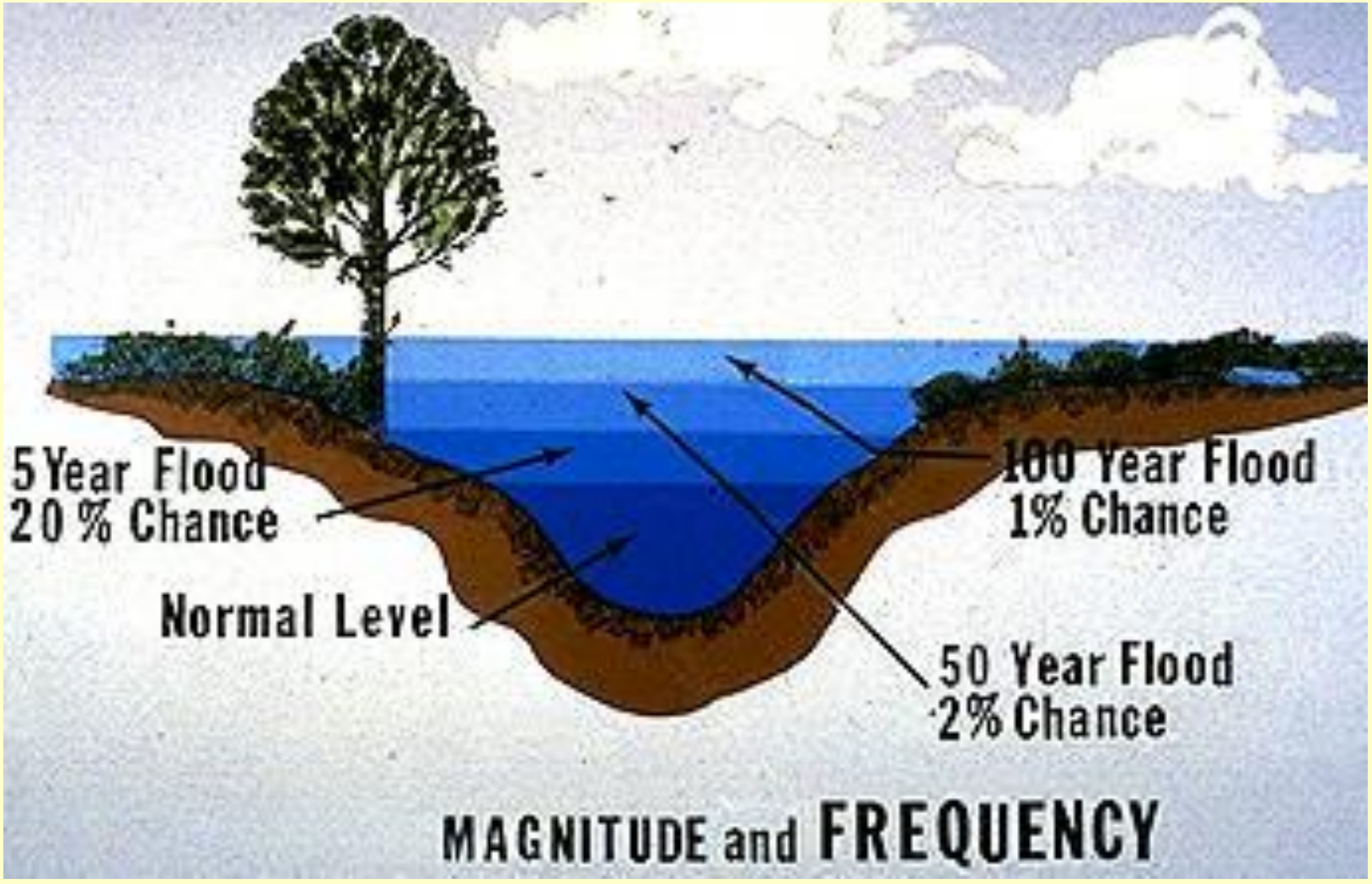
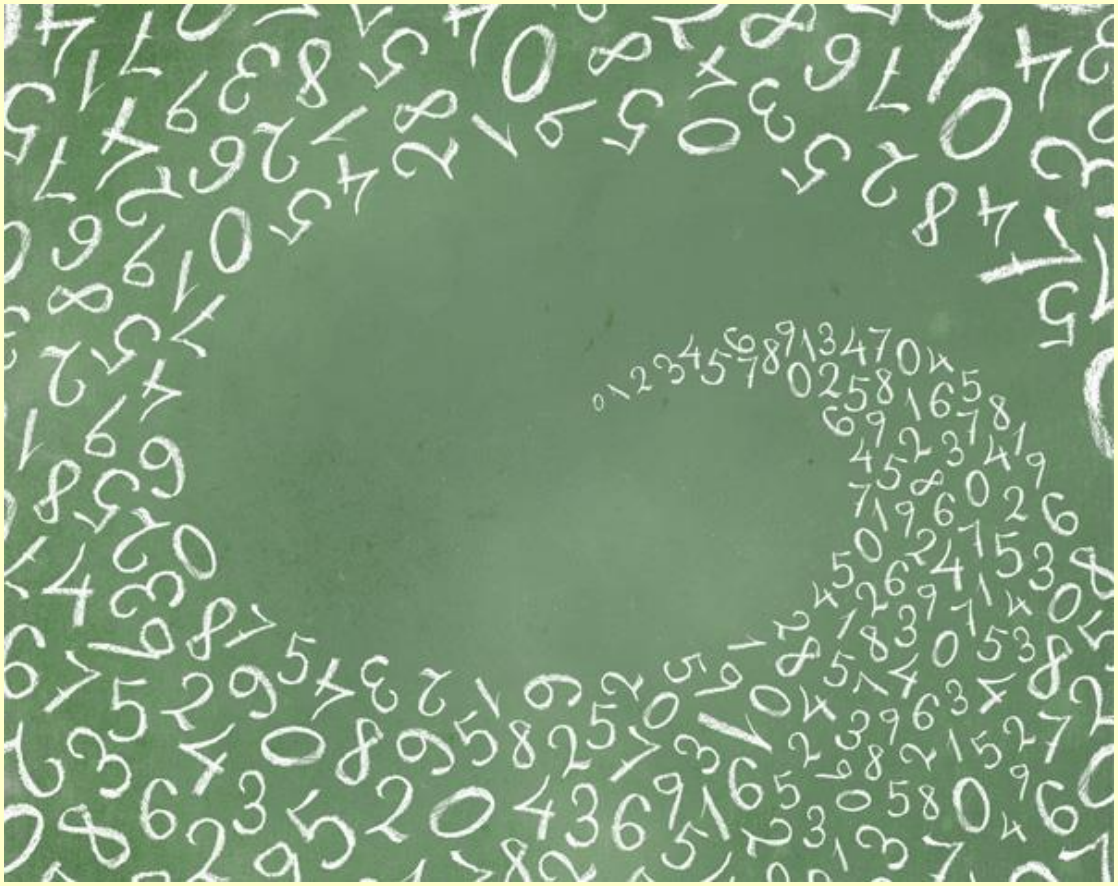
Data Summary & Estimation



Step 1: Select the Variable



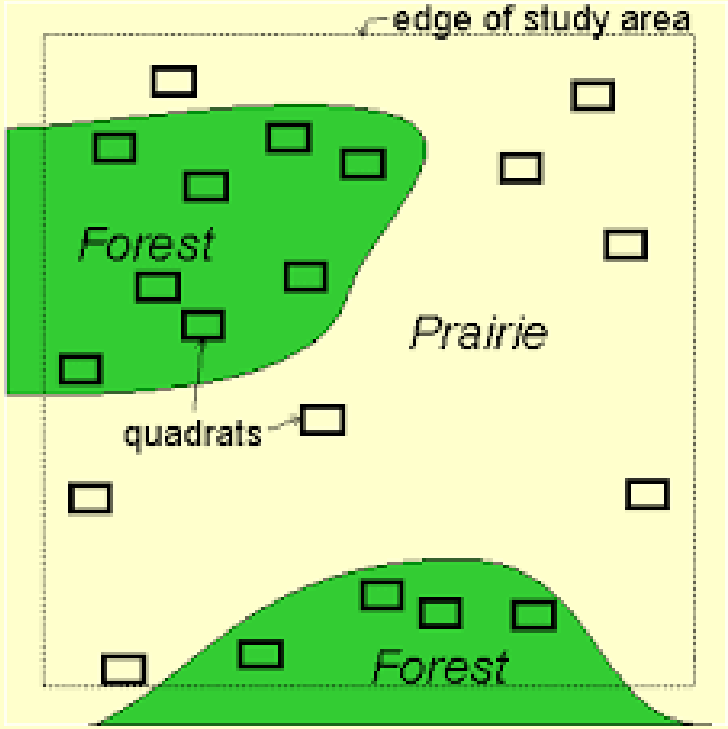
Definition: Anything that can be measured … and can differ across entities or over time

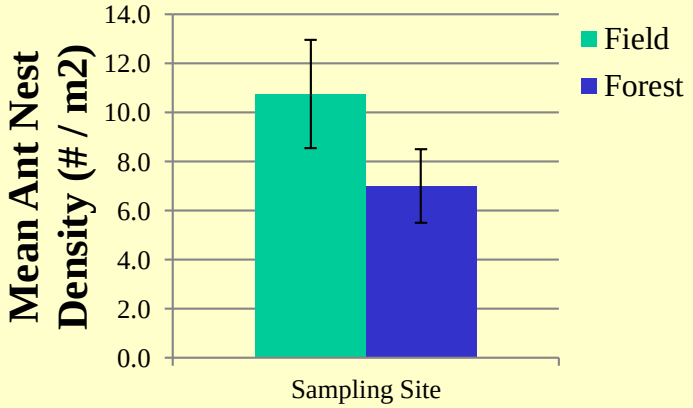
Step 2: Sample the "Biological Population"

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 ?



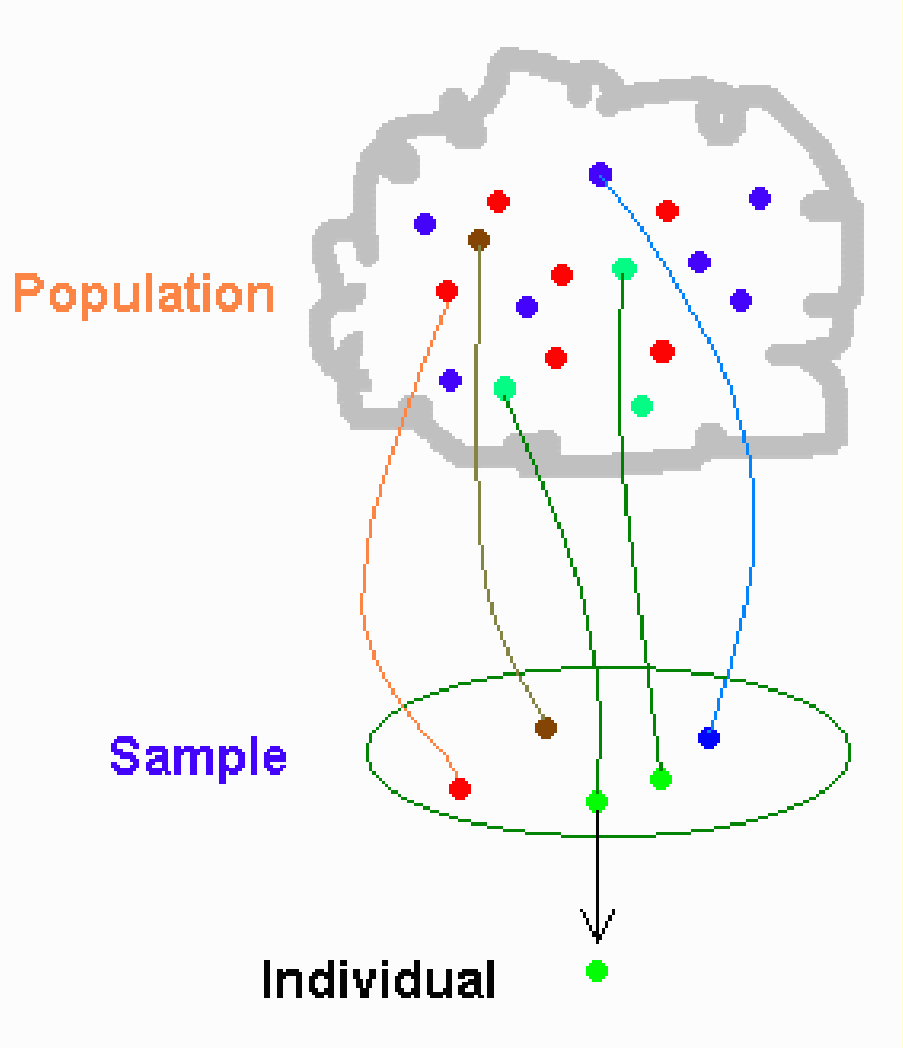


Step 3: Estimate Parameter(s)

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.

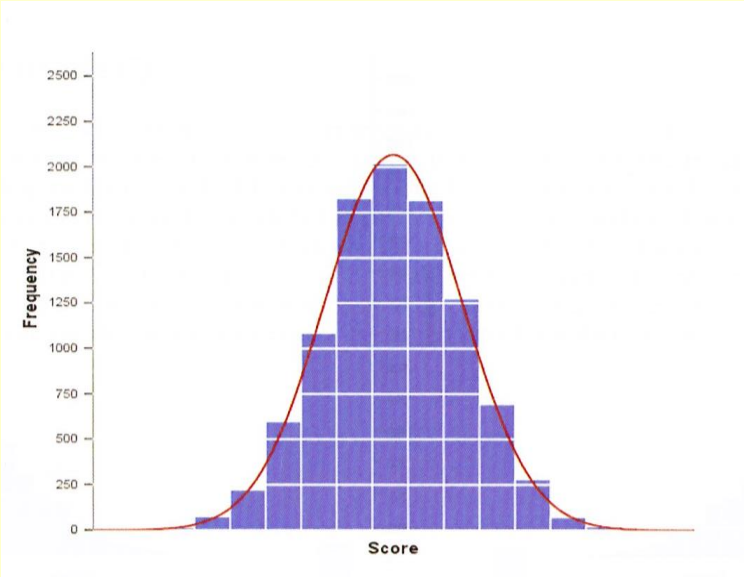
e.g., 95% commonly used – but user decides

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

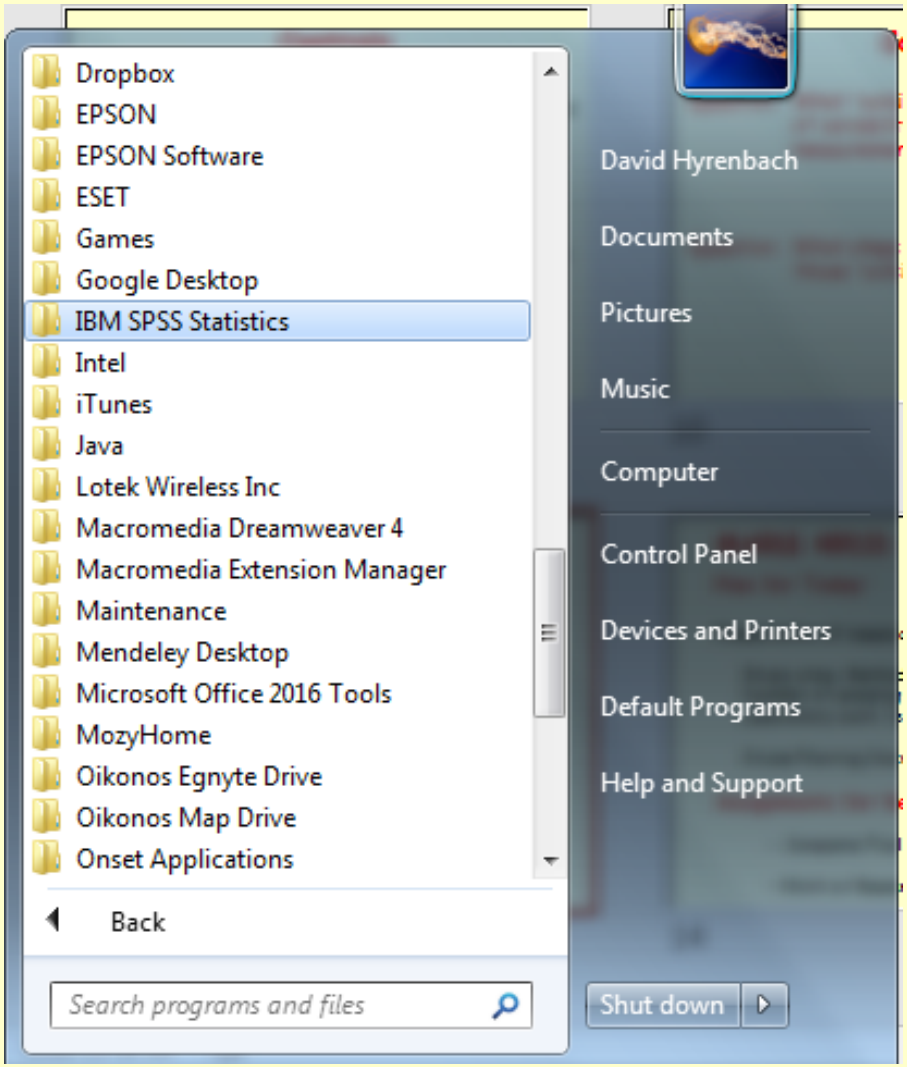


Normal distributions allow to develop inferences, and to build uncertainty around estimates with CIs

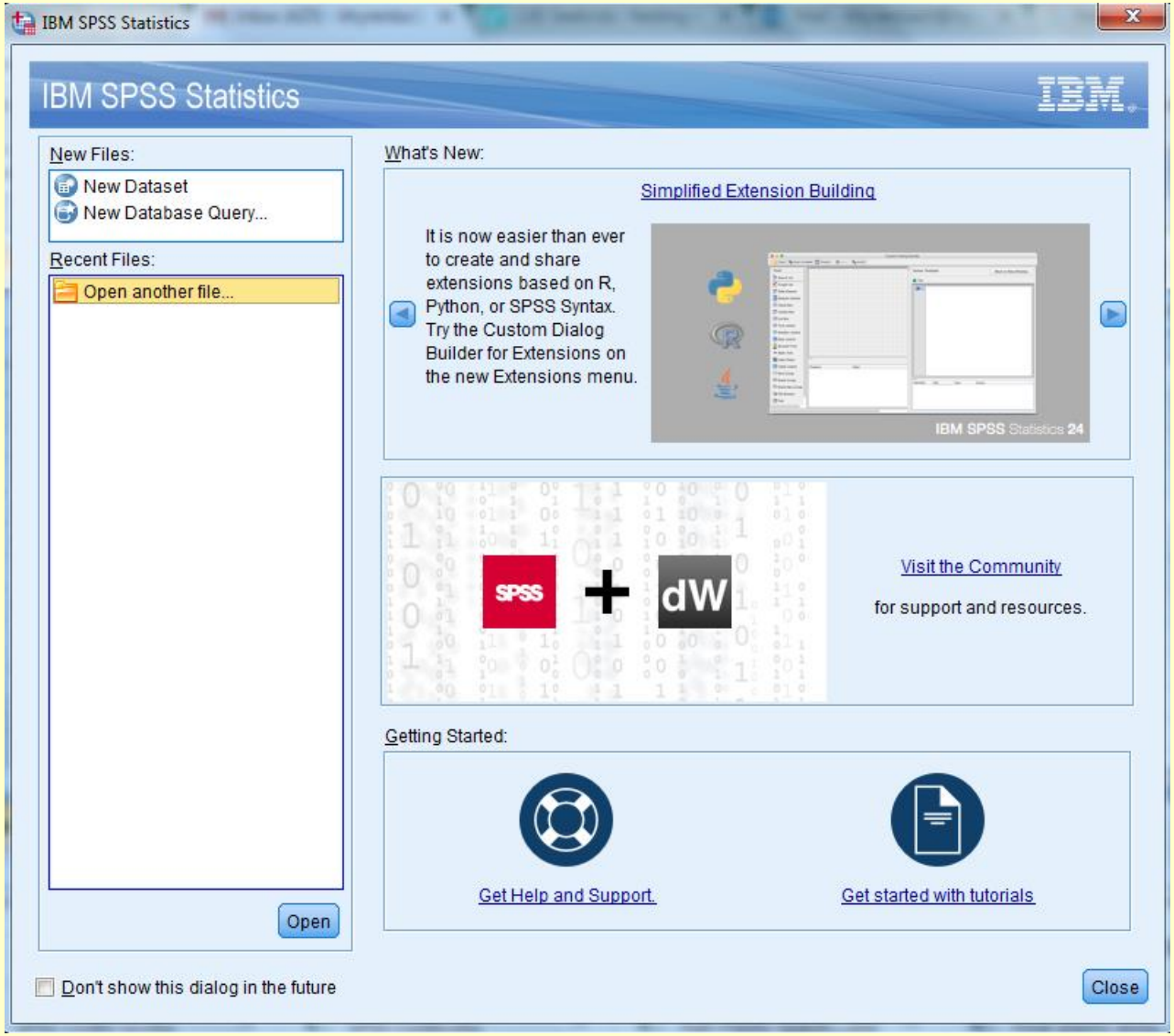
Statistical Review II: Data Summaries

SPSS®

Accessing SPSS



Loading Data

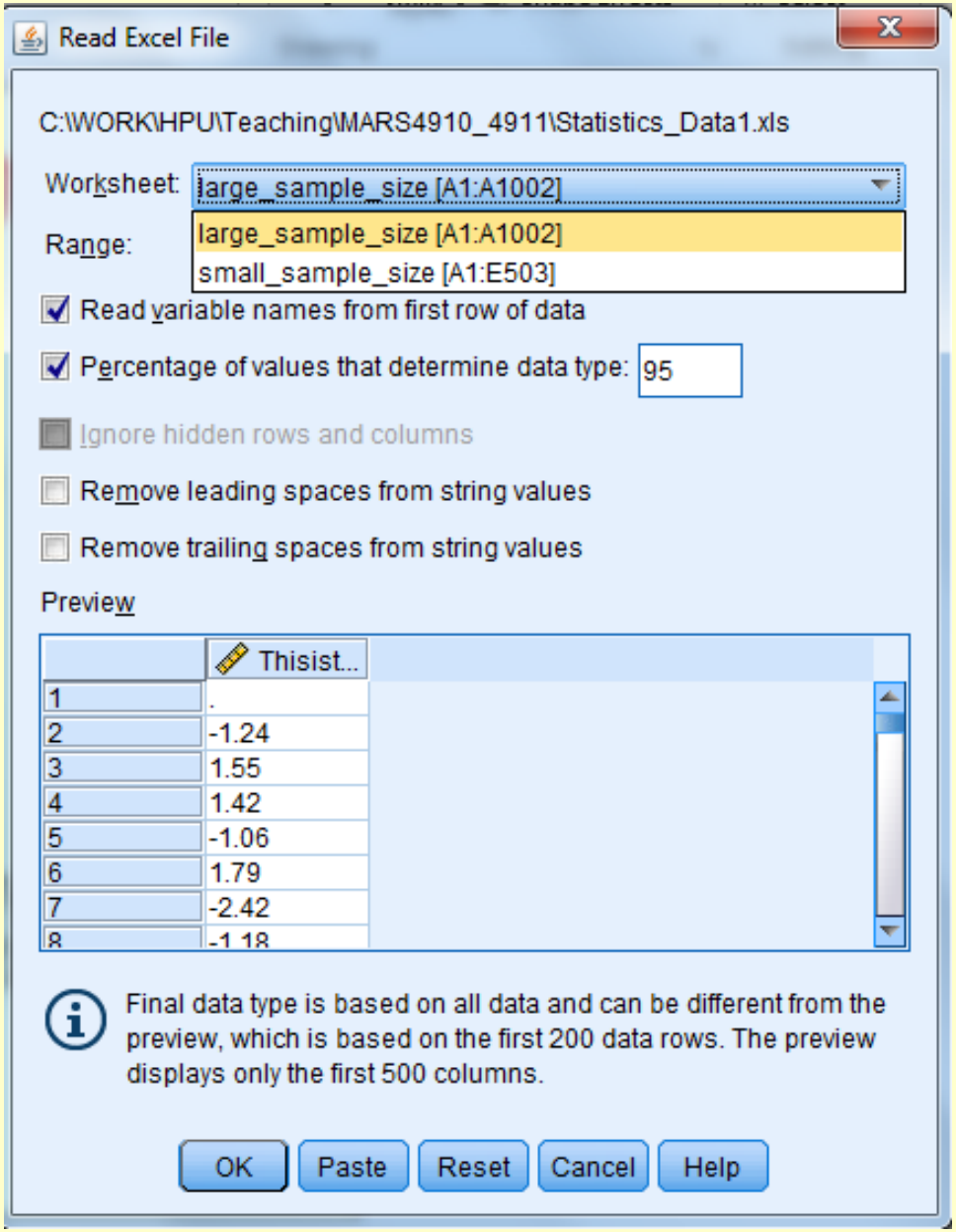


Open existing SPSS Dataset

Open Other Dataset

Select File Type (sps, xls, txt)

Loading Data



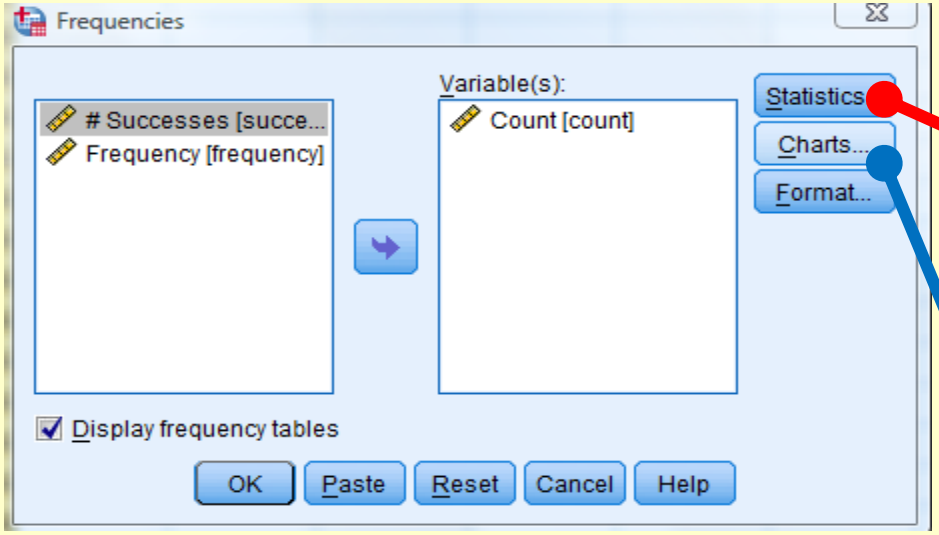
Select a worksheet (with scroll window)

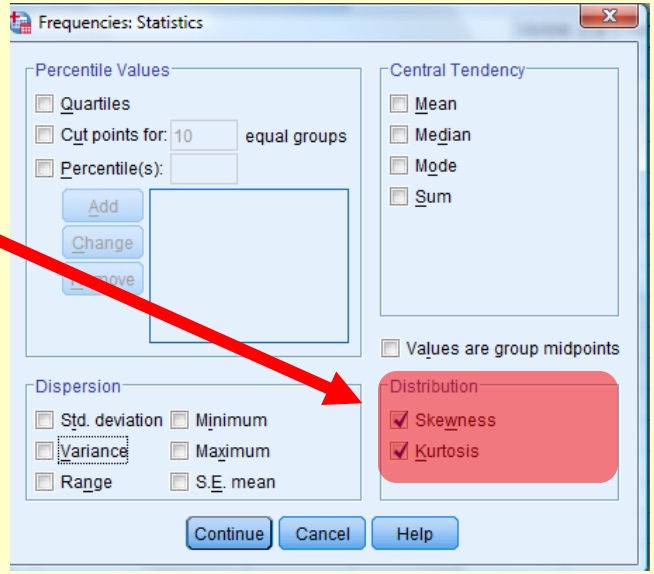
Read variables names in header: YES

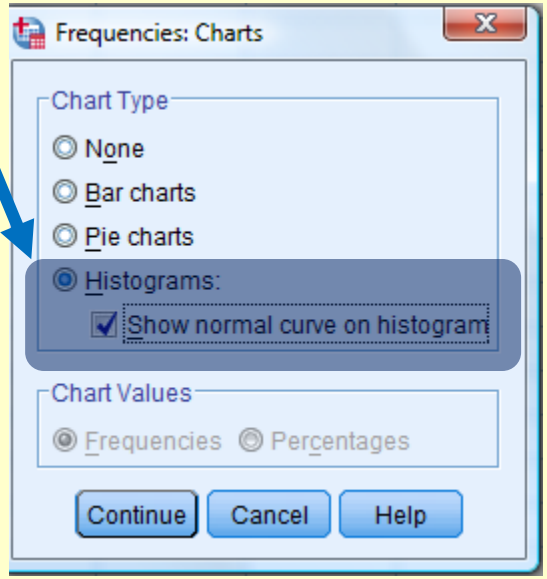
Determine data type automatically: YES

Data Preview Window

Characterizing Data Distributions







Statistics:
skewness / kurtosis

Charts: histograms
(show normal curve)

Quantifying Distributions

Distribution shapes categorized by symmetry (skew)

Skew: Measure of the symmetry of a distribution.

Symmetric distributions have a skew = 0.

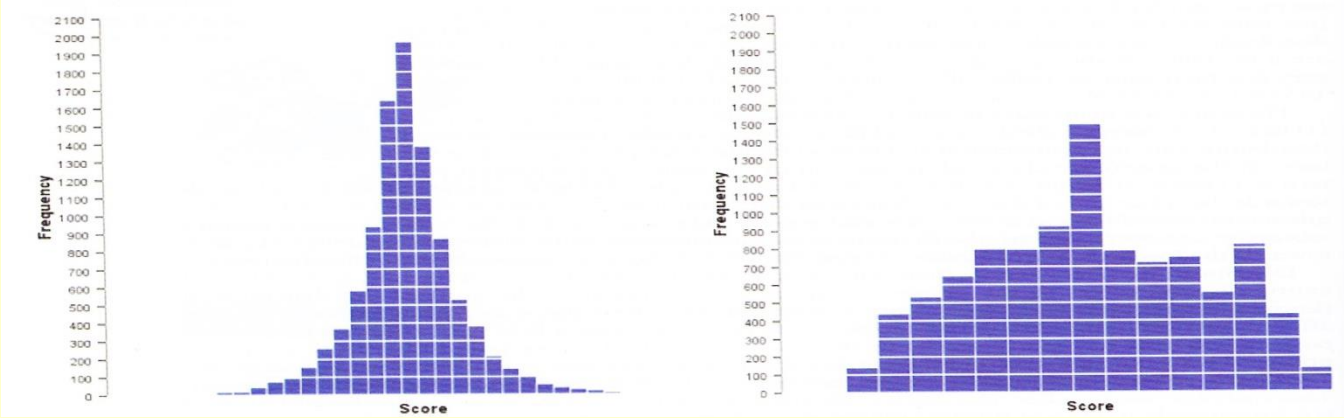
Positive skew:
the mean is larger
than the median,
skewness > 0

Negative skew:
the mean is smaller
than the median,
skewness < 0

Quantifying Distributions

Distribution shapes categorized by kurtosis

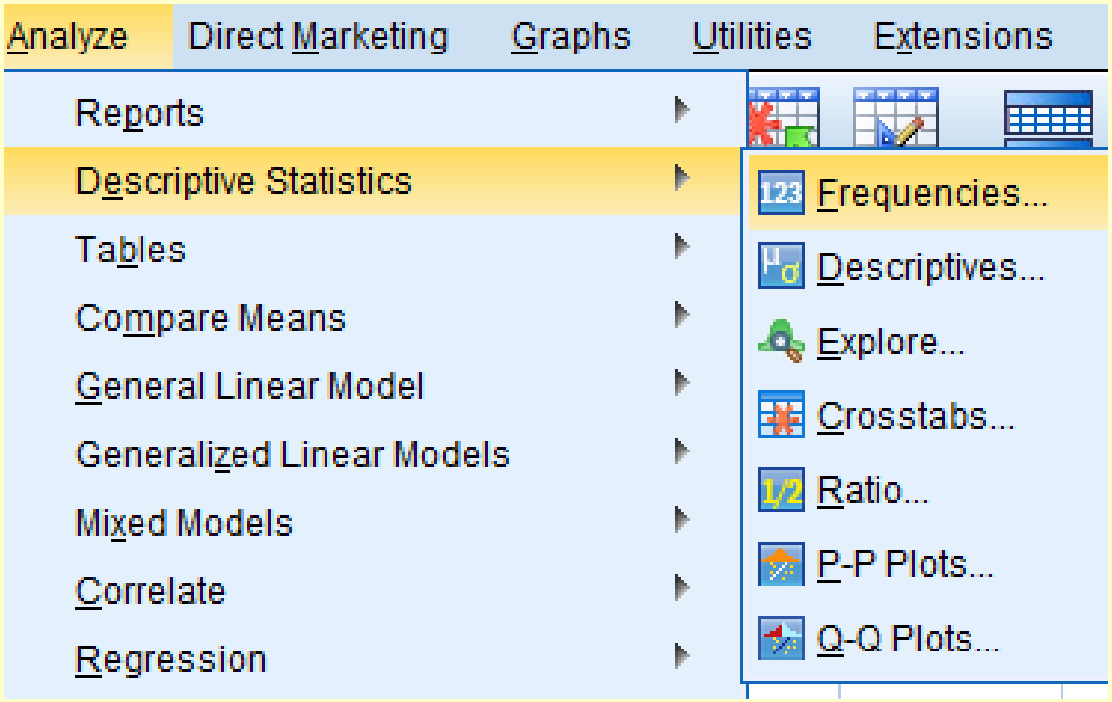
Kurtosis: Measure of the degree to which observations cluster in the tails or the center of the distribution.



Positive kurtosis:
Less values in tails and more values close to mean.
Leptokurtic.

Negative kurtosis:
More values in tails and less values close to mean.
Platykurtic.

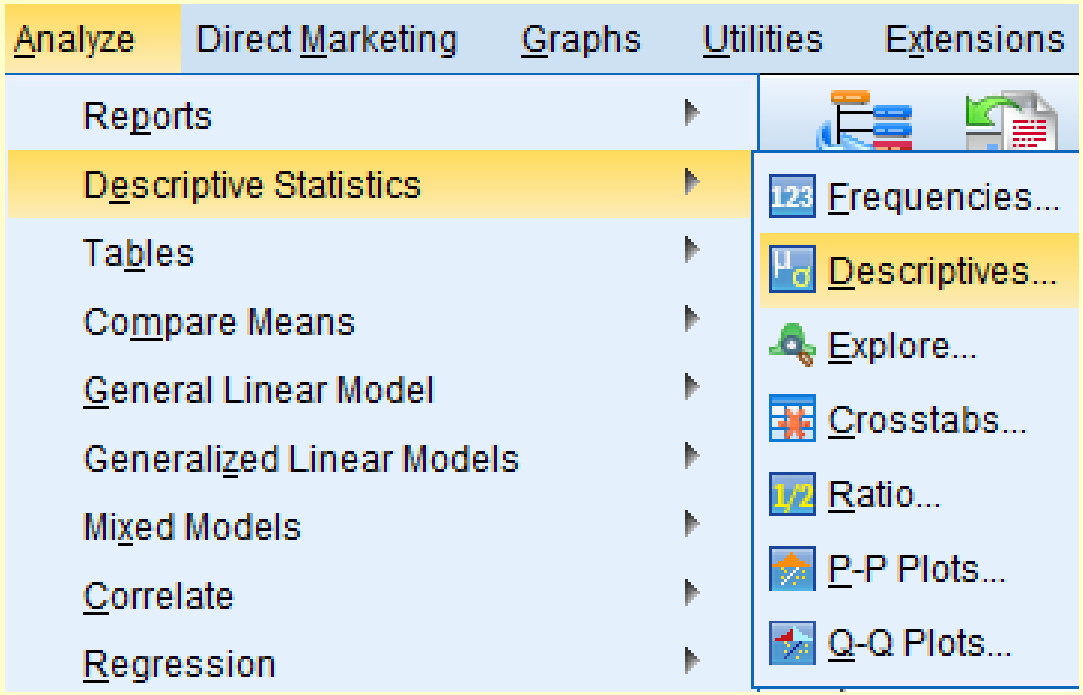
SPSS - Descriptive Statistics



Data Frequencies & Data Summaries

Histograms

SPSS – Descriptive Statistics



Data Summaries

Standardized Values (Z scores)

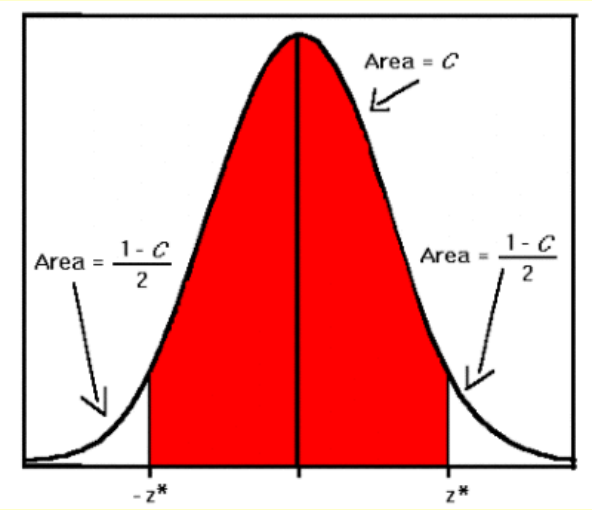
Standardized Data – from Z

Standardizing the distribution (mean= 0, S.D.= 1), develops a random normal distribution:

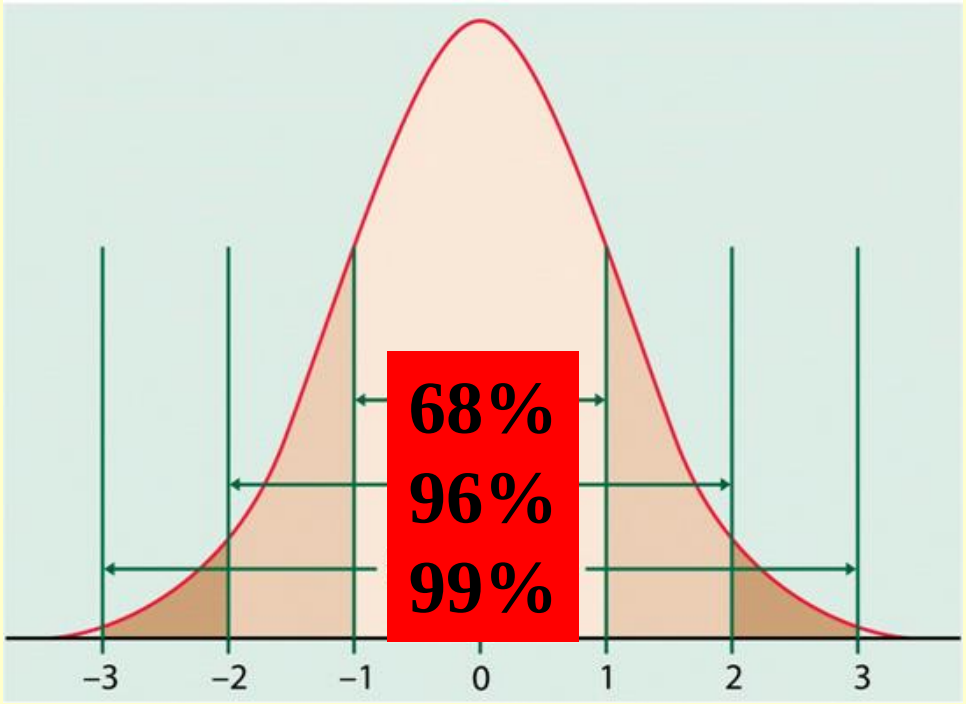
Z Score:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Thus, possible to find two values, between which Z lies with a given probability (usually 0.95).



Is the Basis of Parametric Statistics



Parametric statistical methods require that numerical variables approximate a **normal distribution**.

They compare the **means & S.D.s**

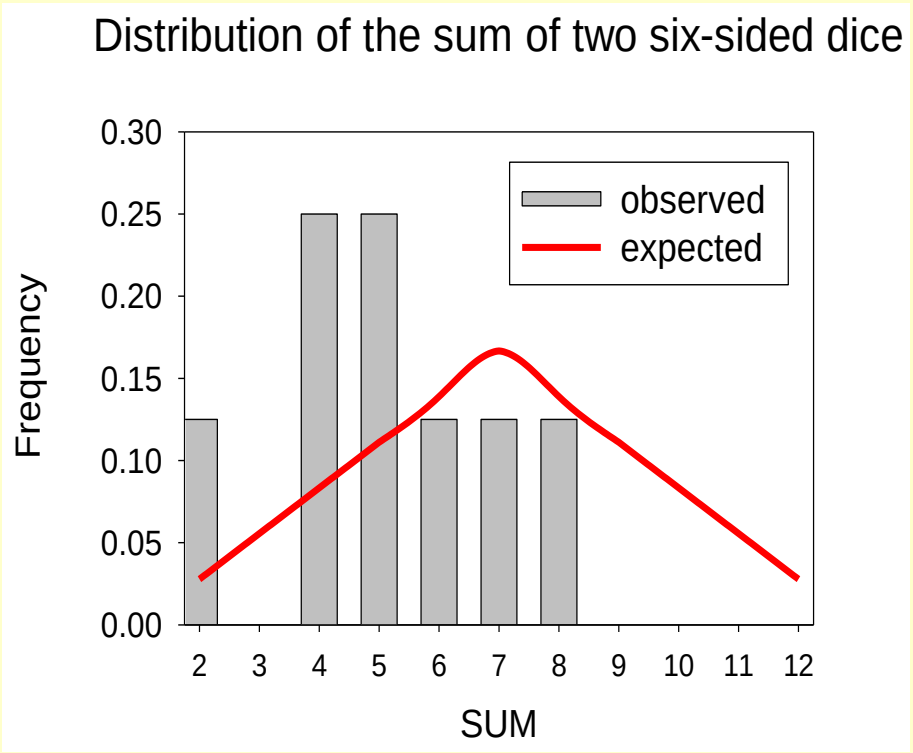
In a normal distribution:

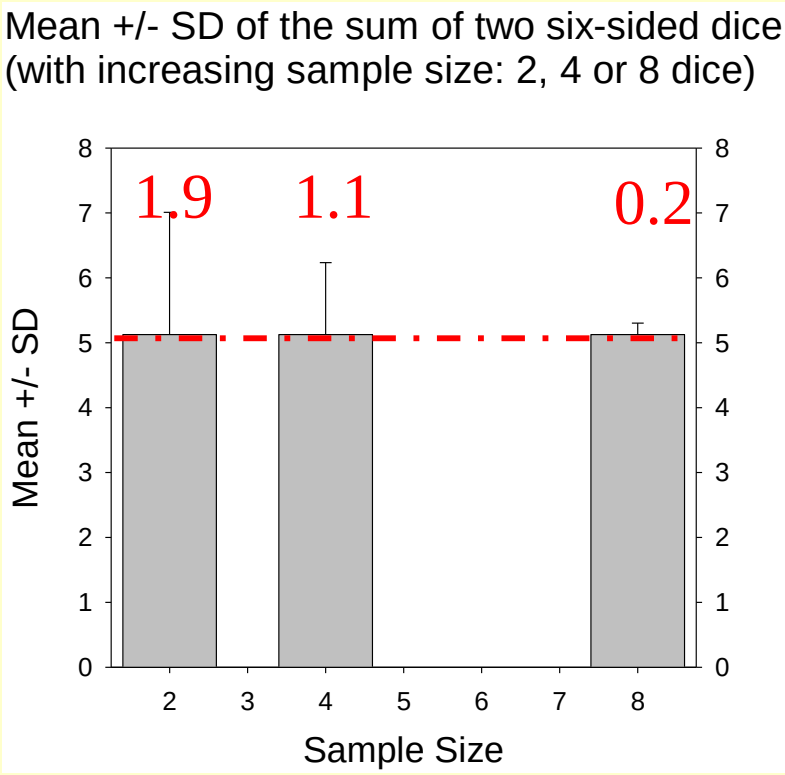
- ~ 68% observations within 1 standard deviation of mean
- ~ 96% within 2 standard deviations
- ~ 99% within 3 standard deviations

Next Step: Going Beyond the Data

A simple Example:

Roll two dice and add the value – 8 times





Mean = 5.1 +/- 1.9 (S.D.)

Next Step: Going Beyond the Data

Sampling allows us to guess about population parameters

However, different samples from the same population will differ… due to random variation (sampling)

Therefore, it is critical to assess how well any given sample represents the population.

To do this, we use the Standard Error (S.E.)

S.E. : Standard Deviation / Sqrt (N)

$$S.E. = \frac{s}{\sqrt{N}}$$

Confidence Intervals

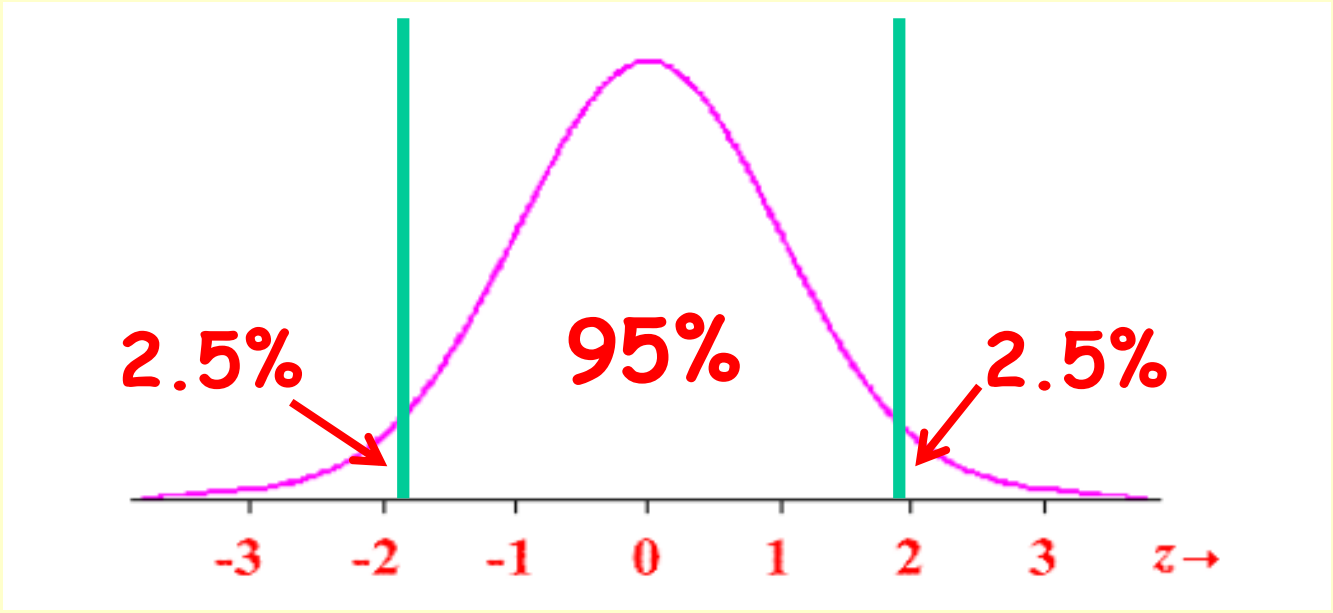
Because different samples produce slightly different estimates, we can assess the accuracy of our estimates by calculating the boundaries within which we believe the true population parameter value lies.

The confidence interval (**CI**) indicates reliability of an estimate. It is the observed interval (i.e. calculated from observations), different from sample to sample, that frequently includes the parameter – if we repeat the experiment

How frequently the observed interval contains the parameter is determined by the **confidence level**.

Measuring Probability

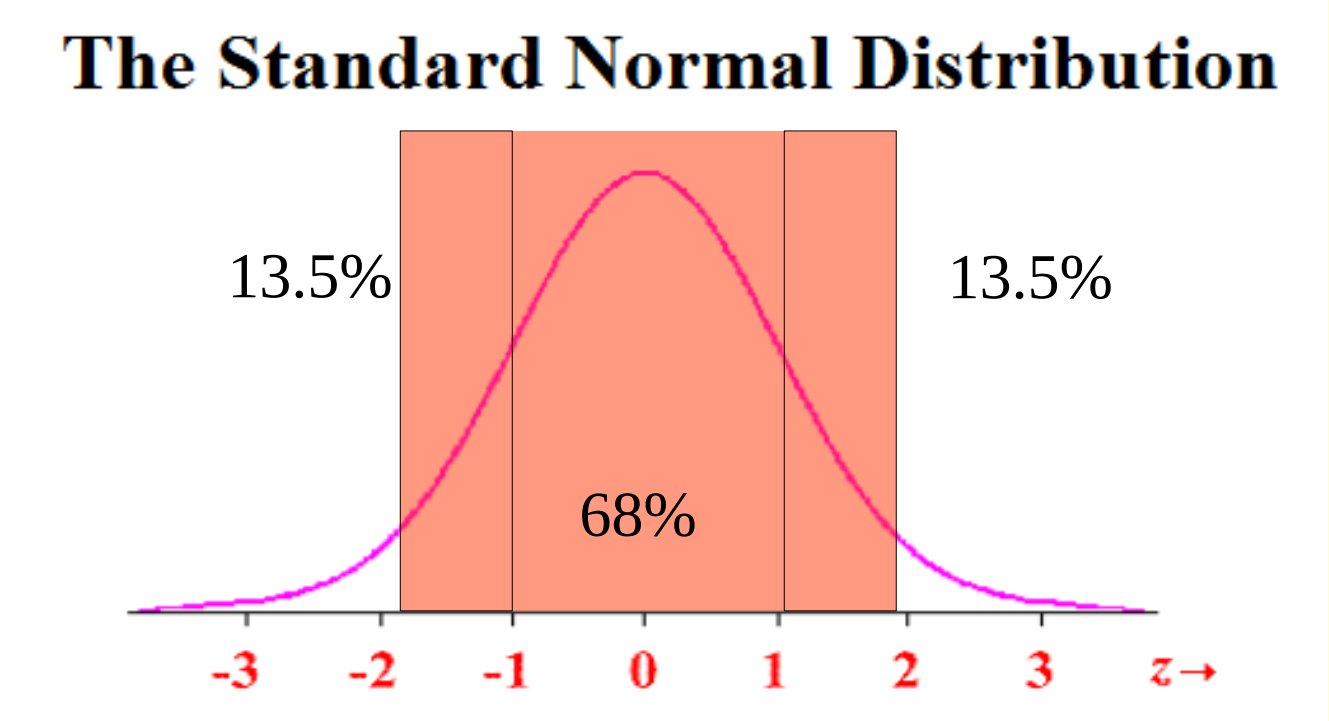
To contain 95% of the mass of the distribution, we must encompass from Z = -1.96 to Z = +1.96



NOTE: Area under entire curve is 1

Confidence Intervals – One Test

Standardized normal distribution (mean= 0, S.D.= 1) is characterized by the following properties:



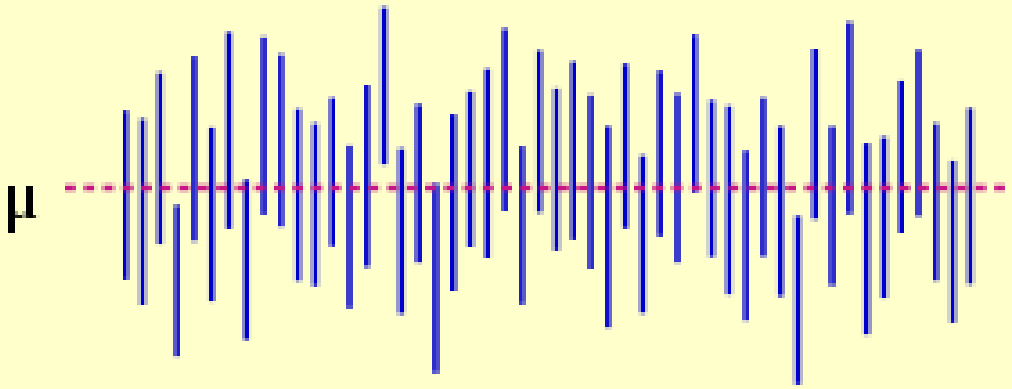
NOTE: Area under curve is 95%

To contain 95% of the mass of the distribution, We must encompass from Z = -1.96 to Z = +1.96

Confidence Intervals – Many Tests

Formulation = 95% confidence intervals

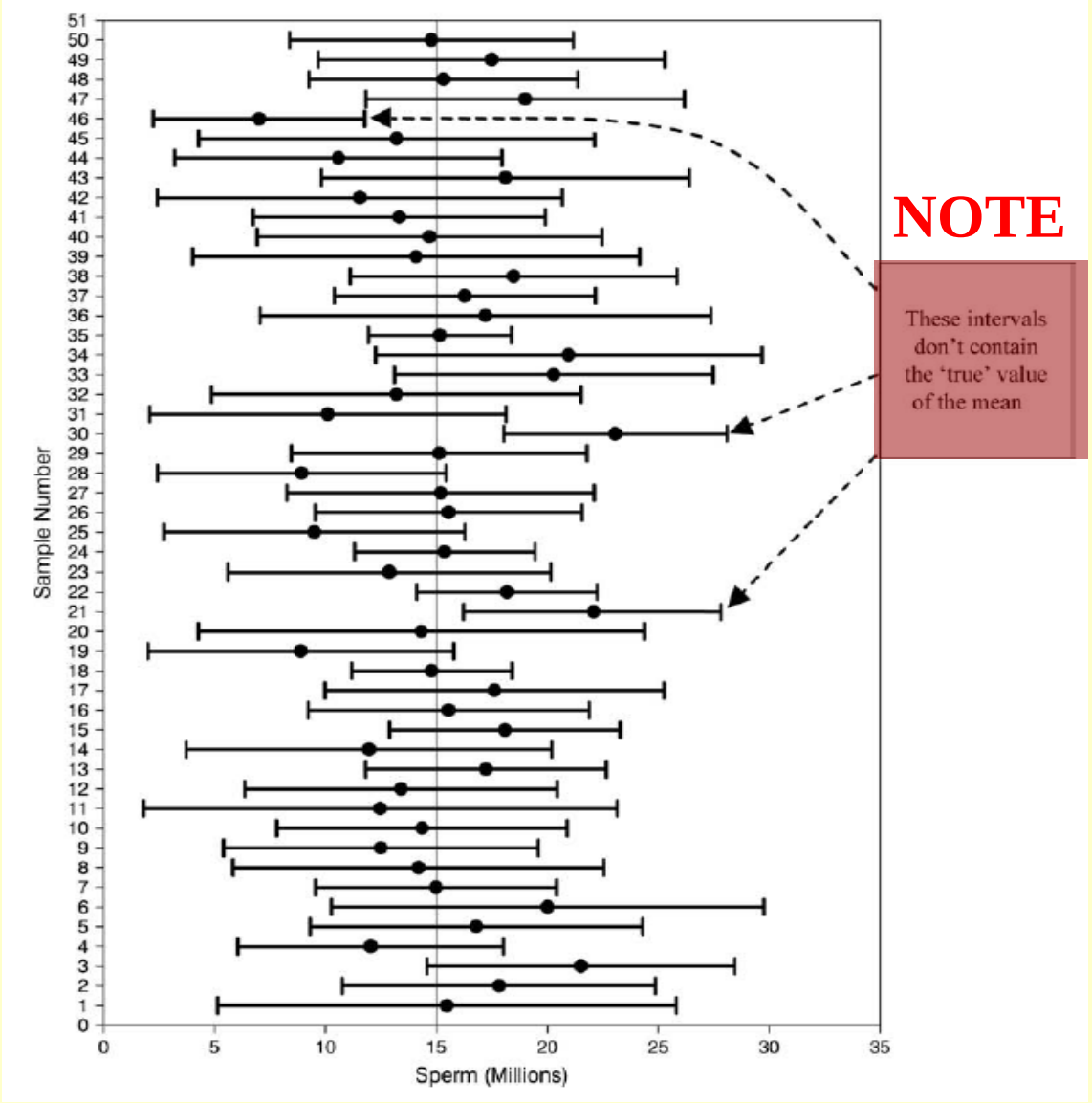
Lower bound: Mean – (1.96 * SE)
Upper bound: Mean + (1.96 * SE)



By definition: 95% of the confidence intervals (from different experiments) will overlap the real parameter μ

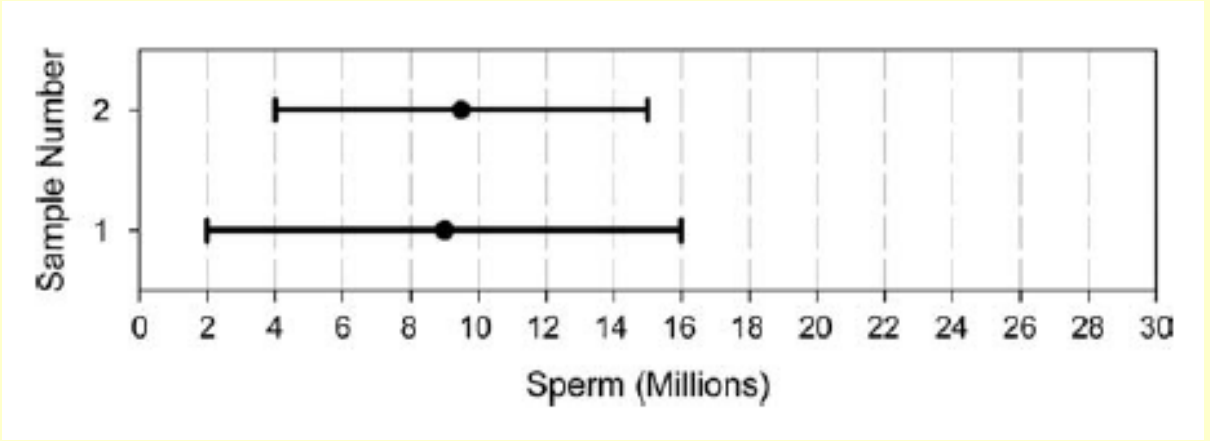
Interpreting Confidence Intervals

The (**CI**) is the interval that includes the estimated parameter, with a probability determined by **confidence level** (usually 95%).

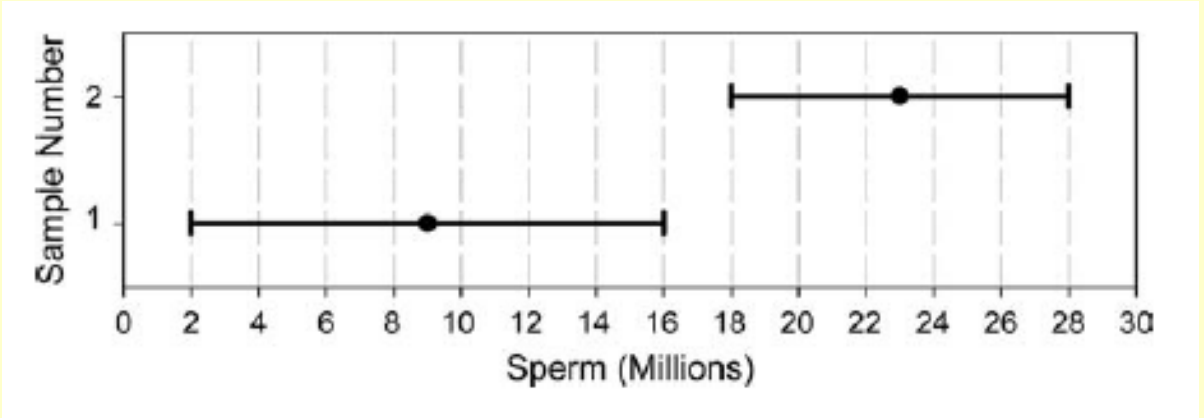


Interpreting Confidence Intervals

Case 1.
Two samples indistinguishable. They are from same population



Case 2.
Two samples different. They are not from same population



5-minute paper

Write your name

What is a point estimate ?

What is a confidence interval ?

How is a Z score calculated (show formula) ?

How is a 95% C.I. calculated (show formula) ?

Summary

- Data Summaries used the describe the samples
- Need to go "beyond the data":
  - calculate point estimate (of given parameter)
  - calculate spread about that estimate
- Normally distributed data amenable to parametric summary statistics (mean / SD) and tests
- Next, we will learn how to analyse data for normality