



Biostatistics IV

An introduction to bootstrap

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.

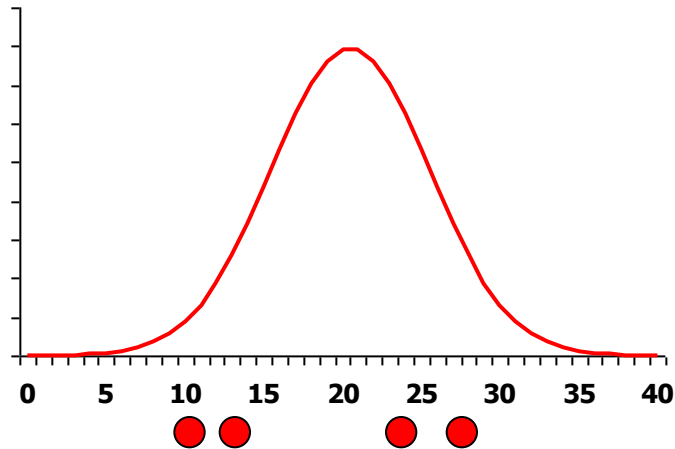
The original version of the tale is in German, where Munchausen actually draws himself up by the hair, not the bootstraps. The figure on the left refers to German version of the story.

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, \dots, 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 $\hat{\Theta} = t(x)$. Some measure of the estimate’s accuracy is as important as the estimate itself; we want a standard error of $\hat{\Theta}$ and, even better a confidence interval on the true value Θ .”
 - Efron and LePage (1992)

Our sample mean



4 sample →

- From the 4 samples we take a mean and calculate a standard deviation!
 - By doing this we are assuming that the population is normally distributed
 - But what if we don't know the distribution?

The solution: Bootstrap

- Generate large numbers of “bootstrap” data sets, $x_1, x_2, x_3, \dots, x_b$, 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.

Resampling the sample

The sample i Data	Bootstrap sample number					
	1	2	3	4 ...	n	
1	3	7	8	5	3	7
2	5	3	9	3	7	11
3	7	3	11	7	8	3
4	8	5	11	5	9	11
5	9	8	5	8	3	9
6	11	9	8	3	3	3
The mean	7.17	5.83	8.67	5.17	5.50	7.33
The std. Deviation	2.86	2.56	2.25	2.04	2.81	3.67

- Within a bootstrap sample the values are drawn randomly from the original sample.
 - Could thus get some measurements more than once within a particular bootstrap sample

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!



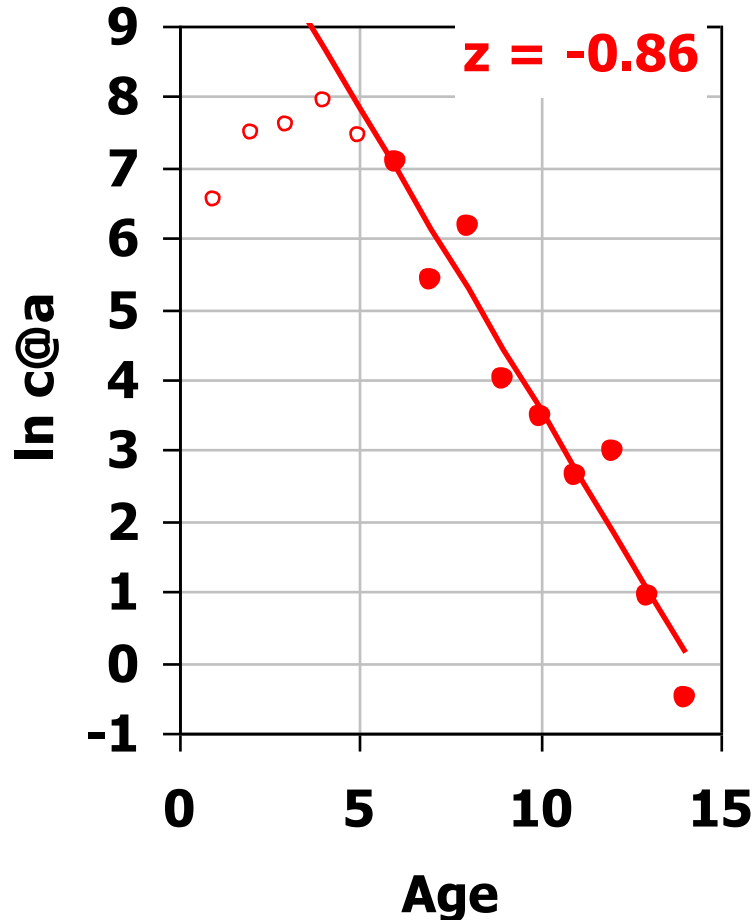
Bootstrapping the residuals

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

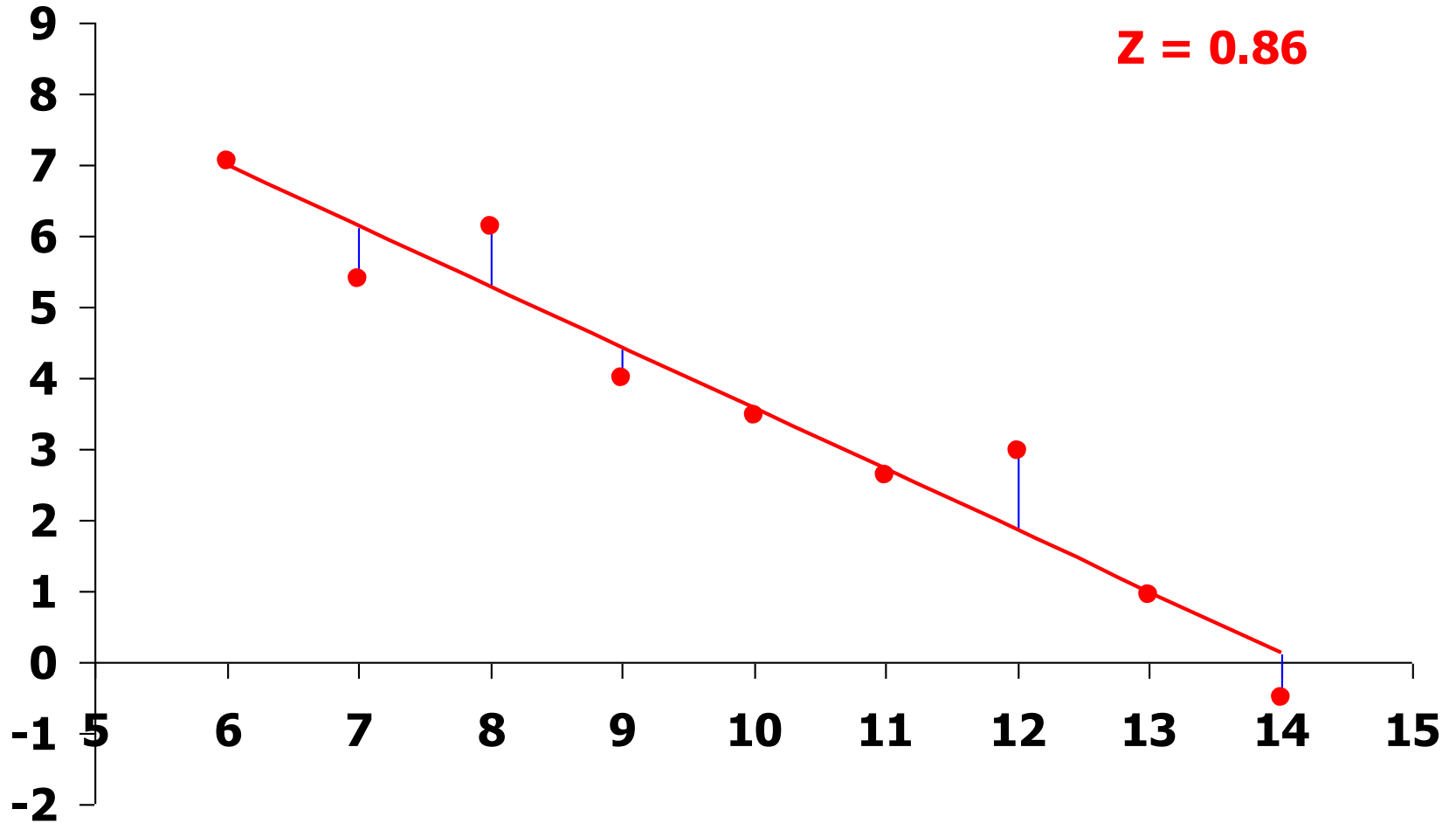


Original data set and statistics



- Data set: $c@a$ of one cohort
- Analysis: Slope estimate of fully recruited fish
- $Z = -\text{Slope} = \text{Total mortality}$
- Task: Obtain some information about the confidence interval of the Z using bootstrapping techniques.

The residuals to resample



One bootstrap sample

Original data set

Age	Observed In c@a	Predicted In 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 from draw	Bootstrap In 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 ...

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

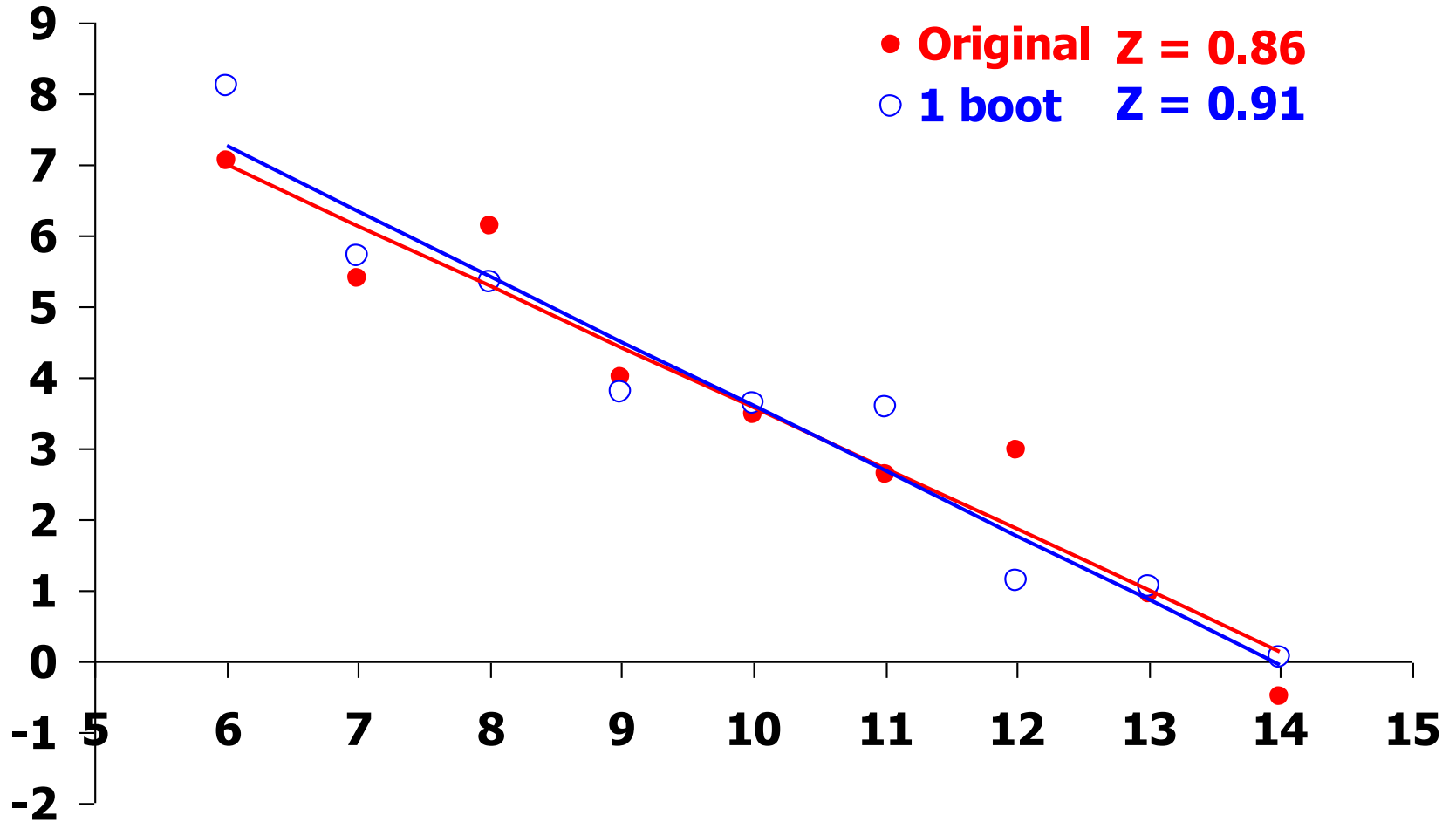
$$\varepsilon_{a^*} = K_a - \hat{K}_a^*$$

where

$$K_a = \ln(C_a)$$

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

Original and 1 bootstrap sample



n bootstrap samples

1 Bootstrap sample

Residual		
Random	value for draw	Bootstrap In 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

Residual		
Random	value for draw	Bootstrap In 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

Bootstrap In c@a

6.346
6.200
6.143
4.489
3.468
2.612
1.795
0.355
1.008

Slope (-Z)
0.82

Bootstrap In c@a

7.855
5.710
4.634
3.695
3.477
2.067
1.921
0.910
-0.501

Slope (-Z)
0.91

Bootstrap In c@a

6.346
6.045
6.143
4.324
4.684
2.777
1.211
0.910
0.054

Slope (-Z)
0.87

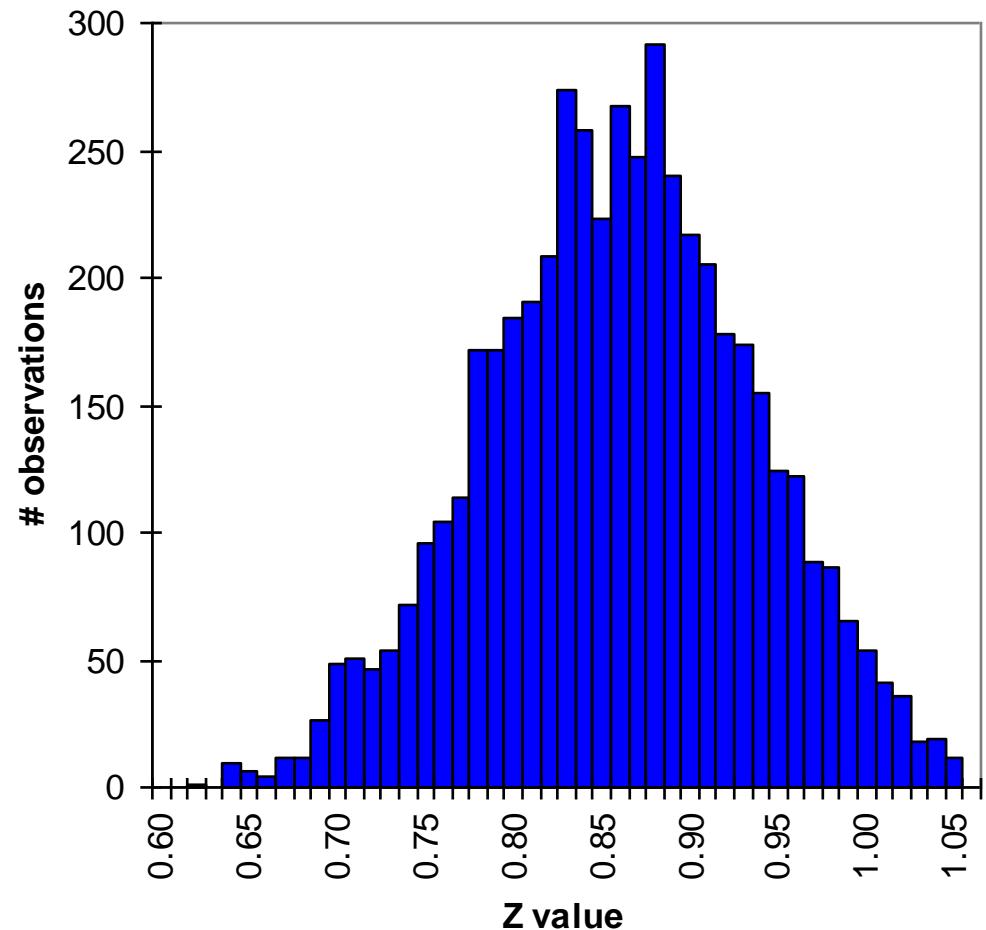
Bootstrap In c@a

7.056
6.999
4.634
5.540
3.506
2.622
1.127
0.271
0.054

Slope (-Z)
0.97

5000 bootstrap samples: Distribution

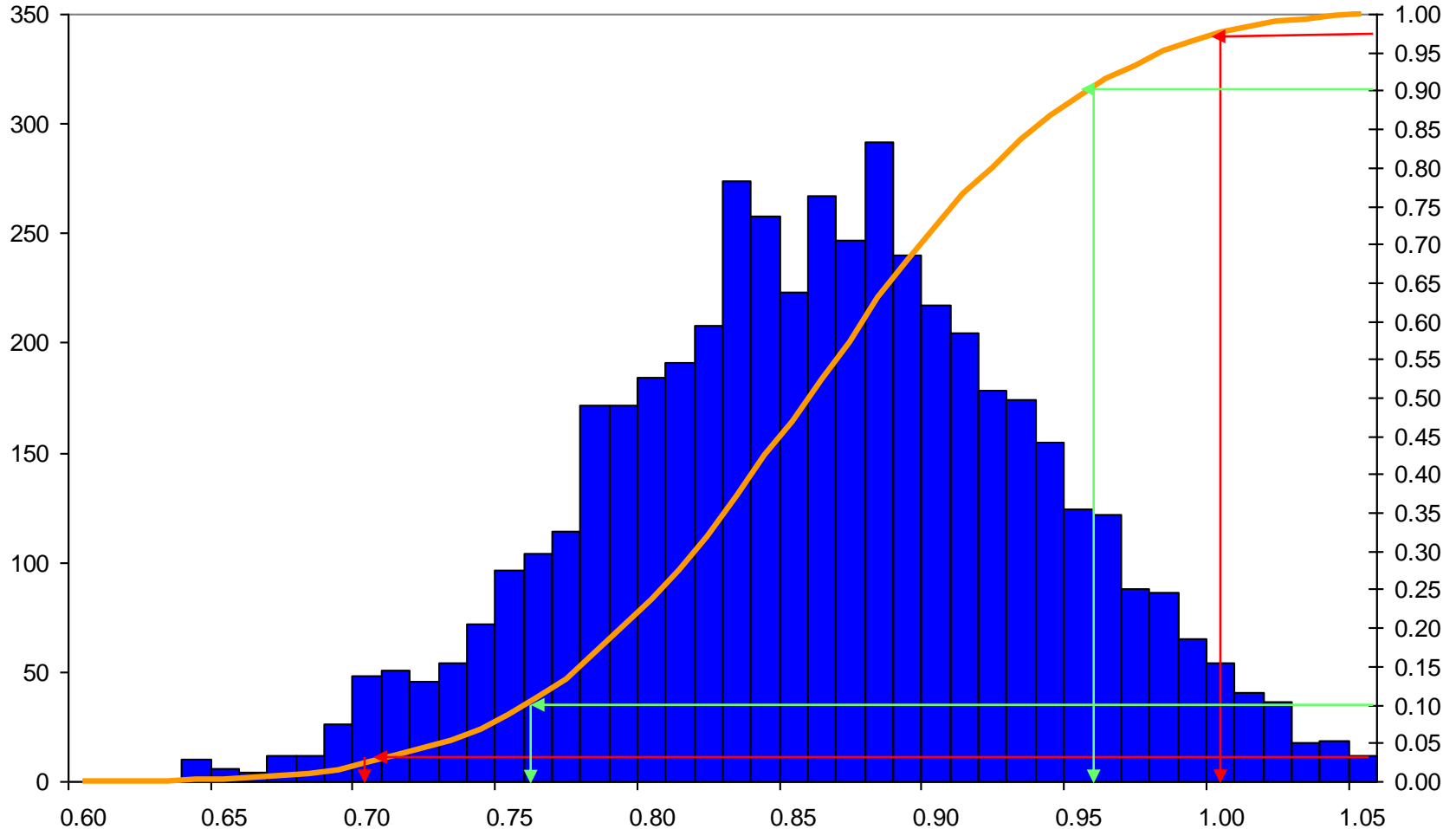
Bootstrap #	Z
1	0.896
2	0.858
3	0.809
4	0.986
5	0.967
6	0.920
7	0.885
8	0.858
9	0.891
10	0.918
11	0.932
12	0.898
13	0.865
14	0.906
15	0.973
...	...
...	...
4992	0.817
4993	0.666
4994	0.814
4995	0.840
4996	0.890
4997	0.756
4998	0.887
4999	0.702
5000	0.716



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 $100(1-\alpha)\%$ CI around the sample estimate of θ is obtained from the two bootstrap estimates that contain the central $100(1-\alpha)\%$ of all b bootstrap estimates.

80% & 95% confidence interval



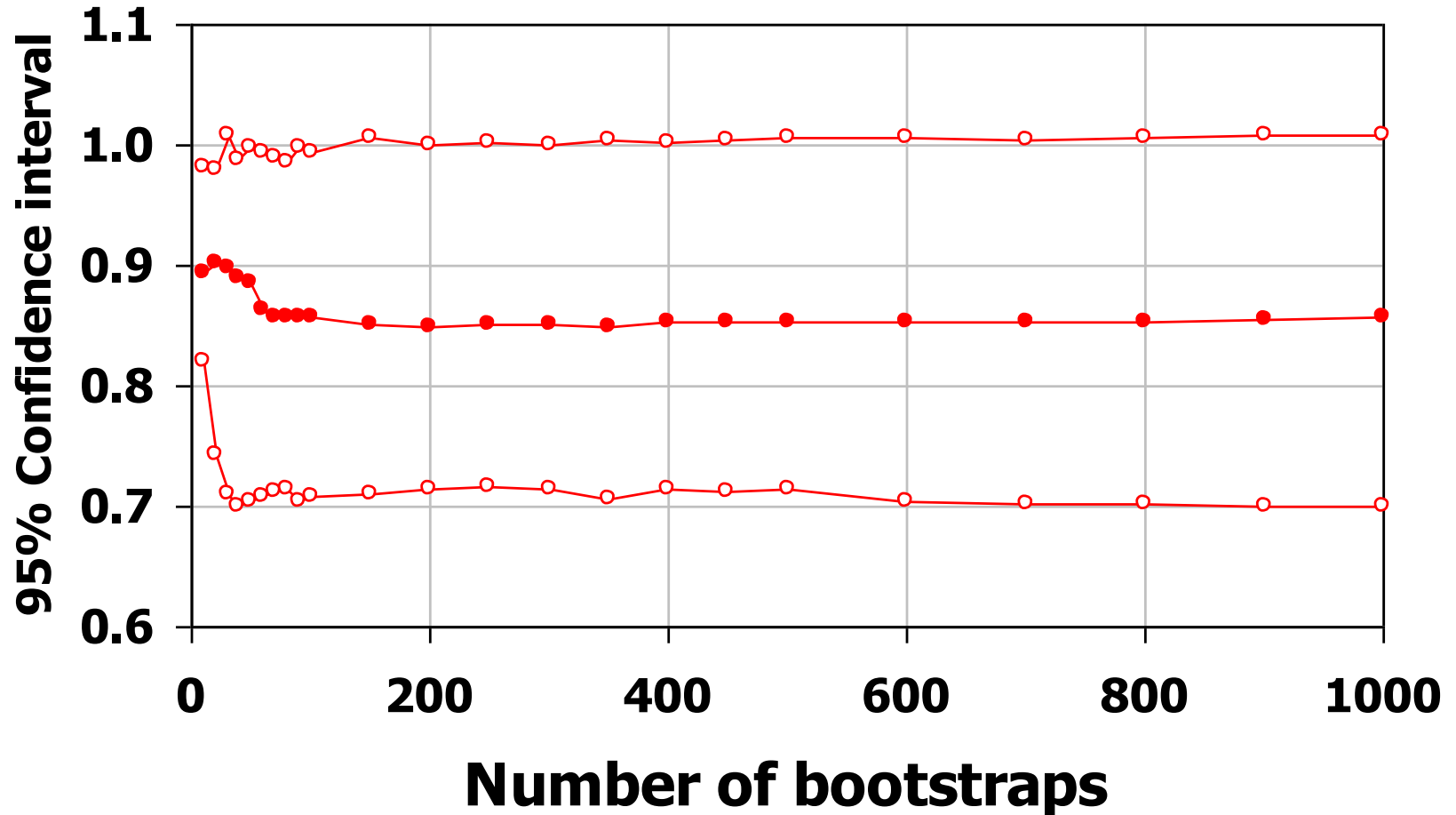
Parametric confidence interval

- If the parameter estimated is **expected to follow a normal distribution**, the C.I. may be obtained from the usual:
 - $CI = \Theta \pm t_{n-1, \alpha/2} se_{\Theta}$
 - where:
 - Θ : sample parameter estimate
 - $t_{n-1, \alpha/2}$: 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 ----- >

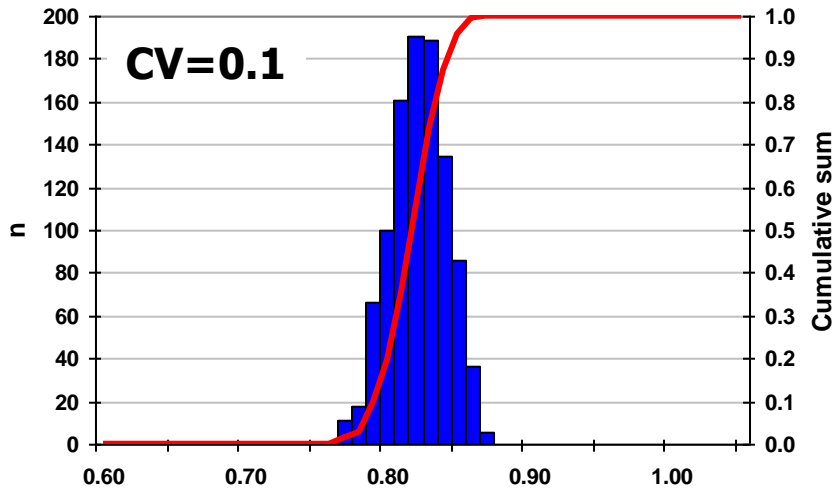
95% CI of Z and bootstrap numbers



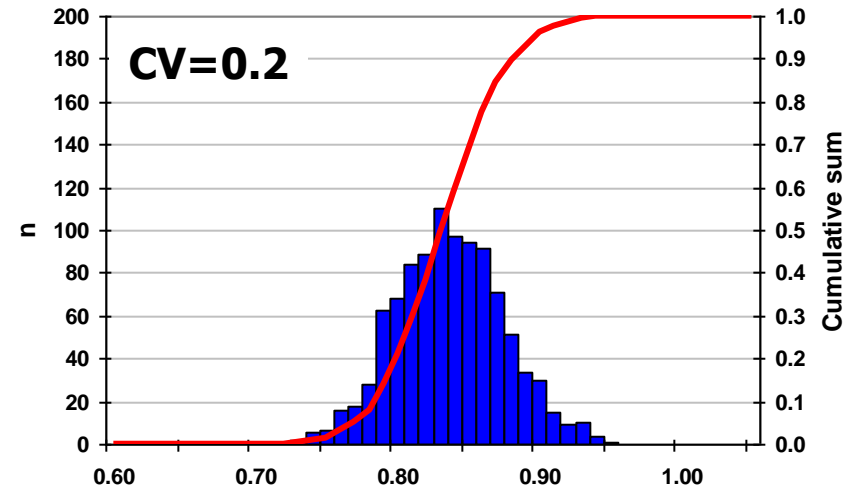
In this simple case do not need much more than around 200 bootstraps to get CI

Different CV in catches: 1000 bootstraps

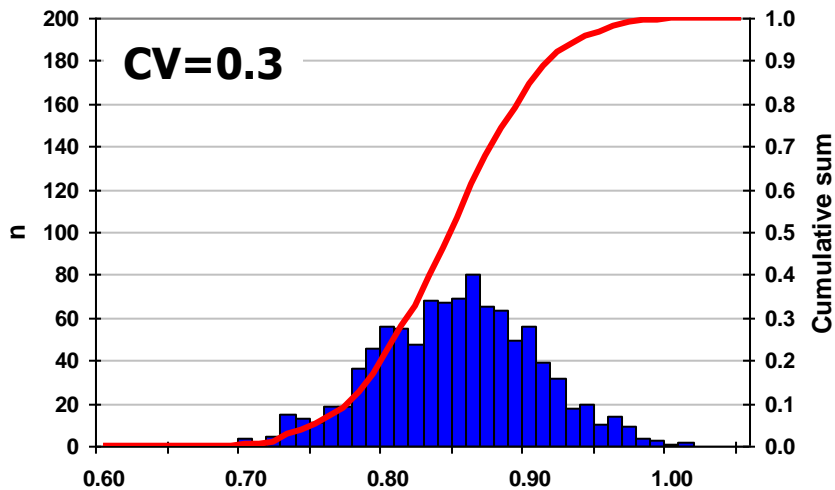
Z bootstrap distribution



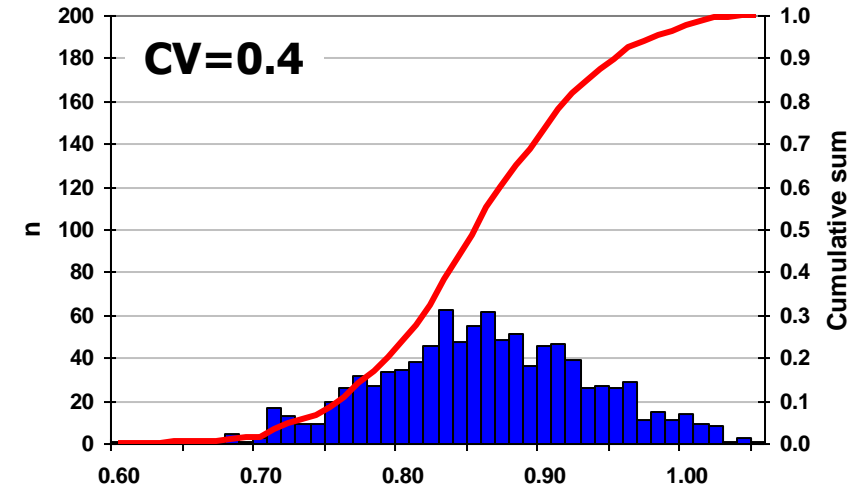
Z bootstrap distribution



Z bootstrap distribution



Z bootstrap distribution



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 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.