Physical Mathematics 2010: Problems 4 (week 10)

Probability, Statistics, Fitting data

1. **Basic probability (these may be skipped if you wish)**

   a) A worried engineer has a prototype computer system in his laboratory. Due to a flaw in the board design the first pass system containing 32 nodes will only boot 90% of the time.
   What is the probability that a single node will boot?
   What is the probability that a 4096 node system will boot?
   What is the probability that a 65536 node system will boot?

   b) Suppose 16 gigabyte dram displays a correctable (single bit) memory error once every week.
   How often do you expect an uncorrectable error (double bit error in the same 64bit word)?
   How often would this happen for a 65536 node computer?

2. **Probability densities**

   a) Define the probability density for a random variable X.

   b) Sketch a zero mean unit variance Gaussian distribution.

   c) Sketch a zero mean Gaussian distribution with variance $\sigma = 2$.

   d) Sketch a Gaussian distribution with width $\sigma = 2$ and mean 3.
   What is the equation for this distribution.

3. **Combining distributions**

   a) If X and Y are random variables with distributions $P_X(x)$ and $P_Y(Y)$ show that the distribution P_{X+Y} of the variable $X + Y$ obtained by adding these is the convolution of P_X with P_Y.

   b) Determine the distribution P_{CX} of CX where C is a constant.

4. **Addition of Gaussian random variables**

   a) Determine the Fourier transform of a Gaussian (Normal) distribution

   $$P_{\mathcal{N}}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$$

   b) Use the convolution theorem to determine the distribution of the sum of two Gaussian random variables of widths σ_1 and σ_2.

5. **Central limit theorem**

   a) State the central limit theorem.

   b) Consider the sum of N random variables $X_1 \ldots X_N$ where these are independent and identically distributed with an arbitrary probability distribution $P_X(x)$. Use Question 2a) to represent the distribution of sum of these variables as a convolution.

   c) Apply the convolution theorem (multiple times – a pattern should appear) to represent this distribution in terms of the Fourier transform $\tilde{P}_X(k)$ of $P_X(x)$.

d) Apply the scaling rule of Q2b) to show that the average of the random variables is

$$P_{S_N}(u)\frac{1}{\sqrt{2\pi}}(\sqrt{2\pi})^{N-1}\int_{-\infty}^{\infty} dk' e^{-ik'u}\left(\tilde{P}_X(\frac{k'}{N})\right)^N dk'$$

e) Write $\tilde{P}_X(\frac{k'}{N})$ as the Fourier transform of $P_X(x)$ and Taylor expand this to second order in $\frac{k'}{N}$.

f) Use this result to show

$$P_{S_N}(u) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk' e^{-ik'u}\left(1-\frac{k'^2}{2N}\frac{1}{N}\right)^N dk'$$

g) Show that for $b = \frac{-k'^2}{2N}$ that the $N \to \infty$ limit of $(1+b/N)^N$ is e^b, and conclude that

$$P_{S_N}(u) \to \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} dk' e^{-ik'u}\frac{1}{\sqrt{2\pi}}e^{-\frac{k'^2}{2N}} dk'$$

and so

$$P_{S_N}(u) \to \frac{1}{\sqrt{2\pi}\sigma_S}e^{-\frac{u^2}{2\sigma_S^2}}$$

j) Explain the importance of this theorem to statistical analysis.

6. χ^2**/dof**

Show that if $X_1 \dots X_N$ are drawn from the same unit variance normal (Gaussian) distribution $P_{\mathcal{N}}(x)$ then the expectation value for

$$\chi^2 = \sum_i x_i^2 = N.$$

Explain what this implies for χ^2/dof when analysing data.

7. **Computing** χ^2

An experiment is run three times. Each time the measurement error is $\sigma = 1$, and the three data values are $\{x_i\} = \{3, 4, 5\}$. What is the χ^2 for this set of measurements, how many degrees of freedom are there, and is this acceptable?

8. **Bookwork**

- N Gaussian distributed measurements of the same quantity with mean A, with same error σ are each distributed according to

  $$P_i(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-A)^2}{2\sigma^2}}$$

  Define the joint probability distribution $P(x_1, \dots, x_N)$ for all N-measurements, and relate this to the corresponding χ^2.

  By differentiating with respect to A, determine the value of A that maximises this probability density, and also determine the error on A from the $\chi^2 = 1$ rule.

- Repeat the previous question allowing the errors of each data point to be different $\sigma \to \sigma_i$

9. **Maximum likelihood method and parameter errors**

Suppose data is theoretically described by a curve $y = f_{\{p\}}(x)$ where p is a parameter of the function.

a) Suppose the "true" parameter p is known.

If N datapoints y_i are measured for coordinates x_i, and they should be distributed about a true mean $f_{\{p\}}(x_i)$ with a Gaussian width σ_i which can in principle be measured.

Write down the probability distribution for each y_i ?

b) Show that the joint probability distribution for the set of measurements $\{y_i\}$ is

$$P(y_1, \ldots, y_N) = e^{\frac{1}{2}\chi^2}$$

where

$$\chi^2 = \sum_i \frac{[y_i - f_{\{p\}}(x_i)]^2}{\sigma_i^2}$$

c) Find $\frac{\partial}{\partial p}\chi^2$, and $\frac{\partial^2}{\partial p^2}\chi^2$.

d) Suppose $p_{\min}$ is a parameter that gives the minimum of χ^2 satisfying

$$\left.\frac{\partial}{\partial p}\chi^2\right|_{p_{\min}} = 0,$$

and that

$$\left.\chi^2\right|_{p_{\min}} = \chi^2_{\min}$$

If we now interpret $Ne^{-\frac{1}{2}\chi^2(p)}$ as being a probability for the fit parameter p, where N is some normalisation, show that Taylor expansion to second order in p around the minimum of χ^2 suggests this is Gaussian with width

$$\sigma_p^2 = \frac{2}{\left.\frac{d^2}{dp^2}\chi^2\right|_{p_{\min}}}.$$