Estimating Distributions

Problem Motivation

Example 1: You have data on how many people order cake 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.

Solution:
Estimate a probability distribution based on data

Estimation Methods

In this lecture:

1. Method of Moments (MOM)
2. Maximum Likelihood Estimation (MLE)
3. Maximum a Posteriori (MAP)
4. Kernel Density Estimation (KDE)

Parametric Techniques

Methods of estimating *parameters* of specific probability distributions:

1. Method of Moments (MOM)
2. Maximum Likelihood Estimation (MLE)
3. Maximum a Posteriori (MAP)

Method of Moments

Summary:

1. Assume a distribution (e.g. Poisson, Normal, Binomial, etc.)
2. Compute the relevant sample moment (e.g. mean, variance)
3. Plug that sample moment into the PDF/CDF of the assumed distribution

Method of Moments

What's a *moment*?

Generally refers to the mean (1st moment) or variance (2nd moment) of a distribution.

Method of Moments

Example:
You flip a biased coin 100 times. 52 times it comes up heads. What's the MOM estimate of P[heads]?

1. Assume a Bernoulli distribution.
2. Compute the sample mean (1st moment): 52/100 = .52 = p

Method of Moments

Example 2:

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

1. Assume a Poisson distribution.
2. Compute sample mean: lambda = (4 + 5 + 6) / 3 = 5
3. P(zero users) = exp(-5)

Maximum Likelihood Estimation

Law of Likelihood:
If $P(X \mid H1) > P(X \mid H2)$, then evidence supports H1 over H2.

Question:
What hypothesis does the evidence most strongly support?

Answer:
The hypothesis H that maximizes $P(X \mid H)$, which is found via *maximum likelihood estimation*.

Maximum Likelihood Estimation

What's the likelihood (probability) of data given our model?

$$f(x_1, x_2, ..., x_n|\theta) = f(x_1|\theta) * f(x_2|\theta) * f(x_3|\theta) * ... * f(x_n|\theta)$$

MLE finds distribution parameters to maximize *likelihood function*.

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

$$\hat{\theta}_{mle} = \underset{\theta \in \Theta}{argmax}\ log\mathcal{L}(\theta|x_1, ..., x_n)$$

Maximum Likelihood Estimation

$$X_i \sim Binomial(N, p),\ \ i = 1, 2, \ldots, n$$

As with MOM, assume data comes from some distribution

$$\Rightarrow \quad f(x_i) = \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i}$$

$$\Rightarrow \quad \mathcal{L}(p|x) = \prod_{i=1}^{n} \binom{N}{x_i} p^{x_i} (1-p)^{N-x_i}$$

Define Likelihood

$$\Rightarrow \quad log\mathcal{L}(p|x) = \sum_{i=1}^{n} \log \binom{N}{x_i} + x_i \log p$$

$$+ (N - x_i) \log (1-p)$$

Log Likelihood

$$\Rightarrow \quad \frac{\partial\, log\mathcal{L}(p|x)}{\partial p} = \sum_{i=1}^{n} \left[\frac{x_i}{\widehat{p}} - \frac{N - x_i}{1 - \widehat{p}} \right] = 0$$

$$\Rightarrow \quad \hat{p} = \frac{\bar{x}}{N}$$

Estimate parameter using some calculus!

Maximum A Posteriori (MAP)

Recall Bayes Rule:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)}$$

MAP finds H to maximize P(H | X):

$$\operatorname*{argmax}_{H} P(X|H)P(H)$$

Note: if each hypothesis is equally likely, the MAP is same as MLE.

Maximum A Posteriori

Example

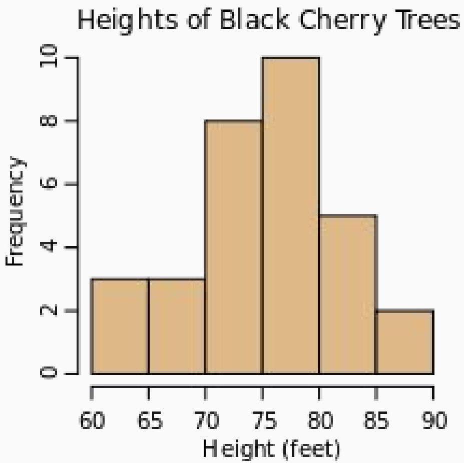
Non-parametric Techniques

Question: How can you model data that does not follow a known distribution?

Answer: Use non-parametric techniques.

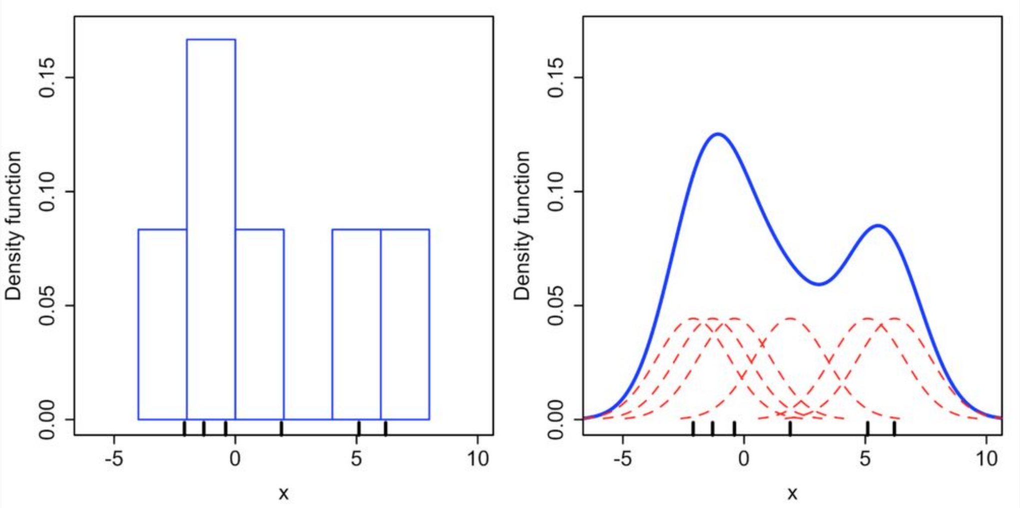
Histograms

- A histogram groups continuous data into discrete intervals and displays relative frequencies
- But, it's not a smooth distribution



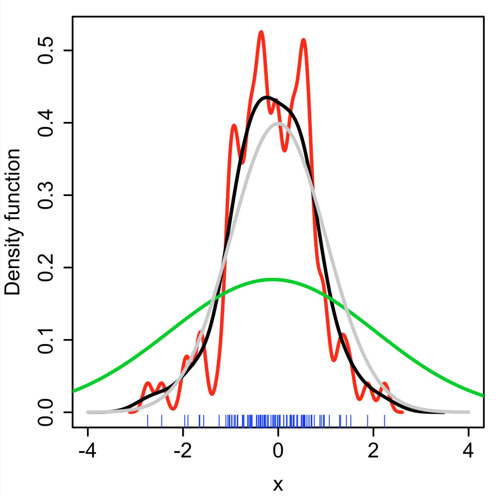
Kernel Density Estimation

- Non-parametric way to estimate PDF of a random variable
- KDE smoothes the histogram by summing Gaussians instead of rectangles



Kernel Density Estimation

- Kernel functions have a *bandwidth* parameter to control over-/under-fitting
- Each curve below shows an estimated PDF with different bandwidths



Kernel Density Estimation

- The result of KDE is a continuous probability density function

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i)$$

- The kernel function K is often a Gaussian, but it can be any positive function that integrates to 1 and has mean 0
- The Gaussian has a bandwidth parameter *h* corresponding to its variance

Other kernel functions: https://en.wikipedia.org/wiki/Kernel_(statistics)#In_non-parametric_statistics