SAS Workshop: Large Claims Modelling

Please install solver and analysis toolbox add-in

Introductions

Introduction

- The Working Group
- You…
- Getting the Spreadsheets Working on Your PC

Installation

- Two spreadsheets
- Remember to
  - Switch on macros
  - Switch on Microsoft’s analysis add-in

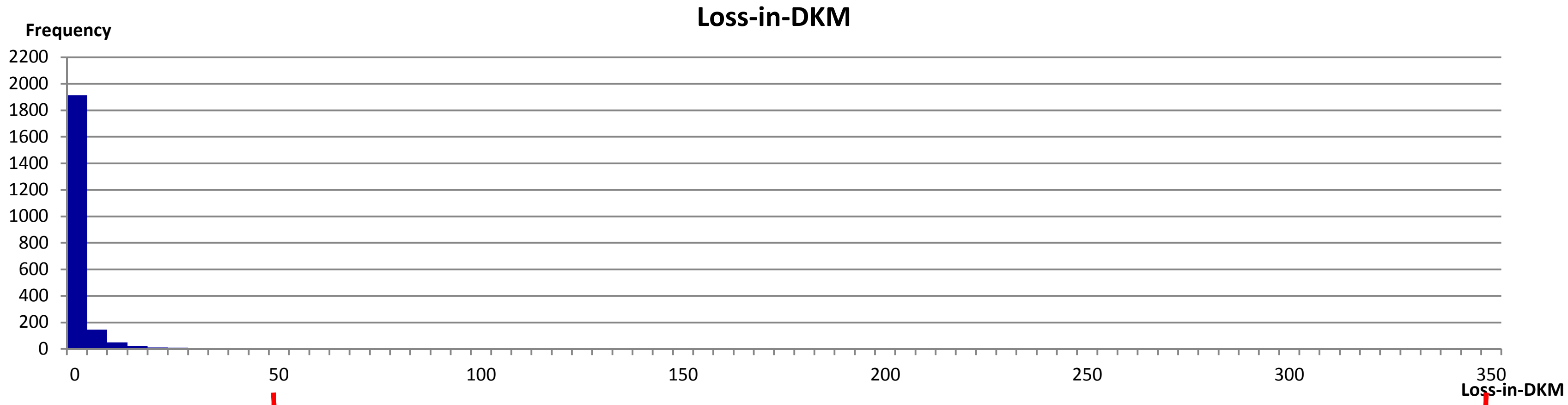
What Is This Workshop About?

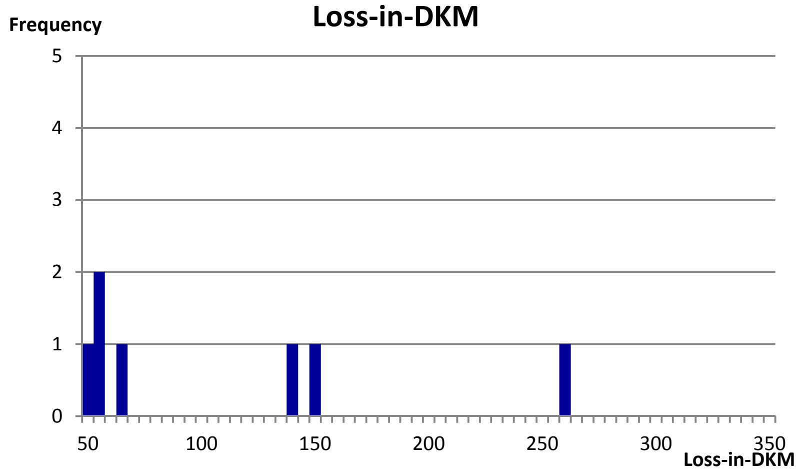
Why Fit Distributions to Tails?

Why do we need to fit a tail to a distribution – some real life examples

- Pricing non-proportional covers
- Reinsurance requirements and optimisation
- Capital modelling & Enterprise risk management
- Capacity management, including PML estimates

How unreliable the historical data in estimating large loss – A closer look at the scarcity of data





- More than 99.7% of the data centred around smaller losses (losses below $50m)
- Hence, how should we price for the largest and rare losses (the tail)?
- A right model is necessary to estimate the tail (losses above threshold)

What are we trying to achieve today?

Task

- Choosing a Distribution
- Choosing a Cutoff
- Choosing Distribution Parameters
- Parameter Uncertainty (Bootstrapping)
- Pricing an XOL
- Estimating PMLs

The Data

Two Data Sources

- Danish Fire Data
  - collected at Copenhagen Reinsurance
  - 2167 fire losses over the period 1980 to 1990
  - adjusted for inflation to reflect 1985 values
  - http://www.ma.hw.ac.uk/~mcneil/data.html
- US Motor Data
  - Available on Kaggle, sourced from Allstate
  - Competitors outperformed actuaries by 271%
  - 1000 largest claims
  - 3 years of data
  - http://www.kaggle.com/host/casestudies/allstate

Section

Choosing A Distribution

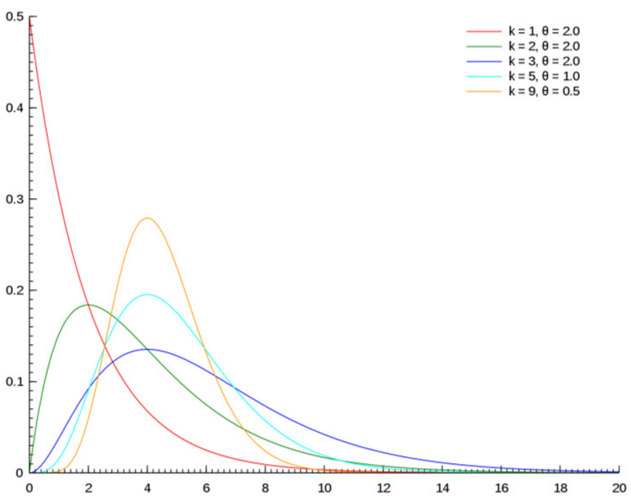
Extreme Value Theory

- It is known that under general conditions and if the threshold is sufficiently high, the excess claim follows extreme value distributions.
- Conditions:
  - No upper bound for claim
  - Continuous pdf

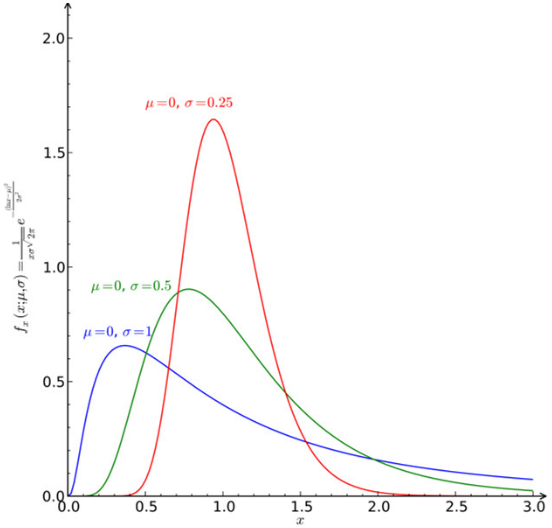
Choose Distribution

Distribution	Parameter	Usage
Gamma	k, θ	Waiting time
Log-normal	μ, σ	Finance price
Pareto 1	α	Distribution of wealth
Weibull	k, λ	Time to failure
Exponential	λ	Time between events
Generalized Pareto	σ, ξ, (μ)	Tail of distribution

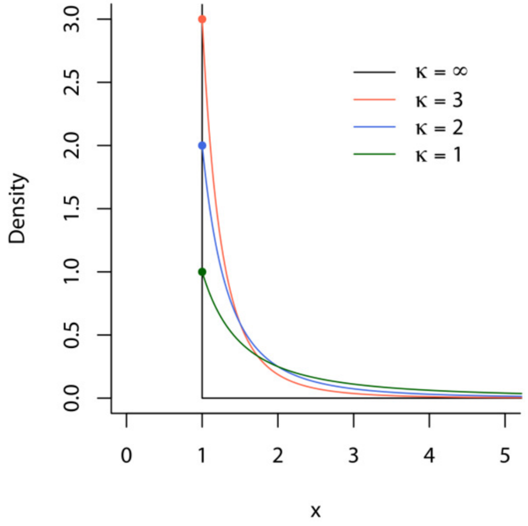
Gamma



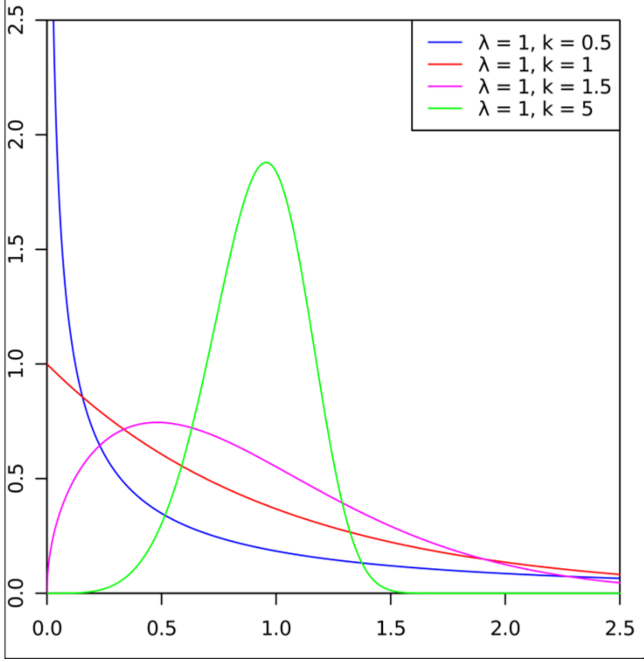
Log-normal



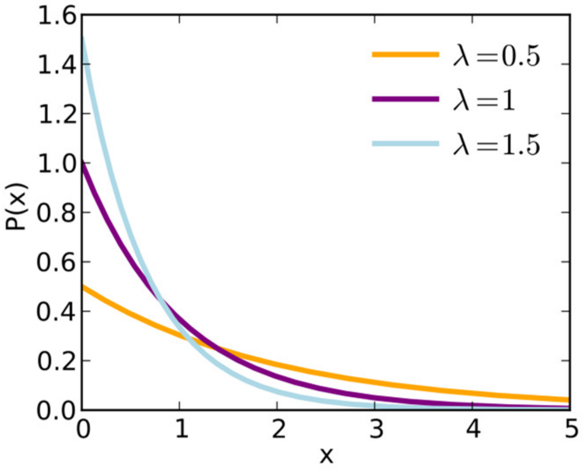
Pareto1



Weibull



Exponential



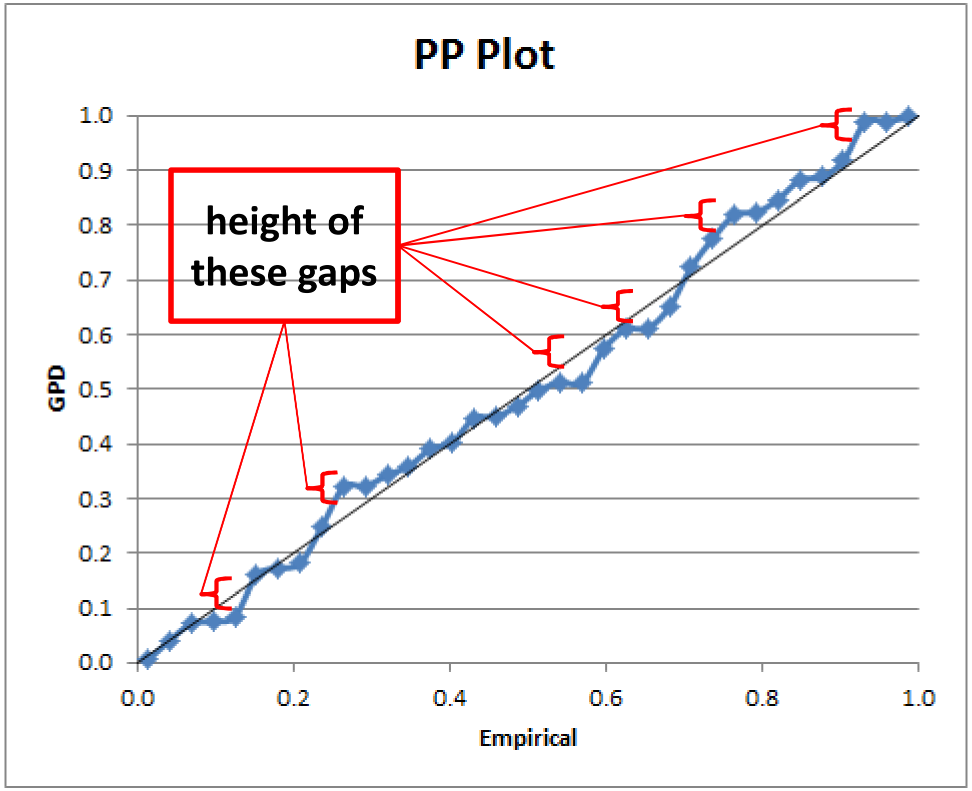
General Pareto

Similar to Pareto 1

Figures from Wikipedia

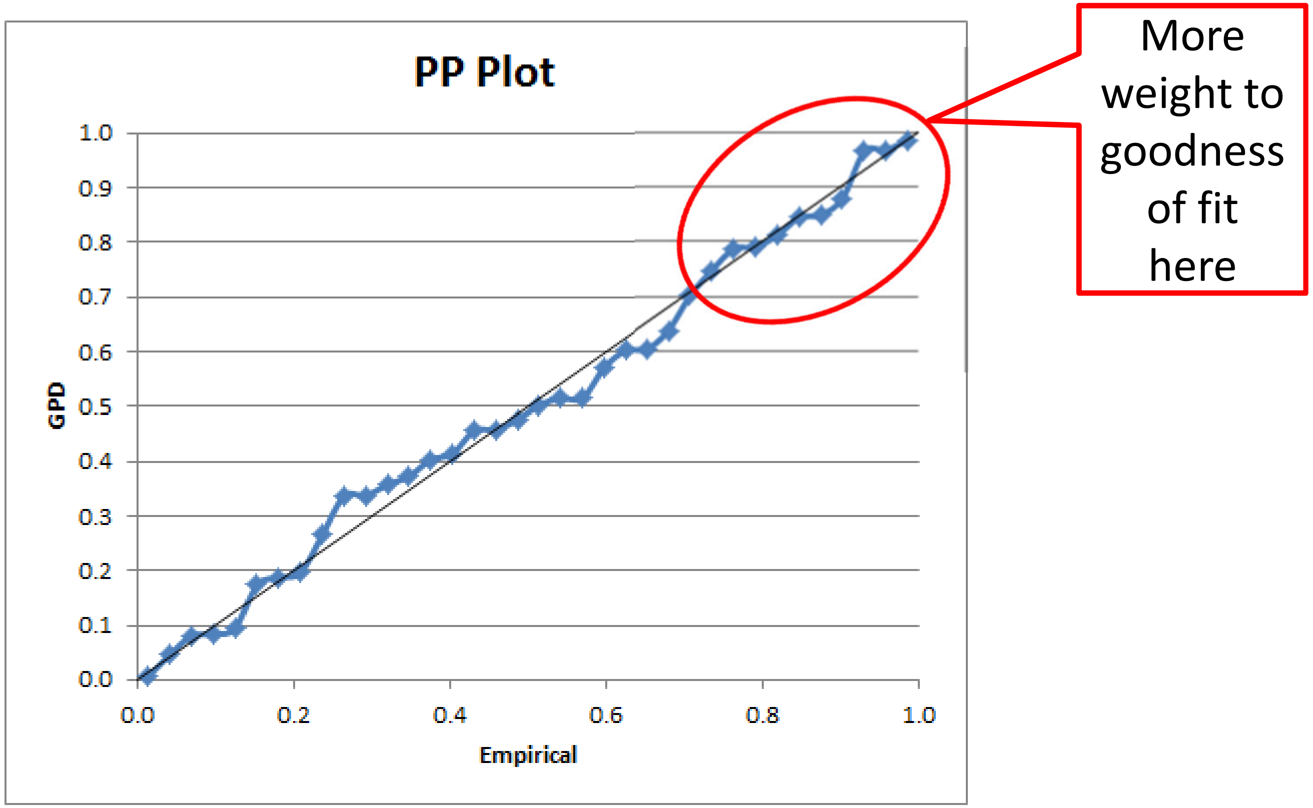
Kolmogorov-Smirnov

- K-S score is largest gap in PP plot



Anderson-Darling

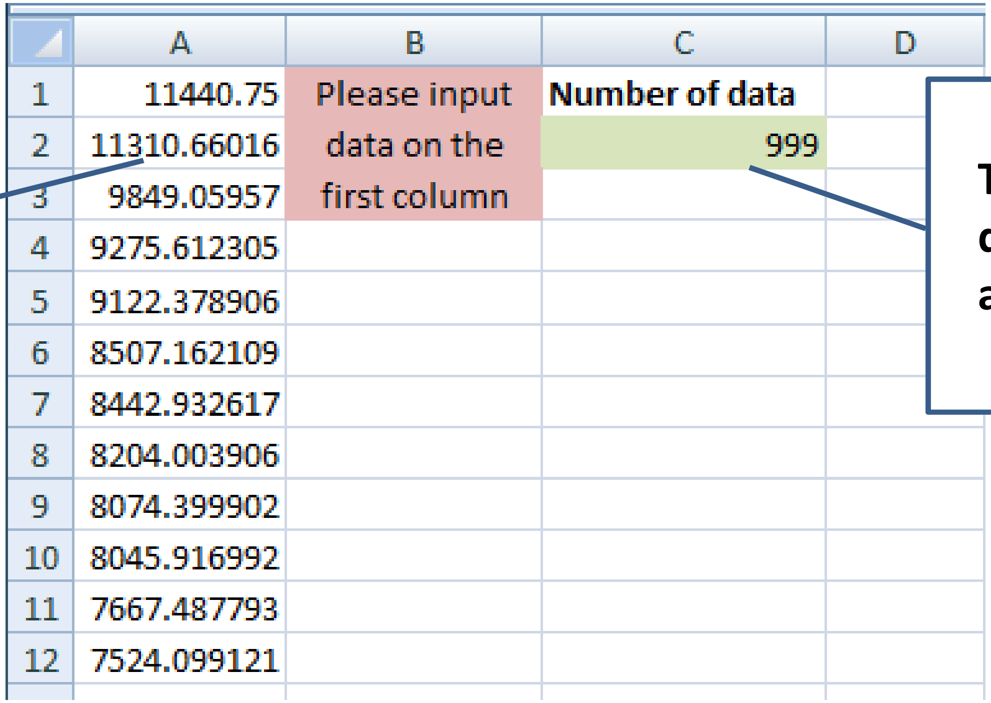
- Similar to Kolmogorov-Smirnov but with greater weight given to fit of higher values



More weight to goodness of fit here

How to use

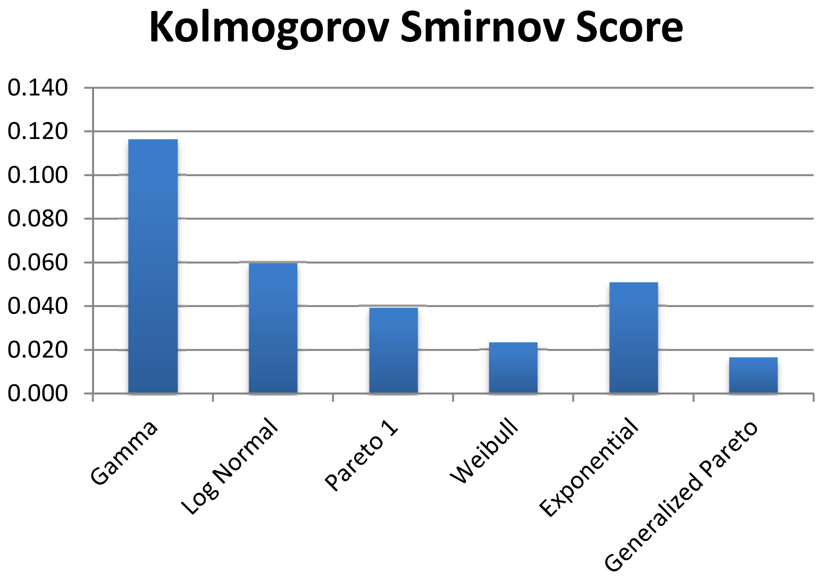
Paste data here, sort by descending order to avoid problem.
All data must be numbers.

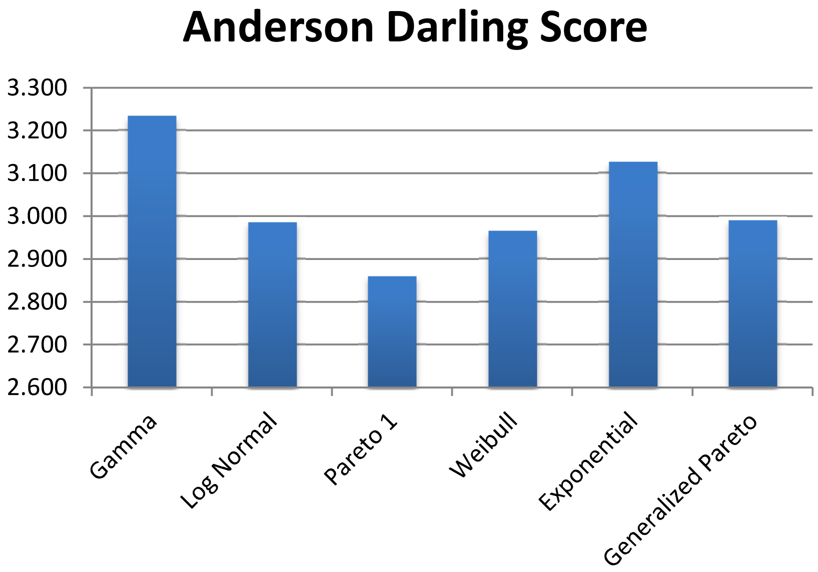


	A	B	C	D
1	11440.75	Please input data on the first column	Number of data	
2	11310.66016		999	
3	9849.05957			
4	9275.612305			
5	9122.378906			
6	8507.162109			
7	8442.932617			
8	8204.003906			
9	8074.399902			
10	8045.916992			
11	7667.487793			
12	7524.099121			

Total number of data calculated automatically

Choose Distribution





- All parameters are calibrated via MLE
- Low KS and AD score means better fit

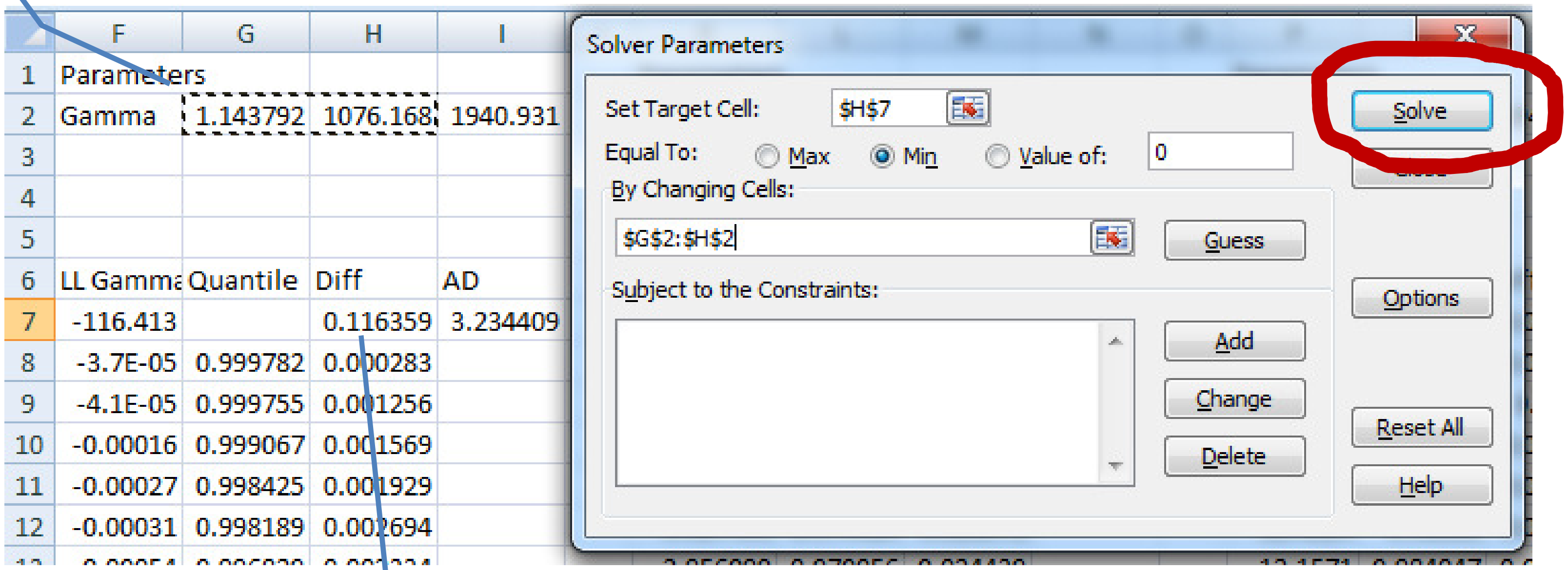
Other calibrations

- MLE is the most efficient to calibrate
- You may aim for best fit for KS/AD
- Use “solver” to calibrate
  Data -> Analysis -> Solver
- You may need to install “solver” add-in

How to use “solver”

Target parameters

Solve it!



	F	G	H	I
1	Parameters			
2	Gamma	1.143792	1076.168	1940.931
3				
4				
5				
6	LL Gamma	Quantile	Diff	AD
7	-116.413		0.116359	3.234409
8	-3.7E-05	0.999782	0.000283	
9	-4.1E-05	0.999755	0.001256	
10	-0.00016	0.999067	0.001569	
11	-0.00027	0.998425	0.001929	
12	-0.00031	0.998189	0.002694	

Solver Parameters

Set Target Cell: $H$7

Equal To: ○ Max ● Min ○ Value of: 0

By Changing Cells: $G$2:$H$2

Subject to the Constraints:

First number under “Diff” is KS score, target this cell

Notes on Solver

- Take MLE estimators as initial guess
- Solver algorithm takes time
- Best fit to AD score is calibrated similarly
- Note the parameters for KS-fit and AD-fit are different

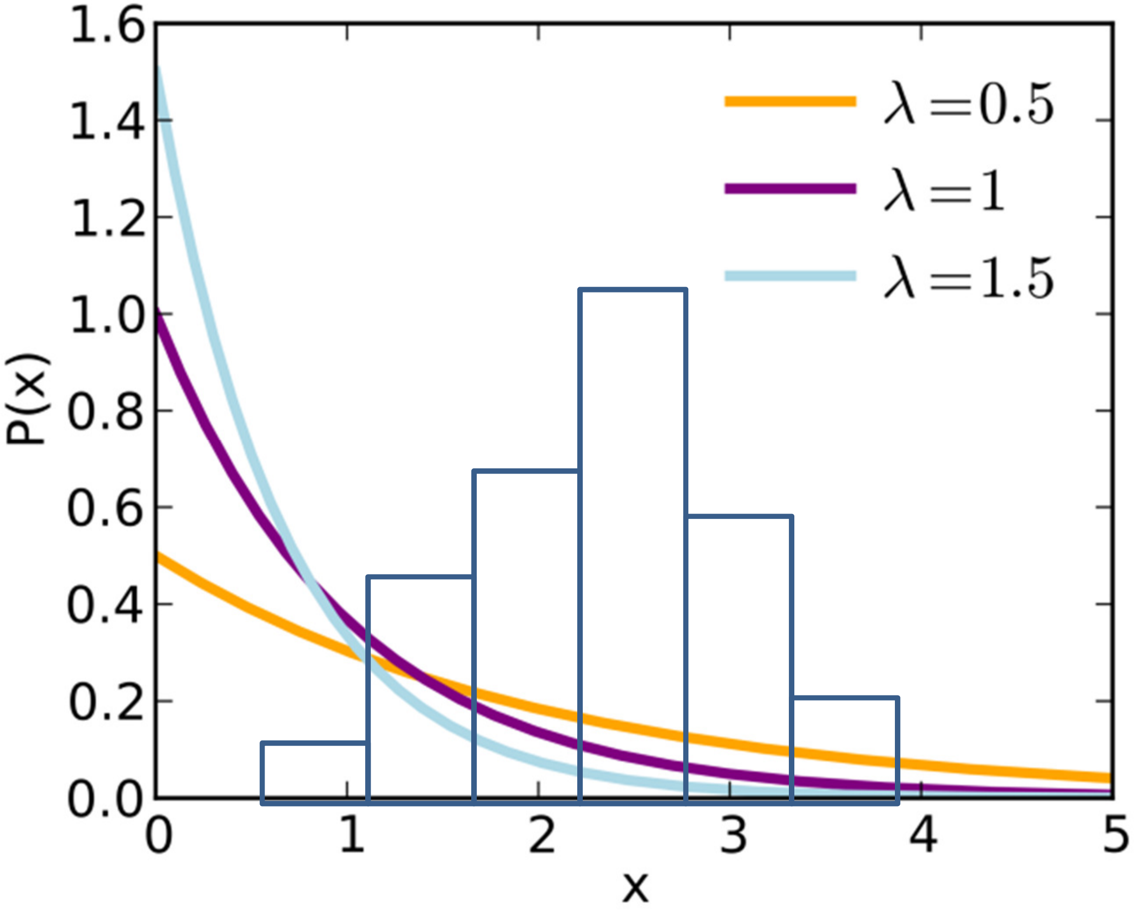
Task 1

- Split into groups, each group take one set of data and present the results
  - Danish data top 1000
  - Danish data top 200
  - US data top 1000
  - US data top 200
- For computation reason, we test 1000 data at most
- Random data will not fit extreme value distributions and estimation will fail

Task 1

- Breakout groups to report back
  - Which distribution?
  - Which distribution parameters to use?

Fit random data

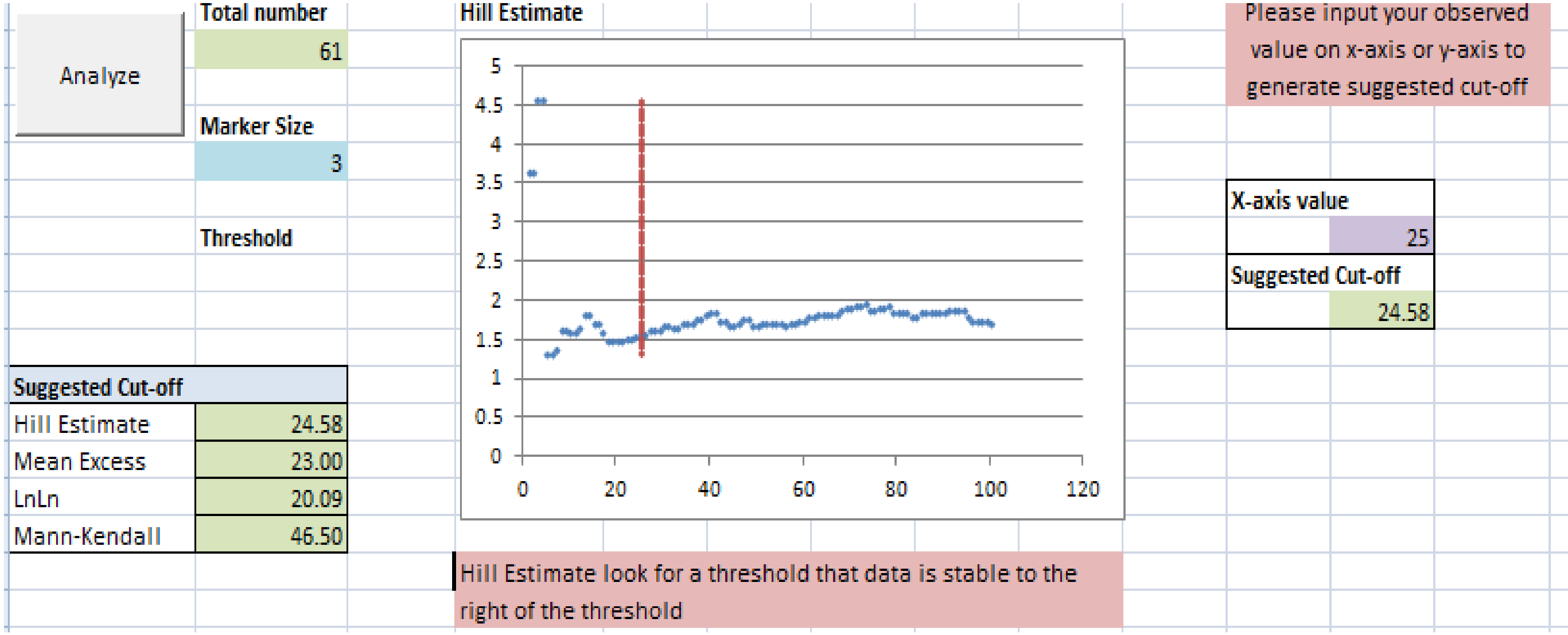


Choosing A Cutoff

Basic structure

- Inputs and outputs

🟪	**Required Inputs**
🟦	**Optional Inputs**
🟩	**Outputs**
🟥	**Notes**



Total number
61

Marker Size
3

Threshold

Suggested Cut-off	
Hill Estimate	24.58
Mean Excess	23.00
LnLn	20.09
Mann-Kendall	46.50

Please input your observed value on x-axis or y-axis to generate suggested cut-off

X-axis value	
	25
Suggested Cut-off	
	24.58

Hill Estimate look for a threshold that data is stable to the right of the threshold

Sheet: DataInput

- 1. Paste data in the first column
  - Data need not be in increasing/decreasing order
  - Ensure all entries are numbers or empty
  - You may remove missing or invalid data by leaving the cells empty, the Macro will ignore them automatically
  - Number of data will be shown in column C

Sheet: Threshold

- 2. Click Analyze at the upper left corner, you will get
  - Plot of Hill Estimate
  - Plot of Mean Excess
  - Plot of LnLn
  - Plot of Mann-Kendall
  - Calculation worksheets are hidden

Hill Estimate

- Consider the order statistics X_{1,n} ≥ … ≥ X_{n,n}, Hill Estimator is defined as

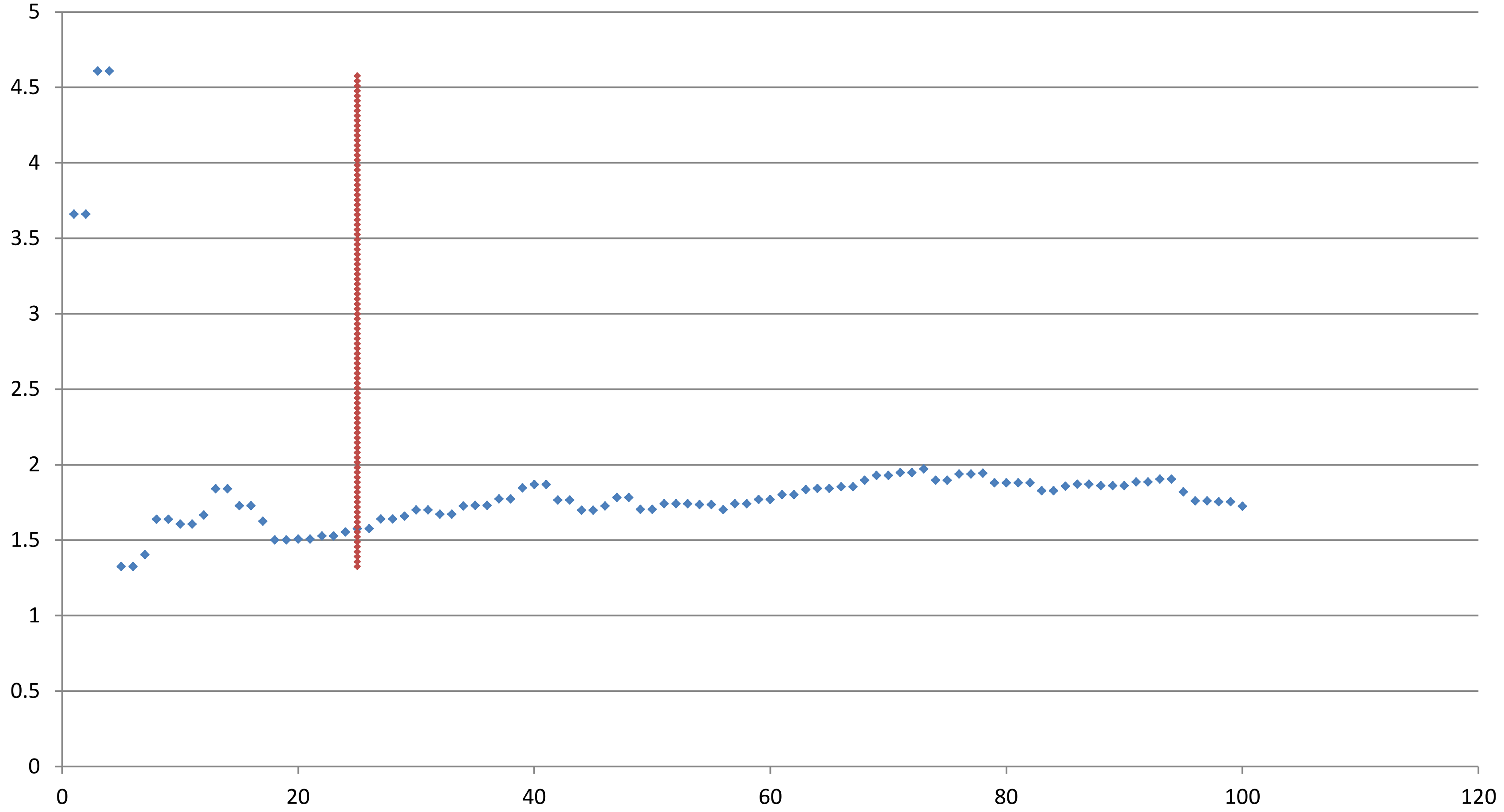
$$\hat{\alpha}_{k,n}^{H} = \left[\frac{\sum_{j=1}^{k} \left(\ln X_{j,n} - \ln X_{k,n} \right)}{k} \right]^{-1}$$

- The inverse of the average log-exceedance above the threshold.

Hill Estimate

- “Regions of the plot that are approximately close to be horizontal lines indicate values of K for which the estimate is essentially stable with respect to the choice of the cut-off.”

Hill Estimate



Mean Excess

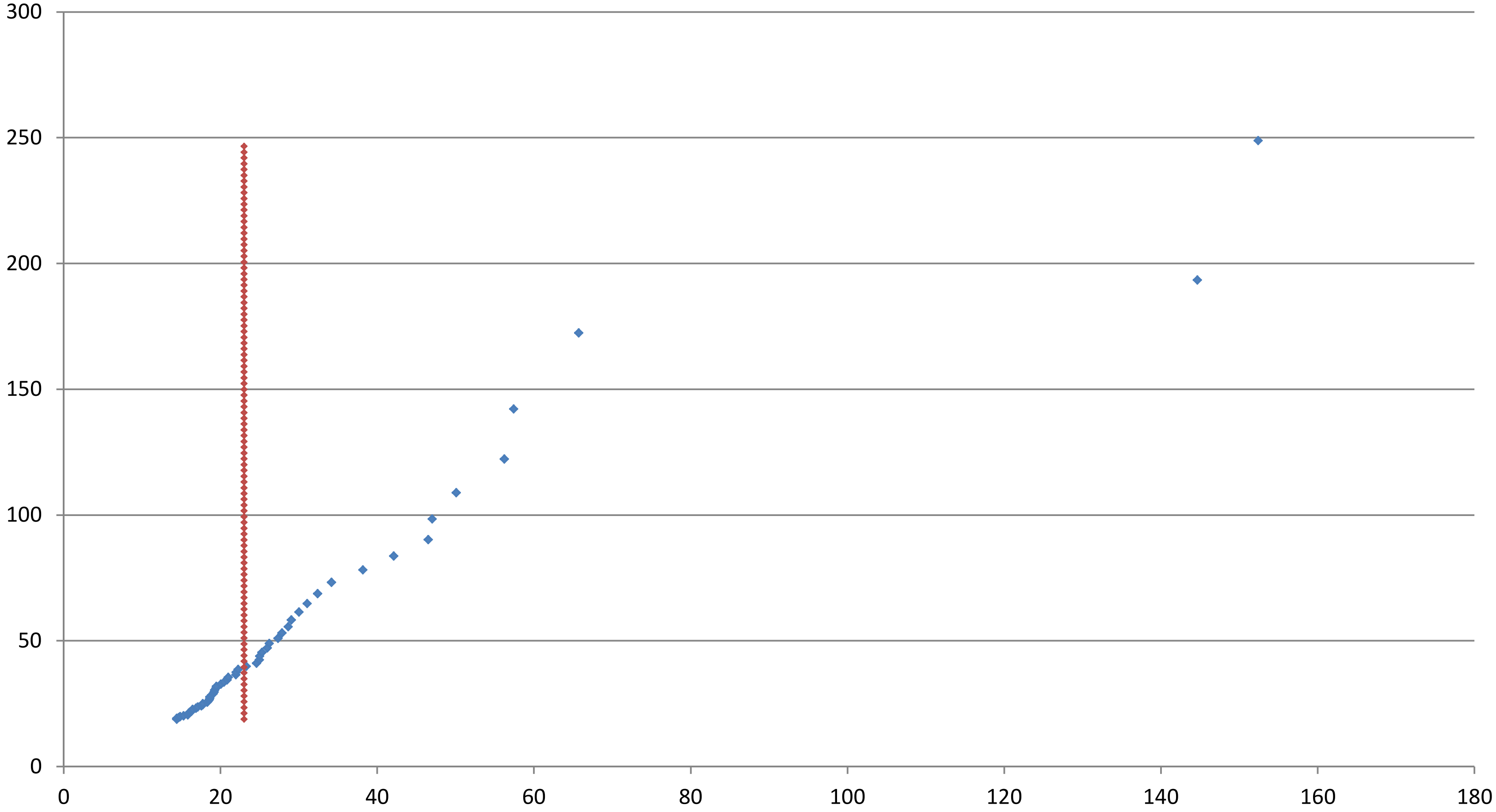
- The sample mean excess function is defined as

$$e_n\left(u\right) = \frac{\sum_{i=1}^{n}\left(X_i - u\right) \cdot I_{\{X_i > u\}}}{\sum_{i=1}^{n} I_{\{X_i > u\}}}$$

Mean Excess

- We look for a threshold that data points to the right of the threshold are roughly linear. In this method, you may ignore the last points that are deviated away from the general trend due to small sample in the highest claims.

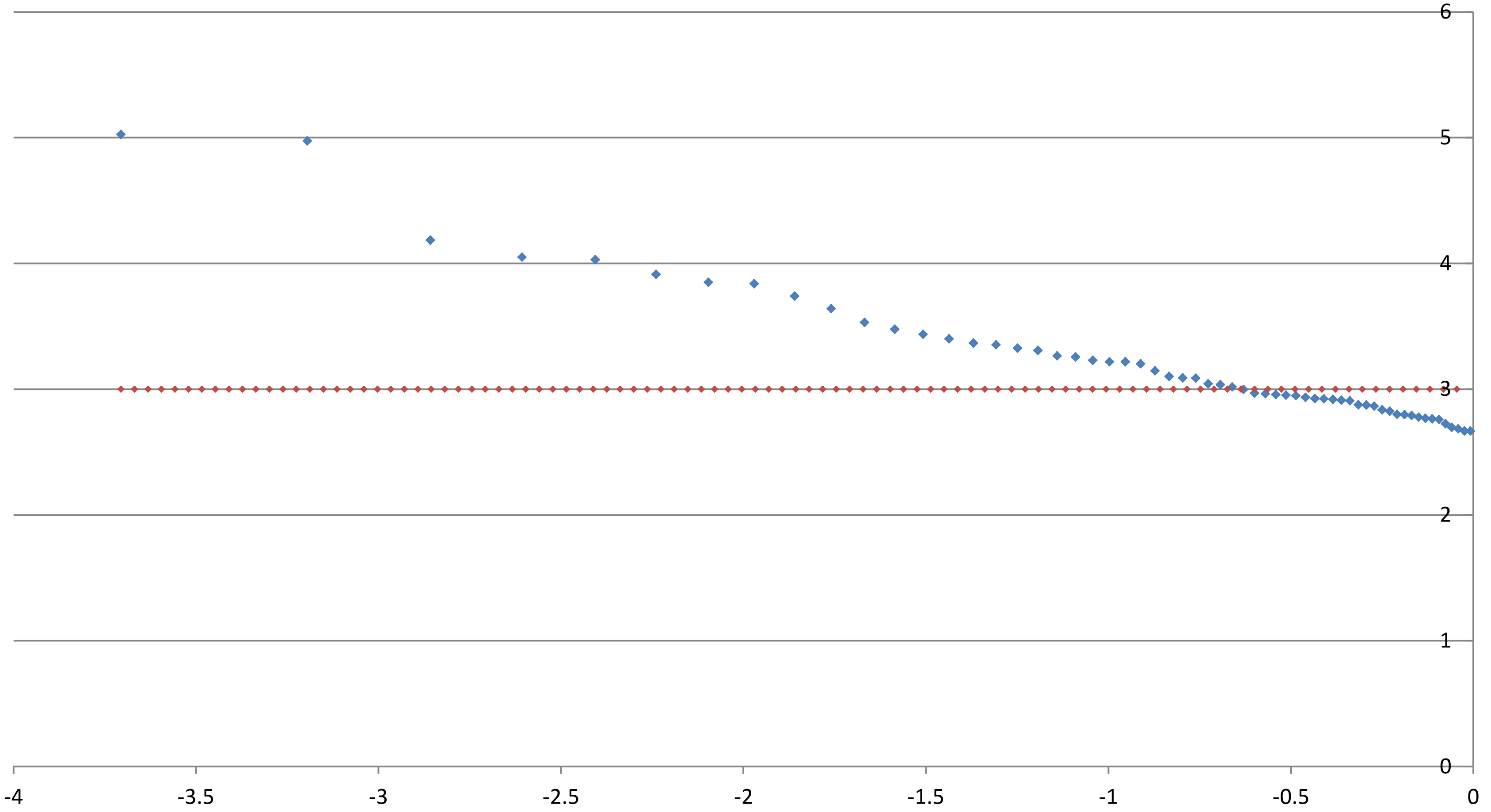
Mean Excess



Ln-Ln plot

- Ln-Ln plot Ln of the data. We look for a linear trend to the left of the threshold.

Ln-Ln plot



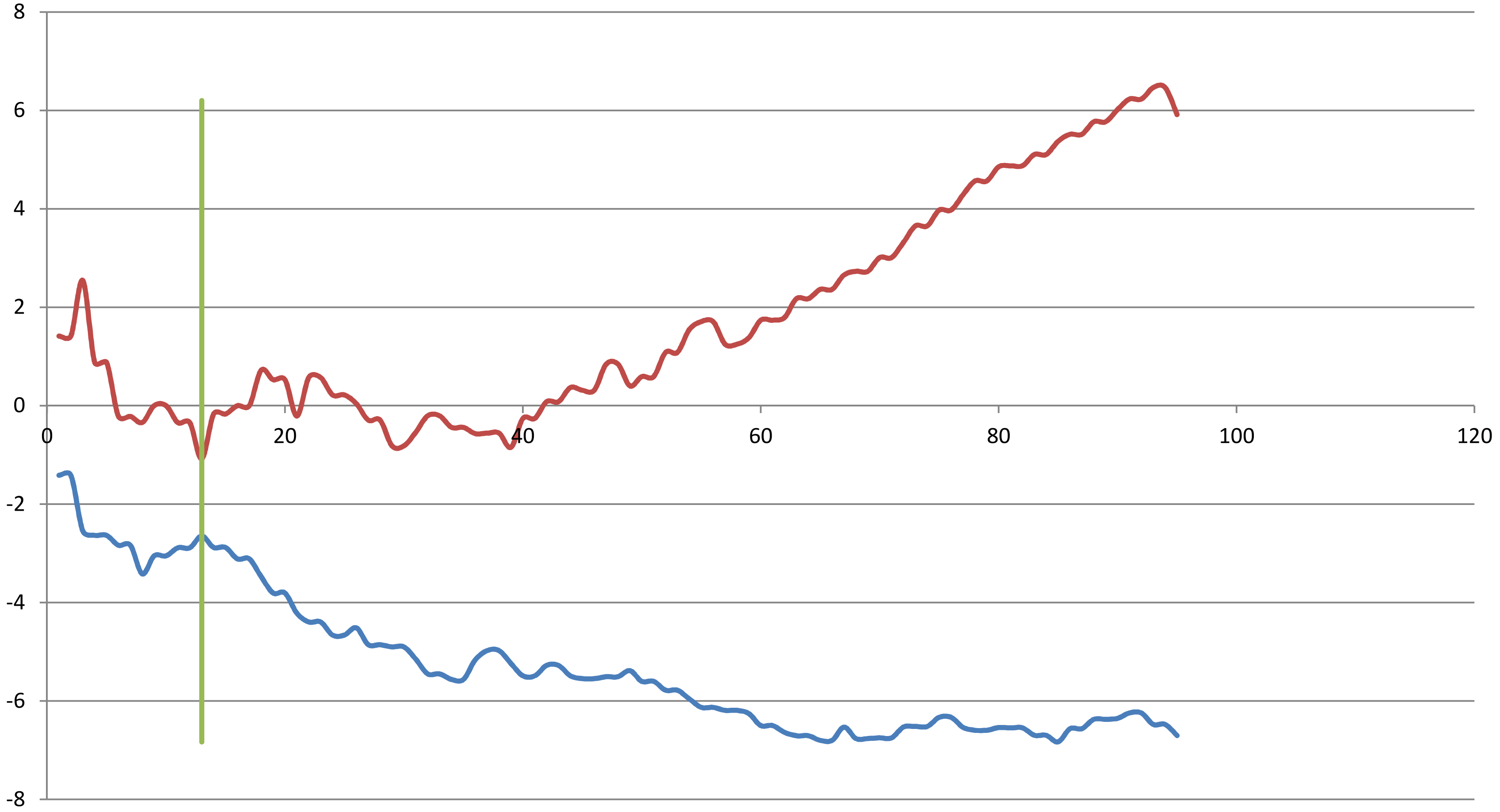
Mann-Kendall Test

- Mann-Kendall plot test the beginning of a trend. The trend begins when the to lines intersect or when they are close to each other. If there are multiple intersections, user should consider the first intersection.
- The closest point in the plot will be generated automatically, user may modify from that threshold according to own intuition

Mann-Kendall Test

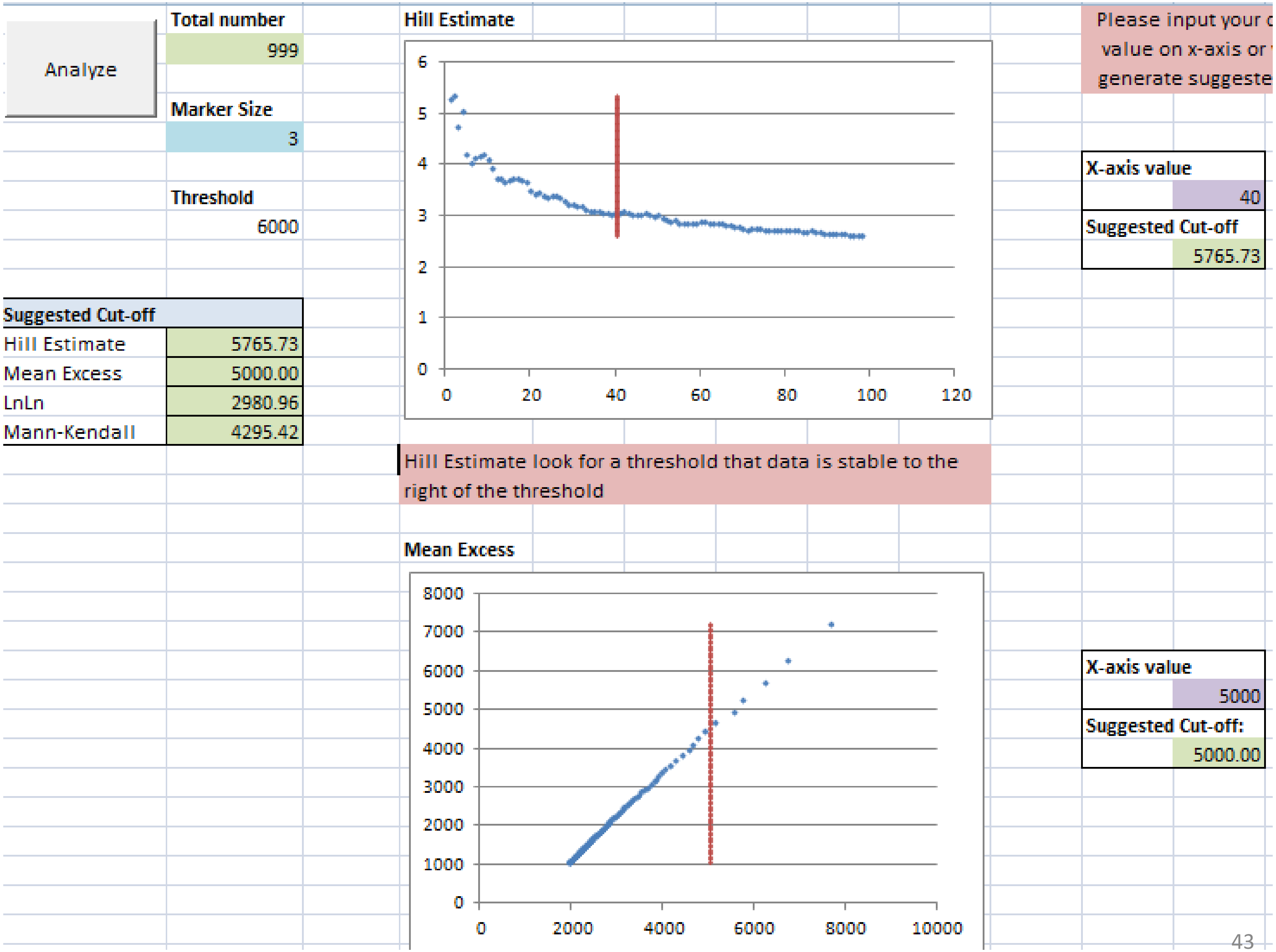
- For details of the algorithm, please refer to "*Estimation of the beginning and end of recurrent events within a climate regime*" by Friedrich-Wilhelm Gerstengarbe*, Peter C. Werner

Mann-Kendall Test



Sheet: Threshold

- 3. You need to choose the threshold based on the four plots
  - Input numbers in the purple boxes of observed x-axis/y-axis values
  - Suggested threshold will be generated and summarized in the table on the left



Total number

999

Marker Size

3

Threshold

6000

Suggested Cut-off	
Hill Estimate	5765.73
Mean Excess	5000.00
LnLn	2980.96
Mann-Kendall	4295.42

Hill Estimate

Please input your c
value on x-axis or
generate suggeste

X-axis value

40

Suggested Cut-off

5765.73

Hill Estimate look for a threshold that data is stable to the right of the threshold

Mean Excess

X-axis value

5000

Suggested Cut-off:

5000.00

Task 2

- Split into groups, each group take one set of data and present the results (what is the cut-off and why?)
  - Danish data top 1000
  - US data top 1000

Task 2

- Breakout groups to report back
  - Selected cutoff
  - Why?

Section

Bootstrapping and XOL Pricing

What is Bootstrapping?

- Bootstrapping is a method to measure uncertainty of sample estimates.
- Assume your distribution choice and fitted parameters are correct, then
  - The history is only one possible random outcome
  - Other possible outcomes would lead to different parameter estimates
  - We can estimate the parameter uncertainty by considering the varying parameter estimates from these other possible random outcomes

What is Bootstrapping?

- We construct a random resample (with replacement) of the claim data with same size
- For each resampled set, we calibrate the MLE estimators

Why Bootstrap?

- The more uncertain the distribution parameters, the more likely are large claims
- So greater uncertainty means
  - Higher XOL prices for the higher layers
  - Higher PMLs

How to use

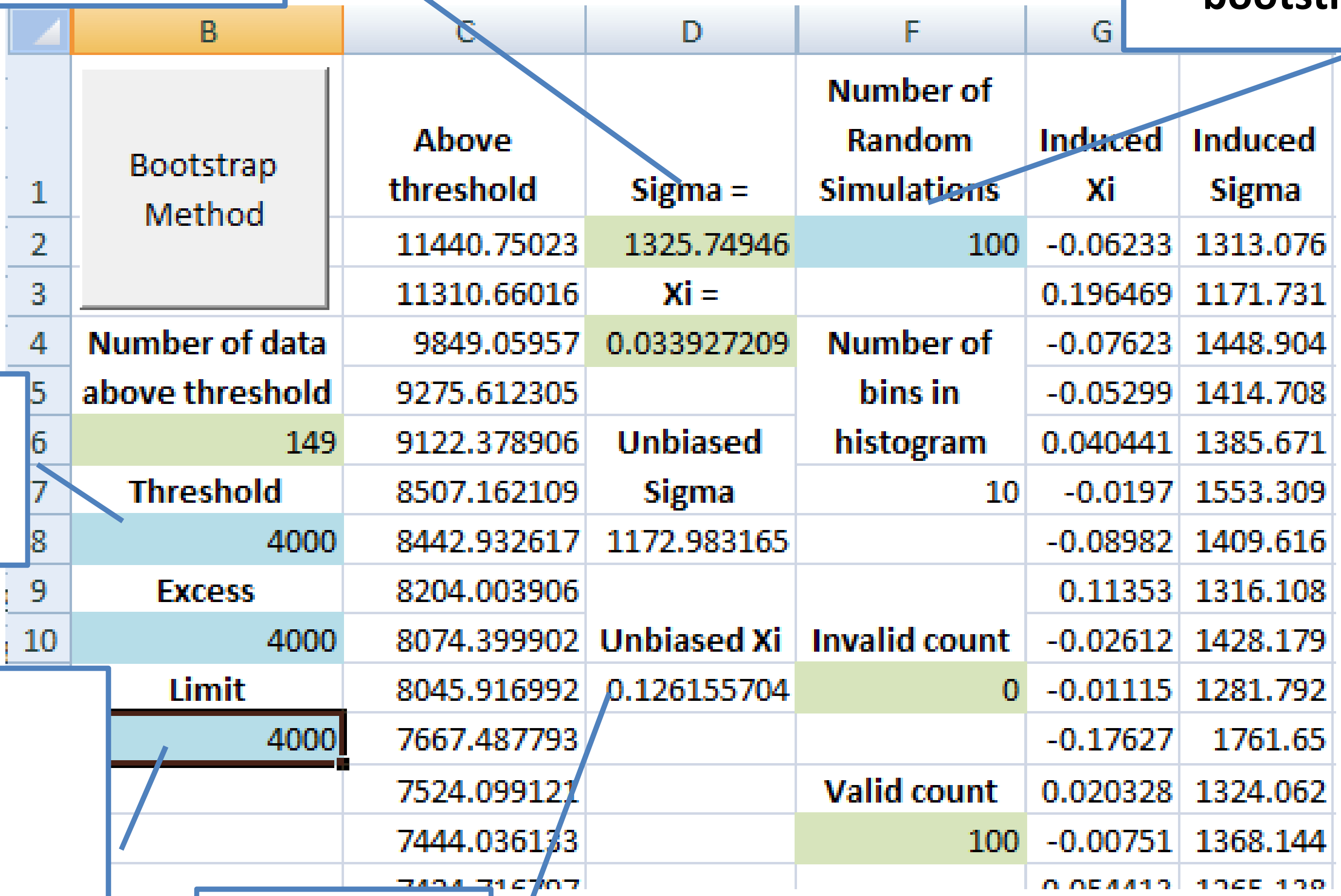
MLE estimator

Number of simulations for bootstrapping

Input cutoff

Price XOL premium between (excess, excess + limit)

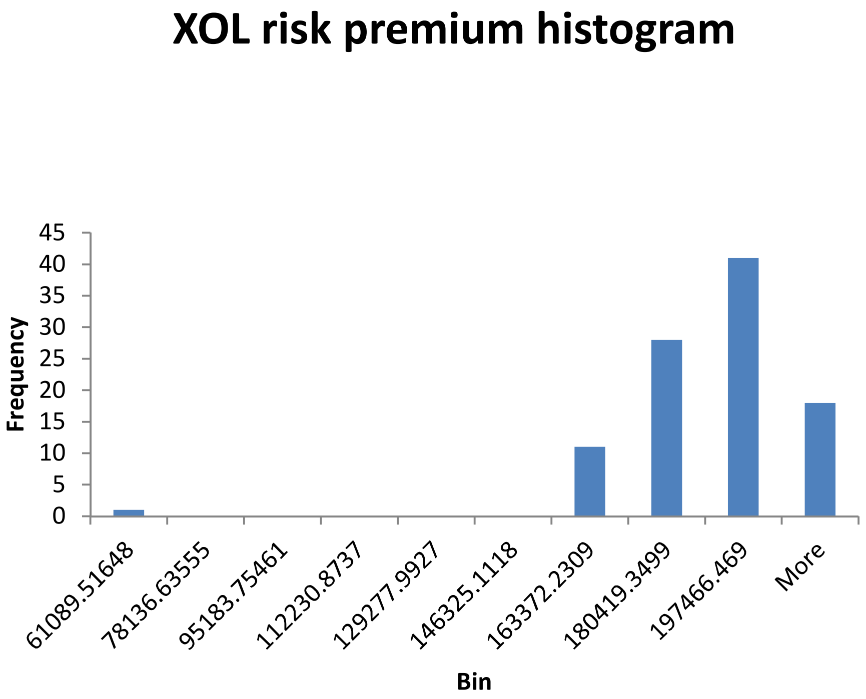
Unbiased estimator

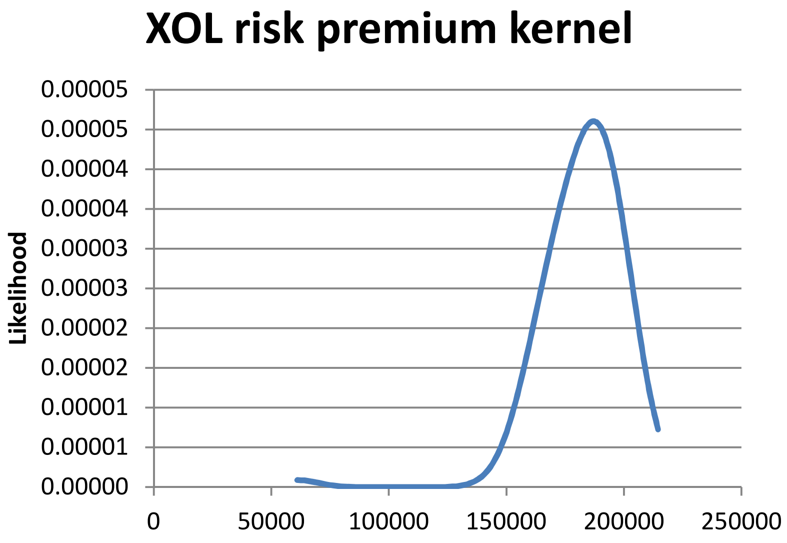


	B	C	D	F	G	
1	Bootstrap Method	Above threshold	Sigma =	Number of Random Simulations	Induced Xi	Induced Sigma
2		11440.75023	1325.74946	100	-0.06233	1313.076
3		11310.66016	Xi =		0.196469	1171.731
4	Number of data above threshold	9849.05957	0.033927209	Number of bins in histogram	-0.07623	1448.904
5		9275.612305			-0.05299	1414.708
6	149	9122.378906	Unbiased Sigma		0.040441	1385.671
7	Threshold	8507.162109		10	-0.0197	1553.309
8	4000	8442.932617	1172.983165		-0.08982	1409.616
9	Excess	8204.003906			0.11353	1316.108
10	4000	8074.399902	Unbiased Xi	Invalid count	-0.02612	1428.179
	Limit	8045.916992	0.126155704	0	-0.01115	1281.792
	4000	7667.487793			-0.17627	1761.65
		7524.099121		Valid count	0.020328	1324.062
		7444.036133		100	-0.00751	1368.144
		7424.716797			0.054413	1265.128

Result

- A column of "viable risk premium"





Task 3

- Split into groups, each group take your data and cut-off and present your XOL premium
  - Low layer XOL premium
    - Danish Data: 3 excess of 2
    - US Top 1000: 500 excess of 3000
  - High layer XOL premium
    - Danish Data: 150 excess of 300
    - US Top 1000: 2500 excess of 12500

Task 3

- Breakout groups to report back
  - XOL risk premium
  - Justification

Bootstrapping

Probable Maximum Loss (PML)

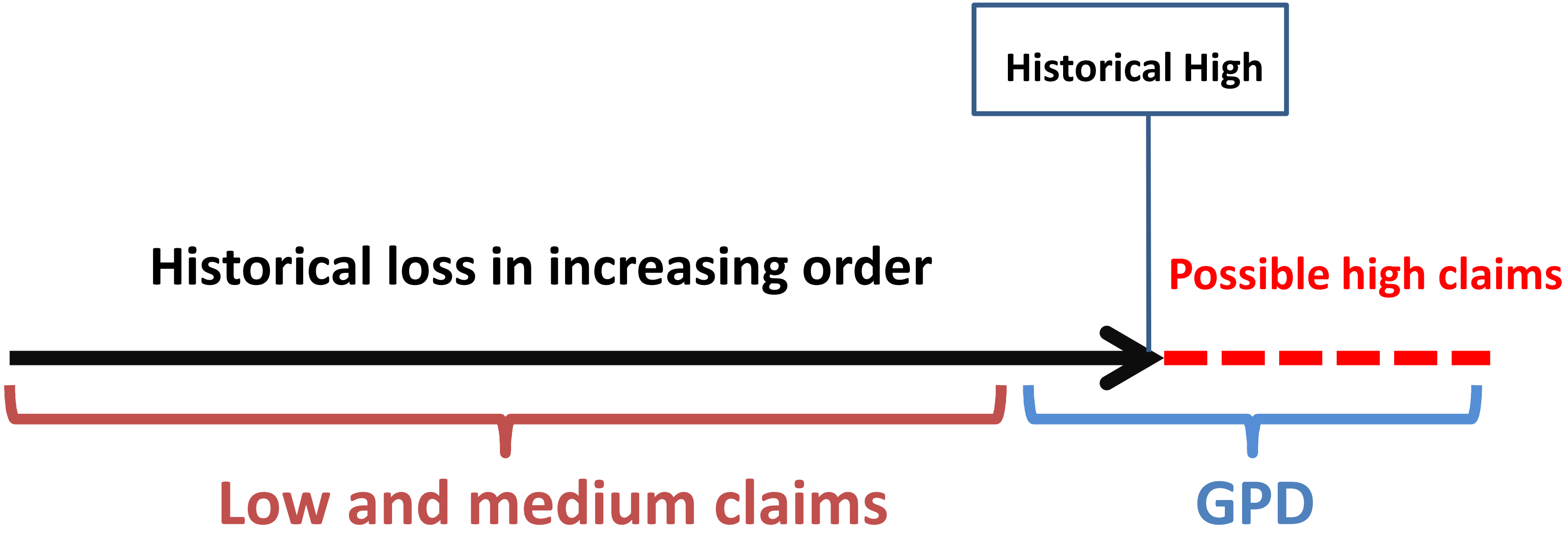
Definition: Maximum loss expected in a given year (given a probability or return period)

Useful functions

- GPDquantile(p,cut-off,sigma,xi)
  - Purpose: Numerical solve for the inverse cdf.
- GPDgencdf(x,cut-off,sigma,xi)
  - Purpose: return the cdf for generalised pareto
- GPDmean(cut-off,sigma,xi)
  - Purpose: returns the mean of a generalised pareto, given its parameters

Estimating PMLs

- Determine Which Region of the Distribution



Estimating PMLs

- For region below cutoff, use retrospective estimate
- For region above cutoff, use distribution

- Remember to adjust for how many years of historical data you have used!

Task 4

- What are the PMLs for next year? (state assumptions)
  - 1:10 year return period
  - 1:100 year return period
  - 1:200 year return period
- Do for both Danish Data and US Top 1000

Task 4

- Breakout groups to report back
  - PML Estimate
  - Justification / assumptions

Section

Wrap Up

What We’ve Done Today

- Learned How to Model the Tail of a Distribution
- Learned How to Use Our Model to
  - Price XOL
  - Estimate PML

Questions?