Problems on *statistics and data analysis (MVComp2)*

lecturer: Björn Malte Schäfer
head tutor: Alessio Spurio Mancini
summer term 2016

Problem sheet 6

To be handed in during the exercise group on 13.June.2016

1. *Shannon entropy and Gaussian distribution* *(10 points)*

Consider the differential Shannon entropy, which is a functional of a probability density $p(x)$ defined on an interval $[a, b]$,

$$S\left[p\right] = -\int_a^b \mathrm{d}x\, p\left(x\right) \ln p\left(x\right), \tag{I}$$

where $p(x)$ is properly normalised, $\int_a^b \mathrm{d}x\, p\left(x\right) = 1$.

(a) Which probability density maximizes $S\left[p\right]$ on the interval $[a, b]$?

(b) Maximise $S\left[p\right]$ on the real interval $[a, b]$ with a fixed variance

$$\int_a^b \mathrm{d}x\, x^2 p\left(x\right) = \sigma^2.$$

and give an explicit expression for $p(x)$.

2. *Skewness of the exponential distribution / Python exercise* *(20 points)*

Consider an exponential distribution $p(x) = \exp(-x)$, $x = 0 \cdots + \infty$.

(a) Calculate *analytically* the value of the skewness s, defined as $s \equiv \left\langle (x - \langle x \rangle)^3 \right\rangle$.

Now write a Python script which takes samples from the the exponential distribution defined above and estimates the centralised skewness from 100 samples. Then, repeat the process and plot the distribution of the samples of the skewness. Also, plot the cumulative distribution and its complement.

The final result should look similar to figure 1.

(b) Write down the mean value of the estimates that you get from your script: Is it close to 2?

(c) Test the script for $s \geq 3$: Imagine that there's one estimate of s which is equal to 3. At what probability does such a thing occur? Also, determine the percentage of samples of the skewness that have value larger than 3.

(d) Are these outliers more or less likely if there are more samples?

(e) Calculate the value of s_{max} such that only 20 % of the samples are larger.

(f) If you have defined the size of the test α to be 0.1, where is the limit s_{max} in the estimate?

(g) Again, let us consider we have one estimate that is equal to 3. Would you reject the hypothesis that it comes from a distribution with skewness=2 at $\alpha = 0.1$?

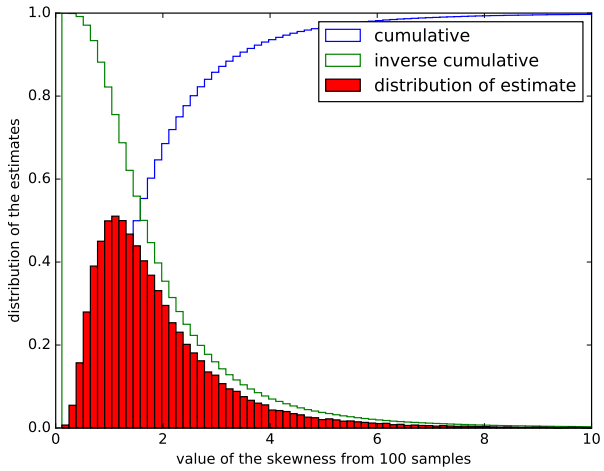

Figure 1: Plots obtained with the Python script in exercise 2

3. *Drug testing* *(10 points)*

Imagine that criminal investigators run a test for drug abuse. The test indicates a positive result (+) if a person has taken drugs (D) with a probability of $P(+|D) = 0.99$ and yields a negative result (−) of a person who has not taken drugs (N) with a probability of $P(-|N) = 0.98$.

(1) What are the probabilities $P(-|D)$ and $P(+|N)$?

(2) What's the total probability $P(+)$ for a positive test result assuming that $P(D) = 0.005$ of the population has taken drugs?

(3) Determine the conditional probability $P(D|+)$ using Bayes' law and explain your derivation.

(4) Did you get a surprising number? Comment on it.

4. *Monkeys* *(10 points)*

Imagine that a troop of monkeys throws bananas (N in total) into M baskets. When the monkeys are done, there are n_i bananas in the basket i, $N = \Sigma_{i=0}^{M} n_i$.

(a) How many possibilities are there to put n bananas into M baskets?

(b) What is the number of possibilities to put n_1 bananas into the first basket?

(c) What is the number of possibilities to put n_1 bananas into basket 1, n_2 bananas into basket 2, $\ldots n_M$ bananas into basket M?

(d) What is the probability that one obtains a specific set of $\{n_i\}$?

(e) Show that for large n with Stirling's formula $\ln n! \simeq n \ln n - n$ one obtains

$$\ln p\left(\{n_i\}\right) = -N \ln M + N \ln N - \sum_i n_i \ln n_i$$

(f) Define the probability p_i that a banana lands in basket i, $p_i \equiv n_i/N$. Show that

$$\ln p\left(\{p_i\}\right) = -N \ln M - N \sum_i p_i \ln p_i.$$