Swarthmore College Honors Exam 2011

Probability and Statistics

Robin Pemantle, University of Pennsylvania

1. Suppose that $X_1, X_2, \ldots$ are an infinite collection of random variables defined on some probability space. Complete the following definitions. Compact mathematical notation is preferred.

   (a) $X_n \to Y$ in probability if and only if ...

   (b) $X_n \to Y$ in distribution if and only if ...

   (c) $\{X_n\}$ are independent if and only if ...

   (d) X_1 and X_2 are bivariate normal if and only if ...

2. Let σ be a random uniformly chosen permutation of the numbers from 1 to n. We write this in cycle notation, so that for example if $n = 5$, the notation $(142)(35)$ stands for the permutation mapping 1 to 4, 4 to 2, 2 to 1, 3 to 5 and 5 to 3.

   (a) What is the probability that the element 1 is contained in a cycle of length precisely k?

   (b) What is the expected number of cycles in σ of length precisely k?

   (c) What is the expected total number of cycles in σ?

3. Suppose X, Y and Z are any three random variables defined on some probability space. What is the greatest possible value of

$$\min\{\mathbf{P}(X > Y), \mathbf{P}(Y > Z), \mathbf{P}(Z > X)\}\,?$$

4. If X and Y are IID uniform on $[0, 1]$, what is the density of the random variable $X - Y$?

5. Suppose $X_1, X_2, \ldots$ are IID with mean 3 and variance 5.

   (a) Give a precise statement of the implication of the Central Limit Theorem when applied to the variables $\{X_n\}$.

   (b) What upper bound does the CLT guarantee for $\mathbf{P}(\sum_{n=1}^{100} X_n \geq 360)$?

   (c) Compute the best upper bound you can for $\mathbf{P}(\sum_{n=1}^{100} X_n \geq 360)$.

6. A purportedly random sample of body weights of 196 residents of Delaware County is recorded in a ledger as $w_1, \ldots, w_{196}$. Based on this, it is asserted that the average body weight of a resident of Delaware County is within 4.4 pounds of 143 pounds, with 95% confidence.

   (a) Explain precisely (but succinctly) what probability statement is being asserted in the above statement of a confidence interval.

   (b) What can you infer was true about the sample average and sample variance of the values $\{w_1, \ldots, w_{196}\}$ given that they resulted in the 95% confidence interval 143 ± 4.4?

7. In a certain chain reaction, occuring in discrete time, at each step a particle will produce either 0, 2 or 3 new particles, with equal probabilities independently of all the other particles. Beginning with one particle, what is the probability that this process will die out?

8. Let Z be a Poisson random variable of mean ν and let $X_1, X_2, \ldots$ be a sequence of exponential random variables of mean 1, independent of each other and of Z. Let Y denote the random sum defined by

$$Y := \sum_{j=1}^{Z} X_j \,.$$

(a) Compute the moment generating function

$$g(\lambda) := \mathbf{E}e^{tY} \,.$$

(b) Use this and Markov's inequality to obtain an upper bound on

$$\mathbf{P}(Y \geq a) \,.$$

(c) Suppose customers arrive according to a Poisson process with rate 1. Let X'_j be the time between the $(j-1)^{st}$ and j^{th} arrivals. Let Z' be the number of arrivals by time ν. Let $Y' = \sum_{j=1}^{Z'} X'_j$. Give YES/NO answers with at most one sentence of justification:

i. Are $\{X'_j\}$ independent mean 1 exponentials?

ii. Is Z' a Poisson of mean ν?

iii. Is Y' distributed the same as Y?

9. (a) Let (X, Y) be the values of a Brownian motion at times 1/3 and 2/3. What is the joint distribution of the pair (X, Y)? Please state your answer in terms of the bivariate normal distribution.

(b) Let (Z, W) be the values of a Brownian *Bridge* at times 1/3 and 2/3. What is the joint distribution of the pair (Z, W)? Please state your answer in terms of the bivariate normal distribution.

(c) Suppose you are observing a Brownian motion or a Brownian Bridge but you don't know which. The observed values at times 1/3 and 2/3 respectively are -1 and 1/2. Does this evidence favor the Brownian motion or the Bridge?

10. This problem concerns an F-test. In case you are not familiar with the details, I have provided a detailed description of the test; this is for your reference and may be more information than you need.

    128 subject are randomly assigned into 8 groups of 16 each. Each group eats a different kind of energy snack and then takes a test. The scientist in charge notices that some snacks seem more helpful than others. In particular, subjects who got apples, power bars, candy or Red Bull did better (average score of 1.17) than subjects who got fruit rolls, ginseng tea, diet coke or nothing (average score of 0.79). The scientist decides to do an F-test with two groups ($I = 2$) of 64 subjects each ($J = 64$), one group being the apple/bar/candy/drink group and the other consisting of the remaining 64 subjects.

    First, the scientist of computes the sum $\sum_{i=1}^{64}(X_i - 1.17)^2$ where X_i is the test score of subject i and the subjects are renumbered so that the first 64 are the ones who got apple, bars, candy or drink. The value obtained is 56.4. Similarly, the sum of $(X_i - 0.79)^2$ for $65 \leq i \leq 128$ (the other group of 64 subjects) is 73.6. The scientist sets

    $$SS_W = 56.4 + 73.6 = 130.0$$

    and sets

    $$SS_B = 64[(1.17 - 0.98)^2 + (0.79 - 0.98)^2] = 4.44\,.$$

    The F-ratio is computed as

    $$F = \frac{SS_B/(I-1)}{SS_W/2(J-1)} = \frac{4.44/1}{130/126} = 4.30\,.$$

    The 95th percentile for the F-statistic with 1 and 126 degrees of freedom is a little under 3.92 (the percentile with 120 degrees of freedom for the denominator; see Appendix B, Table 5, page A11 of Rice). The scientist concludes that with 95% confidence, the effect of a good energy snack (apples, bars, candy, drink) was greater than the effect of a bad snack. [*go to next page*]

Explain: was the finding convincing, that one group of snacks has a differentially higher causal effect on test scores than the other group does? I have provided a "path to inference", in which a number of possible weak links are listed. You may, but do not have to, choose your answer as one of these six steps. In any case, an answer of three to five sentences will suffice.

(a) Not convincing because there is no well defined statistical model in which this test takes place

(b) Not convincing because there is a well defined model but the assumptions of the model are not met

(c) Not convincing because there is a model and the assumptions are met but the computations were not correctly carried out

(d) Not convincing because there is a model, assumptions were met, computations were correct, but the findings were misinterpreted

(e) Not convincing because there is a model, assumptions were met, computations were correct, findings were interpreted correctly, but spurious or reverse causation was not ruled out

(f) Convincing