Last updated October 5, 2016

Sampling – Basic Concepts

G. Bacaro

Design and Analysis of Environmental Monitoring and Experiments
Master Degree in Global Change Ecology
I Year, I term

Inference, Sampling and Confidence

Sampling

Inference defined

Sampling

Statistics and parameters

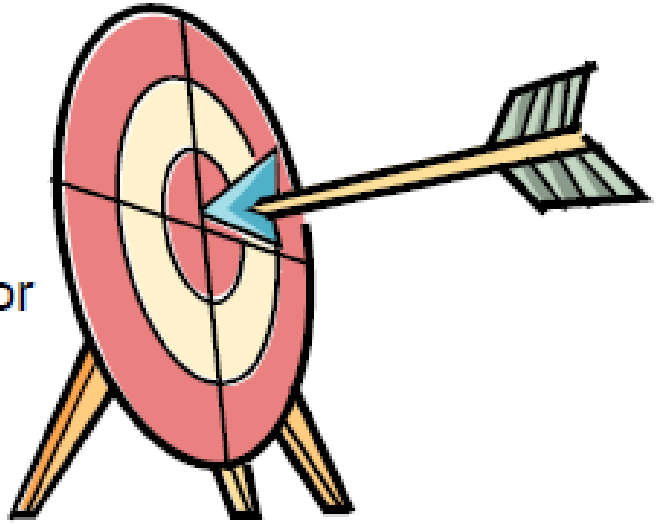
Sampling distribution

Confidence and standard error

Estimation

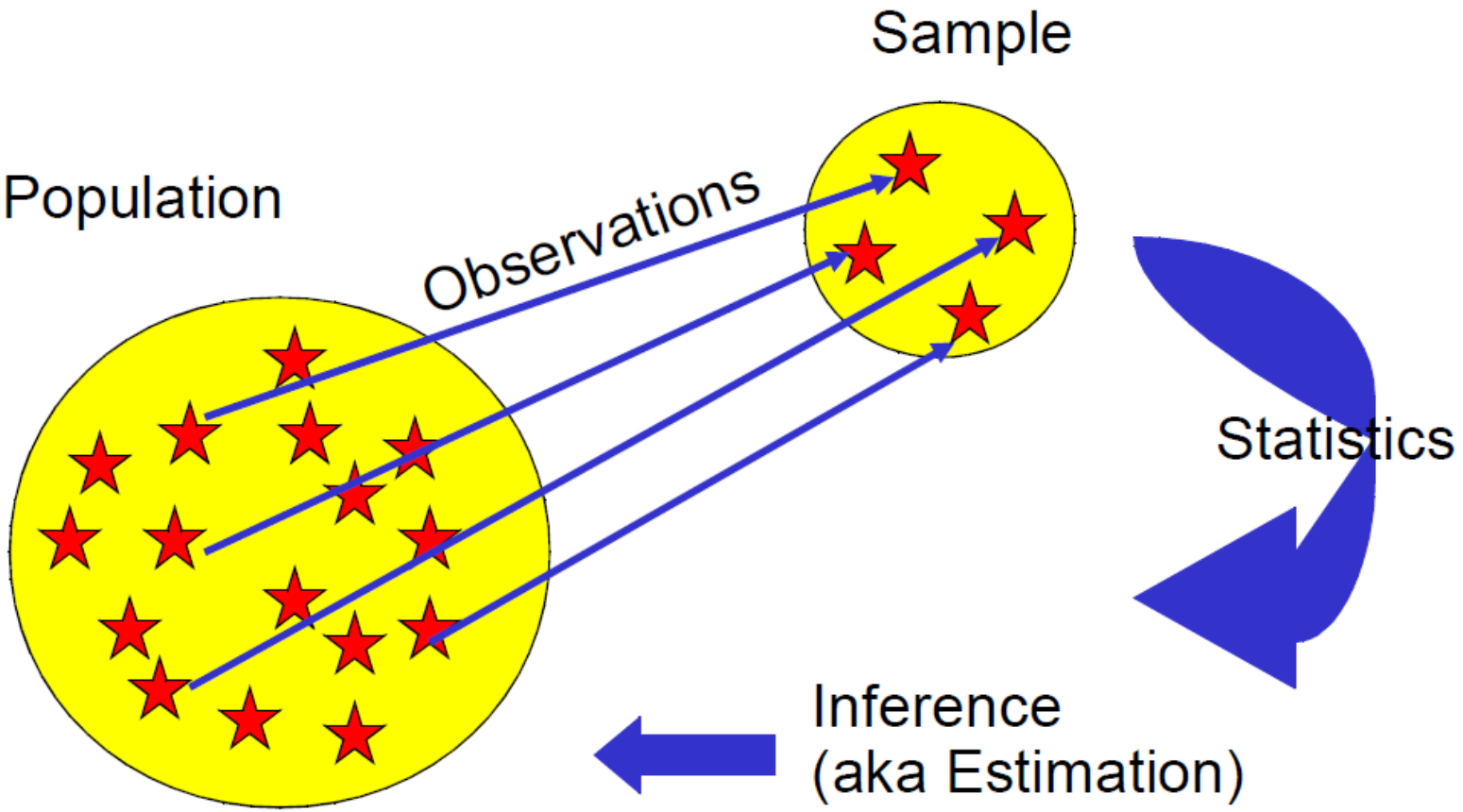
Precision and accuracy

Estimating sample size



Inference

Sampling



Inference

Sampling

"the act or process of inferring: the act of passing from statistical sample data to generalizations (as of the value of population parameters) usually with calculated degrees of certainty"

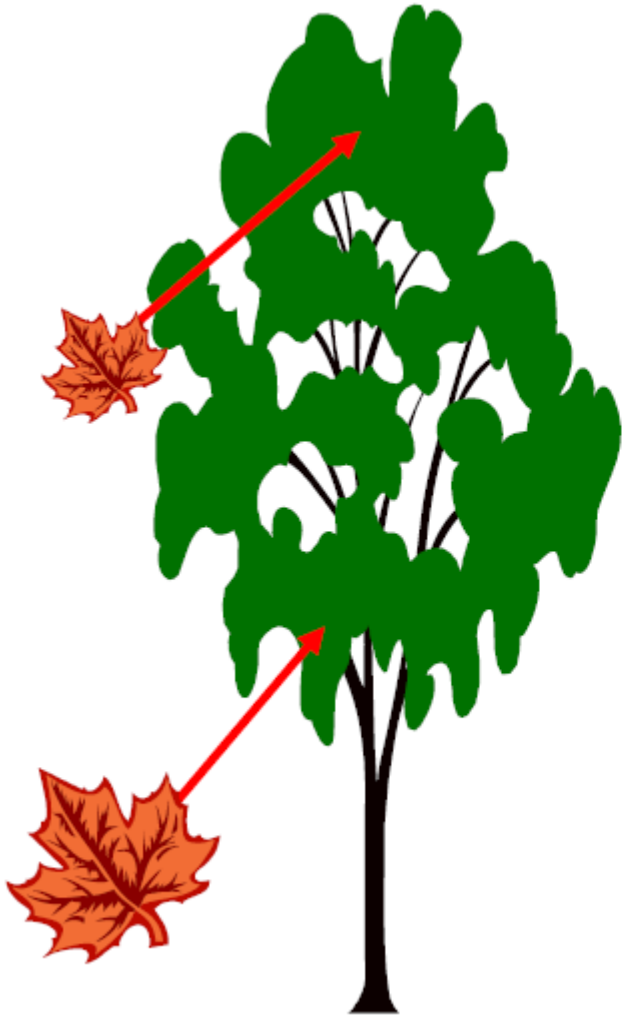
Merriam-Webster 1998

The strength (certainty) of your inference is largely dependent upon your sampling (both methodology and intensity).

But, the sample must be *representative* of the population.

Representative Sampling

Sampling



Consider a maple tree from which you wished to describe mean leaf length from.

Where would you sample from?

Lower leaves = shade leaves
Upper Leaves = sun leaves

Avoid BIASED sampling!
No biological validity if biased!

Obtain a Representative Sampling

RANDOMNESS

Insure that every member of the population has an equal probability of being sampled.

CAUTION !!!
A random sample can still be biased !

Randomness only refers to how the observations were selected for the sample.
No guarantee that sample is representative.
(Chance alone may affect outcome.)

Obtain a Representative Sampling

STRATIFICATION

If there is obvious bimodality, stratify the sample (split) and randomly sample WITHIN each stratified segment.

Example of Stratification:
This is why we sample the tree stratum and seedling stratum separately in vegetation studies.

Random vs. Haphazard

Random

An element of a set whose members all have an equal probability of occurrence.

Haphazard (Sample of Convenience)

By mere chance, accident, or fortuity; without design, casually. (Probabilities unequal.)

The Oxford English Dictionary, 2nd ed.

Statistics and Parameters

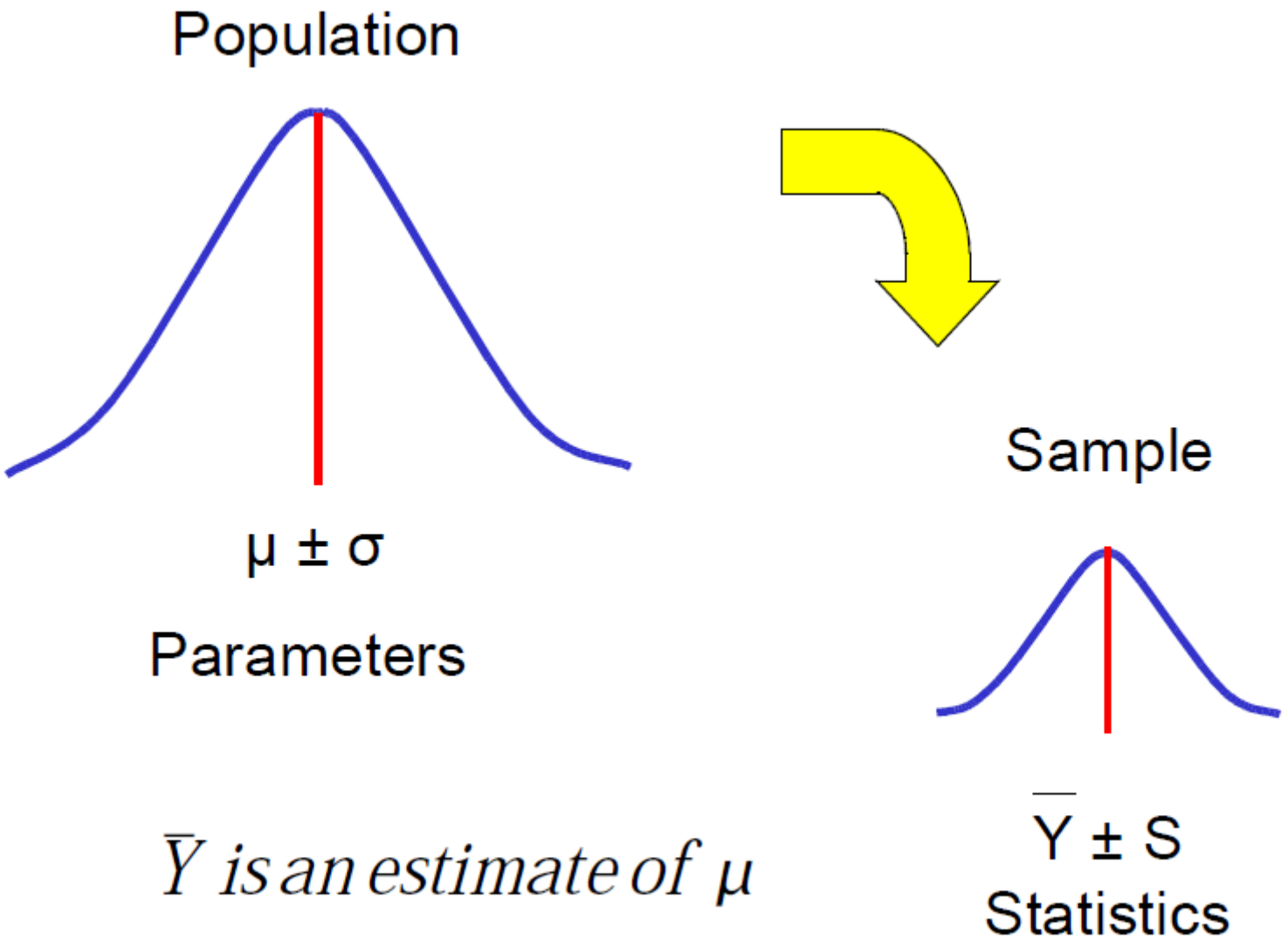
Sampling

Values used to describe populations are referred to as parameters (Greek) (e.g., standard deviation = σ)

Values used to describe samples are referred to as statistics (Roman) (e.g., standard deviation = S)

Statistics and Parameters

Sampling



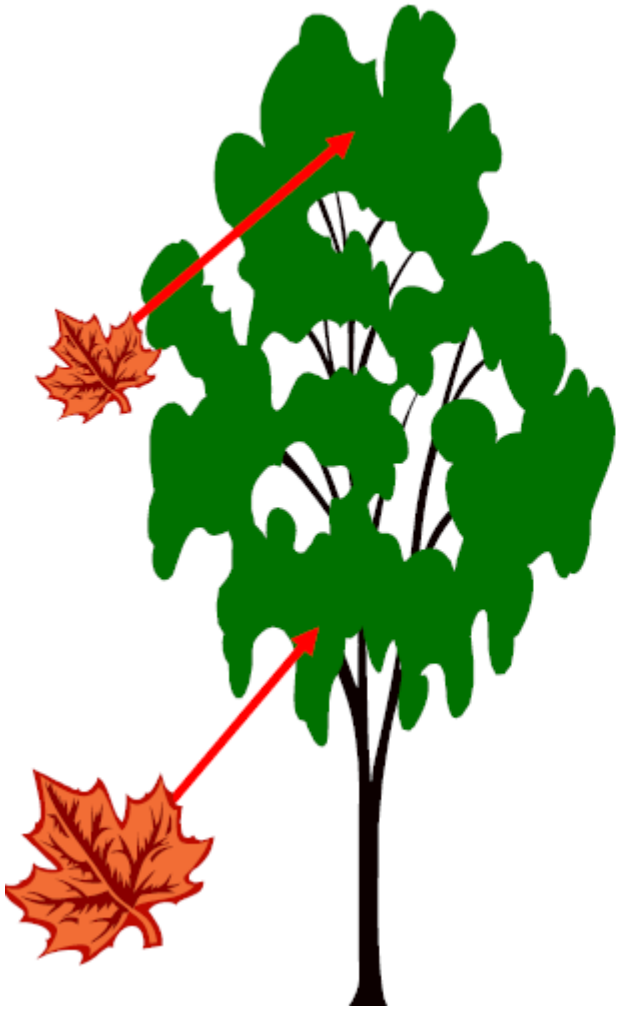
Statistics and Parameters

Sampling

$\bar{Y}$ is said to be an *unbiased* estimate of μ because if we were to draw an infinite number of samples of size *N*, with replacement, the mean of those samples would equal μ.

However, the mean of all the variances S^2 would NOT equal σ^2. The former would be smaller. Thus, we say S^2 is a biased estimate of σ^2.

Why is S^2 a biased estimate of σ^2



Recall the maple leaf example:

If we made 100 samples of $N = 10$ leaves each, it is very improbable that we would ever sample the full range of leaf sizes. We would likely always miss the smallest and largest values from the population.

We are unable to ever sample the full range (variance) of the population.

Sampling Distributions

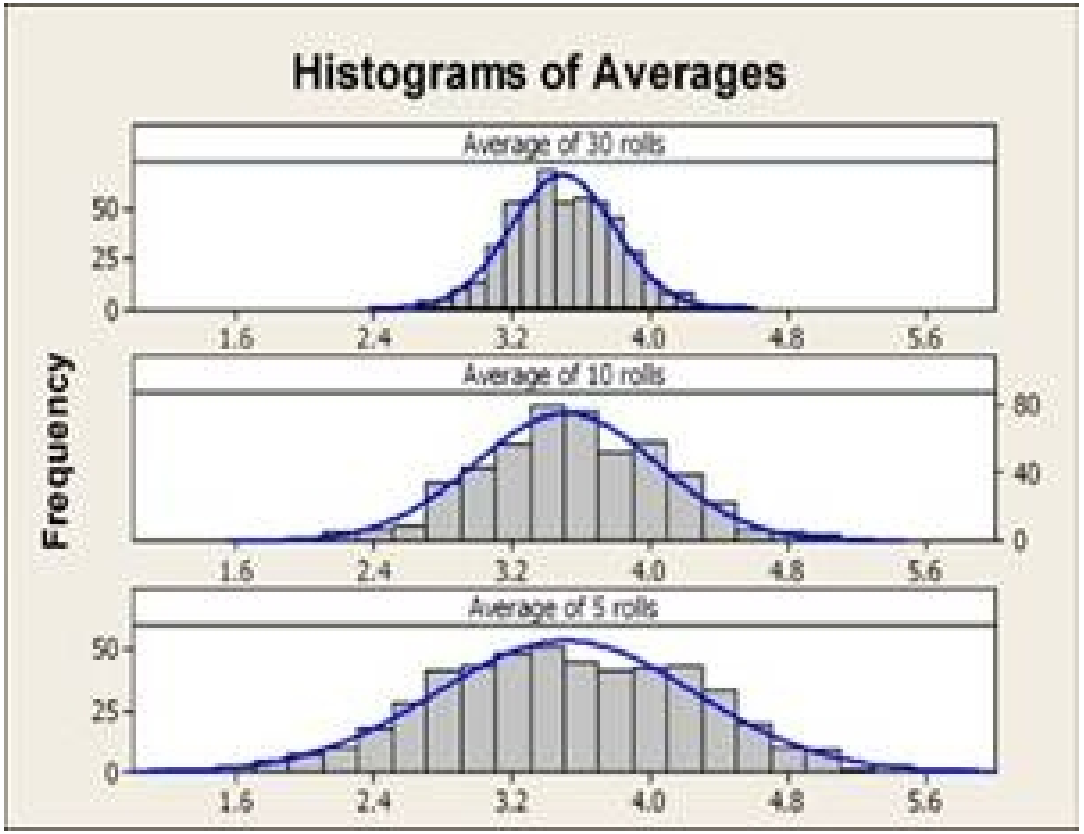
Note that we can use $\overline{Y}$ and S^2 as estimators of μ and σ^2 if the target population is normal, regardless of the sample size.

The **Central Limit Theorem** suggests that as sample sizes (*N*) become large, the shape of most samples begins to approximate a normal distribution.

Note that in addition to the normal distribution, there are many other distributions that can be used for statistical inference (*t*-, *F*-, χ^2, etc.).

Central Limit Theorem

Sampling



In other words, if we repeatedly take independent random samples of size n from any population, then when n is large, the distribution of the sample means will approach a normal distribution.

As the sample size increases, the sampling distribution of the mean, X-bar, can be approximated by a normal distribution with

mean μ

and standard deviation

$\sigma/\sqrt{n}$

where:

μ is the population mean

σ is the population standard deviation

n is the sample size

Accuracy

What is accuracy?

The degree to which a measurement, or an estimate based on measurements, represents the true value of the attribute that is being measured.

Last. A Dictionary of Epidemiology. 1988

In short, obtaining results close to the *TRUTH*.

Associated terms:

- Validity

Precision

Sampling

What is precision?

Precision in epidemiologic measurements corresponds to the reduction of random error.

Rothman. Modern Epidemiology. 1986.

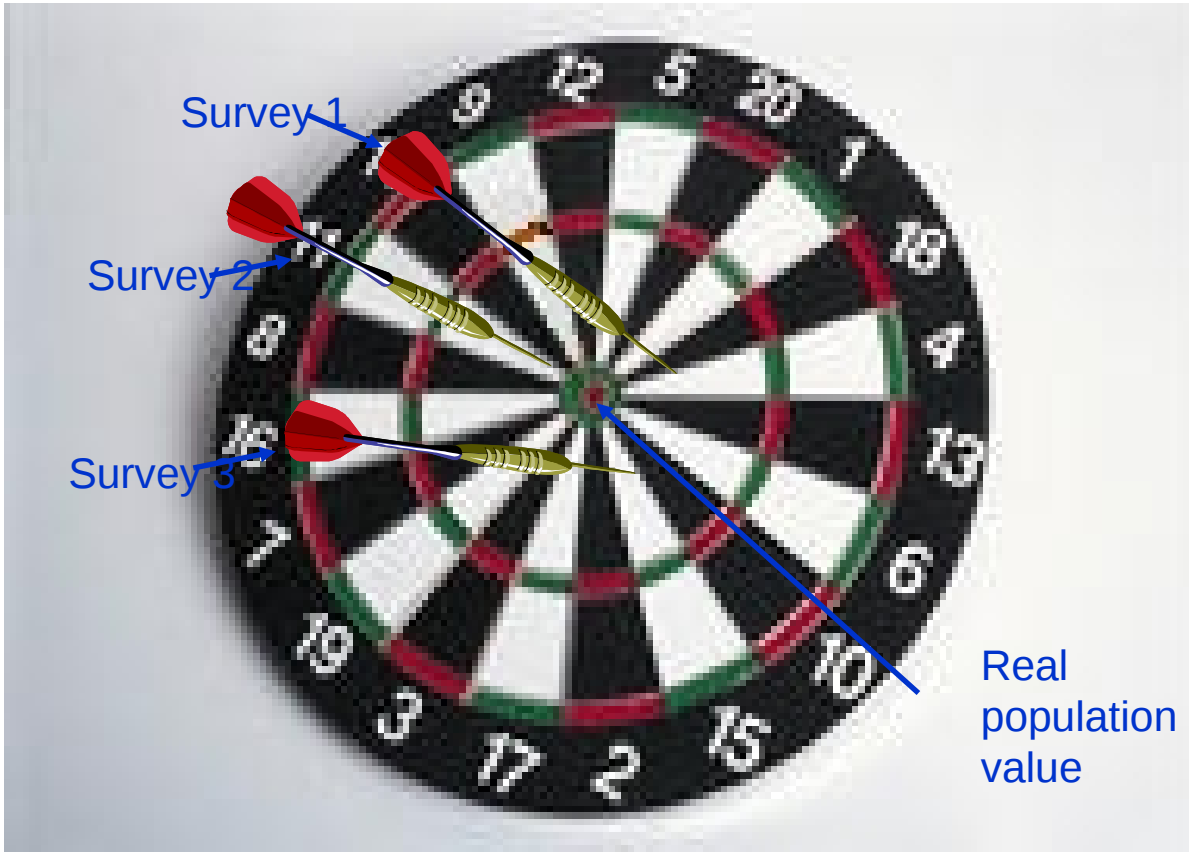
- In short, obtaining similar results with repeated measurement

Associated terms:

- Reliability
- Reproducibility

Accuracy vs. Precision

Accuracy: obtaining results close to truth



Accuracy vs. Precision

- Precision: obtaining similar results with repeated measurement (may or may not be accurate)

Accuracy vs. precision

Poor precision (from small sample size) with reasonable accuracy (without bias):

Accuracy vs. precision

Good precision (from small sample size) with reasonable accuracy (without bias):

Accuracy vs. precision

Good precision (from large sample size), but with <u>poor accuracy</u> (with bias):

Confidence Interval and Standar Errors

Sampling

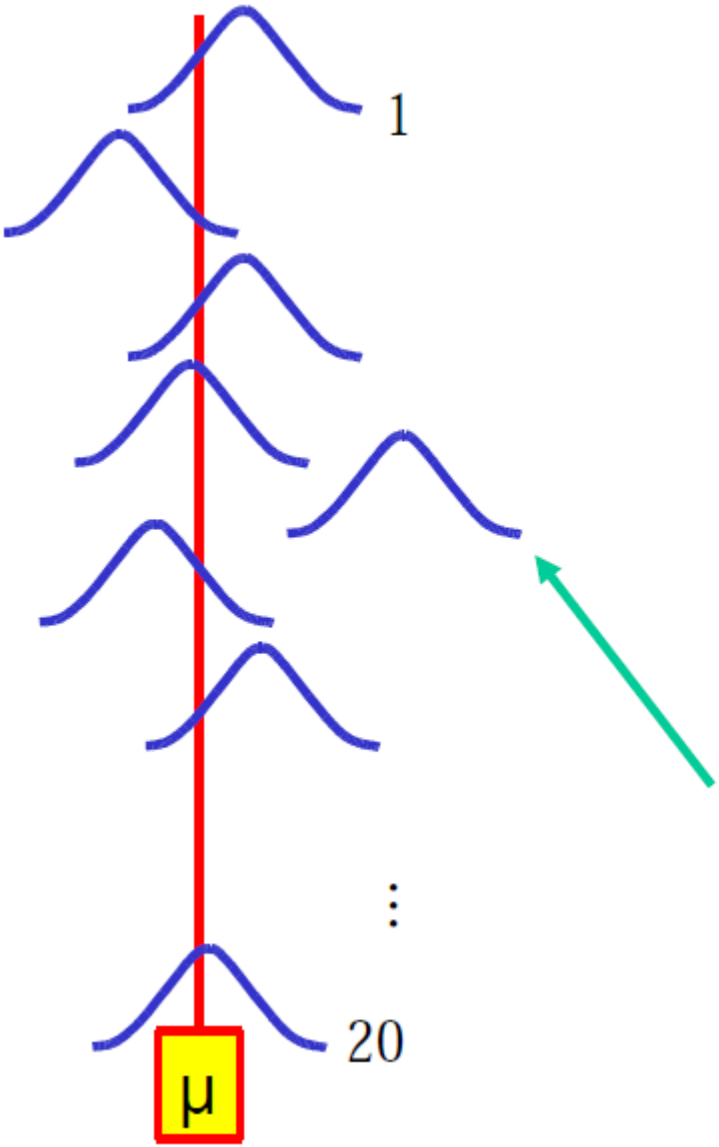
Earlier, we introduced the notion of "strength" or "certainty" of inference.

Formally, this statistical notion is referred to as *confidence*.

In any given experimental situation we can calculate our statistical confidence that our sample statistics accurately reflect our population parameters. In other words, what is the probability that we have made a mistake?

Confidence

Sampling



Suppose we know μ of a population.

Suppose also that we collected 20 samples (*N* = 10 each) from that pop in an effort to describe μ.

If 1 of the 20 samples was unable to describe the mean (one sample did not contain μ), we can say we have 95% confidence in our ability to sample the population and accurately describe its mean.

Standard Error of the mean

Sampling

From the previous illustration we can see that the mean of all the means should closely approximate μ.

We can also approximate the standard deviation of a sampling distribution of means.

This is referred to as the Standard Error (SE):

$$\sigma_{\bar{Y}} \approx \frac{\sigma}{\sqrt{N}} \qquad \text{Approx. as:} \qquad SE \approx \frac{S}{\sqrt{N}}$$

Estimation

Sampling

Now, suppose a sample is drawn from a population with unknown μ and σ :

Sample: $N = 100, \bar{Y} = 50, S = 5$

Assuming an infinite number of samples of $N = 100$, we can calculate SE = S/√N = 5/10 = 0.5

Q: What does a SE of 0.5 represent?
A: A standard deviation of our sampled means.

± 1 SE would represent 68% of the area (± 2 SE ca. 95%)
In other words, 68% of the time the mean would fall between 49.5 and 50.5 (we have 68% confidence).

How much confidence is enough?

From the preceding example, we had 68% confidence in our prediction of the mean.

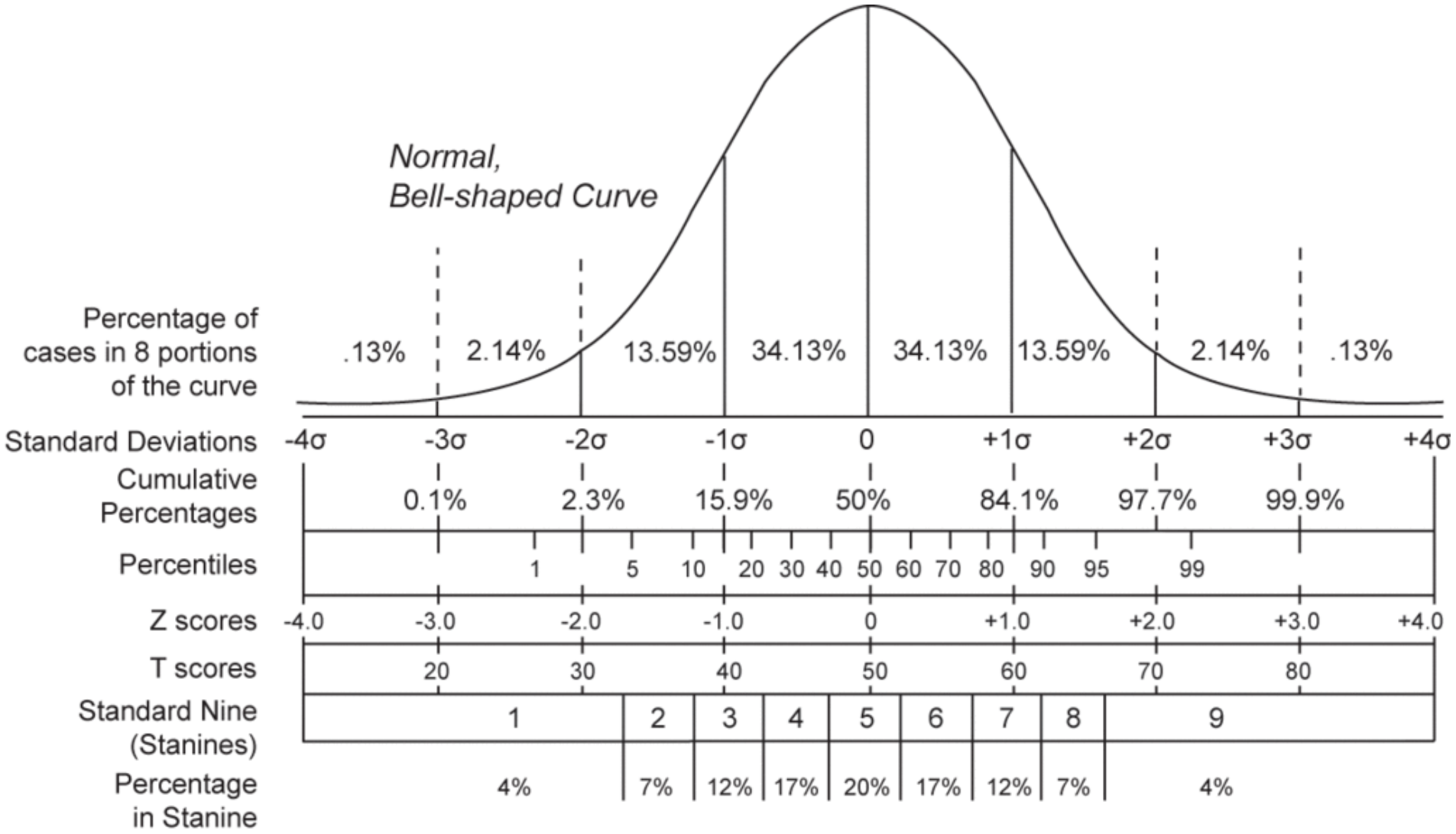
Unfortunately, to be looked upon with favor by our colleagues we would need at least 95% confidence.

Q: How do we get 95%?
A: Use Z-values and our knowledge of the SND.

With 95% of the area, we know there is 5% remaining (2.5% in each tail).

Standard Score

Sampling



How much confidence is enough?

We can now use our determined Z-value of 1.96 to calculate the 95% Confidence Interval around the mean (the "2SE Rule of Thumb):

$$CI_{0.95}: \bar{Y} \pm 1.96 \frac{S}{\sqrt{N}}$$

In our example: $CI_{0.95}$: 50 ± 1.96 (0.50)

$CI_{0.95}$: 49.02 - 50.98

NB: A CI can be calculated for any statistic!

Precision and Accuracy Revised

Sampling

Remember our discussion of precision & accuracy?

Because $SE = S/\sqrt{N}$,
the greater N becomes, the smaller SE becomes.

So, as *N* increases, SE decreases (the precision of our estimate of μ increases)!

This is why we normally wish *N* to be as large as is feasible for a given experiment. (Within TME limits.)

Precision: depends upon sample size (CI)
Accuracy: depends on sampling protocol

Sample Size

Q: So, if sample size is so important, how large of a sample is needed to describe a population?
A: It depends…

★ McCarthy's Quick Guidelines: ★

N	Adequacy
1	Poor
2-4	Fair
5-10	Good
10-15	Better
>15	Best

Index of Precision

General guidelines are useful for general practice but sometimes an explicit estimate of precision is required.

Index of Precision: $D = SE / \bar{Y}$

For example, given $SE = 2$ and $\bar{Y} = 20$

The level of precision is 10% of the mean
($D = 2/20 = 0.10$)

Index of Relative Precision

D would be better expressed as a $CI_{0.95}$:

$$D' = t_{0.05/2,\,df}\left(\frac{SE}{\bar{Y}}\right)100$$

Thus, you are expressing precision as a *percentage* of the sample mean at the traditional 95% CI.
D' is more straightforward to conceptualize than D.

Calculate Relative Precision – Example-

Sample: 5, 3, 4, 4

$N = 4$
$\bar{Y} = 4$
$S = 0.816$
SE = 0.408 (S/√N)

Conclusion:
More sampling is required

D' = 3.182 (0.408 / 4) * 100 = 32.5%

Note$_1$: df = N - 1, $t_{0.025,3} = 3.18$
Note$_2$: 32.5% is high; < 20% preferred, < 10% better

Estimating Desired Sampling Size -pilot data

If a small, preliminary (pilot) study provides a fair description of the mean and variance, then you can estimate N needed for a desired level of precision:

$\hat{N}$ = desired N
t = t-value from App. C (95CI) at df from pilot data
S = std. dev. of pilot data
D'' = D' expressed in decimal form (not %!)
$\bar{Y}$ = mean of pilot data

$$\hat{N} = \left(\frac{t_{0.05/2,\,df}\, S}{D'' \bar{Y}} \right)^2$$

*Caveat: final sample will have different mean and variance!

Estimating Desired Sampling Size -example

Assume a pilot study with $N = 10, \bar{Y} = 4$, S = 2.
Assume also the desired level of precision to be D' = 20% (i.e., a 95% CI ± 20% of $\bar{Y}$).

Thus, D" = 0.20 and $t_{0.05/2,9} = 2.262$

$$\hat{N} = \left(\frac{(2.62)(2)}{(0.20)(4)} \right)^2 = 32$$

Conclusion:
A sample size of 32 is needed to get this level of precision