Poisson Statistics Analysis

May 2021

Abstract

We test the hypothesis that count data from a Geiger counter detecting radiation from a ^{60}Co source follows a Poisson distribution. This includes deriving the equations for quantities that describe the distribution, such as mean and variance.

1 Introduction

2 Analysis

Our sample of ^{60}Co has a number of nuclei on the order of Avagadro's number, $\sim 10^{23}$, and since the half life of ^{60}Co is 5.27 years [1] while the time taken collecting data was around 30 minutes, the total number of nuclei throughout the course of the experiment was approximately constant: N_n. The number of nuclei that decay in any one time step is in general given by the binomial distribution $B(n; p, N_n)$ but since in our case N_n is huge and p is tiny, we can safely approximate this distribution by the Poisson distribution $P(n; \mu)$, where $\mu = pN_n$. To justify this, we can compute the probability p of any one nucleus decaying during one trial of 10 seconds:

The probability that any one nucleus will decay in a trial of length 5.27 years is 0.5. Thus the probability that any one nucleus will decay in a trial of length 10 seconds is

$$\begin{aligned} p &= 0.5\frac{10s}{5.27y} \\ &= 0.5\frac{10s}{166308552s} \\ &= 3.006 \times 10^{-8} \end{aligned}$$

Since the Poisson distribution gives $p = \frac{\mu}{N_n}$ and N_n is very large, p being of order 1×10^{-8} is reasonable.

Our assumption is that the data we have collected is distributed according to a Poisson distribution, that is it has a mean value μ. In order to extract the value of μ for our data we use the Method of Maximum Likelihood (MML), which states that the estimator, in our case $\hat{\mu}$, for a given distribution is found at the maximum of the joint probability distribution function (pdf), so it's a solution to

$$\partial_{\hat{\mu}} \prod_{i=1}^{N} P(x_i; \hat{\mu}, \vec{\theta}) \tag{2.1}$$

where N is the number of data points or trials and $\vec{\theta}$ is the other parameters that the distribution could depend on. For a Poisson distribution this $\vec{\theta}$ is 0, and for a Gaussian distribution it's σ. We find the value of this estimator $\hat{\mu}$ for both the Poisson and Gaussian distributions by first noticing that the maximum of the joint pdf will occur at the same point as the maximum of the log of the joint pdf since log is monotonic. So we solve:

$$
\begin{aligned}
\text{Poisson: } 0 &= \partial_{\hat{\mu}} \ln\left(\prod_{i=1}^{N} P(x_i;\hat{\mu})\right) \\
&= \partial_{\hat{\mu}} \sum_{i=1}^{N} \ln(P(x_i;\hat{\mu})) \\
&= \sum_{i=1}^{N} \partial_{\hat{\mu}} \ln\left(\frac{\hat{\mu}^{x_i} e^{-\hat{\mu}}}{x_i!}\right) \\
&= \sum_{i=1}^{N} \left(\frac{\hat{\mu}^{x_i} e^{-\hat{\mu}}}{x_i!}\right)^{-1} \left(\frac{x_i \hat{\mu}^{x_i-1} e^{-\hat{\mu}}}{x_i!} - \frac{\hat{\mu}^{x_i} e^{-\hat{\mu}}}{x_i!}\right) \\
&= \sum_{i=1}^{N} \left(\frac{1}{\hat{\mu}^{x_i}}\right) \hat{\mu}^{x_i} \left(x_i \hat{\mu}^{-1} - 1\right) \\
&= \sum_{i=1}^{N} \left(\frac{x_i}{\hat{\mu}} - 1\right) = 0 \\
\implies & \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \hat{\mu} \\
\implies & \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \equiv \bar{x}
\end{aligned}
$$

$$
\begin{aligned}
\text{Gaussian: } 0 &= \partial_{\hat{\mu}} \ln\left(\prod_{i=1}^{N} P(x_i;\hat{\mu},\hat{\sigma})\right) \\
&= \sum_{i=1}^{N} \partial_{\hat{\mu}} \ln\left(\frac{1}{\hat{\sigma}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)^2}\right) \\
&= \sum_{i=1}^{N} \partial_{\hat{\mu}} \left[\ln\left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right) + \ln\left(e^{-\frac{1}{2}\left(\frac{x_i-\hat{\mu}}{\hat{\sigma}}\right)^2}\right)\right] \\
&= \sum_{i=1}^{N} \left[0 + \left(-\frac{x_i-\hat{\mu}}{\hat{\sigma}} \frac{-1}{\hat{\sigma}}\right)\right] \\
&= \sum_{i=1}^{N} \frac{x_i - \hat{\mu}}{\hat{\sigma}^2} = 0 \\
\implies & \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} \hat{\mu} \\
\implies & \hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i \equiv \bar{x}
\end{aligned}
$$

So we can say that our estimator to the mean of the Poisson and Gaussian distributions, $\hat{\mu}$ is just $\bar{x}$, the arithmetic mean of our data. Now we ask whether this estimator is biased or not, that is to say whether the expectation of the estimator is the parameter itself, i.e $E[\hat{\mu}] = \mu$. Noting

that the data is assumed to be Poisson distributed, so $E[x_i] = \mu$:

$$\begin{aligned} E[\hat{\mu}] = E[\bar{x}] &= E[\frac{1}{N}\sum_{i=1}^{N} x_i] \\ &= \frac{1}{N} E[\sum_{i=1}^{N} x_i] \\ &= \frac{1}{N}\sum_{i=1}^{N} E[x_i] \\ &= \frac{1}{N}\sum_{i=1}^{N} \mu \\ &= \frac{N}{N}\mu = \mu \end{aligned}$$

Now we consider the uncertainty related to the estimator $\hat{\mu}$, which is the square root of the variance $V[\hat{\mu}]$. This is motivated by the Gaussian distribution $P(x; \mu, \sigma)$ where the uncertainty on x is $V[x] \equiv E[(x-\mu)^2]$:

$$\begin{aligned} V[x] &= E[x^2] - E[x]^2 \\ &= E[x^2] - \mu^2 \\ &= \int_{-\infty}^{\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} dx - \mu^2 \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2} dx - \mu^2 \\ \text{let } t = \frac{x-\mu}{\sqrt{2}\sigma},\ dt &= \frac{dx}{\sqrt{2}\sigma} \\ \implies V[x] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (\sqrt{2}\sigma t + \mu)^2 e^{-t^2} \sqrt{2}\sigma dt - \mu^2 \\ &= \frac{1}{\sqrt{\pi}} \left(\int_{-\infty}^{\infty} 2\sigma^2 t^2 e^{-t^2} dt + \int_{-\infty}^{\infty} 2\sqrt{2}\sigma t \mu e^{-t^2} dt + \int_{-\infty}^{\infty} \mu^2 e^{-t^2} dt \right) - \mu^2 \\ &= \frac{1}{\sqrt{\pi}} \left(2\sigma^2 \int_{-\infty}^{\infty} t^2 e^{-t^2} dt + 2\sqrt{2}\sigma\mu \left(-\frac{1}{2} e^{-t^2}\right)\Big|_{-\infty}^{\infty} + \mu^2\sqrt{\pi} \right) - \mu^2 \\ &= \frac{1}{\sqrt{\pi}} \left(2\sigma^2 \int_{-\infty}^{\infty} t^2 e^{-t^2} dt + 2\sqrt{2}\sigma\mu \cdot 0 \right) + \mu^2 - \mu^2 \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \int_{-\infty}^{\infty} t^2 e^{-t^2} dt \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \left[\left(-\frac{t}{2} e^{-t^2} \right)\Big|_{-\infty}^{\infty} + \frac{1}{2} \int_{-\infty}^{\infty} e^{-t^2} dt \right] \\ &= \frac{2\sigma^2}{\sqrt{\pi}} \frac{1}{2} \sqrt{\pi} = \sigma^2 \end{aligned}$$

And now we want to find the uncertainty on the mean value parameter that we extract from the data, which is the square root of the variance $V[\hat{\mu}] \equiv E[(\hat{\mu}-\mu)^2]$. Note the data is assumed

to be Poisson distributed.

$$\begin{aligned} E\left[(\hat{\mu}-\mu)^2\right] &= E\left[(\bar{x}-\mu)^2\right] \\ &= E\left[(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu)^2\right] \\ &= E\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j - \frac{2\mu}{N}\sum_{i=1}^{N} x_i + \mu^2\right] \\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} E[x_i x_j] - \frac{2\mu}{N}\sum_{i=1}^{N} E[x_i] + E[\mu^2] \end{aligned}$$

Now $E[x_i x_j] = E[x^2] = \mu^2 + \mu$ if $i = j$ and $E[x_i]E[x_j] = \mu^2$ if $i \neq j$. So

$$\begin{aligned} V[\hat{\mu}] &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[(\mu^2+\mu)\delta_{ij} + \mu^2(1-\delta_{ij})\right] - 2\mu\bar{x} + \mu^2 \\ &= \frac{1}{N^2}[N\mu^2 + N\mu + N^2\mu^2 - N\mu^2] - 2\mu^2 + \mu^2 \\ &= \frac{\mu}{N} + \mu^2 - \mu^2 \\ &= \frac{\mu}{N} \backsimeq \frac{\bar{x}}{N} \end{aligned}$$

Thus we have an estimate for the uncertainty of our mean:

$$\Delta\bar{x} = \sqrt{\frac{\bar{x}}{N}} \tag{2.2}$$

And for completeness, the mean count for a data set is given by

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.3}$$

2.1 Running Mean

In order to visualise how the arithmetic mean converges to a value and how the uncertainty for it shrinks, we plot a running mean

$$r_c(j) \equiv \bar{x}_j = \frac{1}{j}\sum_{i=1}^{i=j} x_i \tag{2.4}$$

with the uncertainty for each running mean given by

$$\Delta\bar{x}_j = \sqrt{\frac{\bar{x}_j}{j}} \tag{2.5}$$

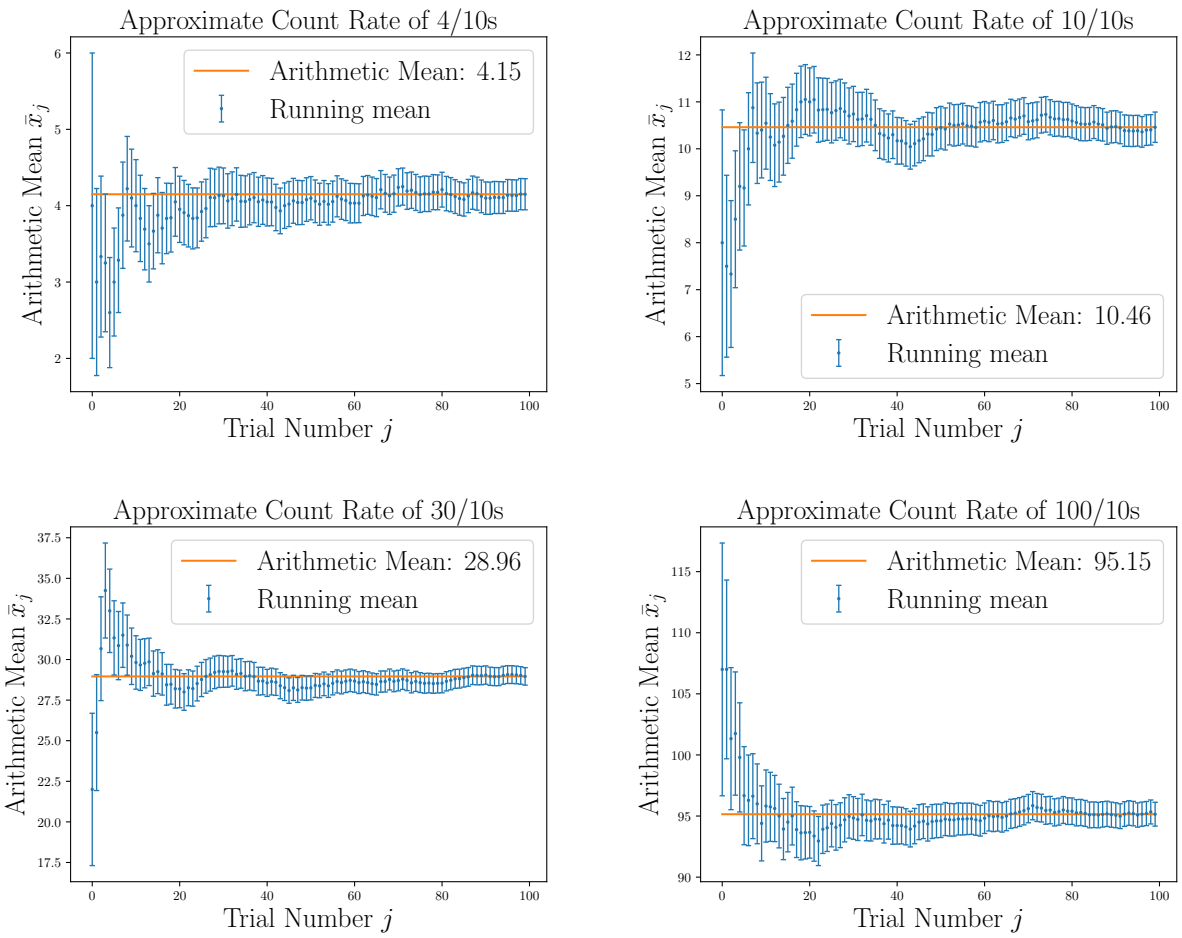

Figure 2.1: Running Means for all 4 approximate count rates. Means calculated using Equation 2.4 and uncertainties calculated using Equation 2.5.

We can clearly see the value of the mean tending to a value for each count rate, even if that value is not the expected value, as well as the uncertainty for each mean shrinking with each step. We also see that in the first 10 or so trials, the estimate for the uncertainty on the mean is very large as the estimation we use depends on a large number of trials, so for small j the estimate is inaccurate.

2.2 Arithmetic Mean and Sample Variance of the Data

In order to test with more certainty that our measured distribution is Poisson or not we must look at the variance of the data. A Poisson distribution has precisely the same mean and variance, as proved in the pre-lab questions, and so now we look for an unbiased estimator for the variance of a Poisson distribution. The sample variance is defined as

$$s^2 \equiv \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 \tag{2.6}$$

and we can check that this is unbiased for a Poisson distribution, i.e. $E[s^2] = \mu$:

$$\begin{aligned}
E[s^2] &= E\left[\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2\right] \\
&= \frac{1}{N-1}E\left[\sum_{i=1}^{N}(x_i - \bar{x})^2\right] \\
&= \frac{1}{N-1}E\left[\sum_{i=1}^{N}x_i^2 - 2\sum_{i=1}^{N}x_i\bar{x} + \sum_{i=1}^{N}\bar{x}^2\right] \\
&= \frac{1}{N-1}E\left[\sum_{i=1}^{N}x_i^2 - 2\bar{x}N\bar{x} + N\bar{x}^2\right] \\
&= \frac{1}{N-1}E\left[\sum_{i=1}^{N}x_i^2 - N\bar{x}^2\right] \\
&= \frac{1}{N-1}\left(\sum_{i=1}^{N}E[x_i]^2 - E[N\bar{x}^2]\right) \\
&= \frac{1}{N-1}\left(\sum_{i=1}^{N}(\mu^2 + \mu) - N\left(\frac{\mu}{N} + \mu^2\right)\right) \\
&= \frac{1}{N-1}(N\mu^2 + N\mu - \mu - N\mu^2) \\
&= \frac{1}{N-1}(\mu(N-1)) = \mu
\end{aligned}$$

We can do the same for a Gaussian distribution, i.e. show $E[s^2] = \sigma^2$. We can skip the first few lines since they are identical no matter the distribution:

$$\begin{aligned}
E[s^2] &= \frac{1}{N-1}\left(\sum_{i=1}^{N}E[x_i]^2 - E[N\bar{x}^2]\right) \\
&= \frac{1}{N-1}\left(\sum_{i=1}^{N}(\sigma^2 + \mu^2) - N\left(\frac{\sigma^2}{N} + \mu^2\right)\right) \\
&= \frac{1}{N-1}(N\sigma^2 + N\mu^2 - \sigma^2 - N\mu^2) \\
&= \frac{1}{N-1}(\sigma^2(N-1)) = \sigma^2
\end{aligned}$$

Thus we have an unbiased estimator for the variance of our data, but we need an uncertainty for it, so we can find $V[s^2] = E[(s^2 - \mu)^2]$. We begin with

$$\begin{aligned} V[s^2] &= \frac{1}{N}\left(\mu(1+3\mu) - \frac{N-3}{N-1}\mu^2\right)[1] \\ &= \frac{1}{N}\left(\mu + 3\mu^2 - \frac{N-3}{N-1}\mu^2\right) \\ &= \frac{1}{(N-1)N}\left((N-1)\mu + 3(N-1)\mu^2 - (N-3)\mu^2\right) \\ &= \frac{1}{(N-1)N}((N-1)\mu + 3N\mu^{-}3\mu^2 - N\mu^2 + 3\mu^2) \\ &= \frac{(N-1)\mu + 2N\mu^2}{(N-1)N} \end{aligned}$$

Thus the uncertainty for s^2 is

$$\Delta s^2 = \sqrt{\frac{(N-1)\mu + 2N\mu^2}{(N-1)N}} \tag{2.7}$$

Now we can finally report a sample mean and variance, with uncertainty, for each approximate count rate in Table 2.1.

Approximate Count Rate	Sample Mean	Sample Variance
4/10s	4.15 ± 0.20	5.12 ± 0.62
10/10s	10.46 ± 0.32	9.8 ± 1.5
30/10s	28.96 ± 0.54	31.2 ± 4.2
100/10s	95.15 ± 0.98	81 ± 14

Table 2.1: Sample Mean and Sample Variance for each approximate count rate, calculated using Equation 2.3, Equation 2.2, Equation 2.6, and Equation 2.7

And finally we can perform our test to see whether the data collected is in fact Poisson distributed, plotting the value of $s^2/\bar{x}$ as a function of $\bar{x}$ for each approximate count rate. The uncertainty on the $s^2/\bar{x}$ value is calculated using

$$\Delta\left(\frac{s^2}{\bar{x}}\right) = \frac{s^2}{\bar{x}}\sqrt{\left(\frac{\Delta s^2}{s^2}\right)^2 + \left(\frac{\Delta\bar{x}}{\bar{x}}\right)^2} \tag{2.8}$$

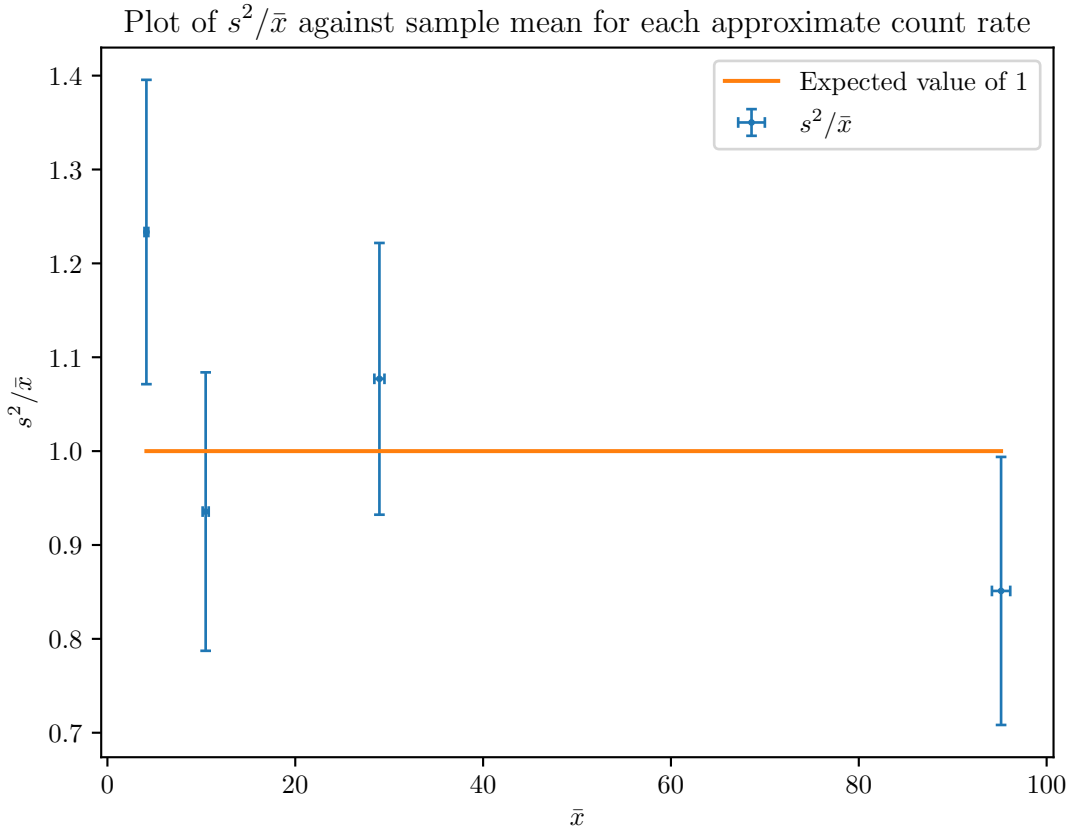

Figure 2.2: The ratio between the sample variance and the sample mean of our data, for each approximate count rate, plotted as a function of the sample mean. Uncertainty on $s^2/\bar{x}$ found from Equation 2.8 and on $\bar{x}$ found from Equation 2.2. The expected value of 1 is shown.

We see that 2 of the 4 count rates produce values of $s^2/\bar{x}$ that agree with the expected value within their uncertainty. This partially confirms our hypothesis but more tests are required.

2.3 Poisson Plots

Now we visualise the Poisson nature of our data in a different way. We will plot a histogram of our data where the x-axis shows the number of counts per trial and the y-axis shows how many trials had that many counts. We generally describe binned data such as this with a binomial distribution $B(n_i; p_i, N)$ where n_i is the number of counts in bin i, N is the total number of trials, and p_i is the probability of a single measurement giving a result in the given bin i. The binomial distribution says that $E[n_i] = Np_i$ and the variance is $V[n_i] = Np_i(1 - p_i)$, so we can estimate the uncertainty of the measured value n_i as

$$\sqrt{n_i\left(1 - \frac{n_i}{N}\right)} \tag{2.9}$$

We could bin our data where each bin has width of 1, but this leads to some inaccurate estimates of means and uncertainties for bins, especially when we have bins with 0 trials. For this reason we tweak our bins in order to get better statistics. Sometimes tweaking is not necessary

however. We also made sure that, on either side, we never had a bin with fewer than 5 trials. Finally we plotted the Poisson distribution with the arithmetic mean from the data as we expect the data to fit to the distribution if our hypothesis that this data obeys a Poisson distribution is correct.

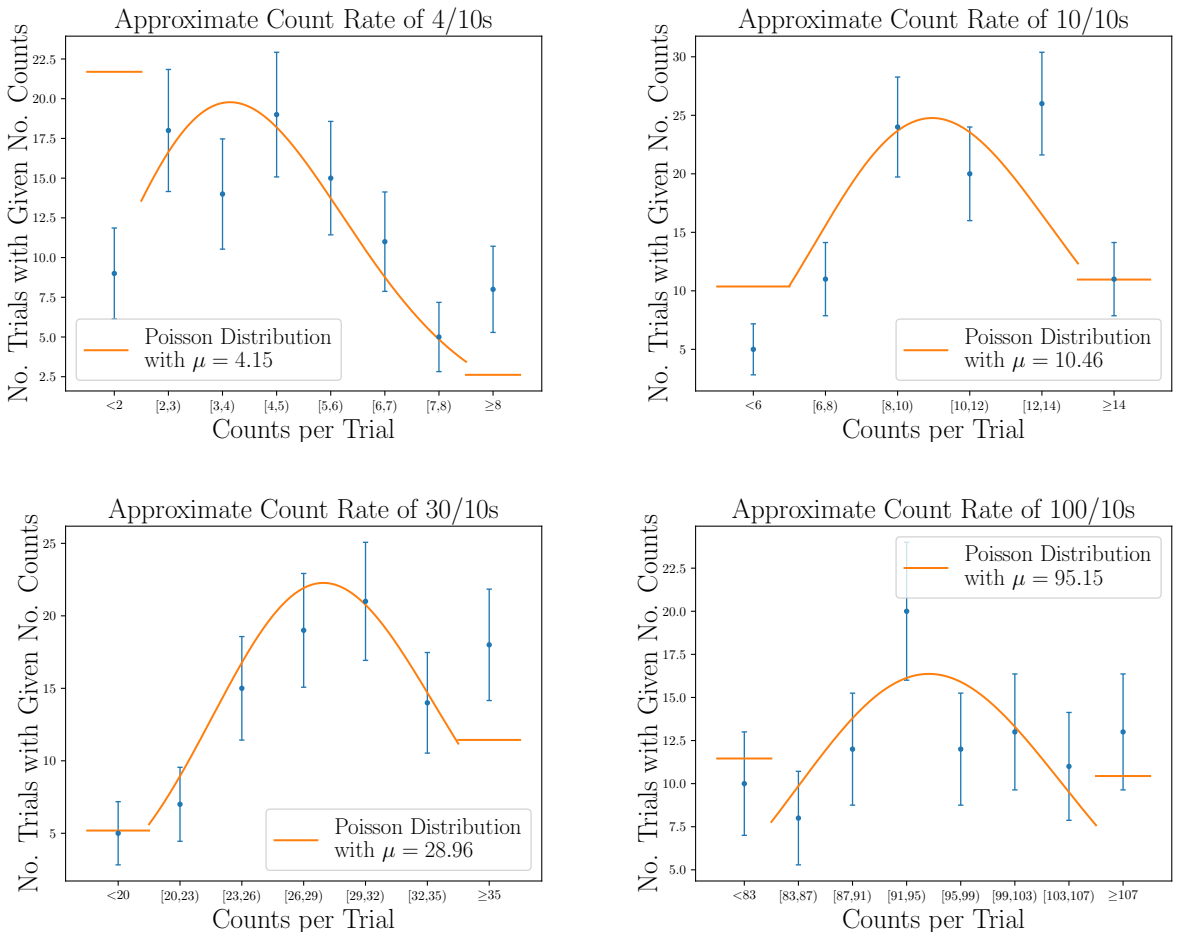

Figure 2.3: Frequency distribution of observed number of counts for all 4 approximate count rates. Uncertainty of Number of Trials given by Equation 2.9

We see that to some degree our data matches the Poisson distribution but a more rigorous approach is needed to confirm of reject our hypothesis.

2.4 Fraction of Bins Not Described by Poisson

We want to have a mathematical test that tells us if our estimates for our error bars was correct, in other words we are checking if our underlying hypothesis is correct as those estimates are based on that hypothesis. Our error bars really do represent the probability that 68% of future measurements will lie within their bounds, then there should a 68% chance that the Poisson prediction made using the arithmetic mean of the data lies within the error bars of each bin. Thus in order to check our hypothesis we can check to see the percentage of bins whose error bars enclose the Poisson curve and see if it is close to 68%.

Counting the number of bins that don't agree with the Poisson curve is easy, we can do it by inspection, but we want to have some kind of uncertainty associated with that number and since

each bin has a probability p_a of agreeing with the curve and there are N_{bins} bins, the number of bins that agree with the curve n_a is given by the binomial distribution and thus the uncertainty of n_a is the square root of the variance of the binomial distribution: $N_{bins}p_a(1-p_a)$. Thus we report the uncertainty of n_a as

$$\sqrt{n_a\left(1-\frac{n_a}{N_{bins}}\right)} \tag{2.10}$$

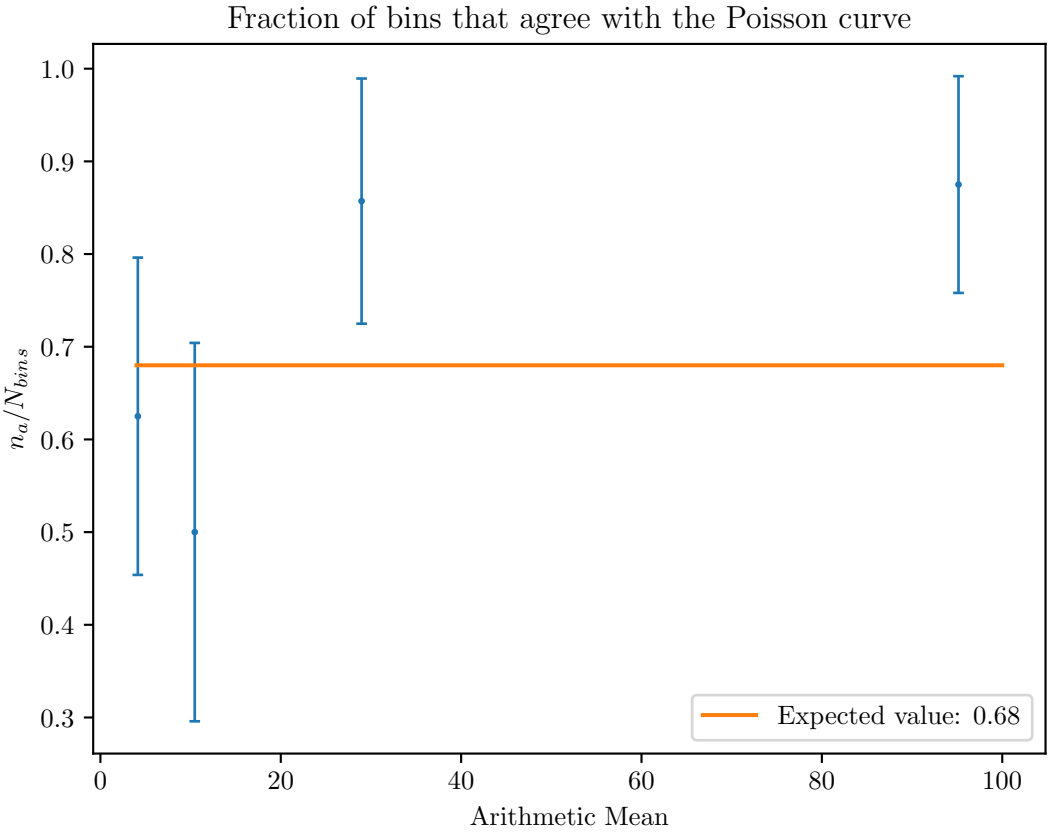

Figure 2.4: Fraction of bins that agree with the Poisson curve within their uncertainty. The expected value of 0.68 is shown. The uncertainty of each fraction is given by Equation 2.10/N_{bins}.

So we see that for 2 of the 4 count rates we are reasonably sure that the Poisson hypothesis is correct. The mean of the fraction of bins that agree across all 4 runs is 0.7143 ± 0.0036 with the uncertainty given by the square root of the sample variance given by Equation 2.6.

3 Conclusion

References

[1] W.A. Horowitz, *Poisson Statistics*, (UCT, 2021)