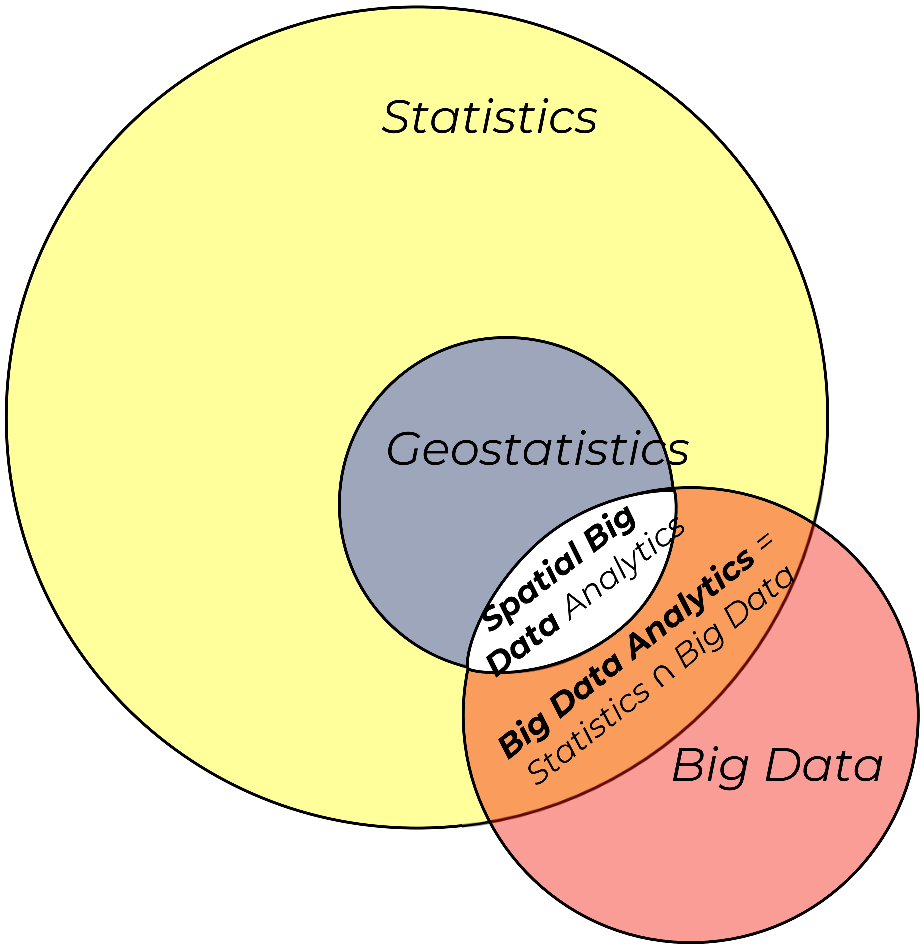
## Lecture Outline

- Basic Data Analytics
- Multivariate Analysis
- Statistical Expectation

# MOTIVATION

- Data Analytics is:
  - the use of statistics and visualization to learn from data

- We will start with some basic statistics and build up




Proposed Venn diagram for spatial big data analytics

# BASIC DATA ANALYTICS

# TYPES OF SUBSURFACE DATA

| Type | Resolution | Coverage | Information Type |
|---|---|---|---|
| *Core* | $\simeq \infty$ | In Well Bore | Lithology, pore and sedimentary structures |
| *Well Log* | 10 cm | Near Bore | Facies, porosity, minerology |
| *Image Log* | 5 mm | Near Bore | Sedimentary structures, faults |
| *Seismic* | 10 m | Exhaustive | Framework, trends, facies, porosity |
| *Production* | 10–100 m | Drainage Radius | Volumes, connectivity, permeability |
| **Analog** | | | |
| *Mature Fields* | 10–100 m | ≤ Complete | Validation, prior for all |
| *Outcrop*<br>*Geomorphology* | $\simeq \infty$<br>$\simeq \infty$ | none<br>none | Concepts, input statistics<br>Concepts |
| *Shallow Seismic* | ≥ Element | none | Concepts, input statistics |
| *Experimental Stratigraphy* | $\simeq \infty$ | none | Concepts |
| *Numerical Process* | ≥ Complex | none | Concepts |

All subsurface data sources have unique scale, and precision.
We most account for this in our models.

# SAMPLING AND STATISTICS

- Variable: any property measured / observed in a study
  - e.g., porosity, permeability, mineral concentrations, saturations, contaminant concentration
  - in data mining / machine learning this is known as a feature
- Population: Exhaustive, finite list of property of interest over area of interest. Generally, the entire population is not accessible.
  - e.g., porosity at each location within a reservoir
- Sample: The set of data that have actually been measured
  - e.g., porosity data from measured by well-logs within a reservoir
- Parameters: summary measure of a population
  - e.g., population mean, population standard deviation, we rarely have access to this
- Statistics: summary measure of a sample
  - e.g., sample mean, sample standard deviation, we use statistics as estimates of the parameters

# PROBABILITY: A FREQUENTIST DEFINITION

- Measure of the likelihood that an event will occur. For random experiments and well-defined settings (such as coin tosses):

$$\mathrm{Prob}(A) = \mathrm{P}(A) = \lim_{n\to\infty}\left(\frac{n(A)}{n(\Omega)}\right)$$

**frequentist approach** to probability is the limit of relative frequency over a large number of trials

- where:
  - n(A) = number of times event A occurred
  - n(Ω) = number of trails
- Exploration probability of success:
  - 25 well drilled last year
  - 5 exploration successes
  - probability of success is 20%

# PROBABILITY: A BAYESIAN DEFINITION

- Measure of the likelihood that an event will occur. For any occurrence:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

**Bayesian approach** probability is interpreted as reasonable likelihood representing a state of knowledge or as quantification of a personal belief. Specify a prior and update with new information.

where:

$P(A)$ = prior

$P(B)$ = evidence

$P(B \mid A)$ = likelihood

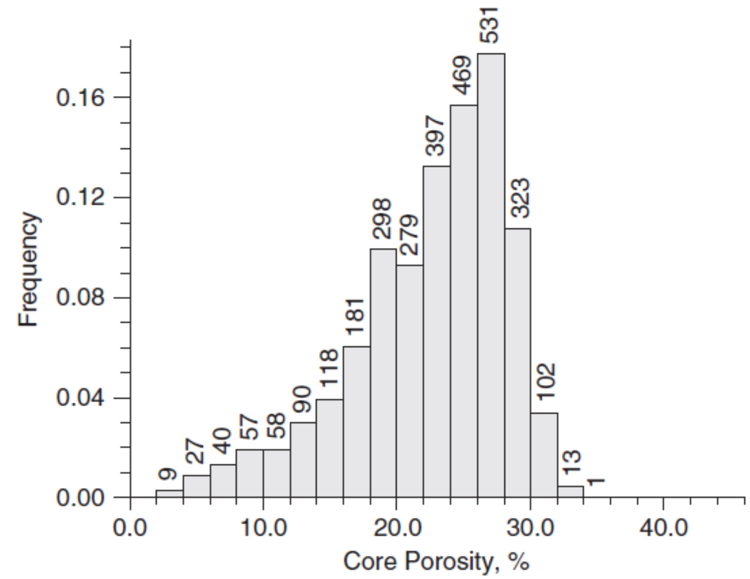
$P(A \mid B)$ = posteriori

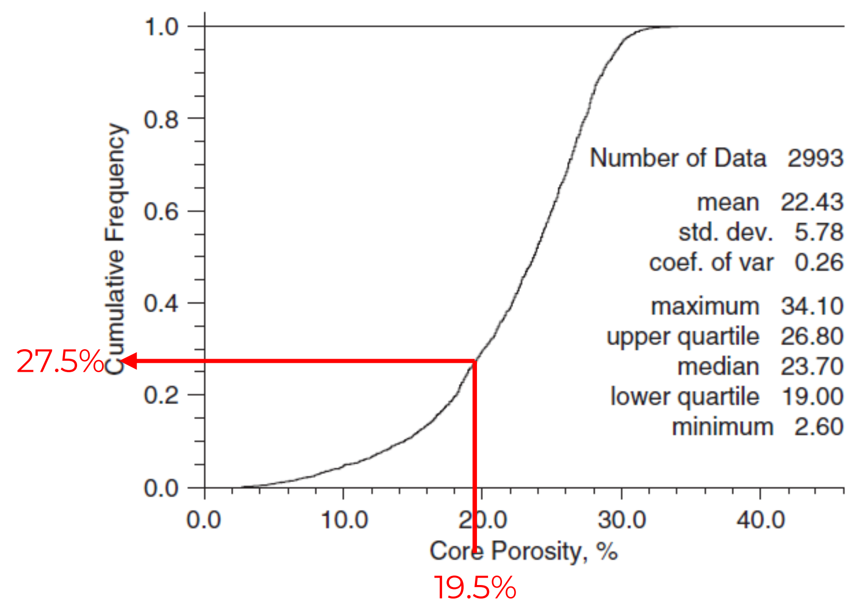
Example: Given a prior probability of 20% based on last years, a likelihood of 10% from recent drilling and an evidence term of 20%, the updated posteriori exploration success rate is 10%.

# STATISTICAL UNIVARIATE, MARGINAL DISTRIBUTIONS

- **Univariate** indicates pertaining to one variable.
- **Histogram** is the frequency for an exhaustive, mutually exclusive set of bins.
- **Probability density function (PDF)**, f(x) is a continuous function indicating the relative likelihood of a specific value x.
- **Cumulative distribution function (CDF),** $F_X(x)$ is the probability that the variable takes a value less than or equal x.

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



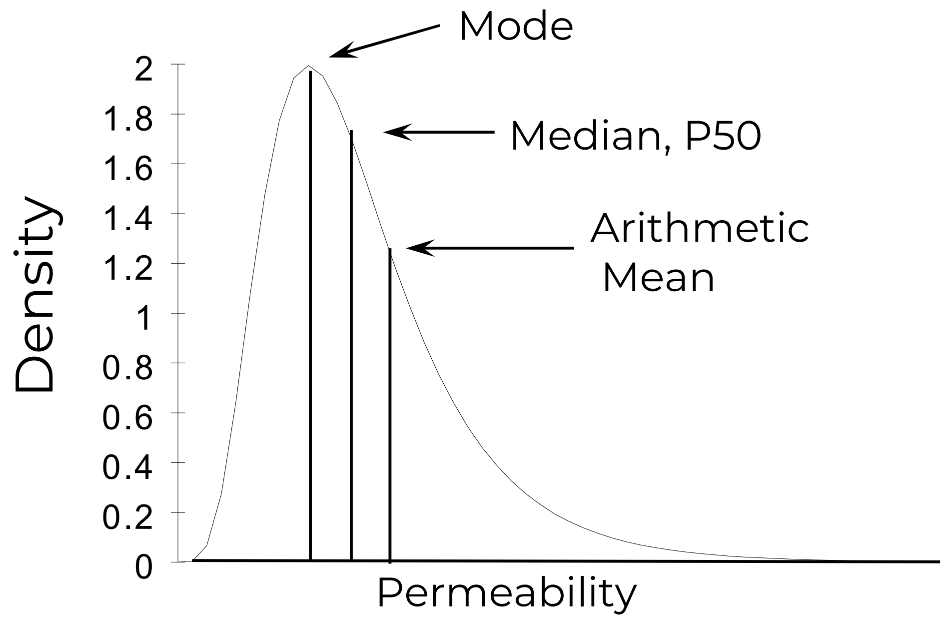



# STATISTICAL SUMMARIES

- Measures of Central Tendency
  - Sample Arithmetic Mean

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

  - Median, 50th Percentile
  - Mode, most common value



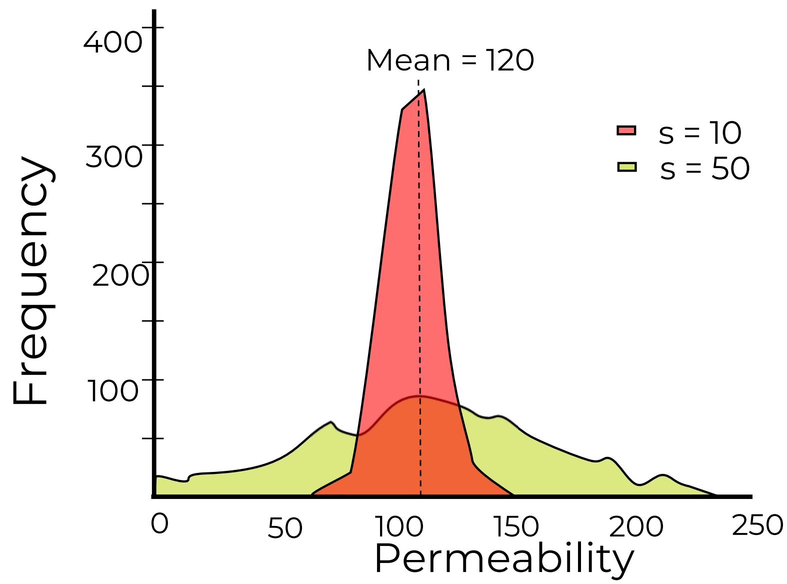

- Measures of Dispersion
  - Sample variance

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

  - Standard deviation

$$s = \sqrt{s^2}$$

  - Range, max – min
  - Interquartile range, P75 – P25



# MULTIVARIATE ANALYSIS

# MOTIVATION FOR MULTIVARIATE METHODS

- We typically need to build reservoir models of more than one property of interest
  - Expanded by whole earth modeling, closing loops with forward models
  - Expanded by unconventionals
- Subsurface properties may include:
  - Rock Classification: lithology, architectural elements, facies, depofacies
  - Petrophysical: porosity, directional permeability, saturations
  - Geophysical: density, p-wave and s-wave velocity
  - Geomechanical: Compressibility / Poisson's ratio, Yong's modulus, brittleness, stress field
  - Paleo- / Time Control: fossil abundances, 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

More about this later…

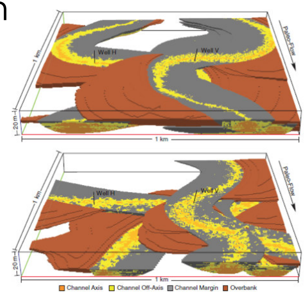
# MOTIVATION FOR MULTIVARIATE METHODS

- 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

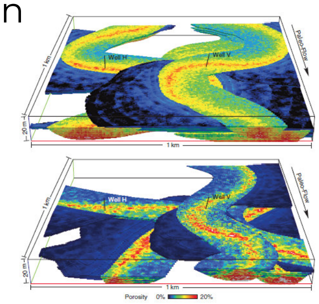
Multiple Facies Realizations



Continuous Simulation

Models Porosity PDF, Trends and Spatial Continuity within each facies

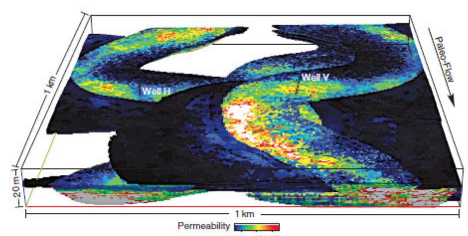
Multiple Porosity By-facies Realizations



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

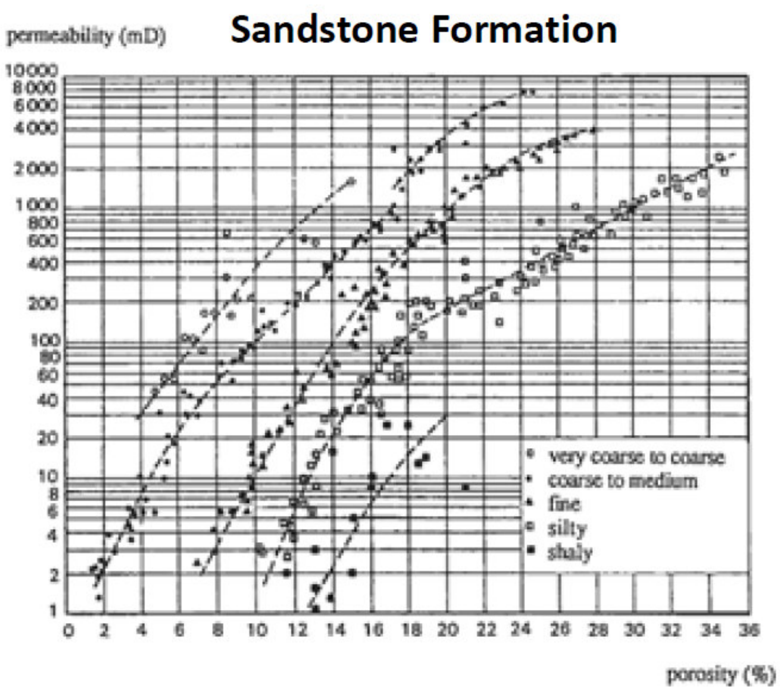
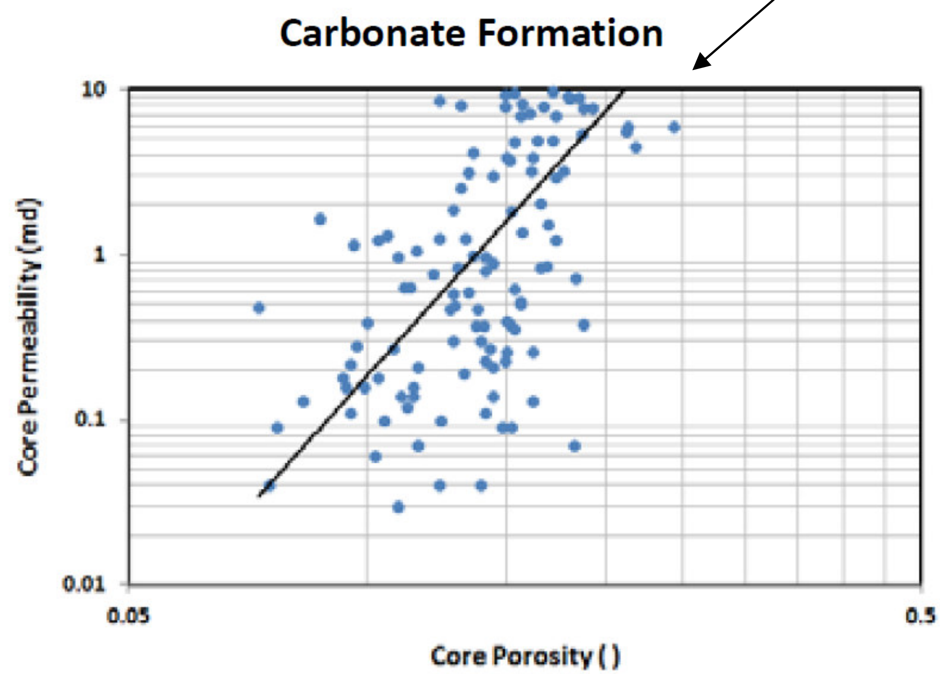



Figure from Peters, E. J., 2012, Advanced Petrophysics.

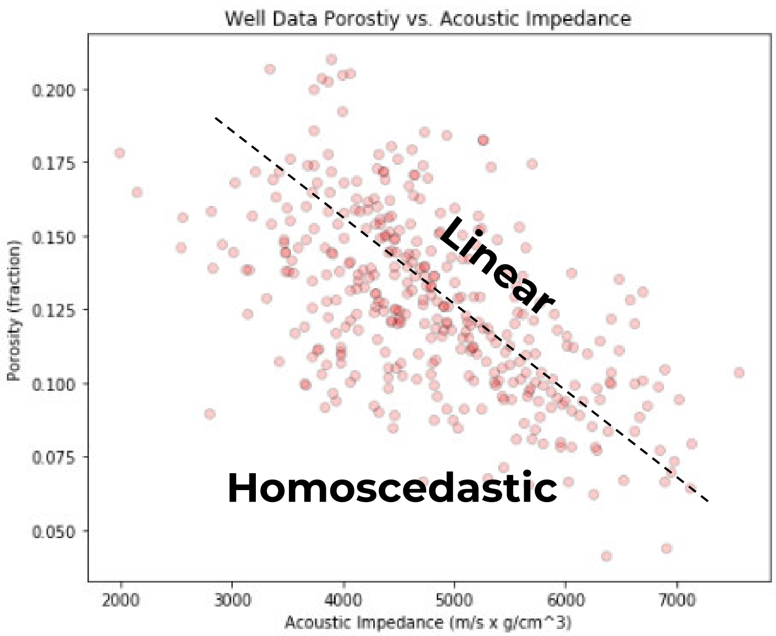


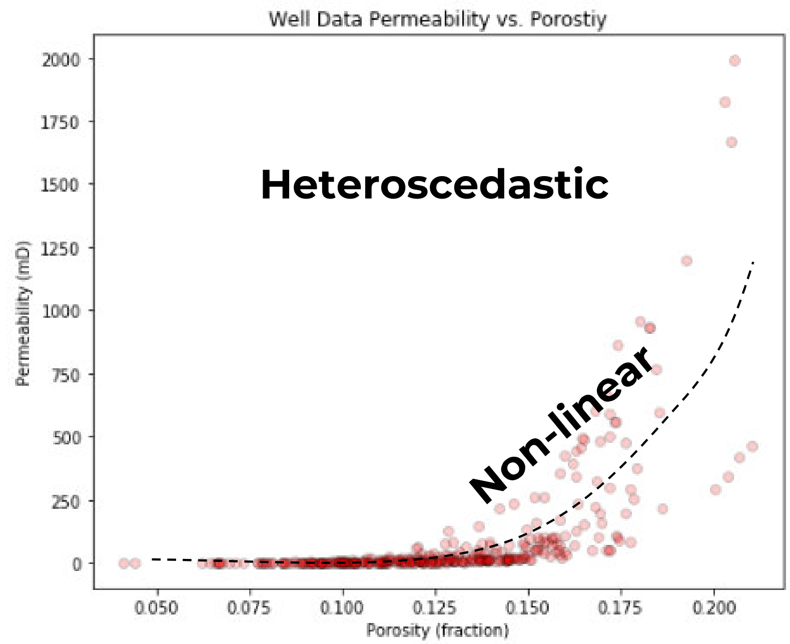

Slide from Dr. Zoya Heidari's PGE 337 Course

# 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

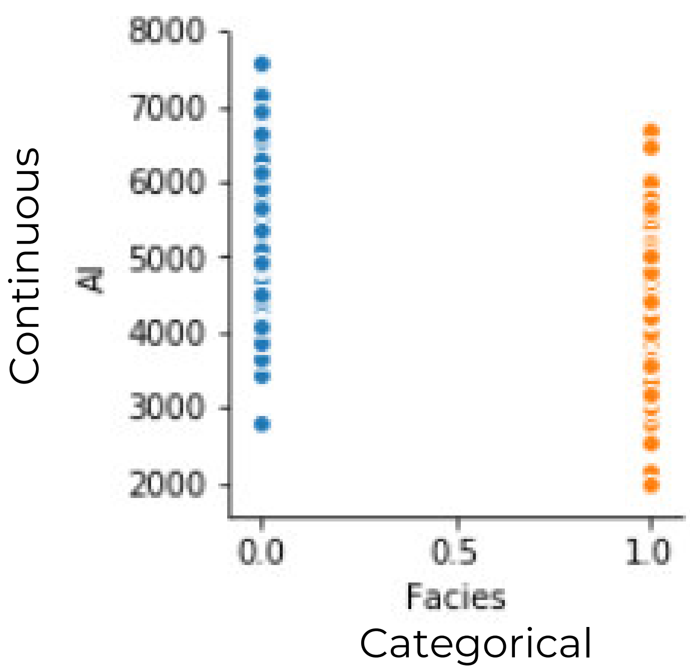


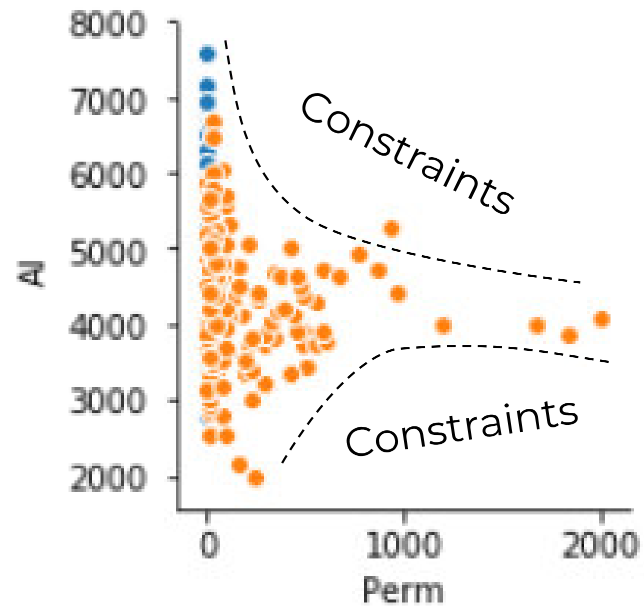



# BIVARIATE STATISTICS

## 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

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

Correlation coefficient of variables x and y

number of data pairs

standard deviation of variables x and y

- Correlation coefficient is a standardized covariance

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

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

    – 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

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

Rank correlation coefficient of variables x and y

number of data pairs

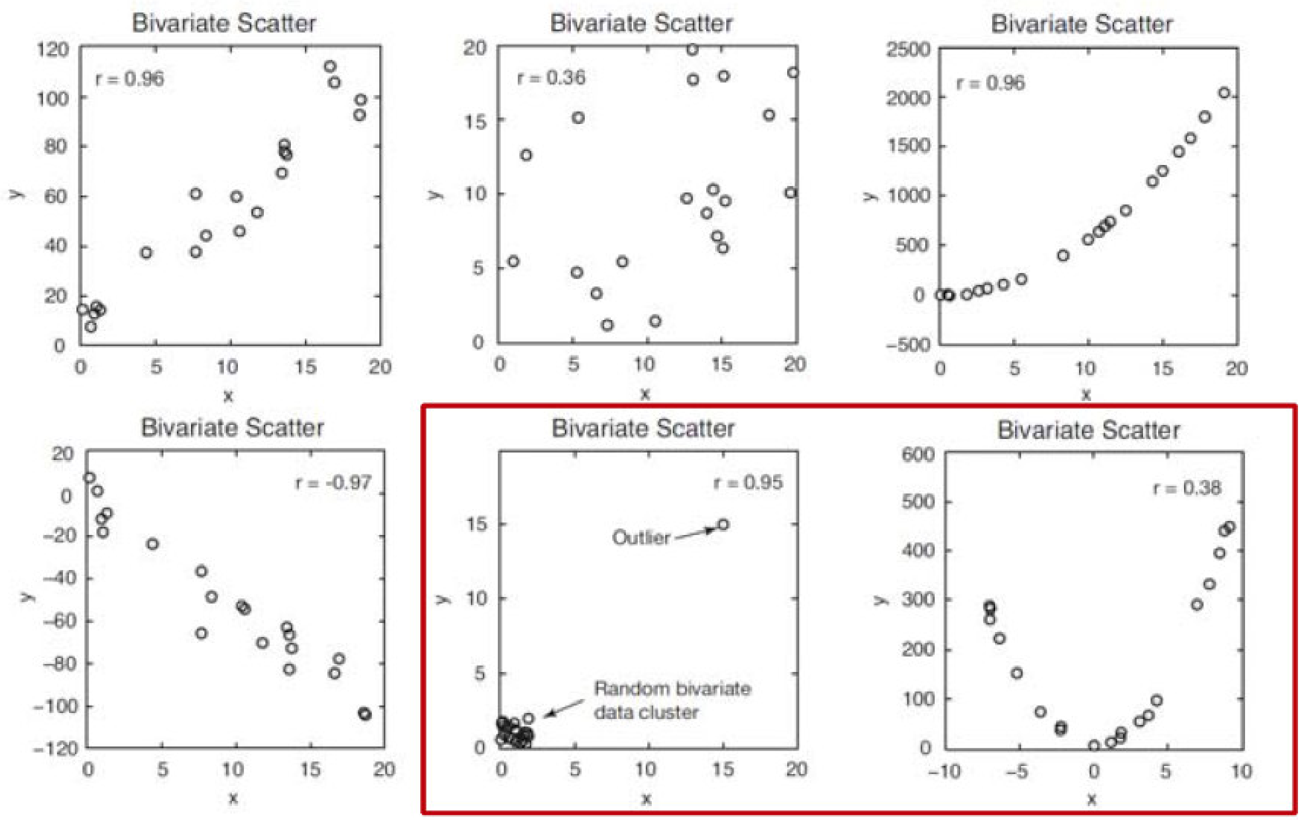
standard deviation of Rank transform of variables x and y

  - Rank transform, e.g. $R_{x_i}$, sort the data in ascending order and replace the data with the index, $i$=1,...,$n$
  - 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?

Slide from Dr. Zoya Heidari's 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!

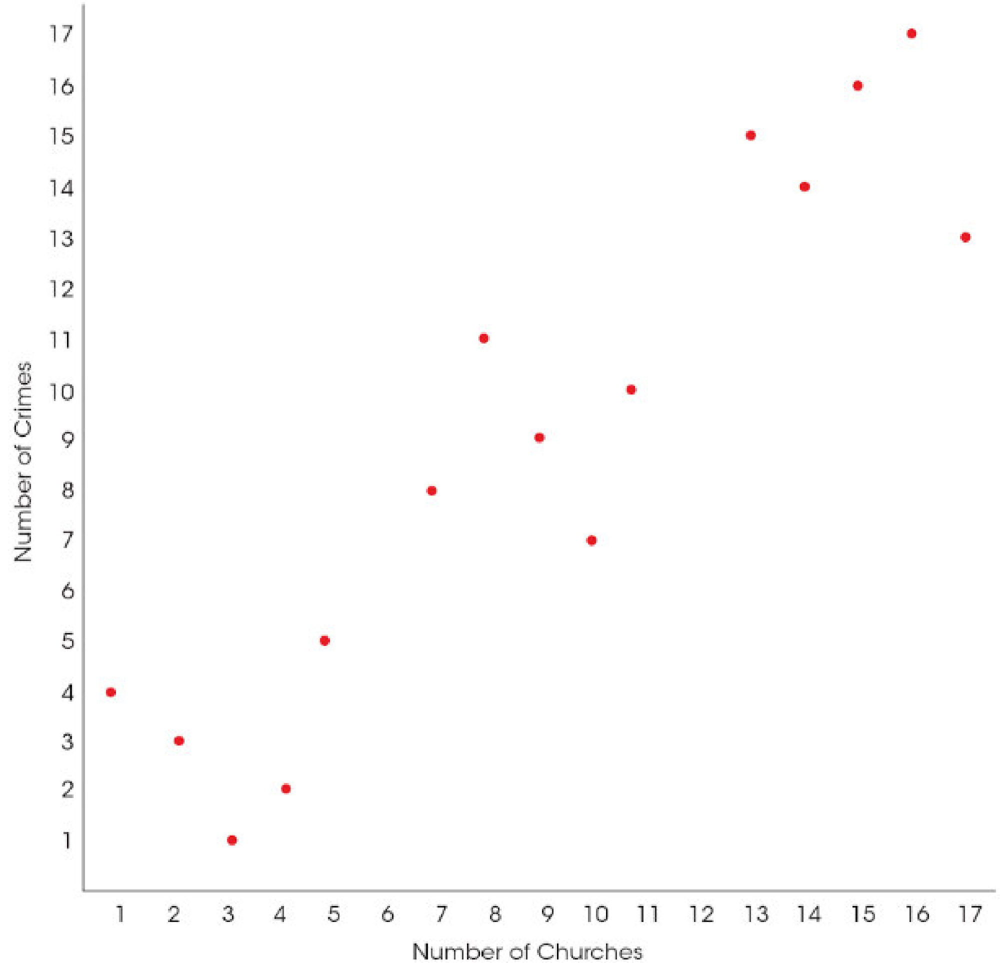



**FIGURE 15.10**
Hypothetical data showing the logical relationship between the number of churches and the number of crimes for three groups of cities: those with small populations ($Z = 1$), those with medium populations ($Z = 2$), and those with large populations ($Z = 3$).

# BIVARIATE STATISTICS

## Correlation and Causation

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

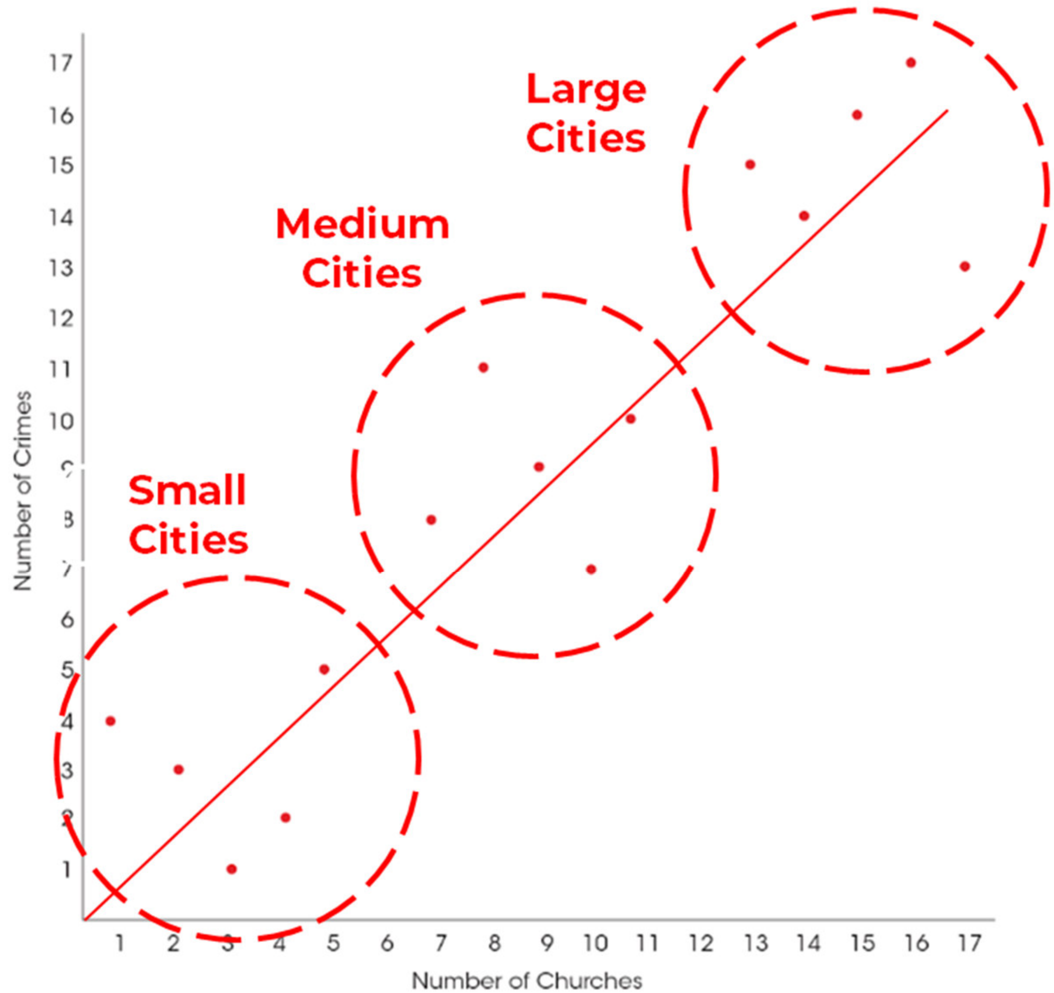



**FIGURE 15.10**
Hypothetical data showing the logical relationship between the number of churches and the number of crimes for three groups of cities: those with small populations ($Z = 1$), those with medium populations ($Z = 2$), and those with large populations ($Z = 3$).

# BIVARIATE STATISTICS

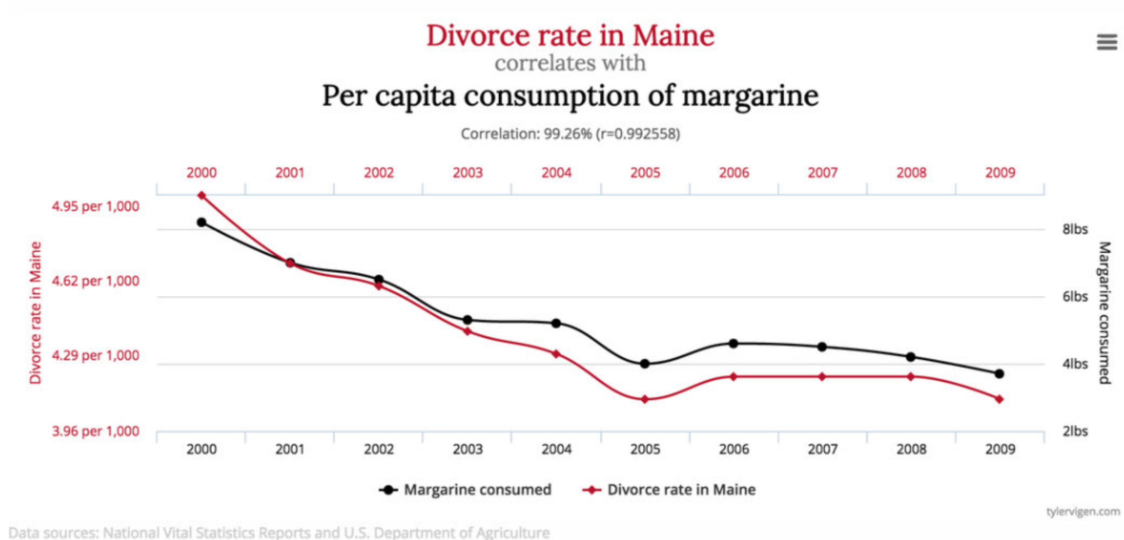
## Comical Examples of Correlation and Causation

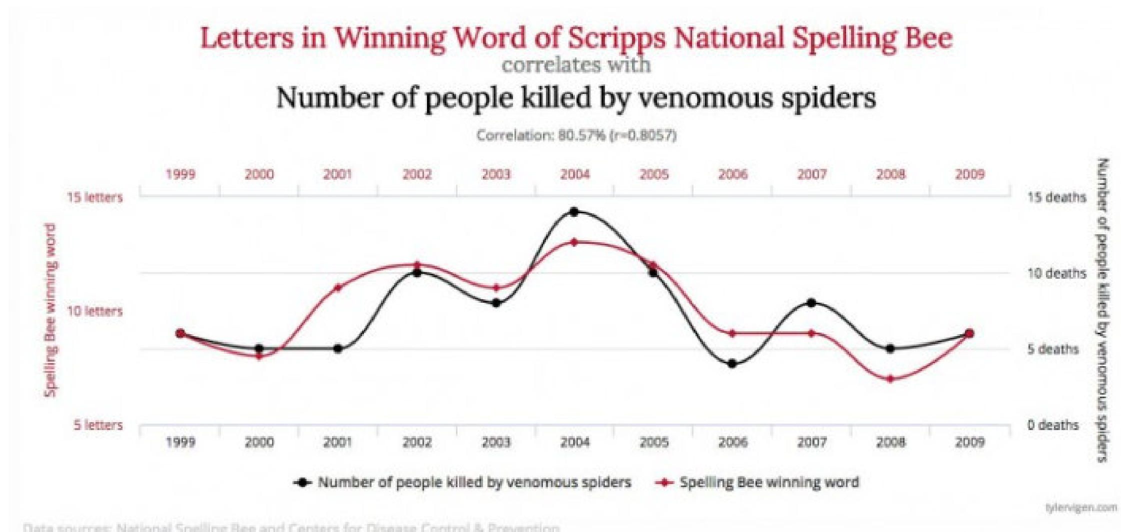
**Margarine** causes **divorce**?

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?**



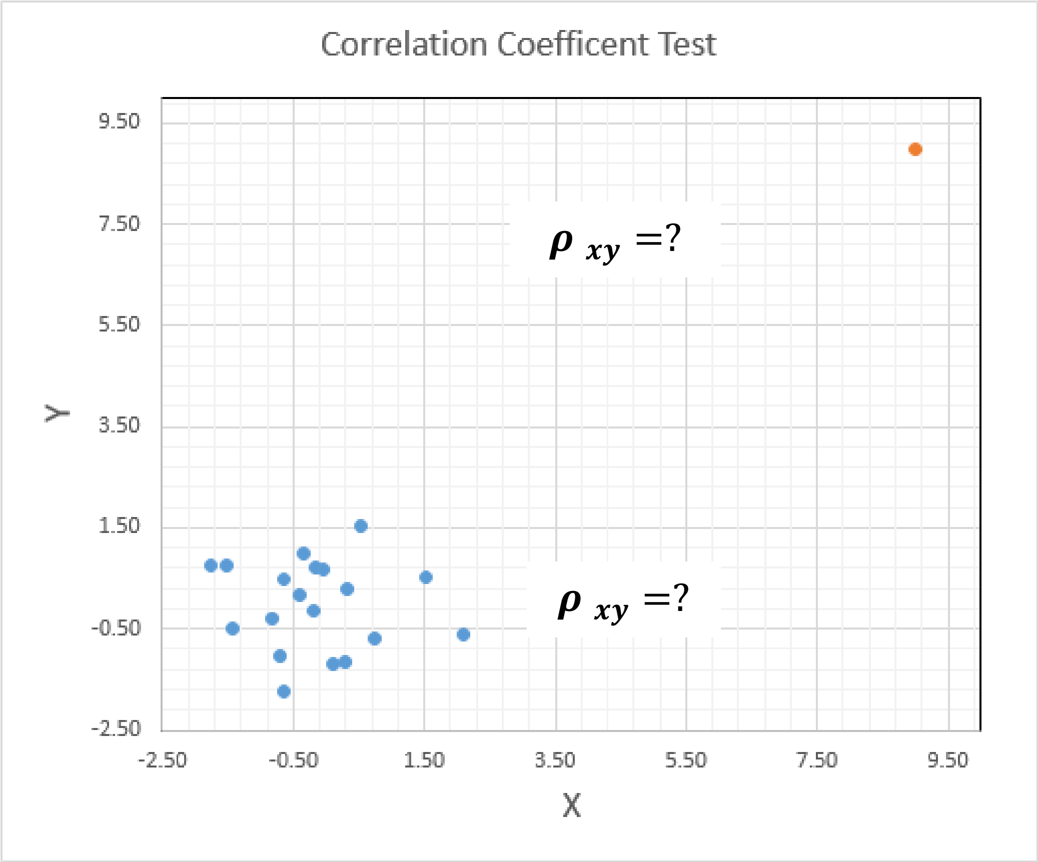
# BIVARIATE STATISTICS

## Demonstration with Pearson's Correlation Coefficient

1. Generate a random data set of x and y variables and estimate their correlation coefficient

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

**How does this outlier affect the correlation coefficient?**



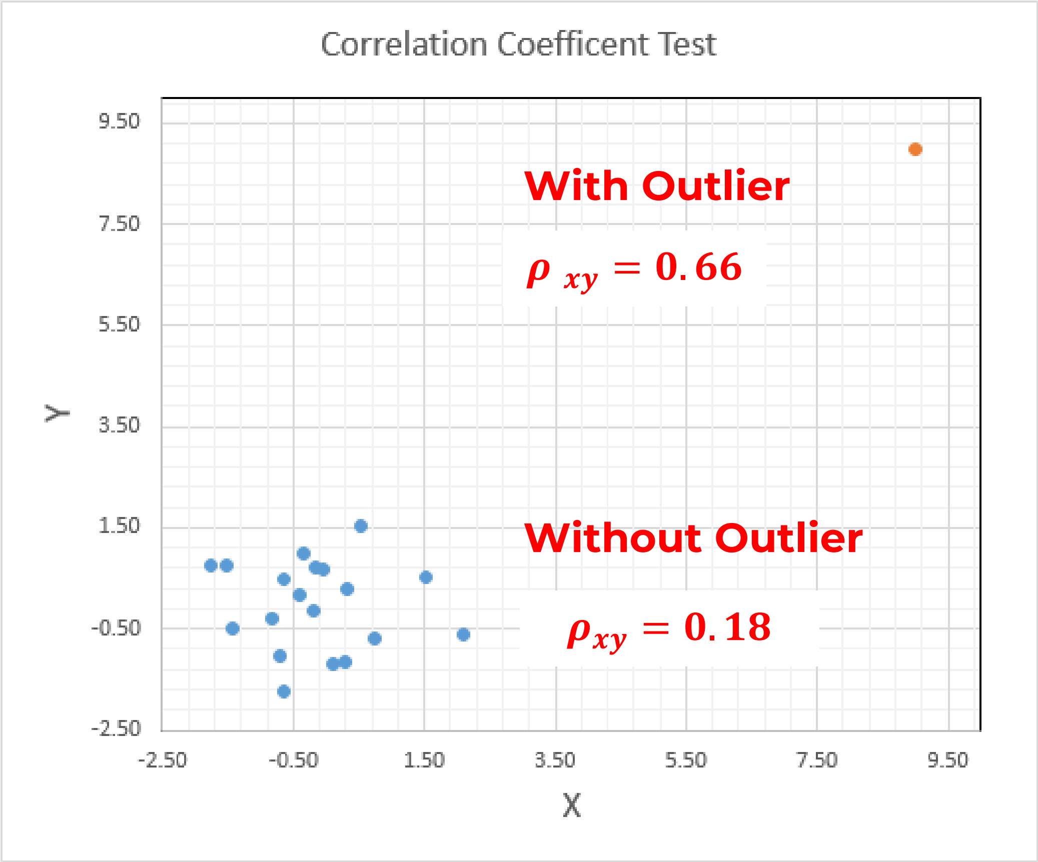
# BIVARIATE STATISTICS

## Demonstration with Pearson's Correlation Coefficient

1. Generate a random data set of x and y variables and estimate their correlation coefficient

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

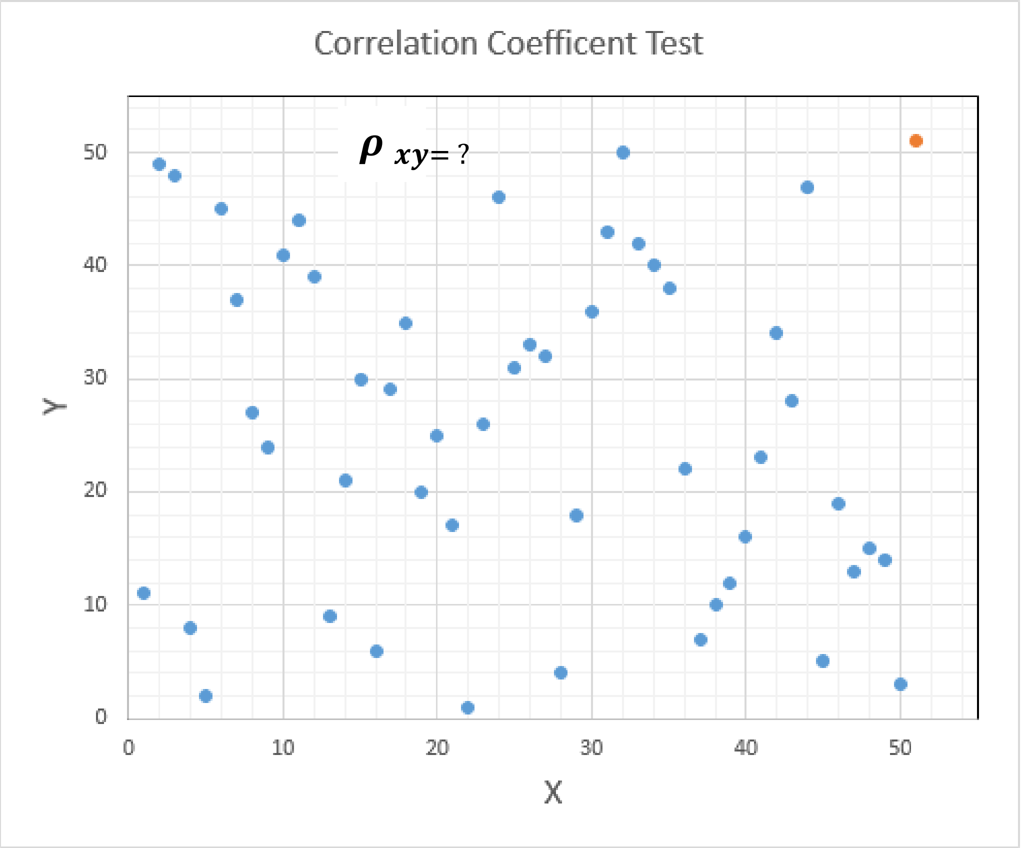
**How does this outlier affect the correlation coefficient?**



# BIVARIATE STATISTICS

## Demonstration with Pearson's Correlation Coefficient

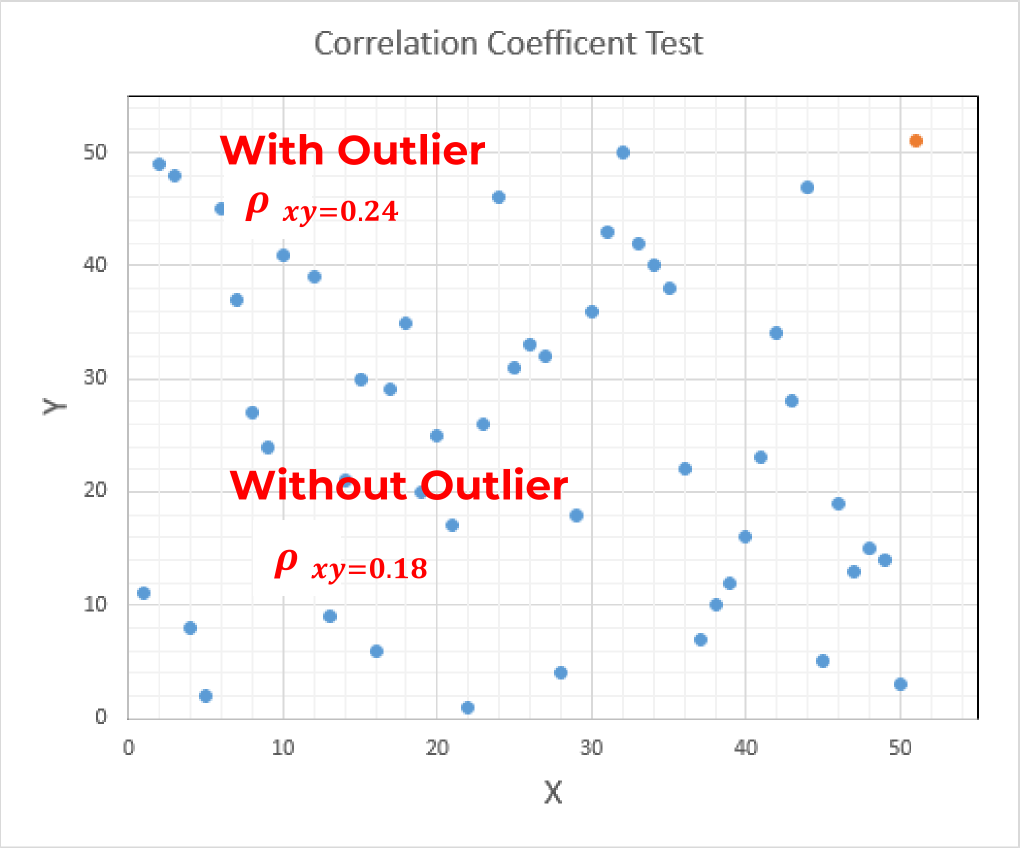
- **Now,** Apply the rank transform to the dataset
- How does this outlier now affect the correlation coefficient?
- This is a more robust form of the correlation coefficient called the rank correlation coefficient



# BIVARIATE STATISTICS

## Demonstration with Pearson's Correlation Coefficient

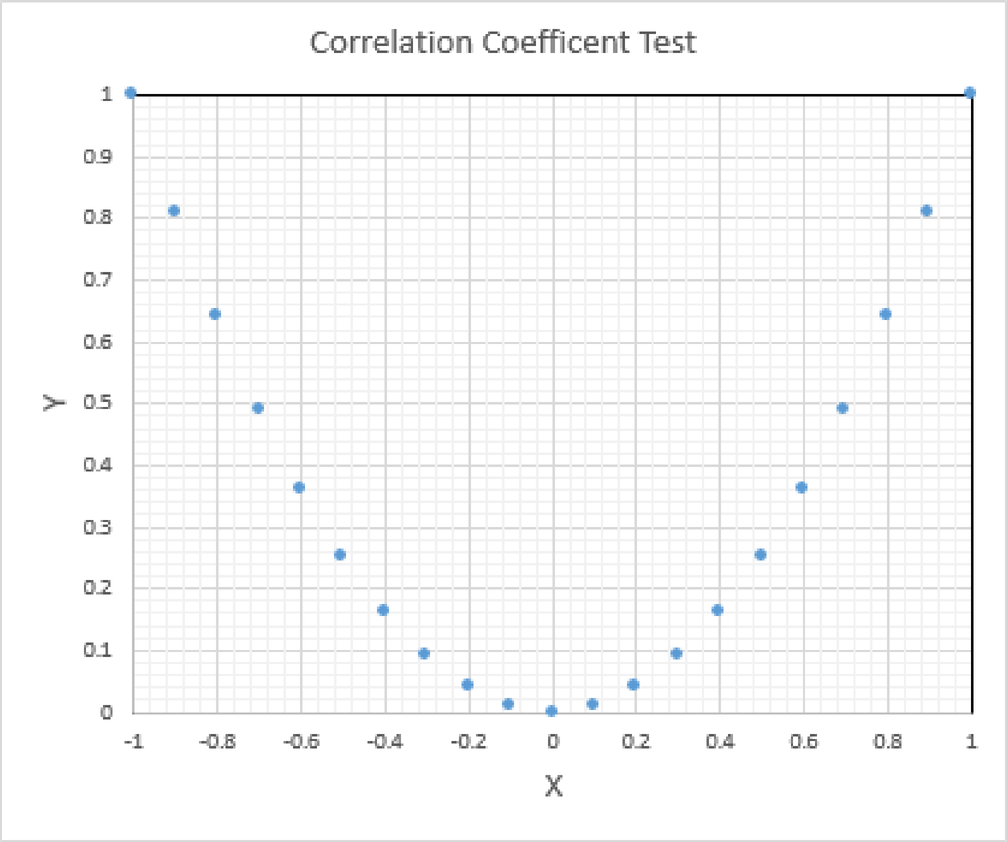
- **Now,** Apply the rank transform to the dataset
- How does this outlier now affect the correlation coefficient?
- This is a more robust form of the correlation coefficient called the rank correlation coefficient



# BIVARIATE STATISTICS

## Measuring Linear Relationships with the Correlation Coefficient

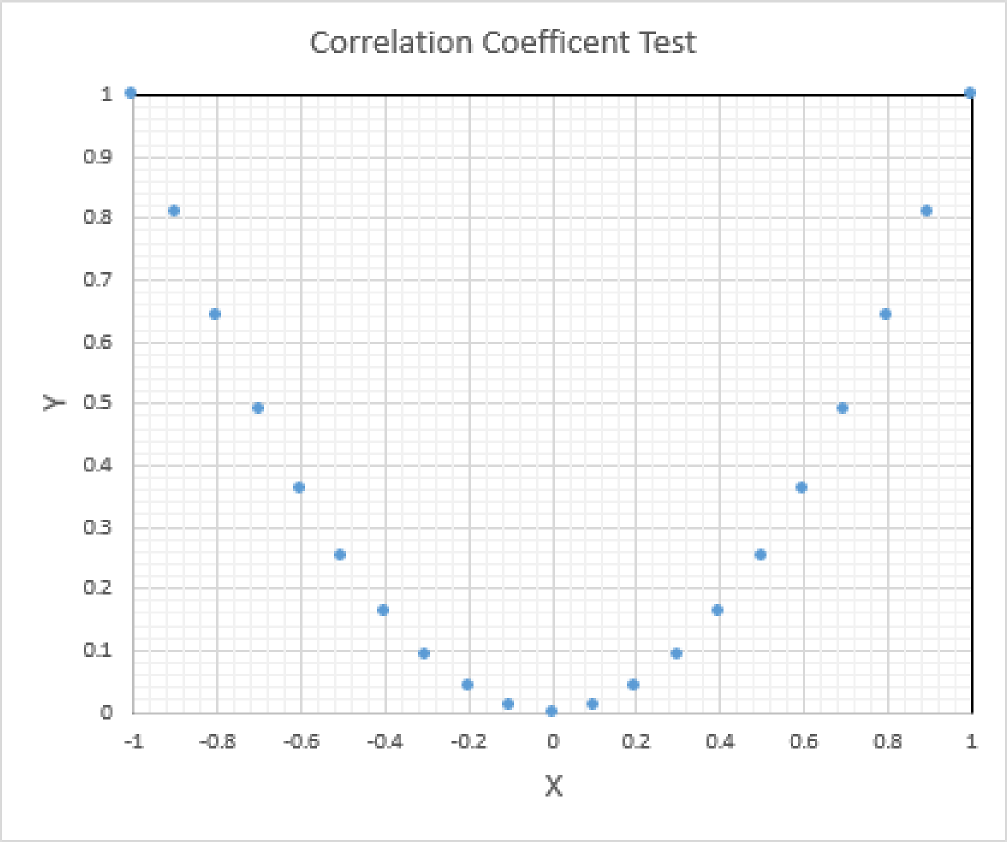
- Correlation / Covariance is a measure of linear relationship
- What is the Correlation / Covariance of
  - y = x^2 over range of [-1, 1]?



# 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, $\rho_{xy}$ = 0.0!
  - Over range [0,1]?
    - Correlation Coefficient, $\rho_{xy}$ = 0.96,
    - Rank Correlation Coefficient, $\rho_{RxRy}$ = 1.0



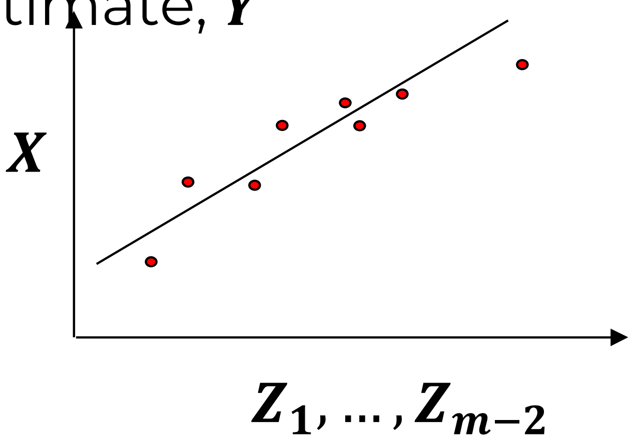
# BIVARIATE STATISTICS

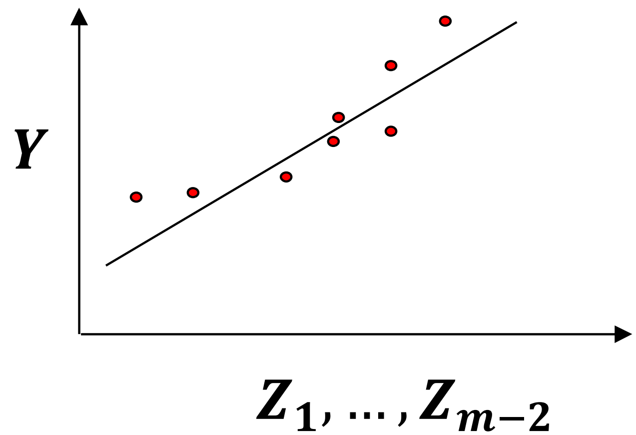
## Partial Correlation

- 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, $\boldsymbol{Y}^*$



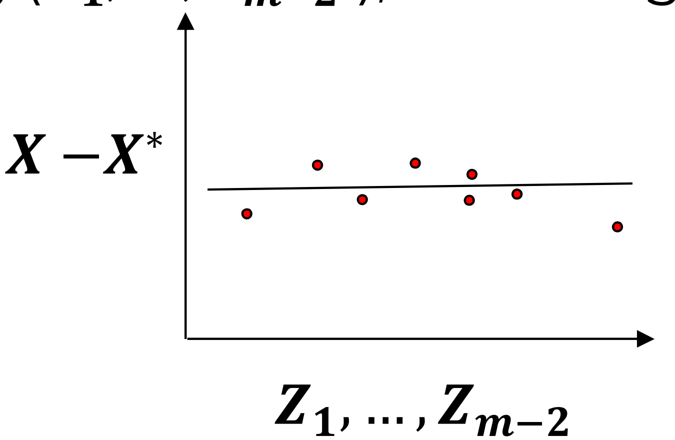
# BIVARIATE STATISTICS

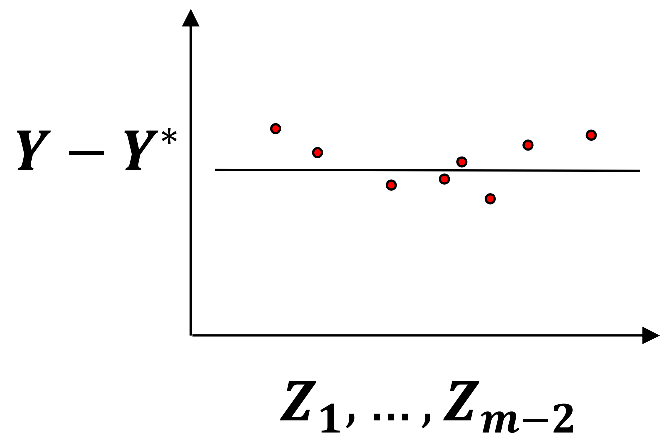
## Partial Correlation

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

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

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



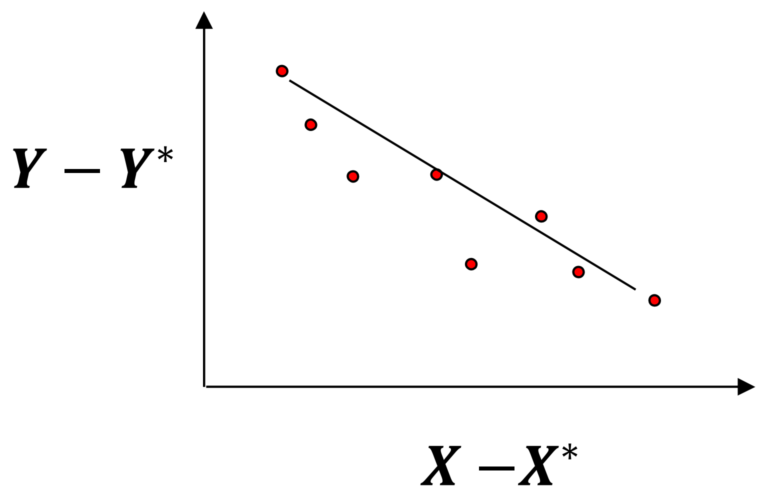



# BIVARIATE STATISTICS

## Partial Correlation

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

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

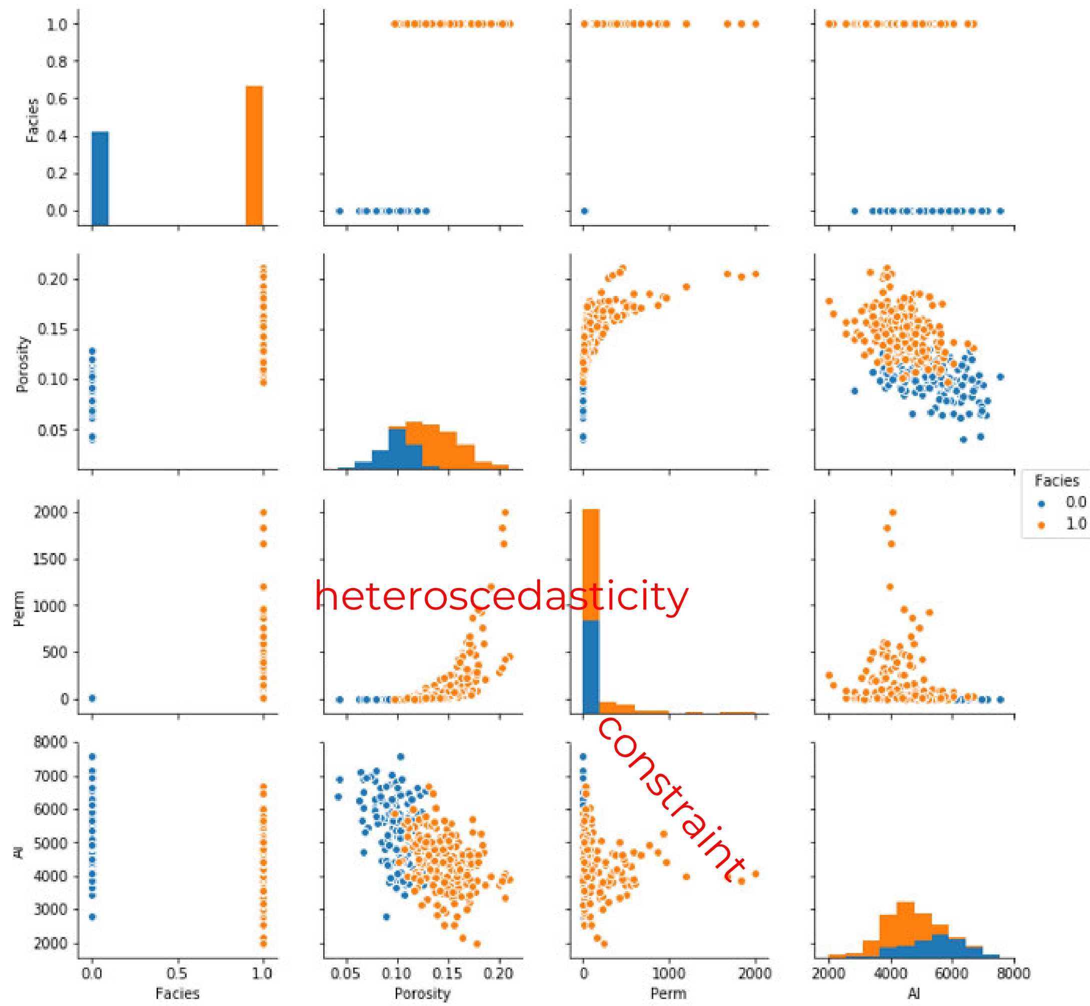



- The partial correlation, provides a measure of the linear relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ while controlling for the effect of $\boldsymbol{Z_1, \dots, Z_{m-2}}$ other features on both, $\boldsymbol{X}$ and $\boldsymbol{Y}$

# 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



# BIVARIATE STATISTICS HANDS-ON

- Walk-through of bivariate statistics.

Daytum +2 Course: Data Analytics, Geostatistics and Machine Learning Deep Dive

Multivariate Analysis Demonstration and Exercise

### Goal

Calculate multivariate analytics for a spatial dataset.

### Description

Here's a simple, documented workflow, demonstration of multivariate analysis for subsurface modeling workflows. This should help you get started with building subsurface models that integrate spatial continuity.

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

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

File is multivariate.ipynb.

# STATISTICAL EXPECTATION

# STATISTICAL EXPECTATION

Statistical expectation is a probability weighted average of a continuous distribution.

- For discrete (binned) continuous random variables (RVs), normalized histogram:

$$\mathrm{E}[X] = \sum_{i=1}^{n} x_i f_x(x_i) = \sum_{i=1}^{n} x_i \, p_i$$

discrete value

probability

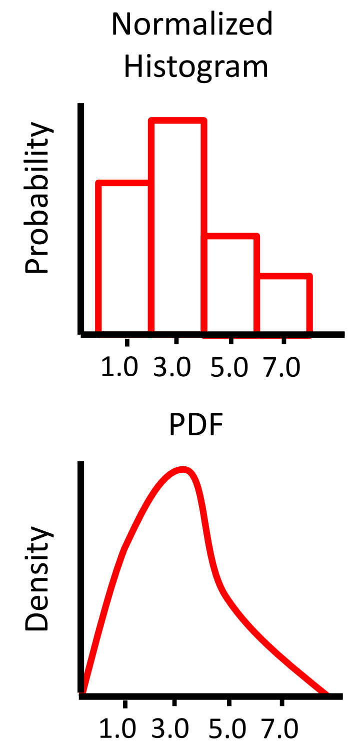
$$\sum_{i=1}^{n} f(x_i) = \sum_{i=1}^{n} p_i = 1 \text{, closure}$$

- For continuous RVs, PDF:

value pdf

$$\mathrm{E}[X] = \int_{-\infty}^{+\infty} x f_x(x) dx$$

- Recall: $f(x)$ is the probability density function of feature $X$, and



Schematic of continuous normalized histogram and PDF.

# STATISTICAL EXPECTATION

Statistical expectation vs. arithmetic average?

- Expected value for a random variable is the long-run (assuming enough samples) average.

- Given the samples and assuming that they are randomly sampled:
  - equiprobable
  - unbiased
  - large enough sample set

For example:

$$Porosity, x_{i=1,\ldots,10} = \{10\%, 14\%, 20\%, 16\%, 5\%, 10\%, 12\%, 22\%, 12\%, 2\%\}$$

$$\underbrace{\text{if } p_i = \frac{1}{n}}_{\text{equal probability for all data}}, \forall \text{ i} = 1, \ldots, \text{n then E[X]} = \sum_{i=1}^{n} x_i \, p_i = \sum_{i=1}^{n} x_i \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}$$

$p_i = \frac{1}{n}$

$$\text{E[X]} = \bar{x} = \frac{113\%}{10} = 11.3\%$$

# STATISTICAL EXPECTATION HANDS ON

- Discrete Continuous Example:
  - The following binned grain sizes (mm) outcomes with probability in brackets

    10 (0.1), 20 (0.5), 30 (0.1), 40 (0.2), 50 (0.1)

- Problem:  Calculate the expected grain size.

# STATISTICAL EXPECTATION HANDS ON

- Discrete Continuous Example:
  - The following binned grain sizes (mm) outcomes with probability in brackets

    10 (0.1), 20 (0.5), 30 (0.1), 40 (0.2), 50 (0.1)

- Problem: Calculate the expected grain size.

$$E[X] = \sum_{i=1}^{N} p_i x_i = 10(0.1) + 20(0.5) + 30(0.1) + 40(0.2) + 50(0.1)$$

$$E[X] = 27 \text{ mm}$$

# STATISTICAL EXPECTATION HANDS ON

Expectation of a constant:

$$E[c] = c$$

Expectation of a random variable + a constant:

Distributive Property

$$E[X + c] = E[X] + E[c] = E[X] + c$$

Expectation of a product of a random variable and a constant:

$$E[cX] = cE[X]$$

By both of these, statistical expectation is a "linear operator"

Expectation of addition of two random variables:

$$E[X + Y] = E[X] + E[Y]$$

Expectation of the product of two random variables (if independent):

$$E[XY] = E[X]E[Y], if\ X \perp\!\!\!\perp Y$$

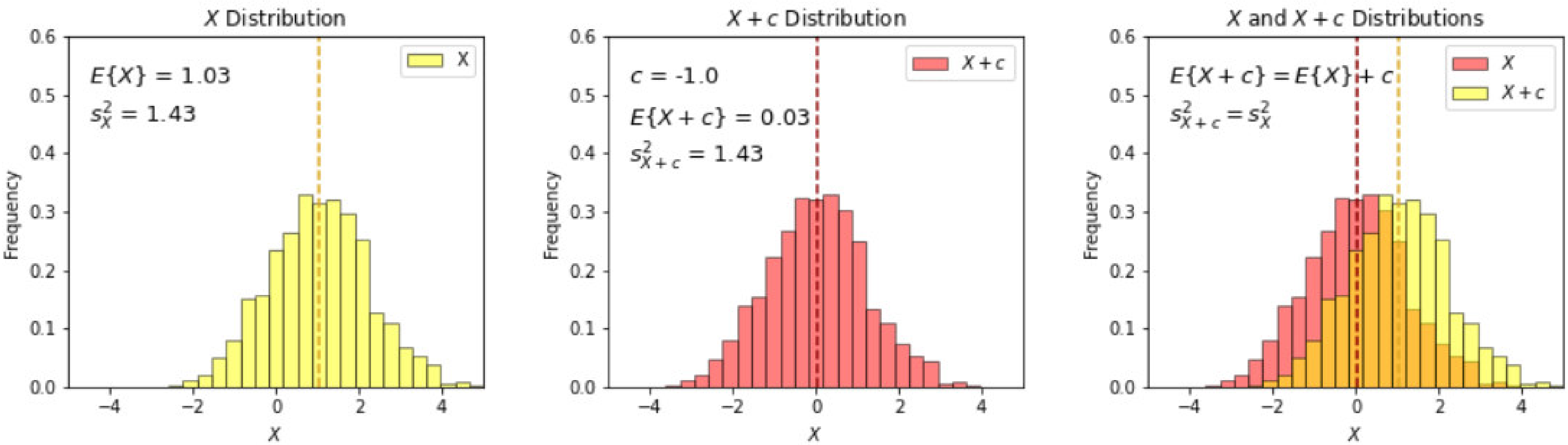
$X$ and $Y$ are independent

# STATISTICAL EXPECTATION HANDS ON

Expectation Operators:

$$E[X + c] = E[X] + E[c] = E[X] + c$$

- Expectation of a random variable + a constant
- Here's an example of a constant added to a random variable.



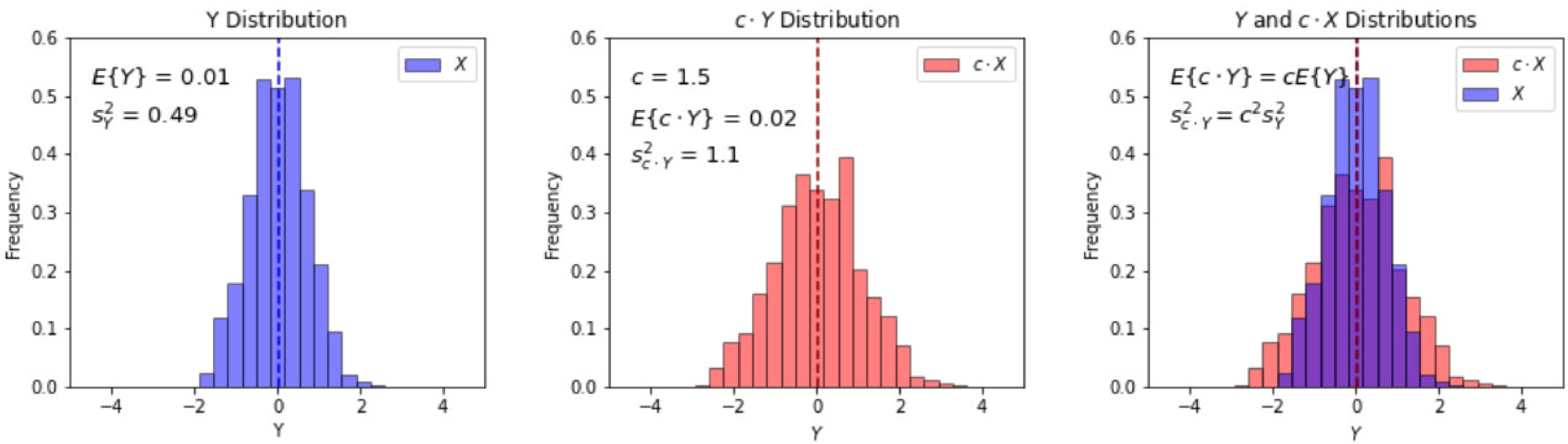

Demonstration of random variable, $X$, + constant, -1.0, file is PythonNumericalDemos_Expectation.ipynb.

# STATISTICAL EXPECTATION HANDS ON

Expectation Operators:

$$E[cX] = cE[X]$$

- Expectation of a product of a random variable and a constant
- Here's an example of random variable multiplied by a constant. The random variable mean is 0.0, so only the variance changes.



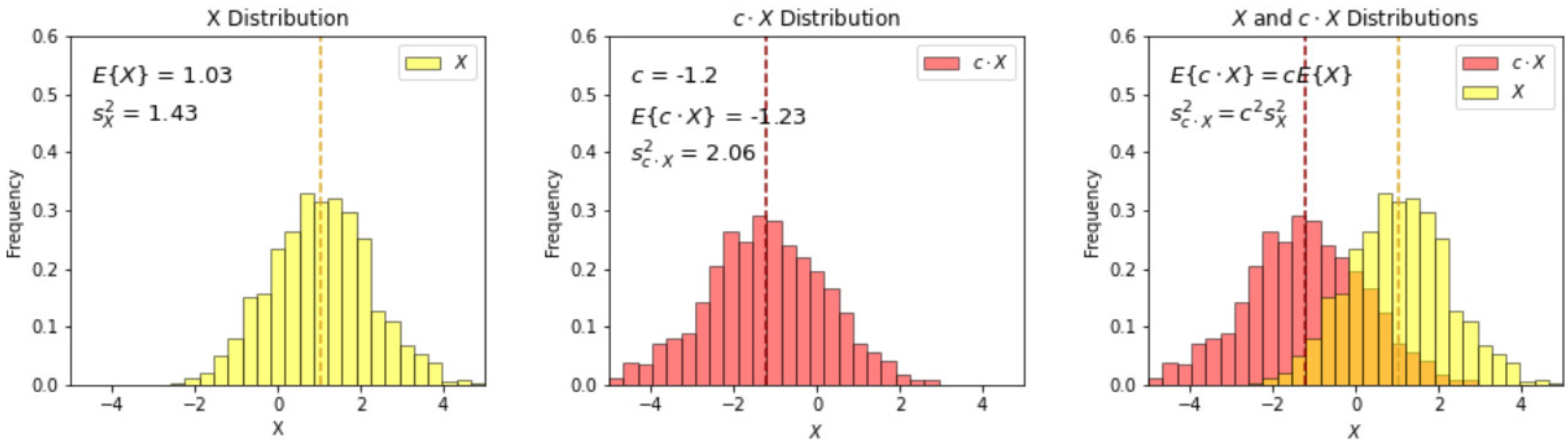

Demonstration of random variable, Y, times a constant, 1.5, file is PythonNumericalDemos_Expectation.ipynb.

# STATISTICAL EXPECTATION HANDS ON

Expectation Operators:

$$E[cX] = cE[X]$$

- Expectation of a product of a random variable and a constant
- Here's an example of random variable multiplied by a constant. The random variable average and variance changed.

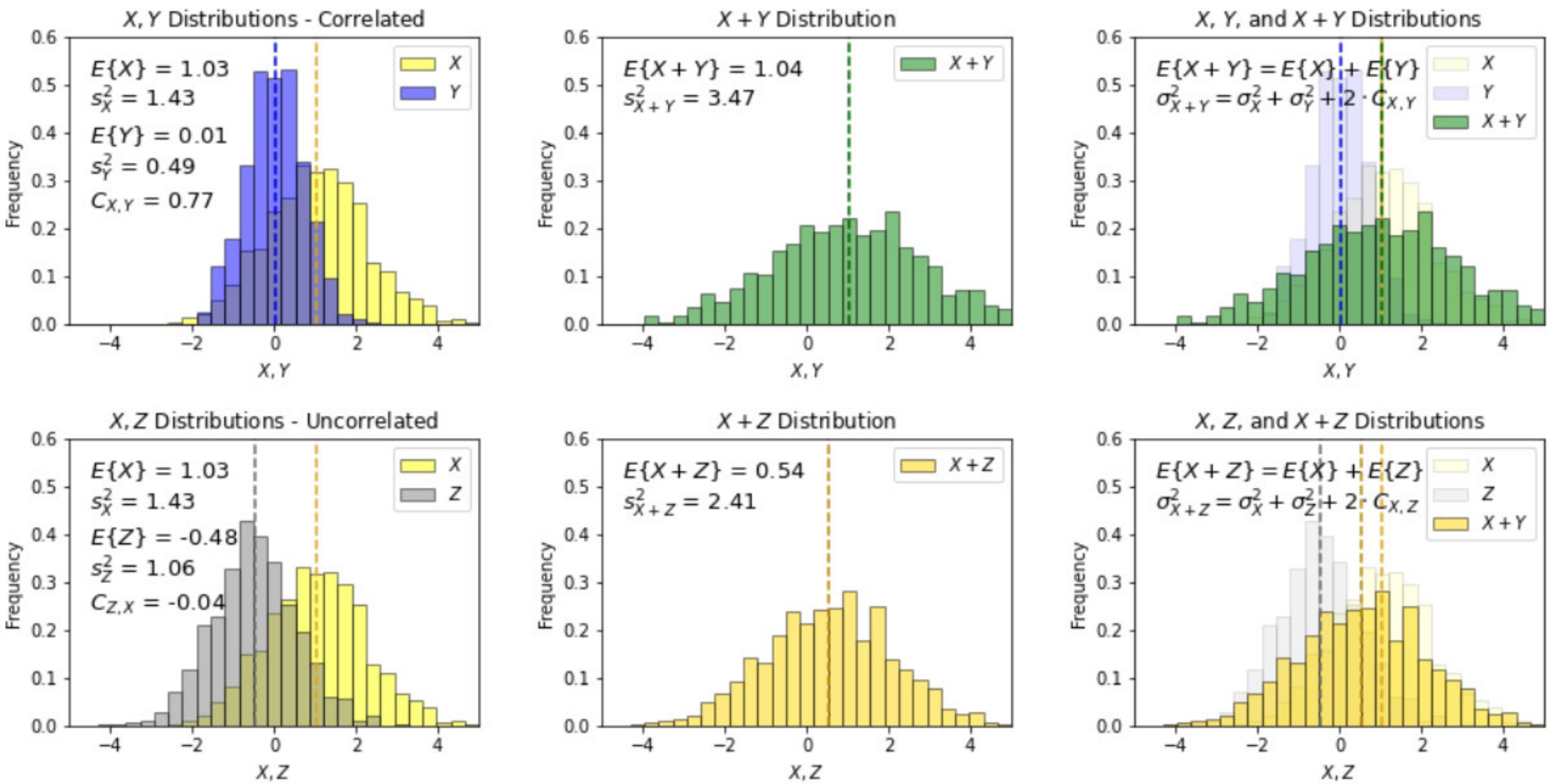



Demonstration of random variable, Y, times a constant, 1.5, file is PythonNumericalDemos_Expectation.ipynb.

# STATISTICAL EXPECTATION HANDS ON

Expectation Operators: $E[X+Y] = E[X] + E[Y]$

- Expectation of addition of two random variables




Demonstration of addition of random variables, $X + Y$ (above), and $X + Z$ (below), file is PythonNumericalDemos_Expectation.ipynb.

# STATISTICAL EXPECTATION HANDS ON

- Some practice with expectation:
  - The expected total porosity is 16%, you estimate a reduction by 3% to all total porosity values to calculate effective porosity. Total porosity ($\varphi_t$) and effective porosity ($\varphi_e$) are random variables.

$$\varphi_e = \varphi_t - c, E\{\varphi_t\} = 16\%, c = 3\%, E\{\varphi_e\} = ?$$

$$E\{\varphi_e\} = E\{\varphi_t - c\} =$$

  - The expected absolute permeability is 100 mD and the relative permeability is 0.25, calculate the expected effective permeability for the reservoir. Absolute permeability ($k$) and effective permeability ($k_i$) are random variables.

$$k_i = k_{ri}k, E\{k\} = 100\, mD, k_{ri} = 0.25, E\{k_i\} = ?$$

$$E\{k_i\} = E\{k_{ri}k\} =$$

# STATISTICAL EXPECTATION HANDS ON

- Some practice with expectation:
  - The expected total porosity is 16%, you estimate a reduction by 3% to all total porosity values to calculate effective porosity. Total porosity ($\varphi_t$) and effective porosity ($\varphi_e$) are random variables.

$$\varphi_e = \varphi_t - c, E\{\varphi_t\} = 16\%, c = 3\%, E\{\varphi_e\} = ?$$

$$E\{\varphi_e\} = E\{\varphi_t - c\} = E\{\varphi_t\} - E\{c\} = E\{\varphi_t\} - c = 16\% - 3\% = 13\%$$

  - The expected absolute permeability is 100 mD and the relative permeability is 0.25, calculate the expected effective permeability for the reservoir. Absolute permeability ($k$) and effective permeability ($k_i$) are random variables.

$$k_i = k_{ri}k, E\{k\} = 100\ mD, k_{ri} = 0.25, E\{k_i\} = ?$$

$$E\{k_i\} = E\{k_{ri}k\} = k_{ri}E\{k\} = 0.25\ (100\ mD) = 25\ mD$$

# STATISTICAL EXPECTATION HANDS ON

- Some practice with expectation:
  - Given random variables (uncertain with a range of possible outcomes) from a water reservoir with:

$$E\{\varphi\} = 15\%, E\{Area\} = 1{,}000{,}000m^2,$$
$$E\{thickness\} = 100m, E\{s_w\} = 1.0$$

  - Assuming independence, calculate the expected water volume in the reservoir .

$$v_w = \varphi \cdot Area \cdot thickness \cdot s_w$$

$$E\{v_w\} = E\{\varphi \cdot Area \cdot thickness \cdot s_w\} = ?$$

# STATISTICAL EXPECTATION HANDS ON

- Some practice with expectation:
  - Given random variables (uncertain with a range of possible outcomes) from a water reservoir with:

$$E\{\varphi\} = 15\%, E\{Area\} = 1{,}000{,}000m^2,$$
$$E\{thickness\} = 100m, E\{s_w\} = 1.0$$

  - Assuming independence, calculate the expected water volume in the reservoir .

$$v_w = \varphi \cdot Area \cdot thickness \cdot s_w$$

$$E\{v_w\} = E\{\varphi \cdot Area \cdot thickness \cdot s_w\} =$$

$$E\{v_w\} = E\{\varphi\}E\{Area\}E\{thickness\}E\{s_w\}$$

$$= 0.15 \cdot 1{,}000{,}000m^2 \cdot 100m \cdot 1.0 = 15\ Mm^3$$

# MULTIVARIATE NEW TOOLS

| Topic | Application to Subsurface Modeling |
|---|---|
| **Multivariate Analysis** | In the presence of multivariate relationships, must jointly model variables.<br><br>*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.<br><br>*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.*<br><br>*Assess the influence of acoustic impedance on predicting porosity away from wells with conditional distributions.* |
| **Statistical Expectation** | *Statistical expectation for math with uncertainty distributions.*<br><br>*Use statistical expectation for optimum decision making in the presence of uncertainty.* |

## Lecture Outline

- Basic Data Analytics
- Multivariate Analysis