1

a) $E[f(x)] = \sum_{x \in X} x \; p(x)$

$= f(a)p(a) + f(b)\,p(b)\; f(c)\,p(c)$

$= 10 \cdot 0.1 + 5 \cdot 0.2 + 10/7 \cdot 0.7$

$= 3$

b) $E[1/p(x)] = \sum_{x \in X} 1/p(x) \cdot px = \sum_{x \in X} 1$

$= 3$

c) For an arbitrary pmf, this expectation value will equal to the size of the outcome space

2)

a) $E[X] = \sum_{i=1}^{m} a_i X_i P(X_i) \qquad X_i = N(\mu_i, \Sigma_i)$

$$p(x_i|\mu_i, \Sigma_i) = \frac{1}{\sqrt{2\pi\Sigma_i^2}} e^{-\frac{1}{2\Sigma_i^2}(x-\mu_i)^2}$$

$$E[x] = \sum_{i=1}^{m} a_i X_i \frac{1}{\sqrt{2\pi\Sigma_i^2}} e^{-\frac{1}{2\Sigma_i^2}(X_i-\mu_i)^2}$$

dimension of x is $d \times 1$

b)

$$cov[X,X] = E[XX^T] - E[X]E[X]^T$$

$$E[XX^T] = \sum_{i=1}^{m} a_i X_i X_i^T p(X_i)$$

$$= \sum_{i=1}^{m} a_i X_i X_i^T \underbrace{\frac{1}{\sqrt{2\pi\Sigma_i^2}} e^{-\frac{1}{2\Sigma_i^2}(X_i-\mu_i)^2}}_{S}$$

$$cov[X] = \sum_{i=1}^{m} a_i X_i X_i^T S - \sum_{i=1}^{m}[a_i X_i S] \cdot \sum_{i=1}^{m}[a_i X_i^T S]$$

$cov[X]$ is $d \times d$

3)

a. With sigma = 1 each of the 3 sample sizes had a mean of close to 0. The 10 and 100 sample sizes had some variance with the means and strayed slightly away from 0 in some cases. The 1000 sample size was almost always very close to 0. With sigma = 10, each of the sample sizes had there means farther away from 0. The sample size of 10 was mostly far away from 0 and the 1000 sample was very close to 0 every time.
From this we can conclude that increasing the standard deviation will increase the error between the sample and actual mean, and increasing the sample size will decrease the error between the sample and actual mean.

b. When the variance matrix is equal to the identity matrix the Gaussian distribution will have zero correlation between the three random variables. They are independent.
c. This variance matrix has the X and Z random variables directly correlated with each other and the Y variable is independent of the other two.

4)

a) $f(\lambda) = \theta e^{-\theta \lambda}$

$= \frac{1}{2} e^{-\frac{x}{2}}$

$\ln f(x) = \ln \frac{1}{2} e^{-\frac{\lambda}{2}}$

$0 = -\frac{\lambda}{2} \cdot \ln\left(\frac{1}{2}\right)$

$0 = -\frac{\lambda}{2} \cdot -0.693$

$= \lambda \cdot 0.3465$

$\boxed{\lambda = 2.885}$

b) $\lambda_{ML} = \underset{\lambda \in (0,\infty)}{\operatorname{argmax}} \; p(D|\lambda)$

$$p(D|\lambda) = \prod_{i=1}^{n} p(x_i|\lambda)$$

$$= \frac{\lambda^{\sum_{i=1}^{n} x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

$$\ln \lambda_{ML} = \ln p(D|\lambda)$$

$$= \ln \lambda \sum_{i=1}^{n} x_i - n\lambda - \sum_{i=1}^{n} \ln(x_i)!$$

$$\frac{d}{d\lambda} = \frac{1}{\lambda} \sum_{i=1}^{n} x_i - n$$

$$0 = \frac{1}{\lambda} 79 - 9$$

$$9\lambda = 79$$

$$\lambda = \frac{79}{9} = 8.78$$

c)

$$\lambda_{map} = \underset{\lambda \in (0,\infty)}{\arg\max} \{ p(D|f) p(f) \}$$

$$\ln \lambda_{map} = \ln p(D|f) p(f)$$

$$= \ln p(D|f) + \ln p(f)$$

$$= \ln \frac{\lambda^{x} e^{-\lambda}}{x!} + \ln \theta e^{-\theta\lambda}$$

$$= \ln \frac{\lambda^{x} e^{-\lambda}}{x!} + \ln \theta - \theta\lambda$$

$$\frac{d}{d\lambda} = \underbrace{\ln \frac{\lambda^{x} e^{-\lambda}}{x!} \frac{d}{d\lambda}}_{} + 0 - \theta$$

$$= \frac{1}{\lambda} 79 - 9 - \frac{1}{2}$$

$$9.5 = \frac{79}{\lambda}$$

$$\lambda = \frac{79}{9.5}$$

$$\boxed{\lambda = 8.316}$$

d. To predict the number of accidents happening tomorrow we would use the λ that we got from our Maximum likelihood estimate. The λ that we got is the estimated mean of the distribution so it is a good estimate of the number of accidents that can happen in a day. Another thing to consider is the variance of the accidents for the nine days that we have data for. This variance can be used with the estimated mean to get an estimate. The λ from the MAP estimate would seem like a better estimate because of our previous knowledge of this scenario. Both of the mean estimates are good to use because they are similar and when more data is accumulated they become closer together.

e. The purpose of a prior is to give estimated distribution some structure using past knowledge of the experiment. When not much data is viewed the prior will help estimate the data, potentially making it more accurate. When lots of data is used the prior makes less of an impact because the data will shape our distribution to equal the actual distribution.

f. To show that the estimated amount of accidents should decrease, θ should be increased to a value greater than 1. A value of 2 reduces the amount of accidents by ~4.

5) a)

Bern PMF: $p(w) = a^{w}(1-a)^{1-w}$

sunny/not $p(s|a) = a^{s}(1-a)^{1-s}$

$$= \begin{cases} a & \text{if } s = 1 \quad \text{sunny} \\ 1-a & \text{if } s = 0 \quad \text{not sunny} \end{cases}$$

tall/not $p(t|s,b)p(s|a) = b^{t}(1-b)^{1-t}\, p(s|a)$

$$= b^{t}(1-b)^{1-t}\, a^{s}(1-a)^{1-s}$$

$$p(t) = b^{t}(1-b)^{1-t}\, a^{s}(1-a)^{1-s}$$

$$\ln p(t) = t\ln b + (1-t)\ln(1-b) + s\ln a + (1-s)\ln(1-a)$$

$$\frac{\partial}{\partial a} = \frac{s}{a} - \frac{(1-s)}{(1-a)} \Rightarrow 0 = \frac{s}{a} - \frac{(1-s)}{(1-a)}$$

$$= \frac{s-1}{a-1} + \frac{(s-1)}{(a-1)}$$

$$0 = \frac{2(s-1)}{a-1}$$

$$a-1 = 2s-2$$

$$\boxed{a = 2s-1}$$

$$\frac{d}{db} = \frac{t}{b} - \frac{(1-t)}{(1-b)} \Rightarrow 0 = \frac{t}{b} - \frac{(1-t)}{(1-b)}$$

$$= \frac{t-1}{b-1} + \frac{(t-1)}{(b-1)}$$

$$= \frac{2(t-1)}{b-1}$$

$$b-1 = 2t-2$$

$$\boxed{b = 2t-1}$$

b) sunny $\Rightarrow$ $S=1$

$$p(t \mid S=1) = b^{t}(1-b)^{1-t} \; a^{S}(1-a)^{1-S}$$

$$= b^{t}(1-b)^{1-t} \; a$$

From our observations we will be able to estimate the values of a and b. The resulting estimations of whether the table is open or not are represented by:

$$p(t=1 \mid S=1) = a \cdot b$$

$$p(t=0 \mid S=1) = a \cdot (1-b)$$

c) To account for the time of day another probability distribution would need to be incorporated into this ML estimate. This distribution would return A Є { a_m, a_a, a_n} what would correspond to how the time of day effects the availability of the table. This would also affect the probability of how if its sunny or not effects the availability. When A = a_n it will always be dark out. There will be some correlation between these distributions. 1