PGE 338 Data Analytics and Geostatistics

Lecture 7: Confidence Intervals for Integrating Uncertainty Models

Lecture outline . . .

- **Concepts**
- **Analytical Confidence Intervals**
- **Bootstrap Empirical Confidence Intervals**
- **Examples**

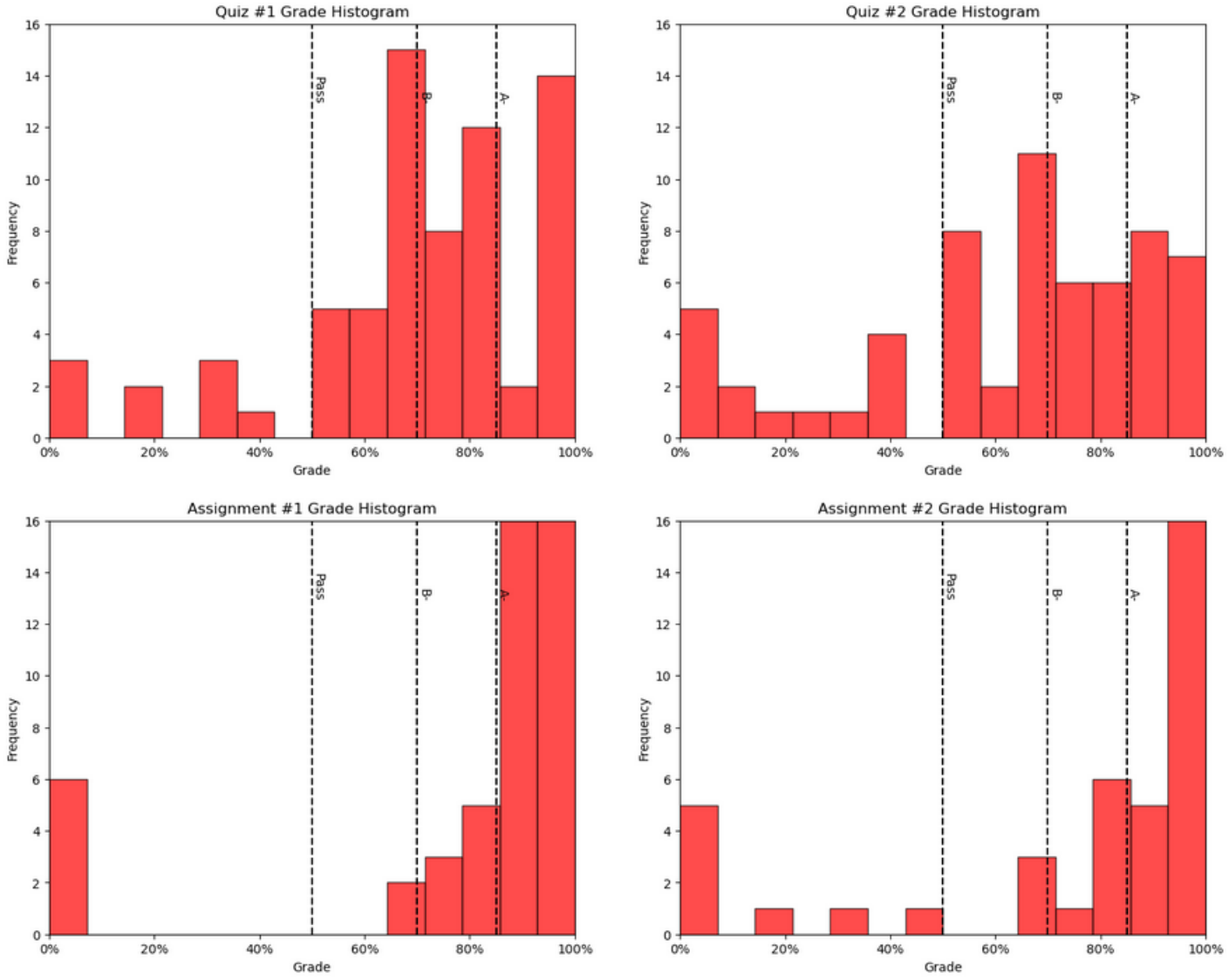
Michael Pyrcz, The University of Texas at Austin

- Introduction
- General Concepts
- Univariate
  - PDF / CDF
  - Statistics
  - Distributions
  - Heterogeneity
  - Hypothesis
- Bivariate
- Time Series Analysis
- Spatial Analysis
- Machine Learning
- Uncertainty Analysis

Checking In

How’s it going?

- If you are stuggling, now is a good time to get help!
- Office hours, TA office hours, course tutor.
- Review and make sure you understand what you got wrong, etc. for the midterm.
- See syllabus for percentage to letter grade assignments.
  - I will apply this as is.



Grades distribution on first 2 quizzes and assignments.

Checking In

From Quiz #2:

5. Bootstrap (4 points):
a) What is the output from bootstrap?

b) List the steps of bootstrap.

Grades distribution on first 2 quizzes and assignments.

Checking In

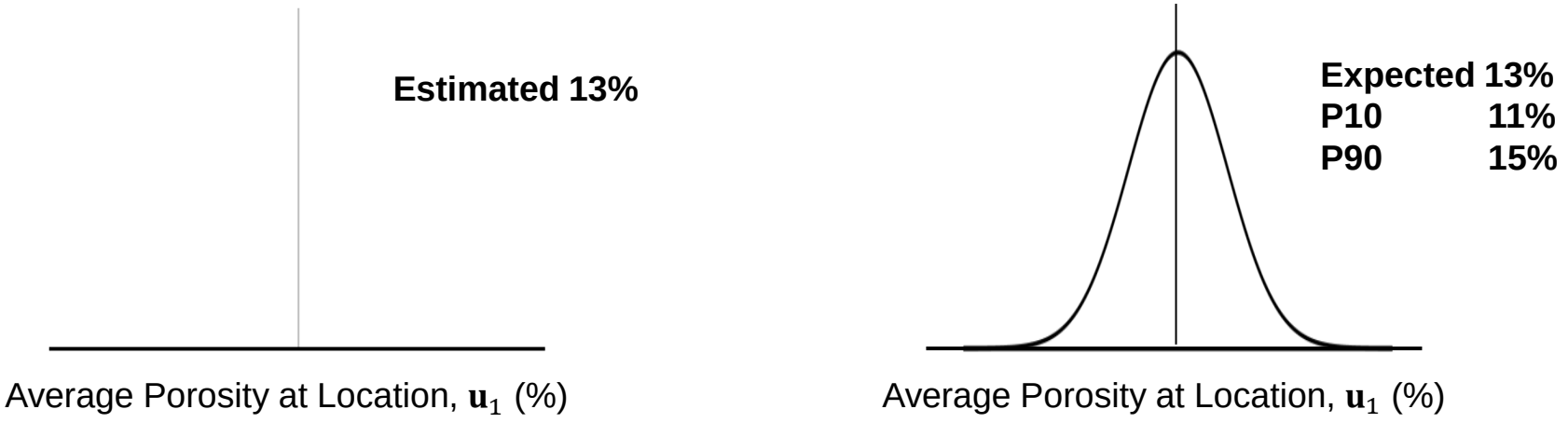
From Quiz #2:

7. Monte Carlo Simulation (4 points):
Your asset manager in Block 14 Angola needs a P10, P50 and P90 assessment of oil in place (OIP). Assuming the following random variables, $\bar{\varphi}$(average porosity) Gaussian (mean = 20%, standard deviation = 2.0%) and h (thickness of the reservoir) Gaussian (mean 50 m, standard deviation = 5 m), A = 1,000,000 $\boldsymbol{m^2}$, $\boldsymbol{s_0}$ = 0.5 and recall OIP = $\bar{\varphi} \times \boldsymbol{A} \times \boldsymbol{h} \times \boldsymbol{s_0}$. Describe a Monte Carlo simulation workflow to accomplish this. Concisely list all steps.

Motivation

We need to report uncertainty with respect to the inferred population parameters!

- Decision making is made with uncertainty models.



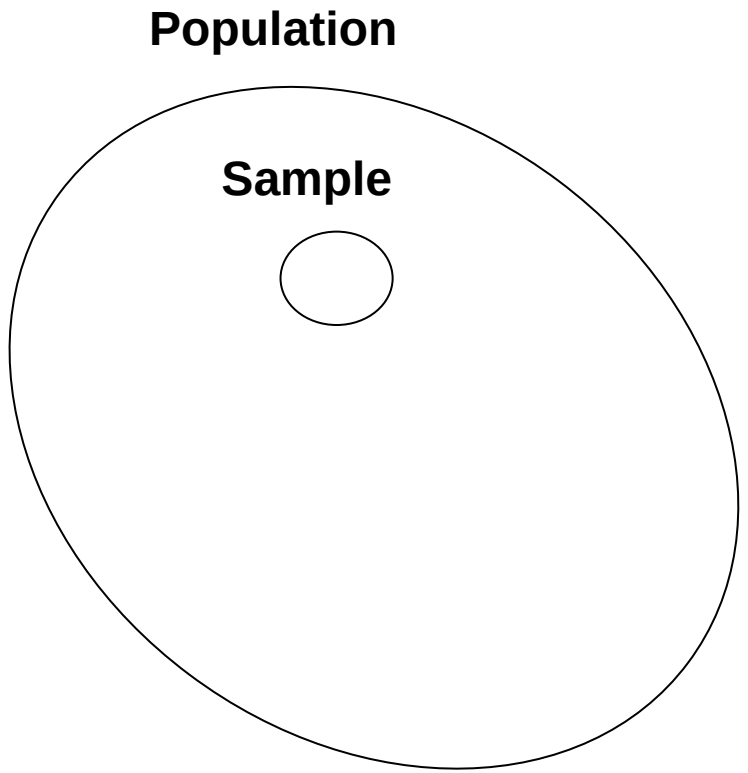
Estimate of average porosity at a location (left) and estimate with uncertainty (right).

- Uncertainty in statistics to access, upside potential, downside risk?
- Yes, we already covered bootstrap, now we cover confidence intervals and analytical methods.

Review of Nomenclature

Recall our Nomenclature for sample statistics and population parameters.

	Sample	Population
proportion	$\hat{p}$	p
mean	$\bar{x}$	μ
standard deviation	s	σ
variance	s^2	σ^2



PGE 338 Data Analytics and Geostatistics

Lecture 7: Confidence Intervals for Integrating Uncertainty Models

Lecture outline . . .

- **Concepts**

Michael Pyrcz, The University of Texas at Austin

Confidence Interval Definition

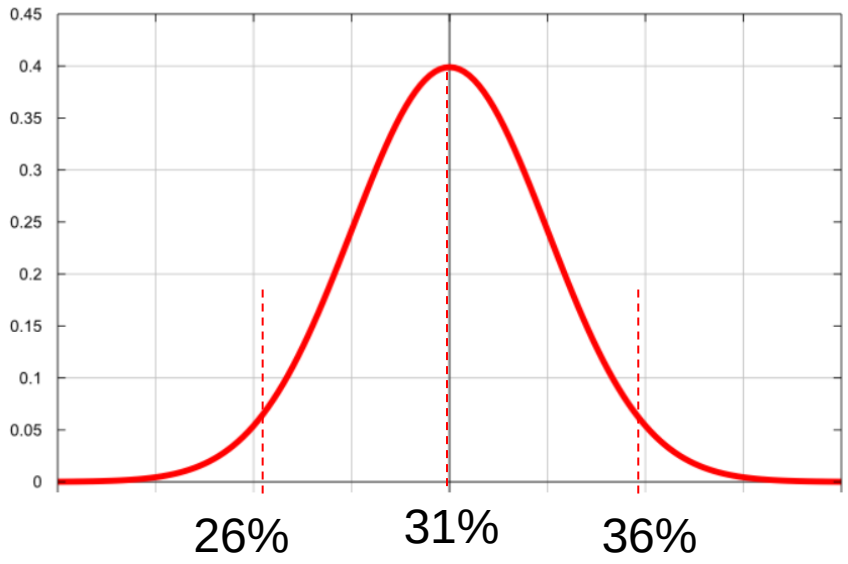
The **uncertainty in an estimated population parameter** from a sample, represented as a range, lower and upper bound, based on a specified probability interval known as the **confidence level**.

We communicate confidence intervals like this:

- **There is a 95% probability (or 19 times out of 20) that the true reservoir [population parameter], proportion of sand, is between 26% and 36%**

We cover analytical methods for population mean and proportion, then we switch to the general bootstrap method!

- Recall:
  - **Bootstrap for uncertainty in any statistic!**
    - We could do this now with bootstrap.
  - **We were calculating uncertainty in the population** parameter! We know the statistics from the sample. But, I didn't fight against the common communication.

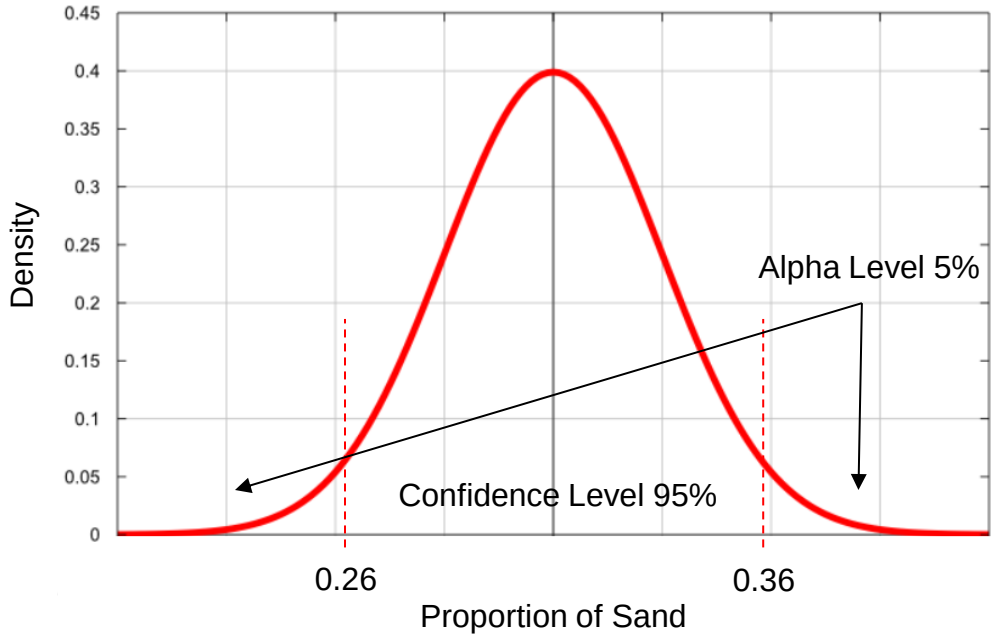


Uncertainty model in sand proportion.

Confidence Level & Alpha Level Definition

Confidence level is the probability of the population parameter being between the assessed lower and upper confidence bounds.

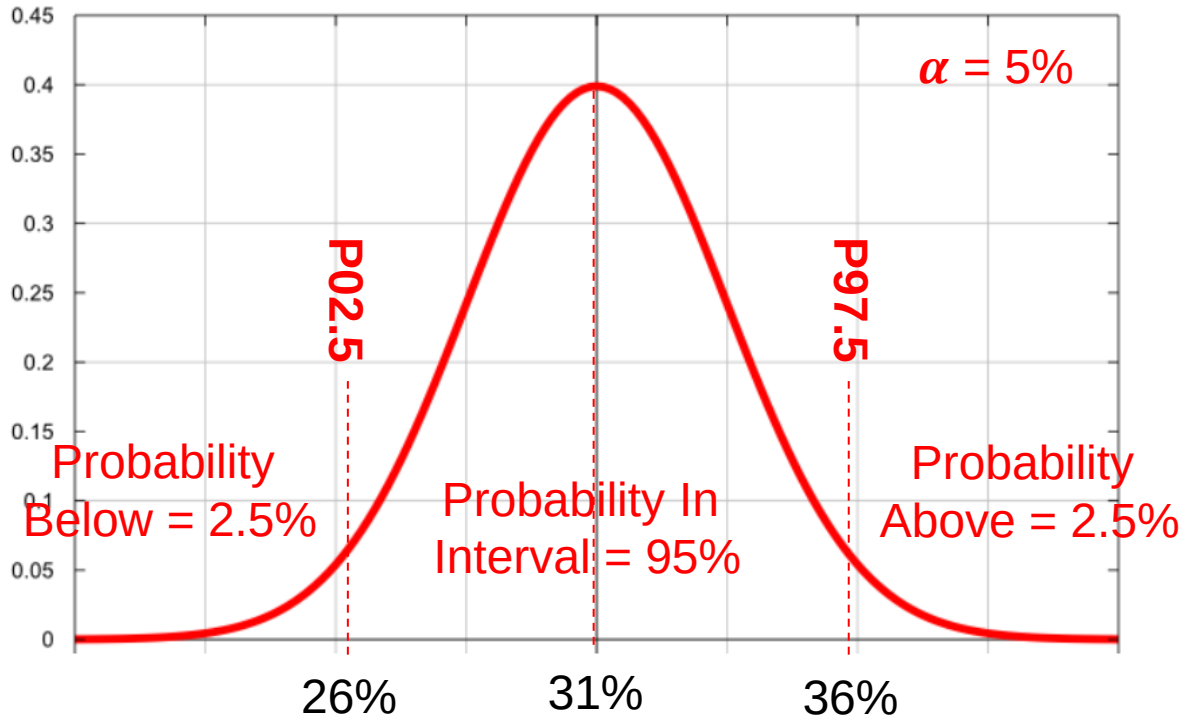
Alpha level (α) = 1 – Confidence Level, the probability the population parameter is outside the confidence interval. Alpha level is also known as significance level.



Confidence interval for proportion of sand

Confidence Level & Alpha Level Definition

What Alpha Level (α) should you use?



29% ⟷ 33%
α = 50%, P25 – P75

26% ⟷ 36%
α = 5%, P02.5 – P97.5

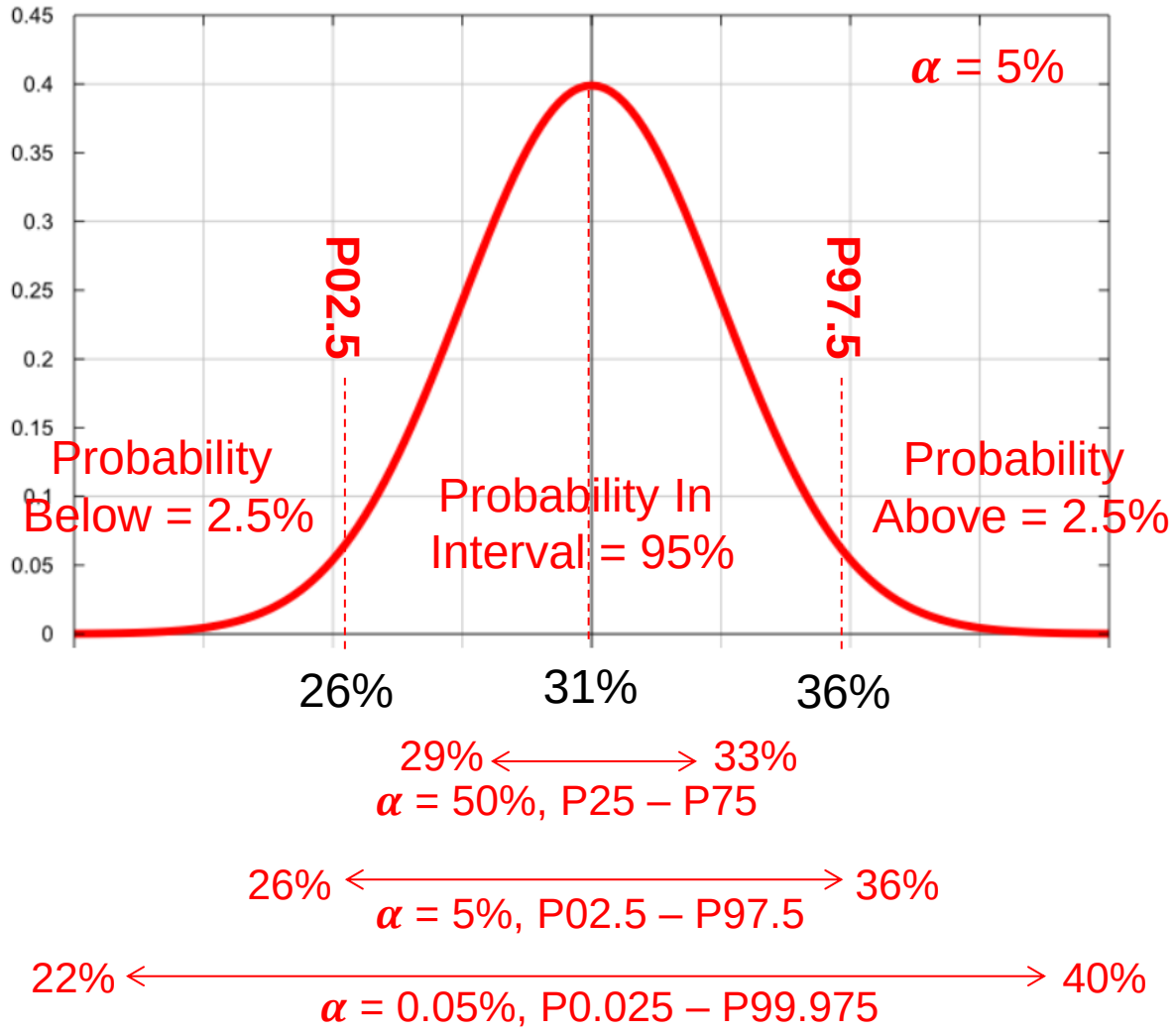
22% ⟷ 40%
α = 0.05%, P0.025 – P99.975

Confidence Level & Alpha Level Definition

What Alpha Level (α) should you use?

- too wide, almost all outcomes are in the interval - not useful!
- too narrow, many outcomes are outside the interval - not useful!
- determine through the question:

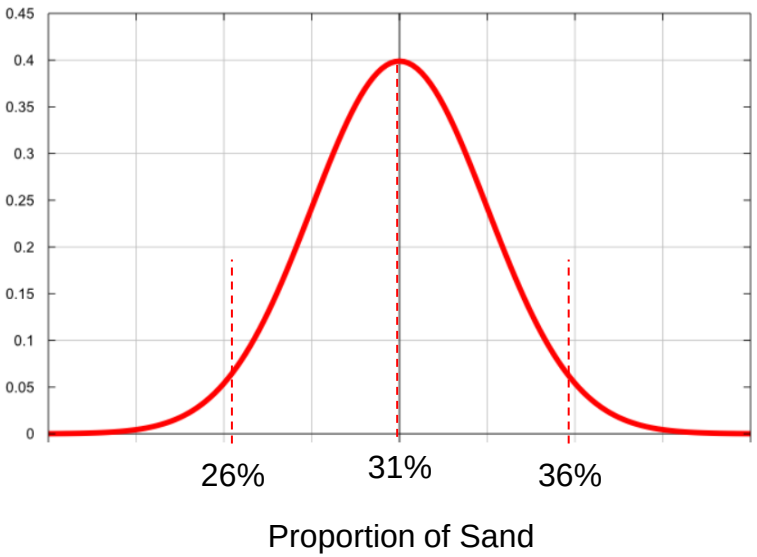
How often should a rare event occur?



Confidence Intervals Explained

What does 95% confidence in the proportion of sand mean?

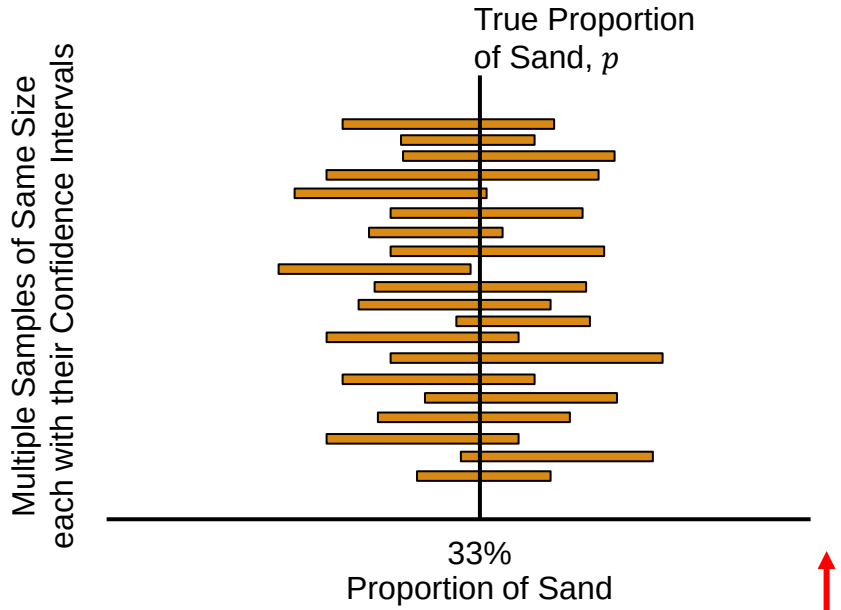
95% Probability of Including True Population Proportion of Sand



95% probability of true reservoir proportion of sand is between 26% and 36%.

Note, Bayesian credible intervals actually calculate the probability of including the true population parameter directly.

95% of Sample Sets w. Associated Confidence Intervals Include the True Population Proportion



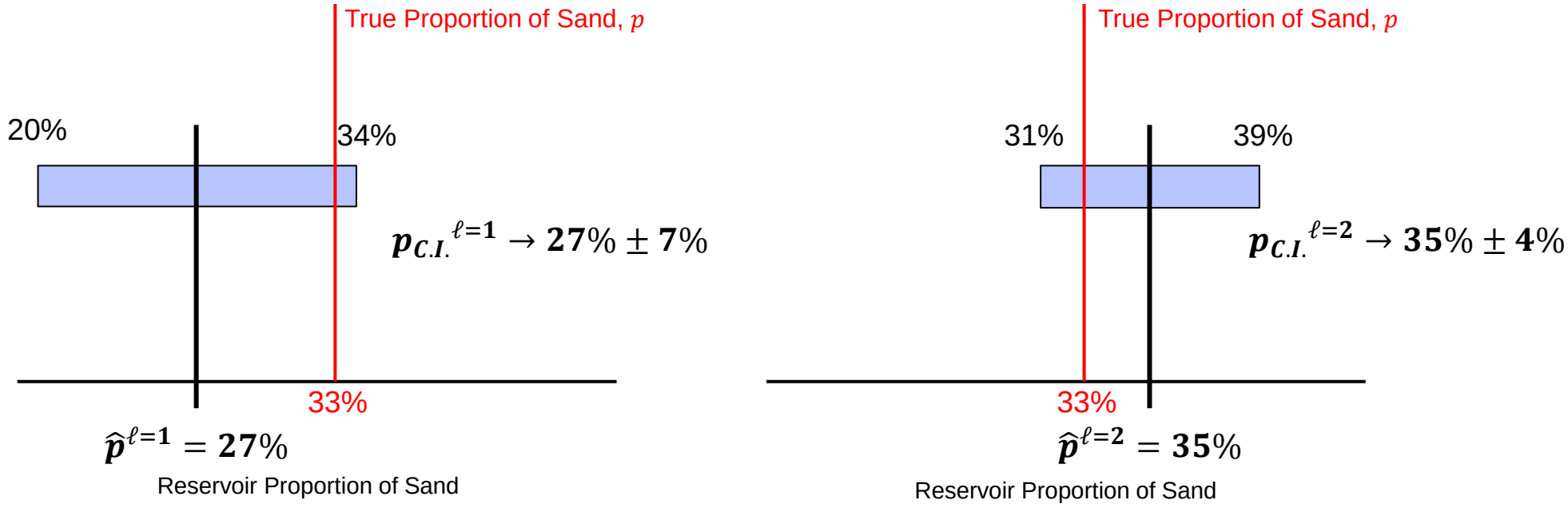
We calculate this
and we assume this.

Confidence Intervals Explained

What does 95% confidence mean? 95% of time true parameter is in the interval.

Let's illustrate further:

- Imagine two different sample sets from the population, we could calculate the 95% confidence interval from each. 95% of these should include the true population proportion.



Confidence intervals from 2 sample sets and the unknown true value.

Standard Error

‘a measure of the statistical accuracy of an estimate, **equal to the standard deviation** of the theoretical distribution of a large population of such estimates’. - Oxford Dictionary

Standard Error in a Mean:

$$\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}$$

What is the impact of more samples, n?

What is the impact of the standard deviation, $\boldsymbol{\sigma}$, of the feature of interest?

Standard Error in a Proportion:

$$\sqrt{\frac{\boldsymbol{p}(\boldsymbol{1}-\boldsymbol{p})}{\boldsymbol{n}}}$$

What is the impact of the proportion, p, of the feature of interest?

While we have these analytical forms we could bootstrap to solve empirically for accuracy for any statistic! Later our ‘Take 3’ will show this…

z-score or t-score

The number of standard deviations to the lower or upper interval of the confidence interval.

- calculated from the $\boldsymbol{\alpha}$ value applied to the standard Gaussian (z-score) or Student's t distribution (t-score), it is the inverse of the CDF, $G_y^{-1}(p)$, where p is the percentile of the lower/upper bound.

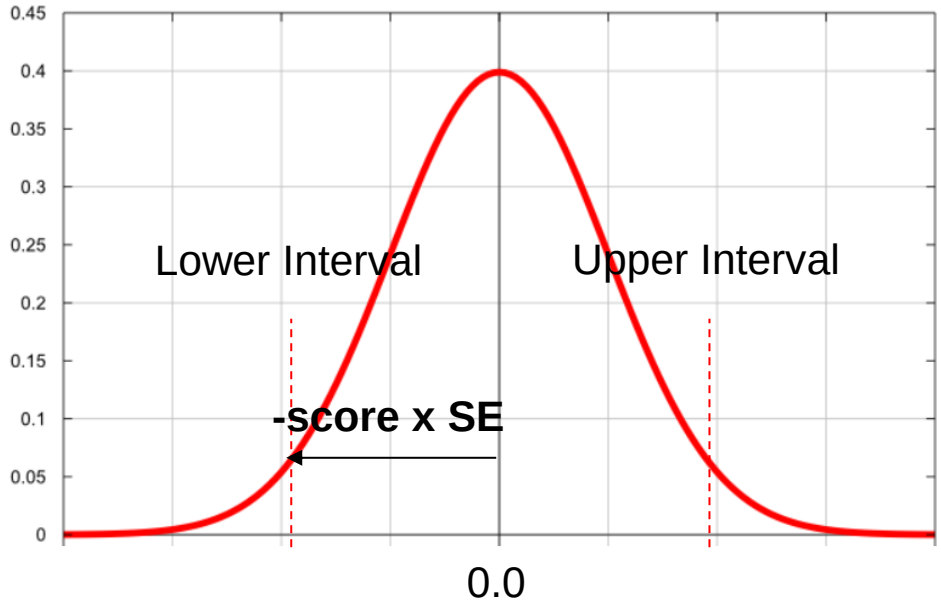
z-score example, given $\alpha = 5\%$ we need the 0.025 percentile.

- Python: SciPy.stats.norm.ppf(0.025) = -1.96
- Excel: norm.inv(0.025,0,1) = -1.96

t-score example, given $\alpha = 5\%$ we need the 0.025 percentile.

- Python: SciPy.stats.t.ppf(0.025,15-1) = -2.14
- Excel: t.inv(0.025,15-1) = -2.14

Accounts for the shape of the theoretical distribution of the confidence interval. More later.



Standard Gaussian or Student's t distribution. The mean is 0.0 and the standard deviation is 1.0.

PGE 338 Data Analytics and Geostatistics

Lecture 7: Confidence Intervals for Integrating Uncertainty Models

Lecture outline . . .

- **Analytical Confidence Intervals**

Michael Pyrcz, The University of Texas at Austin

Confidence Interval Analytical Expression

Now we are ready to consider the analytical expression for the confidence interval (C.I.) of the population proportion.

sample proportion

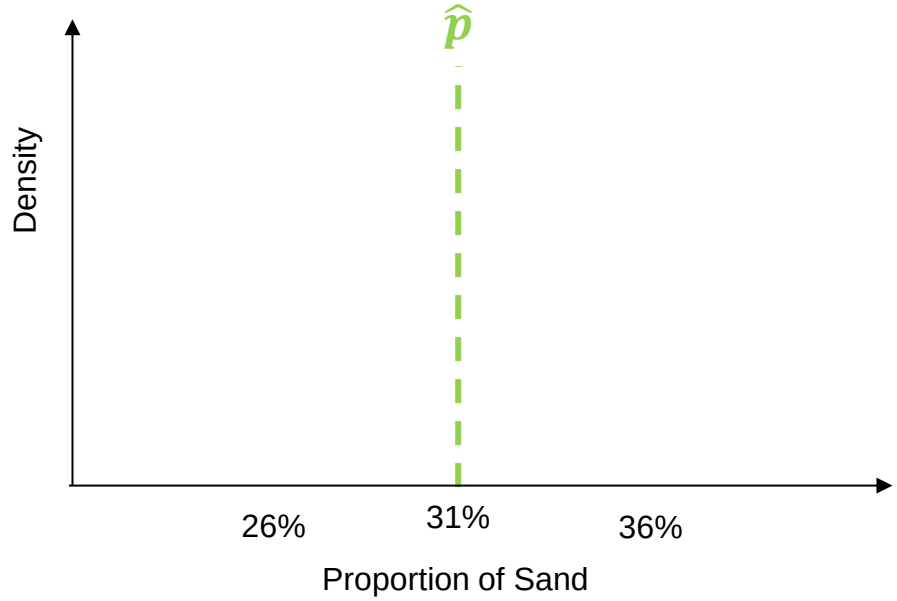
-score

standard error of a proportion

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

Sample Statistic, $\hat{p}$

- Note, the confidence interval is centered on the statistics from the sample.
- We are assuming the statistic is unbiased, the best estimate of the centroid of our uncertainty model



Uncertainty model in sand proportion.

Confidence Interval Analytical Expression

Now we are ready to consider the analytical expression for the confidence interval (C.I.) of the population proportion.

sample proportion

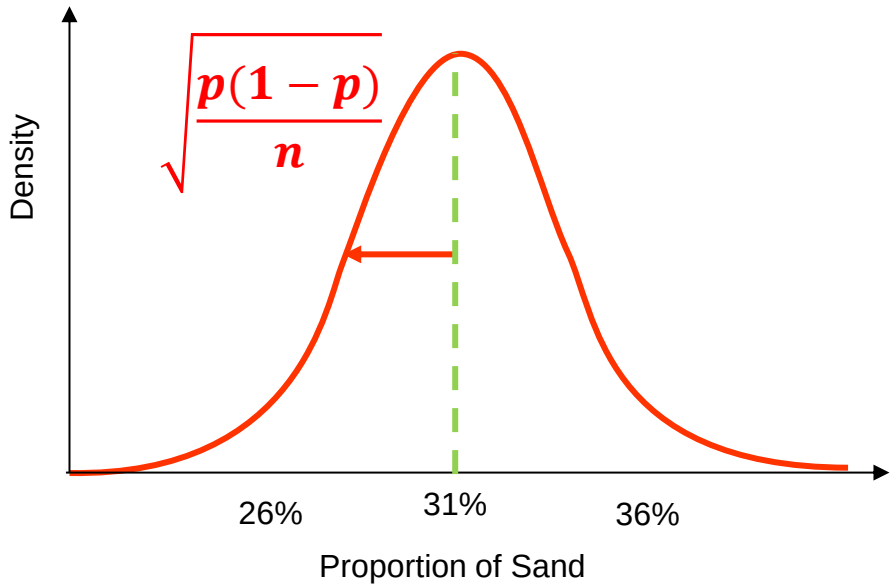
-score

standard error of a proportion

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

Standard Error, $\sqrt{\frac{p(1-p)}{n}}$

- Standard deviation, dispersion, of our uncertainty distribution
- Note, $p(1-p)$ is the variance of a binary feature (e.g., sand and shale).
- The form is standard deviation / square root of number of data
- Standard error for the mean we use, $\sqrt{\frac{\sigma^2}{n}}$, also standard deviation divided by the square root of the number of data



Uncertainty model in sand proportion.

Confidence Interval Analytical Expression

Now we are ready to consider the analytical expression for the confidence interval (C.I.) of the population proportion.

sample proportion

-score

standard error of a proportion

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

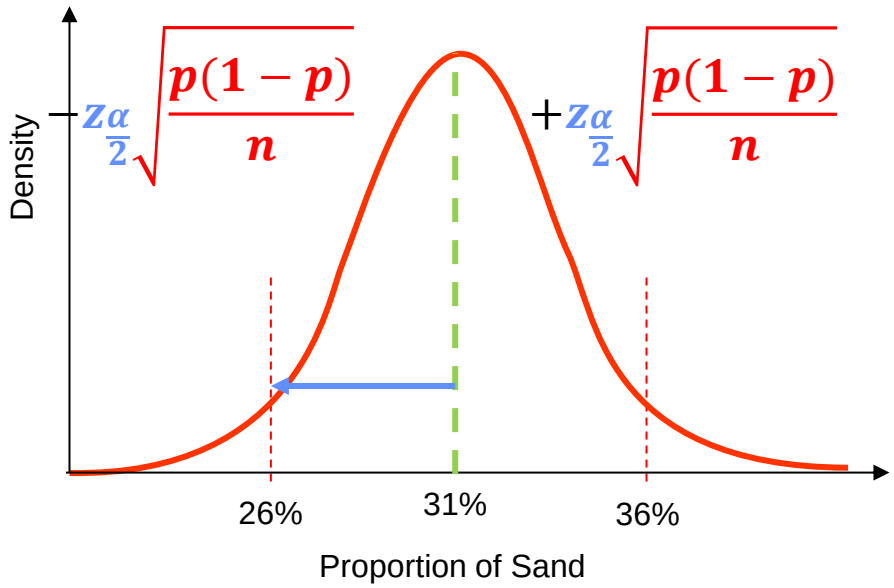
z-score, $z_{\frac{\alpha}{2}}$

- the number of standard deviations to reach the lower / upper confident intervals

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
–3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
–3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
–3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
–3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
–3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
–2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
–2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
–2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
–2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
–2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
–2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
–2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
–2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
–2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
–2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
–1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

Use a function or a table to look up a z-score given $\frac{\alpha}{2}$

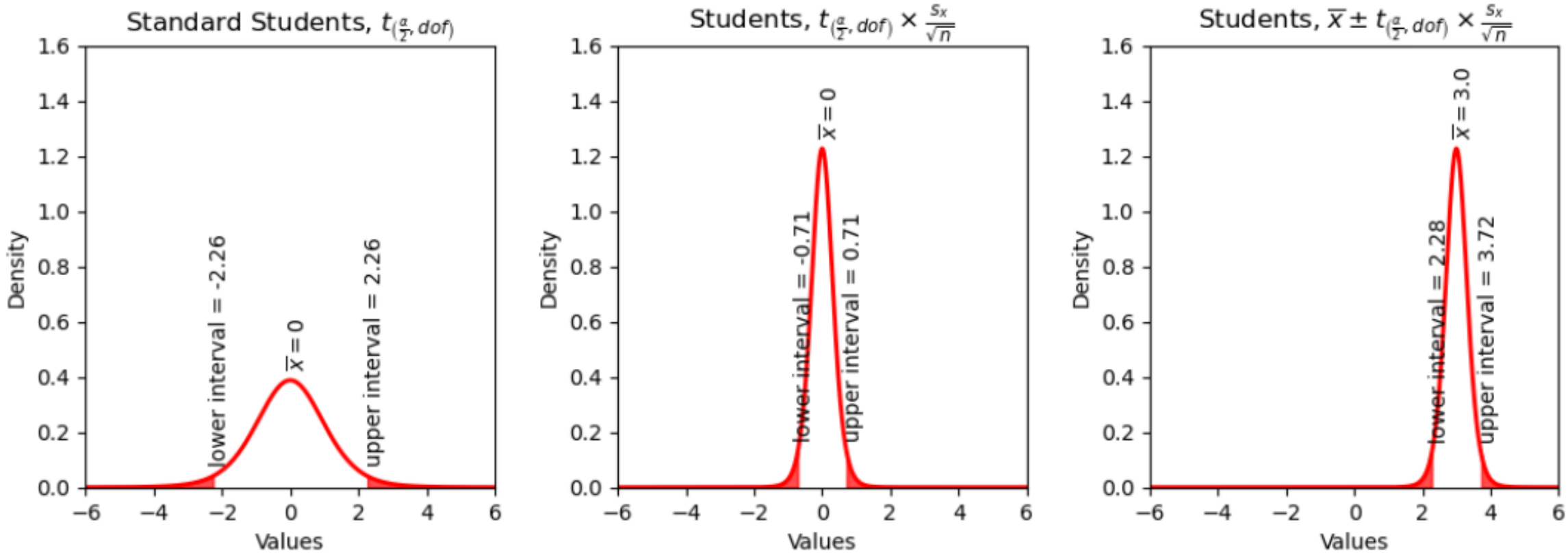
- the offset from the sample statistic, $z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$
- use a function or a table to get the –score.



Uncertainty model in sand proportion.

Confidence Interval Analytical Expression

Now we are ready to consider the analytical expression for the confidence interval (C.I.) of the population proportion.



Visualization of analytical confidence interval for the population mean, standard Student's t distribution (left), with standard deviation = SE (center), centered on statistic (right).

In Other Words Confidence Intervals

A confidence interval is constructed by taking a statistic and adding and subtracting a margin of error.

- Example: I am 95% confident that population mean, $\boldsymbol{\mu}$, will fall in this interval.

Example: then 95% confidence interval will be equal to:

$$\textbf{Statistic} \pm \textbf{Score} \times \textbf{Standard Error}$$

Score depends on the distribution e.g., t-score (Student's t), z-score (Gaussian)

$$\mu_{C.I.} \rightarrow \bar{x} \pm 1.96 \times se$$

se is standard error

1.96 is based on desired interval and distribution shape / Gaussian for 95% confidence interval

- The 95% confidence relates to the reliability of the estimation procedure for the population parameter, e.g., the estimate of the average.

Slide modified from Dr. Zoya Heidari's PGE 337 Course

Confidence Intervals

How does this get used in industry?

- The language of uncertainty is ubiquitous
- Decision makers needs to know:
  - uncertainty in your results → **how much uncertainty in your estimate?**
  - significance of your results → **how much uncertainty relative to observed differences and decision criteria?**

Significance Example:

- your estimate reserves is 10 tonnes > economic hurdle, but the uncertainty in your estimate is +/- 50 tonnes

A lot of scientists and engineers do not know how to do this!

- Add this to your toolbox to differentiate yourself and improve the impact of your work

Confidence Intervals with Small Sample Sizes

For example, consider the confidence interval (C.I.) of the proportion and mean:

$$p_{C.I.} \rightarrow \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \qquad \mu_{C.I.} \rightarrow \bar{x} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}}$$

- But we don't have the true population proportion, p, nor population variance, σ^2! This adds more uncertainty if we don't have enough data for a reliable estimate of the p or σ^2.

$$p_{C.I.} \rightarrow \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}} \approx t_{\frac{\alpha}{2},n-1} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad \mu_{C.I.} \rightarrow \bar{x} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\sigma^2}{n}} \approx t_{\frac{\alpha}{2},n-1} \sqrt{\frac{s^2}{n}}$$

We use the Student's t distribution, t-scores, instead of Gaussian, z-score, to account for small sample size.

Confidence Intervals with Small Sample Sizes

Confidence Interval for a Mean or Proportion Statistic

If enough samples, we can use the Gaussian distribution:

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right), \qquad if\ n \geq 30$$

Common Criteria for Enough Samples

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right), \qquad if\ np \geq 30 \text{ and } n(1-p) \geq 30$$

atleast 30 case of each category.

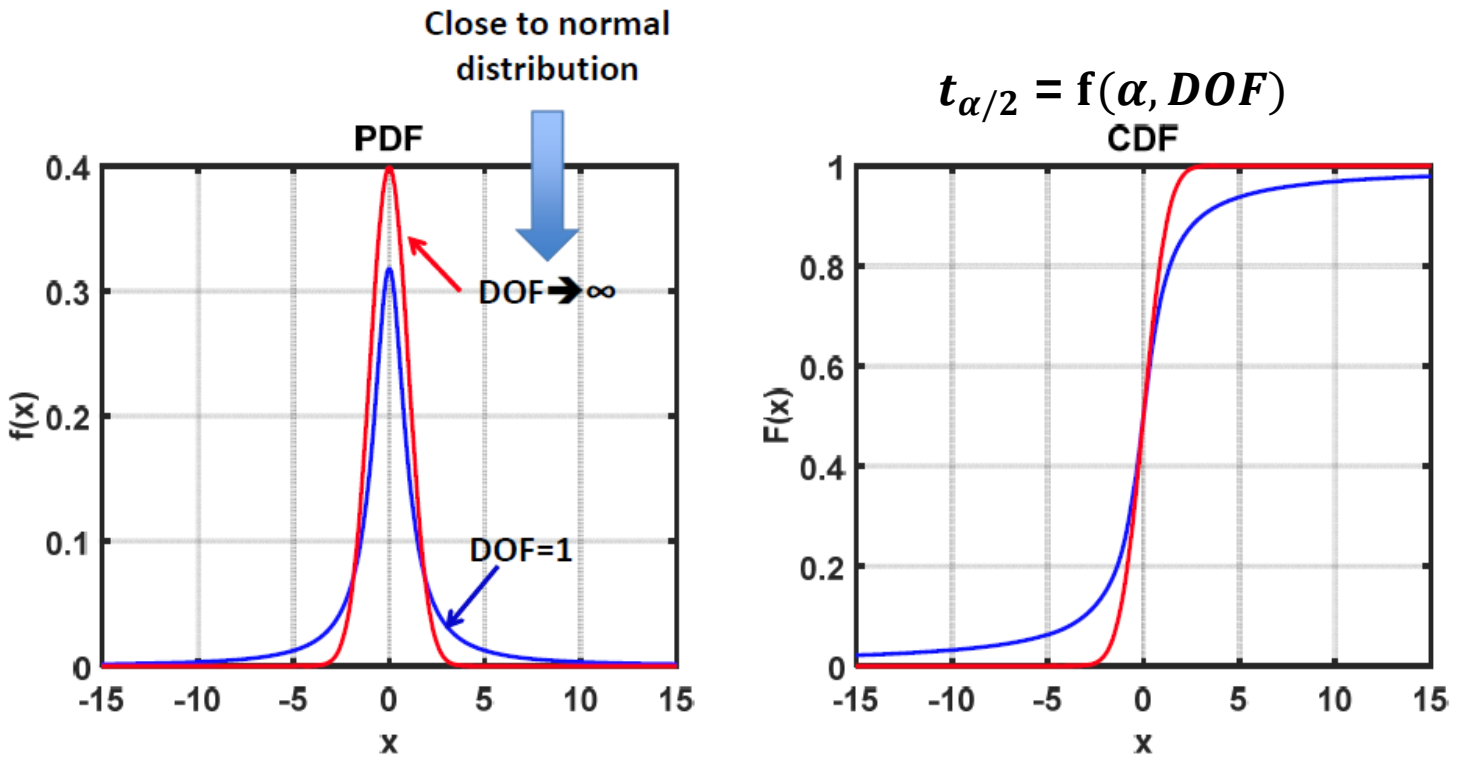
Note: ′∼’ means [,is read as] ‘distributed as’.

Otherwise use Student’s t-distribution.

Student's t-Distribution

Recall: the t-distribution like Gaussian with "fatter tails".

- As sample size is large ($dof = n - 1$) the t-distribution approaches the Normal distribution.
- We use the **student's t distribution for C.I. of mean and proportions** because we don't know the true standard deviation or proportion, if small sample size is small.



For large (enough) degrees of freedom the Student's t distribution approaches Gaussian distribution.

Image from Dr. Zoya Heidari's PGE 337.

Degrees of Freedom Definition

Degree of Freedom (dof):

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

Representation: ν or $d.f.$ or DOF

number of independent pieces of information = number of data – number of assumptions

Example for Estimating Sum of Squares (or variance):

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

DOF = sample size – constraints (estimate of the mean) for estimating the sum of squares:

$$DOF = n - 1$$

Example Confidence Levels

Some Common Confidence Levels, demonstrated for Gaussian distribution

Confidence Level	Alpha Level	Type I Error Probability	z-Score	Confidence Interval
0.90	0.10	0.10	1.645	$\hat{p} \pm 1.645se$
0.95	0.05	0.05	1.96	$\hat{p} \pm 1.96se$
0.99	0.01	0.01	2.58	$\hat{p} \pm 2.58se$

Comments:

- Use of 0.95 confidence level is common.
- **Alpha level** is the probability of probability of **type I error**, rejecting the null hypothesis when the null hypothesis is true (false reject is a **false positive**).
- Use a small alpha is conservative, you will most often fail to reject in hypothesis tests (more later).

Confidence Interval Example

Example 1, using Excel:

Confidence Interval in Sample Mean Example:

if $\bar{x}$ = 14%, s = 2% and n = 100, then for 95% CI,

$$\bar{x} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{s^2}{n}} \text{ or expressed as } \bar{x} \pm z_{\frac{\alpha}{2}}\frac{s}{\sqrt{n}}$$

P2.5

$z_{\frac{\alpha}{2},n-1}$ →**NORM.INV($\alpha/2$) in Excel = 1.96.**

$$14\% \pm 1.96\frac{2\%}{\sqrt{100}}$$

14% ±0.39%

Confidence Interval Example

Example 2, using a table:

if $\bar{x}$ = 400 mD, s = 50 mD and n = 12, then 95% CI,

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}} \qquad t_{\frac{0.05}{2},12-1} = 2.20$$

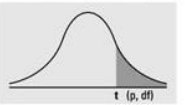
Check with: $t_{\frac{\alpha}{2},n-1}$ →T.INV.2T(α,n-1) or T.INV($\alpha/2$,n-1) in Excel.

$$400 \text{ mD} \pm 2.20\frac{50 \text{ mD}}{\sqrt{12}}$$

$$400 \text{ mD} \pm 31.7 \text{ mD}$$

Numbers in each row of the table are values on a *t*-distribution with (*df*) degrees of freedom for selected right-tail (greater-than) probabilities (*p*).

$\frac{\alpha}{2}$



Degrees of Freedom

df/p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	43178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251
26	0.255955	0.684043	1.314972	1.705618	2.05553	2.47863	2.77871	3.7066
27	0.255858	0.683685	1.313703	1.703288	2.05183	2.47266	2.77068	3.6896
28	0.255768	0.683353	1.312527	1.701131	2.04841	2.46714	2.76326	3.6739
29	0.255684	0.683044	1.311434	1.699127	2.04523	2.46202	2.75639	3.6594
30	0.255605	0.682756	1.310415	1.697261	2.04227	2.45726	2.75000	3.6460
z	0.253347	0.674490	1.281552	1.644854	1.95996	2.32635	2.57583	3.2905
CI	———	———	80%	90%	95%	98%	99%	99.9%

t-score table by alpha/2 (columns) and degrees of freedom rows.

Image from http://www.dummies.com/education/math/statistics/how-to-use-the-t-table-to-solve-statistics-problems/

Confidence Interval Take 2

Let's take another run at confidence intervals.

The 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

The Problem:

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

Average Porosity = 15%

Recall, this is the central tendency, average, of the uncertainty distribution!

- Given 16 samples and sample standard deviation = 2

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

Recall, this is the dispersion, standard deviation, of the uncertainty distribution!

- Uncertainty in Reservoir [Population] Average Porosity

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

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

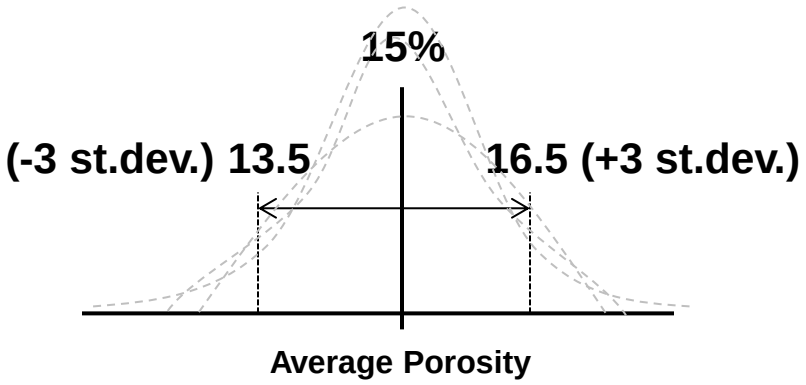
Confidence Interval
Take 2

The Problem:

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

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

  - 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.

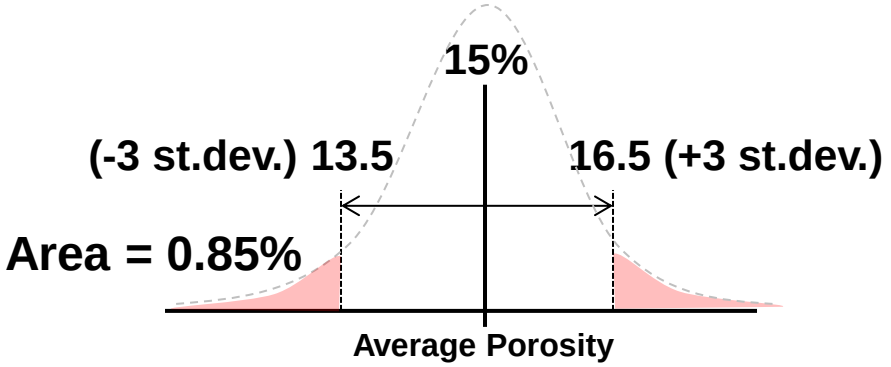
Confidence Interval
Take 2

The Problem:

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

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

  - 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.

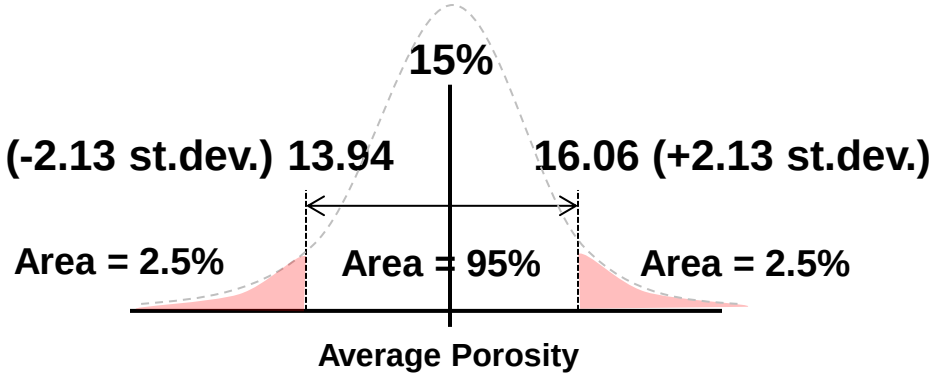
Confidence Interval
Take 2

The Problem:

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

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

  - 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

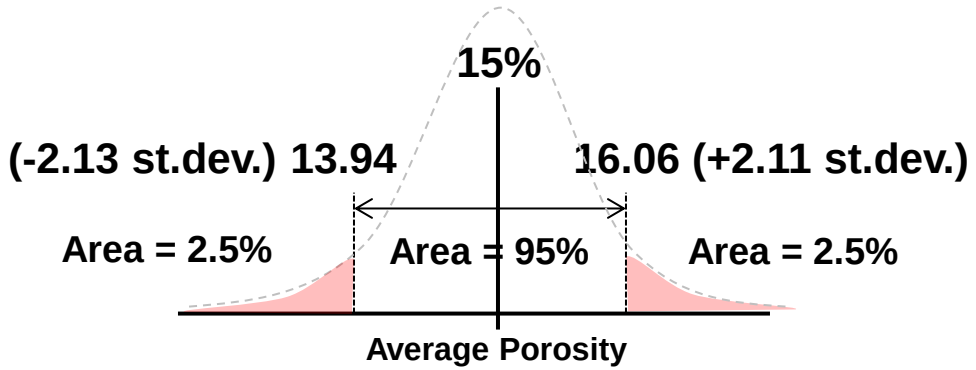
The Problem:

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

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

or

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



$$\mu_{C.I.} \rightarrow \bar{x} \pm t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}$$

Confidence Interval
Take 2

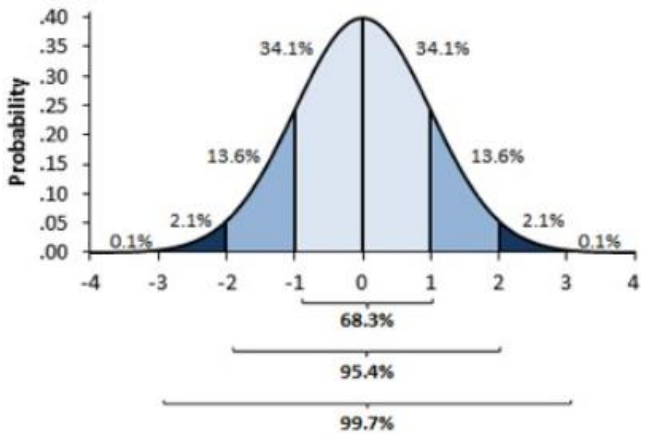
What did we learn?

- **standard error** tells us the dispersion of our uncertainty with respect to the population parameter

- BUT we still need the distribution shape. We know specific distributions occur for different processes. For example:

    1. average / proportion – Gaussian, Student's t if too few samples (<30 is commonly used)
    2. variance or average of squares – Chi-square

- we calculate the score (called e.g., t-score) from that standardized distribution (mean of 0.0, st.dev. of 1.0) scaled (multiplied) by standard error and offset (add to) the estimate to get our confidence interval.

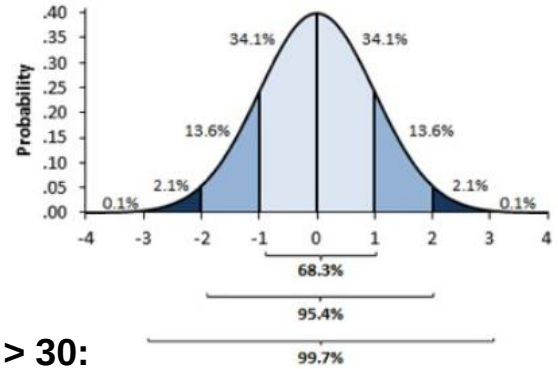


z-statistic / z-score

Confidence Interval Take 2

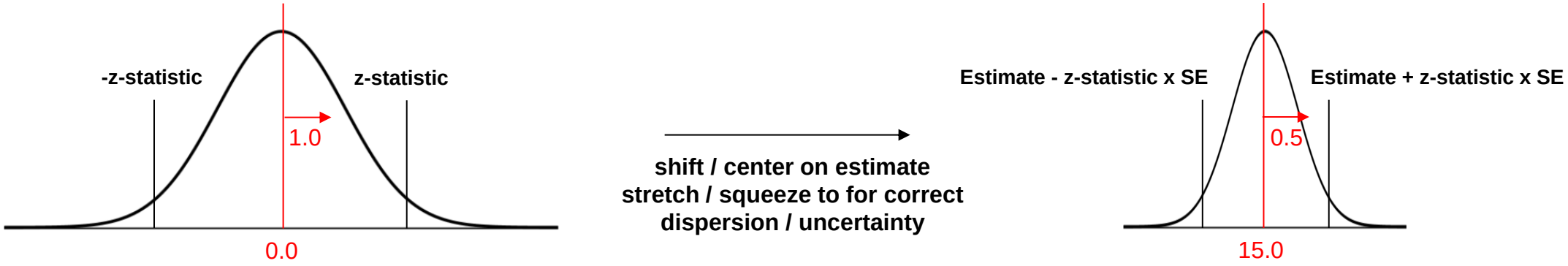
What did we learn?

- we calculate the score (called e.g., t-score) from that standardized distribution (mean of 0.0, st.dev. of 1.0) scaled (multiplied) by standard error and shift (addition) by estimate to get our confidence interval.



z-statistic / z-score

For previous example if n > 30:



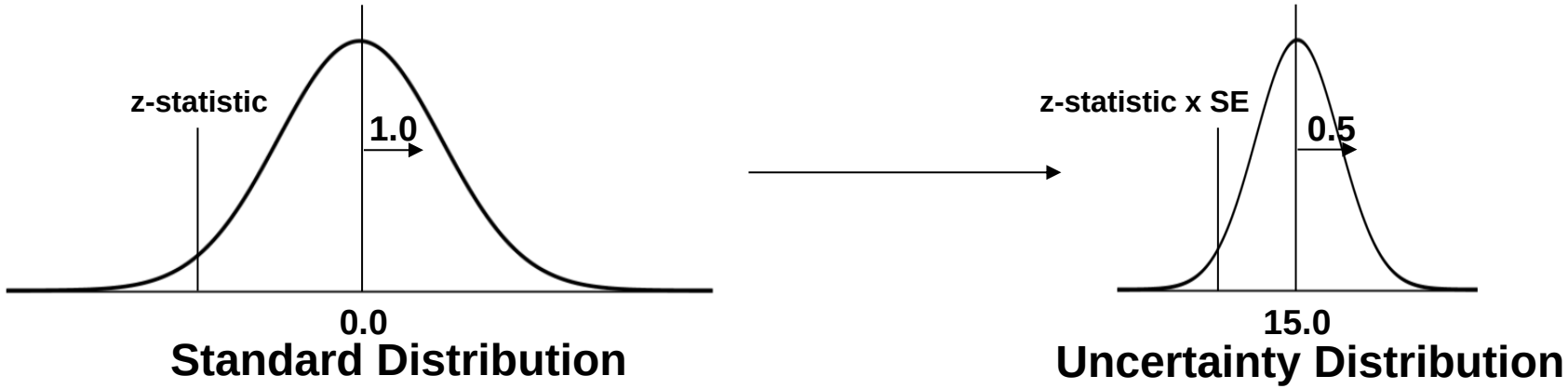
Theoretical Distribution / Standard Normal

Uncertainty Distribution / Confidence Intervals

Confidence Interval Take 2

What did we learn?:

- Distribution scaling – does that really work?



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

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

In our case the distributions' initial mean = 0.0 and standard deviation = 1.0 and we add the estimate.

$$y_{P97.5} = \sigma_{SE}\left(y_{P97.5,\,standard}\right) + \hat{y}$$

✓ **confidence interval = SE x z-score + estimate**

Confidence Intervals in Python

Confidence intervals in Python

- Mean and Proportion
- Using SciPy.stats.t.interval function and checked by-hand with elementary functions

Data Analytics

Confidence Intervals and Hypothesis Testing in Python in Python

Michael Pyrcz, Associate Professor, The University of Texas at Austin

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

This is a tutorial / demonstration of **Confidence Intervals and Hypothesis Testing in Python**. In Python, the SciPy package, specifically the Stats functions (https://docs.scipy.org/doc/scipy/reference/stats.html) provide excellent tools for efficient use of statistics.

I have previously provided these examples worked out by-hand in Excel (https://github.com/GeostatsGuy/LectureExercises/blob/master/Lecture7_CI_Hypoth_eg_R.xlsx) and also in R (https://github.com/GeostatsGuy/LectureExercises/blob/master/Lecture7_CI_Hypoth_eg.R). In all cases, I use the same dataset available as a comma delimited file (https://git.io/fxLAt).

This tutorial includes basic, typical confidence interval and hypothesis testing methods that would commonly be required for Engineers and Geoscientists including:

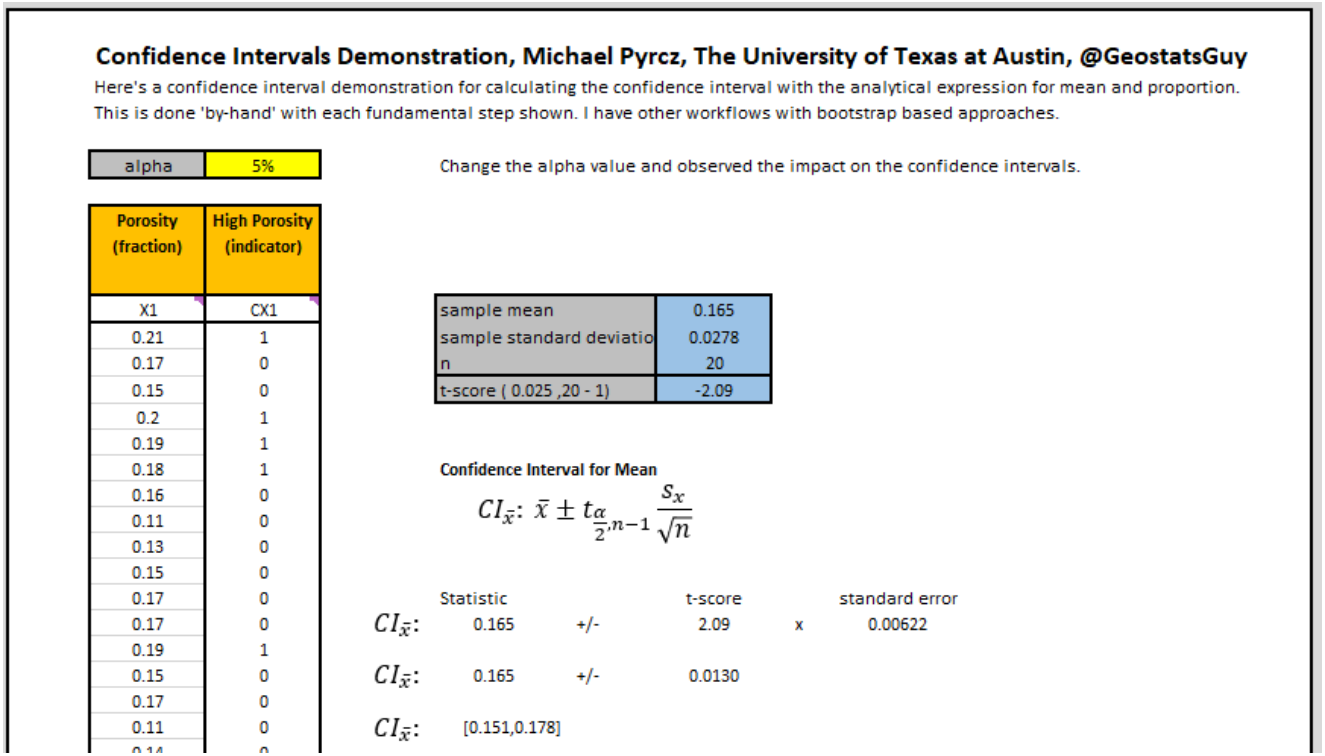
1. Student-t confidence interval for the mean and proportion
2. Student-t hypothesis test for difference in means (pooled variance)
3. Student-t hypothesis test for difference in means (difference variances), Welch's t Test
4. F-distribution hypothesis test for difference in variances

Jupyter notebook Python demonstration,
file is ‘PythonDataBasics_ConfidenceInterval_HypothesisTesting.ipynb’.

Confidence Intervals in Excel

Confidence intervals in Python

- Mean and Proportion
- Using basic Excel function, ‘by-hand’



Confidence Intervals Demonstration, Michael Pyrcz, The University of Texas at Austin, @GeostatsGuy

Here's a confidence interval demonstration for calculating the confidence interval with the analytical expression for mean and proportion.
This is done 'by-hand' with each fundamental step shown. I have other workflows with bootstrap based approaches.

alpha	5%

Change the alpha value and observed the impact on the confidence intervals.

Porosity (fraction)	High Porosity (indicator)
X1	CX1
0.21	1
0.17	0
0.15	0
0.2	1
0.19	1
0.18	1
0.16	0
0.11	0
0.13	0
0.15	0
0.17	0
0.17	0
0.19	1
0.15	0
0.17	0
0.11	0

sample mean	0.165
sample standard deviatio	0.0278
n	20
t-score (0.025 ,20 - 1)	-2.09

Confidence Interval for Mean

$$CI_{\bar{x}}: \bar{x} \pm t_{\frac{\alpha}{2},n-1} \frac{s_x}{\sqrt{n}}$$

	Statistic		t-score		standard error
$CI_{\bar{x}}$:	0.165	+/-	2.09	x	0.00622
$CI_{\bar{x}}$:	0.165	+/-	0.0130		
$CI_{\bar{x}}$:	[0.151,0.178]				

Excel demonstration of confidence intervals, file is Confidence_Intervals.xlsx.

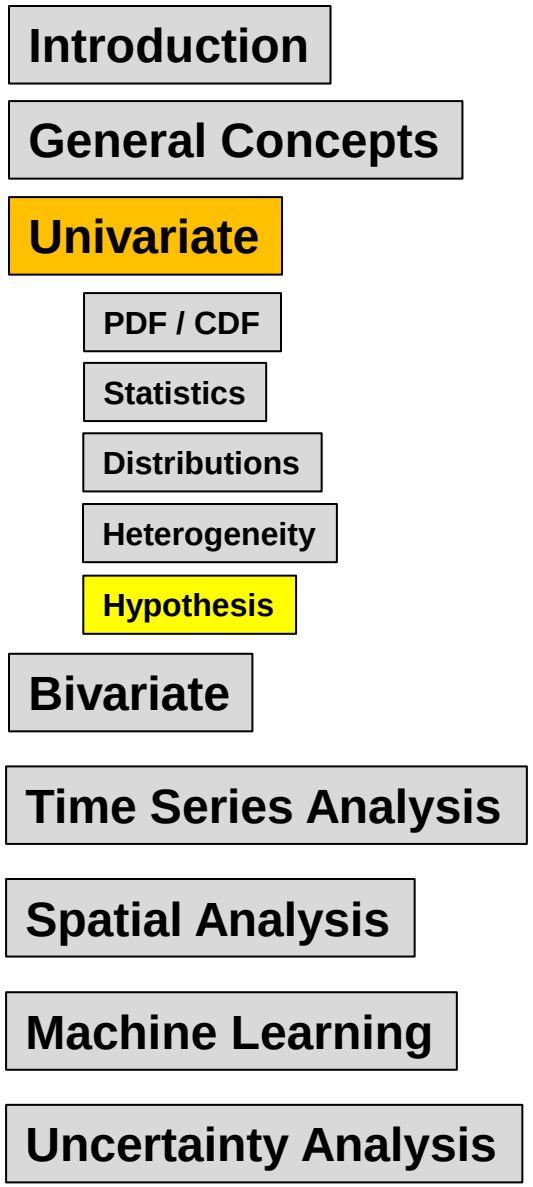
PGE 338 Data Analytics and Geostatistics

Lecture 7: Confidence Intervals for Integrating Uncertainty Models

Lecture outline . . .

- **Bootstrap Empirical Confidence Intervals**

Michael Pyrcz, The University of Texas at Austin



Confidence Intervals by Bootstrap

We can bootstrap calculate the associated distribution of uncertainty for any population parameter, and percentiles for the confidence interval

1. Pool the data, must be unbiased, independent samples
2. Resample ‘n’ (number of data) times with replacement to calculate a **realization of the data**
3. Calculate the statistic of interest (realization of the statistic) on the data realization
4. Repeat ‘L” (number of realizations) times
5. Plot the distribution (Histogram / CDF), calculate percentiles, P10 / P90 for 80% confidence interval

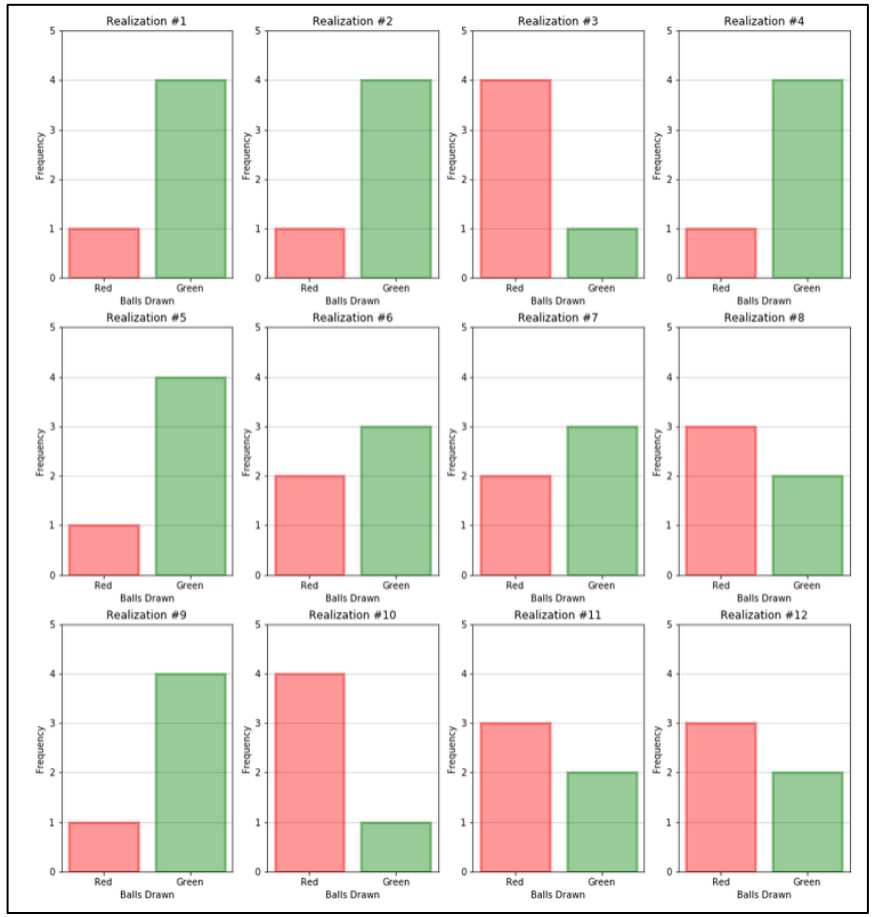
Comments:

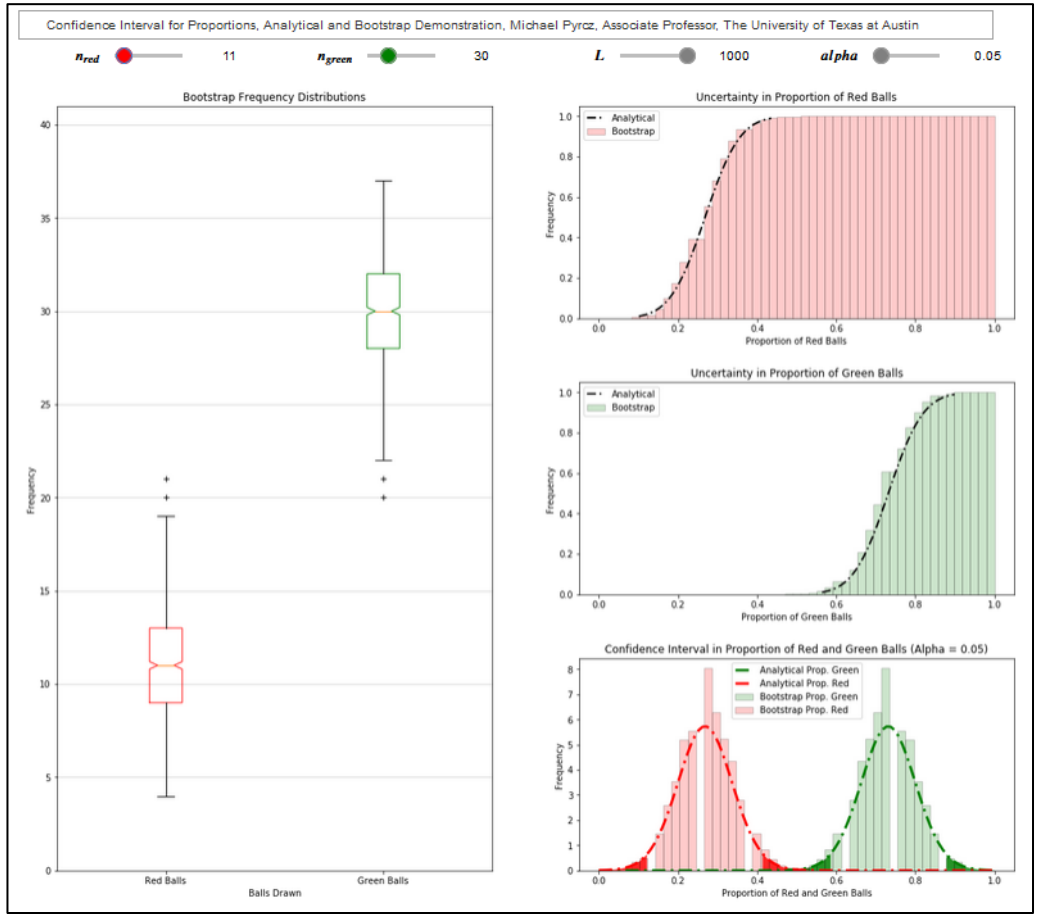
- Even if the statistic of interest is derived from other statistics, don’t stop short, calculate a realization of the statistic of interest with each bootstrap data sets.
- E.g., $dp^{\ell} = \frac{P_k(50)^{\ell} - P_k(16)^{\ell}}{P_k(50)^{\ell}}$

Build a realization of the permeability distribution, calculate P50, P16 realizations then calculate the Dykstra Parsons realization
Repeat L times…

Confidence Interval Take 3!

Interactive Python demonstration for Bootstrap Confidence Interval in Proportions





12 bootstrap realizations of number of green and red samples.

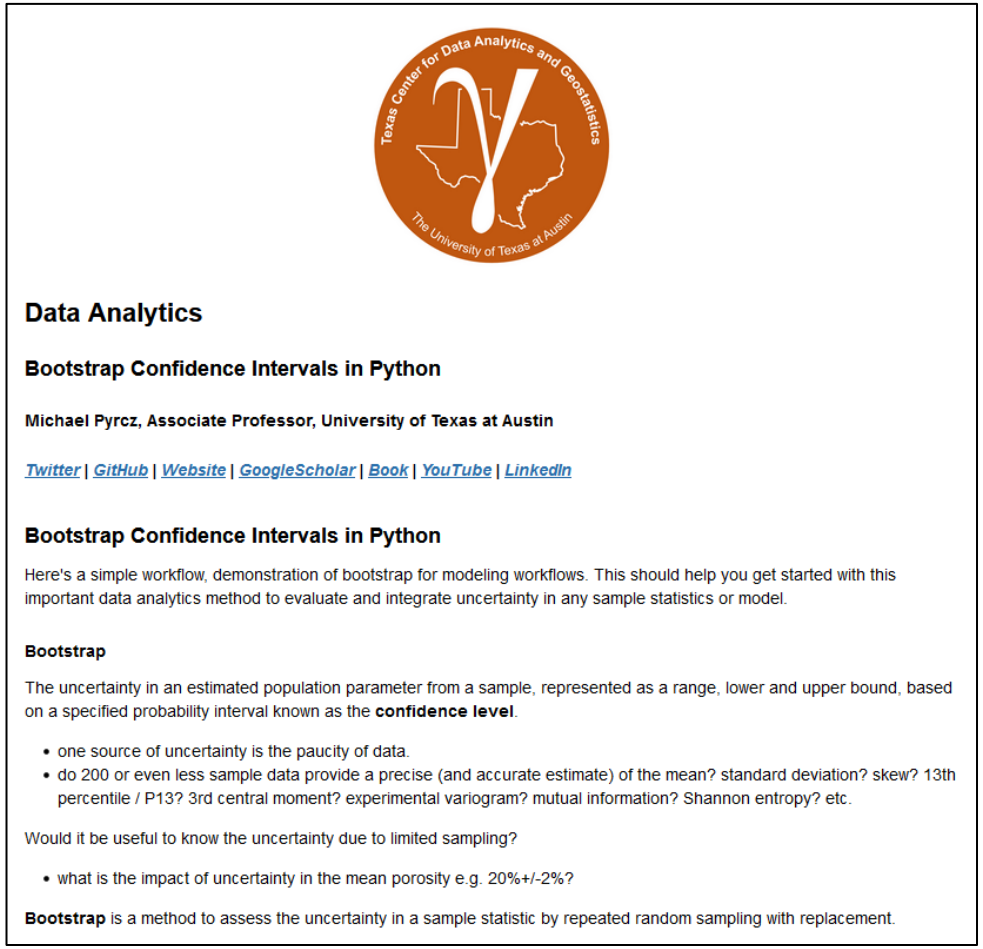
Interactive uncertainty in green and red proportions.

Jupyter notebook Python interactive demonstration ‘Interactive_Confidence_Interval.ipynb’.

Confidence Interval Take 3!

Bootstrap for confidence intervals for a variety of statistics / parameters.

- Mean / arithmetic average
- Proportion
- Interquartile Range
- Coefficient of Variation
- Correlation Coefficient



Data Analytics

Bootstrap Confidence Intervals in Python

Michael Pyrcz, Associate Professor, University of Texas at Austin

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

Bootstrap Confidence Intervals in Python

Here's a simple workflow, demonstration of bootstrap for modeling workflows. This should help you get started with this important data analytics method to evaluate and integrate uncertainty in any sample statistics or model.

Bootstrap

The uncertainty in an estimated population parameter from a sample, represented as a range, lower and upper bound, based on a specified probability interval known as the **confidence level**.

- one source of uncertainty is the paucity of data.
- do 200 or even less sample data provide a precise (and accurate estimate) of the mean? standard deviation? skew? 13th percentile / P13? 3rd central moment? experimental variogram? mutual information? Shannon entropy? etc.

Would it be useful to know the uncertainty due to limited sampling?

- what is the impact of uncertainty in the mean porosity e.g. 20%+/-2%?

Bootstrap is a method to assess the uncertainty in a sample statistic by repeated random sampling with replacement.

Jupyter notebook Python demonstration ‘PythonDataBasics_BootstrapConfidence.ipynb’.

PGE 338 Data Analytics and Geostatistics

Lecture 7: Confidence Intervals for Integrating Uncertainty Models

Lecture outline . . .

- **Examples**

Michael Pyrcz, The University of Texas at Austin

Confidence Interval Exercise 1

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

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

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

Confidence Interval Exercise 1

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

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

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

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

$$CI \rightarrow\ 0.70 \pm 0.063,\ [0.64, 0.76]$$

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

Confidence Interval Exercise 2

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

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

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

Confidence Interval Exercise 2

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

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

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

$$CI \rightarrow \hat{p} \pm t_{\frac{\alpha}{2},9}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.70 \pm 2.26\sqrt{\frac{0.7\,(1-0.7)}{200}}$$

$CI \rightarrow 0.70 \pm 0.328$, [0.37, 1.03]

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

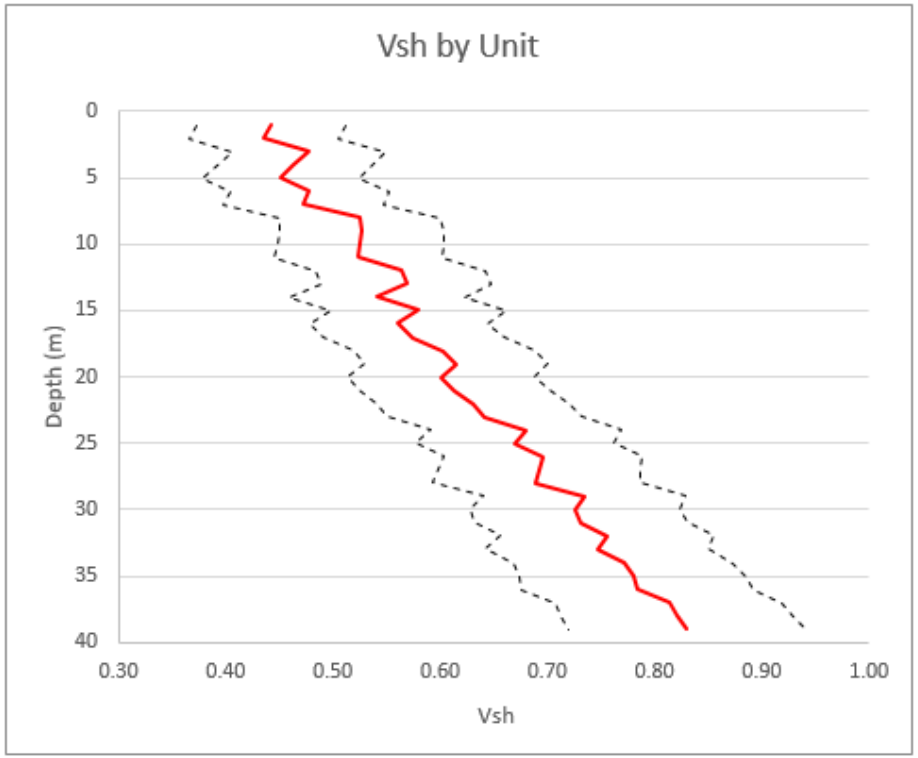
$t_{\alpha/2}$ →T.INV.2T(α,n-1) or T.INV($\alpha/2$,n-1) in Excel.

Confidence Interval
Of a Local Trend (Mean by Depth)

Another Example Confidence Interval Application

Confidence interval on a plot. We have Vsh samples over multiple depths.

Depth	Vsh Est	n	Level	z-stat	SE	Interval	P025	Expected	P975
1	0.44	196	0.05	-1.96	0.04	0.07	0.37	0.44	0.51
2	0.44	192	0.05	-1.96	0.04	0.07	0.37	0.44	0.51
3	0.48	188	0.05	-1.96	0.04	0.07	0.41	0.48	0.55
4	0.46	184	0.05	-1.96	0.04	0.07	0.39	0.46	0.54
5	0.45	180	0.05	-1.96	0.04	0.07	0.38	0.45	0.52
6	0.48	176	0.05	-1.96	0.04	0.07	0.40	0.48	0.55
7	0.47	172	0.05	-1.96	0.04	0.07	0.40	0.47	0.55
8	0.52	168	0.05	-1.96	0.04	0.08	0.45	0.52	0.60
9	0.53	164	0.05	-1.96	0.04	0.08	0.45	0.53	0.60
10	0.53	160	0.05	-1.96	0.04	0.08	0.45	0.53	0.60
11	0.52	156	0.05	-1.96	0.04	0.08	0.45	0.52	0.60
12	0.56	152	0.05	-1.96	0.04	0.08	0.48	0.56	0.64
13	0.57	148	0.05	-1.96	0.04	0.08	0.49	0.57	0.65
14	0.54	144	0.05	-1.96	0.04	0.08	0.46	0.54	0.62
15	0.58	140	0.05	-1.96	0.04	0.08	0.50	0.58	0.66
16	0.56	136	0.05	-1.96	0.04	0.08	0.48	0.56	0.64
17	0.57	132	0.05	-1.96	0.04	0.08	0.49	0.57	0.66
18	0.60	128	0.05	-1.96	0.04	0.08	0.52	0.60	0.69
19	0.61	124	0.05	-1.96	0.04	0.09	0.53	0.61	0.70
20	0.60	120	0.05	-1.96	0.04	0.09	0.51	0.60	0.69
21	0.61	116	0.05	-1.96	0.05	0.09	0.52	0.61	0.70
22	0.63	112	0.05	-1.96	0.05	0.09	0.54	0.63	0.72

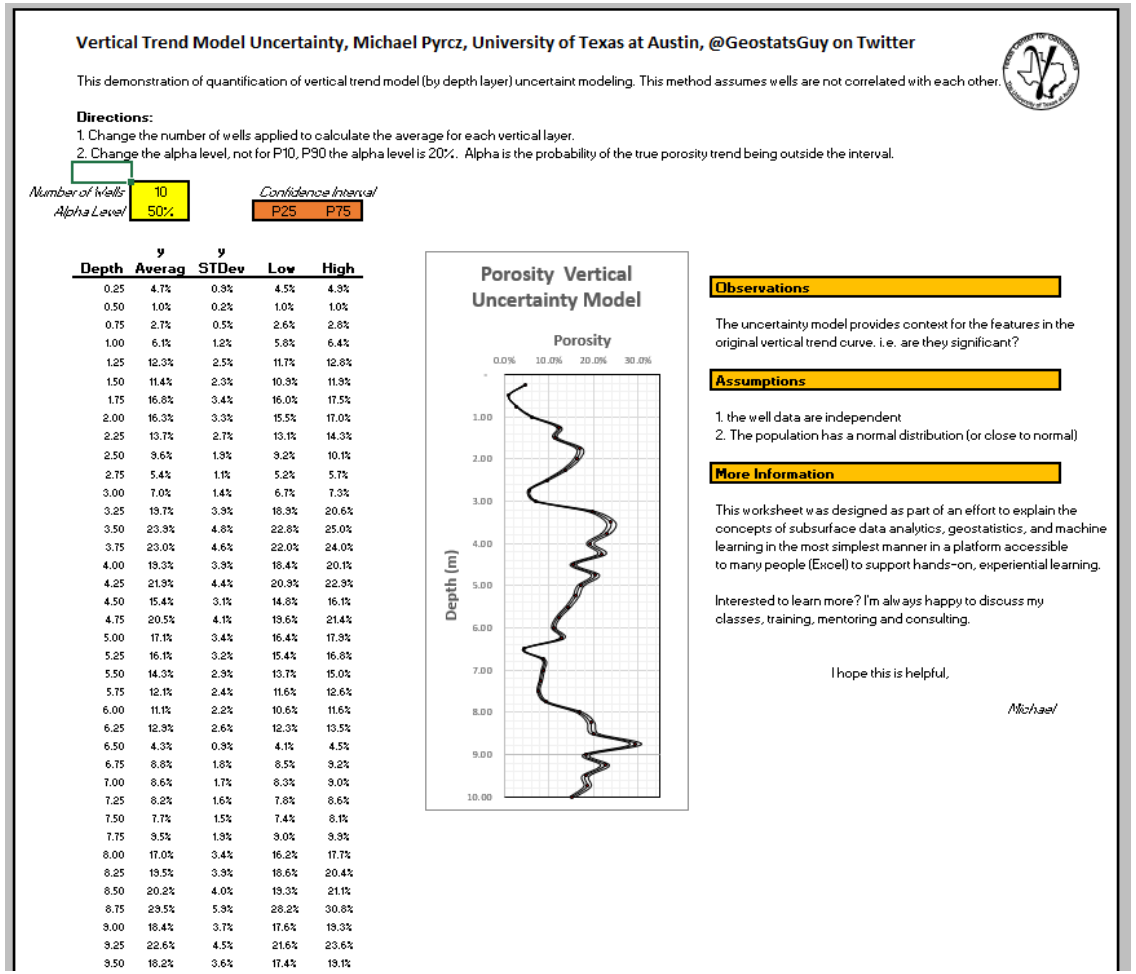


Vsh vs. Depth data (left) and plot (right) with 95% confidence interval, file is Vertical_Trend_Uncertainty_Demo.xlsx.

Confidence Interval Demonstration in Excel

Confidence intervals for uncertainty in a depth trend for fraction of shale (Vsh).

- We have Vsh samples over multiple depths.
- Change number of wells and alpha level.



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: 10
Alpha Level: 50%

Confidence Interval: P25 P75

Depth	y Averag	y STDev	Low	High
0.25	4.7%	0.3%	4.5%	4.9%
0.50	1.0%	0.2%	1.0%	1.0%
0.75	2.7%	0.5%	2.6%	2.8%
1.00	6.1%	1.2%	5.8%	6.4%
1.25	12.3%	2.5%	11.7%	12.8%
1.50	11.4%	2.3%	10.9%	11.9%
1.75	16.8%	3.4%	16.0%	17.5%
2.00	16.3%	3.3%	15.5%	17.0%
2.25	13.7%	2.7%	13.1%	14.3%
2.50	9.6%	1.9%	9.2%	10.1%
2.75	5.4%	1.1%	5.2%	5.7%
3.00	7.0%	1.4%	6.7%	7.3%
3.25	19.7%	3.9%	18.9%	20.6%
3.50	23.9%	4.8%	22.8%	25.0%
3.75	23.0%	4.6%	22.0%	24.0%
4.00	19.3%	3.9%	18.4%	20.1%
4.25	21.9%	4.4%	20.9%	22.9%
4.50	15.4%	3.1%	14.8%	16.1%
4.75	20.5%	4.1%	19.6%	21.4%
5.00	17.1%	3.4%	16.4%	17.9%
5.25	16.1%	3.2%	15.4%	16.8%
5.50	14.3%	2.9%	13.7%	15.0%
5.75	12.1%	2.4%	11.6%	12.6%
6.00	11.1%	2.2%	10.6%	11.6%
6.25	12.9%	2.6%	12.3%	13.5%
6.50	4.3%	0.9%	4.1%	4.5%
6.75	8.8%	1.8%	8.5%	9.2%
7.00	8.6%	1.7%	8.3%	9.0%
7.25	8.2%	1.6%	7.8%	8.6%
7.50	7.7%	1.5%	7.4%	8.1%
7.75	9.5%	1.9%	9.0%	9.9%
8.00	17.0%	3.4%	16.2%	17.7%
8.25	19.5%	3.9%	18.6%	20.4%
8.50	20.2%	4.0%	19.3%	21.1%
8.75	29.5%	5.9%	28.2%	30.8%
9.00	18.4%	3.7%	17.6%	19.3%
9.25	22.6%	4.5%	21.6%	23.6%
9.50	18.2%	3.6%	17.4%	19.1%

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

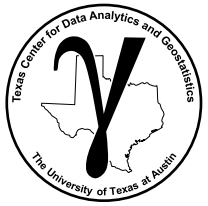
This worksheet was designed as part of an effort to explain the concepts of subsurface data analytics, geostatistics, and machine learning in the most simplest manner in a platform accessible to many people (Excel) to support hands-on, experiential learning.

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

I hope this is helpful,

Michael

Vsh vs. Depth confidence interval workflow, file is Vertical_Trend_Uncertainty_Demo.xlsx.



Conclusion

We need to report uncertainty with respect to the inferred population parameters!

- In some cases, we have analytical solutions.
- We can bootstrap to get any confidence interval.

PGE 338 Data Analytics and Geostatistics

Lecture 7: Confidence Intervals for Integrating Uncertainty Models

Lecture outline . . .

- **Concepts**
- **Analytical Confidence Intervals**
- **Bootstrap Empirical Confidence Intervals**
- **Examples**

Michael Pyrcz, The University of Texas at Austin

- Introduction
- General Concepts
- Univariate
  - PDF / CDF
  - Statistics
  - Distributions
  - Heterogeneity
  - Hypothesis
- Bivariate
- Time Series Analysis
- Spatial Analysis
- Machine Learning
- Uncertainty Analysis