12.010 Computational Methods of Scientific Programming

Lecturers
Thomas A Herring
Chris Hill
Review of last Lecture

• Looked at class projects
• Graphics formats and standards:
  – Vector graphics
  – Pixel based graphics (image formats)
  – Combined formats
• Characteristics as scales are changed
Class Projects

- Class evaluation will be done on Dec 6 as well.

1. 5-in-a-line game: D. Husmann, L. Song
2. Swarming behavior in animals: J. Berdahl, H. Owens
3. Monte-Carlo simulation: Q. Ejaz, P. James
4. Geothem and river incision: J. Stanley, E. Swanson, M. Perignon
5. Group theory: M. Reed, A. Marok
Review of statistics

• Random errors in measurements are expressed with probability density functions that give the probability of values falling between x and x+dx.
• Integrating the probability density function gives the probability of value falling within a finite interval.
• Given a large enough sample of the random variable, the density function can be deduced from a histogram of residuals.
Example of random variables

Random variable

Sample

Uniform

Gaussian
Histograms of random variables

- Gaussian: \( 490/\sqrt{2\pi} \cdot \exp(-x^2/2) \)
- Uniform

Number of samples vs Random Variable x
Characterization Random Variables

• When the probability distribution is known, the following statistical descriptions are used for random variable x with density function f(x):

\[
\text{Expected Value } \quad < h(x) > = \int h(x) f(x) \, dx \\
\text{Expectation } \quad < x > = \int x f(x) \, dx = \mu \\
\text{Variance } \quad < (x - \mu)^2 > = \int (x - \mu)^2 f(x) \, dx
\]

Square root of variance is called standard deviation
Theorems for expectations

• For linear operations, the following theorems are used:
  – For a constant $<c> = c$
  – Linear operator $<cH(x)> = c<H(x)>$
  – Summation $<g+h> = <g>+<h>$

• Covariance: The relationship between random variables $f_{xy}(x,y)$ is joint probability distribution:

$$
s_{xy} = <(x - \mu_x)(y - \mu_y)> = \int (x - \mu_x)(y - \mu_y)f_{xy}(x,y)dxdy
$$

Correlation: $\rho_{xy} = \frac{s_{xy}}{\sigma_x \sigma_y}$
Estimation on moments

• Expectation and variance are the first and second moments of a probability distribution

\[ \hat{\mu}_x \approx \frac{1}{N} \sum_{n=1}^{N} x_n / N \approx \frac{1}{T} \int x(t) dt \]

\[ \hat{\sigma}_x^2 \approx \frac{1}{N} \sum_{n=1}^{N} (x - \mu_x)^2 / N \approx \sum_{n=1}^{N} (x - \hat{\mu}_x)^2 / (N - 1) \]

• As N goes to infinity these expressions approach their expectations. (Note the N-1 in form which uses mean)
Probability distributions

• While there are many probability distributions there are only a couple that are common used:

Gaussian  \[ f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \]

Multivariant  \[ f(x) = \frac{1}{\sqrt{(2\pi)^n |V|}} e^{-\frac{1}{2}(x-\mu)^T V^{-1}(x-\mu)} \]

Chi – squared  \[ \chi_r^2(x) = \frac{x^{r/2-1} e^{-x/2}}{\Gamma(r/2)2^{r/2}} \]
Probability distributions

• The chi-squared distribution is the sum of the squares of $r$ Gaussian random variables with expectation 0 and variance 1.
• With the probability density function known, the probability of events occurring can be determined. For Gaussian distribution in 1-D; $P(|x|<1\sigma) = 0.68$; $P(|x|<2\sigma) = 0.955$; $P(|x|<3\sigma) = 0.9974$.
• Conceptually, people thing of standard deviations in terms of probability of events occurring (ie. 68% of values should be within 1-sigma).
Central Limit Theorem

- Why is Gaussian distribution so common?
- "The distribution of the sum of a large number of independent, identically distributed random variables is approximately Gaussian" 
- When the random errors in measurements are made up of many small contributing random errors, their sum will be Gaussian.
- Any linear operation on Gaussian distribution will generate another Gaussian. Not the case for other distributions which are derived by convolving the two density functions.
Random Number Generators

• Linear Congruential Generators (LCG)
  – \( x(n) = a \times x(n-1) + b \mod M \)
• Probably the most common type but can have problems with rapid repeating and missing values in sequences
• The choice of \( a \), \( b \) and \( M \) set the characteristics of the generator. Many values of \( a \), \( b \) and \( M \) can lead to not-so-random numbers.
• One test is to see how many dimensions of \( k \)-th dimensional space is filled. (Values often end up lying on planes in the space.
• Famous case from IBM filled only 11-planes in a \( k \)-th dimensional space.
• High-order bits in these random numbers can be more random than the low order bits.
Example coefficients

- Poor IBM case: \( a = 65539, \ b = 0 \) and \( m = 2^{31} \).
- MATLAB values: \( a = 16807 \) and \( m = 2^{31} - 1 = 2147483647 \).
- Knuth's Seminumerical Algorithms, 3rd Ed., pages 106--108: \( a = 1812433253 \) and \( m = 2^{32} \).
- Second order algorithms: From Knuth:
  \[
x_n = (a_1 x_{n-1} + a_2 x_{n-2}) \mod m
  \]
  \( a_1 = 271828183, \ a_2 = 314159269, \) and \( m = 2^{31} - 1 \).
Gaussian random numbers

• The most common method (Press et al.)
  Generated in pairs from two uniform random number x and y
  \[ z1 = \sqrt{-2\ln(x)} \cos(2\pi y) \]
  \[ z2 = \sqrt{-2\ln(x)} \sin(2\pi y) \]

• Other distributions can be generated directly (eg, gamma distribution), or they can be generated from the Gaussian values (chi^2 for example by squaring and summing Gaussian values)

• Adding 12-uniformly distributed values also generates close to a Gaussian (Central Limit Theorem)
Conclusion

• Examined random number generators:
• Tests should be carried out to test quality of generator or implement your (hopefully previously tested) generator
• Look for correlations in estimates and correct statistical properties (i.e., is uniform truly uniform)
• Test some algorithms with matlab: randtest.m