EEC 686/785
Modeling & Performance Evaluation of Computer Systems

Lecture 9

Wenbing Zhao
Department of Electrical and Computer Engineering
Cleveland State University
wenbing@ieee.org
(based on Dr. Raj jain's lecture notes)

Outline

- Review of lecture 8
- Simple linear regression models
- Review for midterm #1

Confidence Interval for The Mean

- k samples => k sample means => can't get a perfect estimate of μ from any finite number of finite size samples => use probabilistic bounds c_1 and c_2:

  $$\text{Probability}\{c_1 \le \mu \le c_2\} = 1\text{-}\alpha$$

- **Confidence interval** for the population mean: (c_1, c_2)
- **Significance level**: α
- **Confidence level**: 100(1-α)
- **Confidence coefficient**: 1-α

Determine Confidence Interval

- 100(1-α)% confidence interval for the population mean for large sample (n >= 30):

  $$(\bar{x} - z_{1-\alpha/2} s/\sqrt{n}, \bar{x} + z_{1-\alpha/2} s/\sqrt{n})$$

  $$z_{1-\alpha/2} = (1-\alpha/2)\text{-quantile of N(0,1)}$$

- 100(1-α)% confidence interval for $n < 30$:

  $$(\bar{x} - t_{[1-\alpha/2;n-1]} s/\sqrt{n}, \bar{x} + t_{[1-\alpha/2;n-1]} s/\sqrt{n})$$

  $$t_{[1-\alpha/2;n-1]} = (1-\alpha/2)\text{-quantile of a } t\text{-variate}$$

  with $n-1$ degree of freedom

Comparing Two Alternatives: Paired vs. Unpaired

- Paired: one-to-one correspondence between the *i*th test on system B
  - Example: performance on *i*th workload => use confidence interval of the difference
- Unpaired: no correspondence
  - Example: *n* people on system A, *n* on system B => need more sophisticated method

Approximate Visual Test

- To compare two unpaired samples, simply compute the confidence interval for each alternative separately

Confidence Intervals for Proportions

- Proportion = probability of various categories
  - E.g., P(error)=0.01, P(no error)=0.99
- n_1 for n observations are of type 1 =>
  - Sample proportion = $p = n_1/n$
  - Confidence interval for the proportion $p \mp z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}$
- Assumes normal approximation of Binomial distribution => Valid only if $np \geq 10$
- Need to use binomial tables if $np < 10$, can't use *t*-values

Sample Size for Determining Mean

- Larger sample => Narrower confidence interval => Higher confidence
- Question: How many observations n to get an accuracy of $\pm r$% and a confidence level of 100(1-α)%?
  - The 100(1-α)% confidence interval of the population mean is

$$\bar{x} \mp z\frac{s}{\sqrt{n}}$$

- r% accuracy => $CI = (\bar{x}(1-r/100), \bar{x}(1+r/100))$

=> $\bar{x} \mp z\frac{s}{\sqrt{n}} = \bar{x}\left(1 \mp \frac{r}{100}\right) \Rightarrow z\frac{s}{\sqrt{n}} = \bar{x}\frac{r}{100} \Rightarrow n = \left(\frac{100zs}{r\bar{x}}\right)^2$

Regression Models

- A **regression model** allows one to estimate or predict a random variable as a function of several other variables
- **Response variable**: the estimated variable
- **Predictor variables**, predictors, or factors: the variables used to predict the response
- Regression analysis assumes that all predictors variables are quantitative so that arithmetic operations are meaningful
- **Simple linear regression mode**: linear regression model with a single predictor variable

Definition of a Good Model

- Regression models should minimize the distance measured vertically between the observation point and the model line
- The difference between the observed response and the predicted response is called **residual**, **modeling error**, or simply **error**

Definition of a Good Model

- Two requirements
  - Zero overall error, i.e., negative and positive errors cancel out
  - To choose the line that minimizes the sum of squares of the errors

Definition of a Good Model

- Linear model: $\hat{y} = b_0 + b_1 x$
  - $\hat{y}$ is the predicted response when the predictor variable is x
  - b_0 and b_1 are fixed regression parameters to be determined from the data
- Given n observation pairs $\{(x_1,y_1),\ldots,(x_n,y_n)\}$, the estimated response $\hat{y}_i$ for the ith observation is

$$\hat{y}_i = b_0 + b_1 x_i$$

The error is:

$$e_i = y_i - \hat{y}_i$$

Definition of a Good Model

- The best linear model is given by the regression parameter values, which minimizes the Sum of Squared Errors (SSE): $$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$
- Subject to the constraint that the mean error is zero
$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0$$
- This constrained minimization problem is equivalent to minimizing the variance of errors

Estimation of Model Parameters

- The regression parameters that give minimum error variance are
$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} \quad b_0 = \bar{y} - b_1\bar{x}$$
where
$$\bar{x} = \text{mean of the values of precitor variables} = \frac{1}{\text{n}}\sum_{i=1}^{n} x_i$$
$$\bar{y} = \text{mean response} = \frac{1}{n}\sum_{i=1}^{n} y_i$$
$$\sum xy = \sum_{i=1}^{n} x_i y_i \quad \sum x^2 = \sum_{i=1}^{n} x_i^2$$

Derivation

- Error in the *i*th observation: $e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$
- Mean error:
$$\bar{e} = \frac{1}{n}\sum_i e_i = \frac{1}{n}\sum_i \{y_i - (b_0 + b_1 x_i)\} = \bar{y} - b_0 - b_1\bar{x}$$
- Setting mean error to 0 => $b_0 = \bar{y} - b_1\bar{x}$
- Substituting b_0 =>
$$e_i = y_i - \bar{y} + b_1\bar{x} - b_1 x_i = (y_i - \bar{y}) - b_1(x_i - \bar{x})$$
- Sum of squared errors:
$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [(y_i - \bar{y})^2 - 2b_1(y_i - \bar{y})(x_i - \bar{x}) + b_1^2(x_i - \bar{x})^2]$$

Derivation

$$\frac{SSE}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}[(y_i - \bar{y})^2] - 2b_1\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x}) + b_1^2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
$$= s_y^2 - 2b_1 s_{xy}^2 + b_1^2 s_x^2$$

- Find the value of b_1 that gives min SSE:
$$\frac{1}{n-1}\frac{d(SSE)}{db_1} = -2s_{xy}^2 + 2b_1 s_x^2 = 0$$
- => $$b_1 = \frac{s_{xy}^2}{s_x^2} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2}$$

Example 14.1

- Number of disk I/O's and processor times of 7 programs were measured as {(14,2), (16,5), (27,7), (42,9), (39, 10), (50, 13), (83, 20)}
- Linear model to predict CPU time as a function of disk I/O's, given $n = 7$, $\sum xy = 3375$, $\sum x = 271$, $\sum x^2 = 13{,}855$, $\sum y = 66$, $\sum y^2 = 828$, =38.71, and = 9.43

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = \frac{3375 - 7 \times 38.71 \times 9.43}{13{,}855 - 7 \times (38.71)^2} = 0.2438$$

$$b_0 = \bar{y} - b_1\bar{x} = 9.43 - 0.2438 \times 38.71 = -0.0083$$

Example 14.1

- The linear model:
  CPU time = 0.0083 + 0.2438(number of disk I/O's)

Example 14.1

Allocation of Variation

- The purpose of a model is to be able to predict the response with minimum variability
- Without a regression model, one can use the mean response as the predicted value for all values of the predictor variables

Allocation of Variation

- The errors in this case would be larger than those with the regression model, i.e., the error variance = variance of the response

$$\text{Error} = \varepsilon_i = \text{observed response} - \text{predicted response} = y_i - \bar{y}$$

$$\text{Variance of errors without regression} = \frac{1}{n-1}\sum_{i=1}^{n}\varepsilon_i^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$= \text{variance of } y$$

Allocation of Variation

- **Total sum of squares** (SST): $\sum_{i=1}^{n}(y_i - \bar{y})^2$
  - It is a measure of y's variability and it is called the **variation** of y
  - $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (\sum_{i=1}^{n} y_i^2) - n\bar{y} = SSY - SS0$
- SSR: sum of squares explained by the regression, it is the difference between SST and SSE:

$$SSR = SST - SSE \quad \text{or} \quad SST = SSR + SSE$$

Allocation of Variation

- **Coefficient of determination** R^2: the fraction of the variation that is explained determines the goodness of the regression

$$\text{Coefficient of determination} = R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}$$

  - The higher the value of R^2, the better the regression
  - R^2 is also the square of the sample correlation R_{xy} between the two variables

$$\text{Sample correlation}(x, y) = R_{xy} = \frac{s_{xy}^2}{s_x s_y}$$

$$\text{Coefficient of determination} = \{\text{correlation coefficient}(x, y)\}^2$$

Allocation of Variation

- Shortcut formula for SSE:

$$SSE = \sum y^2 - b_0 \sum y - b_1 \sum xy$$

- Example 14.2: for the disk I/O-CPU time data of ex 14.1. Calculate the coefficient of determination

$$SSE = \sum y^2 - b_0 \sum y - b_1 \sum xy = 828 + 0.0083 \times 66 - 0.2438 \times 3375 = 5.97$$

$$SST = SSY - SS0 = \sum y^2 - n(\bar{y})^2 = 828 - 7 \times (9.43)^2 = 205.71$$

$$SSR = SST - SSE = 205.71 - 5.87 = 199.84$$

$$R^2 = \frac{SSR}{SST} = \frac{199.84}{205.71} = 0.9715$$

Standard Deviation of Errors

- **Mean squared error** (MSE): an estimate of the variance of errors = SSE divided by its degree of freedom: $$s_e^2 = \frac{SSE}{n-2}$$
  - Degree of freedom for SSE is $n - 2$ because errors are obtained after calculating two regression parameters from the data
- **Stand deviation of errors**: square root of MSE

Standard Deviation of Errors

- Degree of freedom
  - $n - 2$ for SSE
  - n for SSY since it is obtained from n independent observations without estimating any parameters
  - 1 for SS0 since it can be computed simply from $\bar{y}$
  - $n - 1$ for SST since one parameter $\bar{y}$ must be calculated from the data before SST can be computed
  - 1 for SSR
  - Thus the degrees of freedom add just the way the sums of squares do: $SST = SSY - SS0 = SSR + SSE$

$$n-1 = \quad n-1 \quad = 1 \quad + \ (n-2)$$

Example 14.3

- For the disk I/O-CPU data of example 14.1, the degrees of freedom of the sums are:

Sums of squares: $SST = SSY - SS0 = SSR + SSE$

$$205.71 = 828 - 622.29 = 199.84 + 5.87$$

Degrees of freedom: $6 = \quad 7 - 1 \quad = 1 + 5$

The mean squared error is:

$$\text{MSE} = \frac{\text{SSE}}{\text{degree of freedom for errors}} = \frac{5.87}{5} = 1.17$$

The standard deviation of errors is: $s_e = \sqrt{MSE} = \sqrt{1.17} = 1.08$

Confidence Intervals for Regression Parameters

- The regression coefficients are random in the same manner as the sample mean or any other parameter computed from a sample
- Using a single sample, only probabilistic statements can be made about true parameters β_0 and β_1 of the population => true model is

$$y = \beta_0 + \beta_1 x$$

  - b_0 and b_1 are "statistics" (mean values) corresponding to the parameters β_0 and β_1, respectively

Confidence Intervals for Regression Parameters

- The standard deviation of b_0 and b_1

$$s_{b_0} = s_e \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - n\bar{x}^2} \right]^{1/2}$$

$$s_{b_1} = \frac{s_e}{\left[\sum x^2 - n\bar{x}^2 \right]^{1/2}}$$

Confidence Intervals for Regression Parameters

- The 100(1-α)% confidence intervals for b_0 and b_1 can be computed using the 1-α/2 quantile of a t variate with n-2 degrees of freedom, i.e., $t_{[1-\alpha/2;n-2]}$
- The confidence intervals are:

$$b_0 \mp t s_{b_0} \quad \text{and} \quad b_1 \mp t s_{b_1}$$

- If confidence interval includes 0, then the regression parameter cannot be considered different from zero at the 100(1-α)% confidence level

Example 14.4

- For the disk I/O-CPU data of example 14.1, we can calculate the standard deviations of b_0 and b_1:

$$s_{b_0} = s_e \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - n\bar{x}^2} \right]^{1/2} = 1.0834 \left[\frac{1}{7} + \frac{(38.71)^2}{13,855 - 7 \times (38.71)^2} \right]^{1/2} = 0.8311$$

$$s_{b_1} = \frac{s_e}{\left[\sum x^2 - n\bar{x}^2 \right]^{1/2}} = \frac{1.0834}{\left[13,855 - 7 \times (38.71)^2 \right]^{1/2}} = 0.0187$$

Example 14.4

- From table A.4, the 0.95-quantile of a t-variate with 5 degrees of freedom is 2.015 => 90% confidence interval for b_0 is

$$-0.0083 \mp (2.015)(0.8311) = -0.0083 \mp 1.6747 = (-1.6830, 1.6663)$$

  - Confidence interval includes 0 => b_0 is essentially 0
- Similarly, 90% confidence interval for b_1 is

$$0.2438 \mp (2.015)(0.0187) = 0.2438 \mp 0.0376 = (0.2061, 0.2814)$$

  - Confidence interval does not include 0 => the slope b_1 is significantly different from 0 at this confidence level

Case Study 14.1

- Performance comparison of RPC mechanism on UNIX and ARGUS

Case Study 14.1

UNIX data

ARGUS data

- The linear models are:

  Time on UNIX = 0.030 (data size in bytes) + 24

  Time on ARGUS = 0.034 (data size in bytes) + 30

Case Study 14.1

- Confidence intervals for intercepts overlap while those of the slopes do not
  - Setup times are not significantly different
  - Per-byte times (slopes) are different

Confidence Intervals for Predications

- Developing regression helps predict the value of the response variable for those values of predictor variables beyond the currently measured
- Given the regression equation, we can predict the response for any given value of predictor variable:

$$\hat{y}_p = b_0 + b_1 x_p$$

  - This formula gives only the mean value of the predicted response based upon the sample

37

Confidence Intervals for Predications

- Standard deviation of the mean of a future sample of *m* observations is

$$s_{\hat{y}_{mp}} = s_e\left[\frac{1}{m}+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum x^2 - n\bar{x}^2}\right]^{1/2}$$

  - Standard deviation of a single future observation (*m*=1)

$$s_{\hat{y}_{mp}} = s_e\left[1+\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum x^2 - n\bar{x}^2}\right]^{1/2}$$

  - Standard deviation of the mean of a large number of future observations at x_p (m=∞):

$$s_{\hat{y}_{mp}} = s_e\left[\frac{1}{n}+\frac{(x_p-\bar{x})^2}{\sum x^2 - n\bar{x}^2}\right]^{1/2}$$

38

Confidence Intervals for Predications

- Confidence intervals for predications from regression models
  - Standard deviation of the prediction is minimal at the center of the measured range

39

Example 14.5

- Again, using the disk I/O-CPU time data of example 14.1, estimate the CPU time for a program with 100 disk I/O's
- Regression equation:

CPU time = $-0.0083+0.2438$(number of disk I/O's)

- Therefore, for a program with 100 disk I/O's, the mean CPU time is:

CPU time = $-0.0083+0.2438(100)=24.3674$

Standard deviation of errors $s_e = 1.0834$

40

Example 14.5

- Standard deviation of the predicted mean of a large number of observations is:

$$s_{\hat{y}_p} = 1.0834\left[\frac{1}{7}+\frac{(100-38.71)^2}{13,855-7(38.71)^2}\right]^{1/2} = 1.2159$$

- 90% confidence interval for predicted mean = 24.3674±(2.015)(1.2159)=24.3674±2.4500=(21.9174,26.8174)

Example 14.5

- Standard deviation of the predicted CPU time of a single future program with 100 disk I/O's:

$$s_{\hat{y}_{1p}} = 1.0834\left[1+\frac{1}{7}+\frac{(100-38.71)^2}{13{,}855-7(38.71)^2}\right]^{1/2} = 1.6286$$

- 90% confidence interval for predicted mean = 24.3674±(2.015)(1.6286)=24.3674±3.2816=(21.0858,27.6489)
  - => confidence interval is wider than that for the mean of a large number of observations

Visual Tests for Verifying the Regression Assumptions

- The true relationship between the response variable *y* and the predictor variable *x* is linear
- The predictor variable *x* is nonstochastic and it is measured without any error
- The model errors are statistically independent
- The errors are normally distributed with 0 mean and a constant standard deviation

Visual Tests for Verifying the Regression Assumptions

- **Linear relationship**: prepare a scatter plot of *y* versus *x*
  - Any nonlinear relationship can be easily seen from this plot

Visual Tests for Verifying the Regression Assumptions

- **Independent errors**: after the regression, compute errors and prepare a scatter plot of ε_i vs. the predicted response $\hat{y}_i$
  - Any visible trends in the plot would indicate a dependence of errors on the predictor variable

Visual Tests for Verifying the Regression Assumptions

- **Independent errors**: additionally, you can plot the residuals as a function of the experiment number
  - Any trend up or down indicate the presence of other factors, environmental conditions, or side effects that varied in different experiments

Visual Tests for Verifying the Regression Assumptions

- Normally distributed errors: prepare a normal quantile-quantile plot of errors
  - If the plot is approximately linear, the assumption is satisfied

Visual Tests for Verifying the Regression Assumptions

- **Constant standard deviation of errors**: also known as **homoscedasticity**. Observe the scatter plot of errors vs. predicted response prepared for the independence test
  - If the spread in one part of the graph seems significantly different than that in other parts, then the assumption is not valid

Example 14.6

- For the same data in example 14.1, check assumptions:
  - Independence check: ok

Example 14.6

- Check for normality assumption: ok

Example 14.6

- Check homoscedasticity: some trend towards lower values of $\hat{y}$. However, the magnitude is small

Example 14.7

- For the RPC performance study in case study 14.1
  - The residual-vs-$\hat{y}$ plot for the ARGUS data shows a higher spread on the rhs of the graph than that on the left side => can be a concern

Example 14.7

- The normal quantile-quantile plot for the same residuals => departure from normality is higher than that in previous example

A Systematic Approach to Performance Evaluation

1. State goals and define the system
2. List services and outcomes
3. Select metrics
4. List parameters
5. Select factors to study
6. Select evaluation technique
7. Select workload
8. Design experiments
9. Analyze and interpret data
10. Present results

Commonly Used Performance Metrics

- Response time, Reaction time
- Turnaround time
- Stretch factor
- Throughput
- Capacity (nominal, typical, knee)
- Efficiency
- Utilization
- Reliability, availability

Selecting Evaluation Techniques

- The life-cycle stage is the key
- Other considerations are:
  - Time available, tools available,
  - Accuracy required,
  - Trade-offs to be evaluated,
  - Cost,
  - Salability of results

Selecting Metrics

- Include:
  - Performance:
    - Time: responsiveness
    - Rate: throughput or productivity
    - Resource: utilization
  - Error rate, probability
  - Time to failure and duration
- Consider including:
  - Mean and variance
  - Individual and global
- Selection criteria:
  - Low-variability
  - Non-redundancy
  - completeness

Types of Work Load

- Test workload
  - Real workload
  - Synthetic workload
- Test workload used
  - Addition instruction
  - Instruction mixes
  - Kernels
  - Synthetic programs
  - Application benchmarks

The Art of Workload Selection

- Services exercised
  - SUT, CUT
  - Metrics and workload depend on the system
  - The requests at the service-interface level of the SUT should be used to specify or measure the workload
- Level of detail
  - Most frequent request
  - Frequency of request types
  - Time-stamped sequence of requests
  - Average resource demand
  - Distribution of resource demands
- Timeliness; Loading level; Impact of other components; Repeatability

Monitors

- **Monitor**: a tool used to observe the activities on a system
- Usage:
  - To improve software performance. Find frequently used segments of the software
  - To measure resource utilizations and to find the performance bottleneck
  - To tune the system
  - To characterize the workload
  - To find model parameters, to validate models, and to develop inputs for models

Monitor Terminology

- **Event**
- **Trace**
- **Overhead**
- **Domain**
- **Input rate**
- **Resolution**
- **Input width**

Monitor Classification

- **By implementation**
  - Software monitor
  - Hardware monitor
  - Firmware monitor
  - Hybrid monitor
- **By trigger mechanism**
  - Event-driven: good for rare events
  - Timer-driven (sampling monitor): good for frequent events
- **By result display ability**
  - Online monitors: display the system state continuously
  - Batch monitors: collect the data that can be analyzed later

Two Special Software Monitors

- **Program-Execution Monitor**: Software monitors designed to observe application software
- **Accounting Logs**: resource usage
  - **Per activation**: average resource consumption per activation of the program
  - **Percentage of total**: resource consumed by all activations of a particular program, expressed as a percentage of the resources consumed by all activations of all programs
  - **Per-second resource-consumption rates**: divide the resource consumption by the elapsed time
  - **Per-CPU-second resource consumption**: divide the resource consumption by the CPU time consumed

Terminology

- **Capacity planning**: ensuring that adequate computer resources will be available to meet the future workload demands in a cost-effective manner while meeting the performance objectives
- **Capacity management**: ensuring that the currently available computing resources are used to provide the highest performance
- Capacity management is concerned with the present while capacity planning is concerned with the future
- **Benchmarking**: to compare the performance of two competing systems in an objective manner, benchmarks are run on these systems using automatic load drivers

Guidelines for Preparing Good Graphic Charts

- **Require minimum effort from the reader**
- **Maximize information**: there should be sufficient information on the graph to make it self-sufficient
- **Minimize ink**: present as much information as possible with as little ink as possible. Too much unnecessary info on a chart makes it cluttered and uninteresting
- **Use commonly accepted practices**: present what people expect
- **Avoid ambiguity**: show coordinate axes, scale division, and origin. Identify individual curves and bars. Do not plot multiple variables in the same chart

Gantt Charts

- **Gantt chart** can be used to show the relative duration of any number of Boolean conditions, i.e., conditions that are either true or false
  - A resource being used or being idle is an example of a Boolean condition
  - Each condition is shown to be a set of horizontal line segments
  - The total length of the line segments represents the relative duration of the condition
  - The position of various segments is arranged such that the overlap between different lines represents the overlap between the conditions

Kiviat Graphs

- **Kiviat graph**: a circular graph in which several different performance metrics are plotted along radial lines
  - In the most popular version of the graph, an even number of metrics are used.
  - Half of these metrics are HB metrics so that a higher value of the metrics is considered better.
  - The other half of the metrics measure are LB metrics, and a lower value is considered better
  - **Kiviat graph for an ideal system is star**

Ratio Games

- **Ratios** provide good opportunities for playing performance games with competitors
- Ratios have a *numerator* and *denominator* (*base*)
- **Two ratios with different bases are not comparable**
- However, many examples in published literature where computer scientists have knowingly or unknowingly compared ratios with different bases
- **Ratio game**: the technique of using ratios with incomparable bases and combining them to one's advantage

Strategies for Winning a Ratio Game

- It is better to use your opponent's system as base for HB metric

Basic Probability and Statistics Concepts

- **Independent events**
- **Random variables**
- **Cumulative Distributed Function** (CDF): $F_x(a) = P(x \le a)$
- **Probability Density Function**: $f(x) = dF(x)/dx$
- **Probability Mass Function**: $f(x_i) = p_i$
- **Mean or Expected Value** $\mu = E(x) = \sum_{i=1}^{n} p_i x_i = \int_{-\infty}^{+\infty} x f(x) dx$
- **Variance (σ^2)**: $Var(x) = E[(x-\mu)^2]$
- **Standard Deviation: σ**

Basic Probability and Statistics Concepts

- **Coefficient of Variation** (C.O.V.): σ/μ
- **Covariance**: $Cov(x,y) = \sigma^2_{xy} = E[(x-\mu_x)(y-\mu_y)] = E(xy) - E(x)E(y)$
- **Correlation Coefficient**: $\rho_{xy} = \sigma^2_{xy}/\sigma_x\sigma_y$
- **Quantile**: $P(x \le x_\alpha) = F(x_\alpha) = \alpha$
- **Medium**: the 50-percentile
- **Mode**
- **Normal distribution**: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}, \quad -\infty \le x \le +\infty$

Selecting among the Mean, Median and Mode

Indices of Dispersion

- **Indices of dispersion**: variability is specified using one of the following measures
  - Range – minimum and maximum of the values observed
  - Variance or standard deviation
  - 10- and 90- percentiles
  - Semi-interquantile range (SIQR)
  - Mean absolute deviation

Selecting the Index of Dispersion