Course of Study Bachelor Computer Science	Exercises Statistics WS 2023/24
Sheet IV - Solutions	

Descriptive Statistics - Linear Regression

1. For the X, Y data:

x_i	2	6	3	4	5
y_i	3	7	4	7	6

draw a scatterplot and compute

   (a) covariance

   (b) coefficient of correlation

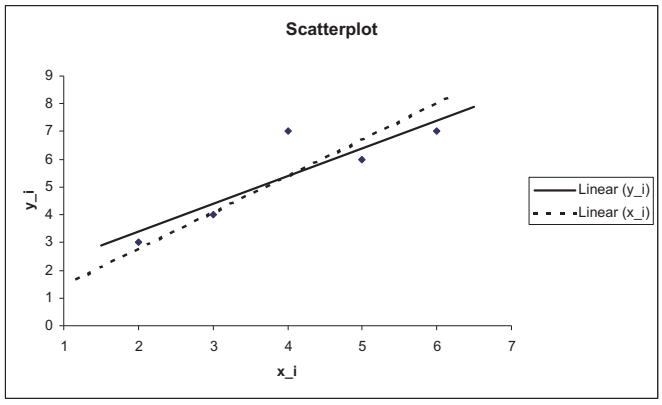
   (c) regression line: criterion variable Y and predictor variable X

   (d) regression line: criterion variable X and predictor variable Y

i	x_i	y_i	x_i * x_i	y_i * y_i	x_i * y_i
1	2	3	4	9	6
2	6	7	36	49	42
3	3	4	9	16	12
4	4	7	16	49	28
5	5	6	25	36	30
Sum	20	27	90	159	118

		X	Y
mean		4	5,4
variance		2,5	3,3
covariance		2,5	
coeff. of corr.		0,87038828	
Y = a+bX :	a =		1,4
	b =		1
X = α + β Y :	α =		-0,09090909
	β =		0,75757576

x	y
2,181818182	3
5,212121212	7



Answer:

```
#########################################################
# Descriptive Statistics: Simple example linear regr.
# Complete Solution
#
# File: des_stat_ex_lin_reg_sol.R
#
#########################################################

# data
x <- c(2,6,3,4,5)
```

```
y <- c(3,7,4,7,6)

# scatterplot
par(mfrow=c(1,1))
plot(x,y,main="scatterplot", xlim=c(0,7), ylim = c(0,8))

# covariance
cov(x,y) # 2.5

# coefficient of correlation
cor(x,y) #  0.8703883

# regression line: y=a+bx
# criterion variable Y and predictor variable X
lm(y~x)
# a = 1.4, b = 1.0
a <- lm(y~x)$coefficients[1]
b <- lm(y~x)$coefficients[2]

# regression line: x = alpha + beta * y
# criterion variable X and predictor variable Y
lm(x~y)
# alpha = -0.0909, beta = 0.7576
alpha <- lm(x~y)$coefficients[1]
beta <- lm(x~y)$coefficients[2]

# transform the regression x = alpha + beta * x to y= a' + b' x
a_strich <- -alpha/beta
b_strich <- 1/beta

# gemeinsames Diagramm
plot(x,y,main="scatterplot",
     sub="regression: y ~ x (blue), x ~y (red)",
     xlim=c(1,7), ylim = c(0,8))
abline(lm(y~x), col="blue")
abline(a=a_strich, b=b_strich,col="red")

# eps-file
#dev.copy2eps(file="../pictures/ex_bi.eps")

# Diagramm mit ggplot()
library(tidyverse)
ggplot(data = tibble(x=x, y=y)) +
  geom_point(mapping = aes(x=x,y=y)) +
  geom_abline(slope = b, intercept = a, color = "blue") +
  geom_abline(slope = b_strich, intercept = a_strich, color = "red") +
  ggtitle("scatterplot", subtitle = "regression: y ~ x (blue), x ~y (red)") +
  theme_classic()


#########################################################
# Descriptive Statistics: Simple example linear regr.
# Complete Solution
#
# File: des_stat_ex_lin_reg_sol.R
#
#########################################################

# data
x <- c(2,6,3,4,5)
y <- c(3,7,4,7,6)

# scatterplot
par(mfrow=c(1,1))
plot(x,y,main="scatterplot", xlim=c(0,7), ylim = c(0,8))

# covariance
cov(x,y) # 2.5

# coefficient of correlation
cor(x,y) #  0.8703883

# regression line: y=a+bx
# criterion variable Y and predictor variable X
lm(y~x)
# a = 1.4, b = 1.0
a <- lm(y~x)$coefficients[1]
b <- lm(y~x)$coefficients[2]

# regression line: x = alpha + beta * y
# criterion variable X and predictor variable Y
lm(x~y)
# alpha = -0.0909, beta = 0.7576
alpha <- lm(x~y)$coefficients[1]
```

```
beta <- lm(x~y)$coefficients[2]

# transform the regression x = alpha + beta * x to y= a' + b' x
a_strich <- -alpha/beta
b_strich <- 1/beta

# gemeinsames Diagramm
plot(x,y,main="scatterplot",
     sub="regression: y ~ x (blue), x ~y (red)",
     xlim=c(1,7), ylim = c(0,8))
abline(lm(y~x), col="blue")
abline(a=a_strich, b=b_strich,col="red")

# eps-file
#dev.copy2eps(file="../pictures/ex_bi.eps")

# Diagramm mit ggplot()
library(tidyverse)
ggplot(data = tibble(x=x, y=y)) +
  geom_point(mapping = aes(x=x,y=y)) +
  geom_abline(slope = b, intercept = a, color = "blue") +
  geom_abline(slope = b_strich, intercept = a_strich, color = "red") +
  ggtitle("scatterplot", subtitle = "regression: y ~ x (blue), x ~y (red)") +
  theme_classic()
```

2. For a certain class, the relationship between the amount of time spent in exercises (X) and the test score (Y) was examined.

x_i	y_i
10	5
9	5
9	4
11	6
10	7
10	5
6	3
10	4
8	5
12	7
9	4
4	2
12	8

(a) Draw a scatterplot of the data.

(b) Is there a positive or a negative association between X and Y?

(c) Compute the covariance and the coefficient of correlation.

(d) Compute the regression line $Y = a + bX$.

(e) Find the predicted test score for someone with 8 units of time spent in exercises.

(f) Interpret the values of the parameters of the regression line.

(g) Compute the proportion of variation explained by the simple linear regression.

(h) Add a the point (20,0) to the data. Inspect how this additional point influences the linear regression.

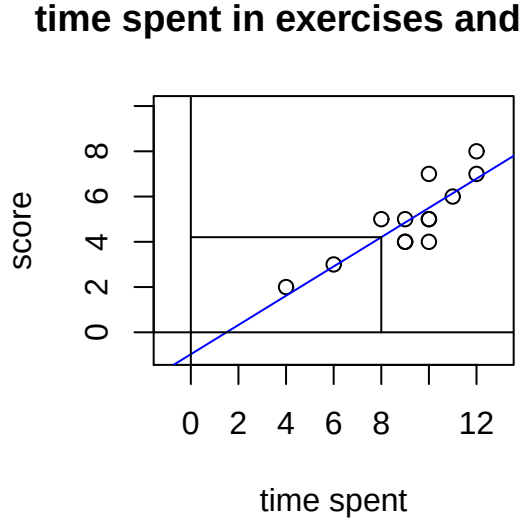
Answer:

(a)

(b) The scatter plot indicates a positive association.

(c) covariance: $s_{xy} = 3.25$, coefficient of correlation: $r = 0.861269$

(d) regression line: $Y = -0.9693878 + 0.6466837 \cdot X$



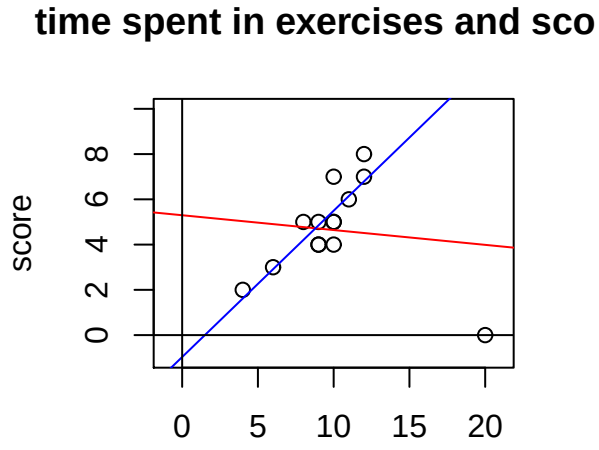
(e) predicted score $= -0.9693878 + 0.6466837 \cdot 8 = 4.204082$

(f) The coefficient a is the intercept, and b is the slope of the regression line. b can be interpreted as the additional score a student

will get if he spent 1 time unit more in exercises. A possible interpretation of a is the score a student will get if he does not spent any time in exercises. Since a is negative this interpretation is not meaningfull.

(g) $B = r^2 = 0.7417842$ proportion of variation explained by the simple linear regression

(h) If we add the point (20,0) to the data additionally we will get the red regression line instead of the former blue regression line.



The new value of coefficient of determination is $\hat{B} = 0.01258843$. We see that linear regression is quite sensitive to outliers.

```
#########################################################
# Descriptive Statistics: time spent in exercise and
#                         score in exam
# Solution
# File: des_stat_exerc_score_sol.R
#
#########################################################
# load package
library(tidyverse)

# For a certain class, the relationship between the
# amount of time spent in exercises (X) and the test
# score (Y) was examined.
results <-
  tibble(
    time =  c(10,9,9,11,10,10,6,10,8,12,9,4,12),
    score = c(5, 5,4,6, 7, 5, 3,4, 5,7, 4,2,8)
  )

# a) Draw a scatterplot of the data.
plot(x = results$time,
     y = results$score,
     main="time spent in exercises and score",
     xlab="time spent", ylab="score")
# applying ggplot()
ggplot(data = results) +
  geom_point(mapping = aes(x=time, y=score)) +
```

```
  ggtitle("relationship: time spent in exercises and score") +
  theme_bw()

# Compute the covariance and the coefficient of correlation.
results %>%
  summarise("covariance" = cov(time,score),
            "coeff. of correlation" = cor(time,score))

# Compute the regression line Y=a+bX.(X = time, Y = score)
reg1 <- lm(results$score~results$time)
# coefficients of the line
a <- reg1$coefficients[1]
b <- reg1$coefficients[2]
a;b

# add the residuals and the fitted values to the tibble results
results <- results %>%
  mutate(res.scores = reg1$residuals,
         pred.scores = reg1$fitted.values)
results

# Find the predicted score for someone with 8 units of
# time spent in exercises.
pred_test_score <- a+b*8
pred_test_score #  4.204082

plot(x = results$time,
     y = results$score,
     xlim = c(-1,13),
     ylim = c(-1,10),
     main="time spent in exercises and score",
     xlab="time spent", ylab="score")
# add the regression line
abline(reg1, col="blue")
# add line segments
segments(8,pred_test_score,0,pred_test_score)
segments(8,0,8,pred_test_score)
# add axis
abline(h=0)
abline(v=0)

# applying ggplot()
ggplot(data = results) +
  geom_point(mapping = aes(x=time, y=score)) +
  geom_smooth(mapping = aes(x=time, y=score), method = "lm",
              se = FALSE, color = "blue") +
  geom_segment(mapping = aes(x=time, y=score, xend=time, yend=pred.scores)) +
  geom_hline(yintercept = 0) +
  geom_vline(xintercept = 0) +
  ggtitle("relationship: time spent in exercises and score") +
  theme_bw()

# Compute the proportion of variation explained by the linear reg.
B <- cor(results$time,results$score)^2
B #  0.7417842

# add the point (20,0)
results <-
  results %>%
  add_row(time=20,score=0)

# calculate again the regression line
reg2 <- lm(results$score~results$time)
# blot both regression lines
plot(x = results$time,
     y = results$score,
     xlim = c(-1,21),
     ylim = c(-1,10),
     main="time spent in exercises and score",
     sub="blue=regr. without (20,0),red=regr. with (20,0)",
     xlab="time spent", ylab="score")
# add regression lines
abline(reg2, col="red")
abline(reg1,col="blue")
# add axis
abline(h=0)
abline(v=0)

# applying ggplot()
ggplot(data = results) +
  geom_point(mapping = aes(x=time, y=score)) +
  geom_smooth(mapping = aes(x=time, y=score), method = "lm",
              se = FALSE, color = "blue") +
```

```
  # changes caused by new point (20,0)
  geom_smooth(data = results %>% add_row(time=20,score=0),
              mapping = aes(x=time, y=score), method = "lm",
              se = FALSE, color = "red") +
  geom_point(x=20,y=0, color="red") +
  geom_hline(yintercept = 0) +
  geom_vline(xintercept = 0) +
  ggtitle("relationship: time spent in exercises and score",
          subtitle = "red changes caused by new point") +
  theme_bw()

# coeffiecient of determination
B_new <- cor(results$time,results$score)^2
B_new
```

Descriptive Statistics - Contingency Tables

1. **Exercise 4.1 from Heumann, Schomaker, p. 90**
   A newspaper asks two of its staff to review the coffee quality at different trendy cafes. The coffee can be rated on a scale from 1 (miserable) to 10 (excellent). The results of the two journalists X and Y are:

Cafe	X	Y
1	3	6
2	8	7
3	7	10
4	9	8
5	5	4

   (a) Calculate Spearman's rank correlation coefficient.

   (b) Does the coefficient differ depending on whether the ranks are assigned in a decreasing or increasing order?

   (c) Suppose that the coffee can only be rated as eiter good (> 5) or bad (≤ 5). Do the chances of good rating differ between the journalists?

   Answer:

   (a) Calculating the ranks of the rates we get the following table

cafe	x	y	R_x	R_y
1	3	6	1	2
2	8	7	4	3
3	7	10	3	5
4	9	8	5	4
5	5	4	2	1

We get the Spearman's rank correlation coefficient R = 0.6 by calculating the correlation coefficient of the variables R_x and R_y. Thus a high rating of X correlates with a high rating of Y.

(b) Now we use a decreasing order of rating, i.e. highest values get the lowest rank.

cafe	x	y	R_x	R_y	Rdesc_x	Rdesc_y
1	3	6	1	2	5	4
2	8	7	4	3	2	3
3	7	10	3	5	3	1
4	9	8	5	4	1	2
5	5	4	2	1	4	5

Using the values of Rdes_x, Rdes_y we get the same coefficient.

(c) The ratings are now:

quality	x	y
bad	2	1
good	3	4

Mention the definition of conditional distributions

$$f_{i|j}^{Q|J} = \frac{f_{ij}}{f_{.j}}, \quad f_{j|i}^{J|Q} = \frac{f_{ij}}{f_{i.}}$$

with Q = quality, J = journalist. The ratio of the relative risks that the journalist X rates a bad coffee resp. a good coffee compared to journalist Y are (Q = quality, J = journalist):

$$\frac{f_{1|1}^{Q|J}}{f_{1|2}^{Q|J}} = \frac{2/5}{1/5} = 2 \qquad \frac{f_{2|1}^{Q|J}}{f_{2|2}^{Q|J}} = \frac{3/5}{4/5} = 0.75$$

The odds ratio, i.e the ratio of the chances of rating a bad coffee and rating a good coffee, is

$$\frac{\frac{f_{1|1}^{Q|J}}{f_{1|2}^{Q|J}}}{\frac{f_{2|1}^{Q|J}}{f_{2|2}^{Q|J}}} = \frac{2}{0.75} \approx 2.666$$

The chance of rating a coffee as bad is 2.666 times higher for journalist X compared to journalist Y.

```
#########################################################
# Descriptive Statistics: Exercise 4.1,
# Heumann, Schomaker, page 90
# Solution
#
# File: des_stat_coffee_sol.R
#
#########################################################

# load packages
library(tidyverse)

# data
rate <-
  tibble(
    cafe = c(1,2,3,4,5),
    x = c(3,8,7,9,5),
    y = c(6,7,10,8,4)
  )
# add the ranks
rate <-
  rate %>%
  mutate(R_x = rank(x)) %>%
  mutate(R_y = rank(y))
# display data
rate

# Spearman's rank correlation coefficient
R <- cor(x = rate$R_x, y = rate$R_y)
R
# or directly using the cor() function
cor(x = rate$x, y = rate$y, method = "spearman")

# change the the order of rating
rate <-
  rate %>%
  mutate(Rdesc_x = rank(-x)) %>%
  mutate(Rdesc_y = rank(-y))
# display
rate

# Spearman's rank correlation coefficient
Rdesc <- cor(x = rate$Rdesc_x, y = rate$Rdesc_y)
Rdesc

# only the ratings good and bad
rate2 <-
  tibble(
  quality = c("bad","good"),
  x = c(2,3),
  y = c(1,4)
  )
# display
rate2

# odds ratio
OR <- ((2/5)/(3/5))/((1/5)/(4/5))
OR
```

2. The following 3x2 contingency table categorizes students according to whether or not they pass an introductory statistics course and their level of attendence:

	Course Result		
Attendance	Pass	Fail	Totals
Over 70%	40	10	50
30%-70%	20	10	30
Under 30%	10	10	20
Totals	70	30	100

   (a) Calculate the expected values in case of no association between

attendance and course result.

(b) Calculate χ^2, C and C_{corr}.

Answer:

(a)

	Course Result		
Attendance	Pass	Fail	Totals
Over 70%	40 (35)	10 (15)	50
30%-70%	20 (21)	10 (9)	30
Under 30%	10(14)	10 (6)	20
Totals	70	30	100

(b) $\chi^2 = 6.349206$, $C = 0.2443389$, $C_{corr} = 0.3455474$

```
#########################################################
# Descriptive Statistics:
# association attendance and course results
# Solution
#
# File: des_stat_attendance_result_sol.R
#
#########################################################
library(tidyverse)

# 3x2-contingency table: attendance, result
tab <- matrix(c(40,10,20,10,10,10),nrow=3,ncol=2,byrow=TRUE)
tab

# indifference table
# indifference table
ind_tab <-
  matrix(rowSums(tab),nrow=3,ncol=1)   %*%
  matrix(colSums(tab),nrow=1,ncol=2) / sum(tab)
# or
chisq.test(tab)$expected

# computation of Chi^2, C and C_corr
chisq.test(tab)$statistic
# or
chi_2 <- sum((tab-ind_tab)^2/ind_tab)
chi_2 # 6.349206
C <- (chi_2/(chi_2+sum(tab)))^0.5
C # 0.2443389
C_korr <- ((min(2,3)/(min(2,3)-1)) *chi_2/(chi_2+sum(tab)))^0.5
C_korr # 0.3455474
```

3. Exercise 4.4 from Introduction to Statistics and Data Analysis from Heumann, Schomaker, p. 91
   The famous passenger liner Titanic hit an iceberg in 1912 and sank. The data set Titanic of the package titanic contains the survival data of the passengers:

   - A total of 325 passengers travelled in first class, 285 in second class, and 706 in third class. In addition, there were 885 staff members on board.
   - Not all passengers could be rescued. The following were not rescued: 122 from the first class, 167 from the second class, 528 from the third class and 673 staff.

(a) Determine the contingency table for the variables "travel class" and "rescue status". This can be done by the following steps

- Create a tibble raw_data with the columns class (possible values: first, second, third, staff) and state (possible value: rescued and not.recued) and fill the rows with the corresponding number of pairs given by the survival data from above.
- Count the number of rescued and non rescued in the classes and the crew by applying table() to the two columns of raw_data.

(b) Use the contingency table to summarize the conditional relative frequency distribution of rescue status given travel class. Could there be an association of the two variables?

(c) What would be the contingency table table from a) look like under the independence assumption? Calculate χ^2, C and C_{corr}. Is there any association between rescue status and travel class?

(d) Combine the categorie "first class" and "second class" as well as "third class" and "staff". Create a contingency table based on these new categories. Determine and interpret χ^2, C and C_{corr}.

Answer:

(a) Contingency table:

	not rescued	rescued	Sum
first	122	203	325
second	167	118	285
third	528	178	706
staff	673	212	885
Sum	1490	711	2201

(b) Conditional frequencies: for example $f_{\text{rescued} \mid \text{first class}} = \frac{202}{337}$

	not rescued	rescued
first	0.38	0.62
second	0.59	0.41
staff	0.76	0.24
third	0.75	0.25
Sum	0.68	0.32

It is obvious that the proportion of passengers being rescued differs by class. May there is an association between the two variables that there are better chance of being rescued among passenger from higher classes.

(c) The elements of the indifference table I are given by $I_{ij} = \frac{f_{i\cdot} \cdot f_{\cdot j}}{f_{\cdot\cdot}}$

	not rescued	rescued	Sum
first	220.01	104.99	325
second	192.94	92.06	285
staff	599.11	285.89	885
third	477.94	228.06	706
Sum	1490	731	2201

The value of the coefficients measuring the association are:

$$\chi^2 = 190.4011, \quad C = 0.27, \quad C_{corr} = 0.4$$

These values are indicating a moderate association which contradicts the hypothesis derived from the values of the conditional freuencies.

(d) Combining the first and second class as well as the the third class and staff we get:

contingency table

	not rescued	rescued	Sum
first+second	289	321	610
third+staff	1201	390	1591
Sum	1490	711	2201

indifference table

	not rescued	rescued	Sum
first+second	412.95	197.05	610
third+staff	1077.05	513.95	1591
Sum	1490	711	2201

The value of the coefficients measuring the association are:

$$\chi^2 = 159.3265, \quad C = 0.26, \quad C_{corr} = 0.37$$

These coefficients indicate a moderate association.

If we consider the relative risks, which are defined by the quotient of the conditional frequencies

$$f_{i|j}^{C|S} = \frac{f_{ij}}{f_{\cdot j}}, \quad f_{j|i}^{S|C} = \frac{f_{ij}}{f_{i\cdot}}$$

with C = class and S = status

$$\frac{f_{1|1}^{S|C}}{f_{1|2}^{S|C}} = \frac{\frac{289}{610}}{\frac{1201}{1591}} \approx 0.63 \quad \frac{f_{2|1}^{S|C}}{f_{2|2}^{C|S}} = \frac{\frac{321}{610}}{\frac{390}{1591}} \approx 2.15$$

we see that the proportion of passengers who were rescued was 2.15 times higher in first and second class compared to third class and staff. Furthermore the proportion of passengers who were not rescued was 0.62 times lower in the first and second class compared to the third class and staff.
The odds ratio

$$\frac{\frac{f_{2|1}^{S|C}}{f_{2|2}^{S|C}}}{\frac{f_{1|1}^{S|C}}{f_{1|2}^{S|C}}} \approx 3.42,$$

Thus the chance of being rescued, i.e. the ratio rescued/not rescued, was was 3.42 times higher for the first and second class compared to the third class and staff.

This is plausible, as passengers in the better classes were accommodated on higher decks and thus had better access to the lifeboats, while 3rd class passengers and staff were further down the ship where the water penetrated first.

The hypothesis of independence of both variables can be checked with a statistical test (χ^2 test). If this is carried out, this hypothesis must be rejected.

```
#########################################################
# Descriptive Statistics: Exercise 4.4,
# Heumann, Schomaker, page 91
# Solution
#
# File: des_stat_titanic_sol.R
#########################################################
# load packages
library(tidyverse)
library(titanic)

# generate the data from the titanic data set
tdata <-
  as.tibble(Titanic) %>%
  spread(key=Survived, value=n) %>%
  select(-Age, -Sex) %>%
  group_by(Class) %>%
  summarise(not.rescued = sum(No),
            rescued = sum(Yes)) %>%
  mutate(Sum = not.rescued+rescued)
tdata

# conditional frequencies rescue status given class
tdata %>%
  mutate(
    not.rescued = not.rescued / Sum,
    rescued = rescued / Sum
  ) %>%
  select(-Sum) -> cond_freq
cond_freq$rescued # proportion of the rescued persons depending on the class

# use of chisq.test() to get the indifference table and chi2
tdata %>%
  select(not.rescued, rescued) %>%
  # chisq.test() needs a matrix as an input
  as.matrix() %>%
  chisq.test() -> test.res

# indifference table
test.res$expected %>% addmargins()
```

```
# chi-square
test.res$statistic -> chi2
chi2

# Pearson's contingency coefficient
nobs <- tdata$Sum %>% sum()
C <- (chi2/(chi2+nobs))**0.5
C

# corrected Pearson's contingency coefficient
C_cor <- (2/1 * chi2/(chi2+nobs))**0.5
C_cor

######################################################################
# group first and second class as well as third class and staff
######################################################################
cont_tab_new <-
  tibble(
    Class = c("first+second","third+staff"),
    not.rescued = c(tdata$not.rescued[1]+tdata$not.rescued[2],
                    tdata$not.rescued[3]+tdata$not.rescued[4]),
    rescued = c(tdata$rescued[1]+tdata$rescued[2],
                    tdata$rescued[3]+tdata$rescued[4])
  )

# indifference table
chisq.test(
  cont_tab_new %>% select(-Class) %>% as.matrix()
)$expected %>% addmargins()

# chi-square
chi2_new <-
  chisq.test(
    cont_tab_new %>% select(-Class) %>% as.matrix()
  )$statistic
chi2_new

# Pearson's contingency coefficients
C_new <- (chi2_new/(chi2_new+nobs))**0.5
C_new
C_cor_new <- (2/1 * chi2_new/(chi2_new+nobs))**0.5
C_cor_new

# Chi^2 Test, p-value < 2.2e-16
chisq.test(
  cont_tab_new %>% select(-Class) %>% as.matrix()
)

# conditional frequencies: rescue status given new classes
cond_freq_new <-
  tibble(
    Class = cont_tab_new$Class,
    not.rescued = cont_tab_new$not.rescued /
      (cont_tab_new$not.rescued + cont_tab_new$rescued),
    rescued = cont_tab_new$rescued /
      (cont_tab_new$not.rescued + cont_tab_new$rescued),
  )
cond_freq_new
cond_freq_new$rescued # proportion of the rescued persons depending on the class

# relative risks
rel_risks <- c(
  cond_freq_new$not.rescued[1] / cond_freq_new$not.rescued[2],
  cond_freq_new$rescued[1] / cond_freq_new$rescued[2]
)
# proportion of not rescued persons among 1./2. class when compared with 3.class and staff
rel_risks[1]
# proportion of rescued persons among 1./2. class when compared with 3.class and staff
rel_risks[2]

# odds ratio
rel_risks[2] / rel_risks[1] # chance to be rescued for 1./2. class when compared with 3.class and staff
```