Recap

• (Ch 6) Drawing general conclusions from a sample of the population
• (Ch 7) Assessing the significance of the evidence against a hypothesis

Today

• (Ch 9) Inferring a probability model from a dataset
  • Maximum likelihood estimation (MLE)
  • Confidence intervals for MLE estimates
  • Bayesian inference
Motivation: binomial example

• Suppose we have a coin with unknown probability of coming up heads

• We toss it $N$ times and observe $k$ heads

• We know that this data comes from a binomial distribution

• What is your best estimate of the probability of the coin coming up heads?
Motivation: geometric example

• Suppose we have a die with unknown probability of coming up six

• We roll it and it comes up six for the first time on the $k$th roll

• We know that this data point comes from a geometric distribution

• What is your best estimate of the probability of the die coming up six?
Motivation: Poisson example

• Suppose we have data on the number of babies born each hour in a large hospital

<table>
<thead>
<tr>
<th>hour</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td># of babies</td>
<td>$k_1$</td>
<td>$k_2$</td>
<td>...</td>
<td>$k_N$</td>
</tr>
</tbody>
</table>

• We can assume that this data comes from a Poisson distribution

• What is your best estimate of the intensity $\lambda$?
The parameter estimation problem

• Suppose we have a dataset $D = \{x\}$ that we know comes from a distribution in a certain family (e.g. binomial, geometric, Poisson, etc.)

• What is the best estimate of the parameters $\theta$ of the distribution?

• Examples
  • For binomial and geometric distributions, $\theta = p$ (probability of success)
  • For Poisson and exponential distributions, $\theta = \lambda$ (intensity)
  • For normal distributions, $\theta = (\mu, \sigma)$
Maximum likelihood estimation (MLE)

• We write the probability of seeing the data $D$ given parameters $\theta$

\[ L(\theta) = P(D|\theta) \]

• The **likelihood function** $L(\theta)$ is not a probability distribution

• The **maximum likelihood estimate** of $\theta$ is

\[ \hat{\theta} = \arg \max_{\theta} L(\theta) \]
Likelihood function: binomial example

• Suppose we have a coin with unknown probability of coming up heads

• We toss it \( N \) times and observe \( k \) heads

• We know that this data comes from a binomial distribution

• What is the likelihood function \( L(\theta) = P(D|\theta) \)?
MLE derivation: binomial example
Likelihood function: geometric example

• Suppose we have a die with unknown probability of coming up six

• We roll it and it comes up six for the first time on the $k$th roll

• We know that this data point comes from a geometric distribution

• What is the likelihood function $L(\theta) = P(D|\theta)$?
MLE derivation: geometric example
MLE with data from IID trials

• If the dataset $D = \{x\}$ comes from IID trials

$$L(\theta) = P(D|\theta) = \prod_{x_i \in D} P(x_i|\theta)$$

• This likelihood function is hard to differentiate (except for the binomial and geometric cases)

• Clever trick: take the (natural) log
Log-likelihood function

• Since log is a strictly increasing function

\[ \hat{\theta} = \arg \max_{\theta} L(\theta) = \arg \max_{\theta} \log L(\theta) \]

• So we can aim to maximize the log-likelihood function

\[ \log L(\theta) = \log P(D|\theta) = \log \prod_{x_i \in D} P(x_i|\theta) = \sum_{x_i \in D} \log P(x_i|\theta) \]

• The log-likelihood function is usually much easier to differentiate
Log-likelihood function: Poisson example

• Suppose we have data on the number of babies born each hour in a large hospital

<table>
<thead>
<tr>
<th>hour</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td># of babies</td>
<td>$k_1$</td>
<td>$k_2$</td>
<td>...</td>
<td>$k_N$</td>
</tr>
</tbody>
</table>

• We can assume that this data comes from a Poisson distribution with unknown intensity $\lambda$

• What is the log-likelihood function $\log L(\theta)$?
MLE derivation: Poisson example
Drawbacks of MLE

• Maximizing some likelihood or log-likelihood functions is intractable

• If there isn’t much data, the MLE estimate may be unreliable
  • If we observe 3 heads in 10 coin tosses, should we accept that $P(\text{heads}) = 0.3$?
  • If we observe 0 heads in 2 coin tosses, should we accept that $P(\text{heads}) = 0$?
Confidence intervals for MLE estimates

• An MLE parameter estimate \( \hat{\theta} \) depends on the dataset that was seen

• We can construct a confidence interval for \( \hat{\theta} \) using the **parametric bootstrap**
  
  • Use the distribution with parameter \( \hat{\theta} \) to generate a large number of datasets
  
  • From each “synthetic” dataset, re-estimate the parameter using MLE
  
  • Use the histogram of these re-estimates to construct a confidence interval
Bayesian inference

• In MLE, we maximized the likelihood function $L(\theta) = P(D|\theta)$

• In Bayesian inference, we will maximize the posterior, which is the probability of the parameters $\theta$ given the observed data $D$

$$P(\theta|D)$$

• Unlike $L(\theta)$, the posterior is a probability distribution

• The value of $\theta$ that maximizes $P(\theta|D)$ is called the maximum a posteriori (MAP) estimate $\hat{\theta}$
The prior

• From Bayes rule

\[ P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \propto P(D|\theta)P(\theta) \]

• We ignore the probability of the data \( P(D) \) because it is constant

• Bayesian inference allows us to incorporate prior beliefs about \( \theta \) in the prior \( P(\theta) \), which is useful
  • when we have some beliefs, such as a coin cannot have \( P(\text{heads}) = 0 \)
  • when there isn’t much data