Assignment 4 — Due October 3, 2003

1. This problem will illustrate the Central Limit Theorem and indicate how simulation studies can be conducted with R. Consider the discrete random variable X with the following distribution:

x	1	2	3	4	5	6	7	8
$p(x)$	0.2	0.1	0	0	0	0.1	0.2	0.4

Suppose first that you wish to generate 100 values from this distribution. The following approach would be appropriate for this:

```
> x = c(1, 2, 6, 7, 8)
> px = c(0.2, 0.1, 0.1, 0.2, 0.4)
> draw100 = sample(x, 100, replace=T, prob=px)
```

You will now have 100 realized values from the random variable X saved as the object `draw100`. You can examine the distribution of these values by using the `stem` or `hist` commands.

To demonstrate the CLT, you will generate samples of size 30 and compute $\bar{x}$ for each sample. You will do this 200 times resulting in 200 realized values of $\bar{x}$. These 200 values of $\bar{x}$ can be examined using `stem` or `hist`. Your job is to judge how well the CLT works for this distribution. (Probably `hist` will work better for this.)

Here is the suggested procedure. Create the objects `x` and `px` as before. Then:

```
> draws = sample(x, 30*200, replace=T, prob=px)
> draws = matrix(draws, 30)
> drawmeans = apply(draws, 2, mean)
```

Now the object `drawmeans` contains 200 values of $\bar{x}$, each based on 30 observations. You can examine the values, make displays, and look at summary statistics. (Feel free to experiment with different sample sizes; to use a sample size of 10, replace 30 with 10 in the `sample` command. See R Appendix 5.7.1 at the course website for further information and explanation.)

Your write-up for this problem should include a histogram of the 200 realized $\bar{x}$ values. How well does the CLT work for this distribution? Are you surprised? Present a concise paragraph summarizing your findings.

2. The Ejection Fraction (EF) of the heart is a measure of the efficiency of the heart as a pump — the higher the EF the better. There are two competing methods for measuring EF: one of which is very expensive, but very precise, and the other which is less expensive, but not very precise. In either case, to estimate the EF of a given patient, the methods will be applied several times, and the results will be averaged.

You have a choice between one of the following two alternatives, each of which has roughly the same total cost:

I. Use the precise method ($\sigma = 5$). Take a random sample of size 2.

II. Use the less precise method ($\sigma = 8$). Take a random sample of size 5.

(a) Which alternative would you select to meet the stated objective? (Give your reason).

(b) Suppose that, for a given patient, it was known that their EF is 63. Consider the population of EF measurements that can be taken on that patient. That population follows a normal distribution with $\mu = 63$. Suppose you select alternative II above. Find the probability that the sample mean EF is between 60 and 65.

(c) As an experiment, the investigators decided to take two measurements on the patient, one using method I and one using method II. The results would then be averaged to provide an estimated EF for the patient. What would the standard deviation of that estimate be?

3. A bag of apples is required to weigh 2.5 lb (= 40 ounces). The apples of this variety have a weight which is roughly normally distributed with mean 2.2 ounces and variance 0.2 (ounces^2). The bag I buy has 20 apples in it. What is the probability that its weight will be less than the required weight? (Assume that the weight of the plastic container is negligible.)

4. The probability of of causing remission of cancer in diseased rats for a particular drug is known to be 0.3. Suppose 180 randomly selected diseased rats are each given the particular drug. Let X be the number of rats in which cancer remission occurs.

   (a) Compute $P(X \geq 50)$.

   (b) Define Y to be the proportion of rats with remission: $Y = X/180$. Compute $P(0.25 < Y < 0.40)$. (Hint: In both (a) and (b), use the normal approximation to the binomial distribution.)

5. Suppose we have observations from a $N(\mu, \sigma^2)$ distribution. The following table is similar to that in Assignment 3; however, here you are asked to consider probability statements about S^2 where S^2 is the sample variance from a sample of size n. Using the chi-squared table, fill in the blanks, treating each row as a separate problem.

	μ	σ^2	n	a	$P(S^2 \leq a)$	b	$P(S^2 \geq b)$
(a)	3	25	4	??	.975	??	.975
(b)	2	25	18	??	.975	??	.975
(c)	-30	24	12	46	??	58	??
(d)	30	??	25	4	.05	??	.99

   (e) For each row of the table, did you make use of all the information given? If some unnecessary information was given, indicate what it was.

   Using R (see for instance R Appendix 5.7.3 at the course website), fill in the blanks in the following:

	μ	σ^2	n	a	$P(S^2 \leq a)$	b	$P(S^2 \geq b)$
(f)	9	6	18	??	.95	??	.75
(g)	70	36	12	31	??	19	??

6. I have 2 cats, Ray and Felix, who behave independently of one another. I know that at 4:00 pm there is a 0.8 chance that Ray will be in my garage and a 0.6 chance that Felix will be in my garage.

   (a) If I walk into my garage some day at 4:00 pm, what is the probability that at most one cat will be there?

   (b) Consider the event "Ray is in the garage" and the event "Felix is not in the garage." Are these events mutually exclusive?

   (c) Consider the random variable X which represents the number of cats I find in my garage at 4:00 pm. Determine the distribution of X (Hint: Start by writing out the sample space.)

   (d) What is the mean number of cats in my garage at 4:00 pm?

Readings: Week 4: Course Notes: Chapter 5