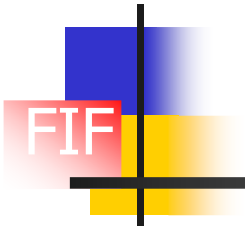


An introduction to bootstrap

Einar Hjörleifsson

The background

Getting something from nothing?

In Rudolph Erich Raspe's tale, Baron Munchausen had, in one of his many adventurous travels, fallen to the bottom of a deep lake and just as he was to succumb to his fate he thought to pull himself up by his own BOOTSTRAP.

The original version of the tale is in german, where Munchausen actually draws himself up by the hair, not the bootstraps. The figure on the left refers to german version of the story.

Efron and LePage gave the method they developed the name Bootstrap in honour of the unbelievable stories that the Baron told of his travels.

The general problem

- "We have a set of real-valued observations x_1, ... x_n independently sampled from an unknown probability distribution F. We are interested in estimating some parameter Θ by using the information in the sample data with an estimator Θ-hat = t(x). Some measure of the estimate's accuracy is as important as the estimate itself; we want a standard error of Θ-hat and, even better a confidence interval on the true value Θ."
  - Efron and LePage (1992)

The solution: Bootstrap

- Generate large numbers of "bootstrap" data sets, x1, x2, x3, ... xb, from the original data.
  - The observations in each data set is generated by a random draw, with replacement, from the observations in the original data set.
  - Each data set has the same number of observation (n) as the original data set.
- Refit the model to each bootstrap data set.
- Compute the statistics of interest (probability profile, standard deviations, confidence intervals) from the results for each model fit.

Bootstrap: Why?

- When the sample contains all the available information about the population one can act as if the sample is really the population for the purpose of estimating the sampling distribution.
- Sampling with replacement is consistent with sampling a population that is effectively infinite - treat the sample as the total population.
- Elegant, powerful and easy :-)

Bootstrap: When?

- When sample cannot be represented by a distribution, especially if the underlying population distribution is not known.

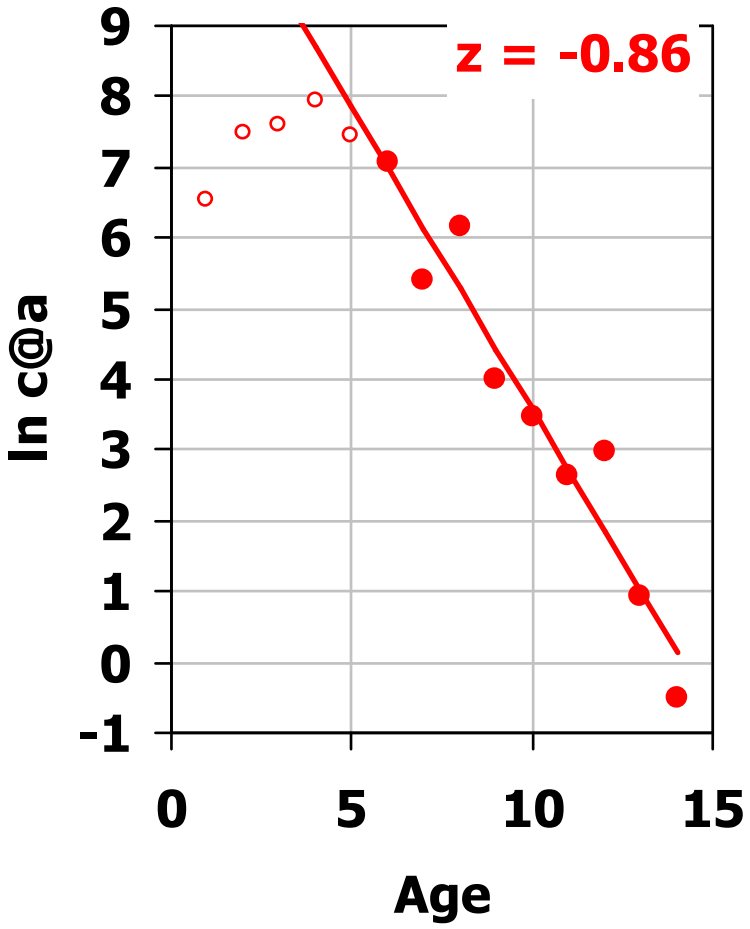
- If one knows the distribution there is little advantage to using bootstrap.
  - There is however no harm in bootstrapping such data sets!

Bootstrapping the residuals

Resampling residuals

- In a time series model one must maintain the time order of the data.
- Thus, resample the residuals with replacement from the optimum fit.
- The randomly sampled residuals are applied to the optimum fitted values to generate new bootstrap samples.
- Process repeated n-times to obtain a probability profile of the value of interest.
- Use a catch curve analysis as an illustration

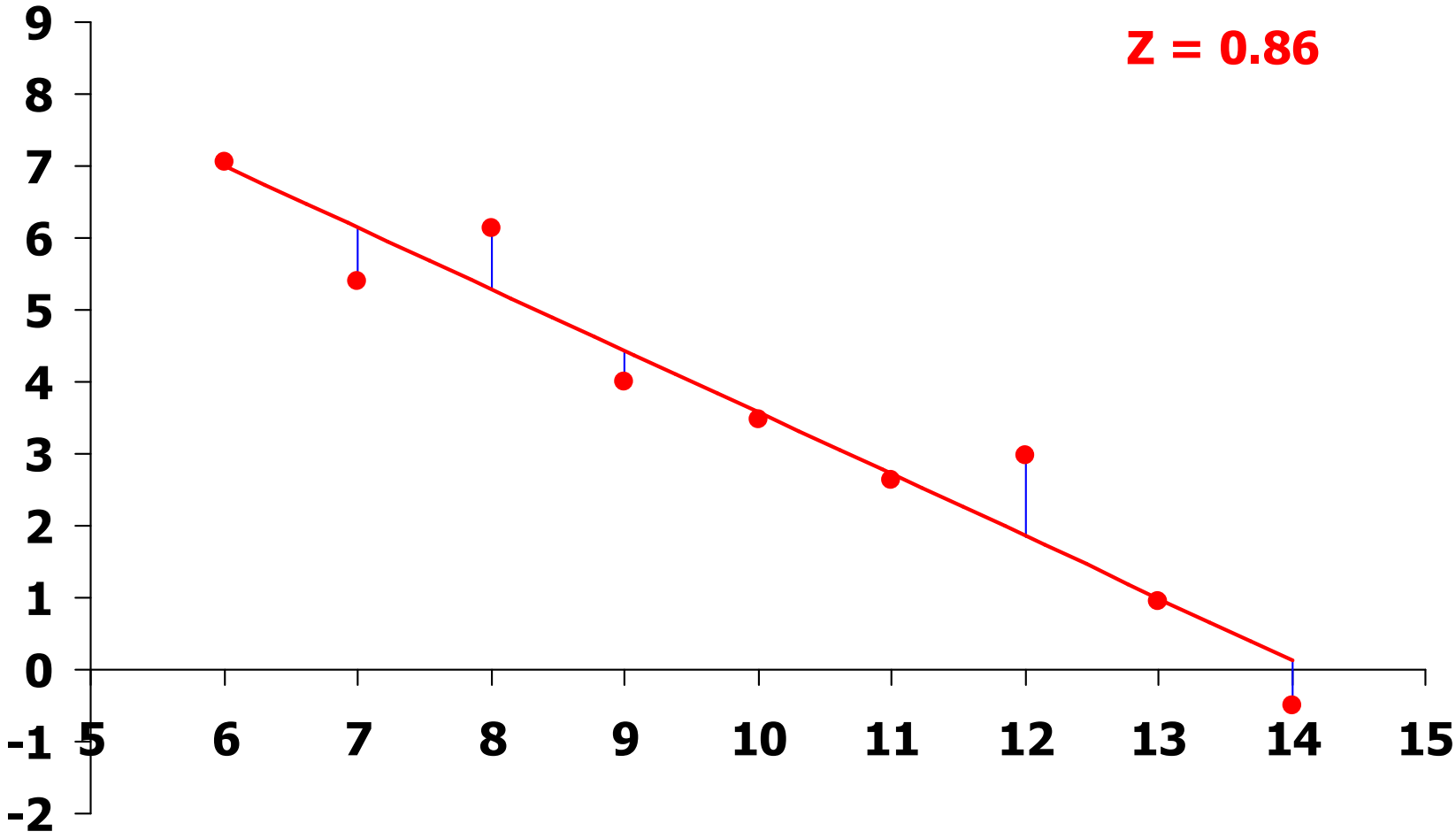
Original data set and statistics



- Data set: c@a of one cohort
- Analysis: Slope estimate of fully recruited fish
- Z = -Slope, if no change in mortality between year
- Task: Obtain some information about the confidence interval of the Z using bootstrapping techniques.

The residuals to resample

einar



One bootstrap sample

Original data set

Age	Observed ln c@a	Predicted ln c@a	obs-pred
6	7.056	6.990	0.066
7	5.406	6.134	-0.728
8	6.143	5.279	0.865
9	3.999	4.423	-0.424
10	3.468	3.567	-0.099
11	2.622	2.711	-0.090
12	2.972	1.855	1.117
13	0.939	1.000	-0.061
14	-0.501	0.144	-0.645

-Slope (Z)
0.86

1 Bootstrap sample

Random draw (age)	Residual value for draw	Bootstrap ln c@a
12	1.117	8.107
9	-0.424	5.710
6	0.066	5.344
14	-0.645	3.778
6	0.066	3.633
8	0.865	3.576
7	-0.728	1.127
6	0.066	1.065
11	-0.090	0.054

-Slope (Z)
0.91

Bootstrap value = Predicted value + random residual value

More formally stated …

$$K_a^* = \hat{K}_a + \varepsilon_{a^*}$$

$$\varepsilon_{a^*} = \left(K_a - \hat{K}_a \right)^*$$

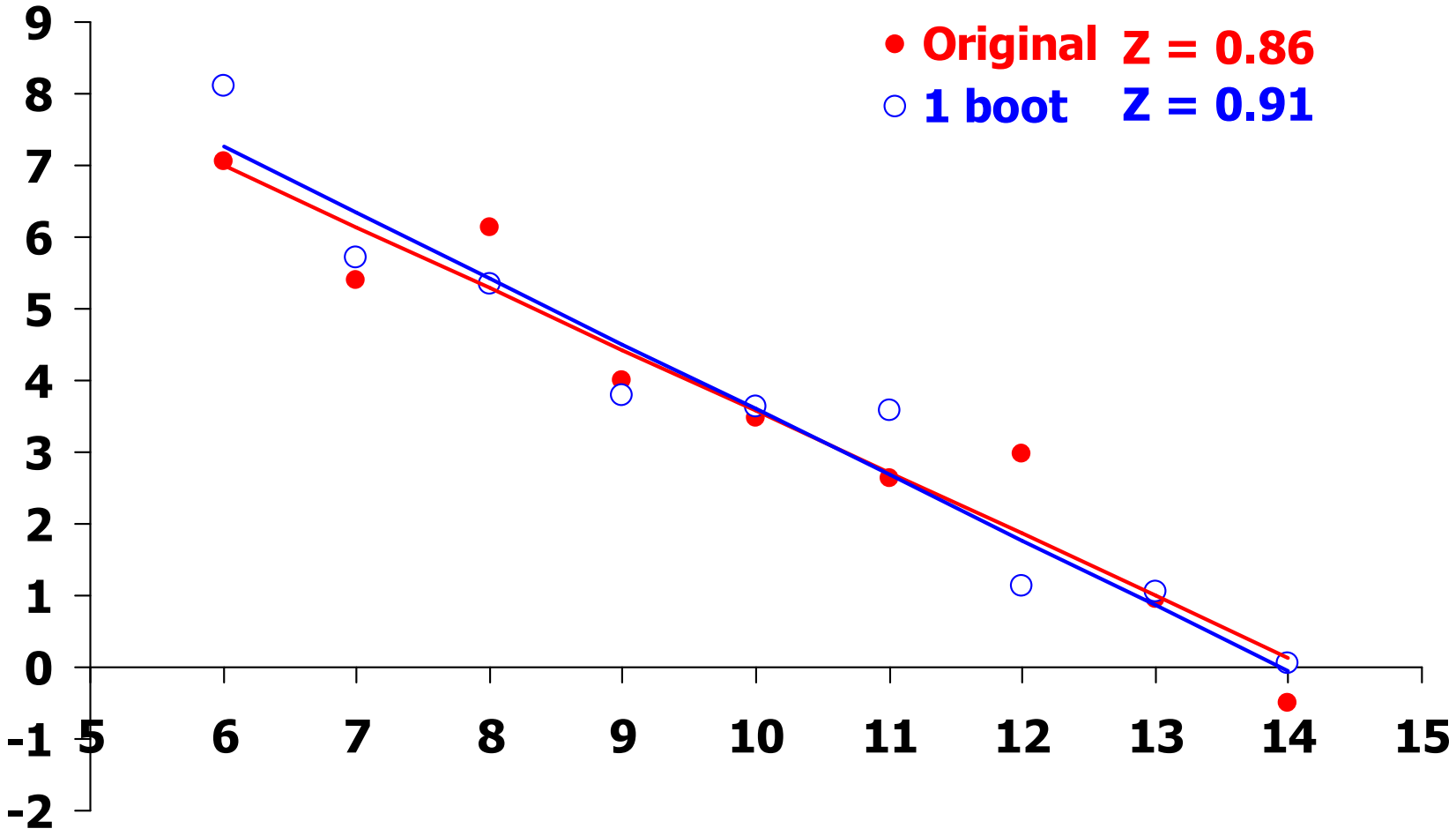
where

$$K_a = \ln(C_a)$$

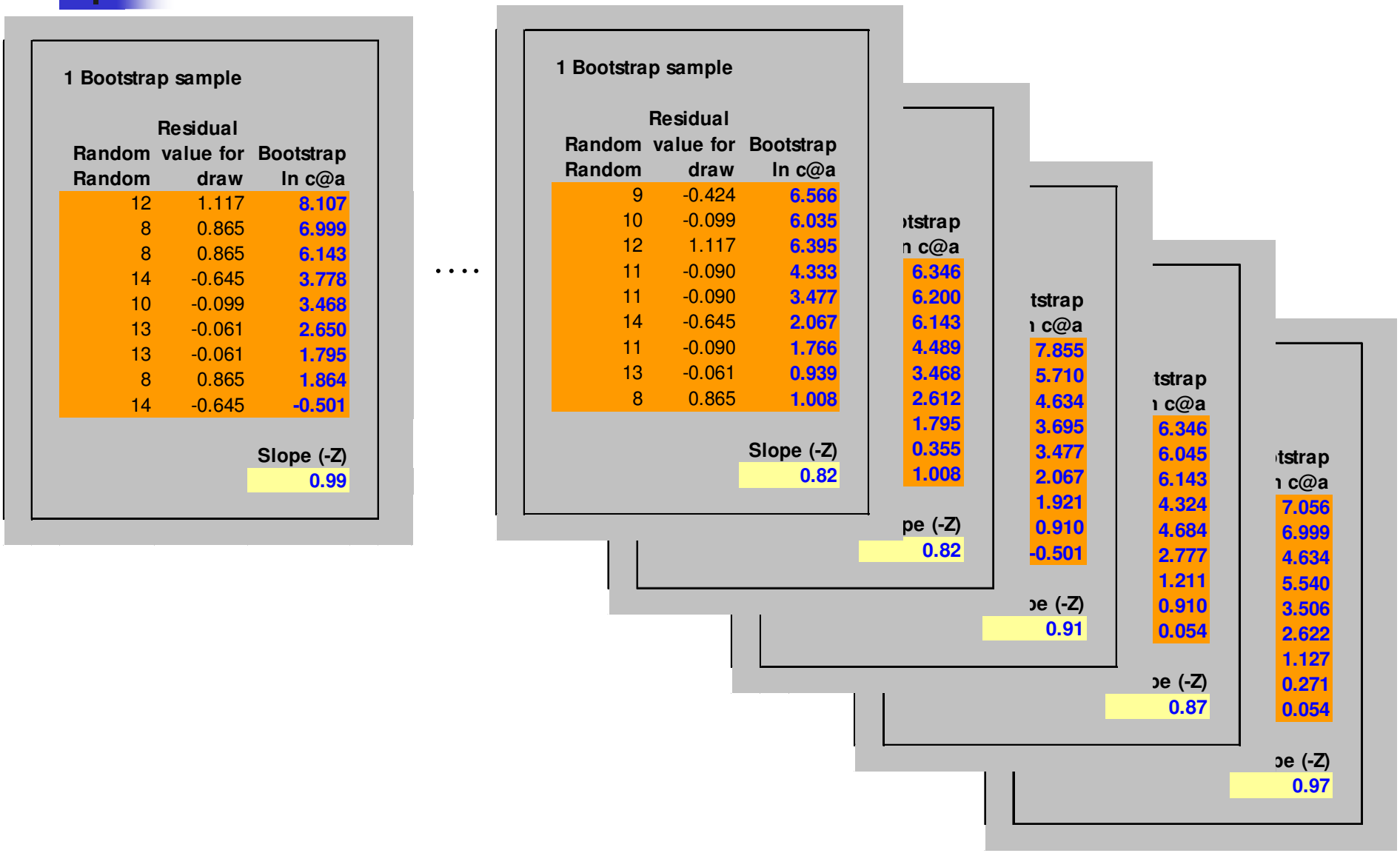
The * represents a random draw of the residuals from the set available

Original and 1 bootstrap sample

einar



n bootstrap samples



1 Bootstrap sample

Random Random	Residual value for draw	Bootstrap ln c@a
12	1.117	8.107
8	0.865	6.999
8	0.865	6.143
14	-0.645	3.778
10	-0.099	3.468
13	-0.061	2.650
13	-0.061	1.795
8	0.865	1.864
14	-0.645	-0.501

Slope (-Z)
0.99

. . . .

1 Bootstrap sample

Random Random	Residual value for draw	Bootstrap ln c@a
9	-0.424	6.566
10	-0.099	6.035
12	1.117	6.395
11	-0.090	4.333
11	-0.090	3.477
14	-0.645	2.067
11	-0.090	1.766
13	-0.061	0.939
8	0.865	1.008

Slope (-Z)
0.82

tstrap
n c@a
6.346
6.200
6.143
4.489
3.468
2.612
1.795
0.355
1.008

pe (-Z)
0.82

tstrap
n c@a
7.855
5.710
4.634
3.695
3.477
2.067
1.921
0.910
-0.501

pe (-Z)
0.91

tstrap
n c@a
6.346
6.045
6.143
4.324
4.684
2.777
1.211
0.910
0.054

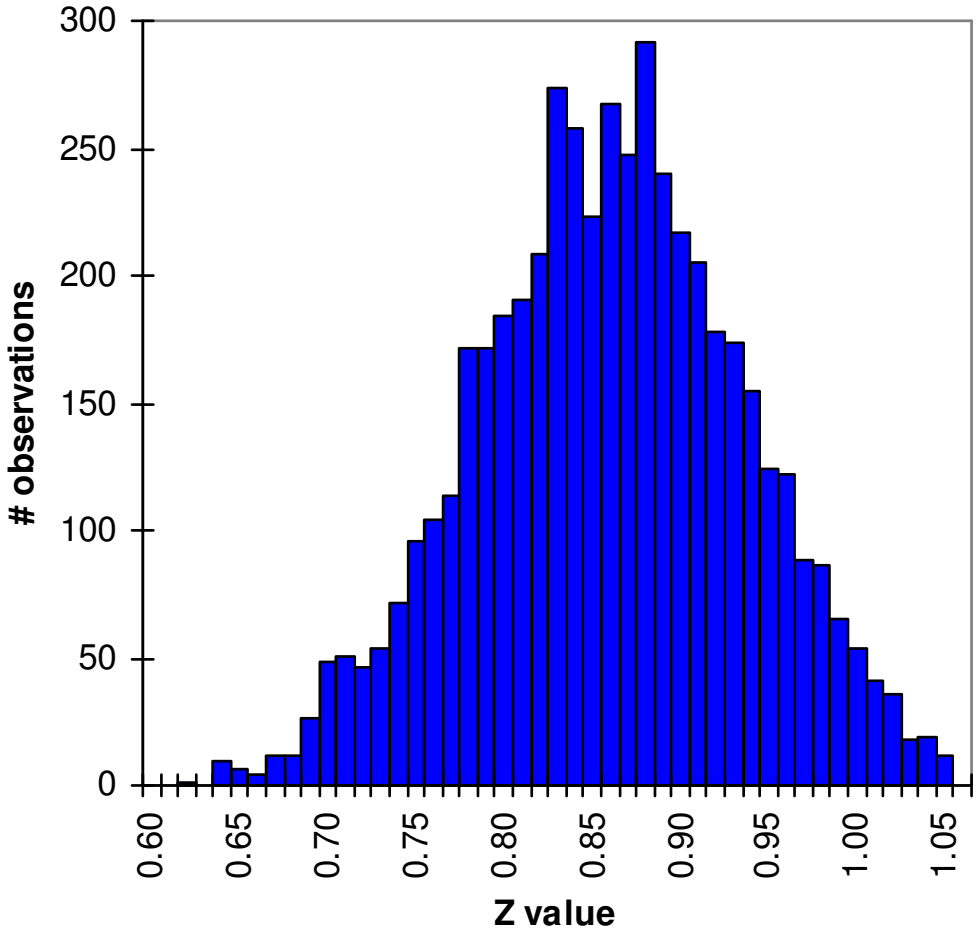
pe (-Z)
0.87

tstrap
n c@a
7.056
6.999
4.634
5.540
3.506
2.622
1.127
0.271
0.054

pe (-Z)
0.97

5000 bootstrap samples: Distribution

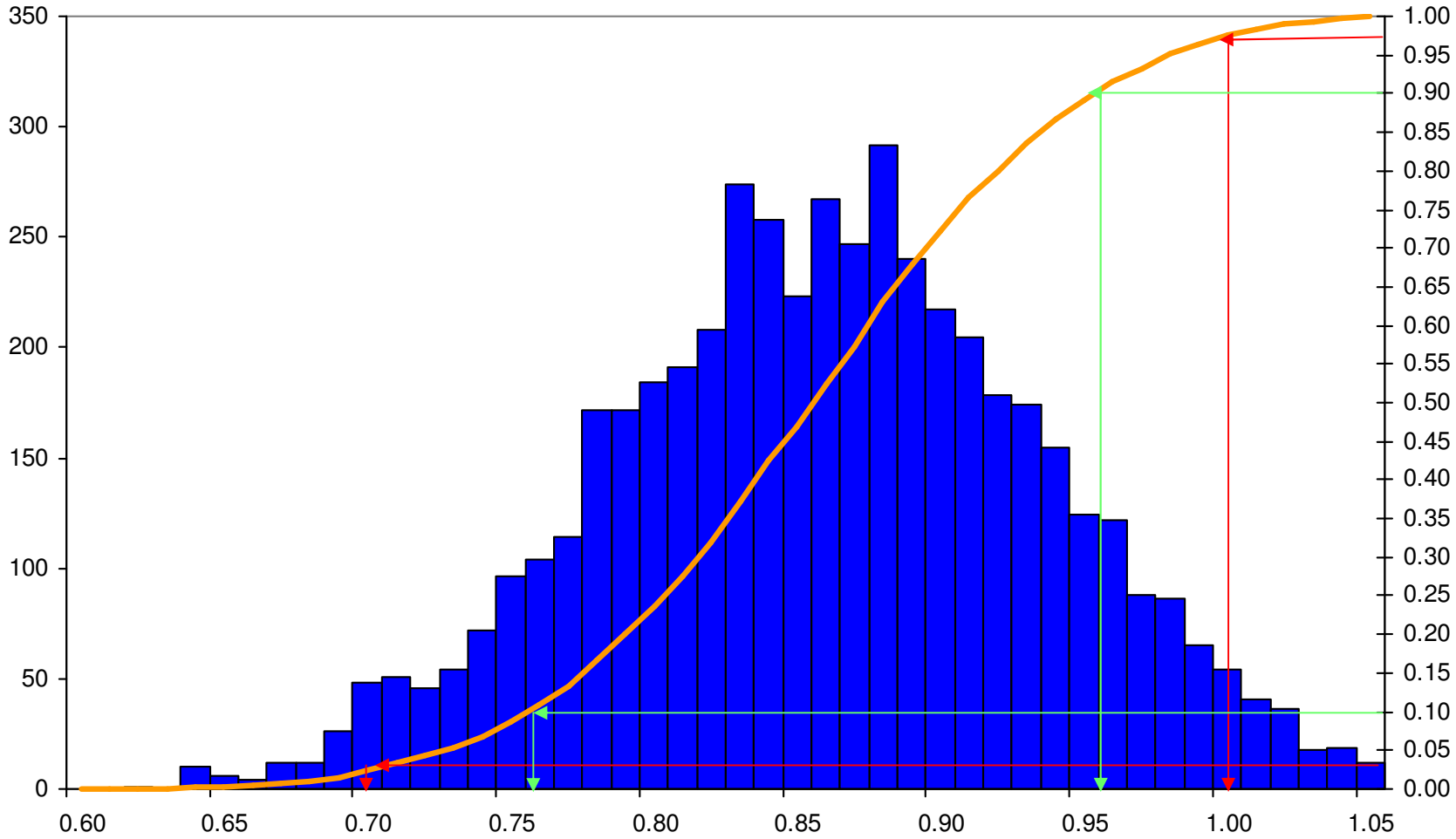
Bootstrap #	Z
1	0.896
2	0.858
3	0.809
4	0.986
5	0.967
6	0.920
7	0.885
8	0.858
9	0.891
10	0.918
11	0.932
12	0.898
13	0.865
14	0.906
15	0.973
...	...
...	...
4992	0.817
4993	0.666
4994	0.814
4995	0.840
4996	0.890
4997	0.756
4998	0.887
4999	0.702
5000	0.716



Bootstrap confidence intervals

- With b bootstrap estimates of the parameter of interest Θ_b:
  - Obtain confidence interval simply by finding the percentile bootstrap estimates that contain the desired confidence.
  - More formally stated: An estimate of the 100(1-α)% CI around the sample estimate of is obtained from the two bootstrap estimates that contain the central 100(1-α)% of all b bootstrap estimates.

80% & 95% confidence interval



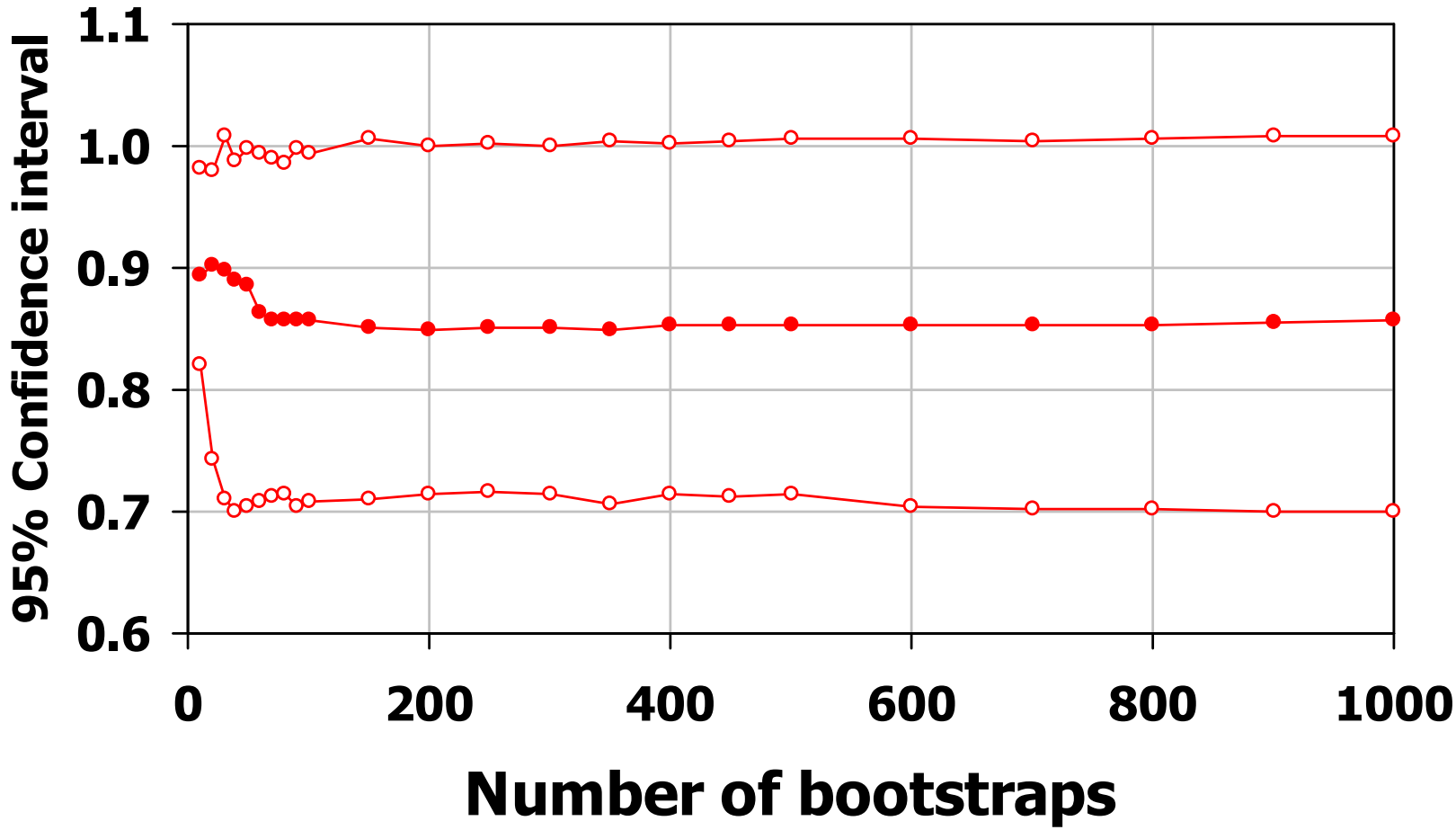
Parametric confidence interval

- If the parameter estimated is expected to follow a normal distribution, the C.I. may be obtained from the usual:
  - $CI = \Theta \pm t_{n-1,a/2} se_{\Theta}$
  - where:
    - Θ: sample parameter estimate
    - $t_{n-1,a/2}$: student t-distribution value for n-1 degrees of freedom
    - n: b, number of bootstrap replicates
    - Since b is generally high, could just use the 1.96 if obtaining 95% confidence interval.

How many bootstraps?

- This was more of an issue prior to the common availability of powerful computers.
- However, in complicated models with many parameters issue is still valid and the number of bootstraps needed is a question of efficiency.
- Can simply test the sensitivity of the parameters in question by running different number of runs ------- >

95% CI of Z and bootstrap numbers



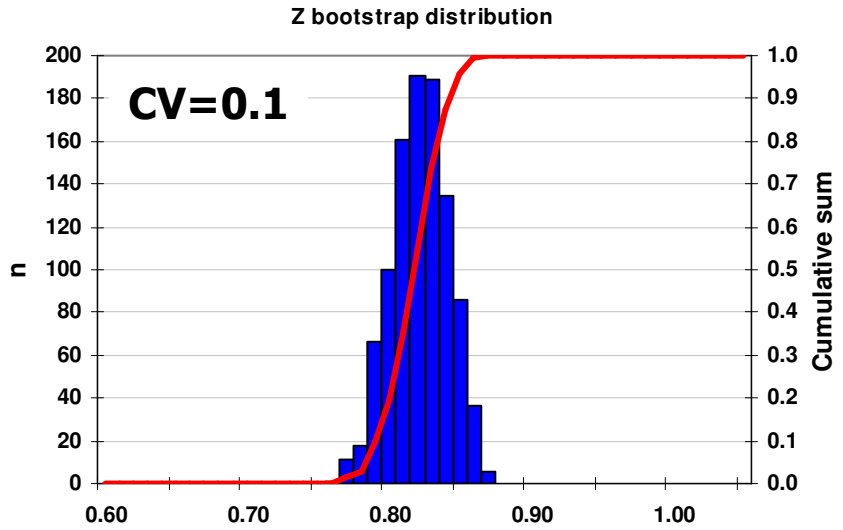
In this simple case do not need much more than around 200 bootstraps to get CI

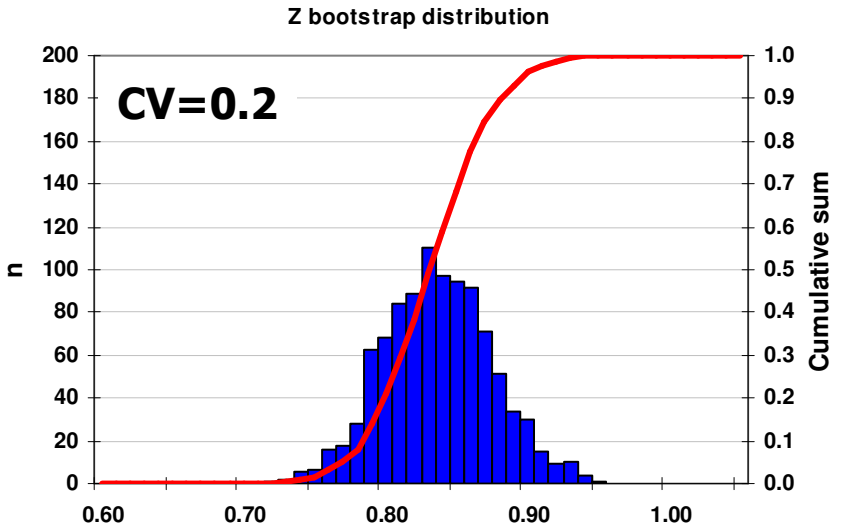
A little hands on experience on bootstrapping

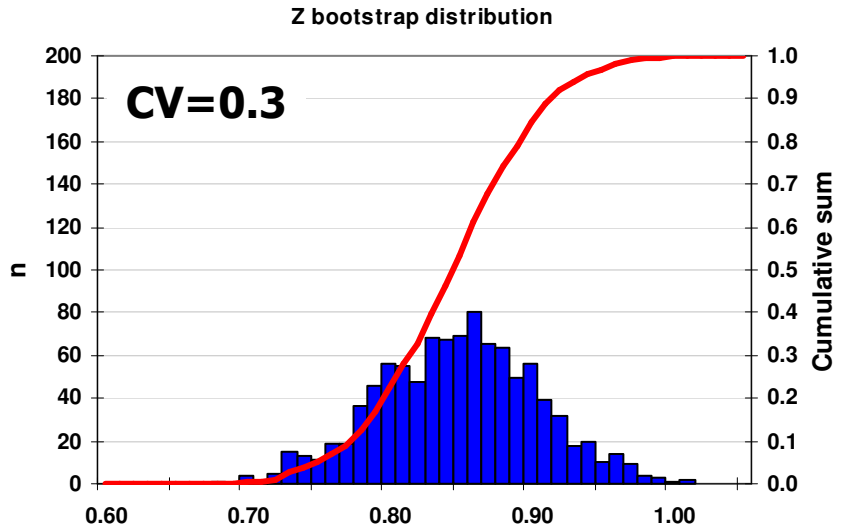
- We have our known state of affairs (xGenerator) where we can set the Z_{ay}.
  - Set constant F_y and M_y, constant fishing pattern with time, but a plateau (=1.00) above some age (set the s_R to a very high value).
  - Set the CV for the catches to a certain value, same value for all age group.
  - Pick a year class, calculate slope of catch curve for the age classes that are fully recruited to the fisheries.
  - Run 1000 bootstrap to estimate the confidence interval of the estimated slope.
  - Repeat the process but change the CV of the measurements of catch at age
    - To ease and speed up the exercise, copy the xBootstrapping101.xls into your BuildingBlock folder (where you have your simulator).
    - I suggest Afull=6, Fy=0.6 and My=0.2 (just because that's how the graphical displays were set up for, but this is not necessary).

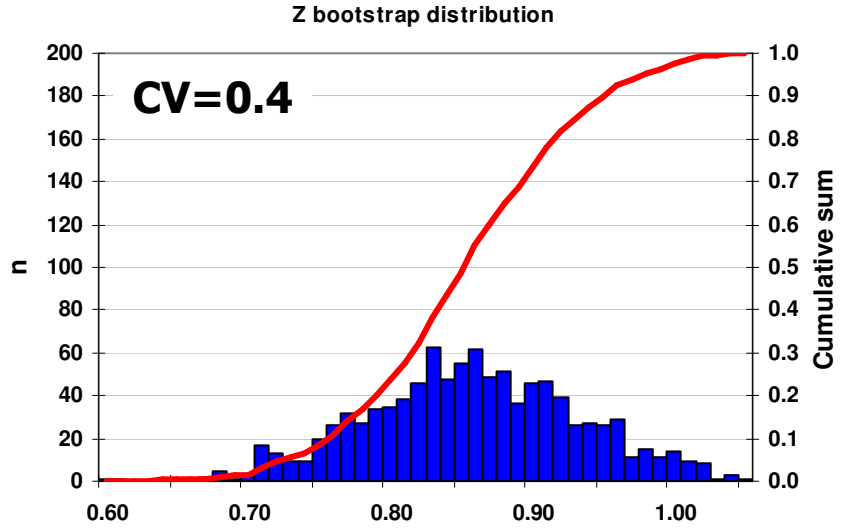
Different CV in catches: 1000 bootstraps

True Z=0.8









Bootstrapping in stock production models

Stock production model (revision)

- Have two equations:

$$B_{y+1} = B_y + rB_y\left(1 - \frac{B_y}{B_{max}}\right) - Y_y$$

$$\hat{U}_y = qB_y$$

  - Data: Y_y, U_y ($CPUE_y$)
  - Parameters: q, r, B_{max}, and B_0
  - Assumptions: .....

  - Note: often assumed $B_0 = B_{max}$, or some ratio there of

Objective function I (revision)

- Lognormal error model:

$$\hat{U}_y = qB_y e^{\varepsilon}$$

- Objective function:
  - SSQ $\sum_y \left(\ln U_y - \ln \hat{U}_y\right)^2 = \sum_y \left(\ln U_y - \ln\left[q\left(B_y\right)\right]\right)^2$

  or
  - LL $-\frac{n}{2}\left(\ln 2\pi + 2\ln\hat{\sigma} + 1\right)$

Objective function II (revision)

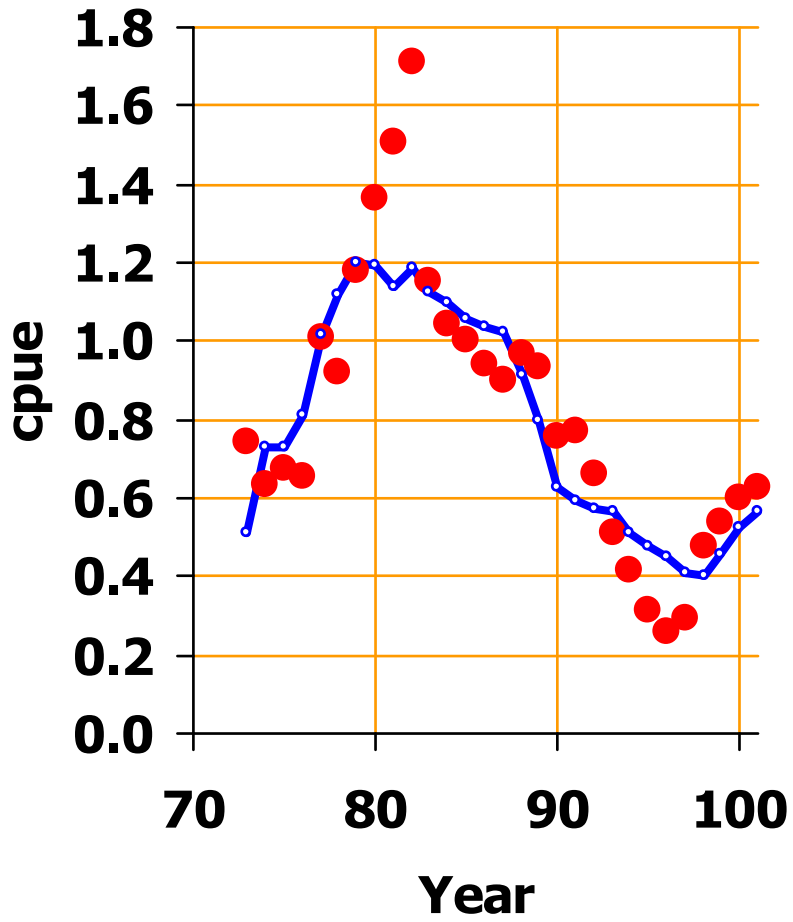
- Note: Full objective function

$$\sum_y \left(\ln U_y - \ln \hat{U}_y\right)^2 = \sum_y \left(\ln U_y - \ln\left[q\left(B_y\right)\right]\right)^2$$

$$= \sum_y \left(\ln U_y - \ln\left[q\left(B_{y-1} + rB_{y-1}\left\{1 - \frac{B_{y-1}}{B_{\max}}\right\} - Y_{y-1}\right)\right]\right)^2$$

- Estimate q, r, B_{max} and B_0 by minimizing the residuals

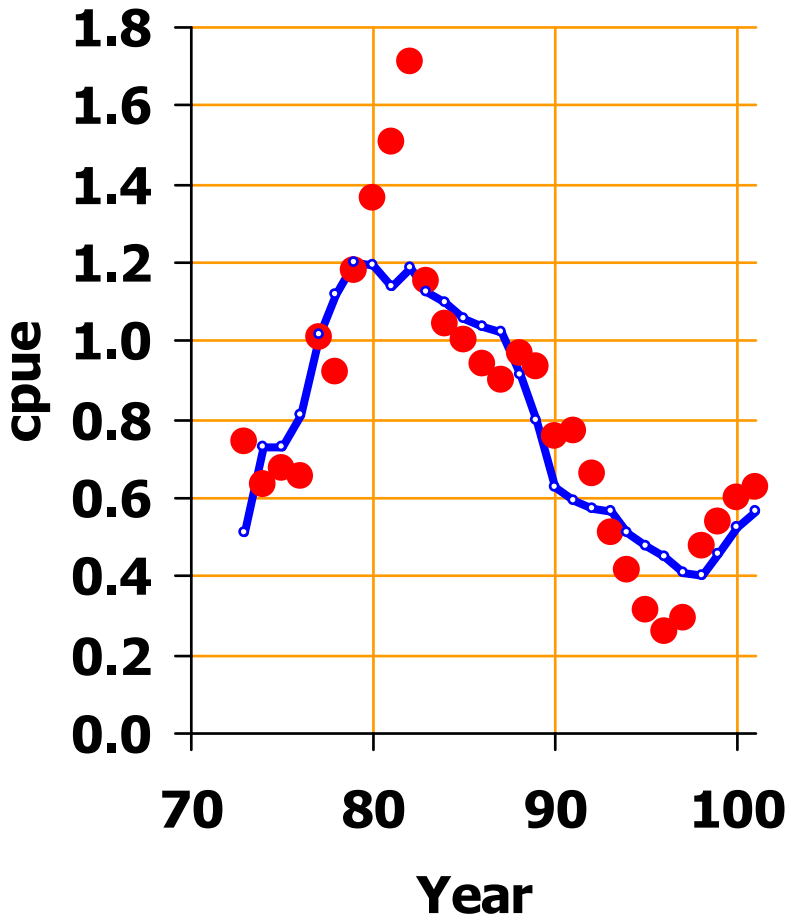
Example: Stock production model

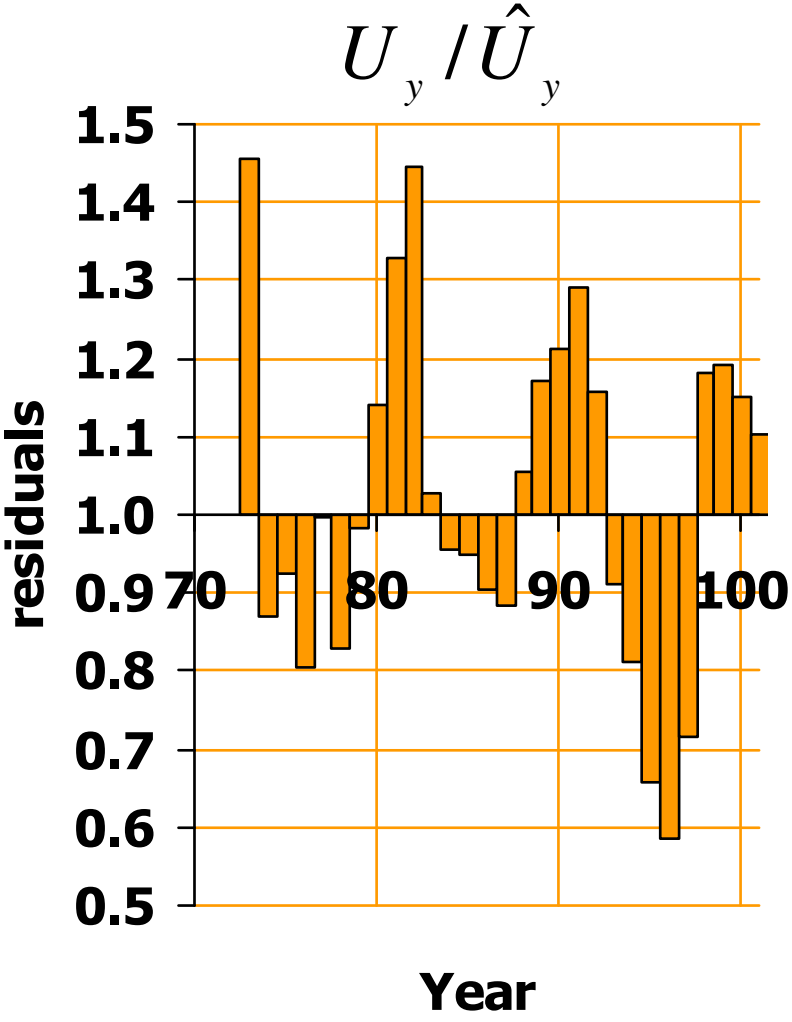


- Data set: West Nordic Greenland halibut
- Analysis: Simple stock production model.
- Parameters: q, r, Bmax, Bo
- Task: Obtain some information about the confidence interval

Example: Residuals of the fit

einar





Bootstrap sampling

einar

$$U_y^* = \hat{U}_y e^{\varepsilon_{y^*}} \qquad \varepsilon_{y^*} = \ln\left(\frac{U_y}{\hat{U}_y}\right)^*$$

*selected at random from 1:n

OPTIMUM MODEL			RESAMPLING W. REPLACEMENT				BOOTSTRAP SAMPLE	
Year	U	Û	#	U/Û	Random number*	(U/U)*	Year	U*
1973	0.75	0.51	1	1.45	9	1.33	1973	0.68
1974	0.63	0.73	2	0.87	5	1.00	1974	0.73
1975	0.67	0.73	3	0.93	3	0.93	1975	0.67
1976	0.65	0.81	4	0.80	11	1.03	1976	0.83
1977	1.01	1.01	5	1.00	27	1.19	1977	1.21
1978	0.92	1.12	6	0.83	12	0.96	1978	1.07
2000	0.60	0.52	28	1.15	23	0.66	2000	0.34

note: $e^{\varepsilon_{y^*}} = e^{\ln\left(\frac{U_y}{\hat{U}_y}\right)^*} = \left(\frac{U_y}{\hat{U}_y}\right)$

Objective function – 1 bootstrap

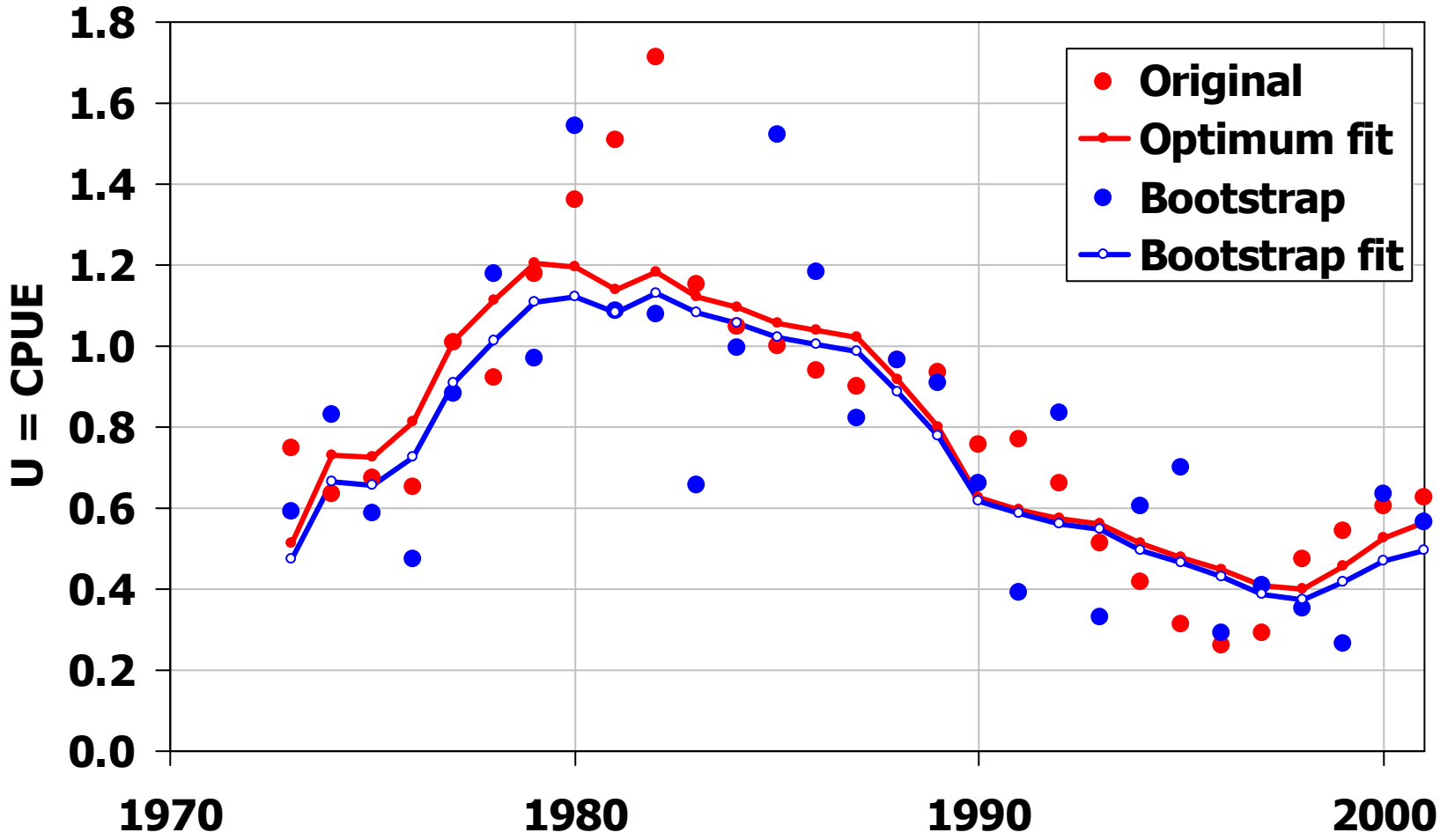
- Note: The star indicates this is a bootstrap sample

$$\sum_y \left(\ln U_y^* - \ln \hat{U}_y\right)^2 = \sum_y \left(\ln U_y^* - \ln\left[q^*\left(B_y\right)\right]\right)^2$$

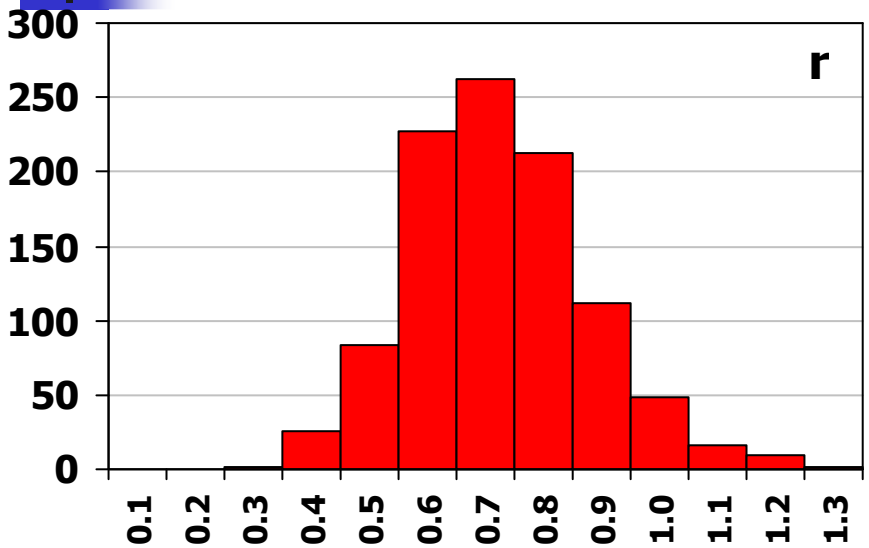
$$= \sum_y \left(\ln U_y^* - \ln\left[q^* \left(B_{y-1} + r^* B_{y-1} \left\{ 1 - \frac{B_{y-1}}{B^*_{\max}} \right\} - Y_{y-1} \right) \right] \right)^2$$

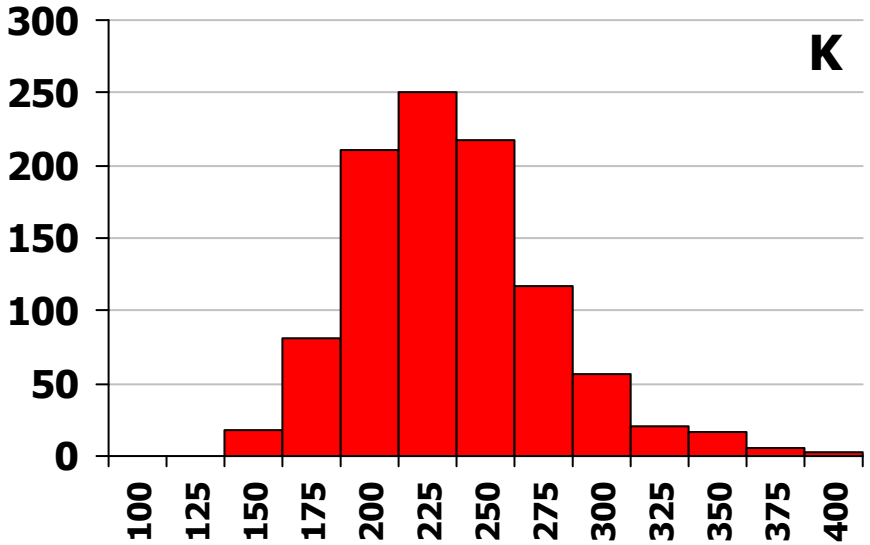
- Estimate the parameters q, r, B_{max} and B_0 for the bootstrap sample by minimizing the objective function.
- Store the parameters

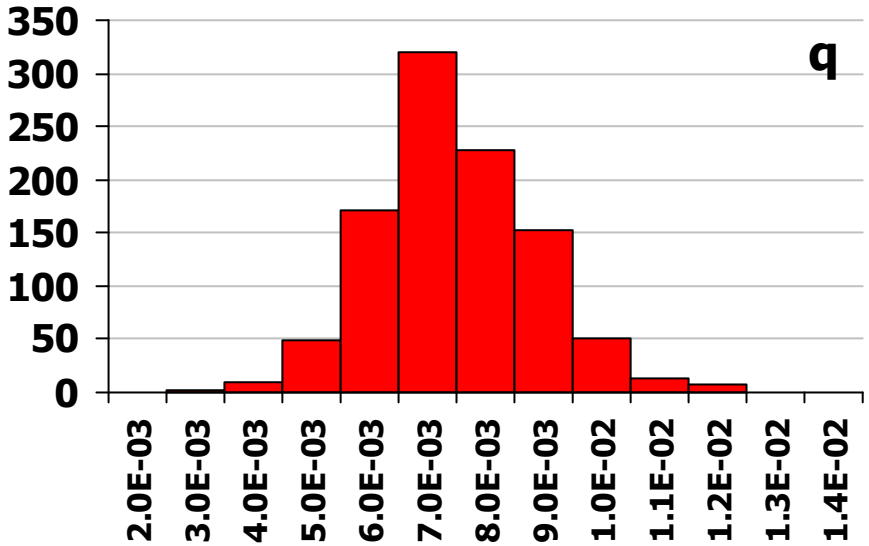
Original vs. Bootstrap sample n

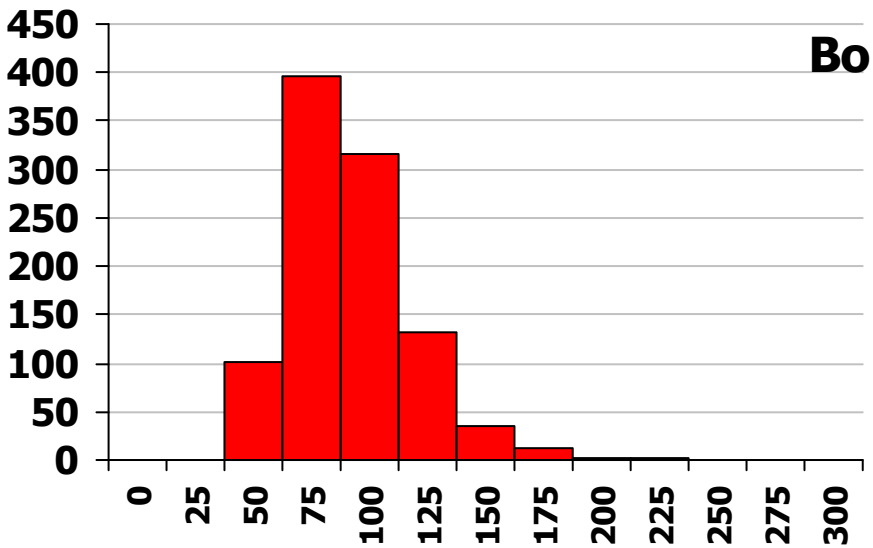


Parameter distribution from 1000 b

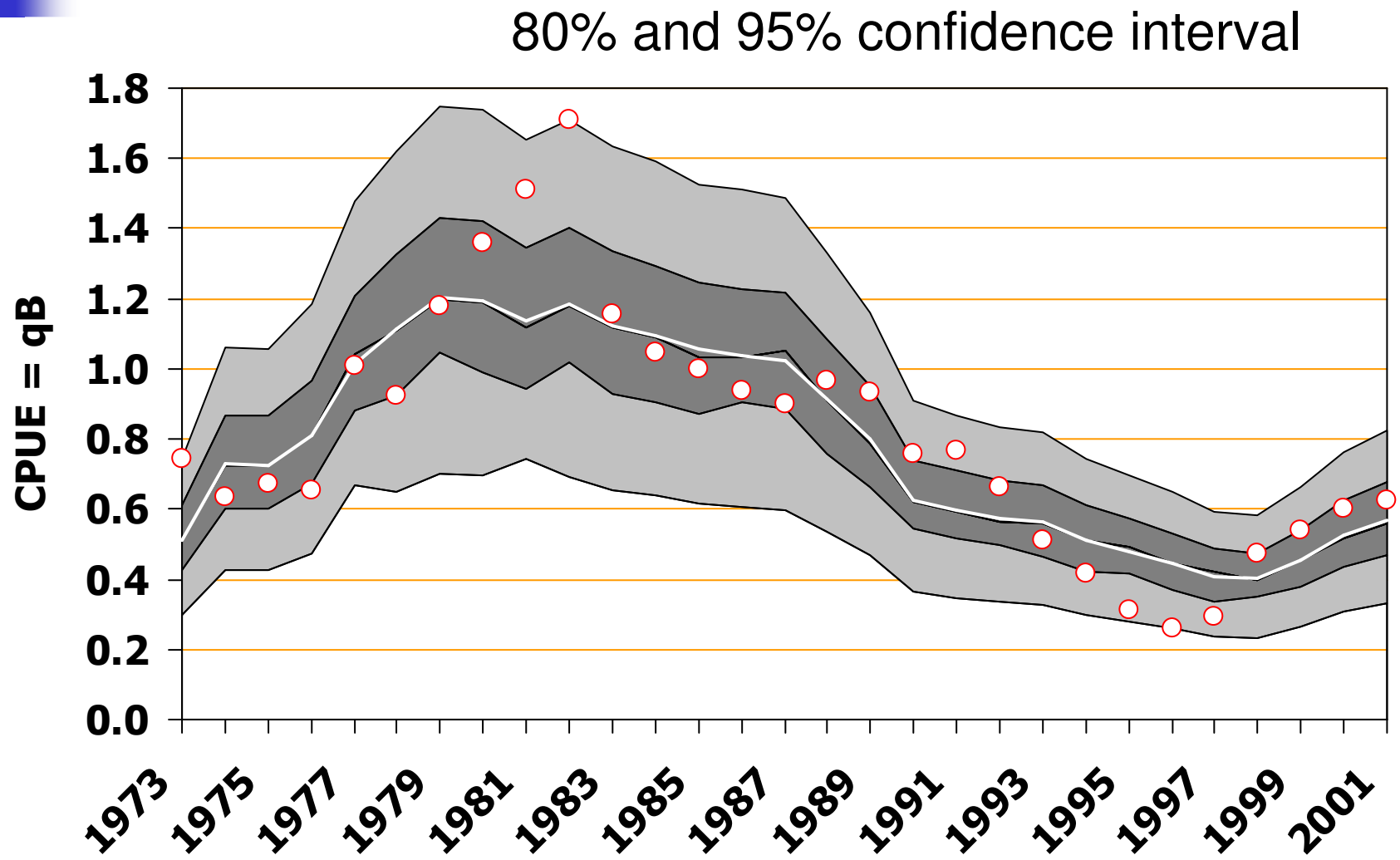








1000 bootstrap runs



Reference values (revision)

- The reference values for stock production model are derived from parameters and are:

$$B_{MSY} = \frac{B_{\max}}{2}$$

$$E_{MSY} = \frac{r}{2q}$$

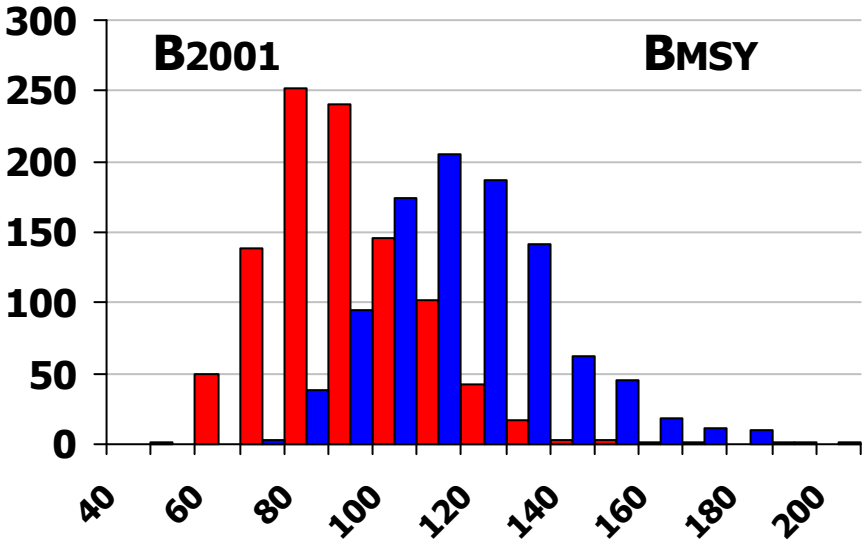
$$MSY = \frac{rB_{\max}}{4}$$

$$F_{MSY} = qE_{MSY} = q\frac{r}{2q} = \frac{r}{2}$$

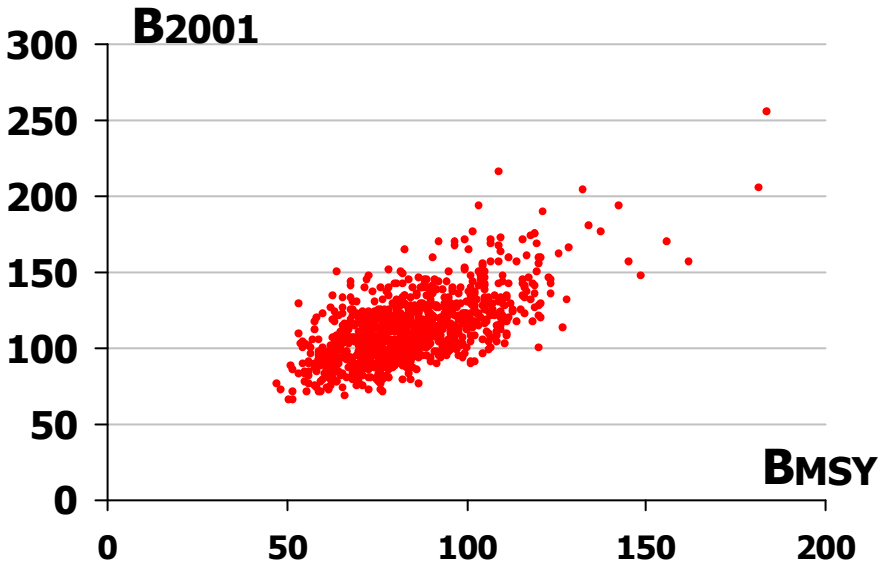
MSY: Maximum sustainable yield
Emsy: Effort at MSY
Fmsy: Max. sustainable fishing mortality

Current stock status relative to B_{MSY}?

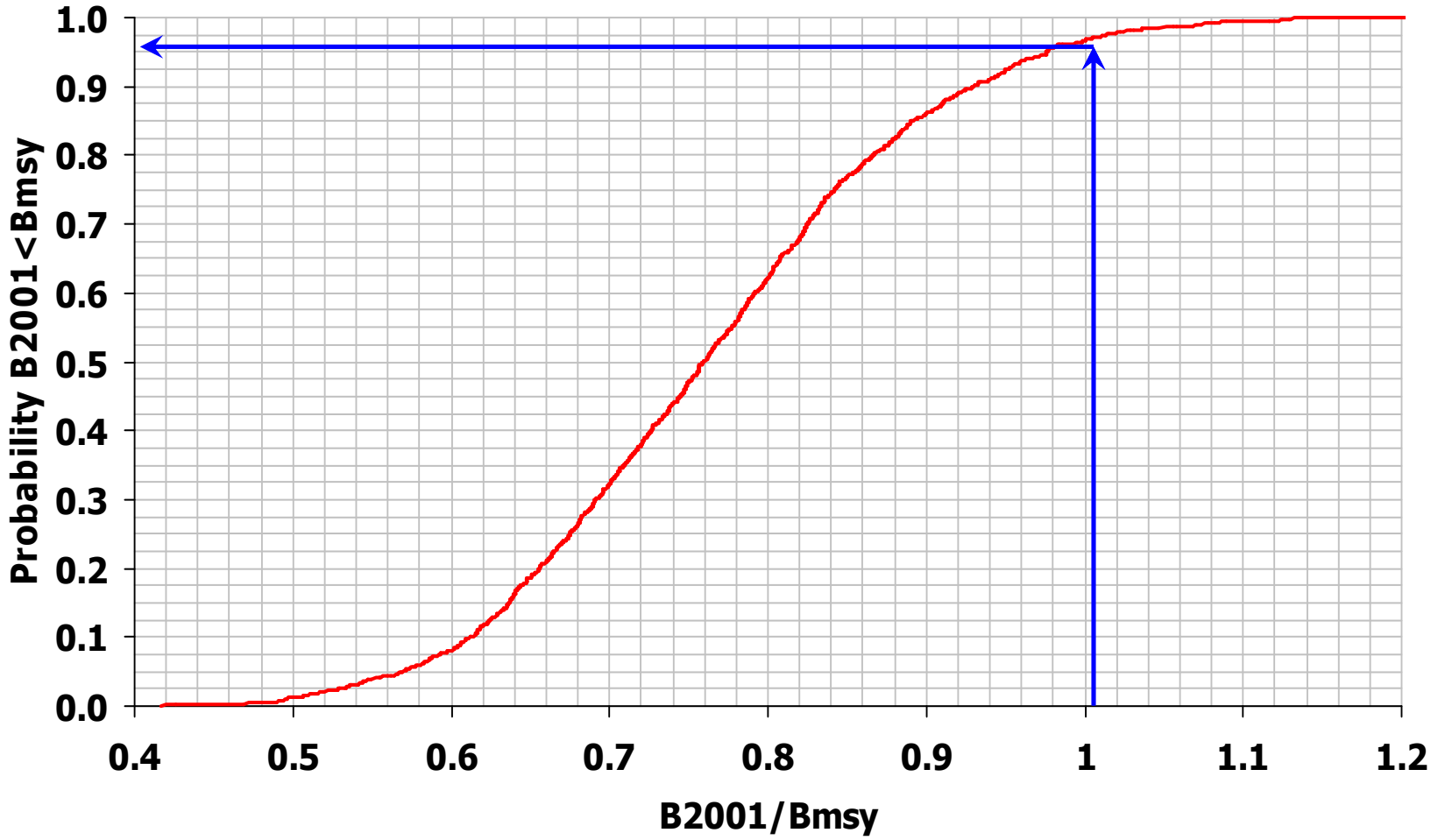
- Have estimates and CI of B_{2001} and B_{MSY}:



Note: The estimates of B_{MSY} and B_{2001} are correlated.
May be better to take the ratio of B_{2001}/B_{MSY} from each run.



p(B2001<BMSY) = 0.97



Bootstrap in age based models

“A visual tour” using xModel

Follow these steps xModel

- Start from the optimum fit
- Store the estimates of c@a, u1@a & u2@a

$$\hat{C}_{a,y} \qquad \hat{U}^{\text{Survey1}}_{a,y} \qquad \hat{U}^{\text{Survey2}}_{a,y}$$

- Store the estimates of the ratio of obs/pre

$$\left(\frac{C_{a,y}}{\hat{C}_{a,y}}\right)^{*} \qquad \left(\frac{U^{Survey1}_{a,y}}{\hat{U}^{1}_{a,y}}\right)^{*} \qquad \left(\frac{U^{Survey2}_{a,y}}{\hat{U}^{2}_{a,y}}\right)^{*}$$

- Random bootstrap sampling is already set up
- Copy the bootstrap sample into the observation area of the model. Run Solver.

Store in Excel speak is equivalent to copy-paste-special-value within the same area

Bootstrap in xModel

0) Obtain the optimum fit. Store the predicted catch at age values and the residuals.
1) In each bootstrap run a new bootstrap sample (area 3) is drawn randomly from the residuals (area 2) and applied to the optimum catch at age (area 1)
2) The bootstrap sample (are 3) is pasted into the data area where the original observations are stored (see end of arrow).
3) The optimum parameters for this particular bootstrap sample are found by minimizing the objective function.
4) The parameters, and whatever else of interest are stored

Steps 1-4 are repeated many times

Bootstrap stuff is in row 240 and below

	K	L	M	N	O	P	Q	R
240								
241								
242								
243			**Optimum predicted Catch-at-Age as values - used to create the boo**					
244			**Year\Age**	**0**	**1**	**2**	**3**	**4**
245			**1**	24.22	67.09	108.35	38.43	87.09
246			**2**	24.50	31.12	74.00	101.57	30.53
247			**3**	26.39	37.51	40.91	82.75	96.38
248			**4**	42.52	43.67	53.07	48.97	83.60
249			**5**	60.78	63.47	55.51	56.76	43.94
250			**6**	94.98	94.78	83.96	61.51	52.51
251			**7**	35.57	146.62	123.48	91.03	55.30
252			**8**	70.11	51.19	177.46	123.79	75.28
253			**9**	56.14	100.87	61.81	176.99	101.56
254			**10**	107.74	86.90	130.37	65.59	153.56
255			**11**	31.88	147.17	98.77	121.08	49.52
256			**12**	149.14	46.54	178.33	97.51	96.92
257			**13**	41.10	173.12	45.01	141.04	62.69
258			**14**	68.13	60.22	211.95	45.28	116.07

$$\hat{C}_{a,y}\left(\frac{C_{a,y}}{\hat{C}_{a,y}}\right)^{*}$$

The random sampler

$$\hat{C}_{a,y}\left(\frac{C_{a,y}}{\hat{C}_{a,y}}\right)^{*}$$

	K	L	M	N	O	P	Q	R	S	T
322										
323			**Catch-at-age: Bootstrap sample - this will be copied into cells F3:N11 by using the macro**							
324			**Year\Age**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
325			**1**	**24.08**	**54.41**	**99.46**	**25.84**	**57.71**	**98.52**	**154.34**
326			**2**	**38.16**	**23.78**	**96.62**	**104.87**	**36.24**	**57.34**	**36.17**
327			**3**	**34.05**	**52.39**	**41.15**	**79.27**	**101.73**	**21.62**	**32.59**
328			**4**	**53.36**	**65.44**	**68.82**	**54.50**	**31.05**	**87.66**	**24.20**
329			**5**	**72.22**	**71.95**	**62.93**	**105.35**	**39.58**	**74.77**	**82.44**
330			**6**	**69.79**	**71.45**	**114.87**	**50.81**	**47.30**	**35.20**	**47.35**
331			**7**	**36.87**	**216.51**	**113.98**	**71.21**	**99.27**	**42.70**	**27.91**
332			**8**	**15.82**	**53.30**	**166.42**	**128.03**	**62.78**	**48.40**	**22.69**
333			**9**	**54.58**	**83.33**	**66.31**	**145.32**	**111.21**	**54.37**	**23.49**
334			**10**	**71.39**	**130.22**	**207.74**	**49.21**	**285.03**	**67.94**	**49.64**
335			**11**	**27.50**	**106.06**	**106.34**	**113.93**	**110.07**	**71.81**	**32.16**
336			**12**	**141.91**	**46.49**	**177.05**	**108.52**	**110.13**	**37.79**	**42.48**
337			**13**	**42.33**	**138.39**	**50.20**	**157.31**	**53.12**	**76.01**	**13.05**
338			**14**	**70.46**	**54.34**	**174.71**	**50.22**	**88.32**	**30.36**	**31.32**

Press F9 to see a new bootstrap sample. Once you get bored copy the data and paste them into the area where the original observation data are. Run the Solver and you have your first bootstrap. How does the result compare with your F, SSB and R profile of your optimum sample?

New parameter estimates for each bootstrap sample

Optimum fit

PARAMETERS

Name	Ln(paran	S	Parameter
Ln Afull	1.8017		6.06
Ln sL	2.3535		10.52
Ln sR	25.0000		#######
LnF 1900	-0.8229		0.44
LnF 1901	-0.8697		0.42
LnF 1902	-0.8344		0.43
LnF 1903	-0.9048		0.40
LnF 1904	-0.7820		0.46
LnF 1905	-0.7198		0.49
LnF 1906	-0.7832		0.46
LnF 1907	-0.7380		0.48
LnF 1908	-0.7434		0.48
LnF 1909	-0.6385		0.53
LnF 1910	-0.6751		0.51
....	....		....
....	....		....

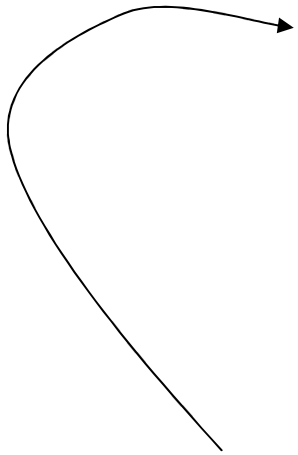
1 bootstrap estimate

PARAMETERS

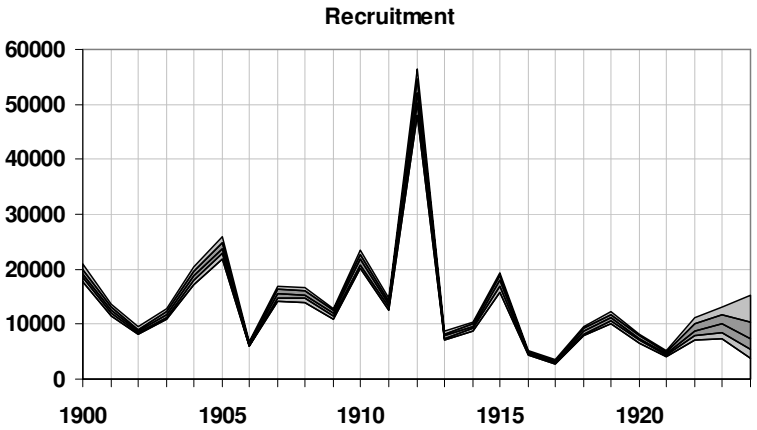
Name	Ln(paran	S	Parameter
Ln Afull	1.8153		6.14
Ln sL	2.4021		11.05
Ln sR	25.7000		########
LnF 1900	-0.8012		0.45
LnF 1901	-0.9061		0.40
LnF 1902	-0.9013		0.41
LnF 1903	-0.9010		0.41
LnF 1904	-0.8386		0.43
LnF 1905	-0.7549		0.47
LnF 1906	-0.7474		0.47
LnF 1907	-0.8347		0.43
LnF 1908	-0.7046		0.49
LnF 1909	-0.6255		0.54
LnF 1910	-0.7053		0.49
....	....		....
....	....		....

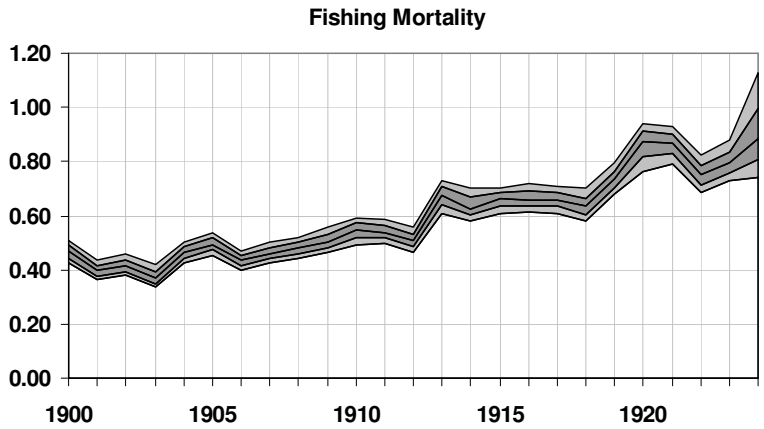
Use a macro to run the bootstrap

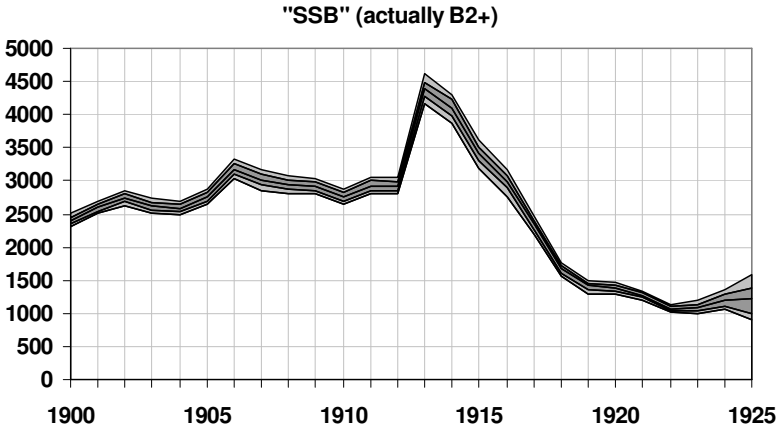

```
Sub Do_Bootstrap()
'
' Do_Bootstrap Macro
' Macro recorded 23/01/2001 by Malcolm Haddon
' Modified by Einar 24/2/2004
' Note: When the Macro ends the values in "The input data"
' area in the Model spreadsheet contains the last bootstap
' sample. Need thus to link again the original measurements
' stored in worksheet CatchAtAge and Survey1 and rerun the
' optimiser (Solver).
'
Dim i As Integer
Application.ScreenUpdating = False

' In the next line one specifies the number of bootstrap runs
  For i = 1 To 500

            ....

            ....

  SolverOk SetCell:="objectives", MaxMinVal:=2, ValueOf:="0",
ByChange:="E37:E38;E40:E99;E101:E111"
    SolverSolve (True)

            ....

Next i
```



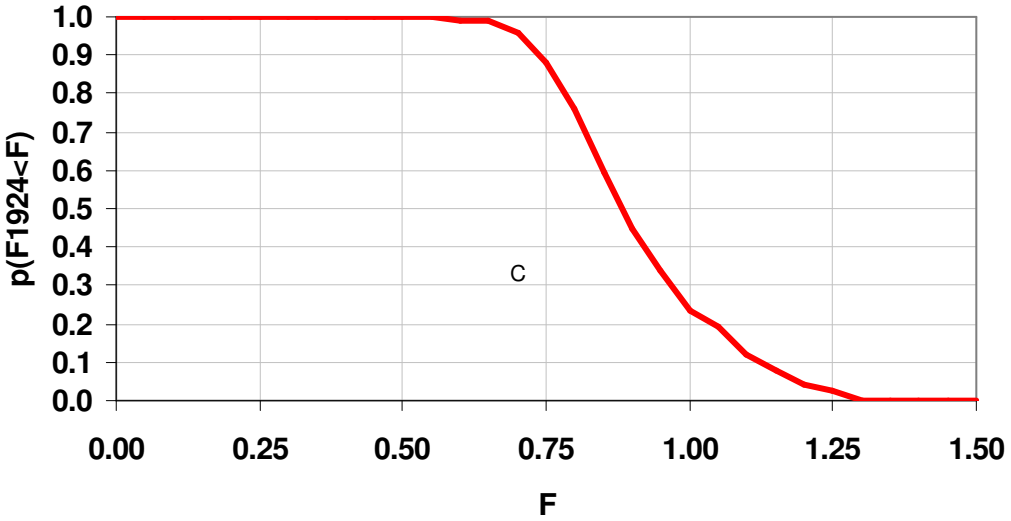
Graphical output

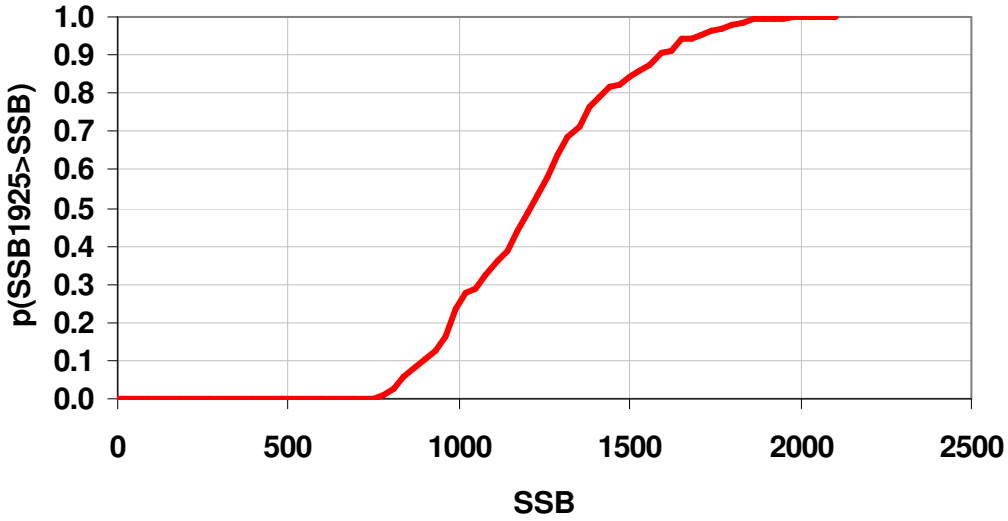




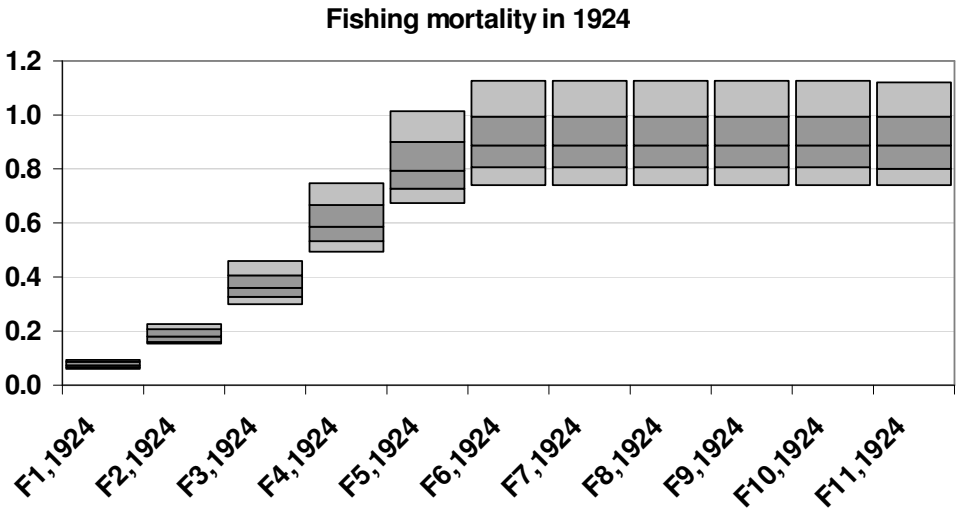


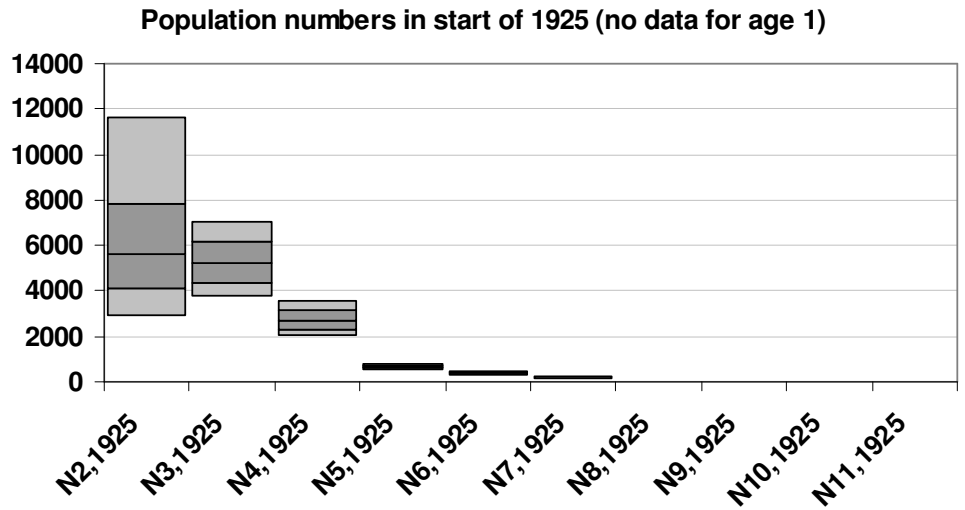
More detailed output …





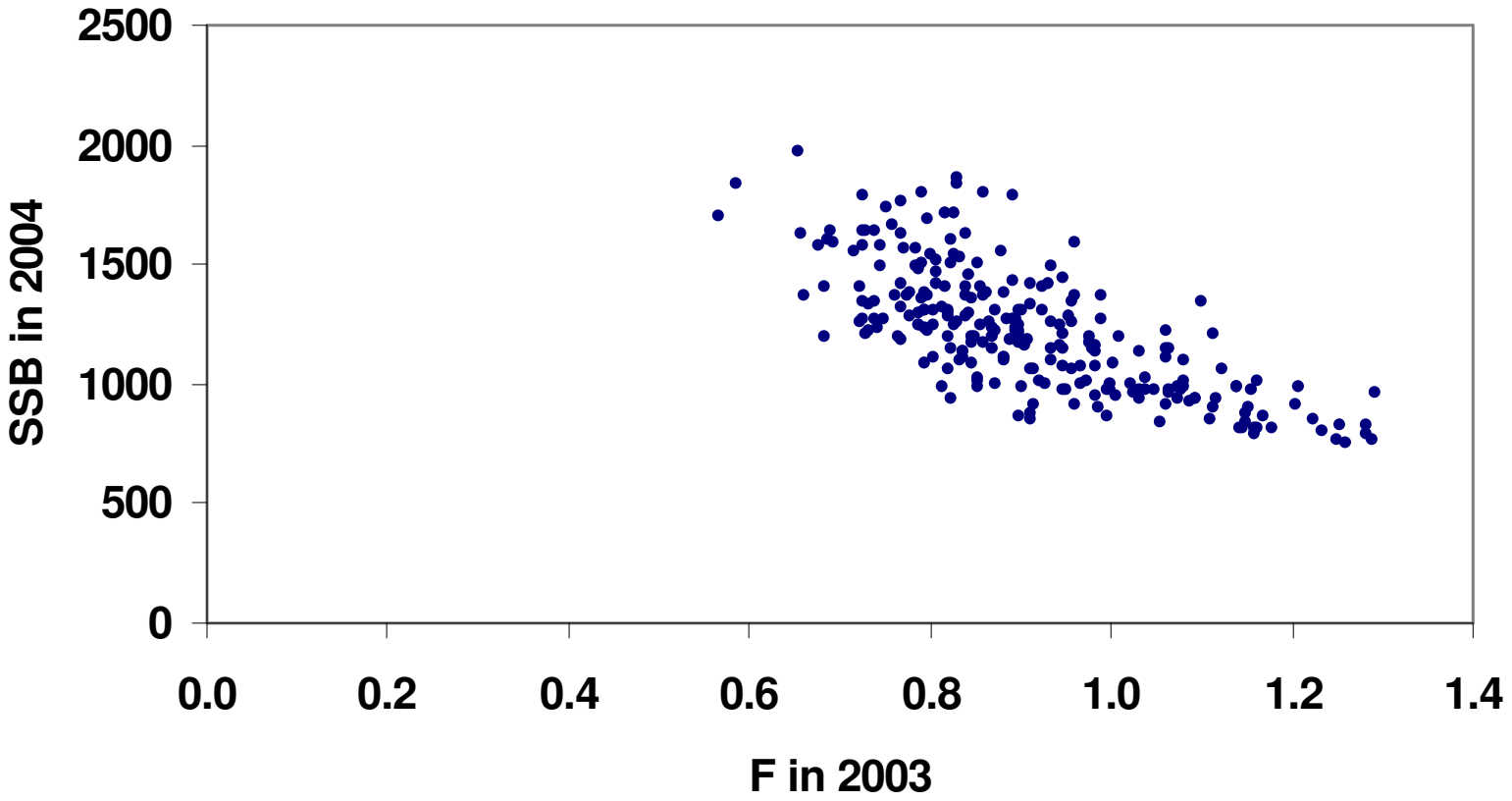
The terminal information





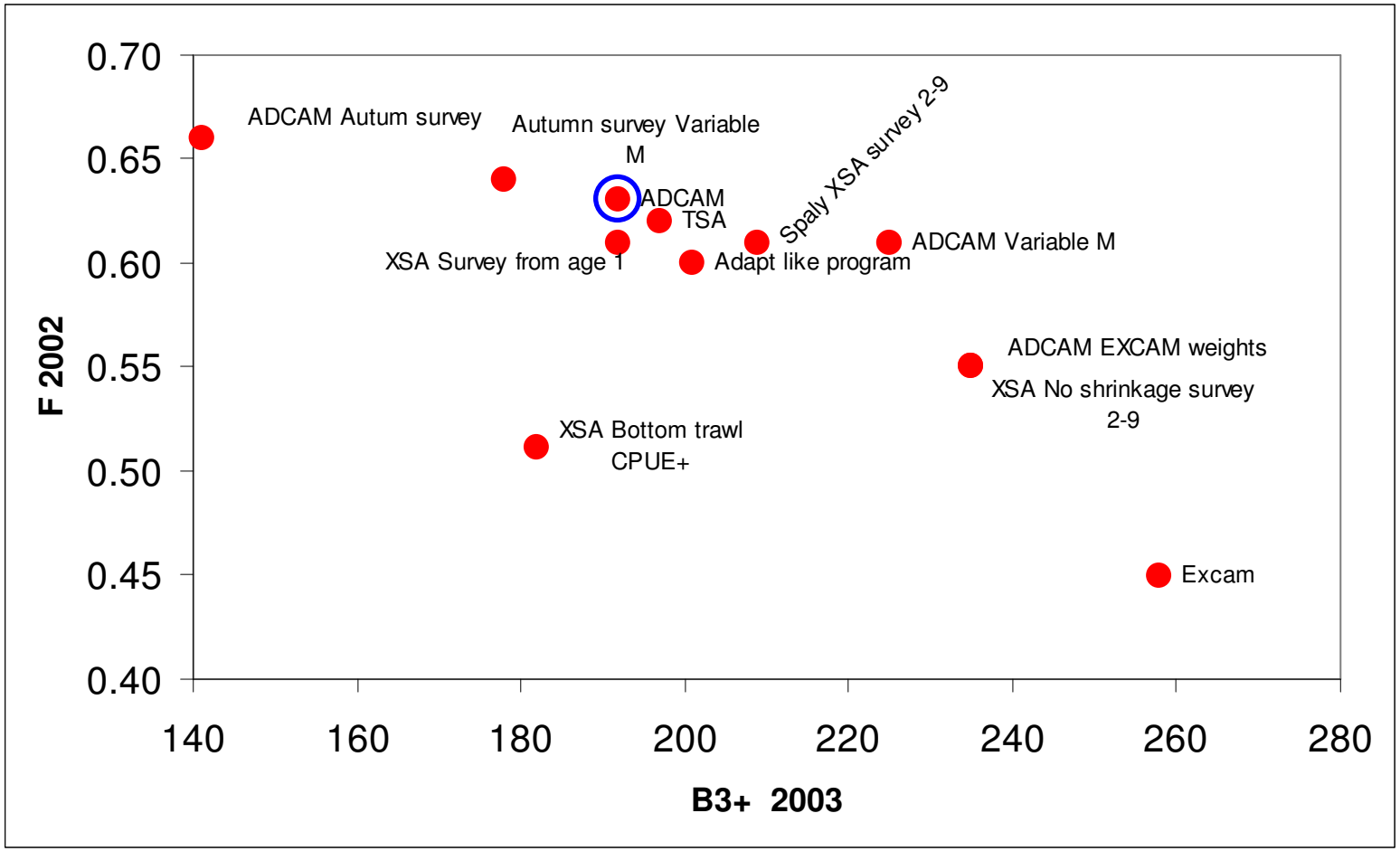
The banana pattern - correlation

einar



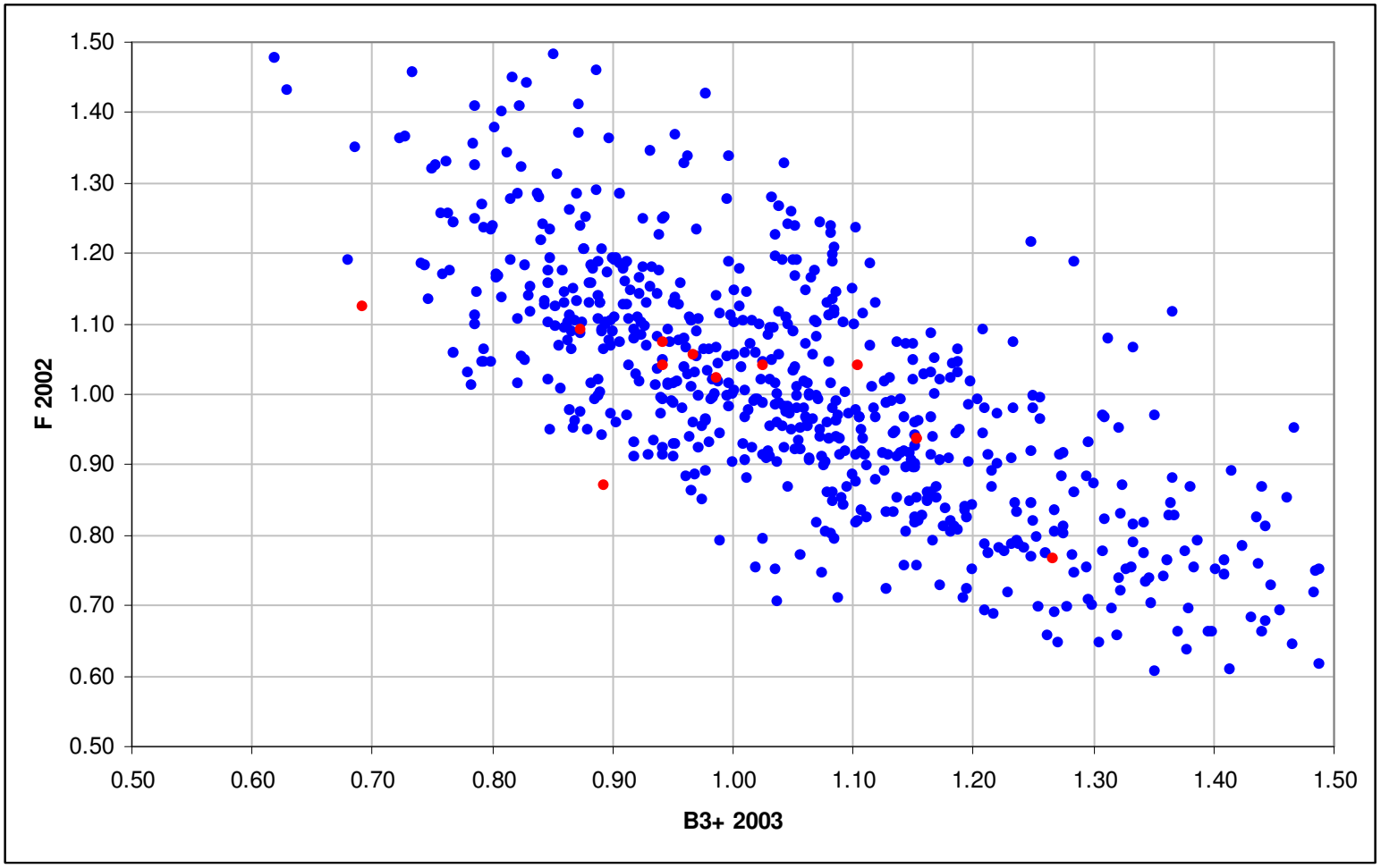
Some “real” life examples

I. haddock - different runs



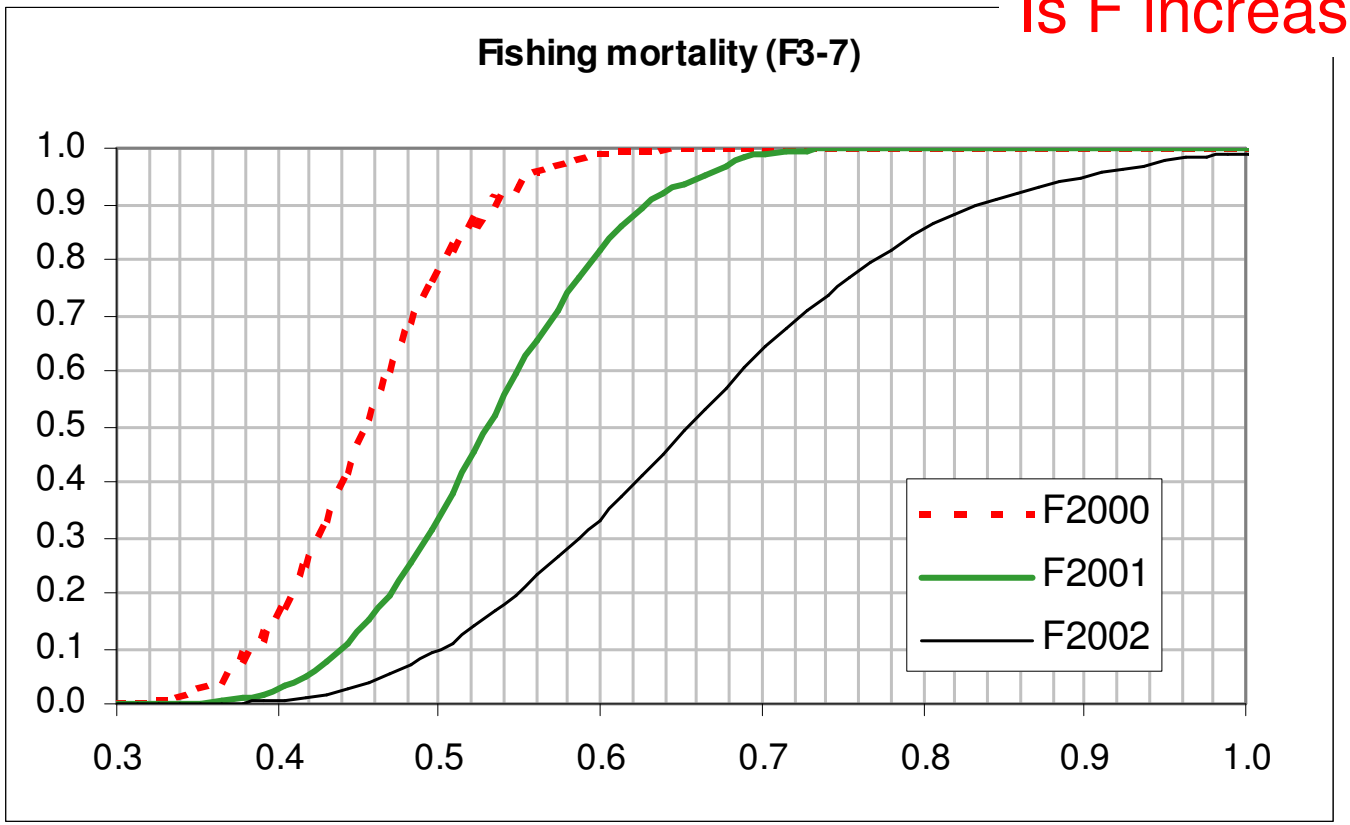
I. haddock - bootstrap noise

Relative scale



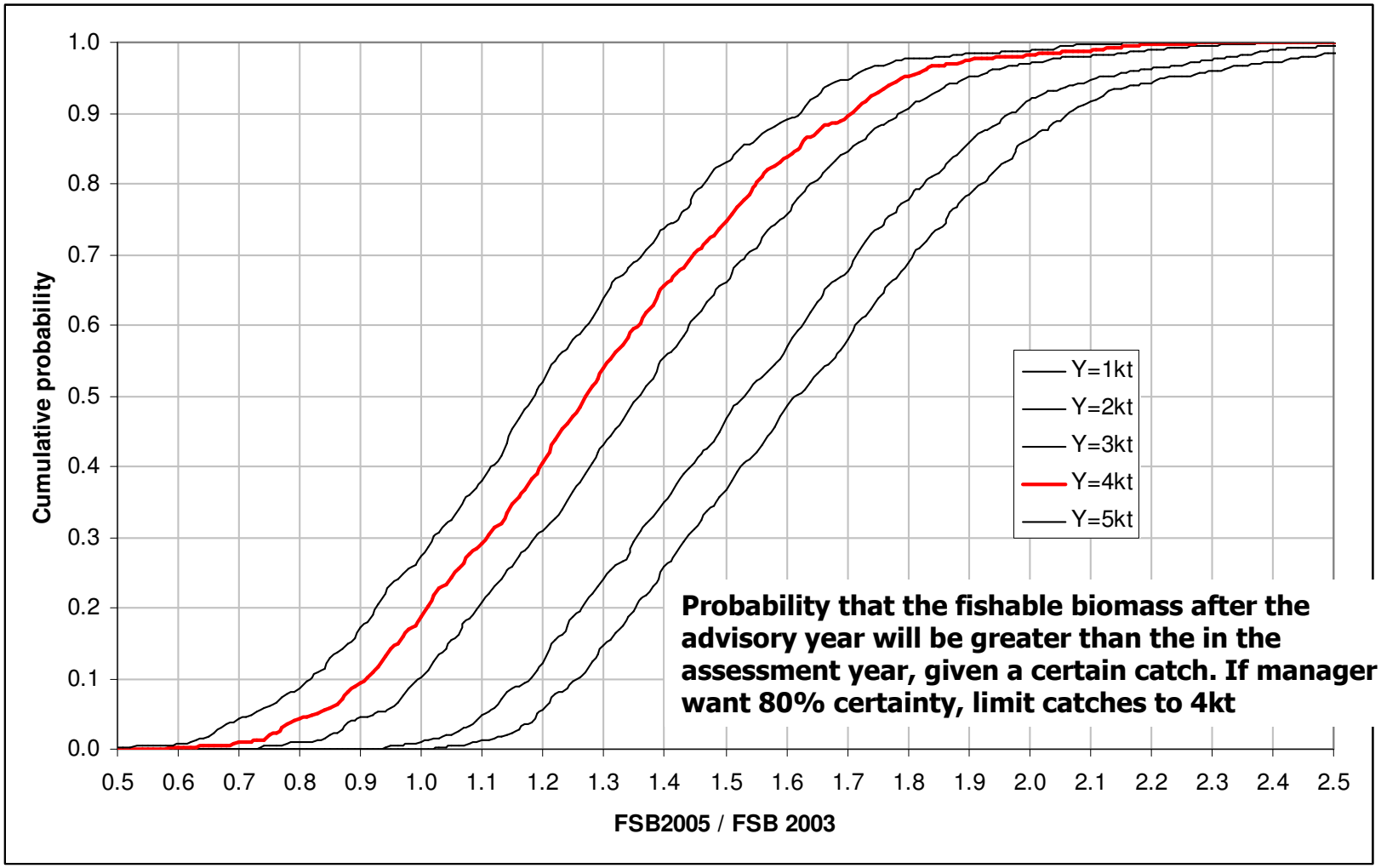
Advantages: Can address questions?

Is F increasing?

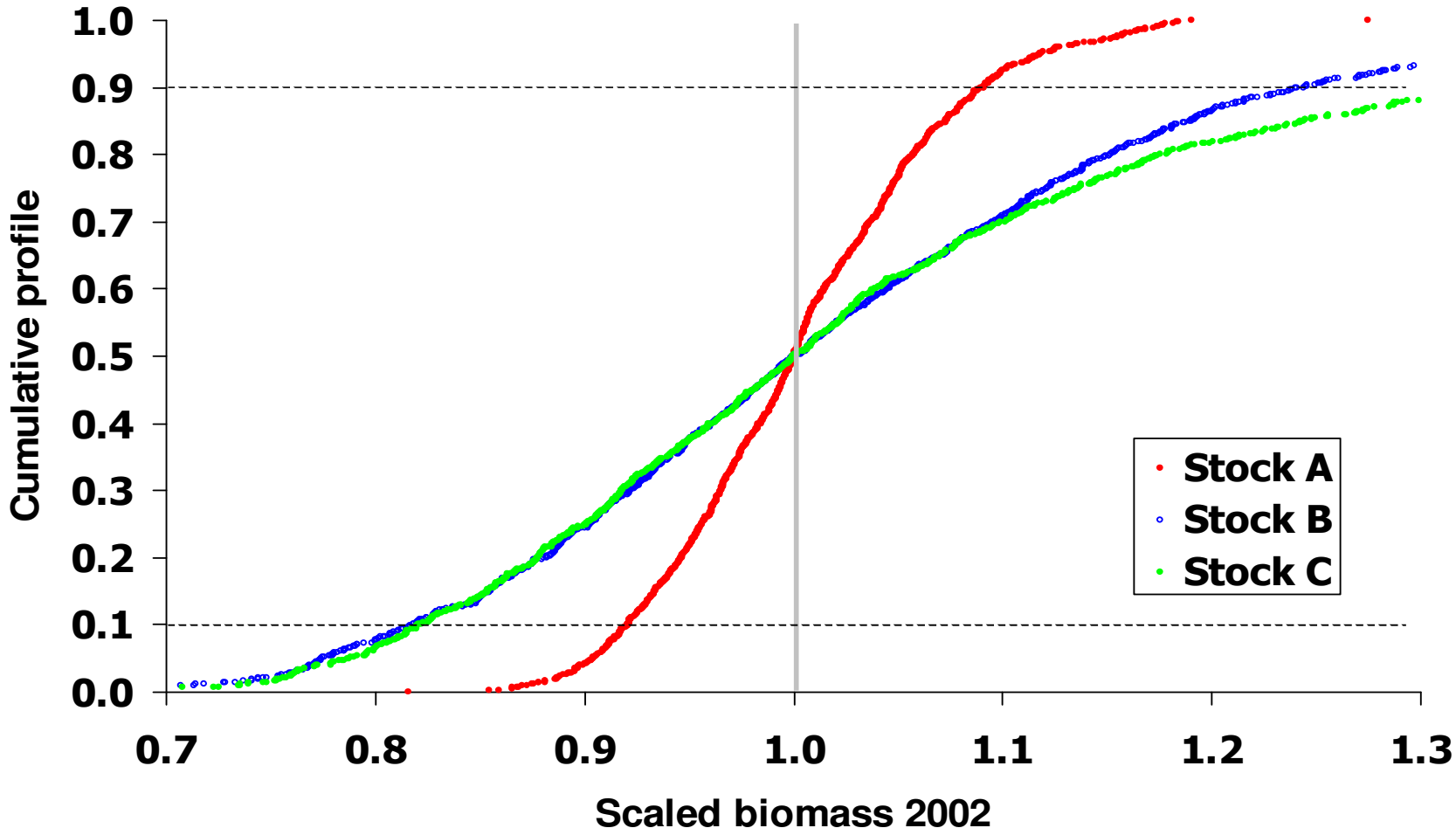


2500 Bootstrapps from xCAM: The results show that the 80th percentile of the fishing mortality in the year 2000 (F2000,80th=0.48) is below the 20th percentile value of the fishing mortality in 2002 (F2002,20th=0.55). This suggests that the increase in fishing mortality from 2000 to 2002 reflects very likely a true increase in fishing mortality over the period. The analysis indicate that the most sensible value of F to use in the assessment year is the terminal year value, not the average of the last three years.

Advantages: Get rid of point estimators



Advantage: Relative error among stocks



Bootstrap does not solve all problems (or for that matter most uncertainty estimating methods)

Assessment uncertainties

- 1) Natural variations
- 2) Observation errors in the input data
- 3) Model misspecifications
- 4) Implementation errors

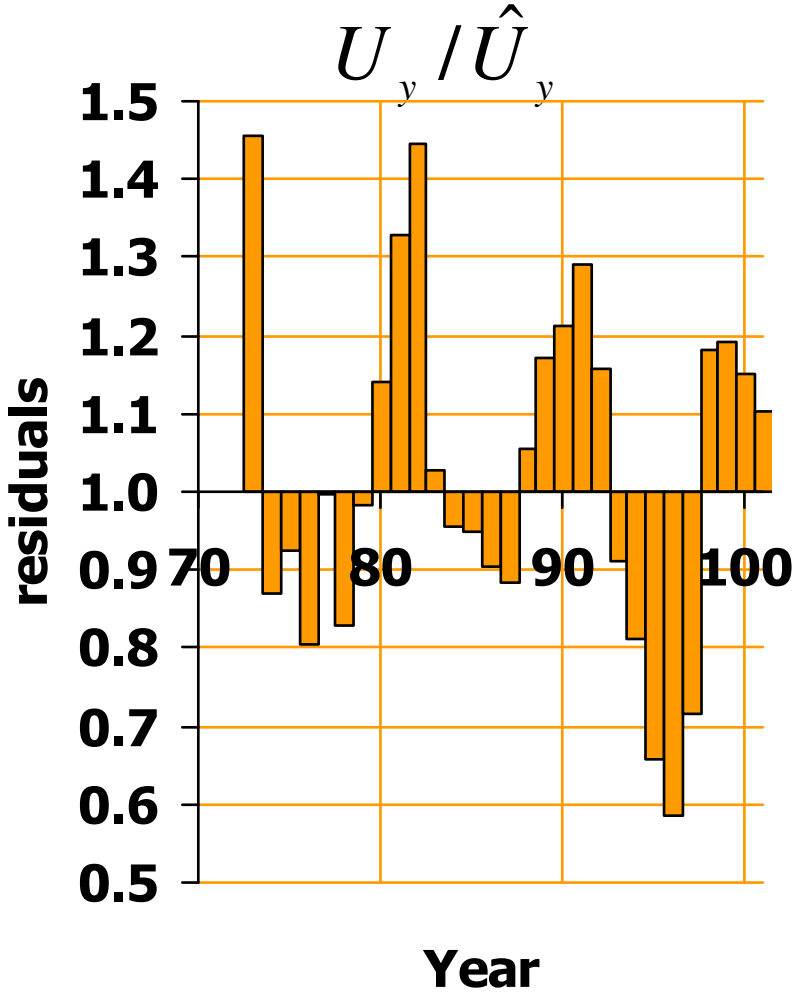
- By bootstrapping we are only estimating the noise in the observations given the model assumptions made.

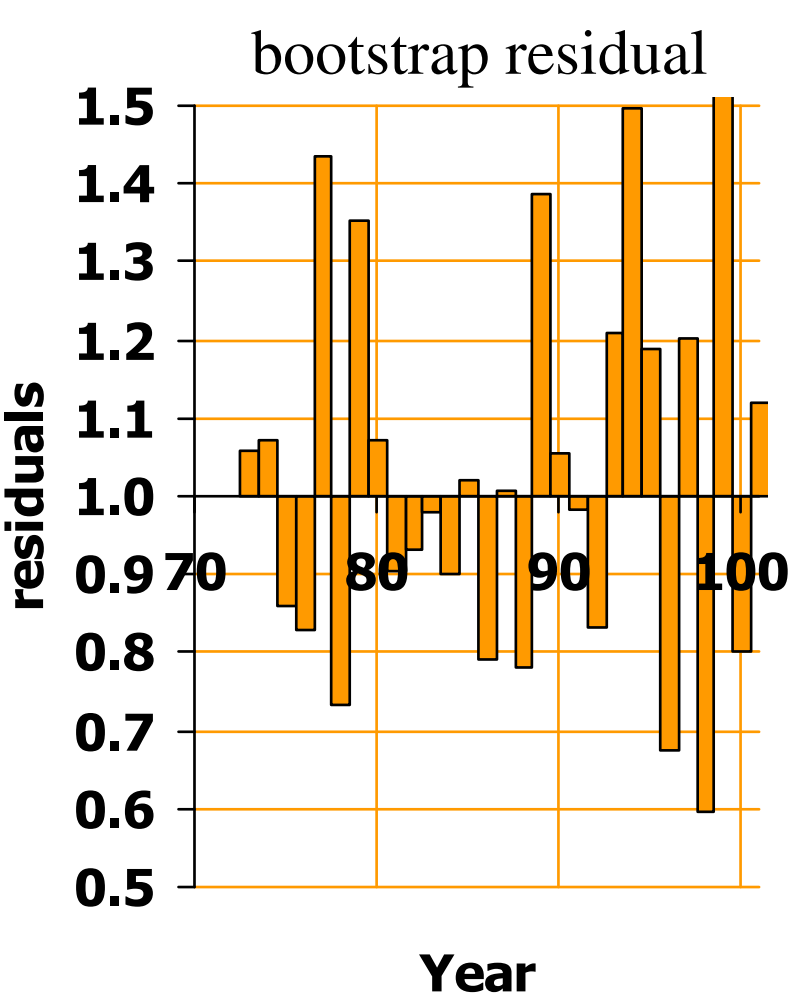
How to resample the residuals?

- The way one resamples the residuals gives different bootstrap confidence profiles.
  - E.g. randomly resampling the whole age year matrix of the survey residuals gives different bootstrap probability profile compared with randomly resampling a whole year block.
  - This problem is in some way analogous to how one creates error structures when generating simulated (artificial) data.
  - No clear answer here, this is a developing field.

Loss of autocorrelation

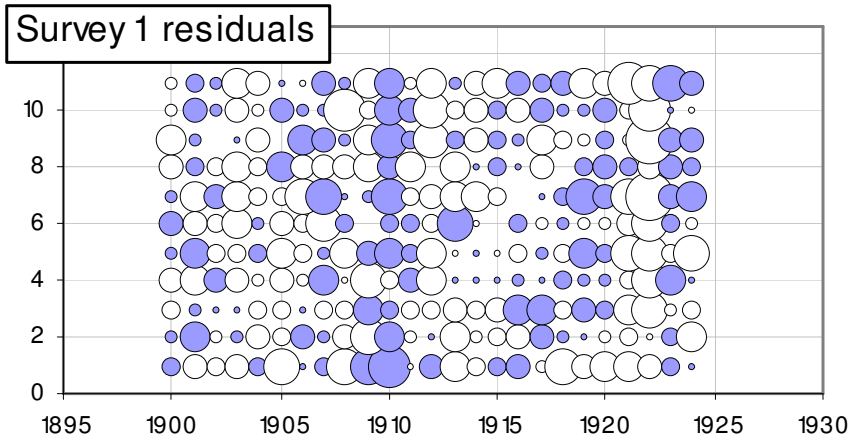
Residuals autocorrelated in the original sample, random in the bootstrap



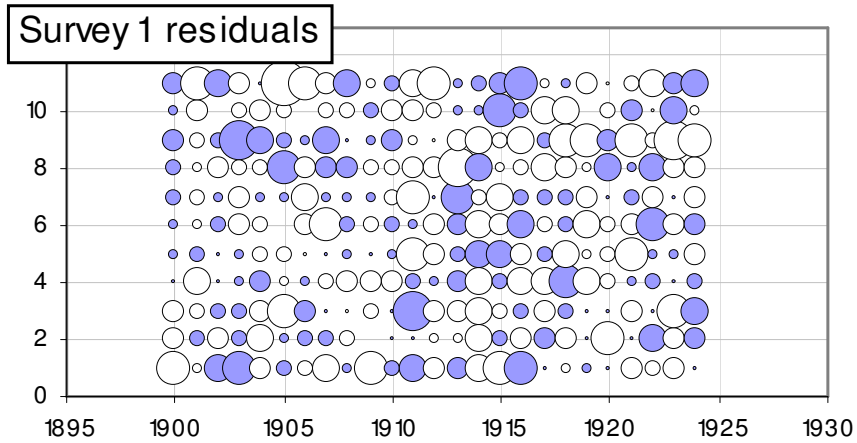


Loss of structure in original residuals

- Original residuals
  - Clear year effects in the residuals from the measured data

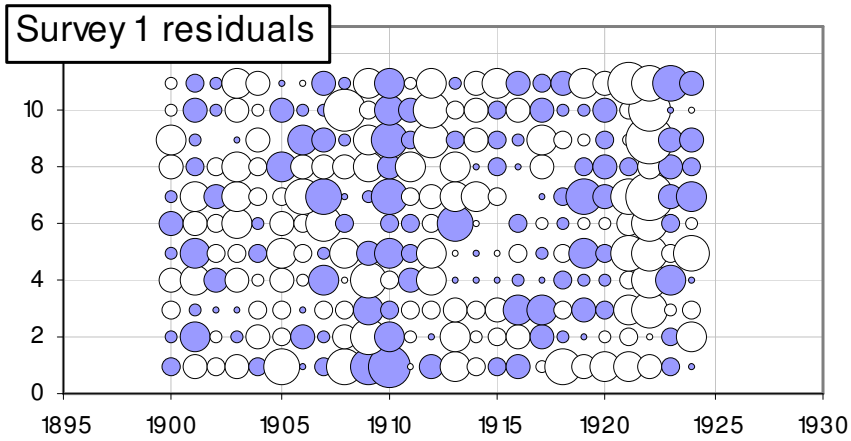


- Bootstrap residuals
  - Loss of structure when randomly select the whole age year matrix

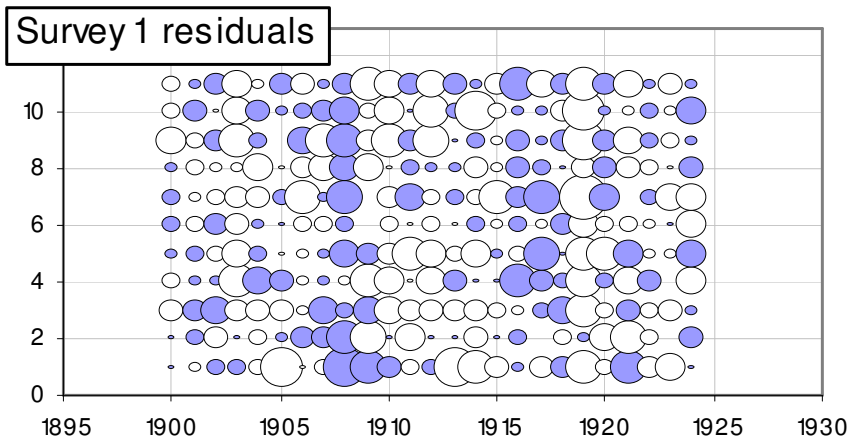


Loss of structure in original residuals

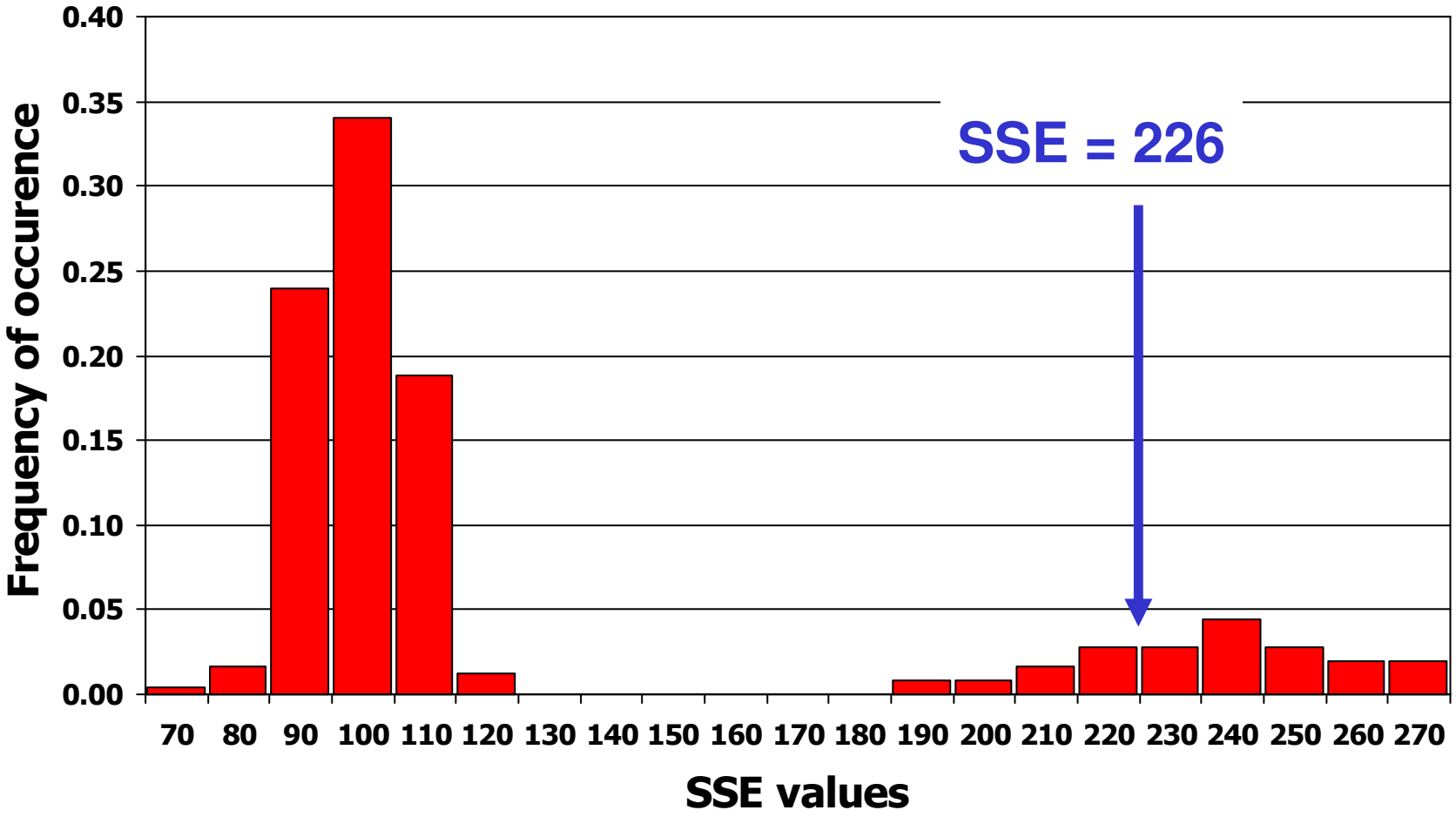
- Original residuals
  - Clear year effects in the residuals from the measured data



- Bootstrap residuals
  - Resampling the whole year retains the year factor in the residuals

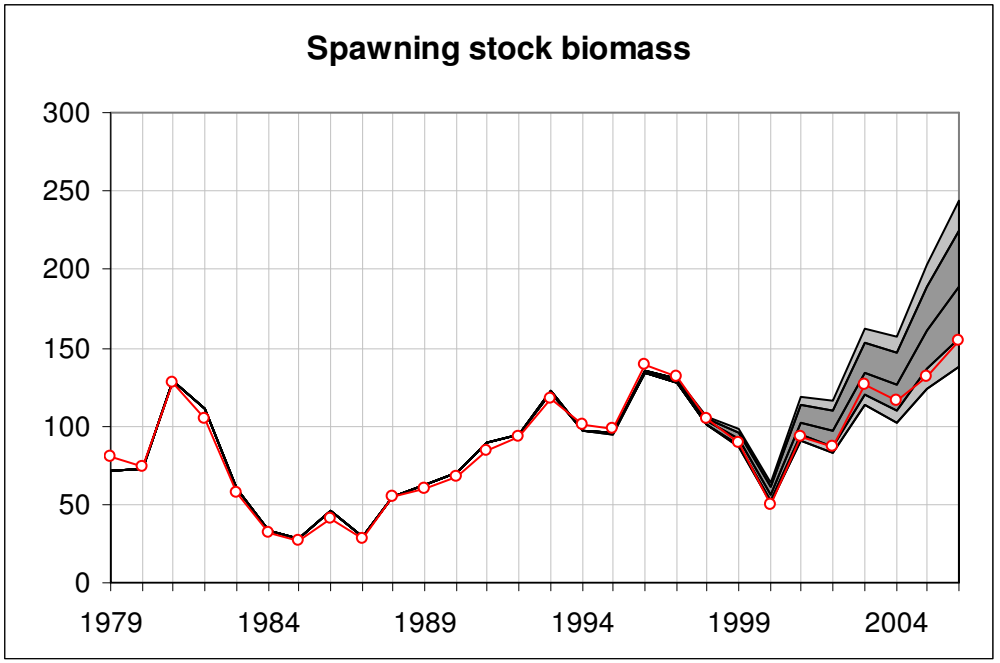


The SSE profile of the bootstrap!



Different models – different results

ADAPT model with bootstrap confidence interval.
Red line: XSA, point estimates.



Mean value not the same, XSA point value on the lower tail of the distribution profile of the ADAPT bootstrap confidence interval. Both models use the same input data – difference in results are related to model assumption.
Sometimes the value from one model does not overlap with the confidence distribution of another model, even when using only one survey. This indicates that the assumptions in the model(s) have a larger Influence on the final results than the noise in the data.

Final point

- Bootstrap seems like the magic thing and it looks very impressive.
  - It is a helpful explorative tool
  - It represent the noise in the data given the model assumption. Thus not a true confidence profile of the total assessment error.
    - Should thus talk of “pseudo-confidence intervals” or “bootstrap confidence interval”.
- The same rule applies with using bootstrap as with any other tools:
  - Understand how it is implemented in the particular software package that you may be using and ask if that is appropriate to the data set that you have.