Intermediate Lab

PHYS 3870

Lecture 5

Comparing Data and Models—Quantitatively

Non-linear Regression

References: Taylor Ch. 9, 12
Also refer to "Glossary of Important Terms in Error Analysis"

Intermediate Lab

PHYS 3870

Errors in Measurements and Models

A Review of What We Know

Quantifying Precision and Random (Statistical) Errors

The "best" value for a group of measurements of the same quantity is the

Average

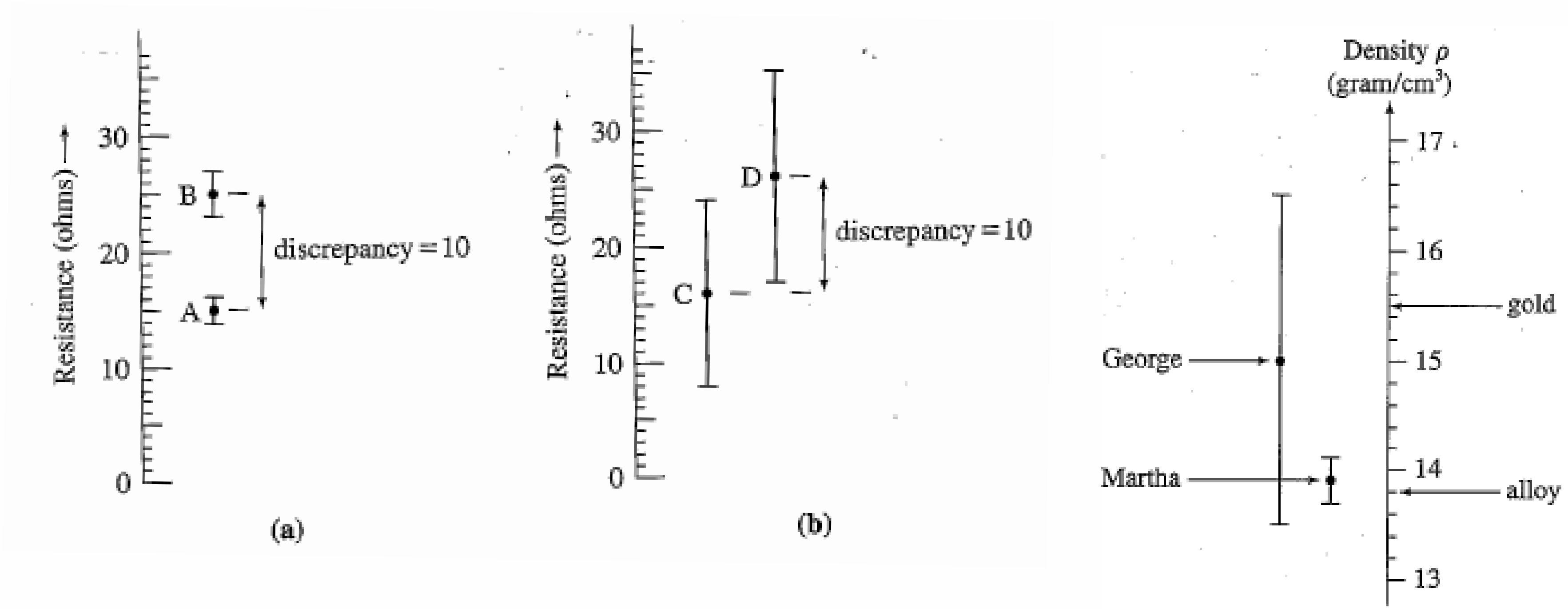
What is an estimate of the random error?

Deviations

A. If the average is the the best guess, then DEVIATIONS (or discrepancies) from best guess are an estimate of error

B. One estimate of error is the range of deviations.

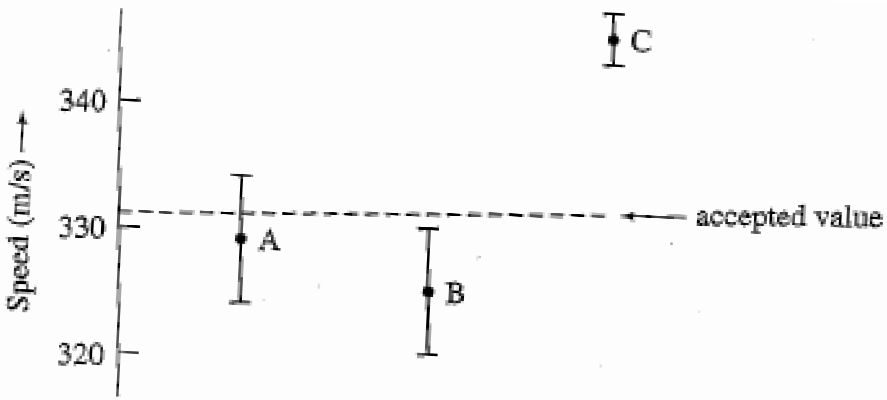
Single Measurement: Comparison with Other Data

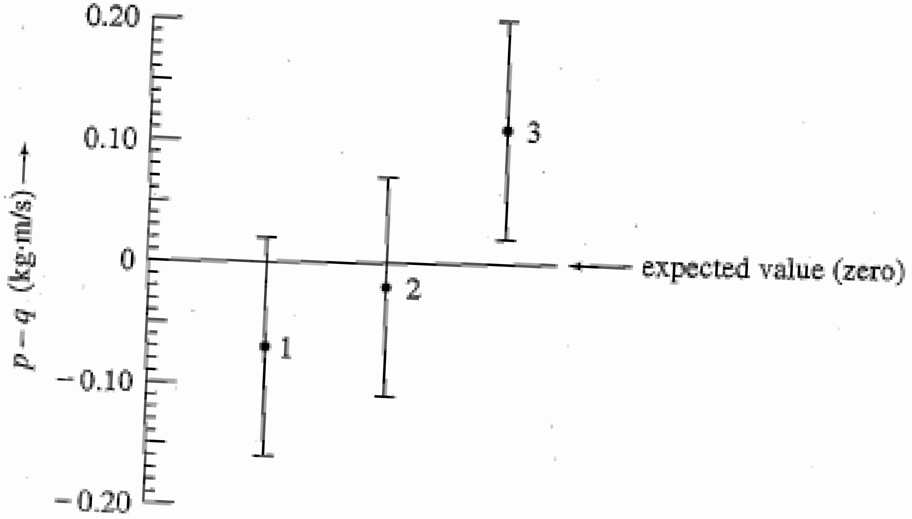


Comparison of precision or accuracy?

$$Precent\ Difference = \frac{(x_1 - x_2)}{\frac{1}{2}(x_1 + x_2)}$$

Single Measurement: Direct Comparison with Standard





Comparison of precision or accuracy?

$$Precent\ Error = \frac{x_{meansured} \quad - x_{Known}}{x_{Known}}$$

Multiple Measurements of the Same Quantity

Our statement of the best value and uncertainty is:

$(\langle t \rangle \pm \sigma_t)$ sec at the 68% confidence level for N measurements

1. **Note the precision of our measurement is reflected in the estimated error which states what values we would expect to get if we repeated the measurement**
2. **Precision is defined as a measure of the reproducibility of a measurement**
3. **Such errors are called random (statistical) errors**
4. **Accuracy is defined as a measure of who closely a measurement matches the true value**
5. **Such errors are called systematic errors**

Multiple Measurements of the Same Quantity

Standard Deviation

The best guess for the error in a group of N identical randomly distributed measurements is given by the standard deviation

$$\ldots\sigma = \sqrt{\frac{1}{(N-1)} \sum_{i=1}^{N} (t_i - \bar{t})^2}$$

this is the rms (root mean squared deviation or (sample) standard deviation

It can be shown that (see Taylor Sec. 5.4) σ_t is a reasonable estimate of the uncertainty. In fact, for normal (Gaussian or purely random) data, it can be shown that

(1) 68% of measurements of t will fall within $\langle t \rangle \pm \sigma_t$
(2) 95% of measurements of t will fall within $\langle t \rangle \pm 2\sigma_t$
(3) 98% of measurements of t will fall within $\langle t \rangle \pm 3\sigma_t$
(4) this is referred to as the confidence limit

Summary: the standard format to report the best guess and the limits within which you expect 68% of subsequent (single) measurements of t to fall within is $\langle t \rangle \pm \sigma_t$

Multiple Measurements of the Same Quantity

Standard Deviation of the Mean

If we were to measure t again N times (not just once), we would be even more likely to find that the second average of N points would be close to <t>.

The standard error or standard deviation of the mean is given by…

$$\sigma_{SDOM} = \frac{\sigma_{SD}}{\sqrt{N}} = \sqrt{\frac{1}{N\,(N-1)}\sum_{i=1}^{N}(t_i - \bar{t})^2}$$

This is the limits within which you expect the average of N addition measurements to fall within at the 68% confidence limit

Errors in Models—Error Propagation

Define error propagation [Taylor, p. 45]

A method for determining the error inherent in a derived quantity from the errors in the measured quantities used to determine the derived quantity

Recall previous discussions [Taylor, p. 28-29]

I. Absolute error: $(<t> \pm \sigma_t)$ sec

II. Relative (fractional) Error: $<t>$ sec $\pm$ $(\sigma_t / <t>)$%

III. Percentage uncertainty: fractional error in % units

Specific Rules for Error Propogation

Refer to **[Taylor, sec. 3.2]** for specific rules of error propagation:

1. Addition and Subtraction **[Taylor, p. 49]**
   For $q_{best}=x_{best}\pm y_{best}$ the error is $\delta q \approx \delta x+\delta y$
   Follows from $q_{best}\pm\delta q = (x_{best}\pm \delta x) \pm (y_{best} \pm\delta y) = (x_{best}\pm y_{best}) \pm(\delta x \pm\delta y)$

2. Multiplication and Division **[Taylor, p. 51]**
   For $q_{best}=x_{best} * y_{best}$ the error is $(\delta q/ q_{best}) \approx (\delta x/x_{best})+(\delta y/y_{best})$

3. Multiplication by a constant (exact number) **[Taylor, p. 54]**
   For $q_{best}= B(x_{best})$ the error is $(\delta q/ q_{best}) \approx |B| (\delta x/x_{best})$
   Follows from 2 by setting $\delta B/B=0$

4. Exponentiation (powers) **[Taylor, p. 56]**
   For $q_{best}= (x_{best})^n$ the error is $(\delta q/ q_{best}) \approx n (\delta x/x_{best})$
   Follows from 2 by setting $(\delta x/x_{best})=(\delta y/y_{best})$

General Formula for Error Propagation

General formula for uncertainty of a function of one variable

$$\delta q = \left|\frac{\partial q}{\partial x}\right| \cdot \delta x$$ [Taylor, Eq. 3.23]

Can you now derive for specific rules of error propagation:

1. Addition and Subtraction **[Taylor, p. 49]**
2. Multiplication and Division **[Taylor, p. 51]**
3. Multiplication by a constant (exact number) **[Taylor, p. 54]**
4. Exponentiation (powers) **[Taylor, p. 56]**

General Formula for Multiple Variables

Uncertainty of a function of multiple variables [Taylor, Sec. 3.11]

1. It can easily (no, really) be shown that (see Taylor Sec. 3.11) for a function of several variables

$$\delta q(x, y, z, \ldots) = \left|\frac{\partial q}{\partial x}\right| \cdot \delta x + \left|\frac{\partial q}{\partial y}\right| \cdot \delta y + \left|\frac{\partial q}{\partial z}\right| \cdot \delta z + \ldots$$

[Taylor, Eq. 3.47]

2. More correctly, it can be shown that (see Taylor Sec. 3.11) for a function of several variables

$$\delta q(x, y, z, \ldots) \leq \left|\frac{\partial q}{\partial x}\right| \cdot \delta x + \left|\frac{\partial q}{\partial y}\right| \cdot \delta y + \left|\frac{\partial q}{\partial z}\right| \cdot \delta z + \ldots$$

[Taylor, Eq. 3.47]

where the equals sign represents an upper bound, as discussed above.

3. For a function of several *independent and random* variables

$$\delta q(x, y, z, \ldots) = \sqrt{\left(\frac{\partial q}{\partial x} \cdot \delta x\right)^2 + \left(\frac{\partial q}{\partial y} \cdot \delta y\right)^2 + \left(\frac{\partial q}{\partial z} \cdot \delta z\right)^2 + \ldots}$$

[Taylor, Eq. 3.48]

Again, the proof is left for Ch. 5.

Error Propagation: General Case

Thus, if x and y are:

a) Independent (determining x does not affect measured y)

b) Random (equally likely for +δx as –δx)

Then method the methods above overestimate the error

Consider the arbitrary derived quantity q(x,y) of two independent random variables x and y.

Expand q(x,y) in a Taylor series about the expected values of x and y (i.e., at points near X and Y).

Fixed, shifts peak of distribution

$$q(x,y) = q(X,Y) + \left(\frac{\partial q}{\partial x}\right)\Big|_X (x - X) + \left(\frac{\partial q}{\partial y}\right)\Big|_Y (y - Y)$$

Fixed **Distribution centered at X with width σ_X**

Error for a function of Two Variables: Addition in Quadrature

$$\delta q(x,y) = \sigma_q = \sqrt{\left[\left(\frac{\partial q}{\partial x}\right)\Big|_X \sigma_x\right]^2 + \left[\left(\frac{\partial q}{\partial y}\right)\Big|_Y \sigma_y\right]^2}$$

Independent (Random) Uncertaities and Gaussian Distributions

For Gaussian distribution of measured values which describe quantities with random uncertainties, it can be shown that (the dreaded ICBST), errors add in quadrature [see Taylor, Ch. 5]

$\delta q \neq \delta x + \delta y$

But, $\delta q = \sqrt{[(\delta x)2 + (\delta y)^2]}$

1. This is proved in [Taylor, Ch. 5]
2. ICBST [Taylor, Ch. 9] Method A provides an upper bound on the possible errors

Gaussian Distribution Function

Distribution Function

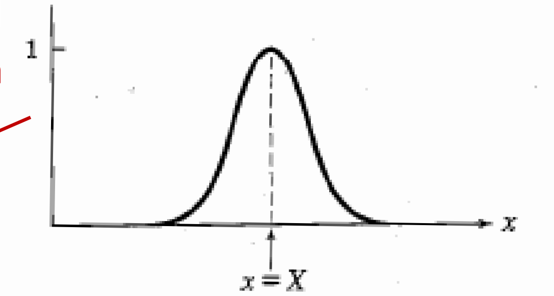
Independent Variable

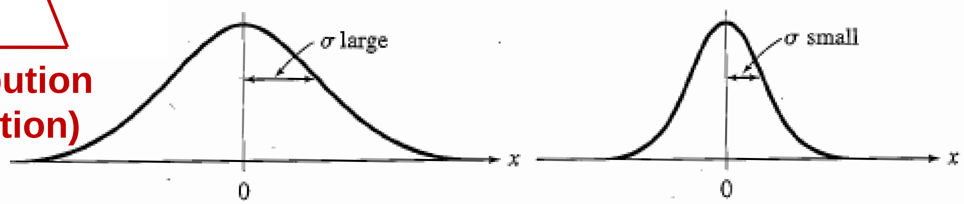
Center of Distribution (mean)

$$G_{X,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2}$$

Normalization Constant

Width of Distribution (standard deviation)





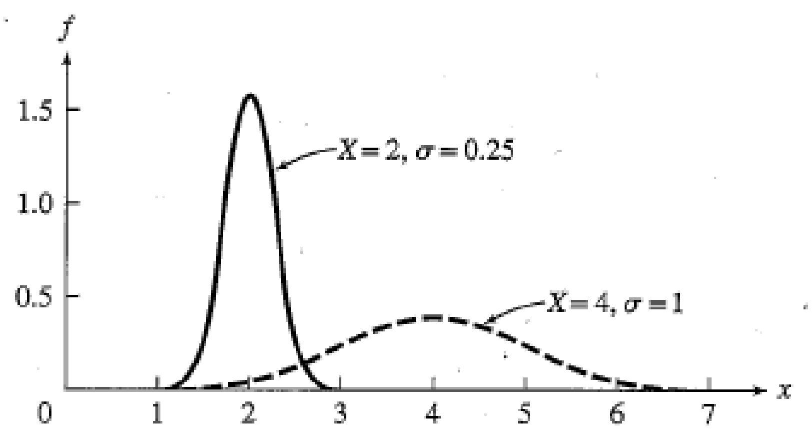


Figure 5.10. Two normal, or Gauss, distributions.

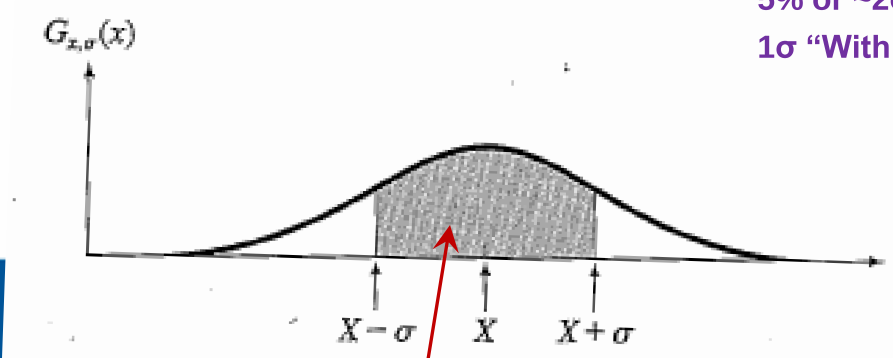
Gaussian Distribution Function

Standard Deviation of Gaussian Distribution

$$Prob(\text{within } \sigma) = \int_{X-\sigma}^{X+\sigma} G_{X,\sigma}(x)\,dx \qquad (5.32)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \int_{X-\sigma}^{X+\sigma} e^{-(x-X)^2/2\sigma^2}\,dx. \qquad (5.33)$$

See Sec. 10.6: Testing of Hypotheses



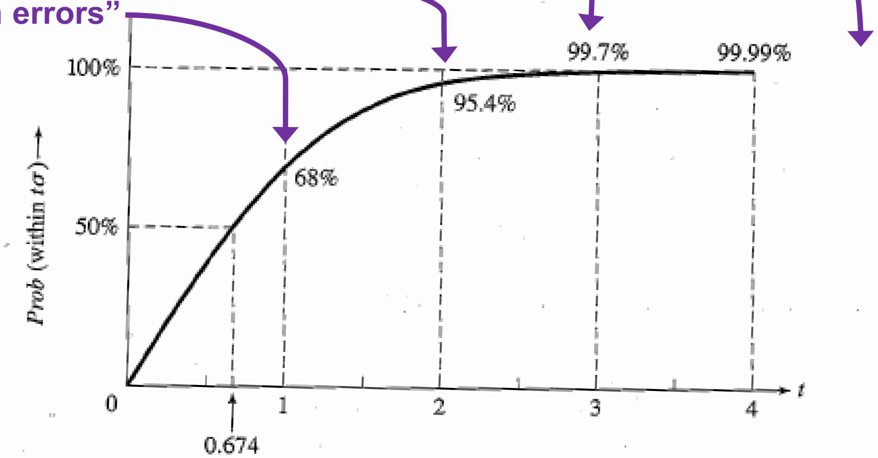
Area under curve (probability that –σ<x<+σ) is 68%

5 ppm or ~5σ "Valid for HEP"

1% or ~3σ "Highly Significant"

5% or ~2σ "Significant"

1σ "Within errors"



t	0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.5	3.0	3.5	4.0
Prob (%)	0	20	38	55	68	79	87	92	95.4	98.8	99.7	99.95	99.99

Figure 5.13. The probability *Prob*(within $t\sigma$) that a measurement of x will fall within t standard deviations of the true value $x = X$. Two common names for this function are the *normal error integral* and the *error function*, erf(t).

Mean of Gaussian Distribution as "Best Estimate"

Principle of Maximum Likelihood

To find the most likely value of the mean (the best estimate of ẋ), find X that yields the highest probability for the data set.

Consider a data set $\{x_1, x_2, x_3 \ldots x_N\}$

Each randomly distributed with

$$Prob_{X,\sigma}(x_i) = G_{X,\sigma}(x_i) \equiv \frac{1}{\sigma\sqrt{2\pi}} e^{-(x_i-X)^2/2\sigma} \propto \frac{1}{\sigma} e^{-(x_i-X)^2/2\sigma}$$

The combined probability for the full data set is the product

$$Prob_{X,\sigma}(x_1, x_2 \ldots x_N) = Prob_{X,\sigma}(x_1) \times Prob_{X,\sigma}(x_2) \times \ldots \times Prob_{X,\sigma}(x_N)$$

$$\propto \frac{1}{\sigma} e^{-(x_1-X)^2/2\sigma} \times \frac{1}{\sigma} e^{-(x_2-X)^2/2\sigma} \times \ldots \times \frac{1}{\sigma} e^{-(x_N-X)^2/2\sigma} = \frac{1}{\sigma^N} \sum e^{-(x_i-X)^2/2\sigma}$$

Best Estimate of X is from maximum probability or minimum summation

Minimize Sum	$\sum_{i=1}^{N} (x_i - X)^2/\sigma$	**Solve for derivative set to 0**	$\sum_{i=1}^{N} (x_i - X) = 0$	**Best estimate of X**	$X_{best} = \sum x_i / N$

Uncertaity of "Best Estimates" of Gaussian Distribution

Principle of Maximum Likelihood

To find the most likely value of the mean (the best estimate of ẋ), find X that yields the highest probability for the data set.

Consider a data set $\{x_1, x_2, x_3 \ldots x_N\}$

The combined probability for the full data set is the product

$$Prob_{X,\sigma}(x_1, x_2 \ldots x_N) = Prob_{X,\sigma}(x_1) \times Prob_{X,\sigma}(x_2) \times \ldots \times Prob_{X,\sigma}(x_N)$$

$$\propto \frac{1}{\sigma} e^{-(x_1-X)^2/2\sigma} \times \frac{1}{\sigma} e^{-(x_2-X)^2/2\sigma} \times \ldots \times \frac{1}{\sigma} e^{-(x_N-X)^2/2\sigma} = \frac{1}{\sigma^N} \sum e^{-(x_i-X)^2/2\sigma}$$

Best Estimate of X is from maximum probability or minimum summation

Minimize Sum $\sum_{i=1}^{N} (x_i - X)^2/\sigma$ **Solve for derivative wrst X set to 0** $\sum_{i=1}^{N} (x_i - X) = 0$ **Best estimate of X** $X_{best} = \sum x_i / N$

Best Estimate of σ is from maximum probability or minimum summation

Minimize Sum $\sum_{i=1}^{N} (x_i - X)^2/\sigma$ **Solve for derivative wrst σ set to 0** See Prob. 5.26 **Best estimate of σ** $\sigma_{best} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - X)^2/\sigma}$

Weighted Averages

Question: How can we properly combine two or more separate independent measurements of the same randomly distributed quantity to determine a best combined value with uncertainty?

Weighted Averages

The probability of measuring two such measurements is

$$Prob_x(x_1, x_2) = Prob_x(x_1)\, Prob_x(x_2)$$

$$= \frac{1}{\sigma_1 \sigma_2} e^{-\chi^2/2} \quad where\ \ \chi^2 \equiv \left[\frac{(x_1 - X)}{\sigma_1}\right]^2 + \left[\frac{(x_2 - X)}{\sigma_2}\right]^2$$

To find the best value for X, find the maximum Prob or minimum X^2

Best Estimate of χ is from maximum probability or minimum summation

Minimize Sum

$$\chi^2 \equiv \left[\frac{(x_1 - X)}{\sigma_1}\right]^2 + \left[\frac{(x_2 - X)}{\sigma_2}\right]^2$$

Solve for derivative wrst χ set to 0

$$2\left[\frac{(x_1 - X)}{\sigma_1}\right] + 2\left[\frac{(x_2 - X)}{\sigma_2}\right] = 0$$

Solve for best estimate of χ

$$X_{best} = \left(\frac{x_1}{\sigma_1{}^2} + \frac{x_2}{\sigma_2{}^2}\right) \Big/ \left(\frac{1}{\sigma_1{}^2} + \frac{1}{\sigma_2{}^2}\right)$$

This leads to

$$x_{W\_avg} = \frac{w_1 x_1 + w_2 x_2}{w_1 + w_2} = \frac{\sum w_i\, x_i}{\sum w_i} \quad where\ w_i = {}^{1}\!/_{(\sigma_i)^2}$$

Note: If $w_1 = w_2$, we recover the standard result $X_{wavg} = (1/2)\,(x_1 + x_2)$

Finally, the width of a weighted average distribution is $\sigma_{weighted\ avg} = \frac{1}{\sum_i w_i}$

Intermediate Lab

PHYS 3870

Comparing Measurements to Linear Models

Summary of Linear Regression

Question 1: What is the Best Linear Fit (A and B)?

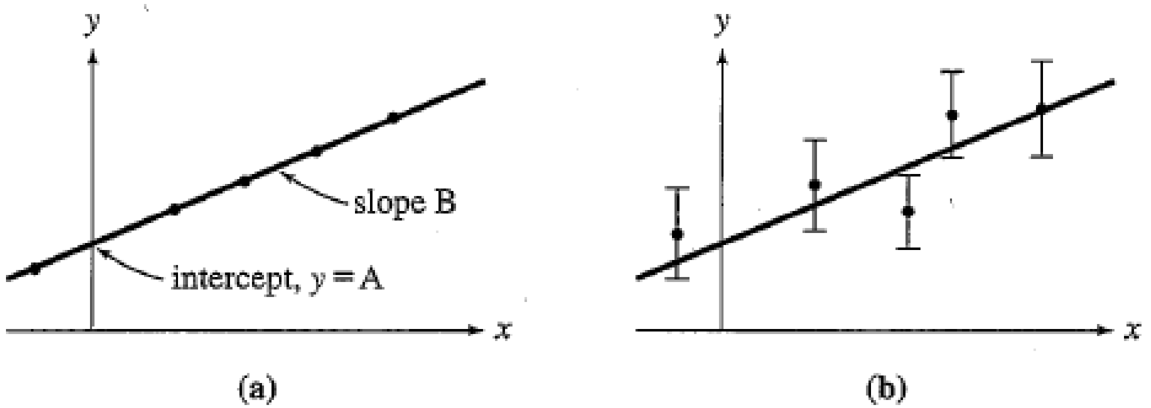


Figure 8.1. (a) If the two variables y and x are linearly related as in Equation (8.1), and if there were no experimental uncertainties, then the measured points (x_i, y_i) would all lie exactly on the line $y = A + Bx$. (b) In practice, there always are uncertainties, which can be shown by error bars, and the points (x_i, y_i) can be expected only to lie reasonably close to the line. Here, only y is shown as subject to appreciable uncertainties.

Best Estimate of intercept, A , and slope, B, for Linear Regression or Least Squares-Fit for Line

For the linear model y = A + B x

Intercept: $A = \frac{\sum x^2 \sum y - \sum x \sum xy}{N \sum x^2 - (\sum x)^2}$ $\sigma_A = \sigma_y \frac{\sum x^2}{N \sum x^2 - (\sum x)^2}$

Slope $B = \frac{N \sum xy - \sum x \sum xy}{N \sum x^2 - (\sum x)^2}$ $\sigma_B = \sigma_y \frac{N}{N \sum x^2 - (\sum x)^2}$

where $\sigma_y = \sqrt{\frac{1}{N-2} \sum [y_i - (A + Bx_i)]^2}$

"Best Estimates" of Linear Fit

Consider a linear model for y_i, $y_i = A + Bx_i$

The probability of obtaining an observed value of y_i is

$$Prob_{A,B}(y_1 \ldots y_N) = Prob_{A,B}(y_1) \times \ldots \times Prob_{A,B}(y_N)$$

$$= \frac{1}{{\sigma_y}^N} e^{-\chi^2/2} \quad where \;\; \chi^2 \equiv \sum_{i=1}^{N} \frac{[y_i - (A + Bx_i)]^2}{{\sigma_y}^2}$$

To find the best simultaneous values for A and B, find the maximum Prob or minimum X^2

Best Estimates of A and B are from maximum probibility or minimum summation

Minimize Sum

$$\chi^2 \equiv \sum_{i=1}^{N} \frac{[y_i - (A + Bx_i)]^2}{{\sigma_y}^2}$$

Solve for derivative wrst A and B set to 0

$$\frac{\partial \chi^2}{\partial A} \equiv \frac{-2}{{\sigma_y}^2} \sum_{i=1}^{N} [y_i - (A + Bx_i)] = 0$$

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

Best estimate of A and B

$$AN + B \sum x_i = \sum y_i$$

$$A \sum x_i + A \sum {x_i}^2 = \sum x_i y_i$$

“Best Estimates” of Linear Fit

Best Estimates of A and B are from maximum probability or minimum summation

Minimize Sum

$$\chi^2 \equiv \sum_{i=1}^{N} \frac{[y_i - (A + Bx_i)]^2}{\sigma_y{}^2}$$

Solve for derivative wrst A and B set to 0

$$\frac{\partial \chi^2}{\partial A} \equiv \frac{-2}{\sigma_y{}^2} \sum_{i=1}^{N} [y_i - (A + Bx_i)] = 0$$

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

Best estimate of A and B

$$AN + B\sum x_i = \sum y_i$$

$$A\sum x_i + A\sum x_i{}^2 = \sum x_i y_i$$

For the linear model y = A + B x

Intercept:	$A = \frac{\sum x^2 \sum y - \sum x \sum xy}{N\sum x^2 - (\sum x)^2}$	$\sigma_A = \sigma_y \frac{\sum x^2}{N\sum x^2 - (\sum x)^2}$	**(Prob (8.16)**
Slope	$B = \frac{N\sum xy - \sum x \sum xy}{N\sum x^2 - (\sum x)^2}$	$\sigma_B = \sigma_y \frac{N}{N\sum x^2 - (\sum x)^2}$	

where $\sigma_y = \sqrt{\frac{1}{N-2}\sum[y_i - (A + Bx_i)]^2}$

Correlation Coefficient

Combining the Schwartz inequality

$$|\sigma_{xy}| \leq \sigma_x \sigma_y$$

With the definition of the covariance

$$\sigma_{xy} \equiv \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) \rightarrow 0$$

The uncertainty in a function q(x,y) is

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

With a limiting value

$$\sigma_q \leq \left|\frac{\partial q}{\partial x}\right| \sigma_x + \left|\frac{\partial q}{\partial y}\right| \sigma_y$$

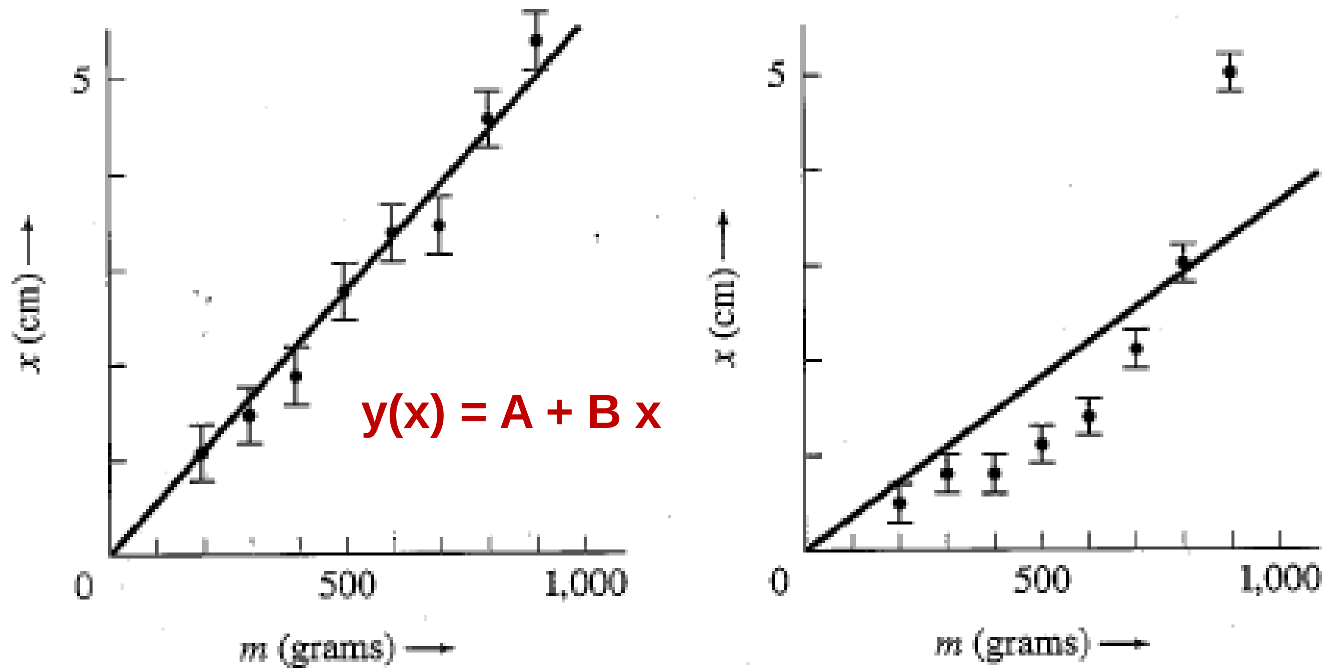
At last, the upper bound of errors is

$$\sigma_q = \left|\frac{\partial q}{\partial x}\right| \sigma_x + \left|\frac{\partial q}{\partial y}\right| \sigma_y$$

And for independent and random variables

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

Question 2: Is it Linear?



(b) (c)

Coefficient of Linear Regression:

$$r \equiv \frac{\sum[(x-\bar{\bar{x}})(y-\bar{\bar{y}})]}{\sqrt{\sum(x-\bar{\bar{x}})^2 \sum(y-\bar{\bar{y}})^2}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Consider the limiting cases for:

- r=0 (no correlation) [for any x, the sum over y-Y yields zero]
- r=±1 (perfect correlation). [Substitute yi-Y=B(xi-X) to get r=B/|B|=±1]

Tabulated Correlation Coefficient

Table C. The percentage probability $Prob_N(|r| \geq r_o)$ that N measurements of two uncorrelated variables give a correlation coefficient with $|r| \geq r_o$, as a function of N and r_o. (Blanks indicate probabilities less than 0.05%.)

r_o ← r value

N data points →

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
4	100	90	80	70	60	50	40	30	20	10	0
5	100	87	75	62	50	39	28	19	10	3.7	0
6	100	85	70	56	43	31	21	12	5.6	1.4	0
7	100	83	67	51	37	25	15	8.0	3.1	0.6	0
8	100	81	63	47	33	21	12	5.3	1.7	0.2	0
9	100	80	61	43	29	17	8.8	3.6	1.0	0.1	0
10	100	78	58	40	25	14	6.7	2.4	0.5		0
11	100	77	56	37	22	12	5.1	1.6	0.3		0
12	100	76	53	34	20	9.8	3.9	1.1	0.2		0
13	100	75	51	32	18	8.2	3.0	0.8	0.1		0
14	100	73	49	30	16	6.9	2.3	0.5	0.1		0
15	100	72	47	28	14	5.8	1.8	0.4			0
16	100	71	46	26	12	4.9	1.4	0.3			
17	100	70	44	24	11	4.1	1.1	0.2			
18	100	69	43	23	10	3.5	0.8	0.1			
19	100	68	41	21	9.0	2.9	0.7	0.1			
20	100	67	40	20	8.1	2.5	0.5	0.1			
25	100	63	34	15	4.8	1.1	0.2				
30	100	60	29	11	2.9	0.5					
35	100	57	25	8.0	1.7	0.2					
40	100	54	22	6.0	1.1	0.1					
45	100	51	19	4.5	0.6						0

N	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45
50	100	73	49	30	16	8.0	3.4	1.3	0.4	0.1
60	100	70	45	25	13	5.4	2.0	0.6	0.2	
70	100	68	41	22	9.7	3.7	1.2	0.3	0.1	
80	100	66	38	18	7.5	2.5	0.7	0.1		
90	100	64	35	16	5.9	1.7	0.4	0.1		
100	100	62	32	14	4.6	1.2	0.2			

Consider the limiting cases for:

- r=0 (no correlation)
- r=±1 (perfect correlation).

To gauge the confidence imparted by intermediate r values consult the table in Appendix C.

The values in Table C were calculated from the integral

$$Prob_N(|r| \geq |r_o|) = \frac{2\Gamma[(N-1)/2]}{\sqrt{\pi}\,\Gamma[(N-2)/2]} \int_{|r_o|}^{1} (1-r^2)^{(N-4)/2}\, dr.$$

See, for example, E. M. Pugh and G. H. Winslow, *The Analysis of Physical Measurements* (Addison-Wesley, 1966), Section 12-8.

Probability that analysis of N=70 data points with a correlation coefficient of r=0.5 is **not** modeled well by a linear relationship is 3.7%.
Therefore, it is very probably that y is linearly related to x.

If

$Prob_N(|r|>r_o)<32\%$ ➔ it is probably that y is linearly related to x

$Prob_N(|r|>r_o)<5\%$ ➔ it is very probably that y is linearly related to x

$Prob_N(|r|>r_o)<1\%$ ➔ it is highly probably that y is linearly related to x

Uncertainties in Slope and Intercept

Taylor:

For the linear model y = A + B x

Intercept: $A = \frac{\sum x^2 \sum y - \sum x \sum xy}{N \sum x^2 - (\sum x)^2}$ $\quad \sigma_A = \sigma_y \frac{\sum x^2}{N \sum x^2 - (\sum x)^2}$ **(Prob (8.16)**

Slope $B = \frac{N \sum xy - \sum x \sum xy}{N \sum x^2 - (\sum x)^2}$ $\quad \sigma_B = \sigma_y \frac{N}{N \sum x^2 - (\sum x)^2}$

where $\sigma_y = \sqrt{\frac{1}{N-2} \sum [y_i - (A + Bx_i)]^2}$

Relation to R^2 value:

$$\sigma_A = \sigma_B \sqrt{\frac{1}{N} \sum x^2}$$

$$\sigma_B = B \sqrt{\frac{1}{N-2} [(1/R^2) - 1]}$$

Intermediate Lab

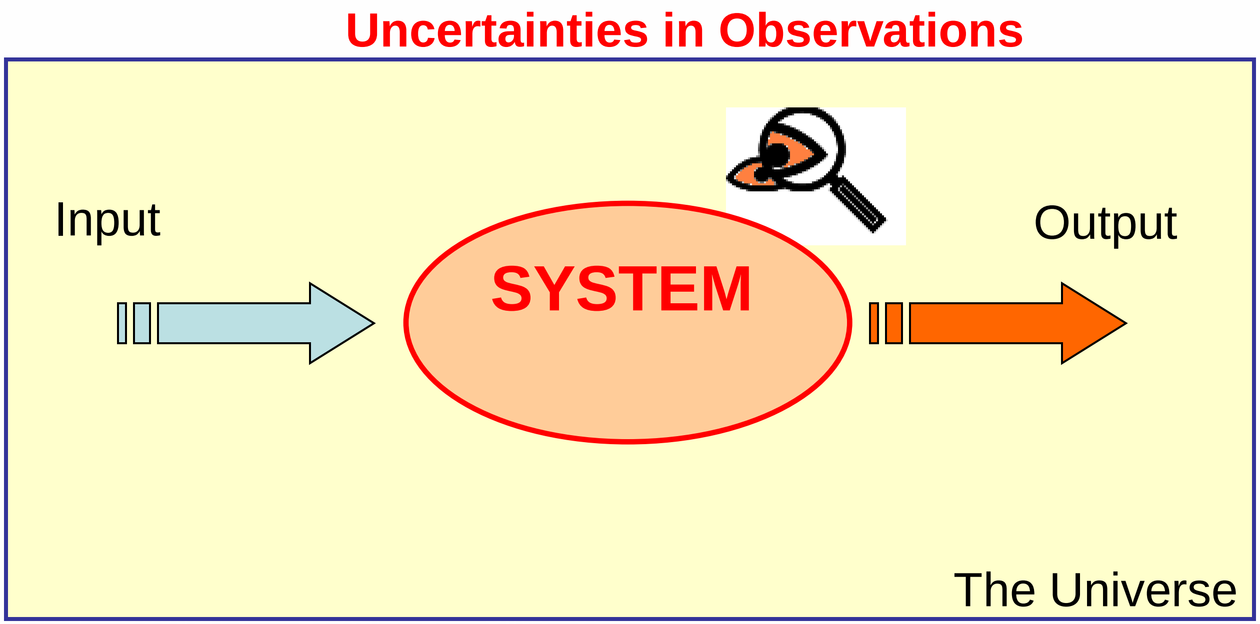
PHYS 3870

Comparing Measurements to Models

Non-Linear Regression

Motivating Regression Analysis

Question: Consider now what happens to the output of a nearly ideal experiment, if we vary how hard we poke the system (vary the input).



A more general model of response is a nonlinear response model

$$y(x) = f(x)$$

Questions for Regression Analysis

A more general model of response is a nonlinear response model

$$y(x) = f(x)$$

Two principle questions:

What are the best values of a set of fitting parameters,

What confidence can we place in how well the general model fits the data?

The solutions is familiar:

Evoke the Principle of Maximum Likelihood,

Minimize the summation of the exponent arguments, that is chi squared?

Recall what this looked like for a model with a constant value, linear model, polynomial model, and now a general nonlinear model

$$[y_i - \bar{Y}] \rightarrow [y_i - (A + Bx_i)] \rightarrow [y_i - (A + Bx_i + Cx_i^2)] \equiv [y_i - f_{fit}(x_i)]$$

$$\chi^2 = \sum_{i=1}^{N} \frac{[y_i - Y]^2}{\sigma_y^2} \rightarrow \sum_{i=1}^{N} \frac{[y_i - (A + Bx_i)]^2}{\sigma_y^2} \rightarrow \sum_{i=1}^{N} \frac{[y_i - (A + Bx_i + Cx_i^2)]^2}{\sigma_y^2} \rightarrow \sum_{i=1}^{N} \frac{[y_i - f_{fit}(x_i)]^2}{\sigma_y^2}$$

Chi Squared Analysis for Least Squares Fit

General definition for Chi squared (square of the normalized deviations)

$$\chi^2 = \sum_{1}^{n} \left(\frac{\text{observed value} - \text{expected value}}{\text{standard deviation}} \right)^2. \qquad (12.11)$$

Perfect model	$\chi^2 \to 0$
Good model	$\chi^2 \lesssim N$
Poor model	$\chi^2 \gg N$

Discrete distribution

$$\frac{O_k - E_k}{\sqrt{E_k}}. \qquad (12.4)$$

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}. \qquad (12.5)$$

Continuous distribution

$$\chi^2 = \sum_{1}^{N} \left(\frac{y_i - f(x_i)}{\sigma_i} \right)^2.$$

Expected Values for Chi Squared Analysis

$$\chi^2 = \sum_1^n \left(\frac{\text{observed value} - \text{expected value}}{\text{standard deviation}}\right)^2. \qquad (12.11)$$

or

$$\chi^2 = \sum_1^N \left(\frac{y_i - f(x_i)}{\sigma_i}\right)^2.$$

If the model is "good", each term in the sum is ~1, so

$$X^2 \approx \sum_{i=1}^{N} i \rightarrow N$$

More correctly the sum goes to the number of degrees of freedom, d≡N-c. The reduced Chi-squared value is

$$X_{red}^{\;2} \equiv \frac{1}{d} X^2 \approx \frac{1}{d} \sum_{i=1}^{N} i \rightarrow \frac{N}{d} \approx 1$$

So X_{red}^2=0 for perfect fit
X_{red}^2<1 for "good" fit
X_{red}^2>1 for "poor" fit
(looks a lot like r for linear fits doesn't it?)

Chi Squared Analysis for Least Squares Fit

Table 12.3. The expected numbers E_k and the observed numbers O_k for the 40 measurements of Table 12.1, with bins chosen as in Table 12.2.

Bin number k	Probability $Prob_k$	Expected number $E_k = NProb_k$	Observed number O_k
1	16%	6.4	8
2	34%	13.6	10
3	34%	13.6	16
4	16%	6.4	6

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

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

$$= 1.80.$$

Table 12.4. The data of Table 12.1, shown here with the differences $O_k - E_k$.

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

From the probability table ~99.5% (highly significant) confidence

Probabilities for Chi Squared

The values in Table D were calculated from the integral

$$Prob_d(\tilde{\chi}^2 \geqslant \tilde{\chi}_o^2) = \frac{2}{2^{d/2}\Gamma(d/2)} \int_{\chi_o}^{\infty} x^{d-1} e^{-x^2/2}\, dx.$$

See, for example, E. M. Pugh and G. H. Winslow, *The Analysis of Physical Measurements* (Addison-Wesley, 1966), Section 12-5.

Table D. The percentage probability $Prob_d(\tilde{\chi}^2 \geqslant \tilde{\chi}_o^2)$ of obtaining a value of $\tilde{\chi}^2 \geqslant \tilde{\chi}_o^2$ in an experiment with d degrees of freedom, as a function of d and $\tilde{\chi}_o^2$. (Blanks indicate probabilities less than 0.05%.)

	$\tilde{\chi}_o^2$														
d	0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	8.0	10.0
1	100	48	32	22	16	11	8.3	6.1	4.6	3.4	2.5	1.9	1.4	0.5	0.2
2	100	61	37	22	14	8.2	5.0	3.0	1.8	1.1	0.7	0.4	0.2		
3	100	68	39	21	11	5.8	2.9	1.5	0.7	0.4	0.2	0.1			
4	100	74	41	20	9.2	4.0	1.7	0.7	0.3	0.1	0.1				
5	100	78	42	19	7.5	2.9	1.0	0.4	0.1						

d	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
1	100	65	53	44	37	32	27	24	21	18	16	14	12	11	9.4	8.3
[illegible]	[illegible]	[illegible]	[illegible]	55	45	37	30	25	20	17	14	11	9.1	7.4	6.1	5.0

Reduced Chi Squared Analysis

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{i=1}^{8}\frac{(O_k - E_k)^2}{E_k}$$

Table 12.6. The percentage probability $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ of obtaining a value of $\tilde{\chi}^2$ greater than or equal to any particular value $\tilde{\chi}_o^2$, assuming the measurements concerned are governed by the expected distribution. Blanks indicate probabilities less than 0.05%. For a more complete table, see Appendix 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	—	—	—	—

d and $\tilde{\chi}_o^2$. For example, with 10 degrees of freedom ($d = 10$), we see that the probability of obtaining $\tilde{\chi}^2 \geq 2$ is 3%,

$$Prob_{10}(\tilde{\chi}^2 \geq 2) = 3\%.$$

Problem 12.2

12.2. ★★ Problem 4.13 reported 30 measurements of a time t, with mean $\bar{t} = 8.15$ sec and standard deviation $\sigma_t = 0.04$ sec. Group the values of t into four bins with boundaries at $\bar{t} - \sigma_t$, $\bar{t}$, and $\bar{t} + \sigma_t$, and count the observed number O_k in each bin $k = 1, 2, 3, 4$. Assuming the measurements were normally distributed with center at $\bar{t}$ and width σ_t, find the expected number E_k in each bin. Calculate χ^2. Is there any reason to doubt the measurements are normally distributed?

4.13. ★★ **(a)** Calculate the mean and standard deviation for the following 30 measurements of a time t (in seconds):

8.16, 8.14, 8.12, 8.16, 8.18, 8.10, 8.18, 8.18, 8.18, 8.24,
8.16, 8.14, 8.17, 8.18, 8.21, 8.12, 8.12, 8.17, 8.06, 8.10,
8.12, 8.10, 8.14, 8.09, 8.16, 8.16, 8.21, 8.14, 8.16, 8.13.

(You should certainly use the built-in functions on your calculator (or the spreadsheet you created in Problem 4.8 if you did), and you can save some button pushing if you drop all the leading 8s and shift the decimal point two places to the right before doing any calculation.) **(b)** We know that after several measurements, we can expect about 68% of the observed values to be within σ_t of $\bar{t}$ (that is, inside the range $\bar{t} \pm \sigma_t$). For the measurements of part (a), about how many would you expect to lie *outside* the range $\bar{t} \pm \sigma_t$? How many do? **(c)** In Chapter 5, I will show that we can also expect about 95% of the values to be within $2\sigma_t$ of $\bar{t}$ (that is, inside the range $\bar{t} \pm 2\sigma_t$). For the measurements of part (a), about how many would you expect to lie *outside* the range $\bar{t} \pm 2\sigma_t$? How many do?

Problem 12.2

$t_{avg} := \text{mean}(t)$ $t_{avg} = 8.149\ s$

$\sigma_t := \text{stdev}(t)$ $\sigma_t = 0.039\ s$

$k := 0..4$

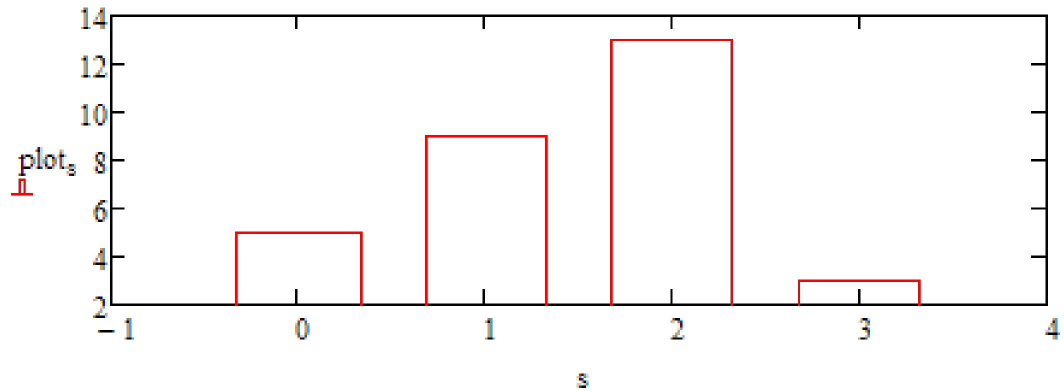
$s := 0..3$

$$bin := \begin{bmatrix} t_{avg} + -3\cdot\sigma_t \\ t_{avg} - \sigma_t \\ t_{avg} \\ (t_{avg} + \sigma_t) \\ t_{avg} + 3\cdot\sigma_t \end{bmatrix} \qquad bin = \begin{pmatrix} 8.033 \\ 8.111 \\ 8.149 \\ 8.188 \\ 8.265 \end{pmatrix} s$$

$$plot := hist\left(\frac{bin}{sec}, \frac{t}{sec}\right)$$

$\tau := 8\cdot sec, 8.01\cdot sec..8.25\cdot sec$

$$plot = \begin{pmatrix} 5 \\ 9 \\ 13 \\ 3 \end{pmatrix}$$



The distribution appears to be a little bit asymmetrical.