

# Data Analytics and Geostatistics: Sparse Data



Lecture outline . . .

- Confidence Intervals
- Hypothesis Testing
- Distribution Transformations

Introduction

Modeling Prerequisites

Fundamental Concepts

Probability

Sparse Data Workflows

Spatial Estimation

Spatial Uncertainty

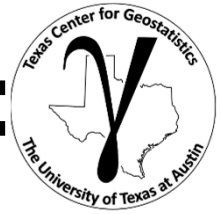
Multivariate, Spatial

Novel Workflows

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin

# Data Analytics and Geostatistics: Sparse Data



## Other Resources:

- Lectures recorded on YouTube.

**Confidence Intervals**

- How does this get used?
  - The language of
  - Decision makers
    - Significance
    - Uncertainty
- A lot of scientists and engineers
  - Add this to your knowledge of the impact of your work

**Hypothesis Testing / Confidence Statistical Hypothesis Testing**

Examples:

1. Collected data from a new well. Does this new data come from the same population as the previous wells or is there a geological discontinuity?
  - Test if the means are the same between the new data and the old data.  
 $H_0: \mu_n = \mu_p$ , the well data is from the same population as the previous wells.  
 $H_1: \mu_n \neq \mu_p$ , the well is from a different population than the previous wells.
2. Applied a new lab measurement for permeability from core data. Does the new method result in consistent amount of permeability variance?
  - Test if the variance is the same between the two methods..  
 $H_0: \sigma_{m1}^2 = \sigma_{m2}^2$ , the permeability variance of the two methods is the same.  
 $H_1: \sigma_{m1}^2 \neq \sigma_{m2}^2$ , the permeability variance of the two methods is different.

Confidence interval and hypothesis testing lectures on YouTube.

- Worked out examples on GitHub in Excel and Python

Introduction

Modeling Prerequisites

Fundamental Concepts

Probability

Sparse Data Workflows

Spatial Estimation

Spatial Uncertainty

Multivariate, Spatial

Novel Workflows

Conclusions

# General Comments on These Methods in a Spatial Context



- The confidence interval and hypothesis testing approaches have a variety of assumptions:
  1. i.i.d. – independent, identically distributed

In spatial context this is:

- no spatial correlation
  - no trends
2. There may also be distribution assumptions. e.g. Student's t-test for difference in means
    - Gaussian distributed means - ok with approximately Gaussian with small sample sizes
    - adequate sample size / representative sampling
    - equal or similar variance

# Data Analytics and Geostatistics: Sparse Data



Lecture outline . . .

- Confidence Intervals

Introduction

**Modeling Prerequisites**

Fundamental Concepts

Probability

**Sparse Data Workflows**

Spatial Estimation

Spatial Uncertainty

Multivariate, Spatial

Novel Workflows

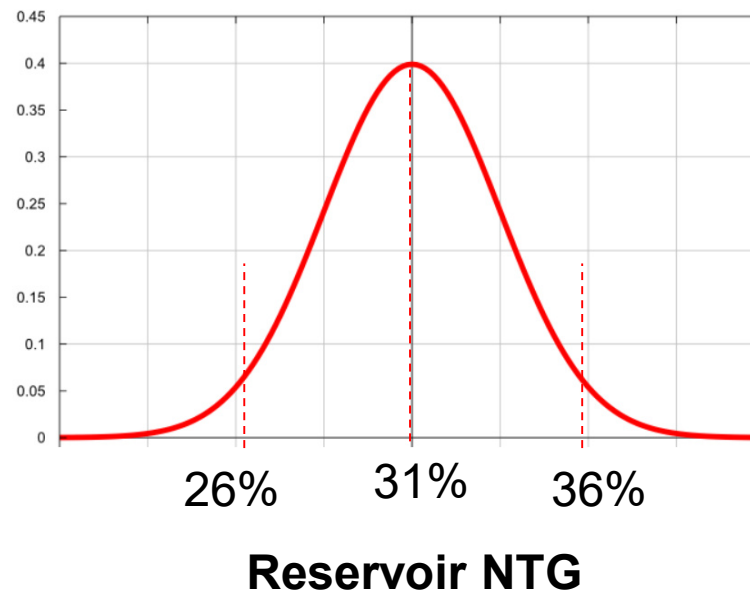
Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin

# Confidence Interval Definition



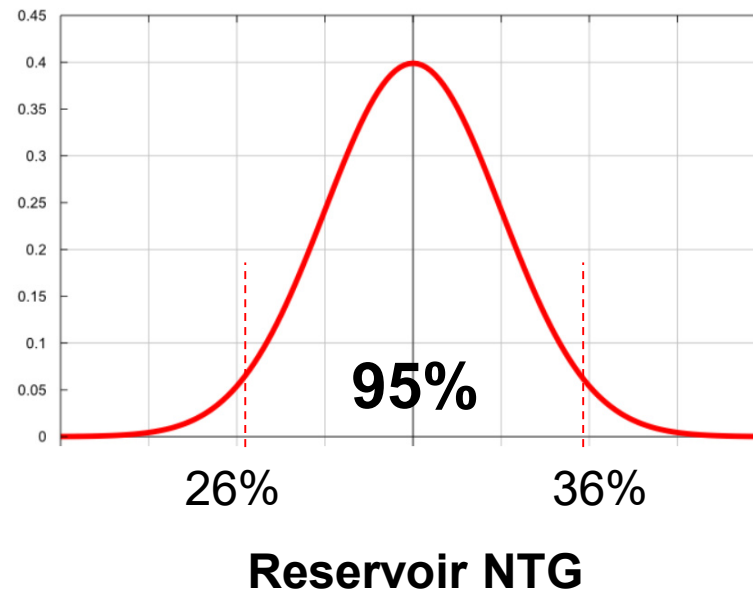
- The uncertainty in a summary statistic represented as a range, lower and upper bound, based on a specified probability interval known as the **confidence level**.
- We communicate confidence intervals like this:
- There is a 95% probability (or 19 times out of 20) that the true reservoir NTG is between 26% and 36%



# Confidence Interval Definition



- The probability of the population parameter being between the assessed lower and upper confidence bounds.
- **Alpha level = 1 – Confidence Level.** The probability the population parameter is outside the confidence interval. Alpha level is also known as significance level.



# Standard Error and Scores (e.g. z-score) Definition

- Take a confidence interval for a proportion:

$$\hat{p} \pm z_{\alpha} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

What is the **Standard Error** and the **z-Score**?

- the standard error, e.g.  $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ , is the uncertainty in the proportion statistics based on the sample traits and the number of samples, available for many statistics
- the -score (z-score if Gaussian),  $z_{\alpha}$ , is the  $F_x^{-1}(\alpha)$  from the standard (mean = 0, standard deviation = 1) of a theoretical sample distribution.
- the standard error uses information from the sample to rescale the standard theoretical sample distribution



# Confidence Intervals

## Degrees of Freedom

### Degree of Freedom (DOF):

1. Sampling - the number of independent pieces of information that go into estimating a parameter (Wikipedia).
2. Dynamic Systems – the number of independent way that a dynamic system can move.

Representation:  $\nu$  or  $d.f.$

Example for Estimating Sum of Squares (or variance):

$$x_1, x_2, \dots, x_n \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

DOF = sample size – constraints

for estimating the sum of squares: DOF =  $n-1$



# Confidence Interval Take 2



- Let's take another run at confidence intervals.
- Problem:
  - You estimated the mean porosity for a reservoir
    - Important because it impacts the OIP (value of the field)

**Average Porosity = 15%**

- What is the uncertainty in that estimate?

**Standard Error is the Uncertainty in an Estimate in Standard Deviations**

For uncertainty in a sample mean:  $SE_{\bar{x}} = \frac{s}{\sqrt{n}}$

as a function of number of samples and sample standard deviation.



# Confidence Interval Take 2

- Problem:
  - You estimated the mean porosity for a reservoir
    - Important because it relates to the OIP (value of the field)

**Average Porosity = 15%**

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

- Given 16 samples and sample standard deviation = 2

$$SE_{\bar{x}} = \frac{2}{\sqrt{16}} = \frac{1}{2}\%$$

- Uncertainty in Average Porosity

$$\text{Average Porosity} = 15\% \pm \frac{1}{2}\%, \quad 14.5\% - 15.5\% \text{ for 1 st. dev.}$$

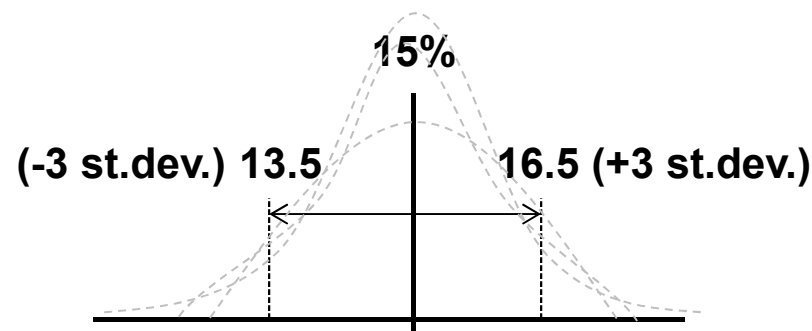
$$\text{Average Porosity} = 15\% \pm 1\frac{1}{2}\%, \quad 13.5\% - 16.5\% \text{ for 3 st. dev.}$$

# Confidence Interval Take 2

- Problem:
  - You estimated the mean porosity for a reservoir
    - Important because it relates to the OIP (value of the field)

***Average Porosity = 15%  $\pm 1\frac{1}{2}$ %, 13.5% – 16.5% for 3 st. dev.***

- Is this good enough? What is the probability of being in this range? We don't know!



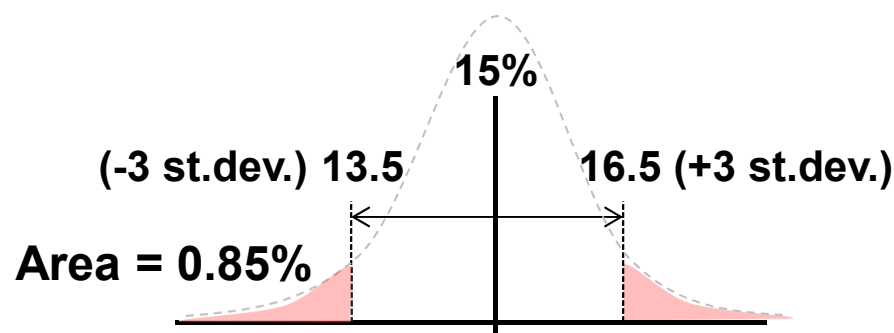
- To calculate that you need the shape! i.e. to know the entire distribution.

# Confidence Interval Take 2

- Problem:
  - You estimated the mean porosity for a reservoir
    - Important because it relates to the OIP (value of the field)

*Average Porosity = 15%  $\pm 1\frac{1}{2}$ %, 13.5% – 16.5% for 3 st. dev.*

- What is the distribution of means with small sample size?



➔ **Student's t**  
by theory we  
know the shape

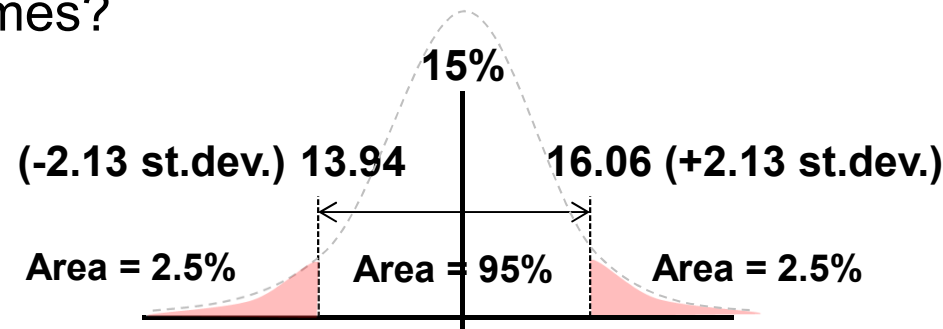
- Use Excel TDIST.2T(3,16) to calculate “prob in interval” = 99.6%
- That confidence interval is too wide, not helpful for decision making! Switch to a 95% significance level.

# Confidence Interval Take 2

- Problem:
  - You estimated the mean porosity for a reservoir
    - Important because it relates to the OIP (value of the field)

***Average Porosity = 15%  $\pm$  1  $\frac{1}{2}$ %, 13.5% – 16.5% for 3 st. dev.***

- How many standard deviations needed to cover 95% of outcomes?



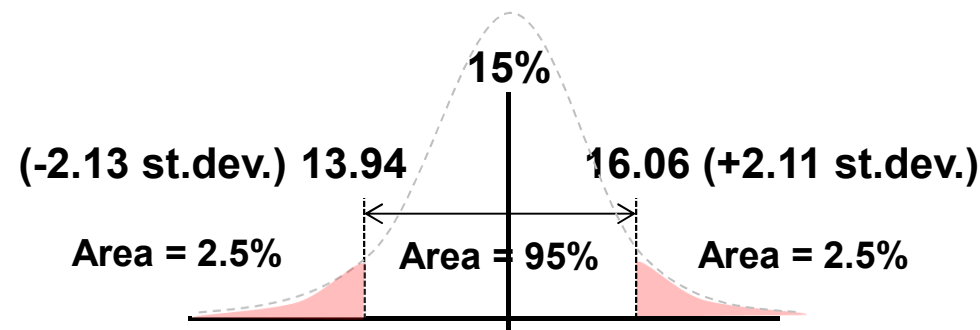
- Use Excel T.INV(0.025,16-1), T.INV(0.975,16-1) to calculate number of standard deviations interval at 95%. Now we have something precise to report:

# Confidence Interval Take 2

- Problem:
  - You estimated the mean porosity for a reservoir
    - Important because it relates to the OIP (value of the field)

*Average Porosity = 15%  $\pm$  2.11  $\left[\frac{1}{2}\right]$ , 15%  $\pm$  1.06, with 95% confidence*

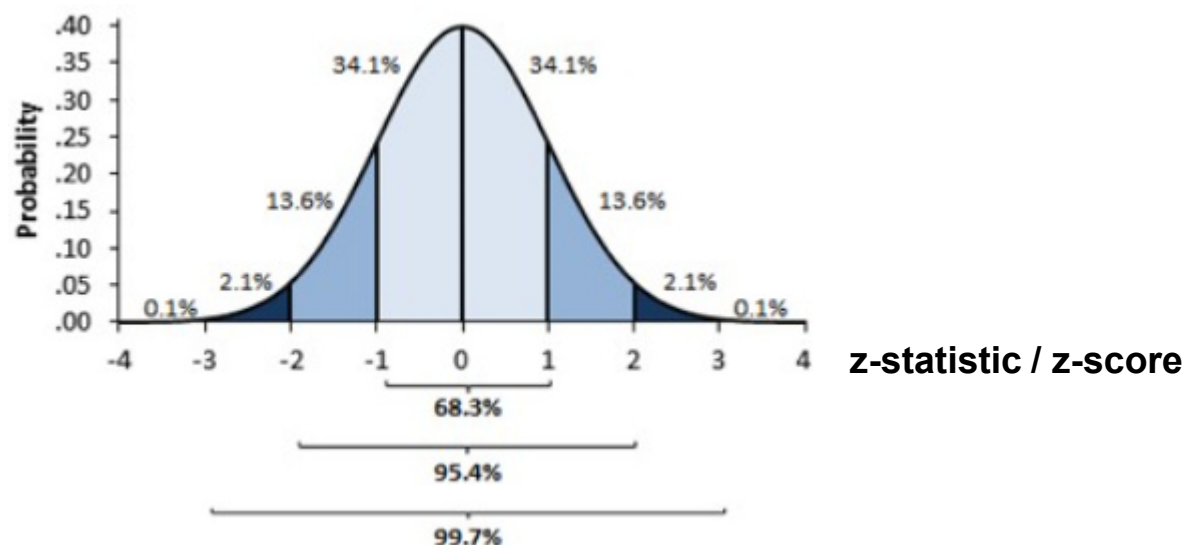
*Average Porosity = 15%  $\pm$  2.11  $\left[\frac{1}{2}\right]$ , 15%  $\pm$  1.06, 19 times out of 20.*



$$CI \text{ for } \mu = \bar{x} \pm t_{\frac{\alpha}{2}, n-1} \frac{s}{\sqrt{n}}$$

# Confidence Interval Take 2

- What did we learn?:
  - Standard error tells us the uncertainty / dispersion of a measure
  - BUT we still need the distribution shape. We know specific distributions occur for different processes. E.g.
    1. average / proportion – Gaussian, student-t if too few samples (<30 is commonly used)
    2. variance or average of squares – chi-square
  - we calculate the statistic / score from that standardized distribution (mean of 0.0, st.dev. of 1.0) multiplied by standard error.





# Confidence Interval

## Example #1

- The geologists has 200 representative samples of facies in your reservoir unit. What is the confidence interval for facies proportion?

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \approx z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- $\hat{p} = 0.7$  proportion of sandstone,  $n = 200$ ,  $\alpha$  level = 5% (confidence level = 95%)





# Confidence Interval

## Example #1

- The geologists has 200 representative samples of facies in your reservoir unit. What is the confidence interval for facies proportion?

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}} \approx z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- $\hat{p} = 0.7$  proportion of sandstone,  $n = 200$ ,  $\alpha$  level = 5% (confidence level = 95%)

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.70 \pm 1.96 \sqrt{\frac{0.7(1-0.7)}{200}}$$

- $\hat{p} = 0.70 \pm 0.063$ , [0.64, 0.76]

$z_{\alpha/2} \rightarrow \text{NORM.INV.S}(\alpha/2)$ in Excel or $\text{NORM.INV}(\alpha/2, 0, 1)$
------------------------------------------------------------------------------------------------------



# Confidence Interval

## Example #2

- The geologists has 10 representative samples of facies in your reservoir unit. What is the confidence interval for facies proportion?

$$\hat{p} \pm t_{\frac{\alpha}{2}, n-1} \sqrt{\frac{p(1-p)}{n}} \quad \text{given small sample}$$

- $\hat{p} = 0.7$  proportion of sandstone,  $n = 10$ ,  $\alpha$  level = 5% (confidence level = 95%)



# Confidence Interval

## Example #2

- The geologists has 10 representative samples of facies in your reservoir unit. What is the confidence interval for facies proportion?

$$\text{C.I. for } p = \hat{p} \pm t_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

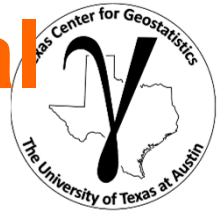
- $\hat{p} = 0.7$  proportion of sandstone,  $n = 10$ ,  $\alpha$  level = 5% (confidence level = 95%)

$$\text{C.I. for } p = \hat{p} \pm t_{\frac{\alpha}{2}, 9} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.70 \pm 2.26 \sqrt{\frac{0.7(1-0.7)}{200}}$$

- $\hat{p} = 0.70 \pm 0.328$ , [0.37, 1.03]
- $\hat{p} = 0.70 \pm 0.284$ , [0.42, 0.98] if we assume a Gaussian instead of student t distribution

$t_{\alpha/2} \rightarrow \text{T.INV.2T}(\alpha, n-1)$  or  $\text{T.INV}(\alpha/2, n-1)$  in Excel.

# Confidence Interval for Vertical Porosity Trend



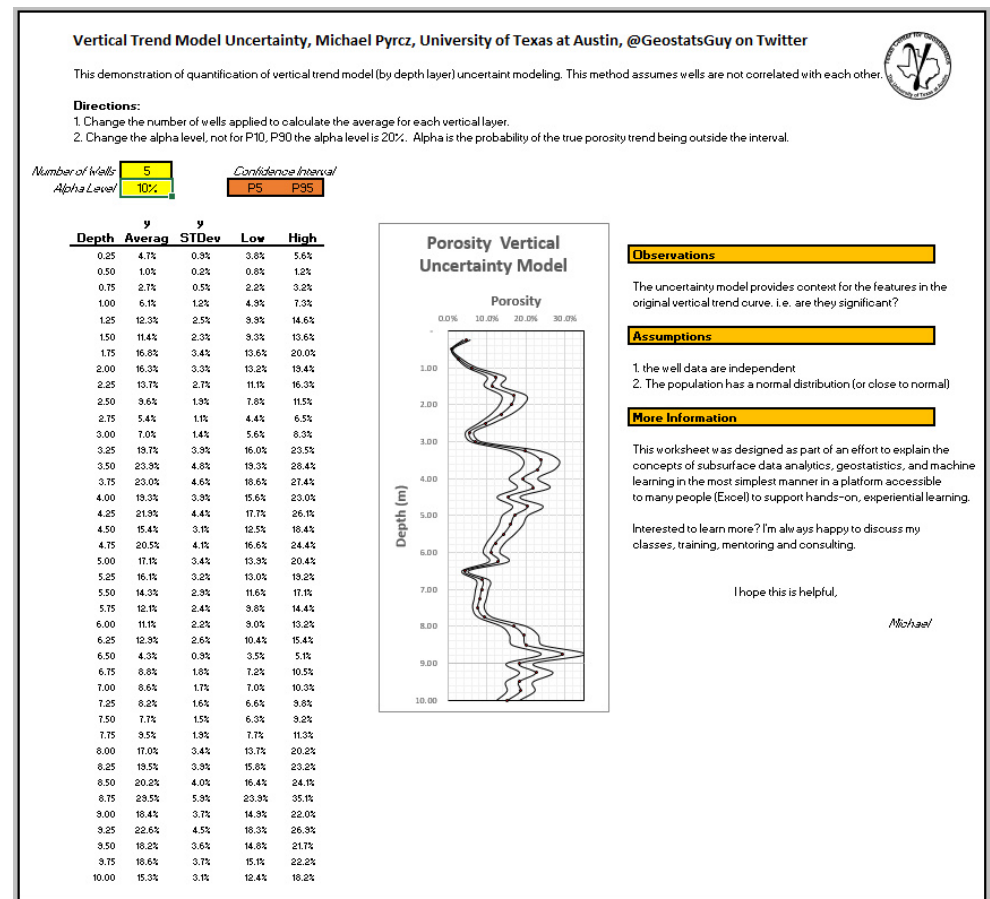
## Confidence Interval of a Trend Model:

### Things to try:

1. Increase the number of wells to 10, 20 etc.
2. Change the alpha level to 20% (P10, P90), and 5%.

Observe the impact on the uncertainty envelope.

What features in the trend are significant?



The file is at: <https://git.io/fhAHX>. The file is Vertical\_Trend\_Uncertainty\_Demo.xlsx.

# Data Analytics and Geostatistics: Sparse Data



Lecture outline . . .

- Hypothesis Testing

Introduction

Modeling Prerequisites

Fundamental Concepts

Probability

Sparse Data Workflows

Spatial Estimation

Spatial Uncertainty

Multivariate, Spatial

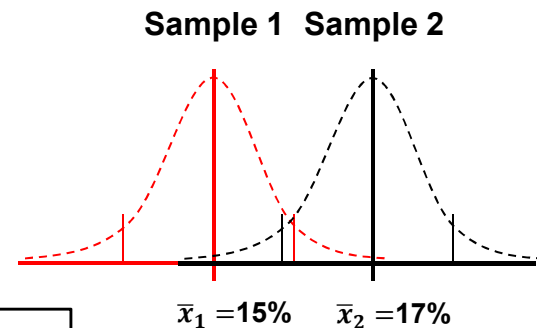
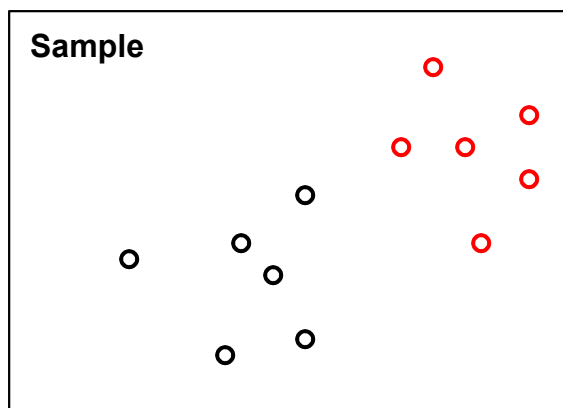
Novel Workflows

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin

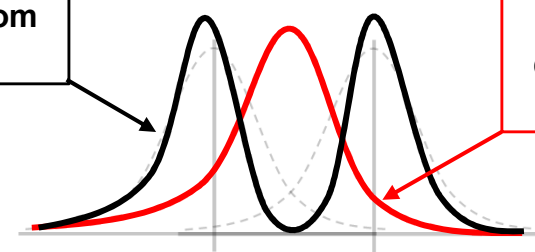
# Hypothesis Testing Take 2

- Problem:
  - You have 2 datasets (1 and 2), did they come from the same population?
  - If you had 2 datasets from the same population. They could look different!



all the same, the  
difference is random  
effect.

all the same, the  
difference is random  
effect.



- There is structure in random!



# Hypothesis Testing Take 2

- Problem:
  - Belief in the law of small numbers

*Psychological Bulletin*  
1971, Vol. 76, No. 2, 105-110

## BELIEF IN THE LAW OF SMALL NUMBERS

AMOS TVERSKY AND DANIEL KAHNEMAN<sup>1</sup>

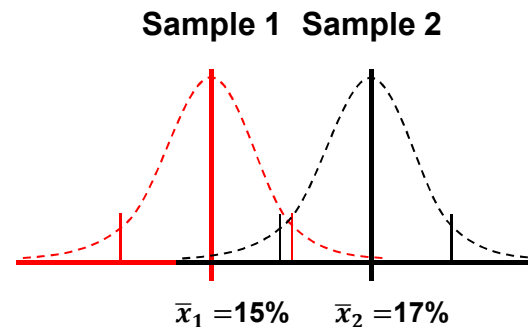
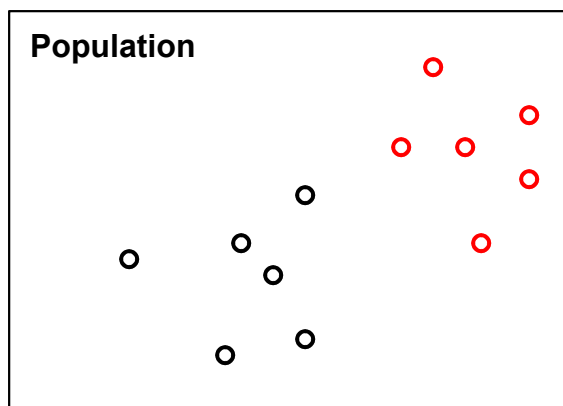
*Hebrew University of Jerusalem*

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

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

# Hypothesis Testing Take 2

- Solution:
  - What are you going to compare?
    - Mean, variance, binned proportion?
  - Set up the hypothesis test:  $\mu_1 = \mu_2$  they come from populations with the same mean; therefore, they could be the same population.





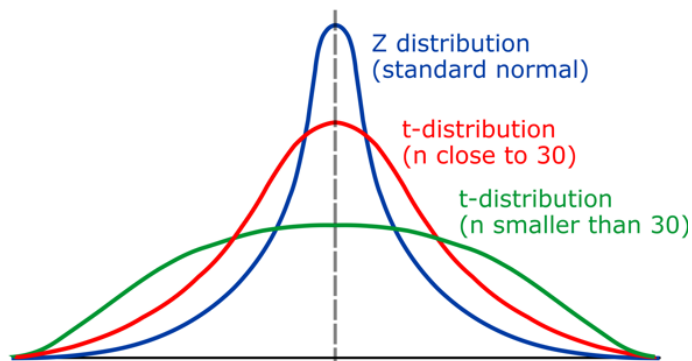
# Hypothesis Testing Take 2

- Solution:
  - Hypothesis test:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2,$$

- Decide on the metric?  $\bar{x}_1 - \bar{x}_2$
- What is the sampling distribution we would expect
  - Difference of 2 Gaussian random variables with small sample size and unknown  $\sigma$



student's t distribution



# Hypothesis Testing Take 2

- Solution:
  - Hypothesis test:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2,$$

- Decide on the metric?  $\bar{x}_1 - \bar{x}_2$
- Distribution – student's t

But how much difference is significant?

- Standard error tells us how much spread due to random, small sample and variability in samples.

$$\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)}$$

# Hypothesis Testing Take 2

- Solution:
  - Hypothesis test:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2,$$

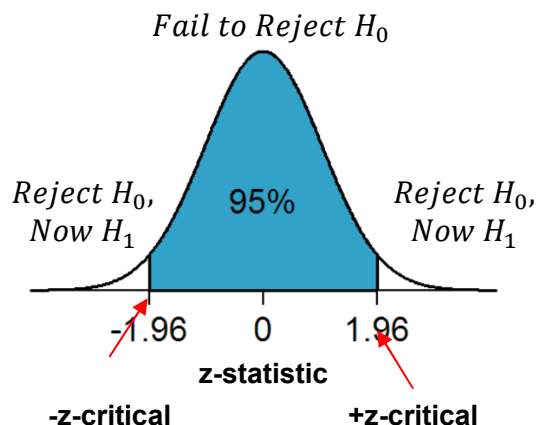
- Decide on the metric?  $\bar{x}_1 - \bar{x}_2$

- Distribution – student's t

- Expected difference from random?

$$\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)}$$

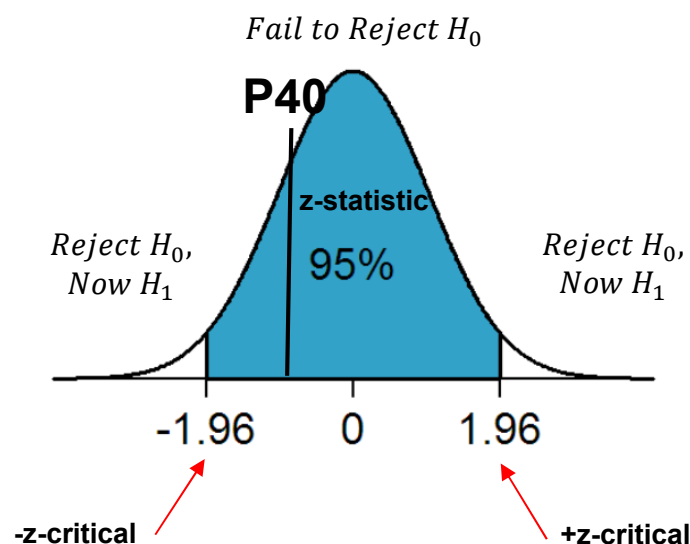
- Look up critical value from distribution to decide reject or fail to reject



this is how our metric should be distributed if both samples were sampled from distributions with the same mean.

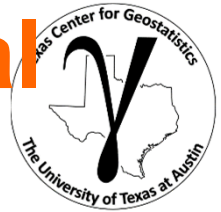
# Hypothesis Testing Take 2

- Solution:
  - What is the p-value?



- Indicates how close you were to rejecting / failing to reject.
- Could conduct the by comparing p-value to  $\frac{\alpha}{2}$  and  $1 - \frac{\alpha}{2}$ .

# Confidence Interval for Vertical Porosity Trend



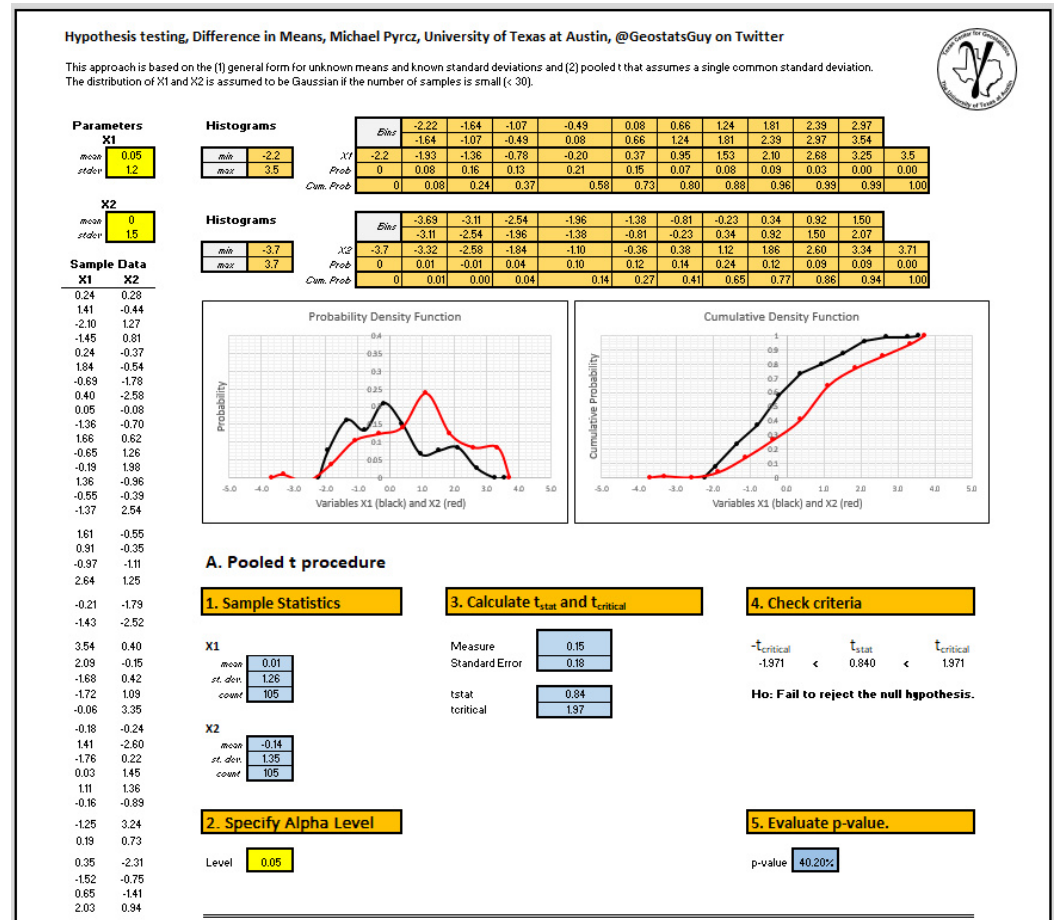
## Confidence Interval of a Trend Model:

### Things to try:

1. Change the alpha level lower and higher (e.g. 0.01 and 0.1).

2. Manually change the means of each group of wells.

Observe the impact on the outcome of the hypothesis test.



The file is at: <https://git.io/fNgBU>.

The file is Difference\_in\_Mean\_Demo.xlsx.



# Probability and Statistics

## What should you learn from this lecture?

- **Fundamentals of Statistics and Probability**
  - **Fundamentals of Probability**
    - » **Basic Definitions and Rules**
    - » **Venn Diagram**
    - » **Conditional Probability**
    - » **Probability tree**
    - » **Bayes' Theorem**
    - » **Applications of Probability in Decision Making**

# Data Analytics and Geostatistics: Sparse Data



Lecture outline . . .

- Distribution  
Transformations

Introduction

**Modeling Prerequisites**

Fundamental Concepts

Probability

**Sparse Data Workflows**

Spatial Estimation

Spatial Uncertainty

Multivariate, Spatial

Novel Workflows

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin

# Distribution Transforms



## Why Cover Distribution Transforms

- In exploration settings it is common to use distribution transforms to integrate analog information with sparse local information.
- Workflow:
  1. Formulate analog distribution from available analog information.
  2. Calculate local summary statistics from available well and seismic information.
  3. Transform the analog distribution to honor the local summary statistics.

Maximize value of analog information to understand the entire shape of the distribution.



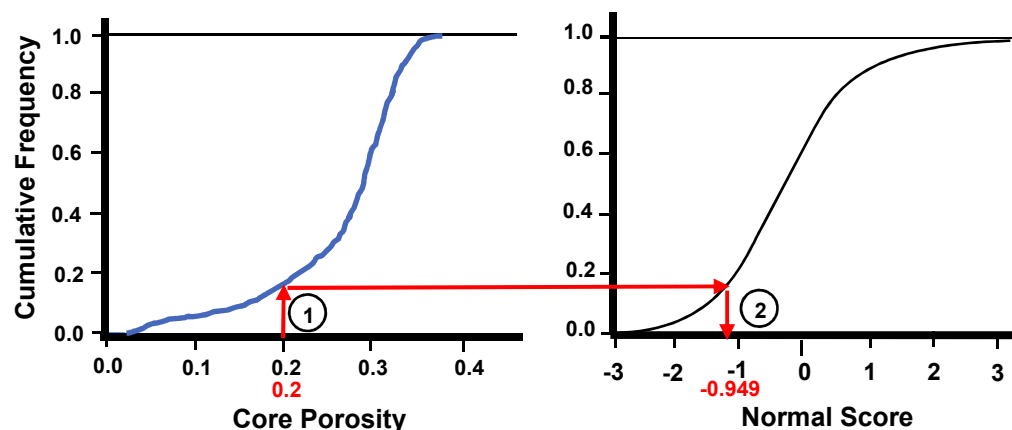
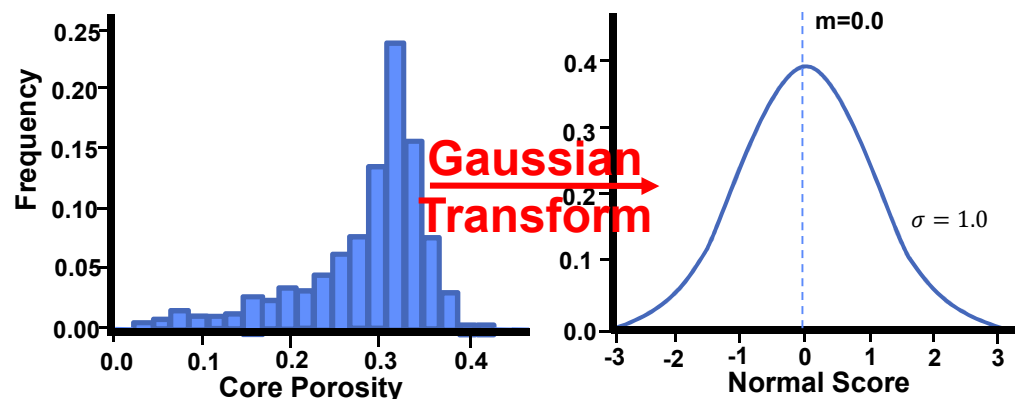
# Distribution Transforms

## Distribution Transforms

- Method may require a specific distribution assumption etc.
- Apply the following to all sample data.

$$Y = G_Y^{-1}(F_X(X))$$

- This may be applied to any parametric or nonparametric.
- Just need to be able to map from one distribution to another through percentiles.





# Distribution Transforms

## Distribution Transform Examples

- Well Log Porosity ( $WL\phi$ )[10, 13, 14, 15, 17]
- Core Porosity ( $Core\phi$ )[6, 9, 10, 13, 17]

How would you transform the Log Porosity to the Core Porosity Distribution?



# Distribution Transforms

## Distribution Transform Examples

- Well Log Porosity ( $WL\varphi$ ) [10, 13, 14, 15, 17]
- Core Porosity ( $Core\varphi$ ) [6, 9, 10, 13, 17]

How would you transform the Log Porosity to the Core Porosity Distribution?

We need to do this.

$$Y = G_Y^{-1}(F_X(X)) \quad \text{where } F \text{ is CDF of Well Log Porosity and } G \text{ is}$$

CDF of core

$$Y = G_{Core\varphi}^{-1}(F_{WL\varphi}(Log\varphi))$$

We need the CDF of both.

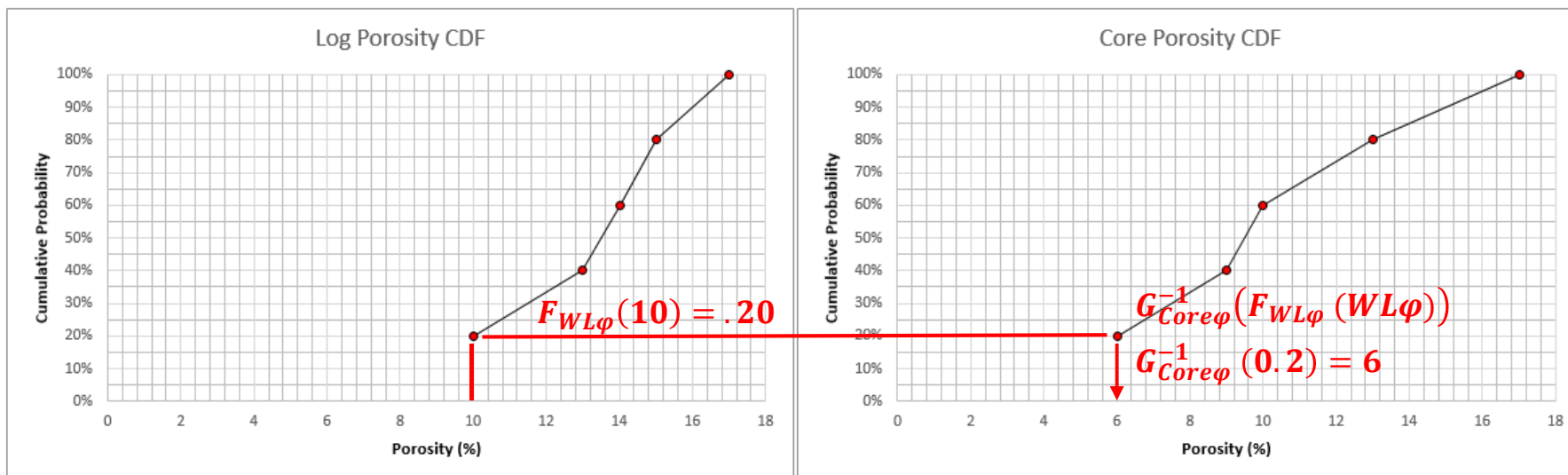
# Distribution Transforms

How would you transform the Log Porosity to the Core Porosity Distribution?

Index	Cum Prob.	Log	Core
1	20%	10	6
2	40%	13	9
3	60%	14	10
4	80%	15	13
5	100%	17	17

**Well Log  
Transferred to Core**

Transformed
6
9
10
13
17



If you have the same number of data, sort the data, data have the same cumulative probabilities so you can

# Distribution Transforms

How would you transform Core Porosity to  $N[0,1]$  Distribution (standard normal with mean = 0.0, standard deviation = 1.0)?

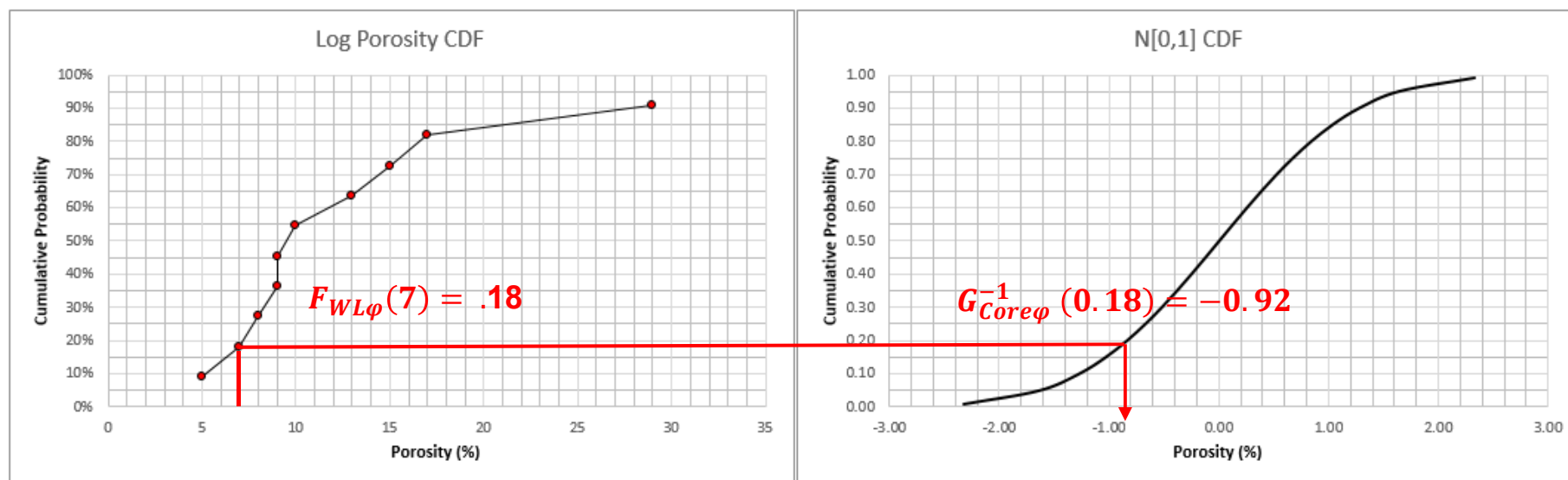
Core Porosity		Sample	Core Porosity	F(Core $\phi$ )
5	→	1	5	9%
7		2	7	18%
8		3	8	27%
9		4	9	36%
9		5	9	45%
10		6	10	55%
13		7	13	64%
15		8	15	73%
17		9	17	82%
29		10	29	91%

Use the  $N+1$  basis so that we assume tails are not known (recall the Gaussian Distribution is unbounded). We don't want to assign have 0 or 1 as a cumulative probability.

cumulative probability ( $F^{-1}(\text{Core}\phi)$ ) = sample index / ( $N+1$ )

# Distribution Transforms

How would you transform Core Porosity to  $N[0,1]$  Distribution (standard normal with mean = 0.0, standard deviation = 1.0)?



Take the cumulative probability for each value in your dataset and apply

`NORM.INV(<Cumulative Probability>, 0.0,1.0)`

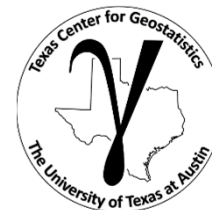
# Distribution Transforms

How would you transform Core Porosity to  $N[0,1]$  Distribution (standard normal with mean = 0.0, standard deviation = 1.0)?

Sample	Core Porosity	F(Core $\phi$ )		N[Core Porosity]
1	5	9%	$G_N^{-1}(F_\phi(\phi))$ <b>Norm.Inv</b>	-1.34
2	7	18%		-0.91
3	8	27%		-0.60
4	9	36%		-0.35
5	9	45%		-0.11
6	10	55%		0.11
7	13	64%		0.35
8	15	73%		0.60
9	17	82%		0.91
10	29	91%		1.34

Take the cumulative probability for each value in your dataset and apply

**NORM.INV**(<Cumulative Probability>, 0.0,1.0)



# Distribution Transforms

Now you try. Transform these values to  $N[10,4]$ .

3, 5, 8, 9, 11, 14, 19% Vsh

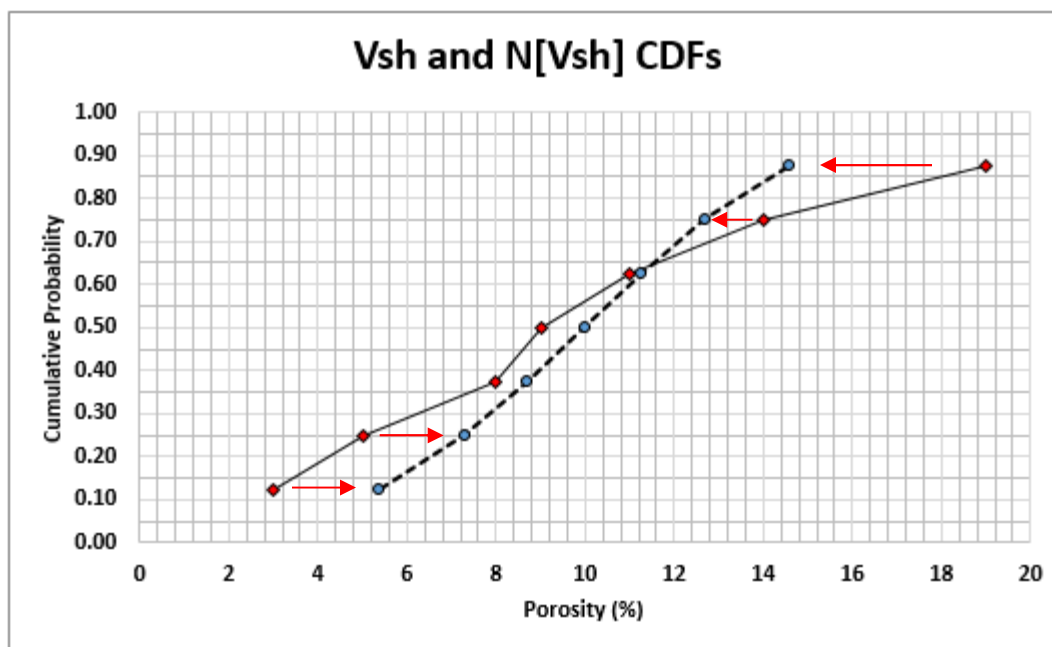


# Distribution Transforms

Now you try. Transform these values to  $N[10,4]$ .

3, 5, 8, 9, 11, 14, 19% Vsh

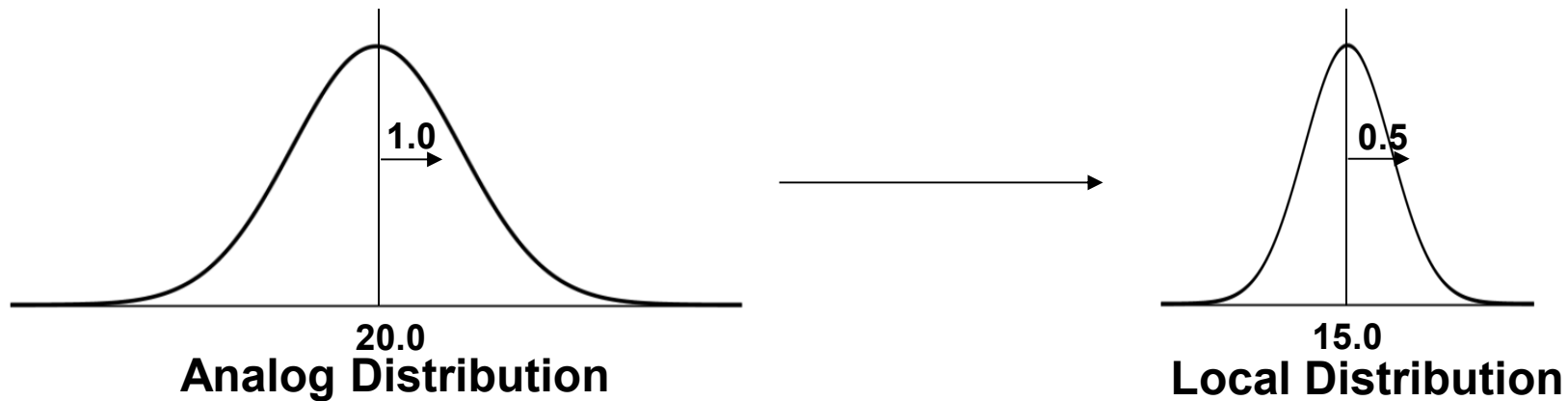
Index	Vsh	Cumul. Prob.	N[Vsh]
1	3	0.13	5.40
2	5	0.25	7.30
3	8	0.38	8.73
4	9	0.50	10.00
5	11	0.63	11.27
6	14	0.75	12.70
7	19	0.88	14.60



# Affine Correction



- Affine Correction, the no shape change distribution recaling:



Affine Correction – to scale values to change mean and standard deviation (shape stays the same):

$$y_{final} = \left( \frac{\sigma_{target}}{\sigma_{original}} \right) (y_{initial} - \bar{y}_{initial}) + \bar{y}_{target}$$

# Data Analytics and Geostatistics: Sparse Data



Lecture outline . . .

- Confidence Intervals
- Hypothesis Testing
- Distribution Transformations

Introduction

Modeling Prerequisites

Fundamental Concepts

Probability

Sparse Data Workflows

Spatial Estimation

Spatial Uncertainty

Multivariate, Spatial

Novel Workflows

Conclusions

Instructor: Michael Pyrcz, the University of Texas at Austin