# Biostatistics 602 - Statistical Inference
# Lecture 01
# Introduction to BIOSTAT602
# Principles of Data Reduction

Hyun Min Kang

January 10th, 2013





## Today's Outline

- Course Syllabus
- Overview of BIOSTAT602
- Sufficient Statistics





## Basic Polls : Home Department

**What is your home department?**

- Biostatistics
- Statistics
- Bioinformatics
- Survey Methodology
- Other Departments





## Basic Polls : Official Roster

**Are you taking the class, or just sitting in?**

- Taking for credit
- Sitting in
- Plan to take, but needs permission

## Basic Polls : 601 History

### Have you taken BIOSTAT601 or equivalent class?

- I took BIOSTAT601.
- I took an BIOSTAT601-equivalent class.
- I do not have BIOSTAT601 equivalent background





## BIOSTAT602 - Course Information

### Instructor

Name Hyun Min Kang
Office M4531, SPH II
E-mail hmkang@umich.edu
Office hours Thursday 4:30-5:30pm

### Course Web Page

- See http://genome.sph.umich.edu/wiki/602
- No C-Tools site will be available in 2013.





## BIOSTAT602 - Basic Information

### Class Time and Location

Time Tuesday and Thursday 1:00-3:00pm.
Location USB 2260

### Prerequisites

- BIOSTAT601 or equivalent knowledge
  (Chapter 1-5.5 of Casella and Berger)
- Basic calculus and matrix algebra





## BIOSTAT602 - Textbooks

### Required Textbook

*Statistical Inference*, 2nd Edition, by Casella and Berger

### Recommended Textbooks

- *Statistical Inference*, by Garthwaite, Jolliffe and Jones.
- *All of Statistics: A Concise Course in Statistical Inference*, by Wasserman
- *Mathematical Statistics: Basics Ideas and Selected Topics*, by Bickel and Doksum.

## Grading

- Homework 20%
- Midterm 40%
- Final 40%





## Important Dates

- First Lecture : Thursday January 10th, 2013
- Midterm : 1:00pm - 3:00pm, Thursday February 21st, 2013
- No lectures on March 5th and 7th (Vacation)
- No lecture on April 2nd (Instructor out of town)
- Last Lecture : Tuesday April 23rd, 2013 (Total of 26 lectures)
- Final : 4:00pm - 6:00pm, Thursday April 25th, 2013 (University-wide schedule)





## Honor code

- Honor code is STRONGLY enforced throughout the course.
  - The key principle is that all your homework and exams must be on your own.
  - See http://www.sph.umich.edu/academics/policies/conduct.html for details.
- You are encouraged to discuss the homework with your colleagues.
- You are NOT allowed to share any piece of your homework with your colleagues electronically or by a hard copy.
- If a break of honor code is identified, your entire homework (or exam) will be graded as zero, while incomplete submission of homework assignment will be considered for partial credit.





## About the style of the class

- In previous years, the instructors wrote the notes on the whiteboard or projected the notes onto a screen during the class
- In this class, we will use prepared slides for the sake of clarity.
- For this reason, the his class has a risk to serve as a slot for after-lunch nap.
- Instructor strongly encourages to copy the slides during the class by hand to digest the material, although all slides will be available online.
- Focusing on the class will be helpful a lot.
- Feedback on the class, especially on the lecture style, would be very much appreciated.

## "Statistical Inference"

### Probability in BIOSTAT601

Given some specified probability mass function (pmf) or probability density function (pdf), we can make probabilistic statement about data that could be generated from the model.

### Statistical Inference in BIOSTAT602

A process of drawing conclusions or making statements about a population of data based on a random sample of data from the population.





## Notations in BIOSTAT602

- $X_1, \cdots, X_n$ : Random variables identically and independently distributed (iid) with probability density (or mass) function $f_X(x|\theta)$.
- $x_1, \cdots, x_n$ : Realization of random variables $X_1, \cdots, X_n$.
- $\mathbf{X} = (X_1, \cdots, X_n)$ is a random sample of a population (typically iid), and the characteristics of this population are described by $f_{\mathbf{X}}(\mathbf{x}|\theta)$.
- The joint pdf (or pmf) of $\mathbf{X} = (X_1, \cdots, X_n)$ (assuming iid) is

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta)$$





## BIOSTAT601 vs BIOSTAT602

### BIOSTAT601

In BIOSTAT601, we assume the knowledge of $\theta$ in making probabilistic statements about $X_1, \cdots, X_n$.

### BIOSTAT602

In BIOSTAT602, we do not know the true value of the parameter $\theta$, and instead we try to learn about this true parameter value through the observed data $x_1, \cdots, x_n$.





## Example of BIOSTAT601 Questions

For a sample size $n$, let $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p_0)$. What is the probability of $\sum_{i=1}^{n} X_i \leq m$?

$$\sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p_0)$$

$$\Pr\left(\sum_{i=1}^{n} X_i \leq m\right) = \sum_{k=0}^{m} \binom{n}{k} p_0^k (1-p_0)^{n-k}$$

## Example of BIOSTAT602 Questions

We assume that the data was generated by a pdf (or pmf) that belongs to a class of pdfs (or pmfs).

$$\mathcal{P} = \{f_X(x|\theta), \theta \in \Omega \subset \mathbb{R}^p\}$$

For example $X \sim \text{Bernoulli}(\theta), \theta \in (0,1) = \Omega \subset \mathbb{R}$.
We collect data in order to

1. Estimate $\theta$ (point estimation)
2. Perform tests of hypothesis about $\theta$.
3. Estimate confidence intervals for $\theta$ (interval estimation).
4. Make predictions of future data.





## BIOSTAT602: Examples of informal questions

1. Estimate $\theta$ (point estimation)
   - What is the estimated probability of head given a series of coin tosses?
2. Perform tests of hypothesis about $\theta$.
   - Given a series of coin tosses, can you tell whether the coin is biased or not?
3. Estimate confidence intervals for $\theta$ (interval estimation).
   - What is the plausible range of the true probability of head, given a series of coin tosses?
4. Make predictions of future data.
   - Given the series of coin tosses, can you predict what the outcome of the next coin toss?





## Data Reduction

**Data**

$x_1, \cdots, x_n$ : Realization of random variables $X_1, \cdots, X_n$.

**Data Reduction**

Define a function of data

$$T(x_1, \cdots, x_n) : \mathbb{R}^n \rightarrow \mathbb{R}^d$$

We wish this summary of data to..

1. Be simpler than the original data, e.g. $d \leq n$.
2. Keep all the information about $\theta$ that is contained in the original data $x_1, \cdots, x_n$.





## Statistic

$$T(X_1, \cdots, X_n) = T(\mathbf{X})$$

- It is a function of random variables $X_1, \cdots, X_n$.
- $T(\mathbf{X})$ itself is also a random variable.
- $T(\mathbf{X})$ defines a form of data reduction or data summary.

## Data Reduction

Data reduction in terms of a statistic $T(\mathbf{X})$ is a partition of the sample space $\mathcal{X}$.

### Example

Suppose $X_i \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$ for $i = 1, 2, 3$, and $0 < p < 1$.
Define $T(X_1, X_2, X_3) = X_1 + X_2 + X_3$, then $T : \{0, 1\}^3 \to \{0, 1, 2, 3\}$.





## Example of Data Reduction

| Partition | $X_1$ | $X_2$ | $X_3$ | $T(\mathbf{X}) = X_1 + X_2 + X_3$ |
|---|---|---|---|---|
| $A_0$ | 0 | 0 | 0 | 0 |
| $A_1$ | 0 | 0 | 1 | 1 |
| | 0 | 1 | 0 | 1 |
| | 1 | 0 | 0 | 1 |
| $A_2$ | 0 | 1 | 1 | 2 |
| | 1 | 0 | 1 | 2 |
| | 1 | 1 | 0 | 2 |
| $A_3$ | 1 | 1 | 1 | 3 |

$$\begin{aligned} \mathcal{T} &= \{t : t = T(\mathbf{X}) \text{ for some } \mathbf{x} \in \mathcal{X}\} \\ A_t &= \{\mathbf{x} : T(\mathbf{X}) = t, t \in \mathcal{T}\} \end{aligned}$$

Instead of reporting $\mathbf{x} = (x_1, x_2, x_3)^T$, we report only $T(\mathbf{X}) = t$, or equivalently $\mathbf{x} \in A_t$.





## Example of Data Reduction

The partition of the sample space based on $T(\mathbf{X})$ is "coarser" than the original sample space.

- There are 8 elements in the sample space $\mathcal{X}$.
- They are partitioned into 4 subsets
- Thus, $T(\mathbf{X})$ is simpler (or coarser) than $\mathbf{X}$.





## Sufficient Statistics

### Definition 6.2.1

A statistic $T(\mathbf{X})$ is a *sufficient statistic* for $\theta$ if the conditional distribution of sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on $\theta$.

In other words, the conditional pdf or pmf of $\mathbf{X}$ given $T = t$, $f_{\mathbf{X}}(\mathbf{x}|\, T(\mathbf{X}) = t) = h(\mathbf{x})$ does not depend on $\theta$

## Sufficient Statistics: Example

- Suppose $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p),\ 0 < p < 1$.
- Claim that $T(X_1, \cdots, X_n) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for $p$.





## Proof : Overview

- $T(\mathbf{X}) = \sum_{i=1}^{n} X_i \sim \text{Binomial}(n, p)$
- Need to find the conditional pmf of $\mathbf{X}$ given $T = t$.
- And show that the distribution does not depend on $p$.





## Detailed Proof

$$
\begin{aligned}
& \Pr\left(X_1 = x_1, \cdots, X_n = x_n | \sum_{i=1}^{n} X_i = t\right) \\
= & \frac{\Pr\left(X_1 = x_1, \cdots, X_n = x_n, \sum_{i=1}^{n} X_i = t\right)}{\Pr\left(\sum_{i=1}^{n} X_i = t\right)} \\
= & \begin{cases} \dfrac{\Pr\left(X_1 = x_1, \cdots, X_n = x_n\right)}{\Pr\left(\sum_{i=1}^{n} X_i = t\right)} & \text{if } \sum_{i=1}^{n} X_i = t \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$





## Detailed Proof (cont'd)

If $\sum_{i=1}^{n} X_i = t,\ t \sim \text{Binomial}(n, p)$

$$
\begin{aligned}
\Pr(X_1 = x_1, \cdots, X_n = x_n) &= \prod_{i=1}^{n} \Pr(X_i = x_i) \\
&= p^{x_1}(1-p)^{1-x_1} \cdots p^{x_n}(1-p)^{1-x_n} \\
&= p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i} \\
\Pr\left(\sum_{i=1}^{n} X_i = t\right) &= \binom{n}{t} p^t (1-p)^{n-t} \\
\Pr\left(\mathbf{X} = \mathbf{x} | \sum_{i=1}^{n} X_i = t\right) &= \frac{1}{\binom{n}{t}}
\end{aligned}
$$

## Detailed Proof (cont'd)

Therefore, conditional distribution

$$\Pr\left(\mathbf{X}=\mathbf{x}|\sum_{i=1}^{n}X_i=t\right) = \begin{cases}\frac{1}{\binom{n}{t}} & \text{if } \sum_{i=1}^{n}X_i=t \\ 0 & \text{otherwise}\end{cases}$$

Because $\Pr(X|\,T(\mathbf{X})=t)$ does not depend on $p$, by definition, $T(\mathbf{X})=\sum_{i=1}^{n}X_i$ is a sufficient statistic for $p$.





## Note from the proof

If $\mathbf{X}$ is a sample point such that $T(\mathbf{X})\neq t$, then $\Pr(\mathbf{X}=\mathbf{x}|\,T(\mathbf{x})=t)=0$ always, so we don't have to consider the case when $T(\mathbf{x})\neq t$ in the future.





## A Theorem for Sufficient Statistics

**Theorem 6.2.2**

- Let $f_{\mathbf{X}}(\mathbf{x}|\theta)$ is a joint pdf or pmf of $X$
- and $q(t|\theta)$ is the pdf or pmf of $T(\mathbf{X})$.
- Then $T(\mathbf{X})$ is a sufficient statistic for $\theta$,
- if, for every $\mathbf{x}\in\mathcal{X}$,
- the ratio $f_{\mathbf{X}}(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta)$ is constant as a function of $\theta$.





## Proof of Theorem 6.2.2 - discrete case

$$\begin{aligned}\Pr\left(\mathbf{X}=\mathbf{x}|\,T(\mathbf{X})=t\right) &= \frac{\Pr\left(\mathbf{X}=\mathbf{x},\,T(\mathbf{X})=t\right)}{\Pr(T(\mathbf{X})=t)} \\ &= \begin{cases}\dfrac{\Pr(\mathbf{X}=\mathbf{x})}{\Pr(T(\mathbf{X})=t)} & \text{if } T(\mathbf{x})=t \\ 0 & \text{otherwise}\end{cases} \\ &= \begin{cases}\dfrac{f_{\mathbf{X}}(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta)} & \text{if } T(\mathbf{x})=t \\ 0 & \text{otherwise}\end{cases}\end{aligned}$$

which does not depend on $\theta$ by assumption. Therefore, $T(\mathbf{X})$ is a sufficient statistic for $\theta$.

## Example 6.2.3 - Binomial Sufficient Statistic

**Problem**

- $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, $0 < \theta < 1$.
- Show that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a sufficient statistic for $\theta$.

This is the same problem from the last lecture, but we would like to solve is using Theorem 6.2.2.





## Example 6.2.3 - Binomial Sufficient Statistic

**Proof**

$$
\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}|p) &= p^{x_1}(1-p)^{1-x_1}\cdots p^{x_n}(1-p)^{1-x_n} \\
&= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \\
T(\mathbf{X}) &\sim \text{Binomial}(n,p) \\
q(t|p) &= \binom{n}{t} p^t (1-p)^{n-t} \\
\frac{f_{\mathbf{X}}(\mathbf{x}|p)}{q(T(\mathbf{x})|p)} &= \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{\sum_{i=1}^n x_i} p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}} \\
&= \frac{1}{\binom{n}{\sum_{i=1}^n x_i}} = \frac{1}{\binom{n}{T(\mathbf{x})}}
\end{aligned}
$$

By theorem 6.2.2. $T(\mathbf{X})$ is a sufficient statistic for $p$.





## Example 6.2.4 - Normal Sufficient Statistic

**Problem**

- $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$
- Assume that $\sigma^2$ is known.
- Show that the sample mean $T(\mathbf{X}) = \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is a sufficient statistic for $\mu$.





## Example 6.2.4 - Proof

**$f_{\mathbf{X}}(\mathbf{x}|\mu)$**

$$
\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}|\mu) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\sum_{i=1}^n \frac{(x_i-\overline{x}+\overline{x}-\mu)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (x_i-\overline{x})^2 + n(\overline{x}-\mu)^2}{2\sigma^2}\right)
\end{aligned}
$$

## Example 6.2.4 - Proof (cont'd)

### $q(T(\mathbf{x})|\mu)$

Remember from BIOSTAT601 that $T(\mathbf{X}) = \overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$.

$$q(T(\mathbf{x})|\mu) = \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left(-n(\overline{x}-\mu)^2/(2\sigma^2)\right)$$





## Example 6.2.4 - Proof

### Putting things together

$$\begin{aligned}\frac{f_{\mathbf{X}}(\mathbf{x}|\mu)}{q(T(\mathbf{x})|\mu)} &= \frac{(2\pi\sigma^2)^{-n/2}\exp\left(-\frac{\sum_{i=1}^n(x_i-\overline{x})^2+n(\overline{x}-\mu)^2}{2\sigma^2}\right)}{(2\pi\sigma^2/n)^{-1/2}\exp\left(-\frac{n(\overline{x}-\mu)^2}{2\sigma^2}\right)} \\ &= n^{-1/2}(2\pi\sigma^2)^{-(n-1)/2}\exp\left(-\frac{\sum_{i=1}^n(x_i-\overline{x})^2}{2\sigma^2}\right)\end{aligned}$$

which does not depend on $\mu$. By Theorem 6.2.2, the sample mean is a sufficient statistic for $\mu$.





## Summary

### Today

- Overview of BIOSTAT602
- Key differences between BIOSTAT601 and BIOSTAT602
- Sufficient Statistics
  - $\mathbf{X}$ is conditionally independent on $\theta$ given $T(\mathbf{X})$
  - If ratio of pdfs between $\mathbf{X}$ and $T(\mathbf{X})$ does not depend on $\theta$, $T(\mathbf{X})$ is a sufficient statistic.

### Next Lecture

- More on Sufficient Statistics
- Factorization Theorem