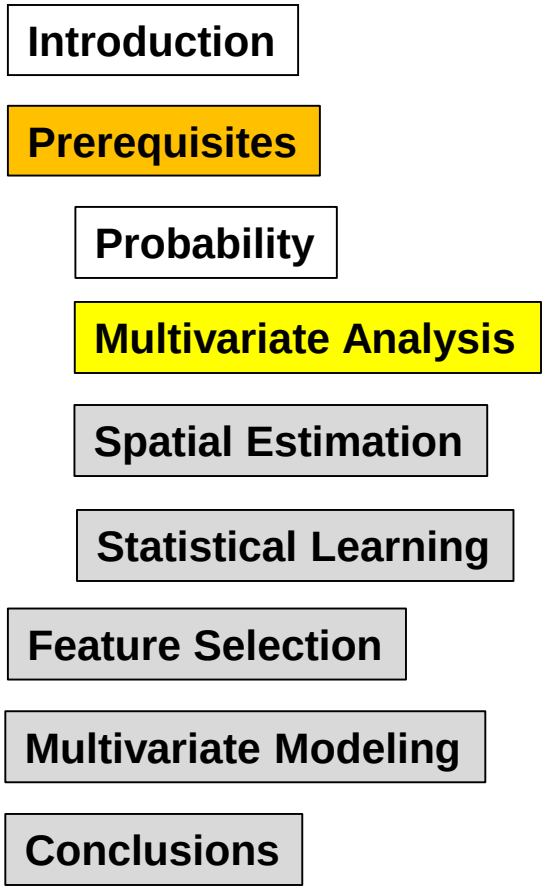
Multivariate Modeling: Multivariate

Lecture outline . . .

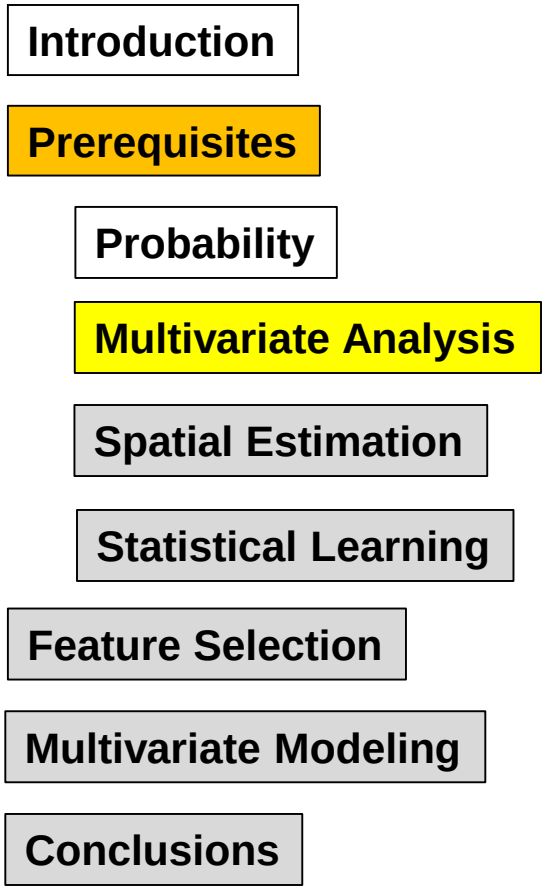
- **Bivariate Analysis**
- **Covariance and Correlation**
- **Marginal, Conditional and Joint**



Multivariate Modeling: Multivariate

Lecture outline . . .

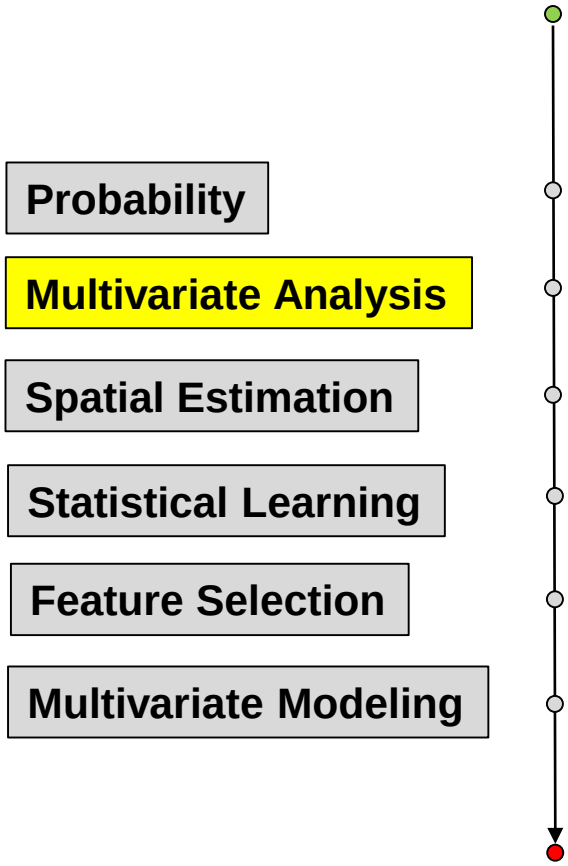
- **Mulitvariate / Bivariate Analysis**



What Will You Learn?

Why Cover Multivariate Analysis?

- We will work with multiple variables
- Will use some of these measures for variable ranking
- We need concepts of marginal, joint and conditional probabilities and distributions.



Multivariate, Spatial Uncertainty

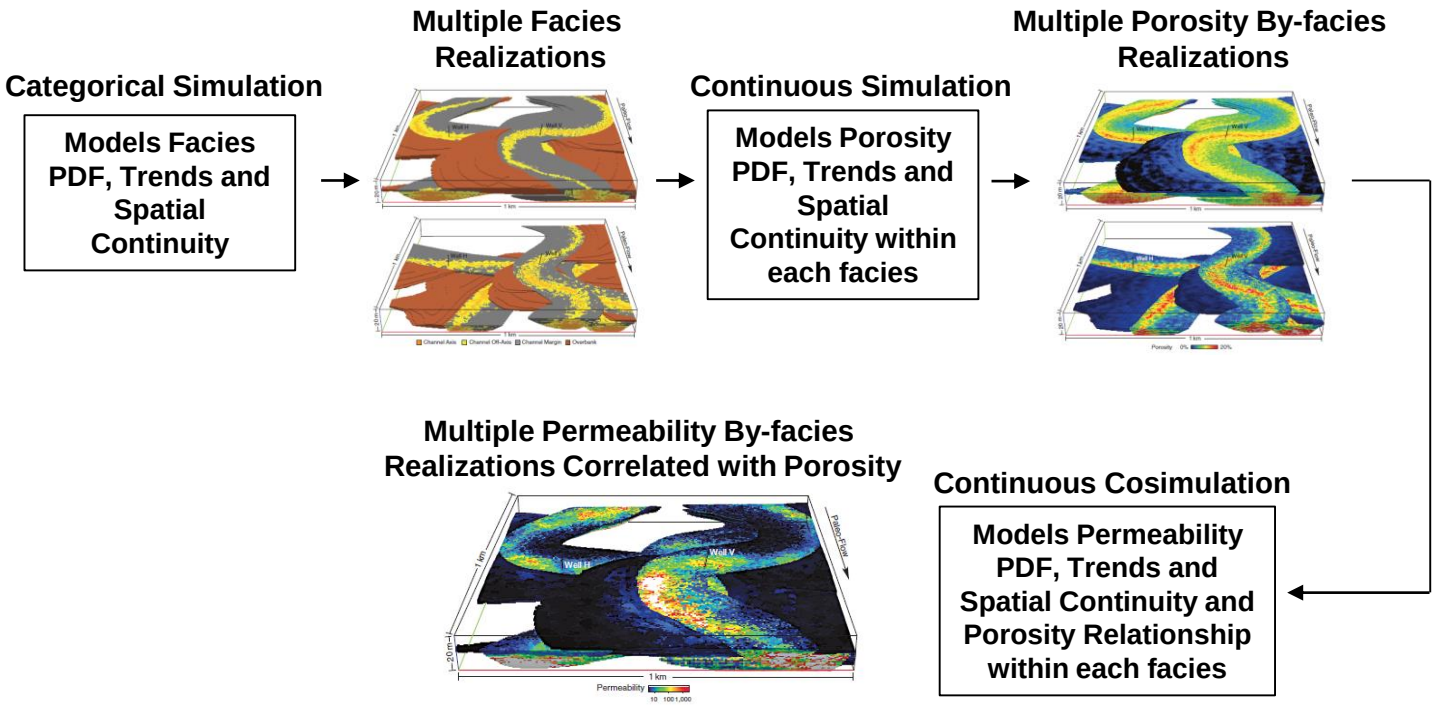
Motivation for Multivariate Methods

- **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

Motivation for Multivariate Methods

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



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

Motivation for Multivariate Methods

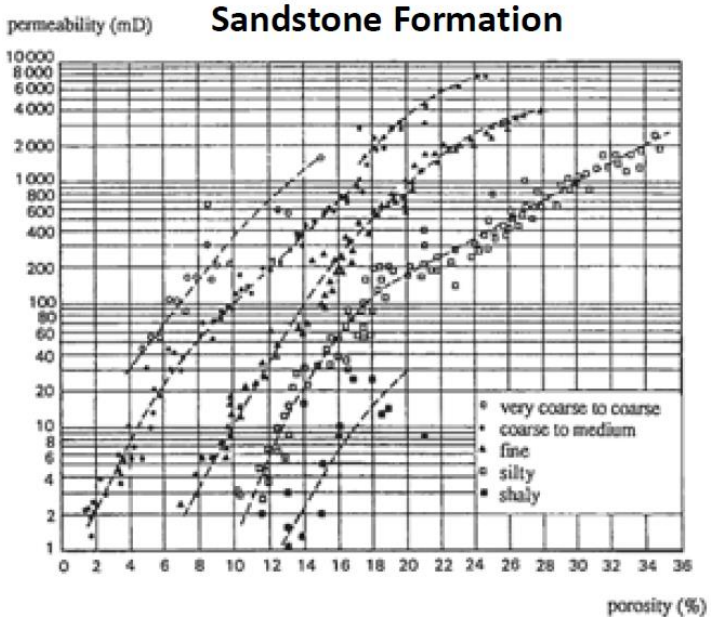
- **Emerging Multivariate Methods Include:**
  - Transforms – remove correlations and then model with independent variables and then back-transform to restore correlation (e.g. step-wise conditional transform).

  Beyond the scope of this course as they are not in common practice.

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



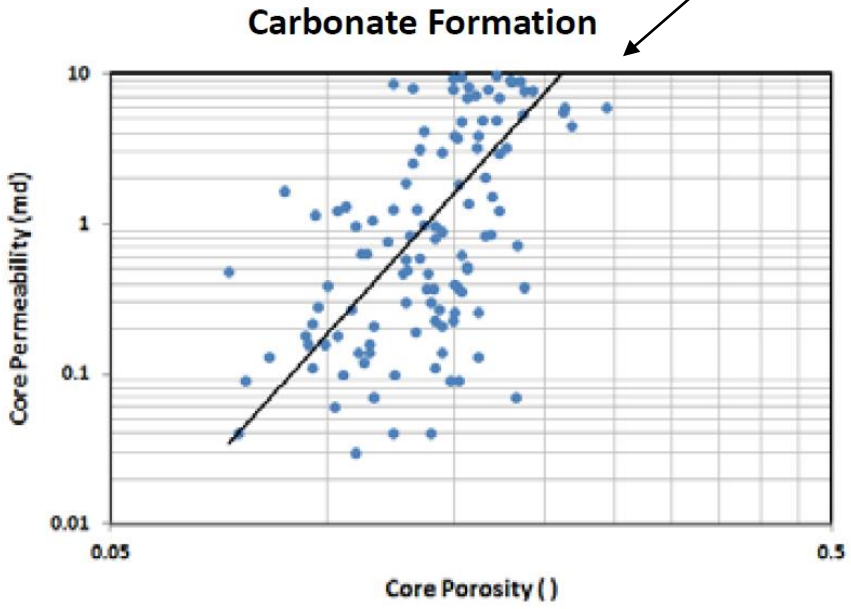
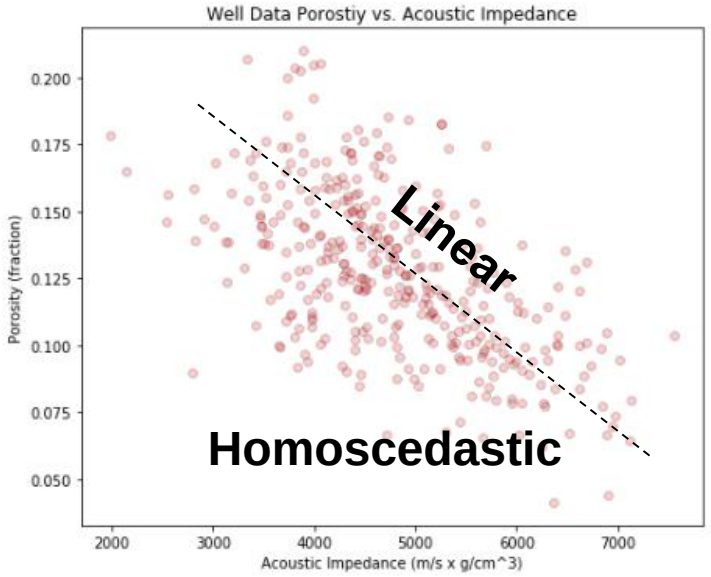


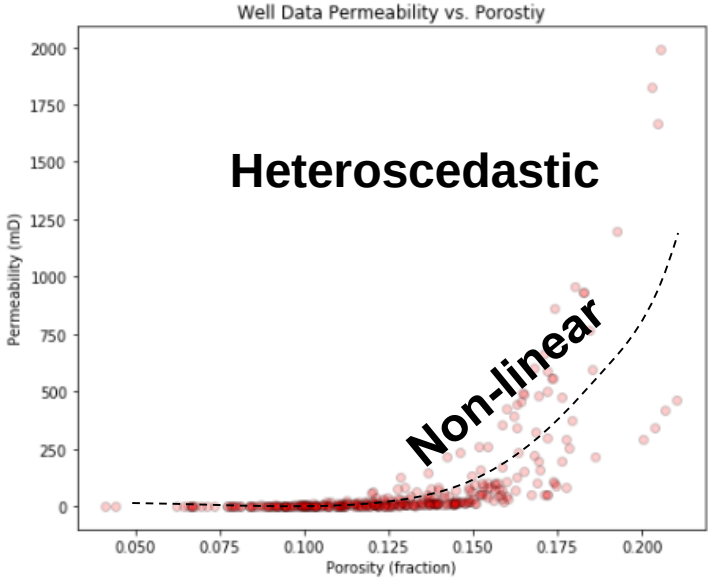
Figure from Peters, E. J., 2012, Advanced Petrophysics.

Slide from Dr. Zoya Heidari's PGE 337 Course

Bivariate Statistics
What is Bivariate Analysis?

- **Examples of bivariate structures**

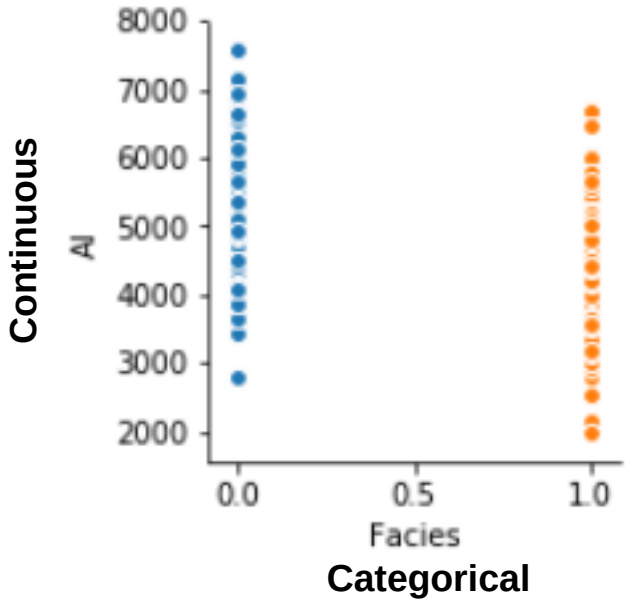




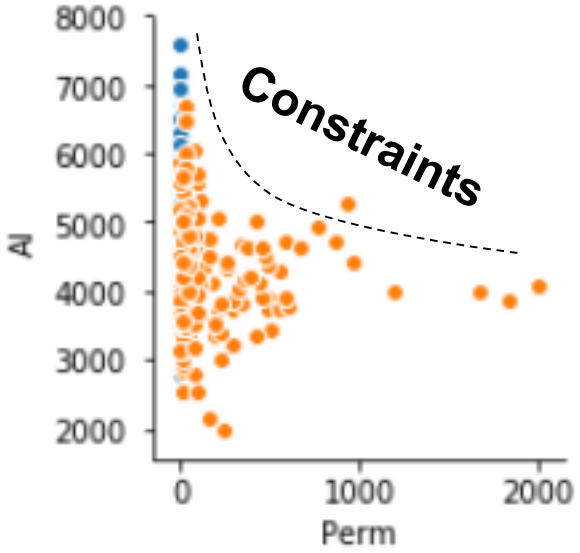
- Linear / Nonlinear – shape of the conditional expectation Y | X
- Homoscedastic / Heteroscedastic – conditional variance of Y | X

Bivariate Statistics
What is Bivariate Analysis?

- **Examples of bivariate structures**



Categorical

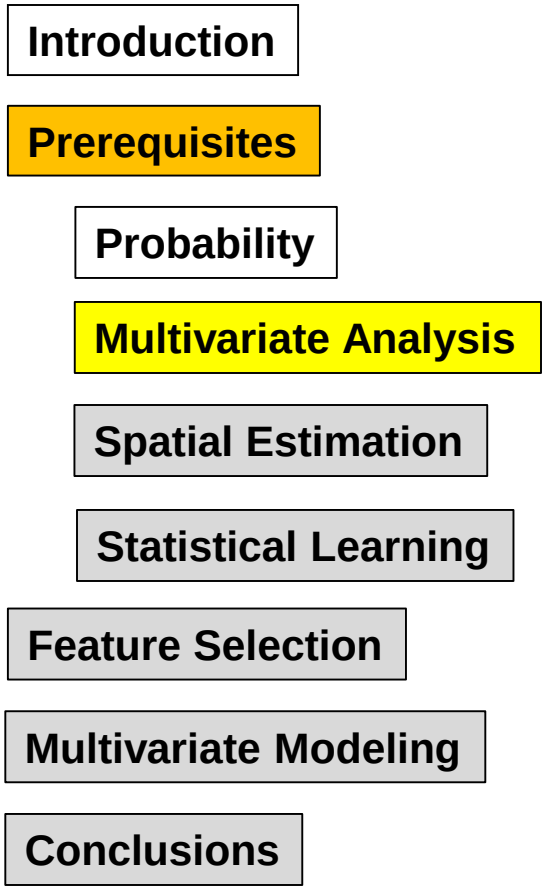


- Categorical variables only have a specified number of possible outcomes, continuous takes on a range of possible outcomes.
- Constraints – specific combinations of variables are not possible.

Multivariate Modeling: Multivariate

Lecture outline . . .

- **Covariance and Correlation**



Bivariate Statistics
Pearson's Correlation Coefficient

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

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

means of variables x and y

Correlation coefficient of variables x and y

number of data pairs

standard deviation of variables x and y

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

$$C_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}$$ Covariance $$\rho_{xy} = \frac{C_{xy}}{\sigma_x\sigma_y}$$

Bivariate Statistics
Variance and Covariance

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

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

A measure of how 2 variables vary together.

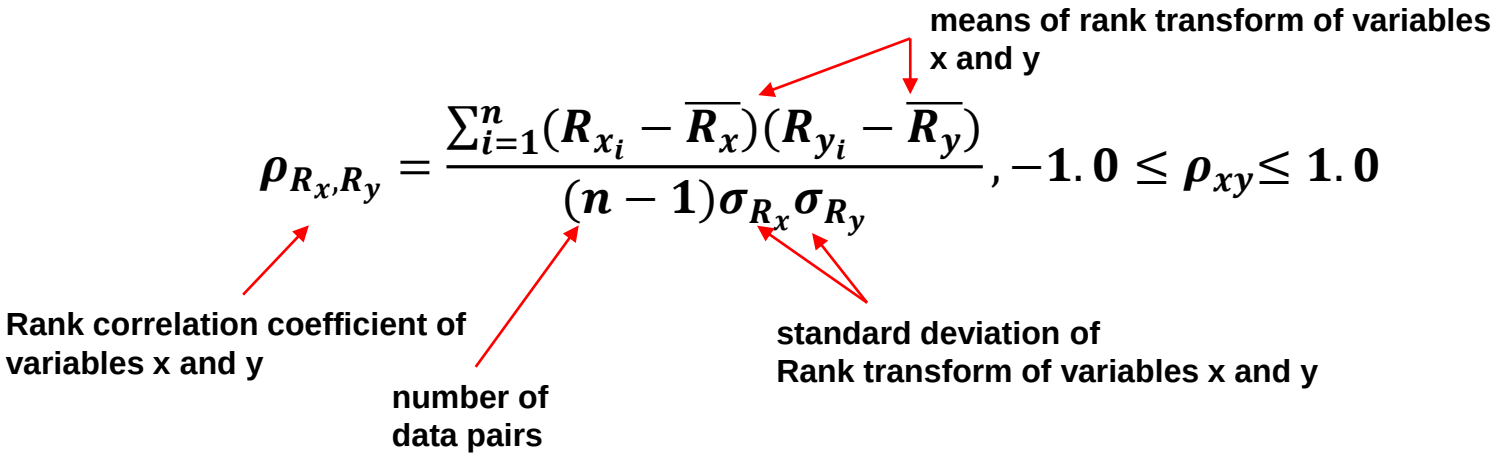
  - **Variance:**

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

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.**



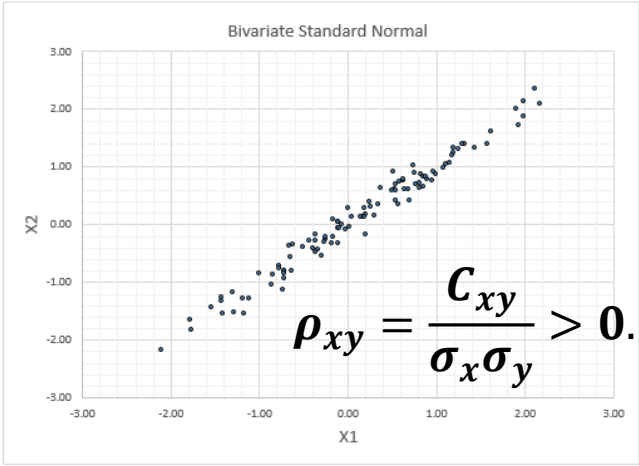
  - **Rank transform, e.g. R_{x_i}, sort the data in ascending order and replace the data with the index, $i = 1, \ldots, n$.**
  - **Spearman's rank correlation coefficient is more robust in the presence of outliers and some nonlinear features than the Pearson's correlation coefficient**

Bivariate Statistics
Covariance

- **Let's think about covariance. For a thought experiment, consider 2 standard normal variables, N[0,1].**

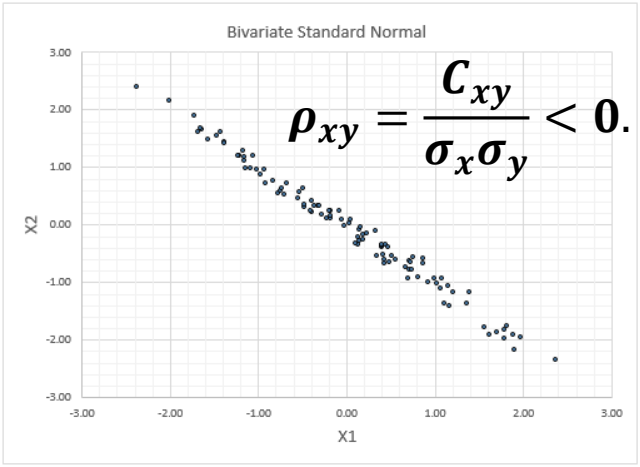
$$C_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)}$$

$$C_{xy} \sim E\{X\,Y\}$$



$$\rho_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} > 0.0$$

$$if\ \rho > 0,\ \uparrow C_{xy},\ \sum_{i=1}^{n/2}[(+x^- \times +y^-] + \sum_{i=n/2}^{n}[(+x^+ \times +y^+]$$

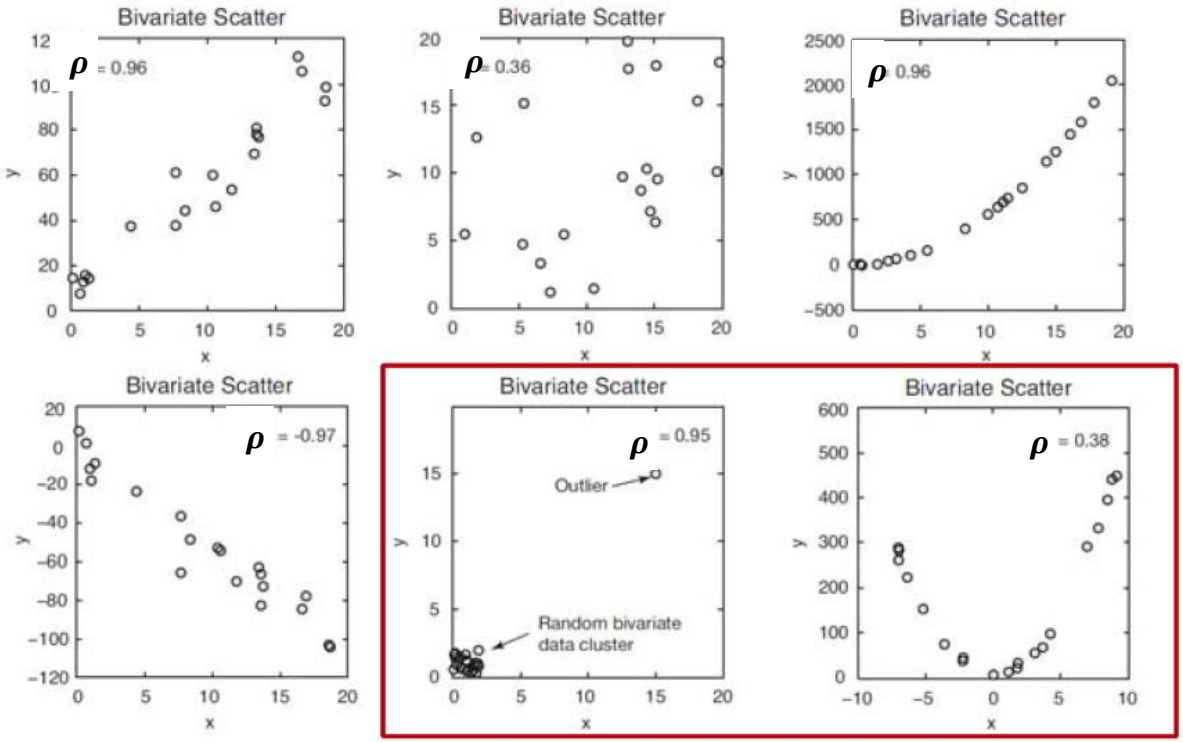


$$\rho_{xy} = \frac{C_{xy}}{\sigma_x \sigma_y} < 0.0$$

$$if\ \rho < 0,\ \uparrow C_{xy},\ \sum_{i=1}^{n/2}[(+x^- \times +y^+] + \sum_{i=n/2}^{n}[(+x^+ \times +y^-]$$

Bivariate Statistics
Pearson's Correlation Coefficient

- **Interpreting the correlation coefficient**



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

Slide from Dr. Zoya Heidari's PGE 337 Course

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!

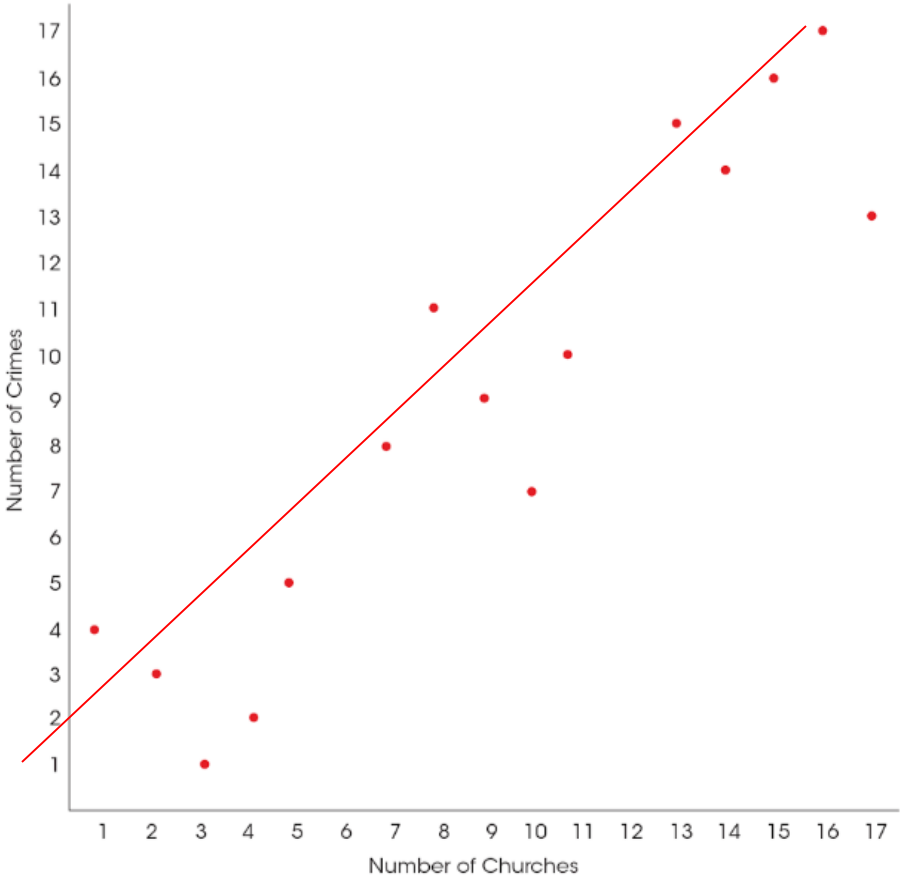


FIGURE 15.10
Hypothetical data showing the logical relationship between the number of churches and the number of crimes for three groups of cities: those with small populations ($Z = 1$), those with medium populations ($Z = 2$), and those with large populations ($Z = 3$).

Bivariate Statistics
Correlation and Causation

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

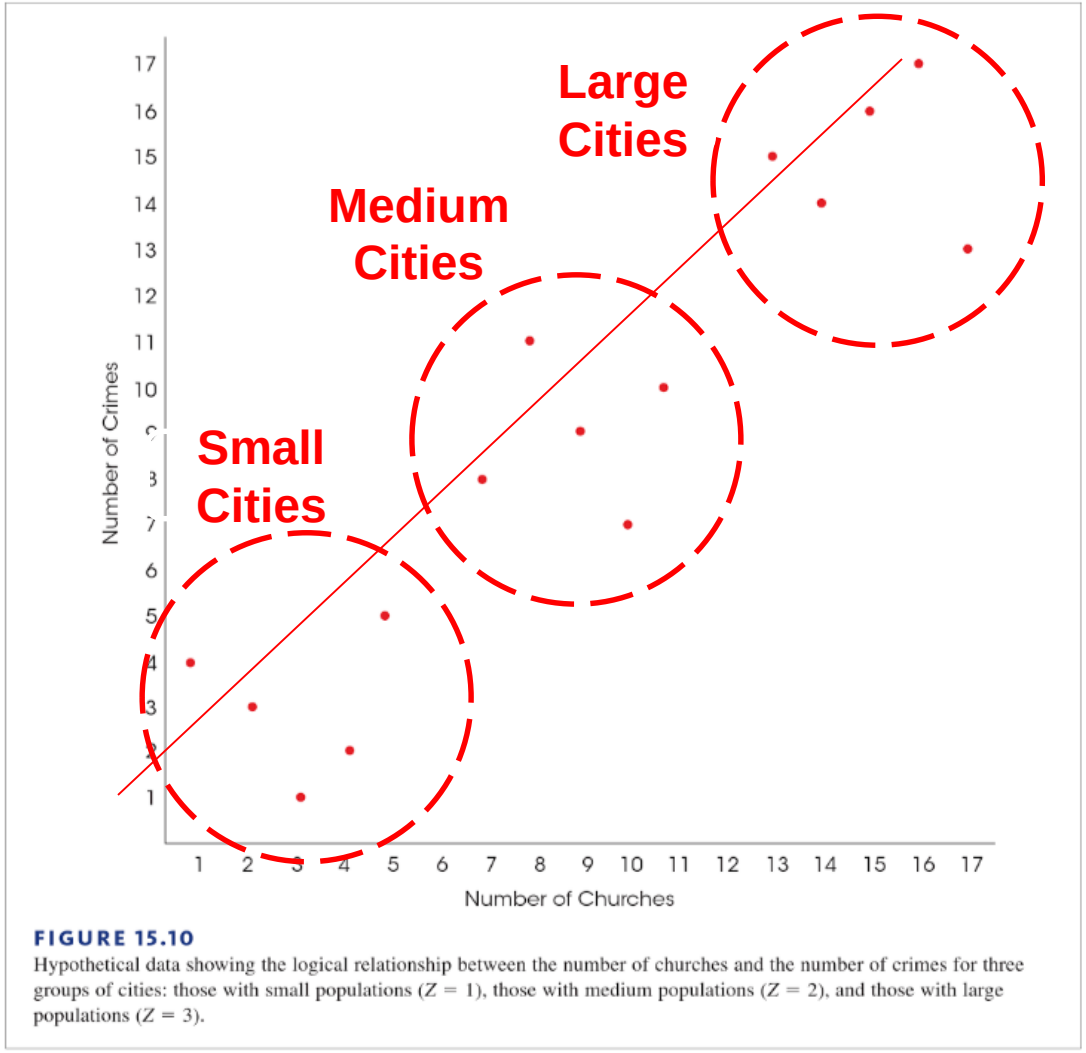


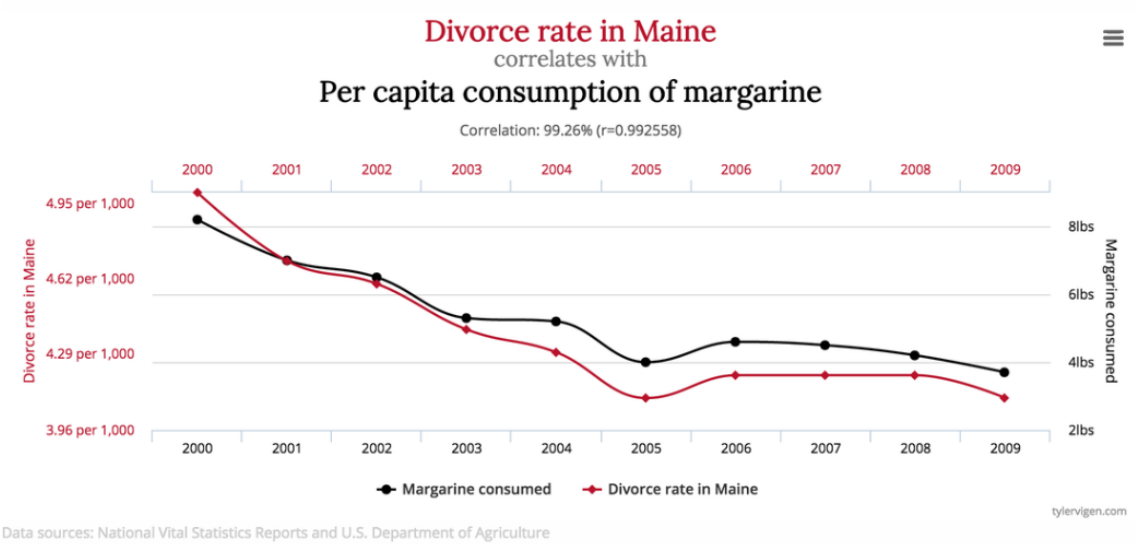
FIGURE 15.10
Hypothetical data showing the logical relationship between the number of churches and the number of crimes for three groups of cities: those with small populations ($Z = 1$), those with medium populations ($Z = 2$), and those with large populations ($Z = 3$).

Bivariate Statistics

Comical Examples of Correlation and Causation

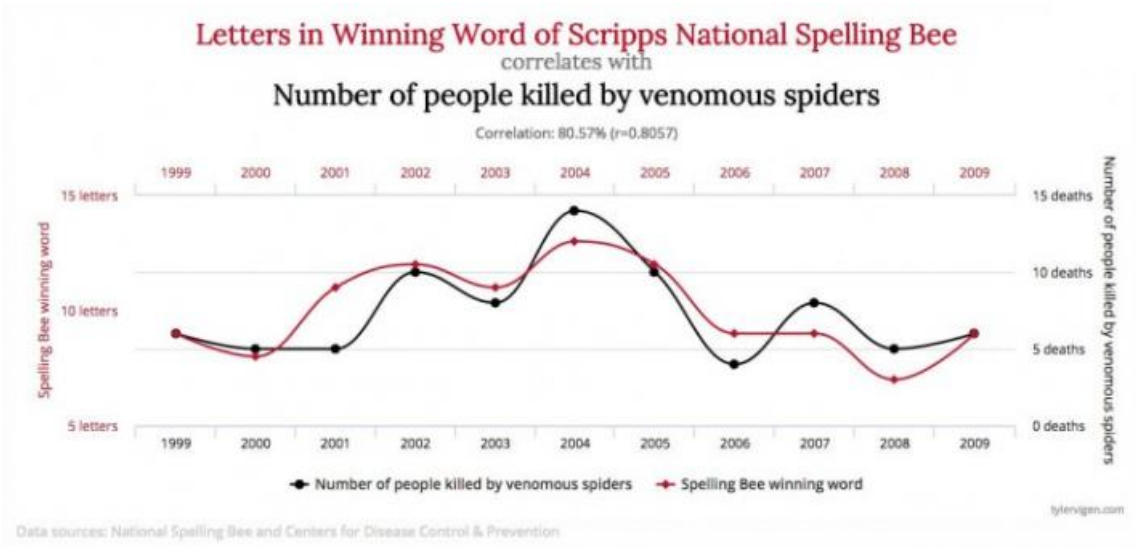
Margarine causes **divorce**?

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**?



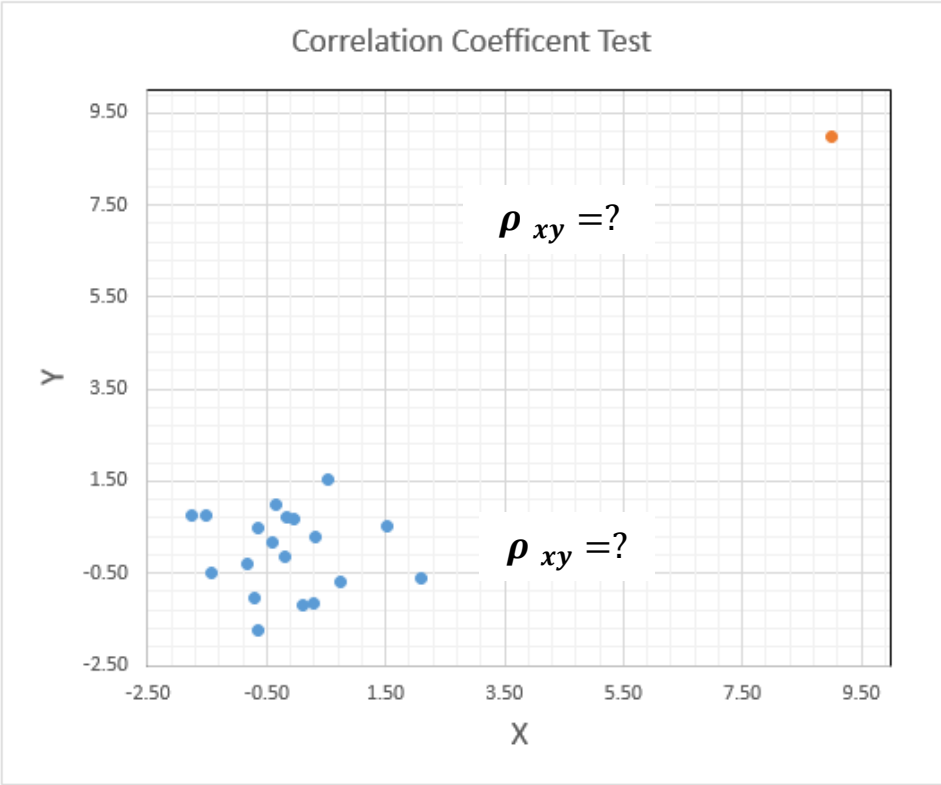
Bivariate Statistics

Exercise with Pearson's Correlation Coefficient

- Task 1: Generate a random data set of 19 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)

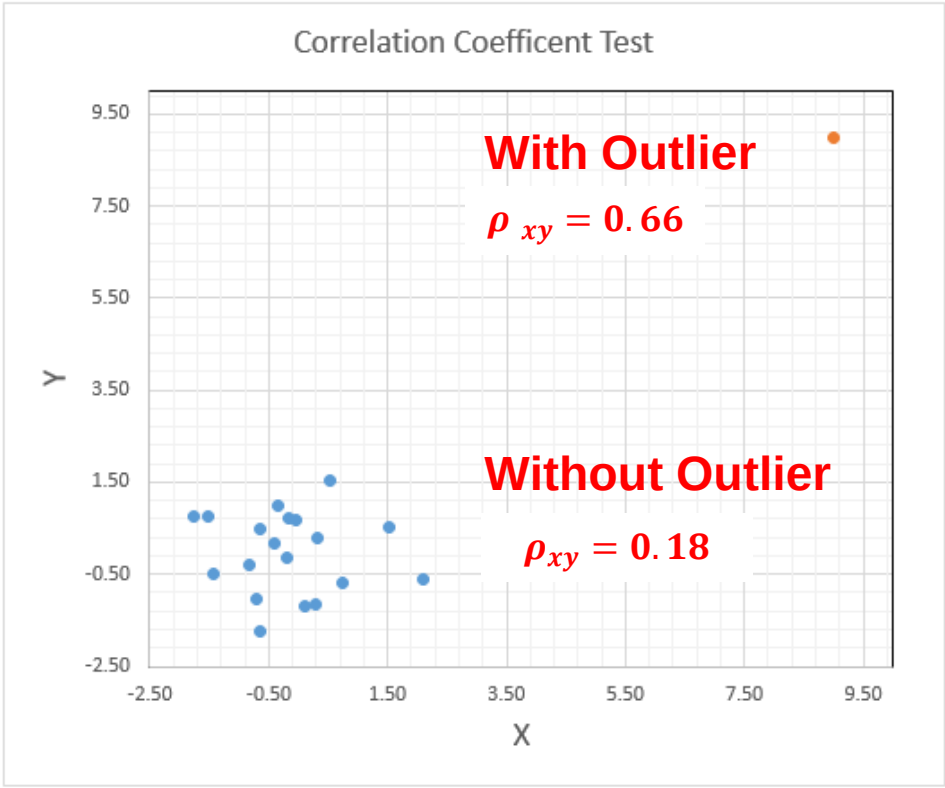
Bivariate Statistics

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?



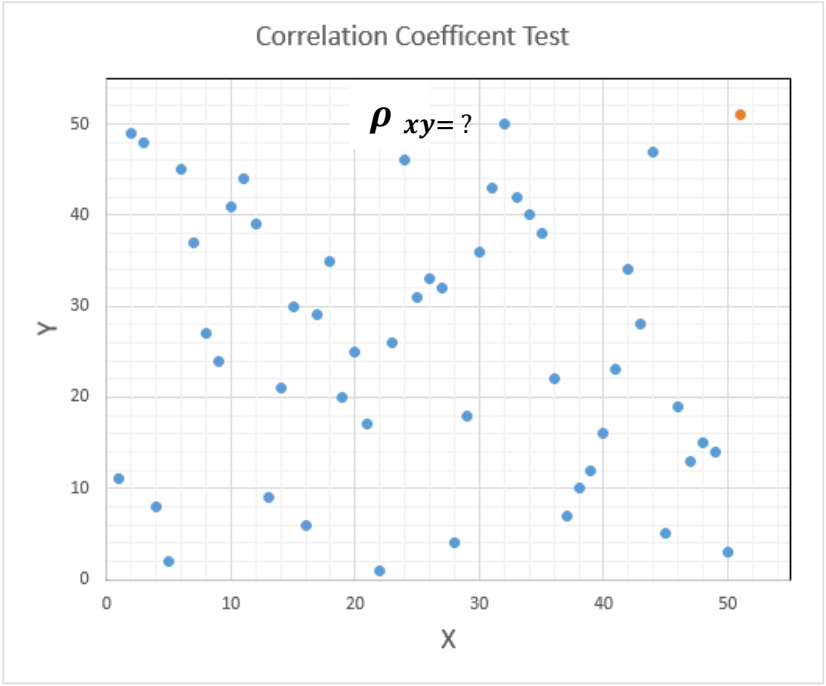
Bivariate Statistics

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.

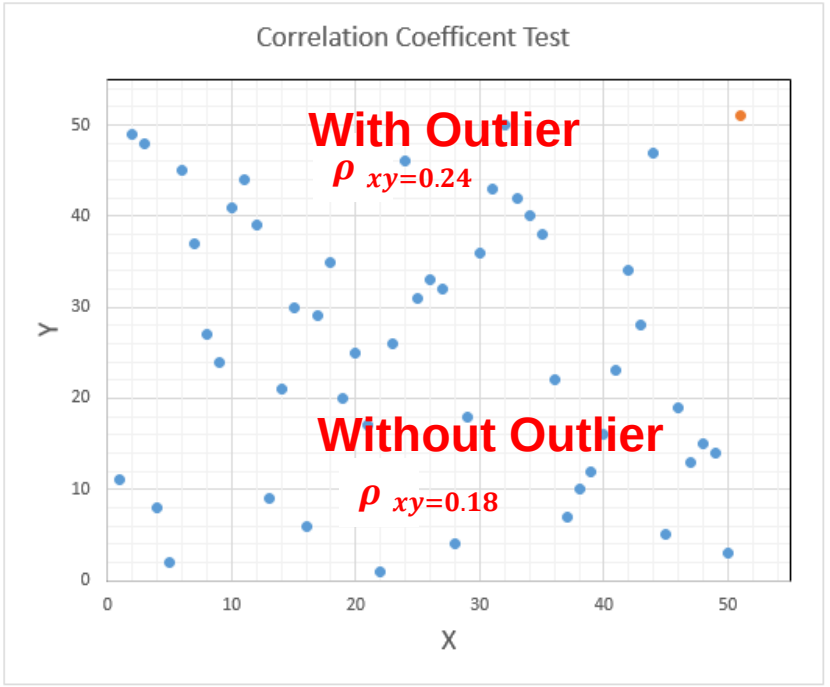


Bivariate Statistics

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)

What is the solution to this issue?

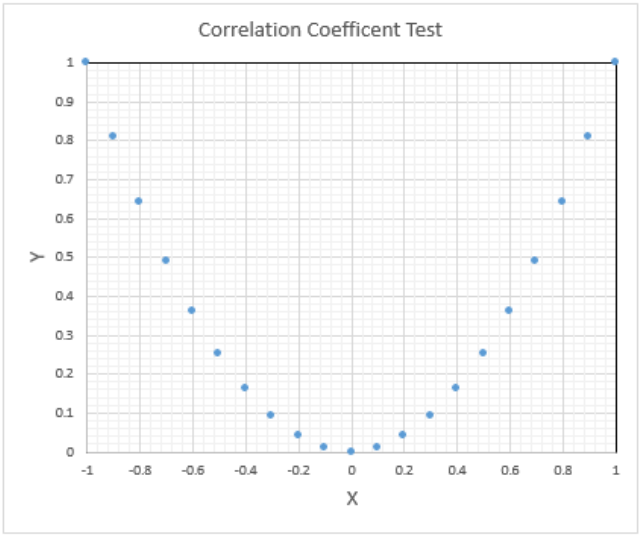
Bootstrap sampling
Jackknife sampling

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)

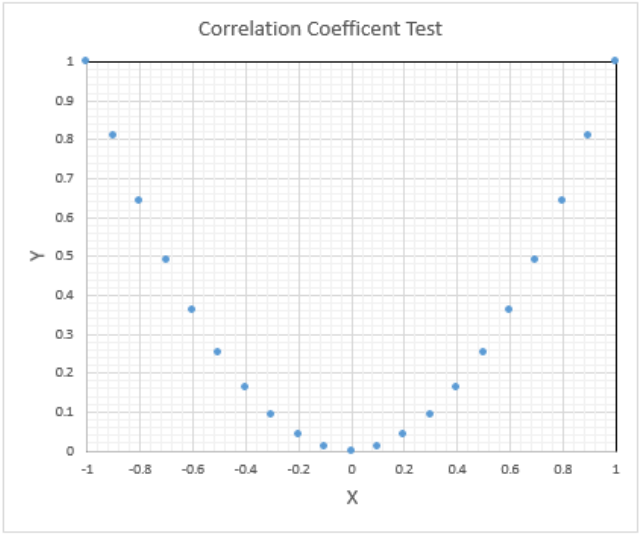
Bivariate Statistics

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]?

  Correlation Coefficient, ρ_{xy} = 0.0!



- Over range [0,1]?

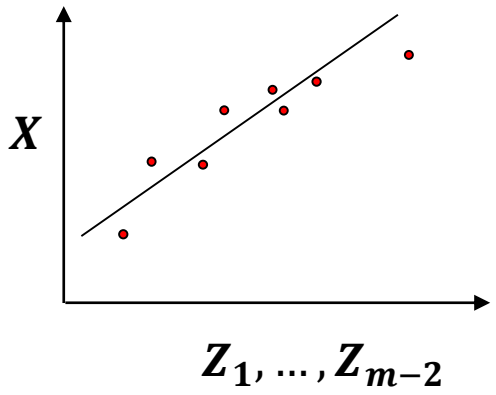
  Correlation Coefficient, ρ_{xy} = 0.96,
  Rank Correlation Coefficient, ρ_{RxRy} = 1.0

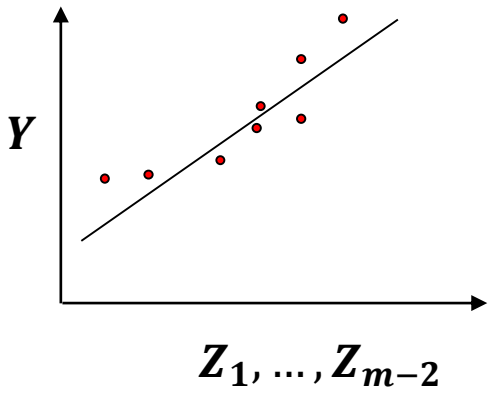
Excel Function Correl(array1,array2)

Bivariate Statistics
Partial Correlation

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

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

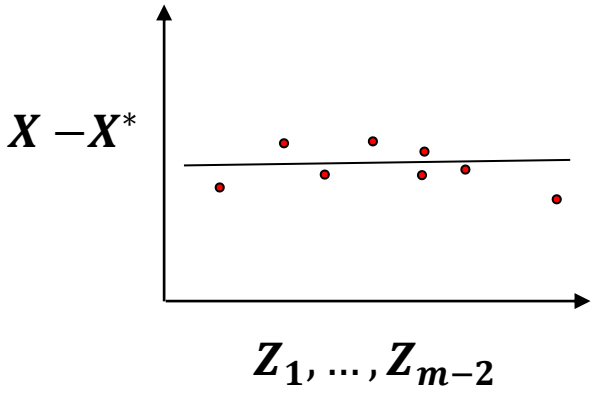


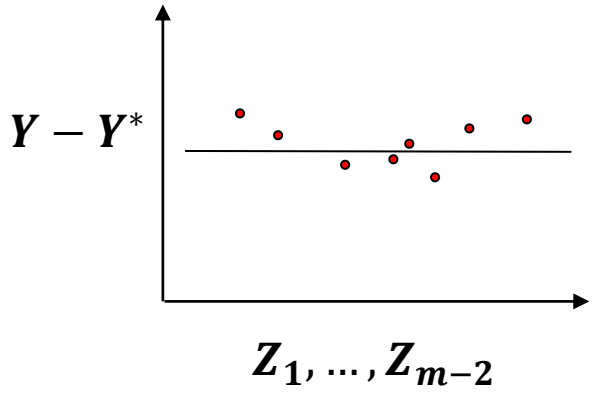


Bivariate Statistics
Partial Correlation

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

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

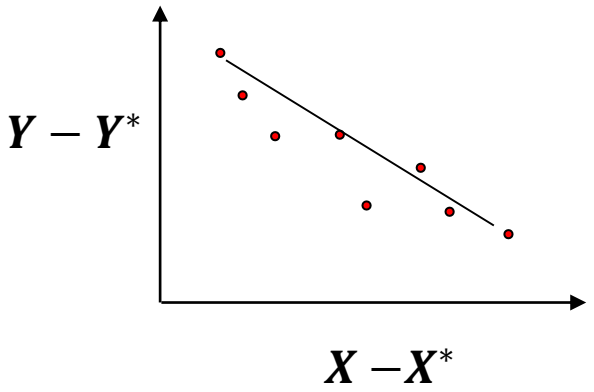




Bivariate Statistics
Partial Correlation

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

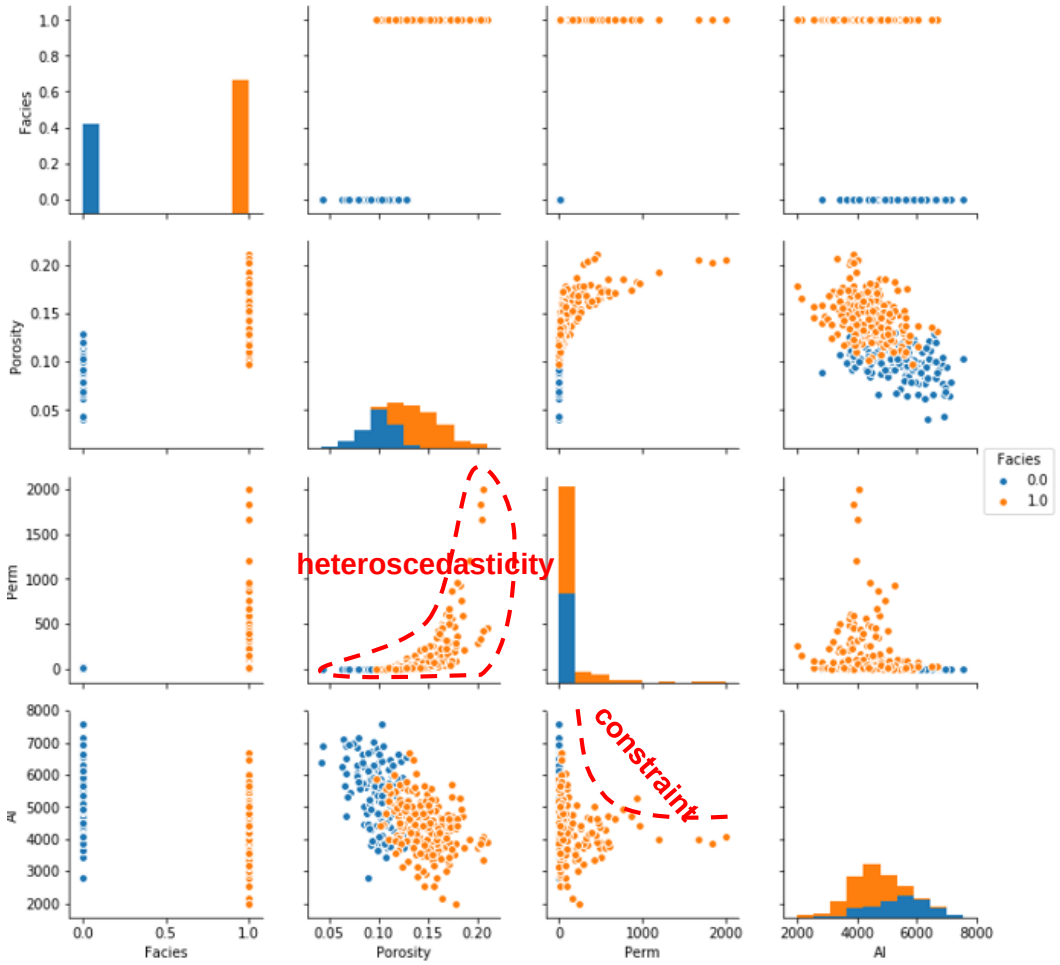
5. calculate the correlation coefficient between the residuals from Steps #3 and #4, $\rho_{\boldsymbol{X-X^*,Y-Y^*}}$



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

Bivariate Statistics
Matrix Scatter Plots

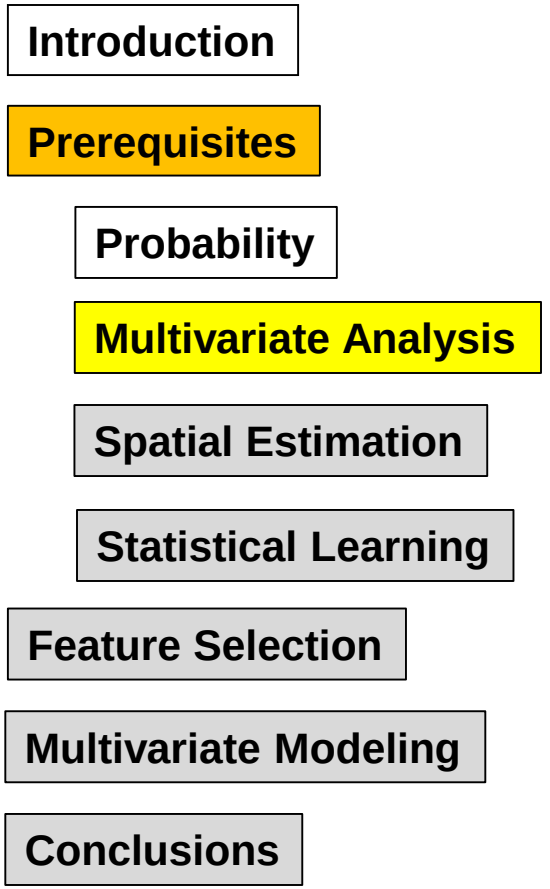
- For more than two variables make matrix scatterplots
  - By hand in Excel or packages in R and Python.
  - Look for linear / nonlinear features
  - Look for homoscedasticity (constant conditional variance) and heteroscedasticity (conditional variance changes with value)
  - Look for constraints



Multivariate Modeling: Multivariate

Lecture outline . . .

- **Working Directly with Marginal, Conditional and Joint**



Probability Definitions
Conditional, Marginal and Joint Probability

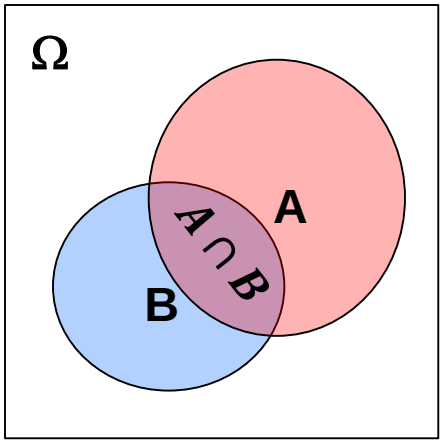
Probability of B given A occurred? $P(B \mid A)$

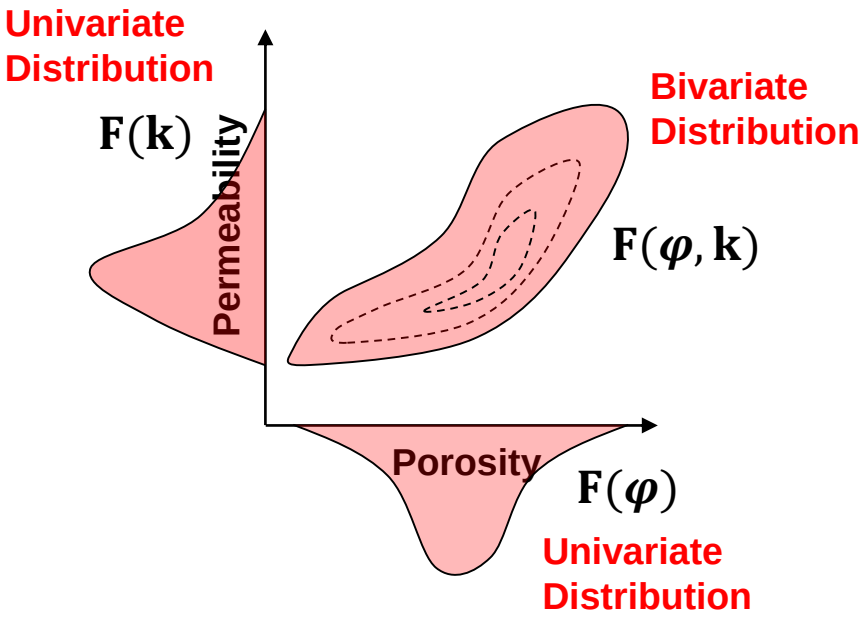
Conditional Probability

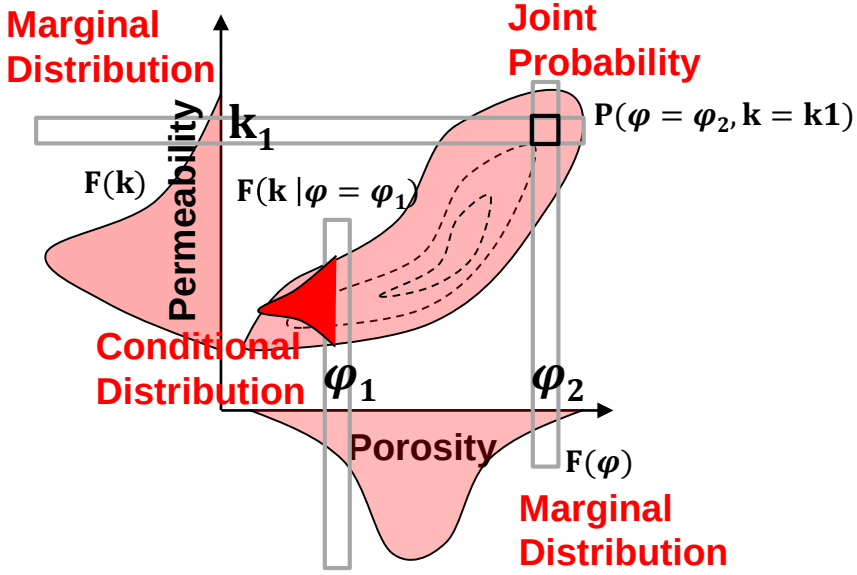
Joint Probability

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A \; and \; B)}{P(A)}$$

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.

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

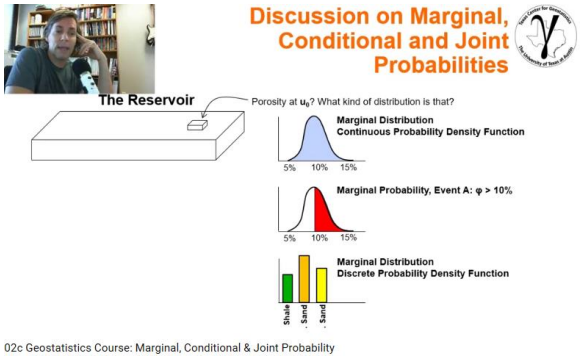
$$P(X \mid Y), P(Y|X)$$

Joint Probability: Probability of multiple events occurring together.

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

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

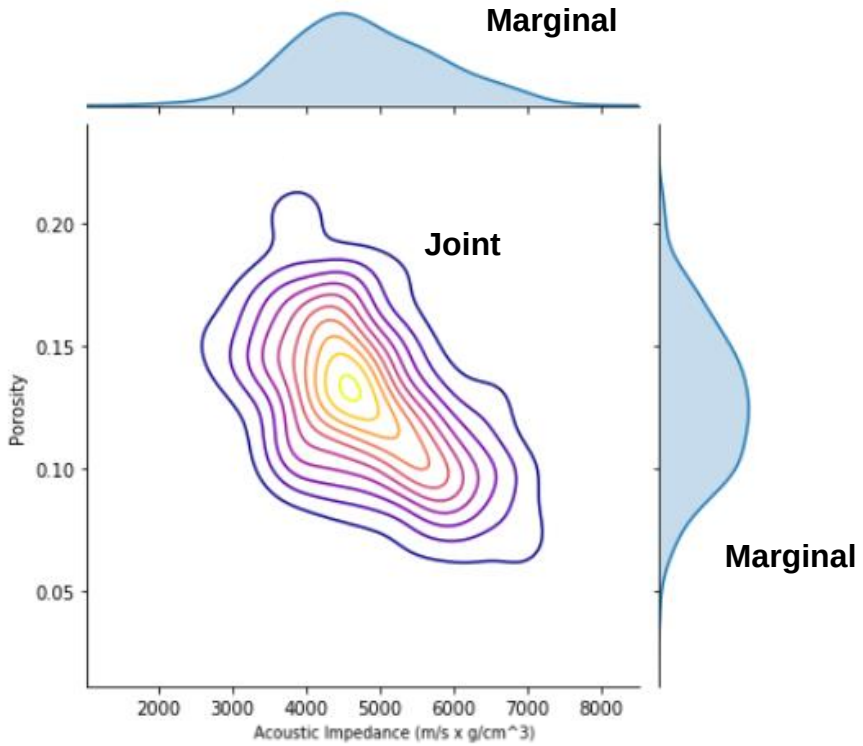
$$P(X, Y), P(Y, X)$$



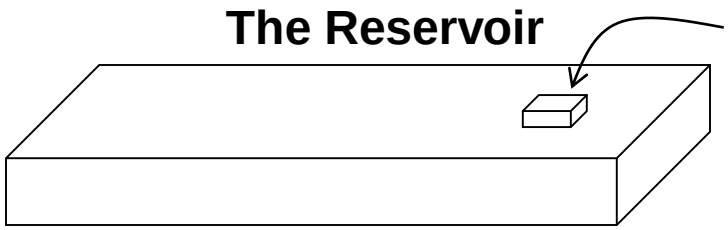
See YouTube Video on Marginals, Conditionals and Joints! https://www.youtube.com/watch?v=bL2gPwMfYpc&index=5&t=0s&list=PLG19vXLQHvSB-D4XKYieEku9GQMQyAzjJ

Marginal, Conditional and Joint Probability

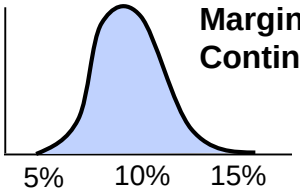
- Working directly with marginal, conditional and joint probability
  - If you have enough data, you can directly calculate all the required probabilities
  - Go beyond statistics like correlation coefficient



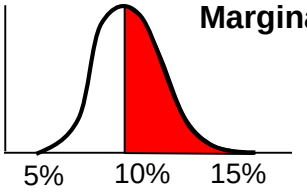
Marginal, Conditional and Joint Probability



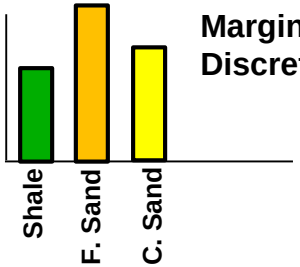
Porosity at $\mathbf{u_0}$? What kind of distribution is that?



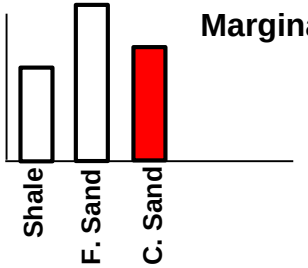
Marginal Distribution
Continuous Probability Density Function



Marginal Probability, Event A: φ > 10%



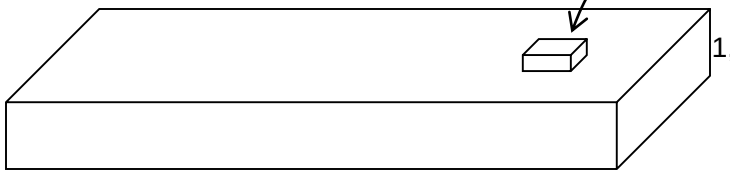
Marginal Distribution
Discrete Probability Density Function



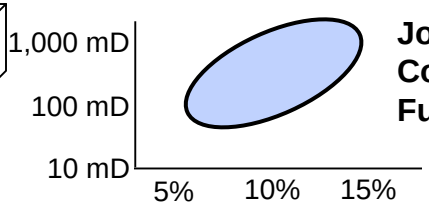
Marginal Probability, Event B: Facies = C. Sand

Marginal, Conditional and Joint Probability

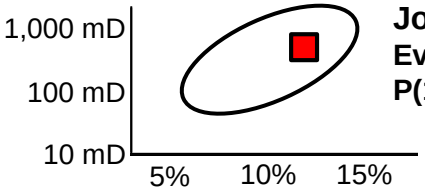
The Reservoir



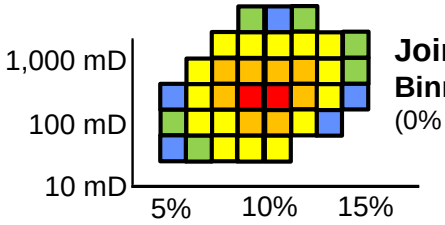
Porosity and Permeability at $\mathbf{u_0}$? What kind of distribution is that?



Joint Distribution
Continuous Joint Probability Density Function

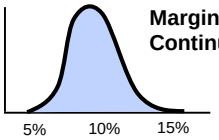


Joint Probability
Event A: 12% < φ < 14% and 600mD < k < 900mD
P(12% < φ < 14% ∩ 600mD < k < 900mD)

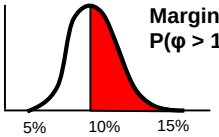


Joint Probability Density Function
Binned
(0% bins removed)

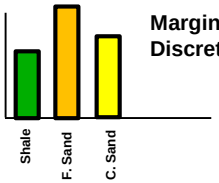
Univariate, Marginal Examples



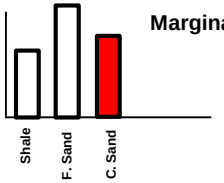
Marginal Distribution
Continuous Probability Density Function



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



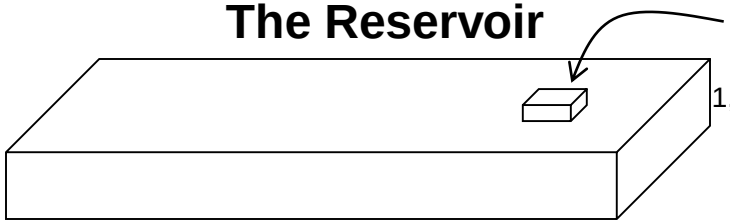
Marginal Distribution
Discrete Probability Density Function



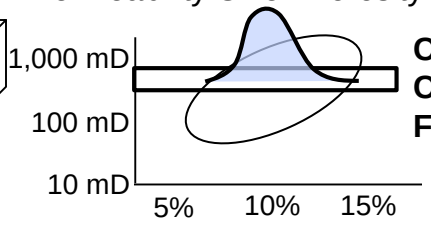
Marginal Probability, Event B: Facies = C. Sand

Marginal, Conditional and Joint Probability

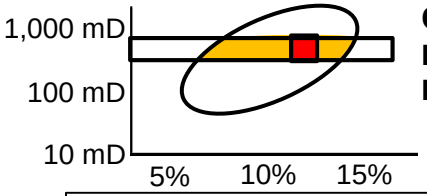
The Reservoir



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

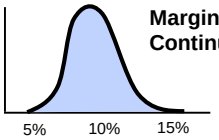


Conditional Distribution
Continuous Conditional Probability Density Function

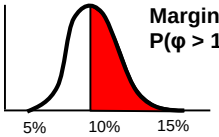


Conditional Probability
Event A: 12% < φ < 14% | 600mD < k < 900mD
P(12% < φ < 14% | 600mD < k < 900mD)

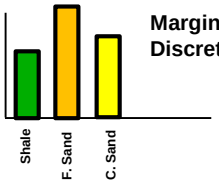
Univariate, Marginal Examples



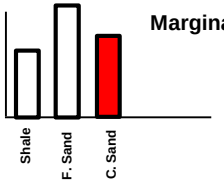
Marginal Distribution
Continuous Probability Density Function



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

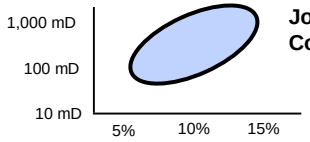


Marginal Distribution
Discrete Probability Density Function

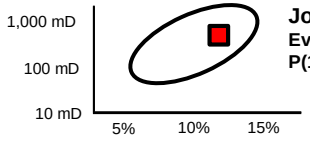


Marginal Probability, Event B: Facies = C. Sand

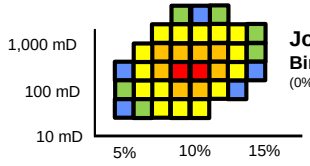
Bivariate, Joint Examples



Joint Distribution
Continuous Joint Probability Density Function



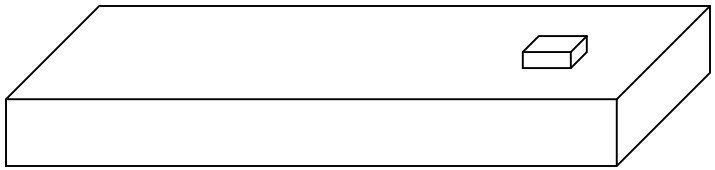
Joint Probability
Event A: 12% < φ < 14% and 600mD < k < 900mD
P(12% < φ < 14% ∩ 600mD < k < 900mD)



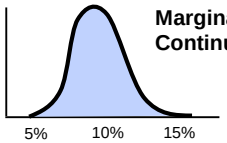
Joint Probability Density Function
Binned
(0% bins removed)

Marginal, Conditional and Joint Probability

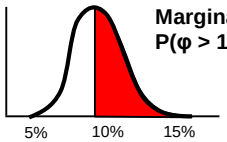
The Reservoir



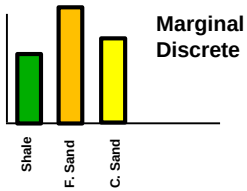
Univariate, Marginal Examples



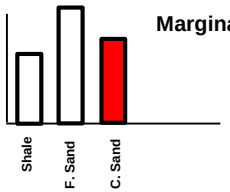
Marginal Distribution
Continuous Probability Distribution Function



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

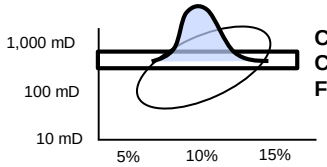


Marginal Distribution
Discrete Probability Distribution Function

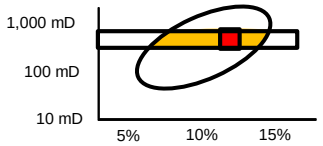


Marginal Probability, Event B: Facies = C. Sand

Bivariate, Conditional Examples

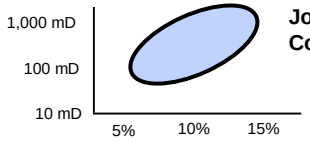


Conditional Distribution
Continuous Conditional Probability Density Function

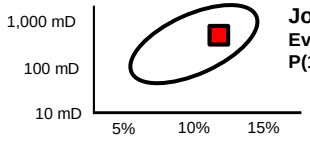


Conditional Probability
Event A: 12% < φ < 14% | 600mD < k < 900mD
P(12% < φ < 14% | 600mD < k < 900mD)

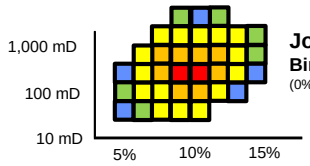
Bivariate, Joint Examples



Joint Distribution
Continuous Joint Probability Density Function

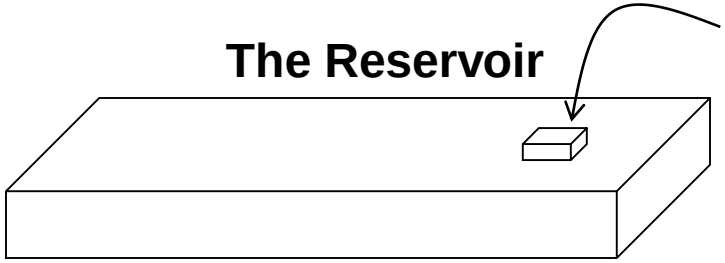


Joint Probability
Event A: 12% < φ < 14% and 600mD < k < 900mD
P(12% < φ < 14% ∩ 600mD < k < 900mD)

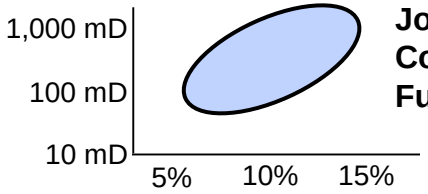


Joint Probability Density Function
Binned
(0% bins removed)

Marginal, Conditional and Joint Probability



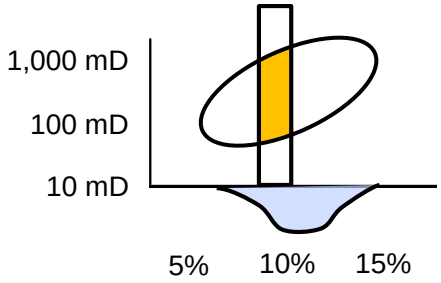
How to Calculate a Marginal Distribution from a Joint Distribution?

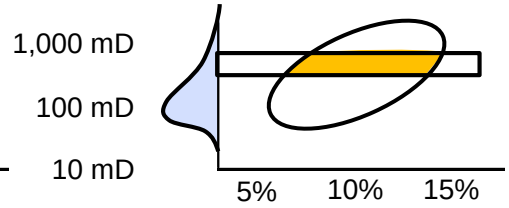


Joint Distribution
Continuous Joint Probability Density Function

Definition of a Marginal Distribution

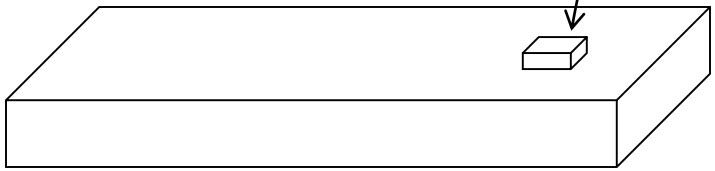
$$f_X(x) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dy \quad \textbf{or} \quad f_Y(y) = \int_{-\infty}^{+\infty} f_{XY}(x,y)dx$$



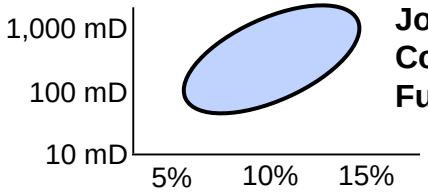


Marginal, Conditional and Joint Probability

The Reservoir



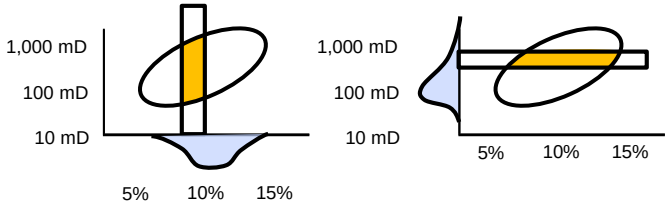
Calculate a Conditional Distribution from a Joint Distribution?



Joint Distribution
Continuous Joint Probability Density Function

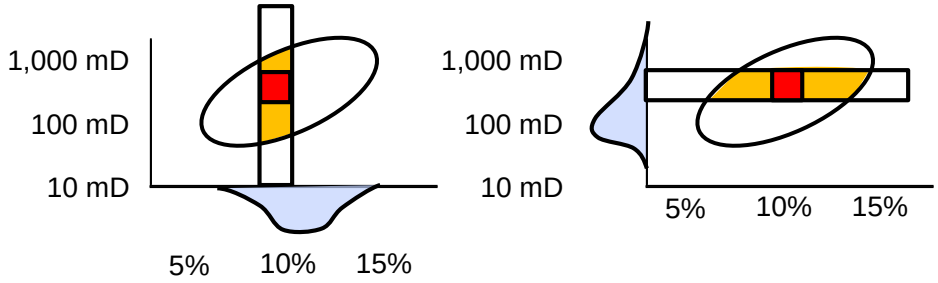
Definition of a Marginal Distribution

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

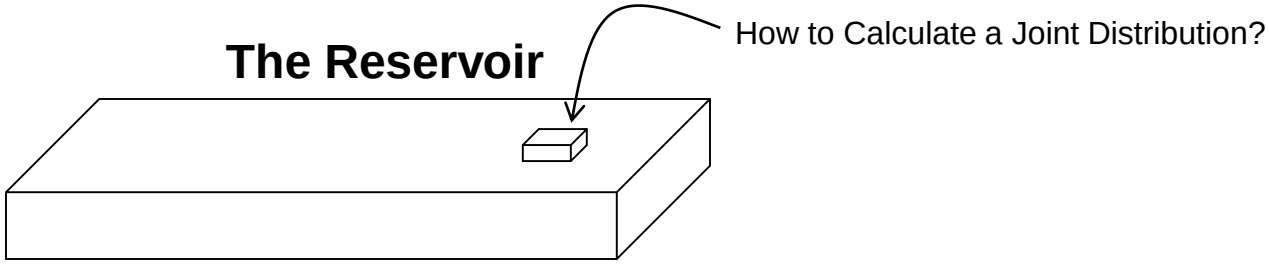


Definition of a Conditional Distribution

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

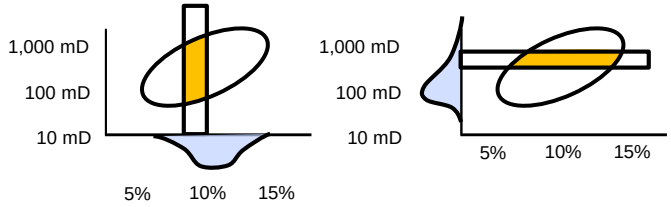


Marginal, Conditional and Joint Probability



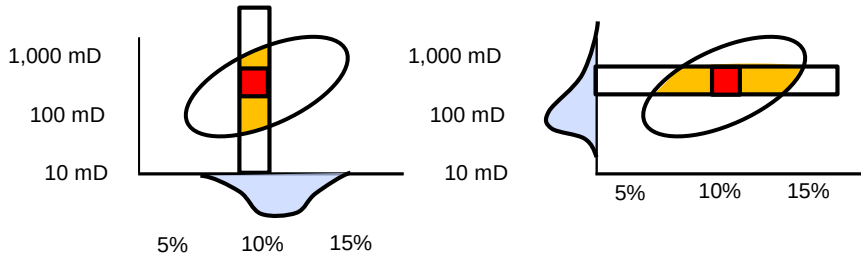
Definition of a Marginal Distribution

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



Definition of a Conditional Distribution

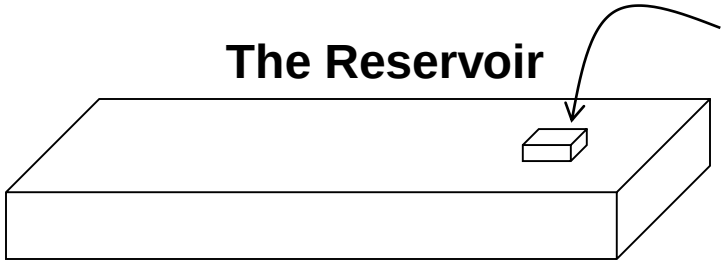
$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)} \quad \textbf{or} \quad f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)}$$



Marginal, Conditional and Joint Probability

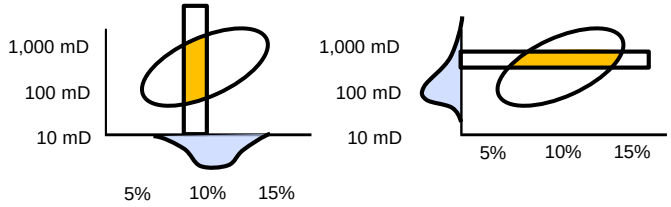
The Reservoir

How to Calculate a Joint Distribution?



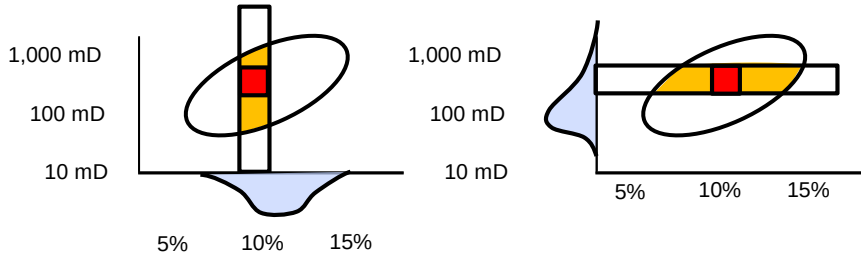
Definition of a Marginal Distribution

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

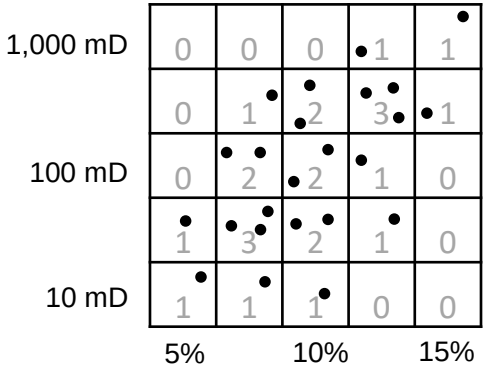


Definition of a Conditional Distribution

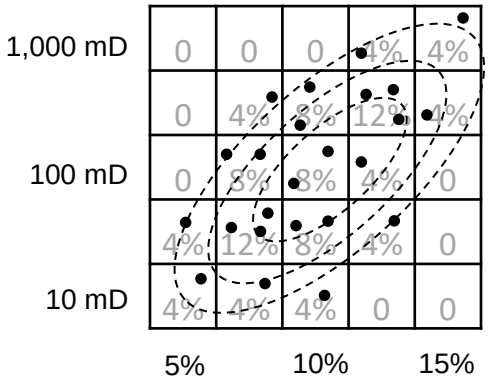
$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x,y)}{f_Y(y)} \quad \text{or} \quad f_{Y|X}(y \mid x) = \frac{f_{XY}(x,y)}{f_X(x)}$$



Non-parametric - Counting Samples in Bins

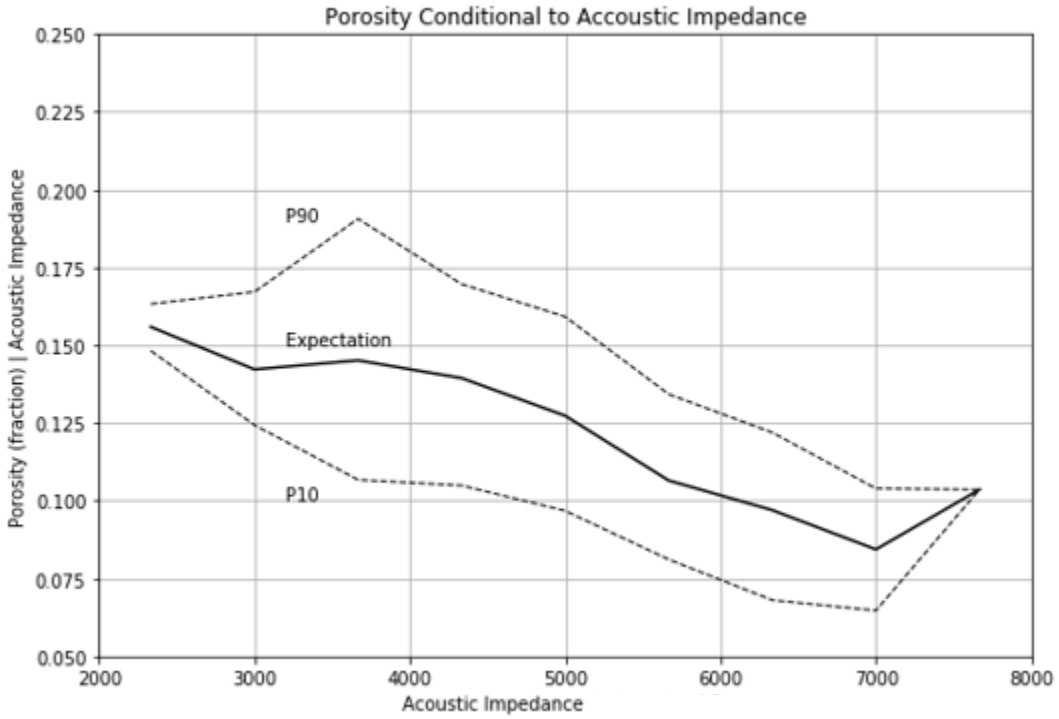


Fitting a Parametric Model



Marginal, Conditional and Joint Probability

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



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 | Google Scholar | 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, $x(\mathbf{u}_\alpha), \forall\, \alpha = \mathbf{0}, \mathbf{1}, \dots, \mathbf{n-1}$.

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

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

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

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.

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

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 analsis

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 . . .

- **Bivariate Analysis**
- **Covariance and Correlation**
- **Marginal, Conditional and Joint**