Universität Bielefeld
Tanja Pasurek
Ehsan Abedi WiSe 2023/24

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?

Point estimation (Chapter 3)

Exercise 3. Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Geo}(p)$ be a random sample taking values in $\{1, 2, 3 \cdots\}$ from a geometric distribution with unknown parameter $p \in (0, 1)$.

(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 (see the **Addendum** to HW 9).

(c) Apply the maximum likelihood method to obtain an estimator $\hat{p}_{\text{ML}}$ for the parameter p. Compare $\hat{p}_{\text{MoM}}$ and $\hat{p}_{\text{ML}}$.

(d) The following dataset with size $n = 15$ comes from a geometric distribution:

```
2, 12, 2, 2, 2, 2, 1, 2, 9, 1, 2, 4, 4, 1, 1.
```

Your task (as a statistician!) is to estimate the underlying parameter p. Calculate $\hat{p}_{\text{ML}}$ for this dataset.

(e) To check your results, find $\hat{p}_{\text{ML}}$ numerically in `R` (like in Ch. 3.5 on pages 14–15). When choosing starting values, you may take into account your estimates obtained in task (d).

Exercise 4. Let $X_1, ..., X_n$ be an i.i.d. random sample from a uniform distribution $\text{Unif}(a, b)$ with unknown parameters $a, b \in \mathbb{R}$ $(a < b)$.

(a) Show that $\widehat{a}_n := \min\{X_1, ..., X_n\}$ and $\widehat{b}_n := \max\{X_1, ..., X_n\}$ are the maximum likelihhod estimators for a and b. (*Hint:* proceed analogously to Example 3 in Ch. 3.4.)

(b) Check that $\widehat{a}_n$ and $\widehat{b}_n$ are asymptotically unbiased, that is,

$$\mathbb{E}[\widehat{a}_n] \to a \quad \text{and} \quad \mathbb{E}[\widehat{b}_n] \to b \quad \text{as} \quad n \to \infty.$$

(*Hint:* In HW 7, Exercise 2(c), we have already found the CDFs of the minimum and maximum of a uniformly distributed random sample. Here you need to compute the corresponding expectations.)

(c) Let $\tau := \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x$, where $f(x)$ is the PDF of Unif(a, b). Find the MLE $\hat{\tau}_n$ of τ.

Exercise 5. Suppose a dataset $x_1, ..., x_n$ is a realization of a random sample $X_1, ..., X_n$ from an Exp(λ) distribution, 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 minimum of $X_1, ..., X_n$. Check (based on HW 7, Exercise 2(d)) that M_n has an Exp$(n\lambda)$ distribution.

Show that $\widetilde{\mu}_n := nM_n$ is an 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.

Exercise 6. Let $X_1, ..., X_n$ be an i.i.d. random sample from a Gamma distribution $\Gamma(\alpha, \beta)$ with unknown parameters $\alpha, \beta > 0$, see p. 24 in Ch. 1.6.

(a) Find the method of moments estimator $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ of the parameter vector $\theta = (\alpha, \beta)$ (*Hint:* proceed as described on p. 10 in Ch. 3.4.)

(b) Define the log-likelihood function $\ell(\alpha, \beta)$ and write down the system of equations to find its maximum. Could you solve it explicitly?