MATH 5010 –Foundations of Statistical Theory and Probability

❖ Point Estimation 1

-Finding Estimators

Instructor: He Wang
Department of Mathematics
Northeastern University

1. Method of Moments

2. Maximum Likelihood Estimators

3. Bayes Estimators

Point Estimator.

Sampling $\{X_1, \dots, X_n\}$ from a population distribution with pdf $f(x|\theta)$

Knowledge of θ or $g(\theta)$ yields knowledge of the entire population.

A **point estimator** is any function (statistic) $W(X_1, \dots, X_n)$ of a random sample.

Examples of point estimators include sample mean, sample variance, sample proportion, etc.

An **estimate** is the realized value of an estimator $W(x_1, \dots, x_n)$, when a sample is actually taken $(X_1, \dots, X_n) = (x_1, \dots, x_n)$.

- **Method of Moments (**Moment Matching**)**

$\{X_1, \ldots, X_n\}$ is a sample from a population distribution with pdf $f\left(x \middle| \vec{\theta}\right)$

Define:

$$
\begin{aligned}
m_1 &= \frac{1}{n}\sum_{i=1}^{n} X_i^1, \quad \mu_1' = \mathrm{E}X^1 \\
m_2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2, \quad \mu_2' = \mathrm{E}X^2 \\
&\vdots \\
m_k &= \frac{1}{n}\sum_{i=1}^{n} X_i^k, \quad \mu_k' = \mathrm{E}X^k.
\end{aligned}
$$

Here m_i is a sample moment and μ_i' is the population moment.

Solving the following linear system

$$\begin{aligned} m_1 &= \mu_1'(\theta_1, \ldots, \theta_k), \\ m_2 &= \mu_2'(\theta_1, \ldots, \theta_k), \\ &\vdots \\ m_k &= \mu_k'(\theta_1, \ldots, \theta_k). \end{aligned}$$

We obtain the method of moments estimator $\tilde{\theta}_i$ of θ_k.

Example (Normal)

Suppose random sample $X_i \sim Normal(\mu, \sigma^2)$ for $i = 1, \ldots, n$

As in Method of Moments, the notation is $\theta_1 = \mu \; and \; \theta_2 = \sigma^2$.

So,

$$m_1 = \bar{X} \qquad\qquad \mu_1' = \theta$$

$$m_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \qquad\qquad \mu_2' = \theta^2 + \sigma^2$$

Solving linear system:

$$\begin{aligned} m_1 &= \mu_1' \\ m_2 &= \mu_2' \end{aligned}$$

We get the method of moments estimators

$$\tilde{\theta} = \bar{X} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{1}{n}\sum X_i^2 - \bar{X}^2 = \frac{1}{n}\sum (X_i - \bar{X})^2.$$

Example (Binomial)

Suppose random sample $X_i \sim Binomial(k, p)$ for $i = 1, \dots, n$ with unknown k, p.

$$P(X_i = x | k, p) = \binom{k}{x} p^x (1-p)^{k-x}, \quad x = 0, 1, \dots, k.$$

Real world example: For a crime (with many unreported occurrences), both the true reporting rate, p, and the total number of occurrences,k, are unknown.)

Method of Moments yields linear system:

$$\bar{X} = kp$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^2 = kp(1-p) + k^2p^2$$

The solution give us the estimators:

$$\tilde{k} = \frac{\bar{X}^2}{\bar{X} - (1/n)\sum(X_i - \bar{X})^2} \qquad \text{and} \qquad \tilde{p} = \frac{\bar{X}}{\tilde{k}}$$

Remarks:

Method of Moments estimators for Normal distribution give the results coincide with our intuition.

Method of Moments estimators for binomial distribution are not the best estimators for the population parameters. In particular, it is possible to get negative estimates of $k \ and \ p$.

Another famous use of Method of Moments is the Satterthwaite approximation for cha square distribution.

Maximum Likelihood Estimator

$\{X_1, \dots, X_n\}$ is a sample from a population distribution with pdf $f\left(x\middle|\vec{\theta}\right)$

Given an observed sample $x_1, \dots, x_n$, **the likelihood function** is defined as

$$L\left(\vec{\theta}\middle|\vec{x}\right) \coloneqq L(\theta_1, \dots, \theta_k | x_1, \dots, x_n)$$

$$= f(x_1, \dots, x_n | \theta_1, \dots, \theta_k)$$

$$= \prod_{i=1}^{n} f\left(x_i\middle|\vec{\theta}\right)$$

The maximum likelihood estimator(MLE) is a parameter value $\widehat{\boldsymbol{\theta}}(\vec{\boldsymbol{x}})$ at which $L\left(\vec{\theta}\middle|\vec{x}\right)$ is maximized, that is

$$\widehat{\boldsymbol{\theta}}(\vec{\boldsymbol{x}}) := \underset{\vec{\theta}}{\operatorname{argmax}}\, L\left(\vec{\theta}\middle|\vec{x}\right)$$

The method of maximum likelihood is, by far, the most popular technique for deriving estimators.,

However, the optimization is a difficult mathematical problem.

One method is using Calculus by solving the critical points:

$$\frac{\partial}{\partial \theta_1} L(\vec{\theta}|\vec{x}) = 0$$

$$\frac{\partial}{\partial \theta_2} L(\vec{\theta}|\vec{x}) = 0$$

$$\vdots$$

$$\frac{\partial}{\partial \theta_k} L(\vec{\theta}|\vec{x}) = 0$$

To make the derivative easier, we will define the **log Likelihood function**

$$l(\vec{\theta}|\vec{x}) \coloneqq \log L(\vec{\theta}|\vec{x})$$

log Likelihood function $l(\vec{\theta}|\vec{x})$ and likelihood function $L(\vec{\theta}|\vec{x})$ will achieve extreme at the same position, that is

$$\widehat{\boldsymbol{\theta}}(\vec{\boldsymbol{x}}) := \underset{\vec{\theta}}{\operatorname{argmax}}\, L(\vec{\theta}|\vec{x})$$

$$= \underset{\vec{\theta}}{\operatorname{argmax}}\, l(\vec{\theta}|\vec{x})$$

Solving

$$\frac{\partial}{\partial\theta_i} l(\vec{\theta}|\vec{x}) = 0 \qquad \text{for } i = 1, \dots, k$$

Then check the local or global maxima.

Review Calculus

Definition: The **Hessian matrix** of $f: \mathbb{R}^n \to \mathbb{R}$ is defined by

$$H(f) := \begin{bmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{bmatrix}$$

Theorem: (Second Derivative Test)

If f: $\mathbb{R}^n \to \mathbb{R}$ is smooth, then a **critical point** $\vec{a} \in \mathbb{R}^n$ (i.e., $\nabla f(\vec{a}) = \vec{0}$,) is

(1) a local minimum if $H(f(\vec{a}))$ is positive definite;

(2) a local maximum if $H(f(\vec{a}))$ is negative definite;

(3) a saddle point if $H(f(\vec{a}))$ contains positive and negative eigenvalues;

(4) there is no conclusion for the other cases.

Example: Normal distribution.

Suppose random sample $X_i \sim Normal(\mu, \sigma^2)$ for $i = 1, \ldots, n$

Step 1. The likelihood function:

$$L(\, \mu, \sigma^2 | \vec{x}) \coloneqq \prod_{i=1}^{n} f(x_i | \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Step 2. The log likelihood function:

$$l(\, \mu, \sigma^2 | \vec{x}) \coloneqq \sum_{i=1}^{n} \log f(x_i | \mu, \sigma^2) = \sum_{i=1}^{n} \left[-\frac{(x_i - \mu)^2}{2\sigma^2} - \frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2\right]$$

$$= -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Step 3. Optimization(Take derivatives and solve)

1. MLE for $\boldsymbol{\mu}$

$$\frac{\partial}{\partial\mu} l(\mu, \sigma^2|\vec{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) \; = \frac{1}{\sigma^2} n(\bar{x} - \mu)$$

Solve $\frac{\partial}{\partial\mu} l(\mu, \sigma^2|\vec{x}) = 0$, we get

$$\hat{\mu}_{MLE} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2. MLE for $\boldsymbol{\sigma^2}$

$$\frac{\partial}{\partial\sigma^2} l(\mu,\sigma^2|\vec{x}) = -\frac{n}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2$$

(Treat σ^2 as θ)

Solve $\frac{\partial}{\partial\sigma^2} l(\mu,\sigma^2|\vec{x}) = 0$, we get

Replace μ by $\hat{\mu}_{MLE} = \bar{x}$

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Notice that this is a **biased** estimator (dividing by n, not $n-1$).

Exercise: Log-Normal

Suppose random sample $X_i \sim LogNormal(\mu, \sigma^2)$ for $i = 1, \dots, n$

$$Y_i = ln\ X_i \sim Normal(\mu, \sigma^2)$$

$$f_{X_i}(x_i) = \frac{1}{x_i\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right) \qquad x_i > 0$$

$$\hat{\mu}_{\mathrm{MLE}} = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$$

$$\hat{\sigma}^2_{\mathrm{MLE}} = \frac{1}{n}\sum_{i=1}^{n} (\ln x_i - \bar{y})^2$$

Example: Bernoulli

Suppose random sample $X_i \sim Bernoulli(p)$ for $i = 1, \dots, n$ with unknown p.

Step 1. The likelihood function:

$$L(\,p|\vec{x}) \coloneqq \prod_{i=1}^{n} f(x_i|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i}$$

Step 2. The log likelihood function:

$$l(\,p|\vec{x}) \coloneqq \sum_{i=1}^{n} \log f(x_i|p) = \sum_{i=1}^{n} [x_i \log p + (1-x_i)\log(1-p)]$$

$$= S \log p + (n-S)\log(1-p)$$

Here, denote $S = \sum_{i=1}^{n} x_i$, total number of successes.

Step 3. Optimization(Take derivatives and solve)

$$\frac{\partial}{\partial p} l(p|\vec{x}) = \frac{S}{p} - \frac{n-S}{1-p}$$

Solve $\frac{\partial}{\partial p} l(p\ |\vec{x}) = 0$, we get

$$\hat{p}_{MLE} = \frac{S}{n} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Example: Binomial Distribution (known number of trails)

Suppose random sample $X_i \sim Binomial(k, p)$ for $i = 1, \ldots, n$ with unknown p and known k.

Similar computation as Bernoulli distribution shows that

$$\hat{p}_{MLE} = \frac{S}{kn} = \frac{\sum_{i=1}^{n} x_i}{kn} = \text{sample proportation of successes}$$

This is intuitive: MLE for p is just the total number of observed successes divided by the total number of trials.

Example: Binomial MLE (unknown number of trials, and known p)

Suppose random sample $X_i \sim Binomial(k, p)$ for $i = 1, \ldots, n$ with known p and unknown k.

The likelihood function:

$$L(\,k|\vec{x}, p) \coloneqq \prod_{i=1}^{n} f(x_i|p) = \prod_{i=1}^{n} \binom{k}{x_i} p^{x_i}(1-p)^{k-x_i}$$

Because factorials are hard to differentiate and $k \in \mathbb{N}$, a **discrete optimization approach** is used instead.

Likelihood Ratio Test

To find where the likelihood increases or decreases

$$\frac{L(\ k|\vec{x},k)}{L(\ k-1|\vec{x},k)} = \frac{(k(1-p))^n}{\prod_{i=1}^{n}(k-x_i)}$$

When this ratio $\geq\ 1$, increasing k increases the likelihood.

When this ratio $<\ 1$, increasing k decrease the likelihood.

Therefore, **the MLE is the largest $\boldsymbol{k}$ such that**:

$$(k(1-p))^n \geq \prod_{i=1}^{n}(k-x_i)$$

They transform the condition by defining $z = \frac{1}{k}$, derive the equation:

$$(1-p)^n = \prod_{i=1}^{n}(1 - x_i z)$$

This gives a **strictly decreasing function** of z on the interval:

$$0 \le z \le \frac{1}{\max_i x_i}$$

So there is a unique solution

$$\hat{z} \in \left(0, \frac{1}{\max x_i}\right]$$

The MLE of k is

$$\hat{k}_{MLE} = \left\lfloor \frac{1}{\hat{z}} \right\rfloor$$

Example [Scale Uniform]

Suppose random sample $X_i \sim Uniform(0, \theta)$ for $i = 1, \ldots, n$

The pdf of a uniform distribution on $[0, \theta]$ is given by

$$f_X(x) = \frac{1}{\theta} \text{ for } x \in [0, \theta]$$

Use the method of maximum likelihood to estimate θ

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i; \theta) = \frac{1}{\theta^n}$$

This is a decreasing function, so it does not have a maximum.

We know that $x_i \leq \theta$ for all $i = 1, 2, \ldots, n$. Hence $\theta \geq \max(x_1, x_2, \cdots, x_n)$

So, if we want to maximize $L(\theta)$, we need to choose $\theta_{MLE} = \max(x_1, x_2, \cdots, x_n)$

Invariance property of maximum likelihood estimators

Theorem: Suppose $\hat{\theta}$ is the MLE of θ. Then, for any function $g(\theta)$, the MLE of $g(\theta)$ is $g(\hat{\theta})$.

Reason:

$$\widehat{\boldsymbol{\theta}}(\vec{\boldsymbol{x}}) := \underset{\vec{\theta}}{\operatorname{argmax}}\, L\left(\vec{\theta} \middle| \vec{x}\right)$$

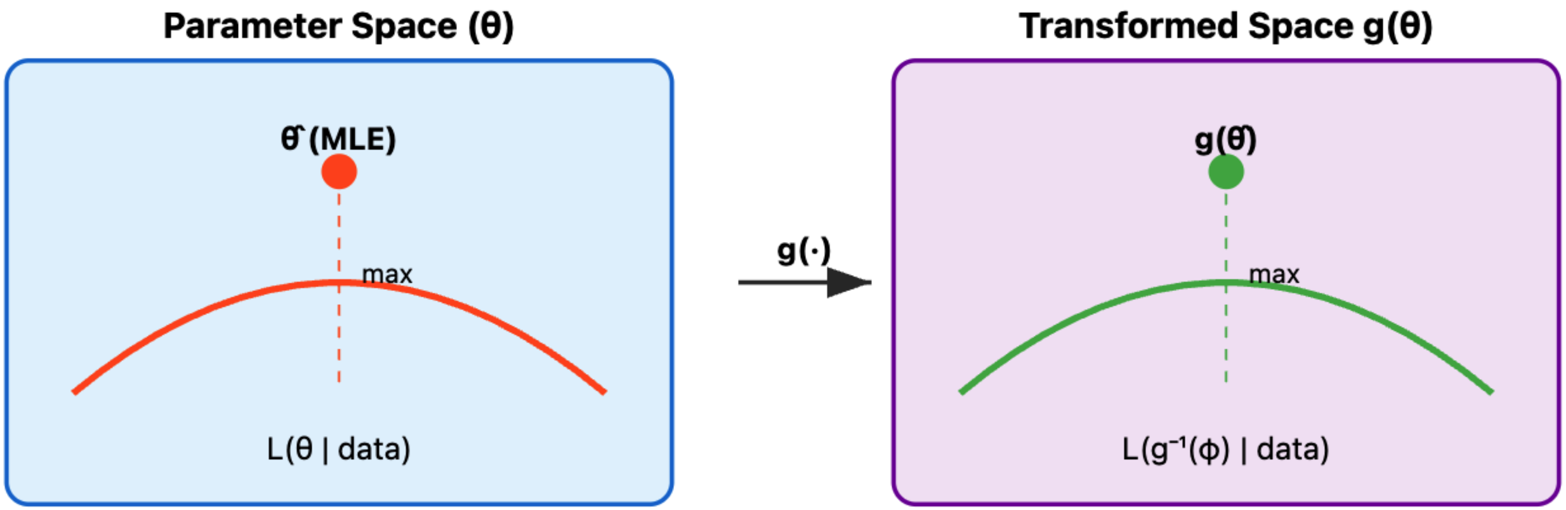
Now, let $\phi = g(\theta)$, for example, $g(\theta) = \sqrt{\theta}$ or $g(\theta) = \log(\theta)$

We could try to reparametrize the likelihood in terms of $\phi = g(\theta)$, , but that just moves the curve horizontally.

The *height* of the likelihood doesn’t change — only the *label* of the x-axis.

So the value of $\phi = g(\theta)$ that corresponds to the maximum is just $g(\hat{\theta})$.

MLE Invariance Property



Key Insight

If $\hat{\theta}$ maximizes L(θ | data), then $g(\hat{\theta})$ maximizes the likelihood in the transformed space

This works because g(·) is a one-to-one transformation that preserves the location of the maximum (the transformation doesn't change which parameter value is best)

Example

If $\hat{\theta} = 2$ is the MLE of a variance parameter θ

Then $\sqrt{2} \approx 1.414$ is the MLE of the standard deviation $g(\theta) = \sqrt{\theta}$

This property tells us that once we've found the MLE of a parameter θ, we don't need to redo the MLE process to estimate any function of θ — we can just **plug in** the MLE into the function.

Example:

Suppose random sample $X_i \sim Normal(\mu, \sigma^2)$ for $i = 1, \dots, n$

$$\widehat{\sigma^2}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Then by invariance, MLE of σ is

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

Bayes Estimators

Random Sample $X_1, \dots, X_n$ from a population distribution with pdf $f(x|\theta)$

In the Method of Moments and MLE, the parameter θ is thought to be an unknown but fixed quantity.

In the **Bayesian** approach, θ is considered to be a quantity whose variation can be described by a probability distribution $\pi(\theta)$ (called the **prior distribution**).

The updated distribution of θ based on the prior $\pi(\theta)$ and the sample information $\vec{x} = (x_1, \dots, x_n)$ is called the **posterior distribution**.

By Bayes' Rule:

$$\pi(\theta|\vec{x}) = \frac{f(\vec{x}|\theta)\pi(\theta)}{m(\vec{x})}$$

$m(\vec{x})$ is the marginal distribution of $\vec{x}$: $\quad m(\vec{x}) = \int f(\vec{x}|\theta)\pi(\theta)\, d\theta$

The posterior distribution is a random quantity for θ.

The **mean** (or **median**, or **max**) of the posterior distribution $\pi(\theta|\vec{x})$ can be used as a point estimate of θ.

$$\hat{\theta}_{MAP} = \operatorname{argmax} \pi(\theta|\vec{x})$$

Bayesian Method Example-Binomial distribution

Suppose there is a coin that may be biased with unknown probability θ of giving a "heads."

Suppose we flip the coin n times and observe x "heads."

The probability of this observation given the value of θ, comes from binomial distribution:

$$P(x|n,\theta) = \binom{n}{x}\theta^{x}(1-\theta)^{n-x}$$

Classical (Frequentist) method:

The frequentist approach is to construct an estimator for θ, which in theory can be any function of the observed data $\hat{\theta}(x,n)$ and show that $\hat{\theta} \to \theta$ as $n \to \infty$.

The classical estimator in this case is the empirical frequency (use MLE)

$$\hat{\theta} = \frac{x}{n}$$

❑ Bayesian approach

Frequentist approach ignores all prior information.

Bayesian approach choose a **prior distribution** $p(\theta)$. A convenient prior in this case is the Beta distribution:

$$p(\theta \mid \alpha, \beta) \ = \mathcal{B}(\theta;\ \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

$$\text{with } \Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt$$

$$Or\ write\ \mathcal{B}(\theta;\ \alpha, \beta) = \frac{1}{B(\alpha, \beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

with normalizing constant $B(\alpha, \beta) = \ \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\,\mathrm{d}\theta$

A **prior probability distribution** $p(\theta)$ to θ, representing your **degree** of belief with respect to θ.

Given our observations $D = (x, n)$, we can now compute the **posterior distribution** of θ by Bayes Theorem:

$$p(\theta \mid x, n, \alpha, \beta) = \frac{P(x|n, \theta)p(\theta|\alpha, \beta)}{P(x|n, \alpha, \beta\ )}$$

$$= \frac{P(x|n, \theta)p(\theta|\alpha, \beta)}{\int_0^1 P(x|n, \theta)p(\theta|\alpha, \beta)\, d\theta}$$

$$= \ \ldots$$

$$= \frac{1}{B(\alpha + x, \beta + n - x)} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}$$

$$= \mathcal{B}(\theta;\ \alpha + x, \beta + n - x)$$

If we assign a different prior distribution, then we arrive at a different posterior distribution. (α, β are hyperparameters.)

Beta distribution Calculator: https://homepage.divms.uiowa.edu/~mbognar/applets/beta.html

Summary of Bayesian approach

$P(\mathcal{D} \mid \vec{\theta})$: the **likelihood** for the data. The parameters of interest $\vec{\theta}$ (unknown)

$P(\vec{\theta})$: density associated with the **prior distribution**. (The **degree of belief** of the distribution before the data.)

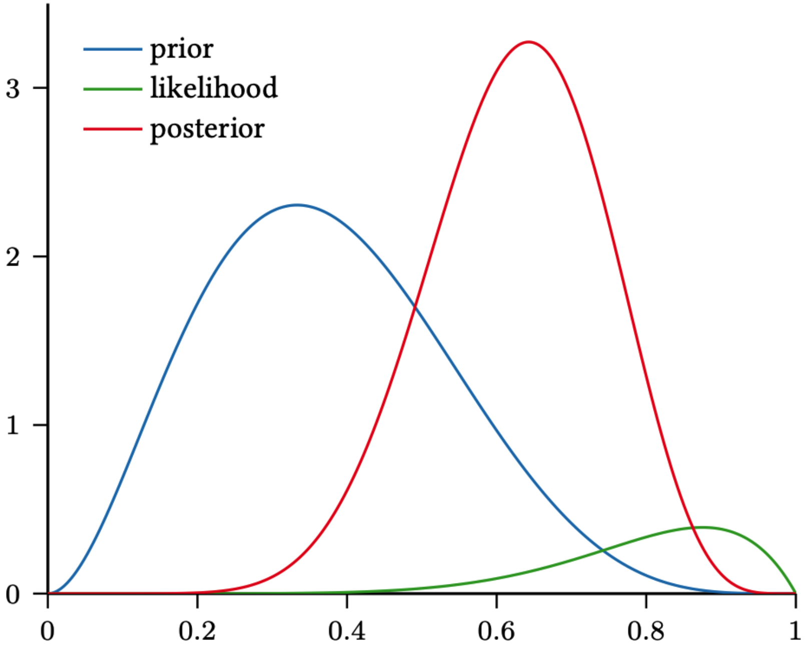
Bayes' theorem converts a prior probability into a **posterior probability**

$$P(\vec{\theta}|\mathcal{D}) = \frac{P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})}{P(\mathcal{D})} = \frac{P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})}{\int P(\mathcal{D} \mid \vec{\theta}')\, P(\vec{\theta}')\, d\vec{\theta}'}$$

Posterior distribution $P(\vec{\theta}|\mathcal{D})$ is the updated degree of belief with respect to θ, based on the data $\mathcal{D}$. The **new degree of belief** is called the posterior probability distribution of θ.

The rather convenient fact that the posterior remains a beta distribution is because the beta distribution satisfies a property known as **conjugacy** with the binomial likelihood.

α, β serve as "pseudocounts," or fake observations we pretend to have seen before seeing the data. They are hyperparameters we need to determine.



$$(\alpha, \beta) = (3, 5)$$

$$(x, n) = (5, 6).$$

We also get **another** "test and confidence interval" (**credible interval**), for example.

$$P\left(\theta < \frac{1}{2} \mid x, n, \alpha, \beta\right) = \int_0^{1/2} p(\theta | x, n, \alpha, \beta) d\theta$$

Once we derive the posterior distribution $p(\theta\ |\mathcal{D})$ using Bayesian approach, we can study the posterior mean and posterior variance.

Posterior mean

$$E[\theta|\mathcal{D}]$$

Posterior variance

$$Var[\theta|\mathcal{D}]$$

For example, the posterior distribution of the probability θ of giving head in coin is

$$E[\theta|\mathcal{D}] = \frac{\alpha + x}{\alpha + \beta + n}$$

$$Var[\theta|\mathcal{D}] = \frac{(\alpha + x)(\beta + n - x)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}$$

In addition, the Mode

Mode [highest point in density curve]= $\frac{\alpha+x-1}{\alpha+\beta+n-2}$

Conjugate prior

$$P(\vec{\theta}|\mathcal{D}) = \frac{P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})}{P(\mathcal{D})}$$

For some likelihood functions $P(\mathcal{D} \mid \vec{\theta})$, if you choose a certain prior $P(\vec{\theta})$, the posterior $P(\vec{\theta}|\mathcal{D})$ ends up being in the same distribution as the prior. Such a prior then is called a **Conjugate Prior** of the likelihood function.

For example, the Beta distribution $\mathcal{B}(\theta;\ \alpha, \beta)$ is the conjugate prior of the binomial distribution $Binomial(x; n, \theta)$.

The conjugate prior of the normal distribution $N(\mu, \sigma^2)$ with fixed σ is also normal distribution, but with different parameters.

If the likelihood function belongs to the exponential family, then a conjugate prior exists, often also in the exponential family.

A Table of conjugate distributions can be found in Wikipedia:

https://en.wikipedia.org/wiki/Conjugate_prior

How does the Conjugate Prior help?

When you know that your prior is a conjugate prior, you can skip the computation.

$$\text{posterior} \propto \text{likelihood} * \text{prior}$$

During the modeling phase, we already know the posterior will also be a beta distribution.

Therefore, after carrying out more experiments, **you can compute the posterior simply by adding the number of acceptances and rejections to the existing parameters α, β respectively**, instead of multiplying the likelihood with the prior distribution**.**

Maximum A Posteriori estimation (MAP) v.s. MLE

Suppose $\vec{\theta}$ are the model parameters, and $\mathcal{D} = \left\{(\vec{x}^{(i)}, \vec{y}^{(i)})\right\}_{i=1}^{N}$ the observed data.

$P(\mathcal{D} \mid \vec{\theta})$: the **likelihood** for the data.

The **Maximum Likelihood Estimate (MLE)** of is

$$\hat{\vec{\theta}}_{MLE} \coloneqq \underset{\vec{\theta}}{\operatorname{argmax}}\, P(\mathcal{D} \mid \vec{\theta}) = \underset{\vec{\theta}}{\operatorname{argmax}}\, \log P(\mathcal{D} \mid \vec{\theta})$$

Most of the model we have are using MLE, e.g., logistics regression, linear regression, generalized linear regression, From Bayesian statistics, based on **prior** $p(\vec{\theta})$, we have calculated the **posterior** distribution:

$$P(\vec{\theta}|\mathcal{D}) = \frac{P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})}{P(\mathcal{D})}$$

The Maximum A Posteriori estimation (MAP) is

$$\hat{\vec{\theta}}_{MAP} \coloneqq \underset{\vec{\theta}}{\operatorname{argmax}}\, P(\vec{\theta}|\mathcal{D}) = \underset{\vec{\theta}}{\operatorname{argmax}}\, P(\mathcal{D} \mid \vec{\theta})\, P(\vec{\theta})$$

$$= \underset{\vec{\theta}}{\operatorname{argmax}}\, \log P(\mathcal{D} \mid \vec{\theta}) + \log P(\vec{\theta})$$

Example (MAP for μ in normal distribution.)

Suppose we have iid data $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$ observed from normal distribution $N(\mu, \sigma^2)$ with known σ.

$$p\left(x^{(i)}|\mu\right) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(x^{(i)} - \mu\right)^2\right)$$

The **MLE** for μ is

$$\hat{\mu}_{MLE} = \underset{\mu}{\operatorname{argmax}} \log P(\mathcal{D}\,|\mu) = \underset{\mu}{\operatorname{argmax}} \log \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(x^{(i)} - \mu\right)^2\right) =$$

$$\dots = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$$

Now find the **MAP** estimate of μ.

The conjugate prior of normal distribution is normal, there is a closed-form solution analytically.

$$\hat{\mu}_{MAP} = \underset{\mu}{\operatorname{argmax}} \log P(\mu|\mathcal{D}) = \underset{\mu}{\operatorname{argmax}} \log P(\mathcal{D}\,|\mu)\, P(\mu)$$

Notice that

$$P(\mu)P(\mathcal{D}\,|\mu) = \frac{1}{\sqrt{2\pi}\,\sigma_0} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right) \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{1}{2\sigma^2}\left(x^{(i)} - \mu\right)^2\right)$$

Hence,

$$\hat{\mu}_{MAP} = \underset{\mu}{\operatorname{argmax}} \log P(\mathcal{D}\,|\mu)\, P(\mu)$$

$$= \underset{\mu}{\operatorname{argm}in} \sum_{i=1}^{N} \frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 + \frac{1}{2\sigma^2}\left(x^{(i)} - \mu\right)^2$$

$$= \frac{\sigma_0^2\left(\sum_{i=1}^{N} x^{(i)}\right) + \sigma^2\mu_0}{\sigma_0^2 N + \sigma^2}$$

$$Var(\mu|\vec{x}) = \frac{\sigma^2\sigma_0^2}{\sigma^2 + N\sigma_0^2}$$

The MAP estimate $\hat{\mu}_{MAP}$ is a linear combination between the prior mean μ_0 and the sample mean $\bar{x}$ weighted by their respective covariances.

$\hat{\mu}_{MAP} \rightarrow \hat{\mu}_{MLE}$ when $\sigma_0 \rightarrow \infty$

References:

- **Book 1. [CB] Statistical Inference**, by Casella, George, Berger, Roger L, 2nd edition
- **Book 2. [W]: All of Statistics: Larry Wasserman**

Online books and courses:

- https://www.probabilitycourse.com/
- https://online.stat.psu.edu/stat415/
- https://stat110.hsites.harvard.edu/
- https://bookdown.org/egarpor/inference/