Subsurface Data Analytics and Machine Learning

Data Analytics

Lecture outline . . .

- Data Analytics
- Confidence Intervals
- Hypothesis Testing
- Bootstrap

Introduction

Data Analytics

Inferential Methods

Predictive Methods

Advanced Methods

Conclusions

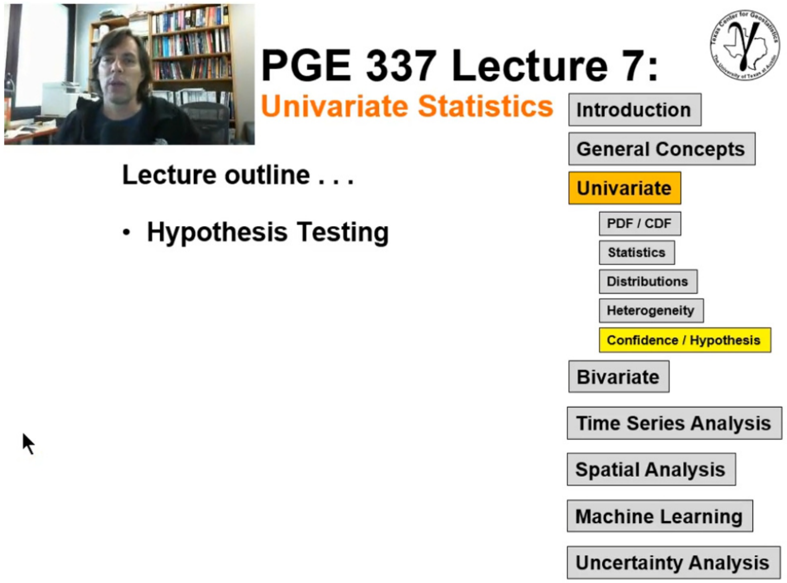
Instructor: Michael Pyrcz, the University of Texas at Austin

Subsurface Data Analytics and Machine Learning

Data Analytics

Other Resources:

- Recorded Lectures on hypothesis testing, confidence intervals, bootstrap etc.



PGE 337 Lecture 7:

Univariate Statistics

Lecture outline . . .

- Hypothesis Testing

Instructor: Michael Pyrcz, the University of Texas at Austin

Goals of This Lecture

- General concepts of data analytics

- Data analytics is the application of geoscience and engineering expertise and statistics with data

- Provide 3 basic workflows that you could use to:
  - add uncertainty
  - report significance

Subsurface Data Analytics and Machine Learning

Data Analytics

Lecture outline . . .

- Data Analytics

Introduction

Data Analytics

Inferential Methods

Predictive Methods

Advanced Methods

Conclusions

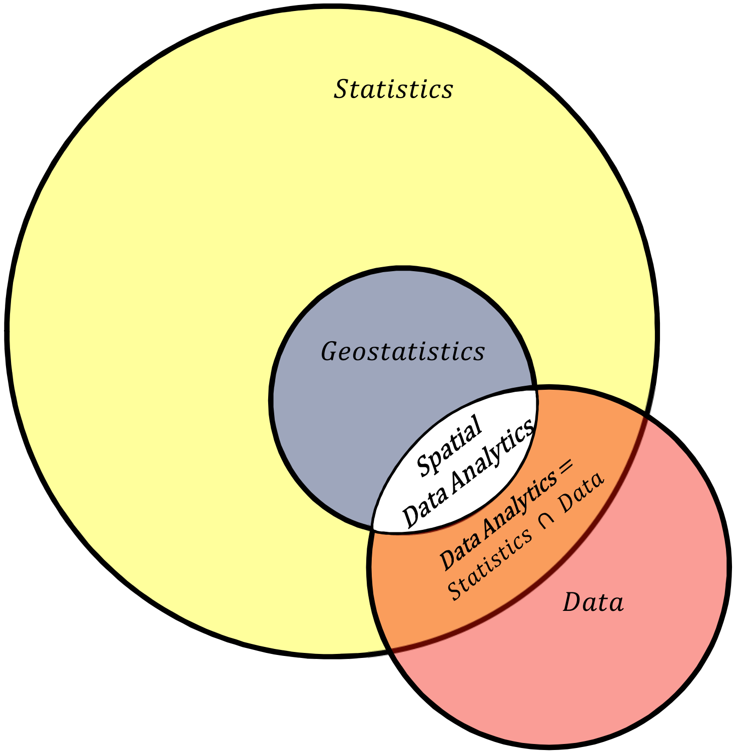
Instructor: Michael Pyrcz, the University of Texas at Austin

Spatial Big Data Analytics

Statistics is collecting, organizing, and interpreting data, as well as drawing conclusions and making decisions.

Geostatistics is a branch of applied statistics: (1) the spatial (geological) context, (2) the spatial relationships, (3) volumetric support, and (4) uncertainty.

Big Data Analytics is the process of examining large and varied data sets (big data) to discover patterns and make decisions.



Venn diagram for spatial big data analytics.

$$\mathit{Data\ Analytics} = \mathit{Geostatistics} \cap \mathit{Data}$$

$$\mathit{Spatial\ Data\ Analytics} = \mathit{Geostatistics} \cap \mathit{Data}$$

$$\mathit{Spatial\ Big\ Data\ Analytics} = \mathit{Geostatistics} \cap \mathit{Big\ Data}$$

Data Analytics

‘Data analytics (DA) is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems and software.’ – Search Data Management

‘Data analysis is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, informing conclusions, and supporting decision-making.’ – Wikipedia

‘Data Analytics / Analysis is Statistics.’

‘Subsurface Data Analytics / Analysis is Geostatistics.’

There is a lot going on concerning **working with and learning from data**.

‘At the core is good practice with statistics with domain expertise’.

Data Analytics

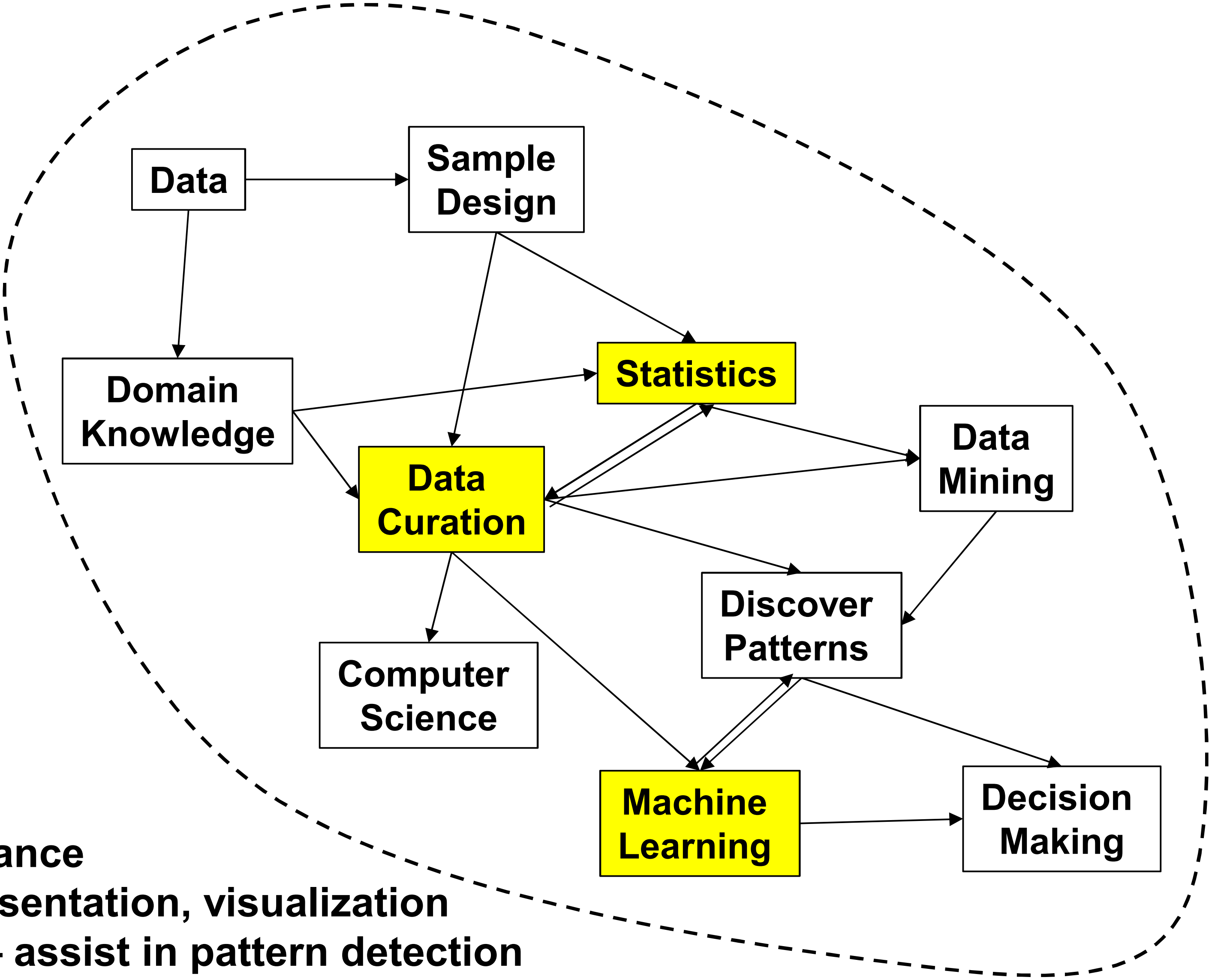
Data

Statistics

+

Domain
Knowldege

Decisions



Data

Sample
Design

Domain
Knowledge

Statistics

Data
Curation

Data
Mining

Discover
Patterns

Computer
Science

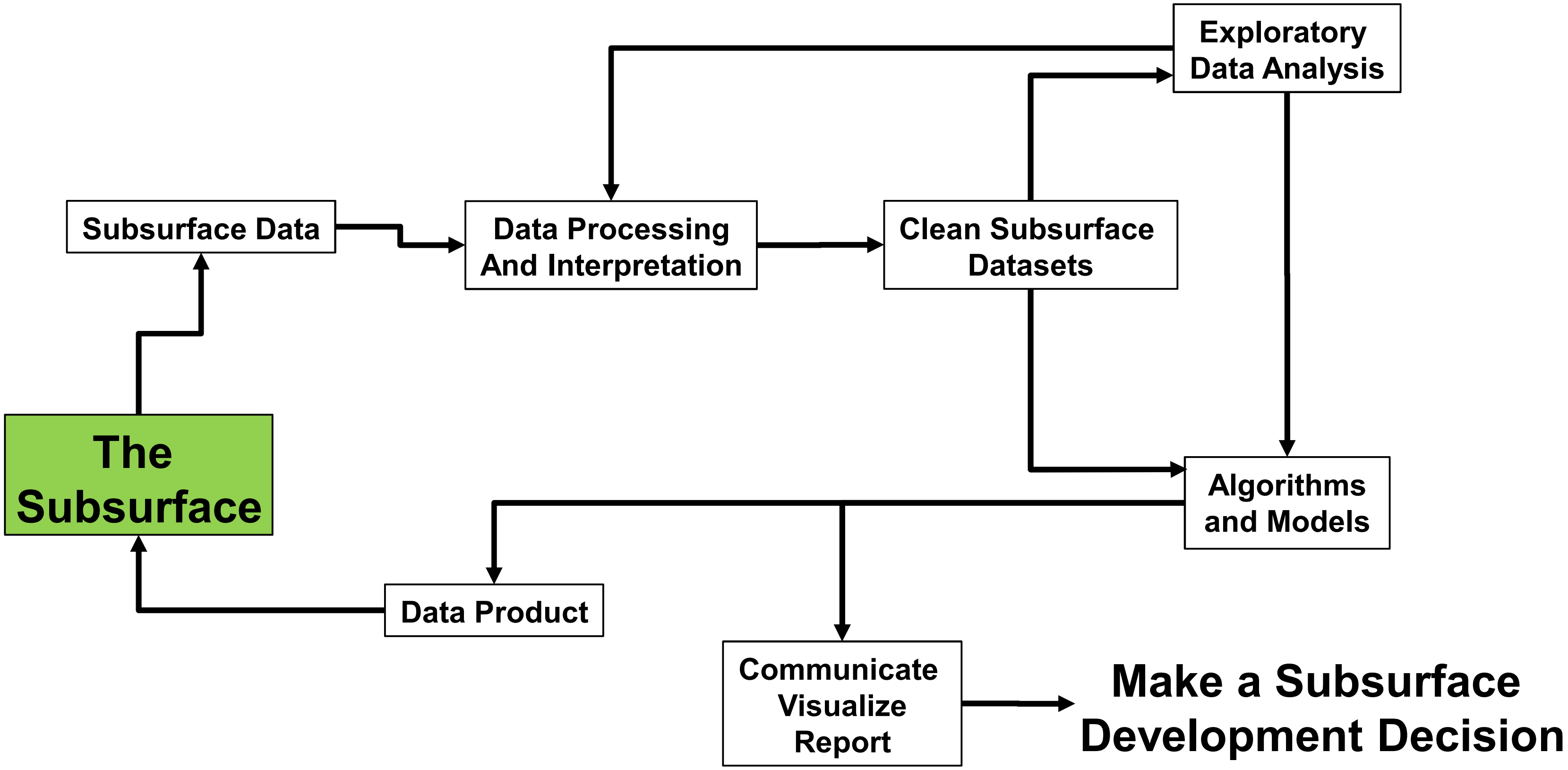
Machine
Learning

Decision
Making

Statistics – significance
Data Curation – presentation, visualization
Machine Learning – assist in pattern detection

Data Science Process

The learning process with data to support decision making.

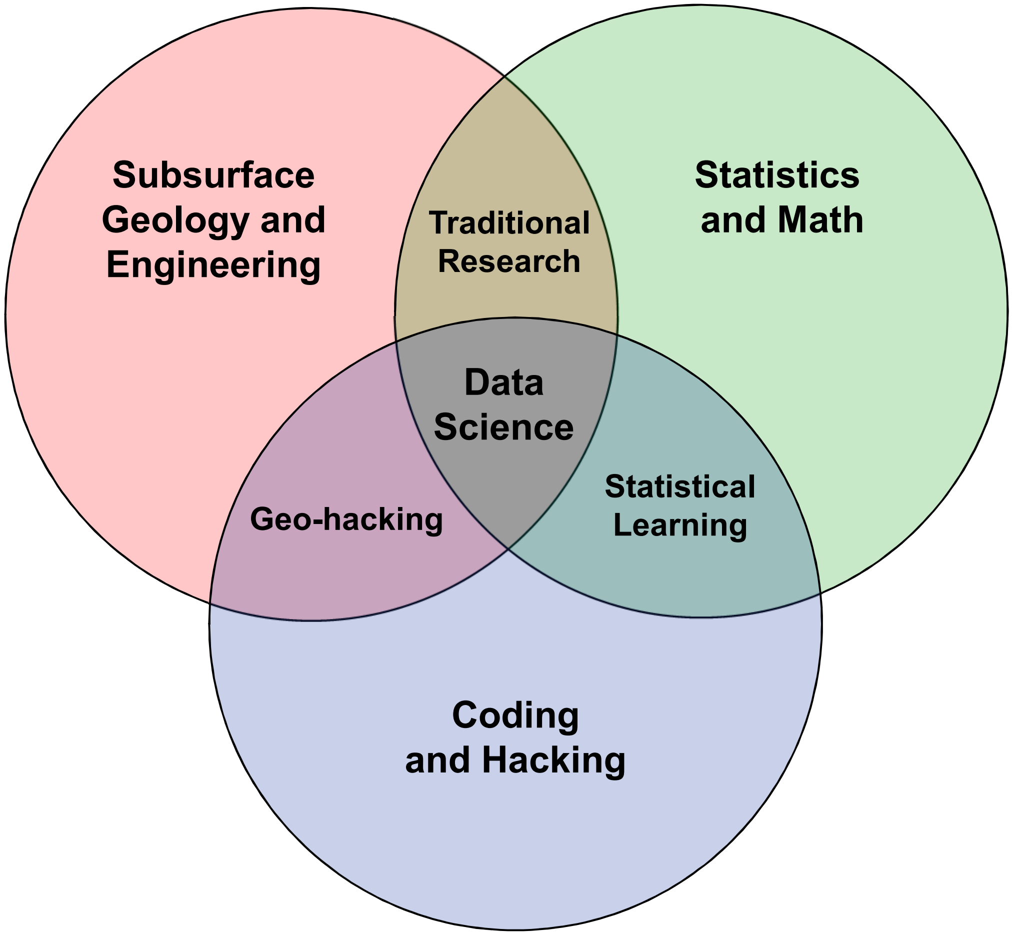

Exploratory Data Analysis
Subsurface Data
Data Processing And Interpretation
Clean Subsurface Datasets
The Subsurface
Algorithms and Models
Data Product
Communicate Visualize Report
Make a Subsurface Development Decision

Analogous to modeling while learning.

Adapted to subsurface data analytics from Doing Data Science, by Schutt & O'Neil (2013)

Data Science Process

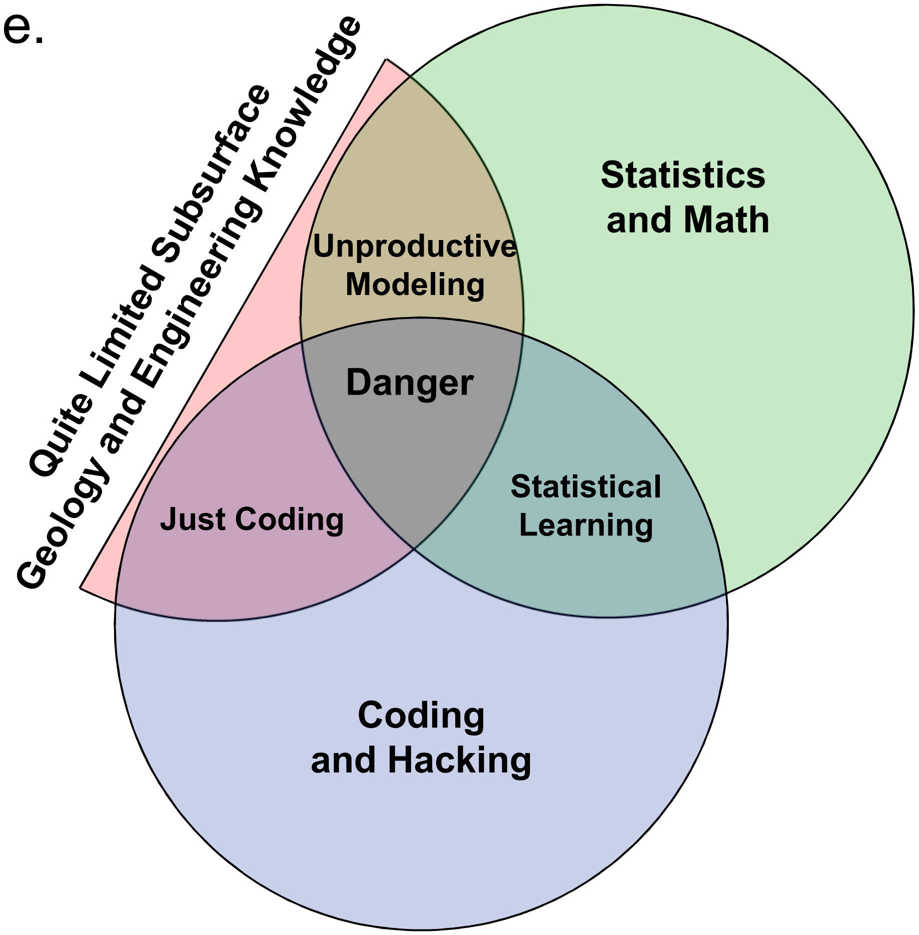
A possible 'ideal' for integrated skill sets from Conway (2015).



Modified from Drew Conway (2015) http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Science Process

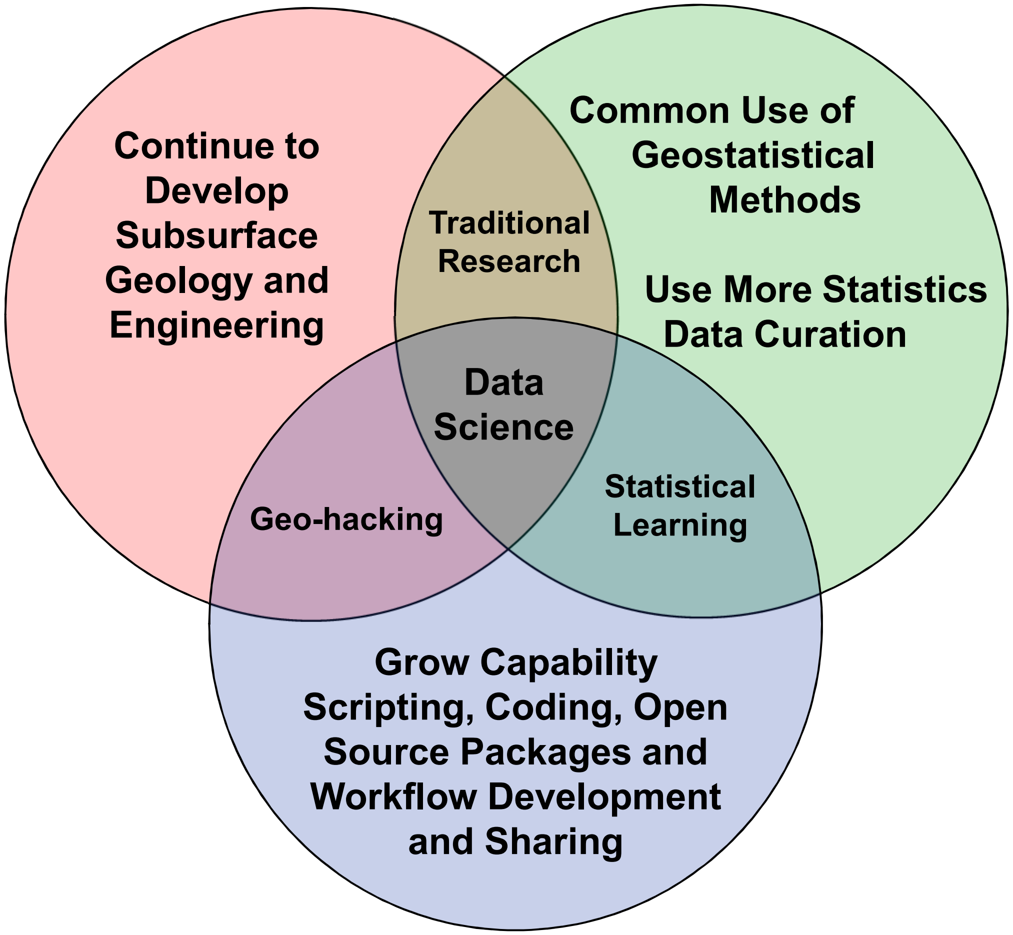
The Digital Crew Change Plan – add a lot of data scientists without domain knowledge.



Modified from Drew Conway (2015) http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Data Science Process

The Grow Analytics Plan – build from core geoscience and engineering capability.


Continue to Develop Subsurface Geology and Engineering
Common Use of Geostatistical Methods
Use More Statistics Data Curation
Traditional Research
Data Science
Geo-hacking
Statistical Learning
Grow Capability Scripting, Coding, Open Source Packages and Workflow Development and Sharing

Modified from Drew Conway (2015) http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

Python for Data Analytics

General Purpose, Popular and High Level Programming Level

- Most popular, active development
- Simple and easy to write, readable computer code
- Simple with automatic memory management, but also integrates complicated coding paradigms like object-oriented.
- Compatible across platforms
- Interpreted programming language, no need to compile the code
- Amazing, standard libraries, and standard packages in Anaconda
- Amazing breadth of open source packages

"With Python, I code less and get more work done."

Python Packages?

We will demonstrate common packages (installed with Anaconda)

- tabular data → pandas
- gridded data → numpy
- data display → matplotlib
- mathematics, trig etc. → math
- statistics → scipy
- geostatistics → geostatspy (my package that we will install)

Recall: I’ll demonstrate to promote accessibility.

The Challenge

Oil and Gas is Unique

Our datasets are massive, multivariate, varied but sparse

Our work spans from pores to basins in scale

We have complex integration of engineering physics, geophysics and geoscience

We face ubiquitous uncertainty that must be managed

There is a layer of irreducible interpretation, subsurface interpretation

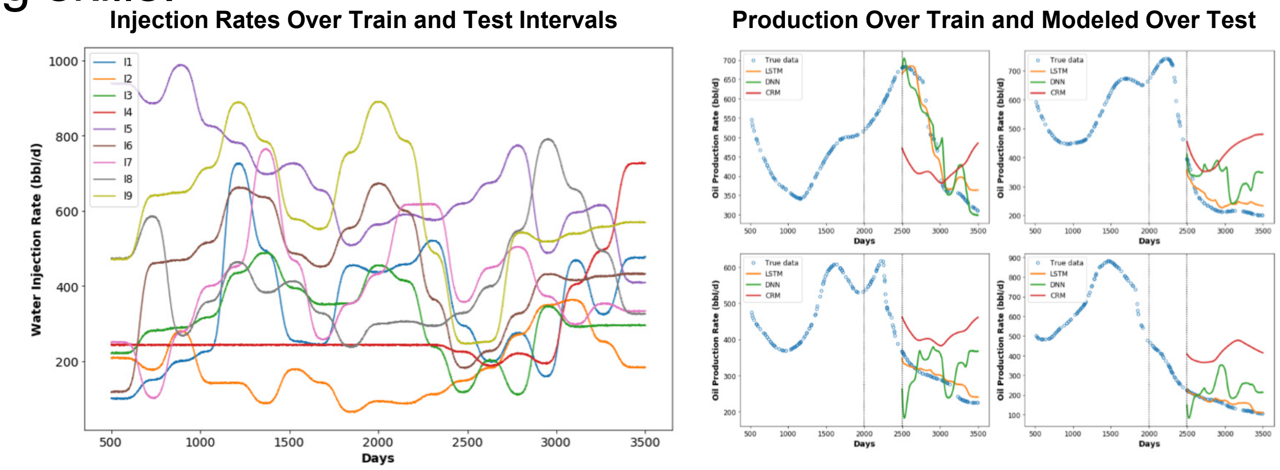
The consequences of our discrete decisions are immense

Opportunities

Opportunity for Academia to Support Industry

1. Development of data analytics and machine learning solutions that account for engineering physics, geophysics and geoscience information

2. Train students for new hires with data analytics and machine learning skills.


Injection Rates Over Train and Test Intervals

Production Over Train and Modeled Over Test

Recurrent neural nets for multiple well injection to prediction model (Nwachukwu et al., 2018)

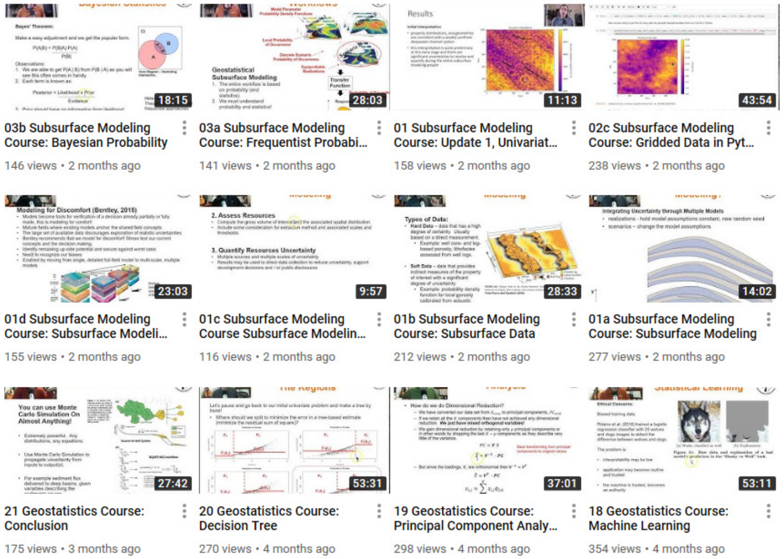
Opportunities

Opportunity for Academia to Support Industry

3. Provide resources to support the growth of operational capability in industry

4. Provide training and resources inside companies

Recent training at Anadarko Offsite



03b Subsurface Modeling Course: Bayesian Probability
146 views • 2 months ago

03a Subsurface Modeling Course: Frequentist Probabi...
141 views • 2 months ago

01 Subsurface Modeling Course: Update 1, Univariat...
158 views • 2 months ago

02c Subsurface Modeling Course: Gridded Data in Pyt...
238 views • 2 months ago

01d Subsurface Modeling Course: Subsurface Modeli...
155 views • 2 months ago

01c Subsurface Modeling Course Subsurface Modelin...
116 views • 2 months ago

01b Subsurface Modeling Course: Subsurface Data
212 views • 2 months ago

01a Subsurface Modeling Course: Subsurface Modeling
277 views • 2 months ago

21 Geostatistics Course: Conclusion
175 views • 3 months ago

20 Geostatistics Course: Decision Tree
270 views • 4 months ago

19 Geostatistics Course: Principal Component Analy...
298 views • 4 months ago

18 Geostatistics Course: Machine Learning
354 views • 4 months ago

GeostatsGuy Lectures YouTube Channel

Data Analytics

Data Analytics Includes Good Practice with Statistics.

- **Statistics are used for Inference and Prediction!**

So We Include a Lecture with Fundamental Statistical Methods to provide practical subsurface.

- **Confidence Intervals** – reporting uncertainty in a sample statistic
- **Bootstrap** – general method for uncertainty in a sample statistic
- **Hypothesis Testing** – reporting significance of differences

Subsurface Data Analytics and Machine Learning

Data Analytics

Lecture outline . . .

- Confidence Intervals

Introduction

Data Analytics

Inferential Methods

Predictive Methods

Advanced Methods

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin

Uncertainty in a Statistic

Why do we care?

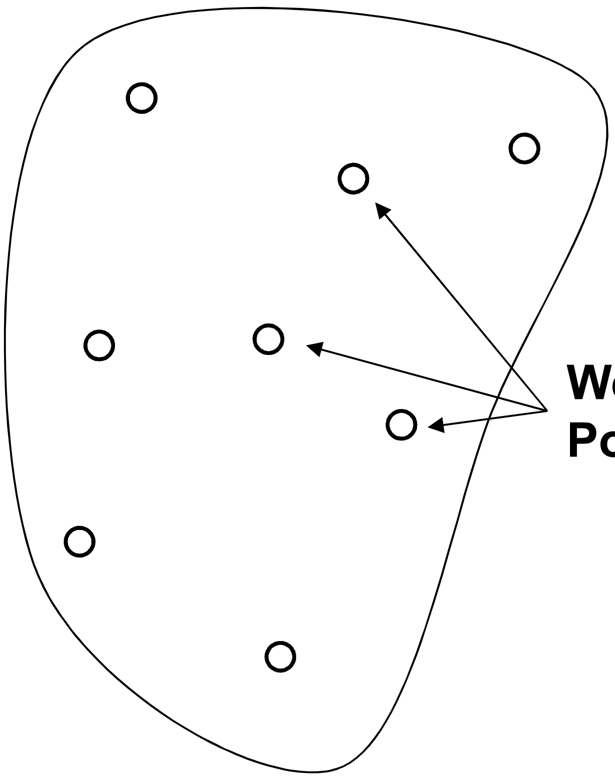
- We use statistics for prediction away from wells!

What do confidence intervals and bootstrap capture?

- Uncertainty due to sparse data

These methods should be used along with expert geoscience and engineering judgement and mapping.

AOI - area of interest



Well-based Porosity

What is the Average porosity over the AOI?

Sample and Population Statistics and Parameters

	Sample	Population
proportion	$\hat{p}$	p
mean	$\overline{x}$	μ
standard deviation	s	σ
variance	s^2	σ^2
	statistic	parameter
	based on limited samples	unavailable

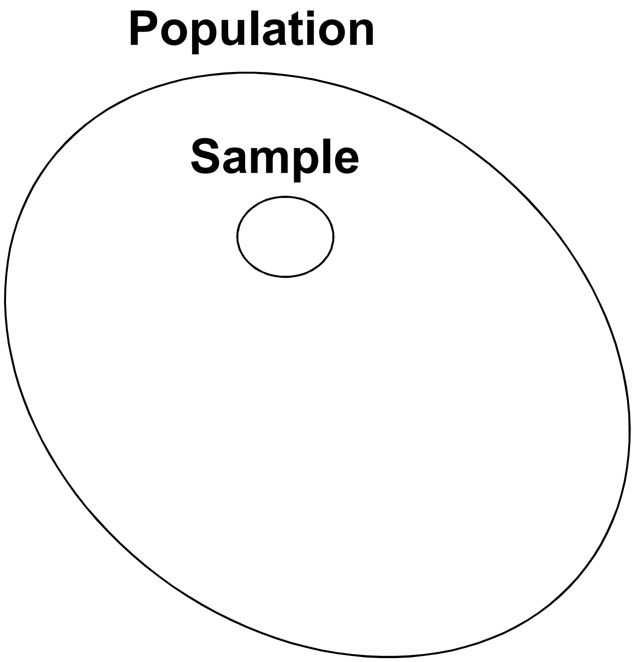


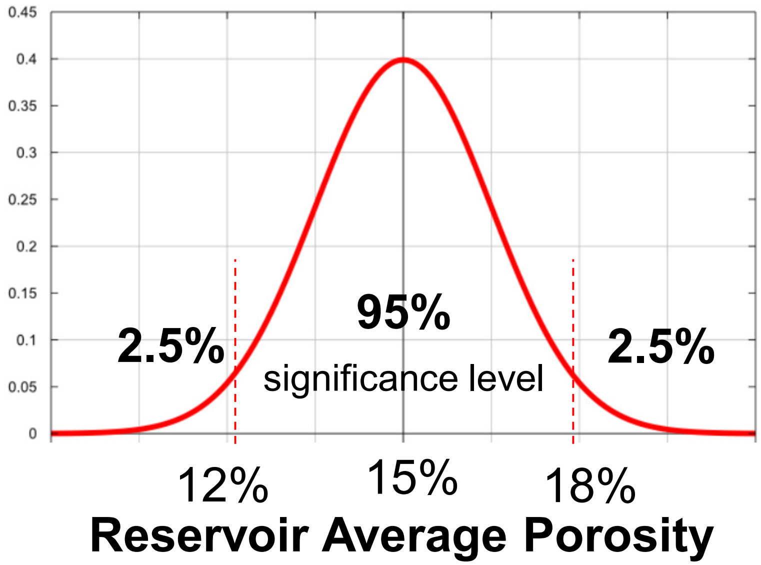
Diagram representing the inaccessible population and sample subset that is available.

Confidence Interval Definition

The uncertainty in a **summary statistic** represented as a range, lower and upper bound, based on a specified probability interval known as the **confidence level**.

We communicate confidence intervals like this:

- There is a 95% probability (or 19 times out of 20) that the true reservoir average porosity is between 12% and 18%



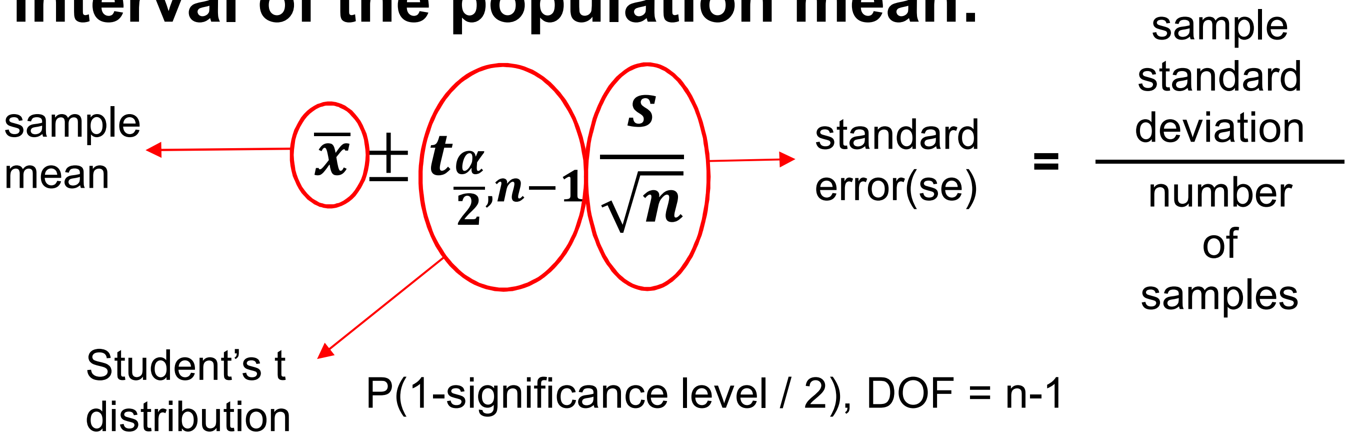
We measured 15%

…but, 95% probability the true result is between 12% and 18%.

Confidence Interval

More on t-score vs. z-score

Confidence interval of the population mean:



$$\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$$

sample mean

Student's t distribution

P(1-significance level / 2), DOF = n-1

standard error(se) = $\frac{\text{sample standard deviation}}{\text{number of samples}}$

Example:

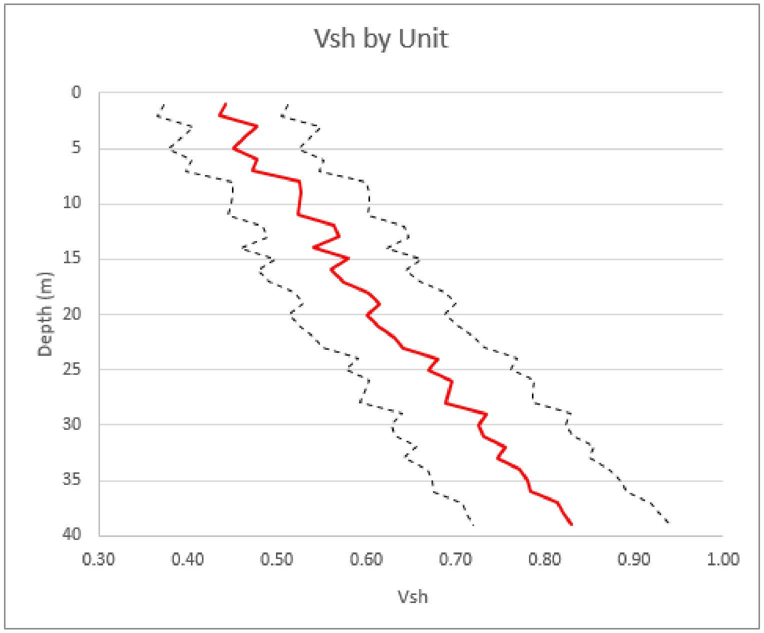
if $\bar{x}$ = 14%, s = 2% and n = 100, then

95% CI, $\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} = \mathbf{14\% \pm 1.98}\frac{\mathbf{2\%}}{\sqrt{\mathbf{100}}}$ =14% ±0.79%

Confidence Interval
Of Vertical Trend (Mean by Depth)

Example Application

- Confidence interval on a plot. We have Vsh samples over multiple depths.



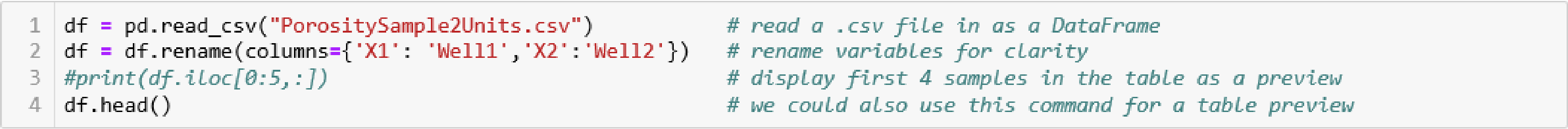
Note: we can open this in Excel.

Vsh vs. Depth with 95% confidence interval.

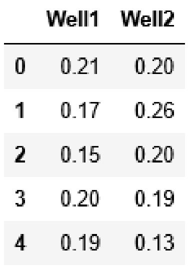
Confidence Interval Demonstration

Let’s load a simple data set with 20 measures of porosity from 2 different units of the reservoir.

- We will calculate the confidence interval for average for each unit



```
df = pd.read_csv("PorositySample2Units.csv")          # read a .csv file in as a DataFrame
df = df.rename(columns={'X1': 'Well1','X2':'Well2'})  # rename variables for clarity
#print(df.iloc[0:5,:])                                # display first 4 samples in the table as a preview
df.head()                                             # we could also use this command for a table preview
```



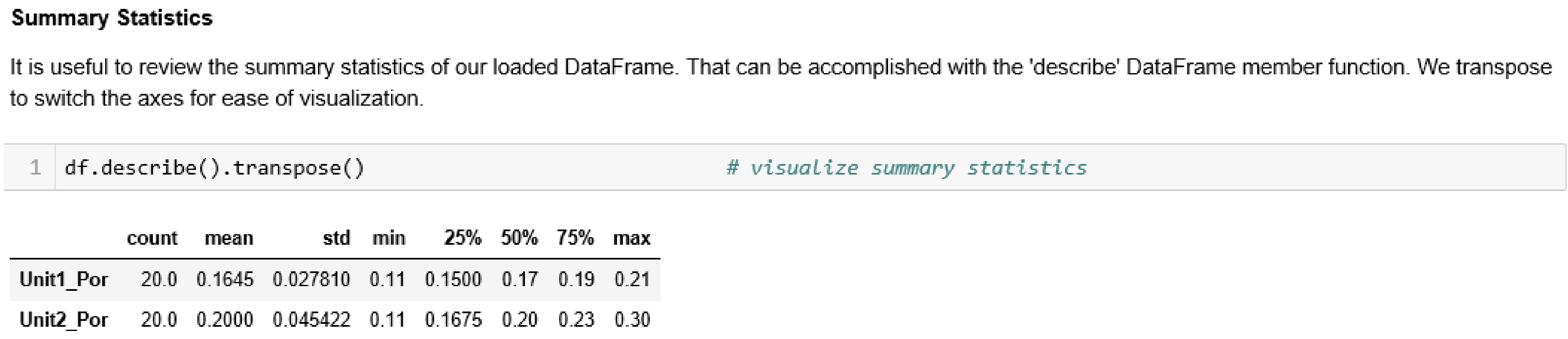
	Well1	Well2
0	0.21	0.20
1	0.17	0.26
2	0.15	0.20
3	0.20	0.19
4	0.19	0.13

- We preview the data set and see porosity measures for X1, X2 as unit 1 and 2.

Confidence Interval Demonstration

Let’s calculate the summary statistics to compare the porosity values from the two units.

- What is the uncertainty in the average porosity at each unit?



Summary Statistics

It is useful to review the summary statistics of our loaded DataFrame. That can be accomplished with the 'describe' DataFrame member function. We transpose to switch the axes for ease of visualization.

```
df.describe().transpose()                                    # visualize summary statistics
```

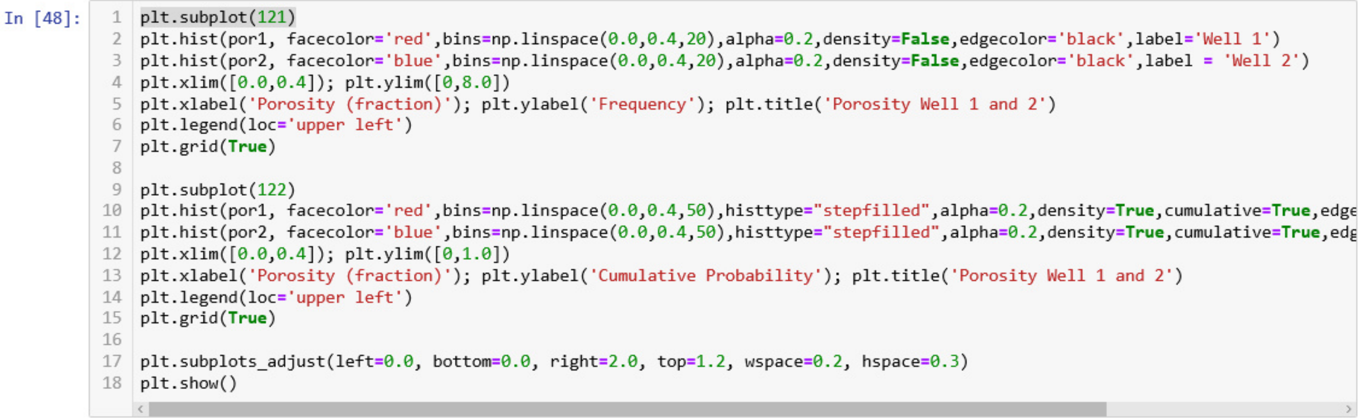
	count	mean	std	min	25%	50%	75%	max
Unit1_Por	20.0	0.1645	0.027810	0.11	0.1500	0.17	0.19	0.21
Unit2_Por	20.0	0.2000	0.045422	0.11	0.1675	0.20	0.23	0.30

- There mean porosity over unit 1 and unit 2 are 16.5% and 20.0%.

Confidence Interval Demonstration

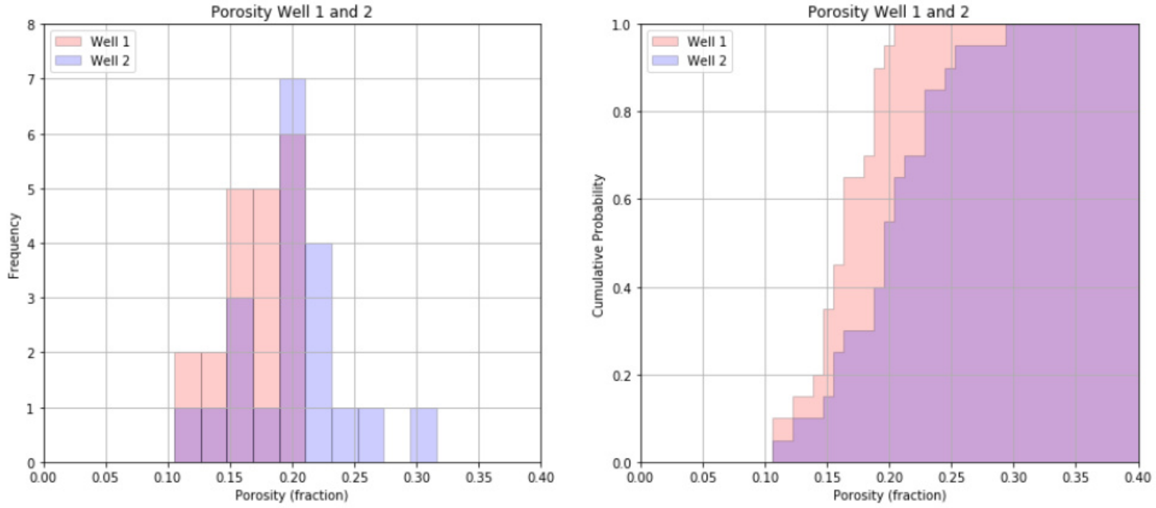
Data visualization

- observe the two distributions from well 1 and 2 porosity together.
- histogram and cumulative distribution function



In [48]:

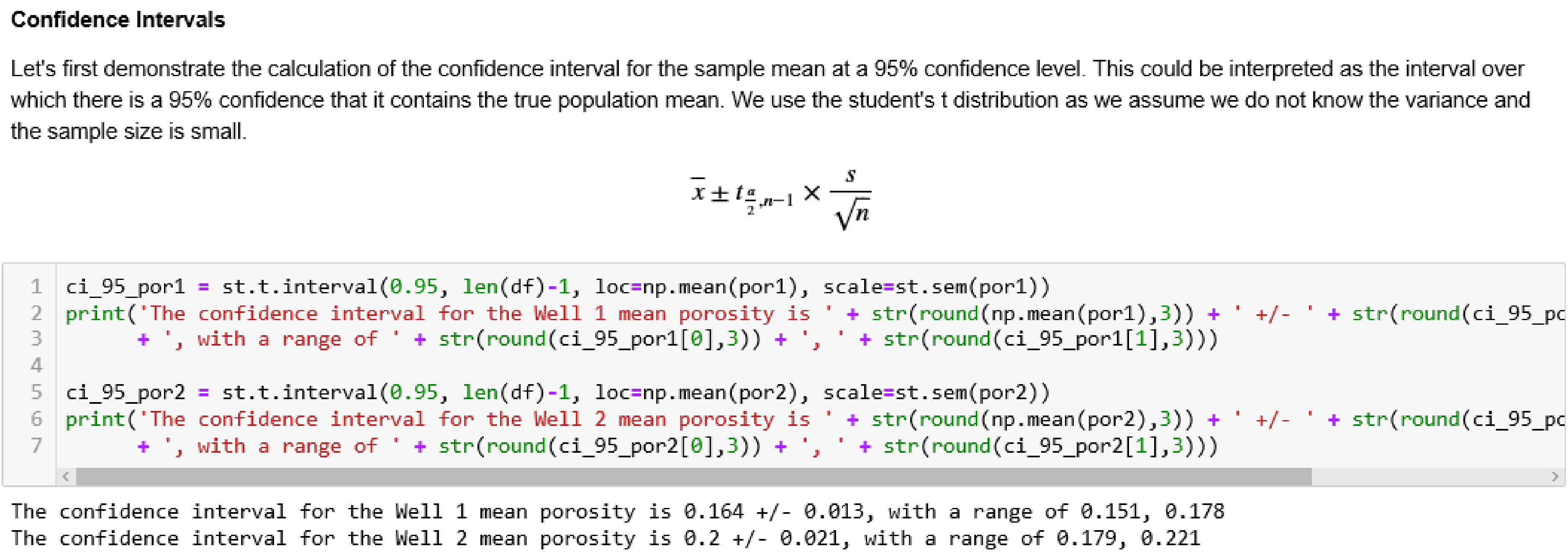
```
plt.subplot(121)
plt.hist(por1, facecolor='red',bins=np.linspace(0.0,0.4,20),alpha=0.2,density=False,edgecolor='black',label='Well 1')
plt.hist(por2, facecolor='blue',bins=np.linspace(0.0,0.4,20),alpha=0.2,density=False,edgecolor='black',label = 'Well 2')
plt.xlim([0.0,0.4]); plt.ylim([0,8.0])
plt.xlabel('Porosity (fraction)'); plt.ylabel('Frequency'); plt.title('Porosity Well 1 and 2')
plt.legend(loc='upper left')
plt.grid(True)

plt.subplot(122)
plt.hist(por1, facecolor='red',bins=np.linspace(0.0,0.4,50),histtype="stepfilled",alpha=0.2,density=True,cumulative=True,edge
plt.hist(por2, facecolor='blue',bins=np.linspace(0.0,0.4,50),histtype="stepfilled",alpha=0.2,density=True,cumulative=True,edg
plt.xlim([0.0,0.4]); plt.ylim([0,1.0])
plt.xlabel('Porosity (fraction)'); plt.ylabel('Cumulative Probability'); plt.title('Porosity Well 1 and 2')
plt.legend(loc='upper left')
plt.grid(True)

plt.subplots_adjust(left=0.0, bottom=0.0, right=2.0, top=1.2, wspace=0.2, hspace=0.3)
plt.show()
```



Confidence Interval Demonstration

Calculuate the condifence intervals for average porosity in each well.

- with scipy we can use a single command to accomplish this.



Confidence Intervals

Let's first demonstrate the calculation of the confidence interval for the sample mean at a 95% confidence level. This could be interpreted as the interval over which there is a 95% confidence that it contains the true population mean. We use the student's t distribution as we assume we do not know the variance and the sample size is small.

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1} \times \frac{s}{\sqrt{n}}$$

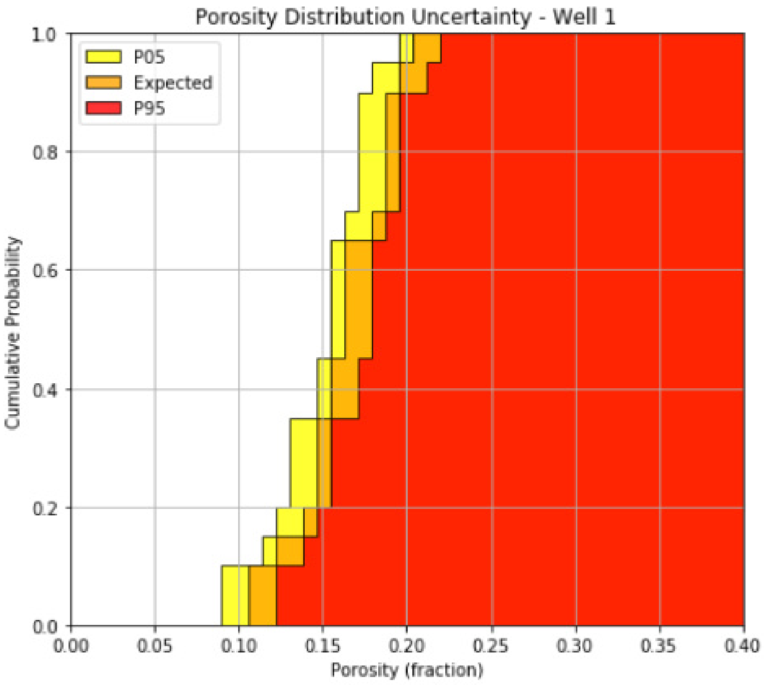
```
ci_95_por1 = st.t.interval(0.95, len(df)-1, loc=np.mean(por1), scale=st.sem(por1))
print('The confidence interval for the Well 1 mean porosity is ' + str(round(np.mean(por1),3)) + ' +/- ' + str(round(ci_95_po
      + ', with a range of ' + str(round(ci_95_por1[0],3)) + ', ' + str(round(ci_95_por1[1],3)))

ci_95_por2 = st.t.interval(0.95, len(df)-1, loc=np.mean(por2), scale=st.sem(por2))
print('The confidence interval for the Well 2 mean porosity is ' + str(round(np.mean(por2),3)) + ' +/- ' + str(round(ci_95_po
      + ', with a range of ' + str(round(ci_95_por2[0],3)) + ', ' + str(round(ci_95_por2[1],3)))
```

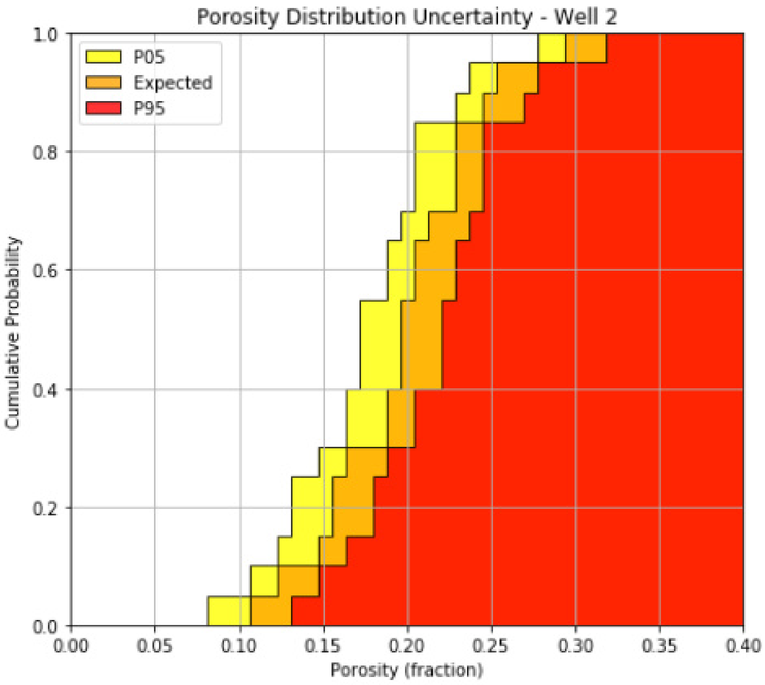
```
The confidence interval for the Well 1 mean porosity is 0.164 +/- 0.013, with a range of 0.151, 0.178
The confidence interval for the Well 2 mean porosity is 0.2 +/- 0.021, with a range of 0.179, 0.221
```

- the t.interval command (alpha level, degrees of freedom, sample mean and standard error in the mean).

Confidence Interval Demonstration

Use the confidence interval of the mean to shift the porosity distribution to calculate the low and high porosity distributions.





- must check for nonphysical values at the tails.

Confidence Intervals Live Demo

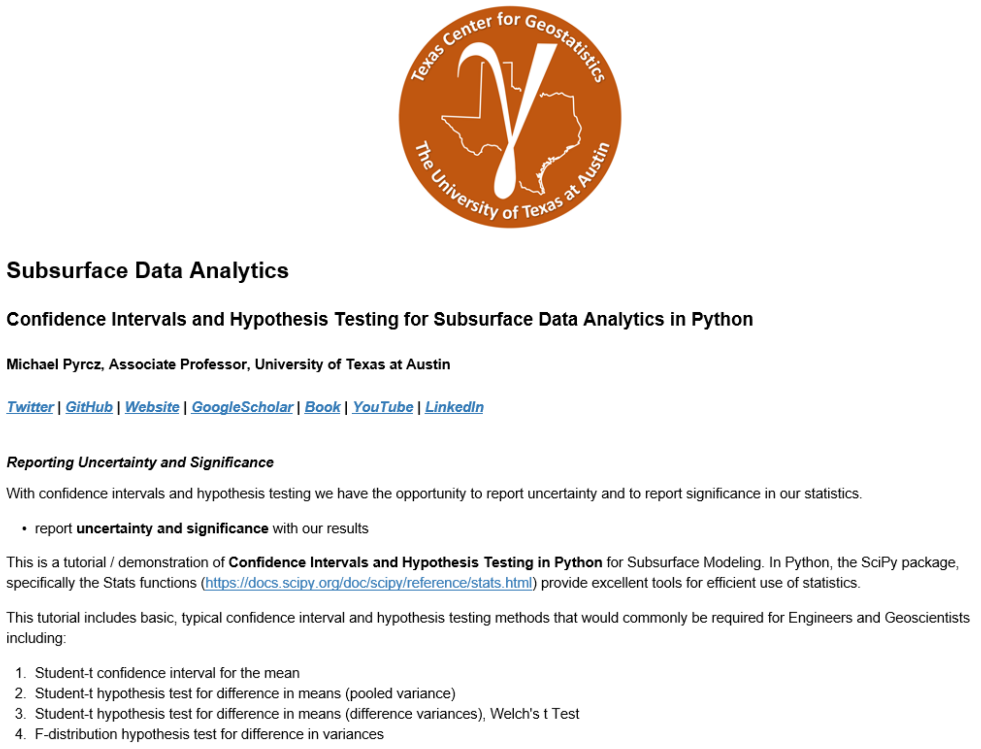
Experiment with

- Data Display
- Confidence Intervals
- Uncertainty Models

in Python Jupyter Notebooks.

Things to try:

1. Change significance level
2. Change histogram display



Subsurface Data Analytics

Confidence Intervals and Hypothesis Testing for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

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

Reporting Uncertainty and Significance

With confidence intervals and hypothesis testing we have the opportunity to report uncertainty and to report significance in our statistics.

- report **uncertainty and significance** with our results

This is a tutorial / demonstration of **Confidence Intervals and Hypothesis Testing in Python** for Subsurface Modeling. In Python, the SciPy package, specifically the Stats functions (https://docs.scipy.org/doc/scipy/reference/stats.html) provide excellent tools for efficient use of statistics.

This tutorial includes basic, typical confidence interval and hypothesis testing methods that would commonly be required for Engineers and Geoscientists including:

1. Student-t confidence interval for the mean
2. Student-t hypothesis test for difference in means (pooled variance)
3. Student-t hypothesis test for difference in means (difference variances), Welch's t Test
4. F-distribution hypothesis test for difference in variances

The file is SubsurfaceDataAnalytics_Confidence_Hypothesis.ipynb at location https://git.io/fjmmJ.

Subsurface Data Analytics and Machine Learning

Data Analytics

Lecture outline . . .

- Hypothesis Testing

Introduction

Data Analytics

Inferential Methods

Predictive Methods

Advanced Methods

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin

Hypothesis Testing

Why do we care?

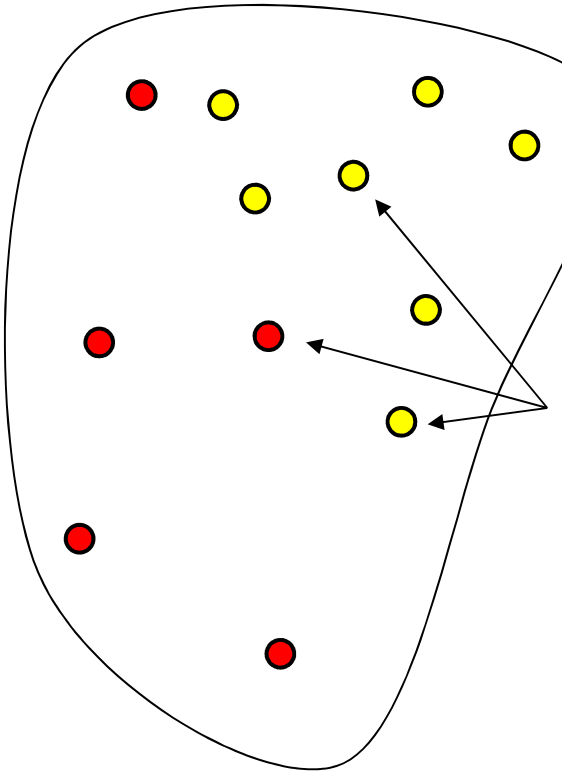
- ‘Difference that makes a difference’ – Sullivan, M.

Hypothesis testing is all about determining if a feature could be due to rand effect!

- Does the resulting porosity distribution form yellow and red wells suggest a different rock population.

These methods should be used along with expert geoscience and engineering judgement and mapping.

AOI - area of interest



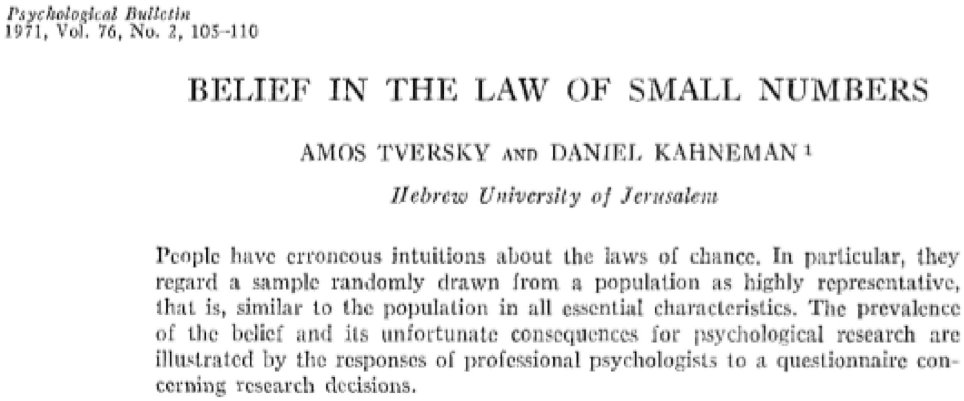
Well-based Porosity

What is the Average porosity over the AOI?

Hypothesis Testing

The Problem – All Differences Look Different!

- Belief in the law of small numbers



Psychological Bulletin
1971, Vol. 76, No. 2, 105–110

BELIEF IN THE LAW OF SMALL NUMBERS

AMOS TVERSKY AND DANIEL KAHNEMAN[1]

Hebrew University of Jerusalem

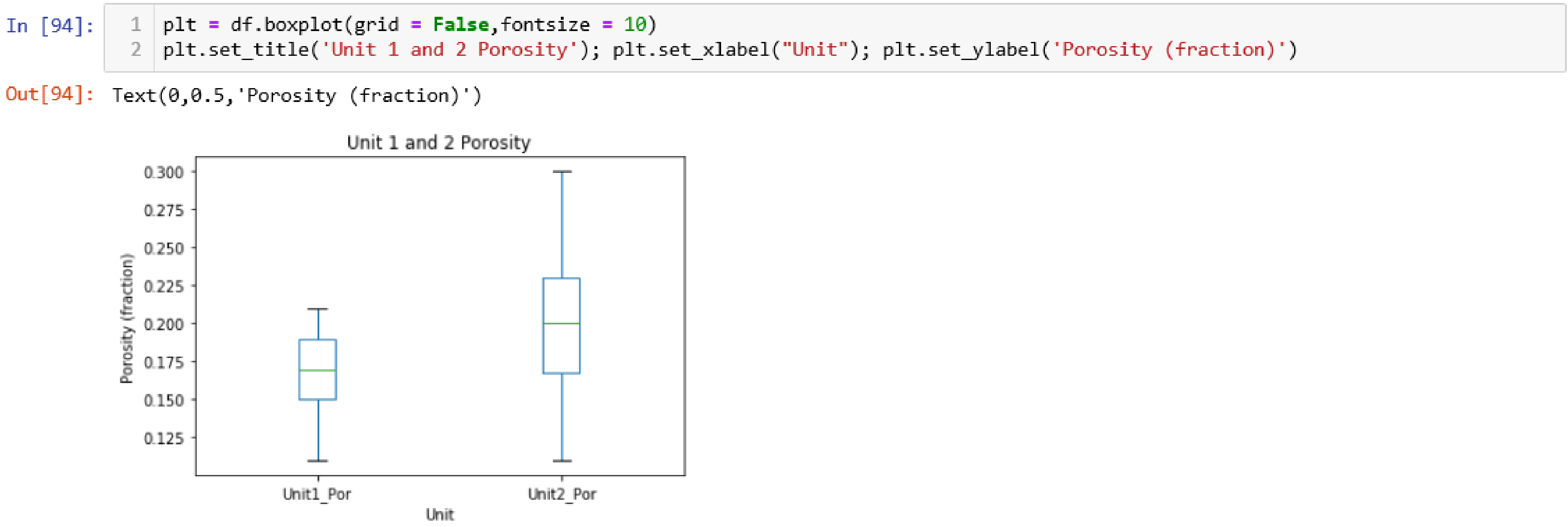
People have erroneous intuitions about the laws of chance. In particular, they regard a sample randomly drawn from a population as highly representative, that is, similar to the population in all essential characteristics. The prevalence of the belief and its unfortunate consequences for psychological research are illustrated by the responses of professional psychologists to a questionnaire concerning research decisions.

  - That samples randomly drawn from a population as highly representative.
  - The mean, variance, P13 will be the same!

- For a small data set this is not the case, confidence intervals and hypothesis testing help us distinguish random form real difference!

Hypothesis Testing Demonstration

We can plot the unit 1 and 2 porosity distributions side by side with a box plot / box and whiskers plot.

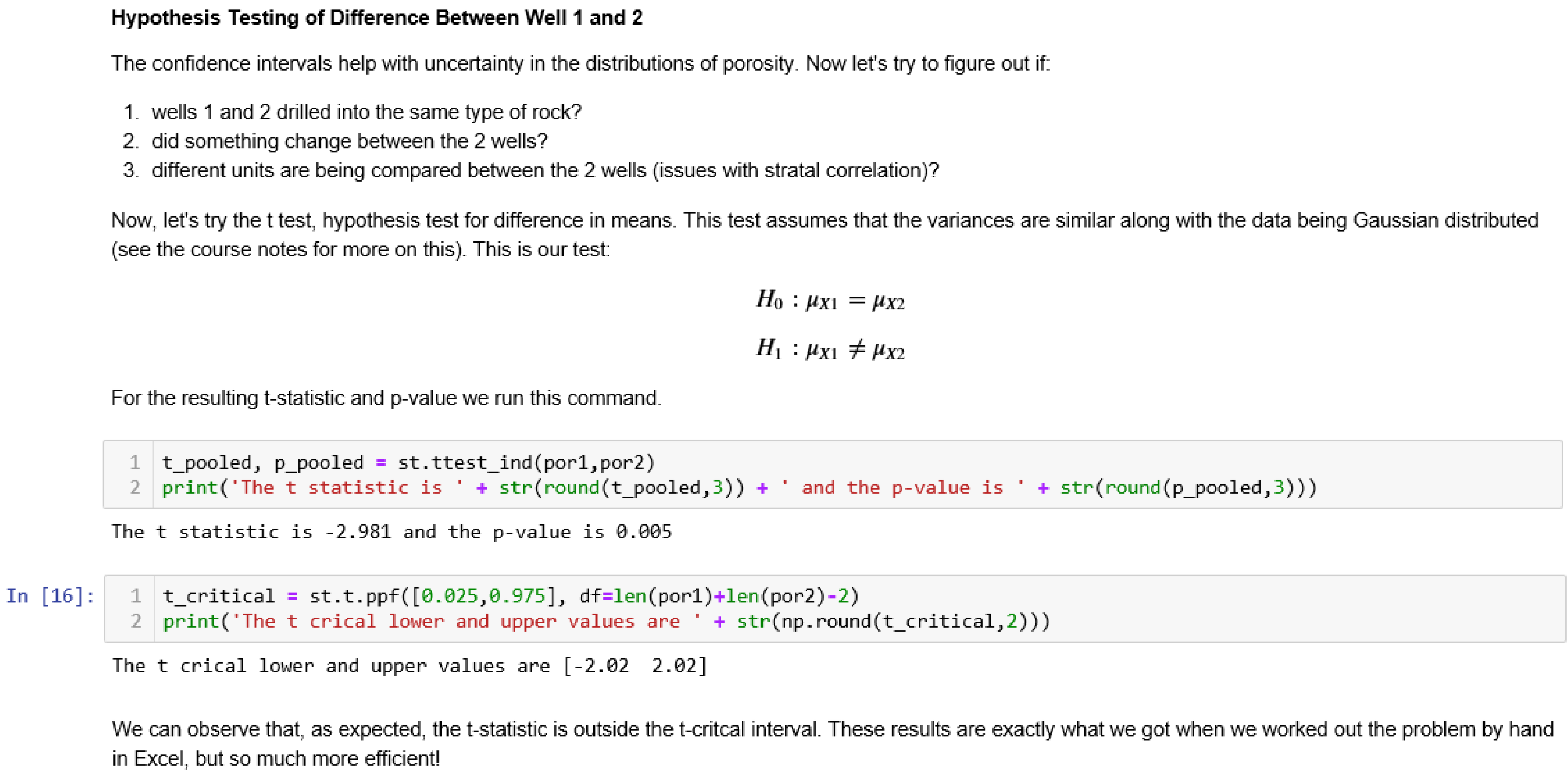


```
In [94]:  plt = df.boxplot(grid = False,fontsize = 10)
          plt.set_title('Unit 1 and 2 Porosity'); plt.set_xlabel("Unit"); plt.set_ylabel('Porosity (fraction)')

Out[94]:  Text(0,0.5,'Porosity (fraction)')
```

- they look quite different, but could it be random effect caused by too few data.

Hypothesis Testing Demonstration

Conduct the hypothesis test to compare the difference in the means between the two wells.



Hypothesis Testing of Difference Between Well 1 and 2

The confidence intervals help with uncertainty in the distributions of porosity. Now let's try to figure out if:

1. wells 1 and 2 drilled into the same type of rock?
2. did something change between the 2 wells?
3. different units are being compared between the 2 wells (issues with stratal correlation)?

Now, let's try the t test, hypothesis test for difference in means. This test assumes that the variances are similar along with the data being Gaussian distributed (see the course notes for more on this). This is our test:

$$H_0 : \mu_{X1} = \mu_{X2}$$

$$H_1 : \mu_{X1} \neq \mu_{X2}$$

For the resulting t-statistic and p-value we run this command.

```
t_pooled, p_pooled = st.ttest_ind(por1,por2)
print('The t statistic is ' + str(round(t_pooled,3)) + ' and the p-value is ' + str(round(p_pooled,3)))
```

```
The t statistic is -2.981 and the p-value is 0.005
```

In [16]:

```
t_critical = st.t.ppf([0.025,0.975], df=len(por1)+len(por2)-2)
print('The t crical lower and upper values are ' + str(np.round(t_critical,2)))
```

```
The t crical lower and upper values are [-2.02  2.02]
```

We can observe that, as expected, the t-statistic is outside the t-critcal interval. These results are exactly what we got when we worked out the problem by hand in Excel, but so much more efficient!

Hypothesis Testing Live Demo

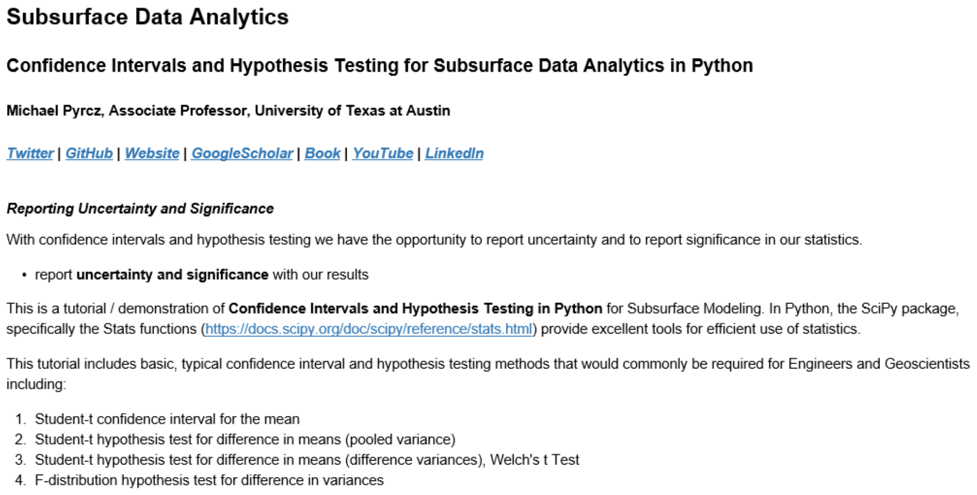
Experiment with

- Hypothesis Testing

in Python Jupyter Notebooks.

Things to try:

1. Change significance level
2. Check out test for difference in variance



Subsurface Data Analytics

Confidence Intervals and Hypothesis Testing for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

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

Reporting Uncertainty and Significance

With confidence intervals and hypothesis testing we have the opportunity to report uncertainty and to report significance in our statistics.

- report **uncertainty and significance** with our results

This is a tutorial / demonstration of **Confidence Intervals and Hypothesis Testing in Python** for Subsurface Modeling. In Python, the SciPy package, specifically the Stats functions (https://docs.scipy.org/doc/scipy/reference/stats.html) provide excellent tools for efficient use of statistics.

This tutorial includes basic, typical confidence interval and hypothesis testing methods that would commonly be required for Engineers and Geoscientists including:

1. Student-t confidence interval for the mean
2. Student-t hypothesis test for difference in means (pooled variance)
3. Student-t hypothesis test for difference in means (difference variances), Welch's t Test
4. F-distribution hypothesis test for difference in variances

The file is SubsurfaceDataAnalytics_Confidence_Hypothesis.ipynb at location https://git.io/fjmmJ.

Subsurface Data Analytics and Machine Learning

Data Analytics

Lecture outline . . .

- **Bootstrap**

Introduction

Data Analytics

Inferential Methods

Predictive Methods

Advanced Methods

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin

Bootstrap Definition

Bootstrap

- method to assess the uncertainty in a sample statistic by repeated random sampling with replacement

Assumptions

- sufficient, representative sampling

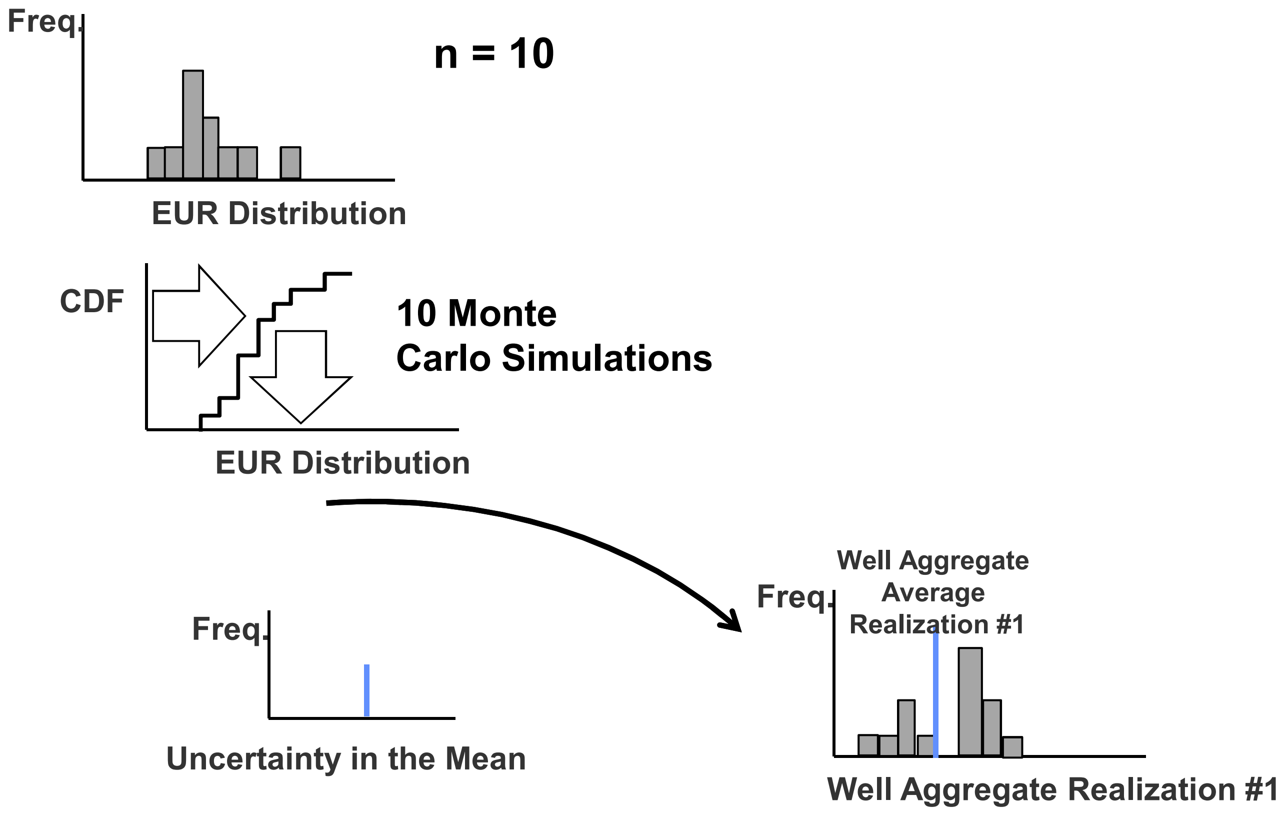
Limitations

- assumes the samples are representative
- assumes stationarity
- only accounts for uncertainty due to too few samples, e.g. no uncertainty due to changes away from data
- does not account for area of interest
- assumes the samples are independent
- does not account for other local information sources

No spatial Context

Univariate Statistics
Bootstrap

Bootstrap for Uncertainty in the Mean



Freq.

n = 10

EUR Distribution

CDF

10 Monte Carlo Simulations

EUR Distribution

Well Aggregate Average Realization #1

Freq.

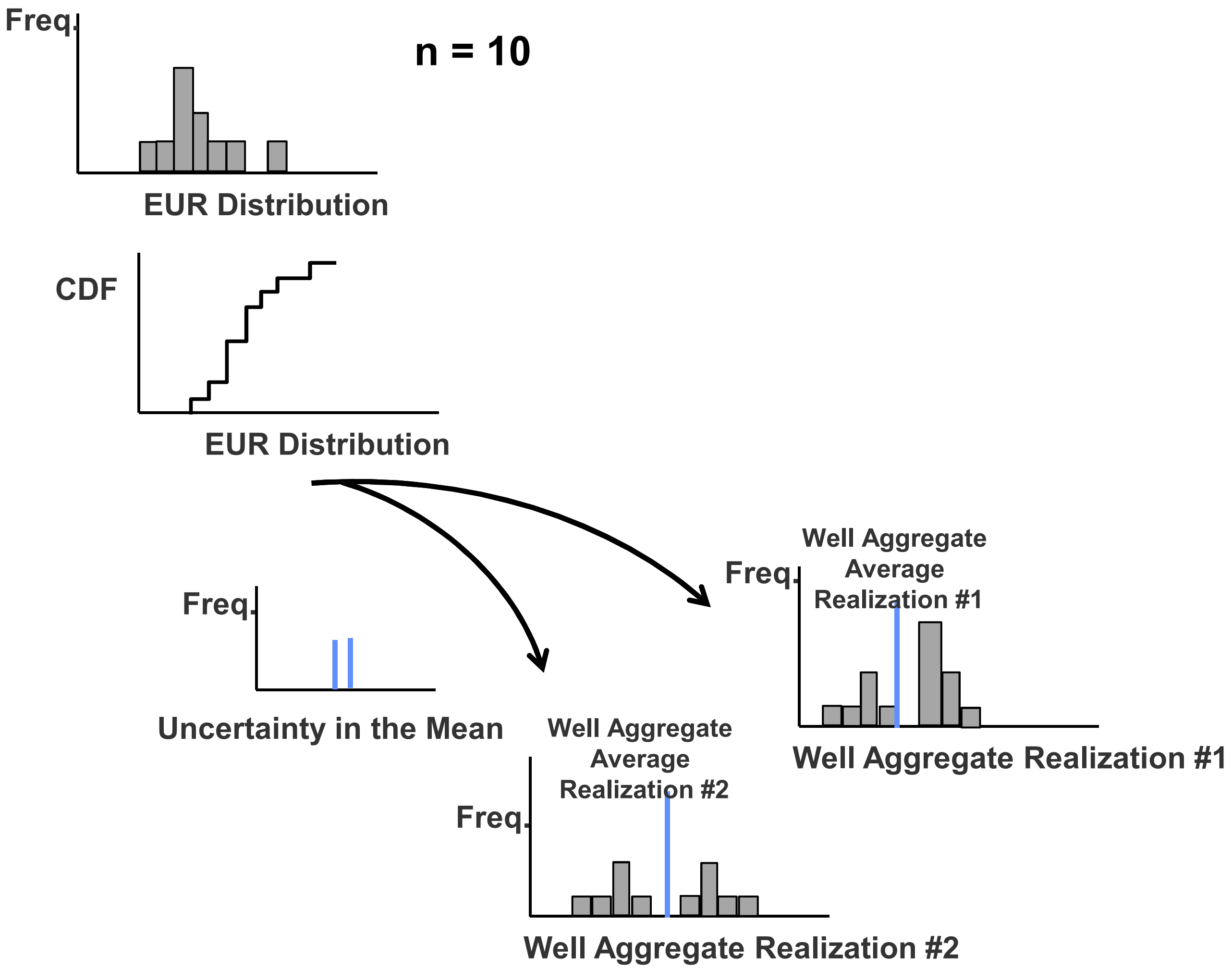
Freq.

Uncertainty in the Mean

Well Aggregate Realization #1

Univariate Statistics
Bootstrap

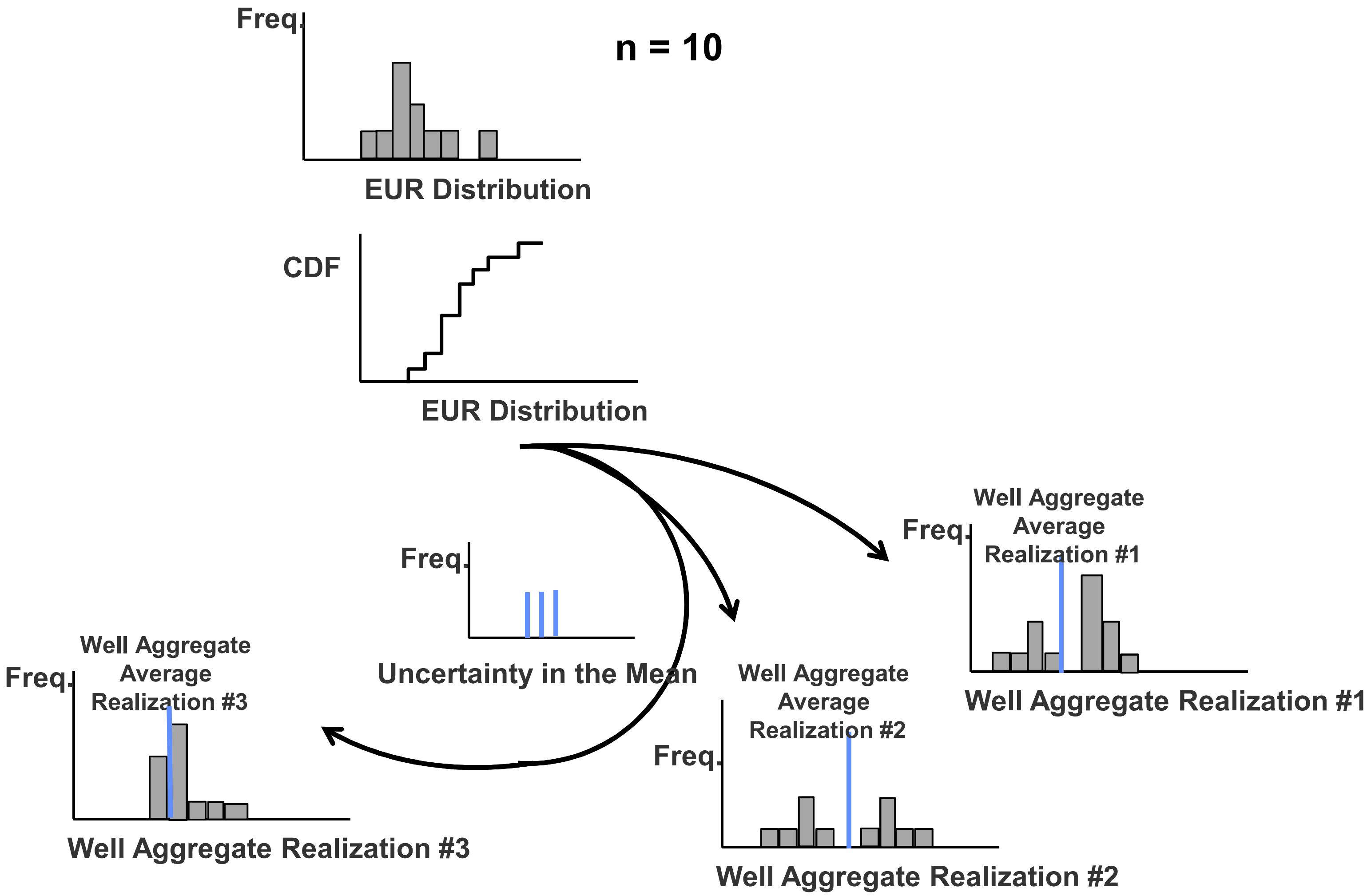
Bootstrap for Uncertainty in the Mean


Freq.
n = 10
EUR Distribution
CDF
EUR Distribution
Freq.
Uncertainty in the Mean
Well Aggregate Average Realization #1
Freq.
Well Aggregate Realization #1
Well Aggregate Average Realization #2
Freq.
Well Aggregate Realization #2

Univariate Statistics

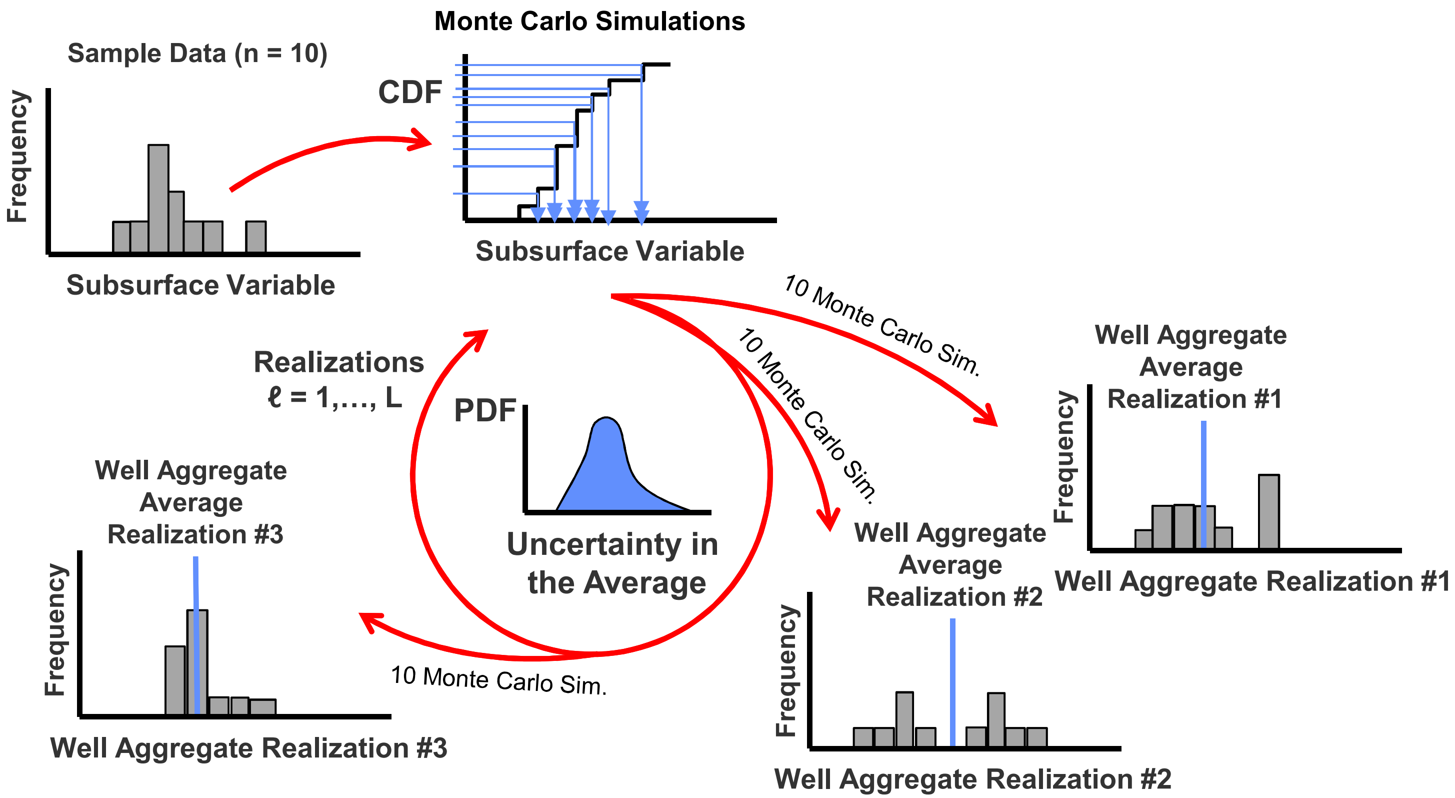
Bootstrap

Bootstrap for Uncertainty in the Mean


Freq.
n = 10
EUR Distribution
CDF
EUR Distribution
Freq.
Uncertainty in the Mean
Freq.
Well Aggregate Average Realization #1
Well Aggregate Realization #1
Freq.
Well Aggregate Average Realization #2
Well Aggregate Realization #2
Freq.
Well Aggregate Average Realization #3
Well Aggregate Realization #3

Univariate Statistics
Bootstrap

Bootstrap for Uncertainty in the Mean


Sample Data (n = 10)
Frequency
Subsurface Variable
Monte Carlo Simulations
CDF
Subsurface Variable
10 Monte Carlo Sim.
10 Monte Carlo Sim.
Realizations
ℓ = 1,…, L
PDF
Uncertainty in the Average
Well Aggregate Average Realization #1
Frequency
Well Aggregate Realization #1
Well Aggregate Average Realization #2
Frequency
Well Aggregate Realization #2
Well Aggregate Average Realization #3
Frequency
Well Aggregate Realization #3
10 Monte Carlo Sim.

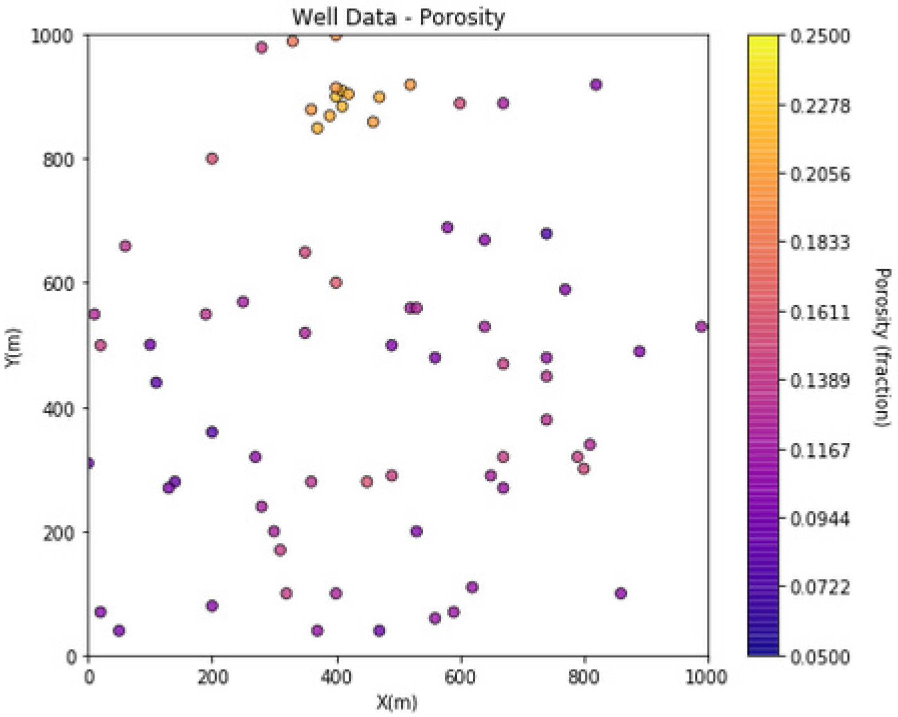
Bootstrap Demonstration

We will work with a complete bootstrap workflow.

1. Load 72 well-averaged porosity
2. Visualize the data
3. Debias the spatial data
4. Bootstrap resampling for uncertainty:
   - mean and standard deviation
   - P10, P50 and P90

Bootstrap Demonstration

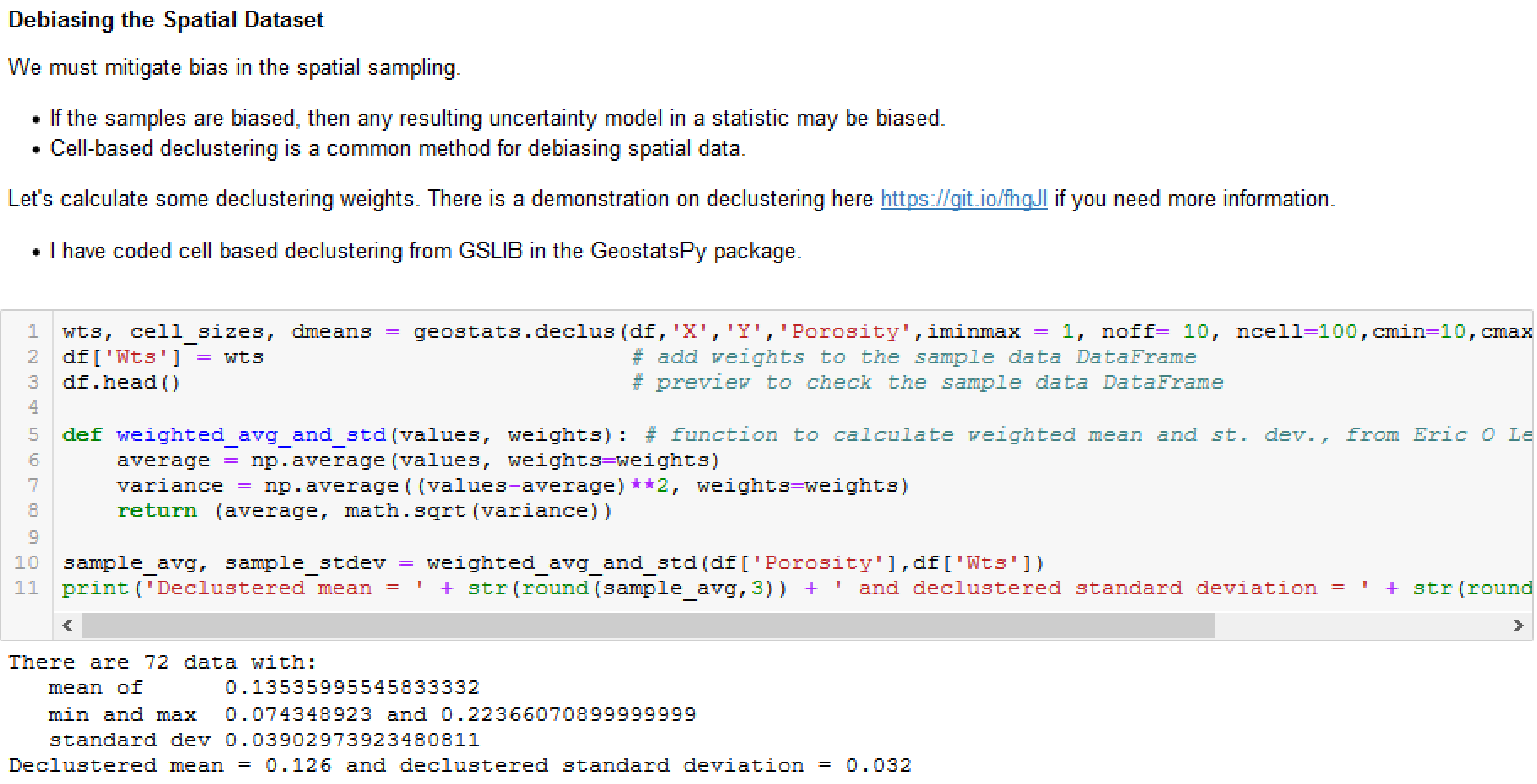
Loading and Visualizing Data



Porosity well averages.

Bootstrap Demonstration

Debiasing to calculate data weights for representative of the subsurface statistics.



Debiasing the Spatial Dataset

We must mitigate bias in the spatial sampling.

- If the samples are biased, then any resulting uncertainty model in a statistic may be biased.
- Cell-based declustering is a common method for debiasing spatial data.

Let's calculate some declustering weights. There is a demonstration on declustering here https://git.io/fhgJl if you need more information.

- I have coded cell based declustering from GSLIB in the GeostatsPy package.

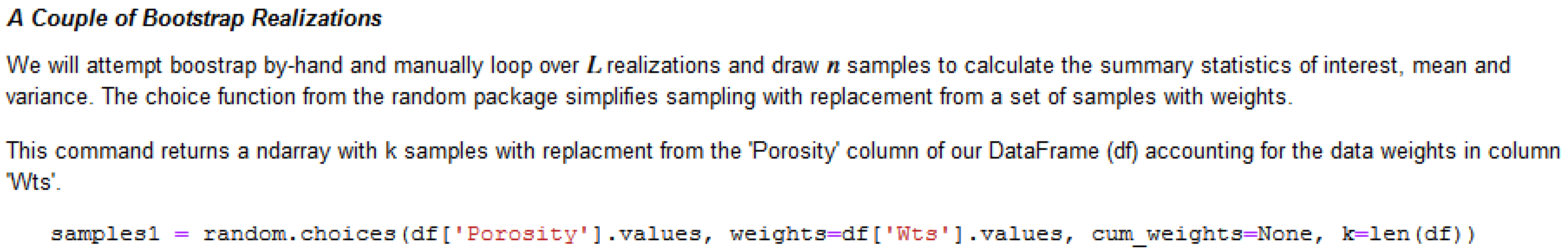
```
wts, cell_sizes, dmeans = geostats.declus(df,'X','Y','Porosity',iminmax = 1, noff= 10, ncell=100,cmin=10,cmax
df['Wts'] = wts                                # add weights to the sample data DataFrame
df.head()                                      # preview to check the sample data DataFrame

def weighted_avg_and_std(values, weights): # function to calculate weighted mean and st. dev., from Eric O Le
    average = np.average(values, weights=weights)
    variance = np.average((values-average)**2, weights=weights)
    return (average, math.sqrt(variance))

sample_avg, sample_stdev = weighted_avg_and_std(df['Porosity'],df['Wts'])
print('Declustered mean = ' + str(round(sample_avg,3)) + ' and declustered standard deviation = ' + str(round
```

```
There are 72 data with:
   mean of      0.13535995545833332
   min and max  0.074348923 and 0.22366070899999999
   standard dev 0.03902973923480811
Declustered mean = 0.126 and declustered standard deviation = 0.032
```

Bootstrap Demonstration

Performing Bootstrap

We have a sampling with replacement method built into ‘numpy’ package.



A Couple of Bootstrap Realizations

We will attempt boostrap by-hand and manually loop over L realizations and draw n samples to calculate the summary statistics of interest, mean and variance. The choice function from the random package simplifies sampling with replacement from a set of samples with weights.

This command returns a ndarray with k samples with replacment from the 'Porosity' column of our DataFrame (df) accounting for the data weights in column 'Wts'.

```
samples1 = random.choices(df['Porosity'].values, weights=df['Wts'].values, cum_weights=None, k=len(df))
```

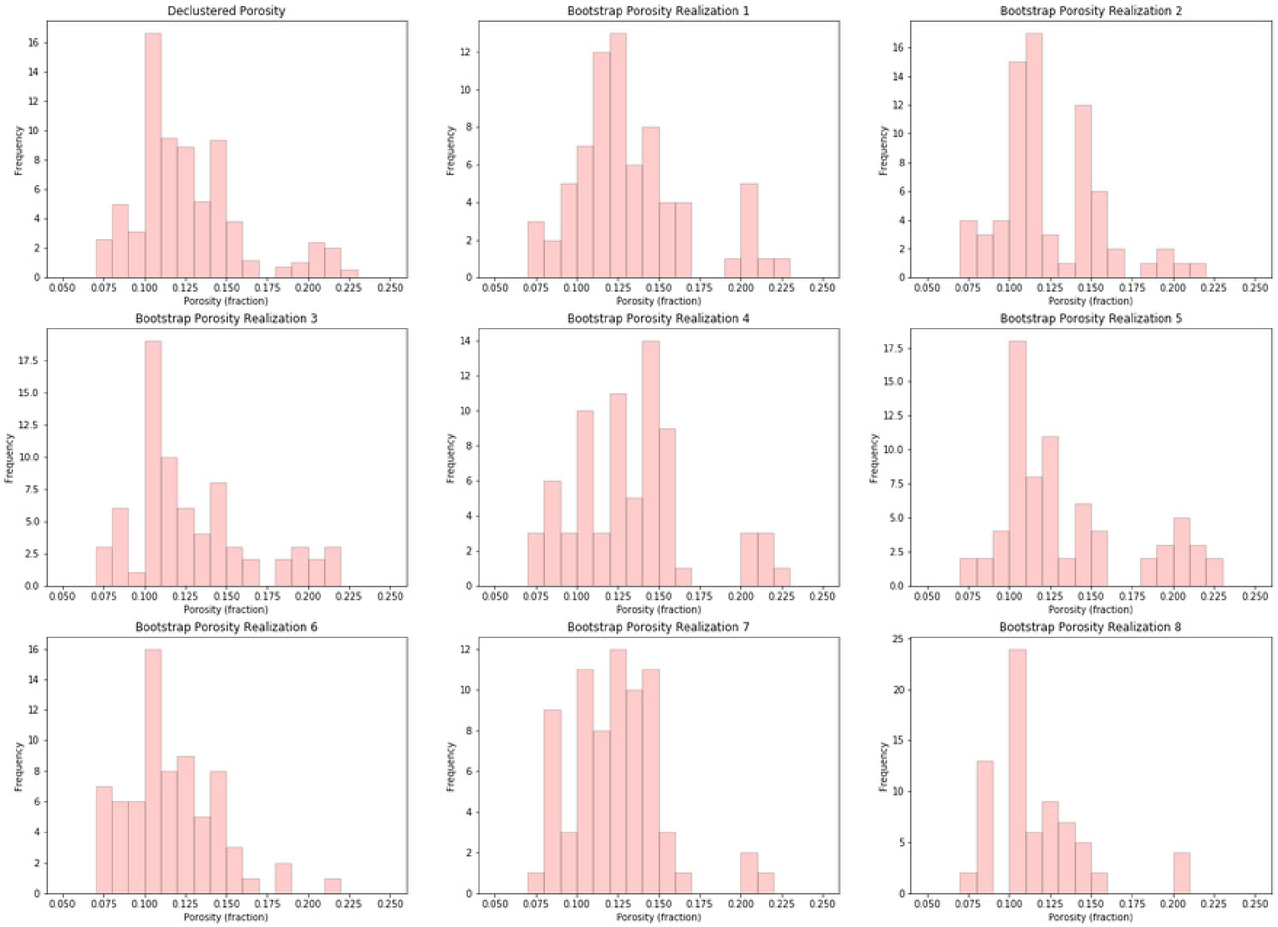
- The method even allows for declustering weights!

- Just provide the data ndarrays and weights (either weights or cumulative weights and ‘k’ number of samples)

Bootstrap Demonstration

Bootstrap Results:

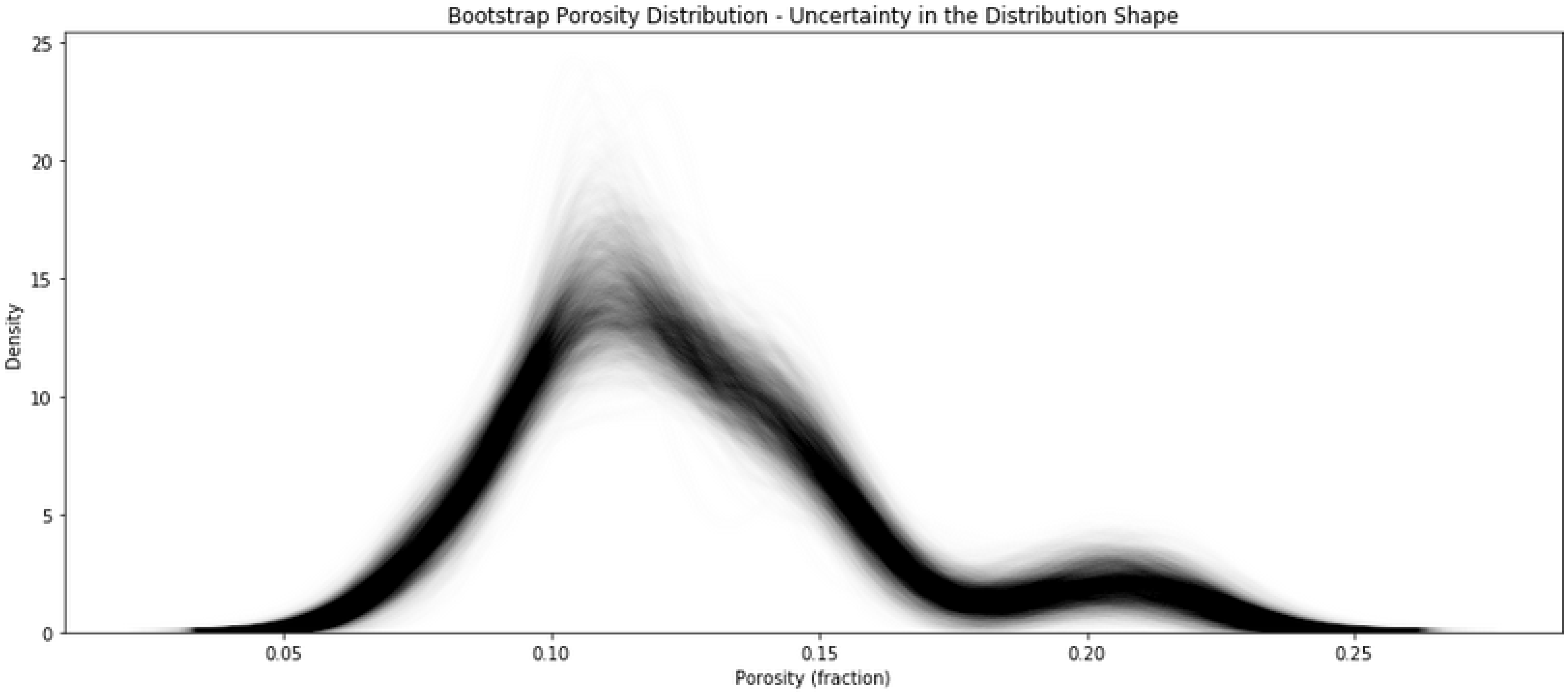
- 9 bootstrap realizations of porosity.



Bootstrap Demonstration

Bootstrap for Uncertainty in the Porosity Distribution

- Here’s 1,500 bootstrap realizations of the porosity distribution.

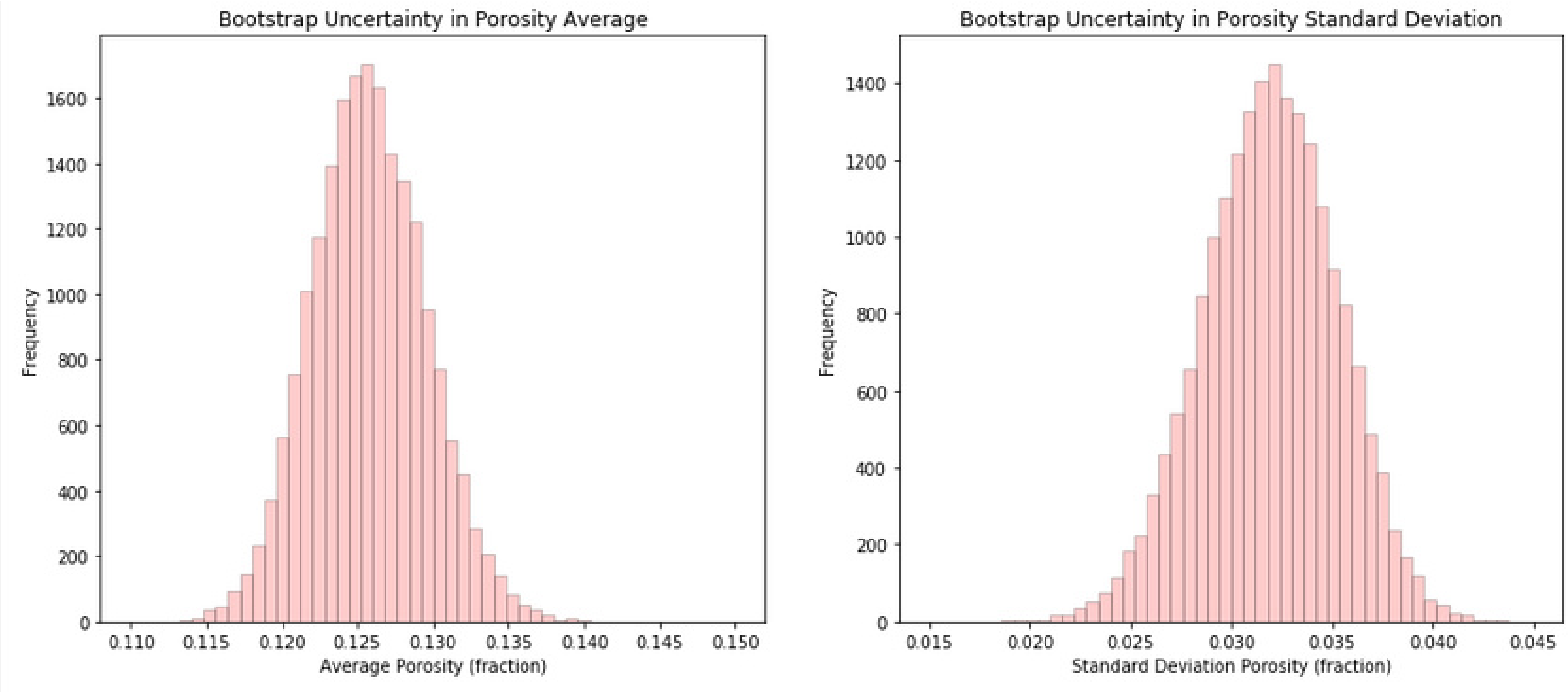


Is the porosity distribution bimodel? Could there be 2 facies?

Bootstrap Demonstration

Bootstrap for Uncertainty in the Porosity Statistics

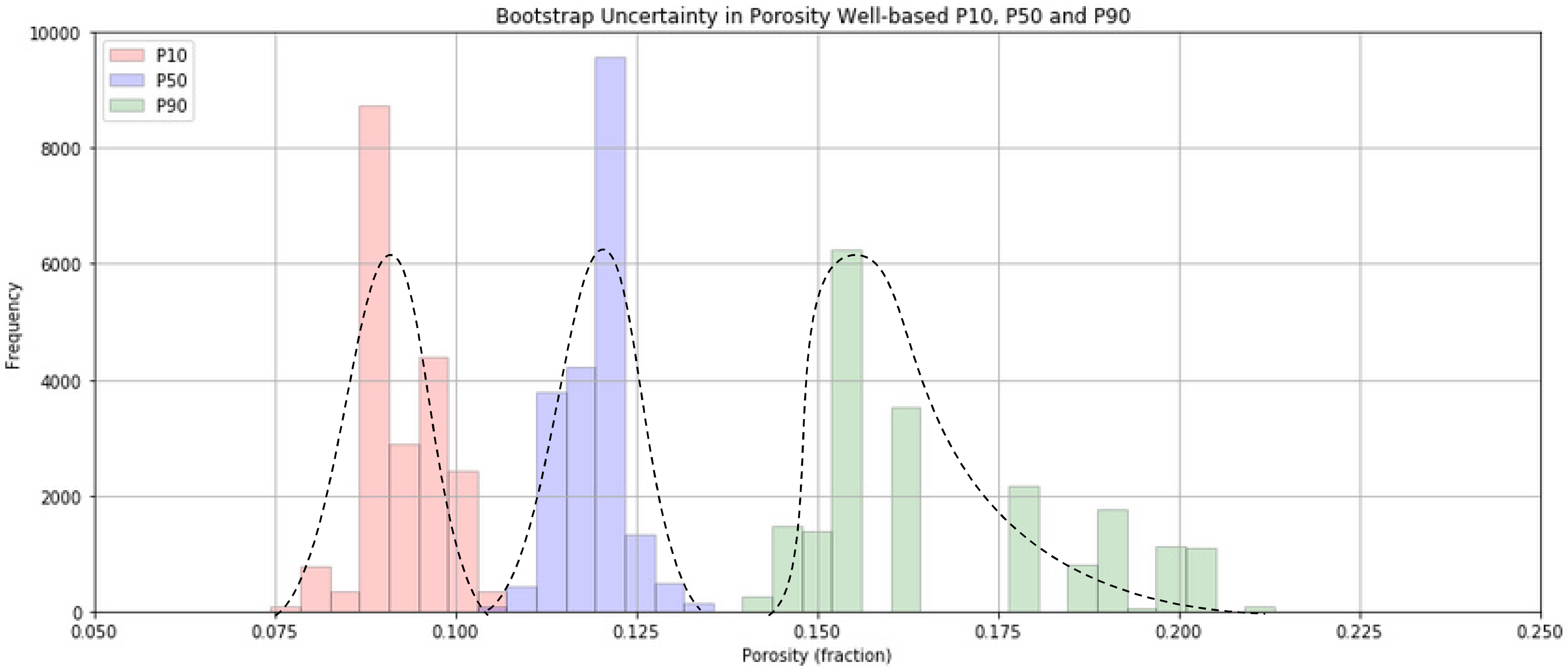
- Uncertainty in the porosity mean and standard deviation.



Bootstrap Demonstration

Bootstrap for Uncertainty in the Porosity Statistics

- Uncertainty in the porosity well-based P10, P50 and P90.



- Could be getting into uncertainty in the uncertainty!

Bootstrap Live Demo

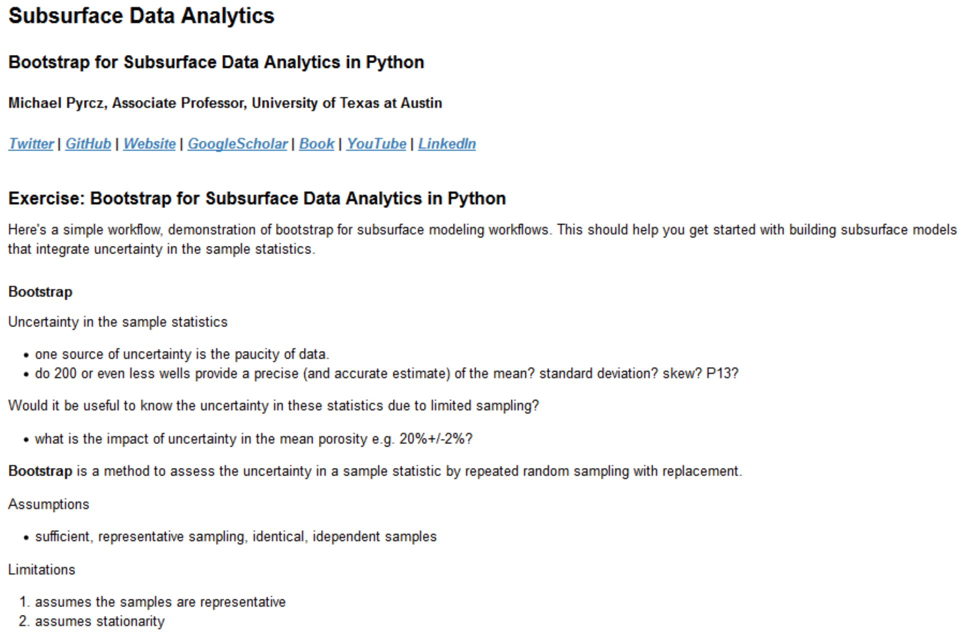
Experiment with

- Bootstrap

in Python Jupyter Notebooks.

Things to try:

1. Change number of samples ‘n’
2. Change number of realizations ‘L’
3. Change the statistic



Subsurface Data Analytics

Bootstrap for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

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

Exercise: Bootstrap for Subsurface Data Analytics in Python

Here's a simple workflow, demonstration of bootstrap for subsurface modeling workflows. This should help you get started with building subsurface models that integrate uncertainty in the sample statistics.

Bootstrap

Uncertainty in the sample statistics

- one source of uncertainty is the paucity of data.
- do 200 or even less wells provide a precise (and accurate estimate) of the mean? standard deviation? skew? P13?

Would it be useful to know the uncertainty in these statistics due to limited sampling?

- what is the impact of uncertainty in the mean porosity e.g. 20%+/-2%?

Bootstrap is a method to assess the uncertainty in a sample statistic by repeated random sampling with replacement.

Assumptions

- sufficient, representative sampling, identical, idependent samples

Limitations

1. assumes the samples are representative
2. assumes stationarity

The file is SubsurfaceDataAnalytics_Bootstrap.ipynb at location https://git.io/fhgUW.

Data Analytics New Tools

Topic	Application to Subsurface Modeling
Confidence Intervals	Use confidence intervals to report uncertainty and propagate it through modeling workflows. *Report uncertainty on every measure.*
Hypothesis Testing	Report the statistical significance of observed differences when possible. *Avoid seeing patterns in random effect due to sparse sampling!*
Bootstrap	Flexible, resampling method for uncertainty in any statistic. *Use bootstrap to quantify and communicate uncertainty due to sparse sampling.*

Reporting uncertainty and significance!’

Subsurface Data Analytics and Machine Learning

Data Analytics

Lecture outline . . .

- Data Analytics
- Confidence Intervals
- Hypothesis Testing
- Bootstrap

Introduction

Data Analytics

Inferential Methods

Predictive Methods

Advanced Methods

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin