Statistical Analysis (I): Estimation and Testing

Manasa Mandava

Indian School of Business

About myself

Electrical Engineering (IIT Bombay)

PhD (Stony Brook University, New York)

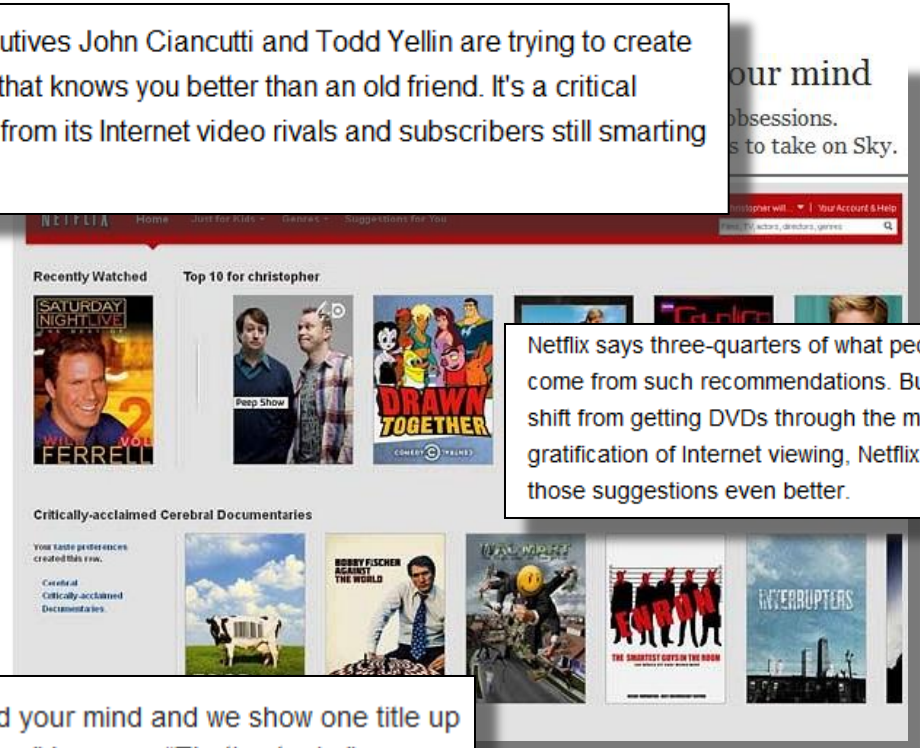
Operations Research

B-School Faculty (ISB)

Research in ‘optimal control in resource constrained settings’

Recommendations at Netflix

LOS GATOS, Calif. – Netflix executives John Ciancutti and Todd Yellin are trying to create a video-recommendation system that knows you better than an old friend. It's a critical mission as Netflix faces pressure from its Internet video rivals and subscribers still smarting from recent price hikes.



Netflix says three-quarters of what people watch now come from such recommendations. But as subscribers shift from getting DVDs through the mail to the instant gratification of Internet viewing, Netflix needs to make those suggestions even better.

"We'll be finished when I could read your mind and we show one title up there and it's exactly right every time," he says. "That's utopia."

- Challenge: Recommend 10 titles from about 35,000 based on customer's recent history
- Metric: Likelihood of customer choosing one of the recommended ones

Managerial Decisions

How many programmers should I staff for this project?

Where should we open our new retail store?

What is the right level of inventory for our new e-reader?

Which consulting company should we hire?

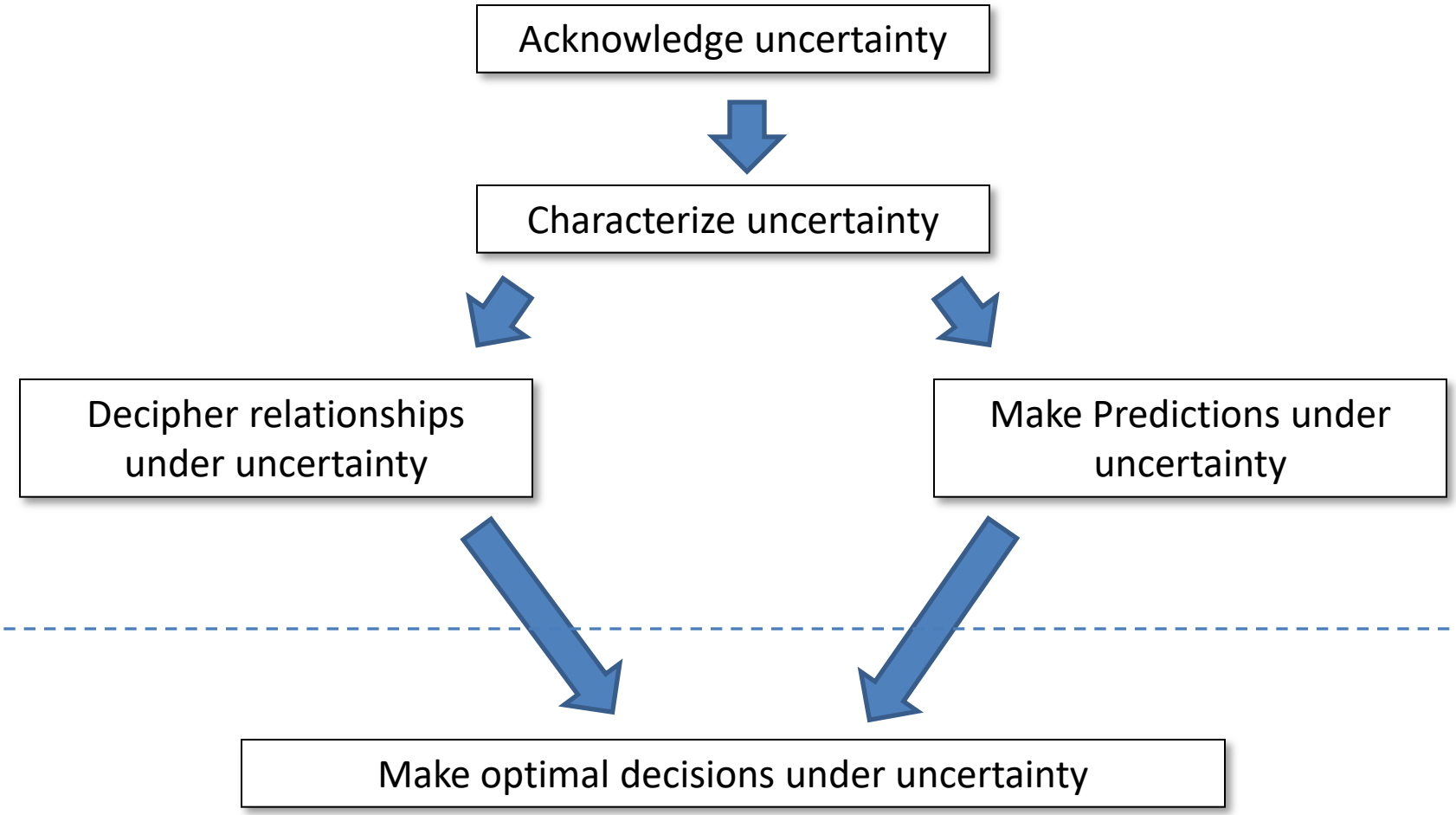
How much should we pay to acquire this business?

How much should we invest in online advertising?

What interest rate should we charge for this loan?

ISB

Overall goals of the course

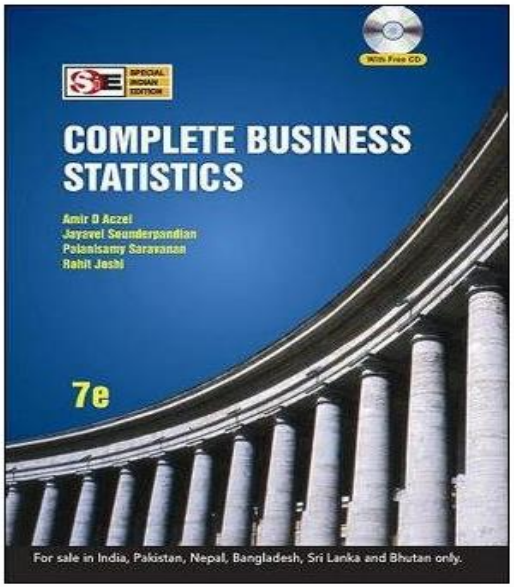


Course Summary

- Descriptive Statistics: Getting a better sense of data
  - Mean, Standard Deviation, Median, Quartiles, Distribution

- Inferential Statistics: Drawing conclusions about the population based on sample data
  - Properties of a single variable

Reference book



- Textbook is much more exhaustive than what we will cover this week
- The best use of the book is as a **reference**, go to specific sections of chapter where you need more clarity
- First solve the exercises from practice problem sets and textbook before thinking of more practice problems

Software for Class (R, RStudio)

- Open source, free, becoming increasingly popular
- Used sporadically in lectures, needed for working through the assignment
  - Important to “get hands dirty” to learn Stats
- You will not be required to analyze data using the software in exams
  - However, you should know how to interpret results of analysis presented by the software
  - You do need to use software to do the two assignments

ISB

Resources

- Class handouts
  - Lecture slides

- Learning Management System (LMS)
  - Assignment
  - Datasets
  - Practice Problem sets and their solutions

ISB

Course Policy

- Grading
  - Quizzes (2) (see LMS for schedule) 20%
  - Assignments (2) (see LMS for schedule) 20%
  - Midterm 30%
  - Final 40%

- Your responsibilities
  - Be on time
  - Participate meaningfully in class
  - Do not disrupt others
  - Bring name-cards to each class
  - Solve practice exercises posted on LMS.
  - Abide by the honor code at all costs

- No make up for missed quizzes/exam/assignment

ISB

Help / Guidance

- Office hours (By appointment)
  - AC 4, Level 1, #4117
- Email (reasonably responsive but not superfast)
- Office phone is not a preferred mode of communication
- Academic Associates
  - Yogesh Khandelwal

Session 1

Descriptive Statistics and Probability Distributions

ISB

Learning objectives

- What is a random variable and a probability distribution?
- What is a normal distribution and what are its properties?
- How to approximate data using normal distribution model?
- How to calculate descriptive statistics of linear combinations of random variables?

Data comes in many flavors …

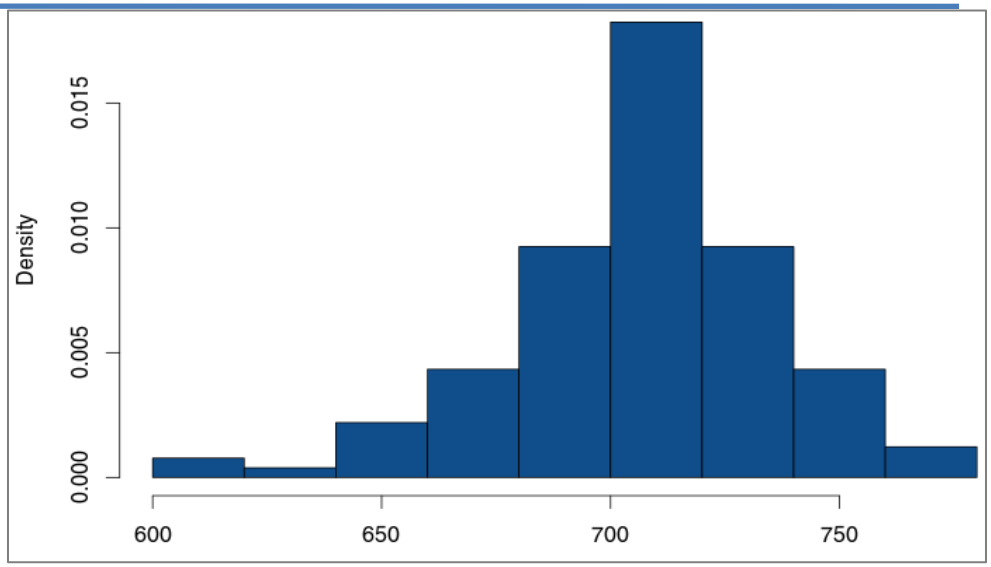
Type of data	Definition	Example
Nominal	Categories	Your previous degree
Ordinal	Can be ranked / ordered but not measured	Business school rankings
Interval scale	Intervals are meaningful but not ratios	Temperature in Fahrenheit or Celsius
Ratio scale	Ratios are meaningful	Sales of a new product

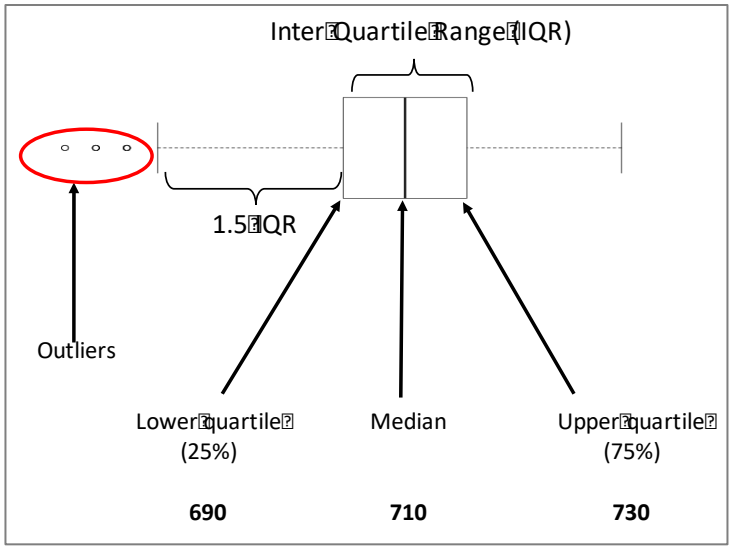
Source of data	Definition	Example
Observational	Analyst does not control data generating process	Stock returns on BSE
Experimental	Analyst has good control over data generation	Drug efficacy in clinical trials

Descriptive Statistics (Revisited)

610	730	590	610	.	.	.	680	630
640	680	540	660	.	.	.	610	540
690	610	520	640	.	.	.	720	680
610	650	660	580	.	.	.	600	730
710	600	760	690	.	.	.	500	720
610	650	660	710	.	.	.	480	600
630	610	680	780	.	.	.	700	690
530	550	730	690	.	.	.	670	540
630	720	610	710	.	.	.	600	600
690	600	730	540	.	.	.	560	770

Data File: mba.csv





Quantiles

100.0	maximu	790
99.5%		780
97.5%		750
90.0%		720
75.0%	quartile	680
50.0%	median	640
25.0%	quartile	600
10.0%		550
2.5%		490
0.5%		416.25
0.0%	minimum	370

Summary Statistics

Mean	638.6326
Std Dev	65.966024
Std Err Mean	2.451608
Upper 95% Mea	643.44572
Lower 95% Mea	633.81948
N	724

Random Variable

- A random variable describes the probabilities for an uncertain future numerical outcome of a random process
- It is a variable because it can take one of several possible values
- It is random because there is some chance associated with each possible value
- Examples?

- Probability
  - Long run average of a random event occurring
  - Different from subjective "beliefs"

Probability Distribution: Discrete and Continuous

- A probability distribution is a rule that identifies possible outcomes of a random variable and assigns a probability to each

- A discrete distribution has a finite number of values
  - e.g. face value of a card, work experience of students rounded off to nearest month

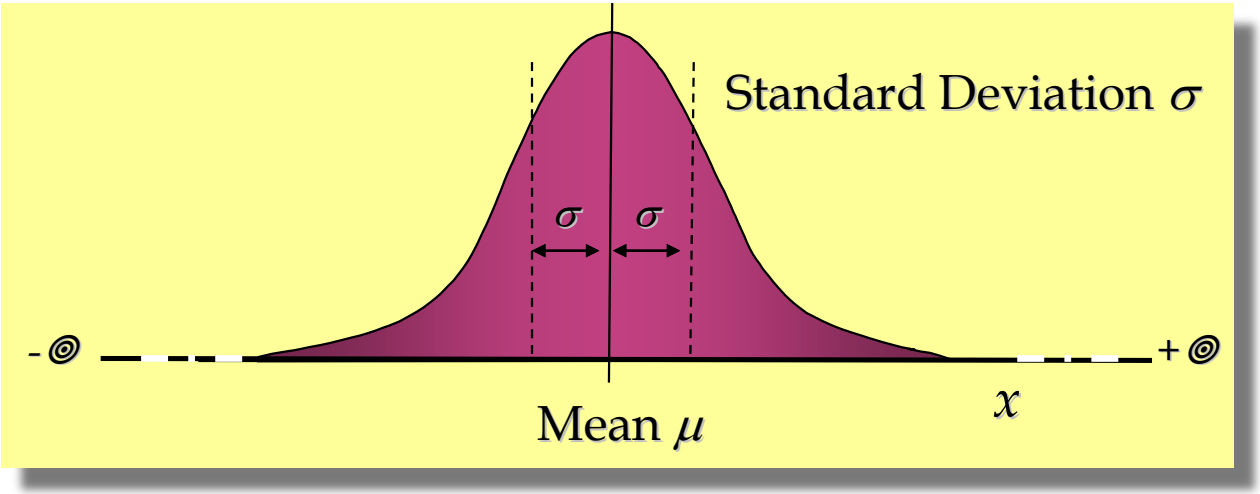
- A continuous distribution has all possible values in some range
  - e.g. sales per month, height of students in this class

- Continuous distributions are nicer to deal with and are good approximations when there are a large number of possible values

Expected Value (Mean), Variance & Standard Deviation

- The expected value or the mean of a random variable is a weighted sum of its values
  - The probabilities serve as weights
  - $\text{Mean}(\mu) = E(X) = \sum_i x_i P(X = x_i)$
- Variance (σ^2): The weighted sum of the squared deviations from the mean
  - Probabilities serve as weights
  - $\sigma^2(X) = \sum_i (x_i - \mu)^2 P(X = x_i)$
  - Units are square of the units of the variable
- Standard deviation (σ): Square root of variance
  - Has same units as the variable

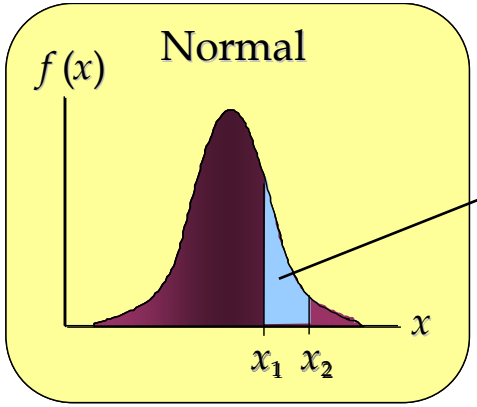
Introduction to Normal Distribution "Model"



- The graph of the pdf (probability density function) is a bell shaped curve
- The normal random variable takes values from - ∞ to +∞
- It is symmetric and centered around the mean (which is also the median and mode)
- Any normal distribution can be specified with just two parameters – the mean (μ) and the standard deviation (σ)
- We write this as $X \sim N(\mu, \sigma^2)$

Probability Calculations for the Normal "Model"

- The probability associated with any single value of the random variable is not defined
- Probability of values being in a range = Area under the pdf curve in that range



$$P(x_1 \le X \le x_2) = P(X \le x_2) - P(X \le x_1)$$

- Area under the entire curve = $P(-\infty \le X \le +\infty) = 1$
- Two methods to calculate $P(X \le x)$
  - Use MS Excel®: NORMDIST(x, μ, σ, 1)
  - Use Z-scores and Standard Normal Distribution

Z-scores, Standard Normal Distribution

- For every value (x) of the random variable X, we can calculate its Z-score:

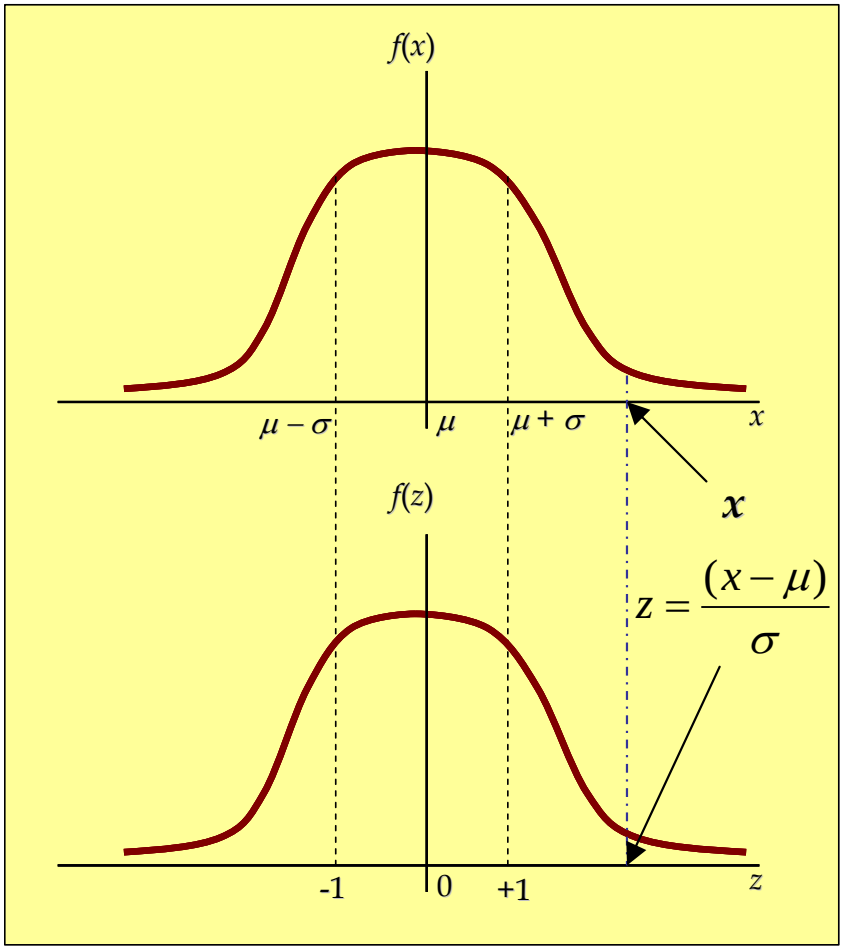
$$z = \frac{x - \mu}{\sigma}$$

- Interpretation – How many standard deviations away is the value from the mean?

- If X~N(μ, σ^2), then
  - Z-scores have a normal distribution with μ=0 and σ=1

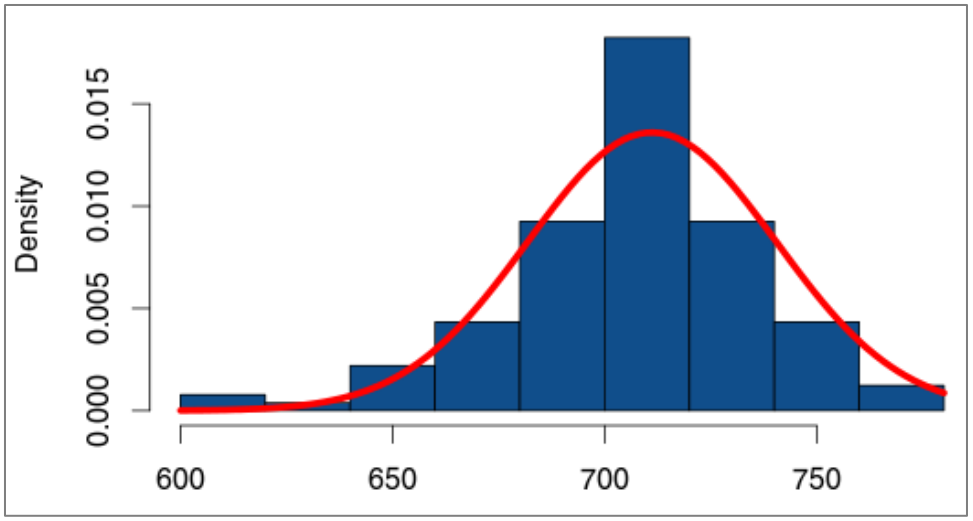
    i.e. Z ~ N(0,1)
  - Standard Normal Distribution

- P(X ≤ x) = P(Z ≤ z)!!



Utility of the Normal "Model" (Example: GMAT Scores)

- Recall the distribution of GMAT scores



- Calculate the following
  - P(X≤ 680)
  - What is P(697≤ X≤740)?

- Now, suppose GMAT scores can be reasonably modeled using $N(638, 66^2)$

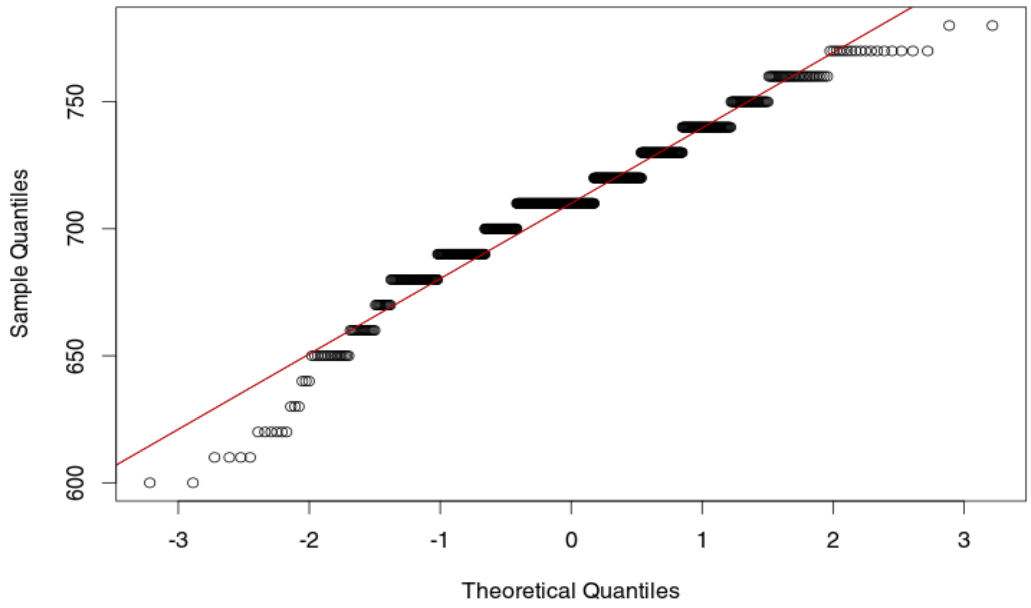
Evaluation of the Normal "Model"

- How can we say that the normal distribution is a reasonable approximation of the data?

- How can data look different from a normal distribution?
  - More than one mode suggesting data come from distinct groups
  - Lack of symmetry
  - Unusual extreme values

- Can identify these differences by looking at
  - Visual inspection of the histogram (not very accurate)
  - Numerical summaries like Skewness and Kurtosis
  - Graphical summaries (Normal Quantile plot)

Normal Quantile (Q-Q) Plot



- Nearly normal if the data track the diagonal reference line on the plot
- Deviations often likely at extremes, and the bands help judge the severity of the deviation

Using Normal "Model" for Managerial Decisions

- Suppose that a packaging system fills boxes of cereal. The package label states the weight of the box as 16 oz. But, the weights of the cereal boxes filled by the packaging system are normally distributed with $\mu = 16.3$ oz and $\sigma = 0.2$ oz.
  - What is the probability that a randomly picked box is underweight?
  - To what weight should the mean of the process be adjusted so that the chance of an underweight box is only 0.005? (Assume $\sigma = 0.2$ oz.)

Linear Combination of Random Variables

Let X_1 and X_2 be two random variables with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. Suppose $Y = aX_1 + bX_2$. Then

1) *The mean of Y is*
2) *The standard deviation of Y is*

- Independent: When the value taken by one random variable does not affect the value taken by the other random variable
  - e.g. Roll of two dice

- Dependent: When the value of one random variable gives us more information about the other random variable
  - e.g. Height and weight of students

Linear Combination of Independent Random Variables

- Suppose Y = aX_1+bX_2
- The mean and variance of Y are given by:
  - $E(Y) = a\mu_1 + b\mu_2$
  - $Var(Y) = a^2\sigma_1^2 + b^2\sigma_2^2$
- Suppose $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$.
- Then above results hold and, in addition, $Y \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$.

Summary of Session I

- A random variable describes the probabilities for an uncertain future numerical outcome of a random process
- A probability distribution is a rule that identifies possible outcomes of a random variable and assigns a probability to each
- The expected value / mean of a random variable is the weighted average of its values
- Variance is the weighted average of the squared deviations from the mean
- A probability distribution can be pictorially represented by a histogram, box-plot, probability density function
- Normal distribution is a symmetric bell shaped continuous probability distribution that is uniquely specified by a mean and standard deviation
  - Every normal distribution can be converted into a standard normal distribution (Z-score)
  - Sum of independent normally distributed random variables is a normally distributed random variable

Software Notes (Session I)

- Frequency table
  - table(variable)
- Bar chart
  - barplot(table(variable))
- Histogram
  - hist(variable)
- Box-plot
  - boxplot(variable)
- Summary statistics
  - Mean: mean(variable)
  - Variance: var(variable)
  - Standard deviation: sd(variable)
  - Skewness: skewness(variable) , needs package 'moments'
  - Kurtosis: kurtosis(variable), needs package 'moments'

Skewness and Kurtosis

- Two additional summary measures of a random variable / probability distribution
- Can be interpreted as the third and fourth moments just as mean and variance are the first and second moments

Skewness

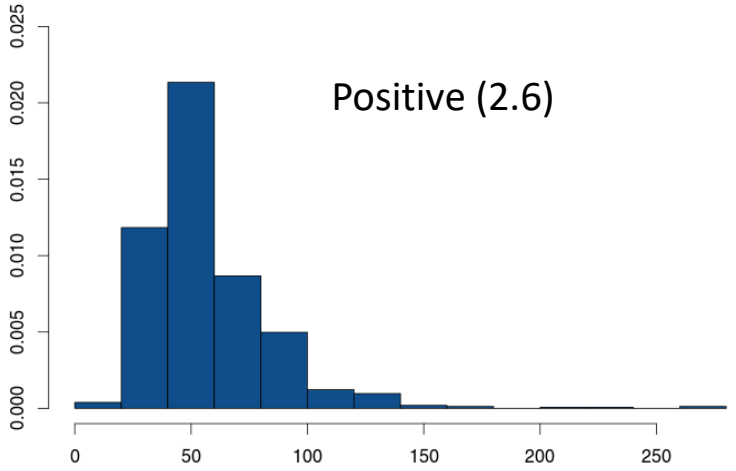
- A measure of "asymmetry" in the distribution
- Mathematically, it is given by
  - $E[(X - \mu)/\sigma]^3$
- Negative skewness implies mass of the distribution is concentrated on the right

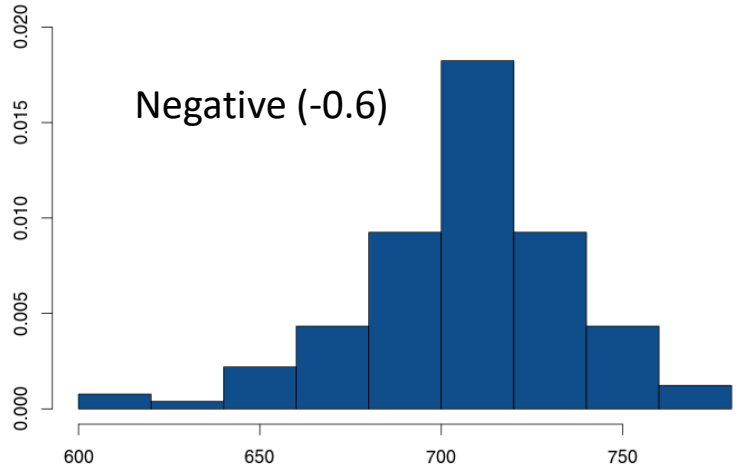
(Excess) Kurtosis

- A measure of the "peakedness" of the distribution (relative to normal)
- Mathematically, it is given by
  - $E[(X - \mu)/\sigma]^4 - 3$
- For symmetric distributions, negative kurtosis implies wider peak and thinner tails

Skewness and Kurtosis (contd.)

Skewness





(Excess) Kurtosis