Министерство образования и науки Российской Федерации

федеральное государственное бюджетное образовательное учреждение высшего профессионального образования

«Российский государственный университет нефти и газа(НИУ) имени И. М. Губкина»

Кафедра «Автоматизированные системы управления»

Дисциплина «Моделирование систем»

О Т Ч Е Т

по лабораторной работе №1

«Статистический анализ одномерных выборок»

Выполнил:

Кононенко Богдан

Группа АА-19-05

Преподаватель: Степанкина О.А.

Москва 2021 г.

1. Провести предварительный анализ данных, включающий:

- оценку числовых характеристик (по 2 в каждой из групп; оформить в виде таблицы);
- графический анализ;
- предварительное заключение о законе распределения каждой случайной величины.

Считывание выборок

```
file = 'ms-data1.xlsx';

A=xlsread(file);
```

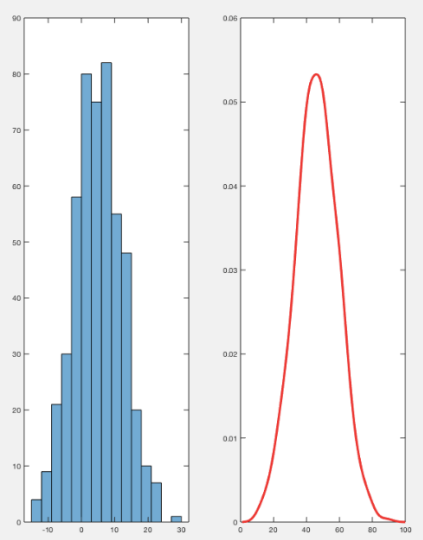
Оценка числовых характеристик

```
% ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ

mat_ozh = mean(A);

mediana = median(A);

sr_otkl = std(A);

sr_oshibka = std(A)/sqrt(length(A));

moda = mode(A);

dispersia = var(A);

axscess = kurtosis(A);

assim = skewness(A);

int = abs(min(A))+abs(max(A));
```

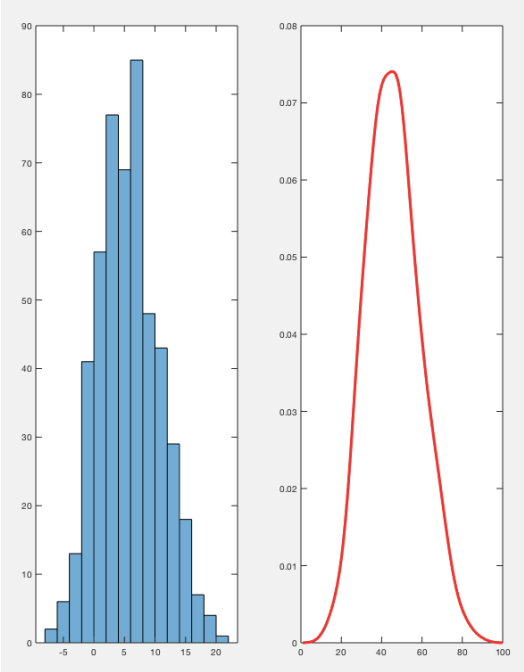
1	2	3	4	5	6	7	
5,00	5,75	7,13	6,78	7,20	4,98	9,98	*мат ожидание*
5,00	5,60	6,79	6,84	7,58	5,00	10,00	*медиана*
7,13	5,01	5,06	3,05	6,34	2,23	2,42	*ср отклонение*
0,32	0,22	0,23	0,14	0,28	0,10	0,11	*ср ошибка*
7,46	-1,18	3,19	0,15	0,92	3,00	9,00	*мода*
2,93	2,81	3,21	2,89	1,82	2,80	2,84	*эксцесс*
0,02	0,28	0,14	-0,04	-0,06	0,35	0,06	*коэффф ассиметрии*
50,80	25,08	25,63	9,32	40,24	4,99	5,86	*дисперсия*
5	6	7	7	7	5	10	*Среднее*

1. Графический анализ выборок

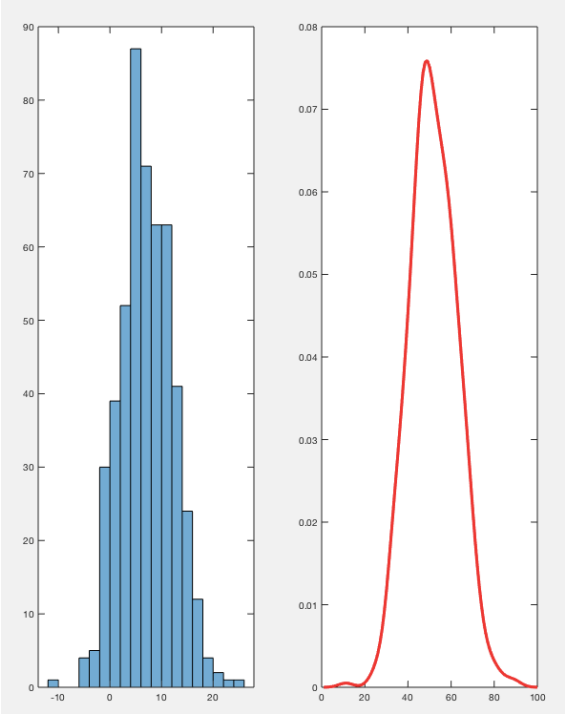
```
% ГРАФИЧЕСКИЙ АНАЛИЗ

% построение гистограммы, т.е. дискретного аналога ненормированной функции распределения

subplot(1,4,1);

h = histogram(x);  % гистограмма

subplot(1,4,2);

plot(ksdensity(x),'-r','LineWidth',3) % плотность распределения

% ПРЕДВАРИТЕЛЬНОЕ ЗАКЛЮЧЕНИЕ:
```



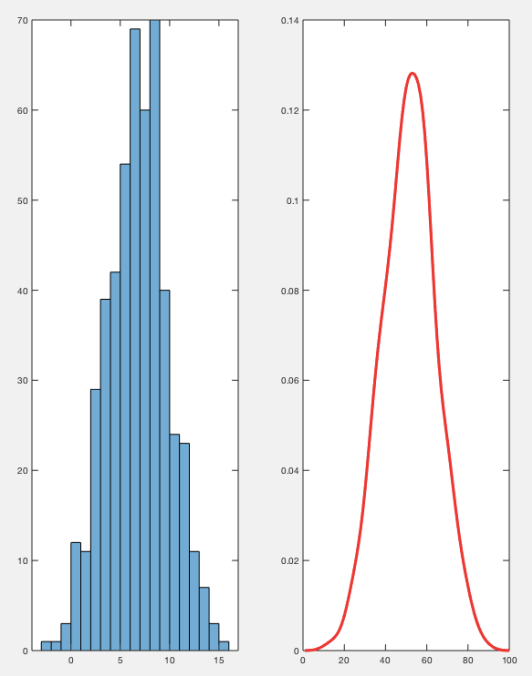

```
% 1 выборка нормальное распределение
```



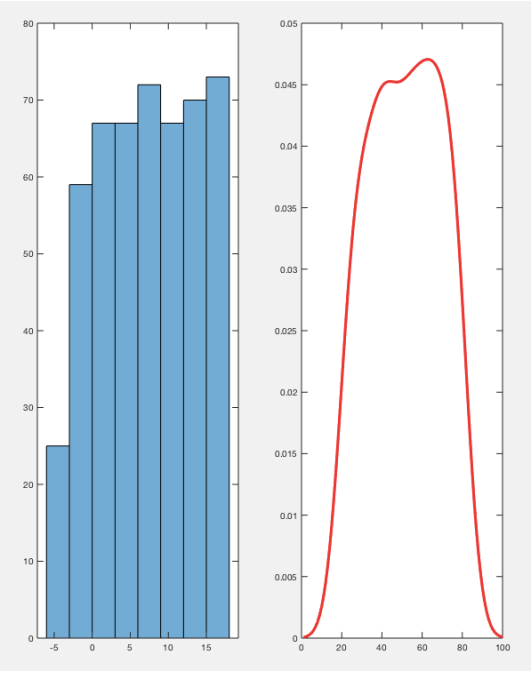
% 2 выборка нормальное распределение



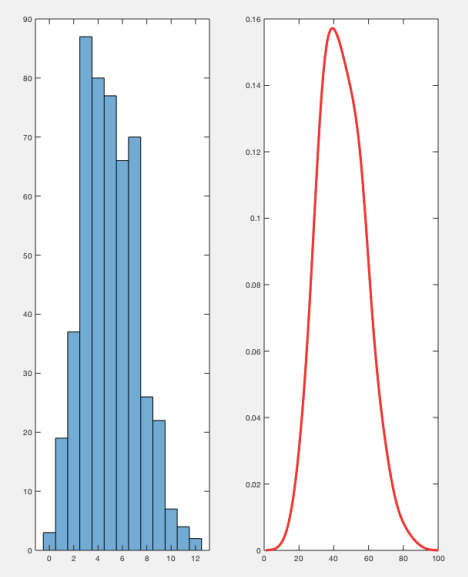
% 3 выборка нормальное распределение



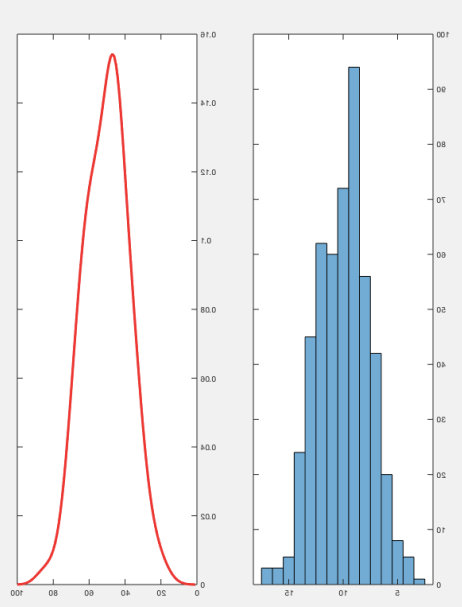
% 4 выборка нормальное распределение



% 5 выборка равномерное распределение



% 6 выборка нормальное распределение



% 7 выборка нормальное распределение

2. Провести проверку на выбросы. В случае выброса – повторить п.1, дополнить таблицу.

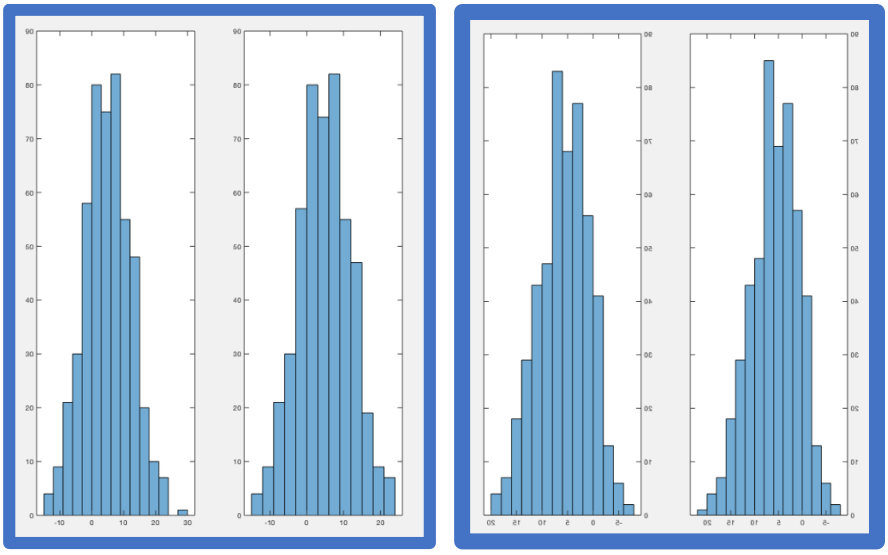
Проверка на выбросы

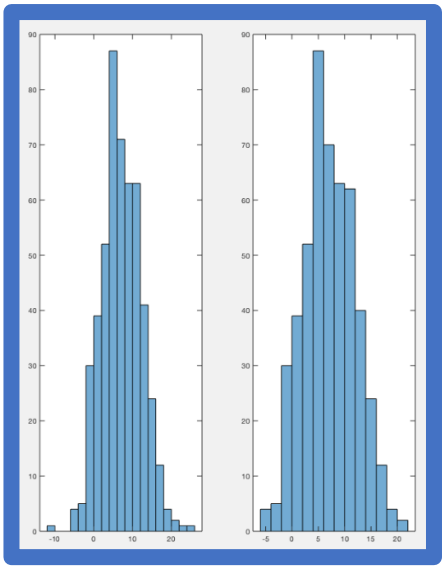
Выброс — в статистике результат измерения, выделяющийся из общей выборки.

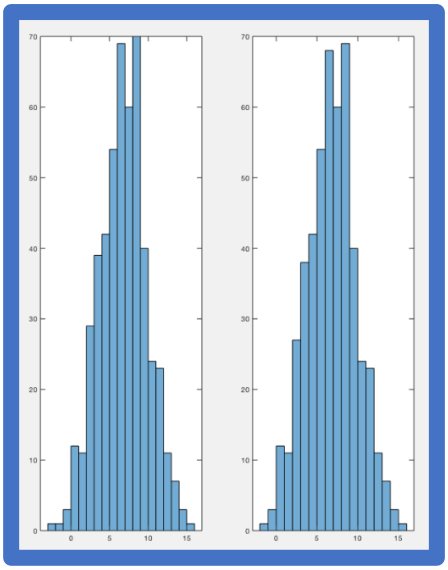
B = rmoutliers(A) обнаруживает и удаляет выбросы из данных в векторе, матрице, таблице или расписании.

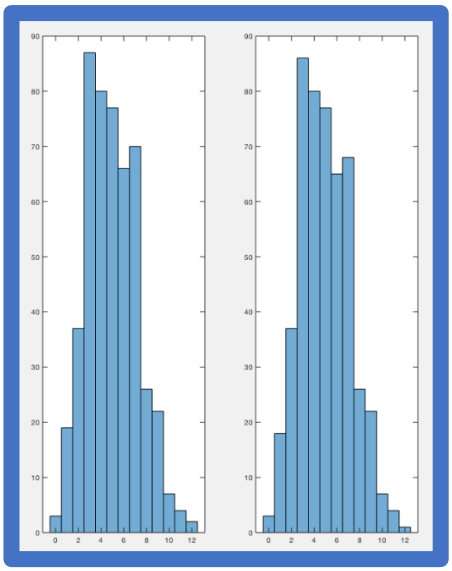
- Если A строка или вектор-столбец, rmoutliers обнаруживает выбросы и удаляет их.
- Если A матрица, таблица, или расписание, rmoutliers обнаруживает выбросы в каждом столбце или переменной A отдельно и удаляет целую строку.

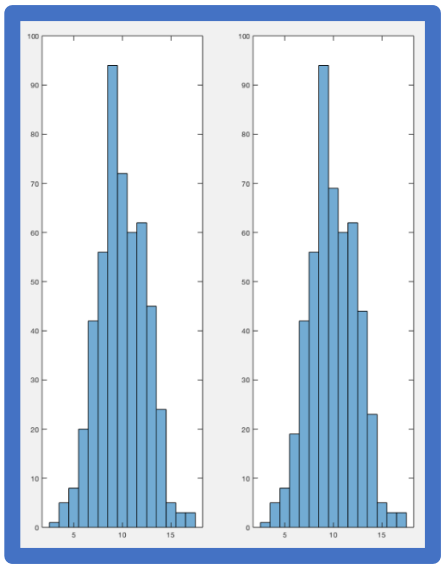
```
% ПРОВЕРКА НА ВЫБРОСЫ
% выброс
A = rmoutliers(B)
x=A(:,n);
subplot(1,3,3);
h = histogram(x);
```











Обновление данных:

1	2	3	4	5	6	7	
5,00	5,75	7,13	6,78	7,20	4,98	9,98	*мат ожидание*
5,00	5,60	6,79	6,84	7,58	5,00	10,00	*медиана*
7,13	5,01	5,06	3,05	6,34	2,23	2,42	*ср отклонение*
0,32	0,22	0,23	0,14	0,28	0,10	0,11	*ср ошибка*
7,46	-1,18	3,19	0,15	0,92	3,00	9,00	*мода*
2,93	2,81	3,21	2,89	1,82	2,80	2,84	*эксцесс*
0,02	0,28	0,14	-0,04	-0,06	0,35	0,06	*коэфф ассиметрии*
50,80	25,08	25,63	9,32	40,24	4,99	5,86	*дисперсия*
5	6	7	7	7	5	10	*Среднее*

3. Проверить каждую выборку на принадлежность к закону распределения, о котором было сделано предположение в пункте 1, используя критерии хи-квадрат и Колмогорова-Смирнова.

Проверить каждую выборку на принадлежность к закону распределения.

В данном случае **критерий Колмогорова** используется для проверки гипотезы о принадлежности наблюдаемой выборки нормальному закону, параметры которого оцениваются по этой самой выборке методом максимального правдоподобия. То есть, проверяется сложная гипотеза и в качестве оценок параметров нормального закона используются выборочные оценки среднего и дисперсии.

Одновыборочный критерий проверки нормальности Колмогорова-Смирнова основан на максимуме разности между кумулятивным распределением выборки и предполагаемым кумулятивным распределением:

$$D_n = \sup_x |F_n(x) - F(x)|$$

$F_n(x)$ - кумулятивное распределение выборки

$F(x)$ - ожидаемое кумулятивное распределение (с известными параметрами)

Если **D** статистика Колмогорова-Смирнова значима, то гипотеза о том, что соответствующее распределение нормально, должна быть отвергнута.

Выводимые значения вероятности основаны на предположении, что среднее и стандартное отклонение нормального распределения известны априори и не оцениваются из данных.

Однако на практике обычно параметры вычисляются непосредственно из данных.

y=normcdf(x,mean(x),std(x)); % возвращает кумулятивную функцию распределения (cdf) стандартного нормального распределения, вычисляемого в значениях в x

CDF = [x y];

k1 = kstest(x,CDF,0.01); % возвращает тестовое решение для нулевой гипотезы что данные в векторном x прибывает из стандартного нормального распределения, против альтернативы, что она не прибывает из такого распределения, с помощью одновыборочного критерия Колмогорова-Смирнова. Результат h 1 если тест отклоняет нулевую гипотезу на 5%-м уровне значения или 0 в противном случае.

Если критерий = 0, то мы можем принять гипотезу о нормальном распределении, иначе (=1) - отклонить.

(H0 – выборка подчиняется нормальному закону распределения, H1 – опровержение предположения о нормальности)

У стандартного нормального распределения почти все значение находятся в пределах ±3 (правило трех сигм). Таким образом, мы получили относительную разность в частотах для одной группы. Нам нужна обобщающая мера. Просто сложить все отклонения нельзя – получим 0 (догадайтесь почему). Пирсон предложил сложить квадраты этих отклонений.

$$\chi_n^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Это и есть знамений **критерий Хи-квадрат Пирсона**. Если частоты действительно соответствуют ожидаемым, то значение критерия будет относительно не большим (т.к. большинство отклонений находится около нуля). Но если критерий оказывается большим, то это свидетельствует в пользу существенных различий между частотами.

«Большим» критерий Пирсона становится тогда, когда появление такого или еще большего значения становится маловероятным. И чтобы рассчитать такую вероятность, необходимо знать распределение критерия при многократном повторении эксперимента, когда гипотеза о согласии частот верна.

CDF1=@(z)cdf('norm',z,mean(x),std(x)); % нормальная функция распределения

h1 = chi2gof(x,'cdf',CDF1); % возвращает тестовое решение для нулевой гипотезы что данные в векторном x прибывает из нормального распределения со средним значением и отклонением, оцененным от x, использование критерия согласия Хи-квадрат.

Если Хи кв. =0, то мы можем принять гипотезу о нормальном распределении, иначе (=1) - отклонить.

(H0 – выборка подчиняется нормальному закону распределения, H1 – опровержение предположения о нормальности)

Параметры законов распределения:

[muhat,sigmahat,muci,sigmaci] = normfit(x) % для нормального

[a,b,a_int,b_int]=unifit(x) % для равномерного

Ст	Предположение о Законе	Параметры	Хи кв.	К.-С.	
1	норм	(5,00; 7,12)	0	0	✓
2	норм	(5,75; 5)	0	0	✓
3	норм	(7,13; 5,06)	0	0	✓
4	норм	(6,78;3,05)	0	0	✓
5	равномерн	(-3,96;17,67)	0	1	~
6	норм		1	1	X
7	норм		1	1	X

Для выборок 1-4 нормальный закон распределения был подтверждён, а для 6-7 – нет.

4. Можно ли в качестве оценки математического ожидания использовать округленное до целого среднее значение?

1	2	3	4	5	6	7	
5,00	5,75	7,13	6,78	7,20	4,98	9,98	*мат ожидание*
5	6	7	7	7	5	10	*Среднее*

Математическое ожидание — это ни что иное, как среднее арифметическое наблюдаемых значений интересующего нас признака.

5. Для нормально распределенных случайных величин проверить: μ1=μ2; s21=s22.

1	2	3	4	
5,0	5,8	7,1	6,8	*мат ожидание*
50,8	25,1	25,6	9,3	*дисперсия*

Можем заметить, что выборки 3 и 4 имеют близкие мат.ожидания, значит, у них пик графика будет находиться примерно на одной линии по оси OX, но так как дисперсии не равны, то графики будут располагаться на разной высоте по OУ.

Выборки 2 и 3 имеют схожие дисперсии, но разные мат.ожидания, значит, графики будут находиться примерно на одной высоте по OУ, но их пик будет находиться на разных координатах по оси OX.

6. Проверить однородность тех же (п. 3) выборок, используя критерии
- Колмогорова-Смирнова,
- Мана-Уитни или Уилкоксона

Проверить однородность выборок

Исходя из предыдущего пункта, попарно объединим 3 и 2 выборки, а также 3 и 4.

Гипотезы об однородности выборок – это гипотезы о том, что рассматриваемые выборки извлечены из одной и той же генеральной совокупности.

Пусть имеются две независимые выборки, произведенные из генеральных совокупностей с неизвестными теоретическими функциями распределения $F_1(x)$ и $F_2(x)$.

Проверяемая нулевая гипотеза имеет вид $H_0: F_1(x) = F_2(x)$ против конкурирующей $H_1 : F_1(x) \neq F_2(x)$. Будем предполагать, что функции $F_1(x)$ и $F_2(x)$ непрерывны и для оценки используем статистику Колмогорова – Смирнова.

Критерий Колмогорова-Смирнова использует ту же самую идею, что и критерий Колмогорова. Однако различие заключается в том, что в критерии Колмогорова сравнивается эмпирическая функция распределения с теоретической, а в критерии Колмогорова-Смирнова сравниваются две эмпирические функции распределения.

Статистика критерия Колмогорова-Смирнова имеет вид:

$$\lambda' = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \cdot \max \left| F_{n_1}(x) - F_{n_2}(x) \right|,$$

где $F_{n_1}(x)$ и $F_{n_2}(x)$ – эмпирические функции распределения, построенные по двум выборкам с объемами n_1 и n_2.

Гипотеза H_0 отвергается, если фактически наблюдаемое значение статистики λ' больше критического λ'_{kp}, т.е. $\lambda' > \lambda'_{\text{kp}}$, и принимается в противном случае.

При малых объемах выборок $(n_1, n_2 \le 20)$ критические значения λ'_{kp} для заданных уровней значимости критерия можно найти в специальных таблицах. При $n_1, n_2 \to \infty$ (а практически при $n_1, n_2 \ge 50$) распределение статистики λ' сводится к распределению Колмогорова для статистики λ. В этом случае гипотеза H_0 отвергается на уровне значимости α, если фактически наблюдаемое значение λ' больше критического λ_α, т.е. $\lambda' > \lambda_\alpha$, и принимается в противном случае.

```
%Критерий Колмогорова-Смирнова
kstest_3_4 = kstest2(A(:,3),A(:,4));
disp("Однородность выборок 3 и 4 п К.-С");
disp(kstest_3_4);

kstest_3_2 = kstest2(A(:,3),A(:,2));
disp("Однородность выборок 3 и 2 п К.-С");
disp(kstest_3_2);
```

H0: Проверяется гипотеза H_0: выборки однородны, т. е. извлечены из одной и той же генеральной совокупности.

Результат: выборки 2,3,4 однородны

Ранговые критерии однородности основаны на использовании номеров наблюдений в вариационном ряду, полученном после упорядочивания объединенной выборки объема Номер, который получает наблюдение в упорядоченной выборке, называется его рангом и обозначается дальше метками.

Предлагаемые ниже критерии состоятельны при проверке гипотезы неоднородности, когда неоднородность порождается различием в параметре положения распределений. Для случая двух выборок альтернативные гипотезы можно записать в виде:

$$\begin{cases} H_{11}: F_1(x) = F_2(x-\mu),\ \mu \neq 0 & \text{— распределения сдвинуты относительно друг друга;} \\ H_{12}: F_1(x) = F_2(x-\mu),\ \mu > 0 & \text{— второе распределение сдвинуто влево по отношению к первому;} \\ H_{13}: F_1(x) = F_2(x-\mu),\ \mu < 0 & \text{— второе распределение сдвинуто вправо по отношению к первому,} \end{cases} \quad (11.25)$$

```
%Критерий Мана-Уитни или Уилкоксона
[p,h] = ranksum(A(:,3),A(:,4));
disp("Однородность выборок 3 и 4 по Мана-Уитни или
Уилкоксона");
disp(p);
disp(h);

[p,h] = ranksum(A(:,3),A(:,2));
disp("Однородность выборок 3 и 2 по Мана-Уитни или
Уилкоксона");
disp(p);
disp(h);
```

H0: Между выборками существуют лишь случайные различия по уровню исследуемого признака.

H1: Между выборками существуют неслучайные различия по уровню исследуемого признака.

p > 0 вторая выборка сдвинута влево по отношению к первой

p < 0 вторая выборка сдвинута вправо по отношению к первой

p = 0 сдвинуты по отношению друг к другу

№ ст	№ ст	μ1= μ2	s21= s22	Критерий К.-С. Выбрана H0?	Ранговый кр. Выбрана H0?	
3	4	✓	✗	1	h =0	p=0.4065
3	2	✗	✓	1	h =1	p=1.1825e-05

Результат: Выборки 3 и 4 имеют случайно схожие различия, а выборки 3 и 2 взяты из одной генеральной совокупности со сдвигом 1.1825e-05.