# CSE 312: Foundations of Computing II

## Section 8: Maximum Likelihood and more Solutions

## 1. Review of Main Concepts

(a) **Realization/Sample**: A realization/sample $x$ of a random variable $X$ is the value that is actually observed.

(b) **Likelihood**: Let $x_1, \dots x_n$ be iid realizations from probability mass function $p_X(\mathrm{x}\,;\theta)$ (if $X$ discrete) or density $f_X(\mathrm{x}\,;\theta)$ (if $X$ continuous), where $\theta$ is a parameter (or a vector of parameters). We define the likelihood function to be the probability of seeing the data.

If $X$ is discrete:

$$L\left(x_1, \dots, x_n \mid \theta\right) = \prod_{i=1}^{n} p_X\left(x_i \mid \theta\right)$$

If $X$ is continuous:

$$L\left(x_1, \dots, x_n \mid \theta\right) = \prod_{i=1}^{n} f_X\left(x_i \mid \theta\right)$$

(c) **Maximum Likelihood Estimator (MLE)**: We denote the MLE of $\theta$ as $\hat{\theta}_{\mathsf{MLE}}$ or simply $\hat{\theta}$, the parameter (or vector of parameters) that maximizes the likelihood function (probability of seeing the data).

$$\hat{\theta}_{\mathsf{MLE}} = \arg\max_{\theta} L\left(x_1, \dots, x_n \mid \theta\right) = \arg\max_{\theta} \ln L\left(x_1, \dots, x_n \mid \theta\right)$$

(d) **Log-Likelihood**: We define the log-likelihood as the natural logarithm of the likelihood function. Since the logarithm is a strictly increasing function, the value of $\theta$ that maximizes the likelihood will be exactly the same as the value that maximizes the log-likelihood.

If $X$ is discrete:

$$\ln L\left(x_1, \dots, x_n \mid \theta\right) = \sum_{i=1}^{n} \ln p_X\left(x_i \mid \theta\right)$$

If $X$ is continuous:

$$\ln L\left(x_1, \dots, x_n \mid \theta\right) = \sum_{i=1}^{n} \ln f_X\left(x_i \mid \theta\right)$$

(e) **Bias**: The bias of an estimator $\hat{\theta}$ for a true parameter $\theta$ is defined as $\text{Bias}\left(\hat{\theta}, \theta\right) = \mathbb{E}[\hat{\theta}] - \theta$. An estimator $\hat{\theta}$ of $\theta$ is unbiased iff $\text{Bias}\left(\hat{\theta}, \theta\right) = 0$, or equivalently $\mathbb{E}[\hat{\theta}] = \theta$.

(f) **Steps to find the maximum likelihood estimator, $\hat{\theta}$:**

(a) Find the likelihood and log-likelihood of the data.

(b) Take the derivative of the log-likelihood and set it to $0$ to find a candidate for the MLE, $\hat{\theta}$.

(c) Take the second derivative and show that $\hat{\theta}$ indeed is a maximizer, that $\frac{\partial^2 L}{\partial \theta^2} < 0$ at $\hat{\theta}$. Also ensure that it is the global maximizer: check points of non-differentiability and boundary values.

(g) **Markov's Inequality**: Let $X$ be a non-negative random variable, and $\alpha > 0$. Then, $\mathbb{P}\left(X \geq \alpha\right) \leq \frac{\mathbb{E}[X]}{\alpha}$.

(h) **Chebyshev's Inequality** (we did not cover this in class): Suppose $Y$ is a random variable with $\mathbb{E}[Y] = \mu$ and $\text{Var}(Y) = \sigma^2$. Then, for any $\alpha > 0$, $\mathbb{P}\left(|Y - \mu| \geq \alpha\right) \leq \frac{\sigma^2}{\alpha^2}$.

(i) **Chernoff Bound (for the Binomial)**: (We will not cover this in class, but it's good to know.) It's stronger than the Chebyshev bound. Suppose $X \sim \text{Binomial}(n, p)$ and $\mu = np$. Then, for any $0 < \delta < 1$,

- $\mathbb{P}\left(X \geq (1+\delta)\,\mu\right) \leq e^{-\frac{\delta^2\mu}{3}}$
- $\mathbb{P}\left(X \leq (1-\delta)\,\mu\right) \leq e^{-\frac{\delta^2\mu}{2}}$

## 2. 312 Grades

Suppose Professor Karlin loses everyones grades for 312 and decides to make it up by assigning grades randomly according to the following probability distribution, and hoping the $n$ students wont notice: give an A with probability $0.5$, a B with probability $\theta$, a C with probability $2\theta$, and an $F$ with probability $0.5 - 3\theta$. Each student is assigned a grade independently. Let $x_A$ be the number of people who received an A, $x_B$ the number of people who received a B, etc, where $x_A + x_B + x_C + x_F = n$. Find the MLE for $\theta$.

**Solution:**

The data tells us, for each student in the class, what their grade was. We begin by computing the likelihood of seeing the given data given our parameter $\theta$. Because each student is assigned a grade independently, the likelihood is equal to the product over students of the chance they got the particular grade they got, which gives us:

$$L(x|\theta) = 0.5^{x_A}\theta^{x_B}(2\theta)^{x_C}(0.5-3\theta)^{x_F}$$

From there, we just use the MLE process to get the log-likelihood, take the first derivative, set it equal to 0, and solve for $\hat{\theta}$.

$$\ln L(x|\theta) = x_A \ln(0.5) + x_B \ln(\theta) + x_C \ln(2\theta) + x_F \ln(0.5-3\theta)$$

$$\frac{\partial}{\partial\theta}\ln L(x|\theta) = \frac{x_B}{\theta} + \frac{x_C}{\theta} - \frac{3x_F}{0.5-3\theta} = 0$$

Solving yields $\hat{\theta} = \frac{x_B + x_C}{6(x_B + x_C + x_F)}$.

## 3. A Red Poisson

Suppose that $x_1, \ldots, x_n$ are i.i.d. samples from a Poisson($\theta$) random variable, where $\theta$ is unknown. Find the MLE of $\theta$.

**Solution:**

Because each Poisson RV is i.i.d., the likelihood of seeing that data is just the PMF of the Poisson distribution multiplied together for every $x_i$. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned}
L\left(x_1, \ldots, x_n \mid \theta\right) &= \prod_{i=1}^{n} e^{-\theta}\frac{\theta^{x_i}}{x_i!} \\
\ln L\left(x_1, \ldots, x_n \mid \theta\right) &= \sum_{i=1}^{n}\left[-\theta - \ln(x_i!) + x_i \ln(\theta)\right] \\
\frac{\partial}{\partial\theta}\ln L\left(x_1, \ldots, x_n \mid \theta\right) &= \sum_{i=1}^{n}\left[-1 + \frac{x_i}{\theta}\right] = 0 \\
-n + \frac{\Sigma_{i=1}^{n} x_i}{\hat{\theta}} &= 0 \\
\hat{\theta} &= \frac{\Sigma_{i=1}^{n} x_i}{n}
\end{aligned}$$

## 4. Independent Shreds, You Say?

(Covered in class.) You are given 100 independent samples $x_1, x_2, \ldots, x_{100}$ from Bernoulli($\theta$), where $\theta$ is unknown. (Each sample is either a 0 or a 1). These 100 samples sum to 30. You would like to estimate the distribution's parameter $\theta$. Give all answers to 3 significant digits.

(a) What is the maximum likelihood estimator $\hat{\theta}$ of $\theta$?

**Solution:**

Note that $\Sigma_{i\in[n]}x_i = 30$, as given in the problem spec. Therefore, there are 30 **1**s and 70 **0**s. (Note that they come in some specific order.) Therefore, we can setup $L$ as follows, because there is a $\theta$ chance of getting a 1, and a $(1-\theta)$ chance of getting a 0 and they are each i.i.d. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned}
L\left(x_1, \ldots, x_n \mid \theta\right) &= (1-\theta)^{70}\theta^{30} \\
\ln L\left(x_1, \ldots, x_n \mid \theta\right) &= 70\ln(1-\theta) + 30\ln\theta \\
\frac{\partial}{\partial\theta}\ln L\left(x_1, \ldots, x_n \mid \theta\right) &= -\frac{70}{1-\theta} + \frac{30}{\theta} = 0 \\
\frac{30}{\hat{\theta}} &= \frac{70}{1-\hat{\theta}} \\
30 - 30\hat{\theta} &= 70\hat{\theta} \\
\hat{\theta} &= \frac{30}{100}
\end{aligned}$$

(b) Is $\hat{\theta}$ an unbiased estimator of $\theta$?

**Solution:**

An estimator is unbiased if the expectation of the estimator is equal to the original parameter, i.e.: $E[\hat{\theta}] = \theta$. Setting up the expectation of our estimator and plugging it in for the generic case, we get the following, which we can then reduce with linearity of expectation:

$$\begin{aligned}
\mathbb{E}[\hat{\theta}] &= \mathbb{E}\left[\frac{1}{100}\sum_{i=1}^{100} X_i\right] \\
&= \frac{1}{100}\sum_{i=1}^{100}\mathbb{E}[X_i] \\
&= \frac{1}{100}\cdot 100\theta = \theta.
\end{aligned}$$

so it is unbiased.

## 5. Y Me?

Let $y_1, y_2, \ldots y_n$ be i.i.d. samples of a random variable with density function

$$f_Y(y|\theta) = \frac{1}{2\theta}\exp(-\frac{|y|}{\theta})$$

.

Find the MLE for $\theta$ in terms of $|y_i|$ and $n$.

**Solution:**
Since the samples are i.i.d., the likelihood of seeing n samples of them is just their PDFs multiplied together. From there, take the log-likelihood, then the first derivative, set it equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned} L\left(y_1, \ldots, y_n \mid \theta\right) &= \prod_{i=1}^{n} \frac{1}{2\theta} \exp(-\frac{|y_i|}{\theta}) \\ \ln L\left(y_1, \ldots, y_n \mid \theta\right) &= \sum_{i=1}^{n} \left[-\ln 2 - \ln \theta - \frac{|y_i|}{\theta}\right] \\ \frac{\partial}{\partial \theta} \ln L\left(y_1, \ldots, y_n \mid \theta\right) &= \sum_{i=1}^{n} \left[-\frac{1}{\theta} + \frac{|y_i|}{\theta^2}\right] = 0 \\ -\frac{n}{\hat{\theta}} + \frac{\Sigma_{i=1}^{n}|y_i|}{\hat{\theta}^2} &= 0 \\ \hat{\theta} = \frac{\Sigma_{i=1}^{n}|y_i|}{n} & \end{aligned}$$

## 6. Laplace MLE

Suppose $x_1, \ldots, x_{2n}$ are iid realizations from the Laplace density (double exponential density): for $x \in \mathbb{R}$,

$$f_X\left(x \mid \theta\right) = \frac{1}{2} e^{-|x-\theta|}$$

Find the MLE for $\theta$. For this problem, you need not verify that the MLE is indeed a maximizer. You may find the **sign** function useful:

$$\operatorname{sgn}(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$$

**Solution:**
We begin by setting up the likelihood like we do in any case. Since these are i.i.d. realizations, we can multiply all their PDFs together. From there, take the log-likelihood, then the first derivative, where we notice that the derivative of $-\ln 2 - |x_i - \theta|$ is just the sign function of $x_i - \theta$. Then, set that equal to 0 and solve for for $\hat{\theta}$.

$$\begin{aligned} L\left(x_1, \ldots, x_{2n} \mid \theta\right) &= \prod_{i=1}^{2n} \frac{1}{2} e^{-|x_i-\theta|} \\ \ln L\left(x_1, \ldots, x_{2n} \mid \theta\right) &= \sum_{i=1}^{2n} \left[-\ln 2 - |x_i - \theta|\right] \\ \frac{\partial}{\partial \theta} \ln L\left(x_1, \ldots, x_{2n} \mid \theta\right) &= \sum_{i=1}^{2n} \operatorname{sgn}(x_i - \theta) = 0 \\ \hat{\theta} &= \text{any value in } [x_n^{'}, x_{n+1}^{'}] \end{aligned}$$

where $x_i^{'}$ is the $i^{\text{th}}$ order statistic: the $i^{\text{th}}$ smallest observation (see 5.10 in the textbook for more details).
If you wanted to argue that this is a global maximizer, note that the log likelihood is the sum of concave functions, so every critical point is a global maximizer.

## 7. What if we lose ?

[This is practice with earlier material] Suppose 59 percent of voters favor Proposition 600. Use the Normal approximation to estimate the probability that a random sample of 100 voters will contain:

(a) at most 50 in favor. Mention any assumption that you make.

**Solution:**

We will make an assumption here. We will assume that the $i^{th}$ person is in favor of the proposition with probability $\frac{59}{100}$. We define $X_i \sim$ Bernoulli($\frac{59}{100}$) representing whether the $i^{th}$ person is in favor or not. We define $X = \sum_{i=1}^{100} X_i$ representing the number of people who are in favor of the proposition. We can approximate $X$ by $Y \sim N(100 \cdot 0.59, 100 \cdot 0.242)$. We need to find $\mathbb{P}(\frac{Y-59}{\sqrt{(24.2)}} < \frac{50.5-59}{\sqrt{(24.2)}})$(after continuity correction and standardization) which is equal to $\Phi(-1.729)$.

(b) more than 100 voters in favor or fewer than 0 voters in favor (again based on this normal approximation). Will the probability be non zero?

**Solution:**

We will use our normal approximation $Y$ from part(a). We are interested in $\mathbb{P}(Y < -0.5) + \mathbb{P}(Y > 100.5)$(after continuity correction) which is the same as

$$\mathbb{P}(\frac{Y-59}{\sqrt{24.2}} < \frac{-0.5-59}{\sqrt{24.2}}) + \mathbb{P}(\frac{Y-59}{\sqrt{24.2}} > \frac{100.5-59}{\sqrt{24.2}}) = \Phi(-12.09) + 1 - \Phi(8.436)$$

. Yes, the probability will be non -zero because the density of the normal distribution is non-zero everywhere. Note that this result is acceptable because the normal distribution is an approximation.

## 8. Law of Total Probability Review

(a) (Discrete version) Suppose we flip a coin with probability $U$ of heads, where $U$ is equally likely to be one of $\Omega_U = \{0, \frac{1}{n}, \frac{2}{n}, ..., 1\}$ (notice this set has size $n+1$). Let $H$ be the event that the coin comes up heads. What is $\mathbb{P}(H)$?

**Solution:**

We can use the law of total probability, conditioning on $U = \frac{k}{n}$ for $k = 0, ..., n$. Note that the probability of getting heads conditioning on a fixed $U$ value is $U$, and that the probability of $U$ taking on any value in its range is $\frac{1}{n+1}$ since it is discretely uniform.

$$\mathbb{P}(H) = \sum_{k=0}^{n} \mathbb{P}(H|U = \frac{k}{n})\mathbb{P}(U = \frac{k}{n}) = \sum_{k=0}^{n} \frac{k}{n} \cdot \frac{1}{n+1} = \frac{1}{n(n+1)} \sum_{k=0}^{n} k = \frac{1}{n(n+1)} \frac{n(n+1)}{2} = \frac{1}{2}$$

(b) (Continuous version) Now suppose $U \sim$ Uniform(0,1) has the *continuous* uniform distribution over the interval $[0, 1]$. What is $\mathbb{P}(H)$?

**Solution:**

We do the same thing, this time using the continuous law of total probability. Note, this time, that we're conditioning on $U = u$ and taking the integral with respect to $u$, and that the density of $U$ for any value in its range is 1 because it is uniformly random.

$$\mathbb{P}(H) = \int_{-\infty}^{\infty} \mathbb{P}(H|U = u) f_U(u) du$$

We can take the integral from 0 to 1 instead because outside of that range the density of $U$ is 0.

$$= \int_0^1 \mathbb{P}(H|U = u) f_U(u) du = \int_0^1 u \cdot 1 du = \frac{1}{2}[u^2]_0^1 = \frac{1}{2}$$

(c) Let's generalize the previous result we just used. Suppose $E$ is an event, and $X$ is a continuous random variable with density function $f_X(x)$. Write an expression for $\mathbb{P}(E)$, conditioning on $X$.

**Solution:**

We use the continuous law of total probability again, this time not deriving it any further and sticking with negative infinity to infinity because we don't know the range of the RV $X$.

$$\mathbb{P}(E) = \int_{-\infty}^{\infty} \mathbb{P}(E|X = x) f_X(x) dx$$

## 9. MAP Estimation*

(Optional: depending on if we have covered this in lecture; Read sections 7.4 and 7.5, if you're interested) Let $x_1, ..., x_n$ be iid realizations from a distribution with common pmf $p_X(x; \theta)$ where $\theta$ is an unknown but **fixed** parameter. Let's call the event $\{X_1 = x_1, ..., X_n = x_n\} = \mathcal{D}$ for data. You may wonder why in MLE, we seek to maximize the likelihood $L(\mathcal{D} \mid \theta)$, rather than $\mathbb{P}(\theta \mid \mathcal{D})$. This is because it doesn't make sense to compute $\mathbb{P}(\theta)$, since $\theta$ is fixed. However, in **Maximum a Posteriori (MAP) estimation**, we assume the parameter is a random variable (denoted $\Theta$), and attempt to maximize $\pi_\Theta(\theta \mid \mathcal{D})$, where $\pi_\Theta$ is the pmf or pdf of $\Theta$, depending on whether $\Theta$ is continuous or discrete. Using Bayes Theorem, we get $\pi_\Theta(\theta \mid \mathcal{D}) = \frac{L(\mathcal{D}|\theta)\pi_\Theta(\theta)}{L(\mathcal{D})}$. To maximize the LHS with respect to $\theta$, we may ignore the denominator on the RHS since it is constant with respect to $\theta$. Hence MAP seeks to maximize $\pi_\Theta(\theta \mid \mathcal{D}) \propto L(\mathcal{D} \mid \theta)\pi_\Theta(\theta)$. We call $\pi_\Theta(\theta)$ the **prior** distribution on the parameter $\Theta$, and $\pi_\Theta(\theta \mid \mathcal{D})$ the **posterior** distribution on $\Theta$. MLE maximizes the likelihood, and MAP maximizes the product of the likelihood and the prior. If the prior is uniform, we will see that MAP is the same as MLE (since $\pi_\Theta(\theta)$ won't depend on $\theta$).

(a) Suppose we have the samples $x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0$ from the Bernoulli($\theta$) distribution, where $\theta$ is unknown. Assume $\theta$ is unrestricted; that is, $\theta \in (0, 1)$. What is $\hat{\theta}_{MLE}$?

**Solution:**

We begin with finding the likelihood by multiplying the probabilities of seeing each of the independent realizations from the Ber($\theta$) distribution. From there, we just use the MLE process to get the log-likelihood, take the first derivative, set it equal to 0, and solve for $\hat{\theta_{MLE}}$.

$$\begin{aligned}
L(x_1, \ldots, x_5 \mid \theta) &= \theta^2(1-\theta)^3 \\
\ln L(x_1, \ldots, x_5 \mid \theta) &= 2\ln(\theta) + 3\ln(1-\theta) \\
\frac{\partial}{\partial\theta} \ln L(x_1, \ldots, x_5 \mid \theta) &= \frac{2}{\theta} - \frac{3}{1-\theta} = 0 \\
& \quad 2 - 2\theta = 3\theta \\
\hat{\theta}_{MLE} &= \boxed{\frac{2}{5}}
\end{aligned}$$

(b) Suppose we impose that $\theta \in \{0.2, 0.5, 0.7\}$. What is $\hat{\theta}_{MLE}$?

**Solution:**

We can compute $L(\mathcal{D} \mid \theta)$ for each value of $\theta$, and take the largest.

$L(\mathcal{D} \mid 0.2) = (1 - 0.2)^3(0.2)^2 = 0.02048$

$L(\mathcal{D} \mid 0.5) = (1 - 0.5)^3(0.5)^2 = 0.03125$

$L(\mathcal{D} \mid 0.7) = (1 - 0.7)^3(0.7)^2 = 0.01323$

So $\hat{\theta}_{MLE} = \boxed{0.5}$.

(c) Assume $\Theta$ is restricted as in part (b) (now a random variable for MAP). Assume a (discrete) prior of $\pi_\Theta(0.2) = 0.1, \pi_\Theta(0.5) = 0.01, \pi_\Theta(0.7) = 0.89$. What is $\hat{\theta}_{MAP}$?

**Solution:**

We compute the objective to maximize for MAP:

$\pi_\Theta(0.2 \mid \mathcal{D}) \propto L(\mathcal{D} \mid 0.2)\pi_\Theta(0.2) = 0.02048 \cdot 0.1 = 0.002048$

$\pi_\Theta(0.5 \mid \mathcal{D}) \propto L(\mathcal{D} \mid 0.5)\pi_\Theta(0.5) = 0.03125 \cdot 0.01 = 0.0003125$

$\pi_\Theta(0.7 \mid \mathcal{D}) \propto L(\mathcal{D} \mid 0.7)\pi_\Theta(0.7) = 0.01323 \cdot 0.89 = 0.0117747$

Hence $\hat{\theta}_{MAP} = \boxed{0.7}$.

(d) Show that we can make the MAP estimator whatever we want it to be. That is, for each of the three candidate parameters above, find a prior distribution on $\Theta$ such that the MAP estimate is that parameter.

**Solution:**

Just assign a prior of $1$ to the desired parameter. If you don't want something degenerate, assign a prior extremely close to $1$, and give uniform probability to the other parameters.

(e) Typically, for the Bernoulli/Binomial distribution, if we use MAP, we want to be able to get any value $\theta \in (0, 1)$ (not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$). So we assign $\theta$ the **Beta distribution** with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = c\theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0, 1)$ and $0$ otherwise as a prior, where $c$ is a normalizing constant which has a complicated form. The **mode** of a $W \sim \text{Beta}(\alpha, \beta)$ random variable is given as $\frac{\alpha-1}{\alpha+\beta-2}$ (the mode is the value with the highest density $= \arg\max_{w \in (0,1)} f_W(w)$). Suppose $x_1, ..., x_n$ are iid samples from the Bernoulli distribution with unknown parameter, where $\sum_{i=1}^n x_i = k$. Recall that the MLE is $k/n$. Show that the posterior $\pi_\Theta(\theta \mid \mathcal{D})$ has a $\text{Beta}(k + \alpha, n - k + \beta)$ density, and find the MAP estimator for $\Theta$. (Hint: use the mode given). Notice that $\text{Beta}(1, 1) \equiv \text{Uniform}(0, 1)$. If we had this prior, how would the MLE and MAP estimates compare?

**Solution:**

We want to maximize $\pi_\Theta(\theta \mid \mathcal{D}) \propto L(\mathcal{D} \mid \theta)\pi_\Theta(\theta) \propto \left(\theta^k(1-\theta)^{n-k}\right)\left(\theta^{\alpha-1}(1-\theta)^{\beta-1}\right) = \theta^{(k+\alpha)-1}(1-\theta)^{(n-k+\beta)-1}$. Hence the posterior $\sim \text{Beta}(k + \alpha, n - k + \beta)$. We are given the mode of any beta distribution, so our estimate is $\hat{\theta}_{MAP} = \boxed{\dfrac{k + \alpha - 1}{n + \alpha + \beta - 2}}$. If $\alpha = \beta = 1$, then this is exactly the MLE, and $\text{Beta}(1, 1) \equiv \text{Uniform}(0, 1)$, so having a uniform prior causes the MLE to equal the MAP estimate.

(f) Since the posterior is also a Beta distribution, we call the Beta distribution the **conjugate prior** to the Bernoulli/Binomial distribution. Intepret what the parameters $\alpha, \beta$ mean as to the prior.

**Solution:**

$\alpha - 1$ is the number of heads you pretend to see beforehand, and $\beta - 1$ is the number of tails you pretend to see beforehand. Why is this? Because our MLE was $\frac{k}{n}$ (heads/trials), and the MAP estimate is $\frac{k+\alpha-1}{n+(\alpha+\beta-2)} = \frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)}$. Hence we add $\alpha+\beta-2$ "fake" trials, $\alpha-1$ which are heads (numerator), and the other $\beta - 1$ which are tails. This should look familiar as our estimates for $\mathbb{P}(word \mid spam)$ and $\mathbb{P}(word \mid ham)$ with a $\text{Beta}(2, 2)$ prior when we did smoothing for Naive Bayes.

(g) Which do you think is "better", MLE or MAP?

**Solution:**

There is no right answer. There are two main schools in statistics: Bayesians and Frequentists. Frequentists prefer MLE since they don't believe you should be putting a prior belief on anything, and you should only make judgment based on what you've seen. They believe the parameter being estimated is a fixed

quantity. On the other hand, Bayesians prefer MAP, since they can incorporate their prior knowledge into the estimation. Hence the parameter being estimated is a random variable, and we seek the mode - the value with the highest probability or density. An example would be estimating the probability of heads of a coin - is it reasonable to assume it is more likely fair than not? If so, what distribution should we put on the parameter space?

Anyway, in the long run, the prior “washes out", and the only thing that matters is the likelihood; the observed data. For small sample sizes like this, the prior significantly influences the MAP estimate. However, as the number of samples goes to infinity, the MAP and MLE are equal.