Assignment II.

probability concepts, hypothesis testing, chi square fitting

due: November 5, 2017 (12 days)

problem 1 PHYS 139/239

(a) show empirically the convergence to the central limit theorem in dice throwing simulations

(b) compare the result with your analytic expectation

(c) If you flip a fair coin one billion times, what is the probability that the number of heads is between 500010000 and 500020000, inclusive? (Give answer to 4 significant figures.)

problem 2 PHYS 139/239

A good statistician has data on how a particular jailer is making their decisions in choosing between prisoners B and C answering the question from prisoner A under repeated circumstances. Based on the data and all observations the statistician knows that the jailer operates in some hardwired way with his preferred x-value and based on that x-value he makes a binary decision choosing between prisoners B and C. However, the statistician has to know the prior probability distribution $p(x|I)$ from all jailers of the universe how they are individually wired for some particular x in making their binary decisions for their x-value. So observing a jailer repeatedly may reveal his x-value to some limited degree but it is folded with the choice x from the probability distribution $P(x|I)$ a particular jailer of the universe is wired for. If we observe another jailer in repeated decision making, he may have a different x value.

Now the statistician knows that the jailer observed chose B $N_B = 13$ times out of $N=37$ observations of the jailer and C is chosen 24 times. Assuming constant prior distribution $P(x|I)$ now the statistician can calculate the posterior distribution $P(x|data)$ of x values based on the observations (N and N_B data).

(1) Calculate this normalized distribution analytically.

(2) Plot it with proper normalization.

(3) Discuss what you expect in the N —> inifinity limit at fixed N_B/N.

problem 3   PHYS 239   bonus for PHYS 139

Repeat Problem 2 when the prior distribution is not constant but $P(x|I) \propto$ x^10(1-x)^9.

problem 4   PHYS 139/239

(a)   prove the additivity of the semi-invariant I_4 analytically and in simulation

(b)   PHYS 239 only show the additivity of I_6 analytically and in simulation to reasonable accuracy for some distribution of your choosing.
Show that $I_6 = 0$ for the normal distribution.

definition of the k_{th} centered moment M_k of a distribution:

$$M_k \equiv \left\langle (x_i - \overline{x})^k \right\rangle$$

following this definition M_2 is the variance of the distribution

Mean and variance are additive over independent random variables:

$$\overline{(x+y)} = \overline{x} + \overline{y} \qquad \mathrm{Var}(x+y) = \mathrm{Var}(x) + \mathrm{Var}(x)$$

note “bar” notation, equivalent to < >

Certain combinations of higher moments are also additive. These are called semi-invariants.

$$I_2 = M_2 \qquad I_3 = M_3 \qquad I_4 = M_4 - 3M_2^2$$

$$I_5 = M_5 - 10M_2M_3 \qquad I_6 = M_6 - 15M_2M_4 - 10M_3^2 + 30M_2^3$$

Skew and kurtosis are dimensionless combinations of semi-invariants

$$\mathrm{Skew}(x) = I_3/I_2^{3/2} \qquad \mathrm{Kurt}(x) = I_4/I_2^2$$

A Gaussian has all of its semi-invariants higher than I_2 equal to zero.
A Poisson distribution has all of its semi-invariants equal to its mean.

problem 5   PHYS 139/239

calculate numerically the t-values and p-values in the table

Let's dispose of the silly (all p's = 0.25):

The test statistic: the value of the observed count under the null hypothesis that it is binomially (or equivalent normally) distributed with p=0.25.

$$\mu = 0.25\,N$$

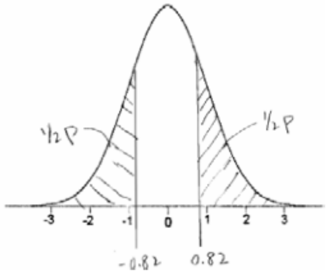
$$\sigma = \sqrt{0.25 \times 0.75\,N}$$

$$t = \frac{n-\mu}{\sigma}$$

$$p = 2[1 - P_{\mathrm{Normal}}(|t|)]$$

t-value = number of standard deviations

p-value = tail probability (here, 2-tailed)



	t-value	p-value
A	174.965	≈ 0
C	–174.715	≈ 0
G	–170.963	≈ 0
T	170.713	≈ 0

The null hypothesis is (totally, infinitely, beyond any possibility of redemption!) ruled out.

problem 6 PHYS 239

bonus for PHYS 139 solutions

explain and calculate numerically the two p-values of the hypothesis and compare with the numbers in the lecture

can you come up with a hypothesis which cannot be killed by the data of the DNA sequence?

The not-silly model: A and T occur with identical probabilities, as do C and G.

The test statistic: Difference between A and T (or C and G) counts under the null hypothesis that they have the same p, which we will estimate in the obvious way (which is actually an MLE).

$$\hat{p}_{AT} = \tfrac{1}{2}(n_A + n_T)/N$$
$$\hat{p}_{CG} = \tfrac{1}{2}(n_C + n_G)/N$$

$$n_A \sim \mathrm{Normal}(N\hat{p}_{AT}, \sqrt{N\hat{p}_{AT}(1-\hat{p}_{AT})})$$
$$n_T \sim \mathrm{Normal}(N\hat{p}_{AT}, \sqrt{N\hat{p}_{AT}(1-\hat{p}_{AT})})$$
$$\Rightarrow n_A - n_T \sim \mathrm{Normal}(0, \sqrt{2N\hat{p}_{AT}(1-\hat{p}_{AT})})$$

the difference of two Normals is itself Normal

the variance of the sum (or difference) is the sum of the variances

problem 7 PHYS 139/239

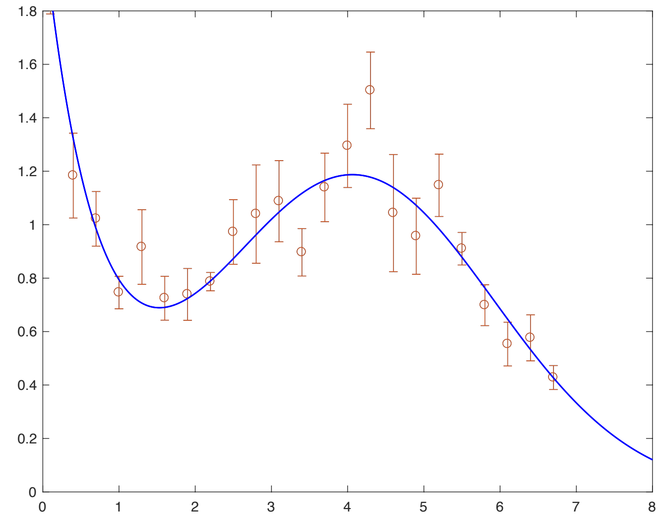
We measure in an experiment at 23 values of x_i the outcome y_i from normal distributions where the results are listed in the data.txt file:

x	y	y error
0.100000000000000	1.955692474636036	0.166896282383792
0.400000000000000	1.183586547503424	0.158551780782385
0.700000000000000	1.022128862295741	0.102145122199102
1.000000000000000	0.746134082944572	0.060820536337125
1.300000000000000	0.916188421395087	0.139506053368529
1.600000000000000	0.724682156536752	0.081793212333357
1.900000000000000	0.739127499035786	0.096894347069944
2.200000000000000	0.786742524422711	0.034353661707974
2.500000000000000	0.972558512530457	0.121213729440151
2.800000000000000	1.039776955766267	0.183845107945299
3.100000000000000	1.087705062846587	0.152064123528747
3.400000000000000	0.896727858969629	0.088835443972525
3.699999999999999	1.139381591276074	0.128022842446142
4.000000000000000	1.294829163615035	0.155588445889791
4.299999999999999	1.502261299770580	0.143493932937373
4.600000000000000	1.043529911555928	0.219186627495748
4.899999999999999	0.956827376670183	0.142469078945670
5.199999999999999	1.147387265711086	0.116683264235504
5.499999999999999	0.909994065501967	0.060876724854546
5.799999999999999	0.698671186235582	0.076323301379691
6.100000000000000	0.553227945238010	0.082132016628130
6.399999999999999	0.576371045690540	0.085922021448737
6.699999999999999	0.427880507687987	0.044877728959367

problem 7 PHYS 139/239

(A) assuming that the experiment is described by the theoretical function f(x) of five parameters, calculate the mean value of $b_3 b_5$ and calculate the error from linear error propagation.

$$f(x) = b_1 \exp(-b_2 x) + b_3 \exp\left(-\frac{1}{2}\frac{(x-b_4)^2}{b_5^2}\right)$$



(B) Calculate the mean value of $b_3 b_5$ and calculate the error from the posterior distribution of $b_3 b_5$