# Statistical Experiments Using the PCIT

This section describes the new experiments you can perform when you have two capabilities:

- a stream of random events in time, typified by the series of photon-event pulses emerging from the Two-Slit apparatus; and
- a computer-interfaced counter, which makes it easy to get a long series of separate measurements, either of pulse count-totals, or of time intervals between pulses.

The events are expected to be random in the technical sense of a Poisson process, which means:

**In any infinitesimal time interval of duration *dt*, the probability of getting one event to occur is $k{\cdot}dt$ (where *k* is a constant), and this probability is independent of what might have happened, or what will happen, in any past or future interval.**

[The probability of getting two or more events is of order $(dt)^2$, and can be neglected.]

The proportionality constant *k*, which gives probability per unit time, also gives the mean rate at which events occur. What you'll observe is a raw integer number of counts, given by the product of an underlying rate *k*, and the duration for which you choose to count. 'Poisson fluctuations' apply to those raw count numbers, (which we will refer to as $C_i$ in our discussion) and not to count rates.

## Count-total Values for a Poisson Process

The Two-Slit apparatus, as with other devices producing events due to radioactivity, is plausibly a source of Poisson-process events. For a fixed set of (stable) settings, such an experimental apparatus is expected to give a series-in-time of individual events, each event statistically independent of all others, yet the events collectively occurring at a steady *mean* rate.

In order for the Two-Slit apparatus to be producing a steady stream of events, it is useful to have it operating at the ***peak*** of a 2-slit interference pattern maximum or the peak of a 1-slit diffraction pattern). [Operation on the *side* of a peak gives a first-order sensitivity to alignment shifts, which clearly would affect the count rate.] It is also necessary to have fixed the photomultiplier tube's high-voltage setting, and the Counter's input discriminator setting, by a procedure like the one described in Section 2. With such a set-up, the actual count rate you'll observe is clearly still related to the bulb power you choose. This might be adjusted to give a mean count rate of (say) 1000 events per second.

What that means, functionally, is that if you use the Manual switch in the Count mode to get a series of individual 1.0-second readings, you'll get a series of numbers (such as 1046, 1013, 995, 1026, . . . ) whose average is near 1000. These are 'individual readings', each one a count of how many separate photon-events occur during 1.0 seconds of gate-open time. Any list you write down is a subset of a much longer list of readings that you probably lack the patience to acquire, write down, and transcribe. The automatic mode of the PCIT, together with its computer interface, will make it easy for you to get a list of perhaps 500 or more such readings, labeled $C_i$ for $i = 1$ to 500.

Now what can you do with such a list? The very first thing to do is to compute the 'sample mean'

$$C_{avg} = \frac{1}{500}\sum_{i=1}^{500} C_i \quad .$$

The particular value the equation yields has no deep significance, as it would change if you were to change the bulb intensity. The deeper significance lies in the *distribution* of $C_i$ -values around the mean value $C_{avg}$.

One measure of the width of this distribution is its variance $V$, defined by

$$V = \frac{1}{500}\sum_{i=1}^{500} (C_i - C_{avg})^2 \quad .$$

The square root of the variance is related to the standard deviation of the distribution, and it measures the degree to which the $C_i$-values spread out around their common mean $C_{avg}$. [If you were to get the *same* $C_i$-value for each reading, the variance would be zero; if the $C_i$-values were spread out randomly but *uniformly* over the range 0 to $2C_{avg}$, you'd get $V = (4/3)\, C_{avg}{}^2$. It is a prediction from the theory for a Poisson process that the expected variance is neither 0, nor proportional to $C_{avg}{}^2$, but instead given by $1 \cdot C_{avg}{}^1$.

This can be *checked*, because $C_{avg}$ is an independent variable under your control. Suppose each $C_i$ is the result of one 1.0-s count, it's easy to attain $C_{avg} \approx 10^2$ (from background count rates in the PMT, or with the bulb dialed down to minimum brightness), or to attain $C_{avg} \approx 10^4$ (with the bulb dialed up to brighter emission). If for each of a series of bulb brightness choices, you get a 'run' of 500 $C_i$ -samples, you will for each brightness level be able to compute a mean $C_{avg}$, and a variance $V$, from your data sets. Then you can make a scatter plot of ($C_{avg}$, $V$) pairs, and display the data on both linear and log-log axes, to test the theoretical prediction.

If we call $\sqrt{V}$ the 'standard deviation' of the count–numbers achieved, what this is saying is that the standard deviation is expected to depend on $C_{avg}$ as $1 \cdot C_{avg}{}^{1/2}$.

So for a:

| | |
|---|---|
| dim bulb setting giving $C_{avg} = 100$ | we expect a standard deviation of about 10; |
| medium setting giving $C_{avg} = 1000$, | we expect a standard deviation of about 32; |
| bright setting giving $C_{avg} = 10{,}000$, | we expect a standard deviation of about 100. |

Note that the standard deviation *grows* absolutely, but *shrinks* relative to $C_{avg}$, as $C_{avg}$ increases. Note too that these statistics results apply to the raw integer *count values*, not to quotients such as the count rate.

But, given even one single series of $C_i$ -values, there is much *more* you can do with the data. One view of the data you should get is a scatter-plot of $(i, C_i)$-pairs. This is a fine way to make a visual check for

- Scatter: you can zoom in on the vertical axis to make a plot showing how the vertical coordinates $C_i$ show scatter, despite their being obtained as successive readings obtained under nominally identical conditions.
- Outliers: you might see some points in such a plot which are much farther from the mean, and the distribution about the mean, than you expect. Below you'll see how to test this question numerically, but for now you will profit from learning to see 'by eye' the cases of gross outliers which can arise for technical reasons.
- Drift: a scatter plot like this is a good way to see if the count-rate is stable in time. The $C_i$ -values you get were taken uniformly spaced in time, and there are imaginable ways in which the count rate could be varying in time (say, if the bulb were becoming systematically dimmer as it aged). Your eye is quite good at seeing a slope in such a plot of data-values, despite the scatter they exhibit.

Here's one method, a chi-squared test, for seeing if the scatter has the expected size, and whether there are outliers. Make the hypothesis that each $C_i$ -value comes as a sample from a distribution of mean $\mu$ and standard deviation $\sqrt{\mu}$. Then perform a 'fit' of this $C_i$-vs.-$i$ data to a model which claims that each $C_i = m$, where $m$ is some mean-value . The quality of fit relative to this one-parameter model can be measured by the smallness of $\chi^2$, defined by

$$\chi^2 = \sum_{i=1}^{500} \left(\frac{C_i - m}{\sqrt{m}}\right)^2 \quad .$$

Practically speaking, you compute the function $\chi^2(m)$ for various hypothetical values of $m$, by computing

$$\chi^2(m) = \sum_{i=1}^{500} \frac{(C_i - m)^2}{C_i} \quad .$$

This $\chi^2(m)$-function will have a minimum, and that will occur near location $m = C_{avg}$. The value you find for $m$ from your data is your best estimate for the underlying mean value of counts per gate interval. The $\chi^2$-function's minimum value is expected *not* to be 0 (as we'd get from a perfect fit of the data to the model), but rather around 500. [In fact we'd expect $\chi^2_{min} \approx 499$, for 500 data points less 1 fit parameter for 499 degree of freedom; and we'd further expect $\chi^2_{min}$ to lie in the range

$$500 \cdot \left(1 \pm \frac{\text{a few}}{\sqrt{2 \cdot 500}}\right) \quad .]$$

If $\chi^2_{min}$ is much larger than this, it would provide good evidence *against* the hypothesis that the data is described by a Poisson process having a fixed mean. In other words, it would show that the data displayed too much scatter. Could this be due to outliers? (You could examine the data set, or the plot, to spot them.) Or could it be due to non-steady count rate? (Next comes a test of that hypothesis.)

Of course if you have the $(i, C_i)$ scatter-plot, you can model it with alternative hypotheses. For technical reasons, there might have been a step-discontinuity in the underlying $C_{avg}$ -value. Or, there might be a uniform-in-time drift in the average count rate. In the latter case, a new model for each $C_i$ would be

$$C_i = m_0 + m_1 \cdot i \quad ,$$

so that the average count would start near $m_0$, but would rise with $i$-index (i.e. with time) at a rate $m_1$. To test this hypothesis, you can evaluate a new $\chi^2$ sum, given by

$$\chi^2(m_0, m_1) = \sum_{i=1}^{500} \frac{[C_i - (m_0 + m_1 \cdot i)]^2}{C_i} \quad .$$

This function of *two* parameters will also have a minimum. It will be located at a pair of variables $(m_0, m_1)$, where $m_0$ and $m_1$ would also be the parameter values emerging from a least-square fit of a linear model to the data in the scatter plot. Finally, this new $\chi^2$-function will have a new minimum value $\chi^2_{min}$, certain to be no higher than the previous $\chi^2_{min}$ – with a new model which is a superset of the old model, you're sure to get at least as good a fit as before. In fact, you'd expect $\chi^2_{min} \approx 498$ (for 500 data points less 2 fit parameters giving 498 degrees of freedom).

If the previous $\chi^2_{min}$ was well above 499, and the new $\chi^2_{min}$ is around 498 [ both in the sense of $500(1 \pm \text{a few}/\sqrt{(2\cdot 500)})$ ], then you have good evidence for a non-zero value of $\mu_1$, which you'd think of as a measure of an (undesired) 'drift rate'.

There is yet one more plot you can make, given a long $\{C_i\}$ list, and that is a *histogram*. For a good-looking histogram, you'd like both

- a long list, perhaps of 1000 or more values of $C_i$;
- a 'stable' list, arising from an experimental configuration that you have shown does *not* suffer from significant drift.

The essence of a histogram is to 'bin' the individual $C_i$-values into a set of bins of equal width. The 'bin width' you pick is your choice, but a bin-width choice of order $(1/5)\sqrt{C_{avg}}$ to $(1/2)\sqrt{C_{avg}}$ would work well. There will be individual occurrences $C_i$ which depart from $C_{avg}$ by several multiples of $\sqrt{C_{avg}}$, so you'll get 10-20 bins with non-zero contents. Now you form and plot $B_j$, the number of occasion of $Ci$-values falling into each bin, as a function of $\mu_j$, the central value for each bin.

[Example: if $C_{avg} = 900$, so that $\sqrt{C_{avg}} = 30$, then a good choice for bin width might be 10. Thus there is one bin whose contents would be incremented by one for each $C_i$ -value falling in the range 895-904 (inclusive), and that bin's center would have a $\mu_j$-value of 900. An adjacent bin would be allocated to counting occurrences of $C_i$ -values in the range 905 to 914 (inclusive), and that bin would be labeled by $\mu_{j+1} = 910$. And so on.]

There are software packages which can form histograms automatically for you out of a $\{C_i\}$-list, though it might be harder to get them to use the bin widths, and boundaries, that you prefer. But however you form the histogram, you should make a scatter-plot, or a bar-chart, for ($\mu_j$, $B_j$) combinations. That's because you expect a Poisson distribution (for large $C_{avg}$, indistinguishable from a Gaussian distribution) to emerge. In fact, the histogram should be described by a model with no more free parameters than the length of the $\{C_i\text{-}\}$ list you're using, and the $C_{avg}$ which you've already obtained from it. Again, your eye can be quite proficient at seeing a misfit between the experimental histogram and the theoretical (Poisson or Gaussian) distribution which can be plotted atop it.

Once you have this technology working, here are two investigations you could try:

1. With the PMT delivering just a background count rate (such as with the shutter closed) of order 100/s, and using a gate time of 0.1 seconds, you'll get $C_i$ -values with a mean count of about 10. You will also be able to computer-record about 10 such readings each second. So it will be easy to get a list of $>10^4$ $C_i$ -values. These you can histogram into bins of width *1*, and still get lots of occurrences $B_j$ in each of the bins, labeled by value $\mu_j = j$, $j$ running from 0 upwards. With a mean value as low at 10 counts, you should plot both a Gaussian and a Poisson distribution, each with the mean (and the variance) expected for a data set having your mean value. You will then be able both to see if the difference between the two models is visible, and to determine if your experimental histogram is consistent with either model.
2. Back to the PMT delivering count rates from a 2-slit interference pattern, you could acquire data set 'A' of $\{C_i\}$-values obtained at an interference maximum, and data set 'B' obtained at an adjacent interference *minimum*. Clearly $C_{avg}{}^{(B)}$ will be less, perhaps well under one-tenth, of $C_{avg}{}^{(A)}$. (The actual ratio will depend on the *fringe contrast* you achieved in the alignment of your Two-Slit apparatus.) But the statistics question is: are both data sets consistent with a Poisson process? If so, this is good evidence in favor of the hypothesis that, in both cases 'A' and 'B', the events being counted arise from the statistically-independent arrival of photons. This, in turn, is good evidence *against* the hypothesis that an interference maximum is a result of cooperation, or correlation, between pairs of photons.

**Time Intervals between successive events**

Suppose you have set up the Two-Slit apparatus (or some other source of random-event pulses) for producing events at an average rate of 2000 per second. Clearly this requires that the average 'waiting time' between an event and its immediate successor in time is (1/2000) of a second, or (1/2) ms, or 500 μs. But what about the *individual* waiting times, say between events #1 and #2, or between events #$M$+1 and #$M$+2, or between a later pair #$N$+1 and #$N$+2? These values are exactly what you can capture using the *Interval* mode of the PCIT.

You should operate the Counter first in the Manual way, for which each depression of the spring-loaded toggle switch will give an individual value of $D_i$ – the 'waiting time', in μs, between the first, and the second, input pulse which arrive at the counter input following your finger command. Transcribe about a dozen of these $D_i$-readings by hand. First, check to see if the mean, or average, waiting time is consistent with what you expect (given the average count rate which you can measure in the pulse-count mode).

Now, look at the *scatter* among these dozen values. You should notice two interesting things.

- There is *much* greater scatter than the $500 \pm \sqrt{500}$ characteristic spread you've seen in counting experiments, and
- The distribution of $D_i$-values is *not* symmetrical about the mean: if the mean of $\{D_i\}$-values is 500 μs, you'll see many more occurrences in the 250-500 μs range than in the 500-750 μs range.

This is meant to whet your appetite to get *much* longer $\{D_i\}$ lists than you'd care to acquire by hand. That's the motivation for the AUTO mode of the PCIT, where $D_i$ values cease to show up on the Counter's display, but appear instead as successive entries in a growing file on a USB-connected computer. In fact, such a list grows rapidly – you need (on average) only wait half a millisecond for an event pair, plus a few more ms for data-transfer time, for each $D_i$ -value to be added to the file. Thus *hundreds* of $D_i$ -values can be recorded in each second of real time.

Impressively long lists of $D_i$-values can be accumulated. [If the PCIT software should 'hang up' after creating a list of finite length, you can get data transfer to resume by hand-toggling the sequence Auto → Manual → Auto. The computer file should then resume getting $D_i$ values from the same underlying distribution. You can piece together enough segments this way to get a $\{D_i\}$-list of any length you choose. A set of at least $10^4$ samples will give interesting results.]

In our example, these $D_i$-values form a list with a mean waiting time $D_{avg}$, of 500 μs. In fact $D_{avg}$ should be the reciprocal of the mean count rate, as measured using the pulse-count mode of the counter's operation on the same stream of physical events.

These $D_i$ -values also have a variance about their mean, defined by $< (D_i - D_{avg})^2 >$, the mean of the square of the deviation from the average. Unlike the case of count data, this variance is ***not*** expected to be as small as $1 \cdot D_{avg}{}^1$. What *is* the expected variance? How *does* the measured variance change as you change the $D_{avg}$ value? (Recall that dimming the bulb gives a lower average count rate, and thus a longer average waiting time.)

Finally, these $D_i$ -values can be displayed in a histogram where they will have *neither* a Poisson nor a Gaussian distribution. Instead, the histogram will be *exponential*, having a mean of $D_{avg}$, but a markedly unsymmetrical distribution about the mean. The mode, i.e. the most common occurrence of this distribution, is *zero*. This means that the shorter the waiting time, the more likely its occurrence! For instance, if $D_{avg}$ is 500 μs and your long list has 25 $D_i$-values falling into the 510-519 μs bin of your histogram, there will be *more* than 25 $D_i$-values falling in the 260-269 μs bin of the histogram, and *still more* than that many $D_i$-values falling in the 10-20 μs bin of the histogram. A bit counter intuitive isn't it?

You'll need to read about waiting-time distributions before you can reconcile this frequent occurrence of short waiting times with your knowledge of the statistical *in*dependence of the individual pulse events in your stream of photon detections. That is to say, you will discover that events which are truly random-in-time are not 'self-avoiding', and therefore must display a degree of apparent 'clumping'. This is *not* evidence against their randomness, but in fact evidence in support of their statistically-independent occurrence.