Weighted Averages

$$A: \quad x = x_A \pm \sigma_A$$
$$B: \quad x = x_B \pm \sigma_B$$

combining separate measurements: what is the best estimate for x ?

assume that measurements are governed by Gauss distribution with true value X

$$\mathrm{Prob}_X(x_A) \propto \frac{1}{\sigma_A} e^{-(x_A - X)^2/2\sigma_A^2}$$

probability that A finds x_A

$$\mathrm{Prob}_X(x_B) \propto \frac{1}{\sigma_B} e^{-(x_B - X)^2/2\sigma_B^2}$$

$$\mathrm{Prob}_X(x_A x_B) = \mathrm{Prob}_X(x_A) \cdot \mathrm{Prob}_X(x_B)$$

probability that A finds x_A and B finds x_B

$$\propto \frac{1}{\sigma_A \sigma_B} e^{-\chi^2/2}$$

find maximum of probability

principle of maximum likelihood

the best estimate for X is that value for which $Prob_X(x_A, x_B)$ is maximum

$$\chi^2 = \left(\frac{x_A - X}{\sigma_A} \right)^2 + \left(\frac{x_B - X}{\sigma_B} \right)^2$$

chi squared – "sum of squares"

find minimum of χ^2

method of least squares

$$\frac{d\chi^2}{dX} = 0 \quad \Rightarrow \quad -2\frac{x_A - X}{\sigma_A{}^2} - 2\frac{x_B - X}{\sigma_B{}^2} = 0$$

$$(\text{best estimate for } X) = \left(\frac{x_A}{\sigma_A^2} + \frac{x_B}{\sigma_B^2} \right) \Big/ \left(\frac{1}{\sigma_A^2} + \frac{1}{\sigma_B^2} \right)$$

$$= \frac{w_A x_A + w_B x_B}{w_A + w_B} = x_{wav}$$

weighted average

$$w_A = \frac{1}{\sigma_A^2} \qquad w_B = \frac{1}{\sigma_B^2}$$

weights

Weighted Averages

$x_1, x_2, \ldots, x_N$ - measurements of a single quantity x with uncertainties $\sigma_1, \sigma_2, \ldots, \sigma_N$

$$x_1 \pm \sigma_1,\ x_2 \pm \sigma_2,\ \ldots,\ x_N \pm \sigma_N$$

$$x_{wav} = \frac{\sum w_i x_i}{\sum w_i}$$ ← weighted average

$$w_i = \frac{1}{\sigma_i^2}$$ ← weights

$$\sigma_{wav} = \frac{1}{\sqrt{\sum w_i}}$$ ← uncertainty in x_{wav}

can be calculated
using error propagation

Example Problem

Two students measure the radius of a planet and get final answers R_A=25,000±3,000 km and R_B=19,000±2,500 km.
The best estimate of the true radius of a planet is the weighted average. Find the best estimate of the true radius of a planet and the error in that estimate.

$$x_{wav} = \frac{w_A x_A + w_B x_B}{w_A + w_B} \qquad w_A = \frac{1}{\sigma_A^2} \qquad w_B = \frac{1}{\sigma_B^2} \qquad \sigma_{wav} = \frac{1}{\sqrt{w_A + w_B}}$$

$$R_{wav} = \frac{\frac{R_A}{\sigma_A{}^2} + \frac{R_B}{\sigma_B{}^2}}{\frac{1}{\sigma_A{}^2} + \frac{1}{\sigma_B{}^2}} = \frac{\frac{25{,}000}{3{,}000^2} + \frac{19{,}000}{2{,}500^2}}{\frac{1}{3{,}000^2} + \frac{1}{2{,}500^2}} = 21{,}459km \rightarrow \underline{21{,}500km}$$

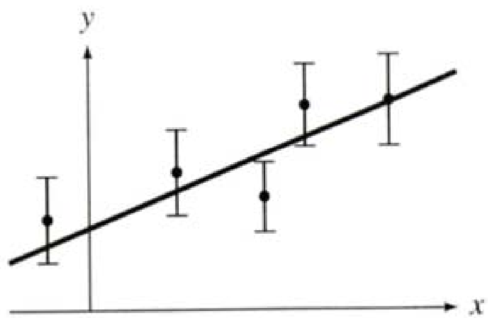
$$\sigma_{wav} = \frac{1}{\sqrt{\frac{1}{\sigma_A{}^2} + \frac{1}{\sigma_B{}^2}}} = \frac{1}{\sqrt{\frac{1}{3{,}000^2} + \frac{1}{2{,}500^2}}} = 1{,}921km \rightarrow \underline{1{,}900km}$$

$$\underline{R_{wav} = 21{,}500 \pm 1{,}900km}$$

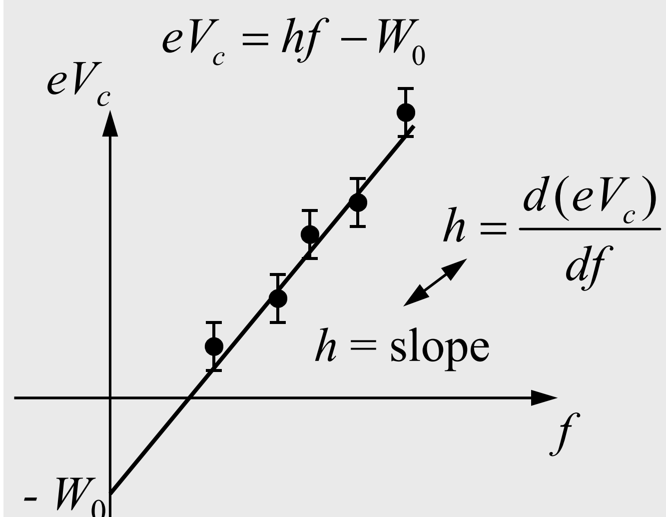
Least-Squares Fitting

consider two variables x and y that are connected by a linear relation

$$y = A + Bx$$



graphical method of finding the best straight line to fit a series of experimental points



The Photoelectric Effect

$$\begin{matrix} x_1, x_2, \ldots, x_N \\ y_1, y_2, \ldots, y_N \end{matrix} \longrightarrow \text{find } A \text{ and } B$$

analytical method of finding the best straight line to fit a series of experimental points is called **<u>linear regression</u>** or **<u>the least-squares fit for a line</u>**

Calculation of the Constants *A* and *B*

$$(\text{true value for } y_i) = A + Bx_i$$

$$\text{Prob}_{A,B}(y_1) \propto \frac{1}{\sigma_y} e^{-(y_1 - A - Bx_1)^2 / 2\sigma_y^2}$$ ← probability of obtaining the observed value of y_1

$$\text{Prob}_{A,B}(y_1,\ldots,y_N) = \text{Prob}_{A,B}(y_1)\cdots\text{Prob}_{A,B}(y_N)$$ ← probability of obtaining the set $y_1, \ldots, y_N$

$$\propto \frac{1}{\sigma_y^N} e^{-\chi^2/2}$$ ← find maximum of probability

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(y_i - A - Bx_i\right)^2}{\sigma_y^2}$$ ← chi squared – “sum of squares”
find minimum of χ^2
<u>least squares fitting</u>

$$\frac{\partial \chi^2}{\partial A} = \frac{-2}{\sigma_y^2} \sum_{i=1}^{N} \left(y_i - A - Bx_i\right) = 0$$

$$\frac{\partial \chi^2}{\partial B} = \frac{-2}{\sigma_y^2} \sum_{i=1}^{N} x_i \left(y_i - A - Bx_i\right) = 0$$

$$\sum y_i - AN - B\sum x_i = 0$$

$$\sum x_i y_i - A\sum x_i - B\sum x_i^2 = 0$$

→

$$\boxed{\begin{aligned} A &= \frac{\sum x^2 \sum y - \sum x \sum xy}{\Delta} \\ B &= \frac{N\sum xy - \sum x \sum y}{\Delta} \\ \Delta &= N\sum x^2 - \left(\sum x\right)^2 \end{aligned}}$$

Uncertainties in *y*, *A*, and *B*

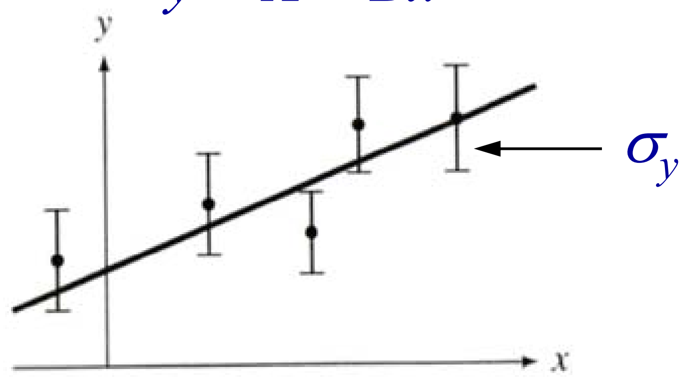
$$\boxed{\begin{aligned}\sigma_y &= \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - A - Bx_i\right)^2} \\ \sigma_A &= \sigma_y\sqrt{\frac{\sum x^2}{\Delta}} \\ \sigma_B &= \sigma_y\sqrt{\frac{N}{\Delta}}\end{aligned}}$$

uncertainty in the measurement of y

uncertainties in the constants A and B

given by error propagation in terms of uncertainties in $y_1, \ldots, y_N$

$y = A + Bx$



Example of Calculation of the Constants *A* and *B*

$T = A + BP$ (y = T, x = P)

if volume of an ideal gas is kept constant,
its temperature is a linear function of its pressure

i	P_i	T_i
1	65	-20
2	75	17
3	85	42
4	95	94
5	105	127

absolute zero of temperature $A = ?$

$$A = \frac{\sum x^2 \sum y - \sum x \sum xy}{\Delta}$$

$$B = \frac{N\sum xy - \sum x \sum y}{\Delta}$$

$$\Delta = N\sum x^2 - \left(\sum x\right)^2$$

$\Sigma P = 425$

$\Sigma P^2 = 37{,}125$

$\Sigma T = 260$

$\Sigma PT = 25{,}810$

$\Delta = N\Sigma P^2 - (\Sigma P)^2 = 5{,}000$

$$A = \frac{\Sigma P^2 \Sigma T - \Sigma P \Sigma PT}{\Delta} = -263.35$$

$$B = \frac{N\Sigma PT - \Sigma P \Sigma T}{\Delta} = 3.71$$

$$\sigma_y = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}\left(y_i - A - Bx_i\right)^2}$$

$$\sigma_A = \sigma_y\sqrt{\frac{\sum x^2}{\Delta}}$$

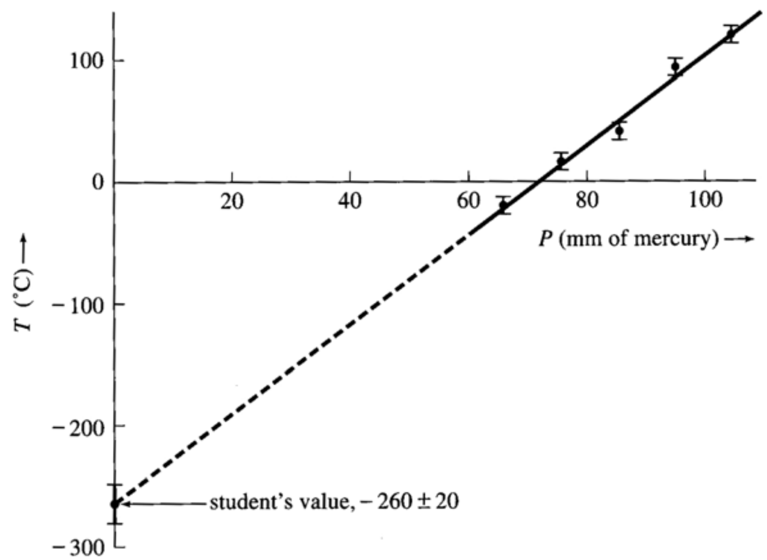
$$\sigma_T = \sqrt{\frac{1}{N-2}\Sigma(T_i - A - BP_i)^2} = 6.7$$

$$\sigma_A = \sigma_T\sqrt{\frac{\Sigma P^2}{\Delta}} = 18$$

$A = -263.35 \pm 18\,^\circ C$

$A = -263 \pm 18^0\ \mathrm{C}$

absolute zero of temperature = – 273.15^0 C



Covariance

$$x = \overline{x} \pm \delta x$$
$$y = \overline{y} \pm \delta y \longrightarrow q(x, y) = \overline{q} \pm \delta q$$

find $\overline{q}$ and δq

N pairs of data $(x_1, y_1), \dots, (x_N, y_N)$

$x_1, \dots, x_N \longrightarrow \overline{x}$ and σ_x

$y_1, \dots, y_N \longrightarrow \overline{y}$ and σ_y

$q_i = q(x_i, y_i)$

$q_1, \dots, q_N \longrightarrow \overline{q}$ and σ_q

$$q_i \approx q(\overline{x}, \overline{y}) + \frac{\partial q}{\partial x}(x_i - \overline{x}) + \frac{\partial q}{\partial y}(y_i - \overline{y})$$

$$\overline{q} = \frac{1}{N}\sum_{i=1}^{N} q_i$$
$$= \frac{1}{N}\sum_{i=1}^{N}\left[q(\overline{x}, \overline{y}) + \frac{\partial q}{\partial x}(x_i - \overline{x}) + \frac{\partial q}{\partial y}(y_i - \overline{y})\right]$$

$$\sum (x_i - \overline{x}) = 0 \Rightarrow \underline{\overline{q} = q(\overline{x}, \overline{y})}$$

$$\sigma_q^2 = \frac{1}{N}\sum (q_i - \overline{q})^2$$
$$= \frac{1}{N}\sum\left[\frac{\partial q}{\partial x}(x_i - \overline{x}) + \frac{\partial q}{\partial y}(y_i - \overline{y})\right]^2$$
$$= \left(\frac{\partial q}{\partial x}\right)^2 \cdot \frac{1}{N}\sum (x_i - \overline{x})^2 + \left(\frac{\partial q}{\partial y}\right)^2 \frac{1}{N}\sum (y_i - \overline{y})^2$$
$$+ 2\,\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\frac{1}{N}\sum (x_i - \overline{x})(y_i - \overline{y})$$

σ_q for arbitrary σ_x and σ_y

σ_x and σ_y can be correlated $\longrightarrow$

$$\boxed{\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2 + 2\,\frac{\partial q}{\partial x}\cdot\frac{\partial q}{\partial y}\,\sigma_{xy}}$$

<u>covariance</u> σ_{xy} $\longrightarrow$

$$\boxed{\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})\cdot(y_i - \overline{y})}$$

when σ_x and σ_y are independent $\sigma_{xy} = 0$ $\longrightarrow$ $\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2$

Coefficient of Linear Correlation

N pairs of values $(x_1, y_1), \ldots, (x_N, y_N)$

$y = A + Bx$ ← do N pairs of (x_i, y_i) satisfy a linear relation ?

$$\boxed{r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} \qquad r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}}$$

← linear correlation coefficient or correlation coefficient

$$-1 \leq r \leq 1$$

suppose (x_i, y_i) all lie exactly on the line $y = A + Bx$

$$y_i = A + Bx_i$$

$$\bar{y} = A + B\bar{x}$$

$$y_i - \bar{y} = B(x_i - \bar{x})$$

$$r = \frac{B \sum (x_i - \bar{x})^2}{\sqrt{\sum (x_i - \bar{x})^2 \cdot B^2 \sum (x_i - \bar{x})^2}} = \frac{B}{|B|} = \pm 1$$

if r is close to $\pm$ 1

when x and y are linearly correlated

suppose, there is no relationship between x and y

$$\sum (x_i - \bar{x})(y_i - \bar{y}) \to 0$$

$$r = 0$$

if r is close to 0

when there is no relationship between x and y

x and y are uncorrelated

Quantitative Significance of *r*

Student i	1	2	3	4	5	6	7	8	9	10
Homework x_i	90	60	45	100	15	23	52	30	71	88
Exam y_i	90	71	65	100	45	60	75	85	100	80

calculate correlation coefficient

$$r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2}}$$

$r = 0.8$
$N = 10$

probability that N measurements of two uncorrelated variables x and y would produce $r \geq r_0$ → **Table C**

Table 9.4. The probability $Prob_N(|r| \geqslant r_o)$ that N measurements of two uncorrelated variables x and y would produce a correlation coefficient with $|r| \geqslant r_o$. Values given are percentage probabilities, and blanks indicate values less than 0.05%.

	r_o										
N	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
3	100	94	87	81	74	67	59	51	41	29	0
6	100	85	70	56	43	31	21	12	6	1	0
10	100	78	58	40	25	14	7	2	0.5		0
20	100	67	40	20	8	2	0.5	0.1			0
50	100	49	16	3	0.4						0

$Prob_N(|r| \geqslant r_o)$

$Prob_{10}(|r| \geqslant 0.8)$
$= 0.5\ \%$

↓

it is very unlikely that
x and y are uncorrelated

↕

it is very likely that
x and y are correlated

↓

the correlation is highly significant

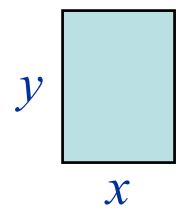
correlation is "significant" if $Prob_N(|r| \geq r_0)$ is less than 5 %

correlation is "highly significant" if $Prob_N(|r| \geq r_0)$ is less than 1 %

Example:

Calculate the covariance and the correlation coefficient *r* for the following six pairs of measurements of two sides *x* and *y* of a rectangle. Would you say these data show a significant linear correlation coefficient? Highly significant?

	A	B	C	D	E	F	
$x =$	71	72	73	75	76	77	mm
$y =$	95	96	96	98	98	99	mm



$$\bar{x} = 74$$

$$\bar{y} = 97$$

covariance $$\sigma_{xy} = \frac{1}{N}\sum(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{6}\left((-3)\times(-2) + \ldots + 3\times 2\right) = \underline{3}$$

correlation coefficient $$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \underline{0.98}$$

Table C $Prob_6(|r| \geq 0.98) \approx 0.2\%$ therefore, the correlation is both significant and highly significant

The square-root rule for a counting experiment

for events which occur at random

but with a definite average rate N occurrences in a time T

the standard deviation is $\sqrt{N}$

$$\boxed{(\text{number of counts in time } T) = N \pm \sqrt{N}}$$

average number of counts in a time T uncertainty

$$\boxed{(\text{fractional uncertainty}) = \frac{\sqrt{N}}{N} = \frac{1}{\sqrt{N}} \quad \text{reduces with increasing } N}$$

<u>Examples</u>

Photoemission:

if average emission rate is 10^6 photons/s, uncertainty is $\sqrt{10^6} = 10^3$ photons/s

and expected number is $10^6 \pm 10^3$ photons/s

Rain droplets on a windshield:

if average rate is 100 droplets/s, uncertainty is $\sqrt{100} = 10$ droplets/s

and expected number is 100 ± 10 droplets/s

fractional uncertainty

$$\frac{1}{\sqrt{N}} = \frac{1}{1000}$$

$$\frac{1}{\sqrt{N}} = \frac{1}{10}$$

Chi Squared Test for a Distribution

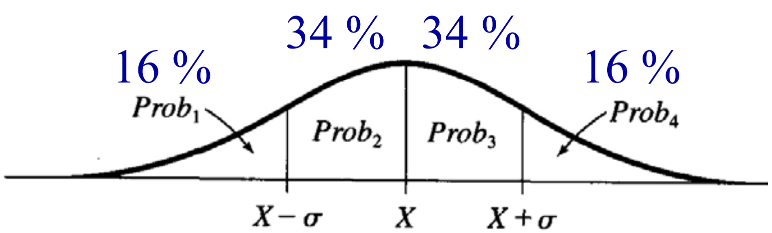
40 measured values of x (in cm)

731	772	771	681	722	688	653	757	733	742
739	780	709	676	760	748	672	687	766	645
678	748	689	810	805	778	764	753	709	675
698	770	754	830	725	710	738	638	787	712

are these measurements governed by a Gauss distribution ?

$$\bar{X} = \frac{\sum x_i}{N} = 730.1 \text{ cm}$$

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{X})^2}{N-1}} = 46.8 \text{ cm}$$



Bin number k	Observed number O_k	Expected number $E_k = NProb_k$	Difference $O_k - E_k$
1	8	6.4	1.6
2	10	13.6	−3.6
3	16	13.6	2.4
4	6	6.4	−0.4

$$\frac{O_k - E_k}{\sqrt{E_k}} = \frac{\text{deviation}}{\text{expected size of its fluctuation}} \sim 1 \ ?$$

$$\boxed{\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}}$$ chi squared

$\chi^2 \lesssim n$ observed and expected distributions agree about as well as expected

$\chi^2 \gg n$ significant disagreement between observed and expected distributions

O_k – observed number

E_k – expected number

$\sqrt{E_k}$ – fluctuations of E_k

$$\chi^2 = \sum_{k=1}^{4} \frac{(O_k - E_k)^2}{E_k}$$

$$= \frac{(1.6)^2}{6.4} + \frac{(-3.6)^2}{13.6} + \frac{(2.4)^2}{13.6} + \frac{(-0.4)^2}{6.4}$$

$$= 1.80 < n \longrightarrow$$ no reason to doubt that the measurements were governed by a Gauss distribution

Degrees of Freedom and Reduced Chi Squared

a better procedure is to compare χ^2 not with the number of bins n but instead with the number of degree of freedom d

n is the number of bins

c is the number of parameters that had to be calculated from the data to compute the expected numbers E_k

c is called <u>the number of constrains</u>

d is <u>the number of degrees of freedom</u>

$$\underline{d = n - c}$$

test for a Gauss distribution $G_{X,\sigma}(x)$ $\rightarrow c = 3$ ← N, X, σ

$$\boxed{\begin{array}{l} (\text{expected average value of } \chi^2) = d = n - c \\ \tilde{\chi}^2 = \chi^2 / d \quad \text{reduced chi squared} \\ (\text{expected average value of } \tilde{\chi}^2) = 1 \end{array}}$$

Probabilities of Chi Squared

quantitative measure of agreement between observed data and their expected distribution

(expected average value of χ^2) $= d = n - c$

$\tilde{\chi}^2 = \chi^2 / d$

(expected average value of $\tilde{\chi}^2$) $= 1$

$\chi^2 = 1.80$

$d = 4 - 3 = 1$

$\tilde{\chi}^2 = 1.80$

$\text{Prob}\,(\tilde{\chi}^2 \geq 1.80) \simeq 18\%$ ← **Table D**

	$\tilde{\chi}_o^2$												
d	0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2	3	4	5	6
1	100	62	48	39	32	26	22	19	16	8	5	3	1
2	100	78	61	47	37	29	22	17	14	5	2	0.7	0.2
3	100	86	68	52	39	29	21	15	11	3	0.7	0.2	—
5	100	94	78	59	42	28	19	12	8	1	0.1	—	—
10	100	99	89	68	44	25	13	6	3	0.1	—	—	—
15	100	100	94	73	45	23	10	4	1	—	—	—	—

probability of obtaining a value of $\tilde{\chi}^2$ greater or equal to $\tilde{\chi}_0^2$, assuming the measurements are governed by the expected distribution

disagreement is "significant" if $Prob_N(\tilde{\chi}^2 \geq \tilde{\chi}_0^2)$ is less than 5 %

disagreement is "highly significant" if $Prob_N(\tilde{\chi}^2 \geq \tilde{\chi}_0^2)$ is less than 1 %

reject the expected distribution