Business Statistics Midterm Exam

Fall 2018: BUS41000

This is a closed-book, closed-notes exam. You may use any calculator.

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."

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 formula:

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

  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

$$P(a \leq X \leq b) = P(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma}).$$

Problem 1: Who's to blame?

In manufacturing a new generation of iPhone, Apple buys a particular kind of microchip from 3 suppliers: 30% from Qualcomm, 50% from Intel, and 20% from Samsung.

Apple has extensive histories on the reliability of the chips and knows that 3% of the chips from Qualcomm are defective, 4% of the Intel are defective and 5% from Samsung are defective.

1. What is the probability that Apple chooses the chip from Intel and that the chip is defective? [5 points]

$$P(defectiveandIntel) = P(defective \mid Intel)P(Intel) = 0.04 \times 0.5 = 0.02$$

2. Write out the joint probability table of two random varibles: Brand of the chip (Brand = {Qualcomm, Intel, Samsung}), and whether the chip is defective (Defect = {Yes, No}). [5 points]

Table	Qualcomm	Intel	Samsung
Defect	$0.03 \times 0.3 = 0.009$	$0.04 \times 0.5 = 0.02$	$0.05 \times 0.2 = 0.01$
Not defect	$0.97 \times 0.3 = 0.291$	$0.96 \times 0.5 = 0.48$	$0.95 \times 0.2 = 0.19$

3. In testing a newly assembled iPhone, Apple found the microchip to be defective. Which provider is most likely to blame? (Justify your answer) [10 points]

$$P(Qualcomm \mid defective) = \frac{0.03 \times 0.3}{0.03 \times 0.3 + 0.04 \times 0.5 + 0.05 \times 0.2} = \frac{0.009}{0.039} = 0.23$$

$$P(Intel \mid defective) = \frac{0.04 \times 0.5}{0.03 \times 0.3 + 0.04 \times 0.5 + 0.05 \times 0.2} = \frac{0.02}{0.039} = 0.51$$

$$P(Samsung \mid defective) = \frac{0.05 \times 0.2}{0.03 \times 0.3 + 0.04 \times 0.5 + 0.05 \times 0.2} = \frac{0.01}{0.039} = 0.26$$

Intel has highest probability given the chip is defective. It's most likely to blame.

Problem 2: Google

Google is testing a new algorithm on personalized recommendation based on state-of-the-art deep learning research. Here is the data from the experiment: for each experiment, there are only two outcomes, success or failure.

Algorithm	current	new
success	1643	1920
failure	857	580

Is the new algorithm better? Justify your answer using either hypothesis testing, or confidence interval. [20 points]

Success rate of current method

$$\frac{1643}{1643+657} = 0.6572$$

Success rate of new method

$$\frac{1920}{1920+580} = 0.768$$

Difference of success rate

$$0.6572 - 0.7372 = 0.1108$$

Standard deviation of the difference of success rates

$$s_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = \sqrt{\frac{0.6572(1-0.6572)}{2500} + \frac{0.768(1-0.768)}{2500}} = 0.0127$$

Confidence interval. The confidence interval of difference at 95% level

$$0.1108 \pm 2 \times 0.0127 = [0.0854, 0.1362]$$

The confidence interval is postive and doesn't contain 0. So we conclude that the new approach is significantly better at 95% level.

Hypothesis testing. Null hypothesis is $\hat{p}_{new} - \hat{p}_{current} = 0$.

$$\frac{0.1108}{0.0127} = 8.72 > 2$$

reject the null hypothesis at 95% level.

Note that because of rounding errors, some of you might get non significant result. I give full credict if your equations are correct.

Problem 3: To build or not to build?

You live in a house that is somewhat prone to mud slides. In the coming rainy season there is a 5% chance of a mud slide occurring and you estimate that a mud slide would do $400,000 in damages.

Before the rainy season you are considering building a retaining wall that would potentially stop the damages from a mud slide. The wall costs $18,000 to build and, if the slide occurs, the wall will hold with 80% probability.

1. If you build the wall, what is the probability that you will lose $400,000 in damages this coming raining season? [10 points]

If build the wall,

$$P(damage) = P(damage \mid mudslide) \times P(mudslide) = 0.2 \times 0.05 = 0.01$$

2. Considering the potential damages on average (in expectation), should you build the wall? (Justify your answer) [5 points]

If build the wall, expected cost

$$E[cost] = 18,000 + 400,000 \times 0.01 = 22,000$$

If not build the wall, expected cost

$$E[cost] = 400,000 \times 0.05 = 20,000$$

So in terms of expectation, do not build the wall

3. Now consider two rainy seasons. Suppose (1) given the wall holds in the first season, it can still function the same way in the second season; (2) if the wall collapses in the first season, you won't build a new wall for the second season. Should you build the wall in the begining? Please justify your answer. (Hint: you may use the probability tree approach. This question might be harder, you can skip and revisit later after finishing other problems) [10 points]

If you do not build the wall. The expected loss is $400,000 * 0.05 * 2 = 40,000$

If you build the wall, list all possible cases of 2 seasons

1. No mud, no mud Probability is $0.95 * 0.95 = 0.9025$. Expected loss is 0
2. Mud, no mud Probability is $0.95 * 0.05 = 0.0475$. Expected loss is $400,000 * 0.2 = 80,000$ because 20% probability that the wall cannot hold.
3. No mud, mud Probability is $0.95 * 0.05 = 0.0475$. Expected loss is $400,000 * 0.2 = 80,000$ because 20% probability that the wall cannot hold.
4. Mud, mud Probability is $0.05 * 0.05 = 0.0025$. Three sub-cases for this situation
   a. hold, hold Loss is 0 with probability $0.8 * 0.8 = 0.64$.
   b. hold, does not hold probability is $0.8 * 0.2 = 0.16$. Loss is $400,000$
   c. does not hold probability is 0.2. Loss is $400,000 * 2 = 800,000$

So the expected loss of "mud, mud" case is $0.16 * 400,000 + 0.2 * 800,000 = 224,000$. The total expected loss of all 4 cases is

$$0.9025 * 0 + 0.0475 * 80000 + 0.0475 * 80000 + 0.0025 * 224000 = 8160$$

Consider cost of a wall, the total expected expenses is $18,000 + 8160 = 26160 < 40000$. We should build the wall.

Problem 4: Breaking bad

Two chemists working for a chicken fast food company, have been producing a very popular sauce. Let's call them Jesse and Mr. White. Gus, their boss, is tired of Mr. White's negative attitude and is thinking about "firing" him and keeping only Jesse on payroll. The problem, however, is that Mr. White seems to produce a higher quality sauce whenever he is in charge of production if compared to Jesse. Before making a final decision, Gus collected some data measuring the quality of different batches of sauce produced by Mr. White and Jesse. The results, measured on a quality scale, are listed below:

Table	average	std. deviation	sample size
Mr. White	97	1	7
Jess	94	3	10

1. Based on this data, can we tell with confidence which one is the better chemist? [15 points]

Method 1

$$97 - 94 \pm 2\sqrt{\frac{1^2}{7} + \frac{3^2}{10}} = 3 \pm 2 \times 1.0212 = [0.96, 5.04]$$

At 95% confidence level, Mr. White is better.

Method 2

$$97 \pm 2\sqrt{\frac{1^2}{7}} = [96.24.97.75]$$

$$94 \pm 2\sqrt{\frac{3^2}{10}} = [92.10, 95.89]$$

At 95% confidence level, the two confidence intervals do not overlap, so we conclude that Mr. White is better.

2. Gus wants to keep the mean quality score of the sauce above 90. In this case, can he get rid of Mr. White, i.e., is Jesse good enough to run the sauce production? [15 points]

$$94 \pm 2\sqrt{\frac{3^2}{10}} = [92.10, 95, 89] > 90$$

Since the 95% confidence interval is higher than 90, he can.

Problem 5

I am trying to build a portfolio composed of SP500 and Bonds. Assume $SP500 \sim N(11, 19^2)$ and $Bonds \sim N(4, 6^2)$.

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? [10 points]

$$11^2 = 0.5^2 \times 19^2 + 0.5^2 \times 6^2 + 2 \times 0.5 \times 0.5 \times \text{Cov}$$

So the covariance is

$$\text{Cov} = 43.5$$

$$\text{Corr} = 43.5/19/6 = 0.382$$

2. Using the covariance you calculated in sub-problem 1, can you show me why

$$sd(0.8SP500 + 0.2Bonds) = 15.697$$

   In addition, which portfolio is better: a 80-20 split between SP500 and Bonds, or 50-50 split? Justify your criteria for a comparison. [10 points]

$$sd(0.8SP500 + 0.2Bonds) = 0.8^2 \times 19^2 + 0.2^2 \times 6^2 + 2 \times 0.8 \times 0.2 \times 43.5 = 15.697$$

Sharpe ratios

$$50/50 \text{ portfolio } \frac{0.5 \times 11 + 0.5 \times 4}{11} = 0.68$$

$$80/20 \text{ portfolio } \frac{0.8 \times 11 + 0.2 \times 4}{15.697} = 0.61$$

50/50 portfolio has higher sharpe ratio, better

3. Which one of the two portolios considered in the above sub-problem 2 has a larger probability of delivering a positive return? [5 points]

$$P(\text{portfolio5 } 50/50 \ > 0) = P(Z > \frac{0-(0.5\times 11 + 0.5\times 4)}{11}) = P(Z > -0.68)$$
$$P(\text{portfolio5 } 80/20 \ > 0) = P(Z > \frac{0-(0.8\times 11 + 0.2\times 4)}{15.697}) = P(Z > -0.61)$$

50-50 has higher probability to generate positive returns

Problem 6: B-17 flying fortress and Wald

During a period in World War II, the U.S. Army Air Forces (AAF) would send over 300 B-17 bombers daily to raid factories in Germany. These missions, originating in the U.K., were very dangerous and, in the peak of the campaign, the return probability for a B-17 crew was only 84%.

In trying to reduce the probability of a failed mission, a Navy statistician (Abraham Wald) was put in charge of studying the damage patterns in the B-17's that successfully made back from a mission. His ultimate goal was to decide where to add extra armor in the planes (you couldn't just add heavy armor everywhere, as the planes would be too heavy to fly!). Wald was able to learn that if a plane made back from a mission there was a 67% probability they were shot in the fuselage, 15% in the fuel systems, 10% in the cockpit area and 8% in the engines.

From experiments, Wald was also able to deduce that during combat, a B-17 would be shot in the fuselage with 58% probability, in the fuel systems with 14%, in the cockpit area 14% and engine 14%.

Based on this information what was Wald's recommendation to the AAF, i.e., if they had to choose one area of the plane, where should they add extra armor to the B-17's? (Hint: Wald suggested to improve on the weakest area — the area with the smallest $P(\textit{success return}|\textit{area being shot})$, the probability of success return condition on that area is shot.) [20 points]

$$\begin{aligned} P(\text{success return} \mid \text{area being shot}) &= \frac{P(\text{success return\&area being shot})}{P(\text{area being shot})} \\ &= \frac{P(\text{area being shot} \mid \text{success return})P(\text{success return})}{P(\text{area being shot})} \end{aligned}$$

So

$$\begin{aligned} \text{fuselage}: & \frac{0.67 \times 0.84}{0.58} = 0.97 \\ \text{fuel}: & \frac{0.15 \times 0.84}{0.14} = 0.90 \\ \text{cockpit}: & \frac{0.1 \times 0.84}{0.14} = 0.60 \\ \text{engines}: & \frac{0.08 \times 0.84}{0.14} = 0.48 \end{aligned}$$

So engines are most weak part, needs extra armor.

Problem 7

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

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

1. Based on these results, what is the probability of the SP500 returning less than 20% next year? In addition, give a 95% prediction interval for next year's SP500 return. [15 points]

$$P(X < 0.2) = P\left(Z < \frac{0.2 - 0.072}{0.13}\right) = P(Z < 0.98) \approx 0.84$$

The predictive interval is

$$[7.2 - 2 * 13, 7.2 + 2 * 13] = [-18.8, 33.2]$$

2. Use a 99% confidence interval, to test the hypothesis that the expected return (true mean) of the SP500 is equal to 4% a year. [5 points]

The standard deviation is $\sqrt{13^2/116} = 1.21$, confidence interval is $7.2 \pm 3 \times 1.21 = [3.58, 10.82]$. We see that 4% is inside the interval so cannot reject the null hypothesis.

3. 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 describes the statistical meaning? [10 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) If we recollect datasets and build confidence intervals many times, 95% of the times, these intervals will cover the true μ.
- (c) We are 95% confident that the true mean is in the interval $[4.7, 9.6]$.

(b)

Statistical meaning of 99% confidence interval: If we simulate data and calculate confidence interval many times, on average, 99% of them can cover the true mean. Note that the true mean is fixed and never change, but the confidence interval moves around in different simulations. About 99% confidence intervals can cover the trun mean.

Common incorrect interprtations. 1. With probability 99%, the true mean lies in the confidence interval. 2. We are 99% confident that the true mean is in the confidence interval.

Problem 8

Assume the linear model: $Y|X \sim N(5 - 2X, 3^2)$.

1. Give a 95% prediction interval for Y given $X = 2$. [5 points]

$$5 - 2 \times 2 \pm 2 \times 3 = [-5, 7]$$

2. What is the mean of the distribution for Y when $X = 3$? How about the variance? [5 points]

Mean: $5 - 2 \times 3 = -1$, variance 9

3. What is the probability $P(Y > 7|X = 0.5)$, given that $X = 0.5$? [5 points]

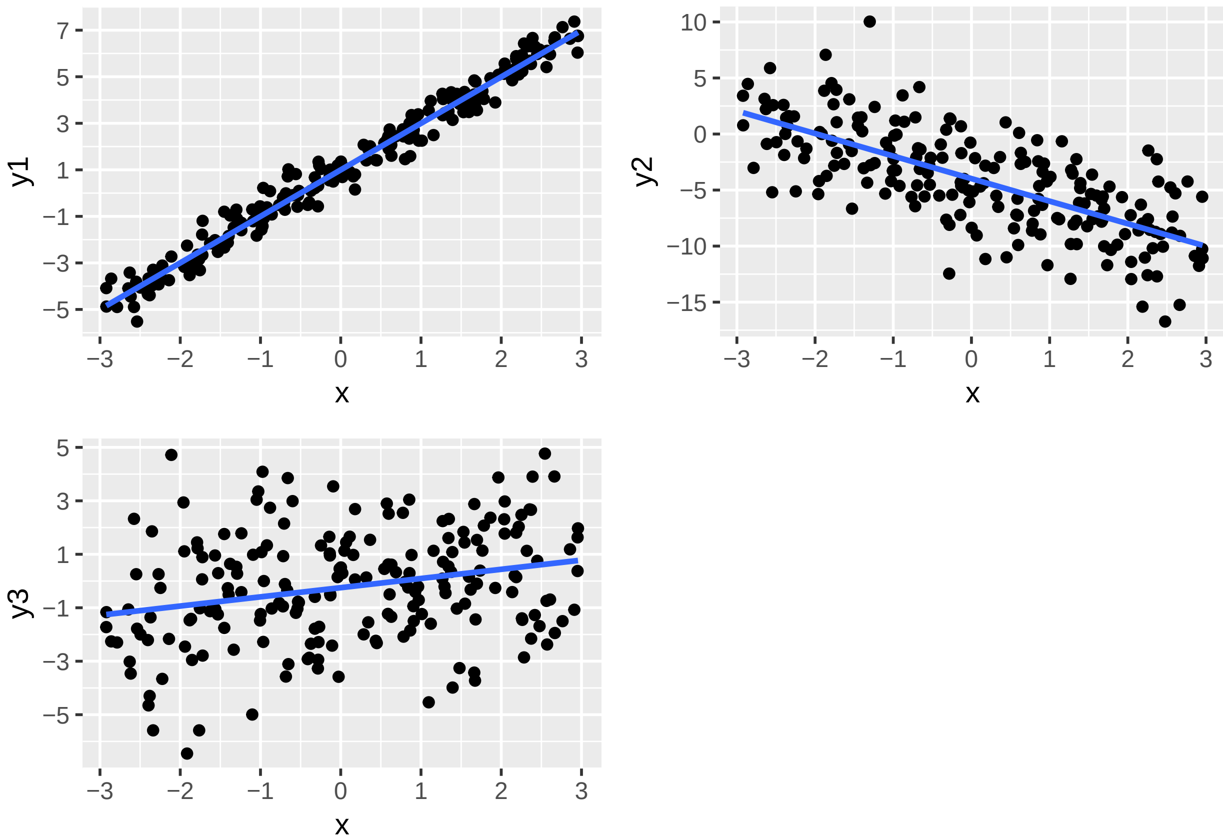
$$P(Y > 7) = P(5 - 2 \times 0.5 + \epsilon > 7) = P(\epsilon > 3) = P\left(Z > \frac{3-0}{3}\right) = P(Z > 1) = 0.16$$

4. What is the probability of $P(Y > 10|X = 0.5)$ and $P(Y = 4|X = 0.5)$, given that $X = 0.5$? [5 points]

$$P(Y > 10 \mid X = 0.5) = P(5 - 2 \times 0.5 + \epsilon > 10) = P(\epsilon > 6) = P\left(Z > \frac{6-0}{3}\right) = P(Z > 2) = 0.025$$

$$P(Y = 4 \mid X = 0.5) = 0$$

Problem 9



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, $Y = b_0 + b_1 X + e, e \sim N(0, s^2)$.

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? [5 points]

- (a) $b_1 = 0.34, b_0 = -0.24$ correct
- (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? [5 points]

- (a) $b_1 = -0.9, s = 6.3$
- (b) $b_1 = -2.0, s = 3.2$ correct
- (c) $b_1 = -4.3, s = 3.1$

3. Which of the following is the correlation (R) and residual standard error (s), for regression $Y1$ on X? [5 points]

- (a) $R = 0.988, s = 0.5$ correct
- (b) $R = 0.707, s = 0.97$
- (c) $R = 0.261, s = 0.1$

4. What is the correlation between $Y2$ and X? [5 points]

- (a) -0.71 correct
- (b) -0.97
- (c) -0.26

5. Using all the information provided so far, give a rough approximation for the 99% prediction interval for $Y1$ given $X = 0$. [5 points]

$$Y1 \mid X = 0 \sim N(1, 0.5)$$

The confidence interval is

$$[1 - 3 \times 0.5, 1 + 3 \times 0.5] = [-0.5, 2.5]$$

6. What is the residual standard error s for $Y3$?
   - (a) 2.05 correct
   - (b) 0.49
   - (c) 3.20

   In addition, give an approximation for $P(Y3 > 4|X = 3)$. [5 points]

Problem 10: Olympics medal and GDP

Using data from Beijing 2008 and London 2012 I run a regression trying to understand the impact of **GDP** (gross domestic product measured in trillions of US$) on the **total number of medals** won by a country in the summer Olympics. The results are

$$\textbf{TotalMedals} = 5.17 + 8.36 \times \textbf{GDP} + \epsilon, \text{ where } \epsilon \sim N(0, 11^2).$$

The following table shows the total medal count for a few countries in Rio 2016 Olympics along with their current GDP:

Country	Total Medals	GDP (in US$ trillions)
US	121	18.5
China	70	11.3
Brazil	19	1.6
UK	67	2.8

Using the results from the simple linear regression, answer the following questions:

1. From the results, what is your prediction (best guess) for the total number of medals for the U.S. in the Rio 2016 Olympics? [5 points]

$$5.17 + 8.36 + 121 = 159.83$$

2. Conditional on their GDP, which of these countries performance in the Rio 2016 is not surprising? Why? (Hint: use prediction interval) [5 points] Construct confidence interval of each contry

$$\text{US } [159.83 - 1.96 \times 11, 159.83 + 1.96 \times 11]$$

$$\text{China } [99.64 - 1.96 \times 11, 99.64 + 1.96 \times 11]$$

$$\text{Brazil } [18.55 - 1.96 \times 11, 18.55 + 1.96 \times 11]$$

$$\text{UK } [28.58 - 1.96 \times 11, 28.58 + 1.96 \times 11]$$

Brazil is not a suprise because it has 19 medals, within the confidence interval.

3. Conditional on their GDP, which of these countries looks like a clear overachiever? [5 points]

UK, gets 67 medals, much higher than the upper bound of confidence interval.

4. Based on the predictions from this model, who did better, China or the U.S.? (Hint: use how many standard deviation away the performance is from predicted value) [5 points]

Calculate Z scores of each contry

$$\text{US } \frac{121 - 160}{11} = -3.55$$

$$\text{China } \frac{70 - 100}{11} = -2.73$$

China does better. US is far left at the normal tail.

(Bonus) Problem 11: Envelope game

At the end of BUS41000 class, your Professor decided to reward you for your hard work, but also to test your probability skills. He placed two checks (one check is \$30, the other is \$70) into two envelopes. Note you have no idea about the value of the checks at all, nor the range of the checks.

1. You decided to randomly pick one envelope, how much is your reward, in expectation? [2 points]

$$30 \times 0.5 + 70 \times 0.5 = 50$$

2. Suppose that the rule changed slightly: You are allowed to choose only one envelope, open it, review the check value. And then decide whether to stick with the opened envelope, or to swap to the other envelope. You recalled that Professor said one can use probability to get more money. Suppose you are going to do the following: draw a random number X using R/Excel from a normal distribution $X \sim N(50, 10^2)$, then compare this X with the value of the check you opened, and only to keep the check if and only if its value is larger than X, otherwise swap to the other emvelope. Using this strategy, how much is your reward, in expectation? How much more money you are going to get compared to sub-problem 1? [6 points]

   Suppose Y is value of check you first pick

Table	$X > Y$	$X < Y$
$Y = 30$	70, with probability 0.5×0.975	30, with probability 0.5×0.025
$Y = 70$	30, with probability 0.5×0.025	70, with probability 0.5×0.975

So the expectation is

$$70 \times 0.5 \times 0.975 + 30 \times 0.5 \times 0.025 \times 2 = 69$$

and $69 - 50 = 19$

3. Suppose you use the same strategy as in the sub-problem 2, but the Professor placed two checks, $70 and $80, respectively. How much is your expected reward? [4 points]

a. 50% probability open with 70,

$$P(X > 70) = P(Z > \frac{70 - 50}{10}) = 0.025$$

So expected return is

$$0.025 \times 80 + (1 - 0.025) \times 70 = 70.25$$

b. 50% probability open with 80,

$$P(X > 80) = P(Z > \frac{80 - 50}{10}) = 0.005$$

So expected return is

$$0.005 \times 70 + (1 - 0.005) \times 80 = 79.95$$

Therefore the total expected return is

$$0.5 \times 70.25 + 0.5 \times 79.95 = 75.1$$