Assignment II.

probability concepts and hypothesis testing

due: October 30, 2019 (14 days)

problem 1 PHYS 139/239

As an example of a statistical test, an experiment is performed to determine whether a coin flip is fair (equal chance of landing heads or tails) or unfairly biased (one outcome being more likely than the other). Suppose that the experimental results show the coin turning up heads 16 times out of 21 total flips. The null hypothesis is that the coin is fair with probability p(head) = 0.5, and the test statistic is the number of heads. If a right-tailed test is considered, the p-value of this result is the chance of a fair coin landing on heads at least 16 times out of 21 flips. The α value of 0.05 (biology) is set as the level of significance. Is the experiment statistically significant to eliminate, based on the right-tailed test, the null hypotheses that the coin is unbiased?

problem 2 PHYS 139/239

(a) If you flip a fair coin one billion times, what is the probability that the number of heads is between 500,100,000 and 500,200,000, inclusive? (Give answer to 4 significant figures.)

(b) How close are the estimates from the normal distribution, or from the Bernoulli binomial distribution which is exact.

problem 3 PHYS 139/239

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

(b) compare the results with your analytic expectation based on the probabilities of unbiased dice outcomes.

problem 4 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 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(xII) 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(xl I) now the statistician can calculate the posterior distribution P(xldata) 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 5 PHYS 239

Repeat Problem 4 when the prior distribution of jailer is not constant but $P(x|\ I\ ) \propto x^9 \cdot (1-x)^{10}$.

problem 6 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 7 PHYS 139/239

We will investigate the DNA sequence of a damaged chromosome linked to Assignment II as damaged DNAdata.txt (see Lecture 5 for introduction to the problem). The goal is to build probability models for the damaged chromosome from four nucleobases ACGT and subject the models to two different null hypotheses discussed in Lecture 5.

(a) determine the number of separate A, C, G, T occurrences.

(b) calculate numerically the 2-tailed t-values and p-values based on the two different hypotheses.