1 Some Statistical Basics.

Statistics treats random errors. (There are also *systematic* errors; e.g., if your watch is 5 minutes fast, you will always get the wrong time, but it won't be random.) The two most important equations which describe the distribution of measured values in the presence of random error are the Gaussian distribution and the Poisson distribution. The first treats measured quantities which have a continuous range of values, while the second applies to cases where only discrete results can occur (e.g., the number of radioactive nuclei that decay in a given interval of time: there might be 41 decays or 42 decays but never 41.3744 ...).

1.1 The Poisson Equation

Let us consider a process which produces discrete events, but with a fixed *average* rate R per unit time. That is, if we count the events for a very long time, then

$$\lim_{t\to\infty} \left\{ \frac{\text{number of events}}{\text{length of time}} \right\} \;\to\; R \tag{1}$$

If we then observe for a (short) time t, then the *expected* number of events will be $N \;=\; R\,t$. But if the individual events are random, the result may not be N, it may be more or less than N. In fact, there is no reason why N should be an integer, while the number n we actually observe in time interval t *must* be integral. Given an expected number N, the Poisson distribution, $P_n(N)$, is the probability that our measurement will find a particular number of events n.

The Poisson equation is given by the expression:

$$\boxed{P_n(N) \;=\; \frac{N^n}{n!}\, e^{-N}} \tag{2}$$

Any observation is certain to yield *some* number of events, either none ($n \;=\; 0$), or $n \;=\; 1, 2, 3, \cdots$. So if we add up all the $P_n(N)$ for a given N, we must obtain unity. Let's check that:

$$\sum_{n=0}^{\infty} P_n(N) \;=\; \sum_{n=0}^{\infty} \frac{N^n}{n!}\, e^{-N} \;=\; e^{-N} \left[1 \;+\; N \;+\; \frac{1}{2!}N^2 \;+\; \frac{1}{3!}N^3 \;+\; \frac{1}{4!}N^4 \;+\; \cdots \right] \tag{3}$$

But the term in $[\ \cdots\ ]$ is just the series expansion of e^N, so

$$\sum_{n=0}^{\infty} P_n(N) \;=\; e^{-N}\, e^{N} \;=\; 1 \;. \tag{4}$$

Now suppose we took a whole series of observations, each for a length of time t, so that the expected value each time is N. The actual number n seen in each observation will vary, but we could average the results of the series to find some mean number $\overline{n}$. If the Poisson distribution is obeyed, then we can compute the mean number we should see: We multiply the number of events observed, n, by the probability $P_n(N)$ that that number occurs, and sum over all possible values of n:

$$\overline{n} \;=\; \sum_{n=0}^{\infty} n\, P_n(N) \;=\; \sum_{n=0}^{\infty} \frac{N^n}{(n-1)!}\, e^{-N} \;=\; N\, e^{-N} \sum_{n=1}^{\infty} \frac{N^{n-1}}{(n-1)!} \tag{5}$$

Now the last sum is just e^N again, as we see if we let $n-1=k$:

$$\overline{n} \;=\; N\,e^{-N}\,\sum_{k=0}^{\infty}\frac{N^k}{k!} \;=\; N\,e^{-N}\,e^{N} \;=\; N \quad . \tag{6}$$

This is just a check that equation (2) makes sense, since we already said that N represents the expected (average) number of events observed.

Lets look at some examples of this distribution:

N	$P_0(N)$	$P_1(N)$	$P_2(N)$	$P_3(N)$	$P_4(N)$	$P_5(N)$	$P_6(N)$	$P_7(N)$	$P_8(N)$
1	0.367879	0.367879	0.183940	0.061313	0.015328	0.003066	0.000511	0.000073	0.000009
1.5	0.22313	0.334695	0.251021	0.125511	0.047067	0.014120	0.003530	0.000756	0.000142
2	0.135335	0.270671	0.270671	0.180447	0.090224	0.036089	0.012030	0.003437	0.000859
2.5	0.082085	0.205212	0.256516	0.213763	0.133602	0.066801	0.027834	0.009941	0.003106
3	0.049787	0.149361	0.224042	0.224042	0.168031	0.100819	0.050409	0.021604	0.008102
5	0.006738	0.033690	0.084224	0.140374	0.175467	0.175467	0.146223	0.104445	0.065278
10	0.000045	0.000454	0.002270	0.007567	0.018917	0.037833	0.063056	0.090079	0.112599

After the mean value $\overline{n}$, the most important quantity which describes the Poisson distribution is some measure of how much, on the average, a particular value n will differ from the mean. You might think that you could just take the quantity of interest, $n-\overline{n}$, multiply by $P_n(N)$, and sum over all n. That would tell us nothing, because the result would be zero. To see why, consider the values of the $\Delta n = n-\overline{n}$: there will be as many negative as positive values. (Try it and see.) The point is that we don't so much care about the sign of Δn as its magnitude. We could compute the mean value of the absolute values $|\Delta n|$. But absolute values are not so easy to work with. It turns out that the best quantity to consider is the mean of the *squares* of the Δn, which are of course always positive. This is called the **variance**, V, of the distribution:

$$V \;=\; \sum_{n=0}^{\infty}(\Delta n)^2\,P_n(N) \;=\; \sum_{n=0}^{\infty}(n-\overline{n})^2\,P_n(N) \;=\; \sum_{n=0}^{\infty}(n-N)^2\,P_n(N) \quad . \tag{7}$$

Let's try to evaluate this. Expand the square to get

$$V \;=\; \sum_{n=0}^{\infty} n^2\,P_n(N) \;-\; 2N\,\sum_{n=0}^{\infty}\,n\,P_n(N) \;+\; N^2\,\sum_{n=0}^{\infty}\,P_n(N) \tag{8}$$

From equations (4), (5), and (6), we already have the values of the second two sums. Thus the last terms become $-2N\times N + N^2 \;=\; -N^2$, and we have

$$V \;=\; \sum_{n=0}^{\infty} n^2\,P_n(N) \;-\; N^2 \quad . \tag{9}$$

Now we must evaluate mean of the squares, $\overline{n^2}$. This is just

$$\overline{n^2} \;=\; \sum_{n=0}^{\infty} n^2\,P_n(N) \;=\; \sum_{n=1}^{\infty} n^2\,\frac{N^n}{n!}e^{-N} \quad , \tag{10}$$

since the $n=0$ term of the series is zero. We need to manipulate this a bit:

$$\overline{n^2} \;=\; Ne^{-N}\,\sum_{n=1}^{\infty}\frac{nN^{n-1}}{(n-1)!} \;=\; Ne^{-N}\,\sum_{k=0}^{\infty}\frac{(1+k)N^k}{k!} \;=\; Ne^{-N}\,\left[\;\sum_{k=0}^{\infty}\frac{N^k}{k!} \;+\; \sum_{k=0}^{\infty}\frac{kN^k}{k!}\;\right]$$

$$\overline{n^2} \;=\; Ne^{-N}\left[\, e^N \;+\; N\sum_{k=1}^{\infty}\frac{N^{k-1}}{(k-1)!}\,\right] \;=\; Ne^{-N}\left[\, e^N \;+\; N\sum_{n=0}^{\infty}\frac{N^{n}}{n!}\,\right] \;=\; N \;+\; N^2$$

Going back to equation (9), we thus obtain our result

$$V \;=\; \overline{n^2} \;-\; N^2 \;=\; \left[\, N \;+\; N^2 \,\right] \;-\; N^2 \;=\; N \tag{11}$$

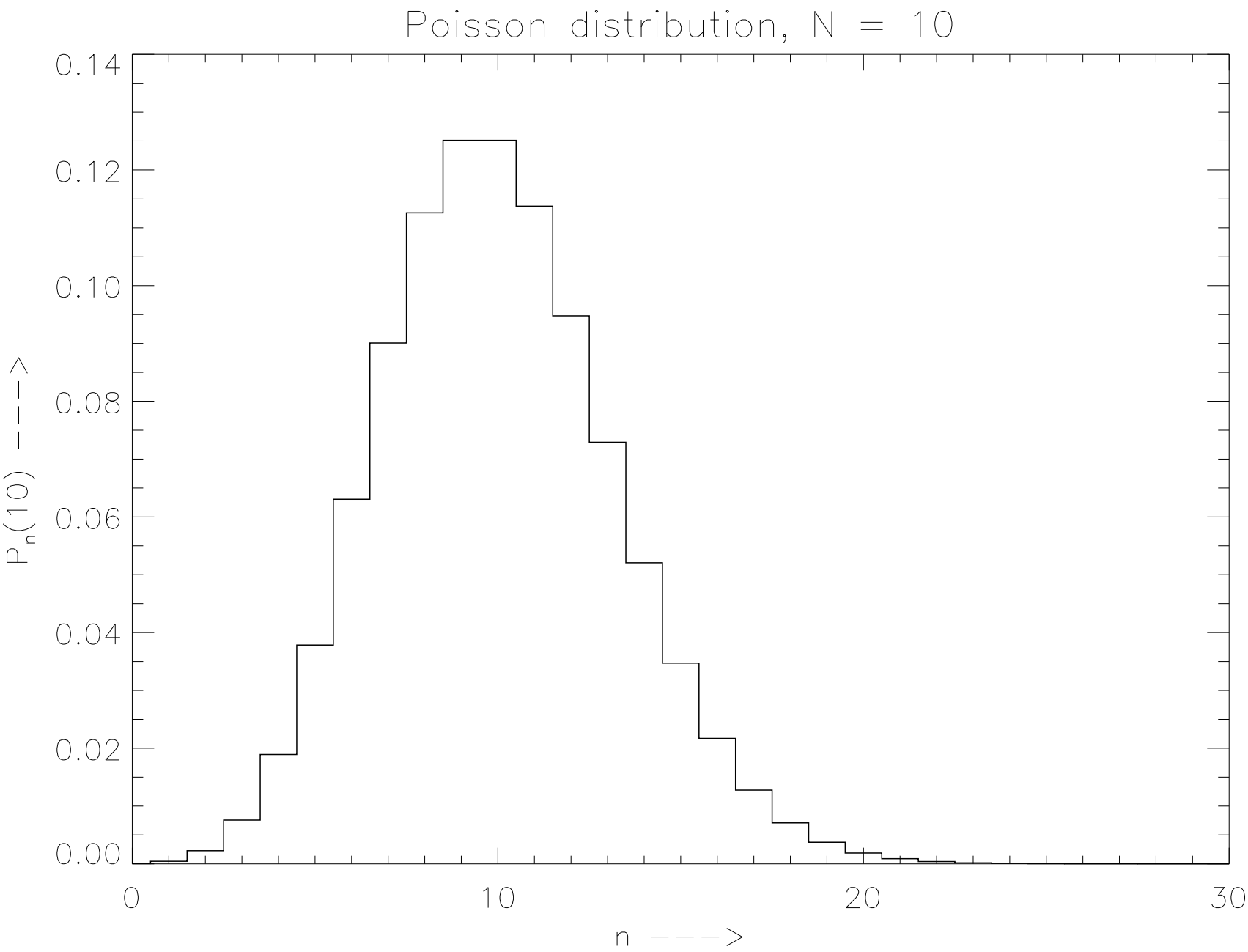
Since the variance involves the square of Δn, we often take the square root of the variance, which is called the *standard deviation* or *root mean square deviation*, and is denoted by the Greek letter σ:

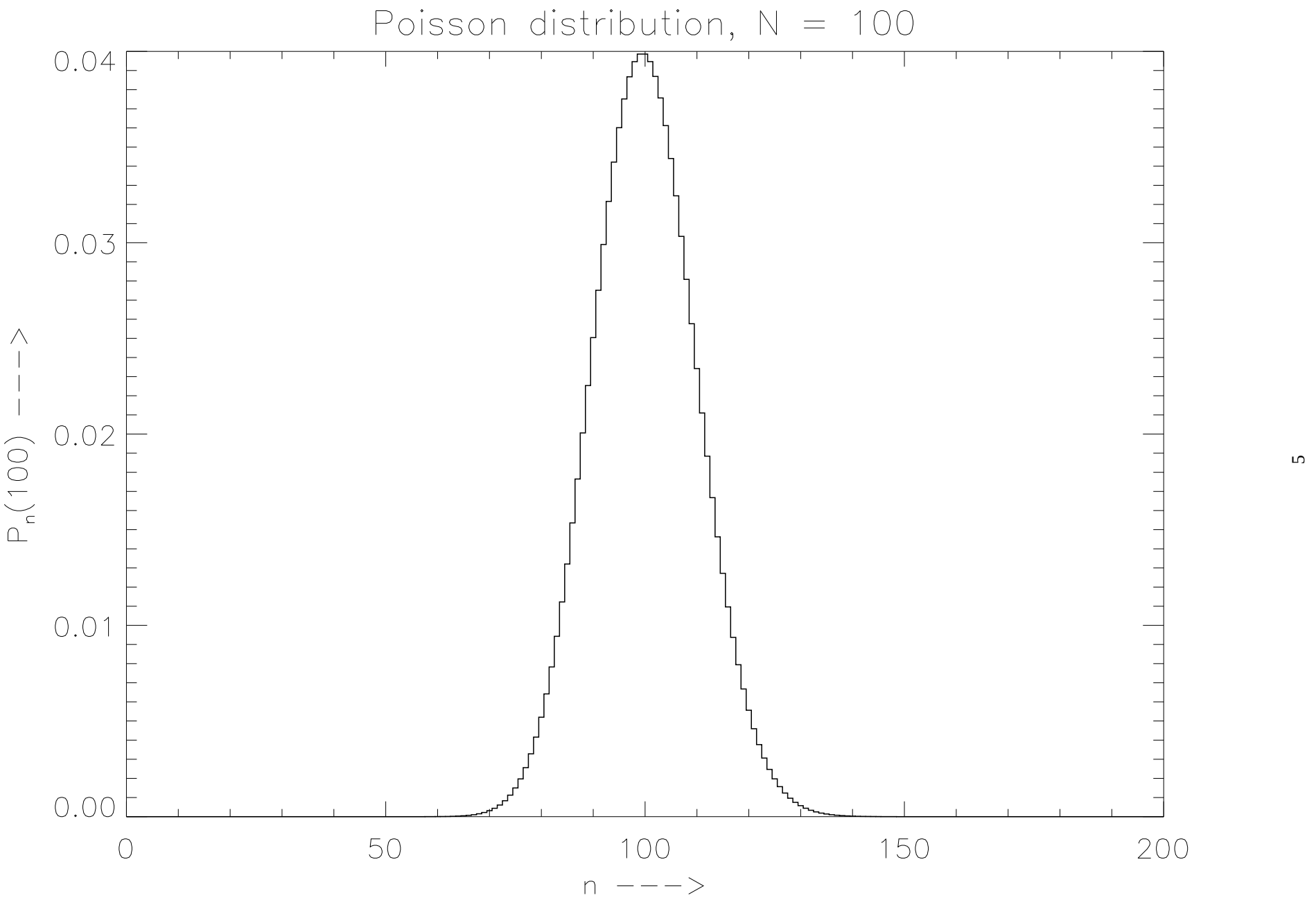
For the Poisson distribution: $\sigma \;\equiv\; \sqrt{V} \;=\; \sqrt{N}$ (12)

There are two important points: First, the spread in the values of n about $\overline{n}$ increases as $\overline{n}$ increases. Second, the spread (σ) does not increase as as fast as $\overline{n}$. Thus, the *relative* scatter in the measured values *decreases* as $1/\sqrt{N}$:

N	$\sigma\ (=\sqrt{N})$	σ/N
10	3.16228	0.316228
100	10	0.100
1000	31.6228	0.0316228
10000	100	0.01000

To see how this applies to astronomical measurements, consider the case where we have a photo-multiplier tube attached to a telescope, and we are counting the photons from a faint star. What we really want to measure is the average rate of photons, R, as this is proportional to the star's brightness. If we count for some length of time t, then we can find R from equation (1). If the number we counted was N, then $R = N/t$ would be the right answer. But what we actually measure is some number n drawn from the Poisson distribution. How much error we make depends on how far n is likely to be from N. Now since N increases in proportion to t, the longer we count, the smaller σ/N will be, and the closer we are likely to be to the true brightness. But the accuracy does not increase very fast, only as $\sqrt{t}$. E.g., if you get 100 counts in one minute, which implies an accuracy of 10%, it will take 100 minutes to get 1% accuracy!


Poisson distribution, N = 10
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
$P_n(10)$ ---->
0
10
20
30
n --->


Poisson distribution, N = 100
0.04
0.03
0.02
0.01
0.00
$P_n(100)$ ---->
0
50
100
150
200
n --->

1.2 The Gaussian Distribution

The Gaussian distribution applies to continuous variables rather than the discrete events described by the Poisson distribution. The two are in fact related, and the histogram of a Poisson distribution for large N looks very much like a Gaussian distribution. Suppose we have a measured quantity x which can take on a continuous range of values $[-\infty < x < +\infty]$. Then the formula for the Gaussian distribution is:

$$\boxed{P_\sigma(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\overline{x})^2}{2\sigma^2}}} \tag{13}$$

The distribution is a function of two parameters: the expected or *mean value* $\overline{x}$ (like $\overline{n}$ in the Poisson distribution), and the *standard deviation* σ. We interpret this formula in the following way: $P_\sigma(x)\,dx$ is the probability that our measurement will produce a value of x in the interval $[x, (x+dx)]$. Since any measurement must yield *some* value of x, integrating over all possible x must give unity:

$$\int_{-\infty}^{\infty} P_\sigma(x)\,dx \;=\; 1 \;. \tag{14}$$

The Gaussian distribution has it's maximum value at $x = \overline{x}$, and that value is $P_\sigma(\overline{x}) = 1/\sqrt{2\pi}\sigma$. If we consider values one σ removed from $\overline{x}$, $x = \overline{x} \pm \sigma$, then from equation (13) we see that the value of $P_\sigma(x)$ will drop by a factor of $e^{-\frac{1}{2}} = 0.606531\cdots$. If we move out to $x = \overline{x} \pm 2\sigma$, the values will drop by $e^{-2} = 0.135335\cdots$.

In astronomy, you will often see a measure of the width of the Gaussian distribution which is a bit easier to visualize. This is called the *full width at half-maximum*, or "FWHM". It is the width of the distribution at the point at which it has dropped to half its maximum value. This will happen for a value of x which satisfies the relation $P_\sigma(x) = \frac{1}{2}P_\sigma(\overline{x})$. From the definition (13), we see that this implies

$$e^{-\frac{(x-\overline{x})^2}{2\sigma^2}} \;=\; \frac{1}{2}$$

thus

$$\frac{(x-\overline{x})^2}{2\sigma^2} \;=\; \ln(2)$$

and

$$(x-\overline{x}) \;=\; \pm\sqrt{2\ln(2)}\,\sigma \;=\; 1.17741\cdots\sigma$$

Since the *full-width* is just twice this value, we see that the relation between the FWHM and the standard deviation is:

$$\text{FWHM} \;=\; 2\sqrt{2\ln(2)}\,\sigma \;=\; 2.35482\,\sigma \tag{15}$$

We also might like to know the probability that the measured x will fall within a certain range. Thus we can look at the probability that x will be within one σ of the mean:

$$P[1\sigma] \;=\; \int_{\overline{x}-\sigma}^{\overline{x}+\sigma} P_\sigma(x)\,dx \;=\; \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-\frac{1}{2}y^2}\,dy \;=\; 0.6828\cdots \tag{16}$$

Likewise, the probability of falling within 2σ of the mean is

$$P[2\sigma] \;=\; \frac{1}{\sqrt{2\pi}} \int_{-2}^{2} e^{-\frac{1}{2}y^2}\, dy \;=\; 0.9544\cdots \tag{17}$$

The corresponding probability of falling further away than 2σ is thus $1 - P[2\sigma] \;=\; 0.0456$. With increasing distance from $\overline{x}$, the probabilities decrease very rapidly; e.g., $P[3\sigma] \;=\; 0.9974$ so that $1 - P[3\sigma] \;=\; 0.0026$.

The standard deviation (if we can estimate it!) is sometimes used to put "error bars" on measured points on plots. Notice that from (16) you would expect almost 1/3 of the points to differ from the true value by *more than the error bars*.

Another quantity sometimes employed is the *probable error* (p.e.). This is the distance from the mean such that one half the values of x should be found within $\overline{x} \pm$ p.e. It turns out that the p.e. = 0.674 σ.

So far we have been talking about $\overline{x}$ and σ as if we knew their values. But for measured quantities, we would need an infinite number of measurements to find them exactly. We can only *estimate* them. It can be shown $\cdots$ — that the best estimate of $\overline{x}$ from N measurements is

$$\overline{x}_N \;=\; \frac{1}{N} \sum_{i=1}^{N} x_i \tag{18}$$

Furthermore, the best estimate of σ is

$$\sigma_N \;=\; \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(x_i - \overline{x}_N\right)^2} \tag{19}$$

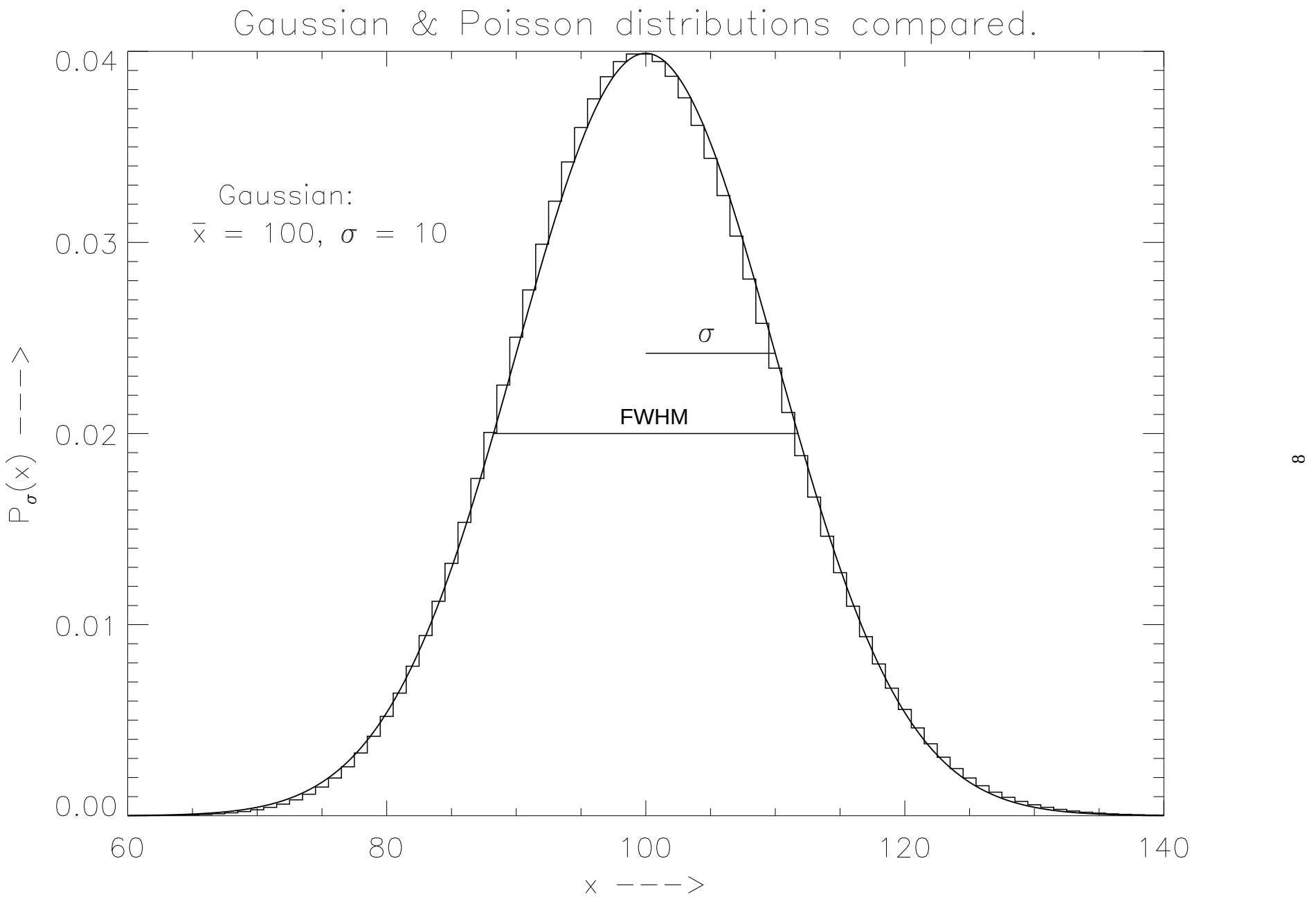
This quantity σ_N is the standard deviation of a single measurement. How good is our estimate $\overline{x}_N$? It turns out that

$$\overline{x} \;=\; \overline{x}_N \;\pm\; \sigma_{\overline{x}} \;=\; \overline{x}_N \;\pm\; \frac{\sigma_N}{\sqrt{N-1}} \tag{20}$$

The quantity

$$\sigma_{\overline{x}} \;=\; \frac{\sigma_N}{\sqrt{N-1}} \tag{21}$$

is called the *standard deviation of the mean*. Note that $\overline{x}_N$ approaches the true value of the mean as $\sim 1/\sqrt{N}$ (just like the error in the Poisson distribution).


Gaussian & Poisson distributions compared.
Gaussian:
$\bar{x}$ = 100, σ = 10
σ
FWHM
0.04
0.03
0.02
0.01
0.00
60
80
100
120
140
$P_\sigma(x)$ ---->
x --->

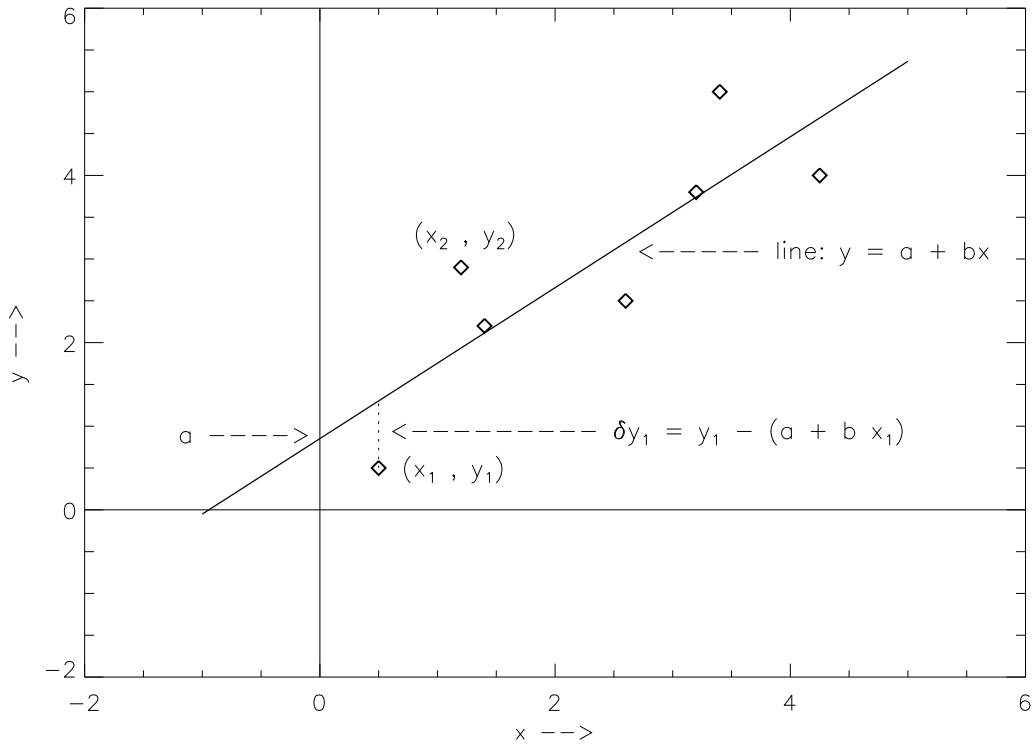
1.3 Least-Squares Fit to a Straight Line

We want to find the line which is the "best fit" to a set of points (x_i, y_i). The line will be defined in terms of its slope b and y-intercept a:

$$y \;=\; a \;+\; b\,x \tag{22}$$

Then the error in y of the first point with respect to the line is

$$\delta y_1 \;=\; y_1 \;-\; (\;a \;+\; bx_1) \tag{23}$$



Let us choose as our measure of error the sum of the squares of all the individual errors:

$$V \;=\; \sum_{i=1}^{N} (\delta y_i)^2 \;=\; \sum_{i=1}^{N} [\; y_i \;-\; (\;a \;+\; bx_i)\;]^2 \tag{24}$$

We want to find the line for which V is a minimum. Thus we want to adjust the parameters a and b to obtain this minimum, and this will happen when

$$\frac{\partial V}{\partial a} \;=\; 0 \quad \text{and} \quad \frac{\partial V}{\partial b} \;=\; 0 \quad . \tag{25}$$

Expanding the square in equation (24), we have

$$V \;=\; \sum_{i=1}^{N} \Big\{\; y_i^2 \;-\; 2y_i a \;-\; 2y_i b x_i \;+\; a^2 \;+\; 2abx_i \;+\; b^2 x_i^2 \;\Big\} \tag{26}$$

We then see that the partial derivatives are

$$\frac{\partial V}{\partial a} \;=\; \sum_{i=1}^{N} \{\; -2y_i \;+\; 2a \;+\; 2bx_i \;\} \tag{27}$$

$$\frac{\partial V}{\partial b} \;=\; \sum_{i=1}^{N} \Big\{\; -2x_i y_i \;+\; 2ax_i \;+\; 2bx_i^2 \;\Big\} \tag{28}$$

Canceling the factors of 2, we see that these equations are just

$$N\,a \;+\; b\sum_{i=1}^{N} x_i \;=\; \sum_{i=1}^{N} y_i \tag{29}$$

$$a\sum_{i=1}^{N} x_i \;+\; b\sum_{i=1}^{N} x_i^2 \;=\; \sum_{i=1}^{N} x_i y_i \tag{30}$$

It saves the trouble of writing the summation signs if we use the over-bar notation:

$$\overline{x} \;=\; \frac{1}{N}\sum_{i=1}^{N} x_i \;\;, \qquad \overline{xy} \;=\; \frac{1}{N}\sum_{i=1}^{N} x_i y_i \;\;, \quad \text{etc.} \tag{31}$$

Then equations (29) and (30) become

$$a \;+\; b\overline{x} \;=\; \overline{y} \tag{32}$$

$$a\overline{x} \;+\; b\overline{x^2} \;=\; \overline{xy} \tag{33}$$

We may use equation (32) to eliminate a from equation (33) to obtain

$$(\overline{y} \;-\; b\overline{x})\,\overline{x} \;+\; b\overline{x^2} \;=\; \overline{xy} \tag{34}$$

Rearranging, we have

$$b\,(\overline{x^2} \;-\; \overline{x}\cdot\overline{x}) \;=\; \overline{xy} \;-\; \overline{x}\cdot\overline{y} \tag{35}$$

Thus, the slope, b, of the *least-squares fit* line is

$$\boxed{b \;=\; \frac{\overline{xy} \;-\; \overline{x}\cdot\overline{y}}{\overline{x^2} \;-\; \overline{x}^2} \;=\; \frac{N\sum_{i=1}^{N} x_i y_i \;-\; \sum_{i=1}^{N} x_i \;\sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 \;-\; \left(\sum_{i=1}^{N} x_i\right)^2}} \tag{36}$$

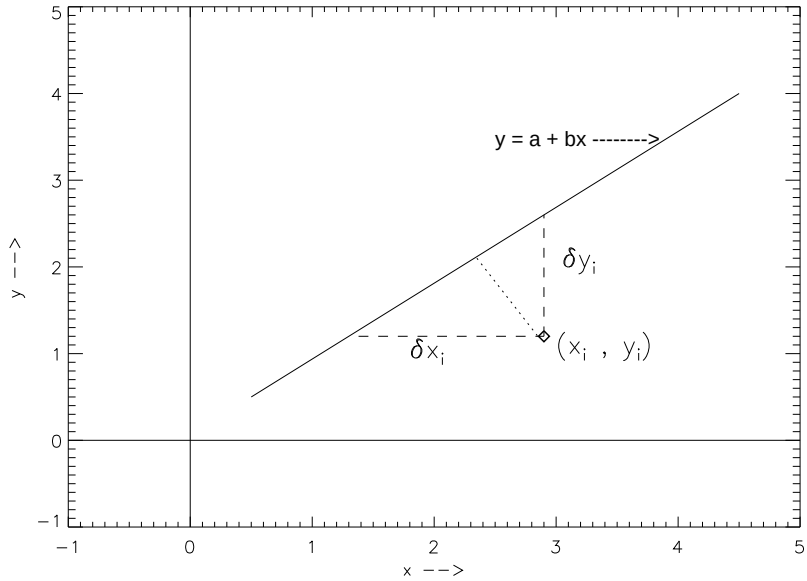
We now substitute this expression for b into equation (32) and solve for the intercept a:

$$a \;=\; \overline{y} \;-\; \frac{\overline{xy} \;-\; \overline{x}\cdot\overline{y}}{\overline{x^2} \;-\; \overline{x}^2}\,\overline{x} \tag{37}$$

which reduces to

$$\boxed{a \;=\; \frac{\overline{y}\cdot\overline{x^2} \;-\; \overline{xy}\cdot\overline{x}}{\overline{x^2} \;-\; \overline{x}^2} \;=\; \frac{\sum_{i=1}^{N} x_i^2 \;\sum_{i=1}^{N} y_i \;-\; \sum_{i=1}^{N} x_i \;\sum_{i=1}^{N} x_i y_i}{N\sum_{i=1}^{N} x_i^2 \;-\; \left(\sum_{i=1}^{N} x_i\right)^2}} \tag{38}$$

Note that this derivation assumes that *all the error is in y!* This may or may not be a good assumption. Perhaps the errors are in x. Would we get the same result if we were to minimize the δx_i? The answer is no.



Consider the point on the line which has the same y-value as some point (x_i, y_i): The value of x on the line at $y = y_i$ is $x = (y_i - a)/b$. Thus the expression for δx_i will be

$$\delta x_i \;=\; x_i \;-\; \frac{y_i \;-\; a}{b} \tag{39}$$

We could go through the same process and minimize

$$V' \;=\; \sum_{i=1}^{N} (\delta x_i)^2 \;=\; \sum_{i=1}^{N} \left[\, x_i \;-\; \frac{1}{b}(y_i \;-\; a) \,\right]^2 \tag{40}$$

However, it is simpler to just write the line equation (22) in the form

$$x \;=\; \left(\; -\; \frac{a}{b} \,\right) \;+\; \left(\, \frac{1}{b} \,\right)\, y \;=\; A \;+\; B\, y \tag{41}$$

Then we see from (36) and (38) that

$$B \;=\; \frac{\overline{x^2} \;-\; \overline{x}^2}{\overline{xy} \;-\; \overline{x} \cdot \overline{y}} \tag{42}$$

and

$$A \;=\; \frac{\overline{xy} \cdot \overline{x} \;-\; \overline{y} \cdot \overline{x^2}}{\overline{xy} \;-\; \overline{x} \cdot \overline{y}} \tag{43}$$

Remember that we have assumed the errors are in y. Now, let us simply rename x as y and y as x. Then the equation of the line is

$$y \;=\; A \;+\; B\, x \tag{44}$$

and the least-squares fit, *with the errors now in* x, are just

$$B \;=\; \frac{\overline{y^2} \;-\; \overline{y}^2}{\overline{xy} \;-\; \overline{x}\cdot\overline{y}} \tag{45}$$

and

$$A \;=\; \frac{\overline{xy}\cdot\overline{y} \;-\; \overline{x}\cdot\overline{y^2}}{\overline{xy} \;-\; \overline{x}\cdot\overline{y}} \tag{46}$$

We see that $A \neq a\ ,\ \ B \neq b$. Furthermore, we might have a case where the errors in the determination of x and of y were comparable. Then perhaps we would want to minimize the squares of the distance from the point to the line d_i:

$$(d_i)^2 \;=\; \frac{\delta x_i^2\ \delta y_i^2}{\delta x_i^2 \;+\; \delta y_i^2} \tag{47}$$

This would give us yet another line, which would lie between the first two.

Finally, let us mention the *correlation coefficient*, r, This quantity is 0 if there is no correlation at all between the points, and reaches ± 1 for a collection of points which lie exactly on a straight line:

$$r \;=\; \frac{N\sum_{i=1}^N x_i y_i \;-\; \sum_{i=1}^N x_i\ \sum_{i=1}^N y_i}{\sqrt{\left[N\sum_{i=1}^N x_i^2 \;-\; \left(\sum_{i=1}^N x_i\right)^2\right]\left[N\sum_{i=1}^N y_i^2 \;-\; \left(\sum_{i=1}^N y_i\right)^2\right]}} \tag{48}$$