Statistical Methods III: Spring 2013

Jonathan Wand

Stanford University

Choice models: mathematical theory

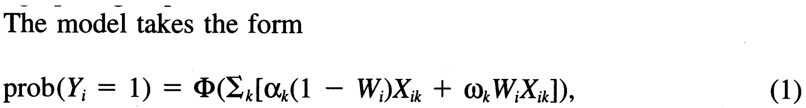
Outline

1. Bartels (1996)
2. Preliminary facts and concepts
   - Follow/ play with /quiz in DB:lecture02/distribution.R
   - Derivation of logistic distribution
3. Two approaches in deriving a theory of choice
   - Theory of comparative judgement: basics
   - Theory of comparative judgement: identification
   - Axioms of choice
4. Upcoming topics

Bartels (1996)

- What is the question? What is at stake?
- What specific choices does he study?
- What types of things does he use to explain choices?
- Some strengths of the study?
- Some weaknesses? How would they potentially change conclusion?

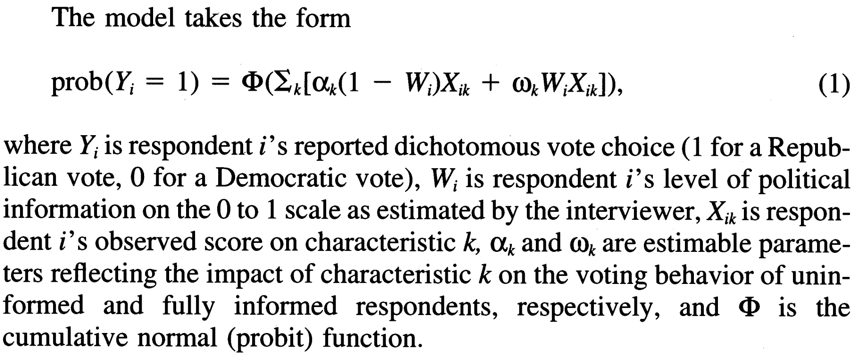
Bartels (1996)



The model takes the form

$$\text{prob}(Y_i = 1) = \Phi(\Sigma_k[\alpha_k(1 - W_i)X_{ik} + \omega_k W_i X_{ik}]), \qquad (1)$$

Bartels (1996)



The model takes the form

$$\text{prob}(Y_i = 1) = \Phi(\Sigma_k[\alpha_k(1 - W_i)X_{ik} + \omega_k W_i X_{ik}]), \qquad (1)$$

where Y_i is respondent i's reported dichotomous vote choice (1 for a Republican vote, 0 for a Democratic vote), W_i is respondent i's level of political information on the 0 to 1 scale as estimated by the interviewer, X_{ik} is respondent i's observed score on characteristic k, α_k and ω_k are estimable parameters reflecting the impact of characteristic k on the voting behavior of uninformed and fully informed respondents, respectively, and Φ is the cumulative normal (probit) function.

Bartels (1996)

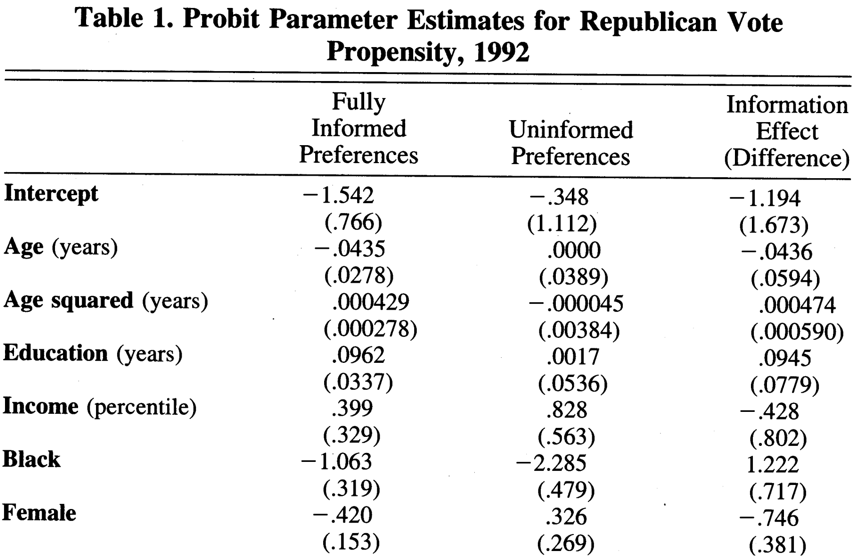


Table 1. Probit Parameter Estimates for Republican Vote Propensity, 1992

	Fully Informed Preferences	Uninformed Preferences	Information Effect (Difference)
Intercept	−1.542	−.348	−1.194
	(.766)	(1.112)	(1.673)
Age (years)	−.0435	.0000	−.0436
	(.0278)	(.0389)	(.0594)
Age squared (years)	.000429	−.000045	.000474
	(.000278)	(.00384)	(.000590)
Education (years)	.0962	.0017	.0945
	(.0337)	(.0536)	(.0779)
Income (percentile)	.399	.828	−.428
	(.329)	(.563)	(.802)
Black	−1.063	−2.285	1.222
	(.319)	(.479)	(.717)
Female	−.420	.326	−.746
	(.153)	(.269)	(.381)

Distributions: notation

Standard Normal(0,1) pdf and cdf as, respectively,

$$\phi(\epsilon) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\epsilon^2}{2}\right\}$$

$$\Phi(\epsilon) = \int_{-\infty}^{\epsilon} \phi(z)\partial z$$

Notes:

- CDF has no closed form expression

Distributions: notation

Type I extreme value (EV I; Gumbel) pdf and cdf as, respectively,

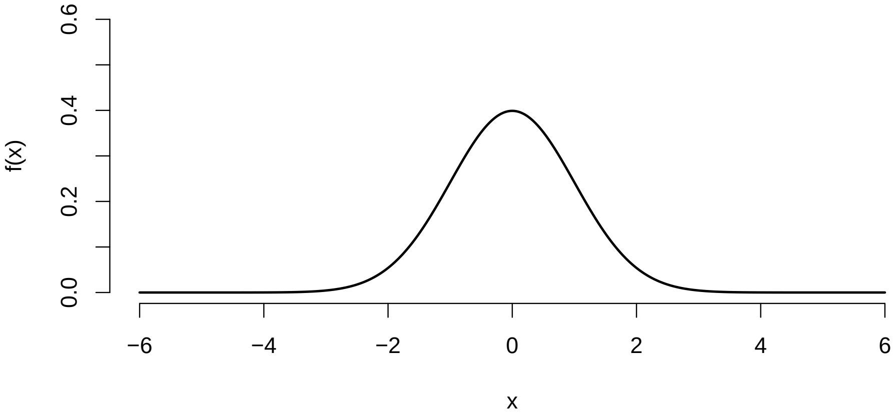
$$\lambda(\epsilon) = e^{-\epsilon} \exp\{-e^{-\epsilon}\}$$

$$\Lambda(\epsilon) = \exp\{-e^{-\epsilon}\}$$

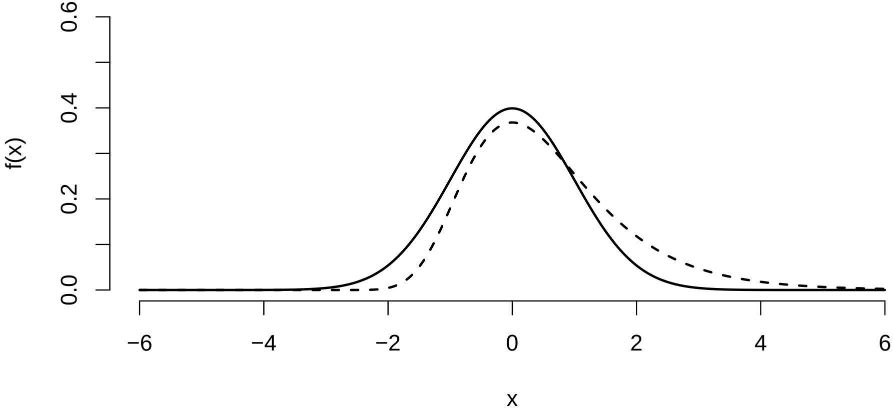
Notes:

- CDF has closed form expression
- standard presentation makes no reference to parameters like the standard normal, they are implicit
- there exists a generalized gumbel,
  and other variations on extreme value distributions

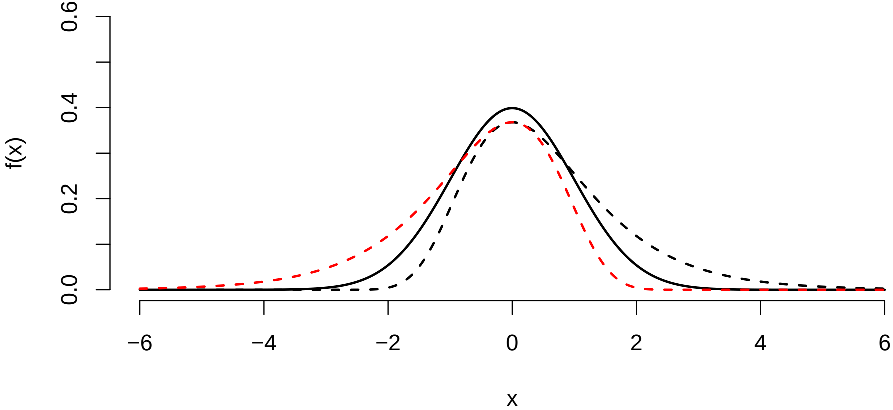
A standard normal densty, $\phi(x)$



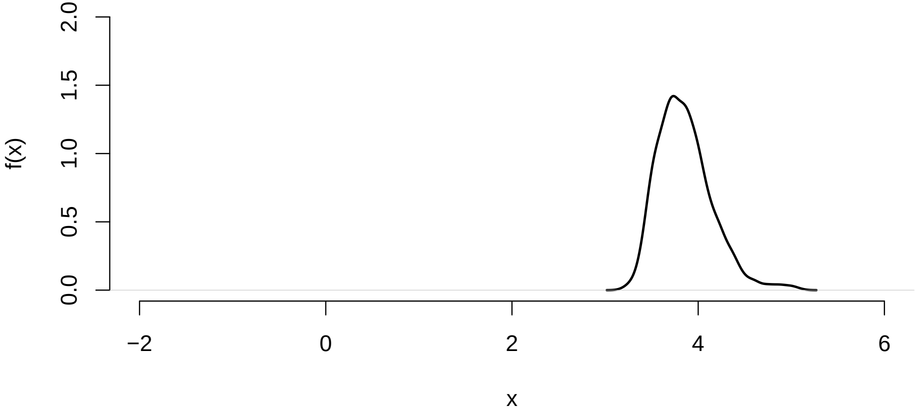
Overlay a gumbel, $\lambda(x)$



Overlay of gumbel $\lambda(-x)$

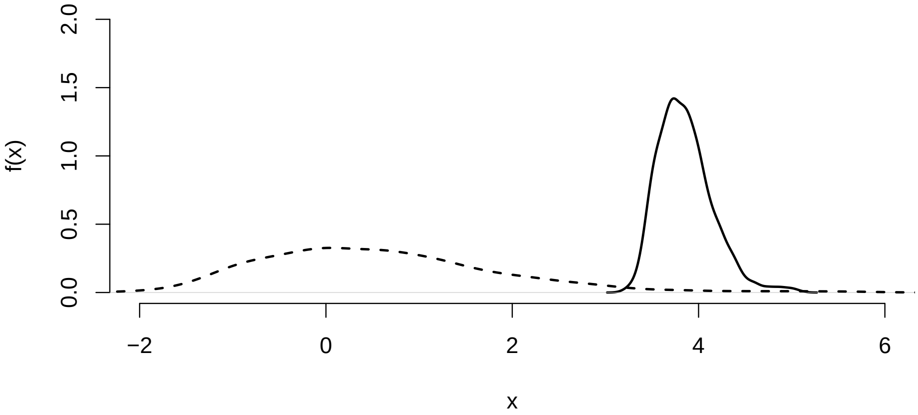


Density of the maximum value from standard normal

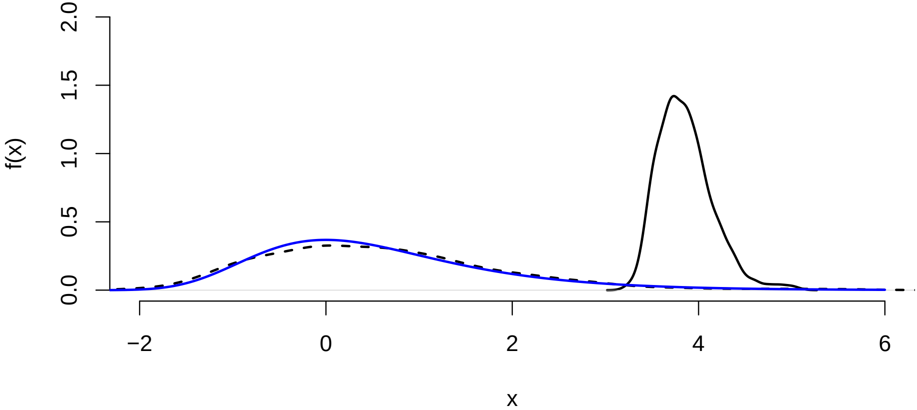


Get *n* draw from $\phi(x)$, keep biggest; repeat *m* times; plot density

Same distribution of maxima, rescaled

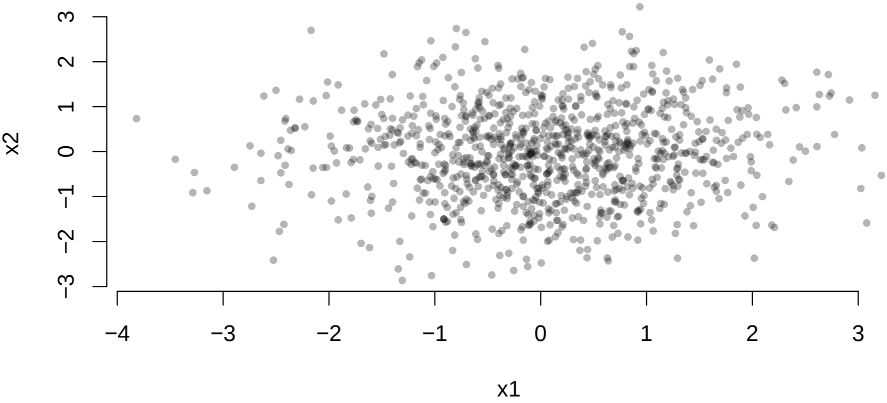


Overlay gumbel

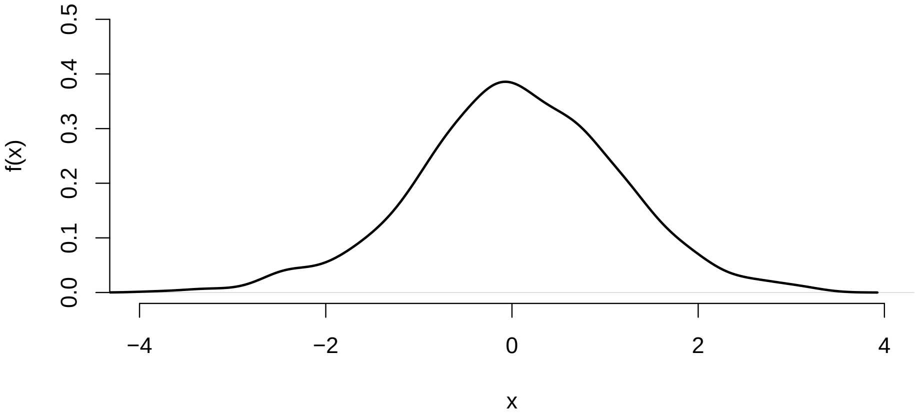


EVD come from maxima with large *n*; $\lambda(x)$ is one of three classes.

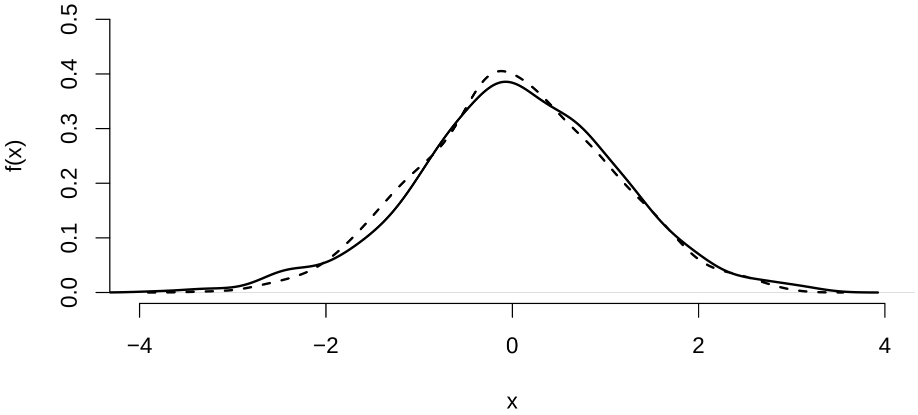
x1,x2 each drawn independently from $\phi(x)$



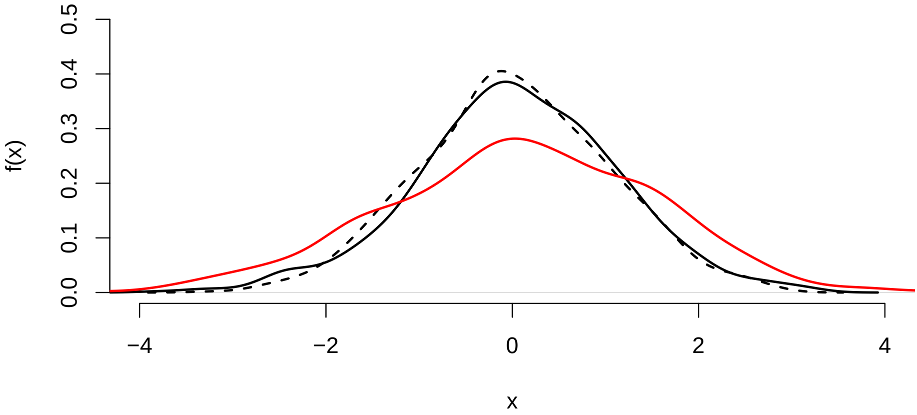
density of x1



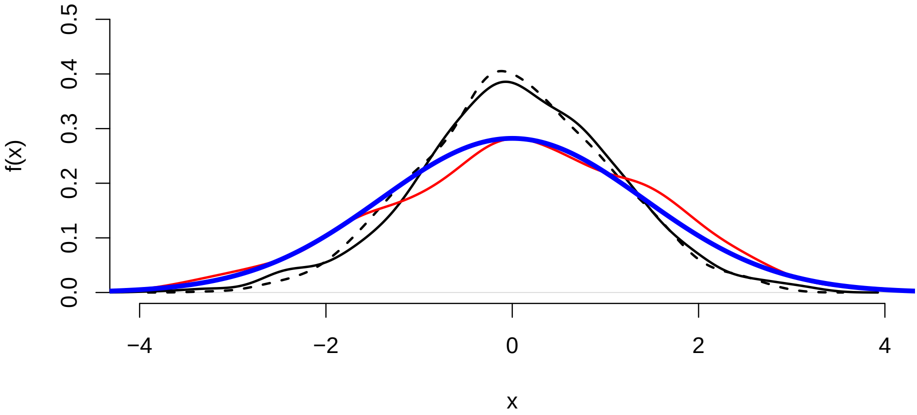
overlay density of -x2



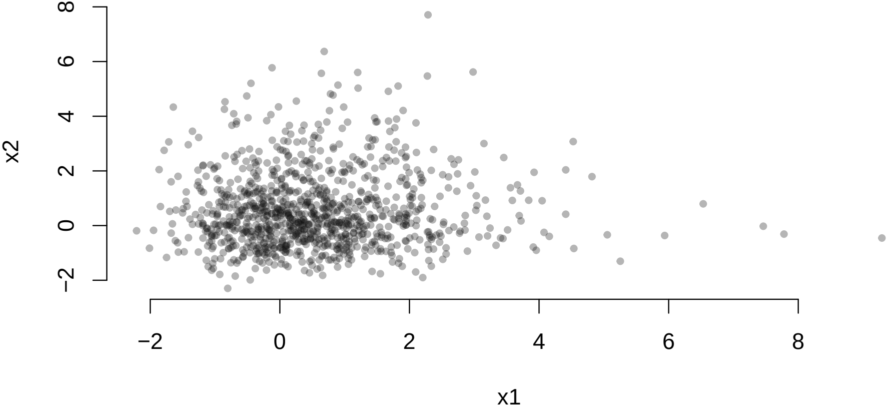
overlay density of x1-x2



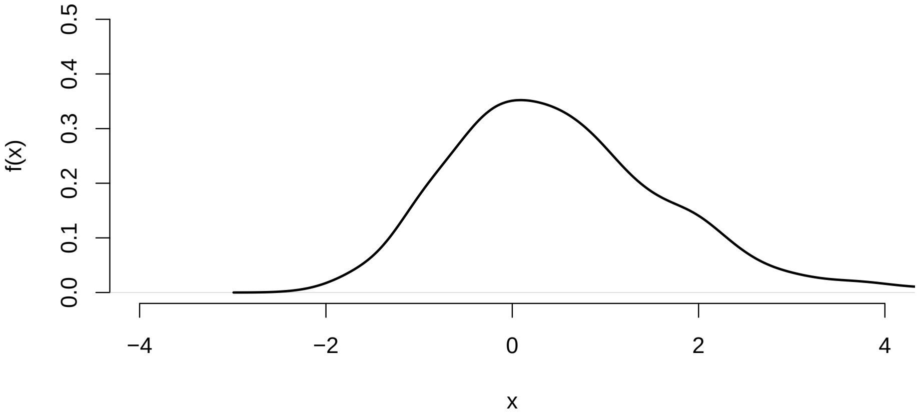
x1-x2 is normally distributed



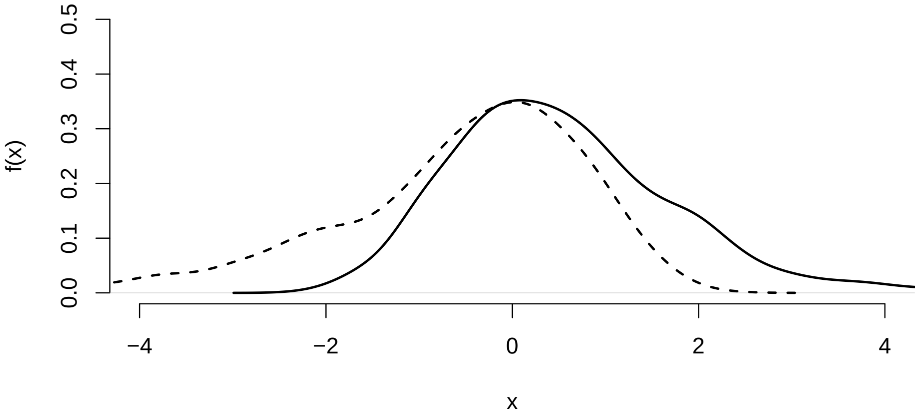
x1,x2 each drawn independently from $\lambda(x)$



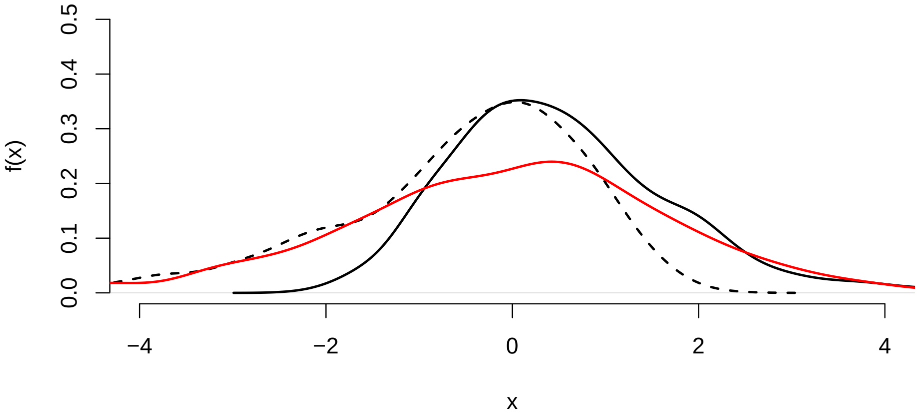
f(x1)



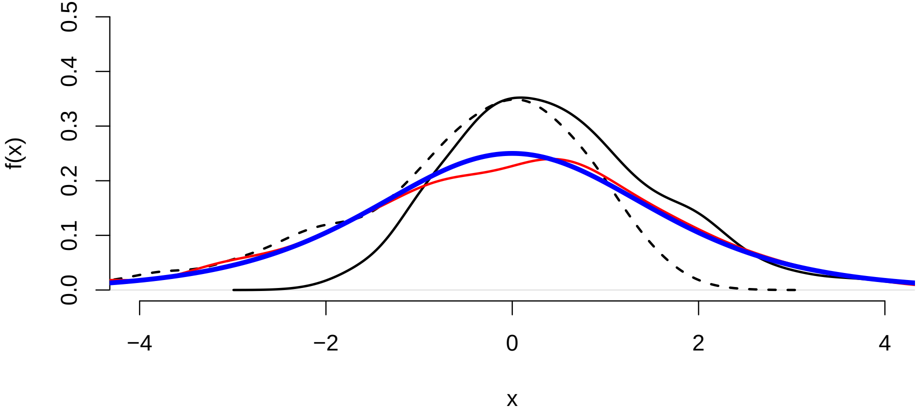
f(-x2) overlay



f(x1-x2)



logistic distribution



Derivation of logistic distribution

Theorem

If X_1 and X_2 are i.i.d. Gumbel RV, then $X_1 - X_2$ has cdf

$$F_{x_1 - x_2}(z) = \frac{1}{1 + e^{-z}} \qquad \textit{(logistic cdf)}$$

Derivation of logistic distribution

Let X be i.i.d. Gumbel RV

- then cdf is,

$$\begin{aligned}\Lambda(z) &= \int_{-\infty}^{z} \lambda(x)dx \\ &= \int_{-\infty}^{z} e^{-x}\exp\{-e^{-x}\}dx \\ &= \exp\{-e^{-z}\}\end{aligned}$$

- and keep in mind, equivalent statements

$$\Lambda(z) = F_X(z) = P(x < z)$$

Derivation of logistic distribution

$$\begin{aligned}
F_{x_1-x_2}(z) &= P(x_1 - x_2 < z) \\
&= P(x_2 < z + x_1) \\
&= \int_{-\infty}^{\infty} \lambda(x_1) \int_{-\infty}^{z+x_1} \lambda(x_2) dx_2 dx_1 \\
&= \int_{-\infty}^{\infty} \lambda(x_1) \Lambda(z + x_1) dx_1 \\
&= \int_{-\infty}^{\infty} e^{-x_1} \exp\{-e^{-x_1}\} \exp\{-e^{-(z+x_1)}\} dx_1 \\
&= \int_{-\infty}^{\infty} e^{-x_1} \exp\{-e^{-x_1} - e^{-(z+x_1)}\} dx_1 && [e^x e^y = e^{x+y}] \\
&= \int_{-\infty}^{\infty} e^{-x_1} \exp\{-e^{-x_1} - e^{-z} e^{x_1}\} dx_1 \\
&= \int_{-\infty}^{\infty} e^{-x_1} \exp\{-e^{-x_1}(1 + e^{z})\} dx_1 && [a + ab = a(1 + b)]
\end{aligned}$$

Derivation of logistic distribution

$$
\begin{aligned}
F_{x_1-x_2}(z) &= P(x_1 - x_2 < z) \\
&= P(x_2 < z + x_1) \\
&= \int_{-\infty}^{\infty} \lambda(x_1) \int_{-\infty}^{z+x_1} \lambda(x_2) dx_2 dx_1 \\
& [\ldots] \\
&= \int_{-\infty}^{\infty} e^{-x_1} \exp\{-e^{-x_1}(1+e^{z})\} dx_1 && [a + ab = a(1+b)] \\
&= \frac{w}{w} \int_{-\infty}^{\infty} e^{-x_1} \exp\{-e^{-x_1} w\} dx_1 && [w = 1 + \exp\{-z\}] \\
&= \frac{1}{w} \int_{-\infty}^{\infty} e^{-x_1} w \exp\{-e^{-x_1} w\} dx_1 && [e^{\log(w)} = w] \\
&= \frac{1}{w} \int_{-\infty}^{\infty} e^{-y} \exp\{-e^{-y}\} dy && [y = x_1 + \log(w)] \\
&= \frac{1}{w} = \frac{1}{1+e^{-z}}
\end{aligned}
$$

Theoretical foundations of Choice models

Two main approaches to deriving models of discrete choice:

1. Discriminal process (Thurstone, 1927)
   Most often associated with models of choice based on normal distributions (probit)
2. Axiomatic derivation (Luce, 1959; McFadden)
   Most often associated with models of choice based on logistic distribution (logit)

Thurstone: discriminal process

Logic of the discriminal process

1. Assume each object has a utility
   - *Thurstone deals with a single attribute, posits a psychological continuum*
   - the utility for item *i* is defined as u_i.
2. Assume utility includes a random component.
   - *Thurstone used Normal distribution, but acknowledged arbitrariness.*
   - $u_i = \mu_i + \epsilon_i$ where $\epsilon_i \sim N(0, \sigma^2)$
3. Individual picks item with the higher utility
   - *A relative comparison, requiring only simple inequality statements.*
   - choose item *j* over *k* if $u_j > u_k$.

Thurstone: discriminal process

- General binary choice, alternative j vs alternative k

$$\begin{aligned} P(j,k) &= P(u_j > u_k) \\ &= P(\mu_j + \epsilon_j > \mu_k + \epsilon_k) \\ &= P(\mu_j - \mu_k + \epsilon_j > \epsilon_k) \\ &= P(\mu_j - \mu_k > \epsilon_k - \epsilon_j) \end{aligned}$$

- If each ϵ is iid Gumbel, then

$$\begin{aligned} P(j,k) &= P(u_j > u_k) \\ &= P(\mu_j - \mu_k > \epsilon_k - \epsilon_j) \\ &= P(z > \delta) = F(z) \end{aligned}$$

where

- $z = \mu_j - \mu_k$
- $\delta = \epsilon_k - \epsilon_j$,
- and F is logistic cdf.

Thurstone: discriminal process

Assume ϵ is iid Gumbel, then

$$
\begin{aligned}
P(j,k) &= P(u_j > u_k) \\
&= P(\mu_j - \mu_k + \epsilon_j > \epsilon_k) \\
&= \int_{-\infty}^{\infty} \lambda(\epsilon_j) \int_{-\infty}^{\mu_j - \mu_k + \epsilon_j} \lambda(\epsilon_k) \\
&= \int_{-\infty}^{\infty} \lambda(\epsilon_j) \Lambda(\mu_j - \mu_k + \epsilon_j) \\
&= \int_{-\infty}^{\infty} \lambda(\epsilon_j) \Lambda(\mu_j - \mu_k + \epsilon_j) \\
&= \frac{1}{w} \int_{-\infty}^{\infty} -e^{-\epsilon_j} w \exp\{-e^{-\epsilon_j} w\} \\
&= \frac{1}{w} = \frac{1}{1 + \exp\{-(\mu_j - \mu_k)\}}
\end{aligned}
$$

Thurstone: discriminal process

Assume ϵ is iid $\mathcal{N}(0, \sigma_i^2)$

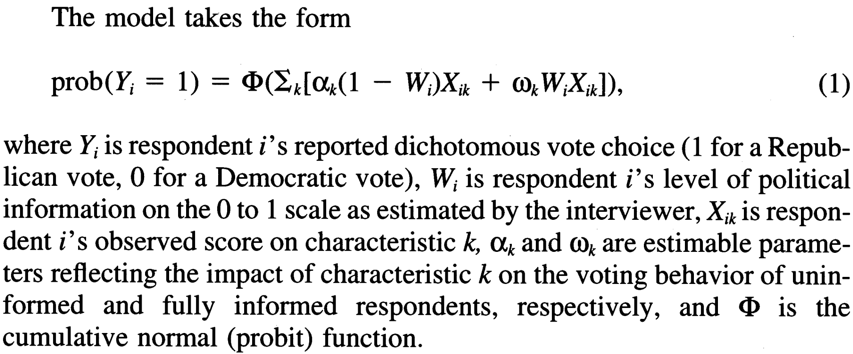
Recall, for independent normal distributions

$$\mathcal{N}(\mu_i, \sigma_i^2) - \mathcal{N}(\mu_j, \sigma_j^2) = \mathcal{N}(\mu_i - \mu_j, \sigma_i^2 + \sigma_j^2)$$

So,

$$\begin{aligned} P(j, k) &= P(\mu_j - \mu_k > \epsilon_k - \epsilon_j) \\ &= \int_{-\infty}^{\mu_j - \mu_k} \frac{1}{\sqrt{2\pi(\sigma_j^2 + \sigma_k^2)}} \exp\left\{ -\frac{z^2}{2\sqrt{\sigma_j^2 + \sigma_k^2}} \right\} \partial z \\ &= \Phi\left(\frac{\mu_j - \mu_k}{\sqrt{\sigma_j^2 + \sigma_k^2}} \right) \end{aligned}$$

Bartels (1996)



The model takes the form

$$\text{prob}(Y_i = 1) = \Phi(\Sigma_k[\alpha_k(1 - W_i)X_{ik} + \omega_k W_i X_{ik}]), \qquad (1)$$

where Y_i is respondent i's reported dichotomous vote choice (1 for a Republican vote, 0 for a Democratic vote), W_i is respondent i's level of political information on the 0 to 1 scale as estimated by the interviewer, X_{ik} is respondent i's observed score on characteristic k, α_k and ω_k are estimable parameters reflecting the impact of characteristic k on the voting behavior of uninformed and fully informed respondents, respectively, and Φ is the cumulative normal (probit) function.

Bartels (1996)

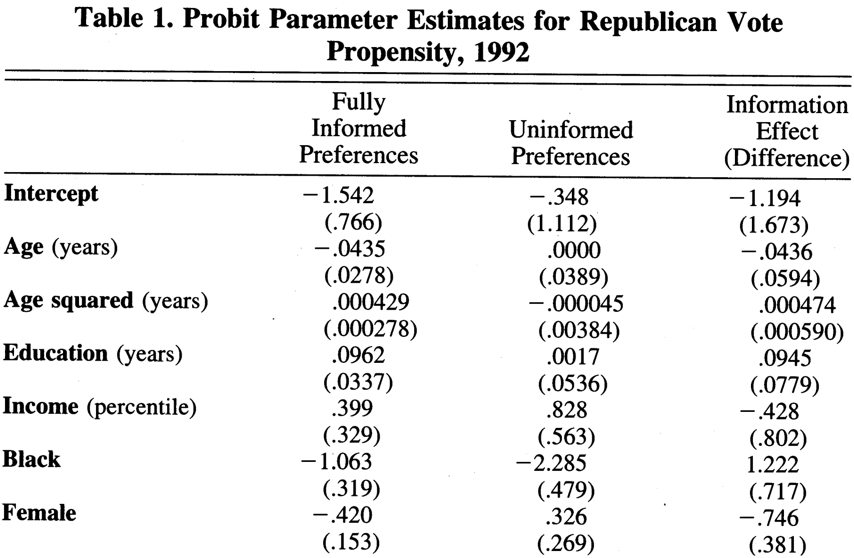


Table 1. Probit Parameter Estimates for Republican Vote Propensity, 1992

	Fully Informed Preferences	Uninformed Preferences	Information Effect (Difference)
Intercept	−1.542	−.348	−1.194
	(.766)	(1.112)	(1.673)
Age (years)	−.0435	.0000	−.0436
	(.0278)	(.0389)	(.0594)
Age squared (years)	.000429	−.000045	.000474
	(.000278)	(.00384)	(.000590)
Education (years)	.0962	.0017	.0945
	(.0337)	(.0536)	(.0779)
Income (percentile)	.399	.828	−.428
	(.329)	(.563)	(.802)
Black	−1.063	−2.285	1.222
	(.319)	(.479)	(.717)
Female	−.420	.326	−.746
	(.153)	(.269)	(.381)

Thurstone: what is identified?

For each individual i, let

- x_i be observed data
- γ_i be a parameter

Consider three scenarios

- Given $\mu_i = x_i\gamma_i$
- Given $\mu_i = x\gamma_i$, i.e., $x_j = x_k = x$
- Given $\mu_i = x_i\gamma$, i.e., $\gamma_j = \gamma_k = \gamma$

Question: what can be identified in a discriminal model?

Thurstone: what is identified?

Given,

$$P(j,k) = \frac{1}{1+\exp\{-(\mu_j - \mu_k)\}}$$

and $\mu_i = x_i \gamma_i$, consider what can be identified?

If $x_j = x_k = x$, then

$$P(j,k) = \frac{1}{1+\exp\{-x(\gamma_j - \gamma_k)\}} = \frac{1}{1+\exp\{-x\beta\}}$$

where $\beta = \gamma_j - \gamma_k$.

If $\gamma_j = \gamma_k = \gamma$, then

$$P(j,k) = \frac{1}{1+\exp\{-\beta(x_j - x_k)\}}$$

where $\beta = \gamma$.

Thurstone: what is identified?

Given,

$$P(j,k) \quad = \quad \Phi\left(\frac{\mu_j - \mu_k}{\sqrt{\sigma_j^2 + \sigma_k^2}}\right)$$

and $\mu_i = x_i\gamma_i$, consider parameters are identified?

If $x_j = x_k = x$, then

$$P(j,k) \quad = \quad \Phi\left(x\frac{\gamma_j - \gamma_k}{\sqrt{\sigma_j^2 + \sigma_k^2}}\right) = \Phi\left(x\beta\right)$$

where $\beta = \frac{\gamma_j - \gamma_k}{\sqrt{\sigma_j^2 + \sigma_k^2}}$.

Thurstone: what is identified?

Given,

$$P(j,k) \quad = \quad \Phi\left(\frac{\mu_j - \mu_k}{\sqrt{\sigma_j^2 + \sigma_k^2}}\right)$$

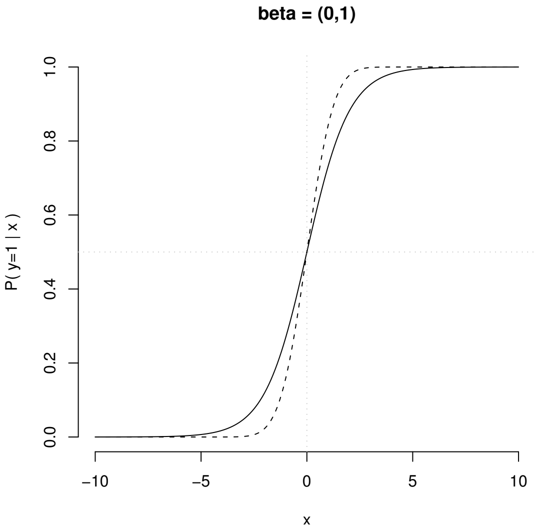
and $\mu_i = x_i\gamma_i$, consider what can be identified?

If $\gamma_j = \gamma_k = \gamma$, then

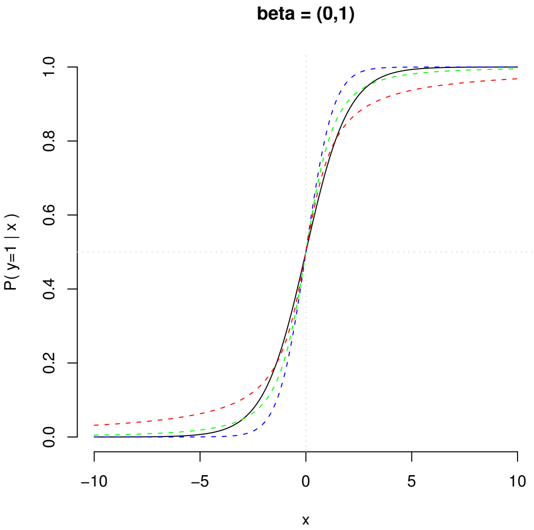
$$P(j,k) \quad = \quad \Phi\left((x_j - x_k)\frac{\gamma}{\sqrt{\sigma_j^2 + \sigma_k^2}}\right) = \Phi\left((x_j - x_k)\beta\right)$$

where $\beta = \frac{\gamma}{\sqrt{\sigma_j^2+\sigma_k^2}}$.

Logit and Probit, CDF



Logit and Cauchit (t), CDF



Axiomatic Foundations of Choice Models

Assumptions

D1 Let $R \subset S \subset T \subset U$.

D2 Let $x, y, z \in T$, arbitrary elements of choice set.

D3 Let $P(x, y)$ be the probability of choosing x instead of y, $0 < P(x, y) < 1$.

D4 $P_S(R)$ is the probability of choosing R given choice from among alternatives in S.

Choice Axiom

(i) $P_T(R) = P_S(R)P_T(S)$

(ii) If $P(x, y) = 0$ for some $x, y \in T$, $P_T(S) = P_{T-\{x\}}(S - \{x\})$

Axiomatic Foundations of Choice Models

Axiom of Choice

- Defines relationship which defines how choices within subsets are related in the context of an individual making probabilistic choices.
- Can be rewritten as $P_T(R \mid S) P_T(S) = P_T(R)$
- Two core implications,
  Lemma 3: Independence of Irrelevant Alternatives (IIA)
  Theorem 3: Probability must satisfy a ratio scale

Axiomatic Foundations of Choice Models

Lemma 3 (Independence from irrelevant alternatives):
For $x, y \in S$,

$$\frac{P(x,y)}{P(y,x)} = \frac{P_S(x)}{P_S(y)}$$

Proof:
By Axiom we have

$$P_S(x) = P(x,y)[P_S(x) + P_S(y)]$$

So

$$\begin{aligned} P_S(x) &= P(x,y)[P_S(x) + P_S(y)] \\ P_S(x) &= P(x,y)P_S(x) + P(x,y)P_S(y) \\ (1 - P(x,y))P_S(x) &= P(x,y)P_S(y) \\ P(y,x)P_S(x) &= P(x,y)P_S(y) \\ \frac{P(x,y)}{P(y,x)} &= \frac{P_S(x)}{P_S(y)} \end{aligned}$$

Axiomatic Foundations of Choice Models

Luce Lemma 3 (Independence from irrelevant alternatives):
What does this mean?

- relative probability of choosing two alternatives is invariant to the composition of the larger set of alternatives.
- Only ratio is invariant, not probabilities themselves
- Might also hear that log-odds of two choices are constant: $\log(P_S(x)) - \log(P_S(y)) = c$.

Axiomatic Foundations of Choice Models

Luce Lemma 3 (Independence from irrelevant alternatives):
Why so cool?

- can estimate parameters defining utility of choices even with only a subset.
- ** Not generally possible if IIA does not hold (e.g., correlation between utilities of choices)—then to estimate any choice must model all choices.
- ** Neither holds in general for models of choice (by design) nor is it plausible that it in general holds empirically.

Axiomatic Foundations of Choice Models

Theorem 3: choice probability is ratio scale
$\exists v : T \rightarrow \Re_+$, unique up to multiplication by $k > 0$, such that

$$P_S(x) = \frac{v(x)}{\sum_{y \in S} v(y)} = \frac{1}{1 + \sum_{y \in S - \{x\}} v(y)/v(x)}$$

McFadden and Yellot have each pointed out the connection to logit models by setting $v(x) = e^x$.

E.g.,

$$P(x, y) = \frac{v(x)}{v(x) + v(y)} = \frac{e^{\mu_x}}{e^{\mu_x} + e^{\mu_y}} = \frac{1}{1 + e^{\mu_y}/e^{\mu_x}} = \frac{1}{1 + e^{-(\mu_x - \mu_y)}}$$

Yellot shows that discriminal process based on Type I discrete value distribution is uniquely equivalent to Choice Axiom.

Axiomatic Foundations of Choice Models

Let $S \in \{1,2,3\}$, and $P_S(j)$ be probability of choosing j from S,

$$\begin{aligned}
P_S(y=1) &= \frac{1}{1+e^{\mu_2-\mu_1}+e^{\mu_3-\mu_1}} = \frac{e^{\mu_1}}{e^{\mu_1}+e^{\mu_2}+e^{\mu_3}} \\
P_S(y=2) &= \frac{1}{1+e^{\mu_1-\mu_2}+e^{\mu_3-\mu_2}} = \frac{e^{\mu_2}}{e^{\mu_1}+e^{\mu_2}+e^{\mu_3}} \\
P_S(y=3) &= \frac{1}{1+e^{\mu_1-\mu_3}+e^{\mu_2-\mu_3}} = \frac{e^{\mu_3}}{e^{\mu_1}+e^{\mu_2}+e^{\mu_3}}
\end{aligned}$$

So,

$$\frac{P_S(y=1)}{P_S(y=2)} = \frac{e^{\mu_1}}{e^{\mu_2}}$$

Axiomatic Foundations of Choice Models

Let $T \in \{1, 2\}$, and $P_T(j)$ be probability of choosing j from T, Recal logit (special case of MNL),

$$\begin{aligned} P_T(y = 1) &= \frac{1}{1 + e^{\mu_2 - \mu_1}} = \frac{e^{\mu_1}}{e^{\mu_1} + e^{\mu_2}} \\ P_T(y = 2) &= \frac{1}{1 + e^{\mu_1 - \mu_2}} = \frac{e^{\mu_2}}{e^{\mu_2} + e^{\mu_2}} \end{aligned}$$

So,

$$\frac{P_T(y = 1)}{P_T(y = 2)} = \frac{e^{\mu_1}}{e^{\mu_2}}$$

Axiomatic Foundations of Choice Models

Comparing probabilities of choosing 1 and 2 in logit and MNL,

$$\frac{P_T(1)}{P_T(2)} = \frac{e^{\mu_1}}{e^{\mu_2}} = \frac{P_S(1)}{P_S(2)}$$

MNL conforms to Choice Axiom/IIA.

See Yellot (1977) and McFadden (1973) for connections between Luce and EV Type 1.

Bartels (1996)

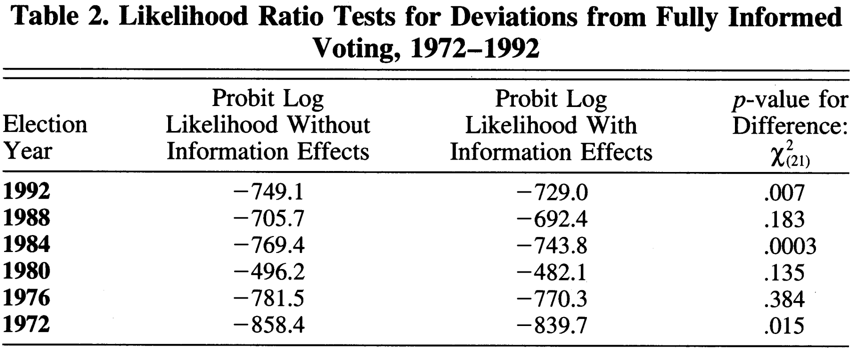
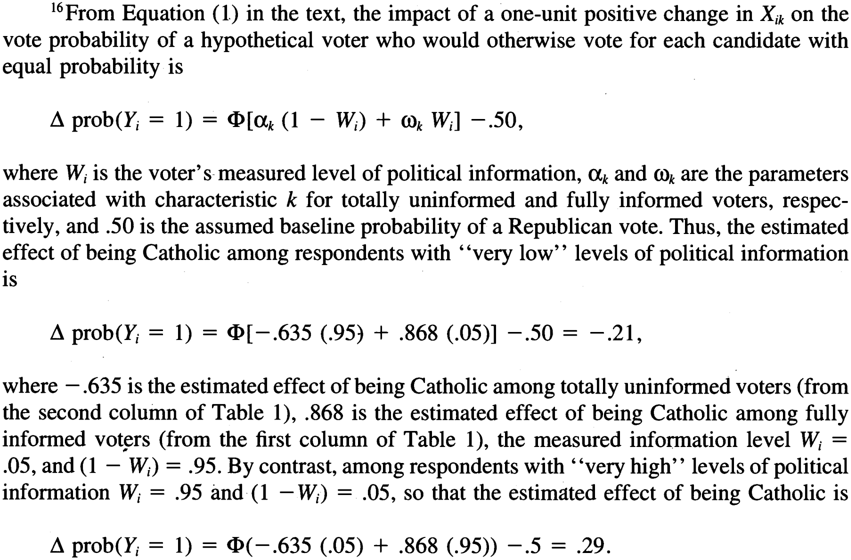


Table 2. Likelihood Ratio Tests for Deviations from Fully Informed Voting, 1972–1992

Election Year	Probit Log Likelihood Without Information Effects	Probit Log Likelihood With Information Effects	*p*-value for Difference: $\chi^2_{(21)}$
1992	−749.1	−729.0	.007
1988	−705.7	−692.4	.183
1984	−769.4	−743.8	.0003
1980	−496.2	−482.1	.135
1976	−781.5	−770.3	.384
1972	−858.4	−839.7	.015

Bartels (1996)



[16] From Equation (1) in the text, the impact of a one-unit positive change in X_{ik} on the vote probability of a hypothetical voter who would otherwise vote for each candidate with equal probability is

$$\Delta \text{ prob}(Y_i = 1) = \Phi[\alpha_k (1 - W_i) + \omega_k W_i] - .50,$$

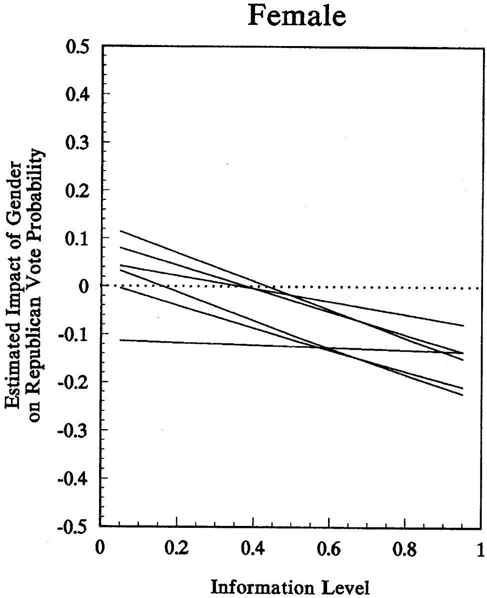
where W_i is the voter's measured level of political information, α_k and ω_k are the parameters associated with characteristic k for totally uninformed and fully informed voters, respectively, and .50 is the assumed baseline probability of a Republican vote. Thus, the estimated effect of being Catholic among respondents with ''very low'' levels of political information is

$$\Delta \text{ prob}(Y_i = 1) = \Phi[-.635\ (.95) + .868\ (.05)] - .50 = -.21,$$

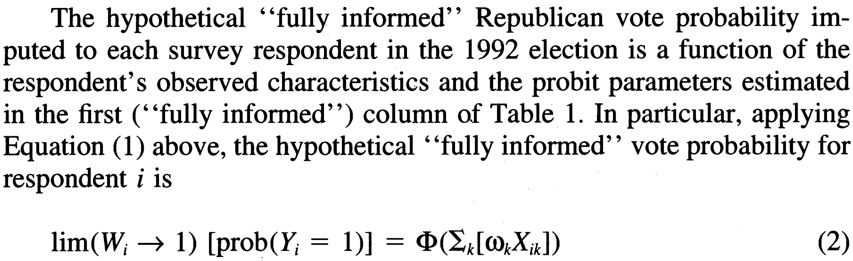
where $-.635$ is the estimated effect of being Catholic among totally uninformed voters (from the second column of Table 1), .868 is the estimated effect of being Catholic among fully informed voters (from the first column of Table 1), the measured information level $W_i =$ .05, and $(1 - W_i) = .95$. By contrast, among respondents with ''very high'' levels of political information $W_i = .95$ and $(1 - W_i) = .05$, so that the estimated effect of being Catholic is

$$\Delta \text{ prob}(Y_i = 1) = \Phi(-.635\ (.05) + .868\ (.95)) - .5 = .29.$$

Bartels (1996)



Bartels (1996)



The hypothetical ‘‘fully informed’’ Republican vote probability imputed to each survey respondent in the 1992 election is a function of the respondent’s observed characteristics and the probit parameters estimated in the first (‘‘fully informed’’) column of Table 1. In particular, applying Equation (1) above, the hypothetical ‘‘fully informed’’ vote probability for respondent i is

$$\lim(W_i \rightarrow 1)\ [\mathrm{prob}(Y_i = 1)] = \Phi(\Sigma_k[\omega_k X_{ik}]) \tag{2}$$

Bartels (1996)

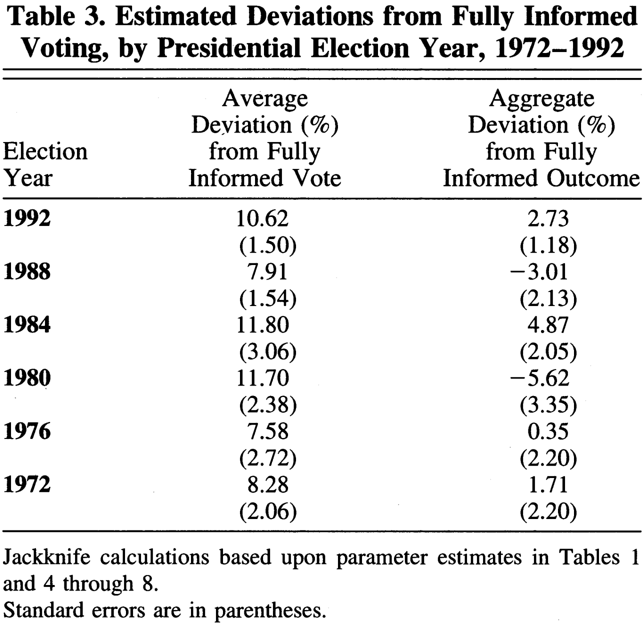


Table 3. Estimated Deviations from Fully Informed Voting, by Presidential Election Year, 1972–1992

Election Year	Average Deviation (%) from Fully Informed Vote	Aggregate Deviation (%) from Fully Informed Outcome
1992	10.62	2.73
	(1.50)	(1.18)
1988	7.91	−3.01
	(1.54)	(2.13)
1984	11.80	4.87
	(3.06)	(2.05)
1980	11.70	−5.62
	(2.38)	(3.35)
1976	7.58	0.35
	(2.72)	(2.20)
1972	8.28	1.71
	(2.06)	(2.20)

Jackknife calculations based upon parameter estimates in Tables 1 and 4 through 8.
Standard errors are in parentheses.