37262 Mathematical Statistics

Lecture 6

Estimation

- In the field of probability, we tend to work from known (or assumed) distributions and then calculate how probable certain events or outcomes would be.

- With classical null hypothesis significance testing, we assume a distribution and then consider whether or not an observed dataset appears compatible with a hypothesised distribution, and hence whether or not the (null) hypothesis should be rejected,

- With **estimation**, rather than simply considering whether or not a given hypothesis should be rejected given a dataset, we instead seek to make statements about what values we believe describe properties or parameters of a distribution, and how much uncertainty we have about these.

Moments

- The ***k*th moment** of a random variable X is defined as $m_k = E(X^k)$.
- In some contexts, it can be more helpful to consider central moments.
- The ***k*th central moment** of a random variable X is defined as $c_k = E((X - m_1)^k)$.
- You may recognise m_1 as the mean of the distribution of X and c_2 as its variance.
- While the first and second moments characterise the mean and variance, higher moments can capture other features of a distribution such as skew (third moment) and kurtosis (fourth moment.)

Sample Moments

- Consider now drawing independent observations $x_1, x_2, ..., x_n$ of a random variable X.
- The ***k*th sample moment** is defined as $s_k = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,k}$.
- Similarly, the ***k*th central sample moment** is defined as $cs_k = \frac{1}{n}\sum_{i=1}^{n} (x_i - s_1)^k$.
- We know (from the Weak Law of Large Numbers) that as the sample size n tends to infinity, the sample moments tend towards the true moments of the distribution of X.
- We can use this fact to estimate unknown properties of the distribution of X given a sample of observations.

Method of Moments

- Using the **method of moments**, we estimate parameters of a distribution by matching the moments of a distribution to be equal to those of a sample drawn from it.

- Consider for example, flipping a (possibly unfair) coin 10 times and observing the number of times it lands Heads. If X is the number of Heads, then $X \sim Bin(10, p)$ where p is the unknown probability of the coin landing Heads.

- Given a sample of realisations of X, $\{x_1, x_2, x_3, x_4, x_5\} = \{3, 2, 5, 4, 3\}$, we can obtain an estimate of the unknown parameter p.

- We do this by calculating a moment of X and a corresponding sample moment and equating these.

Method of Moments

- It is easily shown that, for $X \sim Bin(10, p)$ that $m_1 = E(X) = 10p$.
- The first sample moment of $\{x_1, x_2, x_3, x_4, x_5\} = \{3, 2, 5, 4, 3\}$ is $s_1 = \frac{1}{5}(3 + 2 + 5 + 4 + 3)$.
- Equating $m_1 = 10p = s_1 = \frac{17}{5}$ gives an estimate of p (via the Method of Moments) of

$$\hat{p}_{MM} = \frac{17}{50} = 0.34.$$

Method of Moments

- For two (or more) parameter distributions, we will need to calculate more than one moment.

- Note that the number of parameters we need to estimate is not always the same as the number in the distribution, since we can sometimes fix some of these as known, as we did with the earlier example of flipping 10 coins, leading to $X \sim Bin(10, p)$.

- Although the binomial is a two parameter distribution, the fact that we know that 10 coins were flipped reduces the estimation problem to just one unknown parameter, p.

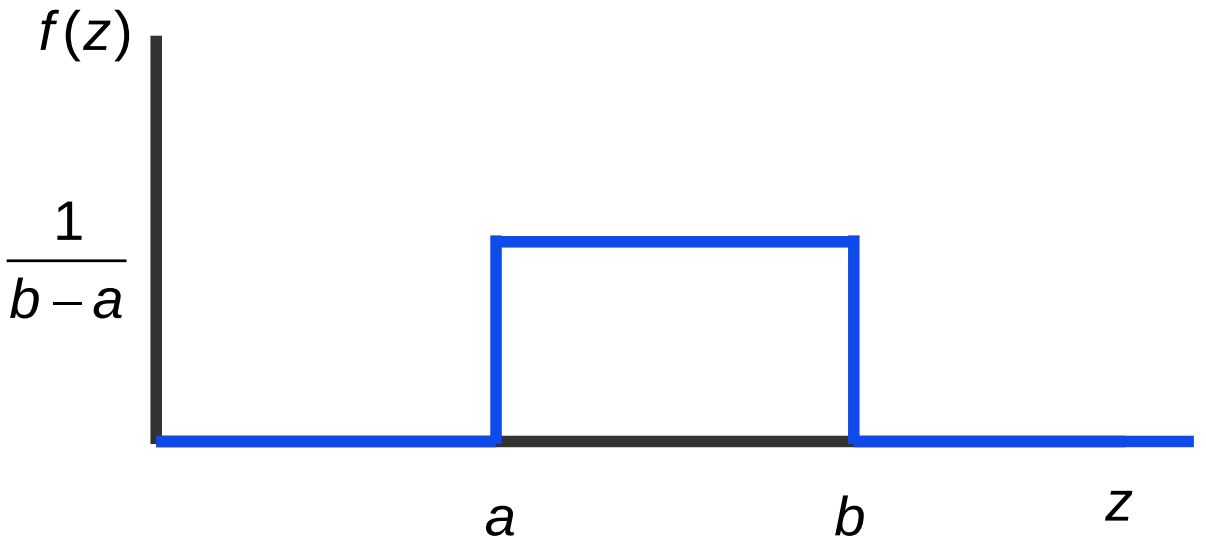
Method of Moments

- Consider now estimating the two parameters a and b ($a < b$) for a uniform variable

  $Z \sim U[a,b]$

- The density function of Z is

$$f(z) = \begin{cases} \dfrac{1}{b-a} & z \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$



Method of Moments

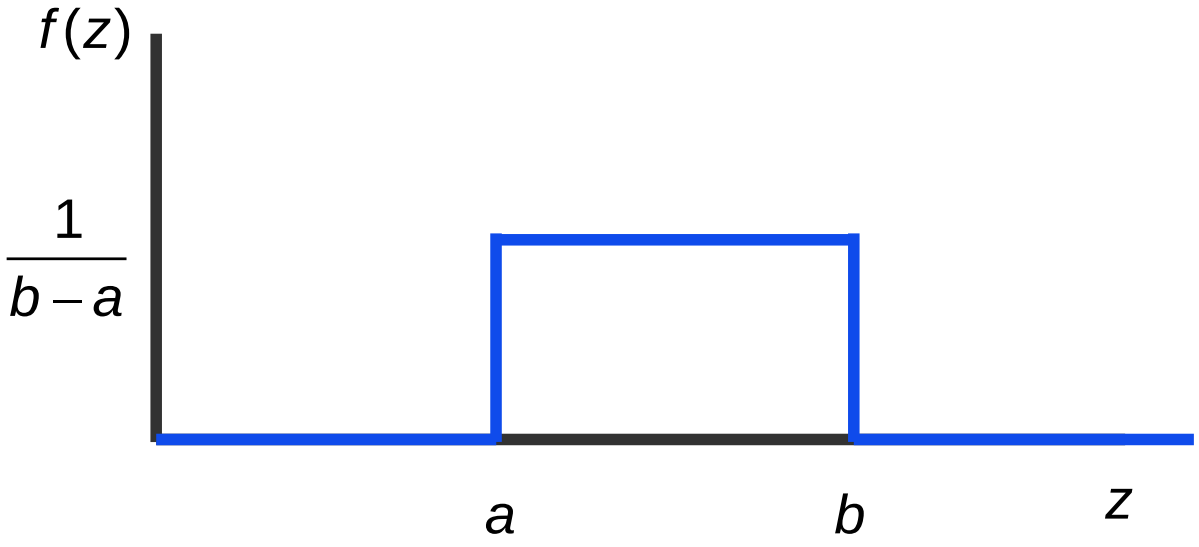
- The first two moments of Z are given by

$$m_1 = E(Z) = \int_a^b \frac{z}{b-a} dz = \left[\frac{z^2}{2(b-a)} \right]_a^b$$

$$= \frac{b^2 - a^2}{2(b-a)} \quad = \frac{b+a}{2}$$

and $$m_2 = E(Z^2) = \int_a^b \frac{z^2}{b-a} dz = \left[\frac{z^3}{3(b-a)} \right]_a^b$$

$$= \frac{b^3 - a^3}{3(b-a)} \quad = \frac{b^2 + ab + a^2}{3}$$



$$f(z) = \begin{cases} \frac{1}{b-a} & z \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

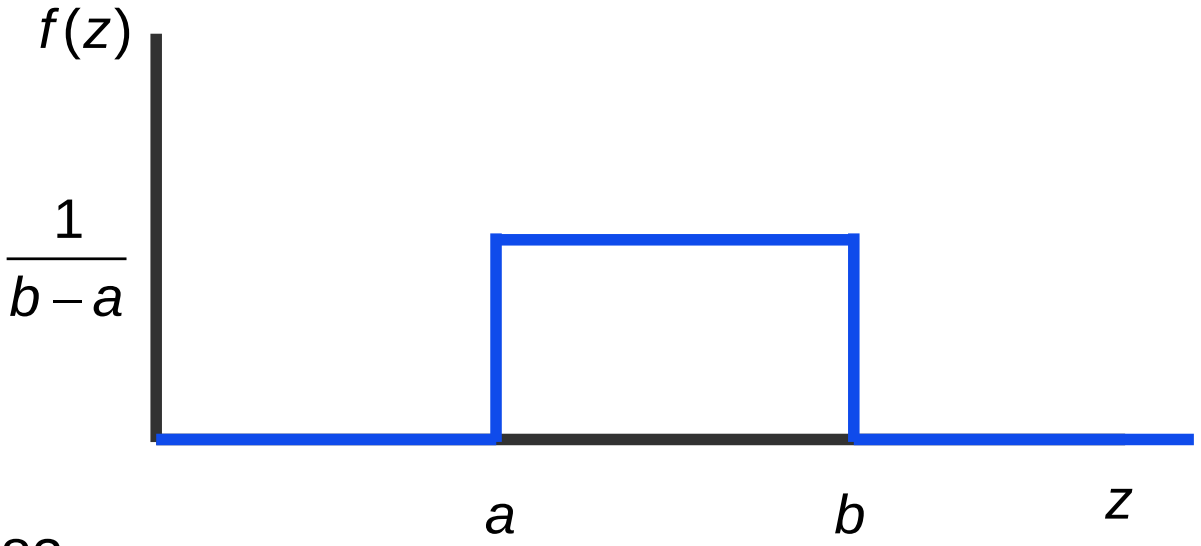
Method of Moments

- Given a set of observations of Z,

$$\{z_1, z_2, z_3, z_4, z_5\} = \{0.5,\ 7.1,\ 4.2,\ 3.7,\ 2.2\}$$

we calculate the first two sample moments as

$$s_1 = \frac{1}{5}(0.5 + 7.1 + 4.2 + 3.7 + 2.2) = \frac{17.7}{5} = 3.54$$



$$\text{and } s_2 = \frac{1}{5}(0.5^2 + 7.1^2 + 4.2^2 + 3.7^2 + 2.2^2) = \frac{86.83}{5} = 17.366$$

$$f(z) = \begin{cases} \dfrac{1}{b-a} & z \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

- We match these to the moments of the distribution of Z.

Method of Moments

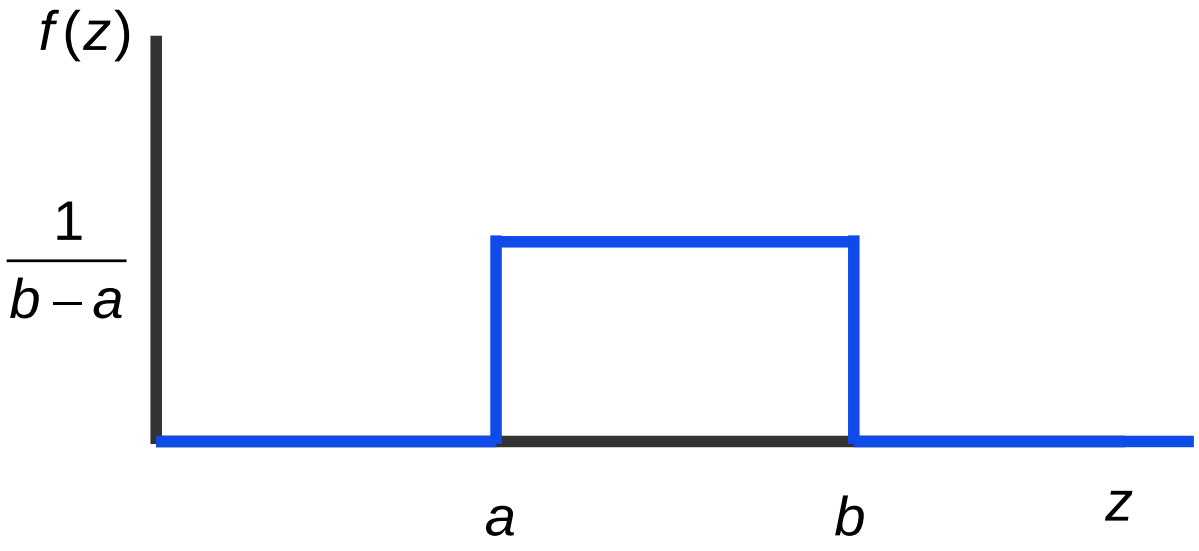
- $m_1 = \frac{b+a}{2} = s_1 = 3.54$ and

  $m_2 = \frac{b^2+ab+a^2}{3} = s_2 = 17.366$

- We solve these simultaneously.

- $\frac{b+a}{2} = 3.54$ hence $a = (7.08 - b)$.

- $\frac{b^2+ab+a^2}{3} = 17.366 = \frac{b^2+(7.08-b)b+(7.08-b)^2}{3}$



$$f(z) = \begin{cases} \frac{1}{b-a} & z \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

Method of Moments

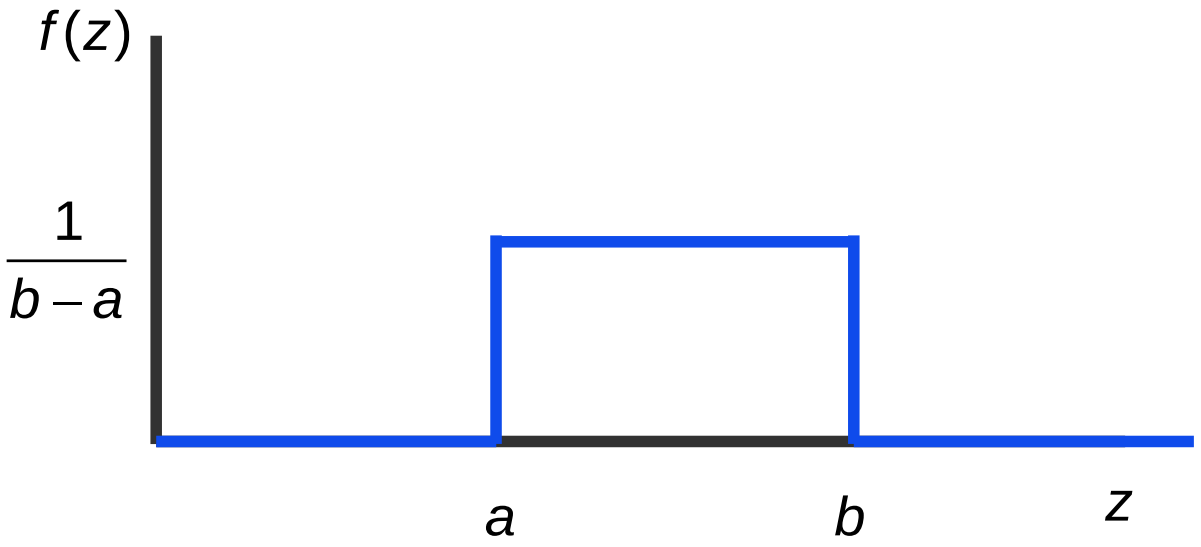
- $17.366 = \dfrac{b^2 + (7.08 - b)b + (7.08 - b)^2}{3}$

- $52.098 = b^2 + 7.08^2 - 7.08b$

  hence $52.098 = b^2 - 7.08b - 1.9716 = 0.$

- Via the discriminant, we obtain that

$$b = \frac{7.08 \pm \sqrt{7.08^2 + 4(1.9716)}}{2}$$

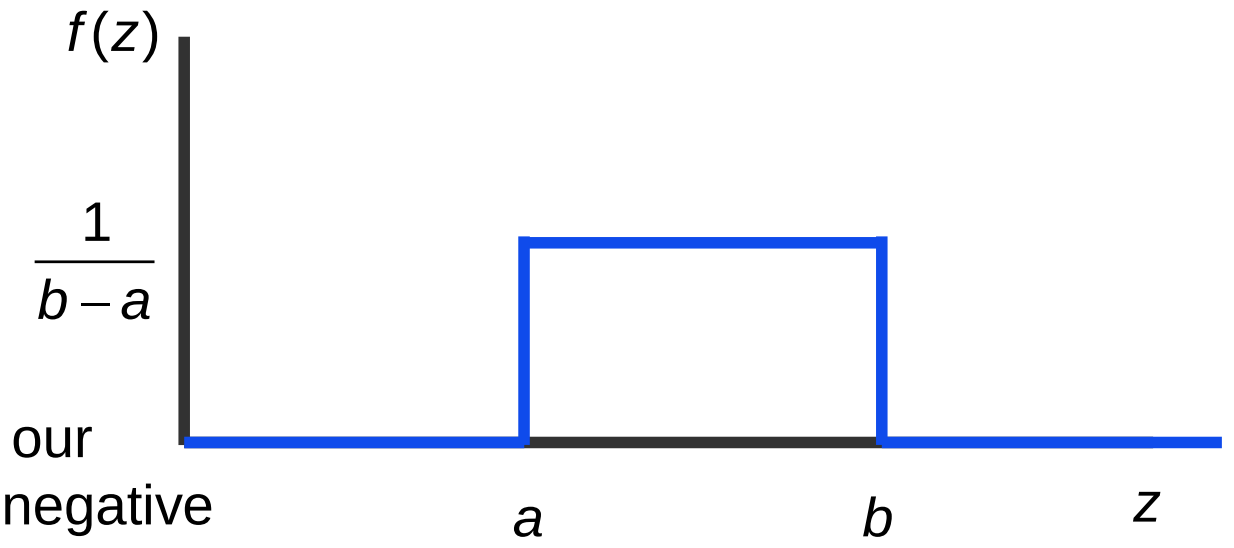


$$f(z) = \begin{cases} \dfrac{1}{b-a} & z \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

Method of Moments

- $b = \dfrac{7.08 \pm \sqrt{7.08^2 + 4(1.9716)}}{2}$
- $b \approx -0.2683$ or 7.3483.
- We know, though, that $Z \in [a,b]$ and that all of our observed values are greater than zero, so the negative root of the equation is not a reasonable estimate.
- We conclude that $\hat{b}_{MM} \approx 7.3483$

  hence $\hat{a}_{MM} \approx 7.08 - 7.3483 = -0.2683$.



$$f(z) = \begin{cases} \dfrac{1}{b-a} & z \in [a,b] \\ 0 & \text{otherwise} \end{cases}$$

A Note on Estimates

- We usually will obtain different estimates of the same parameters, given the same observations but a different choice of moments to match.

- For example, consider a Poisson variable $N \sim Poi(\lambda)$ with probability mass function

$$P(N=k)=\begin{cases}\dfrac{e^{-\lambda}\lambda^{k}}{k!} & k \in \{0,1,2,3,\ldots\} \\ 0 & \text{otherwise}\end{cases}$$

- The first moment of N is therefore $m_1 = E(N) = \lambda$ and the second central moment of N is $c_2 = Var(N) = \lambda$.

A Note on Estimates

- For $N \sim Poi(\lambda)$, the mean and variance of N are both equal to the rate parameter λ.

- Unless they are coincidentally equal, if we estimate λ using the first moment, we will obtain one value and if we use the second moment, we will get a completely different estimate of the same parameter.

- Consider observations $\{n_1, n_2, n_3, n_4, n_5\} = \{3, 2, 5, 4, 3\}$ of N.

$$s_1 = \frac{17}{5} = 3.4 \text{ and } cs_2 = \frac{1}{5}\left((3-3.4)^2 + (2-3.4)^2 + (5-3.4)^2 + (4-3.4)^2 + (3-3.4)^2\right) = \frac{5.2}{5} = 1.04.$$

- Using the first moment to estimate the rate would give $\hat{\lambda}_{MM1} = 3.4$ and using the second moment would give $\hat{\lambda}_{MM2} = 1.04$.

A Note on Estimates

- Estimates can sometimes give values which we know cannot be the true parameters of the underlying distribution.

- Consider that we knew that a fair coin was flipped an unknown number of times and that the variable Y was defined as the number of Heads observed. Clearly $Y \sim Bin(n, 0.5)$ where n is an unknown positive integer.

- Given a sample $\{y_1, y_2, y_3, y_4, y_5\} = \{3, 2, 5, 4, 3\}$ of Y, applying the method of moments would give $m_1 = 0.5n = s_1 = \frac{17}{5}$ so $\hat{n}_{MM} = 6.8$ even though we know that the true n is an integer.

Likelihood Functions

- Let X be a variable with probability density function $f(x)$ which depends on one or more parameter(s), $\boldsymbol{\Theta} = (\theta_1, \theta_2, ... \theta_m)$

- Given a set of independent observations of X, $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$ the **likelihood function** is defined as $L(\boldsymbol{X} \mid \boldsymbol{\Theta}) = f(x_1 \mid \boldsymbol{\Theta}) \times f(x_2 \mid \boldsymbol{\Theta}) \times ... f(x_n \mid \boldsymbol{\Theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\Theta})$

- That is, the likelihood function is the product of the probability densities of each observation given the true values of the parameters.

- A similar definition holds for discrete variables, only using the product of probability masses.

Maximum Likelihood Estimation

- The likelihood function can be used to obtain estimates of the parameters $\boldsymbol{\Theta} = (\theta_1, \theta_2, \ldots \theta_m)$.
- The **maximum likelihood estimator (mle)** of $\hat{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$ is defined as $\arg\max_{\boldsymbol{\Theta}} \left(L(\boldsymbol{X} \mid \boldsymbol{\Theta}) \right)$.
- In other words, the maximum likelihood estimator of the parameters is obtained by finding the parameter values for which the observed data have greatest likelihood.
- Note, using maximum likelihood estimation does not give the parameters which are most likely given the observations, rather gives the parameter values for which those observations were most likely.

Maximum Likelihood Estimation

- To maximise $L(\boldsymbol{X} \mid \boldsymbol{\Theta}) = \prod_{i=1}^{n} f(x_i \mid \boldsymbol{\Theta})$, we usually consider taking the logarithm of both sides.
- We define the **loglikelihood** function as $\ell(\boldsymbol{X} \mid \boldsymbol{\Theta}) = \ln L(\boldsymbol{X} \mid \boldsymbol{\Theta})$.
- Since the logarithm function is a one-to-one transformation, we know that $L(\boldsymbol{X} \mid \boldsymbol{\Theta})$ is maximised wherever $\ell(\boldsymbol{X} \mid \boldsymbol{\Theta})$ is maximised.
- That is $\arg\max_{\boldsymbol{\Theta}} \left(L(\boldsymbol{X} \mid \boldsymbol{\Theta})\right) = \arg\max_{\boldsymbol{\Theta}} \left(\ell(\boldsymbol{X} \mid \boldsymbol{\Theta})\right)$

Loglikelihood Functions

- The reason we maximise the loglikelihood instead of the likelihood is that this turns the problem from one of differentiating a product of functions into one of differentiating a sum, which is much simpler.

- Consider the one parameter problem based on $L(\boldsymbol{X} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$.

- By the product rule, we have $\frac{\partial}{\partial \theta} L(\boldsymbol{X} \mid \theta) = \frac{\partial}{\partial \theta} \prod_{i=1}^{n} f(x_i \mid \theta) = \sum_{j=1}^{n} \left[\frac{\partial}{\partial \theta} f(x_j \mid \theta) \prod_{\substack{i=1 \\ j \neq i}}^{n} f(x_i \mid \theta) \right]$.

- This is clearly quite messy and so we instead consider taking logarithms first.

Loglikelihood Functions

- $L(\boldsymbol{X} \mid \theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$ hence $\ell(\boldsymbol{X} \mid \theta) = \sum_{i=1}^{n} \ln f(x_i \mid \theta)$.
- The derivative is now much tidier, $\frac{\partial}{\partial \theta} \ell(\boldsymbol{X} \mid \theta) = \frac{\partial}{\partial \theta} \sum_{i=1}^{n} \ln f(x_i \mid \theta) = \sum_{i=1}^{n} \frac{\partial}{\partial \theta} \ln f(x_i \mid \theta)$.

Maximum Likelihood Estimation

- Consider now the problem of estimating the probability of a coin landing Heads.

- We say that the number of Heads on a given flip $\sim Bern(p)$ where p is an unknown parameter.

- Say we have observed ten flips and the number of Heads on each is $\{1,0,0,1,1,1,1,0,1,1\}$.

- We know that the probability mass function of $X \sim Bern(p)$ is $P(X = k) = \begin{cases} p & k = 1 \\ 1-p & k = 0 \\ 0 & \text{otherwise} \end{cases}$

Maximum Likelihood Estimation

- $P(X=k)=\begin{cases} p & k=1 \\ 1-p & k=0 \\ 0 & \text{otherwise} \end{cases}$ hence the likelihood of the

the sample $\boldsymbol{X}=\{1,0,0,1,1,1,1,0,1,1\}$ is $L(\boldsymbol{X}\mid p)=p(1-p)(1-p)ppppp(1-p)pp$.

- We therefore obtain the mle for p by finding the p which maximises

$$L(\boldsymbol{X}\mid p)=p^7(1-p)^3 \text{ or alternatively } \ell(\boldsymbol{X}\mid p)=7\ln p+3\ln(1-p).$$

- The point at which the derivative of either of these is zero is clearly a maximum, not a minimum, since it is obvious that setting p too large or too small will give a smaller likelihood. For example, $p=0$ and $p=1$ would both make the likelihood zero.

Maximum Likelihood Estimation

- Differentiation gives

$$\frac{\partial}{\partial p}\ell(\boldsymbol{X} \mid p) = \frac{\partial}{\partial p}\left(7\ln p + 3\ln(1-p)\right)$$

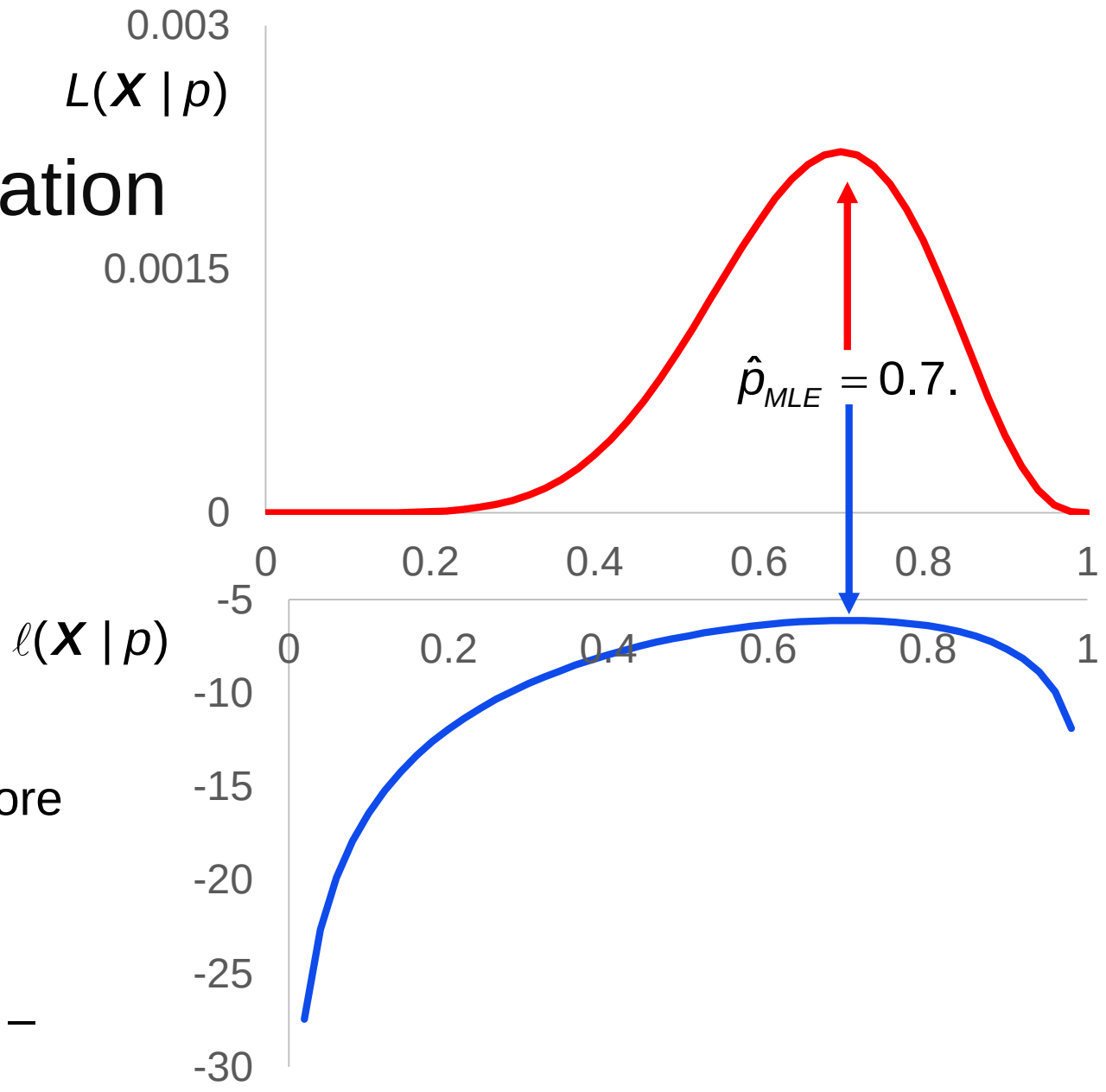
hence $\frac{\partial}{\partial p}\ell(\boldsymbol{X} \mid p) = \frac{7}{p} - \frac{3}{1-p}$

$= 0$ when $p = 0.7$.

- The maximum likelihood estimate is therefore

$\hat{p}_{MLE} = 0.7.$

- This is the same as the estimate which we would obtain by matching the first moment – the simple sample proportion.



Maximum Likelihood Estimation

- Consider a set $\boldsymbol{X} = \{x_1, x_2, ..., x_n\}$ of independent realisations of $X \sim N(\mu, \sigma^2)$ where both the mean and variance are unknown.

- The likelihood function associated with this sample is therefore

$$L(\boldsymbol{X} \mid \mu, \sigma^2) = \prod_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right) = \left(\sqrt{2\pi}\right)^{-n} \left(\sqrt{\sigma^2}\right)^{-n} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}$$

- This gives a loglikelihood function of $\ell(\boldsymbol{X} \mid \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}$.

Maximum Likelihood Estimation

- We can obtain the maximum likelihood estimates of the two parameters by maximising the loglikelihood function $\ell(\boldsymbol{X} \mid \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}$.
- $\frac{\partial \ell(\boldsymbol{X} \mid \mu, \sigma^2)}{\partial \mu} = -\sum_{i=1}^{n} \frac{2\mu - 2x_i}{2\sigma^2}$.
- $\frac{\partial \ell(\boldsymbol{X} \mid \mu, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2(\sigma^2)^2}$.
- Setting both of these derivatives to zero, we maximise the loglikelihood.

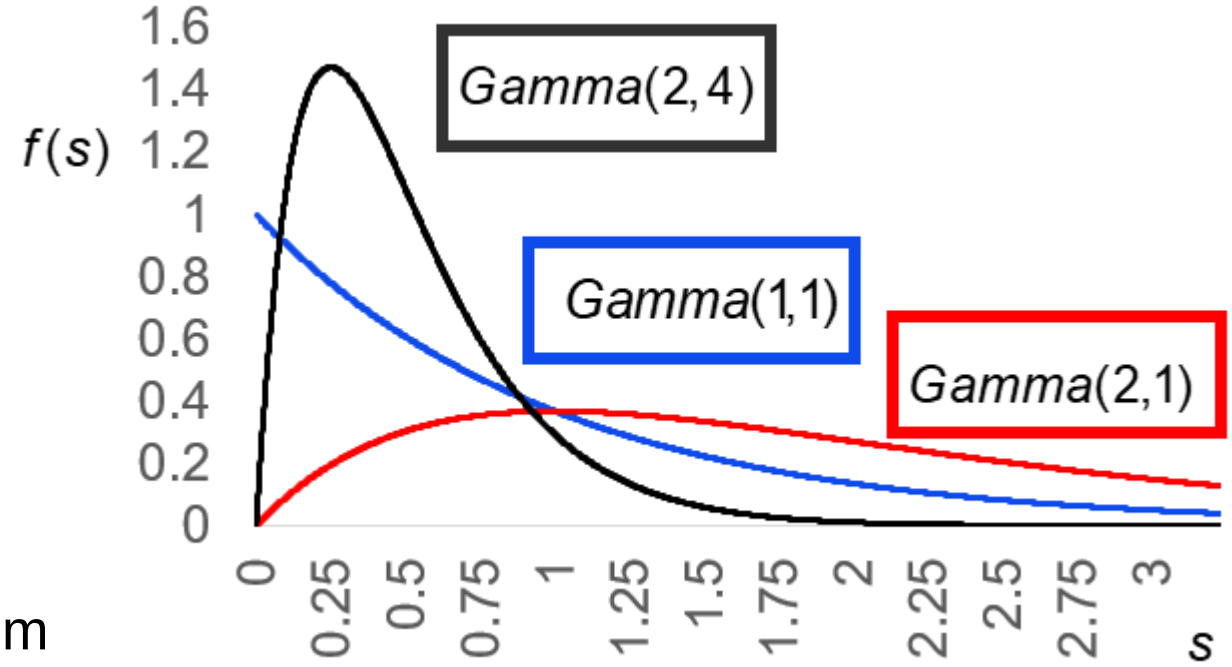
Maximum Likelihood Estimation

- $-\sum_{i=1}^{n} \frac{2\mu - 2x_i}{2\sigma^2} = 0$ hence $2n\mu - 2\sum_{i=1}^{n} x_i = 0.$
- This gives $\hat{\mu}_{MLE} = \frac{\sum_{i=1}^{n} x_i}{n}$.
- $-\frac{n}{2\sigma^2} + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2(\sigma^2)^2} = 0$ hence $-n\sigma^2 + \sum_{i=1}^{n} (x_i - \mu)^2 = 0.$
- This gives $\hat{\sigma}^2_{MLE} = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu}_{MLE})^2}{n}$.
- These are also equal to the first sample moment ($\hat{\mu}_{MLE} = s_1$) and second central sample moment ($\hat{\sigma}^2_{MLE} = cs_2$.)

Maximum Likelihood Estimation

- Recall the gamma distribution with density

function $f(s) = \begin{cases} \dfrac{\beta^{\alpha} s^{\alpha-1} e^{-\beta s}}{\Gamma(\alpha)} & s \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$.



- Given a sample of independent observations of S, $\{s_1, s_2, \ldots, s_n\}$ we can obtain the maximum likelihood estimators for both parameters – α and β.

Maximum Likelihood Estimation

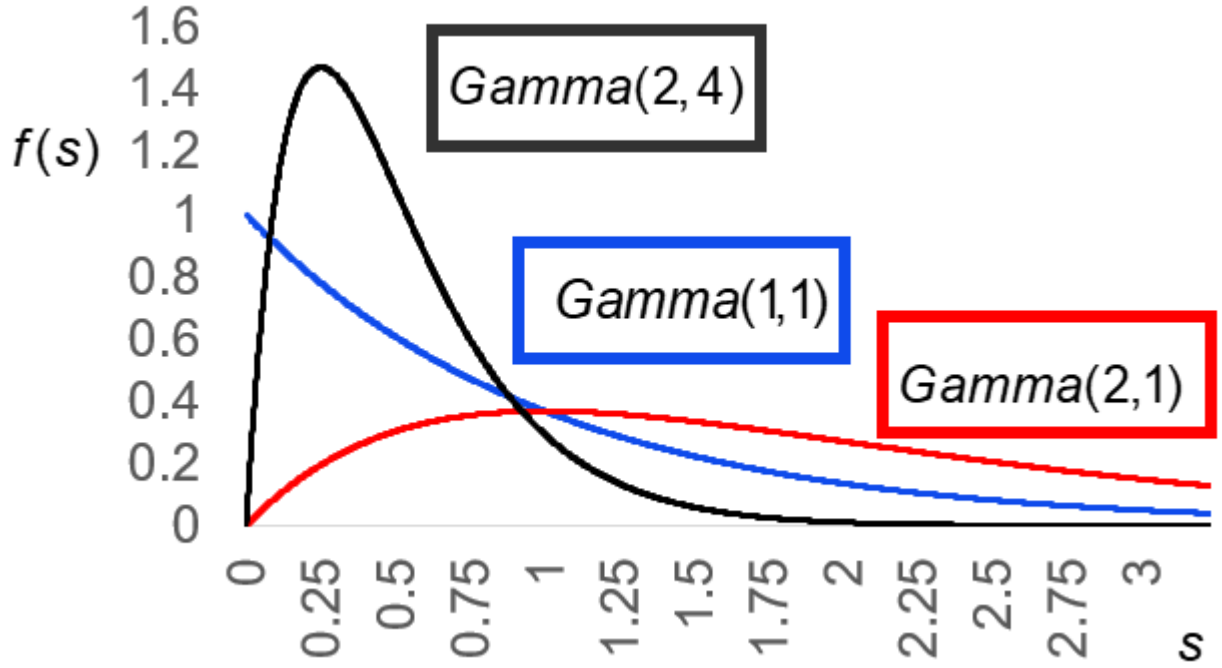
- The likelihood of a sample $\boldsymbol{S} = \{s_1, s_2, \ldots, s_n\}$

  is therefore $L(\boldsymbol{S} \mid \alpha, \beta) = \prod_{i=1}^{n} \frac{\beta^{\alpha} s_i^{\alpha-1} e^{-\beta s_i}}{\Gamma(\alpha)}$.

- $L(\boldsymbol{S} \mid \alpha, \beta) = \frac{\beta^{n\alpha}}{(\Gamma(\alpha))^n} \prod_{i=1}^{n} s_i^{\alpha-1} e^{-\beta \sum_{i=1}^{n} s_i}$.

- The loglikelihood is then

$$\ell(\boldsymbol{S} \mid \alpha, \beta) = \sum_{i=1}^{n} (\alpha - 1) \ln s_i - \beta \sum_{i=1}^{n} s_i + n\alpha \ln(\beta) - \ln\left(\Gamma(\alpha)^n\right).$$



Maximum Likelihood Estimation

- Differentiating

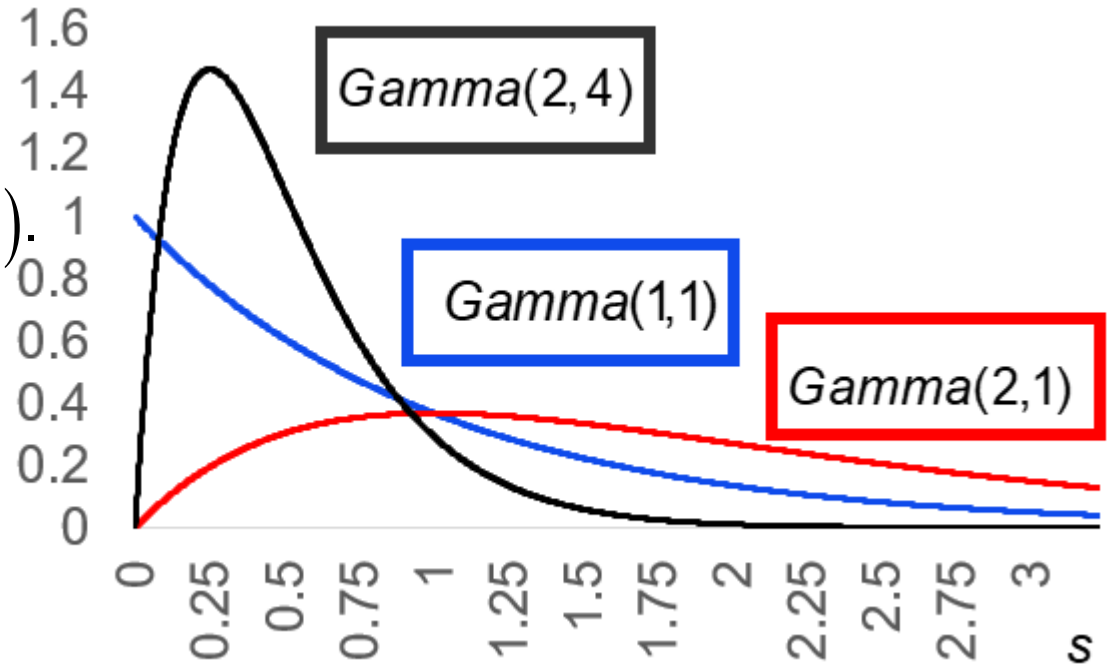
$$\ell(\boldsymbol{s} \mid \alpha, \beta) = \sum_{i=1}^{n} (\alpha - 1) \ln s_i - \beta \sum_{i=1}^{n} s_i + n\alpha \ln(\beta) - \ln\left(\Gamma(\alpha)^n\right).$$

gives

$$\frac{\partial \ell}{\partial \beta} = -\sum_{i=1}^{n} s_i + \frac{n\alpha}{\beta}$$

and

$$\frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} \ln s_i + n \ln(\beta) - n \frac{\Gamma'(\alpha)}{\Gamma(\alpha)}.$$

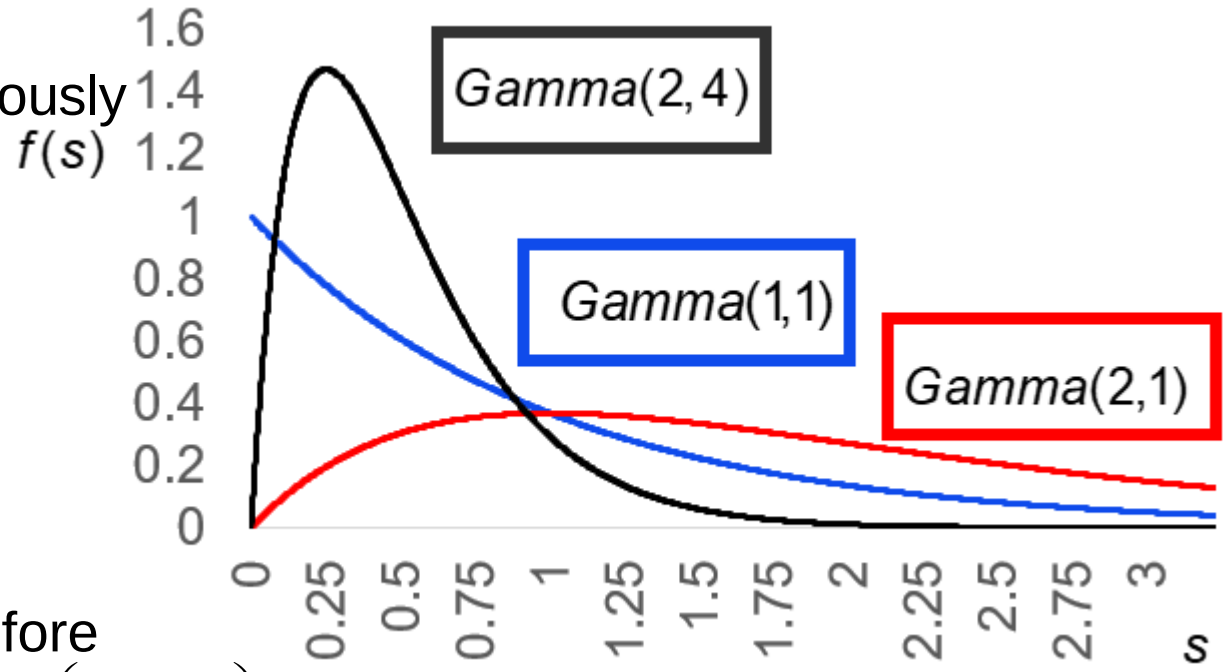


Maximum Likelihood Estimation

- Setting both derivatives to zero we simultaneously solve

$$\frac{\partial \ell}{\partial \beta} = -\sum_{i=1}^{n} s_i + \frac{n\alpha}{\beta} = 0$$

$$\text{and } \frac{\partial \ell}{\partial \alpha} = \sum_{i=1}^{n} \ln s_i + n\ln(\beta) - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} = 0.$$



- The maximum likelihood estimators are therefore

$$\hat{\beta}_{MLE} = \frac{n\hat{\alpha}_{MLE}}{\sum_{i=1}^{n} s_i} \text{ where } \hat{\alpha}_{MLE} \text{ satisfies } \sum_{i=1}^{n} \ln s_i + n\ln\left(\frac{n\hat{\alpha}_{MLE}}{\sum_{i=1}^{n} s_i}\right) - n\frac{\Gamma'(\hat{\alpha}_{MLE})}{\Gamma(\hat{\alpha}_{MLE})} = 0.$$

- This nonlinear equation may need to be solved numerically.