Universität Bielefeld
Tanja Pasurek
Ehsan Abedi WiSe 2024/25

Foundations of Statistics

Homework 9

Association of two random variables (Chapter 2.5)

Exercise 1. The famous passenger liner Titanic hit an iceberg in 1912 and sank.
A total of 337 passengers travelled in first class, 285 in second class, and 721 in third class. In addition, there were 885 staff members on board.
Not all passengers could be rescued. Only the following were rescued: 135 from the first class, 160 from the second class, 541 from the third class and 674 staff.

(a) Determine and interpret the contingency table for the variables "travel class" and "rescue status."

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

(c) What would the contingency table from (a) look like under the independence assumption? Calculate Cramer's V statistic. Is there any association between travel class and rescue status?

(d) Given the results from (a) to (c), what are your conclusions?

Exercise 2. Consider the `Animals (MASS)` data set that records average body weight (`body`) and brain size (`brain`) for several species, some quite extinct.

Brain–body mass ratio is hypothesized to be a rough estimate of the intelligence of an animal (although fairly inaccurate in many cases); see:
`https://en.wikipedia.org/wiki/Brain-body_mass_ratio`

We would expect that larger bodies would be paired off with larger brains, leading to a positive correlation closer to 1 than 0.

(a) Check the hypothesis by computing the Bravais-Pearson correlation coefficient. Plot the data both in "normal" and "log-log" scales.

(b) Mark the dinosaur species in the plots. Now, remove the dinosaur species (as outliers) and recompute the aforementioned coefficient.

(c) Compute Spearman's rank correlation coefficient in both cases (with and without outliers). Which coefficient is more robust to the presence of outliers in the dataset?

Exercise 3. The file `bodytemp` in the package `evidence` (Analysis of Scientific Evidence Using Bayesian and Likelihood Methods), which you can install from `CRAN`, contains normal body temperature measurements (degrees Fahrenheit) and heart rates (beats per minute) of 65 males (coded by 1) and 65 females (coded by 2).

(a) For both males and females, make scatterplots of heart rate versus body temperature. Comment on the relationship or lack thereof.

(b) Quantify the strengths of the relationships by calculating Bravais-Pearson and Spearman's rank correlation coefficients.

(c) Does the relationship for males appear to be the same as that for females? Examine this question graphically, by making a scatterplot showing both females and males and identifying females and males by different plotting symbols.

Exercise 4. Let $X, Y : \Omega \to \mathbb{R}$ be two random variables such that $\mathbb{E}(X^2), \mathbb{E}(Y^2) < \infty$ and let $\mathrm{Var}(Y) \neq 0$.

We aim to approximate X by a linear function $a + bY$ of Y in mean square error.

(a) Show that the quadratic deviation $\mathbb{E}[(X - a - bY)^2]$ is minimized by

$$a_* = \mathbb{E}(X - b_* Y) \quad \text{and} \quad b_* = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(Y)}.$$

(b) What does this mean when X and Y are uncorrelated?

Parameter estimation (Chapter 3.1–3.4)

Exercise 5. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Geo}(p)$ be a random sample taking values in $\{1, 2, 3 \cdots\}$ from a geometric distribution with unknown parameter $p \in (0, 1)$.
(Note that there are two definitions for the geometric distribution. Here we work with "the number of Bernoulli trials needed to get one success.")

(a) With the method of moments, find an estimator $\hat{p}_{\text{MoM}}$ for p.

(b) Check that this estimator is biased by using Jensen's inequality.

(c) With the maximum likelihood method, find an estimator $\hat{p}_{\text{ML}}$ for p. Compare $\hat{p}_{\text{MoM}}$ and $\hat{p}_{\text{ML}}$.

Exercise 6. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Exp}(\lambda)$, where $\lambda > 0$ is unknown.

(a) Check that $\widehat{\mu}_n := \overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ is an unbiased estimator for the population mean $\mu := 1/\lambda$.

(b) Let M_n denote the smallest order statistics of $X_1, ..., X_n$. Recall that in HW6, Exc. 3(d), we proved that $M_n \sim \text{Exp}(n\lambda)$. Show that $\widetilde{\mu}_n := nM_n$ is another unbiased estimator for μ as well.

(c) Which estimator would you prefer for estimating μ. To this end, calculate the variances of $\widehat{\mu}_n$ and $\widetilde{\mu}_n$ and justify your answer.