Application of heavy-tailed distribution using PROC IML, NLMIXED, and SEVERITY

PALASH SHARMA M.S.

DOCTORAL STUDENT, DEPARTMENT OF BIOSTATISTICS

THE UNIVERSITY OF KANSAS MEDICAL CENTER

M.S. IN STATISTICS

UNIVERSITY OF NEVADA, RENO

PASSION: APPLIED HEALTH DATA ANALYTICS AND BIOSTATISTICS

JOHN KEIGHLEY, PH.D.

ASSISTANT PROFESSOR, DEPARTMENT OF BIOSTATISTICS

UNIVERSITY OF KANSAS MEDICAL CENTER

UNIVERSITY OF KANSAS CANCER RESEARCH CENTER, KANSAS CANCER REGISTRY

Outline

- Univariate Heavy tailed distribution
- Generalized Pareto distribution
- Simulation technique
- Numerical optimization using PROC NLMIXED
- Application using PROC SEVERITY
  - Cancer patient Data
  - USA Blackout Data.

Univariate Heavy tailed Distribution:

PARETO DISTRIBUTION

Generalized Pareto distribution

- The probability density function of the three parameter Generalized Pareto distribution is

$$f_X(x) = \frac{1}{\sigma}\left(1 + \frac{\xi(x-\mu)}{\sigma}\right)^{-(1+\frac{1}{\xi})} \quad , x \geq \mu, \sigma \geq 0, \xi > 0$$

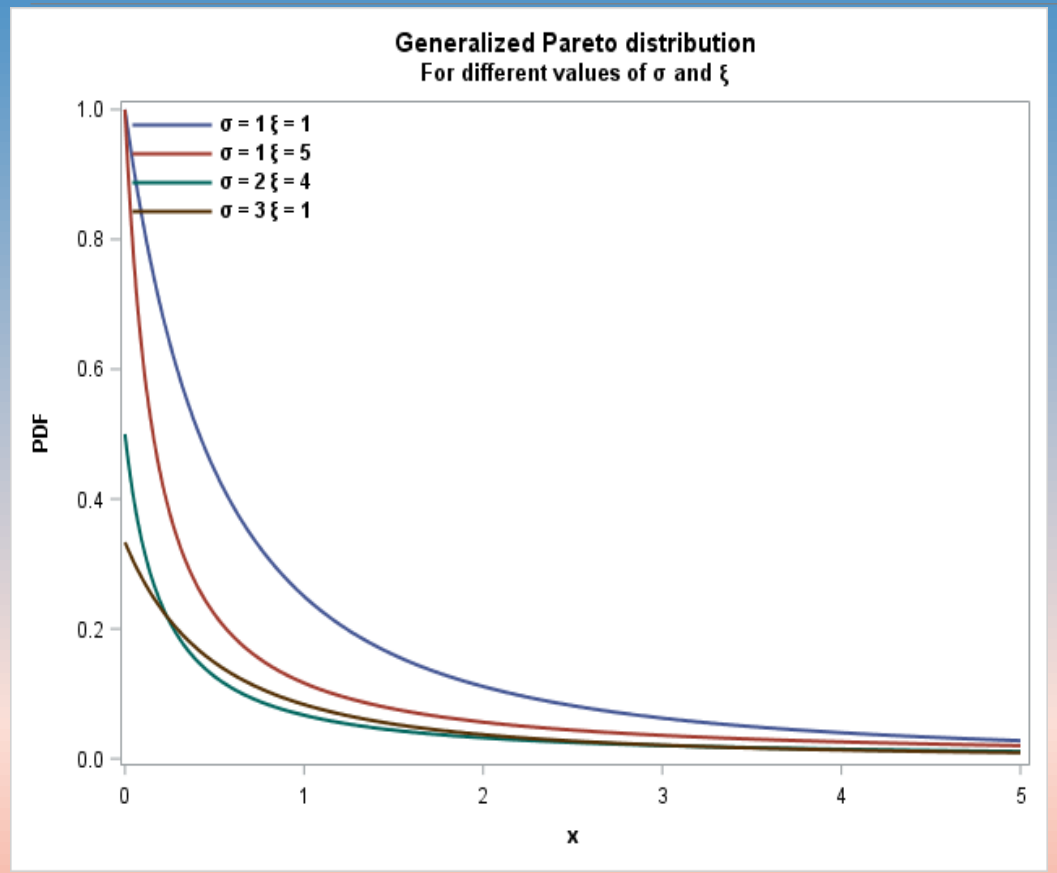
- Two parameter GPD model is,

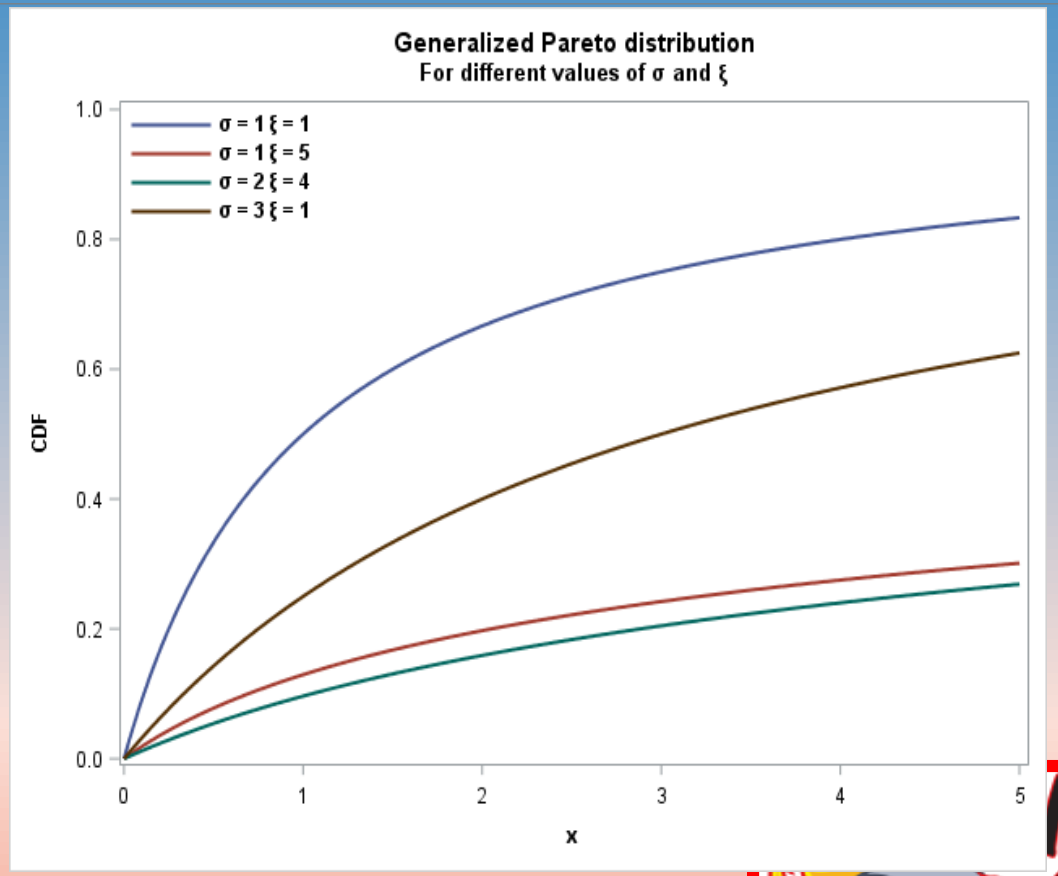
$$f_X(x) = \frac{1}{\sigma}\left(1 + \frac{\xi x}{\sigma}\right)^{-(1+\frac{1}{\xi})} \quad , x \geq 0, \sigma \geq 0, \xi > 0$$

- The cumulative distribution function of the GPD can be written as,

$$F_X(x) = 1 - \left(1 + \frac{\xi x}{\sigma}\right)^{\frac{-1}{\xi}} \quad , \sigma \geq 0, \xi > 0$$

Generalized Pareto distribution





Simulation Technique

- Using the probability integral transform theorem

- The following closed form expression for generating random variable arising from GPD model.

$$X = F^{-1}(U) = \frac{\sigma}{\xi}\left[\frac{1}{(1-U)^{\xi}} - 1\right]$$

Where U ~Uni (0,1),

NUMERICAL OPTIMIZATION USING PROC NLMIXED

Table1: Standard error and estimated parameter when is $\sigma = 2$ and $\xi = 3$

n	Parameters	Value	S.E	95% C.I.
50	$\hat{\sigma}$	2.0062	0.01133	1.9840 – 2.0284
	$\hat{\xi}$	2.9898	0.007977	2.9742 – 3.0054
500	$\hat{\sigma}$	2.0052	0.003587	1.9982 – 2.0123
	$\hat{\xi}$	2.9984	0.002529	2.9935 – 3.0034
800	$\hat{\sigma}$	2.0028	0.002833	1.9972 – 2.0083
	$\hat{\xi}$	2.9997	0.002000	2.9958 – 3.0036
1000	$\hat{\sigma}$	2.0010	0.002532	1.9960 – 2.0059
	$\hat{\xi}$	3.0004	0.001789	2.9969 – 3.0039

Application:

APPLICATION OF GPD MODEL USING PROC SEVERITY

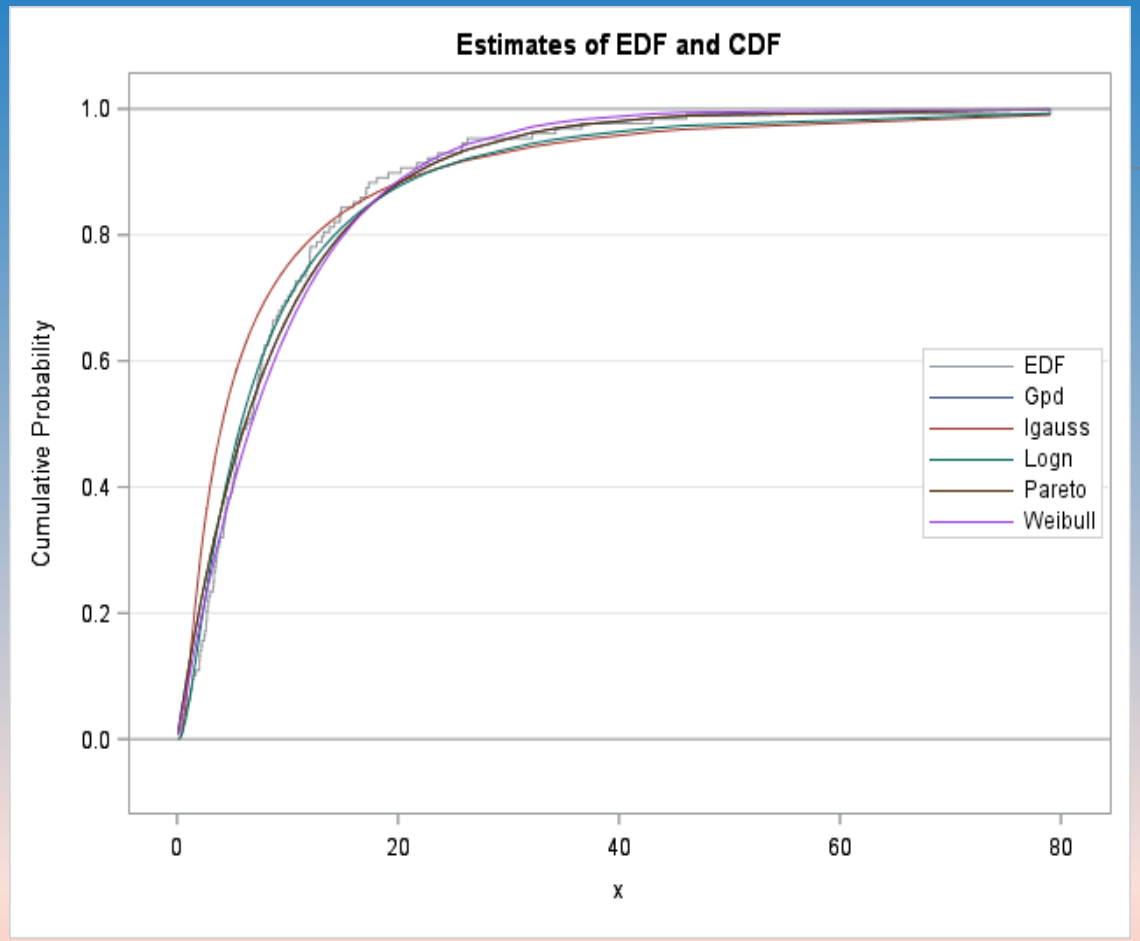
Cancer Patient Data: Model Fit

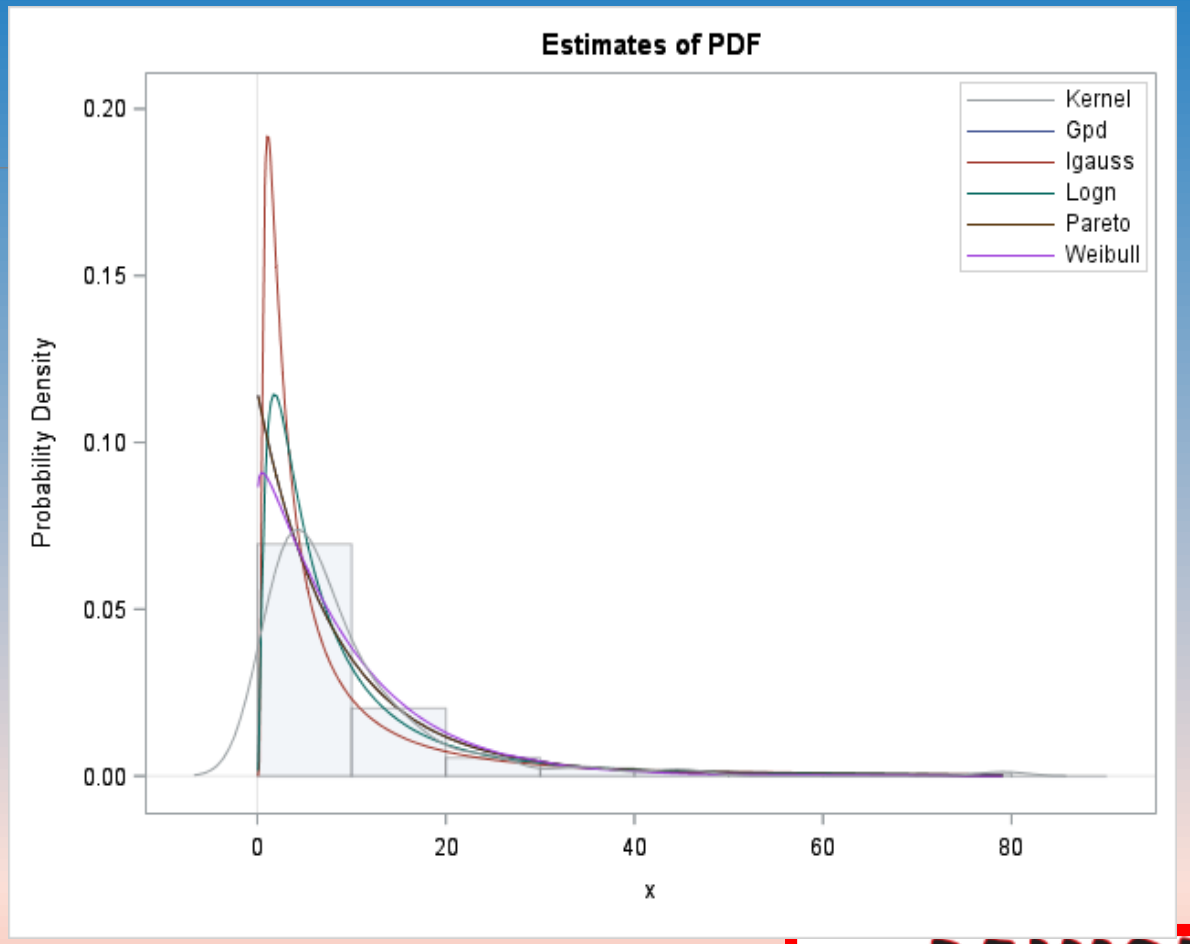
All Fit Statistics

Distribution	-2 Log Likelihood		AIC		AICC		BIC		KS		AD	
Gpd	827.65117	*	831.65117	*	831.74717	*	837.35523	*	1.11074		1.42650	
Igauss	880.57157		884.57157		884.66757		890.27563		2.19822		7.86204	
Logn	830.19194		834.19194		834.28794		839.89600		0.71672	*	0.86238	*
Pareto	827.65117		831.65117		831.74717		837.35523		1.11076		1.42654	
Weibull	828.16494		832.16494		832.26094		837.86900		0.80831		1.02110	

Note: The asterisk (*) marks the best model according to each column's criterion.

Model Assessment



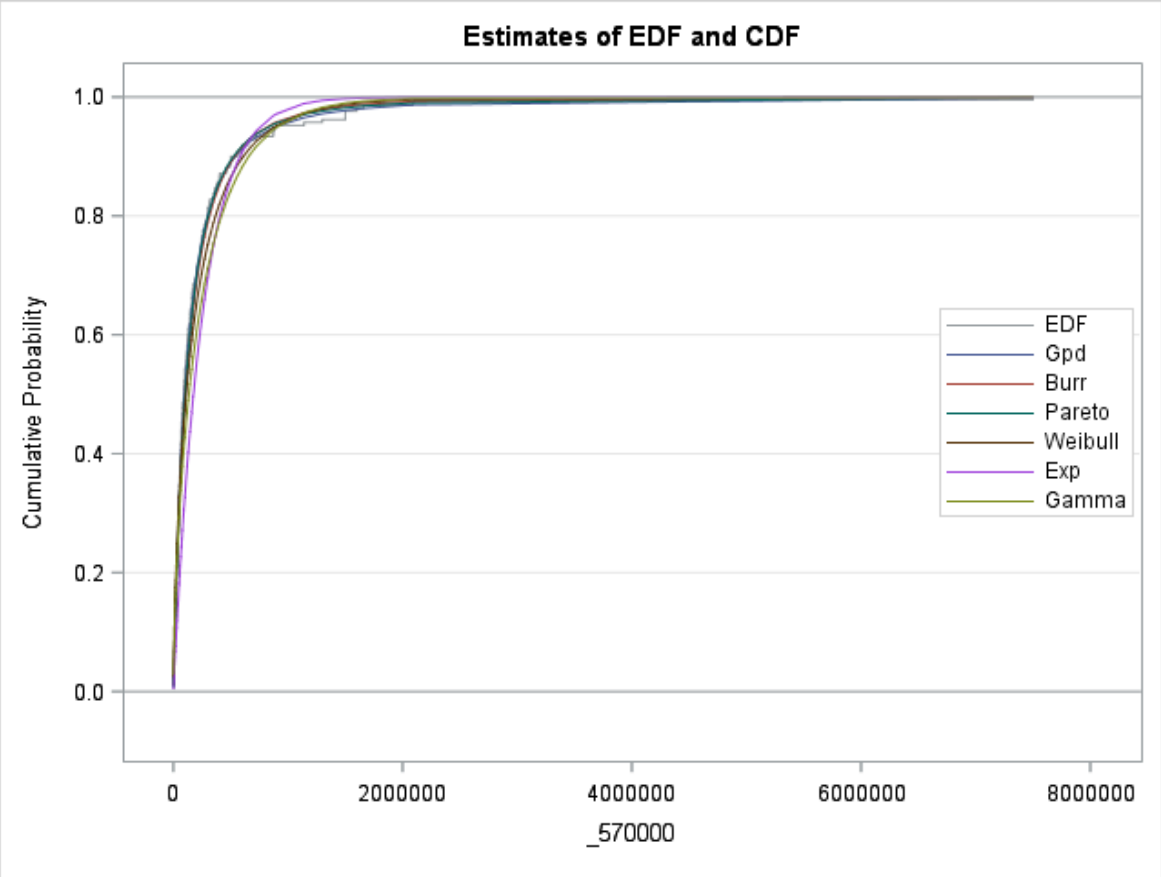


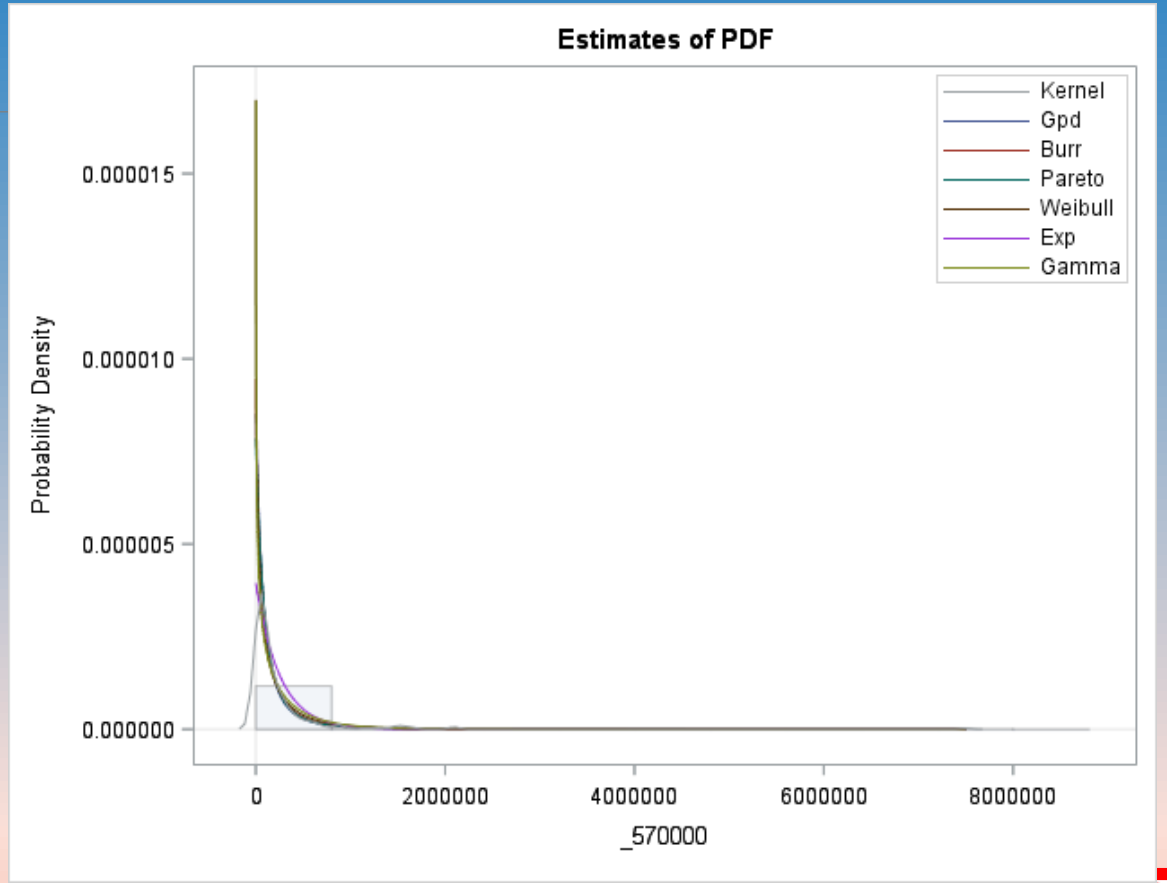
USA Blackout Data : Model Fit

All Fit Statistics													
Distribution	-2 Log Likelihood		AIC		AICC		BIC		KS		AD		CvM	
Gpd	5549	*	5553	*	5553	*	5560	*	0.73019	*	0.96456	*	0.08680	*
Burr	5554		5560		5560		5570		0.92229		1.47112		0.19399	
Pareto	5550		5554		5554		5561		0.86627		1.10749		0.13371	
Weibull	5579		5583		5583		5590		1.16498		3.46025		0.48864	
Exp	5644		5646		5646		5650		3.13880		19.50451		3.79835	
Gamma	5605		5609		5609		5615		1.82088		6.39766		1.15394	
Note: The asterisk (*) marks the best model according to each column's criterion.														

Estimation Details						
Distribution	Converged	Iterations	Function Calls	Gradient Updates	Hessian Updates	Time (Seconds)
Gpd	Yes	29	68	527	495	0.00
Burr	Maybe	100	213	5355	5252	0.02
Pareto	Maybe	100	264	5355	5252	0.00
Weibull	Yes	36	85	779	740	0.00
Exp	Yes	0	4	5	2	0.00
Gamma	Yes	3	16	20	14	0.00

Model Assessment





Conclusion

- Pareto and generalized version of Pareto distribution.
- The method of maximum likelihood is suggested using PROC NLMIXED.
- PROC SEVERITY is used for model fitting and model assessment.
- Generalized Pareto distribution provides significantly better fit than the other distributions.

Future Research

- Develop new probability distribution theory.
- Develop their statistical property.
- Real life Application with the new univariate distribution.
- Apply various datasets and distributions to compare the best fit.

References

- SAS/IML® software 14.2. SAS® Institute Inc. 2016, Cary, NC.
- SAS/STAT® software 14.2. SAS® Institute Inc. 2016, Cary, NC.
- Aban, I.B., Meerschaert, M.M. and Panorska, A.K. 2006. "Parameter estimation for the truncated Pareto distribution", *J. Am. Statist. Assoc.* 101(473) pp. 270–277.
- Akinsete, A., Famoye, F. and Lee, C. 2008. ''The beta-Pareto distribution.'' *Statistics*, *42*(6), 547-563.
- Chen, J., Thorp, J. S. and Parashar, M. 2001, Analysis of electric power disturbance data. *In 34th Hawaii International Conference on System Sciences, IEEE Computer Society*.
- Clauset, A, Shalizi, C.R. and Newman, M.E.J.2009. "Power-law distributions in empirical data" *SIAM Review* 51(4), 661-703. (arXiv:0706.1062).
- Eugene, N., Lee, C. and Famoye, F. 2002. ''The beta-normal distribution and its applications,'' *Commun. Statist. Theory Meth.*, pp. 497–512.
- Levy, M. and Levy, H. 2003. ''Investment talent and the Pareto wealth distribution: Theoretical and experimental analysis,'' *Rev. Econ. Statist*. 85(3), pp. 709–725.
- Lee, E. T. and Wang, J. 2003. ''Statistical methods for survival data analysis''. *volume 476. John Wiley & Sons.*
- Mahmoudi, E. 2011. "The beta generalized Pareto distribution with application to lifetime data.'' *Mathematics and Computers in Simulation*, 81, 11, 2414–2430.
- Newman, M.E.J. 2005. Power laws, Pareto distributions and Zipf's law, *Contemp. Phys.* 46(5), pp. 323–351.

QUESTIONS?

Palash Sharma, M.S

psharma4@kumc.edu

John Keighley, Ph.D.

jkeighle@kumc.edu

Presenter: Palash Sharma

Institution: The University of Kansas Medical Center

Address: 3901 Rainbow Blvd, Kansas City, KS 66160

Phone: 775-203-7920

Email: psharma4@kumc.edu

Web: https:// palash63@github.io