

4 Asymptotics and Robustness

4.1 Consistency

- b **Relative frequency** $R_n \rightarrow P\langle A \rangle = \pi = \mathcal{E}\langle R_n \rangle$. More precisely,

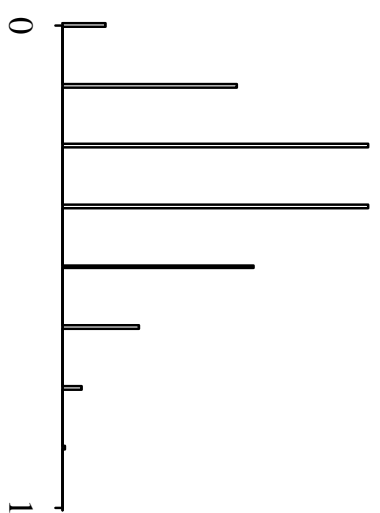
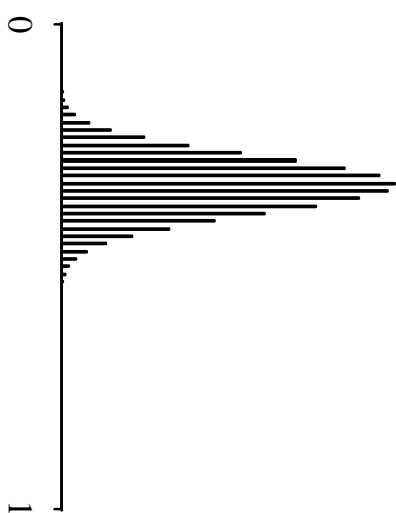
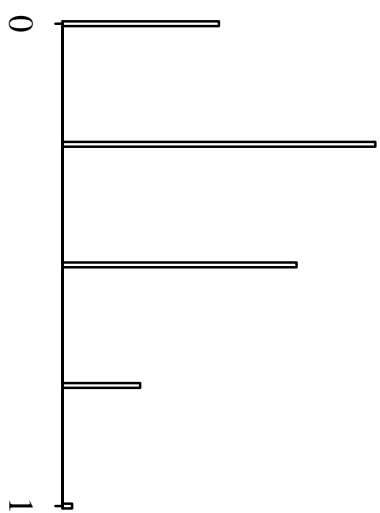
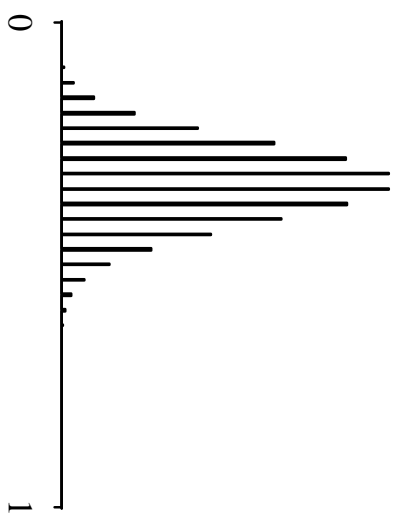
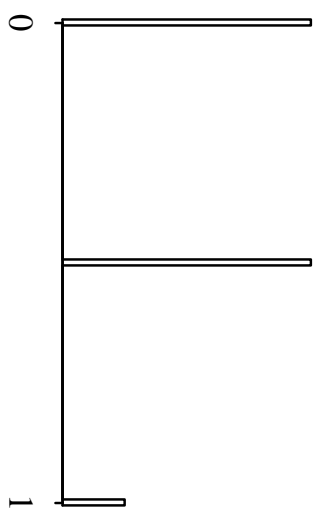
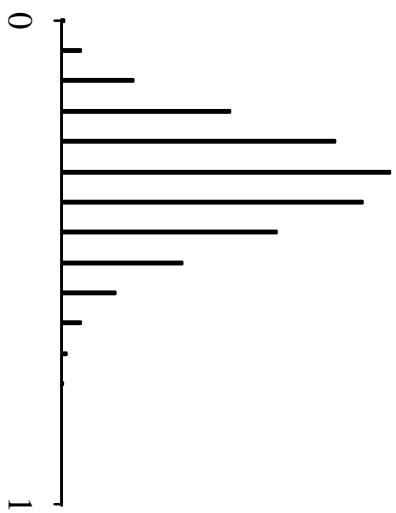
$$\lim_{n \rightarrow \infty} P\{|R_n - \pi| > \varepsilon\} = 0$$

Law of Large Numbers. Jakob Bernoulli (published posth. 1713)

$$X \sim \mathcal{B}\langle n, \pi \rangle \longrightarrow \mathcal{E}\langle X/n \rangle = \mathcal{E}\langle X \rangle / n = n\pi / n = \pi$$

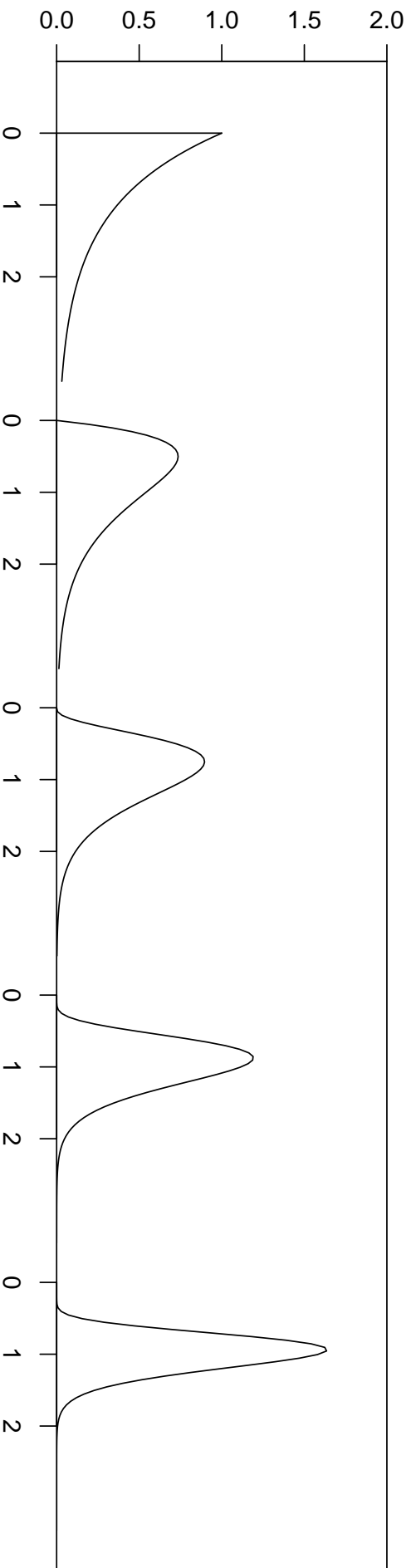
$$\text{var}\langle X/n \rangle = \text{var}\langle X \rangle / n^2 = n\pi(1 - \pi) / n^2 = \pi(1 - \pi) / n$$

$$\rightarrow 0! \implies X/n \rightarrow \pi.$$



c General case: $X_1, X_2, \dots, X_n \dots$ indep. \longrightarrow

$$P\{|\bar{X}_n - \mu| > \varepsilon\} \xrightarrow{n \rightarrow \infty} 0 \quad \text{for each } \varepsilon > 0,$$



d Empirical cumulative **distribution function** \longrightarrow theoretical,

$$\hat{F}_n(x) = \frac{1}{n} \# \langle i : X_i \leq x \rangle \xrightarrow{n \rightarrow \infty} F(x) = P\langle X \leq x \rangle .$$

e **Consistency of the characteristic values.** $\hat{\gamma} \rightarrow \gamma$.

f **Functionals.**

$T\langle F \rangle$ theoretical characteristic value, $T\langle \hat{F}_n \rangle$ empirical.

$$T\langle \hat{F}_n \rangle \xrightarrow{n \rightarrow \infty} T\langle F \rangle$$

* Mathematical conditions, “regularity assumptions”.

g **Integrals** $\mathcal{E}\langle X \rangle = \int x dF\langle x \rangle$

For continuous random variables:

$$\mathcal{E}\langle X \rangle = \int x dF\langle x \rangle = \int x f\langle x \rangle dx$$

For discrete ones:

$$\mathcal{E}\langle X \rangle = \int x dF\langle x \rangle = \sum_x x \langle X = x \rangle$$

For empirical distribution functions:

$$\int x d\hat{F}_n\langle x \rangle = \sum_i x_i \frac{1}{n} = \bar{X}$$

h Let $\mathcal{F}_{\underline{\theta}}$ be a parametric family with par. $\underline{\theta} = [\theta_1, \dots, \theta_p]$

Sample $X_1, \dots, X_n, X_i \sim \mathcal{F}_{\underline{\theta}}$

Estimator of the parameters:

A function of the sample, $T_k \langle \hat{F}_n \rangle$ designed to estimate par. θ_k

→ If $F = F_{\underline{\theta}}$ then we want $T_k \langle F_{\underline{\theta}} \rangle = \theta_k$
Fisher consistency.

j **The Location Model**

$X_1, X_2, \dots, X_n, X_i \sim \mathcal{N} \langle \mu, \sigma_0^2 \rangle, \text{ independent}$

estimator: \bar{X}

– or median? $\text{med} \langle X_1, X_2, \dots, X_n \rangle \rightarrow \text{med} \langle N \langle \mu, \sigma_0^2 \rangle \rangle = \mu$

4.2 Maximum likelihood and M-estimators

a Given a parametric family with density $f_{\underline{\theta}}(x)$

write as $f(x, \underline{\theta})$

Given the observation x , $f(x, \underline{\theta}) =$ likelihood of $\underline{\theta}$.

Same for probabilities $P_{\underline{\theta}}(X = x)$, also written as $f(x, \underline{\theta})$.

Maximum-Likelihood estimator: maximize the likelihood !

$$\hat{\underline{\theta}} = \arg \min \langle f(x, \underline{\theta}) \rangle$$

b Sample $X_1, \dots, X_n \longrightarrow$ joint density $\prod_i f(x_i, \underline{\theta})$.

Maximize this or

$$L(x_1, \dots, x_n; \underline{\theta}) = \sum_i \log \langle f(x_i, \underline{\theta}) \rangle$$

or minimize

$$D(\underline{\theta}) = -2 \sum_i \log \langle f(x_i, \underline{\theta}) \rangle = \sum_i \rho(x_i, \underline{\theta})$$

D : Deviance.

ρ -function: “deviation” of the observation x_i from model $\langle \underline{\theta} \rangle$.

For normal distribution with given variance:

$$\rho(x, \mu) = ((x - \mu)/\sigma)^2 + c$$

\longrightarrow Least Squares

- c Example logistic distribution.

$$f_{\mu, \sigma} \langle x \rangle = \frac{1}{\left(e^{z/2} + e^{-z/2} \right)^2}, \quad z = \frac{x - \mu}{\sigma}$$

location-scale family. Likelihood $-2 \log \langle e^{z/2} + e^{-z/2} \rangle$.

$$\rho \langle x; \mu, \sigma \rangle = 4 \log \left\langle e^{(x-\mu)/2\sigma} + e^{-(x-\mu)/2\sigma} \right\rangle$$

- d Form derivative for θ_k and set to 0

$$\frac{\partial L}{\partial \theta_k} \langle x_1, \dots, x_n; \underline{\theta} \rangle = \sum_{i=1}^n s_k \langle x_i; \underline{\theta} \rangle,$$

„Likelihood scores”

$$s_k \langle x; \underline{\theta} \rangle = \frac{\partial}{\partial \theta_k} \log \langle f \langle x_i, \underline{\theta} \rangle \rangle = -\frac{1}{2} \frac{\partial}{\partial \theta_k} \rho \langle x_i, \underline{\theta} \rangle.$$

$$\sum_{i=1}^n \underline{s} \langle x_i; \hat{\underline{\theta}} \rangle = 0.$$

Maximum-Likelihood estimator: Solve for $\hat{\underline{\theta}}$!

e **Logistic distribution.** Location-scale: $z = (x - \mu)/\sigma \longrightarrow$

$$\partial z / \partial \mu = -1/\sigma \quad \text{and} \quad \partial z / \partial \sigma = -(x - \mu)/\sigma^2 = -z/\sigma$$

$$\begin{aligned} s_\mu \langle x; \underline{\theta} \rangle &= \frac{1}{\sigma} \frac{e^{z/2} - e^{-z/2}}{e^{z/2} + e^{-z/2}} \\ s_\sigma \langle x; \underline{\theta} \rangle &= z s_\mu \langle x; \underline{\theta} \rangle \end{aligned}$$

f **More general:**

$$\hat{\underline{\theta}} = \operatorname{argmin}_{\underline{\theta}} \sum_{i=1}^n \rho \langle X_i, \underline{\theta} \rangle \quad \text{or}$$

$$\hat{\underline{\theta}} = \text{solution of } \sum_{i=1}^n \underline{\psi} \langle X_i, \underline{\theta} \rangle = 0$$

M estimator.

Use scores function of the logistic distribution even if you think (or hope) that the observations follow the normal distribution.

g

M estimators as functionals.

$$T_\rho\langle F \rangle = \operatorname{argmin}_{\underline{\theta}} \int \rho\langle x, \underline{\theta} \rangle dF\langle x \rangle \quad \text{or}$$

$$T_\psi\langle F \rangle = \text{solution of } \int \underline{\psi}\langle x, \underline{\theta} \rangle dF\langle x \rangle = 0.$$

Estimator: Solution of

$$\int \underline{\psi}\langle x, \underline{\theta} \rangle d\hat{F}_n\langle x \rangle = \frac{1}{n} \sum_i \underline{\psi}\langle x_i, \underline{\theta} \rangle = 0$$

$$T\langle F_n \rangle \rightarrow T\langle F \rangle.$$

T estimates θ if $T\langle F_{\underline{\theta}} \rangle = \underline{\theta} \implies$

$$\int \underline{\psi}\langle x, \underline{\theta} \rangle dF_{\underline{\theta}}\langle x \rangle = 0$$

h Example: Location model and Huber estimator

$$f\langle x, \mu \rangle = f\langle x - \mu, 0 \rangle, \quad f\langle z, 0 \rangle = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

log-lik.: $-\frac{1}{2} \log\langle 2\pi \rangle - (x - \mu)^2/2$, scores $s\langle x, \mu \rangle = x - \mu$

→ estimator $\sum_i (x_i - \hat{\mu}) = 0 \rightarrow \hat{\mu} = \frac{1}{n} \sum_i x_i$

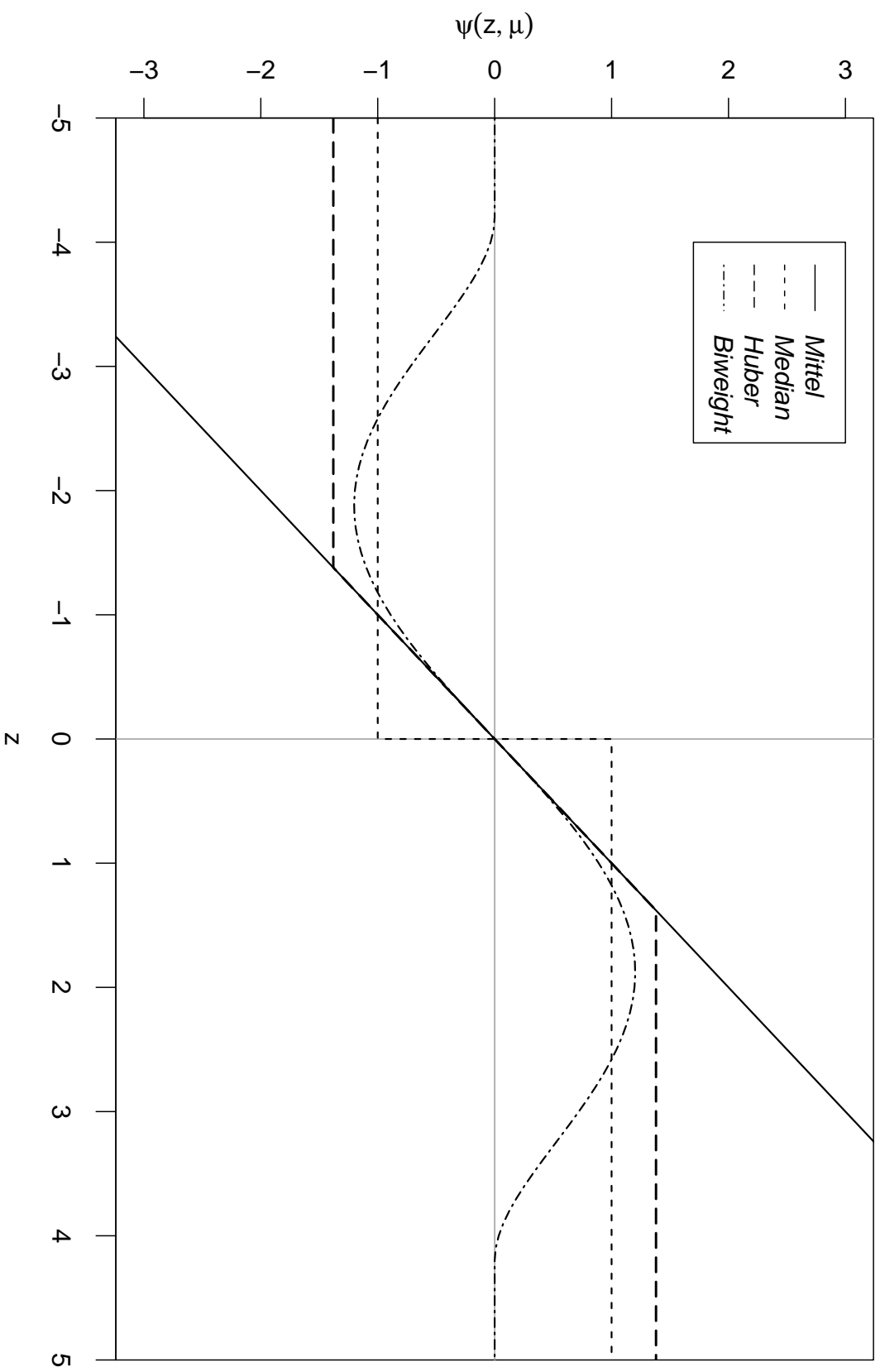
→ M estimator with $\psi\langle x, \mu \rangle = x - \mu$

Median is an M estimator, $\psi\langle x, \mu \rangle = \begin{cases} -1 & x - \mu \leq 0 \\ 1 & x - \mu > 0 \end{cases}$.

Huber estimator: M estimator with

$$\psi\langle x, \mu \rangle = \begin{cases} x - \mu & \text{für } |x - \mu| \leq k \\ -k & \text{für } x - \mu < -k \\ k & \text{für } x - \mu > k \end{cases}$$

k tuning constant.



- i Maximum-Likelihood estimator

$$\int \underline{s}\langle x, \underline{\theta} \rangle f\langle x, \underline{\theta} \rangle dx = 0$$

Fisher-consistent estimators for the parameter(s) of the family!

Proof:

$$\frac{\partial}{\partial \theta} \log \langle f\langle x, \underline{\theta} \rangle \rangle = \frac{1}{\frac{\partial}{\partial \theta} f\langle x, \underline{\theta} \rangle} \frac{\partial}{\partial \theta} f\langle x, \underline{\theta} \rangle = \underline{s}\langle x, \underline{\theta} \rangle \implies$$

$$\frac{\partial}{\partial \theta} f\langle x, \underline{\theta} \rangle = \underline{s}\langle x, \underline{\theta} \rangle f\langle x, \underline{\theta} \rangle$$

$$\int f\langle x, \underline{\theta} \rangle dx = 1 \implies \int \frac{\partial}{\partial \theta} f\langle x, \underline{\theta} \rangle dx = 0$$

4.3 Influence Function

- b **Empirical Influence function.** Example of weights of pigs

107 108 111 101 97 113 109 105 116 122.

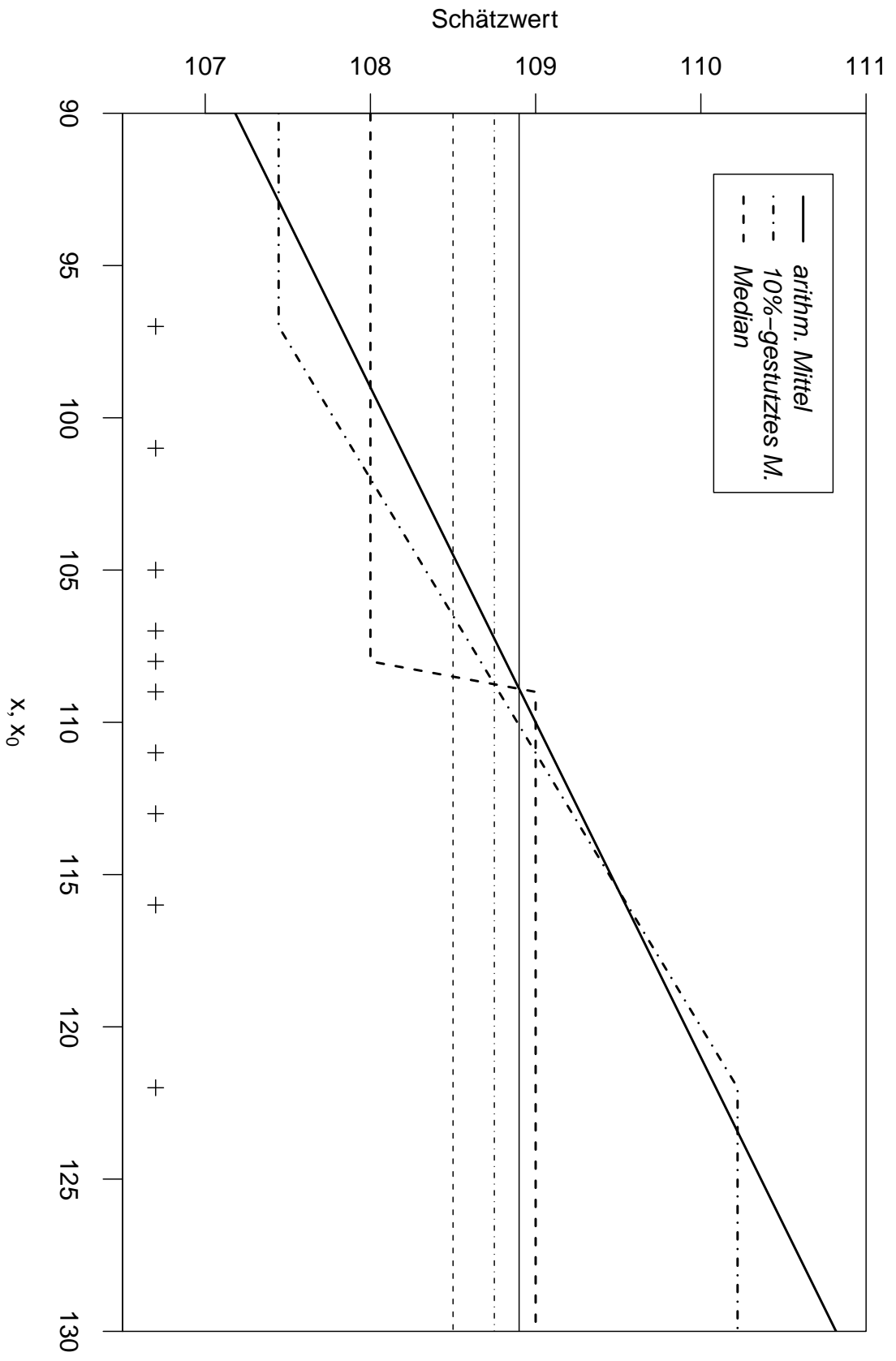
Additional observation $x_0 \rightarrow$

$$\begin{aligned} & (107 + 108 + \dots + 122 + x_0)/11 \\ &= \frac{n\bar{x} + x_0}{n+1} = \frac{(n+1)\bar{x} - \bar{x}}{n+1} + \frac{x_0}{n+1} = \bar{x} + \frac{1}{n+1}(x_0 - \bar{x}) \\ &= 108.9 + (x_0 - 108.9)/11. \end{aligned}$$

Median $(108 + 109)/2 = 108.5$

$\rightarrow 108$, if $x_0 \leq 108$

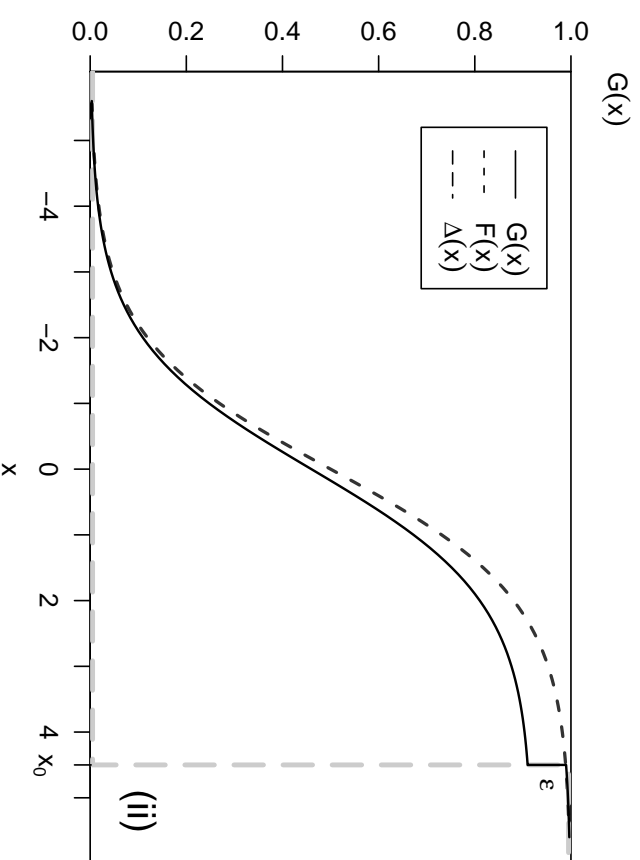
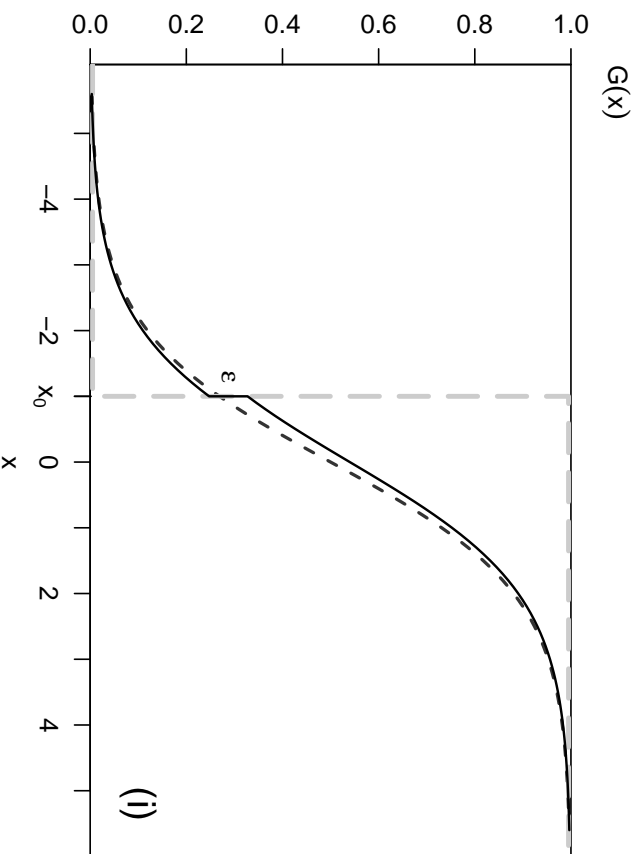
$\rightarrow 109$, if $x_0 \geq 109$



c Empirical Influence Function or Sensitivity Curve

$$SC \langle x_0; T, x_1, \dots, x_n \rangle = n \left(T \langle x_1, \dots, x_n, x_0 \rangle - T \langle x_1, \dots, x_n \rangle \right)$$

d Gross Error Model. $(1 - \varepsilon) F \langle \cdot \rangle + \varepsilon \Delta_{x_0} \langle \cdot \rangle$
 Model for “wrong” observation or “gross error”



More general: $G \langle x \rangle = (1 - \varepsilon) F \langle x \rangle + \varepsilon H \langle x \rangle$

e **Influence Function.**

$$\text{IF}\langle x; T, F \rangle = \lim_{\varepsilon \rightarrow 0} \frac{T\langle (1 - \varepsilon)F + \varepsilon\Delta_x \rangle - T\langle F \rangle}{\varepsilon}.$$

f **IF for \bar{X} :**

$$\begin{aligned} \mathcal{E}\langle (1 - \varepsilon)F + \varepsilon\Delta_x \rangle &= (1 - \varepsilon)\mathcal{E}\langle F \rangle + \varepsilon\mathcal{E}\langle \Delta_x \rangle \\ &= (1 - \varepsilon)\mathcal{E}\langle F \rangle + \varepsilon x \\ \text{IF}\langle x; \bar{X}, F \rangle &= x - \mathcal{E}\langle F \rangle \end{aligned}$$

g_g^* Median $\text{med}\langle F \rangle = F^{-1}\langle 0.5 \rangle$.

If $x > t_\varepsilon^+ = F^{-1}\langle 0.5/(1-\varepsilon) \rangle$ then $\text{med}\langle (1-\varepsilon)F + \varepsilon\Delta_x \rangle = t_\varepsilon^+$

If $x > t_\varepsilon^- = F^{-1}\langle 1-0.5/(1-\varepsilon) \rangle$ then ... = t_ε^-

$$\begin{aligned} \frac{d}{d\varepsilon} F^{-1}\langle 0.5/(1-\varepsilon) \rangle &= \frac{1}{f\langle F^{-1}\langle 0.5/(1-\varepsilon) \rangle \rangle} \cdot \frac{0.5 \cdot (-1)}{(1-\varepsilon)^2} \cdot (-1) \\ &\rightarrow 1/(2f\langle \mu \rangle), \quad \mu = F^{-1}\langle 0.5 \rangle \end{aligned}$$

$$\text{IF}\langle x; \bar{X}, F \rangle = \begin{cases} -1/(2f\langle \mu \rangle) & \text{für } x < \text{med}\langle F \rangle \\ 1/(2f\langle \mu \rangle) & \text{für } x > \text{med}\langle F \rangle \end{cases}$$

is not continuous at the median, but jumps

from $-1/(2f\langle \mu \rangle)$ to $1/(2f\langle \mu \rangle)$.

h **Influence Function for M estimators.**

$$\text{IF}\langle x; F \rangle = \frac{1}{c} \psi\langle x, \theta \rangle \quad \text{mit} \quad c = - \int \frac{\partial}{\partial \theta} \psi\langle x, \theta \rangle f\langle x, \underline{\theta} \rangle dx .$$

Influence function is proportional to ψ .

If T estimates θ , that is, $\int \underline{s}\langle x, \underline{\theta} \rangle f\langle x, \underline{\theta} \rangle dx = 0$

$$c = \int \psi\langle x, \underline{\theta} \rangle s\langle x, \underline{\theta} \rangle f\langle x, \underline{\theta} \rangle dx .$$

Special case: Maximum-Likelihood estimator

$$c = - \int \frac{\partial}{\partial \theta} s\langle X, \theta \rangle f\langle x, \underline{\theta} \rangle dx = \int s\langle X, \theta \rangle^2 f\langle x, \underline{\theta} \rangle dx .$$

* Proof: $T\langle G \rangle$ for the distribution $G = (1 - \varepsilon)F + \varepsilon\Delta_x$:

$$\begin{aligned}
& \int \psi\langle x, T\langle G \rangle \rangle dG\langle x \rangle \\
&= (1 - \varepsilon) \int \psi\langle x, T\langle G \rangle \rangle dF\langle x \rangle + \varepsilon \psi\langle x, T\langle G \rangle \rangle \\
&\psi\langle x, T\langle G \rangle \rangle \approx \psi\langle x, T\langle F \rangle \rangle + \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle (T\langle G \rangle - T\langle F \rangle) \\
&\int \dots \approx \int \psi\langle x, T\langle F \rangle \rangle dF\langle x \rangle \\
&\quad + (T\langle G \rangle - T\langle F \rangle) \int \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle dF\langle x \rangle \\
&\approx (1 - \varepsilon)(T\langle G \rangle - T\langle F \rangle) \int \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle dF\langle x \rangle \\
&\quad + \varepsilon(\psi\langle x, T\langle F \rangle \rangle + (T\langle G \rangle - T\langle F \rangle) \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle) = 0 \\
&T\langle G \rangle - T\langle F \rangle \approx - \frac{\varepsilon(\psi\langle x, T\langle F \rangle \rangle + (T\langle G \rangle - T\langle F \rangle) \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle)}{(1 - \varepsilon) \int \frac{\partial}{\partial \theta} \psi\langle x, T\langle F \rangle \rangle dF\langle x \rangle}
\end{aligned}$$

i **Linearization.**

$$\begin{aligned} T\langle \hat{F}_n \rangle &\approx T\langle F \rangle + \frac{1}{n} \sum_{i=1}^n \text{IF}\langle X_i; T, F \rangle \\ T\langle G \rangle &\approx T\langle F \rangle + \int \text{IF}\langle x; T, F \rangle d(G - F) \end{aligned}$$

4.4 Asymptotic Distribution

a Central limit theorem.

The distribution of the standardized average

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}}$$

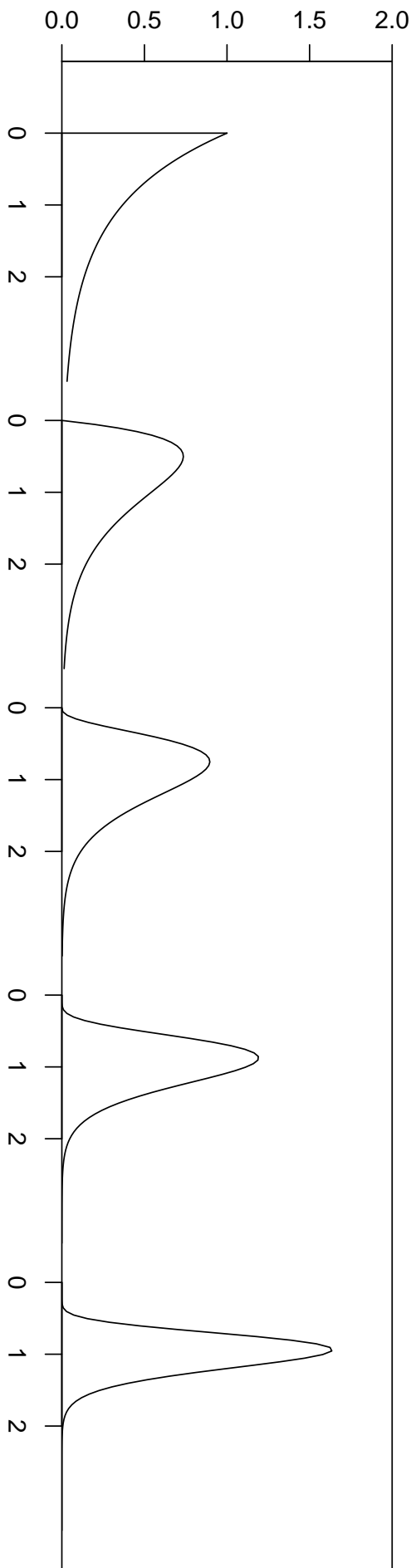
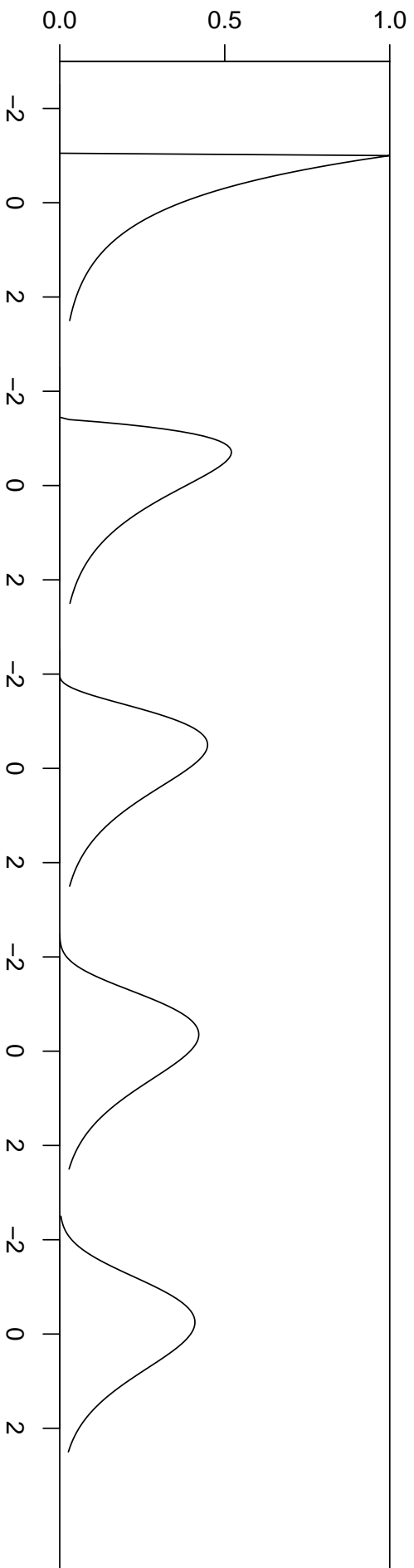
approaches the standard normal distribution as n grows.

$$P\langle Z_n \leq z \rangle \xrightarrow{n \rightarrow \infty} \Phi\langle z \rangle$$

(Φ : cum. distribution function of the standard normal)

... if X_i are independent and equally distributed
and the variance is finite.

b Expressed differently: $\bar{X}_n \approx \mathcal{N}\langle \mu, \sigma^2/n \rangle$



c **Central Limit Theorem for functionals.**

$$\begin{aligned} T\langle X_1, X_i, \dots, X_n \rangle &\approx T\langle F \rangle + \frac{1}{n} \sum_{i=1}^n \text{IF}\langle X_i; T, F \rangle \\ &\approx \sim \mathcal{N}\langle T\langle F \rangle, v/n \rangle \\ v &= \text{var}\langle \text{IF}\langle X; F \rangle \rangle \end{aligned}$$

We always have $\mathcal{E}\langle \text{IF}\langle X; T, F \rangle \rangle = 0$. \longrightarrow

$$v = \mathcal{E}\langle \text{IF}\langle X; T, F \rangle^2 \rangle .$$

d **Asymptotic variance for M estimators.**

$$v = \frac{1}{c^2} \int \psi\langle x, \theta \rangle^2 dF\langle x \rangle .$$

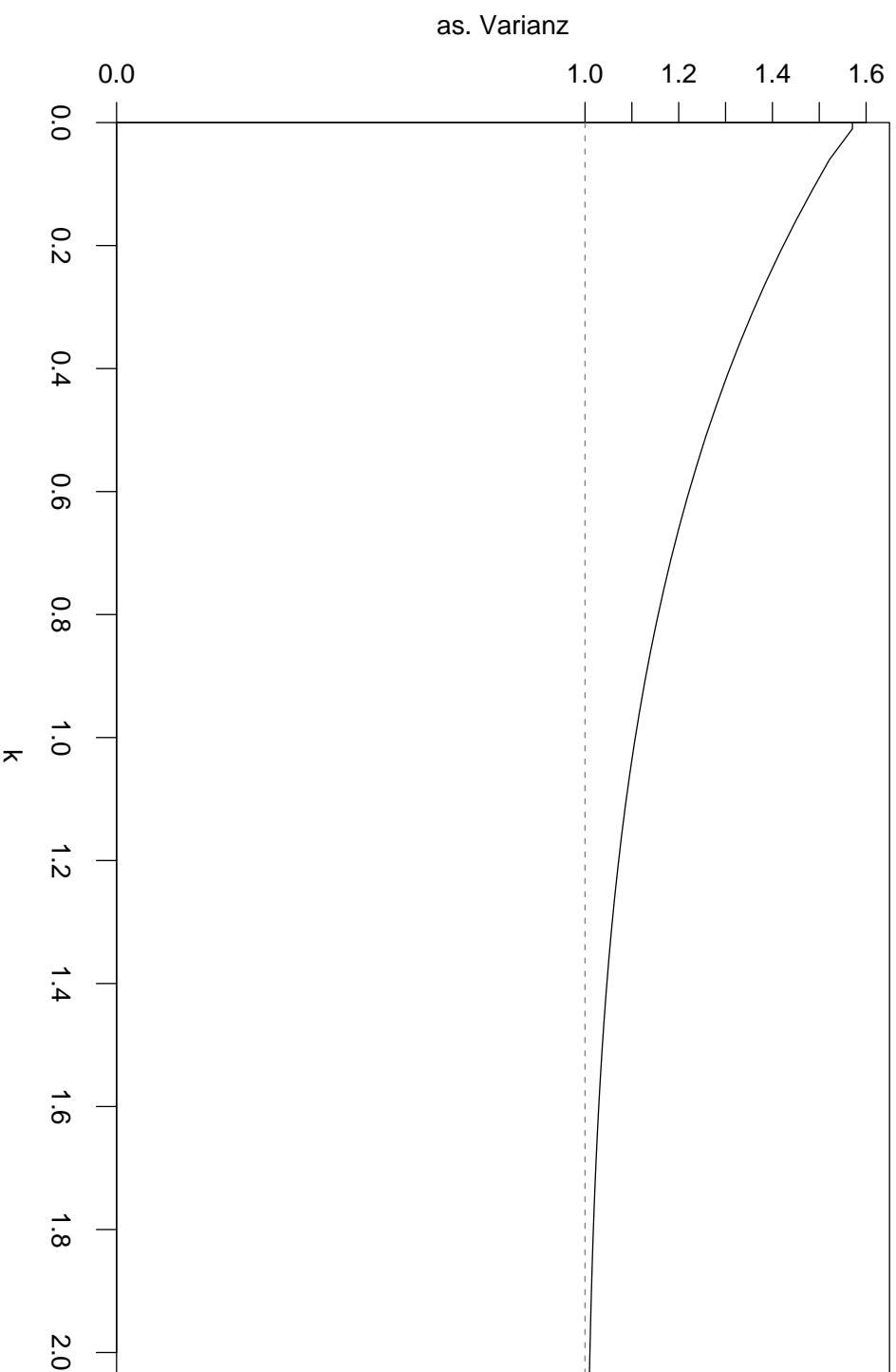
Maximum-Likelihood estimator: Integral $= c \longrightarrow$

$$v = 1/c, \quad c = \int s\langle x, \theta \rangle^2 dF_\theta\langle x \rangle .$$

Fisher-Information.

e **Example Huber estimator** for standard normal distribution

$$v = \frac{\int \psi \langle x \rangle^2 d\Phi \langle x \rangle}{\left(\int \psi' \langle x \rangle d\Phi \langle x \rangle \right)^2}$$



f **Example Logistic distribution**

Maximum-Likelihood estimator for $\mathcal{L}\langle \mu = 0, \sigma = 1 \rangle$

$$\int \frac{(e^{z/2} - e^{-z/2})^2}{(e^{z/2} + e^{-z/2})^4} dz = 0.333$$

= 1 for $\sigma = 1.732$.

g **Maximum likelihood estimator is the best est. asymptotically.**

$v_T \geq 1/c$ for all Fisher consistent estimators.

h Tests and confidence intervals.

- Standardized test statistic $T = (\hat{\theta} - \theta_0) / \sqrt{v/n} \approx \Phi$
 - Confidence interval for θ : $\hat{\theta} \pm 1.96 * \sqrt{v/n}$.
- i Sums of squares \longrightarrow chisquared distribution!.

$$T = \sum_k (N_k - \mu_k)^2 / \mu_k, \quad \mu_k = \mathcal{E}\langle N_k \rangle$$

$$T \approx \sim \chi^2$$

Contradiction to the Central Limit Theorem?

4.5 Likelihood Ratio Tests

a **Basic Idea.**

The plausibility of a model in the light of data is measured by the likelihood.

A null hypothesis usually restricts a parameter to a specific value (or one side of a given value, for one-sided case).

The restriction deteriorates the “fit” of the data to the model.

The likelihood decreases.

If it decreases too much, the null hypothesis must be rejected.

→ Test statistic:

- **likelihood ratio**, or
- log likelihood difference, or
- **deviance** – difference of deviance values between “full model” (free parameter) and “reduced model” (parameter fixed at “null value”)

- b Example: simple regression, scale known. Log likelihood:

$$c - \frac{1}{2\sigma^2} \sum_i (y_i - \beta_1 x_i - \beta_0)^2$$

Maximum likelihood = Least Squares.

Null hypothesis $\beta_1 = 0$, β_0 unspecified.

→ **Log likelihood difference**

$$\begin{aligned} c - \frac{1}{2\sigma^2} \sum_i (y_i - \beta_1 x_i - \beta_0)^2 - \left(c - \frac{1}{2\sigma^2} \sum_i (y_i - \beta_1 x_i - \beta_0)^2 \right) \\ = \frac{1}{2\sigma^2} \left(\sum_i (y_i - \beta_1 x_i - \beta_0)^2 - \sum_i (y_i - \beta_0)^2 \right) \end{aligned}$$

= difference of Sums of Squares (total minus residual)

= **Sums of Squares of Model**

... divided by $2\sigma^2$.

σ^2 unknown \rightarrow estimate from residuals!
multiply by 2 \rightarrow **Difference of deviances** \rightarrow **F-Test.**

(To be precise, σ is estimated under the alternative,
not under the null hypothesis...)

Also applicable for multiple regression,
more than one coefficient to be tested.

- c Same properties for deviance differences – asymptotically – in general (under conditions):

Under the null hypothesis, **the deviance difference**

(= twice the log likelihood ratio)

is distributed asymptotically $\sim \chi_{df}^2$.

Degrees of freedom **df** = number of parameters that are fixed by null hypothesis.

Only applies to “nested” models:

The reduced model is obtained by restricting the full model.

4.6 Robust Estimators

a Influence Function should be bounded!

b **Gross error sensitivity.**

$$\gamma^*(T, F) = \sup_x \langle |IF(x; T, F)| \rangle .$$

c Examples:

- Median: $\|IF(x; T, F)\| = 1/(2f(\text{med}\langle F \rangle)) = \gamma^*$
- $\gamma^*(\bar{X}, F) = \infty$.

- d **Maximal Bias.** Gross Error distribution $(1 - \tilde{\varepsilon}) F\langle \cdot \rangle + \tilde{\varepsilon} H\langle \cdot \rangle$
 \longrightarrow Gross Error “neighborhood” $U\langle F, \varepsilon \rangle$

$$b\langle \varepsilon; T, F \rangle = \sup_{G \in U\langle F, \varepsilon \rangle} \langle G\langle T \rangle \rangle .$$

- e **Breakdown point** $\varepsilon^*\langle T, F \rangle$:
 minimal “radius” ε of a neighborhood around F ,
 for which T breaks down,

$$\varepsilon^*\langle T, F \rangle = \inf_{\varepsilon} \langle b\langle \varepsilon; T, F \rangle = \infty \rangle ,$$

f **Empirical breakdown point.**

Sample x_1, x_2, \dots, x_n . Plus q arbitrary $x_1^*, x_2^*, \dots, x_q^*$.

$T\langle x_1, x_2, \dots, x_n, x_1^*, x_2^*, \dots, x_q^* \rangle - T\langle x_1, x_2, \dots, x_n \rangle$ infinite?

→ Proportion $q/(n + q)$ such that ... remains bounded.

Usually independent of x_1, x_2, \dots, x_n .

10% trimmed mean → 10%

g Importance.

h Compare to a bridge:

can vibrate under low wind – more or less strongly –

and break down under a storm – more or less violent

- i If we want minimal Gross Error Sensitivity \longrightarrow Median
 \longrightarrow asymptotic variance 1.571, $\gamma^* = 1.253$.

Compare to insurance: Premium for covering risk.

- j **Optimal compromise:** Hampel

For normal distribution with fixed variance: Huber estimators.

“Premium” of 5% more variance for $k = 1.345$.

$\gamma^* = 1.637$. The compromise pays off!

4.7 Outlook

- a Block on robust regression
- b **Multidimensional estimators.** \longrightarrow **Multivariate statistics**
- c **Small Sample Asymptotics.**