4 Part B starts here and is due on Monday, April 14th at 5 PM.

4.1 Identify Your Lab Partner

This is a multiple choice question. Please select **ONE** of following options that best describes how you complete **Part B** of this lab.

- I am doing Part B of this lab by myself and I don't have a partner.
- My partner for Part B of this lab is [PARTNER'S NAME] with email [berkeley.edu email address]. [SUBMITTER'S NAME] will submit to Gradescope and add the other partner to the group on Gradescope after submission.

Please copy and paste **ONE** of above statements and fill in blanks if needed. If you work with a partner, make sure only one of you submit on Gradescope and that the other member of the group is added to the submission on Gradescope. Refer to the bottom of the notebook for submission instructions.

I am doing Part B by myself

4.2 Section 3: Thinning

Before you begin this section, please review the Summary at the end of Section 7.2 of the textbook.

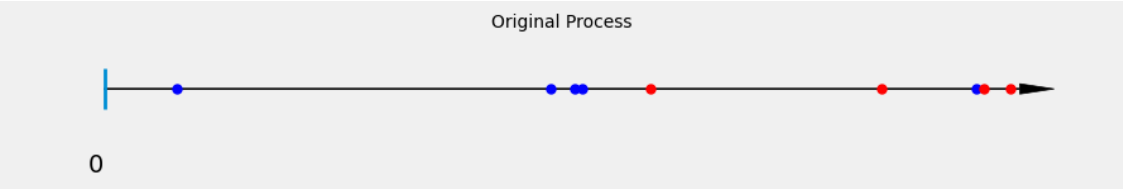
In some Poisson processes, each arrival can be in one of several categories. For example, cars arriving at a toll booth could be domestic or imported. Items arriving at a campus Lost and Found desk could be electronics, notebooks, or everything else.

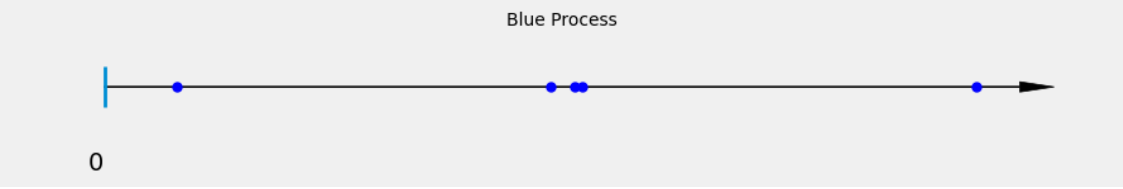
We will work in the case of just two categories, but you will see that there is an immediate generalization to any finite number of categories as in Section 7.3.

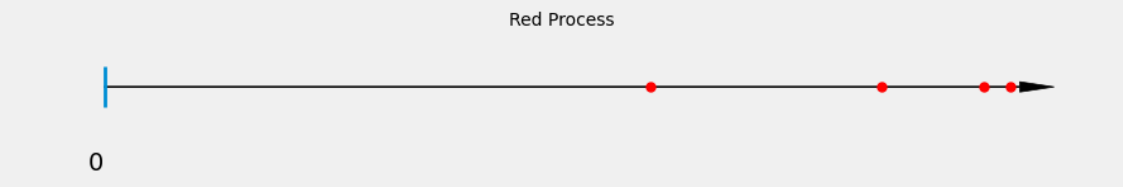
Consider a Poisson process with rate λ, with arrivals represented by red points. Suppose that each arrival is colored blue with chance p, independently of all other arrivals. Then the blue points form a process of their own, as do the remaining red points.

Run the cell below a few times to see what this looks like. We have taken $\lambda = 10$ and $p = 0.6$, but you can change these parameters.

[16]:
```
lam = 10
p_blue = 0.6
Plot_thinned_processes(lam, p_blue)
```







The original process has been separated into a blue process and a red one. Therefore the blue and red processes have a lower intensity than the original process. Their points are more thinly spread out. That is why they are called *thinned* processes.

If you *superpose* the blue process and the red one, that is, if you place the two processes on top of each other, you will recover the original process.

4.2.1 a) The Blue Process

For a fixed time t, let $N_{(0,t)}$ be the total number of points in $(0, t)$, and let $B_{(0,t)}$ be the number of blue points in $(0, t)$. Fill in the blanks with distribution names and parameters. **Explain briefly.**

For each $t > 0$, the total number of points $N_{(0,t)}$ has the _____ distribution, and the number of blue points $B_{(0,t)}$ has the _____ distribution.

Your answer here.

Let I_1 and I_2 be two disjoint time intervals. For each $j = 1, 2$, let B_{I_j} be the number of blue points in the interval I_j. Are B_{I_1} and B_{I_2} independent? **Explain briefly.**

Your answer here.

Fill in the blank:

The blue process is a Poisson process with rate _____ per unit time.

Your answer here.

4.2.2 b) The Blue and Red Processes

For a fixed $t > 0$, let $R_{(0,t)}$ be the number of red points in the interval $(0, t)$.

Fill in the blanks:

$R_{(0,t)}$ has the _____ distribution.

Your answer here.

For each $t > 0$, are the random variables $B_{(0,t)}$ and $R_{(0,t)}$ dependent or independent? Explain.

Your answer here.

Fill in the blanks. The first blank should be filled with one of the words *dependent* or *independent*.

The blue process and the red process are _____ Poisson processes. The blue process has rate _____ and the red process has rate _____.

Your answer here.

4.2.3 c) Lost and Found

Items arrive at a lost-and-found desk according to a Poisson process at the rate of 3 items per day. Each item is a cell phone with chance 5%, independently of all other items.

Use only `stats.poisson.pmf`, `stats.poisson.cdf`, and arithmetic operations to find the chance that in 5 days at least one cell phone and at most 10 other items arrive.

[17]:
```
total = 3 * 5
phone_prop = 0.05
phone = total * phone_prop
```

```
other = total * (1 - phone_prop)

at_least_one = 1 - stats.poisson.pmf(0, phone)
at_most_ten = stats.poisson.cdf(10, other)
total_prob = at_least_one * at_most_ten
total_prob
```

[17]: 0.084261469351954846

4.2.4 d) The Blue and Red Sequence

This exercise is about the lost-and-found processes in Part **c**.

(i) Find the chance that the first cell phone arrives after exactly 15 other items.

Your calculation will be much simpler if you notice that this question is only about the "blue" and "red" sequence, and not about the times at which the points arrive or the gaps of time between them.

[18]:
```
# geometric
fifteen = ((1 - phone_prop)**15) * phone_prop
fifteen
```

[18]: 0.023164561507987652

(ii) Find the chance that the third cell phone arrives after at least 20 other items. Fill in the blank in the comment cell first, and think about what distribution you should use.

[19]:
```
"""Among the first 22 items that arrive, there should be 0, 1, or 2 cell phones.
↪"""

prob = stats.binom.cdf(2,22,0.05)
prob
```

[19]: 0.90517695408156928

4.3 Section 4: Earthquakes in California

"All models are wrong. Some are useful." This trenchant assessment of modeling is attributed to the British statistician George Box and has been the subject of much discussion.

What Box was saying was that while no statistical model can capture all the complexity of a physical system, sometimes simple representations can help us begin to understand systems without taking all of the complexity into account.

Keep that in mind as you skim this paper in a 2016 issue of The American Statistician, a publication of The American Statistical Association. It includes the use of the homogeneous Poisson process to model times of earthquakes in California. It's not surprising that the model might be an over-simplification; see for example this lecture by Prof. Philip Stark of Berkeley's Statistics department. However, it's OK as a rough starting point, and you can use it to make some rough predictions.

Here are the data used in the paper. The table `quakes` consists of data on all earthquakes in California of magnitude 4.9 or greater, between January 1857 and August 2014. There were 51 such quakes. Of course there were smaller earthquakes too, but those are not part of the data set.

In what follows, the word "quake" means "earthquake of magnitude 4.9 or more".

`Magnitude` is a measure of the strength of an earthquake, and `MMI` stands for the Modified Mercalli Intensity scale that measures the intensity of shaking.

An important attribute for the analysis is the `Gap` in time, measured in days, between the quake and the previous one.

[20]:
```
quakes = Table.read_table('Lab07_data/big_CA_earthquakes.csv')
```

[21]:
```
quakes.num_rows
```

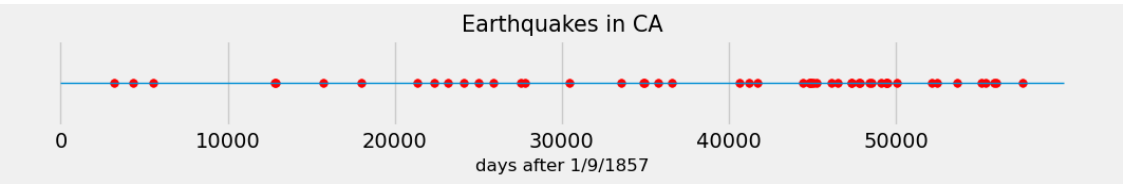
[21]: 51

[22]:
```
quakes.show()
```

```
<IPython.core.display.HTML object>
```

The figure below plots the dates of the 51 earthquakes.

[23]:
```
plt.figure(figsize=(12, 1))
plt.scatter(quakes.column(5), 0*np.ones(quakes.num_rows), color='red', s=30)
plt.hlines(0, 0, 60000, lw=1)
plt.xticks(np.arange(0, 50001, 10000))
plt.yticks([])
plt.xlabel('days after 1/9/1857', fontsize=12)
plt.title('Earthquakes in CA', fontsize=15);
```

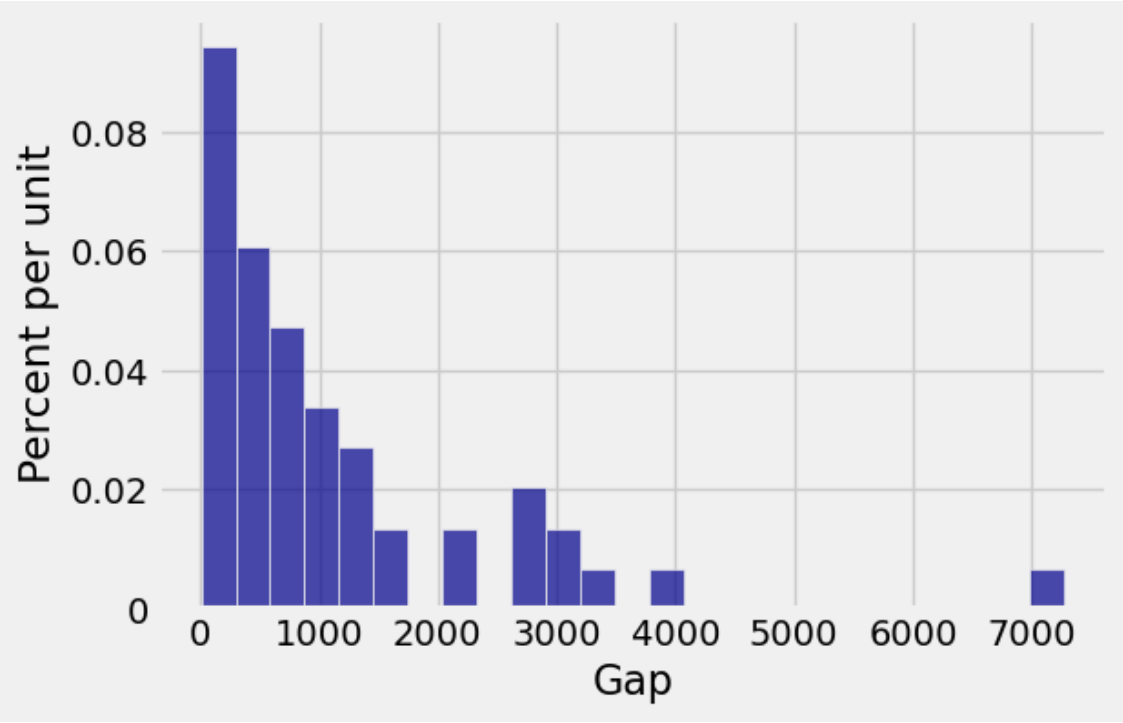


Can this reasonably be modeled by a Poisson process? To answer this, go back and look at the Second Description of the Poisson Process, in **2b**. In the exercise below you will see whether the gaps look exponential.

4.3.1 a) The Gaps and the Rate

Draw a histogram of the gaps.

[24]:
```
quakes.hist('Gap', bins=25)
```



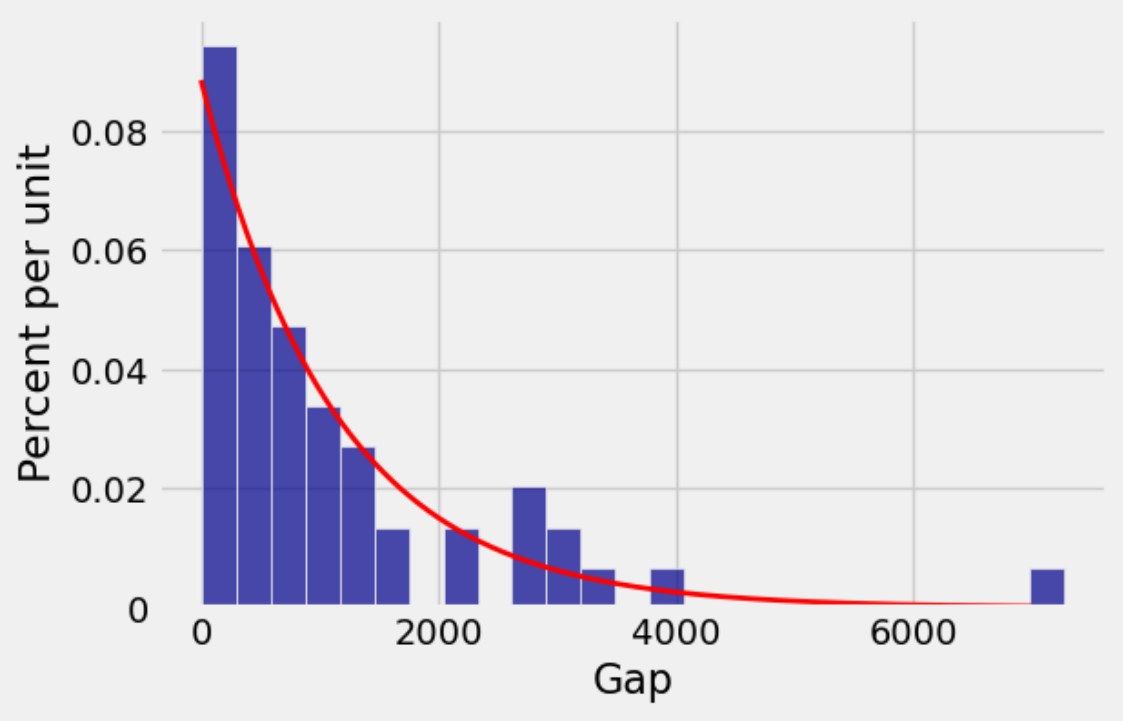
If you model the gap sizes as i.i.d. exponential variables, what is your estimate of the rate λ? Answer this by comparing the mean of the exponential distribution and the average of the observed gaps.

[25]:
```
gaps = quakes.column('Gap')
mean_gap = np.mean(gaps)

lam = 1/mean_gap

mean_gap, lam
```

[25]:
```
(1128.8235294117646, 0.0008858780614903596)
```

Complete the cell below to superpose the appropriate exponential density on the histogram of gaps. Remember that the `scale` argument of `stats.expon` is $1/\lambda$, not λ.

[26]:
```
quakes.hist('Gap', bins=25)
t = np.arange(7000)
f = stats.expon.pdf(t, scale = 1/lam)
plt.plot(t, f, color='red', lw=2);
```



If you have done your calculations correctly, the model of i.i.d. exponential gaps should look plausible. Of course with only 51 observations there will always be some doubt.

Now suppose that you model future quakes as a Poisson process with the rate that you assigned to `lam` at the start of this exercise.

Caution: The answers that you get based on this model should only be taken as a very rough starting point towards a deeper analysis. Apart from the need to check whether all the assumptions of the model apply, there is also the need to account for the variability in the estimate of λ even assuming that the model works. All of that is the domain of time series or geostatistics classes; we will stick to the probability calculations.

4.3.2 b) A Prediction for the Next Year

In the calculations below, you can assume that years have 365 days. Don't worry about details such as leap years or whether you should include the day when you start your observations. Also assume that "within a year from now" means "within 365 days from the day you do the calculation".

(i) Find the probability that there is at least one quake within a year from now.

[27]:
```
# P(at least one quake within a year from now)

prob = 1 - stats.poisson.pmf(0, lam*365)
prob
```

[27]:
```
0.2762762299248045
```

4.3.3 c) Another Prediction for the Next Year

Find the observed proportion of quakes that had magnitudes of at least 6.0.

[28]:
```
# number of quakes with magnitude at least 6.0
num_big = quakes.where('Magnitude', are.above_or_equal_to(6.0)).num_rows

# proportion of quakes with magnitude at least 6.0
prop_big = num_big / quakes.num_rows

num_big, prop_big
```

[28]:
```
(37, 0.7254901960784313)
```

Suppose that each future quake has magnitude at least 6.0 with chance `prop_big` independently of all others. As in the previous part, you can ignore the variability in that estimate; leave that for a time series or geostatistics class.

Under the assumption above, find the chance that within 365 days from now there is one quake of magnitude at least 6.0 and at least one with magnitude in the interval $[4.9, 6)$. Remember that all the quakes in our process have magnitude at least 4.9.

[29]:
```
# P(1 quake of magnitude at least 6.0 and at least one quake of magnitude in [4.
↪9, 6.0))

lam_big = lam * 365 * prop_big
lam_small = lam * 365 * (1 - prop_big)

prob_1 = stats.poisson.pmf(1, lam_big)
prob_2 = 1 - stats.poisson.pmf(0, lam_small)

total_prob = prob_1 * prob_2
total_prob
```

[29]:
```
0.015758422561682667
```

4.3.4 d) Predicting A Big One

Under the same assumptions as in Part **c**, find the number of years n such that there is 99% chance that within n years from now there will be at least one quake of magnitude at least 6.0. It's fine if n is not an integer.

[30]:
```
# number of years n such that
# P(at least quake of magnitude 6.0 will happen in n years) = 0.99
lamb_n = -1 * np.log(0.01)
n = lamb_n / lam_big
n
```

[30]: 19.631221592546048

4.4 Conclusion

What you have learned in this lab:

- Properties of one of the most commonly studied stochastic processes
- How simple assumptions about randomness can lead to a powerful theory
- A physical model for gamma (r, λ) random variables when r is an integer
- Ways in which the major distribution families interact with each other
- An approach to modeling physical phenomena

This lab required quite a bit of mental agility. Congratulations on a job well done!

5 Part B ends here and is due on Monday, April 14th at 5 PM.

5.1 Submission Instructions

Many assignments throughout the course will have a written portion and a code portion. Please follow the directions below to properly submit both portions.

5.1.1 Written Portion

- Scan all the pages into a PDF. You can use any scanner or a phone using applications such as CamScanner. Please **DO NOT** simply take pictures using your phone.
- Please start a new page for each question. If you have already written multiple questions on the same page, you can crop the image in CamScanner or fold your page over (the old-fashioned way). This helps expedite grading.
- It is your responsibility to check that all the work on all the scanned pages is legible.
- If you used LaTeX to do the written portions, you do not need to do any scanning; you can just download the whole notebook as a PDF via LaTeX.

5.1.2 Code Portion

- Save your notebook using `File > Save and Checkpoint`.
- Generate a PDF file using `File > Download As > PDF via LaTeX`. This might take a few seconds and will automatically download a PDF version of this notebook.
  - If you have issues, please post a follow-up on the general Lab 7B Ed thread.

5.1.3 Submitting

- Combine the PDFs from the written and code portions into one PDF. Here is a useful tool for doing so.
- Submit the assignment to Lab 7B on Gradescope.

- **Make sure to assign each page of your pdf to the correct question.**
- **It is your responsibility to verify that all of your work shows up in your final PDF submission.**

If you are having difficulties scanning, uploading, or submitting your work, please read the Ed Thread on this topic and post a follow-up on the general Lab 7B Ed thread.

[ ]:

3a. total # points $N_{(0,t)} \sim$ Poisson (λt)

$\therefore$ # arrivals with a Poisson (λ) distribution over an interval $(0,1)$ follows a Poisson $(\lambda * t)$ distribution.

blue points $B(0,t) \sim$ Poisson $(\lambda + p)$

$\therefore$ each arrival is independent and blue with Bernoulli(p) distribution. Since the total # arrivals $N_{(0,t)} \sim$ Poisson (λt), we have a random number of Bernoulli trials, allowing for poissonization of the binomial.

B_{I_1} and B_{I_2} are independent because the time intervals for I_1 and I_2 are disjoint, meaning that the arrivals in one interval don't affect the other. Since each arrival is independent and marked blue with probability p, the # blue arrivals in each interval is also independent of each other.

The blue process is a Poisson process with rate λp per unit time

3b. $P(\text{red}) = 1 - P(\text{blue})$
$= 1 - p$

$\therefore R_{(0,t)} \sim$ Poisson $(\lambda(1-p))$

For each $t > 0$, the RVs $B_{(0,t)}$ and $R_{(0,t)}$ are dependent because they are counts of blue and red points from the same Poisson (λt) over fixed interval $(0,t)$. Their sum is always equal to the total # arrivals $N_{(0,t)}$. If the # trials N were random, B and R would be independent (textbook 7.2.4)

The blue process and the red process are dependent Poisson processes. The blue process has rate λp and the red process has rate $\lambda(1-p)$.