распределение Релея с модой, равной 10.
10	Приволжский ФО	0,065	6500	**(-2; 6,25)**	90	F-распределение с числом степеней свободы $k_1=k_2=7$.
11	Уральский ФО	0,025	7000	**(-1; 1)**	70	Хи-квадрат распределение с числом степеней свободы k =6.
12	Сибирский ФО	0,01	1500	**(2; 3)**	65	Распределение Стьюдента с числом степеней свободы k =7.
13	Дальневосточный ФО	0,015	3300	**(- 1; 8)**	85	Треугольное распределение на отрезке (1, 4) и модой, равной 2
14	Центральный ФО	0,025	4400	**(0,5 ; 3)**	90	Распределение Релея с модой, равной 7.
15	Южный ФО	0,03	5700	**(-3; 1)**	95	Логистическое распределение с параметрами масштаба и сдвига 9 и 4, соответственно.
16	Приволжский ФО	0,035	3400	**(-2;9)**	100	Показательное распределение с математическим ожиданием, равным 2.
17	Уральский ФО	0,045	4300	**(0, 75; 3)**	55	F-распределение с числом степеней свободы $k_1=3$ и $k_2=7$.
18	Сибирский ФО	0,06	5200	**(3; 9)**	66	Логистическое распределение с параметрами масштаба

						и сдвига 5 и 2, соответственно.
19	Дальневосточный ФО	0,055	4400	**(1; 2)**	90	F-распределение с числом степеней свободы k_1=8 и k_2=7.
20	Северо-Западный федеральный округ	0,065	5600	**(2; 1)**	70	Распределение Стьюдента с числом степеней свободы k =5.
21	Центральный ФО	0,075	5900	**(-1; 3)**	80	Распределение Релея с модой, равной 10.
22	Южный ФО	0,01	3300	**(-2; 9)**	60	Хи-квадрат распределение с числом степеней свободы k =3.
23	Приволжский ФО	0,03	4600	**(3; 7)**	58	F-распределение с числом степеней свободы k_1=5 и k_2=7.
24	Уральский ФО	0,04	5400	**(- 3; 25)**	68	Показательное распределение с математическим ожиданием, равным 100.
25	Сибирский ФО	0,05	6000	**(0; 4)**	73	Логнормальное распределение со средним, равным 2 и стандартным отклонением, равным 1.
26	Дальневосточный ФО	0,025	6500	**(-3; 36)**	45	Распределение Стьюдента с числом степеней свободы k =8.
27	Северо-Западный федеральный округ	0,07	2700	**(3;1)**	56	Распределение Релея с модой, равной 12.
28	Северо-Западный федеральный округ	0,035	3600	**(-3; 1)**	67	Логистическое распределение с параметрами масштаба и сдвига 6 и 2 соответственно.
29	Центральный ФО	0,045	4800	**(4; 4)**	78	Хи-квадрат Хи-квадрат распределение с числом степеней свободы k =4.
30	Южный ФО	0,05	5100	**(- 4; 16)**	87	F-распределение с числом степеней свободы k_1=4 и k_2=7.
31	Приволжский ФО	0,06	4800	**(3; 25)**	66	Распределение Лапласа с параметром масштаба, равным 1, и параметром сдвига, равным 0.