MSIN0096 Homework 2
Due Nov 1st 3pm

1. (10pt) During the last flu season, 425 out of 3,542 patient in a hospital diagnosed with the flu presented severe symptoms. During the same flu season, a new anti-flu drug was being tested. The drug was given to 526 people with the flu and only 25 of them developed severe symptoms. Based only on this information, can you conclude that the drug is a success and at which level?

   Now, it turns out that the people who received this drug were all undergraduate students. Can you infer that the new anti-flu drug alleviate the occurrence of sever symptoms? What are some potential factors that may influence whether someone develop severe symptoms or not?

2. (10pt) This is a continuation of Questions 5(d) and 5(e) in Assignment 1.

   Please download the data q2_data.csv (the same data used in Q5 in Assignment 1) from moodle, which contains 100 random numbers drawn from $N(\mu, 2.5^2)$, where μ is unknown.

   In 5(d), you were asked to conduct a test $H_0 : \mu = 5, H_1 : \mu \neq 5$ at 5% level. If you are told that the random sample actually is drawn from$N(5.2, 2.5^2)$, 5(e) asks you to compute the probability of Type II error. Now please use **bootstrap** to estimate the probability of Type II error of the hypothesis test above. Compare the probability of Type II error computed in 5(e) in Assignment 1, what can you find? (There is no need to redo the hypothesis testing $H_0 : \mu = 5$,$H_0 : \mu \neq 5$ and you can copy the results of 5(d) from your answer or my solution.)

3. (a) (6pt) Assume a data set contains 5 observations $(x_1, x_2, x_3, x_4, x_5)$. $(x_1^*, x_2^*, x_3^*, x_4^*, x_5^*)$ is a bootstrap sample of $(x_1, x_2, x_3, x_4, x_5)$. What is the probability that x_1 is not in $(x_1^*, x_2^*, x_3^*, x_4^*, x_5^*)$? What is the probability that x_3 is not in $(x_1^*, x_2^*, x_3^*, x_4^*, x_5^*)$?

   (b) (6pt) Assume a data set contains 100 observations $(x_1, x_2, \cdots, x_{100})$ and $(x_1^*, x_2^*, \cdots, x_{100}^*)$ is a bootstrap sample. What is the probability that x_i is not in the bootstrap sample? On average, how many observations of $(x_1, x_2, \cdots, x_{100})$ are included in the bootstrap sample $(x_1^*, x_2^*, \cdots, x_{100}^*)$?

4. (10pt) An oil company has purchased an option a field in North Sealand. Preliminary geological studies found out the following probabilities of finding oil in the field:

$$p(\text{High oil reserves}) = 0.5$$

$$p(\text{Low oil reserves}) = 0.2$$

$$p(\text{No oil reserves}) = 0.3$$

   After purchasing the option, the company decided conducted a soil test. They found certain type of soil (denoted as "A") on the seabed.

   According to previous drilling data, the probabilities of finding this particular type of soil are as follow:

$$p(\text{soil type ``A''}|\text{High oil reserves}) = 0.25$$

$$p(\text{soil type ``A''}|\text{Low oil reserves}) = 0.6$$

$$p(\text{soil type ``A''}|\text{No oil reserves}) = 0.15$$

(a) Given the information from the soil test what is the probability the company will not find oil in this field?

(b) Before deciding to drill in the land the company has to perform a cost/benefit analysis of the project. They know drilling and operation cost of this field will be \$50,000,000. Under current oil prices, the value of high oil reserves in this field will be \$100,000,000 and the value of low oil reserves in this field will be \$20,000,000.

Should the company exercise the option, ie, should the company drill?

5. (11pt) On average there are 164 rainy days in London per year. When it actually rains, the weather forecast is correct 90% of the time. When it doesn't rain, the weatherman incorrectly forecasts rain 10% of the time. The weather forecast says that it is going to rain tomorrow. What is the probability that tomorrow will be rainy?

6. (8pt) Gmail categorizes incoming emails into 3 groups, "Primary", "Social" and "Promotion". Suppose 70% of my emails are primary, 20% are social and 10% are promotion. Meanwhile, I searched the keyword "sales" in 3 groups and found that 5% primary emails, 30% social emails and 95% promotion emails include this keyword respectively.

   Suppose an incoming email contains the keyword "sales", what is the probability of being categorized as "primary"?

7. (10pt) There is a coin and the probability of getting head by flipping this coin is θ. Assume that your prior about θ is $p(\theta) = \frac{1}{9}$, for $\theta = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$.

   Now this coin is tossed 50 times and get 16 heads. Please update your prior based on the data and compute the posteriors. Based on the posteriors, what are the two most likely values for θ?

8. (10pt) Target is a top US retailer. By analyzing consumer's purchase history, Target data scientist could figure out if she is pregnant. Following is a true story cited from the New York Time article `http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=2&hp=&pagewanted=all`

   A man walked into a Target outside Minneapolis and demanded to see the manager. He was clutching coupons that had been sent to his daughter, and he was angry, according to an employee who participated in the conversation.

   "My daughter got this in the mail!" he said. "She's still in high school, and you're sending her coupons for baby clothes and cribs? Are you trying to encourage her to get pregnant?"

   The manager didn't have any idea what the man was talking about. He looked at the mailer. Sure enough, it was addressed to the man's daughter and contained advertisements for maternity clothing, nursery furniture and pictures of smiling infants. The manager apologized and then called a few days later to apologize again.

   On the phone, though, the father was somewhat abashed. "I had a talk with my daughter," he said. "It turns out there's been some activities in my house I haven't been completely aware of. She's due in August. I owe you an apology."

   In 2014, there were 3,988,076 new births in the US. In that year, the US female population is 125.9 million. Assume all pregnancy leads to birth and assume all female population is in reproductive age. Please calculate the prior p(a woman is pregnant).

Target identified that purchasing 25 specific products together can indicate pregnancy. Suppose data scientists at Target found that among pregnant female customers, 95% of them purchase these 25 products all together and among non-pregnant female customers, 0.5% of them purchase these 25 products together. Using Bayesian inference to calculate the probability that a woman is pregnant if Target observes that she purchases all these 25 products.

9. (10pt). Download data set q9_data.csv from moodle and use OLS to estimate coefficients b_0, b_1 , b_2 and b_3 in the following equation. e is the disturbance.

$$Y = \frac{b_0 exp(b_1 X_1 + e)}{X_2^{b_2}(1 + X_3)^{b_3}}$$

10. (a) (2pt) We estimate the following equation to explore different user group's contribution to a game app's revenue.

$$rev = \beta_0 + \beta_1 age_1 + \beta_2 age_2 + \beta_3 age_3 + \beta_4 age_4 + \epsilon$$

where rev is the revenue of this game app. age_1 is the share of users younger than 18 years old (inclusive). age_2 is the share of users between 19-35 years old. age_3 is the share of users between 36-55 years old. age_4 is the share of users older than 56 (inclusive). Can you estimate this model? Please justify your answer.

(b) (3pt) Following previous questions, we estimate the equation as below

$$rev = \beta_0 + \beta_1 age_1 + \beta_2 age_2 + \beta_3 age_3 + \epsilon$$

Student A says β_1 measures how much the revenue will change as the share of teenage users increases by 1 percentage, if user shares of age 19-35, 36-55 and above 56 all keep unchanged. Do you agree with student A?

(c) (4pt) Suppose a student estimated the equation above, but she/he made mistakes on generating age_1, age_2 and age_3. Instead variables this student created are AGE_1 the share of users younger than 18 years old (inclusive), AGE_2 the share of users younger than 35 years old (inclusive) and AGE_3 the share of users younger than 55 years old (inclusive). This student gets this estimation result

$$\hat{rev} = 0.87 + 1.20 AGE_1 + +1.08 AGE_2 + 0.67 AGE_3$$

Without re-estimate another regression, write down estimation result of the original regression equation $rev = \beta_0 + \beta_1 age_1 + +\beta_2 age_2 + \beta_3 age_3 + \epsilon$.