Southern University of Science And Technology
School of Life Sciences
BIO210 Biostatistics
(Spring, 2024)

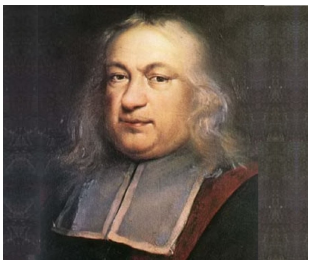
Assignment 3
Due on 26th Mar, 11 p.m.

1. **Rolling dice**: Two fair four-sided tetrahedral dice are rolled simultaneously. Let the random variable $\boldsymbol{X}$ be the absolute difference of the two rolls.

   1.1) **(5 points)** Calculate the PMF, the expected value, and the variance of $\boldsymbol{X}$.

   1.2) **(5 points)** Plot the PMF of $\boldsymbol{X}^2$ and compute $\mathbb{E}\left[\boldsymbol{X}^2\right]$.

2. **The Problem of Points**: let's look at a problem that played an important historical role in the development of probability theory. **The problem of points**, also called **the problem of division of the stakes**, was posed by the French nobleman and gambler Chevalier de Méré (a.k.a Antoine Gombaud) in the 17th century to Pascal, who introduced the ideas that the stake of an interrupted game should be divided in proportion to the players' conditional probabilities of winning given the state of the game at the time of interruption. Pascal worked out some special cases and through a correspondence with Fermat[1].



   The problem led Pascal to the first explicit reasoning about what today is known as an **expected value** or **expectation**. The problem may look easy now, but back in the time when probability theory was non-existent, it required some greatest mathematical minds to solve it. Now let's look at a different but similar version of the problem:

   Han Meimei and Li Lei play a round of golf (18 holes) for a ¥50 stake, and their probabilities of winning on any one hole are p and $1-p$, respectively, independent of their results in other holes. When a person wins one hole, he

[1] Check in the pictures, which one is Pascal and which one is Fermat?

or she will get one point. After a whole round (18 holes) the person with the highest points gets ¥50, and the other person gets nothing. If there is a draw, each of them will get ¥25. At the end of 10 holes, Han Meimei has 6 points and Li Lei has 4 points. Then Li Lei receives an urgent call and has to leave for work, and he suggests dividing the stake. They both agree that a fair way of dividing the stake is to split the ¥50 based on their probabilities of winning had they completed the whole round. The question is: how exactly?

2.1) **(5 points)** Let the random variable $\boldsymbol{H}$ represent the final score of Han Meimei had they completed the whole round. Compute the PMF $\mathbb{P}_{\boldsymbol{H}}(h)$ and the expected value of $\boldsymbol{H}$.

2.2) **(5 points)** Let the random variable $\boldsymbol{L}$ represent the final score of Li Lei had they completed the whole round. Compute the PMF $\mathbb{P}_{\boldsymbol{L}}(l)$ and the expected value of $\boldsymbol{L}$.

2.3) **(10 points)** Let event $\boldsymbol{A}$={Han Meimei would win if they were to complete the whole round}, event $\boldsymbol{B}$={They would end up with a draw if they were to complete the whole round} and event $\boldsymbol{C}$={Li Lei would win if they were to complete the whole round}. Compute $\mathbb{P}(\boldsymbol{A})$, $\mathbb{P}(\boldsymbol{B})$ and $\mathbb{P}(\boldsymbol{C})$. **Note:** algebraic or numerical expressions do not need to be simplified in your answers.

2.4) **(7.5 points)** Let the random variable $\boldsymbol{X}$ represent the money Han Meimei would get had they completed the whole round, and $\boldsymbol{Y}$ that of Li Lei. They both agree that Han Meimei and Li Lei should get the amounts of $\mathbb{E}[\boldsymbol{X}]$ and $\mathbb{E}[\boldsymbol{Y}]$, respectively. Compute $\mathbb{E}[\boldsymbol{X}]$ and $\mathbb{E}[\boldsymbol{Y}]$. **Note:** algebraic or numerical expressions do not need to be simplified in your answers.

3. Who is correct? In a class of 25 students, 11 of them have type **O** blood, 6 type **A**, 5 type **B** and 3 type **AB**. If we randomly select 5 students from them, and let the random variable $\boldsymbol{X}$ represent the number of students with type **B** in them. Now we want to compute the PMF $\mathbb{P}_{\boldsymbol{X}}(x)$.

3.1) **(2.5 points)** Han Meimei approaches the problem in this way: the selection is random, and all outcomes are equally likely. Therefore, she can use the discrete uniform law to calculate probabilities. The total number of outcomes in the sample space is $|\Omega| = \binom{25}{5}$. To figure out the the number of outcomes of having k students with type **B** blood in the sample, she divides the process into two stages. The first stage is to choose k people

from the 5 students with blood type **B** in the class, and the second stage is to choose $5-k$ people from the other 20 students. The total number of outcomes is the simple multiplication of the number of choices in each stage. Now, write the PMF constructed by Han Meimei.

3.2) **(2.5 points)** Li Lei thinks in a different way: the probability is kind of a relative frequency. Therefore, the probability of having a random student with blood type **B** is 5 out of 25. That is 0.2. If 5 students are chosen, the process of choosing one student can be treated as a Bernoulli trial, and there are a total of 5 Bernoulli trials. Therefore, the probability of observing k students with blood type **B** in the sample can be simply calculated using a binomial distribution. Now, write the PMF constructed by Li Lei.

3.3) **(5 points)** Based on the previous two PMFs you just computed, finish the following table to see if they are different or not (***Hint:*** use the `COMBIN` function in **Excel** to help calculate the binomial coefficients.):

k	$\mathbb{P}(\boldsymbol{X}=k)$ by Han Meimei	$\mathbb{P}(\boldsymbol{X}=k)$ by Li Lei
0		
1		
2		
3		
4		
5		

3.4) **(2.5 points)** Who do you think is correct and who is wrong? Explain your answers.

3.5) **(5 points)** In a different class of 256 students, of which 110 of them have type **O** blood, 60 type **A**, 50 type **B** and 36 type **AB**. If we randomly select 5 students, and let the random variable $\boldsymbol{Y}$ represents the number of students with type **B** in them. Repeat the analysis in **3.1)**, **3.2)** and **3.3)** (*i.e.* compute the PMF $\mathbb{P}_{\boldsymbol{Y}}(y)$ using Han Meimei's and Li Lei's methods, respectively, and compare them in a table). What do you notice about the difference between probabilities calculated by Han Meimei's and Li Lei's ways?

4. **Checking independence of a collection of events (2.5 points)**: During the lecture, we made a definition on the independence of a collection of events by using a multiplication equation (Lecture 8 Slide 6). Now suppose we have a collection of $\boldsymbol{n}$ events, how many times do you need to use the equation in order to check if they are independent or not?

5. **Telecommunication (2.5 points)**: In a terrible environment, the probability of success in sending a character by wireless is $\frac{3}{7}$. What is the probability that 22 characters out of 44 are sent successfully, assuming the results of sending each character are independent?

6. **Renal Disease**: The presence of bacteria in a urine sample (bacteriuria) is sometimes associated with symptoms of kidney disease in women. Suppose a determination of bacteriuria has been made over a large population of women at one point in time and 5% of those sampled are positive for bacteriuria.

   6.1) **(2.5 points)** If a sample size of 5 is selected from this population, what is the probability that 1 or more women are positive for bacteriuria?

   6.2) **(2.5 points)** Suppose 100 women from this population are sampled. What is the probability that 3 or more of them are positive for bacteriuria?

   One interesting phenomenon of bacteriuria is that there is a turnover; that is, if bacteriuria is measured on the same woman at two different time points, the results are not necessarily the same. Assume that 20% of all women who are bacteriuric at time 0 are again bacteriuric at time 1 (1 year later), whereas only 4.2% of women who were not bacteriuric at time 0 are bacteriuric at time 1. Let $\boldsymbol{X}$ be the random variable representing the number of bacteriuric events over the two time periods for 1 woman and still assume that the probability that a woman will be positive for bacteriuria at any one exam is 5%.

   6.3) **(2.5 points)** What is the probability distribution of $\boldsymbol{X}$?

   6.4) **(2.5 points)** What is the expected value of $\boldsymbol{X}$?

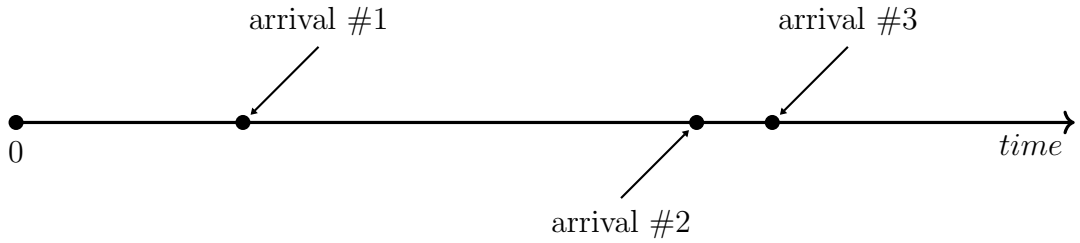
   6.5) **(2.5 points)** What is the variance of $\boldsymbol{X}$?

7. **Otolaryngology**: Assume the number of episodes per year of otitis media, a rare disease of the middle ear in early childhood, follows a Poisson distribution with parameter $\lambda = 1.6$ episodes per year.

   7.1) **(2 points)** Find the probability of getting 3 or more episodes of otitis media in the first 2 years of life.

   7.2) **(2 points)** Find the probability of not getting any episodes of otitis media in the first year of life.

8. **The Poisson process, the exponential distribution and memorylessness:** We have introduced the **binomial** random variable, which models the number of successes after 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" as an example. Empirically, it is reasonable to assume that the number of visitors to a website in a given time window follows a Poisson distribution. This Poisson PMF only has one parameter, and it is **λ visitors per hour**. Now let's keep tracking visitors along a time axis, starting at 0:



   The visitors are the events 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 certain number of events occur within an interval is proportional to the length of the interval and is only dependent on 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. For a particular interval, the events occur independently both within and outside that interval.

8.1) **(1.5 points)** Now, let the random variable (***r.v.***) $\boldsymbol{A}$ represent the number of arrivals in the first 15 minutes, write the PMF of $\boldsymbol{A}$.

8.2) **(1.5 points)** Let the ***r.v.*** $\boldsymbol{B}$ represent the number of arrivals in a particular time interval with a length of $\boldsymbol{t}$ hours. Write the PMF of $\boldsymbol{B}$.

8.3) **(1.5 points)** Now, we observed that there are 5 arrivals during the first hour. Given that has occurred and let the ***r.v.*** $\boldsymbol{C}$ 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;

$\vdots$

8.4) **(1.5 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

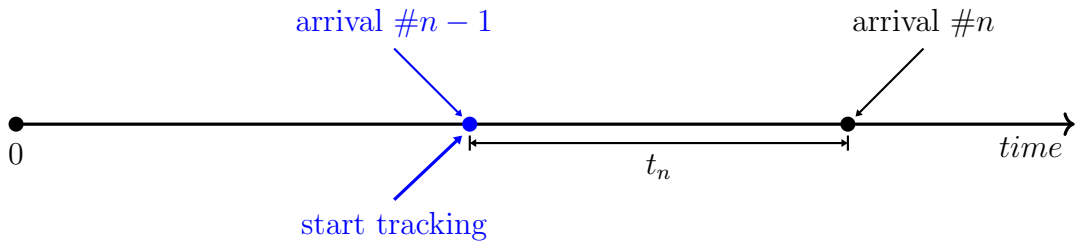
[ ] They are Bernoulli random variables

[ ] They are Binomial random variables

[ ] They are continuous random variables

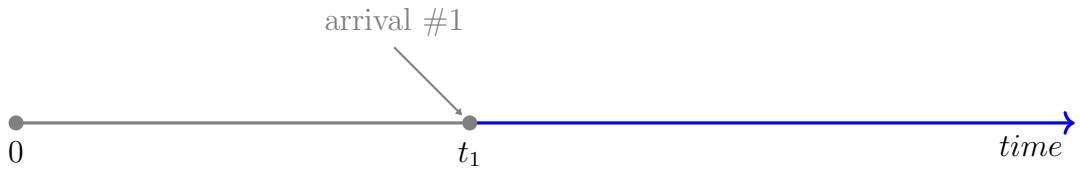
[ ] They all have the same type of distribution

I'm going to tick one answer for you for the above question: their distributions are the same! Why? Think about this 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, $\boldsymbol{T_1}$. **One common trick to get the probability distribution of a random variable is to compute its CDF, and then take the derivate**. Now, look at the following picture:



At the time u, 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 $\mathbb{P}(T_1 > t_1)$.

8.5) **(1.5 points)** $\mathbb{P}(T_1 = t_1) =$ _____.

8.6) **(1.5 points)** How to calculate the probability of the event $\{$ $\boldsymbol{T_1}$ takes a value greater than t_1 $\}$? Think about this: the blue 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 $[0, t_1]$

8.7) **(4 points)** Based on your choice, compute $\mathbb{P}(T_1 > t_1)$. Express the probability using λ and t_1. **Hint:** use the Poisson PMF for the calculation.

$$\mathbb{P}(T_1 > t_1) =$$

8.8) **(1.5 points)** Recall that the CDF of a random variable is defined as $\mathbb{F}_{\boldsymbol{X}}(x) = \mathbb{P}(\boldsymbol{X} \leqslant x)$. Compute the CDF of the random variable T_1:

$$\mathbb{F}_{\boldsymbol{T_1}}(t_1) =$$

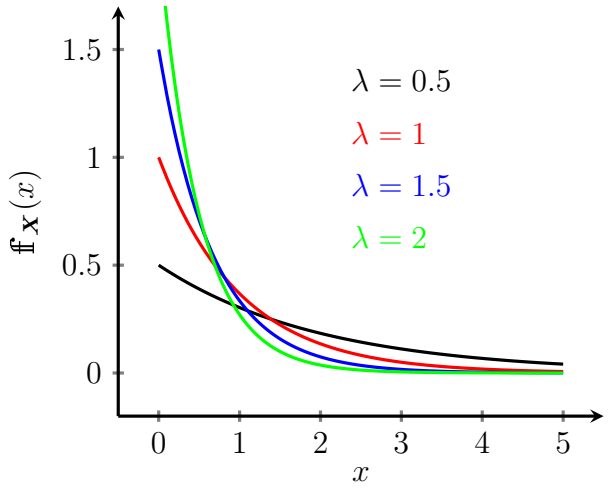
8.9) **(4 points)** Recall that the PDF of a random variable is just the derivative of its CDF. Compute the PDF of the random variable $\boldsymbol{T_1}$:

$$\mathbb{f}_{\boldsymbol{T_1}}(t_1) =$$

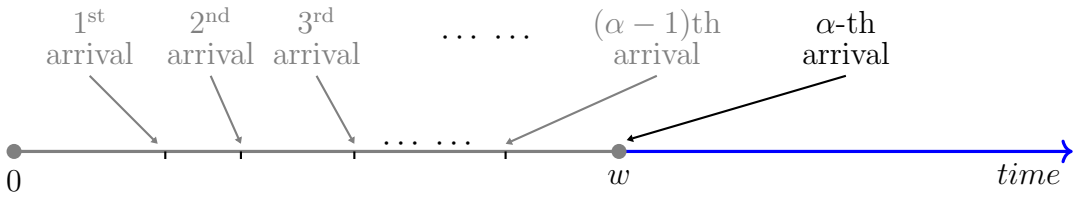
Since all of $\boldsymbol{T_1}$, $\boldsymbol{T_2}$, $\boldsymbol{T_3}$, $\cdots$ have the same distribution. You just computed the PDF for all of them. You can write it in a more general form by replacing $\boldsymbol{T_1}$ with $\boldsymbol{X}$, and it should look like this:

$$\mathbb{f}_{\boldsymbol{X}}(x) = \lambda e^{-\lambda x}$$

That is the PDF of the ***exponential distribution***. It is a very useful probability distribution to model the waiting time between independent events in a Poisson process. 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:



9. * Gamma Distributions: let's look at a slightly different example about the waiting time. In a ***Poisson process***, we introduce a new random variable $\boldsymbol{W}$ to denote the waiting time until the α-th arrivals:



9.1) **(2.5 points)** The blue event { $\boldsymbol{W}$ takes a value greater than w } is equivalent to which of the following:

[ ] Exactly α arrivals in the time interval $[0, w]$

[ ] Exactly $\alpha - 1$ in the time interval $[0, w]$

[ ] No more than α ($\leqslant \alpha$) arrivals in the time interval $[0, w]$

[ ] No more than $\alpha - 1$ ($\leqslant \alpha - 1$) arrivals in the time interval $[0, w]$

9.2) **(1 point)** Based on your answer, compute:

$$\mathbb{P}(\boldsymbol{W} > w) =$$

9.3) **(1 point)** Compute the **CDF**:

$$\mathbb{F}_{\boldsymbol{W}}(w) = \mathbb{P}(\boldsymbol{W} \leqslant w) = 1 - \mathbb{P}(\boldsymbol{W} > w) =$$

9.4) **(1 point)** Compute the **PDF** by taking the derivative. You need to be patient so that many terms will be cancel out. You should get:

$$\mathbb{f}_{\boldsymbol{W}}(w) = \frac{\lambda^{\alpha}}{(\alpha-1)!} e^{-\lambda w} w^{\alpha-1}, \text{ where } w > 0, \lambda > 0, \alpha > 0$$

This is the **Erlang PDF of order** α. Since there is an $(\alpha - 1)!$ term in the denominator, the values of α need to be integers. Since the 18th century, many brilliant mathematicians have been working on extending the **factorial** to non-integers. You can check the Wikipedia page if you are interested in the history.

Eventually, we have what is known today as the **Gamma function**:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} \mathrm{d}x, \text{ where } \alpha > 0$$

When α is an integer, $\Gamma(\alpha) = (\alpha - 1)!$

Now, if we replace the $(\alpha - 1)!$ in the denominator of the Gamma distribution with $\Gamma(\alpha)$, we have the general form of the PDF of the **Gamma distribution**:

$$\mathbb{f}_{\boldsymbol{X}}(x) = \frac{\lambda^{\alpha}}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1}, \text{ where } x > 0, \lambda > 0, \alpha > 0$$

In this case, α is the **shape parameter**, and λ is the **rate parameter**. Alternatively, if we let $\theta = \frac{1}{\lambda}$, the Gamma distribution can be written as the following form, which is also used in many cases:

$$\mathbb{f}_{\boldsymbol{X}}(x) = \frac{1}{\Gamma(\alpha)\,\theta^{\alpha}} e^{-\frac{x}{\theta}} x^{\alpha-1}, \text{ where } x > 0, \theta > 0, \alpha > 0$$

In this form, α is the **shape parameter**, and θ is the **scale parameter**. Those are still valid probability distributions, but the good thing here is: the shape parameter α can take non-integers as well. This makes the Gamma distributions having more flexible shapes which can be used to model many continuous data. Apparently, the **exponential distribution** is just **Gamma distributions** with $\alpha = 1$, and the **Erlang distribution** is just **Gamma distributions** with integer shape parameters α. The distributions with different shape parameters α and scale parameters θ looks like this: