Statistics and Estimation for Computer Science

İstanbul Teknik Üniversitesi

Mustafa Kamasak, PhD

Version: 2022.2.22

Point Estimation

Point Estimation

- A point estimate of a population parameter parameter θ is a single numeric value
- Point estimate: A single best guess about population parameter
- Point estimator: The function/statistic that produces point estimate
- Statistic: Any function of data
- Notation

$$\theta \rightarrow \text{population parameter}$$

$$\hat{\theta} \rightarrow \text{population parameter point estimate}$$

$$\hat{\Theta} \rightarrow \text{point estimator}$$

- Hence

$$\hat{\theta} = \hat{\Theta}(x; \theta)$$

  - x is the sample
  - θ is the parameter of interest

Example: Population Mean

- population parameter: $\theta = \mu$
- point estimate: $\hat{\theta} = \overline{x}$
- point estimator: $\hat{\Theta} = \frac{1}{N} \sum_i x_i$

What is Typically Estimated?

Any population parameter can be estimated. Frequently estimated population parameters are:

- Mean (μ)
- Std. dev. (σ)
- Proportion of a certain attribute in a population (p)
- Difference of means in two populations ($\mu_1 - \mu_2$)
- Difference of proportions in two populations ($p_1 - p_2$)

Unbiased Estimators

- A point estimator is unbiased if $E(\hat{\Theta}) = \theta$
- Bias of an estimator $\hat{\Theta}$ is defined as:

$$\text{bias} = E(\hat{\Theta}) - \theta$$

Standard Error (SE)

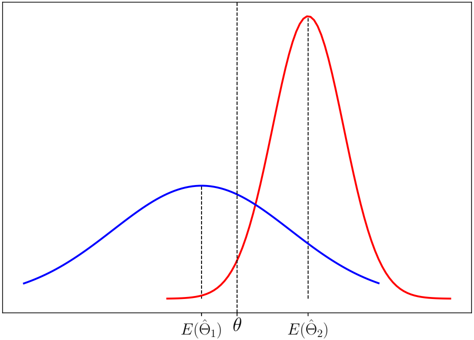
- Std dev of a point estimator is called **standard error (se)**

$$\text{s.e.} = \sigma_{\hat{\Theta}}$$

Example: Bias vs SE

Compare $\hat{\Theta}_1$ and $\hat{\Theta}_2$

- Which one has lower bias?
- Which one has lower se?



Minimum Variance Unbiased Estimator (MVUE)

- A population parameter may have multiple estimators
- Some of these estimators are unbiased
- Amongst unbiased estimator, the estimator with smallest variance is called **minimum variance unbiased Estimator (MVUE)**.

- Sample $= \{x_1, x_2, \cdots, x_n\}$
- Both x_1 and $\overline{x}$ are unbiased estimators of population mean μ
- However, $\overline{x}$ has lower se than x_1
- Infact $\overline{x}$ is MVUE

Mean Squared Error (MSE)

- MSE of an estimator is

$$\begin{aligned}\text{MSE} &= E(\hat{\Theta} - \theta)^2 \\ &= \text{variation} + \text{bias}^2 \\ &= \text{se}^2 + \text{bias}^2\end{aligned}$$

- If estimators are marked on a bias^2 vs variance graph, MSE of each estimator is its distance from origin
- Minimum MSE estimator is the one that is closest to the origin

Sampling Distribution of Sample Mean

- Typically sample mean $\overline{x}$ is used as an estimator of population mean (μ)
- Sampling distribution?

$$\overline{x} = \frac{1}{N} \sum_{i}^{N} x_i$$

- x_i are iid $\rightarrow$ CLT

$$\overline{x} \sim \mathcal{N}(?, ?)$$

- Expected value, variance ?

Expected Value of Sample Mean

$$\overline{x} \sim \mathcal{N}(?, ?)$$

Expected value of sample mean $E(\overline{x})$

$$\begin{aligned}
E(\overline{x}) &= E(\frac{1}{N}\sum_{i}^{N} x_i) \\
&= \frac{1}{N}\sum_{i}^{N} E(x_i) \\
&= \frac{1}{N}\sum_{i}^{N} \mu \\
&= \frac{1}{N}(N\mu) \\
&= \mu
\end{aligned}$$

Standard Error of Sample Mean

$$\overline{x} \sim \mathcal{N}(\mu, ?)$$

Standard error (se) of sample mean $E(\overline{x})$

$$\begin{aligned}
V(\overline{x}) &= V(\frac{1}{N}\sum_{i}^{N} x_i) \\
&= \frac{1}{N^2} V(\sum_{i}^{N} x_i) \\
&= \frac{1}{N^2} \sum_{i}^{N} V(x_i) \qquad \text{(Due to independence)} \\
&= \frac{1}{N^2}(N\sigma^2) \\
&= \frac{\sigma^2}{N}
\end{aligned}$$

Sampling Distribution of Sample Mean

$$\overline{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

- Sample mean is an unbiased estimator of population mean regardless of sample size
- Sample mean gets more precise (low se) with larger sample size

Bias and SE of Sample Variation

- Typically sample variation $\overline{x}$ is used as an estimator of population variation (μ)
- Sampling distribution?

$$\overline{x} = \frac{1}{N}\sum_{i}^{N} x_i$$

- x_i are iid $\rightarrow$ CLT

$$\overline{x} \sim \mathcal{N}(?, ?)$$

- Bias?
- Se?

Bias and SE of Sample Variance

- Typically sample variance s^2 is used as an estimator of population variance (σ^2)

$$s^2 = \frac{1}{N-1}\sum_{i}^{N}(x_i - \overline{x})^2$$

- Why divide with $N-1$ instead of N?

Bias and SE of Sample Variance

- Consider the following estimator

$$s_1^2 = \frac{1}{N}\sum_i^N (x_i - \overline{x})^2$$

- The expected value of this estimator is

$$\begin{aligned}
E(s_1^2) &= \frac{1}{N}E(\sum_i^N (x_i - \overline{x})^2) \\
&= \frac{1}{N}E(\sum_i^N x_i^2 - 2x_i\overline{x} + \overline{x}^2) \\
&= \frac{1}{N}\Big(\underbrace{E(\sum_i^N x_i^2)}_{A} - \underbrace{2E(\sum_i^N x_i\overline{x})}_{B} + \underbrace{E(\sum_i^N \overline{x}^2)}_{C}\Big)
\end{aligned}$$

Bias and SE of Sample Variance

$$
\begin{aligned}
A &= E(\sum_{i}^{N} x_i^2) \\
&= \sum_{i}^{N} E(x_i^2) \\
&= \sum_{i}^{N} (\sigma^2 + \mu^2) \\
&= N(\sigma^2 + \mu^2)
\end{aligned}
$$

Bias and SE of Sample Variance

$$
\begin{aligned}
B &= 2E(\sum_i^N x_i\overline{x}) \\
&= 2E(\frac{1}{N}\sum_i^N x_i^2) + 2E(\frac{1}{N}\sum_i^N \sum_{j,\ j\neq i} x_i x_j) \\
&= 2\frac{1}{N}N(\sigma^2+\mu^2) + 2\frac{1}{N}N(N-1)\mu^2 \\
&= 2\sigma^2 + 2N\mu^2
\end{aligned}
$$

Bias and SE of Sample Variance

$$
\begin{aligned}
C &= E(\sum_i^N \overline{x}^2) \\
&= \sum_i^N E(\overline{x}^2) \\
&= N(\mu_{\overline{x}}^2 + \sigma_{\overline{x}}^2) \\
&= N\mu^2 + N\frac{\sigma^2}{N} \\
&= N\mu^2 + \sigma^2
\end{aligned}
$$

Bias and SE of Sample Variance

$$E(s_1^2) = \frac{1}{N}\Big(\underbrace{E(\sum_i^N x_i^2)}_{A} - \underbrace{2E(\sum_i^N x_i\overline{x})}_{B} + \underbrace{E(\sum_i^N \overline{x}^2)}_{C} \Big)$$

$$= \frac{1}{N}\big(N(\sigma^2 + \mu^2) - 2(\sigma^2 + N\mu^2) + N\mu^2 + \sigma^2\big)$$

$$= \frac{N-1}{N}\sigma^2$$

- $E(s_1^2) \neq \sigma^2 \rightarrow s_1^2$ is a biased estimator
- s_1^2 is asymtotically unbiased $\rightarrow \lim_{n\rightarrow\infty} E(s_1^2) = \sigma^2$

Bias and SE of Sample Variance

- Typically sample variance s^2 is used as an estimator of population variance (σ^2)

$$s^2 = \frac{1}{N-1}\sum_{i}^{N}(x_i - \overline{x})^2$$

- Why divide with $N - 1$ instead of N? **Answer:** For unbiased estimation

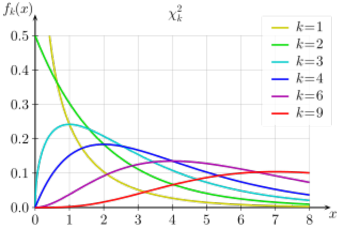
$$s^2 = \frac{SSE}{dof}$$

- SSE: sum of squared error
- dof: degrees of freedom is $N - 1$

Sampling Distribution of Sample Variance

$$\frac{(N-1)s^2}{\sigma^2} \sim \chi^2_{N-1} \qquad \text{dof=N-1}$$

- Chi square distribution
- Proof at `https://online.stat.psu.edu/stat414/node/174/`



Sampling Distribution of Sample Variance

$$
\begin{aligned}
E\left(\frac{(N-1)s^2}{\sigma^2}\right) &= E(\chi^2_{N-1}) \\
\frac{N-1}{\sigma^2}E(s^2) &= N-1 \\
E(s^2) &= \sigma^2
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}\left(\frac{(N-1)S^2}{\sigma^2}\right) &= \mathrm{Var}(\chi^2_{N-1}) \\
\frac{(N-1)^2}{\sigma^4}\mathrm{Var}(S^2) &= 2(N-1) \\
\mathrm{Var}(S^2) &= \frac{2(N-1)\sigma^4}{(N-1)^2} \\
&= \frac{2\sigma^4}{(N-1)},
\end{aligned}
$$

Estimation Methods

Method of Moments Estimation (MoM)

- Developed by Pearson at 1902.
- Derive first k moments of the population parameter – $E(X^k)$
- Compute first k moments of the sample parameter – $\frac{1}{N}\sum_i^N x_i^k$
- Equate theoretical and empirical moments $\rightarrow$ k equations and k unknowns

MoM Example: Bernoulli Distr

- Consider Bernoulli distribuion

$$X \in \{0, 1\} \text{ and } P(X = 1) = p$$

- 1st theoretical moment is

$$E(X) = p$$

- 1st empirical moment is

$$\overline{x} = \frac{1}{N} \sum_{i}^{N} x_i$$

- Equate them

$$\hat{p}_{MoM} = \frac{1}{N} \sum_{i}^{N} x_i$$

MoM Example

- X_i iid $\sim$Bernoulli(p)
- Observed outcomes = [1, 0, 0, 1, 1, 0, 1, 1, 1, 0]
- p=?

MoM Example: Normal Distr

- Consider Normal distribuion
- 1st theoretical moment is

$$E(X) = \mu$$

- 2nd theoretical moment is

$$E(X^2) = \mu^2 + \sigma^2$$

- 1st empirical moment is

$$\overline{x} = \frac{1}{N} \sum_{i}^{N} x_i$$

- 2nd empirical moment is

$$\frac{1}{N} \sum_{i}^{N} x_i^2$$

MoM Example: Normal Distr

- Equate them

$$\hat{\mu}_{MoM} = \overline{x}$$

$$\hat{\sigma}^2_{MoM} = \frac{1}{N}\sum_{i}^{N}(x_i - \overline{x})^2$$

- Biased (asymptotically unbiased) estimator for population variance

MoM Example: Poisson Distr

- Consider Poisson distribution
- 1st theoretical moment is

$$E(X) = \lambda$$

- 2nd theoretical moment is

$$E(X^2) = \lambda^2 + \lambda$$

- Equate 1st empirical moment

$$\hat{\lambda}_{MoM} = \overline{x}$$

- Equate 2nd empirical moment

$$\hat{\lambda}_{MoM} = \frac{1}{2}\left(\sqrt{\frac{4}{n}\sum_{i}^{N} x_i^2 + 1} - 1\right)$$

MoM

- Sometimes not useful (eg. Poisson distr)
- Sometimes moments do not exist

Maximum Likelihood Estimators (MLE)

- Let x_i iid with pdf $f(x_i; \theta)$
- Joint distribution is called likelihood of θ

$$\mathcal{L}(\theta) = \prod_i^N f(x_i; \theta)$$

- MLE of θ is the value that maximizes likelihood

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

- Notation:
  - $\max \mathcal{L}(\theta)$: maximum value of likelihood
  - $\arg\max \mathcal{L}(\theta)$: value of θ that maximizes likelihood
- Scaling and translation do not change $\arg\max$

$$\arg\max a\mathcal{L}(\theta) + b = \arg\max \mathcal{L}(\theta)$$

Log Likelihood

- We do not like optimize expressions with multiplication $\rightarrow$ use logarithm
- Log is a monotonic function

$$\mathcal{L}(\theta_1) > \mathcal{L}(\theta_2) \implies \log \mathcal{L}(\theta_1) > \log \mathcal{L}(\theta_2)$$

- Log likelihood is

$$\ell(\theta) = \log \mathcal{L}(\theta) = \sum_{i}^{N} f(x_i; \theta)$$

- MLE is

$$\hat{\theta}_{MLE} = \arg\max \ell(\theta)$$

Maximum Likelihood Example: Bernoulli Distr

- Let $x_i \sim \text{Bernoulli}(p)$ iid – $\theta = p$
- Define $s = \sum_i^N x_i$
- Log likelihood is

$$\begin{aligned}\ell(p) &= \log(\prod_i^N p^{x_i}(1-p)^{(1-x_i)}) \\ &= \log(p^s(1-p)^{(N-s)}) \\ &= s\log p + (N-s)\log(1-p)\end{aligned}$$

- Maximize wrt to p

$$\begin{aligned}\frac{d}{dp}\ell(p) &= 0 \\ 0 &= \frac{s}{p} - (N-s)\frac{1}{1-p} \\ 0 &= s(1-p) + p(N-s) \\ \hat{p}_{MLE} &= \frac{s}{N}\end{aligned}$$

Maximum Likelihood Example: Normal Distr

- Let $x_i \sim \mathcal{N}(\mu, \sigma^2)$ iid – $\theta = (\mu, \sigma)$ is a vector
- Log likelihood is

$$\ell(p) = \log(\prod_i^N \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\})$$

$$= -\frac{N}{2}\log(\pi) - N\log\sigma - \frac{1}{2\sigma^2}\sum_i^N (x_i - \mu)^2)$$

- Ignore $-N\pi$ and maximize wrt to θ

$$\frac{d}{d\mu}\ell(\theta) = 0$$

$$0 = \frac{1}{\sigma^2}\sum_i^N (x_i - \mu)$$

$$\hat{\mu}_{MLE} = \overline{x}$$

Maximum Likelihood Example: Normal Distr

- Maximize wrt to θ

$$\frac{d}{d\sigma}\ell(\theta) = 0$$

$$0 = \frac{-n}{\sigma} + \frac{1}{\sigma^2}\sum_{i}^{N}(x_i - \mu)^2$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i}^{N}(x_i - \overline{x})^2$$

Maximum Likelihood Example: Uniform Distr

- Let $x_i \sim \text{Uni}(0, \theta)$ iid
- pdf

$$f(x;\theta) = \begin{cases} \frac{1}{\theta} \text{ if } 0 \leq x \leq \theta \\ 0 \text{ otherwise} \end{cases}$$

- Let $x_{max} = \max\{x_i\}$
- Likelihood

$$\mathcal{L}(\theta) = \begin{cases} \frac{1}{\theta^N} \text{ if } \theta \geq x_{max} \\ 0 \text{ otherwise} \end{cases}$$

- $\ell(\theta)$ is zero before x_{max} and it monotonically decreases after x_{max}
- Hence,

$$\hat{\theta}_{MLE} = x_{max}$$

Properties of MLE

- Produces consistent estimators $\theta_{MLE} \xrightarrow{p} \theta$
- Equivariant. MLE of $g(\theta)$ is $g(\theta_{MLE})$ if $g(.)$ is a one-to-one function
- Asymptotically minimum variance estimator (asymtotically optimal/efficient
- Asymptotically normal sampling distribution

Fisher Information

- Define score function as:

$$s(x;\theta) = \frac{\partial \log f(x;\theta)}{\partial \theta}$$

- Score function is a measure of log likelihood sensitivity on θ
- Expected value of score function is zero

$$\begin{aligned} E(s(x;\theta)) &= \int_x \frac{\partial}{\partial \theta} \log f(x;\theta) f(x;\theta) dx \\ &= \int_x \frac{\frac{\partial}{\partial \theta} f(x;\theta)}{f(x;\theta)} f(x;\theta) dx \\ &= \frac{\partial}{\partial \theta} \underbrace{\int_x f(x;\theta) dx}_{1} \\ &= 0 \end{aligned}$$

Fisher Information

- Defined as

$$I(\theta) = E\big(s^2(x;\theta)\big)$$

- With some manipulation

$$I(\theta) = -E\big(\frac{\partial^2}{\partial\theta^2}\log f(x;\theta)\big)$$

- Theoretical standard error of MLE is

$$\mathsf{se} = \sqrt{\frac{1}{I(\theta)}}$$

- Estimated standard error of MLE is

$$\hat{\mathsf{se}} = \sqrt{\frac{1}{I(\hat{\theta})}}$$

Mixture Models

- Sometimes the observation comes from a mixture of different distributions
- Toy example[1]:
  - Consider 2 coins (A, B) with head probabilities p_1 and p_2
  - First pick a random coin with probabilities π_1, and π_2 ($\sum_i \pi_i = 1$).
  - Toss this coin 10 times and write the result
  - Repeat for 5 sets
  - Consider the following observation x:

```
x=[ [H, T, T, T, H, H, T, H, T, H],
    [H, H, H, H, T, H, H, H, H, H],
    [H, T, H, H, H, H, H, T, H, H],
    [H, T, H, T, T, T, H, H, T, T],
    [T, H, H, H, T, H, H, H, T, H] ]
```

  - What is the distribution of observations x?

[1] Example adopted from https://www.nature.com/articles/nbt1406

Mixture Models

- What is the distribution of observations x?
- Let $\theta = [\pi_1, \pi_2, p_1, p_2]$

$$
\begin{aligned}
f(x;\theta) &= \prod_i^5 \pi_1 f_1(x_i) + \pi_2 f_2(x_i) \\
&= \prod_i^5 \sum_k \pi_k f_k(x_i)
\end{aligned}
$$

where x_i is 10 tosses in observation i and f_k is the joint probability of x_i

$$
f_k(x_i) = \prod_m^{10} p_k^{I_H(x_{im})}(1 - p_k)^{I_T(x_{im})}
$$

and I is the indicator function

$$
I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{if } x \notin A \end{cases}
$$

MLE of Mixture Models

- Consider the following example, what is $\hat{\theta}_{MLE}$?

$$
\begin{aligned}
\hat{\theta}_{MLE} &= \arg\max f(x;\theta) \\
&= \arg\max \prod_i \sum_k \pi_k f_k(x_i) \\
&= \arg\max \sum_i \log \sum_k \pi_k f_k(x_i)
\end{aligned}
$$

- No closed form expression for $\hat{\theta}_{MLE}$?
- Need an iterative optimization for maximization of likelihood

Iterative Computation of MLE

Iterative Computation of MLE

- Start with an initial parameter value θ^0
- Until convergence
  - Update θ:

$$\theta^{k+1} \leftarrow \text{Update } \theta^k$$

  - (Log) likelihood value should increase monotonically with each update

$$\mathcal{L}(\theta^k) \leq \mathcal{L}(\theta^{k+1})$$

- How to update θ?
- Risk of local maxima
- Result may change with initial parameter value θ^0

Newton Rapson Method

- Define

$$\mathcal{L}(\theta)' \triangleq \frac{d}{d\theta}\mathcal{L}(\theta)$$

  and

$$\mathcal{L}(\theta)'' \triangleq \frac{d^2}{d\theta^2}\mathcal{L}(\theta)$$

- Using Newton-Rapson approximation

$$\mathcal{L}(\theta)' = \mathcal{L}(\theta^k)' + \mathcal{L}(\theta^k)''(\theta - \theta^k) + \text{higher order terms}$$

- For MLE we want $\mathcal{L}(\theta)' = 0 \approx \mathcal{L}(\theta^k)' + \mathcal{L}(\theta^k)''(\theta - \theta^k)$
- Hence, choose

$$\theta^{k+1} \leftarrow \theta^k - \frac{\mathcal{L}(\theta)'}{\mathcal{L}(\theta)''}$$

- If initial point is not good, ie. $\mathcal{L}(\theta^0)'' > 0$, then $\mathcal{L}(\theta)$ is minimized !!!
- Remember $\mathcal{L}(\theta)'$ is the direction where $\mathcal{L}(\theta)$ increases

Newton Rapson Method

- Other gradient based optimization methods can be used

$$\theta^{k+1} \leftarrow \theta^k - \alpha \frac{\mathcal{L}(\theta)'}{\mathcal{L}(\theta)''}$$

  or

$$\theta^{k+1} \leftarrow \theta^k + \alpha \mathcal{L}(\theta)'$$

  where α is the step size

- Selection of α is not easy (adaptive)
- Computation of $\mathcal{L}(\theta)'$ and/or $\mathcal{L}(\theta)''$ is not easy (may not exist)

Expectation Maximization (EM)

- Find another rv z_i such that $\log \prod_i f(x_i, z_i; \theta)$ is very easy to maximize
- z is called hidden/latent/missing data
- EM has two steps:
  - E-step: Fill the missing data with its expected value
  - M-step: Maximize log-likelihood

EM - Toy Example

For the toy example:

- if information of thrown coins (that is selected each time) is known, it is very easy to maximize log likelihood.
- Let selected coins be as follows

Coin	Observation
B	HTTTHHTHTH
A	HHHHTHHHHH
A	HTHHHHHTHH
B	HTHTTTHHTT
A	THHHTHHHTH

- Then

	n	Head	Tail
Coin A	3	24	6
Coin B	2	9	11

EM - Toy Example

- Likelihood of $z \sim \text{Multinomial}(\pi_1, \pi_2)$

$$\mathcal{L}(\pi_1, \pi_2) = \frac{5!}{3!2!}\pi_1^3 \pi_2^2$$

  with constraint of $\pi_1 + \pi_2 = 1$
- Use Lagrange multiplier

$$\ell(\pi_1, \pi_2, \lambda) = 3 \log \pi_1 + 2 \log \pi_2 + \lambda(1 - \pi_1 - \pi_2)$$

- Maximize $\ell(\pi_1, \pi_2, \lambda)$ with respect to π_k

$$\hat{p}_k = \frac{N_k}{\sum_k N_k} = \frac{N_K}{N}$$

- Hence, $\hat{\pi}_1 = \frac{3}{5}$ and $\hat{\pi}_2 = \frac{2}{5}$

EM - Toy Example

- Log likelihoods

$$\ell(p_1) = 24 \log p_1 + 6 \log(1 - p_1)$$
$$\ell(p_2) = 9 \log p_2 + 11 \log(1 - p_2)$$

- Then

$$\hat{p}_1 = \frac{24}{24 + 6}$$
$$\hat{p}_2 = \frac{9}{9 + 11}$$

EM - Toy Example

- Unfortunately selected coins are not known
- Let z_{ki} be an rv such that

$$z_{ki} = \begin{cases} 1 & \text{if Coin k is selected} \\ 0 & \text{otherwise} \end{cases}$$

- Initialize θ
- Iterate
  - E-step: Find expected values of z_{ki}
  - M-step: Maximize log likelihood using expected values of z_{ki} and update θ

EM - Toy Example

- Initialize θ: $\hat{\pi}$ and $\hat{p}$
- E-step:

$$\begin{aligned}E(z_{ki}) &= 1 \times P(Coin = k \mid x_i) + 0 \times P(Coin \neq k \mid x_i) \\ &= \frac{P(x_i \mid Coin = k) \times P(Coin = k)}{P(x_i)} \\ &= \frac{f_k(x_i)\hat{\pi}_k}{\sum_j f_j(x_i)\hat{\pi}_j}\end{aligned}$$

- Following table has to be filled
- Lets compute z_{12}

Observation	1	2	3	4	5
z_1		z_{12}			
z_2					

EM - Toy Example

- Let $\theta^0 = [\pi_1 = 0.5, \pi_2 = 0.5, p_1 = 0.6, p_2 = 0.5]$
- Lets compute z_{12} for coin A, observation 2
- Second observations is HHHHTHHHHH – 9H and 1T
- With 9H and 1T, one would expect z_{12} to be significantly larger then z_{22}, Why?

$$z_{12} = \frac{\pi_1 f_1(9H, 1T)}{\pi_1 f_1(9H, 1T) + \pi_2 f_2(9H, 1T)}$$

where

$$f_1(9H, 1T) = p_1^9 \times (1 - p_1)^1 = 0.6^9 \times 0.4^1 = 0.00403$$
$$f_2(9H, 1T) = p_2^9 \times (1 - p_2)^1 = 0.5^9 \times 0.5^1 = 0.00098$$

Then

$$z_{12} = \frac{0.5 \times 0.00403}{0.5 \times 0.00403 + 0.5 \times 0.00098} = \frac{0.00201}{0.0025} \approx 0.8049$$

EM - Toy Example

Observation	1	2	3	4	5	N_k
z_1	0.45	0.80	0.73	0.35	0.65	2.99
z_2	0.55	0.20	0.27	0.65	0.35	2.01

where

$$N_k \triangleq \sum_i z_{ki}$$

EM - Toy Example

Observation	1		...	5				
	z	H	T	...	z	H	T	$N_k(H)$	$N_k(T)$
z_1	0.45	5	5	...	0.65	7	3	2.13	0.86
z_2	0.55	5	5	...	0.35	7	3	1.17	0.84

- Compute total z values

$$N_k(H) = \sum_i z_{ki} \frac{1}{M} \sum_m I_H(x_{im})$$

$$N_k(T) = \sum_i z_{ki} \frac{1}{M} \sum_m I_T(x_{im})$$

EM - Toy Example

- M-step:
- Update probabilities

$$\hat{p}_k(H) = \frac{N_k(H)}{Nk}$$
$$\hat{\pi}_k(H) = \frac{N_k}{N}$$

EM - Toy Example

- E-step

$$z_{ik} = \frac{\pi_k f_k(x_i)}{\sum_{j=1}^{K} \pi_j f_k(x_i)}$$

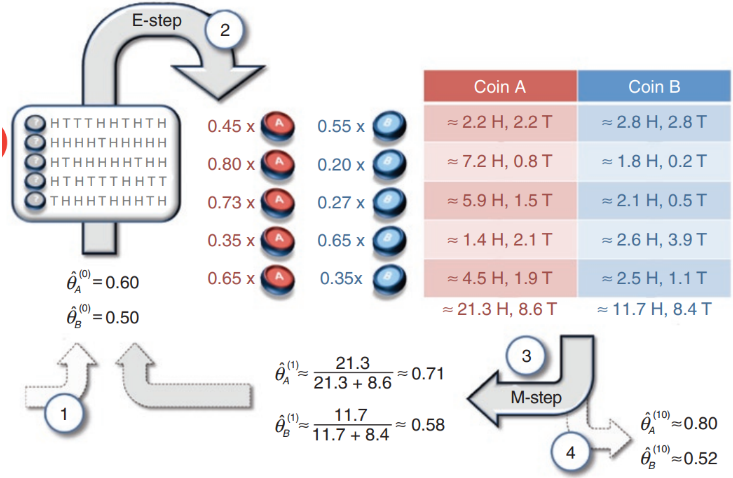
- M-step

$$N_k \triangleq \sum_i^N z_{ik}$$

$$\hat{\pi}_k = \frac{N_k}{N}$$

$$\hat{p}_k = \frac{1}{N_k} \sum_i^N z_{ik} \frac{1}{M} \sum_m^M I_H(x_{im})$$

EM - Toy Example



Taken from `https://www.nature.com/articles/nbt1406`

Gaussian Mixture Model (GMM)

- Instead of a coin toss with Bernoulli distr. use mixture of normal distr with different parameters
- The observation comes from K different normal distr. with (μ_k, σ_k^2)

$$f(x;\theta) = \prod_i \sum_k^K \pi_k \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left\{-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right\}$$

- Parameters are

$$\theta = [\pi_1, \cdots, \pi_K, \mu_1, \cdots, \mu_K, \sigma_1^2, \cdots, \sigma_K^2]$$

EM - Gaussian Mixture Model (GMM)

- E-step

$$z_{ik} = \frac{\pi_k f(x_i|\mu_k, \sigma_k^2)}{\sum_{j=1}^{K} \pi_j f(x_i|\mu_j, \sigma_j^2)}$$

- M-step

$$\hat{N}_k \triangleq \sum_i^N z_{ik}$$

$$\hat{\pi}_k = \frac{\hat{N}_k}{N}$$

$$\hat{\mu}_k = \frac{1}{\hat{N}_k} \sum_i^N z_{ik} x_i$$

$$\hat{\sigma}_k = \frac{1}{\hat{N}_k} \sum_i^N z_{ik} (x_i - \hat{\mu}_k)^2$$

Generalized Expected Maximization

- Similar to EM

- E-step: same as EM
- M-step: Instead of maximizing log likelihood, an increase is achieved
  - Gradient ascent
  - ...

k-means vs EM

- Instead of using $E(z_{ki})$ for each k, assign $z_{ki} \in \{0,\ 1\}$
- Assign $z_{ki} = 1$ for k with maximum expected value,
- Assign $z_{ki} = 0$ for the rest of mixtures

- This method is called **k-means** that is commonly used for segmentation

Maximum a Posteriori (MAP) Estimation

- Known also as Bayesian Estimation
- MLE: Find the parameter that maximize the likelihood of observation (data)

$$\hat{\theta}_{MLE} = \arg\max f(x; \theta)$$

- MAP: Find the **most likely** parameter with the observations

$$\hat{\theta}_{MAP} = \arg\max f(\theta|x)$$

- Remember Bayesian theorem

$$\begin{aligned}\hat{\theta}_{MAP} &= \arg\max f(\theta|x) \\ &= \frac{f(x;\theta)f(\theta)}{f(x)} \\ &= \frac{f(x,\theta)}{f(x)}\end{aligned}$$

where

- $f(\theta)$ is the prior information about parameter
- $f(x)$ is the joint distribution of data

Controversy over Bayesian Estimation

- Frequencist point of view: Parameter is **not** a random variable
  $\implies f(\theta) = ?$

- Bayesian point of view: There may be a degree of belief for population parameter, How?
  - Belief/Prejudice
  - Physics, geometry
  - Statistics
  - ...
- If $f(\theta)$ is uniform $\implies \hat{\theta}_{MLE} = \hat{\theta}_{MAP}$

MAP Estimator

- Bayesian Estimator is typically expressed as minimization of a cost function,
- Cost function $\mathcal{C}$ is formed by adding a loss function $\mathcal{L}$ to the negative log likelihood

$$
\begin{aligned}
\hat{\theta}_{MAP} &= \arg\max \prod_{i}^{N} \frac{f(x_i|\theta) f(\theta)}{f(x)} \\
&= \arg\max \sum_{i}^{N} \log f(x_i|\theta) + \log f(\theta) - \underbrace{\log f(x)}_{\text{Ignored}} \\
&= \arg\max \ell(\theta) + \log f(\theta) \\
&= \arg\min -\ell(\theta) + \mathcal{S}(\theta)
\end{aligned}
$$

- Loss function $\mathcal{L}$ penalizes parameter estimations that are distant from a prior value

Popular Loss Functions

- ℓ_p norm is defined as

$$\begin{aligned}\ell_p(\theta) &= \|\theta\|_p \\ &= \Big(\sum_i |\theta_i|^p\Big)^{1/p}\end{aligned}$$

- Hence,

$$\ell_p^p(\theta) = \sum_i |\theta_i|^p$$

- If $0 < p < 1$ then $\ell_p^p(\theta)$ is not convex
- If $1 \leq p$ then $\ell_p^p(\theta)$ is convex
- If cost function (both log likelihood and loss) is convex, convex optimization methods can be used

Popular Loss Functions

- Let θ_m is the ideal value population parameter

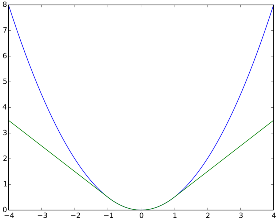
p	$\mathcal{S}(\theta)$	Note
0	$\lim_{p \to 0} \sum_i \theta_i^p$	number of nonzero parameters
1	$\lVert\theta - \theta_m\rVert_1$	absolute loss
2	$\lVert\theta - \theta_m\rVert_2^2$	quadratic loss
∞	$\max_i \{\theta_i\}$	maximum valued parameter

Huber Loss Function

- Quadratic loss penalizes distant parameters quite harshly
- Huber function

$$\mathcal{S}_\delta(\theta, \theta_m) = \begin{cases} \frac{1}{2}(\theta - \theta_m)^2 & \text{if } (\theta - \theta_m) < \delta \\ \delta|\theta - \theta_m| - \frac{1}{2}\delta^2 & \text{otherwise} \end{cases}$$

  - Quadratic for small differences
  - Linear for large differences



Hit or Miss/0 or 1 Loss Functions

- Hit-or-miss loss function

$$\mathcal{S}_\delta(\theta, \theta_m) = \begin{cases} 0 & \text{if } |(\theta - \theta_m)| < \delta \\ 1 & \text{otherwise} \end{cases}$$

- 0-or-1 loss function

$$\mathcal{S}_\delta(\theta, \theta_m) = I_0(\theta - \theta_m)$$

where $I(.)$ is the indicator function:

$$I_A(\Delta) = \begin{cases} 1 & \text{if } \Delta \in A \\ 0 & \text{if } \Delta \notin A \end{cases}$$