MSIN0096 Homework 1
Due Oct 27st 10am (London Time)

1. (10pt) London Marathon is a very popular marathon. People who would like to run London Marathon need to enter a ballot first. Each year, around 17,000 runners are randomly drew from the ballot applicants. The table below shows the number of people entered the ballet and the number of places offered from 2018-2020.

	Number of Places Offered	Number of Ballot Applicants
2018	17,000	386,050
2019	17,000	414,168
2020	17,000	457,861

   (a) (5pt) What is the probability that an applicant failed to get a place in all three years?

   (b) (5pt) Given that an applicant failed to get a place in 2018 and 2019, What is the probability that this applicant gets a place in 2020?

2. (10pt) A jar contains one red ball, one green ball and one blue ball. Each time, one ball is randomly drew without replacement (the ball is not put back into the jar). Given that the first draw is red, what is the probability that the second draw is green?

3. (12pt) An airline charges the following baggage fees: £30 for the first bag and £50 for the second bag. Suppose 60% of passengers have no checked luggage, 28% have one piece of checked luggage and 12% have two pieces. Assume that no people check more than two bags.

   (a) (6pt) Assume X is the baggage revenue per passenger. Find out the probability density function (PDF) and cumulative density function (CDF) of the random variable X.

   (b) (6pt) Compute the average revenue per passenger, and compute the corresponding standard deviation.

4. (13pt) Assume that $X \sim N(0, 3^2)$, Compute following probabilities

   (a) (2pt) $P(X > 3)$

   (b) (2pt) $P(|X| > 3)$

   (c) (3pt) Find x, such that $P(|X| > x) = 0.05$.

   (d) (6pt) For another random variable $Y \sim N(10, 3^2)$, find y such that $P(|Y| > y) = 0.05$. You may not be able to find the precise y using the method learned in lectures and any reasonable approximation will be sufficient.

5. (12pt) A manufacturer of booklets packages them in boxes of 100. It is known that on average, the booklets weigh one ounce, with a standard deviation of 0.05 ounce. The manufactures is interested in calculating

$$P(\text{100 booklets weigh more than 100.4 ounces})$$

   a number that would help detect whether too many booklets are being put into a box. Use Central Limit Theorem to calculate an approximate value of this probability.

6. (13pt) 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 to conduct 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) (5pt) Given the information from the soil test what is the probability the company will not find oil in this field?

(b) (8pt) 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?

7. (10pt) Gmail categorizes incoming emails into 3 groups, "Primary", "Social" and "Promotion". Assume 70% of my emails are primary, 20% are social and 10% are promotion. Meanwhile, I searched the keyword "sales" in 3 groups and found out 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"?

8. (10pt) Target is a big US retailer. By analyzing female consumer's purchase history, data scientists at Target could predict whether she is pregnant or not. Following is a true story cited from New York Time.[1]

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.

[1] `http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?_r=2&hp=&pagewanted=all`

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 2012, there were 3,988,076 new births in the US and US female population was 125.9 million. Assume all pregnancy leads to birth and all female population is in reproductive age. Please use this information to approximately calculate the prior, the probability of pregnancy in female population.

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) To track how users are playing a newly launched game app, we random select 15 players and record how many minutes they spend on this game in one day, which are listed as below.

$$8.6; 9.4; 7.9; 6.8; 8.3; 7.3; 9.2; 9.6; 8.7; 11.4; 10.3; 5.4; 8.1; 5.5; 6.9$$

Please construct a 95% confidence interval of average time consumers spend with this game daily.