

Parametric estimation of prevalence and incidence based on the waiting time distribution

Henrik Støvring^(1,2)

Werner Vach⁽¹⁾

⁽¹⁾Department of Statistics and Demography

⁽²⁾Research Unit of General Practice

University of Southern Denmark

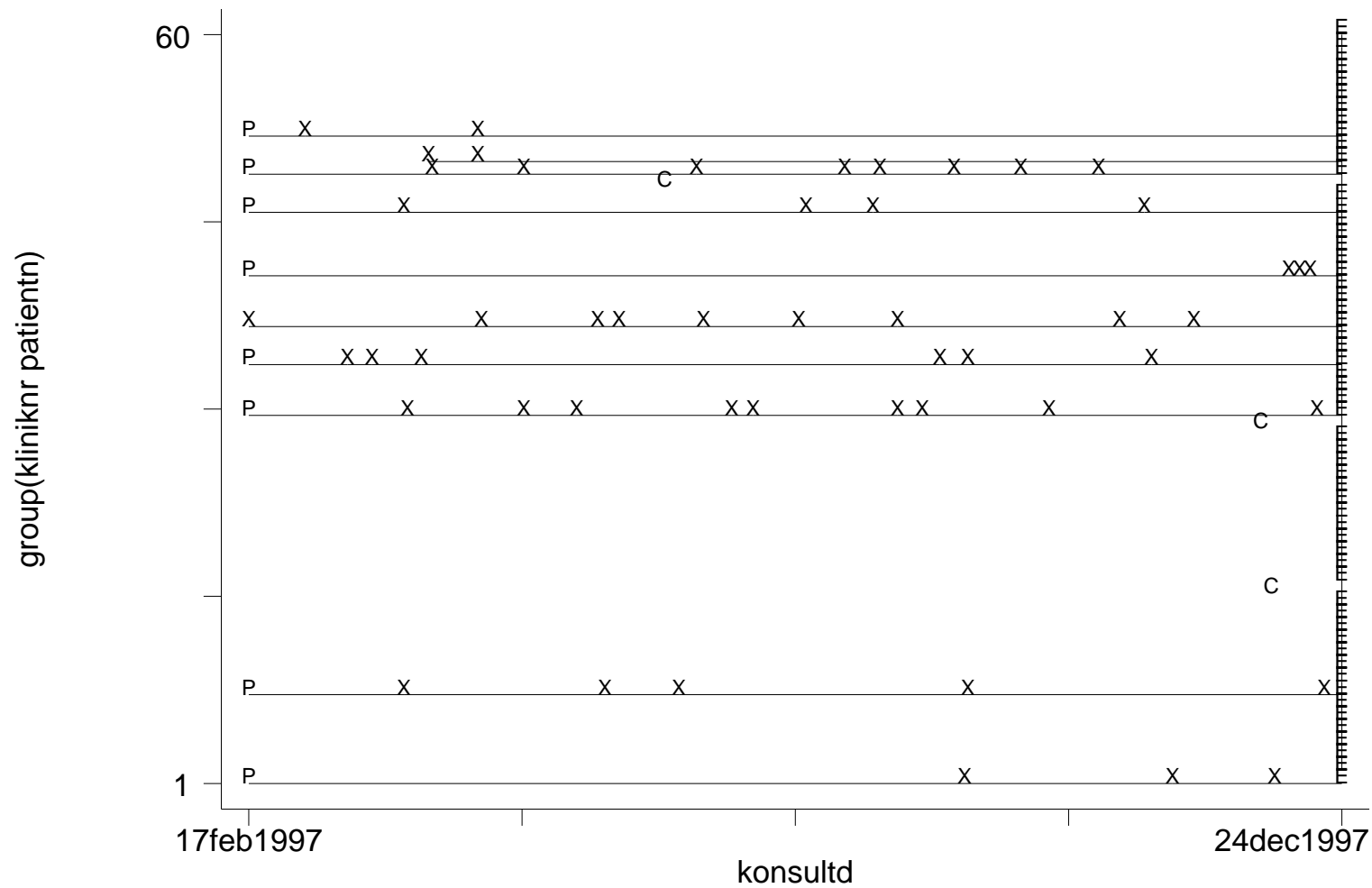
July 2, 2001

1 Outline

- Problem/Objective
- Model
- Application and diagnostics
- Outline of simulation results

2 Introduction

- Basic epidemiologic measures
 1. Prevalence proportion
 2. Incidence intensity
- Traditional estimates are obtained by
 1. Cross sectional: Assessment of individual disease status at time t_0
 2. Follow up: Assessment of individual change from non-diseased to diseased
- **But:** Costly to obtain information on individual disease history



2.1 The challenge

- Databases with health-related events are available, often population wide
- Examples
 - Redemptions of drug prescriptions
 - Diagnose coded contacts with a general practitioner (GP)
- Problem
 - The individual disease/treatment status is *not* observed between events
- Challenge
 - estimate prevalence and incidence based on these databases
 - should be based on general principles

3 Basic definitions

- Consider events in the fixed time window $[0; \delta)$.
- No late entry
- Information on censoring is available
- Basic individual stochastic variables

X : Initial disease or treatment status (unobserved/latent)

T : Time of first event in time window

Z : Exit time

4 The idea of Hallas et al.

- The distribution of T is a two-component mixture density (Hallas et al. 1997):

$$f^T(t) = p \cdot f^{T|X=1}(t) + (1 - p)f^{T|X=0}(t) \quad (1)$$

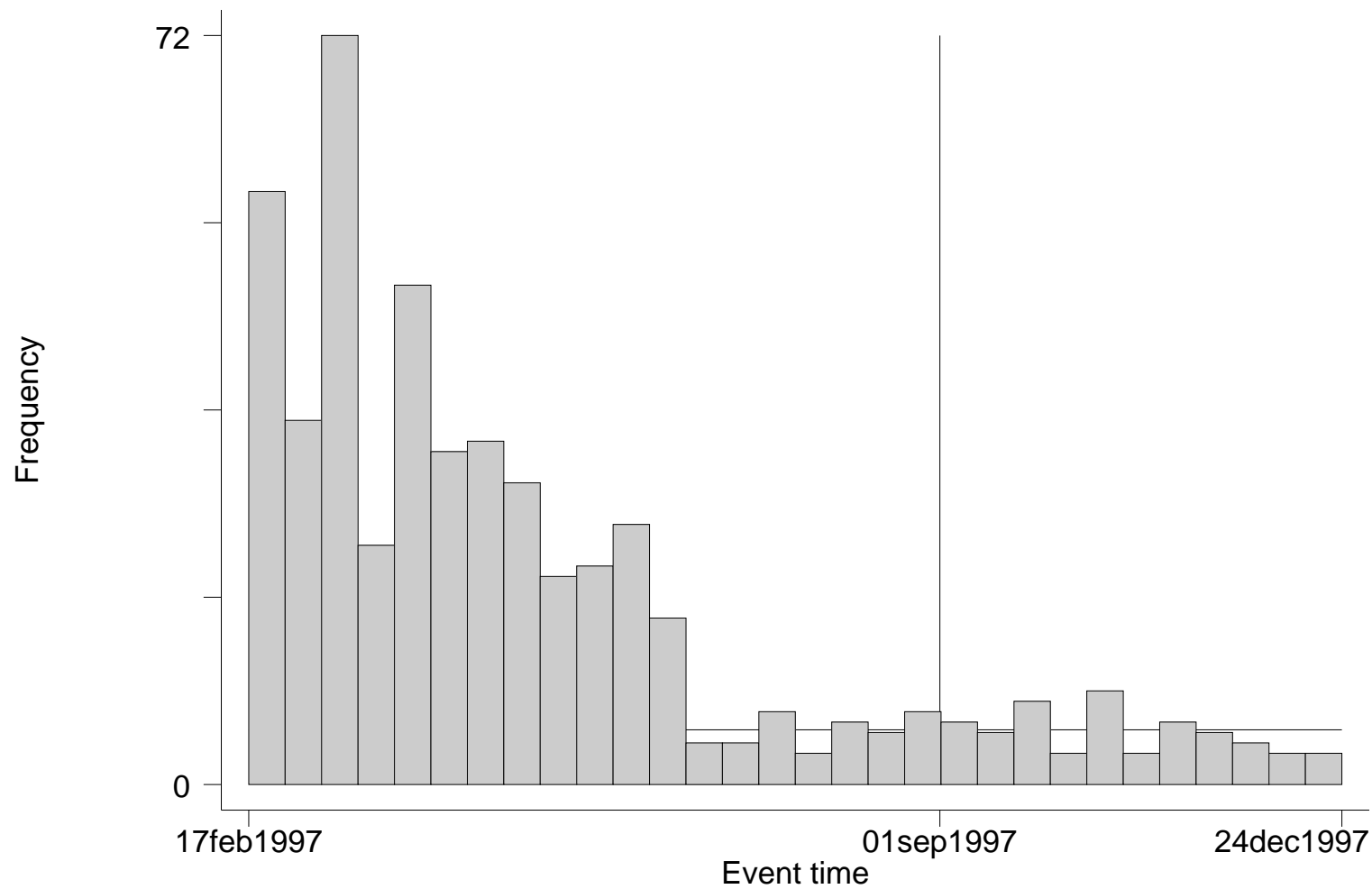
where $p = P(X = 1|T < \delta)$

- The prevalence proportion at 0 is given by

$$\text{Prevalence} = p \frac{n}{N} \quad (2)$$

where $\frac{n}{N}$ is the proportion of observed subjects

- Extension of the model
 - incorporation of censoring
 - appropriate parametric choices for the distributions



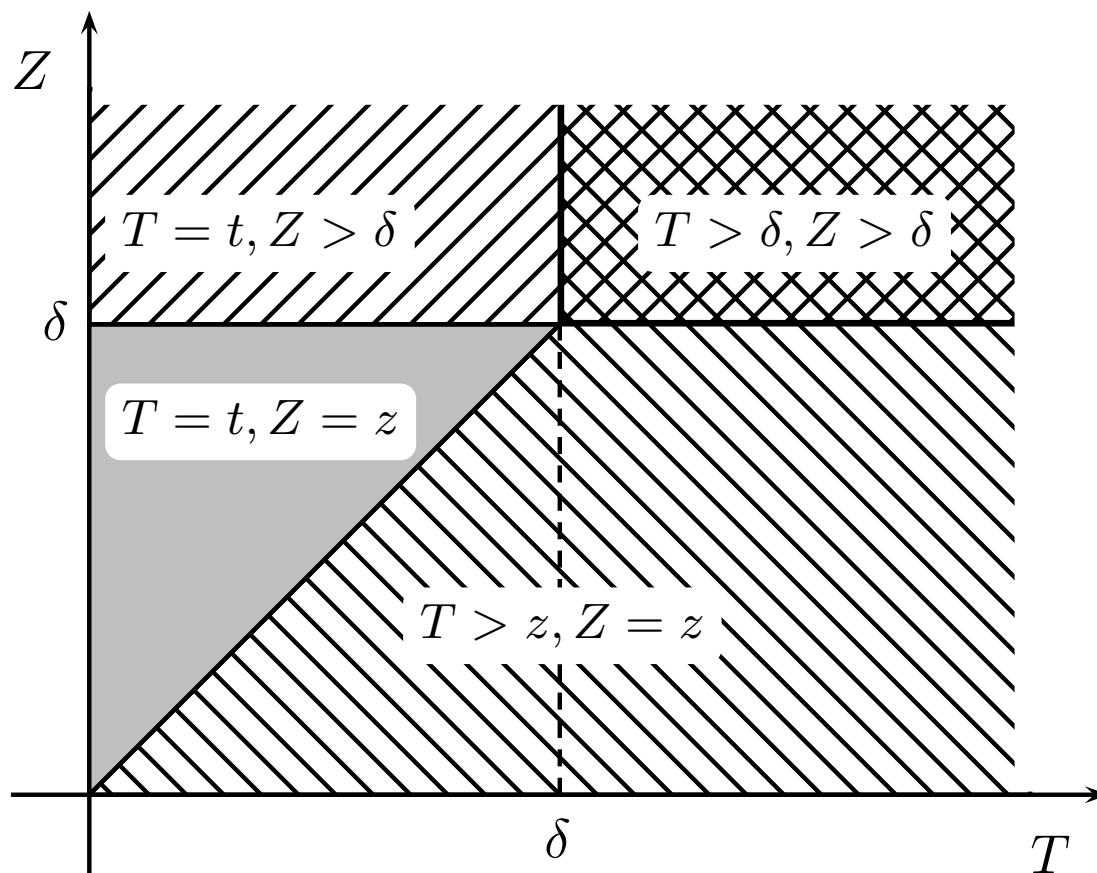
5 Joint model for waiting time and exit time

- In principle we can make a bivariate model for (T, Z)
- The natural starting point is the model

$$f^{(T,Z)} = p f^{(T,Z)|X=1} + (1 - p) f^{(T,Z)|X=0} \quad (3)$$

- Most subjects will contribute incomplete observations

- There are four types of contributions to the likelihood:



5.1 Dependence structure of (T, Z)

- For the prevalent: Independence of T and Z
- For the non-prevalent:
Likely with an underlying frailty influencing both T and Z
- Assumptions for the non-prevalent:
 - T and Z are independent given that both are smaller than δ
 - T is independent of the event $Z < \delta$ given that $T < \delta$
 - Z is independent of the event $T < \delta$ given that $Z < \delta$
- In other words
 - Globally T and Z are dependent
 - Locally they are independent

5.2 Example of likelihood contribution

- Consider the observation $T = t, Z = z$
- The likelihood contribution has the structure

$$l(\theta) = p \cdot g(t) P(Z \leq \delta | X = 1) d_1(z) + (1 - p) P(T \leq \delta, Z \leq \delta | X = 0) h(t) d_0(z) \quad (4)$$

where

- d_1 and d_0 are the densities of $Z|X, Z \leq \delta$,
- g and h are the densities of $T|X = 1$ and $T|X = 0, T \leq \delta$
- In the remaining contributions we integrate over the missing data

5.3 Choice of distributions

- All conditional densities are chosen as $U(0; \delta)$
- g is a forward recurrence density, since:
 - Events of a single prevalent subject are a renewal process
 - We hit an interval at random, ie. length biased sampling
 - Hit time is uniformly distributed on the hit interval
 - Suppose the interarrival density is γ , then

$$g(t; \theta) = \frac{\int_t^\infty \gamma(s; \theta) ds}{\int_0^\infty \gamma(s; \theta) ds} \quad (5)$$

$$= \frac{S_\gamma(t; \theta)}{\mu} \quad (6)$$

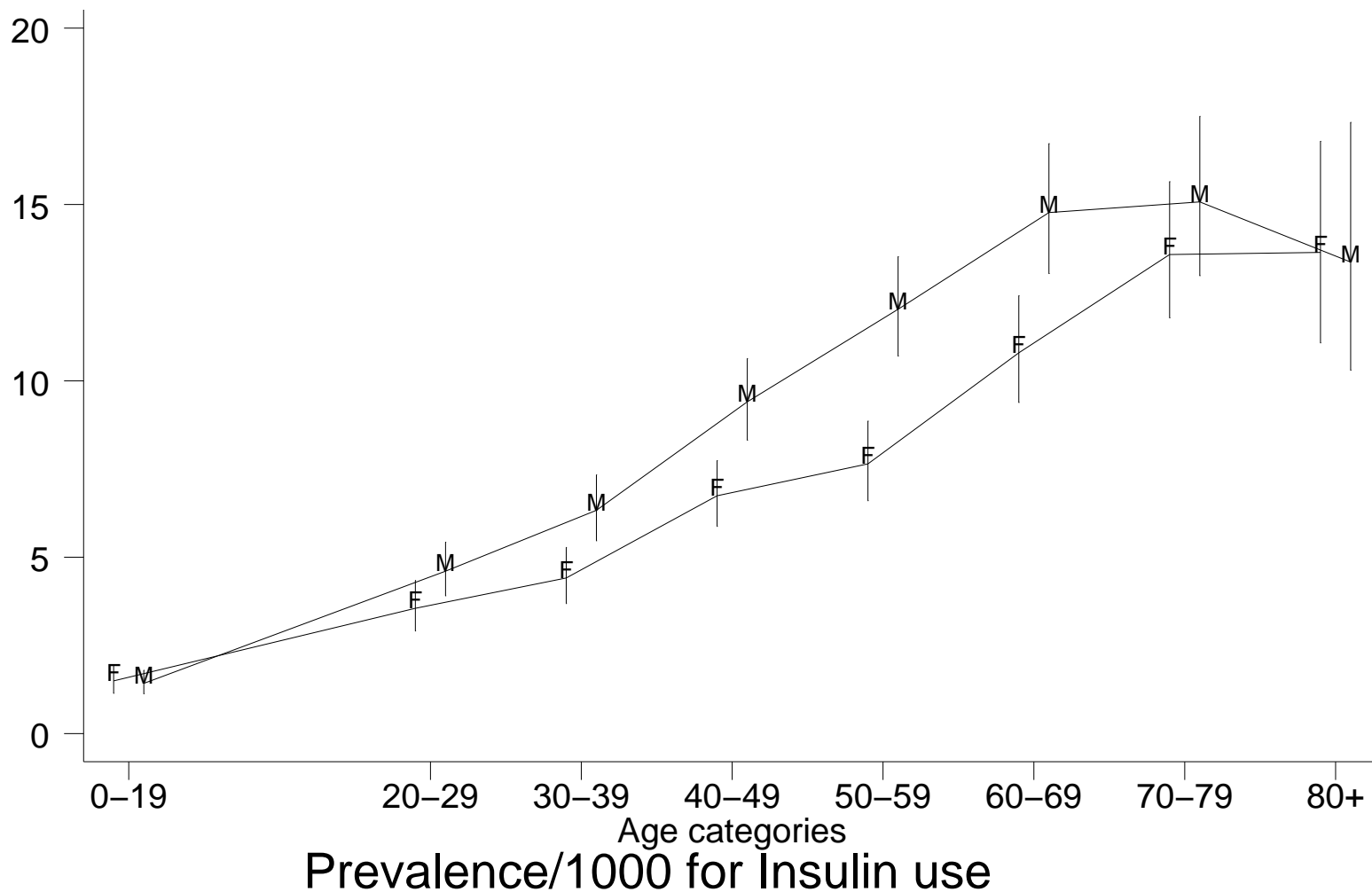
- We have worked with the Exponential, Weibull and Log-Normal as choices for γ

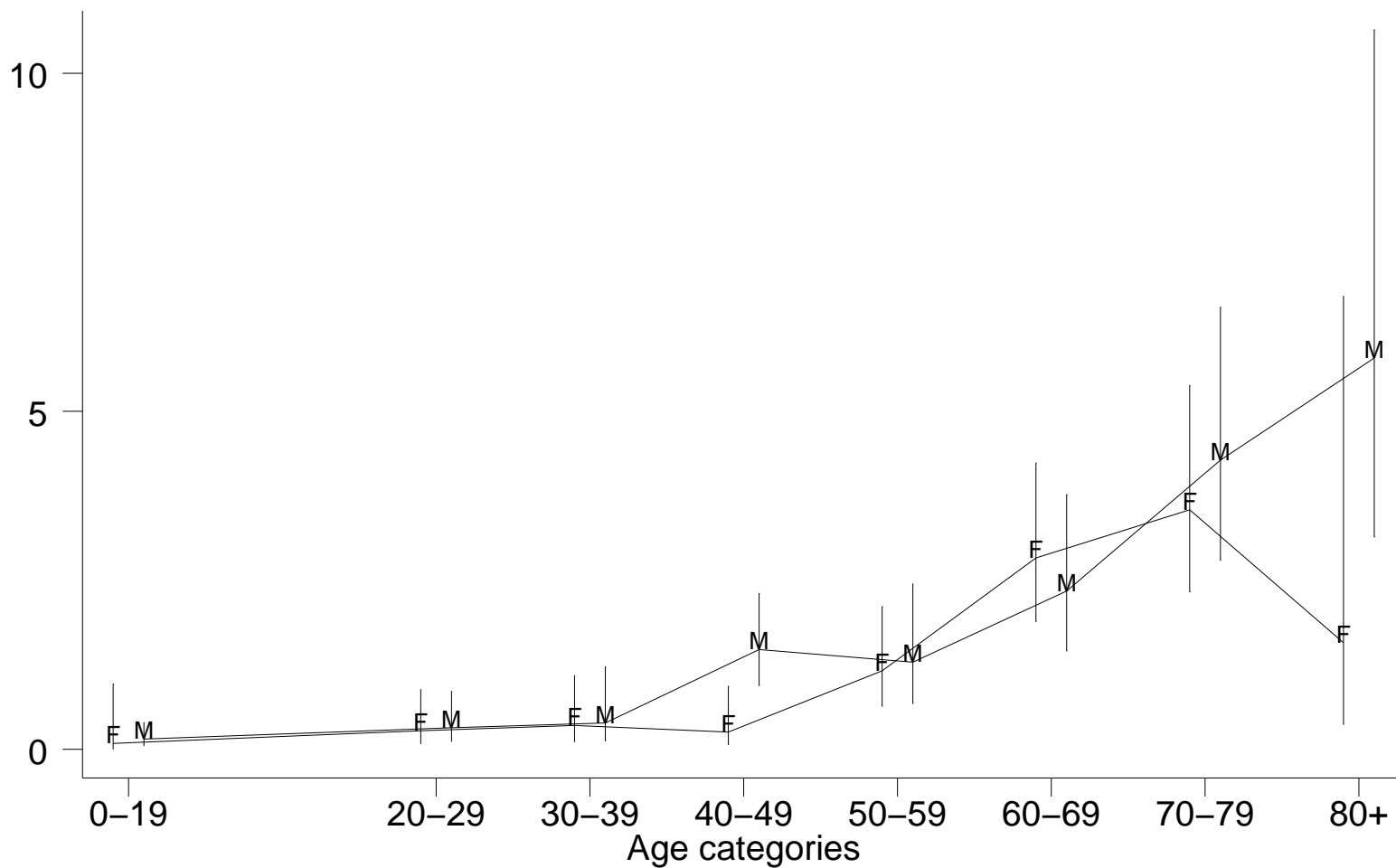
5.4 How to compute prevalence and incidence

- Prevalence proportion = p
- Incidence intensity = $P(T < \delta | X = 0)h(0; \theta)$
- These parameters can be estimated by Maximum Likelihood, which involves
 - reparametrization
 - numerical integration (stratified MC, antithetic)
 - numerical maximization (Stata 7.0, Gould and Sribney 1999)

6 Applications

- Based on Odense University Pharmacoepidemiologic Database (OPED)
- Covers the island Funen
- All Insulin prescriptions in October 1, 1997 – September 30, 1998
- Population size at baseline: 471,265
- 3684 subjects with events





Incidence Intensity/1000 pyr for Insulin use

6.1 Results

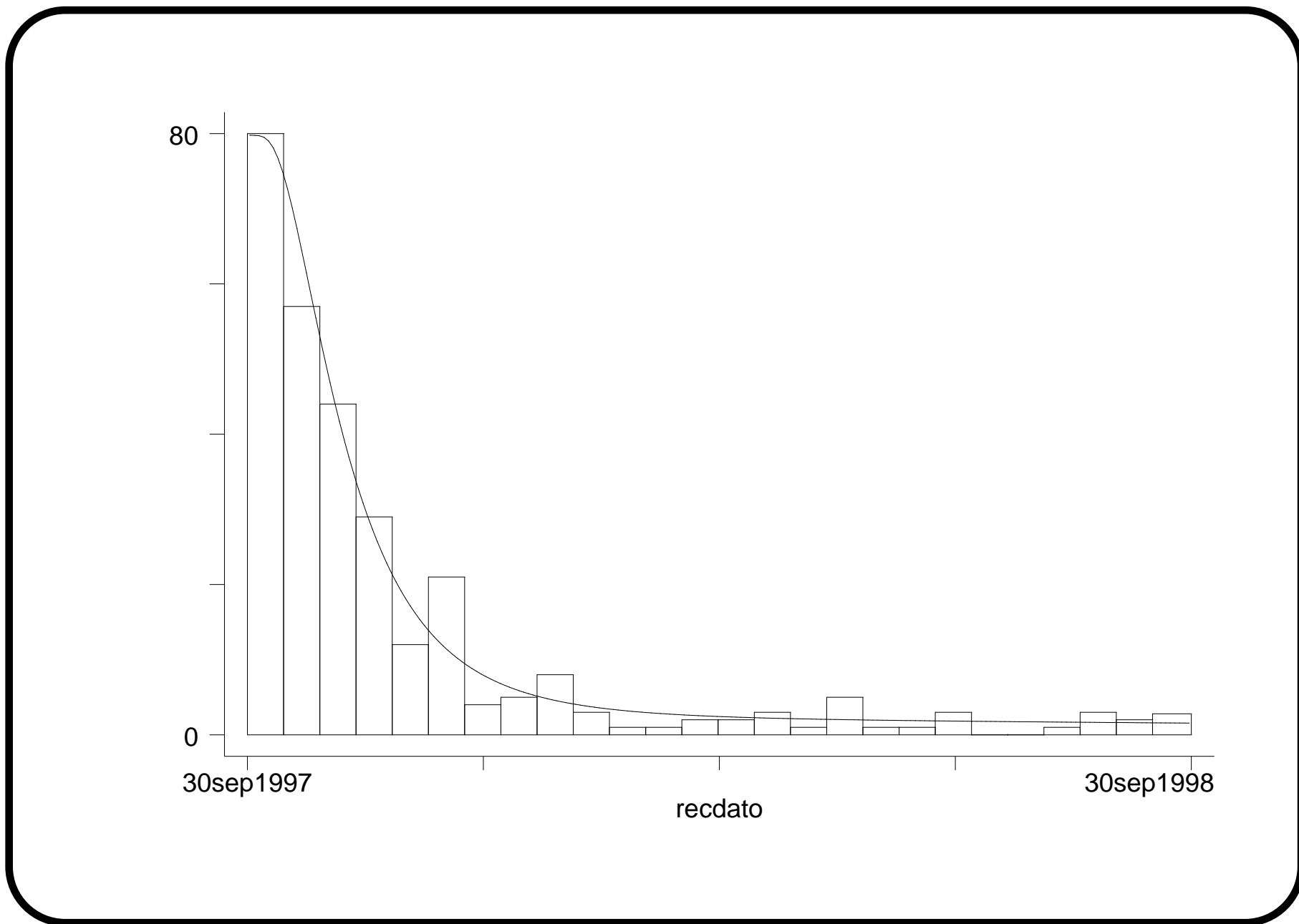
- Incidence increases with age
- Prevalence increases with age below age 70, then it levels off
- Sensible, since mortality is much higher among diabetics
- Higher incidence and prevalence among males

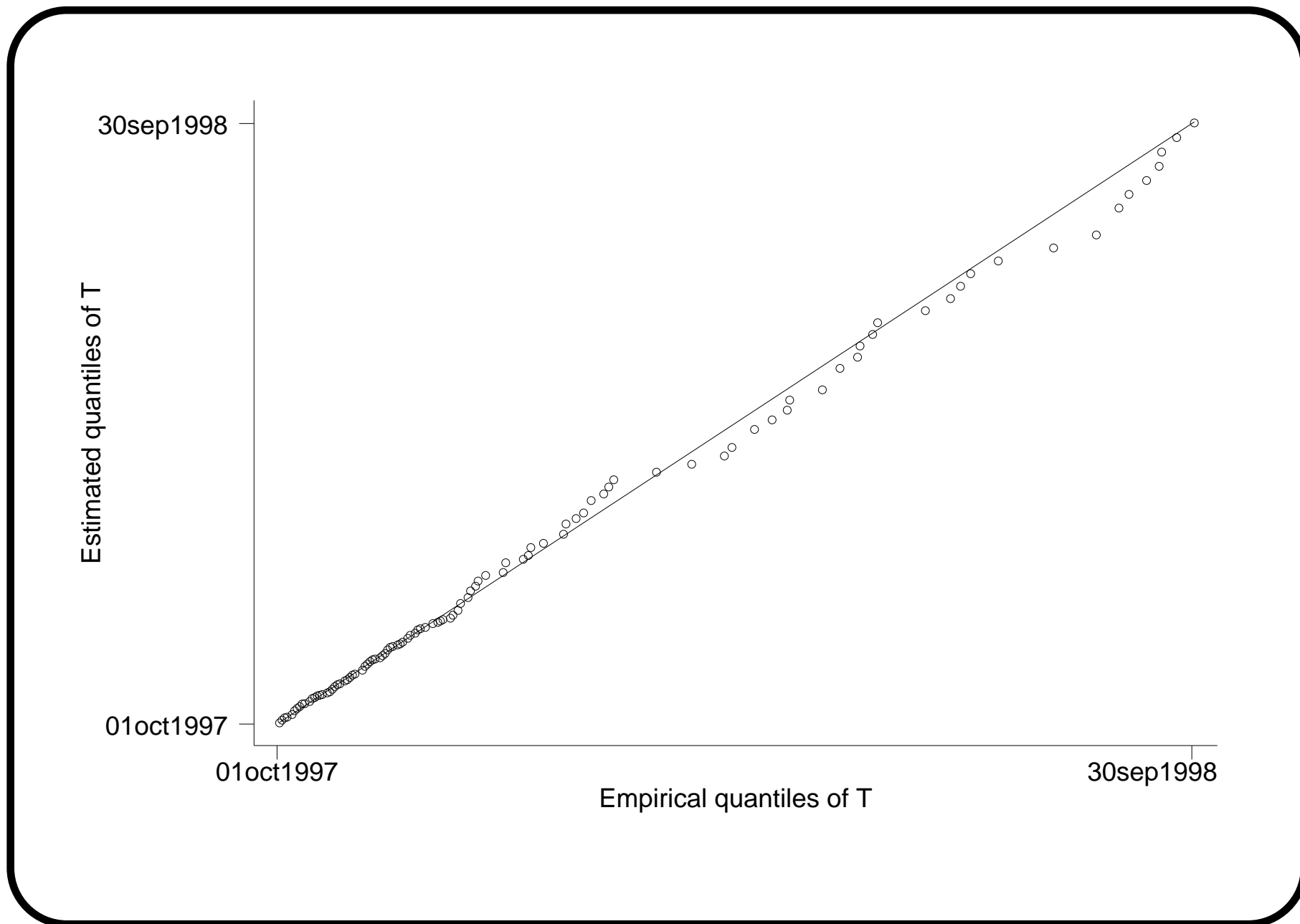
7 Diagnostic plots

- Mainly interested in the fit of the models to the waiting times
- From model it is possible to compute

$$f(t|T \leq Z, T \leq \delta) = \frac{\int_t^\infty f(t, z) dz}{\int_0^\delta \int_s^\infty f(t, z) dz ds} 1(t \leq \delta) \quad (7)$$

- Graph this against empirical distribution of observed T 's
- Alternatively make a Q-Q-plot
- Example is for Insulin, females aged 60-69





8 Outline of simulation results

- For correctly specified models
 - Bias is negligible
 - In comparison to studies with complete disease history, doubling the sample sizes yields similar precision (prevalence 1.5, incidence 2.5)
 - Nominal levels of confidence intervals are maintained
 - Very small impact of censoring levels
 - Caution should be exercised when $\hat{P}(T > \delta | X = 1) > .25\%$
- For misspecified models
 - Bias increases with level of misspecification
 - Precision decreases with level of misspecification
 - Even in extreme situations, CP of 95%-CIs remains approximately 90%

9 Summary

- Formulated an intuitive idea using general principles
- Obtained model, which
 - incorporates censoring
 - allows ML estimation
 - allows construction of diagnostic procedures
- Thus, 'cheap' estimates of incidence and prevalence are available
- Perspectives
 - Non- or semiparametric modelling
 - Use individual estimate of status as predictor for other outcomes (mortality)

References

Gould, W. and W. Sribney (1999). *Maximum Likelihood Estimation with Stata*.
College Station, Texas: Stata Press.

Hallas, J., D. Gaist, and L. Bjerrum (1997). The waiting time distribution as a
graphical approach to epidemiologic measures of drug utilization.
Epidemiology 8, 666–70.

Støvring, H. and W. Vach (2001). Estimation of prevalence and incidence in
pharmacoepidemiological and other health-related databases. *Submitted*.