Probability and Statistics Exercises

1. Box 1 contains 8 apples and 4 oranges. Box 2 contains 10 apples and 2 oranges. Boxes are chosen with equal probability. What is the probability of choosing an apple? If an apple is chosen, what is the probability that it came from box 1?

2. The Prosecutor’s fallacy. The probability that there is a DNA match given that a person is innocent is estimated as $1/100,000$. Assume that the probability that there is a match given that a person is guilty is 1. Suppose that the defendant in a trial lives in a city where there are 10,000 people that could have committed the crime, and that there is a DNA match to the defendant. Calculate the probability that the defendant is indeed guilty, given no other evidence except the DNA match, i.e., $P(\text{guilty}|\text{DNA match})$. How does this vary as the size of the population varies?

3. A study of a very limited population of Aliens reveals the following number of body appendages (limbs):

   $2, 3, 6, 8, 11, 18$

(a) Find the mean $m$ and the standard deviation $\sigma$ of this population.

(b) List all possible samples of two aliens without replacement (“zonder terugleggen”), and find each mean. Do the same with samples of four aliens.

(c) Each of the means above is called a sample mean. Find the mean of all the sample means (denoted by $m_x$) and the standard deviation of all the sample means (denoted by $\sigma_x$) for both the $N = 2$ and $N = 4$ samples.

(d) Verify the Central Limit Theorem: (i) compare the population mean with the mean of both sample means; (ii) compare the population standard deviations divided by the square root of the sample size with the standard deviation of both sample means (i.e.,
does $\sigma_x \approx \sigma / \sqrt{N}$). BTW, a better approximation for small population sizes is $\sigma_x = \sigma / \sqrt{N} \times \sqrt{(M - N)/(M - 1)}$ with $M = 6$ the size of the original population.

(e) Compare the distribution (shape) of the population with the distributions of both sample means using histograms. What happens to the shape of the sample means as the sample size $(N)$ increases?

4. A tire manufacturer claims that its tires will last no less than an average of 50,000 km before they need to be replaced. A consumer group wishes to challenge this claim.

(a) Clearly define the parameter of interest in this problem.

(b) State the null hypothesis and the alternative hypothesis in terms of this parameter.

(c) This is called a one-sided test. Why?

(d) In the context of the problem, state what it means to make a type 1 error.

(e) Suppose we set the significance level of the test at 0.1, what does this number mean?

5. An economist estimates that the average Canadian household saves 15% of its income. In a random sample of 64 households, the average saving rate is found to be 14%, and the standard deviation is 7%.

(a) Clearly state the population, the parameter of interest, the null hypothesis and the alternative hypothesis in this problem.

(b) Define the test statistic. Assume that this test statistic is normally distributed.

(c) This is called a two-sided test. Why?

(d) Use a significance level of 0.05 and Table 1 below. What is the rejection region? Do we have enough evidence to refute the economist’s claim?
(e) An alternative road is through the computation of the p-value. Use e.g. the table on [http://www.cs.ru.nl/~tomh/onderwijs/lrs/lrs_files/normtable.pdf](http://www.cs.ru.nl/~tomh/onderwijs/lrs/lrs_files/normtable.pdf). 

*Hint:* given the value of the test statistic $Z$, find in the table the probability $P(X \leq |Z|)$. The p-value for a one-sided test corresponds to $1 - P(X \leq |Z|)$. But since we have a two-sided test, we have to multiply by 2. In math:

$$P(X < -|Z| \lor X > |Z|) = 2P(X > |Z|) = 2[1 - P(X \leq |Z|)] .$$

(f) Explain what the p-value in part e) means.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>one-sided</th>
<th>two-sided</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.282</td>
<td>± 1.645</td>
</tr>
<tr>
<td>0.05</td>
<td>1.645</td>
<td>± 1.960</td>
</tr>
<tr>
<td>0.01</td>
<td>2.326</td>
<td>± 2.576</td>
</tr>
<tr>
<td>0.001</td>
<td>3.090</td>
<td>± 3.291</td>
</tr>
</tbody>
</table>

Table 1: Critical values of the cumulative normal distribution.