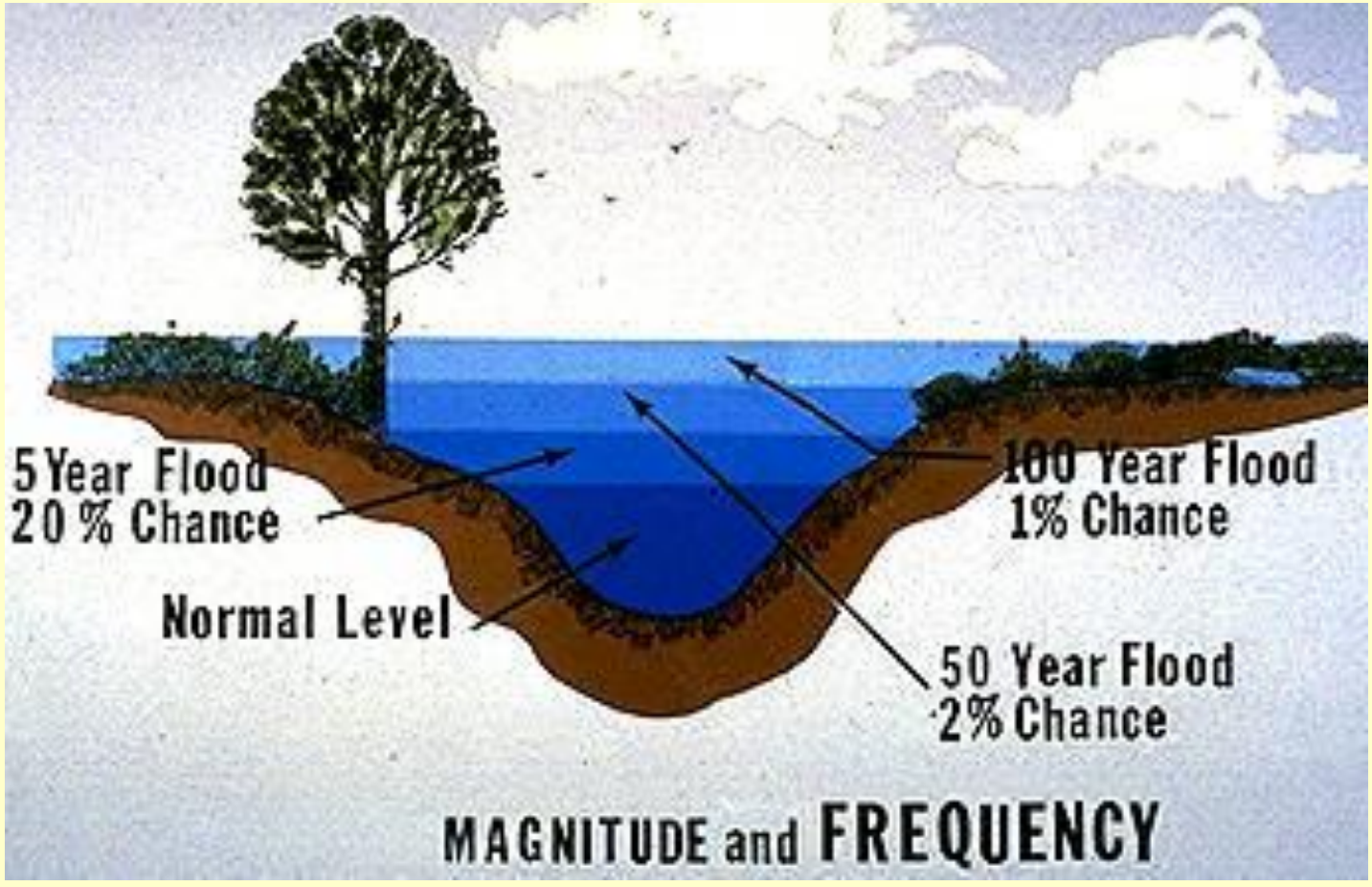
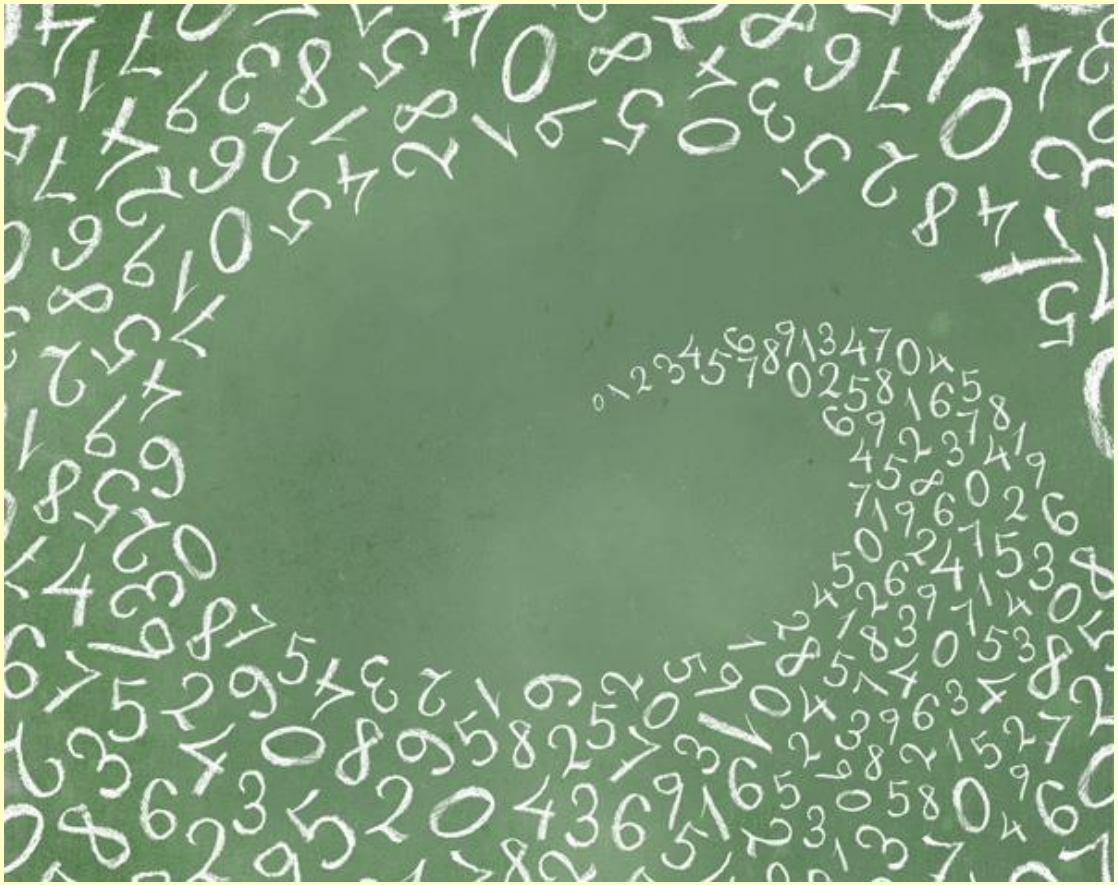
Estimation & Statistical Modelling



http://www.pelagicos.net/classes_biometry_fa16.htm

Step 1: Select the Variable



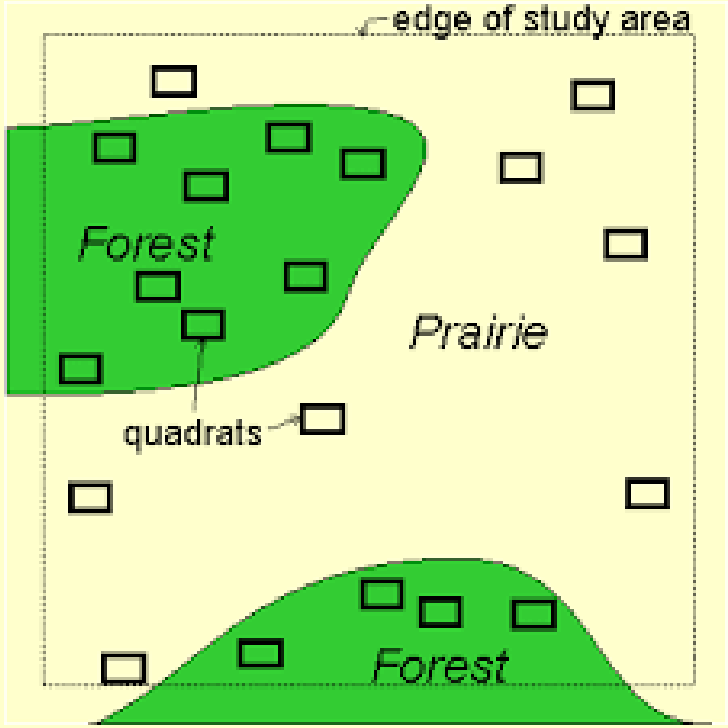
Definition: Anything that can be measured … and can differ across entities or over time

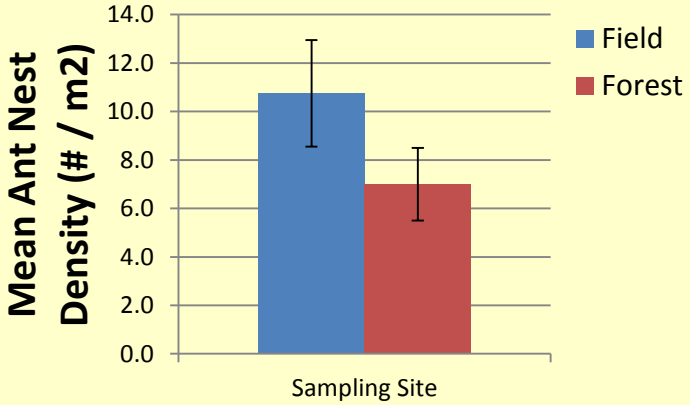
Step 2: Sample the "Biological Population"

What is the biological population (space / time) ?

How do we ensure every item has the same (and independent) probability of being selected (sampled) ?

What statistics do we use to describe the sample we collected and to estimate the population parameters ?



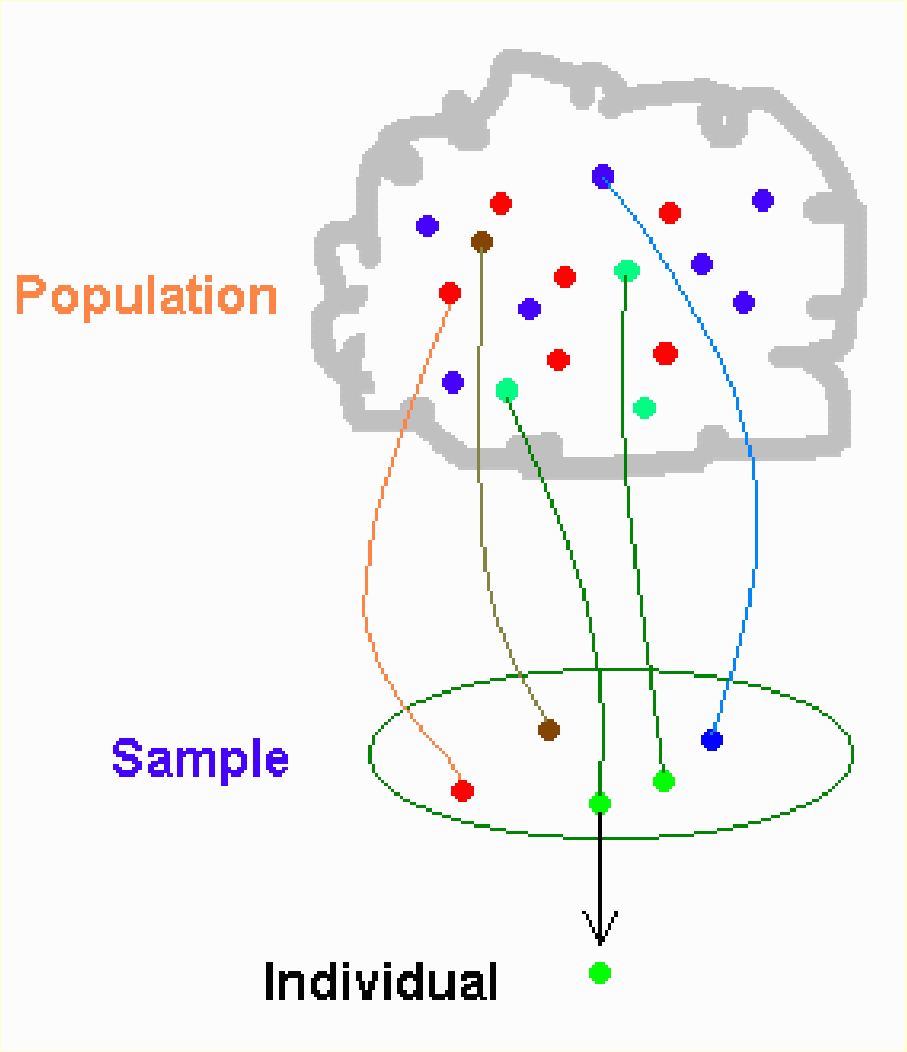


Step 3: Estimate Parameter(s)

Develop a "point estimate" (most likely or best estimate)

Develop a measure of the parameter variability around the "point estimate", which captures a given % of the likely estimate values

The user determines the level of probability captured by the "confidence interval"



An Example in Estimation

How old is your professor ?

N = 18 guesses

Range = 34 – 48

Age (yrs)
34
36
37
37
38
38
38
38
39
40
40
41
41
42
42
42
42
48

An Example in Estimation

N = 18 guesses

Mean = 39.6

Median = 39.5

S.D. = 3.1

value	frequency	relative frequency
34	1	0.056
35	0	0.000
36	1	0.056
37	2	0.111
38	4	0.222
39	1	0.056
40	2	0.111
41	2	0.111
42	4	0.222
43	0	0.000
44	0	0.000
45	0	0.000
46	0	0.000
47	0	0.000
48	1	0.056
sum	18	1

An Example in Estimation

N = 18 guesses

50% = 39.5

5% = 34

25% = 38

75% = 42

95% = 48

value	relative freq.	cumulative freq.
34	0.056	0.056
35	0.000	0.056
36	0.056	0.111
37	0.111	0.222
38	0.222	0.444
39	0.056	0.500
40	0.111	0.611
41	0.111	0.722
42	0.222	0.944
43	0.000	0.944
44	0.000	0.944
45	0.000	0.944
46	0.000	0.944
47	0.000	0.944
48	0.056	1.000
sum	1	9.389

An Example in Estimation

How old is your professor ?

N = 18 guesses

What is the Midpoint Value =

Age (yrs)
34
36
37
37
38
38
38
38
39
40
40
41
41
42
42
42
42
48

#1) Estimates Depend on Sample Size

C.I. Formulation: Mean +/- (Z score * SE)

Mean +/- (1.96 * SE)

S.E. = S.D. / sqrt (n)

n	mean	SD	sqrt(n)	SE	95% CI
3	38.3	1.5	1.7	0.9	1.7
6	40.2	4.4	2.4	1.8	3.5
9	40.1	3.5	3.0	1.2	2.3
12	39.9	3.2	3.5	0.9	1.8
15	39.7	3.0	3.9	0.8	1.5
18	39.6	3.1	4.2	0.7	1.4

#2) Estimates are influenced by chance

Age Estimate: 39.6 years (SD = 3.1)

C.I. Formulation: Mean +/- (Z score * SE)
Mean +/- (1.96 * SE)

S.E. = S.D. / sqrt (n)

n	mean	SD	sqrt(n)	SE	95% CI	lower	upper
9	40.1	3.5	3.0	1.2	2.3	37.8	42.4
9	39.1	2.8	3.0	0.9	1.8	37.3	40.9

Are these two samples from the same population ?

Summary – Statistical Estimation

Note: Information about shape (normality) of the frequency distribution is critical for estimation. Determines the statistic to use (mean or median)

Point estimates describe the most likely value of a parameter of interest, without providing an associated probability level

e.g., Mode, Median, Mean, Variance, S.D.

Confidence intervals describe the probability that the parameter estimate falls within a given range.

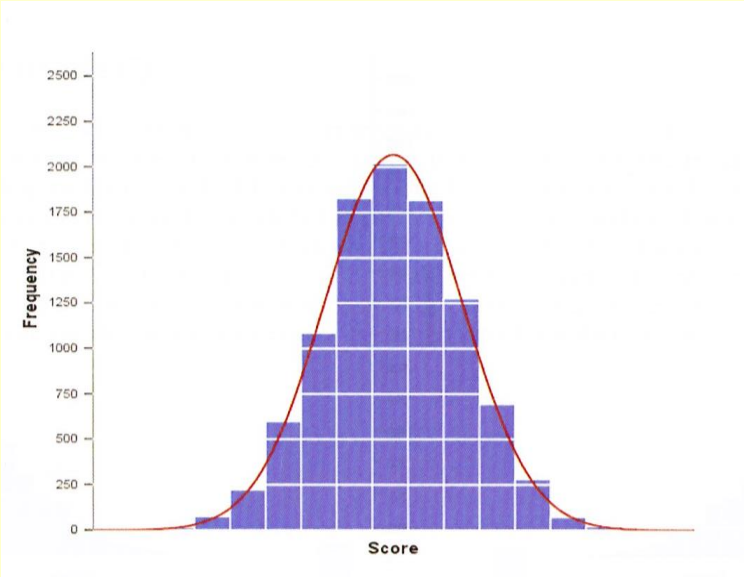
e.g., 95% commonly used – but user decides

Summary - Statistical Models

Reminder - Main goal of statistical sampling:

Calculate parameters from a sample, rather than from entire population

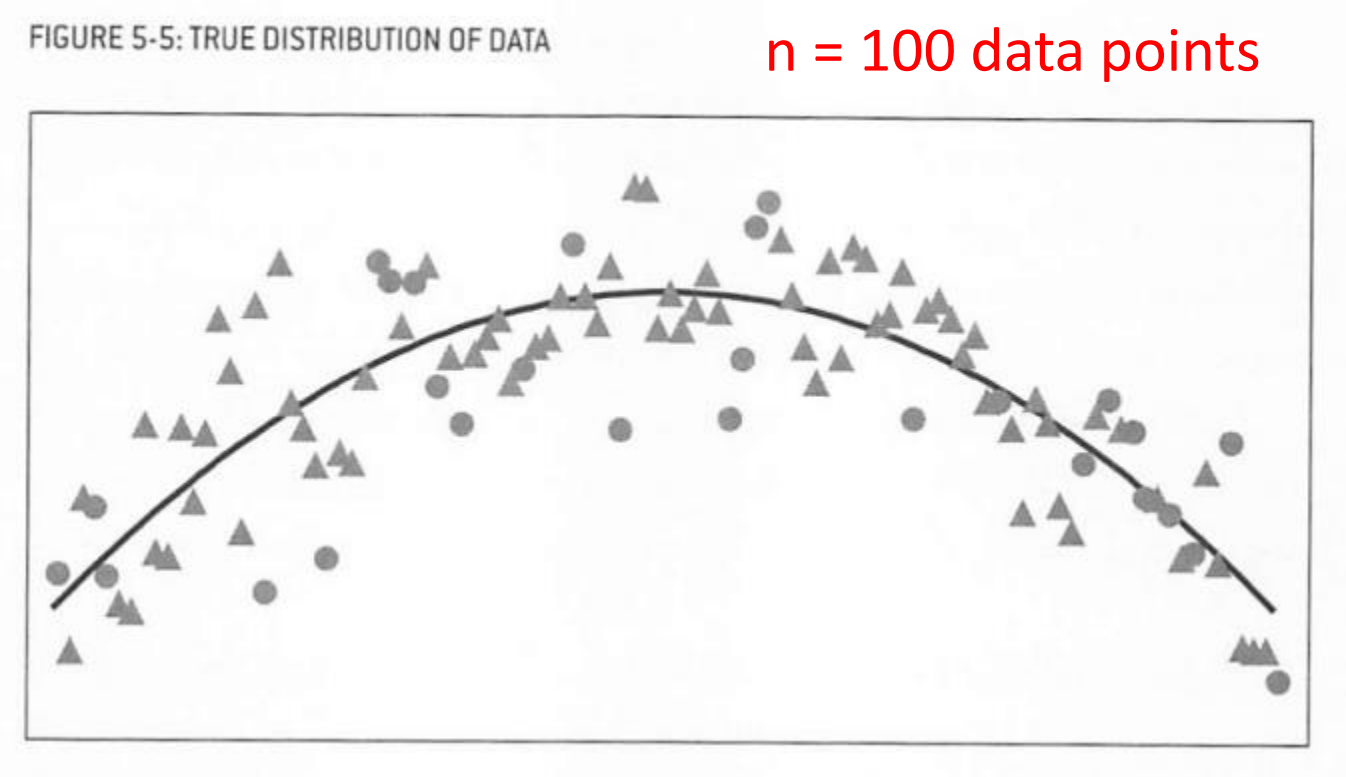
With representative sampling, we can make inferences about the entire population



Normal distributions allow to develop inferences, and to build uncertainty around estimates with CIs

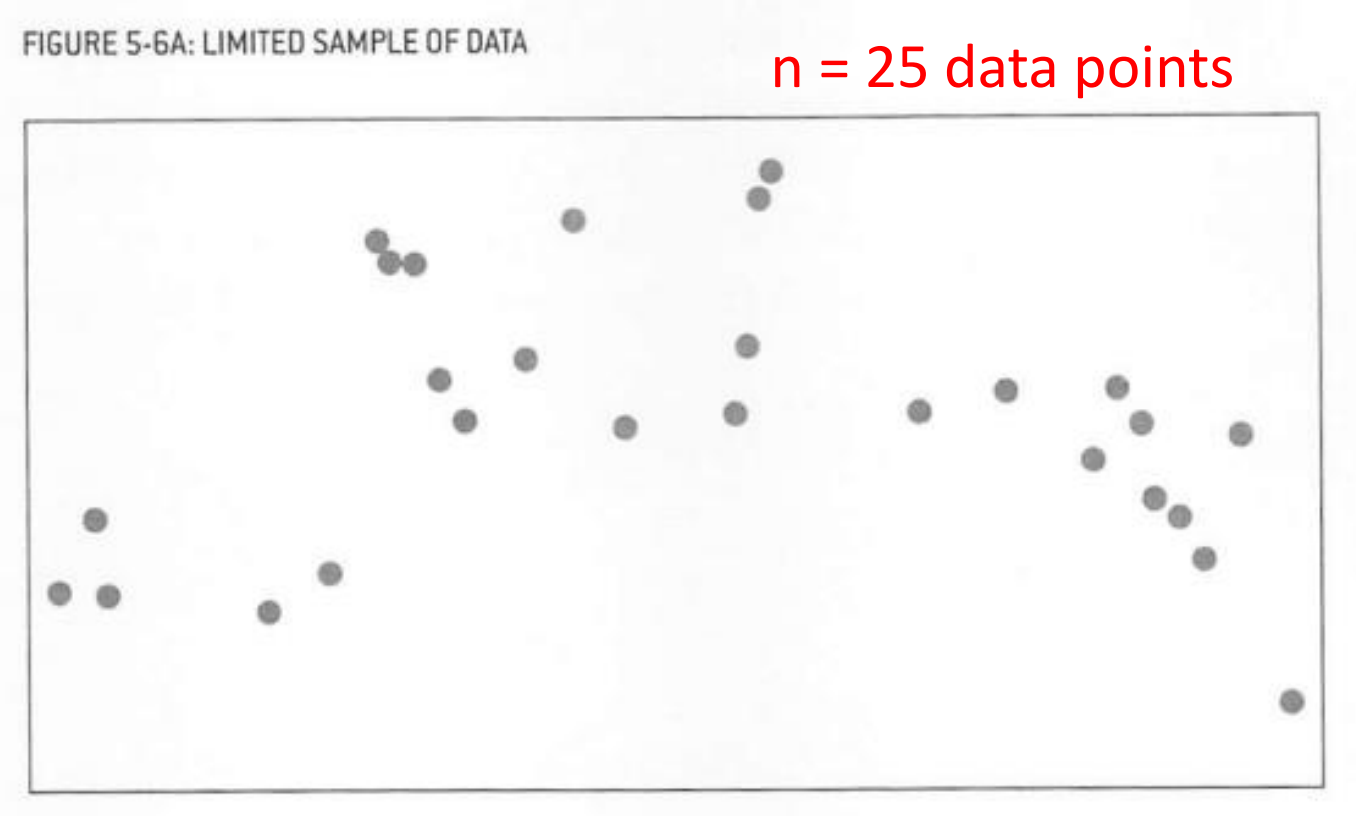
Overfitting: The Most Important Scientific Problem You've Never Heard Of In Statistics

Overfitting = In statistical analysis, the act of mistaking noise for a signal (Silver, 2012)



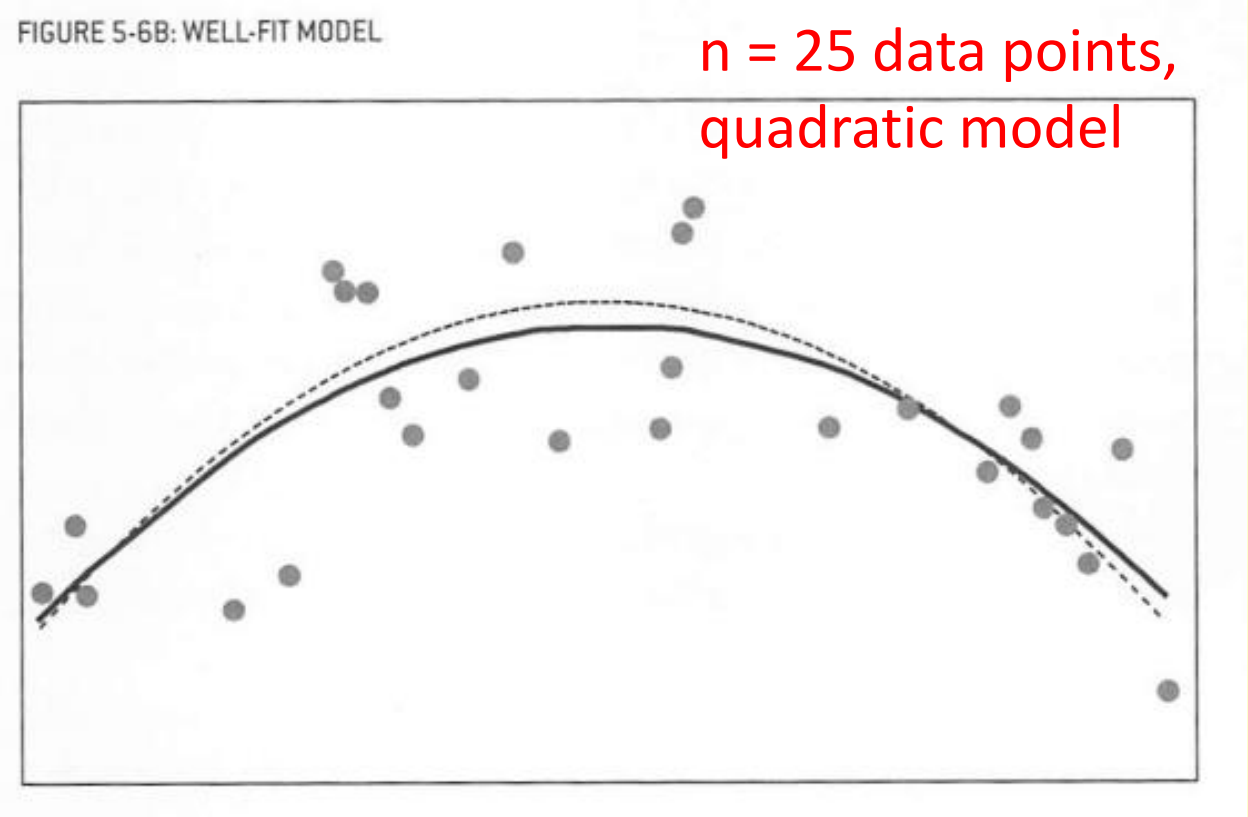
Modeling: Extrapolating from Observations to a Broader Explanation of the Data

Knowing what the real pattern is supposed to be, of course, you will still be inclined to fit the points with some kind of curve shape…



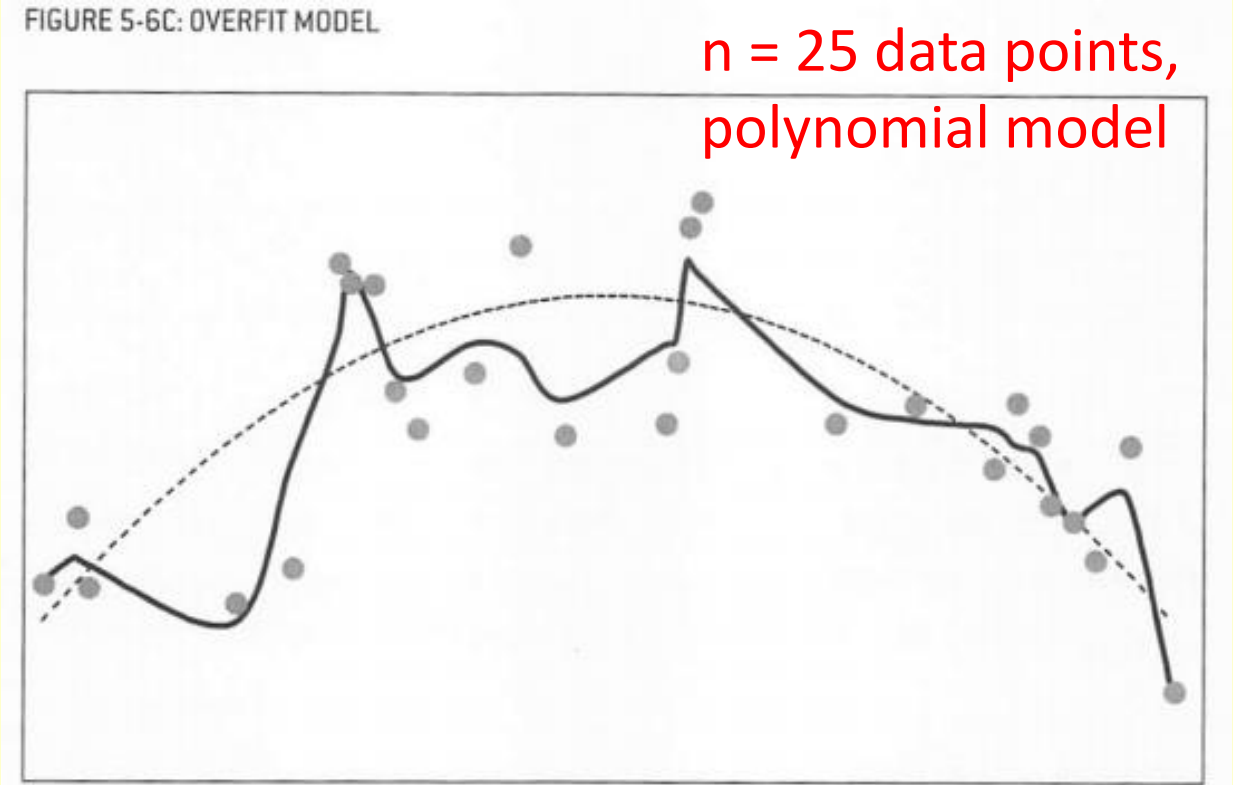
Modeling: Extrapolating from Observations to a Broader Explanation of the Data

Indeed, modeling this data with a simple mathematical expression called a quadratic equation ($y = ax2 + bx + c$) does a very good job of recreating the true relationship



Modeling: Extrapolating from Observations to a Broader Explanation of the Data

However, modeling the data with a more complicated polynomial spline, seems to do a much better job at capturing the underlying pattern. But does it improve our understanding of the processes causing the pattern ?



Overfitting:

Most likely to overfit when data are limited and noisy and when understanding of fundamental relationships is poor.

Both circumstances apply in earthquake forecasting.

Japan, despite being extremely seismically active, unprepared for devastating 2011 quake.

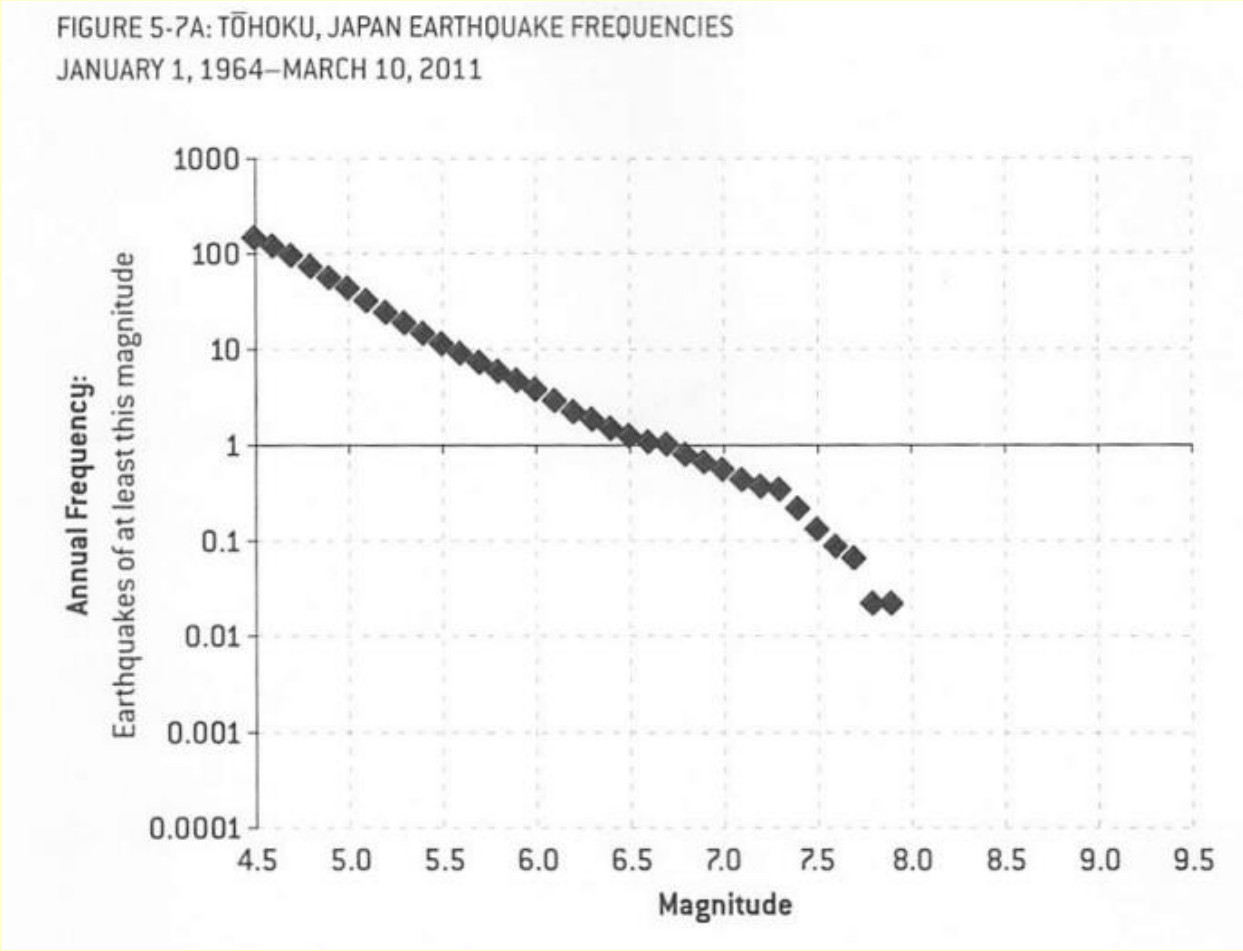
The Fukushima nuclear reactor built to withstand a magnitude 8.6 quake, but not a 9.1 quake.

Modeling Earthquake Magnitude

Relationship almost follows straight-line pattern predicted by Gutenberg and Richter's method.

But, at about 7.5 magnitude, there is a kink in the graph.

Because there had been no quakes as large 8.0 in the region since 1964, the curve bends down accordingly.

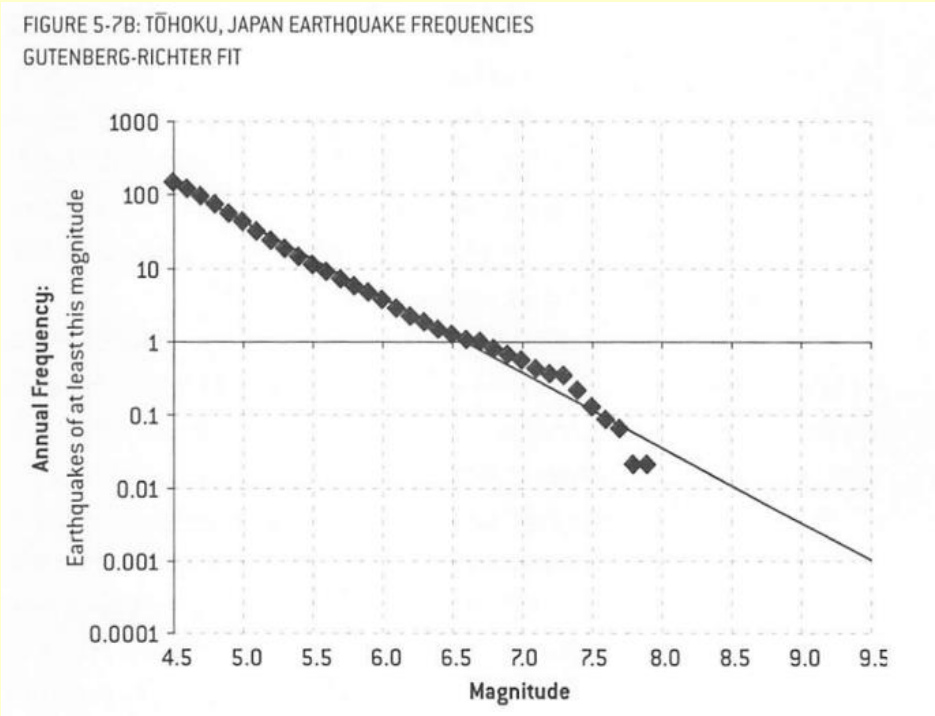


So how to connect the dots ?

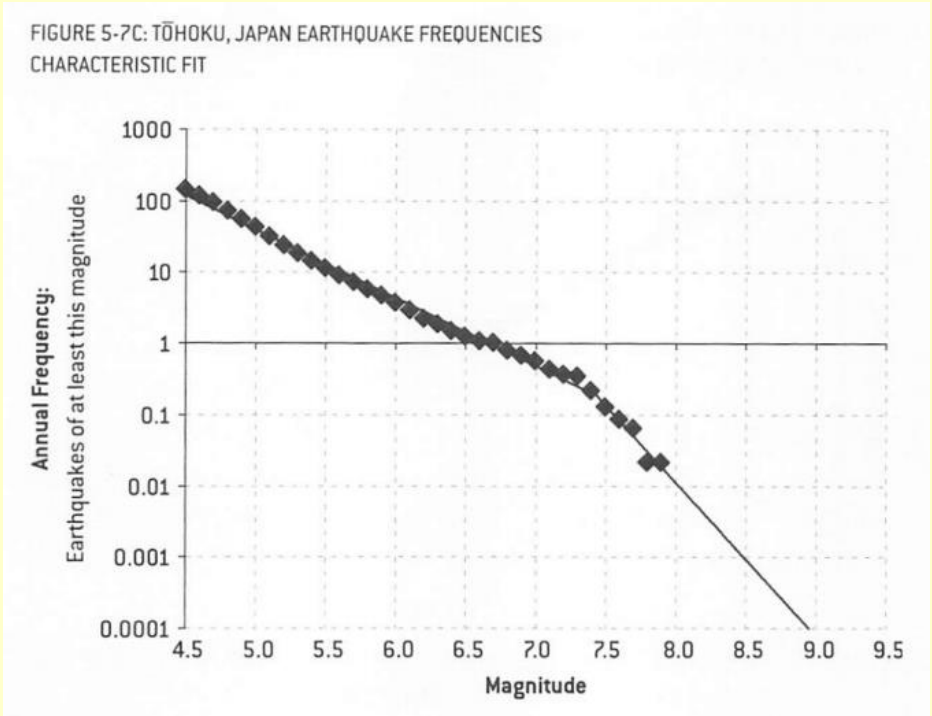
Modeling Earthquake Magnitude

Relationship almost follows straight-line pattern predicted by Gutenberg and Richter's method… but it can be modeled used two different models (with underlying assumptions):

Gutenberg & Richter Fit



Characteristic Fit



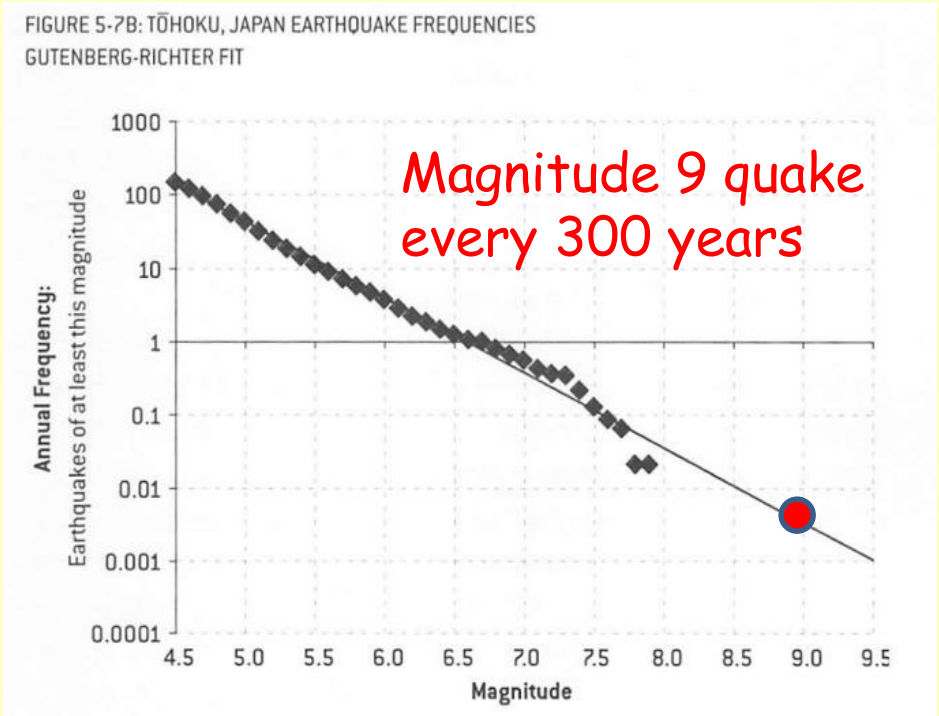
What are the Implications ?

Modeling Earthquake Magnitude

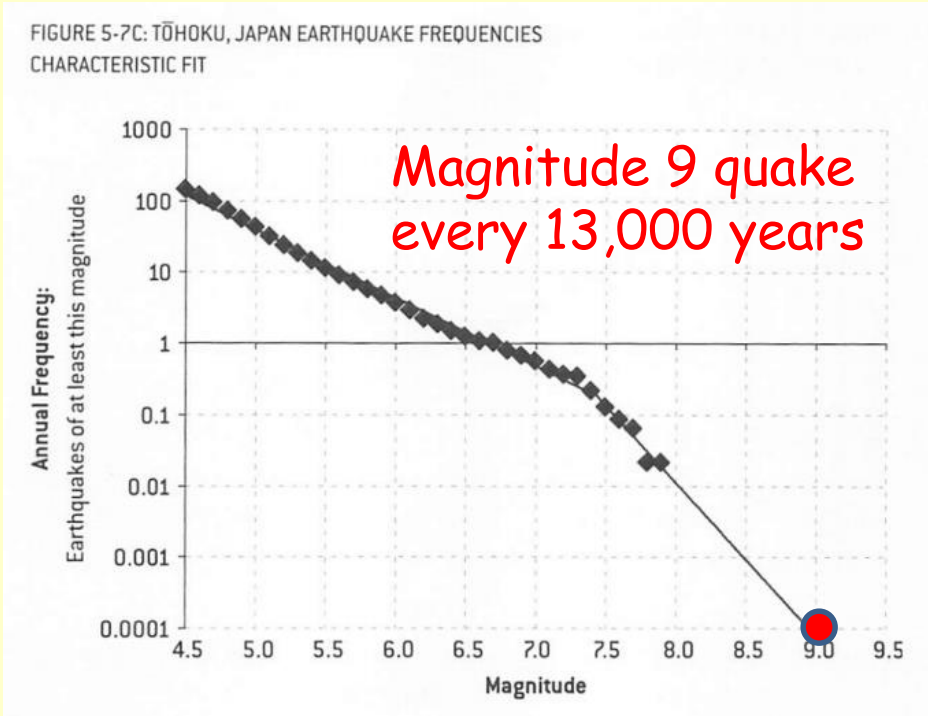
Gutenberg & Richter Law based on empirical patterns, that hold across regions of the globe and over the whole planet.

But, the characteristic fit matched the recent historical record from Tohoku better (line fits the series of points).

Gutenberg & Richter Fit



Characteristic Fit



Statistical Modeling:

Sampling allows us to develop and test models.

Models provide insights into how the world works.

Specific models have underlying assumptions.

For the rest of the class, we will develop and test models using a variety of statistical methods.