

Statistics Homework Problems for 4803L

3rd January 2006

These exercises have to be returned at the following dates:

Exercise 1-3: Thursday, January 19th

Exercise 4-6: Thursday, January 26th

Exercise 7-9: Thursday, February 2nd

The mentioned data files, Matlab routines, and Xmgrace (or xmgr) files are available at

'<http://www.phys.ufl.edu/~mueller/TEACHING/HOMEWORK/>'.

Exercise 1 (To get started)

A Matlab routine 'dice_throws.m' is posted on the web. The routine simulates random throws of one die and two dice. Use it and produce sample sets with 100, 1000, and 10000 data points. Calculate the best estimates for the standard deviation and the sample standard deviations for each of the six data sets. Calculate the normalized sample distribution (histogram) of the generated sample sets and plot it together with the parent distribution. Also plot the residuals of the distribution (difference between histogram and parent distribution) and calculate the standard deviation of this distribution. You can download the Matlab routine 'dice_example_analysis.m' to do all this for the 1-die throws. Feel free to modify the routine to also do the 2-dice throws analysis with Matlab or load the data sets into Excel or any other program to finish the data analysis.

Exercise 2 (Gaussian pdf):

Show by direct integration (no integral tables) that the Gaussian pdf is properly normalized:

$$\int_{-\infty}^{\infty} p(y)dy = 1$$

It can also be shown by direct integration that

$$\begin{aligned}\langle y \rangle &= \int_{-\infty}^{\infty} yG(y, \mu, \sigma) = \mu \\ \langle (y - \mu)^2 \rangle &= \int_{-\infty}^{\infty} (y - \mu)^2 G(y, \mu, \sigma) = \sigma^2\end{aligned}$$

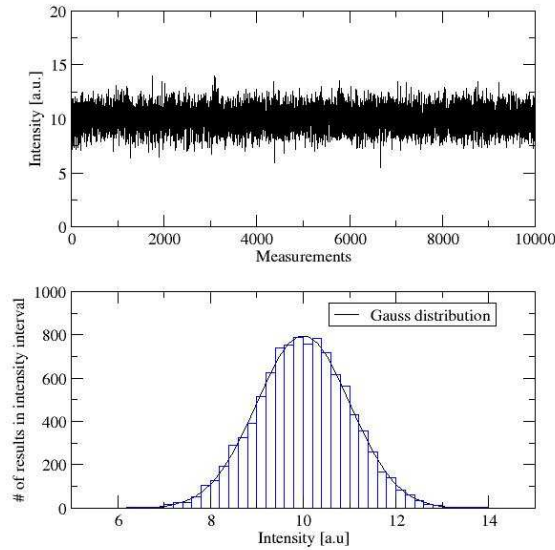


Figure 1: The top curve shows a simulated time series of power measurements. The bottom curve shows a histogram of the data and the probability density function (Gaussian distribution with mean=10 and standard deviation=1).

Exercise 3 (Histograms and Gaussian pdfs):

One example of a measurement of a continuous random variable is the intensity of a laser field. Fig. 1 shows a (simulated) intensity measurement with 10000 data points, a histogram of the data, and a Gaussian probability density function.

Download the Matlab routine 'powernoise.m' to simulate laser intensity noise similar to the one shown in Fig. 1. Produce data sets with 100, 1000, and 10000 data points. Calculate the sample mean and the sample standard deviation for each data set. Make histograms of the generated data and fit Gaussian pdf's to it. Choose the width for the histogram wisely and discuss what would happen if the bin width would be too large or too small. Feel free to use 'exercise_3.m' to calculate the mean and the standard deviation. Matlab has linear and nonlinear fitting routines (see the examples on the web), but you could also use Excel or Xmgrace to load the histograms and fit Gaussian pdf's to it. Discuss the results.

Exercise 4 (weighted averages):

We have prepared $M=6$ data set (labeled $j = 1..6$) each with N_j samples (labeled $i = 1..200$, here all $N_j = 200$) from 3 different Gaussian pdfs. You can download the data sets from the statistics homepage (Exercise4_set#.dat) or open 'Exercise_4_allsets.agr' in xmgrace. Calculate the following quantities for each data set:

- average of j^{th} data set*
- sample standard deviation of j^{th} data set*
- simple average of all data*
- best estimates for the weights*

*best estimate for the weighted average
weighted sample standard deviation*

Exercise 5 (Student-T probabilities):

A few words about the Student-T probabilities: Let us again assume a random variable which follows a Gaussian pdf. In that case 68% of all measurements $\{y_i\}$ will be in the range $\mu_y \pm \sigma_y$ and 95% will be in the $\mu_y \pm 2\sigma_y$ range. We can turn this argument around and construct confidence intervals around measured data averages:

$$\bar{y} \pm z\sigma_y$$

and say that with $z = 1$ there is a 68% probability that μ_y is in the interval $\bar{y} \pm \sigma_y$ and with $z = 2$ there is a 95% probability that μ_y is in the interval $\bar{y} \pm 2\sigma_y$. Such coincidence intervals are seldom reported because they are well known and completely specified once \bar{y} and σ_y are given.

The situation changes when the true standard deviation is unknown. In that case we have to use the sample standard deviation and the sample mean to construct confidence intervals:

$$\bar{y} \pm z s_{\bar{y}}$$

The interval is now constructed with an estimated standard deviation which can be smaller or larger than the true standard deviation. Therefore $z = 1$ (or $z = 2$) does no longer create 68% (or 95%) confidence intervals. William Sealy Gosset, publishing under the pseudonym Student, was the first to take this properly into account. The uncertainty in $s_{\bar{y}}$ decreases with increasing number of degrees of freedom n . The larger the n , the better the estimate and the closer the Student-T intervals will be to the corresponding Gaussian intervals. The number of degrees of freedom is number of data points minus all fitting parameters used to calculate the sample variance.

Table 4 at the end of the statistics write-up shows some Student-T probabilities. As an example of its use, consider the following measured data:

13.4	12.6	13.6	12.5	13.1
------	------	------	------	------

1. First calculate the sample average and the sample standard deviation:

$$\bar{y} = 13.04 \quad s^2 = \frac{\sum_{i=1}^5 (y_i - \bar{y})^2}{4} = 0.233 \quad s = 0.483$$

The number of degrees of freedom is the number of data points minus 1 (because we calculated the mean):

$$d.o.f. = 5 - 1 = 4$$

2. So lets go to the $n = 4$ row in the Student-T table (Table 4). Choose the probability value you want. Let's use 95%, so we go to the second column (the 0.95-column) and the $n = 4$ row and pick the value:

$$z(n = 4; P = 0.95) = 2.77645 = 2.78$$

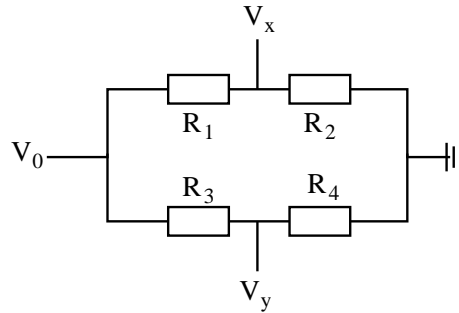


Figure 2: The measured voltages V_x and V_y depend on the values of R_1 to R_4 which in turn depend on the temperature. But they are also correlated through V_0 . Any fluctuation in the supply voltage will certainly also change the two measured voltages.

3. The 95% confidence interval that the true mean is in that interval is:

$$\bar{y} \pm z s_{\bar{y}} = 13.04 \pm 1.34$$

If we would know the true standard deviation, the z-factor would be 2 (for 95% probability interval). If you look again at the Student-T probabilities, you will see that the listed z-value in the 0.95 column is approaching 2 for larger degrees of freedom. Generally, Student-T probabilities are only important for small data sets.

Calculate the 90% confidence interval for the true mean of the following data set:

4.87	4.40	4.43	4.53	4.11	4.63
------	------	------	------	------	------

Exercise 6 (Correlations):

One example for correlated measurements are measurements of voltages in a resistor bridge (see Fig. 2). While the voltages all depend on the values of the resistors (which depend for example on the temperature), the voltages also depend on or are correlated through the supply voltage.

Download a Matlab routine which simulates data obtained from a bridge circuit as shown in Fig. 2 under various conditions: (a) Supply Voltage noise, (b) Resistor Noise, (c) Supply Voltage and Resistor noise. Generate your own set of data (voltage_noise, resistor_noise, voltage_resistor_noise) and calculate the sample variances, the sample covariances, and the sample correlation coefficient of all three pairs of data. Discuss the results.

Exercise 7 (Error propagation):

$$\sigma_a^2 = \sum_{i=1}^M \left(\frac{\partial f(y_1, y_2, \dots, y_M)}{\partial y_i} \right)^2 \sigma_i^2 + 2 \sum_{i=1}^M \sum_{j>i}^M \left(\frac{\partial f(y_1, y_2, \dots, y_M)}{\partial y_i} \right) \left(\frac{\partial f(y_1, y_2, \dots, y_M)}{\partial y_j} \right) \sigma_{ij}^2$$

Derive this equation.

Exercise 8 (Error propagation):

One example for the propagation of errors is the efficiency measurement of the four scintillation counters in the muon experiment. It can be assumed that virtually all muons pass through all four counters potentially triggering each counter. The efficiency of the first counter can be measured by counting coincidences at all four counters (N_{1-4}) and compare it with the number of coincidences of the last three counters (N_{2-4}). The efficiency of the first counter is then:

$$\epsilon_1 = \frac{N_{1-4}}{N_{2-4}}$$

To calculate the standard deviation of the derived efficiency we would need to know the standard deviations or variances of N_{2-4} and N_{1-4} and also the covariance. An easier way is to use the following variables:

$$N_+ = N_{1-4} \quad N_- = N_{2-4} - N_{1-4}$$

The first number is the number of events which triggered all detectors while the last number is the number of events which triggered only the last three detectors. These events are independent from each other and the covariance between the numbers is 0. Now we can assume that the counts N_+ and N_- follow a Poissonian distribution and have a standard deviation of $\sqrt{N_+}$ and $\sqrt{N_-}$, respectively.

A second possibility is to measure first N_{1-4} and then N_{2-4} over separate but equal time intervals. In this case N_{1-4} and N_{2-4} are statistically independent from each other and the covariance is 0. Using both possibilities described above, suppose both gave 100 counts for the coincidence in the three counters and 90 counts for the coincidence in all four counters. Calculate the efficiency and the standard deviation for both cases.

Exercise 9 (Linear regression):

We prepared a set of data where each data point has a different standard deviation. The set of data is 'set_1.dat'. The standard deviation of each point is given in 'set_2.dat'. Both data sets are also available in the Xmgrace file: Exercise_9.agr. Fit a quadratic function to the data in 'set_1.dat' first without using weights, then using weights based on the standard deviation given in 'set_2.dat'.