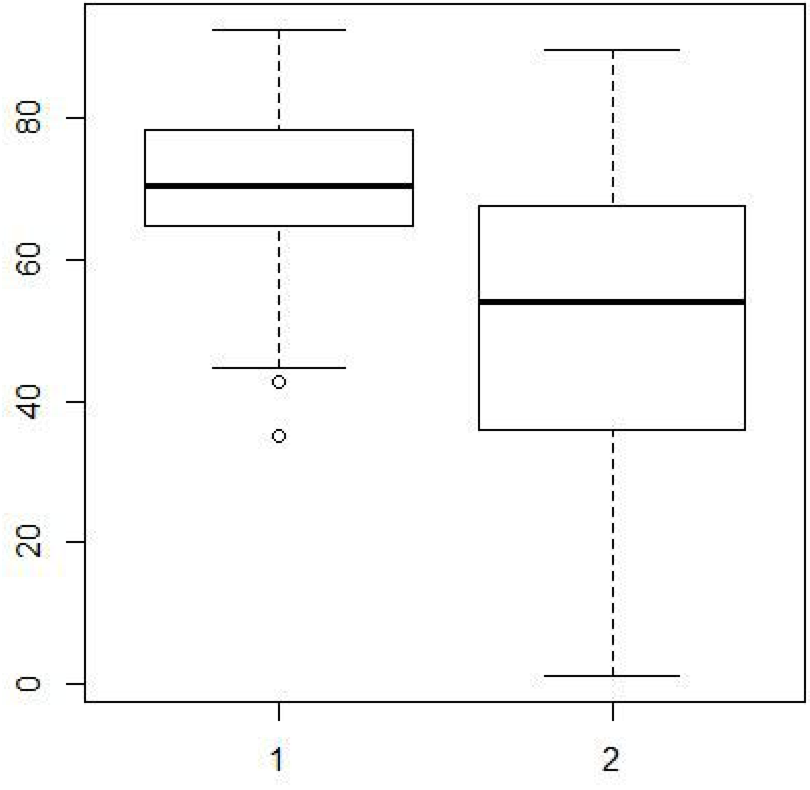
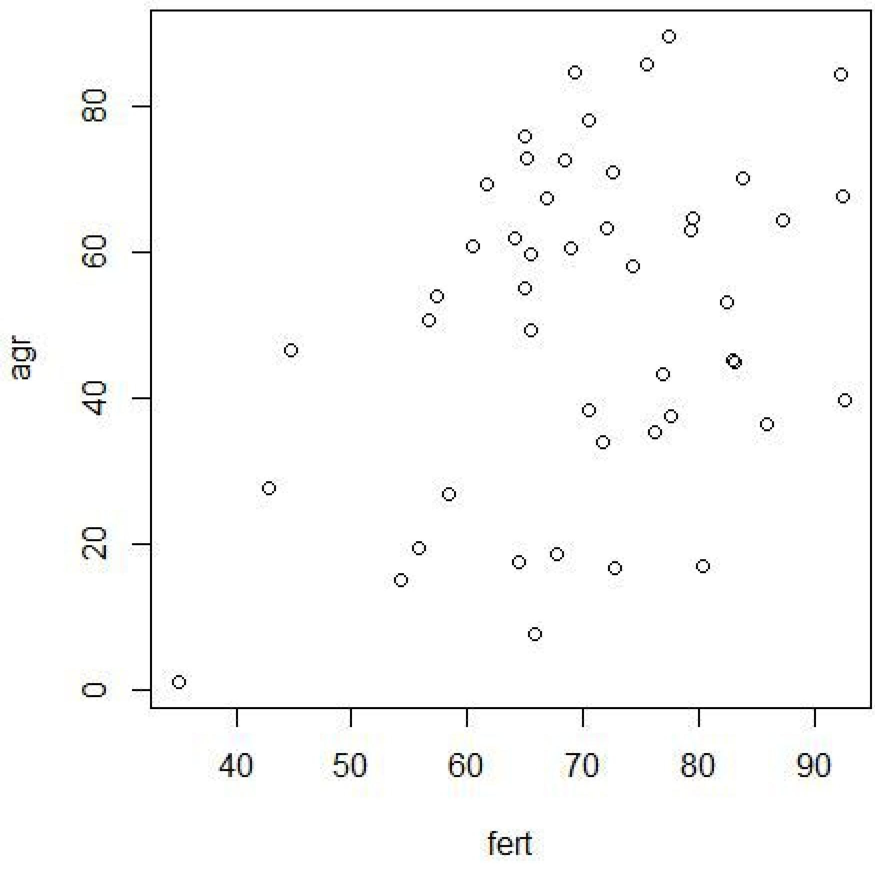
```r
fert <- swiss[,1]
agr <- swiss[,2]
exam <- swiss[,3]
educ <- swiss[,4]
relig <- swiss[,5]
mort <- swiss[,6]

features <- cbind(agr, exam, educ, relig, mort)
features_names <- c("Agriculure:        ", "Examination:      ", "Education:         ", "Religion:

#(i)

res <- rep(0,5)
for (i in (1:5)){
  spear = cor.test(features[,i], fert, method = "spearman")
  kend = cor.test(features[,i], fert, method = "kendall")
  res[i] = paste(features_names[i],
        "correlation (spearman)", toString(round(spear$estimate, 2)),
        " p-value:", toString(round(spear$p.value, 2)),
        "correlation (kendall)", toString(round(kend$estimate, 2)),
        "p-value:", toString(round(kend$p.value, 2)))
  }

res # Коэффициенты корреляции всех признаков с fertility
```

```
>
> res # Коэффициенты корреляции
[1] "Agriculure:        correlation (spearman) 0.24  p-value: 0.1 correlation (kendall) 0.18 p-value: 0.08"
[2] "Examination:       correlation (spearman) -0.66  p-value: 0 correlation (kendall) -0.48 p-value: 0"
[3] "Education:         correlation (spearman) -0.44  p-value: 0 correlation (kendall) -0.33 p-value: 0"
[4] "Religion:          correlation (spearman) 0.41  p-value: 0 correlation (kendall) 0.25 p-value: 0.02"
[5] "Infant mortality: correlation (spearman) 0.44  p-value: 0 correlation (kendall) 0.32 p-value: 0"
> W = 88, p-value = 0.1175
```

```r
#(ii)

boxplot(fert, agr) # диаграмма с "усами"
# Как можно видеть, есть несколько выбросов по fertility, по признаку agriculture выбросов нет.

M <- cbind(fert, features)

q <- boxplot.stats(fert)$out
M <- M[-which(fert %in% q),] # избавляемся от выбросов

spear = cor.test(M[,2], M[,1], method = "spearman")
kend = cor.test(M[,2], M[,1], method = "kendall")
paste("Spearman correlation:", toString(round(spear$estimate, 2)),
      " p-value:", toString(round(spear$p.value, 2)),
      "Kendall correlation:", toString(round(kend$estimate, 2)),
      "p-value:",toString(round(kend$p.value, 2)))

plot(fert, agr)
# Хоть корреляция между признаками и есть, она невелика, к тому же
# в обоих тестах p-value больше 0.1. Поэтому полученный результат незначим.

# Как мы видим, корреляция после удаления выбросов ощутимо уменьшилась.
# Коэффициент корреляции - мера линейной зависимости двух переменных, и мы строим прямую регрессии,
# которая использует сумму квадратов расстояний от наблюдаемых точек до этой прямой.
# Так как выбросы довольно далеки от средних значений, квадраты расстояний от них до прямой
# сильно влияют на результат, поэтому когда мы их убрали, корреляция пересчиталась. Могла и вырасти.
```



0
20
40
60
80
1
2

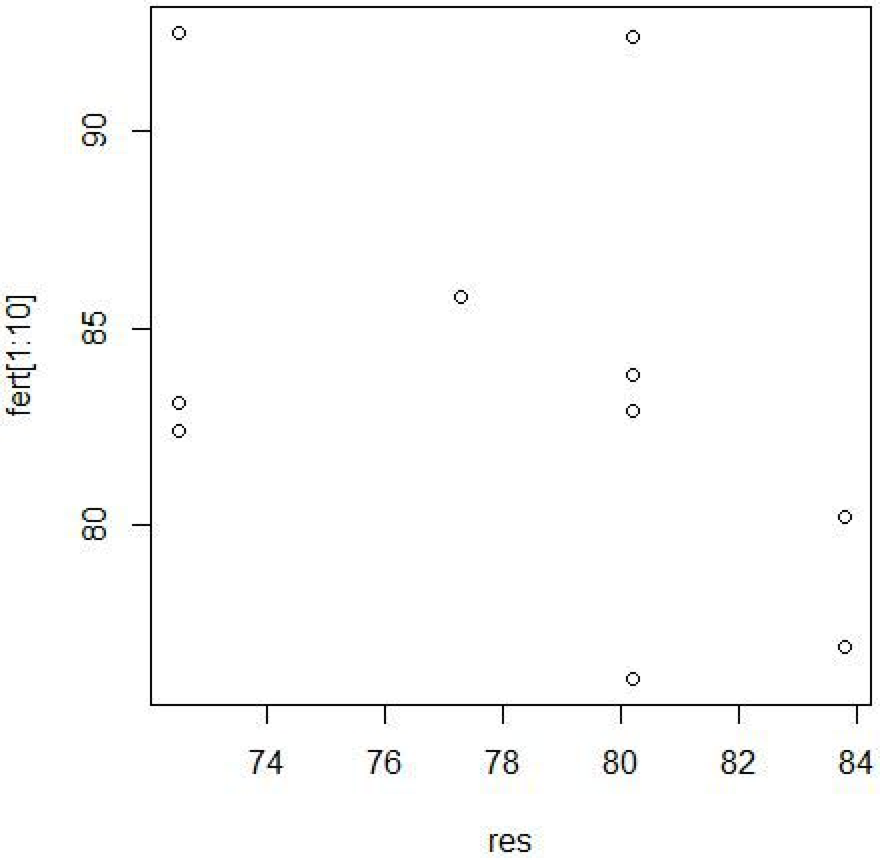


agr
0
20
40
60
80
fert
40
50
60
70
80
90

```
+         "p-value:",toString(round(kend$p.value, 2)))
[1] "Spearman correlation: 0.15  p-value: 0.32 Kendall correlation: 0.12 p-value: 0.24"
>
> plot(fert, agr)
> # Хоть корреляция между признаками и есть, она невелика, к тому же
> # в обоих тестах p-value больше 0.1. Поэтому полученный результат незначим.
>
> # Как мы видим, корреляция после удаления выбросов ощутимо уменьшилась.
> # Коэффициент корреляции - мера линейной зависимости двух переменных, и мы строим прямую регрессии,
> # которая использует сумму квадратов расстояний от наблюдаемых точек до этой прямой.
> # Так как выбросы довольно далеки от средних значений, квадраты расстояний от них до прямой
> # сильно влияют на результат, поэтому когда мы их убрали, корреляция пересчиталась. Могла и вырасти.
> W = 88, p-value = 0.1175
```

```r
#(iii)

scaled_feat <- scale(features)
dst <- rep(0, 37)
res <- rep(0,10)
for (i in 1:10){
  x = scaled_feat[i,]
  for(j in 1:37){
    dst[j] = dist(rbind(x, scaled_feat[j+10,]))
  }
  res[i] = fert[which.min(dst)]
}

plot(res, fert[1:10])
wilcox.test(res, fert[1:10], alternative = "two.sided", paired = TRUE)
# p-value больше чем 0.1, то есть даже на 10%-ном уровне значимости нулевая гипотеза о том,
# что значения fertility в 10 парах похожих провинций одинаковы, не отвергается.
```

```
> wilcox.test(res, fert[1:10], alternative = "two.sided", paired = TRUE)

        Wilcoxon signed rank test with continuity correction

data:  res and fert[1:10]
V = 11.5, p-value = 0.1139
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(res, fert[1:10], alternative = "two.sided",  :
  не могу подсчитать точное p-значение при наличии повторяющихся наблюдений
> # p-value больше чем 0.1, то есть даже на 10%-ном уровне значимости нулевая гипотеза о том,
> # что значения fertility в 10 парах похожих провинций одинаковы, не отвергается.
> W = 88, p-value = 0.1175
```



90
85
80
fert[1:10]
74
76
78
80
82
84
res

```r
#(iv)

C <- swiss[which(relig > 80),]
P <- swiss[which(relig < 20),]
M <- swiss[-which(relig > 80),]
M <- M[-which(M[,5] < 20),]

kruskal.test(list(C[,1], P[,1], M[,1])) # для нескольких (>2) выборок

# Тесты Уилкоксона
wilcox.test(C[,1], P[,1]) # H0 отвергаем
wilcox.test(C[,1], M[,1]) # H0 отвергаем на уровне значимости 5%
wilcox.test(P[,1], M[,1]) # очень большое p-значение!!
# Посмотрим, можно ли проверить альтернативу попроще:
wilcox.test(P[,1], M[,1], alternative = 'greater') # p-value велико
wilcox.test(P[,1], M[,1], alternative = 'less') # p-value порядка 0.9
# Для этой пары нулевая гипотеза не отвергается даже для прочих альтернатив
```

```
> kruskal.test(list(C[,1], P[,1], M[,1])) # для нескольких (>2) выборок

        Kruskal-Wallis rank sum test

data:  list(C[, 1], P[, 1], M[, 1])
Kruskal-Wallis chi-squared = 18.948, df = 2, p-value = 7.681e-05

>
> # Тесты Уилкоксона
> wilcox.test(C[,1], P[,1]) # H0 отвергаем

        Wilcoxon rank sum test with continuity correction

data:  C[, 1] and P[, 1]
W = 372.5, p-value = 2.158e-05
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(C[, 1], P[, 1]) :
  не могу подсчитать точное p-значение при наличии повторяющихся наблюдений
> wilcox.test(C[,1], M[,1]) # H0 отвергаем на уровне значимости 5%

        Wilcoxon rank sum test

data:  C[, 1] and M[, 1]
W = 67, p-value = 0.02496
alternative hypothesis: true location shift is not equal to 0
```

```
> wilcox.test(P[,1], M[,1]) # очень большое p-значение!!

	Wilcoxon rank sum test

data:  P[, 1] and M[, 1]
W = 88, p-value = 0.235
alternative hypothesis: true location shift is not equal to 0

> # Посмотрим, можно ли проверить альтернативу попроще:
> wilcox.test(P[,1], M[,1], alternative = 'greater') # p-value велико

	Wilcoxon rank sum test

data:  P[, 1] and M[, 1]
W = 88, p-value = 0.1175
alternative hypothesis: true location shift is greater than 0

> wilcox.test(P[,1], M[,1], alternative = 'less') # p-value порядка 0.9

	Wilcoxon rank sum test

data:  P[, 1] and M[, 1]
W = 88, p-value = 0.8929
alternative hypothesis: true location shift is less than 0

> # Для этой пары нулевая гипотеза не отвергается даже для прочих альтернатив
> W = 88, p-value = 0.1175
```

```r
#(v)

mort_1st_quart_C <- quantile(swiss[which(relig > 80), 6], 0.25) # смотрим на 0.25-квантиль
med_agr_C <- quantile(swiss[which(relig > 80), 2], 0.5)
mort_1st_quart_P <- quantile(swiss[which(relig < 20), 6], 0.25) # смотрим на 0.25-квантиль
med_agr_P <- quantile(swiss[which(relig < 20), 2], 0.5)
mort_1st_quart_M <- quantile(swiss[which((relig>=20)&(relig<=80)), 6], 0.25)
med_agr_M <- quantile(swiss[which((relig>=20)&(relig<=80)), 2], 0.5)

# Католики
C_1 <- swiss[which((relig > 80) & (agr > med_agr_C) & (mort < mort_1st_quart_C)), 1]
C_2 <- swiss[which((relig > 80) & (agr < med_agr_C) & (mort < mort_1st_quart_C)), 1]
C_3 <- swiss[which((relig > 80) & (agr > med_agr_C) & (mort > mort_1st_quart_C)), 1]
C_4 <- swiss[which((relig > 80) & (agr < med_agr_C) & (mort > mort_1st_quart_C)), 1]

# Протестанты
P_1 <- swiss[which((relig < 20) & (agr > med_agr_P) & (mort < mort_1st_quart_P)), 1]
P_2 <- swiss[which((relig < 20) & (agr < med_agr_P) & (mort < mort_1st_quart_P)), 1]
P_3 <- swiss[which((relig < 20) & (agr > med_agr_P) & (mort > mort_1st_quart_P)), 1]
P_4 <- swiss[which((relig < 20) & (agr < med_agr_P) & (mort > mort_1st_quart_P)), 1]

# Нормальные люди
M_1 <- swiss[which((relig>=20)&(relig<=80) & (agr > med_agr_M) & (mort < mort_1st_quart_M)), 1]
M_2 <- swiss[which((relig>=20)&(relig<=80) & (agr < med_agr_M) & (mort < mort_1st_quart_M)), 1]
M_3 <- swiss[which((relig>=20)&(relig<=80) & (agr > med_agr_M) & (mort > mort_1st_quart_M)), 1]
M_4 <- swiss[which((relig>=20)&(relig<=80) & (agr < med_agr_M) & (mort > mort_1st_quart_M)), 1]
```

```r
avg_C <- c(mean(C_1), mean(C_2), mean(C_3), mean(C_4))
avg_P <- c(mean(P_1), mean(P_2), mean(P_3), mean(P_4))
avg_M <- c(mean(M_1), mean(M_2), mean(M_3), mean(M_4))

avg_C
avg_P
avg_M

friedman.test(cbind(avg_C, avg_P, avg_M))
# Это значит, что нулевую гипотезу мы не отклоняем, и медианы действительно можно считать равными
```

```
>
> avg_C <- c(mean(C_1), mean(C_2), mean(C_3), mean(C_4))
> avg_P <- c(mean(P_1), mean(P_2), mean(P_3), mean(P_4))
> avg_M <- c(mean(M_1), mean(M_2), mean(M_3), mean(M_4))
>
> avg_C
[1] 77.56667 79.30000 78.66000 83.35714
> avg_P
[1] 62.68333 54.30000 67.75714 68.88333
> avg_M
[1]  NaN 35.0 68.3 42.8
>
> friedman.test(cbind(avg_C, avg_P, avg_M))

        Friedman rank sum test

data:  cbind(avg_C, avg_P, avg_M)
Friedman chi-squared = 4.6667, df = 2, p-value = 0.09697

> # Это значит, что нулевую гипотезу мы не отклоняем, и медианы действительно можно считать равн
ыми
> W = 88, p-value = 0.1175
```

Новиков Лев.   НА-03

Т1. Вычислить точное распределение для $\hat{\tau}$ при условии, что переменные независимы, для $n=4$.

$\hat{\tau} = \frac{K}{C_4^2}$, где $K = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Q((X_i,Y_i),(X_j,Y_j))$.

Здесь $Q((a,b),(c,d)) = sgn((c-a)(d-b)) \cdot I_{\{(c-a)(d-b)\neq 0\}}$

Т.е. мы имеем пары $(X_1,Y_1),(X_2,Y_2),(X_3,Y_3),(X_4,Y_4)$.

Упорядочим их по возрастанию X:

$(X_{(1)}, Y_1'), (X_{(2)}, Y_2'), (X_{(3)}, Y_3'), (X_{(4)}, Y_4')$

Рассмотрим все возможные варианты:

y_1'	y_2'	y_3'	y_4'	K	$\hat{\tau}$
1	2	3	4	6	1
1	2	4	3	4	2/3
1	3	2	4	4	2/3
1	3	4	2	2	1/3
1	4	2	3	2	1/3
1	4	3	2	0	0
2	1	3	4	4	2/3
2	1	4	3	2	1/3
2	3	1	4	2	1/3
2	3	4	1	0	0
2	4	1	3	0	0
2	4	3	1	-2	-1/3

y_1	y_2	y_3	y_4	K	$\hat{\tau}$
3	1	2	4	2	1/3
3	1	4	2	0	0
3	2	1	4	0	0
3	2	4	1	-2	-1/3
3	4	1	2	-2	-1/3
3	4	2	1	-4	-2/3
4	1	2	3	0	0
4	1	3	2	-2	-1/3
4	2	1	3	-2	-1/3
4	2	3	1	-4	-2/3
4	3	1	2	-4	-2/3
4	3	2	1	-6	-1

Получаем, что в одном случае $\hat{\tau}=1$, в одном — -1, в трёх $\hat{\tau}=2/3$, в еще трех — $-2/3$, по 5 случаев на $\tau=1/3$ и $\tau=-1/3$, и 6 случаев $\tau=0$.

Поскольку все эти варианты равновероятны и независимы, получаем, что распр-е:

$$\hat{\tau} = \begin{cases} 1, & p=1/24 \\ 2/3, & p=1/8 \\ 1/3, & p=5/24 \\ 0, & p=1/4 \\ -1/3, & p=5/24 \\ -2/3, & p=1/8 \\ -1, & p=1/24 \end{cases}$$

T2. (X_1, Y_1), (X_2, Y_2) — i.i.d.

$p_{(X,Y)}(x,y) = \frac{1}{2} y^2 e^{-x-y} \cdot I_{\{x>0 \,\&\, y>0\}}$

i) Во-первых, $p_{(x,y)}(x,y) = \frac{1}{2} y^2 e^{-y} \cdot I_{\{y>0\}} \cdot e^{-x} \cdot I_{\{x>0\}} =$

~~действительно, ведь~~ $= p_X(x) \cdot p_Y(y) \Rightarrow$

X и Y независимы

Проверяем условие нормировки: $\iint p_{(X,Y)}(x,y)\,dx\,dy$

$= \int\limits_0^\infty e^{-x} dx \cdot \int\limits_0^\infty \frac{1}{2} y^2 e^{-y} dy = \Gamma(1) \cdot \frac{1}{2} \cdot \Gamma(3) = 1 \cdot \frac{1}{2} \cdot 2 = 1.$

Аналогично с условиями нормировки для маржинальных плотностей: $\int\limits_0^\infty e^{-x} dx = \Gamma(1) = 1$,

$\int\limits_0^\infty \frac{1}{2} y^2 e^{-y} dy = \frac{1}{2} \cdot \Gamma(3) = \frac{1}{2} \cdot 2 = 1.$

Еще от функции плотности $p_{(X,Y)}(x,y) \geq 0 \ \forall x, y$, здесь это выполнено.

Следовательно, $p_{(X,Y)}(x,y)$ — действительно p.d.f.

ii) Как мы уже выяснили, X и Y независимы

В этом случае $\tau = 2\mathbb{P}((X_2 - X_1)(Y_2 - Y_1) > 0) - 1 =$

$= 2 \cdot (\mathbb{P}(X_2 - X_1 > 0) \cdot \mathbb{P}(Y_2 - Y_1 > 0) + \mathbb{P}(X_2 - X_1 < 0) \cdot \mathbb{P}(Y_2 - Y_1 < 0)) - 1 =$

$= 2 \cdot (\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2}) - 1 = 2 \cdot \frac{1}{2} - 1 = 0$

Ответ: $\tau = 0$.

Т3. $(S_1, \ldots, S_n)$ – вектор рангов, распределенный равномерно над $n!$ перестановками $\{1, \ldots, n\}$.

i) $\mathbb{E} S_i = \frac{1}{n} \cdot (1 + \ldots + n) = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$, доказано

ii) $\mathrm{Var}\, S_i = \mathbb{E} S_i^2 - (\mathbb{E} S_i)^2 = \frac{1}{n} \cdot (1^2 + \ldots + n^2) - \frac{(n+1)^2}{4} =$

$= \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = \frac{(n+1)(4n+2-3n-3)}{12} =$

$= \frac{(n+1)(n-1)}{12} = \frac{n^2-1}{12}$. Доказано

iii) $\mathrm{cov}(S_i, S_j) = \mathbb{E}(S_i S_j) - \mathbb{E}(S_i) \cdot \mathbb{E}(S_j)$

Находим $\mathbb{E}(S_i S_j)$. Все возможные пары (S_i, S_j) равновероятны $\left(\mathbb{P}(S_i = u,\ S_j = v) = \frac{1}{n(n-1)}\right)$.

Следовательно, $\mathbb{E}(S_i S_j) = \frac{1}{n(n-1)} \cdot \sum\limits_{u \neq v} uv$.

Находим $\sum\limits_{u \neq v} uv$: $\sum\limits_{u \neq v} uv = \sum\limits_{u=1}^{n} \sum\limits_{v=1}^{n} uv - \sum\limits_{w=1}^{n} w^2 =$

$= \sum\limits_{u=1}^{n} u \cdot \sum\limits_{v=1}^{n} v - \sum\limits_{w=1}^{n} w^2 = \left(\frac{n(n+1)}{2}\right)^2 - \frac{n(n+1)(2n+1)}{6} =$

$= \frac{n^2(n+1)^2}{4} - \frac{n(n+1)(2n+1)}{6} = \frac{3n(n+1) - 2 \cdot (2n+1)}{12} \cdot$

$\cdot\, n(n+1) = \frac{n(n+1)}{12} \cdot (3n^2 - n - 2) = \frac{n(n+1)(n-1)(3n+2)}{12}$.

Отсюда $\mathbb{E}(S_i S_j) = \frac{(n+1)(3n+2)}{12}$.

$\Rightarrow \mathrm{cov}(S_i, S_j) = \frac{(n+1)(3n+2)}{12} - \left(\frac{n+1}{2}\right)^2 = \frac{3n^2+5n+2}{12} -$

$- \frac{3 \cdot (n^2+2n+1)}{12} = \frac{-n-1}{12} = -\frac{(n+1)}{12}$. Доказано!

T4. $\{X_i\}_{i=1}^{m}$, $\{Y_j\}_{j=1}^{n}$. $W = \sum_{j=1}^{n} S_j$

i) W минимально $\Rightarrow$ ранги Y минимальны — ранги $1, \dots, n$ $\Rightarrow W = \sum_{j=1}^{n} j = \frac{n(n+1)}{2}$.

W максимально $\Rightarrow$ ранги Y максимальны — ранги $m+1, \dots, m+n$ $\Rightarrow W = \sum_{j=1}^{n} (m+j) = \frac{n(n+1)}{2} + mn = \frac{n(2m+n+1)}{2}$.

Ответ: $W_{MIN} = \frac{n(n+1)}{2}$, $W_{MAX} = \frac{n(2m+n+1)}{2}$

ii) Для начала рассмотрим $x = 0$:

$$\mathbb{P}\left(W = \frac{n(n+1)}{2}\right) = \mathbb{P}\left(S_1 < \min_{1\le i\le m} R_i, \dots, S_n < \min_{1\le i\le m} R_i\right) =$$
$$= \mathbb{P}(S_1 < R_1, \dots, S_1 < R_m) \cdot \ldots \cdot \mathbb{P}(S_n < R_1, \dots, S_n < R_m) =$$
$$= \prod_{i=1}^{m} \prod_{j=1}^{n} \mathbb{P}(S_j < R_i) = \left(\frac{1}{2}\right)^{mn} = 2^{-mn}.$$

$$\mathbb{P}\left(W = \frac{n(2m+n+1)}{2}\right) = \mathbb{P}\left(S_1 > \max_{1\le i\le m} R_i, \dots, S_n > \max_{1\le i\le m} R_i\right) =$$
$$= \prod_{i=1}^{m} \prod_{j=1}^{n} \mathbb{P}(S_j > R_i) = \left(\frac{1}{2}\right)^{mn} = 2^{-mn}.$$

$$\Longrightarrow \mathbb{P}(W = W_{min}) = \mathbb{P}(W = W_{MAX}) = \frac{1}{2^{mn}}$$

$x = 1$: $\mathbb{P}(W = W_{min} + 1) =$
$$= \mathbb{P}\left(S_j < \min_{1\le i\le m} R_i < \max_{1\le j\le n} S_j < R_i\right) = 2^{-(n-1)} \cdot 2^{-1} \cdot$$
$$\cdot 2^{-(m-1)} \cdot 2^{-(m-1)(n-1)} = 2^{-mn} = \frac{1}{2^{mn}}.$$

Аналогично $\mathbb{P}(W = W_{MAX} - 1) = \frac{1}{2^{mn}}$.

Допустим, что W отличается от W_{min} на x. Тогда есть несколько вариантов получить такую сумму рангов. Пример для $m=4$, $n=5$, $x=3$, чтобы было понятно:

YYYYXXXYX, YYYXYYXXX,

YYXYYYXXX. Во всех этих ~~[illegible]~~ случаях W – сумма рангов $Y_j = \frac{n(n+1)}{2}+3$.

Рассмотрим "симметричные" случаи с XYXXXYYYY, XXYXYXYYY, XXXYYYXYY.

Достаточно очевидно, что если для какой-то расстановки $W = W_{min}+x$, то для симметричной $W = W_{MAX} - x$. Причем $\forall m, n, x$.

Строгое док-во: Пусть есть ранги $R_1, \dots, R_m, S_1, \dots, S_n$, и $W = \sum_{j=1}^{n} S_j = W_{min} + x$.

Тогда для симметричной расстановки рангов $R'_1, \dots, R'_m, S'_1, \dots, S'_n$ $W' = \sum_{j=1}^{n} S'_j =$

$$= \sum_{j=1}^{n}(n+m+1-S_j) = n(n+m+1) - \sum_{j=1}^{n} S_j =$$

$$= n(n+m+1) - W = n(n+m+1) - W_{min} - x =$$
$$= n(n+m+1) - \frac{n(n+1)}{2} - x = \frac{n(2m+n+1)}{2} - x =$$
$$= W_{MAX} - x.$$

(Т.к. ранги принимают значения от 1 до $m+n$, $S_j + S_j' = m+n+1$)

Итого $W = W_{MIN} + x$ и $W' = W_{MAX} - x$.

При этом $\mathbb{P}(R_i = r_i, S_j = s_j) =$
$\mathbb{P}(R_i = m+n+1-r_i, S_j = m+n+1-s_j)$.

А это значит, что для любого x

$$\mathbb{P}(W = W_{MIN} + x) = \prod_{\sum_{j=1}^{n} s_j = W_{MIN}+x} \Big(\mathbb{P}(R_i = r_i, S_j = s_j)\Big) =$$

$$= \prod_{\sum_{j=1}^{n} s_j = W_{MIN}+x} \Big(\mathbb{P}(R_i = m+n+1-r_i, S_j = m+n+1-s_j)\Big) =$$

$$= \prod_{\sum_{j=1}^{n} (m+n+1-s_j) = W_{MAX}-x} \Big(\mathbb{P}(R_i = m+n+1-r_i, S_j = m+n+1-s_j)\Big) =$$

$$= \prod_{\sum_{j=1}^{n} s_j' = W_{MAX}-x} \Big(\mathbb{P}(R_i = r_i', S_j = s_j')\Big) = \mathbb{P}(W = W_{MAX} - x)$$

(Подразумевается произведение по всем s_j, r_i, где $\sum_{j=1}^{n} s_j$ фиксирована и равна определенному значению). Доказано.