STA 250/MTH 342 – Intro to Mathematical Statistics

Lecture 1

What is statistics and why study it?

- Statistics is the *prime* information science.
- It deals with the *extraction* of useful *information* from *data*, accounting for the appropriate *uncertainty*.
- Such information can help us *understand how something works*, *make guesses about future observations*, and *make decisions*.
- It has revolutionized virtually all areas of scientific investigation and everyday life in the last 100 years.
  - Biology, physics, psychology, economics, you name it, ...
  - Policy making, the internet, artificial intelligence, ...
- It has becoming ever more important in the past 20 years, due to the computerization of data generation.

Example I: Google and Facebook

- "[W]e create as much information in two days now as we did from the dawn of man through 2003." —Eric Schmidt in 2010.
- That's about five *exabytes* of data every two days, (and that's back in 2010).
- What kind of information can be extracted:
  - The large: Disease epidemic trends, people's social behavior, or even the current stage of the entire human civilization!
  - The "small": What might the user like or want? (What ads are relevant—hotels vs diapers? Which rider is likely to pay more? What is the user trying to say? What does the user like to watch? to wear?)
- All need statistics!

Case study: Image classification ("Deep" learning)

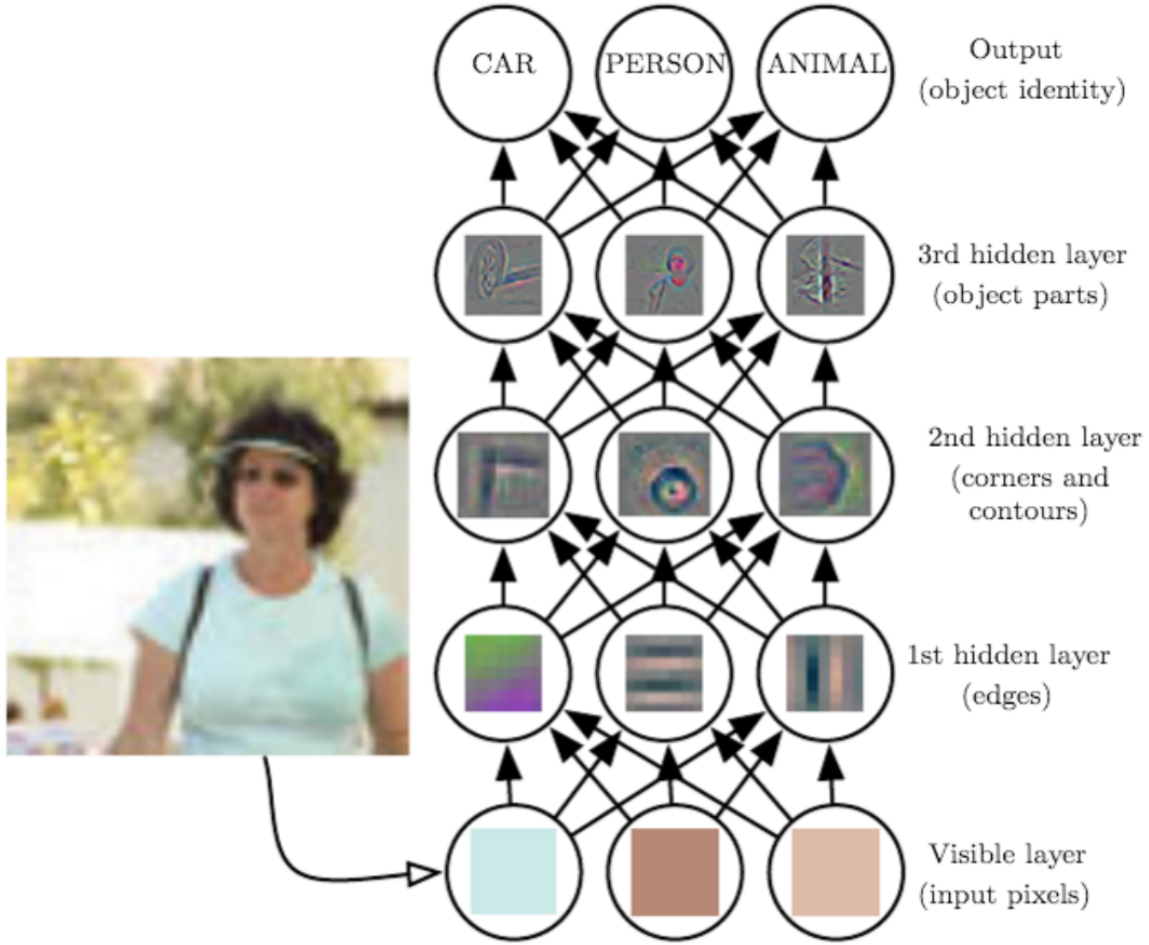

Figure 1.2 from Goodfellow et al (2016)

`http://goodfeli.github.io/dlbook/`

Example II: The century of biology

- We now know that humans are coded in 3 billion DNA letters (A,T,C,G), but what do they mean?
- We know many diseases are highly inheritable (heart problems, cancer, diabetes, etc.) so information regarding these diseases must be contained in these 3 billion letters.
- If we can tell which genes contribute to the disease risk, then we can make *personalized* diagnosis and treatment.
- How to map these genes to diseases?
- We need statistics!

Case study: Microbiome genomics

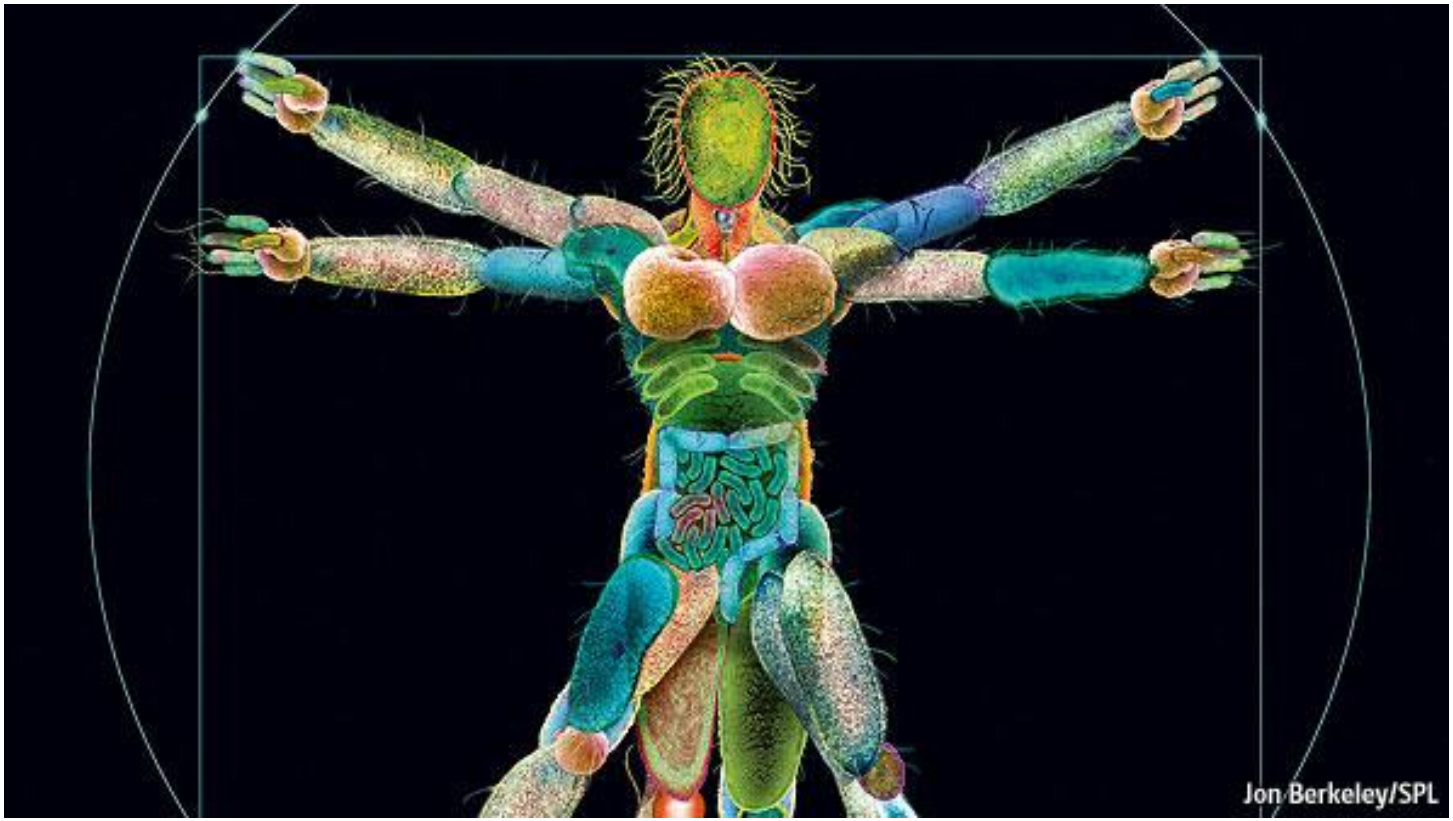

Figure source: `http://lumibyte.eu/`

- Vast majority of the DNAs in your body is not yours!
- Comparing across groups—multiple levels of variation.
- Effects of various treatment and environment factors.
- Dream objective: manipulating microbiome to improve treatment and overall health.

Example III: Statistical arbitrage in high frequency trading

- The prices of tens of thousands of securities are recorded every fraction of a second.
- Can we detect patterns (deviation from equilibrium) in the prices *quickly*, so that an arbitrage opportunity can be created in the next few seconds?
- What are true "signals"? What are just chance? How confident are we?
- We need statistics!

Moral of the story

- We are living in an era of data *explosion*. The data contain all kinds of information that can make the world a better (. . . or worse) place.
- Technological advances have made the generating, transporting, and storing of huge quantities of data possible.
- It is the statistician's job to make sense of the data.
- There has *never* been a better time to become a *statistician*.
  - A world of new exciting possibilities and challenges.

Prerequisites

- Calculus based probability at the level of *STA 230/MTH 230*. Homework 1 is a diagnostic test.
- It will count as 1% of the course grade and will be graded only based on *participation*.
- Please come talk to me if you have not taken STA 230.

- Please review probability theory (Chs 1 to 6 in the textbook).
- It's a diagnostic test so please complete it *independently* to allow me to have an *unbiased estimate* of your background.
- It also should give you a sense of whether this is the right class for you to take without more preparation.
- The material is covered in STA 230 and can be found in Chs 1 to 6 of the textbook.
- If you find Homework 1 difficult, *you will need to master these topics within the first two weeks*. Otherwise this course will become *extremely* challenging at some point.

- The course website: `http://www2.stat.duke.edu/~lm186/Courses/250/F17/sta250.html`
- Make-up lecture on Monday, November 20.

Probability modeling and statistical inference

- Statistics is the science of *extracting information from data*, while accounting for the *uncertainty*.
- Probability is the tool for formulating *uncertainty* in a mathematical fashion.
  - What is a model?
  - What is a probability model?

Probability modeling and statistical inference

- In STA 230 we have learned several families of probability distributions.
  - e.g. Normal, Binomial, Negative Bionomial, Geometric, Exponential, Poisson, Gamma, etc. (You need to be very familiar with these!)
  - These are the basic building blocks for more complex data generative mechnisms.
- Probability modeling is the art of making the appropriate *assumptions* about the underlying randomness in the data.
  - For example: What family of distributions to use? What additional assumptions—such as independence, stationarity, etc?
  - "All models are wrong, but some are useful." —George Box.
- Based on these assumptions, statistics aims at "completing the picture"— drawing inference ("highly educated guess") about various aspects of the underlying random mechanisms.

Common statistical inference problems

- Estimation: What is the value of certain quantities of interest?
  - Prediction: What may be the value of a future observation?
- Hypothesis testing: Are certain assumptions reasonable, or not (in comparison to some alternatives)?

In real problems, statistical inference and probability modeling typically form a *two-way* process.

- Model construction $\longleftrightarrow$ model application/validation.

Example: Coin tossing

- I've got a coin which I tossed 10 times.
- My data: *HHTHHTHTHH*.
- Q: What can be said about the chance for getting *H*?
  - Estimation: What's the chance of getting a head on each toss?
  - Hypothesis testing: Is this a fair coin? Are the tosses independent?

What *probability model* can we use to represent this *experiment*?

Statistical experiment and random variables

- *Experiment*: A process, planned or unplanned, with an (observable or unobservable) outcome, ω.
- *Outcome space*: The collection of all possible outcomes, denoted by Ω.
  - When the outcome is observable, as in the coin tossing example, we often call the outcome space the *sample space* of the experiment.
- *Event*: A subset of the outcome space. (In some cases not all subsets are events, but we don't have to worry about that in this course. See Section 1.4 of textbook.)
- *Random variable*: A real-valued function defined on Ω.
  - $X : \Omega \to \mathbb{R}$
  - The set of values X can take, i.e.,

$$\{X(\omega) : \omega \in \Omega\}$$

  is called the *sample space* of the random variable X.

Example: A coin flipping experiment

- An experiment that consists of a series of n *Bernoulli* trials.
- The outcome of each single trial is either 1, called a success, or 0, called a failure.

What is the outcome space?

$$\Omega = \{0,1\} \times \{0,1\} \times \ldots \times \{0,1\} = \{0,1\}^n.$$

For example, for a fair coin,

- $n = 10$, "1"=H.
- Each possible outcome ω is of the form 1001100101.
- The event that we get six heads is the set of all strings in which there are six 1's.

Random variables defined on this sample space

- For example,

$$X = \#\text{ heads} = \#\text{ 1's}$$

  is a *discrete* random variable.
- More explicitly, $X : \Omega \to \mathbb{R}$, with

$$X(\omega) = \#\text{ 1's in the outcome } \omega.$$

  What is the sample space of X?
- Another example,

$$Y = \#\text{ tails that follow another tail.}$$

A probability model: the Binomial model

- Two *assumptions* regarding the randomness of this experiment:
  - Each trial is independent.
  - The chance of getting a 1 in each trial is the same.
- Based on this model, by the *multiplication rule*, the probability of each observable outcome given the parameter value p is

$$P(1101101011|p) = p \times p \times (1-p) \times \ldots \times p \times p = p^7(1-p)^3.$$

- What is the *probability mass function* (pmf) of X, the number of heads?

The Binomial(n,p) distribution

- The pmf is

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0,1,2,\ldots,n$$
$$= 0 \quad \text{otherwise.}$$

- The *parameter* $p \in [0,1]$ determine this distribution.
- The Binomial distribution is an example of a *parametric family* of probability distributions.
  - A family (or collection) of distributions indexed by a finite number of parameters.
- *Parametric*, *semi-parametric*, and *nonparametric* statistical inference.

Statistical inference (you will learn in this class!)

- Possible *inference* question: What is the head probability p?
- As we will learn in this course, based on this model, a common *estimator* for the p is the value of p that maximizes this probability

$$\hat{p} = \#\ 1\text{'s}/n = X/n,$$

- But just having a point guess is not sufficient. How certain are we about the guess?
- An estimate for the *standard deviation* of X is

$$SE(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})/n}.$$

Another possible inference question

- Suppose instead of guessing the value of p, we have some *hypothesis* about its value, e.g., $p = 0.5$.
- Does the data support/contradict the hypothesis?
- For example, suppose in our observed data, $\hat{p} = 0.7$, what should we conclude?
- Apparently uncertainty quantification is needed to make such a conclusion.
- Now suppose $SE(\hat{p}) \approx 0.14$. What can we conclude?

The two-way process

- If our observed data is *HHHHHHHTTT*, the Binomial model may not be appropriate. Why?
- Recall the two assumptions.
- Can this stickiness be purely due to randomness?
- You will learn ways in which to test *hypotheses* like this.
- If such a hypothesis is *rejected*, one may consider modify the model to better represent the data.

Sampling and Bayesian approaches

- Two schools of thoughts on how inference should proceed.
- One assumes that the unknown parameters are *fixed*. Inference is made based *only* on the probability of the data *given* the "true" but unknown parameter p. (The *sampling* approach.)
- There is another class of statistical thinking called the *Bayesian* approach, which *treat the parameters as random variables as well* and draw inference based on the *joint* distribution of the data and the parameters.
- In this class, we will learn both and examine the strength and weakness of each.
- Which approach do you think is more natural?

Seven pillars of statistical wisdom (Stigler 2016)

Wisdom has built her house;
She has hewn out her seven pillars.
- Proverbs 9:1

1. Aggregation of information.
2. Diminishing information.
3. Mathematical quantification of information/uncertainty.
4. Intercomparisons.
5. Regression and multivariate analysis.
6. Design.
7. Models and residuals.
8. *Smoothing/regularization/multiple testing (avoiding over-interpretation of data).* (Not in Stigler's book.)