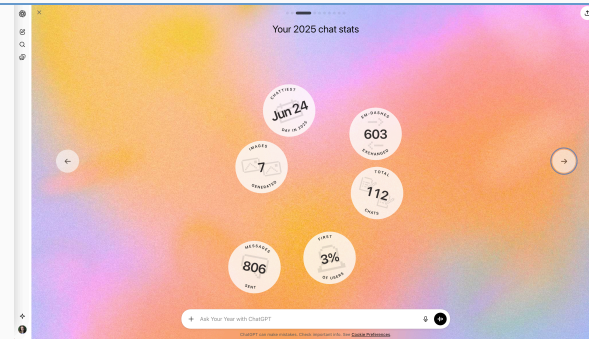


Statistical Theory for Data Science

STA2212H S LEC9101

Week 1

January 6 2025



Settings page for "cursor.com/2025" with a dark background and a grid of icons on the right. It includes the following information:

- Joined 291 Days Ago
- Models:
 - 1 Auto
 - 2 Claude 4.5 Sonnet
 - 3 Claude 3.5 Sonnet
- Agents: 701
- Streak: 6d
- Tokens: 2770M
- Download and Copy buttons at the bottom.



×



Your 2025 chat stats



+ Ask Your Year with ChatGPT  

Joined 291 Days Ago

Models

- 1 Auto
- 2 Claude 4.5 Sonnet
- 3 Claude 3.5 Sonnet

Agents

701

Tabs

2.4K

Tokens

277.0M

Streak

6d



cursor.com/2025

Download

Copy


Joined 291 Days Ago

Models

- 1 Auto
- 2 Claude 4.5 Sonnet
- 3 Claude 3.5 Sonnet

Agents	Tabs
701	2.4K

Tokens	Streak
277.0M	6d

 cursor.com/2025

[Download](#) [Copy](#)

Data \neq Information

1. Course Overview
2. Inference basics: estimation, testing, prediction [AoS Chs 6](#)
3. Likelihood basics: definitions, estimation, nuisance parameters [AoS Ch 6](#); [LaE Ch 1](#)
4. Statistics in the News

Upcoming: Toronto Data Workshop [Link](#)

Wednesday 7 January 2026, noon (EST) on [Zoom](#)

Ciara Zogheib, University of Toronto

“Government Data and AI Strategies: A Case Study of Technology Policies in Canada’s Federal Service”

[Link](#)

STA 2212S: Statistical Theory for Data Science

Tuesday, 5.10pm - 8.00pm

January 6 – March 31 2026

Course description (new!)

This course is a continuation of STA2112H. It is designed for graduate students in statistics and biostatistics. Topics include: Likelihood and Bayesian inference, theory of point estimation, hypothesis and significance testing, inference with missing data, causal inference, estimation and testing for high-dimensional data.

Prerequisite: STA2112H

Course content

Statistical Theory for Data Science II, January 6 2026
The course [Quercus page](#) has

STA 2212S: Statistical Theory for Data Science Syllabus

[Link](#)

Spring 2026

Week	Date	Methods	References
1	Jan 6	Likelihood inference	AoS §9.3ff; LaE §§1.1–1.3, 2.1,2; SM Ch 4
2	Jan 13	Bayesian inference	AoS §§ 11.1–4; LaE §§1.4, 2.8; SM §§11.1,2
3	Jan 20	Point and interval estimation	AoS §§6.3, 7.2; SM §§7.1,2 11.5.2; MS §§4.6–8; §§6.1,3–5
4	Jan 27	Hypothesis testing and significance testing	AoS Ch 10; LaE §1.4; SM §7.3; MS 7.3,4

[Link](#)

Questions Week 1

STA 2212S 2026

1. Suppose $(X_1, Y_1), \dots, (X_n, Y_n)$ are independent pairs of random variables where X_i and Y_i are i.i.d. $N(\mu, \sigma^2)$ random variables:

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x_i - \mu)^2\right\}; \quad f(y_i; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \mu)^2\right\};$$

- (a) Find the maximum likelihood estimators $\hat{\mu}$ and $\hat{\sigma}^2$ of μ and σ^2 .
- (b) Show that $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$ as $n \rightarrow \infty$.
- (c) Suppose now that each pair (X_i, Y_i) has a different expected value, $\mu_i, i =$

Link

Project Guidelines

STA 2212S: Statistical Theory for Data Science, 2026

The final project involves reading and reporting on a paper in the statistical literature, or a paper that uses statistical methods from the course. A list of potential papers will be provided. You will work in teams of two.

Presentation on March 31, 2026.

Report submission due April 14, 2026.

Part 1: Presentation [10 points]

On the last day of class (March 31), you will present your final project; presentations will be 10 minutes long. Detailed guidance on the presentation will be provided.

Part 2: Write-up [30 points]

Your write-up should be: (1) no more than 10 pages, 12 point font, 1.5 vertical spacing; (2) Contain the four sections below, each partner to complete two sections; (3) Include a title page with the title and authors of the paper, the first and last names of the report authors and which section they wrote. (4) Include a list of references.

My likelihood cheatsheet is available [here](#)

STA2212: Inference and Likelihood

A. Notation

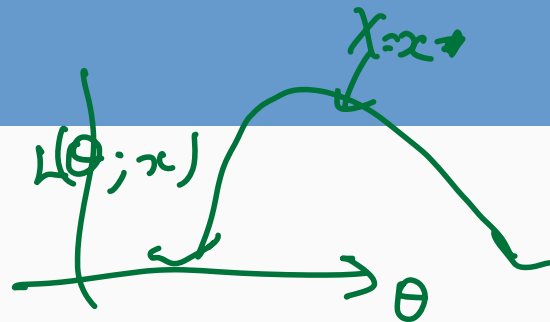
One random variable: Given a model for X which assumes X has a density $f(x; \theta)$, $\theta \in \Theta \subset \mathbb{R}^k$, we have the following definitions:

likelihood function	$L(\theta; x) = c(x)f(x; \theta)$
log-likelihood function	$\ell(\theta; x) = \log L(\theta; x) = \log f(x; \theta) + a(x)$
score function	$u(\theta) = \partial \ell(\theta; x) / \partial \theta \quad s(x; \theta) \quad (9.9)$
observed information function	$j(\theta) = -\partial^2 \ell(\theta; x) / \partial \theta \partial \theta^T$
expected information (in one observation)	$i(\theta) = \mathbb{E}_\theta \{U(\theta)U(\theta)^T\}^{-1} \quad I(\theta) \quad (9.11)$

Independent observations: When we have X_i independent, identically distributed from $f(x_i; \theta)$, then, denoting the observed sample $\mathbf{x} = (x_1, \dots, x_n)$ we

Likelihood function: definition

$$L(\theta; \underline{x}) \propto \underbrace{f(\underline{x}; \theta)}_{\substack{\uparrow \\ X_1, \dots, X_n \text{ iid}}} \leftarrow \text{prob}(\underline{X} = \underline{x} ; \theta)$$



X_1, \dots, X_n iid

Example 1: $X_i \sim \text{Geom}(\theta)$, $i = 1, \dots, n$

$$f(x) = \theta(1-\theta)^{x-1}, x = 1, \dots, 0 < \theta < 1$$

$$L(\theta; \underline{x}) \propto f(\underline{x}; \theta) = \prod_{i=1}^n f(x_i; \theta) = \frac{\theta^n (1-\theta)^{\sum (x_i - 1)}}{\theta^n (1-\theta)^{\sum x_i}}; \theta < 1$$

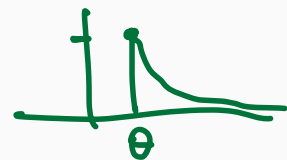
$$l(\theta; \underline{x}) = \log L(\theta; \underline{x}) \quad \text{log-lik.}$$

$$\theta_{\text{true}} \quad \theta_0 \quad \theta_x = 0.5$$

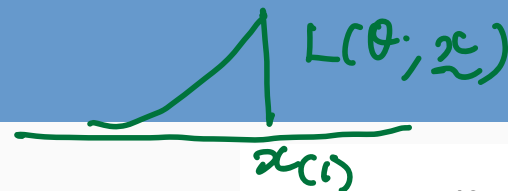
Example 2: $X_i \sim \text{LocExp}(\theta)$, $i = 1, \dots, n$

$$f(x) = \exp\{-(x - \theta)\}, x > \theta, \theta > 0$$

$$f(\underline{x}; \theta) = \prod_{i=1}^n e^{-(x_i - \theta)}, x_i > \theta$$



$$L(\underline{x}; \theta) = e^{-\sum (x_i) + n\theta}, x_{(n)} > \theta \propto e^{n\theta} \mathbb{1}\{\theta < x_{(n)}\}$$



```
geomlik <- function(theta,x){
  theta^length(x)*(1-theta)^(sum(x)-length(x))}

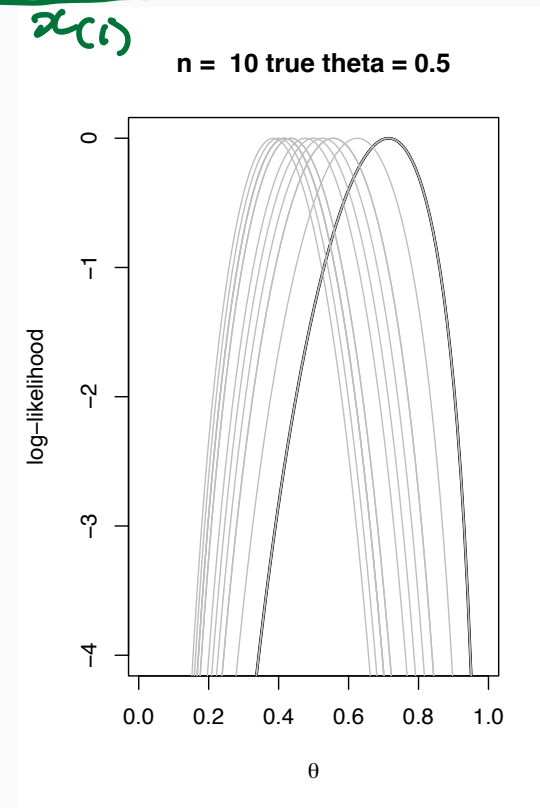
geomllik <- function(theta, x){
  log(geomlik(theta,x))-max(log(geomlik(theta,x)))}

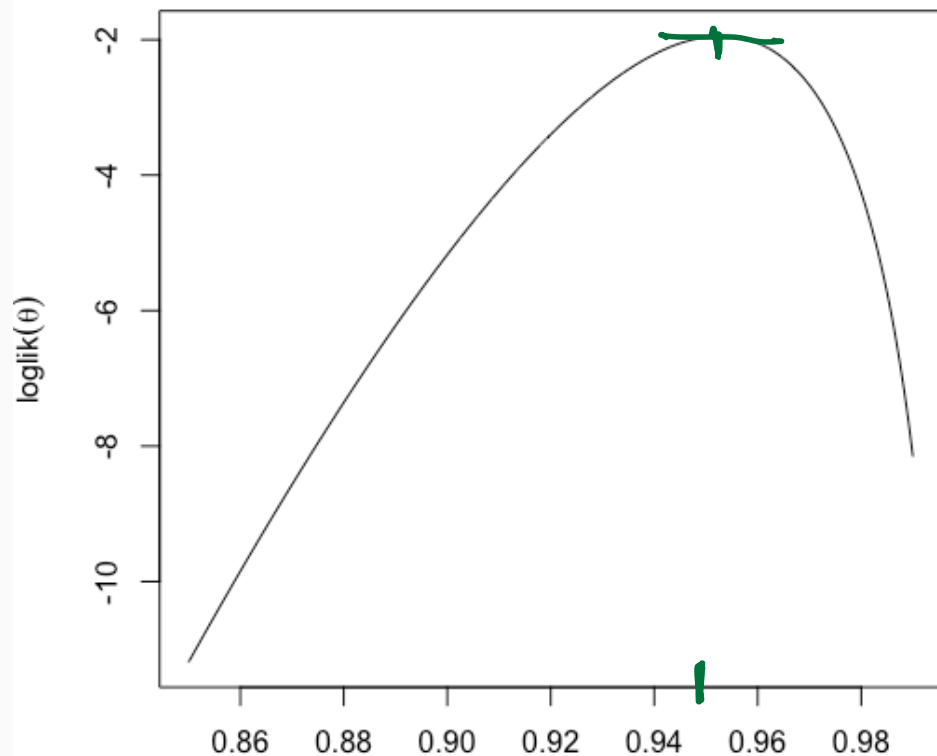
n <- 10; prob <- 0.5
x <- rgeom(n, prob) + 1 #R definition different from mine

thvals <- seq(0,1,length=100)

plot(thvals,geomllik(thvals, x), type="l", lwd=2)

for(i in 1:15){
  x <- rgeom(n,prob)+1
  lines(thvals,geomllik(thvals,x), col="gray") }
```





$$l(\theta; \underline{x}) = \log f(\underline{x}; \theta)$$

$$\underline{\underline{l'(\theta; \underline{x}) = \frac{\partial}{\partial \theta} l(\theta; \underline{x})}}$$

score function

$$l''(\theta; \underline{x}) = \frac{\partial^2}{\partial \theta^2} l(\theta; \underline{x})$$

observed information

$$\underline{\underline{I_n(\theta) = \text{var}_{\theta}\{l'(\theta; \underline{x})\}}}$$

expected information

$$\hat{\theta} = \hat{\theta}(\underline{x})$$

$$\underline{\underline{l'(\hat{\theta}; \underline{x}) = 0}} \text{ defines } \hat{\theta}$$

• model $X \sim f(x; \theta)$, $\theta \in \mathbb{R}^p$ is a column vector $X \in \mathbb{R}^n$

• $L(\theta; \underline{x})$ $L(\underline{\theta}; \underline{x}) = : \mathbb{R}^p \rightarrow \mathbb{R}$ for fixed \underline{x} map from $\mathbb{R}^p \rightarrow \mathbb{R}$

• $l'(\theta; \underline{x})$ $\frac{\partial}{\partial \underline{\theta}} l(\underline{\theta}; \underline{x}) = \begin{pmatrix} \frac{\partial l}{\partial \theta_1} \\ \vdots \\ \frac{\partial l}{\partial \theta_p} \end{pmatrix} = \underline{u}(\theta; \underline{x})$ $p \times 1$ vector

• $-\underline{l}''(\theta; \underline{x})$ $p \times p$ matrix

$$-\frac{\partial^2 l(\theta; \underline{x})}{\partial \underline{\theta} \partial \underline{\theta}^T} = \begin{bmatrix} -\frac{\partial^2 l}{\partial \theta_1^2} & & \\ & \ddots & \\ & & -\frac{\partial^2 l}{\partial \theta_p^2} \end{bmatrix}_{p \times p} \quad \underline{u}(\theta) = \underline{u}(\theta; \underline{x})$$

$$I_n(\theta) = \text{var}_{\theta} \{ l'(\theta; \underline{X}) \} = \text{var}_{\theta} \{ \underline{u}(\theta) \}_{p \times 1} = \int \underline{u}(\theta) \underline{u}(\theta)^T f(\underline{x}; \theta) d\underline{x} - \left[\int \underline{u}(\theta) f(\underline{x}; \theta) d\underline{x} \right] \left[\int \underline{u}(\theta) f(\underline{x}; \theta) d\underline{x} \right]^T$$

- model $X \sim f(x; \theta), \theta \in \mathbb{R}^p$ θ is a **column vector**

- $L(\theta; \underline{x})$ $\ell'(\hat{\theta}; \underline{x}) = 0$ defines $\hat{\theta}(\underline{x})$ map from $\mathbb{R}^p \rightarrow \mathbb{R}$

- $\ell'(\theta; \mathbf{x})$ iid $\sum_{i=1}^n \ell'(\hat{\theta}; x_i) = 0$ ✓ $p \times 1$ vector

- $-\ell''(\theta; \mathbf{x})$ $p \times p$ matrix

$$0 = \sum_{i=1}^n \ell'(\theta; x_i) + (\hat{\theta} - \theta) \sum \ell''(\theta; x_i) + \frac{1}{2} (\hat{\theta} - \theta)^2 \ell''' \dots$$

$$I = E\{-\ell''(\mathbf{x})\} \text{ (Thm 9.17)}$$

var-cov. matrix of $U(\theta; X)$ $\frac{\partial \log f(x; \theta)}{\partial \theta}$
 var of $U(\theta; X)$ $\theta \in \mathbb{R}$ $X \in \mathbb{R}$ $= U(\theta; X)$

$$\begin{aligned} \text{var}_{\theta} U(\theta; X) &= \int U^2(\theta; x) f(x; \theta) dx - \left(\int U(\theta; x) f(x; \theta) dx \right)^2 \\ &= \int \left(\frac{\partial \log f(x; \theta)}{\partial \theta} \right)^2 f(x; \theta) dx - \left(\int \frac{\partial \log f(x; \theta)}{\partial \theta} f(x; \theta) dx \right)^2 \\ &= \int \left(\frac{\partial f(x; \theta) / \partial \theta}{f(x; \theta)} f(x; \theta) \right)^2 dx - \left(\int \frac{\partial f(x; \theta)}{\partial \theta} dx \right)^2 = 0 \end{aligned}$$

$E_{\theta} U(\theta; X) = 0$

$$\begin{aligned} \text{var}_{\theta} U(\theta; X) &= \int \left(\frac{\partial \log f(x; \theta)}{\partial \theta} \right)^2 f(x; \theta) dx \\ &= - \int \frac{\partial^2 \log f(x; \theta)}{\partial \theta^2} \cdot f(x; \theta) dx \\ &= E_{\theta} \left(- \frac{\partial^2 \log f(x; \theta)}{\partial \theta^2} \right) \checkmark \\ &= \int \left\{ \frac{\partial f(x; \theta)}{\partial \theta} \right\}^2 f(x; \theta) dx \end{aligned}$$

$E_{\theta} U(\theta; X) = 0$
 $\text{var}_{\theta} U(\theta; X) = - E_{\theta} \langle l''(\theta; X) \rangle$

$$\int f(x; \theta) dx = 1 \quad \frac{\partial}{\partial \theta} \int \dots = 0 \quad \frac{\partial^2}{\partial \theta^2} \int \dots = 0$$

$$0 = \frac{\partial}{\partial \theta} \int \frac{\partial \log f(x; \theta)}{\partial \theta} f(x; \theta) dx = \int \frac{\partial^2 \log f(x; \theta)}{\partial \theta^2} f(x; \theta) dx + \int \left(\frac{\partial \log f(x; \theta)}{\partial \theta} \right)^2 f(x; \theta) dx$$

Example: logistic regression

$$E(\hat{\beta}_j) + \text{var}$$

```
Boston$scrim2 <- Boston$scrim > median(Boston$scrim) # define binary response
Boston.glm <- glm(crim2 ~ . - crim, family = binomial,
                 data = Boston) #fit logistic regression
summary(Boston.glm)
```

Coefficients:

	$\hat{\beta}$	$\hat{se}(\hat{\beta})$	$\hat{\beta}/\hat{se}$	p-value
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-34.103704	6.530014	-5.223	1.76e-07 ***
zn	-0.079918	0.033731	-2.369	0.01782 *
indus	-0.059389	0.043722	-1.358	0.17436
chas	0.785327	0.728930	1.077	0.28132
nox	48.523782	7.396497	6.560	5.37e-11 ***
rm	-0.425596	0.701104	-0.607	0.54383
age	0.022172	0.012221	1.814	0.06963 .

$j = 1, \dots, 6$

$$\hat{\theta}_j = \hat{\beta}_j \sim \mathcal{N}(0, \hat{se}_j)$$

... Example: logistic regression

(j,j) element \rightarrow
 $\phi \{I_n^{-1}(\theta)\}$

```
Boston.glm <- glm(crim2 ~ . - crim, family = binomial,  
                  data = Boston) #fit logistic regression
```

```
confint(Boston.glm)
```

```
Waiting for profiling to be done...
```

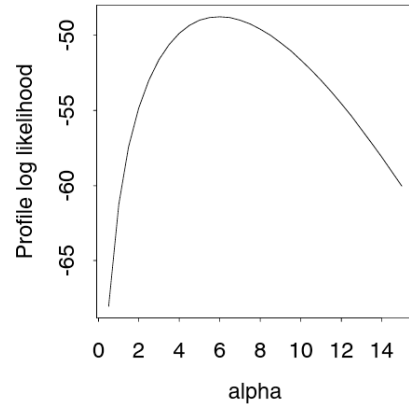
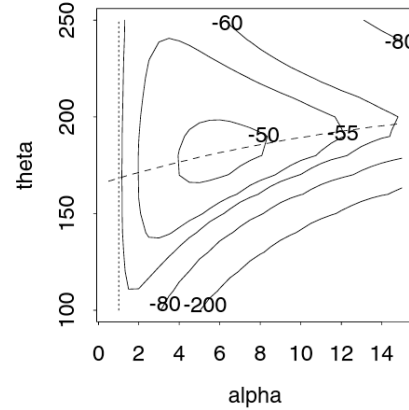
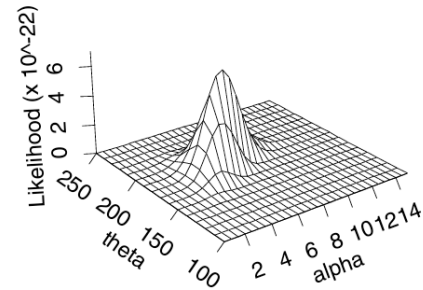
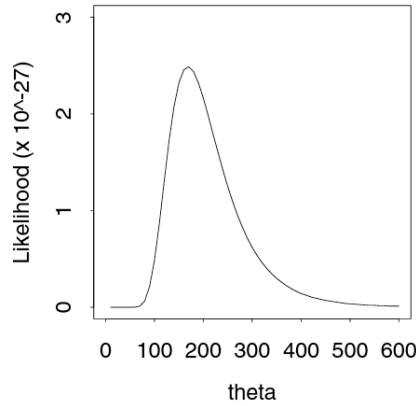
	2.5 %	97.5 %
(Intercept)	-47.480389822	-21.699753794
zn	-0.152359922	-0.020567540
indus	-0.149113408	0.024168460
chas	-0.646429219	2.233443233
nox	34.967619055	64.088411260
rm	-1.811639107	0.950196261
age	-0.001231256	0.046865843
dis	0.280762523	1.140619391
rad	0.376833861	0.975898274
tax	-0.012038221	-0.001324887

Waiting for profiling to be done – what's profiling?

4.1 · Likelihood

95

Figure 4.1 Likelihoods for the spring failure data at stress 950 N/mm^2 . The upper left panel is the likelihood for the exponential model, and below it is a perspective plot of the likelihood for the Weibull model. The upper right panel shows contours of the log likelihood for the Weibull model; the exponential likelihood is obtained by setting $\alpha = 1$, that is, slicing L along the vertical dotted line. The lower right panel shows the profile log likelihood for α , which corresponds to the log likelihood values along the dashed line in the panel above, plotted against α .



Properties of maximum likelihood estimators

if $\frac{\hat{\theta} - \theta}{se} \xrightarrow{d} N(0, 1)$ AoS 9.4,5,6,7

- maximum likelihood estimators are **equivariant**

if $\tau = g(\theta)$ & g is 1-1 then $\hat{\tau} = g(\hat{\theta})$

$\arg \sup_{\theta} L(\theta; \underline{x}) \xrightarrow{\wedge} \theta \xrightarrow{\log} \log \theta$

$E\hat{\theta} = \theta$
 $+$
 $E\hat{\tau} = \tau$

Thm 9.14

special exceptions

- maximum likelihood estimators are **biased**

$E_{\theta} \hat{\theta} \neq \theta$ "except in special cases"

~~" $E_{\theta} \hat{\theta} \approx \theta$ "~~

~~" $E_{\theta}(\hat{\theta}) = \theta$ "~~

Under reg $\hat{\theta} \xrightarrow{P} \theta_*$ when $\underline{X} \sim f(x, \theta_*)$ **Consistent**

- maximum likelihood estimators have no explicit formula

in general

$$l'(\theta; \underline{x}) = 0 \implies \hat{\theta}(\underline{x})$$

- maximum likelihood estimators minimize the KL-divergence to the data

- maximum likelihood estimators minimize the KL-divergence to the data

- KL divergence from f_{θ_*} true to f_{θ} model :

$$D(\theta_*, \theta): D(f_{\theta_*}; f_{\theta}) \equiv E_{f_{\theta_*}} \log \left\{ \frac{f_{\theta_*}(X)}{f_{\theta}(X)} \right\} = -E_{f_{\theta_*}} \log \{f(X; \theta)\} + E_{f_{\theta_*}} \log f_{\theta_*}(X)$$

Handwritten notes: "1 r.v." with an arrow pointing to $f(X; \theta)$; underlines under f_{θ_} , f_{θ} , and $f_{\theta_*}(X)$; arrows pointing to the expectation operators.*

- estimate of $E_{f_{\theta_*}} \log \{f(X; \theta)\}$?

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

Handwritten note: A green oval circles the entire equation above.

- minimize $D(f_{\theta_*}; f_{\theta})$ same as maximize $\ell(\theta; x_1, \dots, x_n)$

Handwritten note: use K-L divergence bet from f_{θ_} to \hat{F}_n empirical cdf.*

- maximum likelihood estimators are (i) consistent, (ii) **asymptotically normal**
- (ii) TS expansion

$$\hat{\theta}_n \Rightarrow \hat{\theta}_n(\underline{X}) = \hat{\theta}(X_1, \dots, X_n) \quad (\theta \in \mathbb{R})$$

$$\xrightarrow{P} \theta_*$$

mean

? $\exists \varepsilon, \delta > 0$ s.t. $P_{\theta_*} \{ |\hat{\theta}_n - \theta_*| > \varepsilon \} < \delta$ for all $n \geq n_0$?
 , and n_0

$\hat{\theta}_n \xrightarrow{p} \theta_*$ when X_1, \dots, X_n iid $f(x; \theta_*)$ cons.

$$\sqrt{n}(\hat{\theta}_n - \theta_*) \xrightarrow{d} \mathcal{N}(0, I^{-1}(\theta_*)) \quad I(\theta) = E_{\theta}\{-l''(\theta; X)\}$$

$$= \text{var}_{\theta}\{l'(\theta; X)\}$$

$$l'(\hat{\theta}) = 0 = l'(\theta) + (\hat{\theta} - \theta)l''(\theta) + R_n$$

$$\frac{1}{2}(\hat{\theta} - \theta)^2 \underline{l'''(\tilde{\theta})}$$

$|\tilde{\theta} - \theta| < |\hat{\theta} - \theta|$

$$(\hat{\theta}_n - \theta) = \frac{l'(\theta)}{-l''(\theta)} + R_n$$

$$\sqrt{n}(\hat{\theta}_n - \theta) = \frac{\frac{1}{\sqrt{n}} \sum_{i=1}^n l'(\theta; X_i)}{-\frac{1}{n} \sum_{i=1}^n l''(\theta; X_i)} \xrightarrow[\text{WLLN}]{\text{(CLT)}} \mathcal{N}\left(0, \frac{\text{var}_{\theta} l'(\theta; X)}{I_1(\theta)}\right)$$

$$\xrightarrow{p} E_{\theta}\{l''(\theta; X)\} = I_1(\theta)$$

$$N(0, I_1^{-1}(\theta))$$

$$\frac{N(0, I_1(\theta))}{I_1(\theta)}$$

$$I_1(\theta)$$

$$\sqrt{n}(\hat{\theta}_n - \theta) I_1^{-1/2}(\theta) = \frac{\frac{1}{\sqrt{n}} \sum \ell'(\theta) \cdot I_1^{-1/2}(\theta)}{\left\{ -\frac{1}{n} \sum \ell''(\theta) \right\} I_1^{-1}(\theta)}$$

$$\frac{1}{\sqrt{n}} \sum \{ \ell'(\theta) I_1^{-1/2}(\theta) \} \xrightarrow{d} N(0, 1) \text{ by CLT}$$

$$-\frac{1}{n} \sum_{i=1}^n \frac{\ell''(\theta)}{I_1(\theta)} \xrightarrow{P} 1$$

Suppose

$$\theta \in \mathbb{R}^p, \mathbf{x} = (x_1, \dots, x_p)$$

$$a_n(\mathbf{x} - \theta) \xrightarrow{d} \mathbf{Z},$$

and $g(\mathbf{x})$ is continuously differentiable at θ , then

$$\{g_1(\mathbf{x}), \dots, g_k(\mathbf{x})\}$$

$$a_n\{g(\mathbf{x}) - g(\theta)\} \xrightarrow{d} D(\theta)\mathbf{Z}$$

where $D(\theta) =$

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\{\mathbf{o}, I^{-1}(\theta)\}$$

$$\sqrt{n}\{g(\hat{\theta}_n) - g(\theta)\} \xrightarrow{d} N\{\mathbf{o}, g'(\theta)^T I^{-1}(\theta) g'(\theta)\}$$

Example

X_1, \dots, X_n i.i.d. Gamma (α, λ)

$$f(x_i; \lambda, \alpha) = \frac{1}{\Gamma(\alpha)} \lambda^\alpha x_i^{\alpha-1} \exp(-\lambda x_i)$$

find $\text{a.var}(\hat{\mu})$ via mv delta method

Newton-Raphson:

$$0 = \ell'(\hat{\theta}) \approx \ell'(\theta_0) + \ell''(\theta_0)(\hat{\theta} - \theta_0)$$

$$\hat{\theta} \approx \theta_0 - \{\ell''(\theta_0)\}^{-1}\ell'(\theta_0)$$

- suggests iteration

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} + \{-\ell''(\hat{\theta}^{(k)})\}^{-1}\ell'(\hat{\theta}^{(k)}) = \hat{\theta}^{(k)} + \frac{S(\hat{\theta}^{(k)})}{H(\hat{\theta}^{(k)})}$$

- requires reasonably good starting values for convergence
- need $-\ell''(\hat{\theta}^{(k)})$ to be non-negative definite
- **Fisher scoring** replaces $-\ell''(\cdot)$ by its expected value $J(\cdot)$
- N-R and F-S are gradient methods; many improvements have been developed
- solution is a **global max** only if $\ell(\theta)$ is concave

E-M algorithm:

procedure

- complete data $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}; \theta)$
- observed data $y = (y_1, \dots, y_m)$, with $y_i = g_i(\mathbf{x})$
- joint density $f_Y(y; \theta) = \int_{A(y)} f_{\mathbf{X}}(\mathbf{x}; \theta) d\mathbf{x}$
- algorithm:

many-to-one

$$A(y) = \{\mathbf{x}; y_i = g_i(\mathbf{x}), i = 1, \dots, m\}$$

1. (E step) estimate the **complete data** log-likelihood function for θ using current guess $\hat{\theta}^{(k)}$
2. (M step) maximize that function over θ and update to $\hat{\theta}^{(k+1)}$ usually by N-R or Fisher scoring

- likelihood function increases at each step
- can be implemented in complex models
- doesn't automatically provide an estimate of the asymptotic variance

but methods exist to obtain this as a side-product

- $f_X(x_i; \lambda, \mu, \theta) = \alpha \frac{e^{-\lambda} \lambda^{x_i}}{x_i!} + (1 - \alpha) \frac{e^{-\mu} \mu^{x_i}}{x_i!}, \quad x = 1, 2, \dots; \lambda, \mu > 0, 0 < \theta < 1$
- Observed data: x_1, \dots, x_n
- Complete data: $(x_1, y_1), \dots, (x_n, y_n); y_i \sim \text{Bernoulli}(\theta)$
- Complete data log-likelihood function:

$$\ell_c(\alpha, \lambda, \mu; \mathbf{y}, \mathbf{x}) = \sum_{i=1}^n y_i \{ \log(\alpha) + x_i \log(\lambda) - \lambda \} + \sum_{i=1}^n (1 - y_i) \{ \log(1 - \theta) + x_i \log(\mu) - \mu \}$$

•

$$E_{\hat{\theta}^{(k)}} \{ \ell_c(\alpha, \lambda, \mu; \mathbf{y}, \mathbf{x}) \mid \mathbf{x} \} = \sum_{i=1}^n \hat{y}_i \{ \log(\alpha) + x_i \log(\lambda) - \lambda \} + \sum_{i=1}^n (1 - \hat{y}_i) \{ \log(1 - \alpha) + x_i \log(\mu) - \mu \}$$

- $\hat{y}_i = E(Y_i \mid x_i; \hat{\theta}^{(k)})$ see p.280 for exact value
- maximizing values of α, λ, μ can be obtained in closed form p.281

AoS likes to work with $\log \mathcal{L}_n(\theta) / \mathcal{L}_n(\hat{\theta}^{(k)})$

General-purpose Optimization

Description

General-purpose optimization based on Nelder–Mead, quasi-Newton and conjugate-gradient algorithms. It includes an option for box-constrained optimization and simulated annealing.

Usage

```
optim(par, fn, gr = NULL, ...,  
      method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B", "SANN",  
                "Brent"),  
      lower = -Inf, upper = Inf,  
      control = list(), hessian = FALSE)
```

```
optimHess(par, fn, gr = NULL, ..., control = list())
```

Arguments

Notes on optimization: Tibshirani, Pena, Kolter CO 10-725 CMU

- Goal: $\max_{\theta} \ell(\theta; \mathbf{x})$
- Solve: $\ell'(\hat{\theta}; \mathbf{x}) = \mathbf{0}$
- Iterate: $\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} + \{j(\hat{\theta}^{(t)})\}^{-1} \ell'(\hat{\theta}^{(t)})$
- Rewrite: $j(\hat{\theta}^{(t)})(\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}) = \ell'(\hat{\theta}^{(t)})$ $B\Delta\theta = -\nabla\ell(\theta)$
- Quasi-Newton:
 - approximate $j(\hat{\theta}^{(t)})$ with something easy to invert
 - use information from $j(\hat{\theta}^{(t)})$ to compute $j(\hat{\theta}^{(t+1)})$
- optimization notes add a step size to the iteration $\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} + \epsilon_t \{j(\hat{\theta}^{(t)})\}^{-1} \ell'(\hat{\theta}^{(t)})$

```
optim(par, fn, gr = NULL, ...,
      method = c("Nelder-Mead", "BFGS", "CG", "L-BFGS-B", "SANN", "Brent"),
      lower = -Inf, upper = Inf, control = list(), hessian = FALSE)
```

Define

$$M_n(\theta) \equiv \frac{1}{n} \sum_{i=1}^n \log \frac{f(X_i; \theta)}{f(X_i; \theta_*)},$$

$$M(\theta) \equiv -D(f_{\theta_*}, f_\theta)$$

Require:

$$\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \xrightarrow{p} \mathbf{0},$$

and, for every $\epsilon > 0$,

$$\sup_{\theta: |\theta - \theta_*| < \epsilon} M(\theta) < M(\theta_*).$$

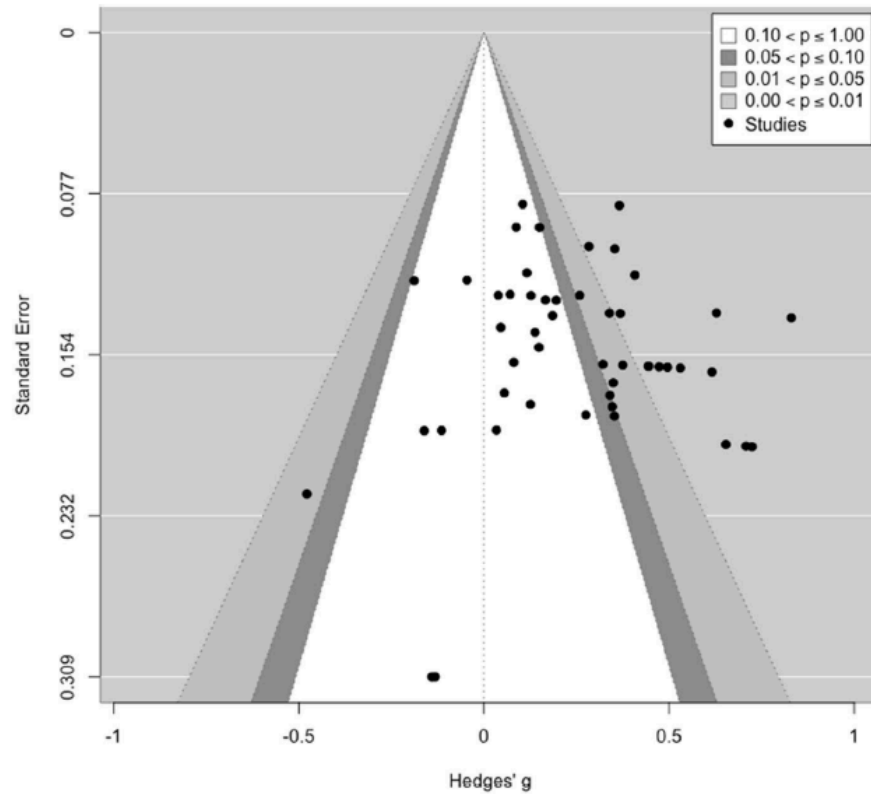
Proof of consistency: Show

$$\text{pr}(|\hat{\theta}_n - \theta_*| > \epsilon) \leq \text{pr}\{M(\hat{\theta}_n) < M(\theta_* - \Delta)\} \rightarrow \mathbf{0}.$$

- (B1) The parameter space Θ is an open subset of \mathbb{R}^p
- (B2) The set $A = \{x : f(x; \theta) > 0\}$ does not depend on θ
- (B3) $\ell(\theta)$ is three times continuously differentiable on A
- (B4) $E_{\theta}\{\ell'(\theta; X_i)\} = 0 \forall \theta$ and $\text{Cov}\{\ell'(\theta; X_i)\} = I(\theta)$ is positive definite $\forall \theta$
- (B5) $E_{\theta}\{-\ell''(\theta; X_i)\} = J(\theta)$ is positive definite $\forall \theta$
- (B6) For each $\theta, \delta > 0, 1 \leq j, k, l, \leq p,$

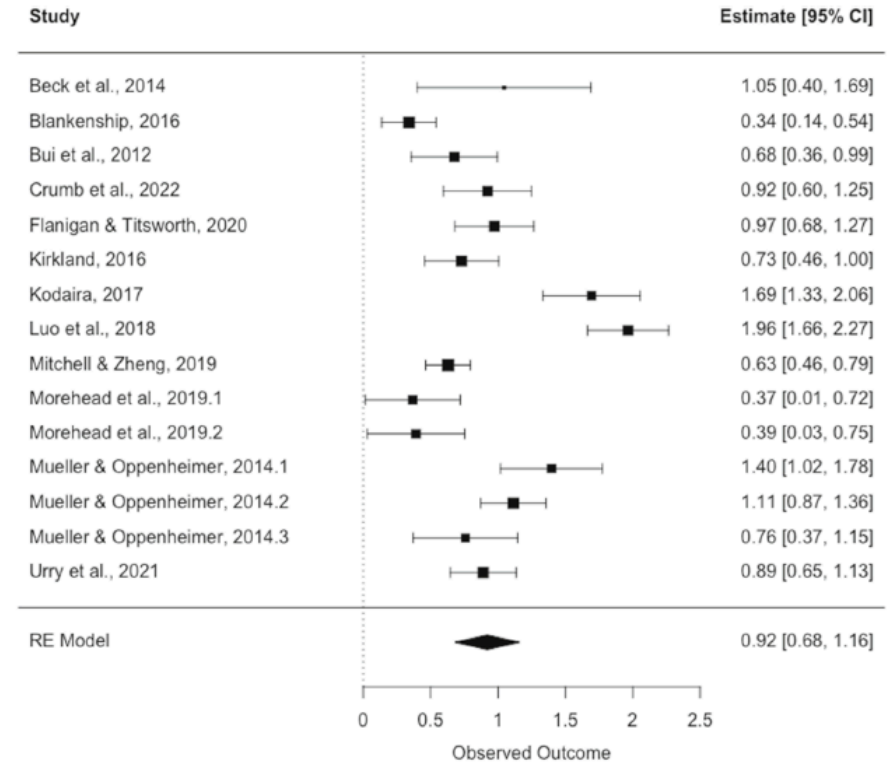
$$\left| \frac{\partial^3 \ell(\theta_*; X_i)}{\partial \theta_j \partial \theta_k \partial \theta_l} \right| \leq M_{jkl}(\theta_*),$$

for $\|\theta - \theta_*\| \leq \delta$, where $E_{\theta}\{M_{jkl}(X_i)\} < \infty$



Note. A positive standardized effect size indicates that students who handwrote their notes had higher achievement than those who typed their notes.

Fig. 1 Funnel plot depicting Hedges' g and precision in assessing achievement: handwritten vs. typed



Note. A positive standardized effect size indicates that students who typed their notes recorded a larger number of words compared to those who handwrote their notes.

SPRINGER NATURE Link

[Find a journal](#)

[Publish with us](#)

[Track your research](#)

[Search](#)

[Home](#) > [Educational Psychology Review](#) > Article

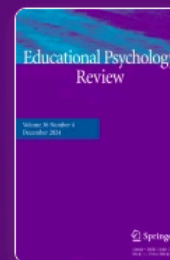
Typed Versus Handwritten Lecture Notes and College Student Achievement: A Meta-Analysis

META-ANALYSIS | Published: 12 July 2024

Volume 36, article number 78, (2024) [Cite this article](#)

Access provided by University of Toronto Robarts Library

[Download PDF](#) 



[Educational Psychology Review](#)

[Aims and scope](#) →

[Submit manuscript](#) →