Packet 1: Introduction to Statistical Inference

Learning objects:

- Learn the relationship between probability and statistical inference.
- Review basic vocabulary and properties of probability.

What is statistics and why statistics?

Statistics is a data-driven information science that concerns the extraction of useful information from the observed data in a principled way, accounting for uncertainty, and such information can help us make decisions.

Examples:

- Business, e.g. transaction data: Wal-Mart data warehouse, credit card companies; shopping mall management.
- Voice, speech recognition e.g. iPhone Siri, Amazon Alexa, Google Home, WeChat.
- Network and communication systems, e.g. Internet links (Google), Purchase recommendation (Amazon), Netflix Prize, an example of recommendation programs ($1,000,000 for an improved recommender algorithm).

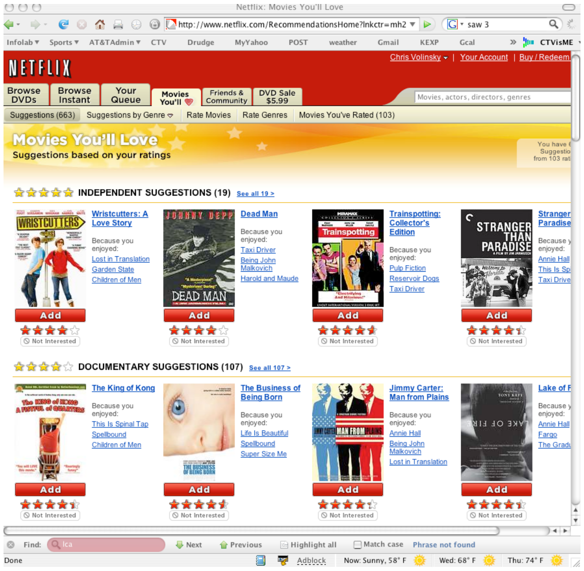

Netflix: Movies You'll Love
http://www.netflix.com/RecommendationsHome?lnkctr=mh2
saw 3
Infolab
Sports
AT&TAdmin
CTV
Drudge
MyYahoo
POST
weather
Gmail
KEXP
Gcal
CTVisME
NETFLIX
Chris Volinsky | Your Account | Buy / Redeem
Browse DVDs
Browse Instant
Your Queue
Movies You'll
Friends & Community
DVD Sale $5.99
Movies, actors, directors, genres
Suggestions (663)
Suggestions by Genre
Rate Movies
Rate Genres
Movies You've Rated (103)
Movies You'll Love
Suggestions based on your ratings
INDEPENDENT SUGGESTIONS (19) See all 19 >
Wristcutters: A Love Story
Because you enjoyed: Lost in Translation, Garden State, Children of Men
Dead Man
Because you enjoyed: Taxi Driver, Being John Malkovich, Harold and Maude
Trainspotting: Collector's Edition
Because you enjoyed: Pulp Fiction, Reservoir Dogs, Taxi Driver
Stranger Paradise
Because y enjoyed: Annie Hall, This Is Sp, Taxi Drive
DOCUMENTARY SUGGESTIONS (107) See all 107 >
The King of Kong
Because you enjoyed: This Is Spinal Tap, Spellbound, Children of Men
The Business of Being Born
Because you enjoyed: Life Is Beautiful, Spellbound, Super Size Me
Jimmy Carter: Man from Plains
Because you enjoyed: Annie Hall, Being John Malkovich, Lost in Translation
Lake of F
Because y enjoyed: Annie Hall, Fargo, The Gradu
Add
Not Interested
Find: lca
Next
Previous
Highlight all
Match case
Phrase not found
Done
Adblock
Now: Sunny, 58° F
Wed: 68° F
Thu: 74° F

Image credit: Chris Volinsky.

- Genomics, e.g. 1000 genome project. We now know that humans are coded in 3 billion DNA letters (A, T, C, G), but what do those DNA sequences tell us?

  If we can tell which genes contribute to the disease risk, then we can make *personalized* diagnosis and treatment.

Image credit: Ivan Chen.

- Image, e.g. Facial, Finger Prints, Handwriting, Brain, Microarray. “Is it you?”

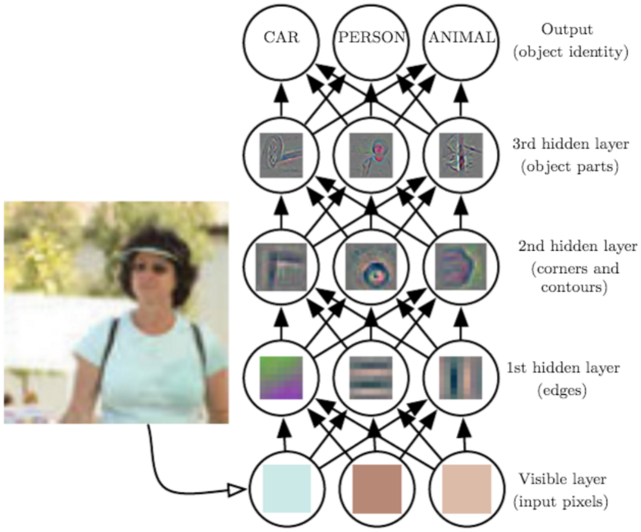
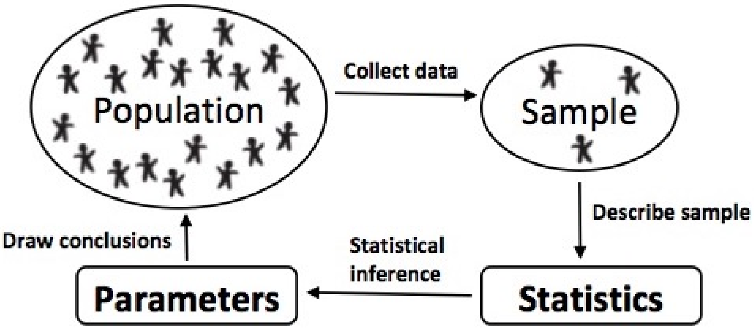


Image credit: Figure 1.2 from Goodfellow et al (2016).

To answer these questions, we need statistics to extract information from the data!

What is Data? Is it information? number? or something else?

Basic terminology in statistical inference



- __________________: The entire group of objects about which we make inferences.

- __________________: A fraction of the population on which we actually collect data.

- __________________: A numerical summary (i.e., a real-valued function) of observed data, i.e., $h(X_1, X_2, \ldots, X_n)$, e.g., mean, median, maximum, variance, etc. Note that $X_1, X_2, \ldots, X_n$ denote the data that we have not yet observed (i.e., r.v.s), and $x_1, x_2, \ldots, x_n$ are the observed (i.e., realized) data.

- __________________: An unknown quantity that indexes a family of distributions.

- ________________: A quantity (e.g., parameter) to be estimated. "It describes what is to be estimated based on the question of interest." —National Research Council (2010).

- ______________: A statistic designed to infer a specific parameter (i.e., an estimand).

- ______________: A value of estimator computed by the data (e.g. $\bar{x}$).

We sometimes do not distinguish between estimators and estimates since we often have to go back and forth between thinking of the data as random and thinking of the data as having "crystallized" into specific values. Also, it is usually clear from the context which is meant.

Probability and statistical inference

A probabilistic model is a set of *assumptions about probability distributions* to represent the randomness of the data. In other words, how were the data generated?

In Math/Stat 414, we have learned several families of probability distributions. For example,

Example: We are interested in measuring the brightness of some galaxies. The brightness of galaxy i ($i = 1, 2, \ldots, n$) is measured by a certain telescope, and is assumed to follow an independent Gaussian distribution. One possible way to describe this probabilistic model is

“All models are wrong, but some are useful.” — George Box.

Overview of this course

Suppose that we obtain data x from a statistical model with an unknown parameter θ (imagine that θ indexes a family of possible distributions for x, and that the true value of θ determines the true data-generating process). We will then consider questions such as:

____________________________: What is a *good* estimator for θ? This depends on the definition of 'good', and we will introduce several criteria by which estimators can be judged. Just providing a point estimator $\hat{\theta}$, without any sense of its uncertainty, is usually unsatisfying. Thus, statistical inference emphasizes accompanying point estimators with information about their uncertainties, e.g., we may be able to describe the distribution of $\hat{\theta}$ or at least say what its standard deviation is (the standard deviation of an estimator is called its standard error).

____________________________: Intuitively, much more informative than just saying something like "I estimate that θ is 2.5" is to provide an interval, such as saying

"I am 95% confident that θ is inclusively between 2.3 and 2.8."

But what does "confident" mean? If θ is a constant, then it either is or isn't in the interval $[2.3, 2.8]$, so what does the 95% mean? In this course we will define precisely what it means to give an interval estimate, and study ways of constructing such estimates.

____________________________: In many applications in the physical, biological, and social sciences, a researcher is interested in testing a hypothesis which can be expressed in the form $\theta = \theta_0$ or, more generally, as $\theta \in H_0$ for some set H_0. Hypothesis testing is closely related to interval estimation (and arguably the latter is more useful), and can be approached via various perspectives.