Дисперсионный анализ. Корреляционный анализ

Сперва напомним что такое статистический критерий - это правило, обеспечивающее принятие истинной и отклонение ложной гипотезы с заданной вероятностью.

По сути это функция $f : \mathcal{D} \to \mathbb{R}$, где $\mathcal{D}$ - некоторые данные, $\mathbb{R}$ - множество действительных чисел, при этом ноль принадлежит области значений этой функции: $0 \in ran f$.

Область принятия гипотезы (ОПГ) - подмножество таких значений критерия, при которых основная гипотеза не может быть отвергнута. Область принятия гипотезы всегда включает в себя значение 0.

Критическая область - подмножество таких значений критерия, при которых основная гипотеза не может быть принята.

Пример - значение t-критерия выполняется по формуле:

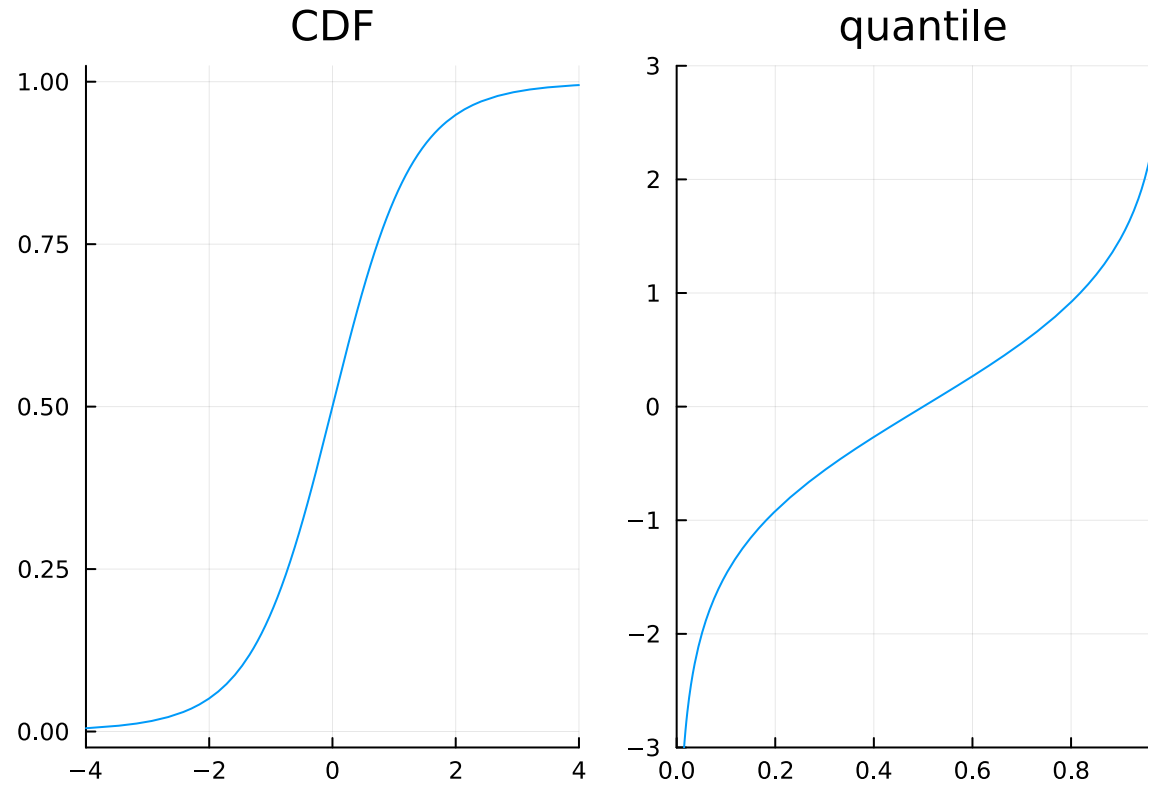
$$\left| \frac{\mu - \mu_0}{\sigma/\sqrt{N}} \right| \leq t_{(\alpha,\nu)}$$

Так-как сатистика критерия (т.е. значение функции) - случайная величина, то и область принятия гипотезы определяется как случайная величина. К примеру критические значения t-критерия находятся из распределения Стьюдента. С помощью вичислений или с помощью таблиц. Приведем примеры нахождения критических значений t-критерия Стьюдента:

Квантиль — значение, которое заданная случайная величина не превышает с фиксированной вероятностью.

In [4]:
```
using Plots, Distributions
dist = TDist(5)
plot(
    plot(x-> cdf(dist, x), legend = false, xlims = (-4,4)),
    plot(x-> quantile(dist, x), legend = false, xlims = (0,1), ylims = (-3, 3)),
    layout = 2, title=["CDF" "quantile"])
```

Out[4]:



In [1]:
```
using Distributions
α = 0.05
dist = TDist(5)

l = quantile(dist, α/2)
u = quantile(dist, 1-α/2)
println("Критические значения", l, " и ", u)
```

```
Критические значения-2.5705818356363146 и 2.5705818356363137
```

In [9]:
```
quantile(TDist(65), 1-.05/2)
```

Out[9]:
```
1.9971379083920036
```

Таблица для **двухсторонних** значений t-критерия Стьюдента

Число степеней свободы d.f.	α		
	00,10	0,05	0,01
1	6,3138	12,706	63,657
2	2,9200	4,3027	9,9248
3	2,3534	3,1825	5,8409
4	2,1318	2,7764	4,5041
5	2,0150	2,5706	4,0321
6	1,9432	2,4469	3,7074
7	1,8946	2,3646	3,4995
8	1,8595	2,3060	3,3554
9	1,8331	2,2622	3,2498
10	1,8125	2,2281	3,1693

Число степеней свободы вычисляется по определенным правилам и обычно равно: $n-1$ для одовыборочного теста, n_1+n_2-2 - для t-критерия для независимых групп.

Давайте применим это правило для тестирования следующей гипотезы:

H_0 : среднее значение по выборке равно 5 $\mu = 5$

H_A : среднее значение по выборке не равно 5 $\mu \neq 5$

Пусть даны данные: $[5, 6, 3, 9, 7, 7]$

Для этой выборки: $\mu = 6.17$ и $\sigma = 2.04$

Напомним μ - это среднее арифметическое, σ - стандартное отклонение.

Используя формулу:

$$\left| \frac{\mu - \mu_0}{\sigma/\sqrt{N}} \right| \leq t_{(\alpha,\nu)}$$

Находим критиченское значение $t = 1.4$. Делаем выводы.

В контексте дисперсионного анализа можно сказать, что t-критерий является частным случаем дисперсионного анализа.

Математическая модель модет быть выражена так:

$x_i = \mu + e_i$

где: x_i - результат измерения i, μ - истинный результат, e - случайная ошибка i-го измерения.

In [5]:
```
d = [5,6,3,9,7,7]
μ = mean(d)
μ₀ = 5
n = length(d)
σ = std(d)
println("μ: ", μ, " σ: ", σ)

t = abs((μ - μ₀)/(σ/sqrt(n)))
println("Значение критерия: ", t)
println("Стандартная ошибка среднео: ", σ/sqrt(n))
```

```
μ: 6.166666666666667 σ: 2.0412414523193148
Значение критерия: 1.4000000000000006
Стандартная ошибка среднео: 0.8333333333333333
```

In [6]:
```
pval = 1 - cdf(dist, t) + cdf(dist, -t)
```

Out[6]:
```
0.22040387992934418
```

In [4]:
```
using HypothesisTests

OneSampleTTest(d, μ₀)
```

Out[4]:
```
One sample t-test
-----------------
Population details:
    parameter of interest:   Mean
    value under h_0:         5
    point estimate:          6.16667
    95% confidence interval: (4.025, 8.309)

Test summary:
    outcome with 95% confidence: fail to reject h_0
    two-sided p-value:           0.2204

Details:
    number of observations:   6
    t-statistic:              1.4000000000000006
    degrees of freedom:       5
    empirical standard error: 0.8333333333333333
```

Дисперсионный анализ

Сущность дисперсионного анализа

Дисперсионный анализ – это статистический метод анализа результатов наблюдений, зависящих от различных одновременно действующих факторов. Может применяться для выбора наиболее важных факторов и оценки их влияния.

Идея дисперсионного анализа заключается в разложении общей дисперсии случайной величины на независимые случайные слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия. Последующее сравнение этих дисперсий позволяет оценить существенность

влияния факторов на исследуемую величину.

Если исследуется влияние одного фактора на исследуемую величину, то речь идет об однофакторном комплексе. Если изучается влияние двух факторов – двухфакторный комплекс. И т.д.

Исходными положениями дисперсионного анализа являются:

- Нормальное распределение значений изучаемого признака в генеральной совокупности;
- Количественный непрерывный тип данных, дискретные данные менее желательны;
- Равенство дисперсий в сравниваемых генеральных совокупностях;
- Случайный и независимый характер выборки.

Гипотеза

Нулевой гипотезой в дисперсионном анализе является утверждение о равенстве средних значений (несмотря на то, что анализ называется "дисперсионным"):

- H_0: $\mu_1 = \mu_2 = \cdots = \mu_j$
- H_A: $\exists\, i, j \in \{1, \ldots, j\}, i \neq j : \mu_i \neq \mu_j$

При отклонении нулевой гипотезы принимается альтернативная гипотеза о том, что не все средние равны, то есть имеются, по крайней мере, две группы, отличающиеся средними значениями.

Однофакторный дисперсионный анализ

Возможен случай, когда в эксперименте одновременно изучаются несколько уровней фактора. В качестве примера можно представит: применение разных лекарственных средств для лечения заболевания, использование разных кормов для животных, применение разных инструментов для обработки деталей.

Важно что бы воздействие фактором было одиночным и не зависело от других факторов. Иногда можно представит ситуацию когда практически на результат влияют несколько факторов, к примеру "препарат" и "режим", но в связи с тем, что комбинации - непересекающиеся, то можно считать это одним фактором. Интерпретация в таком случае должа учитывать, что такой комплексный фактор влияет безразрывно и не раскладывается на составляющие.

Таким образом, мы изучаем - влияет ли выделенный (многоуровневый) фактор на результат. Ответ на этот вопрос можно получить, сравнивая средние значения измерений полученных под влиянием каждого из факторов между собой с последующей оценкой существенности разницы этих средних.

Уровни фактора

Любой фактор исследуется в котексте уровней этого фактора, т.е. - отдельных взаимоисключающих событий. Т.е. фактор называется "Пол", уровни фактора: "мужской", "женский".

Уровней фактора может быть практически любое количество.

К примеру: фактор - "Цвет", уровни: "красный", "фиолетовый", "синий", "зеленый", "черный".

Уровни фактора должны быть уровни из одного множества взаимоисключающих характеристик - т.е. вызывает сомнение осмысленность уровней одного фактора: "синий", "белый", "твердый".

Математическая модель дисперсионного анализа представляет собой частный случай основной линейной модели.

$x_{i,j} = \mu_i + a_{i,j} + e_{i,j}$,

где:

- $x_{i,j}$ — результат i-го измерения в группе A_j (A_j - уровень j фактора A);
- μ_i — точное значение i-го измерения;
- $a_{i,j}$ — систематическая ошибка i-го измерения в группе A_j (т.е. влияние уровня A_j);
- $e_{i,j}$ — случайная ошибка i-го измерения в группе A_j.

Процедура дисперсионного анализа состоит в определении соотношения систематической **(межгрупповой)** дисперсии к случайной **(внутригрупповой)** дисперсии в измеряемых данных.

В качестве показателя изменчивости используется сумма квадратов отклонения значений параметра от среднего: SS.

Общая сумма квадратов SS_{total} раскладывается на межгрупповую сумму квадратов SS_{B} и внутригрупповую сумму квадратов SS_{W}:

$SS_{\text{total}} = SS_{\text{B}} + SS_{\text{W}}$

Пусть E среднее значение генеральной совокупности, а E_j среднее значение уровня j, тогда:

- $SS_{\text{total}} = \sum_{i=1}^{n_j} \sum_{j=1}^{J} (x_{i,j} - E)^2$
- $SS_{\text{B}} = \sum_{j=1}^{J} n_j (E_j - E)^2$
- $SS_{\text{W}} = \sum_{i=1}^{n_j} \sum_{j=1}^{J} (x_{i,j} - E_j)^2$

Степени свободы

Степени свободы вычисляются следующим образом:

- $df_{\text{total}} = df_{\text{B}} + df_{\text{W}}$
- $df_{\text{total}} = N - 1$
- $df_{\text{B}} = J - 1$
- $df_{\text{W}} = N - J,$

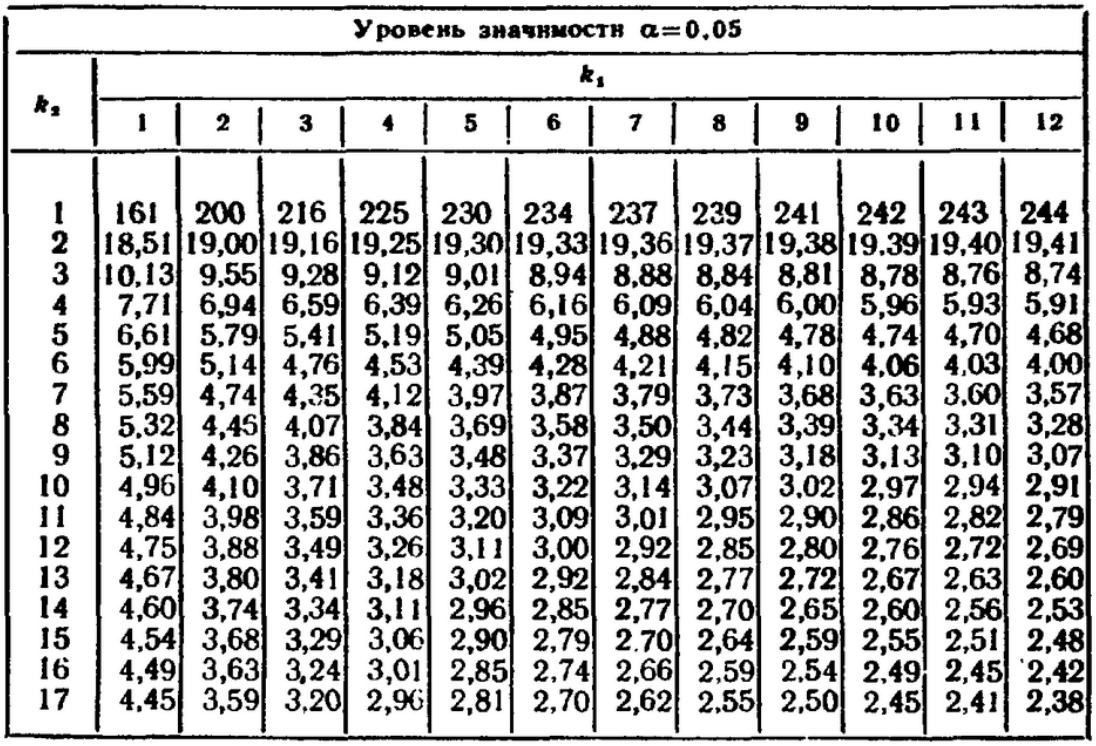
Где и N - объём полной выборки, а J — количество уровней.

Дисперсия каждой части называется - «средний квадрат»: MS - т.е. отношение суммы квадратов к числу их степеней свободы:

- $MS_{\text{total}} = \frac{SS_{\text{total}}}{N-1}$
- $MS_{\text{B}} = \frac{SS_{\text{B}}}{J-1}$
- $MS_{\text{W}} = \frac{SS_{\text{W}}}{N-J}$

Соотношение межгрупповой и внутригрупповой дисперсий имеет F-распределение (распределение Фишера) и определяется при помощи F-критерия Фишера:

$F_{df_{\text{B}}, df_{\text{W}}} = \frac{MS_{\text{B}}}{MS_{\text{W}}}$



Уровень значимости $\alpha = 0.05$

k_2	k_1 = 1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38

In [10]:
```
using Distributions
α = 0.05
dist = FDist(2, 5)
c = quantile(dist, 1-α)
println("Критическое значение ", c)
```

```
Критическое значение 5.786135043349965
```

In [12]:
```
v1 = [3,5,7,6,5,4]
v2 = [6,5,7,8,9,13]
v3 = [6,7,1,7,10,18]
#v1 = [1 3 2 1 0 2 1]'
#v2 = [2 3 2 1 4]'
#v3 = [4 5 3]'
v = vcat(v1, v2, v3)

Mean_Total = mean(v)
Mean_v1 = mean(v1)
Mean_v2 = mean(v2)
Mean_v3 = mean(v3)

SS_total = sum(x->x*x, v .- Mean_Total)

SS_B     = length(v1)*(Mean_v1 - Mean_Total)^2  + length(v2)*(Mean_v2 - Mean_Tot

SS_W     = sum((v1 .- Mean_v1) .^2) + sum((v2 .- Mean_v2) .^2) + sum((v3 .- Mean
println("Mean_Total: ", Mean_Total, ", Mean_v1: ",Mean_v1,", Mean_v2: ",Mean_v2,
println("SS_total: ", SS_total)
println("SS_B: ", SS_B)
println("SS_W: ", SS_W)
println("SS_B + SS_W = ", SS_B + SS_W)
```

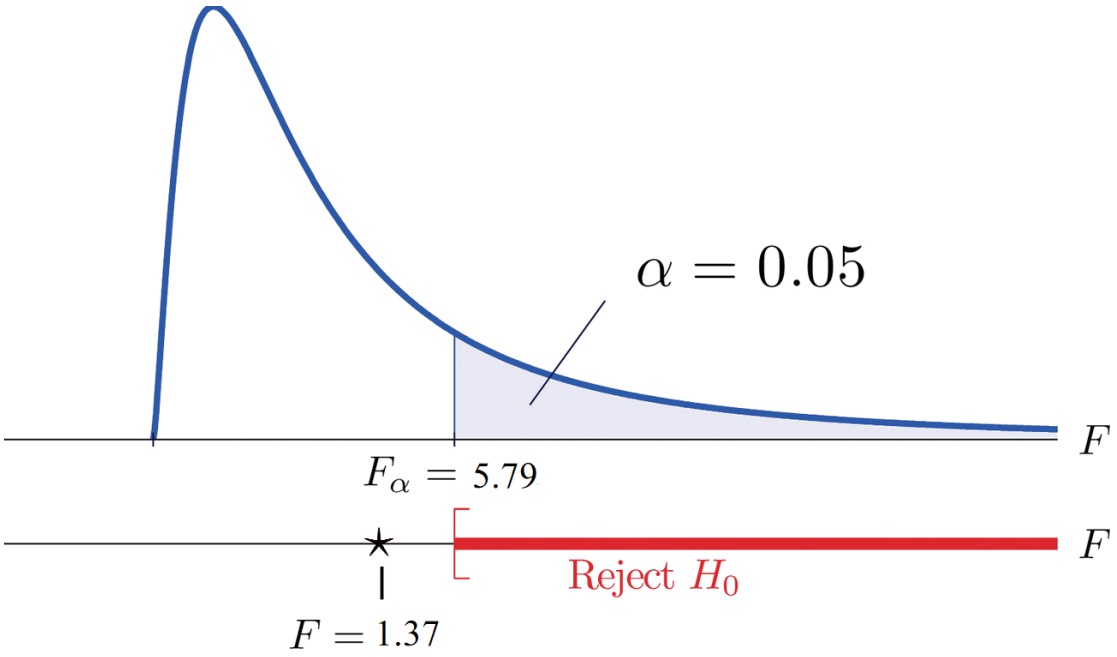
```
Mean_Total: 7.055555555555555, Mean_v1: 5.0, Mean_v2: 8.0, Mean_v3: 8.16666666666
6666
SS_total: 246.94444444444446
SS_B: 38.1111111111111
SS_W: 208.83333333333331
SS_B + SS_W = 246.9444444444444
```

In [13]:
```
df_total = length(v) - 1
df_B     = 3 - 1
df_W     = length(v) - 3
println("df: ", df_B, " ",df_W)
```

```
df: 2 15
```

In [14]:
```
MS_total = SS_total / df_total
MS_B     = SS_B / df_B
MS_W     = SS_W / df_W

F_val =  MS_B / MS_W

dist = FDist(2, 15)

pval  = 1 - cdf(dist, F_val)

println("F-статистика: ", MS_B/MS_W)
println("p-value: ", pval)
c = quantile(dist, 1-α)
println("Критическое значение ", c)
```

```
F-статистика: 1.3687150837988824
p-value: 0.2844488635980029
Критическое значение 3.682320343673239
```



In [9]:
```
using HypothesisTests
OneWayANOVATest(v1, v2, v3)
```

Out[9]:
```
One-way analysis of variance (ANOVA) test
----------------------------------------
Population details:
    parameter of interest:   Means
    value under h_0:         "all equal"
    point estimate:          NaN

Test summary:
    outcome with 95% confidence: fail to reject h_0
    p-value:                     0.2844

Details:
    number of observations: [6, 6, 6]
    F statistic:            1.36872
    degrees of freedom:     (2, 15)
```

Для подтверждения положения о равенстве дисперсий обычно применяется критерий Ливена (Ливиня)(Levene's test).

Апостериорные сравнения

После того как мы оценили всю совокупность средних значений по группам общей оценкой значимости, мы можем поставить вопрос, какие группы отличаются уверенно, а какие нет. Для этого служат апостериорные сравнения: проводятся сравнения по Т-критерию каждой группы с каждой, но показываемая Т-критерием значимость умножается на зависящую от количества сравнений константу (коррекция ошибки первого рода). В результате скорректированные значимости могут интерпретироваться, как интерпретировались бы значимости при отдельных сравнениях, но без риска завысить результат вследствие большого числа сравнений. Существуют и другие подходы к апостериорному тестированию.

Многофакторный анализ

Многофакторный анализ позволяет проверить влияние нескольких факторов на зависимую переменную. Модель многофакторного анализа имеет вид:

$x_{i,j,k} = \mu_i + a_{i,j} + b_{i,k} + \ldots + (ab)_{i,j,k} + e_{i,j,k}$, где:

- $x_{i,j,k}$ — результат i-го измерения;
- μ_i — среднее для i-го измерения;
- $a_{i,j}$ — систематическая ошибка i-го измерения для уровня j фактора A;
- $b_{i,k}$ — систематическая ошибка i-го измерения для уровня k фактора B;
- $(ab)_{i,j,k}$ — систематическая ошибка i-го измерения для уровня j, k комбинации факторов A и B;
- $e_{i,j,k}$ — случайная ошибка i-го измерения.

В отличие от однофакторной модели, где имеется одна межгрупповая сумма квадратов, модель многофакторного анализа включает суммы квадратов для каждого фактора в отдельности и суммы квадратов всех взаимодействий между ними.

В двухфакторной модели межгрупповая сумма квадратов раскладывается на сумму квадратов фактора A, сумму квадратов фактора B и сумму квадратов взаимодействия факторов A и B (такое взаимодействие обычно обозначается: $A * B$):

$$SS_{\text{total}} = SS_{Abtw} + SS_{Bbtw} + SS_{ABbtw} + SS_{\text{W}}$$

Степени свободы раскладываются аналогичным образом:

- $df_{\text{total}} = df_{Abtw} + df_{Bbtw} + df_{ABbtw} + df_{\text{W}}$

где:

- $df_{\text{total}} = N - 1$
- $df_{A} = J - 1$
- $df_{B} = K - 1$
- $df_{AB} = (J - 1)(K - 1)$
- $df_{\text{wg}} = N - JK$

Соответственно трёхфакторная модель включает сумму квадратов фактора A, сумму квадратов фактора B, сумму квадратов фактора C и суммы квадратов взаимодействий факторов A и B, B и C, A и C, а также взаимодействия всех трёх факторов A , B , C:

$$SS_{\text{total}} = SS_{Abtw} + SS_{Bbtw} + SS_{Cbtw} + SS_{ABbtw} + SS_{BCbtw} + SS_{ACbtw} + SS_{ABCbtw} + S$$

При анализе многих факторов рекомендуется использовать общую линейную модель.

Корреляционный анализ

Парная корреляция

Связи между различными явлениями в природе сложны и многообразны. Однако их можно определенным способом классифицировать. В технике и естествознании часто речь идет о функциональной зависимости между переменными X и Y, когда каждому значению X поставлено в однозначное соответствие определенного значения Y.

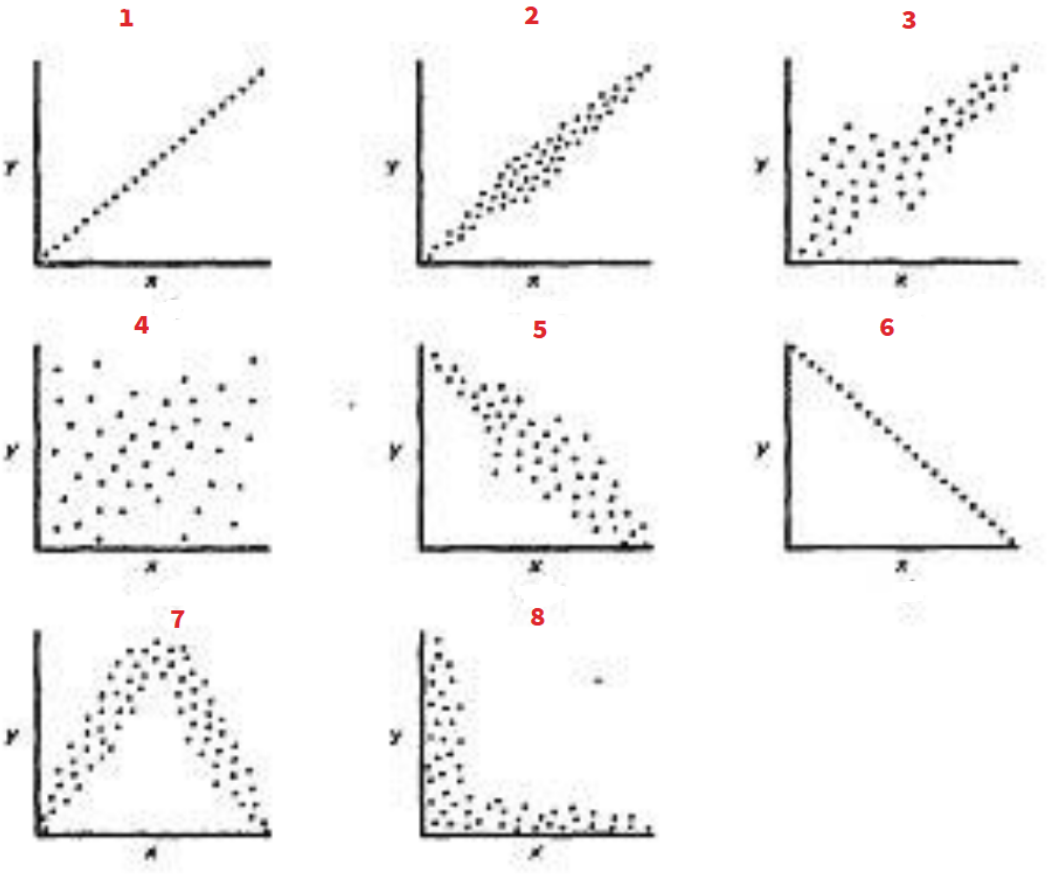
В реальном мире многие явления природы происходят в обстановке действия многочисленных факторов, влияние каждого из которых ничтожно, а число этих факторов велико. В этих случаях связь теряет свою строгую функциональность, и изучаемая физическая система переходит не в определенное состояние, а в одно из возможных.

Здесь речь идет о стохастической связи. Частный случай стохастической связи –

статистическая связь. Об этой связи имеет смысл говорить, когда условное математическое ожидание одной случайной переменной является функцией значения, принимаемого другой случайной переменной. Значения статистической зависимости между случайными переменными имеет большое практическое значение. С ее помощью можно прогнозировать зависимость случайной переменной, в предположении, что независимая принимает определенное значение.

Классификация

- Форма
  - Линейная
  - Нелинейная
- Направление
  - Прямая
  - Обратная
- Сила
  - Сильная
  - Слабая



1. Линейная строгая прямая
2. Линейная прямая

3. Линейная мягкая прямая
4. Отсутствует
5. Линейная обратная
6. Линейная обратная строгая
7. Нелинейная
8. Нелинейная

Для любых коррелированных событий А и Б их отношения включают (причинно-следственная связь):

1. A → B
2. B → A
3. C → A ∪ C → B
4. A ↔ B

Коэффициенты корреляции

1. Для порядковых данных используются следующие коэффициенты корреляции:

- ρ - коэффициент ранговой корреляции Спирмена
- τ - коэффициент ранговой корреляции Кендалла
- γ - коэффициент ранговой корреляции Гудмена – Краскела

2. Для переменных с интервальной и номинальной шкалой используется коэффициент корреляции Пирсона (корреляция моментов произведений).

3. Если, по меньшей мере, одна из двух переменных имеет порядковую шкалу, либо не является нормально распределённой, используется ранговая корреляция Спирмана или τ-Кендалла. Применение коэффициента Кендалла предпочтительно, если в исходных данных имеются выбросы.

Коэффициент корреляции Пирсона

Коэффициент корреляции

Важной характеристикой совместного распределения двух случайных величин является ковариация (или корреляционный момент). Ковариация определяется как математическое ожидание произведения отклонений случайных величин:

$$\operatorname{cov}_{XY} = \mathbf{E}\left[(X - \mathbf{E}(X))(Y - \mathbf{E}(Y))\right] = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$$

где $\mathbf{E}$ — математическое ожидание.

Линейный коэффициент корреляции

Для устранения недостатка ковариации был введён линейный коэффициент корреляции (или коэффициент корреляции Пирсона), который разработали Карл Пирсон, Фрэнсис Эджуорт и Рафаэль Уэлдон в 90-х годах XIX века. Коэффициент корреляции рассчитывается по формуле:

$$\mathbf{r}_{XY} = \frac{\mathbf{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

где $\overline{X} = \frac{1}{n} \sum_{t=1}^{n} X_t$, $\overline{Y} = \frac{1}{n} \sum_{t=1}^{n} Y_t$ — среднее значение выборок.

Коэффициент корреляции изменяется в пределах от минус единицы до плюс единицы

In [10]:
```
v1 = [1,4,3,4,5,6,5,6,7,8,7,8,9]
v2 = [3,4,2,3,4,5,6,7,6,9,8,9,10]

r_xy = cov(v1,v2)/std(v1)/std(v2)

using Statistics
r_xy_2 = Statistics.cor(v1, v2)

println("r метод 1: ", r_xy, ", r используя Statistics: ", r_xy_2)
```

```
r метод 1: 0.9093141848056545, r используя Statistics: 0.9093141848056547
```

Домашнее задание:

1. Вычислить среднее значение (μ) для вектора $[3, 4, 6, 4, 3]$
2. Вычислить выборочное стандартное отклонение (σ) для вектора $[3, 7, 3, 1]$
3. Проверить гипотезу

$H_0 : \mu = \mu_0$

$H_A : \mu \neq \mu_0$

При $\mu_0 = 6$ и двухстороннем уровне значимости $\alpha = 0.1$

Данные: $[10, 5, 4, 11, 8, 9]$

4. Используя дисперсионный анализ протестировать гиптезу о равенстве средних:

$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu$

$H_A : \mu \neq \mu_1 \cup \mu \neq \mu_2 \cup \mu \neq \mu_3$

При двухстороннем уровне значимости $\alpha = 0.05$

Данные: $d_1 = [10, 8, 4, 6, 8]; d_2 = [8, 5, 3, 10, 6]; d_2 = [12, 5, 7, 8, 7]$

5. Найти корреляции Пирсона для векторов

a = [1,6,4,5,7,8,9,6,5] b = [8,6,5,6,4,1,4,3,2]

6. Дать определения для

- ρ - коэффициент ранговой корреляции Спирмена
- τ - коэффициент ранговой корреляции Кендалла
- γ - коэффициент ранговой корреляции Гудмена – Краскела

In [ ]: