Mathematical Statistics II

STA2212H S LEC9101

Week 1

January 7 2025

Could Dark Chocolate Reduce Your Risk of Diabetes?

A new study suggests that it might. We asked experts if that's too good to be true.

Listen to this article · 6:55 min Learn more

Share full article

231

Rosemary Calvert/Getty Images

Today

1. Course Overview
2. Review of Likelihood STA 2112S
3. Properties of maximum likelihood estimators MS Ch. 5.4,5
4. Statistics in the News

Link

STA 2212S: Mathematical Statistics II
Tuesday, 10.00-13.00 January 7 – April 1 2025

Course description:

This course is a continuation of STA2112H. It is designed for graduate students in statistics and biostatistics. Topics include: Likelihood inference, Bayesian methods, Significance testing, Hypothesis testing, Goodness-of-fit, Robust inference, Causality, Classification.

Prerequisite: STA2112H

Course content

The course Quercus page has

- A regularly updated syllabus

STA 2212S: Mathematical Statistics II Link

Syllabus Spring 2025

Week	Date	Methods	References
1	Jan 7	Likelihood inference: review of ML estimation; mis-specified models; computation; nonparametric mle	MS §§5.1–7, SM Ch 4
2	Jan 14	Bayesian estimation; Bayesian inference	MS §5.8; AoS §§ 11.1–4; SM §§11.1,2
3	Jan 21	Optimality in estimation	MS Ch 6; AoS Ch 12; SM §7.1, 11.5.2
4	Jan 28	Interval estimation; Confidence bands	MS §§7.1,2; AoS Ch 7; SM §7.1.4

Link

HW Question Week 1 STA 2212S 2025

Due January 14 *MS, Exercise 5.2*

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 $n \to \infty$.

… Course Overview

Link

Project Guidelines STA 2212S: Mathematical Statistics II 2025

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 April 1, 2025.
Report submission due April 15, 2025.

Part 1: Presentation [10 points]

On the last day of class (April 1), your team will present your final project; presentations will be 10 minutes long. Detailed guidance on the presentation will be provided.

Part 2: Write-up [40 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	$L(\theta; x) = c(x) f(x;\theta)$	$\mathcal{L}(\theta)$
log-likelihood function	$\ell(\theta; x) = \log L(\theta; x) = \log f(x;\theta) + a(x)$	
score function	$u(\theta) = \partial \ell(\theta; x)/\partial\theta$	$\ell'(x;\theta)$
observed information function	$j(\theta) = -\partial^2 \ell(\theta; x)/\partial\theta\partial\theta^T$	$J(\theta) = \mathrm{E}_\theta\{j(\theta)\}$
expected information (in one observation)	$i(\theta) = \mathrm{E}_\theta\{U(\theta)U(\theta)^T\}$[1]	$I(\theta)$ (p.245)

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:

Today: Parametric estimation & inference

Topics covered

- ⋆ Parametric Inference
- ⋆ Method of Moments (MOM) Estimation
- ⋆ Maximum Likelihood Estimation (MLE)
- ⋆ Properties of MLEs

Reading

- ⋆ Recommended: Knight Chp 4.5, 5.1-5.4
- ⋆ Additional: Wasserman Chp 9.1-9.4

Onward: The likelihood function and its log

The Likelihood Function

Let $X_1, \ldots X_n$ be iid with pdf $f(x_i;\theta)$. The **likelihood function** is the joint probability of the observations considered as a function of the parameter,

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

$$L_n(\underset{\sim}{\theta}) \propto \prod f(x_i;\theta)$$

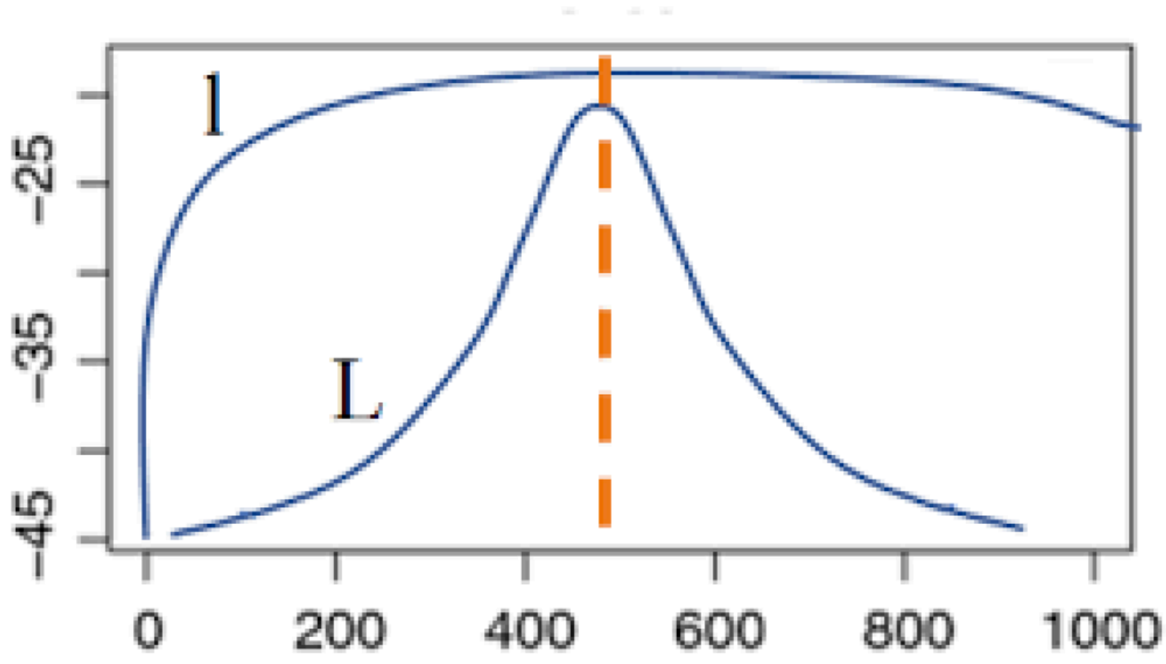
$$= c(\underline{x}) \prod f(x_i;\theta)$$

$$= L_n(\theta; \underline{x}) \propto f(\underset{\sim}{x};\theta)$$

The **log-likelihood function** is

$$\ell_n(\theta) = \log L_n(\theta)$$

Visualizing the likelihood function and its log



Properties of MLEs

Properties

- ⋆ The MLE, $\hat{\theta}_n$ is **consistent** for θ
- ⋆ The MLE for $g(\theta)$ is $g(\hat{\theta}_n)$, that is, the MLE is **equivariant**
- ⋆ The MLE is **asymptotically normal** ←
- ⋆ The MLE is **asymptotically optimal** or **efficient** ←

Example 1: $X_i \sim Geom(\theta), \quad i = 1, \ldots, n$

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

$$L(\theta; \underline{x}) = \theta^n (1-\theta)^{\sum x_i - n}$$

$$\ell(\theta; \underline{x}) = n \ln \theta + (\sum x_i - n) \ln(1-\theta)$$

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

$$\frac{n}{\theta} - \frac{\sum x_i - n}{1-\theta} = 0$$

$$\therefore \hat{\theta} = \frac{1}{\bar{x}}$$

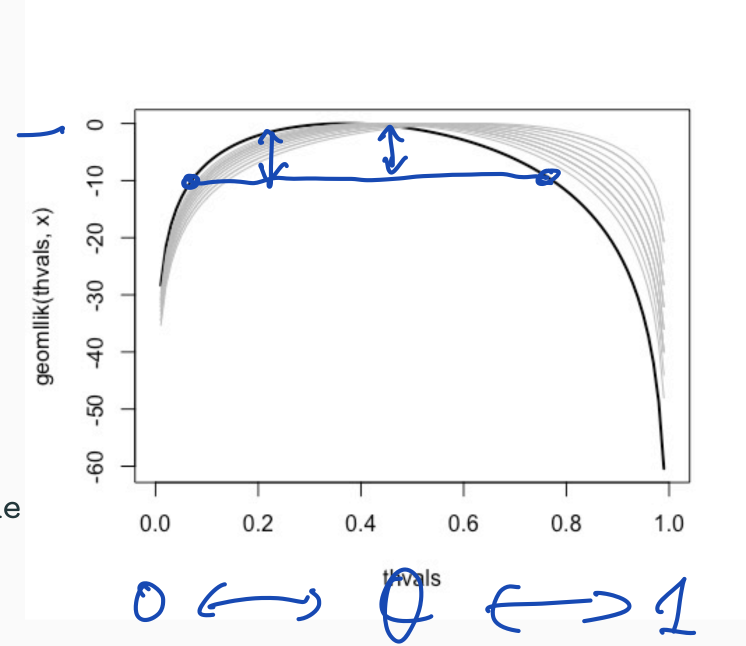
ntbc

Example 2: $X_i \sim LocExp(\theta), \quad i = 1, \ldots, n$

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

Simulated Example

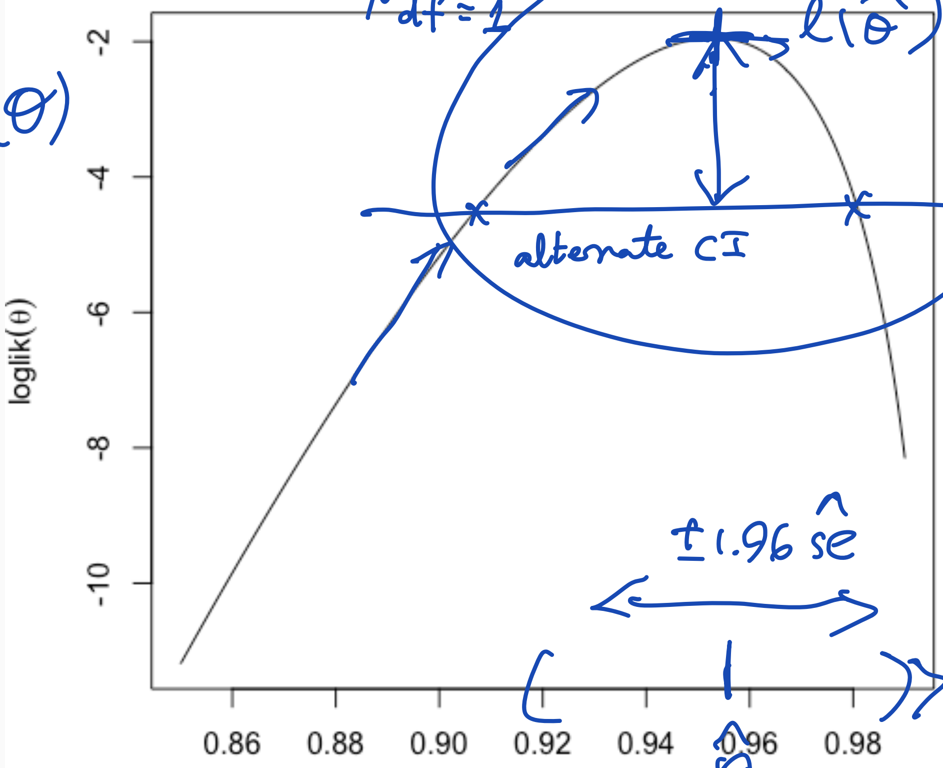
```
geomlik <- function(theta,x){
  theta^length(x)*(1-theta)^(sum(x)-length(x))}

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

$2\{\ell(\hat\theta) - \ell(\theta)\} = w(\theta)$ LRT $\xrightarrow{d} \chi^2_{df=1}$ as $n \to \infty$

$\ell(\theta)$



$\widehat{nI(\theta)}$

$\hat\theta = \arg\sup_\theta \ell_n(\theta; \underset{\sim}{x}) = \hat\theta_n(\underset{\sim}{x})$

$\ell_n'(\hat\theta) = 0$ (usually)

score fn $\ell_n'(\theta) = \ell_n'(\theta; \underset{\sim}{x})$

$U(\theta) = u(\theta; \underset{\sim}{X})$

$-\ell_n''(\hat\theta) = j(\hat\theta)$ obs'd info.

$a.var(\hat\theta) = [I_n(\theta)]^{-1}$

$(I_n(\theta) = nI(\theta)$ if iid $) \doteq [J_n(\theta)]^{-1}$

$X \in \mathbb{R}$

$$E_\theta\, \ell'(\theta; X) = 0$$ — 1st Bartlett identity (last term)

$\ell'(\hat\theta; x) = 0$ — mle

consistency: $\hat\theta_n \xrightarrow{P} \theta$, $n \to \infty$

$$\ell(\theta; X) = \log f(X; \theta)$$

$$E_\theta\{\ell'(\theta; X)\} = E_\theta\left\{ \frac{\frac{\partial}{\partial\theta} f(X;\theta)}{f(X;\theta)} \right\}$$

$$= \int \frac{\frac{\partial}{\partial\theta} f(x;\theta)}{f(x;\theta)} \cdot f(x;\theta)\, dx$$

only if no θ in limits of $\int$?

$$= \frac{\partial}{\partial\theta} \int f(x;\theta)\, dx = \frac{\partial}{\partial\theta} 1 = 0$$

$$0 = \frac{\partial}{\partial\theta} \int f(x;\theta)\, dx$$

$\theta \in \mathbb{R}$

$$= \frac{\partial^2}{\partial\theta^2} \int f(x;\theta)\, dx = \frac{\partial}{\partial\theta} \int \ell'(\theta; x) f(x;\theta)\, dx$$

$$= \int \ell''(\theta; x) f(x;\theta)\, dx + \int \ell'(\theta;x) \frac{\partial}{\partial\theta} f(x;\theta)\, dx$$

$$= \int \ell''(\theta; x) f(x;\theta)\, dx + \int \{\ell'(\theta;x)\}^2 f(x;\theta)\, dx$$

$$0 = E_\theta\{\ell''(\theta; X)\} + \mathrm{var}_\theta\{\ell'(\theta; X)\}$$

$$= -I(\theta) + I(\theta)$$

- model $X \sim f(x;\theta), \theta \in \mathbb{R}^p$ θ is a column vector $\in \mathbb{R}^p$ $X \in \mathbb{R}^n$
- $L(\theta; x) = c(x) f(x;\theta)$ map from $\mathbb{R}^p \to \mathbb{R}$
- $\ell'(\theta; x) = \left[\frac{\partial \ell(\theta; x)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\theta; x)}{\partial \theta_p} \right]^T$ $p \times 1$ vector
- $-\ell''(\theta; x)$ $p \times p$ matrix

$$= \begin{bmatrix} -\frac{\partial^2 \ell(\theta; x)}{\partial \theta_1^2} & \frac{\partial^2 \ell(\theta; x)}{\partial \theta_1 \partial \theta_2} & \cdots \\ & \ddots & \\ & & \frac{\partial^2 \ell(\theta; x)}{\partial \theta_p^2} \end{bmatrix}$$

$$\text{avar}\, \hat{\theta} = \left[-\ell''(\hat{\theta}; x) \right]^{-1}$$

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

A.S. "

- $L(\theta; x)$ — map from $\mathbb{R}^p \to \mathbb{R}$

$\hat{\theta} \pm 1.96\,\hat{se}$ — 95% C.I. for θ "

- $\ell'(\theta; x)$ — $p \times 1$ vector

$\hat{se}^2 = [-\ell''(\hat{\theta}; x)]^{-1}$

- $-\ell''(\theta; x)$ — $p \times p$ matrix

$I_n(\theta) = E_\theta\{J_n(\theta)\} = E_\theta[-\ell''(\theta; X)]$

$I = E(J)$

$\widehat{I_n(\theta)} = [-\ell''(\hat{\theta}; x)]$ estimate of inv. a.var. $(\hat{\theta})$

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|)
                -0.079918
(Intercept) -34.103704   6.530014  -5.223 1.76e-07 ***
zn           -0.059389   0.033731  -2.369  0.01782 *
indus         0.785327   0.043722  -1.358  0.17436
chas         48.523782   0.728930   1.077  0.28132
nox          -0.425596   7.396497   6.560 5.37e-11 ***
rm                       0.701104  -0.607  0.54383
age           0.022      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
```

$$\theta = (\beta_1, \beta_2)$$

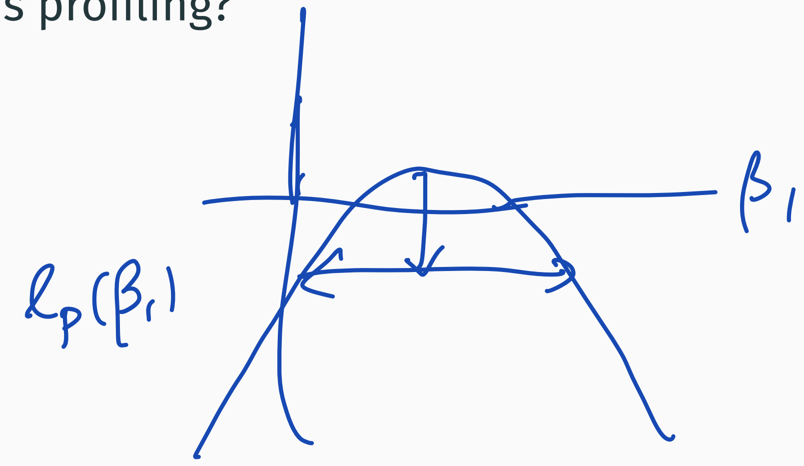
$$\ell(\theta; \underset{\sim}{x}) = \ell(\beta_1, \beta_2; \underset{\sim}{x})$$

$$\max_{\beta_1} \ell(\beta_1, \beta_2; \underset{\sim}{x}) \rightarrow \hat{\beta}_1(\beta_2)$$

$$\left.\frac{\partial \ell}{\partial \beta_1}(\beta_1, \beta_2; x)\right|_{\beta_1 = \hat{\beta}_1} = 0$$

$$\ell_p(\beta_2; \underset{\sim}{x}) = \ell(\hat{\beta}_1(\beta_2), \beta_2; x)$$

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

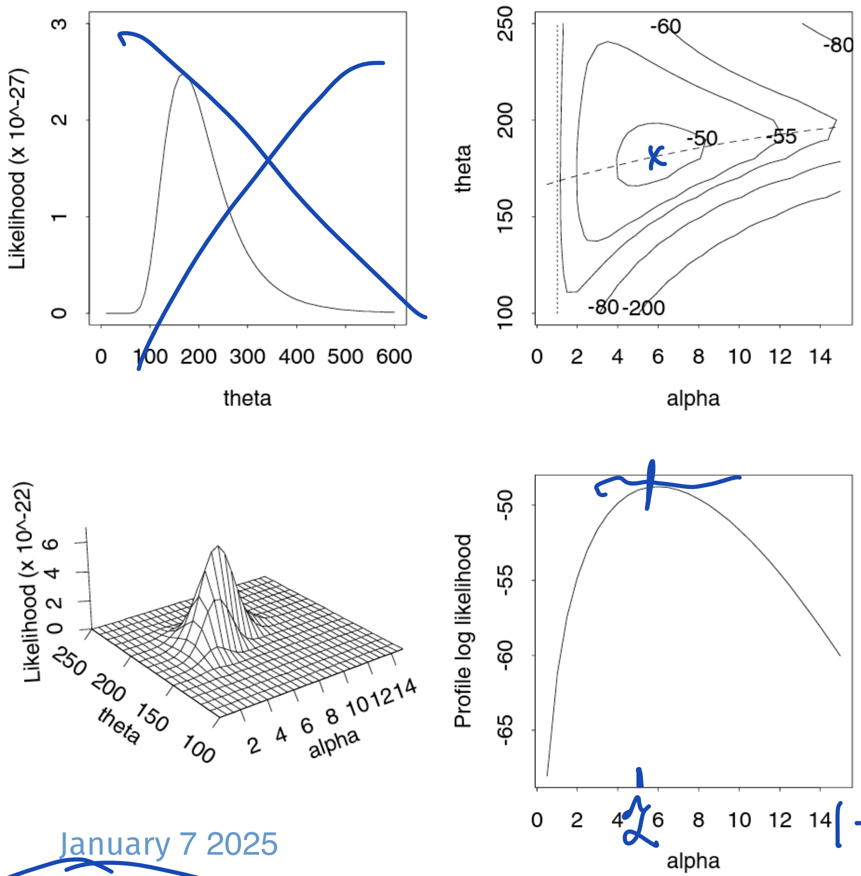


$$2\{\ell_p(\hat{\beta}_1) - \ell_p(\beta_1)\} \leq 1.82$$

↑ χ^2_1 cutoff for 95%

Profile likelihood function

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.



dotted line is a curve of $\theta = \hat{\theta}(\alpha)$

$\max_{\theta} \ell(\theta, \alpha; \underset{\sim}{x})$

$\ell_p(\alpha) = \ell(\hat{\theta}(\alpha), \alpha; \underset{\sim}{x})$

$f(x; \theta, \alpha) = ?$ ntb

$= \alpha \theta x^{\alpha-1} e^{-\theta x^{\alpha}}$

$$l_p(\alpha; \underline{x}) = l(\hat{\theta}_\alpha(\alpha), \alpha; \underline{x})$$ ntbc (exerc.)

$$\frac{\partial}{\partial \alpha} l_p(\alpha) \Big|_{\alpha=\hat{\alpha}} = 0 \qquad (\hat{\alpha}, \hat{\theta}(\hat{\alpha})) \text{ is MLE}$$

$$= (\hat{\alpha}, \hat{\theta})$$

$$\begin{cases} f_1(y_1, y_2) = c_1 \\ f_2(y_1, y_2) = c_2 \end{cases}$$

Wald C.I. is $\hat{\theta} \pm 1.96\, \hat{se}$

LR CI is $\{\theta ;\ w(\theta) \equiv 2\{l(\hat{\theta}) - l(\theta)\} \geq 1.92\}$

$\chi^2_1(0.95)$ quantile $= 3.84$

$$\hat{\theta} \;\dot{\sim}\; N(\theta, j^{-1}(\hat{\theta}))$$

$$w(\theta) \;\dot{\sim}\; \chi^2_1$$

approx's that determine CI

Properties of maximum likelihood estimators

- maximum likelihood estimators are equivariant — example

- maximum likelihood estimators are biased — special exceptions

$E\hat{\theta} \neq \theta$

$E g(\hat{\theta}) \neq g(\theta)$ unless g is linear

~~$\varphi(\theta)$~~ $\varphi = g(\theta)$ $\hat{\varphi} = g(\hat{\theta})$

e.g. $\log(\theta)$

- maximum likelihood estimators have no explicit formula — in general

…Properties of maximum likelihood estimators

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

…Properties of maximum likelihood estimators

model

$f_\theta(x) = f(x;\theta)$

- maximum likelihood estimators minimize the KL-divergence to the data
- KL divergence from f_0 true to f_θ model :

$$KL(f_\theta; f_0) \equiv \mathrm{E}_{f_0} \log \left\{ \frac{f_0(X)}{f_\theta(X)} \right\} = -\mathrm{E}_{f_0} \log\{f(X;\theta)\} + \mathrm{E}_{f_0} \log f_0(X)$$

- estimate of $\mathrm{E}_{f_0} \log\{f(X;\theta)\}$?

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

- minimize $KL(f_\theta; f_0)$ same as maximize $\ell(\theta; x_1, \ldots, x_n)$

Asymptotic properties of maximum likelihood estimators

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

p.256

(i) $\hat{\theta} \xrightarrow{p} \theta$ as $n \to \infty$

$X_1, \dots, X_n$ iid $f(x; \theta_0)$

$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, I_1^{-1}(\theta_0))$

under "true model"

$\hat{\theta}_n \xrightarrow{d} N(\theta_0, I_n^{-1}(\theta_0))$ as $n \to \infty$

$x_1 \dots x_n$

$X \sim f(x; \theta_0)$

$E_{\theta_0} \ell_n'(\theta_0; X) = 0 \leftarrow \ell_n'(\hat{\theta}_n; x) = 0$

$\sum_{i=1}^{n} \frac{\partial \log f(x_i; \theta)}{\partial \theta} \Big|_{\hat{\theta}}$

$\frac{1}{n} \sum (\cdot) \xrightarrow{p} 0$

$\Rightarrow \hat{\theta} \xrightarrow{p} \theta$

$\theta \in \mathbb{R}$

$$\ell_n'(\hat{\theta}_n; \underset{\sim}{x}) = 0$$

$$= \ell_n'(\theta_0; \underset{\sim}{x}) + \underbrace{(\hat{\theta}_n - \theta_0)}\, \ell_n''(\theta_0; \underset{\sim}{x}) + \text{Rem}_n$$

(drop ↑)

$$(\hat{\theta}_n - \theta_0)\left[-\ell_n''(\theta_0; \underset{\sim}{x})\right] \doteq \underbrace{\ell_n'(\theta_0; \underset{\sim}{x})}$$

we want its distⁿ : $\underset{\sim}{x} \longrightarrow \underset{\sim}{X} = (X_1, \dots, X_n)$ r.v.'s

$$\ell_n'(\theta_0; \underset{\sim}{X}) = \sum_{i=1}^{n} \ell'(\theta_0; X_i)$$ iid r.v.'s

$$\frac{1}{\sqrt{n}} \sum \ell'(\theta; X_i) \xrightarrow{d} N(0, \sigma^2)$$

$$\sigma^2 = \operatorname{var}_{\theta_0}\left[\ell'(\theta_0; X_i)\right]$$

$$\hookrightarrow \sqrt{n}(\hat{\theta}_n - \theta_0)\{I(\theta_0)\}^{1/2}$$

$$\doteq \frac{\ell_n'(\theta_0;\underset{\sim}{X})}{-\ell_n''(\theta_0;\underset{\sim}{X})} \cdot \{I(\theta_0)\}^{1/2}\sqrt{n}$$

$$\doteq \frac{\frac{1}{\sqrt{n}}\ell_n'(\theta_0;\underset{\sim}{X})}{I(\theta_0)^{1/2}} \cdot \frac{I(\theta_0)\,n}{-\ell_n''(\theta_0;\underset{\sim}{X})}.$$

$$= \Big(\qquad \Big) \cdot \frac{I(\theta_0)}{-\frac{1}{n}\ell_n''(\theta_0;\underset{\sim}{X})}$$

$$\downarrow P$$

$$1 \quad \text{WLLN}$$

CLT

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

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

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

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

where $D(\theta) =$

… Your friend the delta-method

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

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

See also AoS §9.9

Example

$$X_1, \ldots, X_n \text{ i.i.d. Gamma } (\alpha, \lambda)$$

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

… Example

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

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

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

MS p.270; note change in notation

- 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 $y = (y_1, \dots, y_m)$, with $y_i = g_i(\boldsymbol{x})$ many-to-one
- joint density $f_Y(y;\theta) = \int_{A(y)} f_{\boldsymbol{X}}(\boldsymbol{x};\theta) d\boldsymbol{x}$ $A(y) = \{\boldsymbol{x}; y_i = g_i(\boldsymbol{x}), i = 1, \dots, m\}$
- 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: $(x_1, y_1), \ldots, (x_n, y_n)$; $y_i \sim \mathit{Bernoulli}(\theta)$
- 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

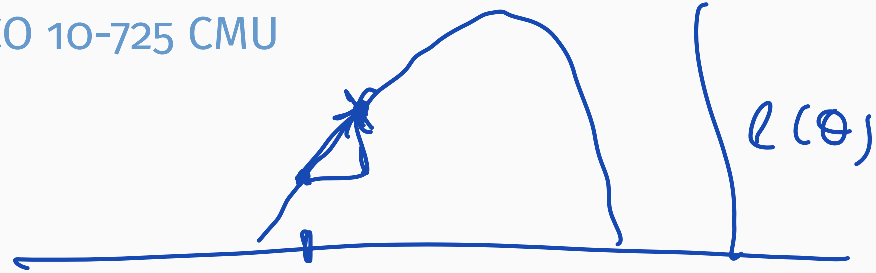
General-purpose optimization based on Nelder–Mead, quasi-Newton and conjugate-gradient algorithms. It includes an option for box-constrained optimization and simulated annealing.

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: $j(\hat\theta^{(t)})(\hat\theta^{(t+1)} - \hat\theta^{(t)}) = \ell'(\hat\theta^{(t)})$
- 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 $\hat\theta^{(t+1)} = \hat\theta^{(t)} + \epsilon_t\{j(\hat\theta^{(t)})\}^{-1}\ell'(\hat\theta^{(t)})$

$$B\Delta\theta = -\nabla\ell(\theta)$$

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

Regularity conditions

- (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

Regularity conditions

- (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) $\mathrm{E}_\theta\{\ell'(\theta;X_i)\} = 0 \quad \forall\,\theta$ 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 $\boldsymbol{\theta}, \delta > 0, 1 \leq j, k, l, \leq p$,

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

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

Regularity conditions

- (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) $\mathrm{E}_\theta\{\ell'(\theta; X_i)\} = 0 \quad \forall\,\theta$ 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 $\boldsymbol{\theta}, \delta > 0, 1 \leq j, k, l, \leq p$,

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

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

Regularity conditions

MS Thm 5.4

Misspecified models

- model assumption $X_1, \dots, X_n$ i.i.d. $f(x;\theta), \theta \in \Theta$
- true distribution $X_1, \dots, X_n$ i.i.d. $F(x)$ notation
- maximum likelihood estimator based on model:

$$\sum_{i=1}^{n} \ell'(\hat{\theta}_n; X_i) = 0$$

- what is $\hat{\theta}_n$ estimating ?

Misspecified models

- model assumption $X_1, \ldots, X_n$ i.i.d. $f(x;\theta), \theta \in \Theta$
- true distribution $X_1, \ldots, X_n$ i.i.d. $F(x)$ — notation
- maximum likelihood estimator based on model:

$$\sum_{i=1}^{n} \ell'(\hat{\theta}_n; X_i) = 0$$

- what is $\hat{\theta}_n$ estimating ?
- define the parameter $\theta(F)$ by

$$\int_{-\infty}^{\infty} \ell'\{x;\theta(F)\} dF(x) = 0$$

- $$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N(0, \sigma^2)$$
- $$\sigma^2 = \frac{\int [\ell'\{x;\theta(F)\}]^2 dF(x)}{(\int [\ell''\{x;\theta(F)\}]^2 dF(x))^2}$$

Misspecified models

- $$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N(0, \sigma^2)$$
- $$\sigma^2 = \frac{\int [\ell'\{x; \theta(F)\}]^2 dF(x)}{(\int [\ell''\{x; \theta(F)\}]^2 dF(x))^2}$$
- more generally, for $\theta \in \mathbb{R}^p$,
$$\sqrt{n}\{\hat{\theta}_n - \theta(F)\} \xrightarrow{d} N_p\{0, G^{-1}(F)\}$$
- $$G(F) = J(F) I^{-1}(F) J(F),$$
- $$J(F) = \int -\ell''\{\theta(F); x_i\} dF(x_i), \quad I(F) = \int \{\ell'(\theta(F); x_i)\}\{\ell'(\theta(F); x_i)\}^T dF(x_i)$$

Godambe information

sandwich variance

Statistics in the News

Could Dark Chocolate Reduce Your Risk of Diabetes?

A new study suggests that it might. We asked experts if that's too good to be true.

Listen to this article · 6:55 min Learn more

Share full article

231

Rosemary Calvert/Getty Images

RESEARCH

OPEN ACCESS

Check for updates

Chocolate intake and risk of type 2 diabetes: prospective cohort studies

Binkai Liu,[1] Geng Zong,[2,3] Lu Zhu,[1] Yang Hu,[1] JoAnn E Manson,[4,5,6] Molin Wang,[4,5,7] Eric B Rimm,[1,4,5] Frank B Hu,[1,4,5] Qi Sun[1,4,5]

[1]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[2]Key Laboratory of Nutrition, Metabolism and Food Safety, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China

[3]Department of Endocrinology and Metabolism, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China

[4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[5]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

[6]Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA

ABSTRACT

OBJECTIVE
To prospectively investigate the associations between dark, milk, and total chocolate consumption and risk of type 2 diabetes (T2D) in three US cohorts.

DESIGN
Prospective cohort studies.

SETTING
Nurses' Health Study (NHS; 1986-2018), Nurses' Health Study II (NHSII; 1991-2021), and Health Professionals Follow-Up Study (HPFS; 1986-2020).

PARTICIPANTS
At study baseline for total chocolate analyses (1986 for NHS and HPFS; 1991 for NHSII), 192 208 participants without T2D, cardiovascular disease, or cancer were included. 111 654 participants were included in the analysis for risk of T2D by intake of chocolate subtypes, assessed from 2006 in NHS and HPFS and from 2007 in NHSII.

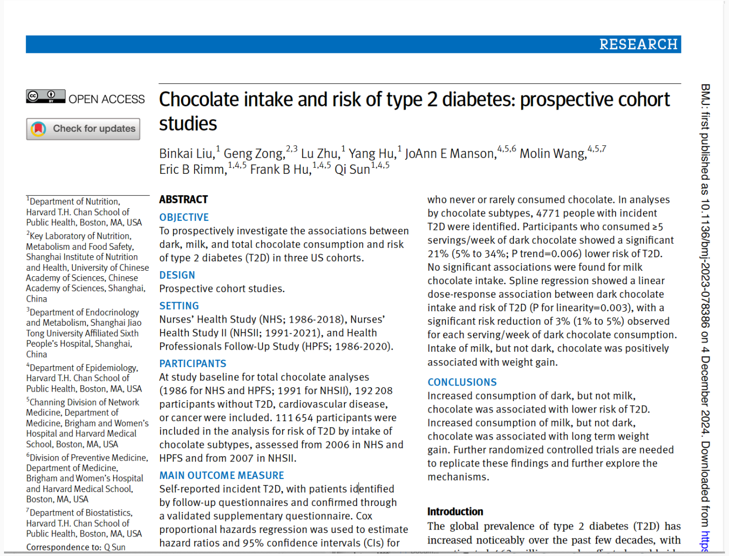
MAIN OUTCOME MEASURE
Self-reported incident T2D, with patients identified by follow-up questionnaires and confirmed through

who never or rarely consumed chocolate. In analyses by chocolate subtypes, 4771 people with incident T2D were identified. Participants who consumed ≥5 servings/week of dark chocolate showed a significant 21% (5% to 34%; P trend=0.006) lower risk of T2D. No significant associations were found for milk chocolate intake. Spline regression showed a linear dose-response association between dark chocolate intake and risk of T2D (P for linearity=0.003), with a significant risk reduction of 3% (1% to 5%) observed for each serving/week of dark chocolate consumption. Intake of milk, but not dark, chocolate was positively associated with weight gain.

CONCLUSIONS
Increased consumption of dark, but not milk, chocolate was associated with lower risk of T2D. Increased consumption of milk, but not dark, chocolate was associated with long term weight gain. Further randomized controlled trials are needed to replicate these findings and further explore the mechanisms.

BMJ: first published as 10.1136/bmj-2023-078386 on 4 December 2024. Downloaded f

… Original study



RESEARCH

OPEN ACCESS

Check for updates

Chocolate intake and risk of type 2 diabetes: prospective cohort studies

Binkai Liu,[1] Geng Zong,[2,3] Lu Zhu,[1] Yang Hu,[1] JoAnn E Manson,[4,5,6] Molin Wang,[4,5,7] Eric B Rimm,[1,4,5] Frank B Hu,[1,4,5] Qi Sun[1,4,5]

[1]Department of Nutrition, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[2]Key Laboratory of Nutrition, Metabolism and Food Safety, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, China
[3]Department of Endocrinology and Metabolism, Shanghai Jiao Tong University Affiliated Sixth People's Hospital, Shanghai, China
[4]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[5]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[6]Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
Correspondence to: Q Sun

ABSTRACT

OBJECTIVE
To prospectively investigate the associations between dark, milk, and total chocolate consumption and risk of type 2 diabetes (T2D) in three US cohorts.

DESIGN
Prospective cohort studies.

SETTING
Nurses' Health Study (NHS; 1986-2018), Nurses' Health Study II (NHSII; 1991-2021), and Health Professionals Follow-Up Study (HPFS; 1986-2020).

PARTICIPANTS
At study baseline for total chocolate analyses (1986 for NHS and HPFS; 1991 for NHSII), 192 208 participants without T2D, cardiovascular disease, or cancer were included. 111 654 participants were included in the analysis for risk of T2D by intake of chocolate subtypes, assessed from 2006 in NHS and HPFS and from 2007 in NHSII.

MAIN OUTCOME MEASURE
Self-reported incident T2D, with patients identified by follow-up questionnaires and confirmed through a validated supplementary questionnaire. Cox proportional hazards regression was used to estimate hazard ratios and 95% confidence intervals (CIs) for who never or rarely consumed chocolate. In analyses by chocolate subtypes, 4771 people with incident T2D were identified. Participants who consumed ≥5 servings/week of dark chocolate showed a significant 21% (5% to 34%; P trend=0.006) lower risk of T2D. No significant associations were found for milk chocolate intake. Spline regression showed a linear dose-response association between dark chocolate intake and risk of T2D (P for linearity=0.003), with a significant risk reduction of 3% (1% to 5%) observed for each serving/week of dark chocolate consumption. Intake of milk, but not dark, chocolate was positively associated with weight gain.

CONCLUSIONS
Increased consumption of dark, but not milk, chocolate was associated with lower risk of T2D. Increased consumption of milk, but not dark, chocolate was associated with long term weight gain. Further randomized controlled trials are needed to replicate these findings and further explore the mechanisms.

Introduction

The global prevalence of type 2 diabetes (T2D) has increased noticeably over the past few decades, with

BMJ: first published as 10.1136/bmj-2023-078386 on 4 December 2024. Downloaded from https

Results: After adjusting for personal, lifestyle, and dietary risk factors, participants consuming ≥ 5 servings/week of any chocolate showed a significant 10% (95% CI 2% to 17%; P trend=0.07) lower rate of T2D compared with those who never or rarely consumed chocolate