UC Berkeley - Physics 5 Series Labs

STATISTICS REFERENCE SHEET

A Note on Notation

One of the common frustrations as a physicist (or a mathematician) is that everyone has their own favorite notations and conventions. We have recently adopted the following textbook that we encourage you use for reference:

- Hughes and Hase, ***Measurements and Their Uncertainties: A Practical Guide to Modern Error Analysis***, https://ebookcentral-proquest-com.libproxy.berkeley.edu/lib/berkeley-ebooks/detail.action?docID=584562

We will be using the notation from this textbook throughout.

SECTION 1: STATISTICAL MEASURES FOR A SINGLE VARIABLE $\{y_i\}$

REFERENCE: HUGHES AND HASE, CHAPTER 2

Consider a quantity y that we are measuring and a set of N measurements $\{y_i\}$ where $i = 1, \cdots, N$.

The set of all values of y for a given population of objects gives the ***parent distribution***.[1] If i runs through every element of the parent distribution, then we will use the parent measures as defined in 1.3a and 1.4a, below. It is more common, however, to just take a subset of the full parent distribution to create a ***sample distribution***. In this case, the sample measures as defined in 1.3b and 1.4b, below, are more appropriate to use. There is a subtle difference in interpretation between the parent and sample statistics. The parent variance or parent standard deviation are properly considered as parameters that partially define the distribution whereas the sample variance and sample standard deviation are statistical measures that depend on the sample.

Mean:
$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{1.1}$$

Deviation from the Mean:
$$d_i = y_i - \bar{y}. \tag{1.2}$$

Variance (Parent):
$$\sigma^2 = \overline{d^2} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \overline{y^2} - \bar{y}^2. \tag{1.3a}$$

Variance (Sample):
$$\sigma_{N-1}^2 = \frac{1}{N-1}\sum_{i=1}^{N} d_i^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2. \tag{1.3b}$$

Standard Deviation (Parent):
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2} = \sqrt{\overline{y^2} - \bar{y}^2}. \tag{1.4a}$$

[1] Hughes and Hase, Section 2.6: "In the theory of statistics, the parent distribution refers to the number of possible measured values."

Standard Deviation (Sample): $$\sigma_{N-1} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2}. \quad (1.4b)$$

Standard Error: $$\alpha = \sigma_{N-1}/\sqrt{N}. \quad (1.5)$$

- The difference between parent and sample statistics only becomes significant if the number of data points is low. If you have roughly 5 or more data points you can pretty safely ignore the distinction.
- In Section 2.7, Hughes and Hase recommend reporting the error to one significant figure but for this class we will **report the error to two significant figures**.
  - Example: If I measure the spring constant k of a spring a number of times to get data $\{k_i\}$ I would report the result as $k = \bar{k} \pm \alpha_k$. Given ten measurements (in N/m) $\{k_i\}$ = {86, 85, 84, 89, 85, 89, 87, 85, 82, 85} the final answer would be presented as k = 85.7 ± 0.7 N/m. The sample standard deviation is σ_k = 2.2 N/m. If I were to perform the same experiment *once* on a different spring, finding a value of k = 71 N/m then I would report k = 71 ± 2 N/m and I would have roughly 68% confidence that the true spring constant was within 2 N/m of 71 N/m.

Hughes and Hase concludes Chapter 2 with the **five golden rules** for reporting an experimentally determined parameter:

1. The best estimate of a parameter is the mean $\bar{y}$.
2. The error is the standard error in the mean α_y.
3. Round up the error to the appropriate number of significant figures.
4. Match the number of decimal places in the mean to the standard error.
5. Include units.

SECTION 2: STATISTICAL MEASURES FOR TWO VARIABLES $\{x_i, y_i\}$

REFERENCE: HUGHES AND HASE, CHAPTER 7

Covariance: $$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}). \quad (2.1)$$

Correlation Coefficients: $$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}. \quad (2.2)$$

The covariance roughly tells you how much x and y change together. If larger y_i values tend to be paired with greater x_i values, then the covariance is positive. If larger y_i values tend to be paired with smaller x_i values, then the covariance is negative. The more *linear* the relationship is between x and y the larger the covariance will be. The units of covariance are the units of x times the units of y.

Note that the variance of a variable is just the covariance of a variable with itself (compare Eqs. 1.3 and 2.1).

The coefficient of linear correlation tells you how linear the relationship between x and y is. The correlation ρ_{xy} will always lie between -1 and 1. The closer $|\rho_{xy}|$ is to 1 the stronger the linear relationship is between x and y. The sign of ρ_{xy} gives the sign of the slope.

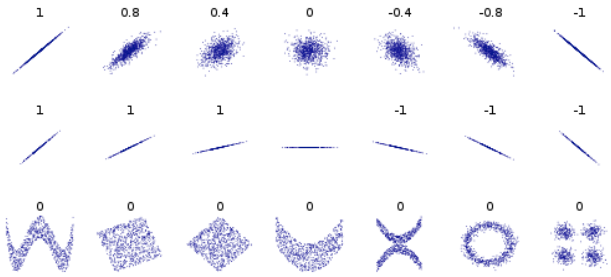

Figure 1: Various data sets and their correlations. http://en.wikipedia.org/wiki/Correlation_and_dependence

SECTION 3: SOURCES OF ERROR AND UNCERTAINTY[2]

REFERENCE: HUGHES AND HASE, CHAPTER 1

Every measurement or piece of data comes with its own uncertainty - we never know the result of a measurement *exactly*. There are many possible sources of error uncertainty in any given measurement and we classify them by how they cause our measurements to deviate from the "true" values.

Random Errors

These errors arise from random variations in the measurement technique or environmental conditions and yield different results each time the experiment is repeated. The uncertainty associated with random fluctuations of the reading is evaluated by the standard error which therefore requires multiple measurements. Random errors are *two-sided* and contribute to the *spread* of the data. The smaller the random uncertainty, the more precise the measurement is – ***Random errors influence precision.***

Random errors are best studied by statistical techniques as the observed distribution of results is randomly generated. The best estimate of the measured quantity is the mean and the spread of results is measured by the standard deviation.

Systematic Errors

These errors cause the measured quantity to be *shifted* in a *one-sided* way (either positive or negative) from the true value and manifest themselves as a consisten *offset* in all of the data. Because of this, these errors can be difficult to discern or detect and cannot be analyzed by statistical techniques. The smaller the systematic error, the more accurate a measurement is – ***Systematic errors influence accuracy***.

One a systematic error has been identified, data should be corrected to account for the error before further analysis.

Examples of systematic errors include *zero errors* (e.g. using a ruler whose end has been worn away or measuring a mass without accounting for the tare), calibration error, and insertion error (e.g. the current through a wire being affected by the insertion of a voltmeter or ammeter into the circuit).

Reading Errors

[2] In the 1990s, a group of professional metrologists published the *Guide to the Expression of Uncertainty in Measurement* to create an international standard on errors, uncertainties, and the distinction between the two. A copy of the revised edition published in 2008 is available on the course website.

Reading errors are also errors in precision that occur even in the absence of other sources of random errors. These errors occur because we cannot read a device's scale to infinite precision. For a digital device, the uncertainty is one in the last digit whereas for an analog one it is half a division. ***Even if repeated measurements yield the same value, precision is limited by the reading error.***

Other Errors

Other forms of error you may encounter include incorrect models, experimenter mistakes (misreading scales or confusion over units), and malfunctioning equipment.

> *When designing and carrying out an experimental procedure, your goal is to reduce systematic and instrumentation errors and to minimize random errors.*

Reporting a Measurement with Uncertainty

As summarized in Hughes and Hase section 2.8, when quoting results keep the following points in mind:

- Analyze the experimental data and calculate the mean $\bar{y}$; keep all significant figures at this stage.
- Calculate the standard error α; keep all significant figures at this stage.
- Determine the appropriate amount of significant figures to retain.

To report the result of a repeated measurement, give both the mean/central value and the standard error,

$$\bar{y} \pm \alpha.$$

If the standard error is zero or significantly smaller than the precision error of your measurement instrument, then the precision error is to be used in place of the standard error.

- Example 1: We are using a ruler with markings every millimeter. Repeated independent measurements of a length gives a mean of 11.89333 cm with a standard error of 0.57211 cm. The reported value is 11.89 ± 0.57 cm. (Keeping two significant figures in the standard error and using that to determine the appropriate precision in the mean).
- Example 2: Repeated independent measurements of a length gives a mean of 11.89333 cm with a standard error of 0.07211 cm. The standard error here is smaller than the precision error of one-half a division (0.05 cm) so the reported value should be 11.89 ± 0.05 cm.

Section 4: Propagation of Uncertainty

Reference: Hughes and Hase, Chapter 4

When using a measured quantity (with uncertainty) to compute a new quantity, we need to take care to propagate the uncertainty.

Propagation of Uncertainty for a Function of a Single Variable

Consider a single variable x and a derived quantity q that can be expressed as a function of x. That is, given a measurement x, the derived value of q for that data point is $q = q(x)$. Given an uncertainty δx in the measurement of x the propagated uncertainty for q is given by

$$\alpha_q = \left|\frac{dq}{dx}\right| \alpha_x. \tag{4.1}$$

- Example: If $q(x) = 1/x$ then $\alpha_q = \left|\frac{1}{x^2}\right| \alpha_x$, or $\frac{\alpha_q}{|q|} = \frac{\alpha_x}{|x|}$.

- Example: If $q(\theta) = \cos\theta$ then $\alpha_q = |\sin\theta|\alpha_\theta$.

Some of the more commonly occuring examples are given below.

Multiplication by a Constant:	$q(x) = cx$	$\alpha_q = \lvert c\rvert\alpha_x.$	(4.2)
Power:	$q(x) = x^n$	$\frac{\alpha_q}{\lvert q\rvert} = \lvert n\rvert\frac{\alpha_x}{\lvert x\rvert}.$	(4.3)
Exponential:	$q(x) = e^x$	$\frac{\alpha_q}{q} = \alpha_x.$	(4.4)
Logarithm:	$q(x) = \ln x$	$\alpha_q = \frac{\alpha_x}{x}$	(4.5)

Propagation of Uncertainty for a Function of Several Variables

Two variables x and y may be considered ***independent*** if their covariance is zero, $\sigma_{xy} = 0$. Consider two independent variables x and y and a derived quantity q that can be expressed as a function of both x and y. That is, given a measurement $\{x_i, y_i\}$, the derived value of q for that data point is $q = q(x,y)$. Given uncertainties α_x and α_y, the the propagated uncertainty for q is given by

$$\alpha_q = \sqrt{\left(\frac{\partial q}{\partial x}\alpha_x\right)^2 + \left(\frac{\partial q}{\partial y}\alpha_y\right)^2}. \tag{4.6}$$

This generalizes to functions of more than two variables in a straightforward manner. Some of the more commonly occuring examples are given below.

Sum or Difference:	$q(x,y) = x \pm y$	$\alpha_q = \sqrt{\alpha_x^2 + \alpha_y^2}.$	(4.7)
Product or Quotient:	$q(x,y) = xy$ or $\frac{x}{y}$	$\frac{\alpha_q}{\lvert q\rvert} = \sqrt{\left(\frac{\alpha_x}{x}\right)^2 + \left(\frac{\alpha_y}{y}\right)^2}.$	(4.8)

If variables x and y *aren't* independent then the actual uncertainty in $q(x,y)$ will be different than that given by Eq. 4.6. For extreme examples, consider the case where x and y are proportional, $y = ax$ (the correlation is $\rho_{xy} = \pm 1$) in the above cases.

Section 5: Regression

Reference: Hughes and Hase, Chapters 5, 6

5.1 – χ^2 and the Normalized Residual

Suppose we have a data set of two variables, an independent variable x with data $\{x_i\}$ and a dependent variable y with measurements and uncertainties $\{y_i \pm \alpha_i\}$ and we want to test a model of mathematical relationship between

the two, $y(x)$. Our model may depend on a set of undetermined parameters. Our goal is to determine how well a given model fits the data and which model parameters give the *best* fit to the data.

We define a dimensionless quantity χ^2 that is a cumulative measure of how far off our data points are from our hypothesis,

$$\chi^2 = \sum \frac{\left(y_i - y(x_i)\right)^2}{\alpha_i^2}. \tag{5.1.1}$$

The numerator in the sum is the square of the *residuals* and tells us how far each measured value y_i is from our hypothesized value $y(x_i)$. Dividing by the square of the standard error gives us a dimensionless number which we can interpret as a measure of how well our hypothesis fits within the error bars of the data.

Another way of interpreting χ^2 is as the sum of squares of the ***normalized residuals***,

$$R_i = \frac{y_i - y(x_i)}{\alpha_i}. \tag{5.1.2}$$

The ***best-fit parameters*** are those that minimize χ^2 – those parameters create the hypothesis which is the "closest fit" to the full set of data points.

For some simple models we can analytically perform the minimization of χ^2 to create closed-form formulas for the best-fit parameters. We will see a few cases below.

For more complicated models, computer-based fitting (using Python, for example!) can be used to minimize χ^2.

5.2 - Simple Least-Squares Linear Regression

If we are fitting the data to a line then we call the procedure a ***linear regression***.

If the errors and uncertainties $\{\alpha_i\}$ are all the same then we can perform a regression based on a ***simple least-squares approach***. The denominator can be pulled out of the sum from Eq. 5.1.1 to become a constant prefact and thus plays no role in minimizing χ^2.

The Linear Model:

Consider the ***linear model*** $y(x) = mx + c$. For a linear simple least-squares regression, we want to minimize

$$\chi^2 = \frac{1}{\alpha^2}\sum(y_i - mx_i - c)^2,$$

with respect to the parameters m and c. The uncertainty α is assumed constant for all data points in this simple least-squares fit. The results are:

Model:
$$y(x) = mx + c, \tag{5.2.1}$$

A useful combination:
$$\Delta = N\sum x_i^2 - (\sum x_i)^2, \tag{5.2.2}$$

Best-fit parameters:
$$m = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{\Delta} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2}, \tag{5.2.3a}$$

$$c = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{\Delta} = \frac{\overline{x^2}\cdot\bar{y} - \bar{x}\cdot\overline{xy}}{\overline{x^2} - \bar{x}^2} = \bar{y} - m\bar{x} \tag{5.2.3b}$$

Common uncertainty: $$\alpha_{CU} = \sqrt{\frac{1}{N-2}\sum(y_i - mx_i - c)^2}. \quad (5.2.4)$$

$$\alpha_m = \alpha_{CU}\sqrt{\frac{N}{\Delta}} = \frac{\alpha_{CU}}{\sqrt{N\sigma_x^2}}, \quad (5.2.5a)$$

Uncertainties in best-fit parameters:

$$\alpha_c = \alpha_{CU}\sqrt{\frac{\sum x_i^2}{\Delta}} = \alpha_{CU}\sqrt{\frac{\overline{x^2}}{N\sigma_x^2}} = \alpha_m\sqrt{\overline{x^2}}. \quad (5.2.5b)$$

In Eq. 5.1.4 we introduce the ***common uncertainty*** which is useful for the uncertainties in the best-fit powers. We can also interpet it as an approximation to the uncertainty in the y measurements assuming that our model is accurate.

The Direct Proportionality Model (Linear Hypothesis through the Origin):

Next, consider the direct proportionality model, which is the linear model constrained to pass through the origin. For a linear simple least-squares regression, we want to minimize

$$\chi^2 = \frac{1}{\alpha^2}\sum(y_i - mx_i)^2,$$

with respect to the parameter m.

Hypthesis: $$y(x) = mx. \quad (5.2.6)$$

Best-fit parameters: $$m = \frac{\overline{xy}}{\overline{x^2}}. \quad (5.2.7)$$

Common uncertainty: $$\alpha_{CU} = \sqrt{\frac{1}{N-1}\sum(y_i - mx_i)^2}. \quad (5.2.8)$$

Uncertainties in best-fit parameters: $$\alpha_m = \alpha_{CU}\sqrt{\frac{1}{\sum x_i^2}} = \frac{\alpha_{CU}}{\sqrt{N\overline{x^2}}}. \quad (5.2.9)$$

5.3 - Weighted Least-Squares Approach[3]

We use a ***weighted least-squares approach*** when we have unequal errors and uncertainty in our data points. We want data points with low uncertainty to "matter more" than data points with high uncertainty so we attach a ***weight*** to each data point,

$$w_i = \frac{1}{\alpha_i^2}. \quad (5.3.1)$$

The Linear Hypothesis:

[3] More information on a weighted least-squares approach can be found in Hughes and Hase, *Measurements and Their Uncertainties*, Section 6.3 or Bevington, *Data Reduction and Error Analysis*, Section 6.3.

Consider the ***linear model*** $y(x) = mx + c$. For a weighted simple least-squares regression, we want to minimize

$$\chi^2 = \sum w_i (y_i - mx_i - c)^2,$$

with respect to the parameters m and c. The results are:

Hypthesis: $$y(x) = mx + b. \tag{5.3.2}$$

A useful combination: $$\Delta' = \sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2, \tag{5.3.3}$$

Best-fit parameters:

$$m = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\Delta'}, \tag{5.3.4a}$$

$$c = \frac{\sum w_i x_i^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i}{\Delta'} = \frac{\sum w_i y_i - m \sum w_i x_i}{\sum w_i}. \tag{5.3.4b}$$

Uncertainties in best-fit parameters:

$$\alpha_m = \sqrt{\frac{\sum w_i}{\Delta'}}, \tag{5.3.5a}$$

$$\alpha_c = \sqrt{\frac{\sum w_i x_i^2}{\Delta'}}. \tag{5.3.5b}$$

The Direct Proportionality Hypothesis (Linear Hypothesis through the Origin), $y(x) = mx$:

Next, consider the direct proportionality model, which is the linear model constrained to pass through the origin. For a linear simple least-squares regression, we want to minimize

$$\chi^2 = \sum w_i (y_i - mx_i)^2,$$

with respect to the parameter m.

Hypthesis: $$y(x) = mx. \tag{5.3.6}$$

Best-fit parameters: $$m = \frac{\sum w_i x_i y_i}{\sum w_i x_i^2}. \tag{5.3.7}$$

Uncertainties in best-fit parameters: $$\alpha_m = \sqrt{\frac{1}{N-1} \frac{\sum (y_i - mx_i)^2}{\sum x_i^2}}. \tag{5.3.8}$$

Uncertainties in the Independent Variable:

Suppose we have uncertainties $\alpha_{x,i}$ and $\alpha_{y,i}$ in our independent and dependent variables, respectively. The first thing we need to do is remove the uncertainty in x. We do this by performing a *simple* least-squares linear regression to find a best-fit slope m_{simple}. Then we exchange the uncertainty in x for additional uncertainty in y,

$$\alpha_{\text{equiv},i} = \sqrt{\alpha_{y,i}^2 + m_{\text{simple}}^2 \cdot \alpha_{x,i}^2}. \tag{5.3.9}$$

5.4 - Other Hypotheses

Linearizing a Non-Linear Relationship

For other functional relationships we can try to "linearize" the problem. For example, consider a power-law hypothesis, $y = Ax^n$, where A and n are our two parameters. To linearlize the problem we define two new variables $w \equiv \ln x$ and $z \equiv \ln y$. Taking the logarithm of both sides of $y = Ax^n$ gives a hypothesis $z = nw + (\ln A)$, which is in the form of a linear relationship, with $\{n, \ln A\}$ serving as parameters $\{m, c\}$. Take care to propagate uncertainties in such a case. If your uncertainties in y were all comparable they in general *won't* be for z and a weighted least-squares approach may be called for.

Non-Linear Relationships which Can't be Linearized

Sometimes you might get a particularly involved fuction that you are trying to fit to that can't be easily linearized. We can still perform a fit by doing a simple or weighted least-squares fit! We just don't necessarily have a closed-form solution to the best-fit parameters as we do with the linear hypotheses. In this case, we can run a Python script to find the parameters that minimize χ^2. Note that all of the normal caveats that come with root-finding algorithms come into play here. For example, you may find a *local* minimum but not the *global* minimum.

Section 6: Testing a Fit

Reference: Hughes and Hase, Chapter 8

6.1 - Agreement Tests

When you are directly comparing two values y and z, you should run an ***agreement test***. We first define the ***discrepancy*** $\epsilon \equiv |y - z|$. Since x and y will typically carry uncertainty, the discrepancy itself will have an uncertainty α_ϵ, which we determine using Eq. 4.7 for propagating errors. If we expect that x and y should be equal, then an experiment should result in $\epsilon < \alpha_\epsilon$ roughly 68% of the time and $\epsilon < 2\delta\epsilon$ roughly 95% of the time. We have to choose a cutoff for when we can reasonably conclude that x and y agree based on our discrepancy test and, as in 5BL, we will choose the $2\delta\epsilon$ criterion. Our agreement test value is therefore

Agreement Test: $$\frac{|y - z|}{2\alpha_\epsilon} = \frac{|y - z|}{2\sqrt{\alpha_y^2 + \alpha_z^2}}. \qquad (6.1.1)$$

We claim agreement when the agreement test value is less than 1.

- Example: Suppose we are using a least-squares approach to compute the best-fit value for the acceleration due to gravity and find $g = 9.72 \pm 0.05$. The accepted value at our latitude is $g_{acc} = 9.80$, which comes with an uncertainty $\alpha_{acc} = 0.01$. The discrepancy is $\epsilon = 0.08$ with uncertainty $\alpha_\epsilon = 0.05$. The agreement test value is therefore 0.8, which is less than 1 so we can claim our result is in agreement with the accepted value. Note that our result fails the more stringent requirement of $\epsilon < 1 \cdot \alpha_\epsilon$.

6.2 - "Goodness of Fit"

Our random errors and uncertainties for data points are meant to represent a 68% confidence interval. That is, in the absense of a systematic error or offset, the data associated with a given data point should fall within the error bars 68% of the time. This gives us a criterion in which we can judge whether a given fit is "good" or not.

If at least 2/3 of the error bars intersect the best-fit curve, then we consider the regression a good fit to the data.

In practice, once you have performed a regression, you should calculate the ***normalized residuals*** associated with your fit as defined in Eq. 5.1.2.

You then create a scatter plot of the residuals with the associated error bars. This scatter plot will provide a nice visual cue as to whether your fit is good or not (with more than 2/3 of the error bars intersecting the x-axis for a good fit) and if there is an extra effect or trend that your fit hasn't captured (for example, if the first half of your data is all below the axis while the second half is all above).

In this lab, we will often start by performing a simple least-squares fit. If this fit does not meet our goodness of fit criterion, we then move on to try a weighted least squares fit if appropriate.

6.3 - The Coefficient of Determination

There are many different measures of how "good" a fit matches the data. The ***coefficient of determination*** r^2, also called the "r-squared value," is a measure of how much of the variance in the dependent variable y is *explained* by the fit model due to the variance in the independent variable x. In other words,

r^2 is a test of whether data falls onto a given line in a "reasonable way."

Coefficient of Determination:
$$r^2 = 1 - \frac{\sum\left(y_i - y(x_i)\right)^2}{\sum(y_i - \bar{y})^2}. \tag{6.3.1}$$

If a simple least-squares linear regression is used to create the fit model then r^2 will always lie between 0 and 1, with low values indicating a particularly poor fit and high values a particularly good fit. Note, however, that r^2 can go outside of these bounds if a different model. In particular, in a *weighted* least-squares linear regression, all sums in Eq. 5.4.1 should be replaced by *weighted* sums in order for r^2 to have the same interpretation.

A flaw in the use of r^2 is that the value can be pushed arbitrarily close to 1 with the addition of more independent variables. Therefore we define the ***adjusted coefficient of determination*** $\bar{r}^2$, also called the "r-bar-squared value,"

Adjusted Coefficient of Determination:
$$\bar{r}^2 = r^2 - \frac{p}{N - p - 1}\,(1 - r^2). \tag{6.3.2}$$

The p in this formula is the number of independent variables (p = 1 in all of the regressions considered in this summary document). Note that there are a *lot* of subtleties in the interpretation of the coefficient of determination.[4]

The coefficient of determination is ***not appropriate*** for comparing predicted values to observed values (you would use an agreement test for that!).

Also note that our calculation of r^2 doesn't incorporate or tell us anything about errors so is most appropriate for a ***simple least-squares fit***.

6.4 - Chi-Squared

[4] For another good summary of the interpretation and limitations of r^2, see http://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit.

Recall the value ***chi-squared*** χ^2 defined in Eq. 5.1.1, which was the sum of squares of the normalized residuals. If our data points all lie within the naturally expected window of the fit curve then each normalized residual in the sum is roughly one or lower. Data points that lie outside the expected variation will contribute terms greater than one.

To adjust for the number of data points we create the ***modified*** or ***reduced chi-squared*** value,

Reduced Chi-Squared: $$\tilde{\chi}^2 = \chi^2/\nu. \qquad (6.4.2)$$

The quantity ν in Eq. 6.4.4 is the number of ***degrees of freedom*** for the system, defined as the number of data points minus the number of parameters in your fit. For example, the linear hypothesis fits two parameters so $\nu = N - 2$ and the direct proportionality hypothesis fits one parameter so $\nu = N - 1$.

For this class, we are more often looking to test whether our *data* is appropriate based on a given model rather than whether a model is a good fit to our data. It is a subtle point but an important one. In this lab class we are really developing our laboratory skills, so we need to hone our data-acquisition methods, using well-established theory to check our progress.

Given good data with well-understood and constrained errors and an appropriate model, $\tilde{\chi}^2$ should be close to one. There are many reasons why $\tilde{\chi}^2$ may be significantly greater (or less than!) 1, however.

- If $\tilde{\chi}^2 > 1$ then the data and/or model is falling outside the expected uncertainty range. The larger $\tilde{\chi}^2$ is, the less likely the discrepancy is due to random statistical variations. Therefore, if $\tilde{\chi}^2 \gg 1$ then our ***results are suspect***. There are a few possibility for why this occurs.
    - Your hypothesis is incorrect.
    - Your *uncertainties* are incorrect. You may have incorrectly evaluated the uncertainties or made invalid assumptions about them. [*This is the more likely scenario for the 5-series labs!*]
- If $\tilde{\chi}^2 < 1$ then the model is falling within the expected uncertainty range. If $\tilde{\chi}^2 \ll 1$ then our results are ***also suspect*** since it indicates that the actual variation of the data is not as large as a normal distribution based on your uncertainty calculations have suggested! This suggests that you have underestimated the errors.

We may ask what "close to 1" means for evaluating $\tilde{\chi}^2$. The answer depends on a number of factors including the numbder of degrees of freedom ν; the more degrees of freedom you are considering the closer you need $\tilde{\chi}^2$ to be to 1. Hughes and Hase Section 8.4 addresses this. The general guidelines suggested by them are:

- *If* $\tilde{\chi}^2 \ll 1$, *check your calcuations for the uncertainties in the measurements.*
- *The hypothesis or data is questioned if*:
    - $\tilde{\chi}^2 > 2$ *for* $\nu \approx 10$.
    - $\tilde{\chi}^2 > 1.5$ *for the approximate range* $50 < \nu < 100$.

◊