Estimation

Morning Objectives

This morning we'll talk about estimating statistical distributions from observed data

- Review what the expected value and variance of a random variable are
- Use Method of Moments (MOM), Maximum Likelihood Estimation (MLE), and Maximum A Posteriori (MAP) to estimate a parametric distribution from observed data
- Understand how Kernel Density Estimation (KDE) estimates a non-parametric distribution from observed data

Why Estimate Distributions?

Why estimate distributions?

- Example 1
  - You have data on how many people order cakes every day at your bakery, and you want to estimate the probability of selling out
- Example 2
  - You have data on how often your car breaks down, and you want to know your chances of safely crossing the country in it
- Example 3
  - You have data on how many people visit your website each day, and you want to know the probability of your servers being overloaded

Econometrician’s Philosophy

If you lack the information to determine the value directly, estimate the value to the best of your ability using the information you do have

Figure 1:At least you tried

Review

Expected Value

If X is a discrete random variable with k possible outcomes and $P(X = x)$ the value of its probability mass function at x, then the expected value of X is

$$E[X] = \sum_{i=1}^{n} x_i P(X = x_i)$$

If X is a continuous random variable and $f(x)$ the value its probability density function at x, then the expected value of X is

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx$$

(Same idea, except replace the summation with an integral and probabilities with probability densities)

Useful Properties of $E[\cdot]$

- If a is a constant:
$$E[a] = a$$
- If X is a random variable and a is a constant:
$$E[aX] = aE(X]$$
- If X and Y are random variables and a and b are constants:
$$E[aX + bY] = aE[X] + bE[Y]$$

Moments of a Random Variable X

- n^{th} Raw Moment:

$$\mu_n = E[X^n]$$

- n^{th} Central Moment:

$$\mu_n' = E[(X - E[X])^n]$$

1^{st} Raw Moment: Mean

$$E[X] = \mu_1 = \mu$$

For a discrete random variable X with n possible outcomes:

$$\mu_1 = \mu = \sum_{i=1}^{n} x_i P(X = x_i)$$

For a continuous random variable X:

$$\mu_1 = \mu = \int_{-\infty}^{\infty} x f(x) dx$$

2^{nd} Central Moment: Variance

The *variance* of a random variable X is the expected value of the square difference from the mean:

$$E[(X - E[X])^2] = \mu_2' = Var(X) = \sigma^2$$

For a discrete random variable X with n possible outcomes:

$$\mu_2' = \sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i)$$

For a continuous random variable X:

$$\mu_2' = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Useful Properties of $Var(\cdot)$

- If a is a constant:

$$Var(a) = 0$$

- If X is a random variable and a is a constant:

$$Var(aX) = a^2 Var(X)$$

- If X and Y are random variables and a and b are constants:

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

μ_2 as a function of μ and σ^2?

μ_2 as a function of μ and σ^2?

Estimating Distributions

Parametric vs. Non-Parametric Methods

Parametric and non-parametric procedures are two broad classifications of statistical methods.

Parametric

- Make assumptions about the shape and parameters of the underlying population distribution the data was sampled from. E.g., $B(n, p)$, $N(\mu, \sigma)$, or $P(\lambda)$

Parametric vs. Non-Parametric Methods

Non-Parametric

- Do not rely on assumptions about the shape and parameters of the underlying population distribution the data was sampled from

Non-Parametric methods are more flexible but generally have less power than corresponding parametric methods. Interpretation with non-parametric methods can also be difficult, e.g., what does the wiggly curve mean?

Estimating Distributions

Parametric

- Method of Moments (MOM)
- Maximum Likelihood Estimation (MLE)
- Maximum a Posteriori (MAP)

Non-Parametric

- Kernel Density Estimation (KDE)

Method of Moments (MOM)

Derive equations related to raw sample and population moments:

$$E[X], E[X^2], E[X^3], \ldots$$

Method

1. Equate the first raw sample moment $M_1 = \frac{1}{N}\sum X_i$ to the first raw population moment $E[X] = \mu_1 = \mu$
2. Equate the second raw sample moment $M_2 = \frac{1}{N}\sum X_i^2$ to the second raw population moment $E[X^2] = \mu_2 = \mu^2 + \sigma^2$
3. Continue until you have as many equations as you have parameters
4. Solve for parameters

Method of Moments (MOM) - Example 1

Your website visitor log shows the following number of visits for each of the last seven days: $[6, 4, 7, 4, 9, 3, 5]$. What's the probability of zero visitors tomorrow?

Method of Moments (MOM) - Example 1

Method of Moments (MOM) - Example 1

Method of Moments (MOM) - Example 2

Suppose we flip a coin N times again, and get H heads. Use MOM to estimate p, the probability of flipping a head.

Method of Moments (MOM) - Example 2

Method of Moments (MOM) - Example 2

Method of Moments (MOM) - Example 3

Suppose we have data sampled from a symmetric uniform distribution with unknown bounds $X \sim U(-b, b)$. Estimate using MLE.

Method of Moments (MOM) - Example 3

Method of Moments (MOM) - Example 3

Maximum Likelihood Estimation (MLE)

Law of Likelihood:

- If $P(X|H_1) > P(X|H_2)$, then the evidence supports H_1 over H_2

Question:

- Which hypothesis does the evidence most strongly support?

Answer:

- The hypothesis H that maximizes $P(X|H)$
  - which is found via MLE...

Maximum Likelihood Estimation (MLE)

Set values of parameters to values that will maximize the likelihood function

- Assume $X_1, X_2, ..., X_n$ are *i.i.d.*, then the likelihood function is their joint density function:

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

Maximum Likelihood Estimation (MLE)

- Maximizing the likelihood function is the same as maximizing the log likelihood function which simplifies calculations

$$\ell(\theta|x_1, x_2, \ldots, x_n) = log\,\mathcal{L}(\theta|x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\arg\max}\; log\,\mathcal{L}(\theta|x_1, x_2, \ldots, x_n)$$

Maximum Likelihood Estimation (MLE) - Example 1

Suppose that $x_1, \ldots, x_n \sim N(\mu, \sigma)$. Estimate μ and σ using MLE.

Maximum Likelihood Estimation (MLE) - Example 1

Ignoring some constants, the likelihood function is

$$\begin{aligned}
\mathcal{L}(\mu, \sigma) &= \prod \frac{1}{\sigma} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2\right) \\
&= \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum (x_i - \mu)^2\right) \\
&= \frac{1}{\sigma^n} \exp\left(-\frac{nS^2}{2\sigma^2}\right) \exp\left(-\frac{n(\bar{x} - \mu)^2}{2\sigma^2}\right)
\end{aligned}$$

where $\bar{x}$ is the sample mean and $S^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$ is a biased estimator for the variance.

Maximum Likelihood Estimation (MLE) - Example 1

The log-likelihood is

$$\ell(\mu, \sigma) = -n \log \sigma - \frac{nS^2}{2\sigma^2} - \frac{n(\bar{x} - \mu)^2}{2\sigma^2}$$

Maximizing this gives estimates $\hat{\mu}_{MLE} = \bar{x}$ and $\hat{\sigma}_{MLE} = S$.

Maximum Likelihood Estimation (MLE) - Example 1

Maximum Likelihood Estimation (MLE) - Example 2

Suppose we flip a coin N times and get H heads. Using MLE, estimate how biased the coin is. I.e., estimate p, the probability of getting head.

Maximum Likelihood Estimation (MLE) - Example 2

Maximum Likelihood Estimation (MLE) - Example 2

Maximum A Posteriori (MAP)

- Generalization of MLE in which we assume a prior distribution g over Θ and go one step further to calculate the posterior distribution

$$f(\theta|x) = \frac{f(x|\theta)g(\theta)}{\int_{\Theta} f(x|\theta)g(\theta)d\theta} \propto f(x|\theta)g(\theta)$$

- To find the optimal θ we find

$$\hat{\theta}_{MAP} = \arg\max_{\theta \in \Theta} f(x|\theta)g(\theta)$$

- To get MLE, assume a uniform prior on θ so that the function g disappears from the arg max above

MLE vs. MAP

- MLE finds θ to maximize $f(x|\theta)$

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\arg\max}\, f(x|\theta)$$

- Whereas MAP finds θ to maximize $f(\theta|x) \propto f(x|\theta)g(\theta)$

$$\hat{\theta}_{MAP} = \underset{\theta \in \Theta}{\arg\max}\, f(x|\theta)g(\theta)$$

Kernel Density Estimation (KDE)

Kernel Density Estimation is used to estimate the pdf of a random variable and is essentially a data smoothing problem.

- KDE estimates a distribution empirically given data by summing kernels centered at each point. The density function of the kernel density estimate is:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

$K(\cdot)$ is a *kernel*: a non-negative function that integrates to one and has mean zero; a kernel is another word for a density function of a distribution with mean 0. The parameter h, a smoothing parameter, is called the *bandwidth*, and it's analogous to the width of bins in a histogram.

Kernel Density Estimation (KDE) - Example

Closely related to histograms, but can be made smooth by using a suitable kernel.

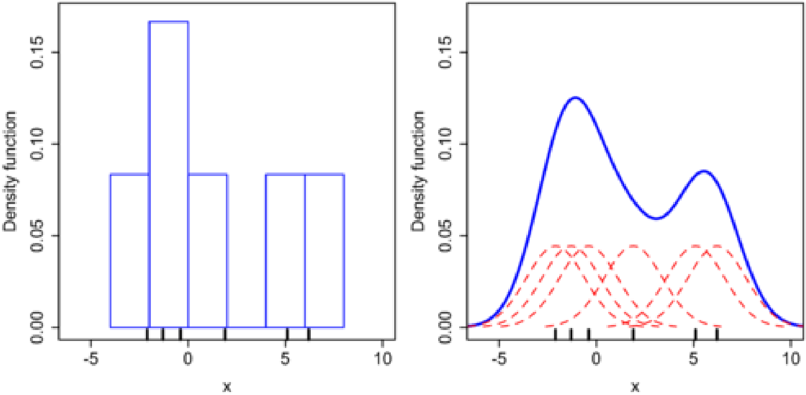

Figure 2:Histograms and KDEs

Histogram Troubles

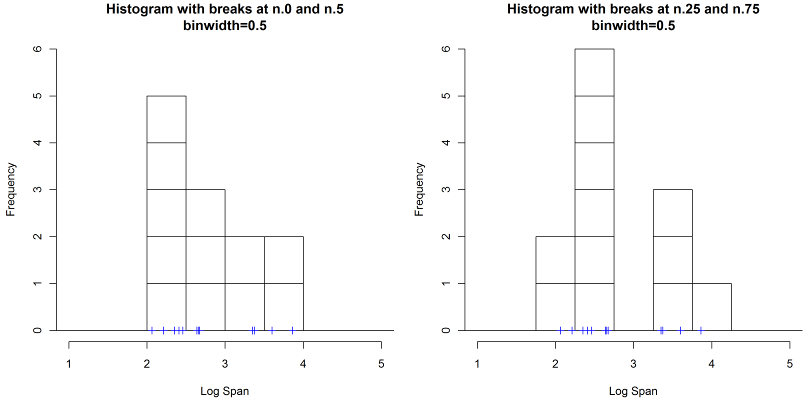

Figure 3:Histogram Troubles

Bandwidth Selection

- A free parameter which exhibits a strong influence on the resulting estimate
- The most common optimality criterion used to select the parameter is the mean integrated squared error

$$MISE(h) = E \int_{-\infty}^{\infty} (\hat{f}_h(x) - f(x))^2 dx$$

Bandwidth Selection

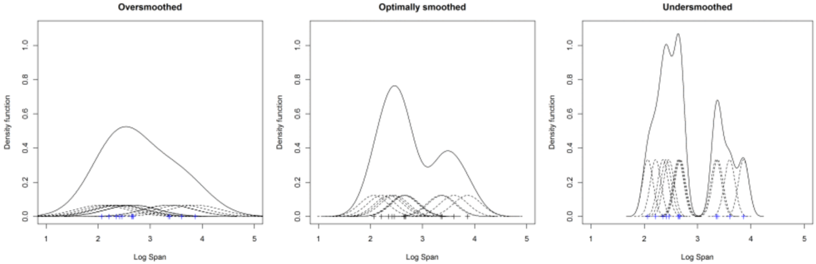

Figure 4:Am I too smooth? Not smooth enough?

Questions

- MOM vs. MLE
  - What do they solve for?
  - How does each approach the tackle the problem?
- How about MAP?
  - How does it relate to the MLE?