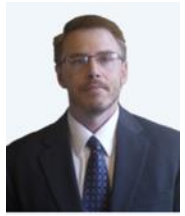
Probability Methods
Data Visualization

Aircraft Airworthiness and Sustainment Conference

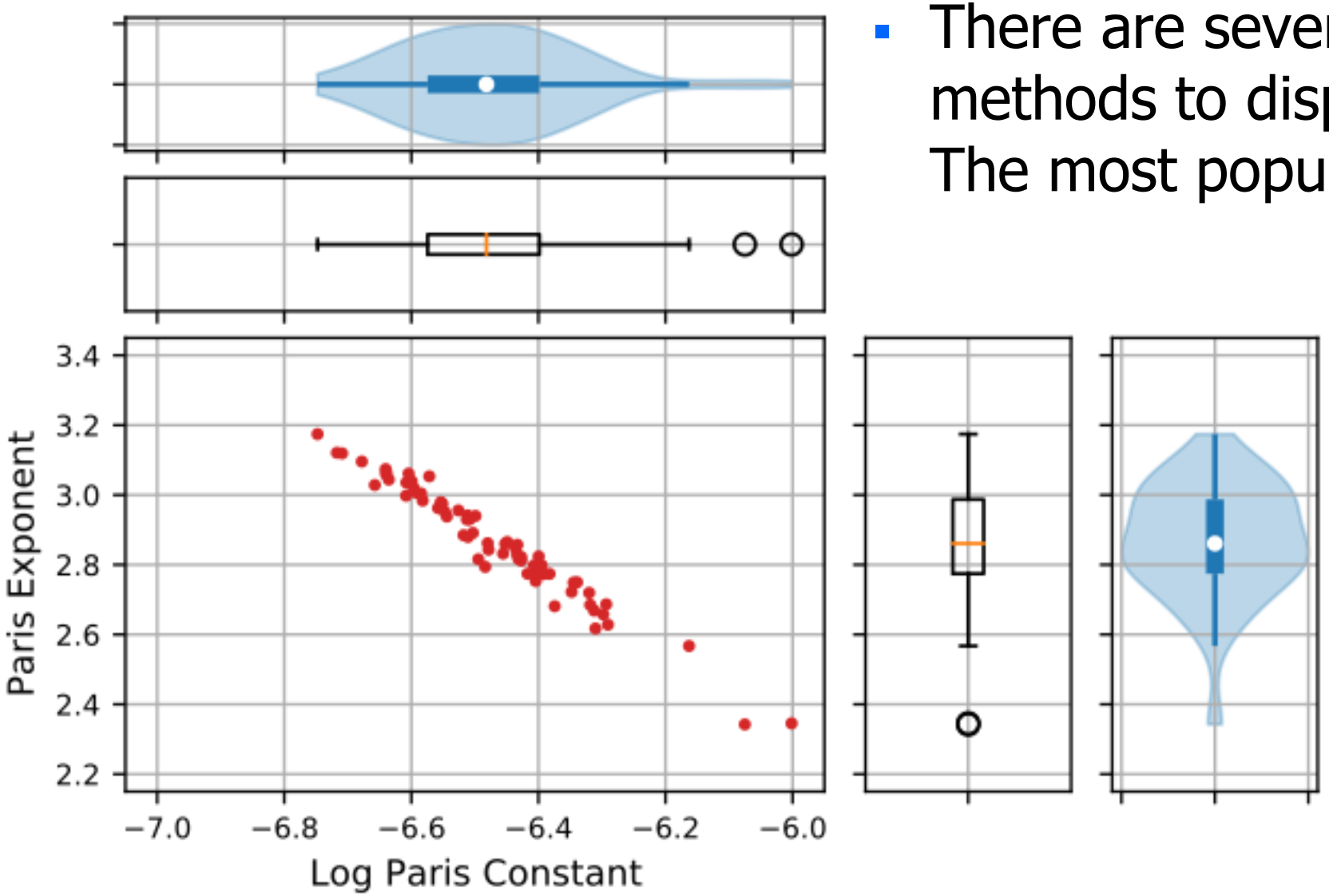
August 29, 2022






AA&S The Aircraft Airworthiness & Sustainment Conference
AUGUST 29 - SEPTEMBER 1, 2022

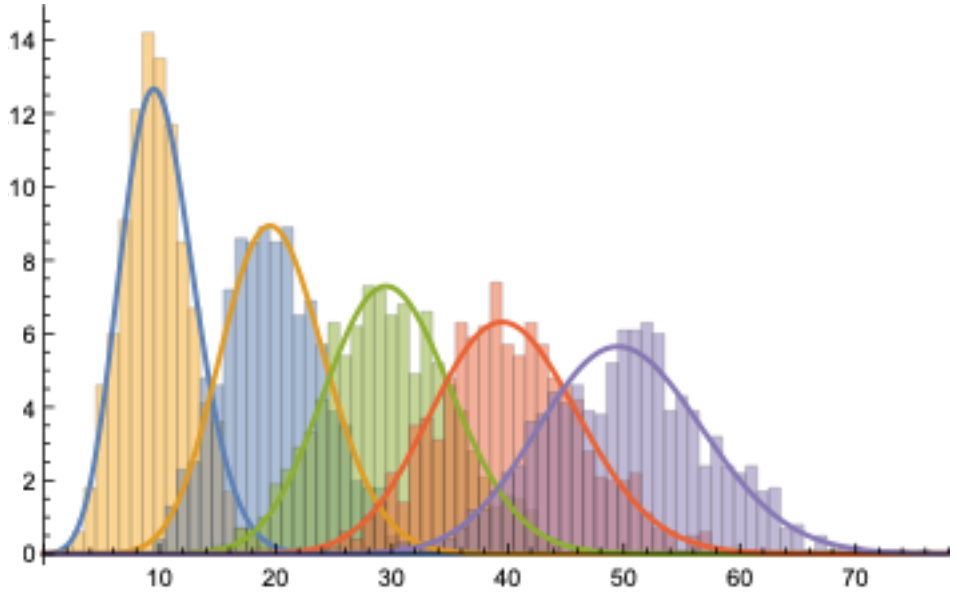
Overview



- There are several popular visualization methods to display probability data. The most popular are:
  - Histogram
  - Box-Whisker
  - Violin plots
  - Scatter plots

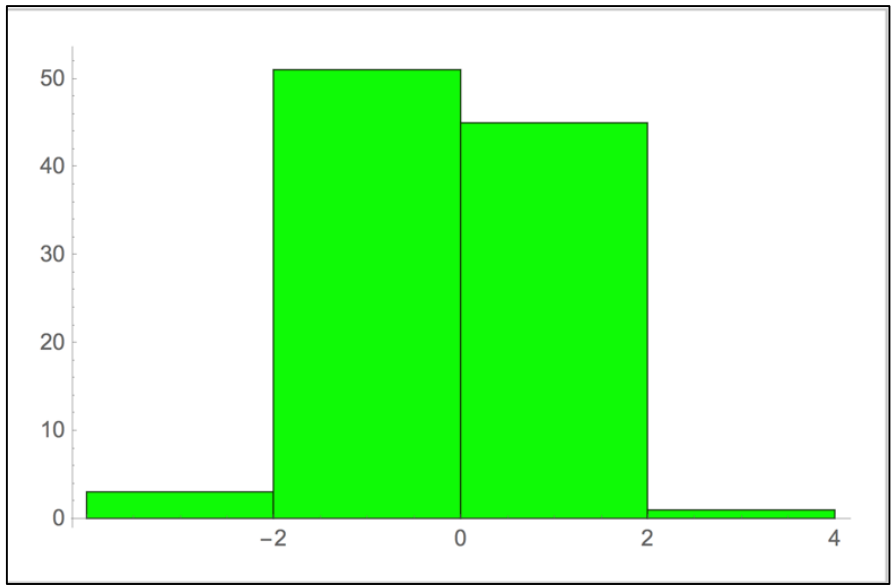
Histogram

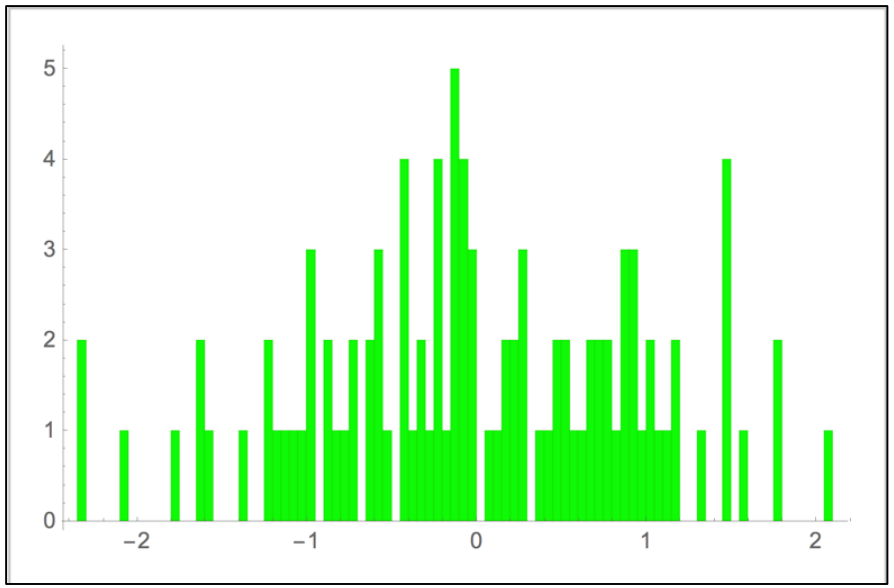
- Histograms are a natural tool to analyze data. The procedure is simple, define bins for your data, count the number of data points is each bin. Plot a column for each bin corresponding to the number of data points in that bin. This gives a natural basis to see where the data is concentrated



Histogram: weakness

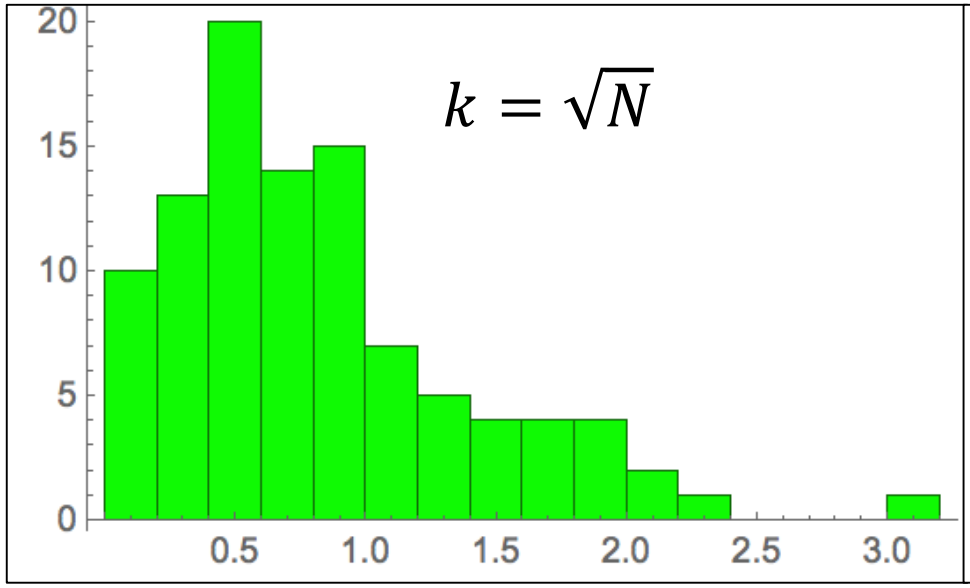
- The weakness with histograms is that their impact depends on the number and width of the bins
  - too few bins and the data is too bunched up
  - too many bins and the data is too spread out.

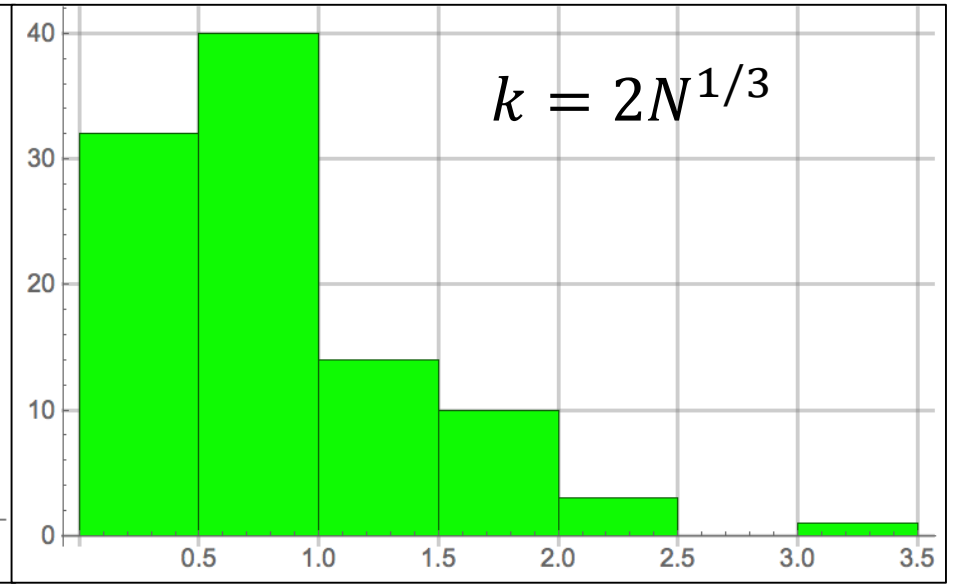




Histogram: Selecting the number of bins

- Rules for picking the number of bins for N data pts:
  - # bins = sqrt(N)
  - # bins = ln(N)+1
  - # bins = $2n^{1/3}$

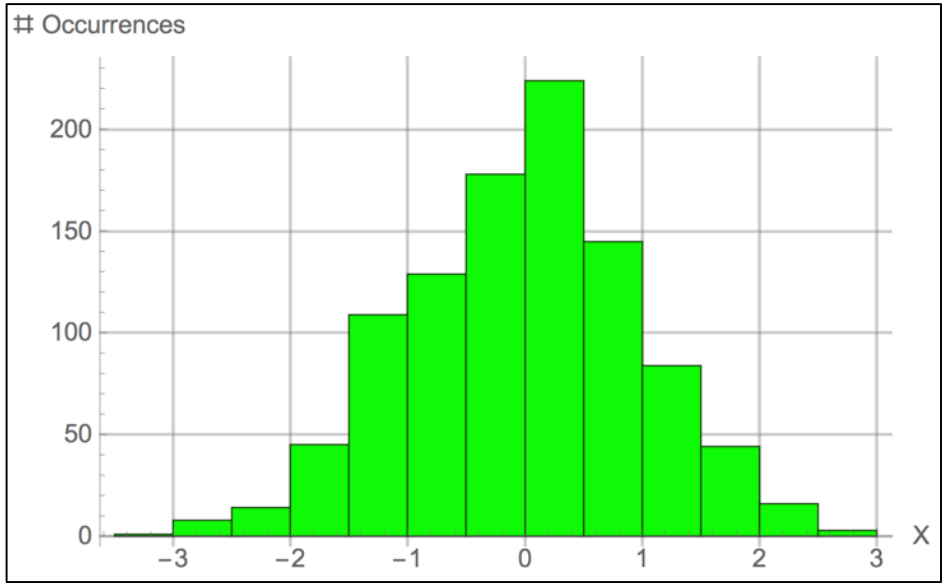




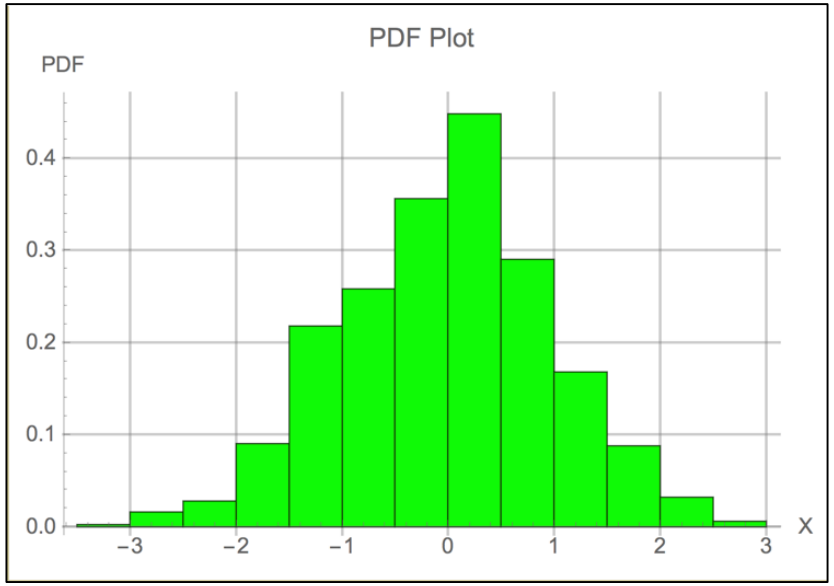
Converting a Histogram to a PDF

- A PDF is defined such that the area under the curve = 1. Hence, one can compute the area of the histogram then divide by that value to create a PDF.
  - If the bin width is constant, then the area is simple the total #pts * the bin width.

Histogram

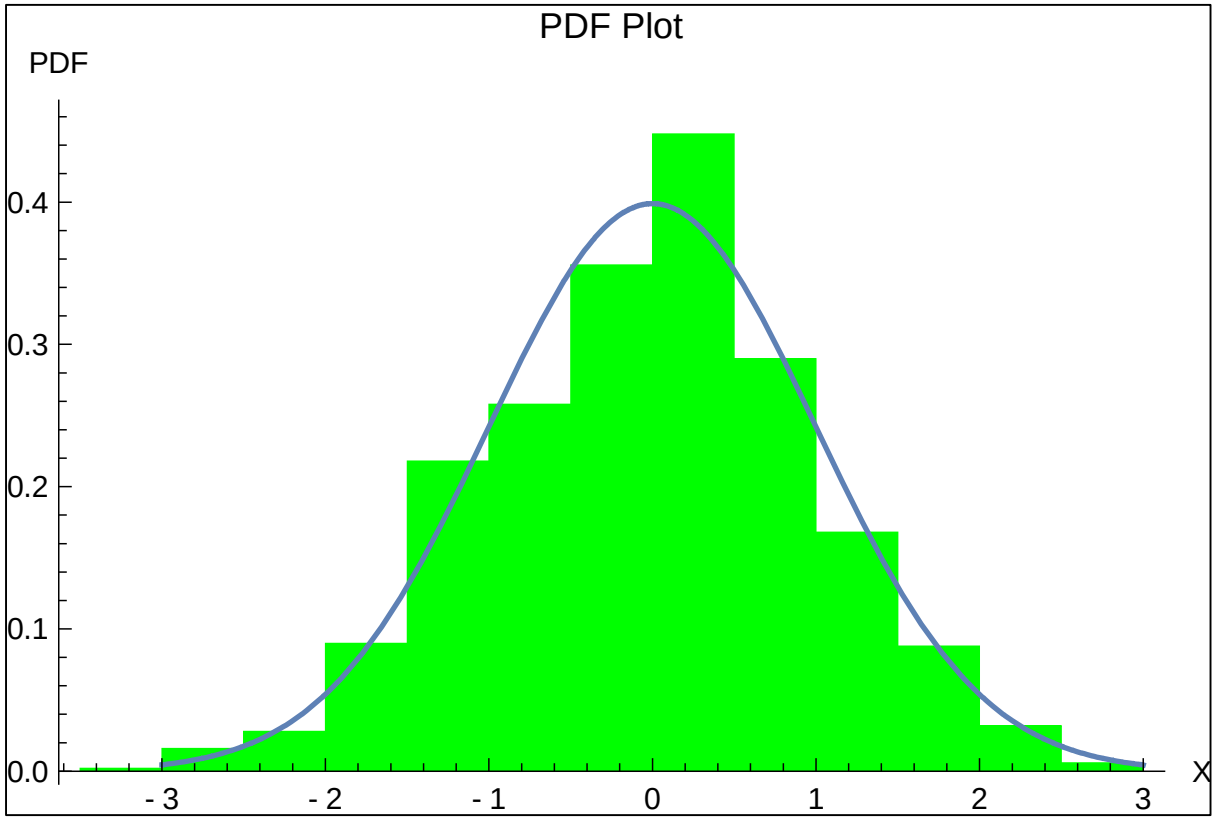


PDF

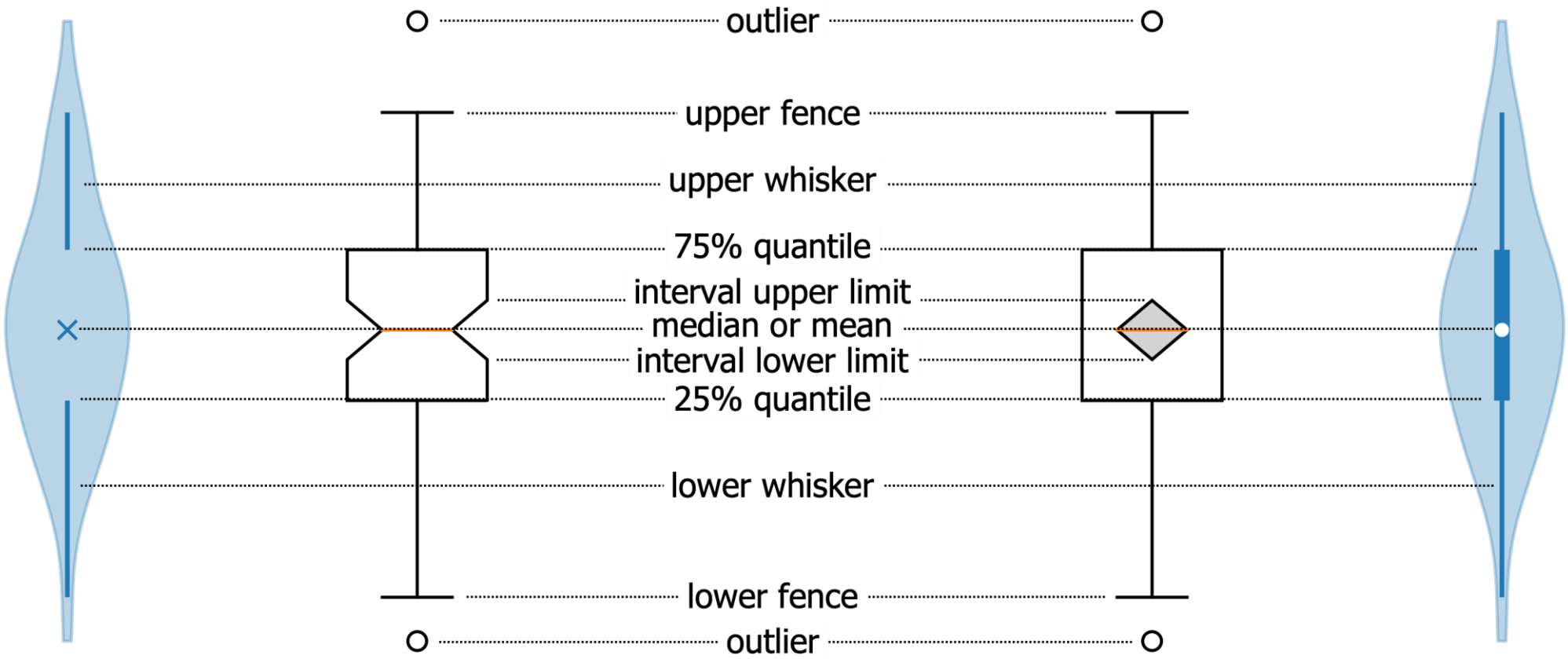


Converting a Histogram to a PDF

- Once a PDF is defined, then a functional form of the PDF can be plotted on top of the data. Note, however, the PDF is still susceptible to the no. of bins issue.



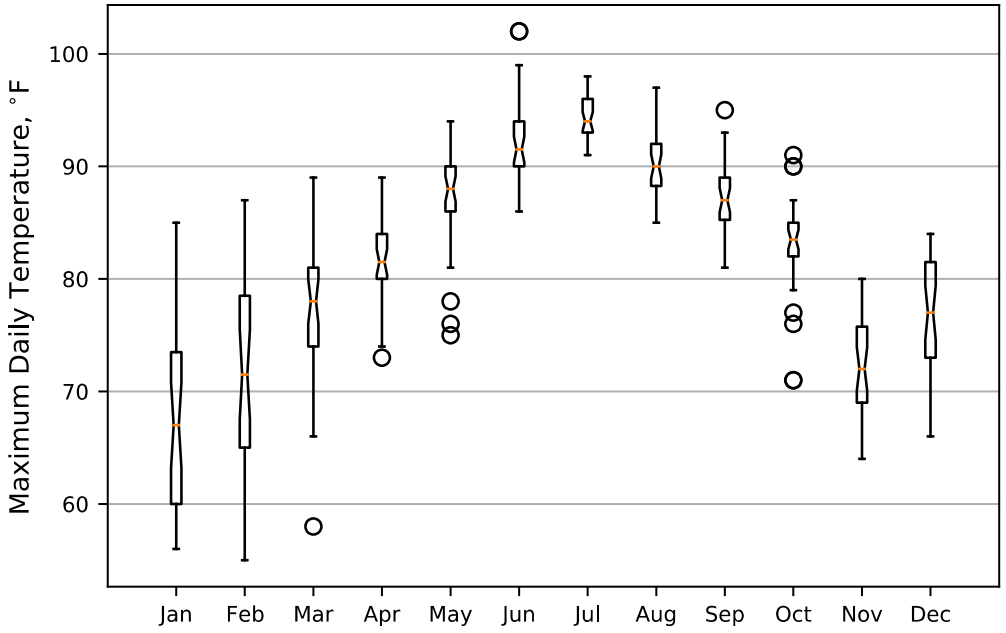
Box-Whisker and Violin Plots



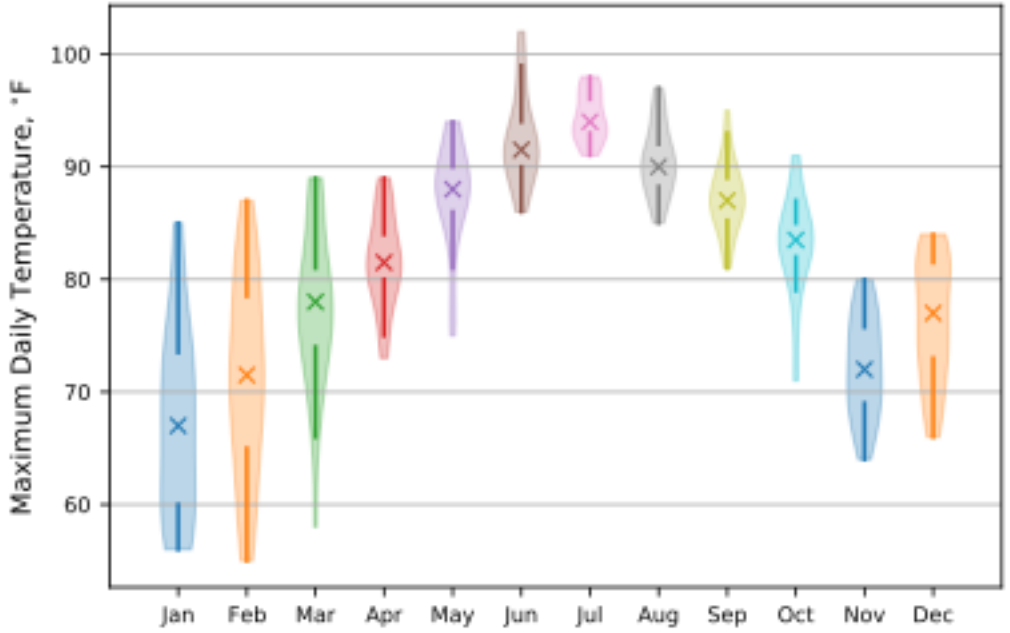
- Box-Whisker plots visualize the data in terms of mean/median, 25/75% quantiles, and outliers.
- Violin plots show most of the same information, providing a representation of the probability density but not showing outliers

Box-Whisker and Violin plots – multiple data sets

Box-Whisker



Violin

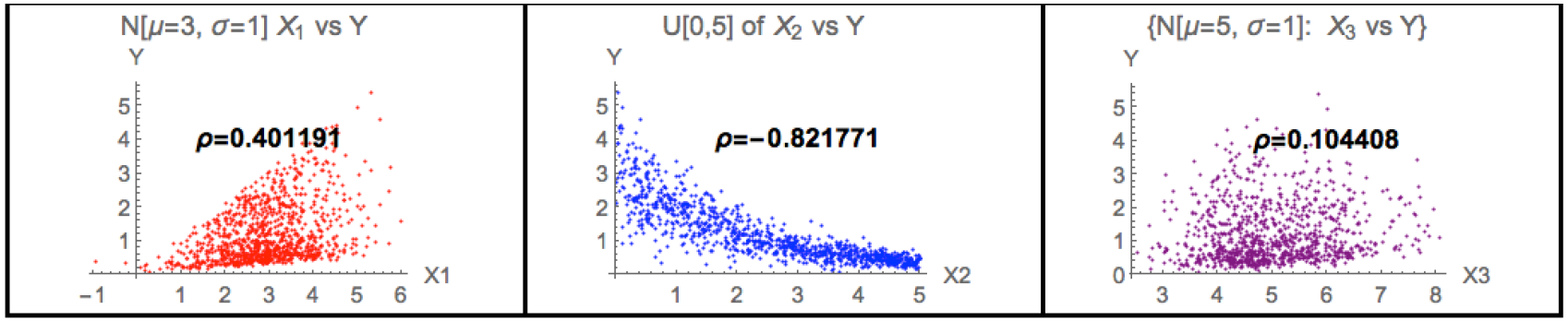


- Compare Box-Whisker and Violin plots of Jacksonville, FL daily high temperatures for each month over the last year
  - Both make it easy to compare the month-to-month temperature ranges
  - Box-Whisker: individual outliers clearly visible (only 1 day over 100 F)
  - Violin: densities for each month can be seen

Scatter Plots

- Data point by data point plot to explore the relationship between two parameters.

$$Y = |x_1|\, e^{-x_2/2} + 0.001x_3^3$$

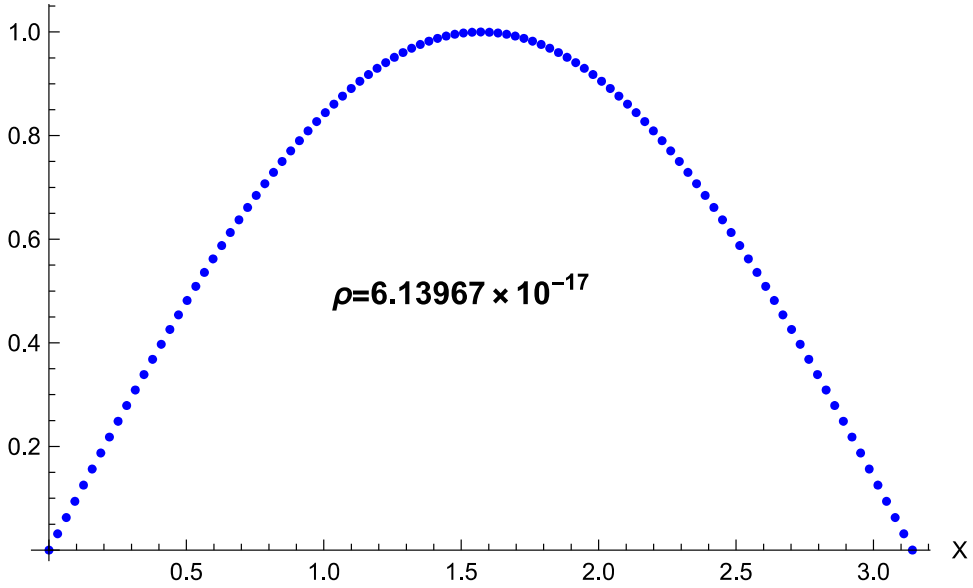


Strong relationship between X_1 and Y and X_2 and Y. X_3 has a minor effect.

Scatter Plots

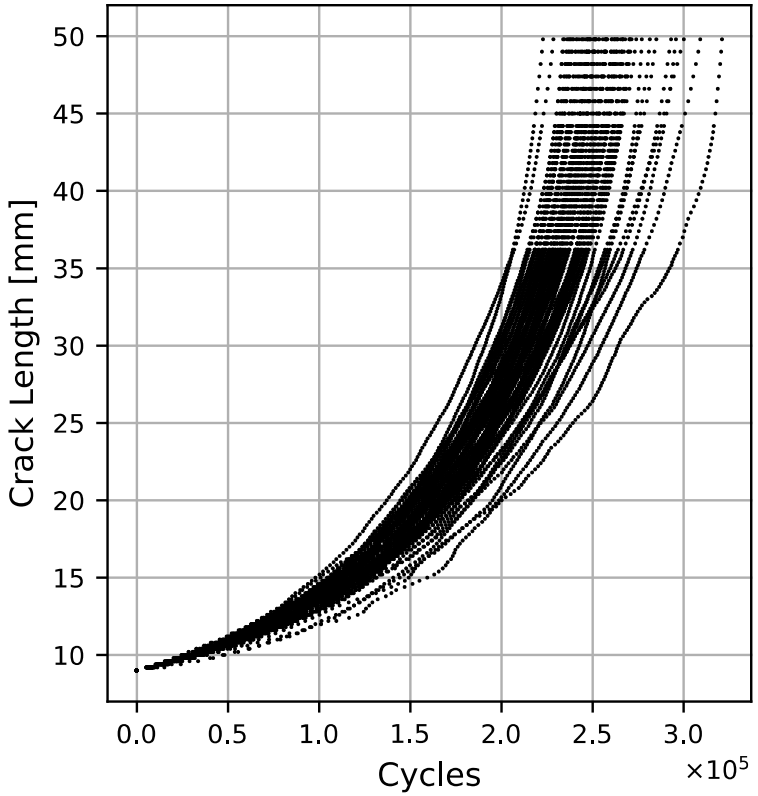
- Cautionary example for the use of the correlation coefficient. The correlation coefficient only shows the "linear" relationship between parameters. Nonlinear relationships cannot be discerned

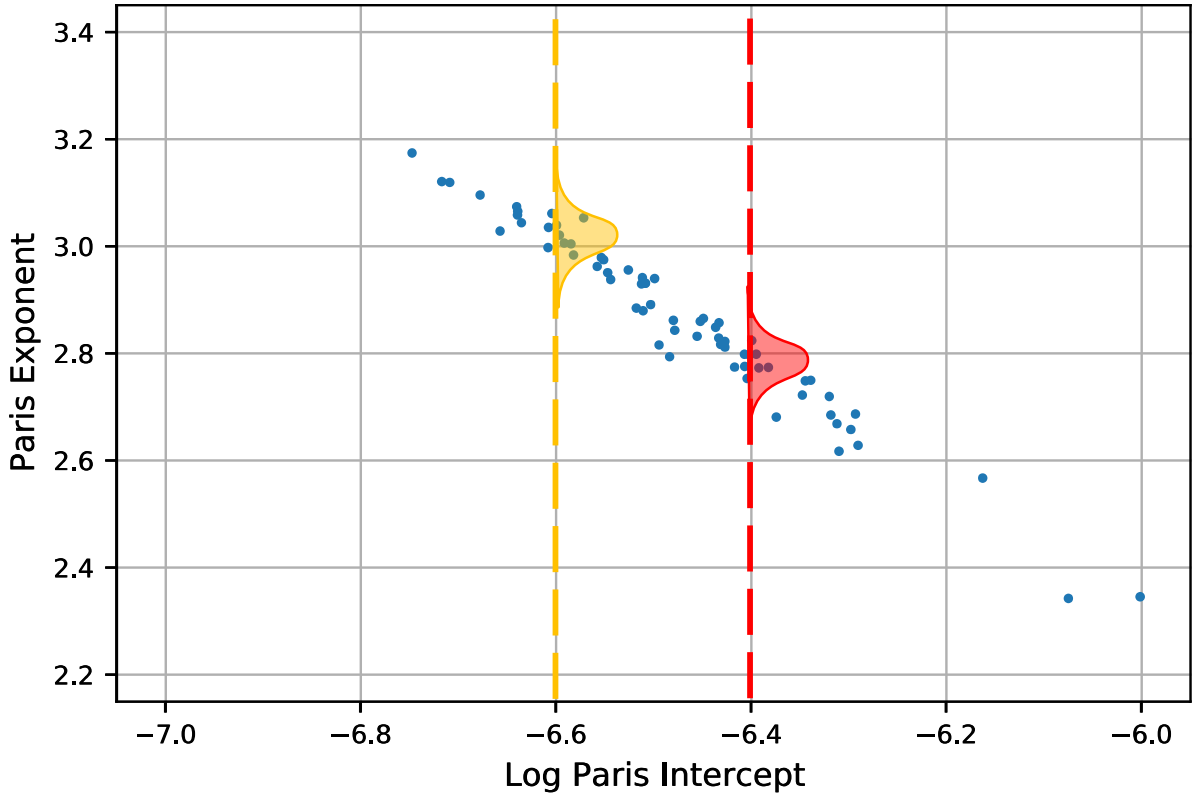
$$Y = \sin(x)$$



$\rho = 0$ even though there is a strong relationship

Correlated Variables

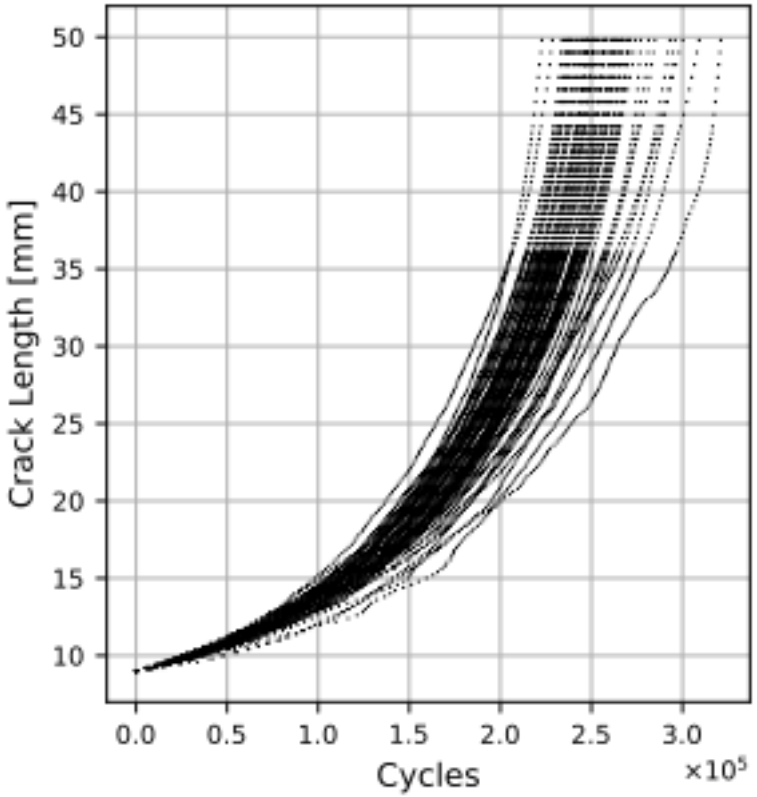


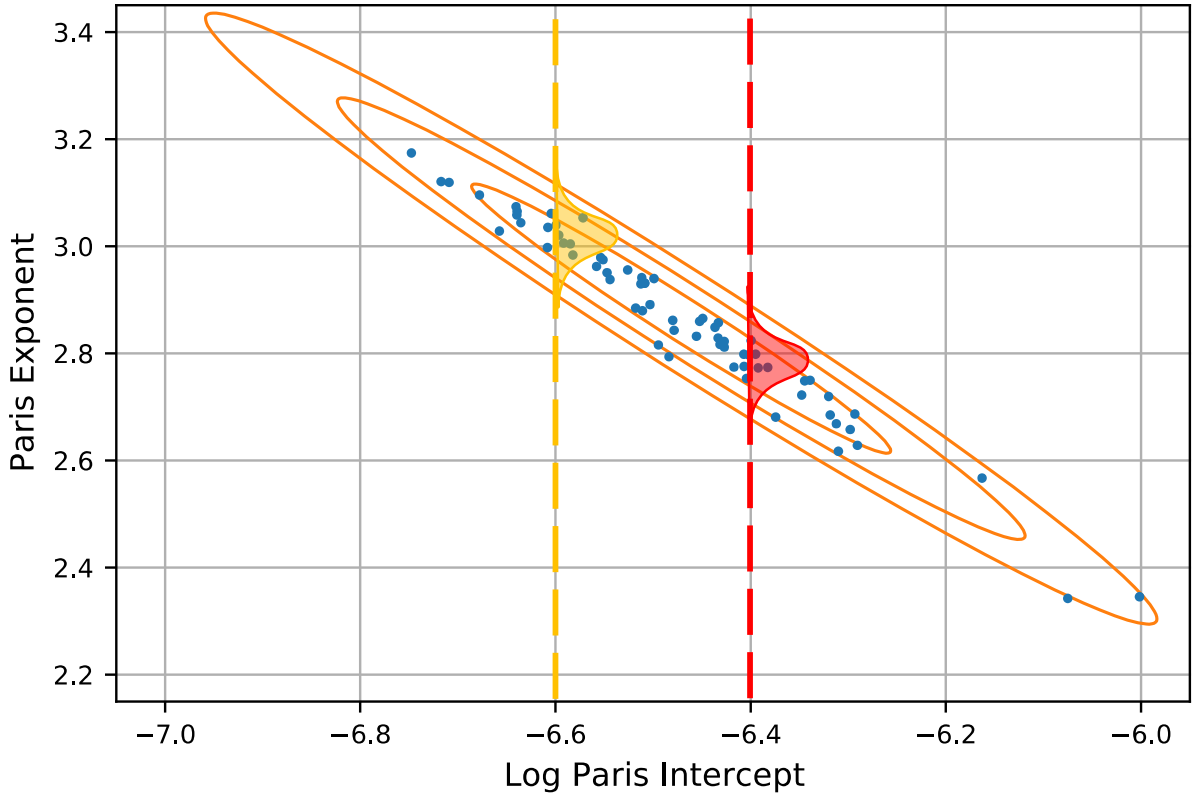


- When the value one variable is likely to take depends on another variable, the variables are correlated
- Virkler's 68 replicated crack growth experiments on 2024-T3 AL showed strong correlation between Paris Law crack growth model coefficients

Virkler D.A., Hillberry B., and Goel P.K., "The statistical nature of fatigue crack propagation.", J. Eng. Mater. Technol. Apr 1979, 101(2): 148-153, 1989

Correlated Variables

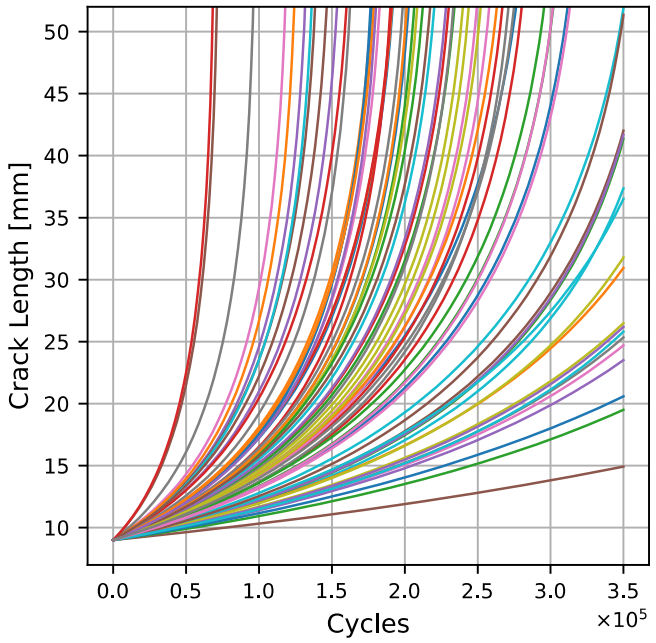




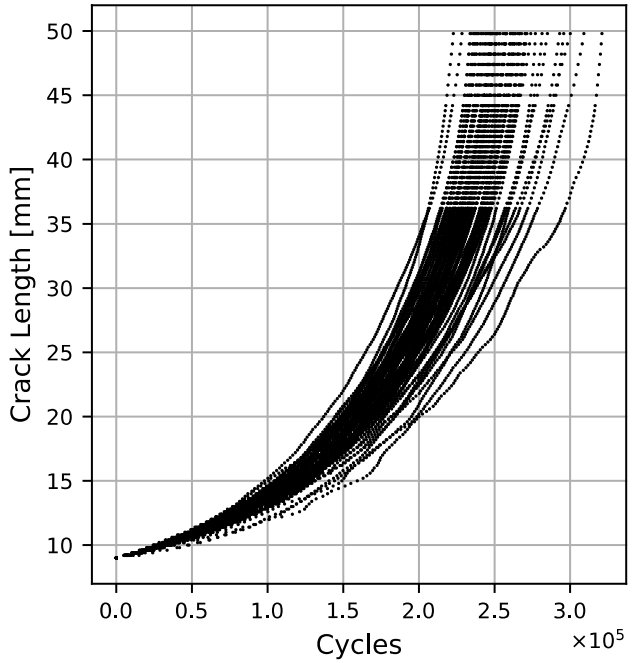
- For uncorrelated variables, the joint PDF is the product of the individual PDFs
  - $f_{\mathbf{X}}(\mathbf{x}) = f_{X_1}(x_1) \cdots f_{X_n}(x_n)$
- For correlated variables, the joint PDF function typically does not reduce to a function of individual random variables
- Multivariate Normal distribution can be written as a matrix function of individual random variables:
  - $f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{2\,\pi\,|\boldsymbol{\Sigma}|}}\, e^{\left(-\frac{1}{2}\,(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{x}-\boldsymbol{\mu})\right)}$

Effect of Correlation on Simulation Results

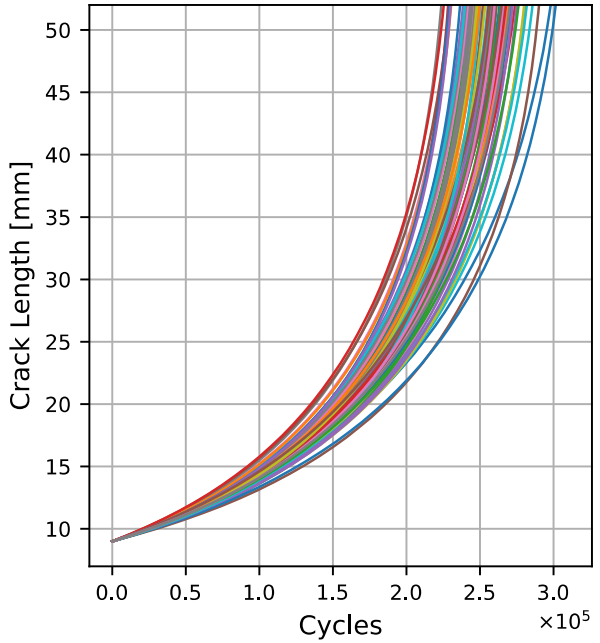
Simulated crack growth without correlation



Virkler crack growth data



Simulated crack growth with correlation



- Neglecting correlation yields much lower life
  - nearly half of the samples fail before the first observed in Virkler's data
- Including correlation captures nearly all of the observed variation

Questions