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?

k/N

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 kth 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?

$1/k$

Motivation: Poisson example

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

hour	1	2	...	N
# of babies	k_1	k_2	...	k_N

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

- What is your best estimate of the intensity λ?

$$\lambda = \frac{\sum_{i=1}^{N} k_i}{N}$$

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 θ 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, $\cancel{\theta = (\mu, \sigma)}$ $\theta = \mu$ or $\theta = \sigma$ depending on context

Maximum likelihood estimation (MLE)

- We write the probability of seeing the data D given parameters θ

$$L(\theta) = P(D|\theta)$$

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

- The **maximum likelihood estimate** of θ is

$$\hat{\theta} = \arg\max_{\theta} L(\theta)$$

Likelihood function: binomial example

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

  $\theta = p$

- We toss it N times and observe k heads $\leftarrow D$

- We know that this data comes from a binomial distribution

- What is the likelihood function $L(\theta) = P(D|\theta)$?

$$L(\theta) = \binom{N}{k} \theta^{k} (1-\theta)^{N-k}$$

MLE derivation: binomial example

$$\frac{d}{d\theta} L(\theta) = \binom{N}{k}\left(\underline{k\theta^{k-1}(1-\theta)^{N-k}} - \underline{\theta^{k}(N-k)(1-\theta)^{N-k-1}} \right) = 0$$

$$k\cancel{\theta^{k-1}}\,\overset{(1-\theta)}{\cancel{(1-\theta)^{N-k}}} = (N-k)\,\overset{\theta}{\cancel{\theta^{k}}}\,\cancel{(1-\theta)^{N-k-1}}$$

$$k - \cancel{k\theta} = N\theta - \cancel{k\theta}$$

$$\hat{\theta} = \frac{k}{N} \text{ is the MLE for } p$$

Likelihood function: geometric example

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

  $\theta = p$

- We roll it and it comes up six for the first time on the kth roll

  $\underline{D}$

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

- What is the likelihood function $L(\theta) = P(D|\theta)$?

  $L(\theta) = (1-\theta)^{k-1}\theta$

MLE derivation: geometric example

$$\frac{d}{d\theta} L(\theta) = (1-\theta)^{k-1} - (k-1)(1-\theta)^{k-2}\theta = 0$$

$$\overset{(1-\theta)}{\cancel{(1-\theta)^{k-1}}} = (k-1)\cancel{(1-\theta)^{k-2}}\,\theta$$

$$1 - \cancel{\theta} = k\theta - \cancel{\theta}$$

$$\hat{\theta} = \frac{1}{k} \text{ is the MLE of } p$$

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

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

$\log(ab) = \log a + \log b$

- 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

hour	1	2	...	N
# of babies	k_1	k_2	...	k_N

- We can assume that this data comes from a Poisson distribution with unknown intensity λ

$$L(\theta) = \prod_{i=1}^{N} \frac{e^{-\theta}\theta^{k_i}}{k_i!}$$

- What is the log-likelihood function $\log L(\theta)$?

$$\log L(\theta) = \log\left(\prod_{i=1}^{N} \frac{e^{-\theta}\theta^{k_i}}{k_i!}\right) = \sum_{i=1}^{N} \log\left(\frac{e^{-\theta}\theta^{k_i}}{k_i!}\right) = \sum_{i=1}^{N}\left(-\theta + k_i \log\theta - \log k_i!\right)$$

MLE derivation: Poisson example

$$\frac{d}{d\theta}\log L(\theta) = \sum_{i=1}^{N}\left(-1 + \frac{k_i}{\theta} - 0\right) = 0$$

$$-N + \frac{\sum_{i=1}^{N} k_i}{\theta} = 0$$

$$N = \frac{\sum_{i=1}^{N} k_i}{\theta}$$

$$\hat{\theta} = \frac{\sum_{i=1}^{N} k_i}{N}$$ is the MLE of λ

MLE for normal distribution

- Suppose we model the dataset $D = \{x\}$ as normally distributed
- There are two parameters to estimate: μ and σ
  - If we fix σ and set $\theta = \mu$

$$\hat{\theta} = \frac{1}{N}\sum_{i=1}^{N} x_i$$ is the MLE of μ

  - If we fix μ and set $\theta = \sigma$

$$\hat{\theta} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$ is the MLE of σ

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 θ given the observed data D

$$P(\theta|D)$$

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

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

The prior

posterior likelihood 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 θ 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