Maximum Likelihood

See Davison Ch. 4 for background and a more thorough discussion.

Maximum Likelihood

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F_\theta,\ \theta \in \Theta$$

$$L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

$$\ell(\theta) = \ln L(\theta) = \sum_{i=1}^{n} \ln f(x_i; \theta)$$

Close your eyes and differentiate?

Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with parameters $\alpha > 0$ and $\beta > 0$

$$f(x; \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} e^{-x/\beta} x^{\alpha - 1}$$

$$\Theta = \{(\alpha, \beta) : \alpha > 0, \beta > 0\}$$

Simulate Some Data: True α=2, β=3

```
> set.seed(3201); alpha=2; beta=3
> D <- round(rgamma(50,shape=alpha, scale=beta),2); D
 [1] 20.87 13.74  5.13  2.76  4.73  2.66 11.74  0.75 22.07 10.49  7.26  5.82 13.08
[14]  1.79  4.57  1.40  1.13  6.84  3.21  0.38 11.24  1.72  4.69  1.96  7.87  8.49
[27]  5.31  3.40  5.24  1.64  7.17  9.60  6.97 10.87  5.23  5.53 15.80  6.40 11.25
[40]  4.91 12.05  5.44 12.62  1.81  2.70  3.03  4.09 12.29  3.23 10.94
> mean(D); alpha*beta
[1] 6.8782
[1] 6
> var(D); alpha*beta^2
[1] 24.90303
[1] 18
```

Alternatives for getting the data into D might be

D = scan(“Gamma.data”) -- Can put entire URL

D = c(20.87, 13.74, …, 10.94)

Log Likelihood

$$
\begin{aligned}
\ell(\alpha,\beta) &= \ln \prod_{i=1}^{n} \frac{1}{\beta^{\alpha}\Gamma(\alpha)} e^{-x_i/\beta} x_i^{\alpha-1} \\
&= \ln \left[\beta^{-n\alpha}\, \Gamma(\alpha)^{-n} \exp(-\frac{1}{\beta}\sum_{i=1}^{n} x_i) \left(\prod_{i=1}^{n} x_i \right)^{\alpha-1} \right] \\
&= -n\alpha \ln \beta - n \ln \Gamma(\alpha) - \frac{1}{\beta}\sum_{i=1}^{n} x_i + (\alpha-1)\sum_{i=1}^{n} \ln x_i
\end{aligned}
$$

R function for the minus log likelihood

$$\ell(\alpha, \beta) = -n\alpha \ln \beta - n \ln \Gamma(\alpha) - \frac{1}{\beta} \sum_{i=1}^{n} x_i + (\alpha - 1) \sum_{i=1}^{n} \ln x_i$$

```
> # Gamma minus log likelihood: alpha=a, beta=b
> gmll <- function(theta,datta)
+      {
+       a <- theta[1]; b <- theta[2]
+       n <- length(datta); sumd <- sum(datta)
+       sumlogd <- sum(log(datta))
+       gmll <- n*a*log(b) + n*lgamma(a) + sumd/b - (a-1)*sumlogd
+      gmll
+      } # End function gmll
```

Where should the numerical search start?

- How about Method of Moments estimates?
- $E(X) = \alpha\beta$, $Var(X) = \alpha\beta^2$
- Replace population moments by sample moments and put a ~ above the parameters.

$$\tilde{\alpha} = \frac{\overline{X}^2}{S^2} \quad \text{and} \quad \tilde{\beta} = \frac{S^2}{\overline{X}}$$

$$\tilde{\alpha} = \frac{\overline{X}^2}{S^2} \quad \text{and} \quad \tilde{\beta} = \frac{S^2}{\overline{X}}$$

```
> momalpha <- mean(D)^2/var(D); momalpha
[1] 1.899754
> mombeta <- var(D)/mean(D); mombeta
[1] 3.620574
```

```
> gammasearch = nlm(gmll,c(momalpha,mombeta),hessian=T,datta=D); gammasearch
$minimum
[1] 142.0316

$estimate
[1] 1.805930 3.808674

$gradient
[1] 2.847002e-05 9.133932e-06

$hessian
         [,1]      [,2]
[1,] 36.68932 13.127271
[2,] 13.12727  6.222282

$code
[1] 1

$iterations
[1] 6

> eigen(gammasearch$hessian)$values
[1] 41.565137  1.346466
```

$-\ell(\widehat{\alpha}, \widehat{\beta})$

$\widehat{\alpha} = 1.805930 \qquad \widehat{\beta} = 3.808674$

$$\left(-\frac{\partial \ell}{\partial \alpha}, -\frac{\partial \ell}{\partial \beta} \right)^{\top}$$

$$\mathbf{H} = \left[\frac{\partial^2 (-\ell)}{\partial \theta_i \partial \theta_j} \right]$$

$$\mathbf{H} = \left[\frac{\partial^2(-\ell)}{\partial\theta_i \partial\theta_j} \right]$$

If the second derivatives are continuous, **H** is symmetric.

If the gradient is zero at a point and |**H**|≠0,

- If **H** is positive definite, local minimum
- If **H** is negative definite, local maximum
- If **H** has both positive and negative eigenvalues, saddle point

A slicker way to define the minus log likelihood function

```
> gmll2 <- function(theta,datta)
+     { gmll2 <- -sum(dgamma(datta,shape=theta[1],scale=theta[2],log=T))
+       gmll2
+     } # End of gmll2

> nlm(gmll2,c(momalpha,mombeta),datta=D)
$minimum
[1] 142.0316

$estimate
[1] 1.805930 3.808674

$gradient
[1] 2.847002e-05 9.133932e-06

$code
[1] 1

$iterations
[1] 6
```

Likelihood Ratio Tests

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} F_\theta,\ \theta \in \Theta,$$
$$H_0 : \theta \in \Theta_0 \text{ v.s. } H_A : \theta \in \Theta \cap \Theta_0^c,$$

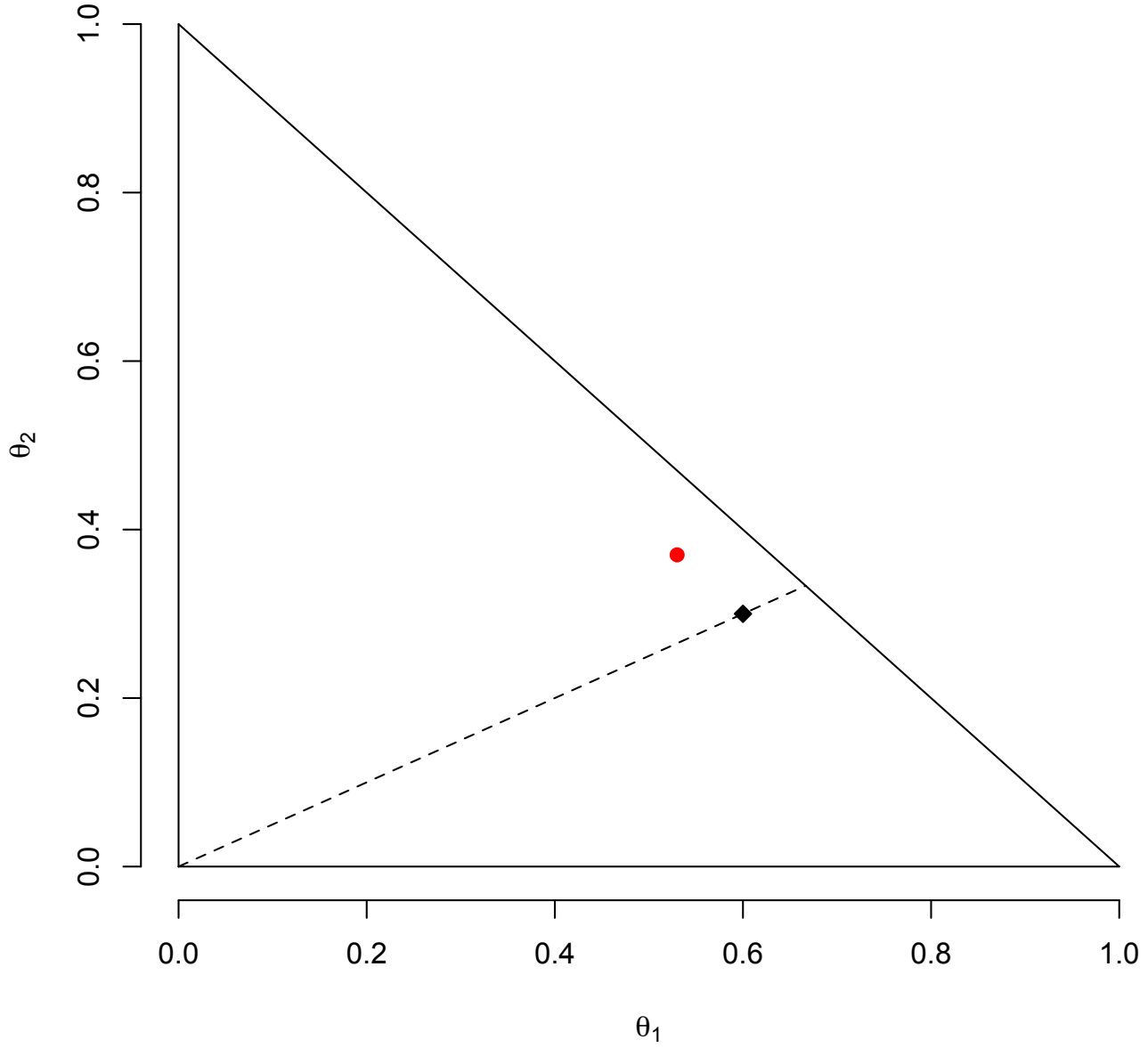
$$G^2 = -2 \log \left(\frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)} \right)$$

Under H_0, G^2 has an approximate chi-square distribution for large N. Degrees of freedom = number of (non-redundant, linear) equalities specified by H_0. Reject when G^2 is large.

Example: Multinomial with 3 categories

- Parameter space is 2-dimensional
- Unrestricted MLE is (P_1, P_2): Sample proportions.
- H_0: $\theta_1 = 2\theta_2$

Parameter space and restricted parameter space



R code for the record

```
# Plotting jobs parameter space with R
# Including MLE and restricted MLE
theta1 = seq(from=0,to=1,by=0.05); theta2=theta1
plot(theta1,theta2,pch=' ', frame.plot=F,
     xlab=expression(theta[1]), ylab=expression(theta[2]))
# Draw boundaries of parameter space
xloc1 = c(0,0); yloc1 = c(0,1); lines(xloc1,yloc1,lty=1)
xloc2 = c(0,1); yloc2 = c(0,0); lines(xloc2,yloc2,lty=1)
xloc3 = c(0,1); yloc3 = c(1,0); lines(xloc3,yloc3,lty=1)
# Restricted parameter space is a line segment
xloc4 = c(0,2/3); yloc4 = c(0,1/3); lines(xloc4,yloc4,lty=2)

points(0.53,0.37, pch=19, col = "red1") # Unrestricted MLE
points(0.60,0.30, pch=23, bg="black") # Restricted MLE
```

Degrees of Freedom

Express H_0 as a set of linear combinations of the parameters, set equal to constants (usually zeros).

Degrees of freedom = number of *non-redundant* linear combinations (meaning linearly independent).

$$\text{Suppose } \boldsymbol{\theta} = (\theta_1, \ldots \theta_7), \text{ with}$$
$$H_0 : \theta_1 = \theta_2,\ \theta_6 = \theta_7,\ \tfrac{1}{3}(\theta_1 + \theta_2 + \theta_3) = \tfrac{1}{3}(\theta_4 + \theta_5 + \theta_6)$$

df=3 (count the = signs)

Can write Null Hypothesis in Matrix Form as H_0: $\boldsymbol{L\theta = h}$

$$H_0 : \theta_1 = \theta_2, \theta_6 = \theta_7, \frac{1}{3}(\theta_1 + \theta_2 + \theta_3) = \frac{1}{3}(\theta_4 + \theta_5 + \theta_6)$$

$$\begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \\ \theta_6 \\ \theta_7 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Rows are linearly independent, so df=number of rows.

Gamma Example: H_0: $\alpha = \beta$

$$
\begin{aligned}
G^2 &= -2 \log \left(\frac{\max_{\theta \in \Theta_0} L(\theta)}{\max_{\theta \in \Theta} L(\theta)} \right) \\
&= 2 \left(-\ell(\widehat{\theta}_0) - [-\ell(\widehat{\theta})] \right)
\end{aligned}
$$

$$
\text{Already have } -\ell(\widehat{\theta}) = 142.0316
$$

Could re-write the function imposing constraints: Not recommended

Make a wrapper function

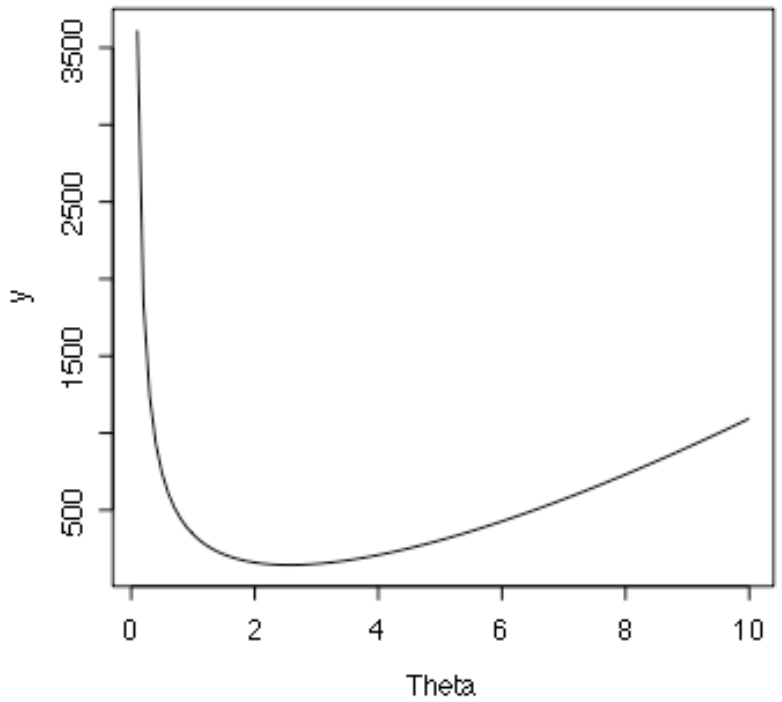
```
> gmll2 <- function(theta,datta)
+     { gmll2 <- -sum(dgamma(datta,shape=theta[1],scale=theta[2],log=T))
+       gmll2
+     } # End of gmll2
```

```
> gmllH0 <- function(theta,datta) # Theta is a scalar this time
+     { gmllH0 <- gmll2(c(theta,theta),datta)
+       gmllH0
+     } # End of gmllH0
> H0start <- sqrt(mean(D)) # Because E(X) = alpha beta
> H0start
[1] 2.622632
```

```
> nlm(gmllH0,H0start,datta=D)
$minimum
[1] 144.1704

$estimate
[1] 2.562369

$gradient
[1] 5.545982e-08

$code
[1] 1

$iterations
[1] 3

Warning messages:
1: NaNs produced in: dgamma(x, shape, scale, log)
2: NA/Inf replaced by maximum positive value
```

It's probably okay, but plot -LL

```
> # It’s probably okay, but plot
> Theta = seq(from=0.1,to=10,by=0.1)
> y = Theta-Theta
> for(i in 1:length(Theta)) y[i] = gmllH0(Theta[i],D)
> plot(Theta,y,type=’l’)
```



Test H_0: $\alpha = \beta$

```
> G2 = 2 * (144.1704-142.0316); G2
[1] 4.2776
> 1-pchisq(G2,df=1)
[1] 0.03861784
```

The actual Theorem (Wilks, 1934)

- There are $r+p$ parameters
- Null hypothesis says that the first r parameters equal specified constants.
- Then under some regularity conditions, G^2 converges in distribution to chi-squared with r df if H_0 is true.
- Can justify tests of *linear* null hypotheses by a re-parameterization using the invariance principle.

How it works

- The invariance principle of maximum likelihood estimation says that the MLE of a function is that function of the MLE. Like

$$\widehat{\sigma} = \sqrt{\widehat{\sigma^2}}$$

- Meaning is particularly clear when the function is one-to-one.
- Write H_0: $\mathbf{L\theta} = \mathbf{h}$, where **L** is r x (r+p) and rows of **L** are linearly independent.
- Can always find an additional p vectors that, together with the rows of **L**, span R^{r+p}
- This defines a (linear) 1-to-1 re-parameterization, and Wilks' theorem applies directly.

Gamma Example H_0: $\alpha = \beta$

Re-parameterize: Let

$$\theta_1 = \alpha - \beta$$

$$\theta_2 = \beta$$

The re-parameterization is one-to-one because

$$\alpha = \theta_1 + \theta_2$$

$$\beta = \theta_2$$

Invariance says $\widehat{\theta}_1 = \widehat{\alpha} - \widehat{\beta}$ and $\widehat{\theta}_2 = \widehat{\beta}$

$$H_0 : \theta_1 = 0$$

Can Work for Non-linear Null Hypotheses Too

$$X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$$

$$H_0 : \mu^2 = \sigma^2$$

Re-parameterize: Let

$$\theta_1 = \mu^2 - \sigma^2$$

$$\theta_2 = \mu$$

The re-parameterization is one-to-one because

$$\mu = \theta_2$$

$$\sigma^2 = \theta_2^2 - \theta_1$$

Invariance says $\widehat{\theta}_1 = \widehat{\mu}^2 - \widehat{\sigma}^2$ and $\widehat{\theta}_2 = \widehat{\mu}$

$$H_0 : \theta_1 = 0$$

Copyright Information