Estimating Missing Data

Software: SuperSMITH 6.0AA 31 Dec 2021

Tip: Use <ctrl> L
To view page by page in full screen mode. Use <esc> to cancel.

Note: For more detail, see the RAMS paper:
Estimating Quantity of Missing Data in Tails of Distribution
RAMS 2021, The 67th Annual Reliability & Maintainability Symposium, May 24-27, 2021, Orlando, FL
Carl Tarum, Bathtub Software LLC, Wes Fulton, Fulton Findings
https://ieeexplore.ieee.org/document/9605766

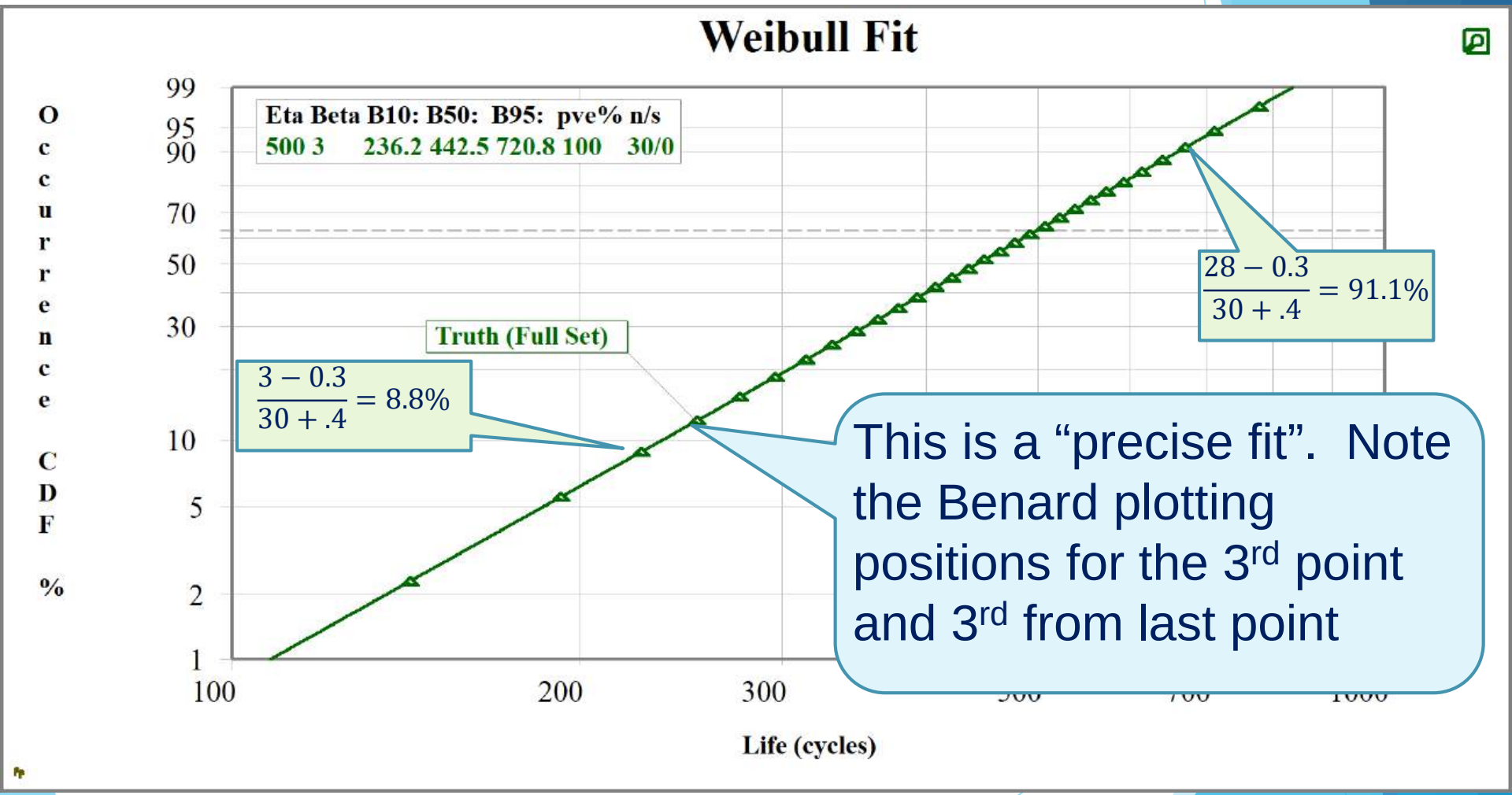
Causes of missing data

- Many distributions are modeled with Weibull, Lognormal, or Normal Distributions.
- Sometimes fits are poor because data are missing
  - Quantity of scrapped parts may not be recorded
  - Sections of data may be missing due to records issues
  - Decision to start recording data may be delayed, and early failures were not recorded.
  - Supplier may remove parts close to the target to ship elsewhere at a premium price

Methods to Analyze Missing Early Data.

- Crow-AMSAA can be used if there is enough new data. The instantaneous failure rate will become correct
- Historically, a 3-parameter Weibull could be used
- With ”snapshot” data, Kaplan-Meier techniques can also be used.
- A new technique, presented in 2021, uses the traits of the distribution to estimate missing data. This can be used on Early, Late, or Intermediate missing data if there are enough points

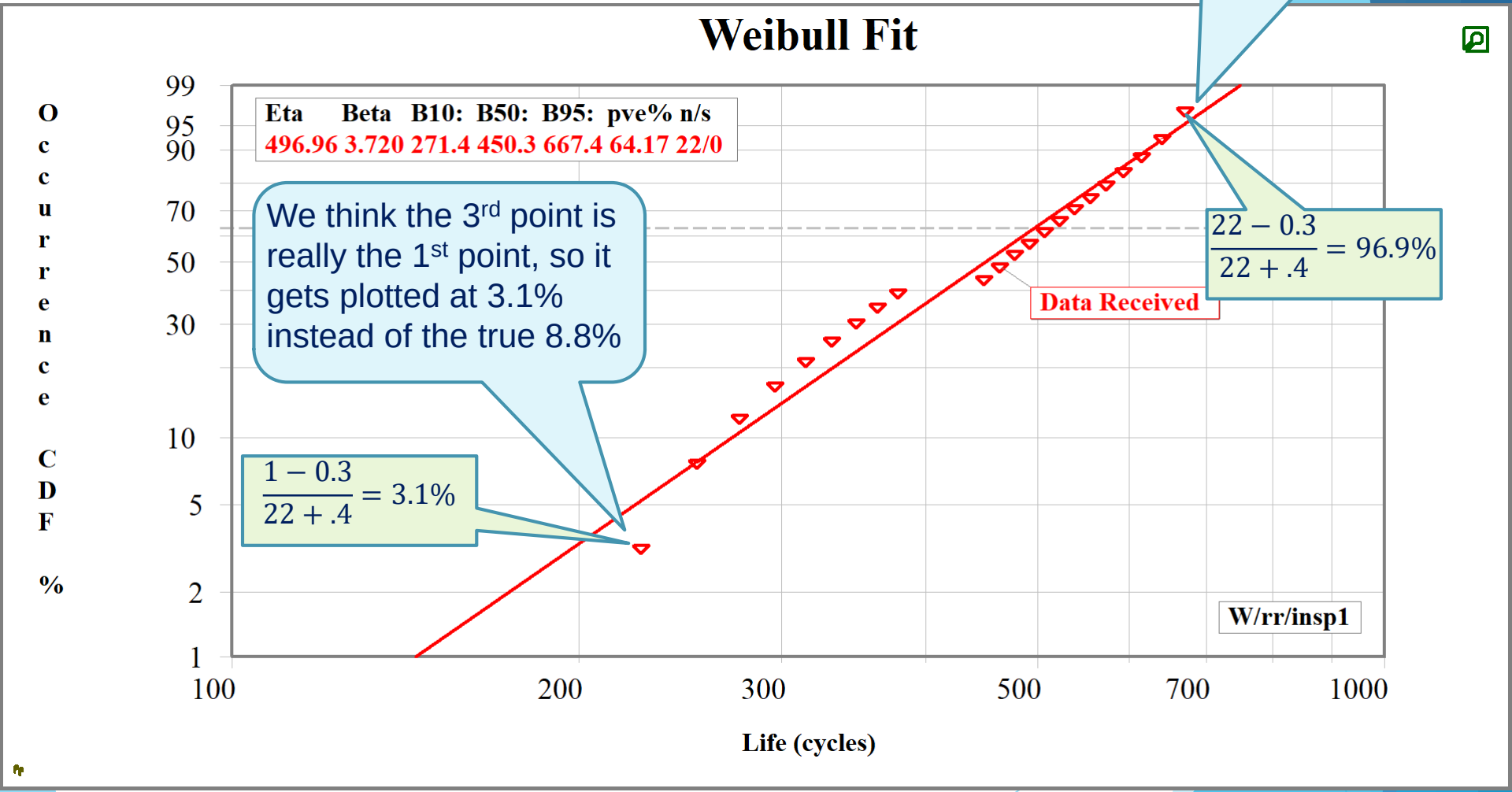
Point Pattern Impact:
Weibull Scaling



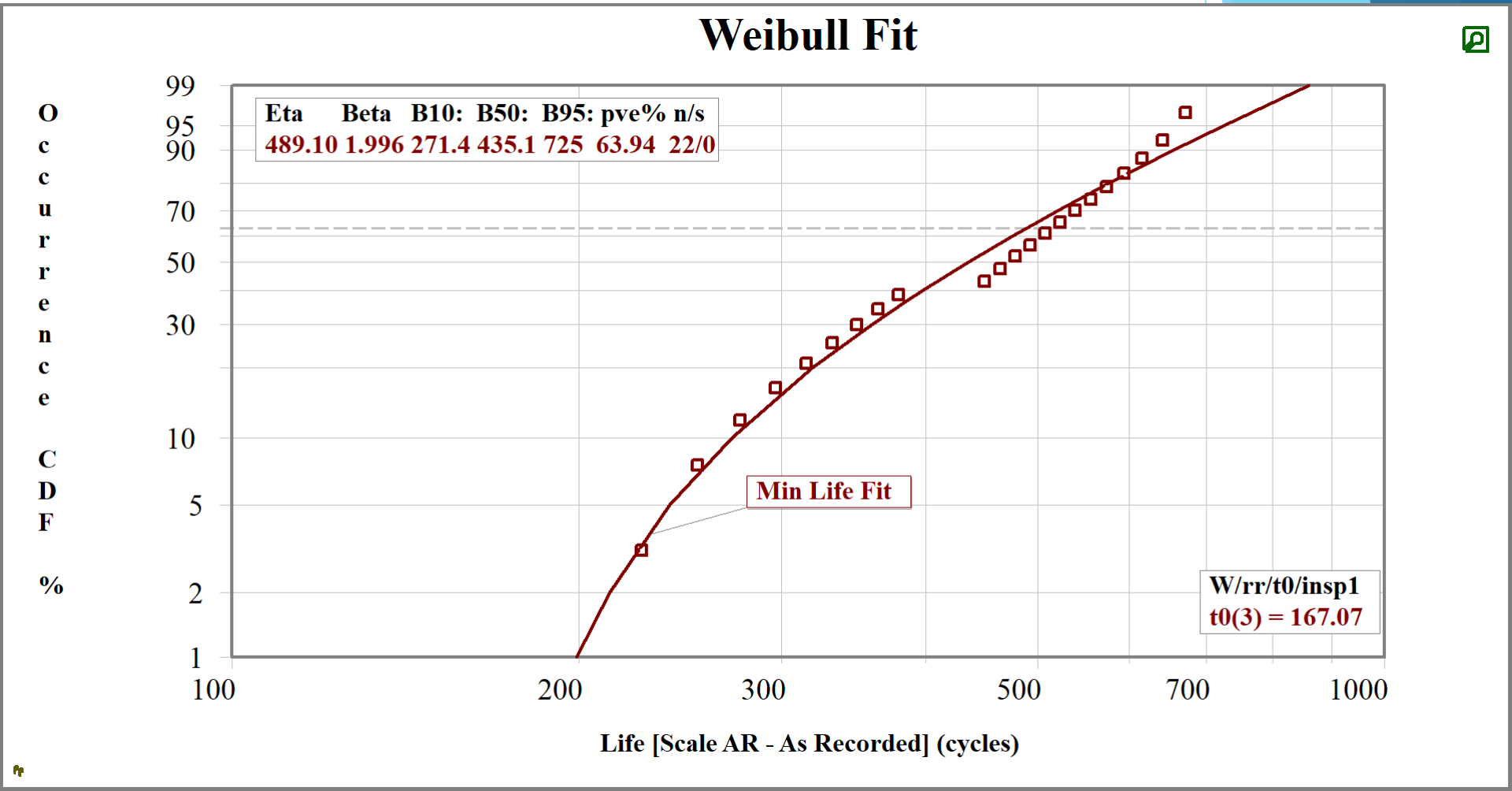
Effect of missing points

- Suppose that the previous slide with 30 points was “Truth”
- But unknown to us, the first two points, 4 points in the middle, and 2 points at the end are missing.
- What would the fit look like?

Data as received



t_0 actually worse pve% than a straight line

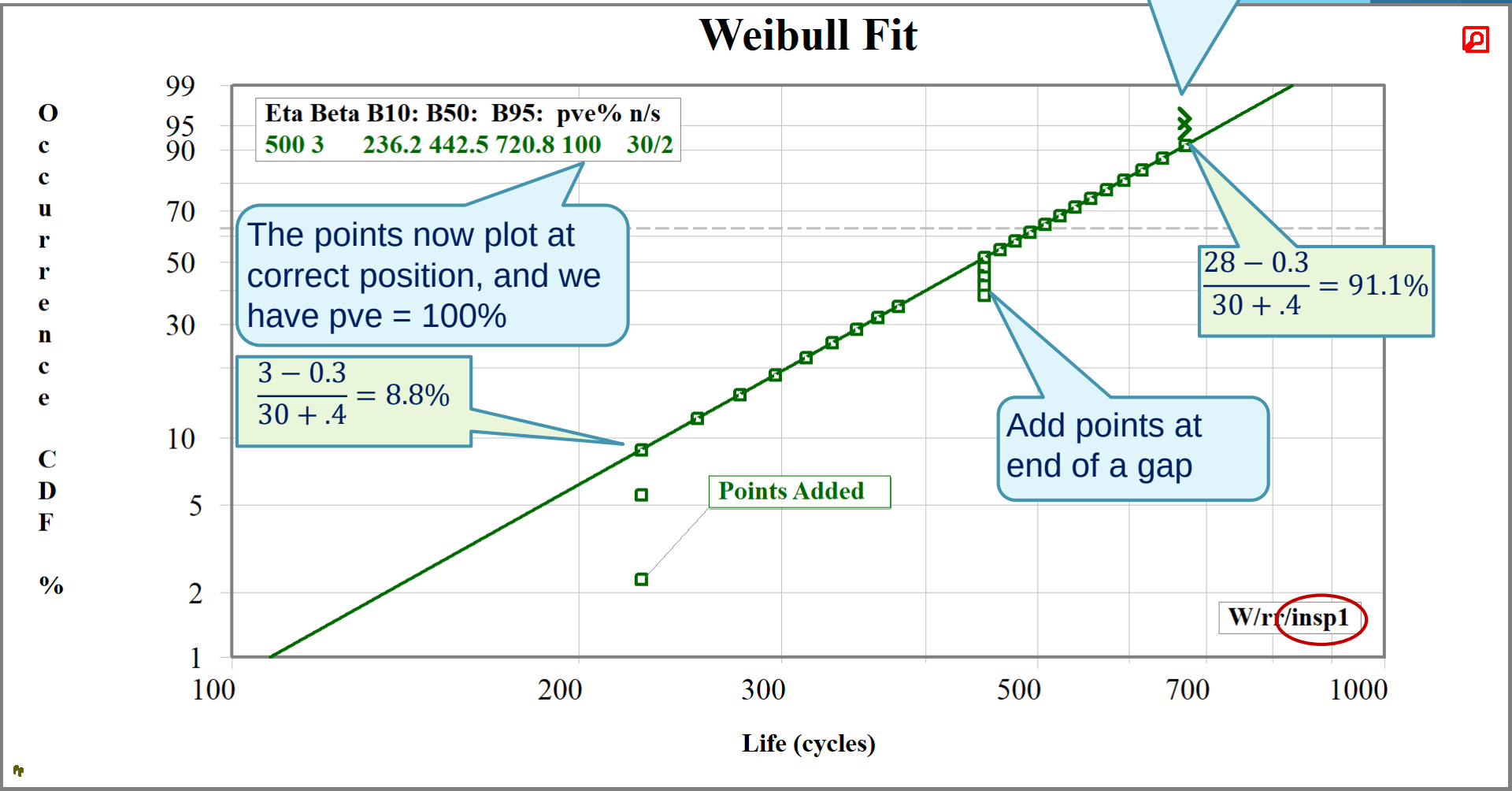


What if we knew how many points were missing?

- Suppose we knew exactly how many points were missing, but not sure of the x values?
- We can add the points and use Inspection Option 1

Add Missing points Restores original fit (Inspection Option 1)

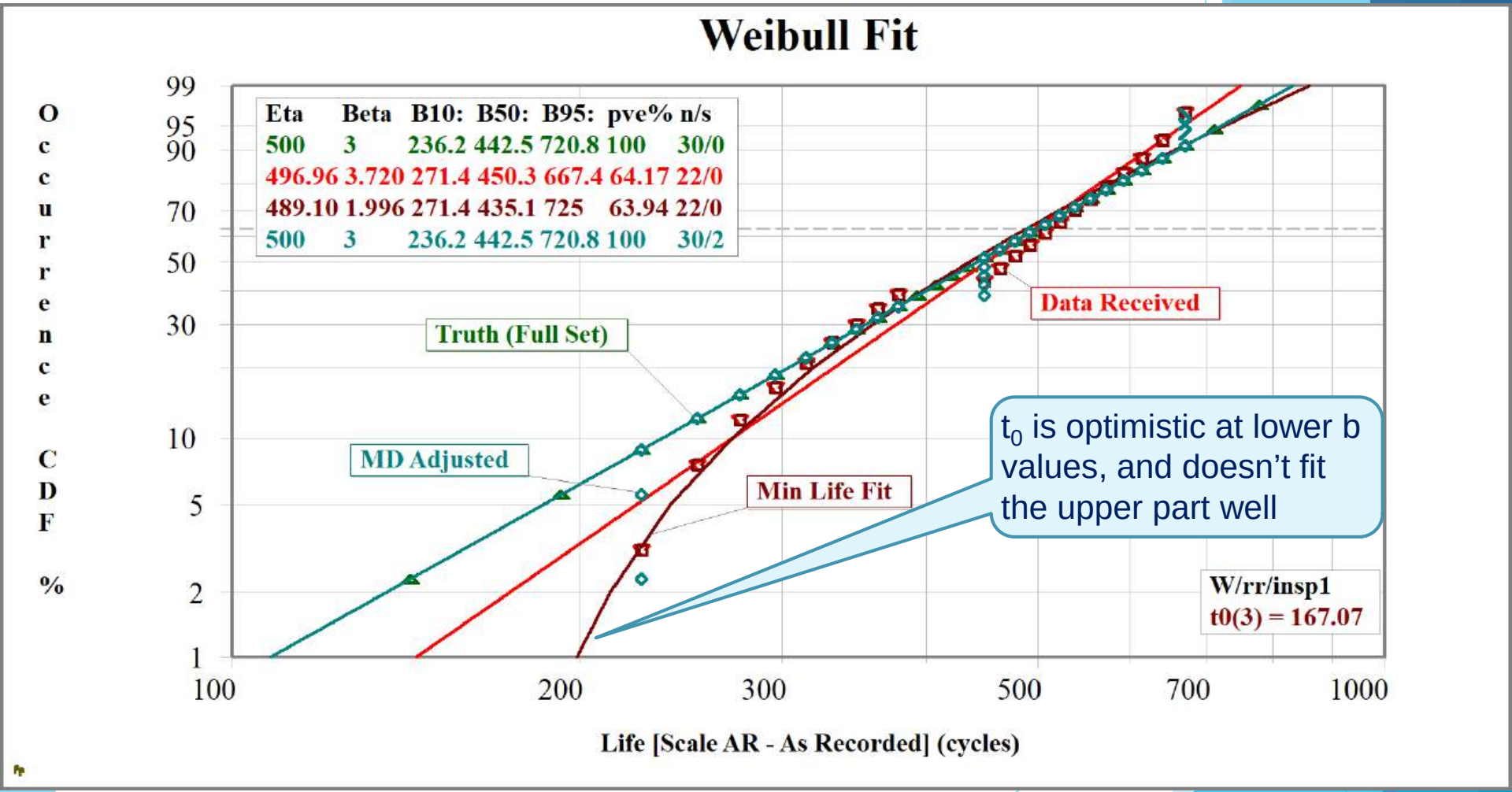
On Right tail, we add suspensions, since inspection plots the top of the stack.



Comparison

- The following slide shows the 4 fits
- Note that the final fit has the same B values and fits
- The set with missing data is optimistic

Set Comparison



How can we decide how many points to add?

- If we knew how many points to add, we could get back to near the original fit
- This would give better estimates of life
- Next slides cover Criteria, Benefits, and Methods
- Tried several methods to estimate points to add

Criteria and Benefits

Criteria

- ✓ Appropriate model (Weibull, Lognormal, Normal)
- ✓ Explanation of missing data
- ✓ At least 100 points, no more than 20% missing
- ✓ Adjusted fit is better
- ✓ Distribution Analysis confirms best model selection.

Benefits

- Alternative to Minimum Life Model
- Parameter prediction closer to true value
- Fit improvement may lead to better model
- Improvement in life prediction

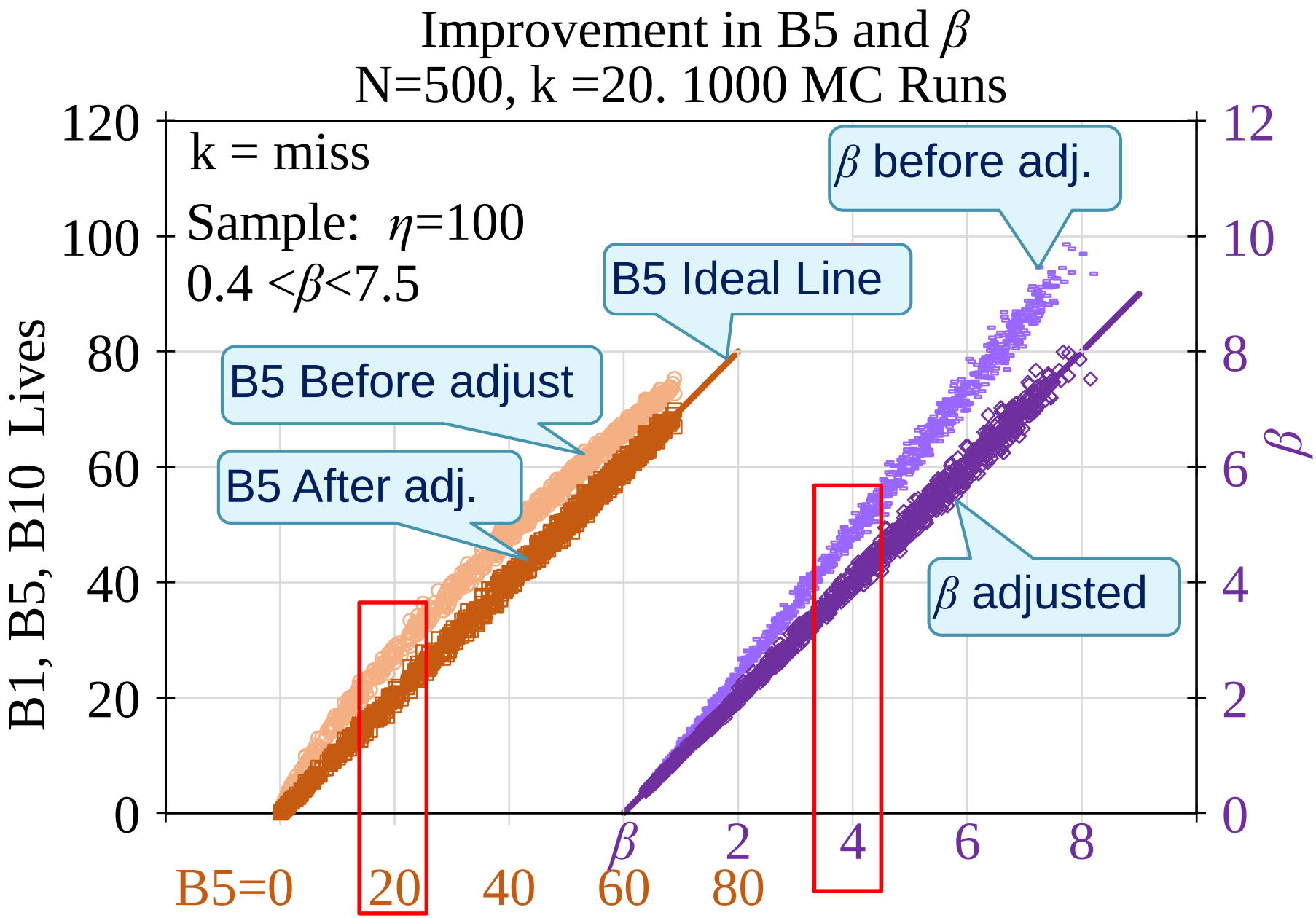
Monte Carlo to check Exit Criteria

- MC sample and fits calculated.
- Points removed, new fit calculated
- Points added per criteria choice
  1. Best r^2
  2. Quadratic=0.
  3. First point above line.
  4. Any of first 3 points above line
  5. Two of first 3 points above line with full data set
  6. **First point above fit line, or points 2 and 3 above line (Best choice for rank regression)**
  7. Minimum of 4 methods:
  8. Average of 1st point above, Quadratic
  - For MLE, increase Discovery points until Minimum Life = 0

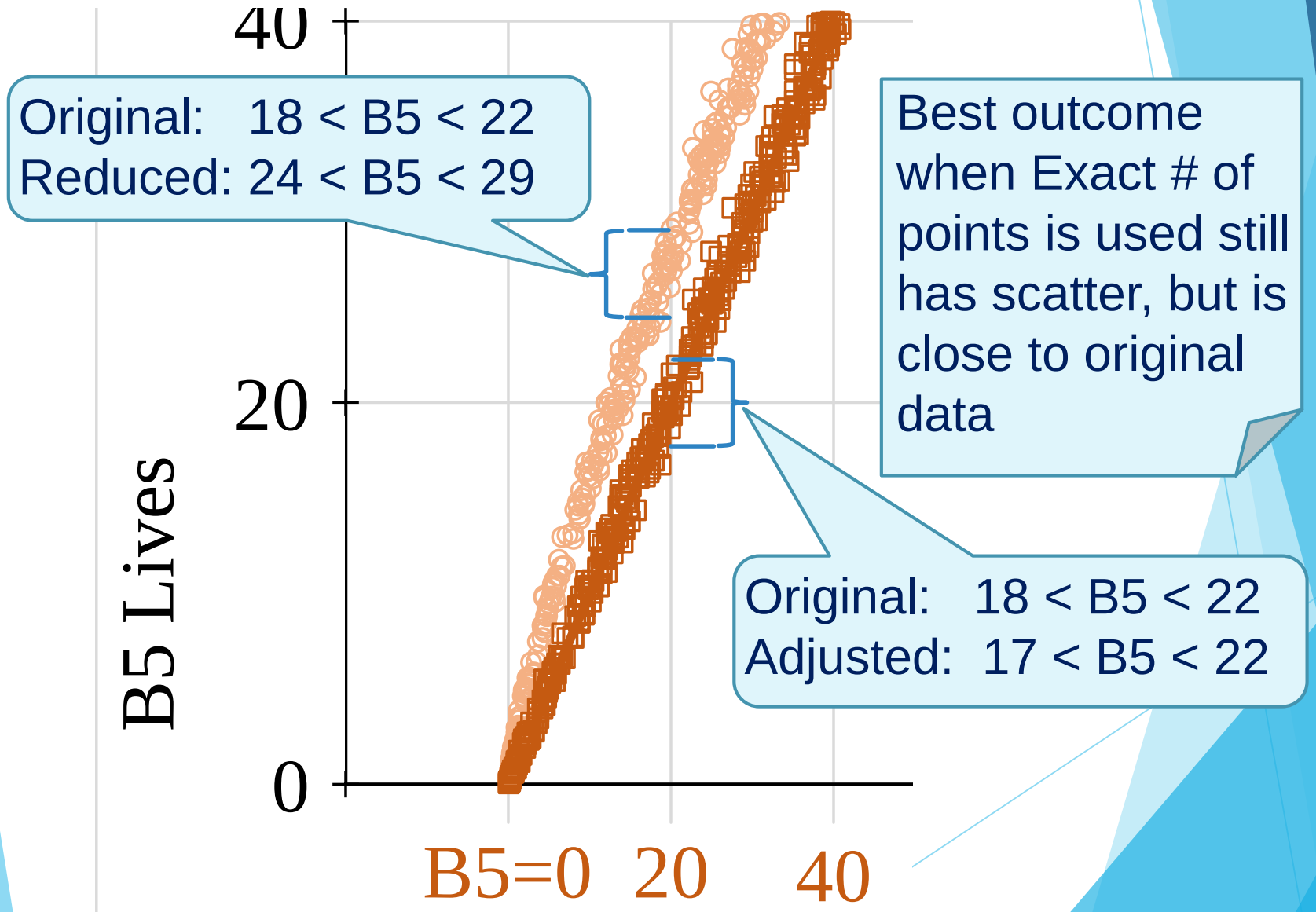
Results Comparison

- Ideally, the adjusted fit would equal the original fit
- There is variation, because the variation of the points removed is lost and unknown.
- Following charts show the median values of the adjusted fits correspond well to the original fits, using MRR (Median Rank Regression)
- Monte Carlo sample size of 500.
  - Remove 20 points on left
  - Add exactly 20 points at first point
  - Switch to Inspect Option 1
  - Gives estimate of resulting variation
  - Look at variation on B5 around 20%, and Beta around 4
  - Repeated for 1000 sets

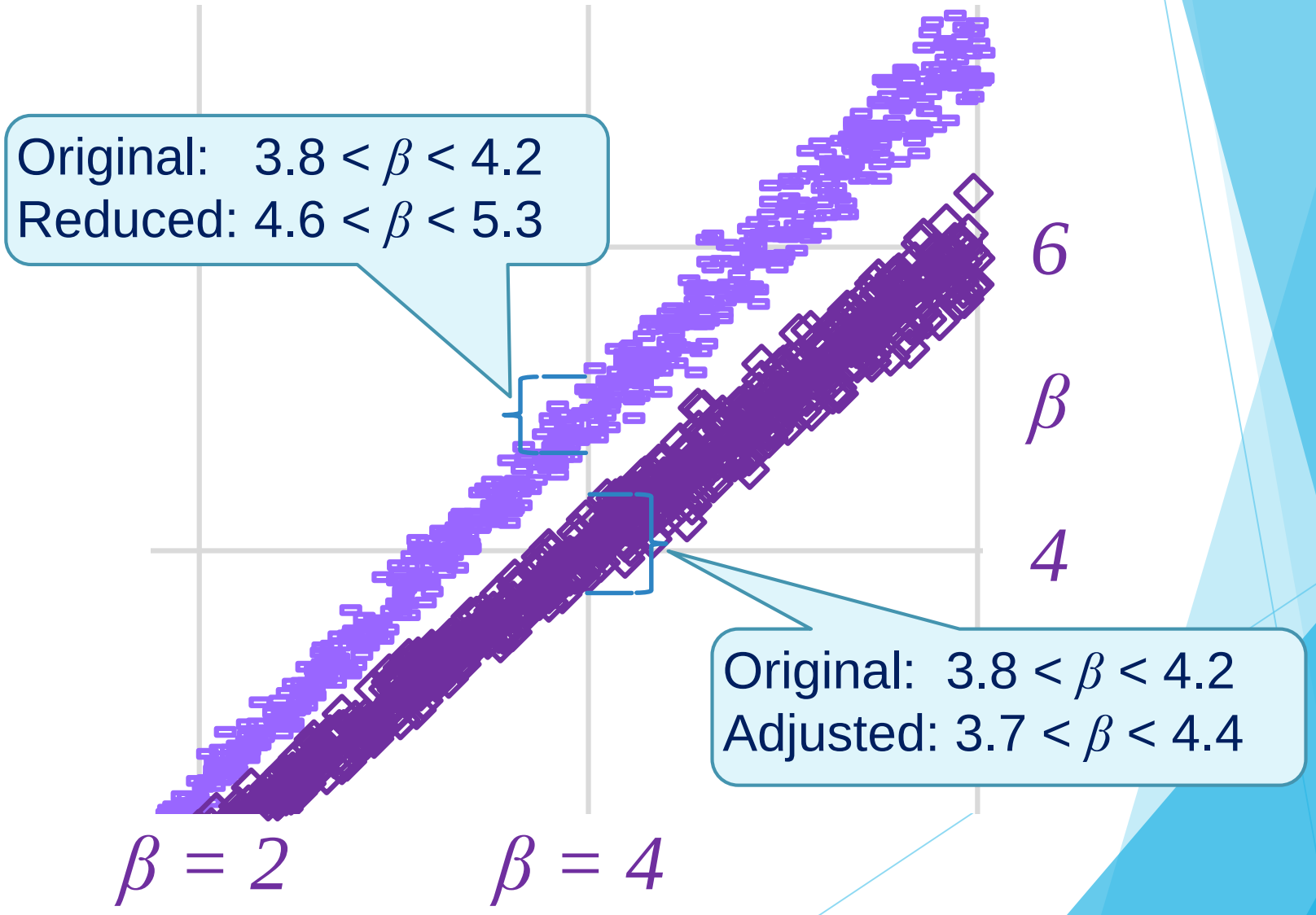
Remove and add 20 points



Remove and add 20 points
Impact on B5



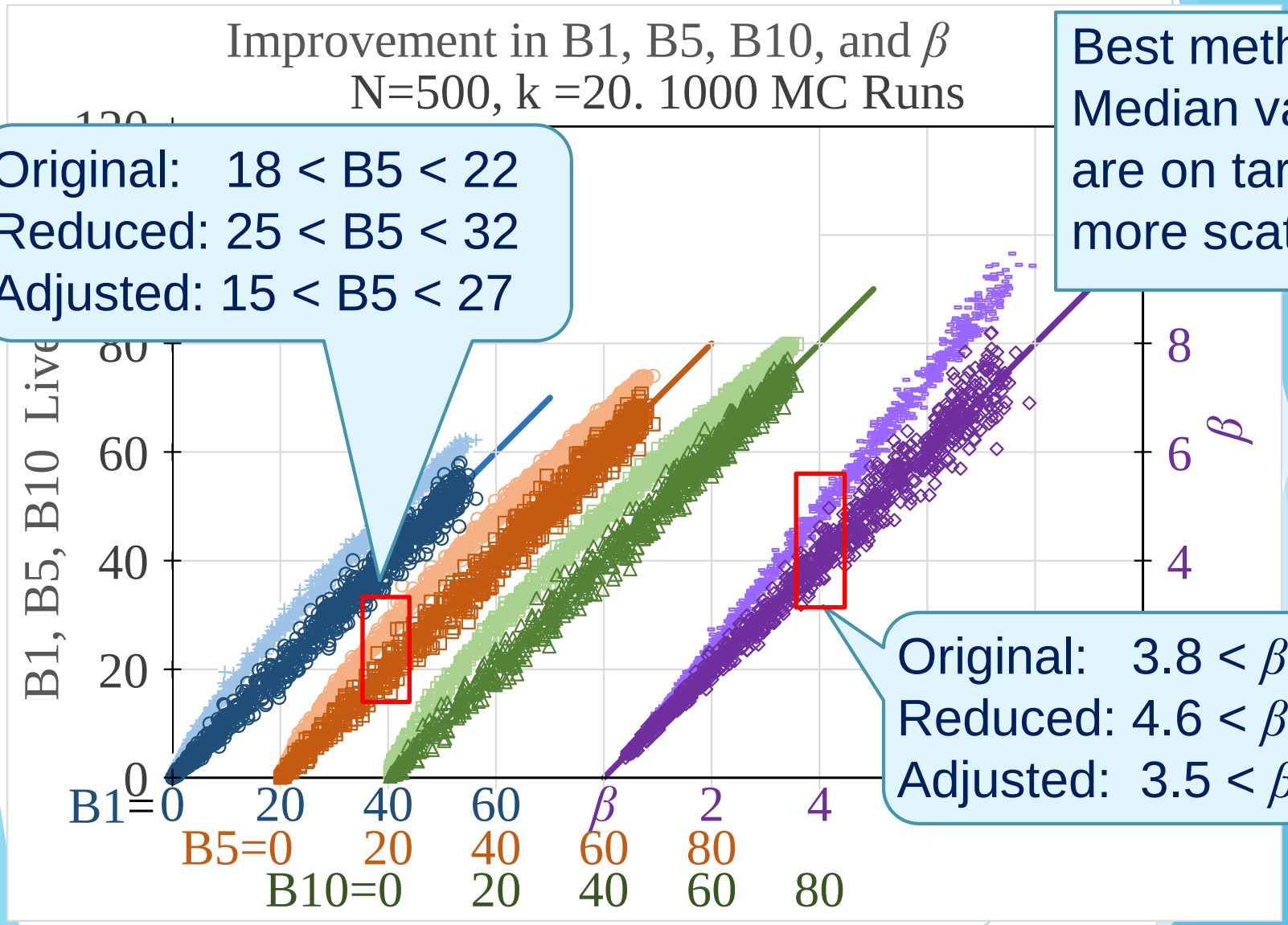
Remove and add 20 points Impact on β



Best Results

- Previous slides show the impact if we knew EXACTLY how many points were missing. The adjusted points don't go back to the straight line, because we have lost the variation in the first 20 points.
- The following slide shows the results of using the best method for left tail missing data:
  - Adding points until the first point goes above the fit line, OR
  - Points 2 and 3 go above the fit line.
  - Since half of the time the first point will be below the fit line, we then adjust the count down by 1 and see if there is a better fit.
  - Similar logic for intermediate and right tail missing data
- Note the variation in B5 and beta when points are missing (Reduced set), and how they are closer to the original values when adjusted

Weibull Distribution Results



Original: 18 < B5 < 22
Reduced: 25 < B5 < 32
Adjusted: 15 < B5 < 27

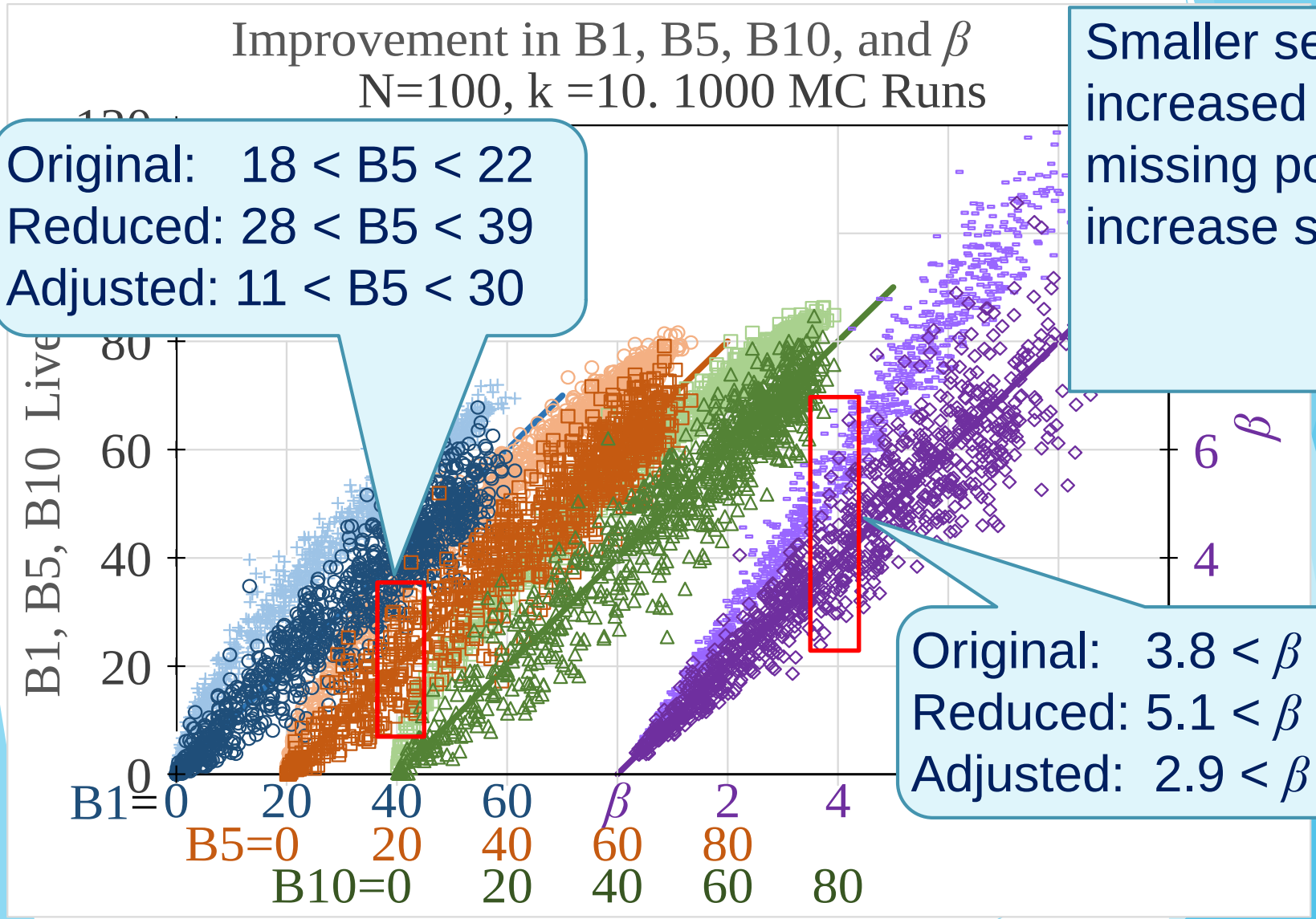
Best method. Median values are on target, but more scatter

Original: 3.8 < β < 4.2
Reduced: 4.6 < β < 5.5
Adjusted: 3.5 < β < 5.0

Large Sample sizes are necessary

- Previous studies were with a sample size of 500.
- Can we use this technique with smaller samples?
- Yes, but there is much more variation. Look at the variation of the sets when we only had 10 missing points on a set of 100.
  - The reduced sets are very optimistic.
    - A B5 of 20 on the original set might be around 30 on the reduced set.
    - Beta of 4 in the original set could be around 6.5 with the reduced set
  - Adding points brought the medians back down to the original values, but there was more variation

More Scatter with N=100



Original: 18 < B5 < 22
Reduced: 28 < B5 < 39
Adjusted: 11 < B5 < 30

Smaller sets and increased % of missing points increase scatter

Original: 3.8 < β < 4.2
Reduced: 5.1 < β < 8.2
Adjusted: 2.9 < β < 6.5

Method Summary

- Method can reasonably estimate missing data (left-tail, intermediate, right-tail) for Weibull, Normal, and Lognormal data
- Suitable data set (>100 observations, reason for missing data, no large quantity of right suspensions)
- Use Rank Regression. Turn off t_0 minimum life, turn on Inspection option
- Add failures(suspensions) until exit criteria are reached: 1st point above, or next 2 points above.
- Review results. (Better fit, <20% added points)

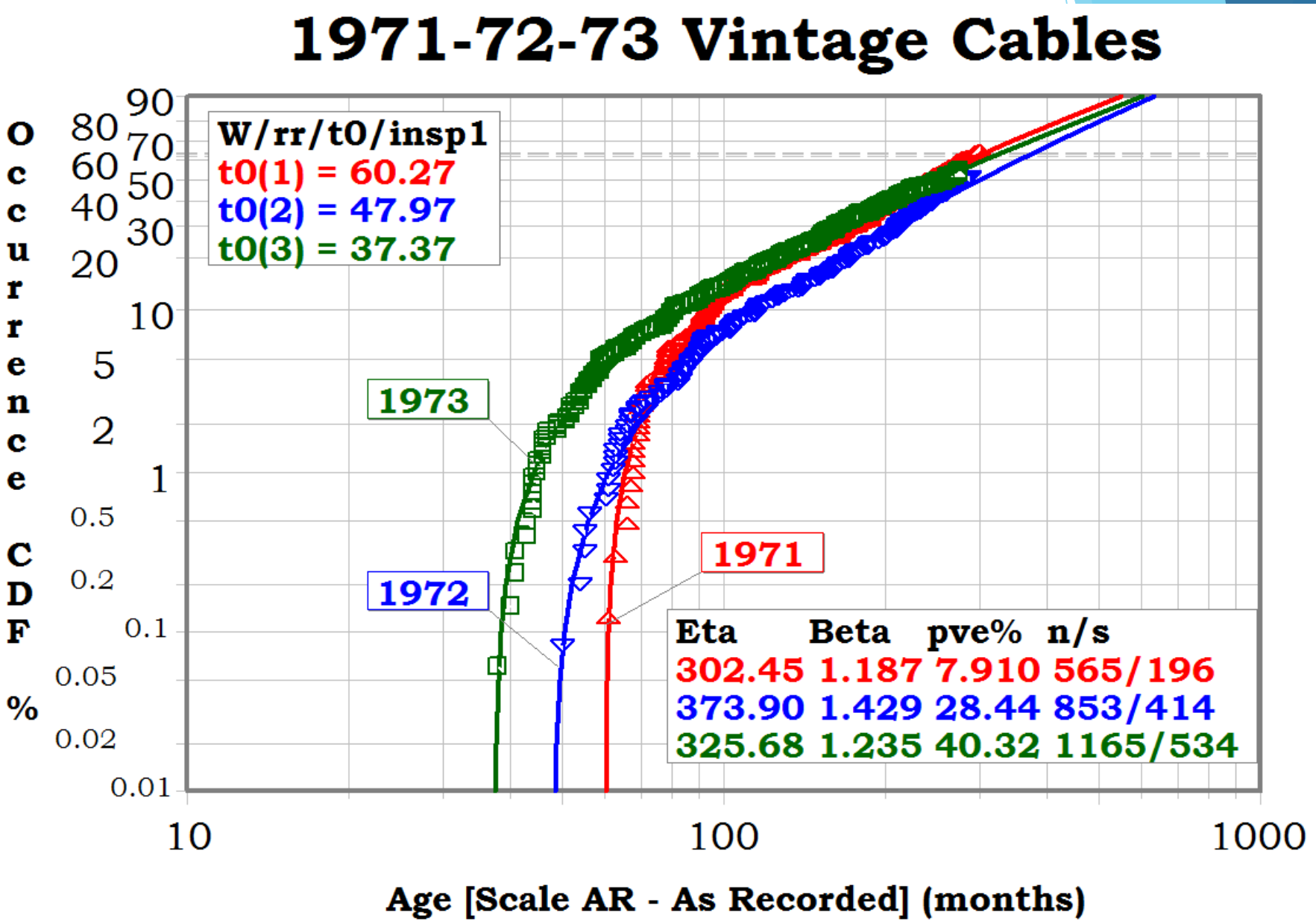
MLE: Maximum Likelihood Estimation

- MLE can be used to estimate left tail missing points
- Calculate MLE with a minimum life, t_0. If t_0 is >0, then there is a curve. Add points as discoveries at the first observation until $t_0 = 0$, indicating no curvature. Then turn t_0 off.
  - Discoveries and t_0 are inherently incompatible because they are two different techniques to address a similar issue. But they are used together in this method to determine points to add.
- MRR (Median Rank Regression) can be used to calculate missing parts, and then an MLE solution can be applied. MRR and MLE will be similar fits.

Example – Fig 3-11 in Handbook

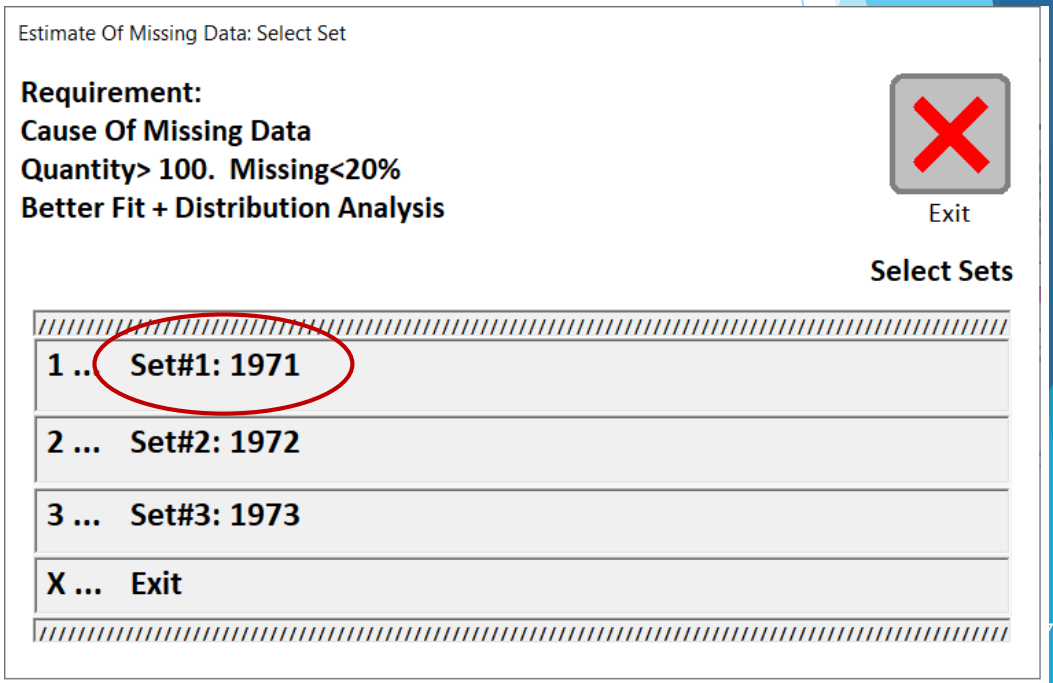
- Buried cables were installed in 1971, 1972, and 1973.
- In 1976, the utility started recording data, but there were no records from prior years.
- Method shown in the handbook is to use a t_0, which is an acceptable method. Lets look at how a Missing Data Analysis would compare

Fig 3-11, NWH

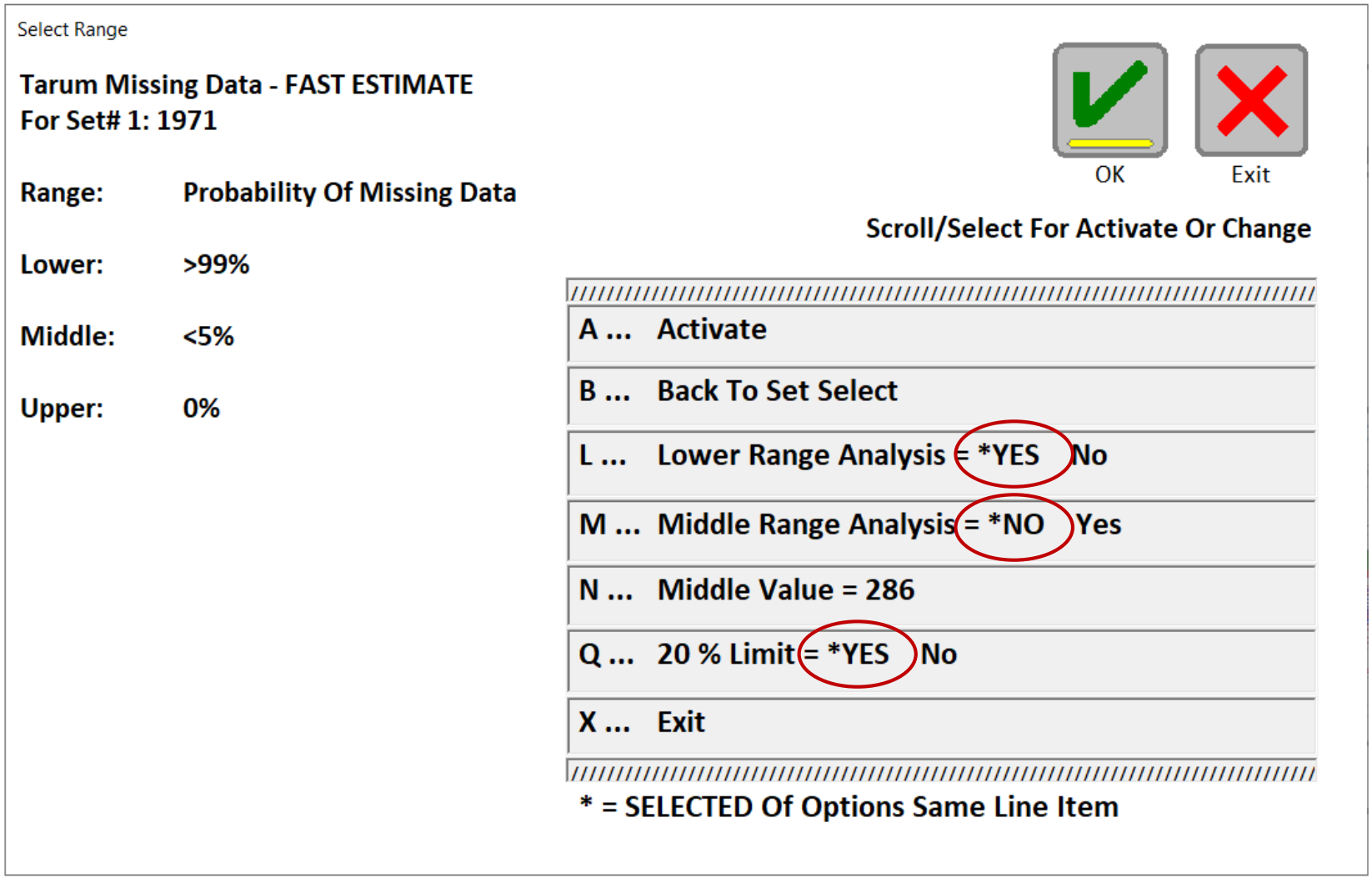

1971-72-73 Vintage Cables
W/rr/t0/insp1
t0(1) = 60.27
t0(2) = 47.97
t0(3) = 37.37
1973
1972
1971
Eta Beta pve% n/s
302.45 1.187 7.910 565/196
373.90 1.429 28.44 853/414
325.68 1.235 40.32 1165/534
Occurrence CDF %
90
80
70
60
50
40
30
20
10
5
2
1
0.5
0.2
0.1
0.05
0.02
0.01
10
100
1000
Age [Scale AR - As Recorded] (months)

Using SuperSMITH™ with missing data

- Open Fig3-11. This is in the SmithDat directory provided with the software
- Clear the t0. Click Icon, and select N, Reset All
- Change the method to rank regression and select inspection option #1
- Select the Missing Data icon under the tools group
- Select 1971



Verify options, then select OK or Activate



Results

SuperSMITH(R)

Tarum Missing Data Estimate

Click to Add to plot as new set

Select For Activate Or Change

Missing Data Estimate

Set:1 1971 D20-M09-YR2022

Eta	Beta	pve	o/s	Set
252.832	2.78394	0.335	369/196	1971 (Original)
277.551	1.43972	78.52	436/196	4 (Modified)

Additional Point Quantity

Quantity	X-Value	Point Type, Range
67	61	Occurrence, Lower
67	Total (12% Of Original Set.)	

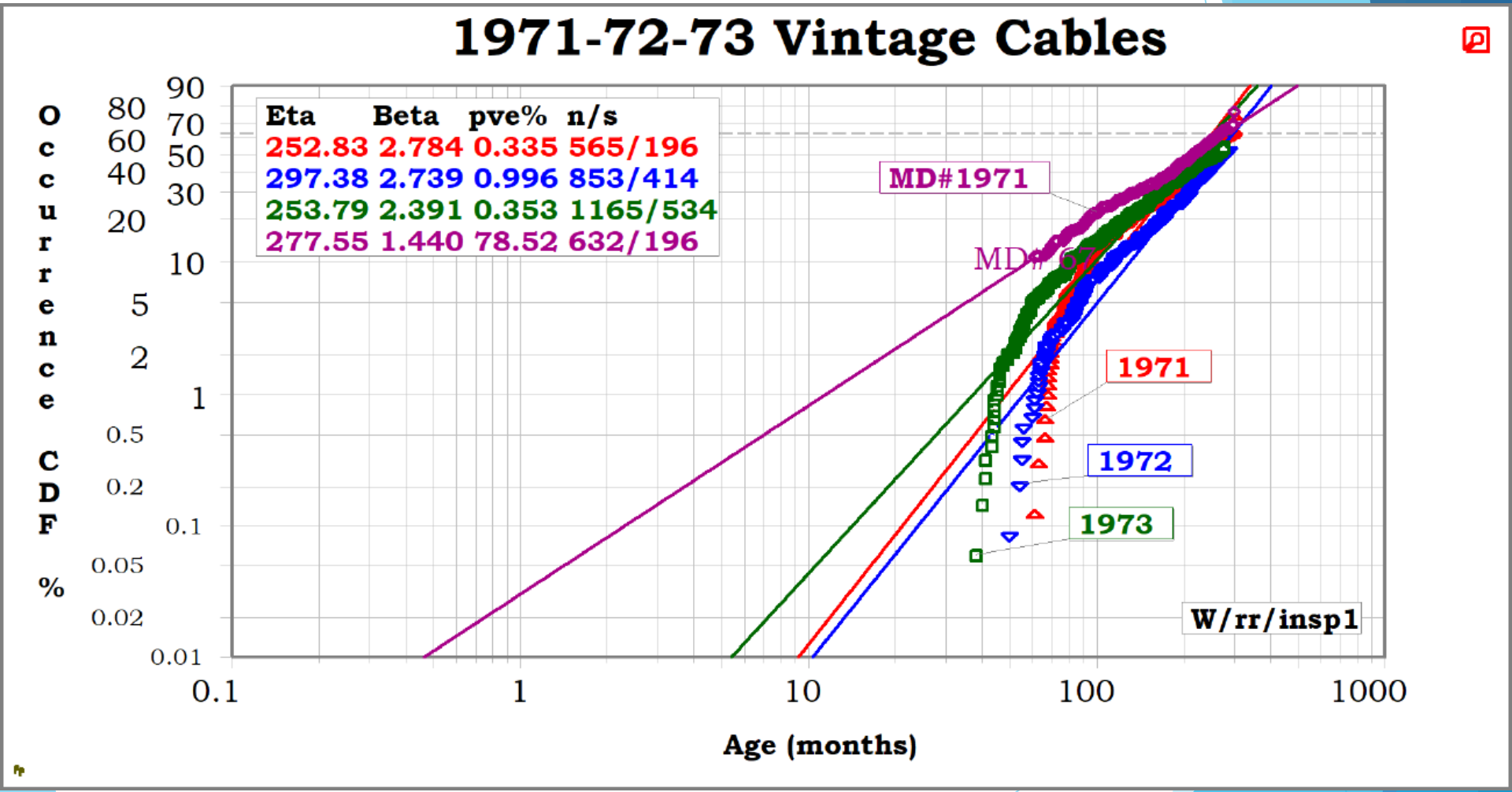
Comparison of parameters

- **A ... Add To Plot**
- **B ... Back To Range Select**
- **C ... Results To Clipboard**
- **L ... Results Add To Logbook**
- **P ... Results To Printer**
- **S ... Save**
- **X ... Exit**

Adds 67 points

Less than 20% points added

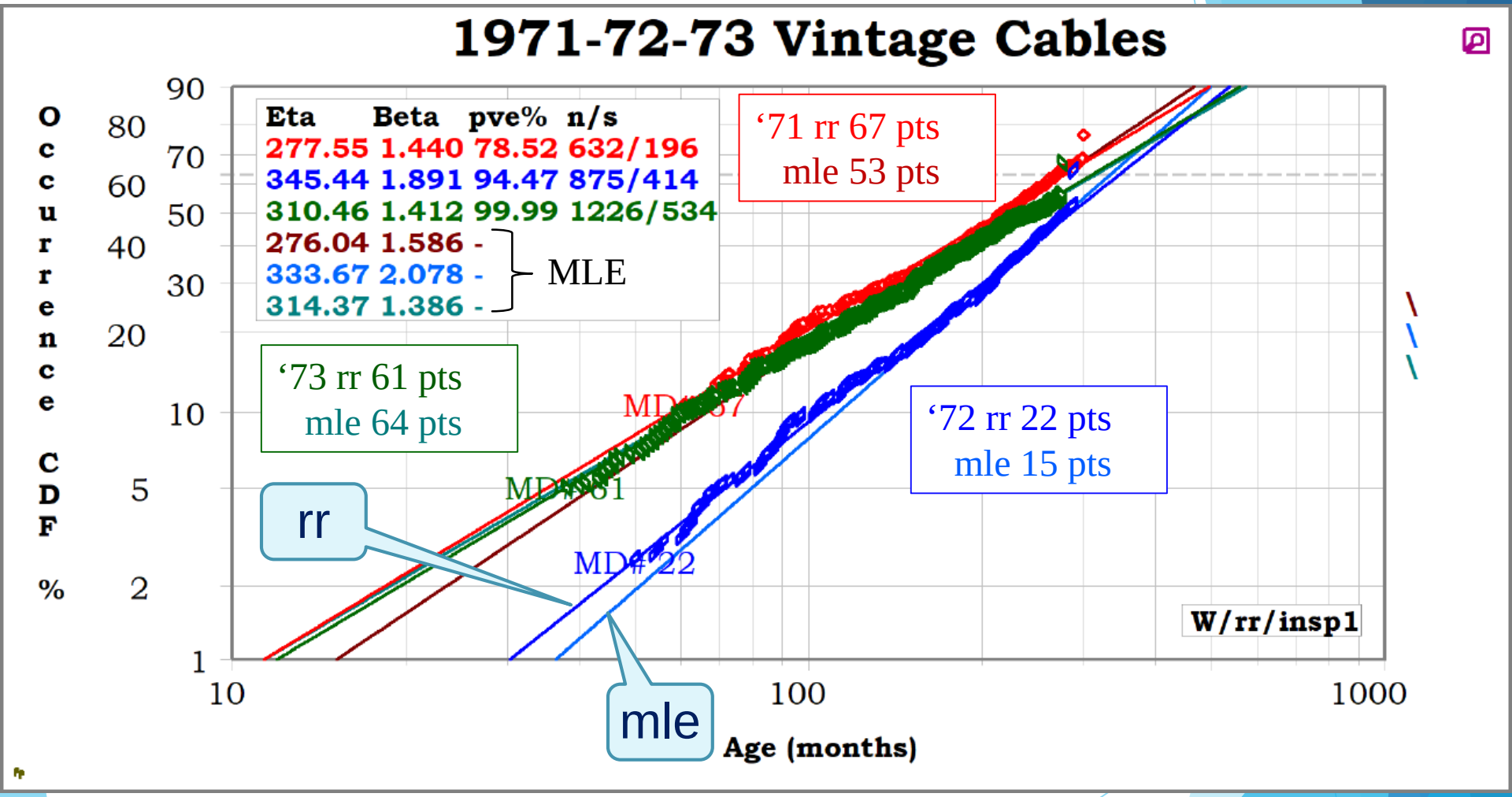
New line for 1971 data on plot



Repeat for other years

- You can repeat for the other years
- The following graphic shows all 3 sets with rank regression and MLE solutions.
  - Fits are very similar (Compare beta, eta, B values)

Vintage Cables missing data



Comments

- Method is based on Rank Regression (Least Squares) fits.
- MLE can be used to optimize for missing left tail data, or to plot MRR results for intermediate or right tail
- Number of required points for analysis and maximum percentage of points added is a recommendation, not a hard number based on simulations. Use judgement in your conditions.
- Use caution if you have many right-tail suspensions. (More than 30%) It could be a batch issue
- Method applies to other distributions (like lognormal), but they are not programmed yet