Probability and Statistics
for Final Year Engineering Students

By Yoni Nazarathy, Last Updated: April 11, 2011.

Lecture 1:
Introduction and Basic Terms

Welcome to the course, time table, assessment, etc..

Basic definitions:

- **Probability** – The mathematical study of random phenomena.
- **Probabilistic Model** – A mathematical model which allows calculation of probabilities (values between 0 and 1) and other related performance measures occurring in random environments.
- **Statistics** – The study of data.
- **Statistical Inference** – The process of using a **sample** obtained from a **population** to understand data properties of the population.
- A **statistic** – a quantity calculated from a sample which can also be modeled using a probabilistic model.

In this course the study of probability and statistics will be coupled closely.

An example to be used for the next few lectures:

A company is manufacturing robotic arms that are to be used in automated mail ordering systems for pick up and placement of ordered-items and placement in boxes. Ordered-items may be lifted by the arm by using one of two modes: (1) suction and (2) grabbing. Items which are to be picked up vary in size, shape and weight, some may be picked up only using mode (1), some only using mode (2) and some using both modes.

One prototype robotic arm which is already in beta-testing in a big automated ordering warehouse, utilizes an embedded algorithm for on-line decision making regarding the mode of pick up: as its electronic sensors view the ordered-item, an on-line decision is made choosing between pick up mode (1) or (2). The prototype arm records its actions and the results of the pick-up on disk, yielding massive data files which are then analyzed by the development team of the company, attempting to improve the performance of the robotic arm and cut down manufacturing costs.

Out of the many data traces which the prototype arm generates, here are two types of scalar data sets (one dimensional) which will be used for illustration in this course:

- Weights of ordered items: $X_1, X_2, \ldots, X_n$.
- Success indications of ordered items with regards to mode (1) (suction) pick-up: $I_1, I_2, \ldots, I_n$.

The first type of data set is measured in Kilograms while the second is a binary data set where 1 indicates success and 0 indicates failure.

One may think of many other types of data sets which may be generated and used for analysis in the robotic arm example, but for purposes of illustration, the first two will be used in the next few lectures. Sections 5s and 6s of the course will look at more complex examples.

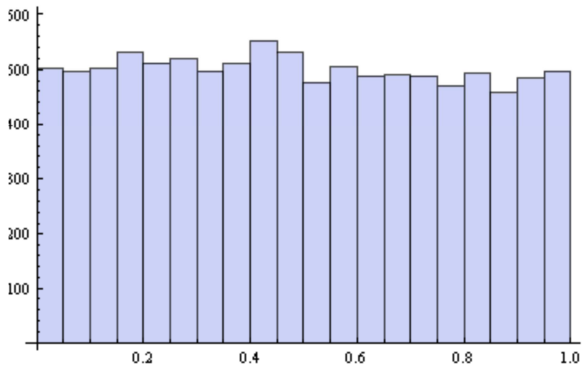
Basic simulation:

A computer may be used to generate a pseudo-random sequence of numbers. E.g. in Excel use the rand() function. The basic sequence contains numbers which are typically either uniformly distributed in the range [0,1] or are discrete numbers in the range 0,….,$2^M - 1$, where M is the number of bits in a computer word (e.g. 32). The basic idea is that the computer generates a deterministic (non-random) sequence based on some initial condtions (seed), yet the "statistical properties" of this sequence is such that it appears random.

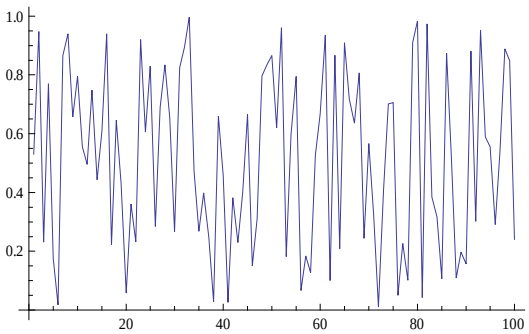
Here for example are 10 pseudo-random numbers generated in Mathematica:

```
{0.815508,0.965318,0.705951,0.13112,0.924486,0.706412,0.735212,0.996408,0.789244,0.625694}
```

And here is a histogram of 10,000 such random values:



Having a "uniform" histogram is not the only measure of quality of a pseudo-random sequence, there are many others, for example, one would like sequential numbers to be independent. Here is a plot of 100 generated numbers:



Generation of pseudo-random numbers is used for a variety of purposes, including

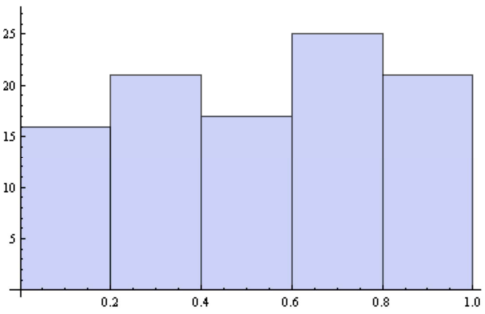
- Monte-Carlo simulation of real-life scenarios.
- Statistical procedures that use randomization.
- Checking of computer code on a variety of inputs.
- Optimization which requires arbitrary choices.
- Probabilistic algorithms.
- Generation of behaviors in computer games.

In this course we will use pseudo-random sequences to generate "fake" data-sets. We will also illustrate some probabilistic and statistical concepts using simulation.

Often, having random numbers (we will omit the "pseudo" prefix from now on), that are uniformly distributed in the range [0,1] is not exactly what we want. For example, we now wish to generate the data of 100 item weights. As a start, we generate 100 random numbers:

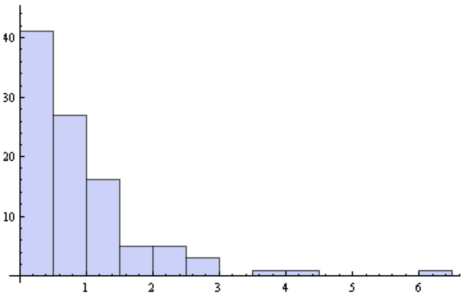
```
{0.317883,0.786902,0.937793,0.256601,0.27541,0.813025,0.553742,0.939716,0.168643,0.938936,0.03618
18,0.830915,0.971512,0.782318,0.15471,0.528937,0.0185043,0.192619,0.599493,0.238083,0.473812,0.30
3966,0.750707,0.997527,0.155929,0.517064,0.812914,0.740925,0.880519,0.704039,0.259172,0.801209,0.
711875,0.765104,0.22299,0.970294,0.740363,0.982786,0.0682806,0.441357,0.721859,0.790166,0.468787,
0.203274,0.248047,0.486201,0.71808,0.205747,0.0921179,0.969137,0.905165,0.464822,0.211599,0.26509
8,0.645993,0.663613,0.499724,0.499994,0.423003,0.693319,0.759361,0.517209,0.354723,0.251962,0.037
5023,0.727042,0.885935,0.0486883,0.789456,0.240841,0.167856,0.842941,0.697338,0.271704,0.26269,0.
37812,0.485739,0.00660644,0.616697,0.714507,0.986015,0.506612,0.193693,0.0211883,0.226654,0.98940
4,0.838971,0.769226,0.189152,0.262362,0.953035,0.720538,0.399696,0.0215204,0.78518,0.877597,0.702
358,0.749816,0.522489,0.499477}
```

A histogram of these numbers indicates to us (what we already know), that they are uniformly distributed in the range [0,1]:

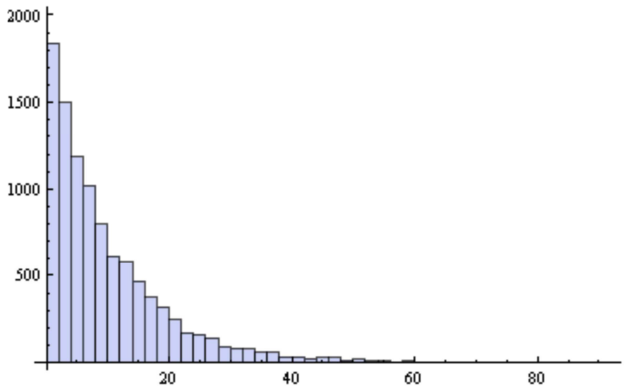


Yet, what if we would like to model weights which are in a different range? One option is to transform the original random variables by for example: $X_i = c\ U_i + d$. This would result in the numbers being in the range [d,c+d] but they would maintain a uniform distribution.

Another option is to make other types of transformations that yield different distributions. For example taking $X_i = -\ ln\ U_i$ generates a sequence which is essentially in the range $[0, \infty)$ and whose histogram is as follows:



For the sequence of weights of items (in the robot arm example) we will use the transformation $X_i = -10\ ln\ U_i$. Here is a histogram of 10,000 such weights:



Thus for our "learning process in this course", we will assume the weights come from a distribution such as the one above. Note that the "center of mass" of this distribution appears to be roughly at around 10. There are many items with weights less than 10 and a few items with weights much bigger than 10.

For the sequence of success indications, we wish to generate a random sequence of 0's or 1'. One way to do that is to make the following transformation $I_i = 1\{U_i \leq p\}$ for some p in the range [0,1]. Here $1\{A\}$ is the notation we use for the "indicator function", a function that takes 1 if the event A occurs and 0 if the event A does not occur. What proportion of the I_i's is going to be 1?

For example, of the 100 data points above, and on choosing p=0.2, we get that there are 16 I_i's which are 1 and the rest are 0 (one way to do/denote this is: $\sum_{i=1}^{100} I_i = 16$. Why is it not exactly 20?

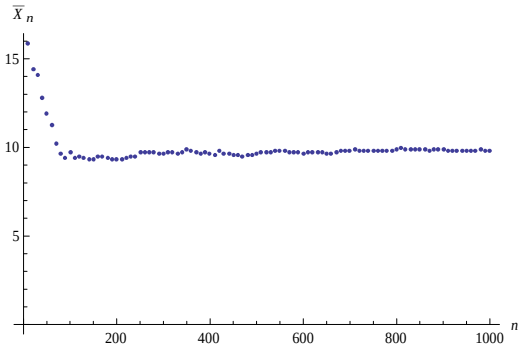
Basic "Statistics":

A "statistic" is a quantity calculated from a random sample, $X_1, X_2, \ldots, X_n$. Here are important examples:

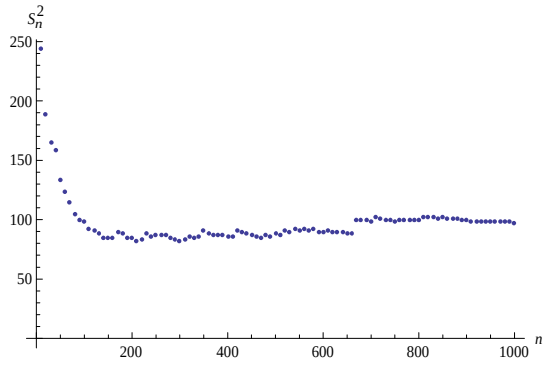
- The minimum or maximum, $Min(X_1, X_2, \ldots, X_n)$ or $Max(X_1, X_2, \ldots, X_n)$.
- The sample mean, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$.
- The sample variance, $S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$
- The sample proportion of values satisfying some property A: $\hat{p} = \frac{\sum_{i=1}^{n} 1\{X_i \in A\}}{n}$.

We often denote a subscript of n, next to the statistic indicating that it was calculated for a sample of n points. We now generate 1,000 weight data points (using the transformation mentioned above) and plot the statistics as increasing values of n. For the proportion statics we use A=[0,10].

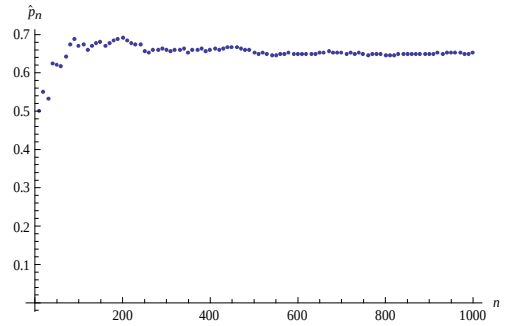
Here is a plot of the sample mean:



And here is the sample variance:



And here is the sample proportion:



For the final values (of n=1000) we got, $\bar{X} = 9.43,\ S^2 = 90.35$, and $\hat{p} = 0.651$. We see that as n grows to infinity these statistics appear to converge. To take a closer look at the values they converge to, take n=1,000,000, let the computer run for a bit and obtain the estimates: We obtain , $\bar{X} = 10.001,\ S^2 = 100.007$, and $\hat{p} = 0.6324$. The point here is that our sample size grows the estimates get closer and closer to the population values (which are actually 10, 100 and 1-1/e – explanation of these values follows in a few pages).

Knowing the population values is a general engineering aim. It will allow the designers of the robotic arm to make efficient design decisions. Estimating the population values from observed data is a basic statistical task. In general, more data is better, but typically data is costly to obtain. A basic question which we will answer (for specific cases) in the first three lectures is: What kind of accuracy in the estimation can we obtain with a given set of data, or alternatively, given a desired accuracy, how many observations do we need?

For brevity and illustration, we will mostly concentrate on the estimation of a proportion. For this we will use the data $I_1, I_2, \dots, I_n$ which indicates success (=1) or failure (=0) with regards to pick up using suction. I.e. we will use the example of estimation of a proportion as a case study.

Probability and random variables:

In order to understand the properties of "statistics" we need to make a probabilistic model. We now formularize some concepts of probability (in brevity).

A **probability** is a number in the range [0,1]. A **probability function** P(A), is a function which takes an **event** A, and returns a probability. For example, P(it will rain today).

In setting up a probability function we think of the situation at hand as "a random experiment". We denote the set of **all possible events** by Ω and have $P(\Omega) = 1$. This is sometimes called the **sample space**. We further denote the empty event (or **empty set**) by $\emptyset$ and have $P(\emptyset) = 0$.

A **random variable**, X, is a random outcome of the experiment typically getting a numeric (or vector) value. For example if the experiment is a temperature reading and X is measured in Kalvin then $\Omega = \{\mathrm{X} \geq 0\}$ and $\emptyset = \{\mathrm{X} < 0\}$.

The **distribution** of the random variable X is a description of the probabilities $P(X \in A)$ for different sets A.

It turns out that a suitable description of the distribution is to use sets A of the form (-∞,x] for increasing values of x. This defines the **cumulative distribution function (CDF**):

$$F(x) = P(X \leq x).$$

Note that $F(-\infty) = 0$ and $F(\infty) = 1$. Why?

Also, F() is a non-decreasing function. To understand this we need to see that

$$\text{if } A \subseteq B \; then \; P(A) \leq P(B).$$

The distribution obtained by generating a pseudo-random variable in the range [0,1] has the following CDF:

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1. \\ 1 & 1 < x \end{cases}$$

This the distribution of a **uniform random variable** in the range [0,1]. Try to understand this.

Let's think now of the first observation (I_1) in our data-set of pickup success as a random variable, denote it X for clarity. Here we have $\Omega = \{X = 0, X = 1\}$. Assume now that there is a number p in the range [0,1] such that P(X=1)=p. In this case how does the CDF F(x) look?

Two important classes of random variables are purely **continuous random** variables and purely **discrete random variables**. For continuous random variables F(x) is continuous (as in the case of the uniform random variable above) for discrete random variables, F(x) has jumps at the values which the random variable may take.

Continuous random variables may be described by the **density function** $f(x) = \frac{d}{dx}F(x)$ and discrete random variables may be described by the **probability mass function** $p(x) = P(X = x)$ (which can also be read off from F(x), how?). Both of these functions are sometimes referred to as PDFs. PDFs can be viewed as histograms and are often the most common way to view a distribution of a random variable.

We always have , $\sum_x p(x) = 1$ or $\int f(x)dx = 1$ where the sum or the integral are taken over the possible values which the random variable may obtain.

Example: Draw the PDF of the random variable relating to the result of the throw of a fair die.

Example: Draw the PDF of the random variable relating to the result of the sum of two die throws.

Example: It turns out (we won't prove it in this course) that if we let $Y = -\mu \ln U$ where U is a uniform random variable in the range [0,1], then Y has CDF: $F(x) = 1 - e^{-\frac{1}{\mu}x}$ for $0 \leq x$. (This is called an **exponential random variable**). What is the density? Show that $\int f(x)dx = 1$.

The **mean** of a random variable, denoted E[X], is the center of mass of the PDF. It is computed as follows:

$$E[X] = \sum_x x\, p(x) \qquad\qquad E[X] = \int x f(x) dx.$$

depending on if the distribution is discrete or continuous.

Example: The mean of the random variable I_1 is p.

Example: Find the mean value of a die throw.

Example: The mean of an exponential random variable:

$$E[X] = \int x f(x) dx = \int_0^\infty x\, \frac{1}{\mu} e^{-\frac{1}{\mu}x} dx = \frac{1}{\mu}.$$

The above can be done by integration by parts (try it!).

The **variance** of a random variable, denoted, Var(X) is a measure of spread:

$$Var(X) = \sum_x (x - E[X])^2\, p(x) \qquad\qquad Var(X) = \int (x - E[X])^2 f(x) dx.$$

Example: The variance of the random variable I_1 is p(1-p).

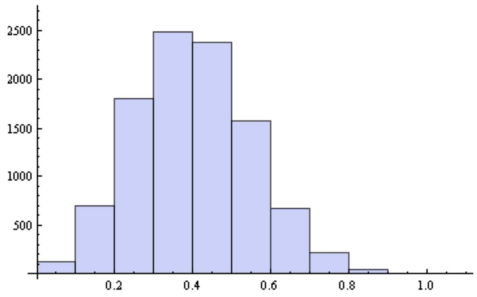
Example: The variance of an exponential random variable is μ^2.

Note the you should take care and not to confuse the mean and variance with the sample mean and sample variance. Yet, they are related in the sense that the sample mean and sample variance are estimates of the mean and the variance.

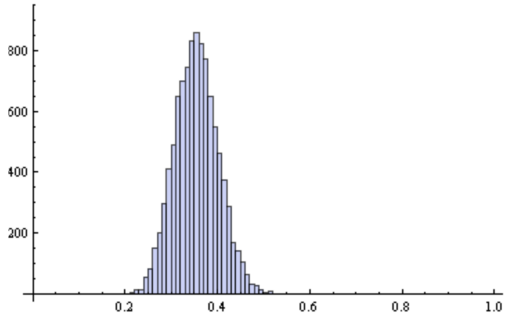
The distribution of a statistic:

Now that we have understood (had a very fast flight through it) the basics of probability modeling, let us revisit the estimation of a proportion again. Assume that we want to estimate the success probability of suction pick-ups, using the random sample $I_1, I_2, \dots, I_n$.

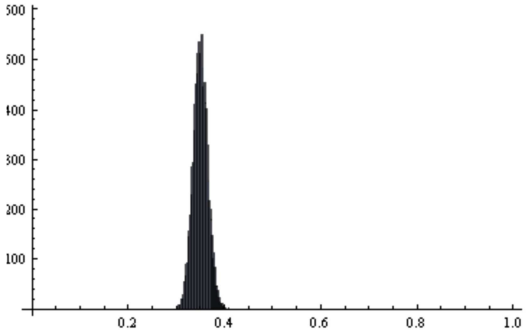
Our estimator is $\hat{p} = \frac{\sum_{i=1}^n I_i}{n}$, this is the observed proportion of pick-up successes. This estimator is actually a random variable and in the next lecture we will see how it is distributed analytically. For now let's do a simulation, showing us the distribution of $\hat{p}$. Fix n = 10. We now need to generate many instances of $\hat{p}$, say 10,000. This can be done by generating 10x10,000 I's and from each 10 getting a new sample of $\hat{p}$. Assume that the actual unknown p, is p=0.35. The results are plotted in the following histogram:



Repeating the same step, but taking n=100, we obtain the following histogram:



And with n=1,000 , we obtain the following histogram:



As can be expected the distribution of the statistic, $\hat{p}$ becomes more and more centered around the actual value of p, when n increases. With n=10, there is a non-negligible chance of obtaining "a completely wrong" value for the estimate $\hat{p}$, but as n increases the chance of getting a wrong value decreases – the next lecture will help quantify this.