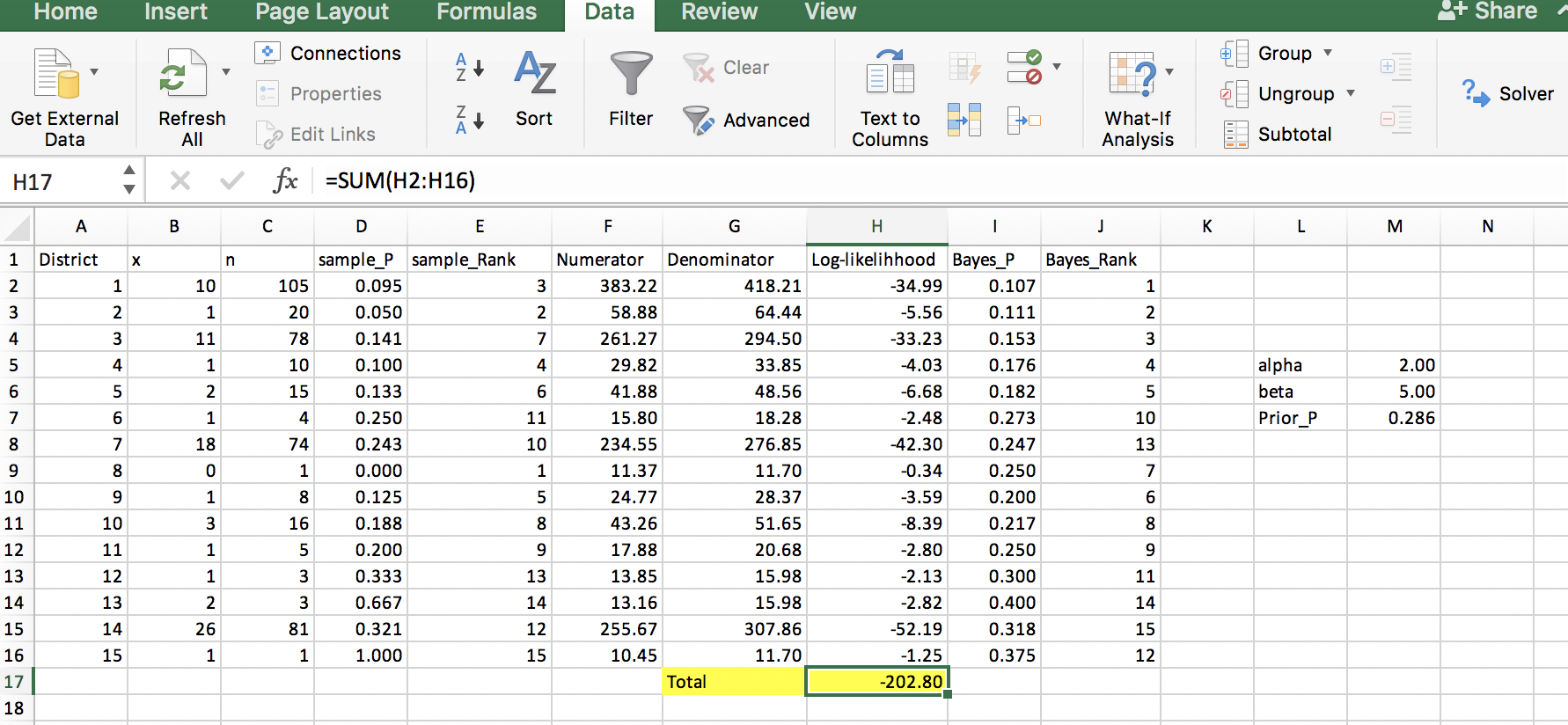
### Empirical Bayes Smoothing of Prevalence by Beta-binomial Model

Suppose in 15 districts, we interviewed 424 participants of whom 79 reported alcohol use in the past week. By stratifying at the district level, prevalence estimates vary from zero to one (100%). Even if we exclude 0% and 100%, the range will be from 5% and 66%.

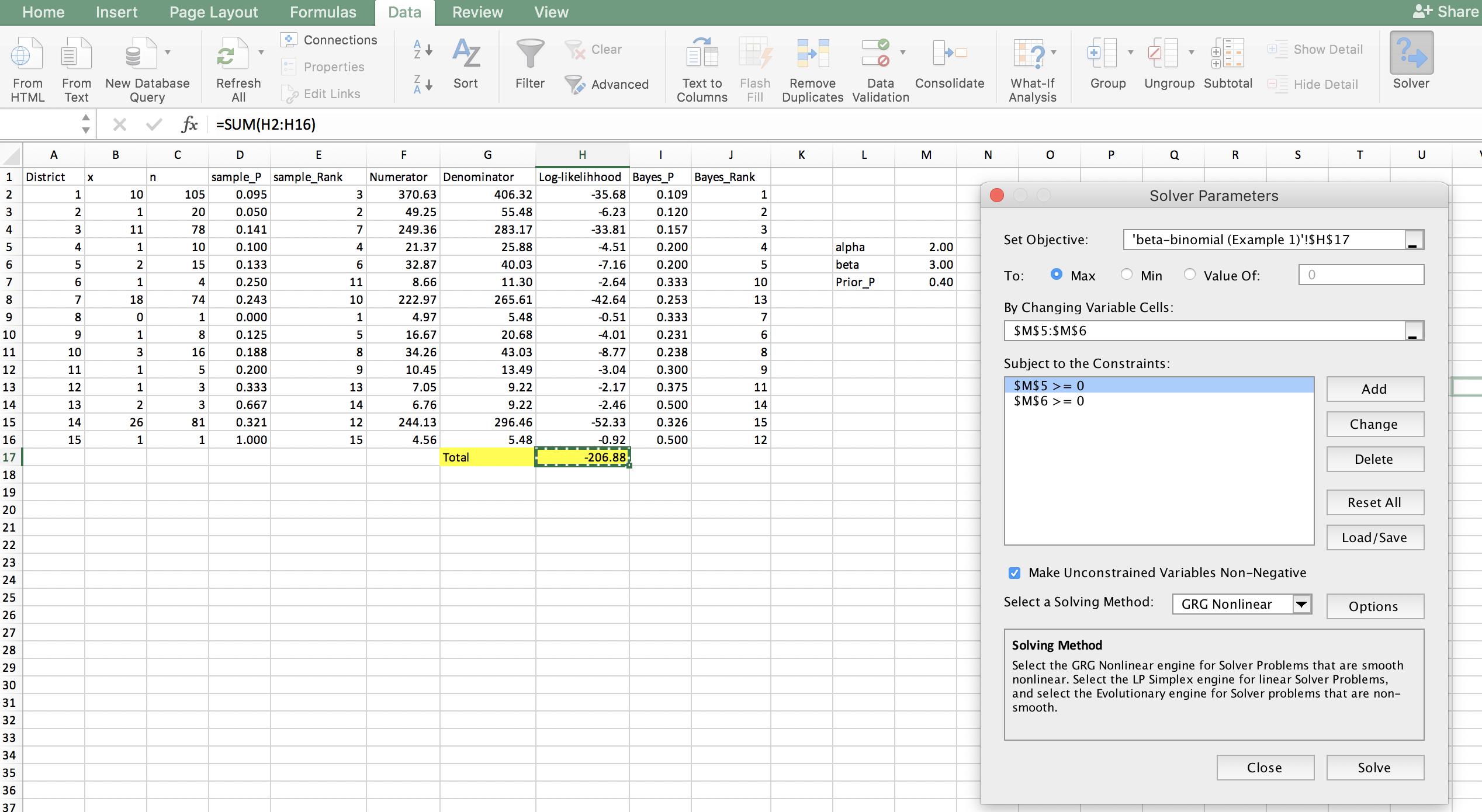
To apply Bayesian framework for this analysis, we should calculate the log-likelihood function based on a beta-binomial distribution. Based on Equation 4.3, we should substitute appropriate values of 𝜶, 𝜷, and P that maximize the log-likelihood function. Keep in mind that . All we need to do is to start with an initial guess of these 𝜶 and 𝜷 parameters and to calculate the log-likelihood function. Let’s start with 𝜶=2, 𝜷=5, which consequently gives P=0.286. These values are entered into M5 to M7 cells in an Excel sheet (Figure 4.1).



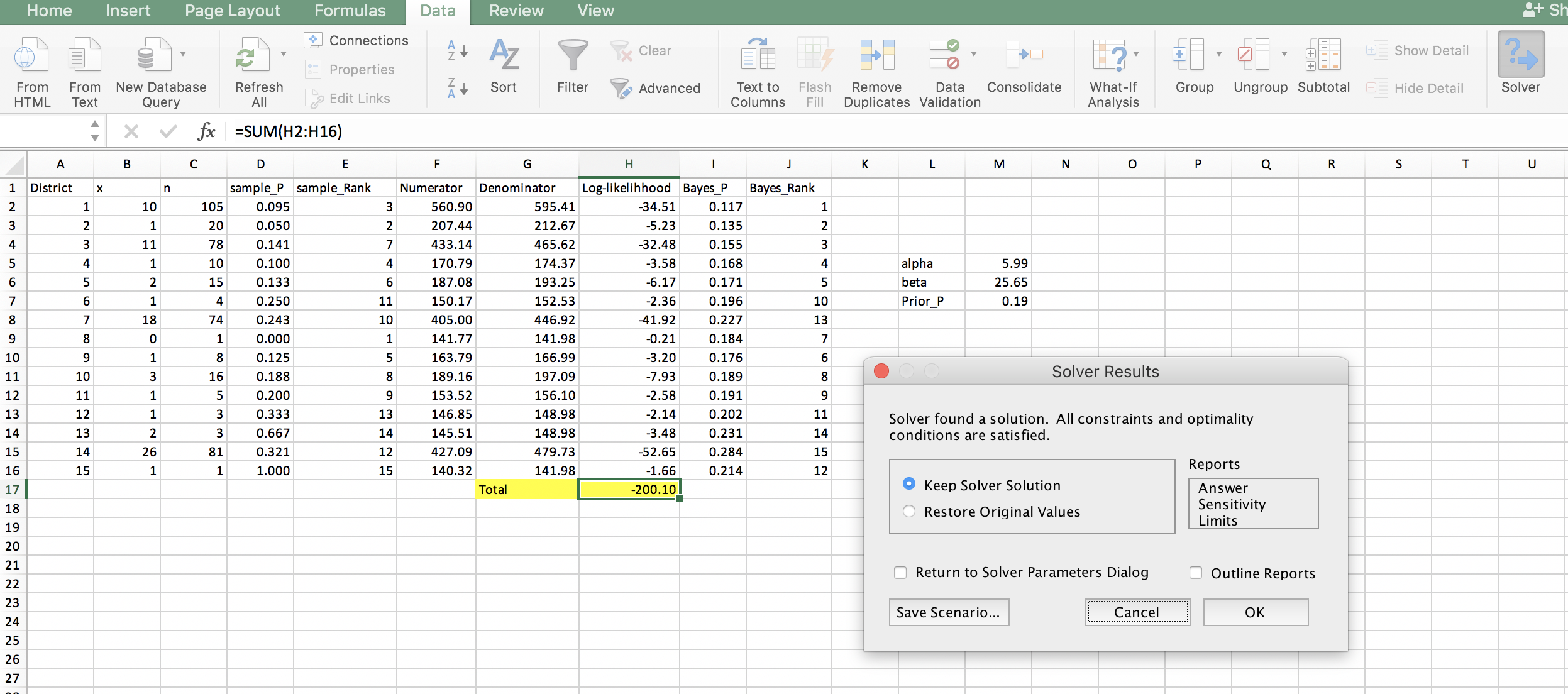
**Figure 4.1: Calculation of nominator and denominator of the beta-binomial likelihood function**

The function GAMMALN returns the logarithm of a gamma function. This function is used to calculate the numerator (Column F) and denominator (Column G) of Equation 4.2 separately. The difference of these two columns was calculated (column H). Summing values in column H gives the log-likelihood function value corresponding to the initial values (cell H17).

Now, we should maximize cell H17 by changing 𝜶 and 𝜷 (M5 and M6). In the data tab, there is a function called Solver. Figure 4.2 shows how to use Solver to address our optimization problem. We call Solver to maximize cell H17, by changing cells M5 and M7. The Solver function estimates parameters at 𝜶=5.99, and 𝜷 =25.65. Using these values, it is now easy to calculate the posterior probabilities. Remember that.



**Figure 4.2. Optimization of initial values – Solver Parameters Window**



**Figure 4.3. Solver Results**

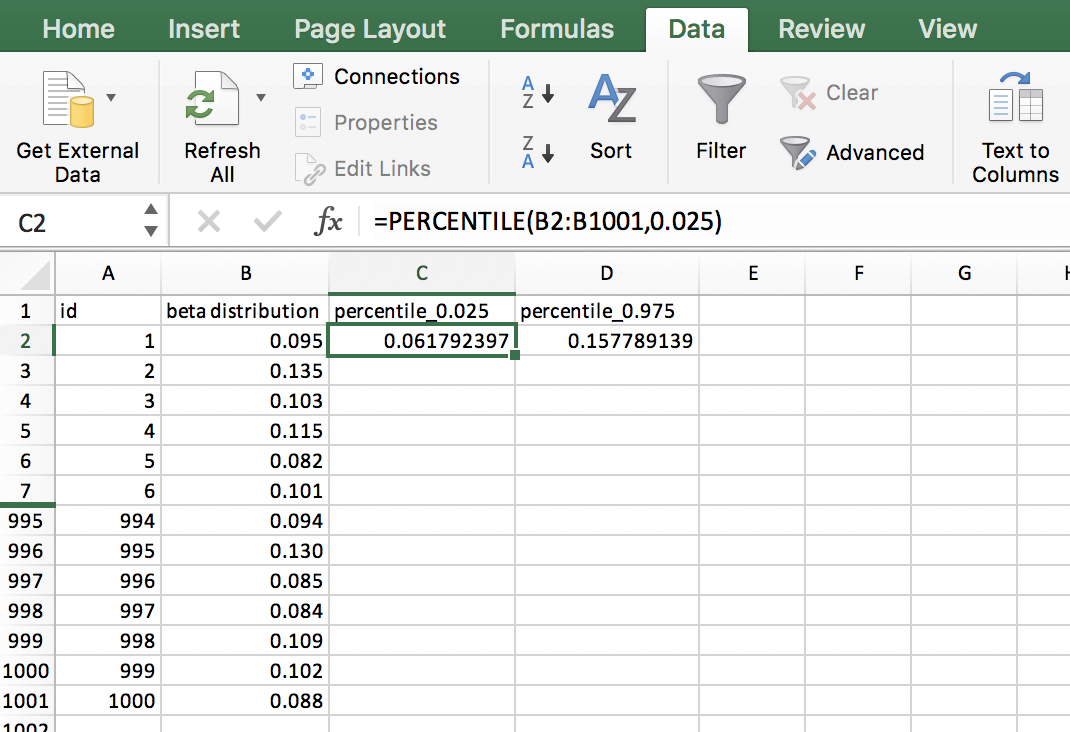
For example, in district 2, the observed (from data) prevalence for alcohol is 0.05, but the posterior prevalence is .

Figure 4.3 and Table 4.1 show the posterior probability estimates and the rank of districts (based on the prevalence of alcohol use). While sample statistics (observed prevalence) varied from 0 to 1, smoothed prevalence ranged from 0.12 to 0.28. Change in the estimations for districts with smaller sample size was more remarkable. As you noted, the ranking of districts also changed. For example, the rank of district 8 changed from first to seventh.

**Table 4.1: Comparison of sample and Bayesian prevalence and ranking of districts**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| District\_No | x | n | Sample\_P | Sample\_Rank | Bayesian\_P | Bayesian\_Rank |
| 1 | 10 | 105 | 0.095 | 3 | 0.117 | 1 |
| 2 | 1 | 20 | 0.050 | 2 | 0.135 | 2 |
| 3 | 11 | 78 | 0.141 | 7 | 0.155 | 3 |
| 4 | 1 | 10 | 0.100 | 4 | 0.168 | 4 |
| 5 | 2 | 15 | 0.133 | 6 | 0.171 | 5 |
| 6 | 1 | 4 | 0.250 | 11 | 0.196 | 10 |
| 7 | 18 | 74 | 0.243 | 10 | 0.227 | 13 |
| 8 | 0 | 1 | 0.000 | 1 | 0.184 | 7 |
| 9 | 1 | 8 | 0.125 | 5 | 0.176 | 6 |
| 10 | 3 | 16 | 0.188 | 8 | 0.189 | 8 |
| 11 | 1 | 5 | 0.200 | 9 | 0.191 | 9 |
| 12 | 1 | 3 | 0.333 | 13 | 0.202 | 11 |
| 13 | 2 | 3 | 0.667 | 14 | 0.231 | 14 |
| 14 | 26 | 81 | 0.321 | 12 | 0.284 | 15 |
| 15 | 1 | 1 | 1.000 | 15 | 0.214 | 12 |
| Total | 79 | 424 |  |  |  |  |

We can also use excel to calculate the Bayesian credibility interval for the posterior alcohol prevalence in each district. The posterior distribution in each district follows a beta distribution with parameters 𝜶+x and 𝜷+n-x. For example, for district 1, parameters of the posterior distribution are 𝜶 =15.99 and 𝜷 =136.64. And so, the posterior estimate for district 1 will be 12% (95% credibility interval: 6%, 15%). Figure 4.3 shows how to use Excel to generate random numbers from a beta distribution, and how to calculate the percentiles 2.5 and 97.5.



**Figure 4.3: generate 1000 random values from a beta distribution (𝜶 =15.99 and 𝜷 =136.64) and calculated the 2.5 and 97.5 percentile**

As shown in the above example, empirical Bayesian smoothing determines the posterior prevalence by a weighted combination of the local and prior prevalence. The weights are based on the sample size. When the sample size is large, it is given a large weight, and the posterior estimate leans toward the prior (data).

### Empirical Bayes Smoothing of Standard Morbidity Ratio (SMR) by Poisson-Gamma Model

The standardized morbidity ratio (SMR) uses the indirect method of standardization to compare the morbidity experience of a given area with a standard. Usually morbidity rate at national level is used as standard. This allows the morbidity experience of each district to be compared to the national average. For each district, SMR is the ratio of observed cases () to expected one (). The expected number of cases is calculated by multiplying the national (overall) prevalence by sample size at district level (Equations 4.4 and 4.5).

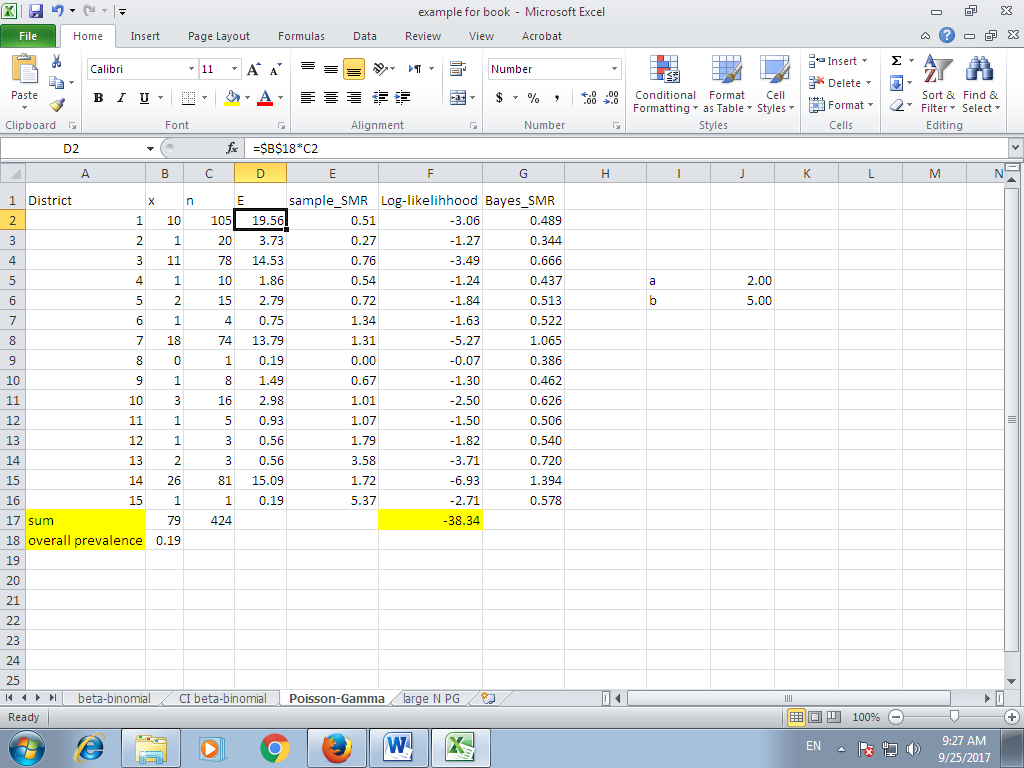
Equation 4.4:

Equation 4.5:

The interpretation of SMR is the same as that of Relative Risk (RR). SMR of one indicates that risk of outcome in a district is the same as the risk for the nation. On the other hand, SMR of five means that the risk of outcome in the district is five times that of the national average. As explained before, in the case of small area analysis, SMR estimates are likely to become very diverse.

To provide a numerical example, we will continue with our previous example. As shown in Figure 4.4, we divided summation of all positive cases to the total sample size to get the national prevalence (). Sample size of each district is multiplied to 0.19 to get expected number of positive cases (Column D). Dividing observed to the expected number of positive cases SMR is calculated (Column E). Due to small sample size in some districts, SMR values are highly diverse around one, with a range of 0 to 5.37.

Figure 4.4: Calculation of crude SMR



To smooth the SMR estimates, we assume that the number of positive cases in district i follow Poisson distribution with mean of . The conjugate for Poisson distribution is . Mean and variance of Gamma distribution is:

Equation 4.6:

In this case, the posterior distribution is distribution. The mean of this posterior distribution is used to smooth district level SMR:

Equation 4.7:

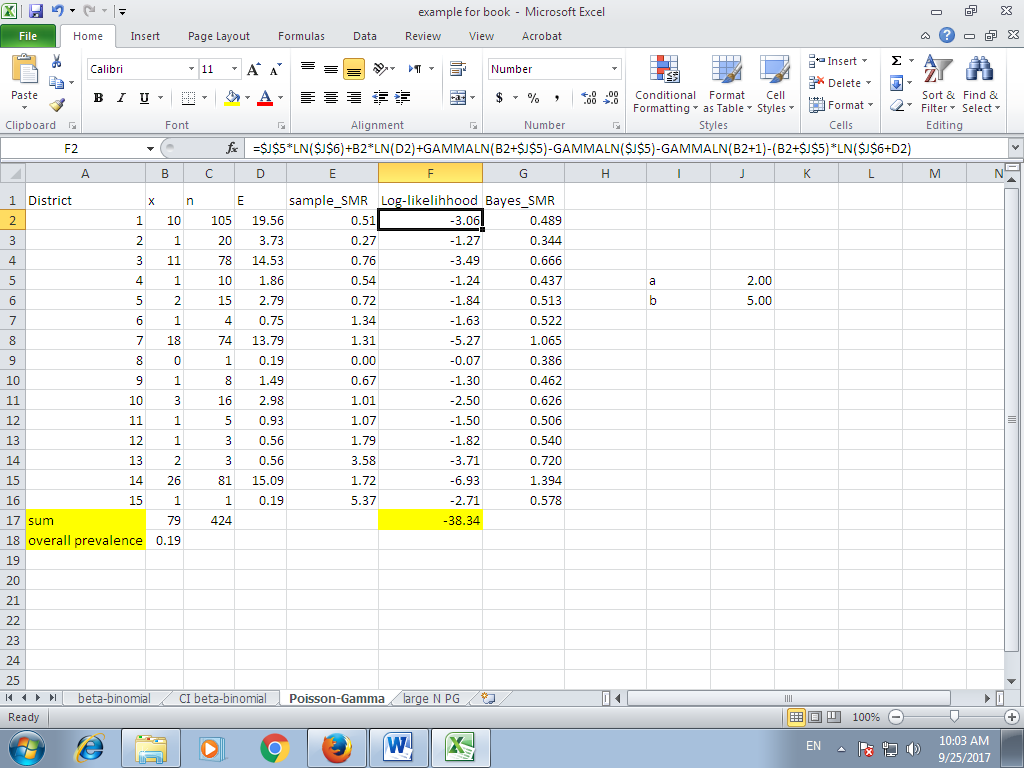
To be able to implement these analyses in Excel, we should construct the likelihood function based on Poisson-Gamma distribution, and maximize the function with respect to hyper-parameters a and b (Equations 4.8 and 4.9).

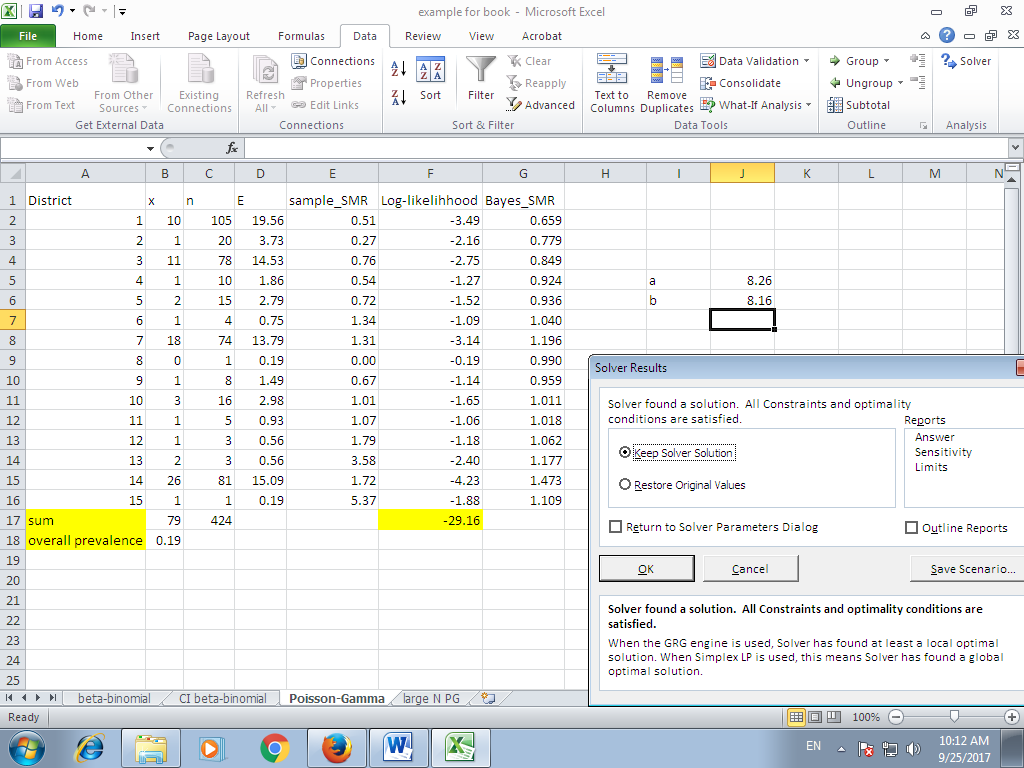
Equation 4.8:

Equation 4.9:

]All we need to do is to start with an initial guess of a and b parameters and to calculate the log-likelihood function. Let’s start with a=2, b=5. These values are entered into J5 to J7 cells in an Excel sheet (Figure 4.4). Using these initial values, log-likelihood function is constructed (Figure 4.5 top panel, Column F). We should now use Solver and maximize cell F17 by changing J6 and J7 cells (Figure 4.5 bottom panel).

Figure 4.5: Calculation of the Gamma-Poisson log-likelihood function (top panel) and maximization of this function (bottom panel)





Based on Solver solution, best estimates for both hyper-parameters are 8.26 (a) and 8.16 (b). Using these values, SMR values are updated (Column G, Figure 4.6). Now, the variance is much lower, and the minimum and maximum values are 0.65 and 1.77.

Figure 4.6: Calculation of the posterior SMR values

