Point Estimation

Ali Torkamani

Slides mostly from Vibhav Gogate from University of Texas at Dallas (Slight changes applied)

Coin Flipping

• A fair coin
  – When flipped: head with probability 0.5

• You flip a fair coin 100 times
  – how many heads would you expect?

• You find a coin and flip it 100 times
  – If 75 of the time the coin turns up tail, then how would you describe the coin?
  – More formally, what is the probability $p$ with which this specific coin turns up head?
Random variables

• **Random variable (RV)**
  – A variable whose value is subject to variations due to chance

• **In different problems RV’s take different values:**
  – **Coin flipping:** {head or tail} (Discrete)
  – **Tossing a die:** {1,2,...,6} (Discrete)
  – **Age of a UO students:** \( x \in \mathbb{N}, x \geq 17 \) (Discrete)
    • Am I wrong???
  – **GPA of a UO student:** \( x \in \mathbb{R}, 0 \leq x \leq 4.00 \)
    (Continuous)

Probability distribution

• **For a random variable** \( x \) (E.g. GPA of a UO students)
  – Which value is more likely? \( x>3.5 \) ?
  – Which value is less likely? \( x<2.7 \) ?
  – In general, what is the probability of a specific value \( x \)?
    • \( P(X=x) =? \)
    • (E.g. What is the probability of randomly picking up a student with GPA \( x=3.93 \)?)

  – **Convention:** Capital letters are used to show the random variable that can take all possible values from its domain, but small letters indicate a specific value that \( X \) can take (E.g. \( x=3.93 \))
Population vs. Sample

- **Population**: all members of a set
- **Sample**: a randomly chosen subset of the population
- **Example 1**:
  - Population: All students of the University of Oregon
  - Sample: Everybody who is in this class now!
- **Example 2**:
  - Population: All Americans who are eligible to vote
  - Sample: 1000 persons who are randomly chosen from this population

Parameter vs. Estimator

- A **parameter** is a numerical measurement that describes characteristic of a population, such as its mean or variance
  - Since we usually don’t have access to all members of the population, we never know the true value of the parameter
  - An **estimator** is a function that approximates the value of the parameter, based on the limited number of samples that are available
Methods of Point Estimation

- **Goal**: obtain the “best estimator”

- **Conditions**: 
  - Least bias (unbiased) 
  - Minimum variance of estimation error 
  - Recognize that we may need to tradeoff these conditions!

- **Applications**: 
  - Estimate distribution parameters e.g., mean, variance

- We now discuss the most “general” techniques to obtain point estimators. 
  - **Maximum Likelihood Estimator**
  - There are several other estimation techniques that we don’t talk about today: Method of Moments, Least Squares, Regularized likelihood, ...

### Maximum Likelihood Estimation (MLE)

- Developed by Sir R.A. Fisher (1920s)

- Preferred method by statisticians particularly if n is sufficiently large

- Maximum likelihood estimation maximizes the likelihood that the observed sample is a function of the possible parameter values.
Binary Variables (1)

• Coin flipping: heads=1, tails=0

\[ p(x = 1 | \mu) = \mu \]

• Bernoulli Distribution

\[
\begin{align*}
\text{Bern}(x | \mu) &= \mu^x (1 - \mu)^{1-x} \\
\mathbb{E}[x] &= \mu \\
\text{var}[x] &= \mu(1 - \mu) 
\end{align*}
\]

Binary Variables (2)

• N coin flips:

\[ p(m \text{ heads}|N, \mu) \]

• Binomial Distribution

\[
\begin{align*}
\text{Bin}(m|N, \mu) &= \binom{N}{m} \mu^m (1 - \mu)^{N-m} \\
\mathbb{E}[m] &= \sum_{m=0}^{N} m \text{Bin}(m|N, \mu) = N \mu \\
\text{var}[m] &= \sum_{m=0}^{N} (m - \mathbb{E}[m])^2 \text{Bin}(m|N, \mu) = N \mu(1 - \mu)
\end{align*}
\]
Binomial Distribution

Your first consulting job

Billionaire in Eugene asks:

– He says: I have thumbtack, if I flip it, what’s the probability it will fall with the nail up?
– You say: Please flip it a few times:

– You say: The probability is:
  • P(H) = 3/5
– He says: Why???
– You say: Because…
Thumbtack – Binomial Distribution

• P(Heads) = \( \theta \), P(Tails) = 1-\( \theta \)

• Flips are i.i.d.:
  – Independent events
  – Identically distributed according to Binomial distribution

• Sequence \( D \) of \( \alpha_H \) Heads and \( \alpha_T \) Tails

\[
P(D \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}
\]

Maximum Likelihood Estimation

• **Data**: Observed set \( D \) of \( \alpha_H \) Heads and \( \alpha_T \) Tails

• **Hypothesis**: Binomial distribution

• **Learning**: finding \( \theta \) is an optimization problem
  – What’s the objective function?
  \[
P(D \mid \theta) = \theta^{\alpha_H} (1 - \theta)^{\alpha_T}
\]

• **MLE**: Choose \( \theta \) to maximize probability of \( D \)
  \[
  \hat{\theta} = \arg \max_\theta P(D \mid \theta)
  \]
  \[
  = \arg \max_\theta \ln P(D \mid \theta)
  \]
Your first parameter learning algorithm

\[ \hat{\theta} = \arg \max_{\theta} \ln P(D | \theta) \]
\[ = \arg \max_{\theta} \ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T} \]

- Set derivative to zero, and solve!

\[ \frac{d}{d\theta} \ln P(D | \theta) = \frac{d}{d\theta} [\ln \theta^{\alpha_H} (1 - \theta)^{\alpha_T}] \]
\[ = \frac{d}{d\theta} [\alpha_H \ln \theta + \alpha_T \ln(1 - \theta)] \]
\[ = \alpha_H \frac{d}{d\theta} \ln \theta + \alpha_T \frac{d}{d\theta} \ln(1 - \theta) \]
\[ = \frac{\alpha_H}{\theta} - \frac{\alpha_T}{1 - \theta} = 0 \quad \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]

Question: For a given function \( f(x) \), what does its derivative \( f'(x) \) tell us about the function?

- Note derivative is positive where green, negative where red, and zero where black.

Source: Wikipedia.com
Parameter Estimation: Summary

• MLE for Bernoulli

Given: \( \mathcal{D} = \{x_1, \ldots, x_N\} \), \( m \) heads (1), \( N - m \) tails (0)

\[
p(\mathcal{D}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \prod_{n=1}^{N} \mu^{x_n}(1 - \mu)^{1 - x_n}
\]

\[
\ln p(\mathcal{D}|\mu) = \sum_{n=1}^{N} \ln p(x_n|\mu) = \sum_{n=1}^{N} \{x_n \ln \mu + (1 - x_n) \ln(1 - \mu)\}
\]

\[
\mu_{ML} = \frac{1}{N} \sum_{n=1}^{N} x_n = \frac{m}{N}
\]
But, how many flips do I need?

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]

- Billionaire says: I flipped 3 heads and 2 tails.
- You say: \( \theta = 3/5 \), I can prove it!
- He says: What if I flipped 30 heads and 20 tails?
- You say: Same answer, I can prove it!
- **He says: What’s better?**
- You say: Umm... The more the merrier???
- He says: Is this why I am paying you the big bucks???
- You say: I will give you a theoretical bound.

---

A bound (from Hoeffding’s inequality)

For \( N = \alpha_H + \alpha_T \), and

\[ \hat{\theta}_{MLE} = \frac{\alpha_H}{\alpha_H + \alpha_T} \]

Let \( \theta^* \) be the true parameter, for any \( \varepsilon > 0 \):

\[ P( | \hat{\theta} - \theta^* | \geq \varepsilon ) \leq 2e^{-2N\varepsilon^2} \]
PAC Learning

- **PAC:** Probably Approximate Correct
- **Billionaire says:** I want to know the thumbtack \( \theta \), within \( \varepsilon = 0.1 \), with probability at least \( 1 - \delta = 0.95 \).
- **How many flips?** Or, how big do I set \( N \)?

\[
P(|\hat{\theta} - \theta^*| \geq \varepsilon) \leq 2e^{-2N\varepsilon^2}
\]

\[
\delta \geq 2e^{-2N\varepsilon^2} \geq P(\text{mistake})
\]

\[
\ln \delta \geq \ln 2 - 2N\varepsilon^2
\]

\[
N \geq \frac{\ln(2/\delta)}{2\varepsilon^2} = \frac{\ln(2/0.05)}{2 \times 0.1^2} \approx \frac{3.8}{0.02} = 190
\]

What if I have prior beliefs?

- **Billionaire says:** Wait, I know that the thumbtack is “close” to 50-50. What can you do for me now?
- **You say:** I can learn it the Bayesian way...
- Rather than estimating a single \( \theta \), we obtain a distribution over possible values of \( \theta \)
Bayesian Learning

Use Bayes rule:

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

Or equivalently:

\[
P(\theta | D) \propto P(D | \theta)P(\theta)
\]

Also, for uniform priors:

\[
\rightarrow \text{ reduces to MLE objective}
\]

\[
P(\theta) \propto 1 \quad P(\theta | D) \propto P(D | \theta)
\]

Bayesian Learning for Thumbtacks

\[
P(\theta | D) \propto P(D | \theta)P(\theta)
\]

Likelihood function is Binomial:

\[
P(D | \theta) = \theta^H (1 - \theta)^T
\]

• What about prior?
  – Represent expert knowledge
  – Simple posterior form

• Conjugate priors:
  – Closed-form representation of posterior
  – For Binomial, conjugate prior is Beta distribution
Beta Distribution

- Distribution over $\mu \in [0, 1]$. 

\[ B(a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \]

\[
\begin{align*}
\text{Beta}(\mu | a, b) &= \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)} \mu^{a-1}(1 - \mu)^{b-1} \\
\mathbb{E}[\mu] &= \frac{a}{a + b} \\
\text{var}[\mu] &= \frac{ab}{(a + b)^2(a + b + 1)}
\end{align*}
\]

\[ B(a, b) = \int_0^1 u^{a-1}(1 - u)^{b-1}du, \quad a>0, b>0 \]

\[ \Gamma(a) = \int_0^\infty u^{a-1}e^{-u}du \]
Beta prior distribution – $P(\theta)$

$$P(\theta) = \frac{\theta^{\beta_H - 1}(1 - \theta)^{\beta_T - 1}}{B(\beta_H, \beta_T)} \sim Beta(\beta_H, \beta_T)$$

- Likelihood function: $P(D \mid \theta) = \theta^{\alpha_H}(1 - \theta)^{\alpha_T}$
- Posterior: $P(\theta \mid D) \propto P(D \mid \theta)P(\theta)$

$$P(\theta \mid D) \propto \theta^{\alpha_H}(1 - \theta)^{\alpha_T} \theta^{\beta_H - 1}(1 - \theta)^{\beta_T - 1}$$

$$= \theta^{\alpha_H + \beta_H - 1}(1 - \theta)^{\alpha_T + \beta_T - 1}$$

$$= Beta(\alpha_H + \beta_H, \alpha_T + \beta_T)$$

Posterior Distribution

- Prior: $Beta(\beta_H, \beta_T)$
- Data: $\alpha_H$ heads and $\alpha_T$ tails
- Posterior distribution:

$$P(\theta \mid D) \sim Beta(\beta_H + \alpha_H, \beta_T + \alpha_T)$$
Bayesian Posterior Inference

- **Posterior distribution:**
  \[ P(\theta \mid D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]

- **Bayesian inference:**
  - No longer single parameter
  - For any specific \( f \), the function of interest
  - Compute the expected value of \( f \)
  \[ E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid D) d\theta \]
  - Integral is often hard to compute

MAP: Maximum a Posteriori Approximation

\[ P(\theta \mid D) \sim \text{Beta}(\beta_H + \alpha_H, \beta_T + \alpha_T) \]
\[ E[f(\theta)] = \int_0^1 f(\theta) P(\theta \mid D) d\theta \]

- As more data is observed, Beta is more certain
- **MAP:** use most likely parameter to approximate the expectation
  \[ \hat{\theta} = \arg \max_{\theta} P(\theta \mid D) \]
  \[ E[f(\theta)] \approx f(\hat{\theta}) \]
MAP for Beta distribution

\[ P(\theta \mid D) = \frac{\theta^{\beta_H+\alpha_H-1}(1-\theta)^{\beta_T+\alpha_T-1}}{B(\beta_H + \alpha_H, \beta_T + \alpha_T)} \sim Beta(\beta_H+\alpha_H, \beta_T+\alpha_T) \]

MAP: use most likely parameter:

\[ \hat{\theta} = \arg \max_\theta P(\theta \mid D) = \frac{\alpha_H + \beta_H - 1}{\alpha_H + \beta_H + \beta_T + \beta_T - 2} \]

Beta prior equivalent to extra thumbtack flips
As \( N \to \infty \), prior is “forgotten”
But, for small sample size, prior is important!

What about continuous variables?

- Billionaire says: If I am measuring a continuous variable, what can you do for me?
- You say: Let me tell you about Gaussians...

\[ P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]
Some properties of Gaussians

- Affine transformation (multiplying by scalar and adding a constant) are Gaussian
  - $X \sim N(\mu, \sigma^2)$
  - $Y = aX + b \rightarrow Y \sim N(a\mu + b, a^2\sigma^2)$

- Sum of Gaussians is Gaussian
  - $X \sim N(\mu_x, \sigma^2_x)$
  - $Y \sim N(\mu_y, \sigma^2_y)$
  - $Z = X + Y \rightarrow Z \sim N(\mu_x + \mu_y, \sigma^2_x + \sigma^2_y)$

- Easy to differentiate, as we will see soon!

Learning a Gaussian

- Collect a bunch of data
  - Hopefully, i.i.d. samples
  - e.g., exam scores

- Learn parameters
  - Mean: $\mu$
  - Variance: $\sigma$

$$P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
MLE for Gaussian:

\[ P(x \mid \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \]

- Prob. of i.i.d. samples \( D=\{x_1, ..., x_N\} \):

\[ P(D \mid \mu, \sigma) = \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{i=1}^N e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \]

\[ \mu_{MLE}, \sigma_{MLE} = \arg \max_{\mu, \sigma} P(D \mid \mu, \sigma) \]

- Log-likelihood of data:

\[
\ln P(D \mid \mu, \sigma) = \ln \left[ \left( \frac{1}{\sigma \sqrt{2\pi}} \right)^N \prod_{i=1}^N e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right] \\
= -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i-\mu)^2}{2\sigma^2}
\]

Your second learning algorithm:

MLE for mean of a Gaussian

- What’s MLE for mean?

\[
\frac{d}{d\mu} \ln P(D \mid \mu, \sigma) = \frac{d}{d\mu} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^N \frac{(x_i-\mu)^2}{2\sigma^2} \right] \\
= \frac{d}{d\mu} \left[ -N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^N \frac{d}{d\mu} \left[ \frac{(x_i-\mu)^2}{2\sigma^2} \right] \\
= -\sum_{i=1}^N \frac{(x_i-\mu)}{\sigma^2} = 0 \\
= -\sum_{i=1}^N x_i + N \mu = 0
\]

\[ \hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^N x_i \]
MLE for variance

- Again, set derivative to zero:

\[
\frac{d}{d\sigma} \ln P(D \mid \mu, \sigma) = \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

\[
= \frac{d}{d\sigma} \left[ -N \ln \sigma \sqrt{2\pi} \right] - \sum_{i=1}^{N} \frac{d}{d\sigma} \left[ \frac{(x_i - \mu)^2}{2\sigma^2} \right]
\]

\[
= -\frac{N}{\sigma} + \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^3} = 0
\]

\[
\hat{\sigma}_{MLE}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]

Learning Gaussian parameters

- MLE:

\[
\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i
\]

\[
\hat{\sigma}_{MLE}^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]

- BTW. MLE for the variance of a Gaussian is **biased**
  - Expected result of estimation is **not** true parameter!
  - Unbiased variance estimator:

\[
\hat{\sigma}^2_{unbiased} = \frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \hat{\mu})^2
\]
Bayesian learning of Gaussian parameters

- Conjugate priors
  - Mean: Gaussian prior
  - Variance: Wishart Distribution

- Prior for mean:

\[
P(\mu \mid \eta, \lambda) = \frac{1}{\lambda \sqrt{2\pi}} e^{-\frac{(\mu - \eta)^2}{2\lambda^2}}
\]