Lab 2

Ondrea Robinson

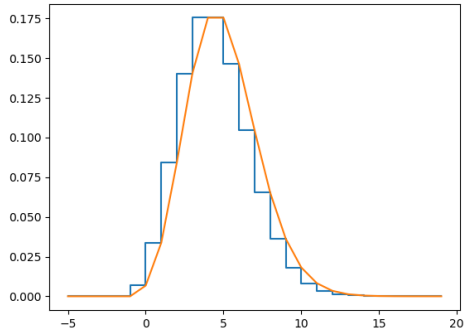
October 2019

1 Problem 1

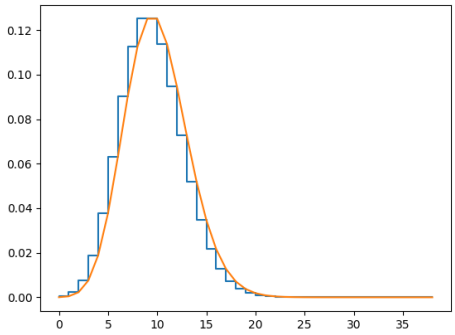
In this section, we look at detecting gamma rays. We want to detect gamma rays against a background of cosmic rays that follow a Poisson distribution. We assume that in one day there's an average of 5 cosmic rays and 8 gamma rays detected by the detector.

1.1

For one day our Poisson background distribution looks like:

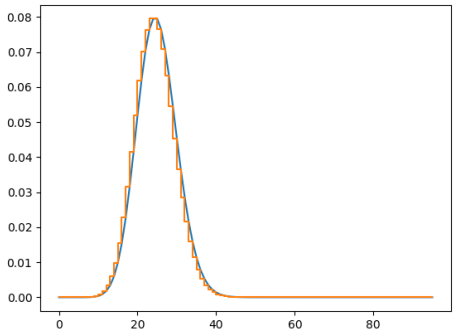


For two days, we can look at the probability as a sum by taking the convolution of our two daily distributions. We then see the probability of events for two days summing to an average number that's twice the average for one day. In our example below, the average for two days is 10 events which is twice the average for one day (5 events).



1.2

If we look at five days we can see a Poisson distribution that averages at 25 events. In the image below one can clearly see this is still a Poisson distribution even after summing over 5 days (step function with the shape of a normal distribution).



Conceptually, it makes sense that over the course of a larger interval, the background would remain Poisson distributed because we're still counting the number of events in a time interval. We haven't changed the measurement method just increased the time interval. Mathematically, we're convolving two of the same function and our result will still be of the same function type.

1.3

As you average days, the probability distribution becomes narrower around the mean. This makes sense since the most common number of events will have an extremely high probability as you average more trials. Conversely, less probable values will become even less likely as you average. This narrows the function

so that it's closer to the mean with decreasing heights as you move outwards. Here's a distribution averaged over 10 days and one averaged over 100 days:

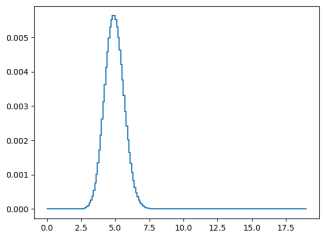


Figure 1: 10 day average

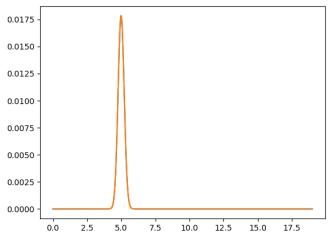


Figure 2: 100 day average

You can see that both have become narrower (the original distribution is the first figure of the paper). The one over 100 days however has become extremely narrow about the mean compared to the one for 10 days.

1.4

Let N be a number of days of observation and Y be the number of gamma rays detected per day. We let $N = 15$ and $Y = 8$. Assuming we saw $M = N*Y = 120$ gamma rays in an N day period, what is the associated sigma value of this measurement? To calculate this we need to compare our measurement value to the appropriate distribution. The correct distribution will also be for an N day period. Since we have the distribution for one day we can take the convolution of this distribution 14 times ($\#ofconvolutions = \#ofdays - 1$). This will give us the distribution of the sum of 15 days with the appropriate average mean of 5 events per day.

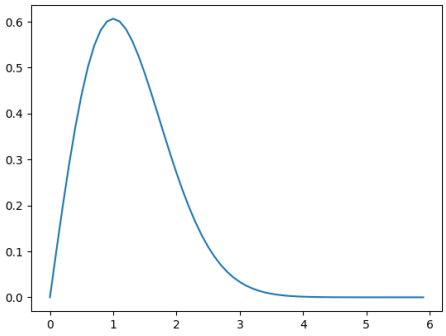
To find the probability of getting our measurement value we integrate our distribution from 120 to ∞. Once we have our probability we can plug this into a ppf function to return our associated sigma value on a normal distribution. See the following code:

```
N=15
M=N*Y
probability = 1 - stats.poisson.cdf(M,X*(15))
sigma = stats.norm.ppf(probability)
print(sigma)
OUT: -4.84
```

So in this case we get a value of 4.84σ.

2 Problem 2

In this section we examined a skewed continuous distribution, the Rayleigh distribution. Here's a standard Rayleigh distribution:



2.1

Now we look at how this distribution changes as we average over more observing intervals. We need to take the convolution with each trial distribution then divide the x-axis by the number of intervals. we use the code:

. We can see the distribution becomes more Gaussian as we average more. Here's a comparison of two distributions with increasing numbers of trials:

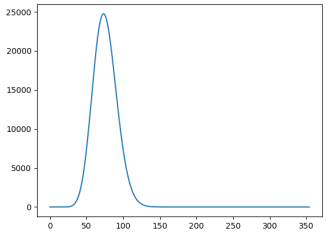


Figure 3: 6 trials averaged Figure 4

As the number of trials go up the distribution looks more and more like a Gaussian distribution with an equal slope on either side of the peaked average.

3 Problem 3

When two neutron stars merge, there's a gravity wave that scientists are able to detect on earth. Recently, astronomers have realized there's also an associated

optical signal. Now they're able to match this optical signal to gravity wave measurements. This is significant because it provides a larger data-set about what actually happens during these mergers. We're going to investigate some of the statistics behind detecting these signals.

3.1 Version 1

Suppose we're looking at optical data-sets from a telescope searching for neutron star signals. Each pixel has a background distribution that is Gaussian with a width $X = 1.65$. We get an alert from a gravity wave detector that the signal is coming from a specific pixel. We look at this pixel and detect a signal of brightness $Y = 10$. To calculate the significance of the signal we have to look at the probability of getting a measurement value of Y or more. We can use the following python code:

```
x = 1.65
Y = 10
prob = 1 - stats.norm.cdf(Y,loc=0, scale=x)
sigma = stats.norm.ppf(prob)
print(prob)
print(sigma)
```

In this case $prob = 6.78047 * 10^{-10}$ and the associated $\sigma = 6.06$. This means that our detected signal Y is a significant discovery since $\sigma(Y) = 6.06 > 5\sigma$. Recall that 5σ is the threshold to be considered a discovery.

3.2 Version 2

In the last example we had a specific pixel to look at. If instead we only have the measurement value Y and a field of Gaussian distributed pixels, it's a little more complicated. Instead we have to look at the distribution of the entire field. We don't want to take the convolution, which we would use if we were summing our distributions, we want to multiply since we're looking at all the pixel contributions at once. To phrase our problem as a statistics question: *what is the significance of a measurement Y in one pixel against a distribution of 10k pixels if all the pixels are Gaussian?* To find the significance of our signal we can use the following, modified, code:

```
x=1.65
Y=10
prob = 1- stats.norm.cdf(Y, loc=0, scale=1.65*10000)
sigma= stats.norm.ppf(prob)
print(prob)
print(sigma)
```

In this case our significance $\sigma = 0.00060606$. which is very very small. Against this distribution the measurement value we got before is insignificant and hardly close to a discovery. This is reasonable if you consider comparing the light measurement in one pixel to the light measurement of 10k cumulative pixels.

4 Problem 4

What we explored in versions 1 and 2 of the last problem is called the trials factor. This effect is powerful but often overestimated. In this section we look closer at its effect. This problem will be done in 4 steps.

4.1 A

First, let's calculate the minimum signal strength needed for a 5σ discovery in the first version of the problem. Recall $x = 1.65$ (the width of our distribution or one σ and $Y = 10$. So in this case we want the ppf of our probability to give us 5σ. $5\sigma = 5 * 1.65 = 8.25$. Let's check:

```
x=1.65
M=5*x
prob=1-stats.norm.cdf(M, loc=0, scale=1.65)
sigma=stats.norm.ppf(prob)
print(prob)
print(sigma)
```

We get out 5σ.

4.2 B

Now let's calculate the minimum signal strength needed for a significant discovery in the second version of the problem. In this case, we multiply the distribution for one pixel by our trials factor (the number of pixels) to get the distribution for 10k pixels. Then looking at this distribution, we can use the inverse cdf to find the measurement value that equals 5σ in our particular distribution.

```
numTrials = 10000
x = 1.65 #width of the distribution
prob5sigma = 1/(3.5e6)
det = stats.norm.ppf((1 - prob5sigma/numTrials), loc=0,
scale=x*numTrials)
print(det)
```

From this our measurement value needed to be of significance 5σ is det = 10809.1935.

4.3 C

The signal in the second case must be 100 times larger rather than 10k.

4.4 D

To summarize, if we perform multiple tests then a measurement value of 1/n is expected to occur after n tests. To plan for this, we can divide our threshold B for one test by the number of tests n, so a result is significant when p ¡ B/n.