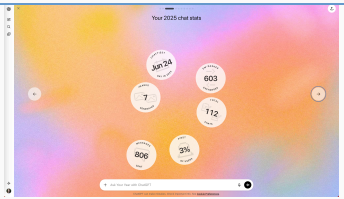
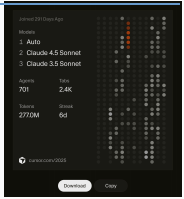
Statistical Theory for Data Science

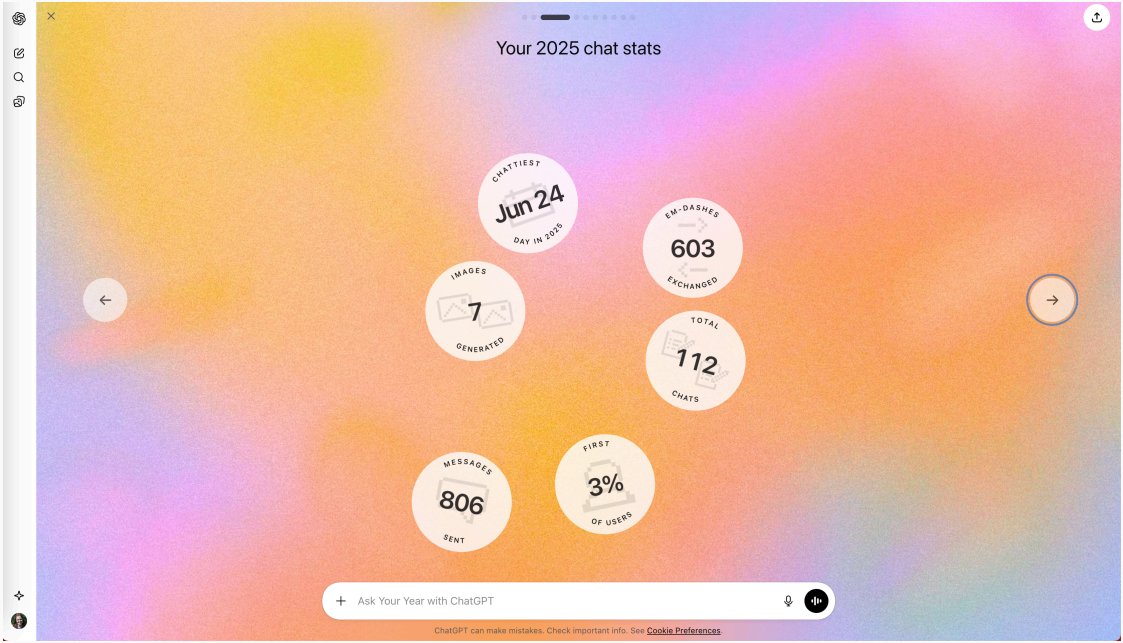
STA2212H S LEC9101

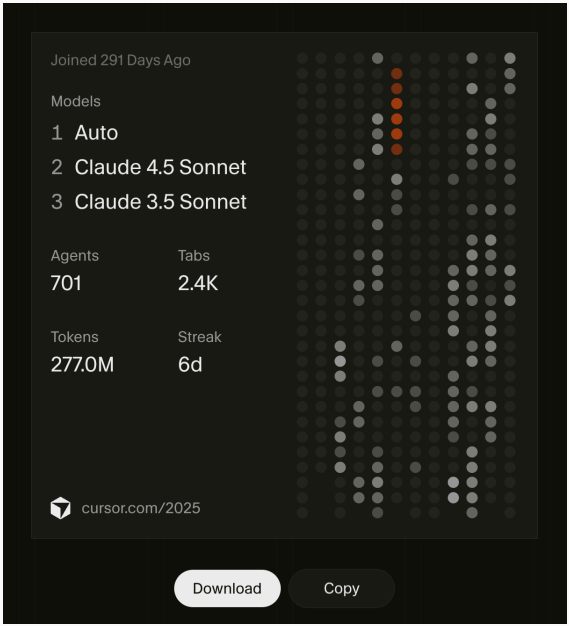
Week 1

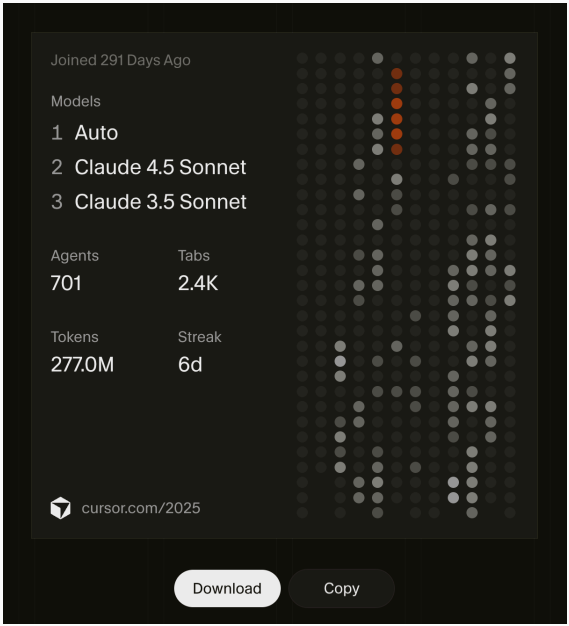
January 6 2025






Your 2025 chat stats
CHATTIEST
Jun 24
DAY IN 2025
EM-DASHES
603
EXCHANGED
IMAGES
7
GENERATED
TOTAL
112
CHATS
MESSAGES
806
SENT
FIRST
3%
OF USERS
Ask Your Year with ChatGPT
ChatGPT can make mistakes. Check important info. See Cookie Preferences


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



Data ≠ Information

Today

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”

Course Overview

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

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
5	Feb 3	Goodness of fit testing, Intro to	AoS §§10.3,4,5,8; MS Ch 9; SM

Link

Questions Week 1 STA 2212S 2026

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

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

   (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 $n \to \infty$.

   (c) Suppose now that each pair (X_i, Y_i) has a different expected value, $\mu_i, i = 1, \ldots, n$. Show that the maximum likelihood estimator $\hat{\sigma}^2 \xrightarrow{p} \sigma^2/2$ as $n \to \infty$.

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 $f(x;\theta)$, $\quad \theta \in \Theta \subset \mathbb{R}^k$, 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	$u(\theta) = \partial\ell(\theta;x)/\partial\theta$	$s(x;\theta)$ (9.9)
observed information function	$j(\theta) = -\partial^2\ell(\theta;x)/\partial\theta\partial\theta^T$	
expected information (in one observation)	$i(\theta) = \mathrm{E}_\theta\{U(\theta)U(\theta)^T\}$[1]	$I(\theta)$ (9.11)

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

Likelihood function: definition

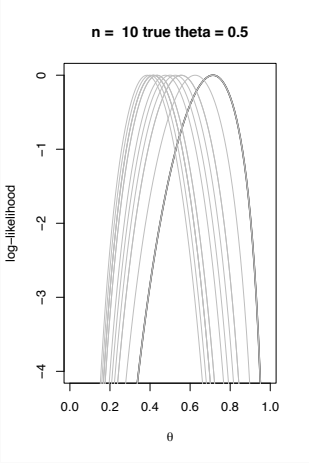
Example 1: $X_i \sim Geom(\theta), \quad i = 1, \ldots, n$ $\qquad f(x) = \theta(1-\theta)^{x-1}, x = 1, \ldots, 0 < \theta < 1$

Example 2: $X_i \sim LocExp(\theta), \quad i = 1, \ldots, n$ $\qquad f(x) = \exp\{-(x-\theta)\}, x > \theta, \theta > 0$

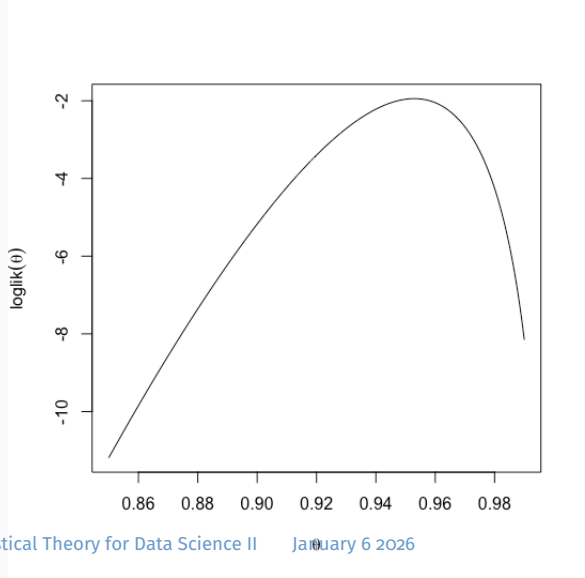
Simulated Example

geomlik.Rmd

```
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") }
```



Likelihood quantities



Vector parameters

AoS §9.10

- model $X \sim f(x; \theta), \theta \in \mathbb{R}^p$ θ is a column vector $X \in \mathbb{R}^n$
- $L(\theta; x)$ map from $\mathbb{R}^p \to \mathbb{R}$
- $\ell'(\theta; x)$ $p \times 1$ vector
- $-\ell''(\theta; x)$ $p \times p$ matrix

- model $X \sim f(x;\theta), \theta \in \mathbb{R}^p$ θ is a column vector
- $L(\theta; x)$ map from $\mathbb{R}^p \to \mathbb{R}$
- $\ell'(\theta; x)$ $p \times 1$ vector
- $-\ell''(\theta; x)$ $p \times p$ matrix

$I = \mathrm{E}\{-\ell''(\boldsymbol{x})\}$ (Thm 9.17)

Example: logistic regression

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

Coefficients:
              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 .
```

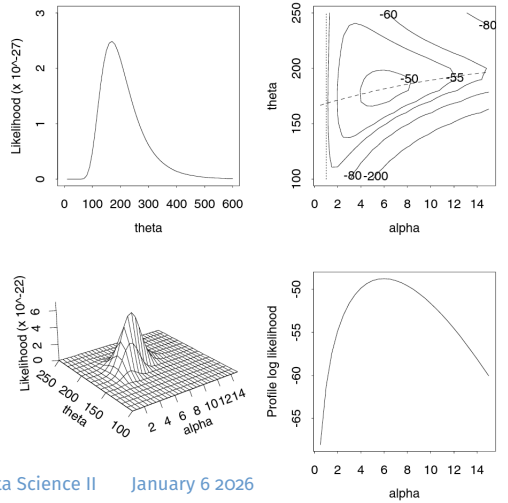
... Example: logistic regression

```
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
```

... Vector parameters

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

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

AoS 9.4,5,6,7

- maximum likelihood estimators are equivariant — Thm 9.14
- maximum likelihood estimators are biased — special exceptions
- maximum likelihood estimators have no explicit formula — in general

...Properties of maximum likelihood estimators

AoS 9.4,5,6,7

- 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 :

$$D(f_{\theta_*}; f_\theta) \equiv \mathrm{E}_{f_{\theta_*}} \log \left\{ \frac{f_{\theta_*}(X)}{f_\theta(X)} \right\} = -\mathrm{E}_{f_{\theta_*}} \log\{f(X;\theta)\} + \mathrm{E}_{f_{\theta_*}} \log f_{\theta_*}(X)$$

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

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

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

Asymptotic properties of MLEs

AoS Thm 9.13, 9.18; MS Thm 5.1–5.4

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

Your friend the delta-method

MS Thm 3.4

Suppose $\theta \in \mathbb{R}^p$, $\boldsymbol{x} = (x_1, \ldots, x_p)$

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

and $g(\boldsymbol{x})$ is continuously differentiable at θ, then $\{g_1(\boldsymbol{x}), \ldots g_k(\boldsymbol{x})\}$

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

where $D(\theta) =$

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\{0, I^{-1}(\theta)\}$$

$$\sqrt{n}\{g(\hat{\theta}_n) - g(\theta)\} \xrightarrow{d} N\{0, g'(\theta)^T I^{-1}(\theta) g'(\theta)\}$$

Example

$X_1, \ldots, X_n$ i.i.d. Gamma (α, λ)

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

... Example

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

Calculating maximum likelihood estimators

AoS 9.13.4

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 $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x}; \theta)$
- observed data $y = (y_1, \ldots, y_m)$, with $y_i = g_i(\boldsymbol{x})$ many-to-one
- joint density $f_Y(y; \theta) = \int_{A(y)} f_{\boldsymbol{X}}(\boldsymbol{x}; \theta) d\boldsymbol{x}$ $A(y) = \{\boldsymbol{x}; y_i = g_i(\boldsymbol{x}), i = 1, \ldots, m\}$
- algorithm:
  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

Example

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

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

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

- $\hat{y}_i = \mathrm{E}(Y_i \mid x_i; \hat{\theta}^{(k)})$ 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 $\log \mathcal{L}_n(\theta) / \mathcal{L}_n(\hat{\theta}^{(k)})$

... Example

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: $\max_\theta \ell(\theta; \boldsymbol{x})$
- Solve: $\ell'(\hat{\theta}; \boldsymbol{x}) = 0$
- 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:

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

and, for every $\epsilon > 0$,

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

Proof of consistency: Show

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

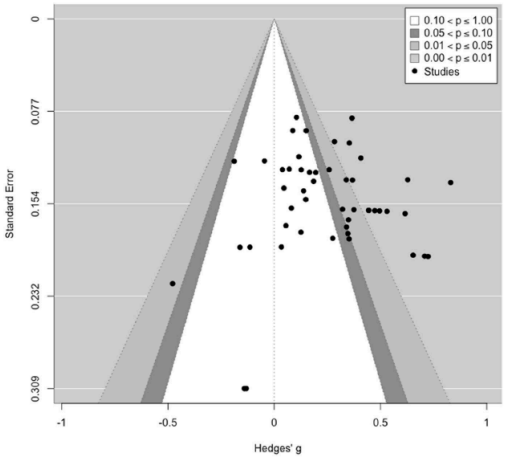
Regularity conditions: asymptotic normality

AoS Thm 9.18; MS Thm 5.4

- (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) $\mathrm{E}_\theta\{\ell'(\theta; X_i)\} = 0 \forall \theta$ and $\mathrm{Cov}\{\ell'(\theta; X_i)\} = I(\theta)$ is positive definite $\forall \theta$
- (B5) $\mathrm{E}_\theta\{-\ell''(\theta; X_i)\} = J(\theta)$ is positive definite $\forall \theta$
- (B6) For each $\theta, \delta > 0, 1 \leq j, k, l, \leq p$,

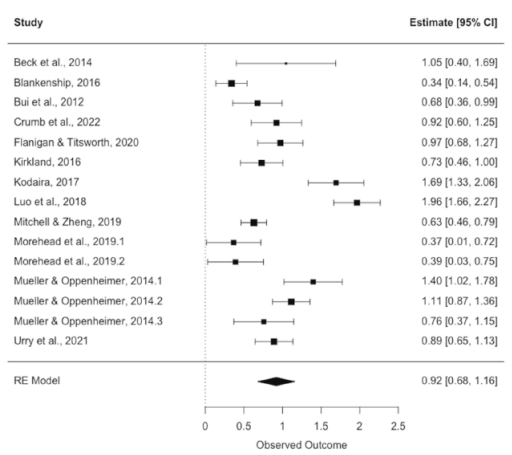
$$\left| \frac{\partial^3 \ell(\theta_*; x_i)}{\partial\theta_j \partial\theta_k \partial\theta_l} \right| \leq M_{jkl}(\theta_*),$$

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



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

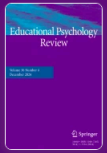
Typed Versus Handwritten Lecture Notes and College Student Achievement: A Meta-Analysis

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 →