What this course is about?

- This course is about Statistical Inference.
- We build on foundations of probability & statistics that we revisited during the last term.

Statistics VS Computer Science

- Some time ago, statisticians were doing their thing in statistics departments, and machine learning scientists were doing their thing in computer science departments. Statisticians thought that computer scientists were reinventing the wheel, computer scientists thought that statistics didn’t apply to their problems.

- Statisticians now recognize that computer scientists are making novel contributions while computer scientists now recognize the generality of statistical theory and methodology.

What is Statistical Inference?

- Statistical Inference, or “learning”, is the process of using data to infer the distribution that generated the data.

- A typical statistical inference question is:

Given a sample $X_1, \ldots, X_n \sim F$, how do we infer F ?

Parametric and Nonparametric Models

- **A statistical model** $\mathscr{F}$ is a set of distributions (or densities or regression functions).
- A **parametric model** is a set $\mathscr{F}$ that can be parametrized by a finite set of parameters.

  For example, if we assume that the data come from a normal distribution, the model is $\mathscr{F} = \left\{ f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right), \ \ \mu \in \mathbb{R}, \sigma > 0 \right\}$

  This is a two-parameter model. x is the random variable, μ, σ are parameters

Parametric and Nonparametric Models

- In general, a parametric model takes the form

  $\mathcal{F} = \left\{ f(x;\theta) : \ \theta \in \Theta \right\}$ where θ is an unknown parameter (or vector of parameters) that can take values in the **parameter space** Θ.

- If θ is a vector but we are only interested in one component of θ, we call the remaining parameters **nuisance parameters**.
- A **nonparametric model** is a set $\mathcal{F}$ that cannot be parametrized by a finite number of parameters.

  For example, $\mathcal{F} = \{\text{all CDF's}\}$ is nonparametric.

Parametric and Nonparametric Models: Notation

- If $\mathscr{F} = \{f(x;\theta) : \theta \in \Theta\}$ is a parametric model, we write

$$\mathbb{P}_\theta(X \in A) = \int_A f(x;\theta)\,dx \quad \text{and} \quad \mathbb{E}_\theta\big(r(x)\big) = \int r(x) f(x;\theta)\,dx.$$

Subscript θ indicates that the probability or expectation is with respect to $f(x;\theta)$, it does not mean we are averaging over θ.

Parametric and Nonparametric Models: Examples

- **Example 1** (One-dimensional Parametric Estimation). Let $X_1, \ldots, X_n$ be independent Bernoulli(p) observations. The problem is to estimate the parameter p.
- **Example 2** (Two-dimensional Parametric Estimation). Suppose that$X_1, \ldots, X_n \sim F$ and we assume normal distribution. There are two parameters, μ and σ. The goal is to estimate the parameters from the data. If we are only interested in estimating μ, then μ is the parameter of interest and σ is a nuisance parameter.

Parametric and Nonparametric Models: Examples

- **Example 3** (Nonparametric estimation of the CDF). Let $X_1, \ldots, X_n$ be independent observations from a CDF F. The problem is to estimate F.
- **Example 4** (Nonparametric density estimation). Let $X_1, \ldots, X_n$ be independent observations from a CDF F and let $f = F'$ be the PDF, that we want to estimate. It is not possible assuming only that F is just some proper CDF – we need to assume some smoothness on f. For example, considering only

$$f \in \left\{ f : \int (f''(x))^2 dx < \infty \right\}$$ – functions from a **Sobolev space**

Parametric and Nonparametric Models: Examples

- **Example 5** (Nonparametric estimation of functionals). Let $X_1, \ldots, X_n \sim F$. Suppose we want to estimate μ assuming only that it exists. The mean μ can be thought of as a function of F:

$$\mu = T(F) = \int x \, dF(x).$$

In general, any function of F is called a **statistical functional**. Other examples of functionals are the variance $T(F) = \int x^2 \, dF(x) - \left(\int x \, dF(x) \right)^2$ and the median $T(F) = F^{-1}(1/2)$.

Parametric and Nonparametric Models: Examples

- **Example 6** (Regression, prediction and classification). Suppose we observe pairs of data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Perhaps X_i is the blood pressure of subject i and Y_i is how long they live.

  X is called a **predictor** or **regressor** or **feature** or **independent variable**.

  Y is called the **outcome** or the **response variable** or the **dependent variable**.

  We call $r(x) = \mathbb{E}(Y|X = x)$ the **regression function**.

  If we assume $r \in \mathscr{F}$ where $\mathscr{F}$ is finite dimensional – the set of straight lines, for example – we have a **parametric regression model**.

  If not – we have a **nonparametric regression model**.

Parametric and Nonparametric Models: Examples

- **Example 6** (Regression, prediction and classification). Suppose we observe pairs of data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Perhaps X_i is the blood pressure of subject i and Y_i is how long they live.

  The goal of predicting Y for a new patient based on their X value is called **prediction**.

  If Y is discrete (for example, live or die) then prediction is instead called **classification**.

  Regression models are sometimes written as $Y = r(X) + \epsilon$ where $\mathbb{E}(\epsilon) = 0$. We can always rewrite a regression model this way.

Fundamental Concepts in Inference: Point Estimation

- **Point estimation** is providing a single "best guess" of some quantity of interest. That could be: a parameter of a parametric model / a CDF F / a PDF f / a regression function r / a prediction for a future value Y of some r.v.
- We denote a point estimate of θ by $\hat{\theta}$ (or $\hat{\theta}_n$ if we inferred it from n data points). Remember that θ is a fixed, unknown quantity. The estimate $\hat{\theta}$ depends on the data, so it is a random variable.
- More formally, let $X_1, \ldots, X_n$ be IID data points from some distribution. A point estimator $\hat{\theta}_n$ of a parameter θ is some function of the data points:

$$\hat{\theta}_n = g(X_1, \ldots, X_n)$$

Fundamental Concepts in Inference: Point Estimation

- The bias of an estimator is defined by $\text{bias}(\hat{\theta}) = \mathbb{E}_\theta(\hat{\theta}) - \theta$

  We say that $\hat{\theta}$ is **unbiased** if $\text{bias}(\hat{\theta}) = 0$. Unbiasedness used to receive much attention, but these days is considered less important; many estimators we will use are biased.

- It is reasonable to require that an estimator converges to the true parameter value as we collect more and more data:

  Definition: A point estimator $\hat{\theta}$ of a parameter θ is consistent if $\hat{\theta}_n \xrightarrow{P} \theta$.

Fundamental Concepts in Inference: Point Estimation

- The distribution of $\hat{\theta}_n$ is called the **sampling distribution.** The standard deviation of $\hat{\theta}_n$ is called the **standard error**, denoted by:

$$\text{se} = \text{se}(\hat{\theta}_n) = \sqrt{\mathbb{V}(\hat{\theta}_n)}$$

- Often this standard error depends on the unknown F. In this case, se is an unknown quantity, but one can usually estimate it. The estimated standard error is denoted by $\hat{\text{se}}$.

Fundamental Concepts in Inference: Point Estimation

- **Example.** Let $X_1, \ldots, X_n \sim$ Bernoulli(p) and let $\hat{p}_n = n^{-1} \sum_i X_i$

Then $\mathbb{E}(\hat{p}_n) = n^{-1} \sum_i \mathbb{E}(X_i) = p$ so $\hat{p}_n$ is unbiased.

The standard error is $\text{se} = \sqrt{\mathbb{V}(\hat{p}_n)} = \sqrt{p(1-p)/n}$.

The estimated standard error is $\hat{\text{se}} = \sqrt{\hat{p}(1-\hat{p})/n}$

Fundamental Concepts in Inference: Point Estimation

- The quality of a point estimate is sometimes assessed by the **mean squared error**, or MSE defined by

$$\text{MSE} = \mathbb{E}_\theta(\hat{\theta}_n - \theta)^2$$

- **Theorem.** The MSE can be written as

$$\text{MSE} = \text{bias}^2(\hat{\theta}_n) + \mathbb{V}_\theta(\hat{\theta}_n)$$

Fundamental Concepts in Inference: Point Estimation

- **Theorem.** The MSE can be written as

$$\text{MSE} = \text{bias}^2(\hat{\theta}_n) + \mathbb{V}_\theta(\hat{\theta}_n)$$

Proof: Let $\bar{\theta}_n = \mathbb{E}_\theta(\hat{\theta}_n)$. Then we have

$$\mathbb{E}_\theta(\hat{\theta}_n - \theta)^2 = \mathbb{E}_\theta(\hat{\theta}_n - \bar{\theta}_n + \bar{\theta}_n - \theta)^2 =$$

$$= \mathbb{E}_\theta(\hat{\theta}_n - \bar{\theta}_n)^2 + 2(\bar{\theta}_n - \theta)\mathbb{E}_\theta(\hat{\theta}_n - \bar{\theta}_n) + \mathbb{E}_\theta(\bar{\theta}_n - \theta)^2 =$$

$$= (\bar{\theta}_n - \theta)^2 + \mathbb{E}_\theta(\hat{\theta}_n - \bar{\theta}_n)^2 = \text{bias}^2(\hat{\theta}_n) + \mathbb{V}(\hat{\theta}_n)$$

Fundamental Concepts in Inference: Point Estimation

- **Theorem.** If bias $\to 0$ and se $\to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is consistent – that is, $\hat{\theta}_n \xrightarrow{P} \theta$.

  Proof: If bias $\to 0$ and se $\to 0$, then MSE $\to 0$.

  So $\hat{\theta}_n \xrightarrow{qm} \theta$ – converges in quadratic mean ($X_n \xrightarrow{qm} X$ if $\mathbb{E}(X_n - X)^2 \to 0$).

  Convergence in quadratic mean implies convergence in probability.

- Recall the **Bernoulli example**. There with $\mathbb{E}_p(\hat{p}_n) = p$ we have bias $= p - p = 0$ and se $= \sqrt{p(1-p)/n} \to 0$. Hence $\hat{p}_n \xrightarrow{P} p$, so $\hat{p}_n$ is a consistent estimator.

Fundamental Concepts in Inference: Point Estimation

- **Definition.** An estimator is **asymptotically normal** if

$$\frac{\hat{\theta}_n - \theta}{\text{se}} \xrightarrow{d} \mathcal{N}(0,1)$$

(standardized, it converges in distribution to the standard normal).

- Central Limit Theorem (CLT) implies asymptotic normality of the sample mean as an estimate of the true mean

Fundamental Concepts in Inference: Confidence Sets

- A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and same for b, are functions of the data such that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \text{ for all } \theta \in \Theta$$

- Typically α is chosen to be 5% – so $1 - \alpha$ is 95%.
- If θ is a vector, then we use a **confidence set** (a sphere or an ellipse) instead of an interval

Fundamental Concepts in Inference: Confidence Sets

- A $1-\alpha$ **confidence interval** for a parameter θ is an interval $C_n = (a, b)$ where $a = a(X_1, \ldots, X_n)$ and same for b, are functions of the data such that

$$\mathbb{P}_\theta(\theta \in C_n) \geq 1 - \alpha, \text{ for all } \theta \in \Theta$$

- C_n is random, and θ is not! A proper interpretation is: if we repeat the experiment over and over again, the interval will contain the parameter 95% of the time
- **Example:** An opinion poll says “83 percent of the population favor arming pilots with guns. This poll is accurate to within 4 points 95% of time”

Fundamental Concepts in Inference: Confidence Sets

- Recall the Bernoulli **example**: for the estimate $\hat{p}_n = n^{-1} \sum_i X_i$ and given α confidence interval would be

  $C_n = (\hat{p}_n - \epsilon_n, \hat{p}_n + \epsilon_n)$, where $\epsilon_n^2 = \log(2/\alpha)/(2n)$

Follows from (Hoeffding's inequality) the fact that for $X_1, \ldots, X_n \sim$ Bernoulli(p) and for any $\epsilon > 0$ we have

$$\mathbb{P}(|\bar{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}$$

Fundamental Concepts in Inference: Hypothesis testing

- In **hypothesis testing**, we start with some default theory – called a **null hypothesis** – and ask if the data provide sufficient evidence to reject the theory. If not – we retain the null hypothesis.
- **Example** (Testing if a coin is fair). Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$ be n independent coin flips. We want to test if the coin is fair. Let H_0 denote the **null-hypothesis** (that the coin is fair) and H_1 – the **alternative hypothesis**:

  $$H_0 : \ p = 1/2, \quad H_1 : \ p \neq 1/2$$

  it is reasonable to reject H_0 if $T = |\hat{p}_n - 1/2|$ is large.