# Introduction to the Theory of Statistics Part 2

## PM522b

Meredith Franklin

Division of Biostatistics, University of Southern California

Slides 3, 2020

# Topics covered

- Point Estimation
- Methods for finding estimators:
  - Maximum Likelihood Estimators
  - Numerical approach to finding Maximum Likelihood Estimators
  - Method of Moments Estimators
  - Bayes Estimators
  - EM Algorithm

# Estimators and estimation

- As we mentioned at the beginning of this course, the purpose of statistics is to provide us with a way to make inference about a population based on information we get from a sample. Populations are characterized by parameters, and we wish to estimate and make inference about these parameters.
- A point estimator is any function of a sample $X_1, ..., X_n$, so a statistic is a point estimator. A point estimator of a parameter $\theta$ is denoted by $\hat{\theta}$.
- Point estimation seeks to find the quantity $\hat{\theta}$ (depending on the values $X_1, ..., X_n$ sampled from a population described by a pdf or pmf $f(x|\theta)$) that is the best representation of the unknown parameter $\theta$.

# Methods for finding estimators

- So far we have looked as estimators based on our intuition.
- For example, we have used the sample mean as an estimator of the parameter $\mu$ for our population mean
- For more complicated situations we need a more rigorous way of estimating parameters
- Methods for finding estimators for one or more population parameters:
  - The *method of maximum likelihood* (ML) is the most popular procedure
  - The *method of moments* (MM) is a simple procedure

# Likelihood

## The Likelihood Function

Suppose $X_1, X_2, ..., X_n \sim f(\mathbf{x}|\theta)$ (iid), given that $\mathbf{X} = \mathbf{x}$ is observed (data). Then, the likelihood function is:

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta)$$

By the assumed independence, the joint distribution of $X_1, X_2, ..., X_n$ is characterized by:

$$f(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

Where "independence means multiply". And more generally for $k$ unknown parameters,

$$f(x_1, x_2, ..., x_n|\theta_1, \theta_2, ..., \theta_k) = \prod_{i=1}^{n} f(x_i|\theta_1, \theta_2, ..., \theta_k)$$

# Likelihood

## The Likelihood Function, con't

Thus, the likelihood function can be more generally defined by:

$$L(\theta_1, \theta_2, ..., \theta_k|\mathbf{x}) = f(x_1, x_2, ..., x_n|\theta_1, \theta_2, ..., \theta_k) = \prod_{i=1}^{n} f(x_i|\theta_1, \theta_2, ..., \theta_k)$$

The distinction between the likelihood function and the pdf is in what is fixed and what is varying. When we consider the pdf $f(\mathbf{x}|\theta)$, $\theta$ is fixed and $\mathbf{x}$ is variable. When we consider the likelihood $L(\theta|\mathbf{x})$, $\mathbf{x}$ is the observed sample which is fixed and $\theta$ is varying over all possible parameter values.

# Likelihood

## Using the likelihood function: a simple example

Given an oddly shaped coin, we don't know the probability of getting heads $\theta$. We perform an experiment of only one flip of the coin and observe a head, so $x = 1$ and $\theta \in [0, 1]$

$$\begin{aligned} L(\theta|x) &= f(x|\theta) = \theta^x(1-\theta)^{1-x} \\ L(\theta|1) &= \theta^1(1-\theta)^{1-1} \\ &= \theta \end{aligned}$$

Suppose we flip it four times and get $x = \{1, 0, 1, 1\}$

$$\begin{aligned} L(\theta|1,0,1,1) &= \theta^1(1-\theta)^{1-1}\theta^0(1-\theta)^{1-0}\theta^1(1-\theta)^{1-1}\theta^1(1-\theta)^{1-1} \\ &= \theta^3(1-\theta) \end{aligned}$$

# Likelihood

The normal likelihood:

- Suppose $X_1, X_2, ..., X_n$ are iid $N(\mu, \sigma^2)$ then

$$\begin{aligned} L(\mu, \sigma^2|x) &= \prod_{i=1}^{n} f(x_i|\mu, \sigma^2) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i-\mu)^2/2\sigma^2} \\ &= (2\pi\sigma^2)^{-n/2} e^{-(1/2\sigma^2)\sum_{i=1}^{n}(x_i-\mu)^2} \end{aligned}$$

- Can be simplified to $L(\mu, \sigma^2|x) = \sigma^{-n} e^{-\sum_{i=1}^{n}(x_i-\mu)^2/2\sigma^2}$
- Note that we dropped the multiplicative term that did not contain the unknown parameter from the likelihood function.
- Generally, if $L(\theta|x) = h(x)g(x|\theta)$ (recall factorization theorem) we can drop the term $h(x)$.

# Likelihood

- The other aspect of the likelihood principle is in relating relative evidence of one value of an unknown parameter to another
- We can compare the likelihood function at two parameter points
- If $L(\theta_1|x) > L(\theta_2|x)$ then $\theta_1$ is more likely to have been responsible for producing the observed $X_1, ..., X_n$

# Likelihood

## A simple example, con't

In our previous experiment, we flipped the coin four times and got $x = \{1, 0, 1, 1\}$

$$L(\theta|1, 0, 1, 1) = \theta^3(1 - \theta)$$

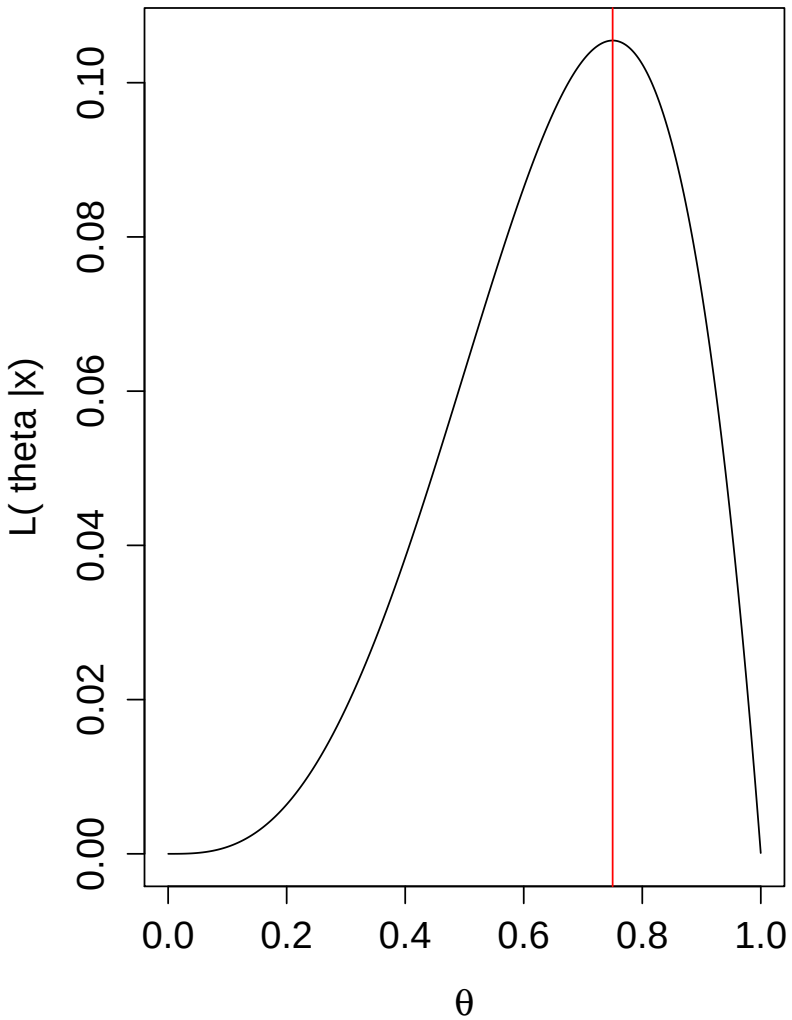
Consider the coin is fair $L(0.5|1, 0, 1, 1)$ versus unfair where we only have 25% chance of getting heads $L(0.25|1, 0, 1, 1)$

$$\frac{L(0.5|1,0,1,1)}{L(0.25|1,0,1,1)} = 5.33$$

There is over 5 times as much evidence supporting the hypothesis that $\theta = 0.5$ over $\theta = 0.25$

This is a relative measure of two possible parameter points, but how do we assess all possible values of $\theta$? We can visualize this with a likelihood plot.

# Likelihood




This plot illustrates 1000 possible parameter values from 0.001 to 0.999 versus the likelihood function evaluated at these parameter values. The red line indicates the value of $\theta$ where the likelihood function (given the observed 4-flip data) $L(\theta|1,0,1,1) = \theta^3(1-\theta)$ is maximized. The $\text{argmax}_\theta L = 0.105$ at $\theta = 0.75$

# Likelihood

- The likelihood principle also states that two likelihood functions contain the same information about $\theta$ if they are proportional to each other
- If $x$ and $y$ are two sample points such that $L(\theta|x)$ is proportional to $L(\theta|y)$ then a constant $C(x, y)$ exists such that

$$L(\theta|x) = C(x, y)L(\theta|y) \text{ for all } \theta$$

- Then conclusions about $\theta$ drawn from $x$ and $y$ should be identical
- This is obvious for $C(x, y) = 1$, and in this case the two sample points result in the same likelihood function and thus contain the same information about $\theta$

# Likelihood

- To put this in a practical context:
  - We saw previously how we compare two parameter values, for our example $L(\theta_1|x) = 5.33L(\theta_2|x)$, so $\theta_1$ was 5.33 times more likely than $\theta_2$
  - If $L(\theta|x) = C(x, y)L(\theta|y)$ for all $\theta$ is also true, then $L(\theta_1|y) = 5.33L(\theta_2|y)$
  - So whether we observe $x$ or $y$, $\theta_1$ is 5.33 times more likely than $\theta_2$

# Method of Maximum Likelihood

- In light of evaluating the likelihood function at various parameter values of $\theta$, we see that it can provide a ranking of their possible values
- We would like the value of $\theta$ where the likelihood reaches a maximum
- This is the value of $\theta$ that is most supported by the data, the maximum likelihood estimate of $\theta$

$$\text{MLE} = \text{argmax}_\theta L(\theta|\mathbf{x}) = \text{argmax}_\theta \log L(\theta|\mathbf{x}) = \hat{\theta}(\mathbf{X})$$

- Note we work with the log of the likelihood in finding MLEs

## MLE, CB 7.2.4

For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $L(\theta|\mathbf{x})$ attains a maximum as a function of $\theta$ with $\mathbf{x}$ held fixed. A maximum likelihood estimator (MLE) of the parameter $\theta$ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.

# Method of Maximum Likelihood

- Issues in finding the MLE of $\theta$ include:
  - the general problem of finding the maximum of a function, i.e. in finding the global maximum and ensuring that the global maximum has been found.
  - the sensitivity of the MLE to changes in the data.
- The problem in finding the maximum can be tricky and may require numerical methods.
- In the simplest cases, we find the MLE by writing down the likelihood, taking the log then differentiating with respect to the parameter of interest and setting this equal to zero:

$$\frac{\partial log L(\theta|x)}{\partial \theta} = 0$$

# Method of Maximum Likelihood

- The likelihood function must be differentiable with respect to the unknown parameter of interest.
- The first derivative being 0 is only a necessary condition for a maximum, not a sufficient one.
- Taking the zeros of the first derivative locate extreme points on the interior of the domain of a function. If extrema exist on the boundary of the function then the first derivative may not be zero. So the boundaries must be checked separately for extrema.
- With the first derivative equal to zero it is possible to have local maxima or minima, global maxima or minima or inflection points.

# Method of Maximum Likelihood

- The maximization of the likelihood function represents a necessary condition for the existence of the MLE, but an additional condition must also be met to ensure that we are estimating a maximum. Necessary conditions for a point to be a global maximum:

$$\ell''(\theta) \leq 0$$

  We evaluate this at the candidate value for the MLE.
- Furthermore, the boundaries (limits at $+\infty$ and $-\infty$) must be checked.
- Finding global maxima can be difficult especially when there are multiple parameters.

# Maximum Likelihood Estimation

## Example, Binomial distribution

The likelihood function is $L(\theta|x_i) = \prod_{i=1}^{n} \theta_i^x (1-\theta)^{n-x_i}$

$$\log L(\theta|x_i) = \sum_{i=1}^{n} x_i \log\theta + (n - \sum_{i=1}^{n} x_i)\log(1-\theta)$$

$$\frac{\partial \log L(\theta|x_i)}{\partial\theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - \frac{n - \sum_{i=1}^{n} x_i}{1-\theta}$$

$$\frac{\partial^2 \log L(\theta|x_i)}{\partial\theta^2} = -\frac{\sum_{i=1}^{n} x_i}{\theta^2} - \frac{n - \sum_{i=1}^{n} x_i}{(1-\theta)^2}$$

Note that the derivatives exist only when $0 < \theta < 1$. Setting

$$\frac{\partial \log L(\theta|x_i)}{\partial\theta} = 0$$

# Maximum Likelihood Estimation

## Example, Binomial distribution con't

When $\sum_{i=1}^{n} x_i = 0$ or $\sum_{i=1}^{n} x_i = n$ we get $\hat{\theta} = 0$ or 1 and are thus on the boundary of the parameter space (endpoints).

$$\log L(\theta|x_i) = n \log(1-\theta) \qquad \sum_{i=1}^{n} x_i = 0$$

$$\log L(\theta|x_i) = n \log(\theta) \qquad \sum_{i=1}^{n} x_i = n$$

# Maximum Likelihood Estimation

## Example, Binomial distribution con't

Since $\log\theta$ is an increasing function, $\log(1-\theta)$ is a decreasing function. Thus, when $\theta$ takes on its smallest possible value, 0 (when $\sum_{i=1}^{n} x = 0$) the likelihood is maximized. In other words, $\hat{\theta} = \sum_{i=1}^{n} x_i/n$ is also the MLE when $\sum_{i=1}^{n} x_i = 0$.

Similarly, when $\theta$ takes on the largest possible value, n, $n\log(\theta)$ is maximized and again $\hat{\theta} = \sum_{i=1}^{n} x_i/n$ is also the MLE when $x = n$.

In all cases, $\hat{\theta} = \sum_{i=1}^{n} x_i/n$.

# Maximum Likelihood Estimation

- Maximum likelihood estimation for multi-parameter distributions is approached similarly to those with a single parameter, but can be technically more difficult due to more complicated likelihood functions
- We have to use multivariate differentiation: all partial derivatives (first and second) must exist and be continuous
  - the vector of first partial derivatives is called the *gradient* and is denoted $\nabla \ell(\theta)$
  - the matrix of second partial derivatives is called the *Hessian* and is denoted $\nabla^2 \ell(\theta)$
- We also have to generalize the conditions for local and global maxima

# Maximum Likelihood Estimation

- The necessary conditions for the multi-parameter case:
  - $\nabla \ell(\theta) = 0$ where $\theta$ is an interior point and a local max
  - $\nabla^2 \ell(\theta)$ is a negative semi-definite matrix
  - the boundaries should also be checked (take limits of the parameter space using the (log) likelihood)

# Maximum Likelihood Estimation

Negative definite and negative semi-definite matrices

- A positive definite matrix is symmetric and all eigenvalues are positive.
- We can check if a matrix is positive definite or semi-definite by looking at the n upper left determinants of the matrix.
- A matrix $M$ is negate semi-definite or negative definite if $-M$ is positive semi-definite or positive definite.
- $M$ has a quadratic form $X^tMX$ and is positive definite if $X^tAX > 0$ for all non-zero vectors ($x \neq 0$). $M$ is positive semi-definite if $X^tAX \geq 0$

Example in class.

# Properties of MLE

Invariance property

- This property states that if $\hat{\theta}$ is the MLE of $\theta$ then $g(\hat{\theta})$ is the MLE of $g(\theta)$
- This is useful as it allows us to estimate a lot of different characteristics of the distribution
- Example, if $\theta$ is the mean of a normal distribution, the MLE of $\sin(\theta)$ is $\sin(\bar{X})$
- This result holds for any function $g$ but is easiest to prove when $g$ is a one-to-one function
- A one-to-one function means that for each value of $\theta$ there is a unique value of $g(\theta)$ and vice versa
- Thus it makes no difference whether we maximize the likelihood as a function of $\theta$ or as a function of $g(\theta)$

# Invariance property of MLE

## Proof of MLE Invariance

If $g$ is a one-to-one function, let $\eta = g(\theta)$
Since $g$ is one-to-one, the inverse function $g^{-1}$ exists and thus $g^{-1}(\eta) = \theta$
The pdf of $f(x|\theta) = f(x|g^{-1}(\eta))$ and the likelihood function in terms of $\eta$ is given by

$$\begin{aligned} L(\eta|x) &= \prod_{i=1}^{n} f(x_i|g^{-1}(\eta)) \\ &= L(g^{-1}(\eta)|x) \end{aligned}$$

and maximizing the likelihoods,

$$\begin{aligned} \sup_{\eta} L(\eta|x) &= \sup_{\eta} L(g^{-1}(\eta)|x) \\ &= \sup_{\theta} L(g^{-1}(\eta)|x) \end{aligned}$$

# Invariance property of MLE

## Proof of MLE Invariance con't

Thus to maximize $L(\eta|x)$, choose $\hat{\eta}$ such that $g^{-1}(\hat{\eta}) = \hat{\theta}$

i.e. take $\hat{\eta} = g(\hat{\theta})$ therefore showing that the MLE of $g(\theta)$ is $g(\hat{\theta})$

# Invariance property of MLE

- In many cases we have functions that are not one-to-one, for example we may wish to estimate the square of the normal mean, $\theta^2$
- $\theta \rightarrow \theta^2$ is not one-to-one
- When $g$ is not one-to-one, the new parameterization $\eta = g(\theta)$ does not give enough information to define $f(x|\eta)$ and we cannot define the likelihood.
- See additional notes for Berk (1967) proof when $g$ is not one-to-one

## Proof of MLE Invariance con't

We define the induced likelihood function

$$\sup_{\eta} L^*(\eta|x) = \sup_{\{\theta:g(\theta)=\eta\}} L(\theta|x)$$

and the value of $\hat{\eta}$ that maximizes the induced likelihood $L^*(\eta|x)$ is called the MLE of $\eta = g(\theta)$. The maxima of $L^*$ and $L$ coincide.

# Numerical Solutions for MLE

- In the examples shown thus far we found a closed form solution when maximizing the likelihood function in order to estimate the unknown parameter $\theta$
- However, in many practical problems there may be more than one solution to $\frac{\partial log L(\theta|x)}{\partial \theta} = 0$, or no nice formula for solving for $\hat{\theta}$
- In such cases we use numerical methods as an "optimization strategy" for solving MLEs
- One of the most widely used methods for optimization is the Newton-Raphson Method (often just called Newton's Method)

# Numerical MLE: Newton-Raphson Method

- The idea is to approximate the likelihood with a quadratic function
- Take an initial guess at the unknown parameter value, then adjust it to maximize the quadratic approximation
- The value is adjusted iteratively until it stabilizes
- The quadratic approximation is derived from the Taylor series expansion of a function

## Taylor Series expansion for a function of one variable

The Taylor series expansion of a function $f(x)$ about a point $x_0$ is:

$$f(x) \approx f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2!}(x - x_0)^2 f''(x_0) + ...$$

The higher order terms are ignored as $x - x_0$ gets small and terms beyond the quadratic become less important. To maximize the quadratic approximation for $f(x)$, we take the derivative, set it equal to zero and solve:

$$f'(x) \approx f'(x_0) + (x - x_0)f''(x_0) = 0$$
$$x = x_0 - f'(x_0)/f''(x_0)$$

# Numerical MLE: Newton-Raphson Method

- We translate this to our likelihood by replacing the generic function f with the log likelihood function $\ell$

$$\theta^{(1)} = \theta^{(0)} - \ell'(\theta^{(0)})/\ell''(\theta^{(0)})$$

- We start with an initial guess for the unknown parameter value ($\theta^{(0)}$)
- Calculate $\theta^{(1)}$
- Iterate the computation: $\theta^{(t+1)} = \theta^{(t)} - \ell'(\theta^{(t)})/\ell''(\theta^{(t)})$ until $|\theta^{(t+1)} - \theta^{(t)}|$ is small, i.e. less than some tolerance level such as $1e-8$, or until a maximum number of iterations is reached (e.g. 1000).

# Numerical MLE: Newton-Raphson Method

## Newton Algorithm for the logistic distribution

The pdf for the logistic distribution is $f(x|\theta) = \frac{e^{-(x-\theta)}}{(1+e^{-(x-\theta)})^2}$.

$$\ell(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

$$= n\theta - n\bar{X} - 2\sum_{i=1}^{n} \log(1 + e^{-(x_i-\theta)})$$

$$\ell'(\theta) = n - 2\sum_{i=1}^{n} \frac{e^{-(x_i-\theta)}}{1 + e^{-(x_i-\theta)}}$$

$$\ell''(\theta) = -2\sum_{i=1}^{n} \frac{e^{-(x_i-\theta)}}{(1 + e^{-(x_i-\theta)})^2}$$

Note: there is no formula for the solution to $\ell'(\theta) = 0$ so Newton's method is necessary.

# Numerical MLE: Newton-Raphson Method

## Newton Algorithm for the logistic distribution con't

With $\ell'(\theta)$, $\ell''(\theta)$ and data $x_1, ..., x_n$, we can choose a starting value $\theta^{(0)}$ and plug it in to Newton's algorithm

Example data generated from: `rlogis(40,location=5,scale=1)`
Since we are simulating data for this example, we know the true value of $\theta = 5$
However, we choose to start our algorithm with $\theta^{(0)}$ =`mean(x)`
The algorithm converges quickly to $\hat{\theta} = 4.923$

# Numerical MLE: Newton-Raphson Method

- For distributions with multiple parameters, the approach is the same

## Taylor Series expansion for a function of multiple variables

The Taylor series expansion of a function $f(x)$ about a point $x_0$ is:

$$f(x) \approx f(x_0) + (x - x_0)^T f'(x_0) + \frac{1}{2!}(x - x_0)^T f''(x_0)(x - x_0)^T$$

To maximize the quadratic approximation for $f(x)$, we take the derivative, set it equal to zero and solve:

$$f'(x) \approx f'(x_0) + (x - x_0)^T f''(x_0) = 0$$
$$x = x_0 - [f''(x_0)]^{-1} f'(x_0)$$

Note: $(x - x_0)^T$ is the vector transpose, and $[f''(x_0)]^{-1}$ is the matrix inverse

# Numerical MLE: Newton-Raphson Method

- Now, $\ell(\theta)$ is a function of $p > 1$ parameters, $\theta = (\theta_1, \theta_2, ..., \theta_p)$
- $\ell'(\theta)$ corresponds to the gradient vector of partial (first) derivatives
- $\ell''(\theta)$ corresponds to the hessian matrix of mixed second derivatives

In the notation for the likelihood,

$$\theta^{(1)} = \theta^{(0)} - [\ell''(\theta^{(0)})]^{-1}\ell'(\theta^{(0)})$$

# Numerical MLE: Newton-Raphson Method

## Newton Algorithm for the Gamma distribution

The pdf for the gamma distribution is $f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-(x/\beta)}$.

$$
\begin{aligned}
L(\alpha,\beta|x) &= \prod_{i=1}^{n} \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} e^{-(x_i/\beta)} \\
&= \left(\frac{1}{\Gamma(\alpha)\beta^\alpha}\right)^n e^{(\alpha-1)\sum_{i=1}^n \log x_i} e^{(1/\beta)\sum_{i=1}^n x_i} \\
\ell(\alpha,\beta) &= -n\alpha \log\beta - n\log\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n} \log x_i - \frac{1}{\beta}\sum_{i=1}^{n} x_i \\
\frac{\partial \ell(\alpha,\beta)}{\partial \alpha} &= -n\log\beta - n\psi(\alpha) + \sum_{i=1}^{n} \log x_i \\
\frac{\partial \ell(\alpha,\beta)}{\partial \beta} &= -\frac{n\alpha}{\beta} + \frac{n\bar{X}}{\beta^2}
\end{aligned}
$$

where $\psi(\alpha) = \Gamma'(\alpha)/\Gamma(\alpha)$ is the digamma function.

# Numerical MLE: Newton-Raphson Method

## Newton Algorithm for the Gamma distribution con't

There is no closed form to $\nabla \ell(\alpha, \beta) = 0$.
The hessian matrix:

$$\frac{\partial^2 \ell(\alpha, \beta)}{\partial \alpha^2} = -n\psi'(\alpha)$$
$$\frac{\partial^2 \ell(\alpha, \beta)}{\partial \beta^2} = \frac{n\alpha}{\beta^2} - \frac{2n\bar{X}}{\beta^3}$$
$$\frac{\partial^2 \ell(\alpha, \beta)}{\partial \alpha \partial \beta} = -\frac{n}{\beta}$$

where $\psi'(\alpha)$ is the trigamma function

$$\ell''(\alpha, \beta) = -n \begin{pmatrix} \psi'(\alpha) & \frac{1}{\beta} \\ \frac{1}{\beta} & 2\frac{\bar{X}}{\beta^3} - \frac{\alpha}{\beta^2} \end{pmatrix}$$

When implementing Newton's method, we need the inverse of the hessian. This can be computationally burdensome, and the R function `ginv` can do this for us.

# Moment Generating Functions

The moment generating function (MGF) can be used to generate moments. It also helps in characterizing a distribution.

## Definition of MGF

The moment generating function of $X$, $M_X(t)$ is defined to be $E(e^{tX})$. It exists if there is a positive constant $h > 0$ such that $M_X(t)$ is finite for $|t| \leq h$.

For a discrete random variable with pmf $P(X)$

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \sum_x e^{tx} P(X = x) \end{aligned}$$

For a continuous random variable with pdf $f(X)$

$$\begin{aligned} M_X(t) &= E(e^{tX}) \\ &= \int_{-\infty}^{\infty} e^{tx} f(x) dx \end{aligned}$$

Proof: Taylor's series (in class).

# Moment Generating Functions

The MGF is used as a convenient way to describe the moments of $X$. To use the MGF in this manner we differentiate $M_X(t)$

**Theorem**

If $X$ has MGF $M_X(t)$ then

$$E(X^n) = M_X^{(n)}(0)$$

Where

$$M_X^{(n)}(0) = \frac{d^n M_X(t)}{dt^n}\Big|_{t=0}$$

Example with Poisson distribution in class.

# Method of Moments

Given a random sample $X_1, X_2, ..., X_n$ with pdf $f(x|\theta_1, ..., \theta_k)$, the MM estimators can be found by equating the first $k$ sample moments $(m_k^{'})$ to the corresponding $k$ population moments $(\mu_k^{'})$

$$m_1^{'} = \frac{1}{n}\sum_{i=1}^{n} X_i^1 \qquad \mu_1^{'} = E(X^1)$$

$$m_2^{'} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \qquad \mu_2^{'} = E(X^2)$$

$$\vdots \qquad\qquad \vdots$$

$$m_k^{'} = \frac{1}{n}\sum_{i=1}^{n} X_i^k \qquad \mu_k^{'} = E(X^k)$$

# Method of Moments

Thus the MM tells us that

1. the kth population moment is a function of the parameter $\theta$, such that $\mu_k^{'} = \mu_k^{'}(\theta)$
2. the kth population moment $\mu_k^{'}$ is estimated by the kth sample moment $m_k^{'}$
$$\hat{\mu}_k^{'} = m_k^{'}$$
3. if there is more than one parameter to find, then the population moment will be a function of the unknown parameters, i.e. $\mu_k^{'} = \mu_k^{'}(\theta_1, \theta_2, ..., \theta_p)$
4. following above, the estimators of $\theta_1, \theta_2, ..., \theta_p$ are found as solutions to the system of $p$ equations
$$\hat{\mu}_k^{'}(\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_p) = m_k^{'}, k = 1, ..., p$$

# Method of Moments

## Example MM estimators for normal distribution

What are the MM estimators for $\mu$ and $\sigma^2$ from a normal distribution?

- The first population moment is $E(X) = \mu$
- The second population moment is $E(X^2) = \sigma^2 + \mu^2$ (from definition of $Var(X) = \sigma^2 = E[(X - \mu)^2]$)
- The first sample moment is $\frac{1}{n}\sum_{i=1}^{n} X_i$
- The second sample moment is $\frac{1}{n}\sum_{i=1}^{n} X_i^2$

Equating the population and sample moments,

1. $E(X) = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i$
2. $E(X^2) = \mu^2 + \sigma^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$

# Method of Moments

## Example MM estimators for normal distribution con't

Equation (1) tells us $\hat{\mu}_{mm} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}$

Rearranging (2) and substituting the above,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

# Method of Moments

## Example MM estimators for Pareto distribution

The Pareto distribution with parameters $\alpha$ and $\beta$ has pdf
$f(x) = \frac{\beta\alpha^\beta}{x^{\beta+1}}, \alpha < x < \infty, \alpha > 0, \beta > 0$
What are the MM estimators for $\alpha$ and $\beta$?
Recall $E(X) = \int_\alpha^\infty xf(x)dx = \int_\alpha^\infty x\beta\alpha^\beta x^{-(\beta+1)}dx = \int_\alpha^\infty \beta\alpha^\beta x^{-\beta}dx =$
$\frac{\beta\alpha^\beta x^{-\beta+1}}{-\beta+1}\Big|_\alpha^\infty = \frac{\beta\alpha}{\beta-1}$, so

$$E(X) = \frac{\beta\alpha}{(\beta-1)}$$

$$E(X^2) = \frac{\beta\alpha^2}{(\beta-2)}$$

# Method of Moments

## Example MM estimators for Pareto distribution con't

Equating population and sample means,

$$\frac{\beta\alpha}{(\beta-1)} = \frac{1}{n}\sum_{i=1}^{n} X_i = m_1$$

$$\frac{\beta\alpha^2}{(\beta-2)} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = m_2$$

Solving for each of the parameters,

$$\hat{\alpha} = 1 + \sqrt{\frac{m_2}{m_2 - m_1^2}}$$

$$\hat{\beta} = \frac{m_2}{m_1}(1 - \sqrt{\frac{m_2 - m_1^2}{m_2}})$$

# Relationship between MLE and MM

Given a random sample $X_1, ..., X_n$ that are iid with pdf $f(x|\theta)$ where $\theta$ is unknown, the purpose of both MLE and MM is to estimate $\theta$. While they take a different approach, there is a connection between the two methods.
Recall the method of moments we equate the population and sample moments:

$$E(X^k) = \mu_k^{'} = m_k^{'} = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

$$\frac{1}{n}\sum_{i=1}^{n} X_i^k - \mu_k^{'} = 0$$

And recall the log likelihood:

$$\ell(\theta|x) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

for which we solve for

$$\sum_{i=1}^{n} \frac{\partial \log f(X_i|\theta)}{\partial \theta} = 0$$

# Relationship between MLE and MM

Generalizing the method of moments, we let $h(x)$ be a real-valued function such that $E[h(X)]$ exists, giving:

$$\mu_h^{'} = E(h(X)) = \int h(x) f(x|\theta) dx$$

for the population moments and

$$m_h^{'} = \sum_{i=1}^{n} h(X_i)$$

for the sample moments. Equating:

$$\mu_h^{'} = m_h^{'}$$

will give an estimate for $\theta$. If $h(x) = x^k - \mu_k^{'}$ we have the original method of moments result. But if we use $h(x) = \frac{\partial \log f(X_i|\theta)}{\partial \theta}$ we obtain the maximum likelihood result.

# Relationship between MLE and MM

When $h(x) = \frac{\partial \log f(x|\theta)}{\partial \theta}$,

$$
\begin{aligned}
\mu_h^{'} &= E(\frac{\partial \log f(x|\theta)}{\partial \theta}) \\
&= \int \frac{\partial \log f(x|\theta)}{\partial \theta} f(x|\theta) dx \\
&= \int \frac{\frac{\partial f(x|\theta)}{\partial \theta}}{f(x|\theta)} f(x|\theta) dx \\
&= \int \frac{\partial f(x|\theta)}{\partial \theta} dx \\
&= \frac{\partial}{\partial \theta} \int f(x|\theta) dx \\
&= 0
\end{aligned}
$$

# Relationship between MLE and MM

So,

$$m_h^{'} = \sum_{i=1}^{n} \frac{\partial \log f(x|\theta)}{\partial \theta}$$

And equating the two,

$$\sum_{i=1}^{n} \frac{\partial \log f(x|\theta)}{\partial \theta} = 0$$

Which is the same as the maximum likelihood equation. Thus, if $h(x, \theta)$ is chosen such that $E[h(X, \theta)] = 0$ then $h(X, \theta)$ is called the score function. Then, $\frac{1}{n} \sum_{i=1}^{n} h(X_i, \theta) = 0$.
For MM the score function is $h(X, \theta) = X^k - \mu_k^{'}(\theta)$ and for MLE the score function is $h(X, \theta) = \ell^{'}(X|\theta)$.

# Bayes Estimators

Recall, in the **frequentist approach** to estimation the parameter $\theta$ is unknown and $X$ is fixed and used to obtain knowledge of $\theta$. We select random samples from the distribution $X = X_1, X_2, ..., X_n$ where each random variable $X_i$ has the given distribution. Different values of $\theta$ lead to different probabilities of the outcome $x = x_1, x_2, ..., x_n$, i.e. $P(X = x|\theta)$ varies with $\theta$. We use likelihood $L(\theta|x)$ to suggest that the real value of $\theta$ is likely to be one with higher probability $P(X = x|\theta)$.

In the Bayesian approach, $\theta$ is described by a distribution (a prior, $\pi(\theta)$), formulated before the data are seen. This is different from the frequentist approach because the concept of likelihood is replaced by a real probability on $\theta$ and treating it as a random variable rather than an unknown constant. A sample of $X$ is taken and then the prior distribution is updated using Bayes rule. This updated prior distribution is the posterior distribution, $\pi(\theta|x)$.

# Bayes Estimators

## Bayes Rule

CB Theorem 1.3.5

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

CB Equation 7.2.7 and 7.2.8 show this in terms of the prior $\pi(\theta)$, the sampling distribution $f(x|\theta)$ and marginal $m(x) = \int f(x|\theta)\pi(\theta)d\theta$,

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m(x)}$$

# Bayes Estimators

## An example applying CB Theorem 1.3.5

Suppose we have two bowls, A and B where A contains 5 red and 10 green balls, B contains 10 red and 5 green balls. We randomly select one of the two bowls and then sample with replacement from the selected bowl to determine whether we chose bowl A or B. Before sampling, we have prior probabilities on choosing bowl A as $P(A) = 0.5$ and $P(B) = 0.5$.
We take a sample of size n $X = X_1, X_2, ..., X_n$ where k of the outcomes are red balls. Using Bayes rule we want to determine the posterior probabilities $P(A|X)$ and $P(B|X)$. The probabilities $P(X|A)$ and $P(X|B)$ are easy to compute:

$$P(X|A) = \left( \begin{array}{c} 15 \\ k \end{array} \right) \frac{5}{15}^{k} \frac{10}{15}^{n-k}$$

$$P(X|B) = \left( \begin{array}{c} 15 \\ k \end{array} \right) \frac{10}{15}^{k} \frac{5}{15}^{n-k}$$

# Bayes Estimators

## An example applying CB Theorem 1.3.5, continued

So, by Bayes formula

$$
\begin{aligned}
P(A|X) &= \frac{P(X|A)P(A)}{P(X|A)P(A) + P(X|B)P(B)} \\
&= \frac{\binom{15}{k} \frac{5}{15}^{k} \frac{10}{15}^{n-k} 0.5}{0.5 \frac{5}{15}^{k} \frac{10}{15}^{n-k} + 0.5 \frac{10}{15}^{k} \frac{5}{15}^{n-k}} \\
&= \frac{2^{n-k}}{2^{n-k} + 2^{k}} \\
P(B|X) &= 1 - P(A|X) \\
&= \frac{2^{k}}{2^{n-k} + 2^{k}}
\end{aligned}
$$

In 10 trials and we get k=4 reds balls, the posterior probabilities are $P(A|X) = \frac{4}{5}$ and $P(B|X) = \frac{1}{5}$.

# Bayes Estimators

In general, we may have an X that is discrete or continuous, but unlike the simple example above, the parameter will continuous. That means we will be working with probability densities rather than probabilities.
In the discrete case: $P(\theta|x) \propto P(x|\theta)P(\theta)$
In the continuous case: $f(\theta|x) \propto f(x|\theta)f(\theta)$, or using the notation in CB text, $\pi(\theta|x) \propto f(x|\theta)\pi(\theta)$

# Bayes Estimators

A key question is: what should the prior be? If we revisit a Bernoulli process, we know:

- $X = 1$ is a success with probability p, $X = 0$ is a failure with probability 1-p
- $X_1, ..., X_n$ represent independent Bernoulli trials with the same unknown parameter p.
- p lies between 0 and 1
- Beta distribution is often used as a prior.

# The EM Algorithm

The Expectation-Maximization Algorithm is another iterative optimization method used to estimate unknown parameters when the liklihood cannot be solved directly. It is designed to find approximation of MLE for unknown parameters when there are some data missing and/or if the underlying disrtribution of the data is more complex (e.g. mixture of two Gaussians).

- The log likelihood is not able to be solved when there is missing information. The basis for ML is that the data are constant but the parameters may vary.
- We need an approximate method in order to approximate the entire joint likelihood function or to approximate the maximum of the likelihood function. The EM algorithm is an iterative method for approximating the maximum of a likelihood function.
- The EM algorithm introduces a hidden or latent variable to make the likelihood "solvable" when there are missing data.
- The models solved by EM therefore have latent variables in addition to unknown parameters and known data (observations).

# The EM Algorithm

Suppose we have a random vector **x** with joint density $f(x|\theta)$ where $\theta \in \Theta \subseteq \mathbb{R}^p$ ($\theta$ has dimension p) If the complete vector **x** is observed then we calculate the maximum likelihood estimate of $\theta$ the usual way, where the log likelihood is:

$$\log f(x|\theta) = \log L(\theta|x) = \ell$$

and is maximized. When there are missing data, only a function or part of the complete data vector **x** is observed. Thus we have $\mathbf{x} = (x_{obs}, x_{miss})$ where $x_{obs}$ represents the observed but incomplete data and $x_{miss}$ represents the latent variables, or unobserved/missing data. It is assumed that the data are missing at random. We have:

$$\begin{aligned} f(x|\theta) &= f(x_{obs}, x_{miss}|\theta) \\ &= f_1(x_{obs}|\theta) \cdot f_2(x_{miss}|x_{obs}; \theta) \end{aligned}$$

where $f_1$ is the joint density of $x_{obs}$ and $f_2$ is the joint density of $x_{miss}$ given the observed data $x_{obs}$. It follows that the log likelihood is:

$$\ell_{obs}(\theta|x_{obs}) = \ell(\theta|x) - \log f_2(x_{miss}|x_{obs}; \theta)$$

where $\ell_{obs}(\theta|x_{obs})$ is the log likelihood for the observed data.

# The EM Algorithm

The EM algorithm is used when maximizing $\ell_{obs}$ is difficult, but maximizing the complete data log likelihood $\ell$ is straightforward. Since **x** is not observed, $\ell$ cannot be evaluated (or maximized). The EM algorithm maximizes $\ell(\theta|x)$ iteratively by replacing its conditional expectation given the observed data $x_{obs}$. The expectation is computed with respect to the distribution of the complete data evaluated at the current estimate of $\theta$.
We start with an initial value $\theta^{(0)}$ for $\theta$. The first iteration computes:

$$Q(\theta|\theta^{(0)}) = E_{\theta^{(0)}}[\ell(\theta|x)|x_{obs}]$$

So $Q(\theta|\theta^{(0)})$ is maximized with respect to $\theta$, and $\theta^{(1)}$ is found such that

$$Q(\theta^{(1)}|\theta^{(0)}) \geq Q(\theta|\theta^{(0)})$$

By Jensen's Inequality

# The EM Algorithm

Thus the EM algorithm has an expectation step and a maximization step.
E-Step:Compute $Q(\theta|\theta^{(t)})$ where

$$Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}}[\ell(\theta|x)|x_{obs}]$$

M-Step: Find $Q(\theta^{(t+1)})$ such that

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)})$$

The expectation and maximization steps are repeated iteratively until the difference between the likelihoods of $\theta^{(t+1)}$ and $\theta^{(t)}$ is smaller than some tolerance level.

# The EM Algorithm

Example: Consider the experiment we discussed in Slides 1 where we had a set of devices that operate and we record when they fail. These devices follow an exponential distribution with mean $\theta$. To estimate $\theta$ we can test $n$ devices until they fail, where the failure times are $x_1, ..., x_n$.
In a separate experiment, $m$ devices were tested but the individual failure times were not recorded. The experimenter only recorded the number of devices, $r$ that failed at time $t$. The missing data are the failure times of all devices $u_1, ..., u_m$.

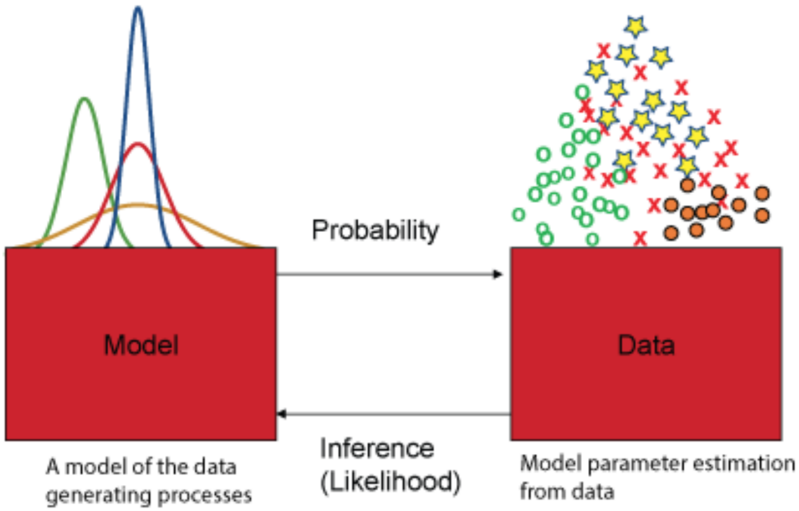
$$Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}}[\ell(\theta|x)|x_{obs}]$$

M-Step: Find $Q(\theta^{(t+1)})$ such that

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta|\theta^{(t)})$$

The expectation and maximization steps are repeated iteratively until the difference between the likelihoods of $\theta^{(t+1)}$ and $\theta^{(t)}$ is smaller than some tolerance level.

# The EM Algorithm in Terms of Latent Variables




This diagram illustrates data that may come from a mixture of $c$ Gaussian distributions with unknown parameters $\boldsymbol{\theta}$. We have $x_1, ..., x_n$ observations and we assume that there are a set of $z_1, ..., z_c$ latent variables (these are now replacing the $x_{miss}$ from above). The latent variables determine the component distribution that the observation comes from.

# The EM Algorithm in Terms of Latent Variables

Given the statistical model that generates a set of observed data $\mathbf{X}$, a set of unobserved latent data or missing values $\mathbf{Z}$, and a vector of unknown parameters $\boldsymbol{\theta}$, along with a likelihood function $L(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ the maximum likelihood estimate (MLE) of the unknown parameters is determined by the marginal likelihood of the observed data:

$$L(\boldsymbol{\theta}|\mathbf{X}) = \int p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z}$$

However, this is often difficult to solve.

# The EM Algorithm in Terms of Latent Variables

Find the likelihood estimators of the previous slide by the EM algorithm:
Expectation Step: Calculate expected value of the log likelihood of **Z** given **X** under a given estimate for $\boldsymbol{\theta}$

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_{Z|X,\theta^(t)}[log(L(\theta|X,Z)]$$

Maximization Step: Find parameters that Maximize

$$\boldsymbol{\theta}^{(t+1)} = \text{argmax}_\theta Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

# The EM Algorithm in Terms of Latent Variables

- The data $\mathbf{X}$ may be discrete or continuous
- The latent variables $\mathbf{Z}$ are discrete, representing classes
- The parameters $\boldsymbol{\theta}$ are continuous and are generally specific to the data from a specific latent variable (also may be associated with all data points)

The R package mclust implements the EM algorithm on Gaussian mixture models, similar to figure above.
The example here `http://rstudio-pubs-static.s3.amazonaws.com/154174_78c021bc71ab42f8add0b2966938a3b8.html` provides code to execute EM "by hand" on coin flipping data.