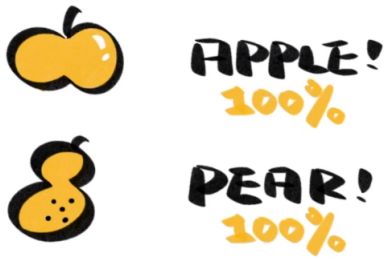
什么是EM(最大期望值算法)

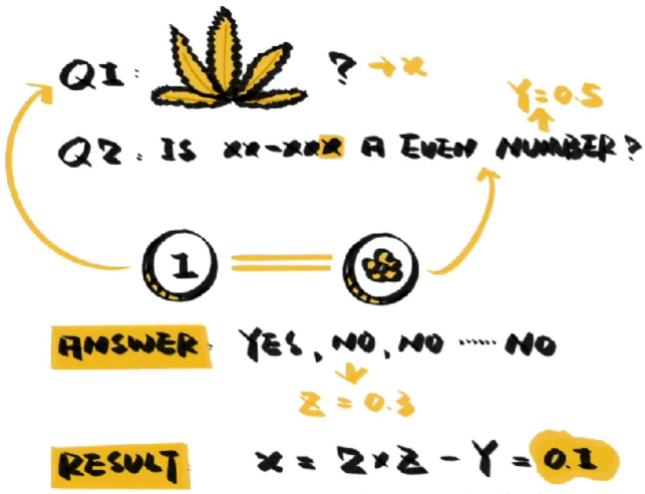
在现实生活中，苹果百分百是苹果，梨百分白是梨。



生活中还有很多事物是概率分布，比如有多少人结了婚，又有多少人有工作，

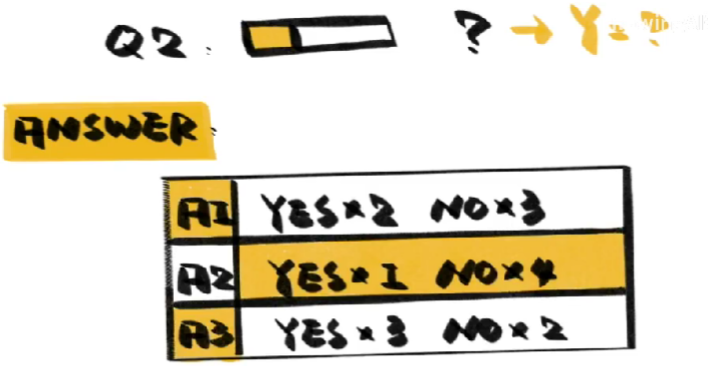
如果我们想要调查人群中吸大麻者的比例呢？敏感问题很难得到真实回答，这时就可以利用概率让调查匿名化，在“你吸大麻吗？”这个问题之外，再提出一个问题，“你的手机尾号是偶数吗？”，然后邀请参与者投掷硬币，正面回答问题1，反面则回答问题2。

调查以电话进行，手机尾号是偶数的比例已经确定，只要调查样本足够多，抛硬币能让回答问题1和问题2的人接近相等，在不知道回答的是哪个问题的情况下，我们依然轻松推测出了人群中吸大麻者的比例，这就是概率的魔法。



现在让我们将问题2稍做变更，将“手机尾号是偶数”替换成“你吸烟吗”这样的未知概率事件，我们还能推断出吸大麻者的概率吗?

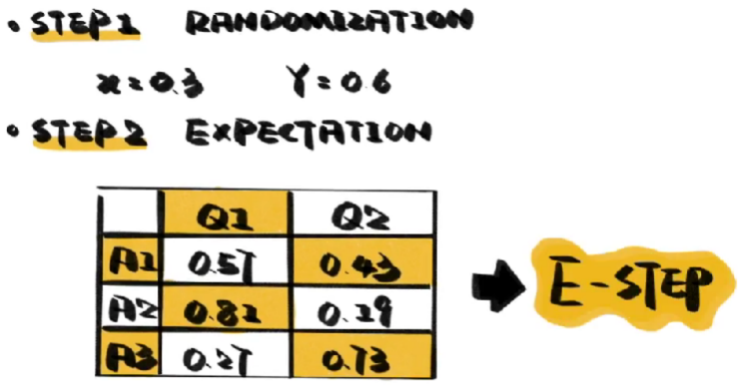
答案依然是能，只不过这次我们改变了调查方法，向每五个人发放同一个问题邀请他们回答，不记录问题是什么，只记录他们的答案，在保证匿名性的同时，我们得到了一些不知归属的成答案。



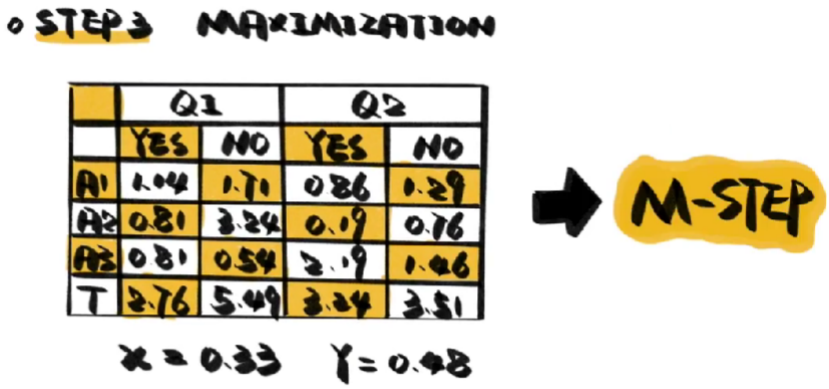
接下来就轮到**EM算法**

EM算法的步骤

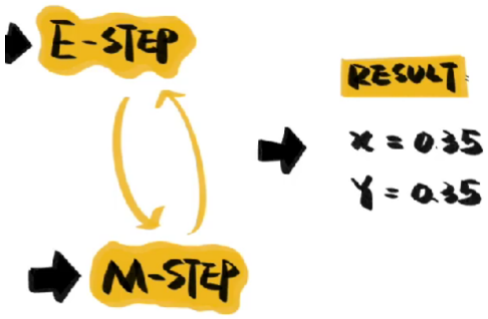
1. 随机化，不知道答案属于拿一个问题，就无法推测吸烟和吸大麻者的比例，不知道这两个比例，就无法推测答案属于哪一个问题，既然如此，我们就随机为吸烟者和吸大麻者赋予一个数值
2. 接下来用这些数值反过来去推测这些成组的答案属于两个问题的可能性，这一步是在估算未知变量也就是问题归属的期望，因此被称为E步



3. 然后我们用这个可能性，反过来估算吸烟者和吸大麻者的概率，由于这个概率是可能性最大的，因此被称为M步。



4. 接下来重复第二步，用新的概率推算答案属于两个问题的可能性，再用可能性反过来推测概率，循环往复，直到估算出较为稳定的数值就停止



就这样，我们推算出了人群中吸烟者和吸大麻者大致的概率，这个过程是不是有点熟悉，K-means的步骤同样是：1.随机赋值、2.反复对照、3.不断逼近。事实上K-means就是EM算法的一个特例，K-means的目标是获得两个中心坐标，从而将梨和苹果作为两种事物进行区分。EM算法则能找到样本的分布规律，在聚类的同时，帮我们找到更多的梨和苹果。

A SPECIAL EM