Методические указания к лабораторной работе

Тема № 1: Точечные оценки параметров методом моментов.

Цель работы – научиться генерировать выборки с заданными параметрами и определять точечные оценки параметров выборок методом моментов.

Реализация нахождения точечных оценок на сгенерированной выборке

Пример. Найдём точечные оценки параметров для выборок с нормальным законом распределения, показательным законом распределения, биномиальным законом распределения. Теоретические моменты для указанных распределений имеют вид: a, σ, 1/λ, np.

Сгенерируем выборки размером 500 со следующими параметрами:

> set.seed(0)

> x1=rnorm(500, 15, 34)

> x2=rexp(500,0.5)

> x3=rbinom(500,20,0.6)

```
set.seed(0)
x1=rnorm(500, 15, 34)
x2=rexp(500,0.5)
x3=rbinom(500,20,0.6)

str(x1)

##  num [1:500] 57.94 3.91 60.21 58.26 29.1 ...

str(x2)

##  num [1:500] 1.5106 0.1232 0.7394 2.4526 0.0542 ...

str(x3)

##  int [1:500] 11 12 14 8 11 13 14 14 11 16 ...
```

Найдем оценки неизвестных параметров

Для нормального закона распределения:

> a=mean(x1); a

> si=sd(x1); si

```
a=mean(x1); a

## [1] 14.9932
```

```
si=sd(x1); si
## [1] 33.64691
```

Для показательного закона распределения:

> lam=1/mean(x2); lam

```
lam=1/mean(x2); lam
## [1] 0.4780344
```

Для биномиального закона распределения:

> p=mean(x3)/20; p

> n=mean(x3)/0.6; n

```
p=mean(x3)/20; p
## [1] 0.6067
 n=mean(x3)/0.6; n
## [1] 20.22333
```

<u>Задание на самостоятельную работу:</u>

1. Сгенерировать выборки с заданными параметрами (см. таблицу 1);
2. Оценить параметры методом моментов;
3. Проанализировать результаты и оформить их в виде отчета.

Таблица 1 Варианты заданий (n − порядковый номер в списке группы)

	1-10	11-20	21-30
НЗР	(600+4n), 2+2n, 3+3n	(400+4n), 10+3n, 3+3n	(700+4n), 15+2n, 3+3n
ПЗР	600, 1+n	400, 3+n	700, 5+n
БЗР	600, 20+n, 0.6	500, 20+n, 0.7	700, 20+n, 0.8

Тема № 2: Сравнение способов оценивания

Целью лабораторной работы является изучение трех способов статистического оценивания.

Лабораторная работа направлена на:

— изучение трех способов оценивания (три оценки):

1. Оценка, полученная методом моментов

$$\widehat{a1} = \frac{2}{n} * \sum_{i=1}^{n} x_{i,}$$

где n – количество независимых испытаний случайной величины $X, x_{i,}$ - значение случайной величины в i- ом испытании.

2. Оценка, полученная методом максимального правдоподобия

$$\widehat{a2} = \frac{n+1}{n} * MAXx_i,$$

где n – количество независимых испытаний случайной величины $X, MAXx_{i,}$ - максимальное значение случайной величины.

3. Оценка, полученная методом порядковых статистик

$$\widehat{a3} = 2 * \widehat{x_{0.5}},$$

где $\widehat{x_{0.5}}$ - выборочная квантиль порядка 0,5, то есть выборочная медианна.

— получение навыков вычисления оценки максимального правдоподобия; медианной и суммарной оценок в R Studio;

— получение навыков вычисления характеристик разброса для оценок в R Studio;

— получений навыков сравнения трех методов оценивания.

Реализация нахождения оценок на сгенерированной выборке

Создадим 20 выборок с n=10, распределенной по НЗР:

```
```{r}

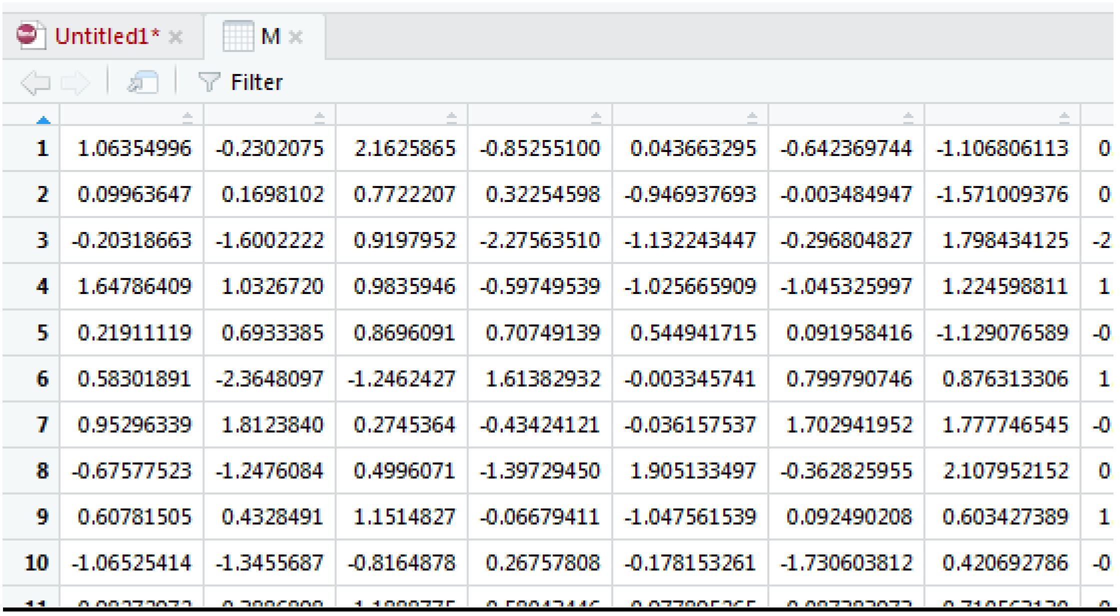
n <- 10

varCount <- 20

foo = rnorm
```
```

````
 xs <- foo (n* varCount )
 M <- matrix (xs , varCount , n)
```
````

Результат сгенерированных выборок:



1	1.06354996	-0.2302075	2.1625865	-0.85255100	0.043663295	-0.642369744	-1.106806113	0
2	0.09963647	0.1698102	0.7722207	0.32254598	-0.946937693	-0.003484947	-1.571009376	0
3	-0.20318663	-1.6002222	0.9197952	-2.27563510	-1.132243447	-0.296804827	1.798434125	-2
4	1.64786409	1.0326720	0.9835946	-0.59749539	-1.025665909	-1.045325997	1.224598811	1
5	0.21911119	0.6933385	0.8696091	0.70749139	0.544941715	0.091958416	-1.129076589	-0
6	0.58301891	-2.3648097	-1.2462427	1.61382932	-0.003345741	0.799790746	0.876313306	1
7	0.95296339	1.8123840	0.2745364	-0.43424121	-0.036157537	1.702941952	1.777746545	-0
8	-0.67577523	-1.2476084	0.4996071	-1.39729450	1.905133497	-0.362825955	2.107952152	0
9	0.60781505	0.4328491	1.1514827	-0.06679411	-1.047561539	0.092490208	0.603427389	1
10	-1.06525414	-1.3455687	-0.8164878	0.26757808	-0.178153261	-1.730603812	0.420692786	-0

Вычислим значения оценок в R и посмотрим получившиеся векторы оценок.

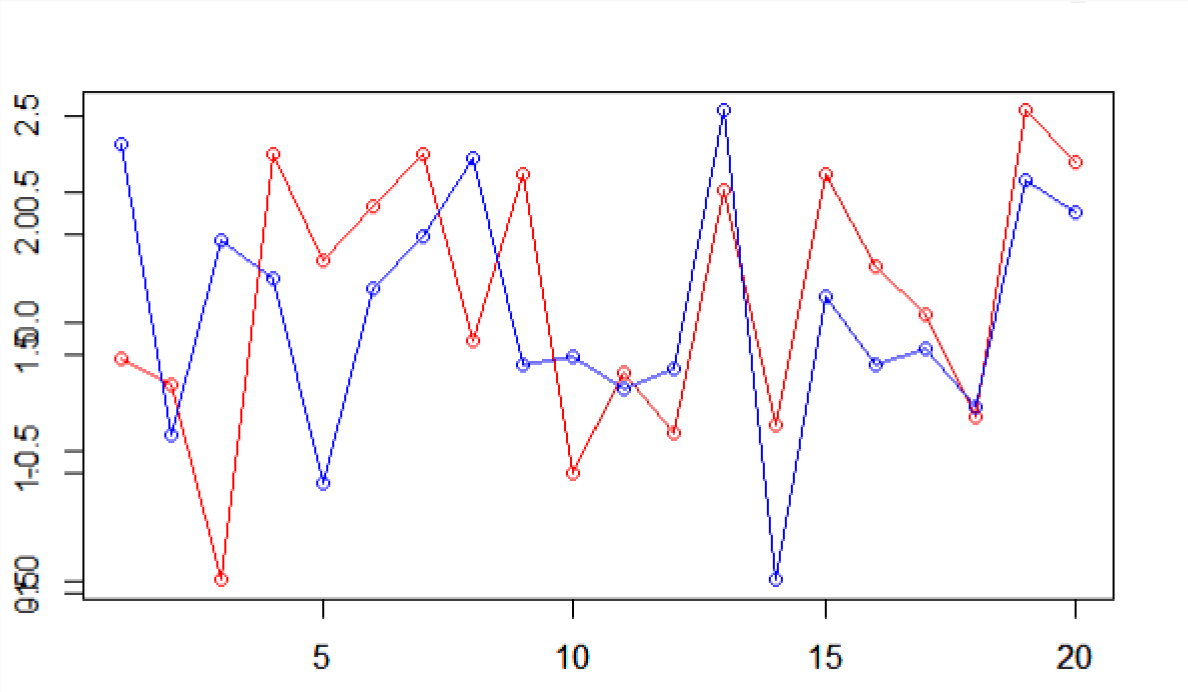
````
```{r}
rez1 <- apply (M , 1, function (x) (2/n )* sum (x) )
rez2 <- apply (M , 1, function (x) (( n +1) /n ) * max (x))
rez3 <- apply (M , 1, function (x) (2* quantile (x , c (0.5) )) )
```


```{r}
str(rez1)
str(rez2)
str(rez3)
```
num [1:20] -0.14 -0.238 -0.994 0.649 0.242 ...
num [1:20] 2.379 1.161 1.978 1.813 0.957 ...
num [1:20] -0.873 0.269 -0.695 1.017 0.764 ...
````
````

Отобразим получившиеся значения на графике:

~~~
```{r}
plot ( rez1 , col = " red ")
lines ( rez1 , col = " red")
par ( new =T )
plot ( rez2 , col = " blue ")
lines ( rez2 , col = " blue ")
par ( new =T )
plot ( rez3 , col = " yellow ")
lines ( rez3 , col = " yellow ")
par ( new =F )

```
~~~



Вычислим среднеквадратическое отклонение, максимальное и минимальное значение, величину размаха:

~~~
```{r}
p11 <- sd( rez1 )
p12 <- sd( rez2 )
~~~

```
p13 <- sd( rez3 )
```

```{r}
max1 <- max( rez1)
max2 <- max( rez2)
max3 <- max( rez2)
```

```{r}
min1 <- min( rez1)
min2 <- min( rez2)
min3 <- min( rez2)
```

```{r}
R1<-max1-min1
R2<-max2-min2
R3<-max3-min3
```

Для выборок объема n=40 и n=160 необходимо также вычислить оценки по трем способам. Свести параметры (минимальное и максимальное значение, величину размаха, среднеквадратическое отклонение между оценками для выборок объема 10, 40, 160) в одну таблицу. Отобразить на графике среднеквадратическое отклонение для всех оценок по всем проанализированным выборкам.

<u>Задание на самостоятельную работу:</u>

1. Сравнить три способа оценивания на выборках объема n=10, n=40 и n=160 для величин, распределенных по нормальному и экспоненциальному законам распределения. Параметры для генерирования выборок необходимо взять из таблицы 2 в соответствии с номером в списке группы.

Таблица 2 - Варианты заданий (n − порядковый номер в списке группы)

	1-10	11-20	21-30
НЗР	m=2+2n σ= 3+3n	m=10+3n, σ=3+3n	m=15+2n, σ=3+3n
ПЗР	λ= 1+n	λ=3+n	λ=5+n

2. Сравнить минимальные и максимальные величины, величину размаха, среднеквадратическое отклонение между оценками для выборок объема 10,40,160 для каждого закона распределения отдельно.
3. Сравнить графически среднеквадратические отклонения трех оценок для значений n=10, n=40 и n=160 для каждого закона распределения отдельно.
4. Сделать выводы о точности найденных оценок.

Тема № 3: Проверка на нормальность распределения

Распределение можно проверить на нормальность следующими способами:

1. Графические способы

Самый простой графический способ проверки характера распределения данных – это построение гистограммы. Построить гистограмму можно использую **hist()**.

Для проверки нормальности распределения также можно воспользоваться функциями **sm.density()** и **sm.density.compare()** из пакета **sm**.

install.packages("sm")

library(sm)

sm.density(X, model = "Normal", xlab="Выборка", ylab="Функция плотности распределения")

2. Формальные тесты

В R реализованы практически все имеющиеся тесты на нормальность – либо в виде стандартных функций, либо в виде функций, входящих в состав подгружаемых пакетов:

ad.test() - тест Андерсона-Дарлинга;

cvm.test() - тест Крамера фон Мизеса;

lillie.test() - тест Колмогорова-Смирнова в модификации Лиллиефорса;

sf.test() - тест Шапиро-Франсия.

В общем виде проверяемую при помощи этих тестов нулевую гипотезу можно сформулировать так: "анализируемая выборка происходит из генеральной совокупности, имеющей нормальное распределение". Если получаемая при помощи того или иного теста вероятность ошибки р оказывается меньше некоторого заранее принятого уровня значимости (например, 0.05), нулевая гипотеза отклоняется.

Рассмотрим практическое применение способов проверки на нормальность распределения в R.

Загрузим пакет ggplot2

```
library(ggplot2)
```

Загрузим пакет sm

```r
library(sm)

## Package 'sm', version 2.2-5.4: type help(sm) for summary information
```

Загрузим данные набора diamonds

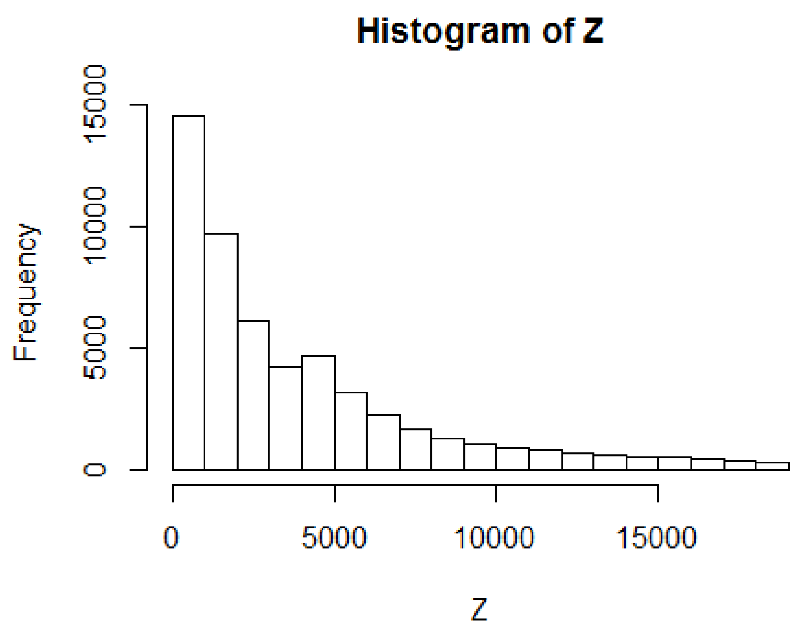
```r
diamond<-diamonds
```

Выберем из набора переменную Price

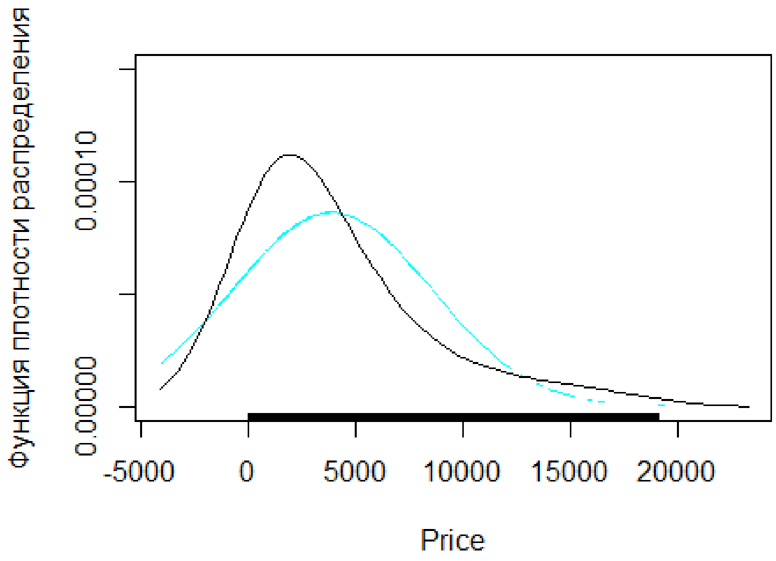
```r
Z<-diamond$price
```

Построим гистограмму ряда Price

```r
hist(Z)
```




```r
sm.density(Z, model = "Normal",xlab = "Price", ylab = "Функция плотности расп
ределения")
```



На этом графике показана кривая плотности для эмпирических данных, позволяющая оценить, насколько правдоподобно предположение о нормальности, и в каких диапазонах анализируемой переменной наблюдаются отклонения от нормального закона распределения (доверительный интервал для нормального закона распределения показан полосой голубого цвета).

Выполним тест Колмогорова-Смирнова в модификации Лиллиефорса на проверку нормальности:

library(nortest)

lillie.test(X)

```
library(nortest)

 lillie.test(Z)

##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Z
## D = 0.18467, p-value < 2.2e-16
```

Данные тест указывает на то, что нулевая гипотеза отвергается.

Задание на самостоятельную работу:

1. Выбрать самостоятельно исходные данные в виде переменной, состоящей из 20 и более наблюдений (данные должны быть реального экономического показателя, например ВВП, объем экспорта, уровень безработицы, привести ссылку на исходные данные).

   Данные можно найти в следующих источниках:

   1. https://university.prognoz.ru/biu/go/main/dataportalinfo/
   2. https://www.kaggle.com/datasets
   3. http://www.gks.ru/
   4. The R Datasets Package - https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html .
2. Определить является ли распределение показателя нормальным, определить показатели статистик Колмогорова-Смирнова и Лиллиефорса. (Вставить Скриншорт результатов)
3. На основе выполненных исследований сделать выводы о статистическом распределении показателя (симметричности, «пиковости» и т.п.), выводы о нормальности распределения показателя на основе критерия Колмогорова-Смирнова и Лиллиефорса.

Тема № 4. Изучение корреляционной зависимости двух случайных величин

Целью является исследование статистической связи двух показателей на основе расчета различных коэффициентов корреляции.

- умение вычислять коэффициент корреляции Пирсона двух случайных величин;
- умение вычислять коэффициент ранговой корреляции Спирмена и Кендалла;
- определение статистической значимости коэффициентов ранговой корреляции на основе соответствующих тестов, формирование вывода о статистической связи (о характере и силе связи) двух показателей на основании анализа коэффициентов корреляции.

Оценка корреляции двух случайных величин*

Введем две переменные X и Y.

X <- c (0.71, 0.17, 1.06, 3.21, 7.26, 0.24, 3.84, 1.96, 0.17, 7.83, 0.02, 0.99, 1.62, 1.15, 0.08, 1.09, 4.56, 0.14, 0.25, 0.53)

Y <- c (14.7, 6.23, 2.33, 14.7, 6.23, 0.96, 6.98, 3.7, 0.43, 3.33, 5.46, 31.63, 5.16, 8.17, 1.2, 15.46, 23.7, 5.16, 7.84, 0.23)

```
X <- c (0.71, 0.17, 1.06, 3.21, 7.26, 0.24, 3.84, 1.96, 0.17, 7.83, 0.02, 0.9
9, 1.62, 1.15, 0.08, 1.09, 4.56, 0.14, 0.25, 0.53)
Y <- c (14.7, 6.23, 2.33, 14.7, 6.23, 0.96, 6.98, 3.7, 0.43, 3.33, 5.46, 31.6
3, 5.16, 8.17, 1.2, 15.46, 23.7, 5.16, 7.84, 0.23)

str(X)

##  num [1:20] 0.71 0.17 1.06 3.21 7.26 0.24 3.84 1.96 0.17 7.83 ...

str(Y)

##  num [1:20] 14.7 6.23 2.33 14.7 6.23 0.96 6.98 3.7 0.43 3.33 ...
```

Для оценки тесноты связи необходимо обратиться к функции cor.test(). По умолчанию вычисляется коэффициент корреляции Пирсона при 95% доверительной вероятности.

cor.test (X, Y)

```
cor.test(X,Y)

##
##  Pearson's product-moment correlation
##
```

```
## data:  X and Y
## t = 0.49633, df = 18, p-value = 0.6257
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3440149  0.5313920
## sample estimates:
##       cor
## 0.1161946
```

Для вычисления коэффициента Спирмена при 95% доверительной вероятности в R при вызове функции cor.test() необходимо воспользоваться аргументом method со значением "spearman":

cor.test (X, Y, method = "spearman")

```
    cor.test (X, Y, method = "spearman")

## Warning in cor.test.default(X, Y, method = "spearman"): Cannot compute
## exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  X and Y
## S = 874.82, p-value = 0.1397
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.3422441
```

Пример расчета коэффициента Кендалла при 95% доверительной вероятности:

cor.test (X, Y, method = "kendall")

```
cor.test (X, Y, method = "kendall")

## Warning in cor.test.default(X, Y, method = "kendall"): Cannot compute exac
t
## p-value with ties

##
##  Kendall's rank correlation tau
##
## data:  X and Y
## z = 1.3005, p-value = 0.1934
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.212769
```

Пример расчета коэффициента Кендалла при 99% доверительной вероятности

cor.test (X, Y, method = "kendall", conf.level = 0.99)

```
cor.test (X, Y, method = "kendall", conf.level = 0.99)

## Warning in cor.test.default(X, Y, method = "kendall", conf.level = 0.99):
## Cannot compute exact p-value with ties

##
##  Kendall's rank correlation tau
##
## data:  X and Y
## z = 1.3005, p-value = 0.1934
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##      tau
## 0.212769
```

Задание на самостоятельную работу:

1. Выбрать самостоятельно исходных данные в виде переменных, состоящих из 20 и более наблюдений, между которыми предположительно есть связь (данные должны быть реального экономических показателей, например, выявить связь между уровнем безработицы и средней заработной платой в каждом регионе, привести ссылку на исходные данные)**.
   Данные можно найти в следующих источниках:
   1. https://university.prognoz.ru/biu/go/main/dataportalinfo/
   2. https://www.kaggle.com/datasets
   3. http://www.gks.ru/
   4. The R Datasets Package - https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html .
2. Вычислить коэффициенты ранговой корреляции Пирсона, Спирмена и Кендалла. (Вставить Скриншорт результатов)
3. Проверить статистическую значимость коэффициентов ранговой корреляции Спирмена и Кендалла при 95% и 99% доверительной вероятности на основе соответствующих тестов (Вставить Скриншорт результатов). Критические значения S-статистики и z-статистики при 95% и 99% доверительной вероятности определить в R.

4. Сделать выводы о силе, направлении и характере статистической связи меду исследуемыми показателями.

* Мастицкий С.Э., Шитиков В.К. (2014) Статистический анализ и визуализация данных с помощью R. – Электронная книга, адрес доступа: http://r-analytics.blogspot.com

**Примечание: временные ряды не рассматривать.

Тема № 5. Коэффициент конкордации

Использование коэффициента конкордации для оценки согласованности мнений экспертов.

Коэффициент конкордации Кендала используется для определения взаимосвязи (согласованности) оценок экспертов. Значение коэффициента конкордации может находится в диапазоне от 0 до 1. Если W=0, считается, что мнения экспертов не согласованны. Если W=1, то оценки экспертов полностью согласованны.

В литературе описание коэффициент конкордации Кендала, ограничивается формулами, в которых не учитываются связанные ранги, а критерий ограничивается требованием стремления W к единице, что существенно затрудняет его практическое использование. Тогда как в абсолютном выражении W может оказаться очень малым, но его значение будет статистически значимым для проверки гипотезы о равномерном распределении рангов (согласии ранжировок). Вычисление коэффициента конкордации без введения поправочных коэффициентов и проверки на статистическую значимость, может привести к существенным ошибкам.

Можно выделить 2 ограничения в использовании коэффициент конкордации Кендала:

- невозможность рассчитать согласованность мнений экспертов по каждой переменной в отдельности.
- коэффициент измеряет согласованность мнений в смысле их коррелированности, но не совпадения.

Реализация нахождения коэффициента конкордации в R:

Для примера, рассчитаем коэффициент конкордации и оценим согласованность мнений экспертов относительно необходимости включения в профиль должности торгового представителя отобранных 13 ключевых компетенций в R Studio.

1. Сформирована выборка из 4 экспертов, задачей которых являлось оценить по 5 балльной шкале необходимость включения в профиль должности торгового представителя 13 компетенций.

Компетенции	Эксперты			
	Эксперт1	Эксперт2	Эксперт3	Эксперт4
Ориентация на результат	5	5	5	5
Организованность	5	5	5	5
Влияние и воздействие	5	5	5	4
Коммуникативность	3	5	5	5
Уверенность в себе	5	4	4	5
Управление конфликтами	4	5	5	4
Адаптируемость	4	5	4	4
Командная работа	5	3	5	4
Ориентация на клиента	5	4	3	5
Активное слушание	4	4	5	4
Открытость	5	4	4	4
Управление рисками	5	4	4	4
Гибкость	4	3	5	4

Для нахождения коэффициента конкордации в R необходимо установ ить пакет/библиотеку *vegan* .

Функция *kendall.global* вычисляет и проверяет коэффициент согласованности между несколькими экспертами (переменными, видами) с помощью теста перестановки.

```
kendall.global(Y, group, nperm = 999, mult = "holm")
```

Функция kendall.post выполняет апостериорные тесты вкладов отдельных экспертов (переменных, видов) в общую согласованность их группы с помощью тестов на перестановку.

```
kendall.post(Y, group, nperm = 999, mult = "holm")
```

Если в таблице данных определены несколько групп экспертов, коэффициенты согласования (kendall.global) или апостериорные тесты (kendall.post) будут вычисляться для каждой группы отдельно.

Загрузим данные в формате *txt.

```
> chem <- read.table(file = "cc.txt", header = TRUE)
> View(chem)
```

Filter

	X1	X2	X3	X4
1	5	5	5	5
2	5	5	5	4
3	3	5	5	4
4	3	5	5	5
5	5	4	4	5
6	4	5	4	4
7	5	3	5	4
8	5	4	3	5
9	4	4	5	4
10	5	4	4	4
11	5	4	4	4
12	4	3	5	4
13	4	5	4	5

Применим функцию

```
> kendall.global(chem, nperm = 999, mult = "holm")
$Concordance_analysis
            Group.1
W         0.1754120
F         0.6381804
```

```
Prob.F    0.7895410
Chi2      8.4197745
Prob.perm 0.7840000


attr(,"class")
[1] "kendall.global"
```

Полученное решение показывает, что согласованность между экспертами низкая относительно необходимости включения в профиль должности торгового представителя всех отобранных компетенций, так как значение коэффициента конкордации - 0,17. Далее необходимо выдвинуть гипотезы о значимости коэффициента конкордации и определить значим ли коэффициент.

К сожалению, коэффициент конкордации не позволяет ответить на вопрос, какие из отобранных компетенций оставить, а какие исключить. Для ответа на этот вопрос можно использовать значения коэффициента вариации.

Задание на самостоятельную работу:

1) Придумать пример, состоящий из 20 и более альтернатив (например, оценка инвестиционных проектов).
2) Выбрать шкалу оценивания и оценить выбранные альтернативы по шкале 5-7 экспертами.
3) Вычислить коэффициент конкордации, сделать вывод о согласованности мнений экспертов и значимости коэффициента конкордации.