Solutions to Assignment #3

1. Use randomization to test the null hypothesis that "there is no difference between calcium supplementation and a placebo" for the experimental data provided in the MS Excel file `CalciumBloodPressureData.xls`, which may be downloaded from `http://pages.pomona.edu/~ajr04747`.

   Describe the procedure that you followed in R to do the simulations and how you estimated the p–value.

   ***Solution**:* Looking at the data we see that the average decrease is systolic blood pressure for the Calcium group was 5, while that for the placebo group was about -0.64 (so, on average, the systolic blood pressure for subjects in the placebo group actually increased). This suggests that Calcium supplementation might have some effect on lowering blood pressure. To see whether this difference in the average decrease in systolic blood pressure is statistically significant, we use randomization to test the the null hypothesis:

   H_o: There is no difference between the two treatments.

   The test statistic in this case is the mean decrease in the systolic blood pressure in a random sample of size 10 picked from the `dec` column in the data. The p-value is the the probability that the mean of randomly select sample is 5 or above.

   Using R to perform $10,000$ repetitions of the sampling we obtain an estimate for the p–value to be about 0.0522. At this point, since our approximation to the p–value is close to the threshold of 5%, we could reject the null hypothesis and conclude that the experiment's result provide statistically significant evidence to conclude that the calcium supplementation reduces blood pressure.

   The answer to this question will depend on the results of the simulation. If the approximate p–value turns out to be around 0.0549 or higher, we might be led to conclude that the experimental findings support the claim that calcium supplementation reduces blood pressure; however, the evidence falls short of the traditional 5% or 1% thresholds. □

   ***Solution**:*[**Alternate Solution**]

   There are several ways to approach the data. For instance, we might decide to focus on whether the treatment actually decreased blood pressure or didn't. In this case, the test statistic would be the proportion of subjects in the calcium group that experienced a decrease in blood pressure: 6 out of 10 subjects in the calcium group, as opposed to 3 out of 11 in the placebo group. Is the difference in proportions statistically significant? Performing an analysis like the one we did in Activity #2 (*Comparing two treatments*), we get a p–value given by the hypergeometric distribution. In this case, the p–value is the probability of picking 6 black cards, or more, at random from a deck of 21 cards containing 9 black cards and 12 red cards. Using the `dhyper()` function in R and the code

```
pVal <- 0
for (i in 6:10) pVal <- pVal + dhyper(i,9,12,10)
```

we obtain the p–value of 0.1420576, or about 14.2%. Thus, in this case we clearly conclude that the evidence provided by the experiment is not statistically significant. □

2. (Spam Topics[1]) A majority of e-mail messages are now "spam." The distribution of topics, according to an article by Robyn Greenspan found on the internet at `http://www.clickz.com/showPage.html?page=3295851`, is given in the following table.

Topic	Adult	Financial	Health	Leisure	Products	Scams
Probability	0.145	0.162	0.073	0.078	0.210	0.142

Choose a spam e-mail message at random.

(a) What is the probability that the selected e-mail does not deal with any of the topics listed in the table?

Solution: Assume that each piece of spam falls into exactly one of the given categories.
The given probabilities have sum 0.81. Thus

$$P(\text{other topic}) = 1 - 0.81 = 0.19 \quad \text{or} \quad 19\%.$$

□

[1] Adapted from Exercise 4.26 in Moore, McCabe abd Graig, *Introduction to the Practice of Statistics,* Sixth Edition, p. 256

(b) What is the probability that the randomly chosen spam e-mail offers adult content or is a scam?

Solution: Again, assume that each piece of spam falls into exactly one of the categories. Then,

$$P(\text{adult or scam}) = P(\text{adult}) + P(\text{scam}) = 0.287.$$

□

3. (PINs[2]) Personal identification numbers (PINs) for automatic teller machines usually consist of four digits. Suppose you notice that most of your pins have the digit 1 in them. You wonder if the issuers of PINs use lots of ones to make the numbers easier to remember. Assume that the choice of digits for a 4–digit PIN is done randomly, so that all digits have the same likelihood of being chosen.

   (a) How many possible PINs are there?

   Answer: There are 10^4 or $10,000$ PINs. □

   (b) What is the probability that a PIN assigned at random has at least one 1 in it?

   Solution: $P(\text{at least a 1}) = 1 - P(\text{no 1's})$. Since we are assuming independence,

$$P(\text{no 1's}) = \frac{9}{10} \cdot \frac{9}{10} \cdot \frac{9}{10} \cdot \frac{9}{10} = 0.6561.$$

   It then follows that $P(\text{at least a 1}) = 0.3439$. □

4. (Nonstandard dice[3]) Assume you have two balanced, six–sided dice. One is a standard die, with faces having 1, 2, 3, 4, 5, and 6 spots. The other die has three faces with no spots and three faces with 6 spots.

   (a) Describe the sample space for the experiment consisting of tossing the two dice simultaneously.

[2] Adapted from Exercise 4.33 in Moore, McCabe abd Graig, *Introduction to the Practice of Statistics,* Sixth Edition, p. 257

[3] Adapted from Exercise 4.57 in Moore, McCabe abd Graig, *Introduction to the Practice of Statistics,* Sixth Edition, p. 268

Solution: The sample space consists of the pairs

$$\begin{array}{cccccc} (0,1) & (0,2) & (0,3) & (0,4) & (0,5) & (0,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array}$$

where the first number in the pair indicates the number of spots in the nonstandard die. □

(b) Let X denote the sum of the spots on the up–faces of the two dice after they are rolled. Give the probability distribution of X.

Solution: Possible values for X are $1, 2, \ldots, 12$. Assuming that the dice are balanced, elements in the sample space are equally likely. Hence

$$P(X = k) = \frac{1}{12} \quad \text{for } k - 1, 2, \ldots, 12$$

and 0 otherwise. □

5. (Foreign–born residents of California[4]) The Census Bureau reports that 27% of California residents are foreign–born. Suppose that you choose three Californians at random so that each has probability $p = 0.27$ of being foreign born, and the choice of each individual is independent from that of any other in the group.

(a) List the elements of the sample space using the letter F to denote foreign–born and D to denote domestic birth.

Solution:

$$\left.\begin{array}{c} \text{FFF} \\ \text{FFD} \\ \text{FDF} \\ \text{FDD} \\ \text{DFF} \\ \text{DFD} \\ \text{DDF} \\ \text{DDD} \end{array}\right\} \text{Sample Space} \tag{1}$$

□

(b) Define a random variable, W, to be the number of foreign–born people in the group of three that are chosen. What are the possible values of W?

[4]Adapted from Exercise 4.59 in Moore, McCabe abd Graig, *Introduction to the Practice of Statistics,* Sixth Edition, pp. 268, 269

Answer: Possible values for W are 0, 1, 2 or 3. □

(c) Give the probability distribution of W

Solution: Using the given fact that $P(\text{F}) = 0.27$, we get that $P(\text{D}) = 1 - 0.27 = 0.73$. We can then use to we compute the probability for each element in the sample space. The results are shown below

Outcome	Probability
FFF	0.019683
FFD	0.053217
FDF	0.053217
FDD	0.143883
DFF	0.053217
DFD	0.143883
DDF	0.143883
DDD	0.389017

The event $(W = 0)$ consists of the single outcome DDD. Thus, $P(W = 0) = 0.389017$. On the other hand, the event $(W = 1)$ consists of the outcomes FDD, DFD and DDF, each of which has probability 0.143883. It then follows that

$$P(W = 1) = 3(0.143883) = 0.431649.$$

Proceeding in this fashion we obtain the probability distribution for W to be

$$P(W = k) = \begin{cases} 0.389017 & \text{if } k = 0; \\ 0.431649 & \text{if } k = 1; \\ 0.159651 & \text{if } k = 2; \\ 0.019683 & \text{if } k = 3. \end{cases}$$

□