# Statistical estimation

- consider a coin with outcome
  - HEADS with a probability $\mu(H) = p$, and
  - TAILS otherwise
- we don't know $p$ (called a **parameter**), and want to estimate it from random trials

- **Maximum Likelihood Principle**:
  - Choose the parameter that maximizes the probability of observed data
  - say random trials gave an outcome $\omega$=(H,H,T,H,T,H,H,T,…,T)

$$\mu(\omega \,|\, p) = p^{\#H}(1-p)^{\#T}$$

# Maximum Likelihood (ML) estimation

- Formulate it as maximization of log-likelihood

$$
\begin{aligned}
p^* &= \arg\max_p \log\big(\mu(\omega)\big) \\
&= \arg\max_p \underbrace{\#H \log p + \#T \log(1-p)}_{\mathscr{L}(p)}
\end{aligned}
$$

- To solve this optimization problem analytically, (which can be done only for some special cases), we take the gradient of the objective function and set it to zero

$$
\frac{\partial \mathscr{L}(p)}{\partial p} = \frac{\#H}{p} - \frac{\#T}{1-p} = 0
$$

- which gives $p^* = \dfrac{\#H}{\#H + \#T}$, which is consistent with our intuition

# Sufficient statistics

- Sufficient statistics of an outcome $\omega$ is a function of $\omega$ that is compact and captures everything we need to know in order to compute $\mu(\omega)$
- in this example, recall that
$$\mu(\omega) \;=\; p^{\#H}(1-p)^{\#T}$$
- So sufficient statistics of $\omega = (H, H, T, H, T, T, T, T, H, \dots, H)$ is $(\#H, \#T)$
- In particular the order of the H's and T's do not matter (because they are independent trials)
- When running the experiment, we do not have to keep track of all sequence of outcomes, but jus the counts

# Bayesian estimation

- **Bayes theorem**
  - $\mu(x_1 | x_2) = \dfrac{\mu(x_1)\mu(x_2 | x_1)}{\mu(x_2)}$
  - proof: $\mu(x_1, x_2) = \mu(x_2 | x_1)\mu(x_1) = \mu(x_1 | x_2)\mu(x_2)$
  - this is useful in assessing
    **diagnostic probability** from **causal probability**:
  - $\mu(cause | effect) = \dfrac{\mu(effect | cause)\mu(cause)}{\mu(effect)}$

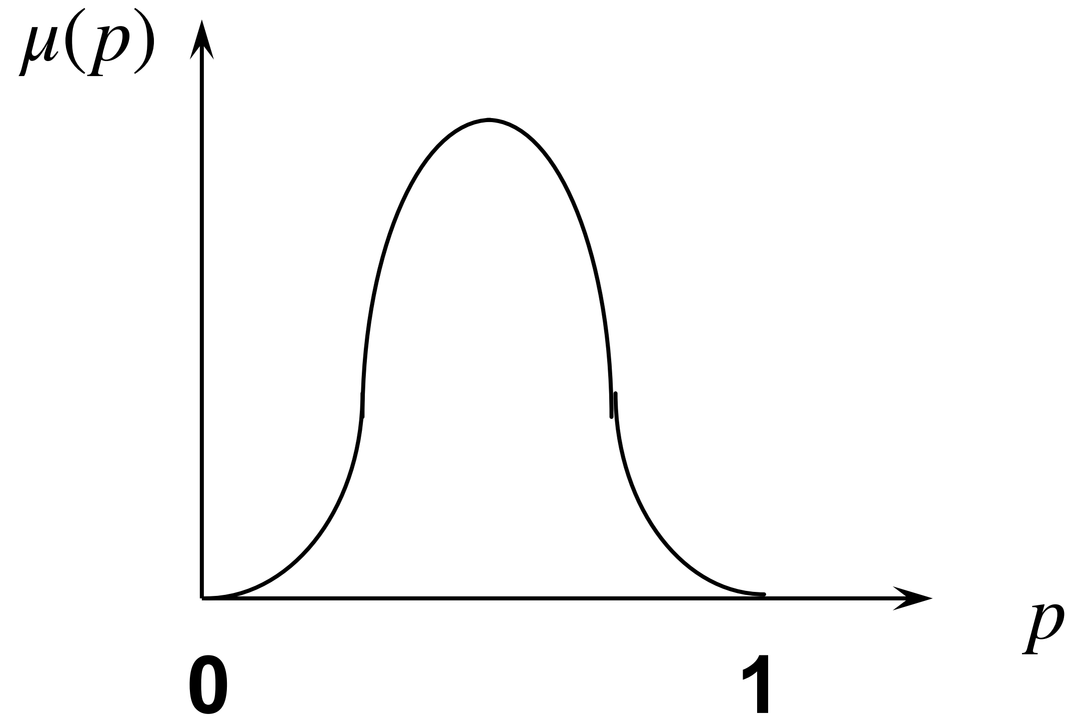
    usually, $\mu(effect | cause)$ is known, whereas $\mu(cause | effect)$ is not
  - For example, if $m$ is meningitis, and $s$ is stiff neck
    $\mu(m = 1 | s = 1) = \dfrac{\mu(s = 1 | m = 1)\mu(m = 1)}{\mu(s = 1)} = \dfrac{0.8 \times 0.0001}{0.1} = 0.0008$
    which is very small (because $\mu(m = 1)$ is small)

  - **prior:** $\mu(m = 1)$ is called a prior distribution, which is the marginal distribution of the cause without any observations
  - **posterior:** $\mu(m = 1 | s = 1)$ is called posterior distribution, which is the conditional distribution of the cause given observation

# Bayesian estimation

- True probability $p$ of the coin is unknown

  but we know it comes from a known probability distribution $\mu(p)$ for example,



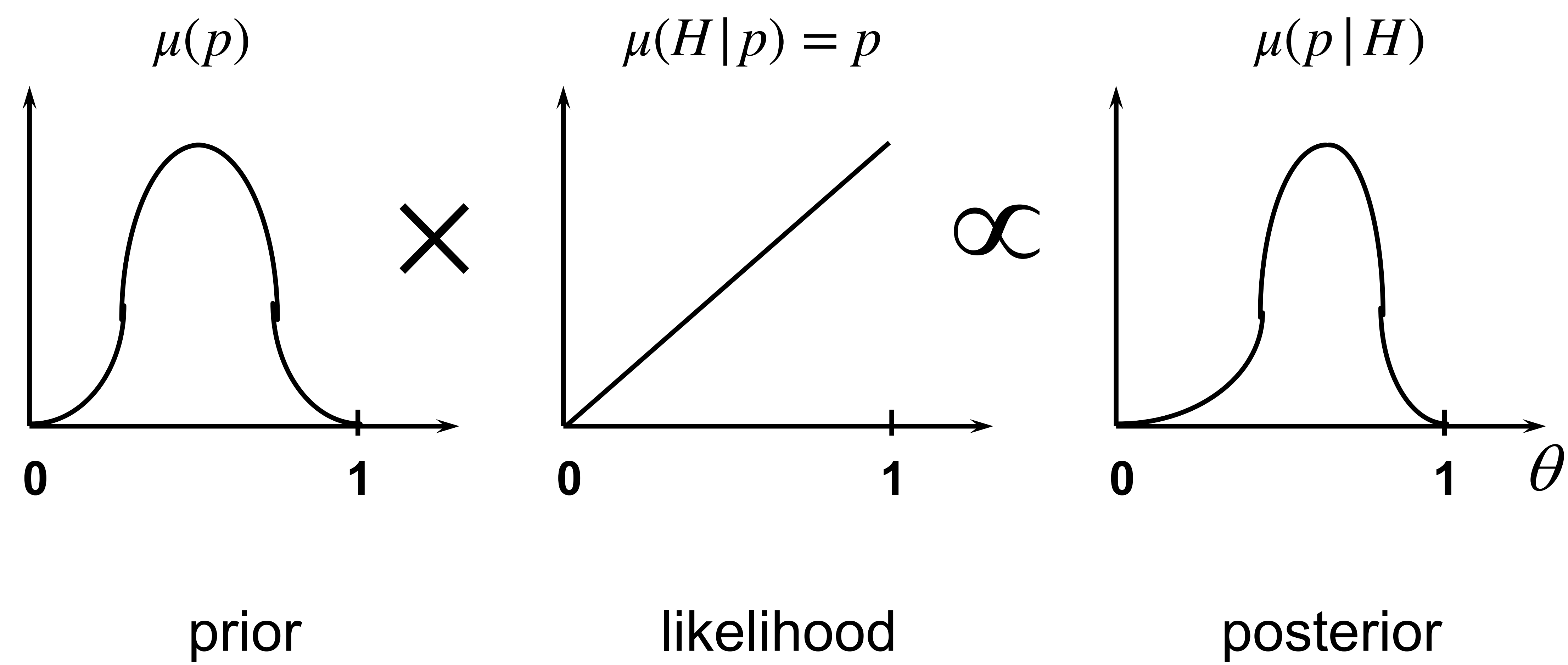

- note that $p$ is a continuous random variable
- applying Bayes theorem, ($p'$ is just a variable that we integrate out)

$$\mu(p \,|\, H) = \frac{\mu(p)\mu(H \,|\, p)}{\int_0^1 \mu(p')\mu(H \,|\, p')dp'}$$

# Bayesian estimation

- $$\mu(p\,|\,H) = \frac{\mu(p)\mu(H\,|\,p)}{\int_0^1 \mu(p')\mu(H\,|\,p')dp'} \;\propto\; \mu(p)\mu(H\,|\,p)$$

  as the denominator does not depend on $p$



# Inference task: Probability of HEADS on next toss

- **Conditional independence** makes solving inference task much more efficient
- consider the task of tossing the coin twice and let the outcome be $x = (x_1, x_2)$
- It is easy to see that the two trials are conditionally independent given $p$, i.e. $\mu(x_1, x_2 \mid p) = \mu(x_1 \mid p)\mu(x_2 \mid p)$
- now, we consider the **inference task** of estimating the probability of HEADS on next toss
- **Inference task** is a task of making a prediction/decision based on some joint distribution, and we will make this notion mathematically precise later on
- First step is to write down what we want to know in terms of the **joint distribution,** because the joint distribution is well defined

$$\mu(x_{n+1} = H \mid x_1^n = (HTHHT\ldots)) \;=\; \int_0^1 \mu(x_{n+1} = H, p \mid x_1^n = (HTHHT\ldots))\, dp$$

which is marginalizing out $p$ from the joint distribution,

- next, we apply chain rule to the term in the integral

$$\mu(x_{n+1} = H, p \mid x_1^n = (HT\ldots)) \;=\; \mu(x_{n+1} = H \mid p, x_1^n)\mu(p \mid x_1^n)$$

- and simplify using conditional independence

$$= \; \mu(x_{n+1} = H \mid p)\mu(p \mid x_1^n) \;=\; p\,\mu(p \mid x_1^n)$$

- Putting together we get

$$\mu(x_{n+1} = H \mid x_1^n = (HTHHT\ldots)) \;=\; \int_0^1 p\,\mu(p \mid x_1^n = (HTHHT\ldots))\, dp \;=\; \mathbb{E}[p \mid x_1^n = (HTHHT\ldots)]$$

# Inference task: Probability of HEADS on next toss

- So, on order to solve this inference task rigorously, we want to compute

$$\mu(x_{n+1} = H \,|\, x_1^n = (HTHHT\ldots)) \;=\; \mathbb{E}[p \,|\, x_1^n = (HTHHT\ldots)]$$

but in many case, computing the integral (i.e. averaging) can be challenging

- instead we use **Maximum a Posteriori (MAP) estimation**:

choosing the value with the highest posterior probability

**MAP estimation**

$$p^* = \arg\max_p \mu(p \,|\, x_1^n)$$

**Maximum likelihood (ML) estimation**

$$p^* = \arg\max_p \mu(x_1^n \,|\, p)$$