Multivariate Modeling: Multivariate

Lecture outline . . .

- Multivariate Analysis
- Joints and Conditionals
- Feature Selection
- Multivariate Estimation

Instructor: Michael Pyrcz, the University of Texas at Austin

Multivariate Modeling: Multivariate

Lecture outline . . .

- Multivariate Analysis

Introduction

Fundamental Concepts

Probability

Data Prep / Analytics

Spatial Continuity / Prediction

Multivariate Modeling

Uncertainty Modeling

Machine Learning

Instructor: Michael Pyrcz, the University of Texas at Austin

Motivation for Multivariate Methods

- **We typically need to build reservoir models of more than one property of interest.**
  - Expanded by whole earh modeling, closing loops with forward models
  - Expanded by unconventionals

- **Subsurface properties may include:**
  - Rock Classification: lithology, architectural elements, facies, depofacies
  - Petrophyscial: porosity, directional permeability, saturuations
  - Geophysical: density, p-wave and s-wave velocity
  - Gemechanical: compressibility / Poisson's ratio, Yong's modulus, brittleness, stress field
  - Paleo- / Time Control: fossil adundances, stratigraphic surfaces, ichnofacies, paleo-flow indicators

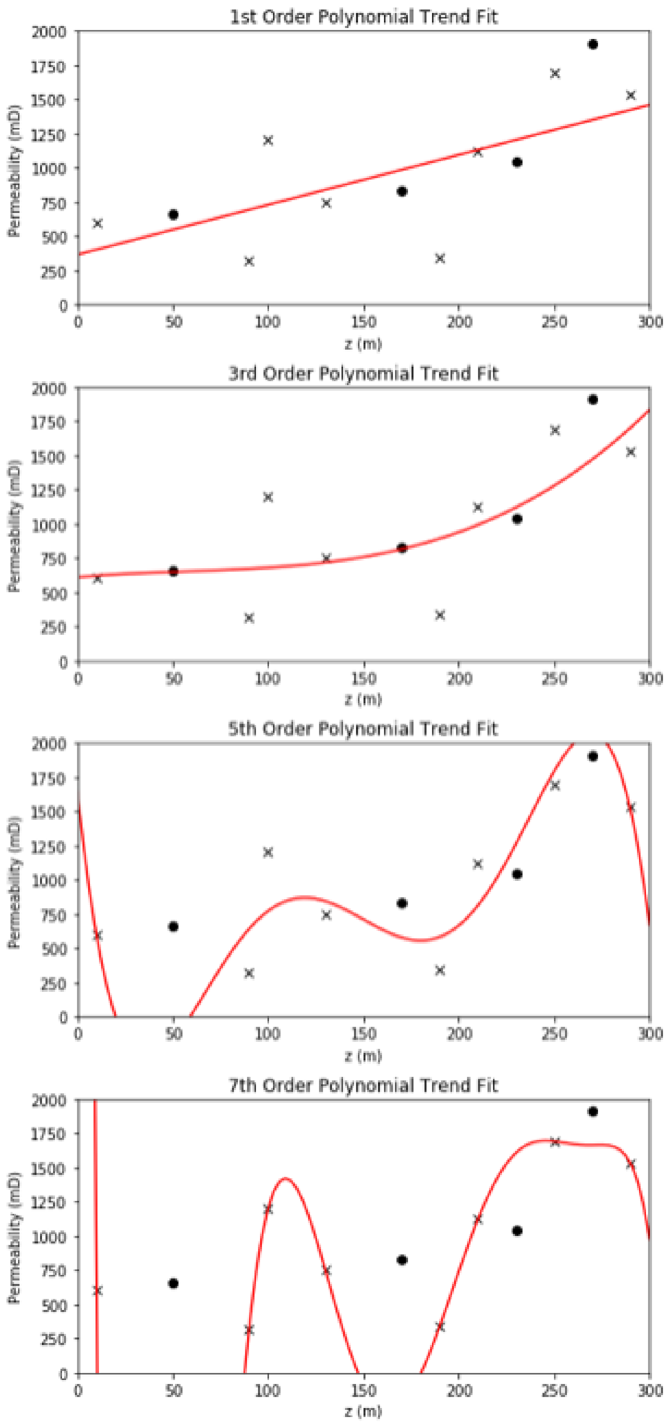
Curse of Dimensionality

Working with more features / variables is harder!

1. More difficult to visualize
2. More data are required to infer the joint probabilities
3. Less coverage
4. More difficult to interrogate / check the model
5. More likely redundant
6. More complicated, more likely overfit

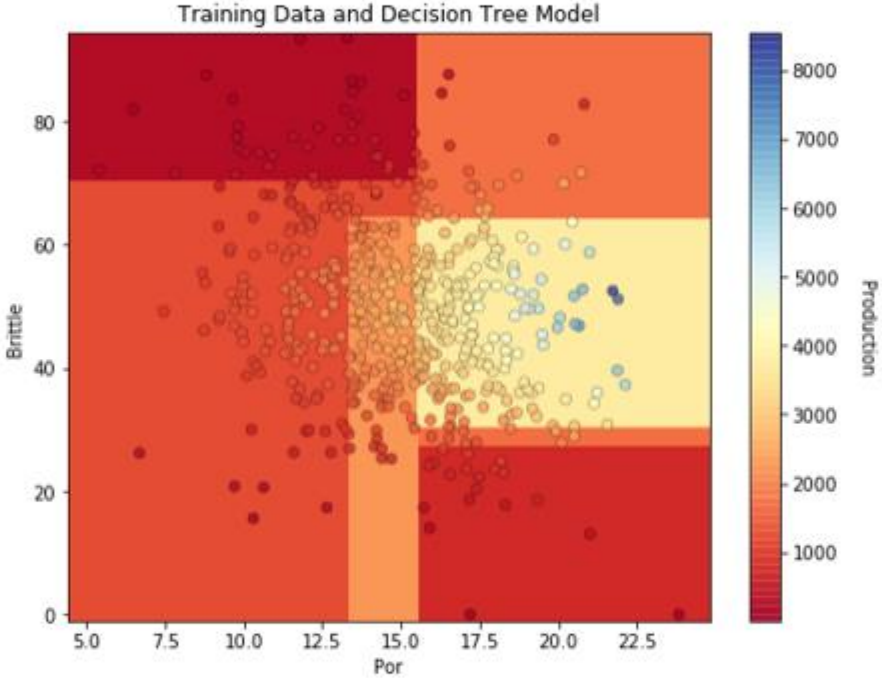
Visualization

- **Consider this simple model:**
  - 1 predictor feature
  - 1 response feature
- How's our model performing?
  - Accuracy in training and testing
- Range of Applicability?
  - Are we extrapolating?
- Overfit
  - Is the model defendable given the data?



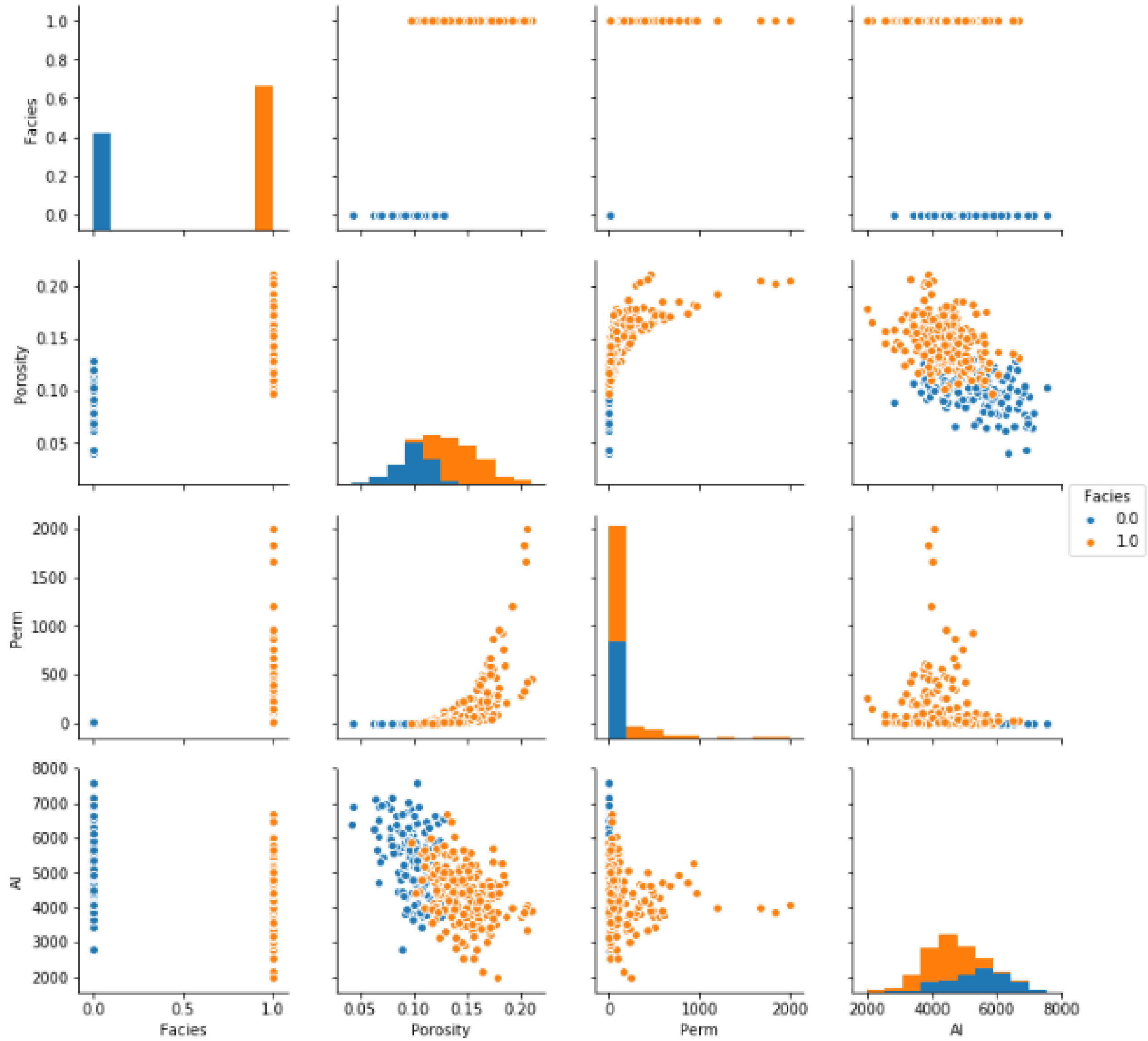
Visualization

- **Consider this simple model:**
  - 2 predictor features
  - 1 response feature
- How's our model performing?
  - Accuracy in training and testing
- Range of Applicability?
  - Are we extrapolating?
- Overfit
  - Is the model defendable given the data?



Visualization

- **Consider this:**
  - 4 predictor features
  - 1 response feature (not shown)
- What are the relationships between features?
- Are there constraints?

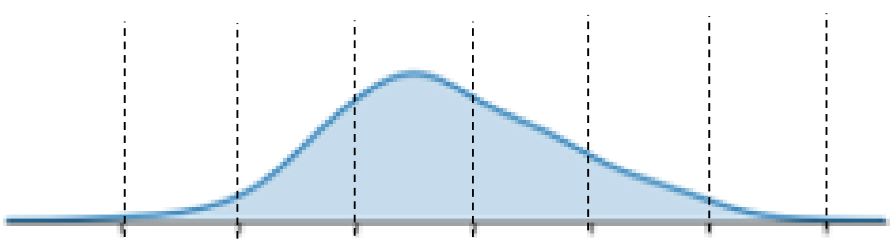


Inferring Joint Probabilities

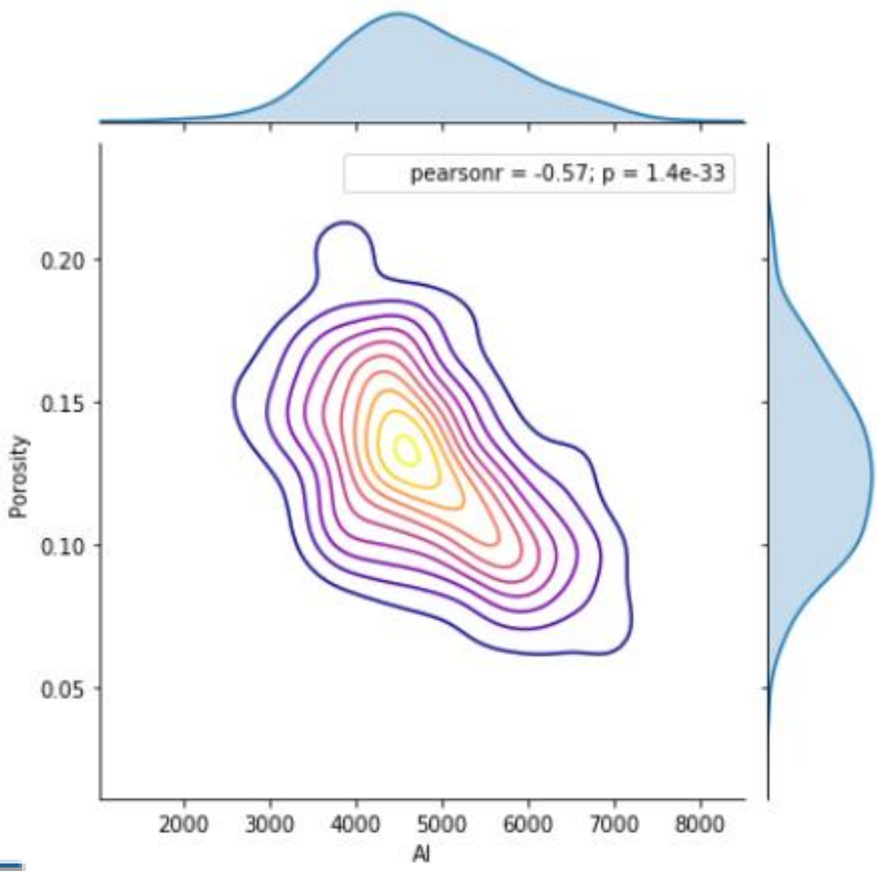
- **Consider any joint probability:**

  $P(X_1 \cap, \ldots, \cap X_m)$ the joint probability of $X_1, \ldots, X_m$

- Let's start with 1 feature (m=1)



$$P\left(X_1^i \leq X \leq X_1^{i+1}\right) = \frac{n(X_1^i \leq X \leq X_1^{i+1})}{n}$$



In each bin we are estimating a probability!
10 data in each bin = 80 data?

Inferring Joint Probabilities

- **Consider any joint probability:**

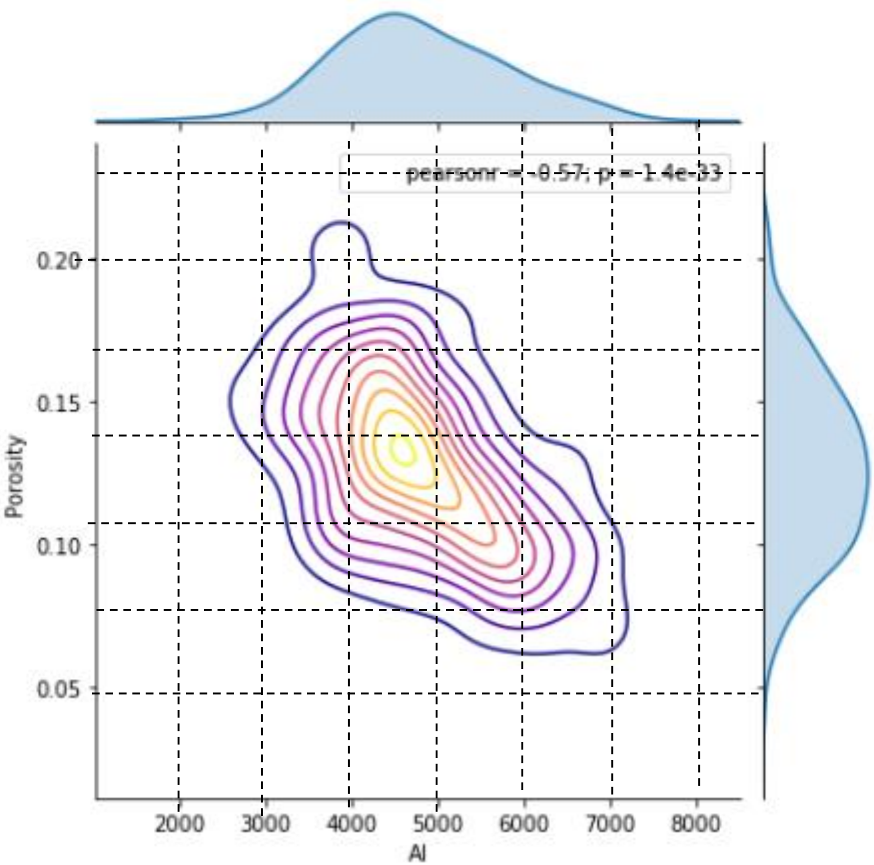
  $P(X_1 \cap, \ldots, \cap X_m)$ the joint probability of $X_1, \ldots, X_m$

- Now move to 2 features (m=2)

$$P\left(X_1^i \le X \le X_1^{i+1}, X_2^j \le X \le X_2^{j+1}\right) = \frac{n(X_1^i \le X \le X_1^{i+1}, X_2^j \le X \le X_2^{j+1})}{n}$$

$$\boldsymbol{n = Data/Bin \cdot Bins^m}$$

- This is optimistic, as it assumes uniform sampling

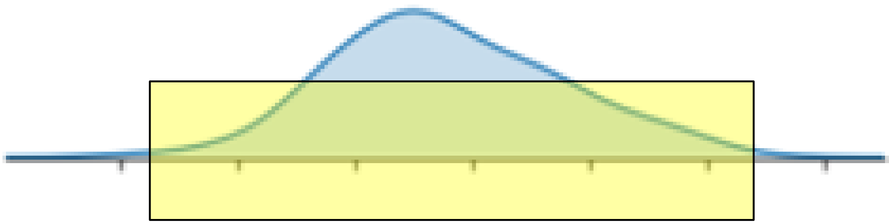


In each bin we are estimating a probability!
10 data in each bin = 640 data?

Coverage

Consider coverage:

- The range of the sample values



- The fraction of the possible solution space that is sampled.

- Let’s return to 1 feature, and assume 80% coverage!

- That’s pretty good right?

Coverage

Consider coverage:

- Now let's move to 2 features, each with 80% coverage

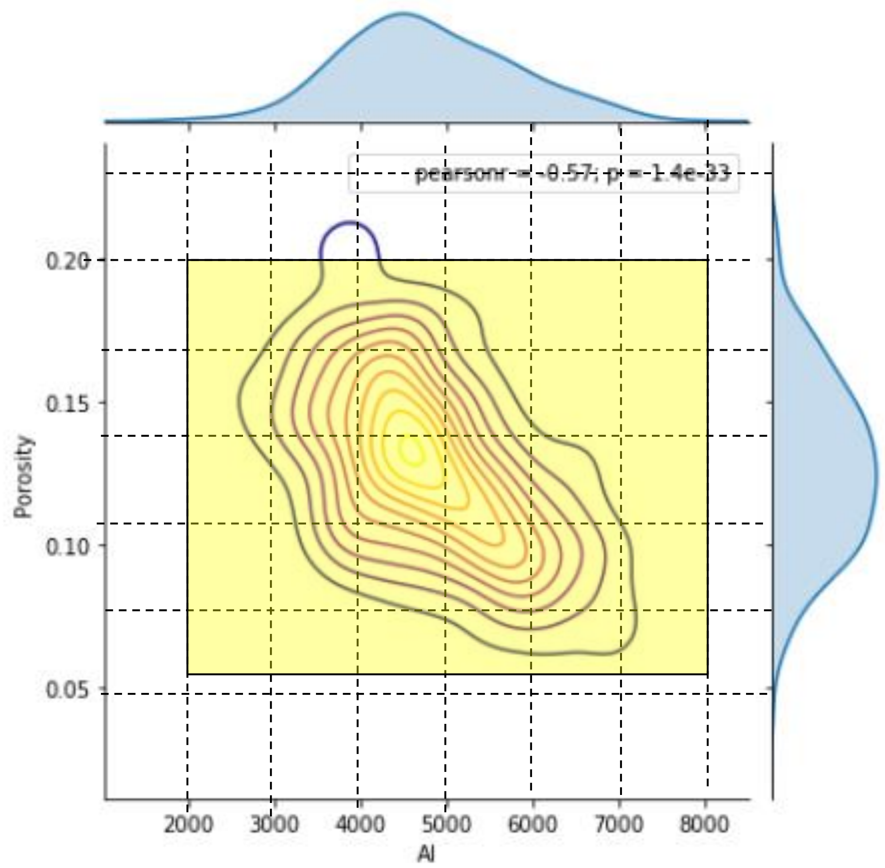
- How much of the solution space is covered?

$$0.8^D, \quad e.g.\, 0.8^2 = 0.64$$

- Even with exponential increase in number of data:

$$\boldsymbol{n = Data/Bin \cdot Bins^m}$$

coverage is decreasing as we increase the number of features!



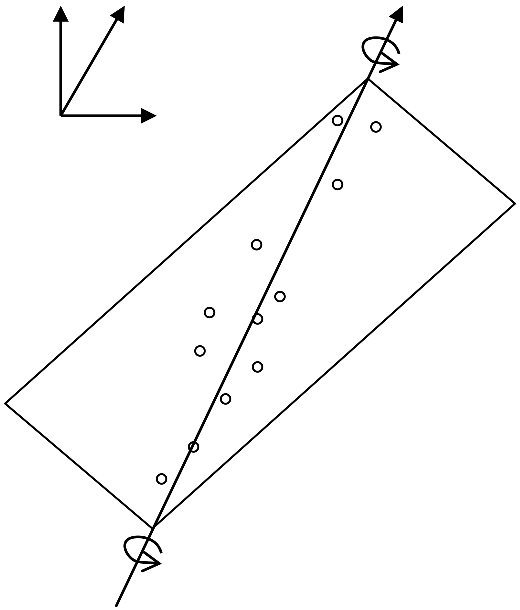
Multicollinearity
Feature Redundancy

"the existence of such a **high degree of correlation between supposedly independent variables** being used to estimate a dependent variable that the contribution of each independent variable to variation in the dependent variable cannot be determined"

- Merriam-Webster Online Dictionary

"In statistics, **multicollinearity** (also collinearity) is a phenomenon in which one predictor variable in a **multiple regression** model can be linearly predicted from the others with a substantial degree of accuracy."

- Wikipedia



It is like fitting a plane to a line!

Motivation for Multivariate Methods

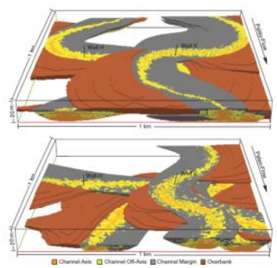
- **A Confession:**
  - Standard geostatistical workflows are bivariate at most
    - » e.g. simulate permeability conditional to porosity

Categorical Simulation

Models Facies PDF, Trends and Spatial Continuity

→

Multiple Facies Realizations



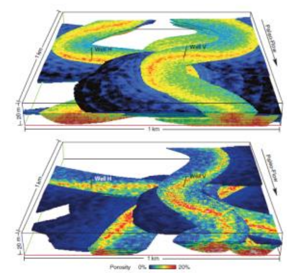
→

Continuous Simulation

Models Porosity PDF, Trends and Spatial Continuity within each facies

→

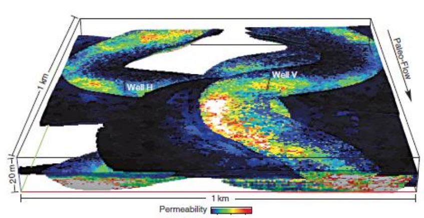
Multiple Porosity By-facies Realizations



Continuous Cosimulation

Models Permeability PDF, Trends and Spatial Continuity and Porosity Relationship within each facies

Multiple Permeability By-facies Realizations Correlated with Porosity



Note: only had 1 realization on hand (should be two in figure).

Motivation for Multivariate Methods

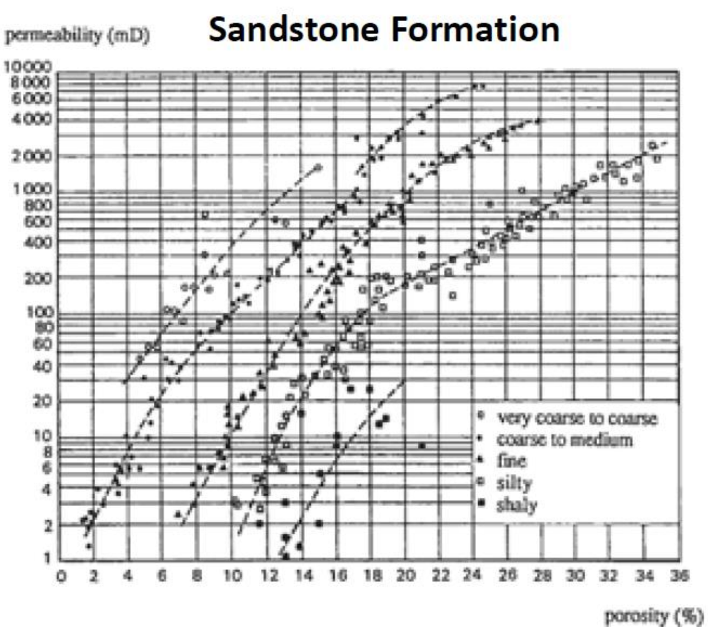
- **Emerging Multivariate Methods Include:**
  - Transforms – remove correlations and then model with independent variables and then back-transform to restore correlation (e.g. step-wise conditional transform).

This is beyond the scope of this course.

Bivariate Statistics
What is Bivariate Analysis?

- **Bivariate Analysis: Understand and Quantify the relationship between two variables**
  - Example: Relationship between porosity and permeability
  - How can we use this relationship?

Scatter Plot



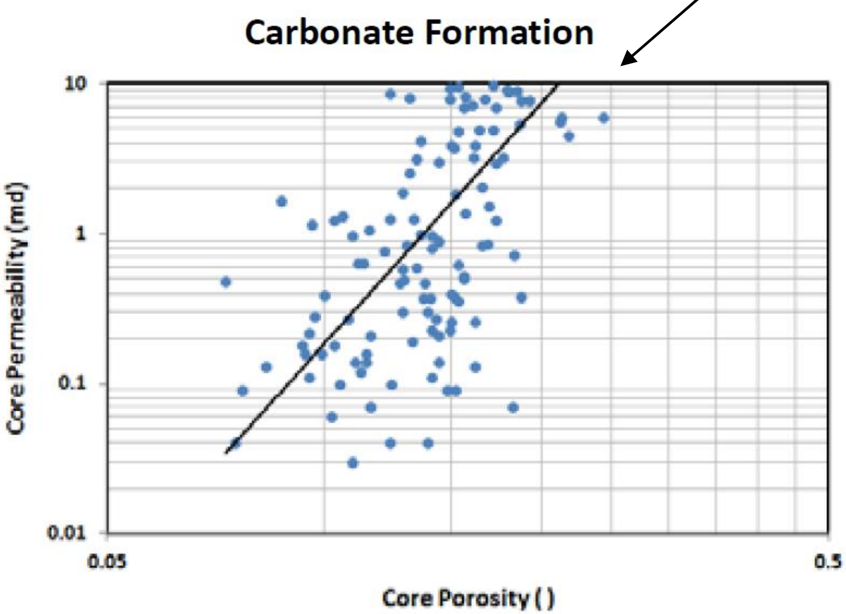
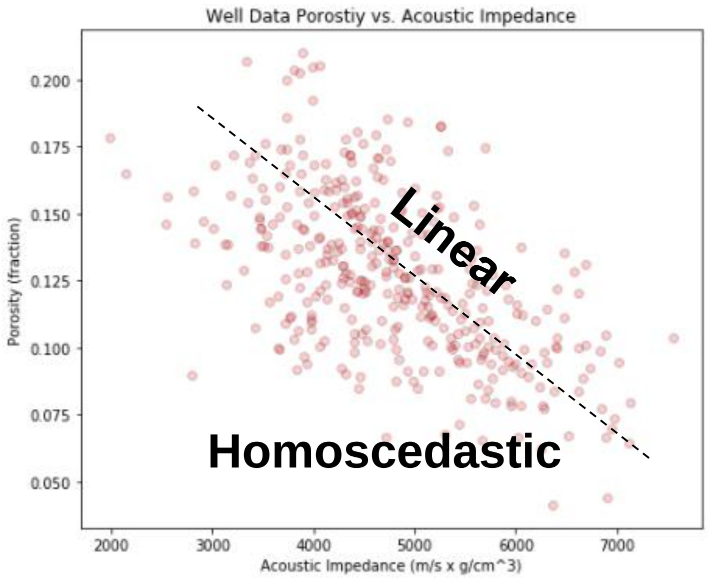


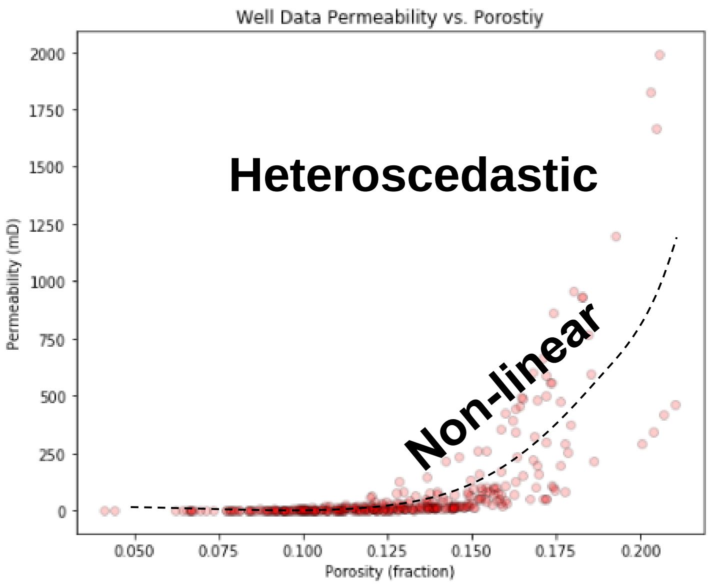
Figure from Peters, E. J., 2012, Advanced Petrophysics.

Slide from Dr. Zoya Heidari's PGE 337 Course

Bivariate Statistics
What is Bivariate Analysis?

- **Examples of bivariate structures**

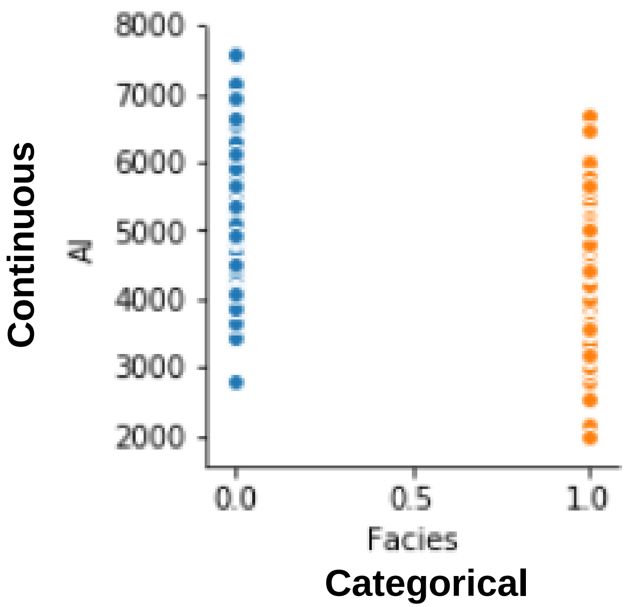


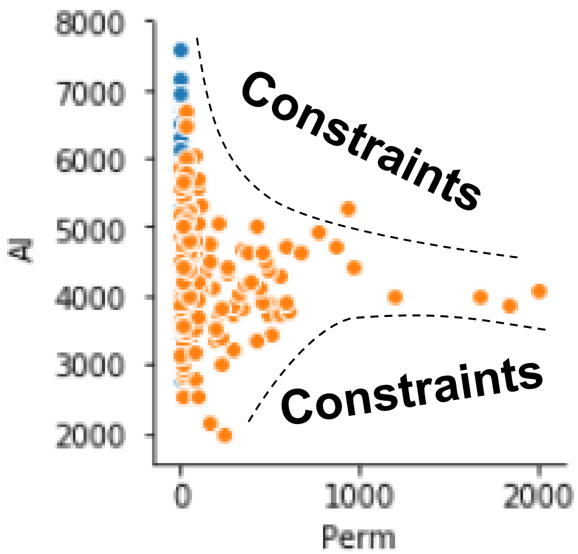


- Linear / Nonlinear – shape of the conditional expectation Y | X
- Homoscedastic / Heteroscedastic – conditional variance of Y | X

Bivariate Statistics
What is Bivariate Analysis?

- **Examples of bivariate structures**





- Categorical variables only have a specified number of possible outcomes, continuous takes on a range of possible outcomes.
- Constraints – specific combinations of variables are not possible.

Bivariate Statistics
Pearson's Correlation Coefficient

- **Definition: Pearson's Product-Moment Correlation Coefficient**
  - **Provides a measure of the degree of linear relationship.**

means of variables x and y

$$\rho_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)\sigma_x\sigma_y}, -1.0 \leq \rho_{xy} \leq 1.0$$

Correlation coefficient of variables x and y

number of data pairs

standard deviation of variables x and y

  - **Correlation coefficient is a standardized covariance.**

$$C_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)}$$

Covariance

$$\rho_{xy} = \frac{C_{xy}}{\sigma_x\sigma_y}$$

- **We can see that covariance and variance are related.**
  - **Replace the second term in the square with another variable.**
  - **Covariance:**

$$\boldsymbol{C_{xy}} = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

A measure of how 2 variables vary together.

- **Variance:**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})$$

A measure of how 1 variable varies with itself.

Bivariate Statistics
Spearman's Rank Correlation Coefficient

- **Definition: Spearman's Rank Correlation Coefficient**
  - Provides a measure of the degree of monotonic relationship.

means of rank transform of variables x and y

$$\rho_{R_x,R_y} = \frac{\sum_{i=1}^{n}(R_{x_i} - \overline{R_x})(R_{y_i} - \overline{R_y})}{(n-1)\sigma_{R_x}\sigma_{R_y}}, -1.0 \le \rho_{xy} \le 1.0$$

Rank correlation coefficient of variables x and y

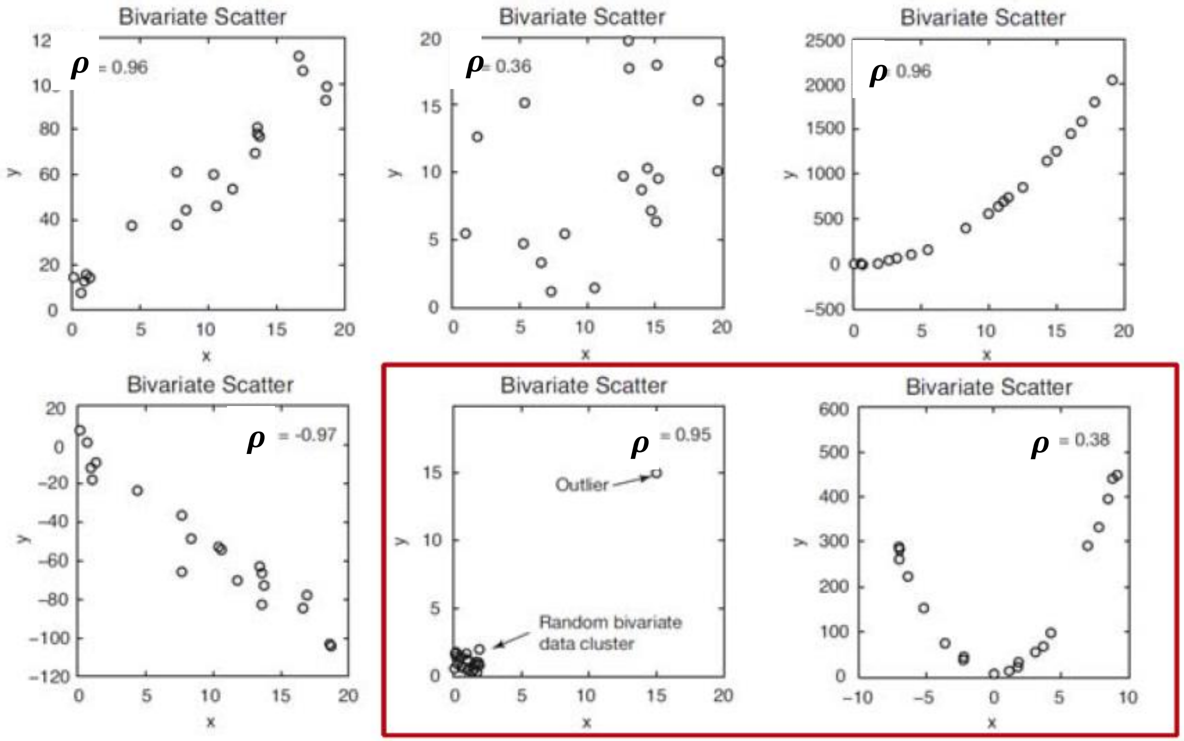
number of data pairs

standard deviation of Rank transform of variables x and y

  - Rank transform, e.g. R_{x_i}, sort the data in ascending order and replace the data with the index, $i = 1, \ldots, n$.
  - Spearman's rank correlation coefficient is more robust in the presence of outliers and some nonlinear features than the Pearson's correlation coefficient

Bivariate Statistics
Pearson's Correlation Coefficient

- **Interpreting the correlation coefficient**



Is Pearson's correlation coefficient a reliable measure of correlation in these cases?

Bivariate Statistics
Correlation and Causation

- Correlation does not imply causation!
  - We require a “true experiment” where one variable is manipulated and others are rigorously controlled!

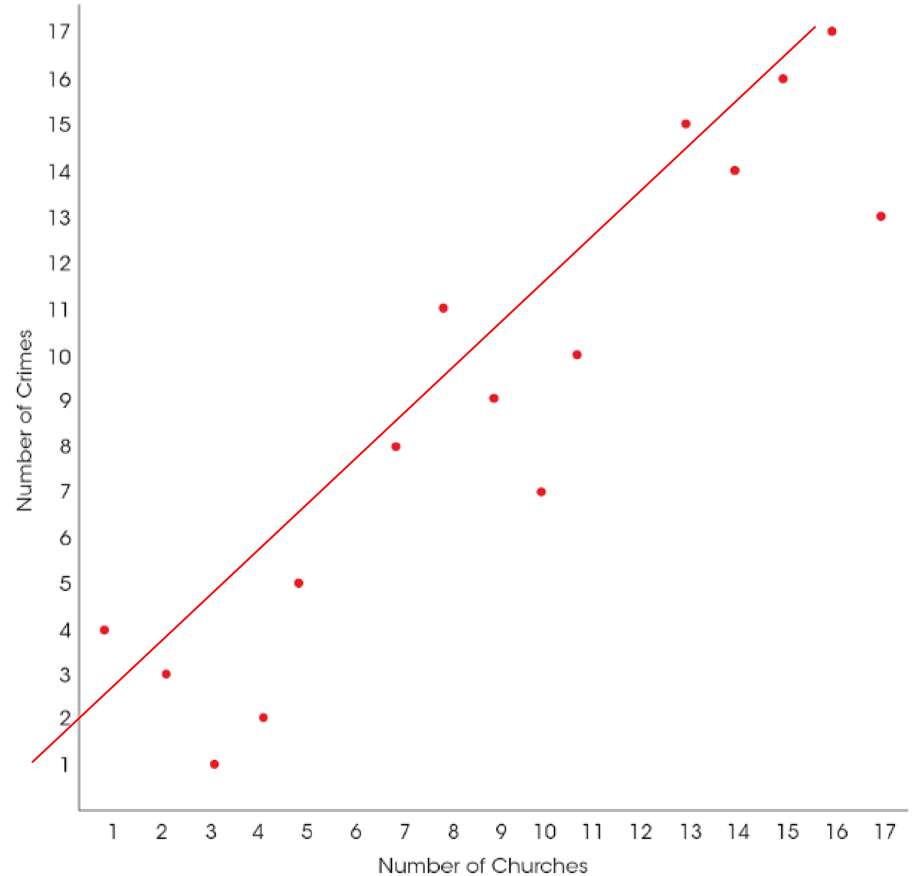


FIGURE 15.10
Hypothetical data showing the logical relationship between the number of churches and the number of crimes for three groups of cities: those with small populations ($Z = 1$), those with medium populations ($Z = 2$), and those with large populations ($Z = 3$).

Bivariate Statistics
Correlation and Causation

- Correlation does not imply causation!
    - Population was not controlled!
    - For each size of city the correlation is nearly zero.

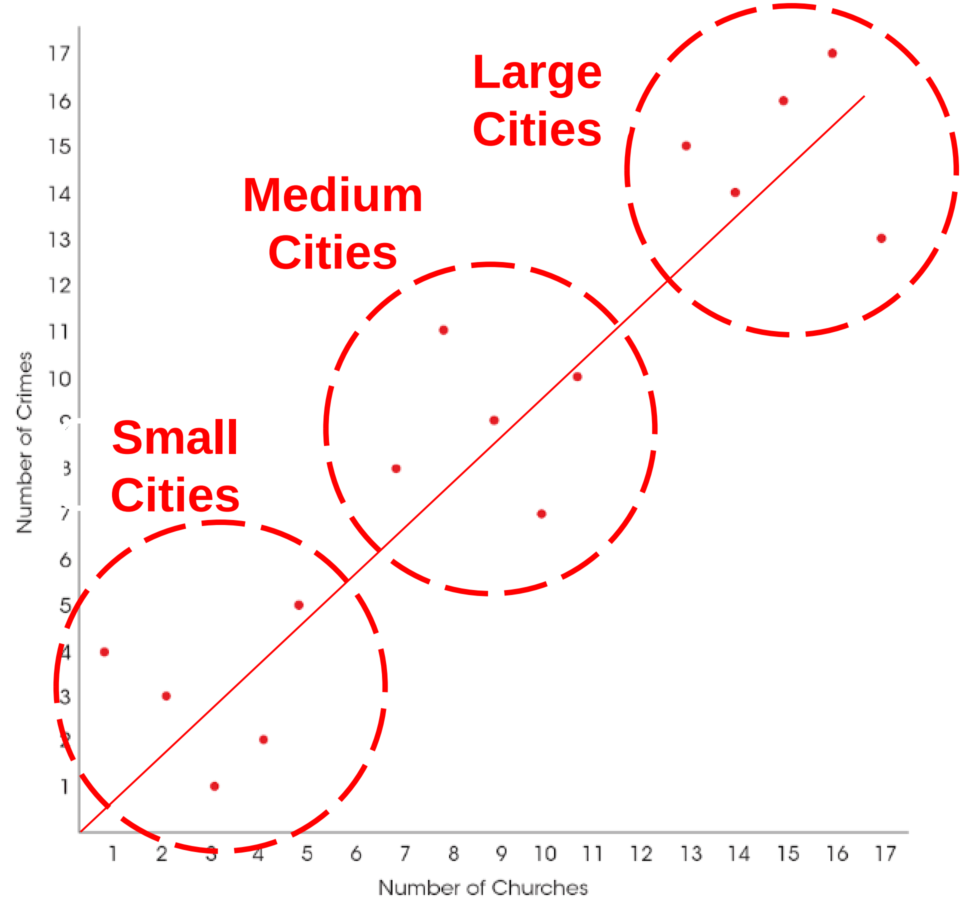


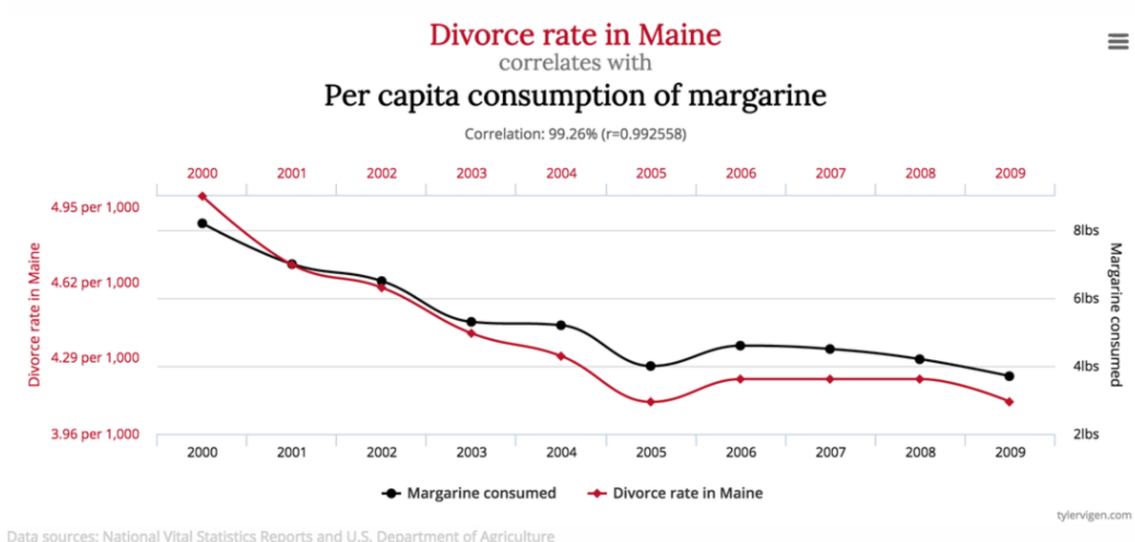
FIGURE 15.10
Hypothetical data showing the logical relationship between the number of churches and the number of crimes for three groups of cities: those with small populations ($Z = 1$), those with medium populations ($Z = 2$), and those with large populations ($Z = 3$).

Bivariate Statistics

Comical Examples of Correlation and Causation

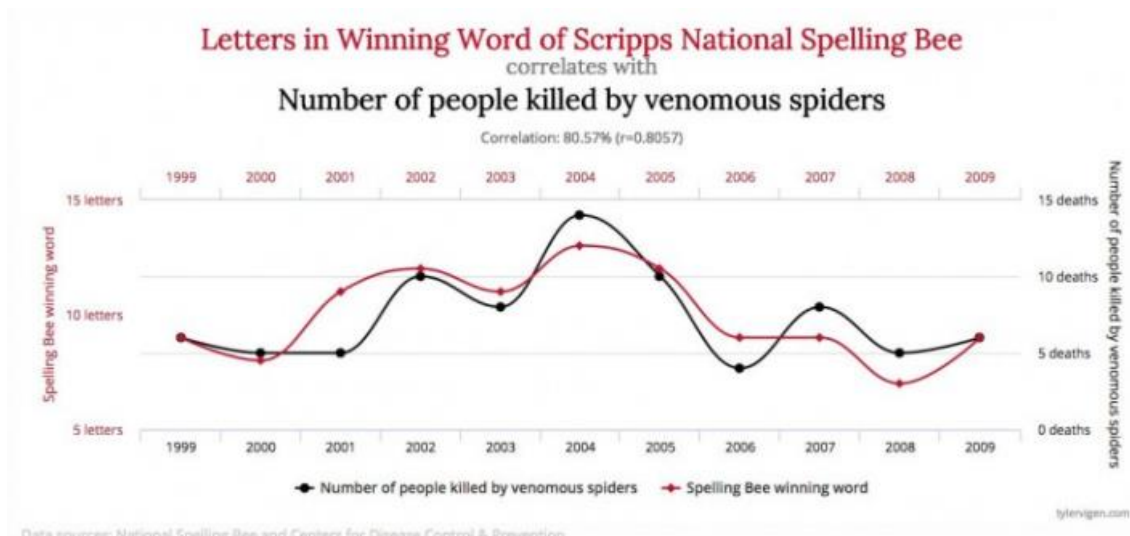
Margarine causes **divorce**?

or **divorce** causes **margarine**?



Spiders killing people causes **longer words in spelling bees**?

or **longer words in spelling bees** causes **venomous spiders to kill people**?

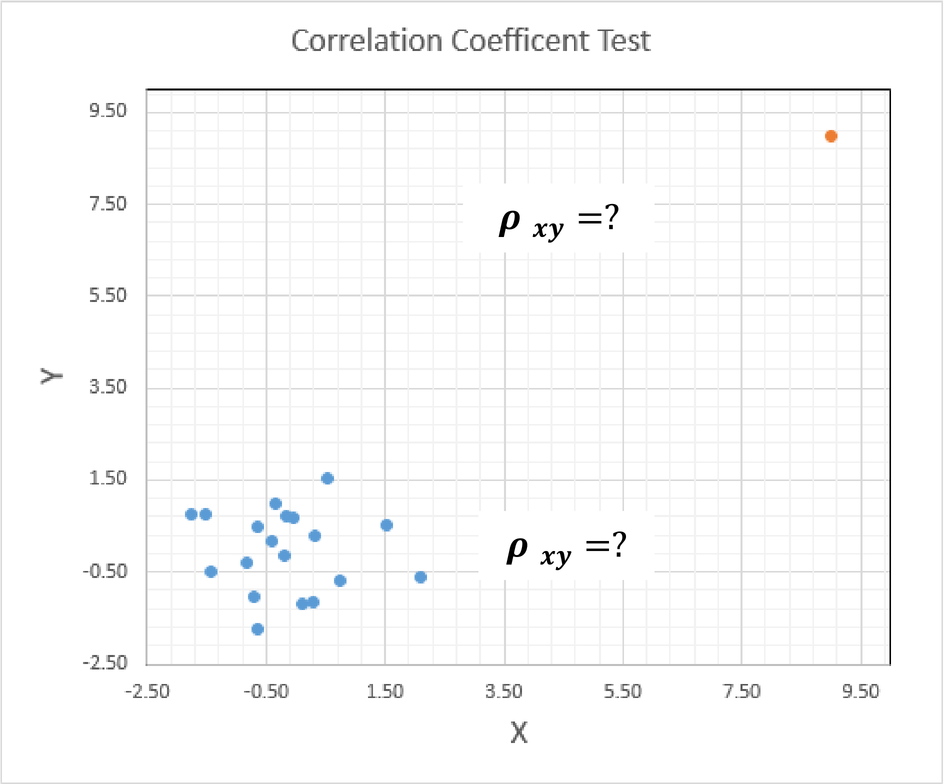


Bivariate Statistics
Exercise with Pearson's Correlation Coefficient

- Task 1: Generate a random data set of x and y variables and estimate their correlation coefficient (Hint: Rand() in Excel with N[0,1]).

- Task 2: Now add any desired outlier to the data and estimate the correlation coefficient (see example).

- How does this outlier affect the correlation coefficient?

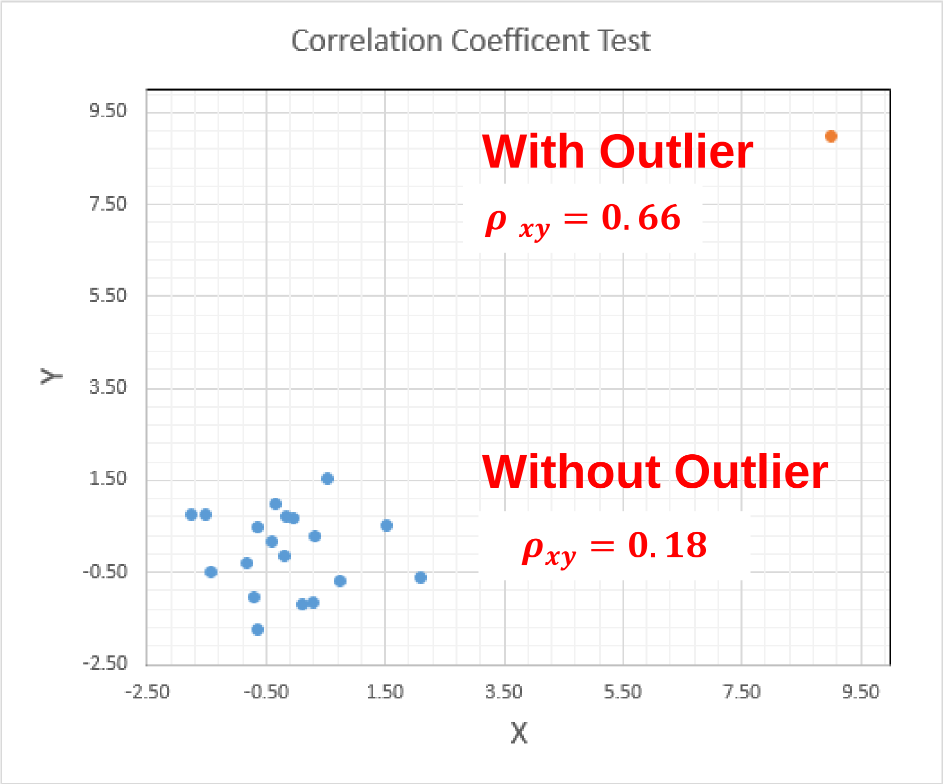


Excel Function NORM.INV(RAND(),0,1)

Bivariate Statistics

Exercise with Pearson's Correlation Coefficient

- Task 1: Generate a random data set of x and y variables and estimate their correlation coefficient (Hint: Rand() in Excel with N[0,1]).
- Task 2: Now add any desired outlier to the data and estimate the correlation coefficient (see example).
- How does this outlier affect the correlation coefficient?

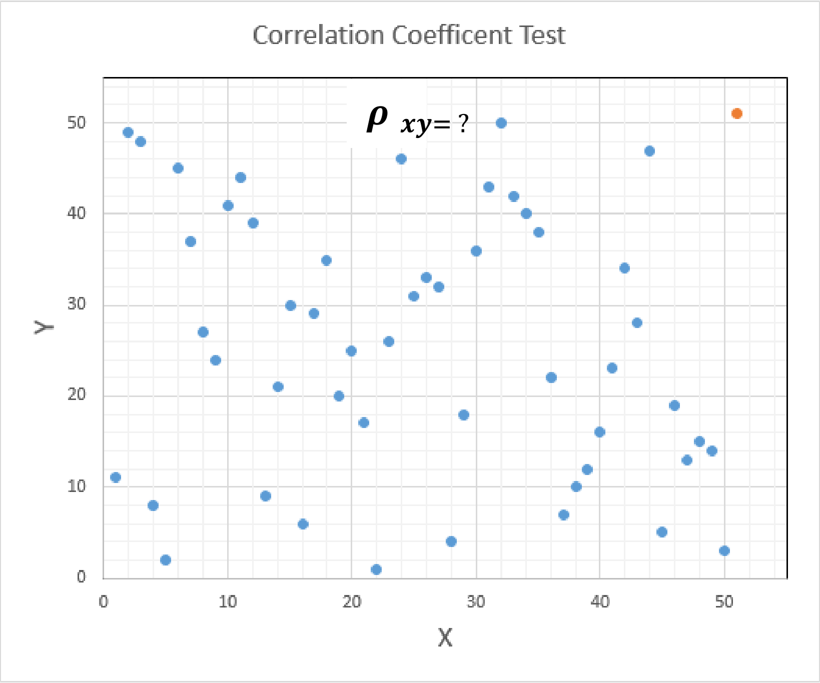


Bivariate Statistics
Exercise with Pearson's Correlation Coefficient

- Task 3: Apply the rank transform to the dataset (Hint: 21-Rank.Avg() in Excel).

- How does this outlier now affect the correlation coefficient?

- This is a more robust form of the correlation coefficient called the rank correlation coefficient.



Bivariate Statistics

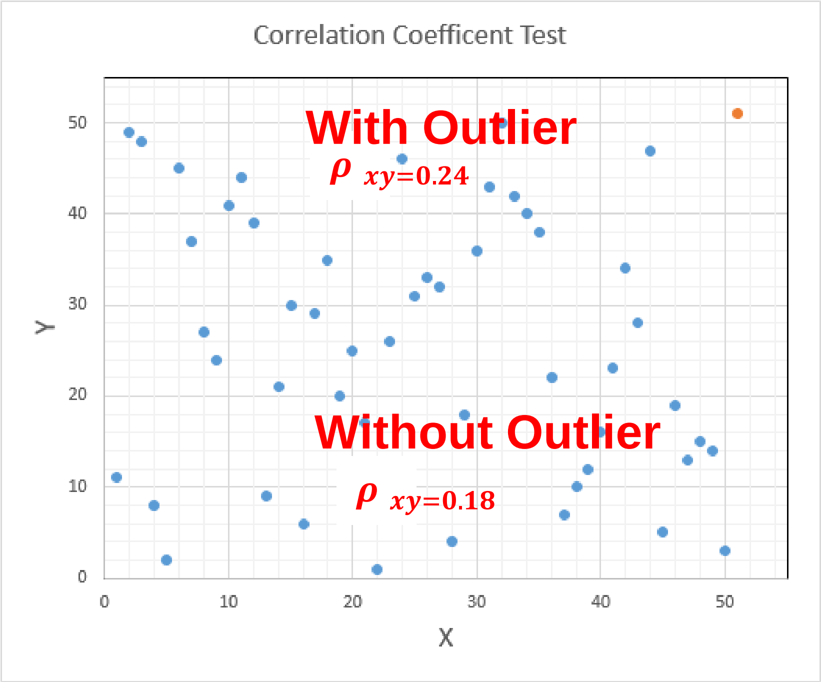
Exercise with Pearson's Correlation Coefficient

- Task 3: Applied the rank transform to the dataset

  (Hint: **52-Rank.Avg()** in Excel).

- How does this outlier now affect the correlation coefficient?

- This is a more robust form of the correlation coefficient called the rank correlation coefficient.



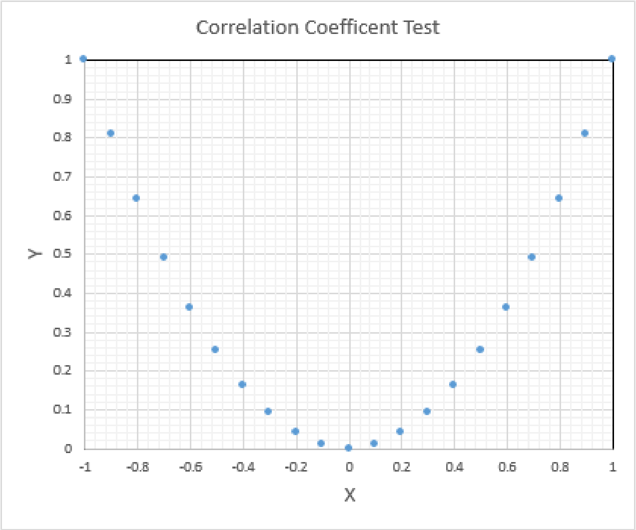
Excel Function =21-RANK.AVG(value,array)

Bivariate Statistics

Measuring Linear Relationships with the Correlation Coefficient

Correlation / Covariance is a measure of linear relationship

- What is the Correlation / Covariance of $y = x^2$ over range of [-1, 1]?

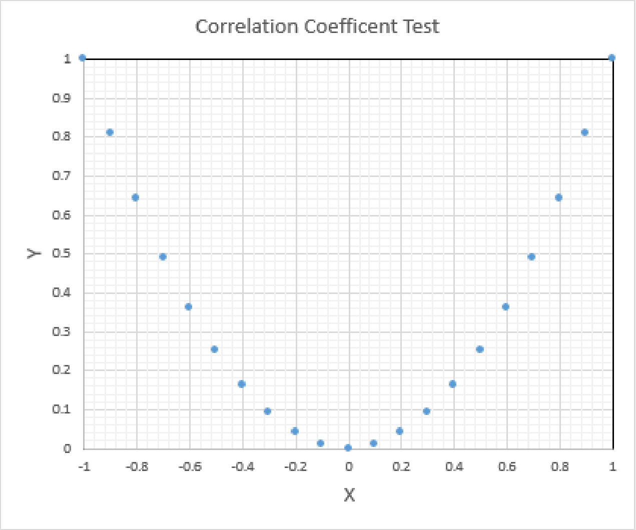


Excel Function Correl(array1,array2)

Correlation / Covariance is a measure of linear relationship

- What is the Correlation / Covariance of y = x^2 over range of [-1, 1]?

  Correlation Coefficient, $\boldsymbol{\rho_{xy}}$ = 0.0!



- Over range [0,1]?

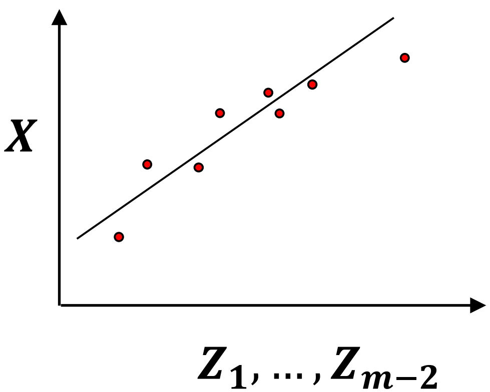
  Correlation Coefficient, $\boldsymbol{\rho_{xy}}$ = 0.96,
  Rank Correlation Coefficient, $\boldsymbol{\rho_{RxRy}}$ = 1.0

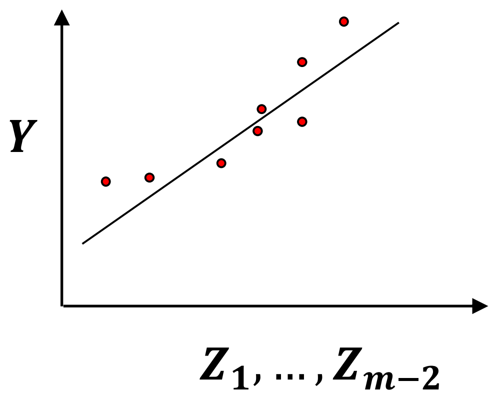
Excel Function Correl(array1,array2)

Bivariate Statistics
Partial Correlation

A method to calculate the correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ after controlling for the influence of $\boldsymbol{Z_1}, \dots, \boldsymbol{Z_{m-2}}$ other features on both $\boldsymbol{X}$ and Y.

1. perform linear, least-squares regression to predict $\boldsymbol{X}$ from $\boldsymbol{Z_1}, \dots, \boldsymbol{Z_{m-2}}$. $\boldsymbol{X}$ is regressed on the predictors to calculate the estimate, $\boldsymbol{X}^*$
2. perform linear, least-squares regression to predict $\boldsymbol{Y}$ from $\boldsymbol{Z_1}, \dots, \boldsymbol{Z_{m-2}}$. $\boldsymbol{Y}$ is regressed on the predictors to calculate the estimate, $\boldsymbol{Y}^*$

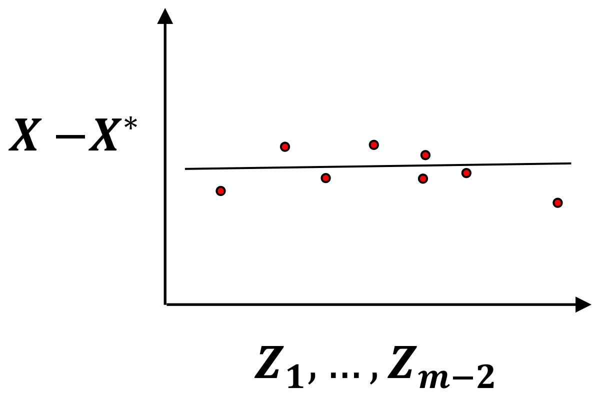


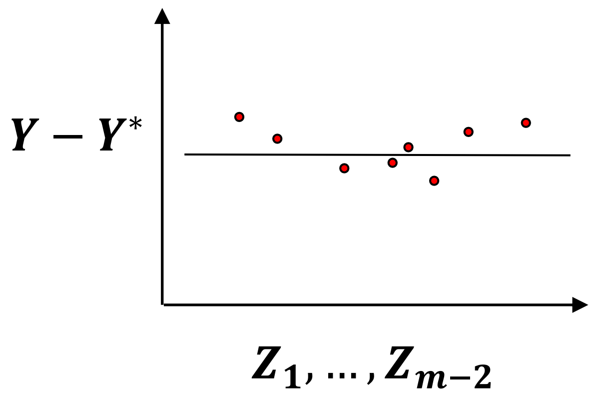


Bivariate Statistics
Partial Correlation

A method to calculate the correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ after controlling for the influence of $\boldsymbol{Z_1, \dots, Z_{m-2}}$ other features on both $\boldsymbol{X}$ and Y.

3. calculate the residuals in Step #1, $\boldsymbol{X - X^*}$, where $\boldsymbol{X^* = f(Z_1, \dots, Z_{m-2})}$, linear regression model
4. calculate the residuals in Step #1, $\boldsymbol{Y - Y^*}$, where $\boldsymbol{Y^* = f(Z_1, \dots, Z_{m-2})}$, linear regression model

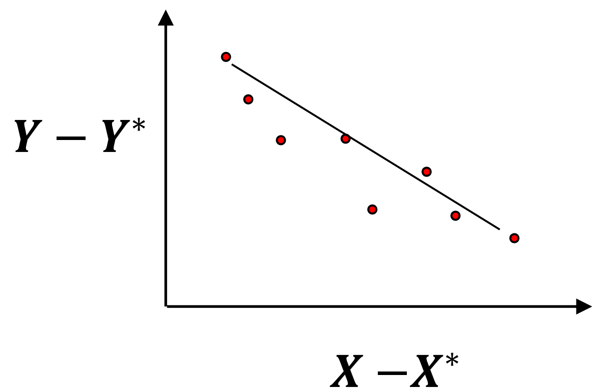




Bivariate Statistics
Partial Correlation

A method to calculate the correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ after controlling for the influence of $\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_{m-2}}$ other features on both $\boldsymbol{X}$ and Y.

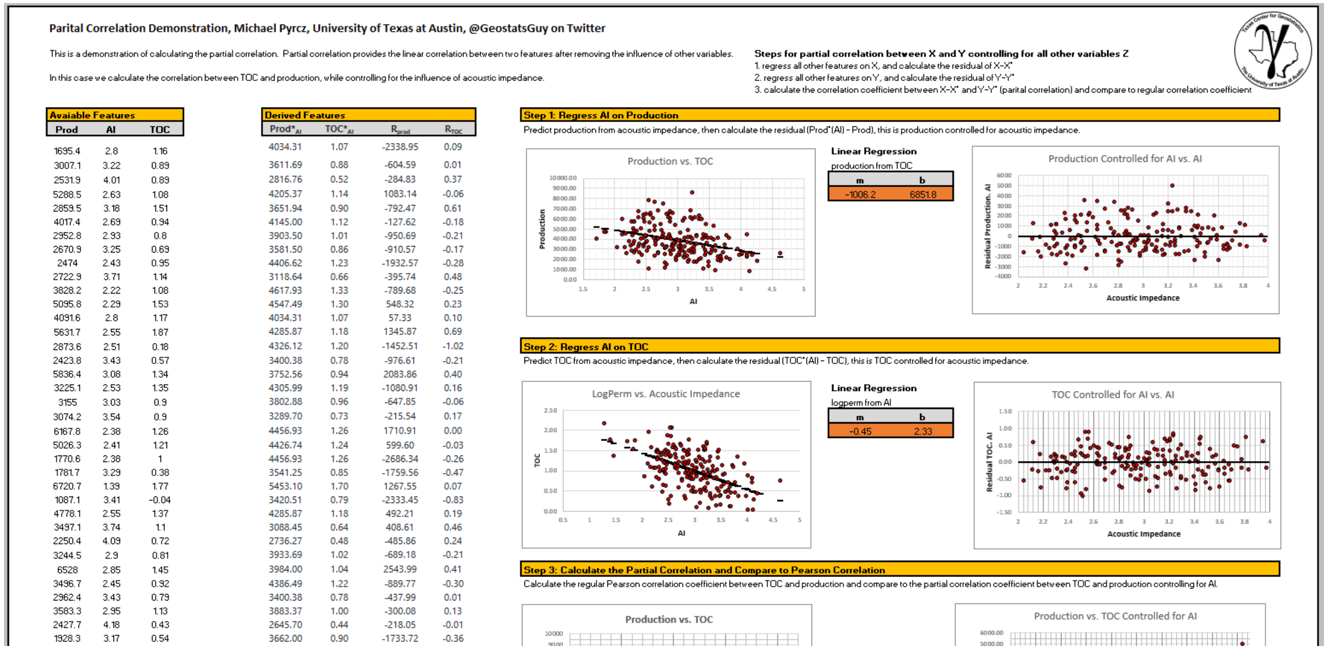
5. calculate the correlation coefficient between the residuals from Steps #3 and #4, $\rho_{\boldsymbol{X-X^*,Y-Y^*}}$



The partial correlation, provides a measure of the linear relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ while controlling for the effect of $\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_{m-2}}$ other features on both, $\boldsymbol{X}$ and $\boldsymbol{Y}$.

Partial Correlation Hands-on in Excel

Experiment with Partial Correlation:



Parital Correlation Demonstration, Michael Pyrcz, University of Texas at Austin, @GeostatsGuy on Twitter

This is a demonstration of calculating the partial correlation. Partial correlation provides the linear correlation between two features after removing the influence of other variables.

In this case we calculate the correlation between TOC and production, while controlling for the influence of acoustic impedance.

Steps for partial correlation between X and Y controlling for all other variables Z
1. regress all other features on X, and calculate the residual of X-X*
2. regress all other features on Y, and calculate the residual of Y-Y*
3. calculate the correlation coefficient between X-X* and Y-Y* (parital correlation) and compare to regular correlation coefficient

Avaiable Features			Derived Features			
Prod	AI	TOC	$Prod^*_{AI}$	TOC^*_{AI}	R_{prod}	R_{TOC}
1695.4	2.8	1.16	4034.31	1.07	-2338.95	0.09
3007.1	3.22	0.89	3611.69	0.88	-604.59	0.01
2531.9	4.01	0.89	2816.76	0.52	-284.83	0.37
5288.5	2.63	1.08	4205.37	1.14	1083.14	-0.06
2859.5	3.18	1.51	3651.94	0.90	-792.47	0.61
4017.4	2.69	0.94	4145.00	1.12	-127.62	-0.18
2952.8	2.93	0.8	3903.50	1.01	-950.69	-0.21
2670.9	3.25	0.69	3581.50	0.86	-910.57	-0.17
2474	2.43	0.95	4406.62	1.23	-1932.57	-0.28
2722.9	3.71	1.14	3118.64	0.66	-395.74	0.48
3828.2	2.22	1.08	4617.93	1.33	-789.68	-0.25
5095.8	2.29	1.53	4547.49	1.30	548.32	0.23
4091.6	2.8	1.17	4034.31	1.07	57.33	0.10
5631.7	2.55	1.87	4285.87	1.18	1345.87	0.69
2873.6	2.51	0.18	4326.12	1.20	-1452.51	-1.02
2423.8	3.43	0.57	3400.38	0.78	-976.61	-0.21
5836.4	3.08	1.34	3752.56	0.94	2083.86	0.40
3225.1	2.53	1.35	4305.99	1.19	-1080.91	0.16
3155	3.03	0.9	3802.88	0.96	-647.85	-0.06
3074.2	3.54	0.9	3289.70	0.73	-215.54	0.17
6167.8	2.38	1.26	4456.93	1.26	1710.91	0.00
5026.3	2.41	1.21	4426.74	1.24	599.60	-0.03
1770.6	2.38	1	4456.93	1.26	-2686.34	-0.26
1781.7	3.29	0.38	3541.25	0.85	-1759.56	-0.47
6720.7	1.39	1.77	5453.10	1.70	1267.55	0.07
1087.1	3.41	-0.04	3420.51	0.79	-2333.45	-0.83
4778.1	2.55	1.37	4285.87	1.18	492.21	0.19
3497.1	3.74	1.1	3088.45	0.64	408.61	0.46
2250.4	4.09	0.72	2736.27	0.48	-485.86	0.24
3244.5	2.9	0.81	3933.69	1.02	-689.18	-0.21
6528	2.85	1.45	3984.00	1.04	2543.99	0.41
3496.7	2.45	0.92	4386.49	1.22	-889.77	-0.30
2962.4	3.43	0.79	3400.38	0.78	-437.99	0.01
3583.3	2.95	1.13	3883.37	1.00	-300.08	0.13
2427.7	4.18	0.43	2645.70	0.44	-218.05	-0.01
1928.3	3.17	0.54	3662.00	0.90	-1733.72	-0.36

Step 1: Regress AI on Production

Predict production from acoustic impedance, then calculate the residual (Prod*(AI) – Prod), this is production controlled for acoustic impedance.

Linear Regression

production from TOC

m	b
-1006.2	6851.8

Step 2: Regress AI on TOC

Predict TOC from acoustic impedance, then calculate the residual (TOC*(AI) – TOC), this is TOC controlled for acoustic impedance.

Linear Regression

logperm from AI

m	b
-0.45	2.33

Step 3: Calculate the Partial Correlation and Compare to Pearson Correlation

Calculate the regular Pearson correlation coefficient between TOC and production and compare to the partial correlation coefficient between TOC and production controlling for AI.

Things to try:

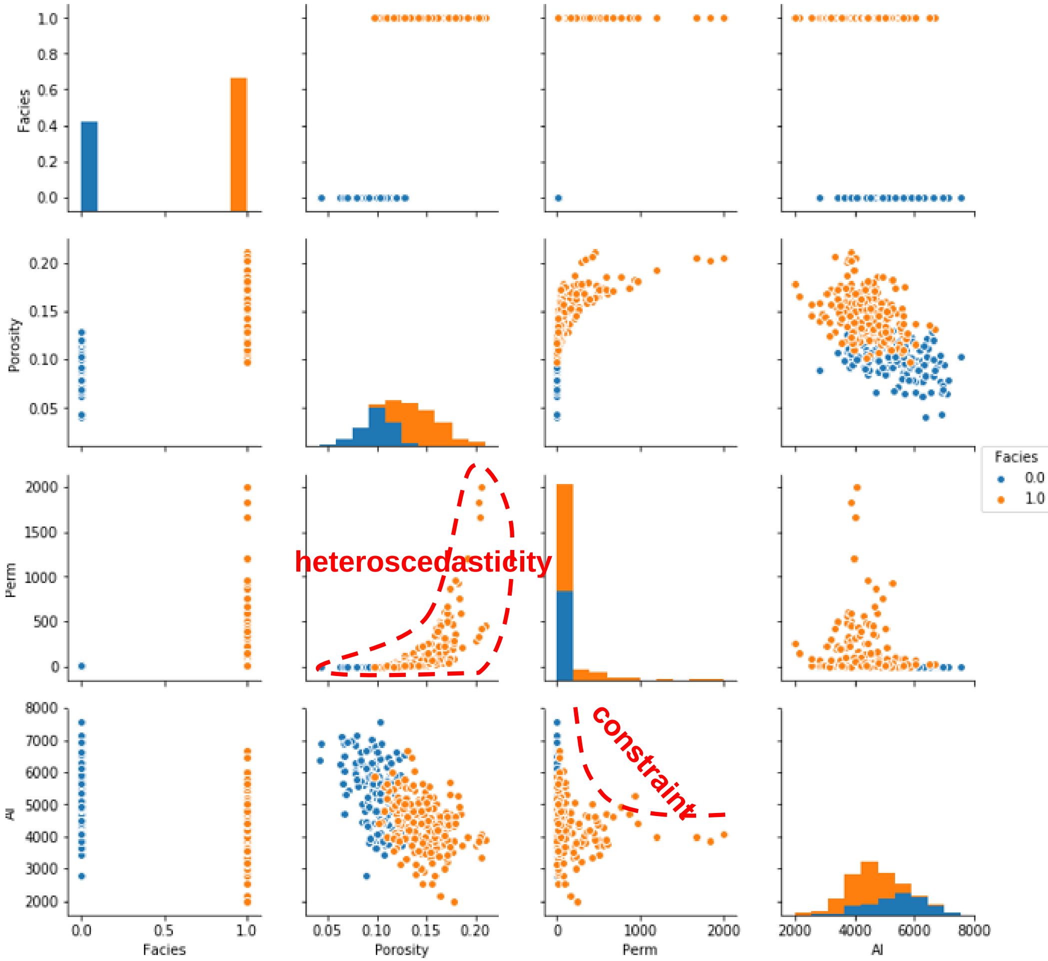
1. Increase the frequency over a region in the joint frequency distribution.
2. Add a TOC, Production outlier. TOC = 10, Production = 9999. What happened?

Does Vsh inform porosity?

The file is Partial_Correlation_Demo.xlsx at location https://git.io/fhA95.

Bivariate Statistics
Matrix Scatter Plots

- **For more than two variables make matrix scatterplots**
  – By hand in Excel or packages in R and Python.
  – Look for linear / nonlinear features
  – Look for homoscedasticity (constant conditional variance) and heteroscedasticity (conditional variance changes with value)
  – Look for constraints



Multivariate Modeling: Multivariate

Lecture outline . . .

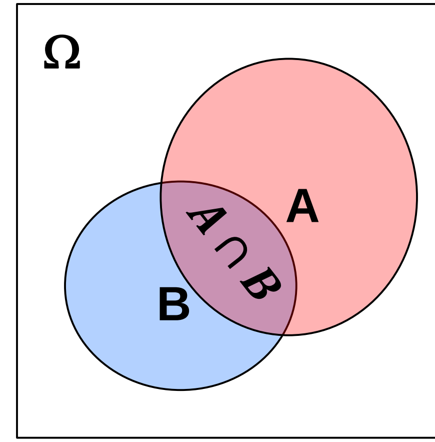
- Joints and Conditionals

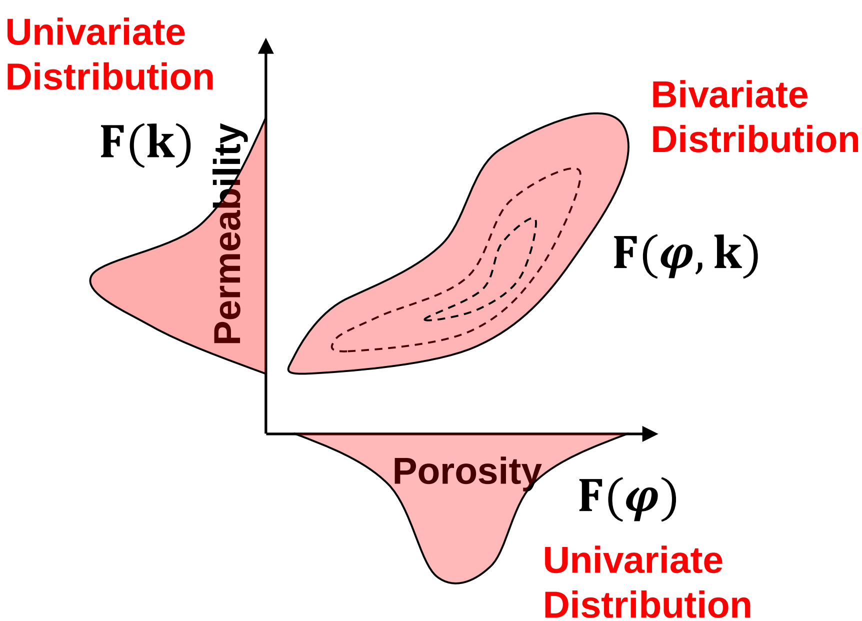
Instructor: Michael Pyrcz, the University of Texas at Austin

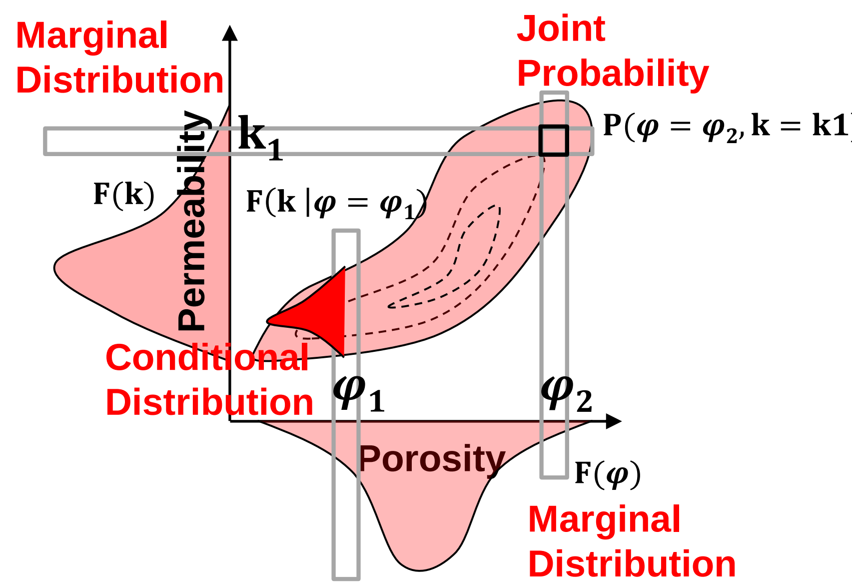
Probability Definitions
Conditional, Marginal and Joint Probability

Probability of B given A occurred? $P(B \mid A)$

$$\underbrace{P(B \mid A)}_{\text{Conditional Probability}} = \frac{P(A \cap B)}{P(A)} = \frac{\overbrace{P(A \text{ and } B)}^{\text{Joint Probability}}}{\underbrace{P(A)}_{\text{Marginal Probability}}}$$







Probability Definitions
Conditional, Marginal and Joint Probability

Marginal Probability: Probability of an event, irrespective of any other event

$$P(X), P(Y)$$

Conditional Probability: Probability of an event, given another event is already true.

$$P(X \; given \; Y), P(Y \; given \; X)$$

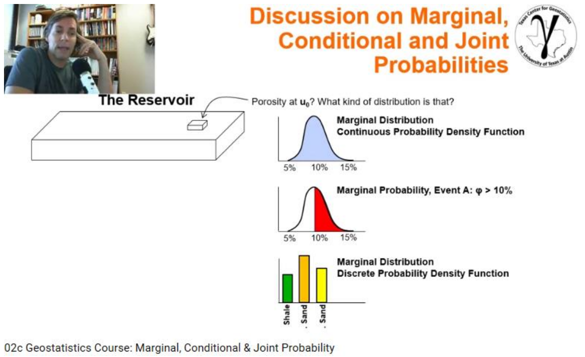
$$P(X \mid Y), P(Y|X)$$

Joint Probability: Probability of multiple events occurring together.

$$P(X \; and \; Y), P(Y \; and \; X)$$

$$P(X \; \cap \; Y), P(Y \cap X)$$

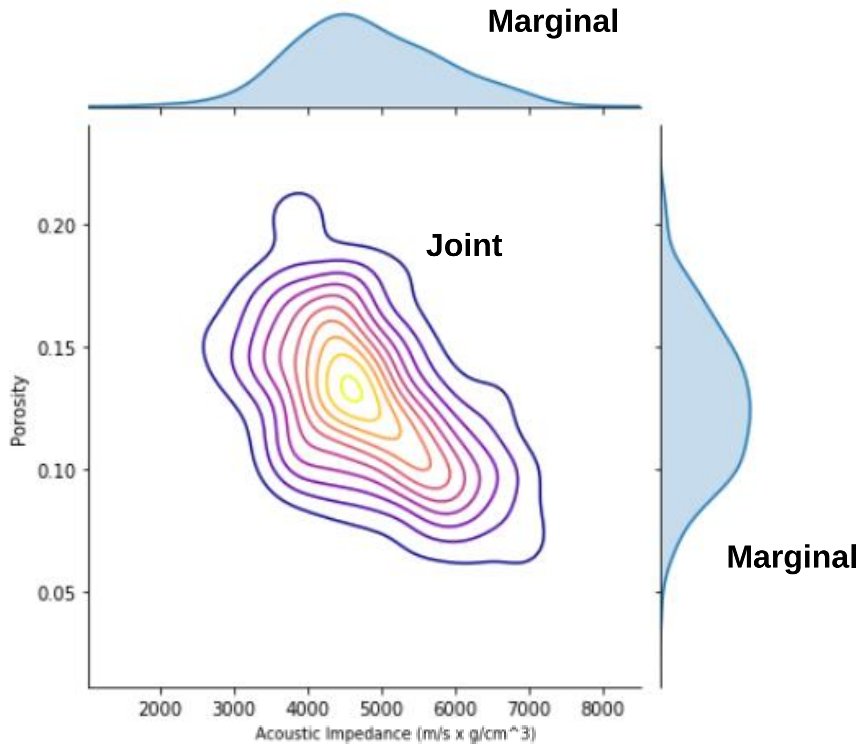
$$P(X, Y), P(Y, X)$$



See YouTube Video on Marginals, Conditionals and Joints! https://www.youtube.com/watch?v=bL2gPwMfYpc&index=5&t=0s&list=PLG19vXLQHvSB-D4XKYieEku9GQMQyAzjJ

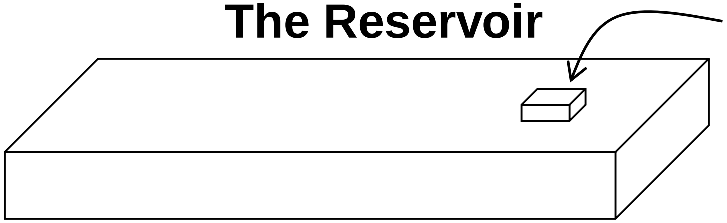
Marginal, Conditional and Joint Probability

- Working directly with marginal, conditional and joint probability
  - If you have enough data, you can directly calculate all the required probabilities
  - Go beyond statistics like correlation coefficient

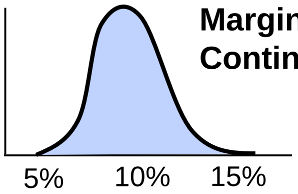


Marginal, Conditional and Joint Probability

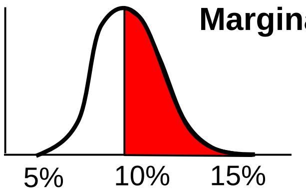
The Reservoir



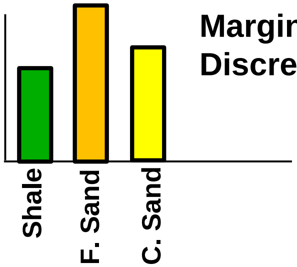
Porosity at $\mathbf{u_0}$? What kind of distribution is that?



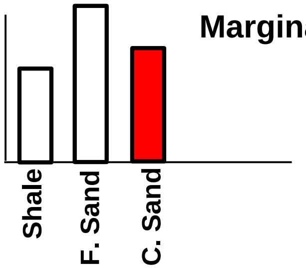
Marginal Distribution
Continuous Probability Density Function



Marginal Probability, Event A: φ > 10%



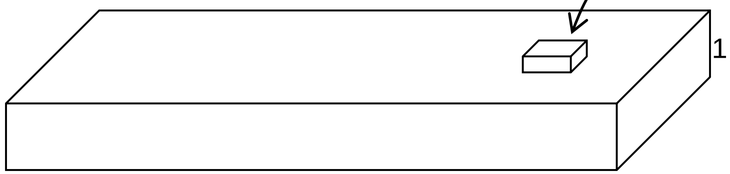
Marginal Distribution
Discrete Probability Density Function



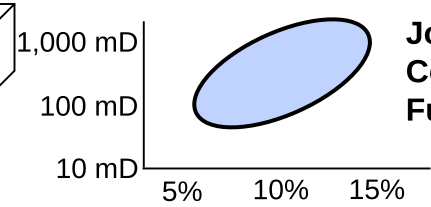
Marginal Probability, Event B: Facies = C. Sand

Marginal, Conditional and Joint Probability

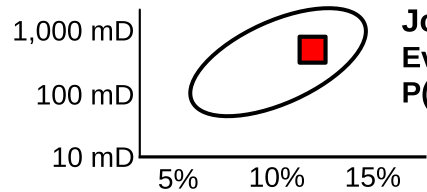
The Reservoir



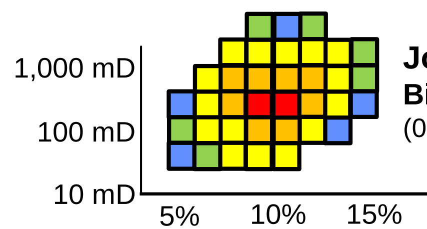
Porosity and Permeability at u_0? What kind of distribution is that?



Joint Distribution
Continuous Joint Probability Density Function

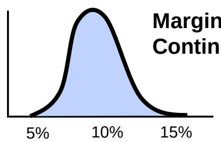


Joint Probability
Event A: 12% < φ < 14% and 600mD < k < 900mD
P(12% < φ < 14% ∩ 600mD < k < 900mD)

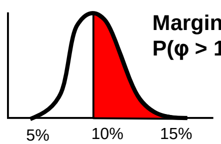


Joint Probability Density Function
Binned
(0% bins removed)

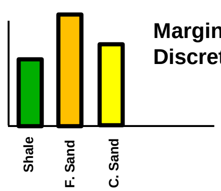
Univariate, Marginal Examples



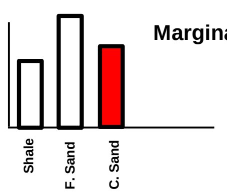
Marginal Distribution
Continuous Probability Density Function



Marginal Probability, Event A: φ > 10%
P(φ > 10%)



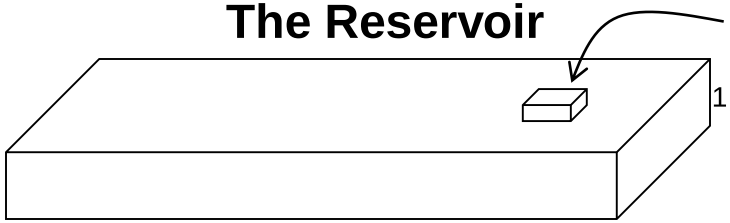
Marginal Distribution
Discrete Probability Density Function



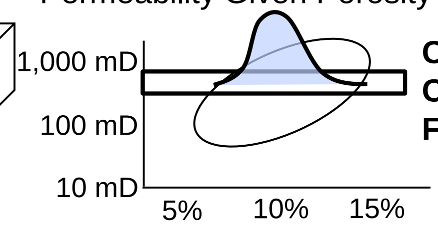
Marginal Probability, Event B: Facies = C. Sand

Marginal, Conditional and Joint Probability

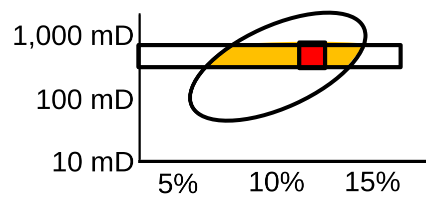
The Reservoir



Permeability Given Porosity = φ_1 at u_0? What kind of distribution is that?

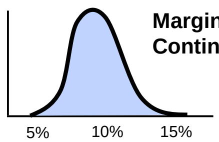


Conditional Distribution
Continuous Conditional Probability Density Function

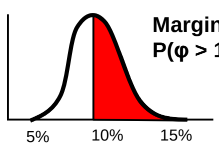


Conditional Probability
Event A: 12% < φ < 14% | 600mD < k < 900mD
P(12% < φ < 14% | 600mD < k < 900mD)

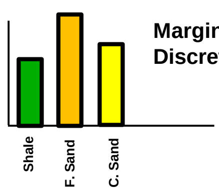
Univariate, Marginal Examples



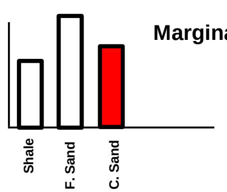
Marginal Distribution
Continuous Probability Density Function



Marginal Probability, Event A: φ > 10%
P(φ > 10%)

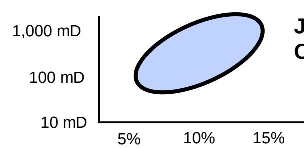


Marginal Distribution
Discrete Probability Density Function

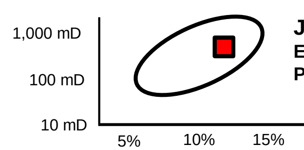


Marginal Probability, Event B: Facies = C. Sand

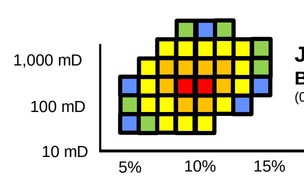
Bivariate, Joint Examples



Joint Distribution
Continuous Joint Probability Density Function



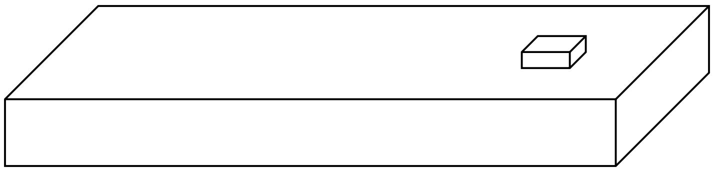
Joint Probability
Event A: 12% < φ < 14% and 600mD < k < 900mD
P(12% < φ < 14% ∩ 600mD < k < 900mD)



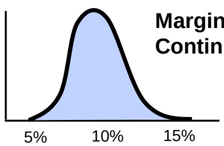
Joint Probability Density Function
Binned
(0% bins removed)

Marginal, Conditional and Joint Probability

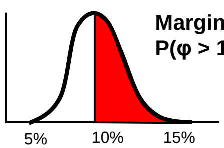
The Reservoir



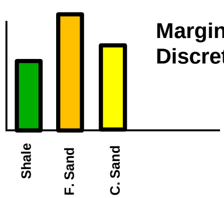
Univariate, Marginal Examples



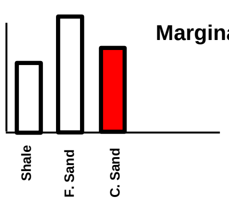
Marginal Distribution
Continuous Probability Distribution Function



Marginal Probability, Event A: φ > 10%
P(φ > 10%)

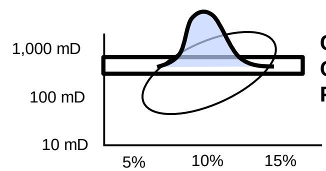


Marginal Distribution
Discrete Probability Distribution Function

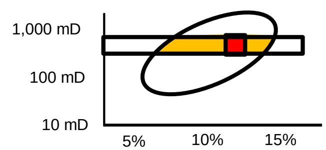


Marginal Probability, Event B: Facies = C. Sand

Bivariate, Conditional Examples

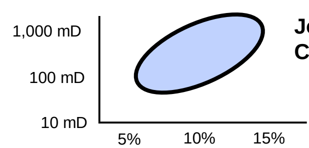


Conditional Distribution
Continuous Conditional Probability Density Function

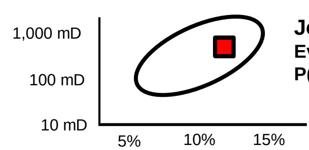


Conditional Probability
Event A: 12% < φ < 14% | 600mD < k < 900mD
P(12% < φ < 14% | 600mD < k < 900mD)

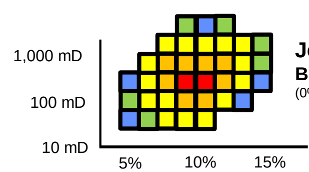
Bivariate, Joint Examples



Joint Distribution
Continuous Joint Probability Density Function



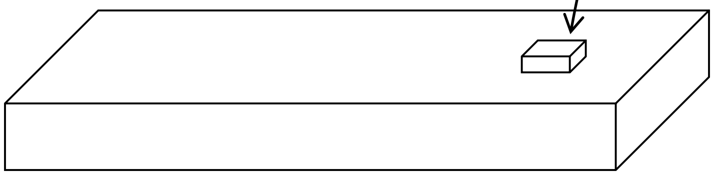
Joint Probability
Event A: 12% < φ < 14% and 600mD < k < 900mD
P(12% < φ < 14% ∩ 600mD < k < 900mD)



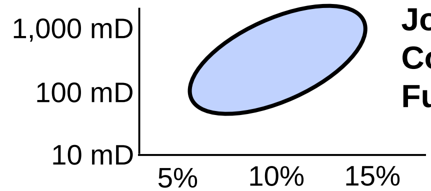
Joint Probability Density Function
Binned
(0% bins removed)

Marginal, Conditional and Joint Probability

The Reservoir



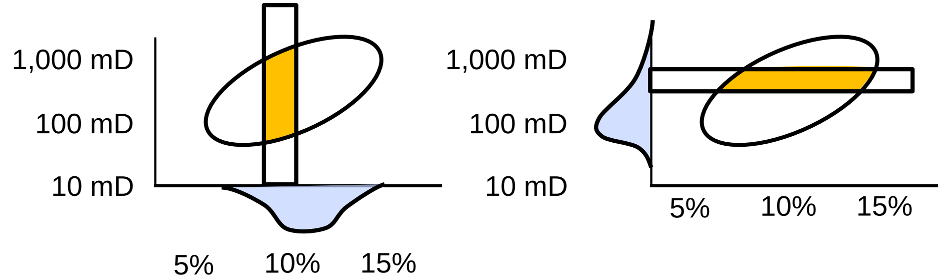
How to Calculate a Marginal Distribution from a Joint Distribution?



Joint Distribution
Continuous Joint Probability Density Function

Definition of a Marginal Distribution

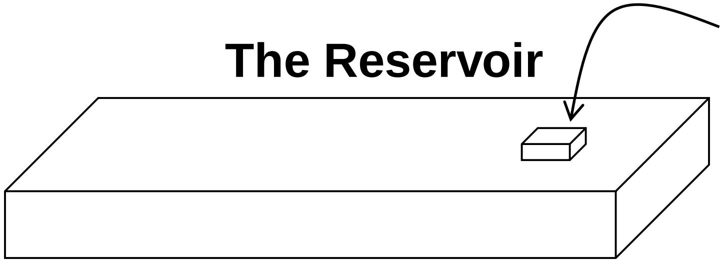
$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy \quad \textbf{or} \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx$$

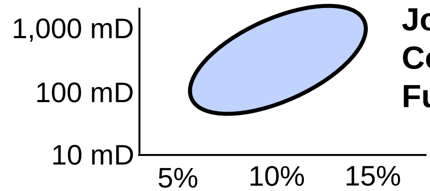


Marginal, Conditional and Joint Probability

Calculate a Conditional Distribution from a Joint Distribution?

The Reservoir

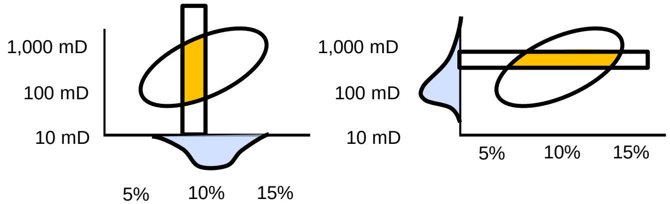




Joint Distribution
Continuous Joint Probability Density Function

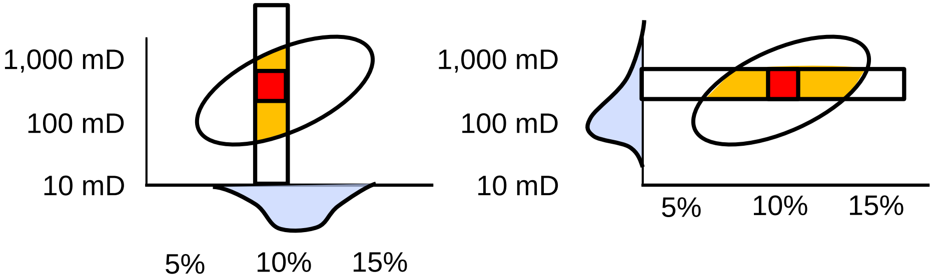
Definition of a Marginal Distribution

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy \quad \textbf{or} \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx$$



Definition of a Conditional Distribution

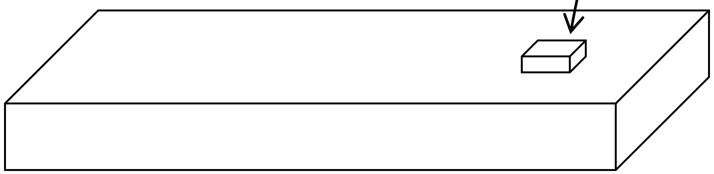
$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)} \quad \textbf{or} \quad f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$



Marginal, Conditional and Joint Probability

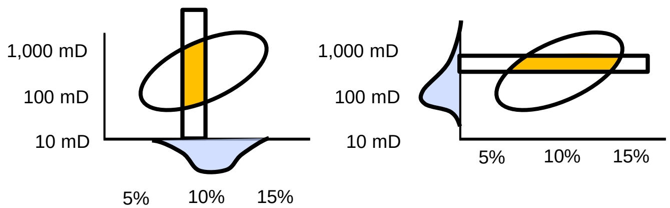
How to Calculate a Joint Distribution?

The Reservoir



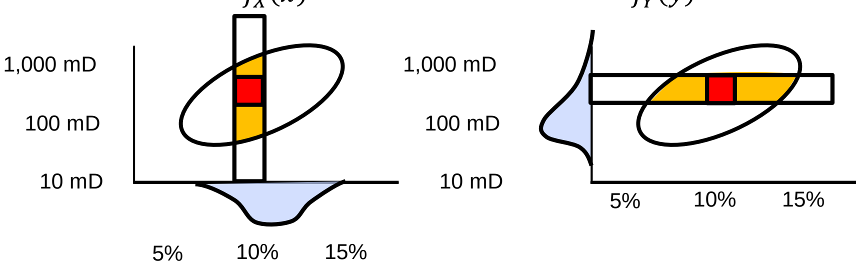
Definition of a Marginal Distribution

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dy \quad \text{or} \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dx$$



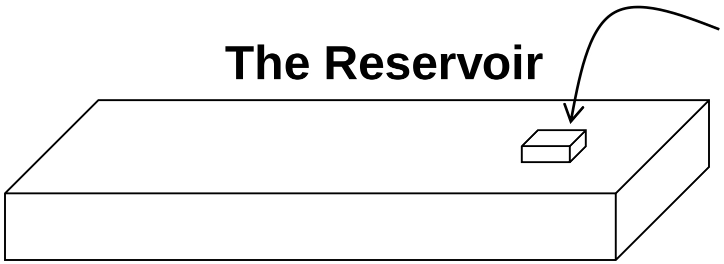
Definition of a Conditional Distribution

$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)} \quad \text{or} \quad f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$



Marginal, Conditional and Joint Probability

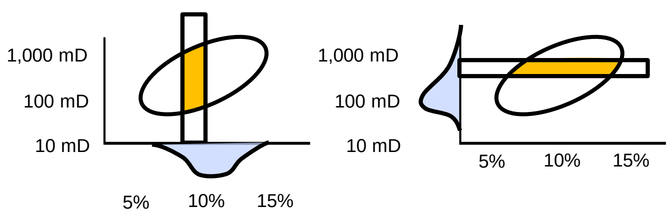
The Reservoir



How to Calculate a Joint Distribution?

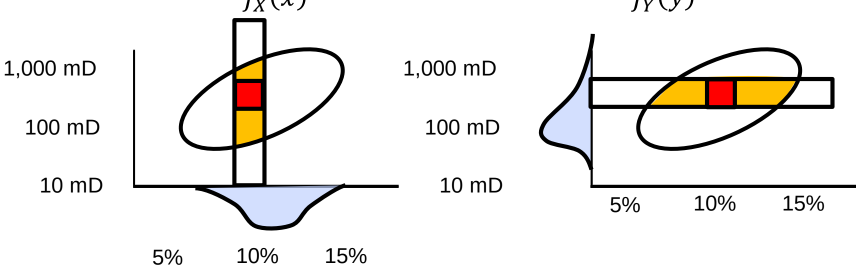
Definition of a Marginal Distribution

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dy \quad \textbf{or} \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)\,dx$$

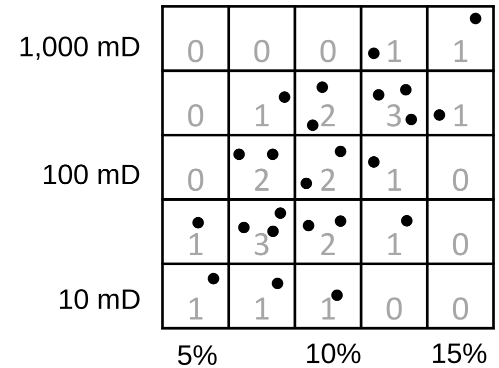


Definition of a Conditional Distribution

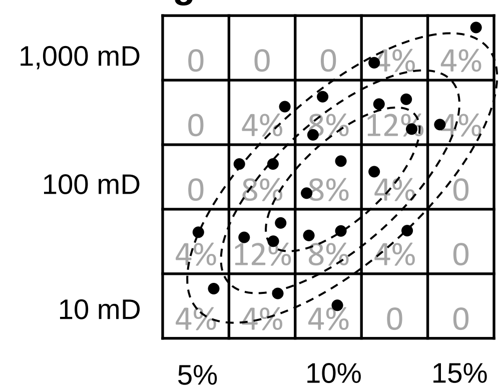
$$f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)} \quad \textbf{or} \quad f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)}$$



Non-parametric - Counting Samples in Bins

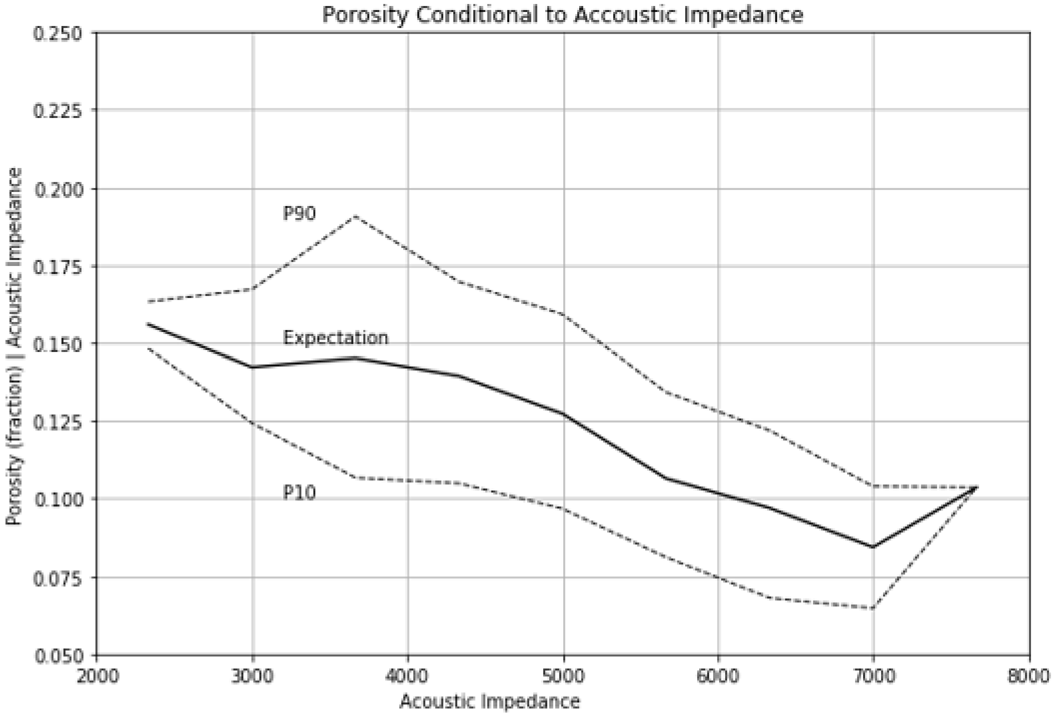


Fitting a Parametric Model



Marginal, Conditional and Joint Probability

- Consider working with conditional statistics.
  - Powerful, flexible assessment of multivariate relationships, without linear assumption



Conditional, Marginal and Joint Hands-on

Joint Distribution:

$$f_{XY}(x, y)$$

Marginal Distribution:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y)dy$$

Conditional Distribution:

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

Table of Frequencies

Porosity (%) \ Fraction Shale (%)	10%	30%	50%	70%	90%
25%	1	1	0	0	0
20%	2	3	2	0	0
15%	1	2	2	1	0
10%	0	0	2	3	2
5%	0	0	1	1	1

Conditional, Marginal and Joint Hands-on

Joint Distribution:

$$f_{XY}(x, y)$$

Marginal Distribution:

$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x, y)dy$$

Conditional Distribution:

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

Table of Joint Probabilities

Porosity (%) \ Fraction Shale (%)	10%	30%	50%	70%	90%
25%	4%	4%	0	0	0
20%	8%	12%	8%	0	0
15%	4%	8%	8%	4%	0
10%	0	0	8%	12%	8%
5%	0	0	4%	4%	4%

Conditional, Marginal and Joint Hands-on

Given these joint probabilities calculate the:

Marginal Distributions:

Vsh	**10%**	**30%**	**50%**	**70%**	**90%**
$f_{Vsh}(v_{sh}) =$					

Porosity	**5%**	**10%**	**15%**	**20%**	**25%**
$f_{\varphi}(\varphi) =$					

Table of Joint Probabilities

Porosity (%)					
25%	4%	4%	0	0	0
20%	8%	12%	8%	0	0
15%	4%	8%	8%	4%	0
10%	0	0	8%	12%	8%
5%	0	0	4%	4%	4%
	10%	**30%**	**50%**	**70%**	**90%**

Fraction Shale (%)

Conditional Distribution:

Vsh	**10%**	**30%**	**50%**	**70%**	**90%**
$f_{Vsh\|\varphi}(v_{sh}\| \varphi = 15\%) =$					

Conditional, Marginal and Joint Hands-on

Given these joint probabilities calculate the:

Marginal Distributions:

Vsh	10%	30%	50%	70%	90%
$f_{Vsh}(v_{sh})$	16%	24%	28%	20%	12%

Porosity	5%	10%	15%	20%	25%
$f_{\varphi}(\varphi) =$	12%	28%	24%	28%	8%

Conditional Distribution:

Vsh	10%	30%	50%	70%	90%
	1/6	1/3	1/3	1/6	0

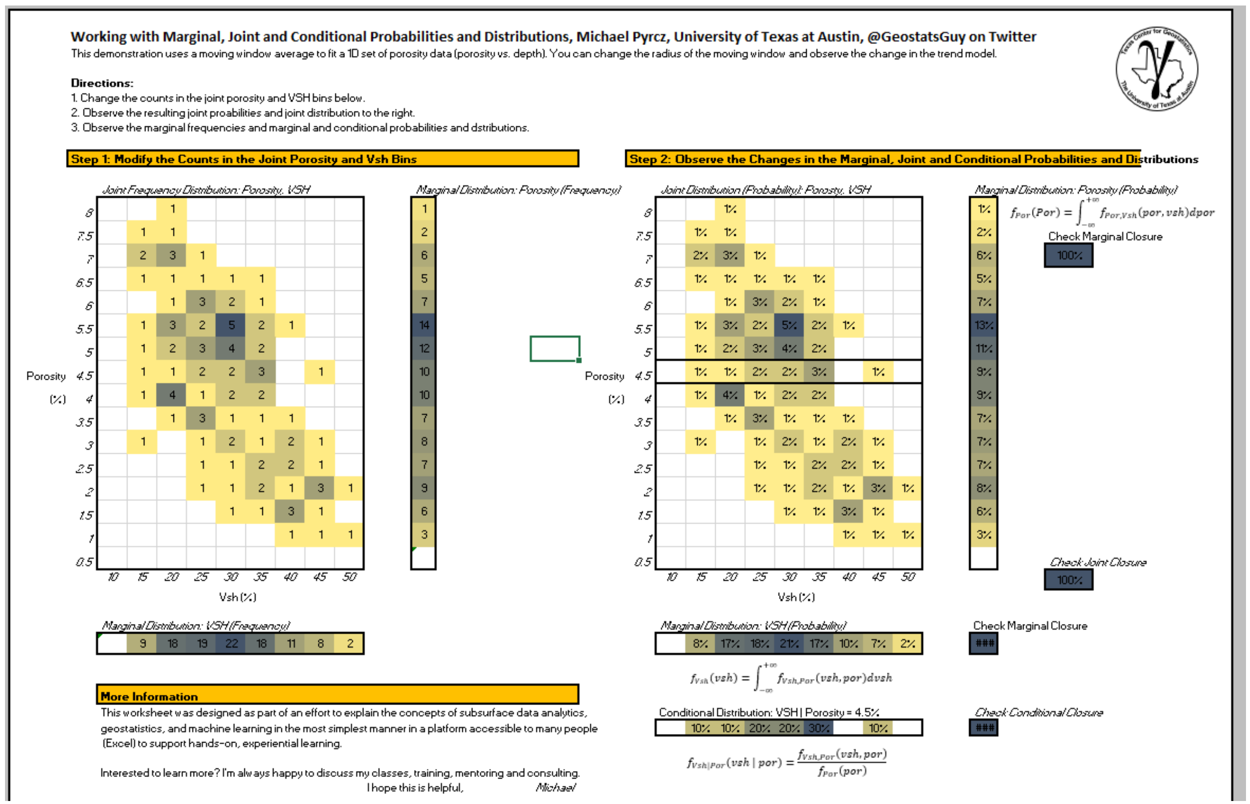
$$f_{Vsh|\varphi}(v_{sh}|\,\varphi = 15\%) = f_{Vsh,\varphi}(v_{sh}, \varphi = 15\%)/f_{\varphi}(\varphi = 15\%)$$

Table of Joint Probabilities

Porosity (%) \ Fraction Shale (%)	10%	30%	50%	70%	90%
25%	4%	4%	0	0	0
20%	8%	12%	8%	0	0
15%	4%	8%	8%	4%	0
10%	0	0	8%	12%	8%
5%	0	0	4%	4%	4%

Spatial Calculation in Hands-on in Excel

Experiment with Marginal, Joint and Conditionals:



Things to try:

1. Increase the frequency over a region in the joint frequency distribution.
2. Does Vsh inform porosity?

The file is Marginal_Joint_Conditional.xlsx at location https://git.io/fhA9X.

Multivariate Analysis Demo

Demo workflow for Multivariate Analysis
https://git.io/fh2DR

GeostatsPy: Multivariate Analysis for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

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

PGE 383 Exercise: Multivariate Analysis for Subsurface Data Analytics in Python

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

Bivariate Analysis

Understand and quantify the relationship between two variables

- example: relationship between porosity and permeability
- how can we use this relationship?

What would be the impact if we ignore this relationship and simply modeled porosity and permeability independently?

- no relationship beyond constraints at data locations
- independent away from data
- nonphysical results, unrealistic uncertainty models

Bivariate Statistics

Pearson's Product-Moment Correlation Coefficient

- Provides a measure of the degree of linear relationship.
- We refer to it as the 'correlation coefficient'

Let's review the sample variance of variable x. Of course, I'm truncating our notation as x is a set of samples a locations in our modeling space, $x(\mathbf{u}_\alpha), \forall \alpha = 0, 1, \dots, n-1$.

$$\sigma_x^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{(n-1)}$$

We can expand the the squared term and replace on of them with y, another variable in addition to x.

$$C_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{(n-1)}$$

We now have a measure that represents the manner in which variables x and y co-vary or vary together. We can standardized the covariance by the product of the standard deviations of x and y to calculate the correlation coefficent.

$$\rho_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{(n-1)\sigma_x \sigma_y}, \; -1.0 \le \rho_{xy} \le 1.0$$

In summary we can state that the correlation coefficient is related to the covariance as:

$$\rho_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y}$$

The Person's correlation coefficient is quite sensitive to outliers and depature from linear behavoir (in the bivariate sense). We have an altenrative known as the Spearman's rank correlations coefficient.

Multivariate Topics

- Other Topics that Could be Covered
  - Methods to remove correlation and model variables independently
  - Methods for dimensional reduction
  - Methods for clustering analysis

Multivariate New Tools

Topic	Application to Subsurface Modeling
Multivariate Analysis	In the presence of multivariate relationships, must jointly model variables. *Summarize with bivariate statistics, and visualize and use conditional statistics to go beyond linear measures.*
Limitations of Correlation	Correlation indicates degree of linear correlation and does not imply causation. *Visualize and use rank correlation coefficient when needed and apply careful experiments (controlled) to establish causation.*
Use Conditional Statistics	*Use conditional distributions to communicate the influence of variables on each other. Provides the value of knowing X to predict Y.* *Assess the influence of acoustic impedance on predicting porosity away from wells with conditional distributions.*

Multivariate Modeling: Multivariate

Lecture outline . . .

- Feature Selection

Instructor: Michael Pyrcz, the University of Texas at Austin

Feature Ranking Motivation

Variable Ranking

- There are often many predictor features, input variables, available for us to work with for subsurface prediction.
- There are good reasons to be selective, throwing in every possible feature is not a good idea!
- In general, for the best prediction model, careful selection of the fewest features that provide the most amount of information is the best practice.

Feature Ranking Motivation

More Motivation to Work with Fewer Variables:

- more variables result in more complicated workflows that require more professional time and have increased opportunity for blunders
- higher dimensional feature sets are more difficult to visualize
- more complicated models may be more difficult to interrogate, interpret and QC
- inclusion of highly redundant and colinear variables increases model instability and decreases prediction accuracy in testing
- more variables generally increase the computational time required to train the model and the model may be less compact and portable
- the risk of overfit increases with the more variables, more complexity

What is Feature Ranking?

More Motivation to Work with Fewer Variables:

- Feature ranking is a set of metrics that assign relative importance or value to each feature with respect to information contained for inference and importance in predicting a response feature.
- There are a wide variety of possible methods to accomplish this.
- My recommendation is a **wide-array** approach with multiple metric, while understanding the assumptions and limitations of each metric.

Here's the general types of metrics that we will consider for feature ranking:

1. Visual Inspection of Data Distributions and Scatter Plots
2. Statistical Summaries
3. Model-based
4. Recursive Feature Elimination

What is Feature Ranking?

Expert Knowledge:

- Also, we should not neglect expert knowledge.
- If additional information is known about physical processes, causation, reliability and availability of features this should be integrated into assigning feature ranks.
- We should be learning as we perform our analysis, testing new hypotheses.

Feature Ranking Metrics

Metric #1: Visual Inspection

- In any multivariate work we should start with the univariate analysis, summary statistics of one variable at a time. The summary statistic ranking method is qualitative, we are asking:
  - are there data issues?
  - do we trust the features? do we trust the features all equally?
  - are there issues that need to be taken care of before we develop any multivariate workflows?

Feature Ranking Metrics

Summary statistics are a critical first step in data checking.

	count	mean	std	min	25%	50%	75%	max
Well	200.0	100.500000	57.879185	1.000000	50.750000	100.500000	150.250000	200.000000
Por	200.0	14.991150	2.971176	6.550000	12.912500	15.070000	17.402500	23.550000
Perm	200.0	4.330750	1.731014	1.130000	3.122500	4.035000	5.287500	9.870000
AI	200.0	2.968850	0.566885	1.280000	2.547500	2.955000	3.345000	4.630000
Brittle	200.0	48.161950	14.129455	10.940000	37.755000	49.510000	58.262500	84.330000
TOC	200.0	0.991950	0.478264	0.000000	0.617500	1.030000	1.350000	2.180000
VR	200.0	1.964300	0.300827	0.930000	1.770000	1.960000	2.142500	2.870000
Prod	200.0	3864.407081	1553.277558	839.822063	2686.227611	3604.303507	4752.637556	8590.384044
const	200.0	1.000000	0.000000	1.000000	1.000000	1.000000	1.000000	1.000000

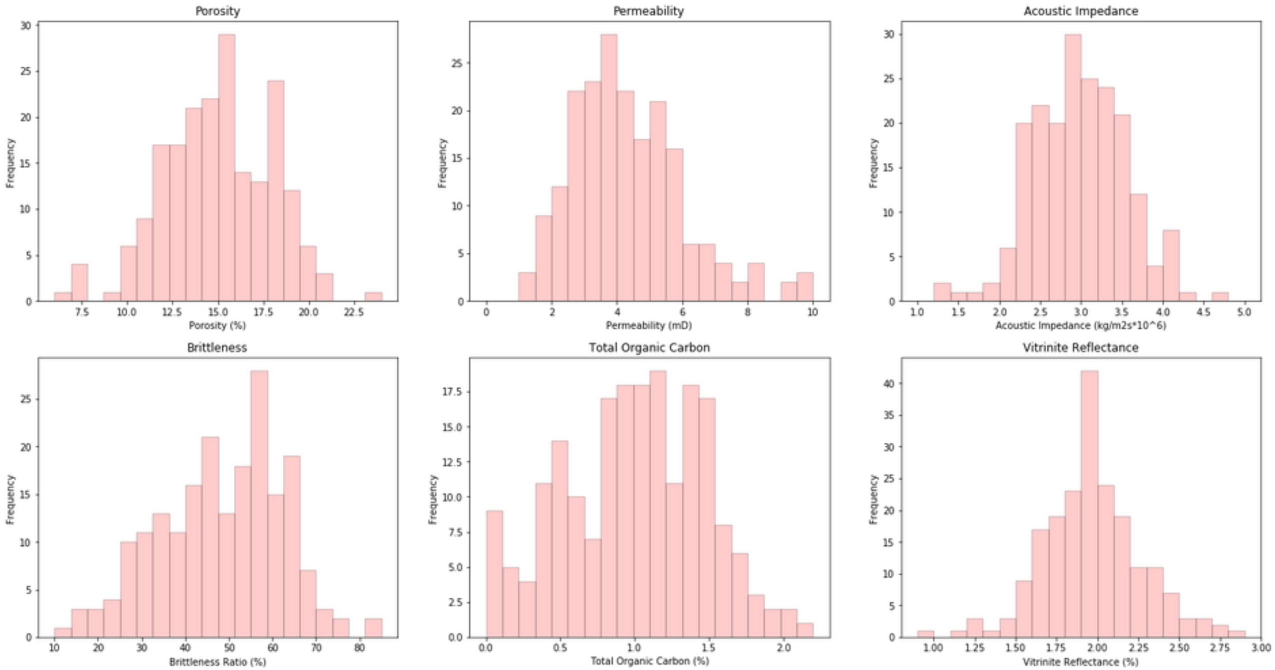
- the number of valid (non-null) values for each feature
- general behaviors such as central tendency, mean, and dispersion, variance.
- issues with negative values, extreme values, and values that are outside the range of plausible values for each property.

Feature Ranking Metrics

Metric #2: Univariate Distributions

- As with summary statistics, this ranking method is a qualitative check for issues with the data and to assess our confidence with each feature.
- It is better to not include a feature with low confidence of quality as it may be misleading (while adding to model complexity as discussed previously).
- Assess our ability to use methods that have distribution assumptions

Feature Ranking Metrics



The univariate distributions look good:

- there are no obvious outliers
- the permeability is positively skewed as often observed
- the corrected TOC has a small zero truncation spike, but it's reasonable
- some departure from Gaussian form, could transform

Feature Ranking Metrics

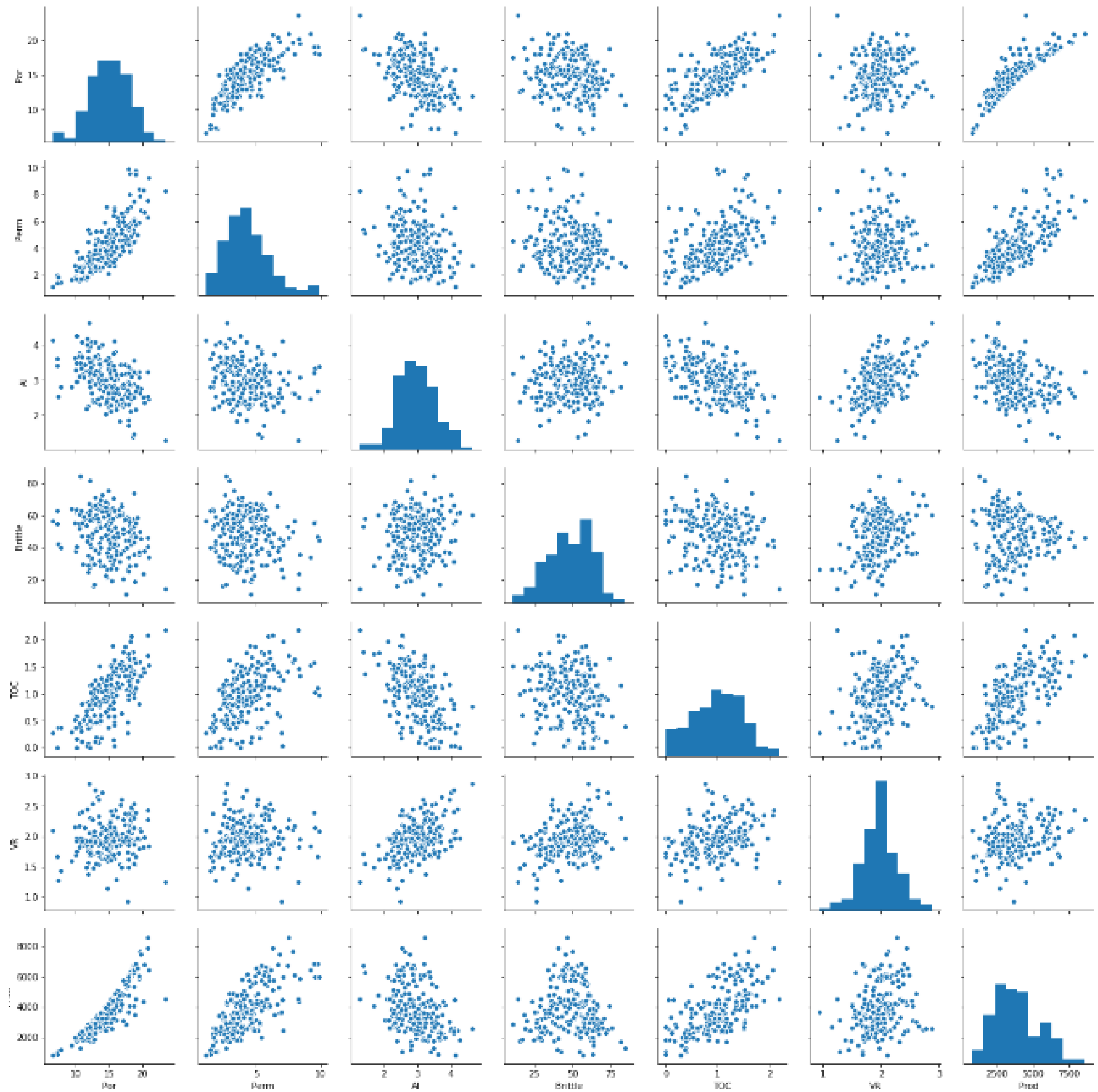
Metric #3: Bivariate

- matrix scatter plots are a very efficient method to observe the bivariate relationships between the variables.
- this is another opportunity through data visualization to identify data issues, outliers
- we can assess if we have collinearity, specifically the simpler form between two features at a time
- Bivariate Gaussian is assumed for methods such as correlation and partial correlation

Feature Ranking Metrics

How could we use this plot for variable ranking?

- variables that are closely related to each other.
- linear vs. non-linear relationships
- constraint relationships and heteroscedasticity between variables.



Feature Ranking Metrics

Metric #3: Bivariate

- bivariate visualization and analysis is not sufficient to understand all the multivariate relationships in the data
- multicollinearity includes strong linear relationships between 2 or more features.
- higher order nonlinear features, outliers and coverage?
- these may be hard to see with only bivariate plots.

Feature Ranking Metrics

Ranking Method #4 - Pairwise Covariance

- Pairwise covariance provides a measure of the strength of the linear relationship between each predictor feature and the response feature.
- We now specify our goal of this study is to predict production, our response variable, from the other available predictor features.
- We are thinking predictively now, not inferentially, we want to estimate the function, $\hat{f}$ to accomplish this

Covariance:

- measures the strength of the linear relationship between features
- sensitive to the dispersion / variance of both the predictor and response

Feature Ranking Metrics

Ranking Method #4 - Pairwise Covariance

- Sensitive to feature variance
- Feature variance is somewhat arbitrary.
  - For example, what is the variance of porosity in fraction vs. percentage or permeaiblity in Darcy vs. milliDarcy. We can show that if we apply a constant multiplier, cc, to a variable, XX, that the variance will change according to this relationship (the proof is based on expectation formulation of variance):

$$\sigma_{cX}^2 = c^2 \sigma_X^2$$

  - By moving from percentage to fraction we decrease the variance of porosity by a factor of 10,000!
  - The variance of each variable is potentially arbitrary, with the exception when all the features are in the same units.

Feature Ranking Metrics

Ranking Method #5 - Pairwise Correlation Coefficient

- Pairwise correlation coefficient provides a measure of the strength of the linear relationship between each predictor feature and the response feature.

- The correlation coefficient:
  - measures the linear relationship
  - removes the sensitivity to the dispersion / variance of both the predictor and response features, by normalizing by the product of the standard deviation of each feature

Feature Ranking Metrics

Ranking Method #6 – Rank Correlation Coefficient

- The rank correlation coefficient applies the rank transform to the data prior to calculating the correlation coefficient. To calculate the rank transform simply replace the data values with the ranks, where n is the maximum value and 1 is the minimum value.
- The rank correlation:
  - measures the monotonic relationship, relaxes the linear assumption
  - removes the sensitivity to the dispersion / variance of both the predictor and response, by normalizing by the product of the standard deviation of each.

Feature Ranking Metrics

Ranking Method #7 – Partial Correlation Coefficient

This is a linear correlation coefficient that controls for the effects all the remaining variables

- $\rho_{XY.Z}$ and is the partial correlation between X and Y after controlling for Z.

1. perform linear, least-squares regression to predict X from $Z_{1,\ldots,m-2}$.
2. calculate the residuals in Step #1, $X - X^*$
3. perform linear, least-squares regression to predict Y from $Z_{1,\ldots,m-2}$.
4. calculate the residuals in Step #1, $Y - Y^*$
5. calculate the correlation coefficient, $\rho_{XY.Z} = \rho_{X - X^*, Y - Y^*}$

Feature Ranking Metrics

Ranking Method #7 – Partial Correlation Coefficient

The partial correlation, provides a measure of the linear relationship between X and Y while controlling for the effect of Z other features on both, X and Y

To use this method we must assume:

- two variables to compare, X and Y
- other variables to control, $Z_{1,\ldots,m-2}$.
- linear relationships between all variables
- no significant outliers
- approximately bivariate normality between the variables

We are in pretty good shape, but we have some departures from bivariate normality.

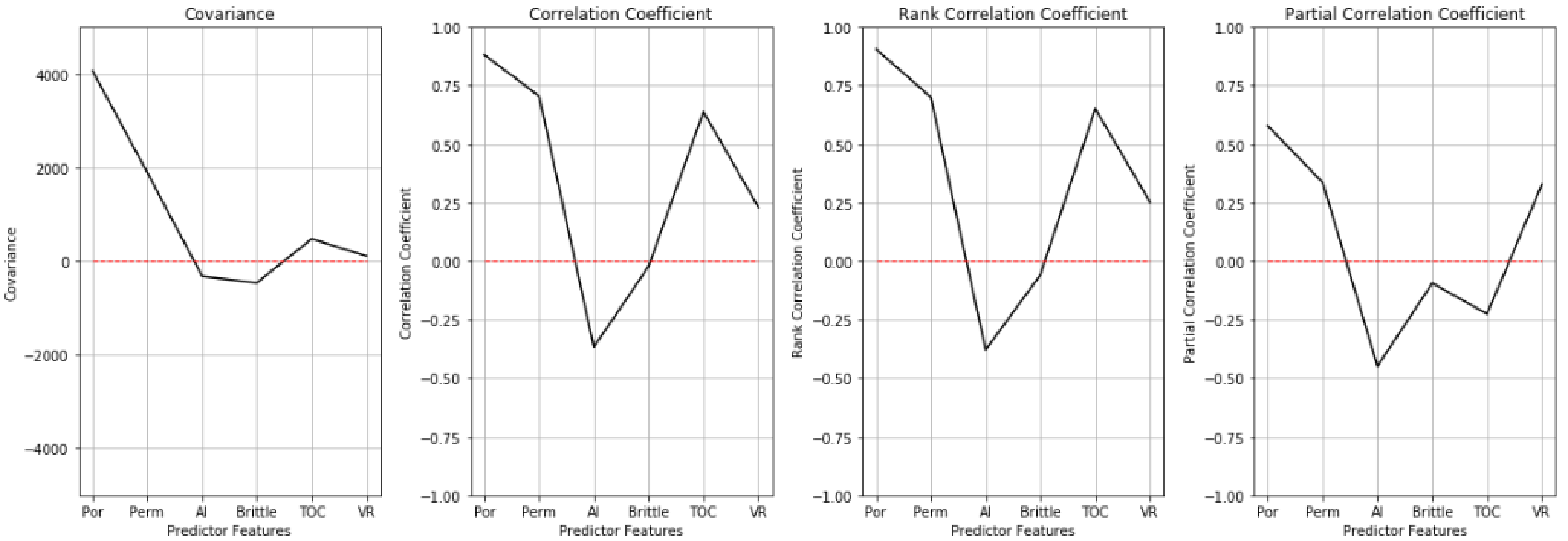
- We apply a Gaussian transform in the demonstration

Feature Ranking Metrics

Ranking Methods #4 - #7 – Results

Are we converging on porosity, permeability and vitrinite reflectance as the most important variables with respect to linear relationships with the production?

- What about brittleness?



Feature Ranking Metrics

Ranking Method # 9 – Model-based Ranking – B coefficients

- We could also consider *B* coefficients from linear regression.

$$Y^* = \sum_{i=1}^{m} B_i X_i + c$$

- These are the linear regression coefficients without standardization of the variables.
- Sensitive to feature variance.
- We are capturing interactions between variables.

Feature Ranking Metrics

Ranking Method # 9 – Model-based Ranking – $\mathbf{B}$ (beta) coefficients

- We could also consider *B* coefficients from linear regression

$$Y^{s*} = \sum_{i=1}^{m} B_i X_i^s + c$$

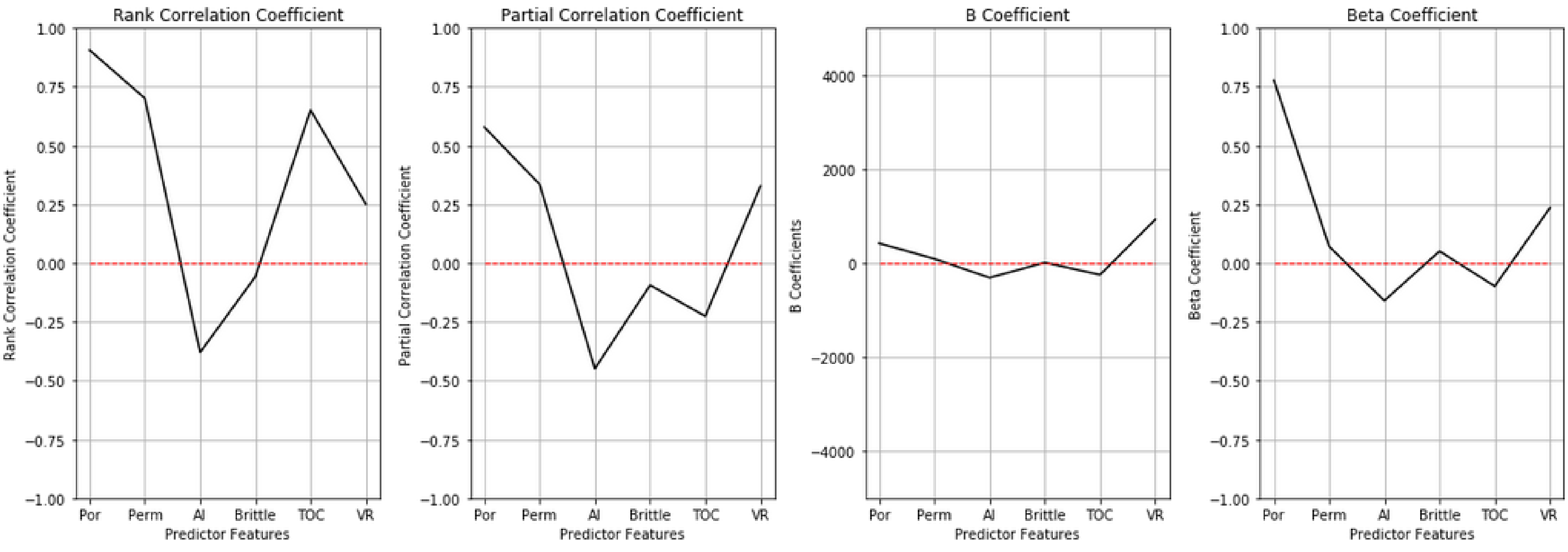
- These are the linear regression coefficients with standardization of the variables, X_i^s and Y^{s*} (variance = 1)
- Not sensitive to variance of the features
- We are capturing interactions between variables.

Feature Ranking Metrics

Ranking Methods #4 - #9 – Results

Now what do we see?

- Beta demotes permeability!
- Porosity, acoustic impedance and vitrinite reflectance retain high metrics



Feature Ranking Metrics

Ranking Methods #11– Recursive Feature Elimination

Recursive Feature Elimination (RFE) method works by recursively removing features and building a model with the remaining features.

- model accuracy is applied to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute
- any model could be used!
- in this example the prediction model based on multilinear regression and indicate that we want to find the best feature based on recursive feature elimination.
- the method assigns rank $1, \ldots, m$ for all features.

Feature Ranking Metrics

Ranking Methods #11– Recursive Feature Elimination

The recursive feature elimination method with a linear regression model provides these ranks:

1. Total Organic Carbon
2. Vitrinite Reflectance
3. Acoustic Impedance
4. Porosity
5. Permeability
6. Brittleness

A couple of the features moved from our previous assessment, but we are close. The advantages with the recursive elimination method:

- the actual model can be used in assessing feature ranks
- the ranking is based on accuracy of the estimate

Feature Ranking Metrics

Ranking Methods #11– Recursive Feature Elimination

The recursive feature elimination method with a linear regression model provides these ranks, but this method is sensitive to:

- choice of model
- training dataset

This method may be applied with cross validation (k fold iteration of training and testing datasets)

- optimize variable selection for prediction with testing data after training with training data

Feature Ranking Demonstration in Python

Demonstration of the wide array approach with a documented workflow.

GeostatsPy: Multivariate Analysis for Subsurface Data Analytics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

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

Subsurface Machine Learning: Feature Ranking for Subsurface Data Analytics

Here's a demonstration of feature ranking for subsurface modeling in Python. This is part of my Subsuface Machine Learning Course at the Cockrell School of Engineering at the University of Texas at Austin.

Variable Ranking

There are often many predictor features, input variables, available for us to work with for subsurface prediction. There are good reasons to be selective, throwing in every possible feature is not a good idea! In general, for the best prediction model, careful selection of the fewest features that provide the most amount of information is the best practice.

Here's why:

- more variables result in more complicated workflows that require more professional time and have increased opportunity for blunders
- higher dimensional feature sets are more difficult to visualize
- more complicated models may be more difficult to interrogate, interpret and QC
- inclusion of highly redundant and colinear variables increases model instability and decreases prediction accuracy in testing
- more variables generally increase the computational time required to train the model and the model may be less compact and portable
- the risk of overfit increases with the more variables, more complexity

What is Feature Ranking?

Feature ranking is a set of metrics that assign relative importance or value to each feature with respect to information contained for inference and importance in predicting a response feature. There are a wide variety of possible methods to accomplish this. My recommendation is a **'wide-array'** approach with multiple metric, while understanding the assumptions and limitations of each metric.

Here's the general types of metrics that we will consider for feature ranking.

1. Visual Inspection of Data Distributions and Scatter Plots
2. Statistical Summaries

Workflow at https://github.com/GeostatsGuy/PythonNumericalDemos/blob/master/GeostatsPy_variable_ranking.ipynb

Multivariate Modeling: Multivariate

Lecture outline . . .

- Multivariate Estimation

Instructor: Michael Pyrcz, the University of Texas at Austin

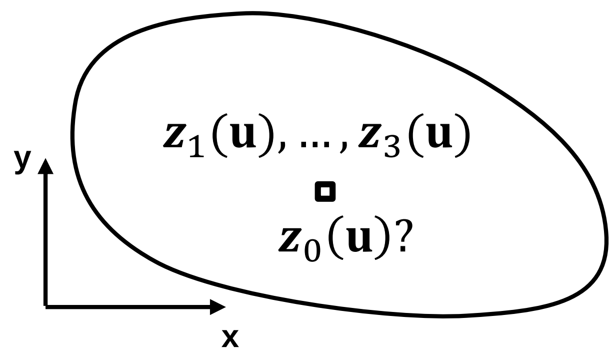
Multivariate Kriging

- Simple kriging may be applied to make estimates given a set of collocated secondary variables at the location to estimate the primary variable.
- This is not spatial estimation, but multivariate estimation!

Covariance between secondary variables | **Covariance between secondary and primary variables**

$$\begin{bmatrix} C(\mathbf{z}_1,\mathbf{z}_1) & C(\mathbf{z}_1,\mathbf{z}_2) & C(\mathbf{z}_1,\mathbf{z}_3) \\ C(\mathbf{z}_2,\mathbf{z}_1) & C(\mathbf{z}_2,\mathbf{z}_2) & C(\mathbf{z}_2,\mathbf{z}_3) \\ C(\mathbf{z}_3,\mathbf{z}_1) & C(\mathbf{z}_3,\mathbf{z}_2) & C(\mathbf{z}_3,\mathbf{z}_3) \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{bmatrix} = \begin{bmatrix} C(\mathbf{z}_o,\mathbf{z}_1) \\ C(\mathbf{z}_o,\mathbf{z}_2) \\ C(\mathbf{z}_o,\mathbf{z}_3) \end{bmatrix}$$

redundancy | **closeness**



- Given the assumption of Gaussian distributed variables we have a complete model of uncertainty for the primary variable at location $\mathbf{u}$!
- We can back transform for uncertainty in the original variable units.

Multivariate Kriging Hands-on

Here's an opportunity for experiential learning with Simple Kriging for multivariate estimation and uncertainty.

- **Things to try:**

Observe the multivariate weights, estimator and variance. Walk through the steps.

Kriging-based Multivariate Prediction
Michael Pyrcz, the University of Texas at Austin

This is an example, demonstration of kriging for multivariate prediction, instead of spatial prediction with comparision to multilinear regression.

X	Y	Facies	Porosity	Perm	AI	St_Por	St_Perm	St_AI	St_Por_Es	Por_est	Por_std	P10	P90
		mean	0.15	183.71	4203.66	0.00	0.00	0.00	0.00	0.15			
		st.dev.	0.05	345.44	1319.09	1.00	1.00	1.00	0.87	0.04			
									8. Estima	9. Back transfo		10. Calculate P10, P90	
100	900	1	0.1002	1.36389	5110.7	-1.0078	-0.5279	0.6876	-0.61957	0.1195	0.025	0.088	0.151
100	800	0	0.1079	12.5768	4671.459	-0.8519	-0.4954	0.3546	-0.35681	0.1326	0.025	0.101	0.164
100	700	0	0.0854	5.98452	6127.548	-1.3057	-0.5145	1.4585	-1.21305	0.09	0.025	0.058	0.122
100	600	0	0.1085	2.44668	5201.638	-0.8416	-0.5247	0.7566	-0.67232	0.1169	0.025	0.085	0.149
100	500	0	0.1025	1.95226	3835.27	-0.962	-0.5262	-0.2793	0.12793	0.1567	0.025	0.125	0.189
100	400	0	0.1106	3.69191	5295.267	-0.7991	-0.5211	0.8275	-0.72657	0.1142	0.025	0.082	0.146
100	300	0	0.0889	1.07358	6744.996	-1.2338	-0.5287	1.9266	-1.57716	0.0718	0.025	0.040	0.104
100	200	0	0.1021	2.39619	5947.338	-0.9695	-0.5249	1.3219	-1.10921	0.0951	0.025	0.063	0.127
100	100	1	0.1375	5.7276	5823.242	-0.2592	-0.5152	1.2278	-1.0349	0.0988	0.025	0.067	0.131
200	900	1	0.1371	14.7713	5621.147	-0.2671	-0.4891	1.0746	-0.91213	0.105	0.025	0.073	0.137
200	800	1	0.126	10.6754	4292.701	-0.4896	-0.5009	0.0675	-0.13584	0.1436	0.025	0.112	0.175
200	700	0	0.1218	3.08583	5397.4	-0.5746	-0.5229	0.905	-0.7867	0.1112	0.025	0.079	0.143
200	600	0	0.0951	0.96257	4619.786	-1.1091	-0.529	0.3155	-0.33216	0.1338	0.025	0.102	0.166
200	500	0	0.0875	1.82327	4949.881	-1.2627	-0.5265	0.5657	-0.52513	0.1242	0.025	0.092	0.156
200	400	0	0.0986	4.57102	5789.623	-1.04	-0.5186	1.2023	-1.01576	0.0998	0.025	0.068	0.132
200	300	0	0.1074	13.5819	7881.899	-0.8621	-0.4925	2.7885	-2.23717	0.039	0.025	0.007	0.071
200	200	0	0.0935	0.4088	6104.849	-1.1413	-0.5306	1.4413	-1.20245	0.0905	0.025	0.059	0.122
200	100	0	0.0799	0.73455	6485.732	-1.4144	-0.5297	1.73	-1.42543	0.0794	0.025	0.048	0.111
300	900	1	0.1115	27.9998	4183.467	-0.7802	-0.4508	-0.0153	-0.06347	0.1472	0.025	0.115	0.179
300	800	1	0.1195	61.0054	5224.544	-0.6205	-0.3552	0.7739	-0.65742	0.1176	0.025	0.086	0.149
300	700	1	0.1342	44.5959	4556.541	-0.3246	-0.4027	0.2675	-0.27401	0.1367	0.025	0.105	0.169
300	600	1	0.1612	190.74	3997.089	0.2179	0.0203	-0.1566	0.124417	0.1566	0.025	0.125	0.188
300	500	1	0.1072	13.3016	5684.915	-0.8675	-0.4933	1.1229	-0.9502	0.1031	0.025	0.071	0.135
300	400	1	0.1117	10.8511	5499.753	-0.7771	-0.5004	0.9826	-0.8429	0.1084	0.025	0.077	0.140
300	300	0	0.1052	6.3122	4715.684	-0.9064	-0.5135	0.3882	-0.38575	0.1312	0.025	0.099	0.163
300	200	0	0.1033	2.81735	6217.196	-0.9461	-0.5237	1.5265	-1.2671	0.0873	0.025	0.055	0.119
300	100	0	0.0991	3.18878	6456.338	-1.0305	-0.5226	1.7078	-1.40703	0.0803	0.025	0.048	0.112
400	900	0	0.0785	1.1098	5440.289	-1.4436	-0.5286	0.9375	-0.81278	0.1099	0.025	0.078	0.142
400	800	0	0.1104	4.11622	5956.295	-0.8017	-0.5199	1.3287	-1.11363	0.0949	0.025	0.063	0.127
400	700	1	0.1228	18.827	4937.337	-0.5538	-0.4773	0.5562	-0.50956	0.125	0.025	0.093	0.157
400	600	1	0.1208	5.97043	4040.573	-0.593	-0.5145	-0.1236	0.009597	0.1508	0.025	0.119	0.183
400	500	1	0.114	132.124	4459.013	-0.7311	-0.1493	0.1936	-0.17455	0.1417	0.025	0.110	0.174
400	400	1	0.0838	5.55878	5953.782	-1.3362	-0.5157	1.3268	-1.11146	0.095	0.025	0.063	0.127
400	300	1	0.118	1.99268	5557.638	-0.6503	-0.5261	1.0265	-0.8811	0.1065	0.025	0.075	0.138
400	200	0	0.1169	6.63131	6735.729	-0.6721	-0.5126	1.9196	-1.56904	0.0722	0.025	0.040	0.104
400	100	0	0.0664	0.03361	6124.087	-1.6869	-0.5317	1.4559	-1.2139	0.0899	0.025	0.058	0.122
500	900	0	0.0934	1.30794	4873.951	-1.1448	-0.528	0.5081	-0.48089	0.1264	0.025	0.095	0.158
500	800	0	0.084	0.7714	5284.836	-1.3326	-0.5296	0.8196	-0.72187	0.1144	0.025	0.083	0.146
500	700	0	0.1098	52.5009	6581.836	-0.8143	-0.3798	1.8029	-1.4567	0.0778	0.025	0.046	0.110
500	600	1	0.122	8.22227	5296.785	-0.5702	-0.508	0.8287	-0.72527	0.1143	0.025	0.082	0.146

Workflow Steps:

0. Standardize the variables, the mean and variance have been corrected to 0 and 1 respectively with affine correction (we should use Gaussian transform for the complete workflow),
1 - 7. Multivariate kriging to calculate kriging estimate and kriging variance.
8. Estimate standardized porosity at each location.
9. Calculate the back transform of the kriging estimate and the kriging variance and 10. the local P10 and P90.

1. Calculate the correlation

	Por	Perm	AI
Por	1.00	0.54	-0.85
Perm	0.54	1.00	-0.49
AI	-0.85	-0.49	1.00

2. Calculate the covariance of the standardized variables

	Por	Perm	AI
Por	1.00	0.54	-0.85
Perm	0.54	1.00	-0.49
AI	-0.85	-0.49	1.00

3. Redudancy Matrix

	Perm	AI
Perm	1	-0.49
AI	-0.49	1

4. Closeness Matrix

	Por
Perm	0.54
AI	-0.85

5. Invert Redundancy Matrix

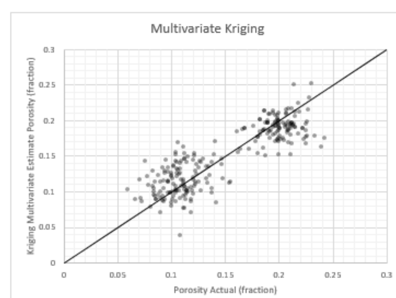
	Perm	AI
Perm	1.31	0.64
AI	0.64	1.31

6. Calculate the Weights

	Weights
Perm	0.17
AI	-0.77

7. Calculate the kriging variance

Kriging Variance 0.25



Comparision to Multilinear Regression

-0.7736	0.1673	-2.06E-16
0.0356	0.0356	0.0310267
0.7507	0.5013	#N/A
388.45	258	#N/A
195.2	64.823	#N/A

b2: slope of fit	-0.774	
se2: standard error of slope	0.036	**se1:** standard er
r2: proportion var. explained	0.75	**sey:** standard er
Fstat: for test of all coefficients	388.4	**d.f.:** d
ssreg: explained variance	195.2	**ssresid:** une

Test Significance of Coefficients

$H_0: b_i = 0$
$H_1: b_i \neq 0$

	tstat b1 = b1/se1		tstat b0 = b0/se0	tcritical
tstat	21.72	*tstat*	4.69	2.25
probability	######	*probability*	######	

	AI	Perm
Result from Hyposthesis tests for coefficients	**Reject H0: Slop**	**Reject H0: Intercept = 0**

File Name: Kriging_Multivariate_Estimation_Demo.xlsx File is at: https://git.io/fhALF

Multivariate New Tools

Topic	Application to Subsurface Modeling
Curse of Dimensionality	Reduce problem to lowest dimension possible. *Feature ranking determined that porosity may be predicted from acoustic impedance and rock type alone.*
Feature Selection	Apply wide array methods to explore the importance of each predictor feature with respect to the response feature. *Partial correlation reveals that rock type provides little additional information to acoustic impedance.*
Multivariate Kriging	*Multivariate kriging combines secondary information sources while accounting for closeness and redundancy.* *Given secondary data the likelihood distribution for local porosity is mean of 15% and standard deviation of 2.5% with a Gaussian distribution.*

Multivariate Modeling: Multivariate

Lecture outline . . .

- Multivariate Analysis
- Joints and Conditionals
- Feature Selection
- Multivariate Estimation

Instructor: Michael Pyrcz, the University of Texas at Austin