

Introduction to Probability, Statistical Interference, Instrumentation, Data Analysis, and Model Fitting

Laura Ingleby, Joel Lamb, Andrew Cowan, Juan Diaz-Toledo

SEPTEMBER 9 2003

Introduction and Theory

This project has two parts. Part I: Probability and Statistics problems, and Part II: Experimental Data Acquisition and Model Fitting. Part I is an introduction to and review of the use of probabilities, permutations, combinations, standard deviations, etc.; simple statistical ideas that can be used in astronomy. Part II is an introduction to the use of laboratory data acquisition and instruments, using elementary data analysis techniques, including fitting models to data and determining associated uncertainties and quality of fit criteria.

Description of Experimental Procedure

Part I: Probability and Statistics Problems

1. Someone [e.g. the State!] offers the following wager: Place a \$1 bet and choose 5 numbers between 1 and 30. If all 5 numbers are guessed correctly (in any order), you win \$1 million. The numbers do not repeat (they are unique).

- a. Is this a good bet? [calculate the odds]

$$P(\text{odds}) = \frac{n!}{(n_{N-1}n_{N-2}n_{N-3}n_{N-4}n_{N-5})} = \frac{5!}{(30 \cdot 29 \cdot 28 \cdot 27 \cdot 26)} = \frac{1}{142506}$$

This means that by investing \$142,506, a shrewd gambler would be guaranteed to win the \$1 million prize, which is a seven-fold return. Indeed, if the lottery were arranged in such a way one would be a fool not to play.

- b. Suppose you only had to choose 4 numbers correctly. Is this a good bet? [odds?]

$$P(\text{odds}) = \frac{n!}{(n_{N-1}n_{N-2}n_{N-3}n_{N-4})} = \frac{5!}{(30 \cdot 29 \cdot 28 \cdot 27)} = \frac{1}{5481}$$

This is an even better bet. An investment of only \$5481 guarantees a 180-fold return. The only check against this being used to make oneself wealthy at the expense of the state is the fact that everybody would be doing it, and the million dollar prize would have to be split among all the millions of winners.

2. Find a coin in your pocket.

a. Toss it 100 times. Record the number of heads.

b. Collect the results from other students in your team; enter these in your lab notebook.

Joel: 43

Laura: 49

Juan: 56

Andy 56

c. Calculate the mean and standard deviation. Compare with the expected results (for a large number of trials).

$$\text{Mean: } \bar{x} = \frac{\sum x_i}{n} = 51$$

$$\text{Standard Deviation: } \sigma = \left[\frac{1}{n-1} \left(\sum x_i^2 - \frac{1}{n} \left(\sum x_i \right)^2 \right) \right]^{1/2} = 5.43$$

d. Calculate the probability that the number of heads exceeds 60.

Integrating the Gaussian distribution gives:

$$\frac{\int_{60}^{100} e^{\frac{-(x-50)}{100}}}{\int_{-\infty}^{\infty} e^{\frac{-(x-50)}{100}}} = 0.078 \approx \frac{1}{12}$$

Is the probability that the number of heads exceeds 60

e. If you had flipped a coin 10^6 times, what is the probability that the number of heads would exceed 600,000?

Integrating this distribution from 600,000 to 1,000,000, and normalizing as above, gives a result that is only formally non-zero. To within the numerical limits of either Mathematica or Mathcad, the result is zero. The careful reader will note that although 60 is only a little more than σ away from the mean for 100 trials, six-hundred thousand is around 140σ greater than the mean for a million trials. As 99.7% of all trials lie within $\pm 3\sigma$, a result that is only within 141σ is fantastically unlikely.

3. Suppose you have a drawer infinitely full of socks, red, green, blue in equal numbers.

a. What is probability of randomly choosing a pair of same color?

2 –red, 2 –green, 2 –blue, 6 total, probability = $2/6 = 1/3$.

b. Now suppose there is exactly one pair of each color. What is the probability of a matching pair now?

The probability of drawing the first sock is unity. The probability that the next sock drawn is the one in five, since there are five socks remaining.

c. Suppose there are 5 pairs Red socks, 10 pairs green, 20 pairs of blue socks. What is probability of picking 3 socks, all the same color?

5 pairs red = 10, 10 pairs green = 20, 20 pairs blue = 40, Total = 70.

$$P(1\text{red}) = 10/70 = 1/7, \quad P(2\text{red}) = 9/69 = 3/23, \quad P(3\text{red}) = 8/68 = 2/17$$

$$P(1\text{green}) = 20/70 = 2/7, \quad P(2\text{green}) = 19/69, \quad P(3\text{green}) = (9/34)$$

$$P(1\text{blue}) = 40/70 = 4/7, \quad P(2\text{blue}) = 13/23, \quad P(3\text{blue}) = 19/34$$

$$P(1\text{red}) \cap P(2\text{red}) \cap P(3\text{red}) = 0.0022$$

$$P(1\text{green}) \cap P(2\text{green}) \cap P(3\text{green}) = 0.021$$

$$P(1\text{blue}) \cap P(2\text{blue}) \cap P(3\text{blue}) = 0.18$$

$$P(\text{total}) = P1 + P2 + P3 = 0.19 = 19\%.$$

4. Galaxies have three morphological types: elliptical (40%), spiral (40%), and irregular (20%). Suppose they are randomly distributed in the Universe.

a. Pick a random elliptical galaxy. What is the probability that its nearest 3 neighbors are all spirals?

$$P(\text{Sp1}) = 0.40, \quad P(\text{Sp2}) = 0.40, \quad P(\text{Sp3}) = 0.40$$

$$P(\text{Sp1}) \cap P(\text{Sp2}) \cap P(\text{Sp3}) = 0.064 = 6.4\%$$

- b. Pick three random galaxies. What is the probability that none of the three are irregular?**

$$P(\text{Ir1}) = 1 - 20/100 = 0.8, \quad P(\text{Ir2}) = 0.8, \quad P(\text{Ir3}) = 0.8$$

$$P(\text{Ir1}) \cap P(\text{Ir2}) \cap P(\text{Ir3}) = (0.8)(0.8)(0.8) = .512 = 51.2\%$$

- c. Consider sampling volumes containing ten galaxies each. How many boxes would you need to sample before having a 50% probability that none of the galaxies was a spiral?**

Each box has a probability of $0.6^{10} = 0.006$ that none of its galaxies are spiral.

$0.5 / 0.006 = 83.33$ boxes to give a 50% chance that at least one box does not contain a spiral galaxy. 

- 5.[Poisson statistics]. Suppose you observe a faint star, counting photons every second. The mean number of photons per second is five.**

General Formula: $P = \frac{e^{-\lambda t} (\lambda t)^k}{k!}$ where

t = length of time (seconds)

λ = average number of occurrences per unit time

k = number of occurrences per unit time

Suppose you observe a faint star, counting photons every second. The mean number of photons per second is five.

- a. What is the probability of receiving 10 photons in a given second?**

$$P_{10} = \frac{e^{-5*1} * (5*1)^{10}}{10!} = .018 = 1.8\%$$

- b. 20 photons?**

$$P_{20} = \frac{e^{-5*1} * (5*1)^{20}}{20!} = .00000026 = 0.000026\%$$

- c. Zero photons**

$$P_0 = \frac{e^{-5*1} * (5*1)^0}{0!} = .0067 = 0.67\%$$

- d. What is the probability that 100 seconds will pass without any one second interval having a count of exactly zero photons?**

$$P_0^{100} = 51\%$$

6. Card tricks.

a. What is the probability of dealing an ace as the first card of a full deck?

$$P = \frac{4}{52} = \frac{1}{13} = .0769 = 7.69\%$$

b. Probability of dealing two straight aces?

$$P = \frac{4}{52} * \frac{3}{51} = \frac{1}{221} = 0.0045 = 0.45\%$$

c. Probability of dealing half the deck [26 cards] without a single ace?

$$P(w/oAce) = \prod_{i=0}^{25} \left(\frac{48-i}{52-i} \right) = 0.055$$

d. Probability of a blackjack (ace plus a face card: K, Q, J or 10) in first two cards of a full deck?

$$P(K) \cap P(A) = (4/52)(16*2/51) = 0.048 = 4.8\%$$

e. Probability that a poker hand (5 cards) dealt from full deck will contain exactly one pair?

$$P(pair) = \frac{n}{\prod_{i=48}^{52} i / 5!} = 0.4224 = 42.2\%$$

7. An astronomer observes an optical and x-ray flare simultaneously in an x-ray binary system. It is known that both optical and x-ray flares occur about once a day for an hour. The astronomer detected the simultaneous flare after observing the system continuously for 10 days. What is the probability that this is a coincidence (i.e., that the simultaneous flares are physically unrelated)? *Note: 'simultaneous' means that there is some overlap in the flare times.*

$$P = 1 * \frac{2}{24} * 10 = \frac{5}{6} = .833 = 83.3\%$$



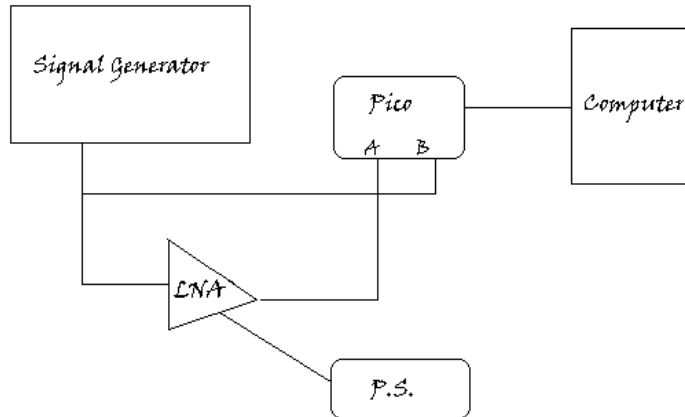
8. What is the probability that in a room of 30 people, at least 2 people share the same birthday?

$$P(ShareBirtday) = 1 - \frac{364 \times 363 \times 362 \times \dots (365 - n + 1)}{365^{n-1}} = 0.703 = 70.3\%$$

Part II: Experimental Data Acquisition and Model Fitting

Description of Experimental Procedure

A signal generator, low-noise amplifier (LNA) and data- acquisition system were set up as in the diagram below. With the signal generator set to 1 MHz, the amplitude attenuated by 40 dB and a sinusoidal waveform selected, the system was tested to verify that all cables were connected properly. Next, the input power level was adjusted to -40 dBm. While increasing the input level, measurements of the input and output levels were taken at 3dBm increments. Input and output levels were recorded until the input



level reached -10 dBm. This set of data was used to plot the output level as a function of input and also to plot the gain as a function of input level.

The input was then reset to -40dBm and the frequency adjusted to 2.0 MHz. The frequency was decreased by 200 KHz, the input and output levels were recorded at each step. This data set was used to graph the gain versus frequency.

Results

Assuming that the output impedance of the signal generator was 600 ohms, the power level was calculated. By looking at the signal on the oscilloscope tool, the peak-to-peak voltage was determined to be 10 mV. The peak-to-peak voltage was divided by $\sqrt{2}$ to find the root mean square voltage. The power was then calculated using the equation,

$$P = \frac{V^2}{R}$$

where V= voltage and R= impedance. $V = \frac{(10\sqrt{2})}{1000}$ $P = 600 \Omega$

$$P = \frac{(0.007V)^2}{600 \Omega}$$

$$P = 8.17 \times 10^{-8} \text{ Watts}$$

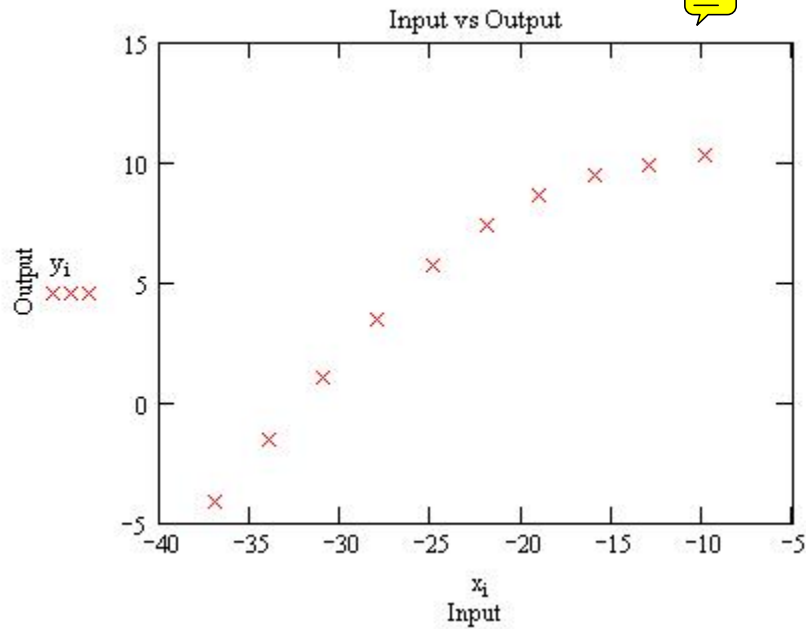
Converting to dBm = $10\log(1000P) = -40.9 \text{ dBm}$.

The results of the input v. output level experiment are included below:

Input vs Output	
Input (Dbm)	Output (Dbm)
-37	-4.10

-34.02	-1.56
-31	1.08
-27.98	3.46
-24.96	5.70
-21.96	7.34
-19.06	8.62
-15.94	9.42
-13.04	9.90
-9.94	10.28

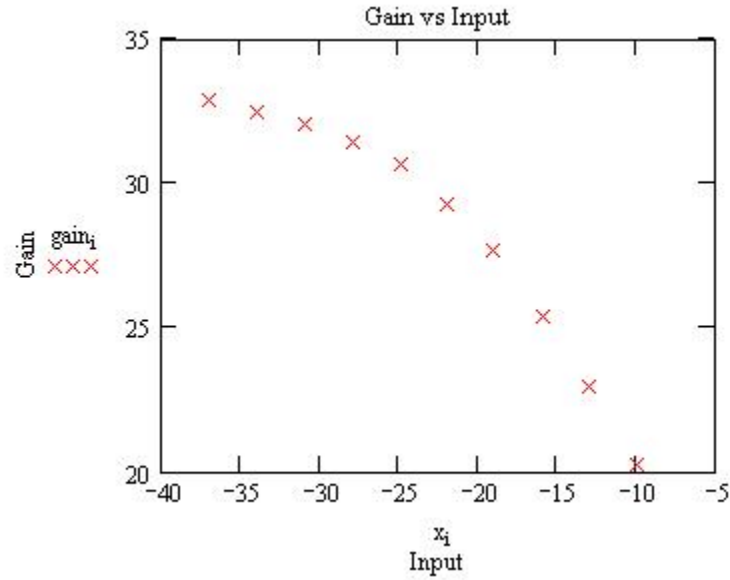
The waveform was noticeably deformed at about -22 dBm.
Mathcad was used to plot the output as a function of the input.



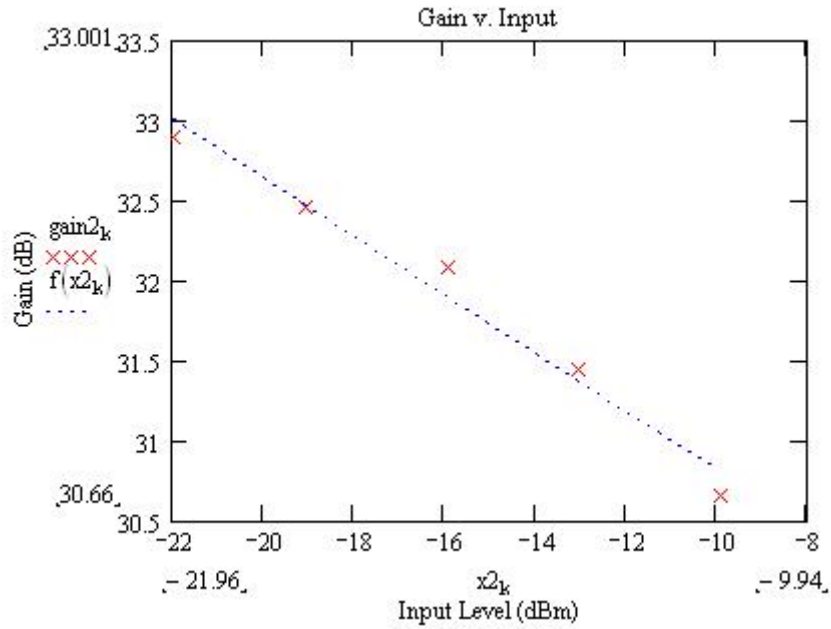
The first four points on the input vs output graph were used to fit a linear function to the data points.

The slope of the linear function is 0.842.

The input and output data was used to plot the gain (output-input) as a function of input.



The last four points on the input vs gain plot were used to fit a linear function.



The slope was calculated as -0.183

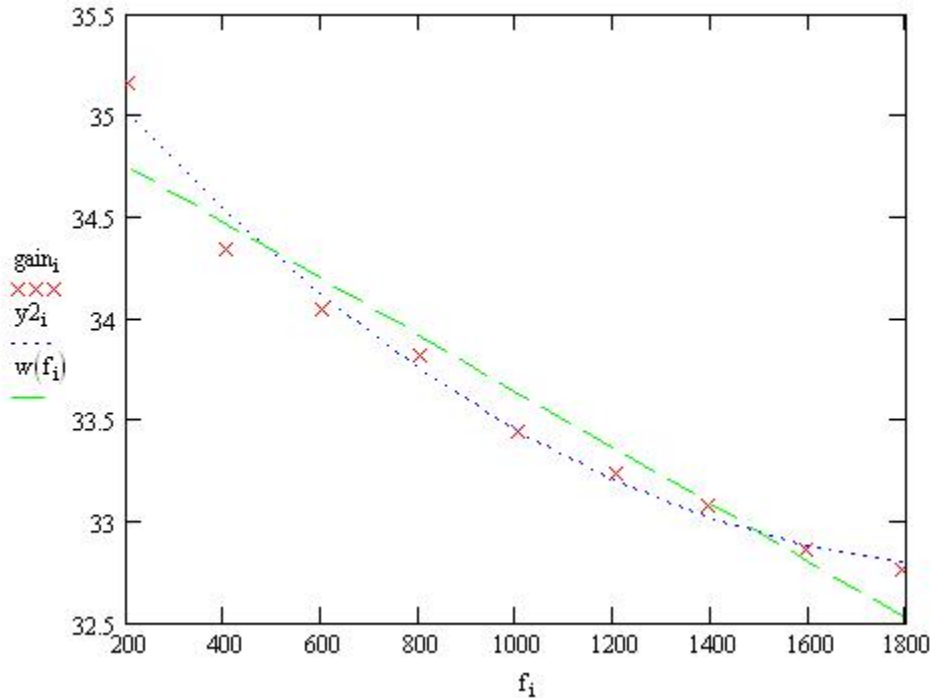
The measurements of input and output as frequency decreased follow.

Gain Vs Frequency			
Frequency (MHz)	Input (dBm)	Output (dBm)	Gain (dBm)
1.8	-40.04	-7.24	32.80
1.6	-40.00	-7.14	32.86
1.4	-40.00	-7.08	32.92
1.2	-40.28	-7.06	33.22
1.0	-40.46	-7.02	33.44




0.8	-40.84	-7.02	33.82
0.6	-41.10	-7.02	34.08
0.4	-41.34	-7.00	34.34
0.2	-42.16	-6.88	35.28

Mathcad was used to graph the gain as a function of the frequency and a linear function was fit to the data.



The slope of the linear fit was calculated as -1.389×10^{-3} .

Uncertainty Calculations:

The uncertainties for this experiment are in themselves somewhat uncertain. The manufacturer of the Pico interface does not provide accuracy data for the instrument, so it is difficult to know the accuracy of our results, however the power-level readouts appeared precise to around ± 3 dBm. 

The absolute power levels given here are all incorrect. Pico assumes the output impedance of its source is 600Ω , where more commonly (and in our case) it is 50Ω . However, we are mainly interested in gain values, which are correct since the error in power levels is consistent.

Discussion:

Analyzing the signal from the signal generator on Pico's spectral analyzer showed strong Fourier components at 2 MHz and all integer multiples thereof. This should not be surprising, since most signal generators do not produce particularly "clean" wave forms.

The amplifier that we tested conformed well to a linear fit for input levels up to around -30 dBm, after which the gain begins to decrease. Observing the oscilloscope display, clipping was noticeable above -21.9 dBm. We are asked to provide a 1 dB gain compression level from our gain v. input plot. It is immediately apparent looking at the plot that the gain begins to roll off almost immediately, with the 1 dB compression point occurring around -31dBm input (1 dBm output). The manufacturer specifies the 1 dB gain compression point as +5 dBm output. Converting 1 dBm measured to the actual power output, given a 50 Ohm output impedance gives +11 dBm, which is far better than the manufacturer's specification.

The amplifier's gain is clearly not constant with frequency over the range of our test, but fits decently to a quadratic regression.

Conclusion:

The beginning portion of this exercise was a useful introduction to and review of the methods of calculating probability as well as the Gaussian and Poisson distributions. It is certainly safe to say that each member of the group learned something from the exercise, and the crash course in Mathcad will also likely prove useful.

The amplifier that we tested was generally found not to be particularly linear for signal strength levels above -30 dBm, and gain compression begins at around the same input level. However, it should be noted that in most cases a low-noise amplifier is selected because the input signal level will be very low, so it might be the case that in the input power regime of the application, the gain and linearity would be sufficient.

Allocation of Effort

All team members contributed in equal proportions to this report.