CSC-591: Foundations of Data Science
T/Th. 12:50-2:05pm. EBI-1005.

Ranga Raju Vatsavai
Chancellors Faculty Excellence Associate Professor in Geospatial Analytics
Department of Computer Science, North Carolina State University (NCSU)
Associate Director, Center for Geospatial Analytics, NCSU
&
Joint Faculty, Oak Ridge National Laboratory (ORNL)

W16: 12/03/15

Today

- Final Review
- Final: 12/8/15, 1:00-4:00pm. (In Class)

Grading

Grading Item	Number of points	X %	Grading Score (out of 100)	Bonus Your Score x %
HW1	100	5	5	
HW2	100	5	5	
HW3	100	5	5	X 2%
HW4	100	5	5	X 2%
HW5	100	5	5	X 4%
MT1	100	15	15	
MT2	100	20	20	X 4%
Final	100	35	35	X 2%
Instructor			5 (class participation + review)	
Total			100	14

Topics

- Probability distributions and estimation
- Sampling distribution and CLT
- Hypothesis testing
- Regression
- Attribute selection

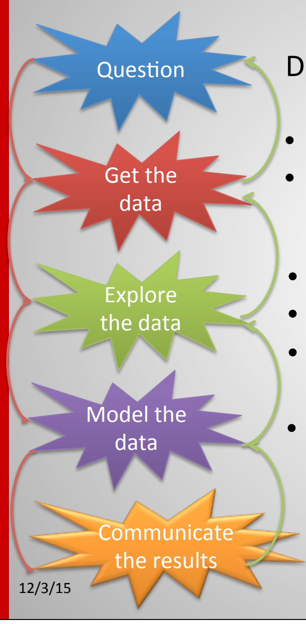
~60%

- Sampling and cross validation
- Bayesian Inference
- Bayesian networks
- Missing data

~40%

NC STATE UNIVERSITY

Data Science Workflow



Does Streptomycin cure Tuberculosis?

- Design a sampling scheme
- Decide on relevant attributes (measurements)

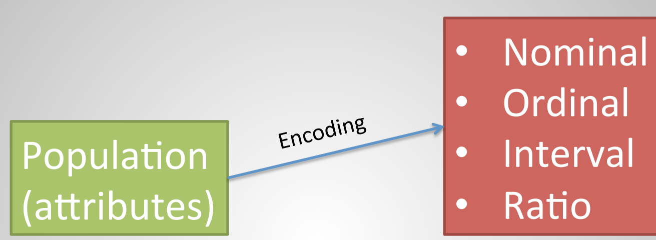
- Data types, Plots
- Missing data, anomalies, …
- Filtering, transformation

- Build, Fit, and Validate the model

12/3/15 © Raju Vatsavai CSC-591. 5

NC STATE UNIVERSITY

DS: Statistical Perspective



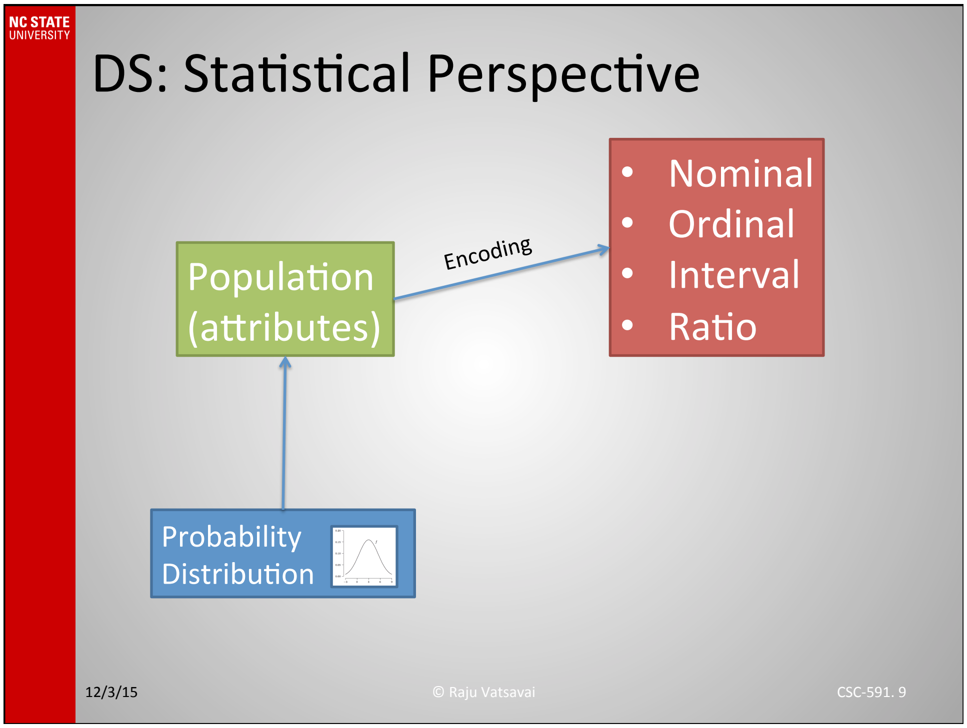
12/3/15 © Raju Vatsavai CSC-591. 6

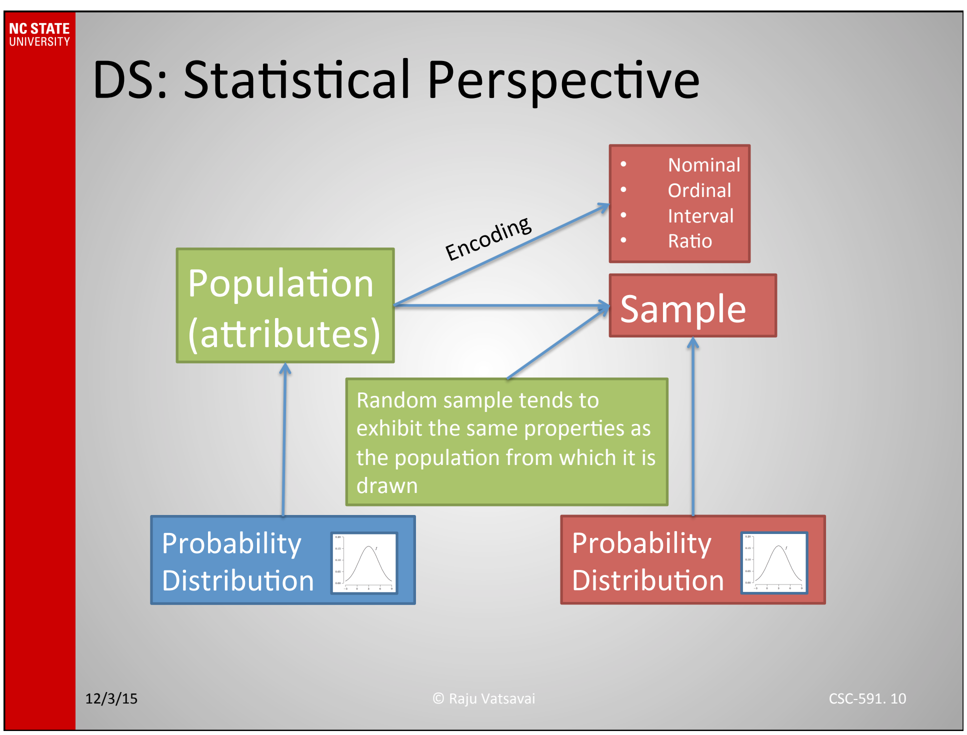
DS: Statistical Perspective

	Attribute Type	Description	Examples	Operations
Categorical Qualitative	Nominal	Nominal attribute values only distinguish. (=, ≠)	zip codes, employee ID numbers, eye color, sex: {*male, female*}	mode, entropy, contingency correlation, $\chi 2$ test
	Ordinal	Ordinal attribute values also order objects. (<, >)	hardness of minerals, {*good, better, best*}, grades, street numbers	median, percentiles, rank correlation, run tests, sign tests
Numeric Quantitative	Interval	For interval attributes, differences between values are meaningful. (+, -)	calendar dates, temperature in Celsius or Fahrenheit	mean, standard deviation, Pearson's correlation, *t* and *F* tests
	Ratio	For ratio variables, both differences and ratios are meaningful. (*, /)	temperature in Kelvin, monetary quantities, counts, age, mass, length, current	geometric mean, harmonic mean, percent variation

DS: Statistical Perspective

- Experiment, outcomes, sample space, events
- Basic set operations
- Probability and three axioms
- Probability rules
- Independent events
- Conditional probability
- Bayes theorem
- Random variables: discrete and continuous
- PMF and Bernoulli distribution


NC STATE UNIVERSITY
DS: Statistical Perspective
Nominal
Ordinal
Interval
Ratio
Encoding
Population (attributes)
Probability Distribution
12/3/15
© Raju Vatsavai
CSC-591. 9


NC STATE UNIVERSITY
DS: Statistical Perspective
Nominal
Ordinal
Interval
Ratio
Encoding
Population (attributes)
Sample
Random sample tends to exhibit the same properties as the population from which it is drawn
Probability Distribution
Probability Distribution
12/3/15
© Raju Vatsavai
CSC-591. 10

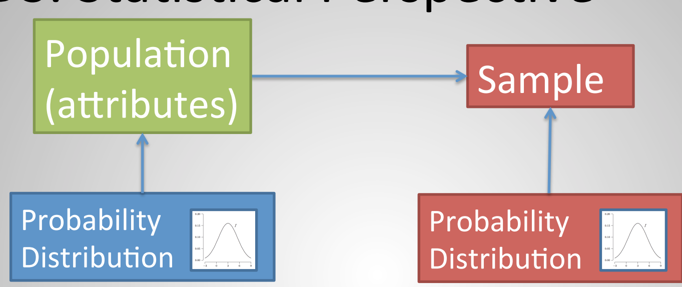
NC STATE UNIVERSITY

DS: Statistical Perspective

- Experiment, outcomes, sample space, events
- Basic set operations
- Probability and three axioms
- Probability rules
- Independent events
- Conditional probability
- Bayes theorem
- Random variables: discrete and continuous
- PMF and Bernoulli distribution
- Continuous uniform distribution, PDF, Normal distribution

NC STATE UNIVERSITY

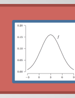
DS: Statistical Perspective



- Draw random samples from a population
- Make inference (conclusions) about populations from noisy data that is drawn from it
- The randomness governing data (samples) is given by densities and mass functions

NC STATE UNIVERSITY

DS: Statistical Perspective

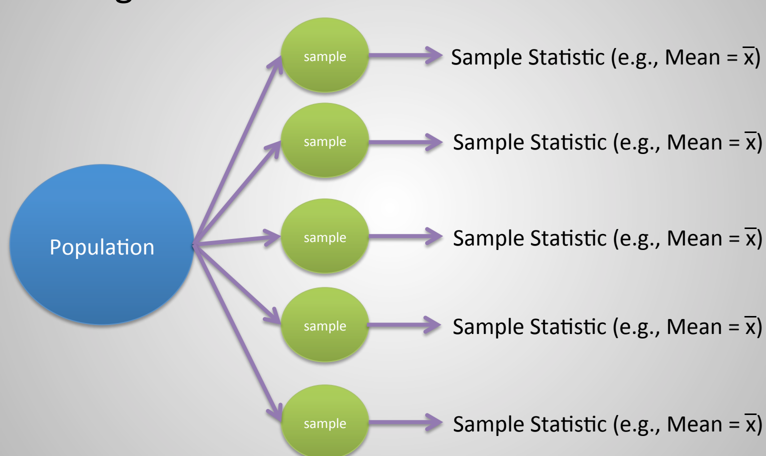
IID Random Samples (random variables) → Probability Distribution  → MLE

- How good are these estimates?

NC STATE UNIVERSITY

DS: Statistical Perspective

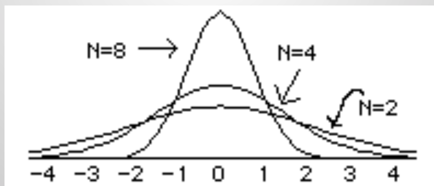
- How good are these estimates?



Sample Distributions Sampling Distribution

DS: Statistical Perspective

- Sampling distribution is used in constructing CI for mean and for significance testing.
- Given a population with a mean of μ and a standard deviation of σ, the sampling distribution of the mean has a mean of μ and a variance of $\sigma_m^2 = \frac{\sigma^2}{N}$



DS: Statistical Perspective

- Given a population with a finite mean μ and a finite non-zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ^2/N as N, the sample size, increases.

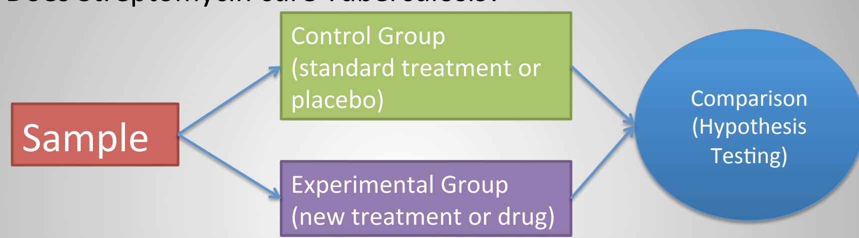
  $\bar{x}$ ~ N(mean = μ, and SE = s/√n)
- What is the significance?

Regardless of the shape of the parent population, the sampling distribution of the mean approaches a normal distribution as N increases.

NC STATE UNIVERSITY

DS: Statistical Perspective

Does Streptomycin cure Tuberculosis?

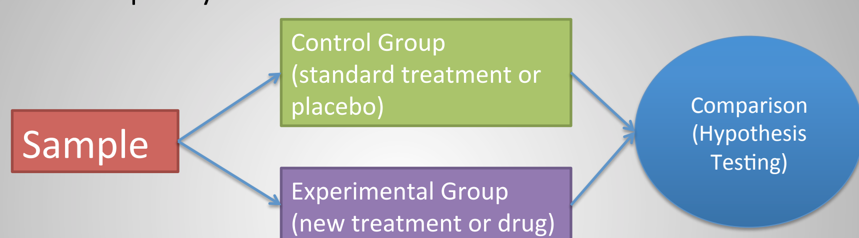


- Hypothesis
  - Is a statement about a population parameter
- Hypothesis testing
  - Based on a sample from population, goal is to decide on which of the two complementary hypothesis is true
- Null and alternative hypothesis
  - Two complementary hypothesis (null and alternative) – both needs to be specified

NC STATE UNIVERSITY

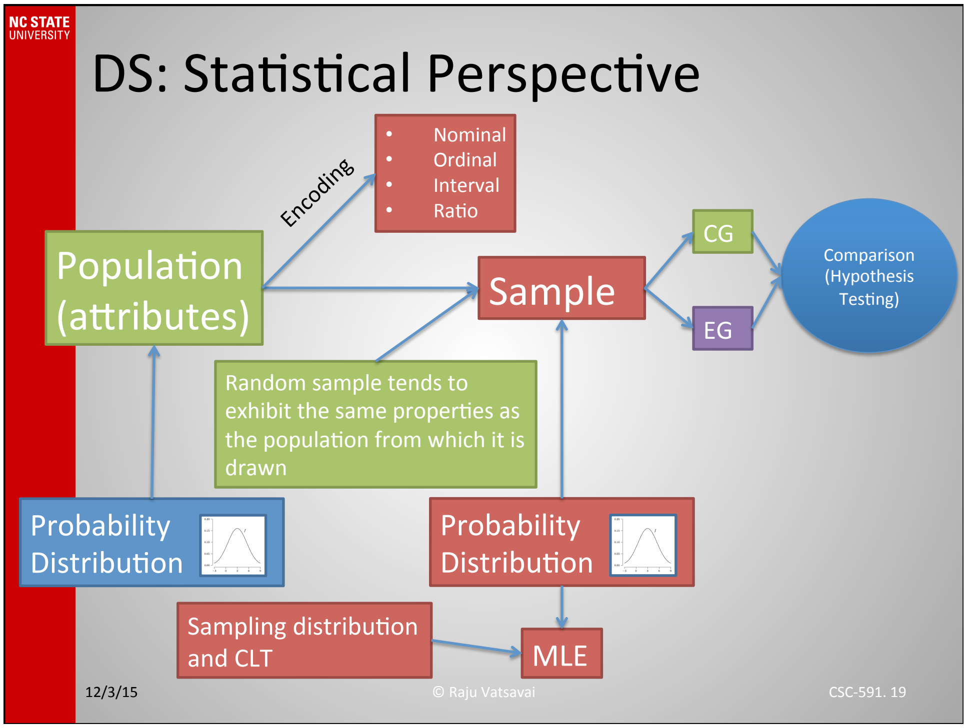
DS: Statistical Perspective

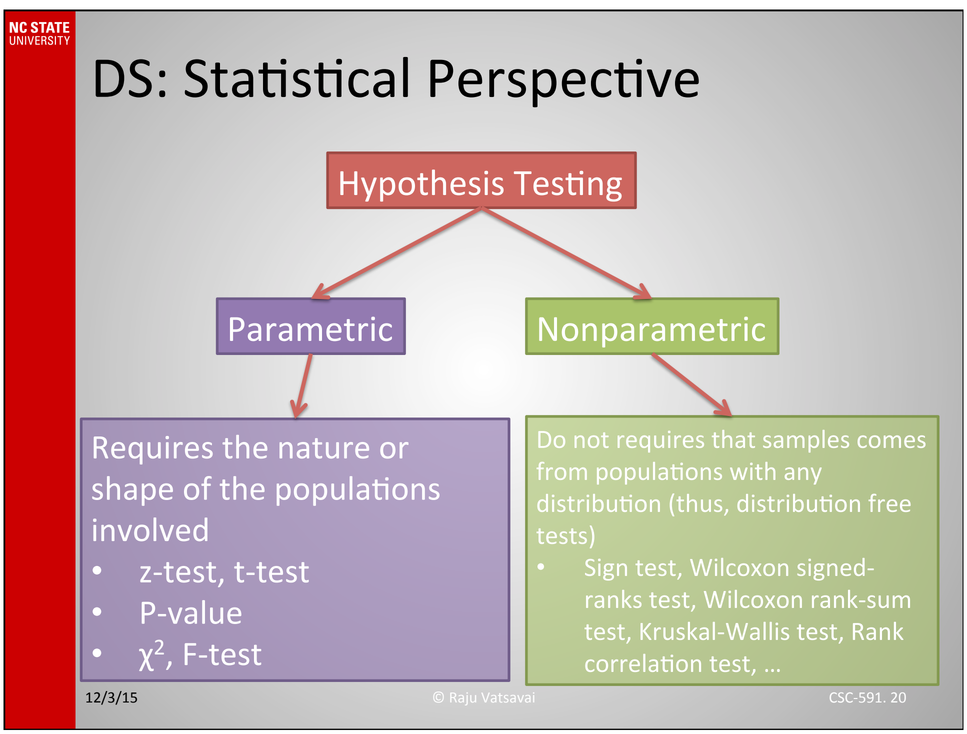
Does Streptomycin cure Tuberculosis?



- General Procedure

Step 1	State the hypotheses and identify the claim.
Step 2	Find the critical value(s) from the appropriate table
Step 3	Compute the test value.
Step 4	Make the decision to reject or not reject the null hypothesis.
Step 5	Summarize the results.


NC STATE UNIVERSITY
DS: Statistical Perspective
Encoding
• Nominal
• Ordinal
• Interval
• Ratio
Population (attributes)
Sample
CG
EG
Comparison (Hypothesis Testing)
Random sample tends to exhibit the same properties as the population from which it is drawn
Probability Distribution
Probability Distribution
Sampling distribution and CLT
MLE
12/3/15
© Raju Vatsavai
CSC-591. 19


NC STATE UNIVERSITY
DS: Statistical Perspective
Hypothesis Testing
Parametric
Nonparametric
Requires the nature or shape of the populations involved
• z-test, t-test
• P-value
• χ^2, F-test
Do not requires that samples comes from populations with any distribution (thus, distribution free tests)
• Sign test, Wilcoxon signed-ranks test, Wilcoxon rank-sum test, Kruskal-Wallis test, Rank correlation test, …
12/3/15
© Raju Vatsavai
CSC-591. 20

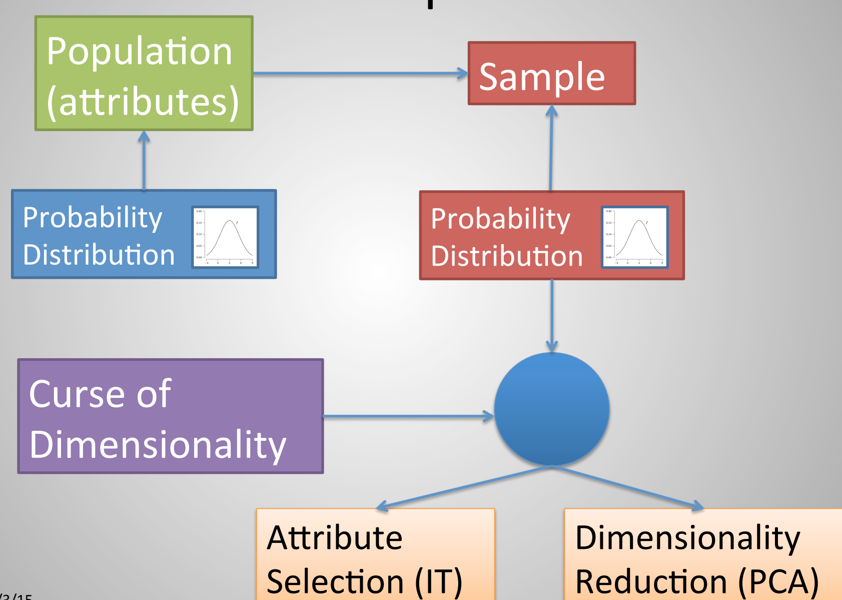
NC STATE UNIVERSITY

DS: Statistical Perspective

- Linear Regression
  - Least Squares
- Correlation
- Regression Parameters
- Properties
- Significance of “r”
- Total Variation (explained + unexplained)
- Coefficient of Determination
- Standard Error of estimate, Prediction Interval
- Multiple Linear Regression
- Multiple Correlation Coefficient (R)
- Testing for significance of R
- Regression as classification: Logistic Regression

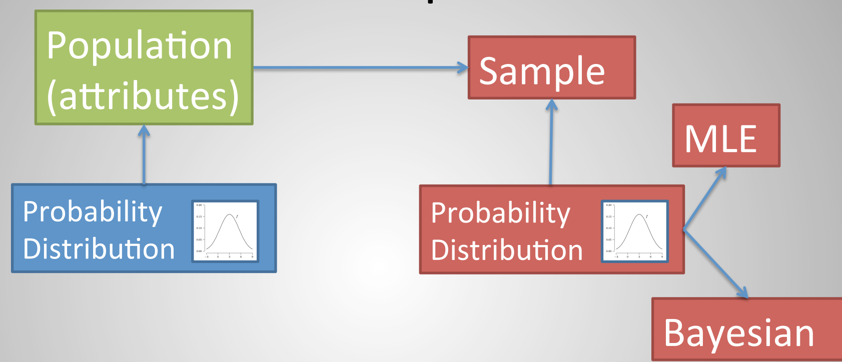
NC STATE UNIVERSITY

DS: Statistical Perspective



NC STATE UNIVERSITY

DS: Statistical Perspective



- Bayesian analysis is a statistical procedure which endeavors to estimate parameters of an underlying distribution based on the observed distribution

NC STATE UNIVERSITY

DS: Statistical Perspective

- Posterior distribution is the most important quantity in Bayesian inference.

$$f(\theta \mid x) = \frac{f(x \mid \theta) f(\theta)}{\int f(x \mid \theta) f(\theta)\, d\theta}$$

- Let X=x denote the observed realization of a uni- or multivariate r.v. X with density function $f(x|\theta)$. Specifying a prior distribution $f(\theta)$ allows us to compute the density function $f(\theta|x)$ of the posterior distribution using Bayes' theorem.

DS: Statistical Perspective

- Bayesian inference allows the probabilistic specification of prior beliefs through a prior distribution.
- It is often useful and justified to restrict the range of possible prior distributions to a specific family with one or two parameters, say. The choice of this family can be based on the type of likelihood function encountered.

DS: Statistical Perspective

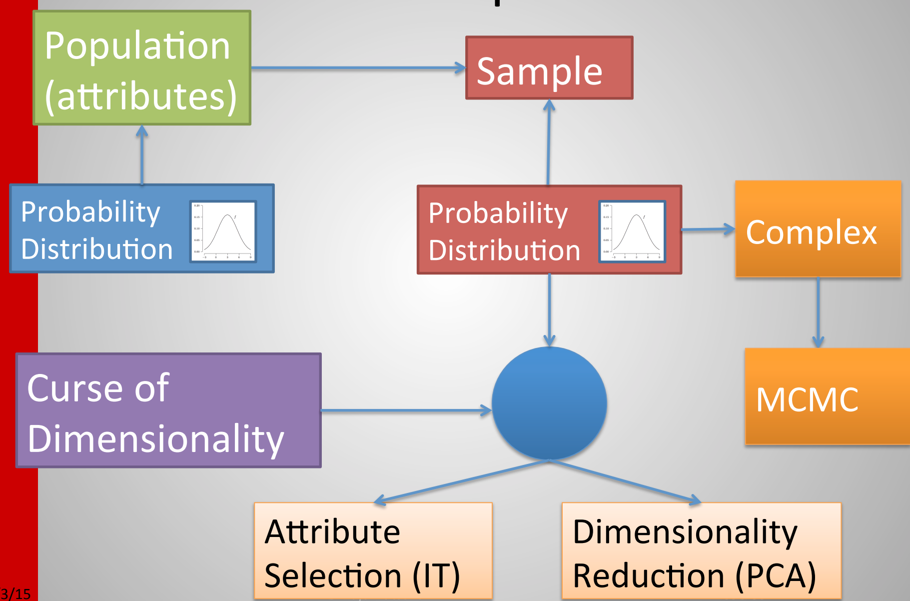
- A pragmatic approach to choosing a prior distribution is to select a member of a specific family of distributions such that the posterior distribution belongs to the same family. This is called a *conjugate prior distribution*.

Summary of conjugate prior distributions for different likelihood functions

Likelihood	Conjugate prior distribution	Posterior distribution
$X \mid \pi \sim \mathrm{Bin}(n, \pi)$	$\pi \sim \mathrm{Be}(\alpha, \beta)$	$\pi \mid x \sim \mathrm{Be}(\alpha + x, \beta + n - x)$
$X \mid \pi \sim \mathrm{Geom}(\pi)$	$\pi \sim \mathrm{Be}(\alpha, \beta)$	$\pi \mid x \sim \mathrm{Be}(\alpha + 1, \beta + x - 1)$
$X \mid \lambda \sim \mathrm{Po}(e \cdot \lambda)$	$\lambda \sim \mathrm{G}(\alpha, \beta)$	$\lambda \mid x \sim \mathrm{G}(\alpha + x, \beta + e)$
$X \mid \lambda \sim \mathrm{Exp}(\lambda)$	$\lambda \sim \mathrm{G}(\alpha, \beta)$	$\lambda \mid x \sim \mathrm{G}(\alpha + 1, \beta + x)$
$X \mid \mu \sim \mathrm{N}(\mu, \sigma^2 \text{ known})$	$\mu \sim \mathrm{N}(\nu, \tau^2)$	$\mu \mid x \sim \mathrm{N}\left(\left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1} \cdot \left(\frac{x}{\sigma^2} + \frac{\nu}{\tau^2}\right), \left(\frac{1}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}\right)$
$X \mid \sigma^2 \sim \mathrm{N}(\mu \text{ known}, \sigma^2)$	$\sigma^2 \sim \mathrm{IG}(\alpha, \beta)$	$\sigma^2 \mid x \sim \mathrm{IG}(\alpha + \frac{1}{2}, \beta + \frac{1}{2}(x - \mu)^2)$

NC STATE UNIVERSITY

DS: Statistical Perspective



12/3/15

NC STATE UNIVERSITY

DS: Statistical Perspective

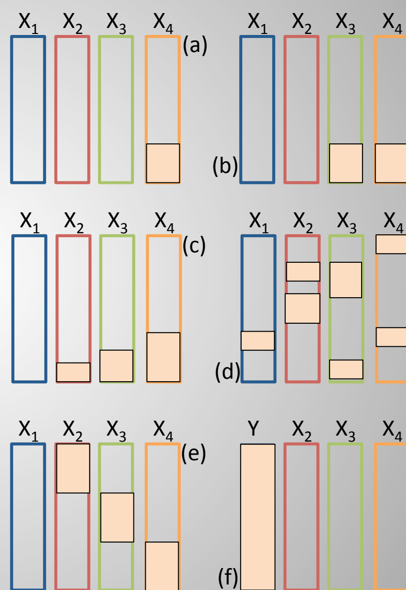
- Model Selection
- Mixture Models
- Expectation Maximization
- Missing Data

12/3/15 © Raju Vatsavai CSC-591. 28

NC STATE UNIVERSITY

Missing Data Patterns

- (a) Univariate pattern
- (b) Unit nonresponse pattern
- (c) Monotone pattern
- (d) General pattern
- (e) Planned missing pattern
- (f) Latent variable pattern



12/3/15 © Raju Vatsavai CSC-591. 29

NC STATE UNIVERSITY

Missing Data Theory

- Rubin, et. al. introduced a classification system for missing data problems
  - Introduced three so-called missing data mechanisms that describe how the probability of a missing value relates to the data, if at all.
    - **MAR**: data are missing at random
    - **MCAR**: missing completely at random
    - **MNAR**: missing not at random

12/3/15 © Raju Vatsavai CSC-591. 30

NC STATE UNIVERSITY

For Final, focus on these topics

- Probability distributions and estimation
- Sampling distribution and CLT
- Hypothesis testing
- Regression
- Attribute selection

~60%

- Sampling and cross validation
- Bayesian Inference
- Bayesian networks
- Missing data

~40%

12/3/15 © Raju Vatsavai CSC-591. 31

NC STATE UNIVERSITY

Exam Structure

- One big with several short questions – to test fundamental understanding (~20%)
- 4-5 Numerical questions (~60%)
- 1-2 Tricky question (~20%)
  - (not to trick you; but solution could be more easy if you think about it little bit). May be numerical, but you can obtain solution without going through routine computations
- Bonus question (2%)
  - Little bit hard
- Roughly 2.5 Hours

12/3/15 © Raju Vatsavai CSC-591. 32

NC STATE UNIVERSITY

Acknowledgements

- Thank you
- Please complete review to get 1% grade point