國立清華大學
NATIONAL TSING HUA UNIVERSITY

EE 306001
Probability

Lecture 24: intro to stats

李祈均

Introduction to statistics

We will cover the following chapters:

Chapt. 8.1 – 8.3

Chapt. 9.1

Statistics

Reality
(e.g., customer arrivals)

Models
(e.g., Poisson)

Data

Note:

- In a sense, there is no ‘new’ probability theory that will be covered in the next couple of lectures
- Statistics (inference problems) can be imagined as exercises using probability theory

However:

- Probability is built upon axioms (rules), given a probability problem, there is a correct (unique) answer
- Statistics does not work that way
  - You are only given data, with only data, say you want to estimate the motion of the planet…

Extremely common:

- Misuse of statistics
- Assumption checked?

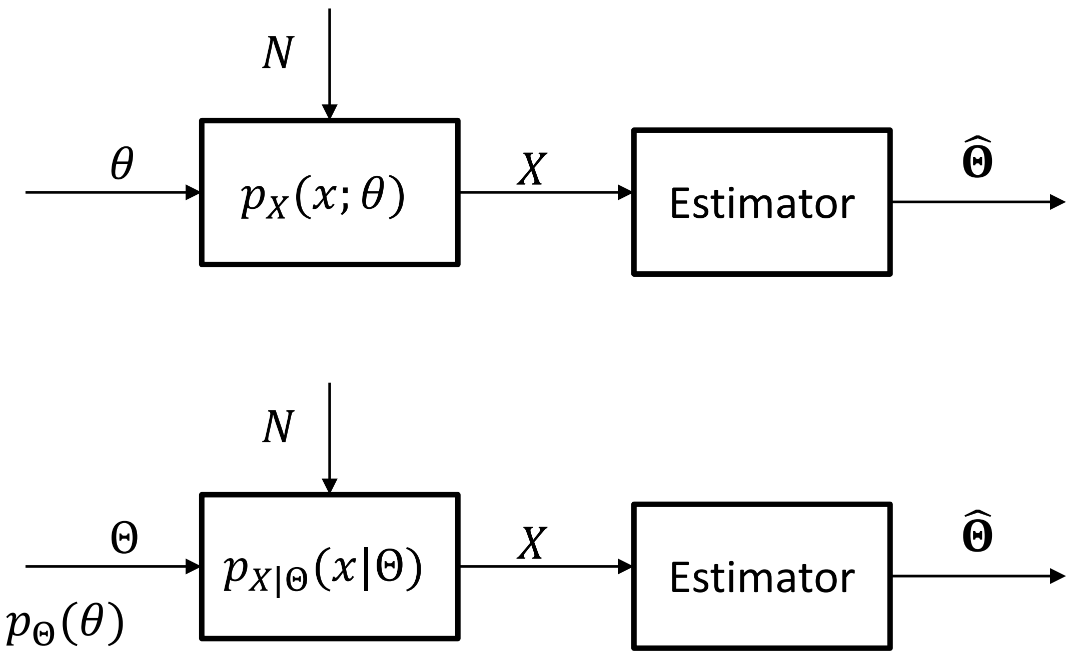
Two different types of statistical estimation problems

- Hypothesis testing (discrete)
  - Unknown takes one of the few possible (discrete)
  - Aim at small probability of incorrect decision
  - Recall the radar – airplane detection problem

- Estimation problem (continuous)
  - Aim at making small probability of error
  - Recall the polling type of problem

Bayesian vs. Classical

- Fundamental philosophical differences
  - Imaging a case of estimating the mass of an electron



Classical statistics
θ Is a parameter (not a r.v.)

Bayesian
Use priors of Θ

Note:

$\widehat{\Theta}$ is a random variable, data is random

estimator is a function on the data

Classical: treat mass as a number

Bayesian: treat mass as though you have certain ‘prior’ belief

These two class of thoughts, debate for 100 years, recently, Bayesian version is a little more prevalent

We know Bayes rule already, and that’s essentially what is involved in inference problem in Bayesian case, we will start with it!

Bayesian inference: Use Bayes rule find posterior pdf as a way to estimate the unknown rv

- Hypothesis testing
  - Discrete data

$$p_{\Theta|X}(\theta|x) = \frac{p_{\Theta}(\theta)p_{X|\Theta}(x|\theta)}{p_X(x)}$$

  - Continuous data

$$p_{\Theta|X}(\theta|x) = \frac{p_{\Theta}(\theta)f_{X|\Theta}(x|\theta)}{p_X(x)}$$

- These are not new, you have learnt much of this from previous chapters

- Estimation: continuous data

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x|\theta)}{f_X(x)}$$

  - $f_{\Theta|X}(\theta|x)$: posterior density -> the pdf of θ after you know something about the measurement
  - $f_{\Theta}(\theta)$: prior density

– Example: estimate the trajectory of a plane

Model (parabola)

- $Z_t = \Theta_0 + t\Theta_1 + t^2\Theta_2$

Data (position measurement)

- $X_t = Z_t + W_t, \quad t = 1,2,3 \ldots, n$

Bayes rule gives:

$$f_{\Theta_0\Theta_1\Theta_2|X_1,\ldots,X_n}(\theta_0\theta_1\theta_2|x_1 x_2, \ldots, x_n)$$

Functional form, and with data, you get the most likely ‘triplets of theta’

Estimation:

- Discrete data

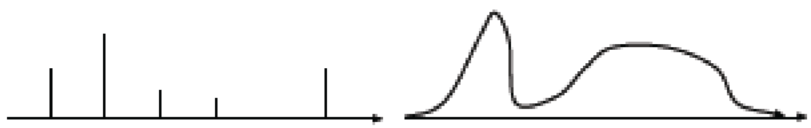
$$f_{\Theta|X}(\theta|x) = \frac{f_\Theta(\theta)p_{X|\Theta}(x|\theta)}{p_X(x)}$$

- Example:
  - Coin with unknown parameter θ (also the polling example)

  - Bayesian statistician?
    - Want to find $f_{\Theta|X}(\theta|x)$ by assuming a prior distribution on theta
    - That prior distribution depends on your initial belief

Output of the Bayesian inference: posterior pdf

- Posterior distribution (the distribution of unknown after your measurement)
  - Pmf $p_{\Theta|X}(.|x)$ or pdf $p_{\Theta|X}(.|x)$



  - If interested in a single answer (you need a single-point number):
    - Imagine polling (fraction =? -> what is that number not a rv not a pdf)
  - Use maximum a posteriori probability (MAP) single-point estimate

$$p_{\Theta|X}(\theta^*|x) = \max_{\theta} p_{\Theta|X}(\theta|x)$$

    - Minimizes probability of error; used in discrete hypothesis testing

- In the continuous estimation case:
  - How do we report
  - We could report

$$f_{\Theta|X}(\theta^*|x) = \max_{\theta} f_{\Theta|X}(\theta|x)$$

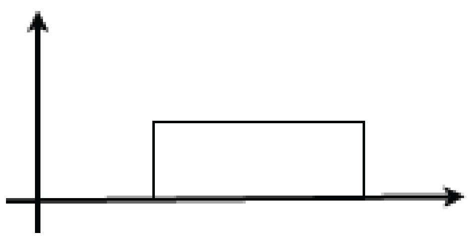
  - The point at the maximum point of the density function
    - In this context, you can't really say this is 'most likely' value of theta – only the probability over a neighborhood of that point is highest
  - We could also report conditional expectation

$$E[\Theta|X = x] = \int \theta\, f_{\Theta|X}(\theta|x) d\theta$$

    - Average of the density function (center of gravity)
  - Both are fair, which is better? Depends!
  - However, single point estimate is tricky and can be misleading

Example estimation 1: Least mean square estimation

- Imagine a case: estimation in the absence of information
  - Single point estimate using prior



- You only have a prior belief on Θ, as uniformly distributed over a range (say 4 – 10)
- You want to have a **point** estimate (single answer) for Θ, how?

Find estimate c, to

$$\text{minimize } E[(\Theta - c)^2]$$

- Essentially, trying to find a c that has minimum error (as measured by expected value of the square difference)
- This is called **least mean square error estimation (LMS)**

$$E[(\Theta - c)^2] = E[\Theta^2] - 2cE[\Theta] + c^2$$

Differentiate with respect to c and set it equal to 0 and solve

$$c = E[\Theta]$$

So in this case, $c = 7$

In this case, how good is your estimate

– Basically: how much **expected error** there is?

$$E[(\Theta - c)^2]$$

What is c now? $E[\Theta]$

$$E[(\Theta - c)^2] = E[(\Theta - E[\Theta])^2] = var(\Theta)$$

Optimal estimate: $\boldsymbol{E[\Theta]}$

Error associated with this estimate: $\boldsymbol{var(\Theta)}$

- Okay now, what would happen if we have data…

LMS estimation of Θ based on X

- Now we have two random variables, Θ and X
- We observe that $X = x$
  - Essentially, we are just now in a new universe, a conditional universe where $X = x$

So again, we want to have an estimate c such that:

$$E[(\Theta - c)^2 | X = x]$$

Is minimized by $c = E[\Theta | X = x]$

We can imagine, that the estimator $E[\Theta|X = x]$ is a function of data X

We know this from the above:

$$E[(\Theta - E[\Theta|X = x])^2|X = x] \leq E\left[\left(\Theta - g(x)\right)^2|X = x\right]$$

This implies

$$E[(\Theta - E[\Theta|X])^2|X] \leq E\left[\left(\Theta - g(X)\right)^2|X\right]$$

Now use law of iterated expectation, take expectation on both sides:

$$E[(\Theta - E[\Theta|X])^2] \leq E\left[\left(\Theta - g(X)\right)^2\right]$$

This means: mean square error is smallest if $c = E[\Theta|X]$ (a function) than any other function $g(.)$ (estimator) on data X

So what is the <u>function</u> that provides the best estimate in least mean square error ? $E[\Theta|X]$ (r.v. of conditional expectation)

Some properties of LMS estimators: property of the estimation error

Use the following notation:

$$\hat{\Theta} = E[\Theta|X], \qquad \tilde{\Theta} = \hat{\Theta} - \Theta$$

- The estimation error $\tilde{\Theta}$ is unbiased, i.e.,

$$E[\tilde{\Theta}] = 0, E[\tilde{\Theta}|X = x] = 0 \text{ for all x}$$

$$E[\tilde{\Theta}|X] = E[\hat{\Theta} - \Theta|X] = E[\hat{\Theta}|X] - E[\Theta|X]$$

$$E[\Theta|X] = \hat{\Theta}, E[\hat{\Theta}|X] = \hat{\Theta}$$

When you know X, you know $\hat{\Theta}$

$$E[\tilde{\Theta}|X] = 0, \text{therefore} E[\tilde{\Theta}|X = x] = 0$$

- Using law of iterated expectation

$$E[\tilde{\Theta}] = E\left[E[\tilde{\Theta}|X]\right] = 0$$

It means that there is no bias in this estimator,

Overall, the expected bias is zero, and there is no clear upward or downward bias

Example

Romeo and Juliet start dating, but Juliet will be late on any date by a random amount X, uniformly distributed over the interval $[0, \theta]$. The parameter, θ, is unknown and is modeled as the value of a random variable, Θ, which is uniformly distributed between zero and one hour.

Now, assume Juliet is late by x on their first date, how should Romeo use this information to update the distribution of Θ

- First note the prior distribution pdf is

$$f_\Theta(\theta) = \begin{cases} 1, \text{if } 0 \le \theta \le 1 \\ 0, \text{otherwise} \end{cases}$$

The conditional pdf of the observation (data given parameter):

$$f_{X|\Theta}(x|\theta) = \begin{cases} \frac{1}{\theta}, \text{if } 0 \le x \le \theta \\ 0, \text{otherwise} \end{cases}$$

Now, we can directly use Bayes rule ($f_{\Theta|X}(\theta|x)$), and note that $f_\Theta(\theta) f_{X|\Theta}(x|\theta)$ is nonzero only if $0 \le x \le \theta \le 1$

- The new posterior probability distribution function is:

$$f_{\Theta|X}(\theta|x) = \frac{f_{\Theta}(\theta) f_{X|\Theta}(x|\theta)}{\int_0^1 f_{\Theta}(\theta')\, f_{X|\Theta}(x|\theta') d\theta}$$

$$= \frac{1/\theta}{\int_x^1 1/\theta' d\theta'} = \frac{1}{\theta|\log x|}, \text{if x} \le \theta \le 1$$

$$f_{\Theta|X}(\theta|x) = \begin{cases} \dfrac{1}{\theta|\log x|}, \text{if x} \le \theta \le 1 \\ 0, \text{otherwise} \end{cases}$$

Okay so with this, can we come up with a single point estimate?

$$f_{\Theta|X}(\theta|x) = \begin{cases} \dfrac{1}{\theta|\log x|}, \text{if x} \leq \theta \leq 1 \\ 0, \text{otherwise} \end{cases}$$

First, lets think about this function,

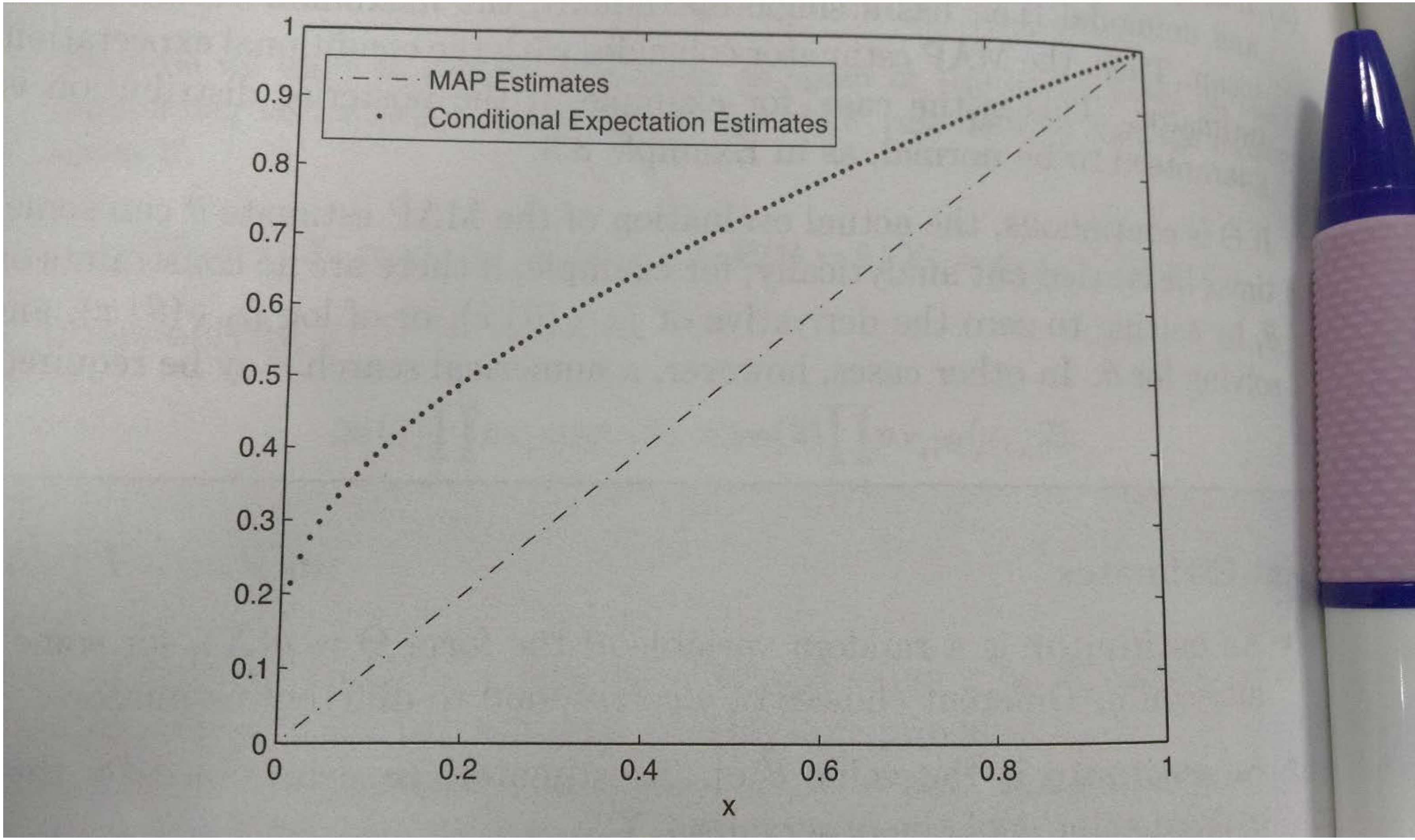
- Any given x, $f_{\Theta|X}(\theta|x)$ is decreasing with θ over the range $[x, 1]$
  - So what's best estimate if we are going for MAP (maximum a-posteriori estimation)?
  - $\theta = x$
  - Note this is an optimistic estimate
    - If Juliet is late by a small amount on the first date (x is small), then the estimated θ is also small! Can it be real?

How about if we have another point estimate?

- Instead of MAP, let's try conditional expectation

$$E[\Theta|X = x] = \int_x^1 \theta \frac{1}{\theta|\log x|} d\theta$$

$$= \frac{1 - x}{|\log x|}$$



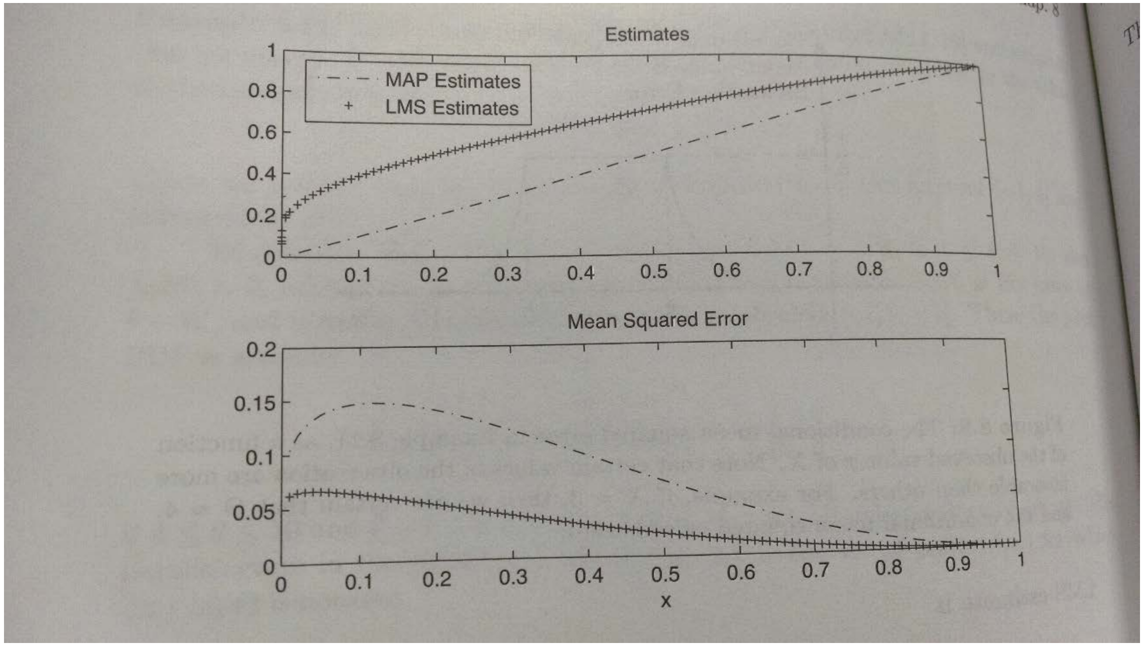
You can see pretty clearly that conditional expectation estimate (which is also the LMS estimate) is much more conservative estimate

Error for MAP estimate

$$E\left[\left(\hat{\theta} - \Theta\right)^2 | X = x\right] = x^2 + \frac{3x^2 - 4x + 1}{2|\log x|}$$

Error for LMS estimate

$$E\left[\left(\hat{\theta} - \Theta\right)^2 | X = x\right] = \frac{1 - x^2}{2|\log x|} - \left(\frac{1 - x}{\log x}\right)^2$$



Example – spam filtering

An email message may be ‘spam’ or ‘legitimate’. We introduce a parameter, Θ, taking value 1 or 2, corresponding to spam and legitimate, respectively, with given probability $p_\Theta(1)$, $p_\Theta(2)$. (Hypothesis testing – discrete)

Let $\{w_1, \dots, w_n\}$ be a collection of special words (or combination of words) whose appearance suggests a spam message. For each i, let X_i be the Bernoulli random variable that models that appearance of w_i in the message ($X_i = 1$ if w_i appears and $X_i = 0$ if w_i does not)

We assume that the conditional probabilities $p_{X_i|\Theta}(x_i|1)$ and $p_{X_i|\Theta}(x_i|2)$, $x_i = 0, 1$ are known. For simplicity, we further assume that conditioned on Θ, the random variable $X_1, X_2, \dots, X_n$ are independent.

To classify spam or non-spam, first let's compute a-posteriori probability

$$P(\Theta = m|X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

$$= \frac{p_\Theta(m) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i|m)}{\sum_{j=1}^{2} p_\Theta(j) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i|j)}, \qquad m = 1,2$$

What is a MAP rule in deciding whether an email is spam or not?

- Minimizes the probability of making error:

So the message is a spam if:

$$P(\Theta = 1 | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) > P(\Theta = 2 | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n)$$

That means:

$$p_\Theta(1) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i|1) > p_\Theta(2) \prod_{i=1}^{n} p_{X_i|\Theta}(x_i|2)$$

Chapter 9.1

Classical statistics

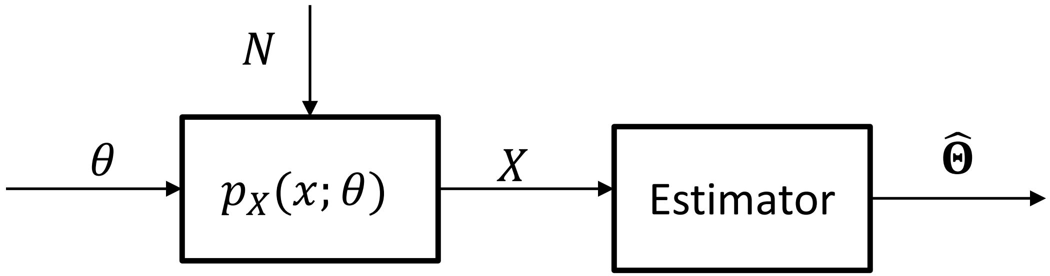
Outline:

Classical statistics

Maximum likelihood estimation

Estimating by sample mean

Confidence interval*



Classical statistics
θ Is a parameter (not a r.v.)

θ: the unknown, nothing random about it, it's just a number

$p_X(x;\theta)$: this distribution depends on θ, however, it is **NOT** conditional pdf (conditional pdf is for r.v.s)!

You can imagine this could just be a parameter to describe the distribution X (somehow depends on θ) – maybe a normal distribution with mean θ

Data X could be a vector $[X_1, X_2, \dots, X_n]$, and θ could be a vector of parameters too!

- You can imagine:

Mathematically: many different probabilistic models, one for each possible value of θ

Problem types:

- Hypothesis testing:
  - H_0: $\theta = 1/2$ versus H_1: $\theta = 3/4$
- Composite hypothesis:
  - H_0: $\theta = 1/2$ versus H_1: $\theta \neq 1/2$
  - A little more complicated (this implies multiple models)
- Today: estimation problem (unknown is continuous)
  - Estimator $\widehat{\Theta}$ is random, though θ is not
  - Desire: design of $\widehat{\Theta}$ to keep the estimation error $\widehat{\Theta} - \theta$ small

Maximum likelihood estimation (ML)

- One way: pick θ, that means pick a specific probability model that the data we observe, $X's$, most likely have occurred
- Assume your data follow a distribution, find the parameter that is most likely result in the X you collect

Mathematically:

- Model with unknown parameter(s), $X \sim p_X(x;\theta)$
- ML: pick θ that *"makes the data most likely" (makes the distribution that generates the data has highest probability)*

$$\hat{\theta}_{ML} = \arg\max_{\theta} p_X(x;\theta)$$

Simple example

- Data sample: $\mathrm{X} = X_1, \ldots, X_n$: i.i.d. exponential($\theta$)

Try to find a good estimate of θ using the ML approach

$$\hat{\theta}_{ML} = \arg\max_{\theta} p_X(x;\theta)$$

$$p_X(x;\theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

Take log

$$\ln p_X(x;\theta) = n \ln \theta - \theta \sum_{i=1}^{n} x_i$$

$$\hat{\theta}_{ML} = \arg\max_{\theta} p_X(x;\theta)$$

$$\hat{\theta}_{ML} = \arg\max_{\theta} \ln p_X(x;\theta) = \arg\max_{\theta} \left(n \ln\theta - \theta \sum_{i=1}^{n} x_i \right)$$

How to solve?

Take derivative with respect to θ and set it equals to 0 and solve:

$$\hat{\theta}_{ML} = \frac{n}{x_1 + \cdots + x_n}$$

What is the expectation of an exponential distribution: $\theta e^{-\theta x_i}$?

It’s $1/\theta$

Let’s abstractly this about this:

In fact, we have just designed an estimator (function on data) of the following form:

$$\widehat{\Theta}_n = \frac{n}{X_1 + \cdots + X_n}$$

Imagine this as an experiment, once you do the experiment, that each X_i will output a number, and you have the ML-estimate

- Comparison to Bayesian approach

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p_{\Theta|\mathrm{X}}(\theta|x)$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \frac{p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)}{p_X(x)}$$

$p_X(x)$ is just a number, can be ignored in the maximization procedure

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p_{X|\Theta}(x|\theta)p_{\Theta}(\theta)$$

If $p_{\Theta}(\theta)$ is uniform (constant)

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p_{X|\Theta}(x|\theta)$$

Is just like the ML estimation (though the derivation logic is a little different if you think about it)

ML estimator for exponential distribution parameter

- Data sample: $\mathrm{X} = X_1, \ldots, X_n$: i.i.d. exponential(θ)

Try to find a good estimate of θ using the ML approach

$$\hat{\theta}_{ML} = \arg\max_{\theta} p_X(x; \theta)$$

$$p_X(x; \theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

Take log

$$\ln p_X(x; \theta) = n \ln \theta - \theta \sum_{i=1}^{n} x_i$$

$$\hat{\theta}_{ML} = \arg\max_{\theta} p_X(x;\theta)$$

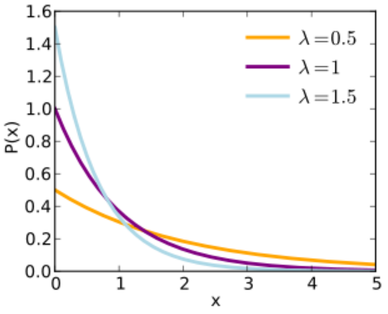
$$\hat{\theta}_{ML} = \arg\max_{\theta} \ln p_X(x;\theta) = \arg\max_{\theta} \left(n \ln \theta - \theta \sum_{i=1}^{n} x_i \right)$$

How to solve?

Take derivative with respect to θ and set it equals to 0 and solve:

$$\hat{\theta}_{ML} = \frac{n}{x_1 + \cdots + x_n}$$

Some desired probability of an estimator



- This estimator, $\widehat{\Theta}_n$, is random
- **Unbiased**: $E\left[\widehat{\Theta}_n\right] = \theta$
  - $\widehat{\Theta}_n$ is a function of data X
  - X is affected by true parameter θ (each θ corresponds to a different model)
  - Is it always true? Not necessary

  Exponential example that we just demonstrated (take $n = 1$)

$$E[1/X_1] = \infty \neq \theta$$

  This ML estimate is biased estimator (biased upward)

  In general ML estimate is just like this, under some condition, it will turn out to be unbiased

- **Consistent**: $\widehat{\Theta}_n \to \theta$ (convergence in probability)
  - This is good property especially if you have large amount of data
  - ML estimate tend to have this properties (given independent data)!

Exponential example:

$$\frac{(X_1 + \cdots + X_n)}{n} \to E[X] = 1/\theta$$

Knowing, we can look at our estimator:

$$\widehat{\Theta}_n = \frac{n}{X_1 + \cdots + X_n}$$

$$\widehat{\Theta}_n = \frac{n}{X_1 + \cdots + X_n} \to \frac{1}{E[X]} = \theta$$

Weak law of large number, this is true no matter what the true theta is

- $\widehat{\Theta}_n$ is a function of data X
- X is affected by true parameter θ (each θ corresponds to a different model)

One more desired property that combines two criterion for an estimator: **small mean square error (MSE)**

$$E\left[\left(\widehat{\Theta}-\theta\right)^2\right] = var\left(\widehat{\Theta}-\theta\right) + \left(E\left[\widehat{\Theta}-\theta\right]\right)^2$$
$$= var\left(\widehat{\Theta}\right) + (\mathrm{bias})^2$$

- Ideally, we want $\widehat{\Theta}$ to be very close θ, so we like the biased term to be zero, and at the same time, the fluctuation (variance of our estimator) is small too!

$$E\left[\left(\widehat{\Theta}-\theta\right)^2\right] = var\left(\widehat{\Theta}-\theta\right) + \left(E\left[\widehat{\Theta}-\theta\right]\right)^2$$
$$= var\left(\widehat{\Theta}\right) + (\mathrm{bias})^2$$

- Let's do a silly example

Assume we have distribution that is normal with unknown mean θ and variance 1

Let's design a simple estimator: just keep saying that mean equals to 100 no matter what

- This estimator has 0 variance, but huge bias term!

- Moral of the story;
  - You can make variance extremely small but pay the price in the bias term
  - There is certain tradeoff between the two
  - We won’t cover this further in this class

Revisit our estimation of mean

- $X_1, \ldots, X_n$: iid mean θ, variance σ^2

$$X_i = \theta + W_i$$

W_i: iid, mean 0, variance σ^2

Design an estimator to estimate mean θ using sample mean:

$$\widehat{\Theta}_n = \text{sample mean} = M_n = \frac{X_1 + \cdots + X_n}{n}$$

Revisit this sample mean estimator

- **Unbiased:**

$$\widehat{\Theta}_n = M_n = \frac{X_1 + \cdots + X_n}{n}$$

$$E[M_n] = \theta$$

- **Consistent:**

By weak law of large number:

Sample mean converge to true mean in probability

$$\widehat{\Theta}_n \to \theta$$

- **MSE**

$$var(\widehat{\Theta}) + (\mathrm{bias})^2 = \frac{\sigma^2}{n} + 0$$

Sample mean is quite good!