Business Statistics Midterm Exam

Fall 2021: BUS41000

This is an closed-book, closed-notes exam. You may use any calculator. However, **you must solve all problems on your own**.

Please answer all problems in the space provided on the exam.

Read each question carefully and clearly present your answers.

Honor Code Pledge: "I pledge my honor that I have not violated the University Honor Code during this examination. I attempt to solve all the problems on my own without external help."

Sign: ____

Name: ____

Useful formulas

- $E(aX + bY) = aE(X) + bE(Y)$
- $Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab \cdot Cov(X, Y)$
- $Cor(X, Y) = \frac{Cov(X,Y)}{sd(X) \cdot sd(Y)}$
- The standard error of $\bar{X}$ is defined as $s_{\bar{X}} = \sqrt{\frac{s_X^2}{n}}$, where s_X^2 denotes the sample variance of X.
- The standard error for the difference in the averages between groups a and b is defined as:

  $$s_{(\bar{X}_a - \bar{X}_b)} = \sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}$$

  where s_a^2 denotes the sample variance of group a and n_a the number of observations in group a.
- The standard error for a proportion is defined by: $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- The standard error for difference in proportion is defined by:

  $$s_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

  where $\hat{p}_1$ and $\hat{p}_2$ denote two independent proportions, and n_1 and n_2 are the number of trials.
- Bayes's formula:

  $$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

  where A, B are two events.
- For $Z \sim N(0, 1)$, $P(-1 \leq Z \leq 1) = 68\%, P(-2 \leq Z \leq 2) = 95\%, P(-3 \leq Z \leq 3) = 99\%$.
- Similarly, $X \sim N(\mu, \sigma^2)$, $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 95\%$.
- Standardization to standard normal: assume $X \sim N(\mu, \sigma^2)$, $Z \sim N(0, 1)$, then

  $$P(a \leq X \leq b) = P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}).$$
- The Sharpe Ratio for Stock S: $\frac{E(S)}{\sqrt{Var(S)}}$.
- For simple linear regression $Y = \beta_0 + \beta_1 X + \epsilon$, with $\epsilon \sim N(0, \sigma^2)$. The residual standard error $s = \sqrt{\frac{SSE}{n-2}}$ estimates the standard deviation σ.

Grading Sheet for TA:

Problem	Score
P1	
P2	
P3	
P4	
P5	
P6	
P7	
P8	
Total	

Problem 1: Choosing an agent. [10 points]

You are considering to purchase a house. On a rating site, you have collected data on the two potential real estate agents in Chicago. For each rating, there are only two categories, YES (recommend) or NO (not recommend).

Recommend?	Agent BIG	Agent SMALL
YES	1644	192
NO	548	48

Is the Agent SMALL better? Justify your answer using either hypothesis testing or confidence interval (with 95% confidence guarantee). [10 points]

Problem 2: Which insurance to purchase? [10 points]

The next step is to choose a house insurance policy. Suppose there are three options available: standard policy, premium policy, and no policy (not insured). If you decide on a policy, you will have to buy it for the whole year.

Policy	Cost per month	Deductible if you file claim for house damage
Standard	\$50	\$5000
Premium	\$55	\$500

Suppose in one year, there is a 1% chance of house damage, and you estimate that the damage will cost you \$200,000. If you are insured, when you are filing a claim for the damage, you are only responsible for the deductible.

1. For one year, which one of the three options you would like to choose in expectation? What are the variances of the potential payoffs for the Standard policy and the Premium policy, respectively? [5 points]

2. Now suppose you want to keep a policy for two years. The insurance company is currently running a promotion: if you do not file a claim in the first year, your monthly cost will be zero (for the second year); otherwise, your monthly fee will stay the same. Suppose the probability of house damage is 1% each year and is independent. Now, which policy you prefer in expectation? [5 points]

Problem 3: Portfolio. [10 points]

I am building a portfolio composed of SP500 and Bonds. Assume that $SP500 \sim N(11, 19^2)$ and $Bonds \sim N(4, 6^2)$. Here we measure the annual return in percentage (i.e., the Bond has an expected annual return of 4%, with a standard deviation of 6%).

1. Consider the 50-50 split between SP500 and Bonds, assume the standard deviation of this 50-50 portfolio is

$$sd(0.5SP500 + 0.5Bonds) = 11.000$$

   Can you figure out the covariance between SP500 and Bonds, as well as the correlation? [3 points]

2. Using the covariance you calculated in sub-problem 1, can you calculate

$$sd(0.8SP500 + 0.2Bonds) =?$$

   Also, which portfolio is better: the 80-20 split between SP500 and Bonds, or the 50-50 split? Justify your answer. [2 points]

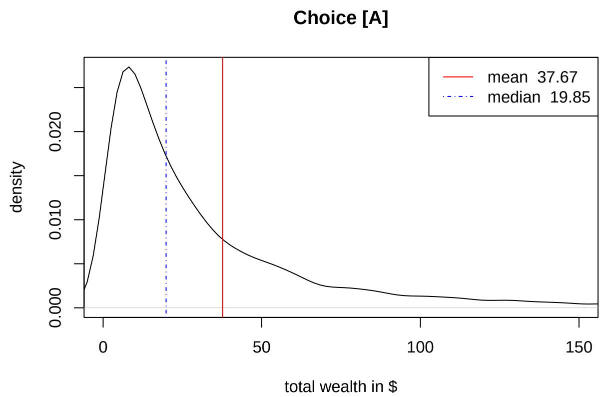
3. Suppose that you decide to invest $50,000 in a 50-50 split portfolio based on SP500 and Bonds, at the beginning of 2022. By the end of 2022, you would need to pay for the property tax, which follows a normal distribution with a mean $9,750, and a standard deviation $2,398 (independent of your portfolio). What is the probability that the return of your portfolio would be enough to cover your 2022's property tax? [5 points]

Problem 4: Investment for retirement. [20 points]

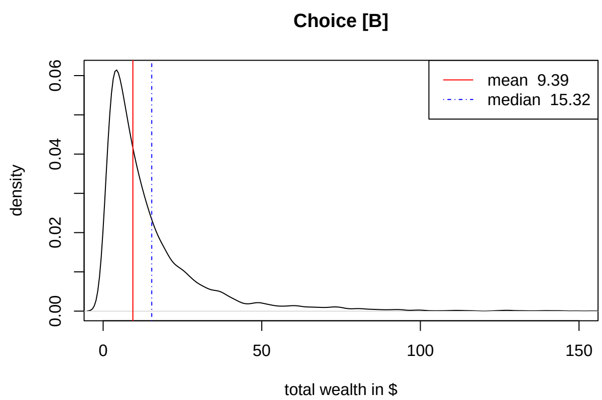
Suppose you are going to invest \$1 in S&P500 the first day you start working, and you are curious about how much that \$1 becomes in 40 years (total wealth) when you retire. You researched and realized that the annual return of S&P500 $\sim N(9.5\%, (19.5\%)^2)$, roughly modeled by a normal distribution.

1. Which one best summarizes the probability density function of the total wealth? [10 points]

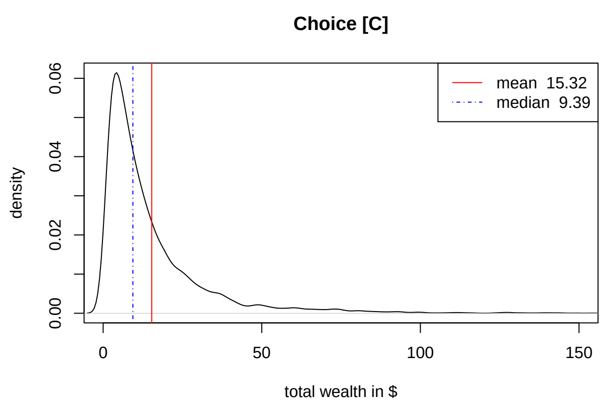
- **[A]**



- **[B]**

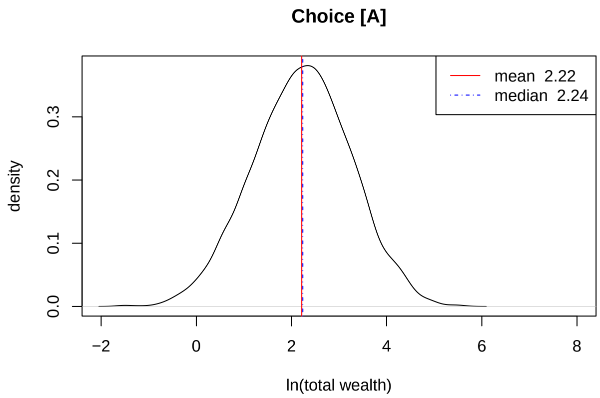


- **[C]**

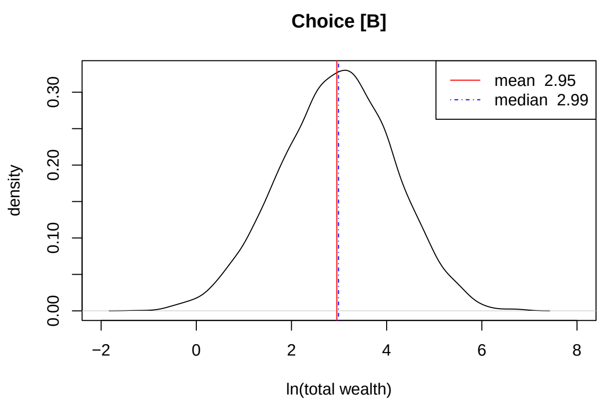


2. Which one best describes the potential outcomes for ln(total wealth)? Hint, here $\ln(x)$ is the natural logarithm function, and you may use the facts: $\ln(1.095) \approx 0.0908$, $\ln(10) \approx 2.3$, and $\ln(20) \approx 3$. To further assist you, see the following plot for $\ln(x)$. [6 points]

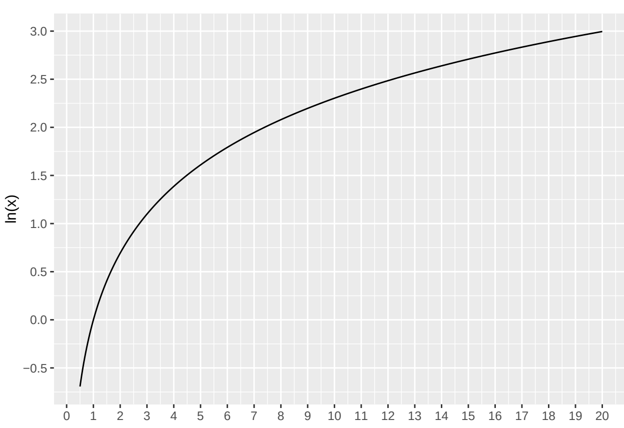
- [**A**]



- [**B**]



Further hint: you may use the graph of $\ln(x)$ as shown below.



3. Can you roughly estimate the probability $P(\text{total wealth} > 60)$? Hint: you may use the fact that $\ln(60) \approx 4.1$. [4 points]

- **[A]** $P(\text{total wealth} > 60)$ is closer to 16%.
- **[B]** $P(\text{total wealth} > 60)$ is closer to 2.5%.
- **[B]** $P(\text{total wealth} > 60)$ is closer to 0.5%.

Problem 5: Confidence interval and hypothesis testing. [15 points]

The following table summarizes the annual returns on the SP500 from 1900 until the end of 2015, in total of 116 years (in percentage terms):

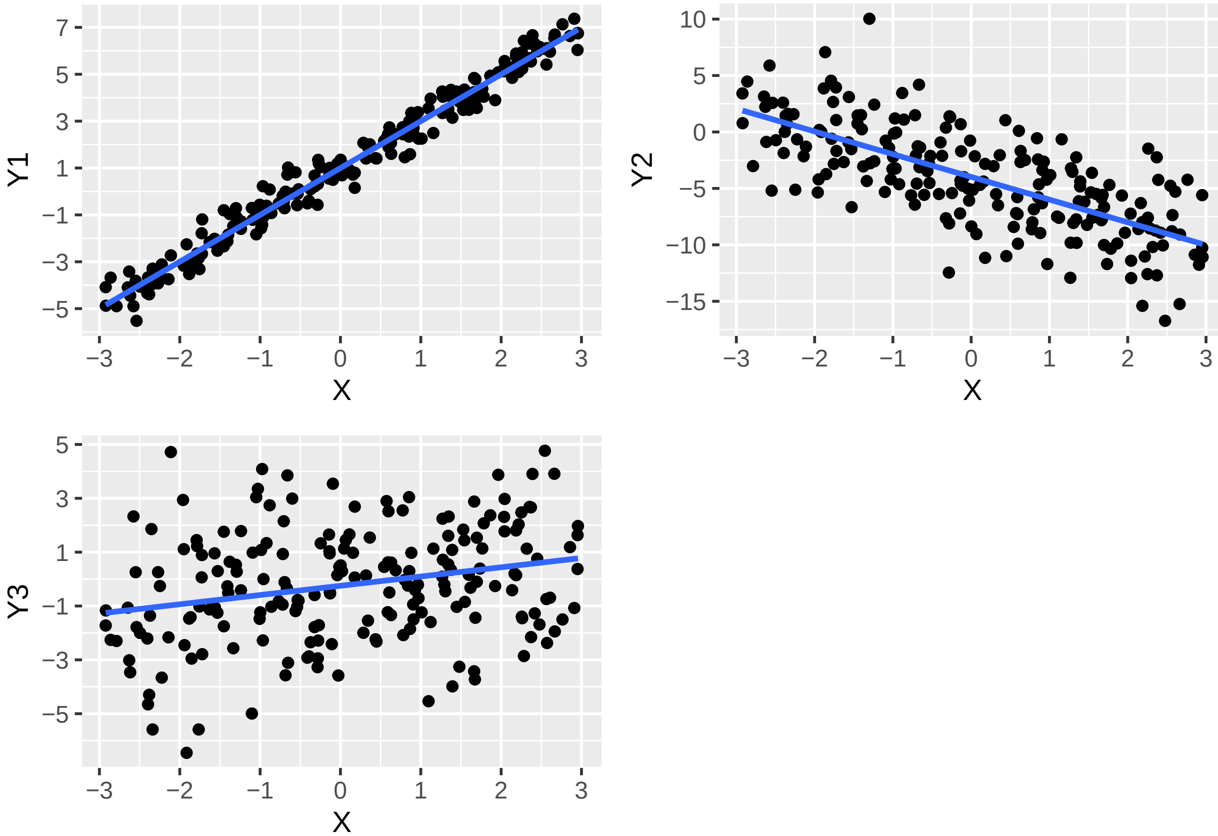
116 years of SP500	
Sample average	7.2
Sample std. deviation	13.0

1. Use a 99% confidence interval, to test the hypothesis that the expected return (true mean) of the SP500 is equal to 4% a year. [10 points]

2. In addition, suppose the 95% confidence interval (constructed based on our dataset) for the population mean of SP500 return μ is $[4.7, 9.6]$. Which one below best describes the statistical meaning? [5 points]

- [**A**] $P(\mu \text{ lies in } [4.7, 9.6]) = 95\%$, in other words, the probability that true mean of SP500 μ lies in the interval $[4.7, 9.6]$ is 95%.
- [**B**] We are 95% sure that the true mean is in the interval $[4.7, 9.6]$.
- [**C**] If we recollect datasets and build confidence intervals many times, 95% of the times, these intervals will cover the true μ.

Problem 6: Regression. [15 points]



In the above scatterplots, three different variables $Y1, Y2, Y3$ are regressed onto the same X (in all three scatterplot we have the exact same n = 200 values for X). The line is the least square regression line. In this question, we can think of residual standard error (s) for each regression as the uncertainty of the error term, $Y = b_0 + b_1 X + \epsilon, \epsilon \sim N(0, s^2)$.

Carefully examine the plots and answer the questions below:

1. Which of the following is the least square estimates of the slope (b_1) and intercept (b_0) for the regression of $Y3$ on X? [3 points]

- **[A]** $b_1 = 0.34, b_0 = -0.24$
- **[B]** $b_1 = 0.78, b_0 = 0.02$
- **[C]** $b_1 = 2.53, b_0 = -0.05$

2. Which of the following is the least square estimates of the slope (b_1) and residual standard error (s), for regression $Y2$ on X? [3 points]

- **[A]** $b_1 = -0.9, s = 6.3$
- **[B]** $b_1 = -2.0, s = 3.2$
- **[C]** $b_1 = -4.3, s = 3.1$

3. What is the correlation between $Y2$ and X? [3 points]

- **[A]** -0.71
- **[B]** -0.97
- **[C]** -0.26

4. Which of the following is the correlation (R) and residual standard error (s), for regression $Y1$ on X? [3 points]

- **[A]** $R = 0.988, s = 0.5$
- **[B]** $R = 0.707, s = 0.97$
- **[C]** $R = 0.261, s = 0.1$

5. What is the residual standard error s for $Y3$? [3 points]

- **[A]** 2.05
- **[B]** 0.49
- **[C]** 3.50

Problem 7: Envelope game. [10 points]

At the end of BUS41000 class, Professor L decides to reward Alice for her hard work. How much reward she can get depends on her probability skills. Professor L places two checks (one check is \$30, the other is \$70) into two envelopes. Note Alice has no idea about the value of the checks.

1. First, Alice decides to pick one envelope randomly. How much is her reward, in expectation? [2 points]

2. Suppose that the rule is changed slightly: Alice is allowed to choose one envelope, open it, and review the value of the check. Then she can decide whether to stick with the opened envelope or to swap to the other one.
   Alice recalls that one could use a randomized strategy to win more money. Here is Alice's idea: she will employ a normal distribution $X \sim N(50, 10^2)$ to help her (using R/Excel)! Let us denote the check's value (inside the envelope she just opened) as x. The randomized strategy is: she will keep the check with probability $P(X < x)$ and swap to the other envelope with probability $1 - P(X < x)$. Using this strategy, how much is Alice's reward, in expectation? How much more money she is going to get compared to sub-problem 1? [8 points]

Problem 8: COVID test messed up [10 points]

20 people took a COVID test. However, the doctor was so careless that she/he forgot to label the test. She/He had to give back all the test results randomly, with one unique sample to each person. In particular, each person gets her/his own test result with probability 1/20. Let X be the *number of people who get their own test results.*

1. What is the average (or expected value) of X? [7 points]

2. What is the variance of X? [3 points]

Extra Page for Calculations