Data Analytics and Geostatistics: Sparse Data

Lecture outline . . .

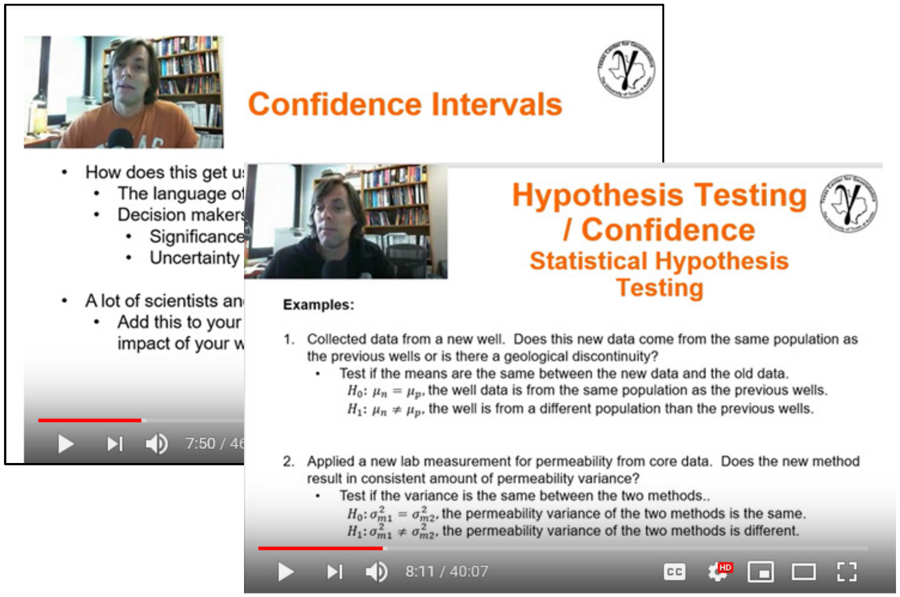
- Confidence Intervals
- Hypothesis Testing
- Distribution Transformations

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 u
  - The language of
  - Decision maker
    - Significance
    - Uncertainty
- A lot of scientists an
  - Add this to your impact of your w

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: $\sigma^2_{m1} = \sigma^2_{m2}$, the permeability variance of the two methods is the same.
     H_1: $\sigma^2_{m1} \neq \sigma^2_{m2}$, 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**

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 . . .

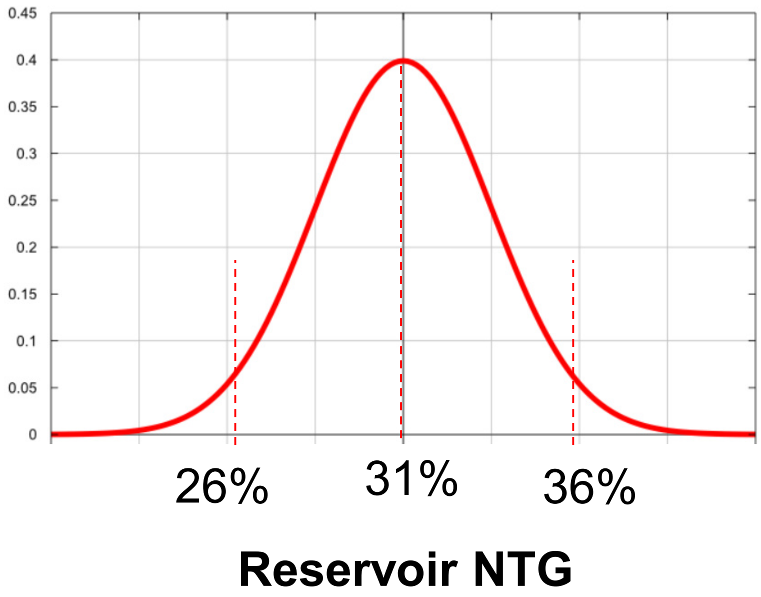
- Confidence Intervals

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**.

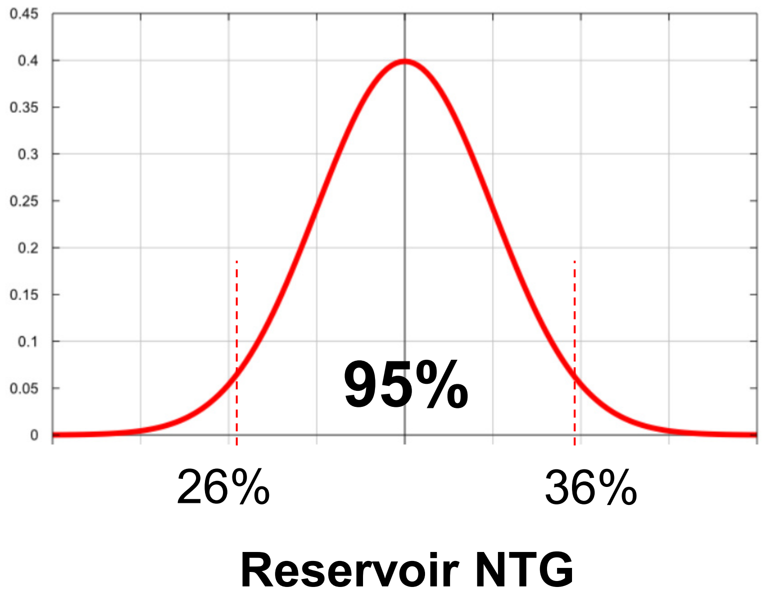
- We communicate confidence intervals like this:
- There is a 95% probability (or 19 times out of 20) that the true reservoir NTG is between 26% and 36%



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:

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

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

- the standard error, e.g. $\sqrt{\frac{\hat{\boldsymbol{p}}(\mathbf{1}-\hat{\boldsymbol{p}})}{\boldsymbol{n}}}$, 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), $\boldsymbol{z_\alpha}$, is the $F_x^{-1}(\boldsymbol{\alpha})$ 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: ν or $d.f$.

Example for Estimating Sum of Squares (or variance):

$$\boldsymbol{x_1, x_2, \ldots, x_n} \qquad\qquad \boldsymbol{s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

DOF = sample size – constraints

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

Figure from Dr. Zoya Heidari's PGE 337.

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

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

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

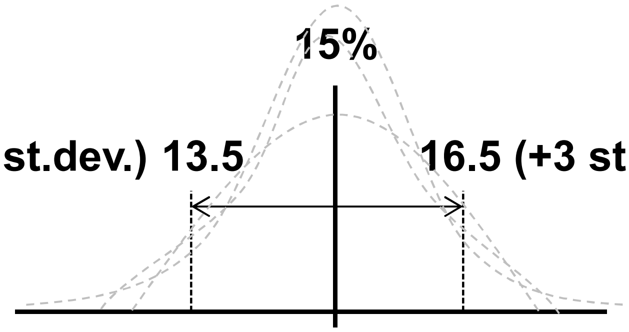
$$Average\ Porosity = 15\% \pm 1\frac{1}{2}\%, \qquad 13.5\% - 16.5\%\ for\ 3\ st.dev.$$

Confidence Interval Take 2

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

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

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



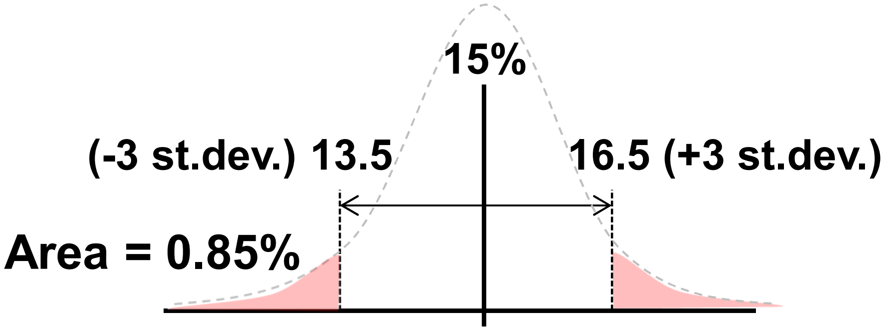
- To calculate that you need the shape! i.e. to know the entire distribution.

Confidence Interval
Take 2

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

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

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



Student’s t by theory we know the shape

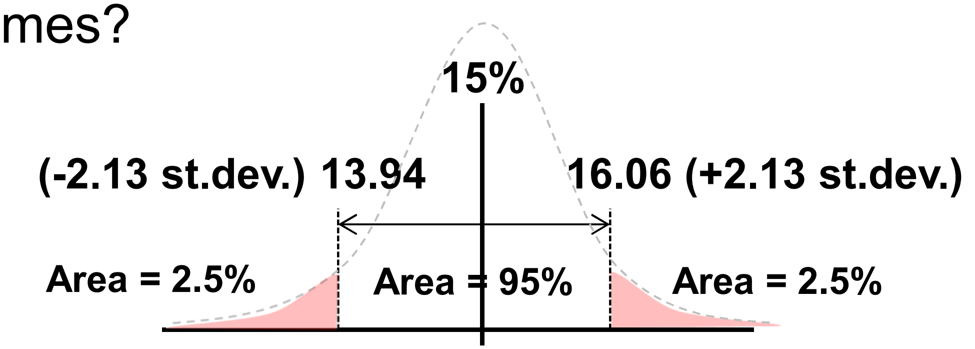
- Use Excel TDIST.2T(3,16) to calculate “prob in interval” = 99.6%
- That confidence interval is too wide, not helpful for decision making! Switch to a 95% significance level.

Confidence Interval Take 2

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

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

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



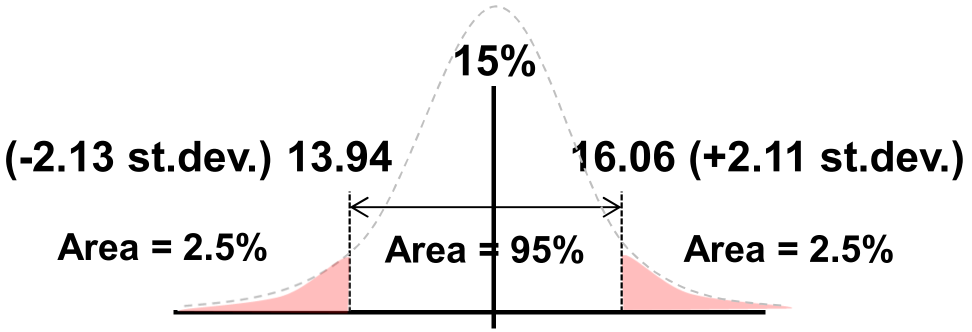
- Use Excel T.INV(0.025,16-1), T.INV(0.975,16-1) to calculate number of standard deviations interval at 95%. Now we have something precise to report:

Confidence Interval
Take 2

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

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

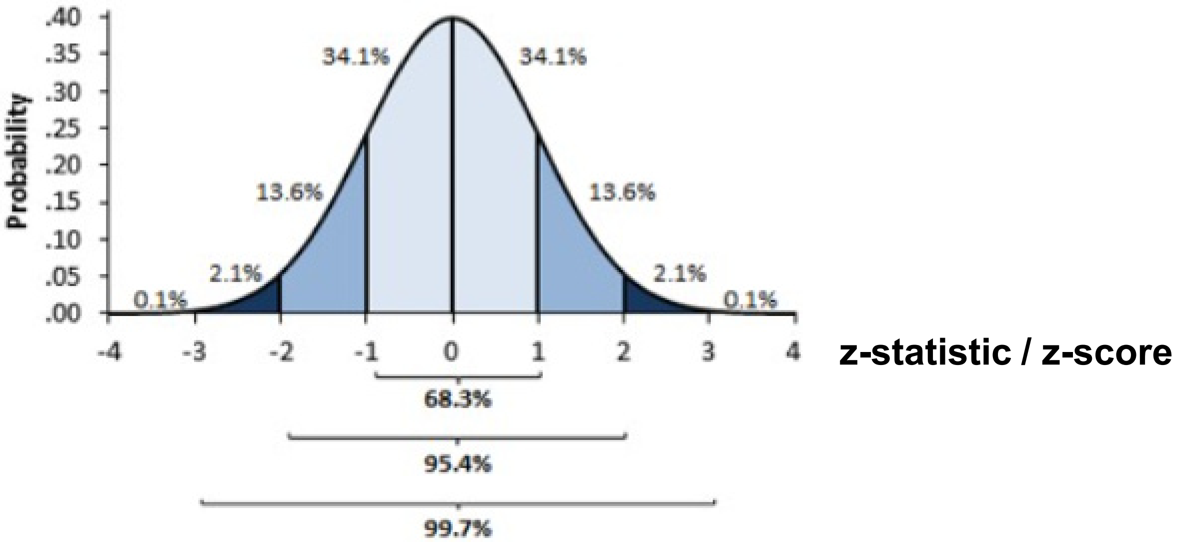
$$\boldsymbol{Average\ Porosity = 15\%\ \pm 2.11\left[\frac{1}{2}\right], 15\% \pm 1.06, 19\ times\ out\ of\ 20.}$$



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

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?

$$\widehat{\boldsymbol{p}} \pm \boldsymbol{z}_{\alpha/2}\sqrt{\frac{\boldsymbol{p}(\mathbf{1}-\boldsymbol{p})}{\boldsymbol{n}}} \approx \boldsymbol{z}_{\alpha/2}\sqrt{\frac{\widehat{\boldsymbol{p}}(\mathbf{1}-\widehat{\boldsymbol{p}})}{\boldsymbol{n}}}$$

- $\widehat{\boldsymbol{p}}$ = 0.7 proportion of sandstone, $\boldsymbol{n}$ = 200, $\boldsymbol{\alpha}$ 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, α 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]

$z_{\alpha/2}$ → **NORM.INV.S($\alpha/2$) in Excel or 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?

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

- $\boldsymbol{\hat{p}}$ = 0.7 proportion of sandstone, $\boldsymbol{n}$ = 10, $\boldsymbol{\alpha}$ 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?

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

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

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

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

$\boldsymbol{t_{\alpha/2}}$ **→T.INV.2T(α,n-1) or T.INV($\alpha/2$,n-1) in Excel.**

Confidence Interval for Vertical Porosity Trend

Confidence Interval of a Trend Model:

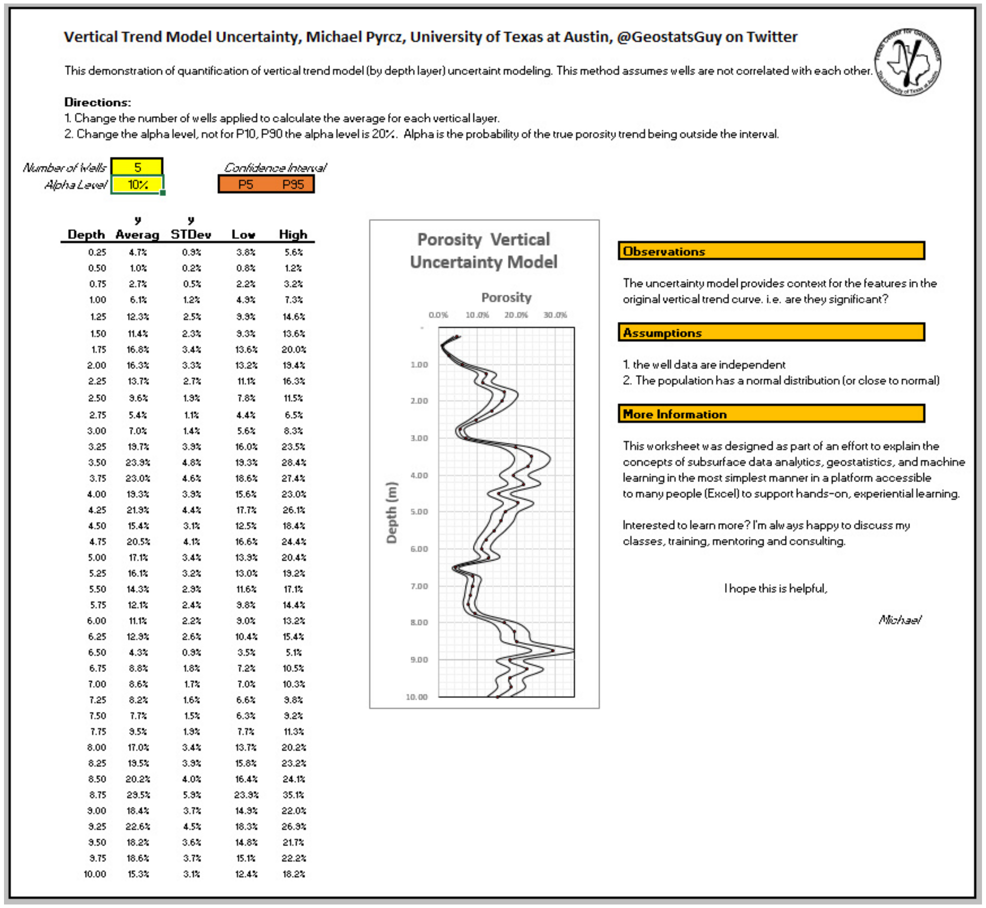
Things to try:

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?



Vertical Trend Model Uncertainty, Michael Pyrcz, University of Texas at Austin, @GeostatsGuy on Twitter

This demonstration of quantification of vertical trend model (by depth layer) uncertaint modeling. This method assumes wells are not correlated with each other.

Directions:
1. Change the number of wells applied to calculate the average for each vertical layer.
2. Change the alpha level, not for P10, P90 the alpha level is 20%. Alpha is the probability of the true porosity trend being outside the interval.

Number of Wells	5	*Confidence Interval*	
Alpha Level	10%	P5	P95

Depth	y Averag	y STDev	Low	High
0.25	4.7%	0.9%	3.8%	5.6%
0.50	1.0%	0.2%	0.8%	1.2%
0.75	2.7%	0.5%	2.2%	3.2%
1.00	6.1%	1.2%	4.9%	7.3%
1.25	12.3%	2.5%	9.9%	14.6%
1.50	11.4%	2.3%	9.3%	13.6%
1.75	16.8%	3.4%	13.6%	20.0%
2.00	16.3%	3.3%	13.2%	19.4%
2.25	13.7%	2.7%	11.1%	16.3%
2.50	9.6%	1.9%	7.8%	11.5%
2.75	5.4%	1.1%	4.4%	6.5%
3.00	7.0%	1.4%	5.6%	8.3%
3.25	19.7%	3.9%	16.0%	23.5%
3.50	23.9%	4.8%	19.3%	28.4%
3.75	23.0%	4.6%	18.6%	27.4%
4.00	19.3%	3.9%	15.6%	23.0%
4.25	21.9%	4.4%	17.7%	26.1%
4.50	15.4%	3.1%	12.5%	18.4%
4.75	20.5%	4.1%	16.6%	24.4%
5.00	17.1%	3.4%	13.9%	20.4%
5.25	16.1%	3.2%	13.0%	19.2%
5.50	14.3%	2.9%	11.6%	17.1%
5.75	12.1%	2.4%	9.8%	14.4%
6.00	11.1%	2.2%	9.0%	13.2%
6.25	12.9%	2.6%	10.4%	15.4%
6.50	4.3%	0.9%	3.5%	5.1%
6.75	8.8%	1.8%	7.2%	10.5%
7.00	8.6%	1.7%	7.0%	10.3%
7.25	8.2%	1.6%	6.6%	9.8%
7.50	7.7%	1.5%	6.3%	9.2%
7.75	9.5%	1.9%	7.7%	11.3%
8.00	17.0%	3.4%	13.7%	20.2%
8.25	19.5%	3.9%	15.8%	23.2%
8.50	20.2%	4.0%	16.4%	24.1%
8.75	29.5%	5.9%	23.9%	35.1%
9.00	18.4%	3.7%	14.9%	22.0%
9.25	22.6%	4.5%	18.3%	26.9%
9.50	18.2%	3.6%	14.8%	21.7%
9.75	18.6%	3.7%	15.1%	22.2%
10.00	15.3%	3.1%	12.4%	18.2%

Observations

The uncertainty model provides context for the features in the original vertical trend curve. i.e. are they significant?

Assumptions

1. the well data are independent
2. The population has a normal distribution (or close to normal)

More Information

This worksheet was designed as part of an effort to explain the concepts of subsurface data analytics, geostatistics, and machine learning in the most simplest manner in a platform accessible to many people (Excel) to support hands-on, experiential learning.

Interested to learn more? I'm always happy to discuss my classes, training, mentoring and consulting.

I hope this is helpful,

Michael

The file is at: https://git.io/fhAHX. The file is Vertical_Trend_Uncertainty_Demo.xlsx.

Data Analytics and Geostatistics:
Sparse Data

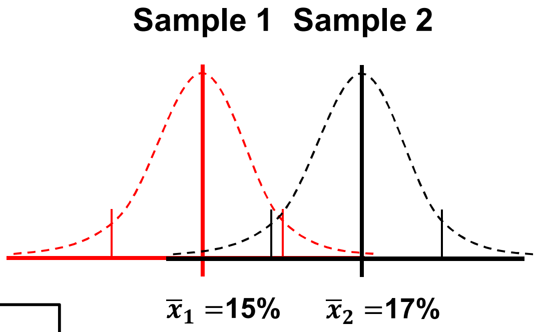
Lecture outline . . .

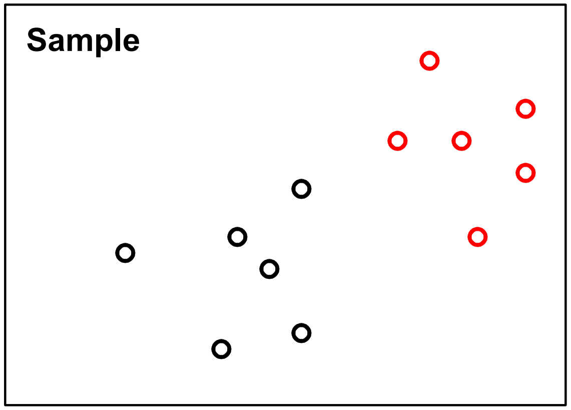
- Hypothesis Testing

Instructor: Michael Pyrcz, the University of Texas at Austin

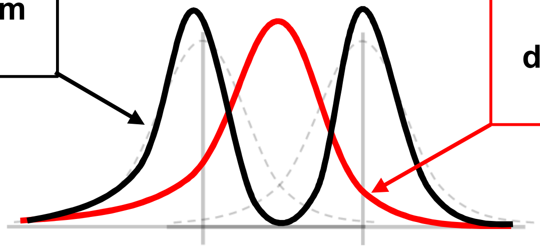
Hypothesis Testing
Take 2

- 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.

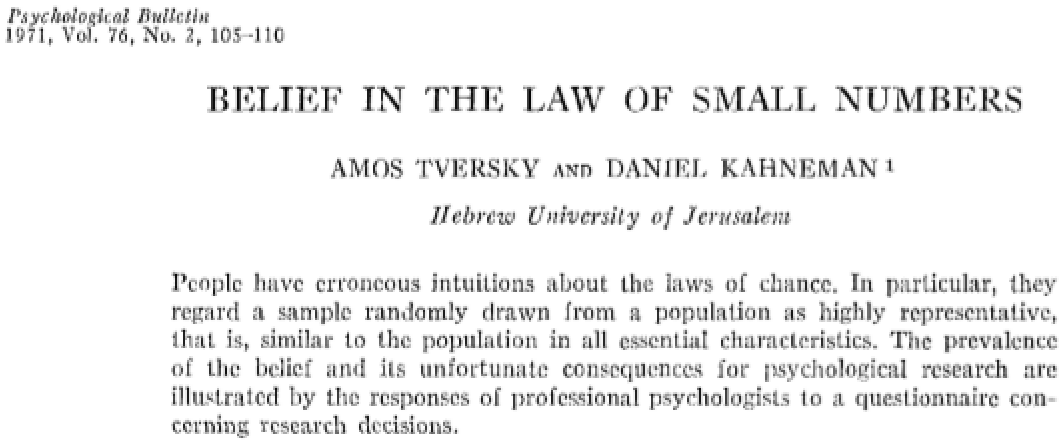


all the same, the difference is random effect.

  - 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 [1]

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.

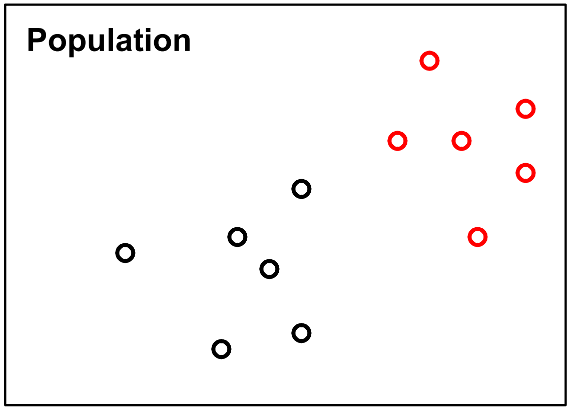
  - That samples randomly drawn from a population as highly representative.

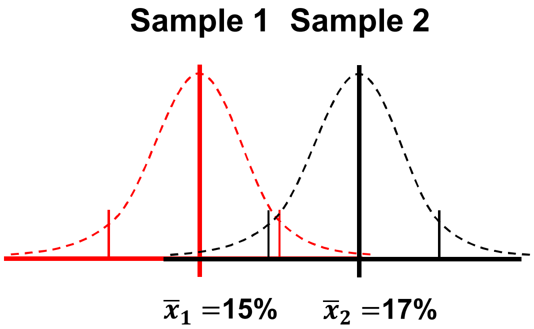
  - The mean, variance, P13 will be the same!

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.





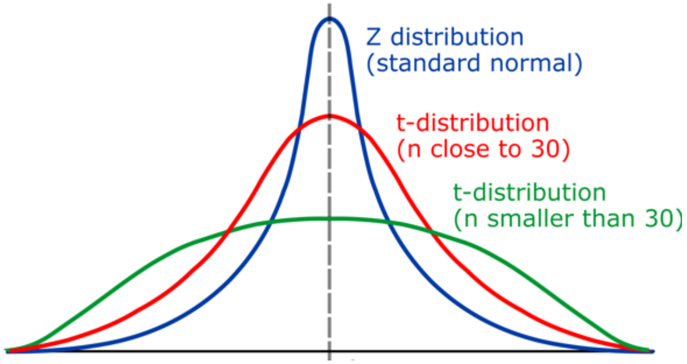
Hypothesis Testing Take 2

- Solution:
  - Hypothesis test:

$$H_0\colon \mu_1 = \mu_2$$
$$H_1\colon \mu_1 \neq \mu_2,$$

  - 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 σ



student's t distribution

Image from https://andyjconnelly.wordpress.com/2017/05/16/uncertainty-and-repeats/

Hypothesis Testing Take 2

- Solution:
  - Hypothesis test:

$$H_0\colon \mu_1 = \mu_2$$
$$H_1\colon \mu_1 \neq \mu_2,$$

  - 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.

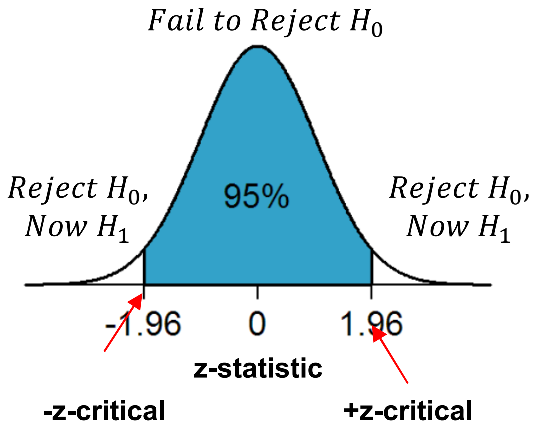
$$\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left(\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\right)}$$

Hypothesis Testing
Take 2

- Solution:
  - Hypothesis test:

$$H_0\colon \mu_1 = \mu_2$$
$$H_1\colon \mu_1 \neq \mu_2,$$

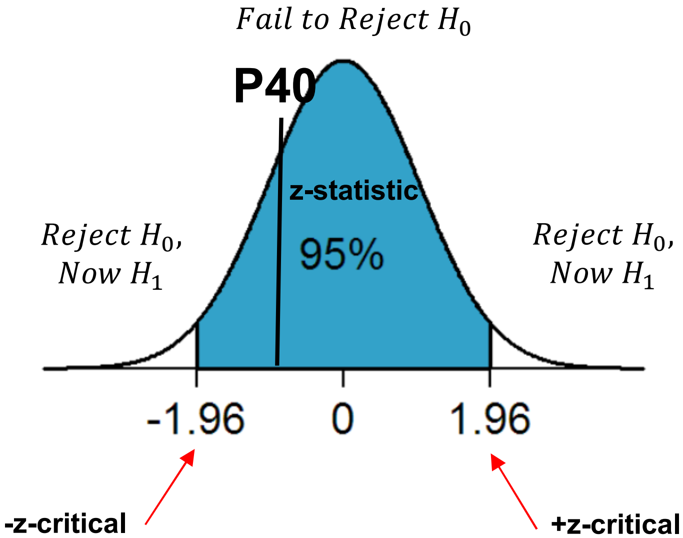
  - Decide on the metric? $\bar{x}_1 - \bar{x}_2$
  - Distribution – student's t
  - Expected difference from random? $\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)\left(\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}\right)}$
  - Look up critical value from distribution to decide reject or fail to reject



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

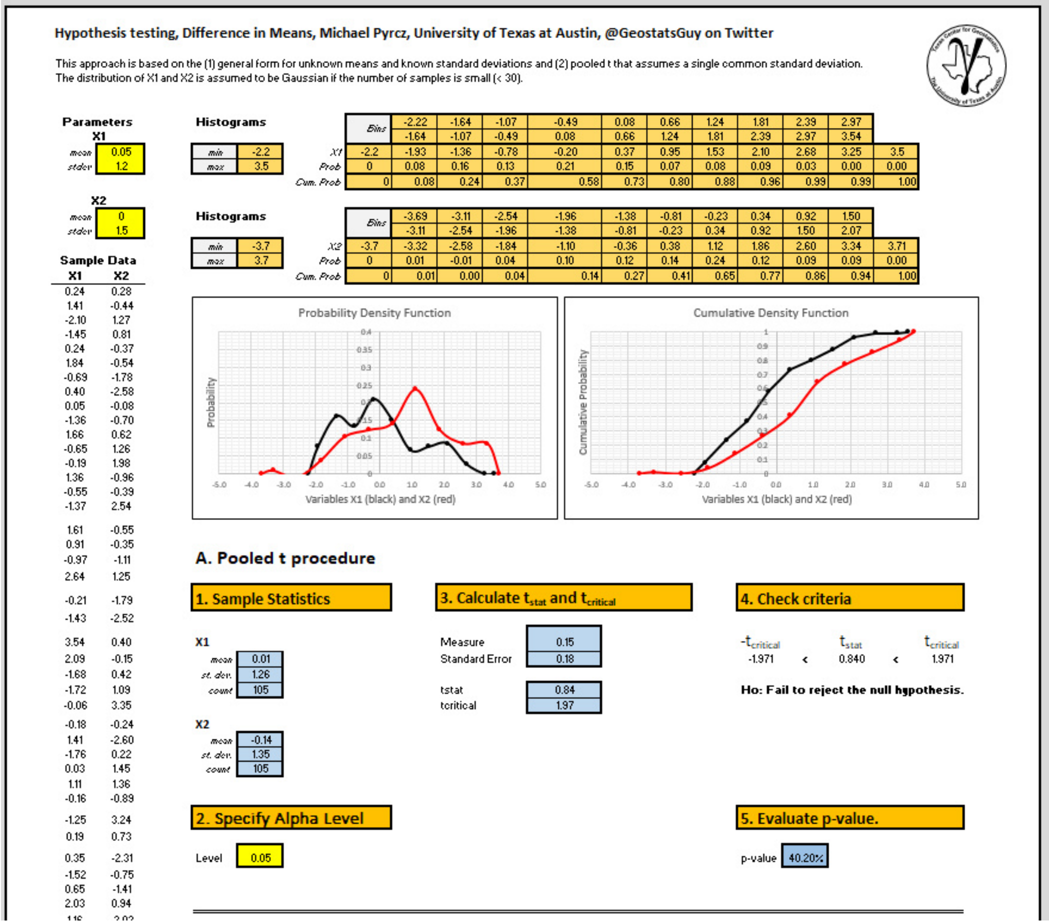
Confidence Interval of a Trend Model:

Things to try:

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

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

Observe the impact on the outcome of the hypothesis test.



Hypothesis testing, Difference in Means, Michael Pyrcz, University of Texas at Austin, @GeostatsGuy on Twitter

This approach is based on the (1) general form for unknown means and known standard deviations and (2) pooled t that assumes a single common standard deviation.
The distribution of X1 and X2 is assumed to be Gaussian if the number of samples is small (< 30).

Parameters

X1	
mean	0.05
stdev	1.2

X2	
mean	0
stdev	1.5

Histograms

min	-2.2
max	3.5

	Bins	-2.22	-1.64	-1.07	-0.49	0.08	0.66	1.24	1.81	2.39	2.97	
		-1.64	-1.07	-0.49	0.08	0.66	1.24	1.81	2.39	2.97	3.54	
X1	-2.2	-1.93	-1.36	-0.78	-0.20	0.37	0.95	1.53	2.10	2.68	3.25	3.5
Prob	0	0.08	0.16	0.13	0.21	0.15	0.07	0.08	0.09	0.03	0.00	0.00
Cum. Prob	0	0.08	0.24	0.37	0.58	0.73	0.80	0.88	0.96	0.99	0.99	1.00

Histograms

min	-3.7
max	3.7

	Bins	-3.69	-3.11	-2.54	-1.96	-1.38	-0.81	-0.23	0.34	0.92	1.50	
		-3.11	-2.54	-1.96	-1.38	-0.81	-0.23	0.34	0.92	1.50	2.07	
X2	-3.7	-3.32	-2.58	-1.84	-1.10	-0.36	0.38	1.12	1.86	2.60	3.34	3.71
Prob	0	0.01	-0.01	0.04	0.10	0.12	0.14	0.24	0.12	0.09	0.09	0.00
Cum. Prob	0	0.01	0.00	0.04	0.14	0.27	0.41	0.65	0.77	0.86	0.94	1.00

Sample Data

X1	X2
0.24	0.28
1.41	-0.44
-2.10	1.27
-1.45	0.81
0.24	-0.37
1.84	-0.54
-0.69	-1.78
0.40	-2.58
0.05	-0.08
-1.36	-0.70
1.66	0.62
-0.65	1.26
-0.19	1.98
1.36	-0.96
-0.55	-0.39
-1.37	2.54
1.61	-0.55
0.91	-0.35
-0.97	-1.11
2.64	1.25
-0.21	-1.79
-1.43	-2.52
3.54	0.40
2.09	-0.15
-1.68	0.42
-1.72	1.09
-0.06	3.35
-0.18	-0.24
1.41	-2.60
-1.76	0.22
0.03	1.45
1.11	1.36
-0.16	-0.89
-1.25	3.24
0.19	0.73
0.35	-2.31
-1.52	-0.75
0.65	-1.41
2.03	0.94

A. Pooled t procedure

1. Sample Statistics

X1	
mean	0.01
st. dev.	1.26
count	105

X2	
mean	-0.14
st. dev.	1.35
count	105

2. Specify Alpha Level

Level 0.05

3. Calculate t_{stat} and $t_{critical}$

Measure	0.15
Standard Error	0.18
tstat	0.84
tcritical	1.97

4. Check criteria

$-t_{critical}$		t_{stat}		$t_{critical}$
-1.971	<	0.840	<	1.971

Ho: Fail to reject the null hypothesis.

5. Evaluate p-value.

p-value 40.20%

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**

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.

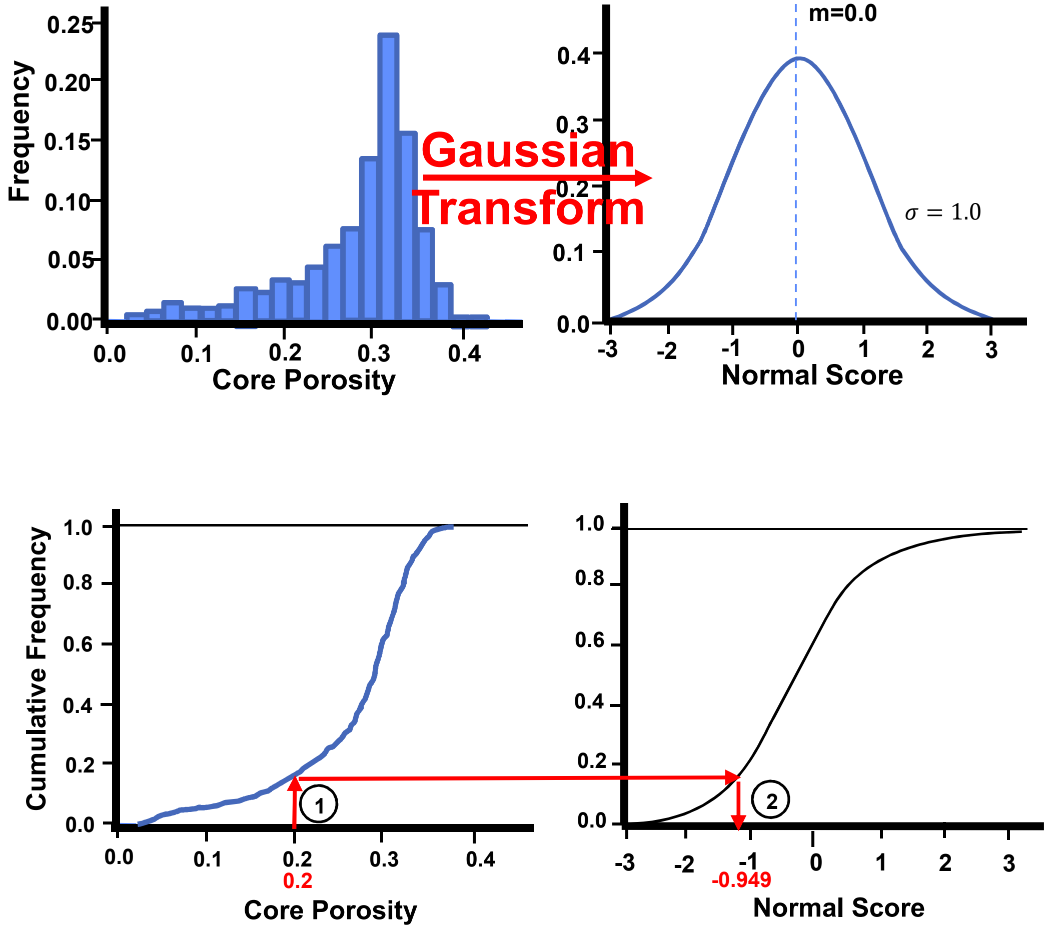
Distribution Transforms

Distribution Transforms

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

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

- 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 ($\boldsymbol{WL\varphi}$)[10, 13, 14, 15, 17]
- Core Porosity ($\boldsymbol{Core\varphi}$)[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 ($\boldsymbol{WL\varphi}$) [10, 13, 14, 15, 17]
- Core Porosity ($\boldsymbol{Core\varphi}$) [6, 9, 10, 13, 17]

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

We need to do this.

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

where F is CDF of Well Log Porosity and G is

CDF of core

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

We need the CDF of both.

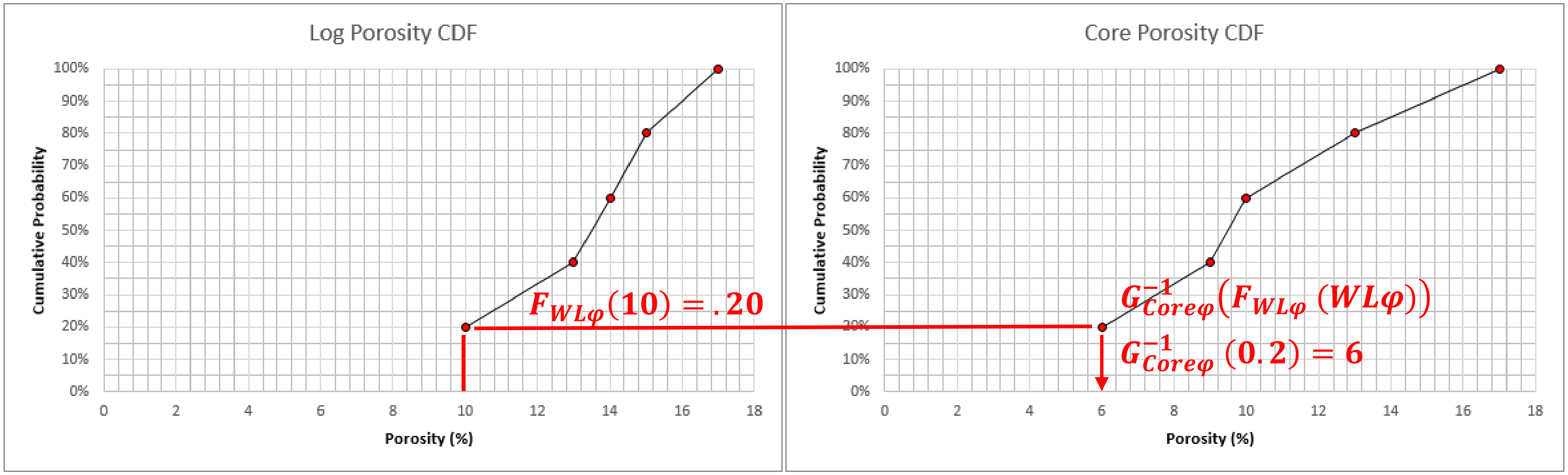
Distribution Transforms

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

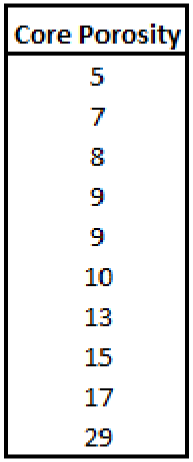
Transformed
6
9
10
13
17



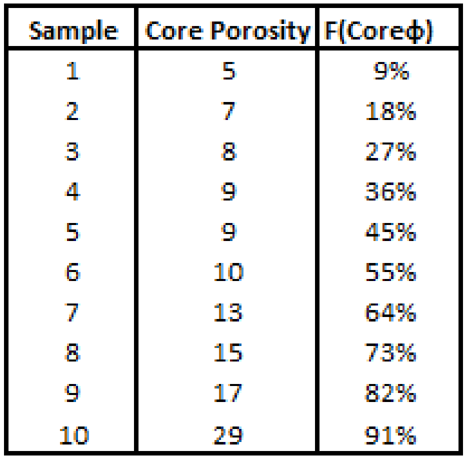
If you have the same number of data, sort the data, data have the same cumulative probabilities so you can

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
5
7
8
9
9
10
13
15
17
29



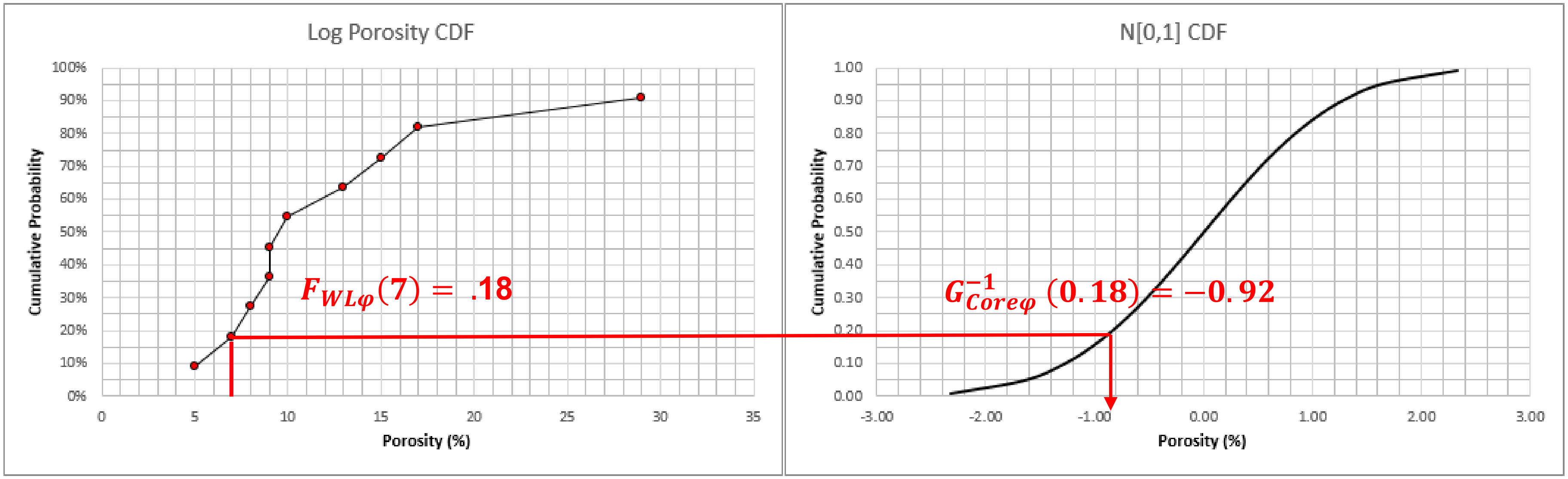
Sample	Core Porosity	F(Coreϕ)
1	5	9%
2	7	18%
3	8	27%
4	9	36%
5	9	45%
6	10	55%
7	13	64%
8	15	73%
9	17	82%
10	29	91%

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}(Coreφ)) = 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)?

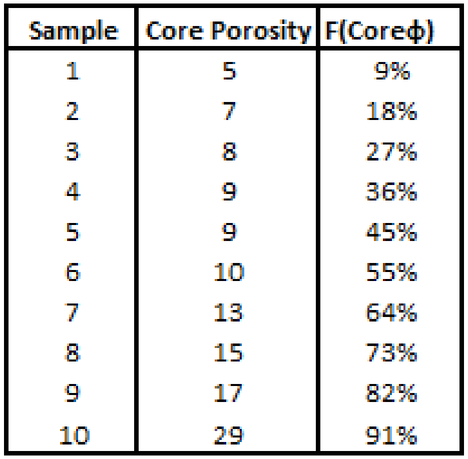


Take the cumulative probability for each value in your dataset and apply

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

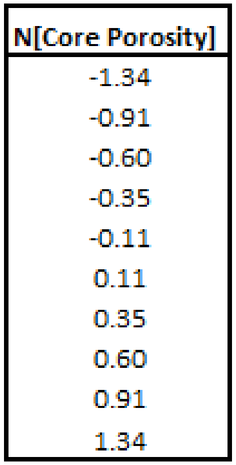
Distribution Transforms

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(Coreϕ)
1	5	9%
2	7	18%
3	8	27%
4	9	36%
5	9	45%
6	10	55%
7	13	64%
8	15	73%
9	17	82%
10	29	91%

$$\xrightarrow[\text{Norm.Inv}]{G_N^{-1}\left(F_{\varphi}\left(\varphi\right)\right)}$$



N[Core Porosity]
-1.34
-0.91
-0.60
-0.35
-0.11
0.11
0.35
0.60
0.91
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].

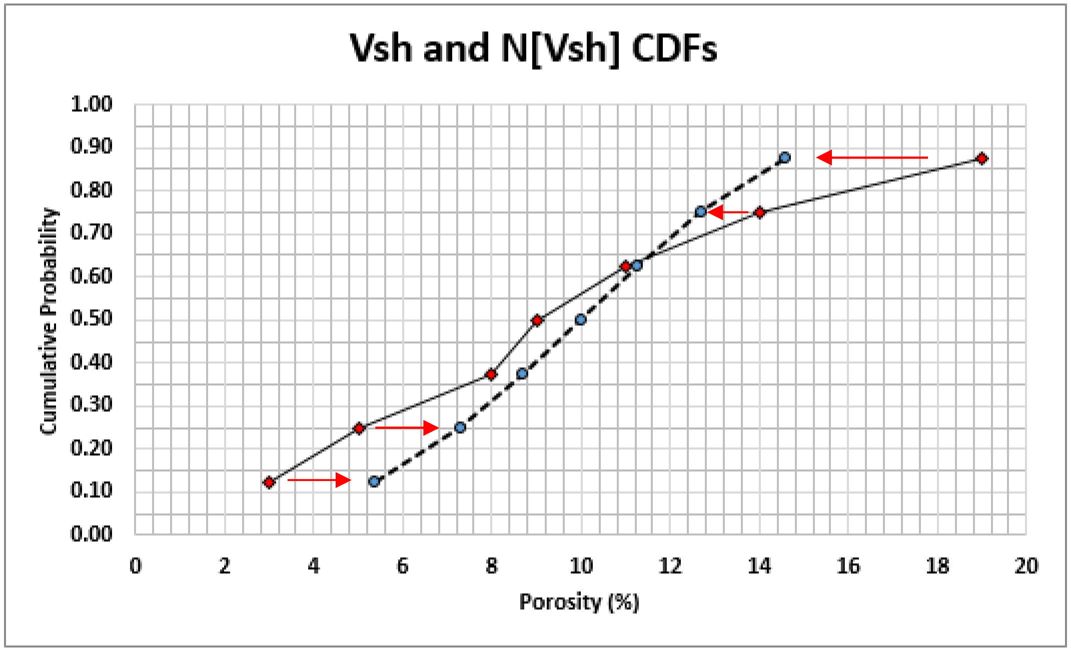
3, 5, 8, 9, 11,14, 19% Vsh

Distribution Transforms

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

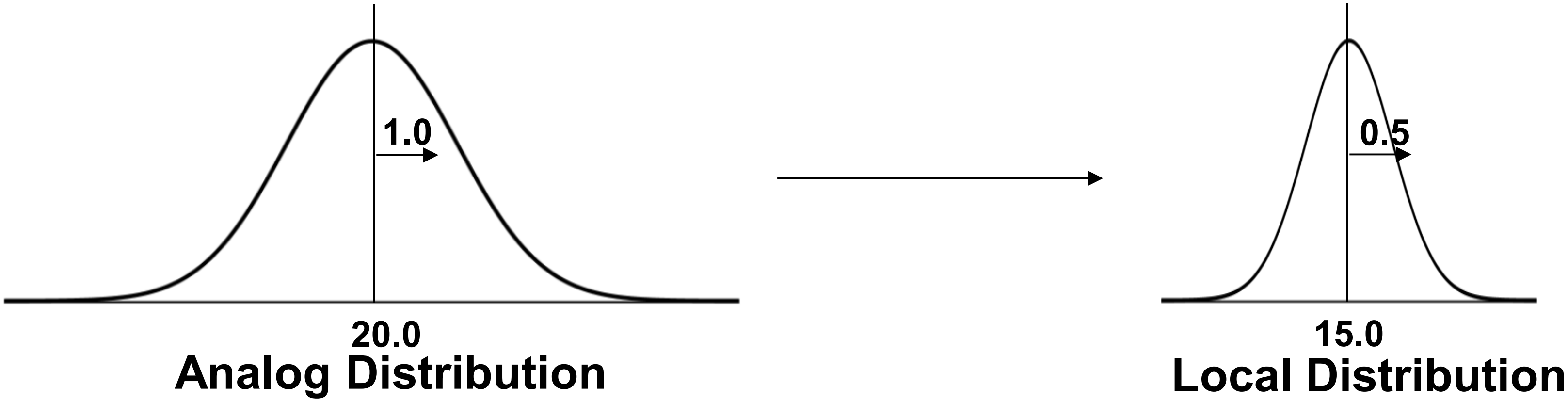
3, 5, 8, 9, 11,14, 19% Vsh

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):

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

Data Analytics and Geostatistics: Sparse Data

Lecture outline . . .

- Confidence Intervals
- Hypothesis Testing
- Distribution Transformations

Instructor: Michael Pyrcz, the University of Texas at Austin