STAT 3503/8109 Lecture 10 Notes

Edoardo Airoldi

Scribe: Srikar Katta[*]

Fall 2020: November 2, 2020

1 Introduction

Today, we will extend our discussion of maximum likelihood estimation by working through an example of maximum likelihood estimation but now utilizing indicator functions ($\mathbb{1}$). Then, we will talk about sufficient statistics and their role in estimation. We will conclude with Bayesian estimation strategies.

2 Maximum Likelihood with Indicator Functions

There are a few famous instances of this problem. One example is known as the "German Tank Problem." In World War II, the Allied forces were destroying German and Italian tank. Each one that was destroyed had a serial number, and each serial number was sequential. A problem of key consideration was estimating how many tanks the German and Italians had. While this is an historic example, there are many current situations that this solution is still applicable to. For example, in finance, oftentimes traders assume they know log returns (a measure used to decide which stocks to buy or sell) are uniformly distributed. If one were to find the maximum negative log return (maximum loss), then this approach that will be laid out in this example will be of key importance.

Example 2.1. *Assume* $x_1 ... x_n \overset{\text{iid}}{\sim} Uniform[0, \theta]$*, where* θ *is an unknown constant. Assume* $x_1 ... x_n$ *are observed. Find* $\hat{\theta}_{MLE}$*.*

[*] Please share any comments or suggestions with Srikar Katta at srikar@temple.edu

Solution: First, let us create the 2x2 table for this situation:

	Observed	**Unobserved**
Variable	$x_1...x_n$	NA
Constant	NA	θ

Because we are estimating with no latent variables and we want to ensure that our estimate is within its parameter space, maximum likelihood estimation is a perfect strategy for this problem. To reiterate, maximum likelihood estimation allows us to find the value of $\hat{\theta}$ that maximizes the chances (likelihood) of the observed data being seen, given some assumptions about the model's distributions.

For maximum likelihood estimation, we need to find the likelihood. Since we have no latent random variables, we can directly find the proper likelihood: $L(\theta) = \mathbb{P}(x_1...x_n|\theta)$. Since $x_1...x_n \overset{\text{iid}}{\sim} Uniform[0, \theta]$, the probability density function for the observed data is $\frac{1}{\theta}$. So,

$$\begin{aligned} L(\theta) &= \mathbb{P}(x_1...x_n|\theta) \\ &= \mathbb{P}(x_1|\theta)...\mathbb{P}(x_n|\theta), \text{ because of the IID assumption} \\ &= \prod_{i=1}^{n} \frac{1}{\theta}, x_i \in [0, \theta], \end{aligned}$$

which is essentially saying that the probability of the observed data occurring given some θ is just $\frac{1}{\theta}$, *as long as* any observed data is not greater than θ. However, having this $x_i \in [0, \theta]$ term is a little cumbersome, which is why we use the indicator function, $\mathbb{1}_{condition}(x)$, which is essentially a shorthanded form of a piece wise function. If the condition is met, then it returns 1, and it returns 0 if the condition is not met. In this example, we want $0 \leq x_i \leq \theta$, so we will the indicator function would work as follows:

$$\mathbb{1}_{[0,\theta]}(x_i) = \begin{cases} 1 & \text{if } 0 \leq x_i \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

So, now we can rewrite the likelihood using this new notation:

$$\begin{aligned} L(\theta) &= \mathbb{P}(x_1...x_n|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\theta}, x_i \in [0, \theta] \end{aligned}$$

$$= \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x_i)$$

$$= \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbb{1}_{[0,\theta]}(x_i),$$

which is

If we refer back to the uniform distribution, θ represents the upper bound for the data, so if there exists some observed data point x_i that is greater than θ, then θ is no longer an upper bound. That means the probability of that specific outcome occurring is 0. Since we take the product of probability of each individual observation occurring given θ (i.e., $\prod_{i=1}^{n} \frac{1}{\theta}$), if one of the observations is greater than θ, the entire likelihood evaluates to 0. For example, suppose $\theta = 5$ and some observation $x_k = 10$. Then, $L(\theta = 10 | x_k = 5) = 0$. Essentially, θ not only dictates the *height* of the probability distribution function but also the support of the observations now.

Now, notice that the likelihood is a function of x_i since it is the input for $\mathbb{1}_{[0,\theta]}(x_i)$. Because we want to maximize the likelihood only with respect to θ, we need to find some way to "swap" $[0, \theta]$ and x_i that is fair *for all* x_i. So, for all x_i, $\theta \geq x_i$. Since x_i belong to a finite set of terms, we can find the maximum of all of the x_i: $\max x_i$. By definition of the maximum, for all x_i, $\max x_i \geq x_i$. So, that means if $\theta \geq \max x_i$, then $\prod_{i=1}^{n} \mathbb{1}_{[0,\theta]}(x_i) = 1$. So, we can rewrite this as $\mathbb{1}_{[0,\theta]}(\max x_i) = 1$, which says that as long as the maximum is between 0 and θ, then θ is valid. Now, we still need to swap the x_i and θ. Well, instead of saying $0 \leq \max x_i \leq \theta$, we can say that $\max x_i \leq \theta < \infty$. So, $\mathbb{1}_{[0,\theta]}(\max x_i) = 1 = \mathbb{1}_{[\max x_i, \infty]}(\theta) = 1$. In other words, as long as θ is between the maximum x_i and infinity, then the likelihood of θ is $\frac{1}{\theta^n} * 1$. So,

$$\textit{Likelihood}(\theta) = \frac{1}{\theta^n} \mathbb{1}_{[\max x_i, \infty]}(\theta),$$

which is now truly a function of θ.

Now that we have a likelihood expressed in terms of θ, we can find $\hat{\theta}_{MLE}$. Let ϵ and a be some numbers greater than 0. Intuitively, we know that when $a < a + \epsilon$, so that means $\frac{1}{a} > \frac{1}{a+\epsilon}$. So, now, replacing that with θ^n, as θ^n increases, $\frac{1}{\theta^n}$ decreases. So, the $\hat{\theta}$ that maximizes the likelihood will be the $\hat{\theta}_{MLE}$ that is less than all other possible $\hat{\theta}$s. Since we have the restriction that $\theta \geq \max x_i$, it is quite easy to see that the smallest $\hat{\theta}$ can be is $\max x_i$, so $\hat{\theta}_{MLE} = \max x_i$.

Analytically, we can take the derivative of the likelihood and find $\hat{\theta}_{MLE}$:

$$\frac{\partial L}{\partial \theta} = -n\left(\frac{1}{\theta^{n-1}}\right),$$

which is less than 0 for all θ, suggesting a decreasing function. So the minimum θ maximizes the likelihood. Thus, $\hat{\theta}_{MLE} = \max x_i$. □

3 Sufficient Statistics and Maximum Likelihood

First,a statistic is any function of the data that is not a function of unknown constants/parameters. Formally, a sufficient statistic is a function of the data $x_1...x_n$, denoted as $T(x_1...x_n)$, that is relevant for computing the maximum likelihood estimator for the unknown constant θ. This means that different models have different sufficient statistics.

Example 3.1. *Suppose x_i is the number of words in document i. Assume $x_i \overset{\text{iid}}{\sim} Poisson(\lambda)$ where $\lambda > 0$. Suppose $x_1...x_n$ are observed and λ is an unknown constant. Find $\hat{\lambda}_{MLE}$.*

Solution: Since $x_1...x_n$ have a theoretical distribution and λ is unknown without a theoretical distribution (note, $\lambda > 0$ is not a distribution), the 2x2 looks as follows:

	Observed	**Unobserved**
Variable	$x_1...x_n$	NA
Constant	NA	λ

So, the likelihood is only a function of λ and the proper likelihood can be found directly:

$$\begin{aligned}
L(\lambda) &= \mathbb{P}(x_1...x_n|\lambda) \\
&= \prod_{i=1}^{n} \mathbb{P}(x_i|\lambda) \\
&= \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}, \text{ since the pdf of } Poisson(\lambda) = \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \\
&= \frac{e^{-\sum_{i=1}^{n}\lambda}\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}, \text{ from exponent rules} \\
&= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n}x_i}}{\prod_{i=1}^{n}x_i!}.
\end{aligned}$$

Now, because we would like to maximize the likelihood, we can instead maximize the log of the likelihood which will most likely be easier to take the derivative of:

$$\begin{aligned}
l &= logL(\lambda) \\
&= log\left[\frac{e^{-n\lambda}\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}\right] \\
&= -n\lambda + \sum_{i=1}^{n} x_i log(\lambda) - log\left(\prod_{i=1}^{n} x_i!\right) \text{ from log rules} \\
&= -n\lambda + \sum_{i=1}^{n} x_i log(\lambda) - \sum_{i=1}^{n} log(x_i!\,) \text{ from log rules}
\end{aligned}$$

First, it is essential to understand how the information in the data affects the likelihood. Well there are two factors in the likelihood with x_i:

$$\sum_{i=1}^{n} log(x_i!\,) \text{ and } \sum_{i=1}^{n} x_i.$$

To find $\hat{\lambda}$, we take the derivative of the log likelikelihood with respect to λ:

$$\begin{aligned}
0 = \frac{\partial l}{\partial \lambda} = \frac{\partial}{\partial \lambda}&\left[-n\lambda + \sum_{i=1}^{n} x_i log(\lambda) - \sum_{i=1}^{n} log(x_i!\,)\right] \\
&= -n\frac{\partial}{\partial \lambda}\lambda + \sum_{i=1}^{n} x_i \frac{\partial}{\partial \lambda} log(\lambda) - \frac{\partial}{\partial \lambda}\sum_{i=1}^{n} log(x_i!\,) \\
&= -n + \sum_{i=1}^{n} \frac{x_i}{\lambda} + 0.
\end{aligned}$$

Notice, $\sum_{i=1}^{n} x_i$ is essentially irrelevant to finding the maximum likelihood estimator since it is not related to λ.

Then, the maximum likelihood estimator can be found from

$$\begin{aligned}
& 0 = \frac{\partial l}{\partial \lambda} \\
\iff\ & 0 = 0 - n + \frac{\sum_{i=1}^{n} x_i}{\lambda} \\
\iff\ & n = \frac{\sum_{i=1}^{n} x_i}{\lambda}
\end{aligned}$$

$$\iff \lambda^* = \frac{\sum_{i=1}^n x_i}{n},$$

so (as long as the second derivative of the log likelihood is less than 0, suggesting that $\lambda*$ is maximum), $\hat{\lambda}_{MLE} = \lambda^*$. Because $\sum_{i=1}^n x_i$ is a function of the data and is relevant to computing the maximum likelihood, it is a sufficient statistic. □

3.1 Factorization Theorem

Suppose $x_1...x_n$ represents the data. Then, the statistic $T(x_1...x_n)$ is sufficient for unknown constant θ if the likelihood can be factored out into two functions:

$$L(\theta) = h(x_1...x_n)g(\theta, T(x_1...x_n)).$$

Example 3.2. *Suppose $x_1...x_n \overset{\text{iid}}{\sim} Poisson(\lambda)$. Find the sufficient statistic.*

Solution: First, we need to find the likelihood:

$$\begin{aligned} L(\lambda) &= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n (x_i!)} \\ &= \frac{1}{\prod_{i=1}^n x_i!} e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}. \end{aligned}$$

Notice, we can write the likelihood as the product of two terms: one dependent on the data and unknown constant and one dependent on only the data. The term dependent on only the data is the sufficient statistic. So, $\frac{1}{\prod_{i=1}^n x_i!}$ is the sufficient statistic for λ. □

Now, we can redefine this definition of sufficient statistics for the log likelihood instead. Recall, the likelihood is the product of two functions, one dependent on the data and unknown constants and another dependent only on the data. So,

$$\begin{aligned} Likelihood(\lambda) &= h(x_1...x_n)g(\lambda, T(x_1...x_n)) \\ \iff log(Likelihood(\lambda)) &= log\left[h(x_1...x_n)g(\lambda, T(x_1...x_n))\right] \\ &\iff \lambda = log(h(x_1...x_n)) + log(g(\lambda, T(x_1...x_n))). \end{aligned}$$

Notice, the derivative of $log(h(x_1...x_n))$ with respect to θ is 0, so this term is not necessary for finding the maximum likelihood estimator. Then, in log likelihood terms, a statistic

$T(x_1...x_n)$ is called sufficient if the log likelihood of θ can be written as

$$logLikelihood(\theta) = h(x_1...x_n) + g(\lambda, T(x_1...x_n)).$$

Example 3.3. *Suppose $x_1...x_n \overset{\text{iid}}{\sim} Normal(\mu, \sigma^2)$, where σ^2 is **known**. Find the sufficient statistics for the unknown $\hat{\mu}_{MLE}$.*

Solution: We just need to find the likelihood and represent it as a product of two terms: one with only the data and another with only the data and μ. So,

$$\begin{aligned}
Likelihood(\mu) &= \mathbb{P}(x_1...x_n|\mu, \sigma^2) \\
&= \prod_{i=1}^{n} \mathbb{P}(x_1|\mu, \sigma^2) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}^n} \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}^n} \prod_{i=1}^{n} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}^n} \prod_{i=1}^{n} e^{-\frac{x_i^2+\mu^2-2x_i\mu}{2\sigma^2}} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}^n} \prod_{i=1}^{n} e^{-\frac{x_i^2}{2\sigma^2}} e^{\mu^2} e^{-2x_i\mu} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\sum_{i=1}^n \frac{x_i^2}{2\sigma^2}} e^{\sum_{i=1}^n \mu^2} e^{-2\sum_{i=1}^n x_i\mu} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\sum_{i=1}^n \frac{x_i^2}{2\sigma^2}} e^{n\mu^2} e^{-2\mu\sum_{i=1}^n x_i}.
\end{aligned}$$

I claim that $\sum_{i=1}^n x_i$ is a sufficient statistic since the likelihood can be rewritten in our form. Let $h(x_1...x_n) = \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\sum_{i=1}^n \frac{x_i^2}{2\sigma^2}}$ let $T(x_1...x_n) = \sum_{i=1}^n x_i$, and let $g(\lambda, T(x_1...x_n)) = e^{n\mu^2} e^{-2\mu T(x_1...x_n)}$. Then,

$$\begin{aligned}
Likelihood(\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\sum_{i=1}^n \frac{x_i^2}{2\sigma^2}} e^{n\mu^2} e^{-2\mu\sum_{i=1}^n x_i} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}^n} e^{-\sum_{i=1}^n \frac{x_i^2}{2\sigma^2}} e^{n\mu^2} e^{-2\mu\sum_{i=1}^n x_i}
\end{aligned}$$

$$= h(x_1...x_n)g(\mu, T(x_1...x_n)).$$

Since there exists a function g that is dependent only on the data and h that is dependent on the data and the unknown constant and their product is equal to the likelihood, $\sum_{i=1}^{n} x_i$ is a sufficient statistic[1] for μ. □

3.2 Benefits of Sufficient Statistics

Discovering the sufficient statistics of a probability distribution can significantly simplify the analytical time necessary to find a maximum likelihood estimator. Suppose we have observed data $x_1...x_n$ that follow some probability distribution and unknown constants θ. We want to estimate $\hat{\theta}_{MLE}$. Suppose we have the sufficient statistic $T(x_1...x_n)$ and functions g and h such that $Likelihood(\theta) = h(x_1...x_n)g(\theta, T(x_1...x_n))$. We can find $\hat{\theta}_{MLE}$ by finding the argument that maximizes the likelihood:

$$\begin{aligned}\hat{\theta}_{MLE} &= arg\max_{\theta} Likelihood(\theta)\\ &= arg\max_{\theta} h(x_1...x_n)g(\theta, T(x_1...x_n)).\end{aligned}$$

Since $h(x_1...x_n)$ does not contain θ, when we take the partial derivative of the likelihood with respect to θ, the $\hat{\theta}$ that maximizes the likelihood will be the same $\hat{\theta}$ that maximizes $g(\theta, T(x_1...x_n))$. So, isolating the sufficient statistics can greatly simplify the amount of computation we actually need to do.

4 Bayesian Estimation

This section will be motivated by the following problem statement: suppose x is a binary random variable that details whether someone's starting salary is greater than \$200,000. Suppose z is a binary random variable that is 1 if someone gets an A in STAT 8109 and 0 otherwise. We observe $(x_1, z_1)...(x_n, z_n)$.

Example 4.1. *Model 1: Suppose* $x_1...x_n \overset{\text{iid}}{\sim}$ *Bernoulli*(θ), *and* $\theta_i = \alpha + z_i\beta$, *and* $0 \leq \alpha + \beta \leq 1$. *This is essentially saying that if a student earned an A in Statistics 8109, then the probability that their starting salary is greater than 200k is β more than someone who didn't take Statistics 8109.*

[1] The sufficient statistic is not unique. For example $\frac{\sum_{i=1}^{n} x_i}{n}$ is also a valid sufficient statistic for μ.

Model 2: Suppose $x_1 ... x_n \overset{\text{iid}}{\sim}$ *Bernoulli*(θ) *and* $log\left(\frac{\theta_i}{1-\theta_i}\right) = \alpha + z_i\beta$*, where* α *and* β *are real valued. This is the same as* $\theta_i = \frac{1}{1+e^{-\alpha - z_i\beta}}$.

In either model, estimate α *and* β*.*

Solution: First, we will find α and β using maximum likelihood estimation. First, we need to find the likelihood:

$$\begin{aligned} Likelihood(\alpha, \beta) &= \mathbb{P}(x_1 ... x_n | \alpha, \beta) \\ &= \mathbb{P}(x_1|\alpha, \beta) ... \mathbb{P}(x_n|\alpha, \beta) \\ &= \prod_{i=1}^{n} \mathbb{P}(x_i|\alpha, \beta) \\ &= \prod_{i=1}^{n} Bernoulli(x_i|\alpha, \beta). \end{aligned}$$

Then, the log likelihood would be

$$\begin{aligned} logLikelihood(\alpha, \beta) &= log\left[\prod_{i=1}^{n} Bernoulli(x_i|\alpha, \beta)\right] \\ &= \sum_{i=1}^{n} log(Bernoulli(x_i|\alpha, \beta)). \end{aligned}$$

For model 1, we specify this as follows since $x_i \sim Bernoulli(\alpha + \beta z_i)$. So,

$$\begin{aligned} logLikelihood(\alpha, \beta) &= \sum_{i=1}^{n} log(Bernoulli(x_i|\alpha, \beta)) \\ &= \sum_{i=1}^{n} log\left((\alpha + \beta z_i)^{x_i}(1 - \alpha - \beta z_i)^{1-x_i}\right) \\ &= \sum_{i=1}^{n} log\left((\alpha + \beta z_i)^{x_i}\right) + log\left((1 - \alpha - \beta z_i)^{1-x_i}\right) \\ &= \sum_{i=1}^{n} x_i log(\alpha + \beta z_i) + (1 - x_i) log(1 - \alpha - \beta z_i) \\ &= \sum_{i=1}^{n} x_i log(\alpha + \beta z_i) + \sum_{i=1}^{n} (1 - x_i) log(1 - \alpha - \beta z_i). \end{aligned}$$

For model 2, we specify the log likelihood as follows:

$$\begin{aligned} logLikelihood(\alpha, \beta) &= \sum_{i=1}^{m} log\left(\left(\frac{1}{1+e^{-\alpha-\beta z_i}}\right)^{x_i}\left(1-\frac{1}{1+e^{-\alpha-\beta z_i}}\right)^{1-x_i}\right) \\ &= \sum_{i=1}^{m} log\left(\frac{1}{1+e^{-\alpha-\beta z_i}}\right)^{x_i} + log\left(1-\frac{1}{1+e^{-\alpha-\beta z_i}}\right)^{1-x_i} \\ &= \sum_{i=1}^{m} x_i log\left(\frac{1}{1+e^{-\alpha-\beta z_i}}\right) + (1-x_i) log\left(1-\frac{1}{1+e^{-\alpha-\beta z_i}}\right). \end{aligned}$$

Now, we need to find α, β that maximize the log likelihoods. By going through the steps that we usually do for maximum likelihood estimation, we can find $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$. □
Now, suppose Professor Airoldi does this every year from 2009 to 2019. In 2020, he has $x_1 ... x_n$ observations, and in each year he has a varying number of students. So, in 2009, we observe $x_1^{2009} ... x_{n_{2009}}^{2009}$ where $n_{2009} = 54$. In 2010, we observe $x_1^{2010} ... x_{n_{2010}}^{2010}$ where $n_{2010} = 95$. This continues so on and so forth. Then, in 2018, we observe $x_1^{2018} ... x_{n_{2018}}^{2018}$ where $n_{2018} = 150$. And in 2019, we observe $x_1^{2019} ... x_{n_{2019}}^{2019}$ where $n_{2019} = 35$. Here, by maximizing the likelihood $L(\alpha, \beta) = \mathbb{P}(x_1^{2009} ... x_{35}^{2019} | \alpha, \beta)$, we can find $\hat{\alpha}_{MLE}$ and $\hat{\beta}_{MLE}$.

But perhaps that in each year, some data was missing. Suppose for model 1, in 2009, $\hat{\alpha}_{MLE} = 0.1$ and $\hat{\beta}_{MLE} = 0.5$. Then, the probability any student earns over 200k is 10%, but a student who earns an A in Statistics 8109 has a 60% probability of getting a starting salary over 200k. Suppose in 2018, $\hat{\alpha}_{MLE} = 0.1$ and $\hat{\beta}_{MLE} = 0.75$. Bayesian estimation will help us answer questions such as, "How can we best use this information to estimate/update the estimate for α and β?"

4.1 Idea

First, we posit distributions for α and β, called $\xi(\alpha, \beta)$ to encode "prior information." In the new Bayesian setup, we have $x_1^{2020} ... x_{n_{2020}}^{2020}$ and $n_{2020} = 54$. Assume $x_i \overset{\text{iid}}{\sim} Bernoulli(\alpha + \beta z_i)$. Further, assume $\alpha = 0.1$ (so α is now a known constant) and $\beta \sim Beta(a, b)$[2]. The first step is to *calibrate the prior distribution*, which is essentially means estimating the unknown constants a, b of the $Beta(a, b)$ using historical data $\hat{\beta}_{MLE}^{2009}, \hat{\beta}_{MLE}^{2010}, ..., \hat{\beta}_{MLE}^{2019}$ using $\hat{\beta}_{MLE}^{i} \overset{\text{iid}}{\sim} Beta(a, b)$. This would allow you to find $\hat{a}_{MLE}$ and $\hat{b}_{MLE}$.

Now, recall we observe $x_1^{2020} ... x_{54}^{2020} \overset{\text{iid}}{\sim} Bernoulli(.10 + z_i\beta)$ and $\beta \sim Beta(\hat{a}, \hat{b})$. We want to estimate β. So a few things have changed from earlier ideas. Let us create the 2x2 table

[2] To see form of Beta distribution, refer to notes 4

to specify how this problem looks:

	Observed	**Unobserved**
Variable	$x_1^{2020}...x_{54}^{2020}$	β
Constant	$z_1^{2020}...z_{54}^{2020}, \hat{a}, \hat{b}$	NA

Recall, the likelihood is a function of the constants (known and unknown) given the data. So, the fact that we do not have an unknown constants should not dissuade our search for the likelihood: $CompleteLikelihood = \mathbb{P}(\text{observations, latent random variables}|\text{ constants})$. However, we are interested in $\mathbb{P}$(latent random variable | observations, constants), which we can find using Bayes' Theorem (i.e., $\mathbb{P}(A|B) = \frac{\mathbb{P}(A,B)}{\mathbb{P}(B)}$. So, compactly, once we obtain $\mathbb{P}(\beta|x_1...x_{54}, z_1...z_{54}, \hat{a}, \hat{b})$, we can obtain β by finding the average of the new distribution, known as the posterior mean (PM) estimate. In other words, we can find $\hat{\beta}$ as,

$$\hat{\beta}_{PM} = \mathbb{E}[\beta|x_1...x_{54}, z_1...z_{54}, \hat{a}, \hat{b}].$$

Now, instead of using the mean of the distribution, we can consider using the mode of the distribution, which is the $\hat{\beta}$ that maximizes the probability distribution. This is known as the maximum a posteriori (MAP) technique. Mathematically,

$$\hat{\beta}_{MAP} = arg\underset{\beta}{max}\mathbb{P}(\beta|x_1...x_{54}, z_1...z_{54}, \hat{a}, \hat{b}).$$

In all of these examples, $Beta(\hat{a}, \hat{b})$ is the prior distribution and $\mathbb{P}(\beta|x_1...x_{54}, z_1...z_{54}, \hat{a}, \hat{b})$ is known as the posterior distribution.

Suppose $x_1...x_n \overset{\text{iid}}{\sim} Bernoulli(0.1 + z_i\beta)$ and are observed. Suppose β is an unknown constant. Then, $L(\beta) = \mathbb{P}(\text{observations}|\beta)$. And we can find $\hat{\beta}_{MLE} = \arg\max_\beta L(\beta)$.

4.2 Frequentist vs Bayesion Approaches

Now, suppose $x_1...x_n \overset{\text{iid}}{\sim} Bernoulli(0.1+z_i\beta)$ and $\beta \sim Beta(a, b)$. Then, $L(a, b) = \mathbb{P}(\text{observations}|a, b)$. However, we can not access this directly: we need to take the integral with respect to β of the complete likelihood:

$$\begin{aligned} Likelihood(a, b) &= \mathbb{P}(\text{observed}|a, b) \\ &= \int_\beta \mathbb{P}(\text{observed}, \beta|a, b) d\beta \end{aligned}$$

$$= \int_{\beta} \mathbb{P}(\text{observed}|\beta)\mathbb{P}(\beta|a,b)d\beta, \text{ from Bayes' Rule.}$$

Then, $\hat{a}, \hat{b} = \arg\max_{a,b} L(a,b)$. To find $\hat{\beta}$, we need to obtain $\mathbb{P}(\beta|\text{ observed }, a, b)$. Then, we have two strategies:

$$\hat{\beta}_{PM} = \mathbb{E}[\beta|\text{ observed }, a, b]$$
$$\hat{\beta}_{MAP} = \arg\max_{\beta} \mathbb{P}(\beta|\text{ observed }, a, b).$$

In the Frequentist approach, we assume β has no prior distribution while we assume that it does in the Bayesian approach. This is essentially the core of the Frequentist and Bayesian statistical differences. By assuming some variability, our unknown is now a latent variable and that can depend on unknown constants, so we now have two problems: estimating unknown constants and the latent variable. This we can compute using the Bayesian approach by finding the posterior distribution.

5 Conclusion

We detailed another approach to Maximum Likelihood Estimation using the indicator function and demonstrated its usefulness in the context of the German Tank Problem. We then discussed sufficient statistics, which allows us to simplify our maximum likelihood calculations. And lastly, we introduced Bayesian estimation strategies and outlined the differences between Frequentist and Bayesian approaches. Moving forward, we will highlight techniques for discovering the posterior distribution and further understand Bayesian inference.