PGE 337 Data Analytics and Geostatistics

Lecture 3: Displaying Distributions

Lecture outline . . .

- **Plotting, Data Visualization**
- **Histograms, Probability Density Functions**
- **Cumulative Distribution Functions**

Introduction

General Concepts

Univariate

- PDF / CDF
- Statistics
- Distributions
- Heterogeneity
- Hypothesis

Bivariate

Time Series Analysis

Spatial Analysis

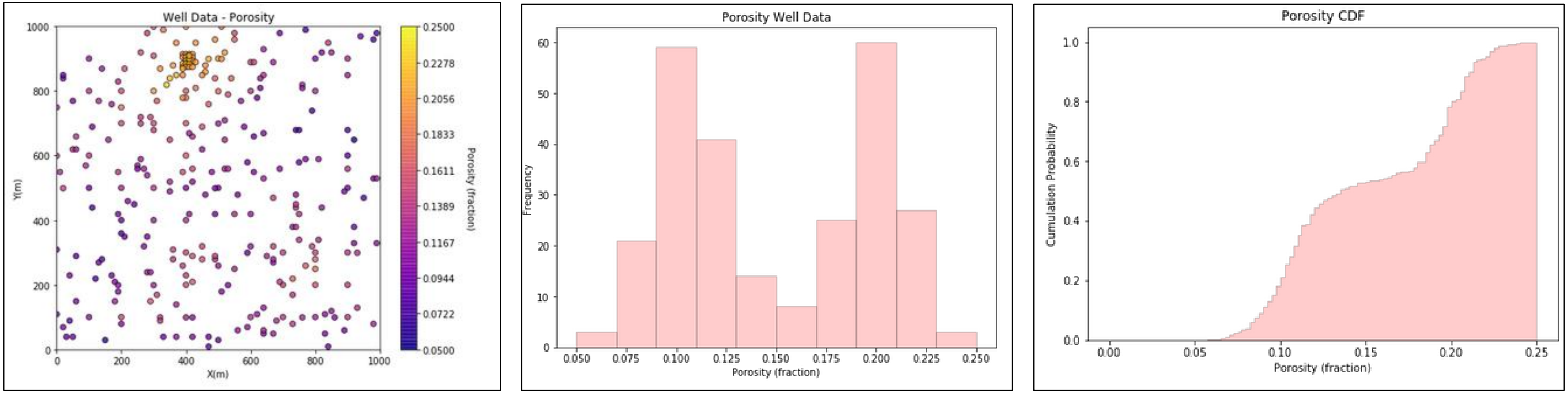
Machine Learning

Uncertainty Analysis

Michael Pyrcz, The University of Texas at Austin

Texas Center for Data Analytics and Geostatistics
The University of Texas at Austin

Motivation



Porosity location map (left), histogram (center), cumulative distribution function (right).

Calculating and visualizing univariate statistical distributions is critical to data analytics for:

- Inference of populations
- Predictions and forecasts of future samples
- Representing uncertainty models
- Supporting decision making

PGE 337 Data Analytics and Geostatistics

Lecture 3: Displaying Distributions

Lecture outline . . .

- Plotting, Data Visualization

Introduction

General Concepts

Univariate

- **PDF / CDF**
- Statistics
- Distributions
- Heterogeneity
- Hypothesis

Bivariate

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

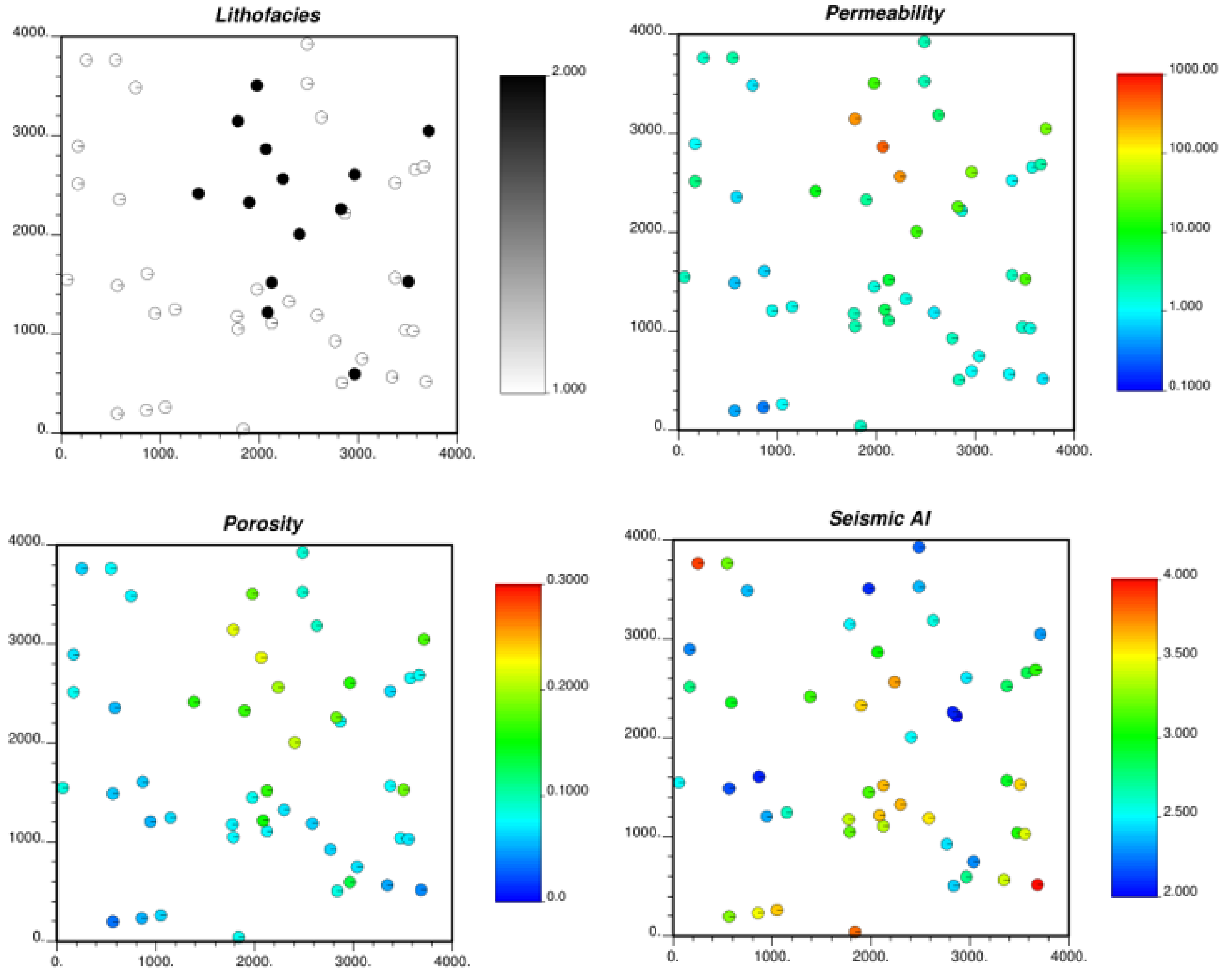
Michael Pyrcz, The University of Texas at Austin

Displaying Spatial Data

Let’s take a step back and consider spatial data visualization.

Raw data postings / location maps

- no model between the data
- straightforward for 2D spatial data sets, for univariate (1 feature at a time), visualization.



Spatial data with 4 features visualized as one location map for each feature.

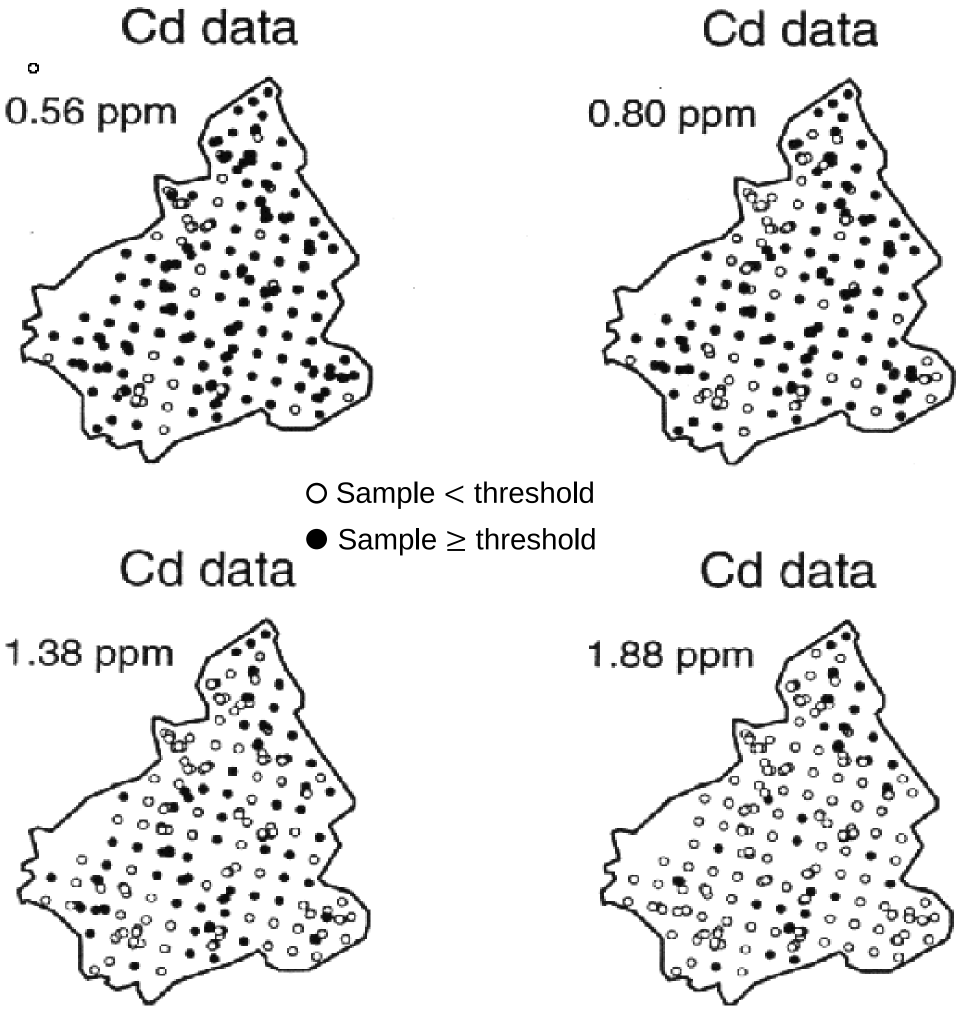
Displaying Spatial Data

Raw data postings / location maps

- may benefit from **feature engineering** to improve communication**.**

- example for contaminants with binary transform

- aside: this is an indicator transform of continuous data (covered during geostatistical facies simulation Lecture 15)



Location map of engineered feature from soil contamination data.
Image modified from Isaaks and Srivastava, 1989.

Feature Engineering Definition

The process of using domain knowledge to select, modify or make the most relevant features from the data for improved inferential and predictive models.

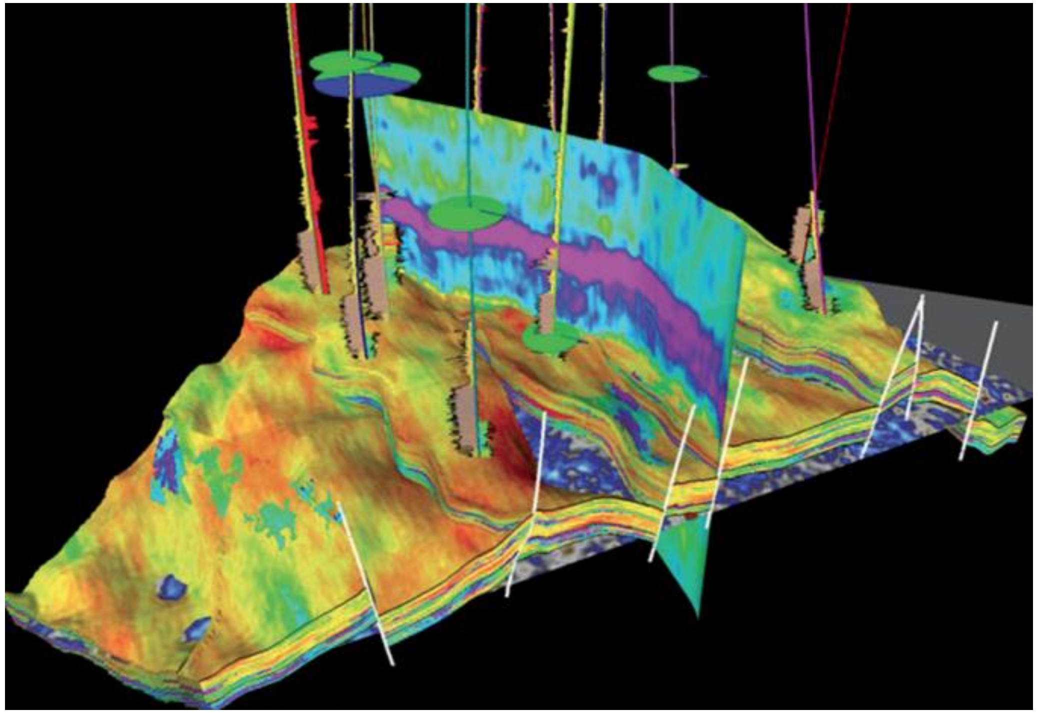
Here’s some examples of feature engineering:

- treatment of data outliers
- Gaussian transformation to support geostatistical modeling
- converting total porosity to effective porosity
- indicator transformations to highlight soil metal concentrations that exceed an environmental standard
- calculating a new feature as a combination of other features, e.g., rock quality index is permeability divided by porosity

Displaying Spatial Data

Raw data posting in 3D with multiple data types and multiple features

- may be difficult, but enabled with interactive 3D displays, superposition of color and shape, and query-able databases
- essential for data and model checking, interpretation, communication with the project team and managers.



Raw data posting in 3D is added with the ability to rotate and zoom, filter data, remove or transparency, color, geometries (well logs and faults), surfaces (contoured) and sections (projected on grid). From Petrel OffShore Engineer Sept 12, 2012.

Image from https://www.oedigital.com/news/459417-integrated-e-p-analysis.

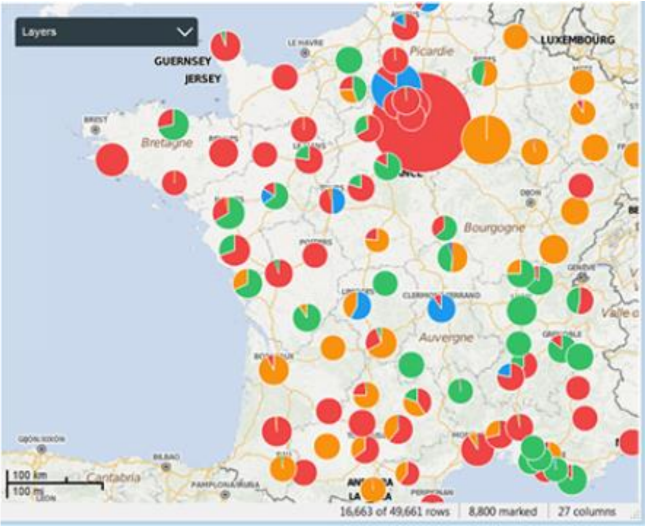
Displaying Spatial Data

Other Data Visualization Examples

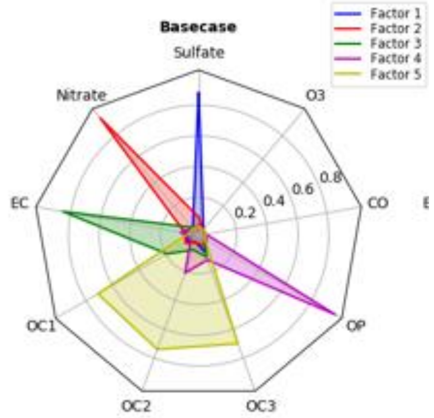
- Spotfire
  - Very flexible display of spatial data

- Python
  - Matplotlib, ggplot2 packages

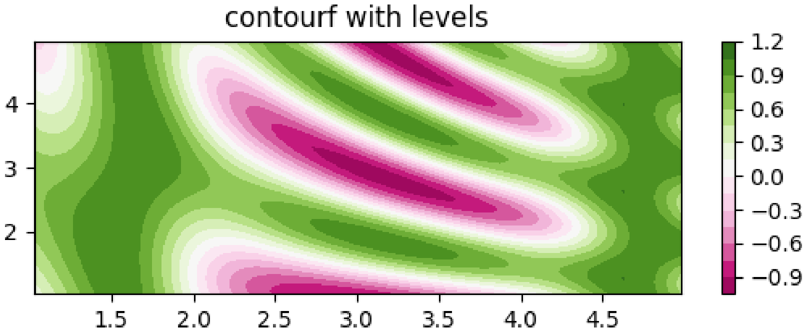
Powerful visualizations improve communication and impact of your work!



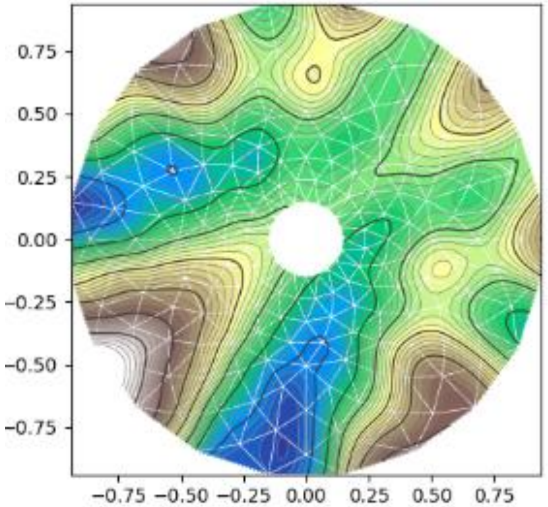
Spotfire Image (spotfire.tibco.com) circle size to size of the market and pie slices for cellular provider share.



Radar plots of machine learning factor loadings (e.g., dimensional reduction).



Spatial contouring (e.g., reservoir trends).



Offset contouring (e.g., variogram map).

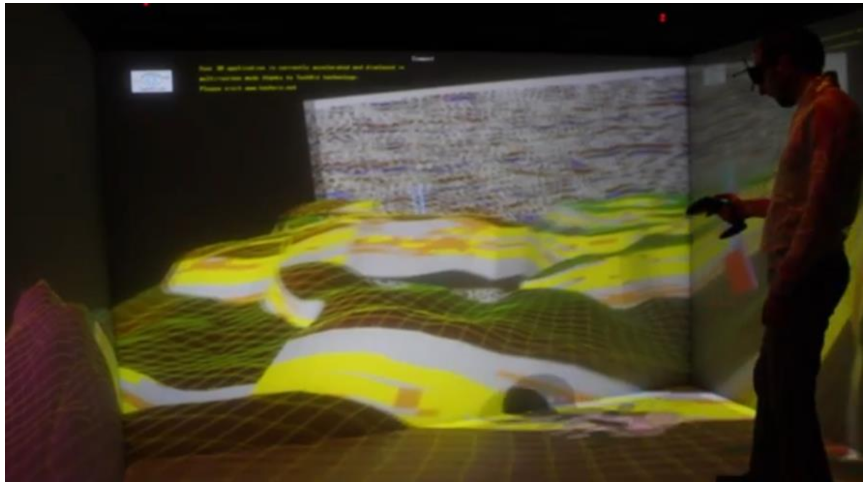
Displaying Data Advanced

Immersive Visualization Virtual / Augmented Reality

- fly around the subsurface, interpret data and check the model
- real-time editing of models and visualization of fluid flow changes!

Virtual Reality is Widely Available in Consumer Markets Now

- I recently walked the White Cliffs of Dover Chalk (SE coast of England on Strait of Dover) in VR
- I own a home in the City of Solitude in Skyrim, too!



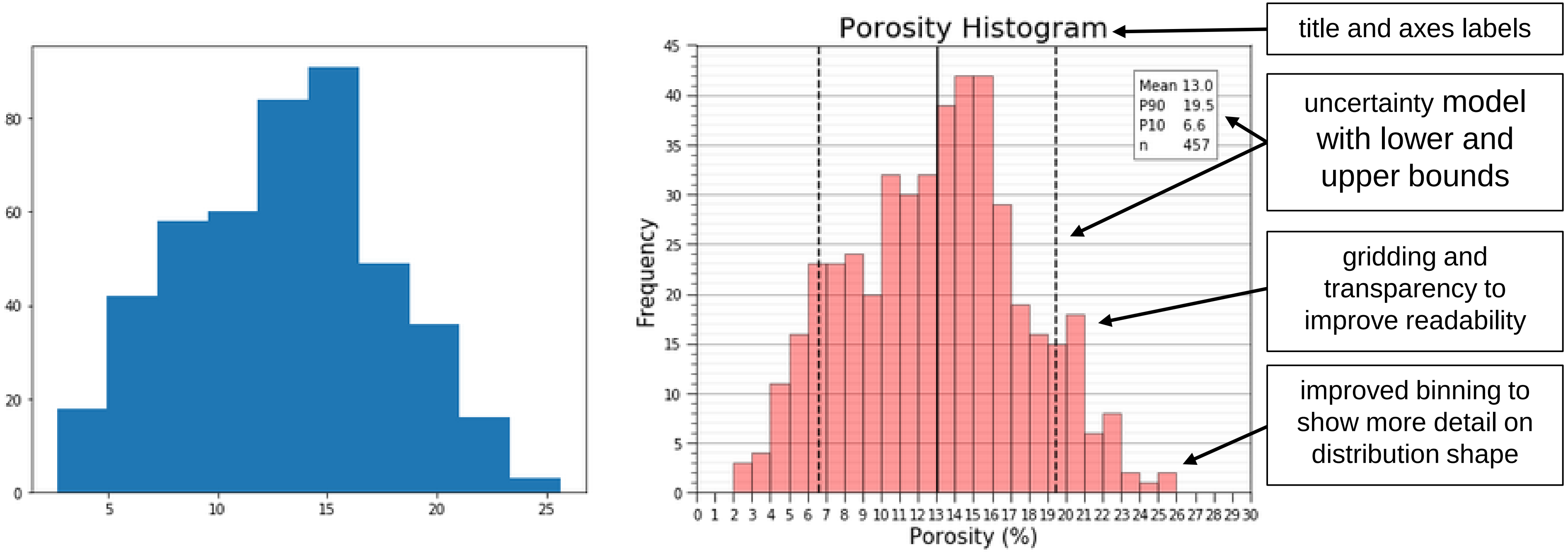
Immersive visualization by TechViz and Schlumberger Petrel.

VR headset, HTC Vive Pro 2, promotional documentation.

Image from (upper) https://www.youtube.com/watch?v=BIYK8prfviU and (lower) https://thegadgetflow.com/portfolio/htc-vive-pro2-vr-headset-120hz-refresh-rate-impressive-5kresolution/

Effective Data Visualization

Design your data visualization to maximize communication. Here’s a histogram, default and resigned, more effective, data visualization.



title and axes labels

uncertainty model with lower and upper bounds

gridding and transparency to improve readability

improved binning to show more detail on distribution shape

Demonstration workflow for univariate data visualization in Python. File is PythonDataBasics_Univariate_Visualization.ipynb

Aspects of effective data visualization includes:

- design the plot space, e.g., label axes, ranges, grids
- compose the plot elements, e.g., bin widths, color, outlines, transparency
- draw attention to critical information, e.g., custom lines and legends

PGE 337 Data Analytics and Geostatistics

Lecture 3: Displaying Distributions

Lecture outline . . .

- Histograms, Probability Density Functions

Introduction

General Concepts

Univariate

- **PDF / CDF**
- Statistics
- Distributions
- Heterogeneity
- Hypothesis

Bivariate

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis

Univariate Definition

Univariate

- involving one variable (feature)

Univariate Statistics

- summary measures based on one feature measured over the samples

Univariate Parameters

- summary measures inferred for one feature measured over the population

We start with univariate, but we will cover bivariate, involving two variables (features) later.

Recall, we mentioned joint probabilities and distributions, they are:

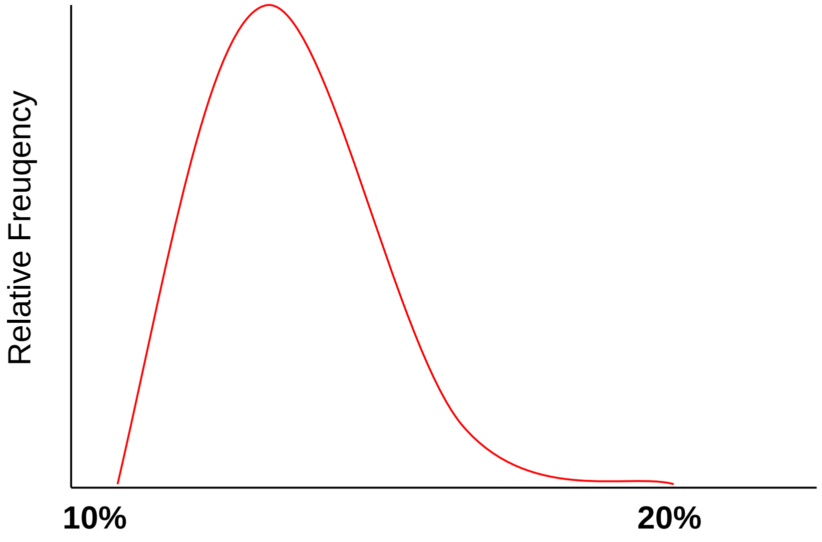
- bivariate (2 features)
- multivariate (general term for >1 features, but often refers to 3 or more).
- massively multivariate, high dimensional (several or more features)

Statistical Distribution

Definition of Statistical Distribution – for a variable / feature a description of the probability of occurrence over the range of possible values.

What do we get from a statistical distribution?

- what is the minimum and maximum?
- do we have a lot of low values?
- do we have a lot of high values?
- do we have outliers (values that don't make sense and need explaining)?



Univariate Statistical Distribution

We will use the term 'distribution' as shorthand.

Definition from WolframMathWorld.

Statistical Distributions Histograms

Histogram is the plot of frequency over an exhaustive set of bins.

Step 1: Divide the continuous feature range of possible values into K equal size bins, Δx:

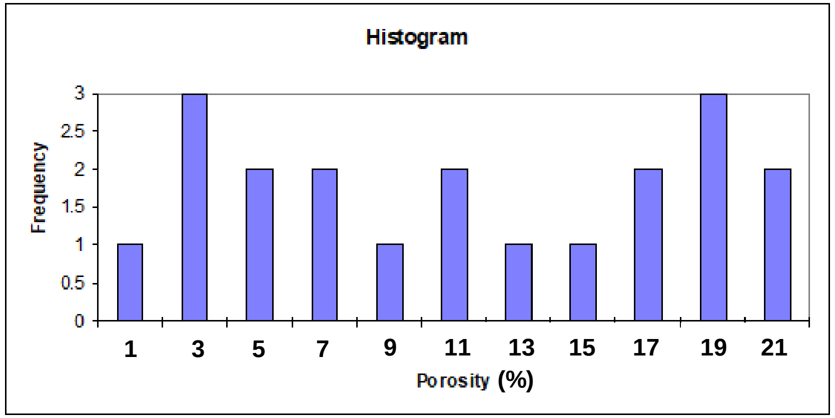
$$\Delta x = \left(\frac{x_{max} - x_{min}}{K}\right)$$

or use available categories .

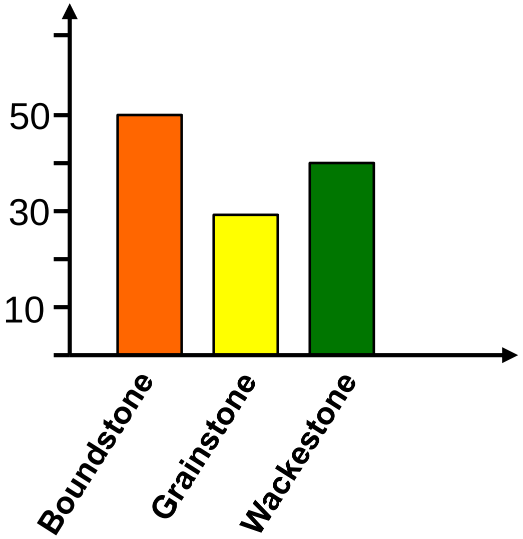
Step 2: Count the number of samples (frequency) in each bin, $n_k\ , \forall\ k = 1, \ldots, K$.

Step 3: Plot the frequency vs. the bin label (use bin centroid if continuous)

Note, typically plotted as a bar chart



Continuous histogram of a porosity data from a set of cores with bin size, $\Delta x = 2\%$, 11 bins are $[0{,}2\%]$, $[2\%, 4\%]$,…,$[20\%, 22\%]$.



Categorical histogram of a facies data from a set of cores.

Statistical Distributions
Normalized Histograms

Normalized Histogram

- the histogram is normalized so frequencies are replaced with the probability that the outcome existing within each bin, k.

$$p_k = \frac{n_k}{n}, \qquad \forall\, k = 1, \ldots, K.$$

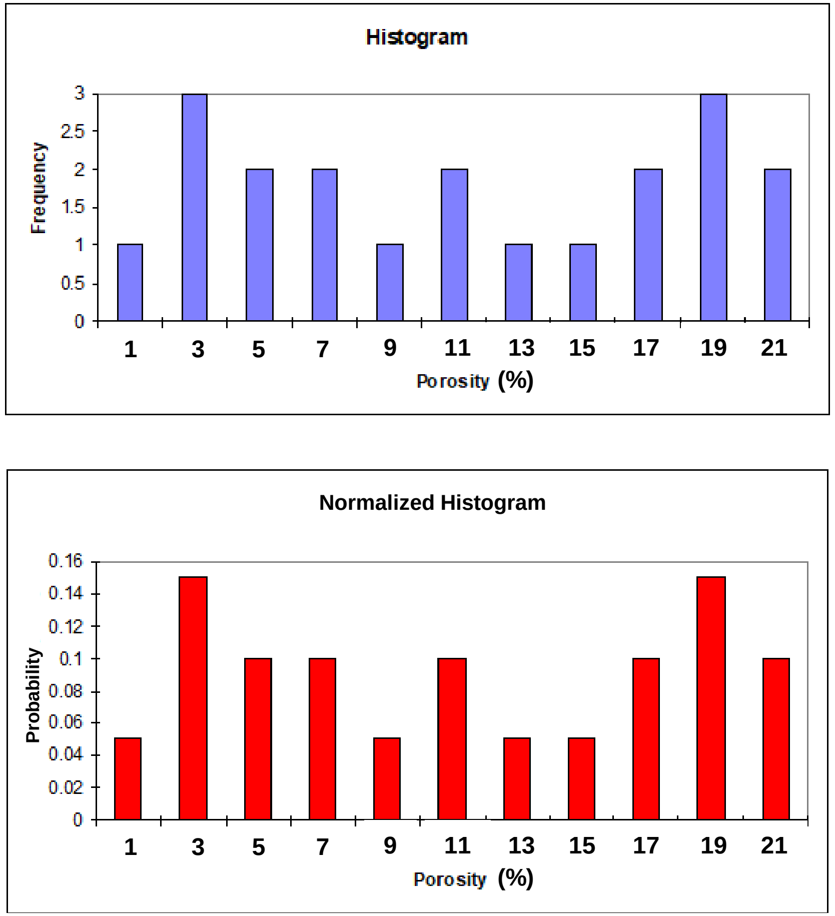
- now for each bin we have probability:

$$0.0 \ \leq p_k \leq 1.0, \ \ \forall\, k = 1, \ldots, K$$

- closure, sum of all bins is one:

$$\sum_{k=1}^{K} p_k = 1$$

Normalized histogram is convenient because we can read probability from the plot.



Histogram and normalized histogram of a porosity data from a set of cores.

Statistical Distributions
Probability Density Function

How to make a normalized histogram?

Step 1: Divide data range (x_{max}- $xmin$) into desired number bins / classes / categories, K, for continuous features:

$$\Delta x = \left(\frac{x_{max} - x_{min}}{K}\right)$$

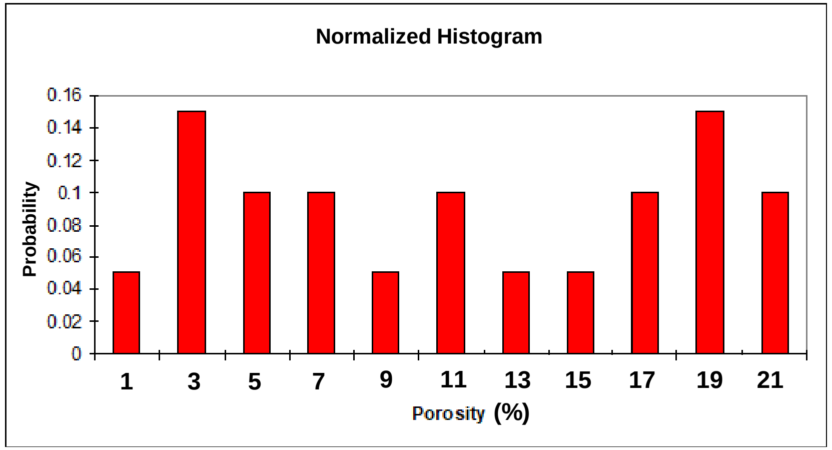
Step 2: Count the number of data in bin, n_k and then compute the probability:

$$p_k = \left(\frac{n_k}{n}\right)$$

were n is the total number of data.

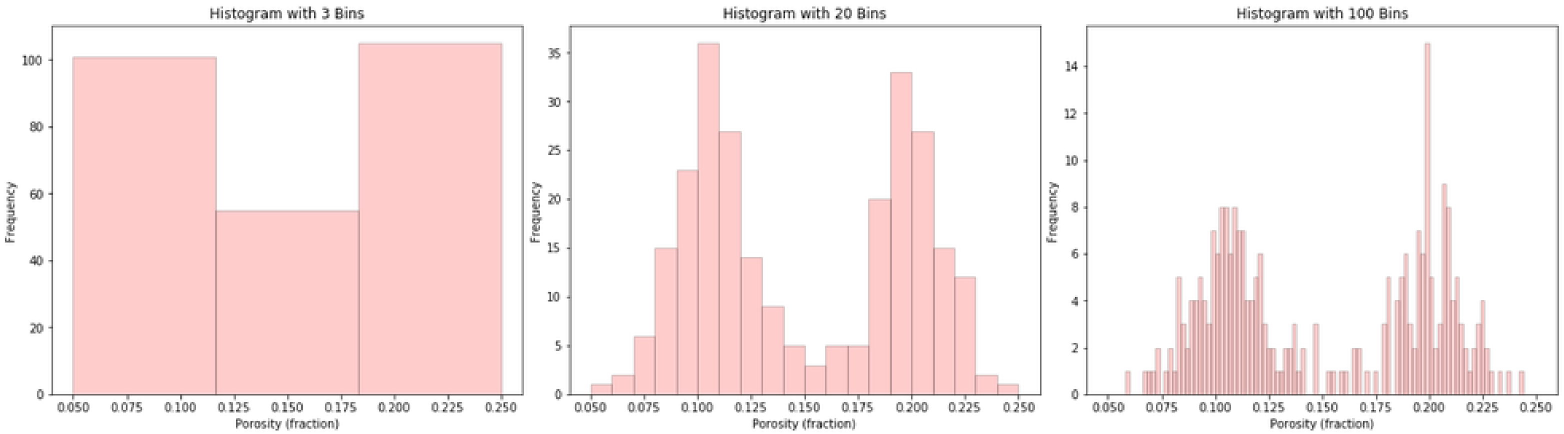
Step 3: Plot bin probability versus mid-range $(x_{k_{,}\ min} + \frac{\Delta x}{2})$ if continuous or categorical label.

Make a histogram and converted it to a normalized histogram, by dividing frequency by total number of data.



Normalized histogram of a porosity data from a set of cores.

Statistical Distributions
Histogram Bin Size



Histograms with different number of bins for the same 261 well log porosity samples form GeostatsPy_datadistributions.ipynb.

Impact of bin size?

- **too large bins / too few bins** often smooth out, mask information
  - lack resolution
- **too small bins / too many bins** are too noisy
  - lack samples in each bin for stable assessment of frequency

choose the highest resolution with lowest possible noise.

Note: very large and very small bins will tend towards equal proportion in each bin (all samples in a single bin or one sample in each bin).

- this will appear as a **uniform distribution** (more on parametric distributions later).

Statistical Distributions
Probability Density Function

A function, $f(x)$, of probability density across the range of all possible feature values, x.

- non-negativity constraint:

$$0.0 \leq f(x)$$

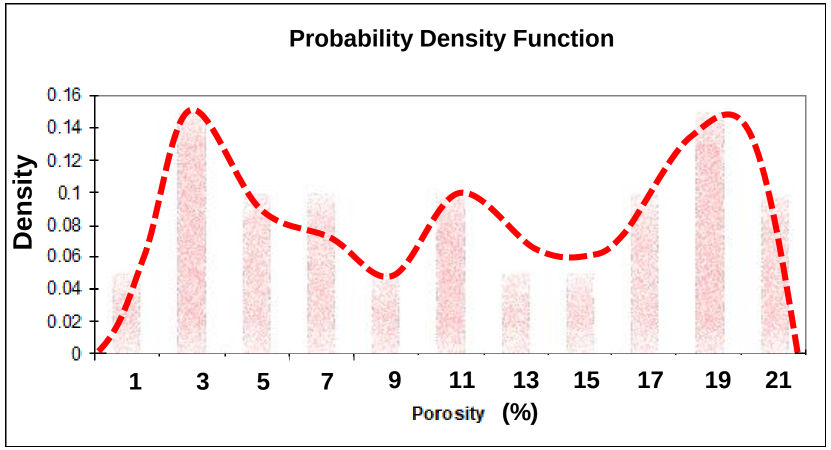
- for continuous measure of density (may be > 1.0)

- integrate to calculate probability

$$0 \leq \int_a^b f(x)dx = P(a \leq x \leq b) \leq 1.0$$

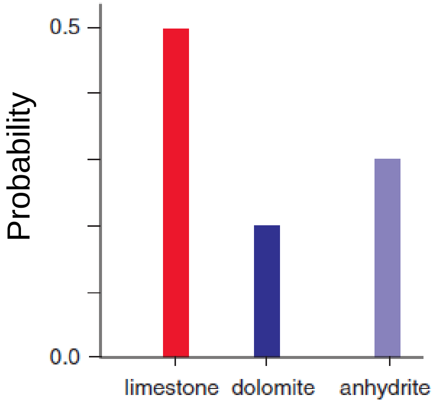
- closure:

$$\int_{-\infty}^{+\infty} f(x)dx = 1$$



Porosity probability density function from a set of well cores.

For categorical features, the normalized histogram is the PDF.



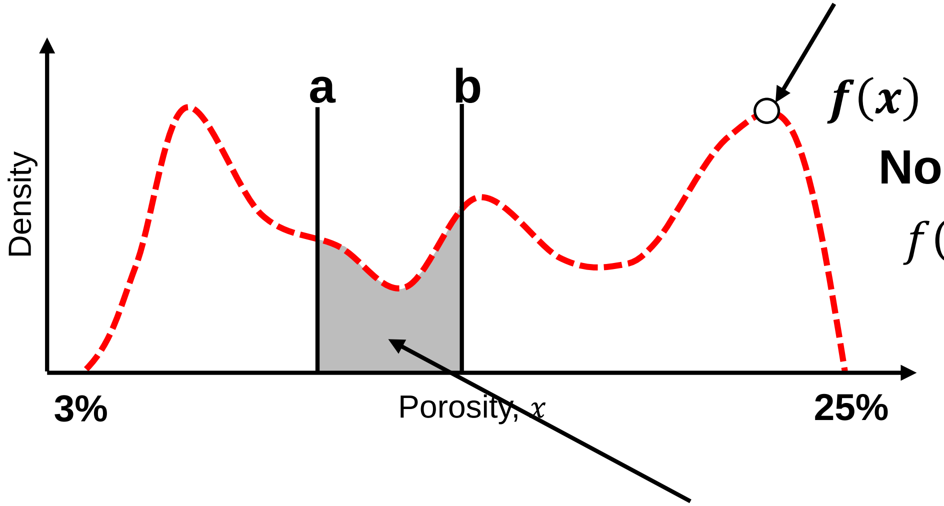
Categorical facies normalized histogram or probability density function.

Statistical Distributions
Probability Density Function

Working with the density measure from a probability density function, $f(x)$.

1. Closure

$$\int_{-\infty}^{+\infty} f(x)dx = 1.0$$



2. Density

Non-negativity Constraint

$$f(x) \geq 0$$

3. Probability

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

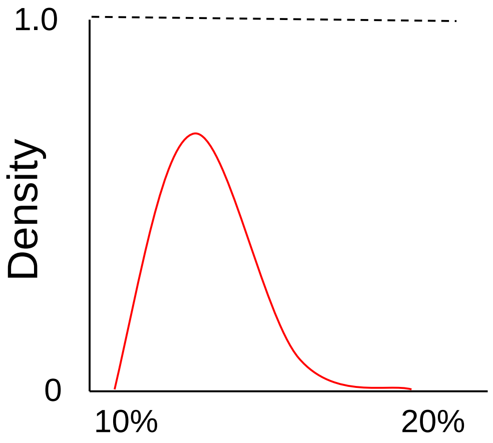
1. Closure – area under the curve is = 1.0

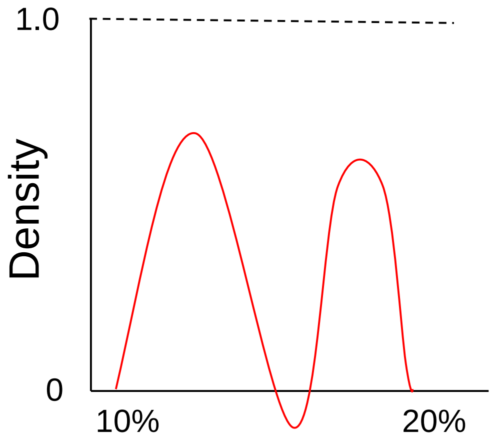
2. $f(x) =$ density, a measure of relative likelihood, may be > 1.0!

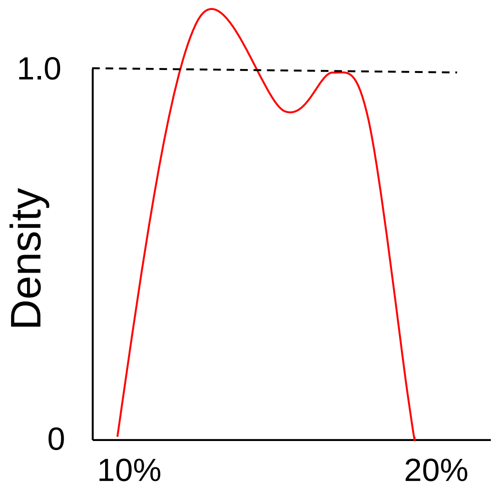
3. Probability is calculated by integrating over an interval.

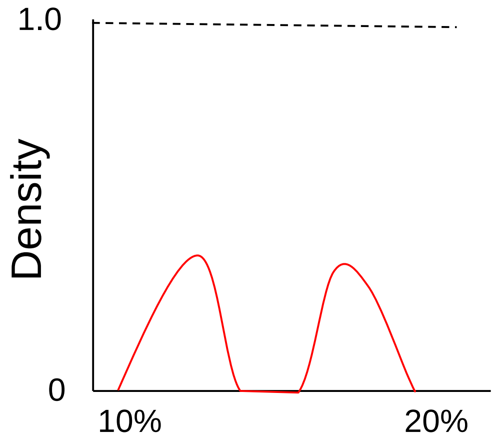
Statistical Distributions
PDF Exercise

Identify the valid PDFs



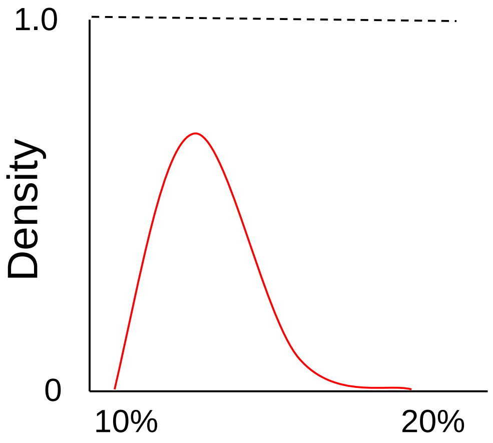




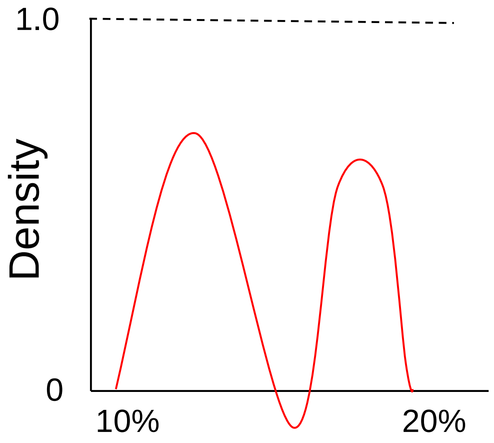


Statistical Distributions
PDF Exercise

Identify the valid PDFs



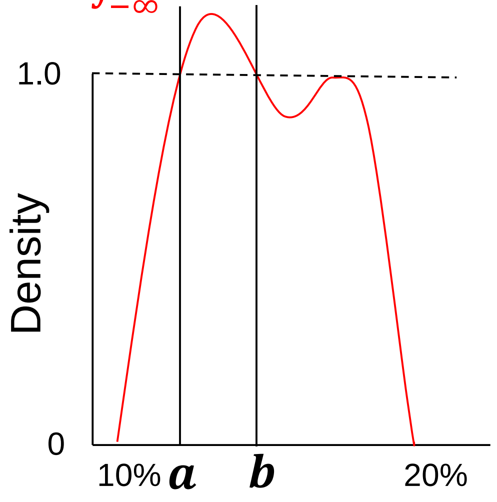
check if $\int_{-\infty}^{+\infty} f(x)dx = 1.0$

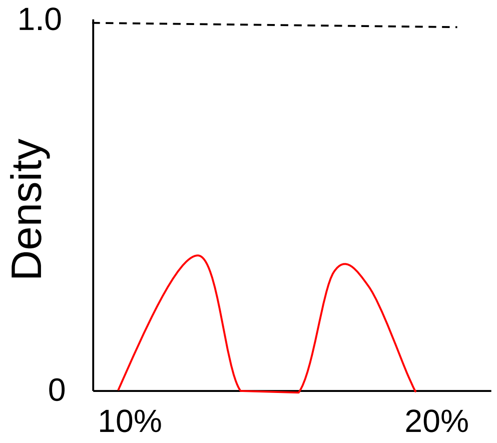


no - negative prob

check if $P(a \le x \le b) \le 1.0$

and $\int_{-\infty}^{+\infty} f(x)dx = 1.0$





check if $\int_{-\infty}^{+\infty} f(x)dx = 1.0$

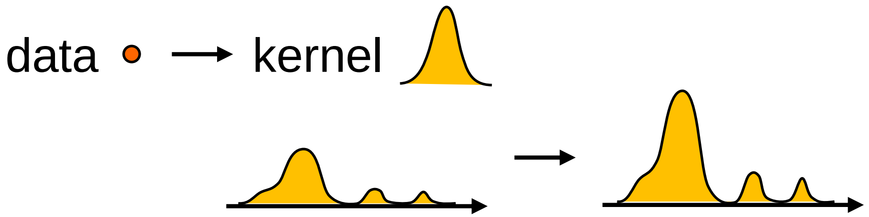
Calculating PDFs from Data

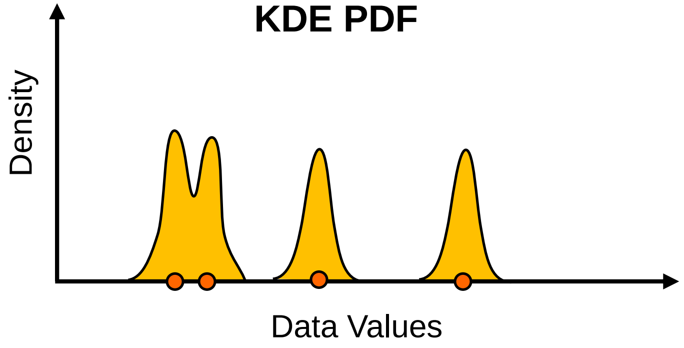
For parametric cases the PDF’s equation is known, but for the nonparametric case, the PDF is calculated from the data (more on parametric and nonparametric distributions in the next unit).

- While a data-derived, nonparametric PDF could be calculated by differentiating a data-derived CDF (discussed next), generally this would be too noisy!
- The common method to calculate a data-derived PDF is to fit a smooth model to the data.

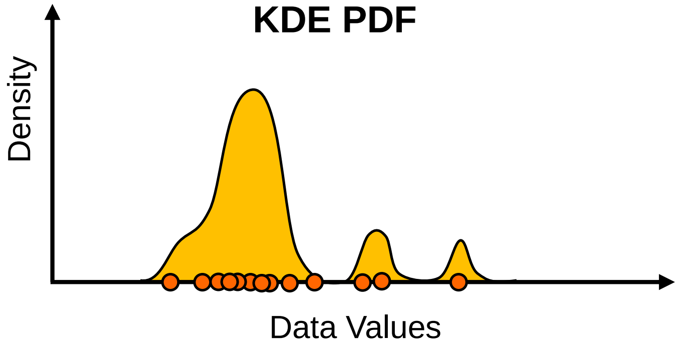
Kernel Density Estimation (KDE) Approach, fit smooth PDF to data:

- Replace all data with a kernel, Gaussian is typical.
- Standardize result to ensure closure, $\int_{-\infty}^{+\infty} f(x)dx = 1$





Schematic of KDE PDF with 4 data.

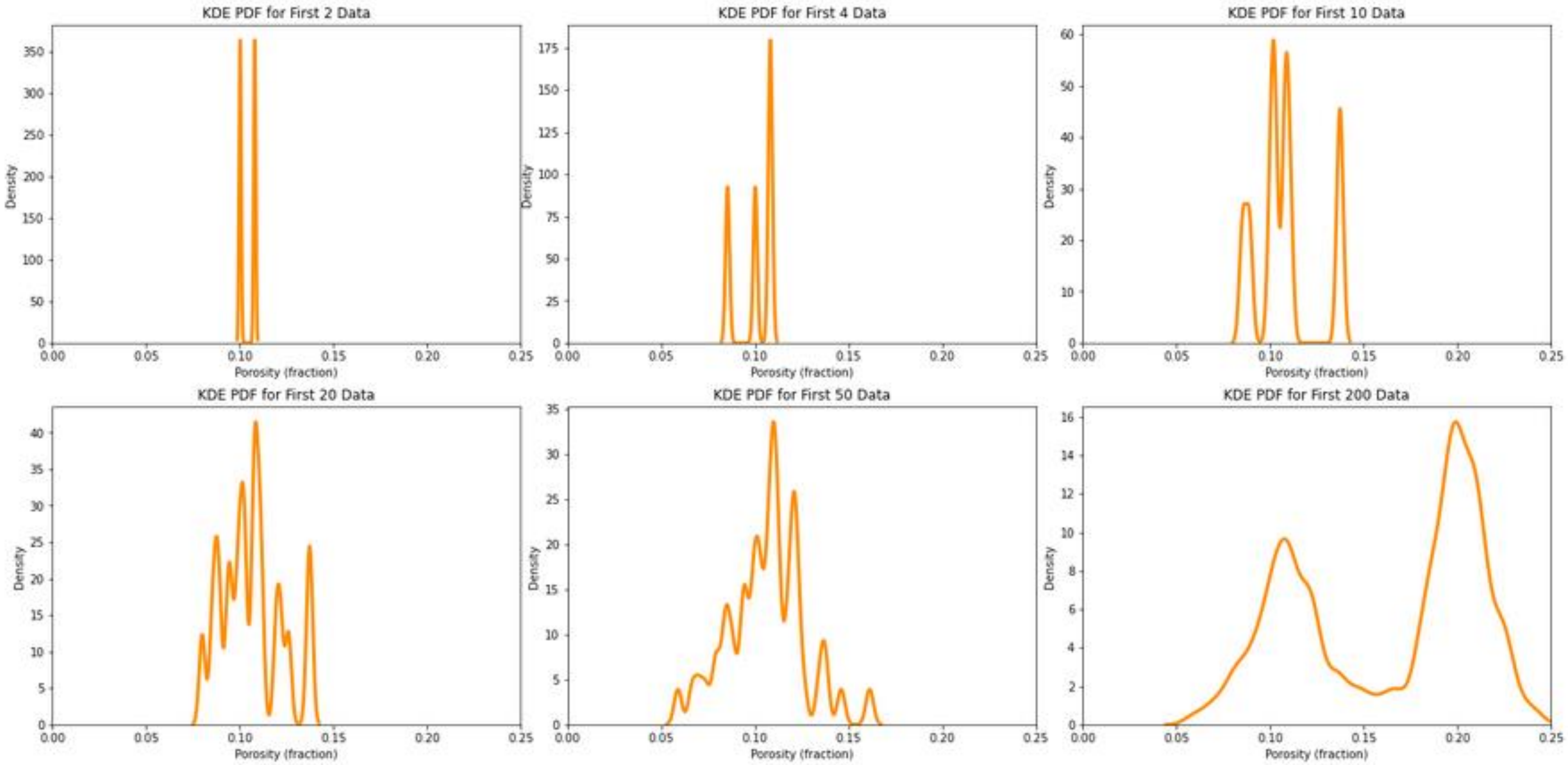


Schematic of KDE PDF with 16 data.

Calculating PDFs from Data

Here’s the resulting PDFs with the KDE method with a Gaussian kernel for the first 2 data, 4 data, 10 data, 20 data, 50 data and 200 data to illustrate the addition of Gaussian kernels and standardization, $\int_{-\infty}^{+\infty} f(x)dx = 1$.

- See the reduction in Y axis limits as more data are added.



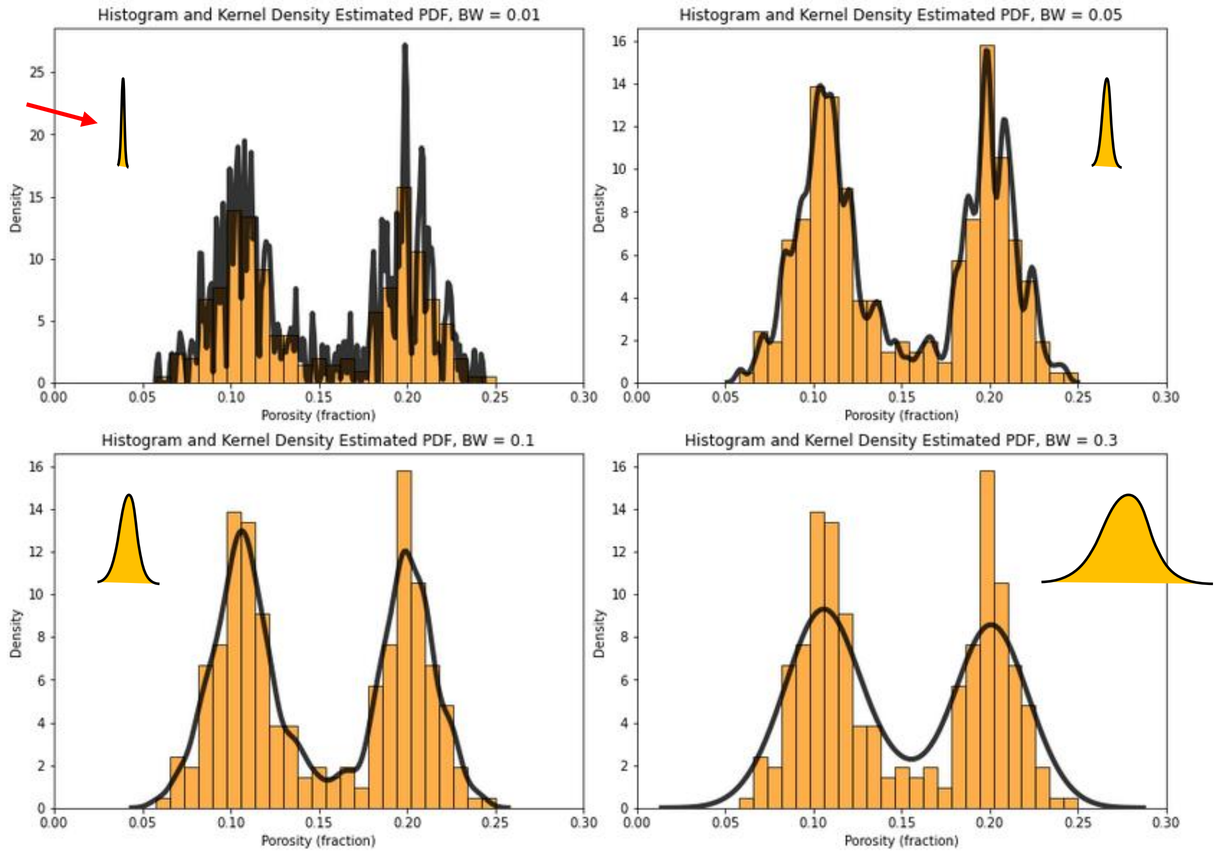
KDE-based PDF for the first 2, 4, 10, 20, 50, and 200 data from dataset sample_data.csv. File is GeostatsPy_datadistributions.ipynb.

Calculating PDFs from Data

What is the impact of changing the kernel width?

- analogous to changing the histogram bin size, attempt to smooth out noise while not removing information

Approximate Kernel



Histogram and Kernel Density Estimated PDF, BW = 0.01

Histogram and Kernel Density Estimated PDF, BW = 0.05

Histogram and Kernel Density Estimated PDF, BW = 0.1

Histogram and Kernel Density Estimated PDF, BW = 0.3

Note, bandwidth is not well documented in Seaborn but seems to be scaled by the original data standard deviation.

E.g.:

$$\sigma_{kernel} = bw \cdot \sigma_x$$

Histogram and KDE PDF for Gaussian kernel with bandwidth of 0.01, 0.1, 1.0 and 10.0 with dataset sample_data.csv, file is GeostatsPy_datadistributions.ipynb.

PGE 337 Data Analytics and Geostatistics

Lecture 3: Displaying Distributions

Lecture outline . . .

- Cumulative Distribution Functions

Michael Pyrcz, The University of Texas at Austin

Statistical Distributions
Cumulative Distribution Function

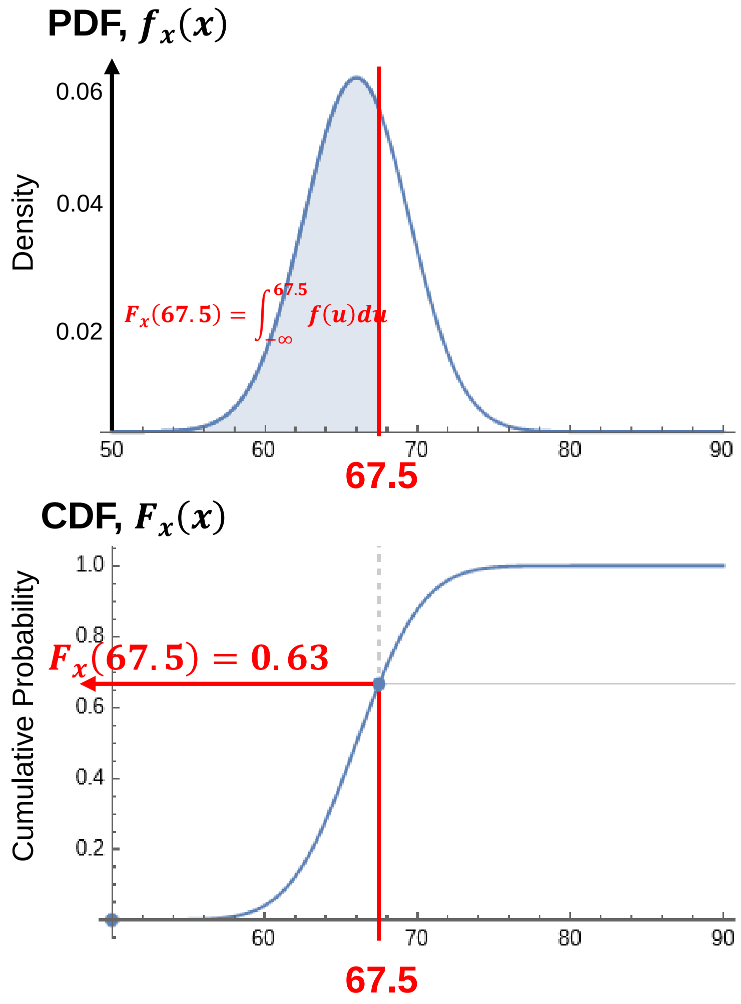
Cumulative Distribution Function:

The cumulative distribution function (CDF) is the sum of a discrete PDF or the integral of a continuous PDF.

- the cumulative distribution function $F_x(x)$ is the probability that a random sample, X, is less than or equal to a value x.

$$F_x(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

- for CDF there is no bin assumption; therefore, bins are at the resolution of the data.
- monotonically non-decreasing function, because a negative slope would indicate negative probability over an interval.



PDF and CDF for the same data illustrating going from PDF to CDF.

Image modified from https://demonstrations.wolfram.com/ConnectingTheCDFAndThePDF/

Statistical Distributions
Cumulative Distribution Function

Cumulative Distribution Function:

To check for a valid CDF.

- non-negativity constraint:

$$F_x(x) = P(X \leq x) \geq 0.0, \forall\, x$$

- valid probability:

$$0.0 \leq F_x\,(x) \leq 1.0, \forall\, x$$

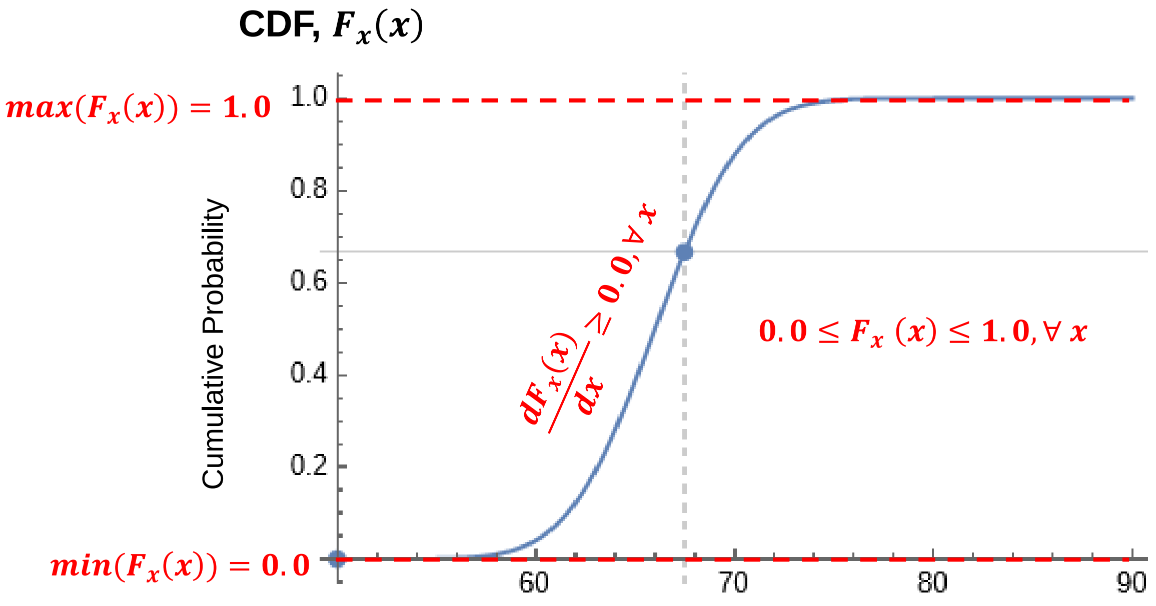
and cannot have negative slope $\dfrac{dF_x(x)}{dx} \geq 0.0, \forall\, x$

- minimum and maximum (closure) values:

$$min(F_x(x)) = 0.0 \qquad max(F_x(x)) = 1.0$$

since the CDF does not have a negative slope we can use limits:

$$\lim_{x \to -\infty} F_x(x) \to 0.0 \qquad \lim_{x \to \infty} F_x(x) \to 1.0$$



CDF for the same data illustrating going from PDF to CDF.

Image modified from https://demonstrations.wolfram.com/ConnectingTheCDFAndThePDF/

Statistical Distributions
Cumulative Distribution Function

Cumulative Distribution Function:

To further demonstrate the calculation of the CDF let's consider a normalized histogram and calculate the CDF at a value at the edge of a bin

- we sum the frequencies for all bins to the left (including values ≤ the bin edge value and divide by the number of data.

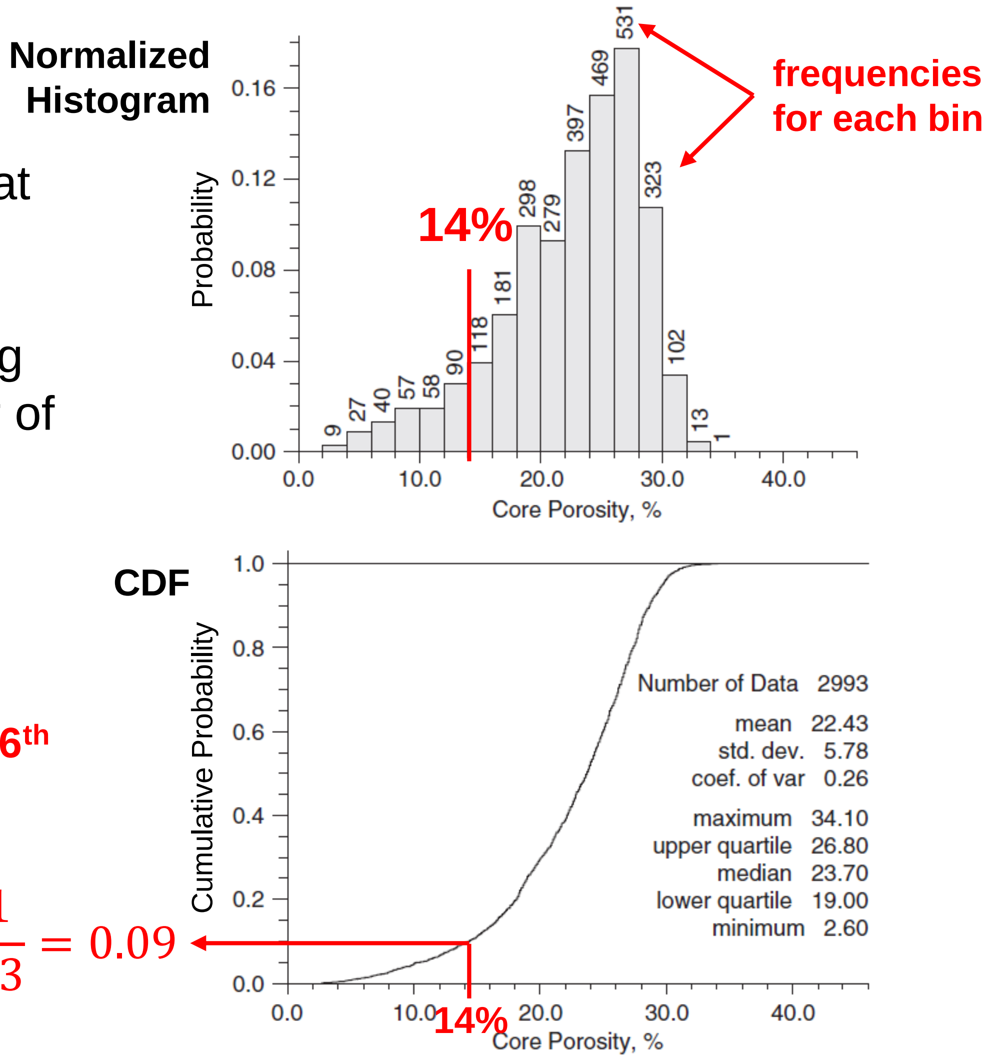
$$\boldsymbol{F_x(x) = P(X \le x) = \int_{-\infty}^{x} f(u)du}$$

For example, if we select 14%, the upper boundary of the 6th histogram bin:

$$F_x(14\%) = \frac{9+27+40+57+58+90}{n} = \frac{281}{2993} = 0.09$$

$$F_x(x) = \frac{n_{\le x}}{n}$$

The CDF of x is the proportion of data equal to or less than x.



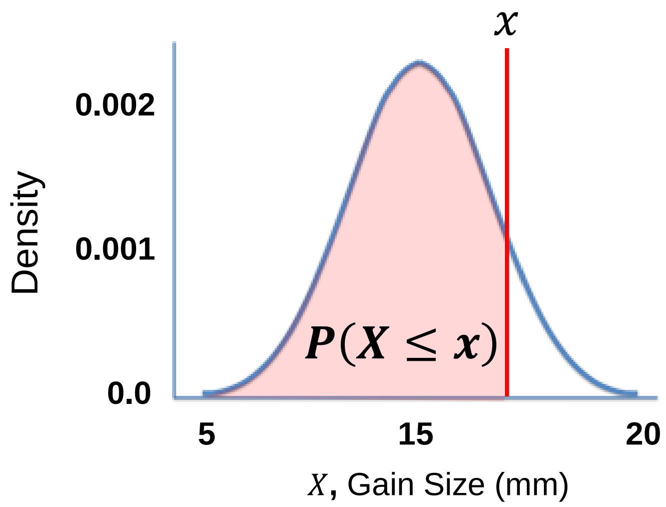
Porosity normalized histogram (with annotated frequencies) and CDF.

Image from Pyrcz and Deutsch (2014), Geostatistical Reservoir Modeling.

Random Variable (RV) Definition

Random Variable (we need this to understand CDF notation)

- we do not know the value at a location / time, it can take on a range of possible values, fully described with a statistical distribution PDF / CDF.
- represented as an upper-case variable, e.g., $\boldsymbol{X}$, while possible outcomes or data measures are represented with lower case, e.g., $\boldsymbol{x}$.
- more latter on this!



PDF with cumulative probability indicated.

Random Variable **A Specific Value or Outcome**

Cumulative Distribution Function $\boldsymbol{F_x(x) = P(X \le x)}$

Probability of any randomly drawn value, $\boldsymbol{X}$, being less than or equal to a specific value, $\boldsymbol{x}$.

Statistical Distributions
Cumulative Distribution Function

How to use a CDF?

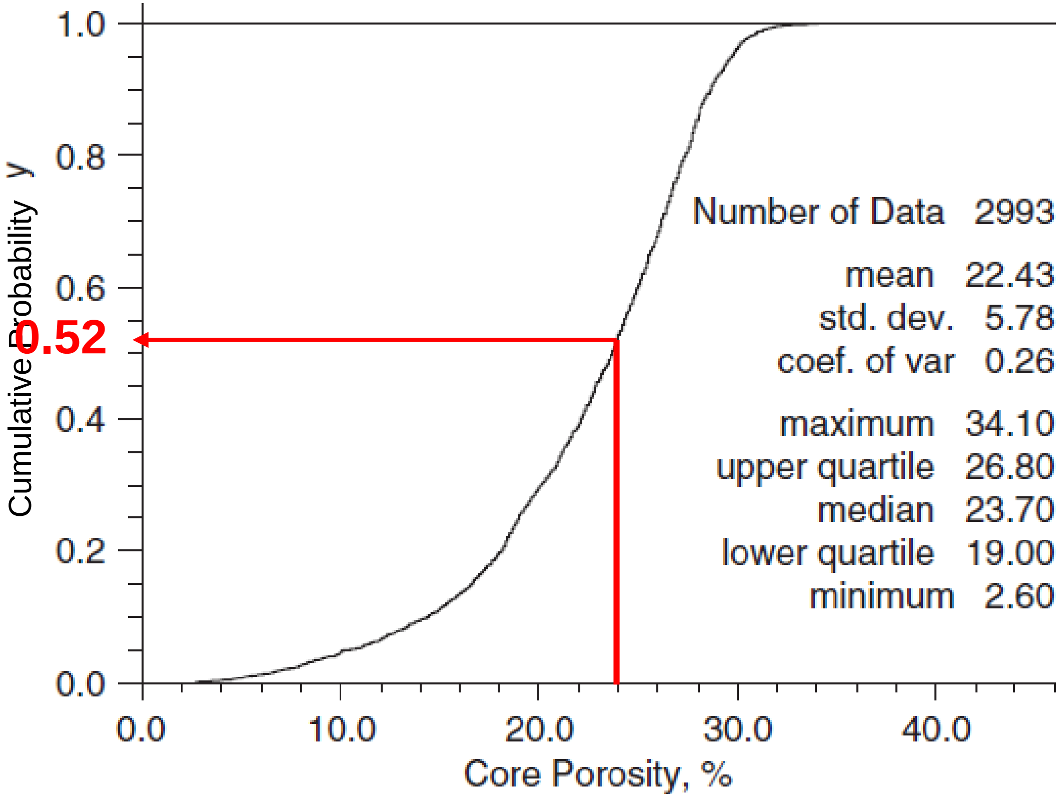
$$F_x(x) = P(X \leq x)$$

Probability of any random sample, X, being less than a threshold, x.

Example:

What is the probability of a random core porosity being less than or equal to 24%?

$$F_x(24\%) \approx 0.52$$



Porosity CDF with graph-based calculation cumulative probability , $F_x(24\%)$ illustrated.

Image from Pyrcz and Deutsch (2014), Geostatistical Reservoir Modeling.

Statistical Distributions
Cumulative Distribution Function

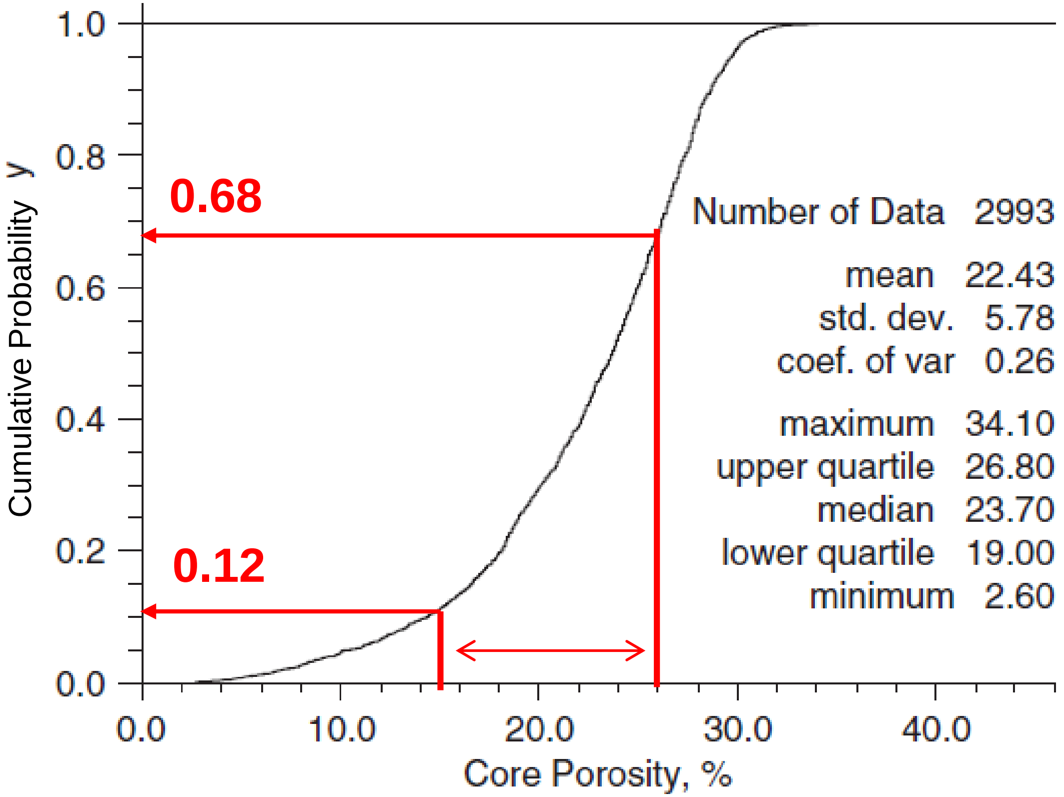
How to use a CDF?

$$P(a \leq X \ \leq b) = F_x(b) - F_x(a)$$

Probability for a random sample, X, to be between a and b.

Another Example:

What is the probability of a random core porosity being greater than 15% and less than or equal to 26%?



Porosity CDF with graph-based calculation interval probability , $F_x(24\%)$ illustrated.

$$P(15\% \leq X \ \leq 26\%) = F_x(26\%) - F_x(15\%) \ \approx 0.68 \ - 0.12 \ \approx 0.56$$

Image from Pyrcz and Deutsch (2014), Geostatistical Reservoir Modeling.

Statistical Distributions
Cumulative Distribution Function

How to use a CDF?

$$\boldsymbol{F_x^{-1}(P) = x_P}$$

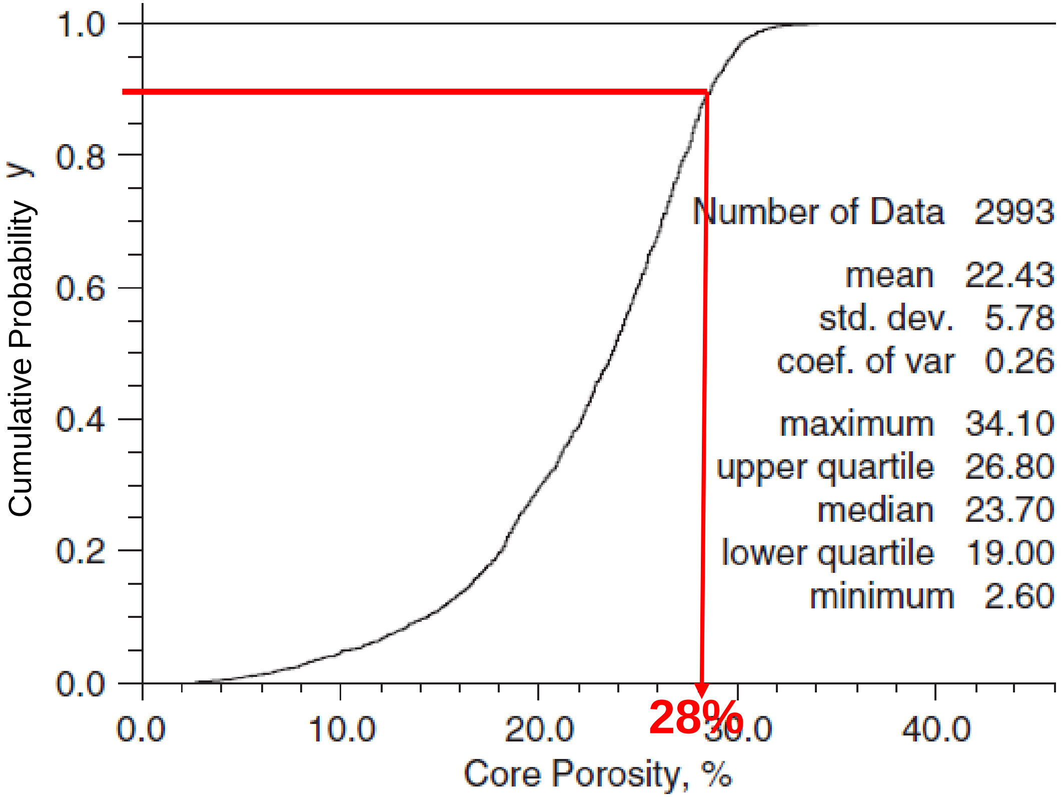
Inverse of a CDF, calculate the percentile value from the cumulative probability.

Example:

What is the 90th percentile (the value for which 90% of random values would be less than it)?

$$P90_x = x_{90} = F_x^{-1}(0.9) \; = \; 28\%$$

Note, F_x^{-1} is the inverse of the CDF, given a cumulative probability find the specific percentile value.



Porosity CDF with graph-based inverse CDF, percentile calculation , $F_x^{-1}(0.9)$ illustrated.

Image from Pyrcz and Deutsch (2014), Geostatistical Reservoir Modeling.

Statistical Distributions
CDF Exercise

Example:

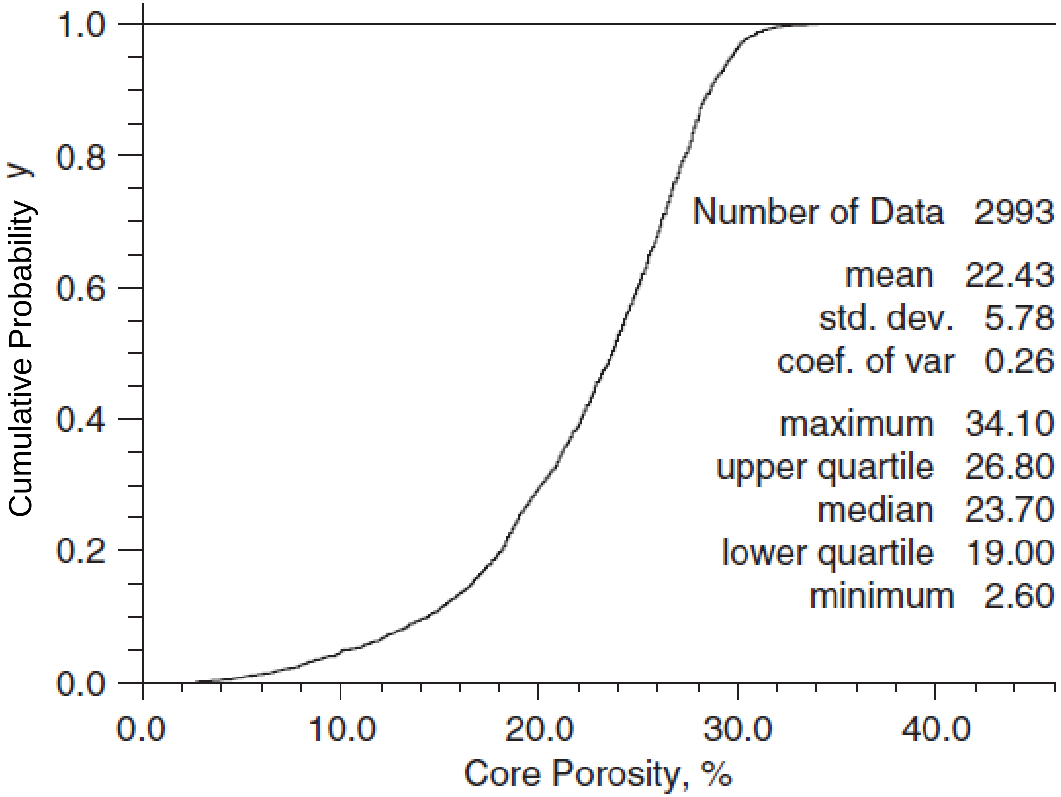
P(Porosity < 20%)?

P(Porosity < 26%)?

P(20% < Porosity < 26%)?

P50?

P20?



Porosity CDF.

Make a rough estimate with the graphical approach.

Image from Pyrcz and Deutsch (2014), Geostatistical Reservoir Modeling.

Statistical Distributions
CDF Exercise

Example:

P(Porosity < 20%) ≈ 0.30

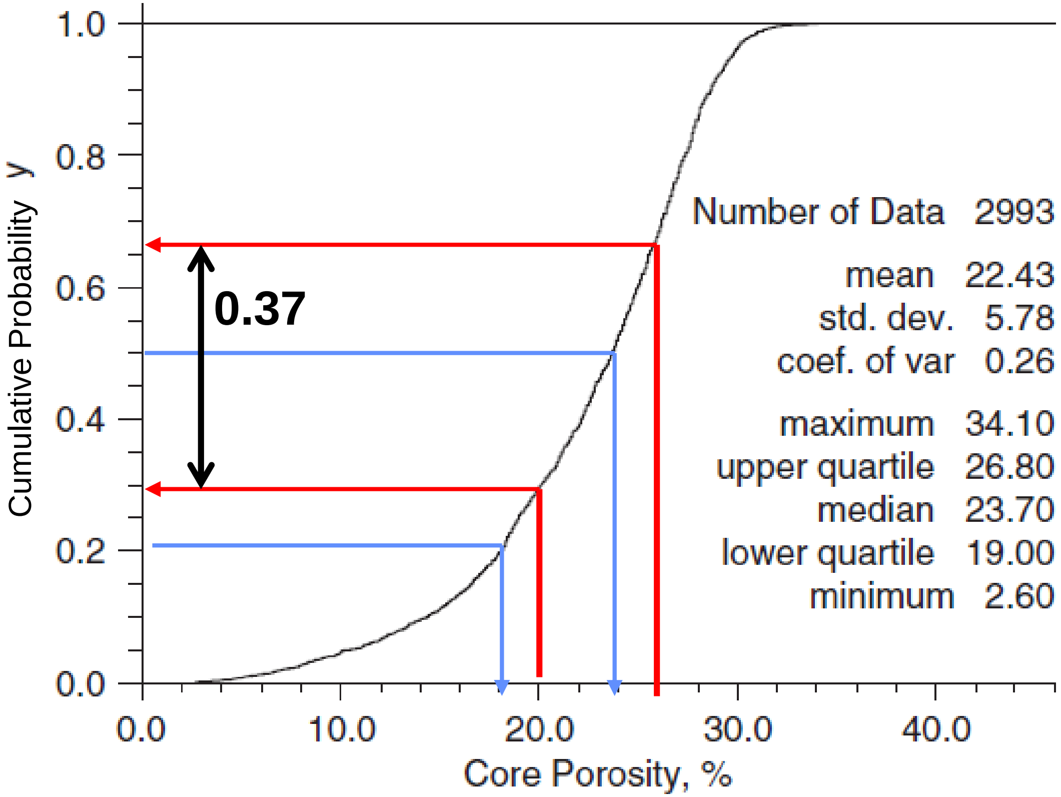
P(Porosity < 26%) ≈ 0.67

P(20% < Porosity < 26%)?

≈ 0.67 – 0.30 ≈ 0.37

P50 ≈ 0.24

P20 ≈ 0.18



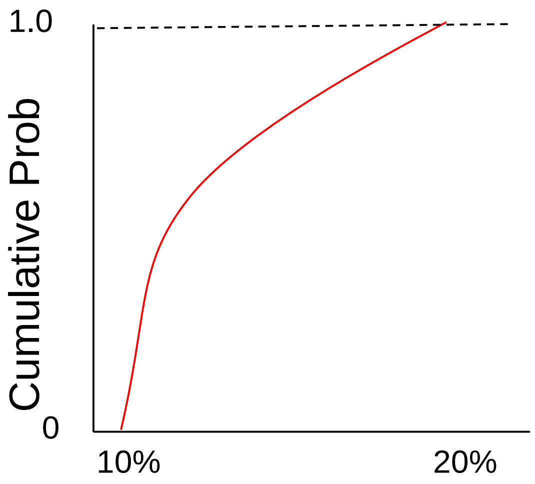
Porosity CDF.

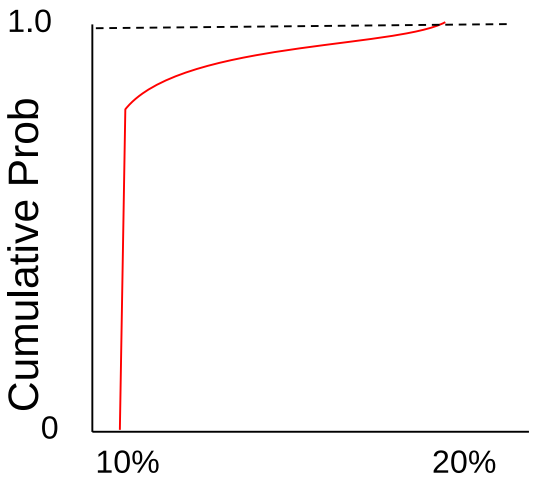
Make a rough estimate with the graphical approach.

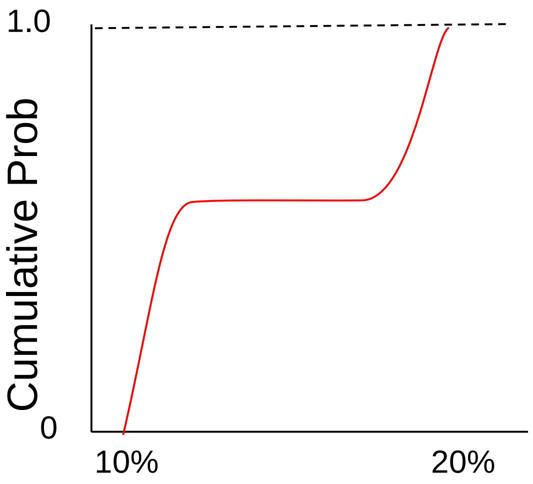
Image from Pyrcz and Deutsch (2014), Geostatistical Reservoir Modeling.

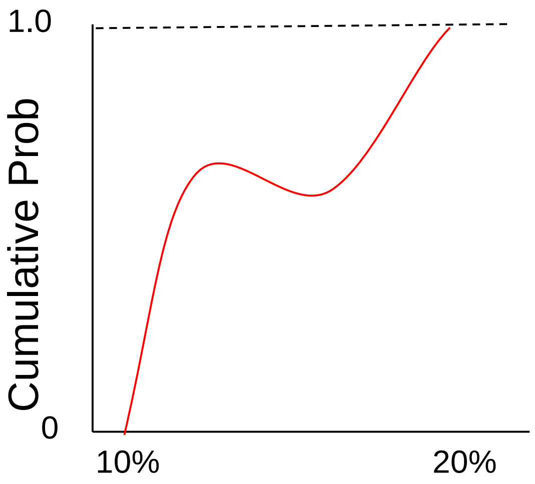
Statistical Distributions
CDF Exercise

Identify valid CDF?





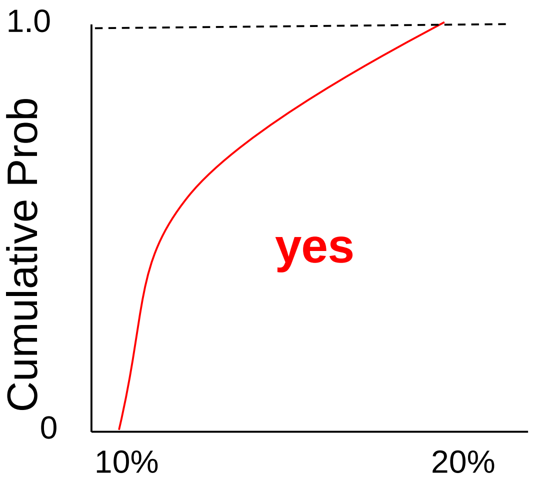




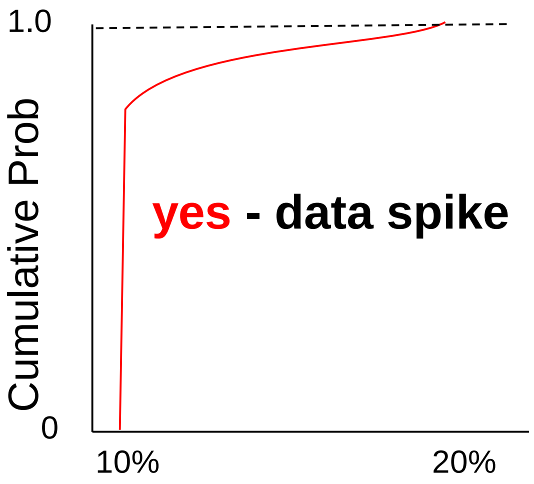
Identify any issues.

Statistical Distributions
CDF Exercise

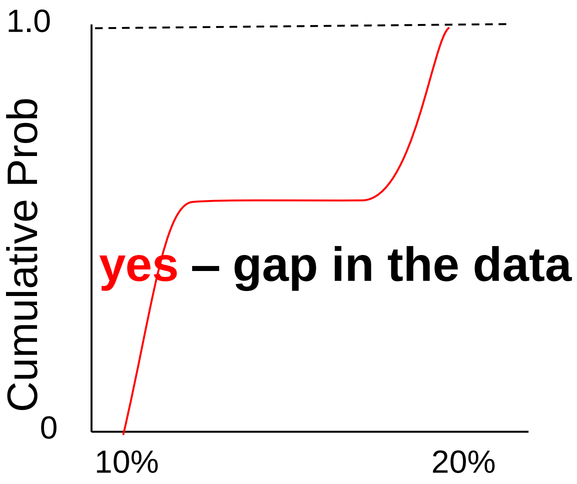
Identify valid CDF?



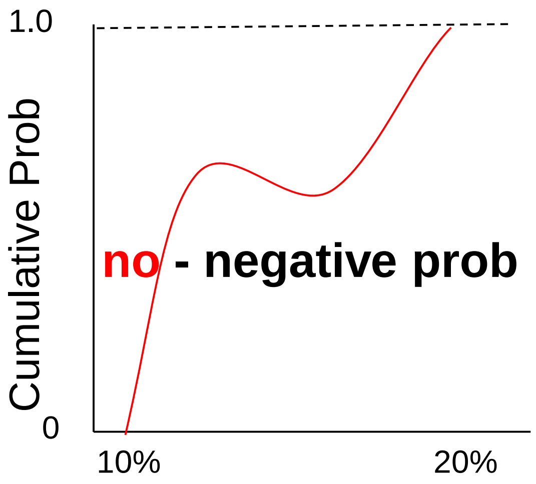
yes



yes - data spike



yes – gap in the data



no - negative prob

Statistical Distributions
Cumulative Distribution Function

How to generate a CDF?

Step1: Sort data samples in an ascending order such that $x_1 \leq x_2 \leq x_3 \leq \ldots \leq x_n$

Step 2: Assign a probability f_i to each sample (for closure $\sum_1^n f_i = 1.0$).

Step 3: Integrate by summing probabilities to calculate cumulative probability of being ≤ than each data value.

$$\boldsymbol{F_i = P(X < x_i) = \sum_{1}^{i} f_i}$$

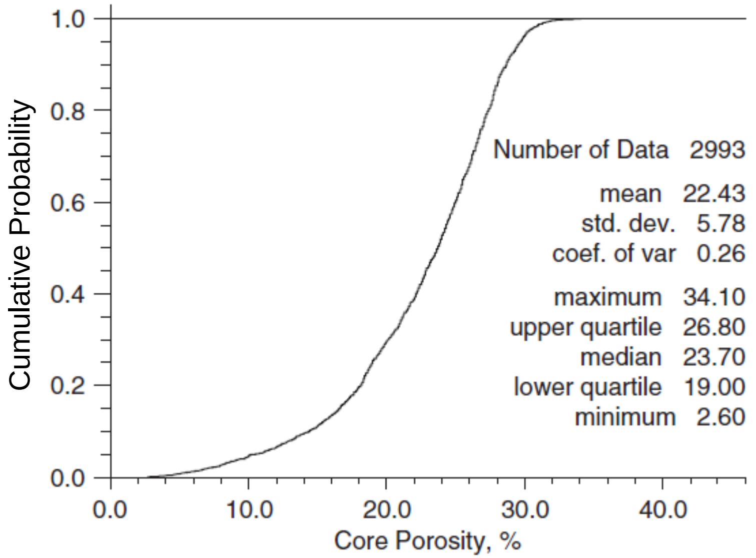
$$\boldsymbol{F_i = \frac{i}{n}}$$

One method, if equal weighted.

Step 4: Plot $\boldsymbol{F_i}$ vs. $\boldsymbol{x_i}$.

Comments:

- Assumption of minimum and maximum.
- Equal weighting is updated with declustering weights (later).
- No bin size assumption.



Porosity CDF, image from Pyrcz and Deutsch (2014).

The i/n method for cumulative probability

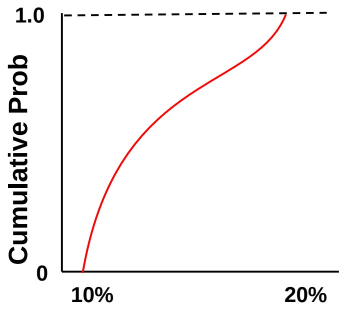
Sorted Value	$x_1, x_2, x_3, \ldots, x_n$
	$F_1, F_2, F_3, \ldots, F_n$
Cumulative Probability	$\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, 1$

Statistical Distributions
Cumulative Distribution Function

Must a CDF be continuous, differentiable to calculate a PDF?

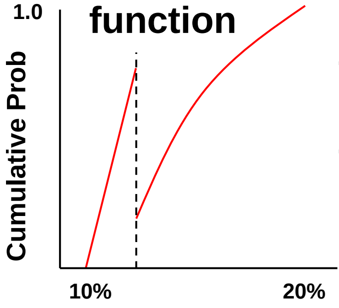
PDF **CDF**

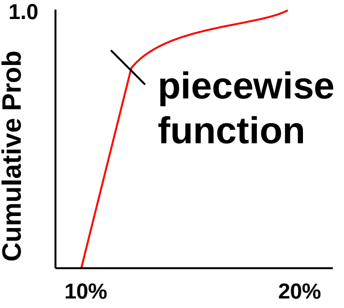
$$f_x(x) = \frac{dF_x(x)}{dx}$$



Continuous? – yes, but the CDF may be a piecewise function, then integrate each segment separately

can't have a jump discontinuity function



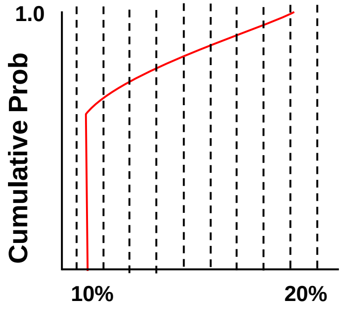


piecewise function

Differentiable? – no, not necessary

What can you do if not?
Use bins and numerical differentiation over the bins or work with raw data.

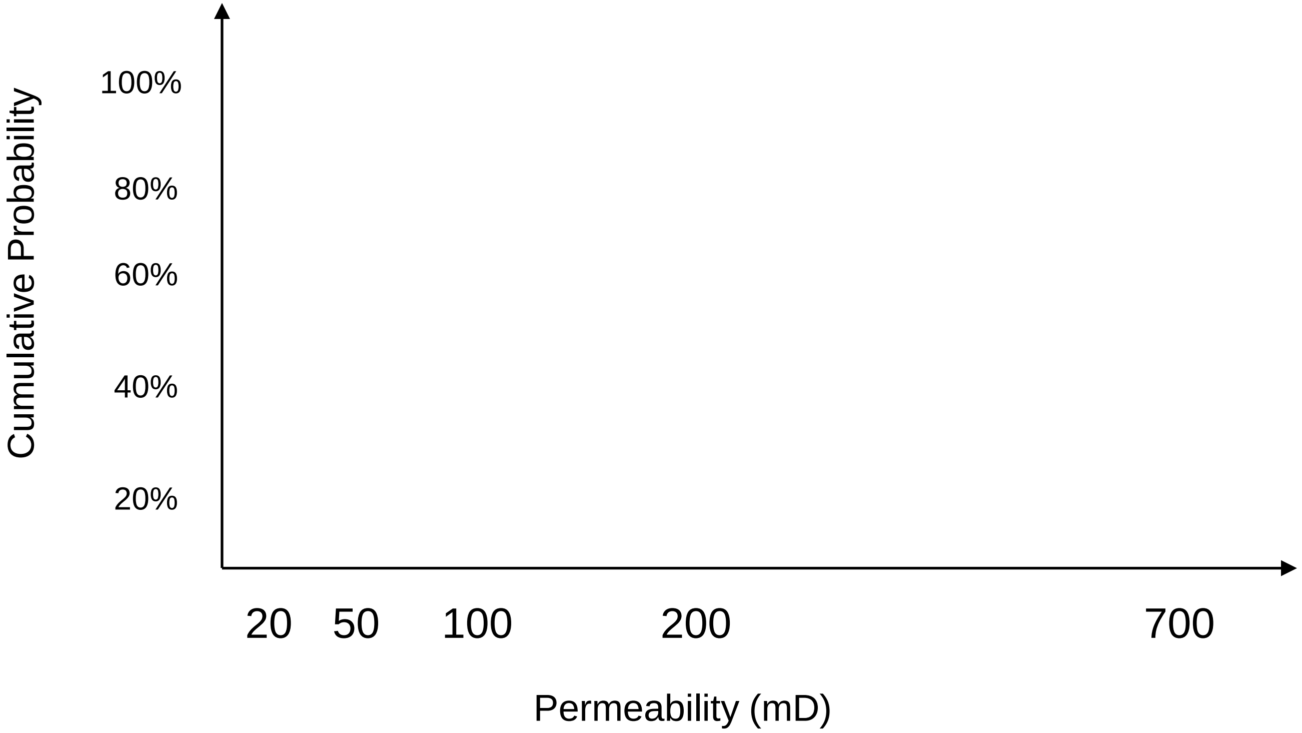
$$f_x(x) = \frac{F_x\left(x + {}^{\Delta}/_{2}\right) - F_x\left(x - {}^{\Delta}/_{2}\right)}{\Delta}$$



Statistical Distributions
CDF Example

Let’s build a CDF?

Permeability Samples:
20, 50, 100, 200, 700 mD



Statistical Distributions
CDF Example

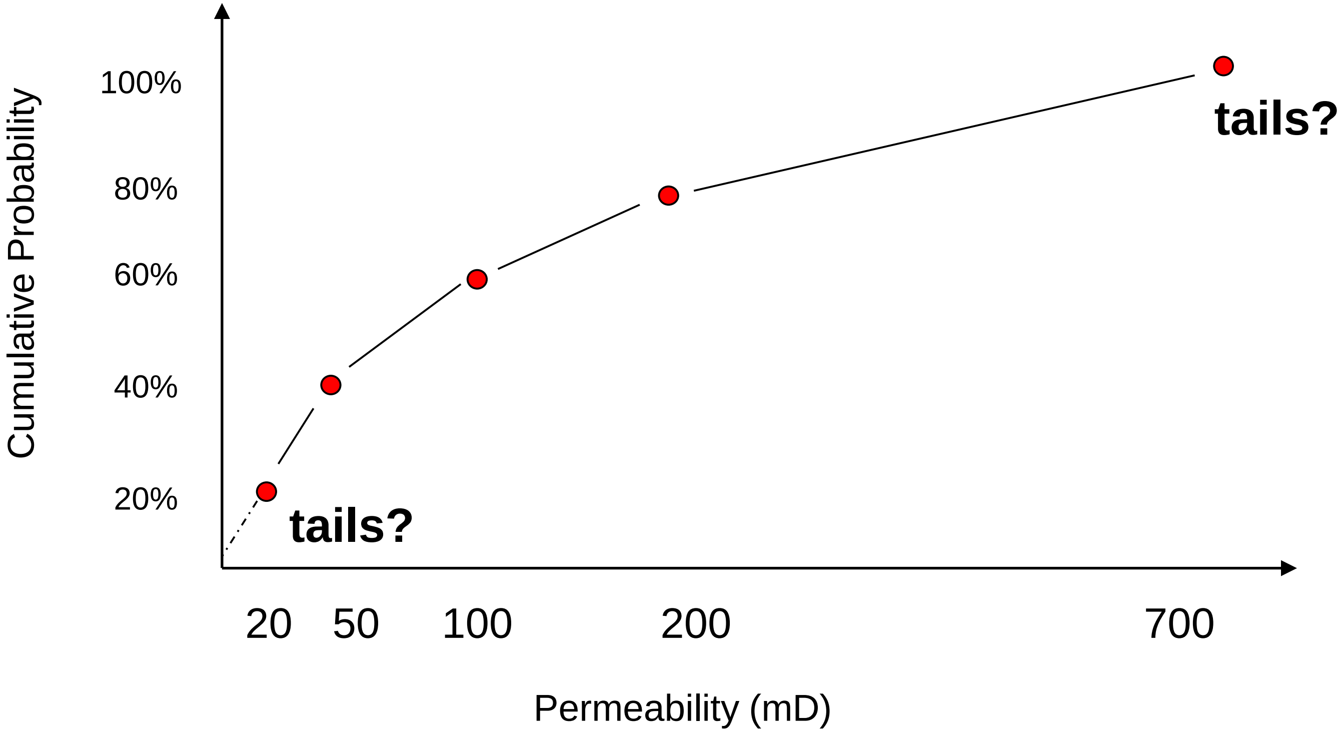
Let's build a CDF?

Permeability Samples:
20, 50, 100, 200, 700 mD

Permeability Samples: 20, 50, 100, 200, 700 mD

$F_k(K)$ = P(K < k_α) 20%, 40%, 60%, 80% 100%

assuming $F_i = \frac{i}{n}$



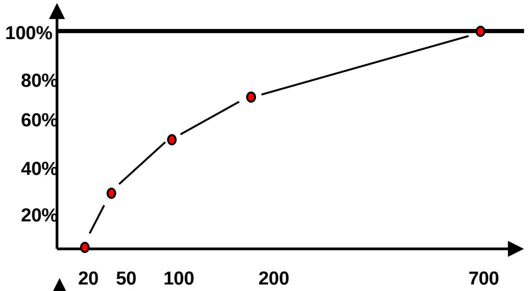
What is tails extrapolation –> the behavior as a distribution reaches the min and max

Statistical Distributions
Tail Assumption

Alternative Tail Assumptions

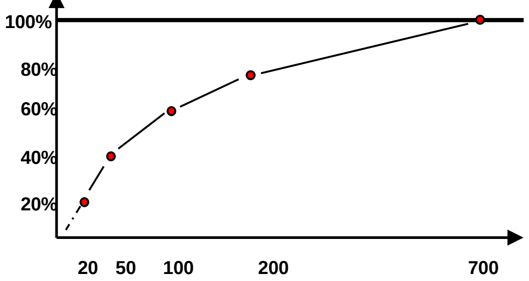
Cumulative Probability

Known Lower and Upper Tail → $F_i = \dfrac{i-1}{n-1}$

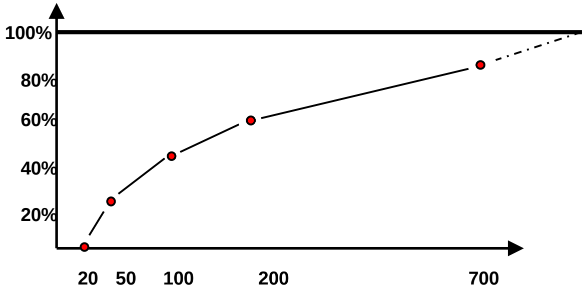


Unknown Lower Tail → $F_i = \dfrac{i}{n}$

Welcome to use this in the class.

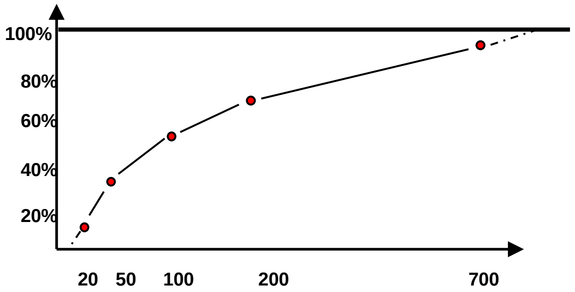


Unknown Upper Tail → $F_i = \dfrac{i-1}{n}$



Unknown Upper and Lower Tail → $F_i = \dfrac{i}{n+1}$

Prefered method, better uncertainty model.



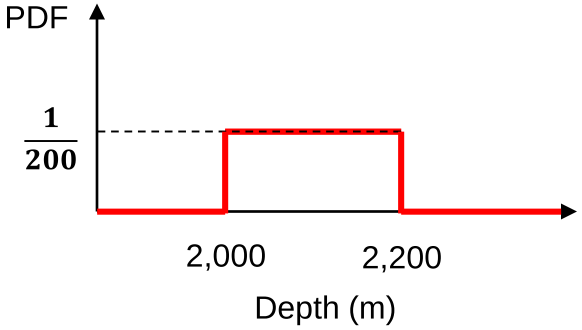
Assuming data in ascending order, $i = 1, \ldots, n$

$$x_1 \le x_2 \le x_3 \le \ldots \le xn$$

Statistical Distributions
PDF / CDF Example

Example: The depth of oil-water contact in a region of a reservoir is described with the following PDF:

$$f(z) = \begin{cases} 0 & z \leq 2{,}000\text{m} \\ 1/200 & 2{,}000\text{m} < z \leq 2{,}200\text{m} \\ 0 & z > 2{,}200\text{m} \end{cases}$$



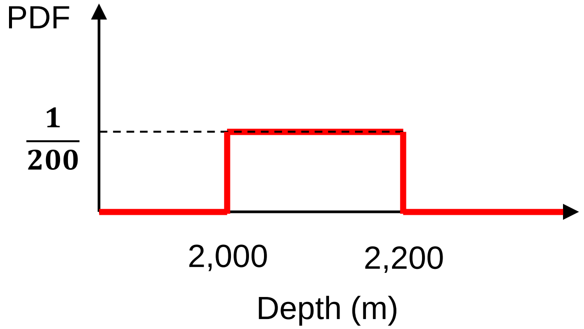
Example:

Is this a valid PDF? What do you need to check?

Statistical Distributions
PDF / CDF Example

Example: The depth of oil-water contact in a region of a reservoir is described with the following PDF:

$$\boldsymbol{f}(\boldsymbol{z}) = \begin{cases} 0 & z \leq 2{,}000\text{m} \\ 1/200 & 2{,}000\text{m} < z \leq 2{,}200\text{m} \\ 0 & z > 2{,}200\text{m} \end{cases}$$



Example:

Is this a valid PDF? What do you need to check?

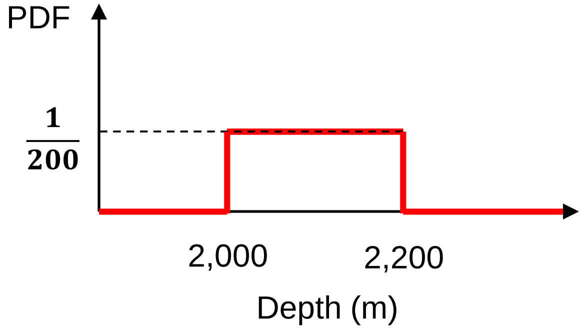
Non-negative, $\boldsymbol{f}(\boldsymbol{z}) \geq \mathbf{0}$.
Closure – probability sum to 1.0.

$$\int_{-\infty}^{+\infty} f(u)du = 1 \qquad \int_{2{,}000}^{2{,}200} \frac{1}{200} dz = \frac{1}{200} z \Big|_{2{,}000}^{2{,}200} = \frac{1}{200} (2{,}200\text{-}2{,}000) = 1.0$$

Statistical Distributions
PDF / CDF Example

Example: The depth of oil-water contact in a region of a reservoir is described with the following PDF:

$$f(z) = \begin{cases} 0 & z \leq 2{,}000\text{m} \\ 1/200 & 2{,}000\text{m} < z \leq 2{,}200\text{m} \\ 0 & z > 2{,}200\text{m} \end{cases}$$



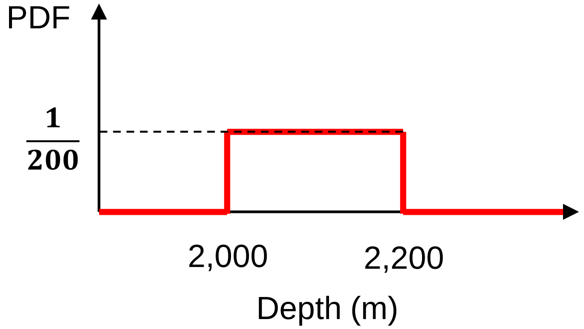
Example:

What is the corresponding CDF?

Statistical Distributions
PDF / CDF Example

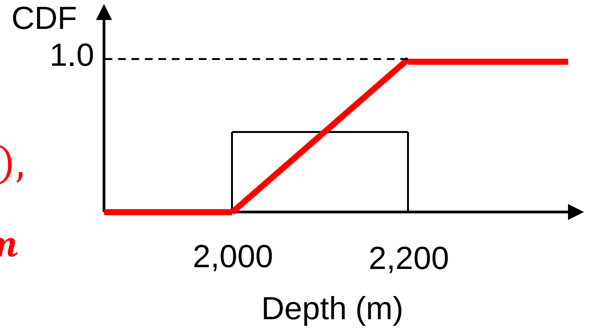
Example: The depth of oil-water contact in a region of a reservoir is described with the following PDF:

$$f(z) = \begin{cases} 0 & z \leq 2{,}000\text{m} \\ 1/200 & 2{,}000\text{m} < z \leq 2{,}200\text{m} \\ 0 & z > 2{,}200\text{m} \end{cases}$$



Example:

What is the corresponding CDF?



$$F_z(z) = \int_{-\infty}^{z} f(u)du = \int_{2{,}000}^{z} \frac{1}{200} du = \frac{1}{200}(z - 2000),$$
$$if\ 2{,}000m\ < z\ \leq 2{,}200m$$

$$F_z(z) = 0, if\ z\ \leq 2{,}000m$$

$$F_z(z) = 1, if\ z\ \geq 2{,}200m$$

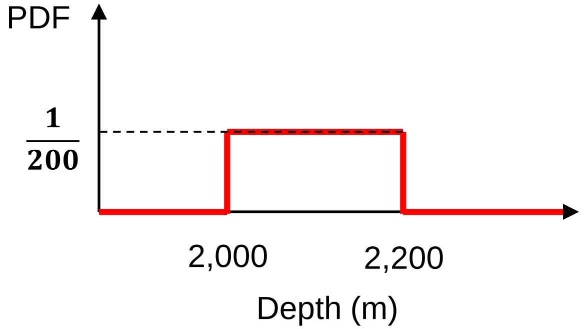
Statistical Distributions
PDF / CDF Example

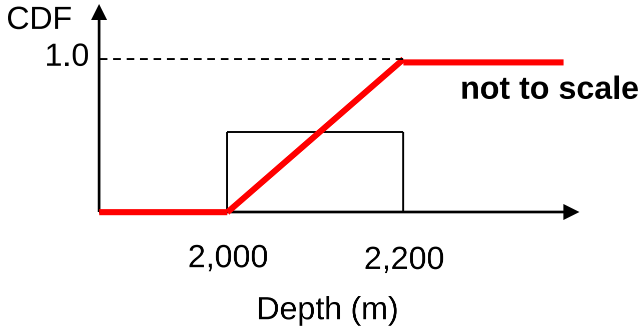
Example: The depth of oil-water contact in a region of a reservoir is described with the following PDF:

$$f(z) = \begin{cases} 0 & z \leq 2{,}000\text{m} \\ 1/200 & 2{,}000\text{m} < z \leq 2{,}200\text{m} \\ 0 & z > 2{,}200\text{m} \end{cases}$$

Example:

What is the probability that the OWC is between 2,050 and 2,150?

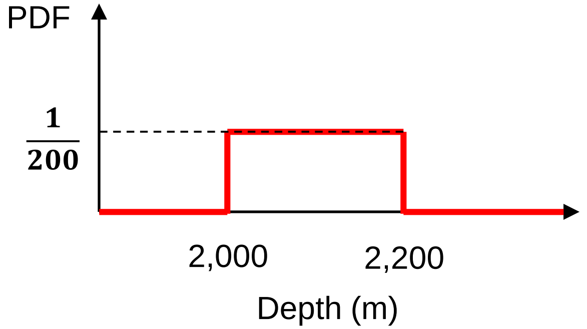




Statistical Distributions
PDF / CDF Example

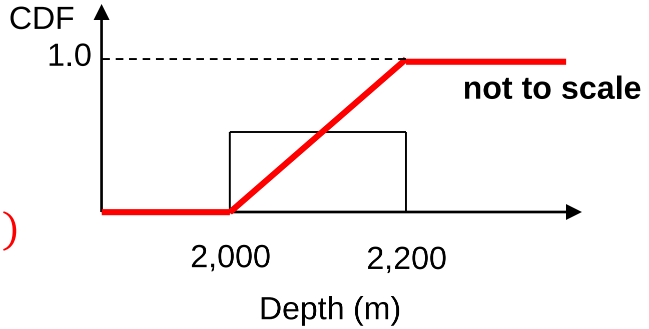
Example: The depth of oil-water contact in a region of a reservoir is described with the following PDF:

$$f(z) = \begin{cases} 0 & z \leq 2{,}000\text{m} \\ 1/200 & 2{,}000\text{m} < z \leq 2{,}200\text{m} \\ 0 & z > 2{,}200\text{m} \end{cases}$$



Example:

What is the probability that the OWC is between 2,050 and 2,150?



not to scale

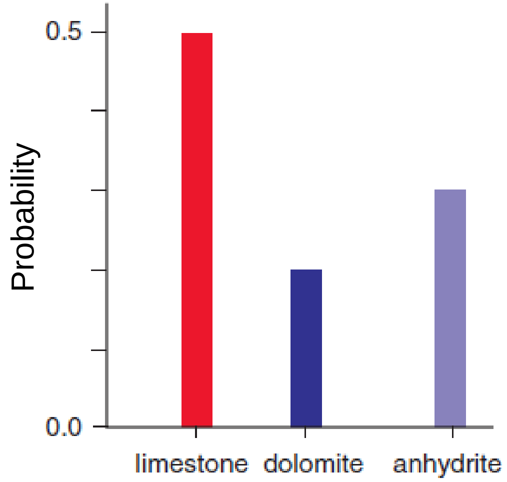
$$\boldsymbol{P(2{,}050 < z < 2{,}1500) = F_z(2{,}150) - F_z(2{,}050)}$$

$$\boldsymbol{= \frac{2{,}150 - 2{,}000}{200} - \frac{2{,}050 - 2000}{200} = 0.5}$$

Statistical Distributions
PDF / CDF in Practice

In practice, what is done with statistical distributions?
For Categorical Variables:

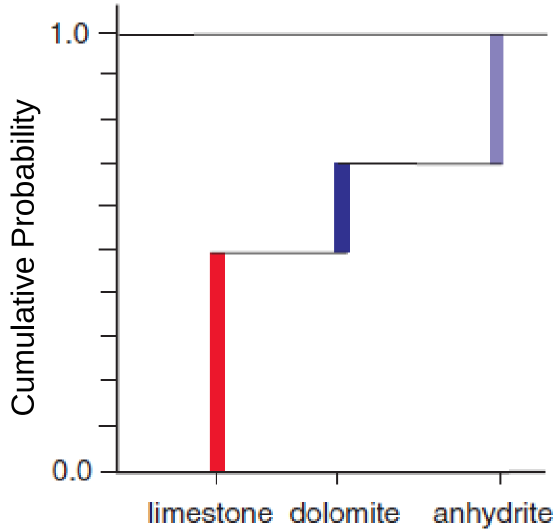
1. Data is pooled, [weighted] proportions calculated for each category = PDF.



Lithofacies categorical PDF.

2. Summary statistics are the proportions.

3. May then calculate a CDF. →



Lithofacies categorical CDF.

4. Note: CDF category ordering is not important for common applications:

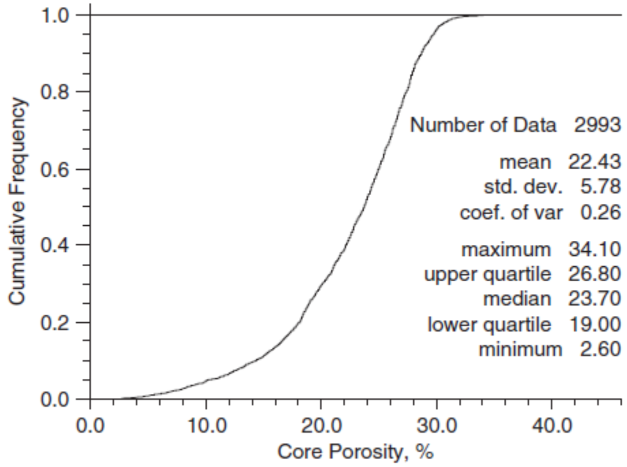
e.g., prediction, Monte Carlo simulation.

Images from Pyrcz and Deutsch (2014), Geostatistical Reservoir Modeling.

Statistical Distributions
PDF / CDF in Practice

In practice, what is done with distributions in subsurface projects?
For Continuous Variables:

1. Data is pooled to calculate a CDF with tail extrapolation.

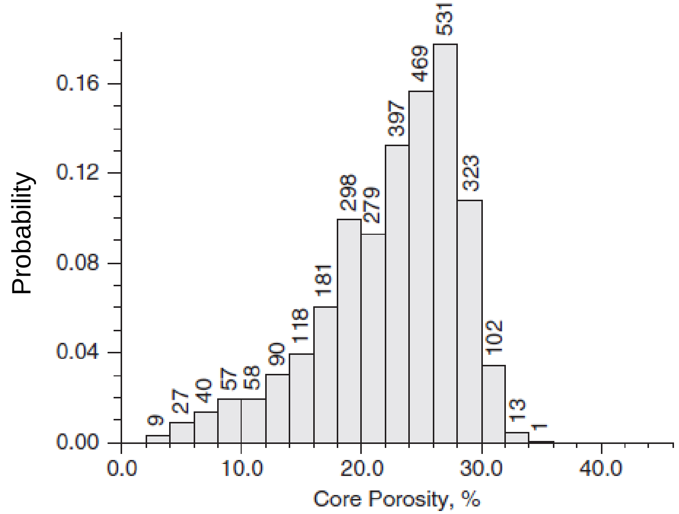


Porosity continuous CDF.

2. Calculate the summary statistics etc.

3. Calculate the histogram by binning the data.

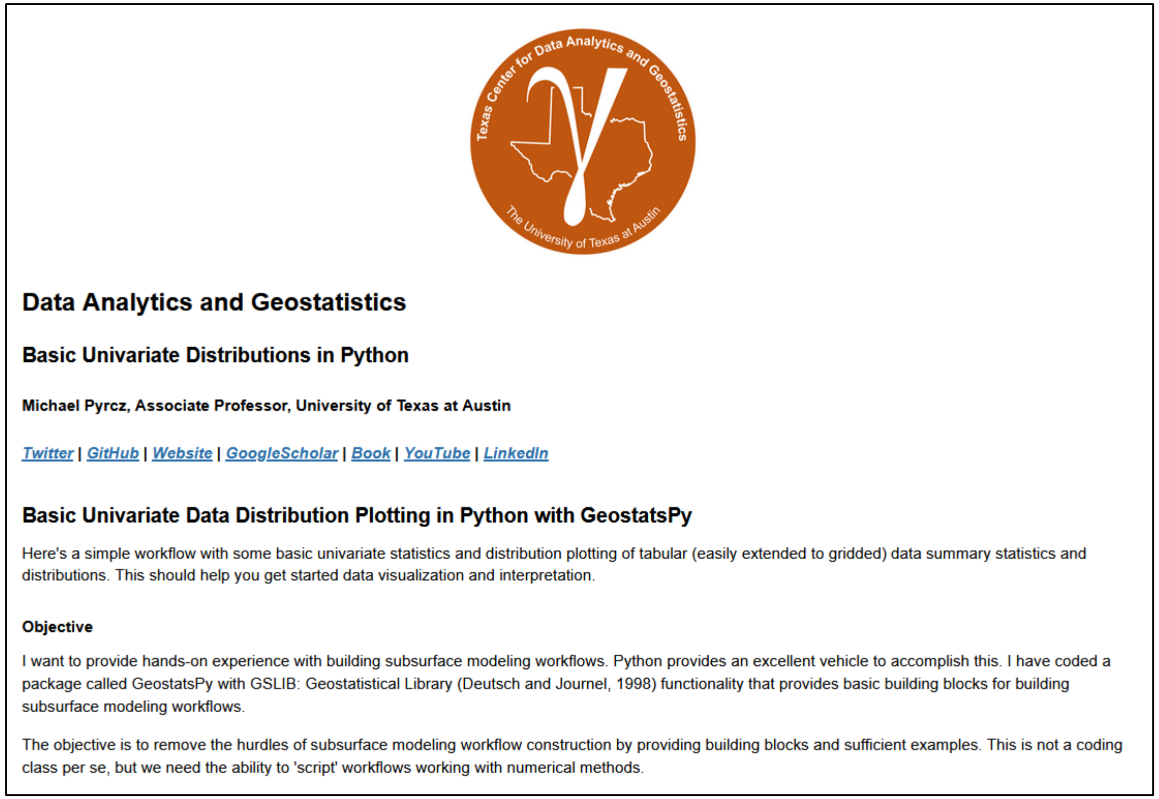
4. Calculate the normalized histogram dividing the frequencies (with weights) by the total number of samples.



Porosity continuous normalized histogram.

Images from Pyrcz and Deutsch (2014), Geostatistical Reservoir Modeling.

Statistical Distributions
PDF / CDF in Practice



Data Analytics and Geostatistics

Basic Univariate Distributions in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

Twitter | *GitHub* | *Website* | *GoogleScholar* | *Book* | *YouTube* | *LinkedIn*

Basic Univariate Data Distribution Plotting in Python with GeostatsPy

Here's a simple workflow with some basic univariate statistics and distribution plotting of tabular (easily extended to gridded) data summary statistics and distributions. This should help you get started data visualization and interpretation.

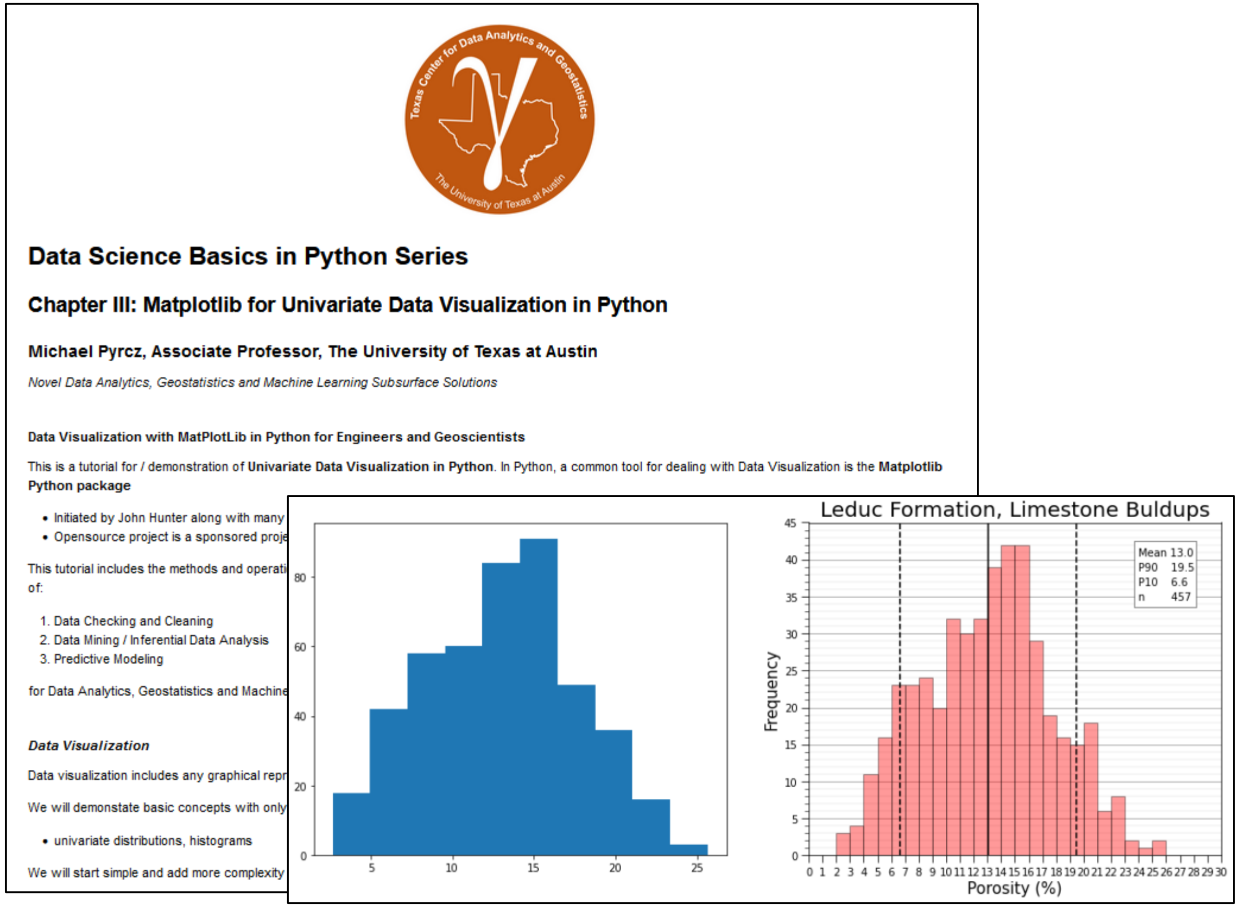
Objective

I want to provide hands-on experience with building subsurface modeling workflows. Python provides an excellent vehicle to accomplish this. I have coded a package called GeostatsPy with GSLIB: Geostatistical Library (Deutsch and Journel, 1998) functionality that provides basic building blocks for building subsurface modeling workflows.

The objective is to remove the hurdles of subsurface modeling workflow construction by providing building blocks and sufficient examples. This is not a coding class per se, but we need the ability to 'script' workflows working with numerical methods.

Example in Python
Jupyter Notebook for histograms, PDFs and CDFs.

File is: GeostatsPy_datadistributions.ipynb



Data Science Basics in Python Series

Chapter III: Matplotlib for Univariate Data Visualization in Python

Michael Pyrcz, Associate Professor, The University of Texas at Austin

Novel Data Analytics, Geostatistics and Machine Learning Subsurface Solutions

Data Visualization with MatPlotLib in Python for Engineers and Geoscientists

This is a tutorial for / demonstration of **Univariate Data Visualization in Python**. In Python, a common tool for dealing with Data Visualization is the **Matplotlib Python package**

- Initiated by John Hunter along with many
- Opensource project is a sponsored proje

This tutorial includes the methods and operati of:

1. Data Checking and Cleaning
2. Data Mining / Inferential Data Analysis
3. Predictive Modeling

for Data Analytics, Geostatistics and Machine

Data Visualization

Data visualization includes any graphical repr

We will demonstate basic concepts with only

- univariate distributions, histograms

We will start simple and add more complexity

Effective univariate data visualization in Python.

File is: PythonDataBasics_Univariate_Visualization.ipynb

PGE 337 Data Analytics and Geostatistics

Lecture 3: Displaying Distributions

Lecture outline . . .

- Plotting, Data Visualization
- Histograms, Probability Density Functions
- Cumulative Distribution Functions

Michael Pyrcz, The University of Texas at Austin

Introduction

General Concepts

Univariate

- PDF / CDF
- Statistics
- Distributions
- Heterogeneity
- Hypothesis

Bivariate

Time Series Analysis

Spatial Analysis

Machine Learning

Uncertainty Analysis