3 Data Processing

3.1 Introduction

The staff and students comprising the research groups in the Department of Physics and Electronics at Rhodes are best described as empirical or experimental physicists. This practical is designed to introduce you to the computational and data analysis tools used most frequently by these groups. Many of the techniques contained in this practical are also relevant to PHY302 experiments that you will do subsequent to this.

3.2 Outcomes

- A practical knowledge of the basic concepts of statistical data analysis and numerical data processing.
- A practical knowledge of how computer simulations can be used to understand the measurement and data analysis operations.
- A practical knowledge of standard numerical analysis methods.

3.3 Basic Statistical Analysis

3.3.1 Important Terms

Systematic Error: Reproducible inaccuracy introduced by faulty or inadequate equipment, calibration or technique.

Random Error: Indefiniteness of result introduced by finite precision of measurement. Measure of fluctuation after repeated experimentation.

Uncertainty: Magnitude of error that is estimated to have been made in determination of results.

Accuracy: Measure of how close the result of an experiment comes to the “true” value.

Precision: Measure of how carefully the result is determined without any reference to any “true” value.

Parent Population: Hypothetical infinite set of data points of which the experimental data points are assumed to be a random sample.

Parent Distribution: Probability distribution of the parent population from which the sample data are chosen.

3.3.2 Significant Figures

1. The leftmost non-zero digit is the most significant digit.
2. If there is no decimal point, the rightmost non-zero digit is the least significant digit.
3. If there is a decimal point, the rightmost digit is the least significant digit, even if it is a zero.
4. All digits between the least and the most significant digits are counted as significant digits.
5. It is sufficient to quote an uncertainty to two significant figures, e.g. if a result is calculated to be 35.218 ± 1.47 it should be recorded as 35.2 ± 1.5.
6. The answer should be quoted to the same number of decimal places as the uncertainty, e.g.:

$$62 \pm 0.019 \quad \text{is wrong}$$

$$62.34821 \pm 0.019 \quad \text{is wrong}$$

$$62.348 \pm 0.019 \quad \text{is correct}$$

3.4 Common Descriptive Quantities

These formulae assume uniform uncertainties in the measured values x_i, so they do not include weighting. They only hold for random errors, and should be used cautiously where $N \leq 10$ or there are significant systematic errors.

Expectation value: Weighted average of a function $f(x)$ over all values of x:

$$\langle f(x) \rangle = \lim_{N\to\infty} \left[\frac{1}{N} \sum_{i=1}^{N} f(x_i) \right] \quad \text{or} \quad \sum_{j=1}^{n} f(x_j) P(x_j) \quad \text{or} \quad \int_{-\infty}^{+\infty} f(x) P(x)\, dx$$

Mean: The population mean μ cannot be obtained in a real experiment.

$$\mu = \langle x \rangle = \lim_{N\to\infty} \left[\frac{1}{N} \sum_{i=1}^{N} x_i \right] \quad \text{or} \quad \sum_{j=1}^{n} x_j\, P(x_j) \quad \text{or} \quad \int_{-\infty}^{+\infty} x\, P(x)\, dx$$

Deviation: $d_i = x_i - \mu$

Variance:

$$\sigma^2 = \langle d_i^2 \rangle = \langle x^2 \rangle - \mu^2 = \lim_{N\to\infty} \left[\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2 \right] = \lim_{N\to\infty} \left[\frac{1}{N} \sum_{i=1}^{N} x_i^2 \right] - \mu^2$$

Standard deviation: $\sigma = \sqrt{\sigma^2}$

The standard deviation is the root mean square (RMS) of the deviations. It is a measure of the precision of a single measurement x_i. For a "Normal" or "Gaussian" distribution of deviations $\sim 68\%$ of the x_i will fall in the range $\bar{x} \pm \sigma$, $\sim 95\%$ in the range $\bar{x} \pm 2\sigma$, and $\sim 99\%$ in the range $\bar{x} \pm 3\sigma$.

Sample mean: (AVERAGE() in Excel) The sample mean is used as a best estimate of the population mean μ:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Sample variance: The sample variance is used as a best estimate of the population variance σ:

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

Sample standard deviation: (STDEV() in Excel)

$$s = \sqrt{s^2}$$

Standard deviation of the mean (also known as the "uncertainty"): This is a measure of the precision of a mean derived from a set of measurements $\{x_i : 1 \leq i \leq N\}$:

$$s_m = \frac{\sigma}{\sqrt{N}} \approx \frac{s}{\sqrt{N}} \tag{0-1}$$

A mean result should usually be quoted as:

$$\bar{x} \pm s_m$$

For Gaussian random deviations there is a 68% probability that $\bar{x}$ is within s_m of the population mean μ, 95% probability that it is within $2s_m$, and 99% probability that it is within $3s_m$ (The range-estimator was used to approximate s_m in PHY101).

3.4.1 Propagation of Uncertainties

As an example, assume the mass M, radius R and length L of a solid cylinder are measured in order to determine the density ρ of its constituent material using the standard formula:

$$\rho = \frac{M}{\pi R^2 L} \tag{0-2}$$

In practice we calculate:

$$\bar{\rho} = \frac{\bar{M}}{\pi \bar{R}^2 \bar{L}} \tag{0-3}$$

where $\bar{\rho}$ is an estimate of the true density calculated from estimates of the mass $\bar{M}$, radius $\bar{R}$ and length $\bar{L}$. A first-order Taylor expansion of this equation yields:

$$\rho = \bar{\rho} + \Delta M \left(\frac{\partial \rho}{\partial M}\right)_{\bar{M},\bar{R},\bar{L}} + \Delta R \left(\frac{\partial \rho}{\partial R}\right)_{\bar{M},\bar{R},\bar{L}} + \Delta L \left(\frac{\partial \rho}{\partial L}\right)_{\bar{M},\bar{R},\bar{L}} \tag{0-4}$$

or:

$$\Delta\rho = \rho - \bar{\rho} = \Delta M \frac{1}{\pi \bar{R}^2 \bar{L}} - \Delta R \frac{2\bar{M}}{\pi \bar{R}^3 \bar{L}} - \Delta L \frac{\bar{M}}{\pi \bar{R}^2 \bar{L}^2} \tag{0-5}$$

Interpreting $\Delta\rho$, ΔM, ΔR and ΔL as deviations of the respective variables, we can express the expected variance of the density in terms of the variances of the constitutive quantities:

$$\sigma_\rho^2 \approx s_M^2 \frac{1}{\pi^2 \bar{R}^4 \bar{L}^2} + s_R^2 \frac{4\bar{M}^2}{\pi^2 \bar{R}^6 \bar{L}^2} + s_L^2 \frac{\bar{M}^2}{\pi^2 \bar{R}^4 \bar{L}^4} \tag{0-6}$$

$$\frac{\sigma_\rho^2}{\bar{\rho}^2} \approx \frac{s_M^2}{\bar{M}^2} + 4\frac{s_R^2}{\bar{R}^2} + \frac{s_L^2}{\bar{L}^2} \tag{0-7}$$

(Bevington & Robinson, sec 3.2). One of the assumptions made in this approximation is that the measurement errors are independent and uncorrelated.

The general rule for propagation of uncertainties (errors) for a formula of the form:

$$\bar{x} = x(\bar{u}, \bar{v}, \bar{w}, \cdots) \tag{0-8}$$

is given by:

$$\sigma_x^2 \approx \sigma_u^2 \left(\frac{\partial x}{\partial u}\right)^2_{\bar{u},\bar{v},\bar{w},\cdots} + \sigma_v^2 \left(\frac{\partial x}{\partial v}\right)^2_{\bar{u},\bar{v},\bar{w},\cdots} + \sigma_w^2 \left(\frac{\partial x}{\partial w}\right)^2_{\bar{u},\bar{v},\bar{w},\cdots} + \cdots \tag{0-9}$$

3.5 Probability Distributions and Histograms

If a random variable x has a parent probability distribution $P(x)$ then this implies that the probability that x lies between x_1 and x_2 is given by:

$$A(x_1, x_2) = \int_{x_1}^{x_2} P(x)\, dx \tag{0-10}$$

Sample distributions for measured or simulated data are usually approximated using histograms, which require the data to be binned. The histogram "columns" record the number of values lying within the bin interval, and the uncertainty in this value is equal to its square-root (because of binomial counting statistics). The histogram can be normalized (i.e. divide each column height by the total number of samples) to provide a probability histogram $P(x)\Delta x$, and then by the bin width Δx to produce a sample probability density function $P(x)$.

3.5.1 Summary of Important Distributions

Binomial distribution: (BINOMDIST() in Excel) Describes the probability of observing x successes out of n tries when the probability for success in each try is p:

$$P_B(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad \mu = np \quad \sigma^2 = np(1-p) \tag{0-11}$$

Poisson distribution: (POISSON() in Excel) Limiting case of the binomial distribution for large n and constant μ; appropriate for describing small samples from large populations:

$$P_P(x; \mu) = \frac{\mu^x}{x!} e^{-\mu} \quad \sigma^2 = \mu \tag{0-12}$$

Gaussian (Normal) distribution: (NORMDIST() in Excel) Limiting case of the binomial distribution for large n and finite p; appropriate for smooth symmetric distributions:

$$P_G(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2} \tag{0-13}$$

3.5.2 Transformation of Probability Distributions

Computer-based random number generators (including the Excel RAND() function) usually return values that are uniformly distributed over the interval $[0, 1]$, i.e.:

$$P_U(x) = \begin{cases} 0 & ;\ x < 0 \\ 1 & ;\ 0 \le x \le 1 \\ 0 & ;\ x > 1 \end{cases} \tag{0-14}$$

This probability is clearly normalized, i.e.:

$$\int_{-\infty}^{+\infty} P_U(x)\, dx = 1$$

In general, however, we require a sample of random numbers drawn from a specific parent probability distribution. In order to produce a sample of random numbers with an arbitrary distribution $P_A(y)$ from a given sample with probability distribution $P_G(x)$ we make use of the "conservation of probability" principle:

$$|P_G(x)\, dx| = |P_A(y)\, dy| \tag{0-15}$$

(Bevington & Robinson, sec 5.3), which allows us to write:

$$\int_{-\infty}^{x} P_G(x')\, dx' = \int_{-\infty}^{y} P_A(y')\, dy' \tag{0-16}$$

Assuming that a random number generator has the uniform probability distribution mentioned above, equation 0-16 simplifies to:

$$x = \int_{-\infty}^{y} P_A(y')\, dy' \tag{0-17}$$

Therefore, if x is drawn from a uniform probability distribution, we can solve the above equation for y to obtain a number which is drawn from the required distribution. This process *transforms* the given random variable x into a new random variable y.

For example, if a normalized uniform distribution of the form:

$$P(y) = \begin{cases} 0 & ;\ y < -1 \\ 0.25 & ;\ -1 \le y \le +3 \\ 0 & ;\ y > +3 \end{cases} \tag{0-18}$$

is required, then equation 0-17 becomes:

$$x = \int_{-1}^{y} 0.25\, dy' = 0.25 \times [y - (-1)] = \frac{y+1}{4}$$

which is easily solved for the required transformation equation:

$$y = 4x - 1$$

(which could have been derived more intuitively). The transformation required to obtain a gaussian random variable is obtained by substituting equation 0-13 into equation 0-17:

$$x = \int_{-\infty}^{y} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(y'-\mu)^2/2\sigma^2}\, dy' = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(y-\mu)/\sigma} e^{-t^2/2}\, dt = \Phi\left(\frac{y-\mu}{\sigma}\right) \tag{0-19}$$

where $\Phi(z)$ is the cumulative normal distribution (NORMSDIST() in Excel), which is transcendental. Inversion of this equation is necessary to obtain the required random variable transformation:

$$y = \Phi^{-1}(x) \times \sigma + \mu \tag{0-20}$$

where $\Phi^{-1}(p)$ is the inverse cumulative normal distribution (NORMSINV() in Excel).

3.5.3 Exercises

Items in boxes like this are deliverables that *must* appear in your prac write-ups for minimum marks. Maximum marks can be obtained by adding insightful comments and discussion.

1. Generate a spreadsheet column with 1000 random values that are drawn from a population that is uniformly distributed over the interval $[0, 1]$.

2. Use the Excel Histogram tool to generate three different sample probability density functions for these random samples. Use bin widths Δx of 0.1, 0.05 and 0.025 respectively, convert the column heights to probabilities $P(x)\Delta x$, and then convert $P(x)\Delta x$ to probability densities, $P(x)$, by dividing by the bin width. Find the standard deviation of the $P(x)$'s in each case and then investigate the dependence of the standard deviations on the bin width. 3 histograms, Table of s vs bin width.
   BEWARE! Excel does not plot a true histogram, only a bar chart! Think CAREFULLY about where the labels are appearing (edge or middle?) at the bottom of the bars and thus what numbers they should be.

3. Transform the 1000 random values into samples that are drawn from a population that is uniformly distributed over the interval $[-5, +5]$. Obtain a histogram of these transformed values using a bin width of 0.5 to verify the success of your transformation. Histogram

4. Transform the original 1000 random values into samples that are drawn from a normal parent distribution with mean $\mu = 5.0$ and standard deviation $\sigma = 1.5$. Obtain a normalized histogram of $P(x)dx$ with a bin width of 0.5 for this gaussian set of values and superimpose a curve of the parent probability distribution.

   Use Equation (0-13) to get the parent distribution, but remember to multiply $P(x)$ by the bin width to get a probability curve for comparison.

   Single graph with data histogram and parent distribution curve.

5. Use the Excel AVERAGE() and STDEV() functions to verify that the sample mean and standard deviation are close to the expected parent values. Do this for both the un-binned (raw) data and the binned data. Why are they different?
   Table comparing results

6. Verify that the cumulative probabilities for the intervals $[-s, +s]$, $[-2s, +2s]$ and $[-3s, +3s]$ are consistent with the expected values. Table comparing results

7. Produce another 9 columns of 1000 gaussian random variables, and compute their sample means, standard deviations and the standard deviation of the means for each set of data. Compare the measured standard deviation of these 10 means with the expected standard deviation of the mean (calculate this from the 10 individual means).
   Table comparing results