Basic Data Concepts

FOR 1001

Dr. Thom Erdle

Today's Objectives

- Data Types (data levels)
- Populations
- Population Parameters (statistical measures)
- Data Quality
- Data Analysis

We will have a data analysis tutorial session as part of Thursday's lab

Data Types

What?

- The nature of data you are recording/observing
- Different data types have different characteristics
- Not all data are the same

Data Types

So What?

- You will be obtaining & using data of different types

  Species
  Size
  Frequency
  Quality

- Valid ***analysis*** varies by data type
- Effective ***presentation*** varies by data type
- Important to know the ***differences***

Data Types

	Summary Statistics	Graphs	Analysis
Nominal			
Ordinal		*Data types differ in these three ways*	
Ratio			

Data Types

	Summary Statistics	Graphs	Analysis
Nominal	Frequency or % occurrences No mean	Pie charts Bar graphs	Lowest level Only names are meaningful
Ordinal	Frequency or % occurrences Mean (questionable)	Bar graphs	Adds order meaning to names
Ratio	Mean Standard Deviation	Bar graphs Scatter plots Time plots	Adds zero so ratios are meaningful

Data Types

Nominal

- Name (or category) assigned to element
- Assigned value (name) simply an identifying label
- No quantitative significance
- No implication of order

Examples

Species

Soil type

Data Types

Ordinal

- Meaningful order
- Intervals between values not equal
- Think of as ranks
- Name (or category) assigned to element but order has meaning (successive values imply directional change)

Examples

Tree health

Tree quality grade

Rank in size

Data Types	Summary Statistics	Graphs	Analysis
Nominal			
Ordinal			
Ratio			

Data Types

Ratio

- **Measured (not just classified)**
- **Discrete or continuous**

  Discrete: Can't be made finer
  Finite possibilities

  Continuous: Can be made finer
  "Infinite" possibilities

- **Mathematically versatile and powerful**

Examples

Number of cavity nest trees per hectare

Mass of carbon in trees

Tonnes of moose browse per hectare

Parts per million of suspended solids in water

Populations

What?

- **Entire group of "individuals" of a specific category within an area of interest**
- **Defined by the context of the problem or issue in question**

So What?

- **It is the entity of interest**
- **Defines what you want to know about**
- **Governs your data acquisition scheme**

e.g.
- stands in Noonan forest
- shade trees on UNB campus
- salmon in Miramichi River
- forestry/ENR students at UNB

Population Parameters

What?

- **Measure describing characteristics of a population (set of data)**
- **Governed by the distribution of values across member of the population**

So What?

- **Informative about that population**
- **Appear in multiple contexts & with great frequency**
- **Key to answering many questions/problems**

Population Parameters

What?

- **Measure describing characteristics of a population (set of data)**
- **Governed by the distribution of values across member of the population**

Examples

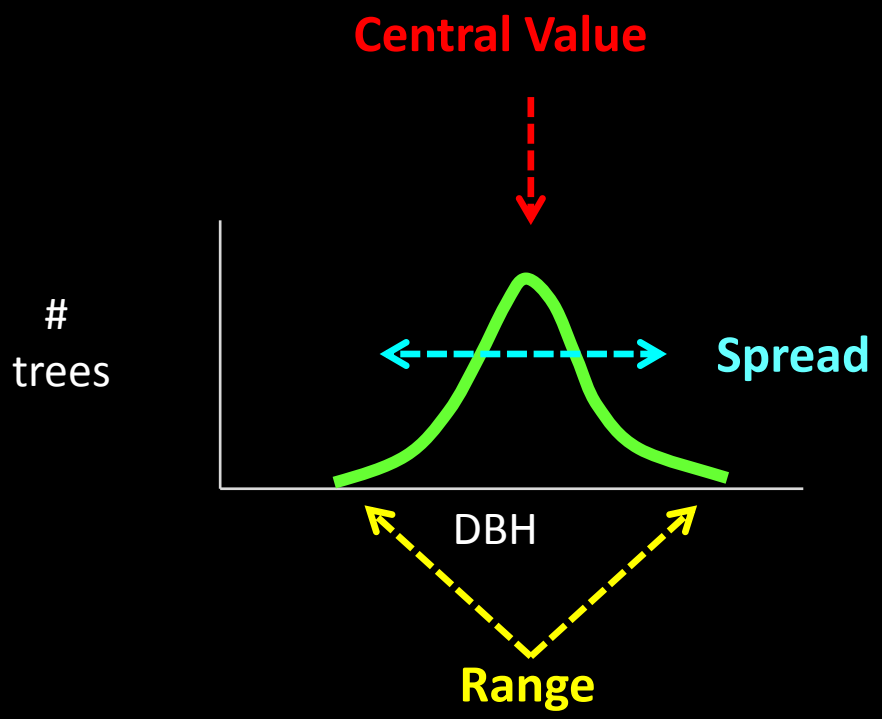
e.g.
- wood volume in stands in Noonan Forest
- # shade trees on UNB campus
- weight of fish in Miramichi River
- summer employment income of forestry /ENR students

Population Parameters

Central Tendency

Range

Spread



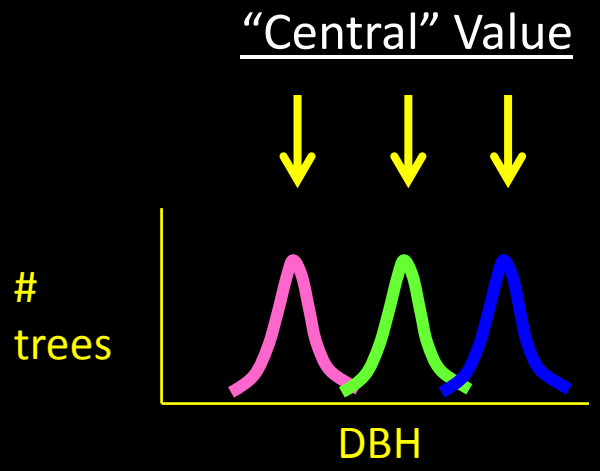
Population Parameters

Central Tendency

- **Describes the “central position” within a set of data**
- **Measures of the middle location in a set of data**

Population Parameters

Central Tendency



Central Values

Mean = $\frac{\text{sum values for each element}}{\text{number of elements}}$

Mode = most frequently occurring value

Median = half of values fall above; half fall below

Population Parameters

Central Tendency

Central Values

Mean = ∑ values / # elements
= 280 / 11
= 25.5 cm

Mode = most frequent value
= 22 cm

Median = midpoint value
= half values above; half below
= 24 cm

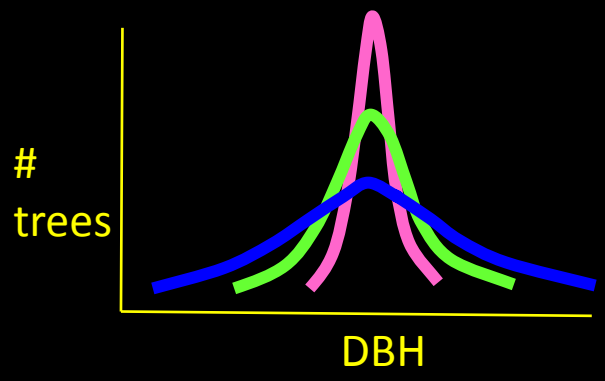
Tree	DBH_{cm}
1.	20
2.	22
3.	22
4.	22
5.	22
6.	24
7.	24
8.	26
9.	28
10.	34
11.	36
∑ = 280	

Population Parameters

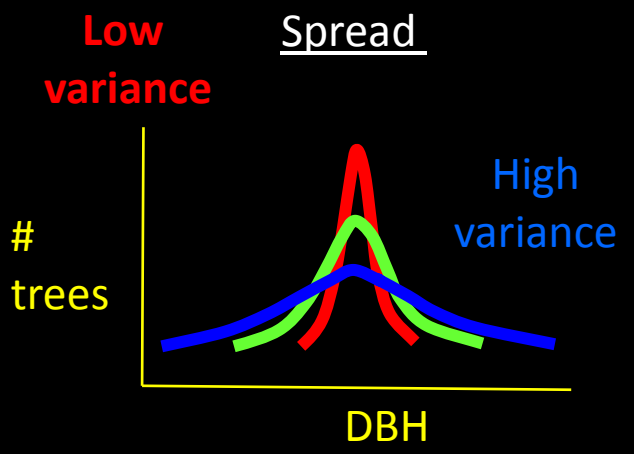
Spread

- **Relates to the variation of values within the population**
- **Of fundamental importance b/c influences how much sampling is required to achieve desired accuracy in estimate**

Spread



Population Parameters



Spread

Variance = average of squared differences from mean value

Variance = $\sum (Y_i - Y_{mean})^2/(n - 1)$

Standard Deviation = SQRT(Variance)

DBH_{cm}	$(Y_i - Y_{mean})^2$
20	$(20-24)^2 = 16$
22	$(22-24)^2 = 4$
24	$(24-24)^2 = 0$
26	$(26-24)^2 = 4$
28	$(28-24)^2 = 16$
$\sum = 120$	$\sum = 40$
Mean = 120/5	Var = 40 / 4
Mean = 24cm	Var = $10cm^2$
	StDev = 3.16cm

Today's Objectives

- Data Types (data levels)
- Populations
- Population Parameters (statistical measures)
- Data Quality
- Data Analysis

We will have a data analysis tutorial session as part of Thursday's lab

Data Quality

What?

- How good are my data?
- What are the characteristics of the data?

So What?

- Critical for determining how much confidence one can place on data
- Identifies potential errors/deficiencies to be guarded against

Data Quality

Accuracy

- **How close are the measured values to the true value**

Precision

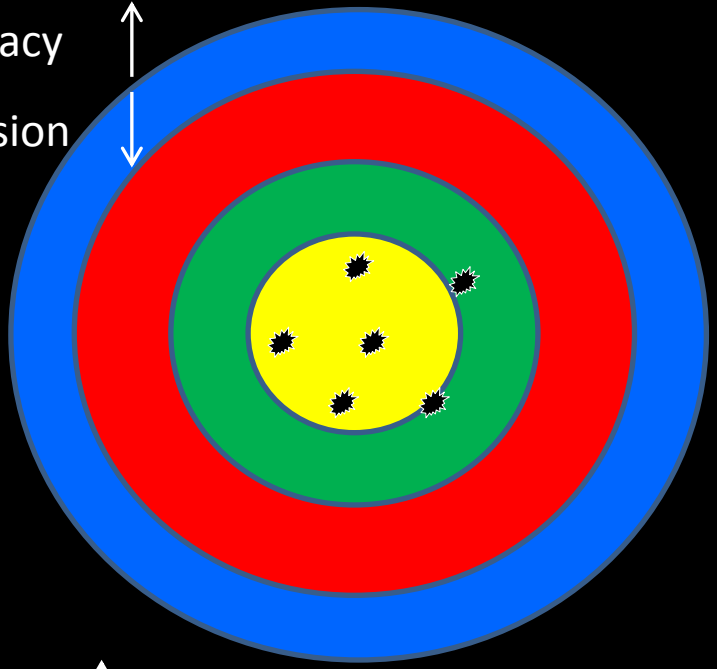
- **How tightly clustered measured values are**

Bias

- **Systematic distortion from the actual value**

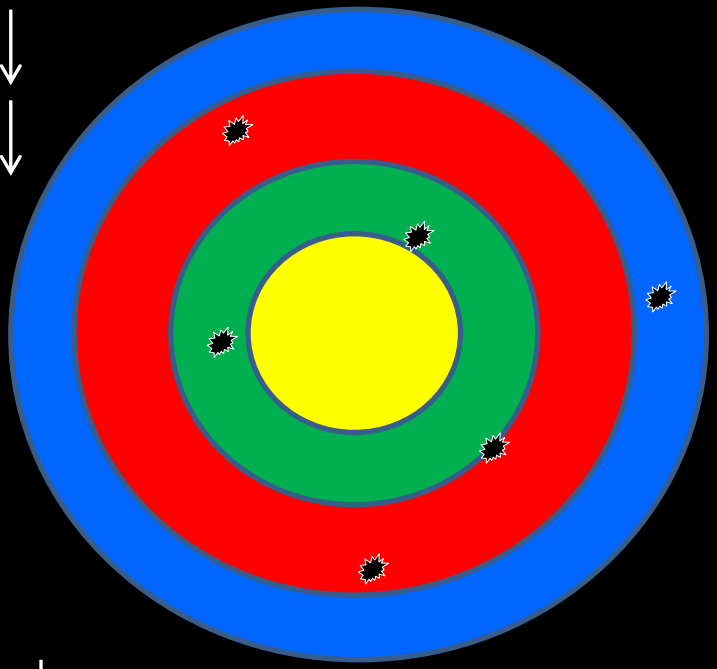
accuracy ↑

precision ↓



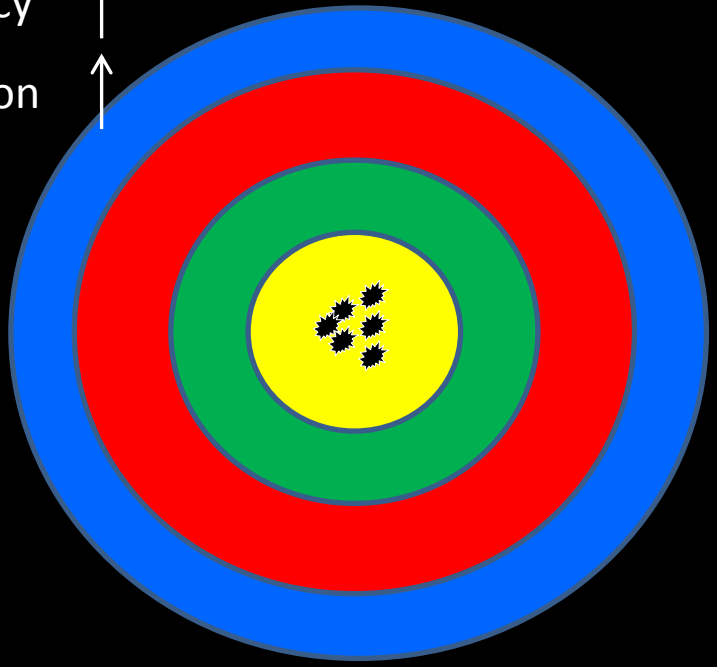
accuracy ↓

precision ↓



accuracy ↑

precision ↑



accuracy ↓

precision ↑