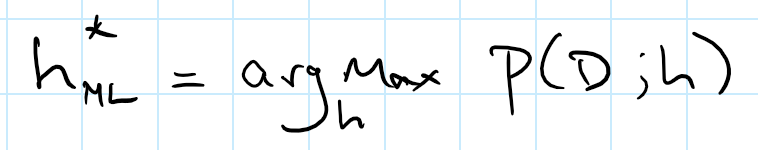
# 解释最大似然、最大后验概率和贝叶斯参数估计

你有一枚硬币。你把它翻三次，三次都是头朝上。下一次抛硬币的概率是多少？这是一个基于数据的参数估计的机器学习问题。在这种情况下，我们要根据数据D估计头部h的概率。

#### 最大似然

这样做的一个方法是找到参数H的值，使观测数据的可能性最大化，即p（d；h）。这里，我们使用；来表示h是分布P的一个参数，这意味着h定义P，但P仅指定观测数据D的可能性。



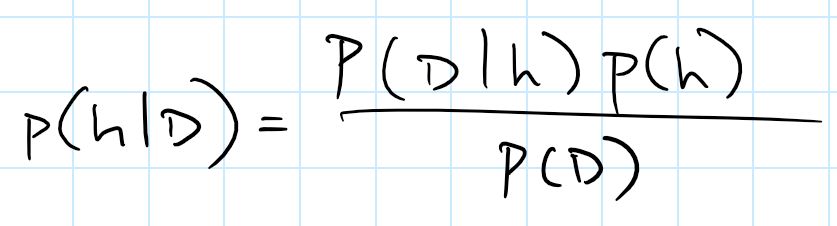
这是一种频率估计方法，被称为极大似然估计。在这种方法下，我们估计h=1.0。

但我们的直觉告诉我们这可能不是真的。我们知道，对于大多数硬币来说，我们也有可能得到尾部，通常我们期望h=0.5。

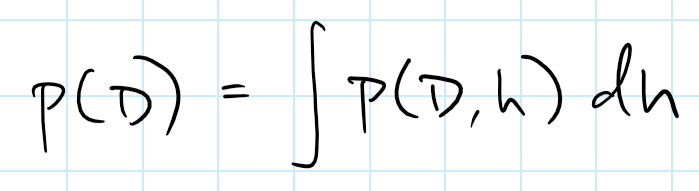
#### 前前后后

我们如何从数学上对这种直觉进行编码？我们可以指定数据上的联合分布和我们的参数：p（D，h）=p（D | h）p（h）。我们指定了一个先验分布p（h），它编码我们关于h应该有什么值的直觉，以及一个给定h，p（D | h）值的条件分布。

我们现在如何估计从D开始的h？我们需要后验分布p（h | D），但我们只有p（D | h）和p（h）。贝耶斯规则来拯救！



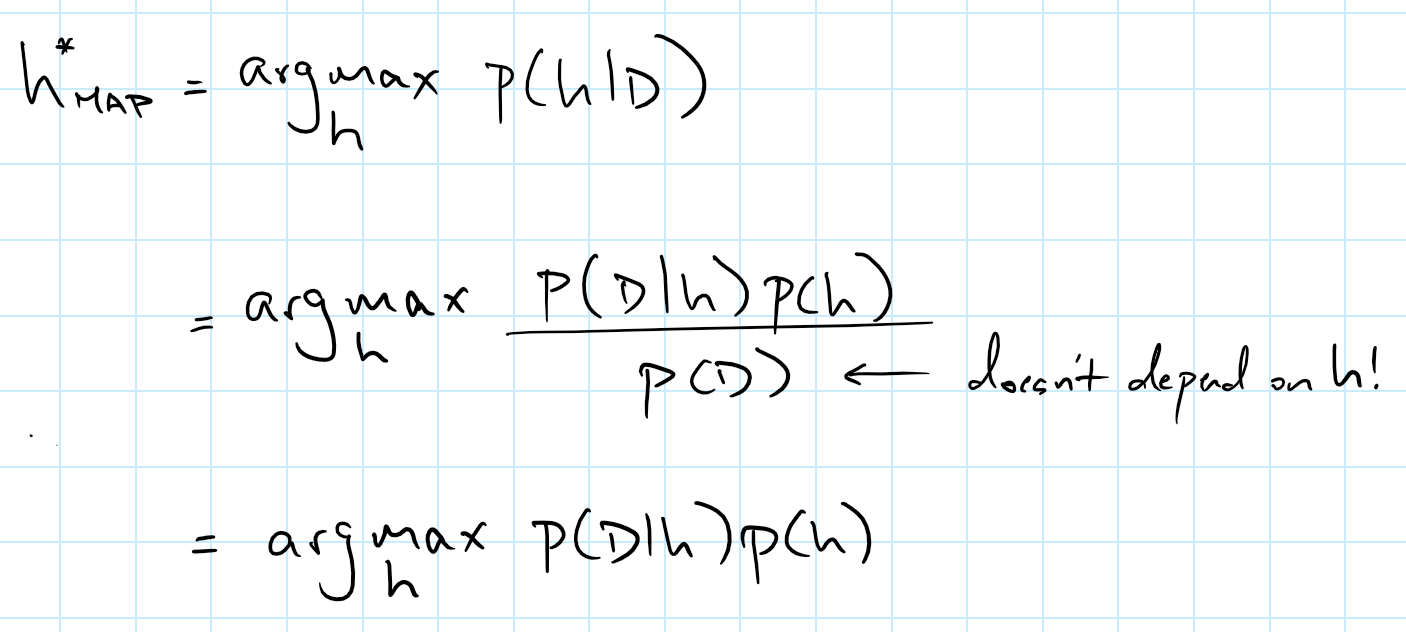
但分母是个问题：



一般来说，不可能计算这个积分。对于这个硬币的例子，如果你使用非常具体的发行版，也就是你可以回避这个问题，但是我们将在其他文章中讨论这个问题。

#### 最大后验概率

但是等等，我们可以说p（h | D）的一些智能的东西，没有标准化常数p（D）！也就是说，标准化常数不会改变分布的相对大小，我们可以在不做积分的情况下找到模式：

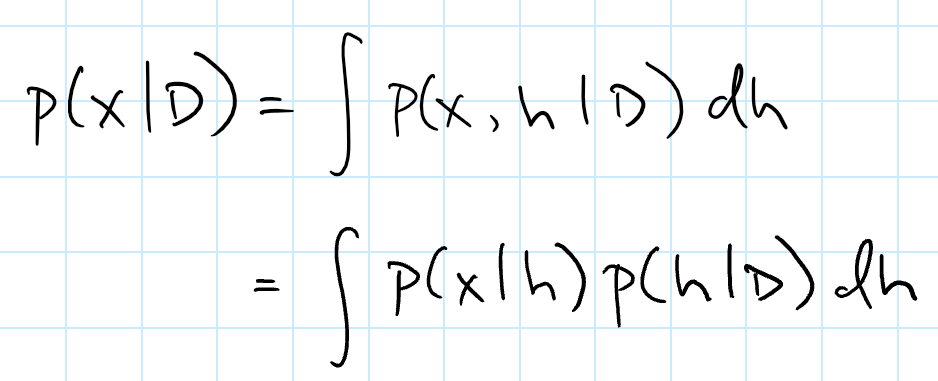


这被称为最大后验概率（MAP）。有很多方法可以找到h的这个特定值，例如使用。

#### 贝叶斯参数估计

在MAP中，我们通过先验结合直觉，忽略了归一化积分，得到了h在后验分布模式下的点估计。

但是如果我们尝试用近似方法求积分呢？如果我们作出通常的假设，我们可以使用这样一个事实，未来的数据x在条件上独立于给定h的观测数据D：



与其使用与后验p（h | D）模式对应的h的单个值来计算p（x | h），这是我们考虑h的所有可能后验值（称为贝叶斯参数估计）的更“严格”的方法。

注意这真的很漂亮。关于概率分布，你可能会关心两件事：

* 推论：给定已知参数的联合分布，通过边缘化和调节其他变量来估计变量子集上的分布。
* 参数估计：根据数据估计概率分布的未知参数。

贝叶斯参数估计将这两个任务作为一枚硬币的两面：

估计一组变量上分布的参数是对原始变量和参数上更大的元分布的推断。

Of course, actually doing this requires us to compute these difficult integrals, which we’ll have to do with approximate methods such as MCMC or variational inference. I’m planning to start a series on variational inference soon, so keep an eye out for that!