MS-C1620 Statistical Inference

Exercise 6

Homework exercise

To be solved at home before the exercise session.

---

1. a. Assume that we have an iid. random sample $x_1, \ldots, x_{1000}$ and we'd like to use the normal Q-Q plot to assess whether the sample came from a normal distibution. How do you expect the normal Q-Q plot to roughly look like (i.e. what general features do you expect it to have and *why*), if the true distribution of the data is
      i. a normal distribution,
      ii. a right-skew distribution,
      iii. a left-skew distribution,
      iv. a bimodal distribution,
      v. a distribution with light tails,
      vi. a distribution with heavy tails?
   b. Recall the differences between the interpretations of the χ^2 homogeneity test and χ^2 test for independence. Come up with a practical situation where the collected data can be expressed as a 2-by-2 table and a related research question for which the correct interpretation is through
      i. the χ^2 homogeneity test,
      ii. the χ^2 test for independence.

Class exercise

To be solved at the exercise session.

*Note: all the needed data sets are either given below or available in base **R**.*

---

1. The data set `rock` contains measurements on 48 rock samples from a petroleum reservoir. Treat the data as an iid. random sample from some distribution and test whether the distribution of `shape` is normal.
   a. Visualize the data to obtain a preliminary idea of the possible normality of the data.
   b. Use the normal Q-Q plot to gain more evidence on the normality/non-normality of the data.
   c. Conduct the Bowman-Shenton (Jarque-Bera) and the Shapiro-Wilk tests of normality on significance level 0.05.
   d. After all the previous, would you conclude the data to be normal (or normal enough for methods with normality assumptions)?
   e. Why is the data not really iid.?

---

2. The data set `randu` contains 400 triples of successive random numbers from the random number generator RANDU. Use the χ^2 goodness-of-fit test to assess whether the first elements in the triplets really obey the uniform distribution on $[0, 1]$.
   a. Extract the first elements in the triplets and visualize their sample distribution.
   b. Discretize the values into a suitable number of categories and calculate the observed category frequencies.
   c. Compute the corresponding expected category probabilites under the uniform distribution on $[0, 1]$.
   d. Recall the hypotheses of the test and conduct it on significance level 0.05.
   e. What are the conclusions of the test? Compare your results with someone who used a different choice of categories for the discretization.

---

3. The data set `Titanic` contains information on the fate of passengers on the fatal maiden voyage of the ocean liner "Titanic". We use the data to study whether there is a connection between the sex (Male/Female) of a passenger and surviving from the ship (No/Yes).
   a. Extract a marginal table containing only the cross-tabulation of the variables `Sex` and `Survived`.
   b. Find a suitable way to visualize the data.
   c. Which test is appropriate for these data (and why?), χ^2 homogeneity test or the χ^2 test for independence?
   d. Conduct your chosen test on significance level 0.05 and state your conclusions.

---

4. **(Optional)** Choose your favorite non-normal distribution and use simulations to study the Type II error probabilities of the Bowman-Shenton (Jarque-Bera) and Shapiro-Wilk tests of normality for that distribution on different sample sizes (e.g. $n = 10, 100, 1000, 10000$). That is, find out the probabilty of falsely concluding that the data comes from a normal distribution when it does not.
