

## Chem 114      Statistics Lectures

In all of our experiments, we are going to try to measure some characteristic of some solution, solid, materials, etc. Say that we label this value  $V$

The final answer that is presented in our report (and in the abstract of our report) is not just  $V$ , but rather

$$V \pm \Delta V \quad (1)$$

In this case,  $\Delta V$  is the uncertainty in  $V$ . This series of lectures is devoted to building an understanding on obtaining a reliable  $\Delta V$ .

Assume we do a lab in which we measure the volume of a box. In this case,

$$\text{Volume} = V_o = H_o \times L_o \times W_o \quad (2)$$

The uncertainty in volume is going to be dependent upon how well we measure the three quantities (height, length, and width) upon which it depends. Since the error associated with each individual measurement is going to propagate our experimental measurement, impacting the error associated with the final measurement, we do what is called a "Propagation of error" calculation. ***In each lab, you will do such a calculation, setting up the equation that is particular to your own lab experiment.***

For the case of the volume of the box, the equation looks something like the following:

$$V = V_o \pm \Delta L (\delta V / \delta L)_{W_o H_o} + \Delta W (\delta V / \delta W)_{L_o H_o} + \Delta H (\delta V / \delta H)_{L_o W_o} \quad (3)$$

which can readily be solved to yield

$$V = V_o \pm \Delta L \times W_o \times H_o + \Delta W \times L_o \times H_o + \Delta H \times L_o \times W_o \quad (4)$$

So now we know that  $\Delta V$  comes from  $\Delta W$ ,  $\Delta L$ , and  $\Delta H$ . Our problem, rather than getting simpler, is rapidly getting more complicated! Furthermore, we also need to define three other new constants if we want to solve (4) --  $W_o$ ,  $H_o$ , and  $L_o$ . Let's deal with these first.

Suppose that we have just measured the width of the box 11 times, and arrived at the following numbers:

Width <sub>i</sub>	Column 1	
82.9		
83.5	Mean	83.52
83.7	Standard Error	0.12
83.5	Median	83.5
83.9	Mode	82.9
83.4	Standard Deviation	0.40
83.2	Variance	0.16
82.9	Kurtosis	-0.788537037
83.6	Skewness	-0.249861798
84	Range	1.2
84.1	Minimum	82.9
	Maximum	84.1
	Sum	918.7
	Count	11

In the table above, the measurements are given in the left hand column. In the middle and right columns are statistical terms and their respective numbers. Let's discuss what these statistical terms are, one at a time.

The **median** is defined as the following: The median of the parent population (called  $\mu_{1/2}$ ) is defined as that value for which, in the limit of an infinite number of determinations of  $W_i$ , half the observations will be less than the median, and half will be greater.

The **mean** of N measurements of W is defined as:

$$W_o = \mu = 1/N(\sum W_i) \quad (5)$$

For the given example, N = 11, and  $\mu$  is the symbol representing the mean. Hopefully, the concepts of the median and the mean are not new to you.

The **mode** is the most likely value. In the above example, it is listed as 82.9, but it could just as easily have been listed as 83.5, since there are two occurrences of each of these values in the table of observations. This value only takes significance in the limit of a large number of observations, and obviously such a limit has not been reached here.

Now we come to numbers that describe the spread in the data. The first number is the **variance**, or  $\sigma^2$ .

$$\sigma^2 = \lim [ (1/N) \sum (x_i - \mu)^2 ] = \lim [ (1/N) \sum x_i^2 ] - \mu^2 \quad (6)$$

The **standard deviation** is the square root of the variance, and is thus denoted as  $\sigma$ .

The **average deviation**, called  $\alpha$ , is defined as the average of the absolute values of the deviations:

$$\alpha = \lim \left[ (1/N) \sum |x_i - \mu| \right] \quad (7)$$

The average deviation is a measurement of the dispersion of the expected observations about the mean. Although it is a fine quantity for defining uncertainties, the fact that (7) contains an absolute value makes  $\alpha$  a difficult quantity to deal with in statistical analysis. The variance and the standard deviation are both much simpler to deal with computationally.

If one takes a large number of measurements of a single quantity, then it is very likely that the measurements, when plotted as (*value vs. frequency of occurrence*), will follow some functional form. This functional form is called the probability distribution. The probability distribution typically has the functional form of a **Gaussian Distribution**, a **Binomial Distribution**, or a **Poisson Distribution**, although other types of distributions are also possible. For now let's not worry about exactly what type of distribution is observed, let's just define a generic probability distribution function **P(x)**. There are two generic types of distributions, a **continuous distribution**, and a **discrete distribution**. We consider discrete distributions first.

### **Discrete Distributions**

If there are  $n$  possible different observable values of  $x$ , and if  $N$  observations have been made, then we have sampled the probability distribution function,  $P(x)$ ,  $N$  times. At the limit of an infinitely large number of observations,

$$\mu = \lim_{N \rightarrow \infty} \left( \frac{1}{N} \right) \sum_{j=1}^n x_j NP(x_j) = \lim_{N \rightarrow \infty} \sum_{j=1}^n x_j P(x_j) \quad (8)$$

$$\text{while the variance, } \sigma^2 = \sum_{j=1}^n (x_j - \mu)^2 P(x_j) = \sum_{j=1}^n [(x_j)P(x_j)]^2 - \mu^2 \quad (9)$$

actually, (9) is directly out of the statistical text by Bevington, and I believe that the  $P(x)^2$  (r.h.s.) is incorrect. The correct translation from the l.h.s. to the r.h.s. should be the following, I think:

$$\begin{aligned} (x_j - \mu)^2 P(x_j) &= (x_j^2 - 2x_j\mu + \mu^2)P(x_j) = x_j^2 P(x_j) - (2x_j\mu + 2\mu^2 - \mu^2)P(x_j) \\ &= 2\mu(x_j - \mu)P(x) - \mu^2 P(x) + x_j^2 P(x_j) \end{aligned}$$

Now, replacing the summation signs:  $\sum_{j=1}^n 2\mu(x_j - \mu)P(x_j) - \sum_{j=1}^n \mu^2 P(x_j) + \sum_{j=1}^n x_j^2 P(x_j)$

The first term goes to zero, leaving the last two terms. The summation over the  $P(x_j)$  in the middle term goes to 1, leaving  $\mu^2$ .

Finally, the expectation value of any function  $f(x) = \langle f(x) \rangle = \sum_{j=1}^n [f(x_j)P(x_j)]$  (10)

### Continuous Distributions

If the probability function is a continuous, smoothly varying function  $P(x)$  of the observed value of  $x$ , we replace the summation signs by an integral over all values of  $x$ , multiplied by the probability  $P(x)$ . The mean  $\mu$  becomes the first moment of the parent distribution:

$$\mu = \int_{-\infty}^{+\infty} xP(x)dx, \quad (11)$$

while the variance becomes  $\sigma^2 = \int_{-\infty}^{+\infty} (x - \mu)^2 P(x)dx = \int_{-\infty}^{+\infty} x^2 P(x)dx - \mu^2$  (12)

and the expectation value of any function of  $x$  is:

$$f(x) = \int_{-\infty}^{+\infty} f(x)P(x)dx \quad (13)$$

Recall from quantum mechanics that similar equations are used to determine the expectation values of various observables. In other words, if  $P(x)$  is the probability distribution, then the expectation value of some observable is simply the operator of that observable multiplied by the  $P(x)$ , and integrated over all space.

So now, let's recall the situation with our box measurement. We have three unknowns that we want to measure - the length, width, and height of the box, and then we want to determine from those measurements the volume of the box. We set up the 'Propagation of error' equation, and find that we need the average values of  $L$ ,  $W$ , and  $H$ , which we denoted  $L_o$ ,  $W_o$ , and  $H_o$ . We also need some measurement of how well we know  $L_o$ ,  $W_o$ , and  $H_o$ . We denoted these uncertainties as  $\Delta L$ ,  $\Delta W$ , and  $\Delta H$ . We have just outlined a way to get at some of these numbers, although we haven't yet quantified how good those numbers are. If we want  $L_o$ , for example, then we simply measure  $L$  many times, and use equation (5) to determine  $L_o$ . The variance can be determined from equation (6), and the standard deviation is the square root of the variance. Actually, this last statement needs to be modified just a little bit, and we will do that now:

It turns out that the sample variance is a calculation which utilizes all of the individual measurements, plus the average. Since the average value, or the mean, is not an independent variable, then we need to multiply our calculated variance by the following correction:

**variance =  $[N/(N-1)]s^2$** , and the standard deviation is the square root of this number. Thus, both the variance and the standard deviation are actually slightly large numbers than those represented by equations (6).

Obviously, in the limit of large numbers of measurements, this definition is the same as that given previously. In the limit of small numbers of measurements, neither the variance nor the standard deviation mean very much to begin with.

Above we utilized the 'generic' probability distribution  $P(x)$ , which we didn't explicitly define. Let's define it now. If we measure the length of a box several times with a meter stick, then chances are we will end up with some particular distribution of measurements that fits a Gaussian probability distribution. If we can then determine just what that characteristic Gaussian function looks like, then we can operate on it with the mean and variance operators, and determine the average values and their corresponding standard deviations. Thus, let's consider various probability distribution functions. Recall that above it was mentioned that probability distributions could often be classified as Binomial distribution functions, Gaussian distribution functions, and Poisson distribution functions. Others, such as Lorentzian distributions, Boltzman distributions, etc., are possible as well. By far the most commonly encountered function will be the Gaussian distribution function, and so let's consider this important function first.

### ***The Gaussian Probability Distribution***

The Gaussian probability function is defined as

$$P_G(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad (14)$$

A plot of the Gaussian is shown below:

There are several ways to quantify the Gaussian curve shown here. First, if a line is drawn such that it is tangent to the steepest part of the curve, it will intersect the curve at  $\pm \sigma$ . It will intersect the x-axis at  $\pm 2\sigma$ .

This width may also be quantified as the  $\exp(-1/2)$  value of the curve. i.e., when the curve is  $e^{-1/2}$  times its value at  $\mu$ , the x-value will be  $\mu \pm \sigma$ . This the same as the standard deviation, or, in other words:

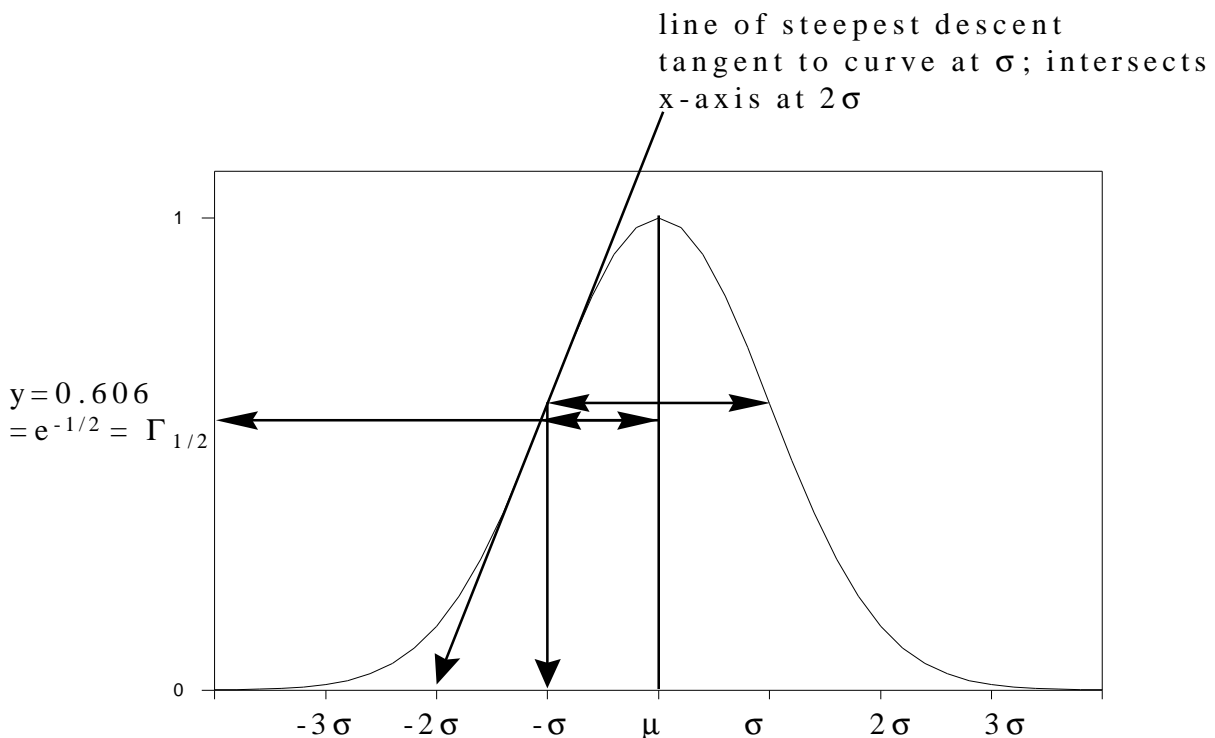
$$P_G(\mu \pm \sigma; \mu, \sigma) = e^{-1/2} P_G(\mu; \mu, \sigma) \quad (15)$$

A second way to specify the width of the curve is to use the full-width-at-half-maximum, of FWHM. This is commonly denoted by the symbol  $\Gamma$ , and may be defined by the following equality:

$$P_G(\mu \pm \frac{1}{2}\Gamma; \mu, \sigma) = \frac{1}{2} P_G(\mu; \mu, \sigma) \quad (16)$$

It turns out that  $\Gamma = 2.354\sigma$ .

The significance of these various ways of measuring the peak width are important. For example, how certain will we be that a value will fall between  $\mu \pm \sigma$ , or  $\mu \pm \Gamma$ ? In fact, this is not a difficult question to answer. Recall that the Gaussian probability distribution was normalized. Thus, integrating the  $P_G(x; \mu, \sigma)$  from  $-\sigma$  to  $+\sigma$ , and dividing by 1, gives the fraction of observations that should fall within a single standard deviation.



$$\int_{\mu-\sigma}^{\mu+\sigma} P_G(x; \mu, \sigma) dx = \text{fraction of observations expected to fall within } 1\sigma \text{ of } \mu.$$

It turns out that  $1\sigma$  is about 68% probability limit, and  $2\sigma$  is about 95% probability limit.  $3\sigma$  will be near 99%.

The other distributions are given by the following equations:

The Binomial distribution, or what are the chances for observing  $x$  successes out of  $n$  tries when the probability for success in each try is  $p$ :

$$P_B(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad (17)$$

$$\mu = np \quad \sigma^2 = np(1-p)$$

**The Poisson distribution.** This is similar to the binomial distribution, although  $n$  is very large and  $\mu$  is constant; i.e. it is appropriate for describing small samples from large populations.

$$P_p(x; \mu) = \frac{\mu^x}{x!} e^{-\mu} \quad \sigma^2 = \mu \quad (18)$$

And, finally, the Lorentzian distribution for describing a natural, homogeneously broadened distribution (used for spectroscopic line shapes)

$$P_L(x; \mu, \Gamma) = \frac{1}{\pi} \frac{\Gamma/2}{(x - \mu)^2 + (\Gamma/2)^2} \quad (19)$$

An example of a Poisson distribution would be die roles. I.e., if you have two dice, then the chances of getting the following numbers are

numbers	probability
2	1/36
3	2/36
4	3/36
5	4/36
6	5/36
7	6/36
8	5/36
9	4/36
10	3/36
11	2/36
12	1/36

The probability of observing  $x$  successes out of  $n$  tries when the probability for success in each try is  $p$ :

$$p_B(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad (20)$$

Thus, to observe a single 4 in five tries:  $p=1/9$ , or 0.111;  $n = 5$ , and  $x = 1$ , and the probability of occurrence is about 35%.

**Return to the Box** So, returning to our problem of determining the volume of a box, based on measurements of the width, height, and length. This is the type of measurement that fill fall under that description of a Gaussian distribution. If we take just a handful of measurements, then we are not going to be able to get a good idea of what the standard deviation is. There is a simple method, called the student t-distribution, which can help us here, and we will get to it soon. To define a true Gaussian, however, it takes many measurements - like 100 or so. With only a few measurements, the standard deviation is going to be fairly large, and the  $2\sigma$ , or 95% confidence limits, is going to seriously limit just what we can finally say about the box. How slowly does the standard deviation decrease as we take more data? Let's look at the t-distribution and try to approximate some answers, before we make a proof.

### ***The student t-distribution***

(Schoemaker, Garland and Nibler 6ed. p. 32-34)

If we just have a few measurements ( $N$  is between 3 and 10), then we can make a (poor) estimate of the standard deviation. First, calculate the range  $R$ , which is the largest minus the smallest observed value. Below is a table of the values  $K_2$ ,  $J$ , and  $N$ . Don't worry about  $J$  for the moment.  $N$  is the number of measurements, and  $K_2$  is a statistical factor.

$N$	$K_2$	$J(95\%)$
3	0.59	1.3
4	0.49	0.72
5	0.43	0.51
6	0.40	0.40
7	0.37	0.33
8	0.35	0.29
9	0.34	0.26
10	0.33	0.23

$\sigma = K_2R$ , and  $m \pm 2\sigma = 95\%$  confidence estimation. If more than a handful of values are collected, then there is probably going to be a better way to estimate the result. However, since the standard deviation will be dependent upon  $K_2$ , let's look at how  $K_2$  changes with number of samples. For 3 samples,  $K_2 = 0.59$ , and for 6 samples,  $K_2 = 0.40$ . Now 6 is twice as many samples as 3, and  $0.4 \sim 0.59/2^{1/2}$ . Thus, as one measures more and more samples, or as  $N$  gets larger, the certainty in  $\mu$  only appears to get better as  $N^{1/2}$ . Let's see if we can quantify this a little bit.

Recall from above that:

$$\sigma^2 = \frac{1}{N-1} \left[ \sum_{i=1}^N (x_i - \mu)^2 \right] \quad (21) \quad (\text{note the } N-1 \text{ in the divisor})$$

$$\text{thus, } \sigma = \frac{1}{\sqrt{N-1}} \left[ \sum_{i=1}^N (x_i - \mu)^2 \right]^{1/2} \quad (22)$$

So, in fact we already knew how the standard deviation behaved with N. It does, in fact, decrease at a rate proportional to  $N^{-1/2}$ . This is one of the cruel aspects of doing experiments. If you take data for an hour, then you will have to continue for 3 more hours in order to improve your 'signal to noise' by a factor of 2. In other words, when one takes lots of data, the statistical improvement in the data is very fast at first (i.e. 4 seconds of data collection is not much worse than 1 second), and agonizingly slow later.

### ***Rejection of bad(?) data***

Before we get into the next topic, which is linear regressions, let's first look at how we know whether to reject a data point. If you take just a few data points, and one of the data points looks pretty bad, then how can you tell whether or not to reject it? Consider the following table, taken from Schoemaker, Garland, and Nibler, 6 ed., p. 41.

N	3	4	5	6	7	8	9	10
Q	0.94	0.76	0.64	0.56	0.51	0.47	0.44	0.41

$$\text{Calculate } q = \frac{|(\text{suspect} \cdot \text{value}) - (\text{closest})|}{\text{range}} \quad (23)$$

If  $q$  is  $< Q$ , keep the data. If  $q > Q$ , toss it out. Note that the range (highest - lowest) may include the suspect value.

### ***The Method of Least Squares***

Let's say that we don't really know the mean of a particular set of experiments, and so we estimate it with some number we'll call  $\mu'$ , and, likewise, we'll suppose some standard deviation  $\sigma'$ . The probability of observing some value  $x_i$  is then giving by the following function:

$$P_i(\mu') = \frac{1}{\sigma' \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{x_i - \mu'}{\sigma'} \right)^2 \right] \quad (24)$$

Now, if we make a set of N observations, then the probability for observing that set of observations is given by the product of the individual probability functions  $P_i(\mu')$ s:

$$P(\mu') = \prod_{i=1}^N P_i(\mu') = \left( \frac{1}{\sigma\sqrt{2\pi}} \right)^N \exp \left[ -\frac{1}{2} \sum \left( \frac{x_i - \mu'}{\sigma} \right)^2 \right] \quad (25)$$

Now, what we would like is for the product probability function to be a maximum - i.e. that we have chosen a  $\mu'$  such that we are most likely to see our set of observations. Thus, we need to maximize (25). It turns out that maximizing (25) is the same as minimizing the argument in the exponential, which we will call X:

$$X = \left[ -\frac{1}{2} \sum \left( \frac{x_i - \mu'}{\sigma} \right)^2 \right] \quad (26)$$

$$\frac{dX}{d\mu'} = -\frac{1}{2} \sum \frac{d}{d\mu'} \left( \frac{x_i - \mu'}{\sigma} \right)^2 = \sum \left( \frac{x_i - \mu'}{\sigma} \right) = 0 \text{ (at the maximum)} \quad (27)$$

$$\text{which, because } \sigma \text{ is a constant, gives } \mu' = \bar{x} \equiv \frac{1}{N} \sum x_i \quad (28)$$

and, of course, we generate the answer that we should have. This is the method of Least Squares, and it is extremely powerful. We will explore this some more.

## END LECTURE TWO

What is the uncertainty  $\sigma$ ?

Assume all data points are from the same parent distribution, so all have the same uncertainty  $\sigma$ . Recall: Propagation of error:

$$\sigma_y^2 = \sigma_a^2 \left( \frac{\delta y}{\delta a} \right)^2 + \sigma_b^2 \left( \frac{\delta y}{\delta b} \right)^2 + \dots \quad (29)$$

where all the variance in each data point  $x_i$  is weighted by the square of the effect  $\frac{\delta \mu'}{\delta x_i}$  that the data has on the result. If all  $\sigma_i = \sigma$ , then substituting  $\mu$  for  $y$  in (29), and  $x_i$  as a variable:

$$\frac{\delta \mu'}{\delta x_i} = \frac{\delta}{\delta x_i} \left( \frac{1}{N} \sum x_i \right) = \frac{1}{N} \quad (30)$$

substituting in  $\mu'$  for the parenthetical term in (30), we get

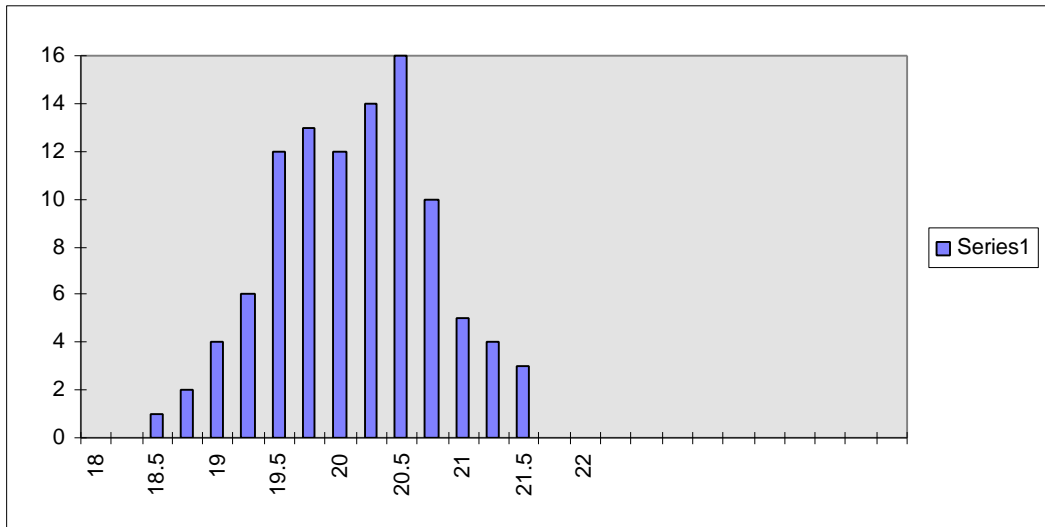
$$\sigma_u^2 = \sum \left[ \sigma_i^2 \left( \frac{1}{N} \right)^2 \right] = N \sigma^2 \left( \frac{1}{N} \right)^2 = \frac{\sigma^2}{N} \quad (31)$$

recalling our argument based on dependent variables

$$\sigma_u = \sqrt{\frac{1}{N-1} \sum (x_i - \bar{x})^2} \quad (32)$$

**Example 4.1: Let's return to the box:**

Assume  $L = \mu_L = 20.000$  cm (known value). The student, after about 100 measurements, now has a data set from which to determine  $L$ . By considering the box, the ruler, and the student's own near sightedness, the student determines that each measurement is good to about +/- 0.5 cm, and has the following set of measurements:



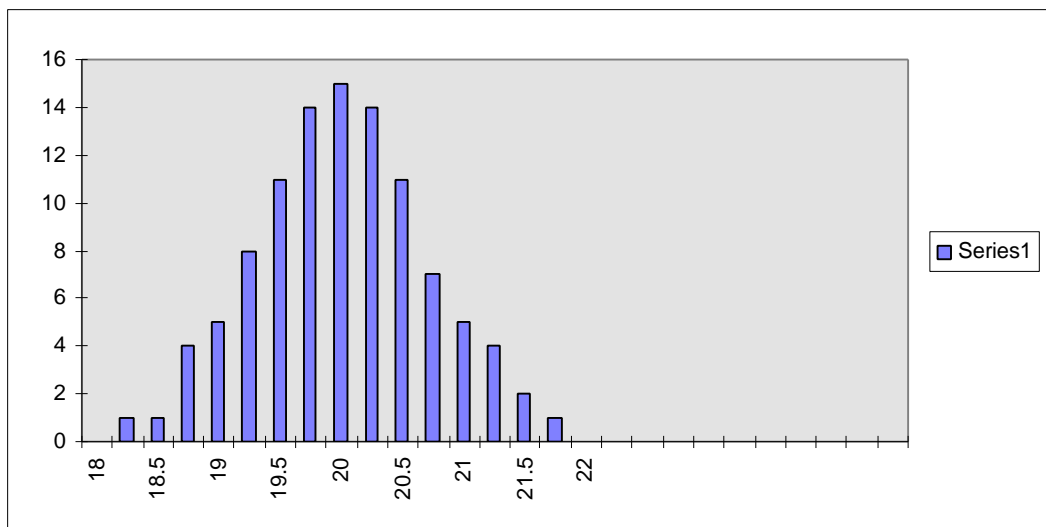
$\sigma_i = \sigma = 0.5$  cm

Assume, for the moment, that the student calculates from the data  $\mu = 19.942$  cm

if  $\sigma = 0.522$ , as calculated from from  $\sigma_u = \sqrt{\frac{1}{N-1} \sum (x_i - \bar{x})^2}$

then  $\sigma_u = \frac{\sigma}{\sqrt{N}} = \frac{0.522}{\sqrt{100}} = 0.0522$ , so the student reports 19.94 +/- 0.05 cm, to within 1  $\sigma$  confidence level.

Now, with many more measurements, say  $10^4$ , the student has generated the following set of data.



Obviously, the data is much cleaner, and the uncertainty of the new data can be readily

calculated to be  $\sqrt{\frac{10^4}{10^2}} = 10$  times better than it was previously, so that now  $\sigma_u = 0.005$  cm. At

this point, other sources of error are probably important. Absolute and relative calibrations may not be good – for example, the meter stick may have a certain amount of absolute error. Perhaps, during the course of the 48 hour period in which the student took the huge number of measurement, the temperature in the room rose and fell, thus causing the box to slightly expand and contract. All sorts of things can happen during this time period.

As a matter of course, if you want to get a better measurement, taking more data points using the same experimental approach is often not the best way to do things. Imagine that we are trying to count photon events, and we take a video camera to record the events. The camera may not be real sensitive, and so we may only be able to see the photon events if several photons arrive simultaneously. We can sit for a long time, and slowly build up statistics. Or, perhaps a better way would be to replace the tv camera with a much more sensitive detector. In this way, we can build up the equally good statistics in a much shorter time. The statistical uncertainty in our data is going to depend linearly on how good the detector is, but only on the square root of the amount of time that we sample. This is a subtle, but important point. If you can imagine a way to make your signal-to-noise improve in a way that is faster than time-averaging, then do it. A better experimental approach is often the way to go.

### ***Nonuniform uncertainties***

Suppose that the student takes the advice from the above paragraph, and, after measuring the box many times, decides to do the measurement differently. This second set of measurement is likely to be characterized by a different uncertainty than the first set. How can the student account for this? Obviously, the student would like to use all of the data that has been measured.

In this case, each data point is weighted by its own uncertainty:

$$P(\mu') = \prod_{i=1}^n \left( \frac{1}{\sigma_i \sqrt{2\pi}} \right) \exp \left[ -\frac{1}{2} \sum \left( \frac{x_i - \mu'}{\sigma_i} \right)^2 \right] \quad (33)$$

once again, we follow the method of least-squares. Minimizing the exponential:

$$-\frac{1}{2} \frac{d}{d\mu'} \sum \left( \frac{x_i - \mu'}{\sigma_i} \right)^2 = \sum \left( \frac{x_i - \mu'}{\sigma_i^2} \right) = 0 \quad (34)$$

solving for  $\mu'$ , we get

$$\mu' = \frac{\sum x_i / \sigma_i^2}{\sum 1 / \sigma_i^2} \quad (35)$$

where each data point is weighted inversely by its own variance

The error in the weighted mean may be calculated by the propagation of error equation. Taking (35), and calculating the change in the mean with respect to the change in individual measurements  $x_i$ , we get:

$$\frac{\delta\mu'}{\delta x_i} = \frac{\delta}{\delta x_i} \frac{\sum \left( \frac{x_i}{\sigma_i^2} \right)}{\sum \left( \frac{1}{\sigma_i^2} \right)} = \frac{1 / \sigma_i^2}{\sum 1 / \sigma_i^2} \quad (36)$$

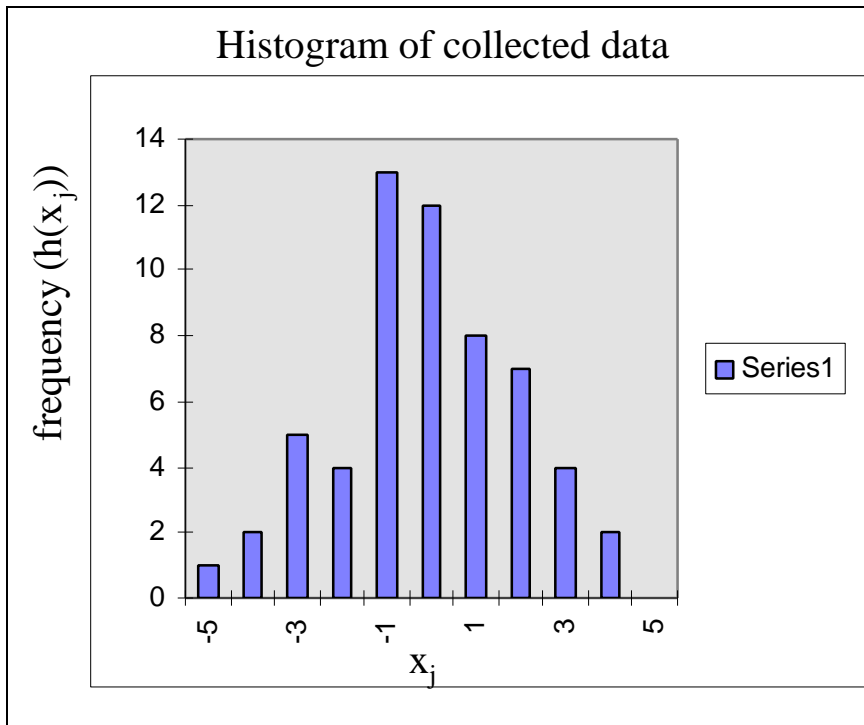
This result, substituted into  $\sigma_{\mu}^2 = \sum \left[ \sigma_i^2 \left( \frac{\delta\mu'}{\delta x_i} \right)^2 \right]$  gives us the general formula for the uncertainty of the mean  $\sigma$ :

$$\sigma_{\mu}^2 = \sum \frac{1 / \sigma_i^2}{\left[ \sum 1 / \sigma_i^2 \right]^2} = \frac{1}{\sum \left( 1 / \sigma_i^2 \right)} \quad (38)$$

**The  $\chi^2$  test of a distribution:**

Assume that we have measured a set of experimental data, and we have obtained both the mean ( $\mu$ ) and the standard deviation ( $\sigma$ ) that describe the data. If we can somehow build up a confidence about the parent population, then we can describe the parent distribution in detail, and predict the outcome of future experiments with some statistical certainty. Thus, we would like to know about the parent population.  $\chi^2$  is a statistical quantity that tells us if we are considering the parent distribution properly.

The experimental data that we have measured consists of a number of measurements, with values  $x_i$ , of some quantity  $x$ . If we have done a number of such measurements, then we can make a histogram, plotting the frequency of observations of  $x_i$ , vs.  $x_i$ :



For this particular set of observations, we will label the frequency of observations of some value  $x_j$  as  $h(x_j)$ .

We will refer to the parent probability distribution as  $P(x_j)$ . Then, the number of expected observations of some value  $x_j$  may be calculated from  $P(x_j)$ , and we call it  $y(x_j)$ :

$$y(x_j) = N P(x_j).$$

Where  $N$  is the total number of measurements taken.

Now, we come to a new concept. For each  $x_j$ , there is a  $\sigma_j(h)$  associated with the uncertainty in  $h(x_j)$ .

$$\text{Note } \sigma_j(h) \neq \sqrt{(\mu - x_j)}$$

Let's explore this new uncertainty a little. If we measured  $P(x_j)$  by doing 10 sets of 100 measurements, then we would get 10 measurements of  $h_k(x_j)$ .

$h_k \equiv k^{\text{th}}$  measurement out of 10 total.

$x_j = j^{\text{th}}$  value of  $x$ .

The expected value of  $h_k(x_j)$  is  $y(x_j) = NP(x_j)$ . This means that if we measure some value  $x_j$  a certain fraction of the time (out of 100 measurements), then we would expect that fraction to be equal to the frequency of  $x_j$  that corresponds to what is predicted by the parent probability distribution. Obviously we will not always get that frequency. As the matter of fact, we are only likely to get  $h_k(x_j) = NP(x_j)$  a relatively small fraction of the time, just like we are only likely to measure  $x_j = \mu$  a small fraction of the time. Thus, each individual histogram bar, or measurement of  $h_k(x_j)$ , has associated with it a mean and a standard deviation.

Since we have already binned our data into a histogram, and since this means that there are only certain possibilities for  $x_j$  (i.e.  $x_j$  is discrete, not continuous), then the statistics that describe  $h_k(x_j)$  are going to be Poisson statistics, even though  $P(x_j)$  may well be a Gaussian distribution function.

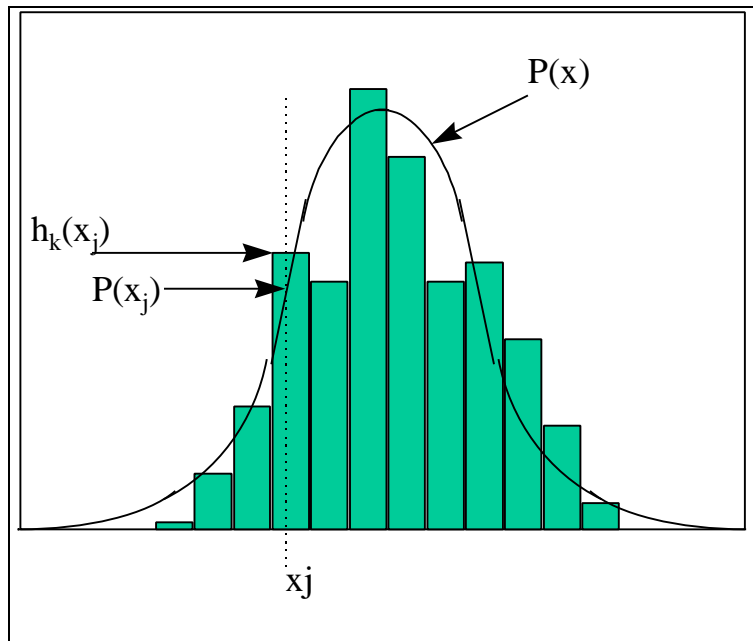
Thus, if we call the mean of the 10 measurements of  $h_k(x_j) = \mu_j$ , and the standard deviation is (according to Poisson statistics) given by:

$$\sigma_j(h) = \sqrt{\mu_j} \quad (39)$$

Then, with those definitions, we can calculate  $\chi^2$ :

$$\chi^2 = \sum_{j=1}^n \frac{[h(x_j) - NP(x_j)]^2}{\sigma_j(h)^2} \quad (41)$$

This definition of  $\chi^2$  implies that  $\chi^2$  is a statistic that characterized the dispersion of the observed frequencies from the expected frequencies. The numerator is a measurement of the spread of the observations, while the denominator is a measurement of the expected spread. Thus, we might expect that in the case of good agreement, the actual spread over the



expected spread should be about equal to 1, and that the optimum value of  $\chi^2$  would be  $n$ , the number of bins in our previous plots. This is almost true.

If each measurement were to reproduce the predicted probability distribution exactly, then  $\chi^2$  would equal 0. However, we recognize from our probability discussions that this is not likely to be the case. Instead, the expectation value for  $\chi^2$  is:

$$\langle \chi^2 \rangle = v = n - n_c \quad (42)$$

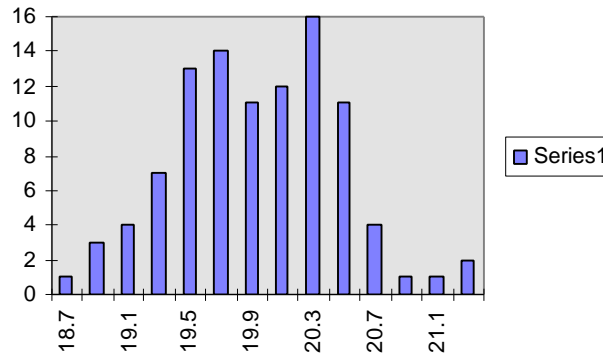
In (42),  $v$  is the number of degrees of freedom, and is equal to the number ( $n$ ) of sample frequencies (in the graph on the previous page,  $n$  is equal to 12) minus  $n_c$ , which is the number of constraints. A constraint is a parameter that has been calculated from the data to describe the probability function,  $NP(x_i)$ . Even if  $P(x_i)$  is chosen completely independent of the sample distribution, it is still normalized to the total number of events in the distribution, so that the expectation value of  $\chi^2$  must, at best, be  $\langle \chi^2 \rangle = n - 1$ . Usually  $\chi^2$  is given as the reduced chi-

square, which is  $\chi^2_v \equiv \chi^2 / v$ , which has an expectation value of  $\langle \chi^2_v \rangle = 1$ . Values that are much larger than 1 result from large deviations from the assumed distribution, and possibility indicate an incorrect choice of the probability distribution. If the values are much smaller than 1 are also indicate something is wrong in the nature of the experiment.

**Problem:** Assume the following data/histogram. The first column of number corresponds to length measurements, the second column corresponds to frequency. If the parent distribution is Gaussian with  $m = 20.00$  and  $s = 0.5$ , then what is  $m$  sample and  $s$  sample? What is  $\chi^2$ ?

Plot the histogram with a curve of the parent distribution  $NP(x)$ .

18.7	1
18.9	3
19.1	4
19.3	7
19.5	13
19.7	14
19.9	11
20.1	12
20.3	16
20.5	11
20.7	4
20.9	1
21.1	1
21.3	2



### Least Squares fit of a Straight Line

One of the most important and commonly used statistical tools is linear regression of a straight line. Fortunately, the technique for doing this is quite general - meaning that it is possible to take a set of data, and fit it to any particular functional form - not just a straight line.

Polynomial fits, exponential fits, fits to Gaussian distributions, etc. are all possible with the technique of Least squares. We have already covered this a little bit, so you should be at least a little familiar with the technique. In this section, we use the technique for fitting to a straight line, and we also try to point out where the technique is general for any functional form.

Assume that you have measured a set of data points  $(x_i, y_i)$ . Define  $a_0$  and  $b_0$  such that:  $y_o(x) = b_o(x) + a_o$ . Each  $y_i$  is drawn from a Gaussian parent distribution, with  $\mu = y_o(x_i)$ , and  $\sigma = \sigma_i$ . The probability of observing some particular value  $x_i$  is given by:

$$P_i = \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{y_i - y_o(x_i)}{\sigma_i} \right)^2 \right] \quad (43)$$

and the probability of making a given set of observations is the product of the individual probability functions for each individual observation:

$$P(a_o, b_o) = \prod P_i = \prod \left\{ \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left[ -\frac{1}{2} \left( \frac{y_i - y_o(x_i)}{\sigma_i} \right)^2 \right] \right\} \quad (44)$$

To maximize this probability product, we need to minimize the term in the exponential (as we did before). If you recall from our previous definition of  $\chi^2$ , this term in the exponential bears a strong resemblance to  $\chi^2$ , and so we label it that:

$$\chi^2 = \sum \left[ \frac{y_i - y(x_i)}{\sigma_i} \right]^2 = \sum \left[ \frac{1}{\sigma_i} (y_i - a - bx_i) \right]^2 \quad (45)$$

Note that we replaced  $y(x_i)$  with the function  $bx_i + a$ . It is at this point that we could have replaced  $y(x_i)$  with just about any function to which we wanted to fit our data. In our case, we are choosing a function of two variables ( $a$  and  $b$ ) that is linear in  $x$ . For example, a nonlinear function of  $x$  that had three variables might be  $y = ax + bx^2 + c$ . Now we minimize  $\chi^2$  with respect to all of the variables in our chosen function. In our case, this is just  $a$  and  $b$ .

$$\frac{\delta \chi^2}{\delta a} = \frac{\delta}{\delta a} \sum \left[ \frac{1}{\sigma_i^2} (y_i - a - bx_i)^2 \right] = -2 \sum \frac{1}{\sigma_i^2} (y_i - a - bx_i) = 0 \quad (46)$$

and

$$\frac{\delta \chi^2}{\delta b} = \frac{\delta}{\delta b} \sum \left[ \frac{1}{\sigma_i^2} (y_i - a - bx_i)^2 \right] = -2 \sum \frac{x_i}{\sigma_i^2} (y_i - a - bx_i) = 0 \quad (47)$$

We can rearrange (46) to give:

$$\sum \frac{y_i}{\sigma_i^2} = a \sum \frac{1}{\sigma_i^2} + b \sum \frac{x_i}{\sigma_i^2} \quad (48)$$

and we can rearrange (47) to give:

$$\sum \frac{y_i \cdot x_i}{\sigma_i^2} = a \sum \frac{x_i}{\sigma_i^2} + b \sum \frac{x_i^2}{\sigma_i^2} \quad (49)$$

for the special case in which all  $\sigma_i = \sigma$ , i.e. all standard deviations of the individual measurements are the same, then we can simplify (48) to give:

$$\sum \frac{y_i}{\sigma_i^2} = \frac{aN}{\sigma^2} + b \sum \frac{x_i}{\sigma_i^2} = \sum y_i = aN + b \sum x_i \quad (50)$$

and we can rearrange 49 to give:

$$\sum x_i y_i = a \sum x_i + b \sum x_i^2 \quad (51)$$

If we multiply the right hand equality of (50) by  $\sum x_i$ , and we multiply (51) by N, then we get the following two equations:

$$\sum y_i \cdot \sum x_i = aN \cdot \sum x_i + b \sum x_i \cdot \sum x_i \quad (52)$$

$$N \sum x_i y_i = aN \sum x_i + bN \sum x_i^2 \quad (53)$$

Now we can subtract (53) from (52), and get a single equation that is only dependent upon the unknown variable b, which is the slope of the straight line. Similarly, we can also solve for the intercept. The equations for a and for b are:

$$a = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{-\sum x_i \sum x_i + N \sum x_i^2} \quad (54)$$

and

$$b = \frac{N \sum x_i y_i - \sum x_i \sum y_i}{-\sum x_i \sum x_i + N \sum x_i^2} \quad (55)$$

In general, this is a problem that is best solved by setting up determinants. For the generic case in which all uncertainties are not equal, and, starting with equations (48) and (49), the determinants are:

$$a = \frac{1}{\Delta} \begin{vmatrix} \sum \frac{y_i}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} \\ \sum \frac{x_i y_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \end{vmatrix} = \frac{1}{\Delta} \left( \sum \frac{x_i^2}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2} \right) \quad (56)$$

$$b = \frac{1}{\Delta} \begin{vmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{y_i}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i y_i}{\sigma_i^2} \end{vmatrix} = \frac{1}{\Delta} \left( \sum \frac{1}{\sigma_i^2} \sum \frac{x_i y_i}{\sigma_i^2} - \sum \frac{x_i}{\sigma_i^2} \sum \frac{y_i}{\sigma_i^2} \right) \quad (57)$$

where

$$\Delta = \begin{vmatrix} \sum \frac{1}{\sigma_i^2} & \sum \frac{x_i}{\sigma_i^2} \\ \sum \frac{x_i}{\sigma_i^2} & \sum \frac{x_i^2}{\sigma_i^2} \end{vmatrix} = \left( \sum \frac{1}{\sigma_i^2} \sum \frac{x_i^2}{\sigma_i^2} - \left( \sum \frac{x_i}{\sigma_i^2} \right)^2 \right) \quad (58)$$

If the equation that we were trying to fit the data to was the second order polynomial function  $y(x) = ax^2 + bx + c$ , then we would have ended up with 3X3 determinants, one each for a, b, and c. Thus, as one goes to more and more complicated equations, the problem gets a little messier. Fortunately, most statistical analysis programs use a single generic routine that sends up N X N determinants for a problem of N variables, and solves for the answers algebraically.

**Problems:**

1. Derive a formula for making a linear fit to data with an intercept at the origin so that  $y = bx$ . Apply your method to fit a straight line through the origin to the following coordinate pairs. Assume uniform uncertainties  $\sigma = 1.5$  for the  $y_i$ 's. Find  $\chi^2$  for the fit, and the uncertainty in b.

$x_i$	2	4	6	8	10	12	14	16	18	20	22	24
$y_i$	5.3	14.4	20.7	30.1	35.0	41.3	52.7	55.7	63.0	72.1	80.5	87.9

2. Find by numerical integration the probability of observing a value from the Gaussian distribution that is:

- more than 1 standard deviation from the mean
- more than 2 standard deviations from the mean
- more than 3 standard deviations from the mean

3. After measuring the speed of sound several times, a student conclude that the standard deviation of his measurements is  $\sigma = 12$  m/s. Assume that the uncertainties are random, and that the experiment is not limited by systematic effects and determine how many measurements would be required to give a final uncertainty in the mean of +/- 2.0 m/s.

4. Find the uncertainty  $\sigma_x$  in x as a function of the uncertainties  $\sigma_u$  and  $\sigma_v$  in u and v for the following functions:

- a.  $x = \frac{1}{2}(u+v)$
- b.  $x = uv^2$
- c.  $x = u^{-2}$
- d.  $x = uv^2$
- e.  $x = u^2 + v^2$

5. If the diameter of a round table is determined to within 1%, how well is its area known? Would it be better to determine its radius to within 1%?

6. Snell's law relates the angle of refraction  $\theta_2$  of a light ray travelling in a medium of index of refraction  $n_2$  to the angle of incidence  $\theta_1$  of a ray travelling in a medium of index  $n_1$  through the equation  $n_2 \sin \theta_2 = n_1 \sin \theta_1$ . Find  $n_2$  and its uncertainty from the following measurements:

$$\theta_1 = (22.03 \pm 0.2)^\circ \quad \theta_2 = (14.45 \pm 0.2)^\circ \quad n_1 = 1.000$$

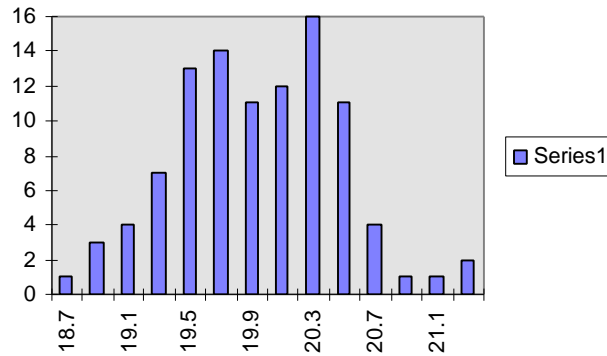
Assume that there is no uncertainty in  $n_1$ .

Problem # 1.

**Problem:** Assume the following data/histogram. The first column of number corresponds to length measurements, the second column corresponds to frequency. If the parent distribution is Gaussian with  $\mu = 20.00$  and  $\sigma = 0.5$ , then what is  $\mu$  sample and  $\sigma$  sample. What is  $\chi^2$ ?

Plot the histogram with a curve of the parent distribution NP(x).

18.7	1
18.9	3
19.1	4
19.3	7
19.5	13
19.7	14
19.9	11
20.1	12
20.3	16
20.5	11
20.7	4
20.9	1
21.1	1
21.3	2



The first thing to do here is to calculate  $\mu$  and  $\sigma$  of the experimental data.

**$m = 19.94$ ;  $s^2 = 0.279$ ;  $s = 0.528$**

From this, it is now possible to calculate an expected distribution, and compare that with the parent.

Taking the values of  $\mu(\text{parent}) = 20.00$  and  $\sigma(\text{parent}) = 0.5$ , we can calculate  $P(x_j)$  for  $x_j = 18.7, 18.9, 19.1, \text{etc.}$  Simply plug  $\mu(\text{parent})$  and  $\sigma(\text{parent})$  into the formula for a Gaussian distribution. Sum up the Gaussian probabilities for the individual histogrammed cells, and you

will get a sum of about 4.97 (i.e.  $\sum_{x_j=18.7}^{x_j=21.3} P(x_j) = 4.97$ ). Then, multiply  $P(x_j)$  by  $(100/4.97)$  to

normalize to the 100 measurements that are listed in the histogram, and this yields  $NP(x_j)$ .

Now, do the same thing for the experimental probability distribution - i.e. take  $\mu = 19.94$  and  $\sigma = 0.528$  and plug those numbers into a Gaussian probability distribution and calculate  $\text{experim.}(x_j)$ . Now, at each  $x_j$ , calculate a  $\mu_j$ , which is simply the square root of  $\text{experim.}(x_j; \sigma = 0.528, \mu = 19.94)$  evaluated at 18.7, 18.9, etc. According to Poisson statistics,  $\sigma_j$  is then  $\mu_j^{1/2}$

Now you have a table of values  $x_j, h(x_j) - NP(x_j)$  and  $\mu_j$ . Calculate

$$\chi^2 = \frac{[h(x_j) - NP(x_j)]^2}{\sigma_j^2} = 12.23.$$

From 12.23, we can calculate  $\chi^2_v$  by dividing  $\chi^2$  by  $(N-1)$ . Since we have taken 14 measurements,  $N-1$  is 13, and  **$\chi^2_v = 0.94$** .

Problem 2.

Derive a formula for making a linear fit to data with an intercept at the origin so that  $y = bx$ . Apply your method to fit a straight line through the origin to the following coordinate pairs. Assume uniform uncertainties  $\sigma = 1.5$  for the  $y_i$ 's. Find  $\chi^2$  for the fit, and the uncertainty in  $b$ .

$x_i$	2	4	6	8	10	12	14	16	18	20	22	24
$y_i$	5.3	14.4	20.7	30.1	35.0	41.3	52.7	55.7	63.0	72.1	80.5	87.9

Here, we just go to back to the least squares of a straight line discussion. Recall equation (45):

$$\chi^2 = \sum \left[ \frac{y_i - y(x_i)}{\sigma_i} \right]^2 = \sum \left[ \frac{1}{\sigma_i} (y_i - a - bx_i) \right]^2$$

we use the same equation, except that we set  $a = 0$ :

$$\chi^2 = \sum \left[ \frac{y_i - y(x_i)}{\sigma_i} \right]^2 = \sum \left[ \frac{1}{\sigma_i} (y_i - bx_i) \right]^2$$

and we minimize with respect to  $b$ :

$$\frac{\delta \chi^2}{\delta b} = \frac{\delta}{\delta b} \sum \left[ \frac{1}{\sigma_i^2} (y_i - bx_i)^2 \right] = -2 \sum \frac{x_i}{\sigma_i^2} (y_i - bx_i) = 0$$

setting all uncertainties to 1.5, we can then solve for  $b$ .

$$\sum x_i y_i = b \sum x_i^2$$

when we do this, we solve for  $b$  and get

x	y	$x_i y_i$	$x_i^2$	$b x_i$
2	5.3	10.6	4	7.2
4	14.4	57.6	16	14.4
6	20.7	124.2	36	21.6
8	30.1	240.8	64	28.8
10	35	350	100	36
12	41.3	495.6	144	43.2
14	52.7	737.8	196	50.4
16	55.7	891.2	256	57.6
18	63	1134	324	64.8
20	72.1	1442	400	72
22	80.5	1771	484	79.2
24	87.9	2109.6	576	
	sums	9364.4	2600	
	$b =$	3.601692		

2. Integrate the equations.

3. To take  $\sigma$  from 12 m/s to 2 m/s, then one has to do  $(12/2)^2$ , or a factor of 36 times more experiments.

4. Easy partial derivatives

5. It is better to determine the diameter to within 1% by a factor of 2

## EMF Lab (Baby Lab)

**Abstract** Various thermodynamic values of the reaction  $\text{HgO} + \text{Zn} \rightarrow \text{Hg} + \text{Zn(OH)}_2$  were measured by doing experiments on a button battery based on this chemical reaction. A voltage of  $1.493 \pm 0.005$  V was measured at 297.2K. The Gibbs free energy ( $\Delta G$ ) was measured to be  $-292.1 \pm 1.0$  KJ/mol, the entropy change of the reaction ( $\Delta S$ ) was measured to be  $-40 \pm 2.1$  JK<sup>-1</sup>mole<sup>-1</sup>. These two numbers yield a change in latent heat of  $\Delta H = \Delta G + T\Delta S = 304 \pm 1.0$  kJ/mol. The internal resistance of the battery was determined to be  $11.2 \pm 1.2$   $\Omega$ .

### Quiz:

A student is trying to determine the time dependence of a radioactive decay signal. According to various fundamental physical laws, the radioactivity should decay exponentially with time.

The student measures the following numbers of counts vs. time:

time	counts
------	--------