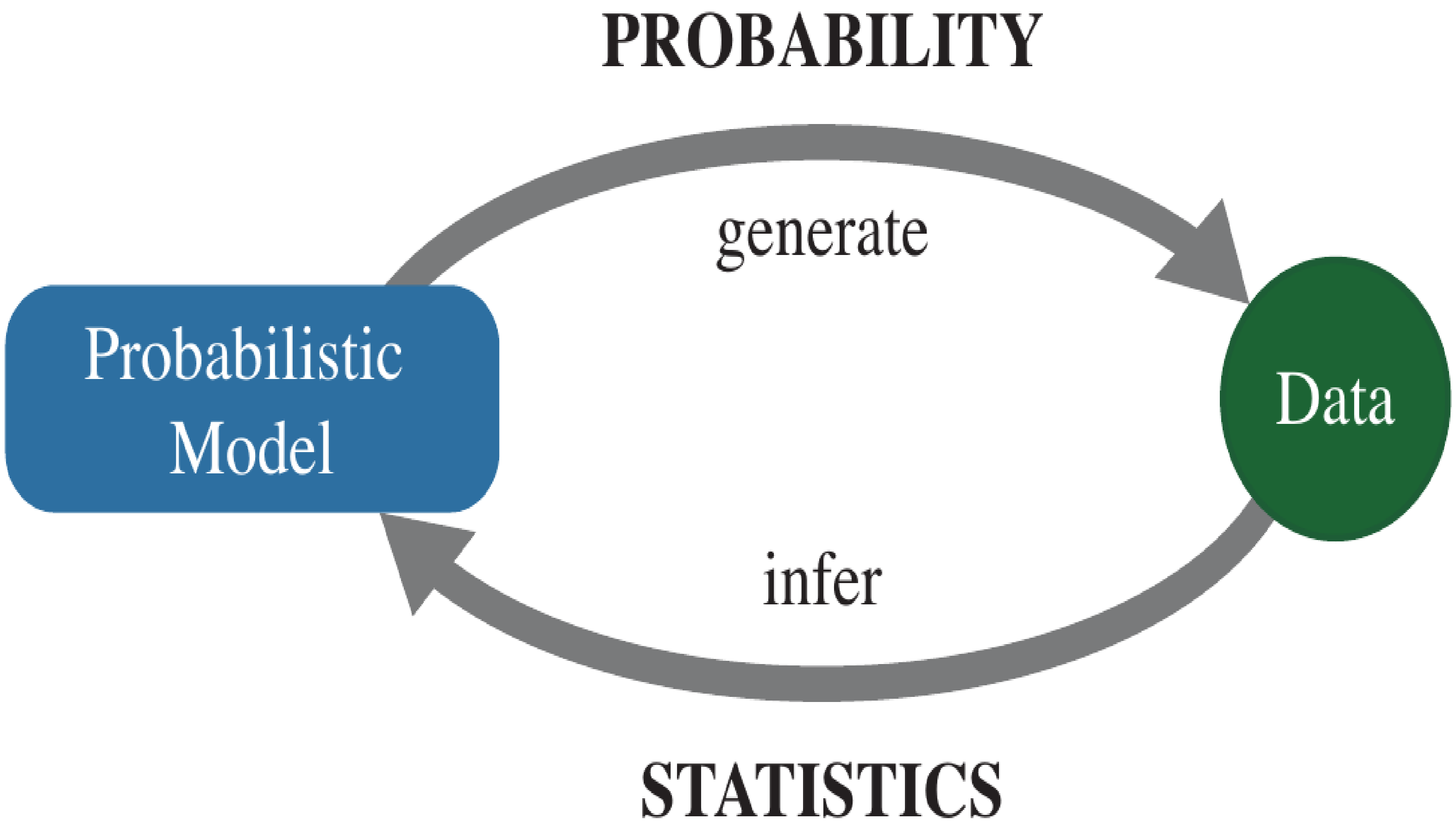
MA 6.101
Probability and Statistics

Tejas Bodas

Assistant Professor, IIIT Hyderabad

Statistics


PROBABILITY
generate
Probabilistic Model
Data
infer
STATISTICS

Statistical Inference

- Statistical Inference methods deal with drawing inference about an unknown model/**random variable**/random process from observations/data.
- There is an unknown quantity θ^* that we would like to estimate using data $\mathcal{D}$. eg: ML, communication systems.
- For the purpose of this course $\mathcal{D}$ will contain samples of a random variable and θ^* could be mean, variance, moments or parameters of the underlying random variable.
- Broadly, you can give 3 types of estimates for θ^*.
  1. Point Estimation: Here you want to give a point estimate which is a single numerical value that is your best guess for θ^*.
  2. Interval Estimation: here you give an interval on say $\mathbb{R}$ where θ^* is bound to lie with some certainty.
  3. Hypothesis testing: In binary hypothesis testing, you have two hypothesis ($H_0 : \theta = \alpha_1$ and $H_1 : \theta = \alpha_2$) and you use data $\mathcal{D}$ to decide which is true.

Statistical Inference

- There are two approaches to Statistical Inference:
  1) Bayesian 2) Frequentist (or classical)
- In Bayesian Inference, the unknown quantity is modelled as a random variable with a distribution that keeps changing as more and more data becomes available.
- Bayesian inference assumes a prior distribution $p_{\Theta}(\theta)$ on the unknown parameter θ^* and uses the likelihood $p_{X|\Theta}(x|\theta)$ for observing data x to obtain the posterior $p_{\Theta|X}(\theta|x)$
- In Bayesian inference, prior and posterior distribution reflect our state of knowledge.

Frequentist or Classical Inference

- Classical Inference models the unknown quantity as a constant and come up with estimators that are deterministic functions of the observed data.
- Given data, these estimators are deterministic functions of the data, but in reality are also random variables.
- For example sample mean as an estimator for the mean.

Classical Inference: Point Estimation

- Let θ^* denote the unknown parameter of a random variable X (typically mean, variance, scale, shape etc) and suppose we observe i.i.d samples of X which are recorded in the dataset $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$.
- In frequentist approach, we estimate θ^*, by defining a point estimator $\hat{\Theta}$ as a function of the random samples $X_1, \ldots X_n$ as $\hat{\Theta} = h(X_1, \ldots X_n)$.
- While $\hat{\Theta}$ is a random variable, given $\mathcal{D}$ the estimator takes the value $\hat{\Theta} = h(x_1, \ldots x_n)$.
- Example : Sample mean $\hat{\mu}_n = \frac{\sum_{i=1}^{n} X_i}{n}$.

Point Estimators: Properties

- The Bias $B(\hat{\Theta})$ of an estimator $\hat{\Theta}$ is defined as

$$B(\hat{\Theta}) = E[\hat{\Theta}] - \theta^*$$

- Unbiased estimators are estimators with zero bias, i.e., $B(\hat{\Theta}) = 0$ and hence $E[\hat{\Theta}] = \theta^*$
- Are all unbiased estimators good ? Let $\hat{\Theta}_1 = X_1$ and $\hat{\Theta}_2 = \frac{\sum_{i=1}^{n} X_i}{n}$. Which estimator is better?
- $Var(\hat{\Theta}_1) = \sigma^2$ while $Var(\hat{\Theta}_2) = \frac{\sigma^2}{n}$.
- We need other measures to determine how good an estimator is, something that looks at the variance of these estimators.

Mean square error of Point Estimators

- The mean squared error of an estimator $\hat{\Theta}$ is defined as

$$MSE(\hat{\Theta}) = E[(\hat{\Theta} - \theta^*)^2]$$

- Note that

$$\begin{aligned} MSE(\hat{\Theta}) &= E[(\hat{\Theta} - \theta^*)^2] \\ &= Var(\hat{\Theta} - \theta^*) + E[\hat{\Theta} - \theta^*]^2 \\ &= Var(\hat{\Theta}) + Bias(\hat{\Theta})^2 \end{aligned}$$

- This means that biased estimators could possibly have lower MSE error if they have extremely low variance!
- Find MSE of $\hat{\Theta}_1 = X_1$ and $\hat{\Theta}_2 = \hat{\mu}_n + 1$.
- Bias-Variance tradeoff talks a lot in machine learning!

Consistency of estimators

- What happens to estimators as the size of the data set ($|\mathcal{D}| = n$) increases? Do all estimators converge to θ^*?
- Not necessarily! For examples $\hat{\Theta}_1 = X_i$ where X_i is picked random from $\mathcal{D}$ does not converge.
- What about $\hat{\mu}_n$. Using SLLN, we see that this does.
- Let $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_n, \ldots,$ be a sequence of point estimators of θ^* (here n denotes the size of the dataset) We say that $\hat{\Theta}_n$ is a **consistent estimator** of θ, if

$$\lim_{n \to \infty} P(|\hat{\Theta}_n - \theta^*| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0$$

- This is convergence in probability. If almost sure convergence holds, it is called strongly consistent.
- Clearly, $\hat{\Theta}_n = \hat{\mu}_n$ is strongly consistent and hence consistent.

RECAP

- A point estimator $\hat{\Theta}$ is a function of the random samples $\hat{\Theta} = h(X_1, \ldots X_n)$
- The Bias $B(\hat{\Theta})$ of an estimator $\hat{\Theta}$ is defined as

$$B(\hat{\Theta}) = E[\hat{\Theta}] - \theta^*$$

- The mean squared error of an estimator $\hat{\Theta}$ is defined as

$$MSE(\hat{\Theta}) = E[(\hat{\Theta} - \theta^*)^2]$$

.

- $MSE(\hat{\Theta}) = Var(\hat{\Theta}) + Bias(\hat{\Theta})^2$
- We say that $\hat{\Theta}_n$ is a **consistent estimator** of θ, if

$$\lim_{n\to\infty} P(|\hat{\Theta}_n - \theta^*| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0$$

Markov's Inequality: Statement

Markov's Inequality: Let X be a non-negative random variable, and let $a > 0$. Then:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

Key Points:

- Applies to **non-negative** random variables.
- Provides an upper bound on the probability of large deviations.
- Useful in analyzing tail probabilities.

Proof of Markov's Inequality

Proof:
Let X be a positive continuous random variable. We start by writing the expectation $\mathbb{E}[X]$ as:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\, dx = \int_{0}^{\infty} x f_X(x)\, dx \quad (\text{since } X \geq 0).$$

For any $a > 0$, we can split the integral as follows:

$$\mathbb{E}[X] = \int_{0}^{a} x f_X(x)\, dx + \int_{a}^{\infty} x f_X(x)\, dx.$$

Thus,

$$\mathbb{E}[X] \geq \int_{a}^{\infty} x f_X(x)\, dx.$$

Proof of Markov's Inequality (cont'd)

Since $x \geq a$ for $x \in [a, \infty)$, we have

$$\int_a^\infty x f_X(x)\,dx \geq \int_a^\infty a f_X(x)\,dx = a \int_a^\infty f_X(x)\,dx.$$

Now, we recognize that $\int_a^\infty f_X(x)\,dx = \mathbb{P}(X \geq a)$, so:

$$\mathbb{E}[X] \geq a \cdot \mathbb{P}(X \geq a).$$

Dividing both sides by a (for $a > 0$), we conclude:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

$\square$