UC Berkeley - Physics 5CL

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. Your author is no exception, unfortunately. I will, however, tell you what my notational conventions mean! This mainly applies to the statistical analyses that we will run. A supplement on data analysis using my notation is available.

- When referencing an arbitrary individual data point I will use a lowercase subscript (usually an i) to indicate the index. Example: The i^{th} measurement of a position variable x is written x_i.
- Curly braces indicate the set of measurements. Example: The set of position measurements is written $\{x_\mathrm{i}\}$.
- The mean of a variable is indicated with triangular brackets. Example: The mean of the position data $\{x_i\}$ is written $\langle x\rangle$.
  - Another commonly used notation is to use an overbar, $\bar{x}$, though I will not be using this notation in these notes. The reason I use the bracket notation is that it is easier (for me at least) to distinguish $\langle x^2\rangle$ from $\langle x\rangle^2$ than it is to distinguish $\overline{x^2}$ from $\bar{x}^2$. Similarly, it is easier for me to distinguish $\langle xy\rangle$ from $\langle x\rangle\langle y\rangle$ that it is to distinguish $\overline{xy}$ from $\bar{x}\bar{y}$.
- Individual sources of error or uncertainty will be labeled with capital deltas in front of the symbol and a subscript indicating the source of error. Example: The reading uncertainty for position measurements x is written Δx_{read}. This error applies for all data points x_i so subscript may be omitted. However, the observational uncertainty differs from measurement to measurement so I will write observational uncertainties as $\Delta x_{\mathrm{obs},i}$.
- The total uncertainty in a quantity will be labeled with a lowercase delta in front of the symbol. Example: The total uncertainty in a position measurement x_i is written δx_i.
- Standard deviations will be written as a lowercase sigma with a subscript indicating the variable. Example: The standard deviation of a set of position measurements $\{x_i\}$ is written σ_x.
- Covariances will be written as a lowercase sigma with two subscripts, indicating the two variables under consideration. Example: The covariance of a data set $\{x_i, y_i\}$ is written σ_{xy}. Note that with this notation the *variance* of the single variable x is σ_{xx} and thus the standard deviation is $\sigma_x \equiv \sqrt{\sigma_{xx}}$.
- Best-fit parameter values will be written with a hat. Example: If we are fitting data $\{x_i, y_i\}$ with a linear regression hypothesis $y = mx$ then the best-fit value for the slope is written $\hat{m}$.

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

Mean:

$$\langle y\rangle = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{1.1}$$

Deviation from the Mean:

$$\varepsilon_i = y_i - \langle y\rangle. \tag{1.2}$$

Variance:

$$\sigma_{yy} = \langle \varepsilon^2\rangle = \frac{1}{N}\sum_{i=1}^{N}(y_i - \langle y\rangle)^2 = \langle y^2\rangle - \langle y\rangle^2. \tag{1.3}$$

Standard Deviation (Population):

$$\sigma_y = \sqrt{\sigma_{yy}} = \sqrt{\langle y^2\rangle - \langle y\rangle^2}. \tag{1.4a}$$

Standard Deviation (Sample): $$\sigma_y = \sqrt{\frac{N}{N-1}}\,\sigma_y = \sqrt{\frac{1}{N-1}\sum \varepsilon_i^2}. \tag{1.4b}$$

Standard Error: $$\sigma_{\langle y\rangle} = \sigma_y/\sqrt{N}. \tag{1.5}$$

The difference between population 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.

The standard deviation represents the uncertainty of a *single* measurement and the standard error represents the uncertainty in the mean of *multiple* measurements.

Example[1]: 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 $k = \langle k\rangle \pm \sigma_{\langle k\rangle}$. 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 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.

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

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

Coefficient of Linear Correlation: $$r_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}. \tag{2.2}$$

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 r_{xy} will always lie between -1 and 1. The closer $|r_{xy}|$ is to 1 the stronger the linear relationship is between x and y. The sign of r_{xy} gives the sign of the slope.

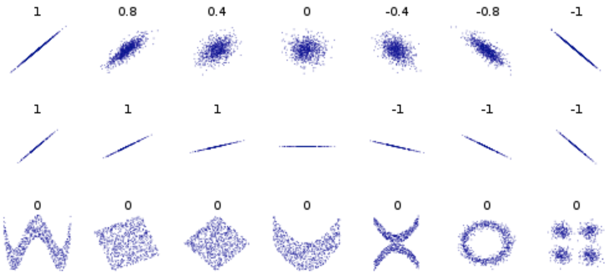

Figure 1: Various data sets and their correlations.[2]

[1] This example is presented in Taylor, *An Introduction to Error Analysis*, Chapter 4.
[2] http://en.wikipedia.org/wiki/Correlation_and_dependence

SECTION 3: SOURCES OF ERROR AND UNCERTAINTY

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 uncertainty in any given measurement. An individual source of uncertainty will be labeled with a capital delta and a subscript for the source.

- ***Reading Uncertainty*** - An uncertainty that is due to the *finite resolution* of our instruments. The reading uncertainty Δy_{read} is plus-or-minus one-half the resolution of the measurement.
- ***Random Uncertainty*** - An uncertainty that is due to random fluctuations in our instruments or readings (very common with digital instruments). In a random fluctuation, your readings will bounce around, but will typically be bounded by some high reading and some low reading. The central value of this measurement y_{cent} is the average of the high and low readings and the random uncertainty Δy_{rand} is half the difference between the high and low value. The expression "$y_{\text{cent}} \pm \Delta y_{\text{rand}}$" thus contains the entire interval of observed values and is the result that should be reported.
- ***Observational Uncertainty*** - Uncertainties due to judgment calls made during an observation. You can estimate the size of your observational uncertainty by average the bounds of your judgement to create a central value and take half the difference between the bounds to create the error.
- ***Counting Uncertainty*** - When dealing with occurrence counts of a random process (such as the decay of a radioisotope) there is an inherent statistical counting uncertainty that grows as the square root of your count. That is, given a count of N the counting uncertainty is $\Delta N_{\text{count}} = \sqrt{N}$.
- ***Instrumental Error*** - Perhaps the instrument was poorly calibrated (a thermometer reading the freezing point of water in standard atmospheric pressure as 2°C and the boiling point as 102°C).
- ***Environmental Error*** - Perhaps there was a power surge or a mild tremor that causing readings to go awry.
- ***Theoretical Error*** - Due to the models or approximations used in your assumptions. For example, modeling a pendulum as a perfect simple harmonic oscillator ignores the deviations that occur when the amplitude of oscillation is large. Such deviations at large amplitudes contribute to the theoretical error.
- ***Blunders*** - There is no fixing this, really. These are just mistakes! Maybe you recorded the wrong value or fell asleep and missed some readings or used the wrong units.

The Total Uncertainty

The total uncertainty in a measurement is written with a lowercase delta, δy. A measurement should always be recorded with the central value along with the uncertainty. If the central value is y and the total uncertainty is δy then the reported measurement would be written "$y \pm \delta y$", with units placed at the end of the expression. Example: If I measure a position of 31.0 cm with a ruler (with a resolution of 2 mm) and find and the reading error were the only source of error or uncertainty I would report the result as 31.0 ± 0.2 mm.

We assume that all sources of error and uncertainty discussed above are independent and therefore the total uncertainty is found by adding the individual errors in quadrature,

$$\delta y = \sqrt{\Delta y_1^2 + \Delta y_2^2 + \cdots}. \tag{3.1}$$

For a quick estimate of the total uncertainty, note that δy will always be larger than the largest individual source of uncertainty (call this Δy_1) and smaller than the sum of all the sources of uncertainty, $\Delta y_1 \leq \delta y \leq \sum \Delta y_i$.

Relevant and Irrelevant Sources of Error and Uncertainty

There will *always* be *many* sources of error and uncertainty for any measurement or calculation. When computing the total uncertainty we can safely ignore some sources of error as long as they don't appreciably change the calculation of the total uncertainty. (Your tolerance for what is an appreciable change is of course subjective, though). The relevance of any individual source of uncertainty for a measurement is based on the *largest* source of uncertainty for that measurement. For example, consider the following table showing possible reading and observational uncertainties for a position measurement of $y = 2.050$ m:

Trial	y (m)	Δy_{read} (m)	Δy_{obs} (m)	δy (m)	Deviation from largest uncertainty.
1	2.050	0.010	**0.200**	0.20025	0.13%
2	2.050	**0.010**	**0.020**	0.02236	11.80%
3	2.050	**0.010**	0.002	0.01020	2.00%

If a given source of uncertainty is roughly an order of magnitude smaller than the largest source of uncertainty then its effects get drowned out, as seen in Trial 1 - where the total uncertainty is only 0.13% larger than the observational uncertainty - and in Trial 3 - where the total uncertainty is only 2.00% larger than the reading uncertainty. When a given source of uncertainty is of the same order of magnitude as the largest source of uncertainty as in Trial 2 then we need to take both sources of uncertainty into account.

SECTION 4: PROPAGATION OF UNCERTAINTY[3]

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_i, the derived value of q for that data point is $q_i = q(x_i)$. Given an uncertainty δx_i in the measurement of x_i the propagated uncertainty for q_i is given by

$$\delta q_i = |q'(x_i)|\delta x_i, \tag{4.1}$$

where $q'(x) \equiv dq/dx$. Some of the more commonly occuring examples are given below.

Multiplication by a Constant:	$q(x) = cx$	$\delta q = \lvert c\rvert\delta x.$	(4.2)
Power:	$q(x) = x^n$	$\dfrac{\delta q}{\lvert q\rvert} = \lvert n\rvert\dfrac{\delta x}{\lvert x\rvert}.$	(4.3)
Exponential:	$q(x) = e^x$	$\dfrac{\delta q}{q} = \delta x.$	(4.4)
Logarithm:	$q(x) = \ln x$	$\delta q = \dfrac{\delta x}{x}$	(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, $r_{xy} = 0$. 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 $q_i = q(x_i,y_i)$. Given uncertainties δx_i and δy_i, the the propagated uncertainty for q_i is given by

[3] For a more thorough discussion see Taylor, *An Introduction to Error Analysis*, Chapter 3.

.

$$\delta q_i = \sqrt{\left(\frac{\partial q}{\partial x}\delta x_i\right)^2 + \left(\frac{\partial q}{\partial y}\delta y_i\right)^2}, \tag{4.6}$$

where the partial derivatives are understood to be evaluated at $\{x_i,y_i\}$. 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: $q(x,y) = x + y$ $\qquad \delta q = \sqrt{(\delta x)^2 + (\delta y)^2}.$ (4.7)

Product or Quotient: $q(x,y) = xy \text{ or } \frac{x}{y}$ $\qquad \frac{\delta q}{|q|} = \sqrt{\left(\frac{\delta x}{x}\right)^2 + \left(\frac{\delta y}{y}\right)^2}.$ (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 $y = ax$ (the correlation is $r_{xy} = \pm 1$) in the above cases.

SECTION 5: LINEAR REGRESSION[4]

Suppose we have a data set of two variables, $\{x_i,y_i\}$. Suppose we guess that the variable y depends linearly on the variable x. That is, we hypothesize a mathematical relationship $y(x) = mx + b$. This hypothesis comes with a number of undetermined parameters. Our goal is to determine which set of parameters $\{\hat{m}, \hat{b}\}$ best fit the data and we further want a measure of how well our data matches the hypothesized relationship with the best-fit parameters. If we are fitting the data to a line, as we are in this example, then we call the procedure a ***linear regression***.

5.1 - SIMPLE LEAST-SQUARES APPROACH

Consider a hypothesis $y(x;a_n)$, where $\{a_n\}$ are a set of parameters for the function $y(x)$. We define a function $Q(a_n)$ that is a cumulative measure of how far off our data points are from a hypothesis with parameters $\{a_n\}$,

$$Q(a_n) = \sum \big(y_i - y(x_i;a_n)\big)^2. \tag{5.1.1}$$

$Q(a_n)$ is the sum of the squares of the *residuals* - how far off each y_i value is from the predicted value $y(x_i;a_n)$. The best-fit values $\{\hat{a}_n\}$ are found by minimizing $Q(a_n)$ simultaneously with respect to all parameters $\{a_n\}$.

Below, a summary of linear regressions based on a simple least-squares approach for common hypotheses is given.

The Linear Hypothesis, $y(x) = mx+b$:

Hypthesis: $y(x;m,b) = mx + b.$ (5.1.2)

Best-fit parameters:

$$\hat{m} = \frac{\sigma_{xy}}{\sigma_{xx}} = \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\langle x^2\rangle - \langle x\rangle^2}, \tag{5.1.3a}$$

$$\hat{b} = \langle y\rangle - \hat{m}\langle x\rangle = \frac{\langle x^2\rangle\langle y\rangle - \langle x\rangle\langle xy\rangle}{\langle x^2\rangle - \langle x\rangle^2}. \tag{5.1.3b}$$

[4] For a more thorough discussion see Taylor, *An Introduction to Error Analysis*, Chapter 8.

.

Uncertainties in *y* based on fit:

$$\delta y = \sqrt{\tfrac{1}{N-2}\Sigma\left(y_i - y\left(x_i;\hat{m},\hat{b}\right)\right)^2}. \tag{5.1.4}$$

Uncertainties in best-fit parameters:

$$\delta\hat{m} = \frac{\delta y}{\sqrt{N\sigma_{xx}}}, \tag{5.1.5a}$$

$$\delta\hat{b} = \sqrt{\langle x^2\rangle}\delta\hat{m} = \delta y\sqrt{\frac{\langle x^2\rangle}{N\sigma_{xx}}}. \tag{5.1.5b}$$

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

Hypthesis:

$$y(x;m) = mx. \tag{5.1.6}$$

Best-fit parameters:

$$\hat{m} = \frac{\langle xy\rangle}{\langle x^2\rangle}. \tag{5.1.7}$$

Uncertainties in *y* based on fit:

$$\delta y = \sqrt{\tfrac{1}{N-1}\Sigma\left(y_i - y(x_i;\hat{m})\right)^2}. \tag{5.1.8}$$

Uncertainties in best-fit parameters:

$$\delta\hat{m} = \frac{\delta y}{\sqrt{N\langle x^2\rangle}}. \tag{5.1.9}$$

5.2 - Weighted Least-Squares Approach

The simple least-squares approach ignored the uncertainties in our data points. For a linear fit this is a valid approach *as long as each measured pair {x_i,y_i} has roughly the same uncertainties {δx,δy}*. When we have differing uncertainties we modify our approach. The first thing we need to do is *eliminate* the uncertainty in *x*. We do this by performing a simple least-squares linear regression to find a best-fit slope $\hat{m}_{\text{simple}}$. Then we exchange the uncertainty in *x* for additional uncertainty in *y*,

$$\delta y_{\text{equiv},i} = \sqrt{\delta y_i^2 + \left(\hat{m}_{\text{simple}}\cdot\delta x_i\right)^2}. \tag{5.2.1}$$

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,

$$w_i = \frac{1}{\left(\delta y_{\text{equiv},i}\right)^2}. \tag{5.2.2}$$

This weight gets attached to our *Q* from earlier,

$$Q(a_n) = \sum w_i\left(y_i - y(x_i;a_n)\right)^2 = \sum\left(\frac{y_i - y(x_i;a_n)}{\delta y_{\text{equiv},i}}\right)^2. \tag{5.2.3}$$

The Linear Hypothesis, *y*(*x*) = *mx*+*b*:

Hypthesis:

$$y(x;m,b) = mx + b. \tag{5.2.4}$$

Best-fit parameters:

$$\hat{m} = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2}, \tag{5.2.5a}$$

$$\hat{b} = \frac{\sum w_i x_i^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2} = \frac{\sum w_i y_i - \hat{m} \sum w_i x_i}{\sum w_i}. \tag{5.2.5b}$$

Uncertainties in best-fit parameters:

$$\delta\hat{m} = \sqrt{\frac{\sum w_i}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2},} \tag{5.2.6a}$$

$$\delta\hat{b} = \sqrt{\frac{\sum w_i x_i^2}{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2}.} \tag{5.2.6b}$$

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

Hypthesis:

$$y(x; m) = mx. \tag{5.2.7}$$

Best-fit parameters:

$$\hat{m} = \frac{\sum w_i x_i y_i}{\sum w_i x_i^2}. \tag{5.2.8}$$

Uncertainties in best-fit parameters:

$$\delta\hat{m} = \sqrt{\frac{1}{N-1} \frac{\sum (y_i - \hat{m} x_i)^2}{\sum x_i^2}}. \tag{5.2.9}$$

Comparisson with Unweighted Least-Squares:

If all of the weights are identical then the weighted least-squares formulas for Eqs. 5.1.2 through 5.1.9 are identical to those in Eqs. 5.2.4 through 5.2.9. In fact, we can make the formulas in Section 5.1 work for the weighted approach by replacing the mean with the *weighted mean*,

Weighted Mean:

$$\langle y \rangle = \frac{\sum w_i y_i}{\sum w_i}. \tag{5.2.10}$$

Of course we have to be careful to use the weighted means in Eqs. 1.3 and 2.1 when variances occur in the formulas.

5.3 - Other Hypotheses

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 $\{m, b\}$. 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.

5.4 - The Coefficient of Determination and the Reduced Chi-Squared Tests[5]

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.

Coefficient of Determination:
$$r^2 = 1 - \frac{\sum\left(y_i - y(x_i;\hat{a}_n)\right)^2}{\sum y_i^2 - \left(\sum y_i\right)^2}. \tag{5.4.1}$$

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:
$$\bar{r}^2 = r^2 - \frac{p}{N-p-1}\left(1 - r^2\right). \tag{5.4.2}$$

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.[6]

Another commonly used test of "goodness of fit" is the ***chi-squared value***,

Chi-Squared:
$$\chi^2 = \sum\left(\frac{y_i - y(x_i;\hat{a}_n)}{\delta y_i}\right)^2. \tag{5.4.3}$$

Note that the chi-squared value in Eq. 5.4.3 is identical to the Q in Eq. 5.2.3 used in the *weighted* least-squares approach. Each term in the sum is a ratio of the actual difference between a data point and the fit and a statistically expected standard variation of the dependent variable from the expected fit value. If our data points all lie within the naturally expected window of the fit curve then each term 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:
$$\tilde{\chi}^2 = \chi^2/d. \tag{5.4.4}$$

The quantity d in Eq. 5.4.4 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 $d = N$-2 and the direct proportionality hypothesis fits one parameter so $d = N$-1. A good fit has $\tilde{\chi}^2$ close to one and a bad fit has $\tilde{\chi}^2$ much greater than one.

◊

[5] For a more thorough discussion see Taylor, *An Introduction to Error Analysis*, Chapter 12.

[6] For a good summary of the interpretation 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.