Analyzing Categorical Time Series in the Presence of Missing Observations

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

C.H. Weiß

Department of Mathematics & Statistics,
Helmut Schmidt University, Hamburg

In real applications, time series often exhibit **missing data**, so impossible to apply standard analytical tools.

For real-valued time series with "missingness", several proposals how to adapt tools, such as sample autocorrelation function (ACF) or spectral estimators, see Scheinok (1965), Bloomfield (1970), Neave (1970), Dunsmuir & Robinson (1981), Yajima & Nishino (1999).

Use idea of Parzen (1963) to understand real-valued time series with missingness as resulting from **amplitude modulation**, where amplitude-modulating process binary.

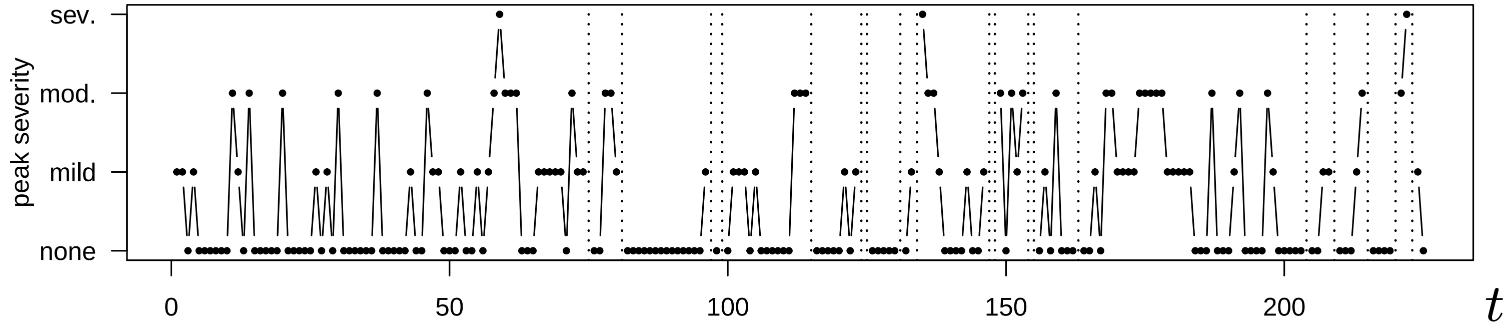
But missingness also happens to **categorical time series**, which consist of qualitative values ordered in time. Completely different analytical tools for cat. t. s. (Weiß, 2018), so aforementioned solutions for missingness not applicable.

Main practical motivation: collaborative project with Marina Vives-Mestres & Amparo Casanova (Curelator Inc.), on categorical time series from migraine patients. Daily data on migraine patients from questionaire in mobile app N1-Headache™. **Missing data** because patients skipped some questions, or stopped before completing questionaire.

Examples: (dotted lines indicate missing data)

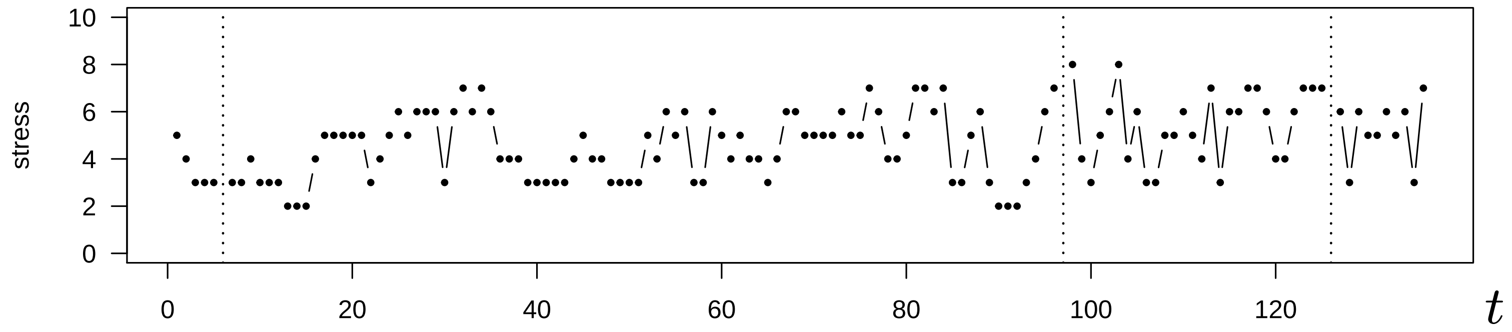
Peak severity (“none”, “mild”, “moderate”, “severe”):

daily levels for patient A; $n = 225$, 19 missing.



Stress (0–10 Likert scale, “not at all” to “a lot”):

daily levels for patient B; $n = 136$, 3 missing.



Outline & Full Paper

- Handle missing data in stationary **categorical** time series, $X_1, \ldots, X_n$ with $n \in \mathbb{N} = \{1, 2, \ldots\}$.
- Outcomes x_t of X_t have qualitative range $\mathcal{S} = \{s_0, s_1, \ldots, s_m\}$; consider both **nominal and ordinal** case.
- Unique approach of incorporating missingness and of deriving asymptotics of proposed statistics.
- Apply novel missing-data approaches to migraine time series.

Full paper: Weiß (2021) Analyzing categorical time series in the presence of missing observations. *Statistics in Medicine* **40**(21), 4675–4690. ($\rightarrow$ open access)

HELMUT SCHMIDT
UNIVERSITÄT
Universität der Bundeswehr Hamburg

Basics of
Categorical Time Series Analysis

nominal vs. ordinal

Nominal range	**Ordinal range**
Marginal PMF: (probability mass function) $\boldsymbol{p} = (p_0, \ldots, p_m)^\top \in [0;1]^{m+1}$ with $p_i = P(X = s_i)$	Marginal CDF: (cumulative distribution fct.) $\boldsymbol{f} = (f_0, \ldots, f_{m-1})^\top \in [0;1]^{m}$ with $f_i = P(X \leq s_i)$
Bivariate lag-h PMF: $p_{ij}(h) = P(X_t = s_i, X_{t-h} = s_j)$	Bivariate lag-h CDF: $f_{ij}(h) = P(X_t \leq s_i, X_{t-h} \leq s_j)$
Binarization $(\boldsymbol{Y}_t)_{\mathbb{N}}$ with $Y_{t,i} = \mathbb{1}_{\{X_t = s_i\}}$, so $E[\boldsymbol{Y}_t] = \boldsymbol{p}$	Binarization $(\boldsymbol{Z}_t)_{\mathbb{N}}$ with $Z_{t,i} = \mathbb{1}_{\{X_t \leq s_i\}}$, so $E[\boldsymbol{Z}_t] = \boldsymbol{f}$
Sample PMF: $\hat{\boldsymbol{p}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{Y}_t$	Sample CDF: $\hat{\boldsymbol{f}} = \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{Z}_t$
Bivariate (cumulative) relative frequencies	
$\hat{p}_{ij}(h) = \frac{1}{T-h} \sum_{t=h+1}^{T} Y_{t,i}\, Y_{t-h,j}$	$\hat{f}_{ij}(h) = \frac{1}{T-h} \sum_{t=h+1}^{T} Z_{t,i}\, Z_{t-h,j}$

Note that both **binarizations** lead to different range:

$$\boldsymbol{Y}_t \in \left\{ \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}, \begin{pmatrix}0\\1\\\vdots\\0\end{pmatrix}, \ldots, \begin{pmatrix}0\\\vdots\\1\\0\end{pmatrix}, \begin{pmatrix}0\\\vdots\\0\\1\end{pmatrix} \right\} =: \{\boldsymbol{e}_0, \ldots, \boldsymbol{e}_m\} \subset [0;1]^{m+1}$$

vs.

$$\boldsymbol{Z}_t \in \left\{ \begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}, \begin{pmatrix}0\\1\\\vdots\\1\end{pmatrix}, \ldots, \begin{pmatrix}0\\\vdots\\0\\1\end{pmatrix}, \begin{pmatrix}0\\\vdots\\0\\0\end{pmatrix} \right\} =: \{\boldsymbol{c}_0, \ldots, \boldsymbol{c}_m\} \subset [0;1]^{m}.$$

Dispersion concepts for qualitative random variables:

- Minimal dispersion iff one-point distribution,

  i. e., $\boldsymbol{p} \in \{\boldsymbol{e}_0, \ldots, \boldsymbol{e}_m\}$ or $\boldsymbol{f} \in \{\boldsymbol{c}_0, \ldots, \boldsymbol{c}_m\}$, respectively.
- Max. nominal dispersion iff uniform, $\boldsymbol{p} = (\frac{1}{m+1}, \ldots, \frac{1}{m+1})^\top$.
- Max. ordinal disp. iff extreme two-point, $\boldsymbol{f} = (\frac{1}{2}, \ldots, \frac{1}{2})^\top$.

Nominal range	**Ordinal range**
Index of qualitative variation: $\mathrm{IQV} \;=\; \frac{m+1}{m}\big(1-\sum_{i=0}^{m} p_i^2\big)$	Index of ordinal variation: $\mathrm{IOV} \;=\; \frac{4}{m}\,\sum_{i=0}^{m-1} f_i(1-f_i)$
	Ordinal skewness: $\mathrm{skew} = \frac{2}{m}\,\sum_{i=0}^{m-1} f_i \;\; - 1$
Nominal Cohen's κ: $\kappa_{\mathrm{nom}}(h) \;=\; \dfrac{\sum_{j=0}^{m}\big(p_{jj}(h)-p_j^2\big)}{1-\sum_{i=0}^{m} p_i^2}$	Ordinal Cohen's κ: $\kappa_{\mathrm{ord}}(h) \;=\; \dfrac{\sum_{j=0}^{m-1}\big(f_{jj}(h)-f_j^2\big)}{\sum_{i=0}^{m-1} f_i(1-f_i)}$

The signed κ-measures serve as substitutes of ACF, where positive (negative) values express extend of (dis)agreement between X_t and X_{t-h}, see Weiß (2018, 2020).

Aforementioned approaches assume time series to be fully observed!

Now, categorical time series **with missing observations**, unified framework for handling missingness in *both ordinal and nominal* time series.

Time restrictions & clarity of talk:

focus on ordinal case, but nominal case in . . .

Full paper: Weiß (2021) Analyzing categorical time series in the presence of missing observations. *Statistics in Medicine* **40**(21), 4675–4690. (→ open access)

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH
STAT

Ordinal Time Series with Missing Data

Amplitude Modulation

Let (X_t) be ordinal process, corresponding binarization $(\boldsymbol{Z}_t)$.

Idea: Adapt amplitude modulation of Parzen (1963) to binarization $(\boldsymbol{Z}_t)$.

Define **amplitude-modulating process** (O_t) as
$O_t = 1$ if X_t observed, and $O_t = 0$ otherwise.
Then, **amplitude modulation** of $(\boldsymbol{Z}_t)$ is $(O_t \cdot \boldsymbol{Z}_t)$

(O_t) might be deterministic or i. i. d. with $E[O_0] = \pi$ or stationary with some dependence structure, $E[O_h\, O_0] = \pi(h)$.

But we assume that (O_t) is independent of (X_t), $(\boldsymbol{Z}_t)$.

Estimation of marginal CDF $\boldsymbol{f}$:

Let $\hat{\mathfrak{f}} := \frac{1}{n}\sum_{t=1}^{n} O_t\,\boldsymbol{Z}_t$, then $\quad E\big[\hat{\mathfrak{f}}\big] \;=\; \big(\frac{1}{n}\sum_{t=1}^{n} E[O_t]\big)\,\boldsymbol{f}$.

Thus, estimate $\boldsymbol{f}$ by

$$\widehat{\boldsymbol{f}}^* \;:=\; \frac{\frac{1}{n}\sum_{t=1}^{n} O_t\,\boldsymbol{Z}_t}{\frac{1}{n}\sum_{t=1}^{n} O_t} \;=:\; \hat{\mathfrak{f}}\,\big/\,\overline{O}.$$

Estimators for IOV and skew by using $\widehat{\boldsymbol{f}}^*$:

$$\widehat{\mathrm{IOV}} \;=\; \frac{4}{m}\sum_{i=0}^{m-1}\widehat{f}_i^*(1-\widehat{f}_i^*), \qquad \widehat{\mathrm{skew}} \;=\; \frac{2}{m}\sum_{i=0}^{m-1}\widehat{f}_i^* \;-1.$$

Estimation of bivariate CDF $f_{ij}(h)$:

Let $\quad \hat{\mathfrak{f}}_{ij}(h) \;=\; \frac{1}{n-h}\sum_{t=h+1}^{n} O_t\, O_{t-h}\, Z_{t,i}\, Z_{t-h,j},$

then $\quad E\big[\hat{\mathfrak{f}}_{ij}(h)\big] \;=\; \big(\frac{1}{n-h}\sum_{t=h+1}^{n} E[O_t\, O_{t-h}]\big)\, f_{ij}(h).$

Thus, estimate $f_{ij}(h)$ by

$$\hat{f}^*_{ij}(h) \;=\; \frac{\frac{1}{n-h}\sum_{t=h+1}^{n} O_t\, O_{t-h}\, Z_{t,i}\, Z_{t-h,j}}{\frac{1}{n-h}\sum_{t=h+1}^{n} O_t\, O_{t-h}} \;=:\; \hat{\mathfrak{f}}_{ij}(h)\,\big/\,\overline{O_t O_{t-h}}.$$

Estimator for ordinal κ by using $\hat{\boldsymbol{f}}^*, \hat{f}^*_{ij}(h)$:

$$\hat{\kappa}_{\text{ord}}(h) \;=\; \frac{\sum_{j=0}^{m-1}\big(\hat{f}^*_{jj}(h) - (\hat{f}^*_j)^2\big)}{\sum_{i=0}^{m-1}\hat{f}^*_i(1-\hat{f}^*_i)} \qquad \text{for } h \in \mathbb{N}.$$

HELMUT SCHMIDT UNIVERSITÄT
Universität der Bundeswehr Hamburg

MATH STAT

Ordinal Time Series with Missing Data

Asymptotics

Theorem 1: Let (X_t) and (O_t) be α-mixing with exponentially decreasing weights.

Then, the corrected CDF-estimator $\hat{\boldsymbol{f}}^*$ satisfies

$$\sqrt{n}\left(\hat{\boldsymbol{f}}^* - \boldsymbol{f}\right) \stackrel{\mathrm{d}}{\to} \mathrm{N}\left(\mathbf{0},\ \boldsymbol{\Sigma}^*\right), \quad \text{with } \boldsymbol{\Sigma}^* = (\sigma_{ij}^*)_{i,j=0,\ldots,m-1}, \quad \text{where}$$

$$\sigma_{ij}^* \ = \ \tfrac{1}{\pi}\left(f_{\min\{i,j\}} - f_i f_j\right) \ + \ \tfrac{1}{\pi^2}\textstyle\sum_{h=1}^{\infty} \pi(h)\left(f_{ij}(h) + f_{ji}(h) - 2\, f_i f_j\right).$$

Furthermore, the bias $E[\hat{f}_j^*] - f_j$ is of order $o(n^{-1})$.

Theorem 2: Let assumptions of Theorem 1 hold. Then, corrected sample IOV and skew asymptotically normal with

$$E\big[\widehat{\mathrm{IOV}}\big] \;\approx\; \mathrm{IOV}\Big(1 \;-\; \tfrac{1}{n}\big(\tfrac{1}{\pi}+\tfrac{2}{\pi^2}\textstyle\sum_{h=1}^{\infty}\pi(h)\,\kappa_{\mathrm{ord}}(h)\big)\Big),$$

$$V\big[\widehat{\mathrm{IOV}}\big] \;\approx\; \frac{1}{n}\,\frac{16}{m^2}\sum_{i,j=0}^{m-1}(1-2f_i)(1-2f_j)\,\sigma^*_{ij},$$

and

$$E\big[\widehat{\mathrm{skew}}\big] \;\approx\; \mathrm{skew} \;+\; o(n^{-1}), \qquad V\big[\widehat{\mathrm{skew}}\big] \;\approx\; \tfrac{1}{n}\,\tfrac{4}{m^2}\textstyle\sum_{i,j=0}^{m-1}\sigma^*_{ij}.$$

Theorem 3: Let assumptions of Theorem 1 hold. Denote

$$\pi(h_1,\ldots,h_r) := E\big[O_0 \cdot O_{h_1} \cdots O_{h_r}\big] \quad \text{with } 0 < h_1 < \ldots < h_r.$$

Under null hypothesis of (X_t) being i. i. d.,

distribution of $\hat{\kappa}_{\mathrm{ord}}(h)$ at lag $h \in \mathbb{N}$ approximately normal

with mean $\quad -\dfrac{1}{n\pi}$

and variance

$$\frac{1}{n}\,\frac{\sum_{i,j=0}^{m-1}\Big(f_{\min\{i,j\}} - f_i f_j\Big)\Big(f_{\min\{i,j\}} - f_i f_j \;+\; 2\Big(1 + \frac{\pi(h,2h)}{\pi(h)} - 2\,\frac{\pi(h)}{\pi}\Big)\, f_i f_j\Big)}{\pi(h)\,\Big(\sum_{k=0}^{m-1} f_k(1-f_k)\Big)^2}.$$

Possible applications: confidence intervals, hypothesis tests.

Example: Theorem 3 to test for serial independence at lag h.

Let (O_t) be i. i. d. ("missing at random"), then simplification

$$V\Big[\hat{\kappa}_{\text{ord}}(h)\Big] \approx$$

$$\frac{1}{n}\,\frac{1}{\pi^2}\,\frac{\sum_{i,j=0}^{m-1}\big(f_{\min\{i,j\}}-f_if_j\big)^2}{\big(\sum_{k=0}^{m-1} f_k(1-f_k)\big)^2} + \frac{2}{n}\,\frac{1-\pi}{\pi^2}\,\frac{\sum_{i,j=0}^{m-1} f_if_j\big(f_{\min\{i,j\}}-f_if_j\big)}{\big(\sum_{k=0}^{m-1} f_k(1-f_k)\big)^2}.$$

Plug-in estimated probabilities $\hat{\pi}, \hat{\boldsymbol{f}}^*$.

Then **critical values** $-1/(n\hat{\pi}) \mp z_{1-\alpha/2}\,\hat{\sigma}_\kappa$,

where $z_{1-\alpha/2}$ denotes $(1-\alpha/2)$-quantile of $\mathrm{N}(0,1)$.

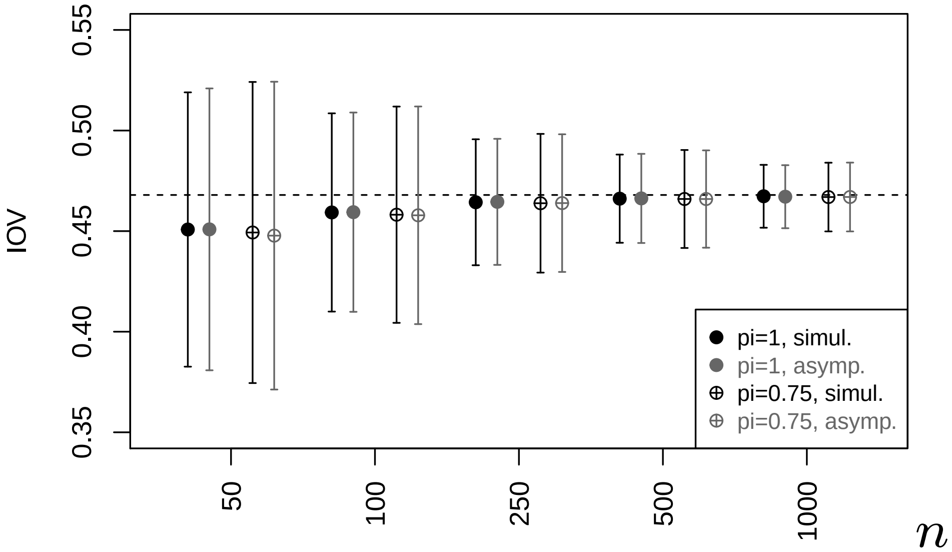
Empirical Illustrations

Simulations & Data Example

Simulation study in full paper, Weiß (2021), confirms good finite-sample performance of asymptotic approximations.

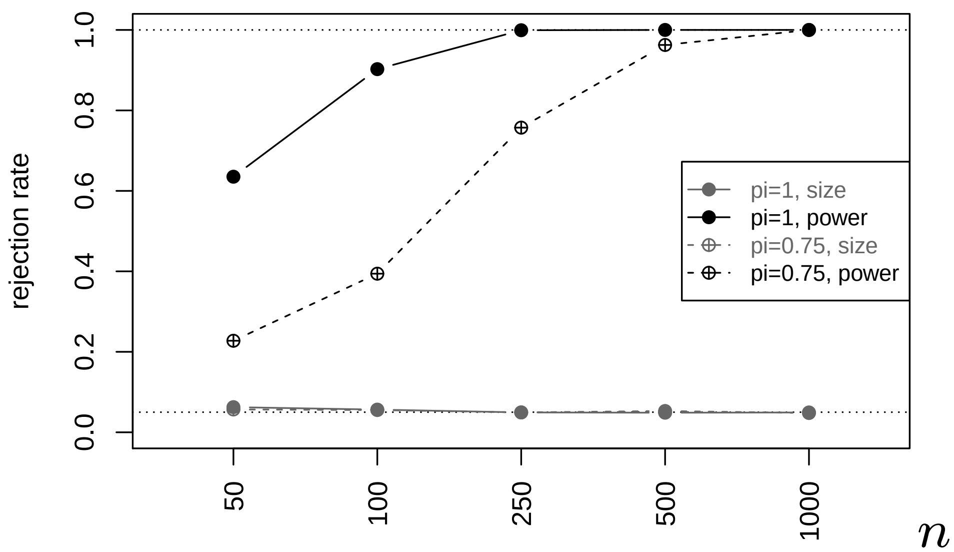
Illustrative examples for models . . .

$(m,p,r) = (3, 0.20, 0.35)$



Mean $\mp$ std. dev. for IOV

$(m,p,r) = (3, 0.20, 0 \text{ vs. } 0.35)$

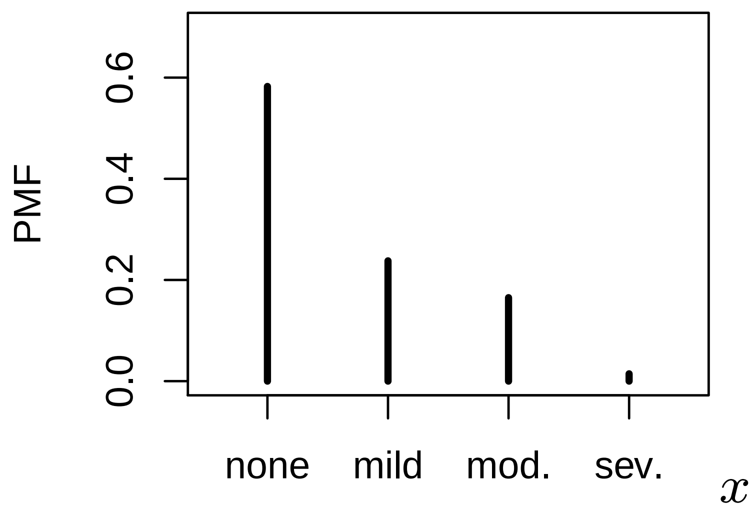


Rej. rate based on $\widehat{\kappa}_{\text{ord}}(1)$

Data application: migraine patients

Estimates and tests *accounting for missingness* for

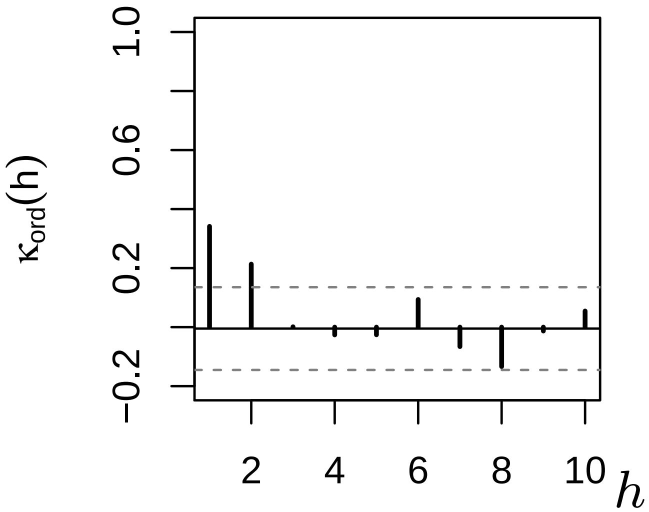
peak-severity time series of patient A ($n = 225$, 19 missing):



Median:
“none”

$\widehat{\text{IOV}}$
≈ 0.540

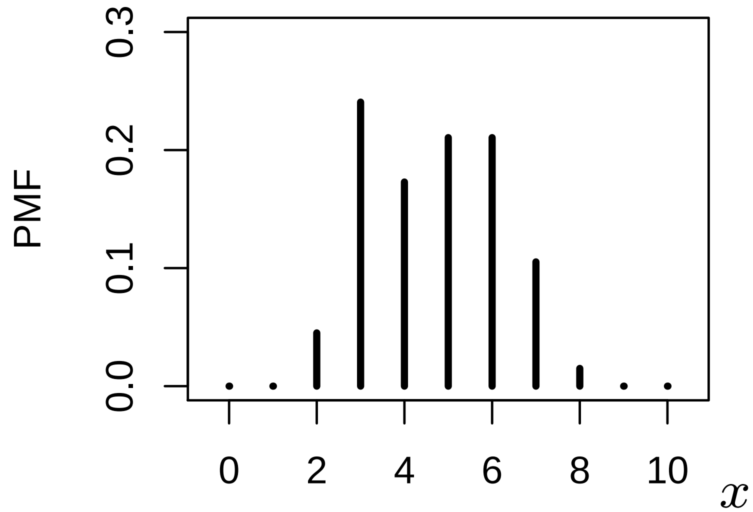
$\widehat{\text{skew}}$
≈ 0.592



MATH STAT

Data application: migraine patients

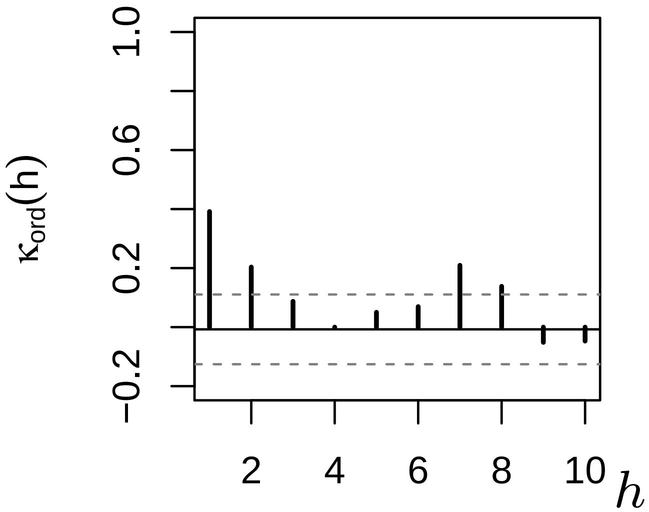
Estimates and tests *accounting for missingness* for **stress** time series of patient B ($n = 136$, 3 missing):



Median: 5

$\widehat{\mathrm{IOV}} \approx 0.335$

$\widehat{\mathrm{skew}} \approx 0.065$



What would happen if just ignore (skip) missing data?

Then complete time series of reduced lengths

$\tilde{n} = 206$ (peak severity) and $\tilde{n} = 133$ (stress).

$\widehat{\kappa}_{\text{ord}}(h)$ changes from $0.341, 0.213, \ldots$ to $0.258, 0.138, \ldots$
(approximate SE from 0.072 to 0.055)

for peak-severity series (19 out of 225),

and from $0.392, 0.203, \ldots$ to $0.370, 0.195, \ldots$
(approximate SE from 0.060 to 0.057)

for stress series (3 out of 136).

$\Rightarrow$ **carefully consider missingness** for time series analysis!

Thank You

for Your Interest!

Christian H. Weiß

Department of Mathematics & Statistics

Helmut Schmidt University, Hamburg

`weissc@hsu-hh.de`

Literature

Weiß (2021) Analyzing categorical time series in the presence of missing observations. *Statistics in Medicine* 40(21), 4675–4690.
($\rightarrow$ open access)

Bloomfield (1970) Spectral analysis with . . . *JRSS B* **32**, 369–380.

Dunsmuir & Robinson (1981) Asymptotic theo. . . *Sankhyā A* **43**, 260–281.

Neave (1970) Spectral analysis of . . . *Biometrika* **57**, 111–122.

Parzen (1963) On spectral analysis with . . . *Sankhyā A* **25**, 383–392.

Scheinok (1965) Spectral analysis with . . . *AMS* **36**, 971–977.

Vives-Mestres & Casanova (2021) Modelling . . . *Stat Med* **40**, 213–225.

Weiß (2018) *An Introduction to Discrete-Valued Time Series*. Wiley.

Weiß (2020) Distance-based analysis . . . *JASA* **115**, 1189–1200.

Yajima & Nishino (1999) Estimation of the . . . *Sankhyā A* **61**, 189–207.