

You are advised to do all homework problems using R whenever possible and then check the results with your own calculations.

STATISTICS

Due November 1, 2011

Homework 3

1. In ecological studies the ecological entropy (or Shannon index) is defined as a measure of biodiversity. Assume that within a system of unlimited individuals, there exists n species of organisms (A_1, A_2, \dots, A_n), and the proportion of individuals found in the j -th species (A_j) is p_j , ($j=1, 2, \dots, n$). The Shannon index is then defined as

$$H = -\sum_{j=1}^n p_j \ln p_j$$

Suppose an ecological system has 8 different species of organisms. The numbers of organisms belonging to individual species are shown in HW3DATA.txt.

- (i) Calculate the Shannon index of the ecological system using the data corresponding to your own class ID.
 - (ii) It can be shown that the Shannon index has its maximum when all species have equal number of organisms, i.e. an organism is equally likely belonging to any species. Calculate the maximum H for ecosystems with the number of species $n = 5, 6, 7, \dots, 15$, and establish a plot of the values of maximum H versus the number of species using R.
2. Use the exemplar R code (for discrete uniform random number generation) shown in class for the following problems:
- (i) Generate a set of 500 random numbers ($n=500$) from a discrete uniform distribution from 21 to 30 (at increment of 1). Calculate the mean and standard deviation of the data set and compare the results against the corresponding theoretical values.
 - (ii) Sketch the histogram of the generated data set using the *hist* command with breaks in increment of 2.
3. Repeat the work of Problem 2 except using $n=50$. Make your comments on the results of Problems 2 and 3.
4. A random experiment whose outcome is random and can be either of two possible outcomes (a success or a failure) is known as the Bernoulli experiment (or the Bernoulli trial).
- (i) A Bernoulli experiment is regularly conducted once in everyday (or say, every 24-hour interval). In a two-year period, the random experiment resulted in a total of 15 successes. Estimate the probability of having more than 3 successes in a period of four months (assuming there are 30 days in each month).
 - (ii) A Bernoulli experiment is regularly conducted once in every month (assuming there are 30 days in a month). In a two-year period, the random

experiment resulted in a total of 15 successes. Estimate the probability of having more than 3 successes in a period of four months.

5. Given a random variable X , calculate $P(|X - \mu_X| \geq 2\sigma_X)$,
 - (i) if X is normally distributed.
 - (ii) if X has a standard Pearson type III distribution (PT3 distribution with zero mean and unit standard deviation). with coefficient of skewness $\gamma = 1.2$.
 - (iii) Compare the results of (i) and (ii) against the results based on the Chebyshev inequality.
6. Let X be a gamma random variable with the following density

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad \alpha, \lambda > 0 \quad \text{and} \quad 0 \leq x < +\infty.$$

A random variable Y is defined as $Y=2\lambda X$. Prove that random variable Y is a random variable of Chi-squared distribution.

7. The travel time from a student's home to NTU campus involves a few factors:
 - (i) Walking to bus stop (stop for traffic lights, crowdedness on the streets, etc.),
 - (ii) Transportation by bus,
 - (iii) Stop by 7-11 or Starbucks for breakfast (long queue), and
 - (iv) Walking to NTU campus.

Let X_i be the time required for completion of the i -th factor. X_1 has a normal distribution with a mean of 15 minutes and a standard deviation of 6 minutes. X_2 has a Gamma distribution with a mean of 30 minutes and a standard deviation of 10 minutes. X_3 is exponentially distributed with a mean of 20 minutes. X_4 has a normal distribution with a mean of 10 minutes and a standard deviation of 5 minutes. Assuming all X_i 's are independent, when should the student leave home in order to achieve a higher than 90% chance of not being late for a class beginning at 9:10 am? [Hint: Generate a random sample with a large sample size (for example, $n=10,000$) for each individual random variables X_i , ($i=1,2,3$, and 4) and then establish a random sample (size 10,000) of the travel time (let it be Y). Then estimate the 0.9 quantile of Y .]

8. Generate 100 random variates (i.e., a random sample of sample size $n = 100$) from each of the following distributions and calculate their sample means and sample variances.
 - (i) Poisson distribution $f_X(x; \lambda) = \frac{e^{-\lambda} \lambda^x}{x!}$, $x = 0,1,2,\dots$ with $\lambda = 5$.
 - (ii) Normal distribution $N(2.0, 4.0)$
 - (iii) X with density function $f_X(x) = 12x^2(1-x)$, $0 < x < 1$.

9. Let X_1 and X_2 be two independent Poisson random variables with parameters λ_1

and λ_2 , respectively. Prove that $Y = X_1 + X_2$ is a Poisson random variable with parameter $\lambda_1 + \lambda_2$.

[Hint: The moment generating function of the sum of *independent* random variables is the product of the moment generating functions of individual random variables.]

10. Suppose that the time span between the occurrences of two consecutive typhoon events (let's consider the time when the Central Weather Bureau categorizes a tropical storm as a typhoon as the time of typhoon occurrence) can be characterized by an exponential distribution. The following table shows the observed time spans (in unit of days) between the occurrences of two consecutive typhoons (this is also known as the *inter-arrival time* in hydrology) over a period of 30 years.

54.22	44.22	9.12	18.46	26.76	8.05	47.90	35.09	5.14	38.25
57.09	6.57	106.64	63.11	49.12	37.75	12.24	14.58	18.81	19.12
11.93	23.10	21.51	3.48	6.06	3.54	8.46	19.66	39.62	55.64
22.71	6.16	8.27	16.21	3.21	36.31	37.01	23.89	5.40	36.73
19.72	2.35	22.99	6.33	5.01	32.73	5.58	8.92	101.67	50.27
18.89	6.52	23.17	166.63	44.09	15.48	4.09	4.38	27.42	68.80
29.82	2.24	0.40	6.02	13.77	89.42	17.88	18.86	23.86	70.88
6.53	5.85	98.01	28.81	9.18	14.19	37.40	10.86	54.74	41.50
36.30	34.35	8.68	36.39	79.13	23.48	102.13	3.22	40.59	36.40
5.41	9.53	0.71	13.90	1.20	27.29	13.18	39.47	4.89	21.75
16.56	5.64	12.27	36.11	23.04	11.83	3.33	53.34	59.07	3.66

- (i) Calculate the mean and standard deviation of the inter-arrival time. Are these two values very close? [Note: For the exponential distribution, the expected value and the standard deviation are numerically the same.]
- (ii) Use the mean value of the interval-arrival time calculated in (i) as the expected value of the inter-arrival time. If a typhoon has just occurred, what is the probability that no typhoons will occur in the next 30 days? [Hint: The cumulative distribution function and probability density function of an exponential random variable are as follows, respectively.]

$$f_X(x) = \lambda e^{-\lambda x}, \quad \lambda > 0, \quad 0 \leq x < +\infty,$$

$$F_X(x) = 1 - e^{-\lambda x}, \quad \lambda > 0, \quad 0 \leq x < +\infty$$