Assignment 4
Due on 31st Oct, 11 p.m.

1. **Gene Expression Distribution:** There are about $30,000$ genes in the human genome. You have the RNA-seq data from a particular cell. It has the expression values of all the genes in that cell. Now you want to have a look at the distribution of these $30,000$ values.

   1.1) **(5 points)** Without consulting any person in the field of genomics, what do you think the shape of the distribution will look like? You only need to draw roughly the shape by hand. Please also have a brief description about why you think the shape of the distribution will look like what you have drawn.

   1.2) **(5 points)** We discussed in the lesson that the expressions of *Pou5f1* across embryonic stem cells follow a normal distribution. What do you think about the expression of another pluripotent marker gene *Nanog*? Draw roughly the shape of distribution of the expression of *Nanog* across embryonic stem cell population.

   1.3) **(5 points)** Are the distributions of **1.1)** and **1.2)** the same? What do they tell you?

2. **Weaver ants:** Weaver ants are eusocial insects of the family Formicidae. They live in trees and are known for their unique nest building behaviour where workers construct nests by weaving together leaves using larval silk. There are actually two types of weaver ants, and the body lengths of these two types of weaver ants are quite different. Suppose the mean body length of all weaver ants is 7 mm, and the standard deviation is 10 mm.

   2.1) **(5 points)** What does the shape of the population distribution of the body length of all weaver ants look like? You only need to draw the shape, roughly.

   2.2) **(5 points)** If you randomly selected 625 weaver ants, what will be the shape of the distribution of their body length look like? Again, you only need to draw the shape, roughly.

   2.3) **(5 points)** The arithmetic mean of the body length of the 625 weaver ants in **2.2)** is 7.5 mm. What is the probability of seeing the mean of the body length is less than or equal to 7.5 mm when you randomly selected 625 weaver ants ?

3. **(2.5 points) A Simple Random Sample:** The weights of 10-year-old girls are known to be normally distributed with a mean of 31.75 kg and a standard deviation of 5.90 kg. If you take a sample of 10-year-old girls, which of the following statement is correct about this sample ?

   (A) If the sample is a good representation of the population, the mean of the sample will be exactly 31.75 kg

   (B) If the mean of the sample is 38 kg, the sample is not really a random sample

   (C) Even if the sample is a good representation of the population, the sample mean can still be different from 31.75 kg

   (D) None of the above

4. **(5 points) The Concept of Population:** Suppose we want to estimate the 5-year survival rate of women who are initially diagnosed as having breast cancer at the ages of $45-54$ and who undergo radical mastectomy at this time. In this case, which of the following are true about the population? Tick all that are appropriate.

   [ ] It is better to think of our population as an abstract concept

   [ ] The population is all women who have ever had a first diagnosis of breast cancer when they were $45-54$ years old

   [ ] The population is all women whoever will have such a diagnosis in the future when they are $45-54$ years old

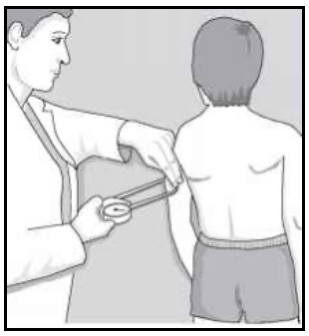
   [ ] The population is all women who have ever had a first diagnosis of breast cancer when they were $45-54$ years old, or whoever will have such a diagnosis in the future when they are $45-54$ years old, and who receive radical mastectomies

   [ ] The population is effectively infinite

5. **(5 points) A Practical Simple Random Sampling Strategy:** Suppose we want to study how effective a hypertension treatment program is in controlling the blood pressure of its participants. We have a list of all 1000 participants in the program, but because of limited resources only 20 can be surveyed. We would like the 20 people chosen to be a simple random sample from all participants in the program. Describe how should we select this random sample? You

can choose whatever computer program you like for this purpose. If you don't know any, check the `RAND` function in **Excel**.

6. **Triceps skinfold thickness**: The thickness of a skinfold is a very useful measurement for nutritional assessment and the prediction of the total body fat mass. Of all skinfold measurements, the triceps skinfold is the most reliable one, because oedema is not often seen in the upper arm. Suppose the triceps skinfold thickness in the whole population follows a normal distribution with a mean of 1.35 cm and a variance of 0.25 cm^2.



   6.1) **(5 points)** If we select a simple random sample of 25 people from the whole population and measure their triceps skinfold thickness, what is the probability of observing the sample mean is less than or equal to 1.05 cm?

   6.2) **(5 points)** A doctor selects a sample of 25 people from the population, but they all suffer from chronic obstructive pulmonary disease (COPD). The mean triceps skinfold thickness of them is 0.92 cm. What can you say from this data? This is an open question, and you can write whatever you think is reasonable.

7. **Decrease Manufacturing Variability:** Suppose that weights of bags of potato chips coming from a factory follow a normal distribution with mean 362 g and standard deviation 20 g.

   7.1) **(5 points)** In general, what percentage of the bags weigh less than 330 g ?

   7.2) **(5 points)** If a simple random sample of size $n = 100$ is taken, what is the probability that the mean of this sample falls between 360 g and 363 g?

   7.3) **(10 points)** The manufacturer is not satisfied with his current production pipeline. He thinks the variation of the weights of bags is too high. To reduce the variation, he has improved his production pipeline. After the improvement, the bag weights follow a normal distribution with the same mean as before, but only 1% of the bags weigh less than 330 g, what is the standard deviation of the bag weight distribution after the improvement ?

8. **A Hypothetical Linear PDF:** The continuous random variable $\boldsymbol{X}$ follows the following PDF with parameter $-1 < \theta < 1$:

$$\mathbb{f}_{\boldsymbol{X}}(x) = \begin{cases} (1-\theta) + 2\theta x & \text{, when } 0 < x < 1 \\ 0 & \text{, otherwise} \end{cases}$$

8.1) **(5 points)** Is it a valid probabilistic model, *i.e.* does it satisfy the normalisation property? Prove your answer.

8.2) **(5 points)** Compute the CDF of $\boldsymbol{X}$.

8.3) **(5 points)** Compute the expected valued of $\boldsymbol{X}$.

We draw a sample of size n from this population, and the observations from this sample are $x_1, x_2, x_3, \cdots, x_n$. Based on this sample data, answer the following questions:

8.4) **(2.5 points)** Write the log likelihood function $\ell(\theta; x_1, x_2, \cdots, x_n)$

8.5) **(5 points)** The maximum likelihood estimate for θ from this sample is $\hat{\theta} = t$. Which of the following condition should t satisfy?

(A) $\prod_{i=1}^{n} \dfrac{2x_i - 1}{1 - t + 2tx_i} = 0$

(B) $\sum_{i=1}^{n} \dfrac{2x_i - 1}{1 - t + 2tx_i} = 0$

(C) $\prod_{i=1}^{n} \dfrac{1}{1 - t + 2tx_i} = 0$

(D) $\sum_{i=1}^{n} \dfrac{1}{1 - t + 2tx_i} = 0$

9. **(5 points) Waiting Time Between Text Messages:** Recall that the exponential probability distribution is used to model the waiting time between consecutive arrivals in a Poisson process. The PDF only has one parameter:

$$\mathbb{f}_{\boldsymbol{X}}(x) = \begin{cases} \lambda e^{-\lambda x} & \text{, when } x > 0 \\ 0 & \text{, otherwise} \end{cases}$$

where $\lambda > 0$. It seems that we can use it to describe how long you will wait before you get another text message. A person has been keeping tracking the time between getting text messages. He collected n data points for $n + 1$ messages, *i.e.* x_1 is the amount of time between the 1st and the 2nd text messages; x_2 is

the amount of time between the 2nd and 3rd text messages, ..., x_n is the time between the n-th and the $(n+1)$-th text messages. Compute the maximum likelihood estimate for λ based on the data.