PGE 338 Data Analytics and Geostatistics

Lecture 4: Univariate Summaries

Lecture outline . . .

- Measures of Centrality
- Measures of Dispersion
- Measures of Shape
- Statistical Expectation

Michael Pyrcz, The University of Texas at Austin

Writing a Report for Your Boss

Concise and easy to understand. Your boss is busy. You need your boss to effortlessly, efficiently get the following:

Executive Summary (this is all they will likely read)

1. **What was the problem?**
   – How does it impact value?
2. **What did you do to address the problem?**
   – Could be a proposal.
3. **What did we learn?**
   – The result / outcome
4. **What is your recommendation?**
   – What should we do in the future?
   – How does this add value?

Most assignments will include one executive summary.

Example Executive Summary

Shale fraction samples were recently collected. If there are

What is the Problem? Why is it important?

anomalous samples they will bias summary statistics, potentially impacting the accuracy of our subsurface assessment. To QC the data we checked for outliers with the Tukey (1.5 times the

How was the problem addressed?

What is the outcome?

interquartile range) approach. No outliers were detected, I recommend that we use the entire dataset for all future analysis.

What is the recommendation? How is value added?

Motivation

Univariate statistics are summaries of our sample data for one feature.

- concise/compact descriptions
- new lens to explore our data, find patterns
- predictive models on their own or the inputs for predictive models.
- later we will talk about significance and confidence intervals for uncertainty models

PGE 338 Data Analytics and Geostatistics

Lecture 4: Univariate Summaries

Lecture outline . . .

- Measures of Centrality

Michael Pyrcz, The University of Texas at Austin

Measures of Central Tendency

- **Arithmetic Average / Mean**

Note, population mean is denoted as μ.

Sample mean, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

- **Median (P50)**

assumes sorted into ascending order

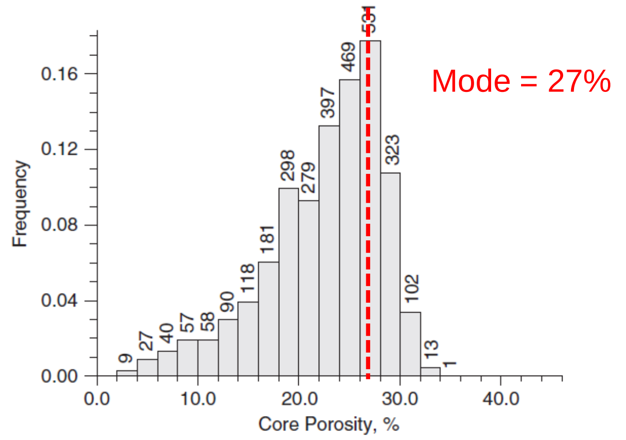
$$\text{Median}(x) = \begin{cases} \boldsymbol{x}_{(\boldsymbol{n}+1)/2} & \text{, if } n \text{ is odd} \\ \dfrac{\boldsymbol{x}_{\boldsymbol{n}/2} + \boldsymbol{x}_{(\boldsymbol{n}/2+1)}}{2} & \text{, if } n \text{ is even} \end{cases}$$

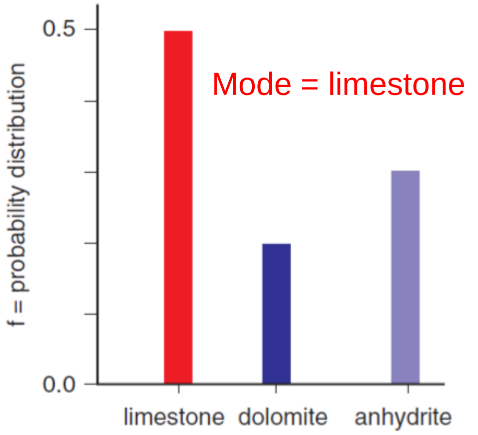
If n is even — **Average of These 2** — x_1 … P50 … x_8

If n is odd — **This Value** — x_1 … P50 … x_7

- **Mode**

  - Most common value
  - Continuous, largest bin, sensitive to bin choice
  - Categorical, category with the highest proportion





Mode for continuous and categorical histograms.

Measures of Central Tendency Example

- The following fraction of shale was measured from 9 core samples. They have been sorted in ascending order.
  - 3%, 4%, 8%, 8%, 8%, 10%, 13%, 16%, 20%
- Question: What is the mean, median and mode?

Measures of Central Tendency Example

- The following fraction of shale was measured from 9 core samples. They have been sorted in ascending order.
  - 3%, 4%, 8%, 8%, 8%, 10%, 13%, 16%, 20%
- Question: What is the mean, median and mode?

Mean: Sum / Count = 90 / 9 = 10%

Median: Sort and take 5^{th} value of the 9 = 8%

Mode: Take most common value = 8%

- could use binning like a histogram or model a PDF for a more robust result

Estimation Error Minimization

We can interpret measures of central tendancy as an estimation problem

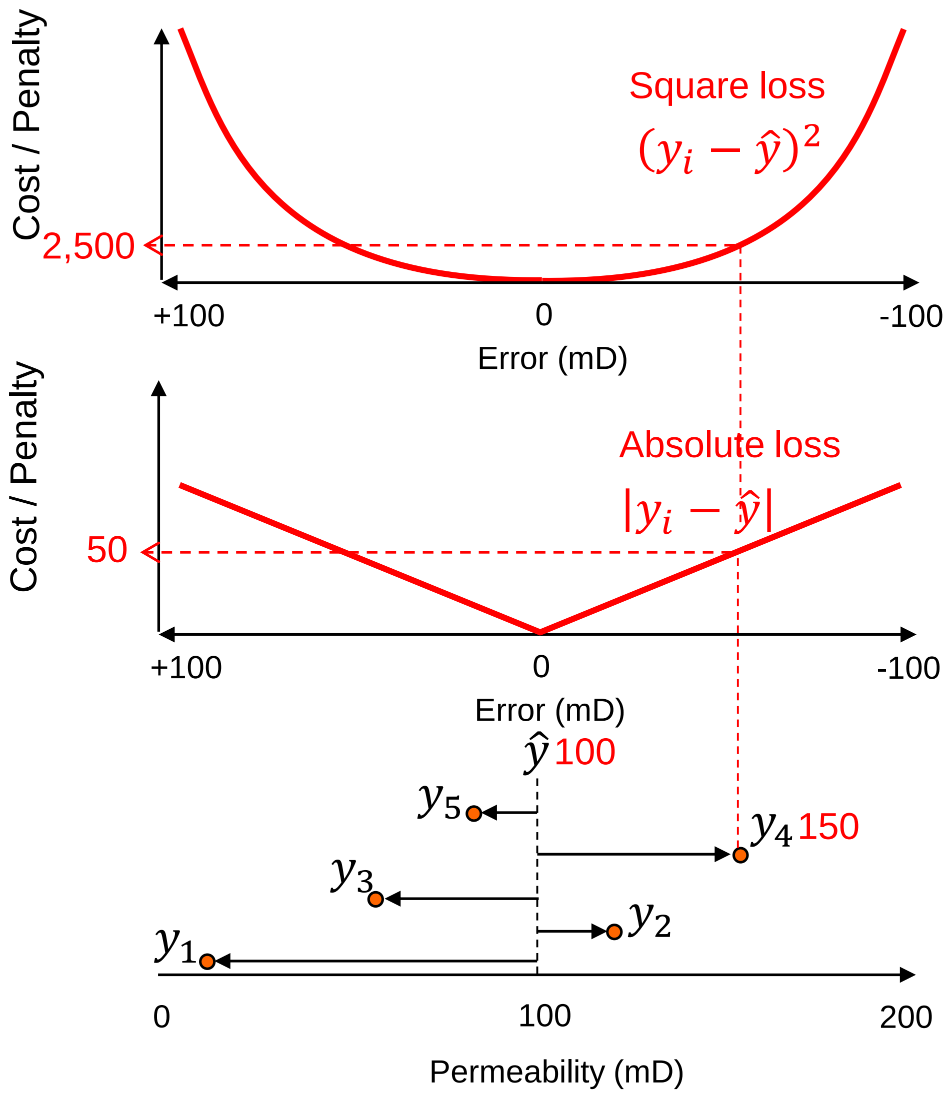
- estimate a single value, $\hat{y}$, to represent the feature, samples, $y_1, \ldots, y_n$.

Arithmetic average / mean is the best estimate to minimize the square error

- if the cost / penalty of an error is squared, $\sum_{i=1}^{n}(y_i - \hat{y})^2$, the mean is the best estimate, where y_i is a data value and $\hat{y}$ is the single estimate.

Median is the best estimate to minimize the absolute error

- if the cost / penalty of an error is the absolute value, $\sum_{i=1}^{n}|y_i - \hat{y}|$, where y_i is a data value and $\hat{y}$ is the single estimate.



Loss functions (above) and data (●) and estimate, $\hat{y}$, with errors shown (below).

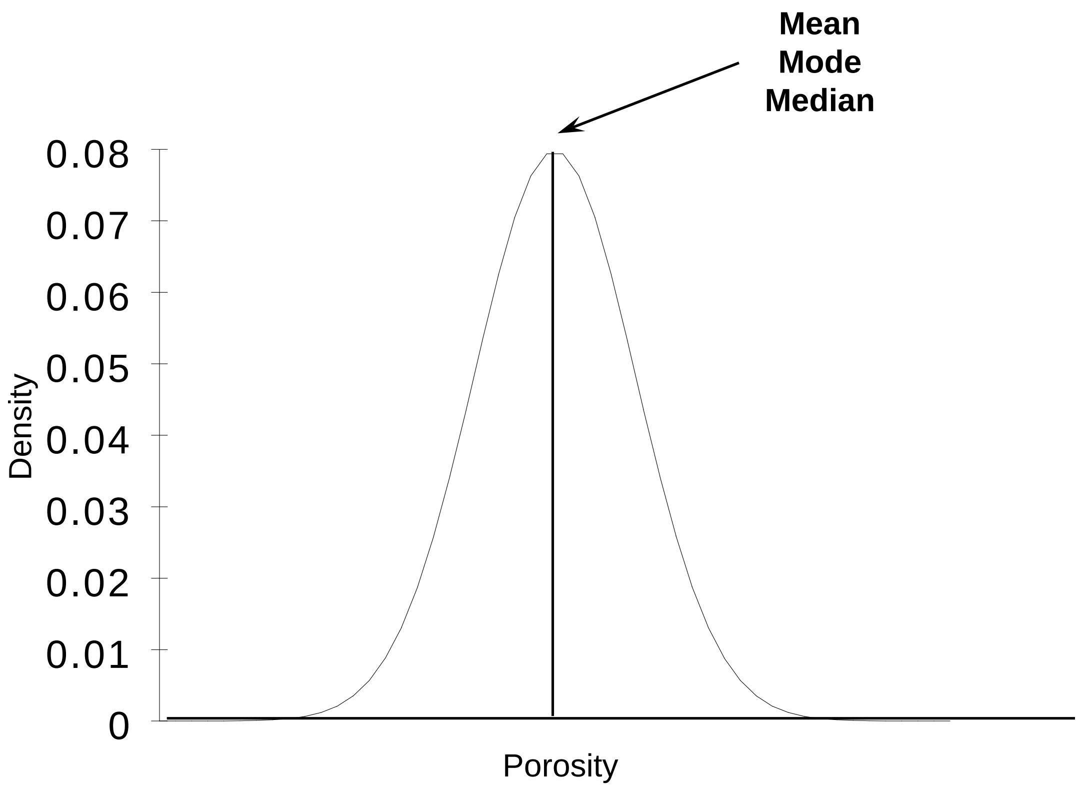
Michael Pyrcz, Associate Professor, The University of Texas

Measures of Central Tendency

Measures of Central Tendency for the Gaussian (Normal) Distribution

- Mode = Median = Mean
- Unimodal, having one mode, and symmetric

We will formalize the Gaussian distribution next lecture.



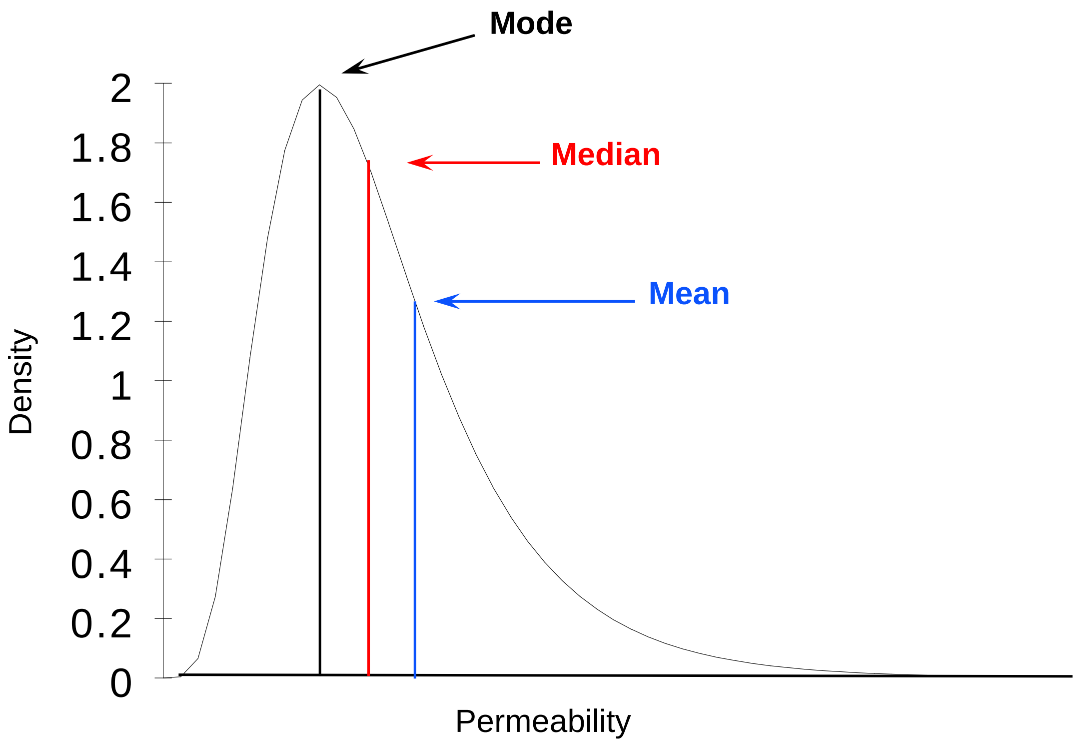
Gaussian parametric distribution with mean, mode and median labeled.

Measures of Central Tendency

Measures of Central Tendency for the Lognormal Distribution

- Mode < Median < Mean
- Unimodal, having one mode, and skewed
- Note, the arithmetic average / mean is very sensitive to extreme values, outliers

We will formalize the lognormal distribution next lecture.

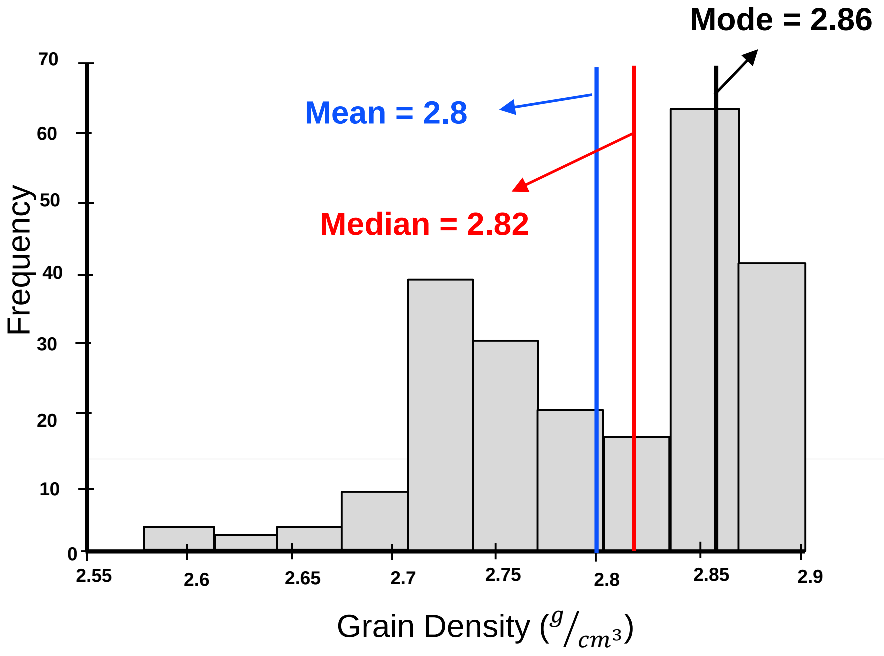


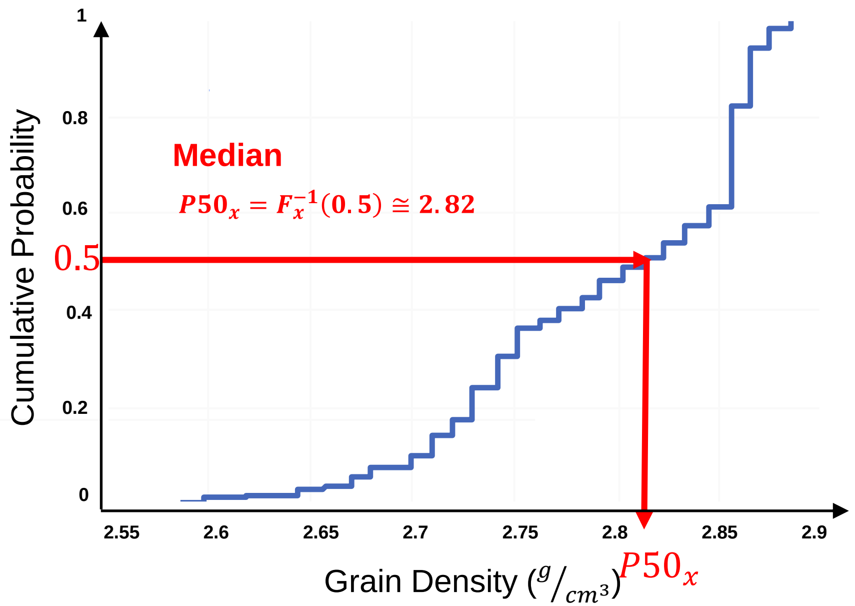
Lognormal parametric distribution with mean, mode and median labeled.

Measures of Central Tendency

Central Tendency Illustrated with Histogram and CDF

- mean, median and model for a nonparametric example
- labeled on histogram, note ordering due to a few very low values
- median can be observed from the CDF





Nonparametric distribution, histogram (left) and CDF (right) with mean, mode and median labeled.

Measures of Central Tendency

Geometric Mean, $\bar{x}_G$

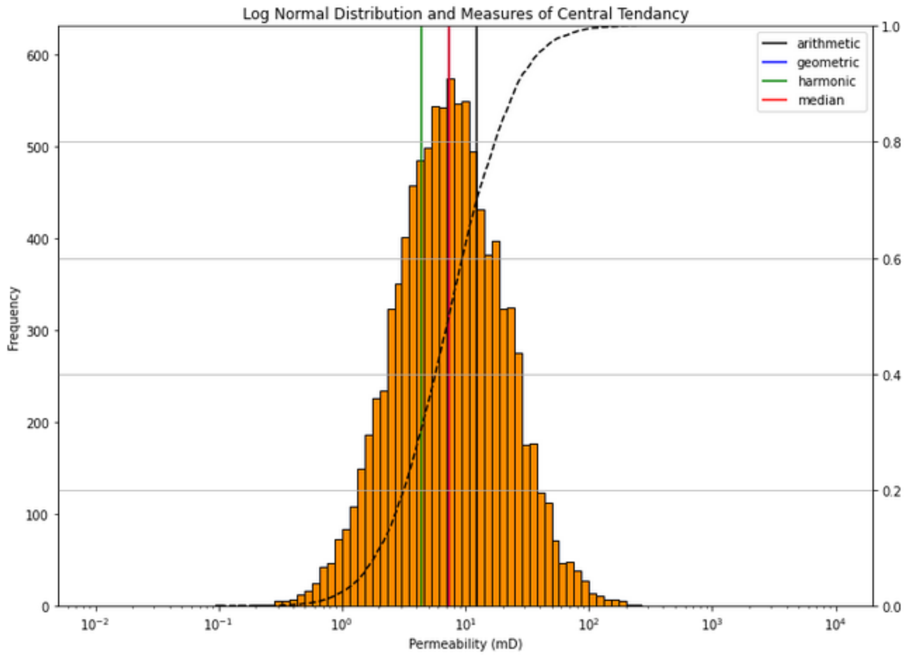
$$\bar{x}_G = (x_1 \cdot x_2 \cdot \cdots \cdot x_n)^{\frac{1}{n}}$$

- commonly applied for central tendency of lognormal distributions
- lognormal distributions converge to $\bar{x}_G$ as the variance shrinks to 0.0.

Harmonic Mean, $\bar{x}_H$

$$\bar{x}_H = \frac{n}{\left(\frac{1}{x_1}+\frac{1}{x_2}+\cdots+\frac{1}{x_n}\right)}$$

- effective permeability for flow perpendicular to layers

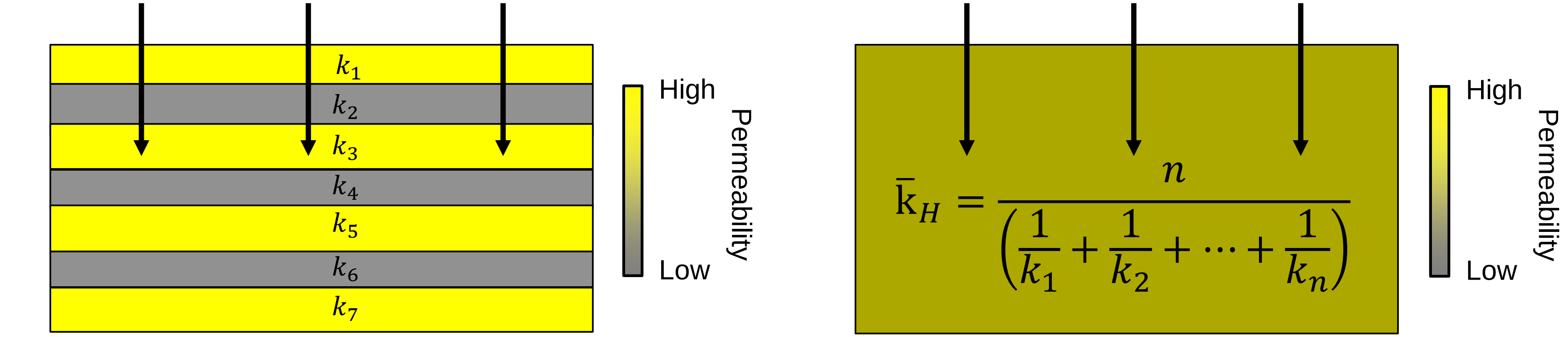


Lognormal distribution with arithmetic average, harmonic mean, geometric mean (under median line) and median (50th percentile). File is: PythonDataBasics_Measures_of_Central_Tendency.ipynb.

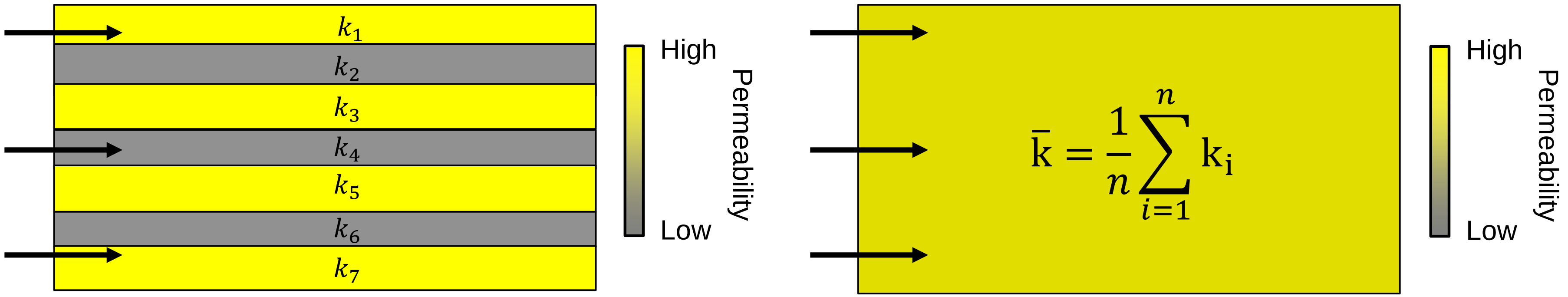
Measures of Central Tendency

Another Interpretation of Central Tendency is Effective Property

- could I replace all the permeabilities of these layers with a single effective permeability?
  - when we apply flow simulation to both models, they flow the same!



Harmonic mean is applied to calculate effective permeability for flow across layers, smallest permeabilities have the greatest impact.



Arithmetic mean is applied to calculate effective permeability for flow along layers, extreme permeabilities have the greatest impact.

Measures of Central Tendency

Another Interpretation of Central Tendency is Effective Property

- a more general form is **power law averaging**

$$\bar{x}_P = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}}$$

Power Law Averaging

$$k_{eff} = \left[\frac{1}{v}\int_v k(u)^p du\right]^{\frac{1}{p}}$$

Power Law Averaging for Volumetric Scale Up of Permeability, k

Example of continuous permeability power law upscaling.

- useful to calculate effective permeability where flow is not parallel nor perpendicular to distinct permeability layers
- flow simulation may be applied to calibrate (calculate the appropriate power for power law averaging)

Power = -1
Harmonic
Average

Power = 0
Geometric
Average*

Power = 1
Arithmetic
Average

* Proof in limit as p→0, see Zanon (2002) on Canvas.

PGE 338 Data Analytics and Geostatistics

Lecture 4: Univariate Summaries

Lecture outline . . .

- Measures of Dispersion

Michael Pyrcz, The University of Texas at Austin

Measures of Dispersion

Variance Related Measures

- **population variance**, average squared difference from the mean

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

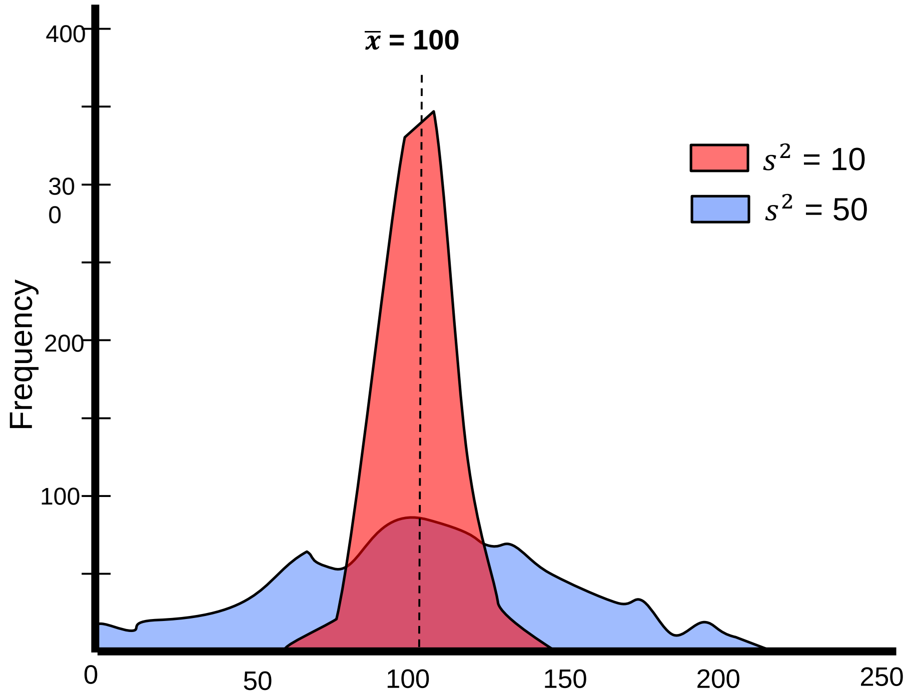
μ is population mean inferred from sample mean, same calculation.

population variance calculated from the entire population.

- **sample variance**

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

sample variance calculated from a sample, $n-1$ is the **degrees of freedom**, accounts for the fact we have n pieces of information, but the population mean is unknown and assumed. More later.



Two sample datasets with the same sample mean and different sample standard deviations, 10 and 50.

- **standard deviation**

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

measure of dispersion in the original units, population and sample standard deviation.

Measures of Dispersion

Range

- calculated as the minimum value subtracted from the maximum value

$$Range_x = max_x - min_x$$

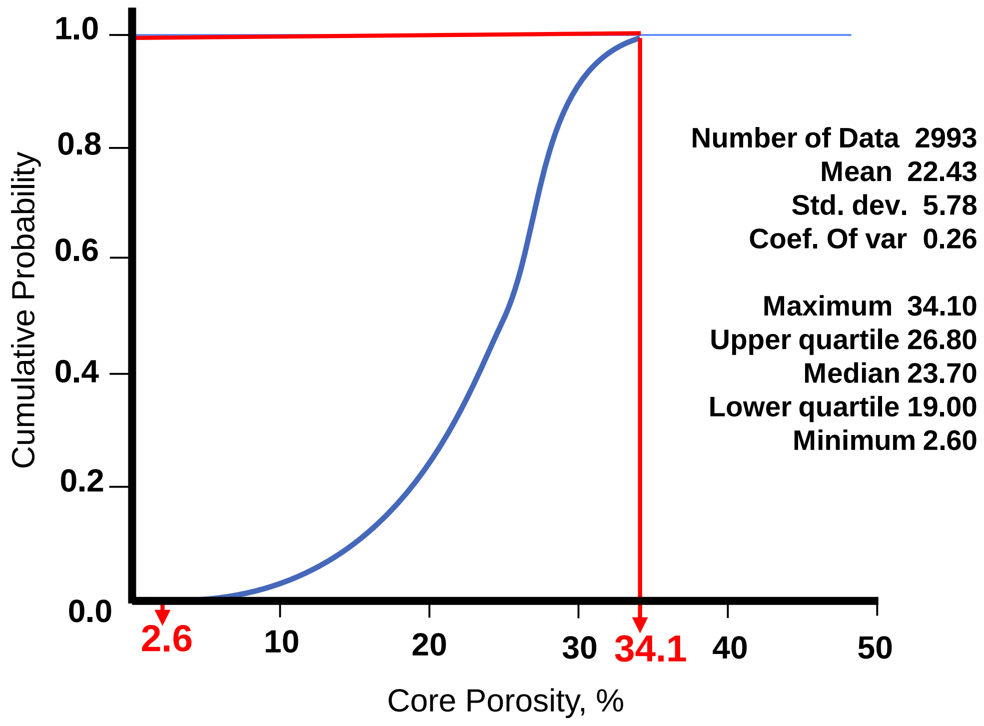
$$Range_x = F_x^{-1}(1.0) - F_x^{-1}(0.0)$$

Recall percentile notation:

$$Pxx = F_x^{-1}(xx)$$

For example:

P25 is the 25th percentile from the CDF, $F_x^{-1}(0.25)$



CDF of 2,993 porosity samples.

$$Range_x = F_x^{-1}(1.0) - F_x^{-1}(0.0)$$

$$Range_x = 34.1 - 2.6$$

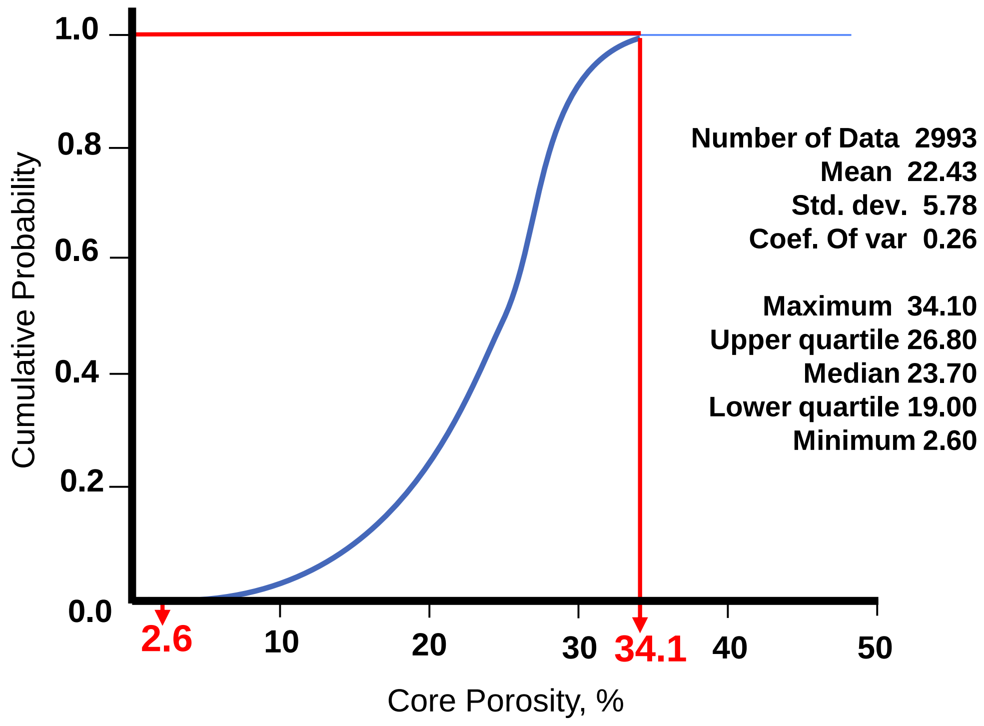
$$Range_x = 31.5$$

Measures of Dispersion

Problems with the Range:

- the minimum and maximum are the most unreliable measures.
- what is the chance that you sampled the extremes? There are very, very little density on the distribution tails!
- outliers could be present (to be discussed)

Safer to work with quartiles, e.g., interquartile range (next)



CDF of 2,993 porosity samples.

$$Range_x = F_x^{-1}(1.0) - F_x^{-1}(0.0)$$

$$Range_x = 34.1 - 2.6$$

$$Range_x = 31.5$$

Measures of Dispersion

Quartiles-based Dispersion

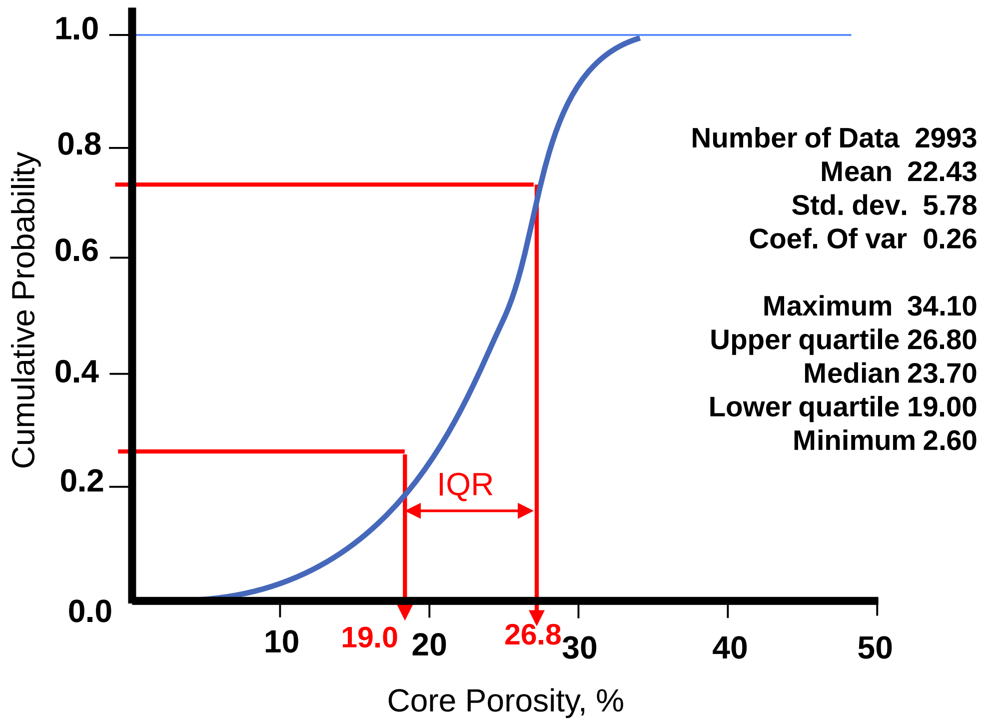
- divide the CDF into 4 equal cumulative probability bins, with bin boundaries:
  - P25, P50, P75

- then use upper and lower bin boundaries, the interquartile range:

$$IQR = F_x^{-1}(0.75) - F_x^{-1}(0.25)$$

Not generally sensitive to outliers or extreme values.

Therefore, often a more reliable measure of dispersion than the range.



CDF of 2,993 porosity samples.

$$IQR = F_x^{-1}(0.75) - F_x^{-1}(0.25)$$

$$IQR = 26.8 - 19.0$$

$$IQR = 7.8$$

Measures of Dispersion

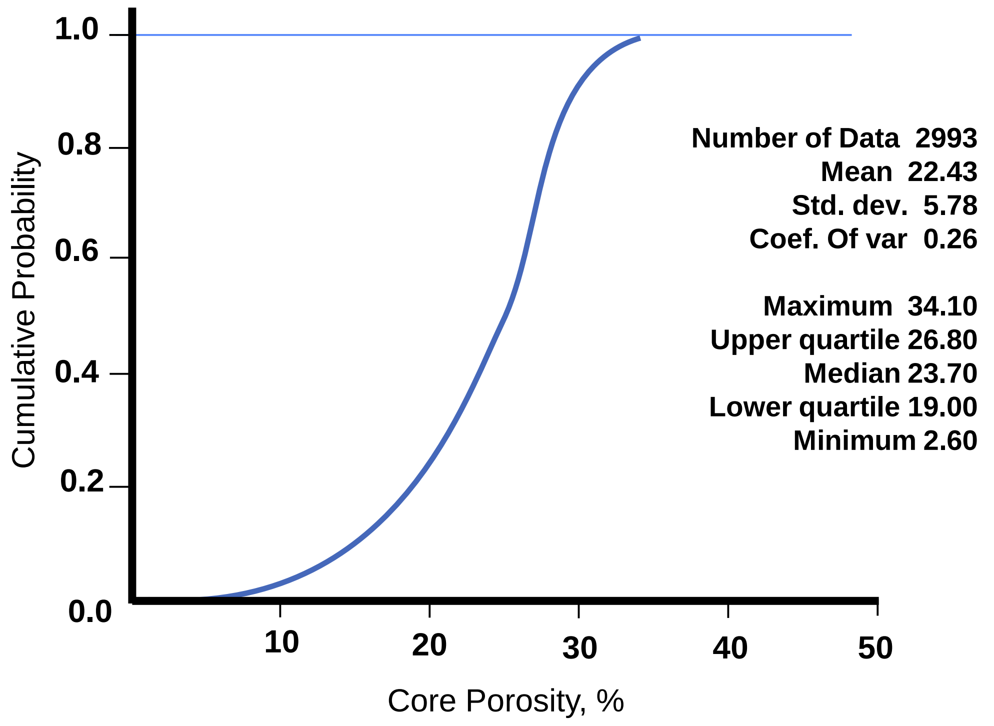
Quantiles

Quantile is any $\left[\frac{1}{q}, \frac{2}{q}, \ldots, \frac{q-1}{q}\right]$ that divides into equal parts.

- Quintiles
  - P20, P40, P60, P80
- Deciles
  - P10,P20,…,P90
- Percentiles
  - P01,P02,…,P99

Could report any percentile values.

- Quartiles with IQR is common, P25 and P75
- So is deciles with P90 and P10
- Dykstra Parsons (later) uses P50 and P16!



CDF of 2,993 porosity samples.

Reporting Percentiles

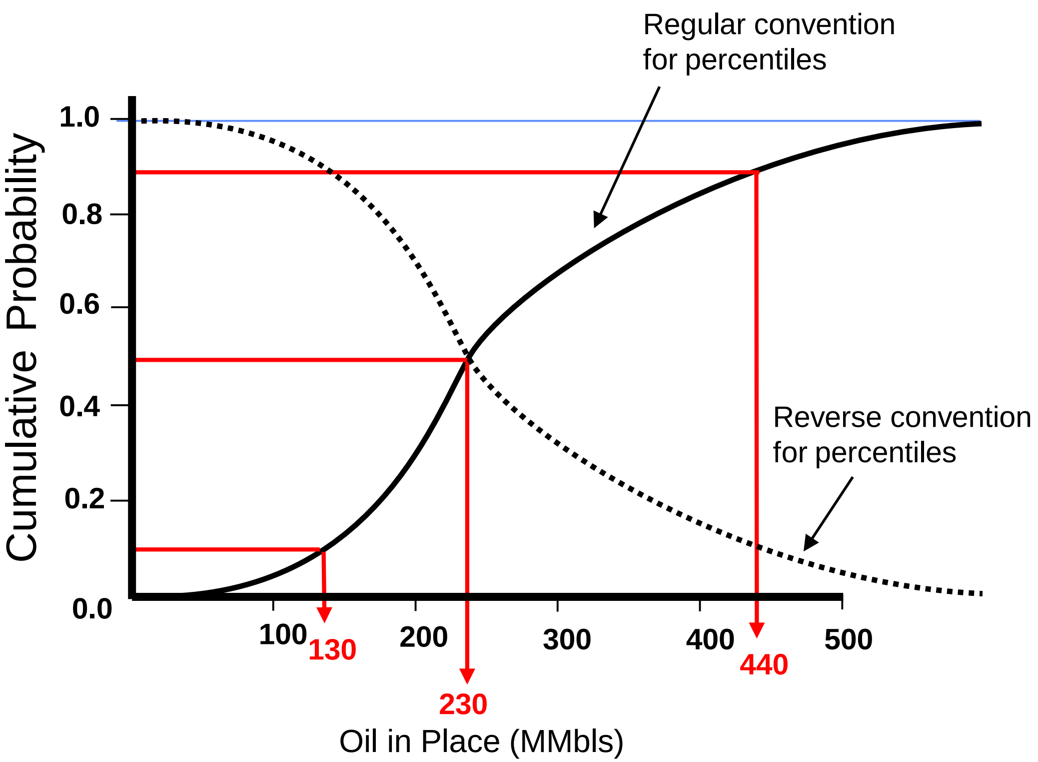
Two Conventions for Percentiles

- most companies report the P10, P50 and P90 as summary statistics from their uncertainty distributions.

- Oil in Place Regular:
    - P10 = 130 MMbls
    - P50 = 230 MMbls
    - P90 = 440 MMbls

- some companies reverse this convention:
    - $\boldsymbol{P_{x,p}} = F_x^{-1}(1 - p)$

- Oil in Place Reverse:
    - P90 = 130 MMbls
    - P50 = 230 MMbls
    - P10 = 440 MMbls



Standard Convention for Percentiles:

$$P_{x,p} = F_x^{-1}(p), \quad where\ F_x(x) = P(X \leq x)$$

Reverse Convention for Percentiles:

$$P_{x,p} = F_x'^{-1}(p), \quad where\ F'_x(x) = P(X \geq x)$$

Measures of Dispersion
Outlier Detection

Tukey Method for Outlier Detection

1. Calculate the following:

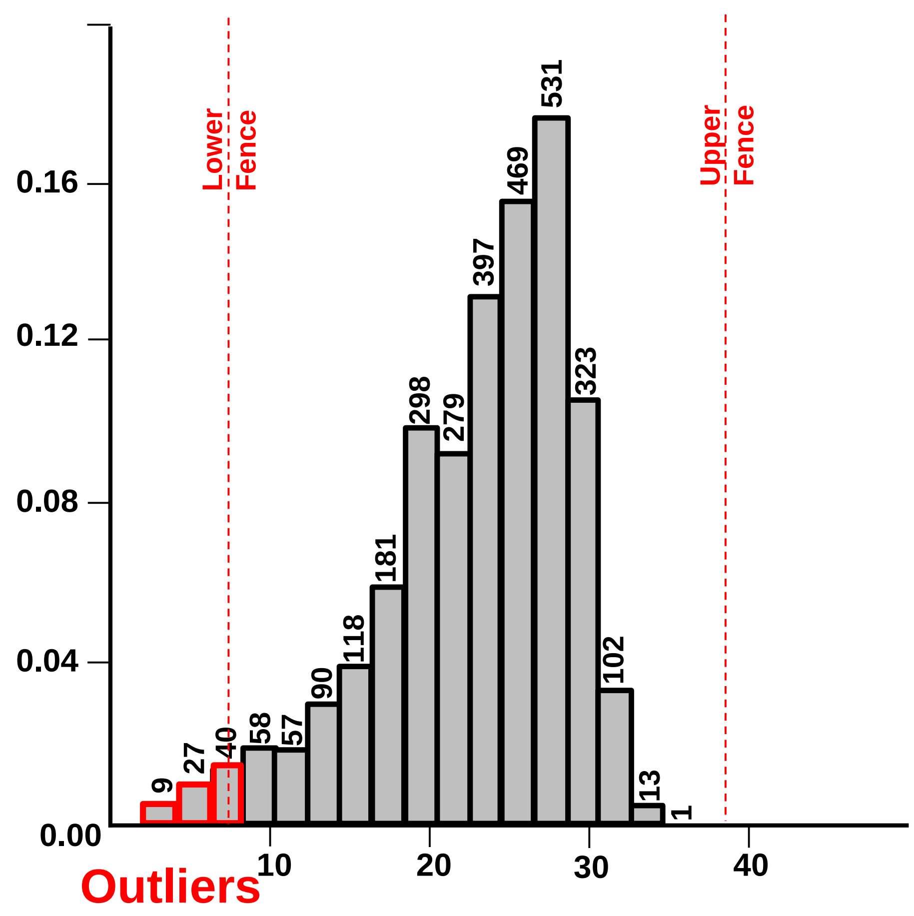
- Lower Fence = $P25 - 1.5 \cdot IQR$
- Upper Fence = $P75 + 1.5 \cdot IQR$

2. Data Samples are Outliers if:
   - $x <$ Lower Fence
   - $x >$ Upper Fence

Example:

P25 = 19.0%, P75 = 26.8%
IQR = P75 – P25 = 7.8%
Lower Fence = 19.0% – 1.5(7.8%) = 7.3%
Upper Fence = 26.8% + 1.5(7.8%) = 38.5%

Samples < 7.3% or > 38.5% are outliers.



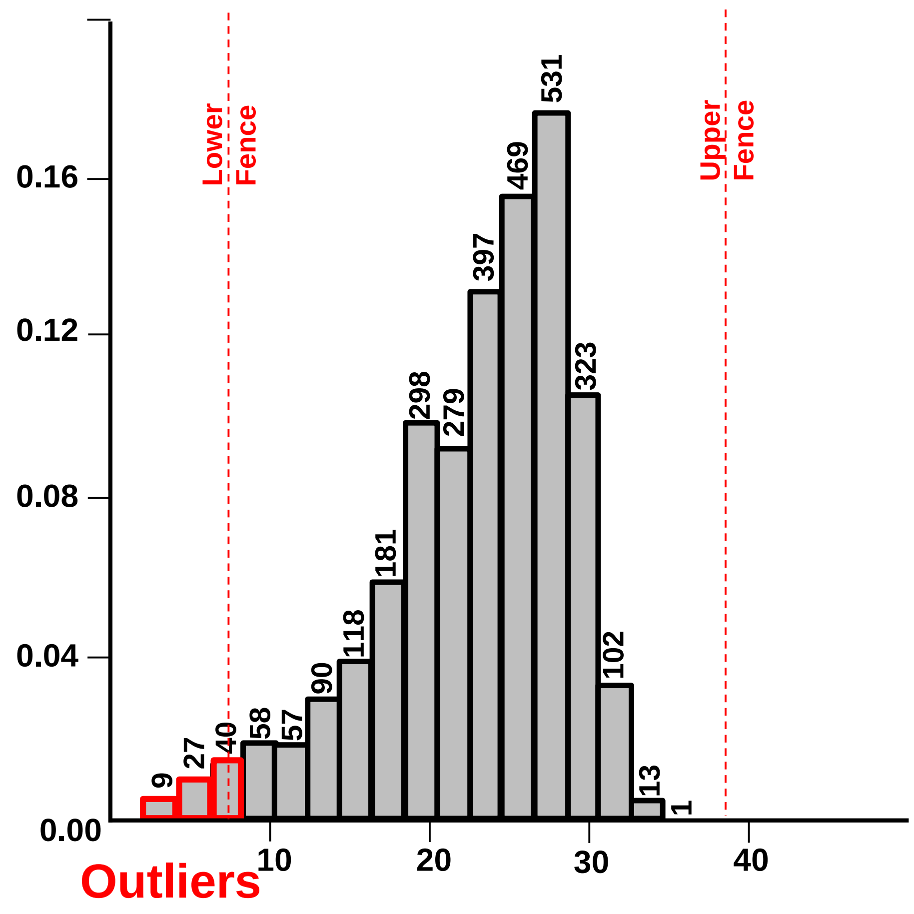
PDF of 2,993 porosity samples, lower and upper fence and bins with outliers.

Measures of Dispersion
Outlier Detection

Outlier Treatment

What to do once outliers are detected?

1. **Remove**: must be able to demonstrate that the data is erroneous
2. **Transform:** (discuss later): reshape the distribution for analysis
3. **Separate**: pull out the outliers and work with them separately. Assumption they are a different population.



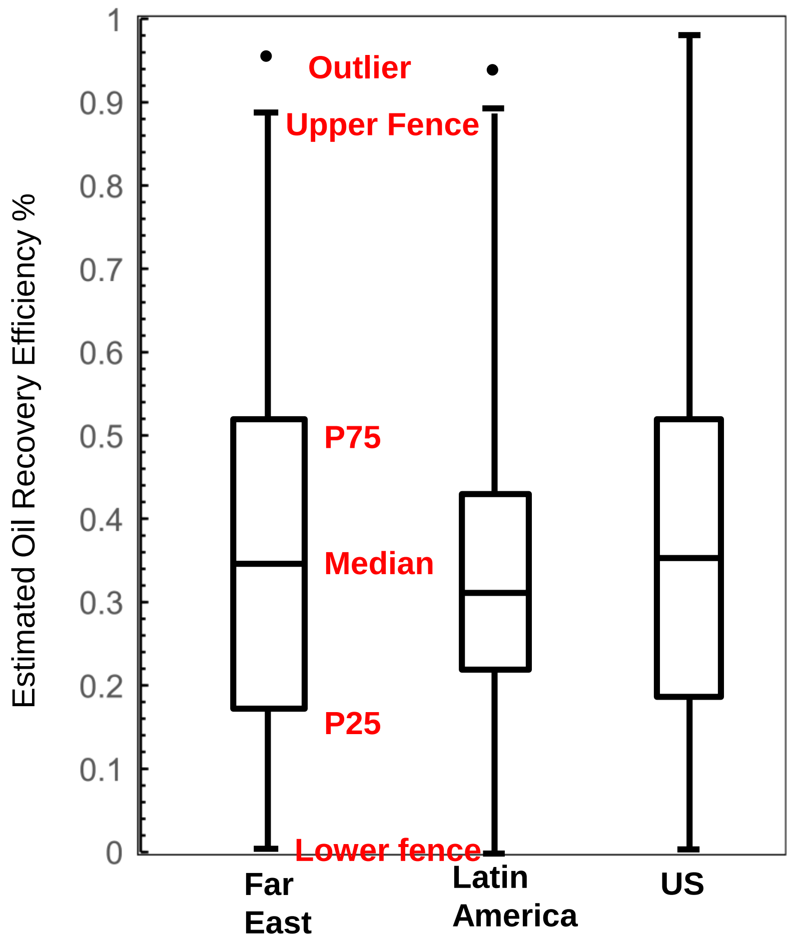
PDF of 2,993 porosity samples, lower and upper fence and bins with outliers.

Box Plots

Visualizing / Comparing Multiple Distributions

- consider box (also known as “box and whisker”) plots
- communicate central tendency, dispersion and outliers
- end of whiskers varies by software. Upper and lower fence (as indicated) are useful for outlier detection.

Sometimes confidence intervals are included to indicate if distributions are significantly different (more on this soon)



Box plot for estimated oil recovery efficiency for wells from three producing

PGE 338 Data Analytics and Geostatistics

Lecture 4: Univariate Summaries

Lecture outline . . .

- Measures of Shape

Michael Pyrcz, The University of Texas at Austin

Moments of a Distribution

General Form and Types of Moments

Quantitative measures of order p related to a shape of a distribution.

All of these are ‘population moments’, ‘sample moments’ are out of scope for brevity.

Moment

$$\frac{1}{n}\sum_{i=1}^{n}(x_i)^p$$

Central Moment

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^p$$

Standardized, Central Moment

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{(x_i-\mu)}{\sigma}\right)^p$$

Moments of a Distribution

Example Moments:

- 1^{st} Moment – Expectation / Average (more on this in the next subsection)

$$E\{X\} = \mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

first moment

- 1^{st} Central Moment = 0

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu) = \cancelto{\mu}{\frac{1}{n}\sum_{i=1}^{n} x_i} - \cancelto{\mu}{\frac{1}{n}\sum_{i=1}^{n} \mu} = \mu - \mu = 0$$

first central moment

- 2^{nd} Central Moment – Variance

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$

second central moment

- 3^{rd} Central Moment – Skew

$$\gamma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3$$

third central moment

- 4^{th} Central Moment – Kurtosis

$$Kurt(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^4$$

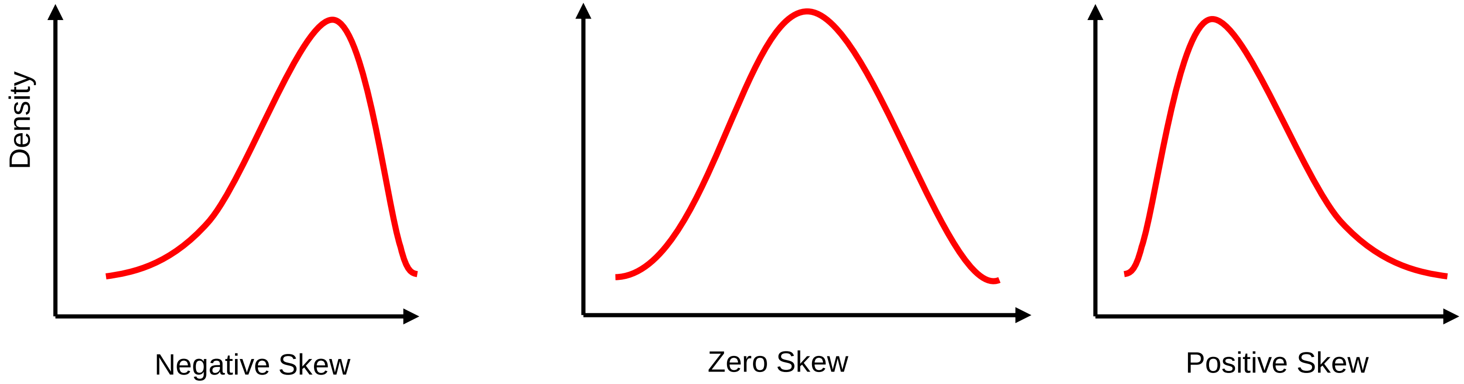
fourth central moment

For Excel and Python check the docs, they may standardized, central moments or central moments.

Moments of a Distribution

3rd Moment – Skew

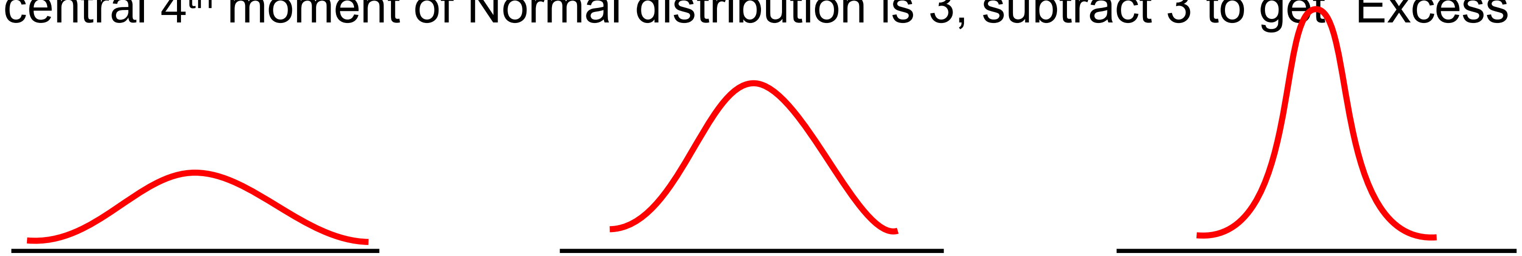
- symmetry of the distribution



PDFs illustrating various skews.

4th Moment – Kurtosis

- ‘peakedness’ of the distribution
- standardized, central 4th moment of Normal distribution is 3, subtract 3 to get “Excess Kurtosis”



Platykurtic distribution
Low degree of peakedness
Excess Kurtosis < 0

Mesokurtic distribution
Same peakedness as Gaussian Distribution”
Excess Kurtosis = 0

Leptokurtic distribution
High degree of peakedness
Excess Kurtosis > 0

PDFs illustrating various kurtoses.

Image (upper) from http://www.janzengroup.net/stats/lessons/descriptive.html and (lower) https://en.wikipedia.org/wiki/Skewness.

Moments of a Distribution

More Skew Measures

Pearson's mode skewness:

$$skewness = \frac{(mean - mode)}{\sigma}$$

departure of mean from mode. Note, Wolfram Mathworld had previously added a spurious '3' multiplication factor.

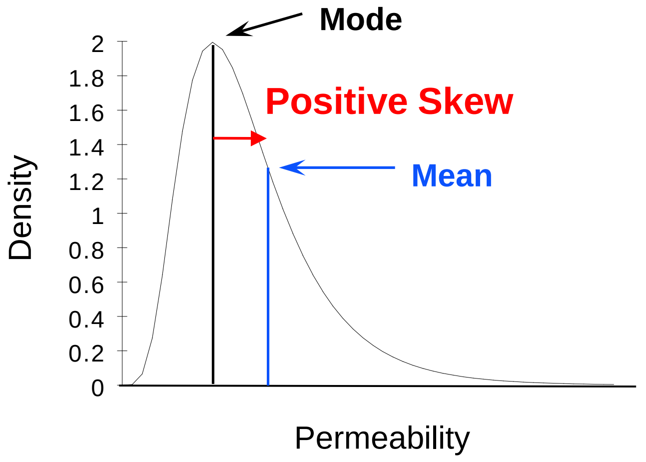


Illustration of inputs for Pearson's mode skewness.

quartile skew coefficient:

$$QS = \frac{(P75 - P50) - (P50 - P25)}{(P75 - P25)}$$

difference in departure of upper and lower quartile from the median value.

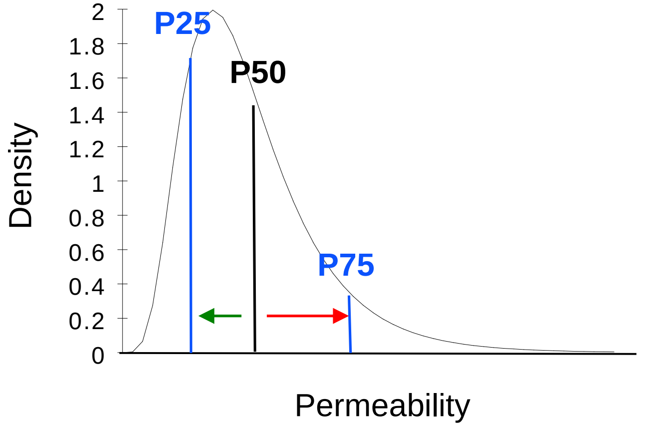


Illustration of inputs for quartile skew coefficient.

Summary Statistics Example

Practice:

- Download the Excel data file called **PorositySample1.xlsx** and work in Excel (PGE 337/Files/Data).
- Plot the CDF.
- Calculate the AVERAGE(), VAR.S(), Quartiles using (PERCENTILE.EXC)
- Check for Outliers
- Calculate Skew() and Kurtosis (Kurt())

Note, for the percentile calculations in Excel there are 2 options:

- PERCENTILE.EXC assumes the $F_i = \frac{i}{n+1}$ cumulative probability method – both tails are not known.

- PERCENTILE.INC assumes the $F_i = \frac{i-1}{n-1}$ cumulative probability method – both tails are known and set to the min and max values of the dataset.

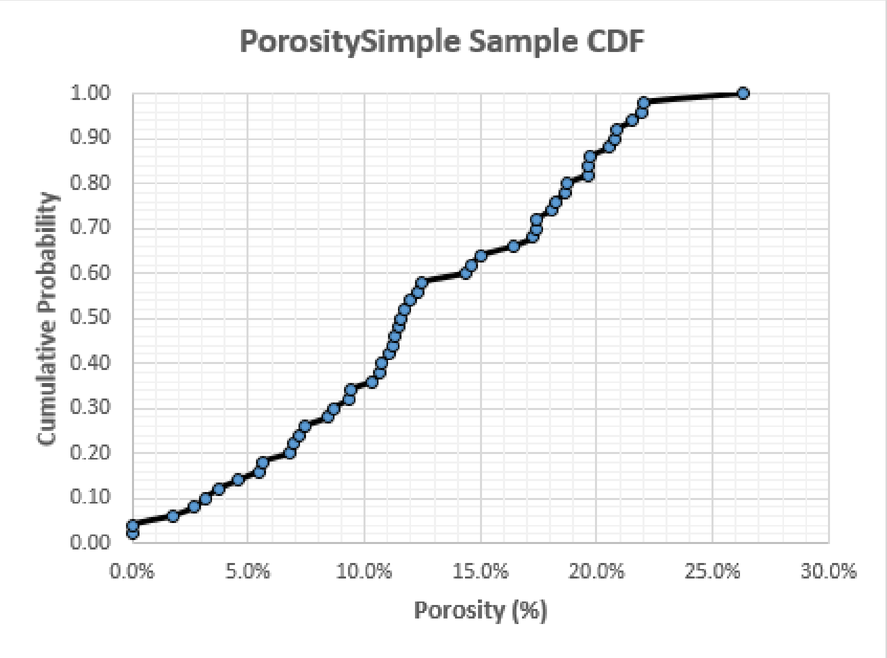
The data file is also at: https://github.com/GeostatsGuy/GeoDataSets

Summary
Statistics Example

Solution:

- What is the shape of this distribution, i.e., what does the PDF look like?

Are there any outliers?



Statistic	By Hand	Excel F(x)	Comments
Average		13%	
Variance (sample)		0.0043	
P25	7.4%	7.4%	Rank 12.8
P50	11.7%	11.7%	Rank 25.5
P75	18.3%	18.3%	Rank 38.3
IQR	10.9%		
Lower Fence	-9.0%		no outliers
Upper Fence	34.7%		
Skew		-0.06	slightly negative skew
Kurtosis		-0.83	platykurtic

Univariate Summary Statistics in Excel

Demonstration / code for calculation of all the discussed univariate statistics in Excel

- measures of central tendency
- measures of dispersion
- Tukey test for outliers
- moments / other measures of shape
- plotting a CDF

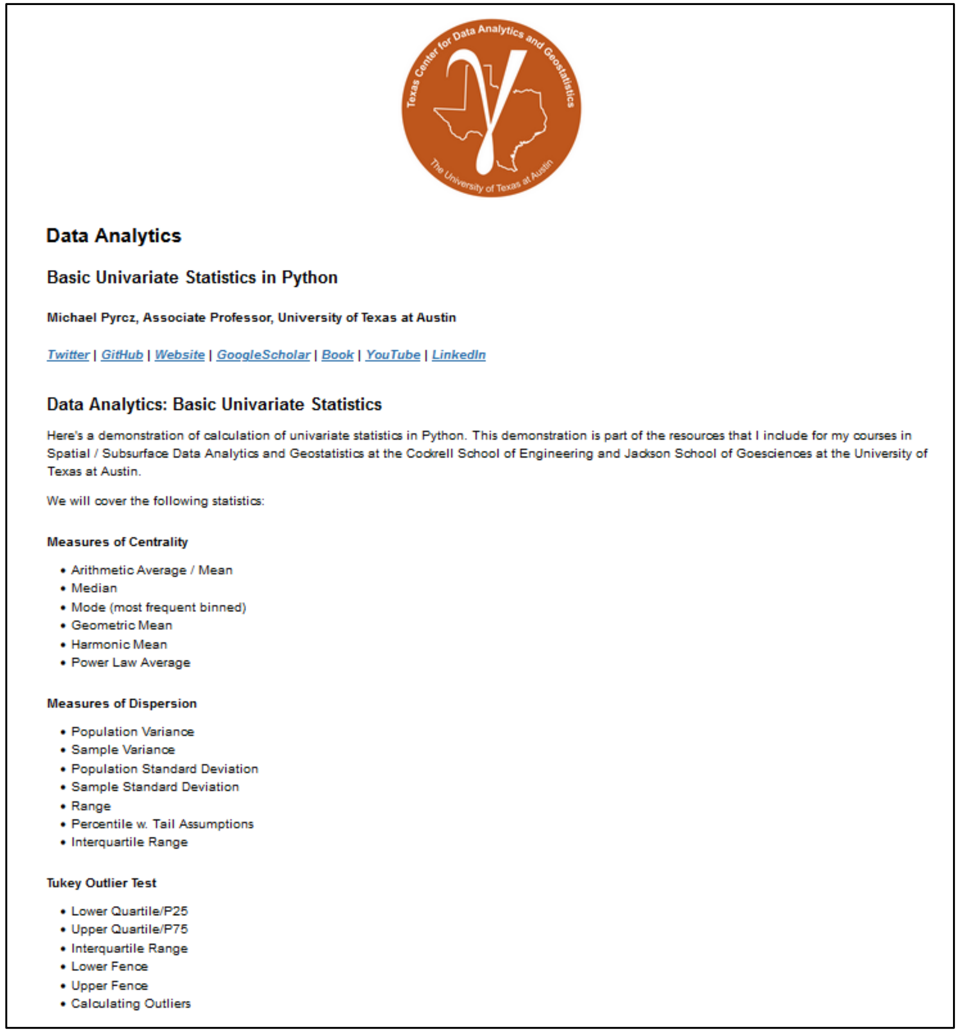
Measures of Centrality	Porosity (ϕ)	Permeability (k)
Arithmetic Average / Mean	11.7	161.0
Median	11.4	144.3
Mode (most frequent binned)	9.0	130.0
Geometric Mean	11.2	143.4
Harmonic Mean	10.6	127.2
Power Law Average	11.6	157.4
Measures of Dispersion		
Population Variance	11.1	6482.5
Sample Variance	11.2	6544.8
Population Standard Deviation	3.3	80.5
Sample Standard Deviation	3.3	80.9
Range	15.3	529.9
Percentile EXC	9.1	103.7
Percentile INC	9.1	104.0
Interquartile Range	4.9	102.6
Tukey Outlier Test		
P25	9.1	104.0
P75	14.0	206.6
Interquartile Range	4.9	102.6
Lower Fence	1.7	-49.9
Upper Fence	21.4	360.4
Number Outliers	0	2
Measures of Shape		
Skew (standardized, sample)	0.2	1.6
Excess Kurtosis (standardized, sample)	-0.5	5.5
Pearson' Mode Skewness	0.8	0.4
Quartile Skew Coefficient	0.1	0.2

Univariate summary statistics in Python. File is: Basic_Statistics_Demo.xlsx.

Univariate Summary Statistics in Python

Demonstration / code for calculation of all the discussed univariate statistics in Python

- measures of central tendency
- measures of dispersion
- Tukey test for outliers
- moments / other measures of shape
- plotting a CDF



Data Analytics

Basic Univariate Statistics in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

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

Data Analytics: Basic Univariate Statistics

Here's a demonstration of calculation of univariate statistics in Python. This demonstration is part of the resources that I include for my courses in Spatial / Subsurface Data Analytics and Geostatistics at the Cockrell School of Engineering and Jackson School of Goesciences at the University of Texas at Austin.

We will cover the following statistics:

Measures of Centrality

- Arithmetic Average / Mean
- Median
- Mode (most frequent binned)
- Geometric Mean
- Harmonic Mean
- Power Law Average

Measures of Dispersion

- Population Variance
- Sample Variance
- Population Standard Deviation
- Sample Standard Deviation
- Range
- Percentile w. Tail Assumptions
- Interquartile Range

Tukey Outlier Test

- Lower Quartile/P25
- Upper Quartile/P75
- Interquartile Range
- Lower Fence
- Upper Fence
- Calculating Outliers

Univariate summary statistics in Python. File is: PythonDataBasics_Univariate_Statistics.ipynb.

PGE 338 Data Analytics and Geostatistics

Lecture 4: Univariate Summaries

Lecture outline . . .

- Statistical Expectation

Michael Pyrcz, The University of Texas at Austin

Statistical Expectation

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

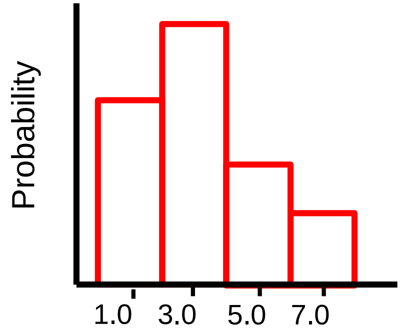
For discrete (binned) continuous random variables (RVs), normalized histogram:

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

discrete value

probability

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



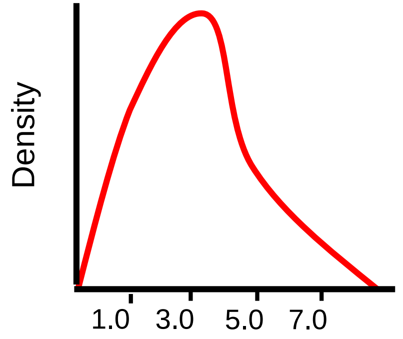
Schematic of continuous normalized histogram.

For continuous RVs, PDF:

value

pdf

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



Schematic of continuous PDF.

Recall: $f(x)$ is the probability density function of feature X, and $\int_{-\infty}^{+\infty} f_x(x) dx = 1$.

Statistical Expectation and Average

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:

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

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

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

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

Statistical Expectation Example

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 Example

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 mm**

Statistical Expectation Operators

Expectation Operators:

Expectation of a constant $\longrightarrow$ $E[c] = c$

Distributive Property

Expectation of a random variable + a constant $\longrightarrow$ $E[X+c] = E[X] + E[c] = E[X] + c$

Expectation of a product of a random variable and a constant $\longrightarrow$ $E[cX] = cE[X]$

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

Expectation of addition of two random variables $\longrightarrow$ $E[X+Y] = E[X] + E[Y]$

Expectation of the product of two random variables (if independent) $\longrightarrow$ $E[XY] = E[X]E[Y], if\ X \perp\!\!\!\perp Y$

X and Y are independent

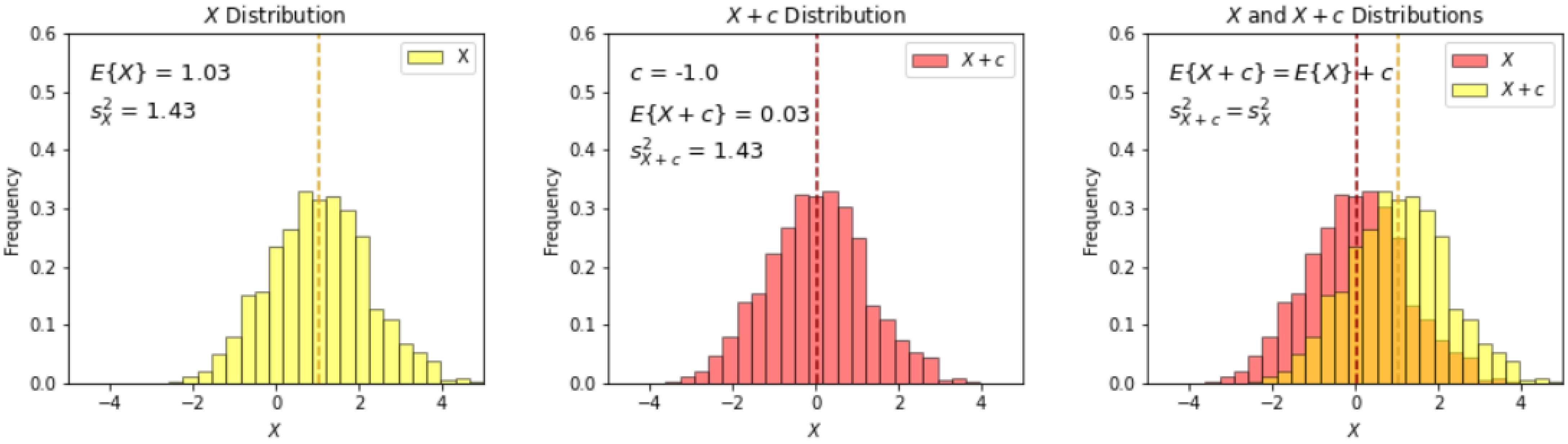
Statistical Expectation Operators

Expectation Operators:

Distributive Property

Expectation of a random variable + a constant → $E[X+c] = E[X] + E[c] = E[X] + c$

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



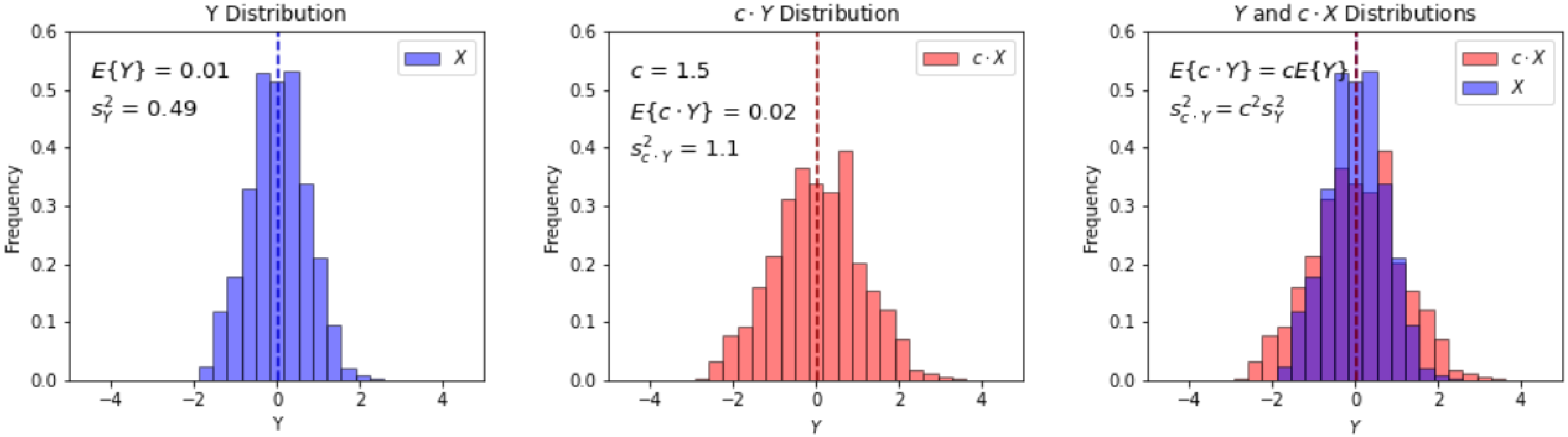
Demonstration of random variable, X, + constant, -1.0, file is PythonNumericalDemos_Expectation.ipynb.

Statistical Expectation Operators

Expectation Operators:

Expectation of a product of a random variable and a constant $\longrightarrow$ $E[cX] = cE[X]$

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



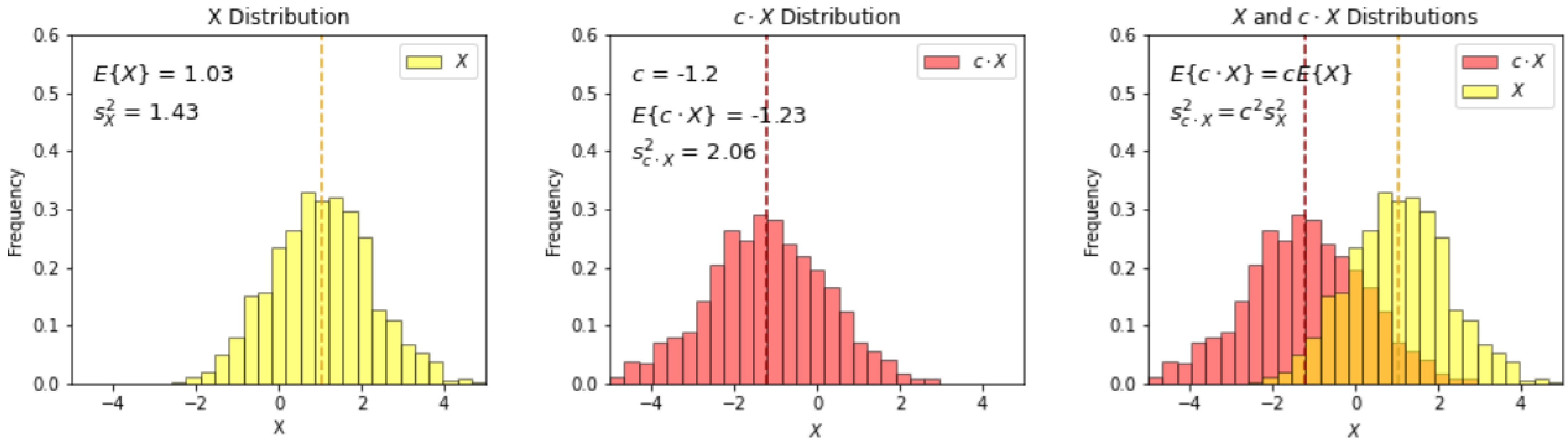
Demonstration of random variable, Y, times a constant, 1.5, file is PythonNumericalDemos_Expectation.ipynb.

Statistical Expectation Operators

Expectation Operators:

Expectation of a product of a random variable and a constant ⟶ $E[cX] = cE[X]$

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



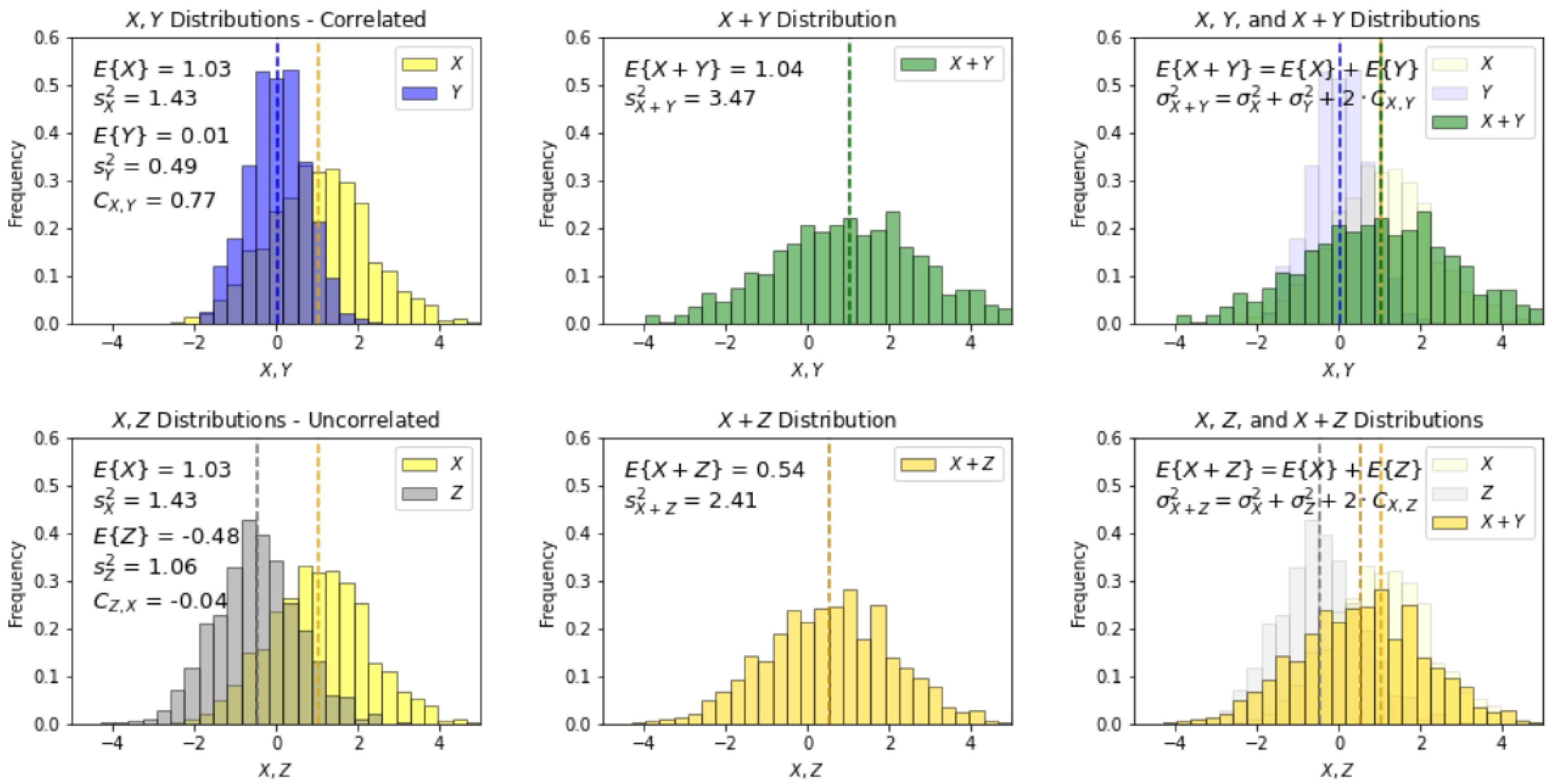
Demonstration of random variable, X, times a constant, 1.5, file is PythonNumericalDemos_Expectation.ipynb.

Statistical Expectation Operators

Expectation Operators:

Expectation of addition of two random variables → $E[X+Y] = E[X] + E[Y]$

- Here's 2 examples of adding random variables.
- X and Y are correlated, and X and Z are independent.
- Note the additivity of statistical expectation is general and still hold for correlated random variables.
- We will cover additivity of variance in Topic 12 kriging.



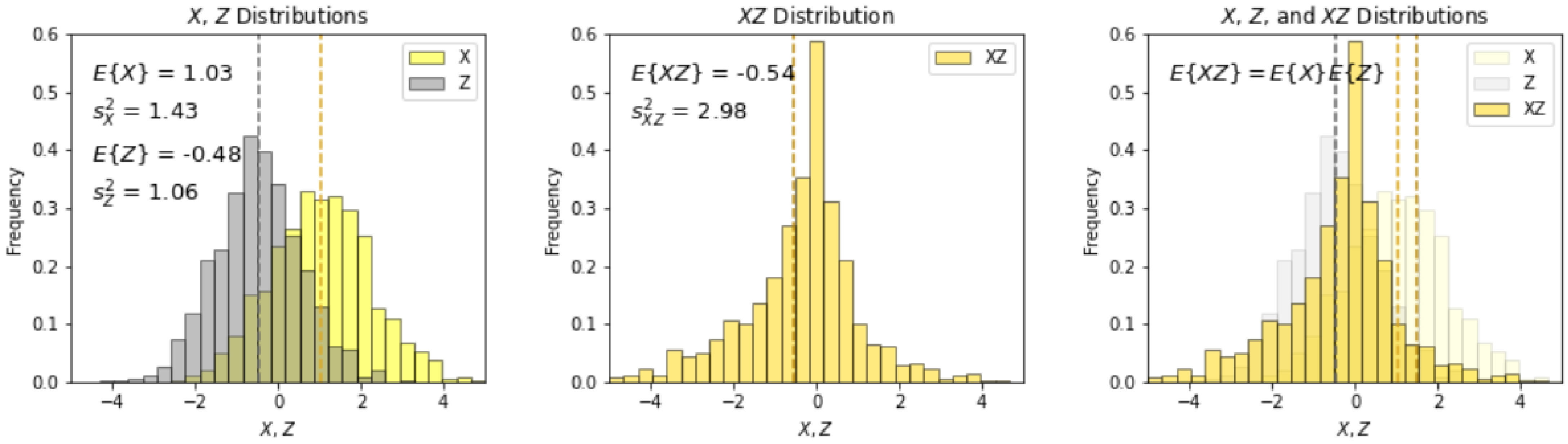
Demonstration of addition of random variables, $X + Y$ (above), and $X + Z$ (below), file is PythonNumericalDemos_Expectation.ipynb.

Statistical Expectation Operators

Expectation Operators:

Expectation of the product of two random variables (if independent) ⟶ $E[XY] = E[X]E[Y], \quad if\ X \perp\!\!\!\perp Y$

- Here's an example of the product of two random variables. Given the random variables are independent we can predict the expectation of the product. Note the shape change!



Demonstration of addition of random variables, XZ, file is PythonNumericalDemos_Expectation.ipynb.

Statistical Expectation Example

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 (φ_t) and effective porosity (φ_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 Example

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 (φ_t) and effective porosity (φ_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.

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

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

Statistical Expectation Example

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 (φ_t) and effective porosity (φ_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 Example

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 Example

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 .

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

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

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

Statistical Expectation Example

Data Analytics and Geostatistics

Some practice with statistical expectation:

Show that $Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$

given $E[E[X]] = E[X]$, $E[cX] = cE[X]$, $E[X^2 + 2X] = E[X^2] + 2E[X]$

Statistical Expectation Example

Data Analytics and Geostatistics

Some practice with statistical expectation:

Show that $Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$

given $E[E[X]] = E[X]$, $E[cX] = cE[X]$, $E[X^2 + 2X] = E[X^2] + 2E[X]$

expectation of a constant is a constant!

constant = mean

constant = mean^2

$$\underbrace{\boldsymbol{E}\left[(\boldsymbol{X} - \boldsymbol{E}[\boldsymbol{X}])^2\right]}_{\text{expand the quadratic}} = \boldsymbol{E}\overbrace{\left[\boldsymbol{X}^2 - 2\boldsymbol{X}\boldsymbol{E}[\boldsymbol{X}] + (\boldsymbol{E}[\boldsymbol{X}])^2\right]}^{\text{distributive property}} = \boldsymbol{E}\left[\boldsymbol{X}^2\right] - 2\boldsymbol{E}\left[\boldsymbol{X}\boldsymbol{E}[\boldsymbol{X}]\right] + \boldsymbol{E}\left[(\mathbf{E}[\mathbf{X}])^2\right]$$

expectation of an expectation is an expectation!

Statistical Expectation Example

Data Analytics and Geostatistics

Some practice with statistical expectation:

Show that $Var(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2$

given $E[E[X]] = E[X]$, $E[cX] = cE[X]$, $E[X^2 + 2X] = E[X^2] + 2E[X]$

$$E[(X - E[X])^2] = E[X^2 - 2XE[X] + (E[X])^2] = E[X^2] - 2E\big[XE[X]\big] + E[\,(\mathrm{E}[\mathrm{X}])^2]$$

$$\boldsymbol{E\left[X^2\right] - 2\underbrace{E[X]E[X]}_{\text{mean x mean = mean}^2} + (\mathrm{E}[\mathrm{X}])^2 = E\left[X^2\right] - \underbrace{2(\mathrm{E}[\mathrm{X}])^2 + (\mathrm{E}[\mathrm{X}])^2}_{\text{combine like terms}}}$$

Statistical Expectation Example

Data Analytics and Geostatistics

Some practice with statistical expectation, given random variable, X:

Show that,

$$\boldsymbol{Var(X) = E\left[(X - E[X])^2\right] = E\left[X^2\right] - (E[X])^2}$$

recall, $E[E[X]] = E[X]$, $\quad E[cX] = cE[X]$, $\quad E[X^2 + 2X] = E[X^2] + 2E[X]$

expectation of a constant is a constant!

constant = mean $\qquad$ constant = mean²

$$\underbrace{E[(X - E[X])^2]}_{\text{expand the quadratic}} = \overbrace{E[X^2 - 2XE[X] + (E[X])^2]}^{\text{distributive property}} = E[X^2] - 2E\left[XE[X]\right] + E[\mathrm{E}[\mathrm{X}])^2]$$

expectation of an expectation is an expectation!

$$E[X^2] - \underbrace{2E[X]E[X]}_{\text{mean x mean = mean}^2} + (\mathrm{E}[\mathrm{X}])^2 = E[X^2] - \underbrace{2(\mathrm{E}[\mathrm{X}])^2 + (\mathrm{E}[\mathrm{X}])^2}_{\text{combine like terms}}$$

$$\boldsymbol{Var(X) = E\left[X^2\right] - (\mathrm{E}[\mathrm{X}])^2}$$

Very convenient as we have can calculate the variance without knowing the mean! - one pass over the data to calculate the expected square and the mean at the same time.

Statistical Expectation Summary

Why is it important to understand expectation:

- Expectation is widely used for decision making, e.g., maximize project expected NPV
- Provides powerful methods work with expectation-based problems
  - e.g., expected value of resource in place over the aggregation of subsurface units
- Many theoretical developments in geostatistics are based on expectation
  - e.g., derivation of the kriging system

PGE 338 Data Analytics and Geostatistics

Lecture 4: Univariate Summaries

Lecture outline . . .

- Measures of Centrality
- Measures of Dispersion
- Measures of Shape
- Statistical Expectation

Michael Pyrcz, The University of Texas at Austin