Statistics 110 - Introduction to Probability

Mark E. Irwin

Department of Statistics

Harvard University

Summer Term

Monday, June 26, 2006 -
Wednesday, August 16, 2006

Personnel

Instructor:	Mark Irwin
Office:	611 Science Center
Phone:	617-495-5617
E-mail:	irwin@stat.harvard.edu
Web-site:	http://www.people.fas.harvard.edu/∼mirwin/stat110/
Lectures:	M-F 11:00 - 12:00, Science Center 109
Office Hours:	Tuesday, 12:00 - 1:00, Thursday 1:00 - 2:00, or by appointment
Teaching Fellow:	TBA
E-mail:	
Section:	Monday, 1:00 - 2:00, Science Center 109

Text Books

Required Text: Rice JA (1994). Mathematical Statistics and Data Analysis, 2nd Edition. Duxbury Press.

Optional Texts: Ross S (1994), A First Course in Probability, 6th Edition. Prentice Hall.

Grading

- Homework (25%): 6 or 7 during the term. Late assignments will not be accepted. The lowest homework grade will be dropped when computing your final grade.

- Quizzes (15%): There will be two 30 minute quizzes during the term. Tentatively scheduled for July 11 and Aug 1.

- Midterm (20%): Tentatively scheduled for Wednesday, July 19 in lecture.

- Final (40%): Wednesday, August 16th, 9:00 am. Location to be announced.

Syllabus

- Basics: sample space, basic laws of probability, conditional probability, Bayes Theorem, independence.
- Univariate distributions: mass functions and densities, expectation and variance, binomial, Poisson, normal, and gamma distributions.
- Multivariate distributions: joint and conditional distribution, independence, covariance and correlation, transformations, multivariate normal and related distributions.
- Limit laws: moment generating and characteristic functions, probability inequalities, laws of large numbers, central limit theorem, delta rule.
- Simulation (Monte Carlo) Methods
- Markov chains: transition probabilities, classification of states, stationary distributions and convergence results.

What is Probability?

An approach (language) for describing randomness, uncertainty, and levels of belief.

If we are going to use probability to describe our levels of belief about a parameter or a process, we need to have an idea what we mean by probability.

Example I have two dice with me, one yellow and one purple. What is

- The probability that the yellow one rolls a 6?
- The probability that the purple one rolls a 6?
- The sum of the two rolls is 12?

Here is a picture of the two dice for those in the back.

So the probabilities are

- $P[\text{Yellow} = 6] = 0$
- $P[\text{Purple} = 6] = \frac{1}{20}$
- $P[\text{Sum} = 12] = \frac{1}{20}$

When determining probabilities and probability model there are two things that need to be considered:

1. What assumptions are you making (e.g. each outcome equally likely for each die and the dice are independent)?

2. What information are you conditioning on? All probabilities are effectively conditional.

What is Probability?

An approach (language) for describing randomness, uncertainty, and levels of belief.

Examples:

- Rolling a fair 6 sided die (Uniform distribution)

$$P[1] = P[2] = \ldots = P[6] = \frac{1}{6}$$

- Radioactive decay (Poisson Process)

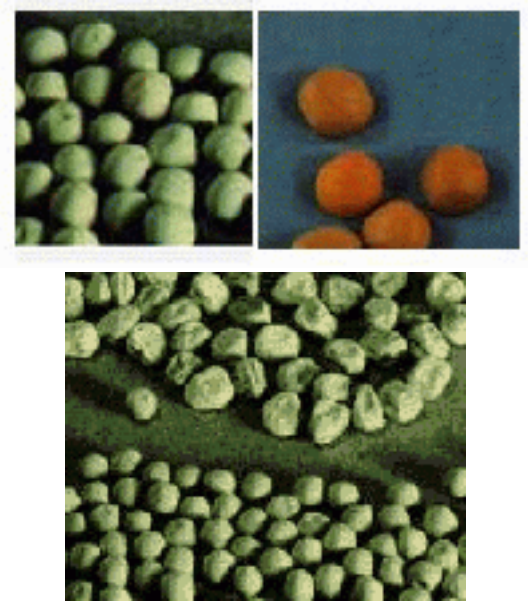
  The probability of having k alpha particles emitted from time T to time $T + t$ is

$$P[k \text{ decays from } T \text{ to } T + t] = \frac{e^{\lambda t}(\lambda t)^k}{k!}$$

  λ is a rate parameter giving the expected number of decays in 1 time unit.

- Genetics - Mendel's breeding experiments.

The expected fraction of observed phenotypes in one of Mendel's experiments is given by the following model



Round/Yellow	$\frac{2+(1-\theta)^2}{4}$
Round/Green	$\frac{\theta(1-\theta)}{4}$
Wrinkled/Yellow	$\frac{\theta(1-\theta)}{4}$
Wrinkled/Green	$\frac{(1-\theta)^2}{4}$

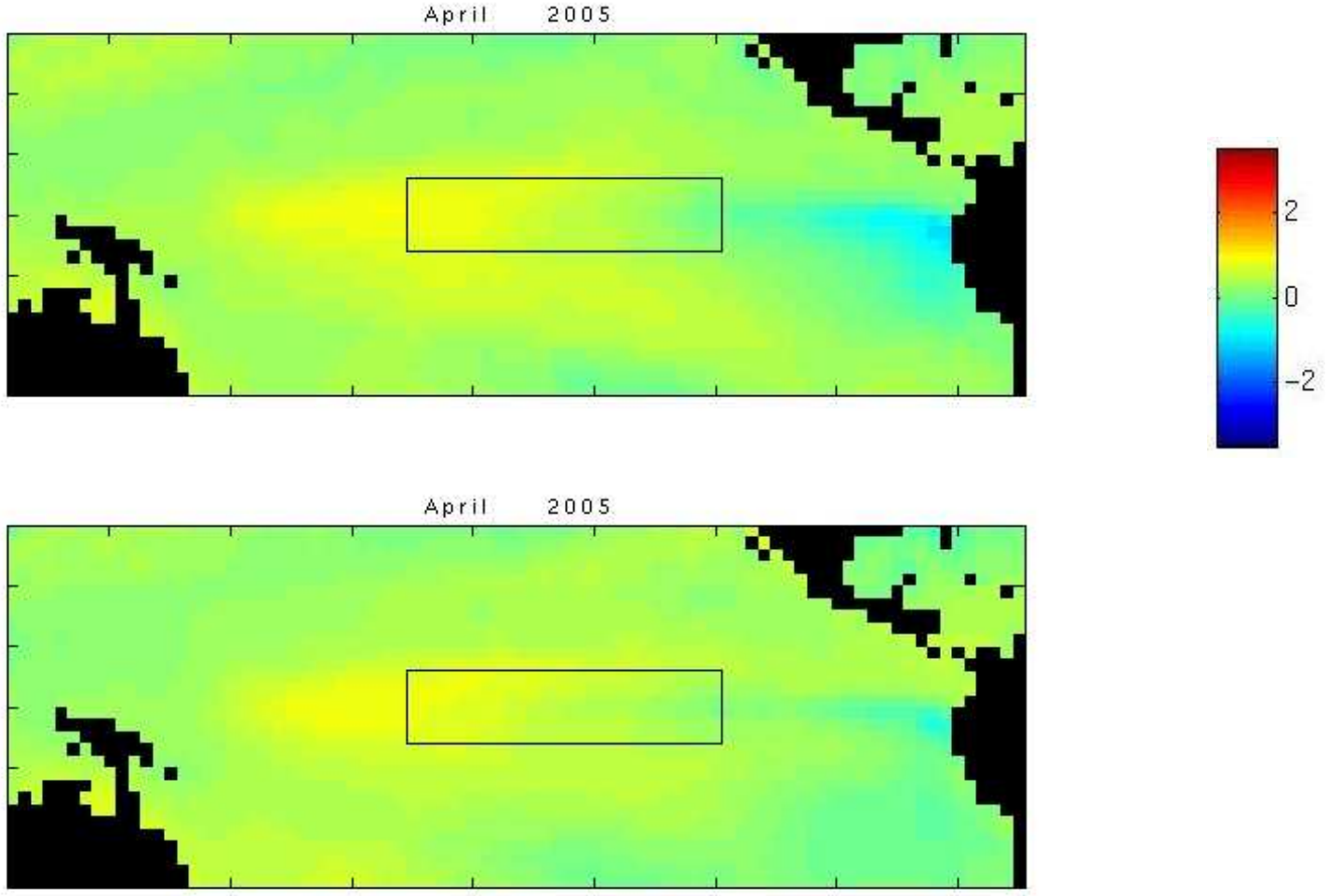
The value θ, known as the recombination fraction, is a measure of distance between the two genes which regulate the two traits. θ must satisfy $0 \leq \theta \leq 0.5$ (under some assumptions about the process of meiosis) and when the two genes are on different chromosomes, $\theta = 0.5$.

- Tropical Pacific sea surface temperature evolution

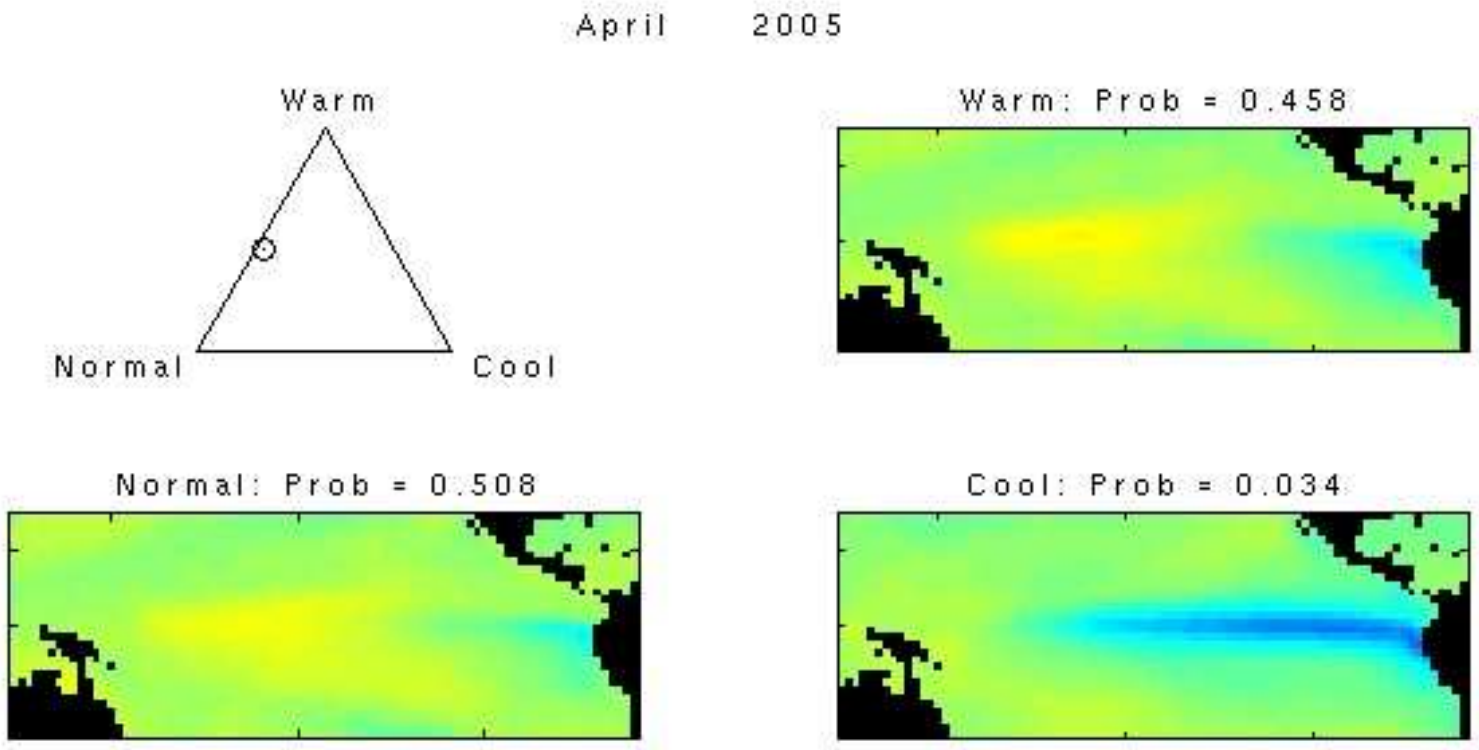
  Probabilistic based model involving wind, sea surface temperature, and the Southern Oscillation Index (air pressure measure) data to forecast sea surface temperatures 7 months in the future.

  A highly complex, hierarchical model was used. In addition to the three sets of input variables into the model, there is a regime component, recognizing that the region goes through periods of below normal, above normal, and typical temperatures. As part of the analysis, we can ask what would temperature field look like conditional on being in a given temperature regime.

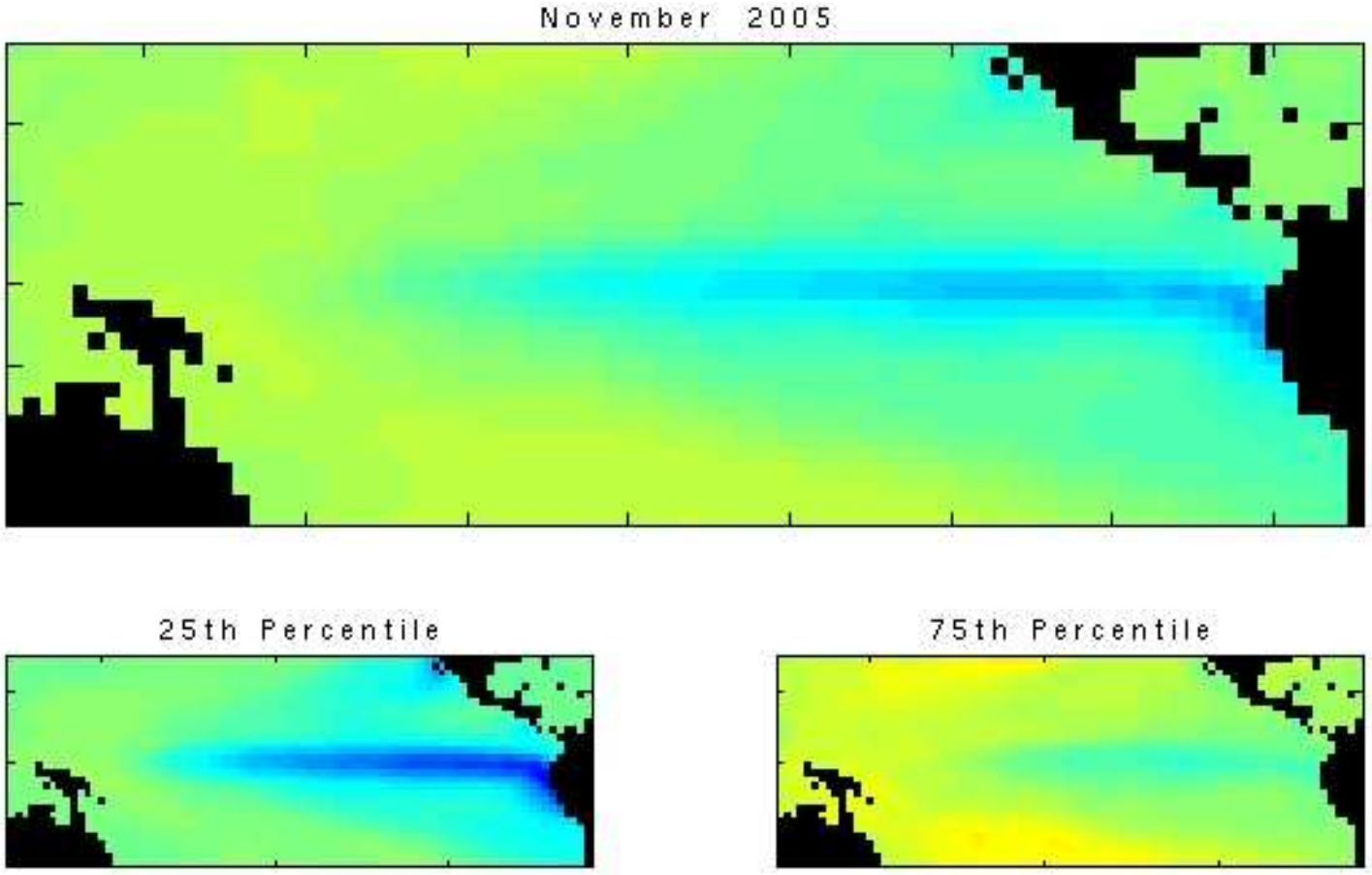
  The forecast map is a summary of the probabilistic model (the posterior mean), which is a weighted average over all possible sea surface temperature maps. As we can determine the likelihood of any temperature field, summary measures of the uncertainty in the forecasts can be determined.



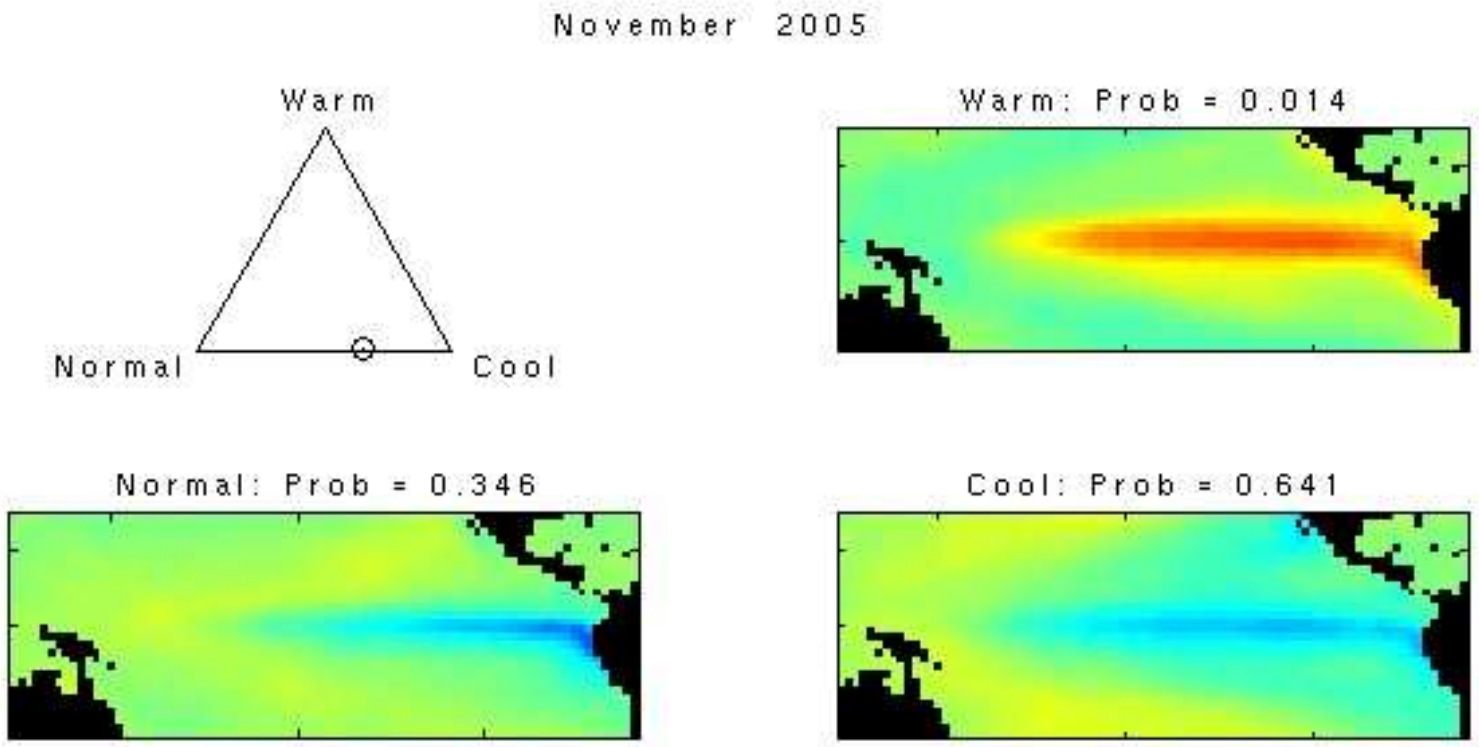
April 2005 anomaly forecast (top) and observed anomaly (bottom) based on Jan 1970 to September 2004 data.



April 2005 regime specific anomaly forecasts based on Jan 1970 to September 2004 data.



November 2005 anomaly forecasts with 25th and 75th percentiles based on Jan 1970 to April 2005 data. For the Niño 3.4 region, the average forecast anomaly is about -0.8°C, suggesting a possible La Niña this winter.



November 2005 regime specific anomaly forecasts based on Jan 1970 to April 2005 data.

Refs:

Berliner, Wikle, and Cressie (2000). Long-Lead Prediction of Pacific SSTs via Bayesian Dynamic Modeling. *Journal of Climate*, **13**, 3953-3968.

http://www.stat.ohio-state.edu/∼sses/collab_enso.php

Note that the model used here is only an approximation of reality. This does not include information about the underlying physics of the meteorology.

The models for the other 3 examples are much closer to the true mechanism generating the data, though there still may be some approximation going on, particularly with the Mendel example.

What is Statistics?

- To gain understanding from data
- Making decisions or taking actions under uncertainty
- Parameter estimation
  - Mendel: recombination fraction θ.
  - SST forecasts: there are a large number of parameters that need to be estimated from the past data. These involve the associations (correlations) of the winds and the SOI with the SST field, how the three variables evolve over time, how likely it is to switch from one regime to another, and so on.

- One approach used is to use data to determine which probability models are plausible

  For the Mendel data

Phenotype	Count	Obs. Rel. Freq.	Exp. Rel. Freq.
Round/Yellow	315	0.5625	$0.5625 = \frac{9}{16}$
Round/Green	108	0.1929	$0.1875 = \frac{3}{16}$
Wrinkled/Yellow	105	0.1875	$0.1875 = \frac{3}{16}$
Wrinkled/Green	32	0.0571	$0.0625 = \frac{1}{16}$

  For this pair of traits, the data is consistent with the two genes being on different chromosomes. In fact the genes for all 7 traits studied by Mendel are each on different chromosomes.

For a second example using the same breeding design (sorry I can't find out what the two traits actually are here)

Phenotype	Count	Obs. Rel. Freq.	Exp. Rel. Freq.
A/B	125	0.6345	$0.5625 = \frac{9}{16}$
A/b	18	0.0914	$0.1875 = \frac{3}{16}$
a/B	20	0.1015	$0.1875 = \frac{3}{16}$
a/b	34	0.1726	$0.0625 = \frac{1}{16}$

This data does not appear to be consistent with the hypothesis that the genes are on different chromosomes. In fact the data is most consistent with $\theta = 0.21$. This estimate is known as the maximum likelihood estimate of the parameter.

Where to probabilities come from?

- Long run relative frequencies

  If an experiment is repeated over and over again, the relative frequency of an event will converge to the probability of the event.

  For example, three different experiments looked at the probability of getting a head when flipping a coin.

  – The French naturalist Count Buffon: 4040 tosses, 2048 heads ($\hat{p} = 0.5069$).

  – While imprisoned during WWII, the South African statistician John Kerrich: 10000 tosses, 5067 heads ($\hat{p} = 0.5067$)

  – Statistician Karl Pearson: 24000 tosses, 12012 heads ($\hat{p} = 0.5005$)

- Subjective beliefs

  Can be used for experiments that can't be repeated exactly, such as sporting event.

  For example, what is the probability that the Red Sox will win the World Series this year.

  Can be done through comparison (i.e. is getting a head on a single flip of a coin more or less likely, getting a 6 when rolling a fair 6 sided die, etc).

  Can also be done by comparing different possible outcomes (1.5 times more likely than the Yankees, 20 more likely than than the A's, 1,000,000 times more likely than the Rockies, etc).

  Often expressed in terms of odds

$$\text{Odds} = \frac{\text{Prob}}{1 - \text{Prob}} \qquad \text{Prob} = \frac{\text{Odds}}{1 + \text{Odds}}$$

- Models, physical understanding, etc.

  The structure of the problem will often suggest a probability model.

  For example, the physics of rolling a die suggest that no one side should be favoured, giving the uniform model discussed earlier.

  However this could be verified by looking at the long run frequencies.

  The probabilities used to describe Mendel's experiments come from current beliefs of the underlying processes of meiosis (actually approximations to the processes).