Selection bias in naturalistic driving studies

Jenny Jonasson, Astra-Zeneca, Gothenburg
Holger Rootzén, Mathematical Sciences, Chalmers
Dmitrii Zholud, Mathematical Sciences, Chalmers
http://www.math.chalmers.se/~rootzen/

SAFER
VEHICLE AND TRAFFIC SAFETY CENTRE AT CHALMERS

Traffic accidents

- 1.3 miljon deaths/year worldwide, 20-50 million severely injured
- Large economic losses
- Less than 1 death/day in Sweden now. Down from 3 deaths/day a few decades ago – at a time with much less traffic
- First simple measures: seatbelts, helmets, follow traffic rules, drunk driving laws, …, then more sophisticated ones: rebuild roads, better tires, improve driver education, airbags, …, then next level of sophistication: more driver training and retraining, ABS, ESP, …, and ??

New and exciting area for statistics

Part A of talk: **Selection bias**

Part B of talk: **Visual behavior/censoring**

Future: **Risk estimation**

Active safety systems for next generation cars. Important for competition with other car makers and for safety (?)

Driver training, traffic laws, car design regulation

…

Naturalistic Driving Research

- *In Situ* investigation of driver performance
  - Use an instrumented vehicle
  - No experimenter or instructions
  - Data continuously collected for extended period

100-car study

- 100 cars, appr 250 drivers, appr 1 year
- Five video cameras, radar sensors; front, rear (for all 100 cars) and each side (for 20 cars), vision-based lane tracker, glare detectors, GPS, accelerometer
- **Still not enough crashes (82) –> try to use near-crashes (761) to learn about crash behavior**

Virginia Tech Transportation Institute
Driving Transportation with Technology

100-Car Naturalistic Driving Study Fact Sheet

Setting up the Study

Study Sponsors
- National Highway Traffic Safety Administration (NHTSA)
- Virginia Department of Transportation (VDOT)
- Virginia Transportation Research Council (VTRC)
- Virginia Tech (VT)

Study Parameters
- 109 primary drivers, 241 total drivers
- Northern Virginia/Metropolitan Washington, DC area
- 12 – 13 months of data collection
- Drivers' ages ranging from 18 to 73; 60% male; 40% female

100-Car Study Features
- First large-scale instrumented-vehicle study undertaken with the primary purpose of collecting pre-crash and near-crash naturalistic driving data.
- Captured a range of crash severities: airbag deployments to minor, low-force, no-property-damage crashes.
- First study to collect detailed information on a large number of near-crash events.
- Drivers were given no special instructions and no experimenter.
- Instrumentation was unobtrusive.

Data Collection Instrumentation Included
- Five channels of digital video
- Front and rear radar sensors
- Accelerometers
- Machine vision-based lane tracker
- GPS
- Vehicle speed sensor

The Database
- Contains many extreme driving cases, including severe drowsiness, impairment, judgment error, risk taking, secondary task engagement, aggressive driving and traffic violations
- Each safety-related conflict was classified as one of the following:
  - Crash – any physical contact between the subject vehicle and another vehicle, fixed object, pedestrian, pedalcyclist or animal.
  - Near-Crash – situations requiring a rapid, severe evasive maneuver to avoid a crash.
  - Incident – situations requiring an evasive maneuver occurring at less magnitude than a near-crash.

The real-world data collected from the 100-Car Study lends itself to multiple additional analyses.

Top Level Database Statistics
- Approximately 2,000,000 vehicle miles
- 42,300 hours of driving data
- 15 police-reported and 67 non-police-reported crashes
- 761 near-crashes
- 8,295 incidents

Types of Driving Behavior Recorded
- Drowsiness
- Driver inattention
- Traffic violations
- Aggressive driving and "road rage"
- Seat belt usage

Discoveries

Driver Inattention
- Nearly 80% of all crashes and 65% of all near-crashes involved driver inattention just prior to (i.e., within 3 seconds) the onset of the conflict.

Rear-End-Striking Crashes
- Visual inattention was a contributing factor for 93% of rear-end-striking crashes.
- In 86% of rear-end-striking crashes, the headway at the onset of the event was greater than 2.0 s.
- Most near-crashes involving conflict with a lead vehicle occurred while

Sherri Box, PR & Marketing Manager
540/231.1549, sbox@vtti.vt.edu
3500 Transportation Research Plaza, Blacksburg, Virginia 24061
www.vtti.vt.edu

Crash

Any contact with an object, either moving or fixed, at any speed in which kinetic energy is measurably transferred or dissipated, and includes other vehicles, roadside barriers, objects on or off of the roadway, pedestrians, cyclists, animals.

Near-crash

Any circumstance **requiring a rapid, evasive maneuver** by the subject vehicle, or any other vehicle, pedestrian, cyclist, or animal to avoid a crash. A rapid, evasive maneuver is defined as a steering, braking, accelerating, or any combination of control inputs that approaches the limits of the vehicle capabilities. As a guide: Subject vehicle braking >0.5 g or steering input that results in a lateral acceleration >0.4 g to avoid a crash constitutes a rapid maneuver.

Selection: “trigger” as above – and then manual selection and annotation

Part A: How can information from near-crashes be used to prevent real crashes? (→ selection bias)

1

Do near-crashes resemble real crashes? Are more extreme near-crashes more like real crashes?

2

Is it possible to find driver behavior or traffic situations which is different in near-crashes than in normal driving? Are these differences even more extreme in real crashes?

Statistical methods used so far:

Odds ratios and logistic regression: Completly dominant – but can't easily extrapolate from less severe events to more severe ones, can't easily judge extent of selection bias.

Regression: Is relative risk the same for crashes and for near-crashes?

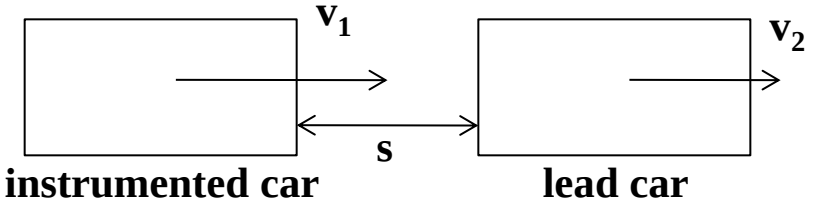
Extreme Value Statistics (almost new): Can near-crashes predict the frequency of real crashes? Do covariates behave in same way for crashes and near-crashes? Requires a continuous crash proximity or crash severity measure.

Underlying philosophy: a traffic accident is a rare and extreme event.

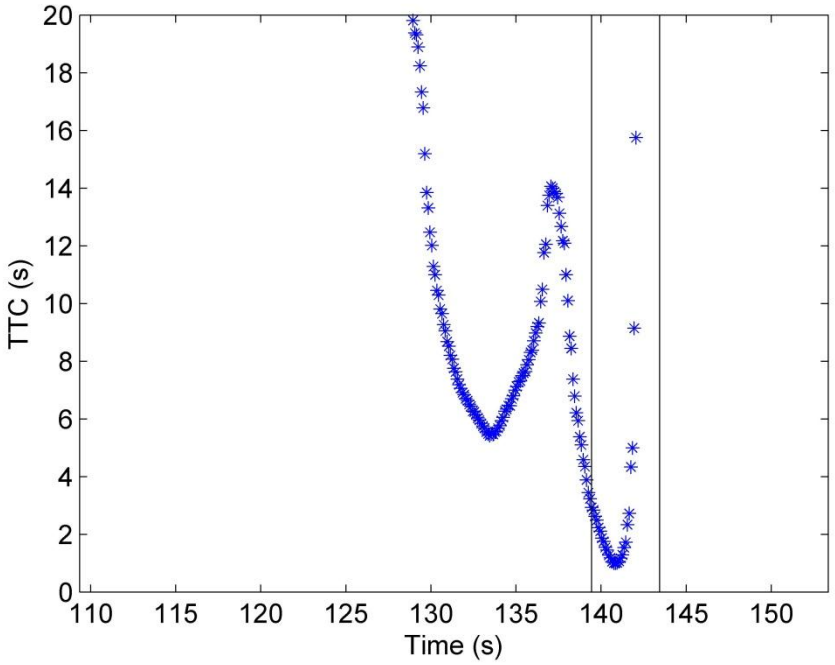
Crash proximity measure

- Measure of how close the near-crash is to a real crash
- Examples: TTEC = Time To edge Crossing, Gap = time between first car leaves conflict area and second car enters conflict area, Time-to-collision (TTC), …
- Here, TTC, the time it takes for the cars to collide when continuing with the same speeds – useful for **rear-ending**

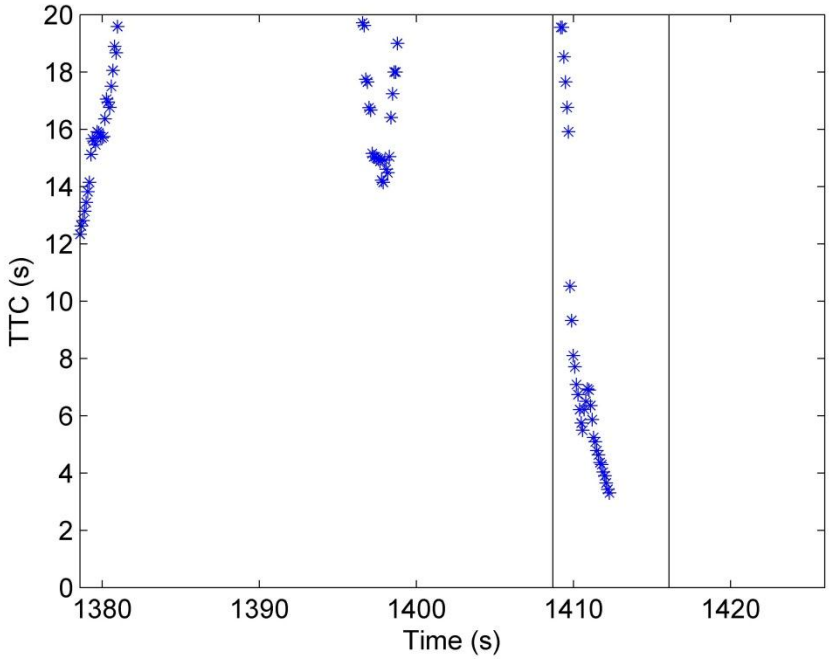
$$TTC = \frac{s}{v_1 - v_2}$$



Examples of TTC computed from radar signals



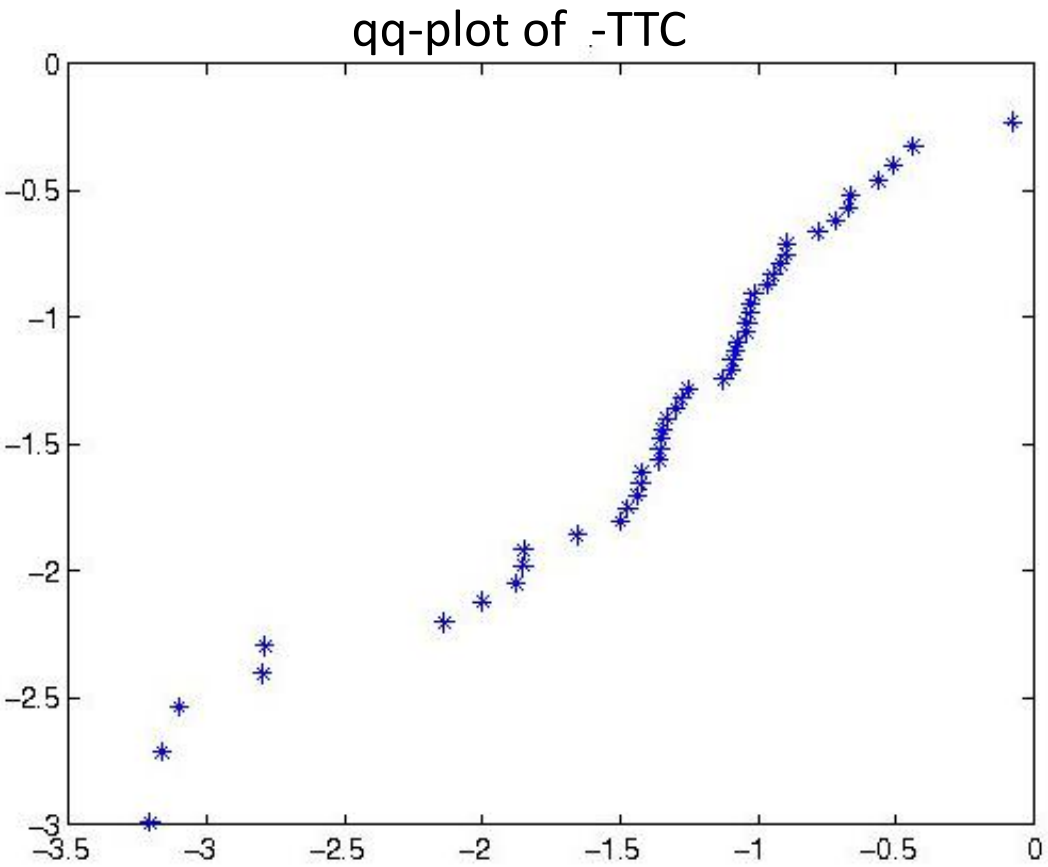
Possible to extract min TTC



Not possible to extract min TTC

100-car data, risk of rear-ending, TTC

384 near-crashes, 29 with good enough radar signals, 14 crashes.



Crash ⇔ TTC < 0

Block maxima 95% confidence interval for expected number of crashes is (0.07, 0,09) (Fitted GEV conditional on –TTC > 0, delta method conf. intervals)

Observed number of crashes = 14

Doesn't match!

Details

recall $GEV(z) = \exp\{-(1 + \frac{\gamma}{\sigma}(z-\mu))_+^{-1/\gamma}\}$

$z_1, \dots z_{29}$ observed values of $-\max(-TTC)$ (= $\max(TTC)$)

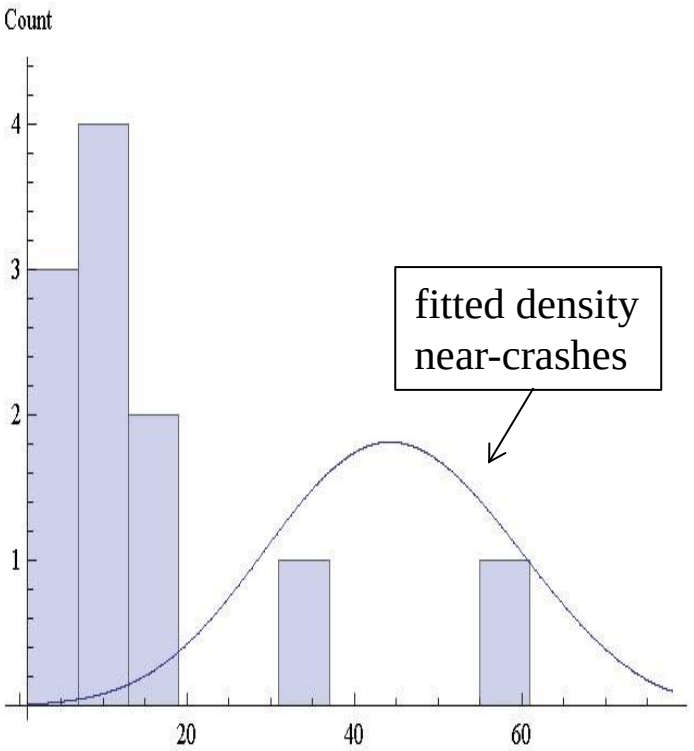
$$l(\mu, \sigma, \gamma; z_1, \dots, z_n) = -29 \log \sigma - \left(\frac{1}{\gamma} + 1\right) \sum_{i=1}^{29} \log\left\{1 + \frac{\gamma}{\sigma}(z_i - \mu)\right\}$$
$$- \sum_{i=1}^{29} \left(1 + \frac{\gamma}{\sigma}(z_i - \mu)\right)^{-1/\gamma} + 29\left(1 - \frac{\gamma}{\sigma}\mu\right)^{-1/\gamma},$$

for $z_1, \dots z_n \cdot \ 0.$

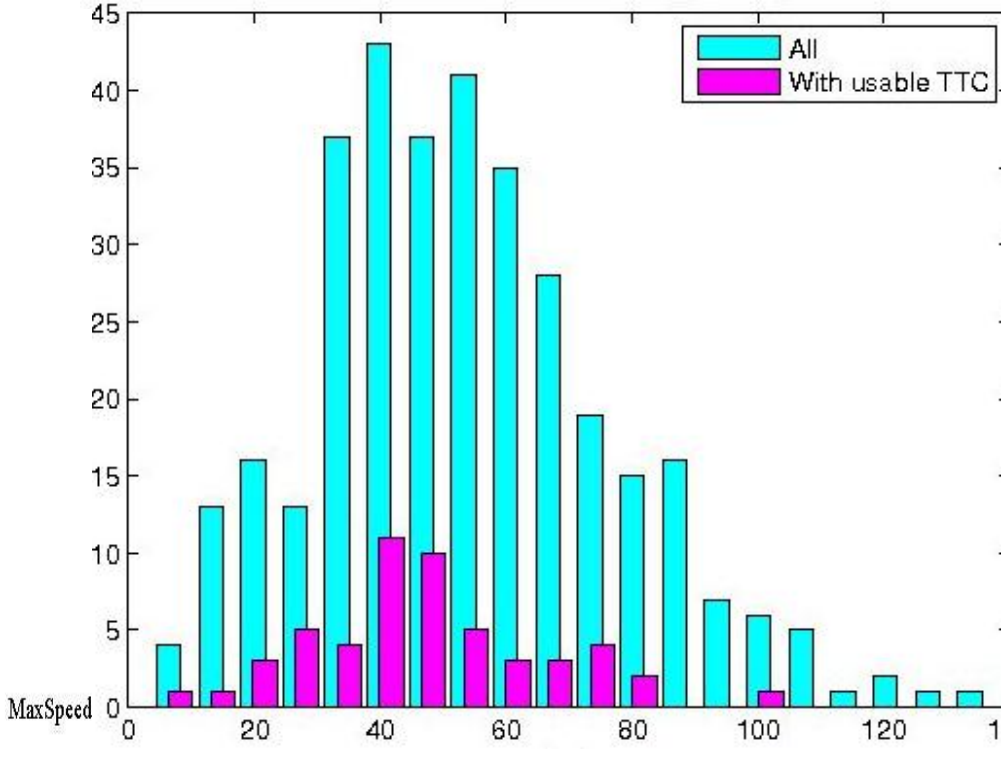
Maximum likelihood estimates $\hat{\mu} = -1.21,\ \hat{\sigma} = 0.21,\ \hat{\gamma} = -0.096$

Confidence interval for expected number of crashes (= $\Pr(-\min(-TTC) < 0) \times \#\{\text{near-crashes and crashes}\}$) via observed information matrix and delta method

Selection bias!



crashes, max speed (km/h)



near-crashes, max speed (km/h)

All but two of the real rear-ending crashes were in start-stop traffic while all the near-crashes with usable TTC were in higher speed situations

So maybe still: → *yes to question 1 (?)*

2

Continuous variables that could influence crash risk:

- Speed
- absolute value of yaw angle
- distance to right and left lane markings
- time the driver looks off-road during last 2 s or 3 s, total length of glances off-road longer than 1.5 s during last 15 s …
- variance of lateral acceleration
- variance of longitudinal acceleration

…..

- Length of overlapping glance off road

Do any of these become more and more extreme as TTC gets smaller and smaller?

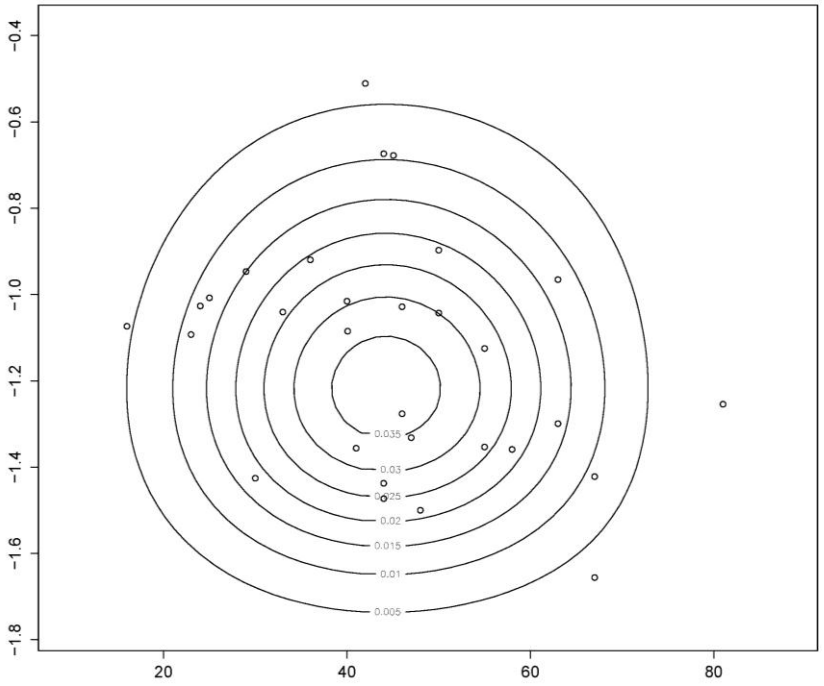
2

- Fitted logistic bivariate extreme value distribution to min/max of each of these variables and TTC for near-crashes, $\alpha \in (0, 1]$ dependence parameter, 1 is independence and 0 is complete dependence

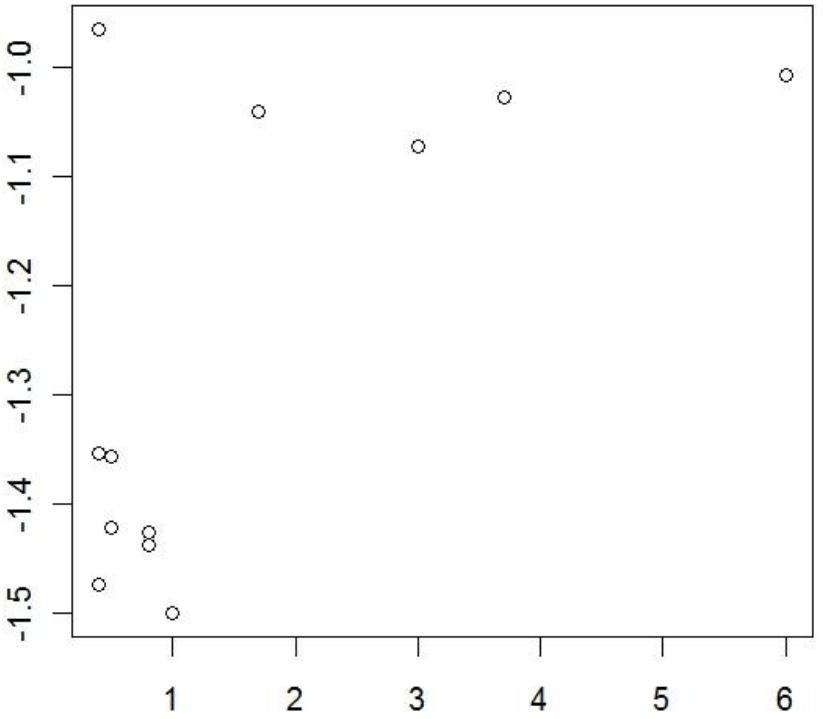
max(eye off road in 3 s window)	α=1.00
max (speed)	α=1.00
max (variance longitudinal acc)	α=1.00
min (dist left markings)	α=1.00
max (dist right markings)	α=0.93

Fitting was not possible for the other variables, however no indication of dependence, except for the last one (length overlapping glance off road)

max -TTC



max speed $\rightarrow \alpha = 1$



length overlapping glance

(12 with overlapping glance, 13 without overlapping glance, 4 vithout video)

Details

Marginal distributions:

$$G_1(x) = \exp\{-(1 + \tfrac{\sigma_1}{\gamma_1}(x - \mu_1)_+^{-1/\gamma_1}\}$$

$$G_2(x) = \exp\{-(1 + \tfrac{\sigma_2}{\gamma_2}(x - \mu_2)_+^{-1/\gamma_2}\}$$

Joint distributions

$$G(x_1, x_2) = \exp\{(G_1^{-1}(x_1)^{-1/\alpha} + G_2^{-1}(x_2)^{-1/\alpha})^{\alpha}\}$$

$$\text{for } \alpha \in (0, 1)$$

Parameters $\mu_1, \mu_2, \sigma_1, \sigma_1, \gamma_1, \gamma_2, \alpha$ estimated by maximum likelihoood

Part B: Visual behavior/censoring

How much do you look off road while driving?

- ❑ 5% of the time
- ❑ 10% of the time
- ❑ 15% of the time
- ☒ 20% of the time

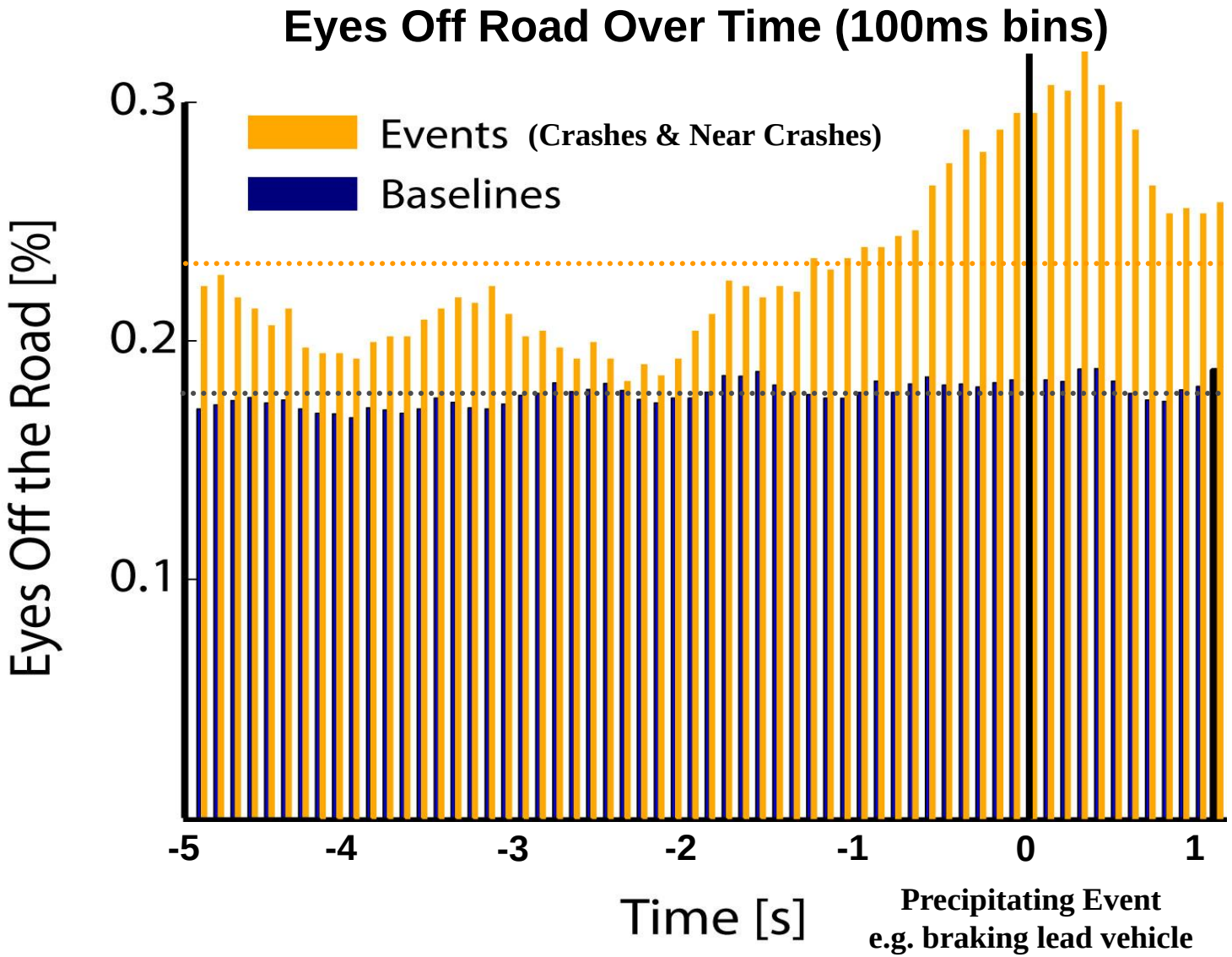
What is the .999 quantile of the lengths of off road glances?

- ❑ 1 second
- ❑ 2 seconds
- ❑ 3 seconds
- ❑ 4 seconds
- ☒ 5 seconds
- ❑ 10 seconds

Is glance behavior different in different circumstances?

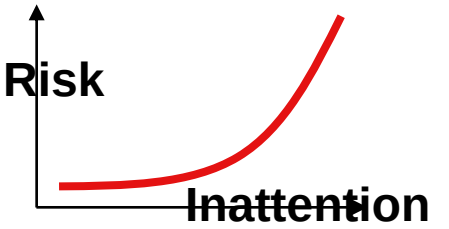
Not well understood

Example Victor & Dozza (2012)



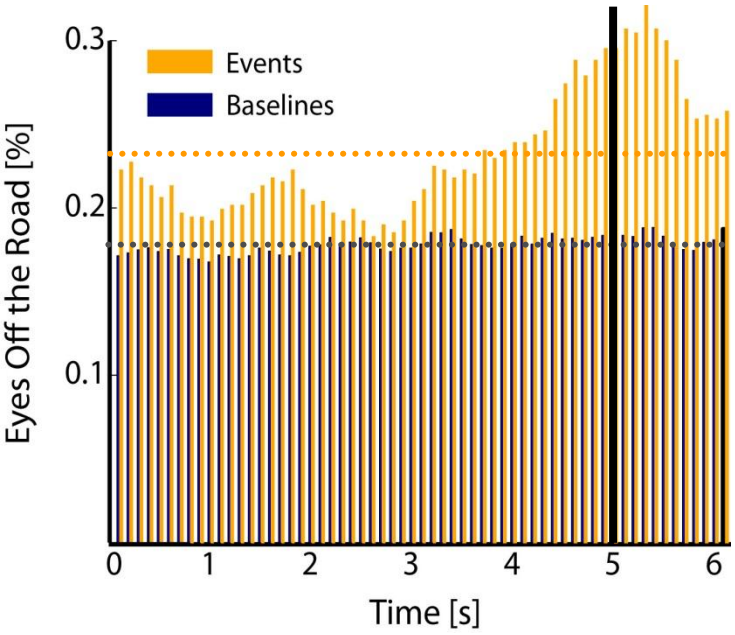
SAFER

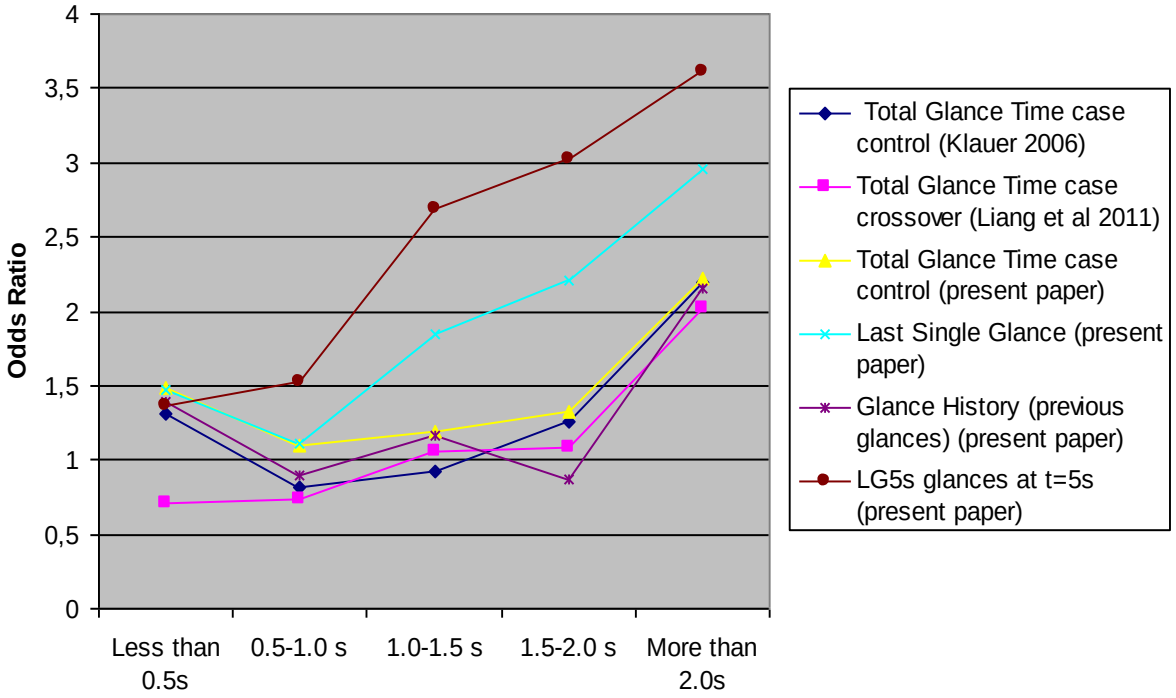
Inattention & Risk example



Victor and Dozza, 2012

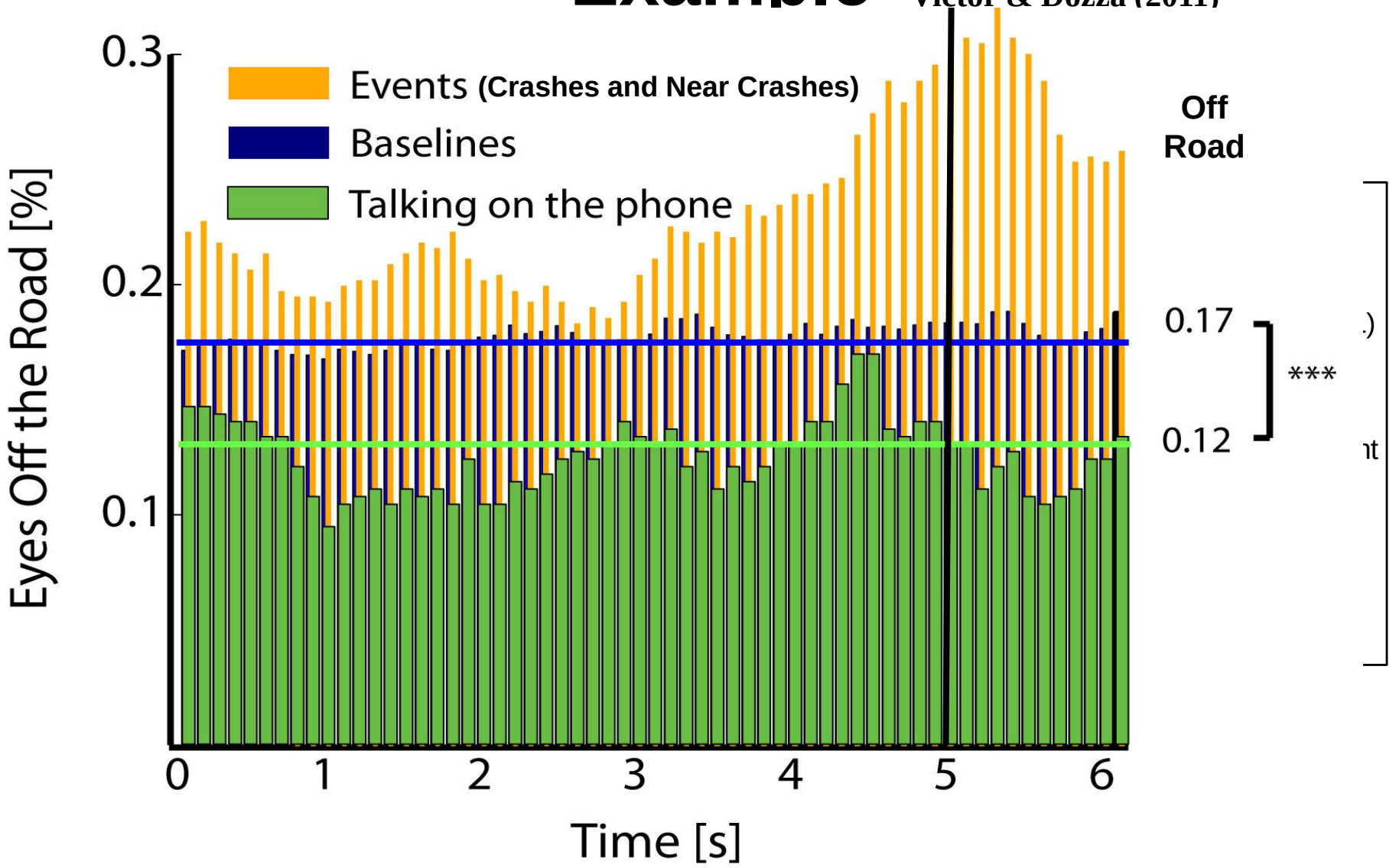
SHRP2





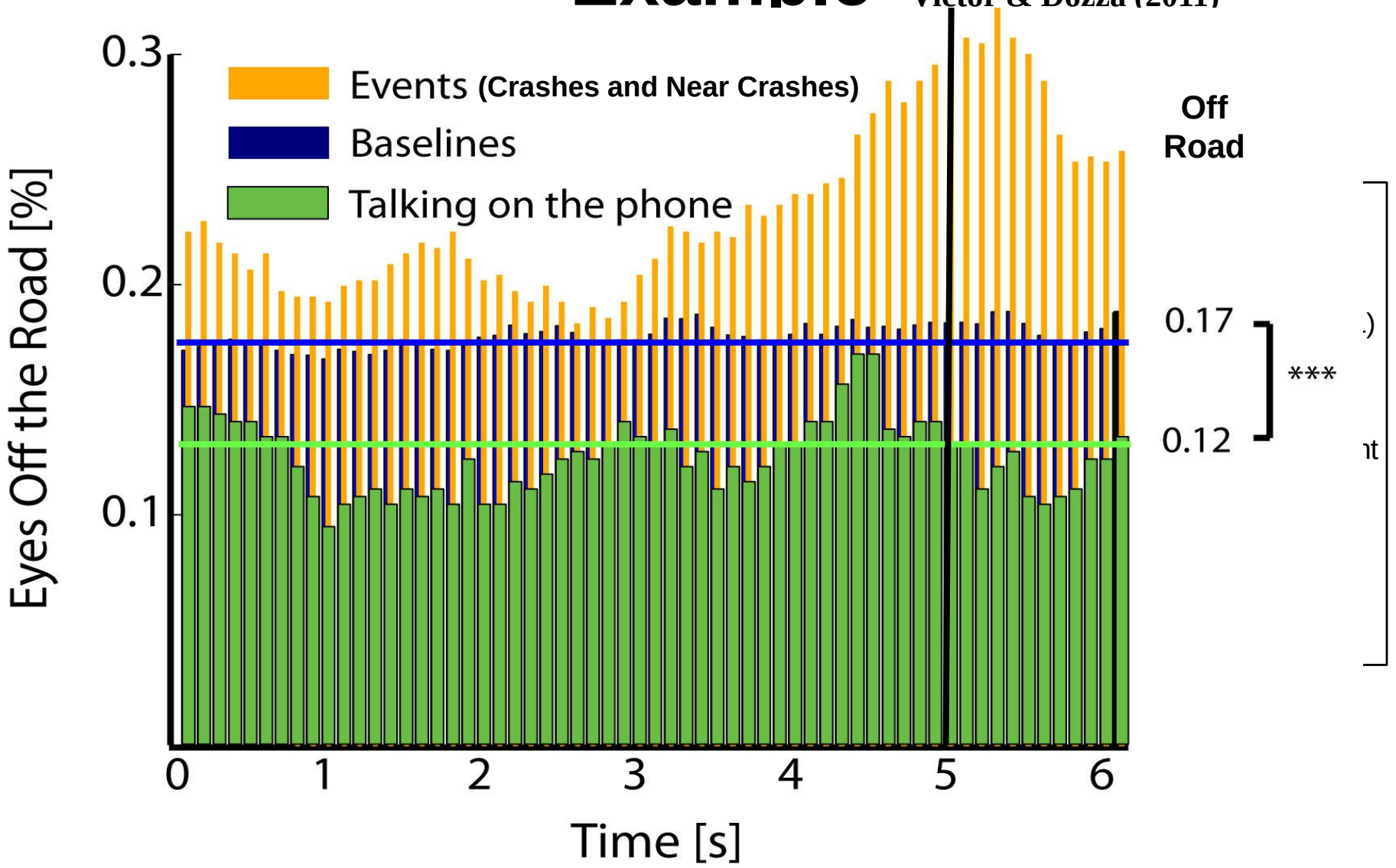
The overlapping glance gives the highest OR-s

Example



SAFER

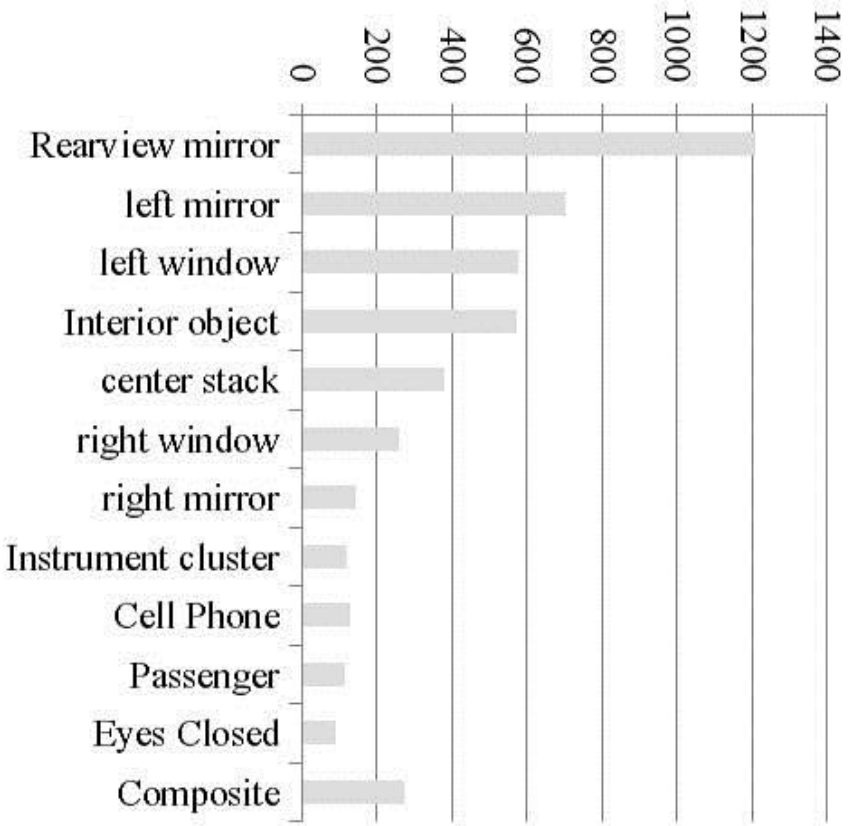
Example



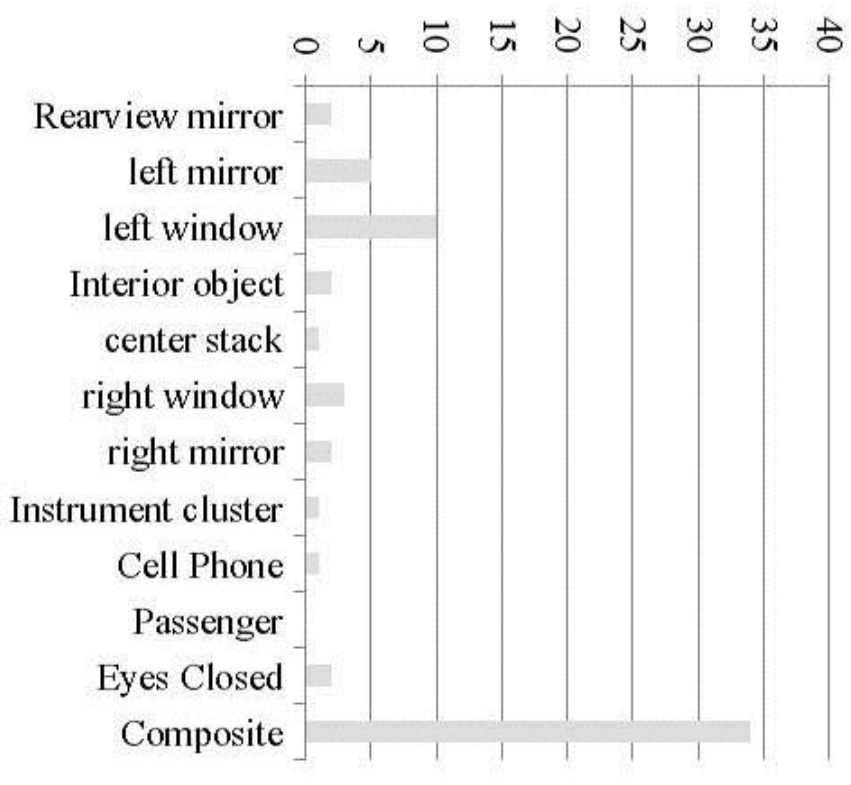
SAFER

Glance behavior in the 100-car study

Raw data: 19,616 annotated 6-second intervals from 100-car study: 4582 with 1 or more off road glances



Glances shorter than 3 seconds



Glances longer than 3 seconds

censoring

Model: think of alternating ”zero-one” renewal process, but valid more generally for stationary ergodic processes

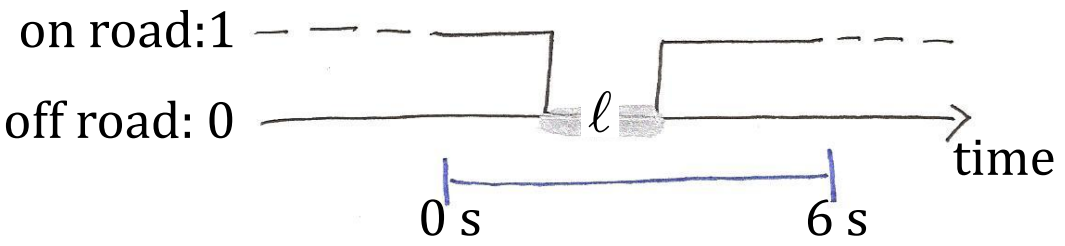
F_0 distribution function of lengths of off-road glances f_0 density function

Only use first glance in observation interval

nc

n_{nc} uncensored glances off road

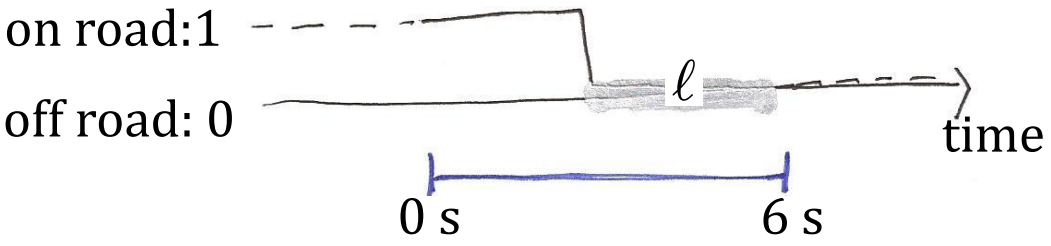
likelihood: $\prod_{k=1}^{n_{nc}} f_0(\ell_{\mathrm{nc},k};\theta)$



rc

n_{rc} right censored glances off road

likelihood: $\prod_{k=1}^{n_{rc}} (1 - F_0(\ell_{\mathrm{rc},k};\theta))$



Sizebiased sampling:

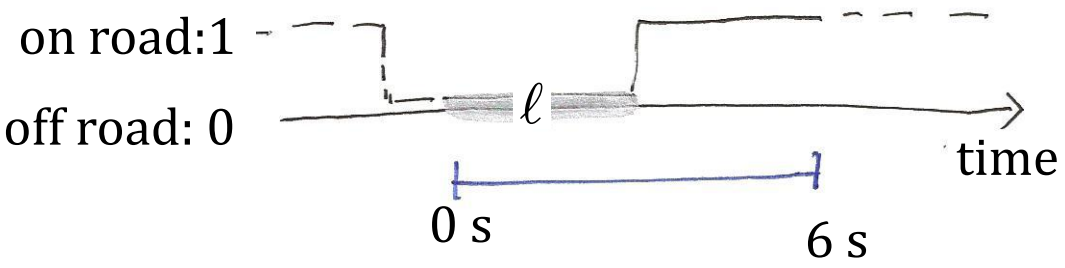
$xf_0(x;\theta)$ density of zero-interval overlapping left endpoint

$f_0^o(x;\theta) = (1-F_0(x;\theta))/\mu(\theta)$ density of overshoot $(\mu(\theta) = \int_0^\infty xf_0(x;\theta)dx)$

lc

n_{lc} left censored glances off road

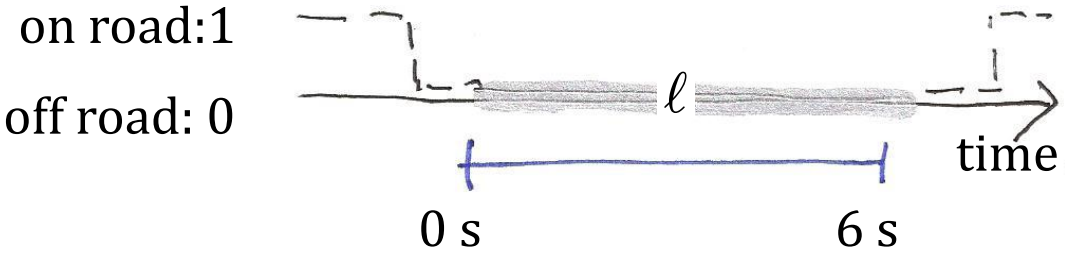
likelihood: $\prod_{k=1}^{n_{lc}} f_0^o(\ell_{\mathrm{lc},k};\theta)$



dc

n_{dc} doubly censored glances off road

likelihood: $n_{dc}(1 - F_0^o(w))$



But: $n_{nc}, n_{rc}, n_{lc}, n_{dc}$ also contains information about θ

$p_0 = Pr$(start in zero-interval), $p_1 = Pr$(start in one-interval)
F_0^o overshoot distribution function of one-intervals
f_1^o overshoot density of one-intervals

$$p_{nc}(w,\theta) = \int_0^w F_0(w-s) f_1^o(s) ds$$

$$p_{rc}(w,\theta) = \int_0^w (1 - F_0(w-s)) f_1^o(s) ds$$

$$L(n_{lc}, n_{dc}, n_{nc}, n_{rc}|n) = C \times (p_0 F_0^o(w,\theta))^{n_{lc}} (p_0(1 - F_0^o(w,\theta))^{n_{dc}}$$
$$(p_1 p_{nc}(w,\theta)^{n_{nc}} (p_1 p_{rc}(w,\theta))^{n_{rc}} / (p_1(1 - F_1^o(w)))^n$$

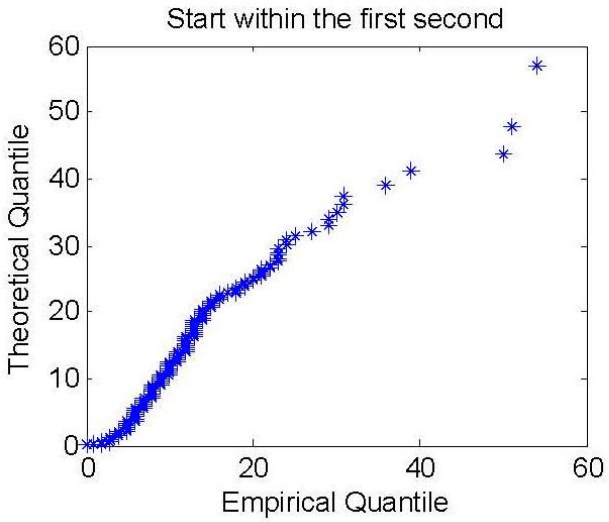
Replace F_0 by estimate, disregard p_0, p_1 and then estimate θ by maximizing the product of the remaining likelihood, and the 4 likelihoods on previous slide

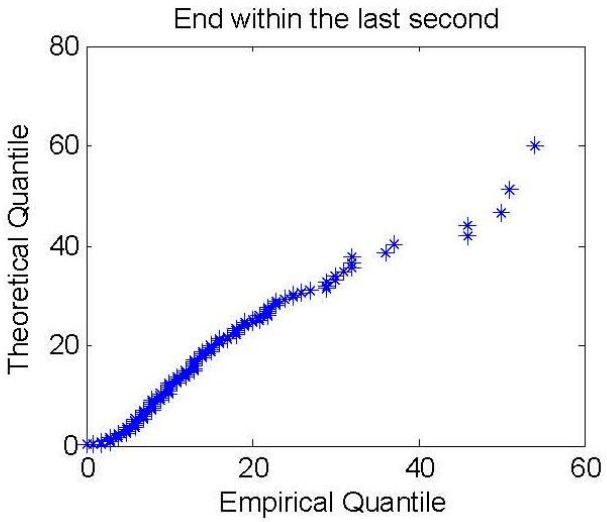
Most important case: $F_0(x) = 1 - e^{-\lambda_0 x}$, $F_1(x) = 1 - e^{-\lambda_1 x}$
(→ overshoot distribution = ordinary distribution)

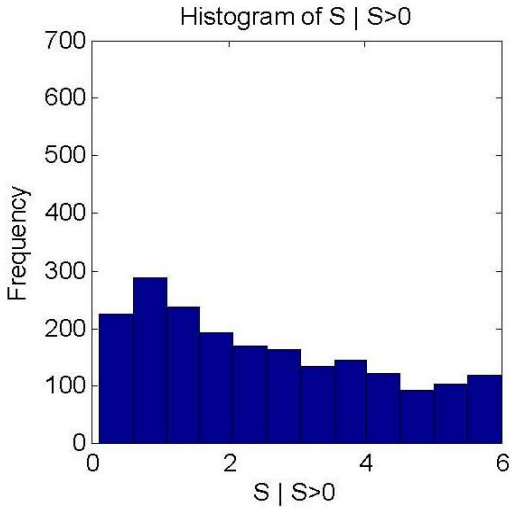
$$
\begin{aligned}
\ell(\lambda_0) \quad = \quad & -\lambda_0 \Big(\sum_{k=1}^{n_{lc}} \ell_{\mathrm{lc},k} + n_{dc} 2w + \sum_{k=1}^{n_{nc}} \ell_{\mathrm{nc},k} + \sum_{k=1}^{n_{rc}} \ell_{\mathrm{rc},k}\Big) + n_{lc} \log(1 - e^{-\lambda_0 w}) \\
& + n_{nc} \log\Big(1 + \frac{\lambda_0}{\lambda_0 - \hat{\lambda}_1} e^{-\hat{\lambda}_1 w} + \frac{\hat{\lambda}_1}{\lambda_0 - \hat{\lambda}_1} e^{-\lambda_0 w}\Big) \\
& + n_{rc} \log\Big(\frac{\hat{\lambda}_1}{\lambda_0 - \hat{\lambda}_1}(e^{-\hat{\lambda}_1 w} - e^{-\lambda_0 w}) + (n_{lc} + n_{nc}) \log \lambda_0 .
\end{aligned}
$$

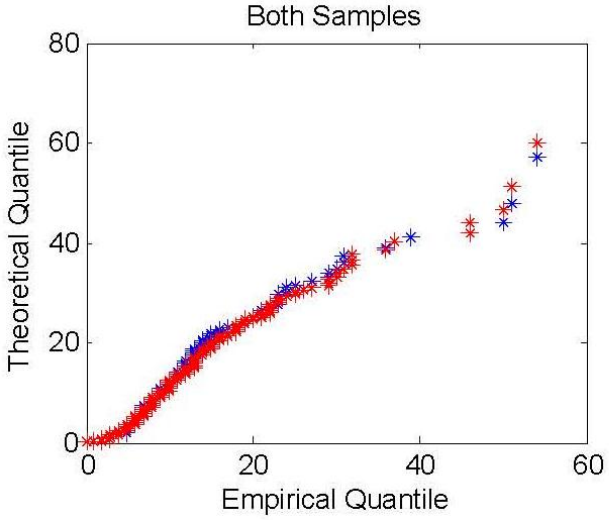
→ $\hat{\lambda}_0 = 0.92$ (λ_1 estimated "externally")

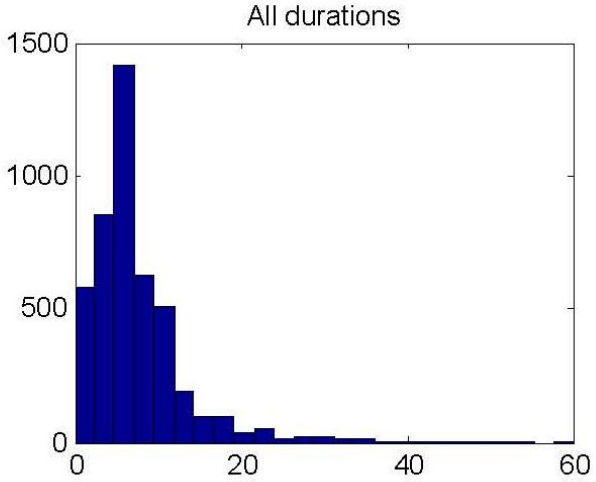
Model fit

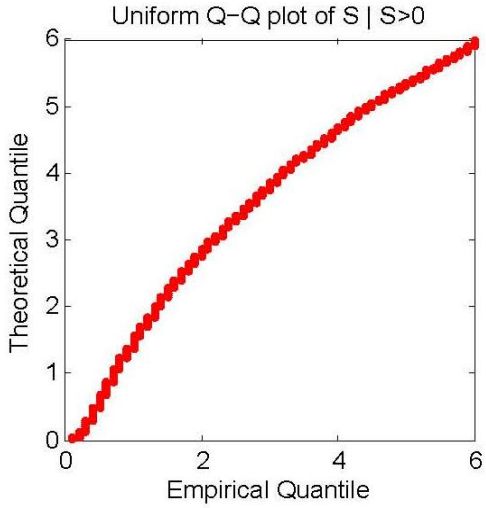












Glance lengths (tenths of seconds)

Starting times (seconds)

Alt. 1: use gamma distribution or phase type distribution

Alt. 2: … but more complicated, and one is only really interested in tails → use tail estimation, i.e. only use observations which are longer than a threshold u and assume excesses $l^u = l - u$ have a Generalized Pareto distribution with d.f. $G(x) = 1 - (1 + \frac{\gamma}{\sigma}x)_+^{-1/\gamma}$ and density $g(x;\sigma,\gamma) = \frac{1}{\sigma}(1 + \frac{\gamma}{\sigma}x)^{-1/\gamma-1}$. Then

$$\bar{F}_0(x;\sigma,\gamma) = \bar{F}_0(u)\bar{G}(x-u;\sigma,\gamma), \quad \text{for} \quad x > u$$

and we get the overshoot density for $l - u$ as

$$f_0^o(x;\sigma,\gamma) = \bar{F}_0(u)\bar{G}(x;\sigma,\gamma)/\mu, \quad x > u \text{ for } \mu = \int_0^u x f_0(x) + u + \bar{F}(u)\int_u^\infty x g(x)$$

Now replace $\bar{F}_0(u)$ and $\int_0^u x f_0(x)$ by their empirical counterparts $\bar{q}_u$ and $\hat{\mu_u}$ to get a ”likelihood” as the product of the 4 factors on the next slide:

nc

$\bar{\boldsymbol{n}}_{nc}$ uncensored excess lengths $\quad \text{likelihood: } \prod_{k=1}^{\bar{n}_{nc}} g(\ell^u_{\mathrm{nc},k};\sigma,\gamma)$

rc

$\bar{\boldsymbol{n}}_{rc}$ right censored excess lengths $\quad \text{likelihood: } \prod_{k=1}^{\bar{n}_{nc}} \bar{G}(\ell^u_{\mathrm{rc},k};\sigma,\gamma)$

lc

$\bar{\boldsymbol{n}}_{lc}$ left censored excess lengths $\quad \text{likelihood: } \prod_{k=1}^{\bar{n}_{lc}} f^o_0(\ell^u_{\mathrm{lc},k};\sigma,\gamma)$

dc

$\bar{\boldsymbol{n}}_{dc}$ doubly censored excess lengths $\quad \text{likelihood: } \bar{n}_{dc}\bar{F}^o_0(\ell^u_{\mathrm{dc},k};\sigma,\gamma)$

$\bar{\boldsymbol{n}}_{nc}$, … , $\bar{\boldsymbol{n}}_{dc}$ doesn't contain further information about σ, γ

Next: divide into different situations, estimate tail separately for each group (or covariate model)

UMTRI (Gordon et al (2010)) “do near-crashes give similar risk estimates as crashes?”

Seemingly Unrelated Regression → *yes to question 1 (?)*

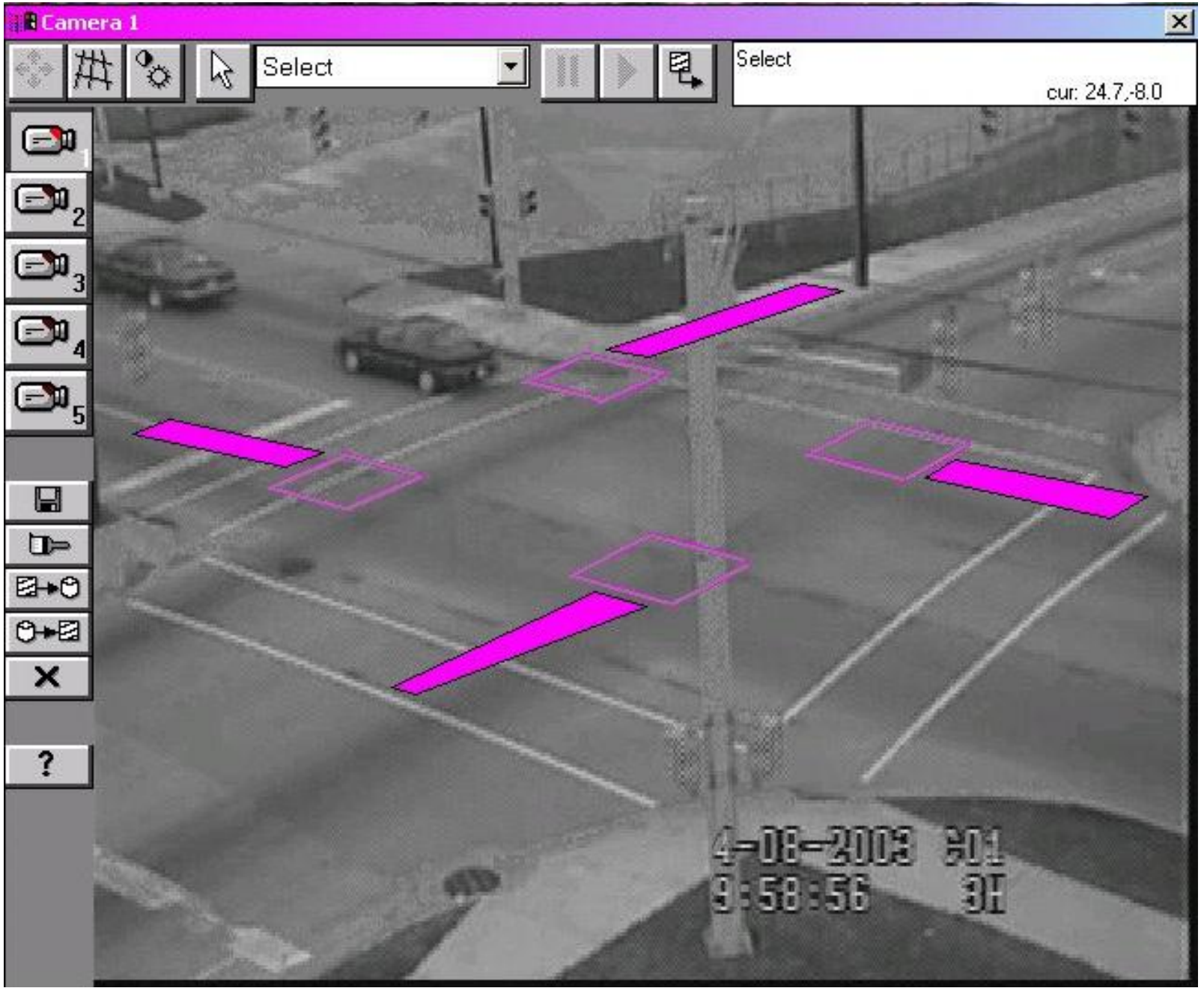
EVS: TTEC → road departure → road way departure crash

2.3 mile segment of US-23 with117 traversals by 43 different drivers in instrumented cars.

EV distribution fit to minimum TTEC values for the 117 traversals → predicts 12 road departures/year

On the average there were 1.8 road way departure crashes/year

→ *yes to question 1 (?)*

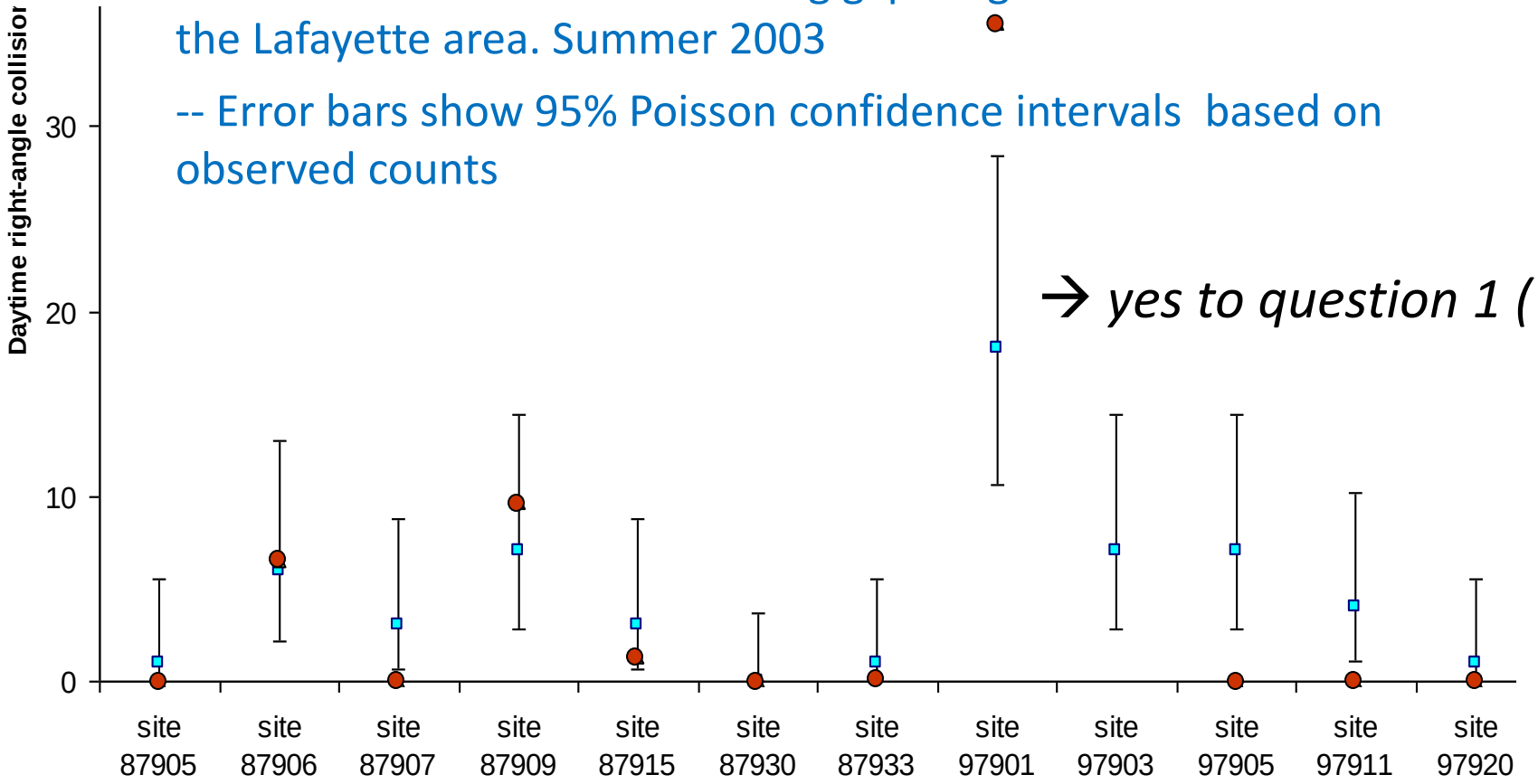

Camera 1
Select
Select
cur: 24.7,-8.0
1
2
3
4
5
?
4-08-2003 501
9:58:56 3I

Slide from presentation by P. Tarko, Purdue university: ”Risk evaluation for intersections”

- □ four year actual counts of daytime right angle collisions
- ● EVS estimate of crash frequency from gap measurements

-- 8-hour observations of crossing gaps. Signalized intersections in the Lafayette area. Summer 2003

-- Error bars show 95% Poisson confidence intervals based on observed counts



→ *yes to question 1 (??)*

One conclusion

different kinds of nearcrashes and crashes; naturalistics studies; vehicles; drivers, all lead to different kinds of

- Selection bias
- Crash proximity measures
- Driver behavior – and ”covariates”

All require separate careful analysis

No omnibus answer to ”is there selection bias in choice of near-crashes”

The future

- Use near-crashes to investigate how (and if) attention measures and other driving and traffic characteristics influence crash risk → highdimensional variable seclection → ***new research questions***

- Develop statistical predictors of crash risk → optimal choice of predictors → ***new research questions***

- Investigate the relation of risk estimates obtained in different naturalistic driving studies (Semifot, 100-car, SHRP 2, …)

- Study the normal driving – near-crash/crash relation in naturalistic driving experiments

More and better data crucial

SHRP 2

- 2000 cars
- 3 years
- Much better instrumentation (?)
- Started a year ago

M. Barnes*, A. Blankespoor, D. Blower, T. Gordon, P. Green, L. Kostyniuk, D. LeBlanc, S. Bogard., B. R. Cannon, and S.B. McLaughlin (2010). Development of Analysis Methods Using Recent Data: A Multivariate Analysis of Crash and Naturalistic Event Data in Relation to Highway Factors Using the GIS Framework. *Final Report* ***SHRP S01,*** *University of Michigan Transportation Research Institute*

Dozza, M. and Trent, V. (2012). Inattention – risk function (lead vehicle crashes). *A SHRP2 S08 analysis report.*

P. Tarko, and P. Songchitruksa (2006). Estimating frequency of chrashes as extreme traffic events. *Report, Purdue University*

J. Jonasson and H. Rootzén (2012). Internal validation of near-crashes in naturalistic driving studies: a continuous and multivariate approach.

K.-F. Wu and P.P. Jovanis. Crashes and crash-surrogate events: Exploratory modeling with naturalistic driving data. *Accident Analysis and Prevention*, **45**, 507–516, 2012.

Einmahl, J., Fils-Villetard, A., and Guillou, A. Statistics of extremes under random Censoring. *Bernoulli*, **14,** 207–227, 2008

Gomes, Y. and Neves, M. Estimation of the Extreme Value Index for Randomly Censored Data. 2010