Statistical Foundations of Data Science
Past Homework 3

The content of this assignment is based on the seventh lecture in the course titled "Bayes Theorem".

1. A test for a rare disease comes up positive. The test is ninety percent accurate when the person is really infected and ninety-five percent accurate when the person is not infected. Surprisingly the results of taking the same test multiple times are independent. One percent of the population has the disease.

   a Compute the probability of being infected given a positive test [3]
   The fraction of the total population that are truly infected and who receive a positive test is $0.01 \times 0.9 = 0.009$, the fraction of the population that are not infected but none the less set of the test is $0.99 \times 0.05 = 0.0495$. Thus the probability of being infected given these prevalences is $\frac{0.009}{0.009+0.0495} = 0.15384615384$ or around fifteen percent.

   b Compute the probability of being infected given a negative test [3]
   The fraction of the total population that are truly infected and who receive a negative test is $0.01 \times 0.1 = 0.001$, the fraction of the population that are not infected and receive a negative test is $0.99 \times 0.95 = 0.9405$. Thus the probability of being infected given these prevalences is $\frac{0.001}{0.001+0.9405} = 0.00106213489$

   c Compute the probability of being infected given two positive tests [3]
   The fraction of the total population that are truly infected and who receive two positive tests is $0.01 \times 0.9 \times 0.9 = 0.0081$, the fraction of the population that are not infected but none the less set of the test twice is $0.99 \times 0.05 \times 0.05 = 0.002475$. Thus the probability of being infected given these prevalences is $\frac{0.0081}{0.0081+0.002475} = 0.7659574468$.

   d Compute the probability of being infected given two negative tests [3]
   The fraction of the total population that are truly infected and who nonetheless receive two negative tests is $0.01 \times 0.1 \times 0.1 = 0.0001$, the fraction of the population that are not infected and who pass the test twice is $0.99 \times 0.95 \times 0.95 = 0.893475$. Thus the probability of being infected given these prevalences is $\frac{0.0001}{0.0001+0.893475} = 0.00011191002$.

   e Compute the probability of being infected given two tests one of which is positive and one of which is negative [3]
   The fraction of the total population that are truly infected and who pass a test and then fail one is $0.01 \times 0.9 \times 0.1 = 0.0009$, the fraction of the population that are not infected and who pass a test and then fail one is $0.99 \times 0.05 \times 0.95 = 0.047025$. Thus the probability of being infected given these prevalences is $\frac{0.0009}{0.0009+0.047025} = 0.01877934272$.

2. If X and Y are independent discrete uniform random variables on the set with parameter 10, (that is equally liekly to take on any of the values in $\{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ and never take on any other values) Compute:

   a $\mathbb{P}(X > 4)$ [1] We enumerate the 6 values $5, 6, 7, 8, 9$ and 10 to see that the answer is $\frac{6}{10}$

   b $\mathbb{P}(XY > 75)$ [3] Again we enumerate (x, y) pairs: $(8, 10)$, $(9, 9)$, $(9, 10)$, $(10, 8)$, $(10, 9)$ and $(10, 10)$ to see that the answer is $\frac{6}{100}$

   c $\mathbb{P}(XY > 65)$ [3] Again we enumerate (x, y) pairs: $(7, 10)$,$(8, 9)$, $(8, 10)$, $(9, 8)$, $(9, 9)$, $(9, 10)$, $(10, 7)$,$(10, 8)$ $(10, 9)$ and $(10, 10)$ to see that the answer is $\frac{10}{100} = \frac{1}{10}$

   d $\mathbb{P}(X > 8|XY > 75)$ [4] $\frac{5}{6}$. We simply enumerate the six cases in part b and observe that five of them have $x > 8$

   e $\mathbb{P}(Y > 7|XY > 65)$ [4] $\frac{9}{10}$. We simply enumerate the ten cases in part c and observe that nine of them have $x > 7$

3. Various students take different IQ tests. These tests may be on different scales, that is may have different means and variances. However in all cases the distributions will be normal

   a Albus gets a Z-score of 3.5. The mean on his test was 110 and the variance 144. Find Albus's score [4]
   The standard deviation $= \sqrt{144} = 12$. This allows us to plug values into the equation $Z = \frac{x-\mu}{\sigma}$ and get $3.5 = \frac{x-110}{12}$ which solves to $x = 152$

   b Voldemort scores 2 standard deviations below normal with a score of 70 on a test with mean 100. Find the standard deviation of scores on the test [4]
   Again we use the equation $Z = \frac{x-\mu}{\sigma}$ and get $-2 = \frac{70-100}{\sigma}$ which solves to $\sigma = 15$

   c Hagrid scores 0.5 standard deviations below the mean. If the tests variance is 100 and Hagrid scored 105 find the tests' mean [4]
   We compute the standard deviation $= \sqrt{100} = 10$ and again use $Z = \frac{x-\mu}{\sigma}$ and get $-0.5 = \frac{105-\mu}{10}$ which solves to $\mu = 110$

   d Harry scores right on the mean with a of 90. Find his Z-score [4]
   As Harry is on the mean his Z-score is 0.

4. The binomial distribution is a sum of IID random variables and is therefore approximately normal large n and fixed p. In our question a fair coin ($p = 0.5$) is flipped one million times. The random variable counting the number of heads is called X.

   a Compute $\mathbb{E}[X]$ [3]
   Using the formula for binomials $\mathbb{E}[X] = np = 500000$

   b Compute $V(X)$ [3]
   Using the formula for binomials $\mathbb{E}[X] = np(1-p) = 250000$

   c Compute the standard deviation of X [3] Using the formula for binomials $\sigma = \sqrt{250000} = 500$

   d Use the CLT to estimate $\mathbb{P}(499500 < X < 501000)$ [5]
   This is the probability that a standard normal random variable is between -1 and 2 standard deviations. This gives $0.9772 - 0.1587 = 0.8185$

5. We have a weighted coin which was sold as coming up heads with probability $\frac{2}{3}$.

   a We flip the coin ten times and get eight heads. Set up a one-sided hypothesis test and compute the p-value. Do not use the normal approximation. [10]
   H_0: $p = \frac{2}{3}$
   H_1: $p > \frac{2}{3}$
   $\alpha = 0.05$
   Let X represent the number of heads. Under the null hypothesis the probability of getting a result this or more extreme is:

$$\mathbb{P}(x=8) + \mathbb{P}(x=9) + \mathbb{P}(x=10) = \binom{10}{8}(\frac{2}{3})^8(\frac{1}{3})^2 + \binom{10}{9}(\frac{2}{3})^9(\frac{1}{3})^1 + \binom{10}{10}(\frac{2}{3})^{10}(\frac{1}{3})^0$$
$$= 0.29914139104$$

   This means we fail to reject the null hypothesis at the five-percent level.

   b We continue to flip until we've flipped a total of a thousand times and we've received 690 heads. Repeat the above exercise, use the normal approximation. [10]
   H_0: $p = \frac{2}{3}$
   H_1: $p > \frac{2}{3}$
   $\alpha = 0.05$

Let X represent the number of heads. Under the null hypothesis this is well approximated by a normal distribution with mean $\frac{2000}{3}$ and variance $\frac{2000}{9}$: This produced a Z-score of $\frac{690-\frac{2000}{3}}{\sqrt{\frac{2000}{9}}} = 1.56524758425$, which corresponds to a p-value of about 0.06 and we again fail to reject the null hypothesis

6. A company sells chocolate bars labelled as weighing 200 grams. We know that the weights are actually normally distributed with variance $100g^2$ We buy 20 such chocolate bars and set a sample mean of $195g$. Perform a one sided hypothesis test to determine if the chocolate bars are underweight. [10]
   Let μ be the true mean weight of all chocolate bars. H_0: $\mu = 200$
   H_1: $\mu < 200$
   $\alpha = 0.05$
   We have the mean weight $\overline{X} = 195$. Under the null hypothesis represent the number of heads. Under the null hypothesis $\overline{X}$ is distributed normally with mean 200 and variance $\frac{100}{20} = 5$: This produced a Z-score of $\frac{195-200}{\sqrt{5}} = -2.2360679775$, which corresponds to a p-value of about 0.0126 and reject the null hypothesis.