1. (10 points) You are an education researcher, and you believe that good students get more sleep. To test this hypothesis, you survey students at a local high school. In a sample of 100 honors students, the average number of hours of sleep per night is 7, with a variance of 0.5. In a sample of 100 non-honors students, the average number of hours of sleep per night is 6.75, with a variance of 0.5. You hypothesize that honors students get more hours of sleep per night than non-honors students.

   (a) State the null hypothesis, alternative hypothesis, and test statistic. Give the form of the rejection region.

   The parameter of interest is $\mu_1 - \mu_2$, the difference between the means of the two populations (honors students and non-honors students). The null hypothesis is $\mu_1 - \mu_2 = 0$, the alternative hypothesis is $\mu_1 - \mu_2 > 0$, and the test statistic is $\bar{Y}_1 - \bar{Y}_2$. Since this is an upper-tail test, the rejection region has the form $\bar{Y}_1 - \bar{Y}_2 \geq k$.

   (b) At the level of $\alpha = 0.05$, is there sufficient evidence to support the hypothesis that honors students get more hours of sleep per night than non-honors students?

   Since this is a large sample test, the estimator $\bar{Y}_1 - \bar{Y}_2$ has an (approximately) normal distribution, so we can use the $Z$-test. Since we have an upper tail test, the value of $z$ we need is $z_\alpha = z_{0.05} = 1.65$ (you could also use 1.64 or 1.645). The standard deviation of the estimator is:

$$\begin{aligned}\sigma_{\bar{Y}_1 - \bar{Y}_2} &= \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \\ &= \sqrt{\frac{0.5}{100} + \frac{0.5}{100}} = \sqrt{\frac{1}{100}} \\ &= \frac{1}{10} = 0.1\end{aligned}$$

   (Note that the problem gives us the variance, not the standard deviation.) The rejection region is therefore:

$$\begin{aligned}\bar{Y}_1 - \bar{Y}_2 &\geq 0 + z_\alpha \sigma_{\bar{Y}_1 - \bar{Y}_2} = 1.65(0.1) \\ \bar{Y}_1 - \bar{Y}_2 &\geq 0.165\end{aligned}$$

   Since our test statistic $\bar{Y}_1 - \bar{Y}_2 = 7 - 6.75 = 0.25$ lies inside the rejection region, we reject the null hypothesis at the level of $\alpha = 0.05$, thus there is sufficient evidence to support the hypothesis that honors students get more hours of sleep per night than non-honors students at this level.

   (c) What is the $p$-value for this test?

   The $p$-value is the smallest value of $\alpha$ for which we will reject the null hypothesis

given our observation.

$$\begin{aligned} p &= \mathbb{P}(\bar{Y}_1 - \bar{Y}_2 \geq 0.25 | \text{null hypothesis is true}) \\ &= \mathbb{P}\left(Z \geq \frac{0.25 - 0}{0.1}\right) \\ &= \mathbb{P}(Z \geq 2.50) \\ &= 0.0062 \end{aligned}$$

(d) Suppose we have reason to believe that honors students get 0.4 more hours of sleep per night than non-honors students. Using this as the alternative hypothesis, and using the same rejection region as found above, what is the value of $\beta$ for this test?

$$\begin{aligned} \beta &= \mathbb{P}(\text{test statistic is outside rejection region} | \text{alternative hypothesis is true}) \\ &= \mathbb{P}(\bar{Y}_1 - \bar{Y}_2 < 0.165 | \mu_1 - \mu_2 = 0.4) \\ &= \mathbb{P}\left(Z < \frac{0.165 - 0.4}{0.1}\right) \\ &= \mathbb{P}(Z < -2.35) \\ &= 0.0094 \end{aligned}$$

2. (10 points) Suppose we have a population which is described by the probability density function

$$f(x) = \begin{cases} \frac{2}{\theta}\left(1 - \frac{x}{\theta}\right) & 0 \le x \le \theta \\ 0 & \text{otherwise} \end{cases}$$

where $\theta > 0$ is an unknown parameter.

(a) Suppose you take $n$ samples $X_1, \ldots, X_n$ from the population. Let $\bar{X}$ be the sample mean. Find the method of moments estimator for $\theta$.

The method of moments estimator is given by setting $\bar{X} = \mu$, where $\mu$ is the population mean. To find $\mu$, we use the formula for the expected value of a continuous random variable.

$$\begin{aligned} \mu &= \int_0^\theta x \frac{2}{\theta}\left(1 - \frac{x}{\theta}\right) dx \\ &= \frac{2}{\theta}\int_0^\theta \left(x - \frac{x^2}{\theta}\right) dx \\ &= \frac{2}{\theta}\left(\frac{x^2}{2} - \frac{x^3}{3\theta}\right)\Big|_0^\theta \\ &= \frac{2}{\theta}\left(\frac{\theta^2}{2} - \frac{\theta^2}{3}\right) \\ &= \frac{2}{\theta}\frac{\theta^2}{6} \\ &= \frac{\theta}{3} \end{aligned}$$

From this we get $\bar{Y} = \theta/3$, so the method of moments estimator for $\theta$ is

$$\hat{\theta} = 3\bar{Y}$$

(b) What is the variance of the method of moments estimator you found in part (a)?

The variance of the estimator from above is:

$$\begin{aligned} Var(\hat{\theta}) &= Var(3\bar{Y}) \\ &= 3^2 Var(\bar{Y}) \\ &= 9\frac{\sigma^2}{n} \end{aligned}$$

where $\sigma^2$ is the population variance. To find that, we use the Magic Variance

Formula. Letting $X$ be a sample from the population, we compute

$$\begin{aligned}
\mathbb{E}(X^2) &= \int_0^\theta x^2 \frac{2}{\theta}\left(1-\frac{x}{\theta}\right) dx \\
&= \frac{2}{\theta}\int_0^\theta \left(x^2 - \frac{x^3}{\theta}\right) dx \\
&= \frac{2}{\theta}\left(\frac{x^3}{3} - \frac{x^4}{4\theta}\right)\Big|_0^\theta \\
&= \frac{2}{\theta}\left(\frac{\theta^3}{3} - \frac{\theta^3}{4}\right) \\
&= \frac{2}{\theta}\frac{\theta^3}{12} \\
&= \frac{\theta^2}{6}
\end{aligned}$$

Since we computed $\mathbb{E}(X)$ in part (a), by the Magic Variance Formula,

$$\begin{aligned}
\sigma^2 &= \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 \\
&= \frac{\theta^2}{6} - \left[\frac{\theta}{3}\right]^2 \\
&= \theta^2\left(\frac{1}{6} - \frac{1}{9}\right) = \frac{\theta^2}{18}
\end{aligned}$$

Substituting this above, we get

$$Var(\hat{\theta}) = \frac{9}{n}\frac{\theta^2}{18} = \frac{\theta^2}{2n}$$

(c) Suppose you take a *single* sample $X$ from the population. Find the maximum likelihood estimator (MLE) for $\theta$.

To get the likelihood function, we plug our sample $X$ into the density function. Then we take the maximize by taking the derivative with respect to $\theta$ and setting it equal to 0.

$$\begin{aligned}
\frac{d}{d\theta}L(X|\theta) &= \frac{d}{d\theta}\frac{2}{\theta}\left(1-\frac{X}{\theta}\right) \\
&= 2\frac{d}{d\theta}\left(\frac{1}{\theta} - \frac{X}{\theta^2}\right) \\
&= 2\left(-\frac{1}{\theta^2} + 2\frac{X}{\theta^3}\right) \\
&= 2\left(\frac{2X-\theta}{\theta^3}\right)
\end{aligned}$$

Setting this to 0, this is only true if the numerator is 0. In this case, we have $\theta = 2X$. Thus the MLE for $\theta$ is $\hat{\theta}_{MLE} = 2X$.

3. (10 points) Suppose we have population whose distribution is a Poisson distribution with unknown parameter $\lambda$. You take a group of $n$ samples $X_1, \ldots, X_n$ and a group of $m$ samples $Y_1, \ldots, Y_m$ from the population. All samples are independent. You form the following estimator for $\lambda$:

$$\hat{\lambda} = a\frac{X_1 + \cdots + X_n}{n} + b\frac{Y_1 + \cdots + Y_m}{m}$$

where $a$ and $b$ are constants.

(a) What condition is needed on $a$ and $b$ so that $\hat{\lambda}$ is unbiased?

Using linearity of expectation and the expected value of the Poisson distribution,

$$\begin{aligned}
\mathbb{E}(\hat{\lambda}) &= \frac{a}{n}\mathbb{E}(X_1 + \cdots + X_n) + \frac{b}{m}\mathbb{E}(Y_1 + \cdots + Y_m) \\
&= \frac{a}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) + \frac{b}{m}\sum_{j=1}^{m}\mathbb{E}(Y_j) \\
&= \frac{a}{n}\sum_{i=1}^{n}\lambda + \frac{b}{m}\sum_{j=1}^{m}\lambda \\
&= \frac{a}{n}n\lambda + \frac{b}{m}m\lambda \\
&= (a+b)\lambda
\end{aligned}$$

This is only equal to the population mean $\lambda$ if $a + b = 1$, which is the condition for the estimator to be unbiased.

(b) What is the mean squared error (MSE) for $\hat{\lambda}$? Do not assume that $a$ and $b$ satisfy the condition you found in part (a), i.e. do not assume the estimator $\hat{\lambda}$ is unbiased.

Recall that MSE is the sum of the bias squared and the variance.

$$\begin{aligned}
Bias(\hat{\lambda}) &= \mathbb{E}(\hat{\lambda}) - \lambda \\
&= (a+b)\lambda - \lambda = (a+b-1)\lambda
\end{aligned}$$

For the variance, we use the fact that the samples are independent and the variance

of the Poisson distribution.

$$\begin{aligned}
Var(\hat{\lambda}) &= \frac{a^2}{n^2} Var(X_1 + \cdots + X_n) + \frac{b^2}{m^2} Var(Y_1 + \cdots + Y_m) \\
&= \frac{a^2}{n^2} \sum_{i=1}^{n} Var(X_i) + \frac{b^2}{m^2} \sum_{j=1}^{m} Var(Y_j) \\
&= \frac{a^2}{n^2} \sum_{i=1}^{n} \lambda + \frac{b^2}{m^2} \sum_{j=1}^{m} \lambda \\
&= \frac{a^2}{n^2} n\lambda + \frac{b^2}{m^2} m\lambda \\
&= \left( \frac{a^2}{n} + \frac{b^2}{m} \right) \lambda
\end{aligned}$$

Alternatively, we can write the estimator as $\hat{\lambda} = a\bar{X} + b\bar{Y}$, and use what we learned in class about the variances of $\bar{X}$ and $\bar{Y}$.

The MSE is then given by:

$$\begin{aligned}
MSE(\hat{\lambda}) &= [Bias(\hat{\lambda})]^2 + Var(\hat{\lambda}) \\
&= (a + b - 1)^2 \lambda^2 + \left( \frac{a^2}{n} + \frac{b^2}{m} \right) \lambda
\end{aligned}$$

4. (10 points) A bag contains $w$ white marbles and $r$ red marbles. You draw a sample of $n$ marbles from the bag without replacement, where $n \leq w + r$.

   (a) What is the probability that your sample contains exactly $y$ red marbles, where $0 \leq y \leq r$?

   If we draw exactly $y$ red marbles, we must draw $n - y$ marbles, so using combinatorics, the probability is

$$\frac{\binom{r}{y}\binom{w}{n-y}}{\binom{r+w}{n}}$$

   (b) Suppose one of the red marbles in the bag is labeled with the number 1. What is the probability that your sample contains the red marble which is labeled with the number 1?

   There are many ways to do this. The probability of the first marble being the Red 1 is $1/(r+w)$. Since order does not matter (so the marbles are exchangeable), the probability of any of the $n$ marbles being the Red 1 must be the same thing, i.e. $1/(r+w)$. Since the $n$ events "The $i$th ball is the Red 1" are disjoint (mutually exclusive), the probability that we get the Red 1 in our sample is:

$$\sum_{i=1}^{n} \mathbb{P}(\text{the } i\text{th ball is the Red 1}) = \sum_{i=1}^{n} \frac{1}{r+w} = \frac{n}{r+w}$$

   You can also use a combinatorics approach to get the probability. There are $n$ balls chosen; if you get the Red 1, one of these the Red 1, and the other $n-1$ of these are chosen from the $r+w-1$ balls which are not the Red 1.

$$\begin{aligned}
\frac{\binom{1}{1}\binom{r+w-1}{n-1}}{\binom{r+w}{n}} &= \frac{\dfrac{(r+w-1)!}{(n-1)![(r+w-1)-(n-1)]!}}{\dfrac{(r+w)!}{n!(r+w-n)!}} \\
&= \frac{\dfrac{(r+w-1)!}{(n-1)!(r+w-n)!}}{\dfrac{(r+w)!}{n!(r+w-n)!}} \\
&= \frac{(r+w-1)!n!}{(r+w)!(n-1)!} \\
&= \frac{(r+w-1)!n(n-1)!}{(r+1)(r+w-1)!(n-1)!} \\
&= \frac{n}{r+w}
\end{aligned}$$

   (c) What is the expected number of red marbles in your sample?

There are again many ways to do this. We will use linearity of expectation together with part (b). This time, label the $r$ red balls $1, 2, \ldots, r$. Define the indicator random variables $R_i$ by

$$R_i = \begin{cases} 1 & \text{red ball } i \text{ is in your sample} \\ 0 & \text{red ball } i \text{ is } \textit{not} \text{ in your sample} \end{cases}$$

Let $R$ be the number of red balls in your sample. Then $R = \sum_{i=1}^{r} R_i$. By linearity of expectation,

$$E(R) = \sum_{i=1}^{r} \mathbb{E}(R_i)$$

For each indicator random variable,

$$\begin{aligned} E(R_i) &= 0 \cdot \mathbb{P}(R_i = 0) + 1 \cdot \mathbb{P}(R_i = 1) \\ &= \mathbb{P}(R_i = 1) \\ &= \mathbb{P}(\text{red ball } i \text{ is in your sample}) \\ &= \frac{n}{r+w} \end{aligned}$$

where the last line is from part (b), since the probability is the same for any red ball $i$ as it is for red ball 1 by symmetry. Thus we have:

$$\begin{aligned} \mathbb{E}(R) &= \sum_{i=1}^{r} \mathbb{E}(R_i) \\ &= \sum_{i=1}^{r} \frac{n}{r+w} \\ &= r\left(\frac{n}{r+w}\right) \\ &= n\left(\frac{r}{r+w}\right) \end{aligned}$$

Note that this expected value is the *same* as that of a binomial random variable with parameter $p = r/(r+w)$, thus the expected number of red balls in the sample is the same whether we sample without replacement (hypergeometric) or with replacement (binomial).

5. (10 points) Consider the following timetable for the train from Zurich to Geneva, Switzerland. Only times between 8:00 and 9:00 are shown.

| 8:00 | 8:30 | 9:00 |
|---|---|---|

Since these are Swiss trains, they are always exactly on time! Suppose you arrive at the Zurich train station uniformly at random between 8:20 and 9:00 and wait for the train to Geneva.

(a) What is the probability density function for the amount of time you spend waiting for the train? Be sure to give appropriate bounds on the density function. You may describe the density function any way you wish, as long as the values and bounds of the density function are clear.

There are a bunch of ways to do this. Here is one of them. If you arrive between 8:20 and 8:30, you wait for the 8:30 train, so the amount of time you wait is Uniform(0, 10). If you arrive between 8:30 and 9:00, you wait for the 9:00 train, so the amount of time you wait is Uniform(0, 30). The probability that you arrive 8:20 and 8:30 is 1/4, and the probability that you arrive between 8:30 and 9:00 is 3/4, since your arrival time is uniformly distributed over a 40-minute interval of time. Thus the density for your waiting time is:

$$\frac{1}{4}\text{Uniform}(0, 10) + \frac{3}{4}\text{Uniform}(0, 30)$$

At this point, it might be easiest to draw pictures of the two uniform densities, scale them appropriately, and add them together. You can then express the density as a graph, which is totally fine. More "mathematically" the Uniform(0, 10) density is:

$$f(x) = \begin{cases} \frac{1}{10} & 0 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

The Uniform(0, 30) density is:

$$g(x) = \begin{cases} \frac{1}{30} & 0 \leq x \leq 30 \\ 0 & \text{otherwise} \end{cases}$$

Then the density for the waiting time is:

$$\begin{aligned}
\frac{1}{4}f(x)+\frac{3}{4}g(x) &= \frac{1}{4}\begin{cases}\frac{1}{10} & 0\le x\le 10\\ 0 & \text{otherwise}\end{cases}+\frac{3}{4}\begin{cases}\frac{1}{30} & 0\le x\le 30\\ 0 & \text{otherwise}\end{cases}\\
&= \begin{cases}\frac{1}{40} & 0\le x\le 10\\ 0 & \text{otherwise}\end{cases}+\begin{cases}\frac{1}{40} & 0\le x\le 30\\ 0 & \text{otherwise}\end{cases}\\
&= \begin{cases}\frac{1}{40}+\frac{1}{40} & 0\le x< 10\\ \frac{1}{40} & 10\le x\le 30\\ 0 & \text{otherwise}\end{cases}\\
&= \begin{cases}\frac{1}{20} & 0\le x< 10\\ \frac{1}{40} & 10\le x\le 30\\ 0 & \text{otherwise}\end{cases}
\end{aligned}$$

(b) What is the expected amount of time you spend waiting for the train?

For this, you can multiply the density from part (a) by $x$ and integrate from 0 to 30. Alternatively, letting $X$ be the waiting time for the next train, by linearity of expectation and using the mean of the uniform distribution:

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}\left(\frac{1}{4}\text{Uniform}(0,10)+\frac{3}{4}\text{Uniform}(0,30)\right)\\
&= \frac{1}{4}\mathbb{E}(\text{Uniform}(0,10))+\frac{3}{4}\mathbb{E}(\text{Uniform}(0,30))\\
&= \frac{1}{4}(5)+\frac{3}{4}(15)\\
&= \frac{5}{4}+\frac{45}{4}\\
&= \frac{50}{4}\\
&= \frac{25}{2}
\end{aligned}$$