3. Maximum Likelihood Estimation

Maximum Likelihood (ML) Estimation

Maximum likelihood identifies the population parameter values that best fit the data

Similar in logic to least squares, the “best” parameters minimize the distance to the data

The normal distribution function captures the discrepancy between the data and parameters

A Simple Analysis Model

Empty regression model with two parameters, the mean and variance

$$Y_i = \mu + \varepsilon_i = E(Y) + \varepsilon_i$$
$$Y_i \sim N\left(E(Y), \sigma^2\right)$$

Y is normally distributed around a predicted value, $E(Y)$, which in this case is the mean

Some Recurring Notation

Predicted value

$$Y_i = \mu + \varepsilon_i = E(Y) + \varepsilon_i$$
$$Y_i \sim N\left(E(Y), \sigma^2\right)$$

Univariate normal curve

Spread

Center

Math Achievement Data

Math achievement data for 250 students

The data set includes pre-test and post-test math achievement scores and academic-related variables such as math self-efficacy, math anxiety, standardized reading scores, socio-demographic variables

math.dat

Variable	Name	Missing %	Scaling
Identifier variable	ID	0	Integer index
Gender	MALE	0	0 = female, 1 = male
Free or reduced lunch	LUNCHASST	4.3	0 = none, 1 = assistance
Achievement group	ACHIEVEGRP	2.0	1 = typically achieving, 2 = low achieving, 3 = learning disability
Standardized reading	STANREAD	10.0	Continuous
Math self-efficacy	EFFICACY	9.7	6-point ordinal scale
Math anxiety	ANXIETY	9.3	Continuous
Pre-test math achievement	MATHPRE	0	Continuous
Post-test math achievement	MATHPOST	18.0	Continuous

Substantive Example

Use maximum likelihood to estimate the mean and variance of math post-test scores

$$MATHPOST_i = \mu + \varepsilon_i$$

The analysis assumes that math scores are normally distributed around the mean

Normal Distribution Function

Likelihood of a score, **given assumed parameter values**

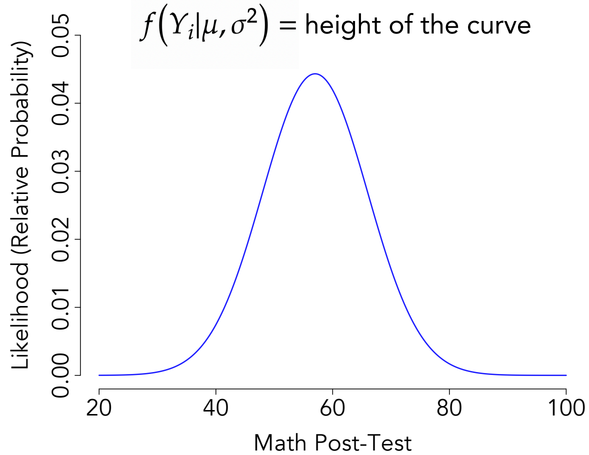
Squared z-score ("fit")

$$f\left(Y_i|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\frac{(Y_i-\mu)^2}{\sigma^2}\right\}$$

Scaling factor that makes area under the curve equal one

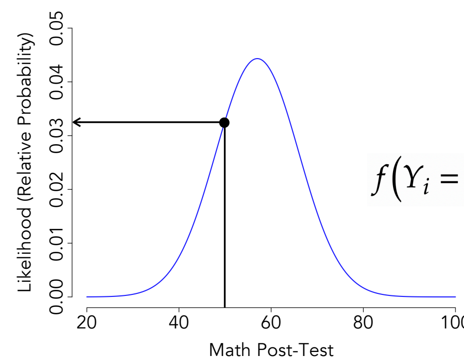
Graphic

$f(Y_i|\mu, \sigma^2)$ = height of the curve



Example

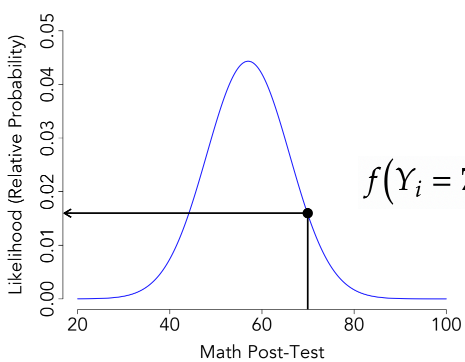
Let's assume for now that μ = 57 and σ = 9



$f(Y_i = 50|\mu, \sigma^2) = .033$

Example

Let's assume for now that μ = 57 and σ = 9



$f(Y_i = 70|\mu, \sigma^2) = .016$

Key Points

A likelihood essentially measures individual fit

The likelihood gets higher (fit improves) when a score is close to the center of the distribution (i.e., is a good match to the parameters)

The likelihood gets lower as scores move away from the center of the distribution

A Sample Of Likelihood Values

Still assume that the population parameters are $\mu = 57$ and $\sigma = 9$

Scores close to the mean have high likelihoods, and vice versa

Y	L
63	0.0354
53	0.0401
71	0.0132
53	0.0401
57	0.0443
55	0.0432
59	0.0432
74	0.0074
44	0.0156
37	0.0037
45	0.0182
63	0.0354

Joint Probability

The joint probability for a set of events is the product of individual probabilities

e.g., The probability of concurrently observing two heads is (.50)(.50) = .25

Strictly speaking, a likelihood is not a probability, but the same rules apply

Sample Likelihood

The sample likelihood is the product of the individual likelihoods

$$f\left(Y|\mu,\sigma^2\right) = \prod_N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(Y_i-\mu)^2}{\sigma^2}\right\}$$

$\prod_N$ is the multiplication operator over all cases

Sample Likelihood

The sample likelihood is the product of 250 individual likelihood contributions

$f\left(Y|\mu,\sigma^2\right) = (.0355)(.0402)\ldots(.0048) =$ really close to zero!

The sample likelihood is too small to work with

Logarithms

Likelihoods are computationally difficult and introduce precision and rounding problems

The product rule for logarithms says log[(a)(b)] = log(a) + log(b)

Using logarithms converts a multiplication problem to a simpler addition problem

Log Likelihood

The log likelihood the natural logarithm of a likelihood

$$\log(L_i) = \log\left(f\left(Y_i|\mu,\sigma^2\right)\right) = \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\frac{(Y_i-\mu)^2}{\sigma^2}\right\}\right)$$

Log likelihood values also quantify relative probability, but they do so on a negative metric

A Sample Of Log Likelihood Values

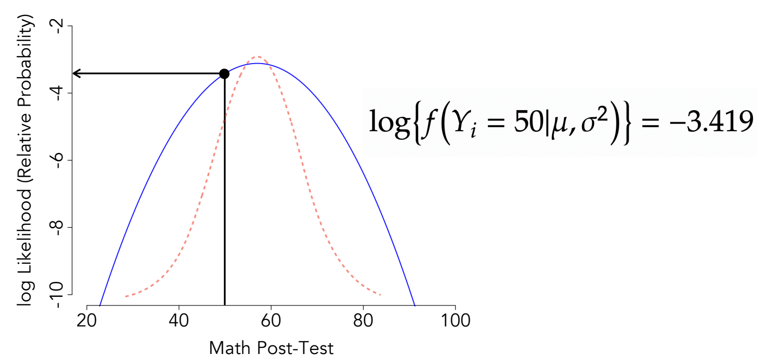
Still assume that the population parameters are $\mu = 57$ and $\sigma = 9$

Scores close to the mean have high likelihoods, and vice versa

Y	L	$\log(L)$
63	0.0354	-1.4510
53	0.0401	-1.3969
71	0.0132	-1.8794
53	0.0401	-1.3969
57	0.0443	-1.3536
55	0.0432	-1.3645
59	0.0432	-1.3645
74	0.0074	-2.1308
44	0.0156	-1.8069
37	0.0037	-2.4318
45	0.0182	-1.7399
63	0.0354	-1.4510

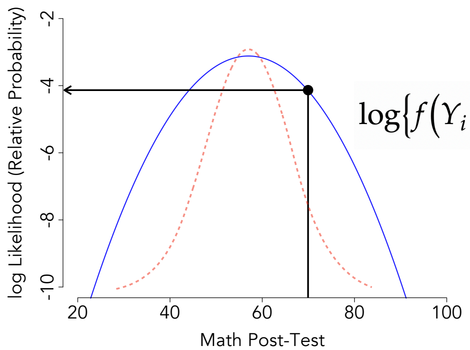
Example

Let's assume for now that $\mu = 57$ and $\sigma = 9$



Example

Let's assume for now that $\mu = 57$ and $\sigma = 9$



$$\log\{f(Y_i = 70|\mu, \sigma^2)\} = -4.159$$

21

Sample Log Likelihood

The sample likelihood is the sum of the individual log likelihood contributions

$$\log L = \sum_N \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2}\frac{(Y_i-\mu)^2}{\sigma^2}\right\}\right)$$

The log likelihood represents the joint probability of the sample data, given the parameters

22

Sample Log Likelihood

The sample log likelihood is the sum of 250 individual log likelihood contributions

$$\log L = (-3.3384) + (-3.2149) + \cdots + (-5.3446) = -914.4729$$

The sample likelihood quantifies the relative probability of obtaining these 250 scores from a normal population with **this particular mean and standard deviation**

23

Interpreting The Log Likelihood

The log likelihood quantifies the fit between the sample data and the population parameters

The log likelihood depends on the sample size, number of variables, number of parameters in the model, missing data, etc.

No absolute criterion for a good or a bad value

24

Estimation Strategy

Thus far we've treated the parameters as known

The sample log likelihood provides a mechanism for estimating unknown parameters

Compute the log likelihood for different parameter values and find the value that produces the highest log likelihood (best fit)

Individual Log Likelihoods At μ = 54

Student	Y	$f(Y)$	log(L)
1	63	0.0269	-3.6162
2	53	0.0441	-3.1223
3	71	0.0074	-4.9001
4	53	0.0441	-3.1223
5	57	0.0419	-3.1717
…	…	…	…
246	49	0.0380	-3.2705
247	54	0.0443	-3.1162
248	61	0.0328	-3.4186
249	51	0.0419	-3.1717
250	38	0.0091	-4.6964
		Sum =	-926.436

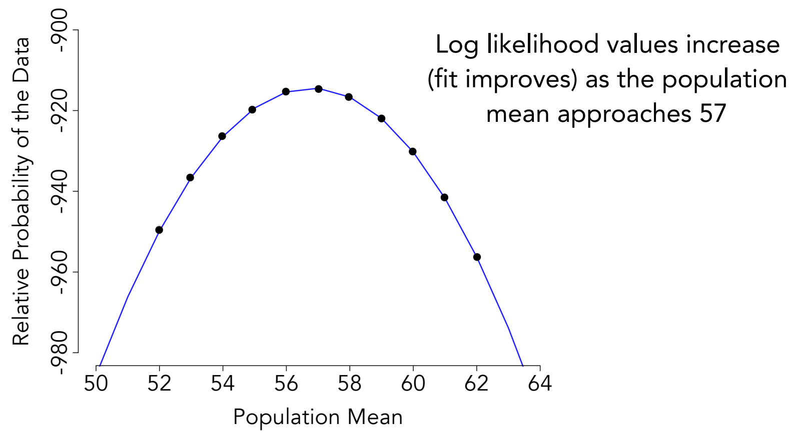
Individual Log Likelihoods At μ = 55

Student	Y	$f(Y)$	log(L)
1	63	0.0299	-3.5112
2	53	0.0432	-3.1409
3	71	0.0091	-4.6964
4	53	0.0432	-3.1409
5	57	0.0432	-3.1409
…	…	…	…
246	49	0.0355	-3.3384
247	54	0.0441	-3.1223
248	61	0.0355	-3.3384
249	51	0.0402	-3.2149
250	38	0.0074	-4.9001
		Sum =	-919.362

Individual Log Likelihoods At μ = 56

Student	Y	$f(Y)$	log(L)
1	63	0.0328	-3.4186
2	53	0.0419	-3.1717
3	71	0.0111	-4.5051
4	53	0.0419	-3.1717
5	57	0.0441	-3.1223
…	…	…	…
246	49	0.0328	-3.4186
247	54	0.0432	-3.1409
248	61	0.0380	-3.2705
249	51	0.0380	-3.2705
250	38	0.0060	-5.1162
		Sum =	-915.374

Auditioning Values Of μ In Integer Increments



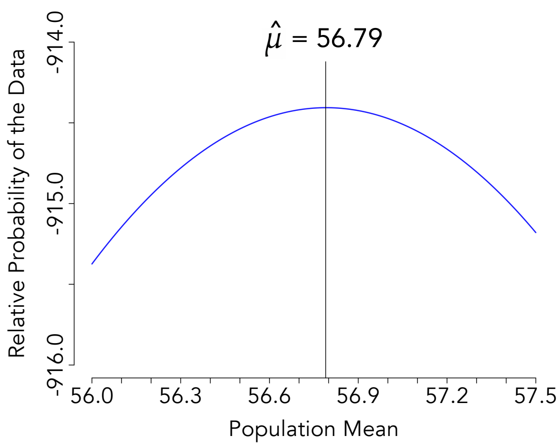
What We Know So Far

Integer search values between 50 and

Log likelihood values improve (get less negative) as the mean increases from 50 to 56, and they get worse from 58 to 64

The mean that maximizes the probability of the data falls somewhere between 56 and 58

Auditioning Values Of μ In .01 Increments



μ	$\log L$
56.76	-914.4077
56.77	-914.4069
56.78	-914.4063
56.79	**-914.4061**
56.80	-914.4062
56.81	-914.4066
56.82	-914.4073

Conclusion

The population mean that maximizes the probability (likelihood) of the data is 56.79

The maximum likelihood estimate is $\hat{\mu}$ = 56.79

Same result as the arithmetic mean

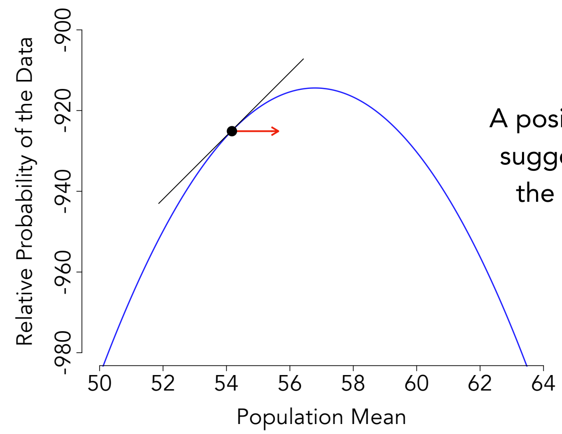
Improving The Search With Calculus

Differential calculus rules give the first derivative (slope) of the log likelihood function at the parameter’s current value

The sign and magnitude of the slope inform the adjustments to the parameters from one iteration to the next

33

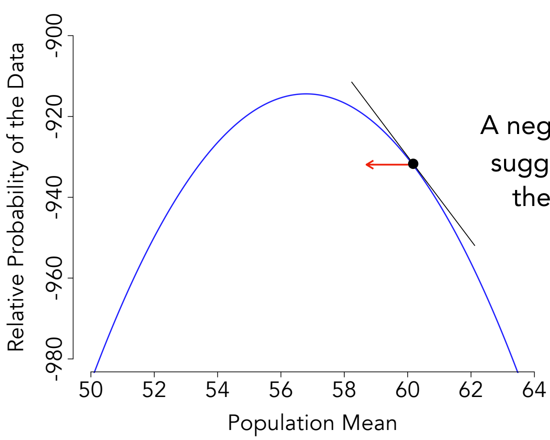
Positive Slope (First Derivative)



A positive first derivative suggests an increase in the parameter value

34

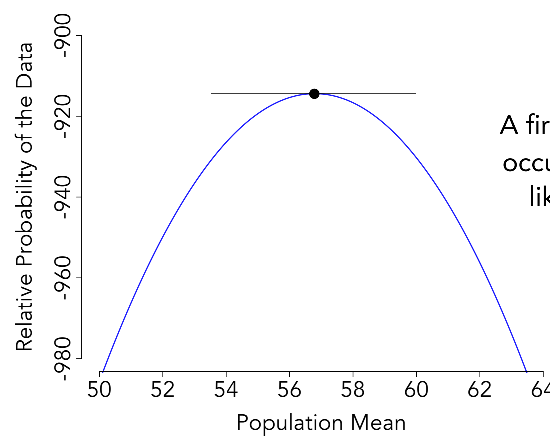
Negative Slope (First Derivative)



A negative first derivative suggests an decrease in the parameter value

35

Zero Slope (First Derivative)



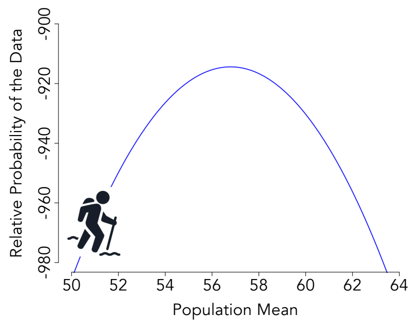
A first derivative of zero occurs at the maximum likelihood estimate

36

Maximum Likelihood = Hill Climbing

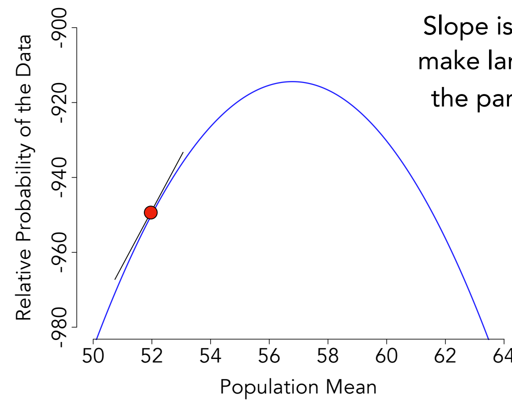
ML estimation is akin to climbing to the top of a hill

Each successive step adjusts the parameters in a direction increases the likelihood until reaching the top of the hill



37

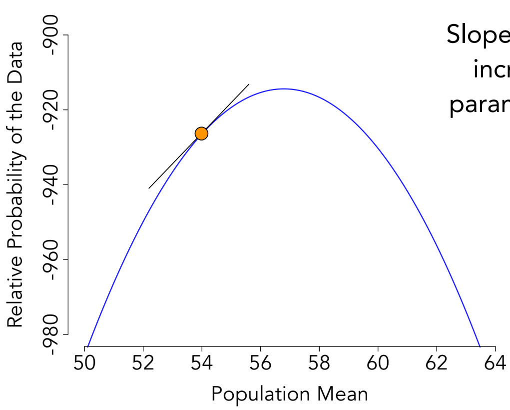
Maximum Likelihood = Hill Climbing



Slope is very positive, make large increase in the parameter value

38

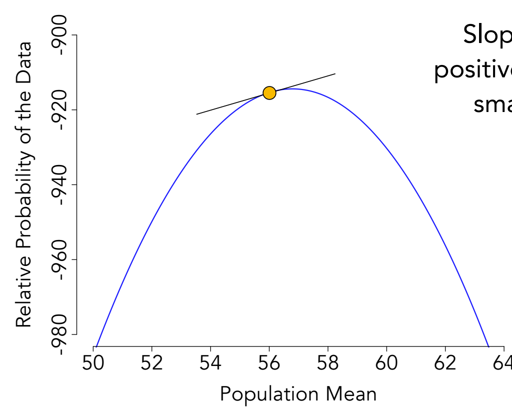
Maximum Likelihood = Hill Climbing



Slope is positive, increase the parameter value

39

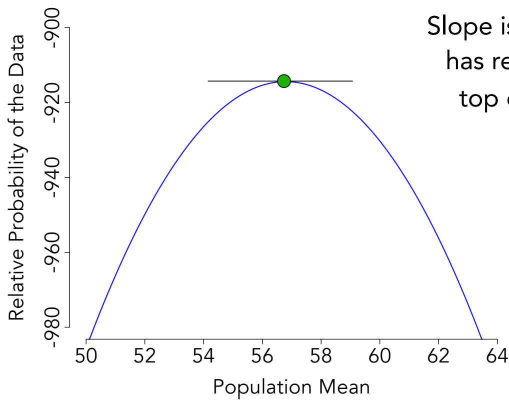
Maximum Likelihood = Hill Climbing



Slope is slightly positive, increase by small amount

40

Maximum Likelihood = Hill Climbing



Slope is zero, climb has reached the top of the hill

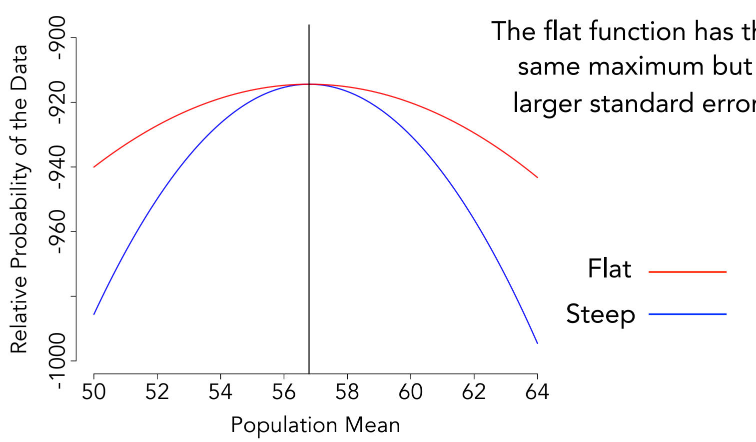
Curvature Of The Log Likelihood

A function's curvature (second derivative) near its maximum determines the standard errors

Steep functions have small standard errors because small changes to the parameter are impactful to the log likelihood

Shallow functions have large standard errors

Two Functions With Different Curvature



The flat function has the same maximum but larger standard error

Mplus Maximum Likelihood Script

```
DATA:
file = math.dat;
VARIABLE:
names = id male lunchasst achievegrp stanread efficacy
anxiety mathpre mathpost;
usevariables = mathpost;
ANALYSIS:
estimator = ml;
MODEL:
[mathpost]; mathpost;
```

Mplus Output

```
MODEL FIT INFORMATION

Number of Free Parameters                         2

Loglikelihood

          H0 Value                         -913.999
          H1 Value                         -913.999
```


Mplus Output

```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 Means
    MATHPOST          56.792      0.592     95.877      0.000

 Variances
    MATHPOST          87.717      7.846     11.180      0.000
```