Statistical Models and Data Analysis

Summer term 2017

Problem Set 9 **13.7.2017**

Please try your hand at these exercises and be ready to pose any questions by 17.7.2017. You can send me email at stemmler@bio.lmu.de if you have questions.

---

1. (Mean and Variance) Suppose the number of visitors to a museum in a given day is N. We have estimates for the mean and variance of N, namely $\mathbb{E}(N) = 400$ and $\mathrm{Var}(N) = 200$. Let Y_i represent the monetary value of souvenirs from the museum store that the i-th person buys. You know, things like t-shirts with the slogans "History, it's about time" or "May the Mass $\times$ Acceleration be with you". These sell for outrageous prices, but you feel good about yourself because you are supporting a worthy cause. Let the Y_i's be independent of N and also of each other. Furthermore, let us suppose that that each person has the same variance and same mean

$$\begin{aligned} \mathbb{E}Y_i &= 20. \\ \mathrm{Var}(Y_i) &= 625. \end{aligned}$$

Note that the standard deviation (square root of the variance) is higher than the mean (some museum visitors buy nothing). Let S be the total sales $S = \sum_{i=1}^{N} Y_i$.

(a) Find the values for $\mathbb{E}(S)$ and $\mathrm{Var}(S)$.

(b) Given your knowledge of statistics, do you think S will be approximately distributed as a Gaussian?

(c) If S were Gaussian, what is the probability that the sales on any given day will be above $10,000$ Euros?

2. (Entropy and Information) Imagine that every letter in the English language occurs with equal likelihood. We ignore spaces and punctuation. There are 26 letters in the alphabet, and you receive a stream of these letters on your mobile device, which you are checking incessantly, instead of doing your homework (tsk, tsk!). Only now, because your wi-fi reception is poor, every 8th letter is garbled, i.e., replaced by a random letter in the alphabet.

(a) Compute the source entropy rate in bits/symbol.

(b) Compute the noise entropy rate. From the noise and source entropy rates, calculate the information rate, once again in bits/symbol.

(c) (No computation:) In reality, not all letters of the alphabet have the same probability. The letter 'e', for instance, is much more common than the letter 'q'. Qualitatively, how will this affect the entropy and information rates above? Assume that errors occur in the same way as before, i.e., a symbol is replaced by a random symbol, with all random symbols having equal likelihood. All possible replacements are equally likely, such that $p = 1/26$ for any letter that results from replacement. Messaging often uses abbreviations like YOLO and LOL and 'u up'? How do abbrevations like this affect the source entropy?

(d) I told my cat YOLO once, and she gave me a quizzical look. Please explain.