More on Law of Total Expectation + Intro Maximum Likelihood Estimation

CSE 312 Spring 26

Lecture 23

Agenda

- Recap + more on Law of Total Expectation
- Introduction to statistical estimation
- Maximum Likelihood Estimation

Law of total probability

Definition. Let A be an event and Y a discrete random variable. Then

$$P[A] = \sum_{y \in \Omega_Y} P(A|Y = y) p_Y(y)$$

Definition. Let A be an event and Y a continuous random variable. Then

$$P[A] = \int_{-\infty}^{\infty} P(A|Y = y) f_Y(y) \mathrm{d}y$$

Conditional Expectation

Definition. Let X be a discrete random variable then the **conditional expectation** of X given event A is

$$\mathbb{E}[X \mid A] = \sum_{x \in \Omega_X} x \cdot P(X = x \mid A)$$

Notes:

- Can be phrased as a "random variable version"

$$\mathbb{E}[X \mid Y = y]$$

- Linearity of expectation still applies here

$$\mathbb{E}[aX + bY + c \mid A] = a\,\mathbb{E}[X \mid A] + b\,\mathbb{E}[Y \mid A] + c$$

Law of Total Expectation

Law of Total Expectation (event version). Let X be a random variable and let events $A_1, \ldots, A_n$ partition the sample space. Then,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i] \cdot P(A_i)$$

Law of Total Expectation (random variable version). Let X be a random variable and Y be a discrete random variable. Then,

$$\mathbb{E}[X] = \sum_{y \in \Omega_Y} \mathbb{E}[X \mid Y = y] \cdot P(Y = y)$$

Example – Computer Failures (a familiar example)

Suppose your computer operates in a sequence of steps, and that at each step i your computer will fail with probability p (independently of other steps).

Let X be the number of steps it takes your computer to fail.

What is $\mathbb{E}[X]$?

Let Y be the indicator random variable for the event of failure in step 1

Then by LTE, $\mathbb{E}[X] = \mathbb{E}[X \mid Y = 1] \cdot P(Y = 1) + \mathbb{E}[X \mid Y = 0] \cdot P(Y = 0)$

$= 1 \cdot p + \mathbb{E}[X \mid Y = 0] \cdot (1 - p)$

$= p + (1 + \mathbb{E}[X]) \cdot (1 - p)$

since if $Y = 0$ experiment starting at step 2 looks like original experiment

Solving we get $\mathbb{E}[X] = 1/p$

Example -- Elevator rides

The number X of people who enter an elevator on the ground floor is a Poisson random variable with mean 10. If there are N floors above the ground floor, and if each person is equally likely to get off at any one of the N floors, independently of where others get off, compute the expected number of stops the elevator will make before discharging all the passengers.

Example -- Elevator rides

The number X of people who enter an elevator on the ground floor is a Poisson random variable with mean 10. If there are N floors above the ground floor, and if each person is equally likely to get off at any one of the N floors, independently of where others get off, compute the expected number of stops the elevator will make before discharging all the passengers.

$$\mathbb{E}[S] = \sum_{i \geq 0} \mathbb{E}[S|\, N = i] \cdot P(N = i)$$

$$= \sum_i \left(1 - \left(1 - \frac{1}{N}\right)^i \right) \cdot e^{-10} \cdot \frac{10^i}{i!}$$

$$E[X] = E[E[X|Y]]$$

Agenda

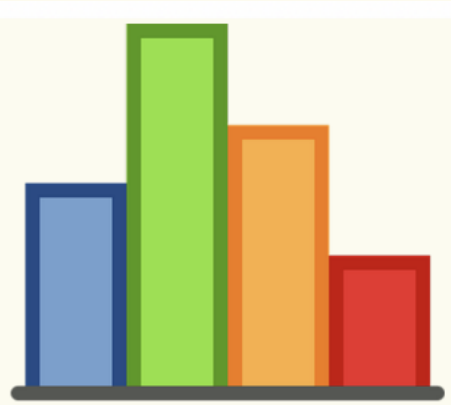
- Recap + more on Law of Total Expectation
- Introduction to statistical estimation
- Maximum Likelihood Estimation

Probability vs Statistics

$\text{Ber}(p = 0.5)$

Probability
Given model, predict data

$P(THHTHH)$



$\text{Ber}(p = ??)$

Statistics
Given data, predict model

$THHTHH$

Recap Formalizing Polls

We assume that poll answers $X_1, \ldots, X_n \sim \mathrm{Ber}(p)$ i.i.d. for unknown p

Goal: Estimate p

We did this by computing $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$

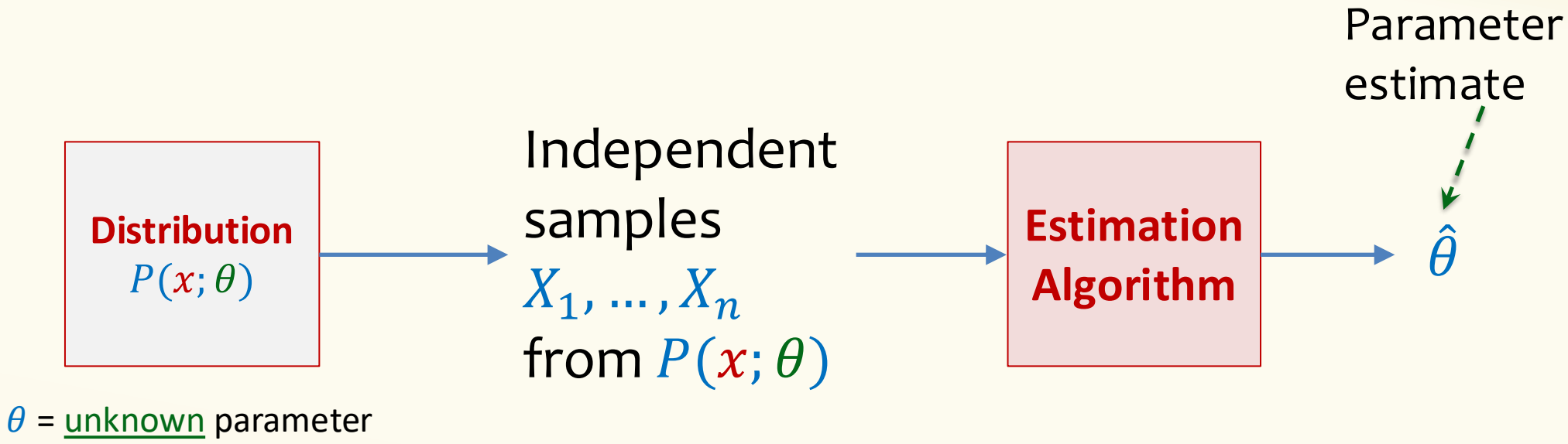
Recap More generally …

In estimation we often ….

- Assume: we know the type of the random variable that we are observing independent samples from
  - We just don't know the parameters, e.g.
    - the bias p of a random coin $\text{Bernoulli}(p)$
    - The arrival rate λ for the $\text{Poisson}(\lambda)$ or $\text{Exponential}(\lambda)$
    - The mean μ and variance σ of a normal $\mathcal{N}(\mu, \sigma)$

- Goal: find the "best" parameters to fit the data

Statistics: Parameter Estimation – Workflow



θ = unknown parameter

Example: coin flip distribution with unknown θ = probability of heads

Observation: $HTTHHHTHTHTTTTHTHTTTTTHT$

Goal: Estimate θ

Example

Suppose we have a mystery coin with some probability p of coming up heads. We flip the coin 8 times, independent of other flips, and see the following sequence of flips

$$TTHTHTTH$$

Given this data, what would you estimate p is?

How can you argue "objectively" that this your estimate is the best estimate?

Agenda

- Recap + more on Law of Total Expectation
- Introduction to statistical estimation
- Maximum Likelihood Estimation

Likelihood

Say we see outcome $HHTHH$.

You tell me your best guess about the value of the unknown parameter θ (a.k.a. p) is 4/5. Is there some way that you can argue "objectively" that this is the best estimate?

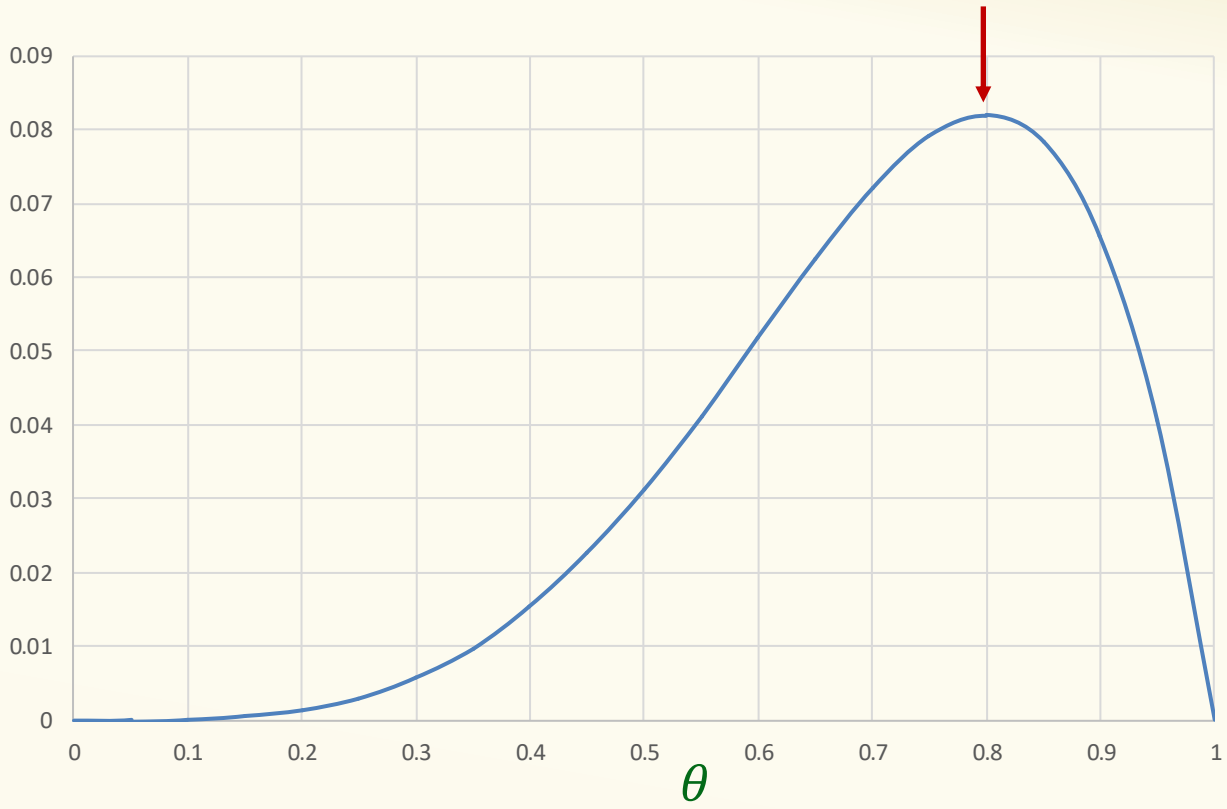
Likelihood

Say we see outcome $HHTHH$.

$$\mathcal{L}(HHTHH\,;\,\theta) = \theta^4(1-\theta)$$

Probability of observing the outcome $HHTHH$ if $\theta =$ prob. of heads.

For a fixed outcome $HHTHH$, this is a function of θ.



Likelihood of Different Observations

(Discrete case)

Definition. The **likelihood** of independent observations $x_1, \ldots, x_n$ is

$$\mathcal{L}(x_1, x_2, \ldots, x_n ; \theta) = \prod_{i=1}^{n} P(x_i; \theta)$$

Example:
Say we see outcome $HHTHH$.

$$\mathcal{L}(HHTHH ; \theta) = P(H; \theta) \cdot P(H; \theta) \cdot P(T; \theta) \cdot P(H; \theta) \cdot P(H; \theta) = \theta^4(1 - \theta)$$

Likelihood vs. Probability

- Fixed θ: **probability** $\prod_{i=1}^{n} P(x_i; \theta)$ that dataset $x_1, \ldots, x_n$ is sampled by distribution with parameter θ
  - A function of $x_1, \ldots, x_n$
- Fixed $x_1, \ldots, x_n$: **likelihood** $\mathcal{L}(x_1, x_2, \ldots, x_n ; \theta)$ that parameter θ explains dataset $x_1, \ldots, x_n$.
  - A function of θ

These notions are the same number if we fix <u>*both*</u> $x_1, \ldots, x_n$ and θ, but different role/interpretation

Likelihood of Different Observations

(Discrete case)

Definition. The **likelihood** of independent observations $x_1, \ldots, x_n$ is

$$\mathcal{L}(x_1, x_2, \ldots, x_n ; \theta) = \prod_{i=1}^{n} P(x_i; \theta)$$

Maximum Likelihood Estimation (MLE). Given data $x_1, \ldots, x_n$, find $\hat{\theta}$ such that $\mathcal{L}(x_1, x_2, \ldots, x_n ; \hat{\theta})$ is maximized!

$$\hat{\theta} = \operatorname*{argmax}_{\theta} \mathcal{L}(x_1, x_2, \ldots, x_n ; \theta)$$

Example – Coin Flips

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

– i.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$\mathcal{L}(x_1, \ldots, x_n ; \theta) = \theta^{n_H} (1 - \theta)^{n_T}$$

Goal: find θ that maximizes $\mathcal{L}(x_1, \ldots, x_n ; \theta)$

Example – Coin Flips

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

– i.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$\mathcal{L}(x_1, \ldots, x_n ; \theta) = \theta^{n_H} (1 - \theta)^{n_T}$$

$$\frac{\partial}{\partial \theta} \mathcal{L}(x_1, \ldots, x_n ; \theta) = ???$$

While it is possible to compute this derivative, it's not always nice since we are working with products.

Log-Likelihood

We can save some work if we use the **log-likelihood** instead of the likelihood directly.

Definition. The **log-likelihood** of independent observations $x_1, \ldots, x_n$ is

$$\ln \mathcal{L}(x_1, \ldots, x_n ; \theta) = \ln \prod_{i=1}^{n} P(x_i; \theta) = \sum_{i=1}^{n} \ln P(x_i; \theta)$$

Useful log properties

$$\ln(ab) = \ln(a) + \ln(b)$$
$$\ln(a/b) = \ln(a) - \ln(b)$$
$$\ln\left(a^b\right) = b \cdot \ln(a)$$

Example – Coin Flips

$$\ln(ab) = \ln(a) + \ln(b)$$
$$\ln(a/b) = \ln(a) - \ln(b)$$
$$\ln(a^b) = b \cdot \ln(a)$$

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

– i.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$\mathcal{L}(x_1, \ldots, x_n ; \theta) = \theta^{n_H} (1 - \theta)^{n_T}$$

Example – Coin Flips

Observe: Coin-flip outcomes $x_1, \ldots, x_n$, with n_H heads, n_T tails

– i.e., $n_H + n_T = n$

Goal: estimate θ = prob. heads.

$$\mathcal{L}(x_1, \ldots, x_n ; \theta) = \theta^{n_H} (1-\theta)^{n_T}$$

$$\ln \mathcal{L}(x_1, \ldots, x_n ; \theta) = n_H \ln \theta + n_T \ln(1-\theta)$$

$$\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n ; \theta) = n_H \cdot \frac{1}{\theta} - n_T \cdot \frac{1}{1-\theta}$$

Want value $\hat{\theta}$ of θ s.t. $\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n ; \theta) = 0$

So we need $n_H \cdot \frac{1}{\hat{\theta}} - n_T \cdot \frac{1}{1-\hat{\theta}} = 0$

Solving gives

$$\hat{\theta} = \frac{n_H}{n}$$

General Recipe

1. **Input** Given n i.i.d. samples $x_1, \ldots, x_n$ from parametric model with parameter θ.
2. **Likelihood** Define your likelihood $\mathcal{L}(x_1, \ldots, x_n ; \theta)$.
   - For discrete $\mathcal{L}(x_1, \ldots, x_n ; \theta) = \prod_{i=1}^{n} P(x_i ; \theta)$
3. **Log** Compute $\ln \mathcal{L}(x_1, \ldots, x_n ; \theta)$
4. **Differentiate** Compute $\frac{\partial}{\partial \theta} \ln \mathcal{L}(x_1, \ldots, x_n ; \theta)$
5. **Solve for** $\hat{\theta}$ by setting derivative to 0 and solving for max.

Generally, you need to do a second derivative test to verify it is a maximum, but we won't ask you to do that in CSE 312.

Example – Geometric distribution

$$\ln(ab) = \ln(a) + \ln(b)$$
$$\ln(a/b) = \ln(a) - \ln(b)$$
$$\ln(a^b) = b \cdot \ln(a)$$

You see data $x_1, \ldots, x_n$ from a geometric distribution with unknown parameter θ

What is the maximum likelihood estimate $\hat{\theta}$?

Definition of Estimator

θ: parameter we're trying to estimate (e.g. success probability p of a Bernoulli r.v.)

These are i. i. d. instances of X

This is a r.v. because it's a function of r.v.s

$X_1, X_2, \ldots, X_n$: sampled data

Sometimes just write $\hat{\theta}$

$\hat{\theta}(X_1, X_2, \ldots, X_n)$: estimator of the unknown θ

This is a constant

$\hat{\theta}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$: estimation of θ based on specific instantiation of the data