17) Statistical Models

We want a way of learning something about the real world using data.

This process involves building of statistical model of the real world, then estimating real world parameters of that model.

Example: 17.1 (Galileo)
Consider a ball rolling on an inclined plane. Let x be the variable giving distance travelled by the ball
t be the variable giving time.

The relationship between x and t is given by the mathematical equation:

$$x = \frac{1}{2}at^2$$

The parameter a appearing in the equation is the acceleration.
To determine value of the parameter, we need to perform the experiment. we let the ball roll and stop it after t seconds and measure distance travelled x in meters.

Then

$$a = \frac{2x}{t^2}$$

But there is a hitch: Each time the experiment is repeated, one gets a slightly different value for parameter a, due to random effects leading to errors both in time and distance measurements.

By repeating experiment n times, one obtains a dataset $x_1, x_2, \ldots, x_n$.

How do we determine the correct value of parameter a from these experiments:

To answers this question, we need to extend the simple deterministic relationship blw x and t to include uncertaintity arising from the random effects.

Such a model is called a statistical model.

The observations $x_1, x_2, \ldots, x_n$ are modelled as values of a random variable $X_1, X_2, \ldots, X_n$.

A possible model for Galileo's experiment would be

$$X_i = \frac{1}{2} a (t + U_i)^2 + V_i$$

The random variables U_i model the random error in the time measurement and the random variable V_i describes the random errors in distance measurement.

We assume that these errors are normally distributed:

$$U_i \sim N(0, \sigma_U^2)$$

$$V_i \sim N(0, \sigma_V^2)$$

Because time measurement is independant of distance measurement and because the repitions of the experiment are independant, we assume U's and V's are independant.

$$U_i \perp\!\!\!\perp V_i$$

With such a concrete model in place, it is now possible to estimate the parameters a from the observed data.

Lets abstract the notion of a statistical model from these examples:

Defn: A (parametric) statistical model (also referred to as probabilistic model or stochastic model) for a numerical dataset:

1) views the data as the values of a set of random variables

2) gives a partial specification of the joint probability distribution for these random variables (and possibly additional unobserved random variables), including in particular information on independance/dependance. This distribution is often referred to as the model distribution

3) contains a set of parameters of the distribution that are unknown but of interest, to be estimated from the dataset. These are model parameters

Statistical inference is the process of inferring the value of the model parameters from the observation.

In summary:

The way a mathematical statician learns information ~~from~~ data is to make a statistical model for the data and to use the data to deduce the value of the unknown parameters in the model