6.0002

Lecture 7

Simulation Models:

- A description of computations that provide useful information about the possible behavior of the system.
- Descriptive, not prescriptive.
- Only an approximation of reality.

Uses:

- To model systems the are mathematically intractable, and extracting useful intermediate results.

Inferential Statistics : It is the process of making inferences about a population based on a sample.

Monte Carlo Simulations : It is a method of estimating value of an unknown quantity using principles of inferential statistics.

Confidence in a prediction is based on:

- Sample Size
- Variance in sample

As variance grows, we need larger samples to have the same degree of confidence.

Law of Large Numbers : The average pf the results obtained from a large number of trials should be close to the expected value and will tend to get closer to the expected value as more trials are performed.

Regression to the Mean : Follow an extreme random event , the next random event is less likely to be extreme.

After getting 10 continuous reds in a roulette, the probability of getting black chance would be relatively low, but not below the expected probability i.e. 5.

Gambler's Fallacy: The gambler's fallacy, also known as the Monte Carlo fallacy, occurs when an individual erroneously believes that a certain random event is less likely or more likely to happen based on the outcome of a previous event or series of events.

Thinking that after getting red 10 times, the probability of getting red becomes less in comparison with black.

$$Variance\ (x) = \frac{\sum_{x\ \epsilon\ X}\ (x - \mu)^2}{|X|}$$

$$Standard\ Deviation = \sigma(x) = \sqrt{\frac{\sum_{x\ \epsilon\ X}\ (x - \mu)^2}{|X|}}$$

Distributions : Captures the notion of a relative frequency with which a random variable takes on certain values.

Distributions are defined by probability density functions - probability of a random variable lying between two values.

Area under curve between two points is probability of example falling within that range.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} * e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

PDF for Normal Distribution

Lecture 8

Central Limiting Theorem (CLT):

- Given a sufficiently large sample:
  - Means of samples in a set of samples will be approximately normally distributed,
  - This Normal Distribution will have a mean closer to the mean of the population,
  - The variance of the *sample means* will be close to the variance of the population divided by the sample size.

Stratified Sampling: The population is divided into non-overlapping subgroups and sample is taken from each subgroup proportional to their size. This helps in reducing the sample size.

Applying CLT surely gives important results. The mean of means of samples is closer to the mean of population, but the standard deviation remains high.

To reduce the SD deviation of this, we increase the size of each sample. In comparison with increasing number of samples, this reduces the standard deviation by a much higher degree. However, with increasing size of samples, we are looking at data points whose number may exceed the size of population, itself.

At this point, the third part of CLT comes into play, we can calculate the variance of population, which be used to extrapolate the SD of a sample, hence the SD of the population.

Standard Error of Mean: Standard error of the mean (SEM) measures how far the sample mean (average) of the data is likely to be from the true population mean.

$$SE = \sigma / \sqrt{n}$$

3rd part of CLT stands true, when sample size reaches a reasonable value, the sample SD is a pretty good estimation of population SD.

Skewness is a measure of the asymmetry of the probability distribution

Estimating mean from a single Sample:

- Choose sample size based on estimate of skew in population,
- Choose a random sample from the population,
- Compute the mean and standard deviation of that sample,
- Use the standard deviation of that sample to estimate the SE,
- Use the estimated SE to generate confidence intervals around the sample mean.

Lecture 9

We are often presented with experimental data in real life, which, in themselves, don't make sense intuitively. To make some meaning of data, we can use data science.

Fitting Curve To Data

- When we fit a curve to a set of data, we are trying to find a fit that relates and independent variable to the estimated value of dependent variable.
- To decide how well a curve fits the data, we need a way to measure the *goodnesss* of the fit - an objective function. This will measure how well a certain curve comes close to the actual values, hence being an appropriate model.
- After defining the objective function, we want to find a curve that minimizes it. Hence, it acts as a beam balance to measure accuracy of the model/curve.

$$\sum_{i=0}^{len(observed)-1} = (observed[i] - predicted[i])^2$$

Least Square Objective Function

Assumption : Required curve is a polynomial curve.

- Use first degree polynomial - $ax + b$
- Find values of a and b for all the x values (independent values) in our experiment such that the objective function is minimized.
- For finding a and b, we use linear regression. We will use the *polyfit()* function for this.

```
pylab.polyfit(observedX, ObservedY, n)
```

n = 1 for line, 2 for parabola, and so on.

returns the value of a and b.

How good is a model, relative to each other?

A better way to find this is by using *coefficient of determination*, R^2.

$$R^2 = 1 - \frac{\sum_i (y_i - p_i)^2}{\sum_i (y_i - \mu)^2}$$

y *are measured values, p are predicted values, and mu is mean of measured values*.

The numerator measures the error in estimates and the denominator measures variability in measured data.

This basically gives us a value between 0 and 1, where 1 means a curve which covers all the data, and 0 covers none.

Why build models?

- Understanding properties of the data.
- Predict the future behavior based on the data.

Training Error

A model, probably of high polynomial order, might fit perfectly on the data it was based on. However, it might not work at all on other data. This is called training error.

To overcome this, we need to cross validate. We should use the model on training data of another model, and vice versa. A truly good fit will give good R2 results on any data set.

A model of high polynomial order, might give great results on training data, but it probably fits the noise rather than the underlying pattern in the data.

Choosing an overly complex model leads to overfitting to the training data. It renders the model useless on other data. Choosing an insufficiently complex model makes the model useless on any data.

Balancing Fit With Complexity

- Fit a low order model to training data,
- Test on new data and record R^2 value,
- Increase Order of model and repeat,
- Continue until fit on test data begins to decline.

In absence of solid theory, we will 2 methods to judge a model.

- Leave-one-out Cross Validation:
  1. For small data:

     Take the data, pop one element out. Create a model. Test the model for that one element. Average the test results of how far the results were from actual answers.

  2. For big data:

     Do the same, but take out a chunk of data, create model on test on others. Compare how far off the models are.
- Repeated Random Sampling:

  Randomly select some chunk. Make model with the rest. Test data on the first chunk. Do it for some number of times.