Министерство науки и высшего образования РФ
ФГБОУ ВО «Ульяновский государственный университет»
Факультет математики, информационных и авиационных технологий

Хрусталев С.А.

МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ЛАБОРАТОРНЫМ РАБОТАМ ПО ДИСЦИПЛИНЕ «СТАТИСТИКА ДЛЯ АНАЛИЗА ДАННЫХ»

для студентов всех направлений и специальностей ФМИАТ

Ульяновск

Методические указания к лабораторным работам по дисциплине «Статистика для анализа данных» / составитель: С.А. Хрусталев. – Ульяновск: УлГУ, 2022.

Настоящие методические указания предназначены в помощь студентам очной формы обучения всех направлений и специальностей ФМИАТ для выполнения лабораторных работ по дисциплине «Статистика для анализа данных». В пособии представлены варианты заданий к лабораторным работам, основные формулы, алгоритмы и формы отчета для успешного выполнения и сдачи лабораторных работ.

Методические указания будут полезны студентам при подготовке к лабораторным занятиям и промежуточной аттестации по данной дисциплине.

Рекомендованы к введению в образовательный процесс Ученым Советом Факультета математики, информационных и авиационных технологий УлГУ (протокол № 3/22 от 19 апреля 2022 г.).

МЕТОДИЧЕСКИЕ УКАЗАНИЯ К ЛАБОРАТОРНЫМ РАБОТАМ

1) Лабораторная работа № 1. *Выборочные характеристики.*

Цель лабораторной работы – нахождение по выборке ее выборочных характеристик.

Задание: построить выборочную функцию распределения $F_n(x)$ и гистограмму, вычислить выборочное среднее, выборочную дисперсию и исправленную дисперсию выборки X. Выборка X берется из файла (файлы прилагаются, номер студента в списке группы соответствует номеру файла). Подготовить отчет к лабораторной работе № 1.

Результатом лабораторной работы № 1 является компьютерная программа, написанная на языке программирования высокого уровня или в статистическом пакете, которая выводит значения выборочных моментов, выборочную функцию распределения и гистограмму по данным из файла.

Для выполнения лабораторной работы необходимо построить следующие эмпирические характеристики наблюдаемого распределения:

1. *График эмпирической функции распределения*:
$F_n^*(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le x\},\ x \in \mathrm{R}$, где $X_i, i = 1,2,..,n,$ – элементы выборки, а I – индикаторная функция.

2. *График гистограммы распределения*:
отрезок значений выборки $[X_{(1)}; X_{(n)}]$, где $X_{(1)} = \min\{X_1,...,X_n\}$ и $X_{(n)} = \max\{X_1,...,X_n\}$, разобьем на $k = 1 + [\log_2 n]$ промежутков $A_1,...,A_k$:
$A_1 = [X_{(1)}; X_{(1)} + \Delta],\ A_j = [X_{(1)} + \Delta \cdot (j-1); X_{(1)} + \Delta \cdot j],\ j = 2,...k,\ \Delta = \frac{X_{(n)} - X_{(1)}}{k}$, далее на каждом промежутке $A_1,...,A_k$ строится «столб», высота которого равна $f_j = \frac{\nu_j}{n \cdot \Delta}$, где ν_j вычисляется по формуле $\nu_j = \sum_{i=1}^{n} I\{X_i \in A_j\},\ j = 1,2,..,k.$

3. *Выборочное среднее*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

выборочную дисперсию

$$\sigma^{2*} = \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

исправленную (несмещенную) выборочную дисперсию

$$S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \overline{X})^2$$

4. В отчете к лабораторной работе № 1 представить построенные графики эмпирической функции распределения и гистограммы распределения, а также найденные выборочные характеристики.

Варианты заданий лабораторной работы № 1

Вариант 1

46	41	15	98	14	95	20	88	65	10
60	42	17	32	17	11	19	10	62	91
90	66	18	27	17	95	62	10	45	58
90	83	20	13	16	82	11	62	91	37
80	45	11	83	12	10	91	24	90	66
16	11	11	91	18	10	91	13	92	83

Вариант 2

23	34	26	24	26	22	24	26	25	28
23	26	25	26	28	23	29	27	23	26
30	28	23	15	22	25	23	26	28	26
27	23	29	26	30	28	25	24	22	23
24	20	23	31	27	21	28	24	23	27
25	22	26	24	22	27	22	26	23	24

Вариант 3

24	33	45	33	47	41	21	28	25	21
23	46	28	39	24	23	49	21	47	42
50	47	22	43	39	21	45	34	43	47
27	23	25	26	34	36	37	30	43	49
35	44	30	35	45	39	25	27	42	41
25	37	40	26	48	45	40	23	27	41

Вариант 4

39	57	58	57	38	36	48	46	37	35
52	57	46	52	35	47	32	49	31	33
45	37	58	43	39	49	39	41	44	42
49	37	47	51	34	36	55	59	48	45
54	40	52	52	42	42	47	50	54	53
31	51	31	38	47	39	54	55	43	54

Вариант 5

41	26	36	39	30	30	42	35	36	34
31	49	39	40	50	35	33	33	39	42
36	34	38	32	37	40	33	43	38	33
35	39	28	32	39	33	38	39	34	26
44	35	37	40	26	31	32	41	34	32
37	44	38	40	45	38	34	37	30	28

Вариант 6

46	41	15	98	14	95	20	88	65	10
61	52	17	32	17	11	19	10	62	91
92	66	18	27	15	15	62	10	45	58
93	83	90	10	16	12	11	62	41	37
80	75	11	88	14	10	91	24	30	56
14	81	11	91	12	10	91	13	92	83

Вариант 7

33	34	26	24	26	22	24	26	25	28
43	26	25	26	28	23	29	27	23	26
30	28	73	85	22	25	93	86	28	56
57	63	29	26	90	20	25	24	22	43
24	20	23	31	27	21	28	74	63	37
25	22	26	24	22	27	22	26	23	24

Вариант 8

34	33	45	33	47	41	21	28	25	21
43	76	28	39	24	23	49	21	47	42
50	47	28	83	39	71	45	34	33	47
57	63	29	26	34	36	77	35	43	21
35	44	30	95	45	39	26	27	42	42
25	37	40	26	48	45	40	23	27	46

Вариант 9

39	57	58	57	38	36	48	46	37	35
52	57	46	52	35	47	32	49	31	43
25	90	50	43	30	49	39	30	24	42
68	30	47	51	64	50	55	59	48	55
54	40	52	92	42	42	47	50	24	63
31	51	31	38	47	39	54	55	43	54

Вариант 10

41	26	36	39	30	30	42	35	36	34
31	49	39	40	50	35	33	33	39	42
36	34	38	32	37	40	33	43	38	33
35	39	28	32	39	33	38	39	34	26
44	35	37	40	26	31	32	41	34	32
37	44	38	40	45	38	34	37	30	28

Вариант 11

39	57	58	57	38	36	48	46	37	35
22	57	46	52	35	47	32	49	31	33
45	37	58	43	37	48	37	41	44	32
34	35	46	51	34	36	55	69	48	25
54	40	52	52	42	42	47	50	54	53
31	51	31	38	47	39	54	55	43	24

Вариант 12

41	26	36	39	30	30	42	35	36	34
31	49	39	40	50	35	33	33	39	42
36	34	37	32	27	40	33	40	38	36
34	36	28	32	10	33	38	32	34	26
44	35	37	40	26	32	32	41	34	72
37	44	38	40	45	38	34	37	30	28

2) Лабораторная работа № 2. *Интервальное оценивание.*

Цель лабораторной работы – построение доверительных интервалов математического ожидания в случае выборки X из нормальной генеральной совокупности, при известной и неизвестной дисперсии.

Задание: считая, что выборка принадлежит нормальному распределению, построить доверительные интервалы для среднего генеральной совокупности в случае известной и неизвестной дисперсии, уровень значимости равен: а) 0.01; б) 0.1. Выборка X берется из файла (файлы прилагаются, номер студента в списке группы соответствует номеру файла). Подготовить отчет к лабораторной работе № 2.

Результатом лабораторной работы № 2 является компьютерная программа, написанная на языке программирования высокого уровня или в статистическом пакете, которая выводит значения границ доверительных интервалов для среднего генеральной совокупности, рассчитанные по выборке X при известной/неизвестной дисперсии для заданных уровней значимости в предположении, что выборка принадлежит нормальному распределению.

Рассмотрим алгоритм построения доверительных интервалов для лабораторной работы № 2.

Пусть выборка $X_1,...,X_n$ объема n распределена нормально с параметрами $a \in \mathrm{R}$ и $\sigma^2 > 0$, $X_i \in N(a,\sigma^2)$, $i = 1,2,..,n$. Предположим, что дисперсия σ^2 известна.

Необходимо построить доверительный интервал для параметра a с уровнем доверия $1-\alpha$, где α – уровень значимости. Согласно следствию из леммы Фишера, случайная величина

$$\eta = \sqrt{n}\left(\frac{\overline{X}-a}{\sigma}\right) \in N(0,1).$$

Тогда вероятность $P\{-c < \eta < c\} = 1-\alpha$, где $c > 0$ – α-квантиль стандартного нормального распределения. Найдем значение c.

$P\{-c < \eta < c\} = 2F_{0,1}(c) - 1 = 1-\alpha$, откуда $F_{0,1}(c) = 1-\frac{\alpha}{2}$, $F_{0,1}(x)$ – функция стандартного нормального распределения. Следовательно, $c = \tau_{1-\frac{\alpha}{2}}$ – квантиль уровня $1-\frac{\alpha}{2}$.

Имеем доверительный интервал для среднего генеральной совокупности из нормального распределения при известной дисперсии с уровнем значимости α:

$$\overline{X} - \tau_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < a < \overline{X} + \tau_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}.$$

В случае неизвестной дисперсии σ^2 случайная величина

$$T = \frac{\overline{X}-a}{S/\sqrt{n}} \in t(n-1),$$

где $t(n-1)$ – распределение Стьюдента с $n-1$ степенью свободы, а число S – несмещенное выборочное стандартное отклонение,

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}.$$

Пусть $t_{\alpha,n-1}$ – α-квантиль распределения Стьюдента, тогда из условия

$$P\left\{-t_{1-\frac{\alpha}{2},n-1} < T < t_{1-\frac{\alpha}{2},n-1}\right\} = 1-\alpha$$

следует формула доверительного интервала для среднего генеральной совокупности из нормального распределения при неизвестной дисперсии с уровнем значимости α :

$$\overline{X} - t_{1-\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}} < a < \overline{X} + t_{1-\frac{\alpha}{2},n-1} \cdot \frac{S}{\sqrt{n}}.$$

В отчете к лабораторной работе № 2 представить значения интервальных оценок среднего нормальной генеральной совокупности в случае известной и неизвестной дисперсии с заданными уровнями значимости.

Варианты заданий лабораторной работы № 2

Вариант 1		Вариант 2		Вариант 3		Вариант 4	
X	Y	X	Y	X	Y	X	Y
4.638967	5.553992	4.419943	3.076396	4.999583	4.857998	5.055534	5.191976
4.089126	5.970502	4.548994	5.292002	4.663880	6.409349	4.442746	4.052010
4.132189	4.551515	4.005450	5.705624	4.704113	4.250146	6.221683	3.506964
4.152145	5.351677	4.209082	4.286663	4.806672	4.024131	5.319702	3.775698
4.645826	4.887804	4.072452	5.917309	5.136867	2.727740	5.348528	4.791589
4.351139	4.981300	4.254048	5.998586	3.937141	3.691379	5.092915	4.865322
4.967062	4.885592	4.361640	4.662639	4.001890	3.978903	6.289782	3.677449
4.105800	4.839377	5.288485	5.741436	4.033692	2.799862	5.444793	5.288073
5.327888	4.784066	4.653330	6.016830	4.670020	6.178269	4.737960	5.062007
4.978439	5.336682	4.320630	5.885262	4.807568	5.954304	5.691439	3.526427
4.222068	4.572902	4.289719	3.555368	4.444367	3.981436	6.011168	4.794958
5.027021	4.704992	3.334606	4.666018	4.505929	3.431112	5.466133	4.476164
4.533238	4.410001	5.295522	3.776803	4.603523	6.202510	4.135636	5.384824
4.161094	5.013749	4.850976	5.800125	4.675631	2.674191	4.945810	3.826881
4.689492	4.140494	3.940006	3.314530	3.885421	5.191473	4.852538	3.240476
3.988046	3.992917	4.242548	4.644782	4.756538	4.949176	5.481944	3.089173
4.721502	4.334615	4.872242	4.257029	4.759050	4.710497	5.642508	4.477848
4.475280	5.562022	3.695416	3.175627	4.337255	5.489007	5.076067	3.384481
3.410654	4.050031	4.583352	4.397477	3.548028	5.661473	4.866458	4.307270
4.932859	3.955142	4.931963	4.889673	4.582474	3.804387	4.197598	4.711493
4.630417	4.291431	4.653260	5.030411	4.094886	4.360500	4.186838	3.500882
4.579480	4.498089	4.464856	3.291653	4.572529	4.078886	5.327632	5.216211
4.358248	5.819647	3.216372	5.551372	5.346949	4.987529	4.237562	4.441075
4.544732	5.772039	3.950228	3.106513	4.276971	5.827909	5.404188	4.721224
4.717440	3.725623	3.819444	4.309444	3.963361	6.350614	4.565123	4.856527
4.685312	4.908290	4.518748	6.021270	3.321524	5.265645	5.093837	5.545297
4.863511	5.443042	4.108789	4.313884	4.158769	5.116094	4.649510	5.105049
4.605265	4.503989	3.252901	4.730788	4.449788	5.630960	5.108003	4.389892
5.183908	5.128877	4.079489	4.632716	4.434878	4.591821	4.392779	3.693075
4.611328	5.548747	5.498051	5.124815	4.259183	4.812892	5.063349	2.754660
4.766582	5.327340	2.486262	5.429554	4.651718	3.672565	5.094495	4.376886
4.690993	5.495485	5.083891	4.557521	4.084310	2.909493	5.116772	4.210238
5.058987	4.425816	4.344382	5.932464	4.969735	3.279029	6.143696	5.045071
4.309048	5.532031	5.514066	3.181418	4.805277	3.851664	5.340153	3.532977
4.852944	5.417558	5.319868	3.202655	4.389871	2.969434	4.764729	5.199836
5.348331	5.543421	4.514862	5.630622	3.413095	5.481650	5.170935	3.398797
4.697955	5.124862	4.503800	3.429688	4.737927	3.827784	5.628836	3.733778
4.506967	5.243186	4.920289	5.091706	3.994193	4.319856	5.275345	5.254762
5.407230	4.974089	5.206850	5.744428	4.787106	3.817292	4.496465	4.359201

Вариант 5		Вариант 6		Вариант 7		Вариант 8	
X	*Y*	*X*	*Y*	*X*	*Y*	*X*	*Y*
5.281689	4.089051	4.605805	2.949930	4.256544	3.898083	4.362190	5.146098
5.066087	4.715469	4.861973	3.985668	4.384818	6.158232	4.519163	3.889866
5.201108	5.220933	4.209134	5.268052	4.196824	3.281270	4.691275	3.830971
4.193511	5.131418	5.246013	3.624995	4.088477	3.085748	3.126319	3.374886
5.278241	3.581046	4.434275	4.366081	3.839396	2.910191	3.860564	2.863365
6.023275	3.837825	4.619586	5.738058	4.221758	2.967016	4.509837	3.077858
4.904945	3.288027	5.826499	5.545265	3.728300	5.978304	4.209502	4.852035
5.147641	5.461146	4.871204	3.208649	3.735244	4.665525	4.040082	3.146390
5.772398	5.421707	4.320415	5.468894	3.659475	6.485718	4.042151	3.119784
5.534241	4.714904	5.826696	5.443598	4.227907	5.665231	5.417557	5.078801
5.803777	5.106098	4.042286	4.599695	4.172860	6.129524	4.389052	5.057467
5.159464	3.350246	5.421813	3.829766	4.312751	3.523459	4.381926	4.926663
4.863457	4.831528	5.172592	3.231742	4.207900	4.618150	4.531429	5.007716
4.630714	4.355055	4.549318	5.083595	3.924240	6.395577	4.187778	4.209381
4.440453	5.378126	3.628714	4.256365	4.808291	4.695414	5.171210	4.988276
2.985251	3.872088	3.671452	4.235763	4.297098	5.551698	5.278497	4.068692
4.694641	5.730727	5.162868	3.641380	3.842002	6.065374	3.976291	2.983955
4.311236	5.076353	4.041153	2.799106	4.093832	3.628840	5.117583	4.203780
4.749141	4.974037	4.435064	5.564908	3.825412	4.931105	5.009902	3.625869
4.610925	3.699176	4.827710	3.654604	4.870298	4.285821	4.204641	5.062163
5.253718	2.919613	5.767499	4.851131	4.089081	5.469708	5.233014	3.935499
4.367430	4.075968	5.311336	3.880552	4.997985	5.966372	4.318184	4.514481
4.766729	2.994726	4.603111	2.780900	3.441144	5.143049	4.808135	4.997008
4.392226	4.864002	5.018687	4.448679	4.895105	5.618802	4.995469	4.622139
4.563418	3.604295	5.324082	3.523519	3.889804	3.640654	3.342634	3.554948
4.859523	5.778261	6.233567	4.352666	3.193811	4.045168	4.784507	5.228469
4.691661	4.084439	5.490405	3.601997	4.968126	3.023487	4.104858	5.174598
6.242247	3.104102	5.447568	5.759426	4.660930	5.493573	4.870563	3.936076
4.636382	3.929788	5.860401	2.896749	3.121998	5.392090	4.639569	3.735998
4.007847	5.757177	5.516620	5.412647	3.735101	5.056807	3.747842	4.241176
4.502763	4.847059	4.636389	5.134955	5.537353	4.000747	4.748564	4.484169
5.083091	5.011500	5.653951	3.478579	4.672340	3.748753	4.183328	5.261252
4.497426	4.857884	5.337584	4.179132	4.217594	6.273301	4.482317	5.093216
4.342907	3.460563	4.704554	3.531664	3.513958	5.125092	3.804129	5.138108
5.473684	4.547263	5.678918	5.315963	4.609426	3.364087	3.668699	2.741704
4.605296	4.706339	5.842530	5.740262	4.056170	4.967138	4.741865	3.414588
4.580597	4.137056	4.741642	2.734235	3.736373	5.999805	4.522710	3.600169
4.052529	4.581149	5.576802	5.276580	3.715809	5.394807	5.258399	4.154687
4.363632	3.951437	4.224607	3.791150	4.002394	6.427947	5.744475	3.881958

3) Лабораторная работа № 3. *Проверка статистических гипотез. Критерий согласия Пирсона*.

Цель лабораторной работы – проверка статистических гипотез с помощью критерия согласия Пирсона.

Задание: для выборок X и Y проверить статистические гипотезы с помощью критерия согласия Пирсона:
– о нормальном распределении, параметры оцениваются по выборке, уровень значимости равен: а) 0.01, б) 0.1;
– о равенстве дисперсий в предположении, что обе выборки принадлежат нормальному распределению: $X \sim N(a_1, \sigma_1)$, $Y \sim N(a_2, \sigma_2)$;
– о равенстве средних в предположении, что обе выборки принадлежат нормальному распределению: $X \sim N(a_1, \sigma_1)$, $Y \sim N(a_2, \sigma_2)$, дисперсии равны, но неизвестны.

Подготовить отчет к лабораторной работе № 3.

Выборки X и Y берутся из файла (файлы прилагаются, номер студента в списке группы соответствует номеру файла).

Результатом лабораторной работы № 3 является компьютерная программа, написанная на языке программирования высокого уровня или в статистическом пакете, или таблицы, сформированные в MS Excel, с представлением исходных данных, этапов проверки гипотез (промежуточных расчетов, критических значений полученных распределений) и конечных результатов (основная гипотеза верна или нет).

а) *Критерий согласия Пирсона, или* χ^2*–критерий*.

Проверим, согласуются ли данные наблюдений $X_1, ..., X_n$ с гипотезой $H_0 : F(x) = F^*(x)$ для любого x. Гипотеза H_1 является отрицанием H_0.

Множество значений случайной величины X разобьем на k непересекающихся интервалов $\Delta_i = (d_{i-1}\,;\, d_i]$, $i = 1,2,...,k$. Для каждого интервала необходимо найти:
– число наблюдений ν_i, которые принадлежат интервалу Δ_i;
– вероятность p_i^* попадания случайной величины X, имеющей функцию распределения $F^*(x)$, в интервал Δ_i:

$$p_i^* = P\{X \in \Delta_i \,|\, H_0\} = F^*(d_i) - F^*(d_{i-1})$$

Поскольку наблюдения случайны, то эмпирические частоты ν_i / n будут отличаться от теоретических вероятностей p_i^*. В качестве меры расхождения между ними можно взять случайную величину

$$\chi^2 = \sum_{i=1}^{k} \frac{(\nu_i - np_i^*)^2}{np_i^*},$$

которую называют статистикой Пирсона или статистикой χ^2 (хи-квадрат) с $k - 1$ степенью свободы.

Английский статистик Пирсон доказал, что, в предположении справедливости гипотезы H_0, распределение случайной величины χ^2 при n, стремящемся к бесконечности, сходится к известному распределению χ^2 с $k-1$ степенью свободы.

Большие значения статистики χ^2 указывают на то, что эмпирическое распределение не согласуется с предполагаемым распределением $F^*(x)$. «Большими» считаем значения статистики, удовлетворяющие неравенству $\chi^2 \geq \chi^2_{1-\alpha,r}$. Критическое значение $\chi^2_{1-\alpha,r}$ находится при заданном уровне значимости α из условия $P\{\chi^2 \geq \chi^2_{1-\alpha,r}\} = \alpha$. Таким образом, критическая область $D_1 = [\chi^2_{1-\alpha,r}; +\infty)$ является правосторонней. Значение квантили $\chi^2_{1-\alpha,r}$ порядка $1-\alpha$ можно найти по таблице квантилей распределения хи-квадрат с r степенями свободы.

На практике критерий Пирсона состоит в следующем: вычисляем наблюдаемое значение статистики χ^2. Если $\chi^2 < \chi^2_{1-\alpha,r}$, то принимаем гипотезу H_0, т. е. считаем, что данные наблюдений согласуются с выбранным распределением. Если $\chi^2 \geq \chi^2_{1-\alpha,r}$, то отвергаем гипотезу H_0.

Критерий Пирсона используется, если объем выборки достаточно большой ($n \geq 50$), при этом в каждом интервале число наблюдений ν_i должно быть не меньше 5. Интервал Δ_i, для которого $\nu_i < 5$, объединяют с соседним.

б) *Проверка гипотезы о равенстве дисперсий двух нормально распределенных генеральных совокупностей. Критерий Фишера (F – критерий).*

Рассмотрим независимые случайные величины $X \sim N(a_1, \sigma_1)$, $Y \sim N(a_2, \sigma_2)$ с неизвестными математическими ожиданиями и дисперсиями.

Пусть имеются независимые выборки $X_1, ..., X_{n_1}$ и $Y_1, ..., Y_{n_2}$, отвечающие соответственно X и Y. По выборкам найдем несмещенные оценки S_1^2 и S_2^2 дисперсий σ_1^2 и σ_2^2. Проверим при уровне значимости α гипотезу о равенстве дисперсий величин X и Y $H_0 : \sigma_1^2 = \sigma_2^2$ при альтернативе $H_1 : \sigma_1^2 \neq \sigma_2^2$.

Критерий основан на статистике

$$F_{n_1-1,n_2-1} = \frac{S_1^2}{S_2^2},$$

которая, при условии справедливости гипотезы H_0, имеет распределение Фишера с ($n_1 - 1$, $n_2 - 1$) степенями свободы.

Пусть $k = n_1 - 1$, $m = n_2 - 1$.

Если гипотеза H_0 верна, и $\sigma_1^2 = \sigma_2^2$, то отношение $\frac{S_1^2}{S_2^2}$ должно быть близким к 1. Поэтому при альтернативе H_1 гипотеза H_0 должна быть отвергнута при «малых» и «больших» значениях статистики $F_{k,m}$. Следовательно, критическая область является двусторонней, ее

вид определяется критическими значениями $f_{k,m}(\alpha/2)$ и $f_{k,m}(1-\alpha/2)$, где $f_{k,m}(p)$ – квантиль порядка p статистики Фишера. «Малые» ($f_{k,m} < f_{k,m}(\alpha/2)$) и «большие» ($f_{k,m} > f_{k,m}(1-\alpha/2)$) значения статистики $F_{k,m}$ образуют критическую область.

При составлении таблицы квантилей учитывается, что числитель статистики Фишера всегда равен большей из двух выборочных дисперсий, пусть для определенности $S_1^2 > S_2^2$. Следовательно, возможные значения статистики Фишера больше 1. Поскольку при малых значениях α $f_{k,m}(\alpha/2) < 1$, то контролируем попадание наблюдаемых значений статистики Фишера $f_{k,m}$ в правую часть критической области ($f_{k,m}(1-\alpha/2)$; $+\infty$). Если наблюдаемое значение тестовой статистики $f_{k,m} > f_{k,m}(1-\alpha/2)$, то выборочные дисперсии различаются значимо, и гипотеза H_0 отклоняется в пользу гипотезы H_1.

в) *Проверка гипотезы о равенстве математических ожиданий двух нормально распределенных генеральных совокупностей в случае равных, но неизвестных дисперсий*.

Рассмотрим независимые случайные величины $X \sim N(a_1, \sigma_1)$, $Y \sim N(a_2, \sigma_2)$ с неизвестными математическими ожиданиями и неизвестными, но равными дисперсиями.

Пусть имеются независимые выборки $X_1, ..., X_{n_1}$ и $Y_1, ..., Y_{n_2}$, отвечающие соответственно X и Y. По этим выборкам найдены оценки $\overline{X}$ и $\overline{Y}$ неизвестных математических ожиданий a_1 и a_2:

$$\overline{X} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_i \text{ и } \overline{Y} = \frac{1}{n_2}\sum_{j=1}^{n_2} Y_j$$

Проверим гипотезу о совпадении математических ожиданий $H_0 : a_1 = a_2$ при альтернативе $H_1 : a_1 \neq a_2$ в предположении, что дисперсии неизвестны и равны, $\sigma_1^2 = \sigma_2^2$. По выборкам найдем несмещенные оценки S_1^2 и S_2^2 дисперсий σ_1^2 и σ_2^2. Объединим две выборки в одну выборку объема $n_1 + n_2$ и найдем для нее смешанную выборочную дисперсию:

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Выборочная дисперсия S^2 является лучшей оценкой для $\sigma^2 = \sigma_1^2 = \sigma_2^2$.

Гипотеза H_0 проверяется с помощью статистики Стьюдента:

$$T_{n_1+n_2-2} = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

При условии справедливости гипотезы H_0 статистика $T_{n_1+n_2-2}$ имеет распределение Стьюдента с $n_1 + n_2 - 2$ степенями свободы. Тогда при $H_1 : a_1 \neq a_2$ критическая область имеет вид $D_1 = (-\infty ; t_{\alpha/2,m}] \cup [t_{1-\alpha/2,m} ; +\infty)$, где $m = n_1 + n_2 - 2$.

Замечание. Прежде чем проверять гипотезу о равенстве математических ожиданий, надо проверить гипотезу о равенстве дисперсий случайных величин X и Y.

В отчете к лабораторной работе № 3 представить таблицы с этапами проверки гипотез (промежуточные расчеты, критические значения соответствующих распределений) и конечными результатами (основная гипотеза верна или нет).

Варианты заданий лабораторной работы № 3 совпадают с вариантами заданий лабораторной работы № 2.

4) Лабораторная работа №4. *Регрессионный анализ.*

Цель лабораторной работы – нахождение по выборке уравнения линейной регрессии, проверка адекватности построенной модели и значимости коэффициентов.

Задание: найти для выборок *X* и *Y* уравнения линейной регрессии, проверить адекватность построенной модели и значимость коэффициентов. Подготовить отчет к лабораторной работе № 4.

Выборки *X* и *Y* берутся из файла (файлы прилагаются, номер студента в списке группы соответствует номеру файла).

Результатом лабораторной работы № 4 является компьютерная программа, написанная на языке программирования высокого уровня или в статистическом пакете, или таблицы, сформированные в MS Excel, с представлением исходных данных, этапов построения регрессии (промежуточных расчетов коэффициентов регрессии) и конечных результатов (уравнение линейной регрессии с графическим отображением).

Модель регрессии

Рассмотрим модель, в которой наблюдаемая случайная величина *X* зависит от случайной величины *Y*.

Пусть зависимость математического ожидания *X* от значений *Y* определяется формулой $E(X|Y=t)=f(t)$, где f – неизвестная функция.

После *n* экспериментов с входными данными получены значения $X_1,...,X_n$, заданными в виде

$$X_1 = f(t_1)+\varepsilon_1, \dots, X_n = f(t_n)+\varepsilon_n,$$

где ε_i – ошибки наблюдения, равные разнице между реальным и усредненным значением случайной величины *X* при значении $Y=t_i$:

$$\varepsilon_i = X_i - f(t_i) = X_i - E(X|Y=t_i).$$

Требуется по значениям $t_1,...,t_n$ и $X_1,...,X_n$ оценить как можно точнее функцию *f*.

Метод наименьших квадратов

Функцию *f* можно восстановить лишь приближенно, в виде полинома. Предположим, что функция f – полином с неизвестными коэффициентами. Метод наименьших квадратов состоит в выборе коэффициентов, при которых сумма квадратов ошибок минимальна:

$$\sum_{i=1}^{n}(X_i - f(t_i))^2$$

Пусть $X_i = \theta + \varepsilon_i$, $i = 1,...,n$, где θ – неизвестный параметр.

Найдем оценку $\widehat{\theta}$ для параметра θ, при которой достигается минимум величины

$$\sum_{i=1}^{n} (X_i - \theta)^2$$

из условия

$$\frac{\partial}{\partial \theta} \sum_{i=1}^{n} (X_i - \theta)^2 \Big|_{\theta=\hat{\theta}} = 0,$$

откуда $\hat{\theta} = \overline{X}$.

Оценка $\widehat{\theta}$ параметра θ, при которой достигается $\min\limits_{\theta} \sum\limits_{i=1}^{n} \varepsilon_i^2$, называется оценкой метода наименьших квадратов.

В лабораторной работе № 4 в качестве уравнения линейной регрессии выберем

$$X_i = \theta_1 + Y_i \theta_2 + \varepsilon_i,\ i = 1,...,n,$$

где $\theta = (\theta_1, \theta_2)$ – неизвестный параметр.

Найдем оценку метода наименьших квадратов для параметра θ.

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (X_i - \theta_1 - t_i \theta_2)^2$$

Приравняв к нулю частные производные по θ_1 и θ_2, найдем оценку $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$:

$$\hat{\theta}_1 = \overline{X} - \hat{\theta}_2 \bar{t},$$

$$\hat{\theta}_2 = \frac{\bar{r} - \overline{X} \cdot \bar{t}}{\bar{s} - (\bar{t})^2},$$

где $\bar{r} = \frac{1}{n} \sum\limits_{i=1}^{n} X_i t_i$, $\bar{s} = \frac{1}{n} \sum\limits_{i=1}^{n} t_i^2$, $\bar{t} = \frac{1}{n} \sum\limits_{i=1}^{n} t_i$.

В отчете к лабораторной работе № 4 представить уравнение линейной регрессии со всеми найденными коэффициентами и графическим отображением.

Варианты заданий лабораторной работы № 4 совпадают с вариантами заданий лабораторной работы № 2.