Exercises for
Computational Statistics

Valerio Marra and Luca Amendola, ITP, Heidelberg
16/09/2013
www.thphys.uni-heidelberg.de/~amendola/teaching.html

ASSIGNMENT 1

You are supposed to verify numerically the so-called "birthday paradox", see below. To be more specific, you should verify the theoretical prediction (if it is a good approximation) as a function of the number N of people. The results should be shown with plots.

The birthday "paradox"

Let us estimate the probability that in N random people there are at least two with the same birthday. A person B has the same birthday of person A only once in 365. Then $P(coinc., N = 2) = 1/365$ and the probability of non-coinc. is e$P(non - coinc., N = 2) = 1 - 1/365 = 364/365$. Let's add a third person. His/her birthday will not coincide with the other two 363 times over 365. The joint probability that the 3 birthdays does *not* coincide is then

$$P(non - coinc., N = 3) = \frac{364}{365}\frac{363}{365} \tag{1}$$

It is clear then that for N persons we have

$$P(non - coinc., N) = \frac{365}{365}\frac{364}{365}\frac{363}{365}...\frac{365 - N + 1}{365} \tag{2}$$

We can now use

$$e^{-x} \approx 1 - x \tag{3}$$

to write

$$\frac{365 - N + 1}{365} = 1 - \frac{N - 1}{365} \approx e^{-(N-1)/365}$$

and therefore

$$P(non - coinc, N) = e^{-1/365}e^{-2/365}e^{-3/365}...e^{-(N-1)/365} = e^{-\frac{N(N-1)}{2}\frac{1}{365}} \tag{4}$$

Finally, the probability of having at least one concidence must be the complement to unity to this, i.e.

$$P(coinc, N) = 1 - e^{-\frac{N(N-1)}{2}\frac{1}{365}} \approx 1 - e^{-\frac{N^2}{730}} \tag{5}$$

For $N = 20$ one has, perhaps surprisingly (this is the "paradox") $P(N) = 0.5$ i.e. almost 50%.

ASSIGNMENT 2

1. Find the mean and covariance matrix of the dataset "bivariate-measurements.txt" available in the dropbox folder "data". You are not supposed to use the built-in function to compute the covariance matrix but to write your own function from scratch. Compute also the correlation matrix.

2. Plot the dataset. Using the covariance matrix previously estimated, draw the major axis of the ellipse formed by the points.

3. Bonus question. Consider the ellipse of point (2), of which you have found the eigenvectors. The eigenvectors are normalized to unity and the eigenvalues tell you how long is the corresponding axis. If you stretch this basic ellipse by the factor 1.5152, you obtain the 68.27% confidence-level contour. Your assignment is then to show that the number of elements of "bivariate-measurements.txt" within this ellipse are indeed approximately 68.27% of the total.

ASSIGNMENT 3

The goal is to perform a maximum likelihood analysis on a set of supernova data (see Dropbox/Tutor-r/lectures/data/supernovae.csv). At the end, we want the position of the maximum of the likelihood, which is defined in eq. (3.16) in the notes. This gives us the most likely values of the matter density of the Universe, Ω_M, and the dark energy density, Ω_L.

First, define the following functions in R:

1. The comoving distance:

$$d_C(z, \Omega_M, \Omega_L) = \int_0^z \frac{dz'}{\sqrt{\Omega_M(1+z')^3 + \Omega_L}}$$

2. The luminosity distance:
If $\Omega_k = 0$: $d_L(z, \Omega_M, \Omega_L) = (1+z) d_C(z, \Omega_M, \Omega_L)$
If $\Omega_k > 0$: $d_L(z, \Omega_M, \Omega_L) = (1+z) \sinh(\sqrt{\Omega_k} d_C(z, \Omega_M, \Omega_L))/\sqrt{\Omega_k}$
If $\Omega_k < 0$: $d_L(z, \Omega_M, \Omega_L) = (1+z) \sin(\sqrt{-\Omega_k} d_C(z, \Omega_M, \Omega_L))/\sqrt{-\Omega_k}$
Here, $\Omega_k = 1 - \Omega_M - \Omega_L$.
3. The distance modules as predicted by the theory:

$$\mu(z, \Omega_M, \Omega_L) = 5 \log_{10} d_L(z, \Omega_M, \Omega_L)$$

4. The auxiliary function S_n:

$$S_n(\Omega_M, \Omega_L) = \sum_i \frac{(m_i - \mu(z_i))^n}{\sigma_i^2}$$

Here, m_i is the distance modules of the i-th super nova at redshift z_i.

5. The log-Likelihood:

$$L(\Omega_M, \Omega_L) = -\frac{1}{2}(S_2 - \frac{S_1^2}{S_0})$$

Next, find the position of the maximum for L in two cases: The flat case, where the curvature of the Universe vanishes, i.e. $\Omega_k = 0$, and the general case. In either case, both parameters should be between 0 and 1. Create a 2D grid of $N \times N$points on the domain $[0, 1] \times [0, 1]$ and evaluate L on each point. Then find where the maximum is. Do the same for the flat case on the interval $[0, 1]$. Attention: In the general case, the computation time goes as N^2 and can be quite long (roughly an hour for $N = 100$). Choose small N's or use a small subset of the entire supernova catalog to test your code. Your result should be close to $(\Omega_M, \Omega_L) = (0.3, 0.7)$. Then you can do a more precise run. If you want, you can afterwards compare the performance with finding the maximum by using the R-package "maxLik".

Hints:

- If an R function expects a scalar, you can't pass a vector in its place and think R will apply the function to each element. Use *sapply* instead.
- You can't naively compare floating point numbers. To check if a number is zero, check if it is very small instead, i.e. if $abs(x) < 1e^{-12}$ or so. -
- *sapply*() or *integrate*() expect a function as an argument, that only takes one argument. Define a new function within the function you are calling *sapply* or *integrate* from to get rid of extra parameters.

ASSIGNMENT 4

Using the posterior $L(\Omega_m)$ (flat case) that you coded before (third assignment), find the 95% confidence-level range for Ω_m, i.e. $\Omega_m^{min} \leq \Omega_m \leq \Omega_m^{max}$. You should find $\Omega_m^{min/max}$ by trial and error such that

$$L(\Omega_m^{min}) = L(\Omega_m^{max}) \int_{\Omega_m^{min}}^{\Omega_m^{max}} L(\Omega_m) d\Omega_m = .95 \int_0^1 L(\Omega_m) d\Omega_m$$

ASSIGNMENT 5

1. Load the dataset "constancy-measurements.txt" (in the dropbox) and plot its error plot (scatter plot plus error bars). Find and plot its linear-model fit (the R function is "lm"). Calculate the residuals.

2. It seems as there is a trend in the data, i.e. the measurements are not constant. Calculate the p-value (5.6 of lecture notes) so as to determine if the null hypothesis that the measurements

are constant is rejected at 99% confidence level. The alternative hypothesis is that the measurements are not constant and so the observed variance is larger than what should be given the measurements errors. Use the chi2 statistics to calculate the p-value: the chi2 distribution is indeed the distribution of the variance modulo some factors.

3. Repeat the test of (2) using the residuals as data.

ASSIGNMENT 6

1. Load the dataset "goals-measurements.txt" (in the dropbox). It is about the total number of goals scored per game in four seasons of World Cup soccer matches (years 1990, 1994, 1998, and 2002). Plot the number of games relative to the total number of goals scored.

2. The distribution seems to follow approximately the Poisson distribution. Check for example that mean and variance are approximately the same.

3. Based on the information provided, is there reason to believe at the 95% confidence level that the number of goals is not a Poisson random variable? Use the Pearson chi2 goodness-of-fit test. The chi2 is obtained using the following formula:

$$\sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

where k is the number of bins, n_i is the observed frequency, n is the total number of games and p_i is the predicted probability from the Poisson distribution.

4. Plot the Poisson distribution on top of the plot of point (1)

ASSIGNMENT 7

Fisher vs. full likelihood.

SN likelihood

The dimensionless Hubble function is:

$$E^2(z) \equiv \frac{H^2(z)}{H_0^2} = \Omega_M (1+z)^3 + \Omega_L \,, \tag{6}$$

where $\Omega_L = 1 - \Omega_M$, that is we assume here flatness. The luminosity distance and distance modulus are then:

$$d_L(z, \Omega_M) = (1+z) \int_0^z \frac{dz'}{E(z)} \quad \text{and} \quad m_t(z, \Omega_M) = 5 \log_{10} d_L(z, \Omega_M) \,. \tag{7}$$

The helper function S_n is:

$$S_n(\Omega_M) = \sum_i \frac{[m_t(z_i) - m_i]^n}{\sigma_i^2} \,, \tag{8}$$

where m_i and σ_i are the distance modulus and error of the i-th supernova at redshift z_i. We can then build the likelihood:

$$\ln L(\Omega_M) = -\frac{1}{2}\chi^2 \quad \text{where} \quad \chi^2 = S_2 - \frac{S_1^2}{S_0}\,. \tag{9}$$

Fisher approximation

The likelihood L can be approximated with the following likelihood

$$L \approx L_f \equiv \exp\left[-\frac{1}{2}(\theta_\alpha - \hat{\theta}_\alpha)F_{\alpha\beta}(\theta_\beta - \hat{\theta}_\beta)\right]\,, \tag{10}$$

where $\boldsymbol{\theta}$ is the vector of theoretical parameters and the Fisher matrix is defined as

$$F_{\alpha\beta} \equiv -\left.\frac{\partial^2 \ln L(\boldsymbol{\theta})}{\partial\theta_\alpha \partial\theta_\beta}\right|_{\rm ML}\,. \tag{11}$$

In our case $\theta = \Omega_M$, and the previous expressions simplify to:

$$\ln L_f \approx -\frac{1}{2}(\Omega_M - \Omega_M^{\rm b.f.})^2 F\,, \tag{12}$$

and

$$F = -\left.\frac{\partial^2 \ln L(\Omega_M)}{\partial {\Omega_M}^2}\right|_{\Omega_M^{\rm b.f.}}\,. \tag{13}$$

Assignment

1. Find F and plot/compare L versus L_f. Calculate the needed derivatives analytically, or numerically if you find it too complicated. Use 1 entry every 15 of the dataset "supernovae.csv" available in the dropbox.

2. Compute the 95% c.l. interval for Ω_M using L_f and L. Compare the two constraints. What is the origin of the difference?

3. [bonus question] What is the meaning of F in this very simple case? Are there other ways to compute a statistical quantity similar to F?

ASSIGNMENT 8

1. Choose a distribution (e.g. normal, chi2,..) and draw from it a dataset of N elements. Save this dataset, it will be your "data".

2. Estimate by bootstrapping the error on the mean and variance of this dataset. Call M the number of bootstrap iterations.

3. Compare the bootstrapping error to the actual error which you should find in a Monte Carlo fashion, that is by generating many other datasets from the fiducial distribution you have chosen. When the normal distribution is used, the agreement should be very good.

Write your code in a flexible way, so that it can be applied to an arbitrary distribution and arbitrary N and M.