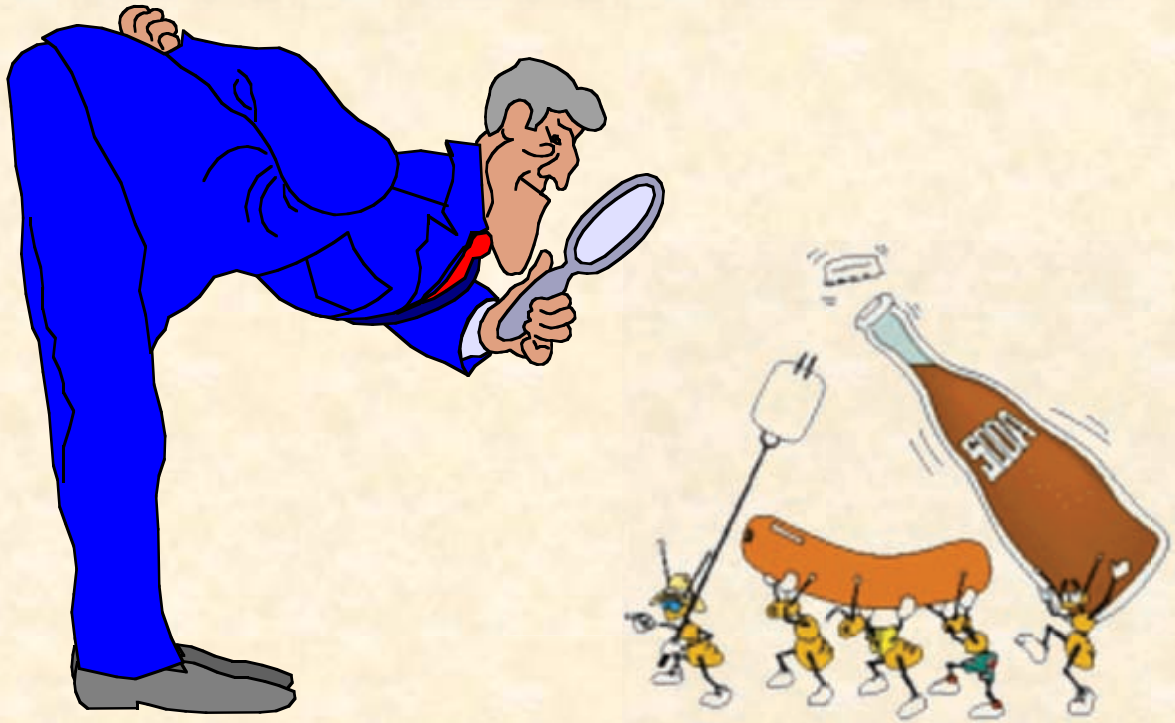
STATISTICS
For Research

Why Statistics?

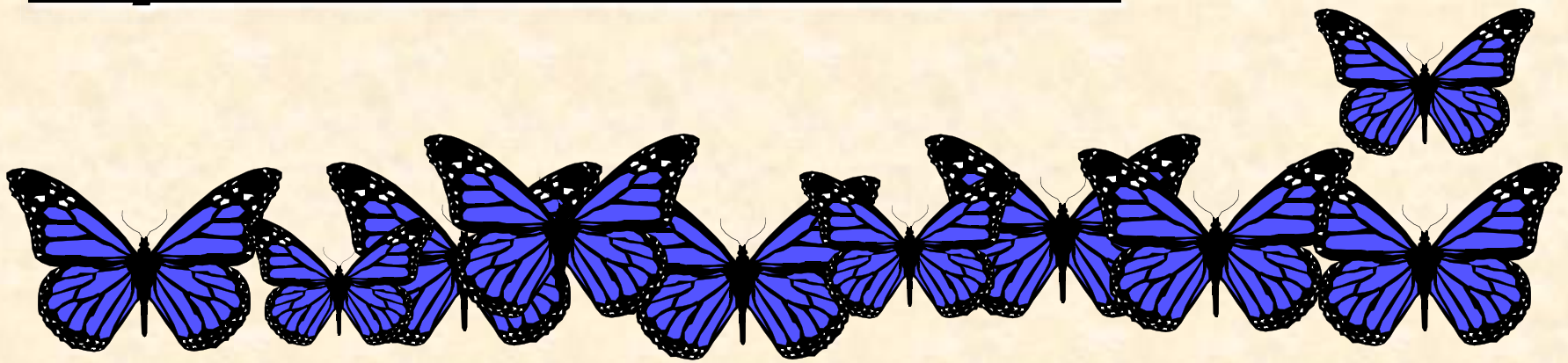
A Researcher Can:

1. *Quantitatively* describe and summarize data



A Researcher Can:

2. Draw conclusions about large sets of data by sampling only *small portions of them*



A Researcher Can:

3. *Objectively* **measure differences** and relationships between sets of data.

Random Sampling

- Samples should be taken at ***random***
- Each measurement has an ***equal opportunity*** of being selected
- Otherwise, sampling procedures may be ***biased***

Sampling Replication

- **A characteristic CANNOT be estimated from a single data point**

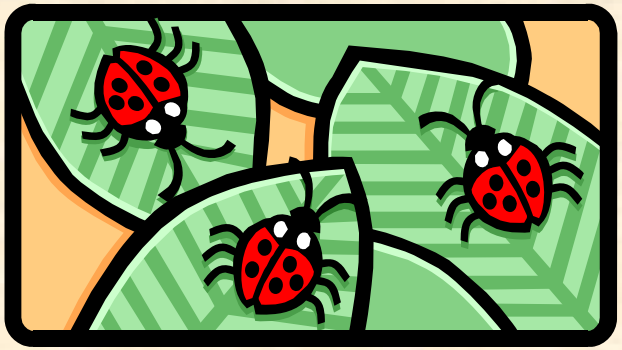
- ***Replicated*** **measurements should be taken, at least 10.**

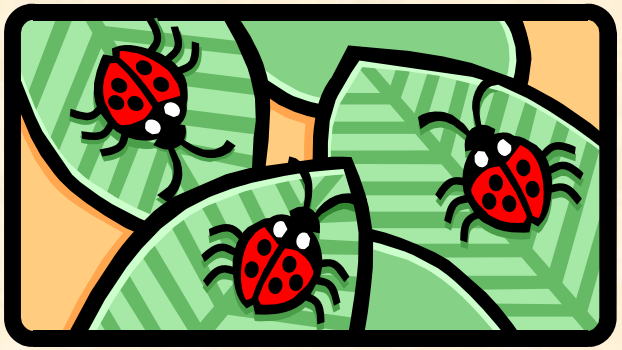
Mechanics

1. *Write down a **formula***
2. ***Substitute numbers** into the formula*
3. ***Solve** for the unknown.*

The Null Hypothesis

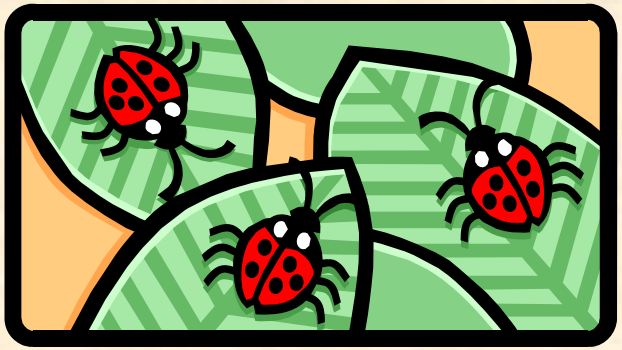
- H_o = There is no difference between 2 or more sets of data
  - any difference is due to chance alone
  - Commonly set at a probability of 95% ($P \leq .05$)

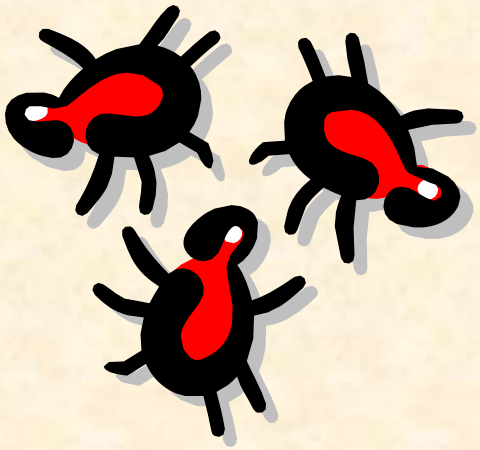




The Alternative Hypothesis

- H_A = There is a difference between 2 or more sets of data
  - the difference is due to more than just chance
  - Commonly set at a probability of 95% ($P \leq .05$)





Averages

- Population Average = mean (**x**)

- a ***Population mean*** = (χ)
  - take the mean of a *random sample* from the population (**n**)

Population Means

To find the population mean (χ),

- add up ( ) the values

  (**x** = grasshopper mass, tree height)

- divide by the number of values (**n**):

$$\chi = \frac{\sum x}{n}$$

Measures of Variability

- Calculating a mean gives only a *partial* description of a set of data
  - Set A = 1, 6, 11, 16, 21
  - Set B = 10, 11, 11, 11, 12
    - Means for A & B ??????

- ***Need a measure of how variable the data are.***

Range

- Difference between the largest and smallest values
  - Set A = 1, 6, 11, 16, 21
    - Range = ???
  - Set B = 10, 11, 11, 11, 12
    - Range = ???

Standard Deviation

Standard Deviation

- **A measure of the deviation of data *from their mean.***

The Formula

$$SD = \sqrt{\frac{N \sum X^2 - (\sum X)^2}{N (N-1)}}$$

SD Symbols

SD = Standard Dev

$\sqrt{}$ = Square Root

$\sum X^2$ = Sum of x^2'd

$\sum(X)^2$ = Sum of x's, then squared

N = # of samples

The Formula

$$SD = \sqrt{\frac{N \sum X^2 - (\sum X)^2}{N (N-1)}}$$

X	X^2
297	88,209
301	90,601
306	93,636
312	97,344
314	98,596
317	100,489
325	105,625
329	108,241
334	111,556
350	122,500
$\sum X = 3,185$	$\sum X^2 = 1,016,797$

$$\text{standard deviation} = \sqrt{\frac{N\sum X^2 - (\sum X)^2}{N(N-1)}} = \sqrt{\frac{10(1{,}093{,}597) - (3185)^2}{10(10-1)}}$$

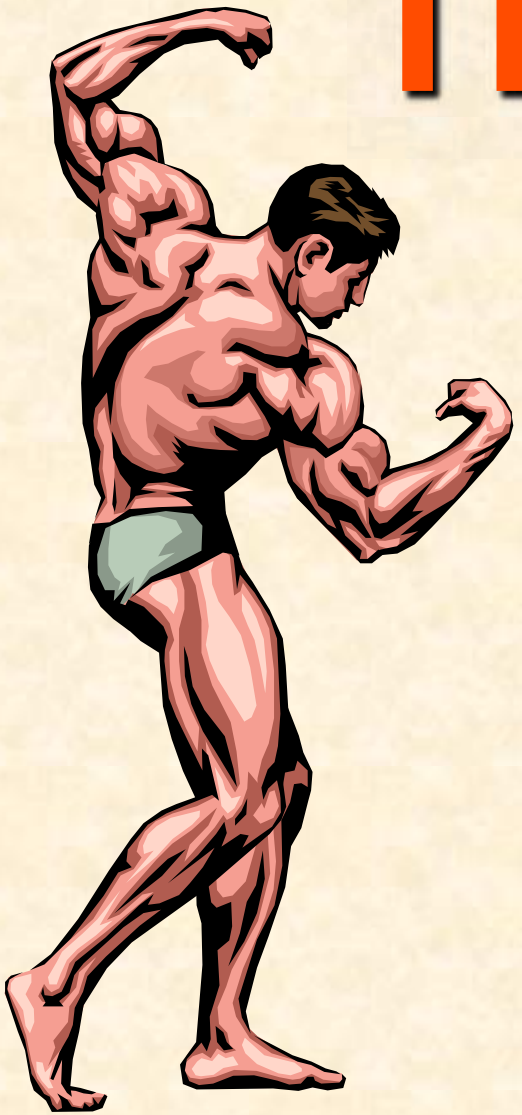
$$= \sqrt{\frac{10{,}935{,}970 - 10{,}144{,}225}{10(9)}} = \sqrt{\frac{781{,}645}{90}}$$

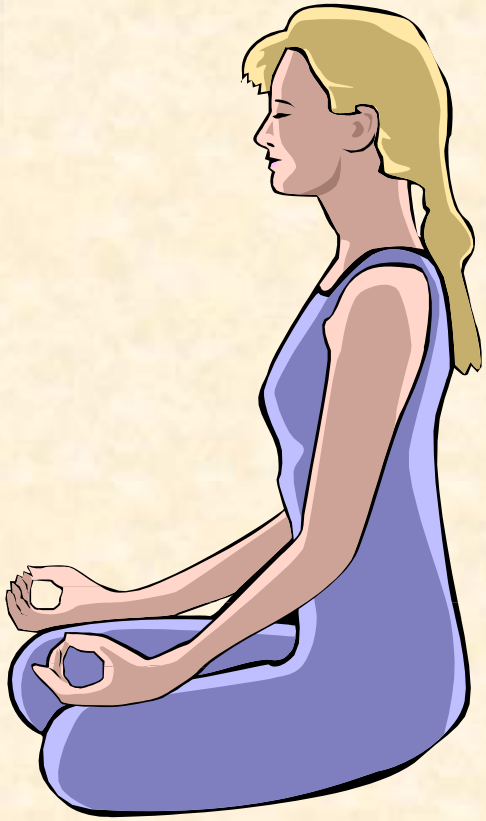
$$= \sqrt{8684.944} = 93.19$$

Once You've got the Idea:

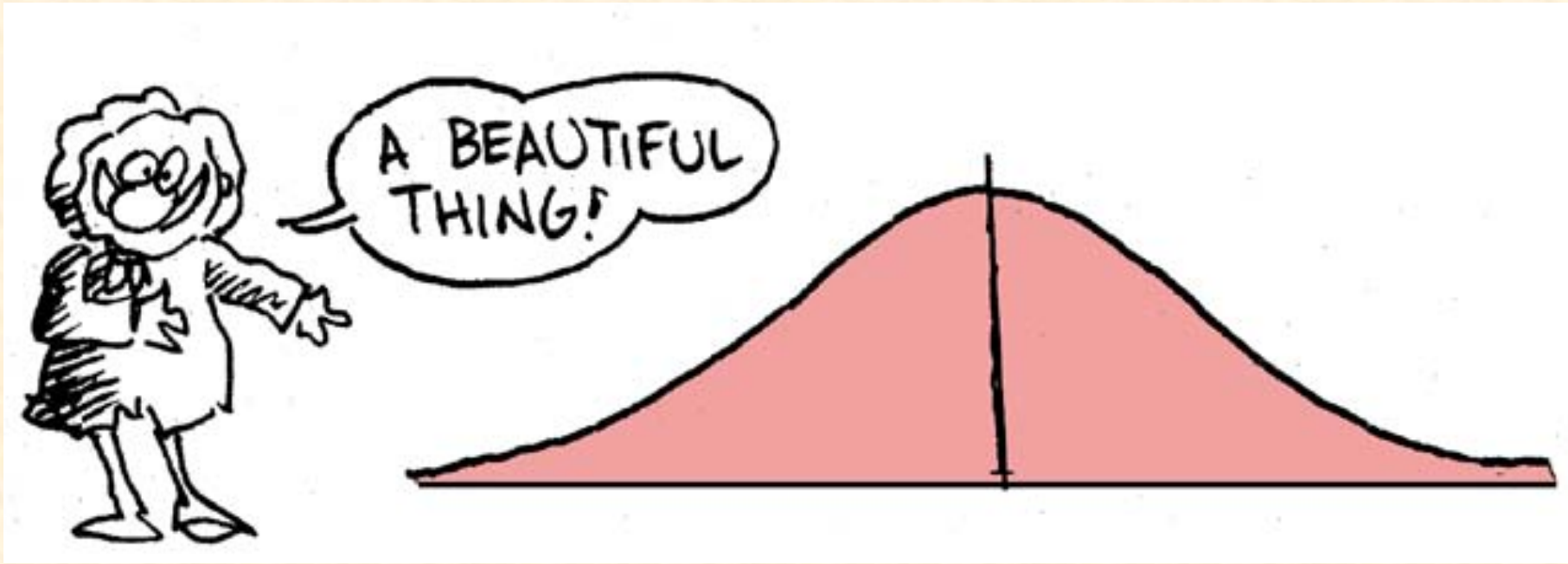
You can use your calculator to find SD!

The Normal Curve

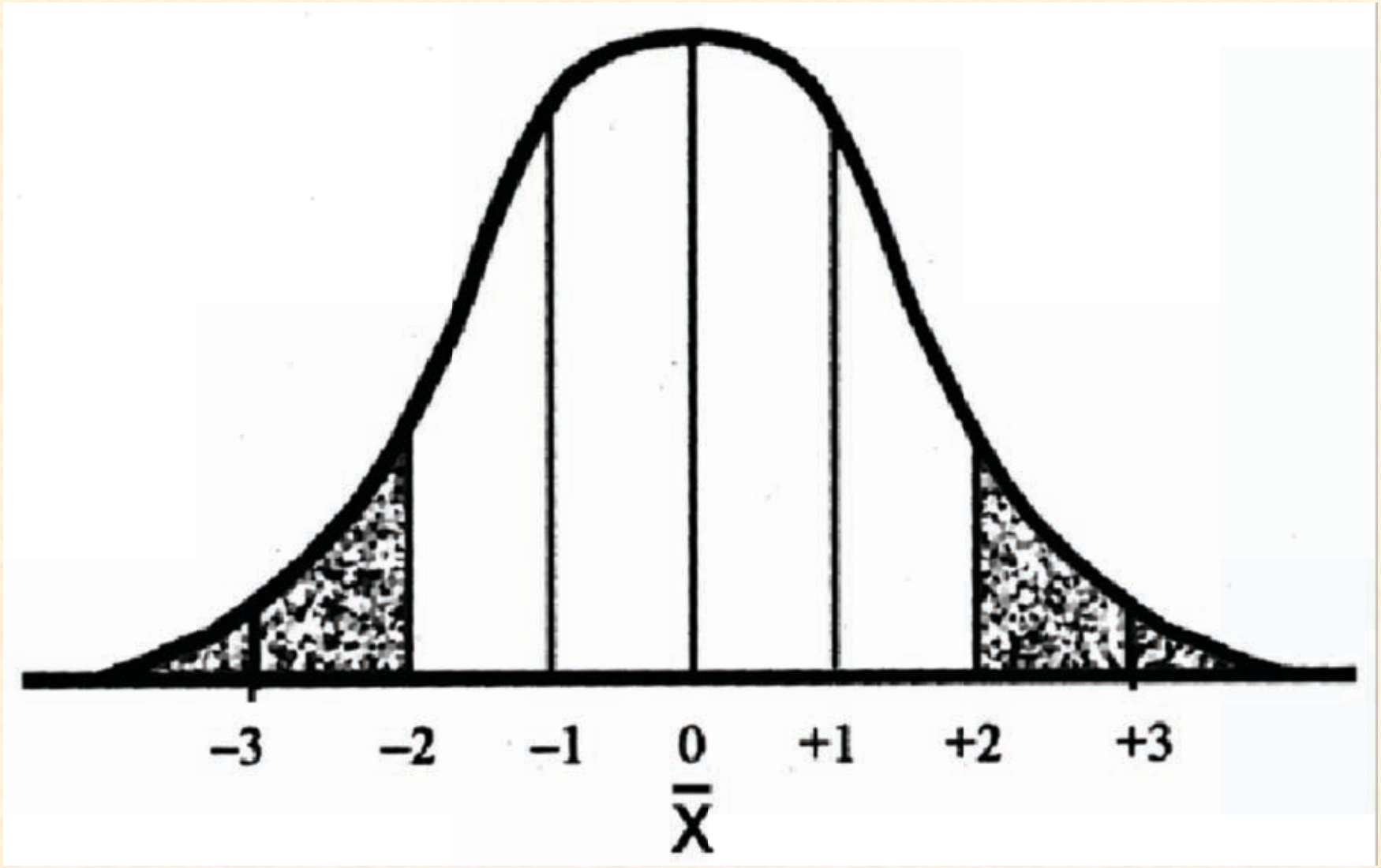




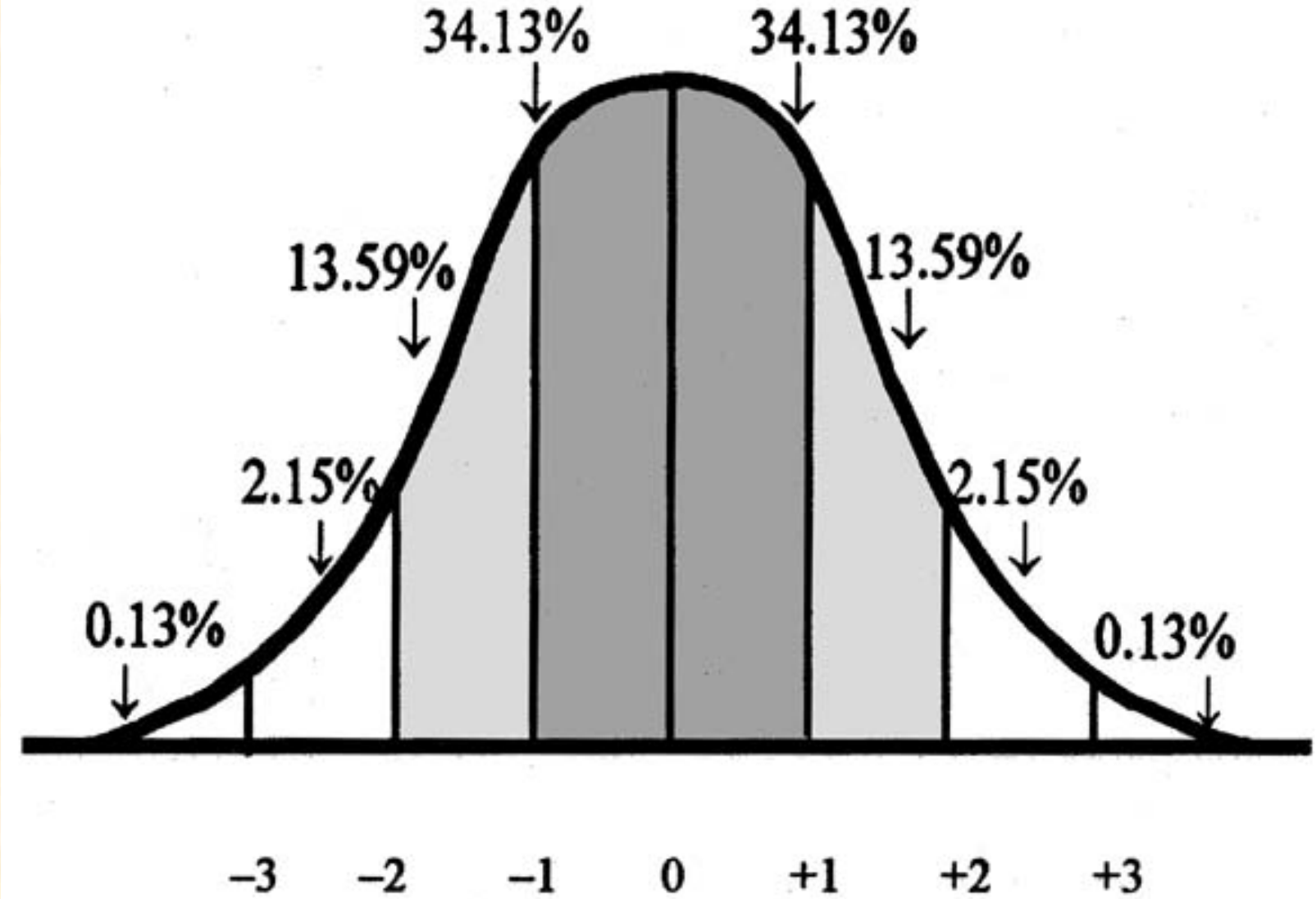
The Normal Curve

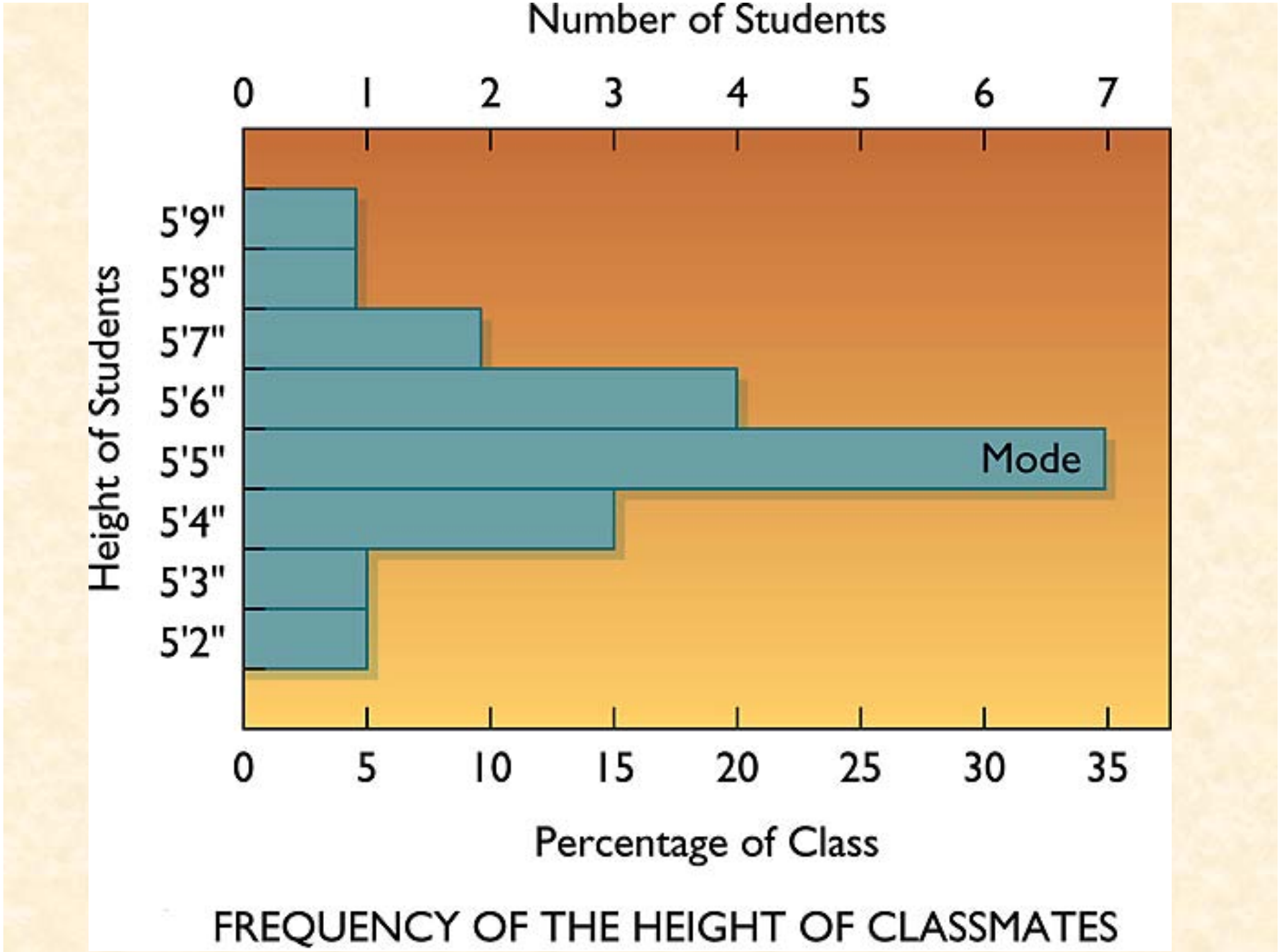


SD & the Bell Curve



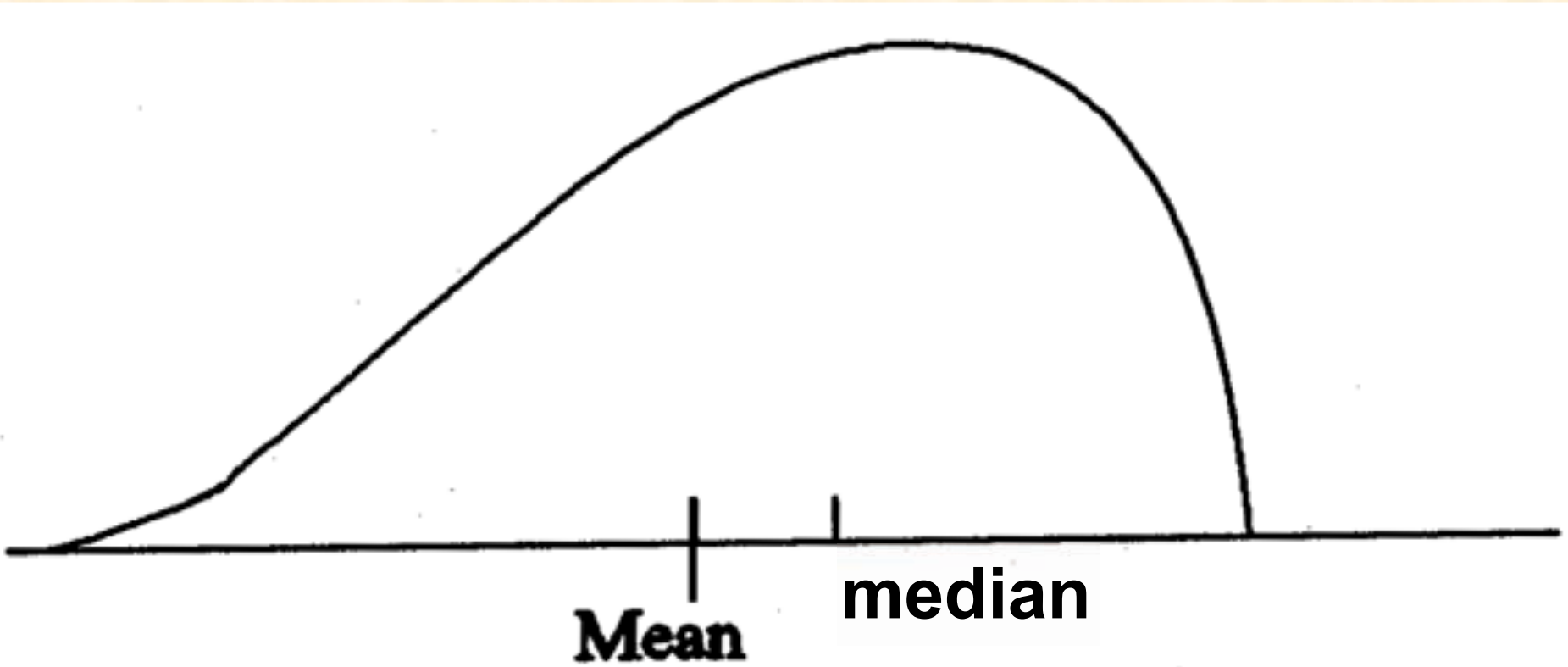
% Increments





FREQUENCY OF THE HEIGHT OF CLASSMATES

Skewed Curves



Critical Values

Standard Deviations ≥ 2 SD

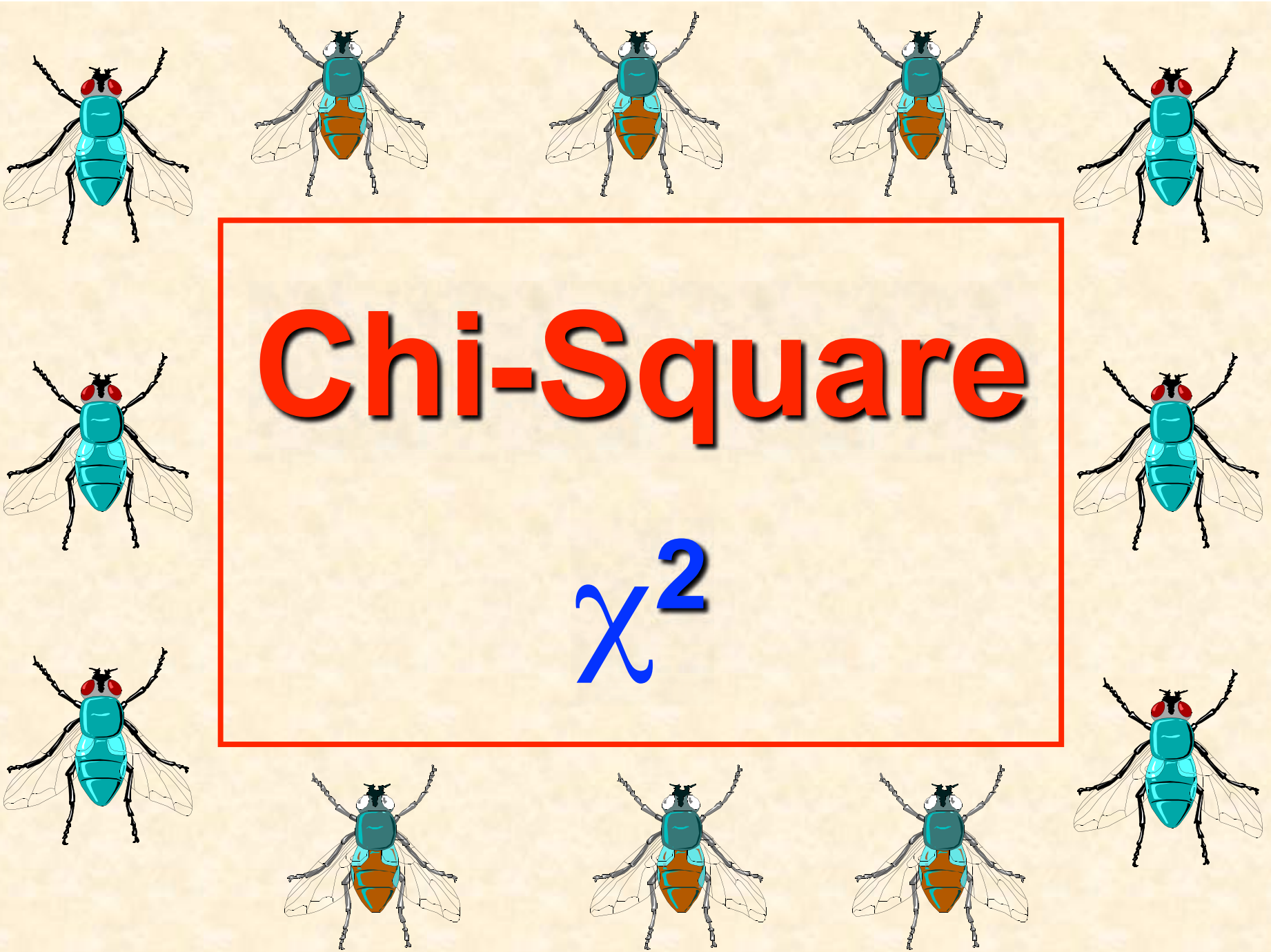
above or below the mean =

due to MORE THAN

CHANCE ALONE.

Critical Values

The data lies *outside* the 95% confidence limits for probability.


Chi-Square
χ^2

Chi-Square Test Requirements

- Quantitative data
- Simple random sample
- One or more categories
- Data in frequency (%) form

Chi-Square Test Requirements

- Independent observations
- All observations must be used
- Adequate sample size ($\geq$10)

Example

Table 1 - Color Preference for 150 Customers for Thai?s Car Dealership

Category Color	Observed Frequencies	Expected Frequencies
YELLOW	35	30
RED	50	45
GREEN	30	15
BLUE	10	15
WHITE	25	45

Chi-Square Symbols

$$\chi^2 = \frac{(O - E)^2}{E}$$

O = Observed Frequency

E = Expected Frequency

= sum of

df = degrees of freedom (n-1)

χ^2 = Chi Square

Chi-Square Worksheet

CATAGORY	*O*	*E*	*(O - E)*	$(O - E)^2$	$\frac{(O - E)^2}{E}$
YELLOW	35	30	5	25	0.83
RED	50	45	5	25	0.56
GREEN	30	15	15	225	15
BLUE	10	15	-5	25	1.67
WHITE	25	45	-20	400	8.89

$\chi^2 = 26.95$

Chi-Square Analysis

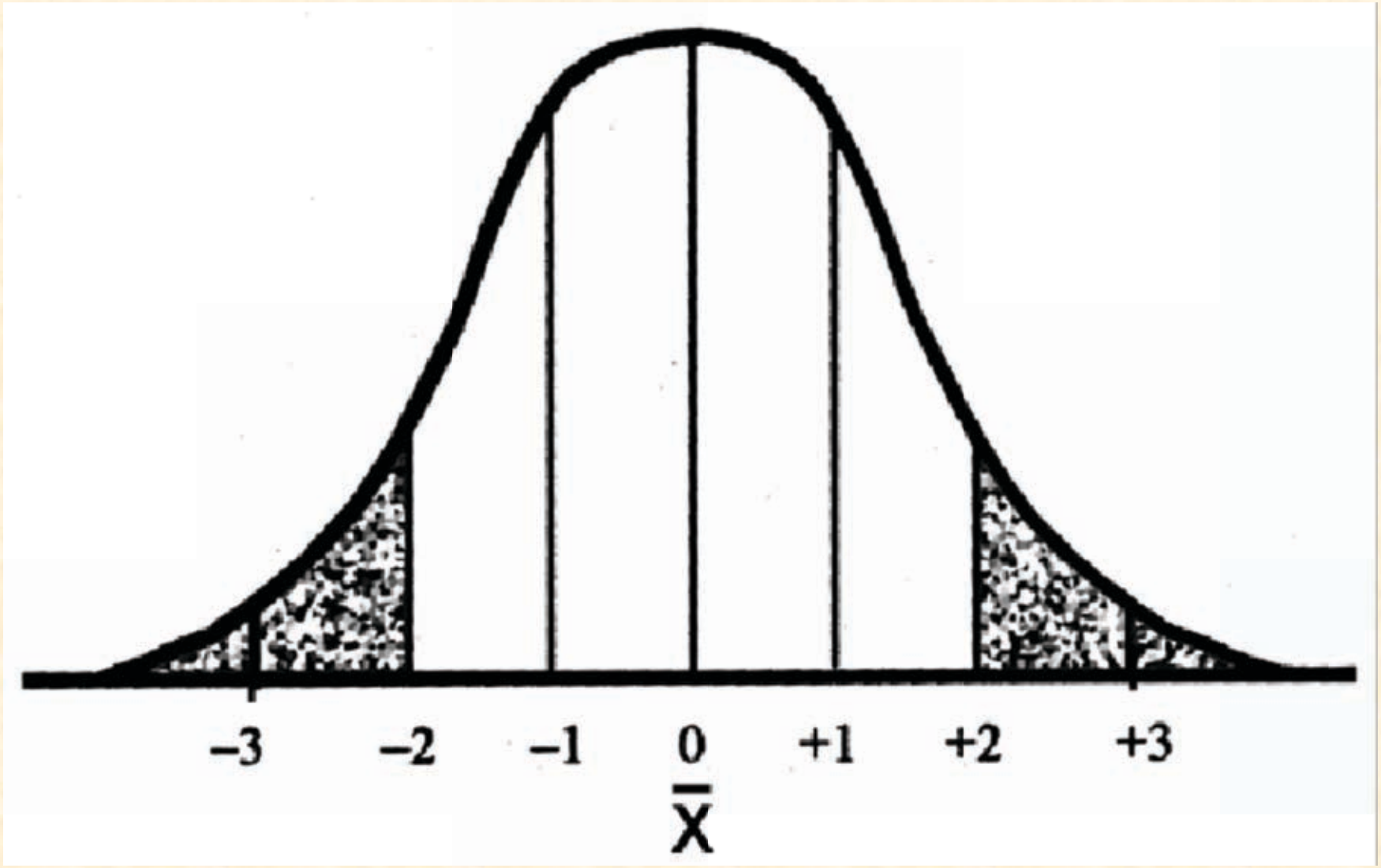
Table value for Chi Square = 9.49

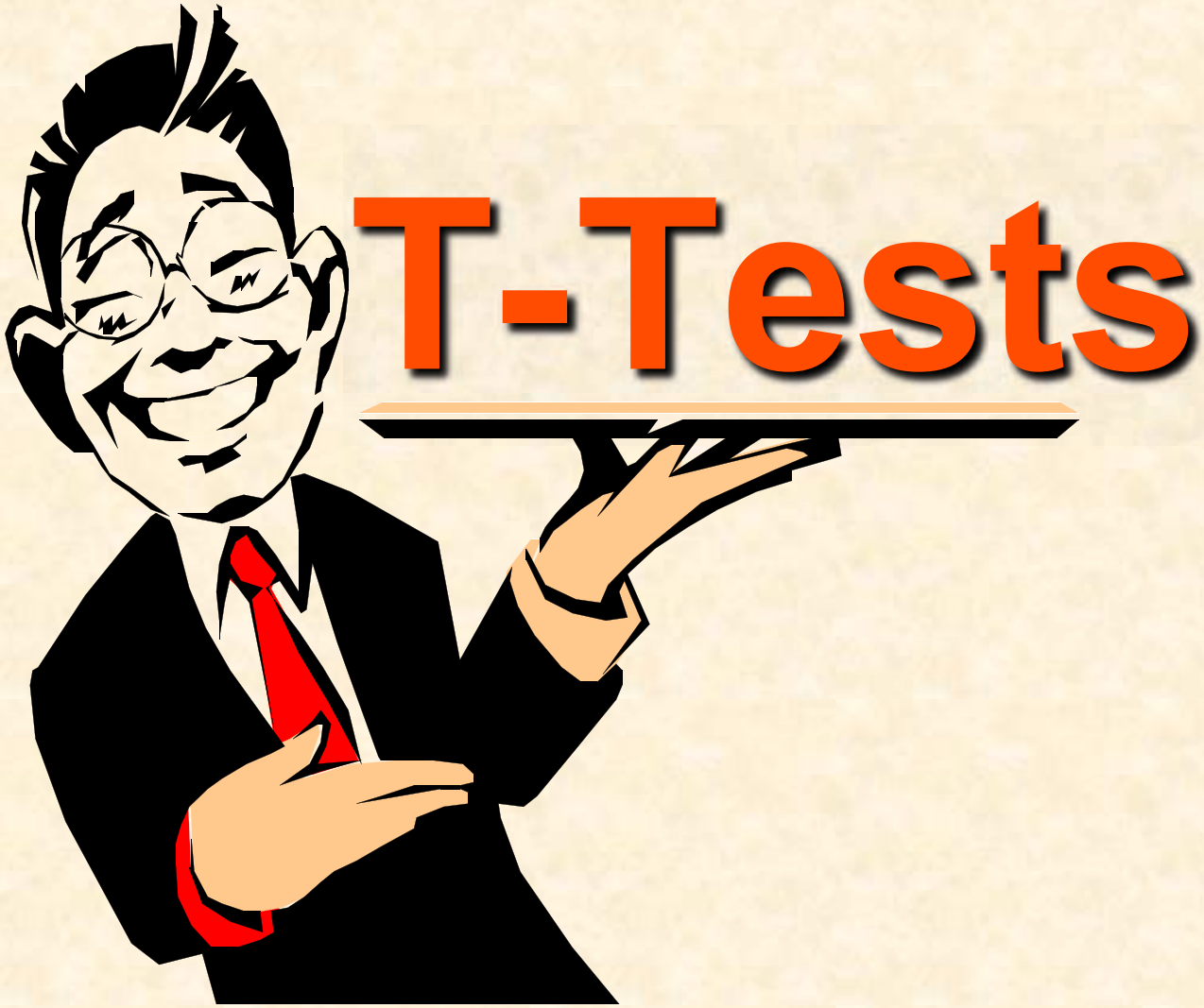
4 d f

P=.05 **level of significance**

Is there a significant difference in car preference????

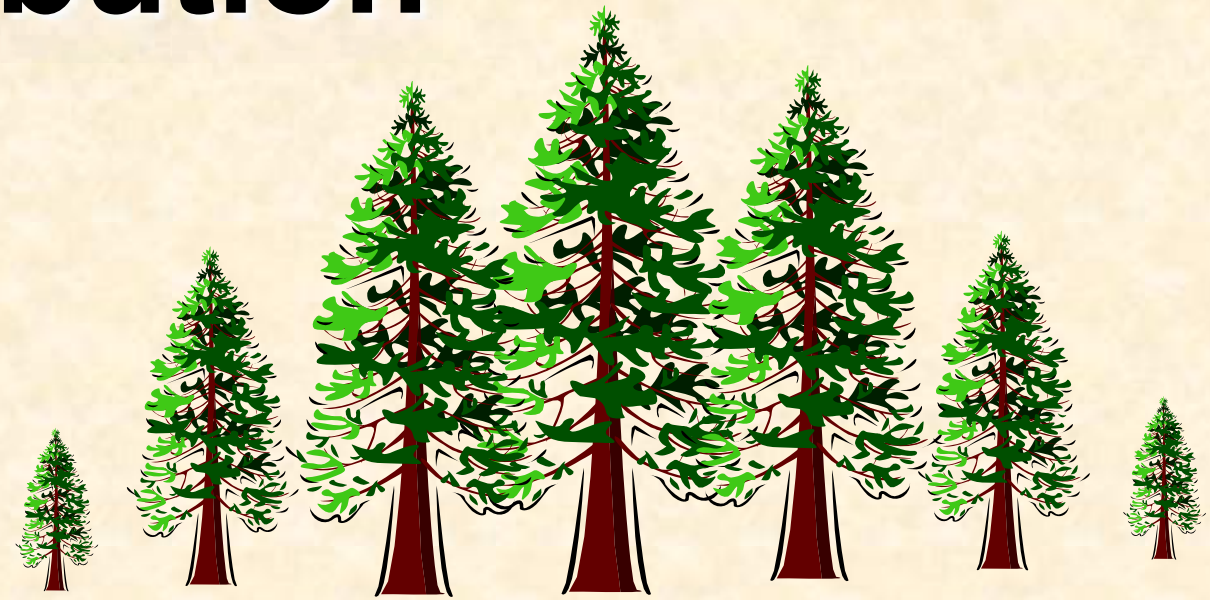
SD & the Bell Curve




T-Tests

T-Tests

For populations that *do* follow a normal distribution



T-Tests

Drawing conclusions about similarities or differences between *population means*

(χ)

T-Tests

- **Is average plant biomass the same in two different geographical *areas* ???**

- **Two different *seasons* ???**

T-Tests

- ***COMPLETELY*** **confident answer =**
  - **measure *all* plant biomass in each area**

- **Is this PRACTICAL?????**

Instead:

- Take one sample from each population
- *Infer* from the sample means and SD whether the populations have the **same** or **different** means.

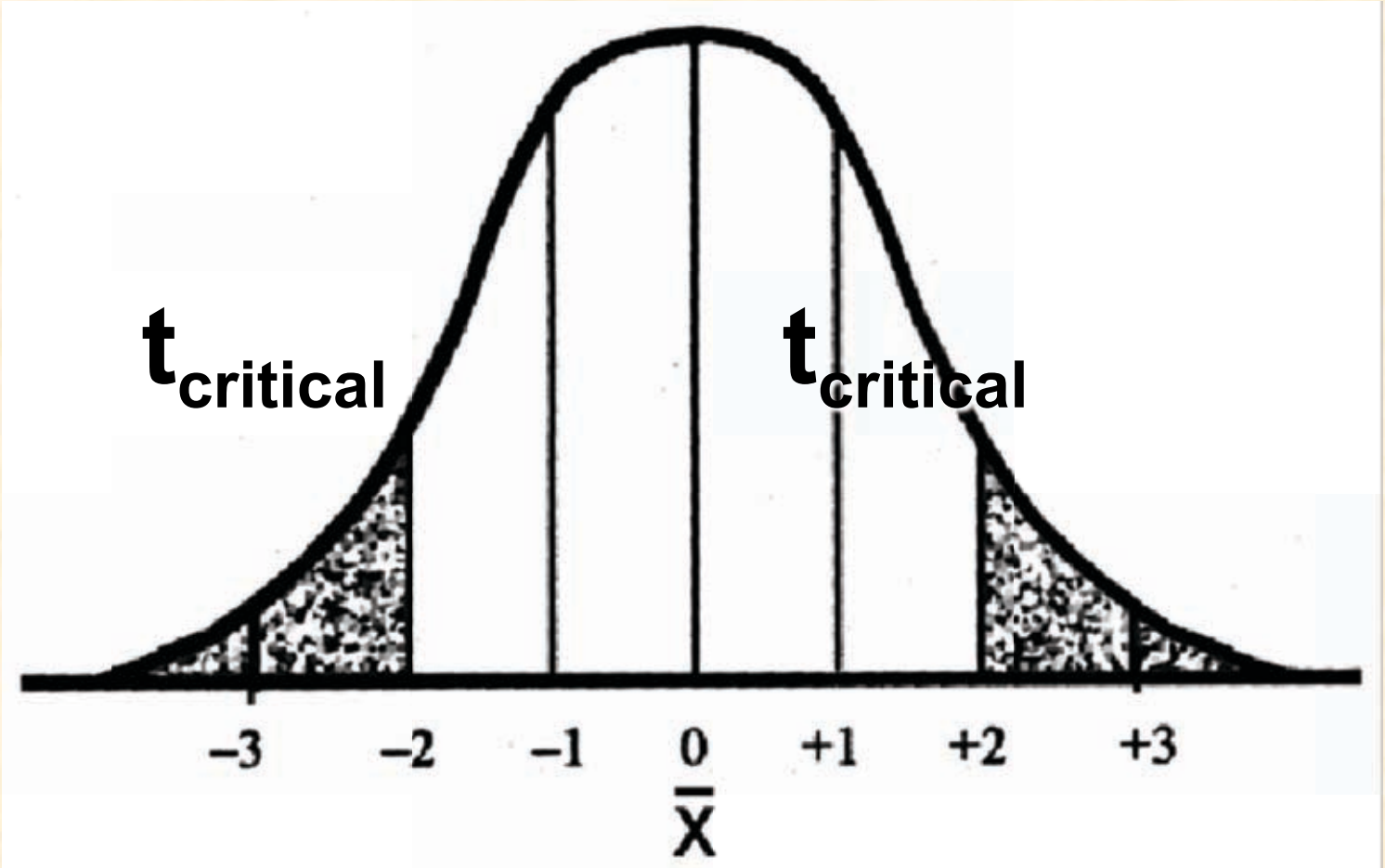
Analysis

- **SMALL *t*** values = **high** probability that the two population means are the same

- **LARGE *t*** values = **low** probability (means are different)

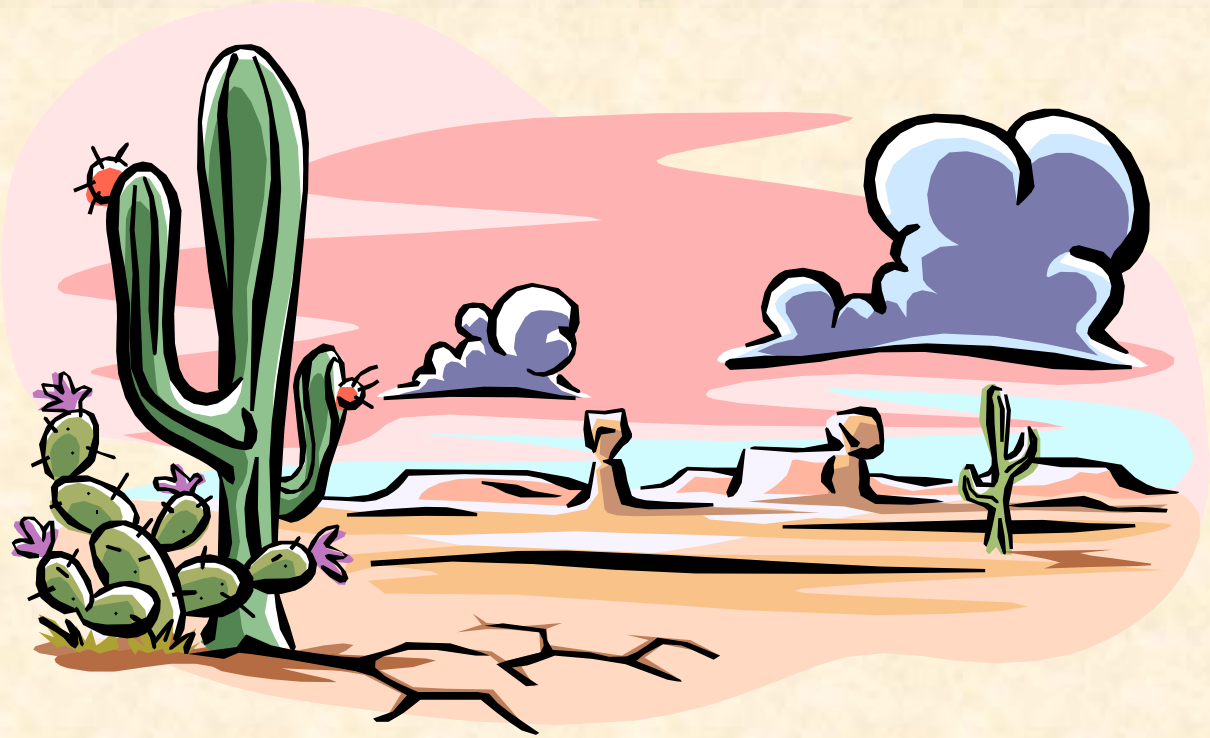
Analysis

$T_{calculated} > t_{critical}$ = reject H_o



We will be using computer analysis to perform the *t*-test

Simpson's Diversity Index

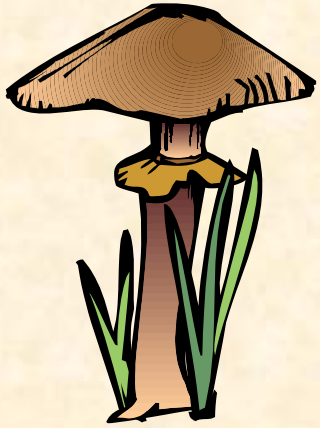


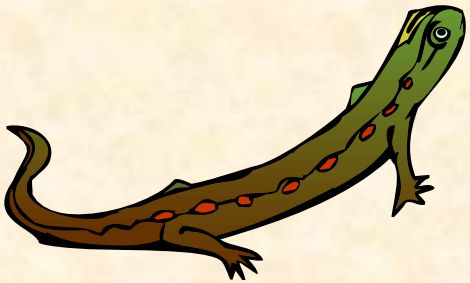
Nonparametric Testing

- For populations that ***do NOT*** follow a normal distribution
  - includes **most wild populations**

Answers the Question

- **If 2 indiv are taken at RANDOM from a community, what is the probability that they will be the SAME species????**





The Formula

$$D = 1 - \frac{\sum n_i (n_i - 1)}{N (N-1)}$$

Example

Species, *N*	Abundance, n_i	Relative Abundance, P_i
1	50	50/85 = 0.588
2	25	25/85 = 0.294
3	10	10/85 = 0.118
N = 3	n = 85	

Example

$$D = 1 - \frac{50(49)+25(24)+10(9)}{85(84)}$$

$$D = 0.56$$

Analysis

- Closer to 1.0 =
  - more Homogeneous community
- Farther away from 1.0 =
  - more Heterogeneous community

- You can **calculate by hand** to find "D"

- School Stats package ***MAY*** calculate it.

Designed by

Anne F. Maben

Former AP Science Coach, LACOE

for the

Los Angeles County Science Fair