# Lecture #5

- The central idea of statistics is to infer from data.
- That is, let $X$ be some unknown r.v. If we observe $X_1, X_2, \dots, X_n \overset{iid}{\sim} X$, what can we say about $X$ itself?
- Possible questions about $X$:
  - What is the density of $X$? This is often very hard
  - Does $X$ look like it comes from a certain family of distributions?
  - What is $\mathbb{E}(X)$? $Var(X)$? More generally,

$$\int_{\mathbb{R}} g(x) \cdot f_X(x)\,dx \quad \text{for some } g(x).$$

- Two broad directions: parametric v. nonparametric

parametric $\downarrow$
- makes strong assumptions
- easy to work with

nonparametric $\downarrow$
- few assumptions $\rightarrow$ very flexible
- more difficult computationally

<u>Defn</u>: A collection of functions $\mathcal{F}$ is a parametric model if it can be parametrized by a finite number of parameters i.e. There exists a finite dimensional parameter space $\Theta = \{(\theta_1, \theta_2, \dots, \theta_D)\}$ s.t. every element of $\mathcal{F}$ is associated with an element of $\Theta$.

<u>ex</u>: Everything we saw in MATH 165 (more or less).

ex: Normal distributions in $\mathbb{R}$ is a parametric model with two parameters: $f(x)$ is a normal density if and only if there exist $\mu \in \mathbb{R}$ ("mean"), $\sigma^2 \in \mathbb{R}_+$ ("variance")

s.t. $$f(x) = \frac{1}{\sqrt{2\sigma^2}} \cdot \exp\left( \frac{-[x-\mu]^2}{2\sigma^2} \right).$$

- Parametric models make identifying a good estimate for the true density easier... just estimate the finitely many parameters! How hard can that be...

- If we consider $\mathcal{F}$ to not have a finite parametrization, this is non-parametric statistics.

ex: Let $\mathcal{F} = \left\{ f \in \mathcal{C}(\mathbb{R}) \,\middle|\, f(x) \geq 0 \text{ and } \int_{\mathbb{R}} f(x)\,dx = 1 \right\}$ be the space of all continuous densities. This is a big space of functions, and doesn't have a useful finite parametrization.

- Contrast: Given data $\{x_i\}_{i=1}^n \overset{iid}{\sim} \underline{X}$ for $\underline{X}$ unknown,

  (1.) Parametric: Find the best Gaussian to model $\underline{X}$

  (2.) Nonparametric: Find the best continuous density to model $\underline{X}$

- Obviously (2.) is less constrained.

- An underlying issue with statistics is how to decide if I have predicted well!
- That is, given data $X_1, \dots, X_n \overset{iid}{\sim} \underline{X}$, how do I know if my predicted density or predicted [redacted] $\mathbb{E}(\underline{X})$ is good or not? This is a problem that never really goes away, and can be studied on many levels.
- We will focus on the "classical" approach to understanding error in statistical estimation: bias-variance tradeoff.

/

- Density estimation is hard, and we will focus on an easier problem for now: point estimation, i.e. estimating a particular quantity of interest. We are especially interested in estimating the expected value of $\underline{X}$, given only observations $X_1, \dots, X_n \overset{iid}{\sim} \underline{X}$.
- More generally, let $\theta$ be an underlying quantity of interest, quite of a parameter in a parametric model (e.g. $\theta = \mathbb{E}(\underline{X})$). $\theta$ itself is unknowable, so we would like to estimate it given (random) data $X_1, \dots, X_n \sim \underline{X}$.
- We can think of a method of estimation as a function on the data. This is a fundamental insight of statistics, because it allows for the machinery of analysis and probability to give us control.
- So, we will write [redacted] $\hat{\theta}(X_1, \dots, X_n)$ as a function that estimates $\theta$, given

data $X_1, \dots, X_n \overset{iid}{\sim} \mathbb{X}$.

ex: Let $X_1, \dots, X_n \sim \mathbb{X}$, where $\mathbb{E}(\mathbb{X})$ exists. A classical problem is to estimate $\mathbb{E}(\mathbb{X})$. It is almost an article of faith that we should use $\frac{1}{n}\sum_{i=1}^{n} X_i$.

In 1-dimension, a quick simulation shows this to be effective for many $\mathbb{X}$ if $n$ is large enough.

In fact the WLLN says this is a good idea! Indeed, WLLN says $\forall \varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(\mathbb{X})\right| > \varepsilon\right) = 0.$$

So, if

$$\left.\begin{aligned} \theta &= \mathbb{E}(\mathbb{X}) \\ \underbrace{\hat{\theta}(x_1, \dots, x_n)}_{\hat{\theta}_n} &= \frac{1}{n}\sum_{i=1}^{n} x_i, \end{aligned}\right\} \xrightarrow{\text{WLLN}} \forall \varepsilon > 0, \ \lim_{n\to\infty} \mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) = 0.$$

Pretty good!

This motivates the idea of consistency.

Defn: A point estimator $\hat{\theta}(x_1, \dots, x_n) = \hat{\theta}_n$ is consistent if $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$.

- A key insight is that $\hat{\theta}_n$ is a r.v. So, we can consider its expectation and variance.

**Defn**: Let $\hat{\theta}_n$ be an estimator of $\theta$. The **bias** of $\hat{\theta}_n$ is $\text{bias}(\hat{\theta}_n) = \mathbb{E}(\hat{\theta}_n) - \theta$, where the expectation is taken over $X_1, \dots, X_n \sim X$. We say $\hat{\theta}_n$ is <u>unbiased</u> if $\text{bias}(\hat{\theta}_n) = 0$.

- Let $\text{Var}(\hat{\theta}_n) = \mathbb{E}\left(\left[\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right]^2\right)$ be the variance of $\hat{\theta}_n$ with respect to the sample $X_1, \dots, X_n \sim X$. We are typically interested in understanding the "total" error of estimating:

$$\text{MSE}(\hat{\theta}_n) = \mathbb{E}\left(\left[\hat{\theta}_n - \theta\right]^2\right), \quad \text{the mean square error.}$$

- Amazingly, this is interpretable into two components: model part (bias) v. random part (variance).

**Theorem (Bias-Variance Tradeoff)**: $\text{MSE}(\hat{\theta}_n) = \text{bias}(\hat{\theta}_n)^2 + \text{Var}(\hat{\theta}_n)$.

**Proof**: $\mathbb{E}([A+B]^2) = \mathbb{E}A^2 + 2AB + \mathbb{E}B^2$

$$\mathbb{E}\left(\left[\hat{\theta}_n - \theta\right]^2\right)$$

$$= \mathbb{E}\left(\left[\overbrace{\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)}^{=A} + \overbrace{\mathbb{E}(\hat{\theta}_n) - \theta}^{=B}\right]^2\right)$$

$$= \underbrace{\mathbb{E}\left(\left[\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right]^2\right)}_{=\text{Var}(\hat{\theta}_n)} + 2\,\mathbb{E}\left(\left[\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right]\overbrace{\left[\mathbb{E}(\hat{\theta}_n - \theta)\right]}^{\text{constant}}\right) + \mathbb{E}\left(\overbrace{\left[\mathbb{E}(\hat{\theta}_n) - \theta\right]^2}^{\text{constant}}\right)$$

$$= \text{Var}(\hat{\theta}_n) + 2\left(\mathbb{E}(\hat{\theta}_n) - \theta\right) \cdot \overbrace{\mathbb{E}\left[\hat{\theta}_n - \mathbb{E}(\hat{\theta}_n)\right]}^{=0} + \overbrace{\left[\mathbb{E}(\hat{\theta}_n) - \theta\right]^2}^{=\text{bias}(\hat{\theta}_n)^2}$$

$$= \text{Var}(\hat{\theta}_n) + \text{bias}(\hat{\theta}_n)^2. \quad \blacksquare$$

- $\text{bias}(\hat{\theta}_n) \approx 0$ if there is enough capacity in the model to fit it well. $\text{Var}(\hat{\theta}_n) \approx 0$ if we have enough data to reliably estimate.