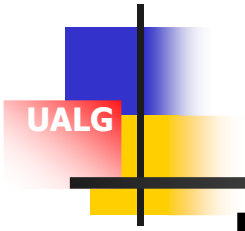


Data generation in Excel: Some technical guidelines

Einar Hjörleifsson

The simulator in math (almost)

einar

$$N_{a,y} = \begin{cases} R_y = N^{STO}_{A.y} = N^{DET}_{A,y} e^{\varepsilon_R} & \varepsilon \sim N\left(0, \sigma^2_R\right) \\ N_{a-1,y-1} e^{-(s_{a-1}F_{y-1} + M_{a-1,y-1})} \\ N_{a-1,y-1} e^{-(s_{a-1}F_{y-1} + M_{a-1,y-1})} + N_{a,y-1} e^{-(s_a F_{y-1} + M_{a,y-1})} \end{cases} \quad \begin{cases} a = A \\ A < a \le a_{plus} \\ a = a_{plus} \end{cases}$$

$$F_{ay} = s_a F_y$$

$$C^{TRUE}_{ay} = \frac{s_a F_y}{s_a F_y + M_{ay}} \left(1 - e^{-(s_a F_y + M_{ay})}\right) N_{ay}$$

$$C^{DET}_{ay} = \left(1 - p_{ay}\right) C^{TRUE}_{ay}$$

$$C^{OBS1}_{a,y} = C^{DET}_{a,y} e^{\varepsilon_{C_a}} \qquad \varepsilon_{C_a} \sim N\left(0, \sigma^2_{C_a}\right)$$

$$C^{OBS2}_{a,y} = C^{OBS1}_{a,y} \frac{Y^{OBS2}_y}{Y^{OBS1}_y}$$

$$Y_y = \sum_a C_{a,y} w^{catch}_{a,y}$$

$$R_y = f(time)$$

$$F_y = f(time)$$

$$s_{ay} = f(age, time)$$

$$q_a = f(age, time)$$

$$M_{ay} = f(?)$$

$$U^{TRUE}_{ay} = q_a N^{\beta_a}_{ay}$$

$$U^{OBS}_{ay} = U^{TRUE}_{ay} e^{\varepsilon_{U_a}} \qquad \varepsilon_{U_a} \sim N(0, \sigma_{U_a})$$

One objective: Generate measurements

einar

Survey indices

Uay	1	2	3	4	5	6	7	8	9
85	28	33	18	24	27	4	11	5	6
86	124	109	59	13	16	13	1	3	1
87	22	296	164	57	13	11	8	1	1
88	16	41	185	89	23	1	2	2	0
89	11	23	42	147	45	13	1	1	0
90	70	32	27	39	92	31	3	1	0
91	90	146	42	18	20	33	8	0	0
92	18	211	138	36	17	13	16	2	0
93	30	38	245	87	11	4	2	5	1
94	59	61	40	143	42	7	3	1	4
95	36	83	48	20	68	8	1	0	0
96	95	66	121	37	19	40	6	1	0
97	9	119	51	53	11	7	11	1	0
98	23	18	108	28	23	5	3	5	0
99	81	86	26	98	13	10	1	2	1
00	61	90	45	9	25	3	2	0	0
01	81	148	115	22	4	11	1	1	0
02	21	298	201	113	23	4	7	0	0
03	112	98	283	245	112	18	3	4	1
04	326	292	71	209	109	34	7	1	1
05	58	693	288	45	157	57	16	3	0

Catch at age (absolute numbers)

Cay	1	2	3	4	5	6	7	8	9
85		0.0	2.0	5.0	6.0	1.0	2.0	3.0	2.0
86		0.0	4.0	4.0	5.0	6.0	1.0	1.0	1.0
87		2.0	8.0	8.0	3.0	2.0	1.0	0.0	0.0
88		0.0	10.0	16.0	6.0	1.0	1.0	1.0	0.0
89		0.0	3.0	23.0	10.0	3.0	1.0	1.0	0.0
90		0.0	3.0	8.0	24.0	7.0	1.0	0.0	0.0
91		3.0	1.0	4.0	7.0	14.0	3.0	0.0	0.0
92		3.0	7.0	4.0	4.0	4.0	6.0	1.0	0.0
93		0.0	12.0	13.0	3.0	2.0	2.0	2.0	0.0
94		0.0	3.0	27.0	11.0	2.0	1.0	0.0	1.0
95		2.0	6.0	6.0	23.0	6.0	1.0	0.0	0.0
96		2.0	9.0	7.0	5.0	14.0	2.0	0.0	0.0
97		1.0	4.0	11.0	5.0	3.0	5.0	1.0	0.0
98		0.0	8.0	6.0	8.0	2.0	2.0	2.0	0.0
99		1.0	2.0	17.0	5.0	5.0	1.0	1.0	1.0
00		2.0	7.0	2.0	14.0	2.0	2.0	0.0	0.0
01		2.0	11.0	7.0	2.0	6.0	1.0	1.0	0.0
02		1.0	11.0	16.0	5.0	1.0	3.0	0.0	0.0
03		0.0	6.0	16.0	13.0	3.0	1.0	1.0	0.0
04		1.0	4.0	18.0	19.0	9.0	2.0	1.0	1.0
05									

Generating data: Using functions

- Empirical specifications
  - Simplest way to generate data is to punch in some starting numbers for each parameter
  - Cumbersome and inefficient, need to punch in a lot of values. But sometimes valuable.
- Use functions
  - Many commonly observed time trends in fish stocks can be described by 2-5 parametric functions

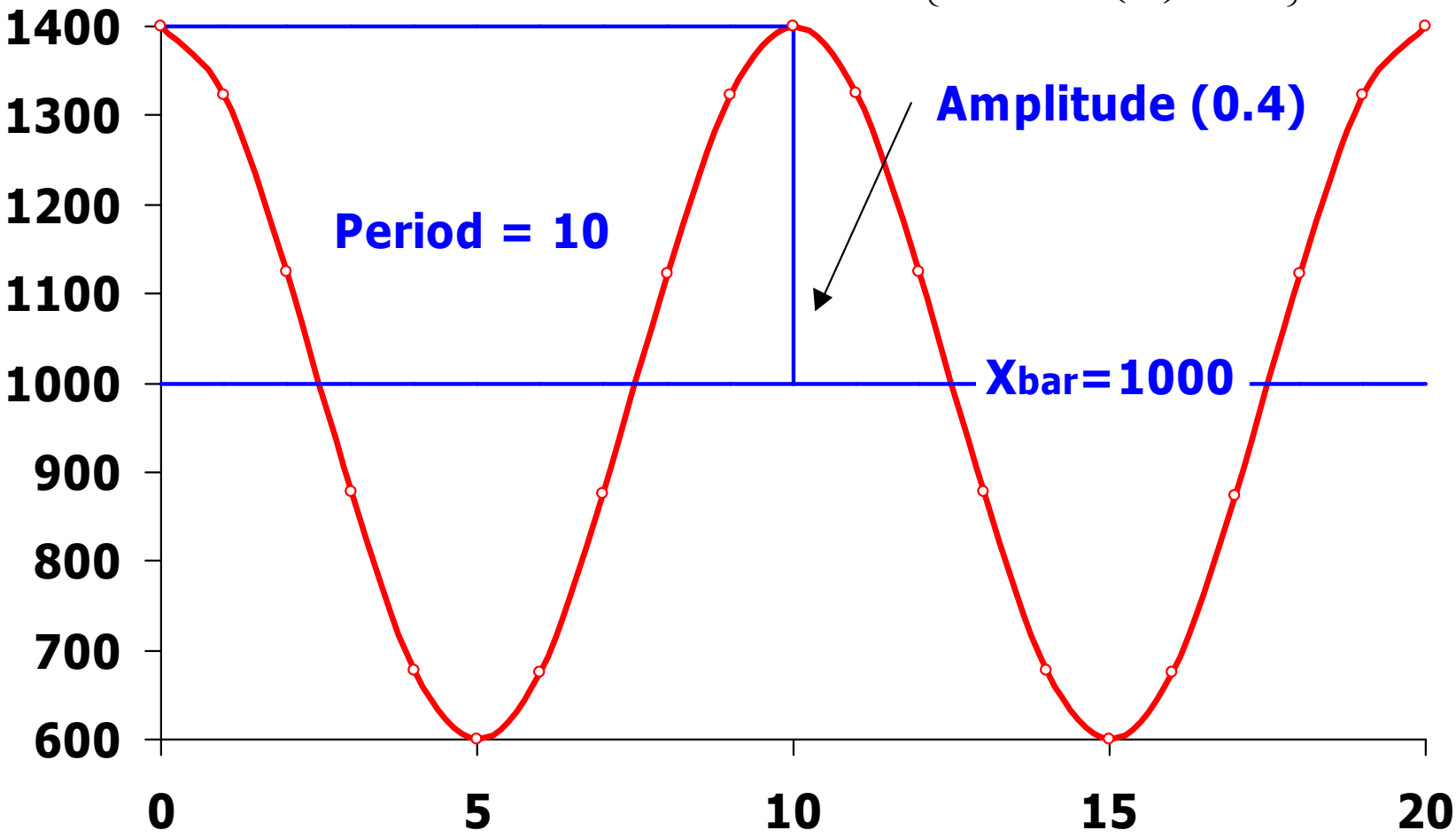
Periodic function

$$Y_t = \bar{Y}\left\{1 + A * \cos\left(\frac{T}{P}\right) * 6.28\right\}$$

- Ybar: Mean level (over the time of 1 period)
- A: Amplitude
- T: Time
- P: Period
- 6.28 -- convertion factor

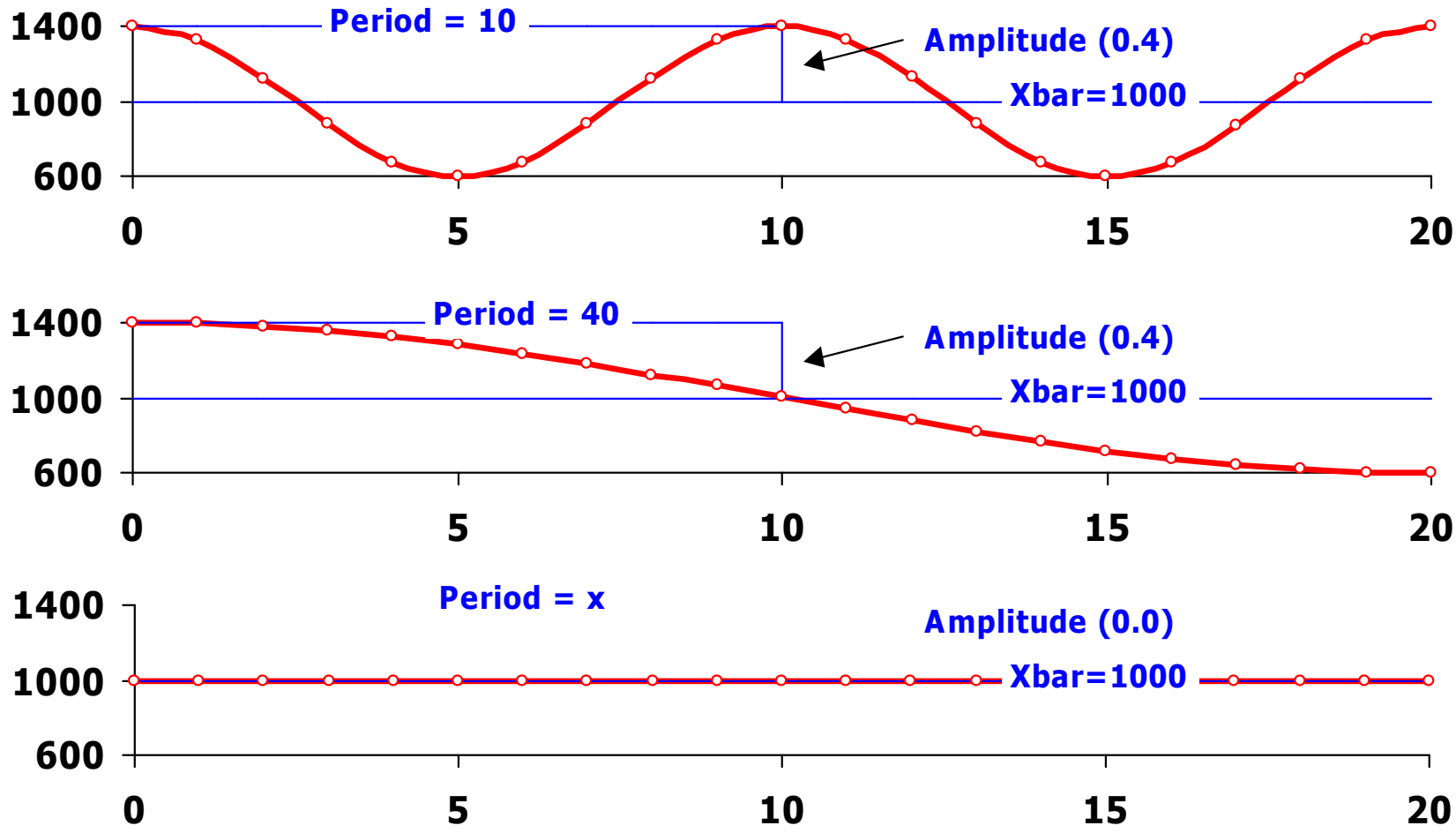
Periodic function: What it looks like

$$Y_t = \bar{Y}\left\{1 + A * \cos\left(\frac{T}{P}\right) * 6.28\right\}$$



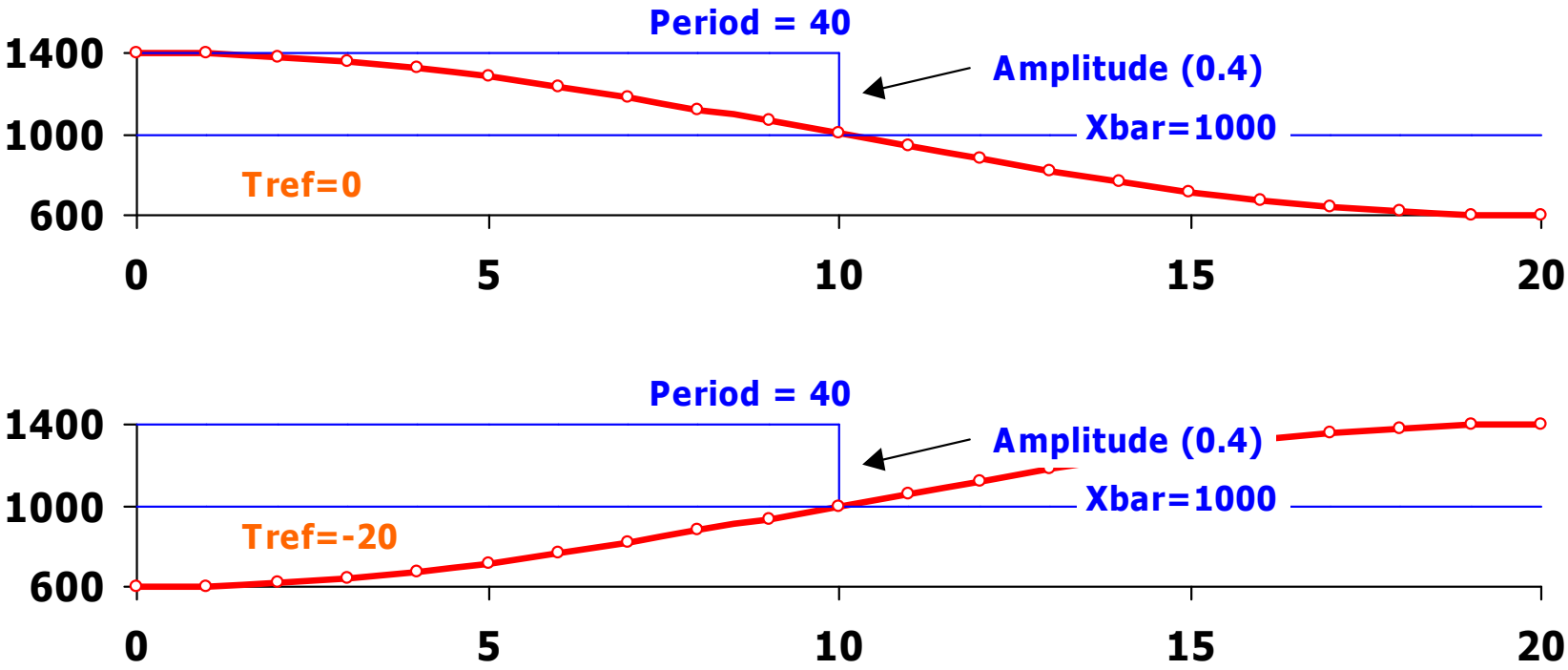
Periodic function: Very flexible

$$Y_t = \bar{Y}\left\{1 + A * \cos\left(\frac{T}{P}\right) * 6.28\right\}$$



Periodic function: Phase shift

$$Y_t = \bar{Y}\left\{1 + A * \cos\left(\frac{T - T_{ref}}{P}\right) * 6.28\right\}$$

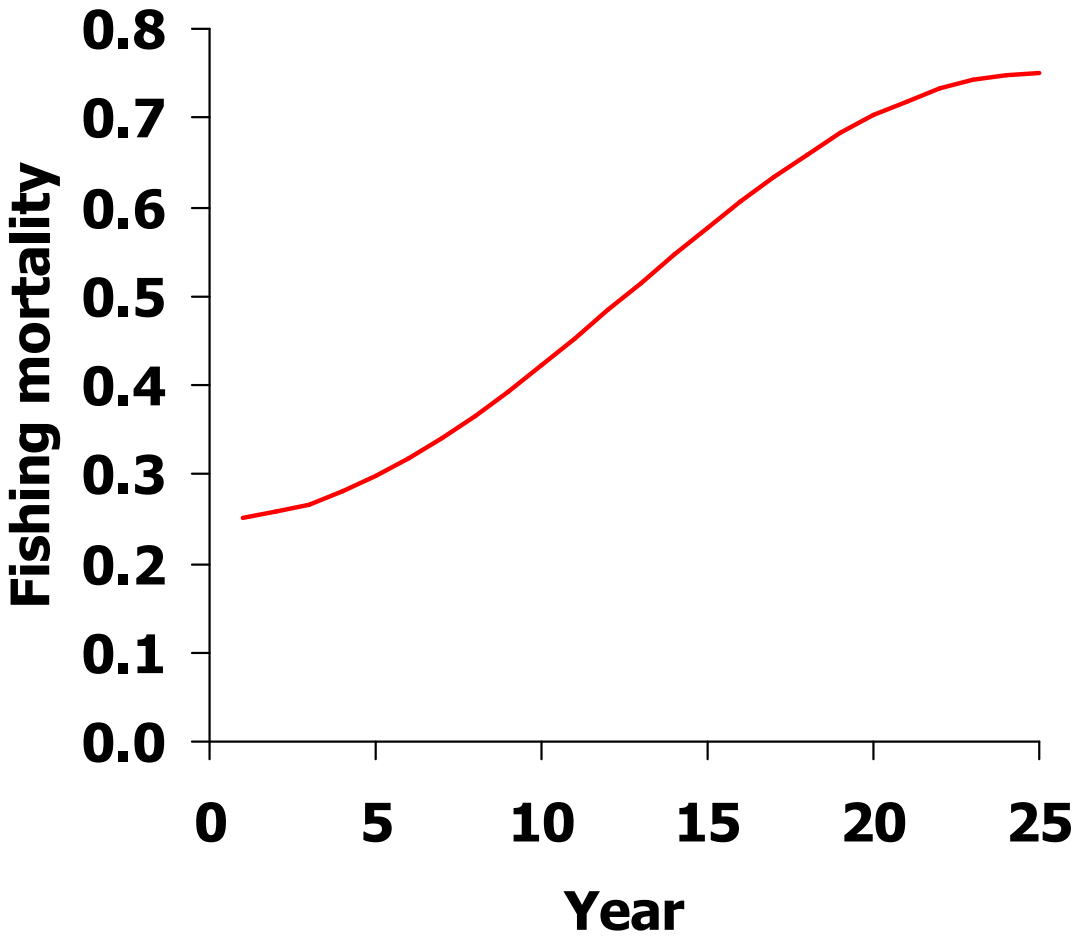


Development in fishing mortality (Fy)

einar

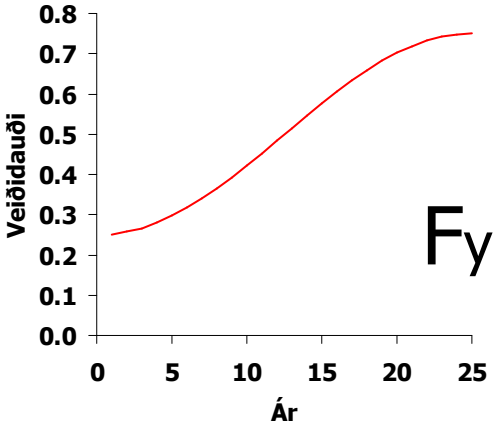
Fishing mortality	
	Parameters
A	0.5
P	50
Ybar	0.5
Tshift	-25

Year	Fy
1	0.25
2	0.26
3	0.27
4	0.28
5	0.30
6	0.32
7	0.34
8	0.37
9	0.39
10	0.42
11	0.45
12	0.48
13	0.52
14	0.55
15	0.58
16	0.61
17	0.63
18	0.66
19	0.68
20	0.70
21	0.72
22	0.73
23	0.74
24	0.75
25	0.75

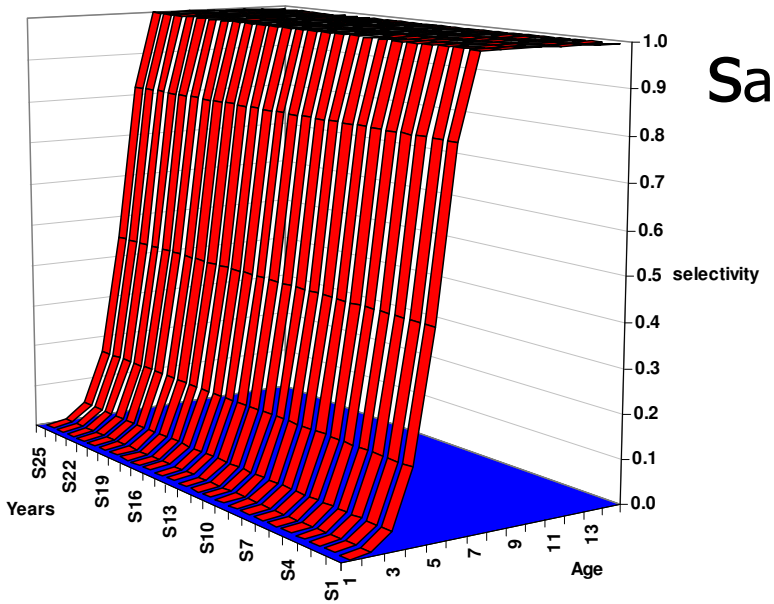


Fishing mortality by age and time: Fay

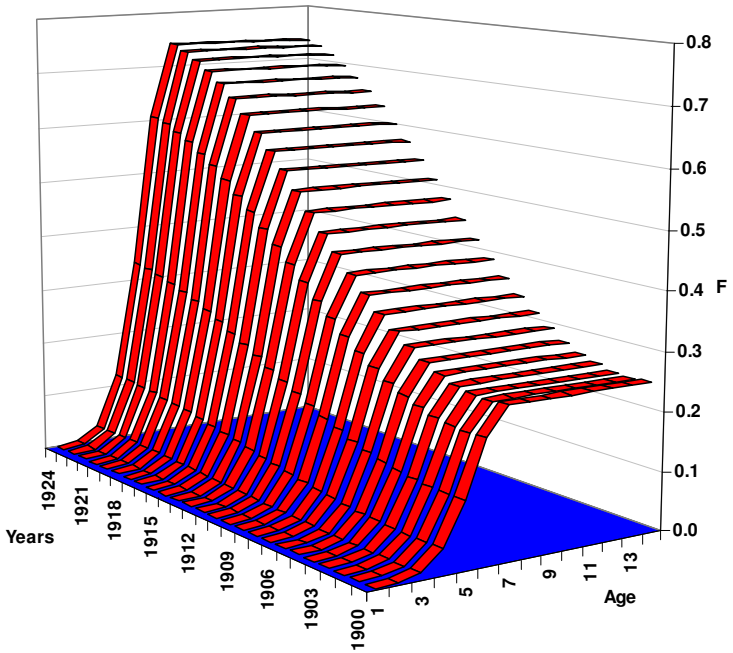
einar



Fy



Sa



Fay = SaFy

Note: Here selection pattern fixed

Selectivity - Logistic function

- Various forms describing the same thing.

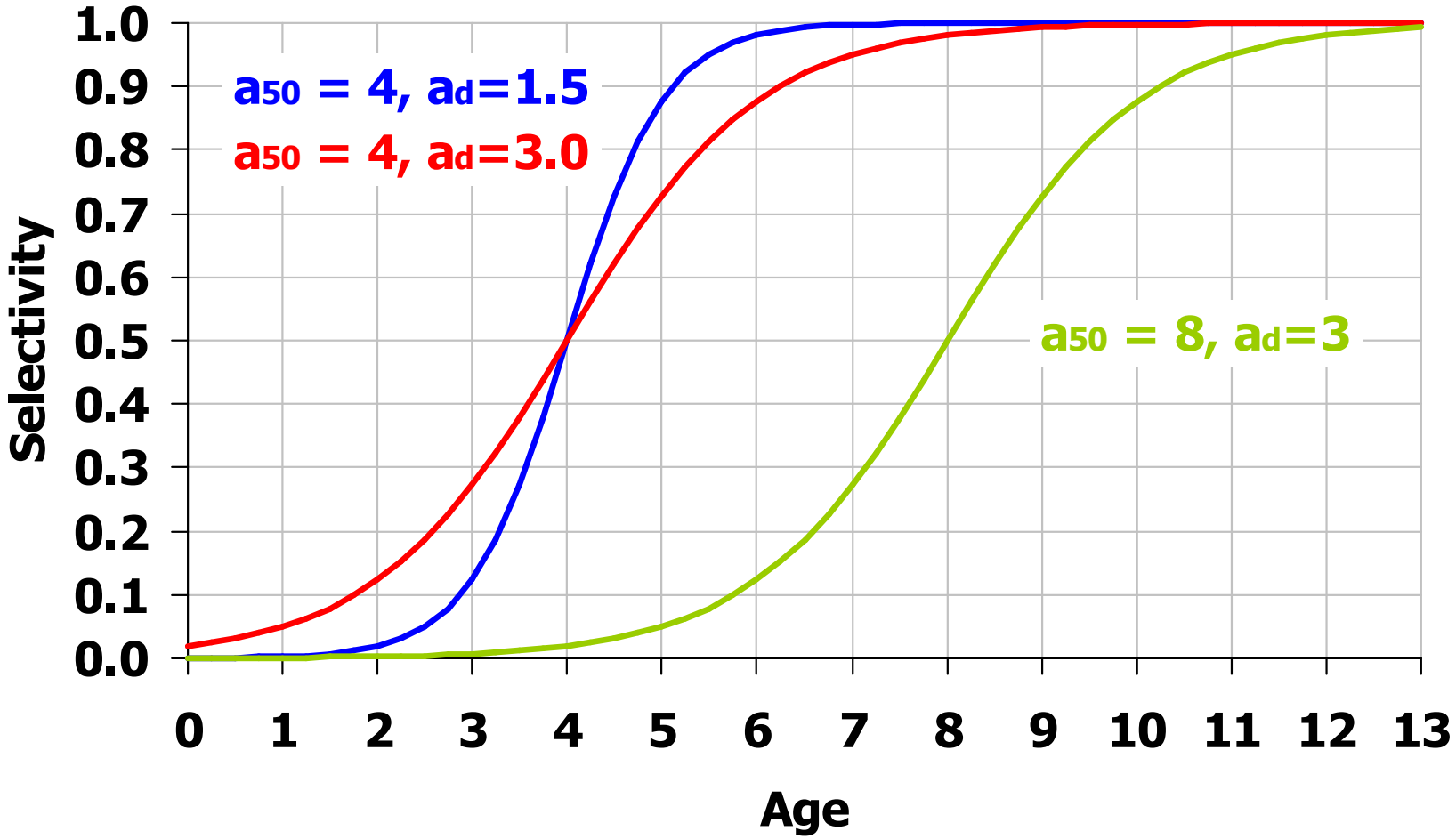
$$s_a = \frac{1}{1+e^{-k(a-a_{50})}}$$

$$s_a = \frac{1}{1+e^{-\ln 19 \frac{a-a_{50}}{a_{95}-a_{50}}}}$$

$$s_a = \frac{1}{1+e^{-\ln 19 \frac{a-a_{50}}{a_d}}}$$

Selectivity - Logistic

Note symmetry



Selectivity - double half-Gaussian

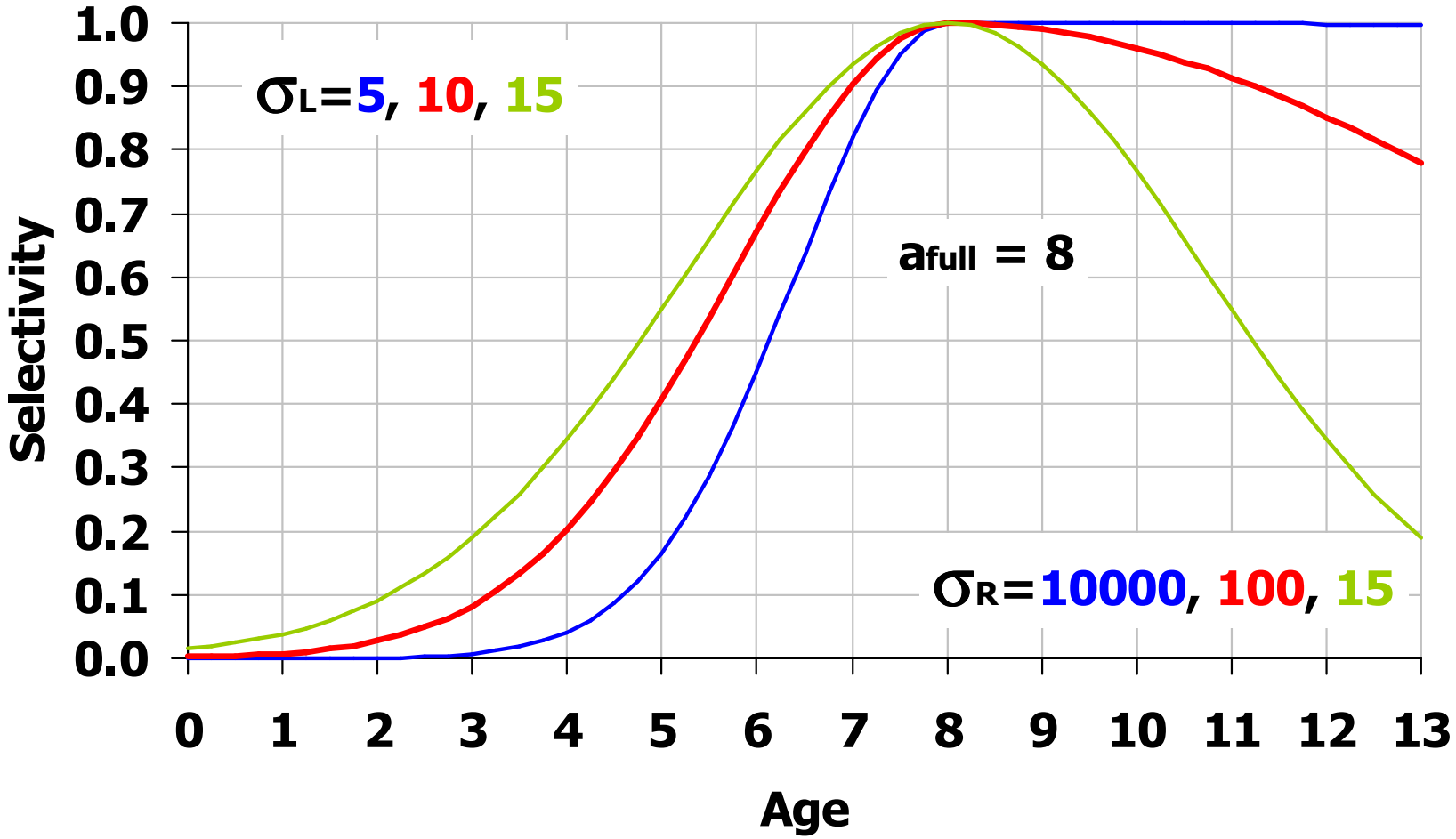
- Are simply two half-normal curves:

$$s_a = \begin{cases} e^{\frac{-(a-a_{full})^2}{\sigma_L}} & \text{for } a \le a_{full} \\ e^{\frac{-(a-a_{full})^2}{\sigma R}} & \text{for } a > a_{full} \end{cases}$$

  - a_{full}: age at full selectivity
  - σ_R: Shape factor (variance) for right hand curve
  - σ_L: Shape factor (variance) for left side of curve

Selectivity - double half-Gaussian

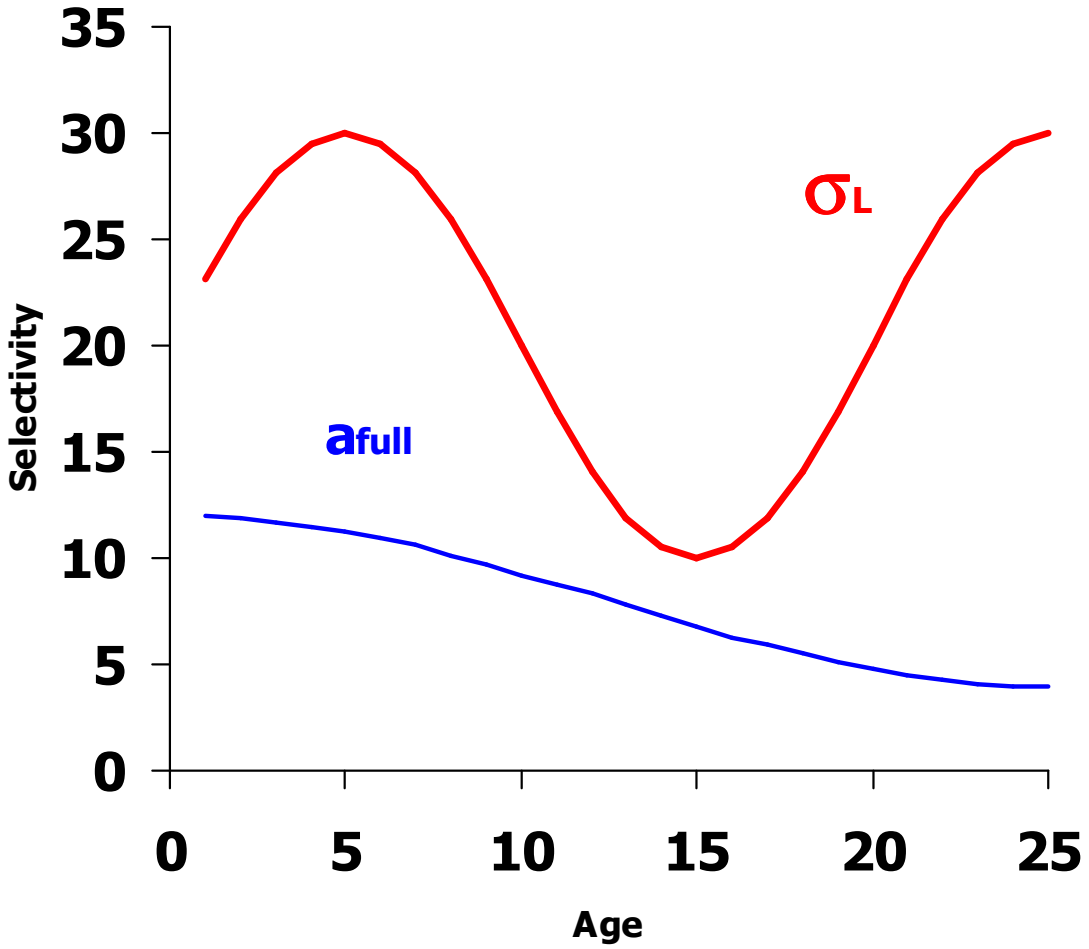
Note asymmetry



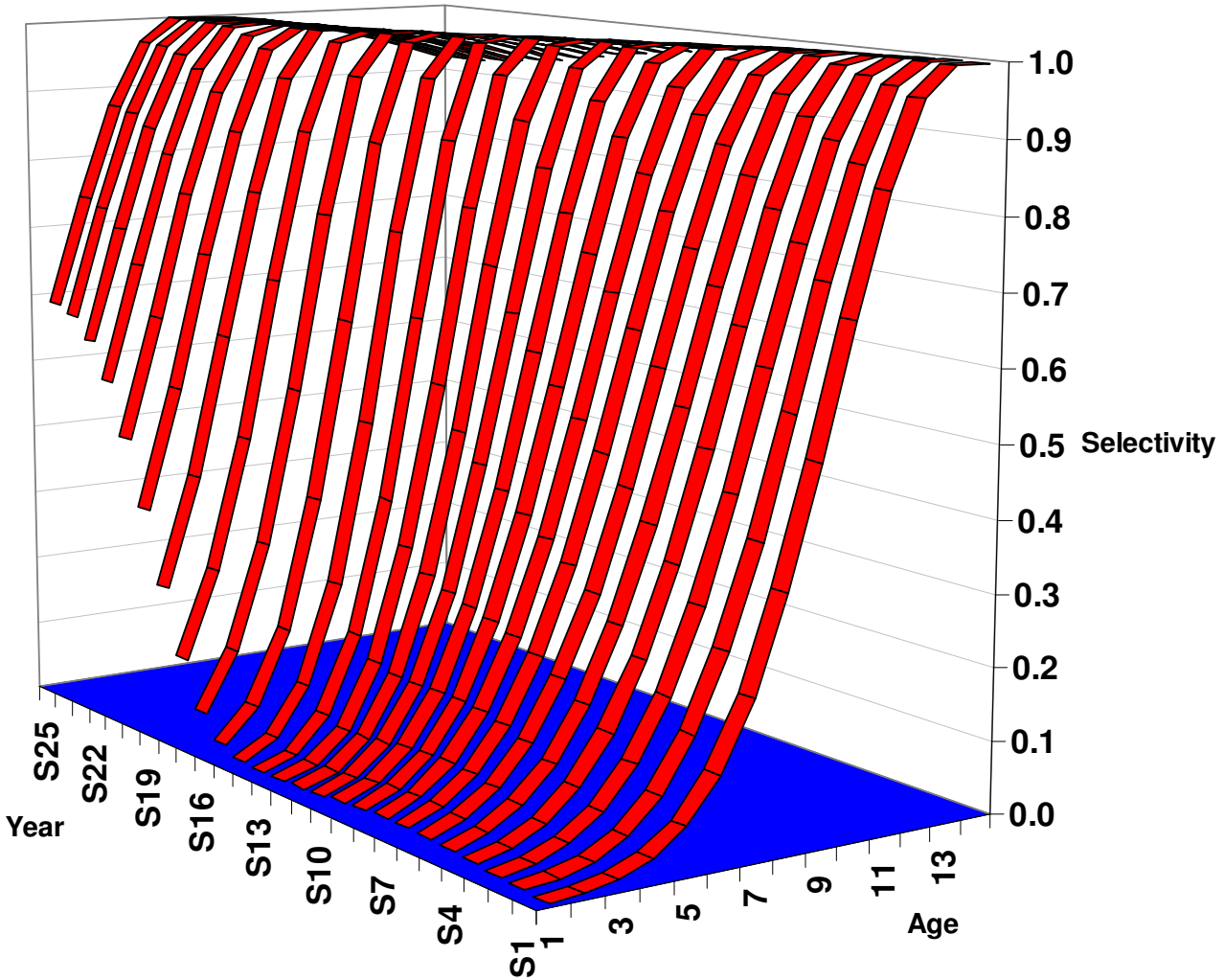
Selectivity - adding periodic function

Double half Gaussian selectivity			
	Afull	σL	σR
Amplitude	0.5	0.5	0
Period	50	20	50
Level	8	20	1000
Tshift	0	25	25

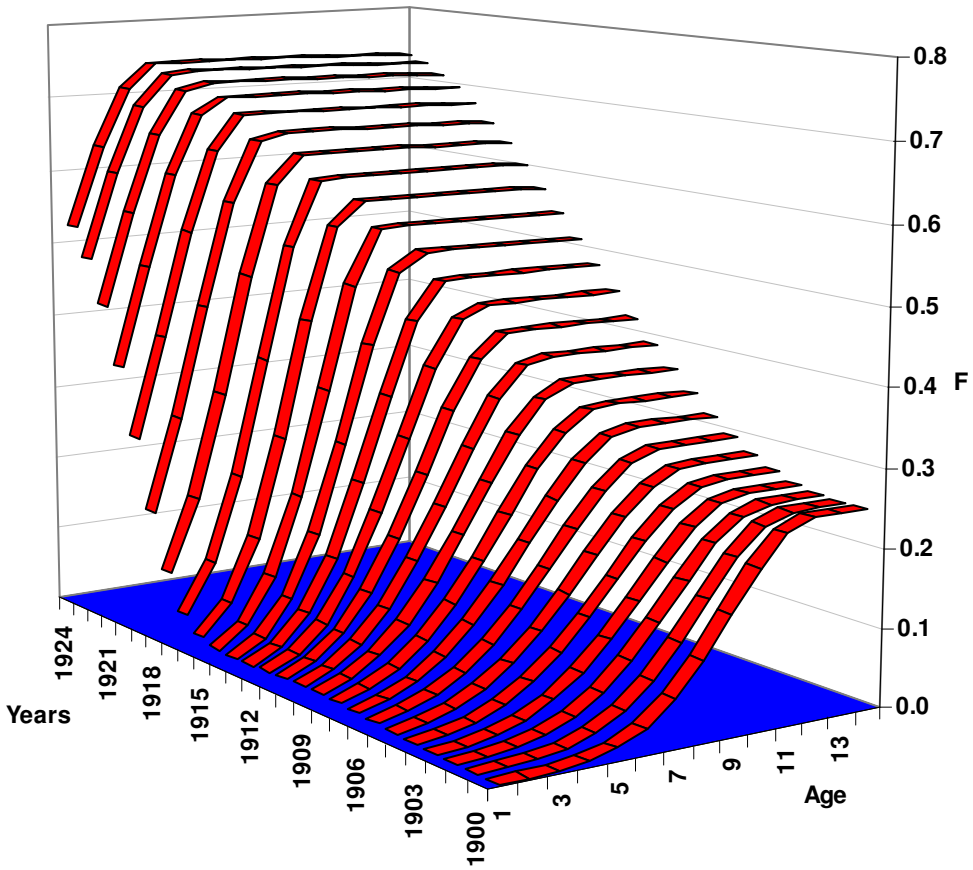
Year	Afull	σL	σR
1	12.0	23.1	1000
2	11.9	25.9	1000
3	11.7	28.1	1000
4	11.5	29.5	1000
5	11.2	30.0	1000
6	10.9	29.5	1000
7	10.6	28.1	1000
8	10.1	25.9	1000
9	9.7	23.1	1000
10	9.2	20.0	1000
11	8.8	16.9	1000
12	8.3	14.1	1000
13	7.8	11.9	1000
14	7.3	10.5	1000
15	6.8	10.0	1000
16	6.3	10.5	1000
17	5.9	11.9	1000
18	5.5	14.1	1000
19	5.1	16.9	1000
20	4.8	20.0	1000
21	4.5	23.1	1000
22	4.3	25.9	1000
23	4.1	28.1	1000
24	4.0	29.5	1000
25	4.0	30.0	1000



Resulting selectivity pattern



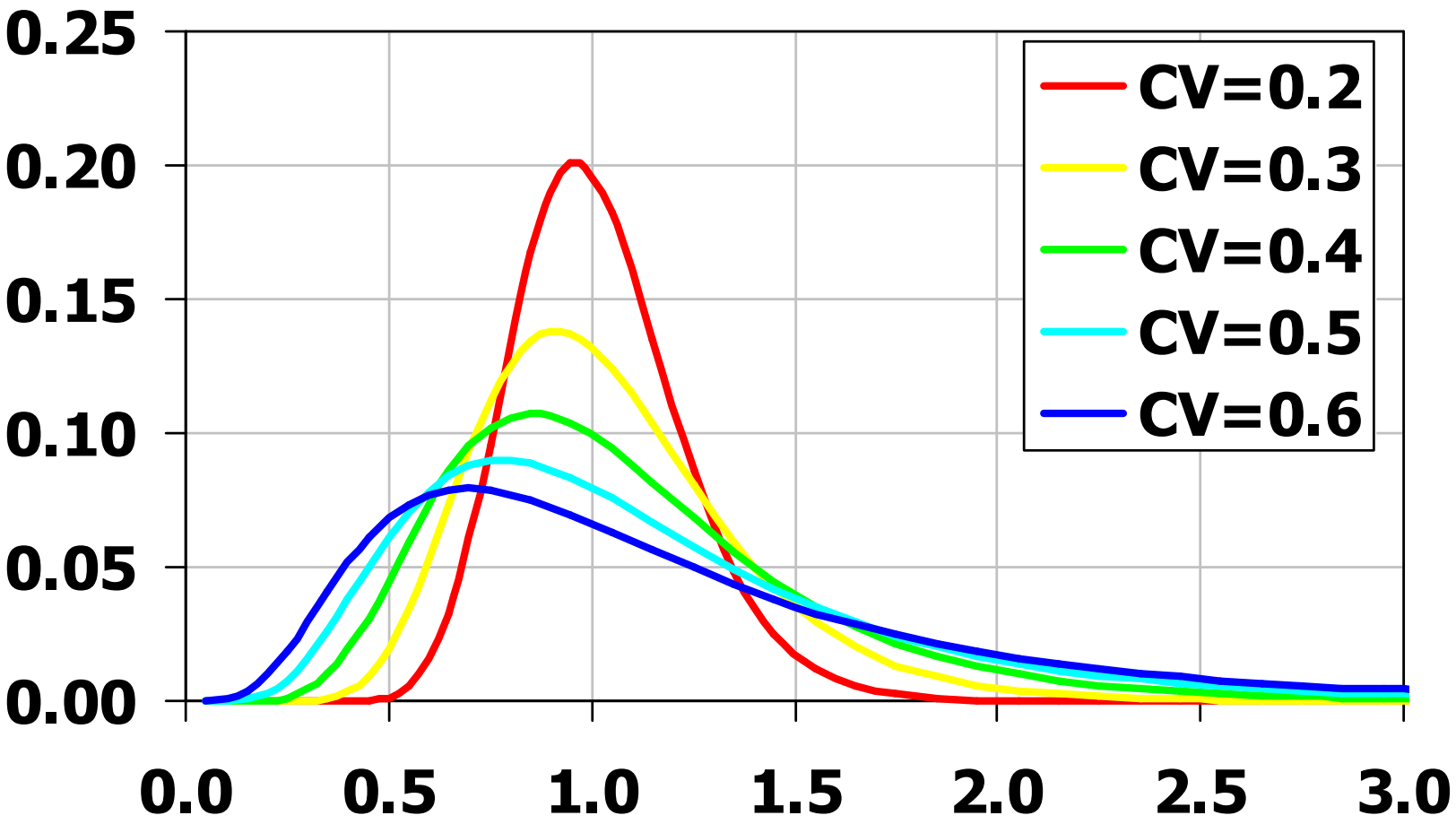
$F_{ay} = s_{ay} F_y$



Pattern generated with relatively few parameters

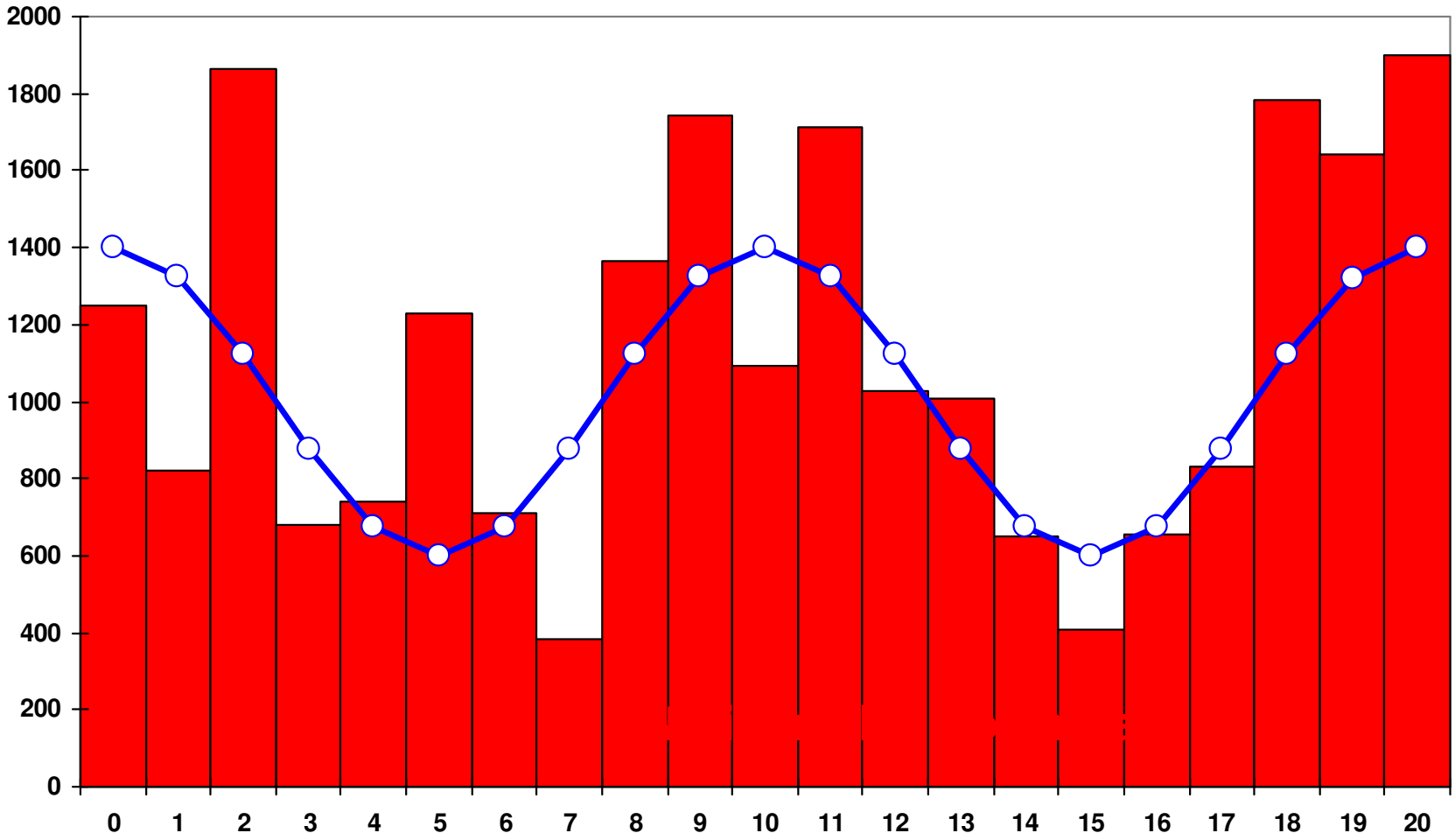
Adding stochastic / noise

XCEL SPEAK: = LOGINV(RAND();0;CV)



Recruitment: Deterministic -> stochastic

XCEL SPEAK: =Ry*LOGINV(RAND();0;CV)



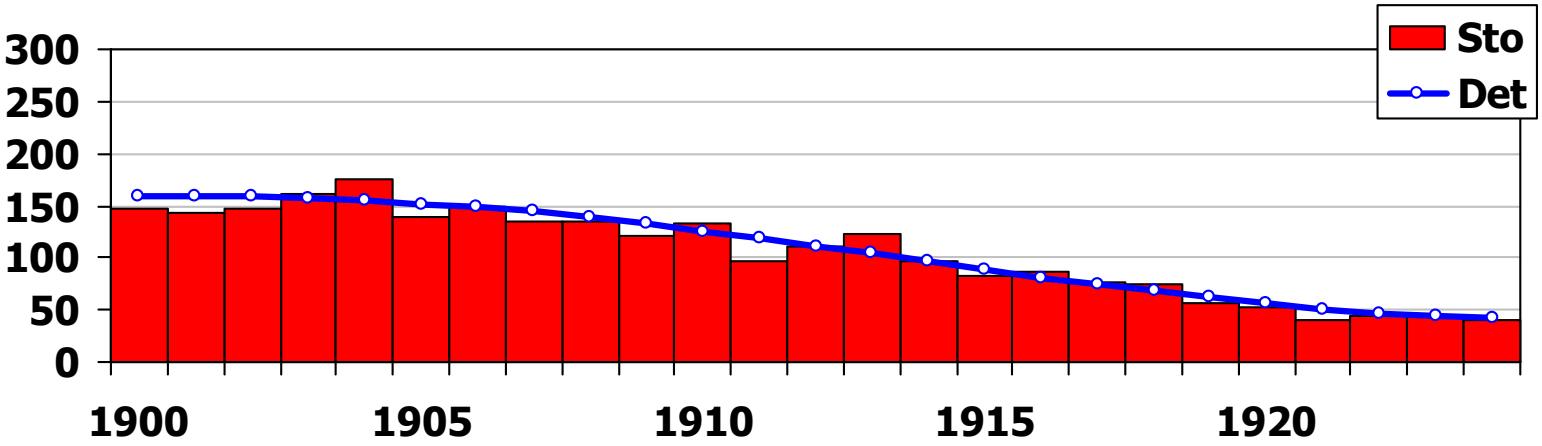
Stated more formally

$$N_{A,y}^{DET} = \overline{N}_{A,y}\left\{1 + A * \cos\frac{y - T_{initY}}{P} * 6.28\right\}$$

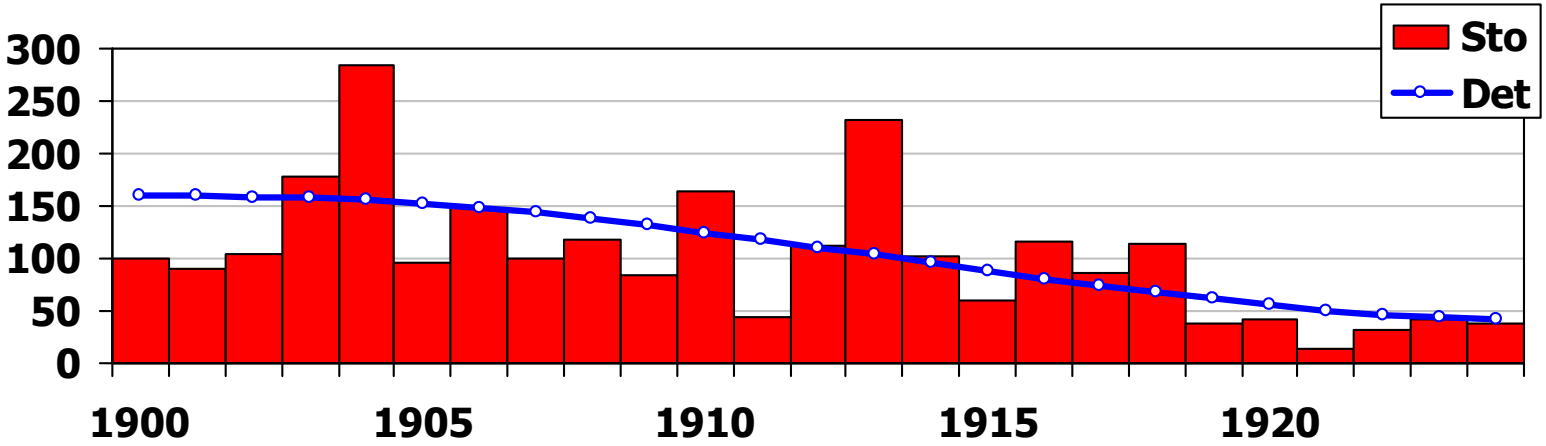
$$N_{A,y}^{STO} = N_{A,y}^{DET} e^{\varepsilon_R} \qquad \varepsilon_R \sim N\left(0, \sigma_R^2\right)$$

- ε_R normally distributed random number
  - Mean = 0
  - Standard deviation = σ_R
- Normally distributed numbers set to the power of e gives lognormal distribution.
  - Note! Loginv(rand(),0,s) = exp(norminv(rand(),0,s))
- By changing the value of σ_R we can control the variance in recruitment.
- Often refer to as σ_R CVs.

Recruitment: Different σ_R

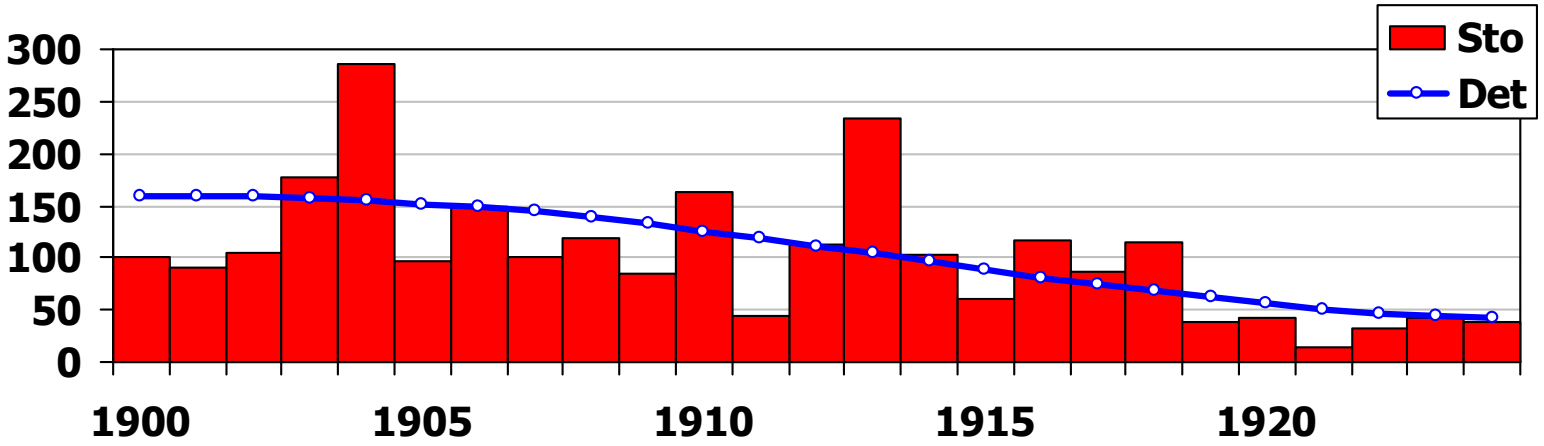


σ_R=0.1

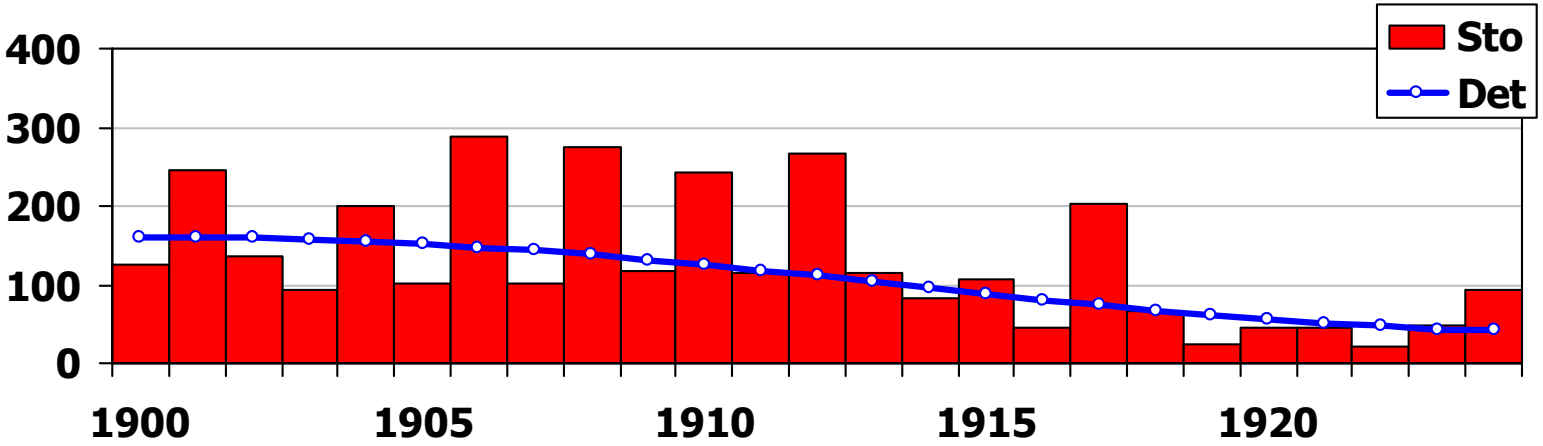


σ_R=0.5

Recruitment: Same σ_R, new random numbers



σ_R=0.5



σ_R=0.5

Simulating measurement error in C_{ay}

- Measurement error for Cay are often assumed log normally distributed
- In mathematical terms this is described as:

True catch

Measured catch

Lognormal random numbers

Normally distributed random numbers with: mean = 0 standard deviation = σ

$$C_{a,y}^{OBS} = C_{a,y}^{TRUE} e^{\varepsilon_{C_a}} \qquad \varepsilon_{C_a} \sim N\left(0, \sigma_{C_a}^2\right)$$

- In Excel speak:

$$C_{a,y}^{OBS} = C_{a,y}^{TRUE} \text{ x } \text{LOGINV}(\text{RAND}(),0,\sigma))$$

Generating error and bias in catches

$$C_{ay}^{TRUE} = \frac{s_a F_y}{s_a F_y + M_{ay}} \left(1 - e^{-(s_a F_y + M_{ay})}\right) N_{ay}$$

$$C_{ay}^{DET} = \left(1 - p_{ay}\right) C_{ay}^{TRUE}$$

$$C_{a,y}^{OBS1} = C_{a,y}^{DET} e^{\varepsilon_{C_a}} \qquad \varepsilon_{C_a} \sim N\left(0, \sigma_{C_a}^2\right)$$

$$C_{a,y}^{OBS2} = C_{a,y}^{OBS1} \frac{Y_y^{DET}}{Y_y^{OBS1}}$$

$$Y_y^{DET} = \sum_a C_{a,y}^{TRUE} w_{a,y}^{catch}$$

$$Y_y^{OBS1} = \sum_a C_{a,y}^{OBS1} w_{a,y}^{catch}$$

- **The actual catch taken**
- **The actual catch**
- **Measured catch, lognormal error**
- **Scaling to total yield**
- **Formula for total yield**

p_{ay}: proportion of unrecorded catch

Survey indices vs. stock size

- Simple model $U_{ay} = q_a N_{a,y}$
- More complex $U_{ay} = q_a N_{a,y}^{\beta_a}$
  - q_a: catchability for age a
  - β_a: power parameter for age a
    - NB, if $\beta=1$ it is a simple model
- Simulating measurement errors

$$U_{ay}^{OBS} = U_{ay}^{TRUE} e^{\varepsilon_{U_a}} \qquad \varepsilon_{U_a} \sim N(0, \sigma_{U_a})$$

- In the simulator we can also add year effect and q –trends.
  - Lets not worry to much about that for now

The simulator in math (almost)

einar

$$N_{a,y} = \begin{cases} R_y = N^{STO}_{A.y} = N^{DET}_{A,y} e^{\varepsilon_R} & \varepsilon \sim N\left(0, \sigma^2_R\right) \\ N_{a-1,y-1} e^{-(s_{a-1}F_{y-1} + M_{a-1,y-1})} & \\ N_{a-1,y-1} e^{-(s_{a-1}F_{y-1} + M_{a-1,y-1})} + N_{a,y-1} e^{-(s_a F_{y-1} + M_{a,y-1})} & \end{cases} \qquad \begin{cases} a = A \\ A < a \le a_{plus} \\ a = a_{plus} \end{cases}$$

$$F_{ay} = s_a F_y$$

$$C^{TRUE}_{ay} = \frac{s_a F_y}{s_a F_y + M_{ay}} \left(1 - e^{-(s_a F_y + M_{ay})}\right) N_{ay}$$

$$C^{DET}_{ay} = \left(1 - p_{ay}\right) C^{TRUE}_{ay}$$

$$C^{OBS1}_{a,y} = C^{DET}_{a,y} e^{\varepsilon_{C_a}} \qquad \varepsilon_{C_a} \sim N\left(0, \sigma^2_{C_a}\right)$$

$$C^{OBS2}_{a,y} = C^{OBS1}_{a,y} \frac{Y^{OBS2}_y}{Y^{OBS1}_y}$$

$$Y_y = \sum_a C_{a,y} w^{catch}_{a,y}$$

$$R_y = f(time)$$

$$F_y = f(time)$$

$$s_{ay} = f(age, time)$$

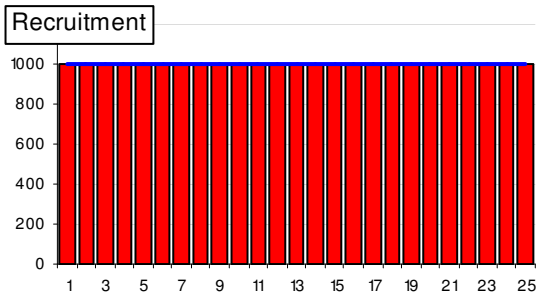
$$q_a = f(age, time)$$

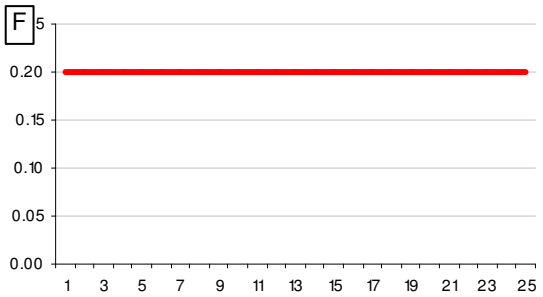
$$M_{ay} = f(?)$$

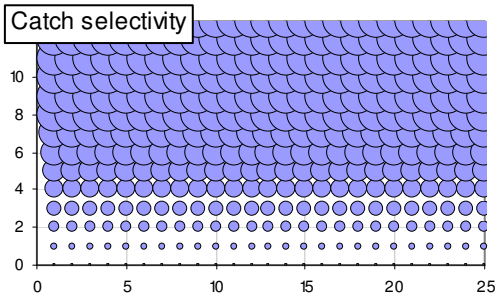
$$U^{TRUE}_{ay} = q_a N^{\beta_a}_{ay}$$

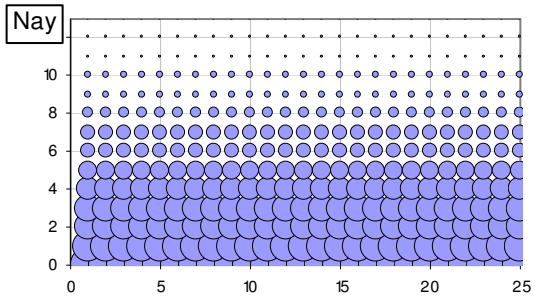
$$U^{OBS}_{ay} = U^{TRUE}_{ay} e^{\varepsilon_{U_a}} \qquad \varepsilon_{U_a} \sim N(0, \sigma_{U_a})$$

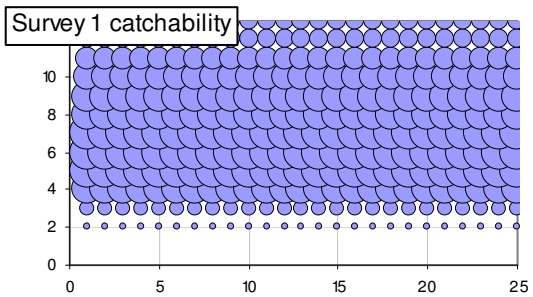
Example of simulations: Static system

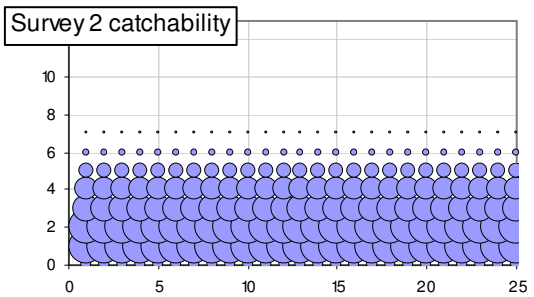


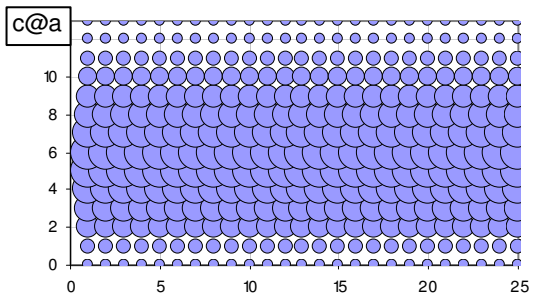


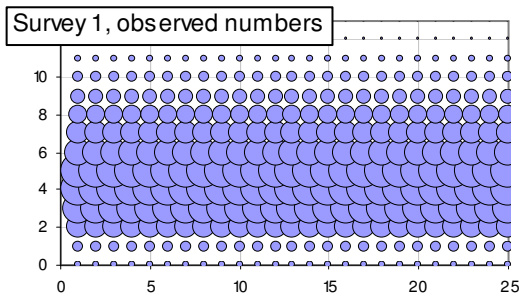


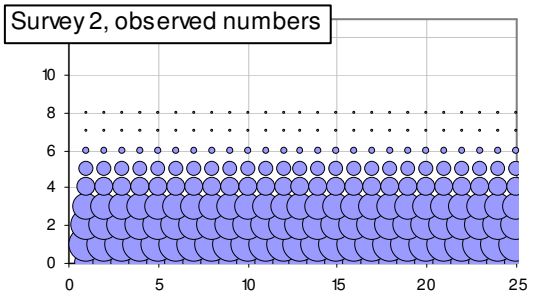




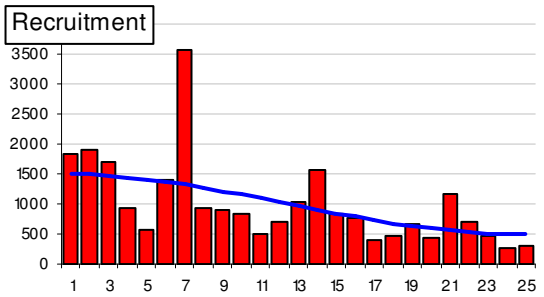


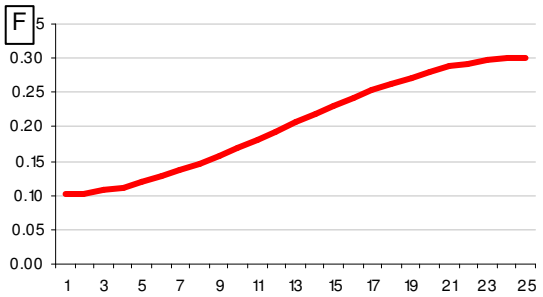


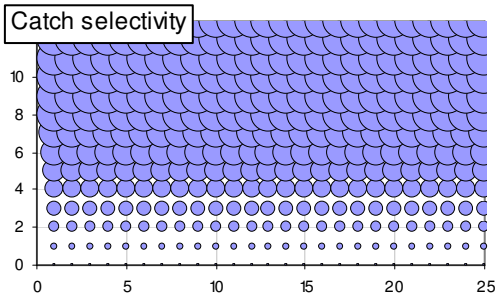


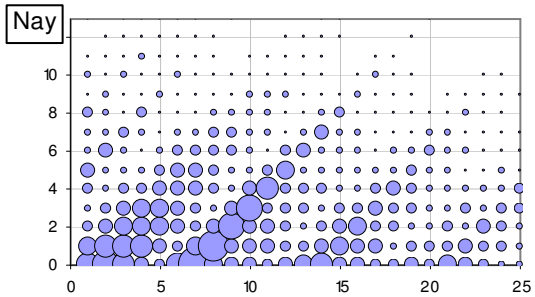


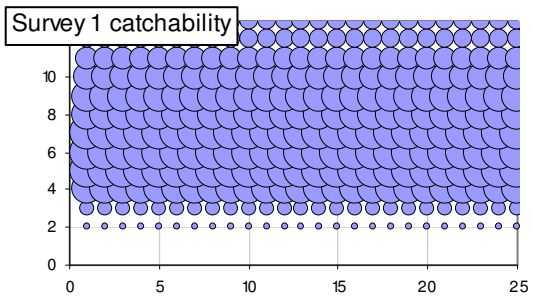
Adding stochasticity & time trends

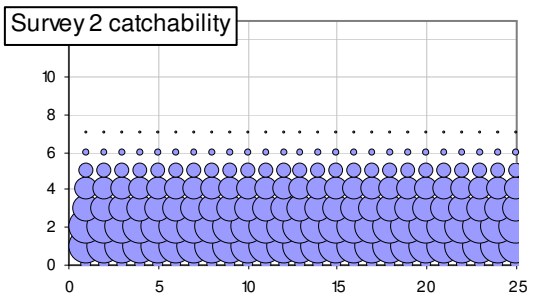


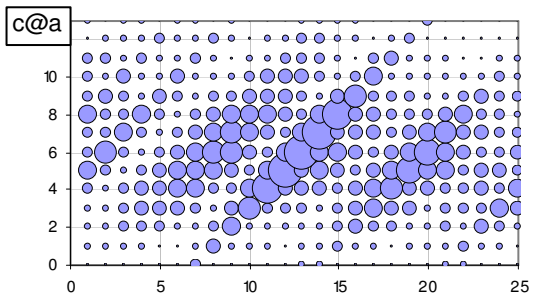


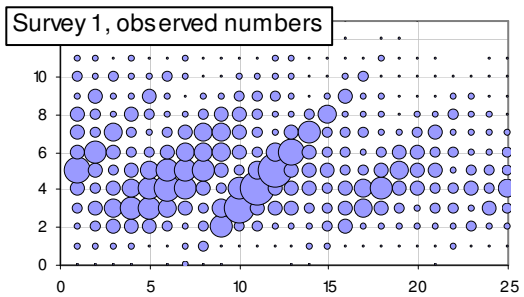


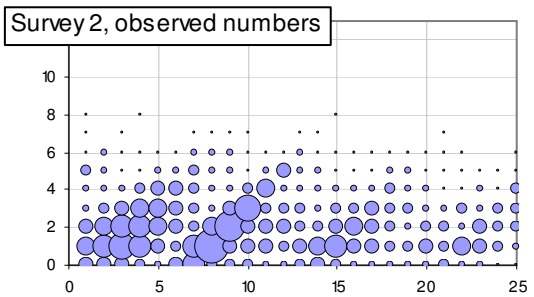






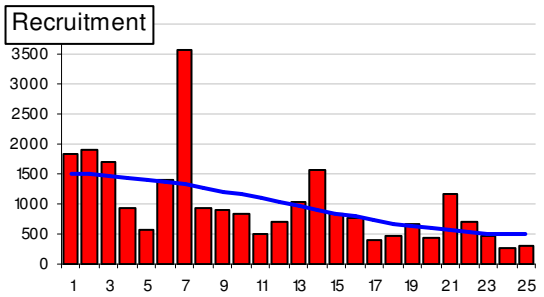


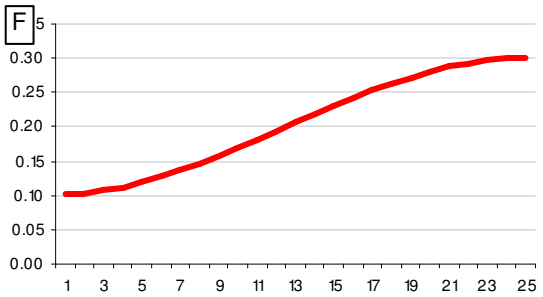


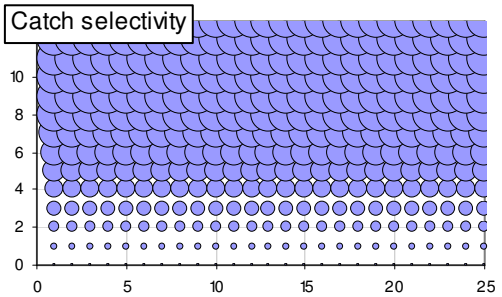


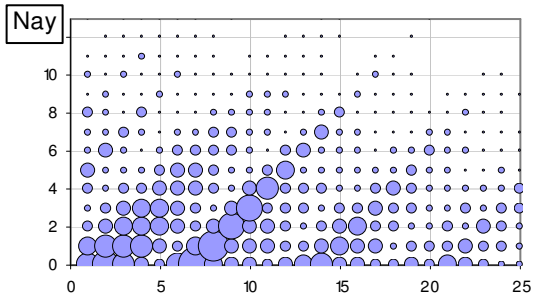
Adding observation noise

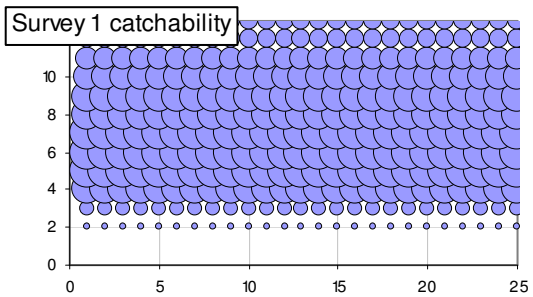
einar

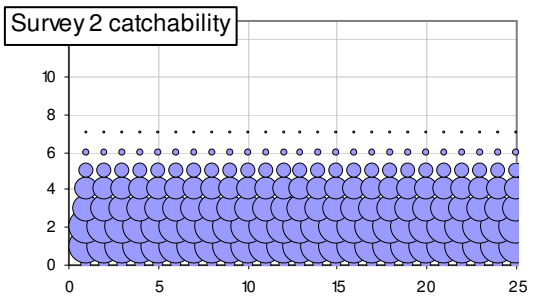


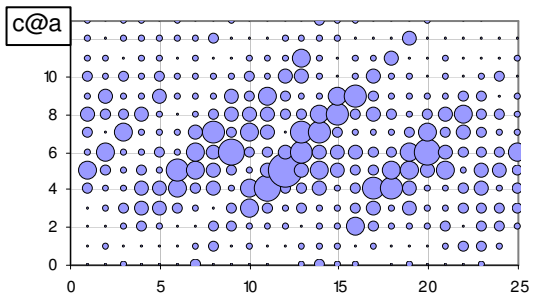


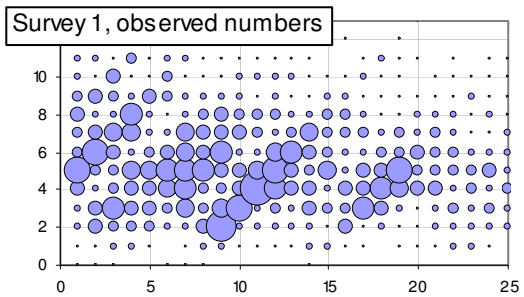


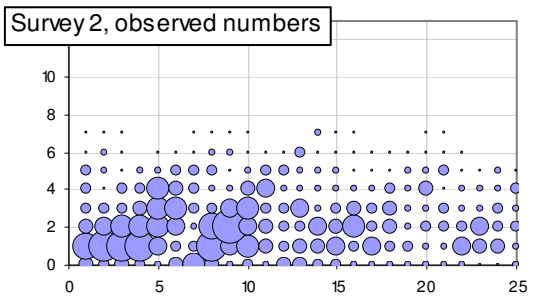




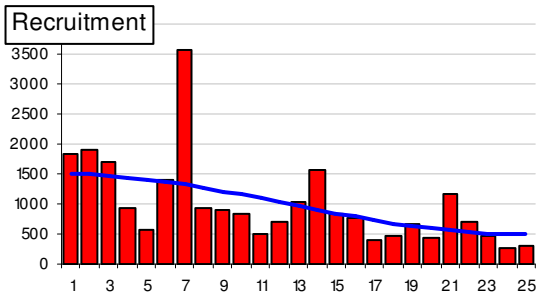


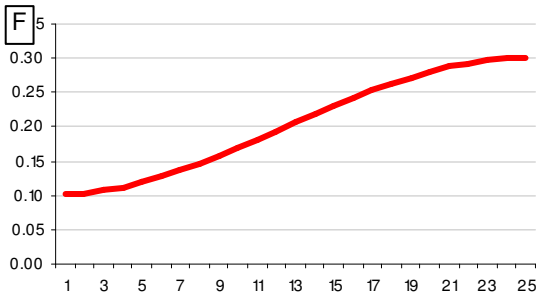


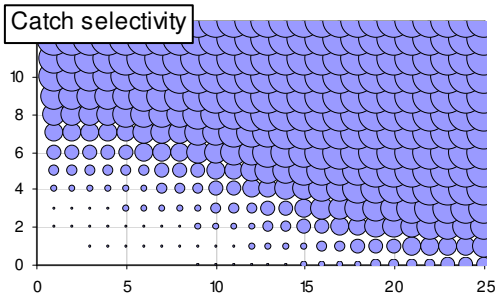


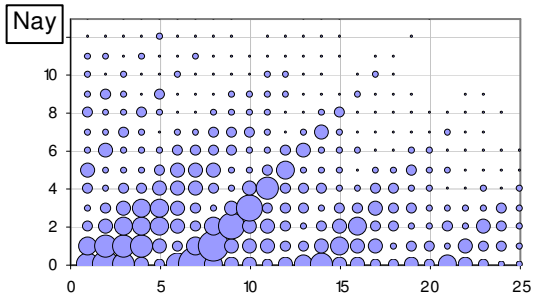


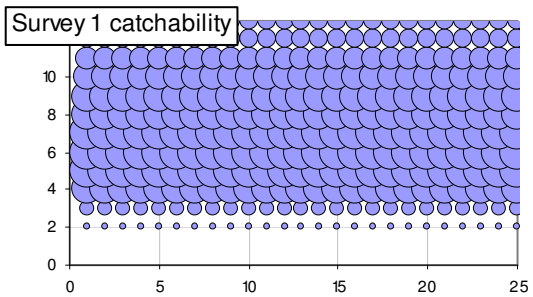
Adding change in fleet selectivity

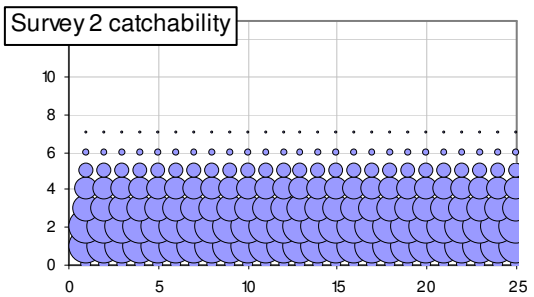


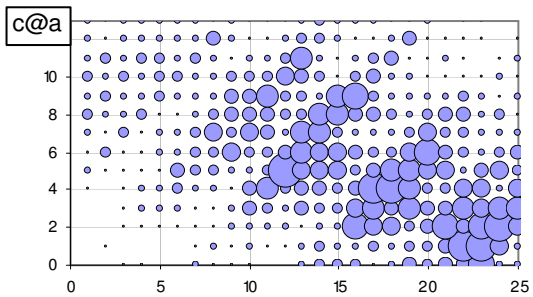


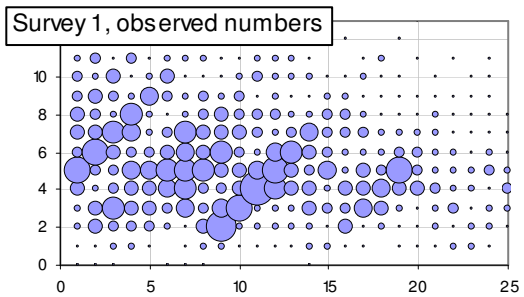


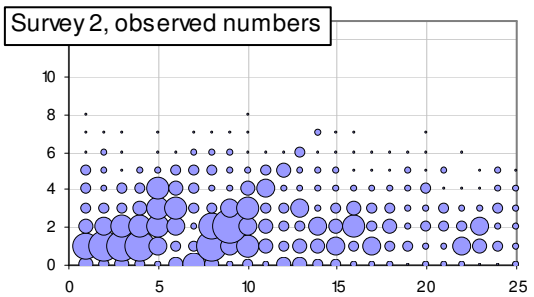




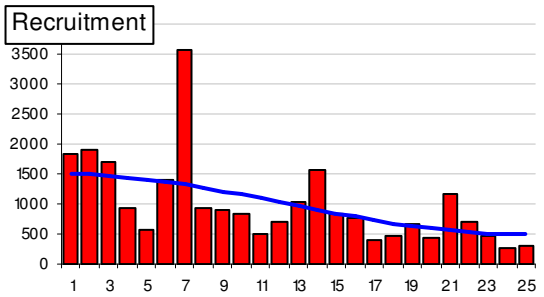


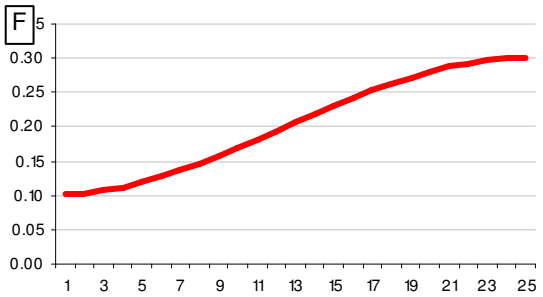


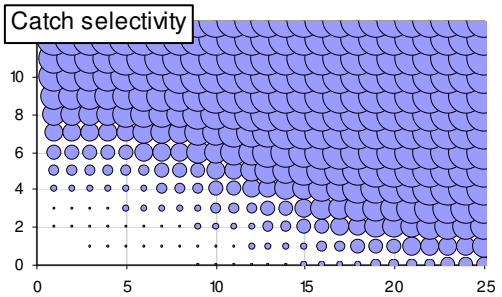


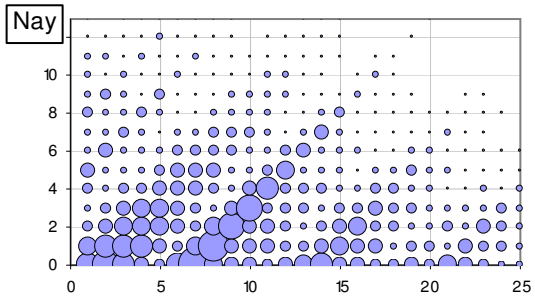


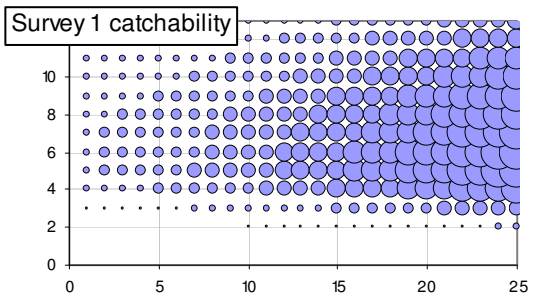
Adding trend in survey 1 (q increases)

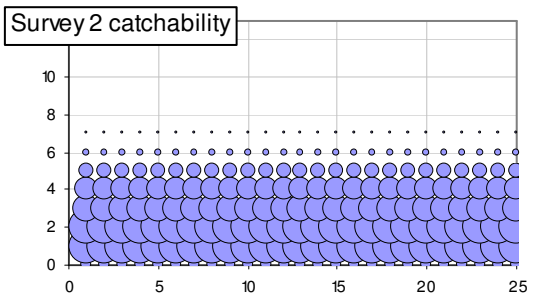


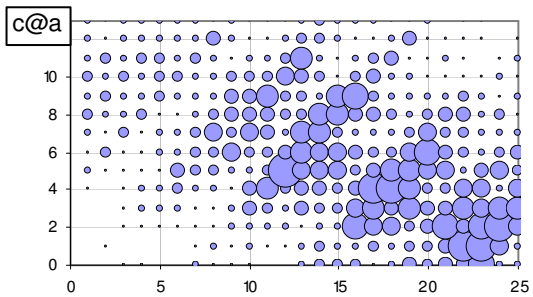


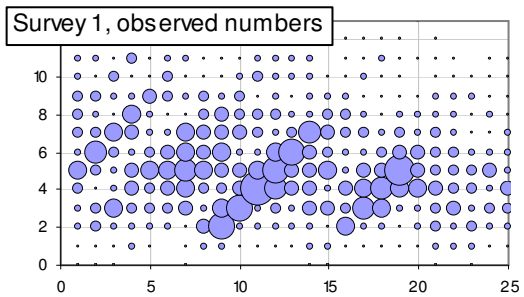


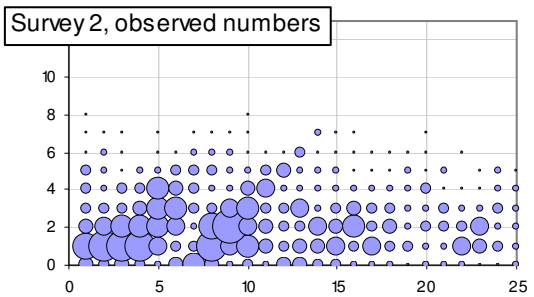




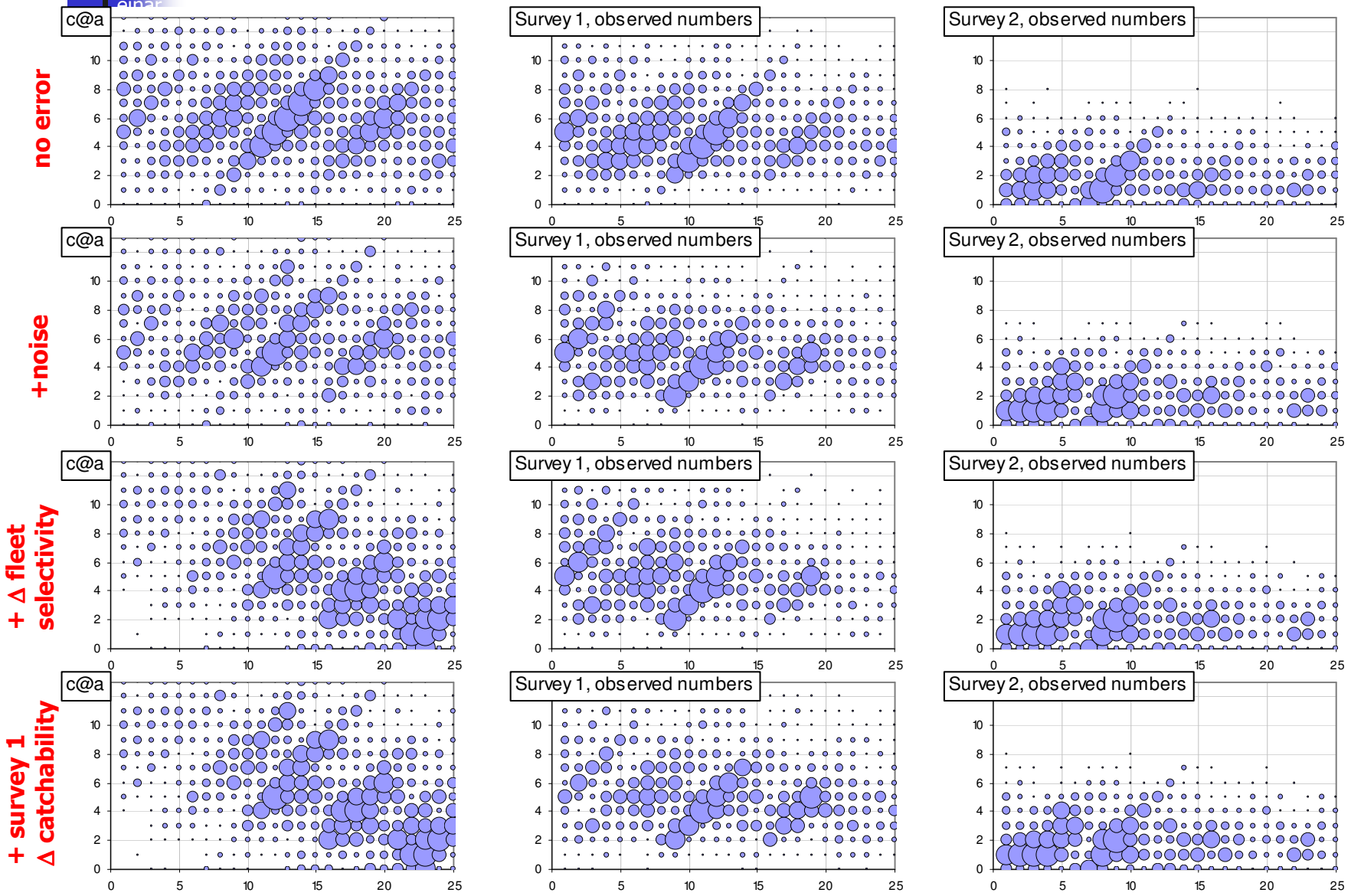








Same R input - different observation



A summary of things done

What we have: Part 1

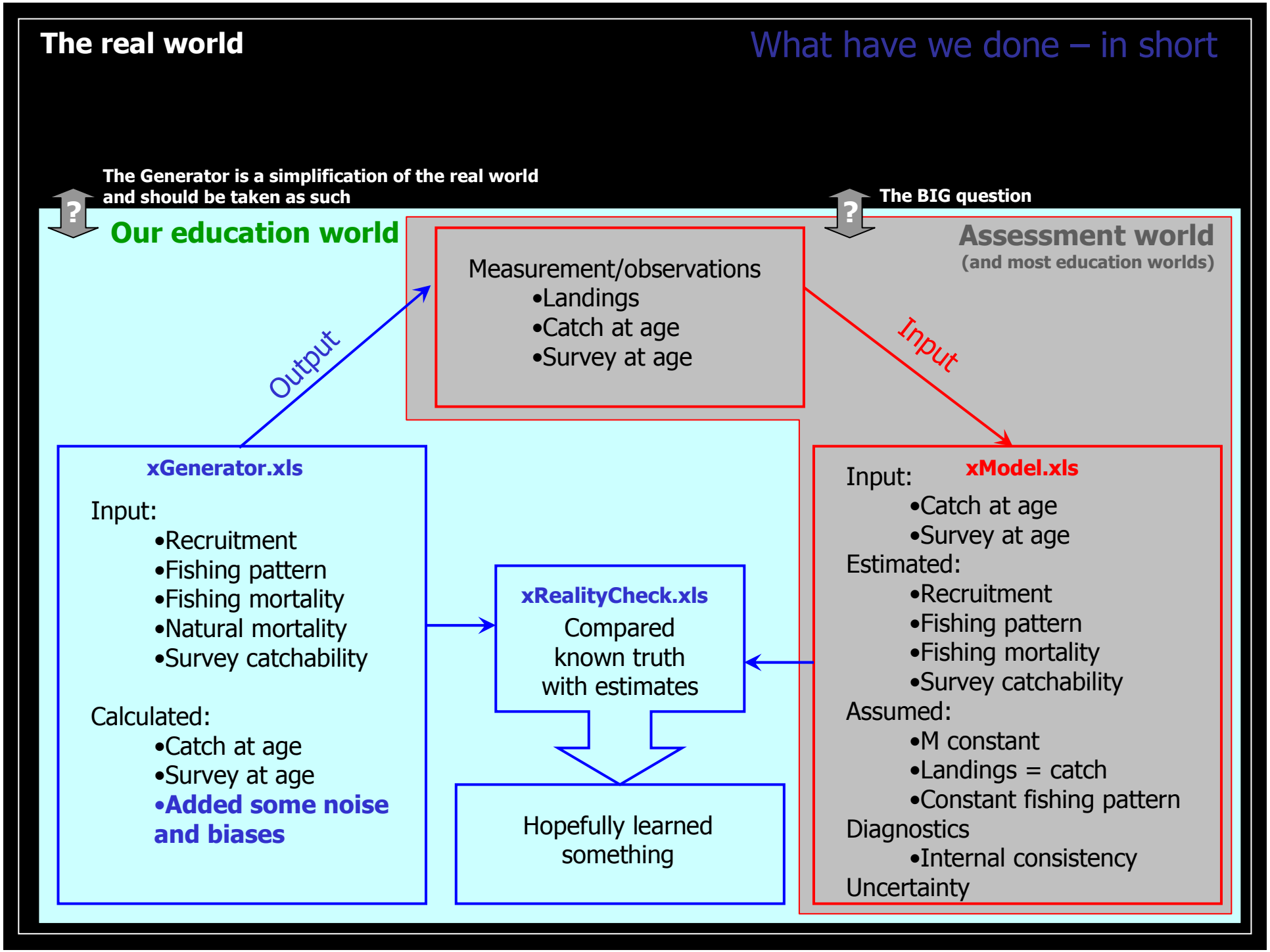
- Population, fisheries and observation simulation – Operational model
  - Introduced the concept and the mathematics for simulating population dynamics, fisheries and typical observations in "a data rich" situations.
  - Used a spreadsheet (xGenerator.xls) to get a hands on feel with the intent to improve the understanding of the concept and mathematics in the simulator.
  - Generated observables with given measurement errors:
    - One c@a matrix – catch at age by years in numbers (Cay)
    - Two u@a matrices – age based survey indices (Uay)
    - Note: These simulated measurements could be used in any conventional off-the shelf assessment software package, be it XSA, ADAPT, TSA, ADCAM, AMCI, ICA, …

What we have: Part 2

- Set-up a simple separable assessment model (xModel.xls)
  - Use observation (Cay, Uay), generated with the simulator to estimated stock size (Nay) and fishing mortalities (Fay), selection pattern, and survey catchability.
    - Note: Could have used actual observations (Cay, Uay) from any of the stocks where we have c@a and age based survey indices, e.g. Icelandic cod, Southern hake, …
    - In our case we are however more interested in understanding how assessment software works, where it goes wrong and how it goes wrong. For that purpose we use the observation generated in the simulator.

What we have: Part 3

- Provided a spreadsheet (xRealityCheck.xls) that made a comparison on some key feature (recruitment, SSB, selection pattern, fishing mortalities) as simulated in the generator ("the truth") with that estimated from the Cay and Uay observation in the assessment model.


The real world
What have we done – in short
The Generator is a simplification of the real world and should be taken as such
The BIG question
Our education world
Assessment world
(and most education worlds)
Measurement/observations
•Landings
•Catch at age
•Survey at age
Output
Input
xGenerator.xls
Input:
•Recruitment
•Fishing pattern
•Fishing mortality
•Natural mortality
•Survey catchability
Calculated:
•Catch at age
•Survey at age
•Added some noise and biases
xRealityCheck.xls
Compared known truth with estimates
Hopefully learned something
xModel.xls
Input:
•Catch at age
•Survey at age
Estimated:
•Recruitment
•Fishing pattern
•Fishing mortality
•Survey catchability
Assumed:
•M constant
•Landings = catch
•Constant fishing pattern
Diagnostics
•Internal consistency
Uncertainty

Just to make sure we have this right

- As it currently is set up the only information that is passed from the xGenerator to the xModel are the observed catch at age and observed survey at age, including noise. I.e. any other settings are NOT passed from the xGenerator to the xModel.
  - The xModel thus does not know anything about recruitment, selection patterns, mortalities (including unaccounted ones) or survey catchabilities that were used to generate the observations in the xGenerator.xls
- In our daily world we are effectively limited to the "Assessment world".
- The generator is a simplified, and possibly a wrong reflection of the real world, the assessment ever more so.

Assignment work

- You will be given a set of exercises, which you are expected to keep notes on and make a report on, to be evaluated as you assignment.
- How to proceed:
  - Keep a backup of the original sets of Excel spreadsheets
  - Make a separate folder for each exercise
    - xGenerator.xls
    - xModel.xls
    - xRealityCheck.xls
- Keep good notes from classroom discussion, they will help in writing the report

Exercise 1: The effect of discards

- One of the biggest problems in fisheries science is that one normally has only estimates of actual landings but the true removal is unknown. Discards of a target species is normally assumed to be size related, the rate of discards normally being confined to the smallest fishes.
- Part 1. Start by fitting the model (run solver) without discards.
  - Discarding is controlled in the simulator, worksheet CatchAtAge in area starting in cell B90.
  - For this part you want to make sure that all the values are = 0.0
- Part 2. Set fix discard rate on younger ages through the whole time period.
  - Discarding is controlled in the simulator, worksheet CatchAtAge in area starting in cell B90.
  - Make sure that there is actually some fishing taking place on these younger ages!
  - Suggestion for discard rate:
    - Age 1 & 2: 0.99 – there is no fishing taking place to speak of in these age groups, see the selection pattern (worksheet xx)
    - Age 3: 0.75
    - Age 4: 0.50
    - Age 5: 0.25
    - Age 6: 0.10
- Can one detect any abnormal (nonrandom) patterns in residuals in either fit? Where are the major differences in the parameter estimates of the two fits?
- Compare the fit in Part 2 with that of your simulated population. Which of the key population measures (recruitment, biomass (SSB) and fishing mortality) deviate mostly from the known truth? Explain why this pattern is observed?
- Part 3. Feel free to try any other combination on discard rates.

Further exercises

- Tomorrow we will work further with the tools provided. We will study how:
  - Changing discard rate with time
  - Underreporting
  - Change in the catchability of the survey
  - Values of M
  - , and other factors
- affect our assessment results.
- The brave ones are encourage to proceed right away.