Population, Sample, Parameter

1. In statistics, the population is any subject (not necessarily a group) which a researcher try to analyze.
2. The sample is a collection of data related to the population. In a typical situation, the sample is assumed to be randomly and independently extracted from the population.
3. The parameter represents a property of the population to be analyzed. The parameter is unknown to the researcher.
4. The goal of statistics is to obtain useful insights about the parameter of the population with the sample extracted from it.

Population Distribution

We regard the population as a probability distribution and call it the population distribution. Then we can interpret the sample as a set of random variables following the population distribution, and the parameters as variables which determine the "shape" of the population distribution.

Let $\boldsymbol{D = (x_1, \ldots, x_n)}$ denote the sample, and $\boldsymbol{\theta}$ denote the parameter of the population distribution. To indicate that the shape of the population distribution depends on $\boldsymbol{\theta}$, the population p.f. or p.d.f. is denoted by $\boldsymbol{p(x_i|\theta)}$ where each $\boldsymbol{x_i}$ $\boldsymbol{(i = 1, \ldots, n)}$ is called an observation and supposed to be a realized value of the random variable following the population distribution. $\boldsymbol{n}$ is often called the sample size.

Likelihood

Suppose the sample $\boldsymbol{D} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are taken from a population distribution where the parameter $\boldsymbol{\theta}$ is a set of unknown parameters. The joint p.f. or the joint p.d.f of $\boldsymbol{D}$ is denoted by

$$\boldsymbol{p}(\boldsymbol{D}|\boldsymbol{\theta}) = \boldsymbol{p}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n|\boldsymbol{\theta}).$$

In particular, if observations are independent of each other,

$$\boldsymbol{p}(\boldsymbol{D}|\boldsymbol{\theta}) = \boldsymbol{p}(\boldsymbol{x}_1|\boldsymbol{\theta}) \times \cdots \times \boldsymbol{p}(\boldsymbol{x}_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} \boldsymbol{p}(\boldsymbol{x}_i|\boldsymbol{\theta}).$$

When we regard $\boldsymbol{p}(\boldsymbol{D}|\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$, it is called the likelihood or likelihood function.

Example: Bernoulli Distribution i

Let us define a random variable X_i $(i = 1, \ldots, n)$ corresponding to tossing a coin such that

$$X_i = \begin{cases} 1, & \text{Head is obtained;} \\ 0, & \text{Tail is obtained,} \end{cases}$$

and

$$\Pr(X_i = 1) = \theta, \quad \Pr(X_i = 0) = 1 - \theta.$$

Then X_i follows the Bernoulli distribution and its p.f. is given by

$$p(x_i|\theta) = \theta^{x_i}(1-\theta)^{1-x_i}, \quad x_i = 0, 1.$$

Example: Bernoulli Distribution ii

Then the joint p.f. of $D = (x_1, \ldots, x_n)$ is

$$p(D|\theta) = \prod_{i=1}^{n} p(x_i|\theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i}$$
$$= \theta^{y}(1-\theta)^{n-y}, \quad y = \sum_{i=1}^{n} x_i.$$

Suppose we have $(x_1, x_2, x_3, x_4, x_5) = (1, 0, 1, 1, 1)$. The value of $p(D|\theta)$ depends on the value of θ.

θ	0.1000	0.2000	0.3000	0.4000	0.5000
$p(D\|\theta)$	0.0001	0.0013	0.0057	0.0154	0.0312
θ	0.6000	0.7000	0.8000	0.9000	
$p(D\|\theta)$	0.0518	0.0720	0.0819	0.0656	

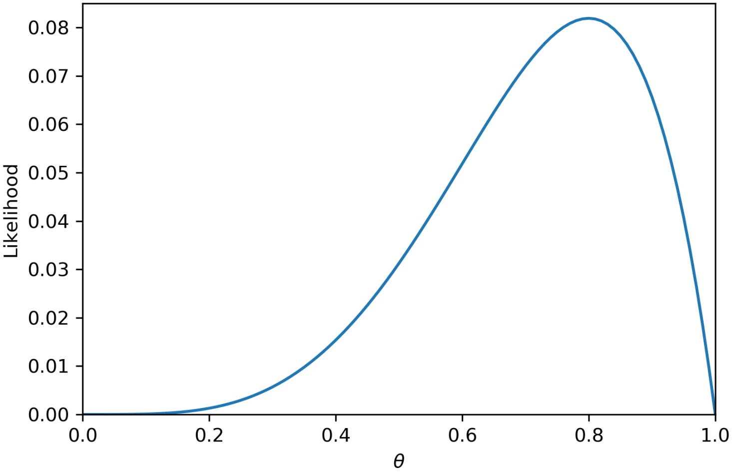

Figure 1: The likelihood of θ in the Bernoulli distribution

Interpretation Of The Likelihood

Given the sample $\boldsymbol{D}$, the likelihood $\boldsymbol{p(D|\theta)}$ is regarded as a kind of "plausibility" of a specific value of $\boldsymbol{\theta}$.

For example, the likelihood of $\boldsymbol{\theta = 0.9}$ is $\boldsymbol{0.656}$ while that of $\boldsymbol{\theta = 0.4}$ in the previous example is $\boldsymbol{0.0154}$. We may say that $\boldsymbol{0.9}$ is about 4 times more plausible than $\boldsymbol{0.4}$ as the true value of $\boldsymbol{\theta}$.

To make comparison between two competing values of $\boldsymbol{\theta}$, say $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_1}$, we introduce the likelihood ratio:

$$\text{likelihood ratio} = \frac{\boldsymbol{p(D|\theta_0)}}{\boldsymbol{p(D|\theta_1)}}.$$

Prior Knowledge On Parameters

In practice, researchers often have information on unknown parameters before they start analysis. For example,

- $\boldsymbol{\theta}$ must take a value between 0 and 1 because it is probability;
- in case of tossing a coin, $\boldsymbol{\theta}$ is supposed to be 50% if the coin is fair.

In Bayesian statistics, we construct a distribution of unknown parameters that reflect our prior knowledge on their true values. This is call the prior distribution. Let $\boldsymbol{p(\theta)}$ denote the prior distribution.

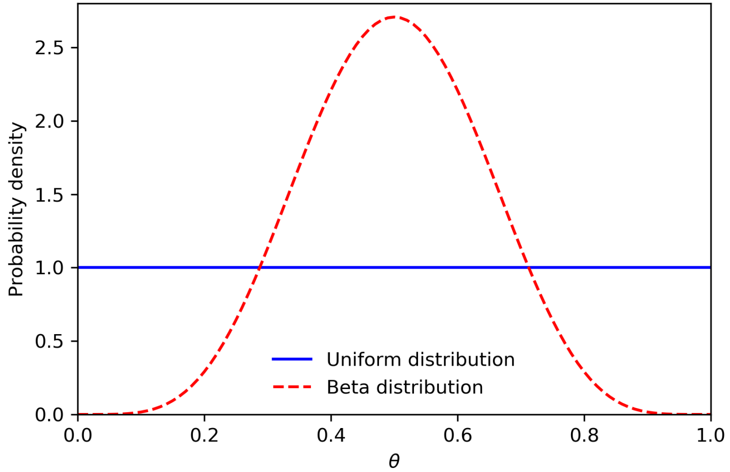

Figure 2: Prior distributions of θ in the Bernoulli distribution

The uniform distribution Uniform(a, b) is

$$p(x|a,b) = \begin{cases} \frac{1}{b-a}, & (a \leqq x \leqq b); \\ 0, & (\text{otherwise}). \end{cases}$$

In the above figure, we set $a = 0$ and $b = 1$.

The beta distribution Beta(α, β) is

$$p(x|\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)},\ 0 \leqq x \leqq 1.$$

where $B(\alpha, \beta)$ is the beta function:

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx.$$

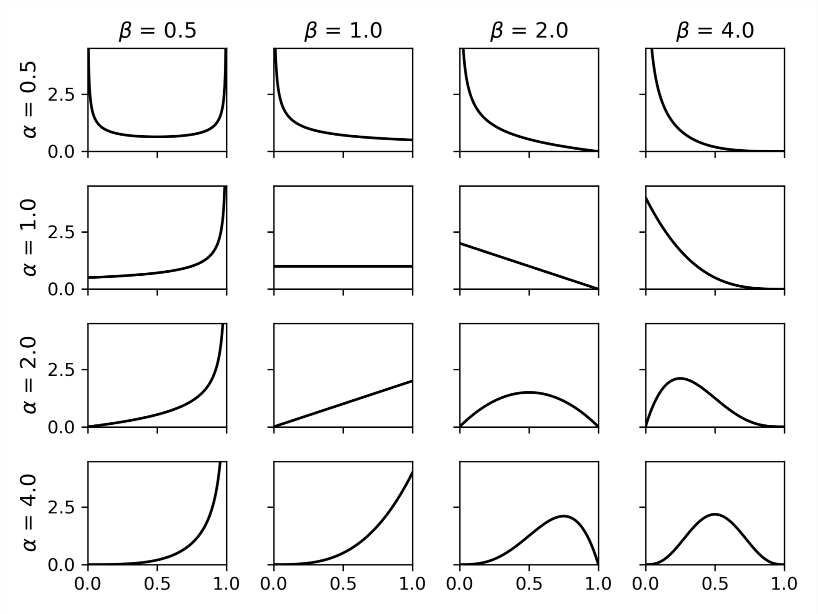

Figure 3: Beta distributions with various (α, β)

Bayes' Theorem And Posterior Distribution

Suppose $\boldsymbol{p(\theta, D)}$ is the joint distribution $\boldsymbol{\theta}$ and $\boldsymbol{D}$. Using the definition of the conditional distribution, we have

$$\boldsymbol{p(\theta, D) = p(\theta|D)p(D) = p(D|\theta)p(\theta).}$$

Arranging the middle and the right-hand side of the above equation, we have

$$\boldsymbol{p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)}.}$$

- The above formula is called Bayes' theorem.
- $\boldsymbol{p(\theta|D)}$ is called the posterior distribution.
- $\boldsymbol{p(D)}$ is called the normalizing constant.

Marginal Likelihood

By marginalizing the joint distribution $\boldsymbol{p(\theta, D)}$, $\boldsymbol{p(D)}$ is given as

$$\boldsymbol{p(D) = \int p(\theta, D) d\theta = \int p(D|\theta) p(\theta) d\theta.}$$

Thus $\boldsymbol{p(D)}$ is interpreted as "averaged likelihood" in terms of the prior $\boldsymbol{p(\theta)}$. In this sense, $\boldsymbol{p(D)}$ is called the marginal likelihood.

Then Bayes' theorem is rewritten as

$$\boldsymbol{p(\theta|D) = \frac{p(D|\theta)p(\theta)}{\int p(D|\theta)p(\theta)d\theta}.}$$

Example: Bernoulli Distribution

Suppose the prior distribution is Beta(α_0, β_0).

The posterior distribution of θ is given by

$$\begin{aligned}
p(\theta|D) &\propto p(D|\theta)p(\theta) \\
&\propto \theta^{y}(1-\theta)^{n-y} \times \theta^{\alpha_0 - 1}(1-\theta)^{\beta_0 - 1} \\
&\propto \theta^{y+\alpha_0-1}(1-\theta)^{n-y+\beta_0-1} \\
&\propto \theta^{\alpha_\star - 1}(1-\theta)^{\beta_\star - 1}, \\
\alpha_\star = y + \alpha_0, &\quad \beta_\star = n - y + \beta_0.
\end{aligned}$$

This is the beta distribution $\textit{Beta}(\alpha_\star, \beta_\star)$.

Bayesian Learning

Bayes' theorem is rearranged as

$$\frac{\boldsymbol{p(\theta|D)}}{\boldsymbol{p(\theta)}} = \frac{\boldsymbol{p(D|\theta)}}{\boldsymbol{p(D)}} \propto \boldsymbol{p(D|\theta)}.$$

The left-hand side is

$$\begin{cases} \dfrac{\boldsymbol{p(\theta|D)}}{\boldsymbol{p(\theta)}} > \mathbf{1}, & \text{plausibility of } \boldsymbol{\theta} \text{ is increased;} \\[2em] \dfrac{\boldsymbol{p(\theta|D)}}{\boldsymbol{p(\theta)}} < \mathbf{1}, & \text{plausibility of } \boldsymbol{\theta} \text{ is decreased.} \end{cases}$$

In other words, an update in belief is proportional to the likelihood $\boldsymbol{p(D|\theta)}$.

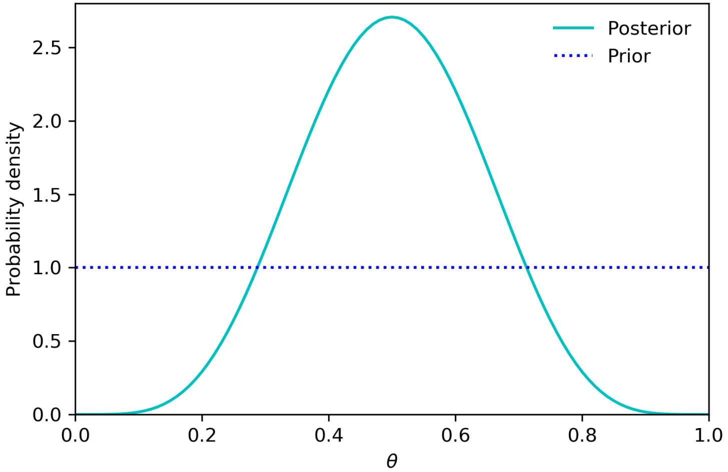

Figure 4: Posterior distributions of the probability of success

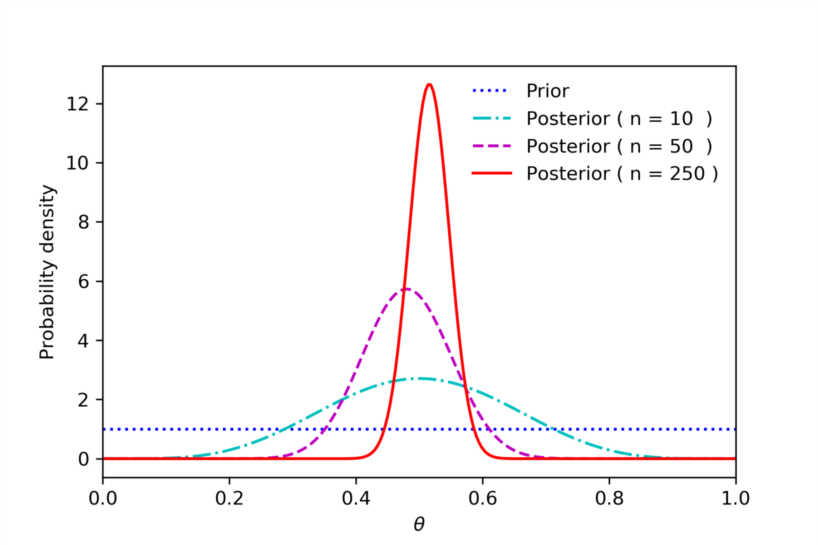

Figure 5: Sequential updating of the posterior distribution