This assignment is due **August 20, 2025**. You may work in teams of up to four people. If you work as a team, choose a leader to sign up your team on the homework 3 groups page of Bruin Learn. Please submit one html or pdf file. Upload your solutions to Bruin Learn. Show all your work. You may use the modules `math`, `numpy`, `pandas`, `scipy`, and `matplotlib`.

1. We will create a toy model to illustrate how difficult it is to determine an investor's skill. This model assumes that investors are either skilled or unskilled, each has control of an equal amount of capital, and that market beats are independent. Let us say that skilled investors have a 55% chance of beating the market, and unskilled investors have a 46% chance of beating the market. Use the class `binom` within `scipy.stats` to solve this problem.

   (a) Because of transaction costs and fees, the average investor should beat the market slightly less than 50% of the time. If we assume that the average investor beats the market 47% of the time, what fraction of investors must be skilled?

   (b) Given that an investor beat the market last year, use Bayes' rule to determine the probability that she is skilled.

   (c) Use Bayes' rule to determine the probability that an investor is skilled given that she beats the market in at least two of the last three years.

   (d) Verify (c) using a simulation. Suppose there are 1,000,000 investors and determine the number of skilled and unskilled investors. Then simulate the number of times each investor beats the market over the last three years for each skill level. Use these results to find the ratio of skilled investors that beat the market at least twice to the total number of investors that beat the market at least twice.

   (e) Using either Bayes' rule or a simulation create a graph of the number of years examined to the probability of being skilled. Let the horizontal axis represent the number of years n. Then use `plot` in `matplotlib.pyplot` to graph the probabilities of being skilled given at least $n-2$, $n-1$, and n market beats in the last n years. Suppose $2 \leq n \leq 10$. Add a legend to help distinguish your three plots.

2. Assume that R is the random variable representing the monthly return of a stock and $X = \ln(1+R) \sim \mathcal{N}(\mu, \sigma^2)$, i.e. the corresponding continuous return X follows a normal distribution.

(a) Suppose $X_k \sim \mathcal{N}(\mu, \sigma^2)$ for $k = 1, 2, \ldots, 12$. Assume X_i and X_j are independent for $i \neq j$. Because you can add continuous returns to find the total return over multiple periods, the total return over a twelve-month period is

$$Y = X_1 + X_2 + \ldots + X_{12}.$$

Find formulas for $E[Y]$ and $\text{Var}(Y)$ using our assumptions about X_k.

Month	April	May	June
Return	-1%	6%	5%

We would like to determine the annualized mean and standard deviation using the monthly returns in the table. The most common way to annualize monthly returns and variances is via $\bar{R} \mapsto 12 \cdot \bar{R}$ and $s^2 \mapsto 12 \cdot s^2$, where $\bar{R}$ and s^2 denote the respective sample mean and variance of R. This is not the best method from a mathematical perspective because it ignores compounding. However, this technique is more reasonable for continuous returns due to (a). With that in mind, given the discrete expected return and variance, we can calculate μ and σ^2, annualize them, and then convert them to discrete annual returns. To accomplish this, note that

$$E[R] = E\left[e^X\right] - 1 = e^{\mu+\sigma^2/2} - 1 \qquad \text{and} \qquad \text{Var}(R) = \text{Var}(e^X) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}.$$

(b) Use the data in the table to estimate $E[R]$ and $\text{Var}(R)$ using sample statistics. Then use algebra to find the corresponding estimates of μ and σ^2. Call the respective estimates $\hat{\mu}$ and $\hat{\sigma}^2$. Annualize $\hat{\mu}$ and $\hat{\sigma}^2$ via $\hat{\mu} \mapsto 12 \cdot \hat{\mu}$ and $\hat{\sigma}^2 \mapsto 12 \cdot \hat{\sigma}^2$.

(c) Calculate the annualized discrete return and variance using your annualized results from (b).

x_i	1	2	3	4	5	6	7	8	9	10
y_i	-1.91	0.02	2.05	4.11	6.09	8.45	10.05	11.99	13.99	16.02

3. In the previous homework, we learned how to estimate the coefficients α and β in the linear equation $y_i = \alpha + \beta x_i + \varepsilon_i$. In this assignment, we will analyze ways to construct confidence intervals for these coefficient estimates at a significance level of 10%. The random variable ε_i represents the error inherent in using $\alpha + \beta x_i$ to predict y_i. The challenge in a confidence interval construction is making valid assumptions about the distribution of ε_i. The standard assumption is that $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. However, this assumption is frequently invalid for financial data.

(a) Find estimates a and b for respective coefficients α and β using the data in the table. To simplify the calculations for a and b, use `np.polyfit` with the `degree = 1`; the output of this function is in the order b, a.

(b) Calculate the residuals $e_i = y_i - (a + bx_i)$. Then use your residuals to calculate s_e using `np.std`. Since you estimated two parameters, i.e. α and β, set `ddof = 2`.

(c) If we assume errors are normally distributed, we can approximate the distribution of ε_i using $\mathcal{N}(0, s_e^2)$. Set `np.random.seed` for reproducibility and repeat the following steps 100,000 times:

(i) Generate $n = 10$ values of $\tilde{e}_i$ using `norm` in `scipy` and your result from (b).

(ii) Use (i) to calculate $\tilde{y}_i = a + bx_i + \tilde{e}_i$.

(iii) Estimate $\tilde{a}$ and $\tilde{b}$ using `np.polyfit` and your x- and $\tilde{y}$-values. Save your results.

Create two histograms to analyze the distributions of the $\tilde{a}$- and $\tilde{b}$-values that you generated. Make sure you set `density = True`.

(d) Using your results in (c), calculate confidence intervals for α and β using the formulas

$$\Big(2a - Q_{1-\alpha/2}(\tilde{a}),\ 2a - Q_{\alpha/2}(\tilde{a})\Big) \qquad \text{and} \qquad \Big(2b - Q_{1-\alpha/2}(\tilde{b}),\ 2b - Q_{\alpha/2}(\tilde{b})\Big),$$

where $Q_p(\tilde{\theta})$ denotes the $100 \times p\%$ quantile of the samples $\tilde{\theta}$. The simpler formulas

$$\Big(Q_{\alpha/2}(\tilde{a}),\ Q_{1-\alpha/2}(\tilde{a})\Big) \qquad \text{and} \qquad \Big(Q_{\alpha/2}(\tilde{b}),\ Q_{1-\alpha/2}(\tilde{b})\Big),$$

will also work fine, but only because the normal distribution is symmetric. You can use the function `np.quantile` to calculate the quantiles. If you saved the $\tilde{a}$- and $\tilde{b}$-values in a `pandas` data frame, the method `pd.DataFrame.quantile` is another option.

(e) Under the assumption that $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, the theoretical confidence intervals are

$$\Big(a + SE(a) \cdot t_{\alpha/2},\ a + SE(a) \cdot t_{1-\alpha/2}\Big) \qquad \text{and} \qquad \Big(b + SE(b) \cdot t_{\alpha/2},\ b + SE(b) \cdot t_{1-\alpha/2}\Big),$$

where

$$SE(a) = s_e \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \qquad \text{and} \qquad SE(b) = \frac{s_e}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}.$$

Use the formulas above to check the intervals in (d). Because $n = 10$ is small, the confidence intervals from (d) and (e) are unlikely to match perfectly.

(f) Another approach is to use "bootstrap" confidence intervals. To do this, we assume ε_i has a discrete uniform distribution on the set $\{e_1, e_2, \ldots, e_n\}$. You can simulate values from this distribution using `np.random.choice` and the residuals you found in (b); the default setting of `replace = True` is necessary for this procedure to be valid. Generate 100,000 $\tilde{a}$- and $\tilde{b}$-values using the same steps as (c), but change step (i) so that you are sampling uniformly from $\{e_1, e_2, \ldots, e_n\}$ instead of the normal distribution. Once again, draw histograms of the distributions. You already set the random seed in (c), but it might be easier to use `np.random.seed` to reset it, so that you do not need to do the calculations sequentially.

(g) Use the $\tilde{a}$- and $\tilde{b}$-values you obtained from (f) and the quantile formulas

$$\Big(2a - Q_{1-\alpha/2}(\tilde{a}),\ 2a - Q_{\alpha/2}(\tilde{a})\Big) \qquad \text{and} \qquad \Big(2b - Q_{1-\alpha/2}(\tilde{b}),\ 2b - Q_{\alpha/2}(\tilde{b})\Big),$$

to calculate the confidence intervals. Unlike in (d), the simpler formula will *not* work because the distributions are not symmetric.