# PLSC 504 – Fall 2022
# Review: Likelihoods, Optimization, etc.

August 22, 2022

# Welcome!

- "Proseminar" in methods
- We'll be meeting Monday afternoons from 13:00-16:00 ET in the Boucke Building, Room 302
- Texts: Various (mostly articles; see the syllabus)
- All course materials: `https://github.com/PrisonRodeo/PLSC504-2022-git`
- Teaching Assistant (*not* "the preceptor"): Michael Burnham
- Software: R...
- Grading: Ten homework assignments (@ 50 points), plus a final project (500 points)
- Contact me: `zorn@psu.edu`, or `@prisonrodeo`, or text (803) 553-4077

# A Very Simple Model

$$Y \sim N(\mu, \sigma^2)$$

$$\begin{aligned} \mathrm{E}(Y) &= \mu \\ \mathrm{Var}(Y) &= \sigma^2 \end{aligned}$$

# Some Data

$$
\begin{aligned}
Y \quad = \quad & 64 \\
& 63 \\
& 59 \\
& 71 \\
& 68
\end{aligned}
$$

$$\Pr(Y_i = y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(Y_i - \mu)^2}{2\sigma^2}\right]$$

So

$$\begin{aligned}
\Pr(Y_1 = 64) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(64 - \mu)^2}{2\sigma^2}\right] \\
\Pr(Y_2 = 63) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(63 - \mu)^2}{2\sigma^2}\right]
\end{aligned}$$

$$\ldots$$

$$\Pr(A, B | \Pr(A) \perp \Pr(B)) = \Pr(A) \times \Pr(B)$$

So:

$$\begin{aligned}\Pr(Y_1 = 64, Y_2 = 63) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(64-\mu)^2}{2\sigma^2}\right] \times \\ &\quad \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(63-\mu)^2}{2\sigma^2}\right]\end{aligned}$$

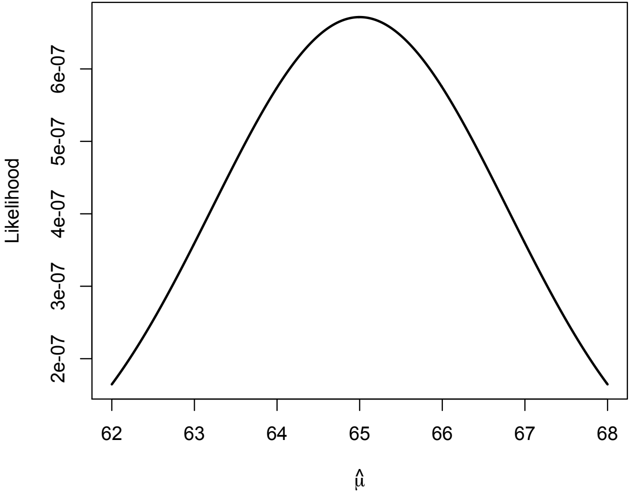
$$\begin{aligned}
\Pr(Y_i = y_i \; \forall \; i) &\equiv L(Y|\mu, \sigma^2) \\
&= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i - \mu)^2}{2\sigma^2}\right]
\end{aligned}$$

$$L(\hat{\mu}, \hat{\sigma}^2 | Y) \propto \Pr(Y | \hat{\mu}, \hat{\sigma}^2)$$

For $\hat{\mu} = 68$, $\hat{\sigma} = 4$:

$$\begin{aligned} L &= \frac{1}{\sqrt{2\pi 16}} \exp\left[-\frac{(64-68)^2}{32}\right] \times \\ & \quad \frac{1}{\sqrt{2\pi 16}} \exp\left[-\frac{(63-68)^2}{32}\right] \times \\ & \quad \frac{1}{\sqrt{2\pi 16}} \exp\left[-\frac{(59-68)^2}{32}\right] \times \ldots \\ &= \text{some reeeeeally small number...} \end{aligned}$$

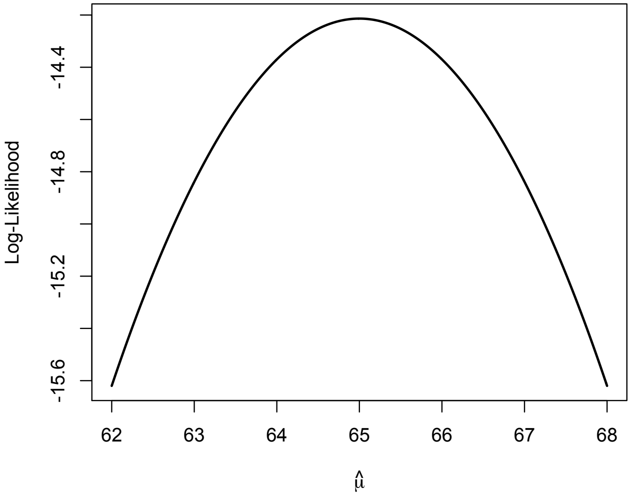
# What a Likelihood Looks Like



# Log-Likelihood

$$
\begin{aligned}
\ln L(\hat{\mu}, \hat{\sigma}^2 | Y) &= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(Y_i-\mu)^2}{2\sigma^2}\right] \\
&= \sum_{i=1}^{N} \ln\left\{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(Y_i-\mu)^2}{2\sigma^2}\right]\right\} \\
&= -\frac{N}{2}\ln(2\pi) - \left[\sum_{i=1}^{N} \frac{1}{2}\ln\,\sigma^2 + \frac{1}{2\sigma^2}(Y_i-\mu)^2\right]
\end{aligned}
$$

# What a Log-Likelihood Looks Like



# The "Maximum" Part

For $L = f(Y, \theta)$,

- Calculate $\frac{\partial \ln L}{\partial \theta}$,
- Set $\frac{\partial \ln L}{\partial \theta} = 0$, solve for $\hat{\theta}$,
- Calculate $\frac{\partial^2 \ln L}{\partial \theta^2}$,
- Verify $\frac{\partial^2 \ln L}{\partial \theta^2} < 0$.

# Example: Normal $Y$

$$\ln L(\hat{\mu}, \hat{\sigma}^2 | Y) = -\frac{N}{2} \ln(2\pi) - \left[ \sum_{i=1}^{N} \frac{1}{2} \ln\ \sigma^2 + \frac{1}{2\sigma^2} (Y_i - \mu)^2 \right]$$

$$\begin{aligned} \frac{\partial \ln L}{\partial \mu} &= \frac{1}{\sigma^2} \sum_{i=1}^{N} (Y_i - \mu) \\ \frac{\partial \ln L}{\partial \sigma^2} &= \frac{-N}{2\sigma^2} + \frac{1}{2} \sigma^4 \sum_{i=1}^{N} (Y_i - \mu)^2 \end{aligned}$$

# Example: Normal $Y$ (continued)

$\longrightarrow$

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

# Example: Linear Regression

$$
\begin{aligned}
\mathrm{E}(Y) \equiv \mu &= \beta_0 + \beta_1 X_i \\
\mathrm{Var}(Y) &= \sigma^2
\end{aligned}
$$

$$
L(\beta_0, \beta_1, \sigma^2 | Y) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right]
$$

$$\begin{aligned}\ln L(\beta_0, \beta_1, \sigma^2 | Y) &= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right] \\ &= -\frac{N}{2}\ln(2\pi) - \sum_{i=1}^{N}\left[\frac{1}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right]\end{aligned}$$

*Kernel*:

$$-\sum_{i=1}^{N}\left[\frac{1}{2}\ln\sigma^2 + \frac{1}{2\sigma^2}(\underbrace{Y_i - \beta_0 - \beta_1 X_i}_{\hat{u}_i})^2\right]$$

# MLE in General

$$\Pr(Y) = f(\mathbf{X}, \theta)$$

$$L = \prod_{i=1}^{N} f(Y_i | \mathbf{X}_i, \theta)$$

$$\ln L = \sum_{i=1}^{N} \ln f(Y_i | \mathbf{X}_i, \theta)$$

$$\ln L(\hat{\theta} | Y, \mathbf{X}) = \max_{\theta} \{\ln L(\theta | Y, \mathbf{X})\}$$

$$\mathbf{g}(\hat{\theta}) = \frac{\partial \ln L(\hat{\theta})}{\partial \hat{\theta}}$$

Taylor series:

$$\frac{\partial \ln L}{\partial \hat{\theta}} \approx \frac{\partial \ln L}{\partial \theta} + \frac{\partial^2 \ln L}{\partial \theta^2}(\hat{\theta} - \theta)$$

$$\begin{aligned} \hat{\theta} - \theta &= \left(-\frac{\partial^2 \ln L}{\partial \theta^2}\right)^{-1} \frac{\partial \ln L}{\partial \theta} \\ &= -\mathbf{H}(\theta)^{-1}\mathbf{g}(\theta) \end{aligned}$$

Need

$$\text{plim}(\hat{\theta} - \theta) = 0$$

So:

- *Assume* $\mathbf{H}(\theta) \xrightarrow{a} \mathbf{A} < \infty$
- *Show* $\mathrm{E}[\mathbf{g}(\theta)] \rightarrow \mathbf{0}$ as $N \rightarrow \infty$

# Consistency (continued)

$$\begin{aligned}
\mathsf{E}[\mathbf{g}(\theta)] &= \frac{1}{N}\,\mathsf{E}\left(\frac{\partial \ln L_1}{\partial \theta} + \frac{\partial \ln L_2}{\partial \theta} + \ldots + \frac{\partial \ln L_N}{\partial \theta}\right) \\
&= \frac{1}{N}\left[\mathsf{E}\left(\frac{\partial \ln L_1}{\partial \theta}\right) + \mathsf{E}\left(\frac{\partial \ln L_2}{\partial \theta}\right) + \ldots\right] \\
&\overset{a}{=} \mathbf{0}
\end{aligned}$$

Cramer-Rao say:

$$\mathsf{Var}(\hat{\theta}) \geq \left[-\mathsf{E}\left(\frac{\partial^2 \ln L(\theta)}{\partial \theta^2}\right)\right]^{-1}$$

$$\begin{aligned}\mathrm{Var}(\hat\theta) &= \mathrm{E}[(\hat\theta-\theta)(\hat\theta-\theta)'] \\ &= \mathrm{E}\left[\left(-\frac{\partial^2 \ln L}{\partial\theta^2}\right)^{-1}\frac{\partial \ln L}{\partial\theta}\frac{\partial \ln L'}{\partial\theta}\left(-\frac{\partial^2 \ln L}{\partial\theta^2}\right)^{-1}\right]\end{aligned}$$

For MLE:

$$\mathrm{E}\left[\frac{\partial \ln L}{\partial\theta}\frac{\partial \ln L'}{\partial\theta}\right] = \mathrm{E}\left[\frac{\partial^2 \ln L}{\partial\theta^2}\right]$$

So,

$$\begin{aligned}\mathrm{Var}(\hat\theta) &= \left[-\mathrm{E}\left(\frac{\partial^2 \ln L}{\partial\theta^2}\right)\right]^{-1} \\ &= [\mathbf{I}(\theta)]^{-1}\end{aligned}$$

By LLN:

$$\frac{\hat{\theta} - \theta}{\sqrt{\mathbf{I}(\theta)^{-1}}} \sim N(\mathbf{0}, \mathbf{1})$$

Or:

$$\hat{\theta} \sim N(\theta, \mathbf{I}(\theta)^{-1})$$

# Invariance: Parameters

For

$$\gamma = h(\theta)$$

$$\hat{\gamma}_{ML} = h(\hat{\theta}_{ML})$$

Suppose

$$\phi^2 = 1/\sigma^2$$

so that

$$Y \sim N(\mu, \phi^2).$$

Then:

$$\ln L(\hat{\mu}, \hat{\phi}^2) = -\left[\sum_{i=1}^{N} \frac{1}{2} \ln\ \phi^2 - \frac{1}{2\phi^2}(Y_i - \mu)^2\right]$$

and:

$$\frac{\partial \ln L}{\partial \phi^2} = \frac{-N}{2\phi^2} + \frac{1}{2}\phi^4 \sum_{i=1}^{N}(Y_i - \mu)^2$$

and:

$$\begin{aligned} \hat{\phi}^2 &= \frac{N}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2} \\ &= \frac{1}{\hat{\sigma}^2} \end{aligned}$$

MLEs:

- Maximize $L(\theta|Y, \mathbf{X})$
- Are consistent in $N$
- Are asymptotically efficient
- Are asymptotically Normal
- Are invariant to (injective) transformations and varying sampling methods

# Optimization

The basic problem: find

$$\max_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^k} \ln L(\hat{\boldsymbol{\beta}} | Y, \mathbf{X})$$

The intuition:

- Start with $\hat{\boldsymbol{\beta}}_{\mathbf{0}}$
- Adjust:

$$\hat{\boldsymbol{\beta}}_{\mathbf{1}} = \hat{\boldsymbol{\beta}}_{\mathbf{0}} + \mathbf{A}_{\mathbf{0}}$$

- Repeat.

*Iterative* process:

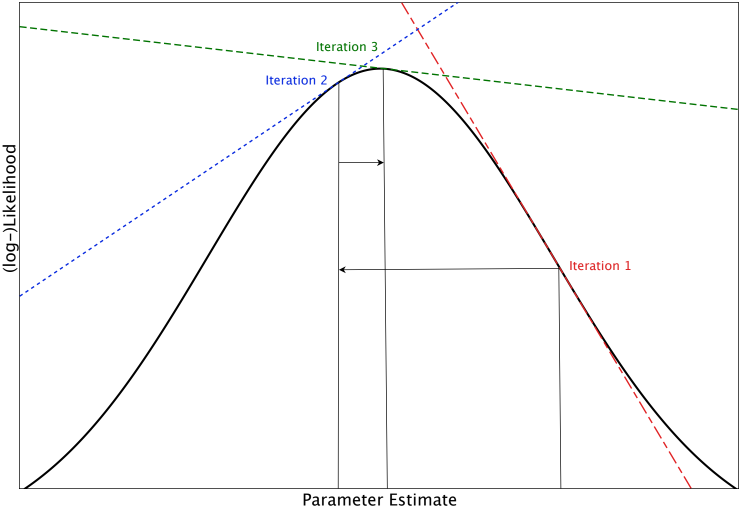
$$\hat{\boldsymbol{\beta}}_\ell = \hat{\boldsymbol{\beta}}_{\ell-\mathbf{1}} + \mathbf{A}_{\ell-\mathbf{1}}$$

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_\ell \ni \hat{\boldsymbol{\beta}}_\ell - \hat{\boldsymbol{\beta}}_{\ell-1}(\equiv \mathbf{A}_\ell) < \tau$$

# What's $\mathbf{A}$?

$$\mathbf{A} = f[\mathbf{g}(\hat{\boldsymbol{\beta}})]$$

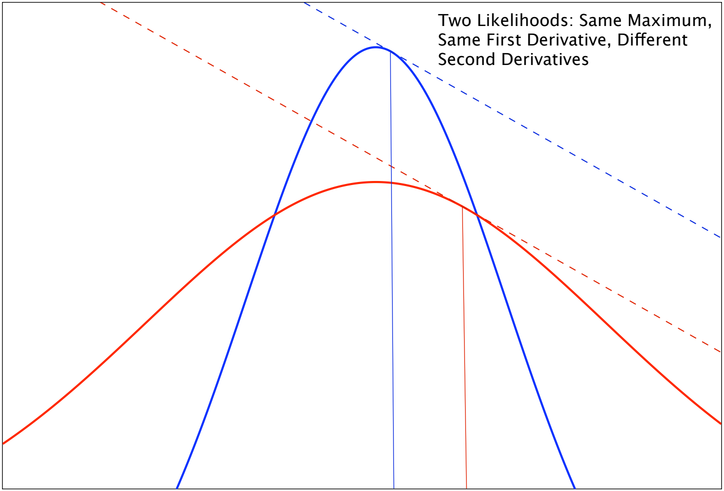
- $\mathbf{g}(\hat{\boldsymbol{\beta}}) =$ "directionality" of change
    - $\mathbf{g}(\hat{\beta}_k) < 0 \rightarrow A_k < 0$
    - $\mathbf{g}(\hat{\beta}_k) > 0 \rightarrow A_k > 0$


Iteration 3
Iteration 2
Iteration 1
(log−)Likelihood
Parameter Estimate

# "Steepest Ascent"

$$\mathbf{A}_\ell = \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_\ell}$$

$$\hat{\boldsymbol{\beta}}_\ell = \hat{\boldsymbol{\beta}}_{\ell-1} + \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}}$$


Two Likelihoods: Same Maximum,
Same First Derivative, Different
Second Derivatives

"Step size":

$$\hat{\boldsymbol{\beta}}_{\ell} = \hat{\boldsymbol{\beta}}_{\ell-\mathbf{1}} + \lambda_{\ell-\mathbf{1}} \boldsymbol{\Delta}_{\ell-\mathbf{1}}$$

- $\boldsymbol{\Delta} \rightarrow$ *direction*
- $\lambda \rightarrow$ *amount* ("step size")

Key: Hessian

$$\mathbf{H}(\hat{\beta}) = \frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}^2}$$

How?

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_\ell &= \hat{\boldsymbol{\beta}}_{\ell-1} - \left(\frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}^2}\right)^{-1} \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}} \\
&= \hat{\boldsymbol{\beta}}_{\ell-1} - \mathbf{H}(\hat{\boldsymbol{\beta}}_{\ell-1})^{-1}\mathbf{g}(\hat{\boldsymbol{\beta}}_{\ell-1})
\end{aligned} \tag{1}
$$

# Sidebar: Newton-Raphson, re-revealed

Taylor series, anyone?

$$f(X) \approx f(a) + f'(a)(x - a)$$

Here,

$$\frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_\ell} \approx \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}} + \frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}^2_{\ell-1}}(\hat{\boldsymbol{\beta}}_\ell - \hat{\boldsymbol{\beta}}_{\ell-1})$$

# What we really want...

$$\frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_\ell} = \mathbf{0}$$

So:

$$\mathbf{0} \approx \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}} + \frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}^2_{\ell-1}}(\hat{\boldsymbol{\beta}}_\ell - \hat{\boldsymbol{\beta}}_{\ell-1})$$

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_\ell &\approx \hat{\boldsymbol{\beta}}_{\ell-1} - \left(\frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}^2_{\ell-1}}\right)^{-1} \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}} \\
&\approx \hat{\boldsymbol{\beta}}_{\ell-1} - \mathbf{H}(\hat{\boldsymbol{\beta}}_{\ell-1})^{-1}\mathbf{g}(\hat{\boldsymbol{\beta}}_{\ell-1})
\end{aligned}$$

Newton-Raphson requires $\mathbf{H}(\hat{\beta})^{-1} \rightarrow$ *calculates* $\mathbf{H}(\hat{\beta})^{-1}$ at every iteration. This can make it somewhat slow / computationally demanding.

Modified Marquardt:

- Used when $\mathbf{H}(\hat{\beta})$ isn't invertable
- Adds a constant $\mathbf{C}$ to diag$[\mathbf{H}(\hat{\beta})]$
- Variants: Add $\mathbf{C}(h_k)$

"Method of Scoring" (due to Fisher) uses:

$$
\begin{aligned}
\hat{\boldsymbol{\beta}}_\ell &= \hat{\boldsymbol{\beta}}_{\ell-1} - \left[ \mathrm{E}\left( \frac{\partial^2 \ln L}{\partial \hat{\boldsymbol{\beta}}^2_{\ell-1}} \right)^{-1} \right] \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}} \\
&= \hat{\boldsymbol{\beta}}_{\ell-1} - \{\mathrm{E}[\mathbf{H}(\hat{\boldsymbol{\beta}}_{\ell-1})]\}^{-1} \mathbf{g}(\hat{\boldsymbol{\beta}}_{\ell-1})
\end{aligned}
$$

# Berndt, Hall$^2$, and Hausman ("BHHH")

Uses:

$$\hat{\boldsymbol{\beta}}_\ell = \hat{\boldsymbol{\beta}}_{\ell-1} - \left( \sum_{i=1}^{N} \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}} \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}}' \right)^{-1} \frac{\partial \ln L}{\partial \hat{\boldsymbol{\beta}}_{\ell-1}}$$

Advantages:

- (Relatively) very easy to compute
- Reasonably accurate...

# Other "Newton Jr.s"

- Davidson-Fletcher-Powell ("DFP")
- Broyden et al. ("BFGS")
- They are:
  - Faster / more efficient
  - Comparatively bad at getting $-\left(\mathbf{H}(\hat{\beta})\right)^{-1}$

# Summary: Optimization & Inference

| Method | "Step size" ($\partial^2$) matrix | Variance-Covariance Estimate |
|---|---|---|
| Newton | Inverse of the observed second derivative (Hessian) | Inverse of the negative Hessian |
| Scoring | Inverse of the expected value of the Hessian (information matrix) | Inverse of the negative information matrix |
| BHHH | Outer product approximation of the information matrix | Inverse of the outer product approximation |

Lots of optimizers:

- `maxLik` package: options for Newton-Raphson, BHHH, BFGS, others
- `optim` (in `stats`) – quasi-Newton, plus others
- `nlm` (in `stats`) – nonlinear minimization "using a Newton-type algorithm"
- `newton` (in `Bhat`) – Newton-Raphson solver
- `solveLP` (in `linprog`) – linear programming optimizer

- *Must* provide log-likelihood function
- Can provide $\mathbf{H}(\hat{\boldsymbol{\beta}})$, $\mathbf{g}(\hat{\boldsymbol{\beta}})$, both, or neither
- Choose optimizer (Newton, BHHH, BFGS, etc.)
- Returns an object of class `maxLik`

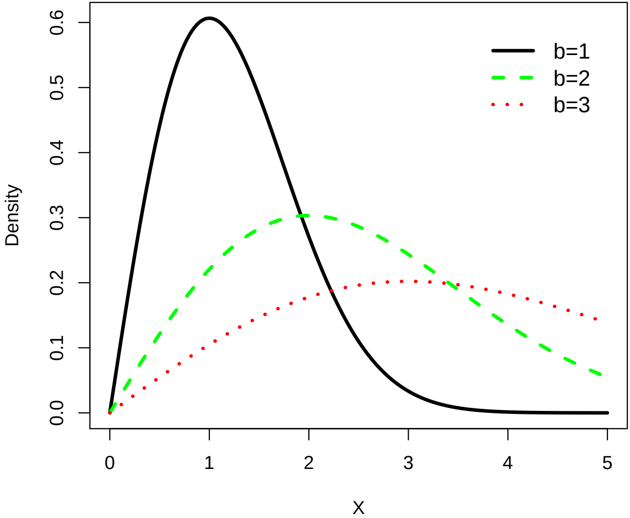
Rayleigh distribution density:

$$\Pr(X = x) = \frac{x}{b^2} \exp\left[\frac{-x^2}{2b^2}\right],\ b > 0$$

Other traits:

- Support $\in [0, \infty)$
- $E(X) = b\sqrt{\frac{\pi}{2}}$
- Mode $= b$
- $Var(X) = \frac{4-\pi}{2} b^2$

# Rayleigh Densities



We can generate a Rayleigh-distributed random variable $X$ with parameter $b$ via inverse transform sampling, as:

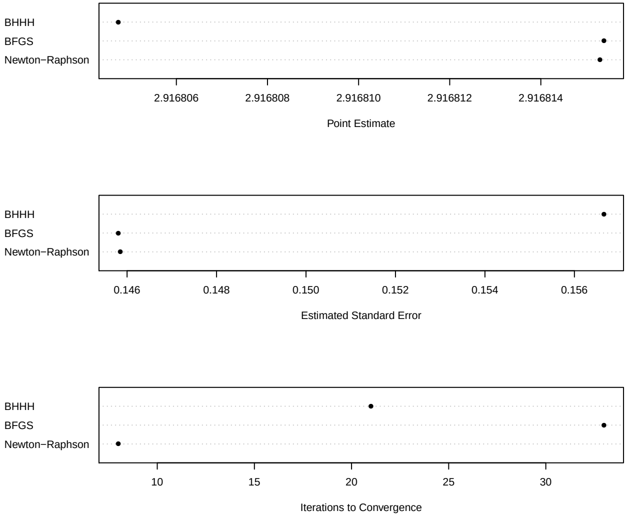
$$X = b\sqrt{-2\ln(1-U)}$$

where $U \in \text{Uniform}[0,1]$. So, for (e.g.) $b = 3$:

```
> library(maxLik,distr)
> set.seed(7222009)
> U<-runif(100)
> rayleigh<-3*sqrt(-2*log(1-U))  # b = 3
> loglike <- function(param) {
+    b <- param[1]
+    ll <- (log(x)-log(b^2)) + ((-x^2)/(2*b^2))
+    ll
+  }
```

```
> x<-rayleigh
> hats <- maxLik(loglike, start=c(1))
> summary(hats)
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 8 iterations
Return code 2: successive function values within tolerance limit
Log-Likelihood: -195.7921
1  free parameters
Estimates:
     Estimate Std. error t value Pr(> t)
[1,]   2.9168     0.1459      20  <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
--------------------------------------------
```

# Comparing Optimizers



```
> Y<-c(0,0,0,0,0,1,1,1,1,1)
> X<-c(0,1,0,1,0,1,1,1,1,1)
> logL <- function(param) {
+   b0<-param[1]
+   b1<-param[2]
+   ll<-Y*log(exp(b0+b1*X)/(1+exp(b0+b1*X))) +
+       (1-Y)*log(1-(exp(b0+b1*X)/(1+exp(b0+b1*X))))
+   ll
+ }
```

# R : What We *Don't* Like To See

```
> Bhat<-maxLik(logL,start=c(0,0))
> summary.maxLik(Bhat)
--------------------------------------------
Maximum Likelihood estimation
Newton-Raphson maximisation, 9 iterations
Return code 1: gradient close to zero
Log-Likelihood: -4.187887
2  free parameters
Estimates:
     Estimate Std. error t value Pr(> t)
[1,]   -104.3        Inf       0       1
[2,]    105.2        Inf       0       1
--------------------------------------------
```

- Potential Problems
- Likely Causes
- Tips

Enemy # 1: Noninvertable $\mathbf{H}(\hat{\boldsymbol{\beta}})$

- "Non-concavity," "non-invertability," etc.
- (Some part of) the likelihood is "flat"
- Why? (Bob Dole...)

# Problems

Identification

- Possible due to functional form alone...
- "Fragile"
- Manifestation: parameter instability

Poor Conditioning

- Numerical issues
- Potentially:
  - Collinearity
  - Other weirdnesses (nonlinearities)

## Potential Causes

- Misspecification.
- Missing data
- Variable scaling
- Typical $\Pr(Y)$

# Hints

- T-h-i-n-k!
- Know thy data
- Keep an eye on your iteration logs...
- Don't overreach