Introduction, Day 2

Peter Solymos

Point count data analysis workshop, BIOS2 2021, March 16-25

Outline

Day 1

- ~~Introduction~~
- ~~We need to talk about data~~
- ~~A primer in regression techniques~~

Day 2

- Behavioral complexities

Day 3

- The detection process
- Dealing with recordings

Day 4

- Putting it all together
- Roadside surveys

What is detectability?

In the most colloquial terms, δ is the probability that a species is detected given it is present:

$$P(Y > 0 \mid N > 0)$$

Occupancy

In an occupancy framework, we can have:

- A detection: true positives (false positive rate is 0)
  - $P(Y > 0) = P(Y > 0 \mid N > 0)P(N > 0) + P(Y > 0 \mid N = 0)P(N = 0)$
  - $P(Y > 0) = \delta\varphi + 0(1 - \varphi) = \delta\varphi$
- A non-detection: false negatives + true negatives
  - $P(Y = 0) = P(Y = 0 \mid N > 0)P(N > 0) + P(Y = 0 \mid N = 0)P(N = 0)$
  - $P(Y = 0) = (1 - \delta)\varphi + 1(1 - \varphi) = 1 - \delta\varphi$,

(These are the marginal probabilities used to estimate the parameters using maximum likelihood.)

Side note on occupancy

People often confuse these two conditional probabilities:

1. Observing 0 given that the species is present:

- $P(Y = 0 \mid N > 0) = 1 - \delta$

2. Presence of a species given that we observe 0 (we ask this question *after* observing the data):

- $P(N > 0 \mid Y = 0) = \frac{P(Y=0|N>0)P(N>0)}{P(Y=0)}$
- $\frac{(1-\delta)\varphi}{\varphi(1-\delta)+(1-\varphi)} = \frac{\varphi(1-\delta)}{1-\delta\varphi}$

Abundance

A lot more combinations of true abundance and observed counts:

	$Y = 0$	1	2	...
$N = 0$	x			
1	x	x		
2	x	x	x	
...	x	x	x	x

Estimating detectability

To estimate δ, we need:

- ancillary information (multiple visits, distance bands, time intervals, multiple observers),
- parametric model assumptions (i.e. δ varies across locations).

The myth of constant detectability

Detectability zealots often view a method that cannot estimate constant detection probability δ (e.g. single-visit occupancy and N-mixture models) as inferior.

Fortunately for the rest of us: δ can only be constant in very narrow situations, e.g. when surveys are conducted:

- in the same region,
- in similar habitat,
- in the same year,
- on the same day,
- at the same time,
- by the same observer,
- using the same protocol.

Constant detectability is rare

Often a consequence of small sample size (i.e. not a lot of detection for a species)[1][2]:

Removal models[a]	Distance models[b] (0) No effects	(1) TREE	(2) LCC	Total
(0) No effects	15	1	4	20
(1) JDAY	9	18	9	36
(2) TSSR	1	2	–	3
(3) JDAY + JDAY^2	1	2	2	5
(4) TSSR + TSSR^2	1	–	–	1
(5) JDAY + TSSR	3	4	–	7
(6) JDAY + JDAY^2 + TSSR	–	1	–	1
(7) JDAY + TSSR + TSSR^2	1	1	–	2
(8) JDAY + JDAY^2 + TSSR + TSSR^2	–	–	–	–
Total	31	29	15	75

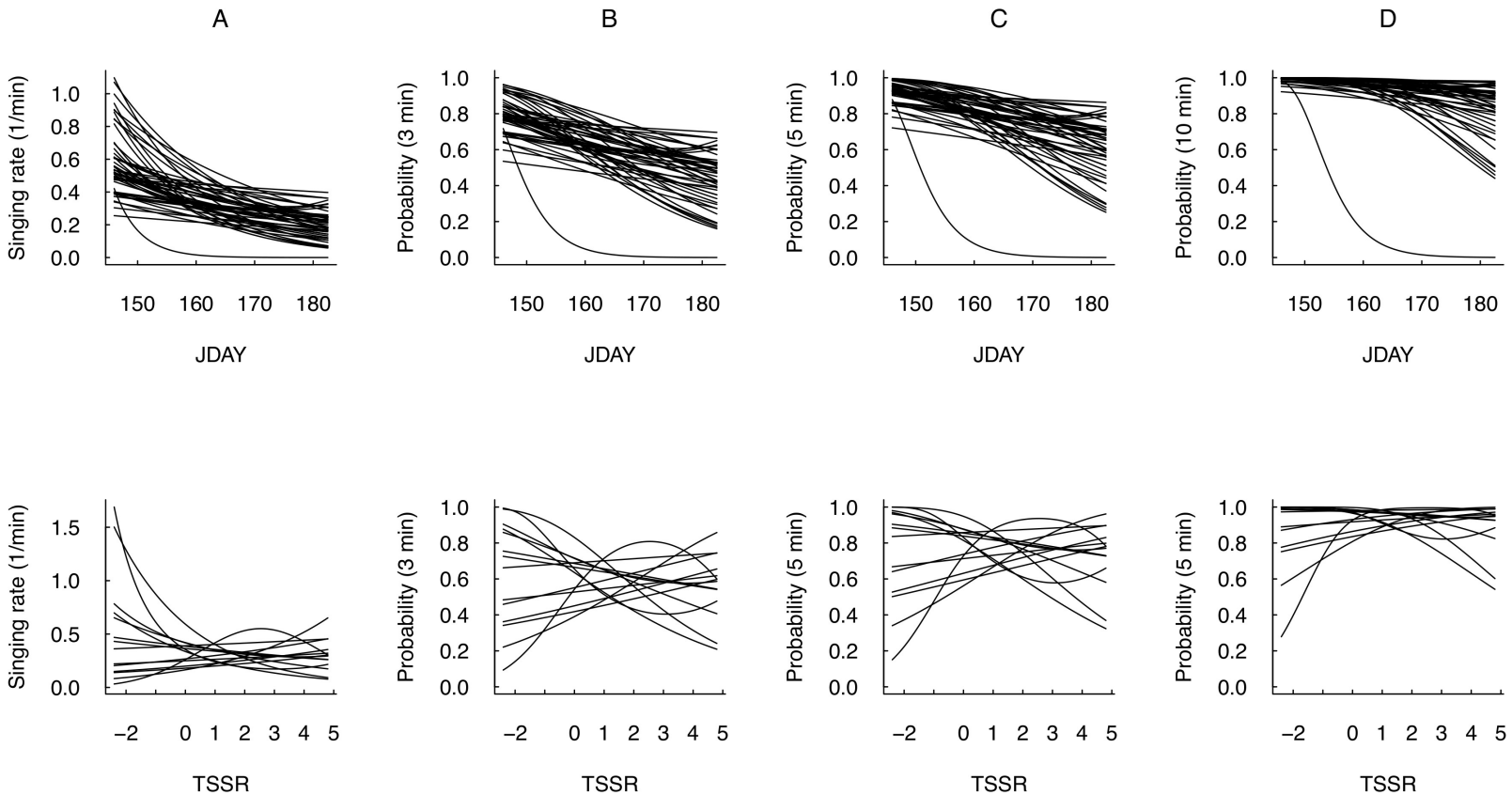
Model ID and covariate effects [b]	$M_0{}^{\varphi}$	$M_f{}^{\varphi}$	$M_f{}^{c}$	Total	Residents
0: Null (time-invariant: M_0 and M_f)	1	0	0	1	0
1: DY	1	6	2	9	2
2: SR	1	5	4	10	2
3: DY + DY^2	2	3	4	9	0
4: SR + SR^2	0	7	17	24	4
5: DY + SR	0	5	3	8	1
6: DY + DY^2 + SR	2	0	1	3	0
7: DY + SR + SR^2	2	2	19	23	3
8: DY + DY^2 + SR + SR^2	1	1	4	6	1
9: LS	0	3	3	6	0
10: LS + LS^2	0	2	10	12	4
11: LS + SR	1	0	6	7	0
12: LS + LS^2 + SR	0	1	4	5	0
13: LS + SR + SR^2	1	3	18	22	6
14: LS + LS^2 + SR + SR^2	0	0	2	2	0
Total	12	38	97	147	23

[b] Covariate effects included linear or quadratic effects of time since sunrise (SR), ordinal day of the year (DY), and the ordinal day relative to local spring (LS).

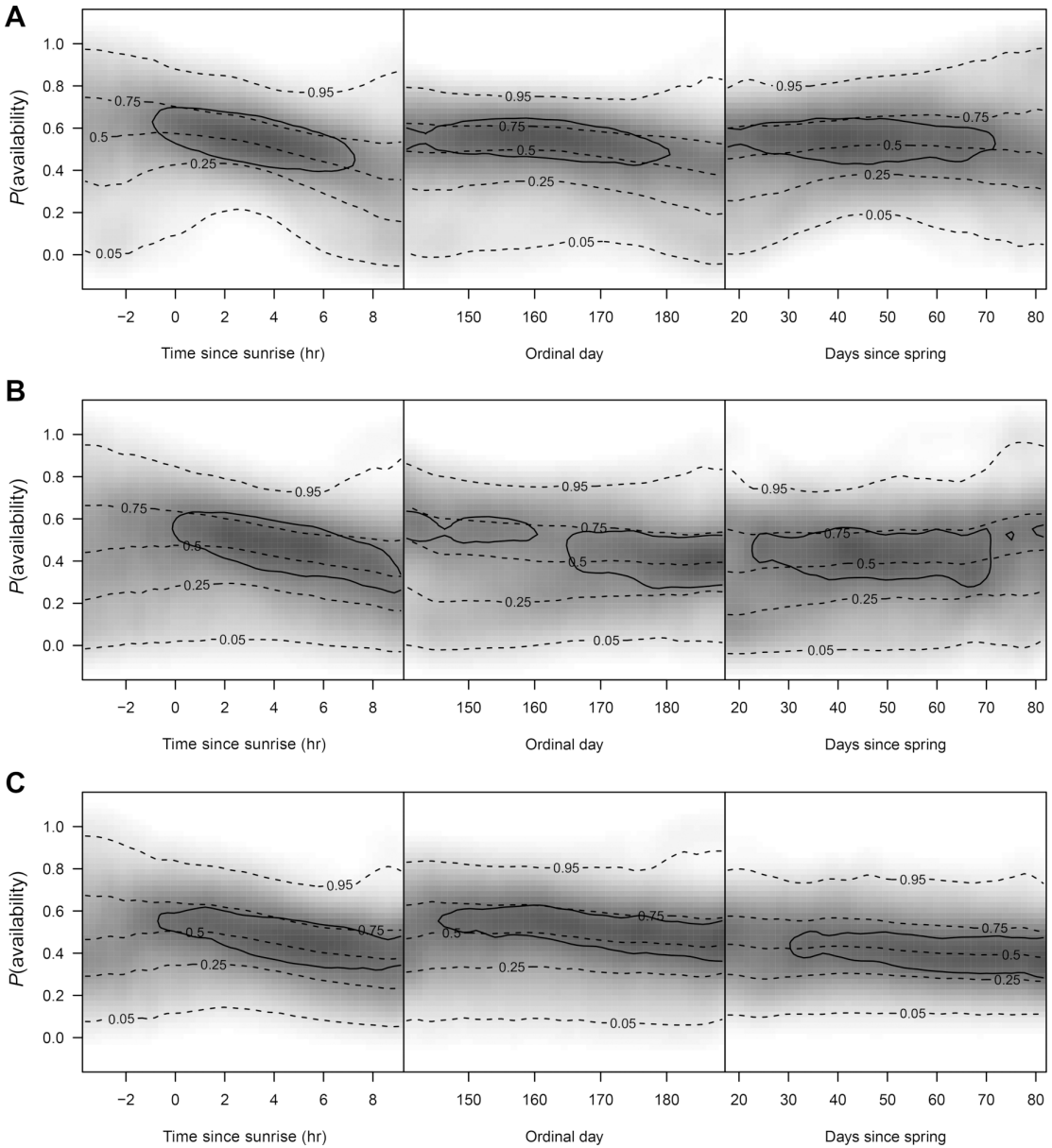
[1] Solymos et al. 2013, Methods. Ecol. Evol. 4:1047–1058.

[2] Solymos et al. 2018, Condor 120:765–786.

Availability varies



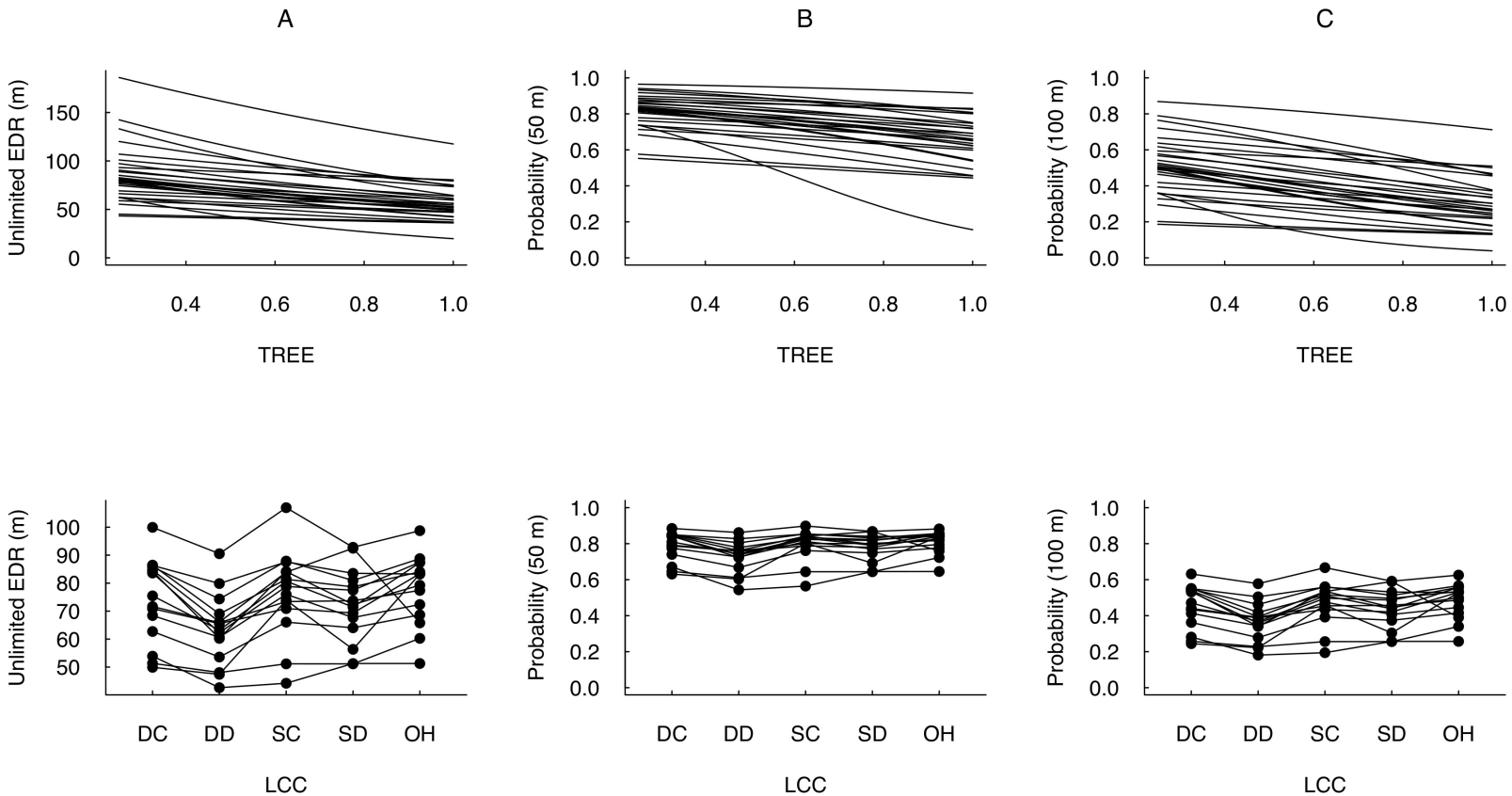
Availability really varies



Biological mechanisms

- Migration timing drives phenology for many species, e.g. ordinal day of year (`DAY`),
- when study spans across biomes, use time since local spring (multi-year average),
- or time since spring green up, last snow day, etc. based on actual survey year,
- time of day,
- time since local sunrise (`TSSR`).

Perceptibility varies too



Physical mechanisms

- Trees block the transmission of sound
- Broad leaves rustle more
- Louder sounds travel farther
- Low frequency sounds travel farther

Let's unwrap δ

1. Once the species/individual is present ($N > 0$)
2. It needs to signal its presence: make itself heard/visible, make itself available (p),
3. Then the signal needs to be received by a sensor: human ear or a microphone (and then the human ear in the lab listening to a recording), perceptibility q

These imply a pre-defined total time duration and maximum counting radius.

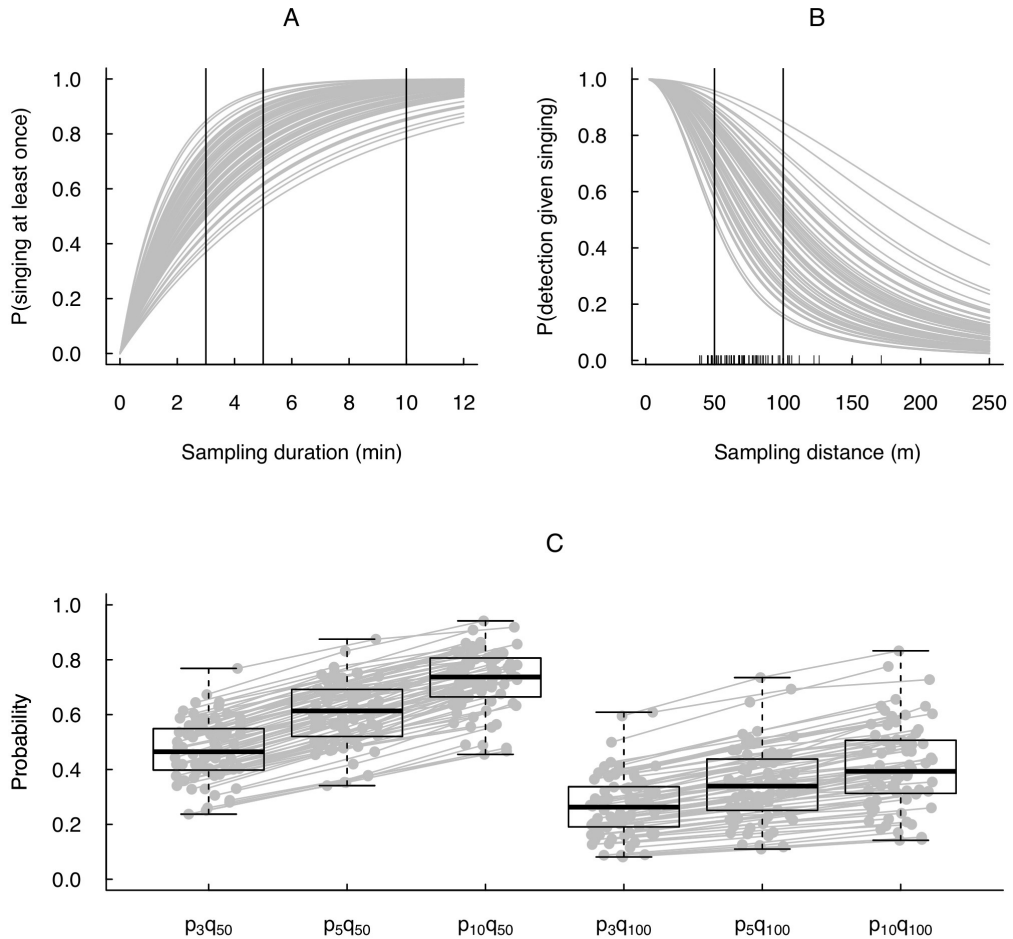
QPAD

Now we can expand our equation:
$E[Y] = NC = (AD)(pq) = qpAD$

The expected value of the observed count becomes a function of the:

- population density (D),
- area sampled (A),
- availability (p),
- and perceptibility (q).

Space and time to the rescue



p is a monotonic function of *time*, while q is monotonic function of *area* (space).

Simulate QPAD

```
D <- 2.0  # inds / unit area
A <- 1.0  # area
p <- 0.8  # availability
q <- 0.5  # perceptibility

n <- 100  # sample size
N <- rpois(n, lambda = A * D)
Y <- rbinom(n, size = N, prob = p * q)
```

Output

```
table(N=N, Y=Y)
```

```
##     Y
## N     0  1  2  3  4
##   0 16  0  0  0  0
##   1 11  9  0  0  0
##   2  7 11  5  0  0
##   3  4  8  9  2  0
##   4  0  2  5  1  1
##   5  0  1  3  2  1
##   6  0  0  0  1  0
##   7  0  0  0  1  0
```

The model

- $(N \mid D, A) \sim Poisson(DA)$
- $(Y \mid N, p, q) \sim Binomial(N, pq)$.

Incorporates key components of reality, but also ignores a lot of details.

Assumptions

- Observations are independent
- Y involves no double counting
- Area is known and measured without error
- Detectability (pq) is independent of N

And a lot more that we'll cover later

Why do we need simulation

Probabilistic simulation is useful to test how well a stat method is working *if* the assumptions are met.

It is not quite as useful for assessing how robust the method is when the assumptions are *not* met.

You need to link to biological mechanisms to do sensitivity analysis.

bSims goals

- Allow easy **testing of statistical assumptions** and explore effects of violating these assumptions
- **Aid survey design** by comparing different options
- And most importantly, to **have fun** while doing it via an intuitive and interactive user interface

bSims design

- **Isolation**: the spatial scale is small (local point count scale) so that we can treat individual landscapes as more or less homogeneous units (but see below how certain stratified designs and edge effects can be incorporated) and independent in space and time
- **Realism**: the implementation of biological mechanisms and observation processes are realistic, defaults are chosen to reflect common practice and assumptions
- **Efficiency**: implementation is computationally efficient utilizing parallel computing backends when available
- **Extensibility**: the package functionality is well documented and easily extensible

bSims verbs

- **Initialize** (`bsims_init`): the landscape is defined by the extent and possible habitat stratification
- **Populate** (`bsims_populate`): the population of finite number of individuals within the extent of the landscape
- **Animate** (`bsims_animate`): individual behaviours described by movement and vocalization events, i.e. the frequency of sending various types of signals
- **Detect** (`bsims_detect`): the physical side of the observation process, i.e. transmitting and receiving the signal
- **Transcribe** (`bsims_transcribe`): the "human" aspect of the observation process, i.e. the perception of the received signal

Behavioral events

Event time (T) is a continuous random variable

In the simplest case, its probability density function is the Exponential distribution: $f(t) = \phi e^{-t\phi}$

The corresponding cumulative distribution function is: $F(t) = \int_0^t f(t)dt = 1 - e^{-t\phi} = p_t$, the probability that the event has occurred by duration t

The parameter ϕ is the rate of the Exponential distribution with mean $1/\phi$ and variance $1/\phi^2$.

Survival and hazard function

The complement of $F(t)$ CDF is called the *survival function* ($S(t) = 1 - F(t)$, $S(0) = 1$), which gives the probability that the event has *not* occurred by duration t.

The *hazard function* ($\lambda(t) = f(t)/S(t)$) defines the instantaneous rate of occurrence of the event (*risk*, the density of events at t divided by the probability of surviving).

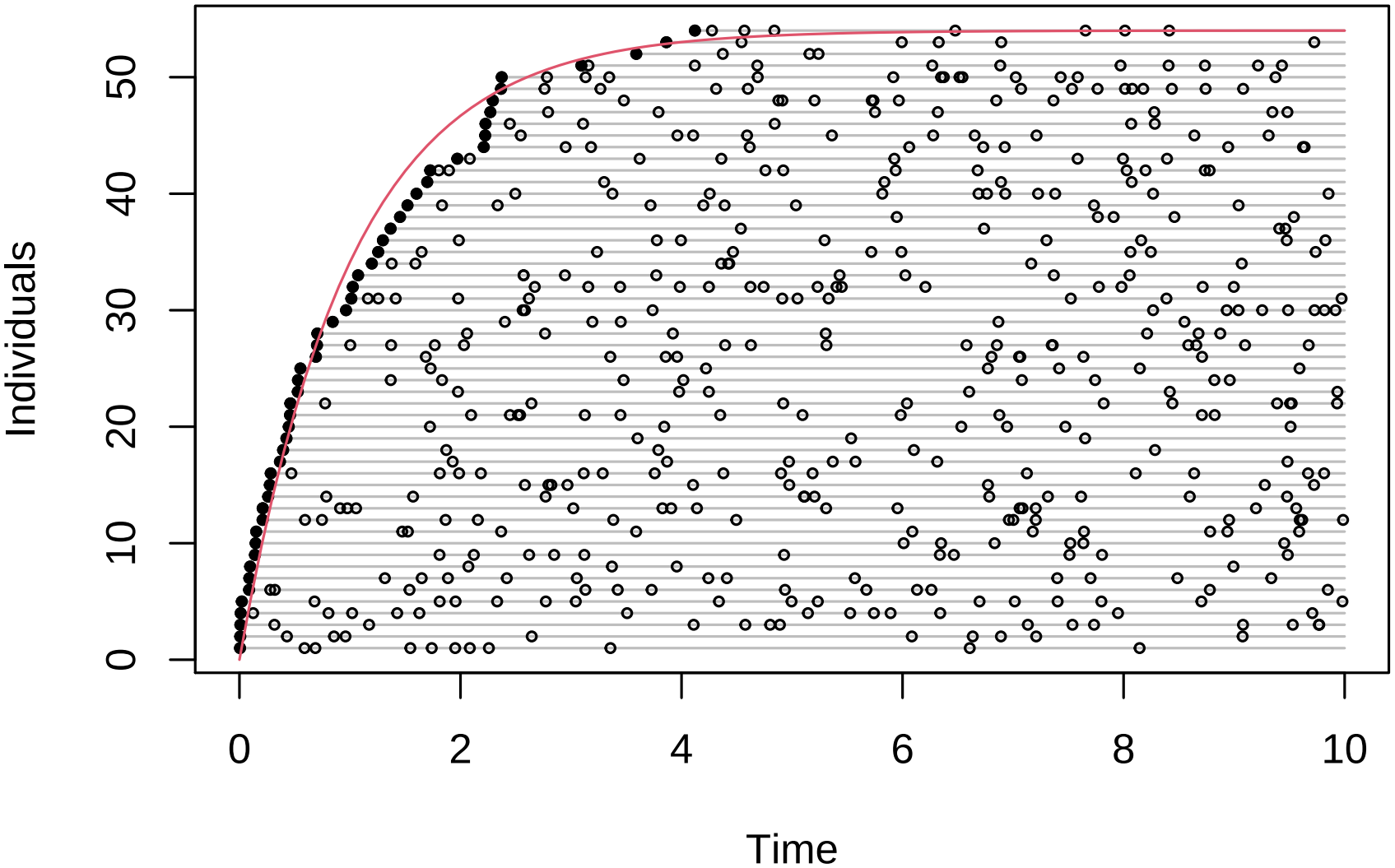
The cumulative hazard (cumulative risk) is the sum of the risks between duration 0 and t ($\Lambda(t) = \int_0^t \lambda(t)dt$).

Exponential model

The simplest survival distribution assumes constant risk over time ($\lambda(t) = \phi$), which corresponds to the Exponential distribution.

The Exponential distribution also happens to describe the lengths of the inter-event times in a homogeneous Poisson process (events are independent, it is a 'memory-less' process).

Exponential model visualized



Zoom in on mechanisms

Today we are focusing on availability, $p \in (0, 1)$

We assume that peceptibility is $q = 1$