Southern University of Science And Technology
School of Life Sciences
BIO210 Biostatistics
(Spring, 2022)

Assignment 5
Due on 3rd Apr, 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 pluripotency marker gene $Nanog$? Draw roughly the shape of distribution of the expression of $Nanog$ across embryonic stem cell population.

   1.3) **(10 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 cm, and the standard deviation is 10 cm.

   2.1) **(10 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 cm. What is the probability of seeing the mean of the body length is less than or equal to 7.5 cm when you randomly selected 625 weaver ants ?

3. **A simple random sample (5 points)**: 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. **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.

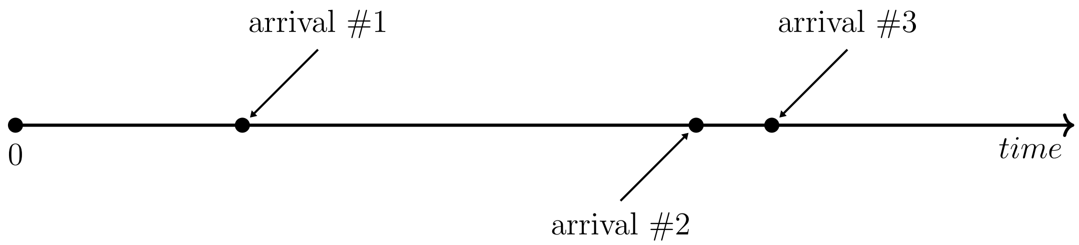
   4.1) **(5 points)** In general, what percentage of the bags weigh less than 330 g ?

   4.2) **(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 ?

   4.3) **(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 ?

5. **The Poisson process, the exponential distribution and memorylessness:** We have introduced the **binomial** random variable, which is performing n independent **Bernoulli** trials. When $n \to \infty$, we see that the binomial PMF becomes the **Poisson** PMF. When $n \to \infty$, how should we interpret n? We said that we could think of it as performing Bernoulli trials continuously in

time/space. Now, let's look at this interpretation in more details using the number of visitors to a website. Empirically, it is reasonable to assume that the number of visitors to a website follows a Poisson distribution. This Poisson PMF only has one parameter, and it is $\boldsymbol{\lambda}$ **visitors per hour**. Now let's keep tracking visitors along a time axis, starting at 0:



The visitors are the event of our interest. Therefore, when the website has a visitor, we say that "an event arrives". Recall from Lecture 11, slide 15, the Poisson distribution has the following properties:

i. The probability that a single event occurs within an interval is proportional to the length of the interval;

ii. Within a single interval, an infinite number of occurrences of the event are theoretically possible, *i.e.* not restricted to a fixed number of trials;

iii. The events occur independently both within the same interval and between consecutive intervals.

5.1) **(2.5 points)** Now, let the random variable $\boldsymbol{A}$ represent the number of arrivals in the first 15 minutes, write the PMF of $\boldsymbol{A}$.

5.2) **(2.5 points)** Let the random variable $\boldsymbol{B}$ represent the number of arrivals in a particular time interval with length of $\boldsymbol{t}$ hours. Write the PMF of $\boldsymbol{B}$.

5.3) **(2.5 points)** Now, we observed that there are 5 arrivals during the first hour. Given that has occurred and let the random variable $\boldsymbol{C}$ represent the number of arrivals in the next time interval of length $\boldsymbol{t}$, write the PMF of $\boldsymbol{C}$.

The above model described is called a **Poisson process**. Due to the independence of each arrival, the average number of arrivals for any time interval only depends on the length of the interval. What has already happened in the past does not matter.

Now let's shift our interest to time. Specifically, we focus on the waiting time between two consecutive arrivals. We have a sequence of random variables as follows:

$\boldsymbol{T_1}$ represents the time between the start (time 0) and the 1st arrival;

$\boldsymbol{T_2}$ represents the time between the 1st and the 2nd arrivals;

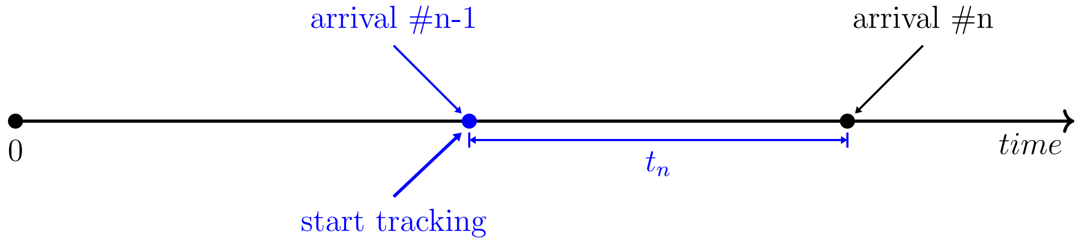
$\boldsymbol{T_3}$ represents the time between the 2nd and the 3rd arrivals;

$\vdots$

5.4) **(2.5 points)** Which of the following are correct about $\boldsymbol{T_1}, \boldsymbol{T_2}, \boldsymbol{T_3}, \cdots$ (tick all that are correct):

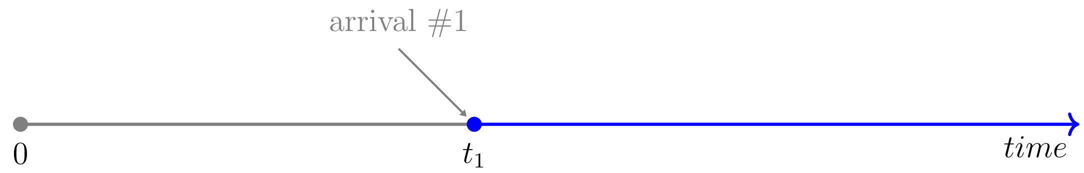
[ ] They are Poisson random variables
[ ] They are Bernoulli random variables
[ ] They are Binomial random variables
[ ] They are continuous random variables
[ ] Their distributions are the same

I'm going to tick one answer for you for the above question: their distributions are the same! Why? Think about in the following scenario:



Imagine your friend has been tracking the number of visitors from the start, and he calls you in when the $(n-1)^{th}$ visitor arrives at the website, and then you replace him and start tracking from there. Since each arrival is independent of each other, what happens in the past does not affect future arrivals. You don't really care about how many visitors have already arrived. It is a fresh start. For you, it is exactly the same ***as if*** you start tracking from the start (time 0). Therefore, all of those random variables have the same distribution. This is called **"the fresh start"** property, or **memorylessness**.

Now, to compute their probability distribution, we only need to figure out one of them. As always, we pick the simplest one. That is, $\boldsymbol{T_1}$. Now, look at the following picture:



At the time t_1, the first visitor arrives. It is easy to see that the blue part represents the event $\{\ \boldsymbol{T_1}$ takes a value greater than $t_1\ \}$. Therefore, the probability of the blue event is $P(T_1 > t_1)$.

5.5) **(2.5 points)** $P(T_1 = t_1) =$ __________.

5.6) **(2.5 points)** How to calculate the probability of the event $\{\ \boldsymbol{T_1}$ takes a value greater than $t_1\ \}$? Think about this: the event is equivalent to which of the following:

[ ] No arrivals in the time interval $[t_1, +\infty)$

[ ] No arrivals in the time interval $[0, t_1]$

[ ] Exactly one arrival in the time interval $[t_1, +\infty)$

[ ] Exactly one arrival in the time interval $[0, t_1]$

5.7) **(5 points)** Based on your choice, compute $P(T_1 > t_1)$. Express the probability using λ and t_1. **Hint:** use the Poisson PMF for the calculation.

$$P(T_1 > t_1) =$$

5.8) **(2.5 points)** Recall that the CDF of a random variable is defined as $F_{\boldsymbol{X}}(x) = P(\boldsymbol{X} \leqslant x)$. Compute the CDF of the random variable T_1:

$$F_{\boldsymbol{T_1}}(t_1) =$$

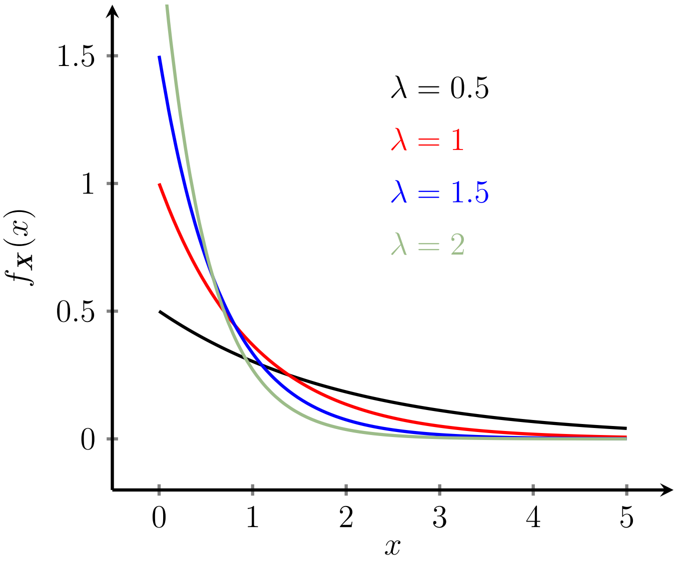
5.9) **(5 points)** Recall that the PDF of a random variable is just the derivative of its CDF. Compute the PDF of the random variable $\boldsymbol{T_1}$:

$$f_{\boldsymbol{T_1}}(t_1) =$$

Since all of $\boldsymbol{T_1}$, $\boldsymbol{T_2}$, $\boldsymbol{T_3}$, $\cdots$ have the same distribution. You just computed the PDF for all of them. You can write it in a more general form by replacing $\boldsymbol{T_1}$ with $\boldsymbol{X}$, and it should look like this:

$$f_{\boldsymbol{X}}(x) = \lambda e^{-\lambda x}$$

That is the PDF of the ***exponential distribution***. It is a very useful probability distribution to model the waiting time between independent events. It only has one parameter: $\boldsymbol{\lambda}$, which should be positive. Here, $\lambda > 0$ is often called the ***rate parameter***. Because as λ becomes larger, the time between two events becomes smaller. In the above example, larger λ means the website get visited more frequently. The shape of the distribution with different λ looks like this:



5.10) **(7.5 points)** Your friend has been keeping tracking the time between consecutive visitors. He collected $\boldsymbol{n}$ data points, *i.e.*, x_1 is the amount of time between the start (time 0) and the 1^{st} visitor, x_2 is the amount of time between the 1^{st} and the 2^{nd} visitor, ..., x_n is the amount of time between $(n-1)^{th}$ and the n^{th} visitors. Use what we learnt during the lecture to get the maximum likelihood estimate for the parameter λ based on these n observations.