1. $\vec{v}=\begin{bmatrix}-1\\-2\\-2\\-2\\10\\-5\\0\end{bmatrix}+\gamma\begin{bmatrix}0.9&0.1&0&0&0&0&0\\0.5&0&0.5&0&0&0&0\\0&0&0&0.8&0&0&0.2\\0&0&0&0&0.6&0.4&0\\0&0&0&0&0&0&1\\0&0.2&0.4&0.4&0&0&0\\0&0&0&0&0&0&1\end{bmatrix}\vec{v}$ gamma = 0.5 时，$\vec{v}=\begin{bmatrix}-2.09\\-3.04\\-2.06\\-0.15\\10\\-5.75\\0\end{bmatrix}$

2.

    1. 先看 v 函数。在上式中，在状态 s 的情况下做出行为 a 的概率乘以 R 函数在状态 s 和行为 a 的情况下获得的奖励，并且累加所有可能行为 a，得到的即为下式中在下一时刻 R 的期望值。并且，在行为 a 的情况下从状态 s 跳转到下一状态的概率乘以下一状态的 v 值，并且对所有可能的下一状态和可能的行为累加，得到的即为下式中的下一状态的 v 值的期望。
    2. 再看 q 函数。R 函数在状态 s，行为 a 的情况下获得的奖励，即为下式中的 R 的期望。并且，在行为 a 下从状态 s 跳转到下一状态的概率乘以在下一状态做出下一行为的概率乘以下一状态做出下一行为的 q 值，对所有可能的下一行为和下一状态累加，得到的即为下式中下一状态下一行为的 q 值的期望。
    3. 上一种形式中明确给出了累加的方式，适合直接求解。下一种形式中明确给出了上一状态和下一状态的关系，适合迭代求解。

3. 不一定，因为 $q(s_t,a_t)=R(s_t,a_t)+\gamma\sum_{s_{t+1}\in S}P^{a_t}_{s_ts_{t+1}}v(s_{t+1})$ ，所以在明确知道奖励函数和转移矩阵的情况下，可以使用 v 函数直接推出 q 函数，从而选择最佳动作。否则不可以。

4.

    1. gamma 为 0.5 的模拟结果：
        - 's1': -1.2974452474857627
        - 's2': -1.9101540189614938
        - 's3': -0.3336104059860154
        - 's4': 2.6550466099276573
    2. gamma 为 1 的模拟结果：
        - 's1': -4.5894
        - 's2': -3.675
        - 's3': 0.3183

- 's4': 2.664

3. gamma 为 0.5 的策略遍历：
    - Choices: {'s1': 0, 's2': 0, 's3': 0, 's4': 0} Result: {'s1': -2.0, 's2': -2.0, 's3':3.0, 's4': 10.0}
    - Choices: {'s1': 0, 's2': 0, 's3': 0, 's4': 1} Result: {'s1': -2.0, 's2': -2.0, 's3':-6.046875, 's4': -9.0234375}
    - Choices: {'s1': 0, 's2': 0, 's3': 1, 's4': 0} Result: {'s1': -2.0, 's2': -2.0, 's3': 0.0, 's4': 10.0}
    - Choices: {'s1': 0, 's2': 0, 's3': 1, 's4': 1} Result: {'s1': -2.0, 's2': -2.0, 's3': 0.0, 's4': -8.0}
    - Choices: {'s1': 0, 's2': 1, 's3': 0, 's4': 0} Result: {'s1': -2.0, 's2': -0.5, 's3': 3.0, 's4': 10.0}
    - Choices: {'s1': 0, 's2': 1, 's3': 0, 's4': 1} Result: {'s1': -2.0, 's2': -5.390442274871127, 's3': -5.745668433171499, 's4': -7.4448171163638746}
    - Choices: {'s1': 0, 's2': 1, 's3': 1, 's4': 0} Result: {'s1': -2.0, 's2': -2.0, 's3': 0.0, 's4': 10.0}
    - Choices: {'s1': 0, 's2': 1, 's3': 1, 's4': 1} Result: {'s1': -2.0, 's2': -2.0, 's3': 0.0, 's4': -5.0}
    - Choices: {'s1': 1, 's2': 0, 's3': 0, 's4': 0} Result: {'s1': -0.6666666666666666, 's2': -1.3333333333333333, 's3': 3.0, 's4': 10.0}
    - Choices: {'s1': 1, 's2': 0, 's3': 0, 's4': 1} Result: {'s1': -0.6666666666666666, 's2': -1.3333333333333333, 's3': -6.781276737119905, 's4': -5.666666666666666}
    - Choices: {'s1': 1, 's2': 0, 's3': 1, 's4': 0} Result: {'s1': -0.6666666666666666, 's2': -1.3333333333333333, 's3': 0.0, 's4': 10.0}
    - Choices: {'s1': 1, 's2': 0, 's3': 1, 's4': 1} Result: {'s1': -0.6666666666666666, 's2': -1.3333333333333333, 's3': 0.0, 's4': -8.75}
    - Choices: {'s1': 1, 's2': 1, 's3': 0, 's4': 0} Result: {'s1': -0.25, 's2': -0.5, 's3': 3.0, 's4': 10.0}
    - Choices: {'s1': 1, 's2': 1, 's3': 0, 's4': 1} Result: {'s1': -2.5938948922248417, 's2': -5.249038639878324, 's3': -6.093712750120364, 's4': -9.688871528723556}
    - Choices: {'s1': 1, 's2': 1, 's3': 1, 's4': 0} Result: {'s1': -1.0, 's2': -2.0, 's3': 0.0, 's4': 10.0}
    - Choices: {'s1': 1, 's2': 1, 's3': 1, 's4': 1} Result: {'s1': -1.0, 's2': -2.0, 's3': 0.0, 's4': -8.75}
4. gamma 为 1 的策略遍历：

- Choices: {'s1': 0, 's2': 0, 's3': 0, 's4': 0} Result: {'s1': -100.0, 's2': -100.0, 's3': 8.0, 's4': 10.0}
- Choices: {'s1': 0, 's2': 0, 's3': 0, 's4': 1} Result: {'s1': -100.0, 's2': -100.0, 's3': -145.0, 's4': -113.0}
- Choices: {'s1': 0, 's2': 0, 's3': 1, 's4': 0} Result: {'s1': -100.0, 's2': -100.0, 's3': 0.0, 's4': 10.0}
- Choices: {'s1': 0, 's2': 0, 's3': 1, 's4': 1} Result: {'s1': -100.0, 's2': -100.0, 's3': 0.0, 's4': -104.0}
- Choices: {'s1': 0, 's2': 1, 's3': 0, 's4': 0} Result: {'s1': -100.0, 's2': 6.0, 's3': 8.0, 's4': 10.0}
- Choices: {'s1': 0, 's2': 1, 's3': 0, 's4': 1} Result: {'s1': -100.0, 's2': -356.0, 's3': -359.0, 's4': -365.0}
- Choices: {'s1': 0, 's2': 1, 's3': 1, 's4': 0} Result: {'s1': -100.0, 's2': -2.0, 's3': 0.0, 's4': 10.0}
- Choices: {'s1': 0, 's2': 1, 's3': 1, 's4': 1} Result: {'s1': -100.0, 's2': -2.0, 's3': 0.0, 's4': -5.0}
- Choices: {'s1': 1, 's2': 0, 's3': 0, 's4': 0} Result: {'s1': -50.0, 's2': -50.0, 's3': 8.0, 's4': 10.0}
- Choices: {'s1': 1, 's2': 0, 's3': 0, 's4': 1} Result: {'s1': -50.0, 's2': -50.0, 's3': -56.0, 's4': -64.0}
- Choices: {'s1': 1, 's2': 0, 's3': 1, 's4': 0} Result: {'s1': -50.0, 's2': -50.0, 's3': 0.0, 's4': 10.0}
- Choices: {'s1': 1, 's2': 0, 's3': 1, 's4': 1} Result: {'s1': -50.0, 's2': -50.0, 's3': 0.0, 's4': -77.0}
- Choices: {'s1': 1, 's2': 1, 's3': 0, 's4': 0} Result: {'s1': 6.0, 's2': 6.0, 's3': 8.0, 's4': 10.0}
- Choices: {'s1': 1, 's2': 1, 's3': 0, 's4': 1} Result: {'s1': -363.0, 's2': -359.0, 's3': -362.0, 's4': -377.0}
- Choices: {'s1': 1, 's2': 1, 's3': 1, 's4': 0} Result: {'s1': -2.0, 's2': -2.0, 's3':0.0, 's4': 10.0}
- Choices: {'s1': 1, 's2': 1, 's3': 1, 's4': 1} Result: {'s1': -2.0, 's2': -2.0, 's3':0.0, 's4': -5.0}

5. 最佳策略均为 quit，study，study，review