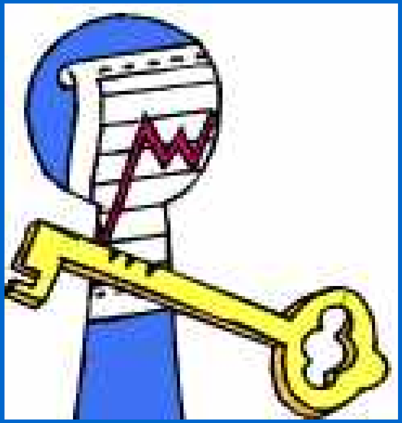
Chapter 2

Data Management and Statistical Techniques

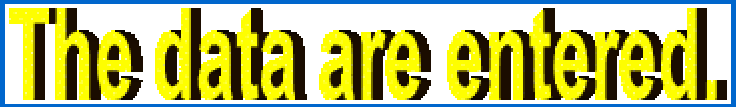
2.1 Introduction

- Manager's responsibility
  - enumerate change
  - assess management actions
  - quantify human influences
- Need statistical tools for these jobs



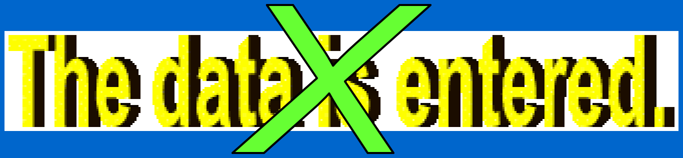
Special Note: Data is the plural form of datum

- so one says, "The data are..."



The data are entered.

- not "The data is..."



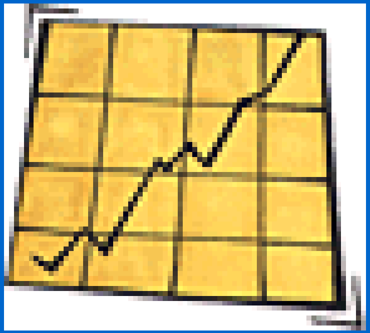
The data is entered.

Audience, Scope, and Limitations

- Always see statistician before data collection
- "Will data answer my question?"

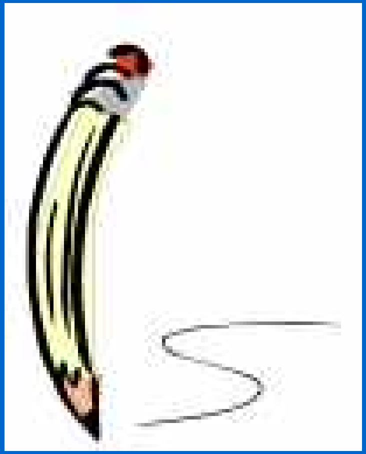
Chapter Covers...

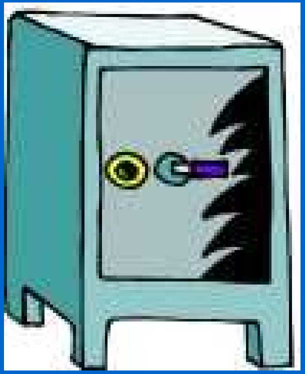
- data collection in the field
- computer management
- overview of stats
- graphing data
- interpretation of data with statistics



2.2 Data Handling and Database Management

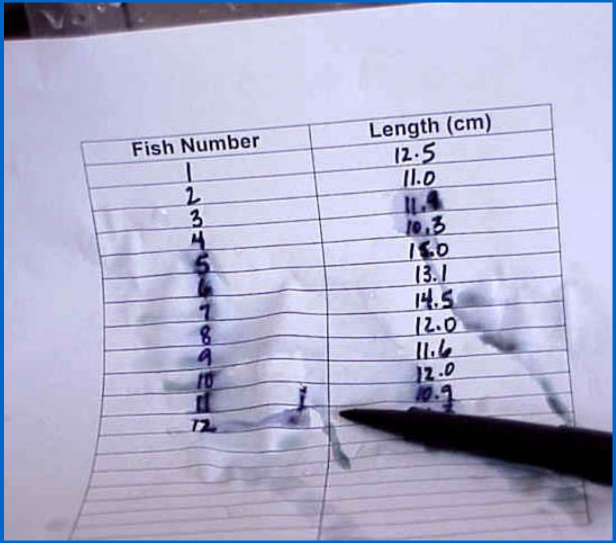
- data are expensive to collect so
  - record accurately
  - keep it safe
  - quickly if possible





Field data sheets are standardized by study

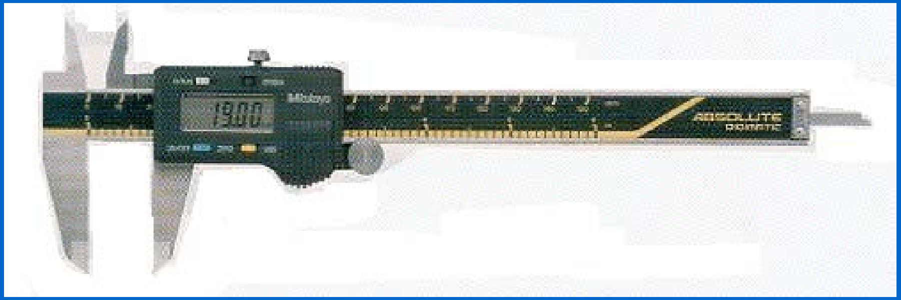
- print on waterproof paper
- write with pencil, ink will run
- write legibly, you may not be one reading
- copy data sheets asap



Fish Number	Length (cm)
1	12.5
2	11.0
3	11.4
4	10.3
5	15.0
6	13.1
7	14.5
8	12.0
9	11.6
10	12.0
11	10.9
12	[illegible]

When possible, make use of new technology

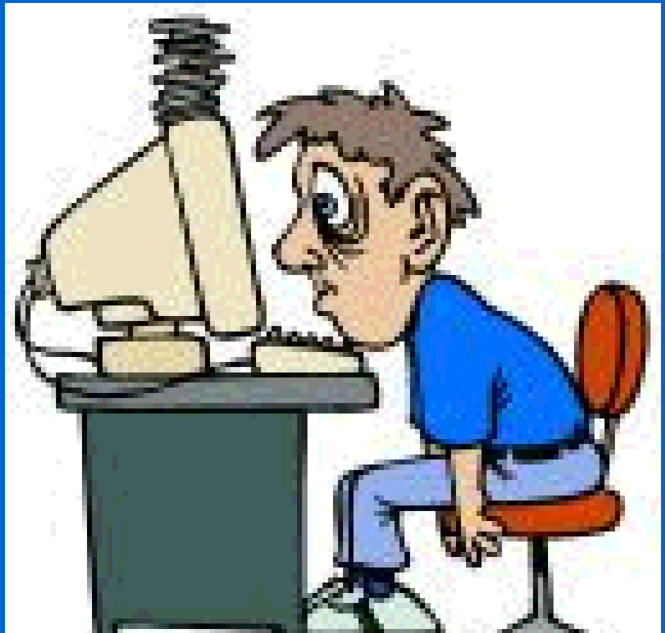
- electronic measuring boards
- digital calipers



- laptop notebooks and dataloggers
- check to be sure data are being recorded

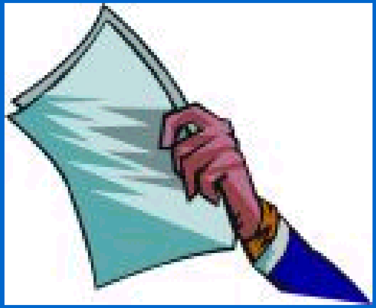
Data Management

- Natural resource agencies use databases. So...
- Biologists need to understand databases
- Also how to enter and retrieve data



Databases are

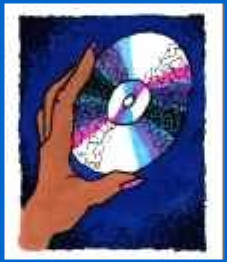
- repositories of information
- logically organized
- facilitate retrieval of specific information
- provide for customized output reports

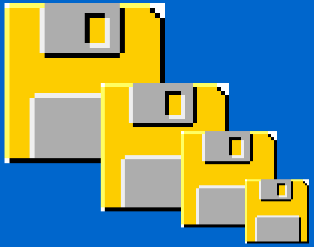


Examples of databases include

- for PC
  - dBase IV
  - Paradox
  - Access
  - Double Helix
- for mainframes
  - Oracle

Storage Considerations



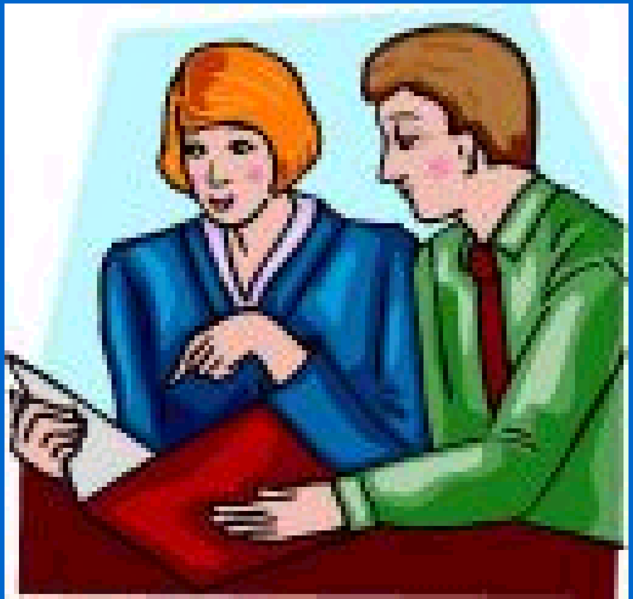


- floppies degrade after 5-10 years
- CDroms may degrade after 30 years
- ALWAYS MAKE BACKUPS
  - daily, weekly, monthly
- old technology becomes obsolete (5 1/4" floppies)

Error management

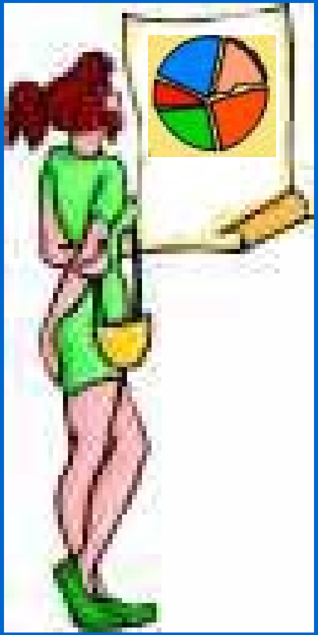
QC

- what quality control exists?
- are data within believable ranges?
- check printouts by hand
- use two people to proofread



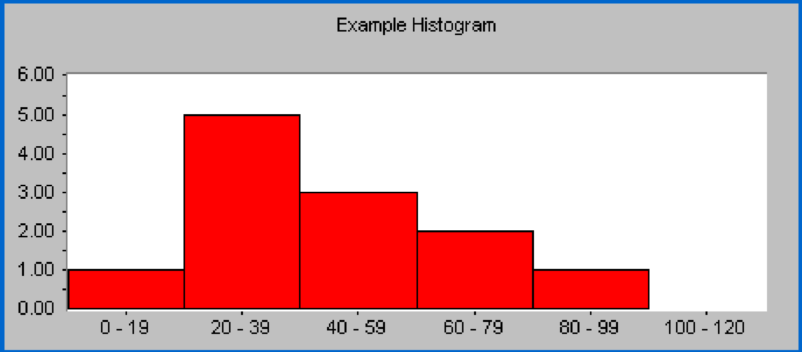
2.3 Data Visualization (i.e. graphs)

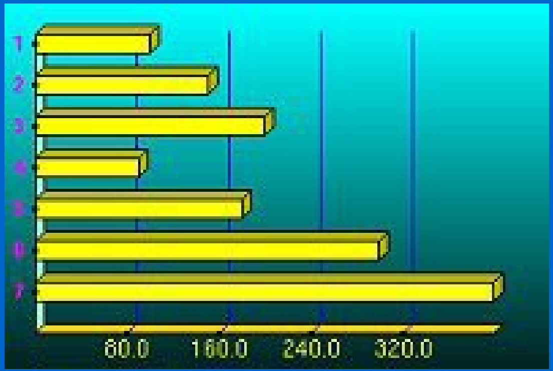
- display all original data
- picture worth 1000 numbers
  - pie chart
  - bar chart
  - histogram (vertical or horizontal)
  - scatter plot
  - line graph
  - (for rules see Box 2.1 pg 23 of text)



Histograms and Bar Charts

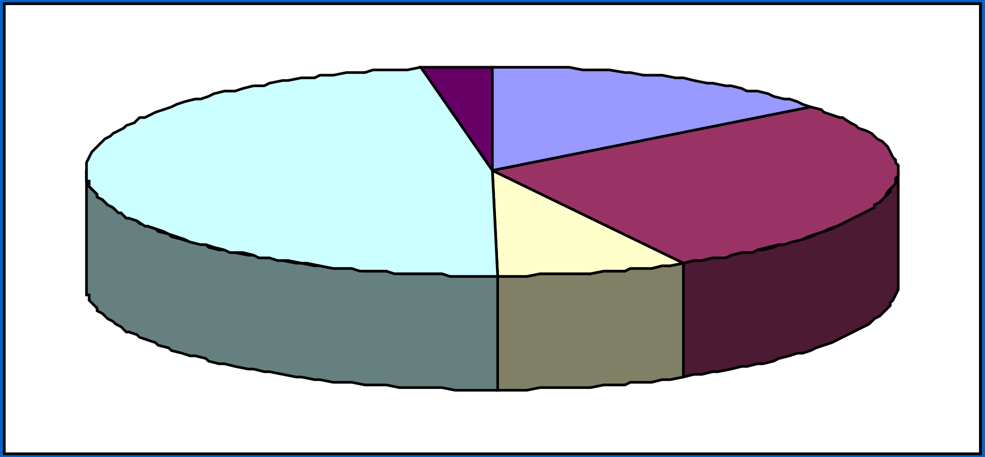
- Histogram
  - – for continuous data
  - – length-frequency data
  - – watch out for bin size bias
- Bar Chart
  - – for category data





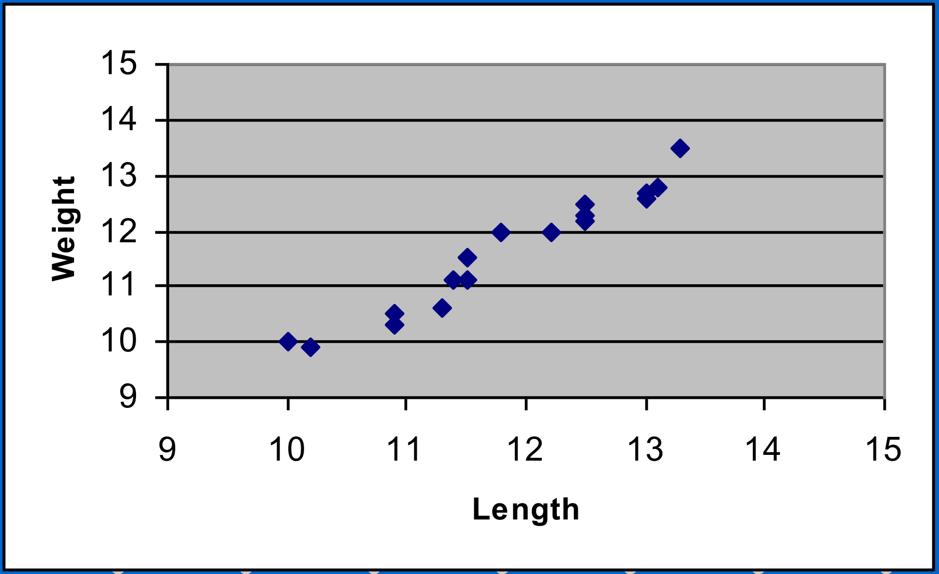
Pie Chart

- also for category data
- like diet components
- size of slice equals relative contribution

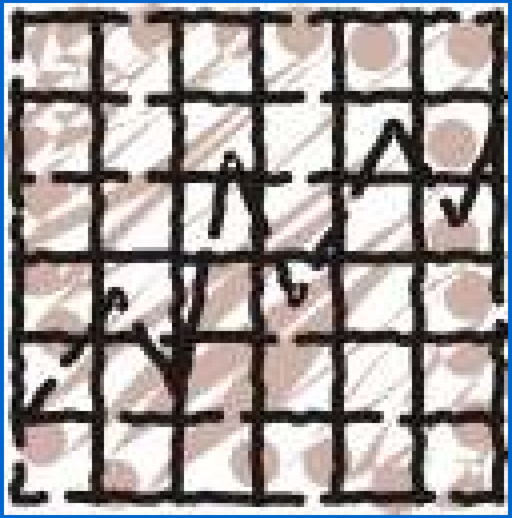


Scatter Plots

- show relation between X and Y
- X (independent variable) on horizontal axis
- Y (dependent variable) on vertical axis
- examples:
  - – length-weight
  - – spawners-recruits
  - – effort-yield



Line Graphs



time

- for ordered data
- time-series with time on X-axis

2.4 Data Terminology and Characteristics

- data set = entire collection of numbers
- case = row of closely associated variables
  - example: L, W, age of single fish
- variable = column describing an attribute of each case
  - example: sex of each fish

Fish	Length	Weight	Age
1			
2			
3			
4			
5			

Qualitative and Quantitative data

- qualitative = category data
  - nominal (sex, species)
  - ordinal (ranked data)
- quantitative = numerical data
  - discrete (integers example:age)
  - continuous (not integers example:length)

Precision, Accuracy, and Bias

- precision = how tight is pattern on shotgun blast?
  - tighter means more precision

- accuracy = how close is pattern to center of bull's eye
  - closer means more accuracy

- bias = consistent inaccuracy

Significant digits

- Minimum accuracy = range / 30
- Maximum accuracy = range/300

3.14159562

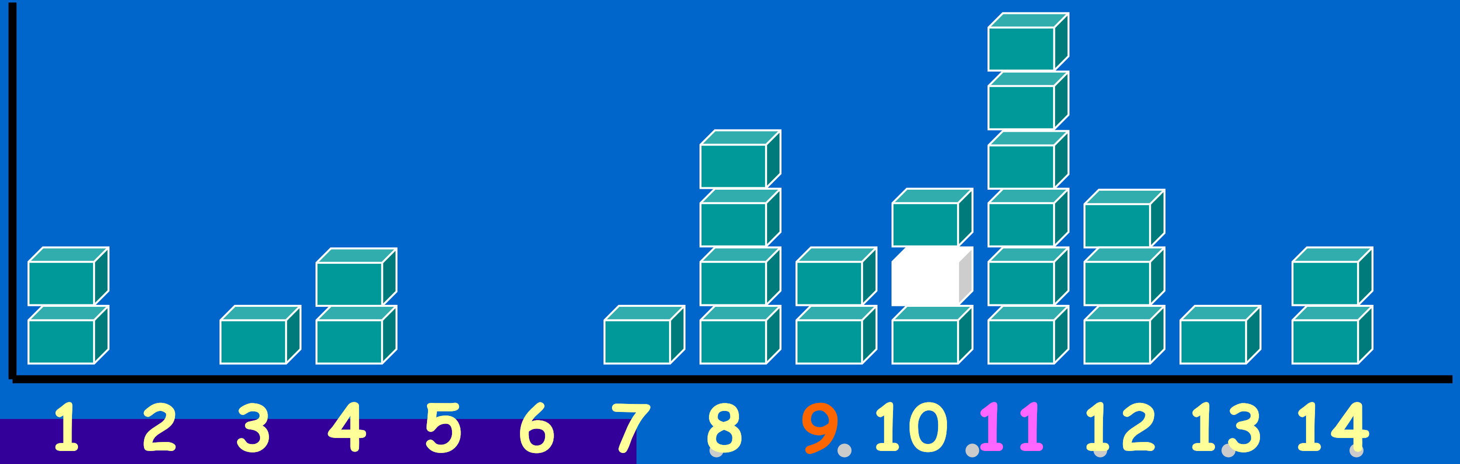
2.5 Statistics

- Analyzing and Interpreting data
- Inferences from a sample to the population

$$\frac{\text{100 Tag Returns}}{\text{500 Tagged Fish}} = \text{Population Exploitation of 20\%}$$

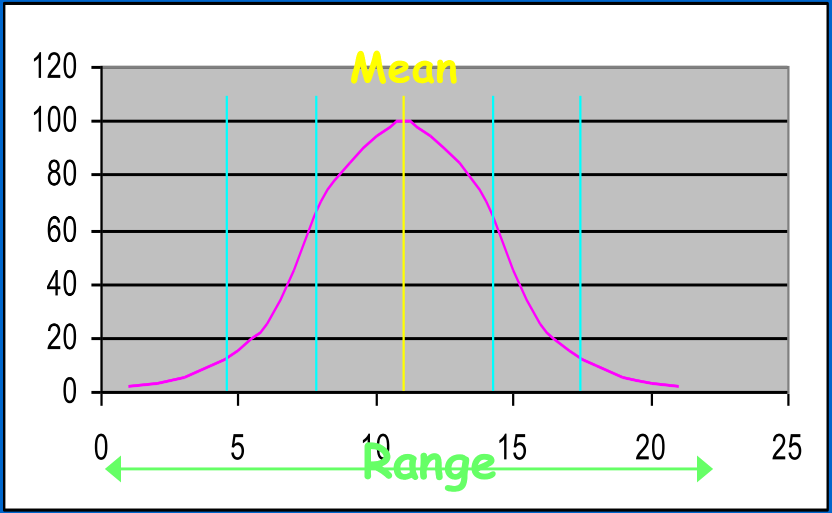
Descriptive Statistics

- summarize lots of measurements
- measures of central tendency
  - mean = arithmetic average
  - median = middle value
  - mode = value occurring the most



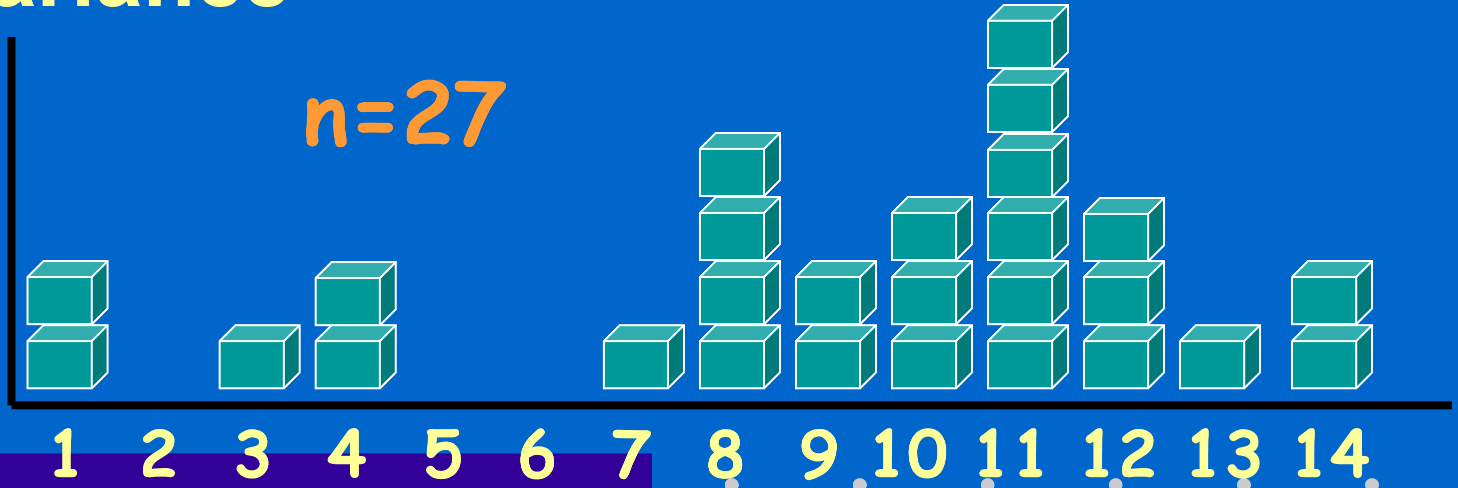
Descriptive Statistics (cont.)

- measures of dispersion
  - range = max - min value
  - variance = sum of squared deviations from sample mean
  - standard deviation = square root of variance
  - standard error of mean = standard deviation divided by root of sample size



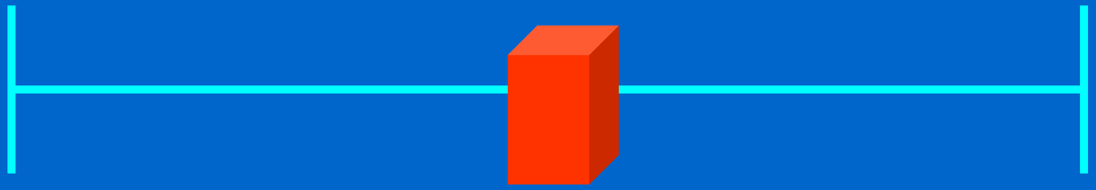
Degrees of Freedom

- number of independent observations in data set
- n-1 where n = number of observations
- increased degrees of freedom reduces variance



Confidence Intervals

- sample average rarely equals population mean
- express estimate as a range of values
- average plus/minus Student's t (n-1 df) times standard error of mean

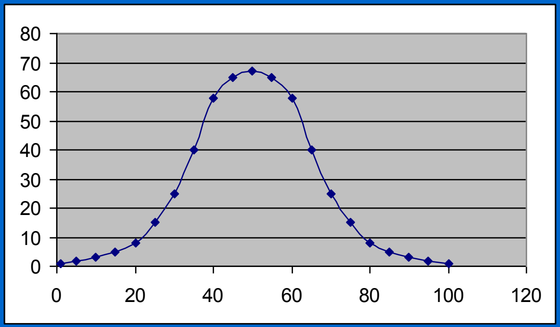


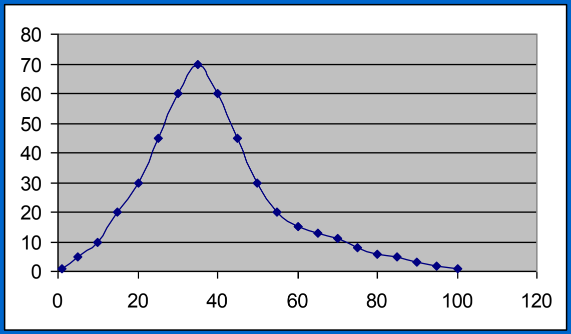
Measures of Precision

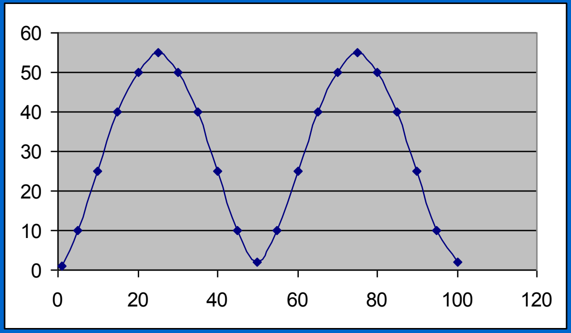
- coefficient of variation = standard deviation divided by sample mean times 100
- reported in percent

Distributions

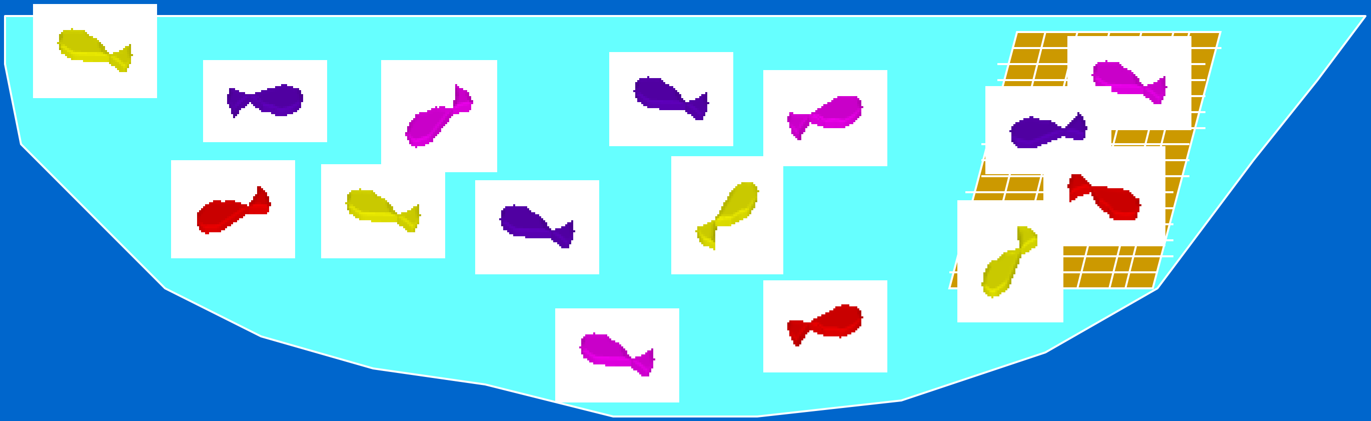
- normal - bell shaped curve
- skewed - data clumped to right or left
- bimodal - two peaks in the range of data





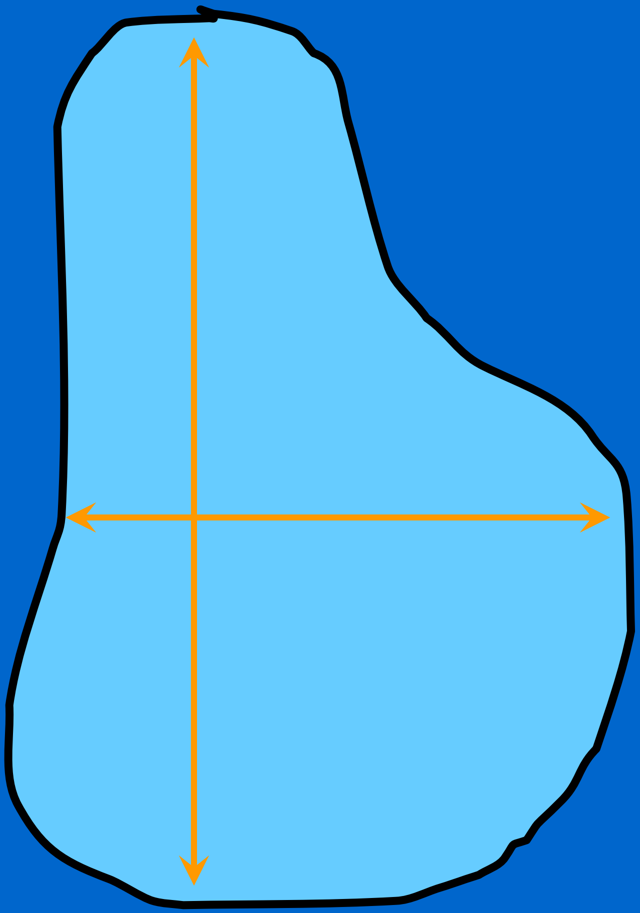


Populations and Samples



- population = all the elements under investigation
- sample = some of the elements
- biological populations sometimes change because fish migrate

Sampling Design Considerations



- size of the sampling area
- sampling units in each sample
- location of sampling units in sampling area
- selection of the sampling unit
- cost/time

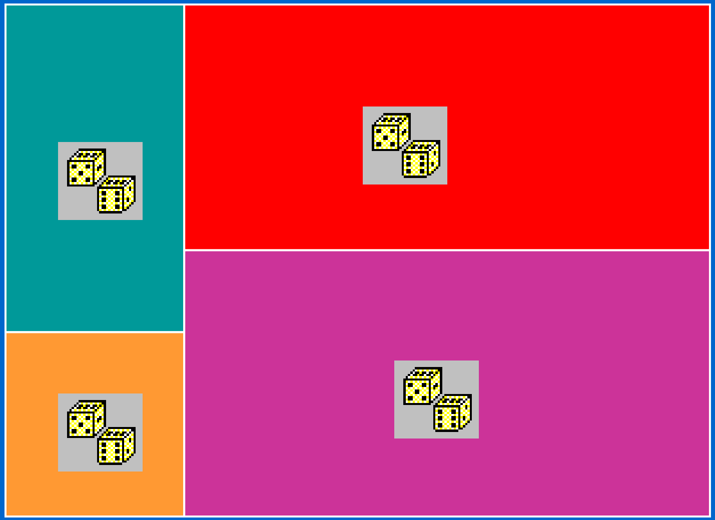
Random sample

- every member of the population has equal opportunity to be sampled

- with or without replacement
- random number table

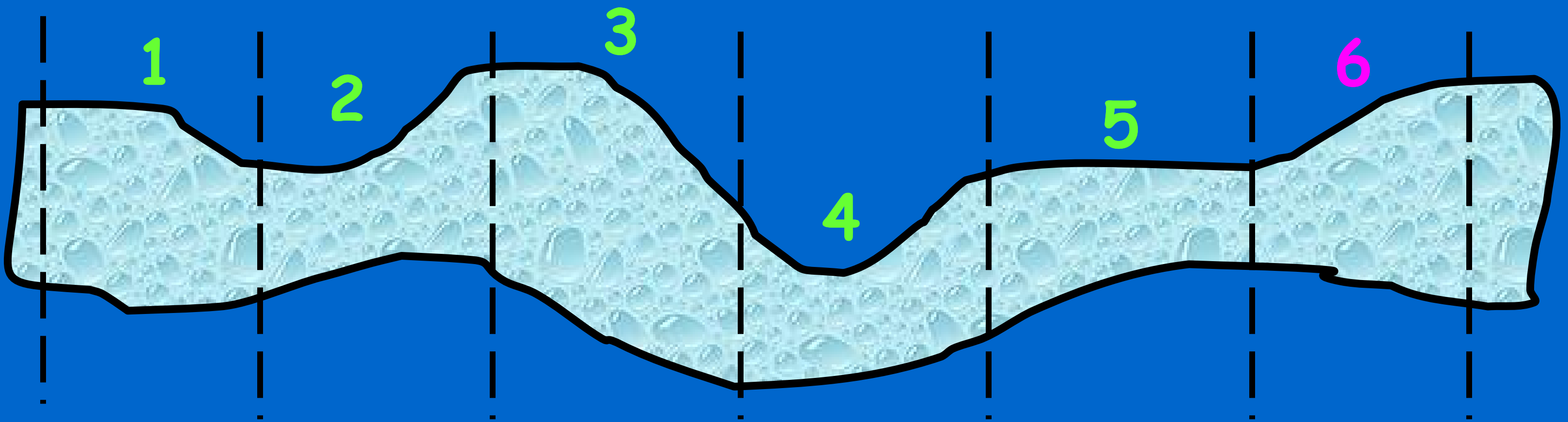
Stratified random sample

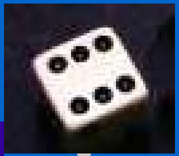
- random samples from subdivisions of populations
- subdivisions are strata based on some unifying characteristic
- account for sources of variation among samples
- strata are homogeneous



Cluster sampling

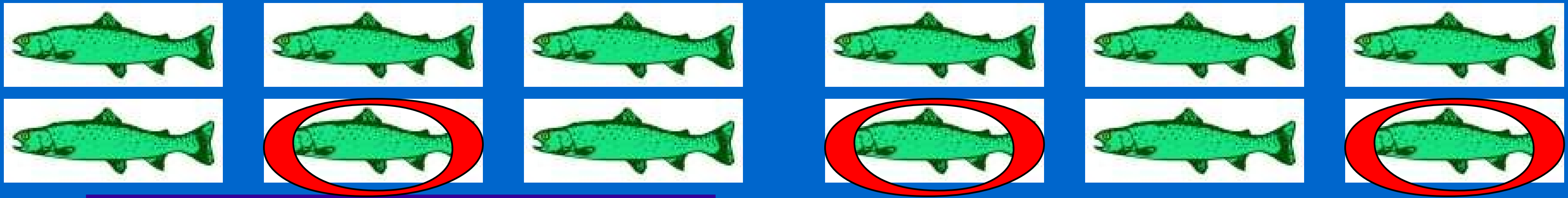
- determine sampling sites
- choose a site randomly
- take all the samples from a single site





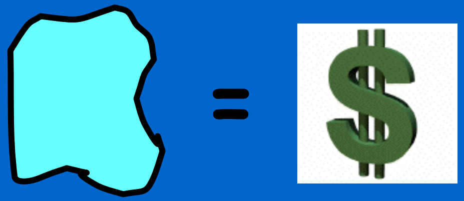
Systematic sampling

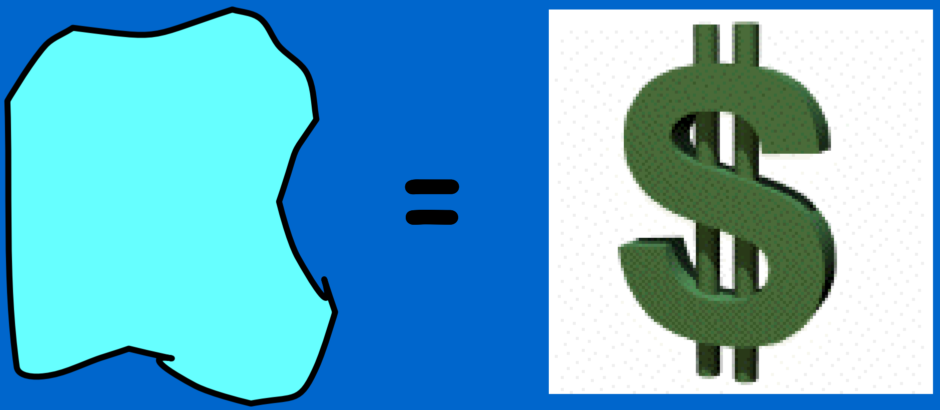
- select sampling units at regular intervals
- examples:
  - – sample every fifth 100-m section of a stream
  - – measure and weigh every 4th fish from a population



Sample Size

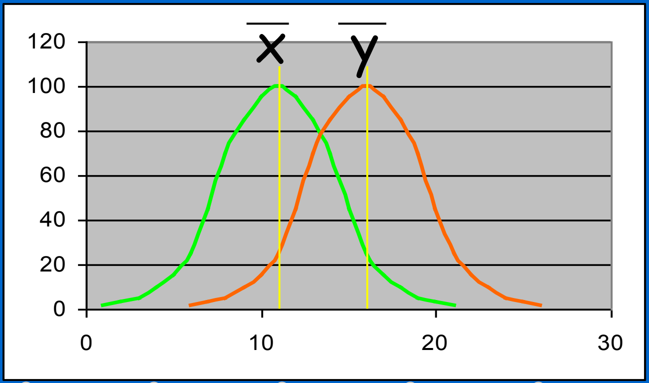
- larger the better, money and time constraints
- stepwise determination (5, 10, 15,...) till mean and CI are stable
- usually n > 30





Inferential Statistics and Hypothesis Testing

- null hypothesis... no difference in pop means
- two-sided alternative hypothesis... yes difference in pop means
- one-sided alternative hypothesis... pop1 > pop2 or vise versa
- the smaller the P-value the more likely that null hyp. is wrong

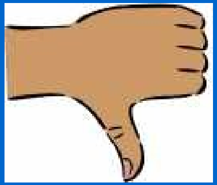


$\bar{x} = \bar{y}$

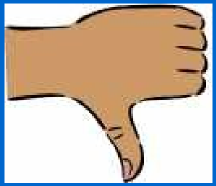
Levels of significance

$P > 0.05$	not significant
$0.01 < P < 0.05$	significant
$0.001 < P < 0.01$	highly significant
$0.0001 < P < 0.001$	very highly sig.

Statistical Errors



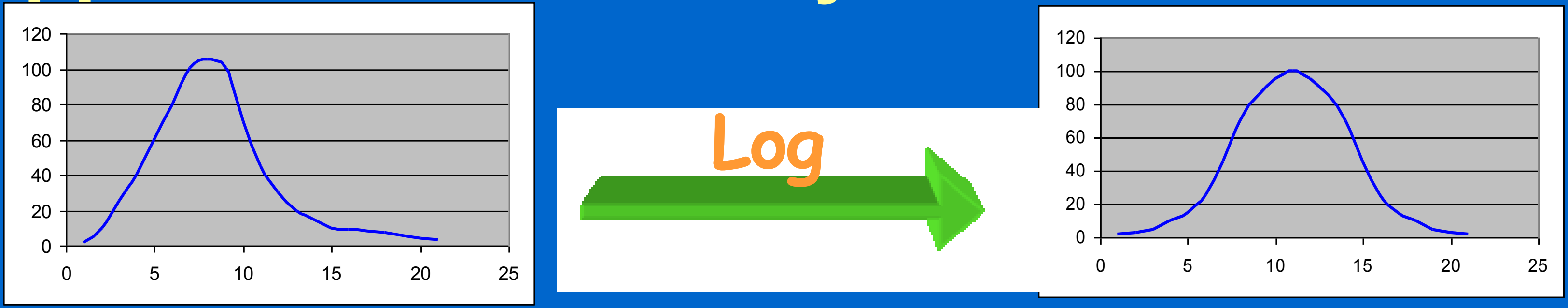
- Null hyp. true but we reject - Type I error (probability = alpha)



- Null hyp. false but we accept - Type II error (probability = beta)
- Power of the test = (1-beta)

Nonparametric and Parametric Tests

- parametric tests assume data distributed normally
- non-parametric tests are distribution-free, uneffected by outliers
- non-normal data might be transformed to approximate normality



Basic Inferential Tests of Significance

- t-Test - are two means different?
- paired t-Test - are means of paired data different?

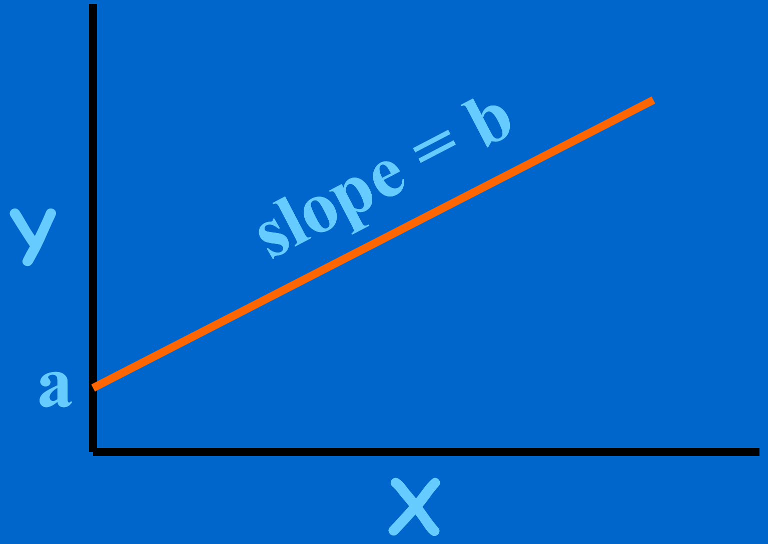
$A \overset{?}{=} B$

- anova - are any of a group of means different from the others?

A = B = C = D

- Chi-square test - does observed freq. dist. differ from expected freq. dist.?

Regression Analysis and Measures of Association



- linear regression - are two variables related according to y = a + b x
- correlation coefficient - ranges from

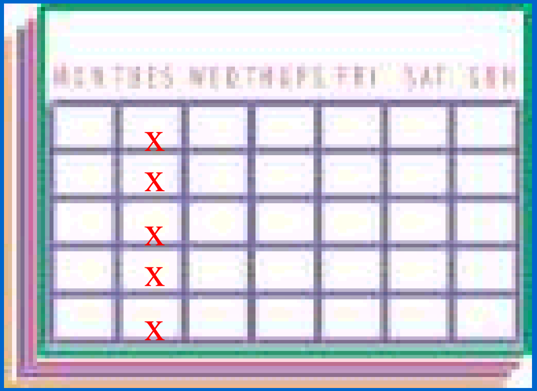
-1 completely opposite to +1 completely similar

- geometric mean regression - central trend line = slope/corr. coef.

Data transformations

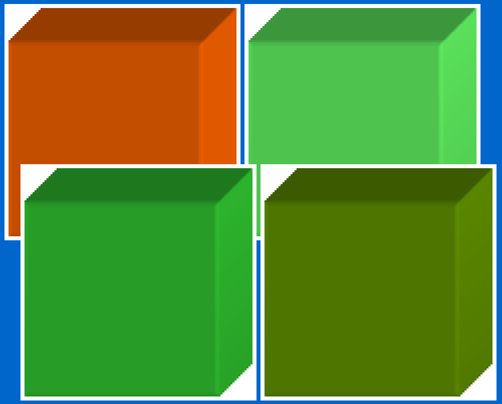
• log10	$\log(x)$
• log e	$\ln(x)$
• square	x^2
• square root	$\sqrt{x}$
• sin	$\sin(x)$
• cube	x^3

2.6 Critical Considerations in Study Design



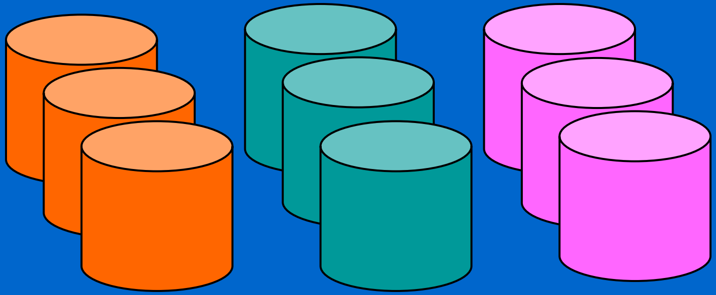
- mensurative design - passive monitoring over time or through space

- manipulative design - some variable is controlled
  - – provide at least 2 treatments
  - – one treatment is control
  - – before/after might be manipulative



Replication

- multiple experimental units per treatment
- controls error occurring in the experiment
- more precise measure of effect of treatments



- pseudoreplication
  - treatments are not truly replicated
  - replicates are not stat. independent