Populations, Samples, & Data Summary in Nat. Resource Mgt.

ESRM 304

Sampling in Natural Resources Management

I. Basic Concepts

II. Tools of the Trade

III. A Most Important Distribution

I. Basic Concepts

A. Why sample?
B. Populations, Parameters, Estimates
C. Variables - continuous, discrete
D. Bias, Accuracy, Precision
E. Distribution functions

I. Basic Concepts

A. Why sample?

1. Partial knowledge is a normal state
2. Complete enumeration is impossible
3. Complete enumeration is too expensive
4. Results are needed in a timely manner

I. Basic Concepts

B. Populations, Parameters, Estimates

1. Population: An aggregate of unit values
2. Parameter: A constant used to characterize a particular population
3. Estimate: A value calculated from a sample in a way that makes it a ‘good’ approximation to a parameter

Statistic: A value calculated from a sample

I. Basic Concepts

C. Variables - continuous & discrete

1. Continuous: A variable that can be measured using a numerical scale that can be subdivided, if desired, into an infinite number of smaller values
2. Discrete: Two (2) types:-

   a) Attributes: binomial –or– multinomial

   b) Counts

I. Basic Concepts

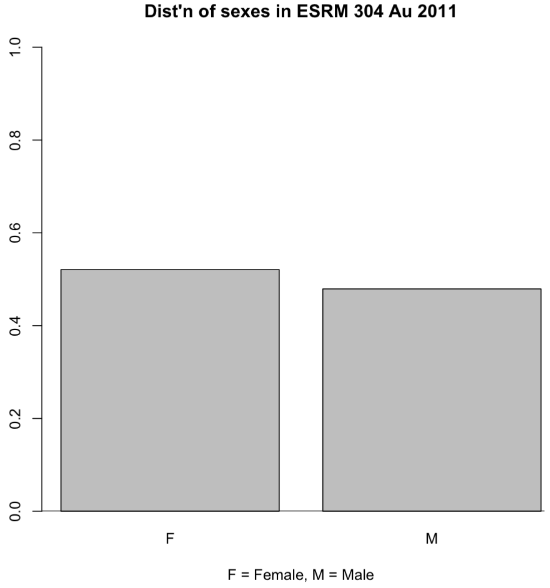
D. Bias, Accuracy, Precision

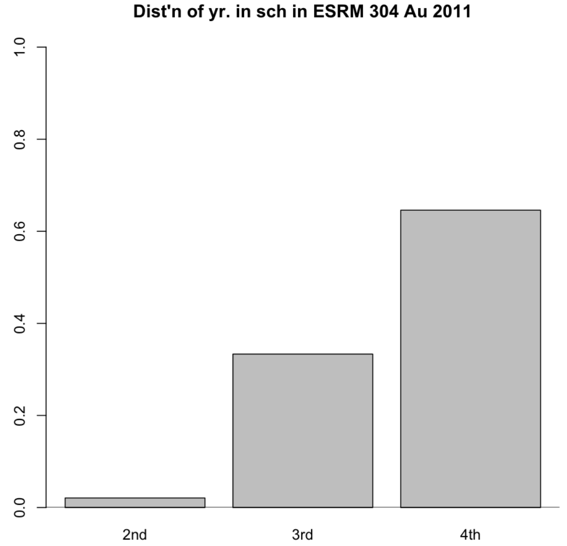
1. Bias:- Systematic distortion
2. Accuracy:- Nearness to true (or population) value
3. Precision:- clustering of unit values to their own mean

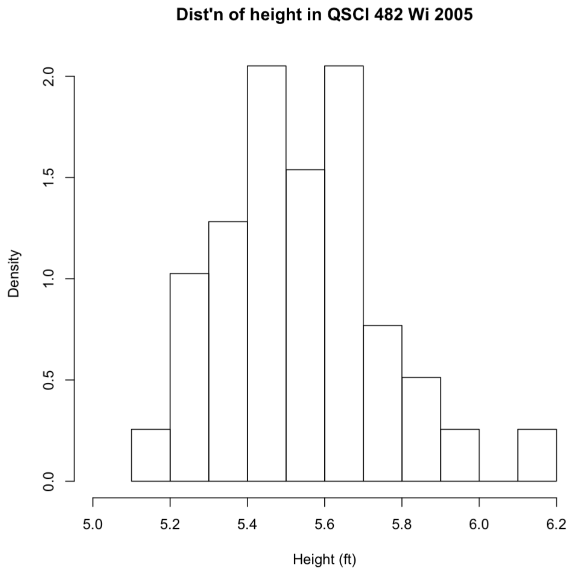
I. Basic Concepts

E. Distribution functions

Show for a sample (or population) the relative frequency with which different values occur

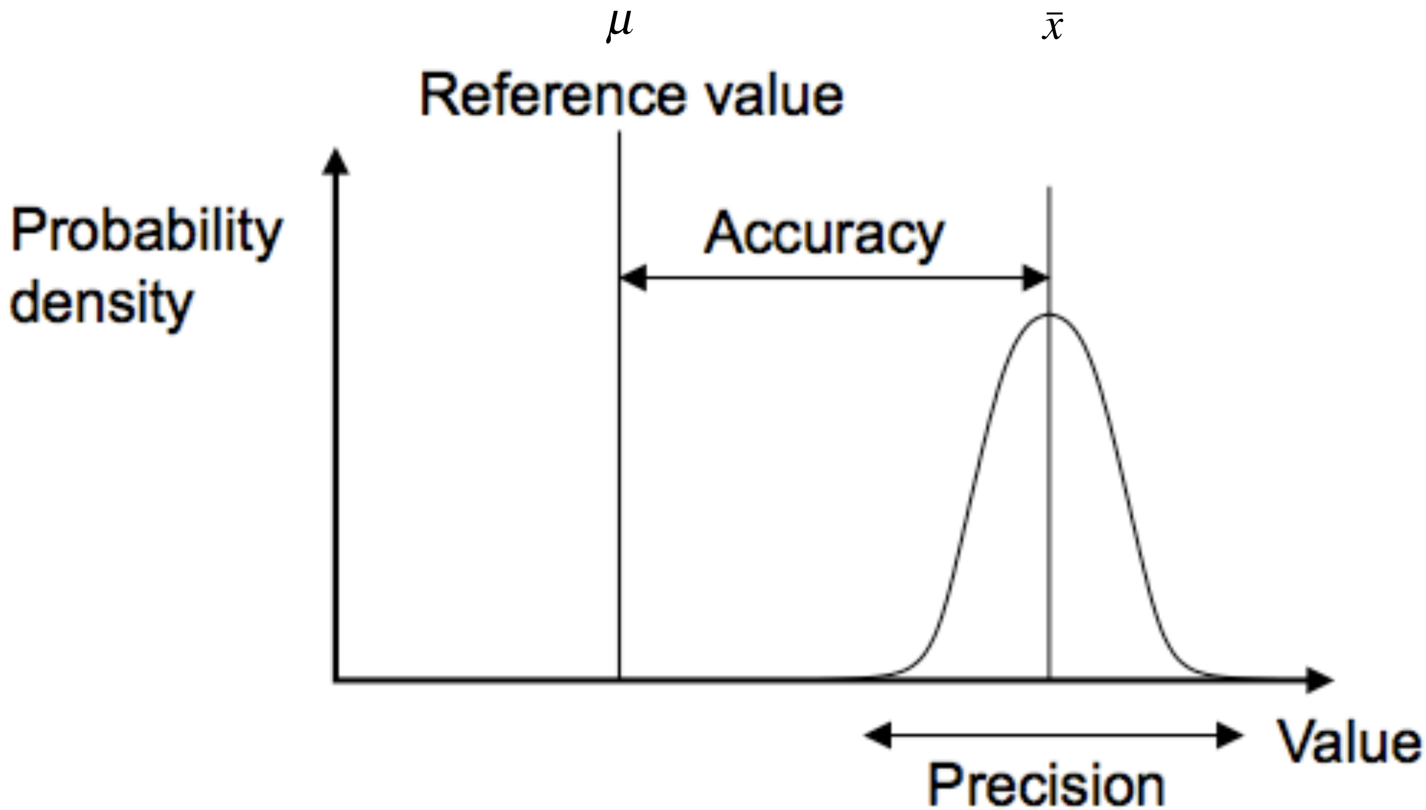






I. Basic Concepts

Another way to look at Bias, Accuracy, Precision



II. Tools of the Trade

A. Subscripts, Summations, Brackets
B. Mean, Variance, Standard Deviation
C. Standard Error of the estimate
D. Coefficient of Variation
E. Covariance, Correlation

II. Tools of the Trade

A. Subscripts, Summations, Brackets

A subscript can refer to a unit in a sample, *e.g.*,

x_1 is value on 1st unit, x_2 is value on of 2nd, etc.,

… it can refer to different populations of values, *e.g.*, x_1 can refer to the value tree height, while x_2 can refer to the value tree diameter,

… there can be more than one subscript, *e.g.*, x_{ij} may refer to the j^{th} individual of the i^{th} species of tree, where $j = 1, \ldots, 50$; i = DF, WH, RC

II. Tools of the Trade

A. Subscripts, Summations, Brackets

To indicate that several (say 6) values of a variable, *x*, are to be added together, we could write

$$\left(x_1 + x_2 + x_3 + x_4 + x_5 + x_6\right)$$

or shorter

$$\left(x_1 + x_2 + \cdots + x_6\right)$$

shorter still

$$\sum_{i=1}^{6} x_i \quad \text{or even} \quad \sum_i x_i \quad \text{or just} \quad \sum x$$

II. Tools of the Trade

A. Subscripts, Summations, Brackets

Order of operations still apply using "sigma" notation, *e.g.*,

$$\sum_{i=1}^{3} x_i y_i = x_1 y_1 + x_2 y_2 + x_3 y_3$$

$$\left(\sum_{i=1}^{3} x_i\right)\left(\sum_{i=1}^{3} y_i\right) = \left(x_1 + x_2 + x_3\right)\left(y_1 + y_2 + y_3\right)$$

$$\left(\sum_{i=1}^{3} x_i^2\right) \neq \left(\sum_{i=1}^{3} x_i\right)^2$$ i.e., $$\left(x_1^2 + x_2^2 + x_3^2\right) \neq \left(x_1 + x_2 + x_3\right)^2$$

II. Tools of the Trade

B. Mean, Variance, Standard Deviation

Mean: $$\bar{x} = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Variance: $$s_x^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{(n-1)} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{(n-1)}$$

Standard Deviation: $$s_x = \sqrt{s_x^2}$$

II. Tools of the Trade

B. Mean, Variance, Standard Deviation - Example

Let's say we have measurements on 3 units sampled from a large population. Values are 7, 8, and 12 ft.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{3}\left(7+8+12\right) = 9\ ft$$

$$s_x^2 = \frac{\left(7^2+8^2+12^2\right)-\frac{1}{3}\left(7+8+12\right)^2}{2} = 7\ ft^2$$

$$s_x = \sqrt{s_x^2} = \sqrt{7\ ft^2} = 2.64\ ft$$

II. Tools of the Trade

C. Standard Error of an estimate

- ✓ The most frequently desired estimate is for the mean of a population
- ✓ We need to be able to state how reliable our estimate is
- ✓ Standard error is key for stating our reliability
- ✓ Standard error quantifies the dispersion between an estimate derived from different samples taken from the same population of values
- ✓ Standard deviation of the observations is the square root of their variance, standard error (of an estimate) is the square-root of the variance of the estimate

II. Tools of the Trade

C. Standard Error of an estimate - Example

Let's say we have a population of (N = 15) tree heights:

7, 10, 8, 12, 2, 6, 5, 9, 3, 7, 4, 8, 9, 11, 5 from which we take 4 units (n = 4) five separate times …

pick 1 (units 10, 8, 3, 11): 7, 9, 8, 4; $\bar{x} = 7; \ s = 2.16$

pick 2 (units 5, 3, 6, 4) : 2, 8, 6, 12; $\bar{x} = 7; \ s = 4.16$

pick 3 (units 8, 11, 3, 13): 9, 4, 8, 9; $\bar{x} = 7.5; \ s = 2.38$

pick 4 (units 9, 14, 11, 5): 3, 11, 4, 2; $\bar{x} = 5; \ s = 4.08$

pick 5 (units 5, 3, 2, 10) : 2, 8, 10, 7; $\bar{x} = 6.75; \ s = 3.40$

… there are 1,365 possible unique samples of size 4 !!!

II. Tools of the Trade

C. Standard Error of an estimate - Example (cont' d)

If we used Simple Random Sampling (SRS), there is a very direct way to calculate standard error of the estimated (sample) mean

In words: standard deviation divided by the square-root of the sample size

In formula: $s_{\bar{x}} = \frac{s_x}{\sqrt{n}}$

$$\textit{pick}\ 1:\ 1.08;\quad 2:\ 2.08;\quad 3:\ 1.19;\quad 4:\ 2.04;\quad 5:\ 1.70$$

Population mean = 7.07; std.dev = 2.91; std.err = 1.457

II. Tools of the Trade

D. Coefficient of Variation

- ✓ Puts variability on a relative scale so we can compare the dispersions of values measured in different units (say feet and meters) or the dispersion of different populations (say heights and weights)
- ✓ Ratio of standard deviation to the mean

II. Tools of the Trade

D. Coefficient of Variation - Example

Using the previous tree height population …

pick 1: $\bar{x} = 7; \; s = 2.16$

$$C = \frac{s}{\bar{x}} = \frac{2.16}{7} = 0.308$$ or, ~ 31 %

If inches had been used, $\bar{x} = 84; \; s = 25.92$

$$C = \frac{s}{\bar{x}} = \frac{25.92}{84} = 0.308$$

II. Tools of the Trade

E. Covariance, Correlation

- ✓ In some situations, we'd like to know if two variables (call one *x*, the other *y*) are associated with each other
- ✓ If the association is direct, covariance is positive
- ✓ If indirect, covariance is negative
- ✓ If not associated, covariance is nearly zero

$$s_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\left(n-1\right)} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\left(n-1\right)}$$

II. Tools of the Trade

E. Covariance - Example

We have a sample of units from a population on which we measured values of two variables

i	1	2	3	4	5	6	Totals
x_i	2	12	7	14	11	8	54
y_i	12	4	10	3	6	7	42

$$s_{xy} = \frac{\left(2\cdot 12 + 12\cdot 4 + \cdots + 8\cdot 7\right) - \frac{1}{6}\left(54\right)\left(42\right)}{6-1} = -14.4$$

II. Tools of the Trade

E. Covariance, Correlation

- ✓ As with variance, the magnitude of the covariance can be related to magnitude of the unit values
- ✓ A measure of the degree of association that is unaffected by size of unit values (like coefficient of variation) is the correlation coefficient
  - ➢ Correlation coefficient varies between -1 and +1
  - ➢ Closer it is to 1 (either sign), the stronger the association it is

II. Tools of the Trade

E. Correlation - Example

$$r = \frac{\text{covariance of x and y}}{\sqrt{(\text{variance of x})(\text{variance of y})}} = \frac{s_{xy}}{\sqrt{(s_x^2)(s_y^2)}}$$

$$r = \frac{s_{xy}}{\sqrt{(s_x^2)(s_y^2)}} = \frac{-14.4}{\sqrt{(12.0)(18.4)}} = -0.969$$

III. A Most Important Distribution

The Normal Distribution

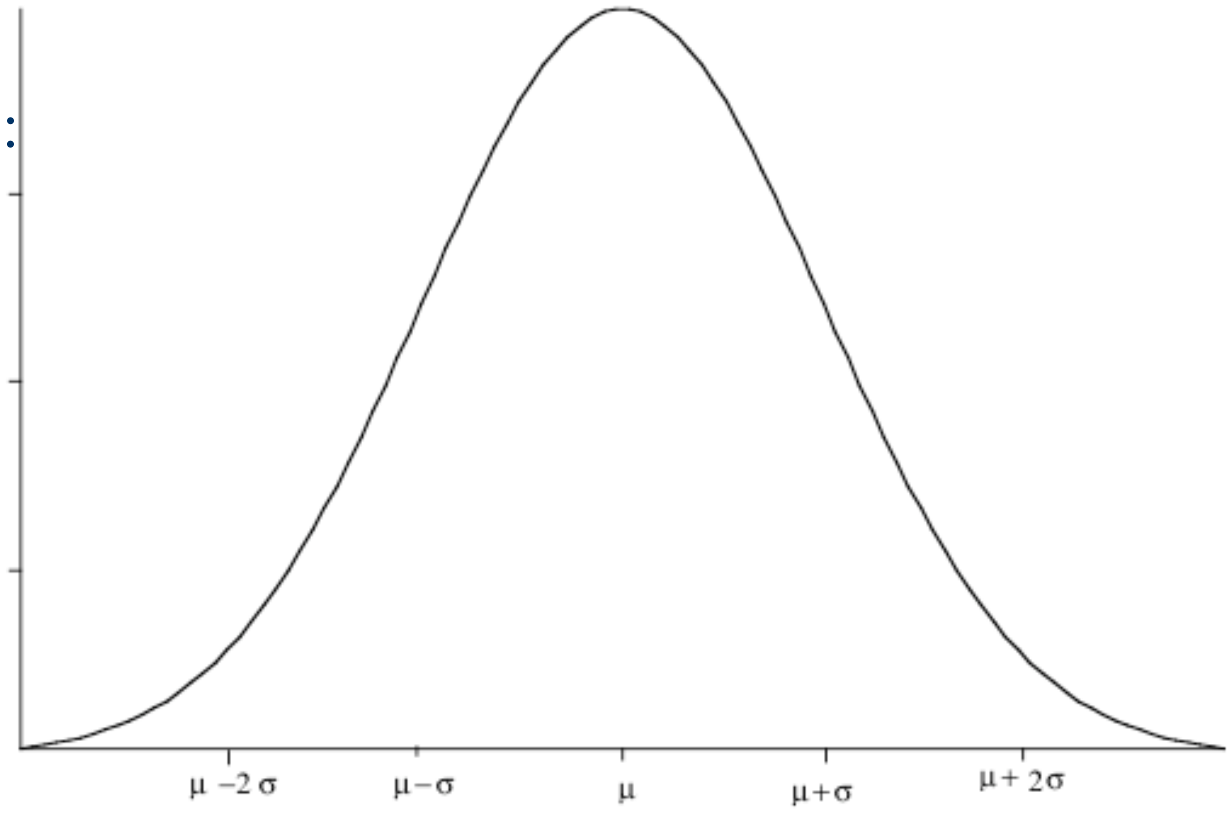
Greek symbols denote parameters:

Mean: μ

Variance: σ^2

English (latin-based) letters denote statistics:

$\bar{x}, \quad s^2$



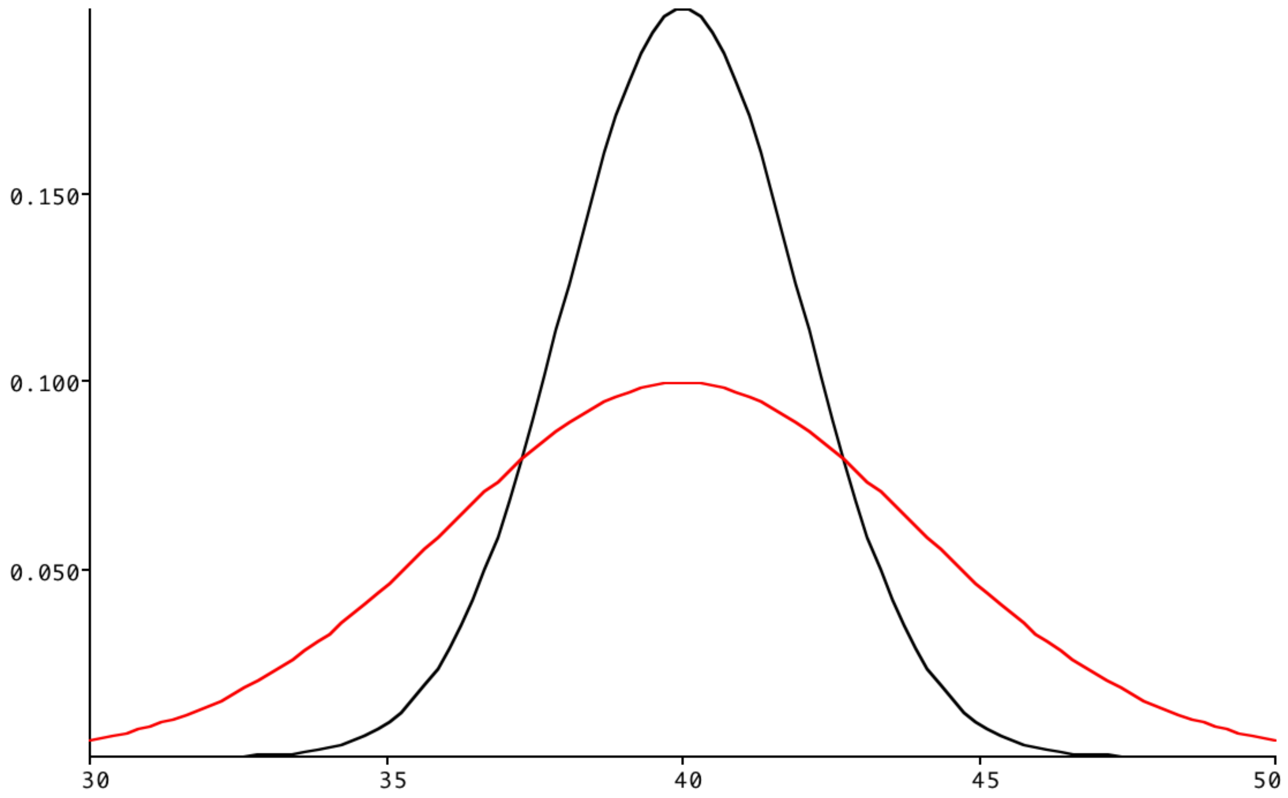
III. A Most Important Distribution

Properties of the Normal Distribution

- The distribution is bell-shaped; symmetrical about mean
- The mean locates the center of the distribution.
- The standard deviation is the distance between the mean and the inflection point of the distribution function.
- The distribution covers the entire real number line, from $-\infty$ to $+\infty$
- It has two parameters: the mean, μ and variance, σ^2

III. A Most Important Distribution

A couple of Normal Distributions



III. A Most Important Distribution

Why all the fuss about the Normal?

It has a variety of uses:

- Many populations found in nature are distributed approximately this way
- Used to calculate the chances a value within a certain range will occur
- Describing experimental error (calculating confidence)
- The distribution of sample means is approximately Normal (Central Limit Theorem)

III. A Most Important Distribution

Why all the fuss about the Normal?

Used to calculate the chances a particular value will be observed within a population (or a range of values)

- Any random variable X following a Normal distribution with mean = μ and variance = σ^2 can be 'mapped' onto the so-called *Standard* Normal (or "Z" distribution, which has a mean of zero and a variance of one) by the following equation:

$$Z = \frac{X - \mu}{\sigma}$$

III. A Most Important Distribution

The Central Limit Theorem:

If the mean, $\bar{X}$ of a random sample $X_1, X_2, X_3, \ldots, X_n$ of size n arising from ANY distribution with a finite mean and variance is transformed into W, using the following equation:

$$W = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}, \qquad \text{where } \sigma_{\bar{X}} = \sqrt{\sigma^2 / n}$$

the distribution of W will *approach* that of a standard Normal deviate with mean = 0, and variance = 1 in the "limit," i.e., as sample size $n \to \infty$.

III. A Most Important Distribution

The Normal distribution does have its limits…

- Application of the normal dist'n assumes σ is known
  - Using it with unknown s.d. will *overstate confidence & reliability, especially* when we also have a small sample (n < ?)
- When we do not know population standard deviation (or variance), use *Student's t* distribution instead
  - The "t" distribution should be used *especially* when we also have a small sample
- Like the normal, "t" is symmetrical, spans $-\infty$ to $+\infty$
- Unlike the normal, a single parameter defines it, ν, i.e., the so-called *degrees of freedom (or df)*

III. A Most Important Distribution

The Central Limit Theorem (unknown σ)

If the mean, $\bar{X}$ of a random sample $X_1, X_2, X_3, \ldots, X_n$ of size *n* (where *n* is small) from a population distributed as a Normal is transformed into W, using the following equation:

$$W = \frac{\bar{X} - \mu}{S_{\bar{X}}}, \qquad \text{where } S_{\bar{X}} = \sqrt{S^2/n}$$

the distribution of W follows the "Student's t" distribution.

If the sample is large enough, W will still map onto the standard Normal (or "Z" distribution) even with unknown variance and unknown population dist'n

Things to Remember- Sampling in Nat. Resources Management

I. Basic Concepts

- ✓ Populations have parameters
- ✓ Samples have statistics (to estimate parameters)

II. Tools of the Trade

- ✓ Standard deviation is the square-root of variance
- ✓ Standard deviation (sd) and Standard Error (se) both quantify dispersion
  - ➢ SD for dispersion of sample values
  - ➢ SE for dispersion of sample mean values

Things to Remember- Sampling in Nat. Resources Management

III. A Most Important Distribution Function

- ✓ The normal distribution has nice properties for describing a population of values measured on a continuous scale (number line)
- ✓ The "Normal" does not do everything for us; we need to use the "t" distribution when pop'n variance is unknown and especially when we have small samples