Лекция-4

Критерий Фишера

Рассмотрим критерий Фишера (*F*- распределение)

$$F(f_1,f_2)=\frac{\chi_1^2}{f_1}\Big/\frac{\chi_2^2}{f_2} \qquad (1)$$

Здесь χ_1^2,χ_2^2 - это случайные величины суммы квадратов дисперсий, а f_1,f_2 и соответствующие им степени свободы.

Распределения Фишера имеет вид

$$P(f_1,f_2,x)=\frac{\Gamma\left(\frac{f_1+f_2}{2}\right)}{\Gamma\left(\frac{f_1}{2}\right)\Gamma\left(\frac{f_2}{2}\right)}f_1^{\frac{f_1}{2}}f_2^{\frac{f_2}{2}}\frac{x^{\frac{f_2}{2}-1}}{\left(f_2+f_1x\right)^{\frac{f_1+f_2}{2}}} \qquad (2)$$

$$P(f1,f2,x):=\frac{\Gamma\left(\frac{f1+f2}{2}\right)}{\Gamma\left(\frac{f1}{2}\right)\cdot\Gamma\left(\frac{f2}{2}\right)}\cdot f1^{\frac{f1}{2}}\cdot f2^{\frac{f2}{2}}\cdot\frac{x^{\frac{f1}{2}-1}}{(f2+f1\cdot x)^{\frac{f1+f2}{2}}}$$

При больших аргументах x>170 переходят к асимптотическому представлению гамма функции. Программа Mathcad дает значения гамма функции только для аргументов меньших x<170. Дальше хорошо работает формула Стирлинга

$$\Gamma(x+1)=\sqrt{2\pi x}\left(\frac{x}{e}\right)^x\ \rightarrow\ \Gamma(x)=\sqrt{2\pi(x-1)}\left(\frac{(x-1)}{e}\right)^{(x-1)}$$

Приведём пример распределения Фишера

$$F\_1(x):=\sqrt{2\cdot\pi\cdot x}\cdot\left(\frac{x}{e}\right)^x$$

$G(x):=F\_1(x-1)$ Проверка работы асимптотики

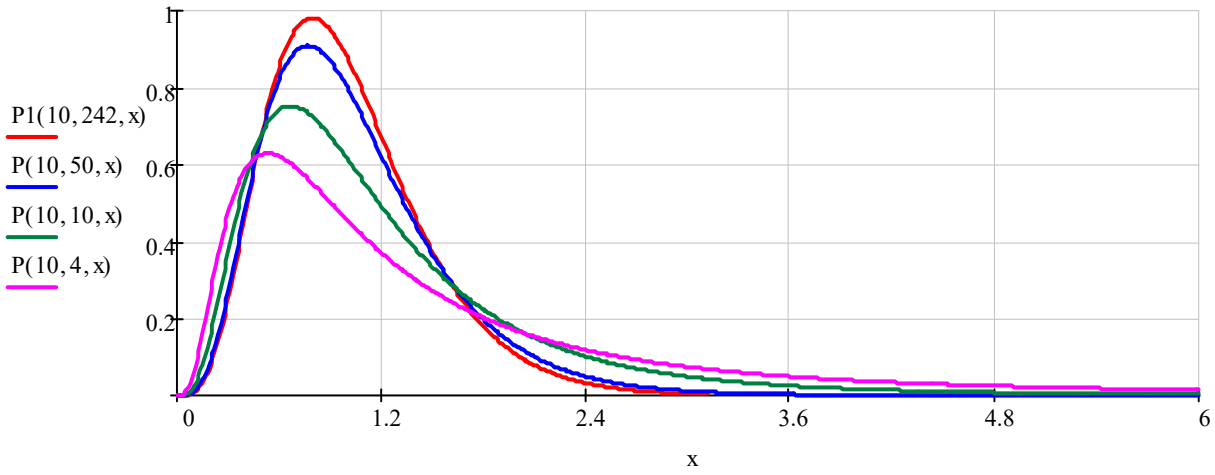
$\Gamma(170)=4.269\times10^{304}$ -**Mathcad** $G(170)=4.267\times10^{304}$ - Формула Стирлига

$\Gamma(5)\cdot5=120$ $\Gamma(6)=120$

$x:=0,0.01..6$

$$P1(f1,f2,x):=\frac{G\left(\frac{f1+f2}{2}\right)}{G\left(\frac{f1}{2}\right)\cdot G\left(\frac{f2}{2}\right)}\cdot f1^{\frac{f1}{2}}\cdot f2^{\frac{f2}{2}}\cdot\frac{x^{\frac{f1}{2}-1}}{(f2+f1\cdot x)^{\frac{f1+f2}{2}}}$$

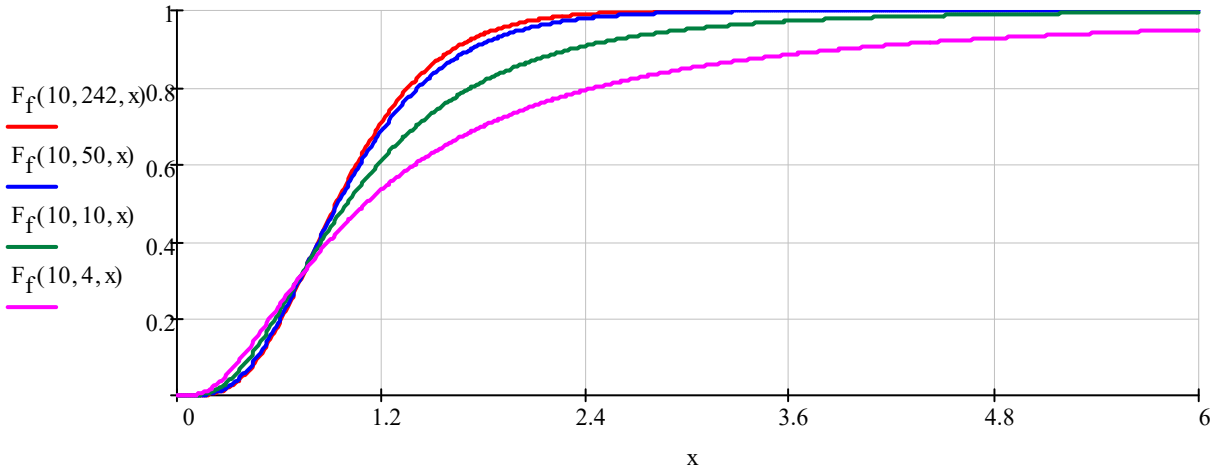
Распределение Фишера



Функция вероятности Фишера

$P2(f1, f2, x) := P(f1, f2, x)$ $P(f1, f2, x) := if(x \geq 170, P1(f1, f2, x), P2(f1, f2, x))$

$F_f(f1, f2, x) := \int_0^x P(f1, f2, x)\, dx$



Параметры распределения Фишера

$f1 := 4 \quad f2 := 10$

$M(f1, f2, m, X_M) := \int_0^\infty P(f1, f2, x) \cdot (x - X_M)^m\, dx$

$M(f1, f2, 0, 0) = 1$ площадь под кривой

$X_M := M(f1, f2, 1, 0) = 1.25$ математическое ожидание

$M_f(f1, f2) := \frac{f2}{f2 - 2}$ $M_f(f1, f2) = 1.25$ аналитическое представление при условии, что f2 >2

Дисперсия распределения Фишера

$M(f1, f2, 2, X_M) = 1.563$

$D_f(f1, f2) := \frac{2 \cdot f2^2 \cdot (f1 + f2 - 2)}{f1 \cdot (f2 - 2)^2 \cdot (f2 - 4)}$ $D_f(f1, f2) = 1.563$ аналитическое представление при условии, что f2 >4

Значения критерия Фишера (*F*- критерия)

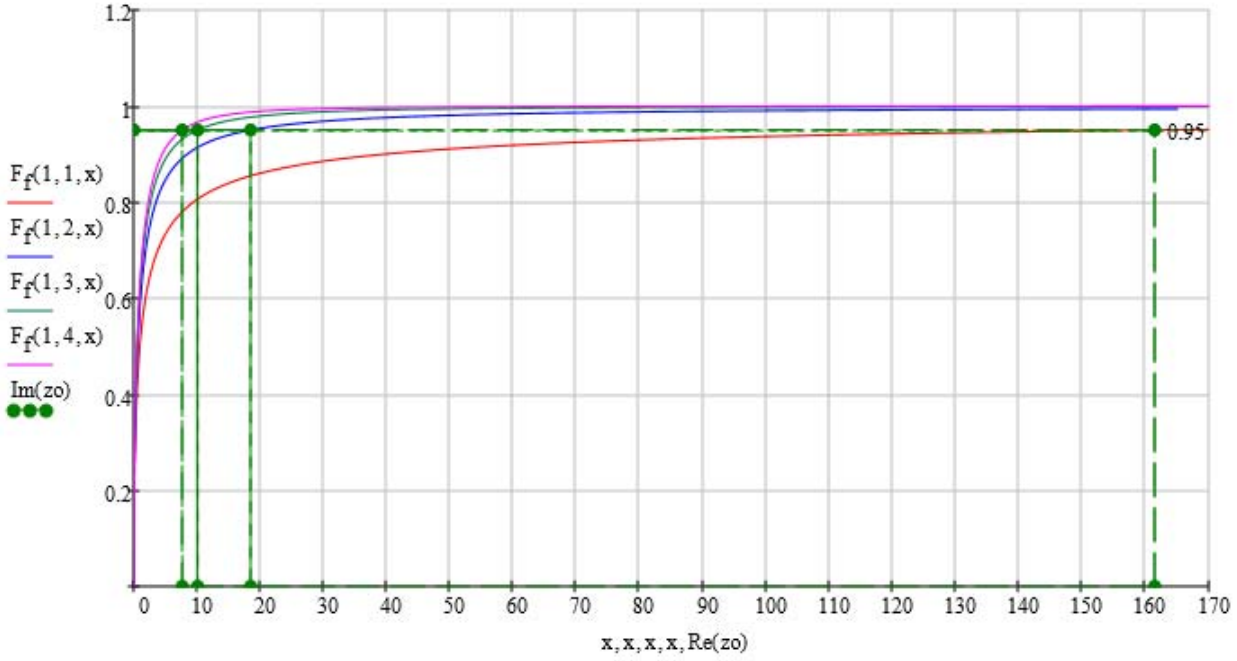
Значения приведены для уровня значимости $\alpha = 0{,}05$.

$d.f_2$	$d.f_1$									
	1	2	3	4	5	6	8	12	24	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	238,9	243,9	249,0	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,557	9,277	9,117	9,013	8,941	8,845	8,744	8,638	8,526
4	7,709	6,944	6,591	6,388	6,256	6,163	6,041	5,912	5,774	5,628
5	6,608	5,786	5,409	5,192	5,050	4,950	4,818	4,678	4,527	4,365
6	5,987	5,143	4,757	4,534	4,387	4,284	4,147	4,000	3,841	3,669
7	5,591	4,737	4,347	4,120	3,972	3,866	3,726	3,575	3,410	3,230
8	5,318	4,459	4,066	3,838	3,688	3,581	3,438	3,284	3,115	2,928
9	5,117	4,256	3,863	3,633	3,482	3,374	3,230	3,073	2,900	2,707
10	4,965	4,103	3,708	3,478	3,326	3,217	3,072	2,913	2,737	2,538
11	4,844	3,982	3,587	3,357	3,204	3,095	2,948	2,788	2,609	2,404
12	4,747	3,885	3,490	3,259	3,106	2,996	2,849	2,687	2,505	2,296
13	4,667	3,806	3,411	3,197	3,025	2,915	2,767	2,604	2,420	2,206
14	4,600	3,739	3,344	3,112	2,958	2,848	2,699	2,534	2,349	2,131
15	4,543	3,682	3,287	3,056	2,901	2,790	2,641	2,475	2,288	2,066
16	4,494	3,634	3,239	3,007	2,852	2,741	22,591	2,425	2,235	2,010
17	4,451	3,592	3,197	2,965	2,810	2,699	2,548	2,381	2,190	1,960
18	4,414	3,555	3,160	2,928	2,773	2,661	2,510	2,342	2,150	1,917
19	4,381	3,522	3,127	2,895	2,740	2,628	2,477	2,308	2,114	1,878
20	4,351	3,493	3,098	2,866	2,711	2,599	2,447	2,278	2,082	1,843
21	4,325	3,467	3,077	2,840	2,685	2,573	2,420	2,250	2,054	1,811
22	4,301	3,443	3,049	2,817	2,661	2,549	2,397	2,226	2,028	1,783
23	4,279	3,422	3,028	2,796	2,640	2,528	2,375	2,204	2,005	1,757
24	4,260	3,403	3,009	2,777	2,621	2,508	2,355	2,183	1,984	1,733
25	4,242	3,385	2,991	2,759	2,603	2,490	2,337	2,165	1,964	1,711
26	4,225	3,369	2,975	2,743	2,587	2,474	2,321	2,148	1,946	1,691
27	4,210	3,354	2,960	2,728	2,572	2,459	2,305	2,132	1,930	1,672
28	4,196	3,340	2,947	2,714	2,558	2,445	2,291	2,118	1,915	1,654
29	4,183	3,328	2,934	2,701	2,545	2,432	2,278	2,104	1,901	1,638
30	4,171	3,316	2,922	2,690	2,534	2,421	2,266	2,092	1,887	1,622
40	4,085	3,232	2,839	2,606	2,449	2,336	2,180	2,003	1,793	1,509
60	4,001	3,150	2,758	2,525	2,368	2,254	2,097	1,917	1,700	1,389
120	3,920	3,072	2,680	2,447	2,290	2,175	2,016	1,834	1,608	1,254
∞	3,841	2,996	2,605	2,372	2,214	2,098	1,938	1,752	1,517	1,000

Формирование таблицы Фишера при вероятности 0.95, то есть при уровне значимости α=0.05 , при фиксированном значении f_1=1, и изменяющемся значении f_2. Это означает, что нужно решить нелинейное уравнение вида

$$F_f(f_1, f_2, x) = \int_0^x P(f_1, f_2, x)dx = 1-\alpha \quad \rightarrow \quad F_f(f_1, 1, x) = \int_0^x P(f_1, 1, x)dx = 1-\alpha$$

Находится квантиль - аргумент функции вероятности такой x, что при этом значении аргумента вероятность равна 1-α=0.95



ORIGIN := 1

$$X(fl,\alpha,m) := \begin{| for\ i \in 1..m-1 \\ \quad | xo \leftarrow m - i \\ \quad | x_i \leftarrow i \\ \quad | X_i \leftarrow root\left[F_f(fl,i,xo) - (1-\alpha), xo\right] \\ z \leftarrow x_l \\ Z \leftarrow X_l \\ for\ i \in 2..m-1 \\ \quad | z \leftarrow stack\left(z, x_i\right) \\ \quad | Z \leftarrow stack\left(Z, X_i\right) \\ augment(x, X)}$$

$$X(1, 0.05, 15) = \begin{pmatrix} 1 & 161.448 \\ 2 & 18.513 \\ 3 & 10.128 \\ 4 & 7.709 \\ 5 & 6.608 \\ 6 & 5.987 \\ 7 & 5.591 \\ 8 & 5.318 \\ 9 & 5.117 \\ 10 & 4.965 \\ 11 & 4.844 \\ 12 & 4.747 \\ 13 & 4.667 \\ 14 & 4.6 \end{pmatrix}$$

$d.f_2$	1
1	161,4
2	18,51
3	10,13
4	7,709
5	6,608
6	5,987
7	5,591
8	5,318
9	5,117
10	4,965
11	4,844
12	4,747
13	4,667
14	4,600

Дисперсионный анализ

Дисперсионный анализ используется тогда когда нужно определить влияние внешнего фактора на наблюдение. Наблюдения принято называть откликом. Фактор влияния может быть температурным, например, когда различные часть эксперимента проводится при разных температурах. Или фактор может быть сезонным, если часть эксперимента проводится в различные сезоны. Факторами влияния могут быть среды, когда часть эксперимента проводится в разных средах (вода, масло, воздух). Что бы увидеть влияния фактора на эксперимент нужно для частей эксперимента, которые проводятся при разных значениях фактора определить статистические параметры - мат. ожидание и дисперсию. Если влияния фактора нет, то статистические параметры не будут зависеть от фактора, то есть дисперсия и мат ожидание для разных экспериментов будут одинаковы. Это означает, что разные части эксперимента относятся к одной генеральной совокупности. Если же различие есть, то в статистических параметрах будет разброс. Степень достоверности этого разброса определятся критерием Фишера, который записывается следующим образом.

$$F = \delta^2 / \sigma^2 = S_\delta^2 / S_\sigma^2 \qquad (3)$$

Здесь δ^2 – межгрупповая дисперсия, дисперсия между группами, σ^2 - внутригрупповая дисперсия, S_σ^2, S_δ^2 - выборочные значения внутригрупповой и межгрупповой дисперсий соответственно.

Приведем пояснения, как вычисляются δ^2 и σ^2.

Предположим, что имеются *m*-групп (выборок). В каждой группе *n*–измерений (количество измерений в одной группе может быть не одинаковым). Общее количество наблюдений $N = n + n + ... = m \cdot n$.

Для определения δ^2 межгрупповой дисперсии, сначала вычисляется среднее значение (мат. ожидание) для всех измерений, то есть все группы воспринимаются как единая генеральная совокупность

$$x_M = \frac{1}{N}\sum_{k=0}^{N-1} x_k \,. \tag{4}$$

Затем вычисляется среднее для каждой группы

$$x_{1M} = \frac{1}{n}\sum_{k=0}^{n-1} x1_k,\ \ x_{2M} = \frac{1}{n}\sum_{k=0}^{n-1} x2_k,\ \ \ldots\ \ \ x_{mM} = \frac{1}{n}\sum_{k=0}^{n-1} xm_k. \tag{5}$$

Теперь можно вычислять межгрупповую дисперсию

$$\begin{aligned} S_\delta^2 &= \frac{1}{m-1}\left(\left(x_{1M}-x_M\right)^2 n + \left(x_{2M}-x_M\right)^2 n + \ldots. + \left(x_{mM}-x_M\right)^2 n\right) = \\ &= \frac{1}{m-1}\sum_{i=1}^{m}\left(x_{iM}-x_M\right)^2 n_i \end{aligned} \tag{6}$$

Здесь m-1, это число степеней свободы.

Для определения внутри группой дисперсии σ^2 нужно использовать среднее в каждой группе (5) и дисперсию в каждой группе. Запишем оценку внутри группой дисперсии (её называют средней дисперсией)

$$\begin{aligned} S_\sigma^2 &= \frac{1}{N-m}\left(\sum_{k=0}^{n-1}\left(x1_k - x_{1M}\right)^2 + \sum_{k=0}^{n-1}\left(x2_k - x_{2M}\right)^2 + \ldots. + \sum_{k=0}^{n-1}\left(xm_k - x_{mM}\right)^2\right) = \\ &= \frac{1}{N-m}\sum_{i=1}^{m}\sum_{k=0}^{n-1}\left(xi_{kM} - x_{iM}\right)^2 \end{aligned}$$

Если число измерений в группах одинаково, можно поступить так, сначала определяю дисперсию в каждой группе

$$S_{\sigma 1}^2 = \frac{1}{n}\sum_{k=0}^{n-1}\left(x1_k - x_{1M}\right)^2,\ \ S_{\sigma 2}^2 = \frac{1}{n}\sum_{k=0}^{n-1}\left(x2_k - x_{2M}\right)^2 \ldots.\ \ S_{\sigma m}^2 = \frac{1}{n}\sum_{k=0}^{n-1}\left(xm_k - x_{mM}\right)^2$$

А затем определяют среднее значение дисперсии, усредненное по количеству групп

$$S_{\sigma 1}^2 = \frac{1}{n}\sum_{k=0}^{n-1}\left(x1_k - x_{1M}\right)^2,\ \ S_{\sigma 2}^2 = \frac{1}{n}\sum_{k=0}^{n-1}\left(x2_k - x_{2M}\right)^2 \ldots.\ \ S_{\sigma m}^2 = \frac{1}{n}\sum_{k=0}^{n-1}\left(xm_k - x_{mM}\right)^2$$

$$S_\sigma^2 = \frac{1}{m}\sum_{i=1}^{m} S_{\sigma i}^2$$

Нужно определить дисперсию

Пример 1. Исследуется эффективность обучения тремя различными методами. Студентам дается задание изучить тему «Робототехника». Для этого 10 студентов конспектируют

первоисточник, 10 изучают ее по учебникам, 10 – с помощью обучающих компьютерных программ. По окончании их уровень знаний проверяется с помощью теста, состоящего из 100 вопросов. Результаты представлены в таблице:

Балы по тесту по трем методикам			
	первоисточник	учебник	компьютер
1	28	39	41
2	33	52	49
3	42	53	56
4	47	54	62
5	48	56	63
6	50	58	64
7	50	59	65
8	51	63	72
9	60	64	77
10	71	77	87

Влияет ли ***методика изучения*** темы на результат? Есть ли значимые различия между тремя выборками по уровню усвоения материала?

Решение. В наблюдении участвуют 30 студентов, которые разделены на 3 группы. Факторным признаком является методика изучения. Вариацию, обусловленную влиянием фактора, положенного в основу группировки, характеризует **межгрупповая дисперсия.**

Будем использовать Mathcad

$x1 := (28\ \ 33\ \ 42\ \ 47\ \ 48\ \ 50\ \ 50\ \ 51\ \ 60\ \ 71)^T$

$x2 := (39\ \ 52\ \ 53\ \ 54\ \ 56\ \ 58\ \ 59\ \ 63\ \ 64\ \ 77)^T$

$x3 := (41\ \ 49\ \ 56\ \ 62\ \ 63\ \ 64\ \ 65\ \ 72\ \ 77\ \ 87)^T$

$n := \text{length}(x1) = 10$ число испытаний на одном уровне

$m := 3$ число уровней фактора.

1. ***Определяем мат. ожидание для каждой группы***

$$M(x,n) := \frac{1}{n}\cdot\sum_{k=0}^{n-1} x_k \qquad x1_M := M(x1,n) = 48 \qquad x2_M := M(x2,n) = 57.5 \qquad x3_M := M(x3,n) = 63.6$$

2.***Определяем общее мат. ожидание***

$$X_M := \frac{x1_M\cdot n + x2_M\cdot n + x3_M\cdot n}{n+n+n} = 56.367$$

Общее среднее и частные средние отличается

3. ***Для того что бы определить является ли это различие существенным и вызвано различными методиками преподавания, определяем межгрупповую дисперсию***

$$S_{\delta 2} := \frac{(x1_M - X_M)^2\cdot n + (x2_M - X_M)^2\cdot n + (x3_M - X_M)^2\cdot n}{3-1} = 618.033$$

И внутригрупповую дисперсию

$$S_{\sigma 2} := \frac{\sum_{k=0}^{n-1} \left(x1_k - x1_M\right)^2 + \sum_{k=0}^{n-1} \left(x2_k - x2_M\right)^2 + \sum_{k=0}^{n-1} \left(x3_k - x3_M\right)^2}{30 - 3} = 140.7$$

4. ***Определяем критерий Фишера***

$F_n := \frac{S_{\delta 2}}{S_{\sigma 2}} = 4.393$ наблюдаемый критерий Фишера

Число степеней свободы k_1=3-1=2 k_2=30-3=27 уровень значимости α=0.05

Смотрим квантиль в таблице Фишера $F_k := 3.354$

$F_n > F_k$ Это значит, что влияние фактора есть

Пример 2.

Известны результаты выборочного обследования пробега автомобильных шин нового типа в различных условиях эксплуатации (см. тб.). Установить, существует ли зависимость между условиями эксплуатации и величиной пробега шин, гарантируя результат с вероятностью 0,95.

Тб. Пробег шин в различных условиях эксплуатации

Условия эксплуатации	Пробег шин, тыс. км.											*f*
Городские	70,5	71,8	69,8	58,9	68,7	72,1	70,3	69,1	72	58,7	66,2	11
Смешанные	58,9	59,1	60,1	62,2	60,5	58,4	59	61,8				8
Загородные	54,2	58,8	56,6	55	56,4							5

Факторный признак – условия эксплуатации (среда).
Результативный признак – величина пробега шин (тыс. километров после которых шина приходит в негодность).
Для каждой группы определяем средний пробег шин.

Для расчетов будем использовать программу Mathcad.

Число испытаний для одного уровня

$x1 := (70.5\ \ 71.8\ \ 69.8\ \ 58.9\ \ 68.7\ \ 72.1\ \ 70.3\ \ 69.1\ \ 72\ \ 58.7\ \ 66.2)^T$ $n1 := \text{length}(x1) = 11$

$x2 := (58.9\ \ 59.1\ \ 60.1\ \ 62.2\ \ 60.5\ \ 58.4\ \ 59\ \ 61.8)^T$ $n2 := \text{length}(x2) = 8$

$x3 := (54.2\ \ 58.8\ \ 56.6\ \ 55\ \ 56.4)^T$ $n3 := \text{length}(x3) = 5$

1. *Определяем мат. ожидание для каждой группы*

$$M(x,n) := \frac{1}{n}\cdot\sum_{k=0}^{n-1} x_k$$

$x1_M := M(x1,n1) = 68.009$ $x2_M := M(x2,n2) = 60$ $x3_M := M(x3,n3) = 56.2$

2. Определяем общее мат. ожидание как для одной генеральной совокупности

$$X_M := \frac{x1_M\cdot n1 + x2_M\cdot n2 + x3_M\cdot n3}{n1 + n2 + n3} = 62.879$$

Заметим, что общее среднее и частные средние отличаются

3. Для того что бы определить является ли это различие существенным и вызвано условиями эксплуатации, определим межгрупповую дисперсию

$$S_{\delta 2} := \frac{(x1_M - X_M)^2\cdot n1 + (x2_M - X_M)^2\cdot n2 + (x3_M - X_M)^2\cdot n3}{m - 1} = 289.425$$

4. Определяем внутригрупповую дисперсию или среднюю дисперсию

$N := n1 + n2 + n3 = 24$ общее количество измерений

$$S_{\sigma 2} := \frac{\sum_{k=0}^{n1-1}(x1_k - x1_M)^2 + \sum_{k=0}^{n2-1}(x2_k - x2_M)^2 + \sum_{k=0}^{n3-1}(x3_k - x3_M)^2}{N - m} = 12.504$$

5. Теперь можно определить критерий Фишера

$$F_n := \frac{S_{\delta 2}}{S_{\sigma 2}} = 23.146$$

Смотрим, какие степени свободы имеем.

Для межгруппового случая имеем m-1=3-1=2

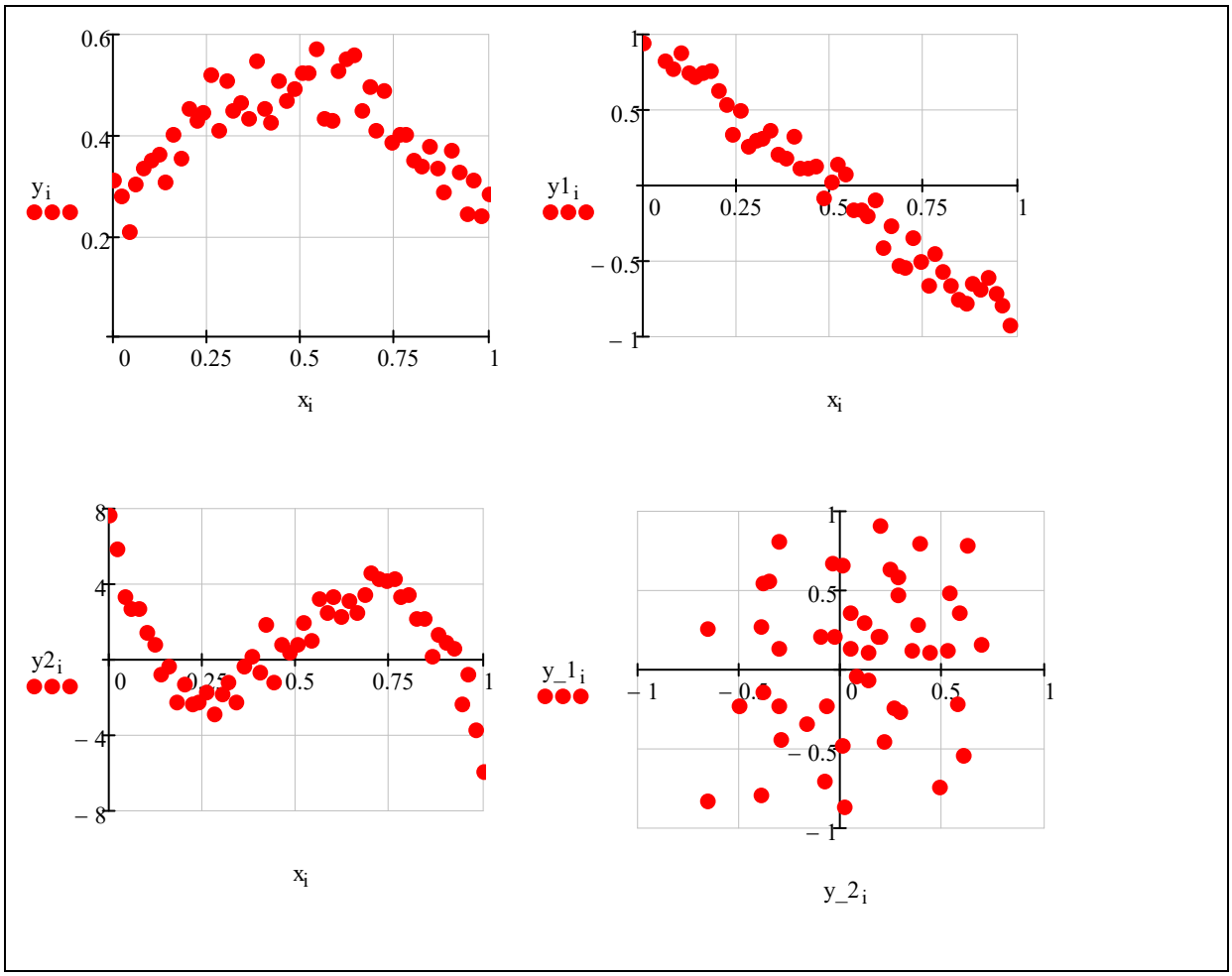
Для внутригруппового случая имеем N-m=24-3=21

По таблице Фишера ищем квантиль соответствующий уровню значимости α=0.05 и степеням свободы f_1=2 , f_2=21

21	4,325	3,467	3,077

$F_k := 3.467$ $F_n > F_k$ Это значить влияние фактора есть

Понятие о статистической и корреляционной связи



На графиках представлены распределения 3х случайных величины. Видно, что на всех графиках есть некий тренд. На четвертом рисунке трудно, что-либо сказать о тренде. То есть мы видим связь между горизонтальной-*x* и вертикальной - *y* величинами. Под случайной связью понимают такую связь, при которой изменения одной величины никак не влияют на изменения другой величины, то есть это фактически отсутствие связи. Корреляционная связь эта такая связь, при которой изменения средних значений одной величины вызывают изменения другой величины. Функциональная связь

(детерминированная) когда есть четка зависимость одной величины от другой. Например – аналитические зависимости одной величины от другой (они могут быть и табличными).