Tutorial 2, Statistical Mechanics: Concepts and applications 2016/17 ICFP Master (first year)

Maurizio Fagotti, Olga Petrova, Werner Krauth

Tutorial exercises

I. WORKSHEET: STATISTICAL INFERENCE

1. **Maximum Likelihood Method for the Bernoulli distribution.**

Reminder: The likelihood function is the joint probability density of the data $X_1, ..., X_N$:

$$L(\theta) = \prod_{i=1}^{N} \pi(X_i; \theta) \tag{1}$$

where θ is some unknown parameter. The value of θ that maximizes $L(\theta)$ is called **maximum likelihood estimator (MLE)** and is used to estimate the true value of θ. Note that it is often easier to compute the maximum of $\log L(\theta)$.

Suppose we have a coin which falls heads up with probability p. Let X_i represent the outcome of the i^{th} flip ($x = 1$ for heads and $x = 0$ for tails). X has a Bernoulli distribution with PDF

$$\pi(x; p) = p^x (1-p)^{1-x} \quad \text{for} \quad x = 0, 1. \tag{2}$$

(a) Estimate p using the MLE.

(b) Find a 95% confidence interval for p.

(c) Let $\tau = e^p$. Find the MLE for τ.
HINT: Use one of the properties of the MLE.

2. **Bootstrap.**

Reminder: The bootstrap is a method for estimating standard errors and computing confidence intervals. The essential idea is to use the empirical distribution to estimate the real distribution of the sample. One can then perform simulations and extract information about errors and confidence intervals by taking numerical averages over the bootstrap realizations.
Given a sample of n data, since the empirical distribution puts weight $1/n$ at each data point, a bootstrap realization can be seen as the result of extracting n points randomly from the original sample. Notice that the same data point can appear more than once in each realization.

We draw a sample of $N = 2^n$ numbers with unknown distribution

$$x_1, x_2, \dots, x_N \tag{3}$$

We wonder which is the minimal value that the random variable can assume and would like to use bootstrap to compute the variance of the minimum. Unfortunately, we left our computer at home, so we can not approximate the variance using simulation. Maybe this is not a big issue: to our great surprise, we note that the numbers have the form 2^{kj}, where $j = 1, 2, \dots, n+1$

$$x_i \in \{2^{kj}\}_{j \in \{1, \dots, n+1\}}, \tag{4}$$

and have multiplicity $m(j) = 2^{n-j}$ for $j \leq n$ and 1 for $j = n+1$.

(a) Determine the result that the simulation would have approached.

(b) Approximate the expression assuming N large (if N is large, also n is large).

(c) Compute the bootstrap $1-\alpha$ confidence interval. With which confidence level can we state that the minimal value is 2^k?

3. **Bayesian inference.**

Reminder: The Bayesian approach is based on three postulates: i) Probability describes degree of belief, not limiting frequencies. ii) We can make probability statements about parameters, even though they are fixed constants. iii) We make inferences about a parameter by producing a probability distribution for the parameter.

In ideal gases of non-relativistic particles the speed v is described by the Maxwell-Boltzmann distribution:

$$\pi_{\mathrm{MB}}(v|m,kT) = \left(\frac{m}{2\pi kT}\right)^{3/2} 4\pi v^2 e^{-\frac{mv^2}{2kT}} \,. \tag{5}$$

We would like to infer the mass of the particles from a small sample $\{v\}$ consisting of n measuraments of the velocities, taken at a given temperature kT.

(a) Construct a prior $\pi_{prior}(m|\pi_{\mathrm{MB}}, kT)$ that encodes our knowledge of the Maxwell-Boltzmann distribution and of the temperature; try to construct a prior that is invariant under reparametrizations, that is to say a prior independent of the particular functions of v that have been measured in the experiment.

Hint: Consider a change of scale.

(b) Estimate the mean and the variance of the mass using Bayesian inference.