Practical Statistics

- Lecture 3 (Sep. 2)
  Read: W&J Ch. 4-5
  - Correlation
  - Hypothesis Testing

- Lecture 4 (Sep. 4)
  - Principle Component Analysis

- Lecture 5 (Sep. 9):
  Read: W&J Ch. 6
  - Parameter Estimation
  - Bayesian Analysis
  - Rejecting Outliers
  - Bootstrap + Jack-knife

- Lecture 6 (Sep. 11)
  Read: W&J Ch. 7
  - Random Numbers
  - Monte Carlo Modeling

- Lecture 7 (Sep. 16):
  - Markov Chain MC

- Lecture 8 (Sep. 18):
  Read: W&J Ch. 9
  - Fourier Techniques
  - Filtering
  - Unevenly Sampled Data
JELLY BEANS CAUSE ACNE!  
SCIENTISTS! INVESTIGATE!  
BUT WE'RE PLAYING MINECRAFT!  
...FINE.

WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).

THAT SETTLES THAT.  
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.  
SCIENTISTS!  
BUT MINECRAFT!
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).

WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).

WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).

WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
Green jelly beans linked to acne!

95% confidence

Only 5% chance of coincidence!
HW 2 Discussion

• Problem 1 Discussion:
  - Are the fit values and uncertainties for 1 a and 1b. consistent with one another? If not, what is wrong with this approach?
  - The purpose of 1 c. is to compare to a linear fit to determine if adding a fit parameter improves the result.

• Problem 2 Discussion:
  - You may want to review the interpretation of chi-squared values in the Aug. 23 notes or Wall and Jenkins p. 87-89.
  - 2 c. should reflect the results you obtained in 1 a. and 1 b.
HW 2 Discussion

- Problem 3 Discussion:
  - You will need to calculate the ML function over a grid of slope ($m$) and intercept ($b$) values. You can use the results of 1 to estimate the right range to search.
  - Think of answering 3 b in terms of which argument you would present in a results talk on the data.

- Problem 4 Discussion:
  - You will reuse your code from problem 3 to calculate parameters for different fake datasets.
  - The programs on the webpage show you how to generate those fake datasets in Matlab.
Some Matlab tips

• Iterating over a variable
  
  for a=1:10
    Var=a*10;
    ....
  end

• How do you find the location of a grid maximum?
  
  - [maxvalue, maxind] =max(matrix(:));
  - [r,c] = ind2sub(size(matrix),maxind);

• Function Call
  
  - [output]=functionname(input);
Example Fitting Problem

- Reduced chi-squared = 0.43
  - What would you expect?

- What assumptions went into achieving these expectations?

- How could we estimate parameter uncertainties?
Monte Carlo Simulation

- Often we may find it easiest just to replicate an experiment or observation in the computer.
- In general these tools are referred to as “Monte Carlo” methods.
- General idea is to simulate randomness and reproduce observations for comparison with data.
- First we need a random number sequence.
Creating Random numbers

- A proper random sequence of numbers is a whole topic in itself. Numerical Recipes discusses this in some detail.

- A simple example of a random number generator is the sequence:

\[ I_{j+1} = \frac{(aI_j + c) \mod m}{m - 1} \]

Where a, c, and m are carefully chosen large numbers. \( I_j \) is a seed value that would always give us the exact same sequence of random numbers.
Example Monte Carlo calculation

• Let’s calculate a difficult, unknown integral:

![Diagram of a circle in a square]

Procedure:
- Draw pairs of uniform random numbers.
- Decide whether the numbers fall within the curve of interest.
- The integral is the fraction of points falling within the curve times the grid searched.
Transformation Method

• Starting from the law for transformation of probabilities:

\[ |p(y)dy| = |p(x)dx| \]

• We can rewrite to solve for the probability we want.

\[ \left| \frac{dy}{dx} \right| = \left| \frac{p(x)}{p(y)} \right| \]

\[ p(y) = \frac{dx}{dy} \]

1. Need to integrate the probability distribution
2. Solve for the new variable \((y)\) in terms of the uniform variable\((x)\)
Random Numbers

• The example gives a “uniform” distribution set of random numbers. That is,

\[ P(x)dx = \begin{cases} \ dx & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases} \]

• We would like useful distributions, such as Poisson, etc. To do so, we need to transform the random numbers.
Transformation Example

- I want to simulate the time it takes between arrival of photons at the detector. This is given by an exponential probability distribution: \[ P(t) = \alpha e^{-\alpha t} dt \]

- Use the transformation of probabilities:

- Need to integrate:

\[ \int e^{-t} dt = \int dx \]

\[ e^{-t} = x \]

\[ t = -\ln x \]

- A random number in the range 0 to 1 will be transformed to one which can be between Inf and 0
Transformation Limitations

• Transformation methods are limited to analytical probability distributions.

• One also needs to be able to integrate the probability distribution and invert the equation to solve for the new variable.
  - Numerical integration can be used as well, but this requires developing a look up table for the results, so the integration is not needed each random number generation.

• Often one of these criteria is not satisfied.
  - You can still generate useful random numbers using the rejection method.
Rejection Method

- Generate two uniform random deviates, x and y.
- Adjust x to span the range of values expected for the random number (x'=f(x)).
- Compare the value of y to the value of the probability distribution at x' (y'=p(x'))
- If y'<y use the value of x' in your simulation, if y'>y reject this pair and start over.
Rejection method
Using Monte Carlo Modeling to check experimental results

- We can check whether we understand our data by using the model we are fitting to generate a whole new experiment (or many equivalent experiments).
  - Model data contamination by systemic parameters.
  - Model effect of data uncertainties on parameter uncertainty.
  - Often easier to do than the data modeling procedures described previously.
We can use this to confirm our results from the white balls bet

Assume: Equal chance of N white balls from 0 to 10

• First round: get 3 white balls

• Second round: probability of getting 3 white balls?
MC simulation Results

First Round

Second Round
Check of Observations

• The bootstrap and jack-knife methods are Monte Carlo techniques which make the minimum assumptions about the data you have.

• Bootstrap only assumes the data points are independent and have identical distributions.

• The jack-knife method assumes the data points also are normally distributed.
Using Monte Carlo to plan observations.

• Say I want to understand how often I am likely to detect a 1-10 MJ planet in a 5-10 AU orbit around the 50 best stars.

  - I can detect 1 MJ planets at >1”, 5 MJ at 0.6” and 10 MJ planets at 0.3” for the best stars.
    ▸ Gets worse for the less optimum stars

  - Planet may be in wrong part of its orbit.
    ▸ We don’t know the inclination of the system.
    ▸ We don’t know the ellipticity of the orbit.
Approach to Modeling the Observations

- Randomly choose uniform separation from 5-10 AU.
- Randomly choose mass with $P(m) = \frac{1}{m}$ from 1-10 MJ.
- Randomly choose inclination with $P(i) = \sin(i)$ from 0-90.
- Randomly choose orbital phase uniform 0-360 degrees.
- Put around a star a certain distance away.
- Is it detectable?
- Repeat.

This approach will tell us how many planets we will detect around our sample if they _all_ had planets.
Can we detect a signal?

• Monte Carlo approaches give you the ability to inject a known signal in the data, to determine whether your data reduction could detect it.

• Example: For a planet detected via radial velocity, can a second, outer planet affect the fit of inner planet?

From Rodigas and Hinz 2009
Simulating unexpected noise sources

- In HW 1 we calculated a trade-off in SNR between taking longer exposures, and measuring the varying background more regularly.
  - For specific noise sources this can be calculated analytically
  - Monte Carlo allows you to just simulate what you would see.

\[
\text{counts}_1 = \mu_1 t + \mu_2 t + \sigma_{RON} \\
\text{counts}_2 = \mu_1 t + \sigma_{RON} + (\text{ugly noise source}) \\
\text{detection} = \text{counts}_1 - \text{counts}_2
\]

This general approach is used for Adaptive Optics simulations to understand the many contributors to wavefront error in a realistic way.
Markov Chain Monte Carlo

- Often want to explore a multi-dimensional parameter space to evaluate a metric.
  - A grid search approach is inefficient.
  - Want algorithm that maps out spaces with higher probability more effectively.

General procedure:
Start with a given set of parameters; Calculate metric
Choose new set of points and calculate new metric.
Accept new point with probability \( P = \frac{\text{metric(new)}}{\text{metric(old)}} \times P(x1,x2) \)

- The procedure provides the optimum parameter values, and also explores the parameter values in a way that allows derivation of confidence intervals.
- More on this next time.
Many Other Uses

- Radiative transfer modeling

- Multi-dimensional integration of complex functions.