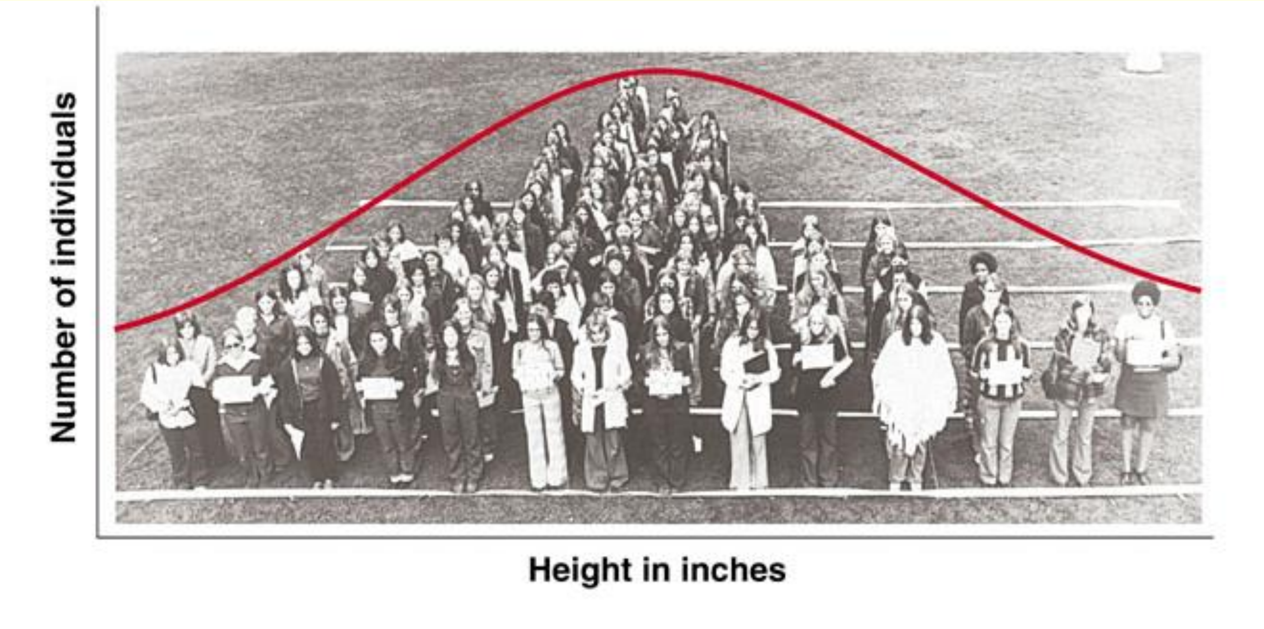
Estimation:
central tendency



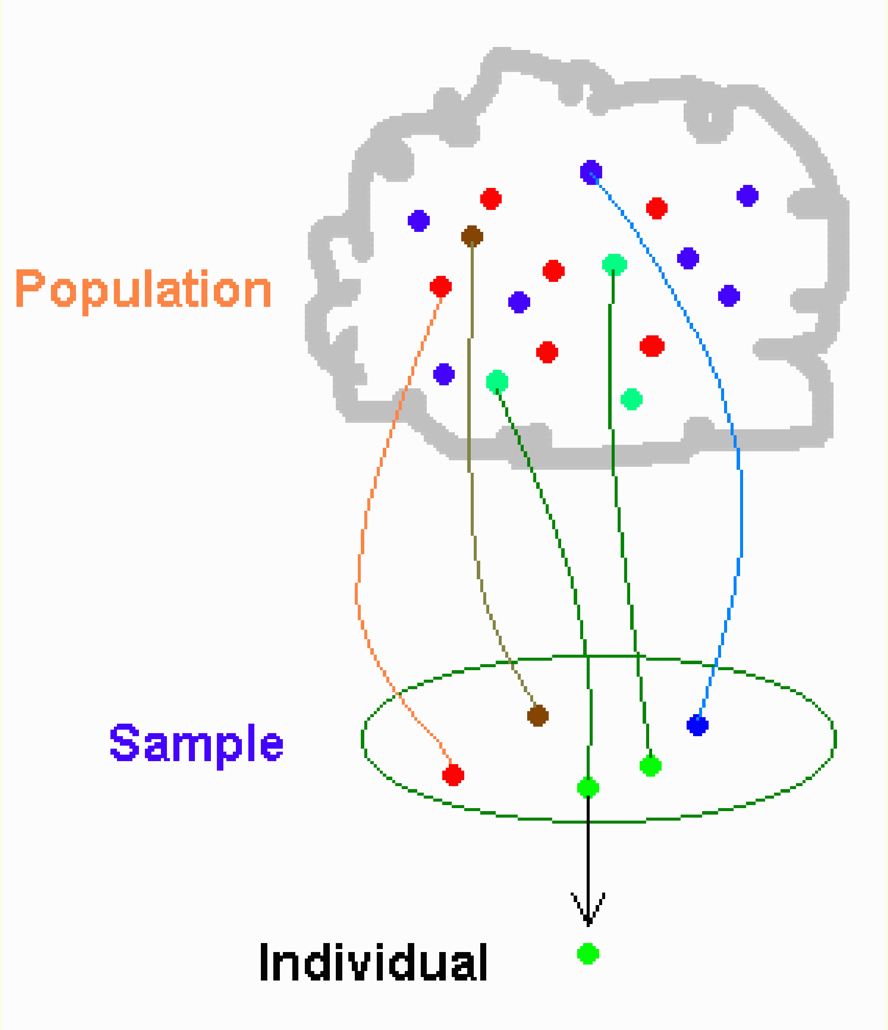
http://www.pelagicos.net/classes_biometry_fa16.htm

Sampling

Why do we sample ?

Why don't we just sample one individual ?

How do we ensure our sample is representative of the entire population ?



Random Sampling

A random sample is a subset of individuals (a sample) chosen from a larger set (a population) such that:

Each individual is randomly chosen (by chance):

- each individual has an equal
- and independent probability

of being chosen during the sampling process,

A random sample is an unbiased surveying technique.

Estimation - Soccer

Number Goals Scored per Game Played:

1, 2, 2, 3, 1, 1, 4, 1, 3

Mean: 18 / 9 = 2

Bracket range of outcomes:

from 1 to 4

The Arithmetic Mean

Take a "random sample" and summarize the observations:

Mean: The average of the scores in the distribution

$$\overline{X} = \frac{\sum_{i=1}^{n} Xi}{n}$$

Does the estimated sample mean ("X-bar") relate to the actual population mean (μ)?

Our estimate will be an unbiased estimator of μ if three conditions are met:

1. Observations taken from randomly selected individuals (from the biological population)
2. Observations are independent from each other
3. Observations taken from a biological population that follows a normal random variable (normally distributed)

The Arithmetic Mean

Note: The mean is a statistical model of the data
(but the mean value may not occur in dataset)

	Observations:	Mean:
Sample1:	1, 2, 3, 4	(1+2+3+4) / 4 = 2.5
Sample2:	0, 1, 2, 7	(0+1+2+7) / 4 = 2.5

Which of these two predictions looks less variable ?

Estimating Variability

Data Series 1: 1,2,3,4

Mean S1:
2.5

Variance S1:
5 / 3 = 1.7

Data Series 2: 0,1,2,7

Mean S2:
2.5

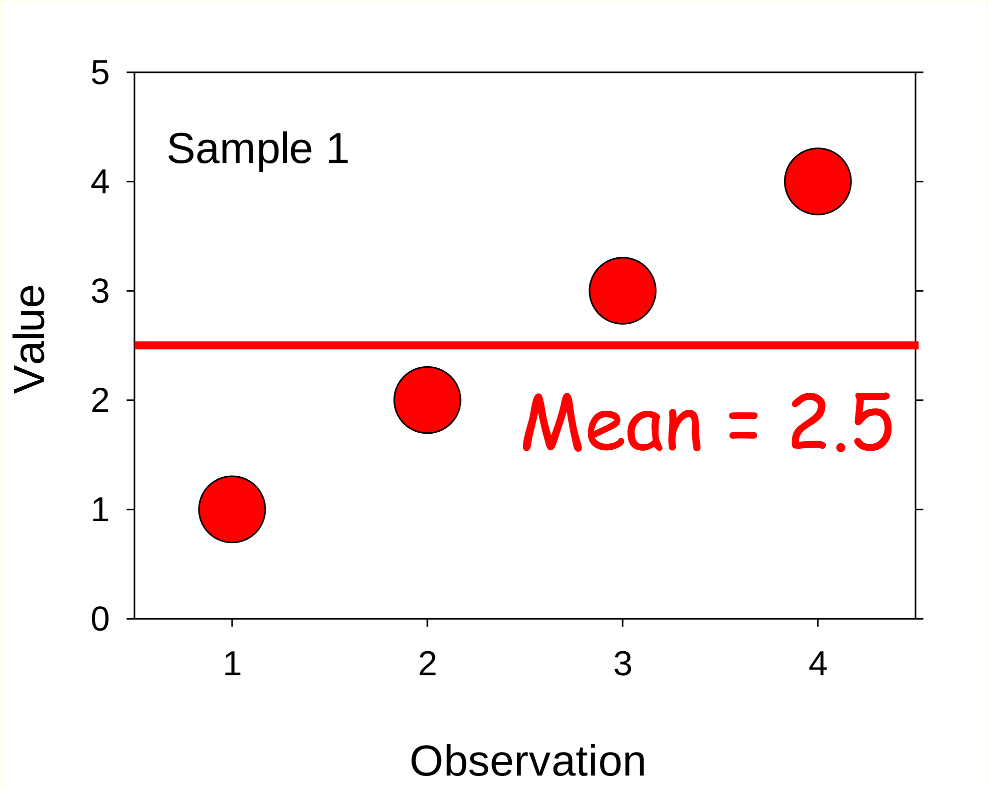
Variance S2:
29 / 3 = 9.7

Variance = $\dfrac{\text{sum of squared deviations from mean}}{\text{degrees of freedom}}$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

The Mean & The Variance

$$\text{Variance} = \frac{\text{sum of squared deviations from mean}}{\text{degrees of freedom}}$$

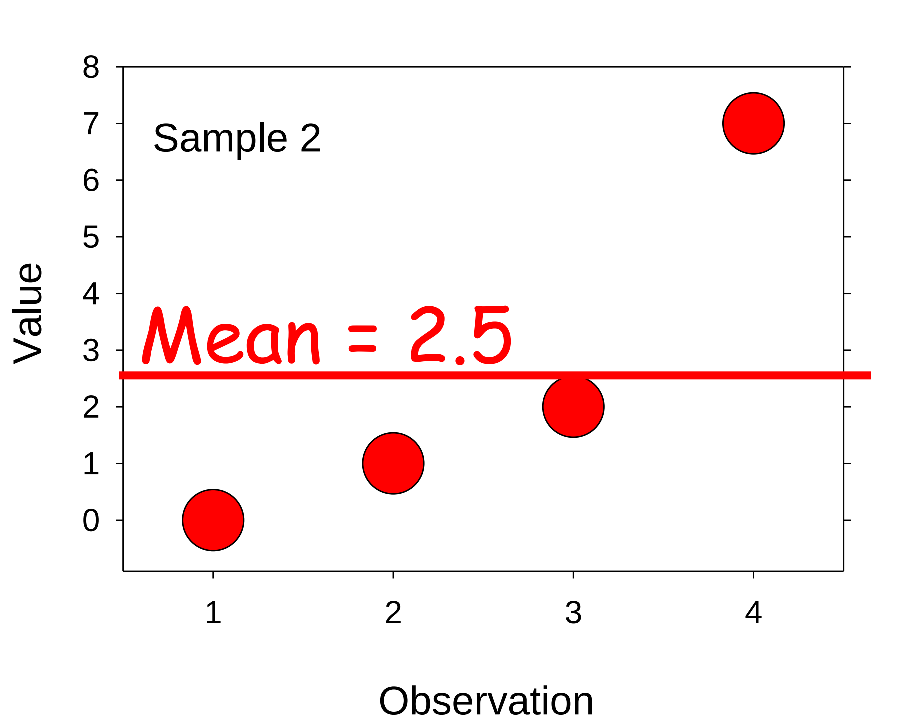


Value	Deviation	Squared Deviation
1	-1.5	2.25
2	-0.5	0.25
3	+0.5	0.25
4	+1.5	2.25
Sum	0	5

Variance = 5 / 3 = 1.7

The Mean & The Variance

$$\text{Variance} = \frac{\text{sum of squared deviations from mean}}{\text{degrees of freedom}}$$



Value	Deviation	Squared Deviation
0	-2.5	6.25
1	-1.5	2.25
2	-0.5	0.25
7	+4.5	20.25
Sum	0	29

Variance = 29 / 3 = 9.7

Why Use n – 1 to Calculate Variance ?

$$\text{Variance} = \frac{\text{sum of squared deviations from mean}}{\text{degrees of freedom}}$$

Two realisations:

1. We are calculating statistics from a sample, rather than from the entire population
2. The observations in the sample are used to calculate the mean first, then the variance

Calculating the mean diminishes our degrees of freedom from n (observations in sample) to n-1.

Degrees of Freedom

	Purple Flower	Red Flower	Totals
Green Seed	50	0	50
Yellow Seed	0	50	50
Totals	50	50	100

Definition: Number of entities that are "free" to vary when estimating some statistical parameter.

Why they Matter? Determine the shape of the probability distribution for many test statistics.

Degrees of Freedom

Number of elements in the set (i.e. how many observations there are) minus the number of different pieces of information you must know about the set to complete the calculation.

Consider a set of $n = 5$ numbers. In the absence of any information about them, all five are 'free' to range from minus infinity and plus infinity.

Suppose, however, you are also told that the sum of the set is 20. Now, only 4 of the numbers are 'free' and the last one is fixed by your knowledge of the total. Hence, there are 4 degrees of freedom.

Degrees of Freedom

Note that it does not matter which 4 numbers are "fixed" first, the final one can always be determined from the total.

Similarly, if there is a set (or sample) of 4 numbers that have a known mean (2.5) and a variance (1.7); only 2 of the numbers are free (there are two degrees of freedom).

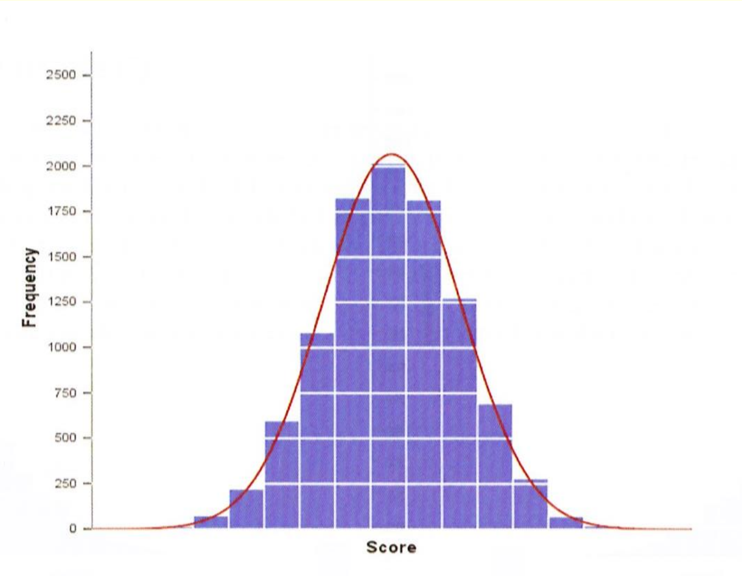
Why? Because once 2 members of the set are known, the others are inevitable … given the mean and variance.

Calculating the Variance

$$\text{Variance} = \frac{\text{sum of squared deviations from mean}}{\text{degrees of freedom}}$$

Reminder - Main goal of statistical sampling:

We calculate statistics from a sample, rather than from entire population



But, if the sampling is representative, we can make inferences about the entire population

Describing Distributions

Mean = 2.5

Variance = 1.7

But …
what are the units?

$$\overline{X} = \frac{\sum_{i=1}^{n} Xi}{n}$$

(goals + goals + … goals)

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2$$

(goals – mean_goals) ^2

Describing Distributions

Standard Deviation = Square root of the variance

The Standard Deviation does two things for us:

1. Allows us to summarize central tendency (location) and the variation (spread) in any sample using same units: mean +/- S.D. (Note: CV = S.D. / mean)

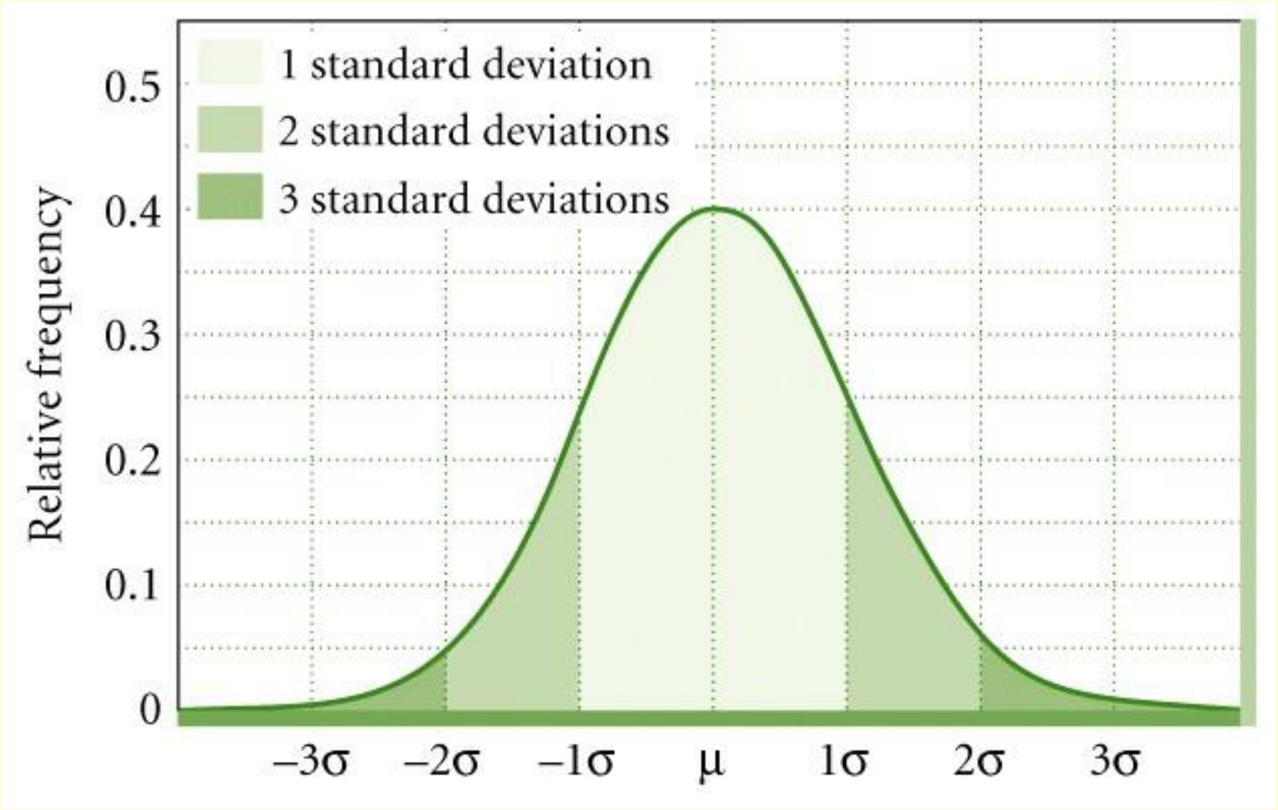
2. Unlocks the ability to estimate population frequency distributions using assumption of normality

The Normal Distribution …

Many population variables follow a normal distribution

Properties: symmetrical, mode = mean = median

mass of distribution follows specific "shape"

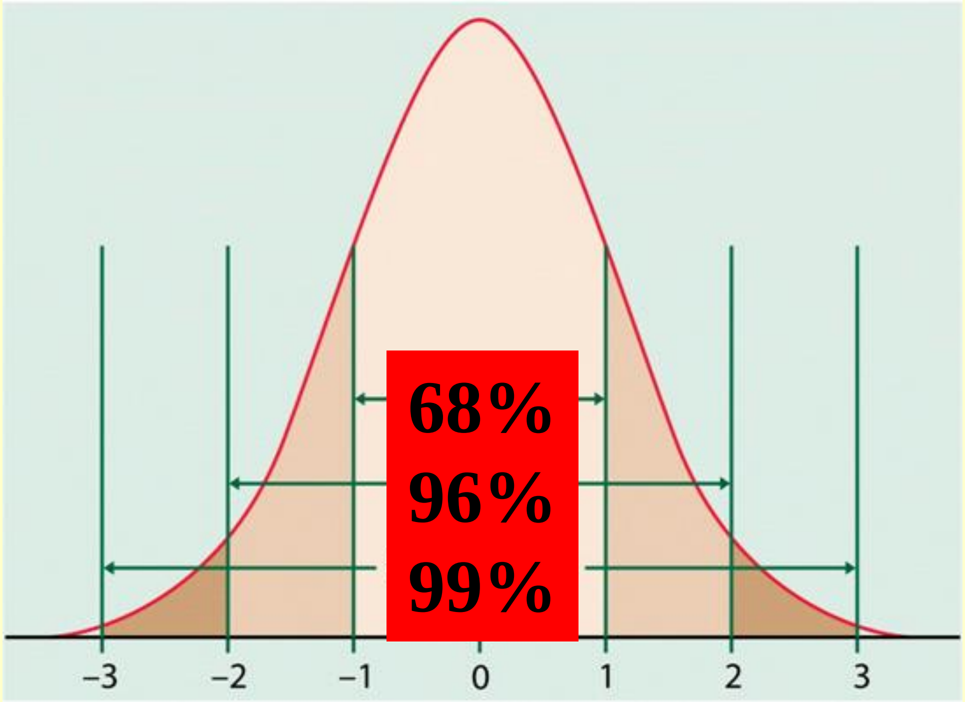


The average of this distribution is the mean, μ

Shape of distribution determined by how observations spread about μ

Spread based on σ

Is the Basis of Parametric Statistics



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

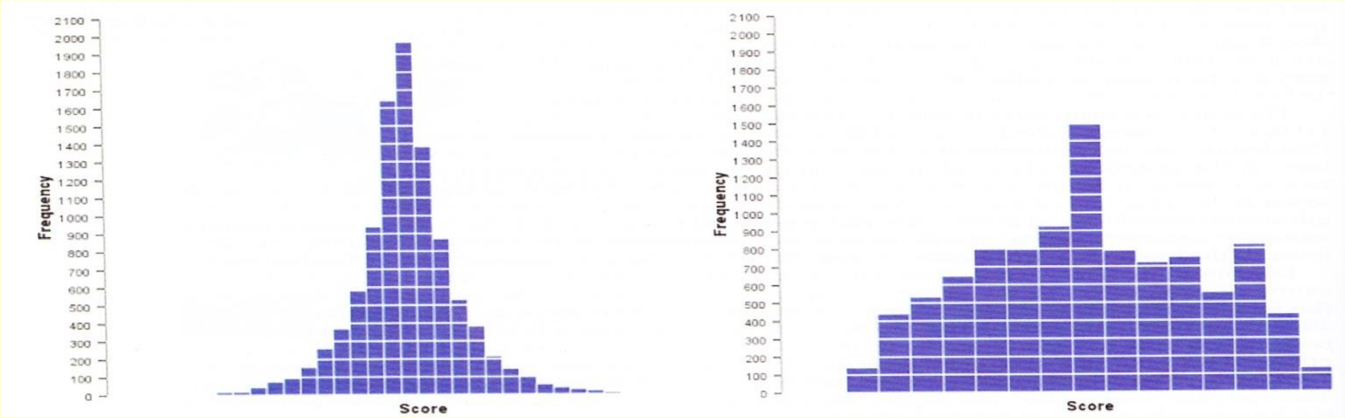
They compare the **means & S.D.s**

In a normal distribution:

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

Normal Distributions as Criteria

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



The ideal shape of the normal distribution is used as the criterion for determining whether any frequency distribution has positive or negative kurtosis.

What is the baseline value for normal distributions? 0

The Power of Normal Distributions

Data Series1: 1,2,3,4 Data Series2: 0,1,2,7

Variance 1: 1.7 S.D. 1: 1.3 Variance 2: 9.7 S.D. 2: 3.1

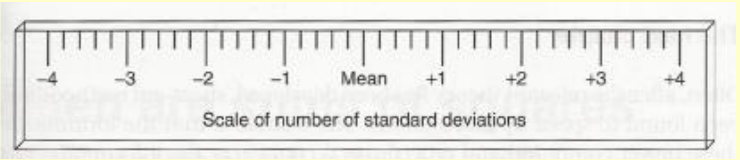
S. D.: Square Root of Variance

$$S.D. = \sqrt{s\text{^}2}$$

Why use the S.D.?

"Standardized" measure of dispersion about the mean used to describe central tendency (mean +/- SD) (CV)

Allows prediction of the distribution of observations in a measured sample



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)

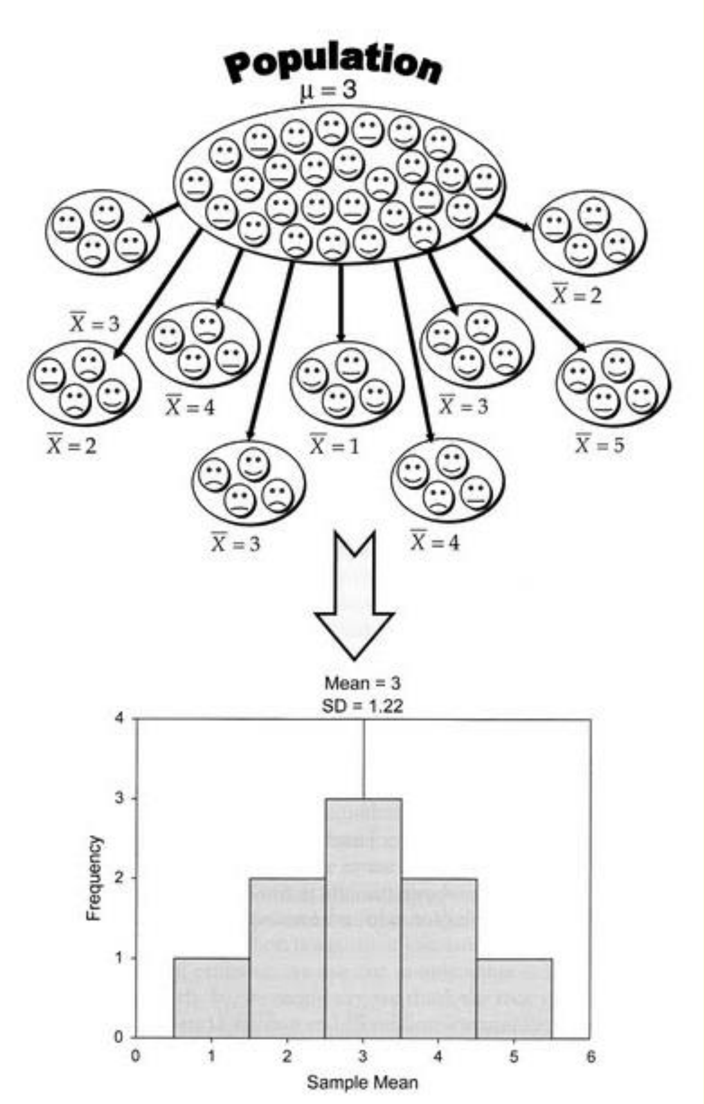
Therefore, it is critical to assess how well any given sample represents the population.

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

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

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

Next Step: Going Beyond the Data



Resample the same population, by looking at small number of individuals at a time… 9 times

This illustrates sampling variation

We plot frequency distribution of the nine estimated means.

The mean of the means provides the best estimate of μ

But what about the S.D. of this distribution?

Confidence Intervals

Developing Predictions with Built-in Margin of Error:

In Grand Forks, thousands of people, prepared for the flood by building sandbag dikes.

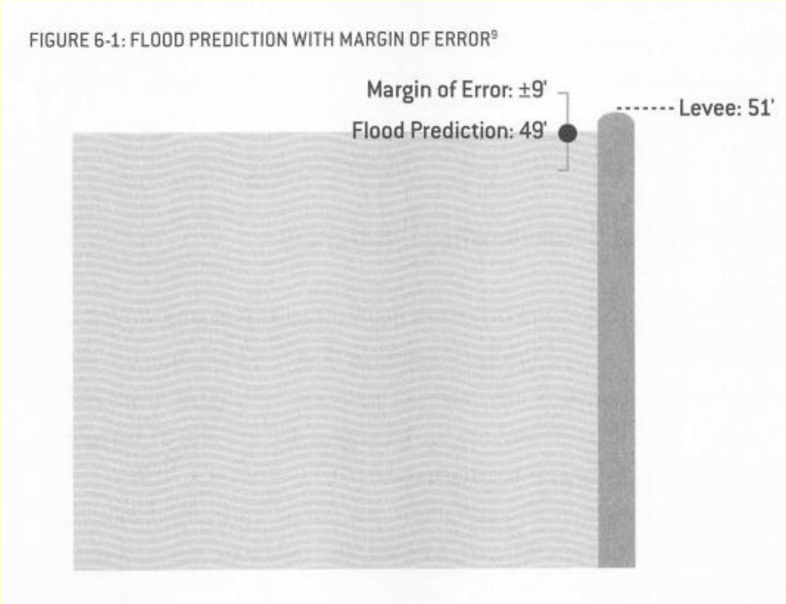
These dikes were based on a 49-foot estimate of flooding by the National Weather Service.

What went Wrong ?

Statistical Inference & Reliability

A point estimate is a single value of a population parameter: 49 ft

Confidence intervals built around point estimates to capture the variability inherent in the data and in the estimation method.



Confidence interval specifies range within which the parameter is estimated to lie (e.g., probability envelope).

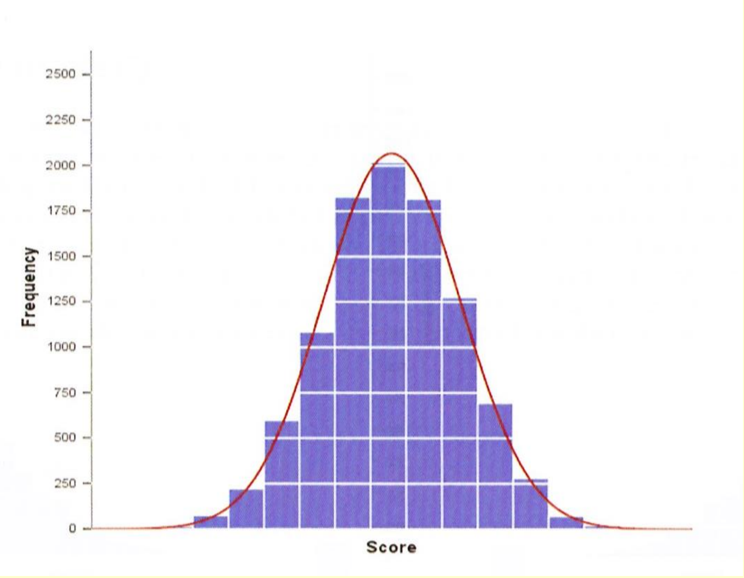
To show reliability of an estimate, confidence intervals are reported along with the point estimate: 49 +/- 9 C.I.

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

Summary – Statistical Estimation

Information about shape of the frequency distribution is critical for estimation.

We use two types of summary statistics:

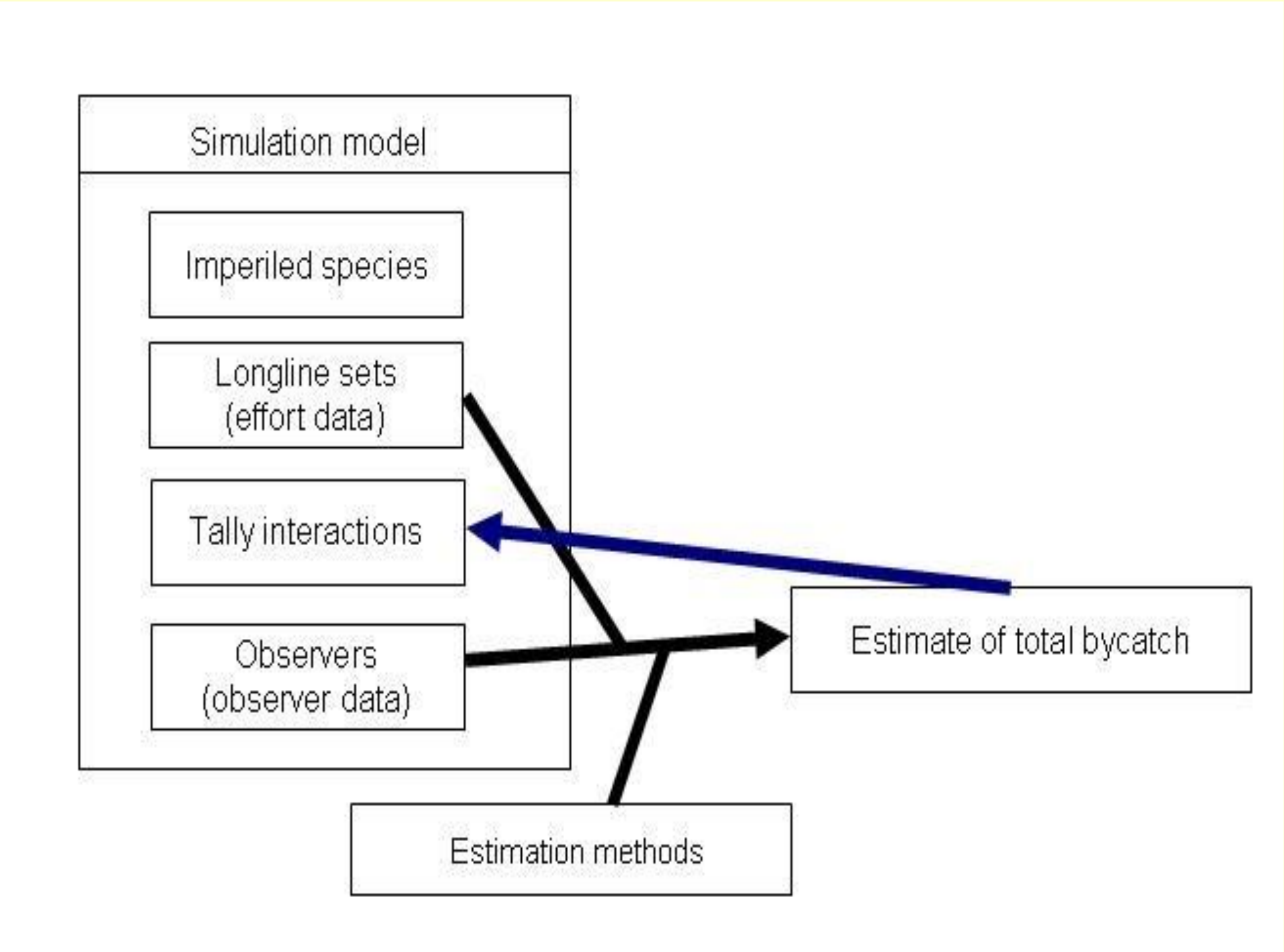
Measures of central tendency (location) describe where majority of the observations are found in the frequency distribution

e.g., Mode, Median, Mean

Measures of spread describe how variable the observations are, about the central location

e.g., Variance, Standard Deviation, IQ Range

Estimation in the Real World - Seaturtle Bycatch



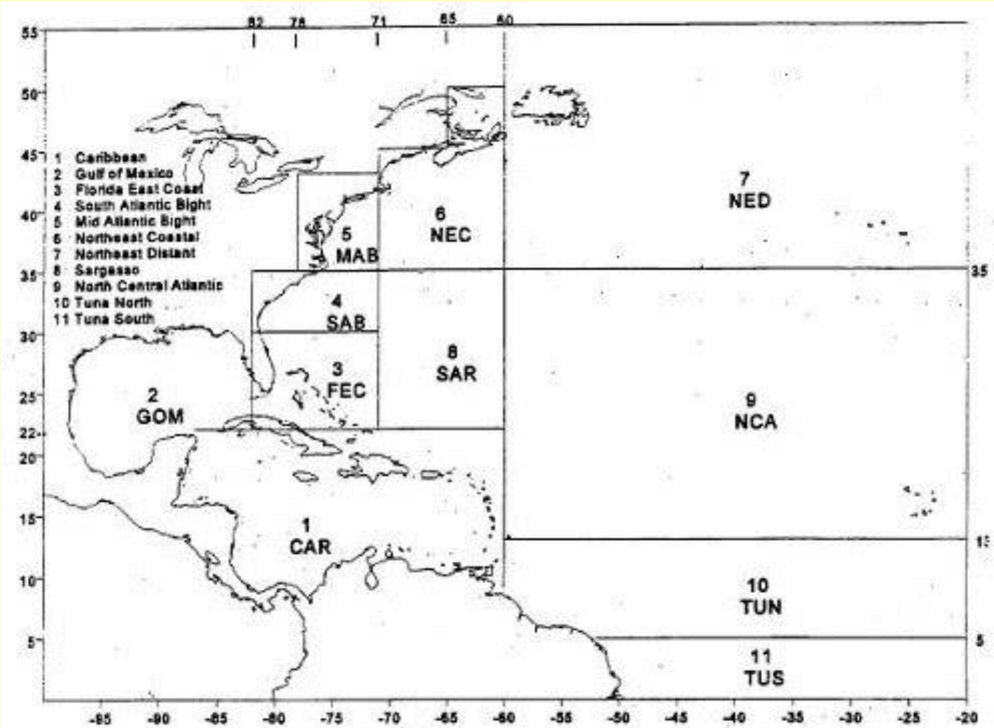
Estimation of "population parameters" using observed data (sample).

(turtles / hooks)

Extrapolation of that rate using fishery-wide effort.

(turtles)

Estimation - Bycatch



Fishing areas used to stratify bycatch estimates in U.S. Atlantic longline fishery

Many difficulties in making fishery-wide bycatch estimates:

Observers deployed on 5-8% of trips.

Some fishing areas have low historical observer coverage.

Variability across fishing areas inhibits extrapolation

Estimation - Bycatch

$$(1)\quad C_t = \frac{m_t}{n_t} e^{L_t} G\left(s_{L_t}^2/2\right),$$

where:

m_t is the number of sets with observed bycatch,

n_t is the total number of observed sets,

L_t is the mean of the log-transformed number of animals taken per 1000 hooks when bycatch occurred,

s_L^2 is the observed sample variance of the log transformed bycatch rate, and

G is the cumulative probability function from the Poisson distribution given as:

$$(2)\quad G\left(s_L^2/2\right) = 1 + \frac{m_t - 1}{m_t}\left(s_L^2/2\right) + \sum_{j=2}^{\infty} \frac{(m_t - 1)^{2j-1}}{m_t^j (m_t + 1)(m_t + 3)\ldots(m_t + 2j - 3)} \times \frac{\left(s_L^2/2\right)^j}{j!}.$$

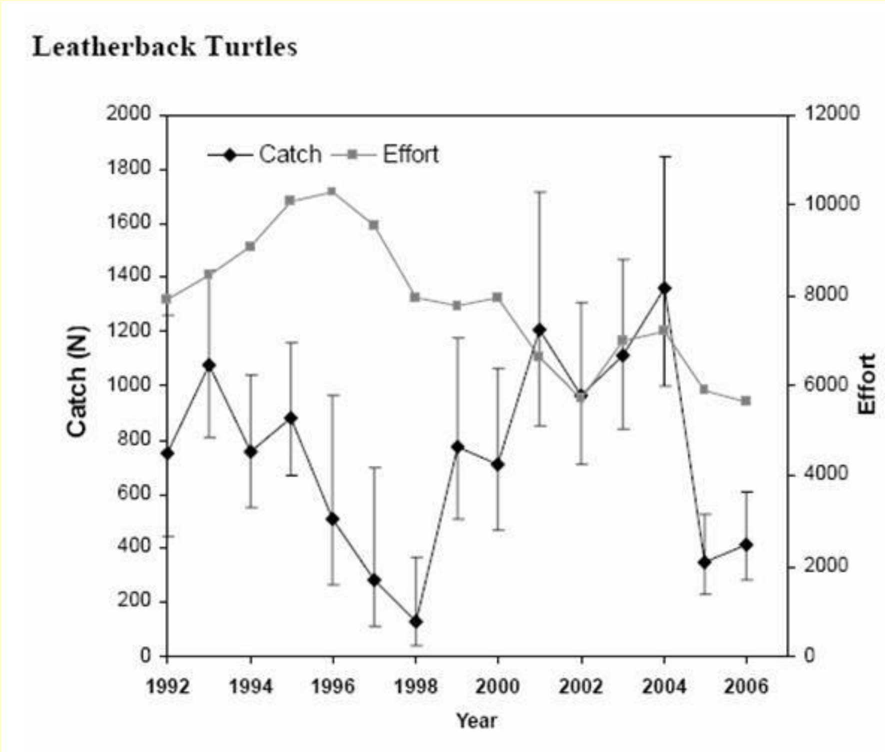
$$(3)\quad \mathrm{var}(C_t) = \frac{m_t}{n_t}\left(e^{2L_t}\left[\frac{m_t}{n_t} G^2\left(s_L^2/2\right) - \left(\frac{m_t - 1}{n_t - 1}\right) G\left(\frac{m-2}{m-1} s_L^2\right)\right]\right).$$

Probability of set with bycatch

Number of bycaught turtles per set

Information about frequency distribution of turtle bycatch numbers per set

Estimation - Bycatch



Estimates of annual Atlantic leatherback sea turtle bycatch represented by the black dots (+/- 95% confidence intervals. The grey line represents fishing effort (X 1000s hooks).

Fairfield-Walsh, C. Garrison, L. 2007. Estimated Bycatch of Marine Mammals and Turtles in the U.S. Atlantic Pelagic Longline Fleet During 2006. NOAA Technical Memorandum NOAA NMFS-SEFSC-560: 54 p.