EXPERIMENTAL DATA ANALYSIS: SELECTED TOPICS

OBJECTIVES: (1) To review some basic statistical concepts; (2) to understand basic data reduction principles, including propagation of error, and (3) to learn how to find the uncertainties of a mean, and the slope and intercept in a simple linear regression model.

ELEMENTARY STATISTICS REVIEW

We know that when repeated measurements are made of some quantity that we do not get the same result each time. We are taking **samples** from some underlying **population**. The **random variation** in our samples is often called **error,** but a better term is **uncertainty.** For example, the figure below is a **histogram** of measured periods of oscillation T of a simple pendulum, the length of which is one meter. We see that the variation in T follows a shape, commonly called a "bell curve" but which is known in statistics as the **normal distribution**.

This probability distribution (or, sometimes, "density") function, or **PDF**, is characterized by two parameters: **location** and **dispersion.** The location is estimated with the **mean** or average of the observations, while the dispersion is estimated with the sample **variance,** which is usually reported as its square root, the **standard deviation**. The location of a PDF is where its center is on the real number line, while the standard deviation (often referred to by its usual symbol, **sigma σ)** tells us how wide the PDF is.

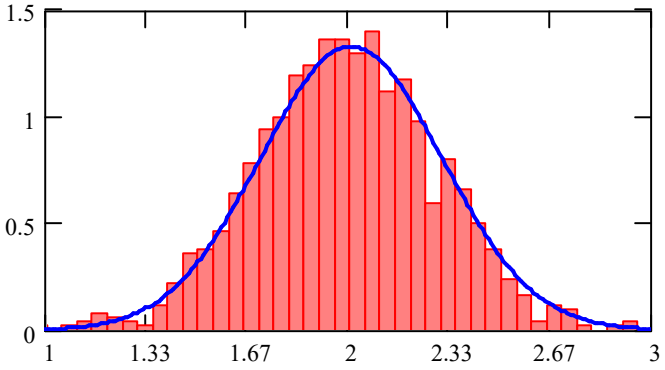

Figure 1. Histogram and gaussian PDF for observations of 1 meter pendulum period T.

In the figure is overlaid the normal (sometimes called **gaussian**) PDF corresponding to the observed mean and sigma for this data. The PDF is given by

$$\frac{d\,\Pr(x)}{dx} = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right] \tag{1}$$

and we can find the probability that the **random variable** x will take on certain values by integrating this PDF. As it happens this integral cannot be evaluated directly, and numerical methods are used. The ability to find these values on your calculators replaces books of tables that used to be needed.

ACCURACY AND PRECISION

When we report the results of experimental measurements, there are two primary concerns: **accuracy** and **precision.** These are sometimes used interchangeably, and this is incorrect. Accuracy is related to the location of the data (i.e, its mean), while precision relates to the dispersion (sigma).

In fact, there are various conflicting definitions for these terms, and we will not place much emphasis on them, but will instead focus on estimating uncertainties (below). However, since this is an introductory course, we should have some idea of what these terms usually mean.

Accuracy is assessed with respect to some known or assumed value. If the location of the data is "far" from this reference value we say that the experimental results are **inaccurate**. The word **bias** is sometimes used to reflect inaccuracy. In elementary lab exercises the "percent difference" is often used to assess the lab's results:

$$\% = 100\frac{(\text{mean} - \text{known})}{\text{known}}$$

This measure gives a very incomplete assessment of the results, and ignores the important role of the dispersion of the data, as we will see in a moment.

Precision is usually assessed with a measure called the **relative error** (also known as **coefficient of variation**), which is the ratio of the standard deviation to the mean (the symbol for which is μ)

$$RE = \frac{\sigma}{\mu}$$

The use of the term "error" here is perhaps unfortunate, but it is very common. *Relative* error is a good way to express the dispersion, since it should be related to the location of the data's PDF. That is, if the sigma was 10 units, the precision would be very poor if the location (mean) was, say, 20 units. On the other hand, if the location was at 2000 units, then a sigma of 10 represents much more precise measurements. This dispersion is assumed to be irreducible, and due only to random variations. It is true, however, that sloppy, careless lab practices can also increase the data's dispersion, another word for which is **scatter**.

Another issue in the lab is **systematic error**, where, to put it bluntly, something is just plain wrong. A measurement device might be miscalibrated or not operating correctly, or the students may have misunderstood how to make a measurement. Systematic errors are usually reflected in a bias (inaccuracy), if the lab practices are otherwise careful. Systematic error can also happen if the model we are using for the data, based on the physics, is incorrect or insufficiently detailed.

A baseball analogy may help to illustrate accuracy vs. precision. You are a manager, choosing from several center fielders, to decide who starts. The fielders go far out into straightaway center (lined up on home plate and second base) and make repeated throws into second base.

Fielder A consistently throws within arm's reach of the bag, so that the second baseman can keep his foot on the base and still catch the ball. Fielder A is both accurate and precise. The ball goes where it is supposed to go, and does so with a small dispersion.

Fielder B throws the ball consistently within arm's reach of the location where the shortstop usually stands. B is precise but inaccurate; small scatter or dispersion, but in the wrong place.

Fielder C scatters the ball all over the infield, anywhere from first to third base. On average, the ball is at second base, but any given throw will probably not be there. C is imprecise. We *could* say he is accurate, since *on average* the ball goes where it is supposed to go. But this doesn't make much sense, and you wouldn't put him in the game.

Thus, **precision is a necessary (but not sufficient) condition for accuracy**.

Consider some students whose lab procedure is sloppy, so that their measurement data is scattered badly. They could, however, just be lucky enough that the average of their data is near the accepted value. If the lab is assessed only on the "percent difference" criterion, it will appear that they did a good job on the lab. Clearly this is not fair to the other students who were more careful in their work.

Lab exercises should be assessed on both relative error and accuracy.

INSTRUMENTAL RESOLUTION

Suppose a graduated cylinder cannot be read any more closely than to the nearest 0.5 mL. This means that your volume measurements will be, say, 11.0 or 11.5 mL, but cannot be 11.24 mL, even if that is the actual, correct value. This **instrumental resolution** is going to affect your data, but it is neither a random error nor a systematic error, it is just a limitation of the measurements. Clearly we need to use instruments that can measure the quantities of interest to sufficient resolution for the purposes of the lab.

This is a part of **experimental design**, in the non-statistical sense. You would not even try to use the classroom wall clock to time a ball that rolls down a ramp in a second or two. You know that you need a stopwatch, and, perhaps, a computer with a timing gate device would be indicated. These kinds of issues must be considered when planning an experiment. *Statistical* experimental design, or analysis of variance, is a discipline in itself, and involves such issues as how many **replicates**, i.e., repeated measurements, to take, and at what settings of the independent variables.

DERIVED QUANTITIES: PROPAGATION OF ERROR (UNCERTAINTY)

So far we have dealt with direct-measurement data. Very often, however, we must do more than just take the average of a few measurements. We may need to use that mean in some derived formula, to get an estimate of the quantity of interest. For example, with the pendulum, we may be estimating the gravitational acceleration g, using observed period data. The starting model for this is, of course,

$$T = 2\pi\sqrt{\frac{L}{g}}$$

We refer to the quantities being **estimated** in this or any other model as **parameters**. Solving this expression for the parameter of interest leads to our estimation model

$$g = \frac{4\pi^2 L}{T^2}$$

The location (mean) of a parameter is sometimes called the **point estimate**, because that's what it is, a single point on the real number line. With this we need to quote its associated uncertainty ("sigma"). In more advanced work we would then proceed to use these numbers to create what is known as a **confidence interval**, and then perhaps conduct **hypothesis tests**.

For now, though, we have created a new random variable g, which is a function of the directly-observed variable T. We need to be able to describe at least the uncertainty ("sigma") of this derived quantity, and if possible, its PDF. The latter is often very difficult and the best approach, with modern computing capabilities, is to use **Monte Carlo simulation**. This is (unfortunately) beyond the scope of this course, so we will be content to find the sigma of the parameter estimate.

The way we do this is to invoke a classical and widely-used formula for the *approximate* variance of a function z of several random variables $x_1, x_2, x_3,$ etc. It looks complicated but is not usually too difficult to apply in practice.

$$\sigma_z^2 = \sum_i \sum_j \left(\frac{\partial z}{\partial x_i}\right)\left(\frac{\partial z}{\partial x_j}\right) \mathrm{cov}\left(x_i, x_j\right) \qquad (2)$$

The sums are taken over all combinations of the independent variables. The partial derivatives are to be evaluated at the mean value of each variable. The covariance of a variable with itself is its variance. This formula is widely used to give an approximate uncertainty in a variable that is a function of several random variables, but *it does not give the PDF of the variable*. A common mistake is to ignore the covariances of the several random variables; sometimes these are significant. Expressions derived from this formula will be provided for your labs, as needed.

To illustrate what happens when we have a function of a random variable, Figure 2 shows a histogram of g estimates, based on the T samples as in Figure 1. In this case, each T observed was used to find a g, and these are histogrammed. The PDF shown is a gaussian, with a sigma found using Eq(2):

$$\sigma_g = \frac{8\pi^2 L}{\bar{T}^3}\sigma_T \tag{3}$$

We see in the figure that the gaussian PDF does not describe this data particularly well- the data is skewed. This is because we have in effect transformed the gaussian data (the *T*'s) into some new random variable (the *g*'s), and we don't know what PDF this new variable follows.

In Figure 3 we have a *g* histogram in which each *g* was estimated using the *mean* of a bunch (say, *N*) of *T* observations. This is the way we would usually do the analysis, and it results in a sigma that is Eq(3) divided by the square root of *N* (more on this below). This time the PDF appears to be more symmetric, and matches the gaussian shape very well. As you might imagine, this is not an accident, and is a consequence of a statistical theory.

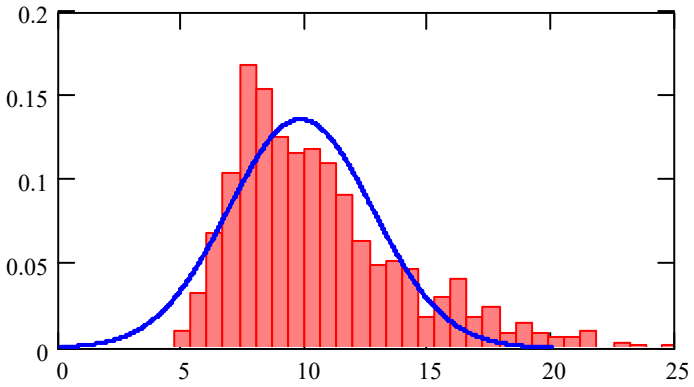

Figure 2. Histogram and gaussian PDF of g estimates using individual T observations.

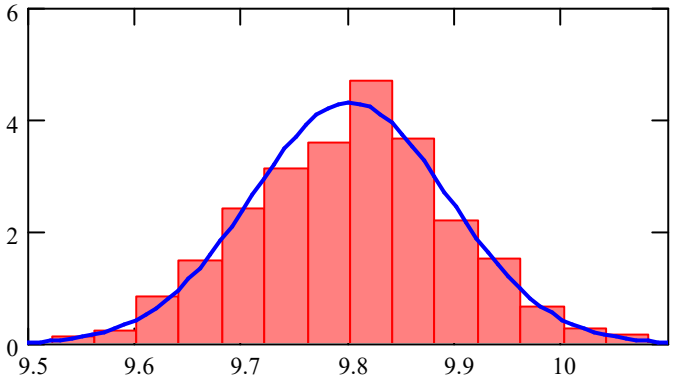

*Figure 3. Histogram and gaussian PDF using the **mean** of T observations for each g estimate.*

Another situation that leads to derived random variables is the need to use some form of **regression analysis** (usually called "curve fitting") in order to get where we need to go. In introductory labs, the regression is often just a simple linear function, or perhaps a quadratic. It should be apparent that the calculations we do in estimating the parameters in these regression functions will create new random variables, and these will have a location and dispersion that we'll need to estimate.

The formulas for this are presented below; anticipating these, next we have histograms of the slopes and intercepts estimated from repeated regressions, on randomly-sampled data at each *x*-value. On the histograms are graphed the expected PDFs for these parameters, using the uncertainty calculations in the next section. We see that they agree very nicely.

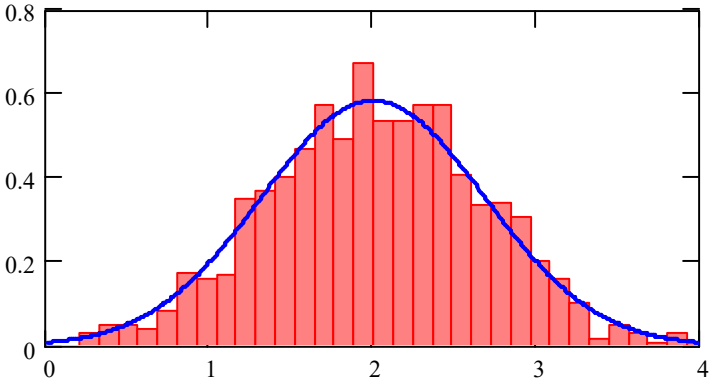

Figure 4. Histogram and PDF for regression intercepts; true value = 2.

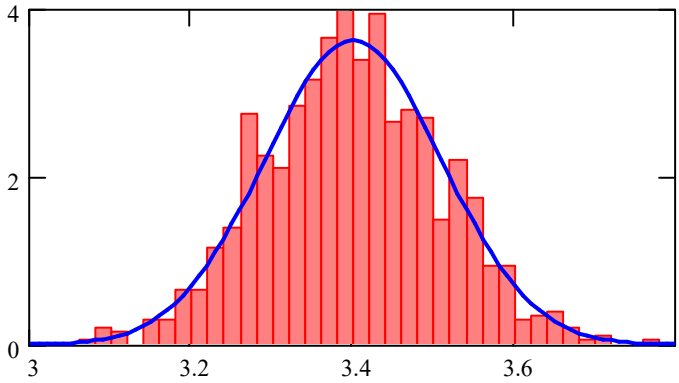

Figure 5. Histogram and PDF for regression slopes; true value 3.4

A final comment on regression; sometimes it is appropriate to **transform** the independent variable, for example by using its square root. This can change the relation between the independent and **response** (dependent) variables so that it is linear, and can then be well-represented by a straight line graph, which might not have been the case using the original independent-variable data. We will pursue this in a lab, later.

COMMENTS ON "SIG FIGS"

You have probably learned about **sig-figs** in an earlier course. We will not use them here. "Significant figures" refers to a crude form of error representation based on a rounding process. The general concept, that the chain is only as strong as the weakest link, is not a problem. But the process of propagation of error is a far better way to deal with this. So, we will quote our experimental results using a point estimate plus or minus its uncertainty.

COMPUTATIONAL FORMULAS

From here we will present, without proof, several useful formulas for application to our labs. The idea is to show the formulas for reference, but we will implement these with your calculators (some of this is built-in already). These formulas are all cases of simple linear regression (SLR).

CASE I. SIMPLE LINEAR REGRESSION $y = b_0 + b_1 x$ SLOPE & INTERCEPT ≠ 0

This is a very common model, with both the slope and intercept to be estimated. The TI calculators will do the point estimation of the slope and intercept directly, but they do not provide the parameter uncertainties. Below are the calculation formulas for this regression, where *n* is the number of data points. See any regression text for details, for example, Draper and Smith, *Applied Regression Analysis*, 2nd Ed., Chapter 1. In modern statistical analysis, this estimation is done using matrix methods rather than these formulas.

Point Estimates, Slope (b_1) and Intercept (b_0)

$$\bar{x} = \frac{1}{n}\sum x_i \qquad \bar{y} = \frac{1}{n}\sum y_i \qquad b_1 = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sum x_i^2 - n\bar{x}^2} \qquad b_0 = \bar{y} - b_1\bar{x} \tag{4}$$

Residual Mean Square Error

This **statistic** is essential to the analysis, and it measures the random variation of the data around the regression line. (A "statistic" is a number calculated from the data.) We should quote this value when doing lab analyses. Note that the difference ε between the data and the line is called the **residual**.

$$\hat{y}_i = b_0 + b_1 x_i \qquad \varepsilon_i = y_i - \hat{y}_i \qquad MSe = \frac{\sum \varepsilon_i^2}{n-2} \qquad s = \sqrt{MSe} \tag{5}$$

Uncertainties

These statistics express the variability in the estimated slope and intercept.

$$\sigma(b_0) = s\sqrt{\frac{\sum x_i^2}{n\left(\sum x_i^2 - n\bar{x}^2\right)}} \qquad \sigma(b_1) = \frac{s}{\sqrt{\sum x_i^2 - n\bar{x}^2}} \tag{6}$$

CASE II. MEAN OF SEVERAL OBSERVATIONS $y = b_0$ SLOPE = 0

This is a very common case, with a simple estimator. What is not so commonly known is that (1) this can be considered a special case of SLR, and (2) the uncertainty in the estimate of the mean, which is called the **standard error**, depends on the number of samples used in the average. The larger this number, the more precise is the estimate of the mean. The statistic "*s*" is an estimator of the standard deviation of the *population* from which the samples are taken. The **standard deviation of the mean** (standard error) is often confused with the population standard deviation. The former depends on the sample size, the latter does not.

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum x_i \qquad \hat{\sigma} = s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum x_i^2 - n\bar{x}^2}{n-1}} \qquad \hat{\sigma}_\mu = \frac{s}{\sqrt{n}} \tag{7}$$

CASE III. ZERO-INTERCEPT REGRESSION $y = b_1 x$ INTERCEPT = 0

Sometimes we know from the physics of the problem, or just from common sense, that when the independent variable is zero, so is the response variable, This leads to what is known as a "zero intercept" (ZI) linear regression. It can be shown that, if this is in fact the case, leaving the intercept in the regression model, even though we know that it isn't necessary, will result in a less precise estimate of the slope. The intercept will test statistically to be zero, which in itself doesn't hurt anything, but the slope estimate will lose precision. Thus, if we can say that the intercept is zero, we should use this model to get the most precise (i.e., minimum variance) estimate of the slope.

Point Estimate, Slope (b_1)

$$b_1 = \frac{\sum x_i y_i}{\sum x_i^2} \tag{8}$$

Uncertainty

$$s = \sqrt{\frac{\sum y_i^2 - \frac{\left(\sum x_i y_i\right)^2}{\sum x_i^2}}{n-1}} \qquad \sigma(b_1) = \frac{s}{\sqrt{\sum x_i^2}} \tag{9}$$

CALCULATOR IMPLEMENTATION

Many of the calculations given above are already implemented in your calculators. While these operations can be carried out directly on list contents, it is simpler to just use the regression functions available. These functions will define some of the quantities (sums) that we need to complete the calculations, and we will need to write short programs for this.

REGRESSION PROCEDURE: CASE I

Enter the independent-variable data into a list (L1).

Enter the dependent-variable data into a second list (L2).

STAT / CALC / 4 (LinReg) L1, L2, Y1 ENTER

Write down the coefficients: a = slope = b1 b = intercept = b0

STAT PLOT	Turn on one of them, plot the L1, L2 data.
ZOOM STAT	Should draw the regression line with the data.
PRGM	Find your regression uncertainty program, run it, write down the results; this would be the "s" and the "sigmas" for the two parameters.

Here is the code for the regression uncertainty calculations. These variables are obtained using VARS / 5 (Statistics) and various entries under the first two columns in that screen. Be aware that this program will operate on the stat variables as they exist at the time of execution. If you've done some other stat work since the regression above, these values will be incorrect. Thus, *run this program immediately after doing the basic regression*.

```
prgmUNCERT
ClrHome
√ ( sum( LRESID^2 ) / (n-2) ) → S          LRESID is found under 2nd LIST / NAMES
Disp "S"
Disp Ans
s * √ ( 1 / n + xbar^2 / ( Sx^2 * (n-1) ) )
Disp "SIG B0"
Disp Ans                                   intercept uncert
s / √ ( Sx^2 * (n-1) )
Disp "SIG B1"
Disp Ans                                   slope uncert
```

REGRESSION PROCEDURE: CASE II

To make this easier to use, first define a list with the name "MEAN" and *always use that list for the data*. Once you have defined this list, its name will appear under 2nd LIST / NAMES so you don't have to type it in.

```
prgmMEANSIG
ClrHome
mean(LMEAN)                    2nd LIST / MATH / 3
Disp "MEAN"
Disp Ans
stdDev(LMEAN)                  2nd LIST / MATH / 7
Disp "POP SIG"
Disp Ans
Ans / √ ( dim(LMEAN) )         2nd LIST / OPS / 3
Disp "MEAN SIG"
Disp Ans
```

REGRESSION PROCEDURE: CASE III

Here is the code for finding the zero-intercept regression and its uncertainty. *This must be executed right after the SLR regression*, in order for the stat variables (sums, etc.) to have the correct values.

```
prgmZI
ClrHome
Σxy / Σx^2
Disp "B1"
Disp Ans                              Slope estimate
√ ( (Σy^2 – (Σxy)^2 / Σx^2 ) / (n-1) )
Disp "S"
Disp Ans                               "s"
Ans * √ ( 1 / Σx^2 )
Disp "SIG B1"
Disp Ans                              Slope uncertainty
```

EXAMPLES

Case I

x = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

y = 7.06, 9.32, 12.03, 14.05, 15.81, 20.04, 22.38, 25.56, 29.69, 30.81

Simple Linear	*Zero-intercept*
$b_0 = 3.57 \pm 0.56$	
$b_1 = 2.75 \pm 0.09$	$b_1 = 3.26 \pm 0.10$
$s = 0.82$	$s = 1.90$

Note that the *s* for the zero-intercept is more than twice that for the regular regression, and that there was no real improvement in the precision of the slope estimate. This indicates that the ZI model is not appropriate for this data. You could also graph the ZI model with the data to see this.

Case II

This is just to check that your code for MEANSIG is correct.

x = 2.791, -0.451, 0.319, 1.258, -1.906, 0.858, -0.0744, 1.583, -0.906, 1.401

x-bar = 0.487 s = 1.373 std err = 0.434 RE ~ 90%

Since the *RE* is so large, we could safely conclude that the true mean of this data is zero. (In fact, it was.)

Case III

This is a ZI case, where that model is appropriate. We will do the SLR anyway, to show the difference.

x = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10

y = 1.6, 4.2, 5.8, 8.5, 11.0, 11.5, 13.8, 15.9, 20.2, 21.0

Simple Linear	*Zero-intercept*
$b_0 = -0.43 \pm 0.53$	
$b_1 = 2.14 \pm 0.085$	$b_1 = 2.08 \pm 0.04$
$s = 0.77$	$s = 0.76$

Observe that the *RE* of the intercept in the SLR is greater than 100%. This is an indication that this parameter does not belong in the model. Generally any *RE* greater than around 50% is a hint that that parameter is not necessary. Also note that the *s* is essentially the same, but that *the uncertainty in the slope has been cut in half for the ZI case*. This improvement in precision is why we would like to use ZI when it is appropriate.