MS-C1620 Statistical Inference

Exercise 7

Homework exercise

To be solved at home before the exercise session.

1. a. Go to the website which lists pairs of variables that have no causal relationship but still exhibit a large correlation. Pick one of the datasets and figure out how the data is presented, i.e., how are the plots constructed from the (x_i, y_i)-data (the plots are *not* scatter plots of the two variables in question), how are individual pairs (x_i, y_i) represented in the plots and what are the lines going through the points?

```
# a.
# In the plots:
# * x-axis is time
# * each time point corresponds to a single pair (x_i, y_i)
# * the x_i-value (y_i-value) of a pair is plotted on the corresponding time point in black (red)
# * the "Correlation" is calculated between the x_i-values and y_i-values
# * the lines are simply smoothed curves running through the x_i-values and y_i-values (they try to visualize the marginal t
rends).
# * Note also that the best fitting line ("y_i = a x_i + b") could not be drawn in the plot in the usual way.
```

```
b. Let $x, y, \varepsilon$ be random variables such that, $$y = x + \varepsilon,$$ where $\mathrm{Var}(x) = 1$, $\mathrm{Va
r}(\varepsilon) = \sigma^2 > 0$ and $x$ and $\varepsilon$ are independent (interpretation: $x$ and $y$ have a perfect linear
relationship but the observed value of $y$ is contaminated with the noise/measurement error $\varepsilon$ having variance
$\sigma^2$). Compute the Pearson correlation $\rho$ between $x$ and $y$ and investigate how it behaves when $\sigma^2$ is in
creased. Interpret this behavior.
```

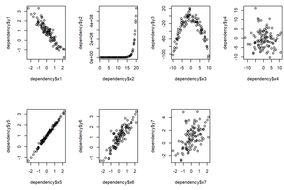
```
# rho(x, y) = cov(x, y)/(sd(x)*sd(y))
#
# Now, cov(x, y) = cov(x, x + e) = cov(x, x) + cov(x, e) = var(x) + 0 = 1,
# where the second equality uses the linearity of covariance and the third equality uses the fact that
# x and e are independent
#
# Also, sd(x) = sqrt(var(x)) = 1 and sd(y) = sqrt(var(x + e)) = sqrt(var(x) + var(e)) = sqrt(1 + sigma^2),
# again by the independence of x and e.
#
# Thus rho(x, y) = 1/sqrt(1 + sigma^2), which decreases towards zero as sigma^2 increases.
# The interpretation for this is that increasing the noise strength (variance) masks the true perfect relationship and the c
orrelation gets weaker (the point-pairs deviate more and more from the straight line). See class exercise 1 for visual versi
on of the same phenomenon.
```

Class exercise

To be solved at the exercise session.

1. The file `data_dependency.txt` contains seven bivariate data sets (the columns `x1` and `y1`, where $i = 1, 2, \ldots, 7$, always form a pair).
   a. Read the file into R using the command `read.table`.
   b. Draw a scatter plot for each pair of variables.
   c. Calculate the Pearson and Spearman correlations of the pairs and compare them to the scatter plots.
   d. The underlying distributions of the samples 5-7 are the same up to the variance of `y1` (the variance is highest in sample 7). What happens to the correlation coefficients as the variance increases and why?

```
# a.
# Replace "params$your_path_here_1" with your path to the .txt file in the code below (and remember that "/" is used to navi
gate sub-folders in R)
dependency <- read.table(params$your_path_here_1, sep = "\t", header = TRUE)


# b.
par(mfrow = c(2, 4))
plot(dependency$x1, dependency$y1)
plot(dependency$x2, dependency$y2)
plot(dependency$x3, dependency$y3)
plot(dependency$x4, dependency$y4)
plot(dependency$x5, dependency$y5)
plot(dependency$x6, dependency$y6)
plot(dependency$x7, dependency$y7)
par(mfrow = c(1, 1))
```



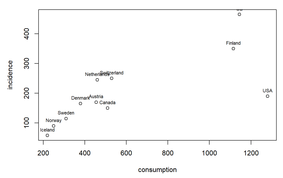

```
# c.
cors <- function(i){
  cor_p <- cor(dependency[, (2*i - 1)], dependency[, 2*i])
  cor_s <- cor(dependency[, (2*i - 1)], dependency[, 2*i], method = "spearman")
  print(paste("Data set ", i, ", Pearson: ", round(cor_p, 3), ", Spearman: ", round(cor_s, 3)))
}

for(i in 1:7){
  cors(i)
}
```

```
## [1] "Data set 1, Pearson: -0.924, Spearman: -0.922"
## [1] "Data set 2, Pearson: 0.51, Spearman: 1"
## [1] "Data set 3, Pearson: -0.035, Spearman: -0.028"
## [1] "Data set 4, Pearson: -0.05, Spearman: 0.004"
## [1] "Data set 5, Pearson: 0.994, Spearman: 0.993"
## [1] "Data set 6, Pearson: 0.879, Spearman: 0.855"
## [1] "Data set 7, Pearson: 0.582, Spearman: 0.587"
```

```
# Interpretation:
# 1. almost perfect decreasing (linear/monotone relationship
# 2. no clear linear relationship but perfect increasing monotone relationship
# 3. symmetric increasing-decreasing relationship -> both correlations zero ("increase masks decrease")
# 4. no discernible relationship -> both correlations zero
# 5.-7. increasing linear/monotone relationship which gets more and more difficult to see because of the increasing y-varian
ce. That is, increasing the y-variance hides the linear relationship under the added "noise", decreasing the correlations. T
his is the same phenomenon as in homework problem 1b.


# d.
# See above.
```

2. The file `data_tobacco.txt` contains data on cigarette consumption and lung cancer incidences from 11 different countries. The variable `consumption` describes the yearly consumption of cigarettes per capita in 1930 and the variable `incidence` tells the lung cancer incidence rates per 100 000 people in 1950. We use correlation to study the connection between these two.
   a. Read the file into R using the command `read.table`.
   b. Draw a scatter plot of `consumption` and `incidence` which also shows the country names.
   c. Using the scatter plot, make an educated guess on the signs and magnitudes of the Pearson and Spearman correlations of the two variables.
   d. Calculate the Pearson and Spearman correlations.
   e. Use permutation test to test whether the two correlations differ significantly from zero, using the significance level 5%.
   f. Drop USA from the data, redo the previous analysis and compare the results to those obtained with the full data. What happened?

```
# a.
# Replace "params$your_path_here_2" with your path to the .txt file in the code below (and remember that "/" is used to navi
gate sub-folders in R)
tobacco <- read.table(params$your_path_here_2, sep = "\t", header = TRUE)


# b.
plot(incidence ~ consumption, data = tobacco)
text(tobacco$consumption, tobacco$incidence, labels = tobacco$country, cex= 0.7, pos=3)
```



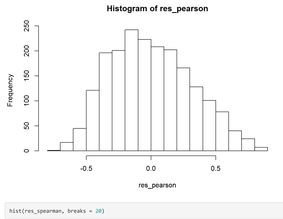

```
# c.
# Apart from USA, the data lies very accurately on a line.
# The outlier most likely decreases the value of Pearson correlation (which is sensitive to outliers) somewhat, and it could
then equal -0.80. The Spearman correlation is more tolerant to outliers (because it ignores the magnitudes of the outliers)
and could then equal something bit larger, e.g. -0.90.


# d.
cor(tobacco$consumption, tobacco$incidence)
```

```
## [1] 0.7409723
```

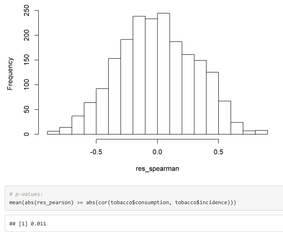
```
cor(tobacco$consumption, tobacco$incidence, method = "spearman")
```

```
## [1] 0.8454545
```

```
# Quite close...


# e.
# Permutation tests
n <- nrow(tobacco)

B <- 2000
res_pearson <- rep(0, B)
res_spearman <- rep(0, B)

for(b in 1:B){
  res_pearson[b] <- cor(tobacco$consumption, sample(tobacco$incidence, n, replace = FALSE))
  res_spearman[b] <- cor(tobacco$consumption, sample(tobacco$incidence, n, replace = FALSE), method = "spearman")
}

# Distributions of the permutation test replicates (distribution of the test statistic under H0)
hist(res_pearson, breaks = 20)
```




```
hist(res_spearman, breaks = 20)
```




```
# p-values:
mean(abs(res_pearson) >= abs(cor(tobacco$consumption, tobacco$incidence)))
```

```
## [1] 0.011
```

```
mean(abs(res_spearman) >= abs(cor(tobacco$consumption, tobacco$incidence, method = "spearman")))
```

```
## [1] 0.002
```

```
# Both smaller than 0.05 -> both correlations differ significantly from 0


# f.
tobacco <- tobacco[-7, ]
# Running the previous code to remove USA and then redoing the steps yields
# Pearson correlation: 0.941, p-value: 0
# Spearman correlation: 0.927, p-value: 0.001
# The correlations are higher and more significant without the "outlier" which masked the "true" relationship.

# Note: In practice, the labeling of USA as an outlier and the consecutive removal of it should be somehow
# justified.
```

3. **(Optional)** Use also the tests given on slides 6.16 and 6.20 to test the null hypothesis $H_0 : \rho = 0$ for Pearson correlation in problem 2e. How do the results compare to the permutation test?

4. **(Optional)** Simulate the distribution of the sample Pearson correlation $\hat{\rho}$ under normality by generating multiple datasets of size n from a bivariate normal distribution of your choice. Then transform the sample Pearson correlations as $\hat{\rho} \mapsto \arctanh(\hat{\rho})$ and inspect the distribution of the transformation. Does it look normal? (it should for large n, as per slide 6.13)